

# 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@df766e2](#) on November 3, 2018.

## Authors

---

- **Loyal A. Goff**

 [0000-0003-2875-451X](#) ·  [loyale](#) ·  [loyalgoff](#)

Solomon H. Snyder Department of Neuroscience, Johns Hopkins University School of Medicine; Kavli Neurodiscovery Institute, Johns Hopkins University; McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine

- **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

- **Rob Patro**

 [0000-0001-8463-1675](#) ·  [rob-p](#)

Department of Computer Science, Stony Brook University

- **Elana J. Fertig**

 [0000-0003-3204-342X](#) ·  [ejfertig](#) ·  [FertigLab](#)

Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, School of Medicine, Johns Hopkins University; Department of Applied Mathematics and Statistics, Whiting School of Engineering, Johns Hopkins University

- **Michael I. Love**

 [0000-0001-8401-0545](#) ·  [mikelove](#) ·  [mikelove](#)

Department of Biostatistics, University of North Carolina at Chapel Hill; Department of Genetics, University of North Carolina at Chapel Hill

- **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

---

- Hicks refs: [\[1\]](#)
- projectR & scCoGAPS: [\[2\]](#)
- Alevin: [\[3\]](#)

## Project Team

---

### PI information

1. Loyal Goff (Submitter)
  - Title: Assistant Professor
  - Degrees: PhD
  - Type of organization: Academic
  - Tax ID: 52-0595110 (JHU)
  - Email: [loyalgoff@jhmi.edu](mailto:loyalgoff@jhmi.edu)
2. Stephanie Hicks
  - Title: Assistant Professor
  - Degrees: PhD
  - Type of organization: Academic
  - Tax ID: 52-0595110 (JHU)
  - Email: [shicks19@jhu.edu](mailto:shicks19@jhu.edu)
3. Elana Fertig
  - Title: Associate Professor
  - Degrees: PhD
  - Type of organization: Academic
  - Tax ID: 52-0595110 (JHU)
  - Email: [ejfertig@jhmi.edu](mailto:ejfertig@jhmi.edu)
4. Casey Greene
5. Tom Hampton

6. Michael Love

- Title: Assistant Professor
- Degrees: Dr. rer. nat.
- Type of organization: Academic
- Tax ID: 56-6001393 (UNC)
- Email: milove@email.unc.edu

7. Rob Patro

- Title: Assistant Professor
- Degrees: PhD
- Type of Organization: Academic
- Tax ID: 16-1514621 (Stony Brook)
- Email: rob.patro@cs.stonybrook.edu

## Description (750 words TOTAL)

1. Loyal Goff

2. 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 [1]. Dr. Hicks will work together with Co-PIs to implement fast search algorithms in latent spaces (Aim 1) and to implement the methods developed into fast, scalable, and memory-efficient R/Bioconductor software packages (Aim 3).

3. Elana Fertig is an Associate Professor of Oncology and Applied Mathematics and Statistics at Johns Hopkins University. She developed the Bayesian non-negative matrix factorization algorithm CoGAPS [4] for latent space analysis. In collaboration with co-PI Goff, she adapted this tool to scRNA-seq data and developed a new transfer learning framework to relate the low-dimensional features in scRNA-seq data across data modalities, biological conditions, and organisms [2]. Dr. Fertig will work with the co-PIs to incorporate the error models from Aim 1 into the latent space representations, dimensionality estimation, and biological assessment metrics in Aim 2. She is developing standardized language for latent space representation in collaboration with co-PIs Goff and Greene [5] that will provide a strong foundation for standardization of these approaches across different unsupervised learning tools.

4. Casey Greene

5. Tom Hampton

6. Michael Love is an Assistant Professor of Biostatistics and Genetics at the University of North Carolina at Chapel Hill. He is a leading developer of statistical software for RNA-seq analysis in the Bioconductor Project, maintaining the widely used DESeq2 [6] and tximport [7] packages. He is a close collaborator with Dr. Rob Patro on bias-aware estimation of transcript

abundance from RNA-seq and estimation of uncertainty during transcript quantification [8]. Dr. Love will work with co-PIs to disseminate versioned reference cell type catalogs through widely used frameworks for genomic data analysis including R/Bioconductor and Python.

