Summary Population-level gene expression combines “transcriptional identity” of cell types and dynamic, biological processes in single cells[1]. Inter-individual genetic variation and environmental / pharmacological perturbations induce further changes. Learning developmental trajectories and timing of state transitions requires analytical techniques to delineate cellular composition from dynamic biological processes in scRNA-seq data.

Genomics data can be represented as a matrix with genes in rows and samples for cells and conditions in columns. Each biological process and transcriptional identity is a basis of that matrix. Non-negative matrix factorization (NMF) methods can learn these bases directly from the input data. Specifically, this class of algorithms decomposes data into a continuously-valued vector associating the relative amount of a gene’s expression in a given biological process with a corresponding continuously-valued vector indicating the activity of that process in each sample (**Fig 1**). CoGAPS uniquely inputs expectation and variance measures of expression, and models sparsity and non-negativity in its atomic prior[2]. It can model both dynamic trajectories[3-6] and transcriptional identity[7,8] in bulk transcriptional data. CoGAPS can also encode disparate error models to integrate data from distinct measurement technologies[6,9]. This enables for input of enhanced quantification of scRNA-seq (e.g., Patro’s methods[10]). Preliminary data in single-cell RNA-sequencing has suggested that CoGAPS effectively distinguish developmental trajectories in different cell types in the retina with such scRNA-seq (**Fig 2**). Together, these results suggest that CoGAPS integration of bulk and single-cell developmental datasets in the HCA will learn patterns that distinguish gene interactions from individual variation from gene interactions along developmental trajectories.

The central aims of this proposal address challenges pervasive to unsupervised algorithms: (1) adapting them to the large sample size of dynamic single cell data (**Aim 1**) and (2) modeling distinct distributions in multimodal data (**Aim 2**). Together, these aims are essential to realize the project goal of “to solve multimodal integration [and] inference of state transitions” in the comprehensive, dynamic HCA datasets. Although optimized for CoGAPS, the parallel processing methods and models of sparsity will be generally applicable to other techniques developed in the consortium. They will also be inter-dependent on preprocessing, signature interpretation, data generation, and visualization efforts of the both proposed collaborative network and broader consortium.

Project Aims, and how they address program goals:

Aim 1 Parallel CoGAPS to learn state transitions and developmental trajectories from large, time-course omics data. Similar to most NMF algorithms, CoGAPS is limited to O(1000) genes for guaranteed convergence. Some apply feature compaction prior to analysis[11,12]. However, associating genes or transcripts with compacted features may be challenging. CoGAPS can be performed directly on genome-wide data using parallel analysis across random gene sets[8]. The considerable redundancy between co-regulated genes enables both compaction and parallelization. Single cell omics data introduces large sample sizes with similar convergence issues. Independence of expression in specific cell types or stages may limit sample-wide compaction or parallelization. To ensure both solution optimization and computational efficiency, we will develop a message passing system to parallelize CoGAPS across gene and sample sets. We will apply this algorithm to randomly selected subsets of time-course genomics data benchmark data to assess the sensitivity of the resulting trajectory inference to distributions of cell types, states, and dynamic stages selected for parallel analyses. The resulting gene associations will interpreted in context of unsupervised gene interaction inference (e.g., Greene and Krishnaswamy’s methods with Goff’s comparison). **Program goal:**

The P-GAPS algorithm will enable efficient inference of state transitions and developmental trajectories from HCA time-course omics data.

Aim 2 Modeling the impact of sparsity on technical variation between bulk and multi-platform single cell RNA-sequencing. The atomic prior in CoGAPS provides a dynamic sparsity constraint. In this model, lowly expressed genes being constrained towards zero and highly expressed genes constrained away from zero, with corresponding constraints to limit the number of patterns to which each gene is associated[2]. We will modify the hyperparameters in the prior distribution for CoGAPS to model different levels of sparsity between bulk and single cell-RNA sequencing data. We will apply the modified algorithm to time-course bulk and RNA-seq data from samples from similar developmental phases. We will model the resulting common biological patterns across sequencing platforms as a function of the sparsity hyperparameter. **Program goal:** Learning an adaptive, sparsity parameter across bulk and single-cell RNA-seq data will enable inference of state transitions and developmental trajectories from multimodal, time-course omics data.

