

5-2017

STATISTICAL METHODS FOR ASSESSING STRUCTURAL CHANGE IN HUMAN & MICROBIAL GENOMES

Xuan Zhu

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations



Part of the [Investigative Techniques Commons](#)

Recommended Citation

Zhu, Xuan, "STATISTICAL METHODS FOR ASSESSING STRUCTURAL CHANGE IN HUMAN & MICROBIAL GENOMES" (2017). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 767.
https://digitalcommons.library.tmc.edu/utgsbs_dissertations/767

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

STATISTICAL METHODS FOR ASSESSING STRUCTURAL CHANGE IN HUMAN & MICROBIAL GENOMES

by

Xuan Zhu, BS

APPROVED:

Advisory Professor: Sanjay S. Shete, Ph.D.

Xuelin Huang, Ph.D.

Jeffrey T. Chang, Ph.D.

Jian Wang, Ph.D.

Marcos R. Estecio, Ph.D.

APPROVED:

Dean, The University of Texas
MD Anderson Cancer Center UTHHealth Graduate School of Biomedical Sciences

STATISTICAL METHODS FOR ASSESSING STRUCTURAL CHANGE IN HUMAN & MICROBIAL GENOMES

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Xuan Zhu, Bachelor of Science degree at The Ohio State University, Columbus

Houston, Texas

May, 2017

Copyright © 2017 Xuan Zhu

All rights reserved

DEDICATION

This thesis is dedicated to my family, my mentor, and my friends for the support and encouragement they provided.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my mentor, my advisory committee members, and professors in my graduate courses, who offered me invaluable advice and support in my Ph.D. training.

First and foremost, I would like to express my sincerest gratitude to my mentor and advisory committee chair, Dr. Sanjay S. Shete, for giving me the opportunity to work with him and continuously supporting me throughout my graduate research. I am truly grateful for Dr. Shete's continuous instruction, patience and encouragement that helped and inspired me during my research study in the department of biostatistics at the University of Texas MD Anderson Cancer Center. I am also particularly thankful to my co-advisor and friend, Dr. Jian Wang, for the care and openness she showed me whenever I had questions. I would also like to express my deepest gratitude and highest respect to Dr. Xuelin Huang, Dr. Jeffrey Chang, Dr. Jianhua Hu, Dr. Veerabhadran Baladandayuthapani, and Dr. Marcos R. Estecio, who supported me as members of my advisory committee and routinely gave me advice and inspiration for my research projects. I am especially grateful to Dr. Peter F. Thall for providing me guidance and insights in clinical trial study.

STATISTICAL METHODS FOR ASSESSING STRUCTURAL CHANGE IN HUMAN & MICROBIAL GENOMES

by

Xuan Zhu, BS

Academic Advisor: Sanjay S. Shete, Ph.D.

Genetic sequencing has been recognized as an effective approach to accurately address biological problems, such as clinical detection of disease, mutation discovery, and targeting specific biomarkers associated with complex diseases. Compared with conventional Sanger sequencing, next-generation sequencing costs much less due to massively parallel high-throughput sequencing. However, due to large numbers of short read sequences, the accuracy of high-throughput sequencing data remains a challenge in that the data obtained from next-generation sequencing often has higher error rates, which may impact downstream genomic analysis. Even if the downstream genomic analysis performs well, the quality of the result will still be impacted by the quality of the data. Before proceeding to downstream analysis, data quality assessment is necessary. Error rate estimation studies have been able to describe the quality of sequencing reads obtained from a sample. However, these studies may have limitations when sequencing a new genome or under a linearity assumption that the number of sequences containing errors increases linearly with the number of error-free read counts, which may not be available for all types of sequencing data. Therefore, it is necessary to estimate sequencing error rates in a more reliable way. In this dissertation, we proposed an empirical error

