Summary

Single-cell analysis has demonstrated that population-level gene expression and the 'transcriptional identity' of individual cells, arises from combinations of basis vectors[1]. Reuse and exaptation of co-regulated modules of genes or other cellular features can contribute to diverse phenomena as patterning, tissue organization, cellular physiology, and paralogous functions in disparate tissues. The extent to which basis vectors are shared/reused throughout the human body remains under-explored. Exploring these features at single-cell resolution provides an opportunity to identify and characterize the reuse of co-regulated features.

While many methods exist to deconvolve gene expression into patterns, most methods do not scale to large datasets with complex sources of variation. Further, basis vector identification and evaluation of models is limited to technical metrics with little consideration for the common or disparate biological properties described by each approach. Tools are needed to benchmark the biological activity described by models derived from independent algorithms. Current computational limitations necessitate the ability to rapidly explore basis vectors learned on smaller datasets across larger datasets, and requires the development of statistical and visualization frameworks upon which to evaluate and compare learned models derived from different computational approaches.

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 [2]. We implemented TLM methodologies to perform integrated analysis of high dimensional multi-omic data in the R package ProjectoR. ProjectoR uses relationships defined within a given data set, to interrogate related biological phenomena in an new data set. Importantly, ProjectoR is agnostic to the source or type of basis vectors (e.g. principal components, metagenes, modules, latent spaces, etc). Instead ProjectoR uses the weights of learned vectors across 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 artefactual basis vectors, corresponding to technical sources of error in the test dataset, result in little to no information content when projected into the target set. Conversely, biological basis vectors stratify samples consistent with their underlying biological processes. Furthermore, basis vectors learned by independent methods on disparate training sets can be projected into a common test dataset and directly compared. We propose to adapt these TLMs to enable rapid comparisons of multiple data types, bulk and single cell library preparation techniques, developmental time, sex, cell types, and even across species in a well characterized model system that provides an ideal setting to compare. Additionally as part of an open collaborative network, we propose to develop and extend ProjectoR as a statistical framework to evaluate and compare basis vectors learned from disparate algorithms.

Project aims, and how they address program goals

Our overall aim for this project is to provide a statistical framework and benchmark datasets to compare methods for identification of basis vectors from single cell data, and determine the biological or artefactual relevance of these vectors using a well-studied mammalian system. We will expand ProjectoR to explore pattern use/reuse at the scale of the HCA with minimal computational effort. Specifically, we will address several of the project goals by 1) developing tools to extract and analyze data organized by genes, cells, or tissues of interest, 2) supporting analytical methods and machine learning approaches, 3) generating curated benchmark datasets, and experimental datasets that directly address computationally-guided questions in quality control, reproducibility, or multimodal integration, and 4) developing new computational approaches comparing and normalizing genomic and imaging data across assays, subjects, and species.

Aim I: *To extend existing benchmark single cell RNA-Seq datasets of the developing retina across library techniques, developmental stages, and species.* Thus, allowing for both discrete cell type identification at

multiple hierarchical levels, as well as continuous properties such as pseudotemporal state, pseudo-spatial state, differentiation state and progenitor competency.

A) We will extend our current catalog of mouse retina single cell data to include sci-RNA-Seq[3] and sci-nucRNA-Seq in mouse developing retina.
B) We will conduct bulk RNA-Seq and single cell sci-RNA-Seq in de-identified human postmortem tissue.
C) We will evaluate the potential for cross-species basis vector comparison using ProjectoR to project developing mouse retina basis vectors into publicly available and newly generated human single-cell RNA-seq and ATAC-seq datasets from retina and other developmentally related tissues.
    a) We will determine whether basis vectors learned in mouse retina, can identify human cells undergoing similar biological processes.

Aim II: *To benchmark ProjectoR using basis vectors (models) from developing mouse and human retina learned from tools across collaborative network*

A) We will evaluate ProjectoR performance on output from various collaborative network models
    a) Benchmark computation speed on projections in exponentially scaled data sets with corresponding increases in dimensionality and complexity.
    b) Using benchmark datasets and *a priori* knowledge we will access accuracy of biological assignments of basis using metrics of sensitivity and specificity of projections to evaluate statistical power.
B) We will identify technical vs biological models and determine methods to QC individual cells via projection of models.

Aim III: *To develop model comparison statistics, pathway enrichment testing, and novel basis vector visualizations in ProjectoR*

A) We will compare collaborative network analytical tools using ProjectoR projections on benchmark datasets to highlight optimal usage for specific biological questions.
    a) We will develop a statistical framework to test for discriminatory power for major cell types or lineages by a given pattern, and develop tools to identify technical artifact patterns.
B) We will develop ProjectoR visualizations to explore shared biological features across benchmark datasets, as well as public and published single cell RNA-Seq datasets to create comprehensive model via projection (e.g. PCA as example).

