# Computational Methods for Longitudinal Microbiome Analysis:

# Identification, Modeling, and Classification

BY

Ahmed Metwally
M.Sc. - Computer Science, University of Illinois at Chicago, 2018
M.Sc. - Biomedical Engineering, Cairo University, 2014
B.Sc. - Biomedical Engineering, Cairo University, 2010

Thesis

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2018

Chicago, Illinois

Defense Committee:

Yang Dai (Chair and Co-advisor), Department of Bioengineering
David L. Perkins (Co-advisor), Departments of Bioengineering, Medicine, and Surgery
Jie Liang, Department of Bioengineering
Patricia W. Finn, Departments of Medicine, and Microbiology/Immunology
Bhaskar DasGupta, Department of Computer Science
Jie Yang, Department of Mathematics, Statistics, and Computer Science

To my family for their love and support…

## ACKNOWLEDGMENTS

I want to convey my sincerest gratitude to my advisors - Drs. Yang Dai, David Perkins, and Patricia Finn for their mentorship and boundless support and patience. Dr. Dai, who spent an enormous amount of time thinking about my projects, lent her expertise and critical thinking in order to help me develop my computational expertise. Dr. Perkins, who fostered my enthusiasm in science, taught me how to take the less traveled path and branch out in novel directions in research. Thank you for your patience in teaching me the biological aspects of the microbiome and host immune response. Dr. Finn, who made my graduate study a blessing, imparted a wide breadth of knowledge, from science to leadership. I will never forget her invaluable pearls and recommendations on each of our projects, papers, and grants.

I am also thankful to the members of my thesis committee, Drs. Jie Liang, Bhaskar DasGupta, and Jie Yang for their time and expertise in shaping my dissertation. Your encouragement and insightful comments are greatly appreciated. I would especially like to thank Dr. Jie Liang, who has been a fabulous mentor and continually pushes me to challenge myself with taking on complex problems in bioengineering and bioinformatics. I also want to thank our wonderful collaborators, whose computational and clinical expertise have taught me so much; Dr. Philip Yu (Computer Science, UIC) and Dr. Thomas Ferkol (Washington University School of Medicine).

I am forever indebted to my parents. Without their love and support, I would never be the person I am now. Their acumen in engineering and science and their own deep intellectual curiosity inspired me to pursue my Ph.D. studies. To me, you are the greatest human beings on the

face of this planet. Thanks to my awesome brother, Sherif, my beautiful sister, Rehab, and my fabulous brother-in-law, Mohamed. The three of you will always be in my heart.

Thanks to my stunning wife, Sally, who I consider to be an important co-author of this dissertation. Her generosity, sacrifice, kindness, understanding, support, and patience are priceless. Thanks to my beloved son, Ezz, who, despite all his crying (and subsequent sleepless nights), also gave me the strength and motivation to work as hard as I can to make him proud.

Over the past two years, I have had the pleasure to serve as the elected global IEEE EMBS student representative. During my tenure, I had the honor to work closely with the executive and administrative boards on multiple projects. This experience has had a significant impact on my career and personal development. I am thankful for the mentorship of Drs. Steve Wright (Texas A&M University), Bin He (Carnegie Mellon University), May Wang (Georgia Institute of Technology), Shankar Subramaniam (University of California, San Diego), Jim Patton (UIC), Ahmed Morsy (Cairo University), Lisa Lazareck-Asunta (Wellcome Trust), and Bruce Hecht (Analog Devices, Inc.).

During my two summer internships at Thermo Fisher Scientific, I had the opportunity to work with a group of renowned computational scientists. All of them have strengthened my computational genomics experience; a big thank you to Drs. Ann Mongan, Fiona Hyland, Sameh El-Difrawy, Earl Hubbell, Timothy Looney, Alex Pankov, Lifeng Lin, Simon Cawley, Dumitru Brinza, Ruchi Chaudhary, and Manimozhi Manivannan. Their passion towards revolutionizing the genomics industry is impressive!

**ACKNOWLEDGMENTS (Continued)**

Last but not least, I am grateful and fortunate to be surrounded by very smart and ambitious colleagues at UIC (some of whom have ventured onto new exciting career adventures); Dr. Jennifer Kwan (thanks for all your helpful edits & leadership discussions), Cody Schott, Ben Turturice, Kathryn Dominguez "our lifesaver", Dr. Yue Huang, Dr. Christian Ascoli, Derek Reiman, Ankit Jambusaria, Shang Gao, Mladen Rasic, Brian Nguyen, Ronnie Knowlton, Dr. Ravi Ranjan, Dr. Asha Rani, Dr. Halvor McGee, Dr. Zahraa Hajjiri, Yang Chen, Anna Salapatas, Amira Kefi, Dr. Gamze Gursoy, Anna Terebus, and Alan Perez-Rathke. Your scientific advice and fun/memorable interactions actually helped make my PhD training an enjoyable journey. Thank you for your collegiality and friendship!

## Contribution of Authors

**Chapter 1** is a brief introduction to my dissertation problems and highlights the significance of my research outcome.

**Chapter 2** is partially based on published work in collaboration with former postdoctoral members at Finn-Perkins' laboratory and a Ph.D. student at Dai's lab. Drs. Ravi Ranjan, Asha Rani, and Halvor McGee collected, processed, and sequenced biological samples used in the first paper. I was the major driver in developing computational strategies and applying them to analyze the data. In the second paper, Derek Reiman extended the core idea that I suggested and implemented the method. This chapter is partially based on the following publications (The necessary permissions are provided in the appendices):

- Ranjan R, Rani R, **Metwally AA**, McGee H, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, 2016.
- Reiman D, **Metwally AA**, Dai Y. *PopPhy-CNN*: A Phylogentic Tree Embedded Architecture for Convolution Neural Networks for Metagenomic Data. *bioRxiv*, 2018.

**Chapter 3** is based on published work for which I was the primary author and major driver of the research. I designed, developed, and evaluated the method. I also wrote the first draft of the manuscript. Dr. Yang Dai helped me with graph theory concepts. Drs. Finn and Perkins helped me in emphasizing the biological significance and writing the first manuscript. Asem Alaa

## Contribution of Authors (Continued)

developed some visualization graphs that are presented in the second manuscript. This chapter is based on the following publications (The necessary permissions are provided in the appendices):

- **Metwally AA**, Dai Y, Finn PW, Perkins DL. *WEVOTE*: Weighted Voting Taxonomic Identification Method of Microbial Sequences. *PLoS ONE*, 2016.
- Alaa A, **Metwally AA**. Cloud-based Solution for Improving Usability and Interactivity of Metagenomic Ensemble Taxonomic Identification Methods. *IEEE EMBS Biomedical and Health Informatics*, 2018.

**Chapter 4** is based on published work for which I was the primary and key driver of the research. I initiated the core idea of the method, prepared the mathematical derivation, performed simulation studies, and implemented the R-package, and wrote the first draft of the papers. Drs. Yang Dai, Patricia Finn, and David Perkins supervised me in the process of developing the method and writing the papers. Dr. Jie Yang (from Mathematics, Statistics, and Computer Science department) helped me validating the mathematical derivation of the method. Dr. Christian Ascoli, a pulmonary fellow at Finn-Perkins laboratory, helped me emphasize the clinical significance of the method outcome. This chapter is based on the following publications (The necessary permissions are provided in the appendices):

- **Metwally AA**, Yang J, Ascoli C, Dai Y, Finn PW, Perkins DL. *MetaLonDA*: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. *Microbiome*, 2018.

## Contribution of Authors (Continued)

- **Metwally AA**, Finn PW, Dai Y, Perkins DL. Detection of Differential Abundance Intervals in Longitudinal Metagenomic Data Using Negative Binomial Smoothing Spline ANOVA. *In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017.

**Chapter 5** is based on unpublished work that I was the primary author and key driver of research. Dr. Philip Yu helped me in formulating the problem. Derek Reiman helped me in the method evaluation. Drs. Yang Dai, Patricia Finn, and David Perkins supervised me in the process of developing the method and writing the papers. This chapter is based on the following manuscript:

- **Metwally AA**, Yu PS, Reiman D, Dai Y, Finn PW, Perkins DL. Utilizing Longitudinal Gut Microbiome Taxonomic Profiles to Predict Food Allergy via Sparse Autoencoder and Long Short-Term Memory Network. *Under-review*.

**Chapter 6** is based on unpublished work for which I was the primary author and key driver of research. Dr. Christian Ascoli helped me in interpreting clinical measurements. Drs. Asha Rani and Ravi Ranjan prepared and sequenced the samples. Dr. Thomas Ferkol provided us with the samples from Washington University School of Medicine in St. Louis. This chapter is based on the following publication:

- **Metwally AA**, Ascoli C, Rani A, Ranjan R, Ferkol TW, Finn PW, Perkins DL. Lower Airway Microbiome Dynamics as a Predictor of Bronchiolitis Obliterans Syndrome after Pediatric Lung Transplantation in Cystic Fibrosis. *Manuscript in-preparation*.

## Contribution of Authors (Continued)

**Chapter 7** is a review of the main topics of this dissertation, highlighting the novel contribution of our methods, and discussing the strengths and weaknesses of our approaches. We also provide perspectives on potential future studies.

## TABLE OF CONTENTS

CHAPTER PAGE

## TABLE OF CONTENTS (Continued)

**CHAPTER** **PAGE**

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF FIGURES (Continued)

**FIGURE** **PAGE**

## LIST OF ABBREVIATIONS

| | |
|---|---|
| ACR | Acute Cellular Rejection |
| BAL | Bronchoalveolar Lavage |
| BH | Benjamini-Hochberg |
| BMI | Body Mass Index |
| BOS | Bronchiolitis Obliterans Syndrome |
| CF | Cystic Fibrosis |
| CFTR | Cystic Fibrosis Transmembrane conductance Regulator |
| CLAD | Chronic Lung Allograft Dysfunction |
| CNN | Convolutional Neural Network |
| DBN | Deep Belief Network |
| DNN | Deep Neural Network |
| FEV1 | Forced Expiratory Volume in 1 second |
| FVC | Forced Vital Capacity |
| GLMMs | Generalized Linear Mixed Models |
| HLA | Human Leukocyte Antigen |

## LIST OF ABBREVIATIONS (Continued)

| | |
|---|---|
| HMM | Hidden Markov Model |
| HMP | Human Microbiome Project |
| IgE | Immunoglobulin E |
| IMMs | Interpolated Markov Models |
| IRB | Institutional Review Board |
| KL | Kullback Leibler |
| KS-test | Kolmogorov-Smirnov test |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| LCA | Lowest Common Ancestor |
| LOWESS | Locally Weighted Scatterplot Smoothing |
| LPS | Lipopolysaccharides |
| LSTM | Long Short-Term Memory |
| LTx | Lung Transplant |
| MetaLonDA | Metagenomic Longitudinal Differential Abundant |
| MGS | Metagenomic Shotgun |
| NB | Negative Binomial |
| NCBI | National Center for Biotechnology Information |

## LIST OF ABBREVIATIONS (Continued)

| | |
|---|---|
| NGS | Next Generation Sequencing |
| NMDS | Non-metric Multi-Dimensional Scaling |
| OTU | Operational Taxonomic Unit |
| ReLU | Rectified Linear Unit |
| RF | Random Forest |
| RNN | Recurrent Neural Network |
| SS-ANOVA | Smoothing Spline ANOVA |
| SVM | Support Vector Machine |
| T1D | Type 1 Diabetes |
| T2D | Type 2 Diabetes |
| WEVOTE | WEighted VOting Taxonomic idEntification |
| ZIP | Zero-Inflated Poisson |

## SUMMARY

The microbiome plays a vital role in host-immune responses resulting in significant effects on host health. Dysbiosis of the microbiome has been linked to diseases including asthma, obesity, diabetes, and inflammatory bowel disease. Over the past decade, culture-independent sequencing methods have revolutionized microbiome studies through identification of the genetic content of microbial communities in the form of millions to billions of short DNA sequences. The sequences derived from the microbiome originate from thousands of different species that need to be identified, quantified, and compared over time among disease phenotypes. These analyses can detect biomarkers that may be used for microbial reconstitution through bacteriotherapy, probiotics, or antibiotics.

Current taxonomic identification methods that achieve high precision can lack sensitivity in some applications. Conversely, methods with high sensitivity can suffer from low precision and require long computation time. Thus, highly accurate and sensitive taxonomic identification methods are needed. Furthermore, in longitudinal studies, sample collection suffers from all forms of variability such as a different number of subjects per phenotypic group, a different number of samples per subject, and samples not collected at consistent time points. These inconsistencies make current analysis methods unsuitable and create opportunities for the development of new methods. In addition, given the strong association between microbiome and disease, computational models can be built to predict disease status or prognosis using longitudinal microbial profiles.

## SUMMARY (Continued)

In this thesis, we discuss the computational methods and tools we have developed that improve both the characterization and longitudinal analysis of the microbiome. The first method, *WEVOTE*, classifies microbial sequences into taxonomic units with both high precision and high sensitivity. The second method, *MetaLonDA*, identifies time intervals of differentially abundant microbial features in longitudinal studies. The third method is a computational framework to predict host clinical phenotype from longitudinal microbiome profiles via deep learning approach. Finally, using these methods and tools, we identified microbiome dynamics suggestive of the development of bronchiolitis obliterans syndrome in pediatric lung transplant recipients, insights that can be leveraged to improve lung transplant outcomes across life span.

# CHAPTER 1

# INTRODUCTION

The microbiome plays a vital role in a broad range of host-related processes and has a significant effect on host health, and its dysbiosis has been linked to various diseases such as asthma, obesity, diabetes, inflammatory bowel disease, etc. Over the past decade, culture-independent sequencing has revolutionized microbiome studies by quickly deciphering the genetic content of microbial communities in the form of millions to billions of short DNA sequences. Moreover, the exponential decay in sequencing cost makes large-scale longitudinal studies affordable and appealing. These sequences, originating from thousands of different species, need to be identified, quantified, and compared over time between phenotypes in order to extract biomarkers, which may be used for microbial reconstitution through bacteriotherapy, probiotics, or antibiotics. Given the strong association between microbiome and disease, computational models can be built to predict diseases status or prognosis using longitudinal microbial profiles.

## 1.1 Problem Identification

One of the key steps in microbial data analysis is the taxonomic classification of sequence reads in a metagenomic dataset. Different methods can generate variation in taxonomic output profiles for the same input dataset. Sample type, sequencing error, and read length are the main factors that cause variation. This variation in the predicted taxonomic annotations presents a challenge to investigators in the selection of identification methods and the interpretation of an-

notations. Hence, developing a method that has a high precision of annotating the metagenome shotgun sequencing (MGS) sequence reads is relevant and merits investigation. In this thesis, we first studied the problem of taxonomic classification of microbial sequences. We developed *WEVOTE* (WEighted VOting Taxonomic idEntification), a phylogenetic-based ensemble method that classifies MGS DNA sequence reads based on an ensemble of existing methods using k-mer-based, marker-based, and naive-similarity based approaches.

The recent advances in DNA sequencing technologies and rapid reduction in costs have fostered longitudinal analyses, which include multiple samples per subject over time. These longitudinal studies provide increased insights into the underlying biological mechanisms of the microbiome role in health and disease. In addition to identifying differentially abundant features, detecting the time intervals where these features exhibit changes in their abundance between two phenotypes in longitudinal studies adds insights into disease pathogenesis. Thus, there is a need to develop a method that can accurately identify the time intervals (start or end) of microbial features (taxon, genes, or pathways) wherein they are differentially abundant between the two phenotypes in a longitudinal study. The method should also be able to handle the inconsistencies of the sample collection process such as the number of samples per subject maybe unequal and obtained at inconsistent time points. In this thesis, we study the problem of detection of differential abundant microbial features and their significant time intervals that are associated with phenotypes. We developed a statistical method, *MetaLonDA* (Metagenomic Longitudinal Differential Abundant), that is based on a semi-parametric Smoothing Spline ANOVA and negative binomial distribution to model the time-course of the features between two phenotypes. The method ac-

curately identifies the time intervals when the features are differentially abundant between two phenotypic groups.

Another focus of the microbiome analysis is the classification and prediction of host phenotype based on microbiome profiles. Since the microbiome has been linked to various diseases, there is an opportunity to develop novel methods that predict the host phenotype based on their microbiome profiles. Recently, multiple deep learning frameworks have been applied to the microbiome phenotype prediction; including Convolutional Neural Networks and Deep Belief Networks. Another deep learning architecture that is being used in time-series prediction applications is called Recurrent Neural Networks. It takes as its input, not only the current input, but also what it perceived in the previous step in time, which is valuable in the analysis over time since it is possible to express a change in the state of the network without having an explicit state. In this thesis, we developed a deep learning framework to predict phenotypes from longitudinal microbiome taxonomic profiles. The method considers the dependency between adjacent longitudinal microbiome profiles.

Apotentail intersting application of these methods is to analyze the dynamics of microbiome in cystic fibrosis (CF) lung transplant recipients. CF is a rare genetic disease that is caused by various mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene which functions as a chloride transporter. It is uncommon for people with cystic fibrosis to live into their 40s, 50s, or beyond. In an effort to improve survival and quality of life, lung transplantation has become an effective therapeutic option and is the only definitive therapy for selected patients with end-stage CF. Median survival in pediatric patients who undergo bilateral lung transplant

is about 5.6 years. Bronchiolitis Obliterans Syndrome (BOS) is the most common cause of re-transplantation and death. 50% of transplant recipients develop BOS in the first 5 years post transplantation; however the etiology is unknown. Since transplant patients are treated with immunosuppression and antibiotics, we hypothesized that the microbiome plays a role in BOS development which is not clearly defined. Using methods developed in this thesis, we analyzed the longitudinal microbiome profiles of pediatric CF patients post-transplant to illuminate the role of the microbiome in BOS development.

## 1.2 Thesis Outline

The focus of this thesis is the development of novel computational methods to analyze longitudinal microbiome profiles for a better understanding of microbiome associated diseases. The research described in this dissertation is organized as follows:

In **Chapter 2**, we provide a background of microbiome research and diseases associated with the microbiome, and covers recent major longitudinal microbiome projects. We also discuss the current advances in computational methods that are being used to analyze microbiome data. This chapter is partially based on the following publications (The necessary permissions are provided in the appendices):

- Ranjan R, Rani R, **Metwally AA**, McGee H, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, 2016.

- Reiman D, **Metwally AA**, Dai Y. *PopPhy-CNN*: A Phylogentic Tree Embedded Architecture for Convolution Neural Networks for Metagenomic Data. *bioRxiv*, 2018.

In **Chapter 3**, we briefly discuss taxonomic identification methods including state of the art of methods and their limitations. Then, we introduce *WEVOTE*, a phylogenetic based ensemble method that classifies microbial sequences with very high sensitivity and the highest precision among state-of-the-art methods. Next, we introduce *WEVOTE-web*, a cloud-based version of the *WEVOTE* algorithm. This chapter is based on the following publications (The necessary permissions are provided in the appendices):

- **Metwally AA**, Dai Y, Finn PW, Perkins DL. *WEVOTE*: Weighted Voting Taxonomic Identification Method of Microbial Sequences. *PLoS ONE*, 2016.
- Alaa A, **Metwally AA**. Cloud-based Solution for Improving Usability and Interactivity of Metagenomic Ensemble Taxonomic Identification Methods. *IEEE EMBS Biomedical and Health Informatics*, 2018.

In **Chapter 4**, we introduce *MetaLonDA*, a method that can identify significant time-intervals of differentially abundant microbial features such as taxonomies, genes, or pathways associated with phenotypes. We show our benchmarking of *MetaLonDA* using both simulated and biological datasets. We also, introduce the R-package that implements the *MetaLonDA* method. This chapter is based on the following publications (The necessary permissions are provided in the appendices):

- **Metwally AA**, Yang J, Ascoli C, Dai Y, Finn PW, Perkins DL. *MetaLonDA*: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. *Microbiome*, 2018.
- **Metwally AA**, Finn PW, Dai Y, Perkins DL. Detection of Differential Abundance Intervals in Longitudinal Metagenomic Data Using Negative Binomial Smoothing Spline ANOVA. *In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017.

In **Chapter 5**, we introduce our development of a deep learning framework that has the capacity to predict food allergy from longitudinal microbiome profiles. The framework is based on sparse autoencoder and Long Short-Term Memory networks. We also report results of applying the developed framework to the DIABIMMUNE study. This chapter is based on the following publication:

- **Metwally AA**, Yu PS, Reiman D, Dai Y, Finn PW, Perkins DL. Utilizing Longitudinal Gut Microbiome Taxonomic Profiles to Predict Food Allergy via Sparse Autoencoder and Long Short-Term Memory Network. *Under-review*.

In **Chapter 6**, we analyze the association between changes in the composition of the lower airway microbiome and the BOS development. This chapter is based on the following publication:

- **Metwally AA**, Ascoli C, Rani A, Ranjan R, Ferkol TW, Finn PW, Perkins DL. Lower Airway Microbiome Dynamics as a Predictor of Bronchiolitis Obliterans Syndrome after Pediatric Lung Transplantation in Cystic Fibrosis. *Manuscript in-preparation*.

Finally, in **Chapter 7**, we review the main topics of this dissertation, highlight the novel contribution of our methods, and discuss the strengths and weaknesses of our approaches. We also provide perspectives on potential future studies.

# CHAPTER 2

# OVERVIEW OF MICROBIOME STUDIES AND DATA ANALYSIS METHODS

## 2.1 Microbiome Studies

The microbiome, a dynamic ecosystem of microorganisms (bacteria, archaea, fungi, and viruses) that live in and on us, plays a vital role in host-immune responses resulting in significant effects on host health. Dysbiosis of the microbiome has been linked to diseases including asthma, obesity, diabetes, transplant rejection, and inflammatory bowel disease (1; 2; 3; 4; 5; 6). These observations suggest that modulation of the microbiome could become an important therapeutic modality for some diseases. For example, fecal transplants have been shown to alleviate diarrhea caused by *Clostridium difficile* infection and temporarily improve insulin sensitivity (7; 8). Specifically, the gut microbiome, which has been the most extensively studied of the human microbiomes, is highly diverse and has been shown to include thousands of different bacterial species (9; 10). This diverse community of bacteria is composed of a few species that are highly abundant and a large amount of species that are found in trace amount (11). The human microbiome can be divided into the core microbiome and the variable microbiome (12). The core microbiome is the set of taxa or genes that present in a given body location (gut, kidney, skit, oral, etc.) in almost all humans. The variable microbiome arises from various factors such as host physiological status, host environment, host genotype, host lifestyle, host pathobiology, etc.

The number of studies investigating the microbiome has risen exponentially since the technological advances in high-throughput sequencing that have led to culture- and cloning-independent analysis (13). Sequencing technologies are able to identify the genetic content of microbial communities in the form of millions to billions of short DNA sequences. These technical advances have been paradigm shifting since the majority (>90%) of microbial species cannot be readily cultured using current laboratory culture techniques (14; 15; 16).

The most common sequencing approach to analyze the microbiome is amplicon analysis of the 16S ribosomal RNA (rRNA) gene (17; 18). In this method, a 16S rRNA region is amplified by PCR with primers that recognize highly conserved regions of the gene and sequenced (19). The limitations of this method are that the annotation is based on a putative association of the 16S rRNA gene with taxa defined as an operational taxonomic unit (OTU). In general, OTUs are analyzed at the phyla or genera level and can be less precise at the species level. In 16S rRNA sequencing, genes are not directly sequenced, but rather predicted based on the OTUs. Due to horizontal gene transfer and the existence of numerous bacterial strains (20; 21; 22), the lack of direct gene identification potentially limits understanding of a microbiome.

An alternative approach to the 16S rRNA amplicon sequencing is metagenome shotgun (MGS) sequencing in which random fragments of genome are sequenced. MGS is more expensive and requires more extensive data analysis (13; 23; 24; 25). A major advantage of the MGS sequencing is that the taxa can be more accurately defined at the species level. In addition, to identify and understand the bacterial genes in a taxon, it may be necessary to sequence a genome with high coverage (23).

## 2.2 Major Microbiome Projects

In this section we summarize three major microbiome projects to date; the Human Microbiome Project (HMP) (12; 17; 26), the Integrative Human Microbiome Project (iHMP) (27; 28; 29), and the DIABIMMUNE Project (1; 30; 31).

### 2.2.1 HMP: Human Microbiome Project

**HMP phase I:** The National Institutes of Health Human Microbiome Project (HMP) was established in 2008, with the mission of generating resources that would enable the comprehensive characterization of the human microbiome and analysis of its role in human health and disease. The HMP-I characterized the microbial communities from 242 healthy adults (129 males and 113 females) between the ages of 18 and 40. Women were sampled at 18 body habitats, and men at 15 (excluding the three vaginal sites), distributed among five major body areas: nasal cavity, oral cavity, skin, gastrointestinal tract, and urogenital tract (17; 32).

**HMP phase II:** The second phase of HMP targeted diverse body sites with multiple time points in 265 individuals. Strain identification revealed subspecies clades specific to body sites (26). It also quantified species with phylogenetic diversity under-represented in the isolate genomes. Body-wide functional profiling classified pathways into universal, human-enriched, and body site-enriched subsets. Finally, temporal analysis decomposed microbial variation into rapidly variable, moderately variable, and stable subsets. HMP-II enables an understanding of personalized microbiome functions and dynamics (26).

### 2.2.2 iHMP: Integrative Human Microbiome Project

iHMP was established with the aim of creating integrated longitudinal datasets from both the microbiome and host from three different cohort studies of microbiome-associated conditions using multiple omics technologies (27; 28; 29). The three cohorts are:

1. **Pregnancy and Preterm Birth Cohort:** The multi-omic microbiome pregnancy initiative is established to better understand how microbiome and host profiles change throughout pregnancy and influence the establishment of the nascent microbiome in neonates. The study aims to recruit 2000 pregnant women and their neonates (27).
2. **Inflammatory Bowel Disease Cohort:** Inflammatory bowel disease (IBD), which includes both Crohn's disease and ulcerative colitis, is one of the most-studied imbalances between microbes and the immune system. There exist genetic and environmental risk factors that are associated with IBD (33; 34). However, they are inadequate to explain the dramatic increase in IBD over the past 50 years (35). Rather, comprehensive evidence has linked IBD to the gut microbiota (36; 37). In contrast to traditional disease models, no single pathogen seems to cause IBD. The IBD multi-omics project has been established to provide comprehensive insights into the gut microbial ecosystem in the context of IBD. This will improve our ability to understand, diagnose, and treat IBD (28; 29).
3. **Type 2 Diabetes Cohort:** Differences in the gut microbiome have been noted between diabetics and healthy individuals (38; 39), and direct alteration of the microbiome in mice has been shown to lower blood glucose levels (40). The longitudinal multi-omic study is

aimed to better understand the biological changes that occur type 1 diabetes (T2D) disease acquisition. This cohort will consist of approximately 100 individuals at risk for diabetes. Samples will frequently be taken (at 1-to-4-day intervals) during infected and other stress states and less frequently (every $\sim$ 3 months) during healthy periods, with a minimum of 27 timepoints sampled per subject.

### 2.2.3 DIABIMMUNE Project

The DIABIMMUNE project is established with the aim of studying the hygiene hypothesis. The hygiene hypothesis is widely supported theory that accounts for the association between the increase in diseases of the immune system and decreased exposure to pathogens. The DIABIMMUNE project analyzes the microbiome in subjects from three separate countries (Russia, Finland, and Estonia) in order to explore this phenomenon (30; 1; 31). These three populations comprise a "living laboratory," offering a unique opportunity to test the hygiene hypothesis and gene-environmental interactions in the development of autoimmune disease. DIABIMMUNE is divided into three cohorts as described below:

1. **Type 1 Diabetes Cohort:** Characterization of developing microbiome in 33 infants en route to type 1 diabetes (T1D). This is a prospective, longitudinal analysis of developing gut microbiome in infants en route to type 1 diabetes (30). Infants from Finland and Estonia were recruited at birth based on predisposition to autoimmunity determined by human leukocyte antigen (HLA) genotyping. Parents collected their infants' stool samples at approximately monthly intervals. The cohort consists of 33 infants, 11 of which seroconverted

to serum autoantibody positivity and of those, four developed T1D within the three-year time-frame of this study.

2. **Three Country Cohort:** Early infancy is considered to be an important time for the maturation of the immune system. This cohort consists of collected stool samples from infants from Finland, Estonia, and Russia to elucidate mechanisms behind the hygiene hypothesis (1). The three countries have substantial differences in incidence of T1D and allergies. The prevalence of the allergies is highest in Finland and lowest in Russia with Estonia intermediate. 74 infants from each country were selected on the basis of similar HLA risk and matching gender and followed up from birth until the age of three. For each infant, three years of monthly stool samples, laboratory assays, and questionnaires regarding breastfeeding, diet, allergies, infections, family history, use of drugs and clinical examinations were collected.

3. **Antibiotics Cohort:** The gut microbial community is dynamic during the first three years of life, before stabilizing to an adult-like state. However, little is known about the impact of environmental factors on the developing human gut microbiome. This cohort is designed to characterize the development of the infant gut microbiome and the impact of repeated antibiotic exposure on bacterial strain diversity and stability (31). The cohort consists of 39 infants followed up for 3 years. Stool samples and clinical information are collected, approximately half of whom received multiple courses of antibiotics during the first three years of life. The microbiota of antibiotic-treated children was less diverse in terms of both

bacterial species and strains, with some species often dominated by single strains. In addition, short-term composition changes between consecutive samples from children treated with antibiotics. Antibiotic resistance genes carried on microbial chromosomes showed a peak in abundance after antibiotic treatment followed by a sharp decline, whereas some genes carried on mobile elements persisted longer after antibiotic therapy ended.

## 2.3 Computational Microbiome Analysis Methods

### 2.3.1 Taxonomic Classification of Microbial Sequences

Metagenomic sequences come from a number of different species, some of which either have a previously sequenced reference genomes or have a related sequenced species sufficiently close phylogenetically. Other sequences, however, may come from organisms that have no sufficiently close relatives with sequenced genomes, or from DNA fragments that show no significant similarity with any available genomic sequence.

The metagenomic classification problem is to assign each sequence of the metagenome to a corresponding taxonomic unit or to classify it as 'novel.' Methods performing taxonomic classification of metagenome sequences can be, in large, grouped into three categories; alignment-based sequence classifiers, alignment-free sequence classifiers, and ensemble sequence classifiers. In the following section, we provide a summary of each category and list of methods that are based on the corresponding design.

**Alignment-based Sequence Classifiers**

In this category, metagenome sequences are aligned to each of the known genomes from the reference database, in order by using best alignment score as an estimator of the phylogenetic "closeness" between the sequence and the genome. This could be done with generic alignment program, such as *BLAST* (41), *Bowtie* (42), *BWA* (43), or *BLAT* (44). Although this approach is highly sensitive, it tends to produce a lot of false positives and the running time of these methods makes it unfeasible to be used in its naive form for large datasets (45; 46). For this reason, a sub-category of classification methods has been developed with the design of preprocessing the sequence files and database before performing the alignment. This reduces the computational time and resources needed and also increases the annotation specificity. One technique to reduce the time complexity of the alignment is by assembling the short reads into longer contigs then annotate these contigs (47). Although annotating contigs is more efficient than annotating reads, this approach is fragile due to the complexity of the assemblers and the assembly error. Another approach is to use the information preserved in phylogeny to annotate metagenome sequence. Examples of methods covering phylogenetic-based classification are *MetaPhlAn* (48; 49), *MEGAN* (50; 51; 52), *MetaPhyler* (53), and *TIPP* (54).

**Alignment-free Sequence Classifiers**

Another strategy to cope with increasingly large metagenomic datasets is the alignment-free methods. Generally, most of alignment-free methods are based on the analysis of words, which are usually of fixed size ($k$-mer). The basic idea behind the $k$-mers methods is performing metage-

nomic classification of next generation sequencing (NGS) sequence reads based on the analysis of shared $k$-mers between a metagenome read and each genome from a pre-compiled database. Given a taxonomic tree involving the species of the database, these tools "map" each read to a node of the tree, thus reporting the most specific taxon or clade the read is associated with. Based on the obtained counts and tree topology, algorithms assign each read to a tree node "best explaining" the counts. Examples of these methods are *Kraken* (46), *LMAT* (55), *Centrifuge* (56), *CLARK* (57), *Kaiju* (58), *ProPhyle* (59), Seed-*Kraken* (60), *CoMeta* (61), One *Codex* (62), and *SMART* (63).

**Ensemble Read Classifiers**

Ensemble read classifier is the approach we proposed (Chapter 3) to classify microbial sequences based on ensemble of base classifiers.

**Computational Challenge in Classification Methods**

A major computational problem of the taxonomic classification of microbial sequences is that sometime the analysis of same sample by different methods produces different results (Figure 1). This raises controversy on the false positive problem, and it is usually attributed to the design of the method (64; 65; 66; 67; 68). Some methods favor speed over accuracy, while the other favor the sensitivity. The conclusion from these benchmarking studies is that current individual taxonomic methods are not the best approach to identify microbial profile and hence developing ensemble methods may be the solution for such metagenomic sequence classification problem.

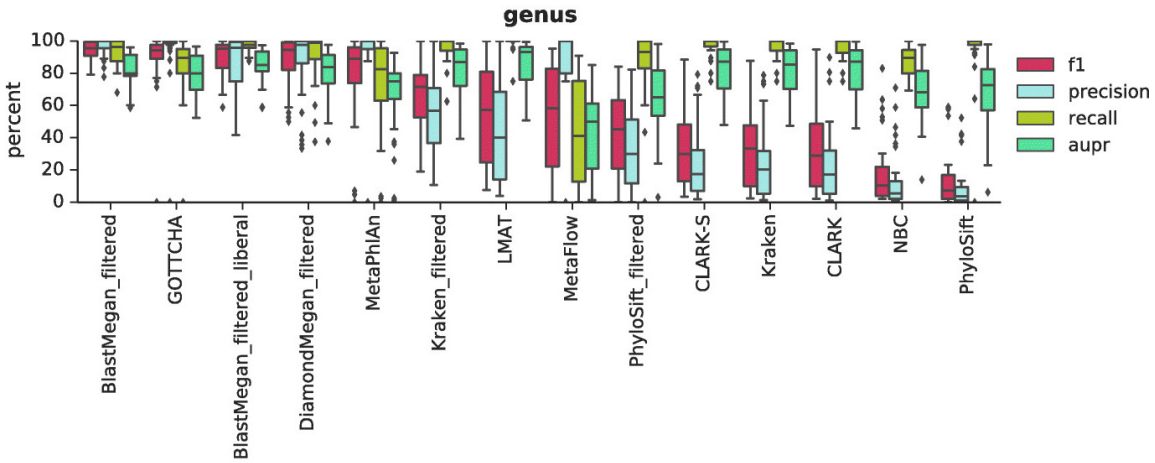



Figure 1: Different methods produce different results for the same input. Evaluation of different taxonomic classification method using 35 simulated and biological metagenome (64).

### 2.3.2 Differential Abundance Analysis

One of the objectives of the microbiome studies is to determine whether there is a particular microbial signature (e.g., taxa or genes) associated with a particular disease state and/or disease outcome. These biomarkers can play an important role in the development of preventative and therapeutic strategies. Modeling metagenomic data for disease-association studies is an active area of research. The standard parametric models may reduce the variance in read counts if the data follows the corresponding parametric distribution, but the models may be substantially biased if the data does not support that distribution. On the other hand, non-parametric models do not assume any prior distribution of the data and thus are not biased towards any distribution, but these models may suffer from a huge model variance (69).

Longitudinal studies of the microbiome have gained tremendous popularity during the past few years due to the ability to detect trends of microbiome changes over time and relate these changes to disease progression in different parts of body, such as gut, kidney, skin, or lung (70; 30; 1; 71; 72; 73). In addition, there has been a drastic reduction in sequencing cost that has made longitudinal studies more affordable on a large scale.

Analyzing longitudinal metagenomic profile data is different from analyzing a single time point profile. An individual's microbiome evolves over time, but its composition has some dependency on its previous structure (i.e., Markov Process), despite the independence between samples. For longitudinal data, two types of differential abundance analysis are widely utilized: (a) Treat data from each time point independently and detect features that have differential abundance be-

tween the phenotypes at individual time points (74); (b) Identify features that have differential abundance during the time-course within a phenotype (75; 76).

Another strategy for longitudinal differential abundance is to identify time intervals of differentially abundant microbial features. To date, two methods have been proposed; the first is *MetaSplines* (77), and the second is *MetaDprof* (78). *MetaSplines* and *MetaDprof* are both based on the Gaussian Smoothing Spline ANOVA (SS-ANOVA) approach (79; 80; 81), where the Gaussian distribution is used to model the number of reads mapped to each microbial feature. *MetaSplines* has a higher sensitivity of detecting time intervals of differentially abundant features than *MetaDprof*, but *MetaDprof* has higher specificity (78). *MetaDprof* has a major drawback, namely, its implementation assumes consistency in longitudinal microbial samples, such that it is only able to perform the analysis on an equivalent number of subjects per phenotypic group, the same number of samples from each subject, and the same elapsed time between adjacent time points. However, these conditions are rarely fulfilled in human microbiome longitudinal studies.

**Challenges in Differential Abundance Analysis**

Longitudinal analysis is usually challenged by variability in longitudinal sample collections, including inconsistencies in the number of subjects per phenotype, number of samples per subject, and sample collection at inconsistent time points. These inconsistencies increase with the level of difficulty with which samples are obtained from the subjects. For example, in humans, the variability decreases in samples collected non-invasively (e.g., stool and urine samples) but increases in the invasive procedures (e.g., bronchoalveolar lavage (BAL) samples which are extracted from the lung by bronchoscopy).

One solution to address this variability is to bin samples into a certain number of windows between the start and end times of the study course by selecting the nearest sample in time for each bin (30), then compare the microbial feature's relative abundance or diversity indices (82; 83) between any pair of time points to characterize any pairwise changes. The limitation of this approach is that it deals with the longitudinal data points as a collection of static snapshots and ignores temporal dependencies. Furthermore, if more than one sample is taken in the same time window, it may result in either retaining only one sample and excluding the others or taking the average of the measured feature's values, which may lead to mischaracterizing the exact microbial behavior.

### 2.3.3 Host Phenotype Prediction

The other primary task of microbiome analysis is the prediction of host phenotype based on microbiome profiles. Since microbiome has been linked to numerous diseases, it opens a door to develop new methods that predict the host phenotype based on their microbiome profiles. Approaches using traditional machine learning models, e.g., Random Forest (RF), LASSO and Support Vector Machines (SVMs), and recently, deep neural networks (DNN), demonstrated the potential of developing microbial biomarker signature for the prediction of disease or phenotype of the host (84; 85; 86; 87; 88). This type of approaches is motivated by the findings that a microbial signature for the host phenotype may be complex, involving simultaneous over- and under-representations of multiple microbial taxa at distinct taxonomic levels and potentially interacting with each other (85; 89). Varying levels of predictive accuracy have been reported. The perfor-

mance of the deep learning models is encouraging, owing to the ability of deep architectures in identifying potential interactions of microbial taxa for disease prediction (85).

Recently, we have proposed a prototype of a novel architecture for convolution neural networks (CNNs) for the prediction of host phenotype from the microbial taxonomic abundance profiles (90; 91). CNNs were originally developed based on the visual cortex in images and have been successful in image processing and speech recognition (92). The major characteristic of a CNN is its ability to generate convolution layers with multiple feature maps that capture the spatial information in training data. However, metagenomic data are represented by relative microbial taxonomic abundance profiles, where taxa can be placed in arbitrary orders. To empower CNNs in metagenomic phenotype prediction, it is important to provide structural input with certain distance metric among the microbial taxa. In our work, we constructed a phylogenetic tree, a natural structure representing the relationship among the microbial taxa in the profiles (90; 91). The tree is embedded in a 2D matrix after populating with the observed relative abundance of microbial taxa in each profile. In this way, the constructed matrices provide a better spatial and quantitative information in the metagenomic data to CNNs, compared to the vectors of relative microbial taxa abundances in an arbitrary order. Our analysis has revealed encouraging predictive ability of CNNs based on metagenomic data taken from different parts of body (90; 91).

