Project Summary

Reuse and exaptation of co-regulated modules of genes or other cellular features are known to contribute to such diverse phenomena as body patterning, tissue organization, cellular physiology, and even paralogous functions in disparate tissues. High-throughput experimental techniques are capable of identifying co-regulated features across broad datasets including 1) gene expression profiling, 2) correlated epigenetic changes (ATAC-Seq), 3) reorganization of chromatin structure (HiC), and 4) quantification of protein expression and modification (proteomics). Exploring these features at single-cell resolution provides the first opportunity to identify and characterize the reuse of co-regulated features at the level of a multicellular organism.

Single-cell analysis of gene expression has demonstrated that population-level gene expression, and the 'transcriptional identity' of individual cells, arises from combinations of basis vectors that describe both dependent and independent contributions (1). Yet, the extent to which these basis vectors are shared/distributed throughout the human body is unknown. While many methods exist to deconvolve gene expression matrices into their constitutive patterns, most methods do not scale well to large datasets with complex sources of variation. This current limitation necessitates the ability to explore basis vectors learned on smaller datasets across larger datasets.

Transfer learning methods (TLMs) use previously learned knowledge from one or more sources to improve learning of a new target data. TLMs are able to relax many of the constraints of other methods by using the fact that if two domains are related, there may exist mappings or features that connect the samples and relationships, respectively(2). Thus, we have implemented TLM methodologies to perform integrated analysis of high dimensional multi-omic data in the R package ProjectoR.

ProjectoR uses the relationships defined within a given high dimensional data set, to interrogate related biological phenomena in an entirely new data set. Importantly, ProjectoR is agnostic to the source or type of basis vectors (e.g. principal components, metagenes, modules, etc). Instead ProjectoR uses the weights of these learned vectors across shared features from one dataset, to establish a feature representation on a target dataset. In this manner, basis vectors corresponding to meaningful biological variation can be compared directly, independent of laboratory of origin or technical artifacts. Projection of artifactual basis vectors result in little to no information content. Conversely, biological basis vectors stratify samples consistent with their underlying biological processes. We propose to adapt these TLMs to enable rapid comparisons of multiple data types, tissues, and even across species. To evaluate ProjectoR's performance at all three levels will require the generation of new datasets from tissues with demonstrated biological relationship in a model organism, as well as the use of publicly available data through the Human Cell Atlas.

Project aims, and how they address program goals

We will expand ProjectoR to allow users to explore pattern use/reuse at the scale of the HCA with minimal computational effort. Specifically, funding of our proposed project will help to address several of the HCA pilot project goals by enabling 1) direct comparisons of different sequencing library preparations, protocols, and resulting datasets, 2) evaluation of experimental replicability, and 3) enable systematic rapid comparison of shared basis vectors across tissues and samples as well as across species, potentially extending insights learned from model organisms directly into human cells.

We will benchmark ProjectoR performance using existing matched bulk RNA-Seq, and two sources of single-cell RNA-Seq data from the developing retina. We will learn basis vectors from these data, in conjunction with our co-investigator, Dr. Elana Fertig, the developer of the CoGAPs algorithm (3-5). Using these vectors, we will estimate the scalability of ProjectoR for use on the entire human cell atlas dataset and demonstrate the utility of this approach for a) identification of technical artifacts, b) classification of major and minor cell-types, c) characterization of continuous biological processes, and d) basis vector reuse via the following specific aims.

Aim I: *To identify and compare basis vectors (modules) of gene regulation in both bulk and single-cell gene expression measurements of the developing retina*

A) Projection of bulk-learned patterns into single-cell datasets to benchmark efficiency and information content of ProjectoR projections in developing mouse retina
    a) We will specifically develop a statistical framework to test for discrimination of major celltypes or lineages by a given pattern, and develop tools to identify technical artifact patterns.
B) To determine the extent of cellular variation that is captured by projection from bulk RNA-Seq, we will next identify basis vectors from the single-cell RNA seq data.
    a) We will compare basis vectors learned in single-cell RNA-Seq with those learned from our bulk RNA-Seq analysis using the PatternMatcher utility in ProjectoR to identify sources of variation not represented in our original basis vectors.

Aim II: *To determine the shared and distinct gene expression modules, at single-cell resolution, between tissues with a common developmental ontogeny, and across species using ProjectoR.*

A) To test whether basis vectors learned from one cellular context can be used to identify gene network reuse in a paralogous cellular context, we will:
    a) Establish a catalog of dissociated mouse hypothalamus single-cell RNA-seq profiles across developmental timepoints.
    b) Project learned basis vectors from both bulk and single-cell developing mouse retina into our hypothalamus cell catalog to identify collections of hypothalamic cells with shared/reused biological processes.
    c) Learn basis vectors from our hypothalamus cell catalog and compare to retina vectors using PatternMatcher and our developing statistical framework.
B) To evaluate the potential for cross-species basis vector comparison, we will use ProjectoR to project developing mouse retina basis vectors into publicly available human single-cell RNA-seq and ATAC-seq datasets.
    a) Determine whether basis vectors learned in mouse retina, can identify human cells undergoing similar biological processes.