Prior contributions in this area and preliminary results

Age	Replicate	Sorting	Condition	Library Type	nCells
E11	1	-	Whole Retina	10x Genomics	9,366
E12	2	-	Whole Retina	10x Genomics	26,576
E14	1	-	Whole Retina	10x Genomics	16,154
E14	3	Chx10-GFP(+)	RPCs	Smart-Seq2	244
E14	2	Chx10-GFP(+)	RPCs	Bulk-RNAseq	-
E14	2	Chx10-GFP(-)	Post-mitotic Cells	Bulk-RNAseq	-
E16	1	-	Whole Retina	10x Genomics	5,401
E18	1	-	Whole Retina	10x Genomics	10,574
E18	3	Chx10-GFP(+)	RPCs	Smart-Seq2	286
E18	2	Chx10-GFP(+)	RPCs	Bulk-RNAseq	-
E18	2	Chx10-GFP(-)	Post-mitotic Cells	Bulk-RNAseq	-
P0	1	-	Whole Retina	10x Genomics	9,926
P2	2	-	Whole Retina	10x Genomics	17,968
P2	3	Chx10-GFP(+)	RPCs	Smart-Seq2	280
P2	2	Chx10-GFP(+)	RPCs	Bulk-RNAseq	-
P2	2	Chx10-GFP(-)	Post-mitotic Cells	Bulk-RNAseq	-
P5	1	-	Whole Retina	10x Genomics	6,756
P8	2	-	Whole Retina	10x Genomics	11,729
P14	1	-	Whole Retina	10x Genomics	8,103
				TOTAL	123,363

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 using tools developed by members of our collaborative network. We will evaluate the projection of these learned basis vectors across the different data sets to benchmark ProjectoR methodologies and develop novel visualizations and statistical tests for model comparison.

We developed the ProjectoR package to enable rapid transfer learning of basis vectors from one biological dataset into another. On top of the generic function, the base projectoR function is currently coded for regression, PCA, CoGAPS, and clustering-derived basis vectors. The ProjectoR package currently contains statistics to evaluate cluster/set overlap (intersectoR), methods to transform learned vectors to more accurately reflect biology (rotatoR), evaluate correlation (correlateR), and standardize gene composition/feature mapping across datasets (geneMatchR).

We have in parallel established a compendium of RNA-Seq libraries (bulk and

single-cell) from dissociated mouse retina at various time points using different library preparations (Table 1). Leveraging our existing catalog of bulk and single-cell RNA-Seq data from the developing mouse retina (Table 1; described below), we identified 18 basis vectors corresponding to both biologically meaningful patterns of gene expression, batch-specific effects, and sources of technical variation (Figure 1). Gene-weights were projected into 10x single-cell RNA-Seq. Projection of >120,000 cells took 55.5 seconds. Using these projections, we identified lineage-specific patterns of gene expression (Figure 1B-C), as well as patterns both correlated and anti-correlated with developmental age (Figure 1D-E). Importantly, modules corresponding to technical artifacts demonstrated no enrichment across cells.

Proposed work and deliverables

We will develop and extend ProjectoR to compare single-cell RNA-sequencing data generated through the HCA project, as well as other single-cell data types (e.g. proteomics, chromatin, and metabolomic data). We will provide definitive benchmarking utilities for analysis at the scale of the HCA.

Benchmarking ProjectoR using existing matched bulk RNA-Seq, two sources of single-cell RNA-Seq data from the developing retina, we will learn basis vectors from these data, in conjunction with our collaborative network of investigators (Fertig, Greene, Patro) [4, 5, 6]. 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. We will classify basis vectors from training sets made using different library preps to:

- Identify vectors corresponding to technique-specific artefacts
- Identify vectors corresponding to quality of input data
    - Doublets -> incompatible vectors in same cell
    - Low quality vectors (correlated with metrics of poor scRNA-Seq performance)

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. Finally, we will develop a statistical approach to extend projections *beyond directly mappable features* by employing a generalized additive model (GAM) in which pattern weights, projected using only a subset of mapped features, are used as explanatory variables to identify other features whose expression may be dependent.

Proposal for evaluation and dissemination of methods, resources, or results.

All primary and benchmark data will be deposited with the HCA data coordination platform, as well as freely accessible online, and archived on the NCBI short read archive along with all available metadata. Learned basis vectors, workflows, and analyses will be publicly available on a custom web site, archived in conjunction with published manuscripts, and posted on Biorxiv. Stable software builds will be released in the ProjectoR package.

Statement of commitment to share proposals, methods, data, and code

We commit to making all scripts, code, and software available via github under the permissive MIT license consistent with the standard operating procedures of the Goff Lab. The proposal has been developed in the open in conjunction with the collaborative network and is publicly shared at https://github.com/gofflab/czi-rfa-2017/. We acknowledge that the proposal will be made public. Proposed work in benchmark data annotation, model identification and characterization, and analysis will be conducted on a publicly available github repository as they are generated.