7. Rob Patro

## Proposal Body (2000 words)

---

High dimensional data can often be compressed into fewer dimensions without a substantial loss of information. For transcriptomic data, compressing on the gene dimension is most attractive: it can be done a single sample at a time, and genes are often co-regulated and thus provide information about each other. In the best case, the fewer dimensions both capture the biological sources of variability while ignoring noise and each dimension or small combination of dimensions aligns to interpretable biological processes.

For the Human Cell Atlas (HCA), low-dimensional representations enable efficient search and transformation, with the benefits becoming particularly pronounced as the number of cells and tissues becomes particularly large. Our **central hypothesis** is that these approaches will also enable search at the tissue and cell level of biologically-meaningful features and transformation in biologically meaningful ways, e.g., from a healthy to a disease context. We propose to advance **base enabling technologies** for low-dimensional representations and to develop solutions for search and transformation that can be practically applied across the entirety of the HCA. We also propose three aims: 1) fast and accurate search for cell, samples, and pathways; 2) statistics to assess, interpret, and define cell types in low-dimensional spaces; and 3) to increase the impact of the HCA and low-dimensional methods by enhancing software and training opportunities.

The *first part of our work on base enabling technologies* is the continued development of techniques that seek to learn interpretable biologically-aligned representations. Our work on the Bayesian, non-negative matrix factorization method scCoGAPS [9] (PI Fertig) has demonstrated its suitability for this domain. The method robustly infers a low-dimensional representation of perturbation [10] and time course [12] data, leading to the winning solution in the HPN DREAM8 challenge [14]. Our recent extension of this method to single cell data in the developing mouse retina simultaneously distinguishes cellular identity, dynamic trajectories, and cell state [9]. Notably for the work proposed in this seed network, this method includes an uncertainty estimate that can be readily modified to account for measurement-specific technical variation [15]. As increasing spatial data becomes available, we will extend these techniques to incorporate this additional source of information.

A complementary strategy for low-dimensional representations uses neural networks, which may consist of multiple layers. These techniques can learn a non-linear mapping into the low-dimensional space. We have previously worked with such methods [16] (PI Greene). However, because so many groups are working in this area (see [17] with new methods appearing each month), we don't propose to do specific work in this area during this project period. Instead, we

plan to continue to use and rigorously evaluate these methods and to incorporate the best performing methods into the search and transformation approaches that we propose. In the event that it becomes clear that work to advance enabling technologies is required in this area, we are well positioned to address such needs.

Members of our team have developed a common language for interpretation of matrix factorization methods to facilitate a unification of latent space methods [5]. The latent space team from the HCA collaborative networks RFA (including PIs Fertig, Goff, Greene, and Patro) is defining common output formats for low-dimensional representations from distinct classes of methods. In Aim 3 we propose to, where required, write wrappers around selected existing and newly developed methods that will facilitate the practical use of methods that learn a low-dimensional representation of the HCA.

The *second part of our work on base enabling technologies* is the improvement of techniques for fast and accurate quantification. Existing approaches for quantification from scRNA-seq data using tagged-end protocols (e.g. 10x Chromium, drop-Seq, inDrop, etc.) have no mechanism for accounting for reads mapping between multiple genes in the resulting quantification estimates. This affects approximately 15-25% of the reads in a typical experiment. It reduces quantification accuracy, and leads to systematic biases in gene expression estimates that correlate with the size of gene families and gene function [3]. We recently developed a quantification method for tagged-end data that accounts for reads mapping to multiple genomic loci in a principled and consistent way [CITE?]. We will expand on this work by, building these capabilities into a production quality tool for the processing of scRNA-seq data. The tool will support: 1. Exploring alternative models for UMI resolution. 2. Developing new approaches for quality control and filtering using the UMI-resolution graph. 3. Creating a compressed and indexable data structure for the UMI-resolution graph to enable direct access, query, and fast search, which will support our Aim 1.