rate estimation approach that employs nonlinear statistical models of cubic smoothing splines and robust smoothing splines to analyze the association between the number of shadow counts and the number of error-free read counts. Traditional approaches to simulation when analyzing sequencing data may not reflect the real structure of the sequencing data. We also proposed a frequency-based simulation approach that mimics the real sequencing count framework and has more computational efficiency. Based on all the simulation scenarios tested, our proposed empirical error rate estimation approach provided more accurate estimations than the shadow regression approach. We also redefined the per-read error rate so that it is more flexible and provides more information according to the sequencing reads. The proposed empirical error rate estimation approach was applied to assess the sequencing error rates of bacteriophage PhiX DNA samples, a MicroArray Quality Control project, a mutation screening study, the Encyclopedia of DNA Elements project. The proposed empirical error rate estimation approach is free from the limitation of a linearity assumption between the number of shadow counts and the number of error-free read counts and demonstrates more accurate error rate estimations for next-generation short read sequencing data.

The errors of next-generation sequencing data discussed in the literature concern sequences with up to two bases that are different from error-free reads by substitution. In this thesis, we also extended the study of error rate estimations to different shadow scenarios, including varying the substitution shadows to be sequences in which only one base is different, only two bases are different, and up to two bases are different from error-free sequencing reads, and extending the investigation to deletion and insertion error rates. The deletion and insertion error rates are calculated differently from multiple approaches. For the extended investigation, both simulation studies and real data analyses were performed using empirical error rate estimation approaches

and shadow regression approaches. Under the simulation scenarios tested, the empirical error rate proved to be more accurate and resulted in less biased estimation for the deletion and insertion analysis.

In this dissertation, we also studied the human microbiome, which has been associated with complex diseases, for example, cardiovascular disease, diabetes, obesity, and specific cancers. To our knowledge, the existing literature has rarely discussed how to process raw human microbiome data and convert it into the format of downstream operational taxonomic units. We reviewed multiple statistical methods for processing, summarizing and analyzing microbiome data, and also provided detailed programming scripts about how to process human microbiome data into a downstream analysis format and assess alpha diversities, beta diversities, and the association between sample diversities and the outcome of interest. For illustration, the statistical approaches were also applied to analyze the foregut microbiome in esophageal adenocarcinoma.

For future directions of research, we provided a potentially effective strategy for analyzing longitudinal human microbiome data associated with complex diseases by using a Bayesian network approach. The Bayesian graphical model approach accounts for interaction with the microbiome and captures the common structure of change within the microbiome, which provides insights to predict the change in the composition of the microbiome over time and find specific microbiome taxa that may be associated with complex diseases.

TABLE OF CONTENTS

STATISTICAL METHODS FOR ASSESSING STRUCTURAL CHANGE IN HUMAN & MICROBIAL GENOMES	1
ACKNOWLEDGEMENTS	v
STATISTICAL METHODS FOR ASSESSING STRUCTURAL CHANGE IN HUMAN & MICROBIAL GENOMES	vi
Academic Advisor: Sanjay S. Shete, Ph.D.	vi
TABLE OF CONTENTS	ix
List of Figures	xi
List of Tables	xii
Chapter 1	1
Introduction	1
1.1 Sanger Sequencing	2
1.2 Next-Generation Sequencing	4
1.3 RNA Sequencing	7
1.4 Human Microbiome	9
1.4.1 16S ribosomal RNA Sequencing	11
1.4.2 16s rRNA versus Whole-Genome Shotgun Sequencing	15
1.5 Motivation and Rationale of the Thesis	16
Chapter 2	20
Empirical estimation of substitution sequencing error rates using smoothing splines	20
2.1 Background	21
2.2 Methods	24
2.2.1 Shadow Regression Overview	24
2.2.2 Empirical Per-Read Error Rate with Smoothing Splines	25
2.2.3 Next-Generation Sequencing Data	28
2.3 Simulation Approaches	30
2.4.1 Simulation Results	40
2.4.2 Next-Generation Sequencing Data Analysis Results	49

