Bachelor Thesis

Port, improve, and benchmark read-based MGnify analysis pipeline for microbiome data using the Galaxy framework

Albert Ratschinski

Examiner: Prof. Dr. Rolf Backofen
Advisor: Dr. Paul Zierep

Albert-Ludwigs-Universität Freiburg
Faculty of Engineering
Department of Computer Science
Chair of Bioinformatics

13. Aug 2024

Writing Period

13. 05. 2024 – 13. 08. 2024

Examiner

Prof. Dr. Rolf Backofen

Advisor

Dr. Paul Zierep

Declaration

I hereby declare, that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work. I also hereby declare that my thesis has not been prepared for another examination or assignment, either in its entirety or excerpts thereof.

Place, date	Signature

Acknowledgements

First and foremost, I express my gratitude to Prof. Dr. Rolf Backofen, who agreed to review this work and gave me the opportunity to work on this interesting topic.
Special thanks are owed to Dr. Paul Zierep for the best supervision one could wish for. Thank you for always being open to my questions, for your invaluable advice, and for keeping me motivated throughout this journey. Working with you was enriching and valuable. I appreciate your dedicated efforts.
I would also like to express my gratitude to Dr. Paul Zierep, Dr. Björn Grüning, and Dr. Matthias Bernt for reviewing my tool wrapper pull requests on GitHub and for their improvement suggestions.
I extend my appreciation to Dr. Alexandre Almeida for addressing my questions regarding the samples from his study.
A special acknowledgment goes also to the MGnify team for being open for questions regarding the pipeline, and it tools, which contributed to this thesis.
Lastly, I would like to thank my family and friends for their support in various ways throughout the writing of this thesis.

Abstract

Metagenomic sequencing is a powerful method used to uncover and characterize microbial communities by analyzing genetic material from environmental samples. This thesis benchmarks the performance of Kraken2, Bracken, MetaPhlAn, and mOTUs using the CAMI challenge dataset as a gold standard within the Galaxy framework. The benchmarking process compared the taxonomic profiling outputs of these tools against the CAMI gold standard using Cami Opal for detailed analysis. Metrics such as taxonomic profiling accuracy, computational efficiency, and memory usage were assessed. Kraken2 and Bracken provided high accuracy, while MetaPhlAn and mOTUs effectively profiled microbial communities using marker genes. Integrating these tools into Galaxy enhances their utility for metagenomic research by providing a reproducible and customizable analysis environment. The study highlights the strengths and limitations of each tool, offering valuable insights for selecting the appropriate pipeline for specific research needs. The availability of these tools within Galaxy offers several advantages for the metagenomics research community, including interoperability, reproducibility, and ease of use, facilitating advanced downstream analyses such as machine learning and differential abundance analysis.

Deutsche Version:

Metagenomisches Sequenzieren ist eine leistungsstarke Methode zur Entdeckung und Charakterisierung mikrobieller Gemeinschaften durch die Analyse genetischen Mate-

rials aus Umweltproben. Diese Arbeit vergleicht die Leistung von Kraken2, Bracken, MetaPhlAn und mOTUs anhand des CAMI-Challenge-Datensatzes als Goldstandard im Galaxy-Framework. Der Benchmarking-Prozess verglich die taxonomischen Profilierungsausgaben dieser Werkzeuge mit dem CAMI-Goldstandard unter Verwendung von Cami Opal für eine detaillierte Analyse. Metriken wie die Genauigkeit der taxonomischen Profilierung, die Recheneffizienz und die Speichernutzung wurden bewertet. Kraken2 und Bracken boten eine hohe Genauigkeit, während MetaPhlAn und mOTUs effektiv mikrobiellen Gemeinschaften mit Markergenen profilierten. Die Integration dieser Werkzeuge in Galaxy erhöht deren Nutzen für die metagenomische Forschung durch eine reproduzierbare und anpassbare Analyseumgebung. Die Studie hebt die Stärken und Schwächen jedes Werkzeugs hervor und bietet wertvolle Einblicke in die Auswahl der geeigneten Pipeline für spezifische Forschungsanforderungen. Die Verfügbarkeit dieser Werkzeuge im Galaxy-Framework bietet der Metagenomik-Forschungsgemeinschaft mehrere Vorteile, darunter Interoperabilität, Reproduzierbarkeit und Benutzerfreundlichkeit, und erleichtert fortschrittliche nachgelagerte Analysen wie maschinelles Lernen und Differenzanalyse der Abundanz.

