**1,600 words max**

**Proposed title:** Bayesian matrix factorization for multimodal integration of time-course omics data

**Summary** Single-cell analysis of gene expression has demonstrated that population-level gene expression and the ‘transcriptional identity’ of individual cells arises from the combinations dependent and independent contributions of discrete biological processes (Wagner, Regev, Yosef, Nat Biotchnology, 2016). The combinations of these processes change dynamically as organisms develop. Inter-individual genetic variation and pharmacological perturbations from treatment of complex diseases induce further changes into these processes. Learning developmental trajectories and timing of state transitions requires analytical techniques to delineate cellular composition from dynamic biological processes in single cell, transcriptional data.

When representing genomics data as a data matrix, the discrete biological processes that distinguish dynamic biological processes from transcriptional identity are each represented as basis vectors in that matrix. Decomposition methods can learn these bases directly from the input data matrix. 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**). Our sparse, Bayesian non-negative matrix factorization algorithm CoGAPS (Coordinated Gene Activity in Pattern Sets)1 accounts for both gene reuse and biological parsimony by encoding sparsity and non-negativity in its atomic prior (SKILLING CITATION). CoGAPS has been shown to learn quantitative trajectories associated with pathway perturbations (HPN-DREAM), therapeutic resistance (TIME COURSE), and ageing (BOOK CHAPTER) from bulk transcriptional data. It can also distinguish individual and tissue-specific differences (GWCoGAPS) from static, bulk transcriptional data. Preliminary data in single-cell RNA-sequencing has suggested that CoGAPS can likewise distinguish developmental trajectories in different cell types in the retina from single-cell data (**Fig 2**). Because the algorithm is Bayesian, it can also encode disparate error models to learn these trajectories across data from distinct measurement technologies (HEDGEHOG). Together, these results suggest that CoGAPS integration of bulk and single-cell developmental datasets in the Human Cell Atlas will learn patterns in the data that distinguish gene interactions associated with individual variation from gene interactions along developmental trajectories.

The central aim of this research proposal is to address challenges pervasive to matrix factorization to transcript-level address the large sample size of dynamic single cell data (**Aim 1**)and model distinct distributions when decomposing multimodal data (**Aim 2**). These approaches build upon gene-level parallelization of CoGAPS in GWCoGAPS (CITE) and modeling distributions of the sparsity hyperparameter in the CoGAPS atomic prior (CITE), respectively. Together, these aims are essential to realize the project goal of “analytical methods … to solve multimodal integration [and] inference of state transitions” in the comprehensive, dynamic datasets of the Human Cell Atlas. Although optimized for CoGAPS, the parallel processing methods and Bayesian models of sparsity will be generally applicable to other unsupervised approaches for genomics data being developed as part of the HCA 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 Gene Activity in Pattern Sets (P-GAPS) 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 TECHNIQUES ADDRESS THIS WITH FEATURE COMPACTION PRIOR TO ANALYSIS THAT MAY LIMIT IMPLICATION OF APPROPRIATE GENES AND/OR SAMPLES THAT ARE MOST RELEVANT. Recently, we showed that CoGAPS can be performed genome-wide using parallel analysis across random gene sets. BOTH THESE COMPACTION AND PARELLELIZATION APPROACHES ARE PERFORMED ON GENES ONLY, AND ALGORITHMS WILL FACE SIMILAR CONVERGENCE ISSUES FOR LARGE SAMPLE SIZES. To ensure both solution optimization and computational efficiency, we will develop a message passing system to extend the GWCoGAPS framework to parallelize pattern detection across large sample sets. We will apply this algorithm to RANDMOLY SELECTED SUBSETS OF HCA benchmark data TO ASSESS THE SENSITIVITY OF THE RESULTING TRAJECTORY INFERENCE TO REPRESENTATION OF CELL TYPES, STATES, AND DYNAMIC STAGES. **Program goal:** The P-GAPS patterns learned from HCA TIME-COURSE DATA WILL BENCHMARK ABILITY inference of state transitions and developmental trajectories WITH PARALLEL PROCESSING OF UNSUPERVISED LEARNING TO ENHANCE EFFICIENCY.

**Aim 2** **Modeling the impact of sparsity on technical variation between bulk and multi-platform single cell RNA-sequencing.** 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:** MULTIMODAL INTEGRATION, INFERENCE OF STATE TRANSITIONS AND DEVELOPMENTAL TRAJECTORIES

**Prior contributions in the area and preliminary results:**

We have performed numerous genomics analyses of data from complex experimental designs in cancer2-22. Our work on cross-study analyses of tumor samples and biopsies with various procurements lead to methods for integrating both technical and biological data types (CITATIONS). We have demonstrated that CoGAPS infers trajectories associated with the dynamics of therapeutic response and acquired therapeutic resistance (CITE HPN-DREAM8, TIME COURSE PAPER). Recently, we have adapted CoGAPS to distinguish the dynamics of cell cycle from state transitions in cell fate decisions from single cell RNA-seq data during retinal development. DESCRIBE THESE RESULTS (Fig 1)!

