Practical search and analysis with low-dimensional representations of the HCA

This manuscript (permalink) was automatically generated from greenelab/czi-seed-rfa@5955617 on November 9, 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: 23-1352685 (UPenn)
- Email: greenescientist@gmail.com

5. Tom Hampton
    - Title: Senior Bioinformatics Analyst
    - Degrees: PhD
    - Type of organization: Academic
    - Tax ID: 02-0222111 (Dartmouth)
    - Email: Thomas.H.Hampton@dartmouth.edu
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 is an Assistant Professor of Computer Science at Stony Brook University. He leads the COMBINE-lab, that develops and maintains numerous open-source genomics tools and methods. He is the primary developer of the popular transcript quantification tools Sailfish [9] and Salmon [8], having collaborated closely with Dr. Love on the latter. He and Dr. Love are actively collaborating on improved methods for transcript quantification, differential testing, and reproducible analysis via tximeta. He has recently been focused on developing improved methods for gene-level quantification from tagged-end single-cell RNA-seq data, as implemented in the alevin tool [3]. Dr. Patro will worth with co-PIs to develop improved single-cell quantification tools that can account for gene-ambiguous reads and provide uncertainty information about the quantification estimates — which is important for accurate and robust creation of reduced-dimensionality latent spaces. He will also work with the co-PIs to develop efficient algorithms and data structures to enable efficient expression search over low-dimensional representations of HCA data.

Proposal Body (2000 words)

The Human Cell Atlas (HCA) provides unprecedented characterization of the molecular phenotypes 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 or cellular attributes. In the best case, the reduced dimensional space captures biological sources of variability while ignoring noise and each dimension aligns to interpretable biological processes.

Scientific Goals

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.

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 techniques, we rely on our Bayesian, non-negative matrix factorization method scCoGAPS [10](PIs Fertig & Goff). This technique learns biologically relevant features across contexts and data modalities [11], including notably the HPN DREAM8 challenge [15]. This technique is specifically selected as a base enabling technnology because its error distribution can naturally account for measurement-specific technical variation [16] and its prior distributions for different feature quantifications or spatial information. For non-linear needs, neural networks with multiple layers, provide a complementary path to low-dimensional representations [17] (PI Greene) that model these diverse features of HCA data. We note that many groups are working in this area for both linear and non-linear techniques (e.g., [18]). Because of the substantial number of groups developing neural network based methods, we do not currently plan additional efforts on methods development beyond scCoGAPS. However, we will continue to use and rigorously evaluate these methods and incorporate the best performing methods into our search and catalog tools. We will extend transfer learning methods, including ProjectR [2] (PIs Goff & Fertig) to enable rapid integration, interpretation, and annotation of learned latent spaces. The latent space team from the HCA collaborative networks RFA (including PIs Fertig, Goff, Greene, and Patro) is establishing common definitions and requirements for latent spaces for the HCA, as well as standardized output formats for low-dimensional representations from distinct classes of methods.

The *second goal of our base enabling technology work* 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 to account for reads mapping between multiple genes in the resulting quantification estimates. This affects approximately 15-25% of the reads generated in a typical experiment, 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 Unique Molecular Identifier (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 prior to secondary analysis.

We will implement the base enabling technologies and methods for search, analysis, and latent space transformations into R/Bioconductor. We will additionally develop platform-agnostic input and output formats for latent space representations of the HCA data to maximize interoperability. The software tools produced will be fast, scalable, and memory-efficient because we will leverage the computational tools previously developed by Bioconductor for single-cell data access to the HCA, data representation (`SingleCellExperiment`, `beachmat`, `LinearEmbeddingMatrix`, `DelayedArray`, `HDF5Array` and `rhdf5`) and data assessment and amelioration of data quality (`scater`, `scran`, `DropletUtils`).

Tools and Resources

By using and extending our base enabling technologies we propose to develop three principle tools and resources for the HCA. These include 1) software to enable fast and accurate search and annotation using low-dimensional representations of cellular features, 2) a versioned and annotated catalog of latent spaces corresponding to signatures of cell types, states, and biological attributes across the the HCA, and 3) educational materials to increase the use and impact of low-dimensional representations and the HCA in general.

Aim 1

Rationale: The HCA provides a reference atlas to human cell types, states, and the biological processes they engage. Scientists will benefit most from the HCA when they can quickly identify cell types and states 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 using low-dimensional representations.*

The primary approach to search in low-dimensional spaces is straightforward: one must create an appropriate low-dimensional representation and identify a distance function or functions that match what biologists seek. Using the low-dimensional representation improves speed and can also reduce noise. We will evaluate representations for their ability to support search and implement the best performing approaches. Current approaches require investigators to perform gene-level quantification on the entirety of a new sample. We aim to enable search during sample preprocessing, prior to gene-level quantification. This will enable in-line annotation of cell types and states and identification of novel features as samples are being processed. We will implement and evaluate techniques to learn and transfer 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 and annotation **[CASEY ADD SHARED LATENT SPACE REF]**. These UMI-graphs will be embedded in the prior of scCoGAPS and architecture of non-linear latent space

techniques. **[Do we need this line? It's a bit more specific than the rest of the paragraph - LAG]**

