SINGLE-CELL ANALYSIS AND MODELLING OF CELL POPULATION HETEROGENEITY

NIKOLAY SAMUSIK

Department of Microbiology & Immunology
Stanford Medical School
Stanford 94305 CA, USA
Email: samusik@stanford.edu

NIMA AGHAEEPOUR

Department of Anesthesiology
Stanford Medical School
Stanford 94305 CA, USA
Email: naghaeep@stanford.edu

SEAN BENDALL

Department of Pathology
Stanford Medical School
Stanford 94305 CA, USA
Email: bendall@stanford.edu

Recent technological developments allow gathering single-cell measurements across different domains (genomic, transcriptomics, proteomics, imaging etc). Sophisticated computational algorithms are required in order to harness the power of single-cell data. This session is dedicated to computational methods for single-cell analysis in various biological domains, modelling of population heterogeneity, as well as translational applications of single cell data.

1. Introduction

Inferring the molecular mechanism of cell behavior and linking it to function and dysfunction is one of the ultimate goals of quantitative biology and medicine. Until recently, most measures to classify and characterize cellular behavior have been performed on the 'bulk samples', whereby a large number of cells were physically homogenized and then assayed. Bulk measurements erase the information about the potentially complex heterogeneity of cellular states within the samples. The problem with such approaches becomes obvious from a simple example: whenever researchers observe a difference in average values of a single parameter between samples, it is quite impossible to differentiate between a scenario where there was a homogenous change of a variable in all cells versus a shift in compositional ratios between a differentially expressing populations Besides, the measurements derived from pooled populations of cells lack the specificity to capture outlier cell behavior that might explain cell differentiation and transitions from normal to disease cellular states. The noise, or variance, between the molecular states of different cells -- even among cells assumed to be homogenous – can be correlated with protein expression and function [1] as well as cell morphology and interaction with neighbors[2]. Emergence of cell heterogeneity might be sporadic (e.g., cell-to-cell variation in cell culture[3]), programmed (e.g., cell differentiation[4] or immune receptor recombination[5]), or a result of adaptive evolution and semi-heritable phenotypic plasticity[6].

The ability to quantify molecular events with single cell resolution is intrinsically linked to analytical advances. Unfortunately, many of those variations could not be systematically studied by traditional molecular biology methods, such as PCR, Western Blotting, IP, genome sequencing, microarrays and RNA-seq, because they lack the sensitivity and the throughput that are required for single cell analysis. One notable exception is immunology, which has enormously benefitted from early adoption of the single-cell analysis by flow cytometry and FACS. Flow cytometry has been pivotal to detailed characterization of various immunological processes, such as blood cell development and activation and has enabled systematic mapping of the roles of various immune cell populations in healthy and disease states. Driven by a need to distinguish multiple cell populations, cytometry placed emphasis on multiparametric analysis whereby the cell populations were defined by increasingly complex combinations of protein markers. More recently, the importance of multiparametric analysis has increased with advent of mass cytometry[7]. Many excellent computational tools have been developed for handling cytometry data, including specialized clustering algorithms for automated mapping of cell population[8], machine learning tools that find cell populations that are correlated to clinical outcome[9], data visualization tools that trace cell differentiation trajectories[10,11], a specialized ontology of cell types[12], algorithms for causal inference of signaling networks by leveraging huge training sets of single-cell data [13], data-driven reference maps of immune cell populations[14] and many others.

For many years the single-cell analysis has been associated with flow cytometry and was limited to measuring protein concentrations using tagged antibodies. Recent advances in experimental techniques and automation have greatly expanded the scope of single-cell analysis and introduced completely novel readouts and modalities. Examples include:

1. Genomic sequencing in single cells [15]
2. Single cell RNA-seq [16]
3. Single molecule RNA sequencing in situ [17]
4. Gene expression profiling by flow cytometry [18] [19]
5. Histo-cytometry [20]
6. Multiplexed ion beam imaging [21]
7. Mapping of chromatin state in single cells [22]
8. Cell morphology and motility analysis in cell cultures [2]
9. Single cell western blotting[23]

These emerging technologies provide an unprecedented opportunity to capture new biological processes and mechanisms at the single cell level. Given the list of analytical methods with a single cell resolving power now available, a wealth of new information, including: protein abundance, methylation patterns, promoter structure, gene expression, copy number variations, gene function and essentiality, DNA structure, evolutionary plasticity, and selective advantage can now be created for integration. Synthesis and interpretation of various modalities of single cell-level data now depends on novel computational approaches that aim to uncover and model the biological principles behind the cell heterogeneity. Data fusion methods that leverage prior biological knowledge for automated cell type annotation. Most importantly, computational methods are needed to provide a system-level view of the interplay of diverse, fluctuating biological components and identify clinically relevant and actionable modules within the biological system. In this session we feature excellent pieces of original research that broadly cover various aspects of single-cell analysis and modelling of cellular heterogeneity.

2. Session contributions

2.1. *Data normalization and quality control*

Quality control is a cornerstone of quantitative data analysis: rigorous filtering of noisy and spurious signals and correction of systematic variability is lays the solid foundation which ensures that the downstream data analysis captures true biological effects.

Aevermann et al. present a quality control pipeline for single-cell analysis which pioneers the use of objective criteria and machine learning for QC of single-nuclei sequencing data. While many researchers today still rely on subjective assessment of data quality, Aevermann and colleagues designed and trained a classifier that implements a random-forest approach with 79 features per sample to stratify samples into 3 quality classes: 1 pass and 2 types of fails. Analysis of 2272 single-nuclei samples successfully screened out 21% low quality data points. Authors demonstrated that removing the low-quality samples had a marked effect on the quality of the results in the downstream multidimensional manifold embedding analysis.

Fread et al. devised an elegant advance for the quality control and filtering of barcoded mass cytometry (CyTOF) data. They are introducing a concept of per-sample filtering of data following the debarcoding, which allows for proper handling of potentially very significant sample-to-sample variations in barcode intensity. Authors are also pioneering the idea of combining multiple cellular features into semi-artificial filtering parameters and writing them into the FCS files, which gives the human analyst an opportunity to set filtering gates using gating software and adjust the positioning of such gates on as sample-by-sample basis, dynamically monitoring the data quality based on biaxial scatterplots for other parameters. This simple yet elegant improvement dramatically streamlines the process of filtering spurious single-cell events and their publicly available software can be expected to be of a great utility to the CyTOF community.

2.2. *Manifold embedding and tracing with single-cell datasets*

One of the most exciting opportunities in the age of single-cell data is the ability to map the complex processes of cell differentiation by tracing the manifold shapes of single-cell distributions and discovering the local trajectories of cell changes in the marker space. This analysis is complicated by the unpredictable nature of manifolds in the data, high dimensionality of feature space and the instability of the local covariance matrix.

Cordero et al. introduce an approach for linear trajectory tracing in single cell RNA-seq data called SCIMITAR that implements morphing Gaussian model and performs simultaneous estimation of the mean expression levels along the trajectory and the local covariance matrix. The authors introduce a new statistical test to select relevant genes based on correlation of gene expression to the trajectory. They convincingly demonstrate that this test is more sensitive and specific that a conventional group-based comparison, picking up more biologically significant genes than the ANOVA-based statistical test in the original paper[24]. While the SCIMITAR algorithm is currently limited by the assumption of a single curvilinear trajectory, the authors anticipate further extension of this approach that would allow capturing more complex manifolds.

Kim et al. present a new scalable algorithm for fast embedding of multidimensional data based on LargeVis algorithm[25]. Unlike most embedding methods, the algorithm works in linear time, which it very useful given the ever-growing datasets. Authors validate the algorithm on CyTOF data from mouse bone marrow and show that the quality of embedding is superior to the slower tSNE algorithm that is currently popular in the single-cell analysis community.

2.3. *Cross-species alignment of single-cell expression patterns*

Traditionally, comparative cytology and histology relied on qualitative descriptions of tissue architectures and cell functions across different organisms. The availability of single-cell data opens a possibility to quantitatively align differentiation trajectories and cell types between species based on their expression profiles and other quantitative functional features. Such mapping could help us understand better the development and evolution of multicellular organisms and also facilitate the transfer of pre-clinical results from model organisms to human.

Johnsons et al. harness the single-cell RNA-seq data from neural precursors in human and mouse for building the cross-species map of neural cell populations. They take a two-step approach, which starts with defining the list of genes which show concordant expression patterns across major neuronal precursor populations in both species. In the second step, the authors co-cluster neuronal cell distributions of the two species based on the concordant gene subset, thus constructing a cross-species map of cell populations. Despite the lack of a perfect overlap, which is expected due to systematic differences in cell distributions between species, the authors show that the obtained cross-species map can be utilized for transferring the functional annotations of cells subsets between the corresponding population of the two species.

2.4. *Modelling of cell heterogeneity in cancer*

While single-cell readouts provide excellent snapshots of population heterogeneity, creating comprehensive mathematical models of cell interactions, somatic transdifferentiation and clonal evolution is key to attaining detailed understanding of dynamic processes that underpin the population heterogeneity in cancer. By identifying the causal chains of events and iterating through various scenarios, mathematical models of cancer cell populations can yield clinically actionable predictions and assist in optimizing treatment strategies.

Kanigel Winner and Costello present a novel modeling technique to model the treatment regimens for people with metastatic bladder cancer. This form of cancer metastasized to the lung has not been previously modeled and hence is an important and realistic problem since overall survival for this disease has not improved in the past three decades. The authors created a computational model to simulate tumor environment by carefully incorporating quantitative data about cell division rates, in vivo drug concentrations, in vitro IC50 curves for cancer cell lines and vascularization patterns of tumor microenvironment. This model was used to analyze different chemotherapeutic regimens much faster than getting in-vivo data. Authors strikingly demonstrated that the standard-of-care chemotherapeutic regimen that alternates gemcitabine and cisplatin inevitably leads to quick emergence of resistant clones, which goes in line with the abysmal 5-year survival rate (6.8%) for this type of cancer following the aforementioned treatment. Authors also found that any conceivable regimen combining the two drugs will eventually lead to resistance because of randomly surviving cancer cell clones. Key factors that contribute to this resistance is the inhomogeneity of drug distribution in the tissue and the ‘dilution effect’ whereby rapidly dividing cells effectively drop the drug concentration by splitting it between daughter cells. With further refinement, this model could help design novel therapeutic regimens that would hopefully lead to disease eradication.

3. Acknowledgements

We thank all of the authors who submitted papers for this session and all of the reviewers who contributed their time and expertise. We acknowledge the NIH grant R01GM109836 and the Rachford and Carlota A. Harris Endowed Chair for support. We are grateful to the PSB organizers for their support and especially Tiffany Murray for meeting coordination.

4. References

1. Newman, J. R. S. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441,** 840–6 (2006).

2. Battich, N., Stoeger, T. & Pelkmans, L. Control of Transcript Variability in Single Mammalian Cells. *Cell* **163,** 1596–1610 (2015).

3. Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M. & Sorger, P. K. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* **459,** 428–32 (2009).

4. Süel, G. M., Kulkarni, R. P., Dworkin, J., Garcia-Ojalvo, J. & Elowitz, M. B. Tunability and noise dependence in differentiation dynamics. *Science* **315,** 1716–9 (2007).

5. Proudhon, C., Hao, B., Raviram, R., Chaumeil, J. & Skok, J. A. Long-Range Regulation of V(D)J Recombination. *Adv. Immunol.* **128,** 123–82 (2015).

6. Quaranta, V. *et al.* Trait variability of cancer cells quantified by high-content automated microscopy of single cells. *Methods Enzymol.* **467,** 23–57 (2009).

7. Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332,** 687–96 (2011).

8. Aghaeepour, N. *et al.* Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10,** 228–38 (2013).

9. Bruggner, R. V, Bodenmiller, B., Dill, D. L., Tibshirani, R. J. & Nolan, G. P. Automated identification of stratifying signatures in cellular subpopulations. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E2770-7 (2014).

10. Bendall, S. C. *et al.* Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell* **157,** 714–725 (2014).

11. Zunder, E. R., Lujan, E., Goltsev, Y., Wernig, M. & Nolan, G. P. A Continuous Molecular Roadmap to iPSC Reprogramming through Progression Analysis of Single-Cell Mass Cytometry. *Cell Stem Cell* **16,** 323–337 (2015).

12. Meehan, T. F. *et al.* Logical Development of the Cell Ontology. *BMC Bioinformatics* **12,** 6 (2011).

13. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. & Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308,** 523–9 (2005).

14. Spitzer, M. H. *et al.* An interactive reference framework for modeling a dynamic immune system. *Science (80-.).* **349,** 1259425–1259425 (2015).

15. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–4 (2011).

16. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343,** 776–9 (2014).

17. Lee, J. H. *et al.* Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10,** 442–58 (2015).

18. Lyubimova, A. *et al.* Single-molecule mRNA detection and counting in mammalian tissue. *Nat. Protoc.* **8,** 1743–58 (2013).

19. Frei, A. P. *et al.* Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods* (2016). doi:10.1038/nmeth.3742

20. Gerner, M. Y., Kastenmuller, W., Ifrim, I., Kabat, J. & Germain, R. N. Histo-cytometry: a method for highly multiplex quantitative tissue imaging analysis applied to dendritic cell subset microanatomy in lymph nodes. *Immunity* **37,** 364–76 (2012).

21. Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20,** 436–442 (2014).

22. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490 (2015).

23. Hughes, A. J. *et al.* Single-cell western blotting. *Nat. Methods* **11,** 749–55 (2014).

24. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 7285–90 (2015).

25. Tang, J., Liu, J., Zhang, M. & Mei, Q. Visualizing Large-scale and High-dimensional Data. *Proc. 25th Int. Conf. World Wide Web* 287–297 (2016). doi:10.1145/2872427.2883041

PRODUCTION OF A PRELIMINARY QUALITY CONTROL PIPELINE FOR SINGLE NUCLEI RNA-SEQ AND ITS APPLICATION IN THE ANALYSIS OF CELL TYPE DIVERSITY OF POST-MORTEM HUMAN BRAIN NEOCORTEX*

BRIAN AEVERMANN[1#], JAMISON MCCORRISON[1#], PRATAP VENEPALLY[1#], REBECCA HODGE[2], TRYGVE BAKKEN[2], JEREMY MILLER[2], MARK NOVOTNY[1], DANNY N. TRAN[1], FRANCISCO DIEZ-FUERTES[1,3], LENA CHRISTIANSEN[4], FAN ZHANG[4], FRANK STEEMERS[4], ROGER S. LASKEN[1], ED LEIN[2], NICHOLAS SCHORK[1], RICHARD H. SCHEUERMANN[1,5,6 †]

[1]J. Craig Venter Institute, 4120 Capricorn Lane, La Jolla, CA 92037, USA, [2]Allen Institute for Brain Science, 615 Westlake Avenue North, Seattle, WA 98103, USA, [3]Centro Nacional de Microbiología, Instituto de Salud Carlos III, Madrid, Spain, [4]Illumina, Inc.,5200 Illumina Way, San Diego, CA 02122, USA, [5]Department of Pathology, University of California, San Diego, 9500 Gilman Drive, La Jolla CA 92093, USA, [6]Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA

Next generation sequencing of the RNA content of single cells or single nuclei (sc/nRNA-seq) has become a powerful approach to understand the cellular complexity and diversity of multicellular organisms and environmental ecosystems. However, the fact that the procedure begins with a relatively small amount of starting material, thereby pushing the limits of the laboratory procedures required, dictates that careful approaches for sample quality control (QC) are essential to reduce the impact of technical noise and sample bias in downstream analysis applications. Here we present a preliminary framework for sample level quality control that is based on the collection of a series of quantitative laboratory and data metrics that are used as features for the construction of QC classification models using random forest machine learning approaches. We've applied this initial framework to a dataset comprised of 2272 single nuclei RNA-seq results and determined that ~79% of samples were of high quality. Removal of the poor quality samples from downstream analysis was found to improve the cell type clustering results. In addition, this approach identified quantitative features related to the proportion of unique or duplicate reads and the proportion of reads remaining after quality trimming as useful features for pass/fail classification. The construction and use of classification models for the identification of poor quality samples provides for an objective and scalable approach to sc/nRNA-seq quality control.

* This work is supported by the Allen Institute for Brain Science, the JCVI Innovation Fund, and the U.S. National Institutes of Health 1R21AI122100.

Contributed equally to this work.

† Corresponding author email: rscheuermann@jcvi.org.

1. Introduction

Single cell genomic analysis is poised to revolutionize our understanding of the diversity and complexity of multicellular organisms. One of the key applications of single cell genomics is the determination of transcriptional profiles using next generation sequencing of amplified cDNA synthesized from the RNA content of single cells or single nuclei (sc/nRNA-seq). By avoiding the averaging phenomenon inherent in the analysis of bulk cell populations, sc/nRNA-seq is revealing a level of cell type complexity and dynamics that is unprecedented in comparison with previous technologies.

sc/nRNA-seq has now been applied to explore a wide range of biological questions. It has been used to understand the heterogeneity of somatic mutations acquired in cancer subclones arising from the same progenitor [Patel 2014][Min 2015], providing insights into therapeutic responses and disease progression. sc/nRNA-seq has been used to track cell state transition dynamics during normal tissue differentiation [Nestorowa 2016], cell cycle progression [Scialdone 2015], and *in vitro* trans-differentiation induced using direct reprogramming methodologies [Treutlein 2016]. It has also been used to investigate the dynamics of X chromosome inactivation in preimplantation embryos [Petropoulus 2016], lineage determination during blastocyst development [Blakeley 2015], T cell receptor repertoires in antigen-specific immune responses [Eltahla 2016], T cell progressive cell states [Proserpio 2016], variability in cellular responses to viral infections [Ciuffi 2016], and the similarities between induced pluripotent stem cell-derived neurons and cells from primary tissue and cortical layers [Handel 2016]. And at its most basic level, sc/nRNA-seq is being used to understand the complexity of steady state cell type distributions in normal human tissues [Zeisel 2015][Wang 2016][Lacar 2016][Li 2016], and abnormal tissue disorders [Ramsköld 2012][Glaublomme 2015][Tirosh 2016].

RNA-seq from single *nuclei* (Grindberg, 2013) provides transcriptomes that strongly reflect those obtained from whole cells. Nuclei can be used in place of cells to assess cell type and state, as well as revealing mRNAs and non-coding RNAs that are differentially enriched in the nucleus. The use of nuclei as a starting material also has the advantage of providing individual transcriptomes without the harsh proteolytic treatment required to disperse single cells from intact tissue specimens, which is known to alter gene expression and damage sensitive cell types. snRNA-seq has enabled single neuron studies even from postmortem human brain tissue (Krishnaswami, 2016). Use of nuclei for RNA-seq enabled the first single neuron analysis of immediate early gene expression associated with memory formation in the mouse hippocampus, whereas proteolytic dissociation of neurons yielded artifactual expression in most cells (Lacar, 2016). In this study we use data from single nuclei RNA-seq, however, the QC analysis proposed should be equally applicable to single cell data.

While the promise of sc/nRNA-seq is enormous, the methods used to isolate and specifically amplify the RNA target material in a manner that preserves the molecular structures and abundance levels pushes the limits of these technologies. As a result, the impact of contaminating nucleic acid templates (e.g. chromosomal and other contaminating DNAs, rRNA, mtDNA), technical variability in laboratory reagents and procedures (e.g. variability in the efficiencies of enzymatic reactions, quality of oligonucleotide reagents, plate position effects, reagent stability), biological variability (e.g. eQTL effects) can introduce noise and bias into the resulting sequence read data that can be difficult to control. Thus, the combination of technical noise and intrinsic biological variability makes the detection of and control for technical artifacts challenging. For this reason, the development and implementation of rigorous quality control procedures throughout the entire laboratory and informatics workflow is essential in order to assess, improve and optimize both the wet lab and dry lab component steps in order to obtain optimal transcript expression values for downstream analysis.

Here we describe an approach to quality control (QC) for sc/nRNA-seq assays in which we capture over 70 different quantitative laboratory and data metrics and use these quality metrics to construct QC classification models that can be used to filter out poor quality samples from downstream analysis. We've applied this QC approach in the context of a project to define the cell type complexity of the human brain neocortex in a collaboration involving the Allen Institute for Brain Science, the J. Craig Venter Institute, and Illumina, Inc.

2. Methods and Results

Laboratory and Informatics workflows

Our standard laboratory workflow for single nuclei RNA-seq is summarized in Figure 1 and is based on the detailed protocol described previously [Krishnaswami 2016]. Single nuclei are sorted into 96- or 384-well plates containing 2 μL 0.2% Triton X-100, 2 Units/μL RNase inhibitor, 1:2000000 dilution of ERCC spike-in RNAs (Life Technologies) per well and frozen immediately in an ethanol/dry ice bath. The ERCC external RNA control, consisting of 92 transcripts derived from NIST-certified plasmids that mimic natural eukaryotic mRNAs, is used to measure limits of detection and dynamic ranges, and can also be used to help quantify differential gene expression. Amplified cDNA is prepared using a Smart-Seq2 modification [Ramsköld 2012, Krishnaswami 2016] to our previous method [Grindberg 2013] to improve amplification of transcript 5' ends. cDNA quality is evaluated using Taqman qPCR for selected housekeeping (ACTB), ERCC, and sample-specific genes. Using the single nuclei amplified cDNA, bar coded libraries are prepared and 60 sample pools are used for next generation sequencing using paired end 2 x 150 chemistry

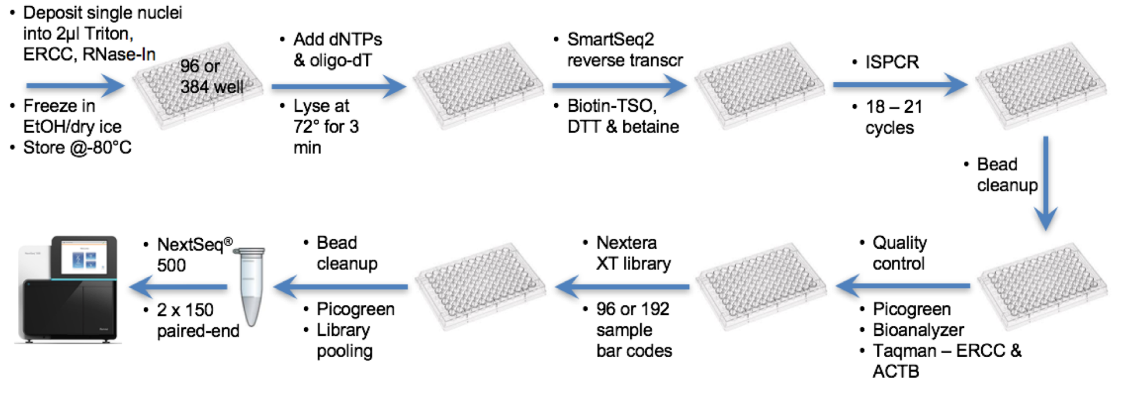


Figure 1. Single cell RNA-seq laboratory workflow – See text and [Krishnaswami 2016] for details. Abbreviations used: External RNA Controls Consortium RNA spike-in control (ERCC), RNase inhibitor (RNase-In), template switching oligonucleotide (TSO), Dithiothreitol (DTT), and beta-actin (ACTB). For Research Use Only. Not for use in diagnostic procedures.

on an Illumina NextSeq® 500 instrument. In each of our pools we also include a small number of positive (diluted, purified human RNA from bulk samples) and negative controls (water only, ERCC only). Sequencing results are quality controlled (QC) as described below, including the use of the laboratory-derived ACTB and ERCC Ct qPCR values, Bioanalyzer length distribution metrics, and picogreen cDNA concentration values.

Our standard operating procedure (SOP) for data processing includes steps for primer and quality trimming, read alignment, transcript assembly, and expression quantification as summarized in Figure 2, and has been described in detail in a recent Nature Protocol publication [Krishnaswami 2016]. After demultiplexing, cDNA, PCR, and library/bar code primer sequences and low quality reads are removed from the primary read-level data using Trimmomatic, producing the input reads for downstream steps. The input reads are fed into several downstream pipelines - RSEM (Bowtie2/EM) for transcript quantification, and TopHat (Bowtie2/Cufflinks), fastQC, MEONCA and SCavenger for quality control metric assessment. MEONCA and SCavenger are in-house developed methods that will be described elsewhere.

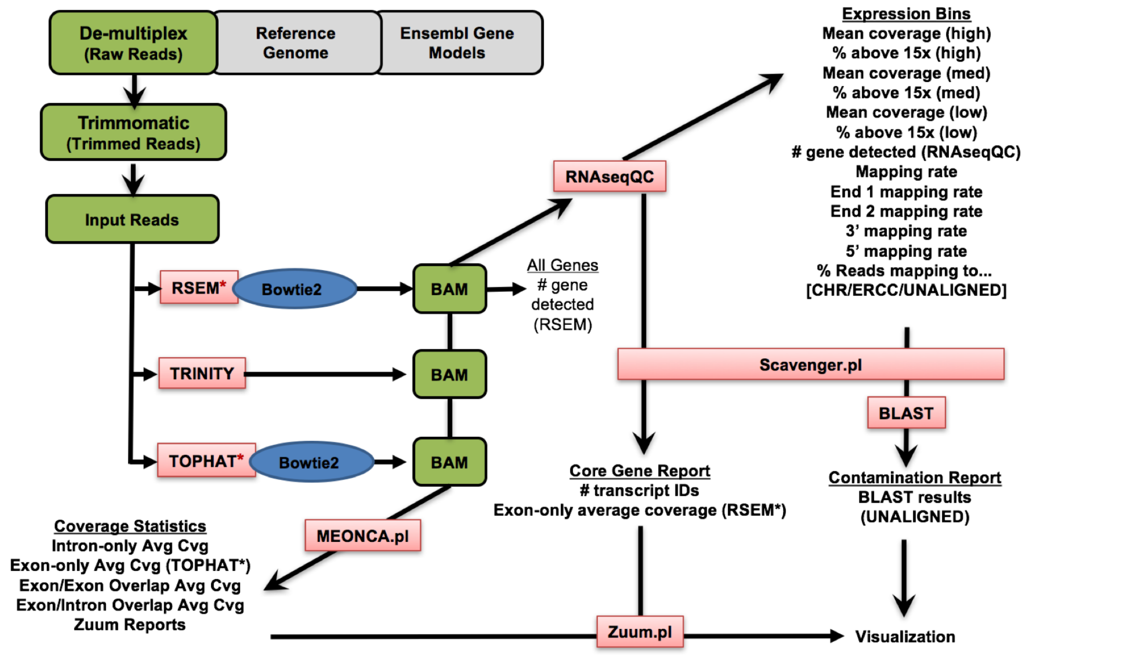


Figure 2. Single cell RNA-seq data processing workflow – Our standard operating procedure is based on the use of the Bowtie2/RSEM combination for sequence alignment, assembly, and transcript quantification. In addition, the ouput of a variety of additional workflows produce quantitative metrics used for assessment of sample and sequence quality. See text for details.

For the data included here, the following software and database versions were used:

- •GENCODE fasta and gtf files (http://www.gencodegenes.org/releases/current.html) Release 21 (GRCh38.p5);
- •FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/download.html) v0.0.14;
- •fastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) v0.10.1;
- •Picard toolkit (http://rseqc.sourceforge.net/) v1.137;
- •Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) v0.35;
- •Bowtie2 (http://sourceforge.net/projects/bowtie-bio/files/bowtie2/) v2.2.7;
- •SAM tools (http://sourceforge.net/projects/samtools/files/samtools/) v1.3;
- •RSEM: (http://deweylab.biostat.wisc.edu/rsem/) v1.2.28;
- •Tophat (https://ccb.jhu.edu/software/tophat/index.shtml) v2.1.0;
- •Cufflinks (https://cole-trapnell-lab.github.io/cufflinks/) v2.2.1.

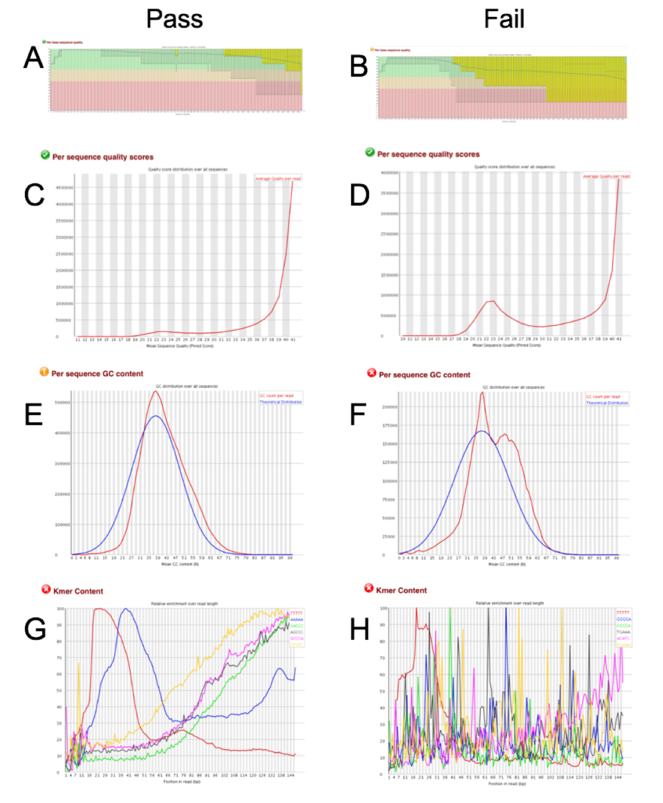


Figure 3. fastQC results used to identify pass and fail samples – Quality statistics produced by fastQC for representative Pass (A, C, E, G) and Fail (B, D, F, H) samples include average Phred score across the length of the read (A, B), average Phred score for the entire read (C, D), GC per read (E, F), and Kmer distribution across the length of the read (G, H).

One of the primary objectives of our informatics pipeline is to identify poor quality samples for possible exclusion, to determine the causes of poor quality for sample preparation process improvement, and to identify marginal quality samples for downstream bioinformatics "normalization". Because the determination of transcriptional profiles at a single cell level pushes the limits of next generation sequencing technologies, the rigorous approach we use for quality control is perhaps the most important aspect of the Single Cell Genomics Lab at JCVI.

Between the laboratory and data processing workflows described above, we collect 79 different quantitative measures that may reflect the quality of the input samples, processing steps, and resulting primary read-level data, which can be used to help address these objectives. Our approach is to use machine learning strategies, specifically random forest approaches, to classify individual sample data as either pass or fail for specific downstream analysis applications. In order to illustrate our approach, we describe the preliminary results from our work to develop a pass/fail classification model for a collaborative project between the JCVI Single Cell Genomics Lab, the Lein Group at the Allen Institute for Brain Science, and Illumina, Inc. to determine the transcriptional profiles for 2272 nuclei isolated from specific neo-cortex regions of post-mortem human brain.