**Proposed work and deliverables:**

**Aim 1 Parallel Gene Activity in Pattern Sets (P-GAPS) to learn state transitions and developmental trajectories from large, time-course omics data.**

**P-GAPS algorithm development.** With GWCoGAPS, we demonstrated that patterns from CoGAPS can be estimated genome wide with a parallel approach across random sets of genes23,24. This algorithm converges for datasets from standard experimental designs capped at O(1000) samples, but will face similar challenges when for the number of samples planned in the Human Cell Atlas. TIME COURSE DATA POSES ADDITIONAL CHALLENGES, REQUIRING REPRESENTATION FROM DIFFERENT STATES / DYNAMICS IN ORDER TO MODEL DEVELOPMENTAL TRAJECTORIES. We will address this challenge by parallelizing CoGAPS across both samples and genes, while assessing algorithm performance as follows.

1. Using the 1,000 genes with largest ratio of inter-sample variability relative to inter-batch as described previously19 in GTeX and TCGA samples, we will apply the parallel approach of GWCoGAPS to samples instead of genes. We will evaluate the similarity of gene weights among parallel runs as a function of the extent of confounding between batch and experimental conditions in each sample set. NOTE ASSESSING THIS ROBUSTNESS HAS BROADER IMPACTS FOR OTHER PATTERN DETECTION METHODS BEYOND COGAPS AND SIMILARITY IMPACTS FOR WHAT CAN BE COMPACTED IN THOSE METHODS.
2. 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.** (1) If P-GAPS fails to converge because of the large sample size, we will implement compaction approaches across samples25,26. (2) Because batches and tissue types are perfectly confounded in TCGA, they may dominate the signal learned from P-GAPS. In this case, we will limit analyses to common tissue types in both dataset to assess biological reproducibility from age, gender, race, and batch.

**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 contains a sparsity parameter tuned for microarray data1 and RNA-sequencing24 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. Although not the focus of this proposal, we note that this same hyperparameter will facilitate matrix factorization of both gene-level and transcript-level data to infer isoform-specific patterns in future work in collaboration with consortia members.

**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 **Table 2**. We run CoGAPS for a range of sparsity hyperparameters for each of the Smart-Seq2 and 10X sequencing 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 preprocessing diffusion methods such as MAGIC27. (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) Raw and processed data from **Table 2**. (4) Manuscript(s) on algorithm and results.

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

**how will benchmark datasets be shared; what testing of computational methods has already been conducted and what new tests are proposed; what engineering support from CZI would advance dissemination of this resource or method**

**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 as part of the open-source CoGAPS Bioconductor package1. Bioconductor is a centralized, peer-reviewed database and release there will support future application and development by the broader community.

**References**

1 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.

2 Afsari B, Geman D & Fertig EJ. Learning dysregulated pathways in cancers from differential variability analysis. *Cancer Inform* **13**, 61-67, (2014). PMC4218688.

3 Cheng H, Fertig EJ, Ozawa H, Hatakeyama H, Howard JD, Perez J, Considine M, Thakar M, Ranaweera R, Krigsfeld G & Chung CH. Decreased SMAD4 expression is associated with induction of epithelial-to-mesenchymal transition and cetuximab resistance in head and neck squamous cell carcinoma. *Cancer Biol Ther* **16**, 1252-1258, (2015). PMC4623002.

4 Fertig EJ, Lee E, Pandey NB & Popel AS. Analysis of gene expression of secreted factors associated with breast cancer metastases in breast cancer subtypes. *Sci Rep* **5**, 12133, (2015). PMC4648401.

5 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.

6 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.

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 Gajula RP, Chettiar ST, Williams RD, Nugent K, Kato Y, Wang H, Malek R, Taparra K, Cades J, Annadanam A, Yoon AR, Fertig E, Firulli BA, Mazzacurati L, Burns TF, Firulli AB, An SS & Tran PT. Structure-function studies of the bHLH phosphorylation domain of TWIST1 in prostate cancer cells. *Neoplasia* **17**, 16-31, (2015). PMC4309734.