**Challenges in Host Phenotype Prediction from Microbial Sequences**

Although DNNs provide incomparable ability of learning non-linear representation from the trainingset that can be used to predict host phenotype, the past results also raise the skepticism that DNNs may not be suitable learning models due to their requirement of excessive amount

of training data, which are impractical in the present metagenomic study. Furthermore, DNNs are often used as black-boxes, making it difficult to extract informative features from the learned models. Therefore, despite the success of DNNs in other biomedical applications (93), it is unclear whether they can outperform the existing models, such as RF, LASSO and SVMs, and whether they can learn a set of informative microbial taxa from metagenomics data.

Developing methods that predict the host phenotype from longitudinal microbiome samples comes with some challenges, e.g., variable sample collection times and uneven number of time points along the subjects' longitudinal study, especially when samples come from human subjects. Hence, using standard prediction methods such as Hidden Markov Models (HMMs) (94) and Auto Regressive (AR) models (95) may not be suitable in these cases.

# CHAPTER 3

# *WEVOTE*: WEIGHTED VOTING TAXONOMIC IDENTIFICATION METHOD OF MICROBIAL SEQUENCES

Previously published as:

- Metwally, A., Dai, Y., Finn, P., Perkins D. (2016) WEVOTE: Weighted Voting Taxonomic Identification Method of Microbial Sequences, *PLoS ONE*, 11(9), e0163527.
- Alaa, A., Metwally, A. (2018) Cloud-based Solution for Improving Usability and Interactivity of Metagenomic Ensemble Taxonomic Identification Methods, *IEEE EMBS Biomedical and Health Informatics*, (pp. 198-201). IEEE

## 3.1 Introduction

Taxonomic classification of sequence reads in a metagenomic dataset is a fundamental step in microbiome data analysis. The existing taxonomic identification methods of MGS data can be primarily classified into four categories: methods based on naive-similarity, methods based on analyzing sequence alignment results, methods based on sequence composition, such as $k$-mers, and marker-based methods. The naive-similarity-based methods rely on mapping each read to a reference database, such as the National Center for Biotechnology Information (NCBI) nucleotide database, and the taxonomic annotation of the best hit is assigned to the read if it passes a pre-set threshold. *Bowtie* (42), *BLASTN* (41), and its faster version *MegaBlast* (96) are the most commonly used algorithms in this category. Since the number of sequences in the database is enormous, these methods have a high probability of finding a match. Therefore, these types of methods usually achieve a higher level of sensitivity compared to other methods (46; 97). However, the

major drawbacks are the increased rate of false positive annotations and the long computational time. Although it has been shown that the taxonomic profile obtained from the naive-similarity-based methods produces a large number of false positives (46; 98), a vast array of researchers are still dependent on them because they do not want to sacrifice the high level of sensitivity to obtain fewer false positives annotations.

The category analyzing the results from sequence alignment includes *MEGAN* (50), and *PhymmBL* (97). These methods consist of a preprocessing step and a post-analysis step. In *MEGAN*, an algorithm involving the Lowest Common Ancestor (LCA) assigns each read an NCBI taxonomic identification number (si. taxon / pl. taxa) that reflects the level of conservation within the sequence. On the other hand, *PhymmBL* constructs a large number of Interpolated Markov Models (IMMs) using a *BLASTN* query against a reference database. It subsequently computes the scores which correspond to the probability of the generated IMMs matching a given sequence. Then it classifies a read using the clade label belonging to the organism whose IMM generated the best score. The methods in this category usually require additional computational time than those in the naive-similarity methods.

The marker-based methods utilize a curated collection of marker genes where each marker gene set is used to identify a unique group of clades. The fundamental difference between these methods and the naive-similarity methods is in the reference databases. Based on how the database of the marker genes is formed, this type of methods is classified into two main subcategories: (i) methods that depend on a universal single copy marker genes database such as *MetaPhyler* (53), *TIPP* (54), and *mOTU* (99), and (ii) methods that depend on a clade-specific marker genes database

such as *MetaPhlAn* (48; 100). These marker-based methods can achieve high accuracy if the reads come from genomes represented by the marker gene database. Otherwise, they only achieve a low-level of sensitivity. The running time varies depending on the statistical algorithm used in each method.

The $k$-mer-based methods use DNA composition as a characteristic to achieve taxonomic annotation. The key idea is to map the $k$-mers of each read to a database of $k$-mers, and then, each read is assigned a taxonomic annotation (46; 55; 57; 58; 101). For example, *Kraken* (46) uses an exact match to align the overlapped $k$-mers of the queries with a $k$-mer reference database, instead of an inexact match of the complete sequence used in the naive-similarity based methods. Because of the exact matching on short $k$-mers, many efficient data structures can be implemented for searching the $k$-mer database; thus the $k$-mer-based methods can be extremely fast. Compared to the naive-similarity methods, it was recently shown that at the genus level, $k$-mer-based methods could achieve a similar sensitivity but with higher precision (65). However, these methods are not robust to sequences that have a high sequencing error rate because they are based on exact matching to the reference database. This limitation is demonstrated in (46). It shows that *Kraken* has the lowest sensitivity compared to other methods when tested on the simBA-5 dataset.

In addition to our benchmarking, it has also revealed that different methods could generate variation in taxonomic output profiles for the same input dataset (65). Sample type, sequencing error, and read length are the main factors that cause variation. This inconsistency in the predicted taxonomic annotations presents a challenge to investigators in the selection of identification methods and the interpretation of annotations.

### 3.1.1 Problem Definition

In this work, we present a novel framework, *WEVOTE* (WEighted VOting Taxonomic idEntification), which takes advantage of three categories of the taxonomic identification methods; naive-similarity methods, $k$-mer-based methods, and marker-based methods. *WEVOTE* combines the high sensitivity of the naive similarity methods, the high precision of the $k$-mer-based methods, and the robustness of the marker-based methods to identify novel members of a marker family from novel genomes.

## 3.2 Methods

### 3.2.1 *WEVOTE* Core Algorithm

The core of *WEVOTE* is a weighting scheme organized as a taxonomic tree tallying the annotations from $N$ different taxonomic identification methods. As shown in Figure 2, the input to *WEVOTE* is the raw MGS reads of a microbiome sample. First, each of the $N$ identification methods independently assigns a taxon for each read. If any method fails to classify the read based on the given threshold, the *WEVOTE* preprocessing phase assigns 0 as a taxon, indicating that the read is unclassified by the corresponding method. Then, *WEVOTE* identifies the taxonomic relationship of the $N$ taxa per read based on the pre-configured taxonomy tree structure and casts a vote to the final taxon, which may be a common ancestor of the $N$ taxa. Although the current version of our method only includes five methods, the voting scheme in our framework is flexible and allows for the inclusion or removal of different methods.

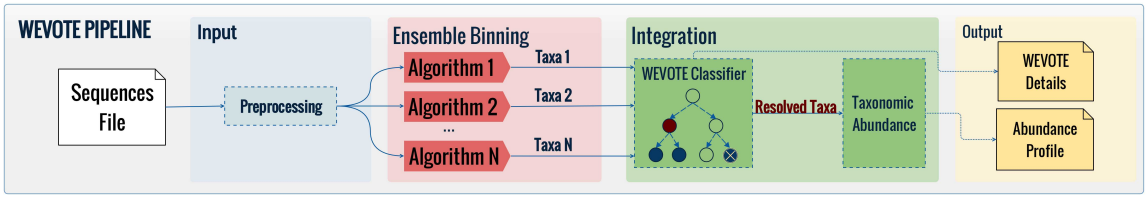



Figure 2: Schematic diagram of the *WEVOTE* framework. The input to the *WEVOTE* is the raw reads of the sample. First, each of the identification methods independently assigns a taxon to each read. Then, *WEVOTE* identifies the taxonomic relationship of the $N$ taxa based on the pre-configured taxonomy tree structure and determines the final taxon assigned to each read (102).

*WEVOTE* utilizes a simplified version of the NCBI taxonomy tree as a backbone for its decision algorithm. This resolved phylogeny tree only contains the nodes that have a taxon corresponding to one of the standard taxonomic levels (Super-kingdom, Phylum, Class, Order, Family, Genus, and Species). This backbone structure facilitates and accelerates the choice of a consensus taxon based on the taxonomic annotations received from each identification method. The decision scheme in *WEVOTE* is shown in Algorithm 1. Here, *N* denotes the number of methods used in the *WEVOTE* pipeline; *C* the number of methods that can classify the read at any taxonomic level, i.e., taxon $\neq 0$; and *A* the number of methods that support the *WEVOTE* decision. The relationship $N \geqslant C \geqslant A$ always holds.

**Algorithm 1** The *WEVOTE* Decision Scheme

```
1:  procedure WEVOTE (N taxa for each read)
2:      for each (Read ∈ sequence file) do
3:          if (C == 0) then
4:              Read.Taxon = 0
5:              Read.DecisionScore = 1
6:              Read.NumSupportedTools = N
7:          else if (C ⩾ 1) then
8:              build a WeightedTree of the reported taxa
9:              Threshold = floor(C/2)
10:             MaxWeight = 0
11:             MaxNode = 0
12:             for each (Node ∈ WeightedTree and weight(Node) > Threshold) do
13:                 if (rootToTaxon(Node) > MaxWeight) then
14:                     MaxWeight=rootToTaxon(Node)
15:                     MaxTaxon=Node
16:                 else if (rootToTaxon(Node) == MaxWeight) then
17:                     MaxTaxon=LCA(Node, MaxTaxon)
18:                 end if
19:             end for
20:             Read.Taxon = MaxTaxon
21:             Read.NumSupportedTools = weight(Read.Taxon)
22:             if (A == C) then
23:                 Read.DecisionScore = A/N
24:             else
25:                 Read.DecisionScore = (A/N) − (1/(m ∗ N))
26:             end if
27:         end if
28:     end for
29: end procedure
```

In the case that no single tool can classify the read, *WEVOTE* will accordingly fail to classify the read and give it a taxon 0 and score of 1. Otherwise, *WEVOTE* starts by building a weighted tree for each read from the taxa reported by individual methods. The weighted tree is a tree that

comprises the nodes of the identified taxa along with their ancestors' taxa including the root. The weight of any node on the weighted tree represents the number of methods that support the identification of this particular node. Next, *WEVOTE* annotates the read with the taxon of the node that has the highest weight from the root to that node (RootToTaxon), with the additional condition that the node itself has more weight than the *WEVOTE* threshold. This threshold can be set as half of the number of methods that classify a read as shown in Figure 3. In the case where more than one node satisfies the *WEVOTE* condition, then the LCA of these nodes will be assigned as the *WEVOTE* decision. For each classified read, a score is also assigned to reflect the confidence of *WEVOTE* decision. The scoring scheme works based on Equation 3.1.

$$\text{Score} = \begin{cases} \frac{A}{N} & \text{If } C = A, \\ \frac{A}{N} - \frac{1}{2N} & \text{otherwise; } A < C \end{cases} \tag{3.1}$$

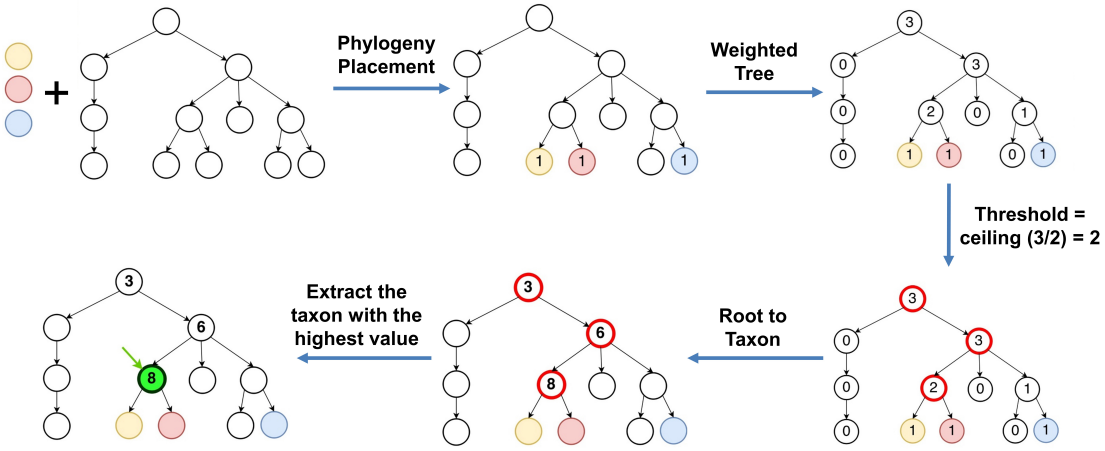



Figure 3: Illustration of *WEVOTE* algorithm.

The choice of the constant $m$ depends on how strongly one elects to penalize the disagreement among individual methods that classify the read but do not agree with the *WEVOTE* decision. A small value of $m$ leads to a small *WEVOTE* score, implying more penalty is placed on the *WEVOTE* decision score, and vice versa. This scoring scheme makes the score satisfy the condition of $\frac{A-1}{N} < score < \frac{A}{N}$. Although the score does not affect the *WEVOTE* decision, it would be useful if the user is interested in assessing the confidence of the taxon assignment made by *WEVOTE*. The default value of $m$ is 2. We have chosen this value because it gives a score exactly in the middle of $\frac{A-1}{N}$ and $\frac{A}{N}$. As $m$ increases, the score skews towards the $\frac{A}{N}$ side. In order to demonstrate the decision and scoring schemes described in the *WEVOTE* algorithm, the case scenarios of *WEVOTE* for $N = 3$ are shown in Figure 4.

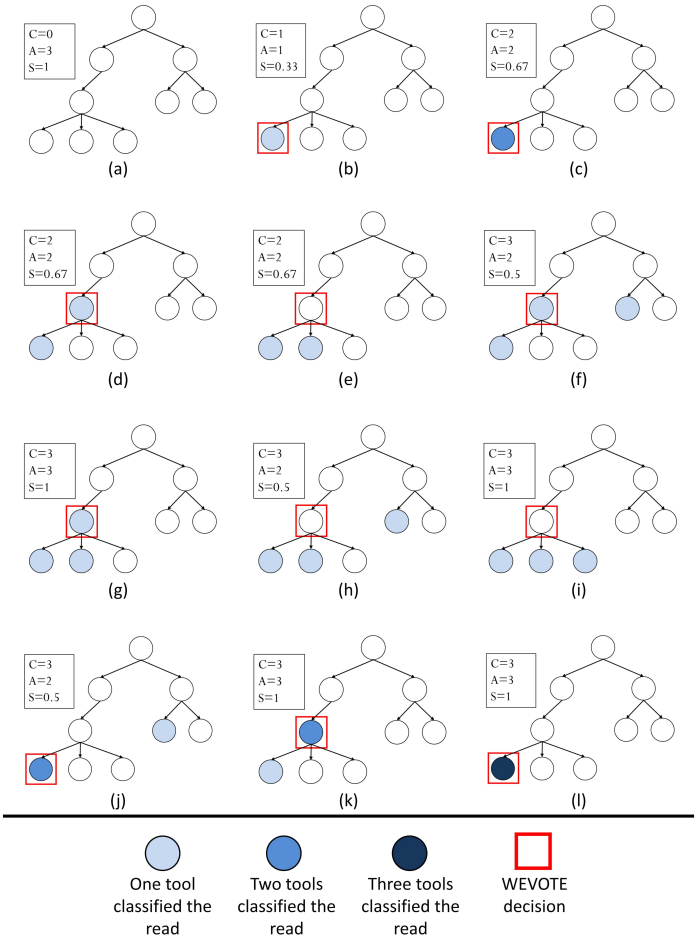



Figure 4: *WEVOTE* case scenarios using three tools. *C* denotes the # tools able to classify the read, *A* represents the # of tools that support *WEVOTE* Decision, and *S* represents the *WEVOTE* score. Scenarios are shown for (a) None of the three tools classified the read; (b) Only one tool classified the read; (c) Two tools classified the read with the same taxon; (d, e) Two tools classified the read with two different taxa; (f-i) Three tools classified the read with three different taxa; (j, k) Three tools classified the read, two taxa are identical, and the other is different; (l) Three tools identified the read with the same taxon (45).

**Methods Used in the Current *WEVOTE* Implementation**

In our current implementation of *WEVOTE*, we used *BLASTN* (41) to represent the naive-similarity-based methods, *Kraken* (46) and *CLARK* (57) as the identification methods representing the $k$-mer methods, and *TIPP* (54) and *MetaPhlAn* (48) representing the marker-based methods. The five methods were chosen since they are widely used and represent the three major categories of taxonomic identification methods. We selected *BLASTN* over *MegaBlast* because of its greater sensitivity. The primary reason for the increased sensitivity in *BLASTN* is the use of a shorter word size as a search seed. Thus, *BLASTN* is better than *MegaBlast* in finding alignments for sequences that have a sequencing error which occurs after a short length of matched bases (i.e., the initial exact match is shorter).

*Kraken* assigns taxonomic annotations to the reads by splitting each sequence into overlapping $k$-mers (46). Each $k$-mer is mapped to a pre-computed database where each node in the database is the LCA taxon of all genomes that contain that $k$-mer. For each read, a classification tree is computed by obtaining all the taxa associated with the $k$-mers in that read. The number of $k$-mers mapped to each node in the classification tree is assigned as a weight for this node. The node that has the highest sum of weights from the root is used to classify the read. *Kraken* is an ultra-fast and highly precise algorithm for reads involving a low rate of sequencing error. *CLARK* is a recently released tool that is very similar to *Kraken* and also based on $k$-mers. It is reported to be faster and more accurate than *Kraken* at the genus/species level (57). The fundamental difference between *Kraken* and *CLARK* is their backbone $k$-mers database. *Kraken* has

only one database that can serve for the classification of metagenomic reads at any taxonomic level. If more than one genome shares the same $k$-mer, *Kraken* assigns this $k$-mer to their LCA taxon. *CLARK*, on the other hand, builds an index for each taxonomic level at which the user wishes to classify. Each level's index has only the discriminative $k$-mers that distinguish its taxa from others.

*TIPP* (Taxonomic Identification and Phylogenetic Profiling) is considered a state-of-the-art method based on a set of marker genes. It uses a customized database of 30 marker genes (103) which are mostly universal and single-copy genes. First, it performs multiple sequence alignment of each marker gene set, then builds a phylogeny tree for each marker gene and constructs a resolved taxonomy tree of these marker genes. Then, it uses *SATe* (104) to decompose the tree of each marker gene to many sub-trees. Subsequently, *TIPP* uses *HMMER* software (105) to build a Hidden Markov Model (HMM) for each of the sub-trees. For each query read, *TIPP* uses *HMMER* again to align the query to the HMMs. Then, *TIPP* uses the alignments to the HMM that have an alignment score and statistical support greater than a group of pre-set values, and places them on the precomputed taxonomic tree using *pplacer* (106) to assign taxonomy to the query. It has been shown that *TIPP* can precisely identify reads containing high sequencing error or novel members of a marker family from novel genomes (54). The other method chosen for this category in our implementation is *MetaPhlAn. MetaPhlAn* has a set of clade-specific marker genes. The marker set was built from the genomes available from the Integrated Microbial Genomes (IMG). For a given read, *MetaPhlAn* compares the read against the precomputed marker set using *BLASTN* searches in order to provide clade abundances for one or more sequenced metagenomes.

### 3.2.2 *WEVOTE-web*: Cloud-based Solution for Improving Usability and Interactivity of *WEVOTE*

Although *WEVOTE* has higher sensitivity and precision than individual methods, several difficulties are imposed as a price to this gain. First, the issue of usability persists, at least for researchers who lack decent computational skills, unless these methods are introduced through graphical user interfaces. Second, installing *WEVOTE* with the corresponding dependencies requires large storage budget and careful dependency management. In addition, much of *WEVOTE* configuration and sorting the reference databases requires caution and minimal scripting background. Third, the execution of specific algorithms requires an unattainable memory space on regular computers. A workaround is to exclude expensive algorithms from the pipeline, however, this will also reduce quality.

In this project, we address the aforementioned obstructions by porting the *WEVOTE* framework into the cloud for attaining several objectives: (a) improve usability by implementing a modular web application, (b) interactive visualization using the rich Javascript visualization libraries, (c) unrestricted deployment of the expensive computational options by leveraging the robust pay-as-you-go infrastructure on the cloud. The modular design of the cloud implementation allowed more use cases of *WEVOTE*.

In our cloud framework, we developed three use cases of the the *WEVOTE* algorithm; (a) apply the whole *WEVOTE* pipeline: the user uploads a sequences file then selects the ensemble methods from the currently available methods, (b) use *WEVOTE* algorithm on an ensemble file:

the user intercepts the pipeline at the integration step and uploads a file containing classified sequences from multiple methods in order to employ the *WEVOTE* classifier followed by generation of the community profile, (c) taxonomic profiling from a classified taxa file: the user intecepts the pipeline from the profiling step and the system generates the community profile from an input of classified sequences. The *WEVOTE* cloud framework consists of two modules; the web module and the visualization module.

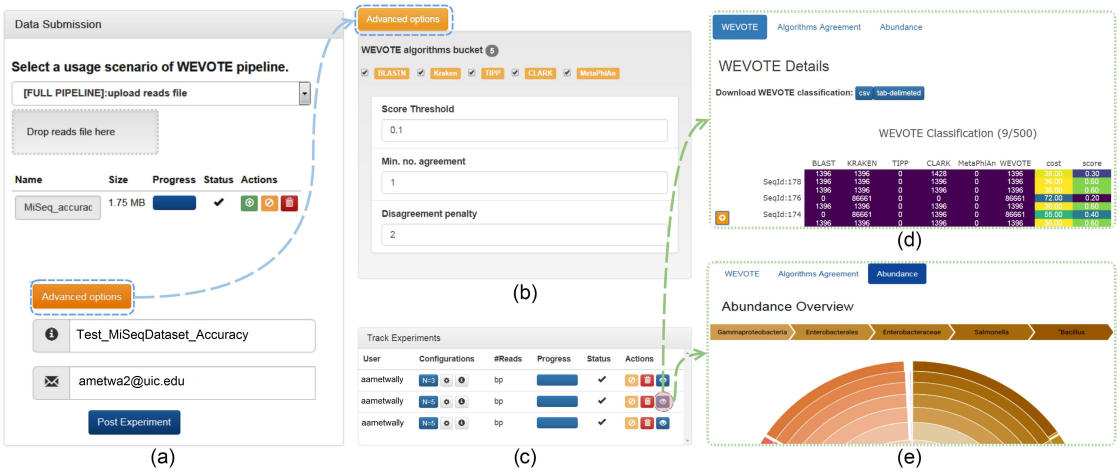



Figure 5: User interface of the cloud implementation of *WEVOTE* pipeline. The panel in (a) is an entry point to the pipeline where the user selects the usage scenario and accordingly uploads the query file. The user may override the configuration by using the secondary panel in (b). Finally, the user is recommended to add a description as a tag to the experiment and a reference e-mail, so the user is notified by e-mail when the results are ready for visualization. Alternatively, the user may track the pipeline progress in the panel in (c). This panel also includes all the meta information corresponding to an experiment (e.g., parameters, incorporated algorithms, and description) so the user can explore and compare previous experiments without confusion. Upon completion, the results are ready for visualization in a dedicated page for different analyses like in (d) and (e). Also, the user may choose to download the results, including the intermediate results as in (d) (102).

**Web Module**

Based on the highlighted usability issues, the user interface was developed with multiple objectives in mind: (a) ease of using the pipeline through intuitive panels with minimal parameters setting, (b) controllability over the pipeline, and (c) user-centered design, so the user is able to keep track of previous experiments. Figure 5 illustrates the current user interface elements in the web application. Furthermore, the web module carries out the communication between the client and the computational web service. It also stores user profiles to keep track of previous experiments. This module is implemented using *TypeScript* language based on *Express.js* as web framework and *MongoDB* as a database driver (107).

**Visualization Module**

In this module, several intuitive and interactive visuals are generated to summarize the taxonomic classification results (Figure 6). Here is a summary of different outputs:


Figure 6: Visualization options on the client-side. The table in (a) lists the intermediate classification results of *WEVOTE* pipeline. The cost column represents the Manhattan distance. The distance is defined as the shortest path between the voting node and the resolved node in the taxonomy tree. The last column is the classification score. (b) The abundance profile is listed as a table. The numerical columns like "cost" and "score" in (a) and "count" in (b) are color-encoded. (d) The interactive radial treemap diagram with the embedded taxon lineage at the top and the abundance percentage inside the diagram center. For all visualization entities in (a), (b), and (d), when the user clicks on a label or a cell corresponds to a taxon, a new tab in the browser will access the taxon page at the NCBI Taxonomy Browser as in (c). In addition, detailed info about the cell is shown upon the mouse hovering on that cell (102).

(a) Tabular list: The tabular list is employed for taxonomic binning and abundance profile results (Figure 6). In the classification results table, two descriptive values are included: (a) Score: each classified read is associated with a score computed based on Equation 3.1, and (b) Cost: a cost value is computed for a classified read$_a$ as a Manhattan distance combining the lengths of shortest paths from the resolved node to the other votes by Equation 3.2.

$$\text{Cost}_a = \sum_{i=1}^{N} ||\textsc{ShortestPath}(\text{WEVOTE}_a, \text{Taxon}_{ai})|| \tag{3.2}$$

Numerical information, e.g., score and cost values in the classification results, are color-coded for seamless data exploration. Moreover, color scale is not global for the whole table, but specific for each column; each cell color is represented by an intensity normalized along the column and log-scaled for better color distribution. However, tables remain an inefficient visual analytic solution for visualizing scalable data as they occupy much space.

(b) Radial treemap: An interactive radial treemap is developed in the visualization module in our implementation to visualize taxonomic abundance profile. A hierarchical clustering is constructed for the taxa in the abundance profile. Each taxon accumulates its abundance value along its ancestry.

(c) Venn diagram: A further precision analysis on the incorporated taxonomic classification methods can be visually assisted using interactive *Venn diagram*. The intersection areas among sets depict the agreement among the taxonomic binning algorithms.

## 3.3 Experiments and Results

### 3.3.1 *WEVOTE* Benchmarking

Simulated datasets have been used in the evaluation of various taxonomic identification methods. In our assessment, we selected fourteen simulated datasets as shown in Table I. Our choice was based on the ability of these datasets to provide the true assignment for each read rather than the true relative abundance at each taxonomic level. This information allows for the evaluation of *WEVOTE* based on various metrics in addition to the assessment of relative abundance.

The first three datasets were used in the evaluation of *Kraken* (46). The HiSeq and MiSeq datasets are simulated from sequences obtained from non-simulated microbial projects but were sequenced using two different platforms, i.e., Illumina HiSeq™ and Illumina MiSeq™. The simBA5 is a simulated dataset with a higher percentage of error to mimic increased sequencing errors. Hence, it can be used to measure the ability of each tool to handle actual sequencing data. The simHC20 dataset was used to benchmark *CLARK* (57) and it contains 20 subsets of long Sanger reads from various known microbial genomes. The other ten datasets were used in *MetaPhlAn* (48) evaluations. HC1 and HC2 consist of reads from high-complexity, evenly distributed metagenomes that contain 100 genomes, and LC1–LC8 consist of reads from low-complexity, log-normally distributed metagenomes that contain 25 genomes. The reads from all ten *MetaPhlAn* datasets were sampled from KEGG v54 (108) with a length of 100 bp and an error model similar to real Illumina reads.

TABLE I: *WEVOTE* benchmarking datasets.

| Source | Dataset | # reads | length (bp) | # genomes |
|---|---|---|---|---|
| **Kraken** | **HiSeq** | 10,000 | 92 | 10 |
| | **MiSeq** | 10,000 | 156 | 10 |
| | **simBA5** | 10,000 | 100 | 1,967 |
| **CLARK** | **simHC20** | 10,000 | 951 | 20 |
| **MetaPhlAn** | **HC1** | 999,998 | 88 | 100 |
| | **HC2** | 999,991 | 88 | 100 |
| | **LC1** | 249,995 | 88 | 25 |
| | **LC2** | 250,000 | 88 | 25 |
| | **LC3** | 250,000 | 88 | 25 |
| | **LC4** | 249,999 | 88 | 25 |
| | **LC5** | 249,999 | 88 | 25 |
| | **LC6** | 250,002 | 88 | 25 |
| | **LC7** | 250,000 | 88 | 25 |
| | **LC8** | 250,000 | 88 | 25 |

Our benchmarking was performed with two variants of *WEVOTE*: (i) *WEVOTE* ($N = 3$) including *BLASTN*, *TIPP* and *Kraken*; and (ii) *WEVOTE* ($N = 5$) including *BLASTN*, *TIPP*, *MetaPhlAn*,

*Kraken*, and *CLARK*. As described previously, *BLASTN* represents the naive-similarity method; *TIPP* and *MetaPhlAn* belong to the category of the marker-based methods; and *Kraken* and *CLARK* belong to the category of the $k$-mer-based methods. The default parameter values were set for the individual method, and the score penalty in *WEVOTE* was set at $m = 2$. Regarding *WEVOTE*, we reported all results in which at least one method supported the *WEVOTE* decision. With this approach, we can evaluate the accuracy of *WEVOTE* at the highest classification rate of the reads. By increasing the threshold, we can generate more precise results as shown later.

We first looked at how accurately each method annotates individual reads at each taxonomic level using sensitivity and precision metrics, which are defined in (Equation 3.3) and (Equation 3.4), respectively. For each level $l$ in a simulated dataset:

$$\text{Sensitivity}_{(l)} = \frac{TP_l}{P_l} \tag{3.3}$$

$$\text{Precision}_{(l)} = \frac{TP_l}{TP_l + FP_l} \tag{3.4}$$

where $P_l$ denotes the number of reads annotated with some taxon at level $l$ in the original dataset; $TP_l$ the number of reads correctly annotated at level $l$; and $FP_l$ the number of reads incorrectly annotated at level $l$.

It could be inappropriate to compare the sensitivity of all the methods used in *WEVOTE*, since the marker-based methods are primarily designed to calculate the microbial abundance of the

sample based on the annotation of the reads that come from genes represented by the marker gene database. Based on this consideration, Figure 7 (I) shows the sensitivity and precision of *Kraken*, *CLARK*, *BLASTN*, and *WEVOTE*; while in Figure 7 (II), we show the precision of *TIPP* and *MetaPhlAn* separately. It is observed from Figure 7 that *WEVOTE* achieves the highest level of precision and a level of sensitivity that is second only to *BLASTN* at the species level. At all other taxonomic levels, *WEVOTE* outperforms all the other individual methods in terms of sensitivity and precision in most datasets (S1 Table at (45)). Note that the reason for the lower precision with $N = 5$ is because the results were reported when the minimum number of methods supported the *WEVOTE* decision was set at 1. If a higher level of precision is required, then the *WEVOTE* reporting threshold should be set at $N/2$ as explained later.

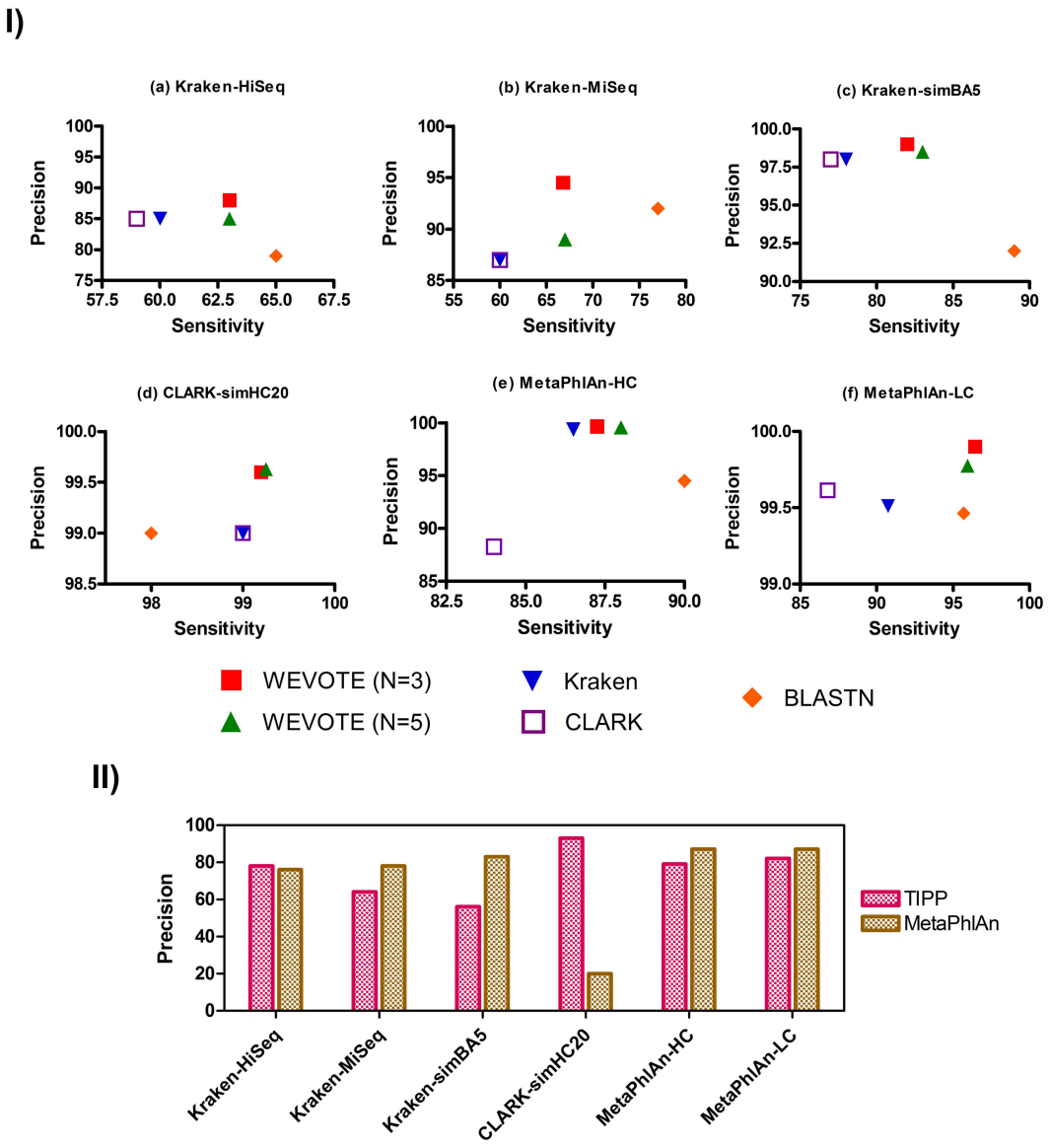



Figure 7: Sensitivity and precision at the species levels. sub-panel (I) shows the sensitivity and precision of methods developed to identify every read; *Kraken*, *CLARK*, *BLASTN*, and *WEVOTE*. sub-panel (II) shows the precision of marker-based methods; *TIPP* and *MetaPhlAn*. The MetaPhlAn-HC and MetaPhlAn-LC datasets are the average of two HC and eight LC datasets, respectively (45).

In addition, we calculated the Hellinger distance (109) ($H_l$) between a sample's metagenomic abundance profile generated by each method and its true abundance profile at each taxonomic level $l$. The Hellinger distance measures the deviation of the predicted profile from the true profile. It is calculated as shown in (Equation 3.5). Here, $C_l$ is the union of all taxa that are in the true and predicted profiles at each taxonomic level $l$. For each taxon $x$ at level $l$, $P_x$ is the predicted relative abundance and $T_x$ is the true relative abundance at taxonomic level $l$. The $\sqrt{2}$ is added to the denominator to keep $0 \leq H_l \leq 1$.

$$H_l = \frac{\sqrt{\sum_{x \in C_l} (\sqrt{P_x} - \sqrt{T_x})^2}}{\sqrt{2}} \tag{3.5}$$

The calculation of the relative abundance (RA) differs among the methods. For methods that are developed to identify every genomic read, such as *BLASTN*, *Kraken*, and *CLARK*, the relative abundance is calculated as shown in (Equation 3.6). As mentioned before, *TIPP* and *MetaPhlAn* are not designed to identify every read. They build metagenomics abundance profile of the sample based on the annotation of the reads that come from genes represented by the marker gene database. In this case, the relative abundance of a taxon *x* is calculated using (Equation 3.7). For *WEVOTE*, we used (Equation 3.6) to calculate the RA. These two forms of relative abundance calculation are implemented in *WEVOTE*. It is the user option to select which method to use. However, the genomic-based method (Equation 3.6) is the default setting.

$$RA_{genomic-based}(x) = \frac{n_x}{n} \tag{3.6}$$

$$RA_{marker-based}(x) = \frac{n_x}{n_c} \tag{3.7}$$

Where $n_x$ is the total number of reads classified at taxon $x$, $n$ the total number of reads, and $n_c$ the total number of classified reads.

As the Hellinger distance represents an error distance, a small value is always preferable. Particularly, $H = 0$ means that the predicted profile is exactly the same as the true profile; while $H = 1$ means that the predicted profile is completely different from the true profile. Figure 8 shows the Hellinger distance between the true relative abundance profile and the profiles generated by all methods at different taxonomic levels (Table S2 at (45)). For all the benchmarking datasets, *WEVOTE*, particularly when $N = 3$, always has the smallest Hellinger distance among all other individual identification methods across all taxonomic levels. Although the Hellinger distance is marginally different for *WEVOTE* and *BLASTN*, the interpretation is quite different. The error that originates from *BLASTN* is due to the false positive annotations while the error that originates from *WEVOTE* is due to the lack of support in annotating the read at the corresponding level. *TIPP* and *MetaPhlAn* have higher Hellinger distance than other methods used in *WEVOTE*. This is mainly because few taxa in the datasets are predicted in low rate by them, i.e., $P_x$ being near zero for few taxa. This has led to the accumulation in the Hellinger distance. One of the reasons for the inability to predict these taxa may be because the current marker gene databases used in *TIPP* and *MetaPhlAn* do not contain sufficient markers of the genomes represented in the simulated datasets.

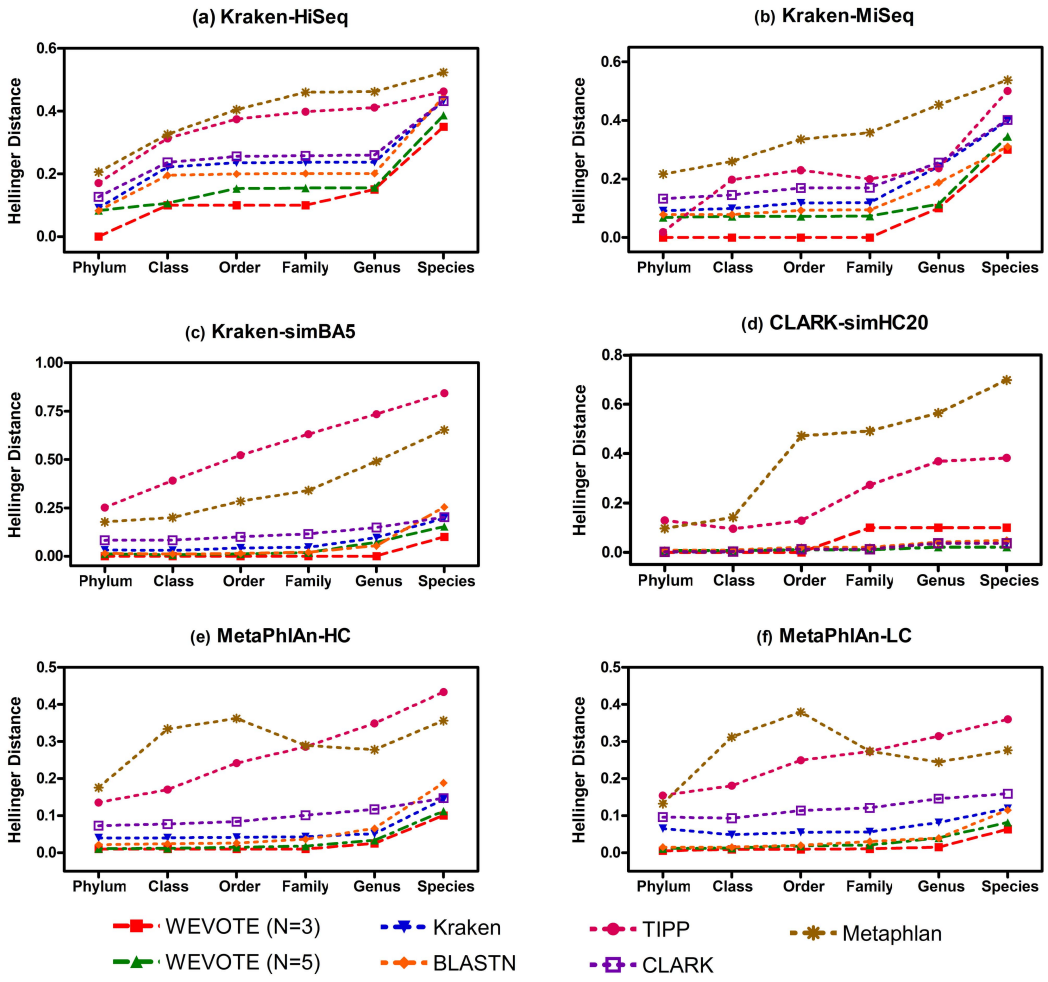



Figure 8: Evaluation of Hellinger distance. The deviation between the predicted and the true abundance profile was measured in terms of the Hellinger distance for each method at different taxonomic levels. Results are shown for: (a) Kraken-HiSeq dataset; (b) Kraken-MiSeq dataset; (c) Kraken-simBA5 dataset; (d) CLARK-simHC20; (e) MetaPhlAn-HC and (f) MetaPhlAn-LC. The lower the error, the more precise the corresponding method is at the corresponding taxonomic level. $H = 0$ means that the predicted relative abundance profile is exactly the same as the true profile; while $H = 1$ means that the predicted profile is completely different from the true profile (45).