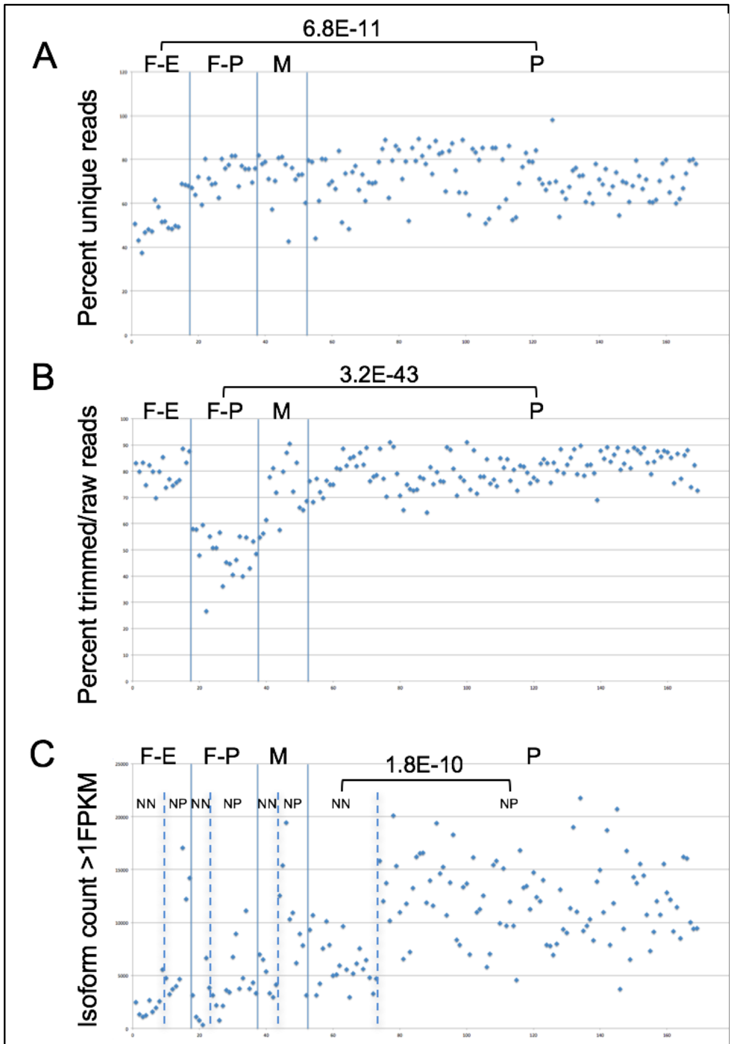


Figure 4. QC metrics in Pass and Fail samples – Single nuclei samples were annotated as Pass (P), Fail-ERCC (F-E), Fail-Phred (F-P), and Marginal (M) based on subjective evaluation of the fastQC results (see text for details). The quantitative levels of three different QC metrics for these four classes of samples are shown. P-values were calculated using a 2-sided student's t-test. NN – NeuN-; NP – NeuN+

Manual evaluation of fastQC results for QC model training

The first step in the development of machine learning classification models is to produce training data for model construction. For our purposes, we used a set of high confidence pass/fail calls for individual samples based on the qualitative assessment of data produced by fastQC, which includes quality Phred scores, GC content, Kmer distributions, and sequence over-representation information, for a random set of selected samples. Examples of these distributions are shown in Figure 3. Pass samples generally exhibit high average quality per read across the entire length of the sequenced fragment (Figure 3A & C). In contrast, Fail samples exhibit a significant number of reads with low mean quality, and quality scores that fall off down the length of the fragments (Figure 3B and D). High quality Pass samples also show an average GC content around 40%, reflecting the GC content of the expressed human transcriptome (Figure 3E). In contrast, some Fail samples show a second peak in the GC content distribution with a mean around 48% GC (Figure 3F); this peak appears to be generated from ERCC reads, which are derived from bacterial genome sequences.

Since we find that some Fail samples show reasonable Phred quality scores but over-representation of ERCC reads and vice versa, we distinguish between Fail samples due to low quality scores (Fail-Phred) and Fail samples due to ERCC over-representation (Fail-ERCC). Finally, Pass samples show a Kmer content distribution in which distinct polyA and polyT peaks can be observed toward the beginning of the read due to the use of oligo-dT priming in 1st strand cDNA synthesis (Figure 3G), whereas Fail sample often show a more random pattern (Figure 3H).

QC metric correlation with QC training data

In order to produce training data for machine learning in the 2272 nuclei study, we selected 196 samples at random, including 169 single nuclei samples and 27 controls (positive and negative), and performed a blinded qualitative evaluation of the fastQC data, producing three classification labels – Pass (152 samples, including all positive controls), Fail-Phred (29 samples), and Fail-ERCC (15 samples) (all negative controls we correctly classified into one of the two Fail categories). Qualitative fastQC evaluation was chosen to produce training data since this approach is independent from the quantitative QC metrics produced by our core data processing workflows described above. A few examples of the correlation between fastQC Pass/Fail calls and the quantitative QC metrics is shown in Figure 4. For Fail-ERCC samples, the "percent unique reads" are significantly lower (p = 6.8E-11) than for the Pass samples (Figure 4A), probably due to the fact that with a greater proportion of ERCC reads, more duplicate reads would result. For Fail-Phred samples, the "percent trimmed/raw reads" are significantly lower than for the Pass samples (Figure 4B, p = 3.2E-43), presumably due to the fact that Trimmomatic removes reads of poor quality. For Pass samples, the number of transcript isoforms detected tends to be generally higher than the number of transcript isoforms detected in either type of failed sample (Figure 4C). However, we noted that there appeared to be a subset of Pass samples that had relatively low isoform counts, similar to what we observed in the Fail samples. It turns out that during the nuclei isolation step, we stain for the expression of a neuron-specific protein, NeuN, to ensure that we get a selection of both neuronal and non-neuronal cell types for our study. When we compared data for NeuN+ and NeuN- passed samples, we found that the isoform counts were significantly different between the two major cell type categories (p = 1.8E-10), with NeuN+ nuclei and NeuN- producing an average of 12,162 and 6,233 transcript isoforms with >1FPKM, respectively.

Machine learning for high throughput QC processing

These quality annotation labels and QC metric values were then used to train the Random Forest algorithm as implemented in KNIME v3.1.2. We generated 100,000 decisions trees that could distinguish the three categories of samples. An example of a high scoring tree is shown in Figure 5 in which "percent trimmed over raw" is used at the first level and is effective at distinguishing Fail-Phred sample from both Pass and Fail-ERCC, and "percent unique reads" is used at the second level to distinguish Pass from Fail-ERCC, as also seen in Figure 4. A summary of the QC features that score high across the entire 100,000 decision tree collection is shown in Figure 6. Using this Random Forest classification model, all 196 samples in the training set were classified correctly with high confidence scores:

- Pass: average confidence = 0.9689; standard deviation = 0.0524
- Fail-Phred (F-P): average confidence = 0.8828; standard deviation = 0.0703
- Fail-ERCC (F-E): average confidence = 0.8286; standard deviation = 0.0959

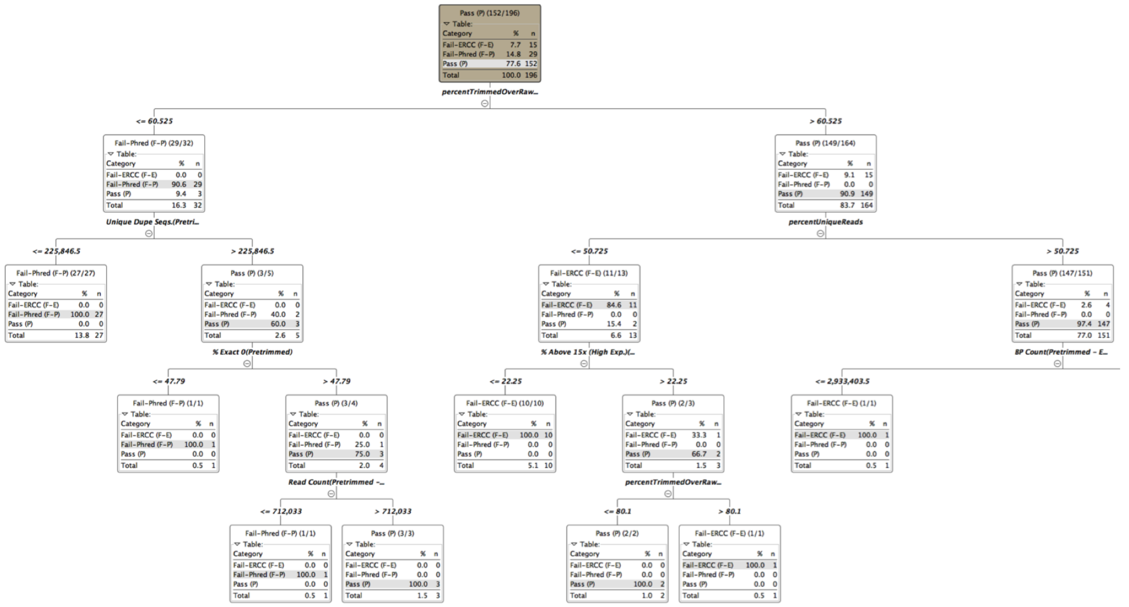
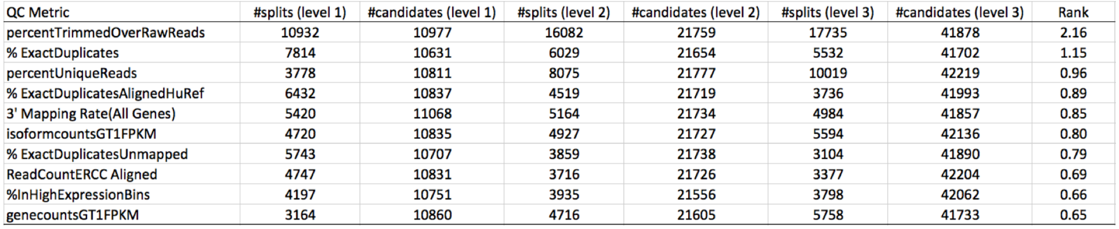


Figure 5. One of the 100,000 decision trees constructed from Random Forest training – The tree shows the different branch point levels, the feature used to segregate the branches at each level, and the segregation of the Pass, Fail-ERCC and Fail-Phred samples at each branch point. The tree was truncated after the first three levels.

To test the classification accuracy of the resulting random forest model, we used an independent test set of 185 single nuclei samples classified using the same fastQC evaluation criteria applied to the training data, with 135 determined to be Pass samples, 29 determined to be Fails and 21 determined to be Marginals. Application of the random forest model to these test Pass and Fail samples resulted in only 8 misclassifications (4.9%), for a classification accuracy of 95%. Marginal samples were split between Pass and Fail classification by the random forest model, with 8 Marginals classified as Pass and 12 classified as Fail.

Using this random forest model applied to the entire dataset, 79% of 2272 single nuclei samples passed quality control. For these samples, the average number of reads after trimming was 16,335,055 (±19,771,224), percent of hg38 mapped read was 33.04 (± 15.50), number of ERCC transcripts detected was 42.43 (± 4.37), and the number of genes detected at a level of >1FPKM was 6794 (± 2131), giving an average coverage of 793 reads per human gene detected. In contrast for Failed-ERCC samples, the average number of reads after trimming was 10,333,560 (±8,589,613), percent of hg38 mapped read was 12.18 (± 13.32), number of ERCC transcripts detected was 42.11 (± 4.73), and the number of genes detected at a level of >1FPKM was 2784 (± 1401), giving an average coverage of 452 reads per human gene detected. For Failed-Phred samples, the average number of reads after trimming was 6,763,387 (±6,167,257), percent of hg38 mapped read was 14.87 (± 12.54), number of ERCC transcripts detected was 39.60 (±12.14), and the number of genes detected at a level of >1FPKM was 2903 (± 1897), giving an average coverage of 346 reads per human gene detected. Removal of these poor quality samples was found to produce tighter cell type clusters in downstream PCA/biSNE analysis (data not shown).



QC Metric	#splits (level 1)	#candidates (level 1)	#splits (level 2)	#candidates (level 2)	#splits (level 3)	#candidates (level 3)	Rank
percentTrimmedOverRawReads	10932	10977	16082	21759	17735	41878	2.16
% ExactDuplicates	7814	10631	6029	21654	5532	41702	1.15
percentUniqueReads	3778	10811	8075	21777	10019	42219	0.96
% ExactDuplicatesAlignedHuRef	6432	10837	4519	21719	3736	41993	0.89
3' Mapping Rate(All Genes)	5420	11068	5164	21734	4984	41857	0.85
isoformcountsGT1FPKM	4720	10835	4927	21727	5594	42136	0.80
% ExactDuplicatesUnmapped	5743	10707	3859	21738	3104	41890	0.79
ReadCountERCC Aligned	4747	10831	3716	21726	3377	42204	0.69
%InHighExpressionBins	4197	10751	3935	21556	3798	42062	0.66
genecountsGT1FPKM	3164	10860	4716	21605	5758	41733	0.65

Figure 6. QC features most useful in Pass/Fail classification trees – The top ten QC metrics found useful for Pass/Fail sample classification are listed together with the number of trees in which they were used for branching at levels 1, 2, and 3, and the number of times they were considered as candidates at that given level (due to the feature down-sampling used by the Random Forest algorithm. For example, percentTrimmedOverRawReads was considered as a candidate feature in 10977 level 1 branches and was selected as the best feature 10932 times.

Discussion/Conclusion

Many groups using sc/nRNA-seq to identify and quantify cellular diversity in complex tissue samples have recognized the critical importance of quality control procedures to obtain optimal results in downstream data analysis, and have used qualitative and quantitative assessment of single quality metrics for this purpose. These include abnormal expression of housekeeping genes (e.g. ACTB, GAPDH) [Ting 2014, Treutlein 2014], outlier clustering [Zeisel 2015, Jiang 2016], median expression value cutoffs [Pollen 2014], and number of genes detected or read mapping rate [Kumar 2014], each with their advantages and disadvantages. In this paper we have demonstrated the use of a machine learning approach, specifically random forest decision trees with a large combination of wet lab and dry lab quantitative metrics, to objectively perform this QC classification. The advantage of this approach is that not only does it provide for an objective, high-throughput pass-fail classification, but it also identifies those quantitative metrics that are most useful in identifying problematic samples.

In this study, we found that there appear to be at least two classes of failed samples, and that the metrics useful in identifying each are different. Failed samples with a second peak in the %GC content plot apparently due to reads derived from the ERCC spike-in control are identified by metrics like the % of exact duplicates and % of unique reads, presumably due to the fact that a relatively small number of transcripts derived from the ERCC control are responsible for a significant proportion of the total reads obtained from those samples. In contrast, failed samples with relatively poor quality scores (low Phred scores) are identified by metrics like the % of trimmed over raw reads, presumably due to the impact of poor quality data trimming by the Trimmomatic software. While there are some metrics that appear to be effective at identifying both classes of failed samples, e.g. the number of transcript isoforms with FPKM values greater than 1, these do not rank as high as the class-specific metrics in the useful feature list. This suggest that identifying and distinguish different types of failure modes may be useful for building QC classification models using machine learning approaches. And while both the three class prediction model used here and a two class prediction model constructed by grouping both fail categories into one showed perfect classification of the training data, the prediction confidence values for calling pass samples were slightly higher using the three class model.

In addition, we also find that the use of metrics related to the number of genes or transcript isoforms detected for quality control purposes should be approached cautiously since these may vary between different cell types, as we observed between our NeuN+ neurons and our NeuN- glial cells, or between different cellular states (e.g. cell cycle phase or activation state).

Recently, Ilicic et al. reported the use of support vector machine modeling to identify stressed/broken/killed cells, empty capture sites and sites with multiple cells in Fluidigm C1 flow cells using microscopic visualization as the gold standard for model training [Ilicic 2016]. They found seven features that were useful for classification independent of cell type and protocol – cytoplasm and mitochondrially-localized proteins, mtDNA-encoded genes, mapped reads, multi-mapped reads, non-exonic reads, and transcriptome variance. Differences between these and the features reported here could be due to the use of different quality metrics as input, the use of nuclei versus whole cells, or that different sorting platforms give rise to different poor quality modes. In any case, the approach reported here is advantageous because it does not require visual microscopic inspection to produce the gold standard results for model training and therefor can be applied in a high throughput fashion to data from any cell sorting platform. While the random forest model developed here has yet to be applied to a completely independent dataset, the test samples used to assess classification accuracy were derived from separate cDNA synthesis, PCR amplification, and library preparation reactions and sequencing runs. The fact that the model gave a 95% classification accuracy on this semi-independent dataset suggests that the feature included in the model are at least robust to technical batch effects.

In conclusion, the use of both wet lab and dry lab metrics for the production of a QC classification model using random forest machine learning appears to be an effective objective strategy for the quality control of sc/nRNA-seq samples, providing further insights into the data features that are most useful for identifying problematic samples.

Acknowledgements

This work is supported by the Allen Institute for Brain Science, the JCVI Innovation Fund, and the U.S. National Institutes of Health 1R21AI122100.

References

Blakeley P, Fogarty NM, del Valle I, et al. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. Development. 2015 Sep 15;142(18):3151-65. doi: 10.1242/dev.123547. Epub 2015 Aug 20. Erratum in: Development. 2015 Oct 15;142(20):3613. PubMed PMID: 26293300; PubMed Central PMCID: PMC4582176.

Ciuffi A, Rato S, Telenti A. Single-Cell Genomics for Virology. Viruses. 2016 May 4;8(5). pii: E123. doi: 10.3390/v8050123. Review. PubMed PMID: 27153082; PubMed Central PMCID: PMC4885078.

Eltahla AA, Rizzetto S, Pirozyan MR, et al. Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells. Immunol Cell Biol. 2016 Jul;94(6):604-11. doi: 10.1038/icb.2016.16. Epub 2016 Feb 10. PubMed PMID: 26860370.

Gaublomme JT, Yosef N, Lee Y, et al. Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. Cell. 2015 Dec 3;163(6):1400-12. doi: 10.1016/j.cell.2015.11.009. Epub 2015 Nov 19. PubMed PMID: 26607794; PubMed Central PMCID: PMC4671824.

Grindberg RV, Yee-Greenbaum JL, McConnell MJ, et al. RNA-sequencing from single nuclei. Proc Natl Acad Sci U S A. 2013 Dec 3;110(49):19802-7. doi: 10.1073/pnas.1319700110. Epub 2013 Nov 18. PubMed PMID: 24248345; PubMed Central PMCID: PMC3856806.

Handel AE, Chintawar S, Lalic T, et al. Assessing similarity to primary tissue and cortical layer identity in induced pluripotent stem cell-derived cortical neurons through single-cell transcriptomics. Hum Mol Genet. 2016 Mar 1;25(5):989-1000. doi: 10.1093/hmg/ddv637. Epub 2016 Jan 5. PubMed PMID: 26740550; PubMed Central PMCID: PMC4754051.

Ilicic T, Kim JK, Kolodziejczyk AA, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016 Feb 17;17:29. doi: 10.1186/s13059-016-0888-1. PubMed PMID: 26887813; PubMed Central PMCID: PMC4758103.

Jiang P, Thomson JA, Stewart R. Quality control of single-cell RNA-seq by SinQC. Bioinformatics. 2016 Apr 10. pii: btw176. [Epub ahead of print] PubMed PMID: 27153613.

Krishnaswami SR, Grindberg RV, Novotny M, et al. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. Nat Protoc. 2016 Mar;11(3):499-524. doi: 10.1038/nprot.2016.015. Epub 2016 Feb 18. PubMed PMID: 26890679; PubMed Central PMCID: PMC4941947.

Kumar RM, Cahan P, Shalek AK, et al. Deconstructing transcriptional heterogeneity in pluripotent stem cells. Nature. 2014 Dec 4;516(7529):56-61. doi: 10.1038/nature13920. PubMed PMID: 25471879; PubMed Central PMCID: PMC4256722.

Lacar B, Linker SB, Jaeger BN, et al. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. Nat Commun. 2016 Apr 19;7:11022. doi: 10.1038/ncomms11022. PubMed PMID: 27090946; PubMed Central PMCID: PMC4838832.

Li J, Klughammer J, Farlik M, et al. Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. EMBO Rep. 2016 Feb;17(2):178-87. doi: 10.15252/embr.201540946. Epub 2015 Dec 21. PubMed PMID: 26691212; PubMed Central PMCID: PMC4784001.

Min JW, Kim WJ, Han JA, et al. Identification of Distinct Tumor Subpopulations in Lung Adenocarcinoma via Single-Cell RNA-seq. PLoS One. 2015 Aug 25;10(8):e0135817. doi: 10.1371/journal.pone.0135817. eCollection 2015. PubMed PMID: 26305796; PubMed Central PMCID: PMC4549254.

Nestorowa S, Hamey FK, Pijuan Sala B, et al. A single cell resolution map of mouse haematopoietic stem and progenitor cell differentiation. Blood. 2016 Jun 30. pii: blood-2016-05-716480. [Epub ahead of print] PubMed PMID: 27365425.

Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014 Jun 20;344(6190):1396-401. doi: 10.1126/science.1254257. Epub 2014 Jun 12. PubMed PMID: 24925914; PubMed Central PMCID: PMC4123637.

Petropoulos S, Edsgärd D, Reinius et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell. 2016 May 5;165(4):1012-26. doi: 10.1016/j.cell.2016.03.023. Epub 2016 Apr 7. PubMed PMID: 27062923; PubMed Central PMCID: PMC4868821.

Pollen AA, Nowakowski TJ, Shuga J, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 2014 Oct;32(10):1053-8. doi: 10.1038/nbt.2967. Epub 2014 Aug 3. PubMed PMID: 25086649; PubMed Central PMCID: PMC4191988.

Proserpio V, Piccolo A, Haim-Vilmovsky L, et al. Single-cell analysis of CD4+ T-cell differentiation reveals three major cell states and progressive acceleration of proliferation. Genome Biol. 2016 May 12;17(1):103. doi: 10.1186/s13059-016-0957-5. Erratum in: Genome Biol. 2016;17(1):133. PubMed PMID: 27176874; PubMed Central PMCID: PMC4866375.

Ramsköld D, Luo S, Wang YC, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012 Aug;30(8):777-82. PubMed PMID: 22820318; PubMed Central PMCID: PMC3467340.

Scialdone A, Natarajan KN, Saraiva LR, et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods. 2015 Sep 1;85:54-61. doi: 10.1016/j.ymeth.2015.06.021. Epub 2015 Jul 2. PubMed PMID: 26142758.

Ting DT, Wittner BS, Ligorio M, et al. Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. Cell Rep. 2014 Sep 25;8(6):1905-18. doi: 10.1016/j.celrep.2014.08.029. Epub 2014 Sep 18. PubMed PMID: 25242334; PubMed Central PMCID: PMC4230325.

Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016 Apr 8;352(6282):189-96. doi: 10.1126/science.aad0501. PubMed PMID: 27124452; PubMed Central PMCID: PMC4944528.

Treutlein B, Brownfield DG, Wu AR, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 2014 May 15;509(7500):371-5. doi: 10.1038/nature13173. Epub 2014 Apr 13. PubMed PMID: 24739965; PubMed Central PMCID: PMC4145853.

Treutlein B, Lee QY, Camp JG, et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. Nature. 2016 Jun 8;534(7607):391-5. doi: 10.1038/nature18323. PubMed PMID: 27281220; PubMed Central PMCID: PMC4928860.

Wang YJ, Schug J, Won KJ, et al. Single cell transcriptomics of the human endocrine pancreas. Diabetes. 2016 Jun 30. pii: db160405. [Epub ahead of print] PubMed PMID: 27364731.

Zeisel A, Muñoz-Manchado AB, Codeluppi S, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015 Mar 6;347(6226):1138-42. doi: 10.1126/science.aaa1934. Epub 2015 Feb 19. PubMed PMID: 25700174.

For Research Use Only. Not for use in diagnostic procedures.

TRACING CO-REGULATORY NETWORK DYNAMICS IN NOISY, SINGLE-CELL TRANSCRIPTOME TRAJECTORIES

PABLO CORDERO

JOSHUA M. STUART

UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California, USA

Keywords: single-cell measurements, Gaussian mixtures, transcriptomics, single-cell trajectory reconstruction

The availability of gene expression data at the single cell level makes it possible to probe the molecular underpinnings of complex biological processes such as differentiation and oncogenesis. Promising new methods have emerged for reconstructing a progression ’trajectory’ from static single-cell transcriptome measurements. However, it remains unclear how to adequately model the appreciable level of noise in these data to elucidate gene regulatory network rewiring. Here, we present a framework called Single Cell Inference of MorphIng Trajectories and their Associated Regulation (SCIMITAR) that infers progressions from static single-cell transcriptomes by employing a continuous parametrization of Gaussian mixtures in high-dimensional curves. SCIMITAR yields rich models from the data that highlight genes with expression and co-expression patterns that are associated with the inferred progression. Further, SCIMITAR extracts regulatory states from the implicated trajectory-evolving co-expression networks. We benchmark the method on simulated data to show that it yields accurate cell ordering and gene network inferences. Applied to the interpretation of a single-cell human fetal neuron dataset, SCIMITAR finds progression-associated genes in cornerstone neural differentiation pathways missed by standard differential expression tests. Finally, by leveraging the rewiring of gene-gene co-expression relations across the progression, the method reveals the rise and fall of co-regulatory states and trajectory-dependent gene modules. These analyses implicate new transcription factors in neural differentiation including putative co-factors for the multi-functional NFAT pathway.

Introduction

Understanding the dynamics of gene expression progression in a cell population as it traverses a biological process such as differentiation has been an outstanding problem in modern cell biology. These dynamics are characterized not only by the changes in cell-to-cell gene expression levels, but by the rewiring of gene regulatory networks as the cells transform from one transcriptional state to another. Tracking these gene regulatory changes would pinpoint coordination of biological function as gene modules are turned on or off throughout the progression.

Single-cell transcriptomics has given important insights into gene expression dynamics, revealing the stochastic nature of gene expression and characterizing in detail the behavior of small genetic networks.[1–4] In their initial incarnation, these measurements were confined to demanding microscopy protocols that assayed gene expression levels through time of only a handful of genes. In recent years, advances in flow cytometry, microfluidics, and sequencing technologies have enabled the interrogation of up to the whole transcriptome in hundreds to thousands of cells.[5–7] Application of these techniques to biological processes such as development provide snapshots of cell states through time and space.

Many computational methods have emerged to infer trajectories of connected state transitions from the static samplings of single-cell transcriptomes. The goal of these methods is to provide a pseudotemporal ordering of cells in which neighboring cells are similar to each other, capturing an overall biological progression. These approaches have been successfully applied to elucidate complex transcriptional patterns and regulators in myoblast differentiation,[8] B cell development,[9] and haematopoiesis.[10] Nevertheless, cell orderings alone give little insight into the state of gene regulatory networks across time. In addition, while most methods use strategies to tackle biological and technical noise, none account for the dynamic, heteroscedastic nature of the data. Further, only a few take into consideration uncertainties in pseudotime assignments,[11] making error estimates difficult to evaluate.

To address these challenges we propose a strategy, Single Cell Inference of MorphIng Trajectories and their Associated Regulation (SCIMITAR), for inferring gene expression network dynamics throughout biological progression from static, single-cell transcriptomes. SCIMITAR gives a detailed, fully probabilistic description of the expression trajectory that, in contrast with previous methods, explicitly accounts for heteroscedastic noise in the data. In addition, it tracks the changes of gene-gene expression correlations at each point in the progression. The probabilistic nature of SCIMITAR transition models allows for evaluating the shape of the multivariate gene expression distribution as a function of biological progression, which we show can be used to pinpoint co-regulatory cell states.

We benchmarked SCIMITAR's inference capabilities in two scenarios. First, we tested its ability to infer cell ordering and network rewiring from simulated transcriptomic measurements where the underlying cell behavior was known. Second, we asked whether SCIMITAR could yield insights in the developmental trajectory of human fetal neurons by analyzing recently published fetal brain single-cell measurements. A likelihood ratio test designed for SCIMITAR revealed 36 genes that significantly varied throughout the progression but that were missed by standard differential expression between cell groups including genes in cornerstone developmental pathways such as the hypoxia inducible factor 1 α (HIF1α), nuclear factor of activated T cells (NFAT), and androgen receptor (AR) pathways. Further, by tracking SCIMITAR coexpression matrices across pseudotime we were able to detect the evolution of co-regulatory states, gene modules, and genes that gained and lost connectivity throughout the trajectory.

Results

Uncovering the full probability distribution progression underlying static single-cell measurements with SCIMITAR

Recently, there has been an explosion of single-cell transcriptomic data in various biomedical contexts and systems. A projection of the data from three such studies (refs[8,10,12]) in Fig 1A using a locally linear embedding reveals that these datasets are characterized by distinct groups of many cells interspersed with cells that fall along what appear to be isolines between groups. This structure suggests a model that combines distributions for cell population density and evolving cell states with heteroscedastic noise. One such model that could describe these data is a continuous mixture of Gaussian distributions with constraints that allow only for smooth, continuous changes in parameters over the course of the progression. We call such a model a Morphing Gaussian Mixture (MGM, see Methods and Fig 1B). The MGM has a mean function, $\mu : [0, 1] \to \mathbb{R}^n$ that threads through the data and is equipped with a covariance matrix function $\Sigma : [0, 1] \to \mathbb{R}^{n\times n}$ that defines a Gaussian distribution at each point in the progression, with n being the number of genes. The mean and covariance matrix functions vary continuously throughout the $[0, 1]$ interval, defining a probability $P(x|\mu, \Sigma, t)$ for each cell gene expression vector x and pseudo time-point $t \in [0, 1]$. To ease inference, these mean and covariance functions can be parametrized with different functional classes, such as polynomials, splines, or Gaussian processes (see Methods). This probabilistic structure maps samples to a smooth curve and allows points to veer away stochastically by modeling the structure of the changing biological and technical noise. $P(x|\mu, \Sigma, t)$ captures the uncertainty of a cell mapping to a particular pseudotime due to the changing covariance nature of the MGM. A key advantage of this approach is that it replaces standard, grouped differential gene expression analysis or differential co-expression analysis with a more sensitive test for potential gene-gene regulatory relationships that change throughout the progression. Details of the MGM model as well as inference of its parameters from data are given in the Methods section.

Benchmarking SCIMITAR in simulated data

To test our strategy, we asked whether SCIMITAR could infer the underlying cell ordering and co-expression networks of simulated data where the ground truth was available. We tested SCIMITAR's cell order inference capabilities in two settings in which noise was added to the system: 1) the noise is *uncorrelated* to the underlying trajectory and 2) the noise is *correlated* with the trajectory. The first setting, adding noise uncorrelated with the trajectory, tests robustness of the method in the presence of genes that are unrelated to the biological progression and that confound ordering inference. The second setting tests how biological and technical noise intrinsic to the system, including gene-gene correlated noise that change over time, affect cell ordering inference.

For the first setting, we simulated data closely following the simulation procedure described in ref.[9] We simulated data in which 3 genes defined the true cell state and 7 genes represented unrelated (uncorrelated) expression programs to the simulated progression. Simulations in this scenario then, 3 dimensions of the data were "signal" while 7 were "noise". To obtain the three-dimensional trajectory, we performed a random walk for 600 steps and sampled a 'cell' from a standardized normal distribution centered at the current point in the walk. We then added seven dimensions of Gaussian noise. We generated several datasets with an increasing noise magnitude (quantified as the standard deviation times the range of the trajectory). We then used SCIMITAR to model these data and obtain the model's optimal cell ordering. We used SCIMITAR with three different functional classes (see Methods): third degree polynomials, cubic splines, and Gaussian Processes with a squared exponential correlation function (GP). We compared SCIMITAR's performance with the cell orderings inferred by two popular methods, Monocle[8] and Wanderlust,[9] and used the Pearson correlation coefficient to compare the approaches (see Fig 2A). The best overall performers were all SCIMITAR models, with Wanderlust coming in close second and Monocle performing slightly worse possibly due to its assumption of linearity in its dimensionality reduction step in agreement with previous studies.[13] All methods were susceptible to the noisy dimensions uncorrelated with the trajectory.

