SINGLE CELL ANALYSIS, WHAT IS IN THE FUTURE?

Lana X. Garmire[†]
Department of Computational Medicine and Bioinformatics, University of Michigan
1600 Huron Parkway, Ann Arbor, 48105, USA
Email: lgarmire@med.umich.edu

Guo-Cheng Yuan
Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard Chan School of Public Health, Boston, MA 02215, USA
Email: gcyuan@jimmy.harvard.edu

Rong Fan
Biomedical Engineering Department, Yale University
55 Prospect Street, MEC 213, New Haven, CT 06520, USA
Email: rong.fan@yale.edu

Gene W Yeo
Cellular and Molecular Medicine, University of California at San Diego
2880 Torrey Pines Scenic Dr. La Jolla, CA92037, USA
Email: geneyeo@ucsd.edu

John Quackenbush[††]
Department of Biostatistics, Harvard University
Dana-Farber Cancer Institute Smith 822A, Boston, MA 02215, USA
Email: johnq@jimmy.harvard.edu

Abstract

Single-cell genomics technology is an exciting emerging area that holds the promise to revolutionize our understanding of diseases and associated biological processes. It allows us to explore processes active in bulk tissue samples, survey tissue complexity, characterize heterogeneous cell populations and explore the role of cellular heterogeneity and interactions in disease. To deal with these new experimental data, new computational methods, software, and data portals to analyze, integrate and interpret the complexity of the system are clearly needed. The many areas where new analytical methods are needed include: (1) computational methods to identify bona fide patterns of gene

[†]LG's work is supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), R01 LM012373 awarded by NLM, and R01 HD084633 award by NICHD.
[††]JQ's work is supported by a grant from the US National Cancer Institute, R35CA220523.

expression, mutations, or DNA methylation among single cells; (2) imaging of gene expression or *in situ* transcriptomic analysis to allow study of the spatial-temporal relationships of single cells in complex tissues; (3) new tools and methods to integrate multi-omics single cell data that can handle the sparsity associated with those data, and (4) new software packages and data portals to enable cloud/HPC deployment to both developers and non-informatics end-users. Here we briefly review the state-of-the-art single cell analysis methods, ranging from clustering to visualization, and discuss the future directions of single cell bioinformatics that overcomes the computational and technical challenges as well as promotes the wide-spread adoption in biomedical research labs.

Keywords: single cell; bioinformatics; software; computation; analysis; sequencing; clustering; visualization; pipeline

1. Background

Single cell genomics represents a major breakthrough in biological science. The technology has challenged both our understanding of how cells function alone and in communities, and the methods we have developed to analyze data from bulk tissue samples[1–3]. The most widely used single-sell technology is single cell RNA-sequencing (scRNA-seq). Platforms, such as Drop-seq, Fluidigm C1 system, and 10x Genomics Chromium System, have made it possible to study a large number of single cells in various biological systems in individual labs as well a world-wide consortium, the Human Cell Atlas, which has as its goal the creation of a reference human cell data resource. Beyond understanding fundamentals of gene expression patterns in each cell, this technology has been utilized in many areas of applications, such as characterizing developmental processes, discovering new cell types, revealing the heterogeneity within tumors, depicting tumor microevolution, as well as identifying novel biomarkers for disease progression and drug resistance[4].

As an exciting frontier of genomics technology, scRNA-seq data analysis is also computationally difficult, due in part to both the technology and basic biology of single cells[5]. For example, as each cell has very limited amount of RNA molecules and the capturing technology is not even close to 100% efficient, specific RNAs may be omitted and appear as “drop-outs”, meaning that the assay fails to capture them and thus their expression value is falsely reported as zero. PCR is sometimes used to amplify RNA as part of the product, “jackpotting” can occur; leading to inflated read counts for other genes. When using droplet based methods, occasionally multiple cells may be incorporated in the droplet, leading to doublets which can confuse data interpretation. Additionally, batch effect is known in single-cell experiments, like other omics assays. All of these factors have impact on estimating true expression values and each requires the use of rigorous modeling methods to estimate the effect and correct for it.

To address various issues such as the ones stated above, we have seen numerous computational methods reported recently. There are also new bioinformatics pipelines, packages and data portals available for public use, depending on users' background and preference[6,7]. A scRNA-seq analysis pipeline usually includes the following preprocessing steps: batch-effect removal, outlier removal, normalization, imputation and gene filtering. Downstream analyses include methods for clustering, differential expression analysis, pathway/ontology enrichment analysis, protein network interaction mapping, and pseudo-time construction. Read counts, the representation of gene expression (GE), are conventionally used as the inputs for bioinformatics analysis. However, some researchers also proposed to use other information, such as small nucleotide variation (SNV) as less bias-prone features to conduct downstream functional analysis[8].

2. Summary of single cell analysis session at PSB 2019

In the single cell analysis session at PSB 2019, four submitted full-length manuscripts were accepted. They cover a range of topics from visualization, pseudo-time inference, and evaluation of clustering methods to probabilistic approach to include gene expression data for metabolic modeling.

The work from Ouyang's group reports on a new method called LISA: Landmark Isomap for Single cell Analysis. It is an unsupervised method that constructs cell trajectory and the pseudo-time relationships. The authors present a thorough comparison to two widely used methods, TSCAN and Monocle2, using both simulated and real data. Their analysis concludes that LISA captures the biology of the system being analyzed more efficiently than Monocle2 or TSCAN, yet is more computationally efficient. Thus, it can be applied to ever-larger scRNA-seq data sets and might potentially useful in the analysis of other single cell omics data.