Lastly, we examined the details of various case scenarios that were encountered in the evaluation of the two *WEVOTE* variants, i.e., $N = 3$ and $N = 5$. The plots in Figure 9 show the percentages of annotations in which the individual methods support the *WEVOTE* decision for all the datasets. It can be observed that the majority of *WEVOTE* annotations are determined based on more than $N/2$ agreements; 2 in the case of $N = 3$ and 3 in the case of $N = 5$. For only a small portion of each dataset, all the used methods agreed on the *WEVOTE* decision. An interesting observation is that a very small portion of all the classified reads by *WEVOTE* are in agreement with one method when $N = 3$, or either 1 or 2 methods when $N = 5$. Therefore, if we set a threshold on *WEVOTE* to report the taxon at which more than half the methods are in agreement with the *WEVOTE* decision, then the precision of *WEVOTE* would increase, and its sensitivity will only be marginally decreased as demonstrated in Figure 10. We have chosen Kraken-HiSeq and Kraken-MiSeq datasets for this investigation because they had low precision among all the used taxonomic identification methods (Figure 7).

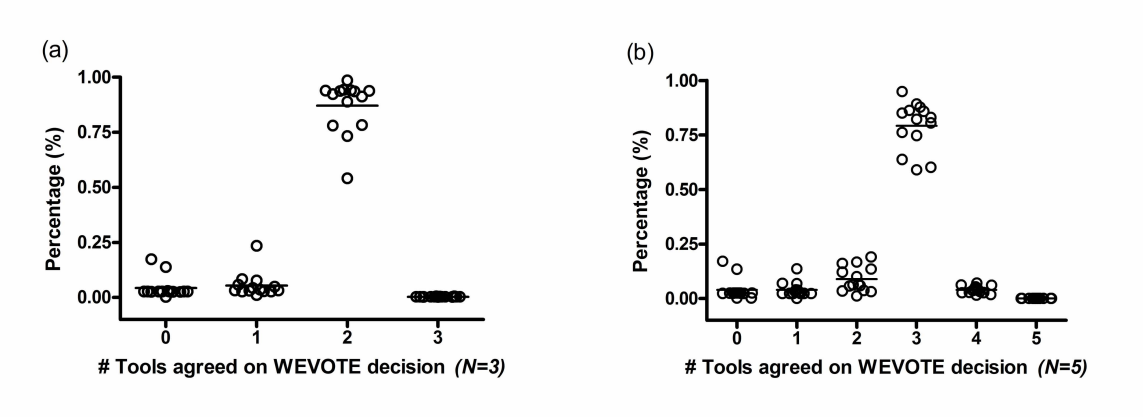



Figure 9: The percentage distribution of the number of individual methods that support the *WEVOTE* decision for the 14 datasets. Here, 0 means that the read was not classified by any methods, 1 means that one method supports the *WEVOTE* assigned taxon for the read, and so on. A=3 in the case of (a) means that all the 3 methods support *WEVOTE* on its assigned taxon for the corresponding read, A=5 in case of (b) means that all the used 5 methods support *WEVOTE* on its assigned taxon for the corresponding read (45).

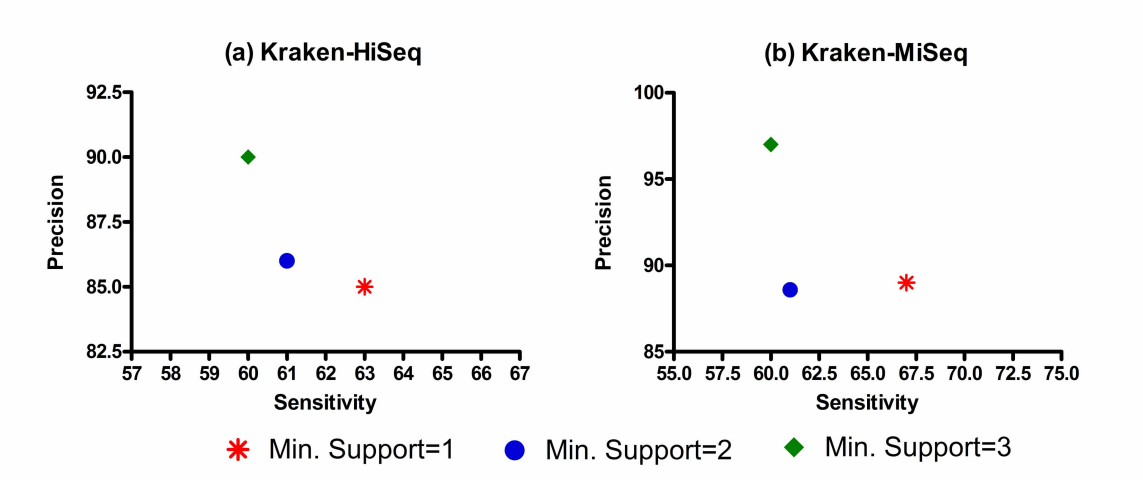



Figure 10: The sensitivity and precision at the species level for the *WEVOTE* (N=5) using different thresholds for the minimum number of methods that support the *WEVOTE* decision. (a) Kraken-HiSeq dataset; and (b) Kraken-MiSeq dataset (45).

### 3.3.2 Computational Resources and Running Performance

All the experiments were performed on the UIC computer cluster (EXTREME) at the University of Illinois at Chicago. To benchmark *WEVOTE*, we used one node with 16 cores (Intel Xeon E5-2670 @ 2.60 GHz, cache size of 20 MB, and 128 GB RAM). Since the *WEVOTE* core algorithm and all the individual methods are parallelizable, we utilized 16 threads for all experiments conducted in this work. Due to the high requirement on the memory for constructing *Kraken* and *CLARK* databases, we used the Highmem node on EXTREME which has specification of 1TB RAM. In order to achieve the maximum performance from *Kraken* and *CLARK*, we used the default versions of the two methods, which require at least 80 GB of RAM. Therefore, if there is only a limited amount of memory available, users can run these methods using their mini versions, i.e., *MiniKraken* and *CLARK-l*, which only require 4 GB of RAM. In this case, the output could be 11%-25% less sensitive, but it will still preserve a high level of precision. The *WEVOTE* algorithm is particularly useful in this case because it can exploit the high precision level of *Kraken* and *CLARK* without using large memory machines and compensate the sensitivity by using *BLASTN*.

Table II shows the running time for each method per dataset. For HC and LC classes of datasets, the running time is presented as the average over the datasets in each class. *Kraken* and *CLARK* finished in less than 3 minutes for any individual dataset. For *BLASTN*, the most time-consuming method that is currently implemented in the *WEVOTE* pipeline, its running time is proportional to the number of reads and the read length in a dataset. The total time of the entire *WEVOTE* pipeline is the summation of the running times of the individual methods and the time

needed to run the *WEVOTE* core algorithm. The *WEVOTE* core algorithm was finished execution in less than 33 seconds for any individual dataset regardless $N = 3$ or $N = 5$. The *WEVOTE* core algorithm is mainly affected by the number of the used methods, and more specifically, the number of methods that identified taxa for the reads. Because the running time of *WEVOTE* pipeline is primarily dominated by the time required by *BLASTN*, the pipeline running time can be reduced if many cores are used to execute *BLASTN*.

TABLE II: *WEVOTE* running time measured in minutes.

| **Simulated Dataset** | ***Kraken*** | ***BLASTN*** | ***TIPP*** | ***CLARK*** | ***MetaPhlAn*** | ***WEVOTE* Pipeline [N=5]** |
|---|---|---|---|---|---|---|
| **HiSeq** | 1 | 2 | 4 | 1 | 1 | 10 |
| **MiSeq** | 1 | 8 | 4 | 1 | 1 | 16 |
| **simBA5** | 1 | 7 | 3 | 1 | 1 | 14 |
| **simHC20** | 1 | 9 | 5 | 1 | 1 | 18 |
| **HC (sd)** | 2 (0.0) | 30 (1.4) | 14 (0.0) | 3 (0.0) | 2 (0.0) | 53 (1.4) |
| **LC (sd)** | 1 (0.0) | 9 (2.9) | 8 (0.5) | 2 (0.0) | 1 (0.5) | 22 (3.5) |

To analyze the time-cost trade-off on the AWS for *WEVOTE-web*, we used Kraken_MiSeq dataset as an experiment for taxonomic classification task. We tested two different AWS machine types; t2.large and t2.2xlarge. The reported time and the total cost according to the reserved instances are listed in Table III.

TABLE III: Time consumed and total cost on different machines specifications. Usage scenario: *full pipeline*; utilized algorithms: *BLASTN*, *Kraken*, *CLARK*, *TIPP*, and *MetaPhlAn*.

| **Instance type** | **Experiment time (minutes)** | Total cost ($) |
|---|---|---|
| t2.large | 26 | 0.10 |
| t2.2xlarge | 14 | 0.38 |

### 3.4 Conclusion

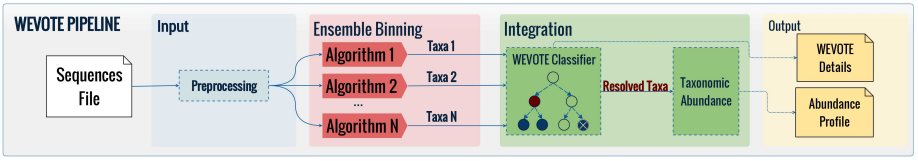
We developed *WEVOTE* (WEighted VOting Taxonomic idEntification), a phylogenetic-based ensemble method that classifies metagenome shotgun sequencing DNA reads based on an ensemble of existing methods using k-mer-based, marker-based, and naive-similarity based approaches. The performance evaluation based on the fourteen simulated microbiome datasets consistently demonstrates that *WEVOTE* achieves a high level of sensitivity and precision compared to the

individual methods across different taxonomic levels. Moreover, the score assigned to the taxon for each read indicates the confidence level of the assignment. This information is especially useful for the assessment of false positive annotations at a particular taxonomic level. The classification score given by *WEVOTE* can be used for any downstream analysis that requires the high confidence of the annotated sequences. Moreover, we introduced a cloud-based solution to address common usability issues in the *WEVOTE* framework. In addition, an interactive visual analytics tool was developed to ease the interpretation of the classification results. We have demonstrated three different use cases of the pipeline that, in turn, reflect the significance of our modular design. *WEVOTE* and *WEVOTE-web* are publicly available on `https://github.com/aametwally/WEVOTE` and `https://github.com/aametwally/WEVOTE-web`, respectively.

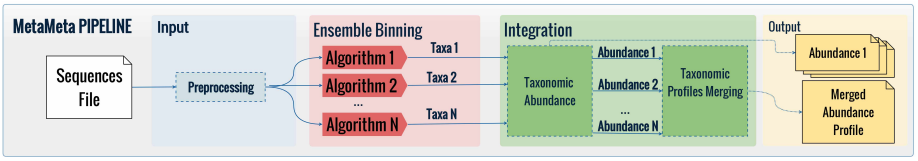
After *WEVOTE* has been developed and showed a spectacular performance in taxonomic identification of microbial sequences, two other methods were developed; *MetaMeta* (110) and *Direct Majority Voting* (64).

*MetaMeta* employs multiple taxonomic binning algorithms producing individual abundance profiles from each algorithm and a final abundance profile (110) (Figure 11.b). The primary difference between *MetaMeta* and *WEVOTE* is that *MetaMeta* merges information at a very late stage. It performs cascading statistical operations to suppress outliers effect in the taxonomic profile. However, It does not guarantee to suppress false positive classifications effect. Moreover, it may prune out low abundant taxa as false positive. Furthermore, *MetaMeta* requires the user to configure more preference parameters that may reduce its usability. Similar to *WEVOTE*, *MetaMeta's*

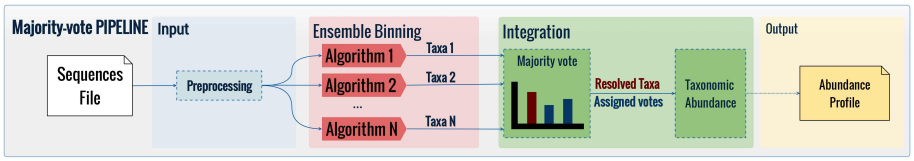
computational resources and running times depend on the utilized algorithms. On the other hand, *Direct Majority Voting* uses multiple combinations of taxonomic binning algorithms (64) by integrating votes (Figure 11.c).




(a)




(b)




(c)

Figure 11: Comparative view three ensemble taxonomic classification methods. The primary variations among these methods are in the integration step (102).

# CHAPTER 4

# *METALONDA*: IDENTIFYING TIME INTERVALS OF DIFFERENTIALLY ABUNDANT FEATURES IN METAGENOMIC LONGITUDINAL STUDIES

Previously published as:

- Metwally, A., Yang, J., Ascoli, C., Dai, Y., Finn, P., Perkins, D. (2018) MetaLonDA: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies, *Microbiome*, 6(1), 32.
- Metwally, A., Finn, P., Dai, Y., Perkins, D. (2017) Detection of Differential Abundance Intervals in Longitudinal Metagenomic Data Using Negative Binomial Smoothing Spline ANOVA, *In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 295-304). ACM.

## 4.1 Introduction

One of the objectives of the microbiome studies is to determine whether there is a particular microbial signature (e.g., taxa, genes, or pathways) associated with a particular disease state and/or disease outcome. These biomarkers can play an important role in the development of preventative and therapeutic strategies. A major challenge in microbiome studies is the variability in microbial taxa among subjects, in addition to variability due to disease influences. A powerful strategy to address this challenge is the analysis of time series data in which the time intervals associated with temporal effects are identified. Modeling metagenomic data for disease-association studies is an active area of research. The standard parametric models may reduce variance if the data follows the corresponding parametric distribution, but the models may be substantially bi-

ased if the data does not support that distribution. On the other hand, non-parametric models do not assume any prior distribution of the data and thus are not biased towards any distribution, but these models may suffer from a large model variance (69).

For longitudinal data, two types of differential abundance analysis are widely utilized: (a) Treat data from each time point independently and detect features that have differential abundance between the phenotypes at individual time points (74), and (b) Identify features that have differential abundance during the time-course within a phenotype (75; 76). Longitudinal analysis is usually challenged by variability in longitudinal sample collections, including inconsistencies in the number of subjects per phenotype, number of samples per subject, and sample collection at inconsistent time points. These inconsistencies increase with the level of difficulty with which samples are obtained from the subjects. For example, in humans, the variability decreases in samples collected non-invasively (e.g., stool and urine samples) but increases in the invasive procedures (e.g., bronchoalveolar lavage (BAL) samples which are extracted from the lung by bronchoscopy).

One solution to address this variability is to bin samples into a certain number of windows between the start and end times of the study course by selecting the nearest sample in time for each bin (30), then compare the microbial feature's relative abundance or diversity indices (82; 83) between any pair of time points to characterize any pairwise changes. The limitation of this approach is that it deals with the longitudinal data points as a collection of static snapshots and ignores temporal dependencies. Furthermore, if more than one sample is taken in the same time window, it may result in either retaining only one sample and excluding the others or taking the

average of the measured feature's values, which may lead to mischaracterizing the exact microbial behavior.

Another strategy is to identify time intervals of differentially abundant microbial features. To date, two methods have been proposed; the first is *MetaSplines* (77), and the second is *MetaDprof* (78).

*MetaSplines* is implemented as an R-script within the *metagenomeSeq* package (111). It starts by fitting a curve for the difference between the means of the read counts of a particular feature from two phenotypic groups across different time points. It uses the standard Smoothing Spline ANOVA (SS-ANOVA) approach (69; 79; 80), where the Gaussian distribution is assumed for the reads. The identification of a significant time interval is based on comparing the areas under the fitted curve to that of the null model, which is generated from the bootstrapping of the sample's group labels. This method is easy to use, and it handles time point inconsistencies in the samples collection, such as, variable sample collection times and the uneven number of time points among the subjects' longitudinal timeframe. However, it assumes the normality of metagenomic read counts, which are not suitable to be modeled by a Gaussian distribution. Additionally, *MetaSplines* has a relatively high false positive rate (78).

*MetaDprof* is also based on the standard SS-ANOVA which assumes the normality of the read count. The difference between the two methods is that *MetaDprof* fits a curve for each phenotypic/treatment group and compares the area between the two curves with the data generated by the permutation of the group labels. The identification of the significant time intervals is accomplished in two steps. Initially, *MetaDprof* tests whether the feature is globally significant or

not. In the event that the feature is significant, it proceeds to identify significant time intervals. Compared to *MetaSplines* and *Next-maSigPro* (112), a tool for differential analysis of longitudinal RNAseq profiles, *MetaDprof* shows a high level of detection power. *MetaDprof* performs very well when the following criteria are met: a) the samples are equally spaced, b) the number of samples taken from each subject is equal, and c) the samples from subjects are collected at the same time points. In animal model studies, the sample collection process can be well controlled to meet the criteria. However, they rarely can be met for samples collected from humans, particularly when an invasive procedure is used, such as bronchoscopy to obtain a bronchoalveolar lavage (BAL) fluid samples.

### 4.1.1 Problem Definition

Develop a method to accurately identify time intervals of differentially abundant features in metagenomic longitudinal studies.

**Significance**

The identified differentially abundant features and their time intervals have the potential to distinguish microbial biomarkers that may be used for microbial reconstitution therapy through bacteriotherapy, probiotics, or antibiotics, and may also suggest timing and duration of the therapy.

## 4.2 Methods

Fixing a feature $f = 1, \ldots, F$, the data under consideration are the random variables $Y_{tki}$ or their observations $y_{tki}$ of mapped reads of the $i$th subject of phenotype $k$ to the feature $f$ at time point $t$, where $t = 1, \ldots, T$, $k = 1, 2$, and subject $i = 1, \ldots, n_k$.

The random variable $Y_{tki}$ is assumed to follow a negative binomial distribution

$$Y_{tki} \sim NB(\alpha, p(t, k)) \tag{4.1}$$

with integer $\alpha > 0$ and success probability $p(t, k) \in (0, 1)$. That is, $Y_{tki}$ stands for the number of failures before the $\alpha$th success in a sequence of Bernoulli trials. Then the probability for observing $y$ number of reads can be written as

$$P(Y_{tki} = y) = \frac{\Gamma(\alpha + y)}{y!\Gamma(\alpha)} \cdot p(t, k)^{\alpha} \cdot (1 - p(t, k))^{y} \tag{4.2}$$

with an expectation and variance

$$E(Y_{tki}) = \frac{\alpha(1 - p(t, k))}{p(t, k)} \tag{4.3}$$

$$Var(Y_{tki}) = \frac{\alpha(1 - p(t, k))}{p(t, k)^2} \tag{4.4}$$

To model the time and phenotypic effect we use a general linear model with a logit link:

$$\eta(t,k) = \log \frac{p(t,k)}{1-p(t,k)} \tag{4.5}$$

From (Equation 4.5), we have

$$p(t,k) = \frac{e^{\eta(t,k)}}{1+e^{\eta(t,k)}} \tag{4.6}$$

$$1-p(t,k) = \frac{1}{1+e^{\eta(t,k)}} \tag{4.7}$$

Assuming $Y_{tki}$'s are independent, the log likelihood given a time-course metagenomic count profiles $\mathbf{y} = \{y_{tki}\}_{t=1,\ldots,T;k=1,2;i=1,\ldots,n_k}$ is calculated as:

$$\begin{aligned}
\mathcal{L} &= \log L(\mathbf{p}, \alpha \mid \mathbf{Y} = \mathbf{y}) \\
&= \sum_{t=1}^{T}\sum_{k=1}^{2}\sum_{i=1}^{n_k}[y_{tki}\log(1-p(t,k)) + \alpha \log p(t,k) \\
&\quad + \log\Gamma(\alpha + y_{tki}) - \log\Gamma(\alpha) - \log(y_{tki}!)] \\
&= \sum_{t=1}^{T}\sum_{k=1}^{2}\sum_{i=1}^{n_k}[y_{tki}\log(1-p(t,k)) + \alpha \log p(t,k) \\
&\quad + \log\Gamma(\alpha + y_{tki}) - \log\Gamma(\alpha)] + \text{constant}
\end{aligned} \tag{4.8}$$

Given the success probabilities $\mathbf{p} = \{p(t,k)\}_{t=1,\ldots,T;k=1,2}$ or equivalently the linear predictors $\boldsymbol{\eta} = \{\eta(t,k)\}_{t=1,\ldots,T;k=1,2}$, the main part of $\mathcal{L}$ involving $\alpha$ is

$$\mathcal{L}_p(\alpha) = \sum_{t=1}^{T}\sum_{k=1}^{2}\sum_{i=1}^{n_k}[\log\Gamma(\alpha + y_{tki}) - \log\Gamma(\alpha) + \alpha\log p(t,k)] \tag{4.9}$$

which will be maximized to update $\alpha$ later.

Given the number of failures $\alpha > 0$, using (Equation 4.6), (Equation 4.7), (Equation 4.8), we have the main part of $\mathcal{L}$ involving $\mathbf{p}$ or $\boldsymbol{\eta}$:

$$\mathcal{L}_\alpha(\boldsymbol{\eta}) = \sum_{t=1}^{T}\sum_{k=1}^{2}\sum_{i=1}^{n_k}[\alpha\eta(t,k) - (\alpha + y_{tki})\log(1 + e^{\eta(t,k)})] \tag{4.10}$$

We seek the estimation of model parameters $\alpha$ and $p(t,k)$ by maximizing (Equation 4.8). Following (Gu, 2013) (81), in order to control the smoothness of the function $\eta$, a roughness penalty $J(\eta)$ is added to the minus log-likelihood together with the smoothing parameter $\lambda > 0$ for the trade-off between the goodness of fit and the smoothness of the spline curve:

$$\min_{p,\alpha} -\mathcal{L} + \lambda \cdot J(\eta) \tag{4.11}$$

In the objective function, $\mathcal{L}$ encourages the goodness of fit; $J(\eta)$ quantifies the smoothness of $\eta$, which is essentially the inner product in a reproducing kernel Hilbert space (Gu, 2013) (81), Section 3.1). The $\lambda$ in expression (Equation 4.11) controls the tradeoff between the goodness of fit

and the smoothness of the spline and can be determined using performance-oriented iterations or cross-validation (Gu, 2013 (81) Section 5.2).

The solution to the optimization problem in (Equation 4.11) leads to the smoothing spline that fits the reads from the samples across multiple time points. With the estimated parameters $\alpha$ and $p(t,k)$, we obtain the estimated mean of $Y_{tki}$ using (Equation 4.3), (Equation 4.6), (Equation 4.7), i.e.,

$$E(\hat{Y}_{tki}) = \hat{\alpha}e^{\hat{\eta}(t,k)} = \frac{\hat{\alpha}\hat{p}(t,k)}{1-\hat{p}(t,k)} \tag{4.12}$$

Connecting the values at each time point using (Equation 4.12) the fitted curve can be constructed in each group. With (Equation 4.4) and (Equation 4.12), the confidence intervals can be obtained for each feature. We use the R package `gss` (Gu, 2013 (81)) to solve problem (Equation 4.11). For readers' reference, a more detailed description for the algorithm used in (81), Section 5.4.6) with a specified $\lambda > 0$ is given below:

$0^\circ$ Given data $\{y_{tki}\}_{t=1,\ldots,T;k=1,2;i=1,\ldots,n_k}$, find the maximum likelihood estimate for the usual logistic regression model with negative binomial responses. That is, determine $\tilde{\alpha}^{(0)}, \tilde{p}^{(0)}(t,k), t = 1,\ldots,T; k = 1,2$ that maximize $\mathcal{L}$ in (Equation 4.8). Denote

$$\tilde{y}^{(0)}_{tki} = y_{tki},\ \tilde{\eta}^{(0)}(t,k) = \log(\tilde{p}^{(0)}(t,k)/(1-\tilde{p}^{(0)}(t,k)))$$

$t = 1,\ldots,T; k = 1,2; i = 1,\ldots,n_k$.

For iteration $s = 1,\ldots,S$, do $1^\circ$, $2^\circ$ and $3^\circ$:

1° Determine $\tilde{\alpha}^{(s)}$ that maximizes

$$\sum_{t=1}^{T}\sum_{k=1}^{2}\sum_{i=1}^{n_k}[\log\Gamma(\alpha+\tilde{y}_{tki}^{(s-1)})-\log\Gamma(\alpha)+\alpha\log\tilde{p}^{(s-1)}(t,k)]$$

2° For $t=1,\ldots,T; k=1,2; i=1,\ldots,n_k$, let

$$\begin{aligned}
\tilde{u}_{tki}^{(s)} &= (\tilde{\alpha}^{(s)}+\tilde{y}_{tki}^{(s-1)})\tilde{p}^{(s-1)}(t,k)-\tilde{\alpha}^{(s)}\\
\tilde{w}_{tki}^{(s)} &= (\tilde{\alpha}^{(s)}+\tilde{y}_{tki}^{(s-1)})\tilde{p}^{(s-1)}(t,k)\cdot(1-\tilde{p}^{(s-1)}(t,k))\\
\tilde{y}_{tki}^{(s)} &= \tilde{\eta}^{(s-1)}(t,k)-\tilde{u}_{tki}^{(s)}/\tilde{w}_{tki}^{(s)}
\end{aligned}$$

3° Use quasi-Newton approach to find $\tilde{\eta}^{(s)}(t,k)$'s that minimize the penalized weighted least squares functional

$$\frac{1}{T(n_1+n_2)}\sum_{t=1}^{T}\sum_{k=1}^{2}\sum_{i=1}^{n_k}\tilde{w}_{tki}^{(s)}(\tilde{y}_{tki}^{(s)}-\eta(t,k))^2+\lambda J(\eta)$$

Let $\tilde{p}^{(s)}(t,k)=e^{\tilde{\eta}^{(s)}(t,k)}/(1+e^{\tilde{\eta}^{(s)}(t,k)})$, $t=1,\ldots,T; k=1,2$.

Once we have the two splines that fits each group's samples, we can then calculate the normalized area between the two curves for each unit time interval of the $T-1$ time intervals. The normalized Area Ratio (AR) is calculated as in (Equation 4.13), where $A_{t,t+1}^{k_1}$ and $A_{t,t+1}^{k_2}$ denote the area under the spline curve from time $t$ to time $t$+1 for group 1 and group 2, respectively, $t=1,\ldots,T-1$.

$$AR_{t,t+1} = \frac{A^{k_1}_{t,t+1} - A^{k_2}_{t,t+1}}{max(A^{k_1}_{t,t+1}, A^{k_2}_{t,t+1})} \tag{4.13}$$

Then, we perform a permutation procedure by permuting the sample group labels to calculate the $AR_b$ for the random samples for each time interval. The procedure is repeated $B$ times. This is essential for calculating the $p - value$ of each interval. The $p\_value$ is calculated using (Equation 4.14)

$$p\_value = \frac{\#(AR_b > AR)}{B} \qquad b = 1, ..., B \tag{4.14}$$

The significant time intervals are those with $p\_value < 0.05$ after multiple testing correction (113) which adjusts for the number of time intervals per feature and for the multiple features that are testing for. Figure 12 gives an illustration of how *MetaLonDA* method works.

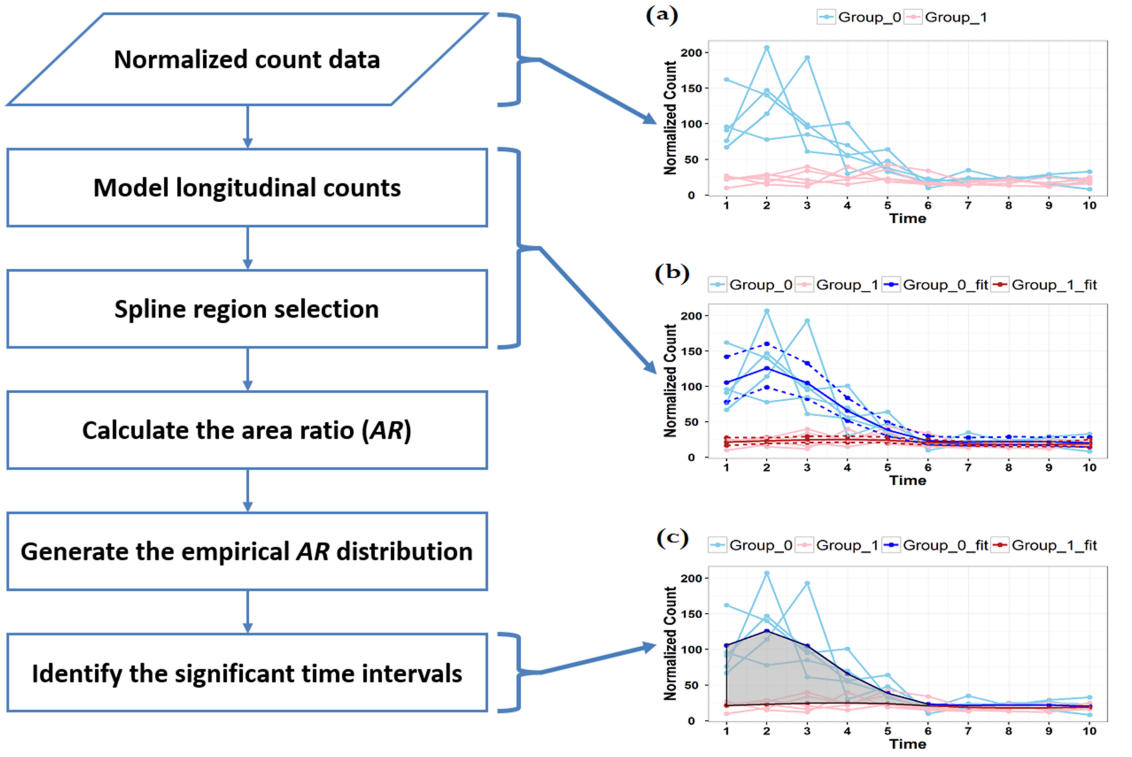



Figure 12: Illustration of how *MetaLonDA* works. (a) The longitudinal samples for one feature from two phenotypic groups. (b) The two fitted NB Smoothing Splines of two groups (The solid dark blue and dark red curves). The dashed curves show the 95% confidence interval. (c) The significant time intervals identified (the grey highlighted regions) (114).

### 4.2.1 *MetaLonDA* R-package Framework

The main components of the *MetaLonDA* framework are shown in Figure 13.

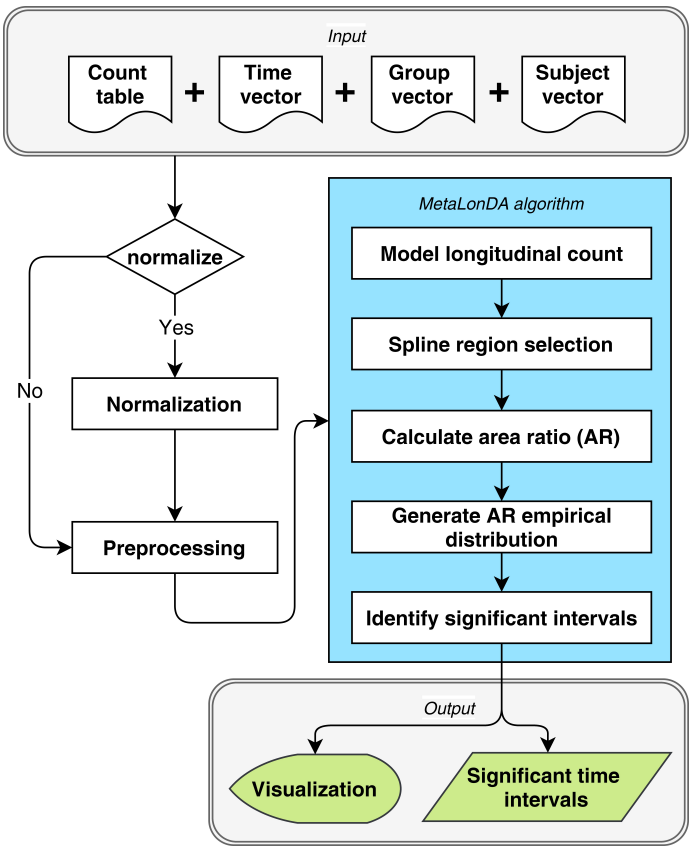



Figure 13: *MetaLonDA* R-package framework (115).

**Input**

Metagenomic reads are processed for each sample to construct taxonomic and/or functional profiles (45; 56; 116; 117). The taxonomic profiles, functional profiles, or both for all samples from different subjects are then integrated into one count table $C$ with dimension of $m \times n$, where $m$ denotes the number of microbial features and $n$ denotes the number of metagenomic samples. $C(i, j)$ represents the number of reads from sample $j$ that mapped to microbial feature $i$. The count table $C$ is the main input to *MetaLonDA*. Additionally, three vectors each of length $n$ are needed for *MetaLonDA* to perform the analysis: (a) time of sample collection vector $T$, (b) phenotypic group vector $G$, and (c) subject ID vector $I$. As previously highlighted, *MetaLonDA* supports unequal numbers of samples between subjects, unequal numbers of subjects between phenotypic groups, and uneven elapsed time between time points.

**Normalization**

Since metagenomic samples may have different sequencing depths, the aggregated metagenomic counts need to be normalized among samples (118; 119; 120). *MetaLonDA* incorporates three different normalization methods into its framework: (a) Cumulative Sum Scaling (111), (b) median-of-ratios scaling factor (121), and (c) Trimmed Mean of M-values (122). If the count table is already normalized, the normalization step should be skipped in *MetaLonDA*. As a preprocessing step for *MetaLonDA* and based on a user-specified threshold, relatively low abundant features are removed from the metagenomic count table. In our model, we assume that the normalized counts of each feature follow a negative binomial (NB) distribution, which is different from mod-

eling the original counts as NB distributed after incorporating a size factor into the mean as in *DESeq2* (121).

***MetaLonDA* Core Algorithm**

MetaLonDA's core algorithm is discussed before in section 4.2.

**Output Format and Visualization**

*MetaLonDA* outputs a table that includes significant features, start and end points of the corresponding significant intervals, the adjusted $p$-value of each significant time interval, and the phenotypic group in which the corresponding feature is more abundant. In addition to the output table, *MetaLonDA* produces two types of visualizations: (a) a figure showing the fitted splines of each group and the associated time interval for each feature that has at least one significant time interval, and (b) a figure visualizing the identified time intervals of the differentially abundant features.

## 4.3 Experiments and Results

### 4.3.1 Evaluation of the Negative Binomial Assumption

One major assumption of *MetaLonDA* is that the number of metagenomic reads mapped to microbial features follows a NB distribution. To evaluate this assumption, we extracted the count data from Caporaso *et al.* (70). In this dataset, microbial samples were taken on a daily basis from a man and a woman over a period of 15 months and 6 months, respectively, from four dif-

ferent body sites. The obtained read counts were normalized using the median-of-ratios scaling factor method (121). After filtering out the relatively rare operational taxonomic units (OTUs) with fewer than 5 reads, a total of 750 OTUs were selected from 1967 samples. The Q-Q plot in Figure 14 exemplifies the suitability of modeling read counts of *Klebsiella* species using different parametric distributions, namely, NB, Poisson, zero-inflated Poisson (ZIP), and lognormal distributions. The theoretical quantiles of each parametric distribution are calculated from random numbers generated from each parametric distribution with parameters estimated from each OTU read count.

Parameters of each distribution are calculated as following, for each vector of feature's reads counts, we used the `fitdistr` function from the *MASS* R-package (123) to estimate the parameters of each parametric distribution used in the project except zero-inflated Poisson (ZIP) distribution. Here are the parameters for each distribution: (a) Negative-binomial distribution: *size* and *mean*, (b) Poisson distribution: *lambda*, (c) Zero-inflated Poisson distribution: *p* and *lambda*, (d) Lognormal distribution: *mean* and *standard deviation*, (e) Normal distribution: *mean* and *standard deviation*, and (f) Exponential distribution: *rate*. For zero-inflated Poisson distribution, we used the `zeroinfl` function from the *pscl* R-package (124; 125) to fit each features read counts with a ZIP. Then we extracted the values of *p* (zero-inflation probability) and the *lambda*. Using the estimated parameters for all aforementioned distributions except ZIP, we simulated $N$ ($N$ = # of samples of the Caporaso *et al.,* study (70)) random numbers are generated using the corresponding parametric distribution. For ZIP, we can generate $N$ random numbers following ZIP with the estimated parameters using `rzipois` function from the *VGAM* R-package (126; 127).

The *p*-value on the top of each sub-figure of Figure 14 represents the BH adjusted *p*-value of the two-sample Kolmogorov-Smirnov (KS) test (128), where a higher *p*-value indicates that the two samples are derived from the same population distribution, and smaller *p*-value indicates that the two samples are drawn from different population distributions. In the case of *Klebsiella*, only the NB distribution is considered suitable (*p*-value = 0.28).

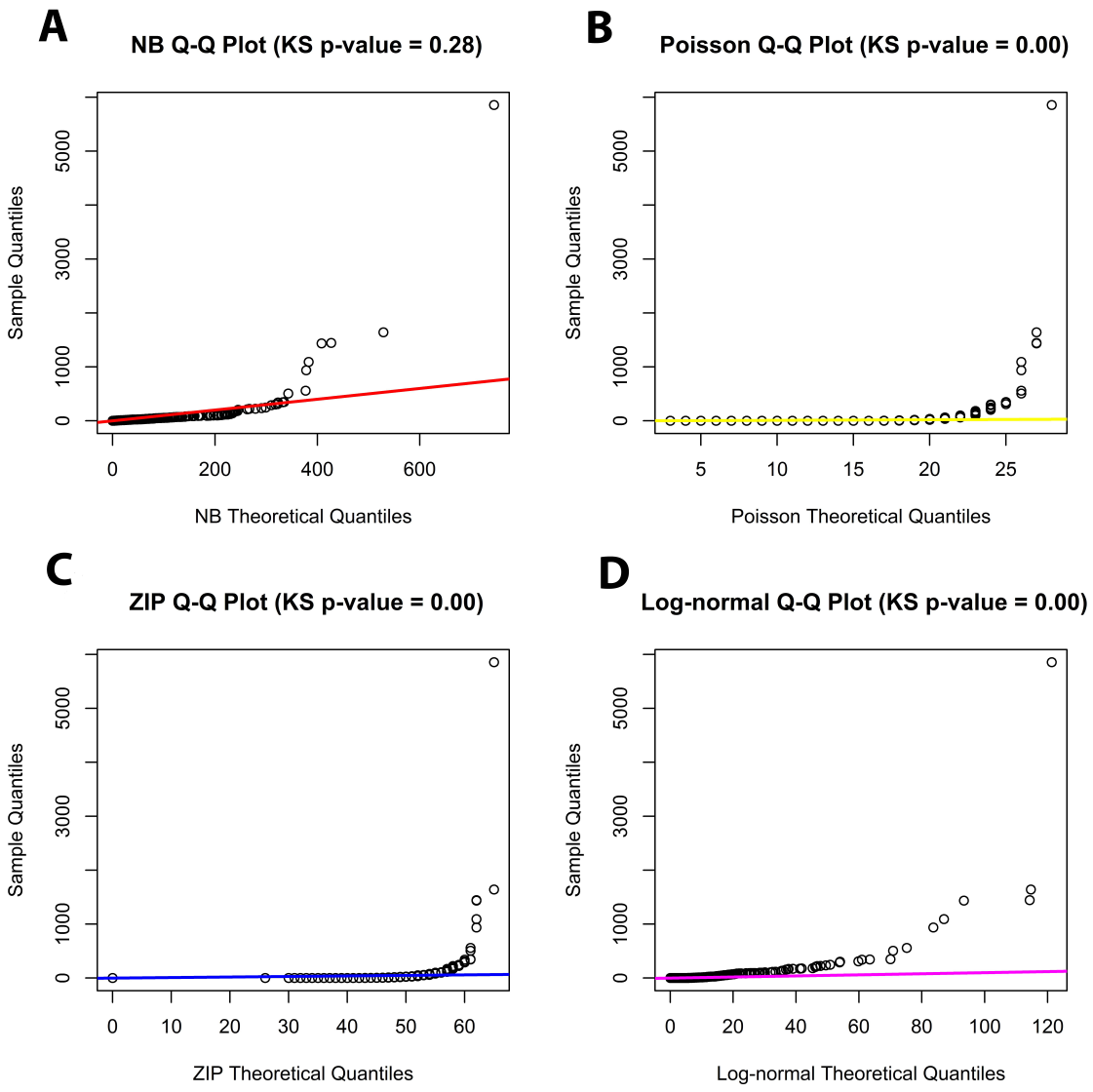



Figure 14: Quantile-Quantile plot between different theoretical distributions on *Klebsiella* read counts. Each sub-figure represents a different distribution: (a) NB distribution; (b) Poisson distribution; (c) ZIP distribution; and (d) Lognormal distribution. The *p*-value above each sub-figure represents the significance of the KS test between the sample quantiles and the theoretical quantiles of the corresponding distribution. The NB distribution is most appropriate to model the OTU count among other standard distributions (115).

