Research Statement

Avi Srivastava | May 3, 2019

My main research interest is to a better understand the dynamic gene expression landscapes through efficient algorithms and statistical inference approaches that account for uncertainty. This broad goal has led me into the field of Computational Biology and the study of RNA Sequencing (RNA-seq) experiments, in particular.

Published Research

RNA-seq technologies are evolving rapidly, and with them, the requirement for fast and accurate tools to determine and analyze the origin of reads in the generated data. The first step of many RNA-seq analyses requires 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 alignments can be crucial for downstream analysis.

I introduced 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 employed 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 the community, and to date, the manuscript has been downloaded $\sim$9k times and cited 44 times [10].

While bulk RNA-seq is an established method to perform genome-wide quantification[9], bulk experiments average-out the expression patterns of individual cells (or cell types) 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, addressing the inherent bias in existing tools which discard gene-ambiguous reads and improving the accuracy of gene abundance estimates. `Alevin` is considerably faster, typically 8 times faster, 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 methods to make them more robust and extend their capabilities. While quasi-mapping is a useful novel concept, it trades-off some accuracy for speed. This does not have a huge impact on typical RNA-seq experiments. However, in *relatively* complex datasets, the loss in accuracy can be significant. In ongoing work, I have updated `RapMap` to employ alignments to improve quantifications in an efficient, dynamic (dataset dependent) way. This optimizes the speed/accuracy tradeoff, making the estimates almost as accurate as end-to-end alignment-based methods, while still using resources similar to `RapMap`. Moreover, as `alevin` internally uses `RapMap` for transcriptome mapping, I believe the improvement in mapping quality will be directly reflected in the accuracy 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

277 Sheep Pasture – East Setauket-11733, New York

(631)974-8087 • *asrivastava@cs.stonybrook.edu*