PeerJ Preprints NOT PEER-REVIEWED

Best practices for conducting benchmarking in the most comprehensive and reproducible way

Serghei Mangul[1$], Lana S. Martin*[1$], Margaret Distler[3], Eleazar Eskin[1,2], Jonathan Flint[3]

[1]Department of Computer Science, University of California Los Angeles, 580 Portola Plaza, Los Angeles, CA 90095, USA

[2]Department of Human Genetics, University of California Los Angeles, 695 Charles E. Young Drive South, Los Angeles, CA 90095, USA

[3]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California-Los Angeles, CA 90095, USA

$ - These authors contributed equally to the paper

* - Corresponding author: lana.martin@ucla.edu

Abstract

Computational biology is rapidly advancing thanks to the many new tools developed and published each month. A systematic benchmarking practice would help biomedical researchers leverage this technological expansion to optimize their projects. Several aspects of algorithm publication and distribution contribute to this challenge. We address these challenges and present seven principles to guide researchers in designing a benchmarking study. Our proposed steps show how benchmarking can create a framework for comparison of newly published algorithms.

Keywords

PeerJ Preprints NOT PEER-REVIEWED

Benchmarking, computational genomics, reproducibility

During the past decade, improvements in genomics and sequencing technologies have led to the rapid development of many new algorithms. A major challenge for today's biomedical researcher is deciding which of these new algorithms will perform best in a particular study. Several aspects of algorithm publication and distribution contribute to this challenge. The bioinformatics software devleopment community currently lacks a systematic procedure for assessing the performance of new algorithms. When novel algorithms are first published, authors independently demonstrate the superiority of their newly developed tool by comparing it to existing algorithms. Potential users lack adequate resources to help them choose the tools that best suit their data, and they must weigh the advantages of adopting a new tool with its potential gains against discarding an exisiting tool with proven capabaility. Unsystematic assessment of new algorithms creates a gap between tool developres and the biomedical researchers, who are the end users of the developed tool.

Benchmarking studies, when systematically performed and reported, allow researchers to choose the most appropriate tool for a project. In a benchmarking study, investigators perform a robust and comprehensive evaluation of existing algorithms' abilities to solve a particular computational biology problem. Benchmarking studies use experimental gold-standard datasets and well-defined scoring metrics to assess the performance and accuracy of each tool when applied to a variety of anaytical tasks and data types. Standardized comparisons of algorithm performance metrics allow a biomedical researcher who is otherwise unfamliar with software development to chose the ideal benchmarked tool for a study.

Currently comprehensive benchmarking studies are not available in many fields of genomics, and researchers have traditionally used empirical rationale for selection of computational tools. This results in little or no overlap with respect to the computational tools utilized by the community.

Problem: Unsystematic Comparison of Tools

At present, assessment of newly published algorithms is unsystematic, and the data used for this comparison is typically generated in-house by a research group. Many computational labs lack adequate resources to generate or access gold-standard experimental data. Novel tools are often compared to several existing tools using simulated data, which is inexpensive to generate and efficiently scales up to represent large numbers of samples. However, comparing tool performance based on simulated data carries several limitations. First, the models under which the simulated data are generated can differentially bias the outcomes of algorithms. For example, the algorithm itself could be trained on simulated data prior to running. Second, simulated data cannot capture true experimental variability and will always be less complex than real data. Finally, not all simulated data are validated with real-world data, and not all methods used to simulate data have been validated. An alternative to the simulated datasets is publically available real datasets composed of a large number of samples. Some of these datasets have been extensively validated and can be used by a large number of groups, helping ensure the reproducibility of results obtained in benchmarking. However, publically available data is often poorly annotated and lacks detailed description of the exact protocols used to generate the data and summary statistics.

Solution: Systematic Benchmarking Practice

A systematic bechmarking practice in bioinformatics would inform the biomedical research community on the strengths and weaknesses of tested tools. Ideally, these repositories suggest the most accurate algorithms available and identify the best specific applications for each benchmarked tool. For example, in our own work, we have found that the best tool for detecting structural variants in low-coverage data may be different than the best tool for detecting structural variants in high-coverage data (unpublished results). In addition, the data produced by each of the benchmarked tools are valuable and serve as a gold standard to be re-used by the developer community. Gold-standard datasets allow comparison of the performance of newly developed algorithms to the performance of previously developed and benchmarked tools. Further, the availability of the results of each benchmarked tool makes this comparison a straightforward task; the developer need only download the results of each tool as raw data or as summary statistics and provide benchmarking data as input for the previously assessed tools. Scripts provided by each benchmarking study would allow researchers to assess the accuracy of new algorithms in comparison to benchmarked ones.

