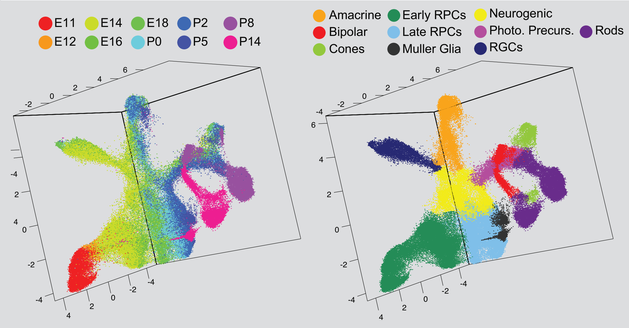
Projection-based deep analysis of retinal development at single cell resolution

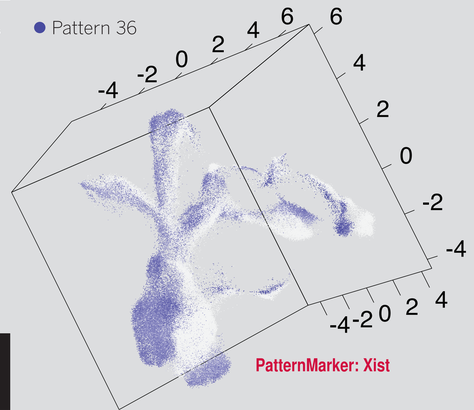
G.L. Stein-O'Brien[1,2,3], B. Clark[1], T. Sherman[3], C. Zibetti[1], C. Colantuoni[4], S. Blackshaw[1,4,5], E. Fertig[3,*], L. A. Goff[1,2,*]

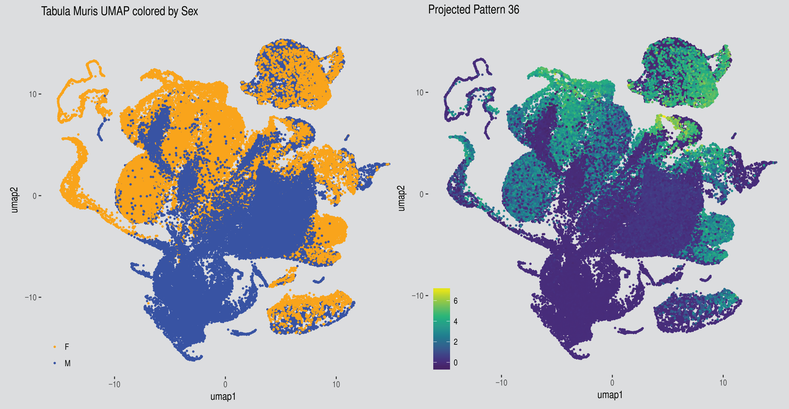
[1]Solomon H. Snyder Department of Neuroscience, [2] McKusick-Nathans Institute of Genetic Medicine, [3]Department of Oncology Biostatistics, [4]Department of Neurology, [5]Department of Ophthalmology, Center for Humans Systems Biology, Department of Ophthalmology, and Institute for Cell Engineering, Johns Hopkins School of Medicine, Baltimore, MD

UMAP of 10x scRNAseq from retina development



Information transfer via **ProjectR** can be used to discover biological properties of latent patterns





