**Proposed title:** Parallel Gene Activity in Pattern Sets (P-GAPS) for systematic characterization of transcriptional variation in the Human Cell Atlas

**Summary** Distinguishing the sources of variation in gene expression between tissue samples from separate individuals is essential to find the hidden sources of genetic variation in complex diseases. However, tissue-specific differences are often more pronounced than inter-individual differences1. This detection problem becomes even more complex when accounting for substantial technical and biological variation in transcriptional data. Artifacts from sample preparation and processing techniques are pervasive to nearly every comprehensive database of genomics data2. Further, the impact of technical variation on genomics data is highly variable within each experiment. Therefore, bioinformatics pattern detection algorithmsthat are robust to technical artifacts are necessary to disentangle tissue and individual specific co-expression patterns. The development of and cross-study and cross-platform analyses with such pattern detection algorithms are essential first steps to assess the biological replicability for the Human Cell Atlas.

Removing technical artifacts within genomics data is a mature field3,4. Batch correction algorithms are largely divided into two classes. The first class of algorithms protect for known biological covariates, but inadvertently remove unmodeled biological variation along with the technical artifacts5,6. Conversely, the second class of algorithms model known batches. While these can preserve unknown biological variability, they can overcorrect on complex study designs leading to false negatives6. Thus, neither class of algorithms can preserve both unknown inter-individual and inter-tissue variation. Newer algorithms that model technical variation among control probes or genes mitigate these challenges7-9 but are incompatible across data platforms. Therefore, techniques to simultaneously delineate technical and biological variation in complex, multi-platform studies are needed.

Previously, we developed the CoGAPS (Coordinate Gene Activity in Pattern Sets) Bayesian, sparse non-negative matrix factorization algorithm10 to discriminate the molecular pathways responding to treatment or other activation11,12, changing along time course13,14, and distinguish tumor subtypes11. Mathematically, this algorithm 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**).

To optimize CoGAPS for whole genome analysis, we also developed a new parallel version of CoGAPS, GWCoGAPS, the first non-negative matrix factorization algorithm that converges for genome-wide data without requiring compaction15,16. GWCoGAPS has been proven to maintain the sensitivity of CoGAPS. Applied to expression data from tissues of healthy, postmortem donors in GTeX1, the algorithm can distinguish genes associated with cell type of origin from genes that are associated with an individual’s genetic background.

Preliminary studies in stem cell differentiation have demonstrated the ability of this genome-wide approach to simultaneous infer robust differentiation patterns and technical artifacts between Ribo0 and PolyA RNA-sequencing data sets (**Fig 2**). Given the high concordance of these technical patters the with artifacts defined by intensive spike-in Ribo0 and PolyA RNA-sequencing experiments17, CoGAPS is uniquely suited to distinguish technical and biological variation in complex study designs. The central aim of this research project is to extend the CoGAPS to large study designs of healthy tissues (**Aim 1**) and between single cell and bulk RNA-seq (**Aim 2**) for the Human Cell Atlas database through the following project aims.

**Project Aims, and how they address program goals**:

**Aim 1 Parallel Gene Activity in Pattern Sets (P-GAPS) for systematic characterization of biological replicability in** **transcriptional data from healthy samples in GTeX and TCGA.** CoGAPS can be implemented in parallel across genes for whole-genome analysis, but is limited to O(1000) samples for guaranteed convergence. 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 bulk RNA-sequencing data from normal samples of multiple human tissue types from GTEX1 and TCGA (**Table 1**). We will apply CoGAPS to determine patterns in gene expression that are associated with technical artifacts between the sample cohorts, individual-specific patterns, and tissue-specific patterns. **Program goal:** The P-GAPS patterns associated with technical variation will be compared with donor age, sex, and tissue to assess biological replicability in human samples.

**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 matched samples in bulk RNA-seq, Smart-Seq2, and 10X single cell sequencing. We will model the resulting common biological patterns across sequencing platforms as a function of the sparsity hyperparameter. **Program goal:** The resulting analysis will enable systematic comparison of data sparsity from distinct sequencing technologies.

