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.