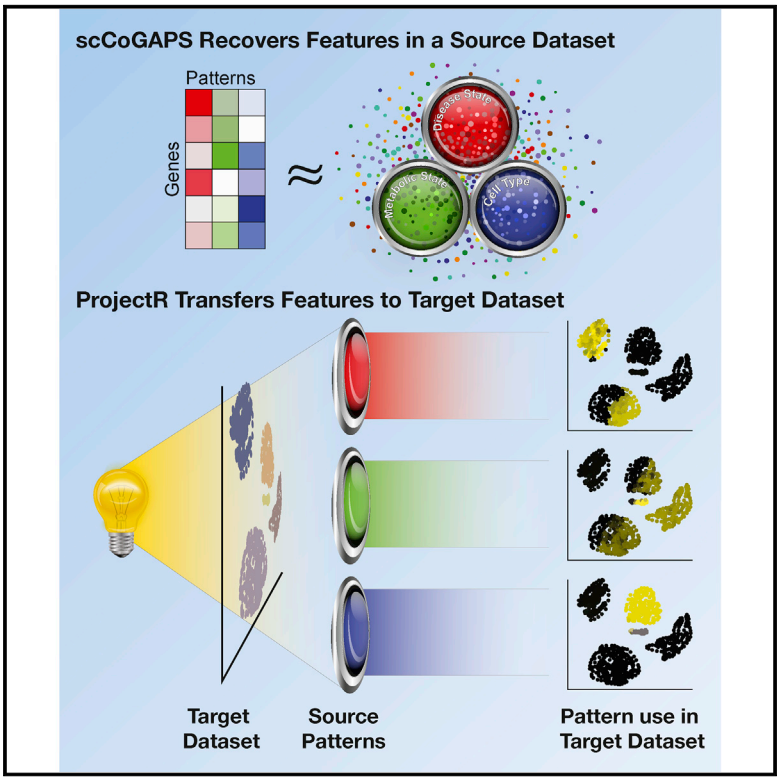
Article

Cell Systems

Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species

Graphical Abstract



Highlights

- Latent spaces provide greater insight into biological systems than marker genes alone
- scCoGAPS learns biologically meaningful latent spaces from sparse scRNA-Seq data
- Transfer learning (TL) enables discovery across experimental systems and species
- ProjectR is a TL framework to rapidly explore latent spaces across independent datasets

Authors

Genevieve L. Stein-O'Brien, Brian S. Clark, Thomas Sherman, ..., Seth Blackshaw, Loyal A. Goff, Elana J. Fertig

Correspondence

ejfertig@jhmi.edu

In Brief

We present tools and workflows for latent space exploration across datasets. scCoGAPS is an implementation of NNMF that is specifically suited for large, sparse scRNA-seq datasets. ProjectR implements a transfer-learning framework that rapidly projects new data into learned latent spaces. We demonstrate the utility of this approach for *de novo* annotation of new datasets, cross-species analysis, linking genomic regulatory and transcriptional signatures, and exploration of features across a catalog of cell types.

Stein-O'Brien et al., 2019, Cell Systems *8*, 395–411
May 22, 2019
https://doi.org/10.1016/j.cels.2019.04.004

CellPress

Cell Systems
Article

Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species

Genevieve L. Stein-O'Brien,[1,2,6,7,15] Brian S. Clark,[2,15,17] Thomas Sherman,[1] Cristina Zibetti,[2] Qiwen Hu,[13] Rachel Sealfon,[14] Sheng Liu,[5] Jiang Qian,[5] Carlo Colantuoni,[2,4] Seth Blackshaw,[2,3,4,5,10] Loyal A. Goff,[2,3,6,16] and Elana J. Fertig[1,6,7,8,9,11,12,16,18,*]

[1]Department of Oncology, Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, USA
[2]Solomon H. Snyder Department of Neuroscience, Johns Hopkins University, Baltimore, MD, USA
[3]Kavli Neurodiscovery Institute, Johns Hopkins University, Baltimore, MD, USA
[4]Department of Neurology, Johns Hopkins University, Baltimore, MD, USA
[5]Department of Ophthalmology, Johns Hopkins University, Baltimore, MD, USA
[6]McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA
[7]Institute for Data Intensive Engineering and Science, Johns Hopkins University, Baltimore, MD, USA
[8]Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, USA
[9]Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, MD, USA
[10]Center for Human Systems Biology, Johns Hopkins University, Baltimore, MD, USA
[11]Institute for Cell Engineering, Johns Hopkins University, Baltimore, MD, USA
[12]Department of Biomedical Engineering and Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA
[13]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA
[14]Flatiron Institute, New York, NY, USA
[15]These authors contributed equally
[16]Senior author
[17]Present address: John F. Hardesty, MD Department of Ophthalmology and Visual Sciences, Washington University, St. Louis, MO, USA
[18]Lead Contact
*Correspondence: ejfertig@jhmi.edu
https://doi.org/10.1016/j.cels.2019.04.004

SUMMARY

Analysis of gene expression in single cells allows for decomposition of cellular states as low-dimensional latent spaces. However, the interpretation and validation of these spaces remains a challenge. Here, we present scCoGAPS, which defines latent spaces from a source single-cell RNA-sequencing (scRNA-seq) dataset, and projectR, which evaluates these latent spaces in independent target datasets via transfer learning. Application of developing mouse retina to scRNA-Seq reveals intrinsic relationships across biological contexts and assays while avoiding batch effects and other technical features. We compare the dimensions learned in this source dataset to adult mouse retina, a time-course of human retinal development, select scRNA-seq datasets from developing brain, chromatin accessibility data, and a murine-cell type atlas to identify shared biological features. These tools lay the groundwork for exploratory analysis of scRNA-seq data via latent space representations, enabling a shift in how we compare and identify cells beyond reliance on marker genes or ensemble molecular identity.

INTRODUCTION

The identity of an individual cell is determined by the combinatorial effects of diverse biological processes. Dimension reduction techniques deconvolve gene expression data into discrete latent spaces, which may correspond to biological and technical influences on the transcriptome (Brunet et al., 2004; Cleary et al., 2017; Kossenkov et al., 2007; Stein-O'Brien et al., 2018; Wagner et al., 2016; Zhu et al., 2017). Latent space techniques are frequently used in the context of novel biological discovery from high-dimensional genomics datasets. Discovery requires evaluation of both the accuracy of the learned latent space and interpretation of biological processes from the low dimensional representation. Both of these tasks are challenging, if not entirely ineffective, using standard analytical methods, requiring biological validation to provide a gold standard (Cleary et al., 2017; Kiselev et al., 2019; Stein-O'Brien et al., 2018). However, in many applications, such a gold standard does not exist. Nonetheless, multiple datasets and measurement assays of the same biological system should reflect a similar set of biological processes. Furthermore, subsets of cellular features may further be preserved across experimental systems from related biological contexts. These properties can be utilized to improve selection, analysis, and interpretation of diverse biological systems by leveraging information learned from different data sources. Specifically, we propose that establishing the biological relevance of

latent spaces requires a 3-fold approach to (1) learn gene-expression signatures associated with biological processes, (2) demonstrate their association with specific cellular features in the dataset from which they are inferred, and (3) test their robustness across related but diverse biological contexts. These latent spaces are best learned from single-cell measures instead of bulk measurements where learned latent spaces may reflect confounded features across cell types and states. The first two steps of this process are prevalent across single-cell RNA-sequencing (scRNA-seq) analyses, but the second often relies on heuristic analysis and expert curation (Zappia et al., 2018). Transfer-learning approaches can be used to perform the last two steps, thereby enabling *in silico* validation, interpretation, and exploration across diverse types of modern high-throughput biological data.

The machine-learning subdomain of transfer learning exploits the fact that if two datasets share common latent spaces, a feature mapping between the two can identify and characterize relationships between the data defined by individual latent spaces (Pan et al., 2008). In this framework, one dataset is the source in which the latent space representation is learned, and another is the target that is mapped into the latent spaces learned in the source. The distribution, domain, or feature space of the source and target data may differ (Pan et al., 2008; Torrey and Shavlik, 2009). Thus, transfer-learning techniques are ideally suited to assess shared latent spaces from one or more sources. Once the robustness of a biological process is established across systems, these approaches can also be applied to use these learned latent spaces to enable exploration of process use across data platforms, modalities, and studies. The established conservation of specific biological processes across systems, such as common developmental pathways across tissues or organisms, can be further leveraged to enable cross-study validation. In this case, the low-dimensional patterns learned from latent space techniques will be shared in samples with biologically meaningful relationships between datasets, while dataset-specific factors and technical artifacts across datasets will not. The challenge then arises in providing a computational tool to enable this *in silico* validation.

We have adapted a transfer-learning approach for high-throughput genomic data analysis with two new methods, scCoGAPS and projectR. These tools provided a framework enabling the identification, evaluation, and exploration of latent-space features in both source and target datasets. To demonstrate this workflow across a variety of contexts, we apply these tools to a time course scRNA-seq dataset from murine retina development and demonstrate recovery of meaningful representations of biological features within individual latent spaces. Application of scCoGAPS identified gene-expression signatures of discrete cell types and biological processes associated with cell-cycle regulation, neurogenesis, and cell-fate specification. We empirically evaluate our transfer-learning approach across a diverse collection of single-cell datasets. In addition to performance assessment, these analyses also demonstrate a wide range of biological applications. We demonstrate how to classify learned cell types in a previously published adult retina scRNA-seq dataset via projectR projection (Macosko et al., 2015). We further illustrate how transfer learning can be used to extract meaningful biological insights across experimental modalities and species by projecting a bulk RNA sequencing (RNA-seq) human retinal development time course (Hoshino et al., 2017) and a mouse bulk Assay for Transposase-Accessible Chromatin for Sequencing (ATAC-Seq) dataset into the learned latent spaces from a developing mouse retina scRNA-seq dataset. To highlight the ability of projected patterns to recover related biological processes and cell types across developmentally related systems, we compare pattern usage between the developing mouse retina and two independent datasets derived from the developing cortex (Nowakowski et al., 2017; Zhong et al., 2018) and another from the developing mouse midbrain (La Manno et al., 2016). Finally, to examine the power of pattern exploration via transfer learning, we identify shared cellular features across a large collection of single cells from an atlas of mouse tissues (Tabula Muris Consortium et al., 2018). In aggregate, these analyses highlight the diversity of potential applications for transfer-learning approaches to rapidly identify and describe related components between a source dataset, in this case derived from the developing mouse retina, and a variety of independent data sources using learned latent spaces.

Using a collection of latent spaces, learned from a dataset of single-cell gene expression estimates, we demonstrate the utility of a combined reduced dimensional representation and transfer-learning approach to identify shared cellular attributes and biological processes across diverse data types in a manner that avoids the complications of normalization or sample alignment. Our approach is able to annotate latent spaces and reveal novel parallels between different tissues, molecular features, and species. Our approach demonstrates that projectR can rapidly transfer annotations, classify cells, and identify the use of biological processes without *a priori* knowledge or annotation within the source dataset. While we focus this application on low dimensional factors learned with scCoGAPS, projectR generalizes as an exploratory analysis and biological interpretation method for other dimension reduction techniques that find latent spaces associated with continuous gene weights.

RESULTS

Adaptive Sparsity for Learning Factors from scRNA-Seq (scCoGAPS): Theory

ScCoGAPS is a non-negative matrix factorization (NMF) algorithm. NMF algorithms factor a data matrix into two related matrices containing gene weights, the Amplitude (**A**) matrix, and sample weights, the Pattern (**P**) matrix (Figure 1A). Each column of **A** or row of **P** defines a factor, and together, these sets of factors define the latent spaces amongst genes and samples, respectively. Each sample-level relationship in a row of the pattern matrix is referred to as a pattern and the corresponding gene weights as an amplitude. In NMF, the values of the elements in the A and P matrices are required to be greater than or equal to zero. This constraint simultaneously reflects the non-negative nature of gene expression data and enforces additivity of factors, generating solutions that are biologically intuitive (Lee and Seung, 1999). The concept of up- or down-regulation reflects a relative difference between two conditions that can, and often is, described by comparing non-negative gene weights between patterns.

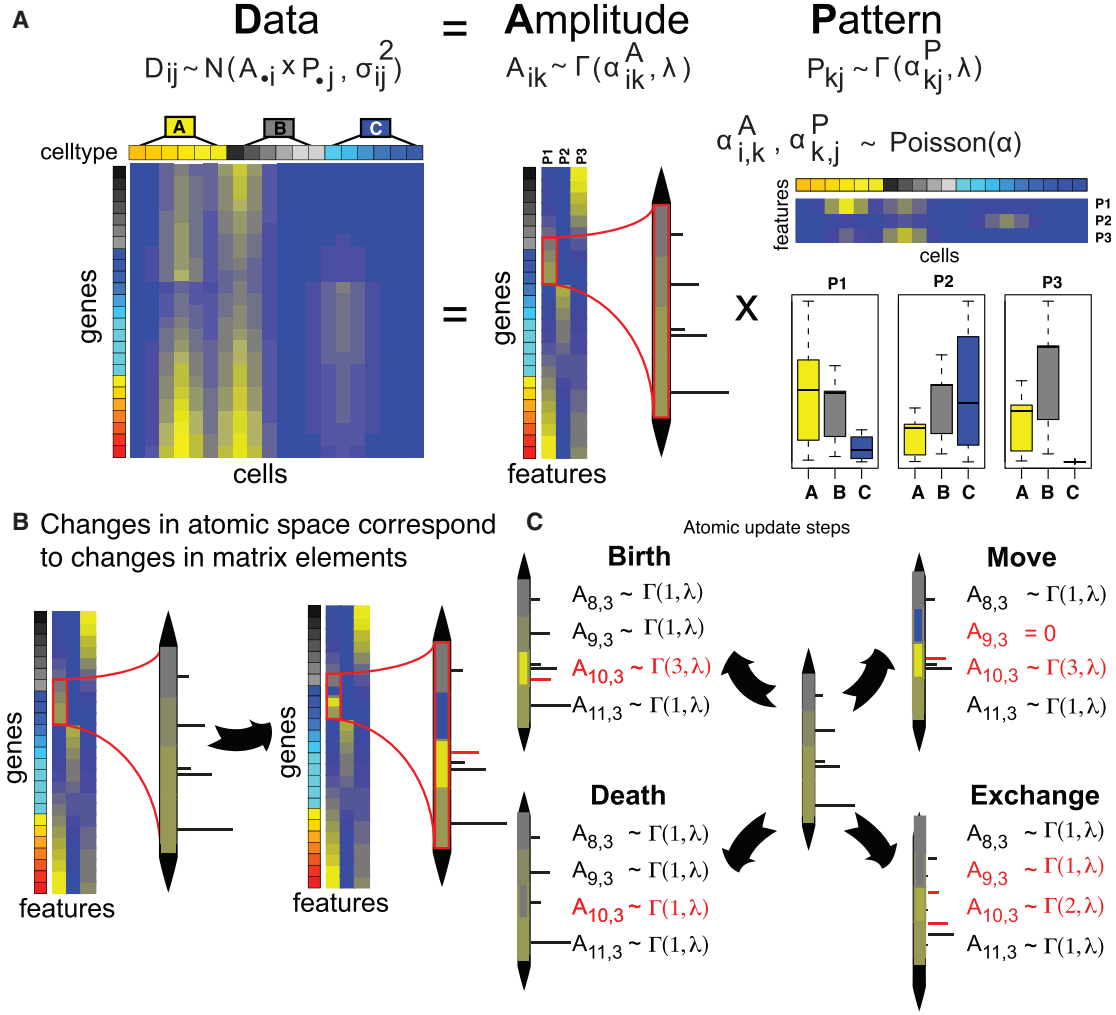


Figure 1. Mathematical Core of the scCoGAPS Algorithm

(A) scRNA-Seq data yields a data matrix that has each sample as a column and each observed gene expression value as a row. scCoGAPS decomposes the preprocessed data matrix into two related matrices. The rows of the amplitude matrix (**A**) quantify the sources of variation among the genes, and the columns of the pattern matrix (**P**) quantify the sources of variation among the cells. The matrix product of A and P approximates the preprocessed input data matrix. The number of columns of A equals the number of rows in **P** and represents the number of dimensions in the low-dimensional representation of the data. Theoretically, each column in the amplitude matrix and the corresponding row of the pattern matrix represents a distinct source of biological, experimental, or technical variation in each cell. The values in the column of the amplitude matrix then represent the relative weight of each gene and the values in the row of the pattern matrix its relative role in each cell. Adaptive sparsity is achieved by placing a Poisson prior on the shape parameter in the gamma distribution for each matrix element $(\alpha_{Aij}, \alpha_{Pij})$ and a fixed scale parameter for all matrix elements (λ_A and λ_P) in A and P, respectively. In expectation, smaller values of α_{ij} will result in smaller values of corresponding matrix element and vice versa for larger values, which will also have a decreased probability of being zero.

