Search for and transformation of human cells and cell types with latent space representations

This manuscript (permalink) was automatically generated from greenelab/czi-seed-rfa@def30c1 on October 19, 2018.

Authors

- **Casey S. Greene**
  0000-0001-8713-9213 · cgreene · greenescientist
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania

- **Stephanie C. Hicks**
  0000-0002-7858-0231 · stephaniehicks · stephaniechicks
  Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

- **Rob Patro**
  0000-0001-8463-1675 · rob-p
  Department of Computer Science, Stony Brook University

- **Elana J. Fertig**
  0000-0003-3204-342X · ejfertig · FertigLab
  Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, School of Medicine, Johns Hopkins University; Department of Applied Mathematics and Statistics, Whiting School of Engineering, Johns Hopkins University

- **Jane Roe**
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

Abstract

Instructions: Describe your collaborative project, highlighting key achievements of the project; limited to 250 words.

Five Key References

Project Team (750 words each)

Stephanie Hicks

- Title: Assistant Professor
- Degrees: PhD
- Type of organization: Academic
- Tax ID: 52-0595110 (JHU)
- Email: shicks19@jhu.edu

Co-PI role:

Dr. Stephanie C. Hicks is an Assistant Professor of Biostatistics at the Johns Hopkins Bloomberg School of Public Health. She is an expert in statistical methodology with a strong track record in processing and analyzing single-cell genomics data, including extensive experience developing fast, memory-efficient R/Bioconductor software to remove systematic and technical biases from scRNA-seq data [1]. This work resulted in a K99/R00 grant from the National Human Genome Research Institute (NHGRI) and funding from the Chan-Zuckerberg Initiative to develop computational tools for the Human Cell Atlas in the previous round of funding.

During Aim 1 of the project period, Dr. Hicks will work together with Co-PIs (Greene, Love, and Patro) to implement fast search algorithms to quantify differences between a reference transcriptome map (the Human Cell Atlas) and non-reference transcriptome maps from other samples of interest, similar to the idea of using a reference genome to identify genomic differences in between a reference and non-reference genome. Globally quantifying differences between transcriptome maps is important because it allows for quantification of differences at the population or individual level between, for example, ten transcriptome maps from individuals with a particular phenotype to be compared to the Human Cell Atlas reference map (or other control transcriptome maps if available). Our metric to quantify differences will depend on the distributions of cell expression within and between individuals, which Dr. Hicks has extensive experience with [2]. We will leverage not only the cell-to-cell correlation structure within one transcriptome map (or human individual), but also the correlation structure across transcriptome maps (or multiple human individuals), which will share common latent spaces across individuals for a particular phenotype.

Our initial approach will be to use linear mixed models to account for the correlation structure within and between transcriptome maps. The statistical method will be fast, memory-efficient and will scale to billions of cells because we work in the latent space with a significantly reduced number of dimensions (instead of billions, just hundreds or thousands). Finally, this work will be in close collaboration with the Co-PIs (Fertig, Goff, Love, and Greene) working on Aim 2, leading to an iterative process of updating the latent spaces generated in Aim 2 and updating the methods to quantify differences between transcriptome maps in the latent spaces.

During Aim 3, Dr. Hicks and other Co-PIs (Love and Greene) will implement the methods developed in Aims 1 and 2 into R/Bioconductor software packages. The software will be fast, scalable, and memory-efficient because will lever the computational tools previously developed by Bioconductor for single-cell data access to the HCA (add name here), data representation (SingleCellExperiment, beachmat, DelayedArray, HDF5Array and rhdf5) and data assessment and ameliorization of data quality (scater, scran, DropletUtils). *add more about Bioc*

Dr. Hicks will hire a dedicated postdoctoral research fellow or software developer to take primary responsibility for the development and continued maintenance of the software packages.

Elana Fertig

- Title: Associate Professor
- Degrees: PhD
- Type of organization: Academic
- Tax ID: 52-0595110 (JHU)
- Email: ejfertig@jhmi.edu

Co-PI role:

Loyal Goff

Casey Greene (Submitter)

Tom Hampton

Michael Love

- Title: Assistant Professor
- Degrees: Dr. rer. nat.
- Type of organization: Academic
- Tax ID: 56-6001393 (UNC)
- Email: michaelisaiahlove@gmail.com

Co-PI role:

Rob Patro

Proposal Body (2000 words)

- Base enabling technologies:
  - Low dimensional representations (Elana)

