# 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@6d5bb32 on October 11, 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

- **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. 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 methods 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 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. 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

### Loyal Goff

### Casey Greene (Submitter)

### Tom Hampton

### Mike Love

### Rob Patro

# Proposal Body (2000 words)

- Base enabling technologies:
  - Low dimensional representations (Elana)
  - New methods published by other groups (scVI, etc - other groups)
  - Fast & improved quantification (Rob)
  - Incorporation of uncertainty estimates in low dimensional representations (Rob / Mike / Elana) - note this can be done w/o modification in scCoGAPS

## 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

- Eschewing marker genes: Practical exploration of the HCA in latent spaces
  - Search tool for perturbations / signatures in latent-space(s) (Loyal)
    - Differential analysis of latent space usage across contexts
  - Latent space transformations for progression? Consider jawns for semi-supervised learning? (Elana)
  - Transfer learning of signatures *between/across* tissues (Loyal)
- 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

- Delivery
  - Training / teaching (scRNAseq, low-dimensional representations, RFA-developed tools) (Tom)
  - Software hardening/testing (Casey - software eng)
  - Bioconductor integration (Stephanie, Mike)

# References