(B) Each iteration of the Markov Chain Monte Carlo sampling employed in CoGAPS updates the atomic space, which corresponds to an update in matrix elements.

(C) There are four possible update steps to the atomic domain that preserve both the prior distribution in (A) and detailed balance: (1) birth to add an atom, (2) death to remove of an atom, (3) moving an atom from one position to another, and (4) exchanging the mass of two atoms. During the update, the probability of selecting birth or death is selected based on the Poisson prior reinforcing the adaptive sparsity. All heatmaps are colored on a blue-yellow scale, where yellow indicates higher expression values and blue lower.

Bayesian NMF techniques can embed biological and technical structure in the data in prior distributions on the **A** and **P** matrices (Kossenkov et al., 2007; Ochs and Fertig, 2012). To accomplish this for bulk data, we previously developed the Bayesian NMF Coordinated Gene Activity in Pattern Sets (CoGAPS) method (Fertig et al., 2010). CoGAPS uses an atomic prior (Sibisi and Skilling, 1996; Skilling and Sibisi, 1996) to model three biological constraints: non-negativity reflective of pleiotropy, sparsity reflective of parsimony, and smoothness reflective of gene co-regulation and smooth dynamic transitions. The atomic prior in CoGAPS is unique in enforcing a sample- and gene-specific sparsity constraint, which we term "adaptive sparsity." In the atomic prior, each element of the **A** and **P** matrices is either zero or follows a gamma distribution. Adaptive sparsity is achieved by placing a Poisson prior on the discrete shape parameter in the gamma distribution for each matrix element

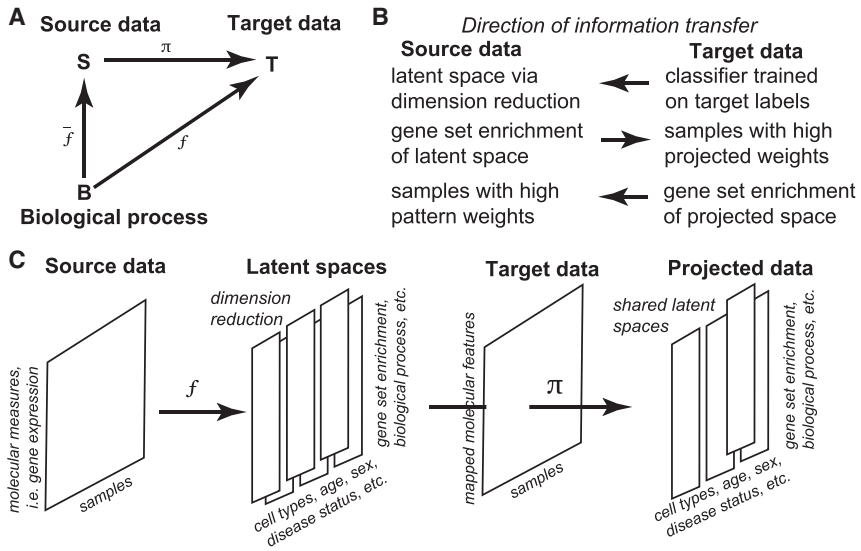


Figure 2. Theoretical Core of the projectR Algorithm

(A) Graphical representation of projection implemented in projectR showing the relationship between the learned functions, or mappings, and the datasets being operated on.

(B) Transfer-learning approaches can be adapted to reveal a variety of insights into both source and target datasets. The type and directionality of knowledge transfer enabled via projectR can vary depending on the experimental question and available annotation for each dataset.

(C) Diagram of the pipeline used to first learn latent spaces and then project them to transfer learning as described.

$(\alpha_{Aij}, \alpha_{Pij})$ and a fixed-scale parameter for all matrix elements (λ_A and λ_P) in **A** and **P**, respectively. Smaller values of α_{ij} result in smaller values of the corresponding matrix elements and vice versa for larger values. Thus, the sparsity constraint on values of latent factors will be relaxed in this model, constraining some matrix elements away from zero (Figure 1B). Adaptive sparsity can also model biological structure in the presence of the technical dropouts and true biological zeros in scRNA-seq. To accommodate the additional sparsity of scRNA-seq data, λ_A and λ_P are set as proportional to the mean of all non-zero values in the data. In contrast, λ_A and λ_P for bulk RNA-seq data are set using the means of the entire dataset. A normal prior on the data enables an empirical solution for the conditional distributions with this gamma prior, enabling efficient Gibbs sampling with this sparsity constraint (STAR Methods). This also models smoothness by grouping closely related dimensions near each other via move and exchange steps that shift a single exponential between adjacent matrix elements (Figure 1C). In practice, these steps retain the global Poisson prior on shape and the gamma prior on matrix elements while altering the shape parameters between adjacent matrix elements to model smoothness.

Parallelization and Data Structures for Cross-Validation and Efficiency: Theory

Bayesian NMF algorithms such as CoGAPS have substantial computing costs that limit their application to the large datasets generated as tissue atlases with scRNA-seq data. As we describe in the STAR Methods, representing the gamma distribution as a sum of exponentials enables efficient Gibbs sampling. We couple this representation with new data structures for their storage and corresponding calculations that are more efficient than previous versions of CoGAPS and greatly reduce the computational cost for scRNA-seq analysis (Figure S1A).

We can leverage our hypothesis that latent spaces learned from scRNA-seq data are reflective of relative gene use in biological processes to enhance the efficiency of Bayesian NMF methods. In this case, distinct subsets of cells sampled from the same condition will have similar factors in a latent space, similar to our previous observation of similar factors across distinct subsets of genes in bulk data (Stein-O'Brien et al., 2018). Inference with Bayesian NMF is parallelized for distinct subsets of cells in the input scRNA-seq data. We selected the ratio of cells in each set to enable inference of latent space factors in highly skewed distributions of samples as can occur with rare cell types. As a result, this approach is a semi-supervised method in which inference of gene weights in factors is unsupervised. Consensus factors are then created across the sets as described previously for random sets of genes (Stein-O'Brien et al., 2018). In addition to gaining efficiency, the factors estimated in parallel across subsets of cells can also be compared to enable cross-validation of the inferred latent spaces (Figure S1B).

Transfer Learning via Dimension Reduction Using projectR: Theory

In our model, known and latent factors of a biological system can be used to compare independent, biologically related datasets. This comparison is made by defining a function from the factors in one dataset and projecting an independent, biologically related target dataset into a lower dimensional space that is common to both. Projection is defined as a mapping or transformation of points from one space to another, often a lower-dimensional space. Mathematically, this can be described as a function $\varphi(x) = y: R^D \mapsto R^d$ s.t for $d \leq D$, $x \in R^D$, $y \in R^d$. The innovation of projectR is the use of a mapping function defined from the latent spaces in a source dataset, which enables the transfer of associated cellular phenotypes, annotations, and other metadata to samples in the target dataset (Figure 2).

We propose that projection of well-defined latent spaces should capture shared biology across independent datasets. In this study, we perform projection in the column space defined by the amplitude matrix from scCoGAPS (factors representing gene weights). This is accomplished by estimating the patterns **P** associated with the amplitude matrix by a generalized least-squares fit to the target data (Fertig et al., 2013a) (STAR Methods). We select this projection approach as a computationally efficient method. Moreover, the lack of the orthogonality constraint allows for greater application of the transfer-learning approach to non-orthogonal latent spaces, allowing for greater

independence of factor projections. Assuming that a given dimension is associated with a specific cellular attribute in the target dataset, the magnitude of the value in this source dataset can indicate its presence within the target dataset. Inversely, if the cellular feature is not shared across the datasets, then projection of the target data into the given latent space will have no significant value. The significance of each projected pattern can be calculated using a Wald test for each sample:latent space interaction. Depending on the distribution or number of the projected sample weights, statistical comparisons between annotated groups can be performed to quantify the presence of these inferred processes in the target data. For example, the mean projected pattern weight between two groups can be compared using standard t tests or regression-based contrasts. Additionally, classifiers can be built using the projected pattern weights, and the predictive value of each pattern assessed globally. This information transfer enables rapid and highly scalable comparison of very different datasets through the lens of a projected latent space learned in a reference dataset. This analysis can leverage the massive amount of publicly available data and their associated metadata to annotate phenotypes in source data more efficiently. Further, the ability to evaluate whether the processes described by latent spaces are shared, despite significant overall differences in the original high dimensional datasets, can enable hypothesis generation and integrated analyses.

Applications

Assessing Latent Spaces and Dimensionality: Lessons from Bulk RNA-Seq

The developing mammalian retina provides an ideal model system to evaluate the degree to which latent spaces reflect known developmental biology. Features such as discrete cell-type signatures, continuous state transitions, signaling pathway usage, developmental age, and sex should each be represented in independent latent spaces. An open question in retinal development is how progenitor cells can generate specific subtypes of neuronal and glial cell types during specific intervals during development—a phenomenon known as progenitor competence (Bassett and Wallace, 2012; Javed and Cayouette, 2017). In an effort to identify genes associated with changes in retinal progenitor cell (RPC) competence, we performed bulk RNA-seq analysis on replicate populations of fluorescence-activated cell sorting (FACS)-isolated RPCs and post-mitotic cells, which were isolated using the *Chx10:*GFP reporter (Rowan and Cepko, 2004) and assessed the fidelity of patterns learned in this bulk analysis across other experimental contexts.

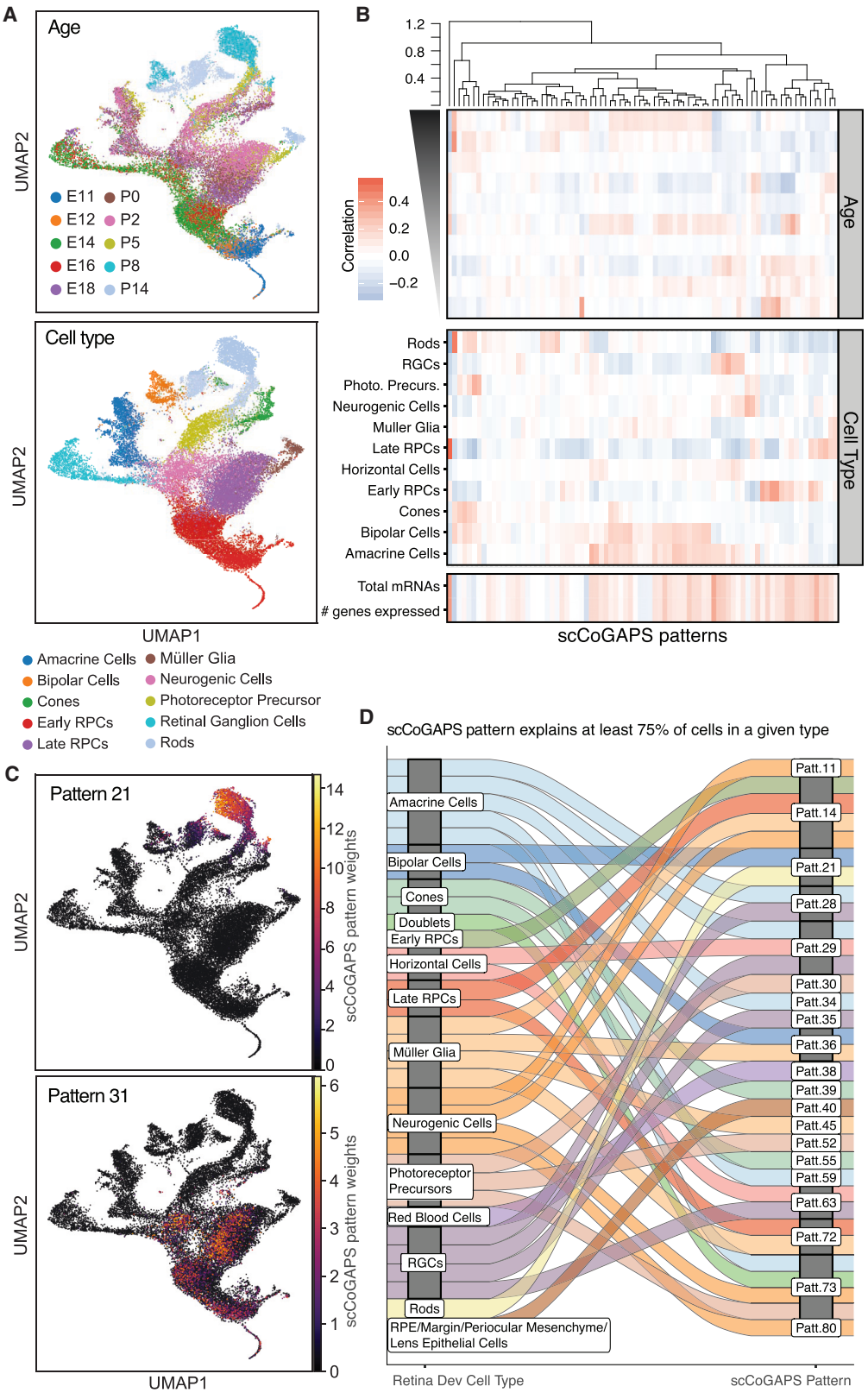
FACS-sorted *Chx10:*GFP+ RPCs and *Chx10:*GFP- post-mitotic retinal neurons (Rowan and Cepko, 2004) were collected from the developing mouse retina at three time points, embryonic day 14 (E14), embryonic day 18 (E18), and postnatal day 2 (P2), and subjected to standard bulk RNA sequencing (Zibetti et al., 2017). We applied our previous genome-wide GWCoGAPS pipeline for bulk RNA-Seq to the normalized FPKM gene expression estimates to identify a latent space consisting of 10 patterns of co-regulated genes (Stein-O'Brien et al., 2017). Dimensionality can be optimized by maximizing the robustness of patterns between dimensions (Moloshok et al., 2002). Moreover, hierarchies of cell types or subtypes can be resolved by comparing patterns across dimensions (Fertig et al., 2013a). Therefore, we applied GWCoGAPS to the bulk data using a range of dimensionalizations to identify patterns associated with specific biological features or cellular states. Final dimensionality was assessed by comparing factorizations of different dimensions using the ClutrFree (Bidaut and Ochs, 2004) algorithm (STAR Methods). Patterns were strongly correlated ($r^2 > 0.7$) between factorizations at different dimensions, indicating the overall robustness of the factors across dimensions (Figure S1C). For example, a pattern broadly associated with all retinal neurons at a lower dimensionality split into two patterns describing photoreceptors and inner retinal cells at a higher dimensionality, as assessed by correlation of cell-type specific marker-gene expression with individual patterns.

We next evaluated whether patterns identified from bulk RNA-seq could describe discrete cell-type signatures obtained from a comprehensive scRNA-seq dataset conducted across retinal development (Clark et al., 2019). In this study, we isolated 120,804 individual cells from whole mouse retina at 10 developmental time points, ranging from embryonic day 11 (E11) to postnatal day 14 (P14). scRNA-seq gene expression profiles were obtained using the 10× Genomics Chromium platform (Clark et al., 2019). To relate the datasets, the scRNA-seq data was projected into the factors learned from the bulk RNA-seq (Table S1) using projectR (STAR Methods). Using the expert-curated cell-type annotations for each single cell, a random forest classifier was trained using projected sample weights as features. Sensitivity and specificity scores were calculated for the relationship between each bulk factor and the annotated cell types detected using scRNA-seq.