For the second test that adds noise correlated with the trajectory, we simulated a curve, μ_{sim} traversing a 10-dimensional space using 10 randomly-generated quadratic polynomials. The correlated noise was simulated from the evolution of randomly generated Watts-Strogatz networks and an additional set of quadratic polynomials with 6 different settings of signal-to-noise ratios (see Supplemental Methods for a detailed description of this benchmark). We found all methods performed similarly (Fig 2B), suggesting that noise intrinsic to the system, including gene-gene statistical dependencies, equally confounds any cell ordering inference method.

In addition to solving the cell ordering problem, SCIMITAR models track evolving gene-gene correlations. We used the correlated noise simulations to test the accuracy of SCIMITAR's gene network rewiring inference. To this end, we compared the covariance functions inferred by the polynomial, spline, and GP SCIMITAR versions. We measured the concordance of trends between each entry of the predicted matrix functions $\Sigma_{ij}^{pred}(t)$ and the corresponding entry of the simulated values $\Sigma_{ij}^{sim}(t)$ using the Pearson correlation coefficient (see Fig 2C). The spline version of SCIMITAR produced the highest correlation coefficients while all versions were substantially better than randomly-generated covariance matrix functions. Closer examination of the three functional classes revealed that the GP version tended to overfit the data locally, closely following local covariance structure even in regions where a few samples were present while the polynomial version lacked the flexibility to model some complex twists and turns in evolving true covariance structures. The spline version struck a balance between smoothing inferences in intervals of the trajectory with few samples and maintaining flexibility to capture non-linear trends. We therefore chose to use the spline functional class for SCIMITAR models in the remainder of this study.

A differentiation model for human fetal neurons

In a previous study, Darmanis et al. obtained a transcriptomic map of the adult and fetal brain using single-cell RNA-seq measurements.[14] One of the findings of the study was a continuous transition the between fetal replicating and quiescent neurons. We applied SCIMITAR to infer cell ordering and network rewiring of these data to elucidate key regulatory changes across the differentiation process. We downloaded these data from the gene expression omnibus (series identifier GSE67835) and obtained the subset corresponding to all fetal neurons. We focused on all transcription factors that were expressed in at least 10% of the cells, log-transformed the data and controlled for cell-cycle effects using scLVM.[15] We then fit SCIMITAR to the data and visualized the results in a two-dimensional locally linear embedding (see Fig 3A). The visualization suggested a single linear trajectory that traversed the fetal replicating and quiescent neurons which was captured by the SCIMITAR model. To obtain progression associated genes, we used a likelihood ratio test tailored for SCIMITAR models with dynamic noise (see Methods). The test revealed 92 genes with expression that was significantly psuedotemporal-dependent (see Fig 3B). To obtain global insights from these genes, we used hierarchical clustering with the Pearson correlation similarity metric to group them into 5 groups and performed Gene Ontology and KEGG pathway enrichment tests on each group (see color groups in Fig 3B). Early-expressed genes (red and green clusters) were associated with glucocorticoid receptors, heat shock factors, and signal transduction; genes expressed in the middle of the progression (yellow and pink clusters) were enriched with Maf-like proteins and cytokines; and the late-expressed genes (cyan cluster) had apoptosis, neurogenesis, and alternative splicing enrichment. These enrichments correspond to multiple observations in the literature. For example, heat shock factor proteins are well known to be involved in early neurodifferentiation[16] while glucocorticoid receptors and Maf-like proteins are found to be expressed at different stages in hippocampal and developmental neurogenesis, respectively.[17,18] Further, neurodifferentiation has been found to be particularly enriched for alternative splicing events.[19]

We then compared SCIMITAR's progression associated genes to those obtained using an ANOVA differential expression test between cells grouped according to their fetal replicating or quiescent annotations. SCIMITAR uncovered 36 genes missed by ANOVA, most of which were highly expressed in the middle of the progression, a detail that is lost when grouping cells into two groups. These missed genes implicate different pathways whose genes were engaged in progression dynamics. For example, five genes, BHLHE40, SMAD3, SP1, and SMAD4, of the hypoxia inducible factor 1 α (HIF1α) pathway, involved in neural development,[20] were revealed to follow an ordered progression by the SCIMITAR model but missed using grouped ANOVA differential expression (see Fig 3C). SCIMITAR revealed that the progression associated genes of this pathway were mostly active in early stages of differentiation. SCIMITAR also illuminated two other pathways: the Nuclear factor of activated T-cells (NFAT) and the Androgen receptor pathway which is critical for neural stem cell fate commitment[21,22] (see Fig 3C).

We note that SCIMITAR's progression associated genes did not include 7 genes from the ANOVA list, false positives for which the variance was too large or where the statistic was skewed by outliers in an otherwise lowly expressed gene. Nevertheless, three genes that seem to be be differentially expressed by manual inspection (BCL11B, AFF1, and REST) were found by ANOVA but missed by SCIMITAR, presumably due to a small subset of cells driving the change between groups.

Evolving co-expression networks reveal defined co-regulatory states

We then used SCIMITAR's inferred covariance functions to track changes in gene-gene connectivity across the progression. We sampled 100 correlation matrices at regular intervals from the covariance function, restricting the matrices to genes deemed progression associated. We calculated a global distance matrix between networks using Frobenius distance to assess their similarities and plotted the similarity values across pseudotime (see Fig 4A). As expected, the strongest similarities were between networks that were neighbors in pseudotime. However, three network clusters could be appreciated in the matrix, suggesting three different co-regulatory states. We obtained the consensus network of each state by averaging the network members of the cluster. Then, we ranked each gene by comparing their co-expression degree in each state to their co-expression degrees in the other two states using z-scores.

The top 20 genes that gained the most connectivity in each state are listed in Fig 4B. All of the gain-of-connectivity genes include genes that have been established as key players in neurodifferentiation, such as PAX6, DLX1, and NEUROD6 and were enriched with neurodevelopmental and neurogenesis GO terms.

To track highly connected gene modules of each state that significantly changed their connectivity, we obtained gene modules for each co-regulatory state using affinity propagation (with a dampening parameter of 0.5), finding 27 gene modules in total. We annotated these modules by gene set enrichment and ordered them across pseudotime (see Fig 4C). This analysis revealed a coordinated functional response across the trajectory: modules in state 1 were annotated with neural stem cell commitment, immune response, and protein trafficking, while state 2 was enriched with embryonic development, neuron regulation, and pallium development. State 3 had more diverse enrichments, from morphogenesis to membrane organelles, suggesting a stage when cells start taking on mature neuron roles depleted of differentiation potential. Importantly, this analysis pinpointed an NFAT-associated module to be most active in co-regulatory state 2 (see Fig 4D). Most NFAT co-factors involved in neural development are still unknown.[23] The uncovered NFAT-associated module provides putative candidates for this function. The full list of modules and their gene networks can be found in the Supplemental Results (see below).

Discussion

An outstanding goal of systems biology is to understand the principles under which the gene regulatory circuitry of a cell changes during a biological process. Single-cell transcriptomes offer a fast way to obtain transcriptome-wide snapshots of these processes. When properly analyzed, these data can be used to recover the principal trends of the biological progression, but current methods do not model the dynamic gene-to-gene correlations in expression that are the hallmarks of the underlying regulatory circuitry. Here, we presented SCIMITAR, a strategy that leverages morphing Gaussian mixtures to track biological progression and model the rewiring of these gene networks from static transcriptomes. SCIMITAR models account for heteroscedastic noise and increase the statistical power to detect progression-associated genes when compared to traditional differential expression tests. Further, the models allow for detecting modes in co-expression structure in the trajectory: defined co-regulatory states that represent potential metastable and transitionary cell states. We note that Gaussian mixtures with non-diagonal covariance matrices suffer from the curse of dimensionality, which we have tried to control for by using shrinkage estimators. Exploring the robustness of other types of regularized estimators such as the graphical LASSO would be a logical next step to improve confidence in the inferred morphing mixture models.

SCIMITAR is part of a recent wave of probabilistic methods for cellular trajectory reconstruction from single-cell measurements.[11,24] These types of models present several advantages, such as assigning uncertainty estimates of cell orderings and providing a natural way for mapping new samples to a trained model — a necessary task for building queryable trajectory maps with multiple progressions. Although SCIMITAR as presented cannot model branched cellular trajectories such as those corresponding to multiple cell fate decisions, the framework can be readily extended by replacing the single-curve parametrization of the mixtures with a branching structure, which deserves further investigation.

Methods

Morphing Gaussian Mixtures: correlated gene progression modeling with no dimensionality reduction

Single-cell transcriptomic measurements are high-dimensional, with the number of variables measured typically ranging from a few markers (generally no less than 48) to the full transcriptome that can be upwards around 30000 transcripts. However, not every gene or transcript is relevant to the biological system of interest and most are not expressed at all. Further, due to the underlying gene regulatory networks, the expression patterns of many genes are correlated and the strength of this correlation changes throughout the progression as the regulatory system changes from one cell state to the next. These biological constraints put the data in some low-dimensional manifold, a property that is used in various ways by cell ordering algorithms to justify reducing the dimensionality of the dataset to a manageable number of dimensions. Monocle, for example, reduces the data's dimensionality to 2 dimensions using independent component analysis and performs its calculations on a lower dimensional manifold. While the procedure captures general aspects of the trajectory, 2 dimensions is generally not enough to capture all of the relevant variability of the data and the reduction leads to loss of information that can impact trajectory reconstruction (see e.g. our benchmarks in the Results sections and other benchmarks in[13,24]). Other methods, such as Wanderlust, reduce the dimensionality in a more principled way through nearest-neighbor calculations but forego capturing the changes in gene-gene expression correlations over time. To address both of these shortcomings, we introduce a model that retains the dimensionality of the dataset and tracks gene-gene correlations throughout the trajectory. To this end, we extended Gaussian graphical models to accommodate time-dependent changes in the mean and covariances of the model with time being a latent variable.

Gaussian graphical models are one of the dominant frameworks for analyzing gene expression data, where the data is assumed to follow a multivariate Gaussian distribution defined by a mean vector and a covariance matrix. Modeling the data becomes more challenging in the presence of population structure where several different populations, each with its own distribution, are intermixed. Gaussian mixture models, which posit that the data comes from a finite combination of multivariate Gaussians, have been used successfully in this scenario.[25] In static single-cell expression from a group of cells continuously undergoing a biological process, such as differentiation, the boundaries between populations are blurred and the data is best described as a continuous transformation between the first and last states. We model this transformation by assuming that the data comes from a *continuous* Gaussian mixture, parametrized by timepoints within the progression (the so-called pseudotime), which are unknown. Let X be the data, p the number of genes, $\mu : [0,1] \to \mathbb{R}^p, \Sigma : [0,1] \to \mathbb{R}^{p \times p}$ the mean and covariance functions of the evolving populations that are time dependent, and γ a probability distribution on the $[0,1]$ interval representing cell population density at each pseudo time-point. Then the probability of the data given the model $M = \{\mu, \Sigma, \gamma\}$ can be written as:

$$P(X|M) = \int_0^1 \gamma(t) P(X|\mu(t), \Sigma(t)) \tag{1}$$

Here, t stands for the pseudotime in the progression. This model, which we name the morphing Gaussian mixture model (MGM), differs from other mixture models in that we require the mean and covariance structures to be described through continuous functions and generalize other related models such as principal curves by inferring local covariance structure in addition to the mean curve. The changing covariance structure allows the model to both keep the dimensionality of the dataset and track co-expression changes throughout the progression.

To fit the model to the data, we use a maximum likelihood approach. As previously defined, the parameters in the MGM model are difficult to infer, since optimization of the likelihood function requires searching the space of all continuous functions. Additionally, the positive-definite requirement on $\Sigma(t)$ makes fitting the matrix function difficult. Therefore, we recast the problem of fitting $\Sigma(t)$ into fitting its pseudotime-dependant Cholesky decompositions: $\Sigma(t) = C(t)^T C(t), \forall t$ and impose a functional form to the $\mu(t)$ and $C(t)$ functions. We consider three different functional classes: polynomials, Gaussian processes with squared exponential correlation models, and cubic, De Boor smoothing splines, a special case of Gaussian processes.

To fit the parameters of the model, we employ coordinate ascent. In the first step, we are given a fixed set values for M and we calculate, for each sample x, the optimal pseudotime t_{opt} in the $[0, 1]$ interval for which $P(x|\mu(t_{opt}), \Sigma(t_{opt}))$ is maximized. In the second step, given optimal pseudotime values, we calculate the cell density γ by fitting kernel density estimator to the assigned pseudo time-points. Finally, in the third step, given density weights γ and pseudotime assignments, we find the μ and Σ functions that best fit the data. To achieve this, we approximate $\mu(t)$ and $C(t)$ locally by obtaining optimal values at the pseudo time-points $0, 0.1, 0.2, ..., 1.0$, inferring the local mean and covariance using each data point weighted by their probabilities as given by γ, and leveraging these values to fit functions from the desired functional class (e.g. a polynomial, spline, or Gaussian process). Because we may have considerably less samples than genes, we use the Ledoit-Wolf-type estimator in the R `corpcor` package to fit the covariance at each pseudo time-point. We repeat this procedure until convergence, as evaluated by the Pearson correlation coefficient of current and past pseudotimes, with stopping criterion $r > 0.9$. As initial values for pseudotime assignments to our optimization routine, we use a de-noised one-dimensional locally linear embedding.[26]

Visualization of the data and SCIMITAR models

To visualize the data and models, we use 2-dimensional locally-linear embeddings, with number of neighbors set to 80% of the number of samples. We plot SCIMITAR means by sampling 100 equidistant points across the mean function and projecting to the embedding. To obtain a projection of the SCIMITAR model's probability density function, we obtain 1000 samples from the model, evenly spaced across pseudotimes in the $[0, 1]$ interval, project to the embedding, and plot a 2-dimensional kernel density estimator of the 1000 points.

A progression association statistical test

To obtain genes whose expression is progression-dependent, we use a likelihood ratio test to compare the SCIMITAR model of each gene's progression and the null hypothesis where the expression of the gene is 'flat-lined', i.e. does not track with the model's path. Specifically, we calculate the statistic:

$$LR = log(L_{null}(\hat{\mu}, \hat{\sigma})) - log(L_{scim}(\mu, \Sigma)) \tag{2}$$

Where L_{scim}, L_{null} are the likelihood functions of the SCIMITAR and null models, respectively, with the null distribution defined as a normal distribution centered at the empirical mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ of all the data representing the case where the data is independent of the progression. To assess whether the null hypothesis should be rejected, we obtain the distribution of LR under the null hypothesis using parametric bootstrapping with 1000 samples and compare the resulting ratios to the LR of the data. We use the Benjamini-Hochberg procedure to correct for multiple comparisons, setting an FDR cutoff of 5%.

Acknowledgments

We thank members of the Stuart lab for feedback on the methodology and Rocio Soto Astorga for graphics support. PC and JMS are supported by a grant from the California Institute of Regenerative Medicine working under the auspices of the Stem Cell Genome Center of Excellence. JMS was also supported by NIGMS grant 5R01GM109031.

Method availability and supplementary material

SCIMITAR code and documentation are freely available at `https://github.com/dimenwarper/scimitar`. Supplementary methods and results can be found at `https://github.com/dimenwarper/scimitar/wiki`.

References

1. M. B. Elowitz, A. J. Levine, E. D. Siggia and P. S. Swain, *Science* **297**, 1183 (16 August 2002).
2. J. M. Raser and E. K. O'Shea, *Science* **304**, 1811 (18 June 2004).
3. H. Maamar, A. Raj and D. Dubnau, *Science* **317**, 526 (27 July 2007).
4. J. Paulsson, *Nature* **427**, 415 (29 January 2004).
5. F. Tang, C. Barbacioru, E. Nordman, B. Li, N. Xu, V. I. Bashkirov, K. Lao and M. A. Surani, *Nat. Protoc.* **5**, 516 (March 2010).
6. S. C. Bendall, E. F. Simonds, P. Qiu, E.-A. D. Amir, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky, R. S. Balderas, S. K. Plevritis, K. Sachs, D. Pe'er, S. D. Tanner and G. P. Nolan, *Science* **332**, 687 (6 May 2011).
7. T. Hashimshony, F. Wagner, N. Sher and I. Yanai, *Cell Rep.* **2**, 666 (27 September 2012).
8. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen and J. L. Rinn, *Nat. Biotechnol.* **32**, 381 (April 2014).
9. S. C. Bendall, K. L. Davis, E.-A. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan and D. Pe'er, *Cell* **157**, 714 (24 April 2014).

10. V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. Macaulay, W. Jawaid, E. Diamanti, S.-I. Nishikawa, N. Piterman, V. Kouskoff, F. J. Theis, J. Fisher and B. Göttgens, *Nat. Biotechnol.* **33**, 269 (March 2015).
11. K. Campbell and C. Yau, *bioRxiv* (5 April 2016).
12. G. Guo, S. Luc, E. Marco, T.-W. Lin, C. Peng, M. A. Kerenyi, S. Beyaz, W. Kim, J. Xu, P. P. Das, T. Neff, K. Zou, G.-C. Yuan and S. H. Orkin, *Cell Stem Cell* **13**, 492 (3 October 2013).
13. J. D. Welch, A. J. Hartemink and J. F. Prins, *Genome Biol.* **17**, p. 106 (23 May 2016).
14. S. Darmanis, S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. Hayden Gephart, B. A. Barres and S. R. Quake, *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7285 (9 June 2015).
15. A. McDavid, G. Finak and R. Gottardo, *Nat. Biotechnol.* **34**, 591 (9 June 2016).
16. M. T. Loones, Y. Chang and M. Morange, *Cell Stress Chaperones* **5**, 291 (October 2000).
17. C. Mirescu and E. Gould, *Hippocampus* **16**, 233 (2006).
18. H. Motohashi, J. A. Shavit, K. Igarashi, M. Yamamoto and J. D. Engel, *Nucleic Acids Res.* **25**, 2953 (1 August 1997).
19. E. V. Makeyev, J. Zhang, M. A. Carrasco and T. Maniatis, *Mol. Cell* **27**, 435 (3 August 2007).
20. Y. Zhao, M. Matsuo-Takasaki, I. Tsuboi, K. Kimura, G. T. Salazar, T. Yamashita and O. Ohneda, *Stem Cells Dev.* **23**, 2143 (15 September 2014).
21. M. Moreno, V. Fernández, J. M. Monllau, V. Borrell, C. Lerin and N. de la Iglesia, *Stem Cell Reports* **5**, 157 (11 August 2015).
22. L. A. M. Galea, M. D. Spritzer, J. M. Barker and J. L. Pawluski, *Hippocampus* **16**, 225 (2006).
23. T. Nguyen and S. Di Giovanni, *Int. J. Dev. Neurosci.* **26**, 141 (April 2008).
24. K. Campbell and C. Yau, *bioRxiv* (23 June 2016).
25. H. H. Chang, M. Hemberg, M. Barahona, D. E. Ingber and S. Huang, *Nature* **453**, 544 (22 May 2008).
26. H. Chen, G. Jiang and K. Yoshihira, Robust nonlinear dimensionality reduction for manifold learning, in *18th International Conference on Pattern Recognition (ICPR'06)*, (ieeexplore.ieee.org, 2006).

Figures

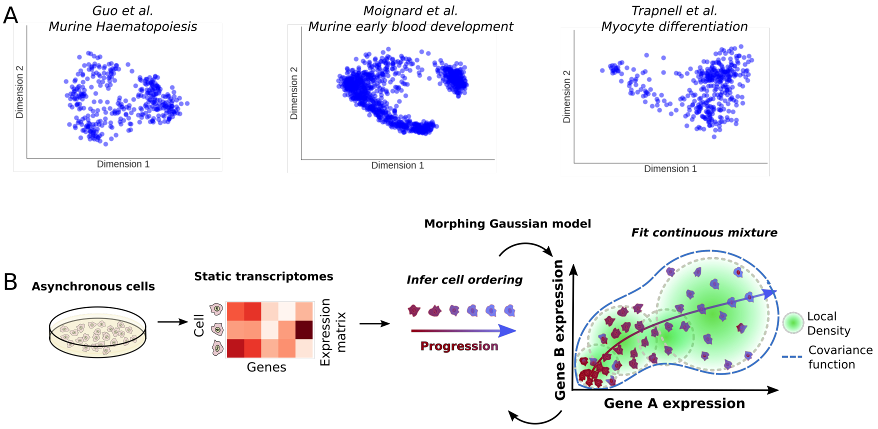


Fig. 1. A. Survey of three different single-cell transcriptomic studies. From left to right: murine haematopoiesis by Guo et al., early blood development by Moignard et al., and myocyte differentiation by Trapnell et al. B. Overview of the SCIMITAR method. Trajectory modeling with dynamic and correlated noise of static transcriptomes of asynchronous cells is achieved by iterating through optimal cell ordering and inference of a continuous set of Gaussian distributions in a morphing mixture of Gaussian models (see Methods in text).

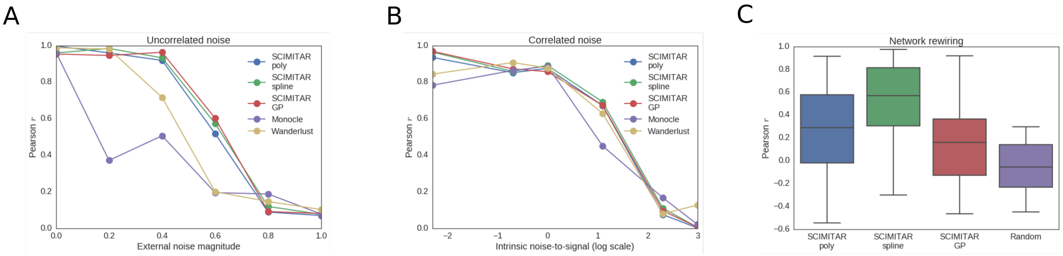


Fig. 2. SCIMITAR *in silico* benchmark. A. Cell ordering results for three functional classes of SCIMITAR (a third degree polynomial, a cubic spline, and Gaussian processes with squared exponential correlation model) and two state-of-the-art methods Monocle and Wanderlust in a setting with noise uncorrelated to the trajectory. B. Cell ordering results for noise correlated with the trajectory. C. Evaluation results of network rewiring across biological progression for SCIMITAR's three functional classes and random covariance functions.

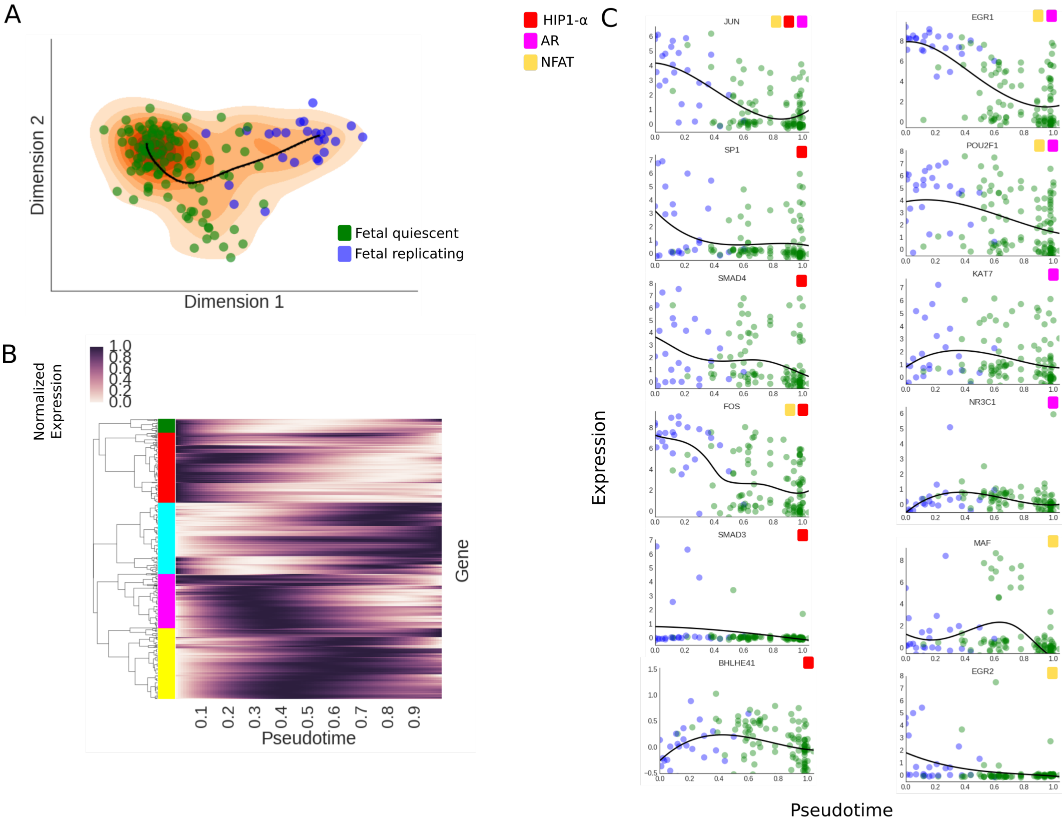


Fig. 3. A. SCIMITAR model for fetal neuron differentiation, projected to a 2-dimensional locally linear embedding. The data is plotted as circles in blue (fetal replicating neurons) and green (fetal quiescent nuerons) while the SCIMITAR model's mean is plotted in black and its projected PDF is plotted in orange. B. Normalized SCIMITAR model means for genes that were deemed progression associated across the progression, clustered into five different clusters using expression correlation throughout psuedotime. C. Expression levels of several genes from three central neurodifferentiation pathways: the HIF1α, NFAT, and Androgen Receptor (AR) pathways that were pinpointed by SCIMITAR associated progression tests.

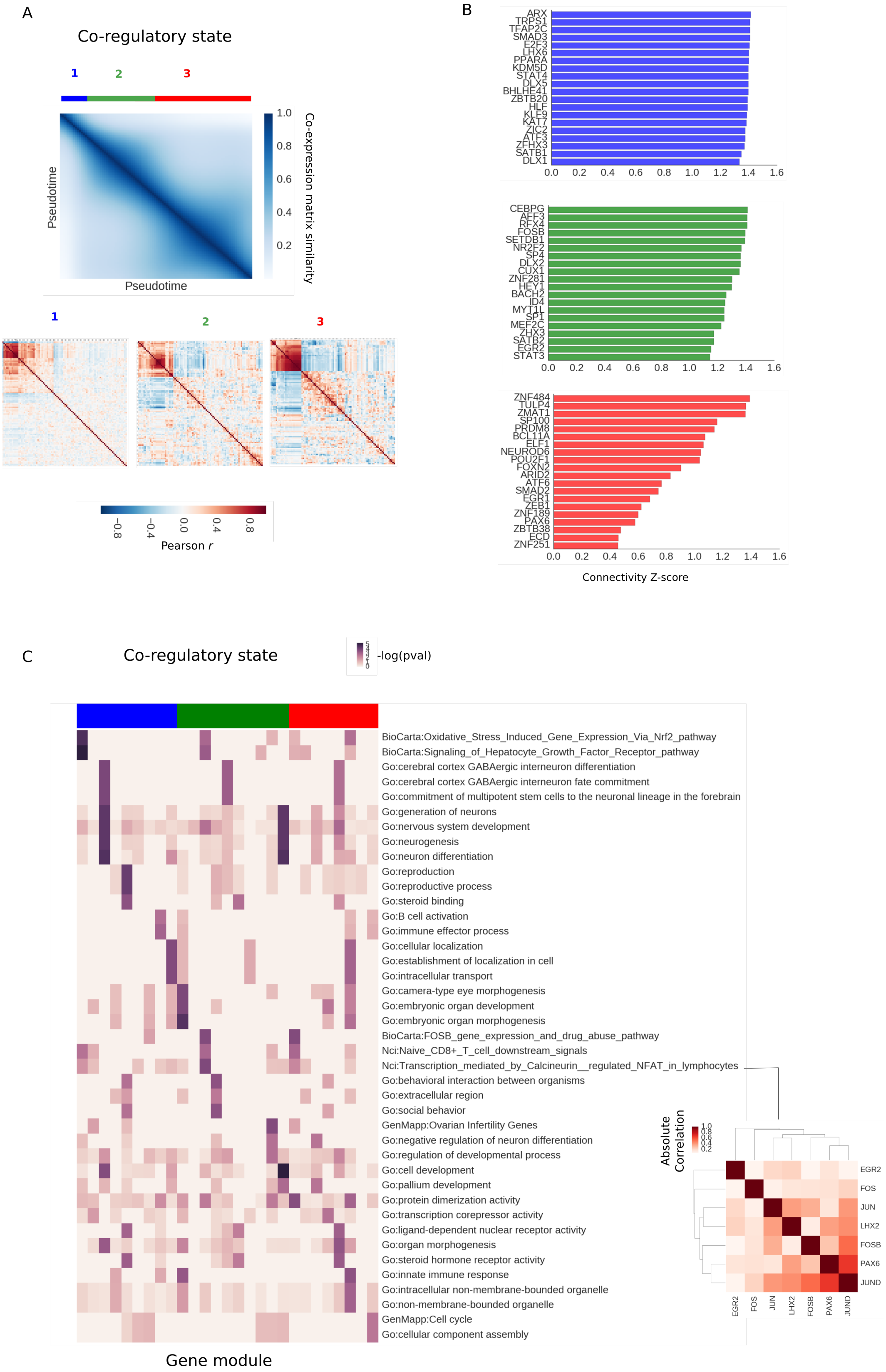


Fig. 4. A. Similarity matrix between co-expression matrices fitted in the SCIMITAR fetal neuron differentiation model across pseudotime. Three different co-regulatory states can be appreciated in the matrix, marked in blue, green, and red. B. Top 20 genes with the most gain-of-connectivity in each co-regulatory state alongside their log co-expression degree. C. Evolution of annotated modules. Each column is a module and each row is a gene annotation — enrichments are shown as $-log(p-value)$ in the heatmap. Column colors denote co-regulatory state. An NFAT-associated module of state 2 is highlighted in the red matrix

AN UPDATED DEBARCODING TOOL FOR MASS CYTOMETRY WITH CELL TYPE-SPECIFIC AND CELL SAMPLE-SPECIFIC STRINGENCY ADJUSTMENT

KRISTEN I. FREAD
Department of Biomedical Engineering, University of Virginia,
Charlottesville, VA 22903, USA
Email: kif5qw@virginia.edu

WILLIAM D. STRICKLAND
Department of Biomedical Sciences, University of Virginia,
Charlottesville, VA 22903, USA
Email: wds2df@virginia.edu

GARRY P. NOLAN
Department of Microbiology and Immunology, Stanford University,
Stanford, California 94305, USA
Email: gnolan@stanford.edu

ELI R. ZUNDER
Department of Biomedical Engineering, University of Virginia
Charlottesville, VA 22903, USA
Email: ezunder@virginia.edu

Pooled sample analysis by mass cytometry barcoding carries many advantages: reduced antibody consumption, increased sample throughput, removal of cell doublets, reduction of cross-contamination by sample carryover, and the elimination of tube-to-tube-variability in antibody staining. A single-cell debarcoding algorithm was previously developed to improve the accuracy and yield of sample deconvolution, but this method was limited to using fixed parameters for debarcoding stringency filtering, which could introduce cell-specific or sample-specific bias to cell yield in scenarios where barcode staining intensity and variance are not uniform across the pooled samples. To address this issue, we have updated the algorithm to output debarcoding parameters for every cell in the sample-assigned FCS files, which allows for visualization and analysis of these parameters via flow cytometry analysis software. This strategy can be used to detect cell type-specific and sample-specific effects on the underlying cell data that arise during the debarcoding process. An additional benefit to this strategy is the decoupling of barcode stringency filtering from the debarcoding and sample assignment process. This is accomplished by removing the stringency filters during sample assignment, and then filtering after the fact with 1- and 2-dimensional gating on the debarcoding parameters which are output with the FCS files. These data exploration strategies serve as an important quality check for barcoded mass cytometry datasets, and allow cell type and sample-specific stringency adjustment that can remove bias in cell yield introduced during the debarcoding process.