2.5 Conclusion	54
2.6 Discussion.....	54
Chapter 3	60
Extension Studies of Empirical Error Rate Estimation Approach for Substitution, Deletion, and Insertion Sequencing Errors.....	60
3.1 Introduction	60
3.2 Method	62
3.3 Result.....	65
3.4 Conclusion.....	81
Chapter 4.....	82
Statistical Approaches for Processing and Analyzing Microbiome Data	82
4. 1. Introduction	82
4.2 Methods in Human Microbiome Analysis.....	84
4.2.1 Next-generation Sequencing Methods for the Human Microbiome	84
4.2.1.1 Foregut Esophageal Adenocarcinoma Study.....	85
4.2.1.2 Processing Raw Microbiome Sequencing Data.....	86
4.3 Diversity Measures of the Microbiome	90
4.4 Association between Microbiome Diversity and Outcome	101
4.5 Challenges in Analyzing Microbiome Data	105
4.6. Processing Methods and Results	108
Chapter 5	139
Conclusions and Future Directions.....	139
5.1 Conclusions	140
5.2 Future Directions.....	143
References.....	149
VITA.....	181

List of Figures

Figure 1. Sample SRR037440 from the MAQC brain experiment 2 data set and corresponding simulated data using frequency-based and Wang et al. simulation approaches	34
Figure 2. Read counts for the SRR032577 sample in the mutation screening re-sequencing study	35
Figure 3. Sample sequencing data from MAQC, Mutation Screening Re-Sequencing, ENCODE, and PhiX DNA data sets	38
Figure 4. Flowchart of basic steps in handling raw microbiome sequencing data	89
Figure 5. Alpha diversity plots for the foregut esophageal adenocarcinoma dataset	121

List of Tables

Table 1. Comparison of the Wang et al. and proposed frequency-based simulation approaches using sample SRR037440*	37
Table 2. Median error rates in MAQC data using shadow linear regression and smoothing spline approaches*	42
Table 3. Median error rates in mutation screening data using shadow linear regression and smoothing spline approaches*	44
Table 4. Median error rates in ENCODE data using shadow linear regression and smoothing spline approaches*	46
Table 5. MedianSimulation error rates in PhiX DNA data using shadow linear regression and smoothing spline approaches*	48
Table 6. Error rates in real MAQC data using shadow linear regression and smoothing spline approaches.....	50
Table 7. Error rates in real mutation screening data using shadow linear regression and smoothing spline approaches	51
Table 8. Error rates in real ENCODE data using shadow linear regression and smoothing spline approaches.....	52
Table 9. Error rates in real PhiX DNA data using shadow linear regression and smoothing spline approaches.....	53
Table 10 Simulation studies for deletion by testing shadows to be only 1 base, only 2 bases, and up to 2 bases deletion.	67
Table 11 Simulation studies for insertion by testing shadow to be only 1 base insertion, only 2 bases insertion, and up to 2 bases insertion.	70

Table 12. Real data application for substitution by varying shadow to be only 1 base different, only 2 bases different, only 3 bases different, up to 2 bases different, and up to 3 bases different.	74
Table 13. Real data application for deletion by varying shadow to be only 1 base deletion, only 2 bases deletion, and up to 2 bases deletion.....	77
Table 14. Real data application for insertion by varying shadow to be only 1 base insertion, only 2 bases insertion, and up to 2 bases insertion.	80
Table 16. Bray-Curtis dissimilarity index using the foregut esophageal adenocarcinoma dataset	123
Table 17. Cao dissimilarity index using the foregut esophageal adenocarcinoma dataset	125
Table 18. Unweighted UniFrac measure for the foregut esophageal adenocarcinoma dataset ..	127
Table 19. Weighted UniFrac measure for the foregut esophageal adenocarcinoma dataset	129
Table 20. Normalized unweighted UniFrac measure for the foregut esophageal adenocarcinoma dataset	131
Table 21. Normalized weighted UniFrac measure for the foregut esophageal adenocarcinoma dataset	133

Chapter 1

Introduction

Biomedical research has identified associations between genetic variants and complex human diseases. Studies on genetic sequences have provided invaluable insights into complex diseases, suggested novel therapeutic strategies, and contributed to effective personalized treatment regimens and disease prevention strategies (*1, 2*). Downstream genomic sequencing analyses have been conducted to identify sequences associated with complex diseases using multiple sequencing methods. However, it remains a challenge to accurately assess the quality of the sequencing data each time a sample is obtained.