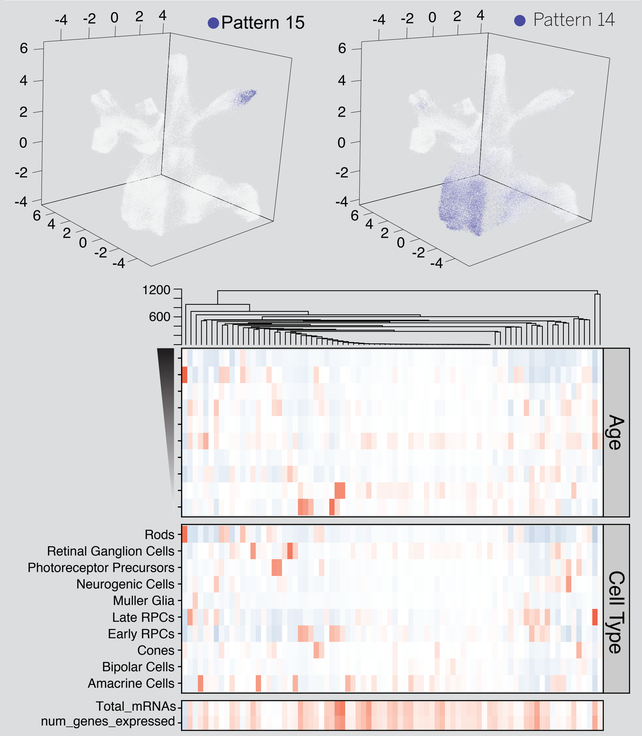
Indentify biomarkers via the **Patternmarkers** stat

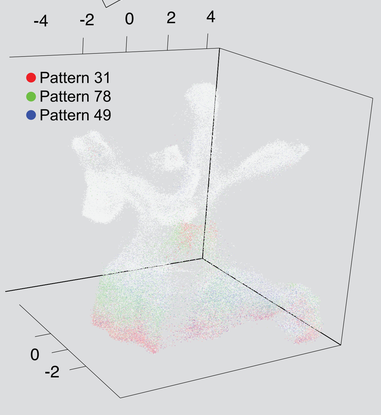
The patternMarkers statistic (s_{ij}) scores the association of the i^{th} gene's values in the amplitude matrix (A_i) with the j^{th} pattern or linear combination of patterns by computing

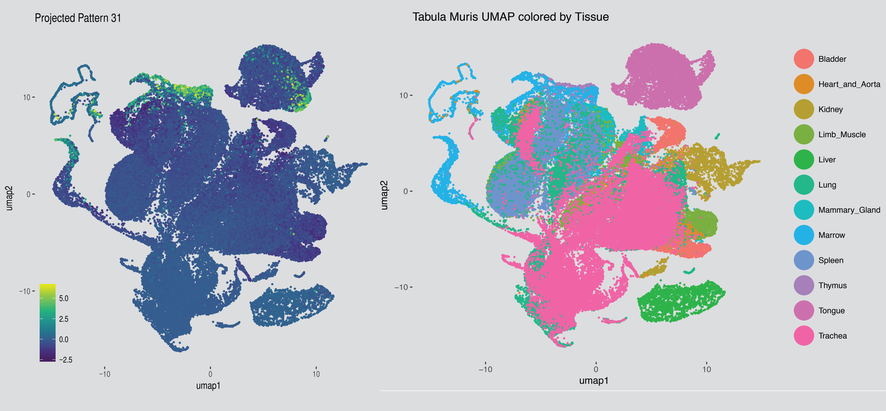
$$s_{ij}(\bar{w}_j) = \sqrt{\sum_k \left(\frac{A_{ik}}{\max A_i} - \bar{w}_{jk} \right)^2}$$

where i indices all the genes in the original data matrix, k indices all the patterns in the NMF solution, and is a vector of components specifying the jth linear combination of patterns that is constrained to sum to 1, and j indices the total number of linear combinations for which patternMarkers statistics are computed. The default setting for Eq. (1) sets j = {1,..., k}, such that w is a set containing a unit vector for each pattern and $s_{ij}(w_j)$ is an $l_2$norm indicating the exclusivity of the contri-bution of gene i to the pattern j and the corresponding $\bar{BP}$. Scaling by the maximum value of each gene in the NMF solution (max A_i) decouples the effect of overall gene expression level without impacting the quality of the factorization. Genes are ranked by increasing $s_{ij}(w_j)$ such that the higher the rank of the gene, the less it is associated with the considered pattern.