Prior contributions in the area and preliminary results:

We have numerous analyses of multi-modal genomics data from distinct procurement techniques leading to methods for integrating both technical and biological data types[6,9,13,14]. We have demonstrated that CoGAPS infers trajectories associated with the dynamics of therapeutic response and acquired therapeutic resistance[4,6,7].

Recently, we have adapted CoGAPS to simultaneously delineate between cell cycle dynamics, technical artifacts, and cell fate trajectories from scRNA-seq data of retinal development from collaborative network member Goff. Like other CNS systems, all the different cell types in the retina arise from a common precursor pool in a stereotyped birth order. Determining the key factors that regulate the selection of an individual retinal progenitor cell (RPC) to exit the cell cycle and differentiate is a critical question in neural development. CoGAPS's can parse these concurrent signals, determining developmental trajectories and fate decisions (**Fig 2**-**4**). The algorithm determines additional patterns associated with technical artifacts (**Fig 5**) demonstrating its use to model pervasive technical and batch variation in scRNA-seq data.

Proposed work and deliverables:

Aim 1 Parallel CoGAPS to learn state transitions and developmental trajectories from large, time-course omics data.

Sensitivity to sample sets Using the 1,000 genes most variable genes, we will apply the parallel approach of GWCoGAPS to samples instead of genes. This approach will be applied to large, time-course HCA benchmark datasets (e.g., Goff's retinal development data). We will compute the similarity of gene weights among parallel runs as a function of the extent of confounding between cell type and cell states in each sample set. Assessing the robustness of such pattern detection by sample composition will also implicate the impact of batch effects on time course data. It will also provide robust quantification of pattern robustness, to assess optimal grouping for sample-level compaction.

Parallelization We will divide the HCA benchmark datasets into groups of random, but overlapping sets of genes and samples. CoGAPS will be run in parallel for each set. During the MCMC iterations in CoGAPS, message passing between the parallel chains will be employed to determine the current state of the factorization. Approximate Bayesian computation across all of

the chains will determine the consensus patterns and gene weights across all random sets of genes and samples, and chains will be continued from the consensus solutions.

Pitfalls and proposed solutions. CoGAPS analysis for sample sets may have imbalanced convergence based upon the distribution of cell types, states, and times in each set. In this case, timing will also be assessed in the sensitivity analysis and incorporated in sample selection.

Aim 2 Modeling the impact of sparsity on technical variation between bulk and multi-platform single cell RNA-sequencing.

Sparsity parameter algorithm development. Currently, CoGAPS is a unique factorization algorithm in modeling both the variance of the data and sparsity. Previously, this algorithm has been tuned for microarray data[15] and RNA-sequencing[8] data. To tune the algorithm for single-cell RNA-sequencing, we will modify the existing algorithm to utilize different sparsity hyperparameters for different samples within a dataset that contains mixed bulk and single-cell RNA-sequencing. This same hyperparameter will facilitate application of matrix factorization to future work beyond this proposal comparing analyses of transcript and gene level quantification.

Statistical analysis for systematic comparison of the impact of sparsity on data from distinct sequencing technologies. We will apply CoGAPS with mixed sparsity hyperparameters to the combined bulk and single cell RNA-sequencing data sets from matched samples in the HCA. We run CoGAPS for a range of sparsity hyperparameters for each of the scRNA-seq technology separately. Robustness will be estimated by comparing the number of shared patterns across data platforms as a function of these hyperparameters.

Pitfalls and proposed solutions. (1) If missing data from single cell RNA-sequencing results in poor convergence, we will apply denoising techniques such as those from collaborative network member Krishnaswamy[16]. (2) If variance stabilization is an ill-fitting error model, we will adapt the MCMC framework of CoGAPS to include a negative binomial error model.