Similarly to the approach by which comparisons to a reference genomes can identify specific differences in a genome of interest, we will use low-dimensional representations from latent spaces to define a reference transcriptome map (the HCA) and use this to quantify differences in target transcriptome maps from new samples of interest. We will leverage common low-dimensional representations and cell-to-cell correlation structure both within and across transcriptome maps from Aim 2 to define this reference. Quantifying the differences between samples characterized at the single-cell level reveals population or individual level differences. **[<– I'm not sure what this sentence means. Please clarify. - LAG]** Comparison of scRNA-seq maps from individuals with a particular phenotype to the HCA reference that is computationally infeasible from the large scale of HCA data becomes tractable in these low dimensional spaces. We (PI Hicks) have extensive experience dealing with the distributions of cell expression within and between individuals [26], which will be critical for defining an appropriate metric to compare references in latent spaces. 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 be scalable to billions of cells using low-dimensional representations.

Aim 2

Rationale: Biological systems are comprised of diverse cell types and states with overlapping molecular phenotypes. Furthermore, biological processes are often reused with modifications across cell types. Low-dimensional representations can identify these shared features, independent of total distance between cells in gene expression space, across large collections of data including the HCA. We will evaluate and select methods that define latent spaces that reflect discrete biological processes or cellular features. These latent spaces can be shared across different biological systems and can reveal context-specific divergence such as pathogenic differences in disease. *We propose to establish a central catalog of cell types, states, and biological processes derived from low-dimensional representations of the HCA.*

By establishing a catalog of cellular features using low-dimensional representations can reduce noise and aid in biological interpretability. However, there are currently no standardized, quantitative metrics to determine the extent to which low-dimensional representations capture generalizable biolobical features. 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] (PIs Goff & Fertig). These provide a strong foundation to compare low-dimensional representations across different low-dimensional data representation technniques. Generalizable representations should transfer across datasets of related biological contexts, while representations of noise will not. In addition, we have found that combining multiple representations can better capture biological processes across scales [28], and that representations across scales capture distinct, valid biological signatures [16]. Therefore, we will

establish a versioned catalog consisting of low-dimensional features learned across both linear and non-linear methods from our base enabling technologies and proposed extensions in Aim 1.

We will package and version low-dimensional representations and annotate these representations based on their corresponding celluar features (e.g. cell type, tissue, biological process) and deliver these as structured data objects in Bioconductor as well as platform-agnostic data formats. Such summaries and annotations have proven widely successful for the ENCODE, Roadmap Epigenome Mapping, and GTEx projects. We are core package developers and power users of Bioconductor (PIs Hicks and Love) and will support on-the-fly downloading of these materials via the *AnnotationHub* framework. 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. We (PI Love) developed hash-based versioning and provenance identification and detection framework for bulk RNA-seq that supports reproducible computational analyses and has proven to be successful [29]. This will help to avoid scenarios where researchers report on matches to a certain cell type in HCA without precisely defining which definition of that cell type. 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. This catalogue will be used as the basis of defining the references for cell type and state, or individual-specific differences with the linear models proposed in Aim 1.

Aim 3

Rationale: Low-dimensional representations for scRNA-seq and HCA data make tasks faster and provide interpretable summaries of complex high-dimensional cellular features. The HCA data associated methods and workflows will be valuable to many biomedical researchers, but their use will require experience with this new toolkit. Furthermore, researchers will need exposure to the conceptual basis of low-dimensional interpretations of biological systems. To address these issues, we propose a scalable education effort that reaches students at and beyond undergraduate level enable faster adoption and interpretation of the HCA, and to maximize its 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. Additionally, we have previously developed didactic course material on single cell RNA-Seq analysis for the annual McKusick Short Course on Human and Mammalian Genetics (PI Goff) at Jackson Labs. For this grant we will extend these educational opportunities by developing topics and materials centered on the HCA and interpretation of low-dimensional

latent spaces. We (PI Hampton) 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 Annotation
- 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. **Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms**
Rob Patro, Stephen M Mount, Carl Kingsford

Nature Biotechnology (2014-04-20) https://doi.org/gfghc2
DOI: 10.1038/nbt.2862 · PMID: 24752080 · PMCID: PMC4077321

10. **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

11. **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

12. **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

13. **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

14. **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

15. **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

16. **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

17. **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

18. **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

19. **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

20. **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

21. **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

22. **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

23. **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

24. **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

25. **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

26. **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

27. **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

28. **Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks**
Jie Tan, Georgia Doing, Kimberley A. Lewis, Courtney E. Price, Kathleen M. Chen, Kyle C. Cady, Barret Perchuk, Michael T. Laub, Deborah A. Hogan, Casey S. Greene

Cell Systems (2017-07) https://doi.org/gcw9f4
DOI: 10.1016/j.cels.2017.06.003 · PMID: 28711280 · PMCID: PMC5532071

29. **tximeta**
Rob Patro Michael Love

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