## Aim 1

*Rationale:* need to add

*Fast search algorithms* (in low dimensions, quantifying differences between case and reference maps, twist: shared latent spaces / k-mers):

*Finding differences between maps:* The fast search algorithms will then be used 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 at the population or individual level 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 if available). Our metric to quantify

differences will depend on the distributions of cell expression within and between individuals, which Dr. Hicks has extensive experience with [22]. 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).

*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

*Rationale:* Biological systems are comprised of diverse cell types with overlapping molecular phenotypes, and biological processes are often reused with modifications across cellular contexts. The functional output of these systems is determined by the interactions between these complex components, rather than a single gene or cell. This suggests that fundamental biological mechanisms may broadly contribute to an observed state, with context-specific modifiers conferring selective susceptibility to disease. Latent space techniques are poised to reveal these fundamental mechanisms in the broad survey of single cell data across model systems and cellular contexts in the Human Cell Atlas. We hypothesize that the features learned from these techniques will define constitutive basis vectors that reflect discrete biological processes or features. Thus, these basis vectors will be shared across different biological systems, with context-specific perturbations indicating pathogenic differences in disease. *We propose a central suite of statistics for assessment and interpretation of latent space tools to define the identity and dimensionality of biological systems.*

*Quantifying latent space estimation with transfer learning:* A critical challenge to latent space methods is the quantification of methods performance. Numerous computational metrics have been developed to assess convergence of the low dimensional estimation. However, these metrics do not quantify whether the features in a low dimensional representation of scRNA-seq data represent biological processes in the measured system. The performance of these methods can be quantified directly in datasets for which cell types and states are known (e.g., perturbation experiments, controlled admixture experiments, etc). However, these annotations are lacking in most biological datasets limiting any such quantification. Transfer learning methods have been developed in machine learning to relate features learned in a source dataset to those in a new,

target dataset in order to transfer annotations from one context to another. In this project, we will adapt these methods to quantify the performance of latent space methods by the extent to which learned low dimensional features from a source dataset transfer to a target dataset in a related biological context. We will benchmark the performance of the resulting metric on simulated datasets, cross-validation in scRNA-seq datasets with known cell types and states, and cross-study validation of systems in related biological contexts with known cell types and states. Gene set enrichment methods will also be used to explore the relevant biological processes described by individual basis vectors, and related bases will be identified through clustering and exploratory approaches in these benchmark datasets. Our transfer learning based metric will be piloted on low dimensional representations learned with scCoGAPS and then applied to a broader suite of latent space tools. We will release software for this transfer learning quantification of latent space representations in R and Python using standard latent space file formats developed by our team in the first year of HCA funding.

*Dimensionality estimation:* Dimensionality reduction methods are sensitive to the number of low features learned in each dataset. Many computational techniques optimize dimensionality by creating a cost function which penalizes models with higher number of features. Similar to the quantification metrics, these penalty terms do not reflect the extent to which features learned at a given dimensionality reflect biology. Moreover, many systems may have more than one biologically accurate low dimensional representation. Such multiple truths in data would be particular prominent in systems that can be subdivided into hierarchical classifications. For example, in the case of cancer we observed that a low dimensional representation of bulk data learned from CoGAPS distinguished cancers from normals whereas a higher dimension distinguished tumor subtypes [15]. Both of these low dimensional representations are equally valid, and each reflects different biological features in the data. To find these multiple truths, we will develop a parallel framework to run scCoGAPS for multiple dimensionalities and quantify performance with our transfer-learning based metric on random subsets of the data. The dimensions with greatest cross-validated feature robustness will be retained as the optimal dimensionalities for each dataset. We will develop software to enable this cross-validation dimensionality estimation across multiple latent space methods. We note that this same software will provide a robust tool to define ensembles of low dimensional representations that reflect underlying biology learned across multiple latent space methods. **Rob: I'm not sure if you want to fill in some of your ideas re persistent homology instead. Very open to that idea and think it may be a nice, more efficient methodology than what's proposed here.**