Deliverables: (1) R/C++ code for P-GAPS encoded in the R/Bioconductor package CoGAPS. (2) R/C++ code to modify the sparsity hyper-parameter by sample encoded in the R/Bioconductor package CoGAPS. (3) R scripts to reproduce all analyses in the project aims. (4) Manuscript(s) on algorithm and results.

Proposal for evaluation and dissemination of methods, resources, or results

Testing of methods CoGAPS has been tested by comparing the inferred patterns to known phenotypes in multiple publications[3,5-9,17]. Similar assessment will be performed in this proposal, notably by comparing pattern robustness across random sets as described above. Results will be compared with trajectories, patterns, and gene signatures inferred with methods (e.g., Krishnaswamy[18] and Greene[19]'s methods) using model comparison methods (e.g., Goff's).

Engineering support Engineering support from CZI in message passing code and access to sufficient cloud-based computing resources would ensure efficiency of the P-GAPS algorithm.

Statement of commitment to share proposals, methods, data, and code with other researchers funded by this RFA and with CZI

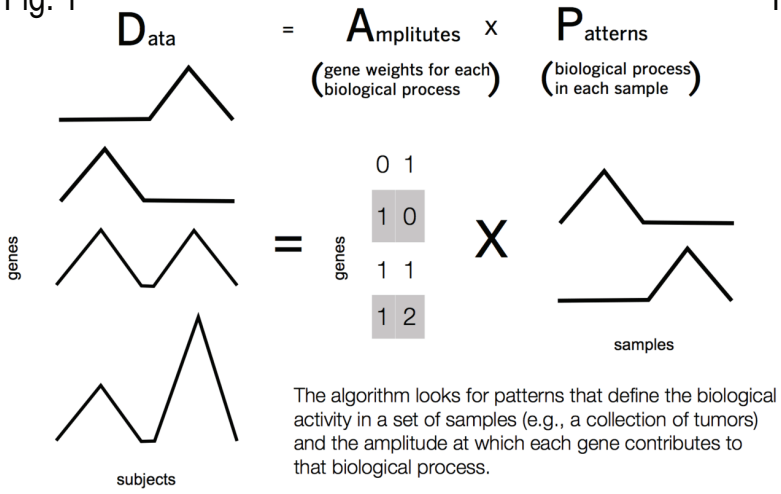
Primary data: All scripts will be openly shared consistent with the PI's standard practice (e.g., https://sourceforge.net/projects/psva/, https://github.com/FertigLab/EGFRFeedback).

Proposal: The proposal has already been developed in collaboration with the collaborative network and shared publicly on https://github.com/FertigLab/HCA.

Methods: Methods will be published, and posted on Bioarxiv during journal submission.

Software: Any algorithms will be released in the CoGAPS Bioconductor package[15].

Fig. 1



The algorithm looks for patterns that define the biological activity in a set of samples (e.g., a collection of tumors) and the amplitude at which each gene contributes to that biological process.