While few patterns had high AUC values for specific cell types, most had moderate values spread across multiple lineages (Figure S1D). One potential explanation for this is that features shared across multiple cells types might dominate the latent spaces found at lower dimensionalization. This finding is consistent with observation that highly expressed genes tend to dominate differential expression analysis in bulk RNA-seq (Ching et al., 2014). An alternative hypothesis is that latent spaces learned in aggregate bulk measures may not cleanly define discrete cell types or states. As bulk RNA-seq is inherently an aggregation, testing these hypotheses requires independent measures of each cell. Since scRNA-seq allows for individual measurements of distinct cells, finding similar latent spaces directly from these data would provide strong evidence of their reflection of biological rather than technical variation. This finding suggests that latent space discovery in scRNA-seq data will better discern biological processes, as well as true cell type and state signatures, than bulk gene expression measurements.

ScCoGAPS Finds Signatures of Cell Types and Continuous Processes in the Developing Retina

To learn patterns directly from our scRNA-seq data across the developing mouse retina, scCoGAPS analysis was performed using the log-transformed, normalized mRNA copies per cell across a previously selected set of high-variance genes (Figure 3A) (Clark et al., 2019). Cells were partitioned into 100 sets of ~1,200 cells using a sampling scheme to ensure representation of all annotated cell types in each set. To eliminate potentially spurious patterns, consensus patterns were derived from



(legend on next page)

at least 25% of the independent sets and required an R^2 value of at least 0.7 to the within-cluster mean (STAR Methods).

We identified a total of 80 scCoGAPS patterns across the full developmental time course (Figure S2; Table S2). Pattern weights were tested for significant differential cell-type representations (Figure 3D) and predictive power (AUC) for each cell-type annotation (Figure S1F). Because performance biases based on the choice of classifier were observed, we calculated a standard contingency table and confusion matrix using the ROCR Bioconductor package to estimate a conservative AUC for each combination of pattern and cell-type annotation (Sing et al., 2005). Learned patterns corresponded to both discrete cell-type signatures and continuous-state transitions, including cycling retinal progenitor populations, a transient neurogenic phase, and intervals of cell-type-specific maturation along developmental trajectories (Figure 3B).

We identified at least one pattern corresponding to each of the 7 major cell types in the developing retina (Figures 3B and S1F). For example, patterns with high weights in annotated horizontal cells (patterns 2 + 16) correlated well and had high predictive power for our manually annotated horizontal cells, despite the relatively sparse number of cells of this type in our dataset. Learned patterns also highlighted gene network reuse across discrete cell types. For example, pattern 37 exhibited high weights in a subset of mature retinal ganglion cells (RGCs) and amacrine cells (AC) (Figure S1F). Additional patterns are specifically associated with mature RGCs (pattern 15) or recover other phenotypic features of these data, such as sex (pattern 36).

The application of scCoGAPS to scRNA-seq data also captured technical aspects of the data as well. Combinations of biologically incompatible patterns (e.g., two patterns for distinct mature cell types within the same cell) can readily delineate doublet cell populations (Figure S5B). In contrast, standard clustering methods would aggregate doublet cells together and separately from each discrete cell type and thereby be unable to recover biological information from them, or otherwise identify them as a unique, discrete cell type. Finally, we also identified patterns associated with technical features in our scRNA-seq dataset such as number of genes expressed (pattern 53) or batch effects (pattern 38). These pattern-phenotype correlations indicate that scCoGAPS recovers a collection of meaningful biological and technical patterns from the developing mouse retina scRNA-Seq data.

These correlations were able to resolve additional biological insights from these data not otherwise discernable from other analysis strategies. For example, pseudotemporal analysis was unable to resolve more closely related cell types or trajectories with a high degree of gene reuse (Clark et al., 2019). Correlation with manual annotation and patternMarker analysis (Stein-O'Brien et al., 2017) of the associated amplitudes allowed us to resolve both differentiating horizontal cells from amacrine cells and rods from cones (Table S3). Additional patterns were identified that correspond to continuous biological processes, i.e., cell-cycle state across RPCs (patterns 14, 31, 33, 62, 49, and 78, 49), with high degrees of gene reuse (Figures 3B, 3C, and S3). Additionally, many shared patterns only account for a small proportion of the cells in later-developing populations, suggesting that these transcriptional programs may be transient, or describe features associated with a subset of cells in a given lineage (Figure S5A).

To evaluate the performance of scCoGAPS relative to other commonly used single-cell deconvolution methods, we compared the patterns learned from scCoGAPS with the rotations learned from singular value decomposition (SVD) and principal-component analysis (PCA), feature weights from a gradient-based NMF (Lee and Seung, 2001), and weights extracted from the Deep Count Autoencoder (DCA) (Eraslan et al., 2018). All methods were evaluated on the same scRNA-seq dataset from the developing mouse retina. PCA and SVD fail to capture individual cell-type patterns and are driven predominantly by technical features that represent the greatest source of variation in these data (Figure S5C). Patterns learned from gradient-based NMF and scCoGAPS are comparable in their maximum correlation; however, the gene weights used to assess biological features for each pattern are more variable across multiple iterations of the gradient-based NMF. Using the Bayesian approach implemented in scCoGAPS, we can derive both mean and variance estimates allowing for variance incorporation into feature weights for gene-set analysis and more robust pattern annotations (Zyla et al., 2017). Both scCoGAPS and gradient NMF outperform DCA in the number of cell-type-specific latent spaces that are identified. Furthermore, DCA does not learn or export interpretable gene weights, which precludes our ability to explore the biological features represented in each of the latent spaces. Indeed, many non-linear deep learning methods using activation functions disambiguate the relationships between gene expression and learned patterns in a way that cannot be easily deconvolved.

We next sought to identify and characterize the specific cellular attributes captured in each pattern. Gene weights (**A**-matrix and their uncertainty) for each learned pattern were used as input for a Gene Ontology (GO) enrichment analysis using the CoGAPS gene-set test (Fertig et al., 2013b) across all Kyoto Encyclopedia of Genes and Genomes (KEGG) and GO gene sets with <100 genes (Figures S1E, S3, and S4). A heatmap of all significant gene-set statistics for all patterns are provided in Figures S3 and S4. Patterns that are well correlated with specific cell types are significantly enriched for appropriate gene ontologies. These include endothelial cells (9, 10, and 56), which are associated with angiogenesis and blood vessel patterning, as well as microglia (5, 6, 24, 25, 27, 57, and 58), which each showed significant enrichment for immune cell activities and processes ($p < 1 \times 10^{-6}$, Figure S4; Table S4). Concordant with their selective expression in rods and cone photoreceptors,

Figure 3. scCoGAPS Analysis of Time Course scRNA-Seq Data from Developing Mouse Retina

(A) UMAP of scRNAseq colored by age (top) and human annotated cell types (bottom).

(B) Heatmap of correlations of each scCoGAPS pattern to each annotated feature.

(C) UMAP of retina development colored by scCoGAPS pattern weights illustrate cell-type-specific (rods, top) and shared (cell cycle, bottom) patterns.

(D) Alluvial of cell-type-specific patterns links manually annotated cell types to scCoGAPS patterns for which at least 75% of the cell of a given type have a pattern weight of > 0.01.

respectively, patterns 21 and 39 are enriched in phototransduction, visual perception, photoreceptor cell maintenance, and photoreceptor outer segment terms ($p < 1 \times 10^{-8}$, Figures 3C and S4; Table S4). RPC-associated patterns (13, 26, 31, 33, 45, 49, 62, 64, 72, and 78) are enriched for cell-cycle regulators and embryonic development terms ($p < 1 \times 10^{-8}$, Figure S4; Table S4). Consistent with the fact that RGCs are the only neuroretinal cells that extend long projection axons, as well as the only cell to undergo high rates of apoptotic cell death during mouse retinal development (Young, 1984), the RGC-associated patterns 15 and 35 are enriched for axon guidance, with Pattern 15 also enriched for negative regulation of apoptosis.

Single-Cell Patterns Learned in One Dataset Can Be Transferred to Another via Projection Analysis

To assess whether learned patterns can be meaningfully transferred across datasets, we used our developing retinal dataset as the source data and compared it to a previously published scRNA-seq dataset from P14 mouse retina, established using a different droplet-based technique (Macosko et al., 2015). The target Drop-seq single-cell dataset was projected into the space of the 80 scCoGAPS patterns from the source 10×-based retinal development time-course data.

We hypothesized that shared latent spaces would stratify target data consistent with their underlying cellular attributes, while artifacts or data-specific features would not. Projected pattern weights were tested for AUC for each cell-type annotation in the target Drop-Seq dataset (Figures 4 and S6A). Because performance biases based on the choice of classifier are known to exist, a standard contingency table and confusion matrix were calculated using the ROCR Bioconductor package to provide a highly conservative AUC for each combination of annotated cell types and patterns (Sing et al., 2005). Using the projected pattern weights and cell types, we evaluated the ability of each pattern to distinguish cell types in the target dataset. (Figure S6A). Consistent with our hypothesis, AUC values confirm that patterns associated with mature cell types present in both the source and target dataset have significant predictive power (AUCs > 0.7, Wald test; BH-correction; $q < 0.01$), while those patterns associated with developmental processes only in the source data did not exhibit significant projections in the more mature (P14) target dataset (AUC < 0.7, Wald test; BH-correction; $q > .01$). For example, pattern 21, which was strongly associated with rods in the retina development time-course data, selectively marked rod photoreceptors in the P14 retina Drop-Seq data (Figure 4A, right panel; AUC = 0.83). Other patterns of mature cell types included pattern 2 (AUC of 0.95 for horizontal cells), Pattern 55 (AUC of 0.91 for amarcines), Pattern 15 and 16 (AUC of .93 and .92, respectively, for RGCs), and Pattern 64 (AUC of .99 for astrocytes) (Figure 4B). In contrast, the RPC pattern 31, which was strongly enriched for GO terms associated with cell cycle, failed to yield any significant signal (Figure 4A, middle panel), consistent with a lack of cycling progenitors in the P14 mouse retina.

Using only the significant patterns associated with mature cell types, we are able to resolve true positive cells from background expression pattern projections in the target dataset as illustrated by AUC curves for the predictive power of each weight for each cell type (Figure 4C) and the distribution of projected pattern weights (Figure 4D). Patterns with poor predictive power, such as Pattern 3, exhibited weights centered around zero, while patterns with high predictive potential, such as the rod-specific pattern 21, exhibit a bimodal distribution (Figure 4D). Cells in the target dataset annotated as rods, however, exhibit a unimodal distribution overlapping with the higher intensity peak of projected pattern weights. The cells contributing to the lower intensity peak, therefore, have some degree of the pattern 21 (rod) signature contributing to their transcriptional profile that likely reflects contamination acquired during dissociation and library preparation. These results validate the biological basis of the scCoGAPS patterns for mature cell types and demonstrate the sensitivity and specificity of projectR as a system to transfer annotations based on factors containing shared biological features across datasets.

projectR Recovers Continuous Processes and Temporal Progression from Disparate Data Types across Species

We next tested whether projection analysis could identify continuous biological features across organisms. Specifically, we projected a publicly available time course analysis of human bulk RNA-Seq from whole retinas into our single-cell scCoGAPS patterns. Homologous genes were used to map the amplitude values across species (STAR Methods). Briefly, $\log_2$-transformed gene expression values from human retina bulk RNA-seq data from gestational day (D) 52 to D136 were projected into the 80 mouse developing patterns. Each projected pattern was evaluated for predictive power for a given human developmental time point with the expectation that the changes in predictive power should reflect the change in pattern utilization over human retinal development. The resulting AUC values revealed a temporal gradient for cell-type-specific patterns, which reflects both developmental age and relative abundance of each cell type in the bulk sample (Figure 5A). Furthermore, the stereotyped birth order of major retinal cell types (Clark et al., 2019) was faithfully recapitulated in the progression of pattern projections in the human time course.

The observed gradient reflects the previously reported three major gene expression epochs of human retina development (Hoshino et al., 2017). The first epoch includes genes with high expression from D52 to D67. Patterns associated with early-born cell types such as horizontal cells (pattern 1) and RGCs (pattern 15) peaked early (D57 and D67, respectively) and then declined, reflecting their decreasing relative abundance as later-born cell types are generated. Patterns with amplitude values significantly enriched in RPC-specific processes such as cell-cycle regulation (Pattern 31) exhibited significant projection in the first epoch (Wald test; BH-correction; $q < .01$) with AUC values greater than 0.7 as well. Furthermore, the increased resolution of the patterns derived from scRNA-seq allowed a more granular association of corresponding biological processes within the larger epoch. These results indicate that shared continuous features associated with developmental programs in both mouse and human retinal development can be identified via transfer learning with projectR.

Species-specific differences were also apparent in this projection analysis. For example, genes that mark mature cone and rod photoreceptors are strongly expressed postnatally in mice (Blackshaw et al., 2001, 2004; O'Brien et al., 2003) but are detected prenatally in humans. Consistent with this, patterns 39 and 21, which are associated with mouse cones and rods,

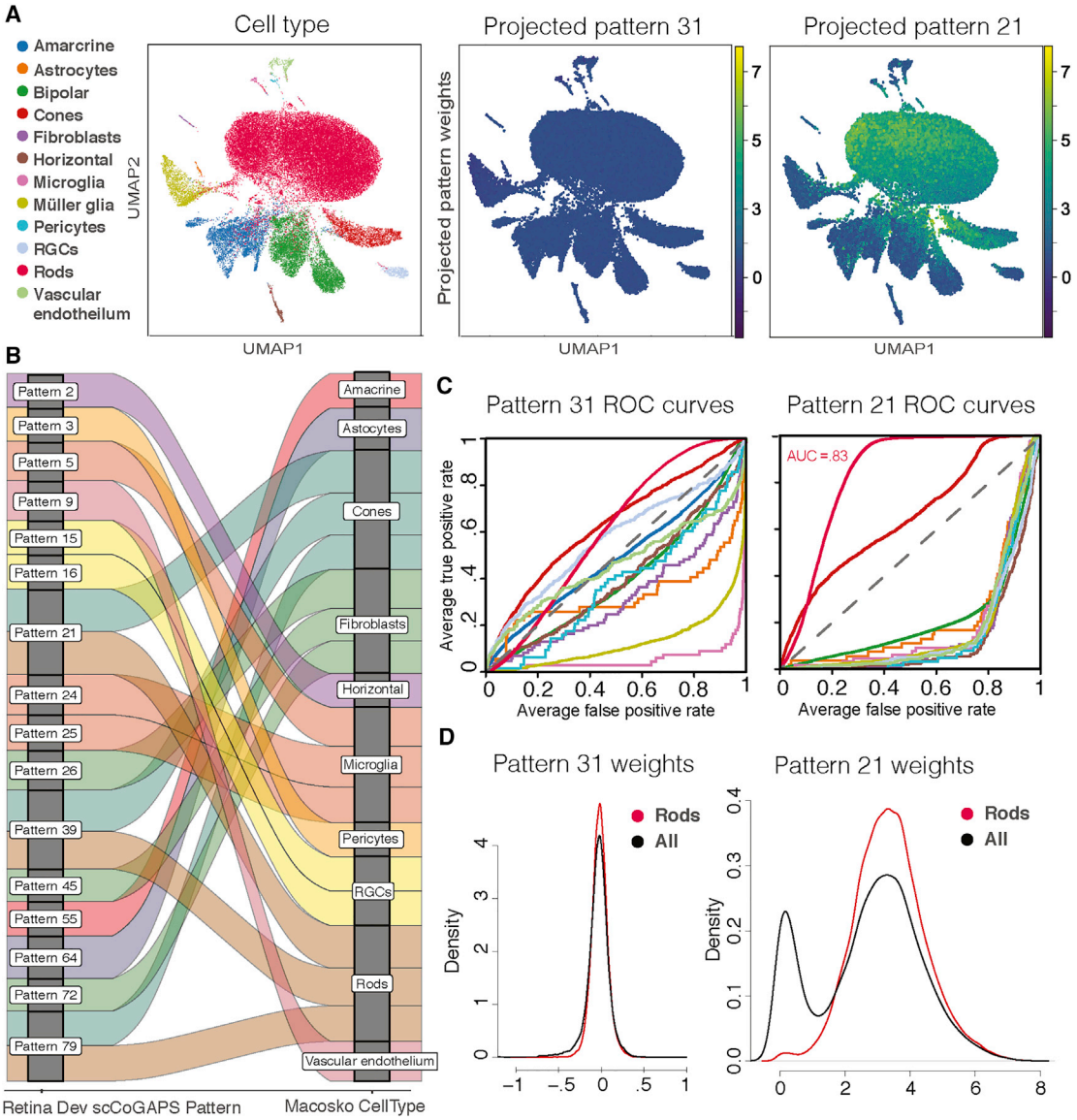


Figure 4. projectR Recovers Shared Cell Types in Independent Murine Retina scRNA-Seq Data

(A) UMAP of DropSeq data from P14 mouse retina colored by annotated cell type (left), projected pattern weights in pattern 31 (center), and projected pattern weights in pattern 21 (right).

