Research Statement

Avi Srivastava | May 3, 2019

My Ph.D. dissertation & research interest is in contributing to a better understanding of dynamic gene expression landscape through analysis efficient algorithms and uncertainty aware graphical models. This broad goal has led me into the field of Computational Biology and the study of RNA Sequencing (RNA-seq).

Published Research

RNA-seq technologies are evolving rapidly, and with them, the requirement for fast and accurate tools to analyze the generated data. The first step of many RNA-seq analyses requires us to solve the problem of read-alignment. When reads are aligned to a collection of reference sequences that share a substantial amount of sub-sequence (near or exact repeats), a single read can have many potential alignments, and considering all such alignment can be crucial for downstream analysis.

I introduce a novel concept, quasi-mapping, and an efficient algorithm implementing this approach called `RapMap` which maps RNA-seq reads (sequences) to the reference sequence(s) (transcriptome). `RapMap` is capable of *mapping* sequencing reads to a target transcriptome substantially faster than existing alignment tools. The algorithm I employ to implement quasi-mapping uses several efficient data structures and takes advantage of the special structure of shared sequence prevalent in transcriptomes to rapidly provide highly-accurate mapping information. `RapMap` is well accepted by community, till date, the manuscript is downloaded $\sim$9k times and cited 44 times [10].

While bulk RNA-seq is an established method to perform genome-wide quantification of gene expressions[9], however, bulk experiments average-out the expression pattern of an individual cell (or a cell type) across millions of cells, losing cell-level heterogeneity which is crucial to understand the gene expression landscape. Moreover, quantification tools for bulk RNA-seq cannot be directly used for droplet based single cell RNA-seq (dscRNA-seq) data[2, 4, 7]. I introduced alevin, a fast end-to-end pipeline to process dscRNA-seq data, addressesing the inherent bias in existing tools which discard gene-ambiguous reads and improves the accuracy of gene abundance estimates. `Alevin` is considerably faster, typically 8 times, than existing gene-quantification approaches, while also using less memory.

Future Research

Given the success of both `RapMap` and `alevin`, I continued pursuing this line of research and collaborated to develop other downstream tools [1, 3, 5, 6, 8]. In addition, I plan to generalize both our methods to make it more robust and extend it further. As quasi-mapping is a novel concept but it trades-off speed with accuracy, which is usually enough for a typical RNA-seq experiment, however, it loses accuracy in *relatively* complex datasets. In my preliminary study, I have already optimized `RapMap` for efficient, dynamic (dataset dependent) speed / accurcay tradeoff. Moreover, as `alevin` internally uses `RapMap` for transcriptome mapping, I believe improvement in mapping would be directly reflected on the performance of `alevin` while also improving the performance of other downstream tools.

277 Sheep Pasture – East Setauket-11733, New York
(631)974-8087 • asrivastava@cs.stonybrook.edu

References

[1] Srivastava, Avi, et al. "Accurate, fast and lightweight clustering of de novo transcriptomes using fragment equivalence classes." arXiv preprint arXiv:1604.03250 (2016).

[2] Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., ... & Trombetta, J. J. (2015). "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets". Cell, 161(5), 1202-1214.

[3] Zakeri, M., Srivastava, A., Almodaresi, F., & Patro, R. (2017). "Improved data-driven likelihood factorizations for transcript abundance estimation". Bioinformatics, 33(14), i142-i151.

[4] Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., ... & Kirschner, M. W. (2015). "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells". Cell, 161(5), 1187-1201.

[5] Almodaresi, F., Sarkar, H., Srivastava, A., & Patro, R. (2018). "A space and time-efficient index for the compacted colored de Bruijn graph". Bioinformatics, 34(13), i169-i177.

[6] Zhu, A., Srivastava, A., Ibrahim, J. G., Patro, R., & Love, M. I. (2019). "Nonparametric expression analysis using inferential replicate counts". BioRxiv.

[7] Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., ... & Gregory, M. T. (2017). "Massively parallel digital transcriptional profiling of single cells". Nature communications, 8, 14049.

[8] Sarkar, H., Srivastava A., Patro R. (2019) "Minnow: A principled framework for rapid simulation of dscRNA-seq data at the read level" ISMB-2019

[9] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". Nature methods, 5(7), 621.

[10] rxivist: webpage https://rxivist.org/papers/5324