Fig. 3

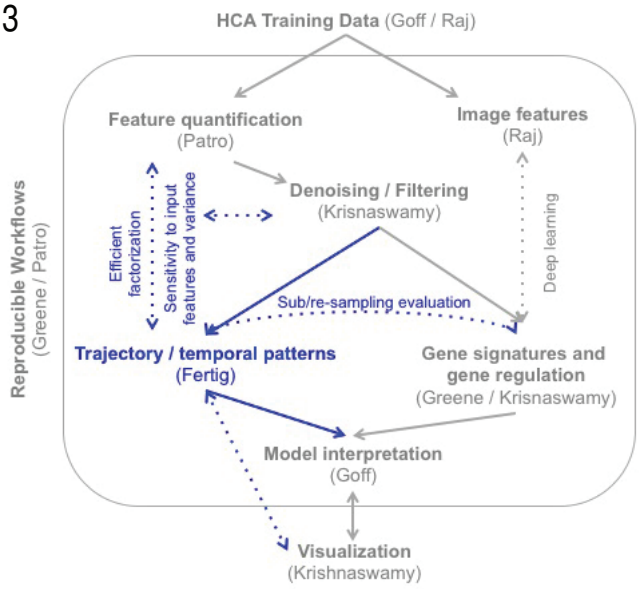
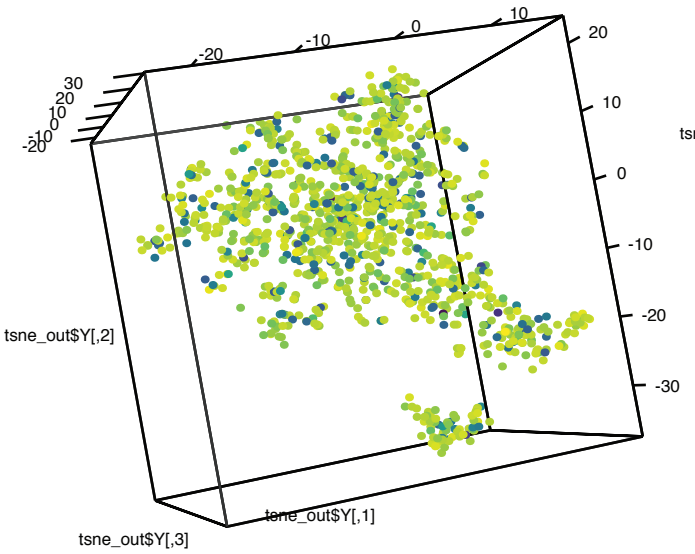
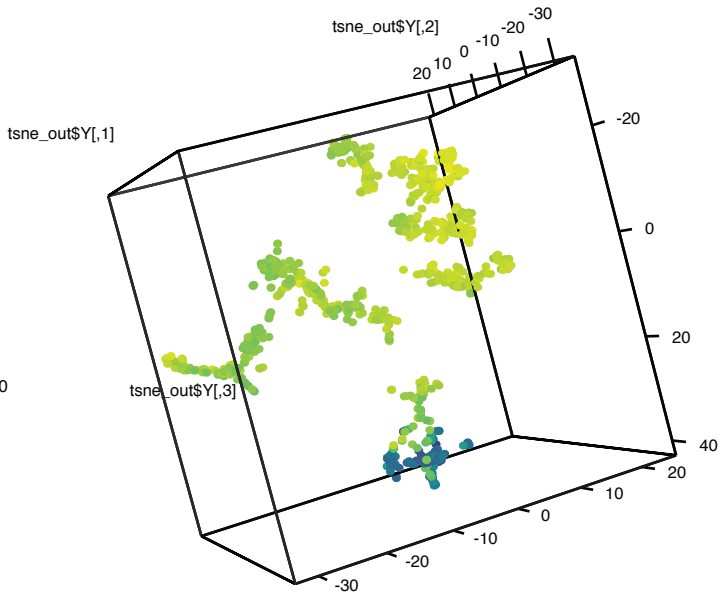


Fig. 2

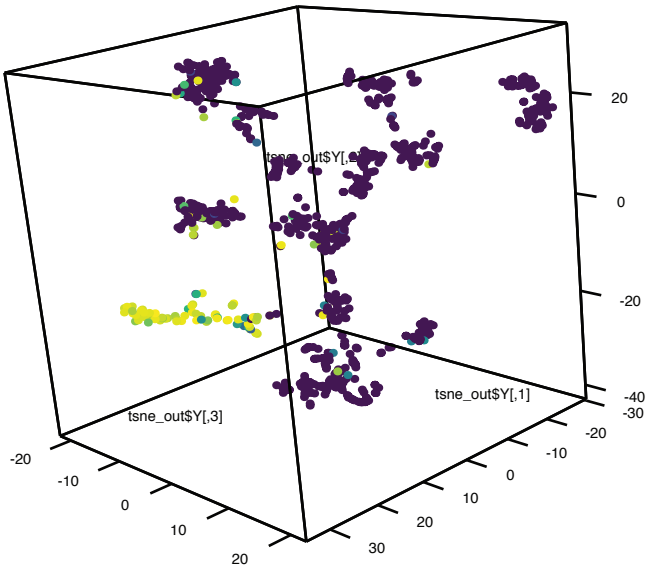
A. T-SNE of raw scRNAseq data colored by # genes expressed



