Assessment of RSubread Using Mock Community Microbiome Data

Nebiyou Kebede[1, 2], Howard Fan[3], Tyler Faits[3], W. Evan Johnson, Ph.D.[4]

[1]University of Massachusetts Amherst, [2]Boston University Bioinformatics BRITE REU Program, Summer 2019, [3]Boston University Graduate Program in Bioinformatics, [4]Boston University School of Medicine Department of Computational Biomedicine

Abstract

Read alignment in the analysis of sequencing data is an ongoing challenge in the processing of microbiome data. Available alignment tools require preliminary computational steps such as installation and integration of dependent packages and familiarity with command-line based interface, which has hindered researchers from performing efficient microbiome analyses. To address this, our lab is working on integrating Subread, the only R-based alignment tool in R (Rsubread), into a user-friendly, intuitive R-based pipeline to analyze microbiome data. However, Rsubread's performance on microbiome data has not been analyzed. Thus, this study was focused on evaluating Rsubread's performance on microbiome data as compared to other established non-R aligners using 16S rRNA sequencing data derived from a mock community composed of 20 bacterial species in equal abundance. To benchmark its performance, we aligned the samples to a reference bacterial database using Rsubread's default parameters. We also optimized the parameters by specifying a range of values for ones that directly impact sensitivity and selection accuracy to attain maximum accuracy. Contrary to our expectations, optimization of the parameters only had minimal impact in this context of 16S rRNA sequence alignment. However, Rsubread detected an average of 70% of expected species using its default parameters, outperforming other aligners (QIIME1, QIIME2, PathoWhole and PathoGreen) that were evaluated on the same mock community and thus providing a solid ground for its integration into the microbiome analysis pipeline.

Background and Motivation

Previous studies have shown that Rsubread provides competitive efficiency and accuracy as compared to other available alignment methods in contexts such as RNA-sequencing. To integrate Rsubread into our lab's microbiome analysis pipeline, though, evaluation of its performance on microbiome data was crucial. The goal of this study was thus to evaluate Rsubread's performance using a mock community dataset and compare its performance with established non-R aligners.

Methods

To benchmark Rsubread's alignment performance, we used 16S ribosomal RNA (rRNA) sequences of a mock community dataset composed of genomic DNA extracted from a mixture of 20 bacterial strains in equal abundance. Using a reference bacteria genome file obtained from NCBI, we first built a reference index using Rsubread's buildindex function to perform read alignment.

The samples were then aligned to the reference index using Rsubread's default parameters. To maximize its accuracy, parameters that directly impact sensitivity and selection accuracy were optimized.

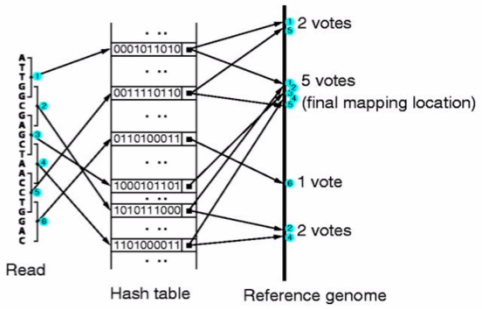

Figure 1: Illustration of RSubread's mapping paradigm using an artificial example

Optimized Parameters:

- **nBestLocations**: maximal number of equally-best mapping locations that will be reported for a multi-mapping read
- **TH**: consensus threshold for reporting a hit
- **maxMismatches**: maximum number of mismatched bases allowed in the alignment
- **nsubreads**: number of subreads extracted from each read

To measure its accuracy, relative abundance count matrices with expected species as rows and samples as columns were extracted from the alignment outputs. For each species, the minima between the measured proportion and the true proportion were taken and summed to give a score between 0 and 100.

Results

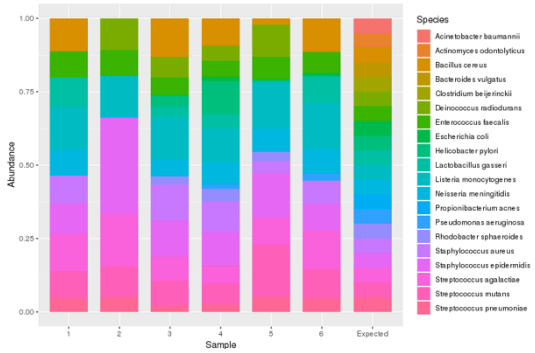

Figure 2: Relative abundance of detected species among all six samples along with expected relative abundance of 20 source species.

- RSubread detected an average 70% of expected species among all six samples
- RSubread measured an accuracy of 62.5%, outperforming Qiime1, Qiime2, PathoGreen and PathoWhole, all of which are read aligners that have been evaluated on the same mock community.

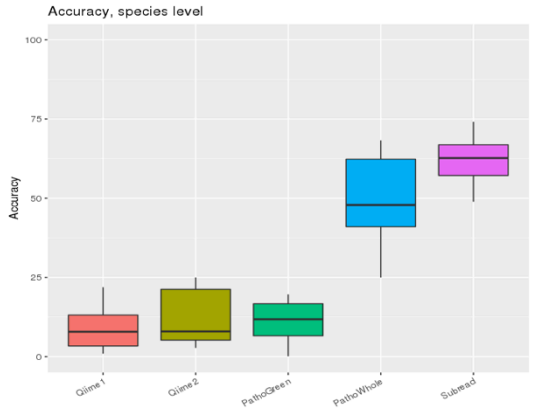

Figure 3: Accuracy measure results for aligners on 16S microbiome mock community

Conclusion and Future Work

Rsubread's alignment performance on 16S rRNA microbiome data provides a solid ground for its integration into our intuitive, user-friendly R-based microbiome analysis pipeline that is in progress. However, understanding the importance of read alignment in the accuracy of downstream analysis, our lab seeks to expand this investigation to other sources of microbiome data, including shotgun sequencing and RNA-seq, where we expect that optimization of alignment parameters will be more influential in maximizing its accuracy even further.

References and Acknowledgments

- Liao Y, Smyth GK, Shi W (2019). "The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads." *Nucleic Acids Research*, **47**, e47. doi: 10.1093/nar/gkz114.
- The following reagent was obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project: Genomic DNA from Microbial Mock Community B (Even, Low Concentration), v5.1L, for 16S rRNA Gene Sequencing, HM-782D

This work was funded, in part, by NSF grant DBI-1559829, awarded to the Boston University Bioinformatics BRITE REU program