Contents

1 Introduction

Metagenomics is an expansive field focused on the analysis of genetic material recovered directly from environmental samples. This approach provides a comprehensive view of the microbial communities present in a given habitat. A critical component of metagenomic analysis involves taxonomic profiling, which identifies and categorizes the various microorganisms within a sample. This profiling is essential for understanding microbial diversity, ecosystem functioning, and the roles of specific microbes in health and disease.

Various tools and pipelines have been developed to facilitate taxonomic profiling, each leveraging different databases and algorithms. Among these tools, Kraken2 and Bracken have gained prominence for their high accuracy and efficiency in classifying reads from metagenomic samples. Kraken2 employs a k-mer based approach to classify reads, while Bracken refines these classifications to improve abundance estimates. Similarly, MetaPhlAn and mOTUs are widely used for profiling microbial communities using marker genes.

Despite the availability of these tools, there is no consensus on a single best approach for metagenomic analysis. Comparative studies are necessary to evaluate their performance in terms of time efficiency, memory usage, and accuracy. These studies help in identifying the strengths and limitations of each tool, providing valuable insights for selecting the appropriate pipeline for specific research needs.

1.1 CAMI Challenge

The Critical Assessment of Metagenome Interpretation (CAMI) challenge provides a rigorous benchmarking platform for metagenomic tools and pipelines. CAMI was initiated to address the need for standardized evaluation in the rapidly evolving field of metagenomics. It aims to provide a common ground for developers and users to assess the performance of different methods under controlled conditions.

By offering a gold standard dataset, CAMI allows researchers to evaluate the performance of different methods in a controlled setting. The CAMI datasets are meticulously curated and designed to reflect realistic microbial community compositions, making them highly relevant for benchmarking purposes. The challenge encompasses various tasks such as taxonomic profiling, genome assembly, and binning, providing a comprehensive evaluation framework.

The importance of the CAMI challenge in my thesis cannot be overstated. It provides a robust, unbiased benchmark against which the performance of Kraken2, Bracken, MetaPhlAn, and mOTUs can be measured. This benchmark is crucial for validating the accuracy and reliability of these tools when applied to real-world metagenomic data. The insights gained from this benchmarking study will not only help in identifying the most effective tool for taxonomic profiling but also guide future improvements in metagenomic analysis pipelines.

1.2 Galaxy project

Galaxy is a versatile, web-based platform designed for accessible and reproducible scientific computing. It supports a wide array of bioinformatics tools and workflows, enabling researchers to process large datasets efficiently. The platform's user-friendly interface and comprehensive workflow management capabilities make it an ideal environment for developing and sharing complex analytical pipelines.

One of the key strengths of Galaxy is its ability to integrate tools and datasets seamlessly. Users can create, import, and modify workflows with ease, facilitating a high degree of customization and flexibility. Galaxy also supports the reproducibility of scientific analyses by maintaining detailed records of the workflows and parameters used in each experiment.

Galaxy's community-driven development ensures that it stays current with the latest advancements in bioinformatics. The platform offers extensive documentation and a wealth of tutorials, covering a wide range of scientific and technical topics. This makes it accessible to researchers with varying levels of expertise and enhances its utility across diverse research applications.

1.3 Aim of this thesis

The primary aim of this thesis is to benchmark the performance of Kraken2, Bracken, MetaPhlAn, and mOTUs using the CAMI challenge dataset as a gold standard. The analysis will be conducted within the Galaxy framework to leverage its robust computational resources and workflow management capabilities. This study will compare the taxonomic profiling outputs of these tools against the gold standard provided by CAMI, using Cami Opal for detailed comparative analysis. By integrating these tools into Galaxy and conducting a thorough benchmarking study, this thesis seeks to provide insights into their relative performance and suitability for different metagenomic research scenarios. Specifically, the thesis will:

- Evaluate the accuracy of Kraken2, Bracken, MetaPhlAn, and mOTUs in taxonomic profiling against the CAMI gold standard dataset.
- Assess the computational efficiency and memory usage of these tools within the Galaxy framework.
- Identify the strengths and weaknesses of each tool, providing recommendations for their use in various metagenomic analysis contexts.
- Contribute to the refinement and improvement of metagenomic analysis pipelines by highlighting areas for future development.