Data driven discovery of cell type specific and shared gene regulatory networks with **scCoGAPS**

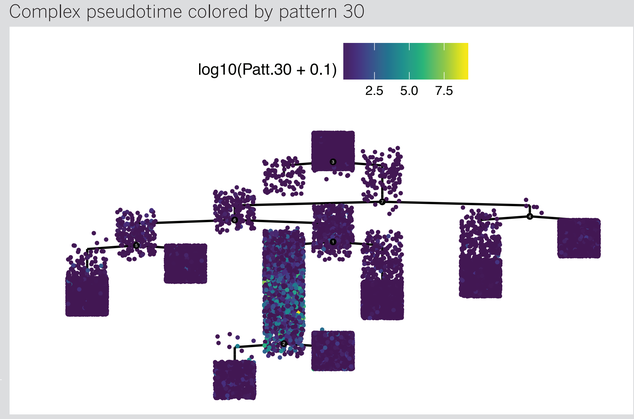




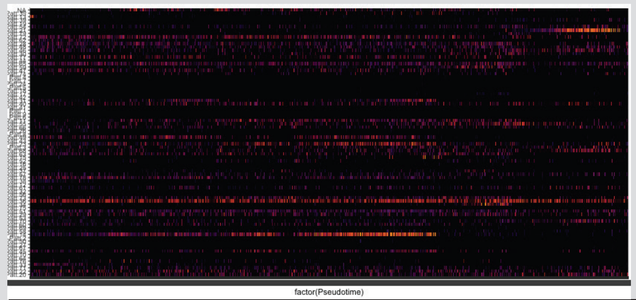


Pseudotemporal ordering of patterns reveals developmental trajectories and cell dynamics

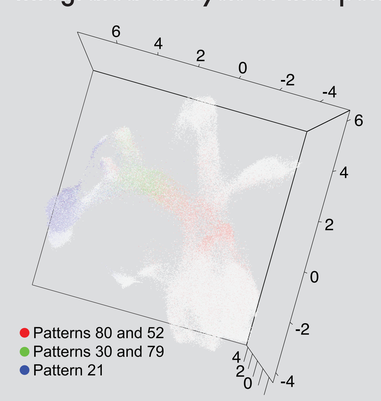
Complex pseudotime colored by pattern 30

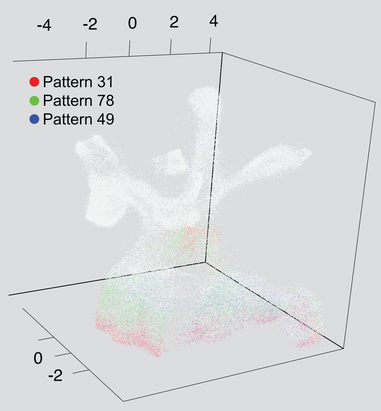


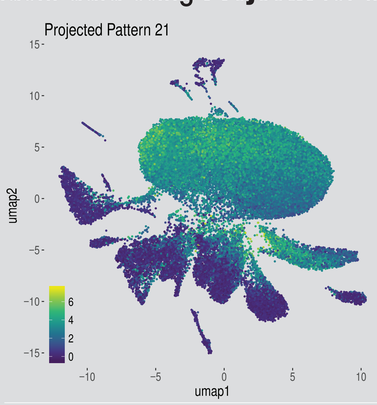
Heatmap of scCoGAPS pattern values correlated with complex pseudotime

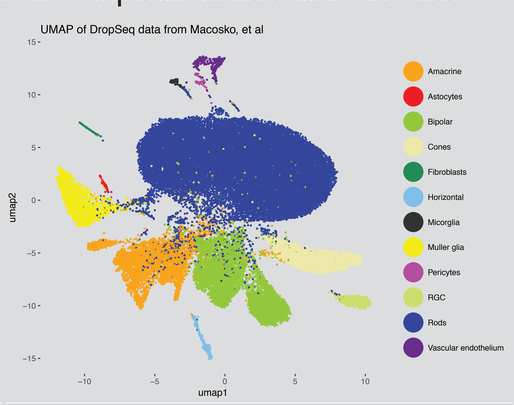


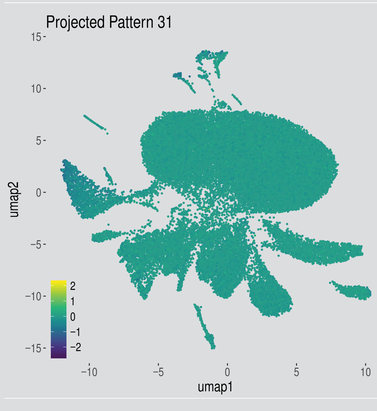
Integrated analysis of independent data using **ProjectR** for in silico experimentation and validation