Latent space determination relies on computational techniques that learn low dimensional structure from high dimensional data. A wide variety of computational methods for dimensionality reduction have been applied to single cell data. These techniques fall largely into two classes: (1) linear techniques based upon matrix factorization and (2) non-linear techniques based upon manifold learning. Each specific technique selected for analysis will reveal distinct structures in the data. Much of the current research emphasizes demonstration of optimality of a single method. In contrast to this prevailing view, we posit that each method reveals a unique set of features associated with a set of *a priori* unknown biological processes. Together, the distribution of the ensemble of low dimensional representations from these methods will uncover the multiple, concurrent biological processes from high dimensional datasets. Previously, we have developed a common language for interpretation of matrix factorization methods to facilitate a unification of latent space methods [3]. The latent space team in the previous round of funding (containing PIs Fertig, Goff, Greene, and Patro) is also developing standardized output formats for these distinct methods. Together, this previous work will facilitate practical unification and biological assessment of latent spaces in this seed network.

Numerous factors impact the biological accuracy of latent spaces learned from dimension reduction techniques. Two notable examples include the preprocessing of the input data (Aim 1) and dimensionality of the low dimensional representation (Aim 2). The impact of each of these components on inference must be quantified and isolated before latent methods can be compared or unified. To that end, we select a single dimension reduction technique as the foundation for the technologies developed in our seed network. We use the Bayesian, non-negative matrix factorization method scCoGAPS [4] (PI Fertig). Previous work has demonstrated the robustness of this approach in inferring biologically relevant low dimensional representations of perturbation [6] and time course [8] from bulk datasets, leading to the winning solution in the HPN DREAM8 challenge [10]. Moreover, recent extension of this method to single cell data in the developing mouse retina distinguished simultaneously cellular identity, dynamic trajectories, and cell state within a single analysis [4]. Notably for the work proposed in this seed network, this method includes an uncertainty estimate that can be readily modified to account for measurement-specific technical variation [11]. Moreover, this same study demonstrated that low dimensional factorizations of tumor populations separate tumor samples relative to normals whereas higher

dimensionalities separate tumor subtypes. Thus, CoGAPS has been shown to yield different factorizations reflective of biological hierarchies at different dimensionalities, rather than having a single dimensionality with a “correct” factorization. While technology development will focus on scCoGAPS, the tools developed as part of the seed network will be generalizable and compared to additional linear and non-linear methods developed in the literature **ADD RELEVANT CITATIONS** and by members of the broader CZI consortium. This single-method focus with subsequent expansion to additional methods will also facilitate educational initiatives to advance broader understanding of the interpretation of latent space solutions from high dimensional data (Aim 3).

- Fast & improved quantification (Rob / Mike)

About fast and improved quant Existing approaches for quantification from scRNA-seq data using tagged-end end protocols (e.g. 10x Chromium, drop-Seq, inDrop, etc.) have no mechanism for accounting for reads mapping between multiple genes in the resulting quantification estimates, and therefore such reads are typically removed **CORRECT?**. This reduces quantification accuracy, and leads to systematic biases in gene expression estimates that correlate problematically with the size of gene families and gene function **Cite new alevin pre-print**. We have recently developed a novel approach for quantification from tagged-end data that accounts for reads mapping to multiple genomic loci (~15-25% of the reads in a typical experiment) in a principled and consistent way. We propose to expand on this work in a number of ways, building these capabilities into a production quality tool for the processing of scRNA-seq data. (1) Exploration of alternative models for UMI resolution, including a maximum likelihood (as opposed to maximum parsimony) model, (2) Development of new approaches for quality-control and filtering using features from the the UMI-resolution graph, (3) Creation of a compressed and indexible data structure for the UMI-resolution graph to allow direct access, query, and, eventually, search.

The technologies to improve quantification will have a critical impact on the outcomes of latent spaces. However, there are currently no standardized, quantitative metrics to determine relative uncovering of biology from low dimensional representations. We have developed new transfer learning methods to quantify the extent to which latent space representations from one set of training data are represented in another [5]. These tools provide a strong foundation to enable biological quantification of latent space representations by quantifying the extent to which those spaces transfer across datasets of related biological contexts.

