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@a06140b on November 4, 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

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
2. Stephanie Hicks
    - Title: Assistant Professor
    - Degrees: PhD
    - Type of organization: Academic
    - Tax ID: 52-0595110 (JHU)
    - Email: shicks19@jhu.edu
3. Elana Fertig
    - Title: Associate Professor
    - Degrees: PhD
    - Type of organization: Academic
    - Tax ID: 52-0595110 (JHU)
    - Email: ejfertig@jhmi.edu
4. Casey Greene
    - Title: Assistant Professor
    - Degrees: PhD

    - Type of organization: Academic
    - Tax ID: TBD
    - Email: greenescientist@gmail.com
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 of 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)

The Human Cell Atlas (HCA) provides unprecedented characterization of the molecular states of each cell across tissues, organisms, and individuals. Computational techniques that provide the ability to rapidly query, characterize, and analyze this atlas will accelerate the pace of discovery in biomedicine. HCA data are high dimensional, but they can often be compressed into fewer dimensions without a substantial loss of information while yielding interpretable features. For transcriptomic data, compressing on the gene dimension is most attractive: it can be applied to single samples, and genes often provide information about other co-regulated genes. In the best case, the reduced dimensional space captures biological sources of variability while ignoring noise and each dimension aligns to interpretable biological processes.

Our seed network aims to create low-dimensional representations that provide search and catalog capabilities for the HCA. The benefit of these approaches will become particularly pronounced as the number of cells and tissues becomes large. Our **central hypothesis** is that these approaches will enable faster algorithms while reducing the influence of technical noise. We propose to advance **base enabling technologies** for low-dimensional representations. We also propose three aims: 1) fast and accurate search for cell, samples, and pathways; 2) a catalog of cell types and biological processes in low-dimensional spaces; and 3) educational materials to increase the impact of low-dimensional representations and the HCA in general.

The first goal of our base enabling technology work is to identify techniques that learn interpretable, biologically-aligned representations. We consider both linear and non-linear techniques. For linear techiques, we rely on our Bayesian, non-negative matrix factorization method scCoGAPS [9] (PI Fertig). This technique learns biologically relevant features across contexts and data modalities [10], including notably the HPN DREAM8 challenge [14]. We will modify the scCoGAPS uncertainty estimate to account for measurement-specific technical variation [15] in the HCA. As spatially annotated data becomes available, we will extend scCoGAPS to incorporate this. For non-linear needs, neural networks with multiple layers, provide a complementary path to low-dimensional representations [16] (PI Greene). We note that many groups are working in this area for both linear and non-linear techniques (e.g., [17]). Because of the substantial number of groups developing neural network based methods, we don't currently

plan additional efforts on this front; however, we will continue to use and rigorously evaluate these methods. We will incorporate the best performing methods into our search and catalog tools. 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.

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 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 indexible data structure for the UMI-resolution graph to enable direct access, query, and fast search.

We will implement the base enabling technologies and methods for search, analysis, and transformation into R/Bioconductor and Python frameworks. The python and R software will use common input and output formats. 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 amelioration of data quality (`scater`, `scran`, `DropletUtils`).

Aim 1

Rationale: The HCA provides a reference atlas to human cells, cell types, and the pathways that they express. Scientists will benefit most from the HCA when they can quickly identify find cells and cell types and compare references to find differences. Low-dimensional representations, because they compress the space, provide the building blocks for search approaches that can be practically applied across very large datasets such as the HCA. *We propose to develop algorithms and software for efficient search over the HCA.*

The primary approach to search in low-dimensional spaces is relatively straightforward: one must create an appropriate low-dimensional representation and identify a distance function or functions that match what biologists seek. However, the most obvious approach to search would require investigators to perform quantification on the entirety of a new sample and select cells or cell types that they wish to search for. Our goal is to enable a streaming search even before investigators

complete the quantification step. This will allow software to identify similar tissues, and in particular to identify cells that are unusual in a sample so that they can be highlighted. We will implement and evaluate techniques to learn shared low-dimensional representations between the UMI-resolution graph and quantified samples, so that samples where either component is available can be used for search **[CASEY ADD SHARED LATENT SPACE REF]**.

Akin to how one uses a reference genome to identify genomic differences between a reference and non-reference genome, we will use the framework that enables fast search to quantify differences between a reference transcriptome map (the HCA and non-reference transcriptome maps from other samples of interest. Quantifying the differences between samples characterized at the single-cell level is important because it allows us to discover population or individual level differences. One could compare ten scRNA-seq maps from individuals with a particular phenotype to the HCA reference. Our metric to quantify differences will depend on the distributions of cell expression within and between individuals, which PI Hicks has extensive experience with [25]. We will leverage common low-dimensional representations and cell-to-cell correlation structure both within and across transcriptome maps, which will often represent multiple humans. We plan to implement and evaluate linear mixed models to account for the correlation structure within and between transcriptome maps. This statistical method will be fast, memory-efficient and will scale to billions of cells because we will use low-dimensional representations.