Huang et al. use a topological analysis method called Mapper to visualize single cell RNAseq subpopulation data. Topological analysis of scRNA-seq is very interesting and allows the delineation of complex relationships that extend beyond the simple clustering that is more commonly used. The authors compared their method to tSNE and showed that Mapper better preserves continuous structure in the data.

In Wolpert and Macready's No Free Lunch Theorem paper, they argued against general purpose algorithms tested on small data sets and built without taking advantage of prior knowledge of the system being analyzed [9]. The work of Greene et al. is a case study in this regard, applied to scRNA-seq analysis. The authors analyzed the effects of parameter tuning in a variational autoencoder

(VAE) on the clustering of simulated scRNA-seq results. They warned that without proper parameter sets, deep learning results can lead to significant error.

Gold et al. presents new application of prior work on sparsely-connected autoencoders (SSCA) and variation autoencoders (SSCVA), in single cell RNA-seq analysis. This paper replaces those statistical methods that were popular in this field with machine learning methods and adds some interpretability by mapping genes to gene sets. The results of SSCVA appear to be better than SSCA, but the gene-set level extraction is not better than raw gene expression.

3. Single cell analysis, what is in the future?

At present, scRNA-seq is the most widely used method of single cell analysis. As we previously noted that there are many choices for each of the various steps along the data analysis pipeline for single cell data. However, there is no clear consensus as to what represents best practices. This, in large part, represents the fact that scRNA-seq is so new that even discoveries of apparently new cell types in a bulk tissue sample need substantial validation using other methods and independent data sets before one can consider them to be reliable. As a result, there are no reliable benchmark data sets that can be used to objectively evaluate the many methods and pipelines that are now available.

Nevertheless, scRNA-seq data sets provide the opportunity to explore tens of thousands of individual cells—data sets that dwarf the number of samples in most other gene expression studies. Such expansive data provides many new opportunities for methods development and the use of creative approaches that can handle massive yet sparse data. Ultimately, these new methods must be critically assessed, and validation will require both careful evaluation of the methods and the design and conduct of experimental studies.

What is most exciting about single cell field is that the technology continues to rapidly evolve, setting the stage for further methodology development. One particularly interesting application is spatially-informed single cell analysis, in which the spatial relationship between various cell types is preserved. Current scRNA-seq protocols first dissociate individual cells and remove debris, followed by single cell encapsulation and sequencing. Analysis of such expression data will require new computational methods to detect spatial patterns and model the relationships between cell types and their associations with various phenotypes.

Another exciting possibility is the development of multi-omic analysis in which genomic, transcriptomic, epigenomic, or other data types are collected on each cell[10,11]. Development of these methods for single cell data presents great challenges as noisy or missing data can lead to incorrect conclusions about the interaction between the sources of those data. Computational methods developed for bulk cell multi-omics integration[12], may be the first-line option for single-cell multi-omics integration, after significant efforts on cleaning, imputation, and normalization that preserve relationships between data types.

Finally, efforts that improve user's experience will be very valuable. One area is increasingly recognized as essential is the development of new methods for visual representation of complex data. With the potential to generate data on millions of cells from hundreds of cell types in a single experiment, there is a clear need for methods that can show the relationships that exist between those cell types that reflect their lineages, relationships, and interacting processes between cell types which are related to the phenotypes. GUI based data portals for interactive scRNA-seq analysis will also help the researchers to navigate through massive amount of information.

Regardless of which area one chooses to focus, it is clear that there are many opportunities for methods development and application in single cell analysis. More importantly, single cell analysis promises to help us understand the complexities of human health and disease—but only if we have appropriate analytical methods.

References

1. Eberwine, J., Sul, J.-Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nat. Methods* **11,** 25–27 (2014).
2. Poirion, O. B., Zhu, X., Ching, T. & Garmire, L. Single-Cell Transcriptomics Bioinformatics and Computational Challenges. *Front. Genet.* **7,** 163 (2016).
3. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16,** 133–145 (2015).
4. Yuan, G.-C. *et al.* Challenges and emerging directions in single-cell analysis. *Genome Biol.* **18,** 84 (2017).
5. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50,** 96 (2018).
6. Zhu, X. *et al.* Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Med.* **9,** 108 (2017).
7. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*

36, 411–420 (2018).

8. Poirion, O. B., Zhu, X., Ching, T. & Garmire, L. X. Using Single Nucleotide Variations in Single-Cell RNA-Seq to Identify Subpopulations and Genotype-phenotype Linkage. *bioRxiv* 095810 (2018). doi:10.1101/095810
9. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1,** 67–82 (1997).
10. Packer, J. & Trapnell, C. Single-Cell Multi-omics: An Engine for New Quantitative Models of Gene Regulation. *Trends Genet.* **34,** 653–665 (2018).
11. Ortega, M. A. *et al.* Using single-cell multiple omics approaches to resolve tumor heterogeneity. *Clin. Transl. Med.* **6,** 46 (2017).
12. Huang, S., Chaudhary, K. & Garmire, L. X. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front. Genet.* **8,** 84 (2017).