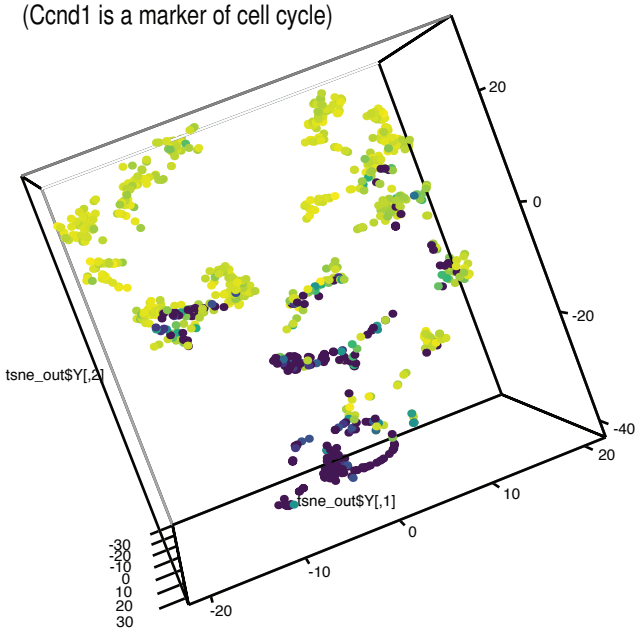
B. T-SNE of CoGAPS patterns for scRNAseq colored by # genes expressed



C. T-SNE of CoGAPS patterns colored by Crx expression
(Crx is a marker of photoreceptors)



D. T-SNE of CoGAPS patterns colored by Ccnd1 expression
(Ccnd1 is a marker of cell cycle)



1) Schematic representation of non-negative matrix factorization implemented in the CoGAPS algorithm. 2) T-SNE dimension reduction of scRNA (A) and CoGAPS pattern weights for smart seq data (B-D) of progenitor enriched cells from days E14, E18, and P2 in mouse retina developement. Pannels (A and B) are colored by number of genes expressed per cell a surrogate for sequence quality and other technical artifacts. CoGAPS patterns (C and D) are colored by marker gene expression to illustrate CoGAPS ability to find and refine patterns of specific lineages (C) and shared dynamic processes(D). 3) Proposed collaborative network with individual contributions highlighted in blue.

References

1 Wagner A, Regev A & Yosef N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* **34**, 1145-1160, (2016). PMC5465644.

2 Sibisi S & Skilling J. Prior distributions on measure space. *Journal of the Royal Statistical Society, B* **59**, 217-235, (1997). PMID PMCID not available.

3 Fertig EJ, Stein-O'Brien G, Jaffe A & Colantuoni C. Pattern identification in time-course gene expression data with the CoGAPS matrix factorization. *Methods Mol Biol* **1101**, 87-112, (2014). PMID 24233779, PMCID not available.

4 Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, Zhang Y, Sokolov A, Paull EO, Wong CK, Graim K, Bivol A, Wang H, Zhu F, Afsari B, Danilova LV, Favorov AV, Lee WS, Taylor D, Hu CW, Long BL, Noren DP, Bisberg AJ, Consortium H-D, Mills GB, Gray JW, Kellen M, Norman T, Friend S, Qutub AA, Fertig EJ, Guan Y, Song M, Stuart JM, Spellman PT, Koeppl H, Stolovitzky G, Saez-Rodriguez J & Mukherjee S. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat Methods* **13**, 310-318, (2016). PMC4854847.

5 Ochs MF & Fertig EJ. Matrix Factorization for Transcriptional Regulatory Network Inference. *IEEE Symp Comput Intell Bioinforma Comput Biol Proc* **2012**, 387-396, (2012). PMC4212829.

