Research Statement

Hirak Sarkar

RNA-seqencing has become the de-facto standard for measuring transcript and gene-level expression across different cells, tissues, organisms and species. Given the tremendous popularity of the RNA-seq assay, there are numerous tools that are designed to store (efficient disk usage), process (map or align), and analyse (quantification, differential expression etc.) this data.

Throughout my doctoral studies, I discovered interesting challenges in each of these steps, and proposed solutions that either improved the performance of existing methods or introduced a new approach altogether. To be more specific, I studied the nature of sequence redundancy in RNA-seq dataset. I observed that the presence of identical sequences in the RNA-seq experiments, often originating from shared exons or paralogous references etc, plays a major role in many challenges related to RNA-seq data analysis. Through different research projects, I tried to both address the challenges of multi-mapping [8, 1, 11] in mapping and quantification of RNA-seq data and, simultaneously have shown ways to turn this challenge into an opportunity to increase the efficiency of the computational pipeline [9, 15].

Published Research

I introduced Quark [8], a semi reference-based compression algorithm for RNA-seq data, which leverages the shared sequences present in the raw fastq files. The core compression algorithm makes use of read level "equivalence classes", a derivative of another popular light-weight algorithm "Quasi-mapping" that I co-authored [14]. Quark stores parts of reference only once that are redundantly present multiple times in the given read dataset, represented in the form of the sequence that induces these equivalence classes. Although the reference is required for compressing the reads, it is not required in the phase of decompression, effectively making the algorithm independent of the reference used for compression (this is where the moniker semi reference-based comes from). Later, I contributed to designing Pufferfish [1], a succinct graph-based indexing scheme for RNA-seq data, that, right now, serves as the basic backbone for the principle mapper in the widely popular quantification tool Salmon [7]. The multi-mappable reads resulting from the sequence redundancy can also lead to uncertain quantification estimates. I addressed this problem in the recently published tool Terminus [9], where we proposed a solution that produces groups of transcripts when read-level evidence is insufficient to provide robust quantification estimates of individual transcripts. Moreover, we have shown that such groups of transcripts can capture biological information (such as gene families) even though the tool itself has no information about underlying annotation.

While bulk RNA-seq provides reproducible, highly-sensitive transcript-level expression measurements, the individual signals from different cells are lost. Single-cell RNA-seq technology, which has become widely popular in the last few years, enables cellular level resolution of RNA-seq expression profiling. As a result, a myriad of publicly available datasets are now available providing gene expression from multiple (ranging from thousands to millions) cells over many organs and tissues. To aid the analysis of these huge datasets, new computational techniques are developed extending the existing tools for bulk RNA-seq datasets. I observed that, in the absence of ground truth knowledge,

5812 Quebec St – Berwyn Heights, MD-20740
(631) 520 8131 • hsarkar@umd.edu • www.hiraksarkar.com

it is often difficult to assess the accuracy of such tools. This lead me to develop a sequence-level simulator Minnow [10] for droplet-based single-cell RNA-seq data. Minnow can mimic the pattern of sequence level multi-mapping of real-world datasets, and is capable of simulating reads from many thousands of cells in a multi-threaded fashion. I have also contributed to improving other methodologies [13, 15, 12, 11] that further enhanced and improved the performance of existing bulk and single-cell RNA-seq pipelines.

Future Research

Although my focus has been the analysis of RNA-seq datasets in both the single-cell and bulk context, I feel it only captures one dimension of the multi-modal and complex cell heterogeneity that exists. Other data modalities, such as DNA methylation (GEM-seq), chromatin accessibility (single-cell ATAC-seq), and spatial information (FISH-seq) not only mitigate the caveat of one protocol (such as missing expression values from one of the assays) but can also validate the conclusions drawn from the data. With this goal in mind, I aim to design computational pipelines that can analyze and integrate sequence-level data from multiple sources. A typical data processing pipeline for a particular assay runs independently. Therefore, even when the results from another assay are present from a matched dataset, this information is not being utilized. A principled method that captures the inherent conflicts in the results from different assays could take advantage of this and further improve the accuracy and robustness of the computational estimates being made. The theoretical underpinning of such a representation can be achieved by defining a transcriptional "latent space", where the set of observed variables can represent the assays at our disposal. On a conceptual level, this idea exists in the field of natural language processing and image processing, widely known as "domain adaptation" [3, 5]. Although similar efforts have been made in single-cell multi-omics dataset [6, 2, 4], such approaches lack generalizability and interpretability. I believe building a large scale repository of the matched publicly-available multi-omics dataset, along with the corresponding transcriptional latent spaces (via both existing and newly proposed methods) can improve the usability and accuracy of the multi-modal analysis.

References

[1] Fatemeh Almodaresi*, Hirak Sarkar*, Avi Srivastava, and Rob Patro. A space and time-efficient index for the compacted colored de bruijn graph. *Bioinformatics*, 34(13):i169–i177, 2018 (* equal contribution).

[2] Matthew Amodio and Smita Krishnaswamy. Magan: Aligning biological manifolds. *arXiv preprint arXiv:1803.00385*, 2018.

[3] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[4] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, 2019.

[5] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by back-propagation. *arXiv preprint arXiv:1409.7495*, 2014.

[6] Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. Jointly embedding multiple single-cell omics measurements. *BioRxiv*, page 644310, 2019.

[7] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417, 2017.

[8] Hirak Sarkar and Rob Patro. Quark enables semi-reference-based compression of rna-seq data. *Bioinformatics*, 33(21):3380–3386, 2017.

[9] Hirak Sarkar, Avi Srivastava, Hector Corrada Bravo, Michael I Love, and Rob Patro. Terminus enables the discovery of data-driven, robust transcript groups from rna-seq data. *bioRxiv*, 2020.

[10] Hirak Sarkar, Avi Srivastava, and Rob Patro. Minnow: a principled framework for rapid simulation of dscrna-seq data at the read level. *Bioinformatics*, 35(14):i136–i144, 2019.

[11] Hirak Sarkar, Mohsen Zakeri, Laraib Malik, and Rob Patro. Towards selective-alignment: Bridging the accuracy gap between alignment-based and alignment-free transcript quantification. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 27–36, 2018.

[12] Avi Srivastava, Laraib Malik, Hirak Sarkar, and Rob Patro. A bayesian framework for inter-cellular information sharing improves dscrna-seq quantification. *bioRxiv*, 2020.

[13] Avi Srivastava, Laraib Malik, Hirak Sarkar, Mohsen Zakeri, Fatemeh Almodaresi, Charlotte Soneson, Michael I Love, Carl Kingsford, and Rob Patro. Alignment and mapping methodology influence transcript abundance estimation. *BioRxiv*, page 657874, 2019.

[14] Avi Srivastava, Hirak Sarkar, Nitish Gupta, and Rob Patro. Rapmap: a rapid, sensitive and accurate tool for mapping rna-seq reads to transcriptomes. *Bioinformatics*, 32(12):i192–i200, 2016.

[15] Avi Srivastava, Hirak Sarkar, Laraib Malik, and Rob Patro. Accurate, fast and lightweight clustering of de novo transcriptomes using fragment equivalence classes. *arXiv preprint arXiv:1604.03250*, 2016.