Progress Report: Rapid exploration, interpretation, and comparison of discrete single cell transcriptional basis vectors

This manuscript was automatically generated from gofflab/goff-czi-report@5cfd51c on August 28, 2018.

Authors

- **Loyal Goff**

  0000-0003-2875-451X · loyale · loyalgoff

  Department of Neuroscience, Johns Hopkins University; Institute for Genetic Medicine, Johns Hopkins University; Kavli Neurodiscovery Institute, Johns Hopkins University · Funded by Grant XXXXXXXX

Abstract

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.

Specific Aims

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.

1. We will extend our current catalog of mouse retina single cell data to include sci-RNA-Seq3 and sci-nucRNA-Seq in mouse developing retina.
2. We will conduct bulk RNA-Seq and single cell sci-RNA-Seq in de-identified human postmortem tissue.
3. 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 and 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

1. We will evaluate ProjectoR performance on output from various collaborative network models.
2. Benchmark computation speed on projections in exponentially scaled data sets with corresponding increases in dimensionality and complexity.
3. 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.
4. 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

1. We will compare collaborative network analytical tools using ProjectoR projections on benchmark datasets to highlight optimal usage for specific biological questions.
2. 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.
3. 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).

CZI Computational Tools Meeting

Developing Retina Benchmark Dataset

Deliverable Products

Latent Spaces Group

Refined Definition of biological latent space

Standardized requirements and file formats for latent spaces

Goff Lab Direct Collaborations

Single cell RNA-Seq atlas of the developing murine retina

Preprint: https://www.biorxiv.org/content/early/2018/07/30/378950 Available: https://github.com/gofflab/developing_mouse_retina_scRNASeq Status:

ProjectR

Preprint: https://www.biorxiv.org/content/early/2018/08/20/395004.1 Available: https://github.com/genesofeve/projectR Status:

scCoGAPS

Preprint: https://www.biorxiv.org/content/early/2018/08/20/395004.1

Computational Tools Network Enhanced Interactions

• Increased collaborative interactions and shared trainees with Dr. Elana Fertig lab - Two groups cross training in experimental and computational single cell analysis

• Postdoctoral trainee Dr. Stein-O'Brien's participation in the Jamboree enabled fruitful discussion and code for projectR.

• Two collaboratively written publications with Dr. Elana Fertig - Release, description, and biological exploration of a comprehensive single cell RNA-seq dataset from the developing mouse retina - Introduction of scCoGAPS and ProjectR for latent space discovery across single cells and transfer learning across experiments.

• Interactions with CZI cellxgene development group Re: possible latent space visualizations

References

1. **Revealing the vectors of cellular identity with single-cell genomics**
Allon Wagner, Aviv Regev, Nir Yosef
Nature Biotechnology (2016-11) https://doi.org/10.1038/nbt.3711

2. **An Integrative Framework for Continuous Knowledge Discovery**
Ding Pan
Journal of Convergence Information Technology (2010-05-31) https://doi.org/10.4156/jcit.vol5.issue3.7