Aim 1

- Fast search: (in low dimensions, quantifying differences between case and reference maps, twist: shared latent spaces / k-mers)
  - Differences b/w maps (Stephanie)
  - New models for UMI deduplication accounting for transcript-level information (Rob)
    - Parsimony & likelihood based, integrated with gene-level uncertainty
  - Everything FAST! API for search against HCA reference? (Rob)

- k-mer / quantified latent spaces (Casey / Rob)

Aim 2

Rationale: Biological systems are comprised of common cell types with overlapping molecular phenotypes. The function of these systems are determined by the interactions between these complex components, rather than a single gene or cell. This suggests that fundamental biological mechanisms may broadly contribute to an observed state, with context-specific modifiers conferring selective suceptiability to disease. Latent space techniques are poised to reveal these fundamental mechanisms in the broad survey of single cell data across model systems and cellular contexts in the Human Cell Atlas. We hypothesize that the features learned from these techniques will define constitutive basis vectors that reflect discrete biological processes or features. Thus, these basis vectors will be shared across different biological systems, with context-specific perturbations indicating pathogenic differences in disease. *We propose a central suite of statistics for assessment and interpretation of latent space tools to define the basis vectors of biological systems.*

Quantifying latent space estimation with transfer learning: A critical challenge to latent space methods is the quantification of methods performance. Numerous computational metrics have been developed to assess convergence of the low dimensional estimation. However, these metrics do not quantify whether the features in a low dimensional representation of scRNA-seq data represent biological processes in the measured system. The performance of these methods can be quantified directly in datasets for which cell types and states are known (e.g., perturbation experiments, controlled admixture experiments, etc). However, these annotations are lacking in most biological datasets limiting any such quantification. Transfer learning methods have been developed in machine learning to relate features learned in a source dataset to those in a new, target dataset in order to transfer annotations from one context to another. In this project, we will adapt these methods to quantify the performance of latent space methods by the extent to which learned low dimensional features from a source dataset transfer to a target dataset in a related biological context. We will benchmark the performance of the resulting metric on simulated datasets, cross-validation in scRNA-seq datasets with known cell types and states, and cross-study validation of systems in related biological contexts with known cell types and states. Gene set enrichment methods will also be used to explore the relevant biological processes described by individual basis vectors, and related bases will be identified through clustering and exploratory approaches in these benchmark datasets. Our transfer learning based metric will be piloted on low dimensional representations learned with scCoGAPS and then applied to a broader suite of latent space tools. We will release software for this transfer learning quantification of latent space representations in R and Python using standard latent space file formats developed by our team in the first year of HCA funding.

Dimensionality estimation: Dimensionality reduction methods are sensitive to the number of low features learned in each dataset. Many computational techniques optimize dimensionality by

creating a cost function which penalizes models with higher number of features. Similar to the quantification metrics, these penalty terms do not reflect the extent to which features learned at a given dimensionality reflect biology. Moreover, many systems may have more than one biologically accurate low dimensional representation. Such multiple truths in data would be particular prominent in systems that can be subdivided into hierarchical classifications. For example, in the case of cancer we observed that a low dimensional representation of bulk data learned from CoGAPS distinguished cancers from normals whereas a higher dimension distinguished tumor subtypes [11]. Both of these low dimensional representations are equally valid, and each reflects different biological features in the data. To find these multiple truths, we will develop a parallel framework to run scCoGAPS for multiple dimensionalities and quantify performance with our transfer-learning based metric on random subsets of the data. The dimensions with greatest cross-validated feature robustness will be retained as the optimal dimensionalities for each dataset. We will develop software to enable this cross-validation dimensionality estimation across multiple latent space methods. We note that this same software will provide a robust tool to define ensembles of low dimensional representations that reflect underlying biology learned across multiple latent space methods. **Rob: I'm not sure if you want to fill in some of your ideas re persistent homology instead. Very open to that idea and think it may be a nice, more efficient methodology than what's proposed here.**

Search tool for latent spaces and reference cell types: **Loyal, Casey – what are the datasets that will be used for this – I would think all healthy cells in a single system to enable quantification of context-specific in the next part of this aim.** Comprehensive identification of basis vectors across conditions is an area of active research for our group in the previous funding period. We will use scCoGAPS and other tools developed within our collaborative network to establish a compendium of basis vectors across our single cell catalog. Ensembles of the low dimensional features that represent robust biological features across methods using methods described above will be preserved as the 'biological basis' of the Human Cell Atlas. The weights of these bases will be correlated across all available metadata attributes for each cell to identify basis vectors that are associated with specific cellular contexts, disease states, technical parameters, or other phenotypic features. A reference catalog of gene weights for specific cell types will be defined by the set of basis vectors associated with cellular identity in datasets with known ground truth. We will adapt the software we developed for transfer learning of features from bulk data recount [12] to facilitate querying of signatures in new user-defined datasets. As datasets accumulate and methods are refined, the biological basis and reference catalog of gene weights will evolve over time. To enable reproducible research leveraging HCA, we will implement a content-based versioning system, which identifies a reference catalog by the gene weights and gene/transcript cDNA sequences using a hash function. Such a versioning and provenance identification and detection framework has proven successful in the bulk RNA-seq context [13].