Experimental plan and deliverables

This project will provide the essential computational framework to enable future research analyzing single-cell multi-omic data from across the human cell atlas. Specifically, we will continue to develop and extend ProjectoR to compare across the single-cell RNA-sequencing data generated through the HCA project, as well as extend to other single-cell data types (e.g. proteomics, chromatin, and metabolomic data). We will provide definitive benchmarking utilities and estimates of required resources for projection at the scale of the HCA.

Age	Replicate	Cells	Condition	Library Type	nCells
E12	1	None	Whole Retina	10x	1640
E14	1	None	Whole Retina	10x	16154
E14	3	Chx10-GFP(+)	RPCs	Smart-Seq2	244
E14	2	Chx10-GFP(+)	RPCs	Bulk-RNAseq	NA
E14	2	Chx10-GFP(-)	Post-mitotic	Bulk-RNAseq	NA
E16	1	None	Whole Retina	10x	5401
E18	1	None	Whole Retina	10x	10,574
E18	3	Chx10-GFP(+)	RPCs	Smart-Seq2	286
E18	2	Chx10-GFP(+)	RPCs	Bulk-RNAseq	NA
E18	2	Chx10-GFP(-)	Post-mitotic	Bulk-RNAseq	NA
P2	3	None	Whole Retina	10x	5835
P2	2	Chx10-GFP(+)	RPCs	Bulk-RNAseq	NA
P2	2	Chx10-GFP(-)	Post-mitotic	Bulk-RNAseq	NA
P2	3	Chx10-GFP(+)	RPCs	Smart-Seq2	280
P5	1	None	Whole Retina	10x	6756
P14	1	None	Whole Retina	10x	8103

Table 1: Summary table of preliminary developing mouse retina gene expression data using three distinct library preparations (bulk, Smart-Seq2, and 10x genomics). These data will be used to derive basis vectors corresponding to CoGAPS modules of gene co-regulation over retinal development. We will evaluate the projection of these learned basis vectors across the different data sets to benchmark the ProjectoR methodologies.

Leveraging our existing catalog of bulk and single-cell RNA-Seq data from the developing mouse retina (Table 1; described below), we have already learned basis vectors from bulk RNA-Seq via CoGAPS. This first-pass has identified 18 distinct basis

vectors corresponding to both biologically meaningful patterns of gene expression, as well as batch-specific effects and other sources of technical variation. We will use our existing Smart-Seq2, and 10x single-cell RNA-Seq catalogs into these 18 basis vectors to benchmark performance of our algorithms.

To derive basis vectors from our established mouse retina single-cell data, we will work with Dr. Elana Fertig to adapt CoGAPS for parallelization across higher-dimensional data (independent study). This will provide the opportunity to identify basis vectors on the significantly higher-resolution Smart-Seq2 and 10x single-cell mouse retina data. We will then use and develop support functions within ProjectoR to compare gene-wise basis vectors between experimental paradigms allowing us to identify discrete sources of variation not represented.

As a proof of concept for ProjectoR analysis across tissues, we will compare our retinal analysis to the mouse hypothalamus. The retina and hypothalamus share a common developmental origin and their specification arises only after retinal progenitors actively repress expression of hypothalamus-specific genes(6,7). Previous work from our collaborator, Dr. Seth Blackshaw, has generated an atlas of molecular markers that demarcate both progenitors and postmitotic cells in many different domains of the developing hypothalamus(8), allowing us to identify the position and identity of many of the cells that will be profiled.

We propose to generate single-cell RNA-Seq libraries from replicate (n=2) samples of whole dissociated mouse hypothalamus during development. Samples will be balanced with respect to sex and species fixed to match our existing mouse retina data.

Prior contributions in this area and preliminary results (not required)

We have developed the ProjectoR package (*in prep*; https://github.com/genesofeve/ProjectoR), to enable rapid transfer learning basis vectors from one biological dataset into another. To further extend this package, and evaluate its use across single-cell RNA-Seq data, we have in parallel established a compendium of RNA-Seq libraries (bulk and single-cell) from dissociated mouse retina at various time points (Table 1).

