# STATISTICAL METHODS FOR FUNCTIONAL METAGENOMIC ANALYSIS BASED ON NEXT-GENERATION SEQUENCING DATA

by

Naruekamol Pookhao

---

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF AGRICULTURAL AND BIOSYSTEMS ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2014

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Naruekamol Pookhao titled "Statistical methods for functional metagenomic analysis based on next-generation sequencing data" and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

_____ Date: 4/17/2014
Lingling An (Agricultural and Biosystems Engineering)

_____ Date: 4/17/2014
Donald Slack (Agricultural and Biosystems Engineering)

_____ Date: 4/17/2014
Dean Billheimer (Interdisciplinary Programs in Statistics)

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 4/17/2014
Dissertation Director: Lingling An

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Naruekamol Pookhao

# ACKNOWLEDGEMENTS

I would like to acknowledge the following for their valuable contributions to this research and to the development of this dissertation:

The National Science Foundation (NSF) and The Cecil Miller Endowment, The University of Arizona Foundation for their financial support.

I am thankful to my advisor, Dr. Lingling An, a good friend and mentor. I am very appreciate for the assistance, patience, and guidance that she gave me through my research.

I would like to extend my appreciation to my committee members, Dr. Donald Slack, Dr. Mark Riley, and Dr. Dean Billheimer for their direction and expertise enabling this project and dissertation to be as successful as it was.

Furthermore, I thank my colleagues in my research group, Michael Sohn, Qike Li, Isaac Jenkins, and Roufei Du for their comments, support, and assistance.

I also extend my thanks to the Agricultural and Biosystems Engineering Department for their compassion which made me feel accepted as a friend and a part of their team.

Finally, I would like to give my special thanks to my family, especially my parents, and friends for their endless love and support.

# TABLE OF CONTENTS

# ABSTRACT

Metagenomics is the study of a collective microbial genetic content recovered directly from natural (e.g., soil, ocean, and freshwater) or host-associated (e.g., human gut, skin, and oral) environmental communities that contain microorganisms, i.e., microbiomes. The rapid technological developments in next generation sequencing (NGS) technologies, enabling to sequence tens or hundreds of millions of short DNA fragments (or reads) in a single run, facilitates the studies of multiple microorganisms lived in environmental communities. Metagenomics, a relatively new but fast growing field, allows us to understand the diversity of microbes, their functions, cooperation, and evolution in a particular ecosystem. Also, it assists us to identify significantly different metabolic potentials in different environments. Particularly, metagenomic analysis on the basis of functional features (e.g., pathways, subsystems, functional roles) enables to contribute the genomic contents of microbes to human health and leads us to understand how the microbes affect human health by analyzing a metagenomic data corresponding to two or multiple populations with different clinical phenotypes (e.g., diseased and healthy, or different treatments). Currently, metagenomic analysis has substantial impact not only on genetic and environmental areas, but also on clinical applications. In our study, we focus on the development of computational and statistical methods for functional metagnomic analysis of sequencing data that is obtained from various environmental microbial samples/communities.

# 1. INTRODUCTION

## 1.1 Next Generation DNA Sequencing

DNA sequencing, a process to determine the precise order of a particular DNA molecule, was initially introduced in the literatures in 1977 (Maxam and Gilbert, 1977; Sanger et al. (1977); ten Bosch et al., 2008; Tucker et al., 2009). The DNA sequence production relied on capillary-based, semi-automated implementations of the Sanger biochemistry (Figure 1a) (Hunkapiller et al., 1991; Shendure and Ji 2008; Swerdlow et al., 1990). Later, in high-throughput shotgun Sanger sequencing production pipelines, genomic DNA to be sequenced is first fragmented, and then cloned to a high-copy-number plasmid vector, which is used to transform *Escherichia coli.* This approach results in an amplified template comprising many 'clonal' copies of a single plasmid insert. For each sequencing reaction, which takes place within a microliter-scale volume, a single bacterial colony is picked; then plasmid DNA is isolated; a ladder of dideoxynucleotides (ddNTP)-terminated, fluorescently dye-labeled products is generated finally. In one run of a sequencing instrument, those ddNTP-terminated, fluorescently dye-labeled products are subjected to high-resolution electrophoretic separation within one of 96 or 384 capillaries. The label on the terminating ddNTP of any given fragment determines the nucleotide identity of its terminal position. And then, when fluorescently dye-labeled DNA fragments pass a detector, a sequencing trace is generated using the four-channel emission spectrum (Shendure and Ji 2008). Capillary-based, semi-automated Sanger sequencing has been used for over the years for many applications, i.e.,

chemistry and clinical study of both small-scale (kilobase) and larger-scale (megabase) projects. Sanger sequencing can achieve read lengths of ~1 Kb and high accuracies at a cost of about $500 per megabase (Mb). However, this approach is not powerful for further increasing in throughput, and the cost of sequencing cannot be reduced since this approach depends on lengthy procedures. An alternative strategy for DNA sequencing is needed for large-scale production of genomic sequence (Tucker et al., 2009). DNA sequencing has been continuously improving in its speed, efficiency, and cost-effectiveness over decades (ten Bosch et al., 2008).

Several new sequencing technologies, called "next generation sequencing (NGS)", have been developed and become such alternative approach (Mardis 2008; Shendure and Ji 2008; Tucker et al., 2009). They are also called "massively parallel sequencing", including the Roche/454 FLX (Margulies et al., 2005), the Illumina/Solexa Genome Analyzer (Bentley 2006), the Applied Biosystems SOLiDTM System (Bentley 2006), and the Helicos HeliScope Single Molecule Sequencer technology (Harris et al., 2008). Although these platforms are quite different in sequencing biochemistry and in the approach to generate the array, the overall work flows of NGS sequencing are conceptually similar (Figure 1b). In high-throughput NGS sequencing production pipelines, genomic DNA to be sequenced is first randomly fragmented, then followed by *in vitro* ligation of common adaptor sequences. These sequences containing fragmented genomic DNA and adaptors are then subjected to one of several protocols, e.g., *in situ* polonies (Mitra and Church 1999), emulsion PCR (Dressman et al., 2003), or bridge PCR

(Adessi et al., 2000; Fedurco et al., 2006) (Figure 2), resulting in an array of millions of spatially immobilized PCR colonies or polonies (Mitra and Church 1999); each composts of many copies of a single shotgun library fragment. These clonal amplification protocols of sequencing features share some commonnality. For any given single library molecule, the corresponding PCR amplicons end up spatially clustered, either to a single location on a planar substrate (*in situ* polonies, bridge PCR), or to the surface of micron-scale beads, which can be recovered and arrayed (emulsion PCR). After that, a single microliter-scale reagent volume is applied to manipulate all array features in parallel. For each array feature, iterations of enzymatic interrogation and imaging are used to build up a contiguous sequencing read (Shendure and Ji 2008).

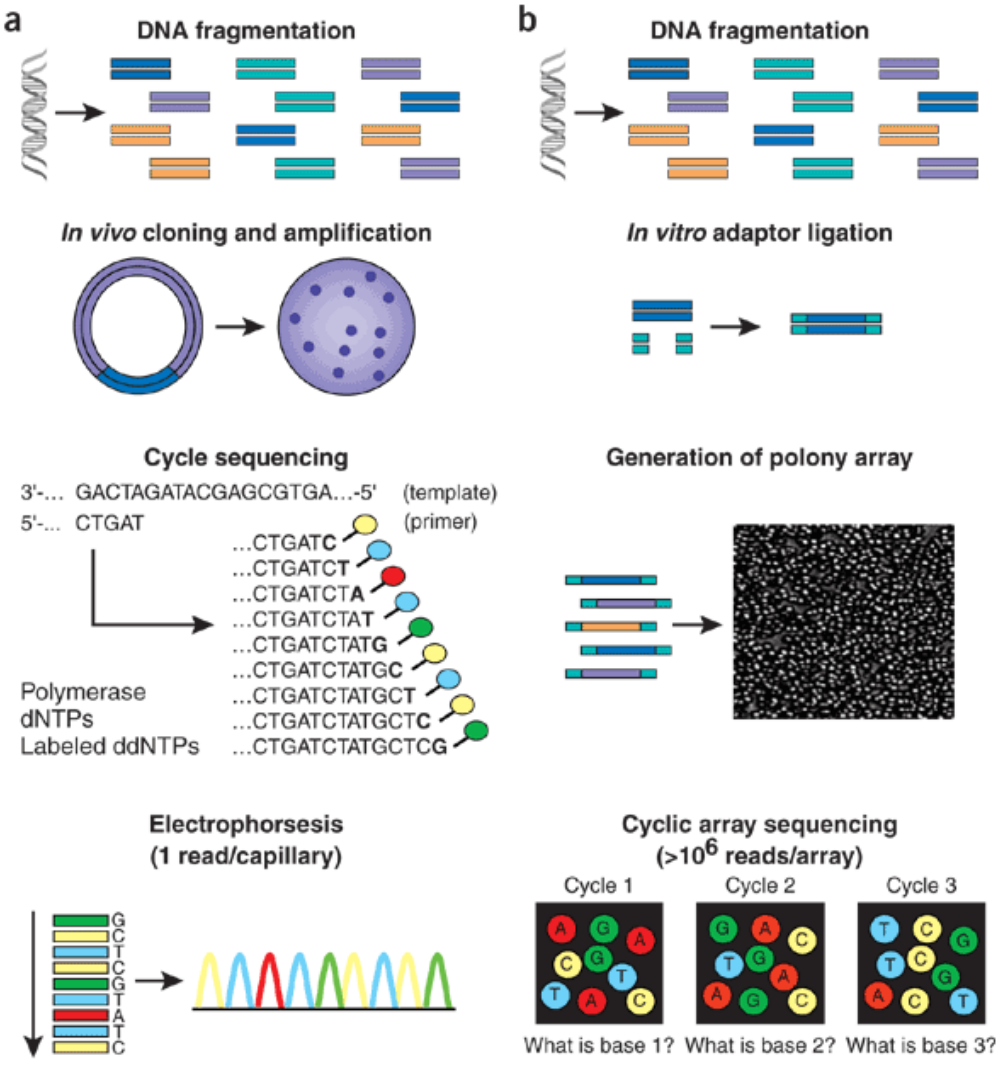


**Figure 1.** Work flow of conventional Sanger versus next generation sequencing. (a) High-throughput shotgun Sanger sequencing production pipelines. (b) high-throughput NGS sequencing production pipelines (Shendure and Ji 2008).

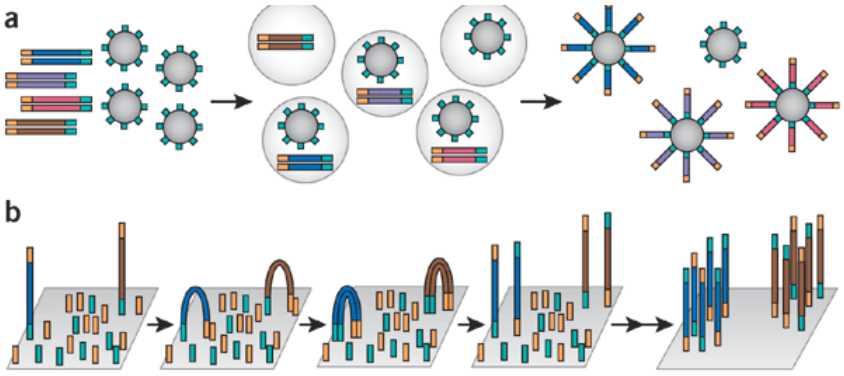


**Figure 2.** Clonal amplification of sequencing features employed in NGS technologies. (a) Emulsion PCR employed in the 454 and the SOLid platforms. (b) Bridge PCR employed in the Solexa platform (Shendure and Ji 2008).

Compared with Sanger sequencing technology, massively parallel sequencers have several advantages: (i) massively parallel sequencers enable to process millions of sequence reads in parallel rather than 96 reads at a time, (ii) next generation sequencers produce reads from fragment 'libraries' that do not require the conventional vector-based cloning and *Escherichia coli*-based amplification stages, and (iii) relatively little input DNA (a few micrograms at most) is needed to produce next-generation sequence-ready libraries and the process is straightforward (Mardis 2008). In other words, NGS technologies require shorter time but produce higher throughput (higher capacity production DNA sequencing) with lower cost. The main reason is that thousands or millions of sequencing reactions can be performed since cloning or template amplification of the DNA fragments are fully automated within the same instrument (Tucker et al., 2009).

## 1.2 Impact of Next Generation DNA Sequencing

NGS technologies have become commercially available since 2004 and have been widely employed in a large number of studies, especially in genetic and clinical studies, to explore and answer genome-wide biological and medical questions. The impact of massively parallel sequencing is substantial, especially for clinical applications since the low-scale, targeted gene/mutation analysis are ultimately replaced by large-scale sequencing of entire disease gene pathways and networks, particularly for the so-called complex disorders (Mardis 2008; ten Bosch 2008). For example, massively parallel sequencing can be used for simultaneous screening to detect mutations in hundreds of loci, for whole-genome screening for novel mutations, and for sequence-based detecting for novel pathogens. Furthermore, if depth of sequence coverage is sufficient, massively parallel sequencing has capability to identify minor alleles, which is a crucial advantage for detecting most tumors, cancer, and complex traits (Tucker et al., 2009).

Importantly, due to the substantial power of NGS technologies enabling to produce hundreds of megabases of data volumes at an affordable cost without requiring a process of cloning or PCR amplification, it facilitates the exploration of the genetic content of entire communities of microbial organisms (Huson et al., 2009; Gilbert et al., 2011; Thomas et al., 2012). This results in a new era of genomics and ecology study, called metagenomics, the study of a collective microbial genetic content recovered directly from natural (e.g., soil, ocean, and freshwater) or host-associated (e.g., human gut and oral) environmental samples on the basis of sequencing data analysis (Riesenfeld

et al., 2004; Gilbert et al., 2011). This advancement of NGS technologies provides a powerful way in metagenomic studies since NGS technologies can be directly applied to an environmental sample without a need of isolating, culturing, and amplifying individual microbial species in a laboratory. Critically, more than 99% of millions microbial species on Earth cannot be cultured in a laboratory (Huson et al., 2007; Rosen et al., 2009). It was also reported that the total number of microbial cells on Earth is estimated to be $10^{30}$, in which prokaryotes represent the largest proportion of individual organisms, which is $10^{6}$ to $10^{8}$. Unfortunately, the genomes of these are mainly uncultured species (Simon and Daniel 2011; Sleator et al., 2008; Turnbaugh et al., 2008). Employing NGS technologies in metagenomics studies not only results in rapidly increase in sequencing the collective genetic content of microbial organisms, but also drives metagenomics go beyond traditional genomics and microbiology, which depend on a process of isolating and culturing individual microbial species in a laboratory. Analyses of the sequencing data about the mixture of micobes have opened a door into the substantial taxonomic and functional knowledge of environmental microbial communities (Simon and Daniel 2011). The field of metagenomics has been responsible for substantial advances in microbial ecology over the past 5 to 10 years to understand the diversity, functions, cooperation and evolution of microbes living in varied natural or host-associated environmental samples (Hugenholtz 2002; Huson et al., 2009, Kunin et al., 2008; Thomas et al., 2012; Wooley and Ye 2010).

### 1.3 Metagenomic Analysis

In principle, there are two main approaches in metagenomic analysis of environmental microbial communities based on nucleic acids: sequence-based and function-based analysis of the collective microbial genomes contained in an environmental sample (add reference). Metagenomics involves constructing a DNA library from an environmental microbial population and then analyzing the sequences (sequence-based analysis) and functions (function-based analysis) in the library (Figure 3) (Daniel 2005; Ferrer et al., 2009; Handelsman 2004; Riesenfeld and Schloss 2004; Simon and Daniel 2011). In this dissertation, an overview of the processes involved in a typical sequence-based metagenome study is addressed (Figure 4).

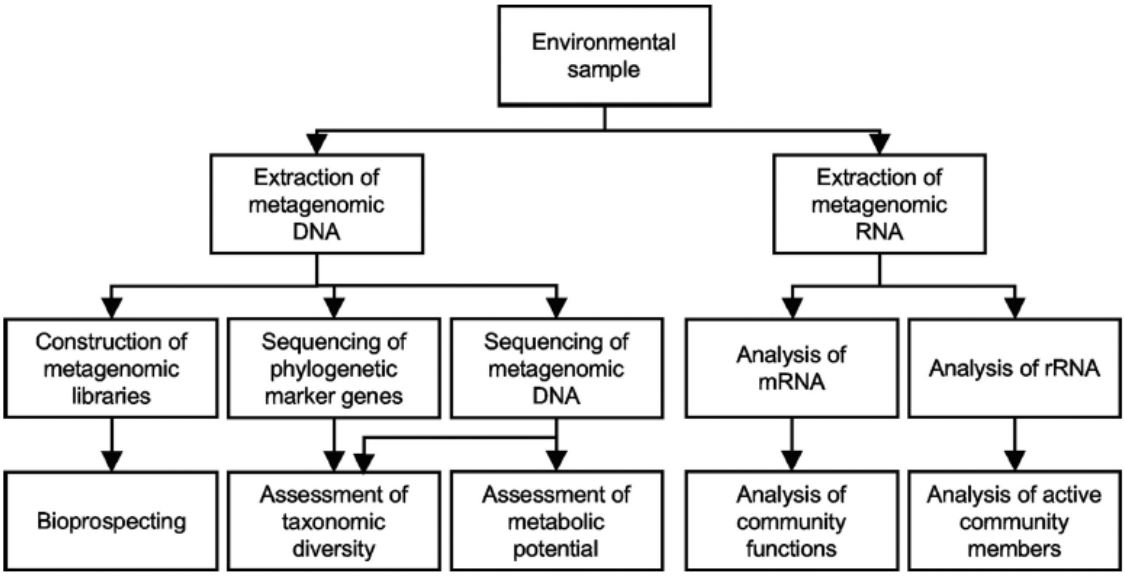


**Figure 3.** Metagenomic analysis of environmental microbial communities based on nucleic acids (Simon and Daniel 2011).

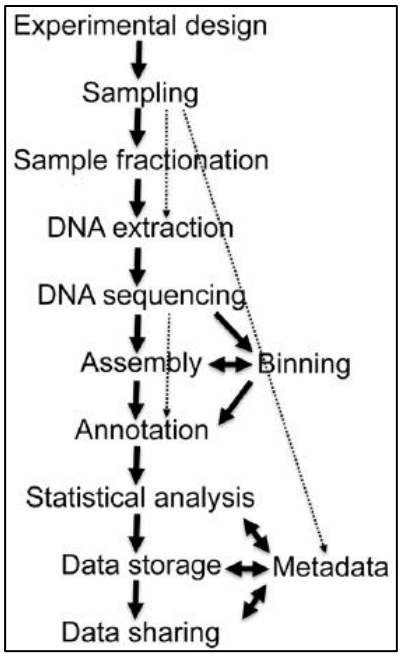


**Figure 4.** Flow diagram of a typical sequence-based metagenome study. Dashed arrows indicate steps that can be omitted (Thomas et al., 2012).

In metagenomic analysis, the sampling process is the first step, which is the most crucial step in any metagenomics project, and if the target community is associated with a host (e.g., human, invertebrate, or plant), DNA fractionation processing (e.g., Burke et al., 2009; Thomas et al., 2010) is required to ensure that minimal host DNA is obtained. DNA is then extracted from the sample in the DNA extraction processing which requires specific protocols (e.g., Burke et al., 2009; Delmont et al., 2011; Venter et al., 2004) for each sample type to certain that (i) DNA extracted represents all cells present in the sample and (ii) sufficient amount of high-quality nucleic acids is obtained for subsequent library production and sequencing (Thomas et al., 2012).

For the DNA sequencing process, although metagenomic shotgun sequencing has shifted from classical Sanger sequencing technology to NGS technologies over the past

10 years, Sanger sequencing is still considered the typical standard for DNA sequencing due to its low error rate and long read length (> 700 bp). However, the overall cost per gigabase (approximate USD 400,000) is still an obvious bottleneck of Sanger sequencing. Of the NGS technologies (detailed in Section 1.1), both the 454/Roche and the Illumina/Solexa systems have now been extensively employed in metagenomic studies (Thomas et al., 2012). In addition, Applied Biosystems SOLiD sequencer has become extensively used in genome resequencing (Gulig et al., 2010).