[*] This work is supported in part by the University of Virginia Department of Biomedical Engineering, CIRM Basic Biology II Grant RB2-01592, and NIH F32 GM093508-01.

1. Introduction

1.1. *Sample multiplexing for flow cytometry and mass cytometry with cell barcoding*

Sample multiplexing, also referred to as pooled sample analysis, is a general approach that has been applied to several biological assays, including ELISA immunoassay[1], next-generation DNA sequencing[2,3], fluorescence-based flow cytometry[4], and mass cytometry[5–7]. In this approach, individual samples are labeled with unique identifiers, and then pooled together for processing and measurement. These unique identifiers can be thought of as sample-specific barcodes. After processing and measurement, the pooled sample dataset is deconvolved using these barcodes to recover individual sample data for further analysis (Fig. 1A).

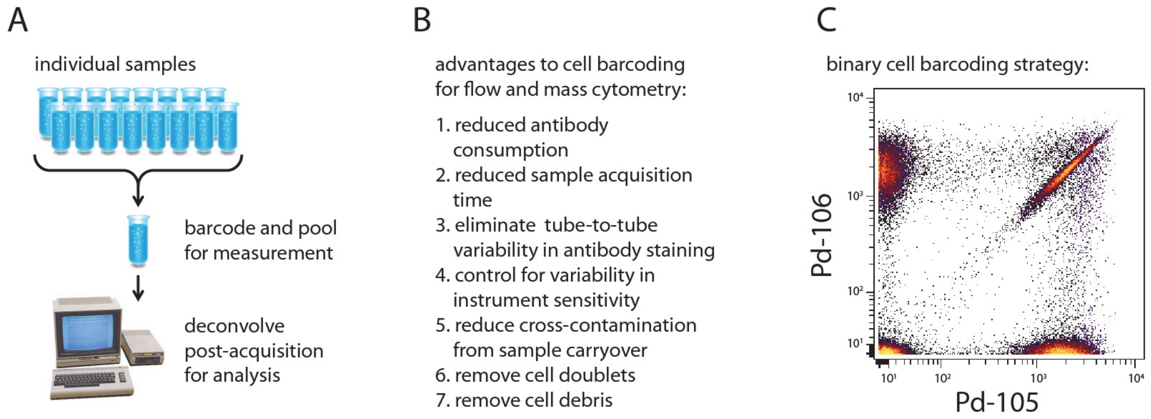

A
individual samples
barcode and pool for measurement
deconvolve post-acquisition for analysis
B
advantages to cell barcoding for flow and mass cytometry:
1. reduced antibody consumption
2. reduced sample acquisition time
3. eliminate tube-to-tube variability in antibody staining
4. control for variability in instrument sensitivity
5. reduce cross-contamination from sample carryover
6. remove cell doublets
7. remove cell debris
C
binary cell barcoding strategy:
Pd-106
Pd-105
10^4
10^3
10^2
10^1

Figure 1. Mass cytometry barcoding overview. (A) General strategy for pooled sample analysis. (B) Flow and mass cytometry-specific advantages to cell barcoding for pooled sample analysis. (C) Binary cell barcoding strategy for flow and mass cytometry, in which every cell is labeled either positively or negatively on barcode-dedicated channels.

The obvious advantages gained by sample multiplexing are a) reducing the time and resources required to analyze multiple samples, and b) improving the comparability between samples, because they are processed identically after pooling. Major advantages specific to flow cytometry and mass cytometry include reduced antibody consumption, increased sample acquisition rate, and the elimination of tube-to-tube variability in antibody staining conditions (Fig. 1B).

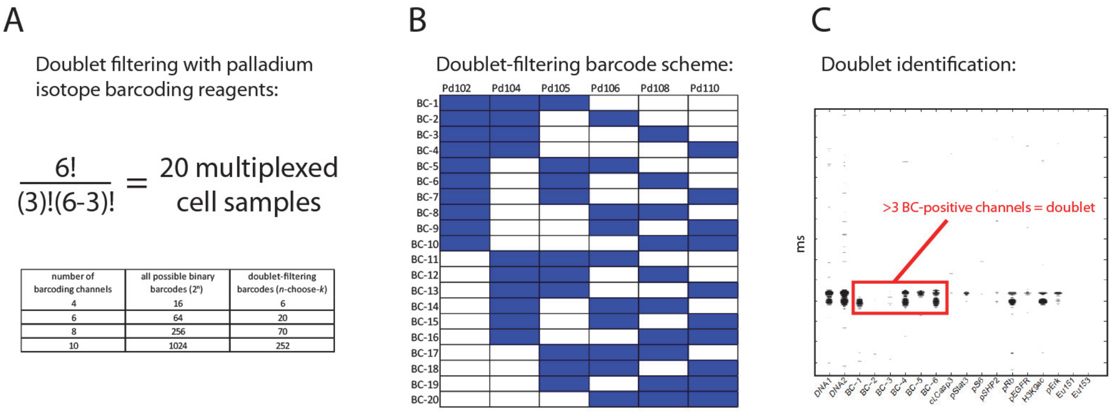
Sample multiplexing for fluorescence-based flow cytometry is performed with cell-reactive dyes that bind irreversibly to accessible nucleophiles on the cell[4]. These accessible nucleophiles include free thiols present on cysteine residues, and free amines present on lysine residues and at the N-terminus of proteins. While not strictly required, cell permeabilization greatly improves cell barcoding performance by increasing the number of accessible nucleophiles available on each cell. Multiple levels of fluorophore labeling can be achieved – previous studies have demonstrated 96-sample multiplexing with only 3 dedicated fluorescence channels: Alexa Fluor 700 (4 staining levels), Pacific Blue (4 staining levels), and Alexa Fluor 488 (6 staining levels)[4]. This multi-level staining approach allows for a high level of multiplexing with limited measurement channels, but relies on uniform levels of dye reactivity between all cell types and samples.

If there is considerable variability in labeling reagent uptake between cell types or sample types, a simpler binary cell barcoding approach can be applied to improve the fidelity of cell sample assignment at the deconvolution step. Because each cell sample is labeled either positively or negatively on each barcode-dedicated channel, the two populations are better separated with less potential for overlap (Fig. 1C). This approach is favored for mass cytometry cell barcoding, because the lanthanide and palladium-based barcode reagents react rapidly with cells even at 4°C[5,7], making the labeling reaction effectively stoichiometric and therefore more sensitive to variability between the samples in cell number, cell type/size, the presence of cellular debris, and residual bovine serum albumin (BSA) from the wash buffer. Using a binary barcode scheme requires more barcode-dedicated measurement channels than multi-level labeling, but allows for greater sample assignment fidelity during deconvolution while still permitting over 40 molecular measurements per cell with a staining panel made up of lanthanide-based mass cytometry antibodies, I127-IdU to mark S-phase cells[8], and cisplatin as a viability stain[9].

1.2. *Doublet-filtering cell barcode scheme*

Cell doublets (as well as triplets, quadruplets, and higher-order cell clusters) pose a significant challenge for single-cell analysis. When analyzing or performing fluorescence-activated cell sorting (FACS) on cell samples with known and well-defined cell types, such as whole blood or primary blood mononuclear cells (PBMCs), cell doublets are for the most part an annoyance that can be gated out using cell surface markers and light scatter properties. In certain defined settings, the study of cell doublets by flow cytometry has even proved to be illuminating with respect to cell adhesion and cell-cell interactions[10]. However, during exploratory analysis of uncharacterized cell samples and cell types, cell doublets are especially problematic, because they may be falsely interpreted as a novel cell type that shares the molecular characteristics of its two component cells.

Fluorescence-based flow cytometry has forward scatter (FSC) and side scatter (SSC) parameters that can be used to identify and remove cell doublets by two-dimensional gating[11]. Mass cytometry does not have a comparable measurement parameter, but a binary barcode scheme has been developed that can identify and remove cell doublets as well as higher-order clusters[7]. Instead of using every possible binary combination, this doublet-filtering barcode scheme uses a limited subset of binary combinations, such that any doublet combination will result in an "illegal" combination that is recognized as a doublet and removed from the dataset. A binary barcode scheme with n dedicated measurement channels will provide 2^n unique barcode combinations, but the doublet filtering binary barcode scheme only uses n-choose-k combinations, where $k = n/2$. 6 palladium isotopes are often used for cell barcoding because they are incompatible with the DTPA-based polymer used to label antibodies with lanthanide metals[12]. Instead of multiplexing 64 samples with all binary combinations (2^6) of the palladium isotopes, the doublet-filtering scheme only allows 20-sample multiplexing (6-choose-3) with palladium-based barcoding reagents (Fig. 2A). Because each barcode combination in this scheme is positive for exactly 3 palladium isotopes (Fig. 2B), any cell that is positive for 4 or more palladium isotopes will be identified as a cell doublet and removed from the dataset (Fig. 2C).


A
Doublet filtering with palladium isotope barcoding reagents:
B
Doublet-filtering barcode scheme:
C
Doublet identification:
>3 BC-positive channels = doublet
ms
DNA1
DNA2
BC-1
BC-2
BC-3
BC-4
BC-5
BC-6
cl,Casp3
pStat3
pS6
pSHP2
pRb
pEGFR
H3K9ac
pErk
Eu151
Eu153

$$\frac{6!}{(3)!(6-3)!} = \text{20 multiplexed cell samples}$$

number of barcoding channels	all possible binary barcodes (2^n)	doublet-filtering barcodes (*n*-choose-*k*)
4	16	6
6	64	20
8	256	70
10	1024	252

	Pd102	Pd104	Pd105	Pd106	Pd108	Pd110
BC-1	■	■	■			
BC-2	■	■		■		
BC-3	■	■			■	
BC-4	■	■				■
BC-5	■		■	■		
BC-6	■		■		■	
BC-7	■		■			■
BC-8	■			■	■	
BC-9	■			■		■
BC-10	■				■	■
BC-11		■	■	■		
BC-12		■	■		■	
BC-13		■	■			■
BC-14		■		■	■	
BC-15		■		■		■
BC-16		■			■	■
BC-17			■	■	■	
BC-18			■	■		■
BC-19			■		■	■
BC-20				■	■	■

Figure 2. Doublet filtering barcode scheme. (A) Sample multiplexing with exhaustive (2^n) and doublet-filtering (*n*-choose-*k*) barcode schemes. (B) Palladium isotope combinations for doublet-filtering barcode scheme. (C) Doublet identification by "illegal" barcode combination viewed in the mass trace scanning window of the mass cytometer.

This doublet-filtering scheme has become part of the standard mass cytometry workflow for many laboratories, and was incorporated into the third-generation Helios™ CyTOF® mass cytometer. Each user should consider the benefits of each approach for their experiment, because in some cases increased sample multiplexing could be more valuable than doublet removal. However, the recent description of ruthenium and osmium-based cell barcoding reagents suggests that high-level multiplexing with simultaneous doublet-filtering is now within reach[13], without having to give up any of the traditional mass cytometry measurement channels such as the lanthanide series metals.

1.3. *Sample deconvolution by sequential gating and Boolean gating strategies*

After pooled sample analysis, sample-specific barcodes are used to recover individual sample data for analysis. Different approaches have been applied to this deconvolution step, including cell type-specific gating followed by sequential 2-D barcode gating[4] or Boolean 1-D barcode gating[5]. Two drawbacks from these gating approaches are 1) time-consuming manual gating, and 2) the potential for cell loss or sample mis-assignment. In situations where the separation between barcoded populations is not large enough to be separable (Fig. 3A), the researcher must decide whether to throw out cells that reside in this intermediate space (Fig. 3B), or to split the populations and accept that some cells may be incorrectly assigned (Fig. 3C). In barcoded samples there is very often at least a small number of cells present in this intermediate zone that cannot be assigned to a specific sample by this debarcoding method. Usually this population is minor as shown in Figure 1C, but results like Figure 3A can also occur, particularly if the cell number in one or more samples is not estimated accurately resulting in uneven barcode labeling between samples. For this 1-D or 2-D gating strategy, boundaries can be drawn algorithmically using distribution shape and percentile cut-points, but the exact placement will depend on how the competing desires for cell yield vs. sample assignment accuracy.

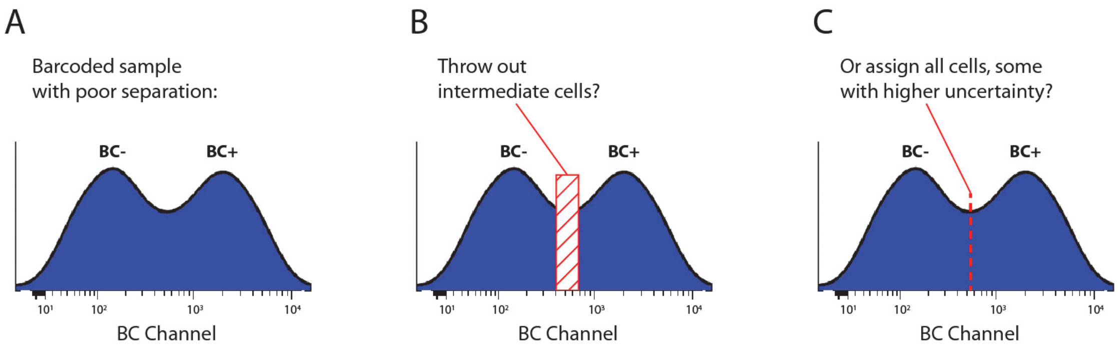


Figure 3. Traditional gating method on poorly-separated barcode sample. (A) Overlapping positive and negative barcode populations. (B) Intermediate cells can be thrown out to increase barcode deconvolution stringency. (C) Intermediate cells can be assigned to increase barcode deconvolution yield.

1.4. Sample deconvolution by single-cell debarcoding algorithm

In order to recover as many cells as possible in an automated and unbiased manner, a novel method for barcode deconvolution was previously developed, termed single-cell debarcoding[7]. This method is designed to perform especially well with the problematic “intermediate zone” cells. Instead of population-based gating, it looks at each cell individually, and asks “which sample barcode does this cell most closely resemble?” Sample assignment and the level of confidence associated with it is calculated by the separation distance between normalized positive and negative barcode channel measurements (Fig. 4A). The choice of separation distance used for this calculation depends on the binary barcode scheme being used. For exhaustive non-doublet-filtering barcode schemes, the largest separation distance is identified. For doublet-filtering barcode schemes, the distance between the top $n/2$ and bottom $n/2$ normalized barcode intensities is used, whether or not this is the largest separation distance present. If the separation distance is large, there is high confidence that the barcode sample assignment is correct. If the separation distance is small, there is low confidence that the barcode sample assignment is correct and these cells may be discarded depending on the deconvolution stringency desired.

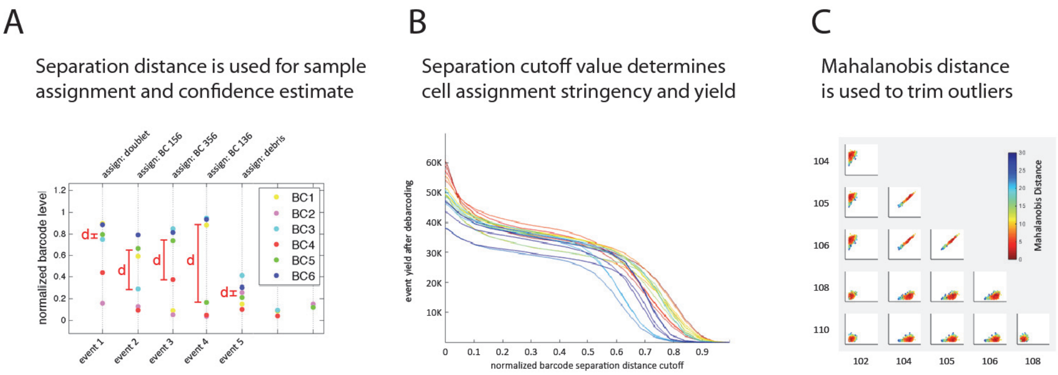


Figure 4. Single-cell debarcoding algorithm. (A) After normalization of the individual barode channel intensities, separation distances (indicated by a red line and the letter "d") are calculated for every cell. In this example, a 6-channel doublet-filtering barcode scheme was used. Therefore, event 1 does not receive a sample assignment because it appears to be a doublet with 4 positive barcode channels and a small separation distance between the top 3 and bottom 3 barcode intensities. Event 5 has low normalized intensities for all 6 barcode measurement channels, and therefore appears to be "debris." (B) The relationship between separation distance cutoff and debarcoder cell yield. Each colored line represents one of the 20 samples in a 6-metal, doublet-filtering, pooled sample dataset. Cell yield decreases with increasing separation distance cutoff stringency, but plateaus somewhat between 0.1 and 0.6. (C) Mahalanobis plots of every barcode-by-barcode biaxial plot for a single assigned cell sample. Every cell is colored by mahalanobis distance, from low (0-red) to high (30-blue).

The single-cell debarcoding software tool was released as a MATLAB standalone executable (https://github.com/nolanlab/single-cell-debarcoder)[7] that does not require a MATLAB installation (http://www.mathworks.com/products/compiler/). This software tool performs debarcoding and sample assignment in a semi-automated manner, presenting the user with visualizations that aid in the choice of two key debarcoding parameters: the separation distance cutoff which affects sample assignment stringency and cell yield (Fig. 4B), and the mahalanobis distance cutoff which is used to trim outliers (Fig. 4C). Standard practice for the single-cell debarcoder is to choose a separation cutoff distance that is as stringent as possible without severe cell loss, such as approximately 0.5 in Figure 4B. Most separation distance plots follow a similar trend, with a plateau in the center flanked by steep declines in the 0-0.1 range (debris and cell doublets) and approaching 1 (all cells will eventually fail the stringency test). Mahalanobis plots are more variable, depending on the mix of cell types in each sample. There is no specific rule or recommendation for setting the mahalanobis distance cutoff, but the default setting of 30 is a good starting point for 6-metal/20-sample palladium-barcoded samples. After the user selects values for the separation distance cutoff and mahalanobis distance cutoff parameters, the single-cell debarcoder tool outputs every deconvolved cell sample as an FCS file.

1.5 Limitations and drawbacks to using fixed-value debarcoding cutoff parameters

Applying the same parameter cutoffs to each sample while debarcoding as previously described[7] is not optimal, because each sample was barcode-stained individually and will therefore vary in barcode staining intensity and population-level variance. If all samples are similar (in terms of cell type, cell number, cellular debris, and residual BSA concentration) and the cell barcoding protocol is performed precisely, then barcode staining will be fairly uniform across every sample. Frequently this is not the case however, resulting in considerable variability of barcode staining between cell samples and large differences in sample behavior with respect to the debarcoding parameters, especially the normalized barcode separation distance cutoff (Fig. 5A).

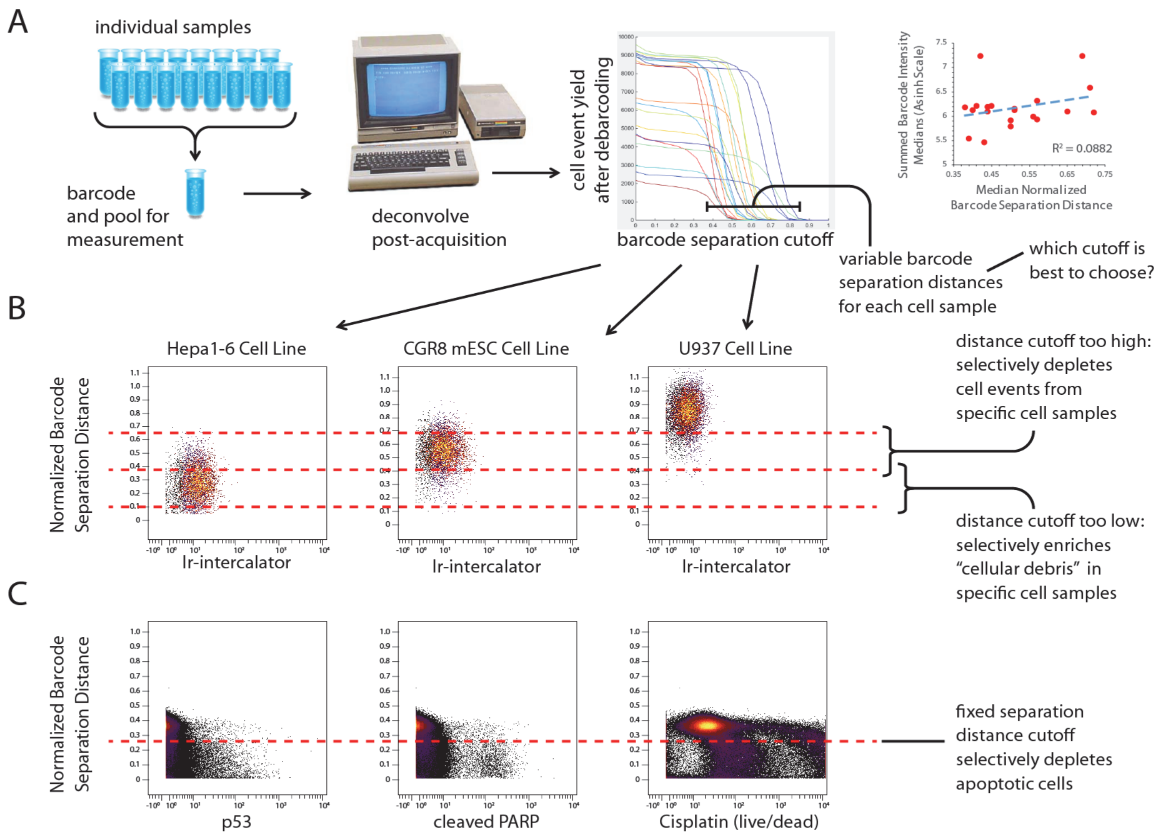


Figure 5. Cell barcode variability and its consequences. (A) Cell samples representing 20 different cell types were barcoded by the 6-palladium doublet-filtering method and then pooled for analysis. The amount of barcode reagent added to each sample was adjusted according to cell number in each sample to normalize barcode staining intensity. Variability in barcode separation distance was observed between the samples, but is only weakly correlated to barcode staining intensity, as measured by the 6-metal summed barcode intensity medians for each sample. (B) Three of the twenty barcoded cell samples, which show highly variable barcode separation distance levels, precluding a single optimal cutoff value. (C) Apoptotic cells with elevated levels of p53, cleaved-PARP, and cisplatin labeling have reduced barcode separation distance, and could be unintentionally discarded from analysis with a typical debarcoding workflow.

In this scenario, no single cutoff value for barcode separation distance is optimal for every sample, forcing the researcher to choose between depleting cells of interest in some samples, or enriching for cellular debris in other samples (Fig. 5B). In addition to cross-sample differences, different cell types can be depleted or enriched within a single sample due to differences in barcode staining behavior based on cell size, cell identity, or cell state (Fig. 5C). These sample-specific and cell type-specific effects are usually minimal, but have the potential to introduce bias into the analysis and conclusions drawn from barcoded mass cytometry experiments. Therefore, each barcoded dataset should be investigated to detect the extent of these effects, and correct for them if necessary.

2. Methods

2.1 Output single-cell debarcoding parameters with each FCS file for visualization and analysis

With the previously released debarcoding tool[7], investigating the possibility for barcode-related enrichment or depletion of specific samples and cell types required laborious and time-consuming back and forth between rounds of debarcoding and FCS file analysis. Side-by-side comparison of FCS files debarcoded with iterative values for the debarcoding parameters was necessary to detect cell type or sample-specific effects. To obviate the need for this slow and inefficient analysis, we have updated the debarcoding software tool to output the debarcoding parameter values for each cell as additional data columns in the FCS file. This update allows for visualization of the barcode parameters, and analysis of how they interact with the other measured parameters and cell types of interest. The MATLAB source code for the updated software tool as well as pre-compiled executable files that do not require MATLAB installation are available to download at https://github.com/zunderlab/single-cell-debarcoder.

2.2 Post-assignment application of debarcode stringency filter and outlier trimming

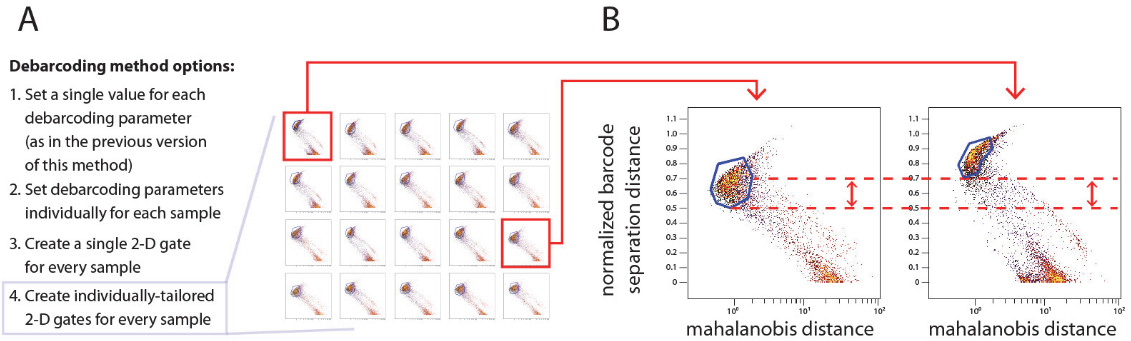


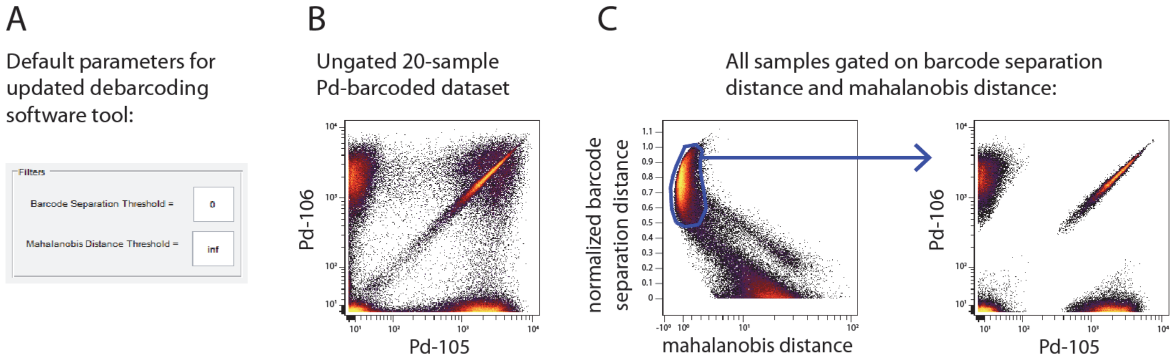
Figure 6. Sample-specific stringency adjustment by individual gating on debarcode parameters. (A) FCS output of the debarcoding parameters allows different strategies for stringency filtering. The option for individually-tailored 2-D gating on normalized barcode separation distance and mahalanobis distance is presented. (B) Two cell samples from Fig. 6A are highlighted to illustrate the population-level differences in barcode parameters between samples.

In addition to visualization and analysis, outputting the debarcode parameters in the FCS file has another practical benefit: stringency filters can be turned off during the debarcoding step and applied after the fact instead. This gives the user flexibility in their choice of stringency filtering: they may apply fixed parameters as in the previous version of this method, or perform sample-specific two-dimensional gating on the debarcode parameters (Fig. 6A). Whichever method is chosen, the user is given the tools to explore these parameters and their relationship to other cell measurements, which will aid in the choice of filtering strategy and its implementation. Some users may prefer fixed parameter stringency filtering because it is simpler and less time consuming, but users with complex, variable samples should consider individually-tailored stringency filtering, which requires more time to implement but helps prevent the introduction of sample-specific biases (Fig. 6B).

3. Results

3.1. Precision Debarcode Stringency Filtering

The newly updated single-cell debarcoding software tool functions identically to the previous version, but with two additions: 1) values for the normalized barcode separation distance and mahalanobis distance are output for every cell, and 2) default parameters for debarcoding are set as "barcode separation threshold = 0" and "mahalanobis distance threshold = inf" (Fig. 7A). These default parameters ensure that every cell is assigned to a sample for FCS output and can be filtered after the fact. This differs from the fixed-parameter filtering which took place at the debarcoding step in the previous software version, resulting in an additional FCS output for unassigned cell events. Outputting the entire dataset (Fig. 7B) with this new method allows for precision stringency filtering by gating on the debarcode parameters (Fig. 7C). This gating will typically be performed using flow cytometry/FCS analysis software, and can be done iteratively and in combination with more fundamental cell type and dataset-specific analyses.



Barcode Separation Threshold = 0

Mahalanobis Distance Threshold = inf

Figure 7. Debarcode stringency gating overview. (A) 20-sample Pd-based doublet-filtering barcode sample, ungated. (B) Barcode stringency trimming by 2-D gating on the normalized barcode separation distance vs. mahalanobis distance.

3.2 Identification and Reduction of Debarcoding-induced Sample Bias

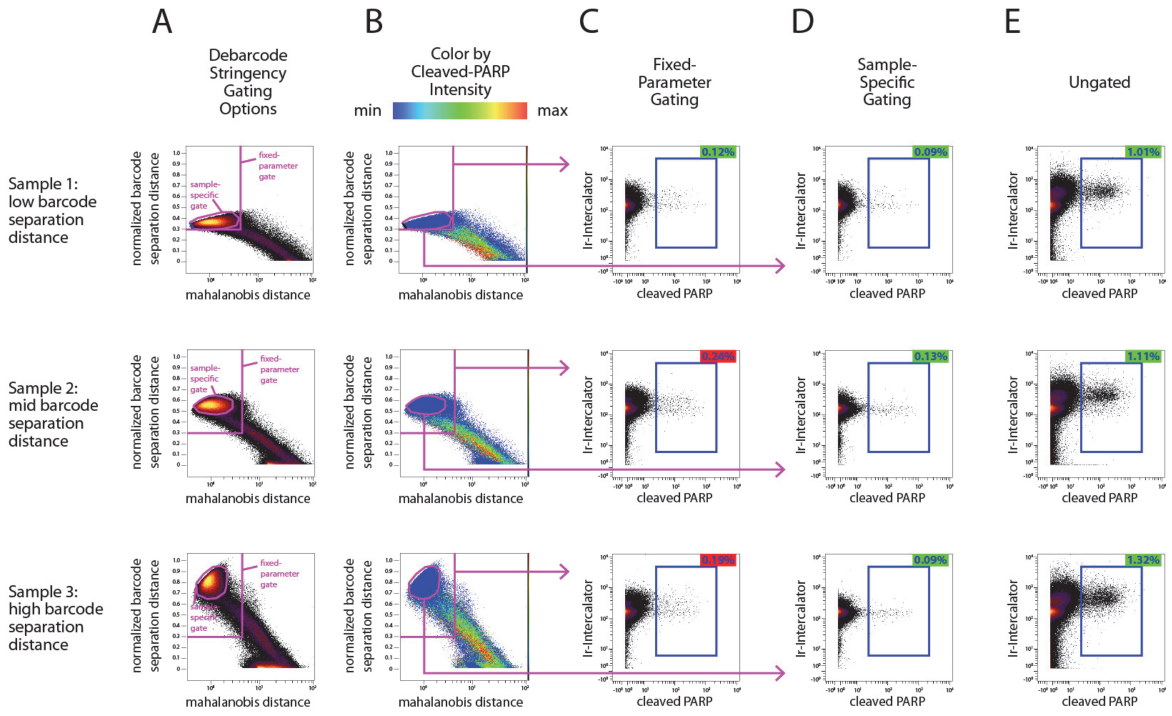


Figure 8. Sample-specific debarcode stringency gating reduces unbalanced enrichment of cleaved-PARP-positive cells. (A) Two options for stringency gating are displayed in magenta: fixed-parameter and sample-specific. (B) Cleaved-PARP intensity color scale applied to the plot from Figure 8A reveals that fewer cells with elevated cleaved-PARP levels fall within the fixed-parameter gate in sample 1 compared to samples 2 and 3. (C-E) The percentages of cleaved-PARP-positive cells present in the bulk-gated, individually-gated, and ungated populations.