*Search tool for latent spaces and reference cell types:* **Loyal, Casey – what are the datasets that will be used for this – I would think all healthy cells in a single system to enable quantification of context-specific in the next part of this aim.** Comprehensive identification of basis vectors across conditions is an area of active research for our group in the previous funding period. We will use scCoGAPS and other tools developed within our collaborative network to establish a compendium of basis vectors across our single cell catalog. Ensembles of the low



dimensional features that represent robust biological features across methods using methods described above will be preserved as the 'biological basis' of the Human Cell Atlas. The weights of these bases will be correlated across all available metadata attributes for each cell to identify basis vectors that are associated with specific cellular contexts, disease states, technical parameters, or other phenotypic features. A reference catalog of gene weights for specific cell types will be defined by the set of basis vectors associated with cellular identity in datasets with known ground truth. We will adapt the software we developed for transfer learning of features from bulk data recount [23] to facilitate querying of signatures in new user-defined datasets (delivery of which is described in the next aim). As datasets accumulate and methods are refined, the biological basis and reference catalog of gene weights will evolve over time. To enable reproducible research leveraging HCA, we will implement a content-based versioning system, which identifies versions of the reference cell type catalog by the gene weights and transcript nucleotide sequences using a hash function. Such a hash-based versioning and provenance identification and detection framework has proven successful in the bulk RNA-seq context to support reproducible computational analyses [24].

*Differentiating context-specific latent spaces from latent spaces that are universal across biological contexts:* The search tool to define reference cell types based upon latent spaces was defined for healthy tissues from XXX (some control). Deviations of common cell types or states from the healthy baseline in other populations will indicate context-specific alterations, which may be associated with disease. To identify potentially pathogenic responses in target datasets, we will implement a random forest classifier into our transfer learning method to segregate cells based on their usage of disease-associated basis vectors after projection. In other cases, disease may arise from changes in variation reflective of inter-cellular heterogeneity. Therefore, we will also develop methods to quantify variation from latent space vectors. Both methods will be incorporated in our latent space search tool. **Loyal: I'm not sure if this is what you had in mind. It may also be that these are reflected in the hierarchy of dimensionality – may want to incorporate here.**

The technologies to improve quantification will have a critical impact on the outcomes of latent spaces. However, there are currently no standardized, quantitative metrics to determine relative uncovering of biology from low dimensional representations. We have developed new transfer learning methods to quantify the extent to which latent space representations from one set of training data are represented in another [2]. These tools provide a strong foundation to enable biological quantification of latent space representations by quantifying the extent to which those spaces transfer across datasets of related biological contexts.

### Aim 3

*Rationale:* Low dimensional representations provide a powerful means to analyze scRNA-seq and HCA data to make tasks faster, perform more biologically grounded analyses, and provide interpretable summaries of complex high-dimensional data. However, using these capabilities to the fullest extent requires integration with widely used toolkits, and in a scalable education effort that can reach students from the undergraduate level and beyond. *We propose to enhance*

*software usability and develop open instructional materials that we will use to deliver short-course training that includes the topics of single cell profiling, machine learning methods, low-dimensional representations, reference cell type catalogs, and tools developed by our group in response to this RFA.*

### **Aim 3.1: Enhance software usability for low-dimensional tools**

We will implement the proposed methods from Aims 1 and 2 into robust software, which will be integrated into the R/Bioconductor and Python frameworks. The software will be fast, scalable, and memory-efficient because will leverage the computational tools previously developed by Bioconductor for single-cell data access to the HCA, data representation (`SingleCellExperiment`, `beachmat`, `DelayedArray`, `HDF5Array` and `rhdf5`) and data assessment and ameliorization of data quality (`scater`, `scraper`, `DropletUtils`).

