Introduction

I am passionate about the design and development of efficient, biotechnology-driven computational methods that can be used to convert large genomic data sets into biologically and clinically useful information. Over the years, this passion has led me through the study of different computational problems associated with various high-throughput sequencing technologies, while the main focus of my research has been the design and development of methods to generate accurate gene expression estimates using bulk and single-cell RNA-sequencing (RNA-seq) data. I believe that transcriptome sequencing, particularly single-cell RNA-seq, is a crucial measurement technique for understanding the Waddington landscape of differentiating cells. My future goal is to map the landscape of cellular states to identify key genes involved in governing a cell's fate using, in part, the efficient infrastructure which I am developing for estimating gene expression.

Current Research: 'Analysis-efficient' RNA-seq quantification

As a Computer Science Ph.D. student at Stony Brook University (since 2014), working with my advisor Dr. Rob Patro, I have focused largely on the problem of RNA-seq quantification. My interest in the field was motivated by the exponential increase in the amount of transcriptomic data [1], along with the variety of available technologies [2] for both bulk and single-cell gene-expression studies. The increase has resulted in many different abundance estimation pipelines, which may not always be optimal. Broadly, I am interested in developing "analysis-efficient" algorithmic approaches that are capable of processing this data in a fundamentally scalable fashion, and which focus on computing only those quantities necessary to answer the scientific question at hand accurately.

Read Alignment: 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 solving 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 in a tool called `RapMap`, which maps RNA-seq reads (sequences) to the reference (transcriptome) sequence(s). `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 accurate mapping information. `RapMap` is well accepted by the community, and to date, the manuscript has been downloaded more than 9k times and cited more than 50 times [3].

Single Cell quantification: Recent advances in biotechnology have enabled us to perform sequencing experiments where we can assay the gene expression profiles across 10s or even 100s of thousands of individual cells. This capability is distinct from previous approaches to bulk RNA-sequencing, since it retains the identity of individual cells. This allows scientists to explore, for example, how the gene expression profiles of cells change during differentiation and development, or how a population of cells responds to drug treatment and the characteristics of a treatment-resistant subpopulation. While bulk RNA-seq is an established method to perform genome-wide quantification[4], these experiments average-out the expression patterns of individual cells (or cell types) across millions of cells, losing cell-level heterogeneity which is crucial to understanding 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[5–7].

While methodological development in this field is moving at a rapid pace, surprisingly little attention has been paid to the first steps in the fundamental processing of scRNA-seq data. I introduced `alevin`, a fast end-to-end pipeline to process dscRNA-seq data which starts from the 'raw' sequencing reads and generates a cell-level profile of gene expression. In addition to being over an order of magnitude faster than existing tools, alevin addresses a conceptual shortcoming in existing approaches. While current tools cannot account for sequencing reads that may have arisen from more than a single gene (∼15 - 23% of the data), and so they simply discard this data, alevin introduces a consistent and principled model to account for this data when quantifying gene expression. Alevin's algorithm enables fundamentally more accurate quantification of single-cell gene expression as it address the inherent bias in existing tools which discard gene-ambiguous reads and improves the accuracy of gene abundance estimates.

Ongoing Research: RNA-seq quantification is *not* solved

After the success of both `RapMap` and `alevin`, I have continued pursuing this line of research and collaborated with peers and PIs from across different institutions. These collaborations resulted in the development of multiple novel methods [8–12] that optimize the efficiency of both bulk and single-cell RNA-seq quantification.

Mappings Matter: Even though quasi-mapping has proven to be useful, it trades-off some of the accuracy of traditional alignment for speed. This led to my ongoing project, where I introduce selective-alignment, a novel method that optimizes between the accuracy of alignment-based approaches and the speed of lightweight mapping methods. Moreover, this project provides a principled comparitive analysis of several read-alignment tools used for estimating gene and transcript abundance, which is the first major step in every gene-expression study. In a preliminary analysis [13], I explore in detail, among various alignment-based approaches, the cause of non-trivial differences between quantifications based upon mapping to the transcriptome (both lightweight and full alignment) and quantifications based upon mapping to the genome and subsequently projecting these alignments into transcriptomic coordinates. Among many relevant observations, the most striking one is that overreliance on simulations has led the community to falsely believe alignment free approaches have equal accuracy as traditional aligners, which is not always true. In fact, in the cases where the two methods differ, traditional aligners seen to report the right alignments and have higher quantification accuracy.

alevin 2.0: A human body contains trillions of cells, and a single progenitor cell can generate the astonishingly high diversity observed in differentiated cell types. Multiple factors, both known and unknown, affect a cell's fate while differentiating, but our understanding of the process is limited. Recent developments in single-cell RNA-seq technologies have been transformative for our knowledge of the process at the molecular level. However, current studies are limited by the use of per-cell gene-counts, which is a gross approximation of the true counts of the molecules in the sample. In my previously proposed method, alevin, the novel UMI deduplication algorithm begins by constructing parsimonious UMI graphs (PUGs) using information from the UMI sequences, the UMI counts, and the transcript equivalence classes. Since PUGs hold more fundamental information (i.e. sequence homology), which is ingrained in its graphical structure, I hypothesize that a set of these subnetworks could be used to better represent a cell's position in the global landscape. I am working on a method to use these PUGs to reconstruct single-cell trajectories more accurately than traditional methods that use the summarized gene-count matrices.

Future Research Interests

Most RNA-seq measurements, including single-cell RNA-seq based gene-expression profiles capture only a static snapshot at a point in time, which limits the power of computational methods for estimating time-dependant phenomena. The difficulties in understanding the dynamic transcriptional process through a static snapshot gets even more extreme when the generated gene-expression profile has estimation bias; for example I show in alevin that in 3'-end single cell protocol how the estimation bias when ignoring gene-ambiguous reads can affect the gene-abundance estimates. Although gene-expression profiling is a strong indicator of the state the cells in the full landscape, I believe RNA-sequencing alone is not enough for accurately understanding the role of individual cells. I am particularly interested in learning how combining different assays, like ATAC-seq or spatially-resolved data, with RNA-seq, can improve the accuracy of analyses and increase our knowledge of the cellular landscape. In the process, I hope to help design novel methods and biotechnology to better characterize and understand transcriptional dynamics in a complex system.

References

[1] Paul Muir, Shantao Li, Shaoke Lou, Daifeng Wang, Daniel J Spakowicz, Leonidas Salichos, Jing Zhang, George M Weinstock, Farren Isaacs, Joel Rozowsky, et al. "The real cost of sequencing: scaling computation to keep pace with data generation." Genome Biology, 17(1): 53, 2016.

[2] Svensson, Valentine, et al. "Power analysis of single-cell RNA-sequencing experiments." Nature methods 14.4 (2017): 381.

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

[4] 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.

[5] 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.

[6] 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.

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

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

[10] 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.

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

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

[13] Srivastava, Avi, et al. "Alignment and mapping methodology influence transcript abundance estimation." BioRxiv (2019): 657874.