To evaluate all other features, we applied the KS test to the read counts of each of the 750 OTUs and the sampled numbers from the corresponding parametric statistical distribution that had the same parameters as estimated from the read counts. Table IV summarizes the number of features that do not show significant divergence ($p$-value $> 0.05$ after BH multiple testing corrections) with NB, ZIP, Poisson, lognormal, exponential, half-normal, and normal distributions. Out of the 750 features, 96% were modeled appropriately using NB distribution. In comparison, ZIP and Poisson were appropriate for 41% and 26% of the OTUs, respectively, whereas the rest of the parametric distributions employed in this analysis barely fit. This indicates the appropriate use of NB as a parametric distribution model for *MetaLonDA* when compared to other standard parametric distributions. Furthermore, this finding is consistent with previous studies that show that cross-sectional differential abundance methods that use a NB distribution to model microbial features outperform methods that rely on other distributions, especially when the number of samples is small (129).

TABLE IV: Number and percentage of species out of 750 species that do not show significant differences (KS $p$-value>0.05) with various standard statistical parametric distributions. The count data is taken from Caporaso *et al.*.

| | **Number** | **Percentage** |
|---|---|---|
| **NB** | 721 | 96.13 |
| **ZIP** | 309 | 41.20 |
| **Poisson** | 201 | 26.80 |
| **Lognormal** | 1 | 0.13 |
| **Exponential** | 0 | 0 |
| **Half-normal** | 0 | 0 |
| **Normal** | 0 | 0 |

### 4.3.2 Performance Evaluation Based on Simulated Datasets

In order to benchmark *MetaLonDA*'s performance, we performed a comprehensive simulation study. Longitudinal features (n=1000) were simulated from NB, Poisson, and ZIP distributions using the *corcounts* R-package (130). Although read counts of metagenomic features follow NB distribution as shown in Table IV, the purpose of simulating data from Poisson and ZIP was to evaluate the robustness of *MetaLonDA* when read counts fail to follow the NB distribution. These

simulated features were categorized into two types: (a) 500 differentially abundant features between the two testing groups, and (b) 500 features that were not differentially abundant between the two testing groups. In the case of the differentially abundant features (demonstrated in Figure 16A), the mean $\mu(t)$ pattern is simulated to be differentially abundant in three regions: (a) at the start of the study course, (b) at the end of the study course, (c) in the middle of the study course (Equation 4.15). In the case of non-differentially abundant features, the $\mu(t) = \mathcal{N}(20, 1)$, where $\mathcal{N}$ denotes normal distribution and $t = 0, \ldots, 20$.

$$\begin{aligned}\mu(t) =& \mathcal{N}(20,1) + [\mathcal{N}(20,1) * (5-t) * I(t<5)] + \\ & [2 * \mathcal{N}(20,1) * (t-8) * I(t>8 \& t \leqslant 11)] + \\ & [2 * \mathcal{N}(20,1) * (13-t) * I(t>11 \& t \leqslant 13)] + \\ & [\mathcal{N}(20,1) * (t-15) * I(t>15)]\end{aligned} \tag{4.15}$$

For features simulated from the NB distribution, we used a size factor equal to $40/\mu(t)$. In the case of Poisson distribution, we used $\lambda = \mu(t)$, and in the case of zero-inflated Poisson distribution, we used $p(y = 0) = 0.3$ for the zero-inflation parameter. Our choice of the zero-inflation probability was based on the analysis of $\hat{p}(y = 0)$ when we fitted all features in the Caporaso *et al.*, study (70) with the ZIP distribution (Table IV). The histogram in Figure 15 shows that 75% of the $\hat{p}(y = 0)$ is less than 0.3 (median of $\hat{p}$ = 0.1). Therefore, our choice of 0.3 is to evaluate how *MetaLonDA* performs in this case of simulated zero inflation.

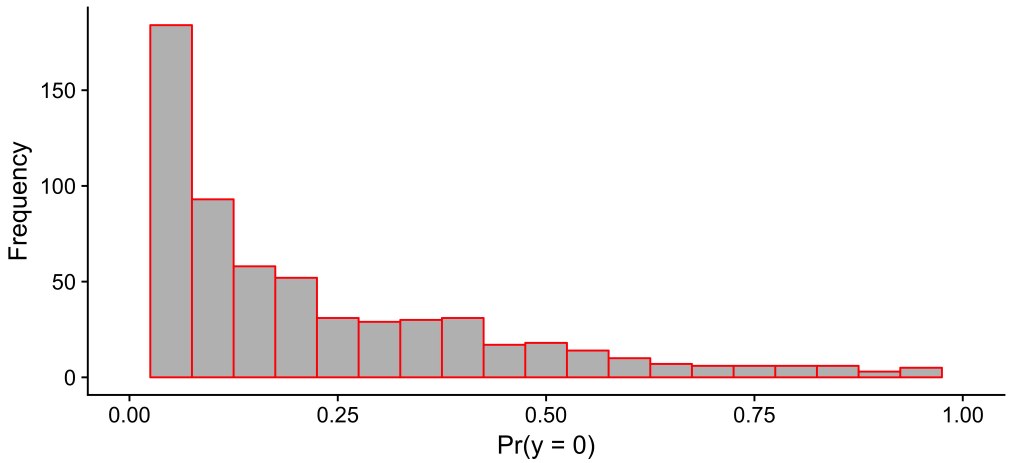



Figure 15: Zero-inflation probability distribution of the fitted ZIP distribution. Read counts are taken from the Caporaso *et al.*, study (115).

In order to mimic the correlation behavior between adjacent time points in longitudinal studies, the simulation of read counts of adjacent samples followed the first-order autoregressive model (131) with a correlation coefficient $\rho = 0.9$. Datasets were simulated for 15 subjects with 20 time points each ($T = 20$). Additionally, to mimic inconsistencies in the number of subjects per group and number of samples per subject, we randomly chose 11 samples from 8 subjects from group (A) and 8 samples from 6 subjects from group (B) (Figure 16A).

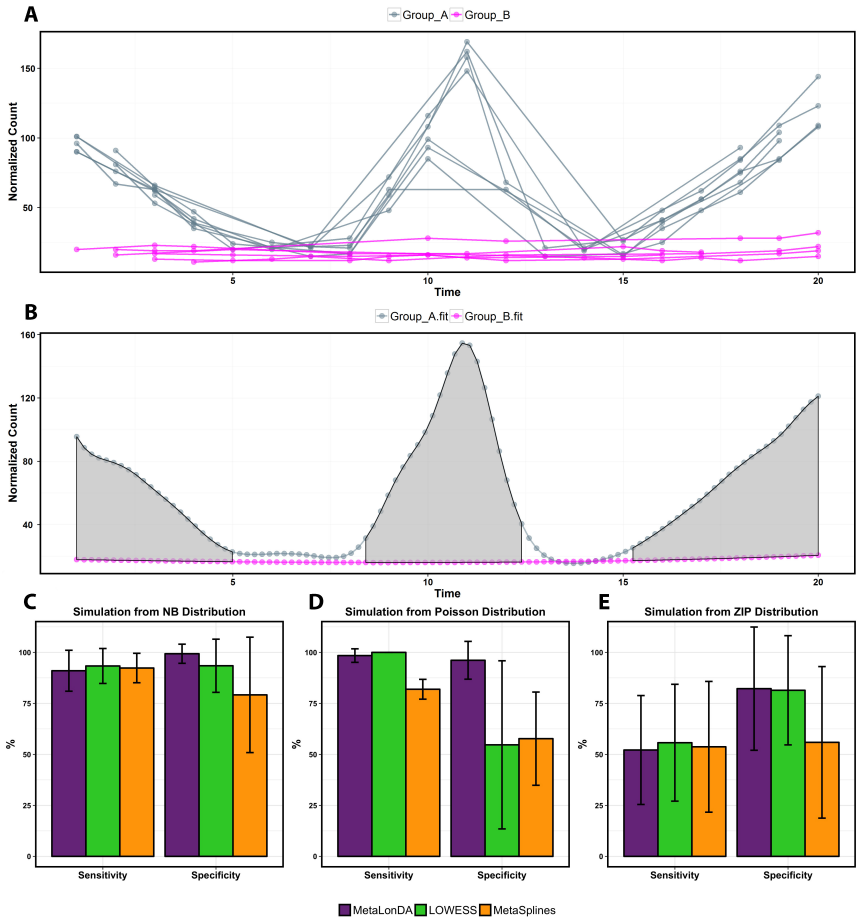



Figure 16: Pattern and performance evaluation of data simulated from various statistical distributions. (a) The pattern of the simulated longitudinal features. Each differentially abundant feature has time intervals between group A and B at [1,5] $\bigcup$ [8,13] $\bigcup$ [15,20] time-unit and non-differential time intervals [5,8] $\bigcup$ [13,15]. The simulated data mimics inconsistencies in sample collection (different number of subjects per group, different number of samples per subject, and samples are not equally spaced.) (b) The fitted smoothing spline of each group and the highlighted significant time intervals between the two groups. (c-e) The performance of different tools using data simulated from NB, Poisson, zero-inflated Poisson, respectively. Each bar represents the mean among 1000 features, and the error bar represents the standard deviation. *MetaLonDA* always has higher specificity than *LOWESS* and *MetaSplines*. This shows *MetaLonDA*'s robustness among different distributions (115).

TABLE V: Performance evaluation on data simulated from various statistical distributions mimicking consistent sampling.

| | **NB** | | **Poisson** | | **ZIP** | |
|---|---|---|---|---|---|---|
| | **Sensitivity (%)** | **Specificity (%)** | **Sensitivity (%)** | **Specificity (%)** | **Sensitivity (%)** | **Specificity (%)** |
| ***MetaLonDA*** | **98** | **95** | 99 | **96** | 84 | 90 |
| ***MetaDprof*** | 94 | 94 | 86 | 94 | 87 | **96** |
| ***LOWESS*** | 96 | 80 | **100** | 47 | **94** | 60 |
| ***MetaSplines*** | 81 | 79 | 85 | 59 | 60 | 64 |

We proceeded to evaluate the performance of *MetaLonDA* in comparison to *MetaSplines*, *MetaDprof*, and *LOWESS* (132). *LOWESS* is a non-parametric local regression model that is based on combining multiple regression models in a *k*-nearest-neighbor-based meta-model. In the context of this project, *LOWESS* refers to using the *LOWESS* regression model to substitute the NB distribution in *MetaLonDA*'s framework. Each method was run for 1000 permutations to construct the AR empirical distribution. The *p*-value threshold was set to 0.05 after multiple testing corrections using BH. The rest of the parameters were set to default. The assessment is based on the $sensitivity = \frac{TP}{TP+FN}$ and $specificity = \frac{TN}{TN+FP}$. In this context, *TP* represents the number of truly identified time intervals of differentially abundant features. *TN* represents the number of truly identified time intervals of non-differentially abundant features, *FP* represents the falsely

identified time intervals of non-differentially abundant features, and *FN* represents the falsely identified time intervals of differentially abundant features.

Table V shows the performance evaluation based on consistent sampling, i.e., the ideal scenario which is rare. *MetaLonDA* has the most balanced prediction in terms of sensitivity and specificity followed by *MetaDprof* and *MetaSplines*.

Next, we benchmarked *MetaLonDA* using the inconsistent sampling scenario. In this experiment, *MetaDprof* was excluded since its package cannot handle the sampling inconsistencies. In the case of data simulated from NB distribution, Figure 16C shows that *MetaLonDA* outperforms *MetaSplines* and *LOWESS* in sensitivity and specificity. On the other hand, in the case of data simulated from Poisson distribution, Figure 16D demonstrates that *LOWESS* has a slightly better sensitivity than *MetaLonDA* (100% vs. 98%). But, the specificity of *LOWESS* and *MetaSplines* is very low when compared to *MetaLonDA* (50% vs. 95%). This is because *LOWESS* and *MetaSplines* over-fit the data. Lastly for the case of the zero-inflated Poisson, Figure 16E shows that *MetaLonDA*, *MetaSplines*, and *LOWESS* have a comparatively low-level of sensitivity($\sim$50%), but *MetaLonDA* has higher specificity. The reason behind this low sensitivity is the high zero inflation probability we chose for ZIP, p(y=0)=0.3. To summarize, *MetaLonDA* always maintains a very high specificity, in contrast to *LOWESS* and *MetaSplines*.

The execution time of *MetaLonDA*, *MetaDprof*, and *MetaSplines* is comparable and depends on the number of permutations used. Analysis of the simulated dataset from a NB distribution with 1000 features took 104 minutes with *MetaLonDA*, 113 minutes with *MetaDprof*, and 99 minutes with *MetaSplines*. The analysis was conducted on a MAC machine with 2.5 GHz Intel Core i7 pro-

cessor and 16 GB 1600 MHz RAM. For the same analysis, *LOWESS* was slightly faster (87 minutes) because it does not have the complex smoothing spline optimization equation (Equation 4.11) that needs to be solved numerically.

### 4.3.3 Performance Evaluation Based on a Biological Dataset: Hygiene Hypothesis Study

In order to assess the biological significance of the identified time intervals of differentially abundant features, we used a publicly available dataset from a longitudinal metagenomic study that investigates the hygiene hypothesis (1). The study was part of the DIABIMMUNE project (`https://pubs.broadinstitute.org/diabimmune`). Stool samples were collected from 222 infants (74 from Russia, 74 from Finland, and 74 from Estonia) from birth to $\sim$ 3 years of age. In our analysis, we identified the time intervals with differentially abundant genera in Russian and Finnish infant guts. We focused on the 585 samples (304 from 70 Russian infants and 281 from 71 Finnish infants) that had been sequenced using Metagenomic shotgun (MGS) sequencing. Figure 17 shows the distribution of time points of the stool samples collected from each group (Figure 18 shows the distribution of time points per subject). Reads from the 585 sequenced samples were quality-controlled by filtering out low-quality reads, short reads ($<$60 bp), and human reads. Taxonomic profiles were constructed using *MetaPhlAn2* (100). The number of reads mapped to each taxonomic feature was then normalized to the reads per kilo-base per million sample reads (RPKM) to correct for bias due to differences in genome size and sequencing depth. The aggregated taxonomic profiles of all 585 samples revealed 128 genera.

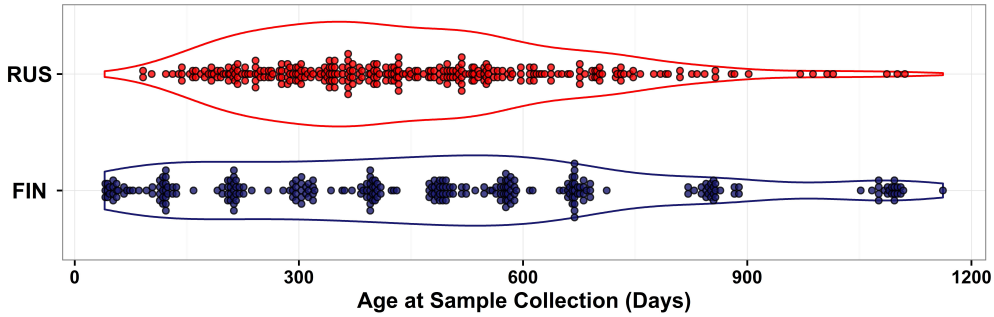



Figure 17: Time distribution of 585 stool samples (304 from 70 Russian and 281 from 71 Finnish) sequenced using MGS in the DIABIMMUNE project. The collected samples have various forms of inconsistencies; different number of subjects per group (70 Russian vs 71 Finnish infants), different number of samples per subject (min=1, max=13), and the samples' time points are not equally spaced (115).

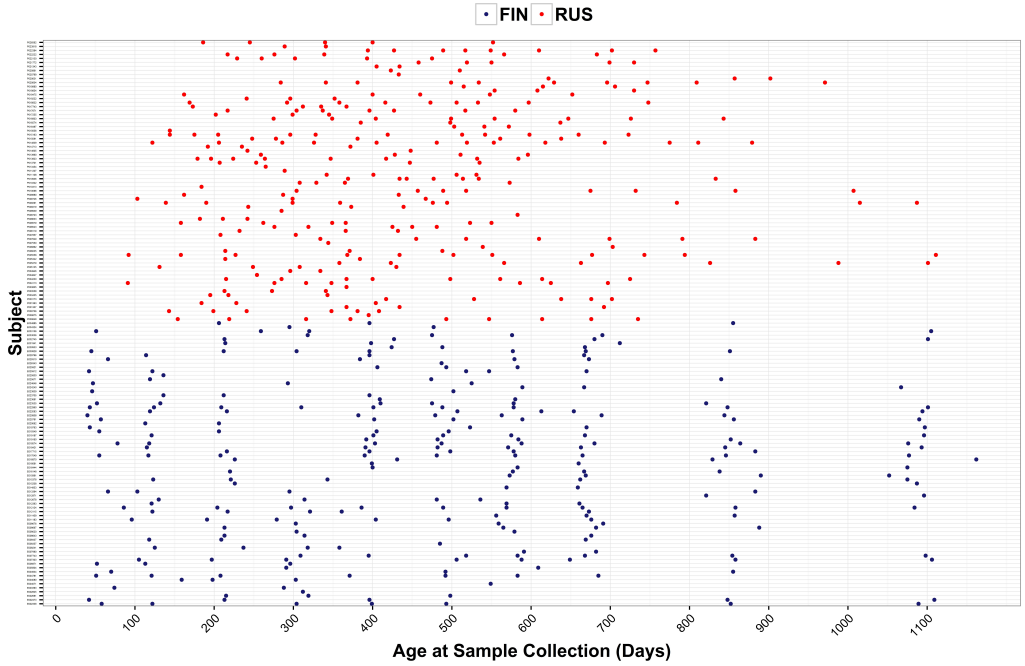



Figure 18: Time points distribution per subject in the DIABIMMUNE study (115).

In order to evaluate the suitability of using NB to model genera read counts before applying *MetaLonDA*, we conducted an analysis similar to the one shown in Table IV. We found that NB can be considered a good fit for 79% of the 128 genera.

We applied *MetaLonDA*, *LOWESS*, and *MetaSplines* to identify the time intervals of the differentially abundant genera. We set the number of permutations for all three methods to 1000, $p$-value threshold = 0.05, multiple testing correction method to BH, and other parameters to default. *MetaLonDA* identified 71 genera that have at least one time interval with differentially abundant genera, *LOWESS* identified 122 genera, and *MetaSplines* identified 80 genera. Although there are 53 mutually inclusive common genera identified by the three methods as shown in Figure 19, this does not necessarily indicate that they share the same identified time intervals as demonstrated in Figure 20. *LOWESS* identified 30 genera that neither *MetaSplines* nor *MetaLonDA* reported. Whereas *MetaLonDA* identified 2 genera that were not reported by either *LOWESS* or *MetaSplines*. These results emphasize the high control of false positive identifications by *MetaLonDA*. This results emphasizes the high control of false positive identifications by *MetaLonDA*. In contrast, *LOWESS* identified 30 genera that neither *MetaSplines* nor *MetaLonDA* reported. The previously discussed simulation study concluded that *LOWESS* and *MetaSplines* have lower specificity compared to *MetaLonDA*. Thus, *MetaLonDA* discovery of few significant time intervals is directly related to its increased specificity compared to the other two methods.

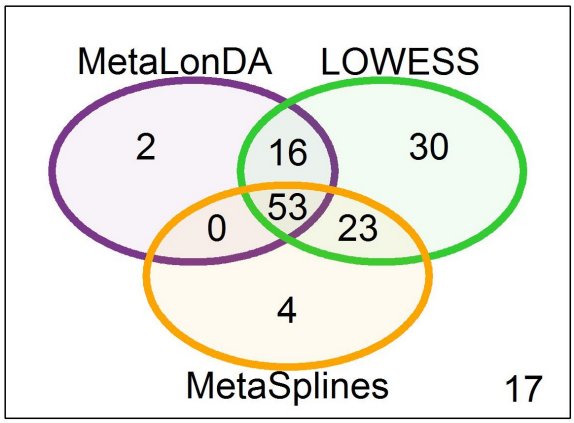



Figure 19: Number of genera identified as differentially abundant between the Finnish and Russian infants. 53 common genera were identified as differentially abundant using the three tools. The ”17” on the lower right corner represents the number of genera that were not identified at any time interval by *MetaLonDA*, *LOWESS*, or *MetaSplines* (115).

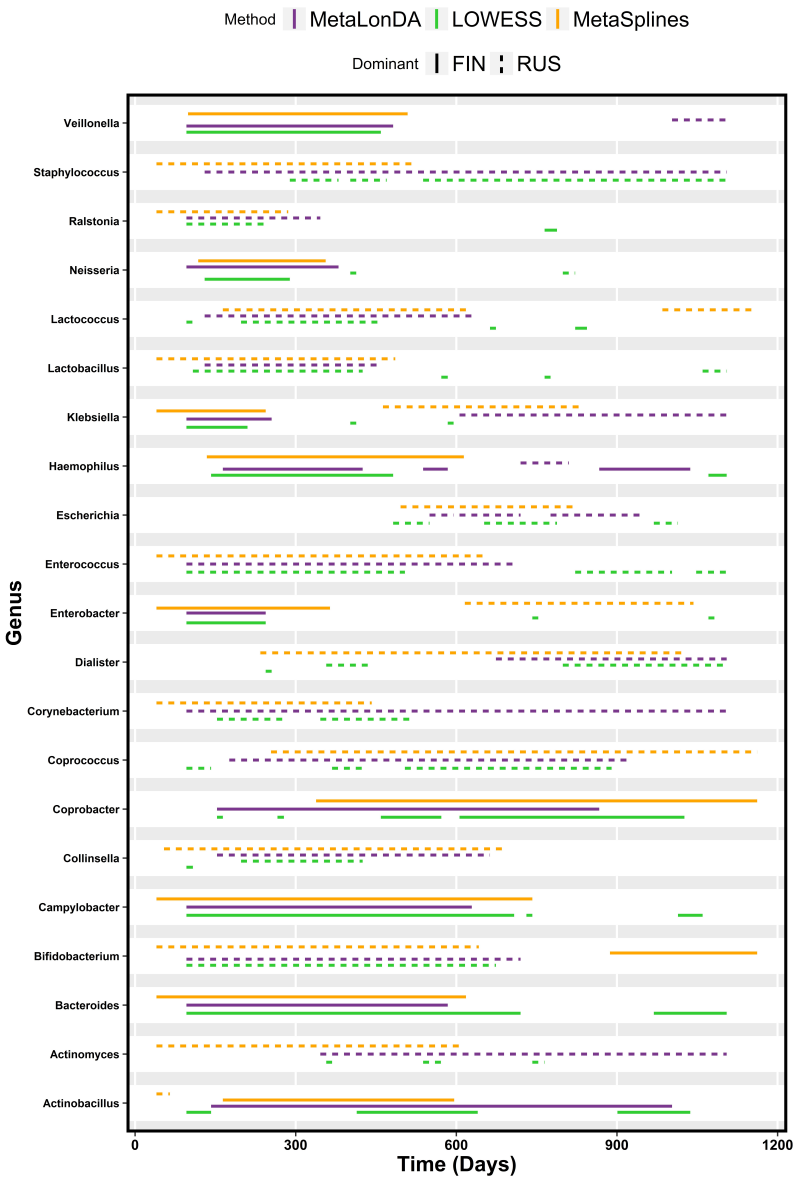



Figure 20: Time intervals of the mutually differentially abundant genera from Finnish and Russian infants identified by *MetaLonDA*, *LOWESS*, and *MetaSplines*. Each line represents significant time interval of the corresponding genus. *MetaLonDA* (purple), *LOWESS* (green), *MetSplines* (orange). The solid lines represent the intervals where samples from the Finnish group have more reads, while the dashed lines represent the differential abundance intervals where samples from the Russian group have more reads (115).

Figure 20 visualizes differences between the time intervals identified by *MetaLonDA*, *LOWESS*, and *MetaSplines* correlating with the major shared genera. In most cases, the time intervals identified by *MetaLonDA* were also identified by either *LOWESS*, *MetaSplines* or both. One critical observation that likely contributes the greater number of false positives observed in *MetaSplines* is that it sometimes identifies time intervals where samples from one group are missing. The absence of one group's samples can make the spline fitting uncontrollable (81). For example, *MetaSplines* identified *Actinobacillus* as relatively more abundant in the Russian infants from day 40 until day 65, although the first Russian sample was collected after 96 days after birth. *MetaLonDA* handles this situation by only reporting significant intervals during the time period when samples from all study groups are available. In the case of the hygiene hypothesis study, individual genera's time intervals identified by *MetaLonDA* are bounded in the range of 96 to 1105 days. Day 96 was the day on which the first sample from a Russian infant was collected, and day 1105 is when the last Russian sample was collected (the first sample from Finnish infants was on day 41, and the last was on day 1162). Since we implemented *LOWESS* on the same *MetaLonDA* framework, it also handles this edge problem. Figure 21 shows the time intervals of differentially abundant genera identified by *MetaLonDA*, *LOWESS*, and *MetaSplines*, while Figure 22 shows time intervals identified by *MetaLonDA* only.

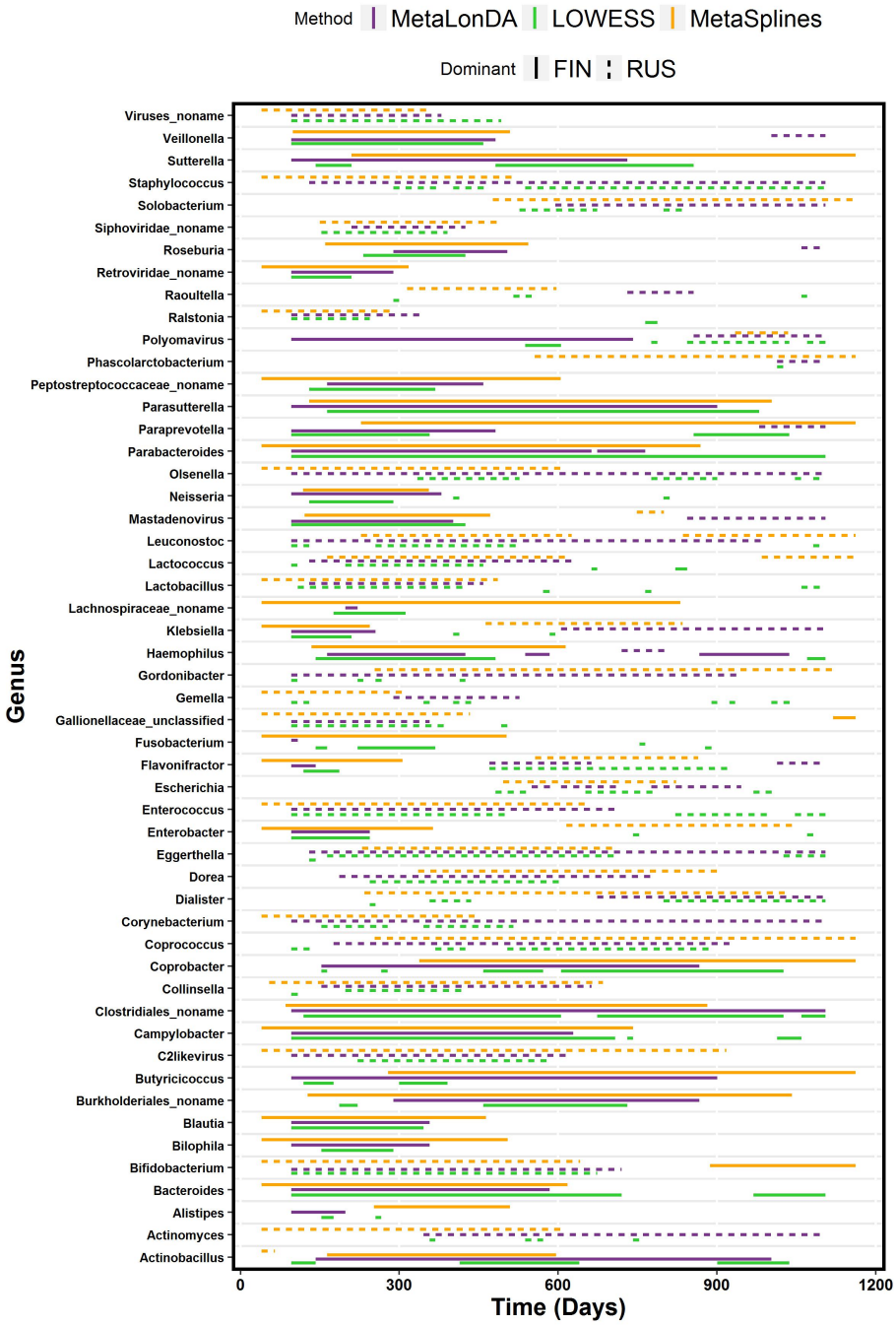



Figure 21: The identified time intervals of the shared differentially abundant genera by *MetaLonDA*, *LOWESS*, and *MetaSplines* between Finnish and Russian infants (115).

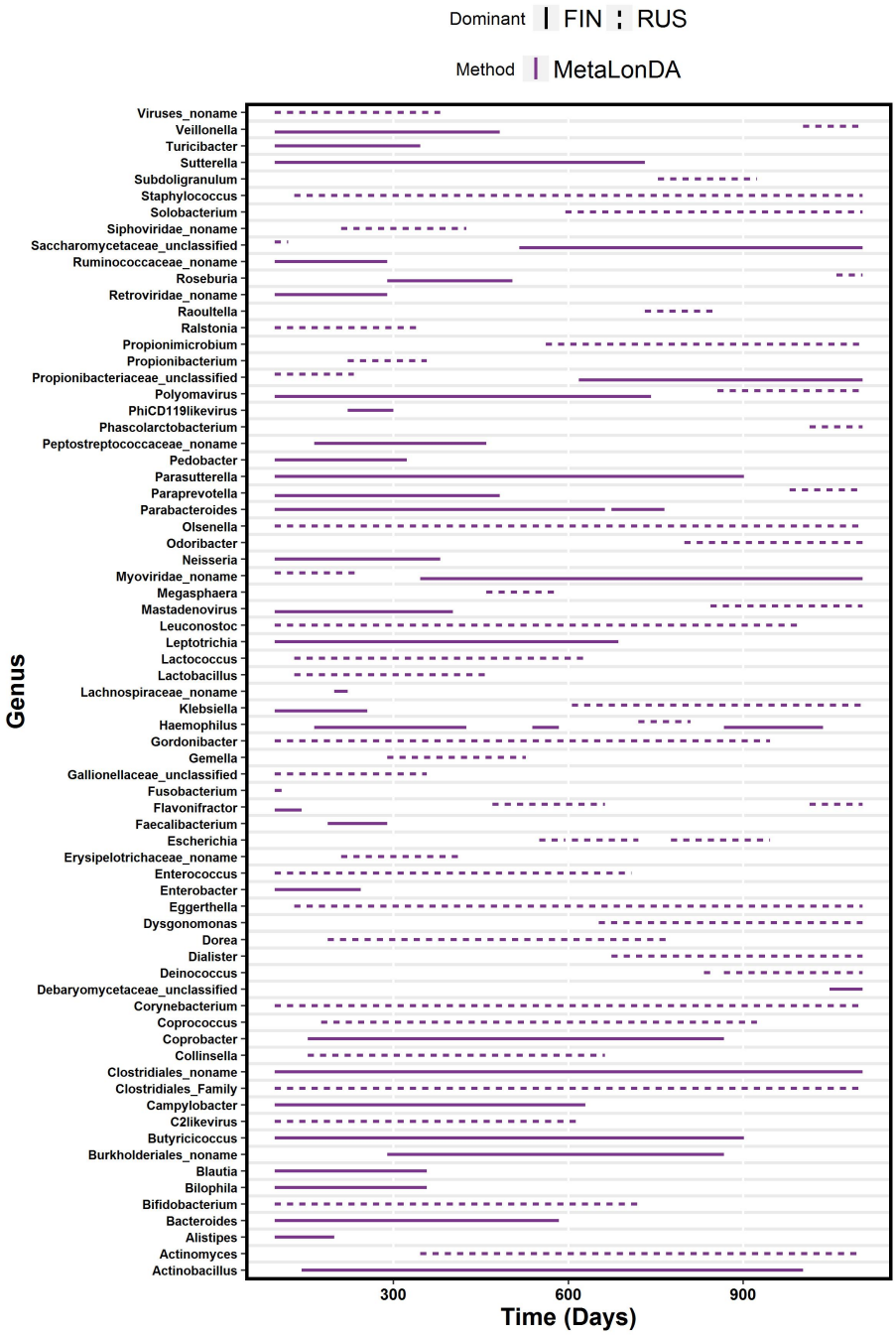



Figure 22: The identified time intervals of the differentially abundant genera by *MetaLonDA* between Finnish and Russian infants (115).

In our analysis, *MetaLonDA* confirms the report by Vatanen *et al.* demonstrating that the genus *Bacteroides* is relatively more abundant during early time points in the Finnish group, whereas the genus *Bifidobacterium* is relatively more abundant in the Russian group (1). *MetaLonDA* specifies that *Bacteroides* were significantly abundant during days 96-584 in Finnish infants, and *Bifidobacterium* were relatively more abundant in Russian infants from day 96 to day 720. Furthermore, in their study, Vatanen *et al.* noted that early life exposure to specific structurally distinct bacterial *lipopolysaccharides* (LPS) influences the development of autoimmune disease. They suggest that in contrast to Russian infants, Finnish infants mount an insufficient immune response due to exposure to *Bacteroides* LPS rather than *Escherichia coli* LPS. Utilization of *MetaLonDA* in this cohort demonstrates that Escherichia establishes a significant community in Russian infants from day 550-946 with little variability. *MetaLonDA* also defined specific time intervals during which other bacterial genera (e.g., *Lactobacillus*, *Leptotrichia*, *Klebsiella*) previously associated as protective or instigating of Type 1 Diabetes (T1D) were differentially abundant (133; 134). Moreover, *MetaLonDA* established that up until day 629, Finnish infants present an additional shift in *Proteobacteria* with an overabundance of genera that are known to be implicated in human disease; including *Campylobacter*, *Haemophilus*, *Klebsiella*, and *Neisseria*. In parallel, when evaluating genera that have previously been associated with protection against T1D, *MetaLonDA* reveals a divergence from *Lactobacillus* and *Lactococcus* to *Veillonella* as the dominant *Firmicutes* genera observed early in the life of Finnish infants. These findings suggest that there is a complex interplay of multiple bacterial genera early in life which may all have immunogenic potential and will allow, in this case, further exploration of the role bacteria specific

LPS as well as other microbial specific stimulators or inhibitors of the host immune response and their role in development of autoimmune disease.

## 4.4 Conclusion

We have developed *MetaLonDA*, a method that can identify significant time-intervals of differentially abundant microbial features such as taxonomies, genes, or pathways. *MetaLonDA* is flexible such that it can perform differential abundance tests on longitudinal samples with different numbers of subjects per phenotypic group, different number of samples per subject, and samples that are not collected at consistent time points. These inconsistencies are often the case for samples collected from human subjects. Inconsistencies increase with the complexity of the procedure utilized to obtain the samples. Usually, there is less inconsistency in samples collected through non-invasive procedures such as stool and urine samples, but increases in the case of invasive procedures such as BAL. *MetaLonDA* relies on two modeling components: the NB distribution for modeling the mapped read counts for each feature and the semi-parametric SS-ANOVA technique for modeling longitudinal profiles associated with different phenotypes.

Extensive experiments on simulated datasets quantitatively demonstrate the effectiveness of *MetaLonDA* with significant improvement over alternative methods. The time needed to execute *MetaLonDA* depends on the number of features being tested and the number of permutations for generating AR empirical distributions. *MetaLonDA* performs significance testing based on unit time intervals that can be hours, days, weeks, months, or years. The identified time intervals of differentially abundant features can be used as preselected features for a machine learning classi-

fier to predict disease prognosis (135; 90; 85). *MetaLonDA* can be applied to any longitudinal count data such as metagenomic sequencing, 16S rRNA gene sequencing, or RNA-Seq. It is worth noting that the NB assumption made for taxonomy would need to be reassessed before *MetaLonDA* can be confidently applied to functional data. In the future, we plan to implement a checker function that evaluates the distributional assumption based on KS test, and accordingly, the best-fitted model can be utilized for the longitudinal differential abundance test.

Furthermore, *MetaLonDA* allows for an in-depth exploration of potential features and establishment of precise time intervals during which individual features may serve as biomarkers from population-based longitudinal studies such as the DIABIMMUNE cohort discussed in this project. Specific significant time intervals can then be utilized to establish targeted timely screening or prevention of individual features and allow for prompt intervention, such as the use of antibiotics or probiotics. Unlike with cross-sectional methods that are incapable of identifying significant time intervals associated with differentially abundant features, *MetaLonDA* may lead to reconstitution of the microbiome and reestablish homeostasis prior to entering the cascade of events that may lead to overt disease.

Although *MetaLonDA* addresses one of the most common limitations in human sample collection inconsistencies, there is still room for improvement. The current version of *MetaLonDA* only finds the association between microbial features, time, and phenotypic group. In the future, we plan to incorporate additional confounding factors (age, gender, race, disease severity, etc.) to the *MetaLonDA* model. Another limitation of *MetaLonDA* is that when samples are sparse over extended time intervals, the fitted smoothing spline has large variation (81). This causes the iden-

tified significant time intervals to be unreliable and should be excluded from the analysis. Thus, identification of these extended intervals based on a statistical method merits further investigation.

*MetaLonDA* is publicly available on the CRAN repository (`https://CRAN.R-project.org/package=MetaLonDA`).

# CHAPTER 5

# UTILIZING LONGITUDINAL GUT MICROBIOME TAXONOMIC PROFILES TO PREDICT FOOD ALLERGY VIA SPARSE AUTOENCODER AND LONG SHORT-TERM MEMORY NETWORK

## 5.1 Introduction

### Food Allergy and Relationship to the Microbiome

Food sensitization and allergy are characterized by an immunologic reaction caused by exposure to antigenic products derived from food, such as Ara h1 (peanuts) or tropomyosin (shellfish). The estimated prevalence of food sensitization and allergy in the US is 8% (136), with peak prevalence between the ages of one and two years old. Food sensitization is often associated with a positive reaction to skin prick testing or by increased levels of serum specific IgE to specific food antigens. Food allergy can be diagnosed by the clinical history of symptoms after food ingestion or by direct food challenge and monitoring of symptoms. Notably, not all individuals who are sensitized develop allergy, but the prevalence of food allergy is substantially higher for individuals with food sensitization. In turn, not all individuals with food allergy are sensitized to food allergens and thus serologic or skin testing alone is not sufficient for diagnosis of the food allergy. There is a need for more objective measures that have predictive value in diagnosing food allergy.