9 Gajula RP, Chettiar ST, Williams RD, Thiyagarajan S, Kato Y, Aziz K, Wang R, Gandhi N, Wild AT, Vesuna F, Ma J, Salih T, Cades J, Fertig E, Biswal S, Burns TF, Chung CH, Rudin CM, Herman JM, Hales RK, Raman V, An SS & Tran PT. The twist box domain is required for Twist1-induced prostate cancer metastasis. *Mol Cancer Res* **11**, 1387-1400, (2013). PMC3833995.

10 Gaykalova DA, Manola JB, Ozawa H, Zizkova V, Morton K, Bishop JA, Sharma R, Zhang C, Michailidi C, Considine M, Tan M, Fertig EJ, Hennessey PT, Ahn J, Koch WM, Westra WH, Khan Z, Chung CH, Ochs MF & Califano JA. NF-kappaB and stat3 transcription factor signatures differentiate HPV-positive and HPV-negative head and neck squamous cell carcinoma. *Int J Cancer* **137**, 1879-1889, (2015). PMC4629062.

11 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.

12 Guerrero-Preston R, Michailidi C, Marchionni L, Pickering CR, Frederick MJ, Myers JN, Yegnasubramanian S, Hadar T, Noordhuis MG, Zizkova V, Fertig E, Agrawal N, Westra W, Koch W, Califano J, Velculescu VE & Sidransky D. Key tumor suppressor genes inactivated by "greater promoter" methylation and somatic mutations in head and neck cancer. *Epigenetics* **9**, 1031-1046, (2014). PMC4143405.

13 Guo T, Gaykalova DA, Considine M, Wheelan S, Pallavajjala A, Bishop JA, Westra WH, Ideker T, Koch WM, Khan Z, Fertig EJ & Califano JA. Characterization of functionally active gene fusions in human papillomavirus related oropharyngeal squamous cell carcinoma. *Int J Cancer* **139**, 373-382, (2016). PMID 26949921, PMCID not available.

14 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.

15 Izumchenko E, Chang X, Brait M, Fertig E, Kagohara LT, Bedi A, Marchionni L, Agrawal N, Ravi R, Jones S, Hoque MO, Westra WH & Sidransky D. Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA. *Nat Commun* **6**, 8258, (2015). PMC4595648.

16 Lee E, Fertig EJ, Jin K, Sukumar S, Pandey NB & Popel AS. Breast cancer cells condition lymphatic endothelial cells within pre-metastatic niches to promote metastasis. *Nat Commun* **5**, 4715, (2014). PMC4351998.

17 Li R, Ochs MF, Ahn SM, Hennessey P, Tan M, Soudry E, Gaykalova DA, Uemura M, Brait M, Shao C, Westra W, Bishop J, Fertig EJ & Califano JA. Expression microarray analysis reveals alternative splicing of LAMA3 and DST genes in head and neck squamous cell carcinoma. *PLoS One* **9**, e91263, (2014). PMC3967989.

18 Ling S, Rettig EM, Tan M, Chang X, Wang Z, Brait M, Bishop JA, Fertig EJ, Considine M, Wick MJ & Ha PK. Identification of methylated genes in salivary gland adenoid cystic carcinoma xenografts using global demethylation and methylation microarray screening. *Int J Oncol* **49**, 225-234, (2016). PMC4902070.

19 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.

20 Rettig EM, Talbot CC, Jr., Sausen M, Jones S, Bishop JA, Wood LD, Tokheim C, Niknafs N, Karchin R, Fertig EJ, Wheelan SJ, Marchionni L, Considine M, Fakhry C, Papadopoulos N, Kinzler KW, Vogelstein B, Ha PK & Agrawal N. Whole-Genome Sequencing of Salivary Gland Adenoid Cystic Carcinoma. *Cancer Prev Res (Phila)* **9**, 265-274, (2016). PMC4818686.

21 Sun W, Gaykalova DA, Ochs MF, Mambo E, Arnaoutakis D, Liu Y, Loyo M, Agrawal N, Howard J, Li R, Ahn S, Fertig E, Sidransky D, Houghton J, Buddavarapu K, Sanford T, Choudhary A, Darden W, Adai A, Latham G, Bishop J, Sharma R, Westra WH, Hennessey P, Chung CH & Califano JA. Activation of the NOTCH pathway in head and neck cancer. *Cancer Res* **74**, 1091-1104, (2014). PMC3944644.

22 Wang Z, Ling S, Rettig E, Sobel R, Tan M, Fertig EJ, Considine M, El-Naggar AK, Brait M, Fakhry C & Ha PK. Epigenetic screening of salivary gland mucoepidermoid carcinoma identifies hypomethylation of CLIC3 as a common alteration. *Oral Oncol* **51**, 1120-1125, (2015). PMC4663116.

23 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.

24 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.

25 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.

26 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.

27 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.