BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Patro, Robert

eRA COMMONS USER NAME (credential, e.g., agency login): rpatro

POSITION TITLE: Assistant Professor of Computer Science

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
University of Maryland, College Park, MD	B.S.	05/2006	Computer Science
University of Maryland, College Park, MD	Ph.D.	08/2012	Computer Science
Carnegie Mellon University, Pittsburgh, PA	Postdoctoral	08/2014	Computational Biology

A. Personal Statement

My research centers around the design and development of efficient algorithms and computational methods for the processing and analysis of high-throughput genomic data. I focus, specifically, on the design of *analysis efficient* approaches, which seek to expend no more computational resources than necessary to accurately answer the biological questions at hand. While my formal computer science training helps me to design asymptotically efficient solutions, I also place a strong emphasis on the practical efficiency of the methods I develop. My extensive programming and software development experience allows me to produce high-quality, production-ready and Free/Libre and Open-Source Software (FLOSS) tools that have found considerable use in both academia and industry. My recent research has focused on computational transcriptomics, where I have introduced the idea of alignment-free transcript quantification from RNA-seq data. These ideas were first put forth in our method, Sailfish, that performed k-mer-based transcript quantification orders of magnitude faster than existing alignment-based solutions. We have recently extended upon these ideas by developing the notion of quasi-mapping, a lightweight proxy for traditional alignment that we have demonstrated can be used for accurate transcript quantification and *de novo* transcript clustering. Our recently-developed tool, *Salmon*, improves on both the speed and accuracy of our approach in Sailfish, and combines quasi-mapping with a novel dual-phase inference algorithm and transcriptome-wide models of sequence-specific, fragment GC and position-specific bias to produce accurate, bias-corrected transcript abundance estimates. Salmon also efficiently generates estimates of the inherent uncertainty present in its abundance estimates. Our recently-developed method alevin extends our quantification tool into the domain of single-cell data, and introduces new efficient and principled algorithms for quantifying gene expression from high-throughput, tagged-end sequencing data. These tools are representative of our dedication to developing computationally efficient and highly-accurate methods for processing high-throughput sequencing data.

1. **Patro, R,** Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature biotechnology 32.5 (2014): 462-464. PMCID:PMC4077321.
2. Avi Srivastava, Laraib Malik, Tom Sean Smith, Ian Sudbery, **Rob Patro**
Alevin efficiently estimates accurate gene abundances from dscRNA-seq data
bioRxiv 335000; doi: https://doi.org/10.1101/335000

3. Srivastava, A., Sarkar, H., Gupta, N., and **Patro, R**. (2016). RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics*, *32*(12), i192-i200. PMCID:PMC4908361.
4. **Patro, R.**, Duggal, G., **Love, M.I.,** Irizarry, R.A., Kingsford, C. (2016). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods, 14(4), 417-419. PMCID: PMC5600148.

B. Positions and Honors

Positions and Employment

2006-2012 Research & Teaching Assistant, Computer Science Department, University of Maryland, College Park.
2012-2014 Postdoctoral Research Associate, Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. (Carl Kingsford, Supervisor)
2014-Present Assistant Professor, Department of Computer Science, Stony Brook University, Stony Brook, NY

Other Experience and Professional Memberships

2013,2015-2018 Member, International Society for Computational Biology
Program committee, ISMB, RECOMB-Seq, HiCOMB, WABI, ACM-BCB, RECOMB (2019), RECOMB 2019 poster chair, IPDPS (2019)
2017 Program committee, ACM-BCB, Asia Pacific Bioinformatics Conference (APBC), HiCOMB 2017, RECOMB-Seq
2016 Program committee, Asia Pacific Bioinformatics Conference (APBC), RECOMB-Seq
2015 Poster award committee, RECOMB
2014 Poster selection committee, RECOMB
2010 Publicity co-chair, ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (i3D)

Honors

2017 ACM-BCB Best Paper Award
2012, 2013 ISMB travel fellowship awardee
2010 Goldhaber Travel Grant Awardee, *University of Maryland*
2009 i3D Student Stipend Award
2007-2008 Verizon Fellowship Recipient
2006-2008 Block Fellowship, Department of Computer Science, *University of Maryland*
2006 CMPS & Computer Science Departmental Honors Graduate, *University of Maryland*
2002 Charles O. Thompson Scholar, Worcester Polytechnic Institute