After obtaining DNA sequencing fragments, usually the assembly process is required to assembly short read fragments to obtain longer genomic contigs, especially when a study focuses on recovering the genome of uncultured organisms or obtain full-length coding sequences for subsequent characterization rather than a functional description of the community. In the assembly process for metagenomics samples, two types of algorithms are employed, including reference-based assembly (co-assembly) and *de novo* assembly. Reference-based assembly algorithms perform well when the sequences in the reference genomes are closely related to sequences in the metagenomic dataset and this type of algorithms require much smaller computational resources compared with *de novo* assembly algorithms. Several software packages have been implemented for reference-based assembly, such as Newbler (Roche), AMOS (http://sourceforge. net/projects/amos/), and MIRA (Chevreux et al., 1999); and these software tools are enabled to perform on laptop-sized machines in a few hours (Thomas et al., 2012).

The binning, the process of sorting DNA sequences into groups representing either an individual genome or genomes from closely related organisms, is then performed. In the binning process, two types of approaches are employed, including compositional-based and similarity-based binning algorithms. The compositional-based binning algorithm uses the fact that genomes have conserved nucleotide composition (e.g., a certain GC or the particular abundance distribution of k-mers) while the similarity-based binning algorithm uses the information of the similarity between an unknown DNA fragment for a gene and known genes in a reference database. Several software and web-based tools have been developed for binning DNA sequences. For instance, Phylopythia (McHardy et al.), S-GSOM (Chan et al., 2008), PCAHIER (Zheng and Wu 2010), and TACAO (Diaz et al., 2009) have been developed based on the compositional-based binning algorithm; the similarity-based binning algorithms include IMG/M (Markowitz et al., 2008), MG-RAST(Glass et al., 2010), MEGAN (Huson et al., 2011; Huson, et al., 2007), CARMA (Krause et al., 2008), SOrt-ITEMS (Monzoorul et al., 2009), and MetaPhyler (Leung et al., 2011). There are also binning algorithms that consider both the compositional-based and similarity-based binning algorithms, including PhymmBL (Brady and Salzberg 2009) and MetaCluster (Leung et al., 2011). Importantly, selecting a binning algorithm depends on the type of available input data and the existence of a suitable training datasets or reference genomes. In general, for short reads (e.g., a 100 base pair (bp) read) the similarity-based binning algorithm is more reliable than the compositional-based binning algorithm because short reads do not contain

enough information for compositional assignment, but they may contain some similarity with a known gene(s) in a reference database (Thomas et al., 2012).

In general, annotation of metagenomic sequence data has two steps: feature prediction and feature annotation. In the prediction step, features of interest are identified by using labeling sequences as genes or genomic elements. Several tools have been specifically developed for feature prediction, including FragGeneScan (Rho et al., 2010), MetaGeneMark (McHardy et al., 2007), MetaGeneAnnotator (MGA)/ Metagene (Noguchi et al., 2008) and Orphelia (Hoff et al., 2009; Yok and Rosen 2011). In addition, a number of tools have also been designed for the prediction of non-protein coding genes such as tRNAs (Gardner et al., 2009; Lowe and Eddy 1997), signal peptides (Bendtsen et al., 2004) or CRISPRs (Bland et al., 2007; Grissa et al., 2007). In the feature annotation step, gene functions and taxonomic neighbors are assigned by mapping to a reference database of gene or protein libraries with existing knowledge (i.e., a non-redundant database) based on homology search tools (e.g., BLASTX, BLAT). A number of reference databases are available for functional annotation, including KEGG (Kanehisa et al., 2004), eggNOG (Muller et al., 2010), COG/KOG (Tatusov et al., 2003), PFAM (Finn et al., 2010), and TIGRFAM (Selengut et al., 2007). However, none of reference database can provide all biological functions. The outputs from homology search tools are formatted in the form of abundance profiles for specific taxa or functional annotations. These outputs support the taxonomic comparison of NCBI taxonomies derived from 16S rRNA gene or whole genome shotgun data and the

functional comparison of relative abundance for COG, KEGG, and SEED (Overbeek et al., 2005) classifications on multiple levels of functions (Thomas et al., 2012).

For the statistical analysis in metagenomics studies, researchers have focused on three typical questions: (i) who is out there?, (ii) what are they doing?, and (iii) whether and how two or more microbial communities differ? (Huson et al., 2007; Rosen et al., 2009; White et al., 2009). For the first question, "who is out there?", a number of methods have been proposed for metagenomic analysis on the basis of taxonomic hierarchy (e.g., class, family, genus, and species) to determine taxonomic compositions in a particular metagenomic sample and to determine the relative species proportions (e.g., Clemente, et al., 2011, Gerlach and Stoye, 2011, Gori, et al., 2011, Huson, et al., 2007, Jiang, et al., 2012, Meinicke, et al., 2011, and Rosen et al., 2009). To answer the second question, scientists detect the functional composition in a particular metagenomic sample and estimate the relative functional abundances presenting in the metagenomic sample. A number of methods have been recently developed for metagenomic analysis on the basis of functional features, particularly at the higher level of functions of a hierarchical functional tree, such as SEED subsystem tree (Dinsdale et al., 2008, Mitra et al., 2011; Parks and Beiko 2010, Sharon et al., 2011, and Yooseph, et al., 2008). For answering the third question, several software tools have been developed to compare different microbial communities based on sequencing data, e.g., UniFrac (Lozupone and Knight 2005), SONs (Schloss and Handelsman 2006), XIPE-TOTEC (Rodriguez-Brito et al., 2006), Metastats (White et al., 2009), MEGAN (Huson et al., 2011; Huson et al., 2007), and

metagenomeSeq (Paulson et al., 2013). In additional to comparing samples with different conditions based on DNA sequencing data, in RNA-Seq studies there is also a goal to compare samples based on RNA sequencing data to recall the genes whose expression levels are different across different conditions. Several statistical tools have been recently developed for this task such as edgeR (Robinson et al., 2010) and DESeq (Anders and Huber 2010), both of which are currently and widely used in RNA-Seq projects.

For data storage and sharing, there are several currently and widely used services, such as IMG/M, CAMERA and MG-RAST. For metadata, a suite of standard languages is required and currently provided by the Minimum Information about any (x) Sequence checklists (MIxS) (Yilmaz et al., 2011). MIxS is an umbrella term to describe MIGS (the Minimum Information about a Genome Sequence), MIMS (the Minimum Information about a Metagenome Sequence) and MIMARKS (Minimum Information about a MARKer Sequence) (Field et al., 2011) and contains standard formats for recording environmental and experimental data (Thomas et al., 2012).

### 1.4 Problem Statement

As mentioned in the Section 1.3, in the statistical analysis of metagenomic projects, researchers have focused on three typical questions: (i) who is out there?, (ii) what are they doing?, and (iii) whether and how two or more microbial communities differ? (Huson et al., 2007; Rosen et al., 2009; White et al., 2009). Currently, there has been increasing interest in metagenomic projects on clinical applications (e.g., Human

Microbiome Project (Turnbaugh et al., 2007)). This research focuses on answering the second and third question, but comparing multiple microbial communities from an angle of functional analysis. Metagenomic analysis on the basis of functional features (e.g., pathways, subsystems, and functional roles) has impact in several areas, such as biological, environmental, and clinical studies. In biological and environmental studies, functional metagenomic analysis provides information not only about the functional components, but also the activities of microbes which describe the role that microbes play in the process of life (Hunter 2004). In clinical studies, comparing different microbial communities corresponding to two or multiple populations with different clinical phenotypes (e.g., diseased and healthy, or different treatments), researchers can determine the activities of microbes relating to the disease and understand the reactions of microbes that respond to different biochemical products. This leads researchers to develop a drug or treatment that specifically affect to either a particular activity or a group of activities of microbes which relate to the disease of interest.

In functional analysis of metagenomes to answer the second question, it relies on the functional information obtained from the prior feature annotation step in the metagenomic analysis pipeline. In the feature annotation step, the outputs from homology search tools (e.g. BLASTX) are formatted in the form of abundance profiles for specific taxa or functional annotations. These outputs support the functional comparison of relative abundance for COG, KEGG, and SEED (Overbeek et al., 2005) classifications on multiple levels of functions (Thomas et al., 2012). The SEED classification is employed

by several well-known and widely used systems, such as MG-RAST and IMG/M, for metagenomic analysis. According to Overbeek et al. (2005), in the SEED gene functions are classified into functional roles and subsystems. A "functional role" is described as a single logical role that a gene or gene product may play in the operation of a cell. A relative set of functional roles that implement a specific biological process or structural complex is described as "subsystem", which might be thought of as generalization of the term called "pathway" in KEGG. The SEED classification is represented hierarchically as a tree with about 13,000 nodes, where the internal nodes represent different subsystems and the leaves represent the functional roles as illustrated in Figure 1 in Appendix A. In functional analysis of metagenomes based on KEGG and SEED classification, therefore, researchers can perform analysis either on functional role level or pathways/subsystems level. To detect the functional composition in a particular metagenomic sample and estimate the relative functional abundances contributed in the metagenomic sample, a number of recently published studies have been developed to achieve this tasks; however, most of them focused on functions at the higher level, including pathway and subsystems. There is a limitation of statistical methods specifically designed for metagenomic analysis at the level of functional role, i.e., low level of the SEED tree. This results in a number of ambiguous questions about functional roles of microbial communities, e.g. do microbial communities consist of extensive genetic diversity, how are they diverse in functional roles, how does the diversity in functional roles of microbial communities affects the interaction with environment they live? Therefore, a specific development of

statistical methods for metagenomic analysis at the level of functional role is substantially needed.

To answer the third question, a number of software tools, e.g. UniFrac (Lozupone and Knight 2005), SONs (Schloss and Handelsman 2006), XIPE-TOTEC (Rodriguez-Brito et al., 2006), Metastats (White et al., 2009), MEGAN (Huson et al., 2011; Huson et al., 2007), and metagenomeSeq (Paulson et al., 2013) have been developed to compare different microbial communities based on sequencing data in metagenomic projects. However, there are still several limitations for those software tools. The most obvious limitation of these tools is that most of them are designed to compare exactly two microbial communities, and only metagenomeSeq can be applied to compare more than two metagenomes. Another limitation is that they are mainly designed for comparison of taxonomic components. There is a lack of statistical method developed for functional compositions of different metagenomes. Among these software tools, only XIPE-TOTEC is developed for metagenomic comparison of functional compositions. There is also a limitation in computational time consumption of XIPE-TOTEC and Metastats because they rely on non-parametric statistical method which employs bootstrap and permutation techniques. In additional to comparing samples with different conditions based on DNA sequencing data, in RNA-Seq studies there is also a goal to compare samples based on RNA sequencing data to recall the genes whose expression levels are different across different conditions. Several statistical tools have been recently developed for this task such as edgeR (Robinson et al., 2010) and DESeq (Anders and Huber 2010), both of

which are currently and widely used in RNA-Seq projects. Although these methods have been well developed and widely used in RNA-Seq projects, and DNA sequencing-based microbiome investigations use the same sequencing machines and represent the processed sequence data in the same manner with RNA-Seq experiments, there is still limitation in directly applying those statistical methods developed for RNA-Seq data to metagenomic data due to the data property

In conclusion, several statistical methods have been developed for metagenomic analysis; however, there is still an urgent need of rigorous statistical methods which are particularly developed for metagenomic analysis on the basis of functional features (e.g., pathways, subsystems, functional roles) to analyze massive sequencing data of metagenomes.

### 1.5 Research Objectives

The main objective of this study is to specifically develop statistical methods for functional metagenomic analysis to analyze metagenomic data obtained from varied natural and host-associated samples. The specific objectives of the study are:

1. To identify all possible functional features (so-called “functional roles”) present in a metagenomic sample/community (Appendix A).
2. To assess an association between differentially abundant functional features (i.e., pathways, subsystems, functional roles, etc.) and their relative phenotypes in different metagenomics datasets (Appendix B).

# 2. LITERATURE REVIEW

This chapter focuses on literature review of the exiting methods proposed for metagenomic analysis on functional identification and for metagenomic comparison of different environmental samples, including the algorithms and the limitations of each method.

## 2.1 Functional Identification

In identifying the functions in a metagenomic sample, researchers aim to identify functional contents presenting in a particular metagenomic sample and also to estimate their relative functional abundances contributed in the metagenomic sample. An output from the feature annotation step usually is formatted in the form of abundance profiles for functional annotations based on KEGG, or SEED classification hierarchy on multiple levels of functions (detailed in Section 1.3) (Thomas et al.,2012). The output of this step is used for further computational and statistical analysis in the downstream of metagenomic analysis. In functional identification of metagenomes, a number of methods have been developed. However, most of them focused on functional identification at the pathway and subsystems level (e.g. Dinsdale et al. 2008, Parks and Beiko 2010, Sharon et al. 2011, Yooseph et al. 2008). There are only a few software tools and systems, such as MG-RAST (Glass et al., 2010) and MEGAN (Huson et al., 2011; Huson, et al., 2007) can perform metagenomic analysis on functional identification of metagenomes at the level

of functional role. Both MG-RAST and MEGAN are well-known and widely used for metagenomic analysis.

MG-RAST is a web-based analytical system that provides fully automated pipeline for quality control, feature prediction, functional annotation, and genomic comparisons. MG-RAST enables us to analyze unassembled reads and also short reads, such as Illumina/Solexa reads of 75 bp, and it requires minimum read length of only 75 bp for gene prediction, similarity analysis, taxonomic binning, and functional classification. For feature identification, MG-RAST uses a two-step approach, FragGeneScan (Rho et al., 2010) and a similarity search for ribosomal RNAs against a non-redundant integration of the SILVA (Pruesse et al., 2007), Greengenes (DeSantis et al., 2006), and RDP (Cole et al., 2009) databases. FragGeneScan is a gene prediction method developed for prediction of protein-coding region in short reads by using sequencing error models and codon usages in a hidden Markov model to improve the prediction (Rho et al., 2010). The output results from feature identification of MG-RAST are expressed in the form of abundance profiles for specific taxa or functional annotations. These output results can be used for comparison of NCBI taxonomies derived from 16S rRNA gene or whole genome shotgun data. Moreover, these output results also support the comparison of relative abundance for KEGG, eggNOG, COG and SEED subsystems on multiple levels of resolution. Another obvious advantage of MG-RAST is that it provides several statistical techniques for comparisons of metagenomes. In addition to provide statistical pipeline for metagenomic analysis, MG-RAST also

provides a large-scale database for storing statistical results and metagenomic datasets (Thomas et al., 2012). So far, MG-RAST has more than 12,000 users, and more than 115,000 metagenomes uploaded and analyzed. Out of the total number of metagenomes, about 16,900 are publicly accessible and 45 Terabases analyzed as of April 2014. Although MG-RAST provides fully automated pipeline for metagenomic analysis, there is a trade-off between accuracy and computational efficiency for short reads using BLAT (Kent 2002) as a homology search tool for functional annotation (Thomas et al., 2012).

MEGAN is a standalone computer software tool used to analyze metagenomic data and to visualize annotation results derived from BLAST searches in a functional or taxonomic dendrogram. One main advantage of MEGAN is the use of dendrograms to display metagenomic data. That a user can collapse network of interpretation at a desired level makes analysis and interpretation of particular functional or taxonomic groups quick and easy. In MEGAN the functional analysis of metagenomes is based on the SEED classification hierarchy. To perform a functional analysis, MEGAN assigns each read to the functional role of the highest scoring gene in BLAST output against a protein database (e.g., NCBI-NR), and then different functional roles are grouped into SEED subsystems as illustrated in Figure 1 in Appendix A. However, there are several limitations in using MEGAN for metagenomic analysis on functional identification. First of all, the best score assignment could miss many putative functions. Because of the existence of sequencing error (Hoff 2009), for the same sequencing read, it could have a function with identical matches of 32 out of 33 codons and also have a function with

match score of 31 out of 33 codons. The MEGAN method will miss the second or even third best scoring functions that the read may have. Even more, MEGAN just assigns one of the best functions (with the same largest match values) to the short read. However, we we know that a gene could play multiple functions at the same time. In other words, MEGAN underestimates the abundance of functional roles.

### 2.2 Metagenomic Comparison

To perform metagenomic comparison, researchers can conduct an experiment to compare genomic features based on either taxonomic compositions or functional components obtained from different metagenomic communities. Several software tools, including XIPE-TOTEC (Rodriguez-Brito et al., 2006), Metastats (White et al., 2009), and MEGAN (Huson et al., 2011; Huson et al., 2007), and metagenomeSeq (Paulson et al., 2013) have been developed to compare different microbial communities based on sequencing data. They have limitations. First of all, most of them are designed to compare exactly two microbial communities and only metagenomeSeq can be applied to compare more than two metagenomes. metagenomeSeq is developed for analysis of sparse high-throughput microbial marker-gene studies. In the marker-gene survey data, most operational taxonomic units (OUTs) are rare and absent. This characteristic is just one specific type of metagenomic data. In general, a metagenomic data also contains OTUs with mixture of relative abundances. Secondly, most of them are mainly designed for comparison of taxonomic components, rather than functional compositions of

different metagenomes and only XIPE-TOTEC is developed for functional comparison. However, there is a limitation in computational time consumption of XIPE-TOTEC since it relies on non-parametric statistical method which employs bootstrap technique. Similarly to XIPE-TOTEC, there is also a limitation in computational time consumption of Metastats because it employs permutation technique in non-parametric statistical method.

XIPE-TOTEC was the first approach used for biomarker discovery in metagenomic samples by testing the statistical significance of difference on subsystems between two metagenomes. XIPE-TOTEC employs non-parametric statistical method relied on bootstrap resampling to build a null distribution for a given subsystem to estimate the corresponding test statistic. There are several steps in the statistical analysis. In the first step, two samples are first pooled together. Then, two samples of $N$ sequences are randomly drawn from the pooled dataset separately. For each resample, the difference in occurrence for each subsystem between two samples is computed. This process is repeated $L$ times. Then, $L$ differences are sorted and median $M$ for each subsystem is obtained. In the second step, two samples of $N$ sequences are randomly drawn from the pooled dataset separately. For each resample, the difference in occurrence for each subsystem between two samples is computed. This process is repeated $L$ times. Then, $L$ differences are sorted and the $5^{th}$ percentile and $95^{th}$ percentile for each subsystem are obtained. The range between the $5^{th}$ percentile and $95^{th}$ percentile is described as a confidence interval for each subsystem. In the last step, for each subsystem if $M$ lies

outside the confidence interval, this subsystem is defined as a statistically significant different subsystem between these two samples. For each subsystem, the resample process is required to repeat multiple times, usually more than thousands times in order to estimate the null distribution accurately. In their experiment, they randomly sampled 10 x $10^3$ sequences and repeated this process in 20 x $10^3$ times to estimate the null distribution. Therefore, computational time consumption is a main bottleneck of XIPE-TOTEC. In addition to the limitation in time consumption, Parks and Beiko (2010) claimed that the difficulty with this approach is that *N* is a free parameter. Increasing *N* reduces the width of the null distribution which in turn increases the false positive error rate.

Metastats was the first software tool developed to address the questions in clinical studies of microbial samples. Metastats identifies features that statistically distinguish from two treatment populations (e.g., healthy vs. disease) employing non-parametric t-test, which is based on permutation to estimate the null hypothesis distribution. Metastats employs Fisher's exact test to separately handle sparsely-sampled features. Metastats requires a count data which is formatted in the form of feature abundant matrix, where rows represent features and columns represent subjects, for the input to the statistical method. For each feature, an empirical non-parametric *t*-test is calculated to compare its abundance across the two treatment groups. This *t*-test statistic is described as the originally observed *t*-test value. Then, the corresponding p-value is obtained by permuting the samples *B* times, recalculating a *t*-statistic for each feature each time, and taking the proportion of *t*-statistics greater than the originally observed

value. Precision of the p-value calculations is obviously improved by increasing the number of permutations used to approximate the null distribution. It is recommended that the number of permutations should be large. Metatats uses 1000 times as the default number of permutations. Since Metastats relies on permutation method to estimate a *t*-statistic, its computational time consumption is the main limitation, and may lose power in testing as it is a non-parametric approach.

metagenomeSeq, implemented and available as a package of the Bioconductor software development project (don't cide the paper for Bioconductor ), has been recently developed to determine features (e.g., OTU, species, and genus) that are differentially abundant between two or more groups of multiple samples (cite the new paper for metagenomeSeq). metagenomeSeq is designated for analysis of sparse high-throughput microbial marker-gene survey data employing a zero-inflated Gaussian (ZIG) mixture model. It requires count data in the form of feature abundant matrix as the input data. The elements of the matrix are the abundance of features observed in subjects. Employing a zero-inflated Gaussian mixture model is motivated by the depth of coverage in a sample, which is directly related to how many features are detected in a sample. Maximum-likelihood estimates are approximated by an EM algorithm. The metagenomeSeq package also includes useful visualization tools. McMurdie and Holmes (2013) claimed that that metagenomeSeq performs well when there is an adequate number of biological replicates, but nevertheless tends toward a higher false positive error rate.