The percentage of reads mapped has been widely used as a quality indicator; however, that does not directly address the fundamental question of how much error is present in the samples to be studied in downstream genomic analyses (*3, 4*). Estimating the error rate typically requires using a reference genome, which may not be appropriate when sequencing a new genome when such a reference genome has not been established.

In this thesis, we developed statistical models to estimate sequencing error rates, including those of substitution, deletion, and insertion in next-generation sequencing data. We also studied and reviewed the statistical approaches used to analyze human microbiome data, and provided a future direction for analyzing longitudinal human microbiome data. The study on the human microbiome may provide novel strategies about effective treatment and prevention of complex diseases.

In Chapter 1, we introduce multiple sequencing approaches, including Sanger sequencing, next-generation sequencing, DNA sequencing, RNA sequencing, and 16S rRNA sequencing of the human microbiome.

1.1 Sanger Sequencing

Automated Sanger sequencing represents the first generation of DNA sequencing (5), and has been widely used in a diverse range of areas to solve various biological problems and identify clinically significant DNA variants associated with disease (6). For example, Sanger sequencing has been used to identify an appropriate treatment strategy that inhibits epidermal growth factor receptor tyrosine kinase for specific patients with non–small-cell lung cancer (6, 7). Sanger sequencing has been used so widely to evaluate non-syndromic hearing loss that most genetic testing strategies for this disorder depend on the gene-specific Sanger sequencing technique (8).

Developed by Sanger and colleagues (9), the original sequencing technology used a chain-termination sequencing approach to determine nucleotide sequences in DNA. Novel technologic developments have been applied to improve Sanger sequencing, resulting in the modern form of Sanger sequencing, which detects fluorescently labeled nucleotide sequences by applying automated sequencing instruments (6). In Sanger sequencing, a DNA sample is sheared into fragments, subcloned into different vectors, and then amplified in bacterial or yeast hosts (10). The amplified DNA is then isolated and sequenced with the Sanger chain termination approach.

Sanger sequencing has a low error rate and was long considered to be the “gold standard” for sequencing genetic mutations (11, 12). Sanger sequencing can achieve up to about 1,000 bp, with

per-base “raw” accuracies as high as 99.999%, which makes it the most accurate sequencing approach currently available (5). Sanger sequencing can read DNA fragments that are approximately 500 bp to 1kb in length. Performing traditional Sanger sequencing requires moderate expertise; whereas sequencing approaches such as next-generation sequencing require a high level of expertise (6).

Despite the advantage of accuracy and strong availability in Sanger sequencing, there are technical limitations in this first-generation sequencing technique. One major limitation involves the limited throughput of the DNA sequence that can be read with each sequencing reaction (13). Traditional Sanger sequencing is generally restricted to one gene at a time; hence it does not easily sequence hundreds of genes in a sample. In addition, traditional Sanger sequencing cannot detect deletions, translocations, or gene copy number alterations. (6) Sanger sequencing requires electrophoretic separation of DNA fragments for DNA sequence reading, which leads to the primary bottleneck for throughput data in this traditional sequencing approach (13). The process of Sanger sequencing is lengthy and requires labor-intensive cloning-based amplification that also limits its application for high-throughput genome sequencing (10). Another limitation is the high cost of traditional Sanger sequencing. With a throughput of Sanger sequencing of approximately 115 kb per day (1,000 bp), the current cost to sequence the entire human genome using Sanger sequencing is estimated to be up to 30 million U.S. dollars. In addition, using one single machine, this task will take approximately 60 years. (13)

These limitations of Sanger sequencing restrict its use to mainly the generation of reference genome sequences of many species, including those of humans. (10)

1.2 Next-Generation Sequencing

New sequencing approaches have been emerging that overcome the limitations of Sanger sequencing. Next-generation sequencing technology has played an essential role in biological research. Next-generation sequencing offers radically fast sequencing capability and produces large amounts of DNA, which has significantly driven down the cost of genetic sequencing (*13*). For example, one typical next-generation sequencing procedure can sequence millions of short reads at the same time and provide high-throughput of parallel DNA sequences in a single reaction.