(B) Alluvial plot of projected patterns links previously annotated cell types to scCoGAPS patterns for which at least 75% of the cell of a given type have a significant projection (Wald test; BH-correction; $q < 0.01$).

(C) ROC curves for classifiers built using the projected pattern weights for Pattern 21 (right) and projected pattern weights in Pattern 31 (left). Cell types are colored according to the legend in (A).

(D) Density plots of projected pattern weights for all cell types (black) and rods only (red).

respectively, exhibit high AUC values during the third epoch of gene expression in our human projection analysis (Figure 5A) (Hoshino et al., 2017). Previous analysis of the bulk RNA-seq data had demonstrated that differentially expressed genes within the the third epoch were enriched for gene ontology terms related to photoreceptors, synaptic connectivity, and neurotransmission (Hoshino et al., 2017). Mouse homologs of the genes annotated with these GO terms were also significantly enriched for higher amplitude values in source patterns 39 and 21 ($p < .001$) confirming that projectR recovered the species-specific temporal differences in the use of these patterns.

To test the ability of projectR to resolve spatiotemporal patterns, we next projected a separate bulk RNA-Seq time course of dissected regions of the human retina from Hoshino et al. (Hoshino et al., 2017). The fovea and macula have been shown to be developmentally ahead of age-matched nasal central and peripheral retina (Hendrickson and Drucker, 1992; Hendrickson et al., 2012; O'Brien et al., 2003) and enriched for both cone photoreceptors and retinal ganglion cells (Curcio and Allen,

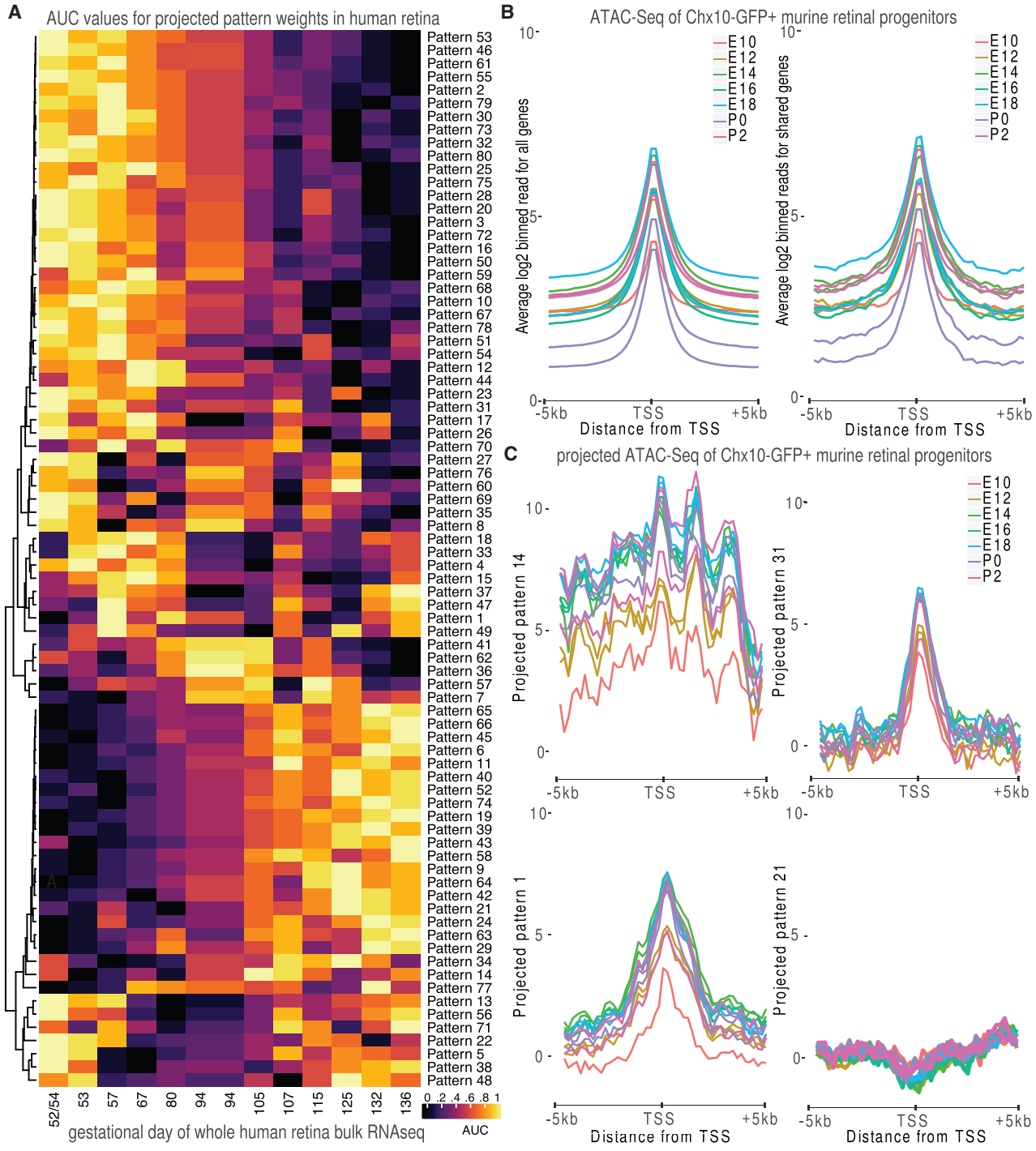


Figure 5. Projection of Retina Time Course Data Reveal Shared Temporal Dynamics across Species and Platforms

(A) Heatmap of AUC values for projected pattern weights in developing whole human retina recapitulates previously established gene expression epochs.

(B) Average ATAC signal for binned read counts overlapping 200-bp interval extending out 5 kb on either side of the transcription start for all genes (left) or the subset of genes from which the scCoGAPS patterns were learned (right).

(C) Projection of binned read counts overlapping 200-bp interval extending out 5 kb on either side of the transcription start into scCoGAPS patterns 14 (top left), 31 (top right), 1 (bottom left), and 21 (bottom right).

1990). A previous differential gene expression analysis of macula versus periphery was underpowered to detect significantly differentially expressed genes at each time point. However, using the projected values for each sample, we could readily identify significant differential pattern usage (Wald test; BH-correction across patterns; $q < .01$) between the macula and peripheral retina at

days 73 and 132. The fovea and macula are enriched in patterns specific to mature neurons, particularly retinal ganglion cells and cones (patterns 1, 15, 39, and 52) and depleted in patterns specific to retinal progenitor cells (patterns 26, 31, and 78) or immature neural precursor cells (patterns 17 and 73) relative to the age-matched peripheral retina (Figure S6B). These results demonstrate the utility of projectR in recovering spatiotemporally regulated differences within tissue and/or organ development.

Projection analysis can also determine pattern usage across a variety of different cellular measurement types. To illustrate this, we determined whether patterns learned from scRNA-seq analysis of the developing mouse retina could be used to identify distinct chromatin accessibility profiles within a mouse retinal ATAC-seq time series obtained from FACS-isolated *Chx10:*GFP^{+} RPCs (Rowan and Cepko, 2004) collected at two-day intervals between E10.5 and P2 (Figures 5B, 5C, and S7). Since ATAC-seq profiles chromatin accessibility, rather than gene expression, projection analysis enabled identification of patterns associated with genes whose local chromatin structure is primed for transcriptional activation. For each gene, ATAC-seq reads were quantified in 200-bp bins −5 Kb to +5 Kb around each canonical transcription start site (TSS) for each time point sampled (STAR Methods). As expected, the naïve signal shows global enrichment over TSSs owing to the increased accessibility at TSS of actively transcribed genes (Buenrostro et al., 2013) (Figure 5B). Overall signal intensity was highly variable with biological replicates from the same time point demonstrating a strong batch effect. These effects persisted when the ATAC-seq data were subset to the same high-variance genes used to define the scCoGAPS patterns (Figure 5B, right). To test the ability of projectR to overcome these effects, no batch correction or further data normalization was performed.

Despite the consistent profile of the observed mean enrichment of ATAC-Seq signal at the TSS across samples, projection of the ATAC-seq into the scCoGAPS patterns revealed several classes of chromatin accessibility patterns. Different accessibility profiles emerged that are lost in aggregate. Furthermore, the shape of the accessible peak and ranking of samples is distinct across different patterns, indicating that projection analysis can recover discrete signatures of accessibility associated with latent spaces learned from gene expression profiles, independent of technical noise. Together, these results suggest that learned accessibility signatures are associated with specific biological processes at distinct developmental time points in the developing mouse retina. Specifically, patterns that reflected missing processes (including non-neuroretinal cell types such as microglia that were not sampled in the ATAC-seq) demonstrate no significant signal in the projection analysis, while shared processes are apparent in both the scRNA-seq and the ATAC-seq data. For those projected patterns with significant ATAC-seq signal, replicates displayed significantly tighter concordance, and the amplitudes of the projected accessibility signatures appropriately reflected temporal progressions.

Broad domains of open chromatin on either side of the transcriptional start site—a hallmark of strongly transcribed genes—are observed exclusively in patterns associated with proliferating RPCs (e.g., patterns 14, 45, 72, and 78; Figure 5C, top left) consistent with the ATAC-Seq sampling of this population. Sharp peaks of open chromatin centered on the TSS corresponded to RPC-specific patterns associated with actively transcribed genes (e.g., patterns 4, 31, and 64; Figure 5C, top right) as well as a subset of patterns associated with maturing retinal subtypes, including cones, RGCs, and ACs (e.g. patterns 1, 2, 15, and 39; Figure 5C, bottom left), and immature rod photoreceptors (pattern 79). Finally, TSS signatures of closed chromatin are associated with patterns specific to cells that are not derived from RPCs, such as microglia (5 and 24) and erythrocytes (28), as well as with the mature rod photoreceptor-specific Pattern 21Figure 5C, bottom right). These data indicate that promoter regions associated with genes specific to RPC-derived cell types exist in an open and poised state in RPCs, with the notable exception of genes specific to mature rods.

projectR Enables Latent Space Comparison across Model Systems from the Developing Retina to the Developing Brain

The retina is often used as model system for neural development. In particular, both retinal neurogenesis and corticogenesis share a stereotyped birth order of different lineages from a single progenitor population (Kohwi and Doe, 2013; Miller and Gauthier, 2007). To test the ability of projectR to identify conserved pattern usage across tissues and model systems, we projected our retinal scRNA-Seq patterns into two datasets derived from developing human cortex (Nowakowski et al., 2017) (Zhong et al., 2018) and an additional dataset of the developing mouse midbrain (La Manno et al., 2016) (Figure 6). Projection of these patterns across all cells in each of the datasets completed in 165.6, 56.0, and 3.0 s, respectively, on a single high performance computing (HPC) node with a 2.5 GHz AMD Opteron Processor 6380 and 40 Gb of RAM. Consistent with a significant degree of conserved developmental programs and tissue composition between retina and select other CNS regions, we identified 87.5% (70 out of 80), 76.3% (61 out of 80), and 98.8% (79 out of 80) of patterns with significant projection ($q <= 0.01$; Wald test) in at least one cell in each of these comparable model systems (Figures 6 and S9), suggesting that many of the processes described by these patterns are reused in other CNS regions.

For the human cortical data, patterns 5, 20, 28, 29, 31, 40, 53, 64, and 65 captured 75% of published annotated cell types (Figure S8A). Consistent with its derivation as a progenitor-associated pattern in the developing retina and GO enrichment for cell cycle, pattern 31 demonstrated significant (AUC > 0.7; $q <= 0.01$; Wald Test; BH-corrected) projection to basal intermediate progenitor cells (IPCs), IPC-derived neuronal precursors of the medial ganglionic eminence (MGE), and dividing radial glia in the cortex (Figure S9A). In cortical data from Nowakowski et al., we observed that pattern 43, which is specific to inhibitory amacrine cells in retina, is also associated with interneurons (Figures 6A and S9A). Newborn excitatory pyramidal neurons are enriched for genes found in both the photoreceptor precursor-enriched pattern 79 (*Unc119*, *Meis2*, and *Cdc43ep3*) as well as the amacrine and horizontal cell-enriched pattern 1 (*Nrxn3*, *Kdm5b*, and *Dusp1*). Additionally, we are able to classify previously unannotated cells (NA) as neurons via significant projection of pattern 7, which is enriched for mature neuronal markers (*Nnat, Tubb2b,* and *Nefl*). In data from Zhong et al., where progenitors and precursors of GABAergic interneurons are annotated as a single class, these cells were significantly associated with patterns specific to GABAergic horizontal and

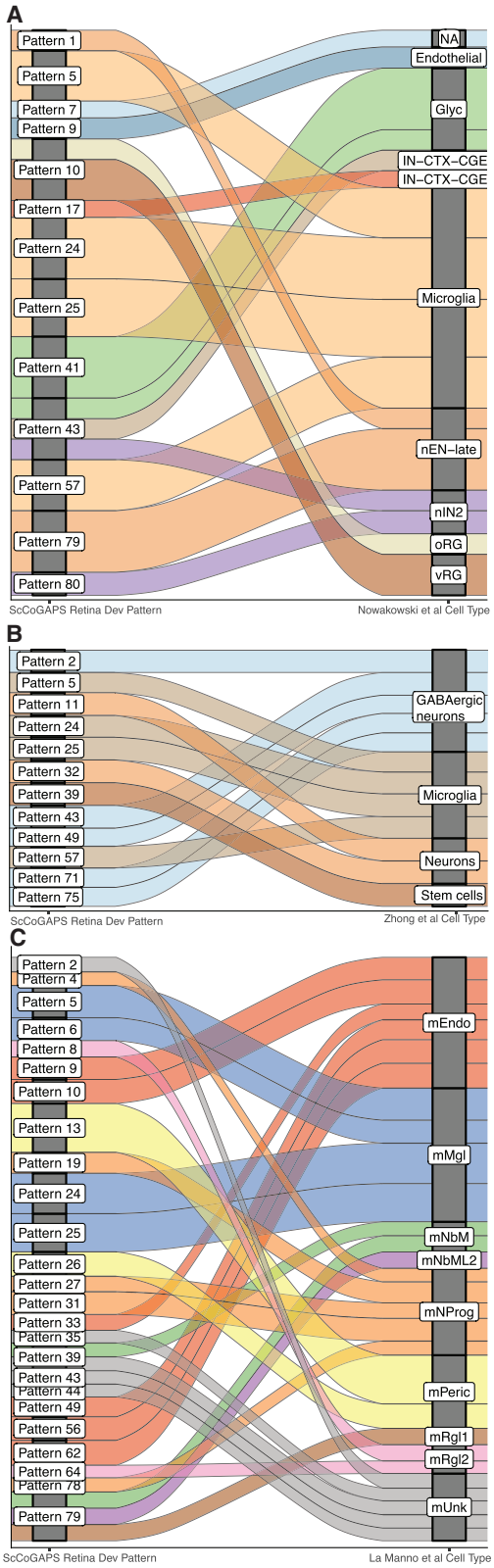


Figure 6. Developing Brain scRNA-Seq Projected in scCoGAPS Patterns of Retina Development

Alluvial plots connecting scCoGAPS patterns to cell types for which at least 25% of all cells are significant (Wald test; BH-correction; q < .01) in a given a projected scRNAseq of human cortical development from (A) Nowakowski et al. and (B) Zhong et al. as well as projected scRNAseq of mouse midbrain development from (C) La Manno et al.

amacrine cells (2 and 43) and RPCs (49 and 71) (Figure 6B). In the mouse midbrain, neural progenitor cells were enriched for retinal progenitor-specific patterns 4, 31, and 78, consistent with their shared roles in these two tissues (Figure 6C). Notably, Glyc cells in human cortex and mUnk cells in mouse midbrain—neither of which could be confidently classified in the original studies—are both enriched for patterns and genes (*Tubb2b*, *Sox4*, *Mapt*, and *Onecut2*) specific to immature amacrine, horizontal, and/or RGC cells, indicating that these both most likely represent as yet undescribed neuronal precursor subtypes (Figure 6C). These associations further demonstrate that projection analysis can be used to identify and annotate comparable cell types and shared cellular attributes across disparate model systems and that information transfer faithfully recovers these associations across species (Figure S9).

