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

Amplicon sequencing is a powerful taxonomic classification method used to discover and characterize microbial communities. In this thesis, the rRNA-prediction subworkflow, a part of the MGnify amplicon pipeline v5.0, was successfully ported to the Galaxy platform. Subsequently, the Galaxy-ported subworkflow was benchmarked against Kraken v2 and its MGnify counterpart using beta diversity and relative taxonomic abundance as benchmark measures. The benchmarking process involved samples from previous MGnify analyses, as well as mock samples.

The Galaxy-ported subworkflow consistently outperformed Kraken v2 for both sample types. The ported subworkflow produces overall similar results to MGnify, a slight discrepancy could be attributed to MAPseq. MAPseq, a tool within the rRNA-prediction subworkflow, appeared to be non-deterministic, which lead to slight differences between the results of MGnify and its Galaxy-ported version.

The availability of the rRNA-prediction subworkflow on Galaxy offers several advantages for the microbiome research community such as interoperability, exchange and modification of specific tools, and shareability, downstream applications such as machine learning and differential abundance analysis.

Deutsche Version:

Die Amplicon-Sequenzierung ist eine leistungsstarke taxonomische Klassifikationsmethode, die zur Entdeckung von mikrobiellen Gemeinschaften verwendet wird. In dieser Arbeit wurde der rRNA-Prediction-Subworkflow, ein Teil des MGnify Amplicon-Pipeline v5.0, erfolgreich auf die Galaxy-Plattform portiert. Anschließend wurde der auf Galaxy

portierte Subworkflow anhand von Beta-Diversität und relativer taxonomischer Häufigkeit mit Kraken v2 und dem MGnify-Subworkflow verglichen. Der Benchmarking-Prozess umfasste Proben aus früheren MGnify-Analysen sowie Mock-Proben.
Der auf Galaxy portierte Subworkflow übertraf Kraken v2 konsistent für beide Probentypen. Der portierte Subworkflow liefert insgesamt ähnliche Ergebnisse wie die MGnify-Ergebnisse, eine leichte Abweichung könnte auf MAPseq zurückzuführen sein. MAPseq, ein Tool innerhalb des rRNA-Prediction-Subworkflow, schien nicht deterministisch zu sein, was zu leichten Unterschieden zwischen den Ergebnissen von MGnify und seiner auf Galaxy portierten Version führte.
Die Verfügbarkeit des rRNA-Prediction-Subworkflows auf Galaxy bietet mehrere Vorteile, wie beispielsweise Interoperabilität, den Austausch und die Anpassung von spezifischen Tools sowie die Möglichkeit zu teilen.

Contents

List of Tables

List of Figures