In general, next-generation sequencing uses double-stranded DNA as the initial input material, although the technology can accommodate multiple types of starting materials. Even though the source of starting material may vary, the data are then converted into a sequencing library in a process that uses common fragmentation steps, size selection, and adapter ligation (*13*). As the sequencing material comes into the fragmentation and size selection stage, DNA templates are broken into smaller fragments for further DNA-pair synthesizing, and adapter ligations are then added with synthetic DNA to the ends of the library fragments that serve as primers for downstream amplification and sequencing reactions (*13*). In the next-generation technique, the sequencing library is either sequenced directly or is amplified and then is sequenced (*13*). DNA sequencing represents a single format, but projects a wide variety of biological phenomena through the analysis of high-throughput sequencing data (*5, 13*).

Compared to other sequencing techniques, next-generation sequencing has significantly decreased the per-base sequencing cost and increased the sequencing volumes, which always

leads to high-throughput genome-wide sequencing data. The next-generation sequencing approach has a number of appealing features, including the high-throughput capacity of parallel sequencing and the ability to simultaneously detect multiple genetic events, including deletions, insertions, copy number alterations, translocations, and exome-wide base substitutions (especially known “hot-spot mutations”) in genes associated with disease susceptibility (6). Moreover, compared with traditional Sanger sequencing, next-generation sequencing is not restricted to sequencing a single gene at a time and can generate millions of sequencing reads simultaneously using a surface area of a reasonable size (5, 6). With the advent of high-throughput next-generation sequencing, genomic investigations have expanded to include a more diverse range of genomic alteration phenomena, which has opened entirely new areas of biological inquiry, such as single-nucleotide polymorphisms (14), epigenetic events (15), copy number variants (16), differential expression (17), alternative splicing, the characterization of ecological diversity, and the identification of unknown etiologic agents (6, 18-20). In addition, next-generation sequencing platforms have enabled researchers to directly sample the nuclear genomes obtained from ancient remains, including those of a cave bear (21), mammoth (22), and Neanderthal (18, 23, 24). Next-generation sequencing has widened the scope of metagenomic analysis to environmentally derived samples (18).

Despite the advantages of next-generation sequencing, the increased capacity of high-throughput parallel sequencing comes at the cost of read length and sequencing accuracy (13). The most prominent shortcomings in next-generation sequencing are the short sequencing read lengths and raw accuracy, which has affected how the reads are utilized in bioinformatic analyses and downstream analyses (5, 18). In addition, researchers face the computational challenge of mapping the sequenced reads against a reference genome to independently verify the alignment

of the reads of interest. Moreover, for comprehensive resequencing of short reads, a much higher read coverage or sampling depth is required to cover the reference sequence at adequate depth and ensure a low gap size in a valid phenotype-genotype association study using the next-generation technique (18). There is a great need for new statistical approaches to use in the generation and analysis of next-generation sequencing since traditional methods have shortcomings when measuring the quality of sequencing data. The commonly used per-base quality score, which counts the probability of a called base in the read being the true sequence base, is quite raw and inaccurate. These misaligned reads and inaccurate quality scores, which propagate into genotyping and single-nucleotide polymorphism (SNP) discovery, have become general and acute problems in many centers that use rapidly evolving experimental processing pipelines, including the 1000 Genomes Project. Even if well-mapped and calibrated reads are given, caution should be taken when statistical modeling is used to resolve the case of simple SNPs, not to mention the more complicated variations such as multi-nucleotide substitutions, insertions, deletions, inversions, and copy number variation. (25)

Next-generation platforms are affecting a complete paradigm shift, including in the organization of large-scale data production, downstream bioinformatics, information technology, computational data storage, and support of laboratory information management infrastructure (18). Existing data analysis pipelines that use next-generation data and the accompanying algorithms must be adjusted to accommodate short read sequencing data. For example, new algorithms and data visualization interfaces are being devised and modified to satisfy the demands in next-generation sequencing (18).

