

Abstract

The HCA provides a reference atlas to human cell types, states, and the biological processes in which they engage. The utility of the reference therefore requires that one can easily compare references to each other, or a new sample to the compendium of reference samples. Because they compress the space, low-dimensional representations provide the building blocks for search approaches that can be practically applied across very large datasets such as the HCA. Our seed network proposes to compress HCA data into fewer dimensions that preserve the important attributes of the original high dimensional data and yield interpretable, searchable features. We hypothesize that using latent space methods to identify low-dimensional representations of HCA data will accurately capture biological sources of variability and will be robust to measurement noise. We propose techniques that learn interpretable, biologically-aligned representations, improve techniques for fast and accurate quantification, and implement these base-enabling technologies and methods for search, analysis, and latent space transformations as freely available, open source software tools. By using and extending our base enabling technologies, we will provide three principle tools and resources for the HCA: 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) short course and educational materials that will increase the use and impact of low-dimensional representations and the HCA in general.

Research Proposal

The Human Cell Atlas (HCA) provides unprecedented characterization of molecular phenotypes across individuals, tissues and disease states – resolving differences to the level of single cells. These data provide an extraordinary opportunity for scientific advancement, enabled by new tools to rapidly query, characterize, and analyze these intrinsically high-dimensional data. To facilitate this, our seed network proposes to compress HCA data into fewer dimensions that preserve the important attributes of the original data and yield interpretable, searchable features. We hypothesize that using latent space methods to identify low-dimensional representations of HCA data will accurately capture biological variation and will be robust to measurement noise. Our network incorporates biologists, computer scientists, statisticians, and data scientists from five leading institutions who will work together to create foundational technologies and educational opportunities that promote effective interpretation of low-dimensional representations of HCA data. We will continue our active collaborations with other members of the broader HCA network to integrate state of the art tools, portals, and annotations to enable broader and more efficient utilization of the HCA.

Scientific Goals

We will create low-dimensional representations that provide search and catalog capabilities for the HCA. Given both the scale of data, and the inherent complexity of biological systems, we believe these approaches are crucial to the long term success of the HCA. 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.

First, we will identify techniques that learn interpretable, biologically-aligned representations. We will consider both linear and non-linear techniques as each may identify distinct components of biological systems. For linear techniques, we rely on our Bayesian, non-negative matrix factorization method scCoGAPS [15] (PIs Fertig & Goff) which learns biologically relevant features across contexts and data modalities [16], including notably the HPN DREAM8 challenge [20]. This technique is selected because its error distribution can naturally account for measurement-specific technical variation [21]. For non-linear needs, neural networks with multiple layers provide a complementary path to low-dimensional representations [8] (PI Greene) that model these diverse features of HCA data. We will make use of the substantial progress in both linear and non-linear techniques (e.g., [22]) and rigorously evaluate emerging methods into our search and catalog tools. We will extend transfer learning methods, including ProjectR [1] (PIs Goff & Fertig) to enable rapid integration 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 formats for low-dimensional representations.

Second, we will improve techniques for fast and accurate quantification. Existing approaches for scRNA-seq data using tagged-end protocols do not account for multi-mapping reads. This affects approximately 15-25% of the reads in a typical experiment, reducing quantification accuracy, and leading to systematic biases [14]. To address this, we will build on our recently developed quantification method for tagged-end data that accounts for reads mapping to multiple genomic loci in a principled way [14] (PI Patro), and extend this into a production quality tool for scRNA-seq preprocessing. Our tool will support: 1. Exploration of alternative models for Unique Molecular Identifier (UMI) resolution. 2. Development of new approaches for quality control and filtering using the UMI-resolution graph. 3. Creation of a compressed and indexable data structure for the UMI-resolution graph to enable direct access, query, and search prior to secondary analysis.

We will implement these base-enabling technologies and methods for search, analysis, and latent space transformations as freely available, open source software tools. We will additionally develop platform-agnostic input and output data formats and standards for latent space representations of the

HCA data to maximize interoperability. The software tools produced will be fast, scalable, and memory-efficient by leveraging the available assets and expertise of the R/Bioconductor project (PIs Hicks & Love) as well as the broader HCA community.