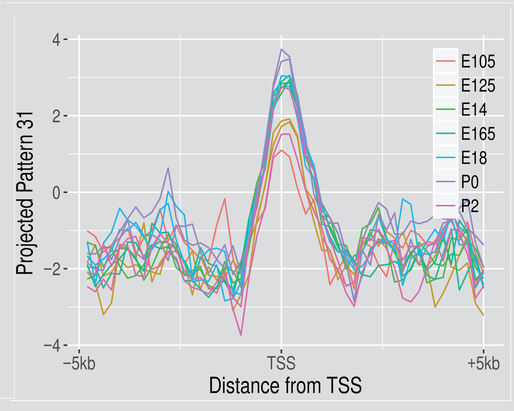










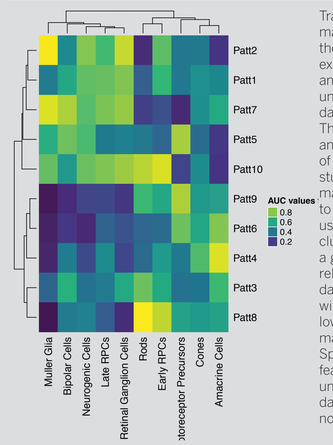


scCoGAPS uniquely accounts for gene reuse and sparsity via Bayesian NMF with an atomic prior

To infer common dimensions from a data matrix **D** of **N** genes and **M** conditions by factorization into a pattern matrix (**P**) and a corresponding amplitude matrix (**A**), the CoGAPS engine on which scCoGAP is built seeks **P** and **A** matrices whose product is from the distribution for **D**. That is, $D_{ij} = AP + \varepsilon_{ij}$ where ε_{ij} is independent, normal noise with mean zero and variance σ^2_{ij}. Thus, the rows of **P** form a set of non-orthogonal basis vectors that describe the patterns of coexpression across the samples in the column of **D**. The rows of **A** quantify the amount of the behavior of a gene explained by each of the pat terns (the rows of **P**). Gibbs sampling of **A** and **P** is implemented via Markov Chain Monte Carlo (MCMC) with simulated annealing. Matrix elements are propsed from a sparse atomic prior:

- Developed in Sibisi and Skilling (1997), the atomic domain is a line of infinite extent
- Each "atoms" weight on this domain is given by an exponential prior
- Elements in the **A** and **P** matrices maps to atoms based on the atoms location in the atomic domain
- Update steps enforce non-negative and sparse elements and thus sparsity of the **A** and **P** matric

Transfer learning methods can quantify specifity and sensitivity of pattern use in projected data



Transfer learning methods (TLMs) are able to relax many of the constraints of other methods by using the fact that if two datasets are related, there may exist mappings or features to connect the samples and relationships. As a result, these TLMs are uniquely suited for integrating -omic analysis across data platforms, modalities, and studies.
Thus, we implemented TLMs to developed ProjectR, an R package for integrated unsupervised analysis of high dimensional omic data from disparate studies. Projection can roughly be defined as a mapping or transformation of points from one space to another often lower dimensional space. ProjectR uses the relationships (e.g. principal components, clusters, metagenes, modules, etc) defined within a given high dimensional data set, to interrogate related biological phenomena in an entirely new data set. By leveraging relative comparisons within data type, ProjectR is able to extract shared low-dimensional features, while circumventing many issues arising from technological variation. Specifically, biologically meaningful relationships/features will stratify new data consistent with their underlying biological processes while artifacts or data specific relationships/features result in little to no information content.