1.3 RNA Sequencing

Another application of next-generation sequencing is RNA sequencing, which uses the next-generation sequencing approach to sequence, map, and quantify a population of transcripts (10, 20, 26). RNA sequencing transcriptomes can be profiled with deep sequencing technologies (27). Deep sequencing technologies for RNA sequencing are invaluable for detecting RNA expression levels and identifying changes in RNA structure (28). At the time of sequencing, a general procedure is to convert a region of RNA to cDNA fragments with adaptors attached to either one or both ends of the read (29). After the initial sequencing, short sequences, ranging from 30 to 400 bases, can be generated in a high-throughput manner (29).

RNA sequencing technology provides a precise measurement of the levels of transcripts and isoforms (27). The transcriptome refers to the complete transcript set in a cell, and the corresponding quantity can be synthesized and then measured in a specific developmental stage or condition via RNA sequencing (27). With RNA sequencing, it is feasible to distinguish specific genomic locations of transcription boundaries in the fairly high resolution of a single base pair as well as the rare variations in transcription regions (29-31). In addition, mRNA expression studies that use RNA sequencing have been utilized in microarrays or quantitative polymerase chain reactions (18). The identification of epigenetic events by investigating changes of RNA expression in RNA sequencing can be used to reveal historical environmental exposures and identify associated impacts in disease (28). Another big contribution of RNA sequencing is to address the heritability component of both common and rare variants, and measure the interactions between the genome and the environment in biological studies (28). RNA sequencing technology has provided unprecedented insights into the transcriptional complexities

of a variety of organisms, including yeast (32), mice (26), Arabidopsis (33), and humans (19, 20). RNA sequencing has revolutionized the analysis of eukaryotic transcriptomes and has been applied to the study of *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, and mouse and human cells (27, 29, 30, 34-37). Benefits of implementing RNA sequencing include a potentially unlimited dynamic range of expression, better sensitivity in genetic sequencing, greater capacity to discriminate regions of high sequence identity, and the ability to profile transcription without prior assumptions on which genomic regions are expressed (38). The analysis of RNA sequencing has generated an unprecedented global view of the transcriptome and the corresponding organization and grouping for species and cell types (27). Compared to DNA sequencing, RNA sequencing provides a lower background signal and does not require an upper limit of quantification that correlates with the number of sequencing reads that would be obtained in the sequencing procedure (27). RNA sequencing also facilitates learning the dynamics of the transcriptome across different tissues or conditions without data normalization (27, 30, 36). For example, researchers have utilized RNA sequencing to accurately monitor gene expression in yeast vegetative growth (34), yeast meiosis (35), mouse embryonic stem cell differentiation (30), and to track and measure the differences in gene expression between various tissues (27, 36).

After obtaining sequencing data, the most urgent challenge arises from mapping the RNA sequencing data to the reference genome, which is regarded as much harder than mapping the data derived from DNA due to the high-throughput sequencing (38). RNA sequencing brings additional challenges of how to efficiently store, retrieve, and process large volumes of sequencing data and how to reduce sequencing errors and remove low-quality sequencing reads (27). Typically, the larger the set of sequencing genomic data, the more complicated the

statistical modeling of the transcriptome. This occurs because sufficient sequencing depth must be controlled to achieve adequate coverage of such a large volume of data (27). Improper control will result in complicated, error-prone estimations of the transcriptome coverage since it is hard to know the usual number and level of different transcript isoforms, and the activity of transcription can vary across the genome (27). For example, when dealing with rare RNA isoforms, it becomes even harder to use RNA sequencing to target these complicated transcriptomes to monitor the changes in gene expression levels across the whole genome (27).