Sample-specific debarcode gating on the normalized separation distance and mahalanobis parameters provides the greatest advantage over the previously used fixed-parameter debarcoding method when there is variability in the debarcode parameters between samples (Fig. 8A), which can lead to uneven distribution of specific cell types across the debarcoded samples. Cells with elevated cleaved-PARP levels are associated with lower separation distance and higher mahalanobis distance (Fig. 8B). This leads to disproportionate enrichment for cleaved-PARP cells in some samples when using fixed-parameter debarcode filtering (Fig. 8C), but is ameliorated by sample-specific gating (Fig. 8D), which more closely matches the ungated sample ratios (Fig. 8E).

4. Discussion

This updated method for single-cell mass cytometry debarcoding allows for visualization and analysis of the debarcoding parameters, and how they specifically relate to every other cell measurement. This can be used to detect any cell type-specific or sample-specific effect of the debarcoding process on the underlying cell data of interest. The source code and Win/Mac executable software are available to download from https://github.com/zunderlab/single-cell-debarcoder. We recommend that this analysis be performed on every debarcoded dataset as a data quality check, particularly when mixed cell types and sample types are barcoded together. In addition to data quality verification, the output debarcoding parameters in every assigned FCS file can be used to guide sample-specific stringency filtering that can be performed after the fact rather than during the debarcoding process. This allows multiple stringency levels to be tested rapidly using flow cytometry/FCS analysis software, where multiple iterations of 1-D or 2-D gating can be used while monitoring the effect on cell type-specific and sample-specific cell yield as well as overall data quality. One limitation to this method is that stringency filtering is not automated, and currently relies on hand-drawn gates. While this method is optimally used to reduce cell yield and enrichment bias between cell samples and cell types that vary in barcode staining behavior, sample-specific or cell type-specific manual gating has the potential to introduce bias. As with any other hand-drawn gating analysis, the barcode gating strategy should always be presented in addition to further analysis in order to mitigate this potential for user-introduced bias. In the future, stringency filtering could be automated with sequential, percentile-based gating steps; or more complex computational methods.

5. Acknowledgments

This work was supported by the University of Virginia Department of Biomedical Engineering, CIRM RB2-01592, and NIH F32 GM093508-01.

6. References

1. Fulton et al. *Clinical Chemistry* **43,** 1749–1756 (1997).
2. Meyer et al. *Nucl. Acids Res.* **35,** e97 (2007).
3. Parameswaran *et al*. *Nucl. Acids Res.* **35,** e130 (2007).
4. Krutzik, P. O. & Nolan, G. P. *Nature Methods* **3,** 361–368 (2006).
5. Bodenmiller *et al*. *Nature Biotechnology* (2012). doi:10.1038/nbt.2317
6. Behbehani *et al*. *Cytometry* n/a-n/a (2014). doi:10.1002/cyto.a.22573
7. Zunder *et al. Nat. Protocols* **10,** 316–333 (2015).
8. Behbehani *et al*. *Cytometry Part A* **81A,** 552–566 (2012).
9. Fienberg *et al*. *Cytometry Part A* **81A,** 467–475 (2012).
10. Snippert *et al. Cell* **143,** 134–144 (2010).
11. Hoffman, R. A. in *Current Protocols in Cytometry* (John Wiley & Sons, Inc., 2001).
12. Majonis *et al*. *Biomacromolecules* **12,** 3997–4010 (2011).
13. Catena *et al*. *Cytometry* **89,** 491–497 (2016).

MAPPING NEURONAL CELL TYPES USING INTEGRATIVE MULTI-SPECIES MODELING OF HUMAN AND MOUSE SINGLE CELL RNA SEQUENCING[*]

TRAVIS JOHNSON MS,
Dept. Biomedical Informatics, Ohio State University,
250 Lincoln Tower, 1800 Cannon Dr. Columbus, Ohio, 43210
Travis.Johnson@osumc.edu

ZACHARY ABRAMS PhD,
Dept. Biomedical Informatics, Ohio State University,
250 Lincoln Tower, 1800 Cannon Dr. Columbus, Ohio, 43210
Zachary.Abrams@osumc.edu

YAN ZHANG PhD,
Dept. Biomedical Informatics, Ohio State University,
250 Lincoln Tower, 1800 Cannon Dr. Columbus, Ohio, 43210
Yan.Zhang@osumc.edu

KUN HUANG PhD,
Dept. Biomedical Informatics, Ohio State University,
250 Lincoln Tower, 1800 Cannon Dr. Columbus, Ohio, 43210
Kun.Huang@osumc.edu

Mouse brain transcriptomic studies are important in the understanding of the structural heterogeneity in the brain. However, it is not well understood how cell types in the mouse brain relate to human brain cell types on a cellular level. We propose that it is possible with single cell granularity to find concordant genes between mouse and human and that these genes can be used to separate cell types across species. We show that a set of concordant genes can be algorithmically derived from a combination of human and mouse single cell sequencing data. Using this gene set, we show that similar cell types shared between mouse and human cluster together. Furthermore we find that previously unclassified human cells can be mapped to the glial/vascular cell type by integrating mouse cell type expression profiles.

[*] This work is partially supported by RGP0053 of the Human Frontier Science Program

1. Introduction

Mouse models are an important part of biomedical research and are routinely used as a stepping-stone towards treatments for humans – gleaning knowledge from high-throughput low risk experiments. Translating this knowledge requires a firm understanding of similarities between these two species [1-2]. Homologous genes exist between these species and these genes often play similar roles in the brain [3]. However, the biochemical pathways within each species have subtle to extreme differences leading to subsets of homologous genes without exact mechanistic overlap in the brain [4]. To address the issue of identifying functionally similar homologous genes we propose the concept of concordant genes defined as gene homologs that mechanistically behave similarly between two species [5]. Specifically, we hypothesize that concordant genes between mouse and human exist and that those genes can be algorithmically derived from combined mouse-human data. We also hypothesize that based off of these concordant genes we can determine cell type matching between mouse and human. Specifically in this study we focus on the comparison of brain cell gene expression profiles between mouse and human to identify concordant gene expression patterns in the brain tissue associated with different cell types taking advantages of recent development in single cell transcriptomics for brain cells. We hope that the single cell granularity of these comparisons will augment the tissue level comparisons of the human and mouse brain transcriptome [6].

RNA sequencing (RNA-Seq) in the past has been used to study brain structure, development, and disease [7]. Recently RNA-Seq has become more granular in the form of single cell RNA sequencing (scRNA-Seq) which is an important tool in the study of tissue heterogeneity due to its unique ability to characterize transcriptomes at the cellular level [8]. Recent advances in single cell transcriptomics in the brain have provided researchers with an influx of new data spanning different brain regions, diseases, and species [9]. Specifically, the Linnarsson group amassed a large single cell dataset from the mouse cortex and hippocampus which was clustered into multiple cell types based expression profiles [10]. Subsequent to the mouse single cell transcriptomic study, the Zhang group created a large human brain scRNA-Seq dataset from postmortem brain tissue and clustered the cells into unique cell types based on expression profiles [11]. Because of the availability of both datasets we believe that in-depth comparative analyses of these two datasets is fundamental to our understanding of neuronal cell types, the distribution of these cell types, and the evolution of brain anatomy in these two species. Furthermore a clear understanding of concordant genes in both human and mouse provides valuable information on how mouse studies can be translated to human research. We provide a methodology and gene set that can be used for these comparative studies and hopefully for future translational research. We demonstrate the method by not only identifying concordant cell types between mouse and human brains with the same set of concordant feature genes, but also matching un-categorized cells in the human brain to a salient cell type based on mouse brain information.

2. Methods

2.1. *Data normalization and cleaning*

The mouse scRNA-Seq unique molecular identifier (UMI) counts [12] were downloaded from the Data section of the Linnarsson lab website (http://linnarssonlab.org/) and human scRNA-Seq transcripts per million (TPM) data was downloaded from the Links section of the SCAP-T website (scap-t.org). Since these data files contain various numbers of genes with different order, we preprocessed the files by scanning matching gene symbols between files then sorting the gene symbols so that the orders were consistent. While this process may not be able to identify all homologous genes, it provides a large list for us to extract concordant genes. The shared gene symbols in the human and mouse datasets were retained for further study (Figure 1). Within the human dataset there were genes that were originally left out of analysis by the original authors due to low expression, resulting in some cells with low number of expressed genes. Because of this, such human cells as well as human cells without annotation in the metadata were also removed from further analysis, resulting in 3,086 human cells each containing 13,355 genes. The mouse dataset resulted in 3,005 cells each containing expression values from 13,355 genes. Both human and mouse data then were transformed into comparable units. Each dataset was log2 transformed and the expression values converted into the within cell z-scores.

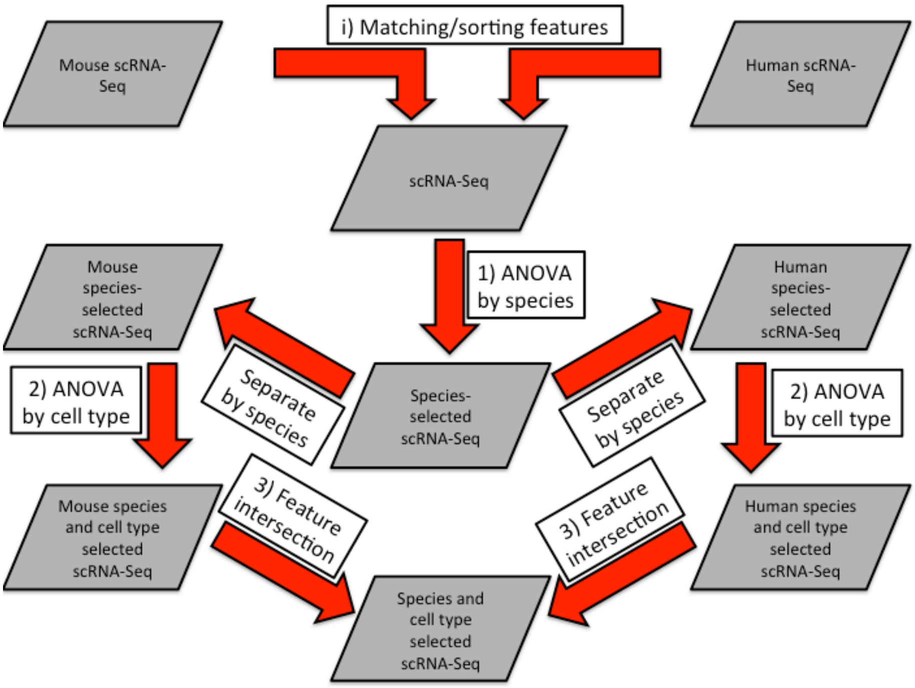

i) Matching/sorting features
Mouse scRNA-Seq
Human scRNA-Seq
scRNA-Seq
1) ANOVA by species
Mouse species-selected scRNA-Seq
Human species-selected scRNA-Seq
Separate by species
Species-selected scRNA-Seq
Separate by species
2) ANOVA by cell type
2) ANOVA by cell type
Mouse species and cell type selected scRNA-Seq
3) Feature intersection
3) Feature intersection
Human species and cell type selected scRNA-Seq
Species and cell type selected scRNA-Seq

Figure 1. Workflow of data normalization (i) and three step feature selection method (1-3).

2.2. *Feature selection*

We developed a three-step approach to find concordant genes between mouse and human based on gene expression profiles (Figure 1). This feature selection was performed to identify genes that were informative at separating cell type but uninformative at separating mouse from human cells. Genes that meet this criterion would be more useful at identifying similar cell types across species.

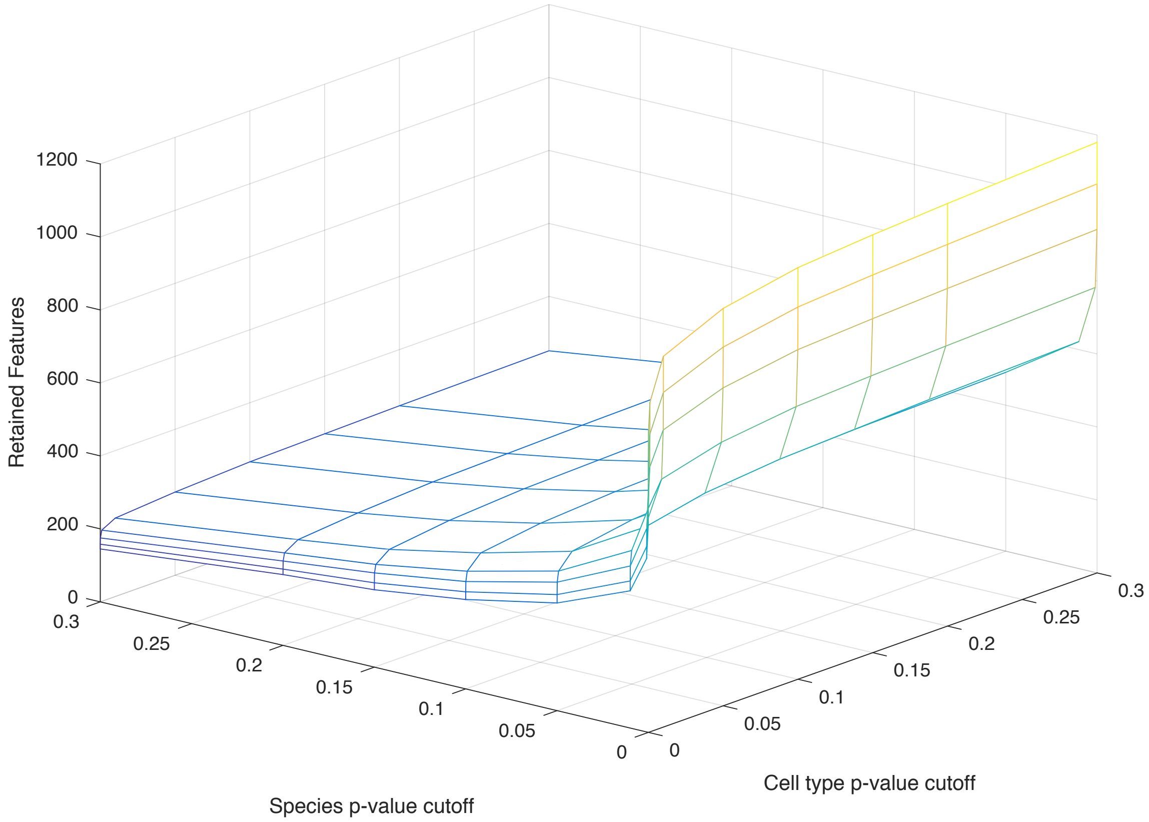


Figure 2. Number of retained features as a function of p-value cutoffs.

First, the human and mouse data matrices were concatenated such that the first 3086 columns consisted of human cells and the last 3005 columns consisted of mouse cells. For each gene in the data matrix, a one-way ANOVA was performed grouped by species to detect genes with significantly different expression level between human and mouse. Only genes with p-values larger than 0.1 were kept. This was done to remove genes that would separate cells by species. Because we are removing the significant genes from our gene set in Step 1, a greater threshold makes our criterion for retaining genes more strict than using a standard significance level. Second, the human and mouse matrices were separated and in each separate matrix a one-way ANOVA was performed on the remaining genes grouped by cell type label and using a threshold p-value of 0.01 – any genes found with a p-value of 0.01 or less were retained. The 0.01 threshold was used to provide stricter criteria for retained genes that were informative about cell type. The 0.1 and 0.01 p-value cutoffs used in the feature selection method are near the inflection point of retained features as a function of cutoff p-value (Figure 2). Third, the intersection of retained genes from human and mouse were retained in the final dataset such that genes that existed in both human and mouse gene sets after Step 2 were retained in the final combined mouse-human gene set.

To compare the differences between cell types and in concordance with previous single cell studies [13], principal component analysis (PCA) was applied to the human and mouse datasets prior to feature selection. The first 2 principal components were then plotted to visually show the differences in cell types and species (Figure 3). After feature selection, principal cross-species cell-type clusters can be viewed in the PCA of the first two principal components colored by species (left) and cell type (right) (Figure 4).

2.3. *Functional annotation of retained concordant genes*

When selecting features, it is important to study the relation of these feature/gene sets to the functional, anatomic, and phenotypic relationships that are being selected for. If there are functional relationships related to a phenotype, then the feature selection method targeting that phenotype is likely more robust. The retained genes from the feature selection step were used as input for the DAVID functional annotation software [13-14]. The functional annotation clusters were reviewed for over represented terms that can be attributed to neural pathways and cell types. We display the three most highly enriched terms within the three most highly enriched clusters from the DAVID functional annotation clustering (Table 1).

2.4. *Clustering cells using Gaussian mixture models*

Gaussian mixture models are effective in clustering microarray expression profiles [16]. We apply Gaussian mixed models (GMMs) in the mouse and human scRNA-Seq data to cluster the cells into principal cell types and to compare the relative proportions of human and mouse cells within each cluster. To perform the GMM we used the first two principal components, the same components used in the PCA plot of cell types. Four GMMs were fit to the data with two, three, four and five components respectively. The cells were clustered into three major cluster using the three component GMM fit in concordance with the three major cell types present in the human dataset. The remaining GMM fits were used in comparison against the three-component GMM fit.

Principal cell types of the mouse and human labels were compared in the PCA space to determine the most similar cell types between both species. To quantitate the mouse-human overlap the mouse and human data were split into three groups from the three major cell types in the original publications. Human cells were split into 3 major groups from their original labels [11]. All "Int" labeled cells were considered Interneuron. All "Ex" labeled cells were considered pyramidal. All "NoN" (No Nomenclature) labeled cells from a C1 Fluidigm chip with reduced mapping rates were without a biologically derived label but were considered a singular group. Similarly, mouse cells were also split based on cell type label mapping to GMM clusters [10]. All cells labeled Interneurons were still considered Interneurons. All S1 Pyramidal and CA1 Pyramidal were considered Pyramidal. All Oligodendrocytes, Microglia, Endothelial, Astrocytes, Ependymal and Mural were considered Glial/Vascular cells. All human and mouse cells that were contained within each GMM cluster were compared by the their original cell type labels to the labels of the GMM cluster. For each cluster a fisher exact test was conducted to calculate the odds ratios and confidence intervals between published cell type labels and GMM predicted cell types.

The VennX package in MATLAB was used to convert the cell type labels into Venn Diagrams to show overlap with both three component GMM predicted cell types and original mouse/human cell type labels from their original publications.

3. Results

3.1. *Feature selection*

Prior to feature selection the human and mouse cells created two clusters separated by species. The mouse cells formed sub-clusters within the major mouse clustering of cells. The human cells formed one main cluster with little differentiation (Figure 3).

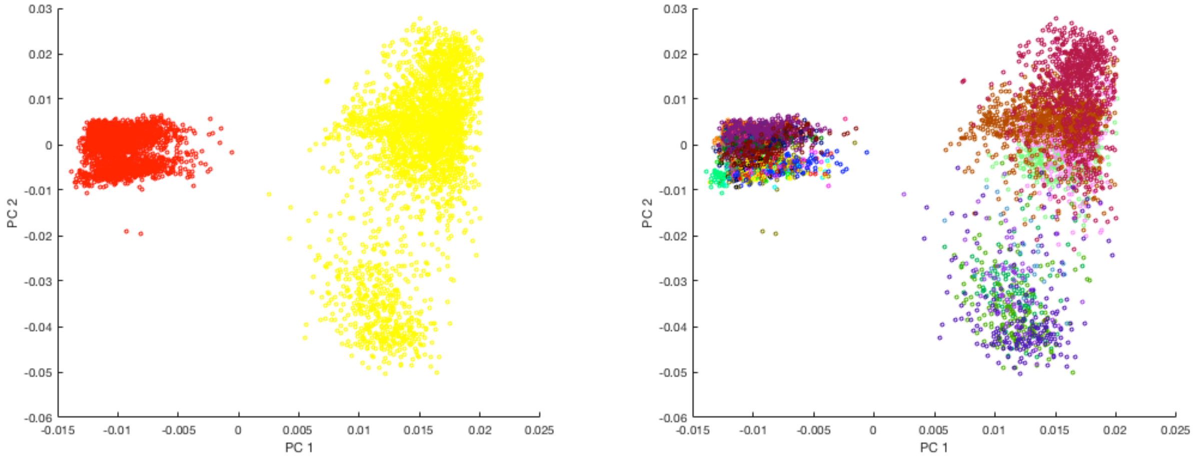


Figure 3. PCA of all human and mouse cells after normalization/cleaning. Left is colored by species, mouse (yellow) and human (red). Right is colored by cell type (36 cell types).

After feature selection, 358 concordant genes were retained, which are informative in terms of distinguishing cell types and uninformative in terms of separating species. As a result, human and mouse cells were no longer completely separate from each other. The mouse cell types still have more variability than the human cell types in the PCA space but cells from both species are contained within the same major clusters of cells (Figure 4).

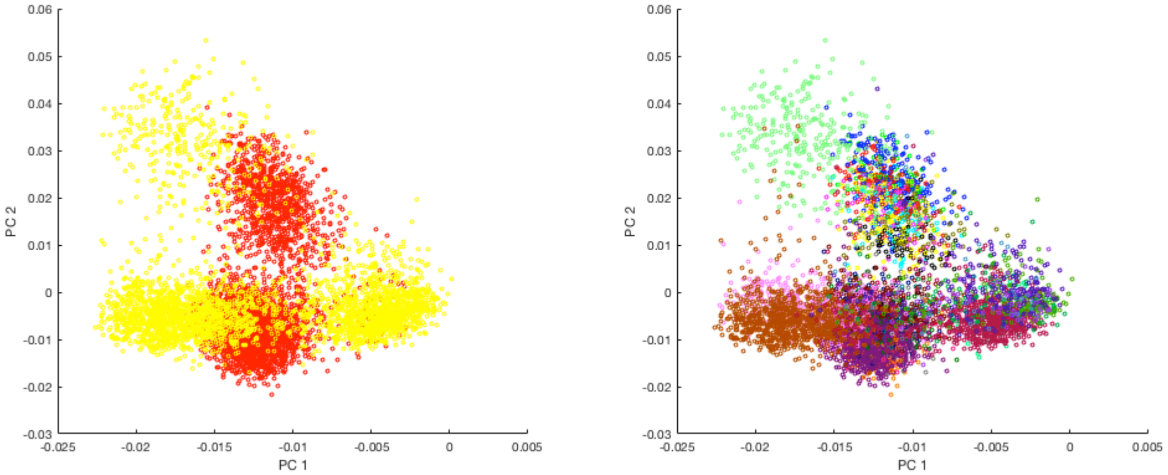


Figure 4. PCA of all human and mouse cells after normalization/cleaning and feature selection. Left is colored by species, mouse (yellow) and human (red). Right is colored by cell type (36 cell types).

3.2. *Functional annotation of concordant genes*

Functional annotation analysis of the concordant gene set revealed GO terms related to binding, ion transport and neural cells. The third most highly enriched annotation cluster was that of the GO terms axon, cell projection and neuron projection with an enrichment score of 1.57 (Table 1). Cluster 7 (not displayed) also contained many neuron related ontology terms.

Table 1. Functional annotation clustering using DAVID. Shown below are the three most highly enriched clusters and three most highly enriched terms within each cluster.

Category	Term	PValue	Fold Enrichment	Bonferroni
Annotation Cluster 1	Enrichment Score: 1.670			
SP_PIR_KEYWORDS	atp-binding	0.008	1.573	0.939
SP_PIR_KEYWORDS	nucleotide-binding	0.010	1.478	0.969
GOTERM_MF_FAT	GO:0032559~adenyl ribonucleotide binding	0.012	1.463	0.997
Annotation Cluster 2	Enrichment Score: 1.594			
GOTERM_BP_FAT	GO:0006826~iron ion transport	0.002	8.868	0.979
SP_PIR_KEYWORDS	iron transport	0.007	10.076	0.919
GOTERM_BP_FAT	GO:0000041~transition metal ion transport	0.012	4.347	1.000
Annotation Cluster 3	Enrichment Score: 1.568			
GOTERM_CC_FAT	GO:0030424~axon	0.010	3.027	0.946
GOTERM_CC_FAT	GO:0042995~cell projection	0.036	1.611	1.000
GOTERM_CC_FAT	GO:0043005~neuron projection	0.056	1.877	1.000

3.3. *Clustering cells using gaussian mixture models*

Gaussian mixture models showed major patterns within the cell profiles. Interneurons from both human and mouse (red and yellow respectively)(Figure 5) clustered in the same GMM. Whereas human pyramidal/projection neurons clustered (green) clustered with the remaining 2 cell types in mouse (S1 pyramidal, CA1 pyramidal). It is also worth consideration that the non-biologically labeled “NoN” human cell types in purple are mapped to a third cluster that begins to appear at 3 GMM components that contains the remaining 6 mouse cell types (mural, endothelial, microglia, ependymal, astrocytes, oligodendrocytes) (Figure 5).

The GMM clustering using three components (BIC = -9.08×10^4) split the cells into three groups that can be roughly defined as Interneurons (red), Pyramidal cells (green) and Glial/Vascular cell types (blue) (Figure 6: Top left). After identifying these three groups and comparing the mouse and human labels the GMM labels it was found that these three groups, Interneurons, Pryamidal cells, and Glial/Vascular cells are very closely mapped between both mouse and human. Also the “NoN” cell type cluster found in the human scRNA-Seq paper were clearly and uniquely clustered with the mouse Glial/Vascular cells (Figure 6 bottom right) with no significant difference between Glial/Vascular mouse cells and “NoN” human cells on PC 1 p-value = 0.41.

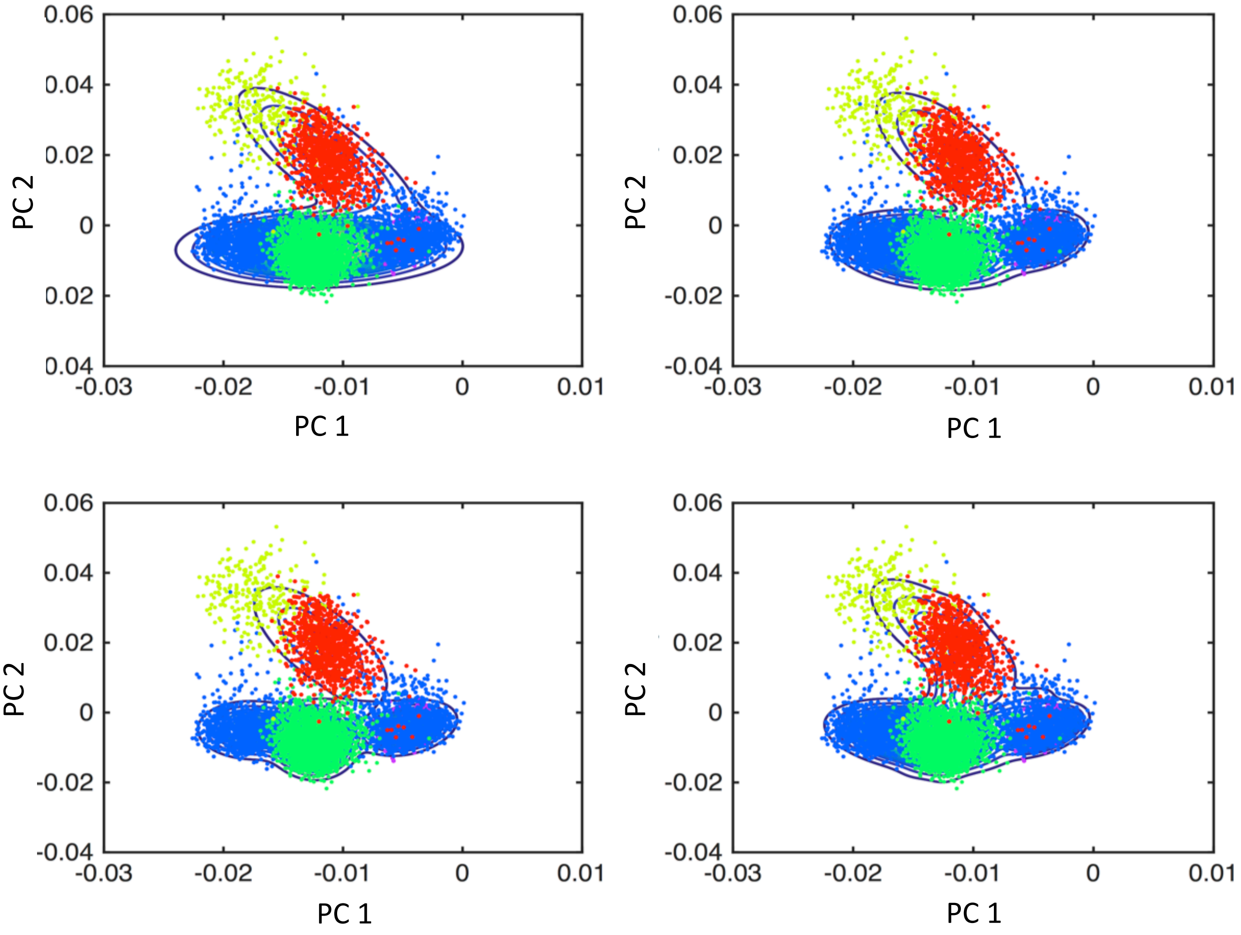


Figure 5. Gaussian mixture model clustering of human and mouse cell types where top left: two components, top right: three components, bottom left: four components and bottom right: five components.

The cell types predicted by the three component GMM were representative of the original cell type labels. The interneuron GMM had an odds ratio of 2.00×10^3 and confidence interval of (1.16×10^3,3.46×10^3), the pyramidal GMM had an odds ratio of 9.93×10^2 and a confidence interval of (6.84×10^2, 1.44×10^3), and the glial/vascular GMM had an odds ratio of 1.15×10^2 and a confidence interval of (91.34, 1.44×10^2) (Figure 6). The GMM cluster for glial/vascular cells had a higher false negative rate than the other GMM clusters due to incorrect clustering of glial/vascular labeled mouse cells.