In microarray and recently RNA-Seq studies, one main objective is to detect the differentially expressed genes across different conditions. In these fields, the features are the abundance of gene(s) or gene expression. Although the statistical methods developed for microarray analysis have matured to a high level of sophistication (Allison et al. 2006), these methods are not directly applicable for analyzing DNA sequencing data. The principal difference is that a microarray data consists of continuous values obtained from the fluorescence intensity of hybridized probes while a DNA sequencing data consists of discrete count values of equivalent sequences. In RNA-Seq data analysis many statistical tools have been recently developed for identifying genes differentially expressed under different conditions. For example, edgeR (Robinson et al., 2010) and DESeq (Anders and Huber 2010) are two widely used methods in RNA-Seq experiments (Anders et al., 2013). Since edgeR and DESeq employ almost the same statistical algorithm of NB model, they have similar strategies to perform differential analysis for count data to identify differentially abundant features between different samples. However, they differ in several ways. The most crucial difference between edgeR and DESeq is the way in estimating the dispersion parameter of NB model. edgeR estimates feature-level dispersion by relying on a trended mean according to the dispersion-mean relationship whereas DESeq estimates the dispersion by taking the maximum of the individual dispersion estimates and the dispersion-mean trend (Anders et al., 2013). However, several literatures have claimed that by comparing the performance of these two methods

via various simulation studies, there is no single method that can dominate the other across all settings (Anders et al., 2013; Nookaew et al., 2012; Soneson et al., 2013).

# 3. PRESENT STUDY

Appendix A and B of this dissertation, both present the manuscripts of the methods, results, and conclusions of both objective one and objective two, respectively. This chapter provides a brief summary of the study.

In Appendix A, the manuscript introduces a mixture model for modelling the translated sequence reads at the codon-level and globally detect the functions within a metagenomic sample with the sequencing error is considered in the statistical model. The proposed method was comprehensively tested on simulated metagenomic data with diverse complexity of microbial community structure, and applied to a real metagenomic dataset. Compared with other available algorithms and tools designated for functional metagenomic analysis, the proposed approach demonstrated greater accuracy in identification and quantitative estimates of relative frequency of contained functionality in a given sample. We also applied the proposed method to two real metagenomic data sets and our findings are consistent with previous reports.

In Appendix B, the manuscript introduces a two-stage statistical procedure for both the selection of informative functional features and identification of differentially abundant functional features (e.g. pathways, subsystems, functional roles) between metagenomic conditions. In the 1$^{st}$ stage of our algorithm, the functional informative features are simultaneously selected using elastic net resulting in dimensional reduction of the metagenomic dataset. In the 2$^{nd}$ stage, we identify the differentially abundant functional features using generalized linear models with negative binomial distribution.

We evaluated the performance of our method through several simulations and applied our proposed method to publicly available metagenomic datasets. According to various simulations, the results showed that our proposed two-stage statistical algorithm can effectively select the informative functional features and efficiently detect the differentially abundant functional features between metagenomic conditions. Compared with previous methods widely used in metagenomic data analysis and even in RNA-seq data analysis, the simulation results show that the proposed approach outperforms the other methods in most situations. We also performed the proposed method on two real metagenomic datasets relating to human diseases. Our findings are consistent with the findings in previous reports. Our method not only compares two microbial populations/conditions, but also can be used for comparison of more than two microbial populations. Therefore, our method can be applicable to more general situations.

In overall summary, we proposed statistical methods which are specifically designed for functional metagenomic analysis to analyze metagenomic data obtained from varied natural and host-associated samples. We expect that the proposed method will help scientists achieve their research objectives based on function detection and function comparison in metagenomic studies.

# 4. REFERENCES


Adessi C et al. (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res. 28, e87.

Allison B, Cui X, Page P, Sabripour M (2006) Microarray Data Analysis: from Disarray to Consolidation and Consensus. Nat Rev Genet 7: 55–65.

Anders S, McCarthy DJ, et al. (2013). "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor." Nat Protoc 8(9): 1765-1786.

Anders S and Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11(10): R106.

Bentley DR. 2006. Whole-genome resequencing. Curr. Opin. Genet. Dev. 16:545–52.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Molec Biol 340(4):783-795.

Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics 8:209.

Burke C, Kjelleberg S, Thomas T (2009) Selective extraction of bacterial DNAfrom the surfaces of macroalgae. Appl Environ Microbiol 75(1):252-256.

Chan CK, Hsu AL, Halgamuge SK, Tang SL (2008) Binning sequences using very sparse labels within a metagenome. BMC Bioinformatics 9:215.

Chevreux B, Wetter T, Suhai S (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information Computer Science and Biology. Proceedings of the German Conference on Bioinformatics 99:45-56.

Clemente,J. et al.(2010) Accurate taxonomic assignment of short pyrosequencing reads. Pac. Symp. Biocomput 15, 3–9.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed- Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37 Database: D141-145.

Daniel R (2005) The metagenomics of soil. Nat. Rev. Microbiol. 3:470–478.

Delmont TO, Robe P, Clark I, Simonet P, Vogel TM (2011) Metagenomic comparison of direct and indirect soil DNA extraction approaches. J Microbiol Methods 86(3):397-400.

Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics 10:56.

Dinsdale,E.A. et al.(2008) Functional metagenomic profiling of nine biomes. Nature 452(3): 629-632.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72(7):5069-5072.

Dressman D, Yan H, Traverso G, Kinzler KW and Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc. Natl. Acad. Sci. USA 100, 8817–8822.

Fedurco M, Romieu A, Williams S, Lawrence I and Turcatti G (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 34, e22.

Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, et al (2011) The Genomic Standards Consortium: Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. PLoS Biol 9(6):e1001088.

Ferrer M, Beloqui A, Timmis KN and Golyshin PN (2009) Metagenomics for mining new genetic resources of microbial communities. J. Mol. Microbiol. Biotechnol. 16:109–123.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. Nucleic Acids Res 38 Database: D211-222.

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A (2009) Rfam: updates to the RNA families database. Nucleic Acids Res 37 Database: D136-140.

Gentleman RC et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.

Gilbert JA, Meyer F, Bailey MJ (2011) The Future of microbial metagenomics (or is ignorance bliss?). ISME J 5(5): 777–779.

Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb Protoc (1), pdb prot5368.

Gori,F. et al.(2011) MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. Bioinformatics, 27: 196-203.

Grissa I, Vergnaud G, Pourcel C: CRISPRFinder (2007) a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35 Web Server: W52-57.

Gulig PA, de Crecy-Lagard V, Wright AC, Walts B, Telonis-Scott M, McIntyre LM (2010) SOLiD sequencing of four Vibrio vulnificus genomes enables comparative genomic analysis and identification of candidate cladespecific virulence genes. BMC Genomics 11:512.

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol. Mol. Biol. Rev. 68:669–685.

McMurdie PJ, Holmes S (2013) Waste not, want not: why rarefying microbiome data is inadmissible. ArXiv:1310.0424v2 [q-bio.QM] 12 Dec 2013.

Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 38(20).

Harris TD et al. (2008) Single-molecule DNA sequencing of a viral genome. Science 320, 106–109.

Hoff KJ, Lingner T, Meinicke P, Tech M (2009) Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res 37 Web Server: W101-105.

Hugenholtz P **(**2002) Exploring prokaryotic diversity in the genomic era. Genome Biol. 3:REVIEWS0003.

Hunkapiller T, Kaiser RJ, Koop BF and Hood L (1991) Large-scale and automated DNA sequence determination. Science 254, 59–67.

Hunter L (2004) Life and its molecules: a brief introduction, AI Magazine, v.25 n.1, p.9-22, Spring 2004.

Huson D, Auch A, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377–386.

Huson D, Mitra S, et al. (2011) Integrative analysis of environmental sequences using MEGAN4. Genome Res 21(9): 1552-1560.

Huson D, Richter D, Mitra S, Auch A, Schuster S (2009) Methods for comparative metagenomics. BMC Bioinformatics 10(Suppl 1):S12.

Jiang, H., et al. (2012) A statistical framework for accurate taxonomic assignment of metagenomic sequencing reads, PLoS ONE. 7(10): e46450.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32 Database: D277-280.

Kent WJ (2002) BLAT-the BLAST-like alignment tool. Genome Res 12(4):656-664.

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J (2008) Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res 36(7):2230-2239.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatics's guide to metagenomics. Microbiol. Mol. Biol. Rev., 72(4):557.

Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M (2011) Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics 12(Suppl 2):S4.

Lowe TM, Eddy SR: tRNAscan-SE (1997) a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25(5):955-964.

Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71: 8228–8235.

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387-402.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–80.

Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36 Database: D534-538.

Maxam AM and Gilbert W (1977) A new method for sequencing DNA. Proc. Natl. Acad. Sci. 74, 560–564.

McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods 4(1):63-72.

Meyer,F. et al.(2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

Mitra RD and Church GM (1999) In situ localized amplification and contact replication of many individual DNA molecules. Nucleic Acids Res. 27, e34.

Mitra, S., *et al.* (2011) Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG, *BMC bioinformatics*, **12**.

Monzoorul Haque M, Ghosh TS, Komanduri D, Mande SS (2009) SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics 25(14):1722-1730.

Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, Bork P (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res 38 Database: D190-195.

Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: detecting speciesspecific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15(6):387-396.

Nookaew I., et al. (2012) A comprehensive comparison of RNA-seq–based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res. 40, 10084–10097.

Overbeek R., et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, Nucleic Acids Res, 33, 5691-5702.

Paulson J, Stine O, Bravo H, Pop M (2013) Differential abundance analysis for microbial marker-gene surveys. Nat Methods 10(12): 1200-1202.

Parks DH and Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26(6): 715-721.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35(21):7188-7196.

Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and errorprone reads. Nucleic Acids Res 38(20):e191.

Riesenfeld CS, Schloss PD, Handelsman J (2004). Metagenomics: genomic analysis of microbial communities. Annu Rev Genet 38: 525-552.

Robinson M, McCarthy D, Smyth Gordon (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1): 139-140.

Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinform. 14, 91.

Rosen, GL, Sokhansanj BA, Polikar R, Bruns MA, Russell J, Garbarine E, Essinger S and Yok N (2009) Signal Processing for Metagenomics: Extracting Information from the Soup. Current Genomics 10(7): 493-510.

Schloss P, Handelsman J (2006) Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. Appl Environ Microbiol 72: 6773–6779.

Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, Richter AR, White O (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res 35 Database: D260-264.

Sanger F, Nicklen S and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 74, 5463–5467.

Sharon, I., et al. (2011) Pathway-based functional analysis of metagenomes, Journal of computational biology : a journal of computational molecular cell biology, **18**, 495-505.

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature Biotechnology 26, 1135 – 1145.

Sleator RD, Shortall C and Hill H. (2008) Metagenomics. Lett. Appl. Microbiol. 47:361–366.

Swerdlow H, Wu SL, Harke H and Dovichi NJ (1990) Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. J. Chromatogr. 516, 61–67.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

ten Bosch JR and Grody WW (2008) Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. J Mol Diagn 10(6): 484-492.

Thomas T, Gilbert J, Meyer F. (2012) Metagenomics – a guide from sampling to data analysis. Microb Inform Exp. 2(1):3.

Thomas T, Rusch D, DeMaere MZ, Yung PY, Lewis M, Halpern A, Heidelberg KB, Egan S, Steinberg PD, Kjelleberg S (2010) Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. ISME J 4(12):1557-1567.

Tucker T, Marra M, and Friedman JM (2009) Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. Am J Hum Genet 85(2): 142–154.

Turnbaugh PJ, Ley R, Hamady M, Fraser-Liggett C, Knight R, et al. (2007) The human microbiome project. Nature 449: 804–810.

Turnbaugh PJ, and Gordon JI (2008) An invitation to the marriage of metagenomics and metabolomics. Cell 134:708–713.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304(5667):66-74.

White J, Nagarajan N, Pop M (2009) Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. PLoS Comput Biol 5(4): e1000352.

Wooley J, Ye Y (2010) Metagenomics: Facts and artifacts, and computational challenges. J of Comp Sci and Tech. 25(1): 71-81.

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 29(5):415-420.

Yok NG, Rosen GL (2011) Combining gene prediction methods to improve metagenomic gene annotation. BMC Bioinformatics 12:20.

Yooseph, S., Li, W.Z. and Sutton, G. (2008) Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering, BMC bioinformatics, **9**.

Zheng H, Wu H (2010) Short prokaryotic DNA fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis. J Bioinform Comput Biol 8(6):995-1011.

# APPENDIX A – STATISTICAL APPROACH OF FUNCTIONAL PROFILING FOR A MICROBIAL COMMUNITY

**Lingling An, Naruekamol Pookhao, Hongmei Jiang, Jiannong Xu**
Submitted to: *PLOS ONE*

**Abstract**

**Backgroud:** Metagenomics is a relatively new but fast growing field within environmental biology and medical science. It enables researchers to understand the diversity of microbes, their functions, cooperation, and evolution in a particular ecosystem. Traditional methods in genomics and microbiology are not efficient in capturing the structure of the microbial community in an environment. Nowadays, high-throughput next generation sequencing technologies are powerfully driving the metagenomic studies. However, there is an urgent need to develop efficient statistical methods to rapidly analyze the massive short sequencing data generated from microbial communities and to accurately detect the features/functions present in a metagenomic sample/community. Although several issues about functions of metagenomes at pathways or subsystems level have been investigated, there is lack of studies focusing on functional analysis at the low level of a hierarchical functional tree, such as SEED subsystem tree.

**Results:** A two-step statistical procedure (metaFunction) is proposed to detect all possible functional roles that are at the low level and present in a metagenomic sample/community. In the first step a statistical mixture model is proposed at the codon level of the genes to globally assign short reads to the candidate functional roles based on the SEED classification, with sequencing error being considered. As a gene could be involved in multiple biological processes the functional assignment is adjusted by utilizing an error distribution in the second step. The performance of the proposed procedure is evaluated through comprehensive simulation studies. Compared with other existing methods in metagenomic functional analysis the new approach is more accurate in assigning reads to functional roles, and therefore at other general levels. The method is also employed to analyze two real data sets.

**Conclusions:** The proposed method provides a new approach on functional metagenomics.

**Keywords**

Functional metagenomics, Next generation sequencing, Codon, SEED classification.

## INTRODUCTION

Metagenomics is the study of genetic material recovered directly from natural (e.g., soil and seawater) or host-associated (e.g., human gut) environmental samples that

contain microorganisms organized into communities or microbiomes. The advancement of high-throughput next generation sequencing technologies provides a powerful way in metagenomic studies since they can be directly applied to an environmental sample without the need of isolating and culturing individual microbial species in a laboratory and more than 99% of millions microbial species on Earth cannot be cultured in a laboratory [1,2]The massively parallel sequencing technologies, such as 454FLX, Illumina Genome Analyzer (GA), and ABI SOLiD, have enabled to generate millions of reads (35-250 base pairs (bp), depending on the platform) at a time [3]

The initial computational analysis of metagenomics focuses on two main questions: (1) who is out there (2) what are they doing [1,2]. To answer the first question, scientists determine taxonomic compositions in a particular metagenomic sample and determine the relative species proportions. Many methods have been proposed [1,4-8] in particular, a new method proposed by Jiang et al. [7] and at the very low level - species.

To answer the question "what are they doing?" scientists need to determine the gene contents and functional categories and estimate the relative functional abundances contributed in the metagenomic sample. According to Overbeek et al. [9], a functional role corresponds roughly to a single logical role that a gene or gene product may play in the operation of a cell, such as 'Aspartokinase (EC 2.7.2.4)', and pathways or subsystems are a collection of related functional roles, and a functional role can be involved in multiple pathways or subsystems (Figure 1). To characterize the functional capacity of a metagenomic community, therefore, researchers can perform analysis either at functional

role level or pathways/subsystems level (Figure 1). Most recently published studies focused on pathways or subsystems level [10-13]. However, a number of questions about functional roles of microbial communities are still ambiguous, e.g., do microbial communities consist of extensive genetic diversity, how are they diverse in functional roles, how does the diversity in functional roles of microbial communities affect their interaction with environment? Performing function analysis of metagenomes at functional roles level, therefore, is a possible approach to address these questions. Through such type of functional analysis of metagenomes, functional roles can be detected and further metabolic pathways or subsystems that the functional roles are involved can be established [12].

Only a few existing methods can deal with the functional role. The representative method is the MEGAN computer program [14]. In MEGAN the functional analysis of metagenomes is based on the SEED hierarchy [15]. The SEED has the most consistent and accurate microbial genome annotations of any publicly available source [9]. To perform a functional analysis, MEGAN assigns each read to the functional role of the highest scoring gene in a BLAST comparison against a protein database (e.g., NCBI-NR), and then different functional roles are grouped into SEED subsystems. The SEED classification can be represented by a rooted tree, where the internal nodes represent subsystems and the leaves represent the functional roles (Figure 1).

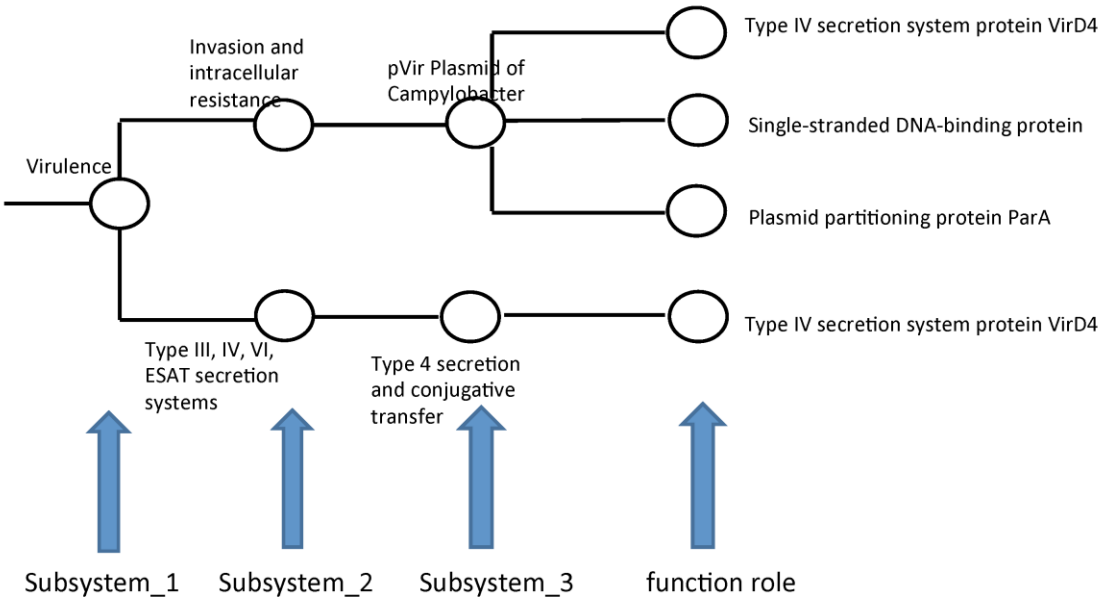


**Figure 1.** Illustration of subsystem tree structure in SEED.

However the MEGAN program has several disadvantages. First of all, the best score assignment might miss putative functions. Because of the existence of sequencing error [16], a sequence read could come from a gene/function with aligned matches of 32 out of 33 codons and could also from a gene/function with aligned match of 31 out of 33 codons. The MEGAN method misses the second or even the third best scoring functions that the read may have. Furthermore, a gene could play multiple functions at the same time. However MEGAN just assigns one function (with the best match value) to the short read even when multiple functions show the same best match values (e.g., the e-value, bitscore, or the number of matched codons). For example, blastx output for a short read shows two functions "*Argininosuccinate lyase (EC 4.3.2.1)*" and "*N-acetylglutamate*

*synthase (EC 2.3.1.1)*" with the same best match values, but MEGAN only assigns the first function (alphabetically) to the read. Thus, MEGAN misses some functions existing in the community and therefore underestimates their abundance.

