### Summary

Certain deep neural networks can generate hypothetical data by learning and decoding a lower dimensional latent space. This latent space enables arithmetic operations that produce realistic output for novel transformations. This allows users to generate hypothetical images1 and to interpolate protein localizations through the cell-cycle2. An accessible example of latent space transformations comes from FaceApp3, which modifies a picture of an individual to produce an image of the subject at an older age, with a different expression, or of a different genders.

Our *overall objective* is to determine how unsupervised deep neural network models can best be trained on single cell expression data from the Human Cell Atlas (HCA) and the extent to which such models define biological latent spaces that capture disease states and targeted perturbations. The *rationale* is that latent space arithmetic for genomic data would enable researchers to predict how the expression of every gene would change in each HCA-identified cell type after drug treatment, in the context of a specific genetic variant, with a specific disease, or a combination of these and other factors.

### Aims

***Aim 1: Develop proof-of-concept unsupervised deep learning methods for single cell transcriptomic data from the HCA.***

***Aim 2: Generate a benchmark dataset of harmonized public data to evaluate the extent to which HCA cell types capture rheumatic disease biology.***

This proposal addresses two RFA points: Aim 1 develops machine learning approaches for solving the inference of state transitions and developmental trajectories, and Aim 2 provides curated benchmark datasets from existing data for evaluating computational methods and designing future assessments.

### Prior Contributions / Preliminary Results

We previously developed neural-network based methods for unsupervised integration of transcriptomic data4–6. We now build to Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) which have a track record of defining meaningful latent spaces for images. We adapted GANs to generate realistic individuals under a differential privacy framework7 and built VAEs over bulk transcriptomic data with the goal of describing a biologically-relevant latent space8. Here, we will apply these unsupervised deep learning methods to single cell transcriptomic data and incorporate novel data augmentation approaches for genomics. We also bring workflow automation experience to the HCA community9.

### Proposed work and deliverables

#### Aim 1: Develop proof-of-concept unsupervised deep learning methods for single cell transcriptomic data from the HCA.

The *objective of this aim* is to implement and test approaches to build deep generative models, such as VAEs10 and GANs11, from HCA single cell RNA-seq data.

Single cell data pose unique opportunities, but also challenges, for deep neural network algorithms. Many cells are often assayed, and many observations are needed to use deep learning effectively. However, transcript abundance estimates for each cell are generally subject to more error than bulk samples.

In our experience with generative deep learning7,8 it can be difficult to predict optimal parameters in advance. We will perform a grid search over VAE architectures and hyperparameters to identify suitable options. We will evaluate zero-inflated loss among more traditional loss functions, as Chris Probert noted potential benefits on our [proposal's GitHub repository](https://github.com/greenelab/czi-rfa/issues/11)12–14. This process will identify a subset of parameters and architectures that are worth exploring further for single cells.

We will also develop data augmentation for single cell RNA-seq data, as no such approaches exist yet for transcriptomes. To understand data augmentation, imagine scanned pathology slides. Each slide may be prepared and scanned with a subtly different orientation or magnification. A deep learning method may identify these measurement differences, or there may be too few slides to train a good model. Applying arbitrary rotations, zooms, and other irrelevant transformations increases the effective amount of training data and reduces the model's propensity to learn such noise.

We plan to use fast abundance estimates for RNA-seq15,16 to perform data augmentation for transcriptomes. Multiple resamples or subsamples of reads during transcript abundance estimation can capture uncertainty in the data, akin to arbitrary rotations. Therefore, we plan to collaborate with Rob Patro's laboratory (Collaborative Network) to implement these and related approaches. We posit that genomic data augmentation will improve latent feature generalization by separating biological from technical features and increasing the effective sample size during training.

We will select high-quality models by choosing those that minimize both reconstruction loss and KL divergence. We will evaluate resulting models for their applicability to rheumatic disease and their suitability for latent space arithmetic (see: Evaluation).

#### Aim 2: Generate a benchmark dataset of harmonized public data to evaluate the extent to which HCA cell types capture rheumatic disease biology.