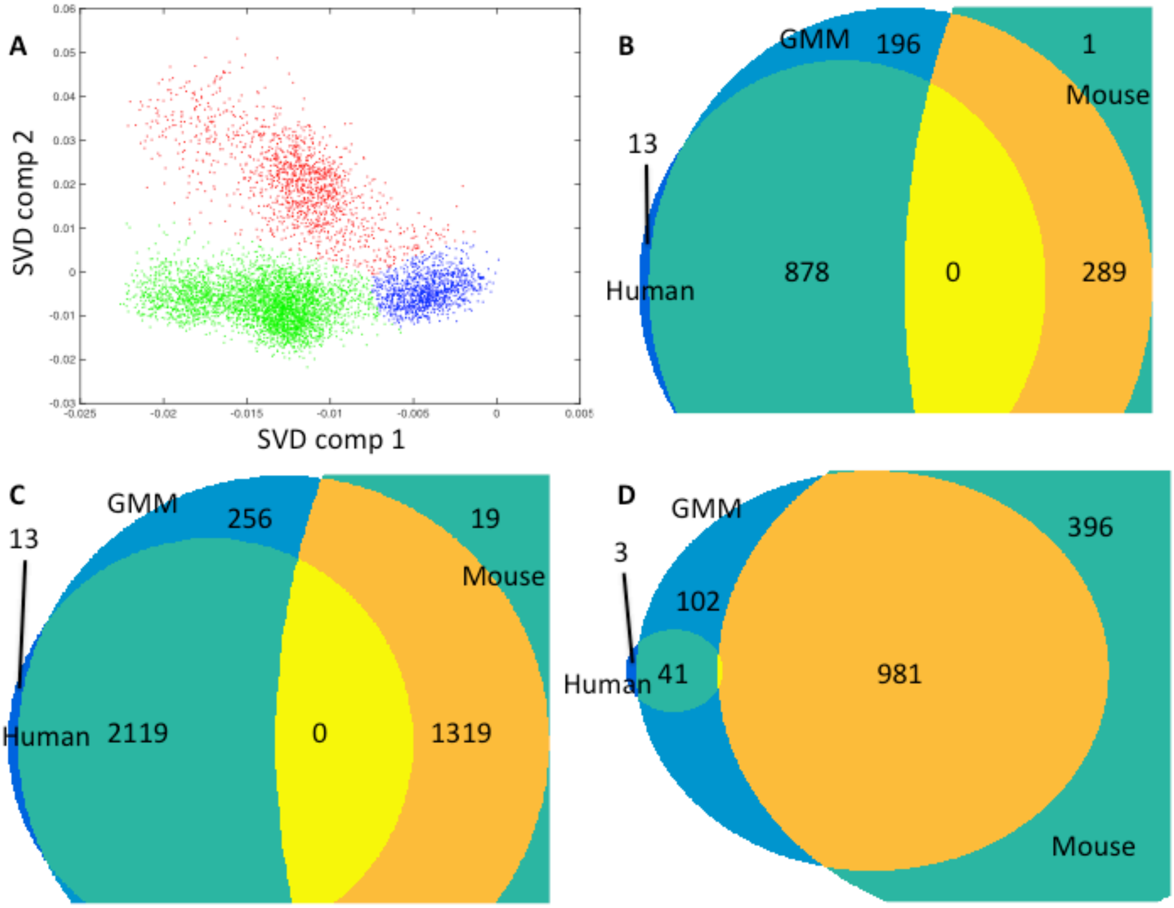


Figure 6. Comparing GMM clustering of human and mouse cells versus reported cell types. A) SVD components colored by GMM predicted clusters red (interneurons), green (pyramidal) and blue (mural/vascular). B-D are Venn diagrams comparing reported human and mouse cell types with GMM predicted cell types. The following superscripts represent if the point was included + or excluded – from species and GMM cluster. The colors from left to right consist of Human$^+$-GMM$^-$ (blue), Human$^+$-GMM$^+$ (green), Human$^-$-Mouse$^-$-GMM$^+$ (blue), Null set (yellow), Mouse$^+$-GMM$^+$ (orange), Mouse$^+$-GMM$^-$ (green). B) GMM Interneurons cluster (red in panel A) with mouse and human interneuron labeled cells. C) GMM Pyramidal (green in panel A) with mouse and human pyramidal labeled cells ("CA1, S1" and "Ex" respectively). D) GMM Glial/Vascular (blue in panel A) with mouse glial/vascular labeled cells and human "NoN" labeled cells.

4. Discussion

4.1. *Insights*

In this study we found that through feature selection it is possible to find informative gene sets that can be used across species. This feature selection of "concordant gene sets" is an important application of single cell data that has multiple downstream applications in relation to cross species modeling, especially in translation of preclinical studies. It is important to note that the data used to find the concordant genes cannot be paired by sample which makes correlation matrices impossible to generate. Without correlation matrices to discover concordant genes, the gene sets must be derived from ulterior methods such as minimizing redundant gene sets through machine learning [17] or grouped statistical tests like ANOVA.

4.1.1. *Scalability*

The feature selection method is based on ANOVA which is calculated across multiple groups. Unlike t-tests, this facet of ANOVA makes the feature selection method scalable in relation to number of species and cell types being studied. Because of this, finding concordant gene sets between many organisms and cell types simultaneously is possible and should be pursued.

4.1.2. *Functional relevance*

The annotated concordant gene set had a clear relationship to the brain through gene ontology which is an important control due to the tissue origin [18]. It is important to note that gene sets with no functional overlap to the phenotype being selected for could potentially be selecting for unknown associated phenotypes. The functional ontology analyses of this concordant gene set shows that there is selection of genes with direct relation to neuronal phenotypes. Because of the enrichment of phenotypically similar ontology terms, a case can be made that seemingly phenotypically dissimilar ontology terms are more likely to have an unknown but direct relationship to our concordant gene set.

4.1.3. *Evolutionary potential*

Concordant gene sets also contain unique evolutionary information. Gene homologs which express differently between two species (Discordant genes) potentially do not share exactly the same functionality. Discordant genes may have the same down-stream effects but the biological mechanism may have changed [6] such that the same quantity of mRNA is not produced across species. Concordant genes are informative because they could represent pathways that are relatively conserved between through the evolution of species.

4.1.4. *Medical and research potential*

In the medical realm concordant gene sets could be of use in translational research. Much of research is conducted in model organisms and using concordant gene sets gives the user an ability to distinguish between transcriptional changes that likely cause similar phenotypes or likely do not between the model and human. Though we do not immediately condone the clinical use of concordant genes at the present these concordant gene sets could help to quickly and efficiently integrate cross-species knowledge to improve translational research.

4.1.5. *Future work*

The scalability of cell type and species number should be tested upon the arrival of comparable data in other species. Aside from the direct feature selection of concordant genes multiple comparisons could be carried out to create hierarchical concordant gene sets for higher granularity. Another option to improve granularity would be to test models that include interaction variables between species, brain location, and cell type. With the generation of concordant gene sets cross-species deconvolution could become more accurate than with more heuristic approaches. Also concordant gene sets can be used in classification of cell types across species. With further refinement of the procedure human cell types could be classified using mouse expression profiles which would require refinement of feature selection and of classification algorithms and validation of such methods on another dataset.

4.1.6. *Importance of single cell granularity*

Single cell technologies in the form of fluorescence-activated cell sorting (FACS) and flow cytometry have been effectively used to model cell heterogeneity [19] before the advent of single cell transcriptomics. Through FACS sorting [20] and flow cytometry [21] deriving the transcriptome of a single cell is much higher throughput than original methodologies that required manual isolation of single cells [22]. Without the single cell granularity of these techniques, it would be impossible to study concordant genes effectively at the cellular level and acquire the sample sizes large enough to properly study concordant gene sets, especially when many species and phenotypes are involved. Only through these recent advances in scRNA-Seq is it possible to properly glean enough information about cell types to model across species.

4.2. *Limitations*

There are some limitations to this study, which included the use of zscores as the measurement of expression. This measurement makes the assumption that the data has a normal distribution. Because of the nature of scRNA-Seq data the distribution is negative binomial. It was important to use zscores because other normalization techniques would not be effective. Quantile normalization introduced artificats in the data that made it unrepresentative. Conversion of UMI counts to TPM alos posed a problem because TPM is based on aligned reads opposed to tag counts from UMIs.
Aside from normalization, the diversity of cell types in each dataset also potentially introduced bias. The human dataset consisted of fewer major cell types than the mouse dataset. The mouse dataset contained more glial cell types while the human dataset had higher granularity within interneurons and pyramidal cells.

5. Conclusion

We were able to find a concordant gene set between mouse and human brain cells that had direct functional ontology relationships to the brain. The concordant gene set allowed us to reduce the distance between cell types of different species allowing separation of cell type regardless of each cell's species. Through the study of these aggregate cell types the biologically unresolved human cell type "NoN" (No Nomenclature) was able to be categorized as Glial/Vascular. Furthermore we show that our methodology is scalable to multiple species and cell types to find concordant gene sets between multiple species and these concordant genes sets are important stepping stones toward evolutionary and translational research goals.

6. Acknowledgements

This work is partially supported by the Human Frontier Science Program (RPG0053/2014) and the NLM T15 training grant. The Ohio Supercomputer Center provided computing support.

References

[1] S. Lin, Y. Lin, J. R. Nery, M. A. Urich, A. Breschi, C. A. Davis, A. Dobin, C. Zaleski, M. A. Beer, W. C. Chapman, T. R. Gingeras, J. R. Ecker, and M. P. Snyder, *Proc. Natl. Acad. Sci.*, **111**, 17224 (2014).

[2] P. P. C. Tan, L. French, and P. Pavlidis, *Front. Neurosci.*, **7**, 1 (2013).

[3] K. Taeho, G. S. Vidal, M. Djurisic, C. M. William, M. E. Birnbaum, C. K. Garcia, B. T.

Hyman, and C. J. Shatz, **341**, 1399 (2013).

[4] S. Matsuda, M. Katane, K. Maeda, Y. Kaneko, Y. Saitoh, T. Miyamoto, M. Sekine, and H. Homma, *Amino Acids* **47**, 975 (2015)

[5] J. W. Rowley, A. J. Oler, N. D. Tolley, B. N. Hunter, E. N. Low, D. a Nix, C. C. Yost, G. a Zimmerman, and A. S. Weyrich, *Blood* **118**, 101 (2011).

[6] J. a Miller, S. Horvath, and D. H. Geschwind, *Proc. Natl. Acad. Sci.* **107**, 12698 (2010).

[7] S. a. Fietz, R. Lachmann, H. Brandl, M. Kircher, N. Samusik, R. Schroder, N. Lakshmanaperumal, I. Henry, J. Vogt, a. Riehn, W. Distler, R. Nitsch, W. Enard, S. Paabo, and W. B. Huttner, *Proc. Natl. Acad. Sci.* **109**, 11836

[8] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll, *Cell* **161**,1202 (2015).

[9] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suvà, A. Regev, and B. E. Bernstein, *Science* **344**, 1396 (2014).

[10] A. Zeisel, A. B. M. Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, and S. Marques, *Science* **347**, 1138 (2015).

[11] B. B. Lake, R. Ai, G. E. Kaeser, N. S. Salathia, Y. C. Yung, R. Liu, A. Wildberg, D. Gao, H.-L. Fung, S. Chen, R. Vijayaraghavan, J. Wong, A. Chen, X. Sheng, F. Kaper, R. Shen, M. Ronaghi, J.-B. Fan, W. Wang, J. Chun, and K. Zhang, *Science* **352**, 1586 (2016).

[12] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson, *Nature methods* **11**, 163 (2014).

[13] D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. a Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, and R. Sandberg, *Nat. Biotechnol.* **30**, 777 (2012).

[14] D. W. Huang, R. a Lempicki, and B. T. Sherman, *Nat. Protoc.* **4**, 44 (2009).

[15] D. W. Huang, B. T. Sherman, and R. A. Lempicki, *Nucleic Acids Res.* **37**, 1 (2009).

[16] P. D. McNicholas and T. B. Murphy, *Bioinformatics* **26**, 2705 (2010).

[17] C. Ding and H. Peng, *Journal of Bioinformatics and Computational Biology* **3**, 185 (2003).

[18] N. A. Twine, K. Janitz, M. R. Wilkins, and M. Janitz, *PLoS One* **6,** e16266.

[19] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs, R. V Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis, *Nat. Biotechnol.* **29**, 886–891 (2011).

[20] N. K. Wilson, D. G. Kent, F. Buettner, M. Shehata, I. C. Macaulay, F. J. Calero-Nieto, M. Sanchez Castillo, C. A. Oedekoven, E. Diamanti, R. Schulte, C. P. Ponting, T. Voet, C. Caldas, J. Stingl, A. R. Green, F. J. Theis, and B. Gottgens, *Cell Stem Cell* **16**, 712 (2015).

[21] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit, *Science* **343**, 776 (2014).

[22] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani, *Nat. Methods* **6**, 377 (2009).

A SPATIOTEMPORAL MODEL TO SIMULATE CHEMOTHERAPY REGIMENS FOR HETEROGENEOUS BLADDER CANCER METASTASES TO THE LUNG

KIMBERLY R. KANIGEL WINNER[1,2], JAMES C. COSTELLO[1,2,3]

[1]Computational Bioscience Program,
[2]Department of Pharmacology,
[3]Univeristy of Colorado Cancer Center
University of Colorado Anschutz Medical Campus
12801 E. 17th Ave. MailStop 8303,
Aurora, CO 80045, USA
Email: kimberly.kanigelwinner@ucdenver.edu, james.costello@ucdenver.edu

Tumors are composed of heterogeneous populations of cells. Somatic genetic aberrations are one form of heterogeneity that allows clonal cells to adapt to chemotherapeutic stress, thus providing a path for resistance to arise. *In silico* modeling of tumors provides a platform for rapid, quantitative experiments to inexpensively study how compositional heterogeneity contributes to drug resistance. Accordingly, we have built a spatiotemporal model of a lung metastasis originating from a primary bladder tumor, incorporating *in vivo* drug concentrations of first-line chemotherapy, resistance data from bladder cancer cell lines, vascular density of lung metastases, and gains in resistance in cells that survive chemotherapy. In metastatic bladder cancer, a first-line drug regimen includes six cycles of gemcitabine plus cisplatin (GC) delivered simultaneously on day 1, and gemcitabine on day 8 in each 21-day cycle. The interaction between gemcitabine and cisplatin has been shown to be synergistic *in vitro*, and results in better outcomes in patients. Our model shows that during simulated treatment with this regimen, GC synergy does begin to kill cells that are more resistant to cisplatin, but repopulation by resistant cells occurs. Post-regimen populations are mixtures of the original, seeded resistant clones, and/or new clones that have gained resistance to cisplatin, gemcitabine, or both drugs. The emergence of a tumor with increased resistance is qualitatively consistent with the five-year survival of 6.8% for patients with metastatic transitional cell carcinoma of the urinary bladder treated with a GC regimen. The model can be further used to explore the parameter space for clinically relevant variables, including the timing of drug delivery to optimize cell death, and patient-specific data such as vascular density, rates of resistance gain, disease progression, and molecular profiles, and can be expanded for data on toxicity. The model is specific to bladder cancer, which has not previously been modeled in this context, but can be adapted to represent other cancers.

1. Introduction

1.1. *Tumor heterogeneity and drug resistance*

Intratumoral heterogeneity is increasingly recognized as a major contributor to cancer progression, metastatic potential, and drug resistance.[1,2] Metastatic tumors that arise from the primary site are generally established from single clones, but may also display initial genetic heterogeneity.[3,4,5] Sub-clonal cell phenotypes with varying metastatic potential and drug resistance have also been shown to develop in 90% of lung metastases within weeks of establishment in mice.[5] This heterogeneity can lead to differential drug response within or among metastases, with newly arising clones developing additional resistance.[5] After the death of sensitive cells and continuing replication of resistant survivors, the spatial dynamics of drug diffusion and accumulation during later drug delivery cycles may change.

A bottleneck in clinical research studies of drug resistance is the lack of tumor sample measurements over the course of treatment from the same patient that can be used to explore the relationship between tumor polyclonality and drug resistance.[6] By building explicit computational models with evolving dynamics, we can manipulate, visualize, and quantitatively analyze patterns of resistance that emerge in a growing tumor. Here, we have created a spatiotemporal model of bladder cancer metastasis to the lung that includes cycles of drug delivery, tumor vascularity, and clumped clonal populations with different drug sensitivities. We model how a heterogeneous tumor responds to the standard first-line regimen of gemcitabine plus cisplatin (GC). Results show that a 100 cell simulated tumor, composed of four clonal populations ranging from highly sensitive to highly resistant cells will not be completely killed by this regimen, and will grow while gaining cross-resistance to both gemcitabine and cisplatin. In this work we aim to model drug response in bladder cancer metastases and establish a baseline set of results that can be extended to model additional visceral sites, determine how varying tumor composition affects drug response, and determine how altering drug scheduling will affect drug response.

1.2. *Prior spatiotemporal models of drug delivery, tumor heterogeneity, and resistance*

Our model is a cellular Potts model, which represents cells and chemical fields on a spatial lattice, interacting and evolving over time. Spatiotemporal models have been used to represent disease development and drug delivery in a variety of cancers, and have generated observations that are not easy to measure in real biological systems.[7–9] They have incorporated parameters such as response to oxygen, information sources provided to the cell such as nutrients and toxicity, and distance from the information source.[8] Spatiotemporal cancer therapy models have used cell cycle, chemotherapy, and radiation data to predict changes in tumor size during treatment. Some have included more specialized events and data, such as bystander effects (in which tumor cells assist in killing damaged cells) resulting from radiotherapy[10] and patient data from CT scans in models of brain cancer.[11,12] These models have successfully produced qualitatively and semi-quantitatively comparable results to *in vitro* studies*,* mouse models, and patient outcomes, showing the promise of spatiotemporal modeling for *in silico* oncology. To our knowledge, there are no existing spatiotemporal models of drug delivery to lung metastases arising from bladder cancer.

Tumor heterogeneity and resistance have been explored with spatiotemporal methods, including two agent-based models (one incorporating game theory for trade-offs between proliferation and migration), field theory, a cellular automaton/cellular Potts model, and a pure cellular automaton. Interestingly, in three of these models,[13,14,15] slowing of the cell cycle was an important predictor of resistance, whether due to cells being driven into quiescence by drugs, by a shortage of oxygen and nutrients, or from initial heterogeneity between clonal populations in their endogenous cell cycles; cells with inherently slow growth were reservoirs for survival during therapies that depend on cell division.[14,15] This last model is the most similar to ours, and is part of a comparison of spatiotemporal implementations, showing that there are trade-offs between performance and resolution for different model types, but that similar types parameterized to the same system will produce cross-validating results. The simulated tumor in ref. 15 was composed of cell populations having heterogeneous cell cycles that changed in response to oxygen, chemotherapy, and radiation (in a 300×300 cellular Potts model). Our model similarly includes cell cycles and chemotherapy, but is different in that it creates a site-specific tumor environment incorporating vascular density specific to metastases to the lung, with *in vivo* concentration curves for drug delivery, and initial and gained resistance modeled using bladder cancer cell lines. In both models, the spatial arrangement of vessels creates a drug concentration unique to each cell in a simulation, allowing spatially driven phenomena to emerge.

1.3. *Bladder cancer drug regimen and cell response*

Annually, it is estimated that there will be nearly 77,000 new cases of bladder cancer with over 16,000 succumbing to the disease.[16] Overall survival has not improved since 1989.[16] The most aggressive form, muscle-invasive bladder cancer, occurs in 30% of patients.[17] Treatment is radical cystectomy, requiring removal of the bladder and sometimes surrounding tissues, followed by chemotherapy. The 5-year survival rate varies from 25-50%. Failure is likely due to occult metastases present before treatment, with the most common visceral metastatic sites in the liver and lungs.[17,18] Patients with inoperable locally advanced or metastatic cancer who undergo GC or methotrexate/vinblastine/doxorubicin/cisplatin (MVAC) regimens have a 5-year overall survival of 13%, but a progression-free survival of 9.8%.[19] Those with lung, liver, or bone[18] metastases have a 5-year overall survival rate of 6.8%.[19] Here, we model this last group of patients, with aggressive metastatic disease localized to the lung.

The standard regimen defined by the National Comprehensive Cancer Network (NCCN) for metastatic bladder cancer includes six 21-day cycles, with GC delivered simultaneously on day 1 (or cisplatin instead on day 2) and gemcitabine alone on day 8.[20] For patients with muscle-invasive or metastatic cancer, who cannot receive cisplatin, monotherapy regimens without cisplatin produce no long-term disease-free survival, with a median survival of six to nine months.[17] This was reflected in initial runs of the model, with rapid acquisition of resistance during cisplatin or gemcitabine monotherapy regimens. Reported efficacy of such regimens is derived from clinical trials. Computational models of drug delivery can additionally be used to generate hypotheses at a small scale where we can explore mechanisms of drug action and drug resistance, as well as adjust the regimen in a consequence-free environment where results for 18 weeks of time course data can be obtained in just hours.

Cisplatin and gemcitabine are genotoxic agents, damaging DNA and causing a cell to undergo apoptosis during cell division. Cisplatin incorporates into DNA as platinum-DNA adducts,[21] whereas gemcitabine is a nucleoside analog that interrupts DNA synthesis and triggers apoptosis.[22] The 50% inhibitory concentration (IC50) is a concentration of drug that inhibits a cellular process by 50%. IC50 for cytotoxicity and drug accumulation in cells are linearly correlated for both cisplatin and gemcitabine, especially at clinically relevant concentrations, which tend to be at the lower end of cytotoxicities measured *in vitro*.[23–25] There is also a linear relationship between tissue platinum concentration and tumor size reduction.[26] These relationships were used to parameterize cellular accumulation of the two drugs.

Synergy between gemcitabine and cisplatin occurs during pre-treatment with gemcitabine or co-treatment with GC in ovarian and neuroblastoma cells.[27,28] In these studies, one in four and one in five cell lines did not respond synergistically. Patients with non-small-cell lung cancer also responded better to a day 1 combination of gemcitabine and cisplatin than to day 1 cisplatin alone (30.4% response compared to 11%, $p<1e-4$), with improved median time to progression and improved overall survival.[29] Synergy in cisplatin during the GC regimen is an important dynamic that we include in the model.

2. Methods

2.1. *Summary of model design*

Our model represents a partially drug-resistant lung metastasis that arose from a primary bladder tumor, containing four clonal cell patches with different sensitivities to gemcitabine and cisplatin. The drugs are delivered through vasculature in the tumor at levels found in patient plasma based on the regimen dosages. Drugs diffuse from vessels with effective diffusion coefficients measured in tumor tissue, and accumulation is a cell-type-specific proportion of drug concentration at the cell site. Synergy between the drugs causes increased intracellular cisplatin accumulation. If cells attempting to replicate have accumulated enough drug to reach their IC50 or greater, they will either die with 50% probability or increase their resistance. Finally, when a cell divides, its accumulated drug is halved between the two child cells. Drug delivery frequency and dosage are from the basic GC drug regimen for metastatic bladder cancer (see Fig. 1 for model).

Tumor and vessel cell types are represented, along with cell division, cell death, and clearance of dead cells as a proxy for the immune system. Vascular density for lung metastases is equal to the ratio of microvessel density between primary and lung metastases in non-clear cell renal cell carcinoma.[30,31] Further biometric parameters, derivations, fits for drug concentrations in patients, and their sources can be found in Table 1. Model permutations include runs with and without synergy, variations on the drug regimen, and variations in rates of resistance gain in the cells.

The modeling platform is Compucell3D (CC3D),[32] an integrated programming and visualization environment for cellular Potts models. Cellular Potts models couple mobile, single-cell agents to a cellular automaton process at the cells' surfaces. Cell agents live on their own 2-D or 3-D lattice, and chemical fields can be layered on in other lattices. Partial differential equations for drug diffusion are solved using the Forward Euler method. For more explicit descriptions of the cellular Potts model for modeling drug delivery in tumors, please see Kanigel Winner, et al.,[33] and Extended Methods are available at https://synapse.org/MetHet. In short, pre-defined biological rules comprise an energy function that drives the behavior of the cellular automaton process at the cell surface during each Monte Carlo time step (MCS). Meeting the rules (by convention) lowers this energy or keeps it the same, allowing biologically reasonable cellular events contributing to growth, division, and death (though stochasticity can be added). Cell death, cell type switches due to drug accumulation, and drug delivery calculated from continuous functions (fits to patient plasma drug concentrations) are expansions of the basic CC3D model coded in a Python wrapper. These processes are non-stochastic. More modeling methods, details of parameter acquisition, and source code that is plug-and-play in CC3D can be found at https://synapse.org/MetHet.

2.2. *Specifics of biological parameters and model dynamics*

- IC50 data for gemcitabine and cisplatin sensitivity in 18 bladder cancer cell lines were acquired from the Genomics of Drug Sensitivity in Cancer (GDSC) database.[34]
- Cell growth and division occurred in all cancer cells. Replication rate was approximated from the averages of 14 cancer cell lines varying in metastatic capacity (31 to 33 hrs.).[22,36,37]
- Cisplatin and gemcitabine in normal cells (lung and phagocytic cells) were given accumulation rates for the bladder cancer cell line (SW780) closest to the middle of the range for both drugs.

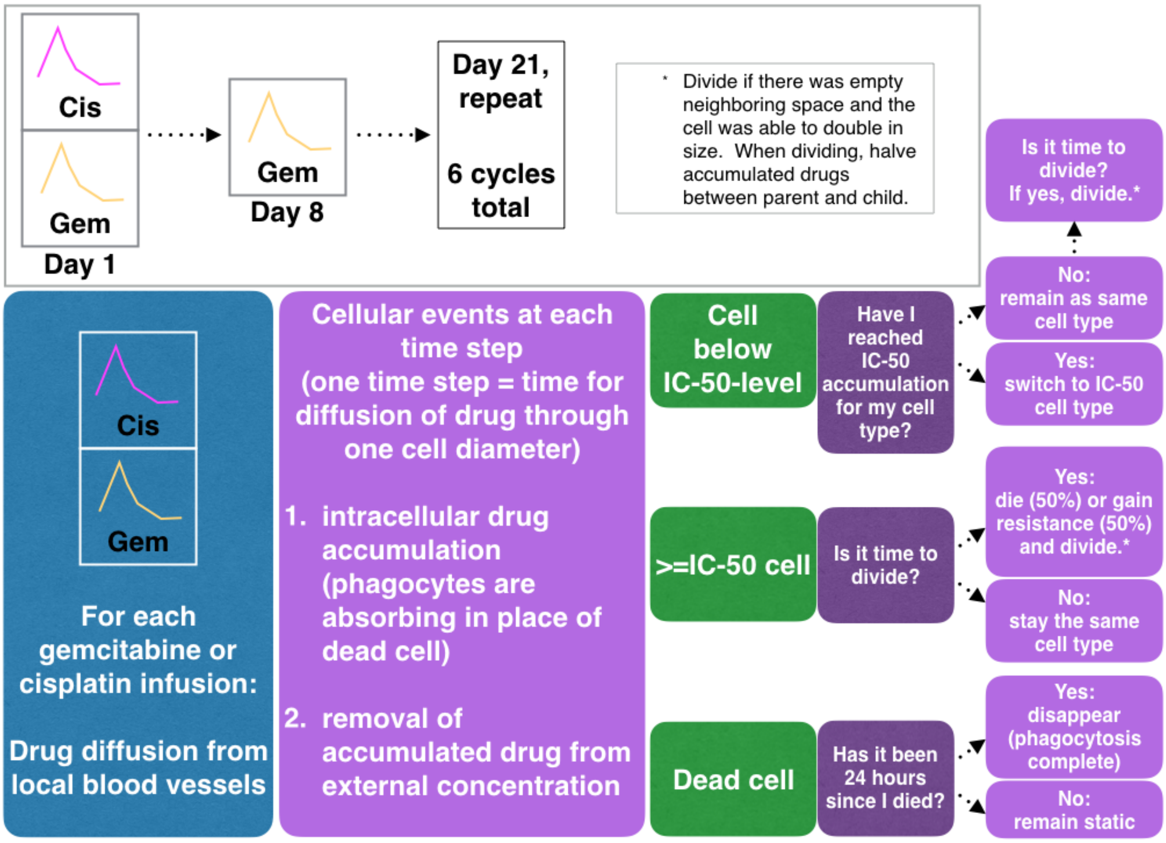

Cis
Gem
Day 1
Gem
Day 8
Day 21, repeat
6 cycles total
* Divide if there was empty neighboring space and the cell was able to double in size. When dividing, halve accumulated drugs between parent and child.
Cis
Gem
For each gemcitabine or cisplatin infusion:
Drug diffusion from local blood vessels
Cellular events at each time step (one time step = time for diffusion of drug through one cell diameter)
1. intracellular drug accumulation (phagocytes are absorbing in place of dead cell)
2. removal of accumulated drug from external concentration
Cell below IC-50-level
Have I reached IC-50 accumulation for my cell type?
Is it time to divide? If yes, divide.*
No: remain as same cell type
Yes: switch to IC-50 cell type
>=IC-50 cell
Is it time to divide?
Yes: die (50%) or gain resistance (50%) and divide.*
No: stay the same cell type
Dead cell
Has it been 24 hours since I died?
Yes: disappear (phagocytosis complete)
No: remain static

Figure 1. Flow chart of events in the model at each time step, reflecting body- and cellular-scale processes

- Acquired resistance was modeled as an increase in the IC50 of any cell that survived an IC50 accumulation of gemcitabine or cisplatin at division time, increasing the chances of being below IC50 and another gain in resistance at the next division time. The quantity to be added to the IC50 for each gain in resistance (Table 1) was derived from bladder cancer cell lines, passaged to increase resistance, as the increase per division required to acquire maximum resistance over one year ("quick") or two years ("slow") for cells with a 30-hour cell cycle.[35]
- Cell accumulation rate and peak of gemcitabine is linearly correlated with concentration *in vitro* and *in vivo*.[38] In bladder cancer cells, cytotoxicity is linearly correlated with gemcitabine concentration,[25] and accumulation is correlated with IC50.[36] Cisplatin DNA lesion counts are linearly correlated with concentration.[27] We therefore fit cellular accumulation rates for both gemcitabine and cisplatin linearly to the IC50 of each cell type, with some modifications.[36,39]
- Cells at IC50 for both gemcitabine and cisplatin at division underwent two chances at death.
- Gemcitabine and cisplatin were modeled with the same effective diffusion coefficient as sodium fluorescein.[40] For details on this choice, please see ref. 33. Both molecules diffused at the same rate in all cell types except for blood vessel, which either took away molecules, ostensibly into flowing blood, or delivered them from the vessel surface.

.

