Abstract

Microbiomics, the investigation of microbial communities at different stages of disease, at specific time-points, or at varying conditions in a particular habitat such as distinct areas of the human body, or environmental samples such as clean rooms, or extremophile ecosystems, is one of the most rapidly growing research areas nowadays. This is mainly facilitated by the development of novel molecular classification approaches, as well as the steadily decreasing sequencing costs during the last decade. As this research area is still evolving by new developments in sequencing techniques and constantly growing knowledge, there is a need for novel or adapted methods, approaches, and tools for all the analysis steps of the community characterization and classification workflow.

This thesis introduces new approaches, methods, and tools for important steps in the entire high-throughput characterization and classification process of complex microbial communities. At the experimental design level, the effects of sequencing library normalization on the final community profile and its diversity was investigated. Subsequently, the *Decontaminator*, an effective tool for the removal of contaminating sequences from the target data sets is introduced as a major improvement during sequence pre-processing. For the core step, the taxonomic classification, an internal transcribed spacer (ITS) reference database, for fungal sequences was created. Tests of the ITS amplicon classification, with a hand curated *in-silico* amplified and fully annotated ITS mock community, showed good results for *reference based* classification and *de-novo* OTU picking approaches based on the UNITE ITS reference sequences. Statistical analysis of determined community profiles was extended by methods for differentially abundant feature detection. Therefor, Metastats, edgeR, and limma+voom, were evaluated using simulated count data, revealing that the linear modeling approaches outperform Metastats for bigger library sizes and fold change values. Based on this evaluation result, real community profiles obtained from analyses conducted within this thesis were tested for differentially abundant features. Finally, with the transcriptome analysis of two *Campylobacter fetus* subspecies, the typical ϵ-proteobacterial promoter motif was also confirmed for *C. fetus sp.* Moreover, this kind of analysis introduces a future direction for more detailed investigation of specific members of a microbial community.

Keywords: High-throughput classification, Microbiome, Sequencing, DA feature detection, Transcriptome analysis, Library normalization

Zusammenfassung

Mikrobiomik, die Erforschung von mikrobiellen Gemeinschaften in verschiedenen Krankheitsstadien zu bestimmten Zeitpunkten, oder unter unterschiedlichen Bedingungen, in einem bestimmten Lebensraum (zB Körperregionen, spezielle Umgebungen wie Reinräume oder extremophile Ökosysteme), zählt zu dem am schnellsten wachsenden Forschungsgebieten. Diese Entwicklung wurde im letzten Jahrzehnt hauptsächlich durch Fortschritte im Bereich der neuen molekularen Klassifikationsansätze, und durch stetig sinkende Sequenzierungskosten unterstützt. Durch die Weiterentwicklung der Sequenzierungstechniken und dem stetigen Zuwachs an Wissen auf diesem jungen Forschungsgebiet besteht ein Bedarf an neuen oder verbesserten Verfahren, Methoden, und Werkzeugen für alle Ebenen des Auswertungsprozesses.

Diese Dissertation stellt neue Ansätze, Methoden, und Werkzeuge für die wichtigsten Schritte des gesamten Hochdurchsatz-Charakterisierungs- und Klassifizierungs-Prozesses von komplexen mikrobiellen Gemeinschaften vor. Auf der Ebene des experimentellen Designs wurden die Auswirkungen auf das mikrobielle Profil anhand von normalisierten Sequenz-Bibliotheken untersucht. Der Vorverarbeitungsschritt wurde um den entwickelten *Decontaminator*, einem effektiven Werkzeug zur Erkennung und Entfernung von verunreinigenden Sequenzen, erweitert. Für den Hauptanalyseschritt, der taxonomischen Klassifizierung, wurde eine Referenz-Datenbank für Pilzsequenzen basierend auf der Internal Transcribed Spacer (ITS) Markerregion erstellt. Um die Qualität und Zuverlässigkeit der ITS-Amplikon Klassifikation zu bewerten, wurde eine von Hand kuratierte und vollständig annotierte ITS Mock Gemeinschaft erzeugt, mit deren Hilfe UNITE als brauchbare Ressource für sowohl referenz als auch *de novo* basierte Klassifizierungsmethoden eignet. Die statistische Analyse der ermittelten mikrobiellen Profile wurde um Methoden zur Identifizierung von differenziell abundanten Gruppen erweitert. Dazu wurden die Methoden Metastats, edgeR, und limma+voom, mit simulierten Count-Daten getestet und evaluiert. Hier konnte gezeigt werden, dass die linearen Modellierungsansätze für größere Bibliotheks- und Effekt-Größen bessere Ergebnisse erzielen als Metastats. Schließlich konnte durch die Transkriptom-Analyse von zwei *Campylobacter fetus* Subspezies das ϵ-Proteobakterium Promotor Motiv auch für diese Subspezies bestätigt werden. Darüber hinaus wurde hier mit der durchgeführten Transkriptom-Analyse eine zukünftige Richtung für weiterführende detaillierte Untersuchungen von speziellen Mitgliedern der mikrobiellen Gemeinschaft vorgestellt.

Keywords: Hochdurchsatz Klassifizierung, Mikrobiom, Sequenzierung, DA Feature Identifizierung, Transkriptom Analyse, Library Normalisierung