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

This manuscript was automatically generated from gofflab/goff-czi-report@0734206 on February 15, 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 vectors1. 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.

References