Food allergies are categorized into three groups: IgE-mediated, non-IgE-mediated, and mixed reactions. IgE-mediated food reactions are caused by the cross-linking of IgE on the surface of

mast cells or basophils by food proteins. This leads to rapid degranulation of these cells and release of histamine which is the primary mediator of IgE symptoms including urticaria, angioedema, and anaphylaxis, which can be life-threatening. These symptoms present acutely within minutes after the ingestion of food allergen. In contrast, non-IgE-mediated and mixed reactions present in a subacute to chronic time-frame and their mechanisms are less defined. Subacute symptoms associated with non-IgE food allergy are localized to the gastrointestinal tract, such as blood/mucus filled stools or vomiting, which can lead to chronic symptoms such as weight loss, dehydration, lethargy, and failure to thrive. Mixed reactions are characterized by food allergens exacerbating IgE-mediated diseases, such as atopic dermatitis. Thus, food allergy represents a spectrum of diseases that are currently diagnosed by subjective measurements during early life.

The increasing incidence of food allergy and other allergic diseases has been attributed to "westernized" life-styles, as prevalence of these diseases is substantially higher in the developed world. One over-arching theme as to why the incidence of allergy is increasing is the loss or disturbance in communities of micro-organisms that live on and in us (i.e., the microbiome). Importantly, differences in composition of the microbiome have been associated with food sensitization and/or IgE and non-IgE-mediated reactions (137; 138; 139; 140; 141), symptom resolution (142), and prevention and treatment (143; 144). This opens the door to develop more rigorous food allergy prediction models that are based on microbiome profiles of newborns, which could be used to predict food allergy and inform early intervention with novel therapies.

**Longitudinal Microbiome Studies**

Longitudinal microbiome studies have been widely utilized to study disease prognosis and microbial dynamics within an ecosystem such as the gut, lung, or kidney (30; 1; 70; 71; 72; 73). The exponential reduction in sequencing cost has resulted in the increase in popularity of longitudinal microbiome studies. Usually, a microbiome study is performed by sequencing the extracted DNA from a biological sample using either metagenomic shotgun (MGS) or 16S rRNA gene sequencing (145). Metagenomic reads are processed for each sample independently to construct the taxonomic and/or functional profiles (45; 56; 116; 117). Developing methods that predict the host phenotype from longitudinal microbiome samples comes with some challenges, e.g., variable sample collection times and an uneven number of time points along the subjects' longitudinal time-line, especially when samples are collected from human subjects. Hence, using standard prediction methods such as Hidden Markov Models (HMMs) (94) and Auto Regressive (AR) models (95) may not be suitable in these cases.

**Deep Learning**

Deep learning has revolutionized various fields by offering robust strategies to extract abstract nonlinear features that are refractory to traditional methods (146; 147). Multiple deep learning frameworks have been developed to predict phenotype from snapshot microbiome profiles (85; 90). On the other hand, a powerful approach to analyze temporal data is the Recurrent Neural Network (RNN). RNNs have shown success in different fields such as natural language processing (148) and speech recognition (149). Although in theory, the RNN can learn depen-

dent representation from distant events, it fails in practice due to problems with vanishing gradients (147). This problem occurs because the error loss is back-propagated through the deep network by multiplying the derivative of the utilized activation function, which is usually the sigmoid or hyperbolic tangent. The derivative of these activation functions is usually less than one. Hence, multiplying the error loss by many of these less than one numbers causes the vanishing gradient problem. Fortunately, Long Short-Term Memory (LSTM) networks, a modified variant of the RNN, have the ability to learn dynamic temporal behavior for a time sequence event and overcome the vanishing gradient problem that occurs in standard RNNs (150).

### 5.1.1 Problem Definition

In this work, we present a deep learning framework to predict food allergies in infants based on longitudinal gut microbiome profiles. In our model, we use all historical samples up to timepoint $t$ (features at timepoint $t$ included) from each subject to predict the phenotype (food allergy vs. non food allergy) at timepoint $t$. We hypothesize that adding the information from past microbiome profiles increases the predictive power of food allergy versus training a model with each timepoint independently. The proposed framework is based on a sparse autoencoder and LSTM network. The proposed model is flexible such that it can analyze subjects with a different number of time points.

## 5.2 Methods

### 5.2.1 Proposed Framework

Figure 23 illustrates an overview of our proposed framework to predict food allergy from longitudinal microbiome taxonomic profiles. It consists of two main components; an autoencoder and an LSTM network. The input to the autoencoder is a vector representing a normalized taxonomic profile of a subject's microbial sample. The aim of this module is to learn a compressed latent representation of a sample's microbial features. The learned latent representations are then passed to the LSTM module to learn temporal dependency between sequence profiles. Subsequently, the output from the last cell of the LSTM model is then fed to a softmax output layer where the prediction can be determined (e.g., food allergy vs. non food allergy). The methodology of obtaining the latent representation and learning temporal dependency is explained in details in the following sections.

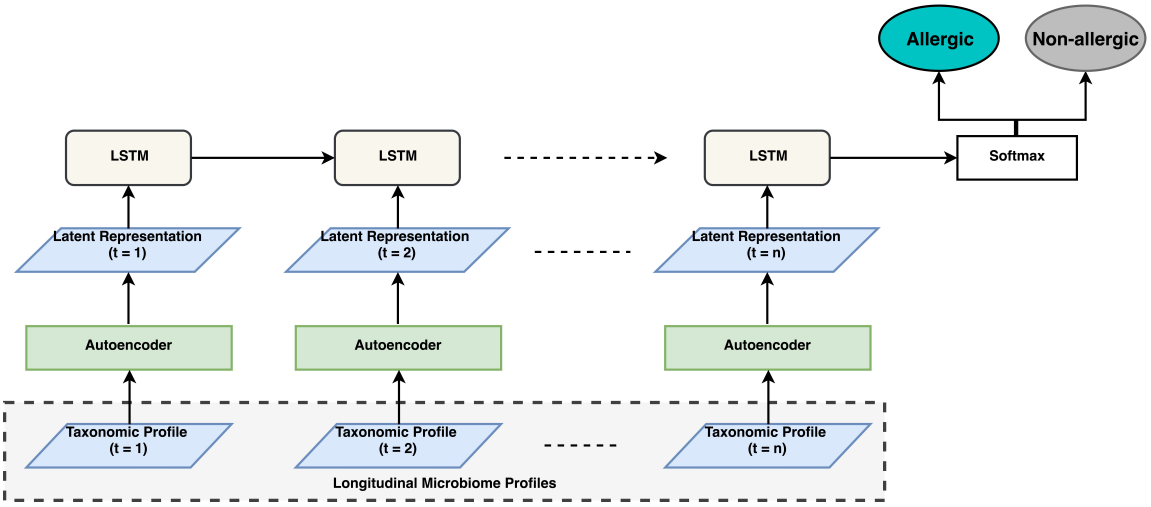



Figure 23: The proposed deep learning framework, where $n$ denotes the number of timepoints from each subject, which is not the same for all subjects.

### 5.2.2 Sparse Autoencoder

Autoencoders are neural network architectures that use unsupervised learning to extract compressed latent representations from unlabeled data. Figure 24 shows a schematic diagram of the autoencoder architecture that we used in our framework. The number of neurons in the input layer equals the number of raw features (220 in our case). It has three hidden layers with 100, 50, and 100 neurons in that order. The number of neurons for the output layer equals the number of raw features (220 in our case).

The output of layer $l$ follows (Equation 5.1), where $\mathbf{x}_l$ is the input feature vector, $\mathbf{W}_l$ is edge weight matrix, and $\mathbf{b}_l$ is the bias. We used the Rectified Linear Unit (ReLU) (Equation 5.2) as the activation function since it makes objective function converges faster (151).

$$f_l(\mathbf{x}_l) = \text{ReLU}(\mathbf{W}_l\mathbf{x}_l + \mathbf{b}_l) \tag{5.1}$$

$$\text{ReLU}(x) = \max(0, x) \tag{5.2}$$

The output of the autoencoder $\mathbf{x}'$ is calculated as in Equation 5.3 where $m$ is the number of layers of the autoencoder (4 layers in our framework)

$$\mathbf{x}' = F_{1\rightarrow m}(\mathbf{x}) = f_1 \circ .... \circ f_m(\mathbf{x}) \tag{5.3}$$

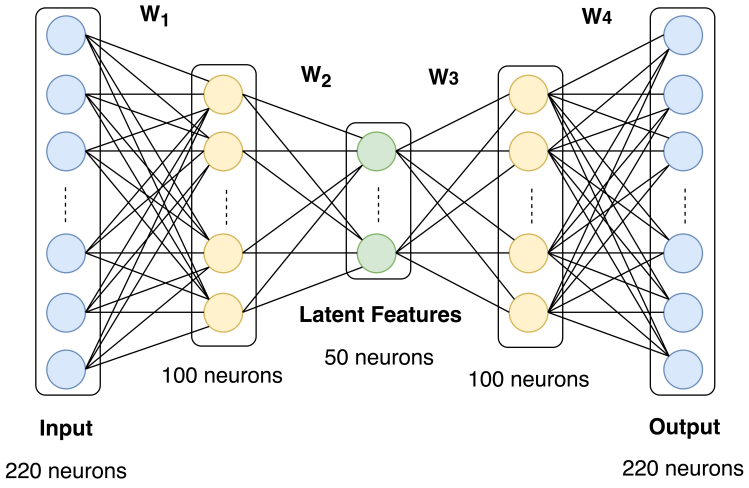



Figure 24: Autoencoder architecture.

The weights and biases of the autoencoder are learned by minimizing the error between the input $\mathbf{x}$ and the reconstructed input $\mathbf{x}'$ as shown in Equation 5.4 where $n$ is the number of datapoints (number of samples in our case).

$$\text{Loss}(\mathbf{x}, \mathbf{x}') = \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \mathbf{x}'_i \right\|^2 \tag{5.4}$$

In order to prevent over-fitting, an L2 regularization on the weights is added to the loss function with regularization parameter $\lambda$ where $m$ denotes the number of layers (Equation 5.5). Additionally, in order to enforce the sparsity on the hidden layer neurons, we added Kullback-Leibler (KL) divergence to the loss function (Equation 5.5) where $\rho$ denotes the sparsity parameter (152), $\beta$ is a parameter that controls the weight of the sparsity penalty term, and $k$ denotes the number of neurons on the latent representation layer.

KL-divergence is a standard function to measure the difference between two distributions. It is calculated as shown in (Equation 5.6), where $\rho'_j$ (Equation 5.7) is the average activation of neuron $j$ in the latent representation layer of the autoencoder across all timepoints and $a_j$ denotes the output of the activation function of neuron $j$. By putting KL-divergence into the loss function, latent representation neurons are forced to activate a small fraction of their neurons (152). This is useful to force the neurons to learn certain patterns of data which in turn increase their specificity in performance contrasted to the more general training.

$$\text{Loss}^{(1)}(\mathbf{x}, \mathbf{x}') = \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \mathbf{x}'_i \right\|^2 + \lambda \sum_{j=1}^{m} \|\mathbf{W}_j\|^2 + \beta \sum_{j=1}^{k} \text{KL}(\rho || \rho'_j) \tag{5.5}$$

$$\mathrm{KL}(\rho||\rho'_j) = \rho * \log \frac{\rho}{\rho'_j} + (1-\rho) \log \frac{1-\rho}{1-\rho'_j} \tag{5.6}$$

$$\rho'_j = \frac{1}{n} \sum_{i=1}^{n} a_j(\mathbf{x}_i) \tag{5.7}$$

The sparse autoencoder is trained via the backpropagation algorithm (152) to minimize the loss function ($Loss^{(1)}$). After the training is completed, the latent features are extracted and passed to the LSTM to train the model for phenotype prediction (food allergy vs. non food allergy).

### 5.2.3 Long Short-Term Memory (LSTM) Network

The LSTM network is a variant of the vanilla RNN that has the ability to learn long sequences (150). This ability is due to the presence of a memory, usually referred to as Cell state $C$ that stores long-term information so that errors will not be propagated through distant states. LSTM networks solve the two major problems of RNNs, the vanishing and exploding gradient descent problems. It accomplishes this by using 3 gates to control the cell state: forget, input, and output.

The forget gate controls the amount of information that should be forgotten from the previous cell state by analyzing the current input $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$. The sigmoid function $\sigma(.)$ in Equation 5.8 gives a value [0-1] representing the proportion of the previous cell state that should be retained.

$$\mathbf{f}_t = \sigma(\mathbf{W}_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f) \tag{5.8}$$

The input gate controls how much of the current input $\mathbf{x}_t$ should be used in training (Equation 5.9). Then, a list of new candidates for the cell state is calculated as in Equation 5.10.

$$\mathbf{i}_t = \sigma(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{b}_i) \tag{5.9}$$

$$\tilde{\mathbf{C}} = \tanh(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t-1} + \mathbf{b}_c) \tag{5.10}$$

Updating the cell state is performed as formulated in Equation 5.11

$$\mathbf{C}_t = \mathbf{f}_t\mathbf{C}_{t-1} + \mathbf{i}_t\tilde{C} \tag{5.11}$$

To calculate the output of the LSTM $\mathbf{h}_t$, usually called the hidden state, that is passed to the next sample in a sequence, we first determine which part of the cell state should be outputted by the following (Equation 5.12) where $\mathbf{o}_t$ denotes the output of the output gate. Subsequently, multiplying $\mathbf{o}_t$ by the squashed cell state $\mathbf{C}_t$ via tanh function (Equation 5.13). $b_f$, $b_i$, $b_c$, and $b_o$ are bias terms.

$$\mathbf{o}_t = \sigma(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{b}_o) \tag{5.12}$$

$$\mathbf{h}_t = \mathbf{o}_t * \tanh(\mathbf{C}_t) \tag{5.13}$$

Since the number of samples for each subject is not identical, we extract the LSTM output of the last sample of each subjects' sequence. This output is then fed into a dense layer with the sigmoid activation and with the dimension of (num_hidden_neurons x num_classes) (64x2 in our case) (Equation 5.14).

$$\mathbf{z}_t = \sigma(\mathbf{W}_z \mathbf{x}_z + \mathbf{b}_z) \tag{5.14}$$

$\mathbf{z}_t$ is then fed to a softmax function in order to give an output probability for each class (Equation 5.15). The class with the highest probability is considered the predicted class.

$$\hat{\mathbf{y}}_t = \text{softmax}(\mathbf{z}_t) \tag{5.15}$$

Since our main target in this project is to predict the phenotype (food allergic vs. non food allergic), we used the cross-entropy between target $y_t$ and predicted output $\mathbf{y}'_t$ (Equation 5.16) to be the loss function.

$$\text{E}(y_t, \hat{y}_t) = -y_t \log(\hat{y}_t) - (1 - y_t) \log(1 - \hat{y}_t) \tag{5.16}$$

To prevent over fitting, we used L2 regularization in the loss function (Equation 5.17), where $J = \{f, i, c, o, z\}$. Similar to the autoencoder, we used the back-propagation algorithm to minimize the loss. Here, $N$ denotes the number of data sequences (subjects in our case)

$$\text{Loss}^{(2)}(y, \hat{y}) = \sum_{i=1}^{N} \text{E}(y_{t_i}, \hat{y}_{t_i}) + \lambda \sum_{j \in J} \|W_j\|^2 \tag{5.17}$$

## 5.3 Experiments

**Dataset: DIABIMMUNE**

In order to evaluate our proposed model, we used the longitudinal microbiome profiles from the DIABIMMUNE project (`https://pubs.broadinstitute.org/diabimmune`), a study that aimed to characterize host-microbe immune interactions contributing to autoimmunity and allergy. These diseases were evaluated in relationship to the hygiene hypothesis, which states that subjects with high bacterial exposure tends to have a more powerful immune system and fewer allergic diseases (1). To test this hypothesis, stool samples were collected from 222 infants (74 from Russia, 74 from Finland, and 74 from Estonia) from birth to 3 years of age. At the time of stool sample collection, various food allergen-specific Immunoglobulin E (IgE) levels were measured for each subject, and based on a predefined threshold, infants were annotated as allergic or non-allergic to the corresponding food allergen. Figure 25 shows the breakdown of the number of subjects with milk, egg, or peanut allergic responses. It is clear that the prevalence of the allergies is highest in Finland and lowest in Russia with Estonia intermediate. This is aligned with the hygiene hypothesis. For the purpose of evaluating our framework, we labeled subjects as food allergy positive if they are allergic to milk, eggs, or peanuts (Figure 25).

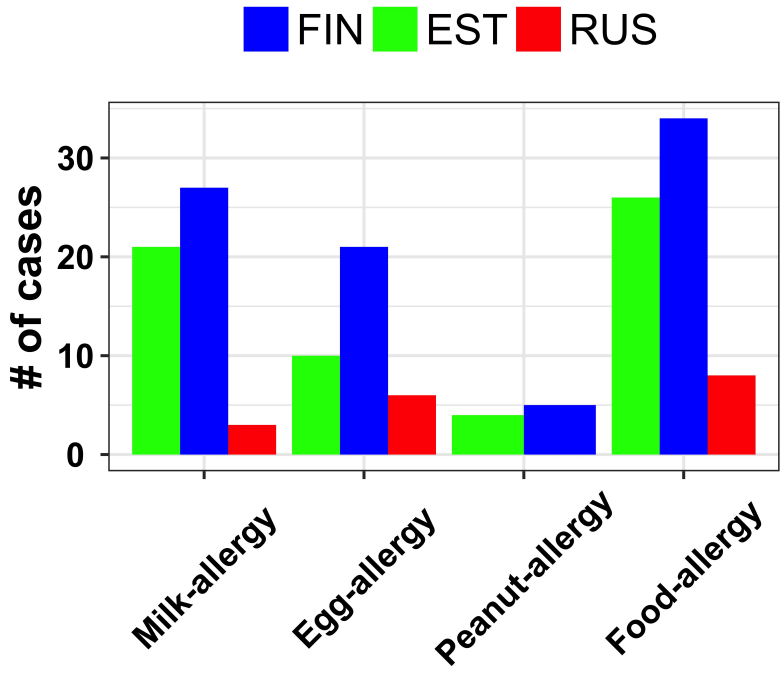



Figure 25: Number of subjects allergic to milk, egg, and peanut within the DIABIMMUNE cohort after filtering out missing data. The Food-Allergy group is the summation of milk, egg, and peanut allergy.

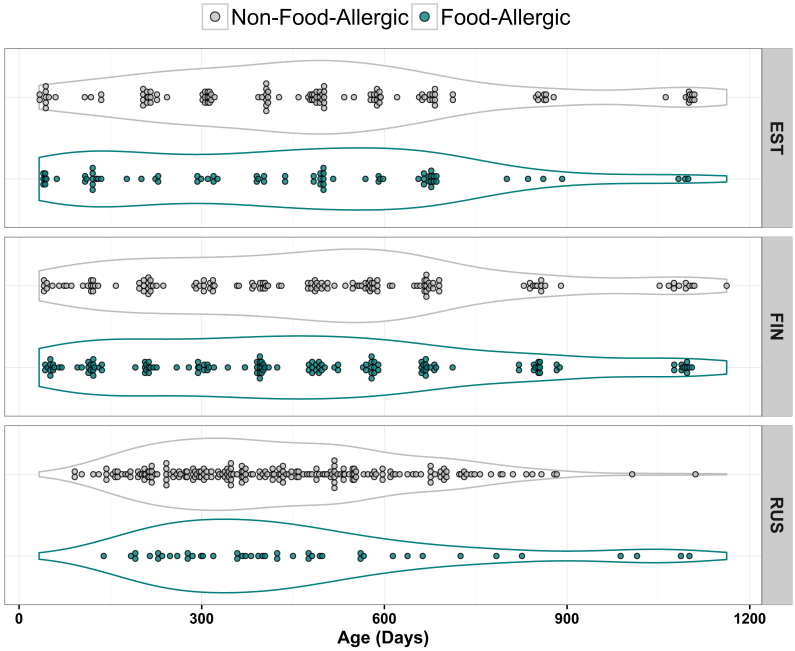



Figure 26: Time distribution of 731 (281 from 71 Finnish, 197 from 70 Estonian, and 253 from 54 Russian) stool samples sequenced using MGS from the DIABIMMUNE project. The collected samples have various forms of inconsistencies, such as different numbers of samples per subject (min = 1, max = 13).

As a preprocessing step, we removed all samples without a food allergy class label, i.e., missing data, resulting in 731 samples from 195 subjects (281 from 71 Finnish, 197 from 70 Estonian, and 253 from 54 Russian). The 195 subjects are categorized as 68 food allergic and 127 non food allergic. Figure 26 shows the distribution of time points of the samples collected from each class from each country.

As shown, these samples suffer from all forms of variability such as a different number of subjects per phenotypic group (food allergy vs non food allergy), a different number of samples per subject, and samples not collected at consistent time points

These samples have been sequenced using MGS sequencing. As previously described in (1), reads from the 731 sequenced samples were quality-controlled by filtering out low-quality reads, short reads ($<$ 60 bp), and human reads. Taxonomic profiles were constructed using *MetaPhlAn2* (100). The number of reads mapped to each taxonomic feature was then normalized to the reads per kilobase per million (RPKM) sample reads to correct for bias due to differences in genome size and sequencing depth. The aggregated taxonomic profiles of all 731 samples revealed 220 genera.

**Benchmarking Procedure**

We benchmarked the proposed framework against other predictive models, such as Support Vector Machine (SVM), Random Forests (RF), and Least Absolute Shrinkage and Selection Operator (LASSO). In our evaluation, we benchmarked two aspects: (1) the effect of extracting and using the latent representation versus using raw features on the prediction, and (2) the effect on the prediction of learning temporal dependency between the sequence of samples, as in LSTM, versus learning from each sampling independently using methods such as SVM, RF, or LASSO.

We implemented our autoencoder-LSTM model using *Tensorflow* (v1.6.0) (153). We trained the autoencoder and LSTM separately. The autoencoder consists of 220, 100, 50, 100, and 220 neurons for the input layer, first hidden layer, second hidden layer (latent representation), third hidden layer, and output layer, respectively. We trained the model with the back-propagation algorithm using the Adaptive Moment Estimation (*Adam*) optimizer (154) with a learning rate of 0.001 and batch size of 5. The model was trained for 50 epochs, and the best model was saved based on the loss value on the test set. For L2 regularization we used $\lambda = 0.05$. For the sparsity constraint, we used $\rho = 0.01$ and $\beta = 3$. For the LSTM module, we used 64 neurons for the LSTM hidden neurons. Similar to the autoencoder, the LSTM model was trained with *Adam* Optimizer with a learning rate of 0.001 and batch size = 5.

The RF, SVM, and LASSO models were all trained using Python's scikit-learn package (`http://scikit-learn.org`). The RF models were trained by setting a maximum of 500 trees. All other parameters were left as the default values. The SVM models were trained using an exhaustive grid search with 5-fold cross-validation over the linear and Gaussian kernels, using the parameters 1, 10, 100, 1000 for error terms and the parameters 0.001, 0.0001 for $\gamma$ values in Gaussian kernels. The LASSO models were trained using iterative fitting with 5-fold cross-validation for the error term $\alpha$ over a set of 50 numbers, evenly log-spaced between 4-10 and 0.5-10.

In the case of RF, SVM, or LASSO, the prediction of the last timepoint of each subject is determined based on a majority voting scheme from the predicted phenotype of all subject's timepoints. This strategy ensures a fair comparison with LSTM, which uses all of a subject's timepoints to predict the phenotype of the last timepoint.

**Evaluation Metrics**

We used 10-fold cross validation to evaluate all of the methods. Given the fact of the imbalanced data (68 allergic and 127 non-allergic) we up-sampled samples from allergic infants and down-sampled samples from non-allergic infants. Up/down sampling was achieved by splitting the training set ($N = 195$) into allergic ($N_a = 68$) and non-allergic ($N_{na} = 127$) subjects. For $l = \frac{N}{10}$ for 10-fold, we calculated the proportion of allergic ($N_a^l = \frac{N_a}{10} = 7$) and non allergic ($N_{na}^l = \frac{N_{na}}{10} = 13$). In the case of a balanced dataset, we should have 10 allergic and 10 non-allergic. To correct for our imbalanced dataset, we up-sampled allergic subjects by sampling $(0.5 * l - N_a^l)$ samples from $N_a^l$ and concatenated them to $N_a^l$. Similarly, we down-sampled non-allergic subjects by sampling $(N_{na}^l - 0.5 * l)$ samples from $N_{na}^l$ and removed them from $N_{na}^l$. This ensures the proper training and more meaningful evaluation metrics.

Various performance metrics were calculated such as $Sensitivity = \frac{TP}{TP+FN}$ and $Specificity = \frac{TN}{TN+FP}$. These metrics have been used to obtain area under the received operating curve (AUROC).

## 5.4 Results and Discussion

### 5.4.1 Analyze the Latent Representation

The first aspect we investigated the similarities of the latent representation to the raw features. Figure 27 shows PCA of the latent features and raw features labeled by country, phenotype, and age. Figure 27.A shows a distinction between samples from Finland and Russia, while the Estonian samples overlap between them. This is aligned with the rate of allergy between the three

countries where Russia and Finland represent the two extremes, and Estonian infants are in between (Figure 25). The distinction is less apparent using the raw features (Figure 27.D). However, changes in age and phenotype are more discriminative using the latent representation. previous studies indicate that the infants' gut microbiome evolves over time (9). At birth, gut microbiomes are similar but then start to differentiate by colonizing different types of bacteria either from breast milk or environmental sources.

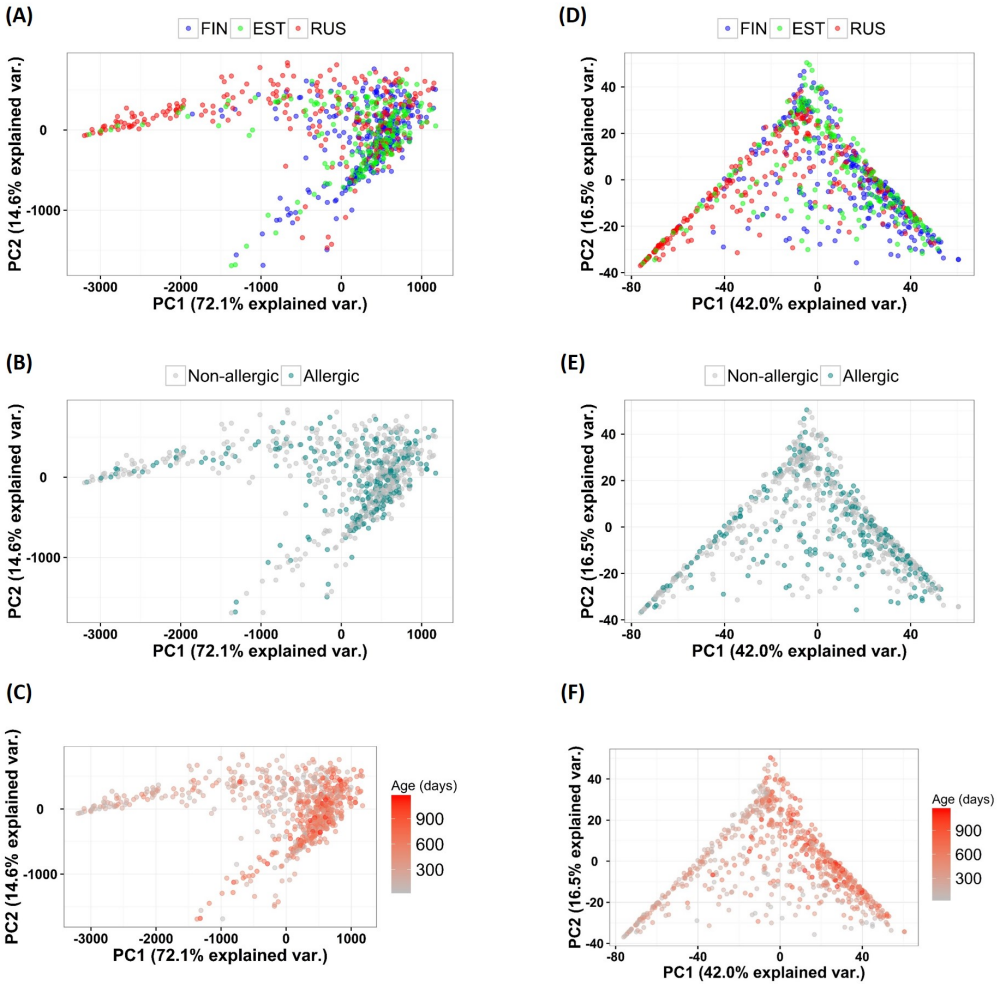



Figure 27: PCA of the latent representation (left panel, A-C) vs. raw features (right panel, D-E). The first row is labeled by country (Russia, Estonia, and Finland), the second row is labeled by phenotype (food allergic vs non food allergic), the third row is labeled by age (0-3 years).

Figure 28 shows the trajectory of the loss by training the autoencoder with more epochs. The loss progressively decreases by more training epochs until it stabilizes after 40 epochs. The smallest loss achieved was 25.8. Ideally, it should be zero or very small value, but this is due to the regularization we put on the autoencoder to prevent over-fitting.

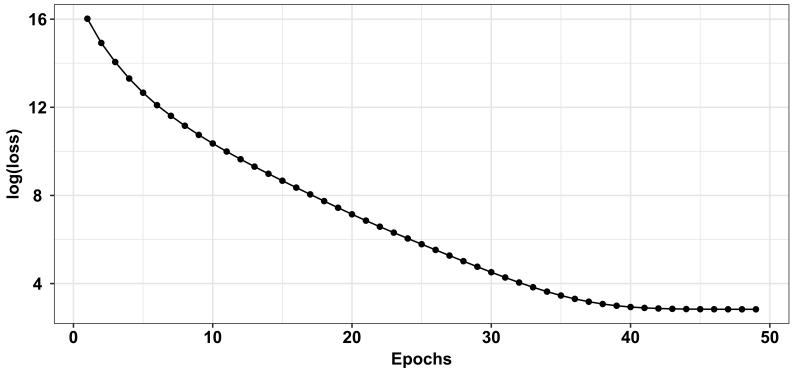



Figure 28: Reduction in autoencoder loss function ($Loss^{(1)}$) with the increasing number of training epochs.

### 5.4.2 Evaluation of Prediction

Subsequently, we evaluated how our proposed method compares with the commonly used classification methods. Figure 29 shows a violin plot of the area under the ROC curve (AUROC) for four classifiers; LSTM, RF, SVM, and LASSO. For each classifier, we evaluated two types of input features; latent features which extracted from the trained autoencoder and raw taxonomic profile features. The violin plot shows the distribution of AUROC for each model after running 10 times 10-fold cross-validation experiments. Samples are shuffled after each 10-fold to test the robustness of each classifier.

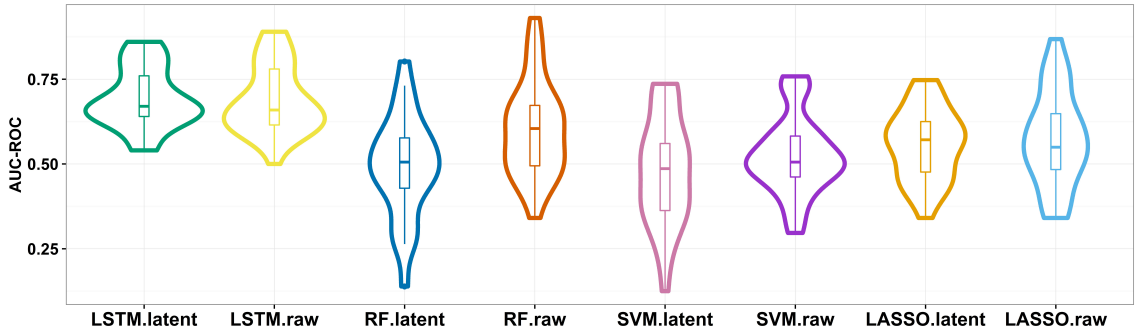



Figure 29: Evaluation of area under ROC curve (AUROC) for the proposed model versus baseline models. In this figure we evaluated four classifiers; LSTM, RF, SVM, and LASSO. For each classifier, we evaluated two types of input features; latent features which extracted from the trained autoencoder and raw taxonomic profile features. The violin plot shows the distribution of AUROC for each model after running 10 times 10-fold cross-validation experiments. The samples are shuffled after each 10-fold cross-validation to test the robustness of each classifier.

TABLE VI: Evaluation of area under ROC curve (AUROC) for the proposed model versus baseline models. P-values are calculated using Mann-Whitney U test between LSTM-latent versus each corresponding method.

| | **mean (sd)** | **p-value** |
|---|---|---|
| **LSTM-latent** | 0.69 (0.08) | - |
| **LSTM-raw** | 0.67 (0.10) | 0.60 |
| **RF-latent** | 0.49 (0.15) | $5.24 * 10^{-7}$ |
| **RF-raw** | 0.60 (0.14) | 0.0027 |
| **SVM-latent** | 0.48 (0.16) | $1.47 * 10^{-6}$ |
| **SVM-raw** | 0.52 (0.11) | $4.83 * 10^{-7}$ |
| **LASSO-latent** | 0.56 (0.10) | $1.03 * 10^{-5}$ |
| **LASSO-raw** | 0.57 (0.14) | 0.0001 |

LSTM shows superior performance compared to the other classifiers, supporting the concept that learning a sequence of events increases the prediction power. Although the median of LSTM trained on latent features and raw features were similar (0.67 vs. 0.65), the prediction was more

stable with the latent representation (Table VI). This highlights the benefit of using compressed features rather than using all raw features which may be redundant information. Alternatively, all other classifiers that do not take time sequence data into consideration perform poorly compared to LSTM. Table VI and Figure 29 show that the AUROC of RF with raw features comes second to LSTM, but it suffers from an unstable performance (sd=0.14).

### 5.4.3 Execution Time

The execution time of training LSTM on both latent representation and raw features is comparable and depends on the number of epochs used and batch size. 10-times 10-fold validation was 92 minutes on the DIABIMMUNE dataset that we used given all the parameters stated above. Alternatively, using raw and latent features, it took 13 and 12 minutes for SVM, 9 and 5 minutes for RF, and 2 and 1 minutes for LASSO, respectively. The prediction time is linear for all algorithms. The evaluation was conducted on a MAC machine with 2.5 GHz Intel Core i7 processor and 16 GB 1600 MHz RAM. No GPU was used for training of the LSTM.

## 5.5 Conclusion

Food allergy is usually difficult to diagnose at young ages, and the inability to diagnose patients with this atopic disease at an earlier age may lead to severe complications due to the lack of treatment. In this work, we have developed a deep learning framework that has the capacity to predict food allergy from longitudinal microbiome profiles. The framework is based on sparse autoencoder and Long Short-Term Memory (LSTM) networks. Sparse autoencoder is devised to extract potential latent structure in microbiome prior to LSTM training. We tested the framework on the DIABIMMUNE dataset (`https://pubs.broadinstitute.org/diabimmune`),

a study that aimed to characterize host-microbe immune interactions contributing to autoimmunity and allergy. Our results demonstrate the increase in predictive power of our proposed model compared to SVM, Random Forest, and LASSO regression.

Although our deep learning framework shows the potential to predict allergic phenotypes from a sequence of microbiome profiles and outperforms other classical methods, it does not reach a prediction level for optimal clinical utilization. This is mainly due to the nature of the training dataset that we used to train our model. The DIABIMMUNE dataset is small (195 subjects) and each subject has few time points (6 on average). With the current reduction in sequencing costs, we anticipate that multiple large longitudinal microbiome projects will be available which in turn could be used to train models like ours for better prediction power.

# CHAPTER 6

# LOWER AIRWAY MICROBIOME DYNAMICS AS A PREDICTOR OF BRONCHIOLITIS OBLITERANS SYNDROME AFTER PEDIATRIC LUNG TRANSPLANTATION IN CYSTIC FIBROSIS

## 6.1 Introduction

Cystic fibrosis (CF) is a highly prevalent autosomal recessive disease associated with reduced life expectancy despite advances in comprehensive care (155; 156; 157). The pulmonary disease remains the primary cause of poor outcomes in those affected by CF and lung involvement is evident even in healthy-appearing infants with as many as 1/3 demonstrating tomographic evidence of bronchiectasis (156; 158; 159). In CF, defective mucociliary clearance, recurrent bacterial infections, and chronic suppurative inflammation of the airways results in end-stage pulmonary disease and accounts for about 80% of deaths (157; 160; 161). In end-stage pulmonary disease, lung transplant (LTx) is considered the final therapeutic option to improve quality of life and prolong survival. Thus, LTx is usually reserved for those with forced expiratory volume (FEV1) $<$30% predicted, poor performance on 6-minute walk test ($<$400m), pulmonary hypertension, or other clinical signs related to end-stage CF pulmonary disease (162; 163).

Overall, median survival in pediatric patients who undergo bilateral lung transplant is about 5.6 years, and functional status with good quality of life is maintained during the first 2-3 years after transplantation (163; 164; 165). Despite increased survival and improved outcomes, LTx

is still characterized by the highest rates of rejection among solid organ transplants. In LTx, chronic lung allograft dysfunction (CLAD) remains the principal cause of poor long-term survival and is common in those who survive beyond 1 year after transplantation (164; 166; 167). Specifically, bronchiolitis obliterans syndrome (BOS), manifesting as a sustained progressive decline in FEV1, is the cause of over 50% of morbidity and 45% of mortality observed at 5 years post-transplantation (164; 167; 168).

To date, the complex immune interactions that contribute to BOS development and reduced long-term survival are poorly understood (169; 170; 171; 172; 173). It is possible that persistent exposure to organisms within microbial communities in this population drives neutrophil activity and promotes the lower airway inflammatory environment that is associated with allograft damage and development of BOS (174; 175; 176). Correspondingly, we recently reported that the presence of a complex lower airway microbial community, enriched with Proteobacteria and Bacteroidetes and relatively devoid of Actinobacteria, is associated with increased risk of BOS in a post-LTx adult cohort of subjects with end-stage lung disease without CF or bronchiectasis (177). Additionally, in CF, others have described reemergence of lower airway pre-transplant microbiota, predominance of recognized CF bacterial pathogen (e.g. Pseudomonas and Staphylococcus), and distinct post-transplant clinical manifestations associated with BOS (178; 179; 180; 181; 182). Hence, in CF, development of BOS during the post-transplant period may be related to the respiratory system's constant exposure to extrinsic and intrinsic microbes (173; 183; 184).

## 6.2 Hypothesis

In this prospective, observational, longitudinal study, we hypothesize that changes in the composition of the pulmonary bacterial microbiome in CF are associated with the progression BOS development and poor overall survival in lung transplant recipients.

## 6.3 Methods

### Ethics Statement

The study protocol and clinical investigation were approved by the Washington University/St. Louis (WASHU) Children's Hospital Cystic Fibrosis Center Institutional Review Board (IRB # 201105467 and 201311048) and the University of Illinois at Chicago (UIC) Institutional Review Board (IRB #2015-0116). Parents, guardians, or other legally authorized representatives provided oral and written informed consent to allow child participation in this study in accordance with the principles expressed in the Declaration of Helsinki.

### Identification of Study Patients

Lung transplant recipients at WASHU were enrolled in a prospective observational registry that included the collection of bronchoalveolar lavage fluid (BAL) samples for research purposes at the time of standard of care bronchoscopies. Standardized medical record abstraction was performed in parallel to obtain demographic and clinicopathologic variables related to transplant outcome. Twelve representative subjects were selected from the registry for this study. BOS was defined as a sustained drop in forced FEV1 by at least 20% from the average of the 2 best post-

transplant FEV1 measurements (167). Subjects were grouped according to presence or absence of BOS, those in the nonBOS group remained without evidence of BOS for at least 3 years following lung transplantation.

**Bronchoscopy**

Lung transplant recipients at WASHU underwent surveillance bronchoscopy 1 day, 1 week, 1 month, 2 months, 3 months, 6 months, 1 year, 2 years, and 3 years post-transplant, and when clinically indicated. BAL was collected in usual fashion. For this longitudinal study, eligible subjects had a surveillance bronchoscopy with the corresponding research BAL sample available in our biorepository.