1.4 Human Microbiome

Another application of next-generation sequencing technology is the sequencing of the human microbiome (18). The advantages offered by next-generation sequencing techniques have initiated a revolution in our understanding of how the human microbiome affects a human health and diseases (18). Recent technologic advances have enabled the deep sequencing and analysis of the composition of the microbiome (39). The microbiome refers to the entire set of microorganisms living within a given habitat and the corresponding genomes and surrounding environmental conditions (40). The metagenome consists of the complex set of genetic sequences of the microbiome plus human DNA sequences (41).

The human microbiome interacts dynamically with our daily environment and is essential to human health (42). Studies have suggested that the human microbiome plays a central role in an individual's nutrition, immunity, and protection from various pathogens within the body (41, 43,

44). The microbiome functions within a complicated microbial community and processes energy and materials for the human body (45, 46). According to the Human Microbiome Project, the majority of microbial species in humans have never been isolated or cultured due to our inability to reproduce the necessary growth conditions in the lab (47).

The human body is inhabited by a complicated collection of microbes that includes bacteria, fungi, viruses and eukaryotes that live symbiotically within several highly specific environments that may vary according to the individual's health status (18, 39, 42). On average, each human hosts a microbiome that consists of 10-100 trillion symbiotic microbial cells, which are almost 10 times the total number of human somatic and germ cells in the body (42, 45, 46). The composition of the microbiome varies according to the individual's genotype, health, diet, lifestyle, social interactions, early use of antibiotic therapy, and exposure to environmental chemicals (45, 48). The variation of bacterial species within the microbiome is significantly different between individuals, and the corresponding composition in the microbiome impacts its habitat (41, 44). The composition and diversity of the microbiome may be further impacted by the mode of birth delivery, and hospitalizations (41). For example, dietary habits may impact the phylogenetic diversity and function of the human microbiome in disease risk and penetrance (41).

In addition, the composition of the microbiome undergoes dynamic change in the human body over time. For example, the intestinal microbiome changes dramatically in composition throughout infancy and childhood (41). A study showed that the composition of the gut microbiome is significantly different between adolescents, children, and adults, which implies that the composition of the human microbiome evolves during the periods of childhood and

adolescence (41). Therefore, the study on longitudinal data in humans may provide essential insights and further understanding of the fundamental interactive structure of the human microbiome and its association with complex diseases.

The understanding and recognition of the importance of the human microbiome has not been fully accomplished due to the complicated and dynamic nature of the microbiome and the limited sequencing analysis related to the microbiome (39). Studies of the human microbiome may provide diagnostic solutions for certain diseases and play a role in standard clinical practice in the future (39).

1.4.1 16S ribosomal RNA Sequencing

One major application in the analysis of the human microbiome using next-generation sequencing methods refers to 16S ribosomal RNA (rRNA) sequencing, which identifies and compares bacteria within a given sample (49)[49]. The 16S rRNAs are the genes encoding from a ribosomal RNA molecule, which are conserved in both bacteria and archaea (40). The 16S rRNA sequencing procedure allows for the identification of species from a complicated mixture of the microbiomic community (40). The terms “16S rRNA gene” and “16S rDNA gene” have been used interchangeably. The current Application Security Manager (ASM) policy uses the term “16S rRNA” (50). The 16S rRNA sequences are often between 500 bp and 1550 bp in length(50). Universal primers are utilized that are complementary to conserved regions. The sequences in the variable region are utilized as comparative taxonomy (50).

With the rapid development of sequencing technology, 16S rRNAs have been widely used as the primary tool to investigate taxonomic assignment and phylogenetic trees (51). For instance, 16S rRNA has been commonly used for phylogenetic study, taxonomic classification, and to infer microbiome diversity of clinical or environmental samples (51).

One prominent advantage that makes 16S rRNA the primary tool in microbiome sequencing is that 16S rRNA is universal in bacterial. 16S rRNA is present in at least one copy over all the bacterial genome, which makes the comparison and measurement in the same window for all the bacteria. Another advantage is the stability in 16S rRNA that the corresponding function in 16S rRNA does not vary over time. In addition, the 16S rRNA genes are adequately large for informatics purposes, and the conserved regions are available for simple sample identification. 16S rRNA sequencing provides reliable information on the bacterial family or species (50-52) .