MG-RAST [15] can assign multiple functions to a read, but just based on some flat cutoffs, e.g, e-value < 1.0e-5 and identity cutoff >60%. Thus assignment of reads to different ranks of taxonomy tree depends on what threshold for bit-score or Expect value is used. Furthermore, it assigns reads one at a time. As a consequence, the results lack specificity.

Motivated by both the advantages and limitations of these methods and inspired by the statistical model in Jiang et al. [7] we propose a mixture model to model the translated sequence reads at the codon-level and globally detect the functions in a metagenomic sample with the sequencing error being considered. As a gene could be involved in multiple biological processes the functional assignment is adjusted by utilizing an error distribution. The proposed two-step method is comprehensively tested on simulated metagenomic data with diverse complexity of microbial community structure, and applied to two real metagenomic datasets. Compared with other available algorithms and tools designated for functional metagenomic analysis, the proposed approach demonstrates greater accuracy in identification and quantitative estimates of relative frequency of contained functionality in a given sample. The R package "metaFunction" is available for download at

http://cals.arizona.edu/~anling/software/metaFunction.htm.

## METHODS

For each sequence dataset we use BLASTX (version 2.2.26+) to search for matched reference sequences (i.e., genes) in the NCBI-NR protein database (downloaded April 2012). Then genes are classified into functional role categories as defined in the SEED database. Based on the sequence reads we need to estimate: (1) the sequencing error (estimated from data) and (2) functional roles contained in the metagenomic sample and their relative abundance. To answer these questions, we set up a mixture model based on the information from BLASTX results to estimate the probability that a given read belongs to a functional role. The flowchart for the whole procedure can be found in Figure 2.

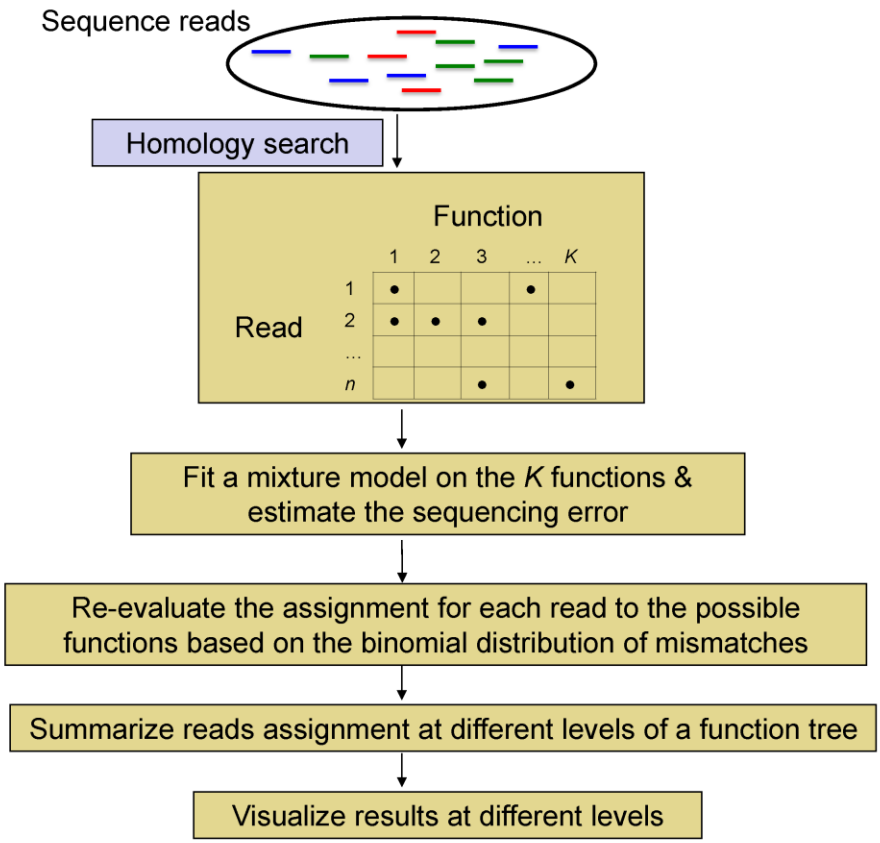


**Figure 2.** Flowchart of the proposed method -metaFunction.

**Estimate sequencing error**

Suppose we have totally *n* sequence reads, $r_1,...,r_n$, which have found sequence homologs in the reference database and these matched genes have totally *K* functions (i.e., gene families). Let $C_{ji}$ denote the number of identical matched codons of read $r_j$ for functional role *i* and $L_j$ denote the maximum aligned codon length for read $r_j$ (i.e,

$L_j = \max_i(L_{ji}))$ across all candidate functions, we have $C_{ji} \leq L_j$. If a read $r_j$ does not have matched sequences in the function *i*, then $C_{ji} = 0$. We assume that the larger the $C_{ji}$ value, the more likely that the read $r_j$ has function *i*. Let $R_i$ denote the proportion of reads having function *i*. Even if the read $r_j$ is from function *i*, it is also possible that $C_{ji}$ is not exactly as the same as $L_j$, the length of the codon sequence, due to sequencing errors and/or single nucleotide polymorphism (SNP) effect. For the simplest model we will consider only one type of mismatch. Let *p* denote the probability of observing a mismatched codon, then 1-*p* is the probability of observing an identity or conserved codon. Therefore the probability that a read $r_j$ has function i with $C_{ji}$ matched codons and $L_j - C_{ji}$ mismatched codons is $R_i p^{L_j - C_{ji}} (1- p)^{C_{ji}}$. Then the probability to observe a read $r_j$ in the dataset is

$$\Pr(r_j) = \sum_{i=1}^{K} \left[ R_i p^{L_j - C_{ji}} (1-p)^{C_{ji}} \right] \quad (1)$$

Hence the likelihood function of the data is:

$$\ell(p, R_1, ..., R_K) = \prod_{j=1}^{n} \Pr(r_j) = \prod_{j=1}^{n} \sum_{i=1}^{K} \left[ R_i p^{L_j - C_{ji}} (1-p)^{C_{ji}} \right] \quad (2)$$

In this likelihood function, the lengths of the reads $L_j$ are given and the values of $C_{ji}$ can be extracted from the BLASTX result. The parameters *p* and $R_i (i = 1, 2, \cdots, K)$ are then

estimated by Expectation Maximization (EM) algorithm [17]. Here $p$ is the probability for the mismatch at the codon level.

**Multiple roles assignment**

One read could get involved in multiple functional roles. For read $j$, assume its best mismatch (i.e., minimum value) is $M_{j0}$, we can determine the maximum allowable mismatch $M_{j1}$ for a given small probability $\alpha$ such that:

$$\Pr(M_{j0}) \le m_j \le M_l) \le \varepsilon \quad (3)$$

where we assume that the mismatch follows a binomial distribution with $(L_j, p)$. Then the read $j$ can be assigned to all the functions with mismatch $\le M_{j1}$. The relative abundance $R_i$ will be updated by this new multiple function role assignment. Figure 3 illustrates the calculation for multiple function assignment based on the binomial distribution. In this illustration the length of the aligned codons is 32 and sequencing error at the codon level is 0.15. If the best mismatch is 0 and the probability $\varepsilon$ = 0.05, then the maximum allowable mismatch is calculated as 1. It means that the functions with matched codons of 32 or 31 (=32-1) in blastx output are possible functions and therefore the read is assigned to these putative functions. The small probability $\varepsilon$ in equation (3) is suggested as one third of sequencing error at the codon level estimated in the first step or just the sequencing error at the nucleotide level, if known or given.

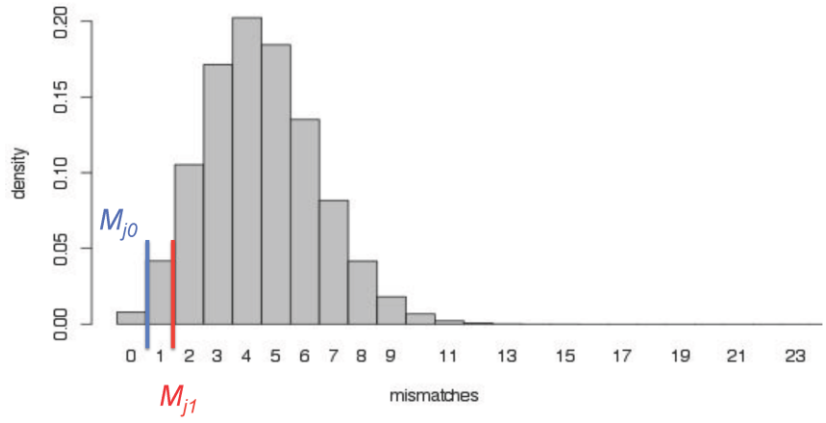


**Figure 3.** Illustration of calculation of multiple function assignment. In this plot $\varepsilon = 0.05$ and the binomial distribution has $p$=0.15 and and $L_j$=*32*.

## Simulation studies

### *Experimental data*

Due to the complexity of metagenomic data, simulation studies with verifiable structure are crucial to benchmark the proposed approach and to conduct comparisons with other existing methods including MEGAN4 standalone software (version 4.60.2) and the MG-RAST (http://metagenomics.anl.gov/). So far there is no literature about how to set up a simulation study for functional metagenomics. We propose to use the SEED database (http://pseed.theseed.org) and conduct six simulation studies. Basic information of these six simulation settings is listed in Table 1. Similar to the studies in MetaSim [18] which contain a small number of genomes in each setting we simulate a small number of functions in each study.

Study 1 contains 10 primary function roles that are far away from each other in the SEED tree structure. For each primary function role, 20% of the total sequences (i.e., sampling rate is 20%) in the SEED database are chosen and chopped randomly into segments of 100 bp long, and then 2% sequencing error is added to each short read. The sequencing error could be due to the substitution, deletion and insertion. For the purpose of method illustration we only consider the substitution error. It is well know that some genes are involved in multiple functions in a microbial community. This is also reflected from the gene sequences in the SEED database, i.e., some sequences are labeled with multiple functions. As expected, a few additional function names are obtained for the short sequences in the 10 primary groups. We name them secondary functions, as they are not our originally chosen functions.

Study 2 uses the same 10 primary functions as study 1 but with various sampling rate. The number of short sequence reads generated for each function is based on the total number of sequences in the function group in the SEED database. Generally, the sampling rate varies between 20% ~ 40%. In studies 3 and 4 we use another set of 10 functions. Different from the studies 1 and 2, the 10 function groups here are very closely related (i.e, some functional roles are belong to the same subsystems). Study 5 contains the same 10 primary function groups as studies 1 & 2 but the sampling rate is much larger, about 4~5 time of the first two studies. Similarly, study 6 contains the same 10 primary function groups as studies 3 & 4 but the sampling rate is about 4~5 times of these two studies.

**Table 1.** Basic information of six simulation studies.

| Study | Characteristic of the 10 primary functional roles | Sampling rate from SEED database |
|---|---|---|
| 1 | Different | fixed 20% |
| 2 | Different | 20~40% |
| 3 | Closely related | fixed 20% |
| 4 | Closely related | 20~40% |
| 5 | Same as study 1 & 2 | Large sample size |
| 6 | Same as study 3 & 4 | Large sample size |

# RESULTS

## Simulation Results

Three methods, MEGAN, MG-RAST and the proposed method metaFunction, are compared through these six simulation studies. The result for the first simulation study is shown in Figure 4 where it plots the relationship between the estimated abundance for each function and its true (i.e., expected) abundance. If all the functions are detected and their abundances are correctly estimated then the Pearson correlation between the expected and estimated abundances is one. From the plot it is obvious that the proposed approach has largest correlation. Table 2 displays the summary information for the correlations in all six studies. The new method outperforms other methods in all studies in terms of correlation between the true and estimated abundances.

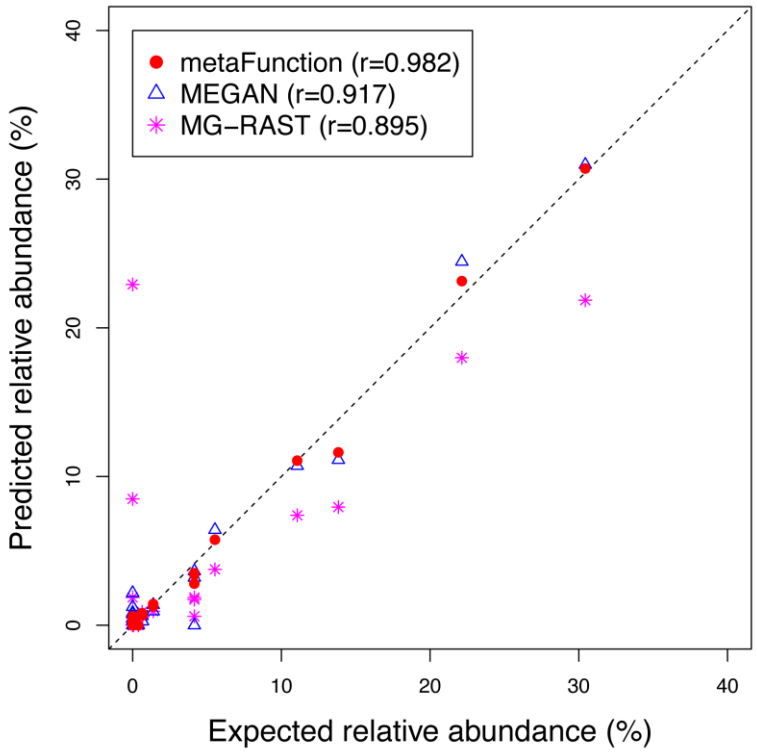


**Figure 4.** Plot of the predicted relative abundance of the detected functions under different methods vs their corresponding expected relative abundance (i.e., the truth) in the first simulation study.

**Table 2.** Summary of the correlation values between the expected and estimated abundance for the simulated functions in all six studies, for three methods.

| | Simulation 1 | Simulation 2 | Simulation 3 | Simulation 4 | Simulation 5 | Simulation 6 |
|---|---|---|---|---|---|---|
| MEGAN | 0.986 | 0.968 | 0.852 | 0.839 | 0.973 | 0.917 |
| MG-RAST | 0.711 | 0.696 | 0.880 | 0.857 | 0.750 | 0.895 |
| metaFunction | 0.996 | 0.993 | 0.953 | 0.943 | 0.997 | 0.982 |

We also evaluate the performance of three methods via the same simulations using another metric. A common measure for error is root mean square of relative error [6,19]. In this definition each feature group is assumed the same weight in the error

calculation, regardless the abundance of features in each group. In function analysis of metagnomics a function group estimated with tiny number of counts actually should be much less likely exists in the sample than a group with large number of read counts. We modify the error measure to weighted root mean square of relative error, i.e.,

$$WRRMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(w_i\left(\frac{a_i - t_i}{t_i}\right)^2\right)}, \text{ where the weight } w_i = \frac{\log(e_i)}{\sum_{j=1}^{m}\log(e_j)},$$

$e_i$ is the estimated number of reads for function $i$, $a_i$ is the estimated relative abundance (i.e., estimated proportion) and $t_i$ is the true relative abundance, and $m$ is the number of true function groups. The WRRMSE results for six studies are shown in Figure 5. In each of subplots the x-axis is the SEED system level. Compared to the MEGAN and MGRAST, the proposed method has the lowest error at any level of the subsystems and for all simulation studies.

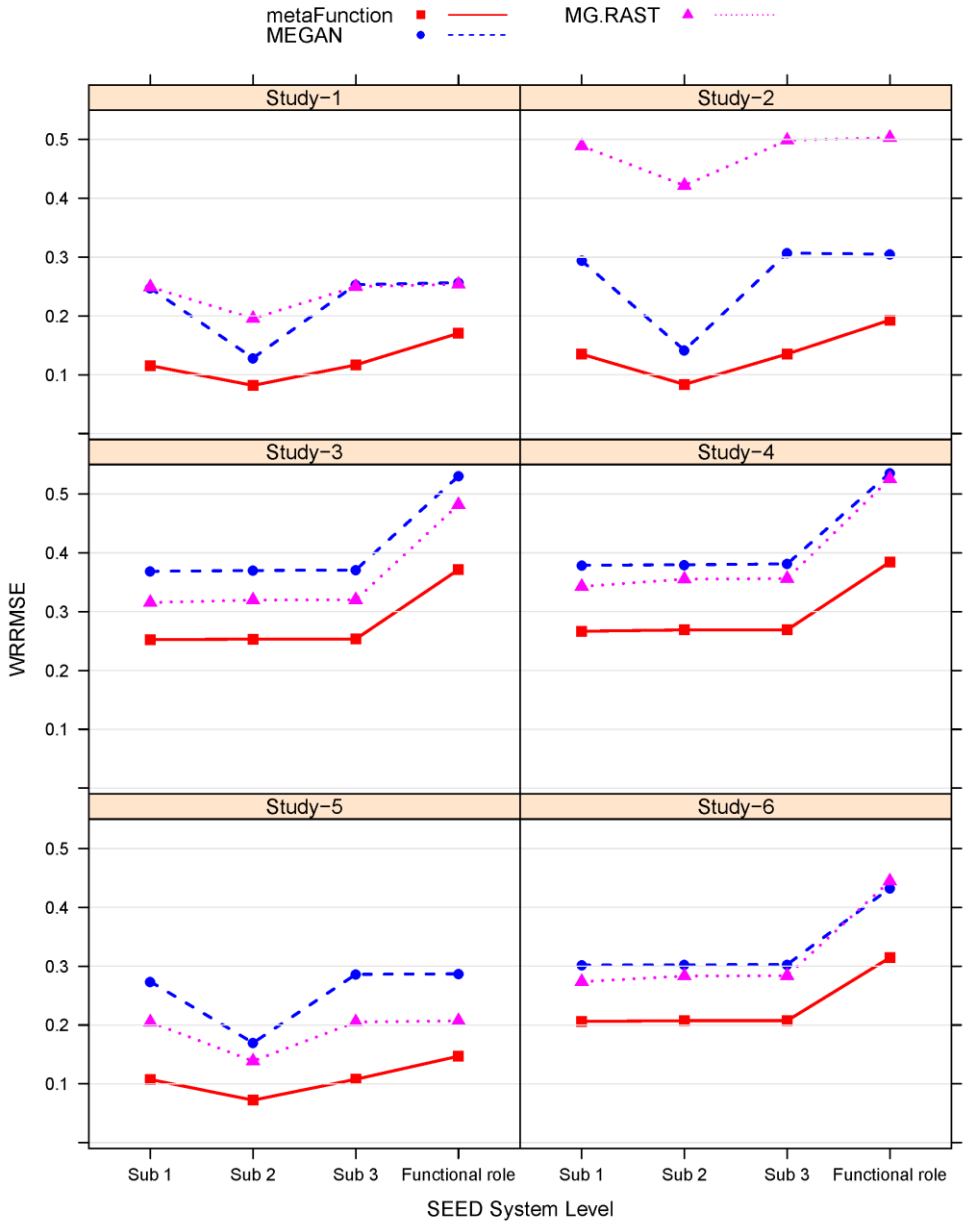


**Figure 5.** Weighted Root of Mean Square Relative Error (WRRMSE) between the estimated functionality profiles by each method (MEGAN, MG-RAST, and metaFunction) and the true proportion profiles at different SEED tree levels for six simulation studies.