Table 1. Model parameters and fits to data

Parameter	Value	Units	Source
Cell diameter (BC* T24 line, aggressive/invasive)	30	μm	[42]
Eff. diffusion coefficient sodium fluorescein	6.40E-06	cm^2/s	[40]
Division time (mean, S.D.)	30, 1	h	[22,36,37]
Time from death to complete phagocytosis	24	h	[43]
Fraction cross-sectional microvessel area in metastasis from urinary system cancer to lung	0.146		[30]
Pixel dimension	1	cell	
Cisplatin resistance gain per survived division	0.125 – 0.25	+ IC50	[35]
Gemcitabine resistance gain per survived division	0.05 – 0.1	+ IC50	[35]
IC50 cis. accumulation for initial cell lines. Seed gem. & cis. sensitive, Seed res. gem./sens. cis., Seed res. cis./sens. gem., Seed gem. & cis. resistant	0.8106177157, 3.774888444, 6.586828431, 5.923917064	μM per cell	calculated using fit from [44]
IC50 gem. accumulation for initial cell lines. Seed gem. & cis. sensitive, Seed res. gem./sens. cis., Seed res. cis./sens. gem., Seed gem. & cis. resistant	0.000017923, 270.913928515, 0.145644144, 46.134163935	μM per cell	calculated using fit from [36]
Accumulation rates of cis. in initial cell lines. Seed gem. & cis. sensitive, Seed res. gem./sens. cis., Seed res. cis./sens. gem., Seed gem. & cis. resistant	7.98701E-05, 6.82909E-05, 7.42347E-06, 5.46716E-05	* cis. (μM) at cell site per MCS	fit from [44]
Accumulation rates of gem. in initial cell lines· Seed gem. & cis. sensitive, Seed res. gem./sens. cis., Seed res. cis./sens. gem., Seed gem. & cis. resistant	4.41575E-04, 2.68443E-04, 4.41518E-04, 4.22858E-04	* gem. (μM) at cell site per MCS	fit from [36]
Fit for cisplatin plasma concentrations during 3h infusion (top) and decay (bottom)	$= 0.11*hrs^3 - 0.83*hrs^2 + 2.2*hrs - 2.6E-16$ $= 57.4124 * e^{(-1.0927 * hrs)}$	μM	[46]
Fit for gemcitabine plasma concentrations during 30m infusion (top) and decay (bottom)	$= 6.8*(min/15 - 1) + 7.3$ $= 101.3452 * e^{(- 0.0676 * min)}$	μM	[45]
Synergy multiplier for cisplatin accumulation	2.5		[27,28]
Total Monte Carlo (simulation) Steps (126 days)	11,916,800		
Time in one Monte Carlo Step (MCS)	0.914	s	

* Bladder Cancer

3. Results

3.1. *No standard or alternate regimen prevents regrowth of a drug resistant tumor*

In preliminary simulations containing only GC dual-sensitive cells, cells declined over time and the population was killed late on day 48, five days after the third round of GC. However, for an initial tumor with three additional cell types that had increased resistance to gemcitabine, cisplatin, or both, neither the standard GC regimen (Figs. 2,3), nor an unrealistically high rate of delivery of gemcitabine could kill all cells.

The initial 2-D tumor of 100 cells consistently quadrupled to 400 cells in 14 to 15 days. At simulation end, 0 to 18 days after the last round of drug (depending on the simulation) the domain was completely filled with drug-resistant tumor cells, primarily cisplatin-resistant and GC dual-resistant cells, as well as a sub-population of the most GC dual-resistant seed population. Within six hours of regimen start, the two most sensitive cell types reached the IC50 for gemcitabine accumulation. While some of these sensitive cells died, some went on to propagate as sub-clones with gained resistance.

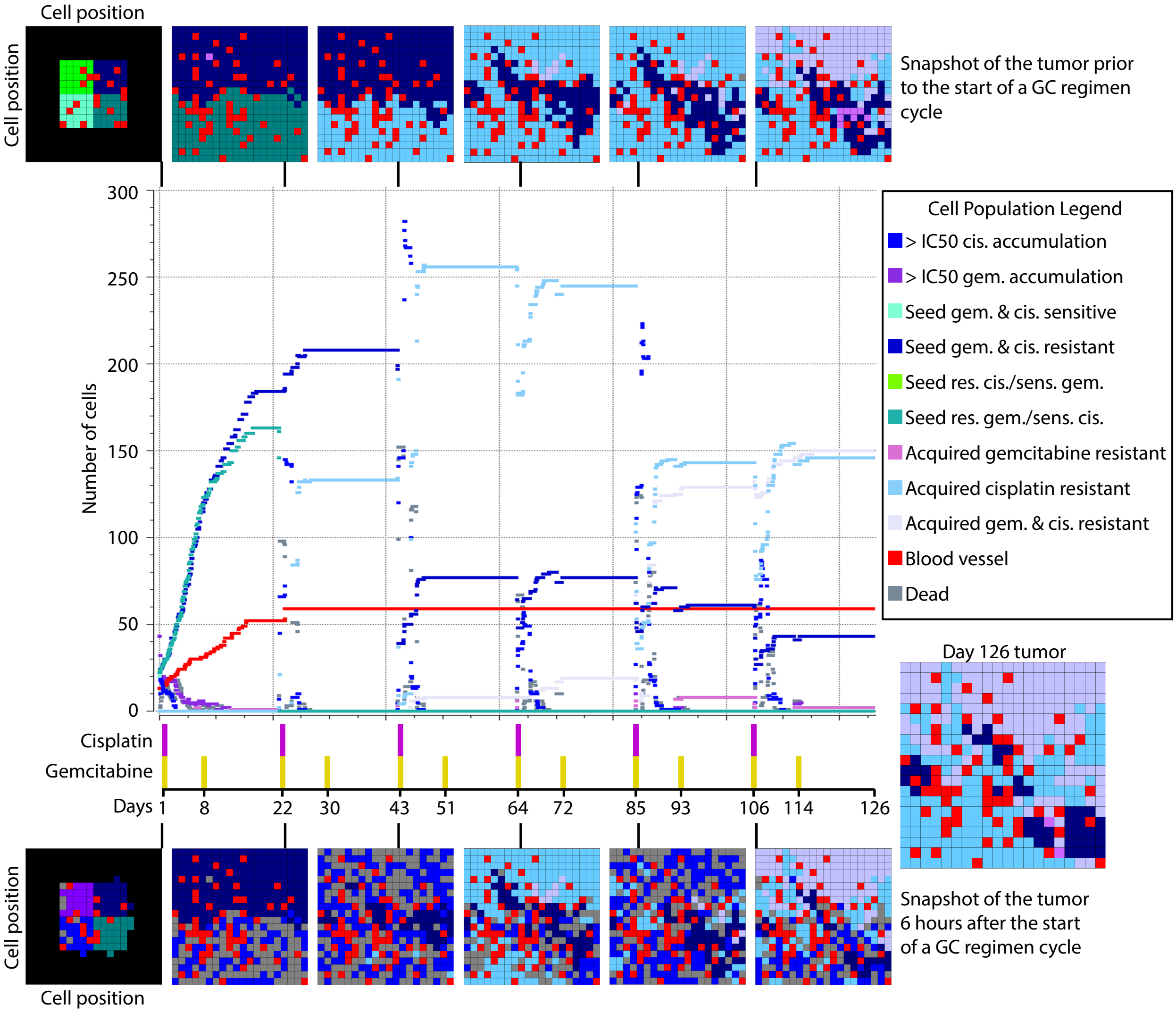


Figure 2. A simulation with random uniform "slow" to "quick" acquired resistance for all cells, and drug synergy in all cells. Pulses of gemcitabine and cisplatin or gemcitabine alone for the first-line chemotherapy regimen are displayed and matched to the simulation. Cells in the simulation were seeded in 100-cell tumors shown in the top-leftmost simulated tumor diagram. The row of simulated tumors on the top represent the state of the tumor before the start of a chemotherapy cycle; the row of simulated tumors on the bottom represent the state of the tumor 7 hours after a GC cycle. Resistant seed cells, cells with dual resistance, and cells with cisplatin resistance composed the final population as shown as the final simulated tumor after 126 days of treatment.

3.2. *Effects of acquired resistance*

3.2.1. *Ability of cells to gain resistance increases likelihood of dual resistance*

When acquired resistance was allowed to arise in the cell populations, cells with acquired resistance comprised the majority of the final tumor (Figs. 2, 3). GC dual-resistant sub-clones arose at day 43, after the third cycle of GC, suggesting that increased dosage or delivery rate prior to this time point may help to keep cross-resistant strains from arising. Interestingly, the fastest rate of acquired resistance for both gemcitabine and cisplatin drove cells with acquired resistance to cisplatin to dominate the population.

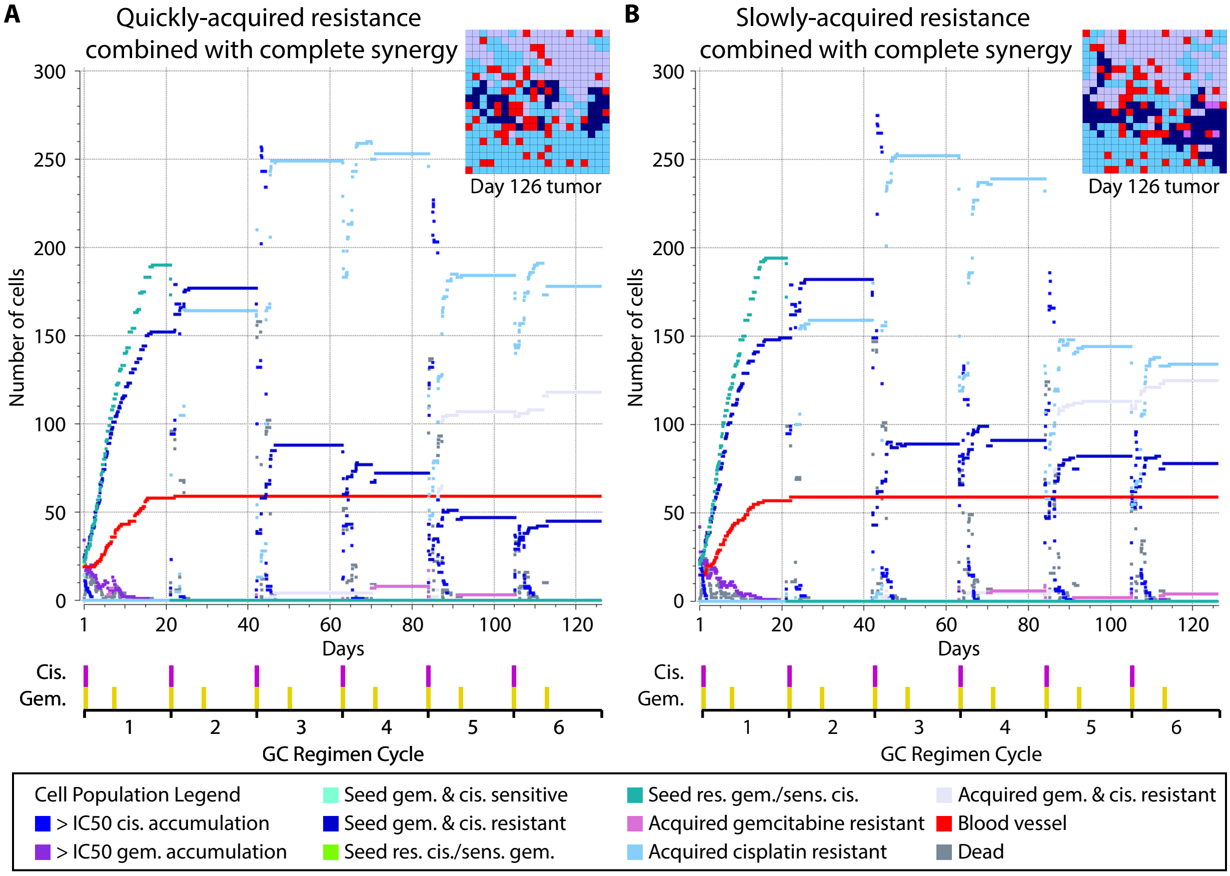


Figure 3. (A) Quickly-acquired resistance and (B) slowly-acquired resistance resulted in tumors composed primarily of cells with newly acquired resistance, with a smaller population of highly GC dual-resistant seed cells. Quickly-acquired resistance drove the tumor toward greater cisplatin resistance.

3.2.2. *Simulated tumors show complete resilience to even intense treatment*

In simulations with an added pulse of gemcitabine at day 18 during each cycle (we mirrored the timing of the 28-day regimen, which has an additional gemcitabine infusion on day 18), we found an earlier rise of the GC dual-resistant phenotype, and more gemcitabine-resistant cells. We also applied single-drug regimens with cisplatin or gemcitabine alone at standard frequencies. Cells with resistance to the treatment drug were the majority of the final population.

To try treatment prior to all cells entering a new cell cycle (30 hrs) while using a potentially tolerable regimen, we shifted the three pulses of gemcitabine to the first three days of each 21-day cycle, at every 24 hours, in addition to the standard cisplatin every 21 days. This caused the end state to be dominated by GC dual-resistant cells. Finally, we pulsed gemcitabine every 24 hours for 126 days, with cisplatin every 21 days. The tumor was not killed, and the simulated tumor area was fully repopulated, primarily with cisplatin-resistant cells after 126 pulses of gemcitabine reduced the gemcitabine-resistant populations.

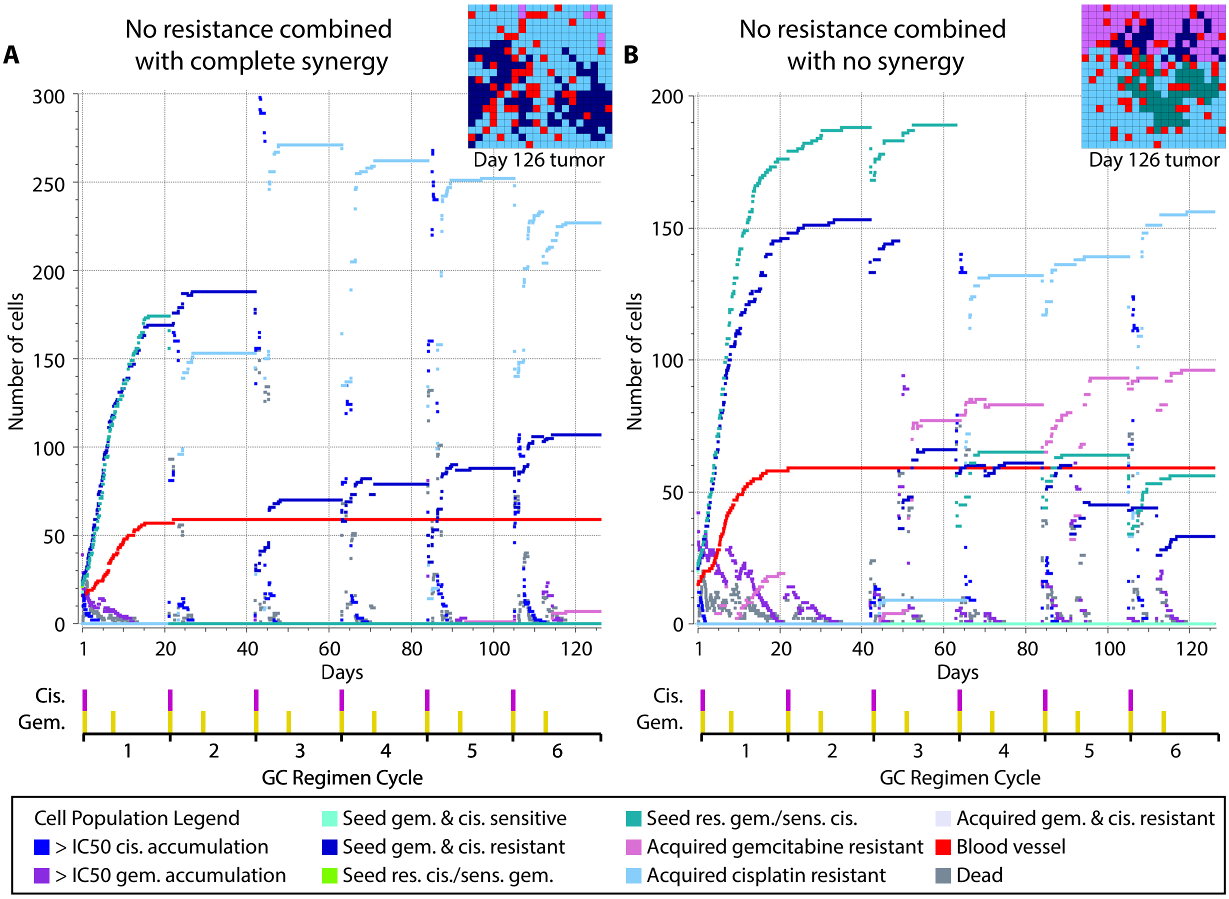


Figure 4. In simulations without acquired resistance, models were considered (A) with drug synergy between gemcitabine and cisplatin and (B) without synergy. The final tumor was composed of cells that survived division after reaching either cisplatin IC50 accumulation, or gemcitabine IC50 accumulation. Cells that survived both gemcitabine and cisplatin IC50 levels did not arise. When there was no synergy in cisplatin (2.5× normal accumulation rate, B), an extra cell type (teal-colored) derived from the seed populations remained.

3.3. *In the absence of acquired resistance, diffusion of drug via cell division allows survival*

In simulations where cells did not acquire resistance, populations primarily composed of cells that randomly survived an IC50-cisplatin division (Fig. 4A) repopulated the simulation space. A subpopulation of initial highly GC dual-resistant seed cells also survived. Because of the division of drug equally between two progeny cells, acquired resistance was not required for tumor repopulation, suggesting that cells reaching IC50 accumulation may survive *in vivo* without newly acquired resistance. In simulations with synergy and resistance (Figs. 2, 3), one clone in the original tumor died during the second round of GC at day 21 (teal-colored; $\mathrm{IC50_{cisplatin}} = 14.0\mu\mathrm{M}$ in range 2.6μM to 225.2μM). When synergy and the ability to gain resistance were absent, this cell type comprised a substantial portion of the final tumor, the most heterogeneous final tumor in our models (Fig. 4B). Hence, for the most resistant seed cells, and in less resistant seed cells in which synergy may not be active, no acquisition of extra resistance was required for repopulation.

4. Discussion

In this work, we were able to capture population-level responses to chemotherapy stress in a model of lung metastasis arising from the bladder. Unless the initial tumor was comprised of highly sensitive cells, the *in vivo* concentration and timing of the standard first-line regimen did not kill the metastasis. Cells were then able to proliferate and fill the simulation space after completion of treatment. A striking result was that in tumors without any ability to acquire resistance, some cells survived the IC50 threshold and were able to repopulate the space. When tumors were allowed to acquire resistance, there was consistent emergence of cells that had coordinately increased resistance to both gemcitabine and cisplatin around 43 days. This occurred after the third cycle of GC, suggesting that early aggressiveness in treatment may be important in avoiding cross-resistant sub-clones. In terms of drug-directed cell selection, when cells were given the ability to acquire resistance, even at slower rates described *in vitro*, final tumors were composed of a majority of cells with acquired resistance. Because metastases starting from single clonal populations in the lung have been shown to develop sub-clones within weeks of establishment,[5] and because cell lines and living tumors are known to gain resistance mutations over time, metastases with large proportions of cells with acquired resistance is a likely scenario in a patient, and the model likely reflects selection *in vivo*.

Qualitative comparisons can be made between prior data and model outcomes. Overall, the acquired resistance model produced rounds of cell death under drug concentrations in patients, showing that the parameters are biologically reasonable. The results are consistent with survival data for patients with inoperable locally advanced or metastatic bladder cancer undergoing a GC or MVAC regimen; those who had lung, liver, or bone metastases had a 5-year overall survival rate of 6.8%.[19] The likelihood of a patient presenting with a completely drug-sensitive metastatic population is low, creating low likelihood of complete cell killing in the tumor. Similarly, in the model, we saw only the most sensitive populations being eradicated by the standard regimen. A patient's metastatic population might have been completely sensitive if metastasis was recently established from a sensitive primary cell and lacked the time to gain genetic heterogeneity. Less likely still, several weeks or more after establishment, the metastases may have either not gained new genetic heterogeneity, or simply not acquired resistance through genetic aberrations. Finally, cells in the model had IC50s derived from cell lines, and some cells died at *in vivo* drug concentrations, suggesting that cell line data reasonably reflects the range of resistances found in patients' tumor cells. While these comparisons to patients and cells are speculative, they are valuable observations for generating hypotheses and represent opportunities for empirical validation as we develop the model further.

When acquisition of resistance was removed, some cells that had initial resistance survived and propagated. This "resistance" occurred because accumulated drug was divided in half between offspring, giving both sensitive and resistant primary sub-clones more time to grow and replicate before reaching IC50, with proliferation outpacing the delivery of drug. This result, in which cells randomly evade death without incorporating new resistance mechanisms, emphasizes the importance of considering growth rate in an aggressive metastatic population.

To estimate the number of doses required to actually kill a metastasis, we simulated delivery of gemcitabine every 24 hours over 126 days, along with synergistic cisplatin every 21 days. Even this unrealistic regimen did not kill the tumor, and drove it to gain cisplatin resistance. Increasing gemcitabine dosage, in combination with increasing the frequency of cisplatin at lower doses should be explored in the future. Additionally, drug regimens that incorporate other drugs besides cisplatin and gemcitabine will be explored in future iterations of the model.

There are caveats to this approach that we must consider. Our model is small (20×20×1 cells) for relatively fast computation so that many scenarios could be explored. Although this size still allowed differential effects to emerge between different drug scenarios, and computational costs scaled proportionally to the number of cells during growth from 100 to 400 cells, larger grids will be part of future work, hopefully approaching the clinical detection limit for lung metastases. The system modeled is specific to bladder cancer; however, lung is a common metastatic site for many other cancers. This and the available data on vascularity at urogenital metastatic sites helped justify the choice of the system modeled. Additionally, the model is simple and general, in part because a GC regimen is used in a variety of cancers, and can be relatively easily adapted to other metastatic or primary sites by replacing parameters in the code. The primary bottleneck to adaptation to other cancers will be the availability of empirical data to derive model parameters.

The model may be allowing consistent tumor survival despite an aggressive drug regimen due to a cell cycle time of 30h +/-1h (S.D.); slower- (or even faster-) cycling cells may create different dynamics. Drug is not delivered from vessels outside of the tumor, inherent cell death rates are not included, and the immune system is not directly considered. Most importantly, although the model can be manipulated unrealistically, useful hypothetical regimens must include practical considerations for regimens given to patients. If a new regimen kills more cells, perhaps the immune system will have a greater chance to reduce a smaller residual population. A simple increase in cell kill under an organizational and dosing scheme reasonable for patients is therefore a goal for this modeling process and the subject of future studies.

Finally, our model results recapitulate prior work by Powathil *et al.*[15] regarding the importance of accounting for cell cycle in drug delivery. Also our results concur with aspects of Waclaw *et al.*,[41] showing that after cell kill opens up space in the tumor, it takes only one or two cells to repopulate the vacant area with a new more resistant sub-clone. Such cell behavior is extremely difficult to track through time in a patient, and even in experimental models such as mice. Therefore, the importance of spatiotemporal models incorporating realistic parameters, with behavior that can be tracked over time to clinically relevant outcomes, cannot be underestimated.

Acknowledgments

We would like to thank members of the Costello lab, in particular Andrew Goodspeed, and Anis Karimpour-Fard for constructive comments. This work is supported by the Boettcher Foundation (J.C.C.) and NIH grant 2T15LM009451 (K.R.K.W.).

References

1. Saunders, N. A. *et al. EMBO Mol. Med.* **4,** 675–84 (2012).
2. Tabassum, D. P. & Polyak, K. *Nat. Publ. Gr.* **15,** 1–11 (2015).
3. Yamamoto, N. *et al. Cancer Res.* **63,** 7785–90 (2003).
4. Talmadge, J. E. & Zbar, B. *J Natl Cancer Inst* **78,** 315–320 (1987).
5. Poste, G. *et al. Proc. Natl. Acad. Sci. U. S. A.* **79,** 6574–8 (1982).
6. Burrell, R. A. *et al. Mol. Oncol.* **8,** 1095–111 (2014).
7. Andasari, V. *et al. J. Math. Biol.* 1-31–31 (2010).
8. Mansury, Y. *et al. J. Theor. Biol.* **219,** 343–370 (2002).
9. Martins, M. L. *et al. Phys. Life Rev.* **4,** 128–156 (2007).
10. Powathil, G. G. *et al. J. Theor. Biol.* **401,** 1–14 (2016).
11. Tracqui, P. *et al. Cell Prolif.* **28,** 17–31 (1995).
12. Stamatakos, G. & Antipas, V. *IEEE Trans.* (2006).
13. Mansury, Y. *et al. J. Theor. Biol.* **238,** 146–156 (2006).
14. De La Cruz, R. *et al*. arXiv:1607.01449v1 (2016).
15. Powathil, G. G. *et al. Semin. Cancer Biol.* **30,** 13–20 (2015).
16. Siegel, R. L. *et al. CA. Cancer J. Clin.* **66,** 7–30 (2016).
17. Piergentili, R. *et al. Curr. Med. Chem.* **21,** 2219 (2014).
18. Sternberg, C. N. *Ann. Oncol.* **17,** 23–30 (2006).
19. von der Maase, H. *et al. J. Clin. Oncol.* **23,** 4602–4608 (2005).
20. National Comprehensive Cancer Network. Bladder Cancer v.1.2016, accessed 5/3/2016
21. Johnstone, T. C. *et al. Philos. Trans. A. Math. Phys. Eng. Sci.* **373,** (2015).
22. van Moorsel, C. J. *et al. Br. J. Cancer* **80,** 981–90 (1999).
23. Mistry, P. *et al. Cancer Res.* **52,** 6188–6193 (1992).
24. Henderson, P. T. *et al. Int. J. cancer* **129,** 1425–34 (2011).
25. Torres, M. P. *et al. PLoS One* **8,** e80580 (2013).
26. Kim, E. S. *et al. J. Clin. Oncol.* **30,** 3345–52 (2012).
27. Moufarij, M. *et al. Mol. Pharmacol.* **63,** 862–9 (2003).
28. Besançon, O. G. *et al. Cancer Lett.* **319,** 23–30 (2012).
29. Sandler, A. B. *et al. J. Clin. Oncol.* **18,** 122–130 (2000).
30. Fukata, S. *et al. Cancer* **103,** 931–942 (2005).
31. Papadopoulos I. *et al. J. Clin. Pathol.* **57,** 250 (2004).
32. Swat, M. H. *et al. Methods Cell Biol.* **110,** 325–66 (2012).
33. Kanigel Winner, K. *et al. Cancer Res.* 0008-5472.CAN-15-1620- (2015).
34. Yang, W. *et al. Nucleic Acids Res.* **41,** D955-61 (2013).
35. Vallo, S. *et al. Transl. Oncol.* **8,** 210–216 (2015).
36. Damaraju, S. *et al. Biochem. Pharmacol.* **79,** 21–9 (2010).
37. Ning S. *et al. Int. J. Oncol.* (2004).
38. Grunewald, R. *et al*. *Cancer Chemother. and Pharmacol*. **27,** 258 (1990).
39. Köberle, B. & Piee-Staffa, A. *Bladder Cancer - From Basic Science to Robotic Surgery.* **Chapter 13**, 265 (2012).
40. Nugent, L. J. & Jain, R. K. *Cancer Res.* **44,** 238–244 (1984).
41. Waclaw, B. *et al. Nature* **525,** 261–264 (2015).
42. Gilloteaux, J. *et al. Anat. Rec. A. Discov. Mol. Cell. Evol. Biol.* **288,** 58–83 (2006).
43. Wagner, B. J. *et al. J. Cell. Sci.* **124**,1644 (2011).
44. Koberle, B. *et al. Biochem. Pharmacol.* **52,** 1729–1734 (1996).
45. Fan, Y. *et al. Acta Pharmacol. Sin.* **31,** 746–52 (2010).
46. De Jongh, F. E. *et al. J. Clin. Oncol.* **19,** 3733–3739 (2001).

SCALABLE VISUALIZATION FOR HIGH-DIMENSIONAL SINGLE-CELL DATA

JUHO KIM

Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
Urbana, Illinois, 61801, USA
Email: juhokim2@illinois.edu

NATE RUSSELL

Institute of Genomic Biology, University of Illinois at Urbana-Champaign
Urbana, Illinois, 61801, USA
Email: ntrusse2@illinois.edu

JIAN PENG

Department of Computer Science, University of Illinois at Urbana-Champaign
Urbana, Illinois, 61801, USA
Email: jianpeng@illinois.edu

Single-cell analysis can uncover the mysteries in the state of individual cells and enable us to construct new models about the analysis of heterogeneous tissues. State-of-the-art technologies for single-cell analysis have been developed to measure the properties of single-cells and detect hidden information. They are able to provide the measurements of dozens of features simultaneously in each cell. However, due to the high-dimensionality, heterogeneous complexity and sheer enormity of single-cell data, its interpretation is challenging. Thus, new methods to overcome high-dimensionality are necessary. Here, we present a computational tool that allows efficient visualization of high-dimensional single-cell data onto a low-dimensional (2D or 3D) space while preserving the similarity structure between single-cells. We first construct a network that can represent the similarity structure between the high-dimensional representations of single-cells, and then, embed this network into a low-dimensional space through an efficient online optimization method based on the idea of negative sampling. Using this approach, we can preserve the high-dimensional structure of single-cell data in an embedded low-dimensional space that facilitates visual analyses of the data.

1. Introduction

Many traditional biological experiments have been conducted on bulk-cell populations[1] with an assumption that cells in the same group share homogeneous properties. However, some evidence[1-3] shows that heterogeneity can exist even within a small group of cells. The assumption based on homogeneity of each cell group can mislead averages and does not properly explain small but critical changes in individual cells. Each cell can have different biological properties such as cell sizes, gene expression levels, RNA transcripts, and bio marker expressions. These variations can be very important to answer previously unsolved questions in stem cell research, cancer biology, and immunology. Single-cell data analysis has contributed to understand the various and important behaviors of individual cells[1-15].

The recent development of single-cell technologies has also improved the analysis to be more reliable and reasonable. For example, mass cytometry[4,16] can measure up to 60 parameters at the same time for tens of thousands of individual cells. In addition, single-cell RNA sequencing techniques[17,18] also have been widely used, which deal with hundreds of or thousands of parameters per cell.

Even though the advanced single-cell technologies can provide quality data, such data sets are still difficult to analyze. Traditionally, single-cell data are analyzed in a biaxial scatter plot for two variables at once[19]. However, this method requires the order of dimension squared to represent all pairwise relationships between variables, which is computationally expensive. In addition, scatter plots cannot capture multivariate relationships between more than two variables. Thus, new computational methods have been developed for analyzing single-cell data. For instance, SPADE[6] tries to find hierarchies of high-dimensional single-cell data showing cellular heterogeneity by clustering of down-sampled cytometry data, constructing minimum spanning trees, and up-sampling. However, this method considers not each cell itself but cell groups and their behaviors on average. X-shift[12] is recently developed to discover cell subsets and visualize them based on a weighted k-nearest neighbor density estimation.

Another approach to deal with the high-dimensionality of single-cell data is to use dimensionality reduction techniques. Some researchers applied principle component analysis (PCA)[20] to find low-dimensional projections of single-cell data[21,22]. Although PCA is possibly the most popular method of dimensionality reduction, it is a linear projection method. Thus, it cannot capture nonlinear structures in single-cell data. In order to address this issue, advanced methods based on nonlinear dimensionality reduction have been developed. Both viSNE[8] and ACCENSE[10] are based on an algorithm called t-Distributed Stochastic Neighbor Embedding (t-SNE)[23]. viSNE applies t-SNE to mass cytometry data and reveals biologically meaningful relationships from bone marrow and leukemia data. ACCENSE combines the results of t-SNE with kernel-based density estimation and finds subpopulations of given single-cell data sets. However, the runtime complexity of t-SNE is $O(n^2)$, and that of its accelerated version, Barnes-Hut-SNE[24] is $O(n \log n)$ where n is the number of cells. Thus, both methods require excessive computational time for large-scale single-cell data sets with hundreds of thousands or millions of cells.