**Experimental plan and deliverables:**

**Aim 1 Parallel Gene Activity in Pattern Sets (P-GAPS) for systematic characterization of biological replicability in** **transcriptional data from healthy samples in GTeX and TCGA.**

**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 genes19,20. 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. 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 previously6 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.
2. We will divide the combined GTeX and TCGA 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.

**Experimental design for GTEX and TCGA data analysis.** We will obtain gene-level count data for normal samples in GTeX and TCGA using common preprocessing pipelines from ReCount18. This will provide gene-level data for XX samples (Table 1).

**Statistical analysis for characterization of systematic characterization of biological replicability in P-GAPS patterns.** Sample patterns will be analyzed with linear models against sample metadata with the R/CRAN package multcomp21 to assess association with individual donor (GTeX only), smoking status (TCGA only), processing batches within each study, tissue type, age, gender, race, and batch.

**Pitfalls and proposed solutions.** (1) If P-GAPS fails to converge because of the large sample size, we will implement compaction approaches across samples15,16. (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 data10 and RNA-sequencing20 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.

**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 MAGIC22. (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.

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

We have performed numerous genomics analyses of data from complex experimental designs in cancer6,11-13,23-39. Our work on cross-study analyses of tumor samples and biopsies with various procurements leaded to methods for integrating both technical and biological data types. Specifically, we developed a novel batch correction algorithm, pSVA, to preserve biological variation for unsupervised analysis6 and 2) a concomitant analysis of DNA methylation and gene expression with CoGAPS11. Recently, we have adapted CoGAPS to study biological variation within healthy human stem cells and postmortem tissues 19,20.

**Description of commitment to full sharing of primary data, metadata, methods, and software.**

**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>).

**Methods:** All methods will be fully described and submitted for peer-reviewed and Bioarxiv publications.

**Software:** Any algorithms will be released as part of the open-source CoGAPS Bioconductor package10. Bioconductor is a centralized, peer-reviewed database and release there will support future application and development by the broader community.

**References**

1 GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, (2015). PMC4547484.

2 Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K & Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**, 733-739, (2010). PMC3880143.

3 Johnson WE, Li C & Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127, (2007). PMID 16632515, PMCID not available.

4 Leek JT & Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**, 1724-1735, (2007). PMC1994707.

5 Parker HS & Leek JT. The practical effect of batch on genomic prediction. *Stat Appl Genet Mol Biol* **11**, Article 10, (2012). PMC3760371.

6 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.

7 Gagnon-Bartsch JA & Speed TP. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539-552, (2012). PMC3577104.

8 Li S, Labaj PP, Zumbo P, Sykacek P, Shi W, Shi L, Phan J, Wu PY, Wang M, Wang C, Thierry-Mieg D, Thierry-Mieg J, Kreil DP & Mason CE. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat Biotechnol* **32**, 888-895, (2014). PMC4160374.

9 Tung PY, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK & Gilad Y. Batch effects and the effective design of single-cell gene expression studies. *Sci Rep* **7**, 39921, (2017). PMC5206706.

10 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.

11 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.

12 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.

13 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.

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

15 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.

16 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.

17 Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer PA, Gao Y, Kim D, Boland J, Hicks B, Kim R, Chhangawala S, Jafari N, Raghavachari N, Gandara J, Garcia-Reyero N, Hendrickson C, Roberson D, Rosenfeld J, Smith T, Underwood JG, Wang M, Zumbo P, Baldwin DA, Grills GS & Mason CE. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* **32**, 915-925, (2014). PMC4167418.

18 Frazee AC, Langmead B & Leek JT. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics* **12**, 449, (2011). PMC3229291.

19 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.

20 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.

21 Hothorn T, Bretz F & Westfall P. Simultaneous inference in general parametric models. *Biom J* **50**, 346-363, (2008). PMID 18481363, PMCID not available.

22 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.

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

24 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.

25 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.

26 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.

27 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.

28 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.

29 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.

30 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.

31 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.

32 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.

33 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.

34 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.

35 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.

36 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.

37 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.

38 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.

39 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.