The accuracy of estimation of relative abundance plays an important role in metagenomic analysis, the accuracy of assignment of short reads is also very interesting to biologists in functional metagnomics as they need the information of what reads do what kind of functions. As the MG-RAST does not give the information of the assignment we compare the performance of MEGAN and the proposed method metaFunction regarding the assignment details. In each simulation study we calculate the proportion of correctly assigned (CA), wrongly assigned (WA), and not assigned (NA, i.e., not aligned to the reference database) across all functions. The assignment details are also examined at other levels of the subsystem. The results of simulation study 1 are displayed in Table 3. At any level of the subsystems (including the function level) the proportions of NA using metaFunction are lower than those from the MEGAN result. Though the WAs for metaFunction are higher than the ones for the MEGAN they are comparable (all <1%). The new approach metaFunction results in much higher CA rate than MEGAN (about 90% vs 70%). Consistent conclusions are obtained for other simulation studies (data not shown).

**Table 3.** Proportion of correctly assigned (CA), wrongly assigned (WA), not assigned (NA) simulated reads under two methods at different levels of the SEED tree for the first simulation study.

| | **MEGAN** | | | **metaFunction** | | |
|---|---|---|---|---|---|---|
| | CA (%) | WA (%) | NA (%) | CA(%) | WA (%) | NA (%) |
| Function | 77.45 | 0.22 | 22.55 | 91.06 | 0.39 | 8.94 |
| Subsystem 3 | 77.22 | 0.22 | 22.78 | 91.05 | 0.38 | 8.95 |
| Subsystem 2 | 76.66 | 0.21 | 23.34 | 91.11 | 0.37 | 8.89 |
| Subsystem 1 | 76.66 | 0.21 | 23.34 | 91.11 | 0.37 | 8.89 |

**Real data analysis**

Real metagenomic data from an environmental study and human health study are analyzed using the proposed method.

***Environmental study***

Metagenomic functions were compared between Lake Erie (North America) and Lake Taihu (China) [20]. Toxic cyanobacterial blooms appear to be a global problem as toxins produced by bloom-associated cyanobacteria can have drastic impacts on the ecosystem and surrounding communities; in addition, the produced bloom biomass can disrupt aquatic food webs and act as a driver for hypoxia. Freshwater samples were collected from different lakes to examine the bloom associated microbial communities. We select two lakes - Lake Erie and Lake Taihu as they represent different continents – to examine the gene contents. After quality checking totally 750 thousands reads with an average length of 425bp are aligned to the NCBI non-redundant database. The proposed method is applied to the two lake samples. In order to compare our results to the findings in the original paper, the functionality profiles of microbial communities in these two lakes are summarized at the level 1 of subsystem (Figure 6).

Though the COG classification was used in the original paper and we used the SEED subsystems, the results from these two systems are highly consistent. For instance, "*Amino Acid and Derivatives*", "*Carbohydrates*", "*Nucleosides and Nucleotides*", and "*Membrane Transport*" are found more abundant in Lake Taihu than in Lake Erie. "*Cell*

*Division and Cell Cycle*", "*Regulation and Cell signaling*", and "*Iron acquisition and metabolism*" are lower in Lake Taihu. Even more, our method can give the detailed comparison between two lakes at a lower level of subsystems, e.g., the profile of subsystem level 3 (see the supplementary Figure_S1).

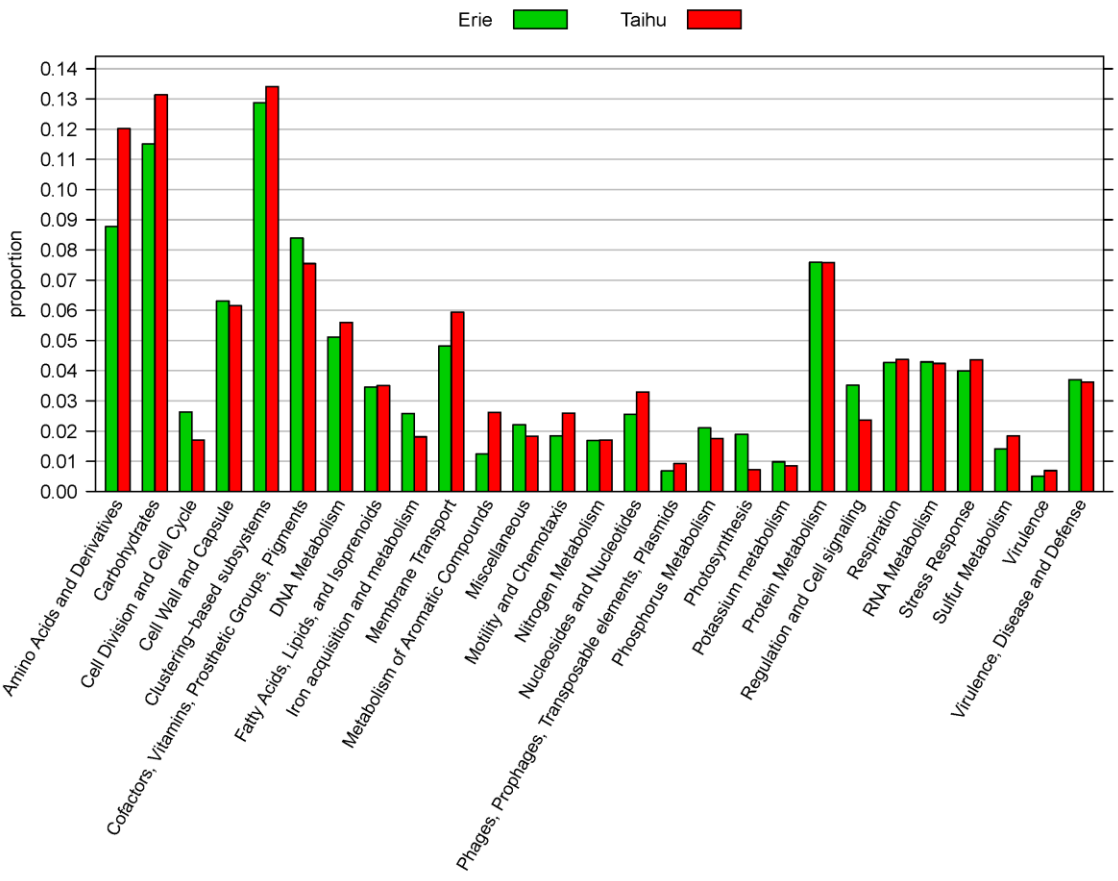


**Figure 6.** Proportions of the detected subsystems at the level 1 in the SEED structure. The top 28 subsystems with proportion >0.005 in at least one of samples are listed. Red bars represent the Taihu Lake sample and green represents the Erie Lake sample.

***Human health study***

Human oral microbial samples were studied for oral cavity problem using 454 pyrosequencing [21]. Two healthy samples and two cavity samples are selected for our analysis, with one at an intermediate stage and the other one at an advanced stage of caries development. After quality checking, 0.5Gbp of sequence with the average read length 425bp are BLASTXed to NCBI-NR protein database for searching matched reference sequences (i.e., genes). Then reads are classified into functional role categories as defined in the SEED database using the proposed method. The results of functionality profiling for all four samples at the subsystem level 3 are shown in Figure 7.

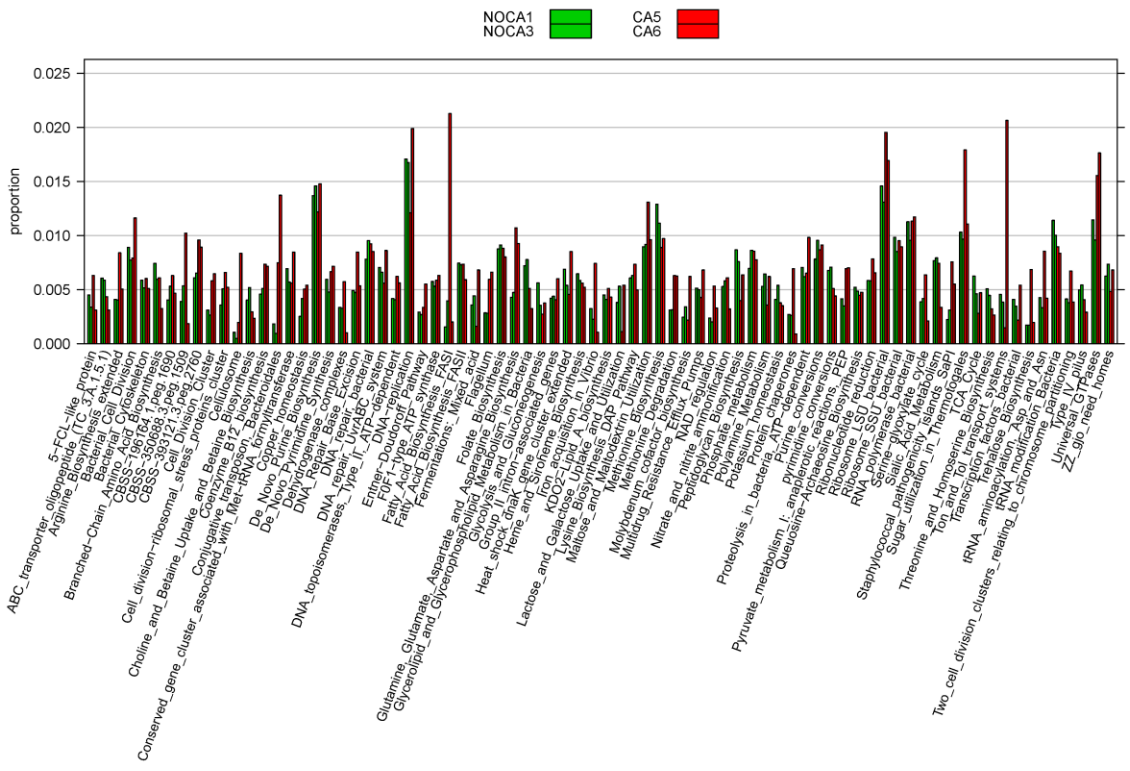

**Figure 7.** Proportions of the detected subsystems at the level 3 in the SEED structure. The top 78 subsystems with proportion >0.005 in at least one of samples are listed. Red bars represent the cavity samples and green represents the healthy samples.

In this plot the abundance of "*Conjugative transposon Bacteroidales*" is much higher in the cavity samples than in the healthy orals, which is also confirmed in other literature [22]; "*Fatty Acid Biosynthesis FASI*" also shows a higher value in the diseased samples than in the healthy samples, which is consistent with the finding in [23]; That the "*Flagellum*" is abundant in the cavity samples is also reported in Seshadri et al. [24]; high values of "*Glutamine Glutamate, Aspartate and Asparagine Biosynthesis*" and of "*Methionine degradation*" in the oral cavity samples are also mentioned in other publications [25,26]; the abundance of "*Universal GTPases*" is higher in the cavity samples than in the healthy orals, which is also found in other literature [27]. In conclusion, the results from the proposed method provide us the findings consistent with the previous literatures.

## DISCUSSION

One of the main challenges in metagenomic studies is how to accurately identify all possible functional roles present in an environmental sample and precisely estimate their abundance. Due to the complexity of metagenomics and the huge volume of sequencing reads of short lengths obtained from the next generation sequencing technologies, the need of efficient statistical tools to accomplish this challenge is increasing. We proposed a two-step procedure to perform functional analysis on a

metagenome: mixture model coupled with the adjustment of multiple role assignment, to assign reads to relative functional roles by utilizing the SEED classification of functional roles and subsystems.

Compared to MEGAN and MG-RAST through comprehensive simulation studies, our procedure metaFunction can be more effective in assigning reads to a relative functional role. In the simulation study 1 and 2, the results show that MEGAN cannot assign any reads to one of the functional roles while in the simulation study 3 and 4, MG-RAST cannot assign any reads to one of the functional roles. This kind of phenomenon has never happened to our approach. In addition, the proposed method can correctly assign higher percentage of reads to functional roles than MEGAN does. Based on the hits obtained by aligning sequence reads against the reference database, in our method, multiple functional roles can be identified and assigned to a read by evaluating the likelihood of alignment. However, MEGAN and MG-RAST utilize the best bit-score for assignment. If a read returns with multiple best scores, then the first function (alphabetically) is chosen for the assignment. In our method all of them with the same best score are assigned to the read. Different from other existing methods, in the proposed method confidence interval can be calculated for all the parameters by using bootstrap.

We also applied the proposed method to two real metagenomic data sets and our findings are consistent with previous reports. A future work is to integrate the taxonomic analysis and functional analysis, in other words, to consider these two issues

simultaneously, so that the power can be improved for both taxonomic and functional profiling a metagenomic sample.

## ACKNOWLEDGEMENTS

*Funding*: This work was supported by National Science Foundation DMS-1043080 to L.A. and H.J and DMS-1222592 to L.A. H.J and J.X.

## REFERENCES

1. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377-386.
2. Rosen GL, Sokhansanj BA, Polikar R, Bruns MA, Russell J, et al. (2009) Signal Processing for Metagenomics: Extracting Information from the Soup. Current Genomics 10: 493-510.
3. Mardis ER (2008) Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9: 387-402.
4. Clemente JC, Jansson J, Valiente G (2011) Flexible taxonomic assignment of ambiguous sequencing reads. BMC Bioinformatics 12.
5. Meinicke P, Asshauer KP, Lingner T (2011) Mixture models for analysis of the taxonomic composition of metagenomes. Bioinformatics 27: 1618-1624.
6. Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F (2011) Accurate genome relative abundance estimation based on shotgun metagenomic reads. Plos One 6: e27992.
7. Jiang H, An L, Lin SM, Feng G, Qiu Y (2012) A statistical framework for accurate taxonomic assignment of metagenomic sequencing reads. Plos One 7: e46450.
8. Lindner MS, Renard BY (2013) Metagenomic abundance estimation and diagnostic testing on species level. Nucleic Acids Research 41: e10.
9. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Research 33: 5691-5702.
10. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. Nature 452: 629-632.
11. Parks DH, Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26: 715-721.
12. Sharon I, Bercovici S, Pinter RY, Shlomi T (2011) Pathway-based functional analysis of metagenomes. J Comput Biol 18: 495-505.
13. Yooseph S, Li WZ, Sutton G (2008) Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. BMC Bioinformatics 9.

14. Mitra S, Rupek P, Richter DC, Urich T, Gilbert JA, et al. (2011) Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. BMC Bioinformatics 12.
15. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9.
16. Hoff KJ, Lingner T, Meinicke P, Tech M (2009) Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Research 37: W101-W105.
17. Dempster APea (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39: 38.
18. Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008) MetaSim-A Sequencing Simulator for Genomics and Metagenomics. Plos One 3.
19. Engeman RM, Sugihara RT, Pank LF, Dusenberry WE (1994) A Comparison of Plotless Density Estimators Using Monte-Carlo Simulation. Ecology 75: 1769-1779.
20. Steffen MM, Li Z, Effler TC, Hauser LJ, Boyer GL, et al. (2012) Comparative metagenomics of toxic freshwater cyanobacteria bloom communities on two continents. Plos One 7: e44002.
21. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simon-Soro A, et al. (2011) The oral metagenome in health and disease. ISME J.
22. Ready D, Pratten J, Roberts AP, Bedi R, Mullany P, et al. (2006) Potential role of Veillonella spp. as a reservoir of transferable tetracycline resistance in the oral cavity. Antimicrobial Agents and Chemotherapy 50: 2866-2868.
23. Fozo EM, Scott-Anne K, Koo H, Quivey RG (2007) Role of unsaturated fatty acid biosynthesis in virulence of Streptococcus mutans. Infection and Immunity 75: 1537-1539.
24. Seshadri G, Myers GSA, Tettelin H, Eisen JA, Heidelberg JF, et al. (2004) Comparison of the genome Treponema denticola with of the oral pathogen other spirochete genomes. Proc Natl Acad Sci U S A 101: 5646-5651.
25. Park SN, Kong SW, Kim HS, Park MS, Lee JW, et al. (2012) Draft Genome Sequence of Fusobacterium nucleatum ChDC F128, Isolated from a Periodontitis Lesion. Journal of Bacteriology 194: 6322-6323.
26. Yoshimura M, Nakano Y, Yamashita Y, Oho T, Saito T, et al. (2000) Formation of methyl mercaptan from L-methionine by Porphyromonas gingivalis. Infection and Immunity 68: 6912-6916.
27. Karlsson C, Malmstrom L, Aebersold R, Malmstrom J (2012) Proteome-wide selected reaction monitoring assays for the human pathogen Streptococcus pyogenes. Nature Communications 3.

# SUPPLEMENTARY

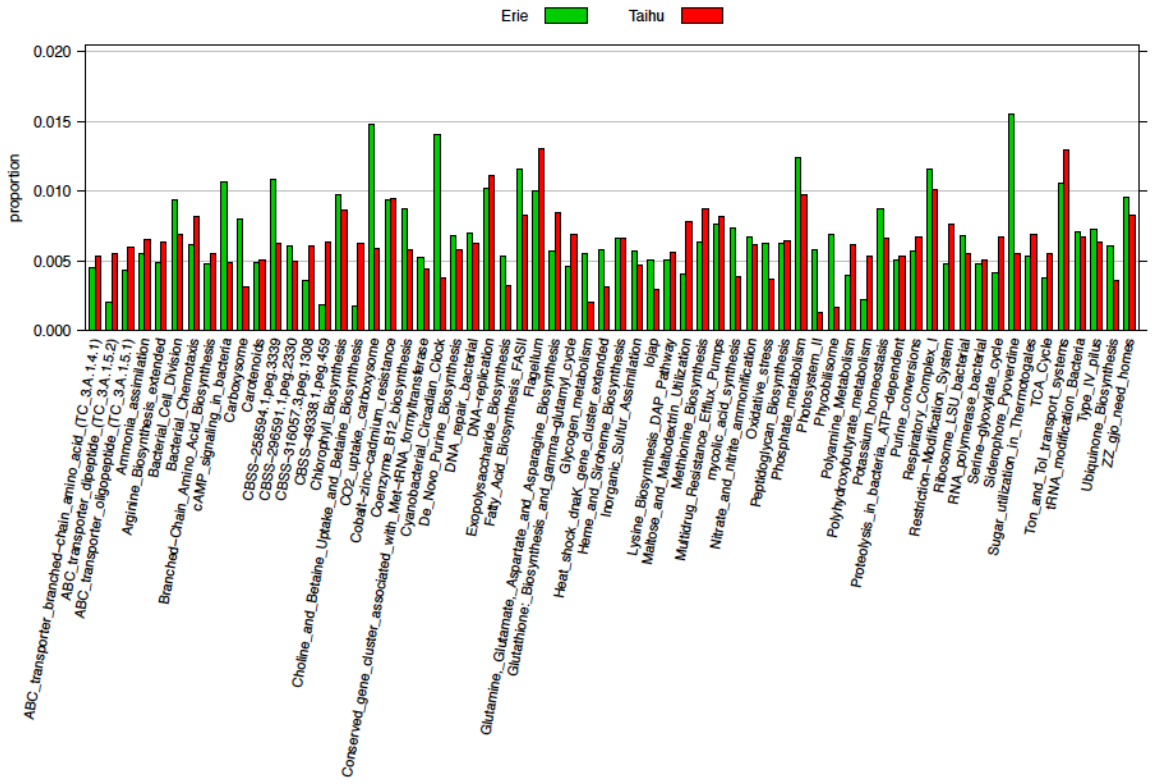


**Figure S1.** Proportions of the detected subsystems at the level 3 in the SEED structure. The top 64 subsystems with proportion >0.005 in at least one of samples are listed. Red bars represent the Taihu Lake sample and green for the Erie Lake sample.of the detected subsystems at the level 3 in the SEED structure. The top 64 subsystems with proportion > 0.005 in at least one of samples are listed.

# APPENDIX B – A TWO-STAGE STATISTICAL PROCEDURE FOR FEATURE SELECTION AND COMPARISON IN FUNCTIONAL ANALYSIS OF METAGENOMES

**N. Pookhao, L. An, M. Sohn, Q. Li, I. Jenkins, R. Du, H. Jiang**
Submitted to: *Bioinformatics*

**Abstract**
**Motivation:** With the advance of new sequencing technologies producing massive short reads data, metagenomics is rapidly growing, especially in the fields of environmental biology and medical science. The metagenomic data are not only high dimensional with large number of features and limited number of samples, but also complex with a large number of zeros and skewed distribution. Efficient computational and statistical tools are needed to deal with these unique characteristics of metagenomic sequencing data. In metagenomic studies, one main objective is to assess whether and how multiple microbial communities differ under various environmental conditions.
**Results:** We propose a two-stage statistical procedure for selecting informative features and identifying differentially abundant features between two or more groups of microbial communities. In the functional analysis of metagenomes the features may refer to the pathways, subsystems, functional roles, and so on. In the first stage of the proposed procedure, the informative features are selected using elastic net as reducing the dimension of metagenomic data. In the second stage the differentially abundant features are detected using generalized linear models with a negative binomial distribution. Compared with other available methods the proposed approach demonstrates better performance for most of the comprehensive simulation studies. The new method is also applied to two real metagenomic datasets related to human health. Our findings are consistent with those in previous reports.