New models for UMI deduplication accounting for transcript-level information: (Rob) *Parsimony & likelihood based, integrated with gene-level uncertainty*

Aim 2

Rationale: Biological systems are comprised of diverse cell types with overlapping molecular phenotypes, and biological processess 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 suceptiability 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 [26] 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 [27].

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.

We will also integrate catalogs of reference cell types (Aim 2). Such summaries and annotations have proven widely successful for the ENCODE, Roadmap Epigenome Mapping, and GTEx projects. We will package and version reference cell types and low-dimensional representations and deliver these as structured data objects in Bioconductor and Python. We are core package developers and power users of Bioconductor and will support on-the-fly downloading of these materials via the *AnnotationHub* framework. We will develop *F1000Research* workflows demonstrating how HCA-defined reference cell types and tools developed in this RFA can be used within a typical genomic data analysis.

Aim 3

Rationale: Low-dimensional representations for scRNA-seq and HCA data make tasks faster and provide interpretable summaries of complex high-dimensional data. The HCA data associated methods, will be valuable to many biomedical fields, but their use will require experience with this new toolkit. A scalable education effort that reaches students at and beyond undergraduate level will be needed to prepare students and maximize impact. *We propose short-course training for the HCA, single cell profiling, machine learning methods, low-dimensional representations, and tools developed by our group in response to this RFA.*

Our educational program is based on a one-week short course that we (PI Hampton) have run annually at Mount Desert Island Biological Lab over the last **X TOM FILL IN** years. The course covers R, gene expression analysis, statistical interpretation, and introduces machine learning (PI Greene). Attendees rate the course well and report that they incorporate new knowledge into their research and teaching. For this grant we will add topics centered on the HCA and increase the frequency of the course. We will run the course at locations distributed throughout the US and provide open course materials on GitHub to allow others to replicate the course. New topics will include:

Comparison of Bulk and Single-cell Assays and Data The Human Cell Atlas Project scRNA-seq: Expression Quantification and Cell Type Discovery scRNA-seq: Low-dimensional Representations scRNA-seq: Search and Analysis in Low-dimensional Representations

We aim to provide a force-multiplier for the HCA and low-dimensional methods as course attendees transmit what they learn to tens of students each year at their own institutions. We will run this course on a cost recovery model, but to maximize the multiplier effect we budget at least *ten scholarships* per offering to cover the room, board, and tuition of faculty who are primarily engaged in undergraduate instruction. This will allow faculty who will disseminate these materials in their own reaching to attend at very low cost. We will 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. Materials will include recorded videos (intended for a refresher for 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.

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 · PMID: 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

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

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 · PMID: 20810601 · PMCID: 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 · PMID: 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 · PMID: 25516281 · PMCID: 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 · PMID: 26925227 · PMCID: 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 · PMID: 28263959 · PMCID: 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

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 · PMID: 22549044 · PMCID: 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 · PMID: 27650546 · PMCID: 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 · PMID: 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 · PMID: 29792227 · PMCID: PMC5966898

14. **Inferring causal molecular networks: empirical assessment through a community-based effort**
Steven M HillLaura 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 · PMID: 26901648 · PMCID: 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 · PMID: 24223768 · PMCID: 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

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

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

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

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

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

22. **Efficient Generation of Transcriptomic Profiles by Random Composite Measurements**
Brian Cleary, Le Cong, Anthea Cheung, Eric S. Lander, Aviv Regev

Cell (2017-11) https://doi.org/gcrjhc
DOI: 10.1016/j.cell.2017.10.023 · PMID: 29153835 · PMCID: PMC5726792

23. **Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization**
Xun Zhu, Travers Ching, Xinghua Pan, Sherman M. Weissman, Lana Garmire

PeerJ (2017-01-19) https://doi.org/gfgr7c
DOI: 10.7717/peerj.2888 · PMID: 28133571 · PMCID: PMC5251935

24. **Integrative inference of brain cell similarities and differences from single-cell genomics**
Joshua Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, Evan Macosko

Cold Spring Harbor Laboratory (2018-11-02) https://doi.org/gfgr7b
DOI: 10.1101/459891

25. **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 · PMID: 26040460 · PMCID: PMC4495646

26. **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

27. **tximeta**
Rob Patro Michael Love

Bioconductor (2018) https://doi.org/gfddxw
DOI: 10.18129/b9.bioc.tximeta