We will provide three principal 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) short course and educational materials that will 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 in which they engage. The utility of the reference therefore requires that one can easily compare references to each other, or a new sample to the compendium of reference samples. 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 distance functions that enable biologically meaningful comparisons. Ideal low-dimensional representations are predicted to be much faster to search, and potentially more biologically relevant, as noise can be removed. In this aim, we will evaluate novel, low-dimensional representations to identify those with optimal qualities of compression, noise reduction, and retention of biologically meaningful features. Current scRNA-seq approaches require investigators to perform gene-level quantification on the entirety of a new sample. We aim to 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 raw or lightly processed data (e.g., kmer representations or UMI-graphs) and quantified samples, so that samples where either quantified or raw data are available can be used for search and annotation [30].

Similar to the approach by which comparisons to a reference genome can identify differences in a genome of interest, we will use low-dimensional representations to define a reference transcriptome map of the HCA and use this to quantify differences in target transcriptome maps from new samples of interest. We will leverage shared low-dimensional features, cell-to-cell correlation structure, and transfer learning methodologies to define this reference. Quantifying differences between samples characterized at the single-cell level reveals population or individual-level differences. Comparison of scRNA-seq maps from individuals with a particular phenotype to the HCA reference, which 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 [31], which will be critical for defining an appropriate metric to compare latent space references. We will 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 that are often reused with modifications across cell types. Low-dimensional representations can identify these shared features across large datasets. We will evaluate and select methods that define latent spaces that reflect discrete biological processes or cellular features. *We propose to*

establish a versioned catalog of cell types, states, and biological processes derived from low-dimensional representations of the HCA.

Establishing a reference catalog of cellular features using low-dimensional representations will facilitate wider adoption of the HCA. However, there are currently no standardized, quantitative metrics to determine the extent to which low-dimensional representations capture generalizable biological features. We have developed transfer learning methods to quantify the shared use of latent spaces across datasets [1] (PIs Greene, Goff & Fertig). This provides a foundation to compare different low-dimensional representations through cross-validation by learning representations in source datasets and testing their ability to transfer into a target dataset. Generalizable representations will be robust in cross-study validation, transferring across datasets of related biological contexts while representations of noise will not. In addition, we have found that combining multiple representations better captures biological processes across scales [6], and that representations across scales capture distinct, valid biological signatures [21]. By establishing a reference set of latent spaces for the HCA, we can provide a uniform resource to represent diverse cellular types, states, and relevant cellular attributes.

We will package and version low-dimensional representations of the HCA and annotate these via their corresponding cellular features. We will deliver these as structured data objects in Bioconductor as well as platform-agnostic data formats. Where applicable, we will leverage the tools developed by Bioconductor for single-cell data access to the HCA, data representation (SingleCellExperiment, beachmat, LinearEmbeddingMatrix, DelayedArray, HDF5Array and rhdf5) and data assessment and quality amelioration (scater, scran, DropletUtils). We are core package developers for 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 catalog by the gene weights and transcript sequences using a hash function. We (PIs Love and Patro) previously developed a successful hash-based versioning and provenance detection framework that supports reproducible computational analyses [13]. Our versioning and dissemination of reference latent space catalogs will help to avoid scenarios where researchers report on matches to a certain feature in HCA without precisely defining which representation of that feature. We will develop *F1000Research* workflows demonstrating how these low-dimensional representations and tools can be used for genomic data analysis. This catalog 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 of scRNA-seq data make tasks faster and provides interpretable summaries of complex, high-dimensional cellular features. The HCA-data-associated methods and workflows will be valuable to many biomedical fields, but their use will require an understanding of basic bioinformatics, scRNA-seq, and how the tools being developed work. Furthermore, researchers will need exposure to the conceptual basis of low-dimensional interpretations of biological systems. This aim addresses these needs in three ways.

First, we will develop a training program for biologists at all levels, including those with no experience in bioinformatics. Lecture materials will be extended from existing materials from previous bioinformatic courses we (PI Hampton) have run since 2009. These courses have trained over 400 scientists and achieved approval ratings of over 90%. The success of these learning experiences is related to our instructional paradigm which includes a very challenging course project coupled with one-on-one support from instructors. We will develop a new curriculum specifically tailored to HCA that incorporates: 1) didactic course material on single cell gene expression profiling (PI Goff), 2) machine learning methods (PI Greene), 4) statistics for genomics (PIs Fertig and Hicks), 4) search and analysis in low-dimensional representations, and 5) tools developed by our group in response to this RFA.