### **Aim 3.: Data Integration into Bioconductor (Stephanie and Mike)**

We will integrate the catalogs of reference cell types (Aim 2) into the R/Bioconductor and Python frameworks. The use of consortia data summaries as annotation and interpretive scaffold for outside datasets has proven widely successful for the ENCODE, Roadmap Epigenome Mapping, and GTEx projects. We will use HCA data to define similar summaries as annotation and scaffolding for local genomic datasets. Additional information about biological variability that can be transferred from individual or ensemble latent spaces, can also be used landmarks to ordinate differences among populations of cells observed in a local datasets, whether bulk or single-cell experiments. We will package and version reference cell types, including measures of technical and biological variability transferred from the latent spaces defined in earlier aims, and deliver these as structured data objects in Bioconductor and Python. We will leverage our expertise in working with Bioconductor both as core package developers and power users to enable on-the-fly downloading of reference cell types and variability via the *AnnotationHub* framework, which includes rich specification of metadata including the provenance and versioning of the catalog. Python workflows leveraging existing frameworks for genomic data analysis (Biopython, BioRanges, and bx-python) will also be supported via cross-platform packages such as Feather that facilitate transfer of data objects between languages. We will develop *F1000Research* workflows for R and python demonstrating how HCA-defined reference cell types and tools developed in this RFA can be used within a typical genomic data analysis.

### **Aim 3.2: Training the next-generation of single cell data scientists**

The HCA data, and low-dimensional representation methods that work with them, could increase the rate of discovery across many biomedical fields, but using the data and methods effectively will require experience with a new toolkit. We have designed an education and outreach effort to fill this gap.

We have developed and run an annual “Applied Bioinformatics” one-week short course at Mount Desert Island Biological Lab over the last **X TOM FILL IN** years (PI Hampton). The course covers R, fundamentals of gene expression analysis, statistical interpretation, and provides an introduction to machine learning (PI Greene). In addition to students and postdocs, attendees include faculty from research intensive universities and faculty from primarily undergraduate institutions. Attendees give the course high marks and report that they will use what they learned at the short course in their research and teaching.

Under this grant we will enhance this short course by adding topics required to successfully use the HCA. We will also increase the frequency of the course and run the course at locations distributed throughout the US. We will provide open course materials on GitHub to allow others to replicate the course. We will aim to provide at least *five scholarships* to individuals who are from communities that are underrepresented. The expanded topics will include:

- UNIX
- The R Statistical Programming Environment
- Visualization and Exploration of High Dimensional Data
- Statistical Approaches for High Dimension Biomedical Data
- Gene Set and Pathway Analysis for Bulk Gene Expression Data
- Low-dimensional Representations of High Dimensional Data
- Compare and Contrast Bulk and Single-cell Biology
- scRNA-seq Assay Methods and Data
- The Human Cell Atlas Project
- scRNA-seq Computational Tools for Quantification and Cell Type Discovery
- scRNA-seq Statistical Tools for Low-dimensional Representations
- Tools for Search and Analysis in Low-dimensional Representations

The short courses will be run on a cost recovery model, but we will increase exposure to the HCA at the undergraduate level by providing at least *ten scholarships* per course that cover room, board, and tuition to faculty who are primarily engaged in undergraduate instruction. This, combined with the geographically distributed locations, will allow faculty with this mission to attend at very low cost. We will also develop a one-week module that can be added in to an undergraduate class on single-cell profiling and the HCA, which we will distribute via GitHub. The materials will include recorded videos (primarily intended for a refresher for the instructors), slides, and exercises. We expect that this module will support faculty who attend with an easy enhancement to any bioinformatics or computational biology instruction that they are already providing at their institution. This structure aims to produce a force multiplier, where attendees to the course can eventually transmit what they learn to tens of students each year.