Differentiating context-specific latent spaces from latent spaces that are universal across biological contexts: The search tool to define reference cell types based upon latent spaces was defined for

healthy tissues from XXX (some control). Deviations of common cell types or states from the healthy baseline in other populations will indicate context-specific alterations, which may be associated with disease. To identify potentially pathogenic responses in target datasets, we will implement a random forest classifier into our transfer learning method to segregate cells based on their usage of disease-associated basis vectors after projection. In other cases, disease may arise from changes in variation reflective of inter-cellular heterogeneity. Therefore, we will also develop methods to quantify variation from latent space vectors. Both methods will be incorporated in our latent space search tool. **Loyal: I'm not sure if this is what you had in mind. It may also be that these are reflected in the hierarchy of dimensionality – may want to incorporate here.**

- Reference Cell types
    - Cell-type summarized expression profiles (Mike, Loyal)
        - A 'reference catalog' of reduced dimensional representations
    - Versioning & provenance of cell types / features as the reference dataset changes (Mike)
    - 'Power-user' application of latent-space discovery in novel dataset and projection of HCA into new learned spaces.

Aim 3

Rationale: Low dimensional representations in provide a powerful means to analyze scRNA-seq and HCA data to make tasks faster, perform more biologically grounded analyses, and provide interpretable summaries of complex high-dimensional data. However, using these capabilities to the fullest extent requires an investment in robustness and evaluation, integration with widely used toolkits, and in a scalable education effort that can reach students from the undergraduate level and beyond. *We propose to enhance software usability and develop open instructional materials that we will use to deliver short-course training that includes the topics of single cell profiling, machine learning methods, low-dimensional representations, and tools developed by our group in response to this RFA.*

- Delivery
    - Training / teaching (scRNAseq, low-dimensional representations, RFA-developed tools) (Tom) Although the HCA data set will greatly increase the rate of biological discovery across many biomedical fields, a background in bioinformatics will be required to these data effectively. Our mission is to make this obstacle easy for most biologists and physician scientists to overcome, and to enable scientists who already possess a background in bioinformatics to use software created in this project effectively. We will address these goals by offering short courses and creating distributed communities.
        - Short Courses Short courses (3-5 days) will introduce all features of bioinformatics required to access HCA and use the tools developed in this application successfully, such as
        - Visualization and Exploration of High Dimensional Data
        - R Statistical Programming Environment

        - UNIX
        - Statistical Approaches for High Dimension Biomedical Data
        - Gene Set and Pathway Analysis
        - Low dimensional representations of high dimensional data
        - R and UNIX Tools Specific to HCA Data Analysis
        - scRNAseq tools
        - RFA-developed tools
        - Course Project The course project will provide a hands-on experience of retrieving data from HCA, performing quality control, exploratory data analysis, identification of differentially abundant clusters, identification of martker genes associated with clusters, differential gene expression, and interpretation of results in the context of other data such as Gene Ontology and KEGG pathways. Groups of participants share their findings in a final presentation.
        - Destributed Communities Where short courses jump start participants'ability to perform basic tasks and learn the tools relevant to scRNA-seq, distributed communities provide the opportunity of mastery of these techniques, which includes the ability to teach others. Our support of these communities will include
        - Workshops and Seminars For Group Leaders
        - Access to Community-Developed Curricula and Documentation
        - Scholarships for Short Courses
    - Software hardening/testing (Casey - software eng)
    - Bioconductor integration (Stephanie, Mike)

References

1. **Missing data and technical variability in single-cell RNA-sequencing experiments**
Stephanie C Hicks, F William Townes, Mingxiang Teng, Rafael A Irizarry

Biostatistics (2017-11-06) https://doi.org/gfb8g4
DOI: 10.1093/biostatistics/kxx053 · PMID: 29121214

2. **quantro: a data-driven approach to guide the choice of an appropriate normalization method.**
Stephanie C Hicks, Rafael A Irizarry

Genome biology (2015-06-04) https://www.ncbi.nlm.nih.gov/pubmed/26040460
DOI: 10.1186/s13059-015-0679-0 · PMID: 26040460 · PMCID: PMC4495646

3. **Enter the Matrix: Factorization Uncovers Knowledge from Omics**
Genevieve L. Stein-O'Brien, Raman Arora, Aedin C. Culhane, Alexander V. Favorov, Lana X. Garmire, Casey S. Greene, Loyal A. Goff, Yifeng Li, Aloune Ngom, Michael F. Ochs, … Elana J. Fertig