**DNA Isolation, Library Construction, and Sequencing**

Around 750 µL of BAL was processed for DNA isolation. Samples were spun for 10min at 1000rpm to remove lymphocytes, and the supernatant was centrifuged at 22,000 rpm for 2h at 4°C (Thermo Fisher Scientific Sorvall WX Ultra 80). The pellet was saved, and supernatant was processed through 3KD amicon Ultra 4 for 30min at 4000g. The retentate was saved and used to resuspend the pellet and was subjected to DNaseI treatment. The contents were subjected to enzymatic treatment using lysozyme, followed by Proteinase K treatment. DNA was purified using the Qiagen QIAmp® MinElute® Virus Spin Kit (Qiagen cat. no. 57704) and subsequently quantified using the Qubit 2.0 fluorometer (Invitrogen). Since, the DNA was in low amounts, it was amplified using the WGA Picoplex kit (New England Biolabs). The amplified DNA was used for library preparation using NEBNext DNA Library Prep Mastermix Set for Illumina according to the manufacturer protocol (New England Biolab). The quantity and quality was anaylsed by

Qubit and Agilent 2100 Bioanalyzer. The libraries were sequenced on the Illumina MiSeq using v3 600 cycle kit for 301 PE read length.

**Statistical Analysis**

Statistical analyses were performed using R 3.3.0 software (The R Foundation for Statistical Computing). The null hypothesis of continuous responses with no prior distributional information was tested using *Mann-Whitney U* test (185). Generalized linear mixed models (GLMMs) (186) were constructed to test association between clinical variables and BOS development in our longitudinal study.

**Taxonomic Profiling and Diversity Measures**

Metagenomic sequencing reads were processed with a custom pipeline that is hosted on the UIC computer cluster "Extreme". Briefly, we first performed quality control by filtering out all low-quality reads ($<$25 on Phred quality score), short reads ($<$100 bp), or any human reads from the generated sequences. The remaining high-quality microbial short-reads were assembled into longer contigs using *MetaVelvet* (187). Subsequently, microbial taxonomic profile for each sample was constructed using *WEVOTE* (45) with *Kraken* (46), *Clark* (57), and *BLASTN* (41) as base classifiers for the *WEVOTE* platform. Fisher's index was chosen as a measure of diversity because it accounts for microbial evenness and richness and utilized to summarize the microbial community in each sample. The *Phyloseq* R-package (188) was utilized to aggregate and summarize taxonomic profiles and metadata for all samples.

**Longitudinal Analysis**

Microbial taxa were normalized using the median-of-ratios method (121) to remove the batch effect between multiple runs. Subsequently, differentially abundant taxa and their significant time intervals were identified using *MetaLonDA* R-package (115). In performing the *MetaLonDA* analysis, the study period was divided into 100 intervals with 1000 permutations to construct the empirical distribution. Specifically, the negative binomial distribution was used to identify time intervals of taxonomic features whereas locally weighted scatterplot smoothing (*LOWESS*) was used to identify time intervals of diversity measures and pulmonary function (FEV1%). Microbial taxa and time intervals were considered statistically significant if a p-value$<$0.05 was observed after correction for the multiple testing using Benjamini-Hochberg (BH) (113).

**Functional Annotation**

*HUMANn2* (117) (`http://huttenhower.sph.harvard.edu/humann2`) was used to construct metagenomic functional profiles for each sample with default parameters. We used the UniRef90 and MetaCyc databases (189; 190) to identify gene families and metabolic pathways, respectively. Functional profiles were normalized to counts per million (CPM) mapped reads.

## 6.4 Results and Discussion

### 6.4.1 Clinical characteristics

The study involved subjects who underwent cadaveric bilateral lung transplant for end-stage lung disease related to cystic fibrosis at WASHU (n=12). In addition to clinical data, an average of

7.4 bronchoalveolar lavage (BAL) samples (range 4-9) for a total of 83 were collected per subject from the time of LTx until development of bronchiolitis obliterans syndrome (BOS) or discharge from WASHU. Actual sample collection timepoints distribution is shown in Figure 30.A. Samples from subjects without BOS (nonBOS) are shown in blue (n=53) whereas those with BOS are shown in red (n=30). Initial BAL samples were collected significantly later in the BOS group ($\bar{x} = 31 \pm 12.8$ days) compared to the BOS-free group ($\bar{x} = 10.7 \pm 13.5$ days) after lung transplant (*Mann-Whitney U* test p-value=0.022). No difference was observed with regards to timing of final sample between the BOS and nonBOS group ($\bar{x} = 393.6 \pm 73$ vs. $608.3 \pm 233$ days, respectively; *Mann-Whitney U* test p-value=0.1).

Baseline subject characteristics comparing BOS versus nonBOS groups are described in Table 1. No differences were observed with regards to gender, body mass index (BMI), pre-transplant diagnosis, recipient age at time of transplant, FEV1 (% predicted) prior to transplantation or right lung perfusion after LTx. Longitudinal differential analysis by *MetaLonDA* shows significant reduction in FEV1 (%) predicted, consistent with development of BOS, is noted to start from day 427 to day 455 (Figure 30.B). Conversely, no significant difference throughout the study period was identified by *MetaLonDA* for other pulmonary function measures, and FVC (%) predicted and FEV1/FVC.

Choice of induction immunosuppression regimens consisting either of tacrolimus (T), mycophenolate (MMF), prednisone (P) or T, MMF, P, and daclizumab (DAC) at acquisition of the first BAL sample did not differ between BOS and nonBOS groups ($\chi^2$ p-value=0.68). Transbronchial biopsy was performed on all subjects within the BOS group and 86% (6/7) nonBOS subjects dur-

ing the first 50 days post-transplant. Comparison of histopathologic acute cellular rejection (ACR) grades A and B proved to be similar between groups ($\chi^2$ p-value=0.68).

TABLE VII: Baseline characteristics of BOS and nonBOS groups. No clinical parameter demonstrated a statistically significant difference by Mann-Whitney U test for continuous data or chi-square test for categorical data. Continuous variables are reported as range (average ± standard deviation) and categorical variables as frequency (percentage). Definition of abbreviations: CF: Cystic fibrosis; BMI: Body mass index; FEV1: Forced expiratory volume in 1 second; ΔF508: Delta F508 mutation.

| | **BOS (5)** | **nonBOS (7)** | **p-value** |
|---|---|---|---|
| **Gender** | | | 0.97 |
| **Male** | 3 (60%) | 4 (57.1%) | |
| **Female** | 2 (40%) | 3 (42.9) | |
| **Ethnicity (Caucasian)** | 5 (100%) | 7 (100%) | 1 |
| **BMI (kg/m2)** | 14.9 - 20.8 (17 ± 2.6) | 13.4-19.9 (16 ± 2.3) | 0.43 |
| **Pre-transplant diagnosis (CF)** | 5 (100%) | 7 (100%) | 1 |
| **Genotype** | | | 0.67 |
| **Homozygous ΔF508** | 1 (20%) | 5 (71.5%) | |
| **Heterozygous ΔF508** | 2 (40%) | 2 (18.5%) | |
| **NA** | 2 (40%) | 0 (0%) | |
| **Recipient age at transplantation (years)** | 13-17 (16 ± 1.8) | 11-16 (14 ± 1.8) | 0.5 |
| **FEV1 (%) predicted before transplantation** | 20-34 (25.2 ± 5.7) | 17-33 (24 ± 6) | 0.52 |
| **Bilateral lung transplantation** | 5 (100%) | 7 (100%) | 1 |
| **Lung allograft type (cadaveric donor)** | 5 (100%) | 7 (100%) | 1 |
| **Right lung perfusion (%)** | 32-78 (60 ± 19) | 46-70 (58 ± 8) | 0.69 |

BOS was diagnosed on average at 222.52 days (range 276-475) post-transplant. Expectedly, BOS subjects demonstrated a lower FEV1% compared to nonBOS subjects ($\bar{x} = 55.5\%$ vs 83.2%, respectively; *Mann-Whitney* test p-value=0.23) at the time of final BAL sample collection. No differences were observed with regards to acute cellular rejection (ACR) grade, immunosuppressive regimen or antibiotic use between BOS subjects and nonBOS subjects at the time of final bronchoscopy.

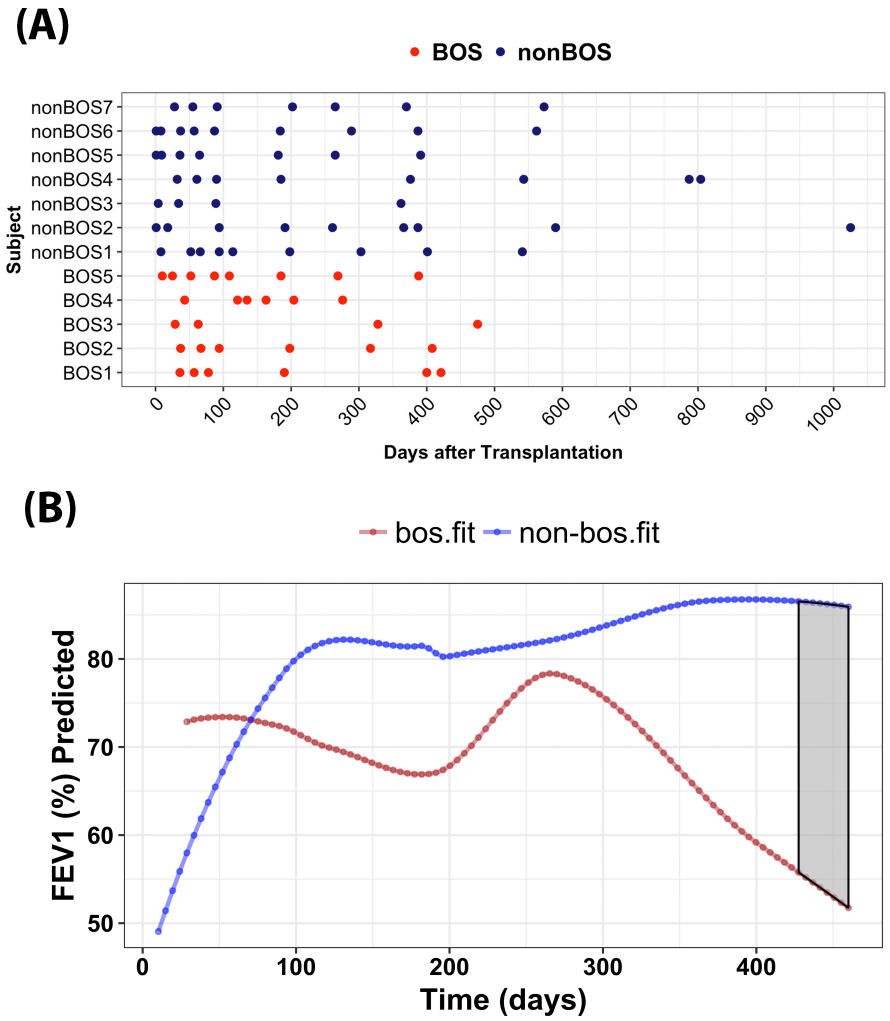



Figure 30: (A) Shown are the timepoints of bronchoalveolar lavage (BAL) sample collection (in days) per subject. (B) Longitudnal differential analysis by *MetaLonDA* demonstrates the time interval during which FEV1 (%) predicted is significantly is different between the BOS and nonBOS groups. The red spline represents pulmonary function of the BOS group over time (days) and blue spline is representative of the nonBOS group. The gray shaded area represents the significant time interval during which differences between groups were observed.

BOS was reported in 41.67% (5/12) subjects, while 58.33% remained nonBOS at the end of their follow-up period. 3 subjects (25%) died. Two subjects (16.67%), 1 male (subject BOS4) and 1 female (subject BOS3), died from complications of BOS during follow-up at 291 and 680 days after LTx. Among nonBOS subjects, one female (subject nonBOS6) was reported to have died from acute respiratory distress syndrome 2,111 days (5.8 years) after LTx and one male (subject nonBOS1) was reported to have developed lymphoproliferative disorder on day 1,156 (3.2 years) post-transplant. Both events in the nonBOS group occurred after WASHU follow-up had concluded. A single subject in the nonBOS group (subject nonBOS4) received an HLA mismatched allograft and was successfully discharged from the WASHU on post-transplant day 1,588 (4.35 years).

Antibiotic regimens at the time of first BAL sample collection were variable in both groups. The use of trimethoprim-sulfamethoxazole (TMP-SMX) prophylaxis, anti-pseudomonal and anti-staphylococcal antibiotics did not differ between BOS and nonBOS groups ($\chi^2$ test p-value=0.92, 0.30, and 0.56; respectively). Notably, macrolides were not part of the initial antibiotic regimen for any subject in either group.

Conventional bacterial cultures performed on BAL samples at baseline were reported as negative for bacterial growth in all subjects who progressed to develop BOS, whereas 2 nonBOS subjects were found to have bacterial growth (1 subject with S. *aureus*, and the other with co-occurring S. *aureus* and P. *aeruginosa*). A single BOS subject had conventional BAL virology positive for parainfluenza virus. At the time of final bronchoscopy, conventional cultures were positive in 2/5 subjects with BOS (both P. *aeruginosa*) and 2/7 nonBOS subjects (one culture was found to have concomitant infection with P. *aeruginosa*, S. *aureus*, and E. *coli*, and the other was

positive for H. *influenza*). None of the final positive cultures were observed in subjects who later died. Additionally, BAL cell counts were assessed and compared. There were no differences noted in BAL total, absolute neutrophil or absolute lymphocyte counts (p-value$>$0.05) between the BOS and nonBOS groups.

### 6.4.2 Lower Airway Microbial Community Structure and Diversity

Bacterial community structure of the lower airways based on metagenomic sequencing data acquired from BAL was assessed. An average of 2,211,236 reads were identified among all 83 BAL samples (Figure 31.A). After removal of low quality and short reads, an average of 1,584,636 and 1,754,157 raw reads and were identified among the 30 and 53 BAL samples collected from subjects who developed BOS versus those who remained nonBOS (*Mann-Whitney* test p-value=0.11). Reads mapped against the human genome were subsequently filtered out and an average of 5,716 microbial reads were identified across all samples. There was no difference in the percent of reads mapped to microbial genomes in the BOS and nonBOS groups (Figure 31.B).

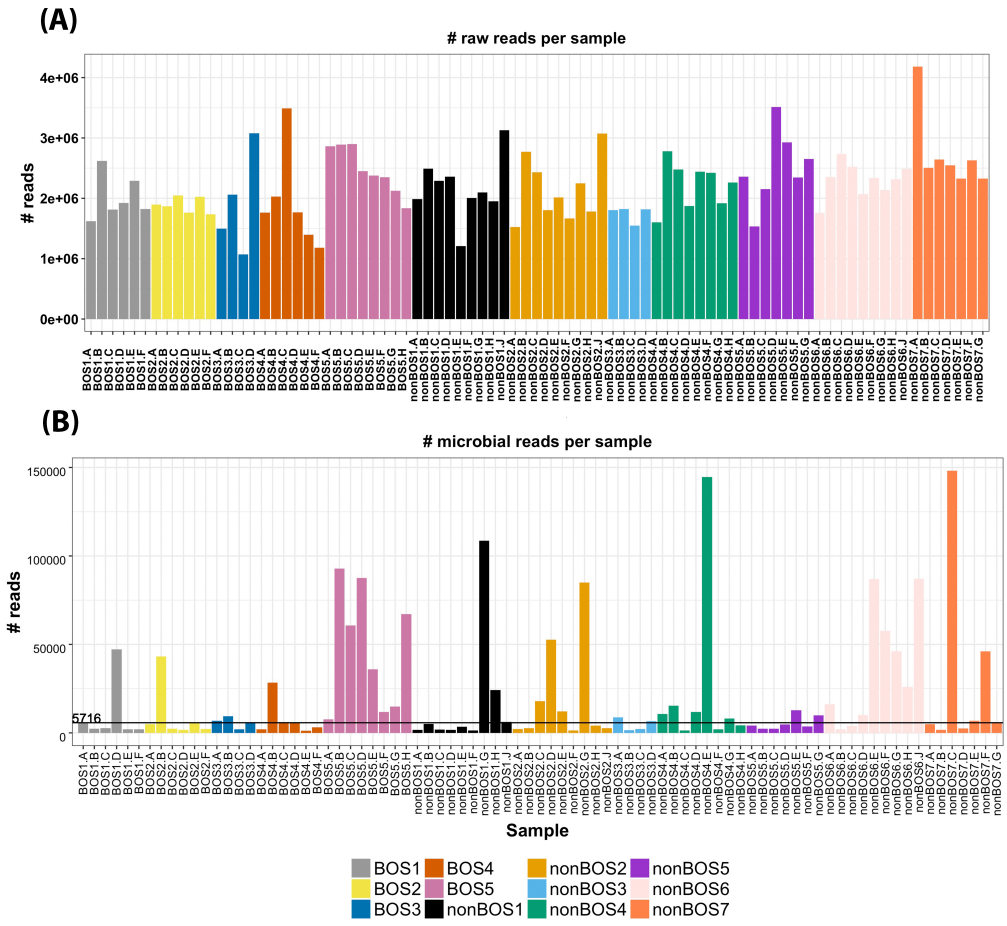



Figure 31: (A) Number of raw MGS sequences. (B) Number of identified microbial sequences.

We identified a total of 15 unique bacterial phyla and 129 bacterial families in all 83 BAL samples. Assessment of bacterial phyla demonstrated predominance of the phylum Proteobacteria, and to a lesser extent the majority of BAL samples demonstrated presence of Actinobacteria, Bacteroidetes, and Firmicutes (Figure 33). Further taxonomic analysis at the family level demonstrated a predominance of the betaproteobacterial family Burkholderiaceae invariably represented in all samples. Many samples also demonstrated high relative abundance of the gammaproteobacterial family Alteromonadaceae and to a much lesser extent Pseudomonadaceae. The genus Burkholderia, belonging to the Betaproteobacteria class, was identified as the predominant microbial genus in the Proteobacteria phylum (Figure 33). Figure 34, shows the most abundant species in the Burkholderia genus. In addition, within the Firmicutes phyla, Streptococcus and Enterococcus were the most abundant genera (Figure 35).

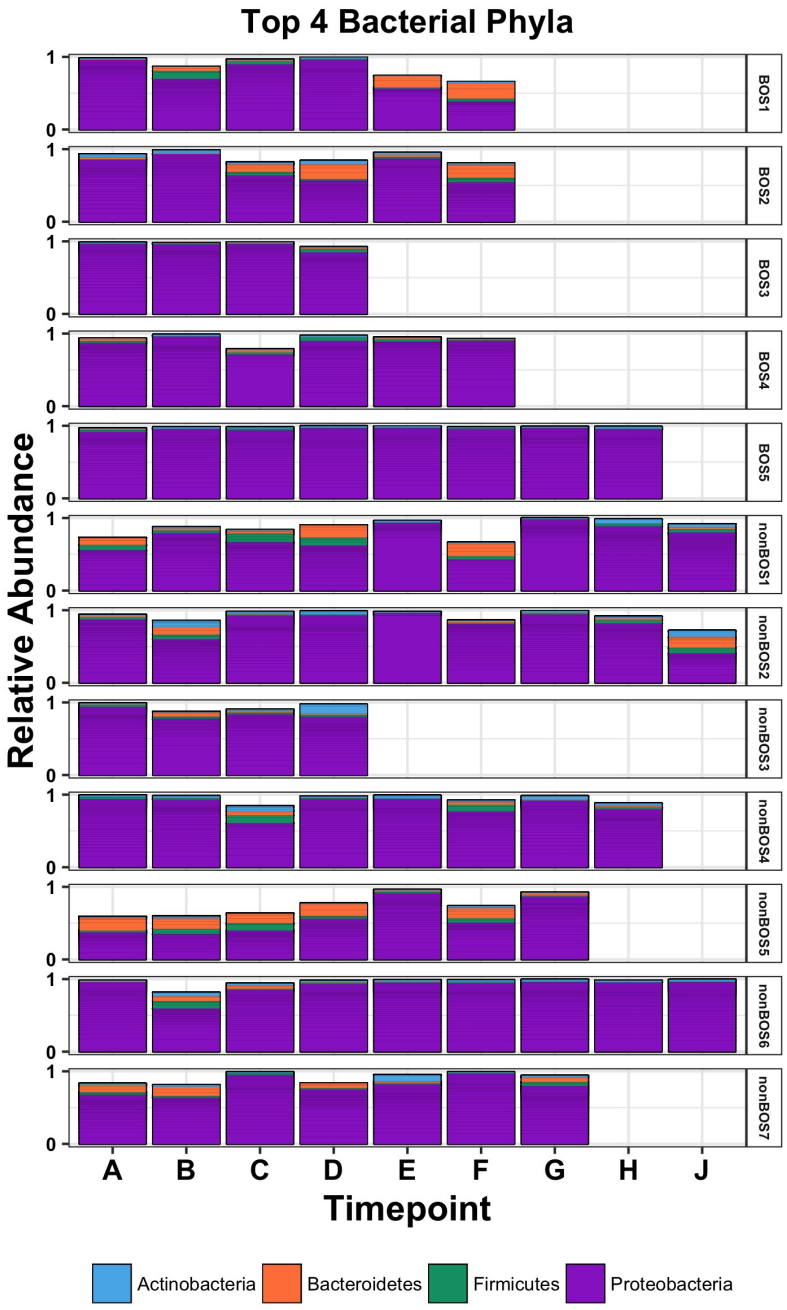



Figure 32: The top four most abundant bacterial phyla were Proteobacteria, Firmicutes, Bacteroides, and Actinobacteria. Stacked bar graphs demonstrate the relative taxonomic abundance per subject across individual BAL collection timepoints in the BOS and nonBOS groups.

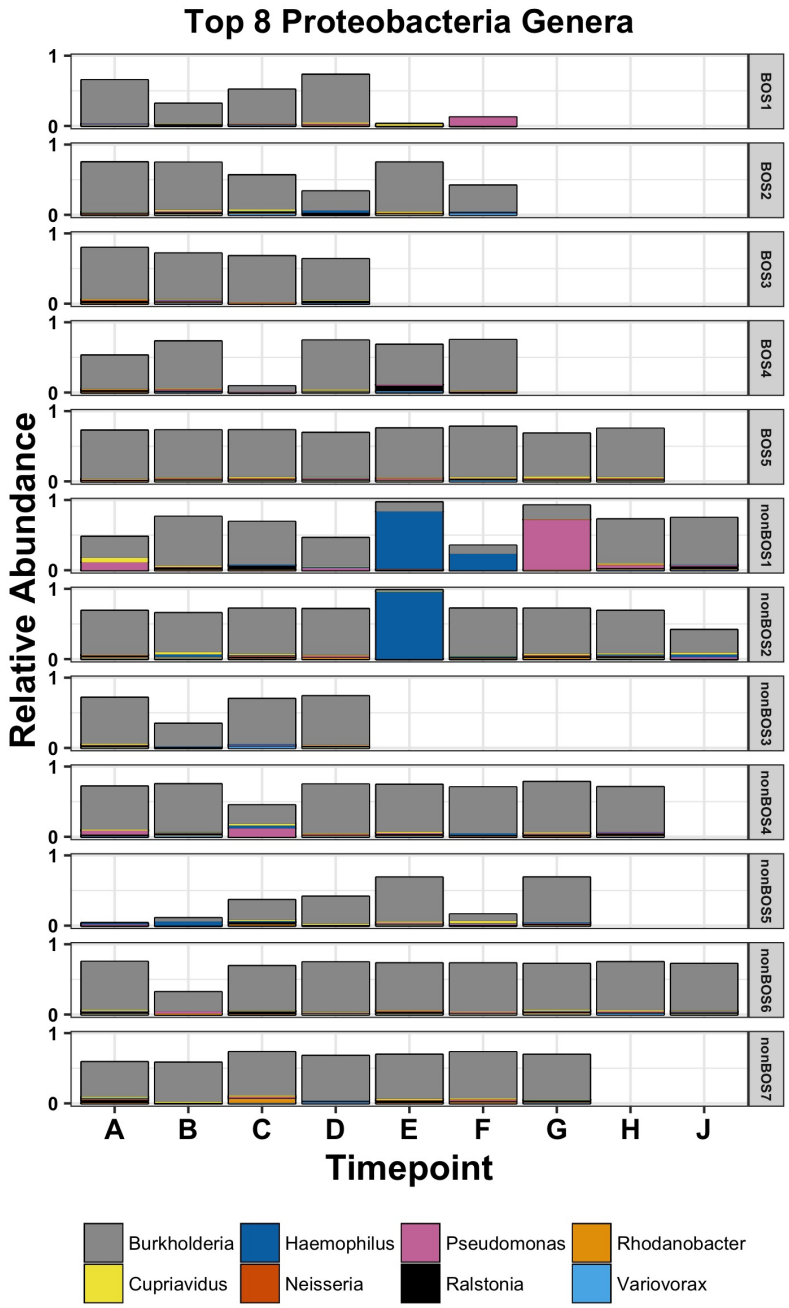



Figure 33: The top 8 most abundant genera in the Proteobacteria phylum. Stacked bar graphs demonstrate the relative taxonomic abundance per subject across individual BAL collection timepoints in the BOS and nonBOS groups. The genus Burkholderia, belonging to the Betaproteobacteria class, was identified as the predominant microbial genera among all identified phyla in the lower airways after bilateral lung transplantation in cystic fibrosis.

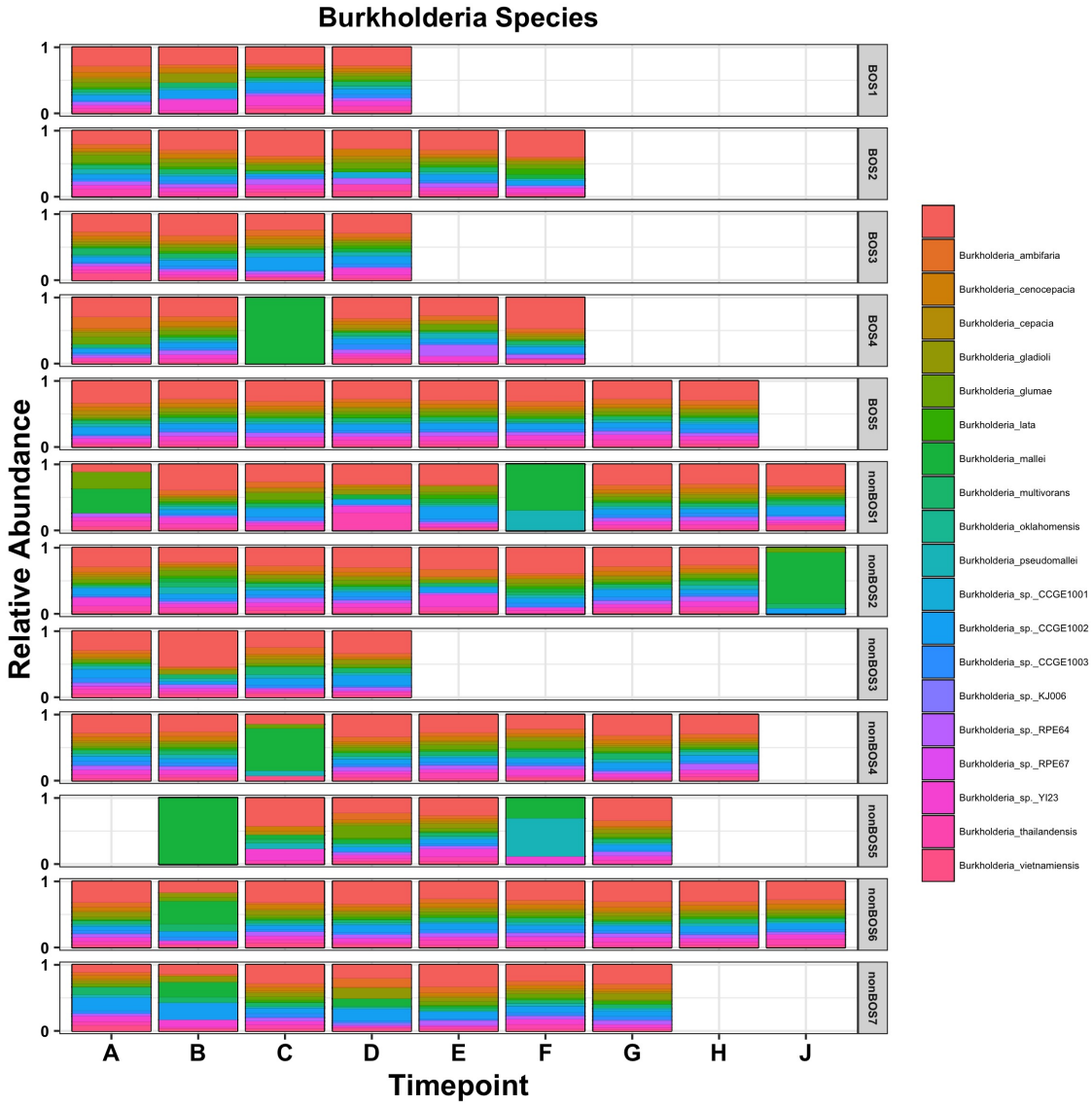



Figure 34: The most abundant species in the Burkholderia genus. Stacked bar graphs demonstrate the relative taxonomic abundance per subject across individual BAL collection timepoints in the BOS and nonBOS groups.

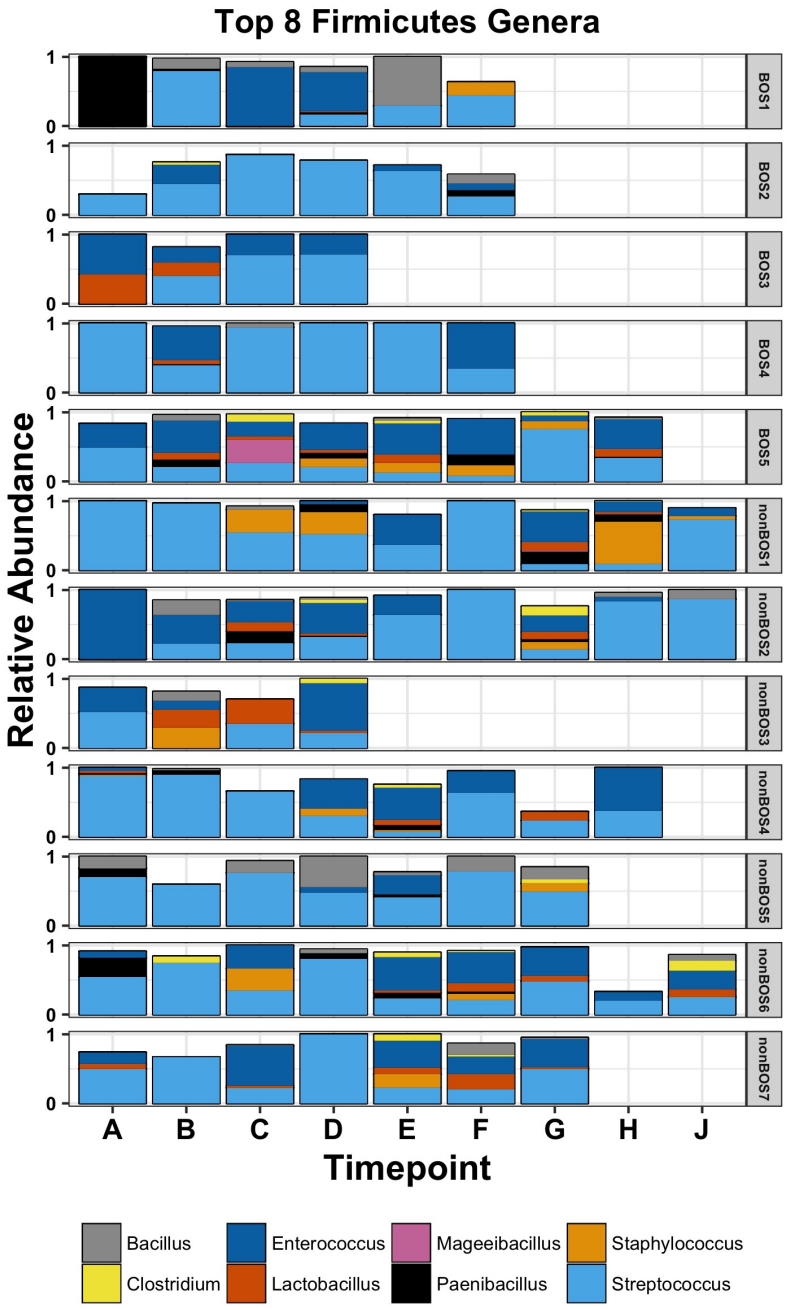



Figure 35: The top 8 most abundant genera in the Firmicutes phylum. Stacked bar graphs demonstrate the relative taxonomic abundance per subject across individual BAL collection timepoints in the BOS and nonBOS groups. Within the Firmicutes phyla, Streptococcus and Enterococcus were the most abundant genera.

Similarity among bacterial communities between all samples was assessed by non-metric multidimensional scaling (NMDS) based on Jaccard index. The unsupervised ordination analysis revealed no distinct clustering of the BOS and nonBOS groups (Figure 36), suggesting that no difference exists with regards to the whole bacterial taxonomic profiles among all samples from either group (BOS or nonBOS). Similarly, we assessed longitudinal variability of the relative abundance of bacterial communities at the phylum level and the family for the most abundant bacterial phyla per subject and identified a core lower airway microbiome comprised mainly of Burkholderia/Burkholderiaceae is preserved throughout the post-transplant state of recipients with cystic fibrosis children. There are a few notable exceptions in lower abundance taxa that correlate with infection identified by conventional culture data (samples at timepoints A, E, F, G, H from subject nonBOS1, and sample at timepoint E from subject nonBOS2).

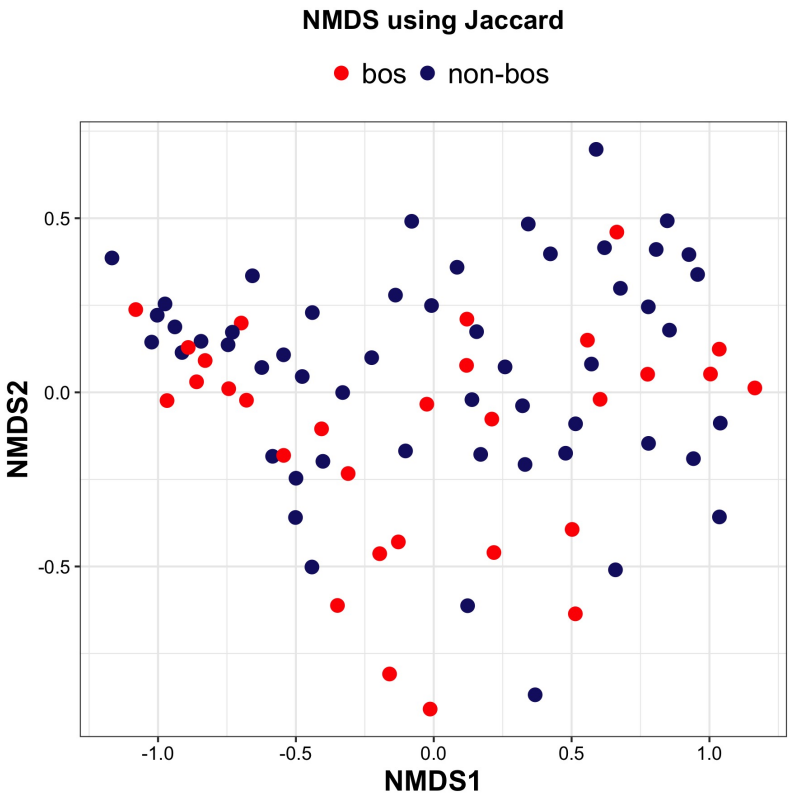



Figure 36: Non-metric multidimensional scaling (NMDS) using Jaccard distance between BOS and nonBOS taxonomic profiles.

Evaluation of Fisher diversity index demonstrated no difference between BOS and nonBOS subjects (p-value=0.11, Figure 37.A). Despite overall similar diversity, longitudinal comparison via *MetaLonDA* analysis of microbial diversity between groups identified the significant time-point (beyond post transplant day 297) during which significantly reduced microbial diversity is observed in the BOS group compared to the BOS-free group (p-value$<$0.05) suggesting that decreased microbial diversity contributes to BOS development (Figure 37.B). In addition, evaluation of individual diversity index trajectories per BAL sample collection timepoints per subject indicates that trends in Fisher's diversity index are related to development of BOS (Figure 38). Specifically, a pattern of relatively low post-transplant Fisher's diversity was observed within the first 50 days post-transplant. This was subsequently followed by an increase in diversity between post-transplant days 50-100, and finally a reduction in diversity at the time of final bronchoscopy and BAL sampling. To further understanding between bacterial diversity and development of BOS, spirometric and BAL parameters were analyzed in relation to Fisher's diversity index.

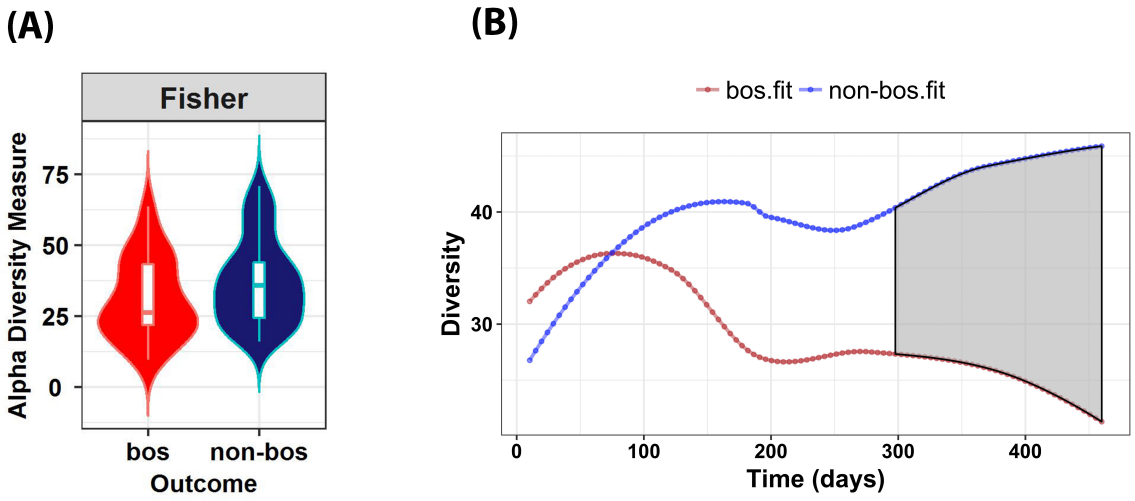



Figure 37: (A) Pooled microbial diversity was assessed by measuring Fisher's index among all BAL samples between BOS and nonBOS groups was found to be comparable (*Mann-Whitney U* test p-value=0.11). (B) Despite overall similar diversity, longitudinal comparison by *MetaLonDA* analysis of microbial diversity between groups identified the significant timepoint (beyond post transplant day 297) during which significantly reduced microbial diversity is observed in the BOS group compared to the BOS-free group (p-value$<$0.05) suggesting that decreased microbial diversity contributes to BOS development. The red spline represents pulmonary function of the BOS group over time (days) and blue spline is representative of the nonBOS group. The gray shaded area represents the significant time interval during which differences between groups were observed.