Patterns 5, 6, 24, 25, and 57 are each associated with microglial cells in the original source dataset. We observe significant differences in the projections of these patterns into microglia from different CNS regions, as well as across species. Patterns 5, 24, and 25 were consistently associated with microglia in all three brain region projections (Figures 6A–6C). However, pattern 57 was significantly ($q < 0.01$; Wald test; BH corrected) associated with microglia in both human cortical projections but not in microglia from the mouse midbrain (Figure 6A, 6B, S9A, and S9B), suggesting a potential difference in microglia signatures derived from different CNS regions. This pattern projection is driven in part by the Cathepsin family member genes *Ctsb* and *Ctsd*, as well as *Cd9*, each of which has been previously shown to be upregulated in a subclass of cortical microglia (Keren-Shaul et al., 2017). Thus, pattern 57 may be specifically associated with the cortically enriched microglia type II and highlighting a region-specific property of microglia detected via projection analysis. Additionally, no significant projections for Pattern 6 were identified in either human CNS dataset (Figures 6C and S8C); 0 out of 68 (0%) annotated microglia in Zhong et al. and 0 out of 77 (0%) microglia in Nowakowski et al. In contrast 76 out of 77 (98.7%) microglia in the human cortical development study have significant ($q <= 0.01$; Wald test; BH corrected) projections into pattern 5. Thus, using projectR, we are able to discriminate region- and species-specific differences in the transcriptional signatures of discrete cell types.

Shared Latent Spaces Identify Novel Cell-Type Associations across an Atlas of Adult Mouse Tissues

Given that latent spaces may reflect the signatures of biological processes in the conditions in which they are learned, we next asked whether we could identify significant use of these processes in more diverse cellular contexts from an atlas of adult mouse tissue scRNA-seq. The Tabula Muris dataset is a collection of 70,118 single-cell gene expression profiles from 12 mouse tissues (Wyss-Coray et al., 2018) collected using the 10× Genomics Chromium platform (Figure 7A). Using projectR, we projected the Tabula Muris dataset into the developing retina

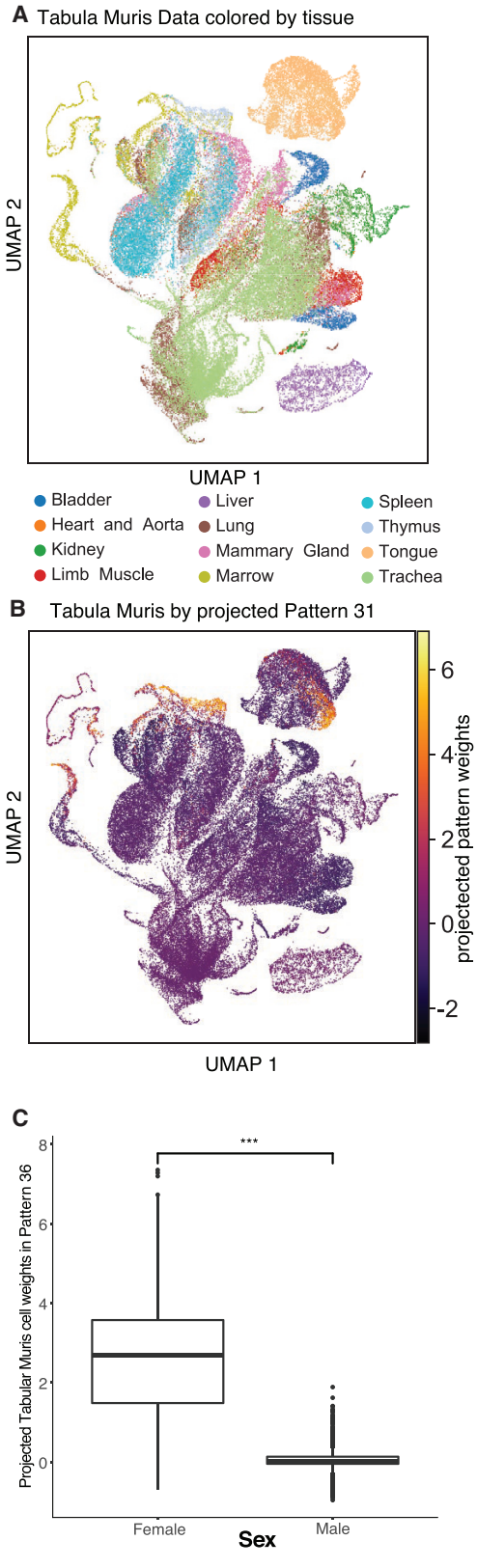


Figure 7. Projection of Retinal scCoGAPS Patterns into Mouse Non-neuronal Cell Dataset

(A) UMAP of scRNA-seq data from the Tabular Muris collection of mouse tissues colored by tissue.

(B) Projected pattern weights in pattern 31.

(C) Boxplot of projected pattern 36 weights stratified by sex demonstrates statistically significant difference corroborating association with genes involved in X-inactivation. ($p < 2.2e\text{-}16$, two-way t-test).

latent spaces. This analysis completed in 107 s on a HPC node with a 2.5 GHz AMD Opteron Processor 6380 and 40 Gb of RAM. Consistent with our hypothesis that biologically meaningful latent spaces will be shared across diverse cell types, 83.8% (67 out of 80) of the patterns demonstrated significant projection ($q < 0.0001$; Wald test) in at least one cell, and significant projections were identified in each of the 12 tissues in the Tabula Muris dataset.

Using only patterns learned in the developing retina, we were able to identify and annotate a variety of cellular features in these data. Many progenitor-associated patterns project into adult tissues with high levels of cell turnover, and specifically within subsets of cells that are actively proliferating (Figure 7B). Consistent with previous projections, pattern 31 is highly predictive of actively mitotic cells and can be used as a proliferative index via projection (AUC > 0.7) in tissues within the Tabula Muris dataset such as marrow, thymus, and tongue (Figure 7B). As previously described (Clark et al., 2019), we identified pattern 36 as specifically associated with sex in our developing retinal source dataset. This association was confirmed by defining biomarkers for each factor, computed using the PatternMarker statistic (STAR Methods) (Stein-O'Brien et al., 2017) (Table S3), and finding *Xist* as the sole PatternMarker for pattern 36. Projection of the Tabula Muris dataset into Pattern 36 almost perfectly segregated cells by sex (Figure 7C, p value $< 2.2e^{-16}$, two-way t-test). While females displayed a range of significant weights, males had uniformly insignificant projected pattern weights. In the source data, pattern 36 has high weights in a large proportion of cells, but sex was not determined *a priori*. The projection of pattern 36 across these two datasets provides an example of how annotations from a target dataset can also be used to annotate latent spaces from the source dataset as well.

Patterns specific to retinal neurons were detected in a number of peripheral tissues (Figure S9A). In the trachea, *Mgp*$^+$ goblet cells expressed genes associated with the neuronal cytoskeleton and neurotransmission (*Gap43*, *Sncg*, *Chgb*, and *Tac1*). In the tongue, *Krt6a*/*Krt16*$^+$ epithelial cells of both the filiform papillae (pattern 37) and *Krt14*$^+$ cells of the basal layer (pattern 41) selectively expressed genes associated with the neuronal cytoskeleton. In the lung, a small number of cells expressed pattern markers associated with amacrine and horizontal cell-enriched patterns 16 and 17 (*Scg5, Tmsb10, Malat1,* and *H3f3a*) (Figure S9A). Notably, this lung subpopulation expressed *Ins1* and *Ins2* and may thus represent a previously uncharacterized subset of pulmonary neuroendocrine cells (Figures S9B–S9D). In each of these cases, none of the most highly selective marker genes of these cells types (*Mgp, Krt6a*/*14*/*16,* and *Ins1*/*2*) were themselves expressed in retina, but rather, the projected patterns identified more complex similarities in gene expression between these peripheral cell types and retinal cells. These findings illustrate the power of this approach to identify biological

processes and cellular attributes shared between otherwise transcriptionally dissimilar cell types.

DISCUSSION

The rapid expansion of high-throughput biological assays has generated massive amounts of data. Single-cell experiments can now involve millions of individual samples adding to the complexity and scalability required to analyze these data. Applying latent space approaches to single-cell analyses has successfully identified and corrected technical errors associated with mRNA dropout (Eraslan et al., 2018) and enabled analysis of cell-cell variation (Loos et al., 2018). However, comparing biologically meaningful molecular features across datasets remains a critical challenge. Context-dependent biological variation and technical variation both challenge the ability to make meaningful interpretations from direct comparisons of biologically distinct datasets (Lê Cao et al., 2009; Tung et al., 2017). Our approach extends the latent space concepts used for data processing to enable the comparison of biological factors across a variety of experimental paradigms and cellular contexts.

By leveraging the structure generated by the co-regulation of genes, we are able to find a reduced set of continuous factors that describe cellular identity, state, and phenotype in a model system where differential expression analysis and marker genes are insufficient alone. This result is not unique to our work (Stein-O'Brien et al., 2018). However, while previous algorithms have focused on resolving differences between samples or groups of samples (Brunet et al., 2004; Cleary et al., 2017; Kim et al., 2017), we focus on optimizing our algorithm's solution to account for gene and pathway reuse in scRNAseq data. scCoGAPS identifies factors using a Bayesian NMF approach with a prior distribution tailored to model the sparsity of scRNA-Seq data. We developed a new computing structure and method for parallelization across all cells in a dataset to allow for computationally tractable factorizations of increasingly larger datasets such as those proposed by the Human Cell Atlas Project (Rozenblatt-Rosen et al., 2017). This parallelization strategy also allows for the independent discovery of patterns across sets of cells or samples and can be exploited to assess confidence in the learned factors, which is not available from other methods. Application of scCoGAPS to time course scRNA-seq data across mouse retina development identified gene-expression signatures of discrete cell types and shared gene networks. When compared to other methods, scCoGAPS outperformed gradient-based NMF and DCA when learning patterns of shared biological processes and SVD, PCA, and DCA when learning cell-type-specific patterns. Because DCA is optimized to denoise data, this use was outside of the intended scope of the algorithm.

projectR quantifies the extent to which the relationships between biological processes, inferred by dimensionality reduction methods, are shared across datasets from different assay technologies, cellular measurements, and species. Using ProjectR, independent and biologically distinct datasets, such as mouse retina and human cortex, can be compared with respect to their use of specific latent spaces. In contrast, existing tools for comparative analysis rely on consensus clustering using marker genes (Kiselev et al., 2018) or visualizations independent of specific molecular features (Cho et al., 2018). CCA (Soneson et al., 2010) and other single-cell dataset comparison tools forcibly align source and target datasets into a common, shared manifold that does not reflect the native state of either dataset. Moreover, these techniques have limited applicability for differences in data dimensionality and distributions (Butler and Satija, 2017; Wang et al., 2015). By mapping target data into a basis set defined by the source data, projectR allows for the direct evaluation of what is shared between, versus what is unique to, the source and target datasets. A key challenge to such cross-study comparison arises from technical variation such as batch effects between datasets, which may be non-linear. In spite of this complexity, projectR can overcome these confounding factors to relate features across datasets from disparate measurement platforms.

Many of the applications of this transfer learning approach including cell-type inference, comparison of factors across distinct conditions, feature discovery, and cross-model and cross-assay integrative analyses are areas of significant future work. The requirement of a feature map for transfer learning with projectR currently precludes its use with multi-layer autoencoders and other nonlinear methods that do not concurrently learn gene and sample weights. However, expansion of projectR to other unsupervised techniques represents an area of current and future work to bridge this gap and other methods exist that work exclusively with autoencoders (Taroni et al., 2019). Likewise, comparison of the least squares projection method employed in projectR to other orthogonal and non-orthogonal projection methods are also critical to determine optimal information transfer between datasets.

Application of scCoGAPS and projectR allows for exploratory analysis of high-dimensional biological data through the lenses of individual biological processes. This approach enables a shift in how we compare and identify cells beyond reliance on marker genes or ensemble molecular identity. Here, we demonstrate the sensitivity of this workflow to recover shared features and annotations across a variety of data types and experimental conditions. Our approach enabled *de novo* annotation and correction of existing cell-type annotations in a target retinal scRNA-seq study. We demonstrate the cross-platform and cross-species sensitivity of this approach to identify paralogous cell types in the retina and other tissues and identify meaningful biological similarities in markedly different cell types in a mouse cell atlas. This approach provides a strong foundation to develop new integrative analysis approaches using low dimensional representations to describe biological systems and how specific cellular attributes are shared across biological contexts.

STAR★METHODS

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

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Single-Cell RNA-Seq Analysis of the Developing Mouse Retina Data Obtained from Clark et al., (2019)
  - Target Public Domain Datasets

- Bulk RNA-Seq of the Developing Mouse Retina
- ATAC-Seq of the Developing Mouse Retina Obtained from Zibetti et al., (2017)
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Pattern Discovery via scCoGAPS
  - CoGAPS Atomic Prior
  - Update Steps for the Atomic Prior
  - Initialization
  - Conditional Distributions for Gibbs Sampling
  - Conditional Distribution for Birth or Resizing of Atoms
  - Conditional Distribution for Exchange between Neighboring Atoms in the Atomic Domain
  - Annealing Parameter
  - Pattern Matching for Consensus Gene Signatures
  - Pattern Curation Using Manual Feature Annotation
  - Benchmarking scCoGAPS against Commonly Used Dimensionality Reduction Tools
  - Gene Set Analysis of scCoGAPS Patterns
  - projectR Analysis
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2019.04.004.

ACKNOWLEDGMENTS

This work was supported by grants from the NIH (R01EY020560 and U01EY027267 to S.B., F32EY024201 and K99EY027844 to B.S.C., R01CA177669, U01CA196390, U01CA212007, and P30CA006973 to E.J.F.); the NSF (IOS-1656592 to L.A.G.); the Chan-Zuckerberg Initiative DAF (2018-182718 for Q.H., 2018-183445 to L.A.G., and 2018-183444 to E.J.F.); an advised fund of Silicon Valley Community Foundation, the Johns Hopkins University Catalyst (E.F. and L.A.G.); and Discovery awards (E.J.F.), and the Johns Hopkins University School of Medicine Synergy Award (S.B., L.A.G., and E.J.F.). Q.H. would like to thank J. Taroni for discussions on transfer learning and low-dimensional representations. The authors would like to thank C.A. Berlinicke and D.J. Zack for assistance with FACS analysis; A. Wolf and F. Theis from the Helmholtz Center, Munich, Germany for productive discussions and introductory scanpy cod; the Johns Hopkins Genetic Resources Core Facility for use of the 10× Genomics Single Cell system; and the Hopkins microarray and Deep Sequencing Core for assistance with sequencing; the CZI Jamboree, C. Greene, K. Korthauer, and A. V. Favorov for invaluable collaborations and discussions; and A. Battle, V. Yegnasubramanian, and J. Bader for comments on the manuscript.