# References

---

**1. Missing data and technical variability in single-cell RNA-sequencing experiments**

Stephanie C Hicks, F William Townes, Mingxiang Teng, Rafael A Irizarry

*Biostatistics* (2017-11-06) <https://doi.org/gfb8g4>

DOI: [10.1093/biostatistics/kxx053](https://doi.org/10.1093/biostatistics/kxx053) · PMID: [29121214](https://pubmed.ncbi.nlm.nih.gov/29121214/)

**2. Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species.**

Genevieve L Stein-O'Brien, Brian S. Clark, Thomas Sherman, Christina Zibetti, Qiwen Hu, Rachel Sealfon, Sheng Liu, Jiang Qian, Carlo Colantuoni, Seth Blackshaw, ... Elana J. Fertig

*Cold Spring Harbor Laboratory* (2018-08-20) <https://doi.org/gd2xpn>

DOI: [10.1101/395004](https://doi.org/10.1101/395004)

**3. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data**

Avi Srivastava, Laraib Malik, Tom Sean Smith, Ian Sudbery, Rob Patro

*Cold Spring Harbor Laboratory* (2018-06-01) <https://doi.org/gffk42>

DOI: [10.1101/335000](https://doi.org/10.1101/335000)

**4. CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data**

Elana J. Fertig, Jie Ding, Alexander V. Favorov, Giovanni Parmigiani, Michael F. Ochs

*Bioinformatics* (2010-09-01) <https://doi.org/cwqsv4>

DOI: [10.1093/bioinformatics/btq503](https://doi.org/10.1093/bioinformatics/btq503) · PMID: [20810601](https://pubmed.ncbi.nlm.nih.gov/20810601/) · PMCID: [PMC3025742](https://pubmed.ncbi.nlm.nih.gov/PMC3025742/)

**5. Enter the Matrix: Factorization Uncovers Knowledge from Omics**

Genevieve L. Stein-O'Brien, Raman Arora, Aedin C. Culhane, Alexander V. Favorov, Lana X. Garmire, Casey S. Greene, Loyal A. Goff, Yifeng Li, Aloune Ngom, Michael F. Ochs, ... Elana J. Fertig

*Trends in Genetics* (2018-10) <https://doi.org/gd93tk>

DOI: [10.1016/j.tig.2018.07.003](https://doi.org/10.1016/j.tig.2018.07.003) · PMID: [30143323](https://pubmed.ncbi.nlm.nih.gov/30143323/)

**6. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**

Michael I Love, Wolfgang Huber, Simon Anders

*Genome Biology* (2014-12) <https://doi.org/gd3zvn>

DOI: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8) · PMID: [25516281](https://pubmed.ncbi.nlm.nih.gov/25516281/) · PMCID: [PMC4302049](https://pubmed.ncbi.nlm.nih.gov/PMC4302049/)

## **7. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences**

Charlotte Soneson, Michael I. Love, Mark D. Robinson

*F1000Research* (2015-12-30) <https://doi.org/gdtgw8>

DOI: [10.12688/f1000research.7563.1](https://doi.org/10.12688/f1000research.7563.1) · PMID: [26925227](https://pubmed.ncbi.nlm.nih.gov/26925227/) · PMCID: [PMC4712774](https://pubmed.ncbi.nlm.nih.gov/PMC4712774/)

## **8. Salmon provides fast and bias-aware quantification of transcript expression**

Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, Carl Kingsford

*Nature Methods* (2017-03-06) <https://doi.org/gcw9f5>

DOI: [10.1038/nmeth.4197](https://doi.org/10.1038/nmeth.4197) · PMID: [28263959](https://pubmed.ncbi.nlm.nih.gov/28263959/) · PMCID: [PMC5600148](https://pubmed.ncbi.nlm.nih.gov/PMC5600148/)

## **9. Comprehensive analysis of retinal development at single cell resolution identifies NFI factors as essential for mitotic exit and specification of late-born cells**