Challenges to Establishing an Effective Benchmarking Platform

There is presently a lack of standardized practices for benchmarking studies in computational genomics, and several aspects of software development in bioinformatics pose challenges to

establishing an effective benchmarking platform. First, a large number of software tools are available, and an increasing number of applications are released each month. For example, over 70 computational tools have been developed to identify structural variants from next-generation sequencing data.

Second, independently conducted benchmarking studies lack a centralized hub for hosting available benchmarking data and scripts to download data. Benchmarking studies and prepared data that are *not* broadly accessible are inherently limited in usability by the biomedical community. Even the most comprehensive studies that follow optimal benchmarking practices do not typically deposit the raw output data of each studied algorithm[1]. Rather, these studies have relied on the "upon request" model, which is a less reliabe and reproducible method of data sharing as it relies on individual authors to share the data. This limits the ability of the community to independently verify and vet the benchmarking data. Inconsistent benchmarking efforts that are not widely disseminated to the academic community cannot be easily expanded upon.

Third, there is little consensus about what consitutes a gold-standard experimental dataset for each particular application. For example, what is the minimum number of samples that should be included in the benchmarking study? What is an adequate level of coverage and/or fidelity of data? Should there be molecular confirmation of the data? These question are presently open and require resolving in order to build a benchmarking practice that can serve as a hub to

store the raw input data, results of the benchmark algorithms (in form of raw output data or summary statistics), and scripts to access accuracy.

Principles for Best Benchmarking Practices

Here we present seven principles to guide researchers in designing a benchmarking study. These guidelines emphasize reproducibility and enabling more effectively utilization of resources and data. Ideally, a benchmarking study will assess all available tools.