AUTHOR CONTRIBUTIONS

G.L.S.-O.'B., B.S.C., S.B., L.A.G., and E.J.F. conceived and directed the study. B.S.C. generated scRNA-Seq data. G.L.S.-O.'B., B.S.C., L.A.G., and E.J.F. analyzed scRNA-seq data, with L.A.G. and E.J.F. as senior bioinformaticians. C.Z. and B.S.C. generated the bulk RNA-Seq, and C.Z. generated the ATAC-seq data. G.L.S.-O.'B., T.S., L.A.G., and E.J.F. developed scCoGAPS. G.L.S.-O.'B., R.S., C.C., L.A.G., and E.J.F. contributed to the development of projectR. Q.H. and G.L.S.-O.'B., developed the random forest classifier for the projections of bulk GWCoGAPS patterns. R.S. and G.L.S.-O.'B. developed the AUC evaluation method for projected pattern weights included in the projectR package. S.L., C.Z., J.Q., and G.L.S.-O.'B. analyzed the ATAC-seq data. G.L.S.-O.'B., B.S.C., S.B., L.A.G., and E.J.F. wrote the paper with input from all co-authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 28, 2018
Revised: January 24, 2019
Accepted: April 17, 2019
Published: May 22, 2019

REFERENCES

Bassett, E.A., and Wallace, V.A. (2012). Cell fate determination in the vertebrate retina. Trends Neurosci. *35*, 565–573.

Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44.

Bidaut, G., and Ochs, M.F. (2004). ClutrFree: cluster tree visualization and interpretation. Bioinformatics *20*, 2869–2871.

Blackshaw, S., Fraioli, R.E., Furukawa, T., and Cepko, C.L. (2001). Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. Cell *107*, 579–589.

Blackshaw, S., Harpavat, S., Trimarchi, J., Cai, L., Huang, H., Kuo, W.P., Weber, G., Lee, K., Fraioli, R.E., Cho, S.H., et al. (2004). Genomic analysis of mouse retinal development. PLoS Biol. *2*, E247.

Brunet, J.P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. USA *101*, 4164–4169.

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

Butler, A., and Satija, R. (2017). Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. bioRxiv.

Ching, T., Huang, S., and Garmire, L.X. (2014). Power analysis and sample size estimation for RNA-Seq differential expression. RNA *20*, 1684–1696.

Cho, H., Berger, B., and Peng, J. (2018). Generalizable and scalable visualization of single-cell data using neural networks. Cell Syst. *7*, 185–191.

Clark, B., Stein-O'Brien, G., Shiau, F., Cannon, G., Davis, E., Sherman, T., Rajaii, F., James-Esposito, R., Gronostajski, R., Fertig, E., et al. (2019). Single cell RNA-Seq analysis of retinal development identifies NFI factors as regulating mitotic exit and late-born cell specification. Neuron *102*.

Cleary, B., Cong, L., Cheung, A., Lander, E.S., and Regev, A. (2017). Efficient generation of transcriptomic profiles by random composite measurements. Cell *171*, 1424–1436.

Curcio, C.A., and Allen, K.A. (1990). Topography of ganglion cells in human retina. J. Comp. Neurol. *300*, 5–25.

Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2018). Single cell RNA-seq denoising using a deep count autoencoder. Nat Commun. *10*, 390.

Fertig, E.J., Ding, J., Favorov, A.V., Parmigiani, G., and Ochs, M.F. (2010). CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. Bioinformatics *26*, 2792–2793.

Fertig, E.J., Favorov, A.V., and Ochs, M.F. (2013b). Identifying context-specific transcription factor targets from prior knowledge and gene expression data. IEEE Trans. Nanobiosci. *12*, 142–149.

Fertig, E.J., Markovic, A., Danilova, L.V., Gaykalova, D.A., Cope, L., Chung, C.H., Ochs, M.F., and Califano, J.A. (2013a). Preferential activation of the hedgehog pathway by epigenetic modulations in HPV negative HNSCC identified with meta-pathway analysis. PLoS One *8*, e78127.

Hendrickson, A., and Drucker, D. (1992). The development of parafoveal and mid-peripheral human retina. Behav. Brain Res. *49*, 21–31.

Hendrickson, A., Possin, D., Vajzovic, L., and Toth, C.A. (2012). Histologic development of the human fovea from midgestation to maturity. Am. J. Ophthalmol. *154*, 767–778.

Hoshino, A., Ratnapriya, R., Brooks, M.J., Chaitankar, V., Wilken, M.S., Zhang, C., Starostik, M.R., Gieser, L., La Torre, A., Nishio, M., et al. (2017). Molecular anatomy of the developing human retina. Dev. Cell *43*, 763–779.

Ishwaran, H., and Rao, J.S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. Ann. Stat. *33*, 730–773.

Javed, A., and Cayouette, M. (2017). Temporal progression of retinal progenitor cell identity: implications in cell replacement therapies. Front. Neural Circuits *11*, 105.

Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T.K., David, E., Baruch, K., Lara-Astaiso, D., Toth, B., et al. (2017). A unique microglia type associated with restricting development of Alzheimer's disease. Cell *169*, 1276–1290.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Kim, D., Langmead, B., and Salzberg, S.L. (2016). HISAT2 Implementation.

Kim, J.W., Abudayyeh, O.O., Yeerna, H., Yeang, C.H., Stewart, M., Jenkins, R.W., Kitajima, S., Konieczkowski, D.J., Medetgul-Ernar, K., Cavazos, T., et al. (2017). Decomposing oncogenic transcriptional signatures to generate maps of divergent cellular states. Cell Syst. *5*, 105–118.

Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet. *20*, 273–282.

Kiselev, V.Y., Yiu, A., and Hemberg, M. (2018). Scmap: projection of single-cell RNA-seq data across data sets. Nat. Methods *15*, 359–362.

Kohwi, M., and Doe, C.Q. (2013). Temporal fate specification and neural progenitor competence during development. Nat. Rev. Neurosci. *14*, 823–838.

Kossenkov, A.V., Peterson, A.J., and Ochs, M.F. (2007). Determining transcription factor activity from microarray data using Bayesian Markov chain Monte Carlo sampling. Stud. Health Technol. Inform. *129*, 1250–1254.

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.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lê Cao, K.A., Martin, P.G.P., Robert-Granié, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinformatics *10*, 34.

Lee, D.D., and Seung, H.S. (1999). Learning the Parts of Objects by Non-Negative Matrix Factorization (Nature Publishing).

Lee, D.D., and Seung, H.S. (2001). Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems 13, T.K. Leen, T.G. Dietterich, and V. Tresp, eds. (MIT Press), pp. 556–562.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liao, Y., Smyth, G.K., and Shi, W. (2014). FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Loos, C., Moeller, K., Fröhlich, F., Hucho, T., and Hasenauer, J. (2018). A hierarchical, data-driven approach to modeling single-cell populations predicts latent causes of cell-to-cell variability. Cell Syst. *6*, 593–603.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Miller, F.D., and Gauthier, A.S. (2007). Timing is everything: making neurons versus glia in the developing cortex. Neuron *54*, 357–369.

Moloshok, T.D., Klevecz, R.R., Grant, J.D., Manion, F.J., Speier, W.F., 4nd, and Ochs, M.F. (2002). Application of Bayesian decomposition for analysing microarray data. Bioinformatics *18*, 566–575.

Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science *358*, 1318–1323.

O'Brien, K.M.B., Schulte, D., and Hendrickson, A.E. (2003). Expression of photoreceptor-associated molecules during human fetal eye development. Mol. Vis. *9*, 401–409.

Ochs, M.F., and Fertig, E.J. (2012). Matrix factorization for transcriptional regulatory network inference. IEEE Symp. Comput. Intell. Bioinforma. Comput. Biol. Proc. 387–396.

Ochs, M.F., Rink, L., Tarn, C., Mburu, S., Taguchi, T., Eisenberg, B., and Godwin, A.K. (2009). Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. Cancer Res. *69*, 9125–9132.

Pan, S.J., Kwok, J.T., and Yang, Q. (2008). Transfer learning via dimensionality reduction. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. 677–682.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Rowan, S., and Cepko, C.L. (2004). Genetic analysis of the homeodomain transcription factor Chx10 in the retina using a novel multifunctional BAC transgenic mouse reporter. Dev. Biol. *271*, 388–402.

Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A., and Teichmann, S.A. (2017). The human cell atlas: from vision to reality. Nature *550*, 451–453.

Sibisi, S., and Skilling, J. (1996). Bayesian density estimation. In Maximum Entropy and Bayesian Methods (SpringerLink), pp. 189–198.

Sibisi, S., and Skilling, J. (1997). Prior distributions on measure space. J. R. Stat. Soc. B *59*, 217–235.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940–3941.

Skilling, J., and Sibisi, S. (1996). Priors on measures. In Maximum Entropy and Bayesian Methods (SpringerLink), pp. 261–270.

Soneson, C., Lilljebjörn, H., Fioretos, T., and Fontes, M. (2010). Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. BMC Bioinformatics *11*, 191.

Stein-O'Brien, G.L., Arora, R., Culhane, A.C., Favorov, A.V., Garmire, L.X., Greene, C.S., Goff, L.A., Li, Y., Ngom, A., Ochs, M.F., et al. (2018). Enter the matrix: factorization uncovers knowledge from omics. Trends Genet. *34*, 790–805.

Stein-O'Brien, G.L., Carey, J.L., Lee, W.S., Considine, M., Favorov, A.V., Flam, E., Guo, T., Li, S., Marchionni, L., Sherman, T., et al. (2017). PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF. Bioinformatics *33*, 1892–1894.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Tabula Muris Consortium; Overall coordination; Logistical coordination; Organ collection and processing; Library preparation and sequencing; Computational data analysis; Cell type annotation; Writing group; Supplemental text writing group; Principal investigators (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367–372.

Taroni, J.N., Grayson, P.C., Hu, Q., Eddy, S., Kretzler, M., Merkel, P.A., and Greene, C.S. (2019). MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of fare disease. bioRxiv.

Torrey, L., and Shavlik, J. (2009). Transfer learning. In Handbook of Research on Machine Learning Applications and Trends Algorithms, Methods, and Techniques, E.S. Olivas, ed. (IGI Global), pp. 242–264.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. *31*, 46–53.

Tung, P.Y., Blischak, J.D., Hsiao, C.J., Knowles, D.A., Burnett, J.E., Pritchard, J.K., and Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. Sci. Rep. *7*, 39921.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nat. Biotechnol. *34*, 1145–1160.

Wang, W., Arora, R., Livescu, K., and Bilmes, J.A. (2015). Unsupervised learning of acoustic features via deep canonical correlation analysis. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), pp. 4590–4594.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Wysoker, A., Tibbetts, K., and Fennell, T. (2013). Picard Tools Version 1.90.

Wyss-Coray, T., Darmanis, S., and Muris Consortium, T. (2018). Single-cell transcriptomic characterization of 20 organs and tissues from individual mice creates a Tabula Muris. bioRxiv.

Young, R.W. (1984). Cell death during differentiation of the retina in the mouse. J. Comp. Neurol. *229*, 362–373.

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

Zheng, C., Zheng, L., Yoo, J.K., Guo, H., Zhang, Y., Guo, X., Kang, B., Hu, R., Huang, J.Y., Zhang, Q., et al. (2017). Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. Cell *169*, 1342–1356.

Zhong, S., Zhang, S., Fan, X., Wu, Q., Yan, L., Dong, J., Zhang, H., Li, L., Sun, L., Pan, N., et al. (2018). A Single-Cell RNA-Seq Survey of the Developmental Landscape of the Human Prefrontal Cortex (Nature Publishing).

Zhu, X., Ching, T., Pan, X., Weissman, S.M., and Garmire, L. (2017). Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. PeerJ *5*, e2888.

Zibetti, C., Liu, S., Wan, J., Qian, J., and Blackshaw, S. (2019). Epigenomic profiling of retinal progenitors reveals LHX2 is required for developmental regulationof open chromatin. Commun. Biol. *2*. Published online April 25, 2019.

Zyla, J., Marczyk, M., Weiner, J., and Polanska, J. (2017). Ranking metrics in gene set enrichment analysis: do they matter? BMC Bioinformatics *18*, 256.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
RNAeasy Mini kit	Qiagen	Cat#74134
Illumina TruSeq kit	Illumina	Cat#RS-122-2001
Deposited Data		
Bulk RNAseq of *Chx10*-Cre:GFP+ cells from a time course of murine retina	This paper	GSE118880
10× scRNAseq time course of murine retina development	Clark et al., 2019	GSE118880
ATAC seq of *Chx10*-Cre:GFP+ cells from a time course of murine retina development	Zibetti et al., 2017	GSE118880
Tabula Muris data	CZI Biohub	https://github.com/czbiohub/tabula-muris
Developing human cortex time course	Nowakowski et al., 2017	https://cells.ucsc.edu/?ds=cortex-dev
Adult murine retina scRNAseq	Macosko et al., 2015	GSE63472
Developing murine midbrain scRNAseq	La Manno et al., 2016	GSE76381
Developing human cortex scRNAseq	Zhong et al., 2018	GSE104276
Bulk RNAseq time course of human retina development	Hoshino et al., 2017	GSE104827
Experimental Models: Organisms/Strains		
Mice:CD1.Tg(Chx10-EGFP/cre/-ALPP)2Clc	Dr. Connie Cepko; {Rowan and Cepko, 2004}	RRID:MGI:3838985
Software and Algorithms		
R version 3.5	The R project	https://www.r-project.org/
scanpy version 1.3	Wolf et al., 2018	https://github.com/theislab/scanpy
scCoGAPS	bioconductor	https://www.bioconductor.org/packages/release/bioc/html/CoGAPS.html
projectR	This paper	https://github.com/genesofeve/projectR
Deep Count Autoencoder (DCA)	Eraslan et al., 2018	https://github.com/theislab/dca
NNLM	Lee and Seung, 2001	https://cran.r-project.org/web/packages/NNLM/vignettes/Fast-And-Versatile-NMF.html

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Elana J. Fertig (ejfertig@jhmi.edu)

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Information about the generation and genotyping of the mouse transgenic lines used in this study can be found in the corresponding original studies: Chx10-Cre:GFP+ (Rowan and Cepko, 2004). All mice were maintained on a CD-1 background. Animals used for bulk RNA-seq and ATAC-Seq ranged from embryonic day 10 (E10) to postnatal day 2 (P2). Both males and females were used in this study. Mice were housed in a climate-controlled pathogen free facility, on a 14 hour-10 hour light/dark cycle (07:00 lights on-19:00 lights off). All experimental procedures were preapproved by the Institutional Animal Care and Use Committee of the Johns Hopkins University School of Medicine.

METHOD DETAILS

Single-Cell RNA-Seq Analysis of the Developing Mouse Retina Data Obtained from Clark et al., (2019)

The developmental time series of scRNA-seq from mouse retina was generated as part of our companion manuscript (Clark et al., 2019), and these data were used for pattern discovery and annotation as described below. UMAP representations (Becht et al., 2018)

were learned on neighbors calculated from the first 32 PCs using scanpy version 1.3 (Wolf et al., 2018) following data preprocessing as described in (Zheng et al., 2017).

Target Public Domain Datasets

All data was downloaded from GEO with the exception of the Tabular Muris data which was downloaded from https://github.com/czbiohub/tabula-muris and the developing human cortex time course from (Nowakowski et al., 2017) which was downloaded from https://cells.ucsc.edu/?ds=cortex-dev. Accession numbers in order of appearance in the manuscript are GSE63472 (Macosko et al., 2015), GSE104827 (Hoshino et al., 2017), GSE104276 (Zhong et al., 2018), and GSE76381 (La Manno et al., 2016).