Brian Clark, Genevieve Stein-O'Brien, Fion Shiau, Gabrielle Cannon, Emily Davis, Thomas Sherman, Fatemeh Rajaii, Rebecca James-Esposito, Richard Gronostajski, Elana Fertig, ... Seth Blackshaw

*Cold Spring Harbor Laboratory* (2018-07-27) <https://doi.org/gdwrzh>

DOI: [10.1101/378950](https://doi.org/10.1101/378950)

## **10. Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma**

Elana J Fertig, Qing Ren, Haixia Cheng, Hiromitsu Hatakeyama, Adam P Dicker, Ulrich Rodeck, Michael Considine, Michael F Ochs, Christine H Chung

*BMC Genomics* (2012) <https://doi.org/gb3fgp>

DOI: [10.1186/1471-2164-13-160](https://doi.org/10.1186/1471-2164-13-160) · PMID: [22549044](https://pubmed.ncbi.nlm.nih.gov/22549044/) · PMCID: [PMC3460736](https://pubmed.ncbi.nlm.nih.gov/PMC3460736/)

## **11. CoGAPS matrix factorization algorithm identifies transcriptional changes in AP-2alpha target genes in feedback from therapeutic inhibition of the EGFR network**

Elana J. Fertig, Hiroyuki Ozawa, Manjusha Thakar, Jason D. Howard, Luciane T. Kagohara, Gabriel Krigsfeld, Ruchira S. Ranaweera, Robert M. Hughes, Jimena Perez, Siân Jones, ... Christine H. Chung

*Oncotarget* (2016-09-16) <https://doi.org/f9k8d8>

DOI: [10.18632/oncotarget.12075](https://doi.org/10.18632/oncotarget.12075) · PMID: [27650546](https://pubmed.ncbi.nlm.nih.gov/27650546/) · PMCID: [PMC5342018](https://pubmed.ncbi.nlm.nih.gov/PMC5342018/)

**12. Pattern Identification in Time-Course Gene Expression Data with the CoGAPS Matrix Factorization**

Elana J. Fertig, Genevieve Stein-O'Brien, Andrew Jaffe, Carlo Colantuoni

*Gene Function Analysis* (2013-10-24) <https://doi.org/f5j7xj>

DOI: [10.1007/978-1-62703-721-1\\_6](https://doi.org/10.1007/978-1-62703-721-1_6) · PMID: [24233779](https://pubmed.ncbi.nlm.nih.gov/24233779/)

**13. Integrated time course omics analysis distinguishes immediate therapeutic response from acquired resistance**

Genevieve Stein-O'Brien, Luciane T. Kagohara, Sijia Li, Manjusha Thakar, Ruchira Ranaweera, Hiroyuki Ozawa, Haixia Cheng, Michael Considine, Sandra Schmitz, Alexander V. Favorov, ... Elana J. Fertig

*Genome Medicine* (2018-05-23) <https://doi.org/gfc4dq>

DOI: [10.1186/s13073-018-0545-2](https://doi.org/10.1186/s13073-018-0545-2) · PMID: [29792227](https://pubmed.ncbi.nlm.nih.gov/29792227/) · PMCID: [PMC5966898](https://pubmed.ncbi.nlm.nih.gov/PMC5966898/)

**14. Inferring causal molecular networks: empirical assessment through a community-based effort**

Steven M Hill, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, ... Sach Mukherjee

*Nature Methods* (2016-02-22) <https://doi.org/f3t7t4>

DOI: [10.1038/nmeth.3773](https://doi.org/10.1038/nmeth.3773) · PMID: [26901648](https://pubmed.ncbi.nlm.nih.gov/26901648/) · PMCID: [PMC4854847](https://pubmed.ncbi.nlm.nih.gov/PMC4854847/)

**15. Preferential Activation of the Hedgehog Pathway by Epigenetic Modulations in HPV Negative HNSCC Identified with Meta-Pathway Analysis**

