SOFTWARE TOOL ARTICLE

A cloud-based learning environment for comparing RNA-seq aligners [version 1; peer review: 2 approved with reservations]

Elizabeth Baskin¹, Peter DeFord², Allison F. Dennis³, Ian Misner⁴, Frederick J. Tan⁵, Ben Busby^{id}⁶

¹Translational Genetics and Genomics Unit, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, USA

²Department of Biological Sciences, Johns Hopkins University, Baltimore, USA

³Program in Genomics of Differentiation, National Institute of Child and Human Development, National Institutes of Health, Bethesda, USA

⁴Bioinformatics and Computational Biosciences Branch, Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, USA

⁵Department of Embryology, Carnegie Institution of Washington, Baltimore, USA

⁶National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, USA

v1 First published: 13 May 2016, 5:888
<https://doi.org/10.12688/f1000research.8684.1>

Latest published: 13 May 2016, 5:888
<https://doi.org/10.12688/f1000research.8684.1>

Abstract

The rapid rise of high-throughput, data intensive experimental techniques has thrust many biologists into the role of data analyst – a role many biologists feel ill equipped to fill. Novices often struggle to find the resources and expertise they need to analyze their experimental results in a wet-lab environment. To fill this need, we developed an educational resource as part of a National Center for Biotechnology Information (NCBI) hackathon. Using RNA-seq as a model, our tutorial guides new users through the steps of data analysis, while placing an emphasis on understanding the motivation behind choices made in the process. To advance the goal of providing a deeper understanding of the analysis process, we developed a new tool, bamDiff. bamDiff allows users to compare the performance of multiple RNA-seq aligners, allowing users to select the most appropriate aligner for the data in question and experimental end-goal. Our tutorial is accessible via a GitHub wiki, with associated data and software provided on an Amazon Machine Image (AMI), which can be completed at no cost to the user through the Amazon Educate Program. Following the hackathon, our tutorial was integrated into the October 2015 offering of NCBI NOW (Next Generation Sequencing (NGS) Online Workshop) a free online experience targeting individuals new to NGS analysis.

Keywords

RNA-seq , SAM/BAM alignments , education , cloud , hackathon , pipeline , workflow , alignment

Open Peer Review

Approval Status ? ?

	1	2
version 1	?	?
13 May 2016	view	view

1. **Malachi Griffith** ^{id}, Washington University in St. Louis, St. Louis, USA
Washington University in St. Louis, St. Louis, USA
Washington University in St. Louis, St. Louis, USA
2. **Timothy I. Shaw**, St Jude Children's Research Hospital, Memphis, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **Hackathons** collection.

Corresponding author: Ben Busby (ben.busby@nih.gov)

Competing interests: No competing interests were disclosed.

Grant information: The work on this project by Ben Busby was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and NCBI. Elizabeth Baskin was supported by the Intramural Research Program of the National Institute of Arthritis and Musculoskeletal and Skin Diseases (Z01-AR041198). Peter DeFord was supported by NIH Training Grant GM007231.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2016 Baskin E *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

How to cite this article: Baskin E, DeFord P, Dennis AF *et al.* **A cloud-based learning environment for comparing RNA-seq aligners [version 1; peer review: 2 approved with reservations]** F1000Research 2016, **5**:888 <https://doi.org/10.12688/f1000research.8684.1>

First published: 13 May 2016, **5**:888 <https://doi.org/10.12688/f1000research.8684.1>

Introduction

With the rise of RNA-seq for exploring biological hypotheses has come an increase in the number of algorithms for aligning RNA-sequences to the genome. The burden of selecting and properly using these algorithms often falls on biologists. However, many biologists do not have training or experience in the skills needed to select the proper tools. To provide an introduction for this audience, the Educational Environment Team sought to develop an interactive learning environment where students with a novice's background in Unix could follow a series of alignment pipelines step-by-step.