The outcomes of this research will not only highlight the strengths and weaknesses of each tool but also contribute to the ongoing efforts to refine metagenomic analysis pipelines, ultimately enhancing the accuracy and reliability of microbial community studies.

2 State of the art

2.1 CAMI

2.2 Kraken v2

Kraken v2 is a k-mer-based taxonomic classification tool. It is the improved version of Kraken v1, it consumes 85% less memory, is five times faster than Kraken v1, and still maintains high accuracy [1]. These improvements are achieved by replacing the previously employed sorted list of k-mer/LCA (Lowest Common Ancestor) pairs, which was indexed by minimizers, with a more efficient probabilistic and compact hash table that maps minimizers to their respective LCAs [1]. In contrast to Kraken v1, storing all k-mers, the data structure of Kraken v2 stores only the minimizers corresponding to every k-mer [1].

The default taxonomy database used in combination with Kraken v2 is the NCBI database, it also supports three non-NCBI taxonomy based databases, such as Greengenes 16S sequences database, RDP 16S database, and SILVA SSU Ref NR 99. Additionally, a custom database can be build using known taxonomies [2].

In a study conducted by Lu and Salzberg in 2020, a comparison was made among Kraken v2, Bracken v2.5, and QIIME 2 v2017.11, all used in conjunction with the RDP 11.5, SILVA 132, and Greengenes v13_8 databases. These tools were evaluated using the same simulated data-sets from a study by Almeida *et al.* [3]. The findings of the study revealed that Kraken v2 and Bracken significantly outperformed QIIME 2 in terms of

database build time for the Greengenes and SILVA databases, being up to 100 times faster. Furthermore, in terms of classification time, Kraken v2 and Bracken were up to 300 times faster, consumed up to 100 times less RAM, and demonstrated higher 16S rRNA profiling accuracy compared to QIIME 2 [4].

Another recent study by Odom *et al.* compared Kraken v2, Mothur, PathoScope 2, QIIME 2, and DADA 2, in conjunction with Greengenes v13_8, Kraken v2, SILVA v138, and Refseq v2020 reference databases. The study was conducted using 16S simulated samples retrieved from multiple sources. The study's results showed that the whole-genome metagenomics tools Kraken v2 and PathoScope 2, demonstrated better performance than the 16S analyses tools DADA 2, QIIME 2, and Mothur [5].

List of Tables

List of Figures

Bibliography

[1] Derrick E. Wood, Jennifer Lu, and Ben Langmead. "Improved metagenomic analysis with Kraken 2". In: *Genome Biology* 20.1 (Nov. 2019), p. 257. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1891-0. URL: https://doi.org/10.1186/s13059-019-1891-0 (visited on 08/09/2023).

[2] *kraken2/docs/MANUAL.markdown at master · DerrickWood/kraken2*. en. URL: https://github.com/DerrickWood/kraken2/blob/master/docs/MANUAL.markdown (visited on 09/02/2023).

[3] Alexandre Almeida et al. "Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments". eng. In: *GigaScience* 7.5 (May 2018), giy054. ISSN: 2047-217X. DOI: 10.1093/gigascience/giy054.

[4] Jennifer Lu and Steven L. Salzberg. "Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2". In: *Microbiome* 8.1 (Aug. 2020), p. 124. ISSN: 2049-2618. DOI: 10.1186/s40168-020-00900-2. URL: https://doi.org/10.1186/s40168-020-00900-2 (visited on 08/09/2023).

[5] Aubrey R. Odom et al. "Metagenomic profiling pipelines improve taxonomic classification for 16S amplicon sequencing data". eng. In: *Scientific Reports* 13.1 (Aug. 2023), p. 13957. ISSN: 2045-2322. DOI: 10.1038/s41598-023-40799-x.