Trends in Genetics (2018-10) https://doi.org/gd93tk
DOI: 10.1016/j.tig.2018.07.003 · PMID: 30143323

4. **Comprehensive analysis of retinal development at single cell resolution identifies NFI factors as essential for mitotic exit and specification of late-born cells**
Brian Clark, Genevieve Stein-O'Brien, Fion Shiau, Gabrielle Cannon, Emily Davis, Thomas Sherman, Fatemeh Rajaii, Rebecca James-Esposito, Richard Gronostajski, Elana Fertig, … Seth Blackshaw

Cold Spring Harbor Laboratory (2018-07-27) https://doi.org/gdwrzh
DOI: 10.1101/378950

5. **Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species.**
Genevieve L Stein-O'Brien, Brian S. Clark, Thomas Sherman, Christina Zibetti, Qiwen Hu, Rachel Sealfon, Sheng Liu, Jiang Qian, Carlo Colantuoni, Seth Blackshaw, … Elana J. Fertig

Cold Spring Harbor Laboratory (2018-08-20) https://doi.org/gd2xpn
DOI: 10.1101/395004

6. **Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma**
Elana J Fertig, Qing Ren, Haixia Cheng, Hiromitsu Hatakeyama, Adam P Dicker, Ulrich Rodeck, Michael Considine, Michael F Ochs, Christine H Chung

BMC Genomics (2012) https://doi.org/gb3fgp
DOI: 10.1186/1471-2164-13-160 · PMID: 22549044 · PMCID: PMC3460736

7. **CoGAPS matrix factorization algorithm identifies transcriptional changes in AP-2alpha target genes in feedback from therapeutic inhibition of the EGFR network**
Elana J. Fertig, Hiroyuki Ozawa, Manjusha Thakar, Jason D. Howard, Luciane T. Kagohara, Gabriel Krigsfeld, Ruchira S. Ranaweera, Robert M. Hughes, Jimena Perez, Siân Jones, … Christine H. Chung

Oncotarget (2016-09-16) https://doi.org/f9k8d8
DOI: 10.18632/oncotarget.12075 · PMID: 27650546 · PMCID: PMC5342018

8. **Pattern Identification in Time-Course Gene Expression Data with the CoGAPS Matrix Factorization**
Elana J. Fertig, Genevieve Stein-O'Brien, Andrew Jaffe, Carlo Colantuoni

Gene Function Analysis (2013-10-24) https://doi.org/f5j7xj
DOI: 10.1007/978-1-62703-721-1_6 · PMID: 24233779

9. **Integrated time course omics analysis distinguishes immediate therapeutic response from acquired resistance**
Genevieve Stein-O'Brien, Luciane T. Kagohara, Sijia Li, Manjusha Thakar, Ruchira Ranaweera, Hiroyuki Ozawa, Haixia Cheng, Michael Considine, Sandra Schmitz, Alexander V. Favorov, … Elana J. Fertig

Genome Medicine (2018-05-23) https://doi.org/gfc4dq
DOI: 10.1186/s13073-018-0545-2 · PMID: 29792227 · PMCID: PMC5966898

10. **Inferring causal molecular networks: empirical assessment through a community-based effort**
Steven M HillLaura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, … Sach Mukherjee

Nature Methods (2016-02-22) https://doi.org/f3t7t4
DOI: 10.1038/nmeth.3773 · PMID: 26901648 · PMCID: PMC4854847

11. **Preferential Activation of the Hedgehog Pathway by Epigenetic Modulations in HPV Negative HNSCC Identified with Meta-Pathway Analysis**
Elana J. Fertig, Ana Markovic, Ludmila V. Danilova, Daria A. Gaykalova, Leslie Cope, Christine H. Chung, Michael F. Ochs, Joseph A. Califano

PLoS ONE (2013-11-04) https://doi.org/gcpgc6
DOI: 10.1371/journal.pone.0078127 · PMID: 24223768 · PMCID: PMC3817178

12. **MultiPLIER: a transfer learning framework reveals systemic features of rare autoimmune disease**
Jaclyn N Taroni, Peter C Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A Merkel, Casey S Greene

Cold Spring Harbor Laboratory (2018-08-20) https://doi.org/gfc9bb
DOI: 10.1101/395947

13. **tximeta**
Rob Patro Michael Love

Bioconductor (2018) https://doi.org/gfddxw
DOI: 10.18129/b9.bioc.tximeta