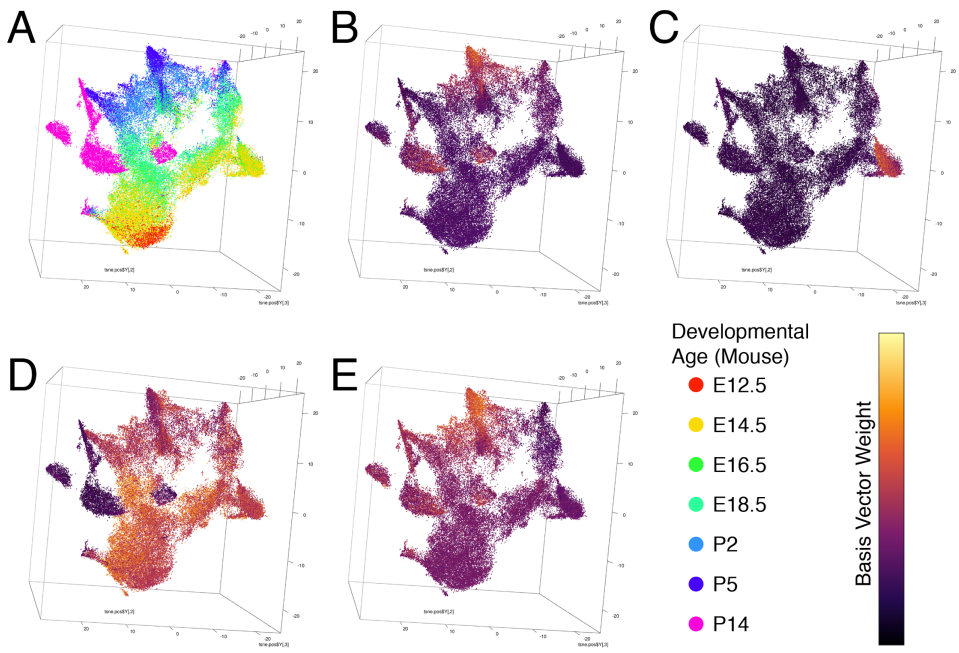

Figure 1: 3D t-stochastic neighbor embedding (tSNE) plot of 54,463 dissociated cells from select timepoints from developing mouse retina showing lineage commitment and maturation of retinal progenitor cells. (A) Cells are colored by developmental age at collection. B-E) cells are colored based on ProjectoR projections of select basis vectors learned from an independent bulk RNA-Seq study of mouse retina development. ProjectoR is able to identify cells with lineage-specific basis vector usage (B & C) as well as basis vectors describing both decreasing usage (D) and increasing usage (E) over developmental time, consistent with

To pilot these analyses, we have derived 18 basis vectors from the bulk RNA-Seq samples using CoGAPS. These learned vectors correspond to distinct sources of variation across our bulk RNA-Seq data: dynamic regulation over developmental time, specific differences between a sorted retinal progenitor population and a more differentiated pool of cells at each timepoint, and technical batch effects. We used the gene-weights of the CoGAPs modules as basis vectors for ProjectoR and projected single-cell RNA-Seq libraries created from our 10x collection of whole dissociated mouse retina across development into these. Projection of >54,000 cells took less than 1 minute (55.5 seconds). Using these projections, we were able to identify lineage-specific patterns of gene expression (Figure 1B-C), as well as patterns both correlated and

anticorrelated with developmental age (Figure 1D-E). Importantly, CoGAPS modules corresponding to technical artifacts demonstrated no enrichment across cells.

Description of commitment to full sharing of primary data, metadata, methods, and software

Results will be published and disseminated through public lectures. All primary data will be archived on the SRA and Gene Expression Omnibus. Source code and software tools will be available through GitHub and will be made available for integration into any required HCA workflows.

References cited

1. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol. 2016 Nov 8;34(11):1145–60.

2. Pan W, Xiang EW, Liu NN, Yang Q. Transfer Learning in Collaborative Filtering for Sparsity Reduction. AAAI. 2010.

3. Fertig EJ, Ding J, Favorov AV, Parmigiani G, Ochs MF. CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. Bioinformatics. 2010 Nov 1;26(21):2792–3.

4. Fertig EJ, Ren Q, Cheng H, Hatakeyama H, Dicker AP, Rodeck U, et al. Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma. BMC Genomics. BioMed Central; 2012 May 1;13(1):160.

5. Stein-O'Brien G, Carey J, Lee W-S, Considine M, Favorov A, Flam E, et al. PatternMarkers and Genome-Wide CoGAPS Analysis in Parallel Sets (GWCoGAPS) for data-driven detection of novel biomarkers via whole transcriptome Non-negative matrix factorization (NMF). bioRxiv. Cold Spring Harbor Labs Journals; 2016 Oct 26;:doi:10.1101–083717.

6. de Melo J, Zibetti C, Clark BS, Hwang W, Miranda-Angulo AL, Qian J, et al. Lhx2 Is an Essential Factor for Retinal Gliogenesis and Notch Signaling. J Neurosci. Society for Neuroscience; 2016 Feb 24;36(8):2391–405.

7. Roy A, de Melo J, Chaturvedi D, Thein T, Cabrera-Socorro A, Houart C, et al. LHX2 Is Necessary for the Maintenance of Optic Identity and for the Progression of Optic Morphogenesis. J Neurosci. Society for Neuroscience; 2013 Apr 17;33(16):6877–84.

8. Shimogori T, Lee DA, Miranda-Angulo A, Yang Y, Wang H, Jiang L, et al. A genomic atlas of mouse hypothalamic development. Nat Neurosci. Nature Research; 2010 Jun 1;13(6):767–75.