The team identified two major goals that could be undertaken within the scope of the hackathon. The first was to produce a user-friendly tutorial that would walk novice computational biologists through the process of aligning RNA-seq data to the genome. The second was to supplement the tutorial with novel methods to compare the results of different RNA-alignment mappers.

Implementing this tutorial on an Amazon Machine Image (AMI) -- a virtual server in the cloud, pre-programmed with all the necessary packages -- allows students to initially bypass the intimidating task of installing software and dependencies, and immediately start performing alignments using a panel of different algorithms. All tasks in the tutorial and the AMI are designed to fall within the \$35 per student provided for free through the Amazon Educate Program (<http://aws.amazon.com/education/awseducate>). By allowing students to run these alignments in succession, we hope to naturally showcase how these aligners vary in their outputs. Here we introduce both a read-based and position-based approach for identifying and evaluating regions of differential performance across the genome.

Methods

Team composition

The team was composed of five members ranging in experience from graduate student to research faculty member. Each member of the team shared the experience of interfacing with non-computational biologists during their daily work and sought to provide a welcoming introduction to mapping tools. All members contributed to the goals of the team, with each selecting tasks to tackle according to their strengths.

Materials and methods

As part of the tutorial, two team members dedicated their time to developing novel methods for comparing the results of RNA-alignment. After each of the town-hall style meetings conducted with the larger hackathon group, we received recommendations from members of other teams for software packages to help streamline the proposed analysis.

Results

The tutorial was constructed as a GitHub Pages wiki -- each new wiki page represents a new step in the workflow¹. We guide students through registration with the Amazon Educate Program, obtainment of data, alignment, and comparison of the performance of four different aligners. The comparison of aligners is achieved

with our own custom written program, bamDiff, in concert with the R (v3.2.1) packages edgeR (v3.10) and csaw (v1.2.1).

The example dataset is sample NA12878 from the Genome in a Bottle Consortium, an extensively curated human-standard². To give a flavor for the alignment workflow without burdening the learner with expensive computational requirements, we limited the genome for alignment to chromosome 20 of the latest human genome reference from RefSeq, GRCh38 (GCF_000001405.30). For the tutorial, we describe the use of four popular aligners: BWA v0.7.12-r1039³, HISAT v0.1.6-beta⁴, STAR v2.4.0j⁵, and BLASTmapper (in preparation). For each, we provide the commands needed to construct an index and align the given data, outlining the expected screen output, files to be created, and time required for each task.

The R package csaw provides an elegant framework for identifying regions of differential expression between RNA-seq experiments, and in this case was extended to identifying regions of differential alignment between mappers⁶. To implement read counting, we used csaw to bin reads into 1KB windows, then filtered windows by count size⁷. Only bins with greater than 10 reads were kept. Windows were then filtered by exons, and only windows/bins containing exons were kept for further analysis (see Figure 1c). edgeR was used to identify windows/regions/bins with significant differences between the four mappers (see Figure 1d)⁸.

In order to systematically compare the performance of the various aligners, we wrote a new program called bamDiff. This Python script takes the output from csaw and edgeR as a CSV file as well as the outputs from each aligner in BAM format. bamDiff will report simple summaries for each BAM file, such as the total number of reads, overall number of alignments, proportion of reported reads that were unmapped by the aligner, as well as proportion of aligned reads that were mapped only once (uniquely).

The real strength of bamDiff comes in its ability to go beyond simple summarization to direct comparisons between BAM files. Internally bamDiff uses SAMtools⁹ to rapidly extract only the reads mapping to a region of interest identified by csaw from one BAM file. These reads are then checked against the other BAM files to see whether they are mapped at all, and if so, whether they are mapped to the same region in the genome as in the first BAM file (see Figure 1e). If they map to a conflicting region outside the region of interest, bamDiff will report the top ten regions reads are mapping to, by aggregating reads mapping within 1kb of each other. These results are reported as text tables. Example usage and output can be viewed at the associated page in our tutorial: https://github.com/NCBI-Hackathons/RNA_mapping/wiki/7.-Compare-alignments-with-bamDiff.