Bulk RNA-Seq of the Developing Mouse Retina

At select developmental time points, cells were collected from biological replicates of FACS-sorted *Chx10*-Cre:GFP+ mouse retinas as previously described (Rowan and Cepko, 2004). RNA was isolated using the RNAeasy Mini kit (Qiagen) with on-column DNase treatment. Isolated total RNA was assessed for integrity on the Bioanalyzer 2100 system, and we required a minimum RNA integrity number of 7. RNA-Seq libraries were created using the Illumina TruSeq kit (Illumina), quantified via PicoGreen assay and fragment size distribution was determined using the Bioanalyzer 2100. Libraries were bar-coded, pooled, and run on a HiSeq2500 instrument to an average sequencing depth of 30.0 million aligned reads per sample. 75–100 bp paired-end reads were mapped to the mouse reference genome (mm10) using Hisat2 (Kim et al., 2015, 2016). Gene expression estimates for the reference transcriptome (Gencode vM5) and differential testing were performed using Cuffdiff2 (Trapnell et al., 2013) with default parameters. Data are available from GEO in GSE118880.

ATAC-Seq of the Developing Mouse Retina Obtained from Zibetti et al., (2017)

Chromatin derived from flow-sorted *Chx10:*Cre-GFP^{+} (Rowan and Cepko, 2004) retinal fractions was processed as previously described (Zibetti et al., 2017). Briefly, chromatin was extracted and processed for Tn5 mediated tagmentation and adapter incorporation, according to the Manufacturer's protocol (Nextera DNA sample preparation kit, Illumina) at 37°C for 30 min. Reduced-cycle amplification was carried out in presence of compatible indexed sequencing adapters. Libraries were quantified using the PicoGreen assay and fragment size distribution was determined using the Bioanalyzer 2100. Up to 4 samples per lane were pooled and run on a HiSeq2500 Illumina sequencer to produce 50-bp paired ends for each sample.

Bowtie2 (version 2.3.2) was used for ATAC-Seq reads alignment to the mouse genome (mm10) (Langmead and Salzberg, 2012). Duplicate reads were removed using Picard tools (version 2.10.7)(Wysoker et al., 2013). Improperly mapped reads were removed using SAMtools (version 1.5). (Li et al., 2009). Read counts for each gene were retrieved using featureCounts program (version 1.5.3). (Liao et al., 2014). Read counts overlapping 200-bp interval extending out 5 kb on either side of the transcription start site were generated with custom scripts using bedtools (version 2.26.0)(Quinlan and Hall, 2010). Data are available from GEO in GSE118880.

QUANTIFICATION AND STATISTICAL ANALYSIS

Pattern Discovery via scCoGAPS

Latent spaces were learned using the scCoGAPS function from the CoGAPS v 3.0 Bioconductor 3.7 package from log transformed cpms of the high variance genes for all samples. Cells were partitioned into 100 sets of ~1200 cells using a sampling scheme to ensure representation of all annotated cell types in each set. Consensus patterns were derived as described in the next section using the patternMatch4scRNASeq function from the CoGAPS v 3.0 Bioconductor 3.7 package and then rerun across all sets using scCoGAPS with fixed = TRUE to ensure reciprocity of the learned weights.

CoGAPS Atomic Prior

CoGAPS decomposes a matrix $\mathbf{D}$ of G genes (rows) and S samples (columns) into two matrices $\mathbf{A} \in \mathbb{R}^{G \times k}$ and $\mathbf{P} \in \mathbb{R}^{k \times S}$ using the model:

$$p(\mathbf{A}, \mathbf{P} | \mathbf{D}, \boldsymbol{\Sigma}) \propto p(\mathbf{D} | \mathbf{A}, \mathbf{P}, \boldsymbol{\Sigma}) p(\mathbf{A}) p(\mathbf{P}),$$

where the elements of $\boldsymbol{\Sigma}$ represent the corresponding standard deviation of each element in the matrix $\mathbf{D}$. Determining the optimal value of k remains an open problem for latent space detection. The CoGAPS model assumes each element of $\mathbf{D}$ is i.i.d. with $p(D_{i,j} | A_{i,\cdot}, P_{\cdot,j}, \Sigma_{i,j})$ a normal distribution with mean $\mu_{i,j} = A_{i,\cdot} \times P_{\cdot,j}$ and variance $\sigma^2_{i,j}$.

In the case of sequencing data, $D_{i,j}$ is log transformed counts. In cases with replicates, $D_{i,j}$ can be replaced with the mean log transformed read counts and standard deviation can be computed across these replicates. In cases without replicates, the standard deviation is assumed to be 10% of the signal in $\mathbf{D}$ with a minimum value of 0.1.

CoGAPS uses an atomic prior (Sibisi and Skilling, 1997) for the $\mathbf{A}$ and $\mathbf{P}$ matrices based upon previous work in Bayesian non-negative NMF for microarrays (Moloshok et al., 2002). The atomic prior (Sibisi and Skilling, 1997) is similar to spike and slab model (Ishwaran and Rao, 2005), in which only a subset of model parameters are nonzero and those that are have a value distributed according to some continuous distribution with non-negative support. As a result, this model results in a ℓ_0 sparsity constraint on these matrices with other constraints depending on the distribution used to model nonzero values in these matrices. The atomic prior

models each nonzero matrix element of **A** or **P** with a gamma distribution. The rate λ^A and λ^P of this distribution is a parameter that is fixed for every matrix element in **A** and **P**, respectively. The shape of the gamma prior for each matrix element is a separate hyper-parameter ($\alpha^A_{i,k}$ for each element of **A** and $\alpha^P_{k,j}$ for each element of **P**), modeled as a Poisson distribution with a fixed parameter α for each matrix element. Zero values for $\alpha^A_{i,k}$ or $\alpha^P_{k,j}$ correspond to $A_{i,j} = 0$ and $P_{k,j} = 0$, modeling the subset of model parameters that are zero.

The expectation of the Gamma distribution is proportional to the sampled values of $\alpha^A_{i,k}$ or $\alpha^P_{k,j}$, introducing a further sparsity constraint on the magnitude of the matrix elements when these values are small. In contrast to standard spike and slab models, the atomic prior also models smoothness by encoding a correlation structure between matrix elements in **A** and **P** during the sampling steps.

Recall that $A_{i,j} \sim \Gamma(\alpha^A_{i,k}, \lambda^A)$ is equivalent to the sum of $\alpha^A_{i,k}$ independent, exponentially distributed random variables with rate parameter λ^A and similarly for $P_{k,j}$. Instead of directly sampling from the Gamma or Poisson distributions, the proposal distribution in the atomic prior updates a single, exponentially distributed random variable $x^A_{i,k,l}$ for A and $x^P_{k,j,m}$ for P at each step. The advantage of sampling a single atom at a time is that the conditional distribution posterior for an exponential prior on each atom and the normal likelihood is a truncated normal, enabling Gibbs sampling. This single random variable is called an "atom" and the set of all such atoms is referred to as the "atomic domain". The value of each matrix element of **A** is then given by

$$A_{i,k} = \sum_{l=1}^{\alpha^A_{i,k}} x^A_{i,k,l}$$

and similarly for **P**. The atoms in the atomic domain are stored in ordered coordinates on a number line ($l^A_{i,k,l}$ for **A** and $l^P_{k,j,m}$ for **P**), which is divided into bins that correspond to each matrix element (Main Figure 1). The set of all atoms for one matrix is referred to as the "atomic domain". If the number of atoms is smaller than the number of matrix elements, this data structure reduces the memory required to keep track of each atom and provides an efficient structure to find all the atoms mapping to a single matrix element. The prior distribution of atom coordinates is uniform, corresponding to an uniform prior for atom membership in each matrix element.

Update Steps for the Atomic Prior

CoGAPS alternates between updating n_A atoms in the **A** and n_P atoms in the **P** matrices. The values of n_A and n_P are sampled from a Poisson distribution with parameter for the total number of atoms in the atomic domain for **A** (N_A) and in the atomic domain for **P** (N_P), respectively. Thus, on expectation all atoms in the domain are updated at each matrix-level iteration. The total number of such update steps is input as a parameter *nEquil* during the burn in stage and *nSamp* during the sampling stage.

In each of these n_A and n_P, we perform one of the four update steps to the respective atomic domains (Main Figure 1). We briefly describe these steps for **A** below, and note that they are defined similarly for **P**.

1. create a single new atom in the atomic domain, so that $N_A \leftarrow N_A + 1$.
2. change the value of a single atom $x^A_{i,k,l} \leftarrow x^A_{i,k,l} - \Delta x^A_{i,k,l}$, and removing it from the atomic domain so that $N_A \leftarrow N_A - 1$ if $x^A_{i,k,l} - \Delta x^A_{i,k,l} = 0$.
3. changing the location of a single atom ($x^A_{i,k,l}$) to a new location between adjacent atoms ($x^A_{m,n,p}$ and $x^A_{q,r,s}$) such that $l^A_{i,k,l} \in (l^A_{m,n,p}, l^A_{q,r,s})$ on the atomic domain.
4. Moving a portion of the value of a single atom ($x^A_{i,k,l}$) to another, adjacent atom ($x^A_{m,n,p}$) so that $x^A_{i,k,l} \leftarrow x^A_{i,k,l} + \Delta x$ and $x^A_{m,n,p} \leftarrow x^A_{i,k,l} - \Delta x$ where $\Delta x \in (-(x^A_{i,k,l} + x^A_{m,n,p}), x^A_{i,k,l} + x^A_{m,n,p})$. Atoms may become small from exchange, but not exactly zero or removed from the atomic to maintain detailed balance.

At each of the n^A or n^P iterations, each of these four steps is chosen at random with 1/3 probability of either birth or death, 1/3 probability of move, and 1/3 probability of exchange. The relative probability of selecting birth or death is selected based on the Poisson prior. Recall for **A** that birth implies $N^A \leftarrow N^A + 1$, the sum of Poisson distributed random variables, and that under the Poisson distribution $P(N+1|N) = (N/(N+\lambda))$ where λ is the Poisson parameter. Together, these three conditions suggest that $P(\text{birth}|N^A) = (N^A/(N^A + \alpha G k))$ for the **A** atomic domain and that $P(\text{birth}|N^P) = (N^P/(N^P + \alpha k S))$ for the **P** atomic domain. The probability of death or resize is then one minus the probability of birth. Metropolis Hastings sampling is used for the move step, whereas Gibbs sampling is used for the other three steps using the conditional distributions.

Initialization

The atomic domains for both **A** and **P** are initialized without any atoms, so that $A_{i,j} = 0$ and $P_{k,l} = 0$. This limits the initial atomic update step to birth step, birth or death when there is at least one atom in the domain, and all four update steps when there are at least two atoms in the domain.

At these initial steps, the estimated fit to the data $\mu_{i,j} = A_{i,}.P_{.j}$ will be zero for most values of i and j. Thus, these initial steps do not change the likelihood and are all accepted. This initialization effectively results in initial conditions which are a random sampling from the prior before Gibbs sampling.

Conditional Distributions for Gibbs Sampling

We would like to sample from Skilling's atomic domain using Gibbs sampling. We will assume that we are seeking the mass of an atom $x^A_{k,l}$ at $A_{k,l,j}$ for the **A** matrix and $x^P_{l,m,j}$ at $P_{l,m}$ for the **P** matrix. We use the variable x in the derivations below to reduced the number of indices in the equations, as the associated matrix element can be clearly inferred from the context of each equation. The initial mass of this atom is x_0, which is 0 if we have decided to birth the atom and > 0 if we have decided to kill it. We retain this term so that we can derive the conditionals for birth and death in a single expression.

Determining the mass of x requires first computing the full conditional distribution $p(x|x_0,\mathbf{D},\mathbf{A},\mathbf{P})$. To do this, we will first consider $P(\mathbf{A},\mathbf{P}|\mathbf{D})$ and examine the resulting distribution. We will begin by recalling that

$$p(\mathbf{A},\mathbf{P}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{A},\mathbf{P})p(\mathbf{A},\mathbf{P}).$$

Putting this in terms of an individual atom, we obtain

$$p(x|x_0,\mathbf{D},\mathbf{A},\mathbf{P}) \propto p(\mathbf{D}|x,x_0,\mathbf{A},\mathbf{P})p(x).$$

We assume that

$$p(\mathbf{D}|x,x_0,\mathbf{A},\mathbf{P}) \sim N(\mathbf{M},\boldsymbol{\Sigma}),$$

where **M** is the mock data matrix given by the product of **A** and **P** that incorporates the change in mass of the atom $x - x_0$ in the updated term. $\boldsymbol{\Sigma}$ is the covariance matrix for **D**. The prior for the mass of each atom x is given by an exponential with parameters λ_A and λ_P, respectively.

In each case, the full conditional distribution simplifies to a normal distribution, which is truncated so the value of the atom $x \geq 0$. Below follows the detailed derivation of this distribution for birth and resizing and exchange.

Conditional Distribution for Birth or Resizing of Atoms

Atomic Domain for A

We will first explore the likelihood in more detail, assuming that the mass of the atom maps to $A_{k,l}$

$$p(\mathbf{D}|x,x_0,\mathbf{A},\mathbf{P}) \propto \exp\left\{-\sum_i\sum_j \frac{1}{2\sigma^2_{i,j}}\left(D_{i,j} - \sum_p A_{i,p}P_{p,j} - (x-x_0)P_{l,j}\right)^2\right\}.$$

Since we are only concerned with computing the conditional for changes to $A_{k,l}$ we note that the other terms in **A** and **P** can be considered as parameters. As a result,

$$\begin{aligned} p(\mathbf{D}|x,x_0,\mathbf{A},\mathbf{P}) &\propto \exp\left\{-\sum_j \frac{1}{2\sigma^2_{k,j}}\left(D_{k,j} - \sum_p A_{k,p}P_{p,j} - (x-x_0)P_{l,j}\right)^2\right\} \\ &= \exp\left\{-\sum_j \frac{P_{l,j}}{2\sigma^2_{k,j}}\left(x - \left(\frac{D_{k,j} - \sum_p A_{k,p}P_{p,j} + x_0P_{l,j}}{P_{l,j}}\right)\right)^2\right\}. \end{aligned}$$

Let $\mu^A_{k,l,j} = \dfrac{D_{k,j} - \sum_p A_{k,p}P_{p,j} + x_0P_{l,j}}{P_{l,j}}$ and $s^A_{k,l,j} = \dfrac{P^2_{l,j}}{2\sigma^2_{k,j}}$. Then, the equation above becomes

$$\begin{aligned} p(\mathbf{D}|x,x_0,\mathbf{A},\mathbf{P}) &\propto \exp\left\{-\sum_j s^A_{k,l,j}\left(x - \mu^A_{k,l,j}\right)^2\right\} \\ &= \exp\left\{-\sum_j s^A_{k,l,j}\left(x^2 - 2\mu^A_{k,l,j}x + \mu^{A2}_{k,l,j}\right)\right\} \\ &= \exp\left\{-\left(x^2\sum_j s^A_{k,l,j} - 2x\sum_j s^A_{k,l,j}\mu^A_{k,l,j} + \sum_j s^A_{k,l,j}\mu^{A2}_{k,l,j}\right)\right\} \\ &\propto \exp\left\{-\sum_j s^A_{k,l,j}\left(x^2 - 2x\frac{\sum_j s^A_{k,l,j}\mu^A_{k,l,j}}{\sum_j s^A_{k,l,j}}\right)\right\}. \end{aligned}$$

If we now incorporate the product with the exponential prior distribution for α,

$$p(x|x_0,\mathbf{D},\mathbf{A},\mathbf{P}) \propto \exp\left\{-\sum_j s^A_{k,l,j}\left(x^2-2x\frac{\sum_j s^A_{k,l,j}\mu^A_{k,l,j}}{\sum_j s^A_{k,l,j}}\right)\right\}\exp\{-\lambda_A x\}$$

$$= \exp\left\{-\sum_j s^A_{k,l,j}\left(x^2-x\left(2\frac{\sum_j s^A_{k,l,j}\mu^A_{k,l,j}}{\sum_j s^A_{k,l,j}}-\frac{\lambda_A}{\sum_j s^A_{k,l,j}}\right)\right)\right\}$$

$$\propto N\left(\frac{2\sum_j s^A_{k,l,j}\mu^A_{k,l,j}-\lambda_A}{2\sum_j s^A_{k,l,j}},\frac{1}{\sqrt{2\sum_j s^A_{k,l,j}}}\right).$$

Within the code, we store values of s and $s \times \mu$ used to avoid dividing by zero in cases where $P_{l,j} = 0$.