Elana J. Fertig, Ana Markovic, Ludmila V. Danilova, Daria A. Gaykalova, Leslie Cope, Christine H. Chung, Michael F. Ochs, Joseph A. Califano

*PLoS ONE* (2013-11-04) <https://doi.org/gcpgc6>

DOI: [10.1371/journal.pone.0078127](https://doi.org/10.1371/journal.pone.0078127) · PMID: [24223768](https://pubmed.ncbi.nlm.nih.gov/24223768/) · PMCID: [PMC3817178](https://pubmed.ncbi.nlm.nih.gov/PMC3817178/)

**16. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics**

Qiwen Hu, Casey S Greene

*Cold Spring Harbor Laboratory* (2018-08-05) <https://doi.org/gdxxjf>

DOI: [10.1101/385534](https://doi.org/10.1101/385534)

**17. Single cell RNA-seq denoising using a deep count autoencoder**

Gökçen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, Fabian J. Theis

*Cold Spring Harbor Laboratory* (2018-04-13) <https://doi.org/gdjcb3>

DOI: [10.1101/300681](https://doi.org/10.1101/300681)

**18. Bayesian Inference for a Generative Model of Transcriptome Profiles from Single-cell RNA Sequencing**

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, Nir Yosef

*Cold Spring Harbor Laboratory* (2018-03-30) <https://doi.org/gdm9jf>

DOI: [10.1101/292037](https://doi.org/10.1101/292037)

**19. Exploring Single-Cell Data with Deep Multitasking Neural Networks**

Matthew Amodio, David van Dijk, Krishnan Srinivasan, William S Chen, Hussein Mohsen, Kevin R Moon, Allison Campbell, Yujiao Zhao, Xiaomei Wang, Manjunatha Venkataswamy, ... Smita Krishnaswamy

*Cold Spring Harbor Laboratory* (2017-12-19) <https://doi.org/gfgrpk>

DOI: [10.1101/237065](https://doi.org/10.1101/237065)

**20. Massive single-cell RNA-seq analysis and imputation via deep learning**

Yue Deng, Feng Bao, Qionghai Dai, Lani Wu, Steven Altschuler

*Cold Spring Harbor Laboratory* (2018-05-06) <https://doi.org/gfgrpm>

DOI: [10.1101/315556](https://doi.org/10.1101/315556)

**21. Transfer learning in single-cell transcriptomics improves data denoising and pattern discovery**

Jingshu Wang, Divyansh Agarwal, Mo Huang, Gang Hu, Zilu Zhou, Vincent B. Conley, Hugh MacMullan, Nancy R. Zhang

*Cold Spring Harbor Laboratory* (2018-10-31) <https://doi.org/gfgrpn>

DOI: [10.1101/457879](https://doi.org/10.1101/457879)

**22. quantro: a data-driven approach to guide the choice of an appropriate normalization method.**

Stephanie C Hicks, Rafael A Irizarry

*Genome biology* (2015-06-04) <https://www.ncbi.nlm.nih.gov/pubmed/26040460>

DOI: [10.1186/s13059-015-0679-0](https://doi.org/10.1186/s13059-015-0679-0) · PMID: [26040460](https://pubmed.ncbi.nlm.nih.gov/26040460/) · PMCID: [PMC4495646](https://pubmed.ncbi.nlm.nih.gov/PMC4495646/)

**23. MultiPLIER: a transfer learning framework reveals systemic features of rare autoimmune disease**

Jaclyn N Taroni, Peter C Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A Merkel, Casey S Greene

*Cold Spring Harbor Laboratory* (2018-08-20) <https://doi.org/gfc9bb>  
DOI: [10.1101/395947](https://doi.org/10.1101/395947)

24. **tximeta**

Rob Patro Michael Love

*Bioconductor* (2018) <https://doi.org/gfddxw>  
DOI: [10.18129/b9.bioc.tximeta](https://doi.org/10.18129/b9.bioc.tximeta)