In this paper, we propose a scalable embedding-based visualization method for large-scale and high-dimensional single-cell data based on a new graph embedding algorithm, LargeVis[25]. The proposed method constructs a k-nearest neighbor (k-NN) network to find the structure of similarities between high-dimensional single-cell data. This process is accelerated by an approximate k-NN construction method based on random projection trees[26] and neighbor exploring[30]. This approach optimizes a probabilistic utility function to embed the high-dimensional single-cell data into a low-dimensional space (2D or 3D). For efficient training, the utility function is approximated using negative sampling[28] that was introduced in word2vec[28]. The runtime complexity of our method is linear with regard to the number of cells, which is faster than previous single-cell visualization tools such as viSNE[8] and ACCENSE[10].

2. Methods

We propose a new approach for visualizing high-dimensional single-cell data via efficient dimensionality reduction based on LargeVis[25]. The algorithm consists of two steps: constructing an approximate k-NN network to find the similarity structure between high-dimensional single-cell data and embedding the constructed network into a 2D or 3D space while preserving the high-dimensional structure in an easily visualized low-dimensional space. Pairwise similarity between single-cell data points is determined by the distance between them in their marker expression representation space. The core assumption is that numerical proximity in the marker space is proportional to cell similarity.

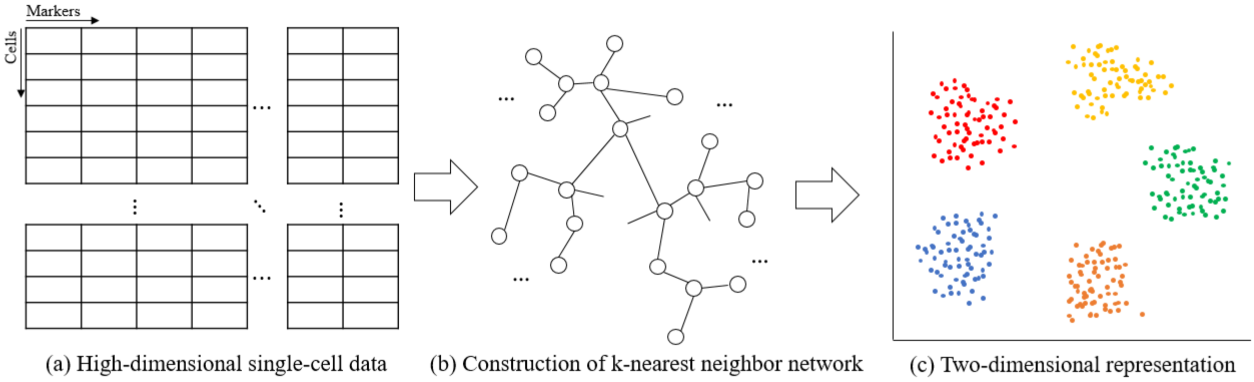

Markers
Cells
(a) High-dimensional single-cell data
(b) Construction of k-nearest neighbor network
(c) Two-dimensional representation

Figure 1. Outline of high-dimensional single-cell data visualization: constructing a k-nearest neighbor network and embedding the network into a 2D space.

2.1. *Notation*

We denote a set of high-dimensional single-cell data as $\mathcal{X} = \{x_i | x_i \in \mathbb{R}^p,\ i \in [n],\ p > 3\}$, where p is the dimension of measurements and n is the number of cells in the data; and the embedded representations of cells are denoted as $\mathcal{Y} = \{y_i | y_i \in \mathbb{R}^2\ or\ \mathbb{R}^3,\ i \in [n]\}$ in a low-dimensional space.

2.2. *Construction of a k-nearest neighbor network*

Constructing a k-nearest neighbor (k-NN) network is a very crucial step in many applications of machine learning such as a distance-based similarity search, manifold learning, and topological data analysis. Finding the exact k-NN network for large-scale single-cell data is time-consuming because it requires $O(n^2)$ time to compute all pairwise distances between all cells in the data set. Approximate methods for constructing a k-NN network have been developed, all of which have a tradeoff between speed and accuracy. Common approaches include locality sensitivity hashing[29], neighbor exploring methods[27], and partitioning methods based on random projection trees[26], k-d trees[31] and k-means trees[31].

As suggested by LargeVis[25], we develop a fast method to construct an approximate k-NN network. We first partition the whole high-dimensional space into two subspaces and generate a tree having only a root node. A set of single-cells in each partitioned subspace belongs to child nodes of the root node. Then, for the two subspaces that each set of single-cells in the child nodes belongs to, we partition each subspace into two sub-subspaces and generate two child nodes for each child node of the root node. The single-cells in each sub-subspace are assigned to each generated child nodes' child node. By continuing to partition the space iteratively, we can build a tree that assigns a group of single-cells belonging to partitioned small subspaces to its nodes. When the number of cells in a certain node is equal to or less than a predefined threshold, we stop the iterations. The single-cells in each leaf node are considered to be a candidate of approximate nearest neighbors. The generated tree is called a random projection tree.

By generating many random projection trees, we can increase the accuracy of the construction of a k-NN network, but it is time consuming. Instead of building many random projection trees, we use a neighbor search method in order to enhance both the accuracy and the efficiency. Specifically, we search the neighbor j of the neighbor of each node i assuming that its neighbor's neighbor is likely to be its neighbor also[30]. If the number of neighbors of node i is less than k, the method pushes some searched neighbor's neighbor j into the set of nearest neighbors of the node i. By iteratively doing this procedure, we can improve the accuracy of the construction and finally find our approximate k-NN network. Regarding the accuracy of the k-NN network construction, one can refer to the paper of largeVis[25], which dealt with several benchmark tests for the accuracy. The k-NN network construction process has linear time complexity because we build only a few random projection trees and because searching a certain node's neighbor's neighbor requires just a few iterations.

We then calculate the weight of each pairwise edge that represents the similarity structure of the constructed network using the Gaussian kernel, which was also used by t-SNE[23,24]. The conditional probability that the edge from data x_i to x_j is observed is first computed by:

$$\begin{aligned} p_{j|i} &= \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{(i,k)\in E} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)} \\ p_{i|i} &= 0 \end{aligned} \tag{1}$$

where the parameter σ_i is determined by setting the perplexity, and E is the set of all edges in the k-NN network. To make the network symmetric, the weights are defined as:

$$w_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{2}$$

where n is the number of input single-cell data. Since the number kn is much smaller than the number of all pairs (n^2), the constructed k-NN network is sparse. The sparsity of the k-NN network can make us compute w_{ij} within linear time complexity. Through the steps, our method can find the similarity structure of high-dimensional single-cell data within linear time complexity $O(kn)$.

2.3. *Network embedding into a low-dimensional space*

Embedding the constructed k-NN network is intended to preserve local and global network topology such that neighbors in the network are near each other in a low-dimensional space. First, for two nodes v_i and v_j, LargeVis[25] defines the probability that they come from the same neighborhood, i.e. the probability that we can observe the edge between two nodes in the k-NN network, as:

$$p(e_{ij} = 1|y_i, y_j) = f(\mathrm{dist}(y_i, y_j)) \tag{3}$$

where f is a transformation function to map the distance between y_i and y_j into a probability value.

The function f satisfies the idea that when the distance between two low-dimensional points is small, the probability observing the connection between them is high. After considering some candidates like a multinomial logistic model and a sigmoid function, we chose $f(x) = \frac{1}{1+\alpha x^2}$ ($\alpha > 0$) due to its computational simplicity. The selected function f does not require any normalization across the data set, thus only $O(n)$ runtime is needed for objective evaluation and gradient calculation in the embedding optimization (see below). In addition, we can control the thickness of the tail of the function f by controlling α. When α becomes smaller, its tail gets thicker. When $\alpha = 1$, f is Student's t-distribution with degree of freedom one except a scaling factor $\frac{1}{\pi}$. On the other hand, t-SNE[23] uses the Gaussian kernel p_{ij} of (1) and a t-distributed kernel $q_{ij} = \frac{(1+\|y_i-y_j\|^2)^{-1}}{\sum_{k\neq l}(1+\|y_k-y_l\|^2)^{-1}}$ to measure its high-dimensional and low-dimensional similarity, respectively. By minimizing the Kullback-Leibler divergence between two similarities through gradient descent, t-SNE finds its low-dimensional embedding. The gradient of its cost function contains the normalization term of q_{ij}. Computing the term requires $O(n^2)$. To avoid inefficiency, accelerated t-SNE[24] uses Barnes-Hut algorithm[32] and reduces its time complexity from $O(n^2)$ to $O(n \log n)$. Two versions of t-SNE are more expensive than our approach.

Like LargeVis[25], we chose Euclidean distance as a distance metric in a low-dimensional space because computing Euclidean distance between embedded single-cell data is simple. In addition, we can map each calculated distance to one of the various probability function values since the range of Euclidean distance is $[0, \infty)$.

To embed the high-dimensional data, we define a log likelihood utility function (4) that considers both the probabilities of all edge connections E of the constructed k-NN network and the probabilities of all negative edges E^C. Negative edges mean that pairwise single-cell connections that are not observed in the k-NN network. This idea originally comes from noise-contrastive estimation (NCE)[33], which considers estimation that differentiates its observed data from noise using nonlinear logistic regression. Using the idea of NCE, we want to discriminate the same type of cells from different types of cells. Specifically, by maximizing the first term of (4), we can make similar single-cells become closer to each other in a low-dimensional space, and by maximizing the second part of (4), we can make dissimilar single-cells move away from each other.

$$J = \sum_{(i,j)\in E} w_{ij} \log p\left(e_{ij} = 1|y_i, y_j\right) + \sum_{(i,j)\in E^C} \gamma \log(1 - p\left(e_{ij} = 1|y_i, y_j\right)) \tag{4}$$

However, considering all negative edges is computationally expensive or even intractable when input data are very large. Thus, instead of using all negative edges, we use the idea of negative sampling[28]. This approach considers only a few samples drawn from a noise distribution. We assumed $P_n(j) \sim d_j^{3/4}$ as the noisy distribution where $d_j$is the degree of node j, which was used in word2vec[28]. By letting M the number of negative samples, we can redefine the utility function as:

$$J = \sum_{(i,j)\in E} w_{ij} \log p\left(e_{ij} = 1|y_i, y_j\right) + \sum_{k=1}^{M} \mathbb{E}_{j_k \sim P_n(j)} \gamma \log(1 - p\left(e_{ij_k} = 1|y_i, y_{j_k}\right)) \tag{5}$$

Then, we optimized (5) by applying asynchronous stochastic gradient descent (ASGD)[34]. It is a powerful optimization technique which can be efficiently parallelized and can make our algorithm more scalable. ASGD can be used in this context because the network constructed by the first step is sparse and there are few memory access conflicts between the threads we used. The learning rate is determined by $\rho_t = \rho(1 - t/T)$ where T is the total number of edge samples[25], and the initial learning rate ρ_0 is determined by considering the properties of input single-cell data. The time complexity of each SGD step of (5) is $O(M)$. For a large number of data set, the number of SGD iterations is usually proportional to the number of the given data set, n. Thus, the time complexity of the optimization is $O(Mn)$, which is linear with respect to the number of samples.

3. Experiments and Discussion

3.1. *Data and data processing*

We used mass cytometry data that are provided by X-shift[12]. They consist of 10 data sets that contain mice bone marrow samples stained with surface markers, and each of them has 51 parameters. Instead of using all of them, we used 39 surface marker expressions[12,35] that were utilized for mass cytometry experiments of the immune system reference framework[35]. In addition, the data was processed through noise thresholding and asinh transformation, i.e. $y = \text{asinh}(\max(x - 1, 0)\ /5)$ like X-shift[12] and viSNE[8] applied. The data sets also offer 24 gating annotations of each cell, which were used to distinguish cells in visualization and compare the clustering performance of viSNE and our method.

3.2. *Experimental setting*

We compared our method with viSNE[8] because it is a state-of-the-art method of single-cell visualization based on nonlinear embedding like our approach. Before implementing both algorithms, we set the parameters of each method. viSNE is based on Barnes-Hut-SNE[24], which has two parameters: perplexity and theta that controls the tradeoff between speed and accuracy. In our experiments, we set the two as 30 and 0.5, respectively. Our method allows for more control and therefore has more parameters: number of trees, number of neighbors, perplexity, number of negative samples, rho, gamma, and alpha. We set the parameters considering our input data set. The first three parameters are related to constructing a k-NN network. The number of trees and neighbors can determine the shapes of a k-NN network, and perplexity is related to computing edge weights of section 2.2. The other parameters are related to network embedding. The number of negative samples is M of (5), rho is the initial learning rate, gamma is the weight of negative edges, and alpha determines the thickness of the tail of f. Table 1 shows the parameters we tuned for our visualization.

Table 1. Parameters for constructing a k-NN network and for network embedding

Parameters for constructing a k-NN network		
Number of trees	Number of neighbors	Perplexity
20 – 100	20 – 150	10 – 50

Parameters for network embedding			
Number of negative samples	Rho	Gamma	Alpha
5 – 10	1 – 10	1 – 10	< 1

All experiments for measuring the computation time were performed on a machine with Intel Xeon E5-2650 CPUs running at 2.30GHz. 40 threads were used except the experiments about the effectiveness of multiple threads.

3.3. ***Results***

3.3.1. *Visualization*

Figure 2 represents the visualization for mice bone marrow replicate 7 data set[12]. Overall, the same type of cells forms a dense subset. The number of a certain class of cells such as HSC in the data set was so small that they were difficult to distinguish from other cells and to find in our visualization. Except these cells, we can see clearly that the same type of cells gathers together and different types of cells move away from each other in a two-dimensional space. In addition, we can find some similar cells to stay together in Figure 2. For example, similar cell types like Intermediate Monocytes (red) and Classical Monocytes (yellow) appear close to each other. Two types of B cells (purple and light green) are also stay near each other.

In addition, we applied viSNE to the same data set. viSNE also represented cell subpopulations very well. The same type of cells was grouped together, and it can clearly distinguish different types of cells. In the experiments, our method tended to form denser and rounder clusters than viSNE but to have more randomly scattered samples. Due to the space limit, the visualization results of viSNE are shown through our web-based visualization tool (see section 4). We also compare our method with other embedding methods such as PCA in the tool.

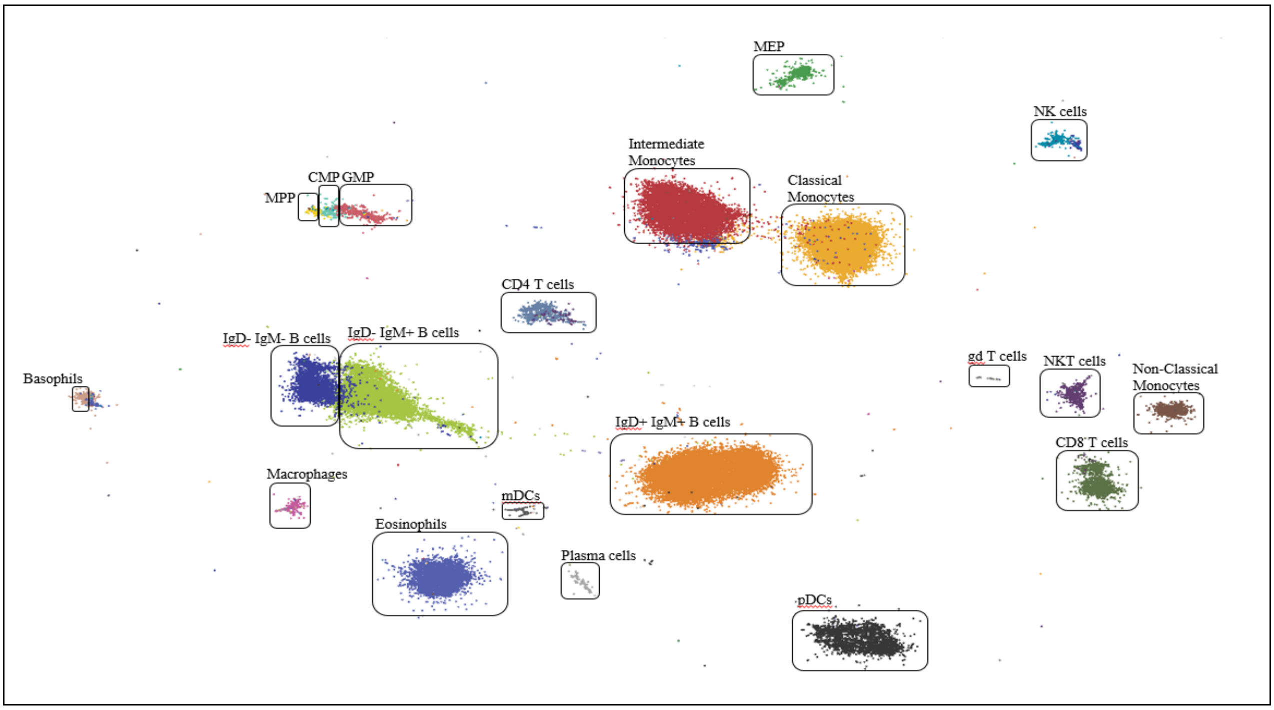


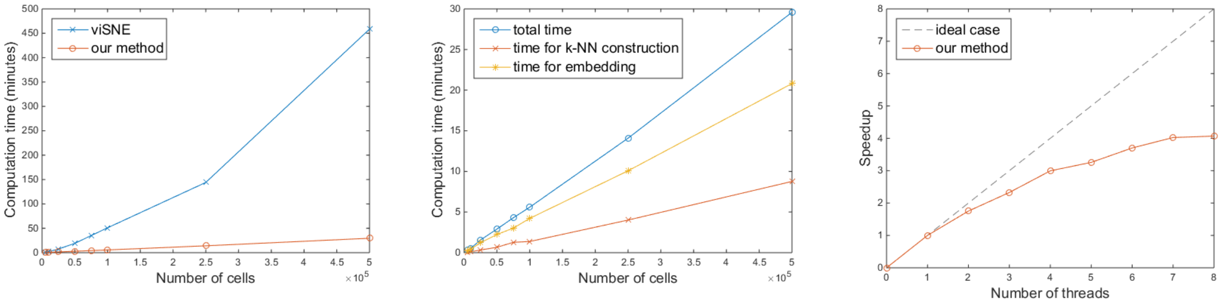
Figure 2. Visualization of our method for bone marrow replicate 7 data set.

3.3.2. *Computation time*

One of the main goals of our method is to make visualization of high-dimensional single-cell data be faster and more scalable. Thus, we compared the computation time between viSNE[8] and our method for various cases. In addition, to test the scalability and parallelizability, we measured the effectiveness of speedup with respect to the number of threads.

To measure the computation time and evaluate the scalability with respect to the size of the data set, we constructed 8 single-cell data sets that contained 5,000, 10,000, 25,000, 50,000, 75,000, 100,000, 250,000, and 500,000 data, respectively. For each data set, cells were uniformly sampled from the union of 10 data sets (total number: 841,644). Each data set contained 39 parameters and were preprocessed by noise thresholding and asinh transformation before sampling. Figure 3(a) shows that our method was faster than viSNE for all 8 sampled data sets and our method is easier to make scalable. The total computation time of our method consists of two computation times: one is for constructing a k-NN network and the other is for embedding the network. Figure 3(b) shows how much time we needed for each step.

In addition, we tested the parallelization of our method in the multi-core setting. Since our method uses asynchronous stochastic gradient descent (ASGD)[34] for training, it can be more accelerated by using multiple threads. We measured the computation time of our method when dealing with the union of all 10 single cell data sets with respect to the number of threads. By increasing the number of threads from 1 to 8, we measured the effectiveness of the multiple threads for our method. When we used 8 threads simultaneously, the speedup rate was 4.1 times faster than single-thread implementation in Figure 3(c). The results show that our method can be easily parallelized and can be made more scalable through a multi-core system.



(a) (b) (c)

Figure 3. (a) Comparison of the computation time of viSNE and our method with respect to the number of single-cell data samples. (b) Separate analysis of the computation time for constructing a k-NN network and for embedding with regard to the number of single-cell data samples. (c) Effectiveness of the multiple threads for speedup of our method.

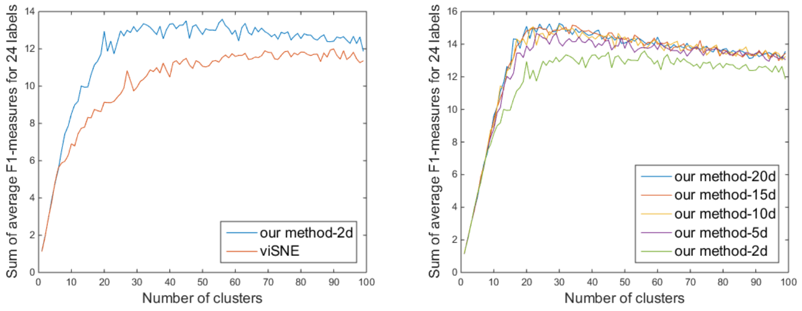
3.3.3. *Clustering*

In this section, we compared the quality of embedding by comparing the performance of clustering. In our experiments, we first applied one of the off-the-shelf clustering algorithms, k-means clustering[20] to the embedded vectors by viSNE[8] and those by our method. Next, we measured the performance of clustering using hand-gated annotations of each cell. Specifically, we followed the process of X-shift[12] to compare the clustering result and hand-gated labels and to calculate F1-measures. As the number of clusters changed from 2 to 100, we computed F1-measures for each cluster that a label was assigned to by the Hungarian algorithm[36]. This process was applied to our 10 data sets, and we obtained an average F1-measure sum. As another performance measure, we obtained maximum F1-measures for each data set across all the number of clusters and took a median.

As the input of clustering, we used the two-dimensional vectors embedded by viSNE[8] and our method. We compared an average F1-measure sum of both methods and a median of maximum F1-measures. Figure 4(a) shows that the clustering performance of our method was better than that of viSNE across all the number of clusters with respect to an average F1-measure sum. In addition, we compared a median of maximum F1-measures of viSNE and our method. Our two-dimensional embedding obtained 14.68 while viSNE obtained 13.23 as its median. Our method also outperformed viSNE for this metric.

Since our method is developed mainly for visualization, two or three dimensional vectors are usually used as a result of embedding. However, the algorithm can embed high-dimensional single-cell data into another arbitrary low-dimensional space other than a two- or three-dimensional space. The vectors embedded in a higher-dimensional space than a space for visualization can lose less intrinsic information about original high-dimensional single-cell data. Thus, they can be used to enhance the performance of clustering. We clustered the data by using 5-, 10-, 15- and 20-dimensional representations obtained by our embedding.

Figure 4(b) shows that the performance of clustering was improved when we used the vectors with higher dimensions than two. The performances as we used 10-, 15-, and 20-dimensional vectors are similar to each other and better than the performance as we used two- or 5-dimensional vectors.



(a)

(b)

Figure 4. (a) Comparison of the clustering performance of viSNE and our method when using two-dimensional vectors with respect to the number of clusters. (b) Comparison of the clustering performances when we changed the dimension of our embedding.

4. Interactive Visualization

To better aide analysis, we also introduce an interactive web browser based visualization tool featured in Figure 5. It allows researchers to examine their own data quickly by enabling functionality like mouse-over, zoom, pan, brushing, and linking on the embedded data. Users can color data by quantities of marker values as well as qualitative gate information. One can select arbitrary groups of single-cell data points, tag them, and save them for downstream analysis. We provide code, documentation, and video demonstrations to reproduce experiments and apply our methods to new single-cell data through the link[a]. All code is made available under an MIT license.

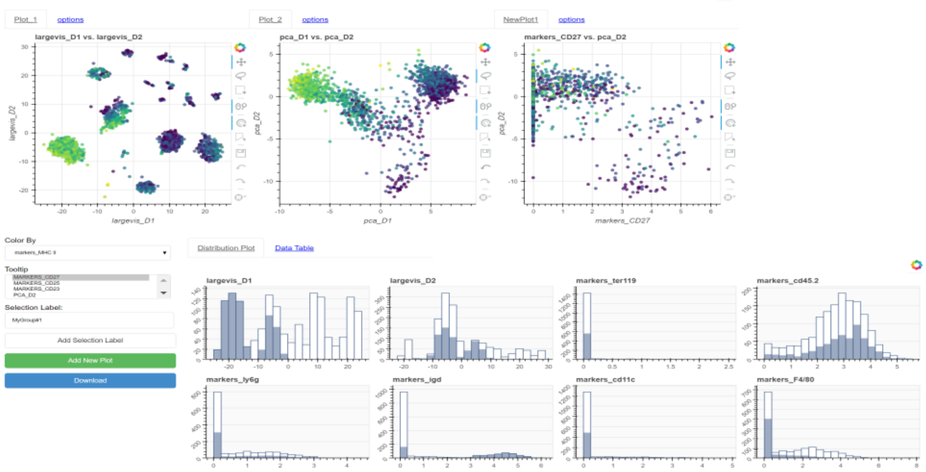


Figure 5. Screen shot of web browser based visualization developed in python. The left scatter plot depicts the result of our proposed method and on the middle, a PCA projection of the data. The right plot describes embedded expressions

[a] https://github.com/nate-russell/SVHD-Single-Cell

of a specific marker with respect to a certain projection. Color assignment and data selection labeling are also available through widgets at the bottom left. Some data statistics and the table to the right show all provided marker data and meta data regarding the single-cell data.

5. Conclusion

In this paper, we introduced a new visualization method for large-scale and high-dimensional single-cell data based on LargeVis[25], which consists of two parts: constructing an approximate k-NN network and embedding the constructed network into a low-dimensional space. Since the both steps have linear time complexity, our method is scalable and readily for analyzing large-scale single-cell data sets with hundreds of thousands or even millions of single cells. Specifically, our experiment results showed that the proposed method is much faster than viSNE[8], a state-of-the-art single-cell visualization method. In addition, through the experiments about clustering, we showed that the quality of our embedding is better than that of viSNE on cell identity mapping with respect to F1-measures. We also provide a web based interactive visualization tool and all necessary code and documentation to extend this approach to new data.

Acknowledgments

This study was supported by a Sloan Research Fellowship and a National Center for Supercomputing Applications (NCSA) Fellowship of University of Illinois at Urbana-Champaign.

References

1. O. Stegle, S. A. Teichmann, and J. C. Marioni, *Nat. Rev. Genet.* **16**, 133-145 (2015).
2. F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, *Nat. Biotechnol.* **33**, 155-160 (2015).
3. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokhare, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, *Nat. Biotechnol.* **32**, 381-386 (2014).
4. O. Ornatsky, D. Bandura, V. Baranov, M. Nitz, M. A. Winnik, and S. Tanner, *J. Immunol. Methods*. **361**, 1-20 (2010).
5. S. C. Bendall, E. F. Simonds, P. Qiu, E. D. Amir, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky, R. S. Balderas, S. K. Plevritis, K. Sachs, D. Pe'er, S. D. Tanner, and G. P. Nolan, *Science.* **332**, 687-696 (2011).
6. P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs Jr, R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis, *Nat. Biotechnol.* **29**, 886-891 (2011).
7. S. C. Bendall, G. P. Nolan, M. Roederer, P. K. Chattopadhyay, *Cell*. **33**, 323-332 (2012).
8. E.-A. D. Amir, K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan, and D. Pe'er, *Nat. Biotechnol*. **31**, 545-552 (2013).
9. S. C. Bendall, K. L. Davis, E.-A. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe'er, *Cell*. **157**, 714-725 (2014).
10. K. Shekhar, P. Brodin, M. M. Davis, and A. K. Chakraborty, *Proc Natl Acad Sci*. **111**, 202-207 (2014).
11. M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. C.

Bendall, N. Friedman, and D. Pe'er, *Nat. Biotechnol*. **34**, 637-645 (2016).

12. N. Samusik, Z. Good, M. H. Spitzer, K. L. Davis, and G. P. Nolan, *Nat. Methods*. **13**, 493-496 (2016).
13. B. Anchang, T. D. P. Hart, S. C. Bendall, P. Qiu, Z. Bjornson, M. Linderman, G. P. Nolan, and S. K. Plevritis, *Nat. Protocols*. **11**, 1264-1279 (2016).
14. A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suvà, A. Regev, and B. E. Bernstein, *Science.* **344**, 1396-1401 (2014).
15. Q. Deng, D. Ramskold, B. Reinius, and R. Sandberg, *Science*. **343**, 193-196 (2014).
16. D. R. Bandura, V. I. Baranov, O. I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J. E. Dick, and S. D. Tanner, *Anal. Chem.* **81**, 6813-6822 (2009).
17. F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani, *Nat. Methods*. **6**, 377-382 (2009).
18. C. Trapnell, *Genome Res*. **25**, 1491-1498 (2015).
19. L. A. Herzenberg, J. Tung, W. A. Moore, and D. R. Parks, *Nat. Immunol*. **7**, 681-685 (2006).
20. C. Bishop, *Springer*. (2006).
21. H. C. Fan, G. K. Fu, and S. P. A. Fodor, *Science.* **347**. 1258367 (2015).
22. D. A. Lawson, N. R. Bhakta, K. Kessenbrock, K. D. Prummel, Y. Yu, K. Takai, A. Zhou, H. Eyob, S. Balakrishnan, C. Wang, P. Yaswen, A. Goga, and Z. Werb, *Nat*. **526**, 131-135 (2015).
23. L. J. P. van der Maaten, and G. E. Hinton, *J. Mach. Learn. Res*. **9**, 2579-2605 (2008).
24. L. J. P. van der Maaten, *J. Mach. Learn. Res*. **15**, 3221-3245 (2014).
25. J. Tang, J. Liu, M. Zhang, and Q. Mei, *Proc. 25th Int. Conf. WWW*. (2016).
26. S. Dasgupta and Y. Freund, *Proc. 40th ACM STOC*. 537-546 (2008).
27. W. Dong, M. Charikar, and K. Li, *Proc. 20th Int. Conf. WWW*. 577-586 (2011).
28. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, *Proc. 26th Adv. NIPS.* 3111-3119 (2013).
29. M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, *Proc. 20th ACM SoCG*. 253-262 (2004).
30. J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, *Proc. 24th Int. Conf. WWW*. 1067-1077 (2015).
31. M. Muja and D. G. Lowe, *IEEE Trans Pattern Anal Mach Intell*. **36**, 2227-2240 (2014).
32. J. Barnes and P. Hut, *Nat*. **324**, 446-449 (1986).
33. M. U. Gutmann and A. Hyvarinen, *J. Mach. Learn. Res*. **13**, 307-361 (2012).
34. B. Recht, C. Re, S. Wright, and F. Niu, *Proc. 24th Adv. NIPS*. 693-701 (2011).
35. M. H. Spitzer, P. F. Gherardini, G. K. Fragiadakis, N. Bhattacharya, R. T. Yuan, A. N. Hotson, R. Finck, Y. Carmi, E. R. Zunder, W. J. Fantl, S. C. Bendall, E. G. Engleman, G. P. Nolan, *Science*. **349**, 1259425 (2015).
36. J. Munkres, *J. Soc. Ind. Appl. Math*. **5**, 32-38 (1957).