Following the hackathon, our tutorial was integrated into NCBI NOW (National Center for Biotechnology Information Next Generation Sequencing (NGS) Online Workshop) a free online experience targeting individuals new to NGS analysis. The first offering of the workshop occurred October 13–23, 2015. Our tutorial was offered as a “take-home” exercise following the sixth lecture, which focused on the analysis of RNA-seq data.

Figure 1. Diagrammatic representation of workflow for csaw, edgeR, and bamDiff. (a) Raw RNA sequencing reads undergo quality control checks and filtering before (b) being aligned using several different aligners. (c) csaw bins the reads and filters for bins based on read counts and overlap with exons. (d) edgeR identifies regions of where the aligners show differential mapping in the csaw filtered regions. (e) from the edgeR identified regions, bamDiff checks where those reads are being mapped in the other BAM files.

Conclusions and next steps

We achieved two distinct goals during the hackathon: first, the development of a novel method for comparing the results of RNA-alignment and second, the creation of a tutorial to guide users through not only undertaking, but also understanding, RNA-seq alignment. Notably, we do not over-simplify the work of data analysis by simply selecting a single aligner for use in our tutorial. Instead, by encouraging the comparison of many different aligners we draw attention to the decision-making process intrinsic to any

bioinformatic work. For our user, this is the selection of an RNA-seq aligner from many frustratingly similar options.

However, both the novel software bamDiff and the tutorial can be further improved.

Currently, bamDiff is best suited to analyzing a small region of the genome/BAM file, rather than undertaking the entire genome/BAM file. Additionally, a graphical summary would make outputs easier to understand.

In the tutorial we assume prior knowledge of the contents, purpose, and format of FASTQ, SAM, and BAM files as we assumed that our tutorial would be used in conjunction with a class or workshop that covers these introductory topics, such as NCBI NOW. The tutorial could be expanded in a number of ways, such as generalizing to pair-end reads or alignments guided by an annotation. Finally, there may be room for improvements in usability, such as streamlining use on platforms such as Google Cloud, Microsoft Azure, or iPlant/CyVerse.

Software availability

Latest source code: https://github.com/NCBI-Hackathons/RNA_mapping.

Source code as at the time of publication: <http://dx.doi.org/10.5281/zenodo.50871>¹⁰

License: **CC0 1.0 Universal**

Author contributions

All of the authors participated in designing the study, carrying out the research, and preparing the manuscript. All authors were

involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

The work on this project by Ben Busby was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and NCBI. Elizabeth Baskin was supported by the Intramural Research Program of the National Institute of Arthritis and Musculoskeletal and Skin Diseases (Z01-AR041198). Peter DeFord was supported by NIH Training Grant GM007231.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors thank Lisa Federer, NIH Library Writing Center, for manuscript editing assistance.

References

1. RNA Mapping Team: **RNA Mapping Github** [Internet]. RNA Mapping Github. [cited 2016 Feb 16]. [Reference Source](#)
2. National Center for Biotechnology Information: **Bioproject** [Internet]. Homo Sapiens ID 236780. 2014. [cited 2016 Feb 16]. [Reference Source](#)
3. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics*. 2009; **25**(14): 1754–60. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nat Methods*. 2015; **12**(4): 357–60. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics*. 2013; **29**(1): 15–21. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Lun AT, Smyth GK: **csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows**. *Nucleic Acids Res*. 2016; **44**(5): e45. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Lun A, Smyth GK: **csaw user manual** [Internet]. 2016. [cited 2016 May 2]. [Reference Source](#)
8. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics*. 2010; **26**(1): 139–40. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics*. 2009; **25**(16): 2078–9. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. DeFord P, Tan F, Misner I, *et al.*: **RNA_mapping: RNA_mapping**. *Zenodo*. 2016. [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 26 September 2016

<https://doi.org/10.5256/f1000research.9346.r16413>

© 2016 Shaw T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Timothy I. Shaw

Department of Computational Biology, St Jude Children's Research Hospital, Memphis, TN, USA

The manuscript summarizes two achievements from NCBI's hackathon of 2015. First is a tutorial to introduce RNAseq mapping. Second is bamDiff, a program for comparing different RNA-seq aligner mapping. This reviewer believes the current state of the tutorial is on the thin side and could benefit from additional expansion. The functionality of bamDiff program is intriguing but based on the current state of the program the reviewer feels the program should be expanded to incorporate other QC metrics. Overall, this work makes great stride for guiding biologist to their first hands-on-experience on NGS.