Atomic Domain for P

Here, we consider atoms whose mass maps to elements $P_{l,m}$. From the likelihood, we get

$$p(\mathbf{D}|x,x_0,\mathbf{A},\mathbf{P}) \propto \exp\left\{-\sum_i \frac{1}{2\sigma^2_{i,m}}\left(D_{i,m}-\sum_p A_{i,p}P_{p,m}-(x-x_0)A_{i,l}\right)^2\right\}$$

$$= \exp\left\{-\sum_i \frac{A_{i,l}}{2\sigma^2_{i,m}}\left(x-\left(\frac{D_{i,m}-\sum_p A_{i,p}P_{p,m}+x_0A_{i,l}}{A_{i,l}}\right)\right)^2\right\}.$$

If $\mu^P_{i,l,m}=\frac{D_{i,m}-\sum_p A_{i,p}P_{p,m}+x_0A_{i,l}}{A_{i,l}}$ and $s^P_{i,l,m}=\frac{A^2_{i,l}}{2\sigma^2_{i,m}}$,

$$p(\mathbf{D}|x,x_0,\mathbf{A},\mathbf{P}) \propto \exp\left\{-\sum_i s^P_{i,l,m}\left(x-\mu^P_{i,l,m}\right)^2\right\}$$

$$= \exp\left\{-\sum_i s^P_{i,l,m}\left(x^2-2\mu^P_{i,l,m}x+\mu^{P2}_{i,l,m}\right)^2\right\}$$

$$\propto \exp\left\{-\left(\sum_i s^P_{i,l,m}\right)\left(x^2-\frac{2\sum_i \mu^P_{i,l,m}s^P_{i,l,m}x}{\sum_i s^P_{i,l,m}}\right)\right\}$$

If we now incorporate the prior distribution for x

$$p(x|x_0,\mathbf{D},\mathbf{A},\mathbf{P}) \propto \exp\left\{-\left(\sum_i s^P_{i,l,m}\right)\left(x^2-\left(\frac{2\sum_i \mu^P_{i,l,m}s^P_{i,l,m}}{\sum_i s^P_{i,l,m}}\right)x\right)\right\}\exp\{-\lambda^P x\}$$

$$= \exp\left\{-\left(\sum_i s^P_{i,l,m}\right)\left(x^2-\left(\frac{2\sum_i \mu^P_{i,l,m}s^P_{i,l,m}-\lambda^P}{\sum_i s^P_{i,l,m}}\right)x\right)\right\}$$

$$\propto N\left(\frac{2\sum_i \mu^P_{i,l,m}s^P_{i,l,m}-\lambda^P}{2\sum_i s^P_{i,l,m}},\frac{1}{\sqrt{2\sum_i s^P_{i,l,m}}}\right)$$

Conditional Distribution for Exchange between Neighboring Atoms in the Atomic Domain

Exchange for A between $A_{k,l}$ and $A_{m,n}$ where k ≠ m

We will refer to the atom corresponding to matrix element $A_{k,l}$ as x, the atom corresponding to the matrix element $A_{m,n} = x_{m,n}$, and x_0 and $x_{0,m,n}$ there initial values, respectively. The value of x after sampling is constrained such that $x \in (0, X)$ and $x_{m,n} = X - x$ where $X = x_0 + x_{0,m,n}$.

If we consider the exponential prior, the exchange step will incorporate both matrix elements. That is,

$$\exp(-\lambda_A x)\exp(-\lambda_A(X-x)).$$

The x terms in this equation cancel, indicating that the conditional depends only on the likelihood. This occurs for all exchange steps, and thus is not described in the remaining subsections on this step.

From the likelihood, we get

$$p(\mathbf{D}|x,x_0,X,\mathbf{A},\mathbf{P}) \propto \exp\left\{-\sum_j \frac{1}{2\sigma_{k,j}^2}\left(D_{k,j}-\sum_p A_{k,p}P_{p,j}-xP_{l,j}\right)^2\right\} \times \exp\left\{-\sum_j \frac{1}{2\sigma_{m,j}^2}\left(D_{m,j}-\sum_p A_{m,p}P_{p,j}-(X-x)P_{n,j}\right)^2\right\}.$$

For simplicity of the equations, we consider only the terms inside of the exponential and formulate them as an equation for x to find the parameters of the truncated normal for value j in the summation.

$$\frac{\left[xP_{l,j}-\left(D_{k,j}-\sum_p A_{k,p}P_{p,j}\right)\right]^2}{2\sigma_{k,j}^2}+\frac{\left[xP_{n,j}-\left(XP_{n,j}+\sum_p A_{m,p}P_{p,j}-D_{m,j}\right)\right]^2}{2\sigma_{m,j}^2}.$$

Letting $\mu_{k,j}=D_{k,j}-\sum_p A_{k,p}P_{p,j}$ and $M_{m,n,j}=XP_{n,j}+\sum_p A_{m,p}P_{p,j}-D_{m,j}$, the above term simplifies to

$$\frac{(xP_{l,j}-\mu_{k,j})^2}{2\sigma_{k,j}^2}+\frac{(xP_{n,j}-M_{m,n,j})^2}{2\sigma_{m,j}^2}.$$

Combining terms, we can write this equation as

$$\frac{\left[\sigma_{m,j}^2P_{l,j}^2+\sigma_{k,j}^2P_{n,j}^2\right]x^2-2\left[\sigma_{m,j}^2P_{l,j}\mu_{k,j}+\sigma_{k,j}^2P_{n,j}M_{m,n,j}\right]x}{2\sigma_{k,j}^2\sigma_{m,j}^2}$$

which can complete the square by

$$\frac{\left[\sigma_{m,j}^2P_{l,j}^2+\sigma_{k,j}^2P_{n,j}^2\right]}{2\sigma_{m,j}^2\sigma_{n,j}^2}\left(x-\frac{\sigma_{m,j}^2P_{l,j}\mu_{k,j}+\sigma_{k,j}^2P_{n,j}M_{m,n,j}}{\sigma_{m,j}^2P_{l,j}^2+\sigma_{k,j}^2P_{n,j}}\right)^2$$

The parameters for the truncated normal can now follow the derivation used for the birth step described above.

***Exchange for* A *between* $A_{k,l}$ *and* $A_{k,n}$**

Considering just the terms inside of the exponent, in this case we will have instead

$$\sum_j \frac{\left(D_{k,j}-\sum_p A_{k,p}P_{p,j}-xP_{l,j}-(X-x)P_{n,j}\right)^2}{2\sigma_{k,j}^2}$$

Collecting the x terms and completing the square we get

$$\sum_j \frac{(P_{l,j}-P_{n,j})^2}{2\sigma_{k,j}^2}\left[x-\frac{D_{k,j}-\sum_p A_{k,p}P_{p,j}-XP_{n,j}}{P_{l,j}-P_{n,j}}\right]^2$$

we let $s_j=\frac{(P_{l,j}-P_{n,j})^2}{2\sigma_{k,j}^2}$ and $\mu_j=\frac{D_{k,j}-\sum_p A_{k,p}P_{p,j}-XP_{n,j}}{P_{l,j}-P_{n,j}}$. The derivation for the terms of the truncated normal follow.

***Exchange for* P *between* $P_{k,l}$ *and* $P_{m,n}$ *where* $l \neq n$**

The derivation for exchange steps in **P** follows that of the derivation for **A** above. In this case,

$$s_i = \frac{\sigma_{i,n}^2A_{i,j}^2+\sigma_{i,l}^2A_{i,m}^2}{2\sigma_{i,l}^2\sigma_{i,n}^2},$$

$$\mu_i = \frac{\sigma_{i,n}^2A_{i,k}\mu_{i,l}+\sigma_{i,l}^2A_{i,m}M_{i,m,n}}{\sigma_{i,n}^2A_{i,k}^2+\sigma_{i,l}^2A_{i,m}^2},$$

where $\mu_{i,l}=D_{i,l}-\sum_p A_{i,p}P_{p,l}$ and $M_{i,m,n}=\sum_p A_{i,p}P_{p,n}+XA_{i,m}-D_{i,n}$.

Exchange for P *between* $P_{k,l}$ *and* $P_{m,l}$

The derivation for the exchange steps for **P** follows that of the derivation for **A**. Thus, in this case

$$s_i = \frac{(A_{i,k} - A_{i,m})^2}{2\sigma_{i,l}^2},$$

$$\mu_i = \frac{D_{i,l} - \sum_p A_{i,p}P_{p,l} - XA_{i,m}}{A_{i,k} - A_{i,m}}.$$

Annealing Parameter

During the equilibration phase, we in fact wish to sample from the conditional distribution

$$p(x|x_0, \mathbf{D}, \mathbf{A}, \mathbf{P}) \propto p(\mathbf{D}|x, x_0, \mathbf{A}, \mathbf{P})^{1/T} p(x),$$

where T is the annealing temperature. This has the effect of multiplying the term σ in each of the equations by a factor of T. As a result, the standard deviation s of the birth and resize terms are the only things to change by as follows.

$$s^A_{k,l,j} = \frac{P_{l,j}}{2T\sigma^2_{k,j}}, \text{ and}$$

$$s^P_{i,l,m} = \frac{A_{i,l}}{2T\sigma^2_{k,j}}.$$

A similar modification of the terms with $\sigma \leftarrow T\sigma$ will also occur in the exchange step, which will modify both the mean and standard deviation terms for this step.

Pattern Matching for Consensus Gene Signatures

Hierarchical clustering was done on gene weights from all sets and the resulting dendrogram is cut so the number of branches is equal to the original number of latent spaces. Each branch then contains the columns(s) of **A** across all the sets that are most related to each other. Well-dimensionalized data will produce robust patterns such that each branch will contain a single contribution from each of the randomly generated sets. As the additional sparsity can cause large clusters driven predominantly by zeros, the minimum and maximum number of patterns contributing to given branch can be specified with defaults of .5 and 1.5 the number of gene sets, respectively. Branches failing to meet the lower bound are dropped, while those exceeding the upper bound are subjected to additional rounds of hierarchical clustering. Additionally, the minimal correlation to the cluster mean for each pattern within a given branch was specified to be 0.7. Consensus signatures were then constructed for each branch by taking a weighted average of the gene signatures for that branch which pass all the criteria. To ease across pattern comparison, the resulting consensus signatures were scaled to have maxima of one. Pattern weights for all the cells were then learned in parallel from these signatures to ensure reciprocity across all the sets.

Pattern Curation Using Manual Feature Annotation

The AUC valued were calculated by inputing either the pattern weights output from scCoGAPs, the projected pattern weights output from projectR, or the p-values output from projectR with a one hot encoded matrix of annotated labels into the prediction function of the ROCR library v 1.0-7. The output of prediction was then evaluate using the performance function with method=auc from the ROCR library v 1.0–7 and the y.values extracted and reported. Note this process has been functionalize and is included in the projectR package v .99 as the auc_mat function. The heatmap in Figure 3B was created using the following. Each feature of contained in the annotation matrix was one hot encoded and the resulting vector correlated against the pattern weights generated by scCoGAPS for each cell.

Benchmarking scCoGAPS against Commonly Used Dimensionality Reduction Tools

SVD was calculated using the svd function with nu=80, nv=80 from the base R package v3.5.2. PCA was calculated with the scale and centered arguments as true using prcomp functions included in the core R stats package v 3.5.2. The gradient-based NMF was run using the both the nmf function from the NMF library v 0.21.0 with method set to ‘‘brunet’’ and k = 80 and the nnmf function from the NNLM library v 0.4.2 with threads set to 24. DCA was run in using dca.api in Python 3.6 the with arguments mode='latent', hidden_size=80, return_info=True, return_model=True.

Gene Set Analysis of scCoGAPS Patterns

Z-scores of gene weights were computed for each pattern in each ensemble by dividing the mean of the **A** matrix estimated across the chain by its standard deviation as previously described (Fertig et al., 2010; Ochs et al., 2009). The resulting matrix of Z-scores

is averaged for sets of patterns determined to match in the the ensemble as described above. A Wilcoxon gene set test with the R/Bioconductor LIMMA package version 3.36.2 (Ritchie et al., 2015) is performed for mouse KEGG and GO sets from the R/Bioconductor packages org.Mm.eg.db version 3.4.0, KEGG.db version 3.2.3, and GO.db version 3.4.0. Gene sets with more than 5 genes and fewer than 100 genes are retained for analysis. p-values for the gene set test are FDR adjusted with Benjamini Hotchberg and available as Table S4. Preranked GSEA was performed for the results of the gradient based NMF from the NMF library gene set test (Subramanian et al., 2005).

projectR Analysis

The R package projectR version 0.99.2 (available from https://github.com/genesofeve/projectR) was used to project the scCoGAPS consensus scCoGAPS patterns of the $\boldsymbol{A}$ matrix into each of the target datasets. These projection are achieved by solving the factorization

$$\boldsymbol{D} = \boldsymbol{A}\widehat{\boldsymbol{P}} + \varepsilon$$

using the least-squares fit to the new data as implemented via a wrapper for the lmFit function in the LIMMA package 3.30.13 (Ritchie et al., 2015). Specifically, a linear model is fit using the $\boldsymbol{A}_{ij}$ weights learned from the source data by scCoGAPS as the design matrix for multiple linear regressions. Each row of the design matrix, $\boldsymbol{A}_{i.}$, corresponds to the features, i.e. genes, which will map between the source and target data. Each column of the design matrix, $\boldsymbol{A}_{.j}$, corresponds to a previously learned individual latent space. The estimated coefficients of these regressions provide $\widehat{\boldsymbol{P}}$ matrix values for the target data. These $\widehat{\boldsymbol{P}}$s score the new samples using a gene-wise weighting, provided by the $\boldsymbol{A}$s, for each latent space. The ranking of the new samples within each pattern, or row of $\widehat{\boldsymbol{P}}$, are then indicative of the relative strength of a given sample's association with the latent space. A wald test to calculate the significance of these coefficients is calculated using the pdf of the negative absolute value of the coefficients scaled by their standard deviation, i.e.

$$W = \frac{-\left|\widehat{\boldsymbol{P}} - \boldsymbol{P}_0\right|}{\widehat{se}\left(\widehat{\boldsymbol{P}}\right)}$$

The AUC values were calculated by inputting either the projected pattern weights output from projectR or the p-values output from projectR with a one hot encoded matrix of annotated labels into the prediction function of the ROCR library v 1.0-7. The output of prediction was then evaluate using the performance function with method=auc from the ROCR library v 1.0–7 and the y.values extracted and reported. Note this process has been functionalized and is included in the projectR package v .99 as the auc_mat function. Additional functionality is included in the latest version of ProjectR (v 1.0) available as part of Bioconductor.

DATA AND SOFTWARE AVAILABILITY

scCoGAPS is available as part of the CoGAPS bioconductor package (3.8) under the GPL license.

ProjectR is available as part of the ProjectR bioconductor package (1.0) under the GPL license. Note, the exact version used for this analysis (v0.99.2) can be freely downloaded from https://github.com/genesofeve/projectR. All code for this analysis is available upon request. The accession number for the bulk RNA-seq data reported in this paper is GEO: GSE118880.