C. Contribution to Science

1. *Lightweight methods for transcript abundance estimation from RNA-seq data.*

A recent focus of my research, since my postdoctoral studies, has been on the development of *lightweight* methods for transcript abundance estimation. Traditional alignment-based approaches for transcript abundance estimation pose a computational bottleneck in the large-scale processing of RNA-seq data, a problem that is only growing with our capabilities for high-throughput data acquisition. This bottleneck also limits the feasibility of re-processing existing data in the light of new biological discoveries (e.g., newly-annotated transcripts). I devised an alignment-free method that infers transcript abundances, in a maximum likelihood framework, based on the counts of k-mers compatible with each transcript, and led the development of the associated software tool, Sailfish. This approach produced accurate abundance estimates orders of magnitude faster than existing alignment-based methods available at the time. Continuing this line of work, I subsequently led the development of an even faster and more accurate transcript quantification methodology, implemented in the software tool Salmon. Salmon introduces and efficient dual-phase inference algorithm that couples stochastic variational inference with batch inference (via an EM or VBEM algorithm) over a reduced representation of the RNA-seq data. It also replaces k-mer counting with an ultra-fast lightweight proxy for alignment (quasi-mapping). The reduced representation adopted by Salmon admits fast estimation of quantification uncertainty via Gibbs sampling. Finally, working in close collaboration with Dr. Michael Love, we

developed expressive bias models that allow Salmon to learn and correct, *in silico*, some of the most pervasive technical biases common in RNA-seq. Salmon represents a state-of-the-art methodology for transcript abundance estimation that is already finding widespread use in the community. As we continue to expand upon its capabilities, we expect that it will become one of the *de facto* tools in many RNA-seq pipelines.

a. **Patro, R,** Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology* 32.5 (2014): 462-464. PMCID:PMC4077321.
b. **Patro, R.,** Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods, 14(4), 417-419. PMCID: PMC5600148.
c. The Salmon software for accurate, fast, and bias-aware transcript abundance estimates using dual-phase inference: https://combine-lab.github.io/salmon
d. Zakeri, M., Srivastava, A., Almodaresi, F., **Patro, R.** (2017). Improved data-driven likelihood factorizations for transcript abundance estimation, *Bioinformatics*, Volume 33, Issue 14, 15 July 2017, Pages i142–i151, https://doi.org/10.1093/bioinformatics/btx262 PMCID:PMC5870700

2. *Algorithmic advancements in processing, indexing and searching high-throughput sequencing data.*

An important component of my current research agenda has been the development of novel data structures and algorithms for solving basic problems in the storage and processing of high-throughput sequencing data. Working in collaboration with colleagues Stony Brook, I helped develop a new approximate membership query (AMQ) data structure called the Counting Quotient Filter (CQF). This space and cache-efficient data structure overcomes many of the limitations of more traditional AMQs (like the Bloom filter), and allows for efficient maintenance of frequency information associated with each item. Subsequently, we successfully applied this data structure to numerous problems that arise in bioinformatics, including k-mer counting (resulting in a tool called Squeakr), weighted de Bruijn graph representation (resulting in a tool called deBGR), and large-scale sequence search (resulting in a tool called Mantis). In addition to applying our novel AMQ data structure, these methods exploit domain-specific observations and data characteristics to solve the corresponding problems in an efficient and scalable manner. For example, Mantis allows sequence-level search across thousands of *raw* (i.e. unassembled) sequencing experiments in less space and up to 100 times faster than existing state-of-the-art approaches, and is exact whereas the previous approaches are approximate. We have also developed a time and space efficient index for the compacted colored de Bruijn graph, called Pufferfish, which allows using this data structure as a sequence search index over one or more reference sequences. The key properties of this data structure are that it provides very fast pattern search while still maintaining a moderate (and tunable) memory footprint. This will allow us to expand the scope of many lightweight sequence analysis algorithms that rely on fast mapping approaches which, currently, are not scalable to large collections of reference sequences given the existing indexing techniques. By making fundamental improvements to the data structures and algorithms used to solve numerous bioinformatics tasks, we anticipate that the research we are doing along these lines will have a broad impact on many different analysis tasks within computational biology.