Major points:

1. The tutorial provides a step by step tutorial from downloading to mapping and some mapping evaluation. The tutorial can be useful to users that find difficulty working on a Unix environment. In the current form, the author introduces basic commands for performing mapping; however, the author should caution and educate the reader that additional vetting of the raw RNAseq mapping is necessary. While mapping evaluation is important, that is just one of the many QC metric necessary in RNAseq data that contribute to the decision making. Here is an incomplete list of RNAseq related issues that should be included in the tutorial:
 1. Whether the RNAseq sample require trimming of low quality or adaptor sequences.
 2. The different parameter tuning that increases the coverage of these hard to map region (i.e. STAR 2-pass).
 3. At what sequencing depth is the RNAseq sample is deep enough for expression analysis, gene fusion detection, splicing detection, and whether additional sequencing is necessary.

(The reviewer recognize that the above comments might not be suitable within the tutorial, but the author should make some attempt to inform the user of these caveats.)

2. Regarding the bam comparison program, the author might want to automatically include output of coverage Bed files that can be displayed in UCSC genome browser or IGV. In the tutorial, the author included some examples of coverage differences; however, the discussions from the tutorial appear to be incomplete. The author failed to discuss the reason that contributes to these differences in coverage such as the parameters or algorithm design. Another discussion point should be to examine where the same read is being mapped to in different program? Here I present a couple factors that might impact the mapping the author might want to consider:
 1. GC content
 2. Highly repetitive region
 3. Paralogous gene bodies such as some ribosome genes and mitochondrial genes or histone genes might have varying coverage in different programs.
3. Within the tutorial, the author mentions “is every mapper allowing reads to map to the intronic region? After all, this is an RNA-seq experiment -- there should be minimal intronic genetic material.” While the statement is largely true; however, this reviewer believe the author should also mention that there are different type of RNA-seq library specifically TruSeq Stranded Total RNA prep would include contain intronic reads. Poly-A enriched RNAseq experiment could also contain intronic reads for intron retention events.
4. Since the author appear to discuss splicing region in the tutorial, a more detailed analysis on how different programs deals with the splicing region could be of tremendous interest to certain readers. The author might also want to consider the impact of RNA-seq library protocols on the variation of splice site mapping.
5. A number of tools have been developed for assessing RNAseq alignment quality. A review tools compared to bamDiff might make a stronger case for the novelty of the program.
6. A comprehensive table in the tutorial or within the manuscript summarizing all the advantages and disadvantages of different mapping programs will definitely enhance the manuscript.

Minor points:

1. Perhaps include other popular aligner such as Tophat as an option.
2. While it is great that the pipeline being on the amazon cloud allows users to bypass installation. Perhaps the author could include installation procedure or reference to individual software if the user wants to install the program on their private server.
3. Make sure all program provide sufficient commenting. Most of the shell script lack sufficient commenting for what the program will do and why the author chose those parameters.
4. Sentence that might benefit from rephrasing “With the rise of RNA-seq for exploring biological hypotheses has come an increase in the number of algorithms for aligning RNA-sequences to the genome.” Particularly the choice of “has come” is a bit awkward of a wording.

5. Another sentence that might benefit from rephrasing “If they map to a conflicting region outside the region of interest, bamDiff will report the top ten regions reads are mapping to, by agglomerating reads mapping within 1kb of each other.”
6. For certain details on RNAseq analysis, the author want to refer the reader to the following paper for additional details on RNAseq analysis.
“<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8>”
7. Regarding the accompanying software bamDiff. The summary statistics merely output summary statistics obtained from other mapping programs. The author might want to consider refining the output to a more presentable format like outputting to an excel file.