Figures

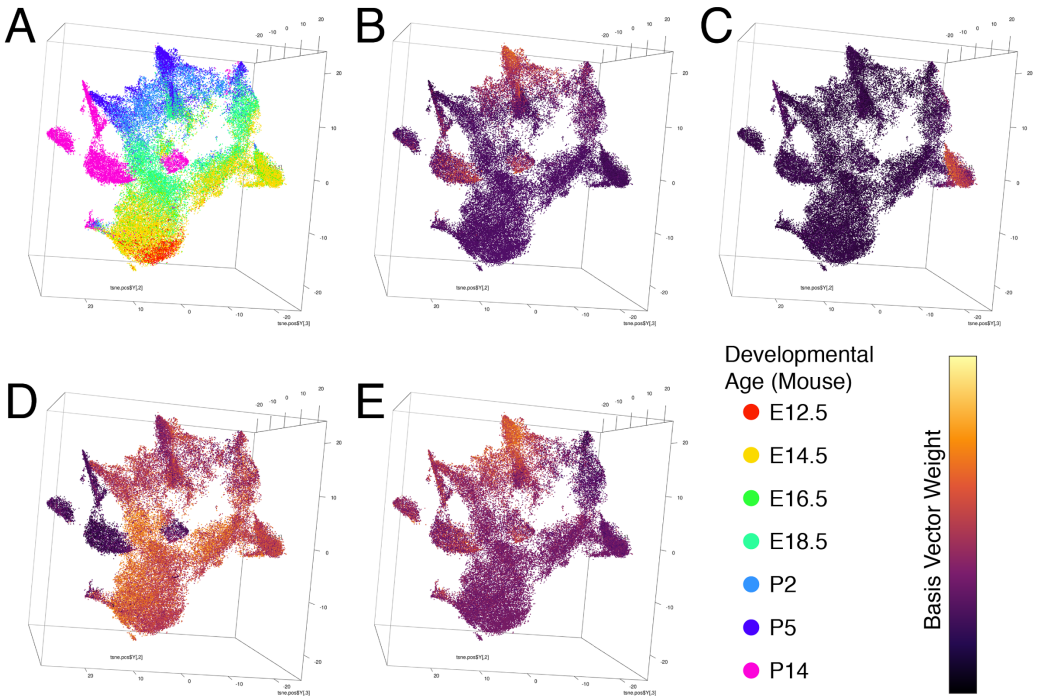

Figure 1: 3D t-stochastic neighbor embedding (tSNE) plot of 54,463 dissociated cells from select timepoints from developing mouse retina showing lineage commitment and maturation of retinal progenitor cells. (A) Cells are colored by developmental age at collection. B-E) cells are colored based on ProjectoR 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

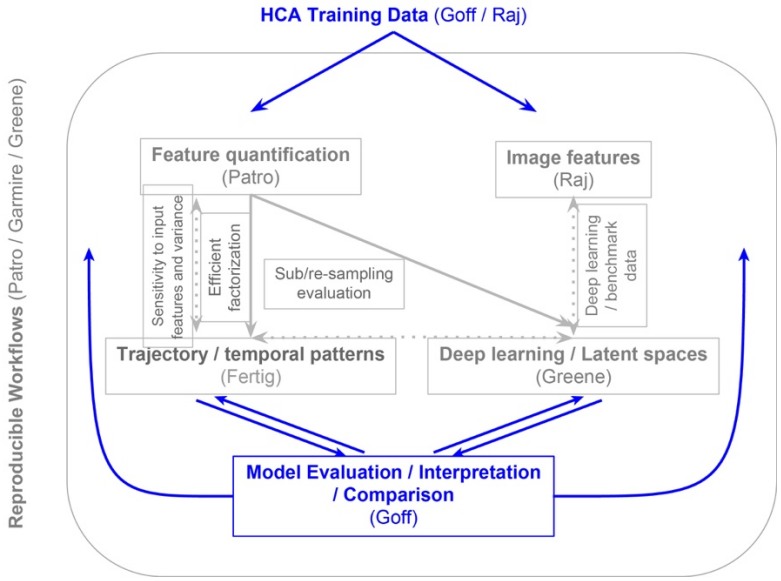

Figure 2: Proposed collaborative network. Contributions from this proposal are highlighted in blue.

References cited

1. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34,** 1145–1160 (2016).
2. Pan, D. An integrative framework for continuous knowledge discovery. *Journal of Convergence Information Technology* (2010).
3. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357,** 661–667 (2017).
4. Fertig, E. J., Ding, J., Favorov, A. V., Parmigiani, G. & Ochs, M. F. CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics* **26,** 2792–2793 (2010).
5. Fertig, E. J. *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* **13,** 160 (2012).
6. Stein-O'Brien, G. *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* doi:10.1101–083717 (2016). doi:10.1101/083717