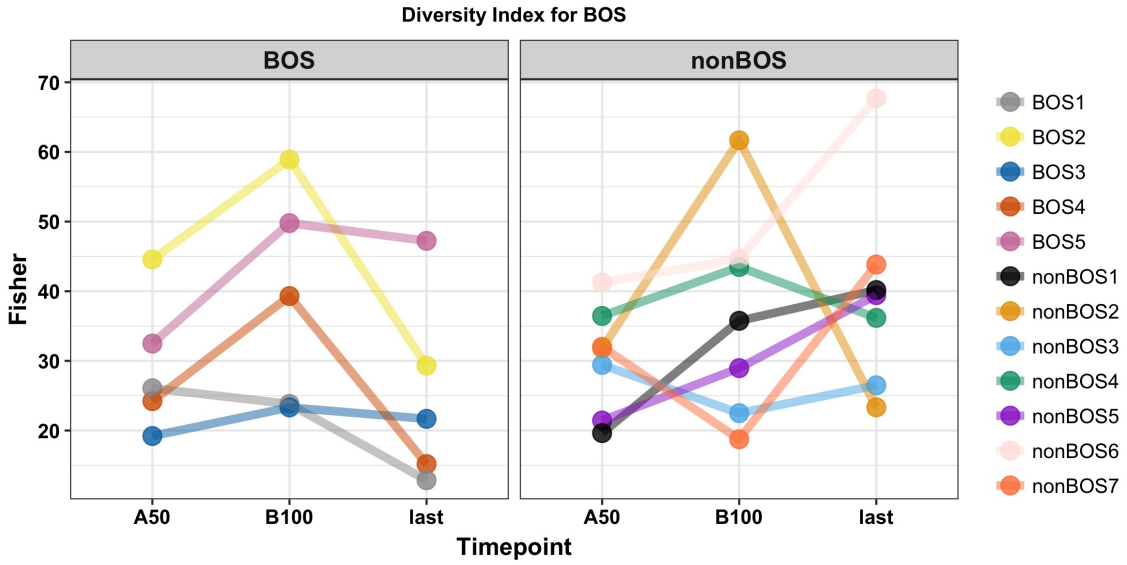



Figure 38: Diversity trajectory for initial and last time points. (A) for BOS group (B) nonBOS group.

Interestingly, among all 83 collected BAL samples, we identified significant negative correlations between increased BAL percent neutrophils (Figure 39; Pearson's $R^2$=0.2, p-value=3.54e-05) and absolute cell counts, but not percent lymphocytes, with decreased Fisher's diversity. This demonstrate that increases in microbial diversity relate to decreased percentage of neutrophils in BAL, suggesting that overgrowth or depletion of specific microbial taxa is directly related with a intensified local inflammatory response.

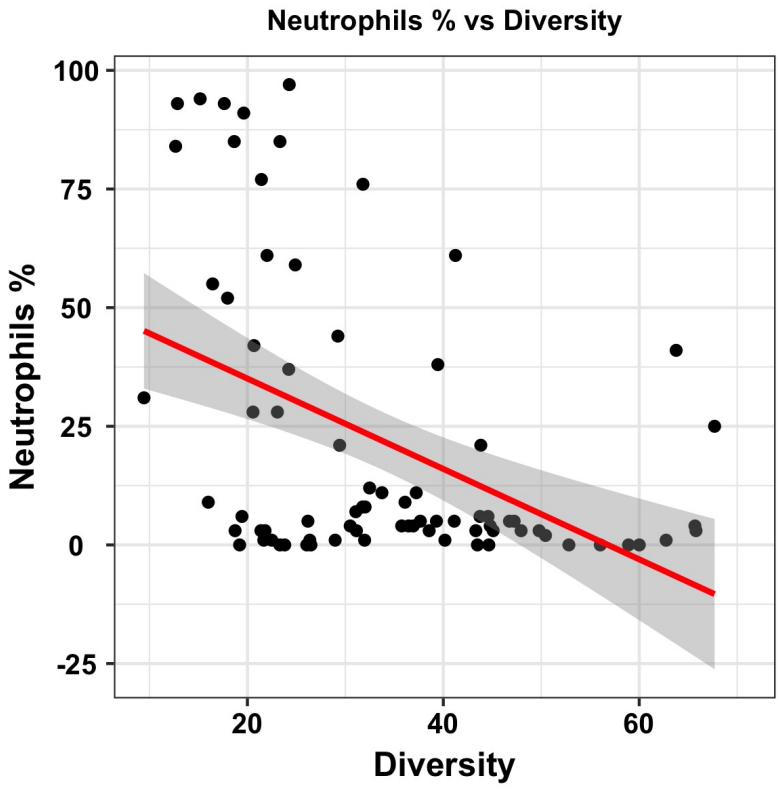



Figure 39: The relationship between microbial diversity (Fisher's index) and bronchoalveolar lavage inflammatory cell counts was explored among all samples. A significant inverse relationship was noted demonstrating that increases in microbial diversity relate to decreased percentage of neutrophils in BAL, suggesting that overgrowth or depletion of specific microbial taxa is directly related with a intensified local inflammatory response (Pearson's $R^2$=0.2, p-value=3.54e-05). Gray shaded area represents the 95% confidence interval.

### 6.4.3 Dynamics of Lower Airway Metagenomics

Dynamics and significance of individual taxonomic features within the lower airway bacterial communities were assessed at the phylum and family levels by applying *MetaLonDA*, a method that identifies the significant time intervals of microbial features in longitudinal studies, to the metagenomics data acquired from the BAL samples. We identified time intervals with differentially abundant phyla (Figure 40) and families in the BOS and nonBOS groups. A total of 2 bacterial phyla, Actinobacteria (p-value=0.016) and Proteobacteria (p-value=0.025) were identified as relatively more abundant in the nonBOS group compared to the BOS group. These two phyla were noted to establish significant communities later time intervals, with Actinobacteria (day 207.9 to 466.5) increasing from an earlier than Proteobacteria (day 352.7 to 466.5) (Figure 40). Actinobacteria and Proteobacteria demonstrate a downward trend in the BOS group in the immediate post-transplant period and ultimately cross-over and become relatively less abundant in the BOS group between days 100-150.

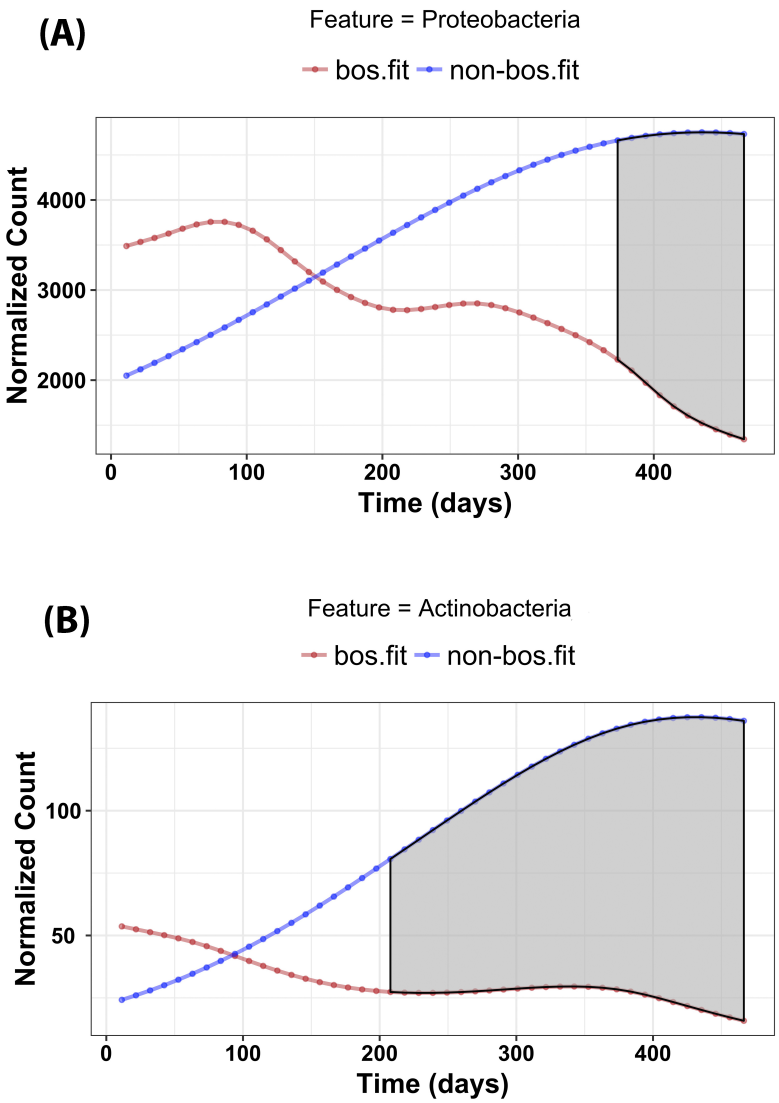



Figure 40: Longitudinal analysis of the identified lower airway microbiome from BAL samples between BOS and nonBOS groups was performed for all phyla with *MetaLonDA*. Differential timepoint analysis identified significant differences in abundance of the predominant (A) Proteobacteria phylum from day 373 to 456 (p-value$<$0.05) and in the less abundant (B) Actinobacteria phylum from day 207 to 456 (p-value$<$0.05). The red spline represents pulmonary function of the BOS group over time (days) and blue spline is representative of the nonBOS group. The gray shaded area represents the significant time interval during which differences between groups were observed.

At the family level, a total of 33 taxa were identified to have significant intervals of increased relative abundance (p-value$<$0.05). 96% (32/33) of these taxa demonstrated a higher relative abundance in nonBOS group and one taxa had a higher relative abundance in BOS (Figure 41). The majority of these bacterial families were noted to be environmental bacteria. However, Staphylococcaceae, Lactobacillaceae, and Neisseriaceae were among the families that have members which have previously been associated with human health and disease. Notably, early high relative abundance of Staphylococcaceae (prior to day 50 post-transplant) and late high relative abundance of Lactobacillaceae are associated with resilience against BOS. Interestingly, late (day 207.9 to 466.45) high relative abundance is associated with resilience against BOS. Similar to the phylum level, a cross-over point was found in 75% of these families related to human health and disease. Higher relative abundance of Staphylococcaceae was maintained in the nonBOS group throughout. These data suggest that a shift in bacterial community structure likely follows a critical event around day 100-150 that possibly determines the development of BOS or resilience against BOS. Specifically, early antibiotic regimens that deplete staphylococci and fail to control colonization with Neisseria may be a contributing factor. However, comparison of clinical events during this time interval did not demonstrate any significant differences between groups.

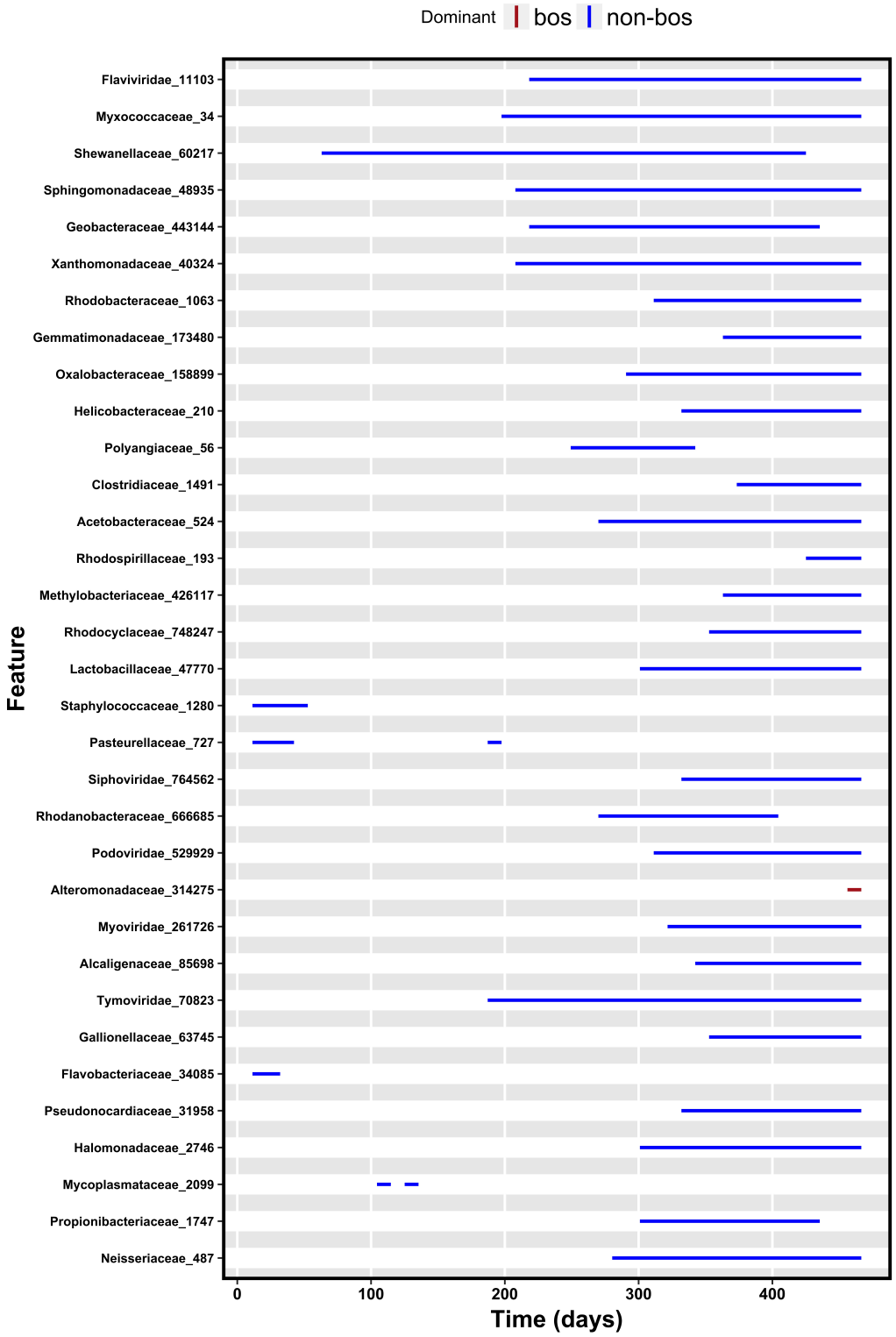



Figure 41: Longitudinal differential abundant families identified via *MetaLonDA*.

## 6.5 Conclusion

We analyzed the longitudinal microbiome profiles of pediatric post-transplant CF patients to illuminate the role of the microbiome in BOS development. We also examined the microbiome role in susceptibility or resilience to BOS. Our results suggest that as time progresses in the post-transplant period, those subjects who develop BOS are prone to a lower airway ecologic shift that leads to reduced bacterial diversity and may indicate an overgrowth of distinct bacteria that promote a persistent inflammatory response and development of BOS. In addition, our results suggest that a shift in bacterial community structure likely follows a critical event around day 100-150 that may determine the development of BOS or resilience against BOS. Specifically, early antibiotic regimens that deplete staphylococci and fail to control colonization with Neisseria may be a contributing factor. Interestingly, comparison of clinical events during this time interval did not demonstrate any significant differences between groups. Thus, it is possible that host-microbe interactions may lead to immune response modulation and exacerbate inflammatory and fibroproliferative changes involved in the development of BOS (191; 192).

Our limitations include the small sample size because the recruitment of cystic fibrosis patients undergoing lung transplantation is limited. Nevertheless, the number of samples and longitudinal analyses using *MetaLonDA*, made it possible to explore microbiome dynamics in small sample size.

A technical issue that arose was the sequencing depth, on average 2 million reads per sample, limiting investigation to only the most abundant taxa on high taxonomic levels such as phyla and

family. Interestingly, with the sequencing depth used in this study, we were able to identify all pathogens that were reported positive in conventional bacterial cultures.

# CHAPTER 7

# CONCLUSIONS

We discussed the computational methods and tools we have developed to improve both characterization and longitudinal analysis of the microbiome.

In this work, we have discussed our computational methods and tools, which were developed in an attempt to improve characterization and longitudinal analysis of the microbiome. The first method, *WEVOTE*, classifies microbial sequences into taxonomic units with both high precision and high sensitivity. The second method, *MetaLonDA*, identifies time intervals of differentially abundant microbial features in longitudinal studies. The third method is a computational framework to predict host clinical phenotype from longitudinal microbiome profiles via deep learning approach. Finally, using these methods and tools, we identified microbiome dynamics suggestive of the development of bronchiolitis obliterans syndrome in pediatric lung transplant recipients, insights that can be leveraged to improve lung transplant outcomes across life span. In the following sections, we summarize our contribution and give some insights to future perspectives.

## 7.1 Taxonomic Identification of Metagenomics Sequences

We developed *WEVOTE* (WEighted VOting Taxonomic idEntification), a phylogenetic-based method that classifies metagenome shotgun sequencing DNA sequence reads based on an ensemble of existing methods using k-mer-based, marker-based, and naive-similarity based approaches. Our performance evaluation, based on fourteen simulated microbiome datasets, con-

sistently demonstrates that *WEVOTE* achieves a high level of sensitivity and precision compared to individual methods across different taxonomic levels ranging from phyla to species. Moreover, the score assigned to the taxon for each read indicates the confidence level of the assignment. This information is especially useful for the assessment of false positive annotations at a particular taxonomic level. The classification score given by *WEVOTE* can be used for any downstream analysis that requires the confidence of the annotated sequences. In addition, we introduced a cloud-based solution to address common usability issues in the *WEVOTE* framework. Lastly, an interactive visual analytics tool was developed to ease interpretation of the classification results. We have demonstrated three different use cases of the pipeline that, in turn, reflects the significance of our modular design. *WEVOTE* and *WEVOTE-web* are publicly available on `https://github.com/aametwally/WEVOTE` and `https://github.com/aametwally/WEVOTE-web`, respectively.

### 7.1.1 Limitations and Future Perspectives

In our current implementation of *WEVOTE*, we have used a uniform weight of voting for each method. While *WEVOTE* outperforms other taxonomic classification methods, examination of the potential of incorporating different weighted votes for individual methods in each specific application merits investigation. A major obstacle in the current *WEVOTE* implementation, is the long computational time. This is mainly caused by the incorporation of BLASTN in the *WEVOTE* framework. In the future, we plan to replace *BLASTN* with *DIAMOND* (193), a method that is as sensitive as *BLAST* but an order of magnitude faster. In addition, we plan to extend WEVOTE algorithm to be able identify the microbial sequences at the strain level.

## 7.2 Identifying Time Intervals of Differentially Abundant Features in Metagenomic Longitudinal Studies

We have developed *MetaLonDA*, a method that can identify significant time-intervals of differentially abundant microbial features such as taxonomies, genes, or pathways. *MetaLonDA* is flexible such that it can perform differential abundance tests on longitudinal samples with different numbers of subjects per phenotypic group, different numbers of samples per subject, and samples that are not collected at consistent time points. These inconsistencies are often the case for samples collected from human subjects in translational studies. Inconsistencies increase with the complexity of the procedure utilized to obtain the samples. Usually, there is less inconsistency in samples collected through non-invasive procedures such as stool and urine samples, but increases in the case of invasive procedures such as bronchoalveolar lavage obtained by bronchoscopy. *MetaLonDA* relies on two modeling components: the NB distribution for modeling the mapped read counts for each feature and the semi-parametric SS-ANOVA technique for modeling longitudinal profiles associated with different phenotypes. Specific significant time intervals of microbial features can then be utilized to establish targeted timely screening or prevention of individual features and facilitate timely interventions, such as the use of antibiotics or probiotics. Unlike with cross-sectional methods that are incapable of identifying significant time intervals associated with differentially abundant features, *MetaLonDA* may lead to reconstitution of the microbiome and reestablishment of homeostasis prior to onset of overt disease. *MetaLonDA* is publicly available on the CRAN repository (`https://CRAN.`

R-project.org/package=MetaLonDA) and an active development is being done on (https://github.com/aametwally/MetaLonDA).

### 7.2.1 Limitations and Future Perspectives

One of *MetaLonDA*'s limitations is that when samples are sparse over extended time intervals, the fitted smoothing spline has large variation (81). This causes the identified significant time intervals to be unreliable and should be excluded from the analysis. Thus, in the future, we plan to develop a statistical method that identifies these extended intervals. Additionally, the current version of *MetaLonDA* only finds the association between microbial features, time, and phenotypic group. Incorporation of additional confounding factors (age, gender, race, disease severity, etc.) to the *MetaLonDA* model will enhance applicability. *MetaLonDA* was developed primarily for identifying time intervals of features in metagenomics studies. Thus, the NB assumption made for taxonomy needs to be reassessed before *MetaLonDA* can be confidently applied to other analyses, e.g., metatranscriptomic, RNAseq, or proteomics data. In the future, we plan to implement a checker function that evaluates the distributional assumption based on the KS test, and accordingly, the best-fitted model can be utilized for the longitudinal differential abundance test.

## 7.3 Predict Host Phenotype from Longitudinal Microbiome Profiles via Deep Learning

Long Short-Term Memory networks have the ability to learn dynamic temporal behavior for a time sequence event. We have developed a deep learning framework that has the capacity of predicting clinical outcomes from longitudinal microbiome profiles, we discussed as an example,

food allergy. Food allergy is difficult to diagnose at young ages, may lead to inability to treat at earlier ages causing severe complications. The framework is based on sparse autoencoder and Long Short-Term Memory networks. Sparse autoencoder is devised to extract potential latent structures in the microbiome prior to LSTM training. We tested the framework on the DIABIMMUNE dataset (`https://pubs.broadinstitute.org/diabimmune`), a study that characterized host-microbe immune interactions contributing to autoimmunity and allergy. Our results demonstrate the proper use of the proposed model and show the increase in predictive power compared to SVM, Random Forest, and LASSO regression.

### 7.3.1 Limitations and Future Perspectives

Although our deep learning framework shows potential to predict phenotype from a sequence of microbiome profiles and outperforms other classical methods, it does not reach a prediction level for optimal clinical utilization. This is mainly caused by the nature of the training dataset that we used to train our model. The DIABIMMUNE dataset is small (195 subjects) and each subject has few time points (6 on average). With the current reduction in sequencing costs, we anticipate that multiple large longitudinal microbiome projects will be available which in turn could be used to train models like ours for better prediction power. Another hypothesis that tested with the DIABIMMUNE dataset but were not successful is predicting disease prognosis from longitudinal microbiome profiles. To train such a model, we needed to have a dataset with subjects that are changing phenotype from one to the other back and forth so that the model learns the pattern associated with the change. Unfortunately, in the DIABIMMUE dataset, subjects maintain the same phenotype across the study period.

## 7.4 Lower Airway Microbiome Dynamics as a Predictor of Bronchiolitis Obliterans Syndrome after Pediatric Lung Transplantation in Cystic Fibrosis

We analyzed the longitudinal microbiome profiles of pediatric post-transplant CF patients to illuminate the role of the microbiome in BOS development. We also examined the microbiome role in susceptibility or resilience to BOS. Our results suggest that as time progresses in the post-transplant period, those subjects who develop BOS are prone to a lower airway ecologic shift that leads to reduced bacterial diversity and may indicate an overgrowth of distinct bacteria that promote a persistent inflammatory response and development of BOS. In addition, our results suggest that a shift in bacterial community structure likely follows a critical event around day 100-150 that may determine the development of BOS or resilience against BOS. Specifically, early antibiotic regimens that deplete staphylococci and fail to control colonization with Neisseria may be a contributing factor. Interestingly, comparison of clinical events during this time interval did not demonstrate any significant differences between groups. Thus, it is possible that host-microbe interactions may lead to immune response modulation and exacerbate inflammatory and fibroproliferative changes involved in the development of BOS (191; 192).

### 7.4.1 Limitations and Future Perspectives

One limitation we have in our study is the small sample size. It is difficult to recruit pediatric patients who have cystic fibrosis, and undergo lung transplant and follow up for three years. Another limitation is sequencing depth, on average 2 million reads per sample. This depth limits

the investigation to only the most abundant taxa on high taxonomic levels such as phyla and family.

# CITED LITERATURE


1. Tommi Vatanen, Aleksandar D. Kostic, Eva d'Hennezel, Heli Siljander, Eric A. Franzosa, Moran Yassour, Raivo Kolde, Hera Vlamakis, Timothy D. Arthur, Anu-Maaria Hämäläinen, Aleksandr Peet, Vallo Tillmann, Raivo Uibo, Sergei Mokurov, Natalya Dorshakova, Jorma Ilonen, Suvi M. Virtanen, Susanne J. Szabo, Jeffrey A. Porter, Harri Lähdesmäki, Curtis Huttenhower, Dirk Gevers, Thomas W. Cullen, Mikael Knip, and Ramnik J. Xavier. Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell*, 165(4):842–853, 5 2016.

2. Kathryn J. Pflughoeft and James Versalovic. Human Microbiome in Health and Disease. *Annual Review of Pathology: Mechanisms of Disease*, 7(1):99–122, 2 2012.

3. Ilseung Cho and Martin J Blaser. The human microbiome: at the interface of health and disease. *Nature reviews. Genetics*, 13(4):260–70, 3 2012.

4. A. Rani, R. Ranjan, H.S. McGee, A. Metwally, Z. Hajjiri, D.C. Brennan, P.W. Finn, and D.L. Perkins. A diverse virome in kidney transplant patients contains multiple viral subtypes with distinct polymorphisms. *Scientific Reports*, 6, 2016.

5. Benjamin A. Turturice, Halvor S. McGee, Brian Oliver, Melissa Baraket, Brian T. Nguyen, Christian Ascoli, Ravi Ranjan, Asha Rani, David L. Perkins, and Patricia W. Finn. Atopic asthmatic immune phenotypes associated with airway microbiota and airway obstruction. *PLOS ONE*, 12(10):e0184566, 10 2017.

6. Ravi Ranjan, Asha Rani, Patricia W Finn, and David L Perkins. Evaluating bacterial and functional diversity of human gut microbiota by complementary metagenomics and metatranscriptomics. *bioRxiv*, 2018.

7. Ciarán P. Kelly. Fecal Microbiota Transplantation — An Old Therapy Comes of Age. *New England Journal of Medicine*, 368(5):474–475, 1 2013.

8. Anne Vrieze, Els Van Nood, Frits Holleman, Jarkko Salojärvi, Ruud S. Kootte, Joep F.W.M. Bartelsman, Geesje M. Dallinga–Thie, Mariette T. Ackermans, Mireille J. Serlie, Raish Oozeer, Muriel Derrien, Anne Druesne, Johan E.T. Van Hylckama Vlieg, Vincent W. Bloks, Albert K. Groen, Hans G.H.J. Heilig, Erwin G. Zoetendal, Erik S. Stroes,

Willem M. de Vos, Joost B.L. Hoekstra, and Max Nieuwdorp. Transfer of Intestinal Microbiota From Lean Donors Increases Insulin Sensitivity in Individuals With Metabolic Syndrome. *Gastroenterology*, 143(4):913–916, 10 2012.

9. Tanya Yatsunenko, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N. Baldassano, Andrey P. Anokhin, Andrew C. Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J. Gregory Caporaso, Catherine A. Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I. Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222, 5 2012.

10. Steven R Gill, Mihai Pop, Robert T Deboy, Paul B Eckburg, Peter J Turnbaugh, Buck S Samuel, Jeffrey I Gordon, David A Relman, Claire M Fraser-Liggett, and Karen E Nelson. Metagenomic analysis of the human distal gut microbiome. *Science (New York, N.Y.)*, 312(5778):1355–9, 6 2006.

11. Michael D. J. Lynch and Josh D. Neufeld. Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology*, 13(4):217–229, 4 2015.

12. Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. The Human Microbiome Project. *Nature*, 449(7164):804–810, 10 2007.

13. Justin Kuczynski, Christian L. Lauber, William A. Walters, Laura Wegener Parfrey, José C. Clemente, Dirk Gevers, and Rob Knight. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47–58, 1 2012.

14. Michael S. Rappé and Stephen J. Giovannoni. The Uncultured Microbial Majority. *Annual Review of Microbiology*, 57(1):369–394, 10 2003.

15. Eric J Stewart. Growing unculturable bacteria. *Journal of bacteriology*, 194(16):4151–60, 8 2012.

16. D Nichols, N Cahoon, E M Trakhtenberg, L Pham, A Mehta, A Belanger, T Kanigan, K Lewis, and S S Epstein. Use of Ichip for High-Throughput In Situ Cultivation of "Uncultivable" Microbial Species. *Applied and environmental microbiology*, 76(8):2445–2450, 2010.

17. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

18. Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R. Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H. Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, Peer Bork, S. Dusko Ehrlich, Jun Wang, Hervé Blottiere, Natalia Borruel, Thomas Bruls, Francesc Casellas, Christian Chervaux, Antonella Cultrone, Christine Delorme, Gérard Denariaz, Rozenn Dervyn, Miguel Forte, Carsten Friss, Maarten van de Guchte, Eric Guedon, Florence Haimet, Alexandre Jamet, Catherine Juste, Ghalia Kaci, Michiel Kleerebezem, Jan Knol, Michel Kristensen, Severine Layec, Karine Le Roux, Marion Leclerc, Emmanuelle Maguin, Raquel Melo Minardi, Raish Oozeer, Maria Rescigno, Nicolas Sanchez, Sebastian Tims, Toni Torrejon, Encarna Varela, Willem de Vos, Yohanan Winogradsky, Erwin Zoetendal, Peer Bork, S. Dusko Ehrlich, and Jun Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 3 2010.

19. Sylvie Sanschagrin and Etienne Yergeau. Next-generation sequencing of 16S ribosomal RNA gene amplicons. *Journal of visualized experiments : JoVE*, (90), 8 2014.

20. Rachel Poretsky, Luis M. Rodriguez-R, Chengwei Luo, Despina Tsementzi, and Konstantinos T. Konstantinidis. Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLoS ONE*, 9(4):e93827, 4 2014.

21. Konstantinos T Konstantinidis and James M Tiedje. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current Opinion in Microbiology*, 10(5):504–509, 10 2007.

22. Konstantinos T. Konstantinidis and Erko Stackebrandt. Defining Taxonomic Ranks. In *The Prokaryotes*, pages 229–254. Springer Berlin Heidelberg, Berlin, Heidelberg, 12 2013.

23. David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, and Chris P. Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2 2014.

24. Chengwei Luo, Luis M. Rodriguez-R, and Konstantinos T. Konstantinidis. A User's Guide to Quantitative and Comparative Analysis of Metagenomic Datasets. volume 531, pages 525–547. 2013.

25. Chengwei Luo, Luis M. Rodriguez-R, and Konstantinos T. Konstantinidis. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Research*, 42(8):e73–e73, 4 2014.

26. Jason Lloyd-Price, Anup Mahurkar, Gholamali Rahnavard, Jonathan Crabtree, Joshua Orvis, A. Brantley Hall, Arthur Brady, Heather H. Creasy, Carrie McCracken, Michelle G. Giglio, Daniel McDonald, Eric A. Franzosa, Rob Knight, Owen White, and Curtis Huttenhower. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*, 550(7674):61, 9 2017.

27. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe*, 16(3):276–89, 9 2014.

28. Marta Wlodarska, Aleksandar D Kostic, and Ramnik J Xavier. An integrative view of microbiome-host interactions in inflammatory bowel diseases. *Cell host & microbe*, 17(5):577–91, 5 2015.

29. Curtis Huttenhower, Aleksandar D. Kostic, and Ramnik J. Xavier. Inflammatory Bowel Disease as a Model for Translating the Microbiome. *Immunity*, 40(6):843–854, 6 2014.

30. Aleksandar D Kostic, Dirk Gevers, Heli Siljander, Tommi Vatanen, Tuulia Hyötyläinen, Anu-Maaria Hämäläinen, Aleksandr Peet, Vallo Tillmann, Päivi Pöhö, Ismo Mattila, Harri Lähdesmäki, Eric A Franzosa, Outi Vaarala, Marcus de Goffau, Hermie Harmsen, Jorma Ilonen, Suvi M Virtanen, Clary B Clish, Matej Orešič, Curtis Huttenhower, Mikael Knip, Ramnik J. DIABIMMUNE Study Group, and Ramnik J Xavier. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell host & microbe*, 17(2):260–73, 2 2015.

31. Moran Yassour, Tommi Vatanen, Heli Siljander, Anu-Maaria Hämäläinen, Taina Härkönen, Samppa J Ryhänen, Eric A Franzosa, Hera Vlamakis, Curtis Huttenhower, Dirk Gev-

ers, Eric S Lander, Mikael Knip, on behalf of the DIABIMMUNE Study DIABIMMUNE Study Group, and Ramnik J Xavier. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science translational medicine*, 8(343):343ra81, 6 2016.

32. The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–221, 6 2012.

33. Bernard Khor, Agnès Gardet, and Ramnik J Xavier. Genetics and pathogenesis of inflammatory bowel disease. *Nature*, 474(7351):307–17, 6 2011.

34. Natalie A Molodecky and Gilaad G Kaplan. Environmental risk factors for inflammatory bowel disease. *Gastroenterology & hepatology*, 6(5):339–46, 5 2010.

35. Natalie A. Molodecky, Ing Shian Soon, Doreen M. Rabi, William A. Ghali, Mollie Ferris, Greg Chernoff, Eric I. Benchimol, Remo Panaccione, Subrata Ghosh, Herman W. Barkema, and Gilaad G. Kaplan. Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review. *Gastroenterology*, 142(1):46–54, 1 2012.

36. Katsuyoshi Matsuoka and Takanori Kanai. The gut microbiota and inflammatory bowel disease. *Seminars in immunopathology*, 37(1):47–55, 1 2015.

37. Georgina L Hold, Megan Smith, Charlie Grange, Euan Robert Watt, Emad M El-Omar, and Indrani Mukhopadhya. Role of the gut microbiota in inflammatory bowel disease pathogenesis: what have we learnt in the past 10 years? *World journal of gastroenterology*, 20(5):1192–210, 2 2014.

38. Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Torben Hansen, Gaston Sanchez, Jeroen Raes, Gwen Falony, Shujiro Okuda, and Mathieu Almeida. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

39. Nadja Larsen, Finn K. Vogensen, Frans W. J. van den Berg, Dennis Sandris Nielsen, Anne Sofie Andreasen, Bente K. Pedersen, Waleed Abu Al-Soud, Søren J. Sørensen, Lars H. Hansen, and Mogens Jakobsen. Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults. *PLoS ONE*, 5(2):e9085, 2 2010.

40. Peter J. Turnbaugh, Fredrik Bäckhed, Lucinda Fulton, and Jeffrey I. Gordon. Diet-Induced Obesity Is Linked to Marked but Reversible Alterations in the Mouse Distal Gut Microbiome. *Cell Host & Microbe*, 3(4):213–223, 4 2008.

41. Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

42. Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, 1 2009.

43. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 7 2009.

44. W James Kent. BLAT–the BLAST-like alignment tool. *Genome research*, 12(4):656–64, 4 2002.

45. Ahmed A. Metwally, Yang Dai, Patricia W. Finn, and David L. Perkins. WEVOTE: Weighted Voting Taxonomic Identification Method of Microbial Sequences. *PLOS ONE*, 11(9):e0163527, 9 2016.

46. Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46, 1 2014.

47. Hanno Teeling and Frank O. Glockner. Current opportunities and challenges in microbial metagenome analysis–a bioinformatic perspective. *Briefings in Bioinformatics*, 13(6):728–742, 2012.

48. Nicola Segata, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–4, 8 2012.

49. Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods*, 12(10):902–3, 10 2015.

50. Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. MEGAN analysis of metagenomic data. *Genome research*, 17(3):377–86, 3 2007.

51. Daniel H Huson, Suparna Mitra, Hans-Joachim Ruscheweyh, Nico Weber, and Stephan C Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome research*, 21(9):1552–60, 9 2011.

52. Daniel H. Huson, Sina Beier, Isabell Flade, Anna Górska, Mohamed El-Hadidi, Suparna Mitra, Hans-Joachim Ruscheweyh, and Rewati Tappu. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLOS Computational Biology*, 12(6):e1004957, 6 2016.

53. Bo Liu, Theodore Gibbons, Mohammad Ghodsi, Todd Treangen, and Mihai Pop. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC genomics*, 12 Suppl 2(2):S4, 1 2011.

54. Nam-Phuong Nguyen, Siavash Mirarab, Bo Liu, Mihai Pop, and Tandy Warnow. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics (Oxford, England)*, 30(24):3548–55, 12 2014.

55. Sasha K Ames, David A Hysom, Shea N Gardner, G Scott Lloyd, Maya B Gokhale, and Jonathan E Allen. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics (Oxford, England)*, 29(18):2253–60, 9 2013.

56. Daehwan Kim, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, 10 2016.

57. Rachid Ounit, Steve Wanamaker, Timothy J Close, and Stefano Lonardi. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16(1):236, 3 2015.

58. P. Menzel, K. Lee Ng, and A. Krogh. Kaiju: Fast and sensitive taxonomic classification formetagenomics. Technical report, 11 2015.

59. Karel Břinda. *Novel computational techniques for mapping and classifying Next-Generation Sequencing data*. PhD thesis, 11 2016.

60. Karel Břinda, Maciej Sykulski, and Gregory Kucherov. Spaced seeds improve k-mer-based metagenomic classification. *Bioinformatics*, 31(22):3584–3592, 11 2015.

61. Jolanta Kawulok and Sebastian Deorowicz. CoMeta: Classification of Metagenomes Using k-mers. *PLOS ONE*, 10(4):e0121453, 4 2015.

62. Samuel S Minot, Niklas Krumm, and Nicholas B Greenfield. One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification. *bioRxiv*, page 027607, 9 2015.

63. Aaron Y. Lee, Cecilia S. Lee, and Russell N. Van Gelder. Scalable metagenomics alignment research tool (SMART): a scalable, rapid, and complete search heuristic for the classification of metagenomic sequences from complex sequence populations. *BMC Bioinformatics*, 17(1):292, 12 2016.

64. Alexa B. R. McIntyre, Rachid Ounit, Ebrahim Afshinnekoo, Robert J. Prill, Elizabeth Hénaff, Noah Alexander, Samuel S. Minot, David Danko, Jonathan Foox, Sofia Ahsanuddin, Scott Tighe, Nur A. Hasan, Poorani Subramanian, Kelly Moffat, Shawn Levy, Stefano Lonardi, Nick Greenfield, Rita R. Colwell, Gail L. Rosen, and Christopher E. Mason. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology*, 18(1):182, 12 2017.

65. Stinus Lindgreen, Karen L. Adair, and Paul P. Gardner. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, 6:19233, 1 2016.

66. Richard J. Randle-Boggis, Thorunn Helgason, Melanie Sapp, and Peter D. Ashton. Evaluating techniques for metagenome annotation using simulated sequence data. *FEMS Microbiology Ecology*, 92(7):fiw095, 7 2016.

67. Darrell O. Ricke, Anna Shcherbina, and Nelson Chiu. Evaluating performance of metagenomic characterization algorithms using in silico datasets generated with FASTQSim. *bioRxiv*, page 046532, 3 2016.

68. Michael A. Peabody, Thea Van Rossum, Raymond Lo, and Fiona S. L. Brinkman. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics*, 16(1):362, 12 2015.

69. Chong Gu. Smoothing Spline ANOVA Models. 2012.

70. J Gregory Caporaso, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, Jeffrey I Gordon, and Rob Knight. Moving pictures of the human microbiome. *Genome biology*, 12(5):R50, 2011.

71. Peter J. Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, William J. Jones, Bruce A. Roe, Jason P. Affourtit, Michael Egholm, Bernard Henrissat, Andrew C. Heath, Rob Knight, and Jeffrey I. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, 1 2009.

72. Jeremy E Koenig, Aymé Spor, Nicholas Scalfone, Ashwana D Fricker, Jesse Stombaugh, Rob Knight, Largus T Angenent, and Ruth E Ley. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, 108 Suppl 1(Supplement 1):4578–85, 3 2011.

73. Alison Morris, Joseph N. Paulson, Hisham Talukder, Laura Tipton, Heather Kling, Lijia Cui, Adam Fitch, Mihai Pop, Karen A. Norris, and Elodie Ghedin. Longitudinal analysis of the lung microbiota of cynomolgous macaques during long-term SHIV infection. *Microbiome*, 4(1):38, 12 2016.

74. Erika K. Ganda, Rafael S. Bisinotto, Svetlana F. Lima, Kristina Kronauer, Dean H. Decter, Georgios Oikonomou, Ynte H. Schukken, and Rodrigo C. Bicalho. Longitudinal metagenomic profiling of bovine milk to assess the impact of intramammary treatment using a third-generation cephalosporin. *Scientific Reports*, 6(1):37565, 12 2016.

75. Francesco Asnicar, Serena Manara, Moreno Zolfo, Duy Tin Truong, Matthias Scholz, Federica Armanini, Pamela Ferretti, Valentina Gorfer, Anna Pedrotti, Adrian Tett, and Nicola Segata. Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems*, 2(1), 2017.

76. Anita Y Voigt, Paul I Costea, Jens Roat Kultima, Simone S Li, Georg Zeller, Shinichi Sunagawa, and Peer Bork. Temporal and technical variability of human gut metagenomes. *Genome Biology*, 16(1):73, 12 2015.

77. Joseph Nathaniel Paulson, Hisham Talukder, and Hector Corrada Bravo. Longitudinal differential abundance analysis of microbial marker-gene surveys using smoothing splines. *bioRxiv*, page 099457, 2017.