The HCA's partnership with the Immunological Genome Project (immgenH) will provide single-cell gene expression-based immunocyte phenotyping at an unprecedented resolution. A compendium comprised of bulk gene expression data from autoimmune/rheumatic diseases is exceptionally well-suited to evaluating the disease relevance of these immunocyte data. The *objective of this aim* is to build and share real and simulated benchmark datasets to evaluate the quality of the cell-type signatures. This will allow CZI to evaluate techniques, including VAEs and other methods, for defining cell-type-specific expression signatures from the HCA's single-cell datasets by measuring their ability to decompose bulk, whole-tissue autoimmune/rheumatic disease data.

We will generate simulated bulk datasets drawn from HCA-identified cell types by combining their expression profiles at different proportions. We will also build a multi-tissue autoimmune/rheumatic disease compendium from existing public datasets that we have curated (currently more than 12,000 samples). This compendium includes samples from patients with systemic lupus erythematosus (SLE), sarcoidosis, and inflammatory bowel disorders among many other diseases. Such a compendium lets us determine the extent to which HCA-derived cell type signatures capture disease-relevant information in a way that matches previous literature. For instance, we expect to detect higher proportions of activated macrophages in lupus nephritis samples than controls17.

These bulk compendia (simulated and real data) will enable HCA participants (computational-method and molecular-assay developers) to directly compare approaches where we expect their most immediate translational impact: application to existing datasets to explain disease-relevant phenomena via a single-cell perspective.

### Proposal for evaluation and dissemination.

We will apply methods that produce low-dimensional representations including VAEs (Aim 1) and other methods to HCA-produced single cell transcriptomes. Source code that generates low-dimensional models will be released via GitHub, and we may produce a manuscript on the topic. Models and datasets will be disseminated via periodic release on Zenodo or a similar platform. We will test these low-dimensional representations via latent space arithmetic and relevance to disease as described below.

#### Evaluate the extent to which low-dimensional representations enable latent space arithmetic in the HCA.

Certain classes of generative deep neural network models, including VAEs and GANs have been shown to imbue intuitive mathematical features to the learned latent features10,11. For instance, a GAN learned latent features that could be manipulated with arithmetic: subtracting out the vector of a smile from a woman and adding it to a man with a neutral face resulted in an image of a smiling man18. We will evaluate the extent to which these properties exist for low-dimensional representations of the HCA's single-cell transcriptomes. We describe two experiments using data proposed by Arjun Raj's group (Collaborative Network), but any HCA benchmark datasets with similar properties will be suitable.

The Raj lab proposes to assay cardiomyocte differentiation from fibroblasts. The driving transcription factors for this process have been identified19. A latent space vector between these two cell types should capture the key transcription factor (TF) networks (Gata4, Mef2c, and Tbx5). To calculate this vector we will subtract the latent space projections of fibroblasts from cardiomyocytes. We will compare the gene composition of this differentiation vector to TF-target calls from cistrome20, which are available for each of these TFs.

Latent space arithmetic can also generate new hypothetical data. We will test the extent to which these models predict the results of perturbations using data that Arjun Raj's homogenized cell type data. For each perturbation, we will hold out one or more cell types and map the rest into the latent space. Subtracting the latent space vector of included cell types from those after perturbation will produce a perturbation vector. We will add the perturbation vector to a withheld cell type to generate synthetic data and compare the synthetic and observed results to determine the prediction error. Comparing low-dimensional methods to a baseline of analogous transformations on raw gene expression can reveal whether or not these approaches more accurately predict perturbations.

#### Rheumatic Disease Evaluation

We will input signatures from low-dimensional projections into existing techniques that decompose bulk data with cell type signatures21,22 and evaluate concordance with ground truth on Aim 2's simulated dataset. Comparing performance with multiple decomposition techniques allows us to benchmark methods' abilities to define bulk-relevant signatures from HCA data. We will also use signatures to decompose the rheumatic disease compendium. We can easily ask which methods produce cell-type signatures that explain the most variance in the compendium. But we can also use experiments within the compendium, such as studies of highly-targeted therapeutics (e.g., a monoclonal antibody to IFN-gamma in the context of systemic lupus erythematosus23, to develop additional data-driven hypotheses. In the case if the IFN-gamma antibody, we can use various methods to predict which cell-types change in proportion or pathway activation during the reduction of this cytokine. These analyses will allow us to identify disease-relevant cases where methods disagree, laying the groundwork for targeted experiments (beyond the one-year timeline) that directly probe these processes to produce informative ground-truth benchmarks.