6 Stein-O'Brien G, Kagohara LT, Li S, Thakar M, Ranaweera R, Ozawa H, Cheng H, Considine M, Favorov A, Danilova L, Califano JA, Izumchenko E, Gaykalova DA, Chung CH & Fertig EJ. *Integrated time-course omics analysis distinguishes immediate therapeutic response from acquired resistance* (Bioarxiv, 2017).

7 Fertig EJ, Ren Q, Cheng H, Hatakeyama H, Dicker AP, Rodeck U, Considine M, Ochs MF & Chung CH. Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma. *BMC Genomics* **13**, 160, (2012). PMC3460736.

8 Stein-O'Brien GL, Carey JL, Lee WS, Considine M, Favorov AV, Flam E, Guo T, Li S, Marchionni L, Sherman T, Sivy S, Gaykalova DA, McKay RD, Ochs MF, Colantuoni C & Fertig EJ. PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF. *Bioinformatics*, (2017). PMID 28174896, PMCID not available.

9 Fertig EJ, Markovic A, Danilova LV, Gaykalova DA, Cope L, Chung CH, Ochs MF & Califano JA. Preferential activation of the hedgehog pathway by epigenetic modulations in HPV negative HNSCC identified with meta-pathway analysis. *PLoS One* **8**, e78127, (2013). PMC3817178.

10 Zakeri M, Srivastava A, Almodaresi F & Patro R. Improved data-driven likelihood factorizations for transcript abundance estimation. *Bioinformatics* **33**, i142-i151, (2017). PMID PMCID not available.

11 de Campos CP, Rancoita PM, Kwee I, Zucca E, Zaffalon M & Bertoni F. Discovering subgroups of patients from DNA copy number data using NMF on compacted matrices. *PLoS One* **8**, e79720, (2013). PMC3835832.

12 Tepper M & Sapiro G. Compressed Nonnegative Matrix Factorization Is Fast and Accurate. *IEEE Transactions on Signal Processing* **64**, 2269-2283, (2016). PMID PMCID not available.

13 Gaykalova DA, Zizkova V, Guo T, Tiscareno I, Wei Y, Vatapalli R, Hennessey PT, Ahn J, Danilova L, Khan Z, Bishop JA, Gutkind JS, Koch WM, Westra WH, Fertig EJ, Ochs MF & Califano JA. Integrative computational analysis of transcriptional and epigenetic alterations implicates DTX1 as a putative tumor suppressor gene in HNSCC. *Oncotarget* **8**, 15349-15363, (2017). PMID 28146432, PMCID not available.

14 Parker HS, Leek JT, Favorov AV, Considine M, Xia X, Chavan S, Chung CH & Fertig EJ. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. *Bioinformatics* **30**, 2757-2763, (2014). PMC4173013.

15 Fertig EJ, Ding J, Favorov AV, Parmigiani G & Ochs MF. CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics* **26**, 2792-2793, (2010). PMC3025742.

16 van Dijk D, Nainys J, Sharma R, Kathail P, Carr AJ, Moon KR, Mazutis L, Wolf G, Krishnaswamy S & Pe'er D. MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv*, (2017). PMID PMCID not available.

17 Fertig EJ, Ozawa H, Thakar M, Howard JD, Kagohara LT, Krigsfeld G, Ranaweera RS, Hughes RM, Perez J, Jones S, Favorov AV, Carey J, Stein-O'Brien G, Gaykalova DA, Ochs MF & Chung CH. CoGAPS matrix factorization algorithm identifies transcriptional changes in AP-2alpha target genes in feedback from therapeutic inhibition of the EGFR network. *Oncotarget* **7**, 73845-73864, (2016). PMC5342018.

18 Moon KR, Dijk Dv, Wang Z, Chen W, Hirn MJ, Coifman RR, Ivanova NB, Wolf G & Krishnaswamy S. *PHATE: A Dimensionality Reduction Method for Visualizing Trajectory Structures in High-Dimensional Biological Data* (2017).

19 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* **5**, 63-71 e66, (2017). PMC5532071.