a. Pandey, P., Bender, M. A., Johnson, R., and **Patro, R.** (2017). deBGR: an efficient and near-exact representation of the weighted de Bruijn graph. *Bioinformatics*, *33*(14), i133-i141. PMID: 28881995
b. Pandey, P., Bender, M. A., Johnson, R., and **Patro, R.** (2018). Squeakr: an exact and approximate k-mer counting system. *Bioinformatics*, 34(4), 568-576. PMID: 29444235
c. Pandey, P., Almodaresi, F., Bender, M. A., Ferdman, M., Johnson, R., and **Patro, R.** (2018). Mantis: A Fast, Small, and Exact Large-Scale Sequence Search Index. *Cell Systems* (also appeared at RECOMB 2018). (doi:https://doi.org/10.1016/j.cels.2018.05.021). PMID:29936185
d. 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. PMID:29949982

3. *Assessment and improvement of de novo transcriptomes.*

Another recent focus of my research has been on how to evaluate and improve *de novo* transcriptome assemblies. Current transcriptome assembly approaches, in the presence of short reads, suffer from fundamental identifiability issues. These problems are exacerbated by bias and non-uniformity in the

underlying sequencing data, and, to some extent, the de Bruijn graph-centric approaches that have been adopted out of computational necessity. With collaborators, I helped to develop the Transrate software, and the associated computational and statistical methodologies, that allow for the *contig-level* scoring of *de novo* transcriptome assemblies. Based on a combination of different metrics that capture the most common pitfalls encountered by short read transcriptome assemblers, and by using the read data itself to evaluate the underlying assembly, Transrate provides a score for each contig that measures its overall quality. Further, *de novo* assemblies can be filtered by these scores, discarding the most problematic contigs, which we demonstrate can improve downstream analysis. My group also developed a lightweight methodology that can accurately aggregate and cluster contigs in *de novo* assemblies into putative gene groups. Our methodology in tackling this problem brings to bear the algorithmic advances that assisted in our lightweight quantification tools. Contigs are linked together by the reads that multimap between them, and the underlying "fragment ambiguity graph" can be filtered and clustered to recover groups of contigs that likely arise from the same underlying gene. We have recently built upon this methodology further, and developed a new approach that uses the same fragment ambiguity graph to apply a graph regularized semi-supervised learning algorithm to the problem of annotating *de novo* transcriptome assemblies using known annotations from related organisms. By sharing information *within* the *de novo* assembly, this approach achieves greater accuracy than existing annotation methodologies. The methods we are developing and the tools that implement them are improving the reliability of *de novo* transcriptome analysis, making this powerful but challenging approach to the study of non-model organisms, and the discovery of novel expressed transcripts in model organisms, fundamentally more useful.

a. Smith-Unna, R., Boursnell, C., **Patro, R.**, Hibberd, J. M., and Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome research*, *26*(8), 1134-1144. PMCID:PMC4971766.
b. The Transrate software for *de novo* transcriptome assembly quality analysis: http://hibberdlab.com/transrate
c. Srivastava, A., Sarkar, H., Malik, L., and **Patro, R**. (2016). Accurate, Fast and Lightweight Clustering of de novo Transcriptomes using Fragment Equivalence Classes. In refereed proceedings of RECOMB-Seq 2016 (*arXiv preprint arXiv:1604.03250*.)
d. Zhang, R., Calixto, C.P., Tzioutziou, N.A., James, A.B., Simpson, C.G., Guo, W., Marquez, Y., Kalyna, M., **Patro, R.**, Eyras, E. and Barta, A., (2015). AtRTD–a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in Arabidopsis thaliana. *New Phytologist*, *208*(1), 96-101. PMCID:PMC4744958.