References

1. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, et al.: A survey of best practices for RNA-seq data analysis. *Genome Biology*. 2016; **17** (1). [Publisher Full Text](#)

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 05 September 2016

<https://doi.org/10.5256/f1000research.9346.r15576>

© 2016 Griffith M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Malachi Griffith

¹ McDonnell Genome Institute, School of Medicine, Washington University in St. Louis, St. Louis, MO, USA

² Department of Genetics, School of Medicine, Washington University in St. Louis, St. Louis, MO, USA

³ Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO, USA

The authors describe the results of an educational effort in which a hackathon event was used to develop an interactive tutorial to help biologists learn a fundamental NGS analysis skill. Specifically, that of selecting an appropriate read aligner, performing alignments, and evaluating the outcome. Overall the tutorial is organized, and the accompanying paper is well written.

Major points:

- The primary goal of this work is commendable. However having reviewed the paper and

tutorial, I was surprised by the lack of discussion/ interpretation of the results. Choosing an appropriate RNA-seq read aligner and evaluating the outcome can indeed be a challenge to those new to the field. The tutorial walks a user through the process of conducting alignments with four possible aligners. Some tools that evaluate the resulting aligners are presented and used during the tutorial. However, the authors offer little interpretation of the results, even for the demonstration data set. What do the results tell us about the quality of each alignment result? What factors might be considered in deciding which is "best"? What are the pitfalls for such assessments? How might the results be visualized to assist interpretation?

- A secondary goal of creating a tool "bamDiff" to assist comparisons between RNA-seq aligners is less well developed. This works seems to be fairly preliminary at this stage, consisting of a single Python script that produces a text summary of a few metrics extracted from RNA-seq BAMs from multiple aligners. Similar to the previous point, additional development would be needed before the results of this tool would be readily useful to most prospective users.
- Considerable resources/ tools for performing quality assessment of BAM files (including RNA-seq alignments) already exist. The authors could provide an overview of these, either in the paper or as an additional section in the tutorial Wiki.
- In the tutorial. The section for each aligner considered (BWA, HISAT, STAR, and blastmapper) should provide a basic description of the aligner, references, link to the aligner documentation, etc.
- The alignment comparisons focus on the number of reads aligned, and how aligners differ in the alignment of particular reads, or reads aligning to particular regions. What other ways might the aligners be different? For example, in their ability to correctly map RNA-seq reads across exon-exon junctions, align reads containing single base sequencing errors or polymorphisms, correctly handle reads containing small insertions or deletions relative to the reference genome, etc.

Minor points:

- Perhaps the abstract should include a URL for the tutorial mentioned in the title.
- The authors have created an AMI to "allow students to initially bypass the intimidating task of installing software and dependencies". This is reasonable, but perhaps the installation task could be provided (with detailed instructions) as an optional exercise.
- On a related note, it would be ideal to have detailed documentation on how the AMI (ami-3590de50) was configured (including all dependencies that were installed).
- In addition, this tutorial could include a "resources/pre-reading" section that referred the reader to additional helpful materials on RNA-seq sequencing and analysis principles (in addition to the hands on pre-requisites already listed in section 1).
- More details on the example RNA-seq data set used in the hands on exercises would be helpful.

- Are there similar efforts for comparison of DNA aligners that could be referenced by this tutorial?
- Other RNA-seq educational pieces that cover many topics relevant to new NGS users (with less focus on aligner comparison specifically) could be cited by this paper (e.g. Griffith M *et al.*, www.rnaseq.wiki).

References

1. Griffith M, Walker JR, Spies NC, Ainscough BJ, et al.: Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLoS Comput Biol.* 2015; **11** (8): e1004393 [PubMed Abstract](#) | [Publisher Full Text](#)

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research