**Availability:** R codes may be requested from the corresponding author.
**Contact:** anling@email.arizona.edu

**Keywords**
Metagenomics, Feature selection, Feature comparative, Generalized linear model, Elastic-net, Negative binomial model.

## 1 INTRODUCTION

Recently next generation sequencing technologies are able to produce high volumes of data at an affordable cost (Gilbert et al., 2011; Huson et al., 2009). The power

of next generation sequencing makes it possible to explore microbial environments, opening a new era of genomics study, called metagenomics (Gilbert et al., 2011). Metagenomics is the study of genomic contents of microbial communities sampled directly from environments (e.g., soil, water, human gut) without prior culturing to understand the true diversity of microbes, their functions, cooperation and evolution in different microbial communities (Wooley and Ye 2010; Huson et al., 2009, Kunin et al., 2008; Hugenholtz 2002). Importantly, since only about 1% of all microbial organisms can be isolated and cultured in a laboratory, metagenomic analysis enables to reveal the genome contents of the majority of microorganisms that cannot be obtained in traditional genomic analysis based on pure culture (Wooley and Ye 2010; Hugenholtz 2002). Metagenomics is broadly applicable to many areas, including ecology and environmental sciences, chemical industry, and biomedicine (Wooley and Ye 2010; Turnbaugh et al., 2007).

In metagenomic analysis, one important aim is to assess whether and how two or more microbial communities differ. To perform metagenomic comparison, researchers can conduct an experiment to compare genomic features based on either taxonomic compositions or functional components obtained from different microbial communities. In this study, we focus on comparison of functions in metagenomes under various conditions. The applications of this research include detection of biological threats and discovery of new bioenergy and new medicine, and so on. For example, comparing microbial communities from human gut corresponding to different phenotypes (e.g.,

diseased and healthy, or different treatments) can help us determine the activities of microbes related to the disease, resulting in understanding the reactions of microbes that respond to different biochemical products. This may lead to drug development or treatment selection that specifically affects to either a particular activity or a group of activities that the disease related microbes might perform.

Statistical procedures play a critical role in detecting differentially abundant features across different microbial conditions. The features here may refer to taxa, functional role, pathway, or subsystems. Several statistical methods or tools have been developed to compare various microbial communities in terms of detecting differentially abundant features, e.g., SONs (Schloss and Handelsman 2006), XIPE-TOTEC (Rodriguez-Brito et al., 2006), Metastats (White et al., 2009), and MEGAN (Huson et al., 2011; Huson et al., 2009). However, all of these methods/tools are designed to compare exactly two microbial conditions; ShotgunFunctionalizeR uses a regression method on comparing multiple samples (Kristiansson et al., 2009) but it assumes Poisson distribution on the count data. It is well known that Poisson model lack flexibility for over-dispersed count data (Rapaport et al., 2013). Another method, metagenomeSeq (Paulson et al., 2013), has been recently developed to assess differential abundance in sparse high-throughput microbial marker-gene survey data. Even though it can compare multiple conditions, metagenomeSeq is designated for comparison of taxonomic compositions of different metagenomes, rather than functional compositions. In this

research, we focus on statistical comparison of functions in metagenomes under various conditions.

Statistical methods developed for RNA-Seq analysis may be applicable to metagenomic analysis also, as both RNA-Seq experiments and metagenomic experiments use sequencing technologies and produce count data. A number of statistical tools have been developed for RNA-Seq data analysis, such as edgeR (Robinson et al., 2010) and DESeq (Anders and Huber 2010). However, there are differences between RNA-Seq data and metagenomic data. Different from RNA-Seq data, one of the common characteristics of metagenomic data is the presence of many features with zero counts. It is because metagenomic samples consist of a mixture of microbes, the species-specific functions may only appear in some microbial conditions, while in typical RNA-Seq experiments the genes are the same for different experimental conditions and only expression levels change. Thus, metagenomic sequencing data may be more sparse than the RNA-seq data.

Our research was motivated by (i) the limitations of existing methods developed for metagenomic analysis, (ii) the increasing focus of metagenomic projects on wide applications in various areas (e.g. Human Microbiome Project (Turnbaugh et al., 2007)), and (iii) the limitations of applying current methods developed for RNA-Seq analysis to metagenomic analysis. In this paper, we propose a two-stage statistical algorithm for selecting informative features and detecting differentially abundant functional features (e.g. pathways, subsystems, functional roles) between different microbial conditions. In the 1$^{st}$ stage of our algorithm, the informative features are selected using elastic net

(Friedman et al., 2010) resulting in dimensional reduction of the metagenomic dataset. In the 2nd stage of our approach, we detect differentially abundant features using generalized linear models (GLMs) with a negative binomial (NB) distribution (Venables and Ripley 2002).

In sparse data, elastic net is a satisfactory variable selection method in the case that the number of predictors ($p$) is much bigger than the number of observations ($N$), that is, when $p >> N$. In addition, another advantage of elastic net is that it is well suitable to data containing a grouping effect, i.e., strongly correlated predictors tend to be in or out of the model together (Friedman et al., 2010; Zou and Hastie 2005). The NB distribution is widely used to model count data. The novelty of our two-step method is that we take the common characteristics of metagenomic data into account and combine the feature selection and feature comparison in metagenomic study to improve the power of feature detection.

Our method can be directly applied to comparison of more than two microbial conditions. Therefore, our method can be applicable to more general situations, e.g., in clinical trials where the goal is to compare multiple treatment conditions or in natural environmental studies where multiple conditions are compared and investigated.

## 2 METHODS

Our approach relies on two assumptions: (i) we are given a metagenomic dataset corresponding to multiple populations/conditions with different phenotypes (e.g. diseased

and healthy human guts, or different locations of sea water); each population consists of multiple individuals (or samples), and (ii) each sample consists of count data representing the relative abundance of specific features within each sample, or number of shotgun reads mapped to a specific biological pathway or subsystem. Our goals are to determine a set of informative features associated with a particular phenotype and to identify statistically significant features whose abundance is different among different populations/conditions.

### 2.1 Data normalization

Due to the high-throughput sequencing technologies, an arbitrary number of reads with large variation across samples is generated under the sampling process. Therefore, a common source of bias in a metagenomic count data is due to different sequencing depths or different levels of reads across multiple individuals (or samples). To proceed with any statistical analysis, a preprocessing of the metagenomic count data is necessary to account for this source of bias, i.e., normalizing the samples in order to make them comparable. For the data normalization, we used the trimmed mean of *M*-values (TMM) (Robinson and Oshlack 2010) which is implemented in the edgeR Bioconductor package.

### 2.2 Two-stage statistical procedure

In the proposed two-stage statistical algorithm, informative features are simultaneously selected in the 1$^{st}$ stage, and then the selected features obtained from the

1st stage are used as the input for the 2nd stage. Differentially abundant features between metagenomic datasets are detected in the 2nd stage.

***1st Stage - Feature selection using elastic net***

The 1st stage aims to detect informative features associated with a particular phenotype. This results in the dimensional reduction of the metagenomic data. As outlined in the introduction, the metagenomic data consist of relative abundances where low abundant microorganisms may be missed due to the sampling process. A statistical method is needed to deal with a sparse data with the presence of a large percentage of zero counts. Elastic net, an algorithm for estimation of generalized linear models with elastic-net penalties, enables to deal efficiently with sparse features (Friedman et al., 2010).

For the 1st stage, two matrices, a *feature count matrix* and a *phenotype matrix*, are required as the input (Figure 1) where $X$ represents the feature count matrix with $p$ features and $N$ samples, and where $Y$ represents the phenotype vector with two categories $y_i \in \{1,2,...,K\}$ where $i = 1,\ldots, N$. When there are only two categories ($K$=2), e.g. diseased and healthy, $y_i \in \{1,2\}$. For the feature count matrix, rows correspond to specific features, and columns correspond to individual metagenomic samples. The element $x_{ij}$ denotes the total number of reads (or relative abundance) of feature $i$ in sample $j$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,N} \\ x_{21} & x_{22} & \cdots & x_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p,1} & x_{p,2} & \cdots & x_{p,N} \end{bmatrix}, \text{ and } Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

**Figure 1.** Format of a feature count matrix and a phenotype matrix.

*Algorithm for elastic net*

In the algorithm of linear model, we have a response variable $Y \in \Re$ and a predictor matrix $X \in \Re^p$, and the regression function is typically determined by $E(Y \mid X = x) = \beta_0 + x^T \beta$ . For $N$ observation pairs ($x_i$, $y_i$), the elastic net solves the following problem

$$\min_{(\beta_0,\beta)\in\Re^{p+1}} R_\lambda(\beta_0,\beta) = \min_{(\beta_0,\beta)\in\Re^{p+1}} \left[ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda P_\alpha(\beta) \right], \quad (1)$$

where

$$P_\alpha(\beta) = \sum_{j=1}^{p} \left[ \frac{1}{2}(1-\alpha)\beta_j^2 + \alpha \mid \beta_j \mid \right] \quad (2)$$

$P_\alpha$ is the elastic-net penalty (Zou and Hastie 2005), and is a compromise between the ridge regression penalty ($\alpha$=0) and the lasso penalty ($\alpha$=1). The elastic net model with $\alpha$=1-$\varepsilon$ for some small $\varepsilon$ ($\varepsilon > 0$) performs much like the lasso, but ignores behavior caused by extreme correlations. This model will tend to pick one feature and ignore the rest if the features are correlated. On the other hand, the elastic net model with $\alpha$=1-$\varepsilon$ for some large $\varepsilon$ ($\varepsilon > 0$) performs much like the ridge regression, which is known as a regression model to shrink the coefficients of correlated predictor variables towards each other,

resulting them to borrow strength from each other. The coordinate descent step used to solve (1) is detailed in Friedman et al. (2010).

*Regularized multinomial regression*

When the response variable is binary ($K$=2), the linear logistic regression model is often used. When the categorical response variable $Y$ has multiple levels ($K$>2), the linear logistic regression model can be generalized to a multi-logit model. The class-conditional probability is represented through a linear function of the predictors:

$$\log \frac{\Pr(G=\ell \mid x)}{\Pr(G=K \mid x)} = \beta_{0\ell} + x^T \beta_\ell, \ell = 1,...,K-1 \tag{3}$$

Here $\beta_l$ is a p-vector of coefficients. For each value of λ, the parameters ($\beta$s) are computed by solving the penalized multinomial log-likelihood problem:

$$\max_{\{\beta_{0\ell},\beta_\ell\}_1^K \in \Re^{K(p+1)}} \left[ \frac{1}{N} \sum_{i=1}^{N} \log(\Pr(y_i = \ell \mid x_i)) - \lambda \sum_{\ell=1}^{K} P_\alpha(\beta_\ell) \right] \tag{4}$$

*Selecting the tuning α and $\lambda$ parameters for regularization path*

As shown in (1), two types of constraints (lasso and ridge constraints) on the parameters are employed in the elastic net. The parameter $\alpha$ controls the relative weight of these constraints. The lasso constraints allow for the selection/removal of variables in the model while the ridge constraints can deal with correlated predictor variables. In our approach, as the second step can deal with feature detection, in the elastic net step we put more weight upon the ridge constraints to deal with correlated features. We use grid

search for $\alpha$ in [0, 0.1] and for each parameter $\alpha$ the corresponding $\lambda$ was determined by cross-validation (CV) (Hastie et al., 2009). The values for the parameters $\alpha$ and $\lambda$ which yield the lowest cross-validated (CV) error were selected.

### ***2nd Stage – Differentially abundant feature detection***

The 2nd stage of our algorithm is to detect features, which are statistically differentially abundant in two or more populations. From examining real metagenomic count data, we discovered that the variance exceeds the corresponding mean of the feature abundance (detailed in Supplementary S1-S4). Negative binomial (NB) distribution, a commonly used model for count data with overdispersion, is used to take the overdispersion into account (Cameron and Trivedi 1998; Venables and Ripley 2002).

*Negative binomial model*

Let *Y* be the output of the 1st stage, thus it corresponds to the total number of reads for feature *i* in sample *j* where $i = 1,2,\ldots, r$ and $j = 1,2,\ldots, N$, assuming *r* out of *p* features are selected in the first stage. The count Y can be modeled by negative binomial distribution:

$$f_Y(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\cdot y!}\cdot\frac{\mu^y\cdot\theta^\theta}{(\mu+\theta)^{y+\theta}} \tag{5}$$

with mean $E(Y) = \mu = \exp(x^T\beta)$ and variance $\text{var}(Y) = \mu(1+\mu/\theta)$. The variance is quadratic in the mean. The NB distribution can also be reparameterized in the term of dispersion by letting $\phi = 1/\theta$. Then, the count *Y* follows NB with $E(Y) = \mu$ and

variance $\text{var}(Y) = \mu(1+\phi\mu)$ where $\phi$ denotes the dispersion parameter. The farther $\phi$ falls above 0, the greater the overdispersion relative to Poisson variability. Clearly, when $\phi = 0$, there is no overdispersion, and the NB distribution reduces to the usual standard Poisson distribution with parameter $\mu$, where $E(Y) = \text{var}(Y) = \mu$. In GLMs, The most convenient way to link the mean response $\mu$ of NB variable to a linear combination of the predictors *X* is the log link, as in Poisson loglinear models, $\log(\mu_i) = \exp(x_i^T \beta)$, where $x_i$ is 1 x *K* row vector of indicator variables, $i$ =1,2,…, $r$, *K* represents the number of microbial conditions in the dataset and $\beta$ is the corresponding $K\times1$ column vector of unknown regression parameters. The covariates can be introduced into a regression model based on the NB distribution via the relationship

$$\log(\mu_i) = \sum_{j=1}^{K} x_{ij}\beta_{j-1} \tag{6}$$

For the NB model, $\beta$ and $\phi$ are estimated by solving the maximum likelihood problem:

$$\ell(\mu_i, \phi; y_i) = \sum_{i=1}^{N}\left\{-\log(y_i) + \sum_{m=1}^{y_i}\log(\phi y_i - \phi m + 1) - (y_i - 1/\phi)\log(1+\phi\mu_i) + y_i\log(\mu_i)\right\} \tag{7}$$

More details on regression models for negative binomial responses can be found in Cameron and Trivedi (1998).

*Hypothesis testing of model parameters in GLMs*

To test the null hypothesis $H_0 : \beta_j = 0$, for the likelihood-ratio approach, denote the maximized value of the likelihood function by $\ell_0$ under $H_0 : \beta_j = 0$ and $\ell_1$ when $\beta_j \neq 0$. The *likelihood ratio* test statistic equals:

$$-2\log(\ell_0 / \ell_1) = -2[\log(\ell_0) - \log(\ell_1)] = -2(L_0 - L_1) \quad (8)$$

where $L_0$ and $L_1$ denote the maximized likelihood functions. Under $H_0 : \beta_j = 0$, this test statistic has a asymptotically chi-squared distribution with 1 degree of freedom.

*Multiple test correction*

A typical metagenomic dataset consists of several hundreds or thousands of features. After comparing multiple metagenomic groups using GLMs with the NB canonical logarithmic link function for simultaneously comparison, multiple comparison correction is needed for controlling the Type I errors. We used the Benjamini–Hochberg false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995) to control the type I error at significance level of 0.05.

## 3 SIMULATION STUDIES

Because of high similarity between RNA-Seq and metagenomic data, the statistical methods developed for RNA-Seq data in detecting differentially expressed genes may be applicable to the analysis of metagenomic data. For this reason we compared our method with two widely used statistical packages for RNA-Seq analysis, edgeR and DESeq, in addition to metagenomeSeq.

### 3.1 Experimental data

In order to make simulated data to reflect the nature of real metagenomic data we examined several types of real datasets from various environmental sources, including human gut, ocean, soil, and fresh water (Supplementary Table S1), and obtained the means and variances of feature abundance in these studies. Very interestingly, we observed strong linear relationships between the means and the variances after log transformation of the feature abundances (Figures S1-S4 in the Supplementary). We used this linear relationship to simulate metagenomic data.

*Experimental Design 1*

We designed a metagenomic simulation study in which subjects are drawn from two populations. Since the sample size affects the performance of statistical methods, we designed metagenomic datasets with various sample sizes, including 10, 25, and 50 subjects drawn from each population. For each dataset, counts were generated using NB distributions, with different means ($\mu$) and variances ($\sigma^2$). The means ($\mu$) of the NB distributions were selected by random sampling from the ranges of the abundant means in four simulation settings (Table 1), and the corresponding variances were computed from the simple linear regression function:

$$E(\sigma^2) = \beta_0 + \beta_1 * \mu \tag{9}$$

(In the first experiment let $\beta_0$ =0.6 and $\beta_1$ =1.8, which are from the observation of four real metagenomic datasets; details can be found in the supplementary file. In next experiment we will flex these two values). In each dataset, we simulated 1000 features for each sample of two populations from NB distributions; 950 of them were generated from the same NB distribution, i.e., $\mu_1 = \mu_2$ with the corresponding variances computed by (9), and the rest 50 were generated from two different NB distributions, i.e., $a * \mu_1 = \mu_2$, where the parameter *a* (i.e., multiplier) is selected from the set of 1.5, 2.5, 5, 7.5, and 10. To prevent bias arising from a specific partition, we simulated the datasets 100 times for each sample size. The performance of four methods were compared using the "Area Under the Curve" (AUC) metric of a Receiver Operator Curve (ROC) and the true positive rate (tpr, i.e., power) were calculated at each level of FDR.

**Table 1.** The ranges of abundant means of the NB distributions in four simulation settings. The feature abundance is prior transformed into log 10 scale. Settings 1-4 reflect the count data of feature abundances with *low* means, intermediate means, high means, and a combination means, respectively. Setting 4 most resembles to the nature of real metagenomic dataset.

| Setting | Minimum Mean (log 10 scale) | Maximum Mean (log 10 scale) |
|---|---|---|
| 1 (low) | -0.5 | 1 |
| 2 (intermediate) | 1 | 2.5 |
| 3 (high) | 2.5 | 5 |
| 4 (combined) | -0.5 | 5 |

*Experimental Design 2*

Different from the first experimental design where the values of $\beta_0$ and $\beta_1$ are fixed, the second experiment allows these two parameters to vary. They were determined

by random sampling from the ranges of [0.1, 1] and [1.5, 2], respectively. These ranges of the estimates for $\beta_0$ and $\beta_1$ were obtained from observing real metagenomic data (details in Supplementary). As the setting 4 resembles most to the nature of real metagenomic dataset, in the second experiment we flexed the $\beta_0$ and $\beta_1$ on this setting. Similar to the first experiment, we simulated 1000 features for each sample of the two populations from NB distributions: 950 of them were generated from the same NB distribution, and the rest were from two different NB distributions.

### 3.2 Simulation Results

*Results from Experimental Design 1*

ROC (receiver operating characteristic) curve is usually used in measuring signal detection. It is created by plotting the true positive rate vs. the false positive rate. Area under the ROC curve, AUC, shows an overall performance of detection methods. The higher the AUC value, the better the method is. Figure 2 displays the AUC results for four methods with different sample sizes (10, 25, and 50) under four simulation settings. AUC values generally increase when the sample size increases; AUC values are greater for higher mean setting. The proposed approach outperforms the other methods in the setting 2 and is well comparable to other methods in the rest of the settings. Very interestingly, the new approach surpasses others very much for the sample size of 10 in the intermediate mean case.