### Statement of commitment to share

We commit to sharing proposals, methods, data, and code publicly under open licenses. We understand that our proposal will be shared if it is funded: we [shared it publicly under a CC-BY license](https://github.com/greenelab/czi-rfa) as it was written.

## References

1. Gregor K, Danihelka I, Graves A, Rezende DJ, Wierstra D. DRAW: A Recurrent Neural Network For Image Generation. arXiv 2015.

2. Osokin A, Chessel A, Salas REC, Vaggi F. GANs for Biological Image Synthesis. arXiv 2017.

3. FaceApp - Free Neural Face Transformation Filters [Internet]. http://www.faceapp.com

4. Tan J, Hammond JH, Hogan DA, Greene CS. ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions. Gilbert JA, editor. mSystems **1**(1), e00025-15, 2016.

5. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, Perchuk B, Laub MT, Hogan DA, Greene CS. Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks. Cell Syst 2017.

6. Tan J, Huyck M, Hu D, Zelaya RA, Hogan DA, Greene CS. ADAGE signature analysis: differential expression analysis with data-defined gene sets. bioRxiv 2017.

7. Beaulieu-Jones BK, Wu ZS, Williams C, Greene CS. Privacy-preserving generative deep neural networks support clinical data sharing. bioRxiv 2017.

8. Way GP, Greene CS. Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes with Variational Autoencoders. bioRxiv 2017.

9. Beaulieu-Jones BK, Greene CS. Reproducibility of computational workflows is automated using continuous analysis. Nat Biotechnol **35**(4), 342–346, 2017.

10. Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv 2013.

11. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative Adversarial Networks. arXiv 2014.

12. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol **16**(1), 241, 2015.

13. McDavid A, Gottardo R, Simon N, Drton M. Graphical Models for Zero-Inflated Single Cell Gene Expression. arXiv 2016.

14. Lin P, Troup M, Ho JWK. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol **18**(1), 59, 2017.

15. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol **34**(5), 525–527, 2016.

16. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods **14**(4), 417–419, 2017.

17. Berthier CC, Bethunaickan R, Gonzalez-Rivera T, Nair V, Ramanujam M, Zhang W, Bottinger EP, Segerer S, Lindenmeyer M, Cohen CD, Davidson A, Kretzler M. Cross-species transcriptional network analysis defines shared inflammatory responses in murine and human lupus nephritis. J Immunol **189**(2), 988–1001, 2012.

18. Radford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

19. Ieda M, Fu J-D, Delgado-Olguin P, Vedantham V, Hayashi Y, Bruneau BG, Srivastava D, Kamp TJ, Reinecke H, Murry CE, al. et. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. Cell **142**(3), 375–86, 2010.

20. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, Zhu M, Wu J, Shi X, Taing L, Liu T, Brown M, Meyer CA, Liu XS. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res **45**(D1), D658–D662, 2017.

21. Chikina M, Zaslavsky E, Sealfon SC. CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. Bioinformatics **31**(10), 1584–1591, 2015.

22. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, Selves J, Laurent-Puig P, Sautès-Fridman C, Fridman WH, de Reyniès A. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol **17**(1), 218, 2016.

23. Welcher AA, Boedigheimer M, Kivitz AJ, Amoura Z, Buyon J, Rudinskaya A, Latinis K, Chiu K, Oliner KS, Damore MA, Arnold GE, Sohn W, Chirmule N, Goyal L, Banfield C, Chung JB. Blockade of Interferon-γ Normalizes Interferon-Regulated Gene Expression and Serum CXCL10 Levels in Patients With Systemic Lupus Erythematosus. Arthritis Rheumatol **67**(10), 2713–2722, 2015.