4. *Computational analysis of 3D genome structure from {3,4,5,Hi}-C data.*

I developed novel computational approaches for the analysis of the 3D structure of genomes from high throughput chromosome conformation capture (3C) and related assays. This work included the development of new approach for the discovery of topologically associated domains (TADs), as well as the first method and tool capable of naturally finding TADs which exist at different characteristic length scales. Using this approach, we also carried out the first statistically rigorous analysis demonstrating that TADs not only exist at different characteristic length scales, but that they are also "nested", forming a hierarchy of chromatin domains within the cell. We also developed novel graph-based methodologies to filter 3C and related data accounting for the fact that the resulting data should adhere to certain metric constraints. We demonstrate that topological characteristics in this graph correlate with known genomic function, and can be used to detect and assess spatial proximity of regions in the underlying genome. This work has advanced our understanding of the genomic structure-function relationship, and has enabled more detailed and fine-grained analysis of topologically associated domains.

a. Duggal, G., **Patro, R.**, Sefer, E., Wang, H., Filippova, D., Khuller, S., & Kingsford, C. (2013). Resolving spatial inconsistencies in chromosome conformation measurements. *Algorithms for Molecular Biology*, *8*(1), 8. PMCID:PMC3655033.
b. Malik, L., & **Patro, R.** (2018). Rich chromatin structure prediction from Hi-C data. *IEEE/ACM transactions on computational biology and bioinformatics*. PMID:29994683
c. Filippova, D.*, **Patro, R.***, Duggal, G., & Kingsford, C. (2014). Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, *9*(1), 14. PMCID:PMC4019371.
* denotes equal contribution

d. The armatus software for the identification of topologically associated domains at multiple scales of resolution: https://github.com/kingsfordgroup/armatus

D. Additional Information: Research Support and/or Scholastic Performance

NSF BIO-1564917 Patro (PI) 7/01/2016 6/30/19
National Science Foundation
Bilateral BBSRC-NSF/BIO: ABI Innovation: Data-driven hierarchical analysis of de novo transcriptomes
This goal of this research is to develop novel methods and associated tools for more accurately analyzing *de novo* transcriptomes. It covers work related to optimizing and evaluating assembly quality, transcript clustering, improved *de novo* expression estimation.
Role: PI

NSF CCF-1750472 Patro (PI) 2/01/2018 1/31/23
National Science Foundation
CAREER: A Comprehensive and Lightweight Framework for Transcriptome Analysis
This goal of this research is to develop methods and tools that close the capability gap between fast, lightweight transcriptome analysis methods and traditional analysis pipelines, but adding capabilities such as splice-aware genome mapping, novel transcript identification and assembly, and single-cell analysis capabilities to current lightweight frameworks.
Role: PI

NSF CNS-1763680: Patro(PI) 8/15/2018 7/31/2022
National Science Foundation
CSR: Medium: Approximate Membership Query Data Structures in Computational Biology and Storage
The goal of this research is to develop algorithms and data structures — especially approximate data structures — for indexing and processing large-scale data. The focus of this particular grant is on applications in computational biology, including the indexing of sequencing experiments and the development of lightweight data structures for assembly, as well as on applications of AMQs in filesystems and storage.
Role: PI

SVCF-182752 Patro(PI) 3/01/2018 02/28/2019
Silicon Valley Community Foundation
Efficient tools for quantifying and simulating transcript-level abundance in single-cell RNA-seq
The goal of this research is to develop efficient methods for (1) simulating realistic transcript-level data from various popular scRNA-seq protocols as well as to (2) investigate the effectiveness of different methods for quantifying transcript abundance uncertainty in scRNA-seq data (particularly in the context of tagged-end sequencing protocols).
Role: PI

NIH R01HG009937 Patro(PI) 09/01/2018 08/31/2023
National Institutes of Health
A Modular Framework for Accurate, Efficient, and Reproducible Analysis of RNA-seq Data
The goal of this research is to develop tools that improve the reproducibility of gene and transcript-level analysis from RNA-seq data by automatically generating and propagating metadata about e.g. the underlying transcriptome though the analysis procedure, as well as to develop improved methods for differential expression analysis taking into account quantification uncertainty and improved methods for allele-specific transcript abundance estimation.

Completed Research Support

NONE