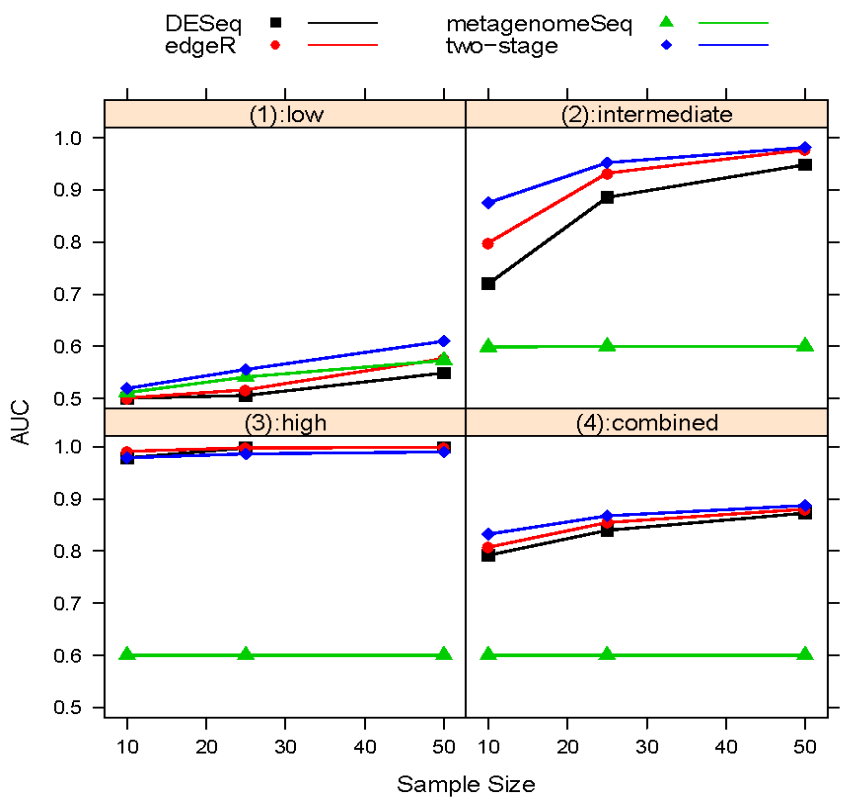


**Figure 2.** The AUC results for sample size of 10, 25, and 50 in each simulation setting in the experimental design 1. (1)-(4) show the AUC results for four settings, i.e, *low* means, intermediate means, high means, and combination of means, respectively.

In addition to the AUC which shows an overall performance of the methods, we also compare our method with other methods in term of power in detecting the true differentially abundant features while the type I error is controlled. The plot of power vs. FDR is more preferred than the above AUC plot as it displays the power under the controlled type I error, which is more important in signal/feature detection. Figure 3 shows the power for sample size of 10, 25, and 50 in each simulation setting in the experimental design 1. Our proposed approach outperforms other methods in most

situations and is well comparable to other methods in the situation where the data contains high abundant features with sufficient large sample size.

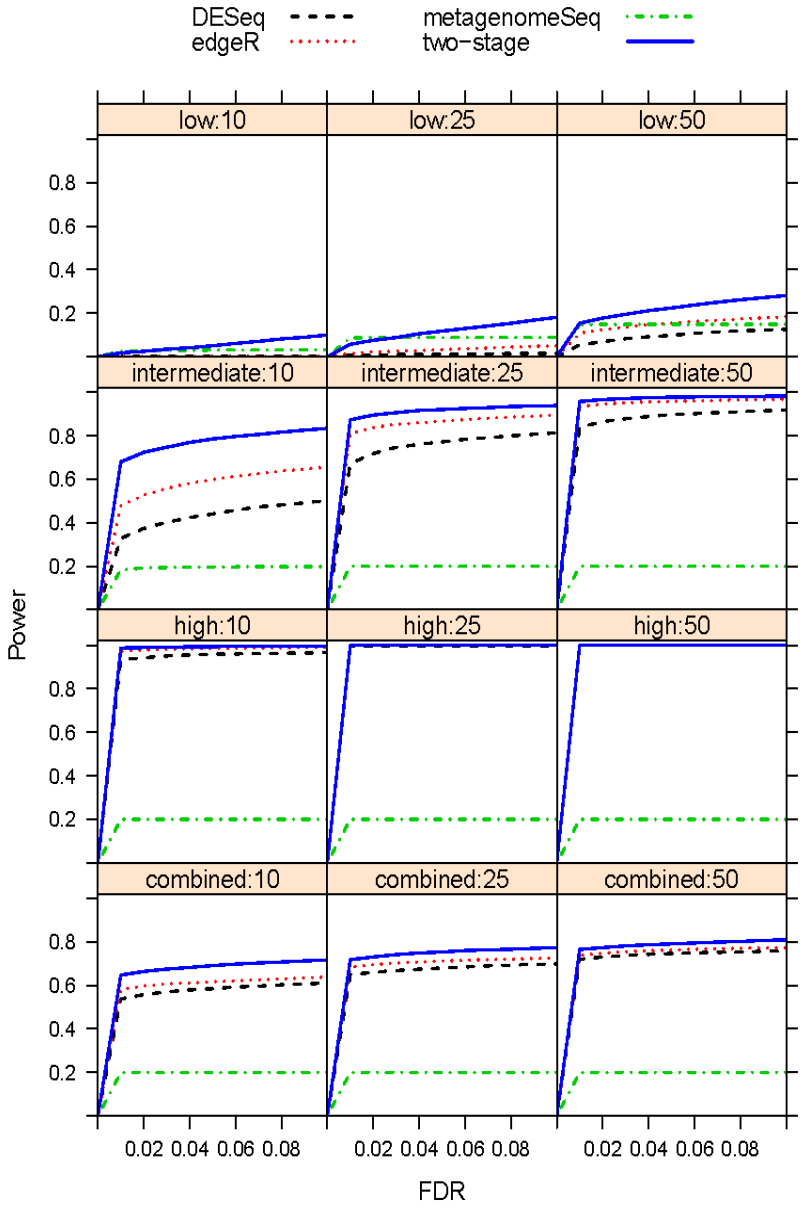


**Figure 3.** The power in detection of the true differentially abundant features for four methods at various level of FDR for sample size of 10, 25, 50. (1)-(4) show the power for four settings in the first experiment, i.e, *low* means, intermediate means, high means, and combination of means, respectively.

*Results from Experimental Design 2*

Figure 4 displays the AUC results, and Figure 5 shows the power detection of the true differentially abundant features obtained from each method for sample size of 10, 25, and 50 in the simulation setting in the experimental design 2. The results show that the proposed method and edgeR outperform DESeq and metagenomeSeq in situations with sample size of 10 and 25 and are comparable with DESeq in a situation with large sample size. The proposed approach has similar performance with edgeR when both of them are compared in term of AUC as shown in Figure 4. However, our proposed method outperforms edgeR and other methods in term of the power in detecting the true differentially abundant features as shown in Figure 5.

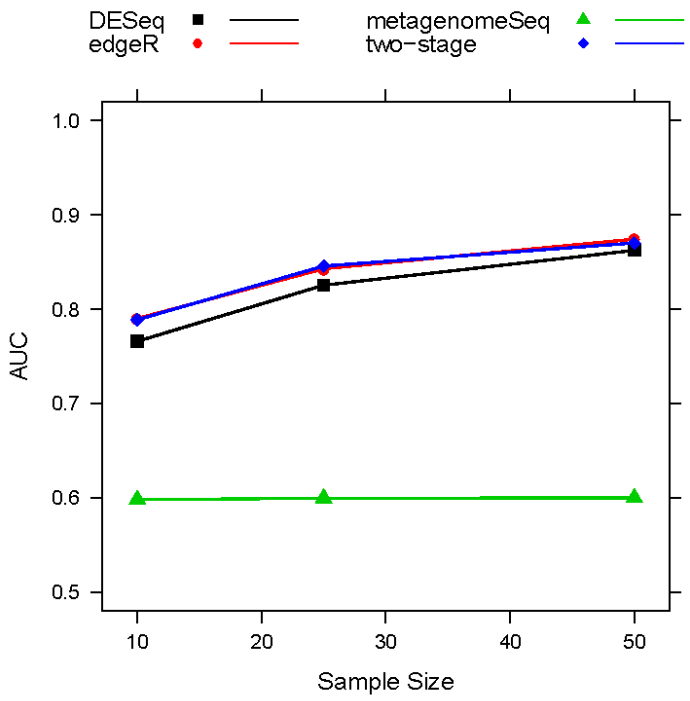

**Figure 4.** The AUC results for sample size of 10, 25, and 50 in the simulation setting in the experimental design 2.

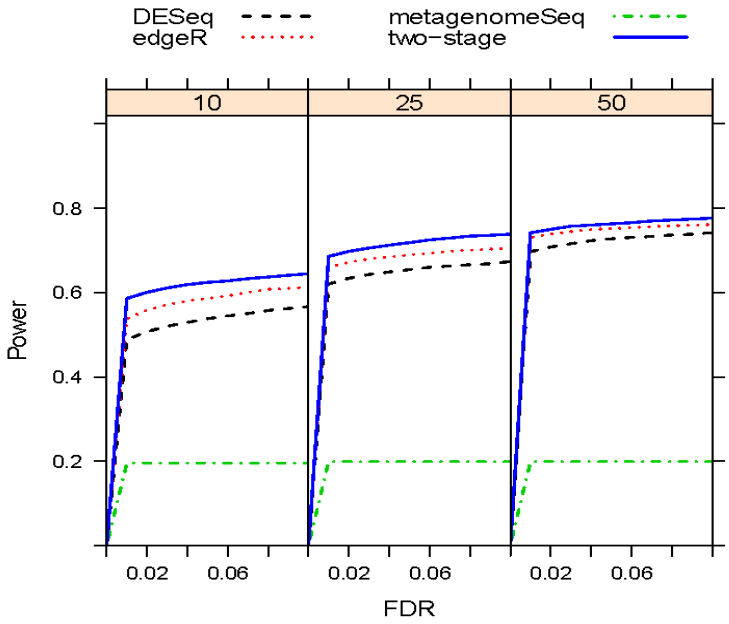


**Figure 5.** The power in detection of the true differentially abundant features obtained from each method for sample size of 10, 25, and 50 in the experimental design 2.

## 4 REAL DATA ANALYSIS

***Human gut data***

We applied our proposed method on human gut metagenomic data from 124 unrelated Danish and Spanish individuals in the Meta-HIT project (Qin et al., 2010) focusing on two human diseases, obesity and inflammatory bowel disease (IBD). The DNA sequences were aligned to the MetaHIT gene catalogue of 3.3 million genes to get the abundance of genes. The genes were annotated to the NCBI non-redundant Clusters of Orthologous Groups (COGs) database and this information was used to transform gene

abundance to COG abundances. Of the 124 individuals, 82 were labeled as lean [body mass index (BMI) < 30] and 42 were labeled as obese (BMI ≥ 30). Moreover, 25 individuals were diagnosed with IBD relative to 99 healthy individuals.

Differences based on our two-stage method with multiple comparison correction of FDR < 0.05 were observed between the lean and obese individuals in COG functional terms. Figure 6 displays the top 25 most significant functions whose abundance differs between the lean and obese groups. We discovered that two functions, including *Cytochrome c-type biogenesis factor* and *Cytochrome c-type biogenesis protein CcmE*, are involved in cytochrome c biogenesis. These functions are enriched in obese individuals. In 2013, it was reported that exogenous *cytochrome c* could be delivered as a potential anti-obesity drug for preventing diet-induced obesity (Hossen et al., 2013). This suggests that the differential abundance of *Cytochrome c biogenesis* between the obese and lean groups may be a contributor to obesity. An *Uncharacterized protein involved in cysteine biosynthesis* was also discovered by our method as differentially abundant function. This function is also found enriched in obese individuals, which is consistent to the finding of Elshorbagy et al. (2012) that *Plasma total cysteine* is independently associated with obesity and insulin resistance in Hispanic children and adolescents.

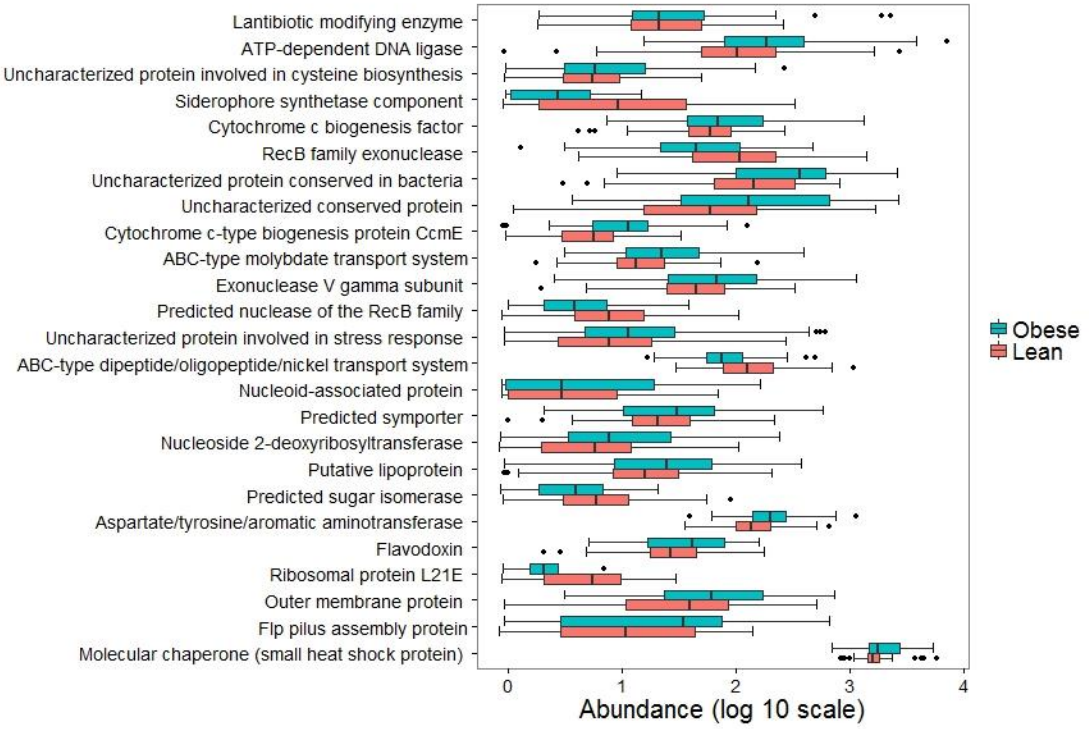


**Figure 6.** Differentially abundant COG functions (in log 10 scale) between lean (BMI<30) and obese individuals (BMI≥30).

Moreover, differences were also observed for IBD patients and healthy individuals in COG functional terms. Figure 7 shows the top 25 most significant differentially abundant functions. We found that *Bacteriophage capsid protein* is enriched in healthy individuals. The dysbiosis theory reviewed by Tamboli et al. (2004) states that an imbalance between putative "harmful" versus "protective" bacterial species might promote chronic intestinal inflammation. The difference of *Bacteriophage capsid protein* between healthy individuals and IBD patients may indicate the bacterial diversity and population change in IBD patients. Moreover, we found two *flagellar proteins* in the list of top 25 significant functions. IBD results from an aberrant and poorly understood

mucosal immune response to the microbiota. Lodes et al. (2004) conducted serological expression cloning to identify commensal bacterial protein that could contribute to the pathogenesis of IBD. The dominant antigen they identified were *flagellins*, molecules known to activate innate immunity via Toll-like receptor 5 (TLR5), and critical targets of the acquired immune system in host defense.

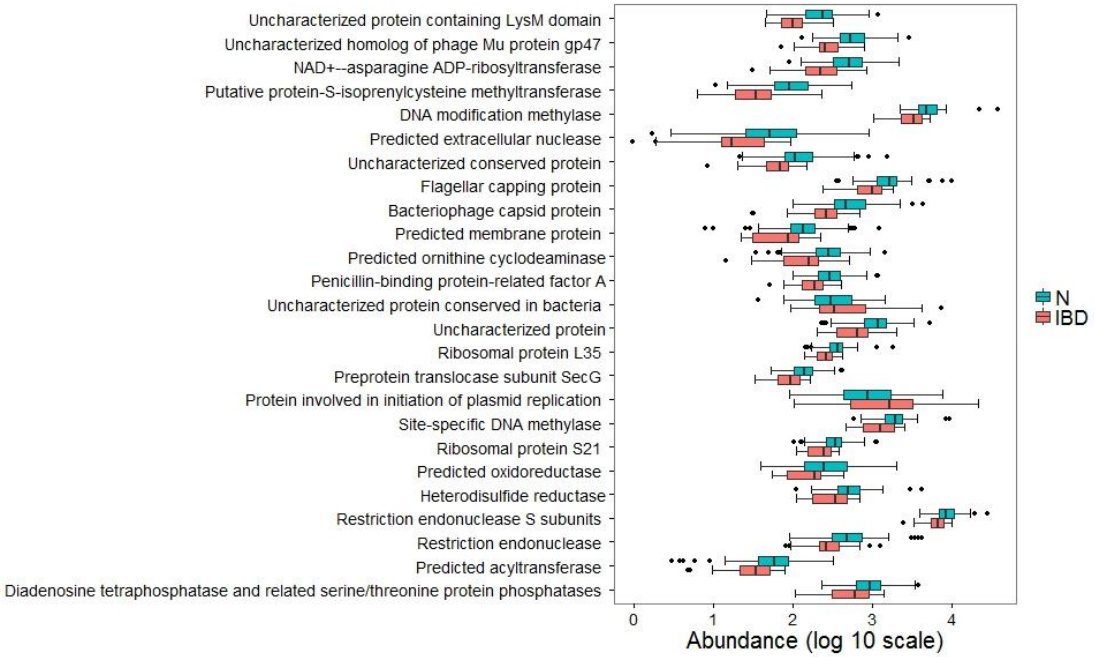


**Figure 7.** Differentially abundant COG functions (in log 10 scale) between IBD and healthy individuals.

### *Human mucus vs. saliva data*

We performed our proposed method on metagenomic shotgun sequence data in the HMP project (Qin et al., 2010) focusing on the functions of microbes in human health and disease through the characterization of microbial communities for two human body

sites: nasal mucus and oral saliva. Out of 42 samples, 30 samples are obtained from human nasal mucus microbial metagenomes and 12 samples from human oral saliva samples. The dataset is stored and maintained on MG-RAST (Supplementary Table S1).

Differentially functional abundances between human nasal mucus and human oral saliva were identified with multiple comparison correction of FDR < 0.05. Figure 8 shows the top 25 most significant differentially abundant functions. Five of them get involved in a biological process of phosphate metabolism and their abundances are more presented in microbial metagenomes of cystic fibrosis (CF) lung patients compared with microbial metagenomes of healthy human saliva individuals. These functions are *Pyrophosphate-energized proton pump (EC 3.6.1.1), Geranyltranstransferase (farnesyldiphosphate synthase) (EC 2.5.1.10), Geranylgeranyl pyrophosphate synthetase (EC 2.5.1.29), Fructose-bisphosphate aldolase class I (EC 4.1.2.13),* and *Maltose-6'-phosphate glucosidase (EC 3.2.1.122)*. Willner et al. (2009) conducted the first metagenomic study of DNA viral communities in the airways of CF diseased and non-diseased individuals and discovered that *Guanosine-5'-triphosphate, 3'-diphosphate pyrophosphatase* are over-representation in CF diseased compared to non-diseased individuals. Several studies, including Jain et al. (2006) and Raskin et al. (2007) discovered that these enzymes are linked to bacterial stringent response, bacterial virulence, antibiotic resistance, biofilm formation, quorum sensing, and phage induction in a variety of bacteria. These findings imply that a unique metagenomic environment of

the CF airway might contribute to functional adaptations, resulting in shifts in metabolic profiles (Willner et al., 2009).

Moreover, we found that *Putative peptidoglycan bound protein (LPXTG motif) Lmo0159 homolog* is enriched in mucus but very rare in saliva metagenomes. This finding is correspondent to the discovery of Quinn et al. (2014) which conducted an experiment to assess how CF lung microbes respond to the biochemistry of the lung environment by identifying pathways, obtained from KEGG classification hierarchy, whose presence enriched in microbial metagenomes of CF lung patients compared to healthy human saliva microbial metagenomes from the HMP. Quinn et al. (2014) reported that peptidoglycan biosynthesis pathway is enriched in human mucus metagenomes of CF lung patients, but rare in healthy human saliva individuals. Furthermore, out of the significant differentially abundant functions, we discovered that three functions, including *Glutamate formyltransferase, Formiminoglutamase (EC 3.5.3.8),* and *Aminobenzoyl-glutamate transport protein* are involved in glutamate protein and are enriched in human mucus. Our finding is also consistent to the findings discovered by Quinn et al. (2014), that D-glutamine and D-glutamate metabolism pathways are enriched in human mucus of CF lung patients compared to healthy human saliva. The results suggests that enrichment of those functions in human mucus of CF lung patients compared with healthy human saliva individuals may be a contributor to CF disease.