Second, the short course will train not only students, but also instructors. We will recruit former participants of this class to return as teaching assistants or module presenters. We have found that course alumni are an invaluable resource in understanding how to improve the course over time. Part of our strategy is to support this community, which includes many people who will drive the next wave of innovation. All of our course materials will be freely available and open source. A capstone session will be included in which we will provide suggestions about how the materials presented in the course can be incorporated into existing course curricula. Course faculty will be available to assist with integration efforts after the course. Finally, the short course will facilitate scientific collaborations by engaging participants in utilizing these tools for collaborative research efforts.

The objectives proposed here will facilitate broader adoption of the HCA and enable rapid discovery and examination of biological features by providing an invaluable index to this cellular atlas.

Figures

Base Enabling

- Aim 1** Low-dimensional
- Aim 2** ● representation of a
- Aim 3** specific cellular feature

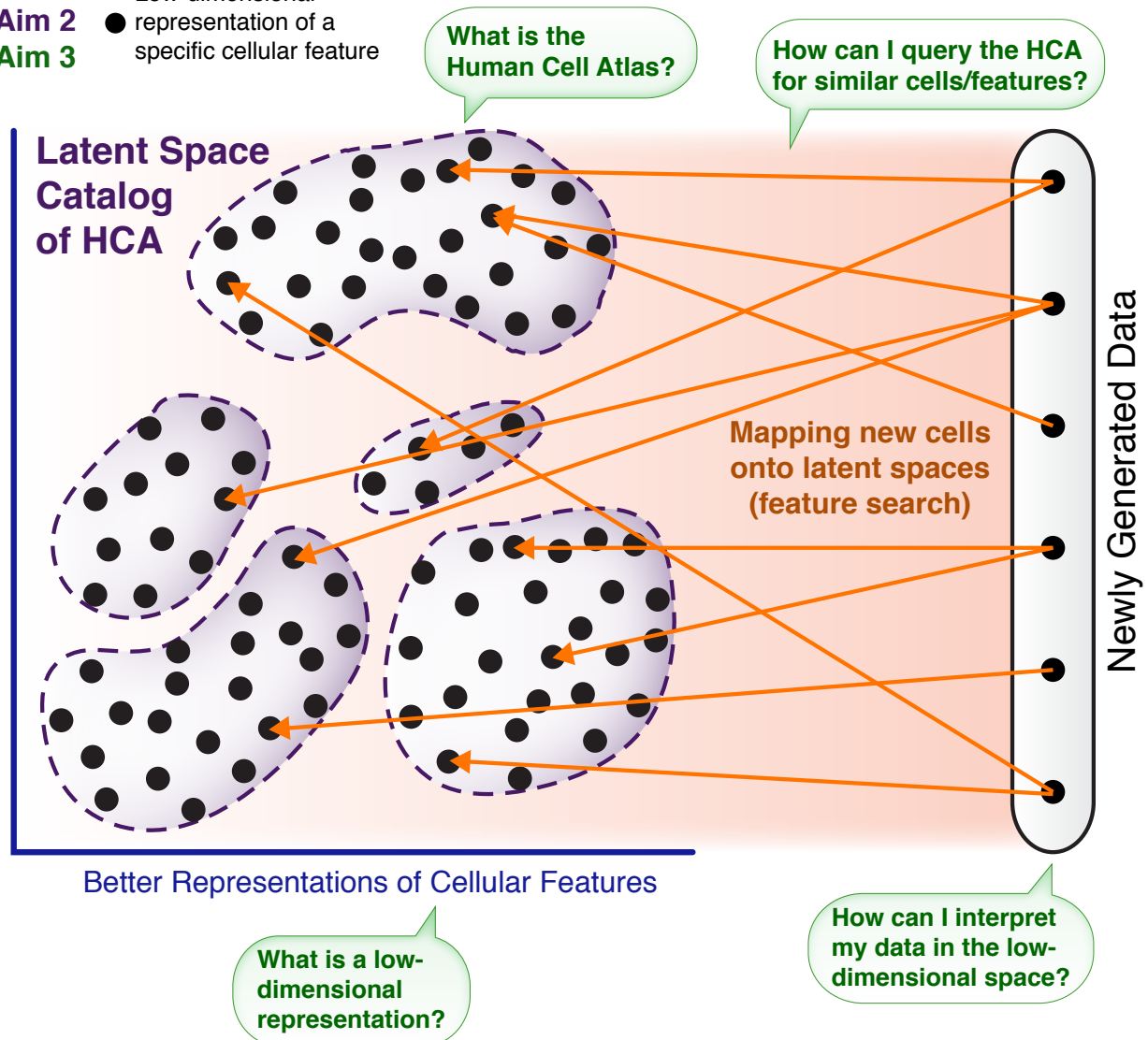


Figure 1: Our seed network will develop compressed and interpretable low-dimensional representations of the Human Cell Atlas data. We will evaluate and annotate efficient feature representations at various levels of data abstraction including transcriptome-wide maps, latent space representations of specific cell types, states, and biological processes, as well as kmers and UMI-graph representations of raw or lightly processed single cell data to enable rapid search of biological features (Aim 1, orange). Processed large-scale HCA datasets will be analyzed with linear and non-linear latent space techniques (Aim 2, purple) and annotated against known and learned metadata in the HCA. Cross-validation on features learned from each technique will be performed with transfer learning to assess the biological robustness of each low dimensional feature. Features which represent robust biological processes across datasets in cross validation will be stored in a versioned catalog, which will contain an ensemble of low dimensional representations across datasets. This catalog will provide a set of gene weights to query for related biologically processes in disparate datasets. Finally, we will enable utilization and interpretation of this catalog through educational modules designed to train users in dimensionality reduction techniques, the latent space catalog, and interpretation of low dimensional spaces (Aim 3, green).

References

1. 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, Loyal A. Goff*, Elana J. Fertig*
Cold Spring Harbor Laboratory (2018-08-20) <https://doi.org/gd2xpn>

DOI: [10.1101/395004](https://doi.org/10.1101/395004) *Co-corresponding Authors

2. 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/) · PMCID: [PMC6215955](https://pubmed.ncbi.nlm.nih.gov/PMC6215955/)

3. 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/)

4. 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/)

5. ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions

Jie Tan, John H. Hammond, Deborah A. Hogan, Casey S. Greene

mSystems (2016-01-19) <https://doi.org/gcgmbq>

DOI: [10.1128/msystems.00025-15](https://doi.org/10.1128/msystems.00025-15) · PMID: [27822512](https://pubmed.ncbi.nlm.nih.gov/27822512/) · PMCID: [PMC5069748](https://pubmed.ncbi.nlm.nih.gov/PMC5069748/)

6. 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](https://doi.org/10.1016/j.cels.2017.06.003) · PMID: [28711280](https://pubmed.ncbi.nlm.nih.gov/28711280/) · PMCID: [PMC5532071](https://pubmed.ncbi.nlm.nih.gov/PMC5532071/)

7. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders.

Gregory P Way, Casey S Greene

Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (2018)

<https://www.ncbi.nlm.nih.gov/pubmed/29218871>

PMID: [29218871](https://pubmed.ncbi.nlm.nih.gov/29218871/) · PMCID: [PMC5728678](https://pubmed.ncbi.nlm.nih.gov/PMC5728678/)

8. 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)

9. 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/)

10. 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/)

11. 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/)

12. 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](https://doi.org/10.1038/nbt.2862) · PMID: [24752080](https://pubmed.ncbi.nlm.nih.gov/24752080/) · PMCID: [PMC4077321](https://pubmed.ncbi.nlm.nih.gov/PMC4077321/)

13. 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)

14. 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)

15. 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, Loyal A. Goff*, Seth Blackshaw*

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

DOI: [10.1101/378950](https://doi.org/10.1101/378950) *Co-corresponding Authors

16. 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/)

17. 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/)

18. 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/)

19. 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/)

20. 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/)

21. 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/)

22. 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)

23. 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)

24. 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)

25. 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)

26. 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)

27. 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](https://doi.org/10.1016/j.cell.2017.10.023) · PMID: [29153835](https://pubmed.ncbi.nlm.nih.gov/29153835/) · PMCID: [PMC5726792](https://pubmed.ncbi.nlm.nih.gov/PMC5726792/)

28. 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](https://doi.org/10.7717/peerj.2888) · PMID: [28133571](https://pubmed.ncbi.nlm.nih.gov/28133571/) · PMCID: [PMC5251935](https://pubmed.ncbi.nlm.nih.gov/PMC5251935/)

29. 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](https://doi.org/10.1101/459891)

30. greenelab/shared-latent-space

chrsunwil

GitHub <https://github.com/greenelab/shared-latent-space>

31. 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/)

32. 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)