78. Dan Luo, Sara Ziebell, and Lingling An. An Informative Approach on Differential Abundance Analysis for Time-course Metagenomic Sequencing Data. *Bioinformatics*, 33(9):1286–1292, 1 2017.

79. Chong Gu. Smoothing Spline ANOVA Models: R Package gss. *Journal of Statistical Software*, 58(5):1–25, 2014.

80. Grace Wahba, Yuedong Wang, Chong Gu, Ronald Klein, and Barbara Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy : the 1994 Neyman Memorial Lecture. *The Annals of Statistics*, 23(6):1865–1895, 12 1995.

81. Chong Gu. *Smoothing spline ANOVA models*. Springer Science & Business Media, New York, 2nd edition, 2013.

82. Micah Hamady, Catherine Lozupone, and Rob Knight. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *The ISME journal*, 4(1):17–27, 1 2010.

83. J. Roger Bray and J. T. Curtis. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, 27(4):325–349, 2 1957.

84. Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H. S. Segler, Anthony Gitter, and Casey S. Greene. Opportunities And Obstacles For Deep Learning In Biology And Medicine. *doi.org*, page 142760, 5 2017.

85. Gregory Ditzler, Robi Polikar, and Gail Rosen. Multi-Layer and Recursive Neural Networks for Metagenomic Classification. *IEEE transactions on nanobioscience*, 14(6):608–16, 9 2015.

86. Qunyuan Zhang, Haley Abel, Alan Wells, Petra Lenzini, Felicia Gomez, Michael A. Province, Alan A. Templeton, George M. Weinstock, Nita H. Salzman, and Ingrid B. Borecki. Selection of models for the analysis of risk-factor trees: leveraging biological knowledge to mine large sets of risk factors with application to microbiome data. *Bioinformatics*, 31(10):1607–1613, 5 2015.

87. Benjamin Wingfield, Sonya Coleman, T.M. McGinnity, and Anthony J Bjourson. A metagenomic hybrid classifier for paediatric inflammatory bowel disease. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 1083–1089. IEEE, 7 2016.

88. Edoardo Pasolli, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLOS Computational Biology*, 12(7):e1004977, 7 2016.

89. Dan Knights, Laura Wegener Parfrey, Jesse Zaneveld, Catherine Lozupone, and Rob Knight. Human-Associated Microbial Signatures: Examining Their Predictive Value. *Cell Host & Microbe*, 10(4):292–296, 10 2011.

90. Derek Reiman, Ahmed Metwally, and Yang Dai. Using Convolutional Neural Networks to Explore the Microbiome. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4269–4272, Jeju Island, South Korea, 2017. IEEE.

91. Derek Reiman, Ahmed A Metwally, and Yang Dai. PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolution Neural Networks for Metagenomic Data. *bioRxiv*, page 257931, 1 2018.

92. Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

93. William Jones, Kaur Alasoo, Dmytro Fishman, and Leopold Parts. Computational biology: deep learning. *Emerging Topics in Life Sciences*, 1(3):257–274, 11 2017.

94. Benjamin Schuster-Böckler and Alex Bateman. An Introduction to Hidden Markov Models. In *Current Protocols in Bioinformatics*, volume 18, pages A.3A.1–A.3A.9. John Wiley & Sons, Inc., Hoboken, NJ, USA, 6 2007.

95. Hirotugu Akaike. Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247, 1969.

96. Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller. A Greedy Algorithm for Aligning DNA Sequences. *Journal of Computational Biology*, 7(1-2):203–214, 2 2000.

97. Arthur Brady and Steven L Salzberg. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature methods*, 6(9):673–6, 9 2009.

98. Liisa B. Koski and G. Brian Golding. The Closest BLAST Hit Is Often Not the Nearest Neighbor. *Journal of Molecular Evolution*, 52(6):540–542, 6 2001.

99. Shinichi Sunagawa, Daniel R Mende, Georg Zeller, Fernando Izquierdo-Carrasco, Simon A Berger, Jens Roat Kultima, Luis Pedro Coelho, Manimozhiyan Arumugam, Julien Tap, Henrik Bjørn Nielsen, Simon Rasmussen, Søren Brunak, Oluf Pedersen, Francisco Guarner, Willem M de Vos, Jun Wang, Junhua Li, Joël Doré, S Dusko Ehrlich,

Alexandros Stamatakis, and Peer Bork. Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods*, 10(12):1196–9, 12 2013.

100. Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902–903, 9 2015.

101. Gail L Rosen, Erin R Reichenberger, and Aaron M Rosenfeld. NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics (Oxford, England)*, 27(1):127–9, 1 2011.

102. Asem Alaa and Ahmed A. Metwally. Cloud-based solution for improving usability and interactivity of metagenomic ensemble taxonomic classification methods. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 198–201. IEEE, 3 2018.

103. Martin Wu and Jonathan A Eisen. A simple, fast, and accurate method of phylogenomic inference. *Genome Biology*, 9(10):R151, 1 2008.

104. Kevin Liu, Tandy J Warnow, Mark T Holder, Serita M Nelesen, Jiaye Yu, Alexandros P Stamatakis, and C Randal Linder. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic biology*, 61(1):90–106, 1 2012.

105. S R Eddy. Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14(9):755–63, 1 1998.

106. Frederick A Matsen, Robin B Kodner, and E Virginia Armbrust. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, 11:538, 1 2010.

107. Kristina Chodorow. *MongoDB : the definitive guide*. O'Reilly Media, second edition, 2013.

108. Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(Database issue):355–60, 1 2010.

109. Elena Deza and Michel Marie Deza. *Encyclopedia of Distances*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

110. Vitor C. Piro, Marcel Matschkowski, and Bernhard Y. Renard. MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*, 5(1):101, 12 2017.

111. Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202, 9 2013.

112. M. J. Nueda, S. Tarazona, and A. Conesa. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics*, 30(18):2598–2602, 9 2014.

113. Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing on JSTOR. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

114. Ahmed A. Metwally, Patricia W. Finn, Yang Dai, and David L. Perkins. Detection of Differential Abundance Intervals in Longitudinal Metagenomic Data Using Negative Binomial Smoothing Spline ANOVA. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB '17*, pages 295–304, Boston, Massachusetts, USA, 2017. ACM Press.

115. Ahmed A. Metwally, Jie Yang, Christian Ascoli, Yang Dai, Patricia W. Finn, and David L. Perkins. MetaLonDA: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. *Microbiome*, 6(1):32, 2018.

116. J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, 5 2010.

117. Sahar Abubucker, Nicola Segata, Johannes Goll, Alyxandria M. Schubert, Jacques Izard, Brandi L. Cantarel, Beltran Rodriguez-Mueller, Jeremy Zucker, Mathangi Thiagarajan, Bernard Henrissat, Owen White, Scott T. Kelley, Barbara Methé, Patrick D.

Schloss, Dirk Gevers, Makedonka Mitreva, and Curtis Huttenhower. Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. *PLoS Computational Biology*, 8(6):e1002358, 6 2012.

118. Mikael Wallroth. Normalization of metagenomic data A comprehensive evaluation of existing methods.

119. J Paul Brooks, David J Edwards, Michael D Harwich, Maria C Rivera, Jennifer M Fettweis, Myrna G Serrano, Robert A Reris, Nihar U Sheth, Bernice Huang, Philippe Girerd, Jerome F Strauss, Kimberly K Jefferson, and Gregory A Buck. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology*, 15(1):66, 12 2015.

120. Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde, and Rob Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27, 12 2017.

121. Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 12 2014.

122. Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

123. W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.

124. Simon Jackman. pscl: Political Science Computational Laboratory, 2015.

125. Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression Models for Count Data in R. *Journal of Statistical Software*, 27(8):1–25, 7 2008.

126. Thomas W. Yee. *Vector Generalized Linear and Additive Models*. Springer Series in Statistics. Springer New York, New York, NY, 2015.

127. Thomas W. Yee. VGAM: Vector Generalized Linear and Additive Models, 2015.

128. Indra M. Chakravarti, R.G Laha, and J Roy. Handbook of methods of applied statistics. John Wiley & Sons, Hoboken, 1967.

129. Viktor Jonsson, Tobias Österlund, Olle Nerman, and Erik Kristiansson. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics*, 17(1):78, 12 2016.

130. Vinzenz Erhardt. corcounts, 2015.

131. Christopher. Chatfield. The Analysis of Time Series: An Introduction. 2016.

132. William S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 12 1979.

133. Yeneneh Haileselassie, Marit Navis, Nam Vu, Khaleda Rahman Qazi, Bence Rethi, and Eva Sverremark-Ekström. Lactobacillus reuteri and Staphylococcus aureus differentially influence the generation of monocyte-derived dendritic cells and subsequent autologous T cell responses. *Immunity, inflammation and disease*, 4(3):315–26, 9 2016.

134. Ningwen Tai, Jian Peng, Fuqiang Liu, Elke Gulden, Youjia Hu, Xiaojun Zhang, Li Chen, F Susan Wong, and Li Wen. Microbial antigen mimics activate diabetogenic CD8 T cells in NOD mice. *The Journal of experimental medicine*, 213(10):2129–46, 9 2016.

135. Dan Knights, Elizabeth K. Costello, and Rob Knight. Supervised classification of human microbiota. *FEMS Microbiology Reviews*, 35(2):343–359, 3 2011.

136. Ruchi S Gupta, Elizabeth E Springston, Manoj R Warrier, Bridget Smith, Rajesh Kumar, Jacqueline Pongracic, and Jane L Holl. The prevalence, severity, and distribution of childhood food allergy in the United States. *Pediatrics*, 128(1):9–17, 7 2011.

137. Andrew T Stefka, Taylor Feehley, Prabhanshu Tripathi, Ju Qiu, Kathy McCoy, Sarkis K Mazmanian, Melissa Y Tjota, Goo-Young Seo, Severine Cao, Betty R Theriault, Dionysios A Antonopoulos, Liang Zhou, Eugene B Chang, Yang-Xin Fu, and Cathryn R Nagler. Commensal bacteria protect against food allergen sensitization. *Proceedings of the National Academy of Sciences*, 111(36):13145–13150, 9 2014.

138. Magali Noval Rivas, Oliver T. Burton, Petra Wise, Yu-qian Zhang, Suejy A. Hobson, Maria Garcia Lloret, Christel Chehoud, Justin Kuczynski, Todd DeSantis, Janet Warrington, Embriette R. Hyde, Joseph F. Petrosino, Georg K. Gerber, Lynn Bry, Hans C. Oettgen,

Sarkis K. Mazmanian, and Talal A. Chatila. A microbiota signature associated with experimental food allergy promotes allergic sensitization and anaphylaxis. *Journal of Allergy and Clinical Immunology*, 131(1):201–212, 1 2013.

139. Jessica H. Savage, Kathleen A. Lee-Sarwar, Joanne Sordillo, Supinda Bunyavanich, Yanjiao Zhou, George O'Connor, Megan Sandel, Leonard B. Bacharier, Robert Zeiger, Erica Sodergren, George M. Weinstock, Diane R. Gold, Scott T. Weiss, and Augusto A. Litonjua. A prospective microbiome-wide association study of food sensitization and food allergy in early childhood. *Allergy*, 73(1):145–152, 1 2018.

140. Xing Hua, James J. Goedert, Angela Pu, Guoqin Yu, and Jianxin Shi. Allergy associations with the adult fecal microbiota: Analysis of the American Gut Project. *EBioMedicine*, 3:172–179, 1 2016.

141. Zongxin Ling, Zailing Li, Xia Liu, Yiwen Cheng, Yueqiu Luo, Xiaojuan Tong, Li Yuan, Yuezhu Wang, Jinbo Sun, Lanjuan Li, and Charlie Xiang. Altered fecal microbiota composition associated with food allergy in infants. *Applied and environmental microbiology*, 80(8):2546–54, 4 2014.

142. Supinda Bunyavanich, Nan Shen, Alexander Grishin, Robert Wood, Wesley Burks, Peter Dawson, Stacie M. Jones, Donald Y.M. Leung, Hugh Sampson, Scott Sicherer, and Jose C. Clemente. Early-life gut microbiome composition and milk allergy resolution. *Journal of Allergy and Clinical Immunology*, 138(4):1122–1130, 10 2016.

143. Jian Tan, Craig McKenzie, Peter J. Vuillermin, Gera Goverse, Carola G. Vinuesa, Reina E. Mebius, Laurence Macia, and Charles R. Mackay. Dietary Fiber and Bacterial SCFA Enhance Oral Tolerance and Protect against Food Allergy through Diverse Cellular Pathways. *Cell Reports*, 15(12):2809–2824, 6 2016.

144. Roberto Berni Canani, Naseer Sangwan, Andrew T Stefka, Rita Nocerino, Lorella Paparo, Rosita Aitoro, Antonio Calignano, Aly A Khan, Jack A Gilbert, and Cathryn R Nagler. Lactobacillus rhamnosus GG-supplemented formula expands butyrate-producing bacterial strains in food allergic infants. *The ISME Journal*, 10(3):742–750, 3 2016.

145. Ravi Ranjan, Asha Rani, Ahmed Metwally, Halvor S. McGee, and David L. Perkins. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, 469(4):967–977, 12 2015.

146. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, 2016.

147. Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, 2013.

148. Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. 8 2015.

149. Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech Recognition with Deep Recurrent Neural Networks. 3 2013.

150. Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

151. Richard H. R. Hahnloser, Rahul Sarpeshkar, Misha A. Mahowald, Rodney J. Douglas, and H. Sebastian Seung. Erratum: Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, 6 2000.

152. Andrew Ng. Sparse autoencoder. 2011.

153. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 3 2016.

154. Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. 12 2014.

155. Philip M. Farrell, Terry B. White, Clement L. Ren, Sarah E. Hempstead, Frank Accurso, Nico Derichs, Michelle Howenstine, Susanna A. McColley, Michael Rock, Margaret Rosenfeld, Isabelle Sermet-Gaudelus, Kevin W. Southern, Bruce C. Marshall, and Patrick R. Sosnay. Diagnosis of Cystic Fibrosis: Consensus Guidelines from the Cystic Fibrosis Foundation. *The Journal of Pediatrics*, 181:S4–S15.e1, 2 2017.

156. Todd MacKenzie, Alex H. Gifford, Kathryn A. Sabadosa, Hebe B. Quinton, Emily A. Knapp, Christopher H. Goss, and Bruce C. Marshall. Longevity of Patients With Cystic Fibrosis in 2000 to 2010 and Beyond: Survival Analysis of the Cystic Fibrosis Foundation Patient Registry. *Annals of Internal Medicine*, 161(4):233, 8 2014.

157. Jessica E Pittman and Thomas W Ferkol. The Evolution of Cystic Fibrosis Care. *Chest*, 148(2):533–542, 8 2015.

158. J Stuart Elborn. Cystic fibrosis. *The Lancet*, 388(10059):2519–2531, 11 2016.

159. Brian P O'Sullivan and Steven D Freedman. Cystic fibrosis. *Lancet (London, England)*, 373(9678):1891–904, 5 2009.

160. Arnon Elizur, Carolyn L. Cannon, and Thomas W. Ferkol. Airway Inflammation in Cystic Fibrosis. *Chest*, 133(2):489–495, 2 2008.

161. Theodore G. Liou, Frederick R. Adler, David R. Cox, and Barbara C. Cahill. Lung Transplantation and Survival in Children with Cystic Fibrosis. *New England Journal of Medicine*, 357(21):2143–2152, 11 2007.

162. David Weill, Christian Benden, Paul A Corris, John H Dark, R Duane Davis, Shaf Keshavjee, David J Lederer, Michael J Mulligan, G Alexander Patterson, Lianne G Singer, Greg I Snell, Geert M Verleden, Martin R Zamora, and Allan R Glanville. A consensus document for the selection of lung transplant candidates: 2014–an update from the Pulmonary Transplantation Council of the International Society for Heart and Lung Transplantation. *The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation*, 34(1):1–15, 1 2015.

163. Stephen Kirkby, Don Hayes, and Jr. Pediatric lung transplantation: indications and outcomes. *Journal of thoracic disease*, 6(8):1024–31, 8 2014.

164. Samuel B Goldfarb, Bronwyn J Levvey, Wida S Cherikh, Daniel C Chambers, Kiran Khush, Anna Y Kucheryavaya, Lars H Lund, Bruno Meiser, Joseph W Rossano, Roger D Yusen, Josef Stehlik, and International Society for Heart and Lung Transplantation. Registry of the International Society for Heart and Lung Transplantation: Twentieth Pediatric Lung and Heart-Lung Transplantation Report-2017; Focus Theme: Allograft ischemic time. *The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation*, 36(10):1070–1079, 10 2017.

165. Anne L Stephenson, Melissa Tom, Yves Berthiaume, Lianne G Singer, Shawn D Aaron, G A Whitmore, and Sanja Stanojevic. A contemporary survival analysis of individuals with cystic fibrosis: a cohort study. *The European respiratory journal*, 45(3):670–9, 3 2015.

166. Angela Koutsokera, Pierre J. Royer, Jean P. Antonietti, Andreas Fritz, Christian Benden, John D. Aubert, Adrien Tissot, Karine Botturi, Antoine Roux, Martine L. Reynaud-Gaubert, Romain Kessler, Claire Dromer, Sacha Mussot, Hervé Mal, Jean-François Mornex, Romain Guillemain, Christiane Knoop, Marcel Dahan, Paola M. Soccal, Johanna Claustre, Edouard Sage, Carine Gomez, Antoine Magnan, Christophe Pison, Laurent P. Nicod, and The SysCLAD Consortium. Development of a Multivariate Prediction Model for Early-Onset Bronchiolitis Obliterans Syndrome and Restrictive Allograft Syndrome in Lung Transplantation. *Frontiers in Medicine*, 4:109, 7 2017.

167. Geert M Verleden, Ganesh Raghu, Keith C Meyer, Allan R Glanville, and Paul Corris. A new classification system for chronic lung allograft dysfunction. *The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation*, 33(2):127–33, 2 2014.

168. Paul D. Robinson, Helen Spencer, and Paul Aurora. Impact of lung function interpretation approach on pediatric bronchiolitis obliterans syndrome diagnosis after lung transplantation. *The Journal of Heart and Lung Transplantation*, 34(8):1082–1088, 8 2015.

169. Marc Estenne and Marshall I Hertz. Bronchiolitis obliterans after human lung transplantation. *American journal of respiratory and critical care medicine*, 166(4):440–4, 8 2002.

170. Sara A. Hennessy, Tjasa Hranjec, Brian R. Swenson, Benjamin D. Kozower, David R. Jones, Gorav Ailawadi, Irving L. Kron, and Christine L. Lau. Donor Factors Are Associated With Bronchiolitis Obliterans Syndrome After Lung Transplantation. *The Annals of Thoracic Surgery*, 89(5):1555–1562, 5 2010.

171. Eric S Weiss, Jeremiah G Allen, Christian A Merlo, John V Conte, and Ashish S Shah. Factors indicative of long-term survival after lung transplantation: a review of 836 10-year survivors. *The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation*, 29(3):240–6, 3 2010.

172. Tereza Martinu, Dong-Feng Chen, and Scott M. Palmer. Acute Rejection and Humoral Sensitization in Lung Transplant Recipients. *Proceedings of the American Thoracic Society*, 6(1):54–65, 1 2009.

173. Keith C Meyer, Ganesh Raghu, Geert M Verleden, Paul A Corris, Paul Aurora, Kevin C Wilson, Jan Brozek, Allan R Glanville, the ISHLT/ATS/ERS BOS Task Force ISHLT/ATS/ERS BOS Task Force Committee, and ISHLT/ATS/ERS BOS Task Force Committee. An international ISHLT/ATS/ERS clinical practice guideline: diagnosis and management of bronchiolitis obliterans syndrome. *The European respiratory journal*, 44(6):1479–503, 12 2014.

174. Gilles Devouassoux, Christian Drouet, Isabelle Pin, Christian Brambilla, Elisabeth Brambilla, Pierre-Emmanuel Colle, and Christophe Pison. Alveolar neutrophilia is a predictor for the bronchiolitis obliterans syndrome, and increases with degree of severity. *Transplant Immunology*, 10(4):303–310, 11 2002.

175. Claus Neurohr, Patrick Huppmann, Benedikt Samweber, Stefan Leuschner, Gregor Zimmermann, Hanno Leuchte, Rainer Baumgartner, Rudolf Hatz, Ludwig Frey, Peter Ueberfuhr, Iris Bittmann, Juergen Behr, and Munich Lung Transplant Group. Prognostic value of bronchoalveolar lavage neutrophilia in stable lung transplant recipients. *The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation*, 28(5):468–74, 5 2009.

176. MARTINE REYNAUD-GAUBERT, PASCAL THOMAS, MONIQUE BADIER, PIERRE CAU, ROGER GIUDICELLI, and PIERRE FUENTES. Early Detection of Airway Involvement in Obliterative Bronchiolitis after Lung Transplantation. *American Journal of Respiratory and Critical Care Medicine*, 161(6):1924–1929, 6 2000.

177. Cody Schott, S. Samuel Weigt, Benjamin A. Turturice, Ahmed Metwally, John Belperio, Patricia W. Finn, and David L. Perkins. Bronchiolitis obliterans syndrome susceptibility and the pulmonary microbiome. *The Journal of Heart and Lung Transplantation*, 4 2018.

178. R Vos, B M Vanaudenaerde, N Geudens, L J Dupont, D E Van Raemdonck, and G M Verleden. Pseudomonal airway colonisation: risk factor for bronchiolitis obliterans syndrome after lung transplantation? *The European respiratory journal*, 31(5):1037–45, 5 2008.

179. Robert P. Dickson, John R. Erb-Downward, Christine M. Freeman, Natalie Walker, Brittan S. Scales, James M. Beck, Fernando J. Martinez, Jeffrey L. Curtis, Vibha N. Lama, and Gary B. Huffnagle. Changes in the Lung Microbiome following Lung Transplantation Include the Emergence of Two Distinct Pseudomonas Species with Distinct Clinical Associations. *PLoS ONE*, 9(5):e97214, 5 2014.

180. Patricia Moran Losada, Philippe Chouvarine, Marie Dorda, Silke Hedtfeld, Samira Mielke, Angela Schulz, Lutz Wiehlmann, and Burkhard Tümmler. The cystic fibrosis lower airways microbial metagenome. *ERJ Open Research*, 2(2):00096–2015, 4 2016.

181. Saad A. Syed, Fiona J. Whelan, Barbara Waddell, Harvey R. Rabin, Michael D. Parkins, and Michael G. Surette. Reemergence of Lower-Airway Microbiota in Lung Transplant Patients with Cystic Fibrosis. *Annals of the American Thoracic Society*, 13(12):2132–2142, 12 2016.

182. Katherine B Frayman, David S Armstrong, Rosemary Carzino, Thomas W Ferkol, Keith Grimwood, Gregory A Storch, Shu Mei Teo, Kristine M Wylie, and Sarath C Ranganathan. The lower airway microbiota in early cystic fibrosis lung disease: a longitudinal analysis. *Thorax*, 72(12):1104–1112, 12 2017.

183. David N O'Dwyer, Robert P Dickson, and Bethany B Moore. The Lung Microbiome, Immunity, and the Pathogenesis of Chronic Lung Disease. *Journal of immunology (Baltimore, Md. : 1950)*, 196(12):4839–47, 6 2016.

184. C. Martin-Gandul, N. J. Mueller, M. Pascual, and O. Manuel. The Impact of Infection on Chronic Allograft Dysfunction and Allograft Survival After Solid Organ Transplantation. *American Journal of Transplantation*, 15(12):3024–3040, 12 2015.

185. H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other.

186. Charles E. McCulloch, S. R. Searle, and John M. Neuhaus. *Generalized, linear, and mixed models*. John Wiley & Sons, Ltd, 2001.

187. T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, 40(20):e155–e155, 11 2012.

188. Paul J. McMurdie and Susan Holmes. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, 8(4):e61217, 4 2013.

189. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, and C. H. Wu. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 3 2015.

190. Ron Caspi, Richard Billington, Carol A Fulcher, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Peter E Midford, Quang Ong, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, and Peter D Karp. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Research*, 46(D1):D633–D639, 1 2018.

191. David A Relman. The human microbiome: ecosystem resilience and health. *Nutrition Reviews*, 70:S2–S9, 8 2012.

192. June L. Round and Sarkis K. Mazmanian. The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology*, 9(5):313–323, 5 2009.

193. Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, 1 2015.

# APPENDICES

# Appendix A

## *WEVOTE* AND *WEVOTE-WEB* PACKAGES

**Source-code**

- `https://github.com/aametwally/WEVOTE`
- `https://github.com/aametwally/WEVOTE-web`

**Publications:**

- **Metwally AA**, Dai Y, Finn PW, Perkins DL. WEVOTE: Weighted Voting Taxonomic Identification Method of Microbial Sequences. *PLoS ONE*, 2016.
- Alaa A, **Metwally AA**. Cloud-based Solution for Improving Usability and Interactivity of Metagenomic Ensemble Taxonomic Identification Methods. *IEEE EMBS Biomedical and Health Informatics*, 2018.

# Appendix B

## *METALONDA* PACKAGE

**Source-code**

- `https://CRAN.R-project.org/package=MetaLonDA`
- `https://github.com/aametwally/MetaLonDA`

**Publications:**

- **Metwally AA**, Yang J, Ascoli C, Dai Y, Finn PW, Perkins DL. MetaLonDA: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. *Microbiome*, 2018.
- **Metwally AA**, Finn PWF, Dai Y, Perkins DL. Detection of Differential Abundance Intervals in Longitudinal Metagenomic Data Using Negative Binomial Smoothing Spline ANOVA. *In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017.

# Appendix C

## COPYRIGHT PERMISSIONS

# Copyright and License

- Copyright on any open access article in a journal published by BioMed Central is retained by the author(s).
- Authors grant BioMed Central a license to publish the article and identify itself as the original publisher.
- Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors, citation details and publisher are identified.
- The Creative Commons Attribution License 4.0 formalizes these and other terms and conditions of publishing articles.
- In accordance with our Open Data policy, the Creative Commons CC0 1.0 Public Domain Dedication waiver applies to all published data in BioMed Central open access articles.

Where an author is prevented from being the copyright holder (for instance in the case of US government employees or those of Commonwealth governments), minor variations may be required. In such cases the copyright line and license statement in individual articles will be adjusted, for example to state '© 2016 Crown copyright'. Authors requiring a variation of this type should inform BioMed Central during or immediately after submission of their article. Changes to the copyright line cannot be made after publication of an article.

## Exceptions to copyright policy

Our policy pages provide details concerning copyright and licensing for articles which were previously published under policies that are different from the above. For instance, occasionally BioMed Central may co-publish articles jointly with other publishers, and different licensing conditions may then apply. In all such cases, however, access to these articles is free from fees or any other access restrictions.

Information specifically regarding permissions and reprints can be found here. Please contact us if there are questions.




# License

**Title:** Detection of Differential Abundance Intervals in Longitudinal Metagenomic Data Using Negative Binomial Smoothing Spline ANOVA

**Author:** Ahmed A. Metwally, et al

**Publication:** Proceedings

**Publisher:** Association for Computing Machinery, Inc.

**Date:** Aug 20, 2017



## Order Completed

Thank you for your order.

This Agreement between Mr. Ahmed Metwally ("You") and Association for Computing Machinery, Inc. ("Association for Computing Machinery, Inc.") consists of your license details and the terms and conditions provided by Association for Computing Machinery, Inc. and Copyright Clearance Center.

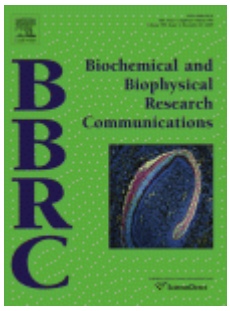
Your confirmation email will contain your order number for future reference.

printable details

| | |
|---|---|
| License Number | 4341531503738 |
| License date | May 03, 2018 |
| Licensed Content Publisher | Association for Computing Machinery, Inc. |
| Licensed Content Publication | Proceedings |
| Licensed Content Title | Detection of Differential Abundance Intervals in Longitudinal Metagenomic Data Using Negative Binomial Smoothing Spline ANOVA |
| Licensed Content Author | Ahmed A. Metwally, et al |
| Licensed Content Date | Aug 20, 2017 |
| Type of Use | Thesis/Dissertation |
| Requestor type | Author of this ACM article |
| Is reuse in the author's own new work? | Yes |
| Format | Electronic |
| Portion | Full article |
| Will you be translating? | No |
| Order reference number | |
| Title of your thesis/dissertation | Computational Methods for Longitudinal Microbiome Analysis: Identification, Modeling, and Classification |
| Expected completion date | Aug 2018 |
| Estimated size (pages) | 220 |
| Attachment | |
| Requestor Location | Mr. Ahmed Metwally<br>909 S. Wolcott Avenue<br><br>CHICAGO, IL 60612<br>United States<br>Attn: Mr. Ahmed Metwally |
| Billing Type | Credit Card |
| Credit card info | Master Card ending in 3081 |
| Credit card expiration | 09/2018 |

| | |
|---|---|
| Total | 8.00 USD |

ORDER MORE CLOSE WINDOW



**Title:** Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing

**Author:** Ravi Ranjan,Asha Rani,Ahmed Metwally,Halvor S. McGee,David L. Perkins

**Publication:** Biochemical and Biophysical Research Communications

**Publisher:** Elsevier

**Date:** 22 January 2016


**Title:** Cloud-based solution for improving usability and interactivity of metagenomic ensemble taxonomic classification methods

**Conference Proceedings:** Biomedical & Health Informatics (BHI), 2018 IEEE EMBS International Conference on

**Author:** Asem Alaa

**Publisher:** IEEE

**Date:** March 2018



## Thesis / Dissertation Reuse

**The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:**

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

| | |
|---|---|
| **Title:** | Using convolutional neural networks to explore the microbiome |
| **Conference Proceedings:** | Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE |
| **Author:** | Derek Reiman |
| **Publisher:** | IEEE |
| **Date:** | July 2017 |



## Thesis / Dissertation Reuse

**The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:**

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

# VITA

# Ahmed A. Metwally

**Education**

***Ph.D. - Bioinformatics/Bioengineering*** (3.85/4.0) **Aug. 2014 – July 2018**
**University of Illinois at Chicago**, Chicago, IL

- Advisors: Prof. Yang Dai & Prof. David L. Perkins
- Thesis: Computational Methods for Longitudinal Microbiome Analysis: Identification, Modeling, and Classification

***M.Sc. - Computer Science*** (3.88/4.0) **Aug. 2016 – May 2018**
**University of Illinois at Chicago**, Chicago, IL

- Advisor: Prof. Philip S. Yu
- Project: Deep Learning Framework for Predicting Food Allergy from Longitudinal Microbiome Taxonomic Profiles

***M.Sc. - Biomedical Engineering*** (3.75/4.0) **Sept. 2011 – May 2014**
**Cairo University**, Giza, Egypt

- Advisor: Prof. Mohamed Abouelhoda
- Thesis: Cloud-based Distributed Suffix Array Construction Algorithms for Indexing Biological Data

***B.Sc. - Biomedical Engineering*** (3.96/4.0 & ranked 1st) **Sept. 2005 – May 2010**
**Cairo University**, Giza, Egypt

- Concentration: Medical Signal/Image Processing

**Research Interests**

Bioinformatics, Systems Biology, Deep Learning, Network Science, Graph Theory, Cancer Genomics, Precision Medicine.

**Honors & Awards**

- American Thoracic Society Scholarship (Pediatric Assembly) **2018**
- UIC Impact Scholar **2018**
- NIH Predoctoral Education for Clinical and Translational Scientists Fellowship (UL1TR002003) ($33,000). **2017**
- American Heart Association Predoctoral Fellowship (Percentile rank: 19%, 2017 funding cutoff: 18%). **2017**
- 2nd place poster winner, ISCB GLBIO, Chicago. **2017**
- Scientific Excellence Award, Department of Medicine, UIC. **2017**
- Best intern's poster award, Thermo Fisher Scientific. **2016**
- 1st place award at UIC Research Forum among participants from Engineering, CS, Math, and Statistics departments. **2016**
- UIC Chancellor's Graduate Research Award ($8,000). **2015**
- Distinctive honor and ranked 1st among undergraduate class. **2010**
- DAAD fellowship to intern at Helmholtz-Zentrum Geesthacht, Germany. **2009**
- Travel awards from: NSF for IEEE BHI in Las Vegas (2018), ACM BCB in Boston (2017), and Phylogenomics Software School in Austin (2016). IEEE for Neural Engineering Summer School in Shanghai (2013). International Society for Computational Biology (ISCB) for InCoB in Sydney (2014). University of Illinois at

Urbana-Champaign for Phylogenomics Symposium in Ann Arbor (2015). International Center for Theoretical Physics (ICTP) for Advanced Techniques for Scientific Programming School in Italy (2014). LinkSCEEM for Conference on Scientific Computing in Cyprus (2013).

**Publications**

- **Metwally AA**, Yang J, Ascoli C, Dai Y, Finn PW, Perkins DL, "MetaLonDA: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies", Microbiome, 2018.
- Alaa A, **Metwally AA**, "Cloud-based Solution for Improving Usability and Interactivity of Metagenomic Ensemble Taxonomic Identification Methods", IEEE Biomedical and Health Informatics, 2018.
- Reiman D, **Metwally AA**, Dai Y, "PopPhy-CNN: A Phylogentic Tree Embedded Architecture for Convolution Neural Networks for Metagenomic Data", bioRxiv, 2018.
- Schott C, Weigt S, Turturice BA, **Metwally AA**, Belperio J, Finn PW, Perkins DL, "Broncholitis Obliterans Syndrome Susceptibility and the Pulmonary Microbiome", Journal of Heart and Lung Transplantation, 2018.
- Ascoli C, Huang Y, Schott C, Turturice B, **Metwally AA**, Perkins DL, Finn PW, ACCESS Research Group, "A Circulating Micro-RNA Signature Serves as a Diagnostic and Prognostic Indicator in Sarcoidosis." American Journal of Respiratory Cell and Molecular Biology, 2018.
- **Metwally AA**, Finn PWF, Dai Y, Perkins DL, "Detection of Differential Abundance Intervals in Longitudinal Metagenomic Data Using Negative Binomial Smoothing Spline ANOVA." ACM BCB, 2017.
- Reiman D*, **Metwally AA***, Dai Y, "Using Convolutional Neural Networks to Explore the Microbiome." IEEE EMBC, 2017.
- Kwan J, Hajjiri Z, Chen Y, **Metwally AA**, Perkins DL, Finn PW, "Donor and Recipient Ethnicity Impacts Renal Graft Adverse Outcomes." Journal of Racial and Ethnic Health Disparities, 2017.
- **Metwally AA**, Dai Y, Finn PW, Perkins DL, "WEVOTE: Weighted Voting Taxonomic Identification Method of Microbial Sequences." PLoS ONE, 2016.
- **Metwally AA**, Kandil AH, Abouelhoda M, "Distributed suffix array construction algorithms: Comparison of two algorithms." IEEE CIBEC, 2016.
- Rani R, Ranjan R, McGee H, **Metwally AA**, Hajjiri Z, Brennan D, Finn PW, Perkins DL, "A diverse virome in kidney transplant patients contains multiple viral subtypes with distinct polymorphisms." Scientific Reports, 2016.
- Kwan J, Hajjiri Z, **Metwally AA**, Finn PW, Perkins DL, "Effect of the Obesity Epidemic on Kidney Transplantation: Obesity Is Independent of Diabetes as a Risk Factor for Adverse Renal Transplant Outcomes." PLoS ONE, 2016.
- Ranjan R, Rani R, **Metwally AA**, McGee H, Perkins DL, "Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing." Biochemical and Biophysical Research Communications, 2016.
- **Metwally AA**, Kandil AH, Abouelhoda M, "Cloud-based parallel suffix array construction based on MPI." IEEE MECBME, 2014.

**Software (Publicly available)**

- **MetaLonDA** (`https://CRAN.R-project.org/package=MetaLonDA`) Identify time intervals of differentially abundant metagenomic features in longitudinal studies.
- **WEVOTE** (`https://github.com/aametwally/WEVOTE`) Weighted voting taxonomic classifier of metagenomic sequences.
- **WEVOTE-web** (`https://github.com/aametwally/WEVOTE-web`) Cloud-based framework of the WEVOTE method.

See `https://github.com/aametwally` for more.

**Professional Experience**

**IEEE Engineering in Medicine and Biology Society (EMBS)**, Piscataway,NJ

***Global Student Representative and Board Member*** **Jan. 2017 – Present**

EMBS is the world's largest international society of biomedical engineers. Being the elected student representative, I speak on behalf of all EMB undergraduate/graduate students and represent the students' views during the EMBS board meetings.

**University of Illinois at Chicago**, Chicago, IL

***Graduate Research Assistant*** **Aug. 2014 – Present**

Developing computational methods to analyze longitudinal microbiome data. Studying the role of lung microbiome in pulmonary diseases progression and transplant rejection. Developing novel deep learning approaches to analyze microbiome data.

**Thermo Fisher Scientific**, South San Francisco, CA

***Bioinformatics Lead Intern*** **May 2016 – Sept. 2016**

Developed a new method that uses simulated annealing optimization to identify a small group of primers to maximally distinguish genes or isoforms of high homology. The output increases the resolution of the Immune-Oncology sequencing panels.

**Thermo Fisher Scientific**, South San Francisco, CA

***Software Engineering Intern*** **May 2015 – Sept. 2015**

Developed a tool called "RunComparator" to evaluate the analysis pipeline of IonTorrent sequencing machines.

**Nile University**, Giza, Egypt

***Junior Scientist*** **Oct. 2012 – Oct. 2013**

Parallelized string matching algorithms that based on suffix tree along with their applications in Bioinformatics alignment algorithms. Developed a cloud-based parallel maximum exact matching algorithm.

**Cairo University**, Giza, Egypt

***Teaching and Research Assistant*** **Sept. 2010 – Aug. 2014**

Taught undergraduate courses: Algorithms, High-Performance Computing, Database, and Healthcare Information Systems. Designed a Radiology Information System.

**Helmholtz-Zentrum Geesthacht**, Berlin, Germany

***Research Intern*** **May 2009 – Aug. 2009**

Engineered an algorithm that can classify various classes of blood vessels by processing microscopic images. Performed biocompatibility tests using HET-CAM methods.

**Talks**

- Harvard Medical School, Boston, "Computational Microbiome Methods in a Longitudinal Pediatric Lung Transplant Cohort". **2018**
- Stanford University, Palo Alto, "Modeling Microbiome Longitudinal Data". **2018**
- IEEE BHI, Las Vegas, "Improving Usability and Interactivity of Metagenomic Ensemble Taxonomic Classification Methods". **2018**
- ACM BCB, Boston, "Detection of Differential Abundance Intervals in Longitudinal Metagenomic studies". **2017**
- IEEE EMBC, Jeju Island, South Korea, "Using Convolutional Neural Networks to Explore the Microbiome". **2017**
- ISCB GLBIO, Chicago, "Microbiome Dynamics as Predictors of Lung Transplant Rejection". **2017**
- COM Research Forum, UIC, "Classification of Metagenomics Data Using Deep Neural Networks". **2016**
- UIC Research Forum, UIC, "WEVOTE: Weighted Voting Taxonomic Identification Method of Microbial Sequences". **2016**
- CBC, Chicago, "Metagenomic Analysis Pipeline: Updating the Microbial Gene Catalog". **2015**
- ISCB InCob'14, Sydney, Australia, "CloudSACA: Distributed Suffix Array Construction Algorithms Package on Cloud". **2014**
- ICTP, Trieste, Italy, "Simulation of Physics Behind the Bike Movement using Genetic Algorithm". **2014**
- MECBME, Doha, Qatar, "Cloud-based Parallel Suffix Array Construction Based on MPI". **2014**
- CSC, Paphos, Cyprus, "Parallel Alignment and Indexing Algorithms based on Distributed and Shared Memory Architecture for Genomics Data". **2013**

**Professional Services**

**Program Committee:** IEEE BHI and IEEE MECBME. **2017 - 2018**

**Paper Review:** IEEE Journal of Biomedical and Health Informatics (JBHI), IEEE MECBME, IEEE BHI, IEEE EMBC. **2014 - Present**

**Organizing Committee:** IEEE EMBC'17, IEEE EMBC'18.

**Conference Chair:** $1^{st}$ IEEE EMBS International Student Conference. **2013**