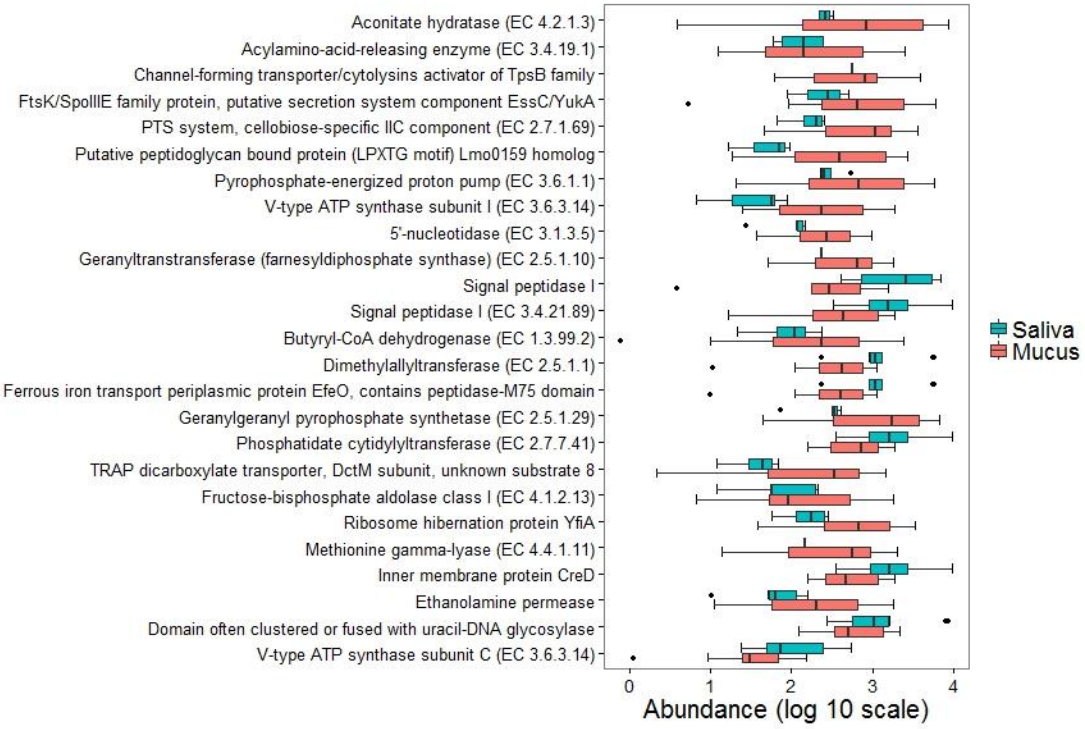


**Figure 8.** Differentially abundant functions (in log 10 scale) between human mucus and human saliva individuals.

## DISCUSSION

Currently, there has been an increasing interest in metagenomic projects with various applications. One typical aim is to assess whether and how two or more microbial communities differ. Comparing microbial genetic contents on the basis of functional features (e.g., pathways, subsystems, functional roles) obtained from different microbial communities with different phenotypes (e.g., diseased and healthy, or different treatments) enables us to identify the gnomic contents of microbes contributing to human

health and disease, which can in turn lead us to understand how the microbes affect human health.

We proposed a two-stage statistical procedure for sequentially selecting informative functional features and detecting differentially abundant functional features between two or more microbial communities/conditions. The proposed method accounts for the specific characteristics of metagenomic data, which are high-dimensional complex data sets consisting of a large proportion of zeros, non-negative counts with skewed distribution, and a large number of features, but limited number of samples. From the results of several simulations, we showed that our proposed method effectively select the informative functional features and efficiently detect the differentially abundant functional features between metagenomic datasets. Also, the simulation results showed that the proposed approach outperforms the other methods, which are widely used in biology and biomedicine, in most situations.

We also applied the proposed method on two real metagenomic datasets related to two human diseases. One of them is related to obesity and inflammatory bowel disease (IBD) and the other one is related to cystic fibrosis lung disease. Our findings are consistent with previous reports. Our method can be directly applied for comparison of more than two microbial samples. Therefore, our method can be applicable to more general situations such as multiple-condition comparison in human health related problems and comparison in multiple microbial communities under different ecological or biological conditions.

## ACKNOWLEDGEMENTS

*Funding*: This work was supported by National Science Foundation [DMS-1043080 to L.A. and H.J] and [DMS-1222592 to L.A. H.J], and partially supported by National Institutes of Health [P30 ES006694 to L.A.] and by The Cecil Miller Endowment at University of Arizona Foundation to N.P.

## REFERENCES

Allison B, Cui X, Page P, Sabripour M (2006) Microarray Data Analysis: from Disarray to Consolidation and Consensus. *Nat Rev Genet* **7**: 55–65.

Anders S and Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* **11**(10): R106.

Benjamini Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc*. B, **57**:289–300.

Bullard J, Purdom E, Hansen K, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**:94.

Cameron A, Trivedi P (1998) Regression Analysis of Count Data. Cambridge University Press.

Donoho D, Johnstone I (1994) Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, **81**:425-455.

Elshorbagy A, Valdivia-Garcia M, et al. (2012). The association of cysteine with obesity, inflammatory cytokines and insulin resistance in Hispanic children and adolescents. *PLoS One* **7**(9): e44166.

Friedman J, Hastie T, Hoeing H, Tibshirani R (2007) Pathwise Coordinate Optimization. *The Annals of Applied Statistics* **2**(1):302-332.

Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software January* **33**(1):1-22

Gilbert JA, Meyer F, Bailey MJ (2011) The Future of microbial metagenomics (or is ignorance bliss?). *ISME J* **5**(5): 777–779.

Hastie T, Tibshirani R, and Friedman J (2009). The Elements of Statistical Learning: Prediction, Inference and Data Mining. 2nd edition. Springer-Verlag, New York.

Hossen M, Kajimoto K, et al. (2013). Therapeutic assessment of cytochrome C for the prevention of obesity through endothelial cell-targeted nanoparticulate system. *Mol Ther* **21**(3): 533-541.

Hugenholtz P **(**2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol.* **3**: REVIEWS0003.

Hunter L (2004) Life and its molecules: a brief introduction, *AI Magazine*, **25**(1): 9-22.

Huson D, Auch A, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* **17**: 377–386.

Huson D, Richter D, Mitra S, Auch A, Schuster S (2009) Methods for comparative metagenomics. *BMC Bioinformatics* **10**(Suppl 1):S12.

Huson D, Mitra S, et al. (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res* **21**(9): 1552-1560.

Jain V, Kumar M, Chatterji D (2006) ppGpp: stringent response and survival. *J Microbiol* **44**: 1–10.

Kristiansson E, Hugenholtz P, Dalevi D (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* **25**(20): 2737–2738.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatics's guide to metagenomics. Microbiol. *Mol. Biol. Rev*., **72**(4):557.

Liu Z, Chen D, Sheng L, Liu A (2013) Class Prediction and Feature Selection with Linear Optimization for Metagenomic Count Data. *PLoS ONE* **8**(3):e53253.

Lodes M, Cong Y, et al. (2004). Bacterial flagellin is a dominant antigen in Crohn disease. *Journal of Clinical Investigation* **113**(9): 1296-1306.

Paulson J, Stine O, Bravo H, Pop M (2013) Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* **10**(12): 1200-1202.

Potrykus K, Cashel M (2008) (p)ppGpp: still magical? *Annu Rev Microbiol* **62**: 35–51.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**(7285):59–65.

Quinn RA, Lim YW, Maughan H, Conrad D, Rohwer F, Whiteson KL (2014) Biogeochemical forces shape the composition and physiology of polymicrobial communities in the cystic fibrosis lung. *mBio* **5**(2):e00956-13.

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* 14(9):R95.

Robinson M and Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**(R25).

Robinson M, McCarthy D, Smyth Gordon (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1): 139-140.

Rodriguez-Brito B, Rohwer F and Edwards R (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**(1):162.

Raskin, DM, Judson N, Mekalanos JJ (2007) Regulation of the stringent response is the essential function of the conserved bacterial G protein CgtA in Vibrio cholerae. *Proc. Natl Acad. Sci*. **104**:4636–4641.

Schloss P, Handelsman J (2006) Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl Environ Microbiol* **72**: 6773–6779.

Tamboli CP, Neut C, Desreumaux P, Colombel JF (2004) Dysbiosis as a prerequisite for IBD. *Gut* **53**:1057.

Turnbaugh P, Ley R, Hamady M, Fraser-Liggett C, Knight R, et al. (2007) The human microbiome project. *Nature **449***: 804–810.

Venables W, Ripley B (2002) Modern Applied Statistics with S. Springer-Verlag, New York, 4th edition.

White J, Nagarajan N, Pop M (2009) Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Comput Biol* **5**(4): e1000352.

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, et al. (2009) Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. *PLoS ONE* **4**(10): e7370.

Wooley J, Ye Y (2010) Metagenomics: Facts and artifacts, and computational challenges. *J of Comp Sci and Tech*. **25**(1): 71-81.

Rodriguez-Brito B, Rohwer F and Edwards R (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**(1):162.

Zhu J, Hastie T (2004). Classification of Expression Arrays by Penalized Logistic Regression. *Biostatistics,* **5**(3), 427-443.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B **67**:301–320.

## SUPPLEMENTARY

### Observing real metagenomic datasets

We examined the relationship between *means* and *variances* of feature abundances in real metagenomic datasets obtained from various environmental sources, including human gut, ocean, soil, and fresh water. Table S1 shows the sources for these metagenomic datasets. The metagenomic datasets are publicly available from MG-RAST

web server (http://metagenomics.anl.gov/). Figures S1-S4 show the scatter plots of variance vs. mean for the feature abundance under each situation.

**Table S1.** Basic information about the real metagenomic datasets used in this experimental study, and all of these datasets are publicly available.

| Metagenomic Dataset | Sample Condition | No. of samples | MG-RAST ID or website |
|---|---|---|---|
| Human | gut | 124 | http://www.sysbio.se/Fantom/ |
| Ocean | Sargasso Sea | 21 | 4449104.3;4494598.3;4494599.3;4494600.3; 4494602.3;4494603.3; 4494604.3;4494605.3; 4494606.3;4494607.3; 4494608.3;4494609.3; 4539503.3;4539504.3;4539506.3;4539507.3; 4539508.3;4539509.3; 4539511.3; 4539512.3; 4539513.3 |
| | Guanabara Bay | 30 | 4453375.3;4453376.3;4453379.3;4453380.3; 4453381.3;4453382.3; 4454501.3;4454502.3; 4454503.3;4454504.3; 4454505.3;4454506.3; 4454686.3;4454687.3;4454688.3;4454689.3; 4454691.3;4454692.3; 4454693.3;4454694.3; 4454695.3;4454696.3;4454697.3;4454698.3; 4454699.3;4454700.3;4454701.3;4454702.3; 4454703.3;4454704.3 |
| Soil | Desert | 9 | 4477803.3;4477805.3;4477872.3;4477873.3; 4477900.3;4477901.3; 4477902.3; 4477903.3; 4477904.3 |
| | Grassland | 13 | 4449249.3;4449252.3;4449255.3;4449256.3; 4449356.3;4449357.3; 4449359.3;4449360.3; 4449362.3;4449363.3; 4449364.3;4449365.3; 4449877.3 |
| Fresh water | | 23 | 4440090.3;4440144.4;4440145.4;4440324.3; 4440325.3;4440416.3; 4440417.3;4440419.3; 4440420.3;4440421.3; 4440426.3;4440427.3; 4440428.3;4440429.3;4440430.3;4440431.3; 4440432.3;4440433.3; 4440434.3;4440435.3; 4440436.3;4440437.3; 4440438.3 |

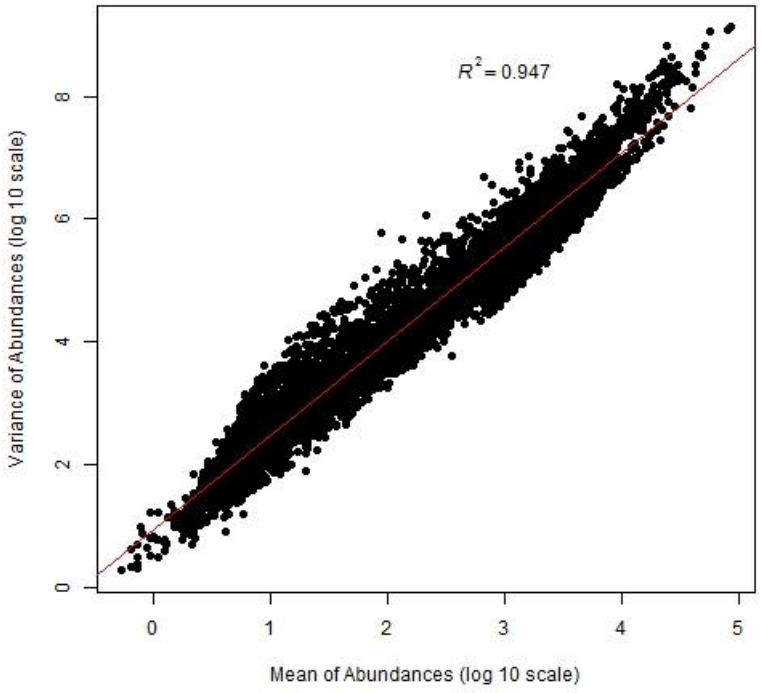


**Figure S1.** Scatter plot between variances vs. means of feature abundances (in log 10 scale) obtained from human gut metagenomic samples. The fitted linear relationship is: $E(\sigma^2) = 0.96 + 1.53 * \mu$, where $\mu$ and $\sigma^2$ are the mean and the variance of the feature abundance (in log 10 scale), respectively.

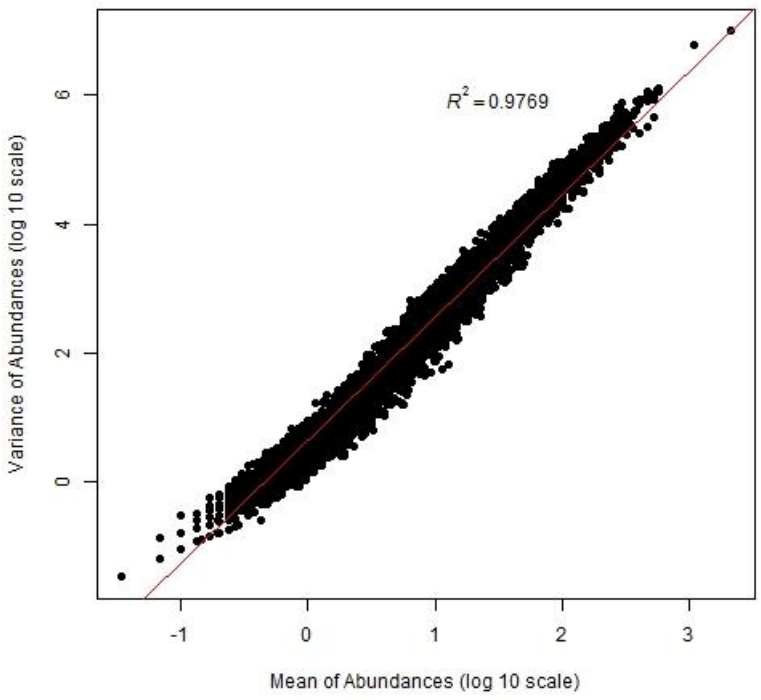


**Figure S2.** Scatter plot between variances vs. means of feature abundances (in log 10 scale) obtained from ocean samples. The fitted linear relationship is $E(\sigma^2) = 0.66 + 1.9 * \mu$.

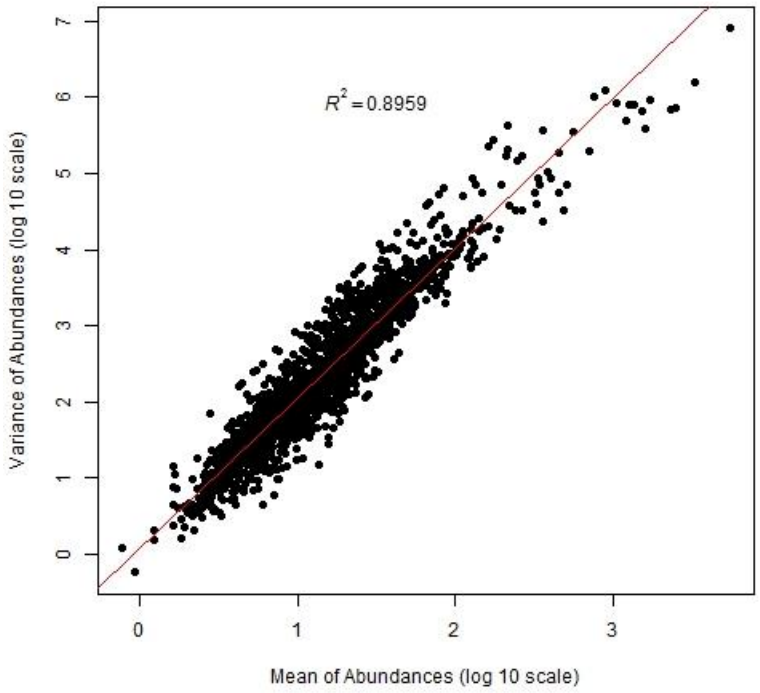


**Figure S3.** Scatter plot between variances vs. means of feature abundances (in log 10 scale) obtained from soil samples. The fitted linear relationship is $E(\sigma^2) = 0.09 + 1.97 * \mu$.

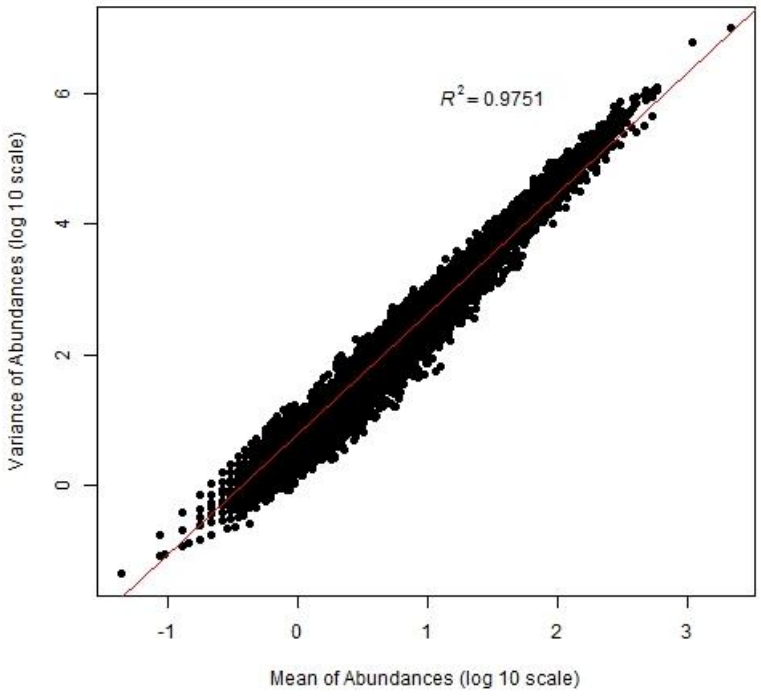


**Figure S4.** Scatter plot between variances vs. means of feature abundances (in log 10 scale) obtained from fresh water samples. The fitted linear relationship is $E(\sigma^2) = 0.79 + 1.85 * \mu$.

The model checking (results not shown) for all the above linear fittings indicates the simple linear pattern is adequate. The estimated regression function obtained from averaging the estimates of simple linear regression parameters $\beta_0$ and $\beta_1$ (S1-S4) is

$$E(\sigma^2) = 0.6 + 1.8\mu$$

where $\mu$ and $\sigma^2$ are the mean and the variance of the feature abundance (in log 10 scale), respectively.