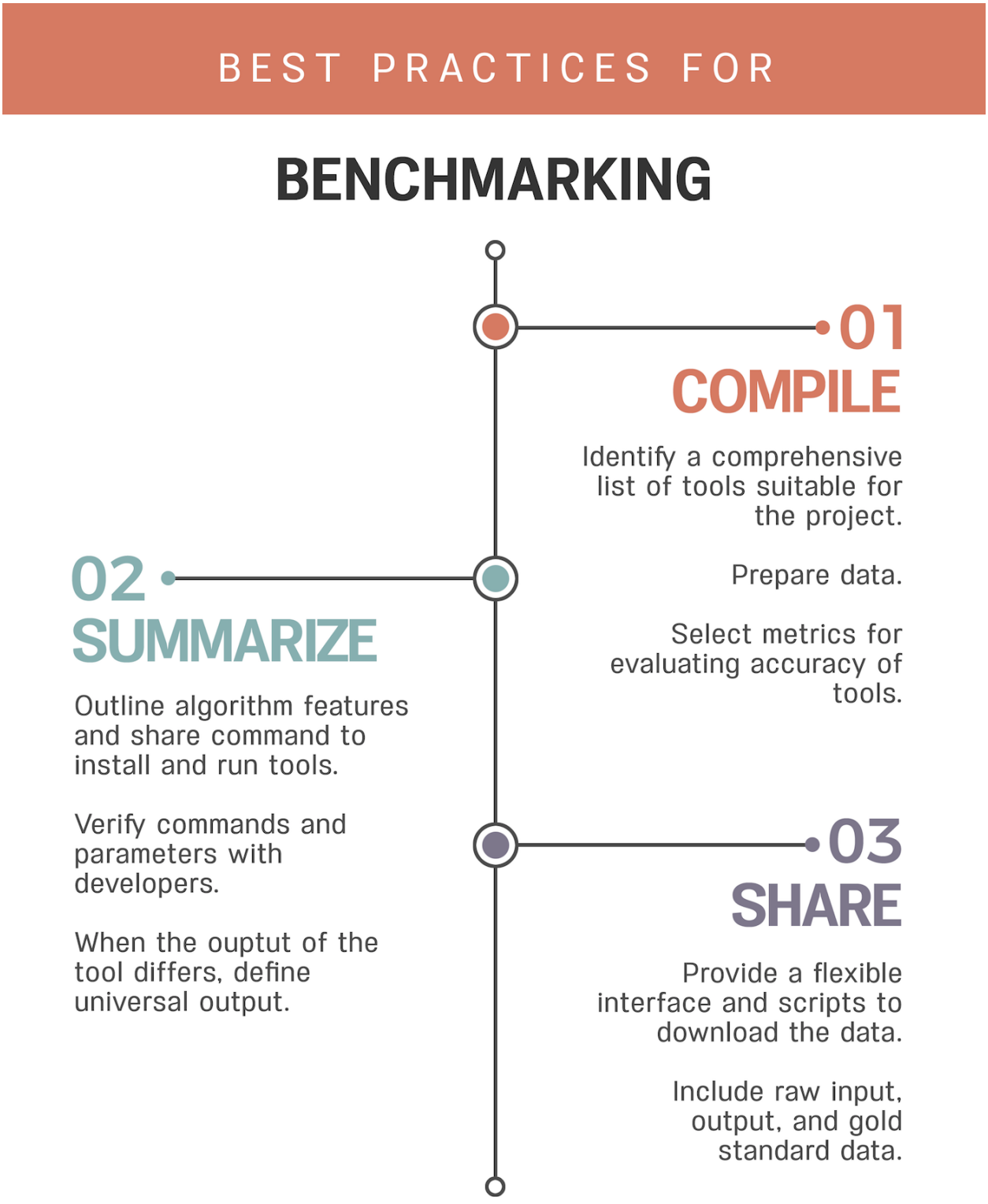

Figure 1. Best practices for benhmarking.

We propose step-by-step instructions to maximize reusability of data generated by a benchmarking study (Figure 1):

1. **Compile a comprehensive list of tools to be benchmarked.** Identify the list of software tools that are suitable for the purposes of the project. One approach is to perform a PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) search for relevant articles. Once relevant publications are identified, review references of each identified publication to complete the compiled list. We advise compiling the most comprehensive list of tools. Selecting the most popular algorithms for the benchmarking study based on the number of citations and/or reputation of the journal is risky, as popularity of the tool does not necessarily imply accuracy for a particular application[1].
2. **Prepare and describe benchmarking data.** Compose a table summarizing the benchmarking data. Explain protocols used for preparing data, obtaining gold standard data, and any potential limitations of the data.
3. **Select evaluation metrics.** Metrics for evaluating the accuracy of the software tools need to be carefully determined and packed in the form of scripts. Such scripts can be used by the community to evaluate the performance of any newly developed algorithms.
4. **Summarize algorithm features and share command to install and run tools.** Prepare a table outlining the benchmarked algorithm's features, the underlying algorithm, the dependencies of the software, and the journal the algorithm was published in. We provide an example in Table S1. Consider verifying table contents with the relevant tool

developers. Finally, providing additional, detailed instructions on how to install and run the benchmarked tools is a valuable resource for the community.

5. **Verify commands and parameters with developers.** Prior to running the software, contact the software developers to ensure the correctness of chosen commands and parameters.
6. **Define universal format (if neccesary).** When the ouptut of each tool is different, develop and share a script with the community to generate a universal format.
7. **Provide a flexible interface to download the data.** In order to facilitate ease of use in the scientific community, we suggest sharing all the data and the commands used in the benchmarking study. Share the interface to download the input raw data, and gold-standard data. In addition to the input and gold standard data, it is recommended to share the raw output data of each benchmarked tool. This additional information will allow the end user to apply their own evaluation metrics to the shared raw output data of the bechmarking study. A script to download the data with a flexible interface will maximize the reusability of the data. Such scripts can be also be used to reproduce the results and figures of the beahmarking study. We provide an example of such scripts in Supplementary Note 1.

Computational biology is rapidly advancing thanks to the many new tools developed and published each month. A systematic benchmarking practice would help biomedical researchers leverage this technological expansion to optimize their projects. Our proposed steps show how benchmarking can create a framework for comparison of newly published algorithms. Under

this framework, users can can compare novel algorithms to all the tools benchmarked in a study simply by downloading the input data and running novel tools on the raw input data. Re-using the results of existing tool assessments and comparisons can save time and increase the number of tools each study can potentially be compared with. Our proposed principles will make computational genomics benchmarking studies more sustainable and reproducible, ultimately increasing the transparency of bioinformatics data and results.

References

1. Baruzzo, G. *et al.* Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods* (2016).