16S rRNA sequencing provides rapid, robust, reproducible, and accurate identification measures for various bacteria, which may lead to clinical improvements, such as the discovery of novel pathogens and identifying noncultured bacteria (50). Studies that analyze 16S rRNA gene sequencing contribute to further understanding of clinical microbiology and associated infectious diseases, enables accurate identification on the level of the microbiome species, and clarifies the associated clinical importance (50).

Despite the advantages associated with using 16S rRNA sequencing, a downside of the 16S rRNA technique may refer to the difficulty in specifying the one-to-one correspondence. As there are more distinct taxonomy than the well-known ones that have names or phenotypic descriptions, it is challenging to assign names to each taxon in a meaningful way (50). Another

downside may refer to the high expertise requirement for technical skills and the high cost of 16S rRNA sequencing (50).

Currently, there are two established approaches for utilizing 16S rRNA to measure and quantify the microbiome and assign sequences to different microbiome groups at different taxonomic levels. These two established approaches are the phylotypes approach and the operational taxonomic units approach (40). The phylotypes approach depends on the comparison with reference sequences and assignment into taxonomic bins. The phylotypes approach searches public databases, such as the Ribosomal Database Project (53), the SILVA database (54), or the Greengenes database (40, 55). The major benefit of using the phylotypes approach is that it enables sequencing classification based on the previously characterized and often clinically related microbiome (40). A limitation is that this approach lacks a coherent definition of the reference sequences for the bacterial species for the whole taxonomic lineage, which makes it impossible to consistently define bacterial taxa (40). There are also conflicts in several widely used taxonomies regarding taxonomic lineages (40). The second approach to measuring microbiome communities refers to operational taxonomic units (OTUs), where 16S rRNA gene sequences are clustered into OTUs at a pre-specified similarity threshold (usually the threshold is defined to be 97%), which are equivalent to species-level clusters (40). The approach of measuring the OTUs overcomes the limitations in the phylotypes approach, which lacks a coherent definition for all taxonomic lineages. Initially, the OTUs measuring approach calculates the taxonomic unit independently from a reference database. Sequences that can be mapped to known bacteria are assigned to the known taxonomy, while sequences that do not match the known taxonomy are clustered based on similarity (40). A table about abundance within the

microbiome can be generated and used for downstream statistical analysis with species counts or compositional structure.

In the use of 16S rRNA for microbiome studies, a harmonious set of guidelines and measures for the interpretation of sequencing data is necessary in order to accurately compare the results between different studies (52). Standard 16S rRNA gene sequencing downstream analysis includes microbiome alpha diversity, beta diversity, and an association study. The diversity is an estimate to characterize and quantify the structure of a microbiome community based on measuring information, including the number of species (richness) and relative abundance (evenness) (40). The alpha diversity estimates the diversity within a sample, such as the Shannon index (56), Simpson index (57), and Chao1 (58) estimators. Furthermore, the beta diversity estimates the between-sample habitat diversity, which describes the distance or similarity between samples (40). Some widely used beta diversity measures include Bray-Curtis (59), Cao dissimilarity index (60), and UniFrac (61).

As adapting 16S rRNA gene sequencing as a tool in species identification is still a relatively new undertaking in most clinical laboratories, the standard measuring approaches will continue to evolve over time (52). Despite the accuracy in 16S rRNA, this type of analysis has not been widely implemented in large clinical and research laboratories due to technical and cost considerations. Thus, further studies in 16S rRNA sequencing may depend on translating 16S rRNA sequencing information into convenient biochemical testing schemes and making it available in a routine analysis format (50).

1.4.2 16s rRNA versus Whole-Genome Shotgun Sequencing

16S rRNA has been widely used to study species composition in bacterial communities (62). However, 16S rRNA sequencing might be biased when unequal amplification occurs in the sequencing process and a direct genetic identification is lacking (62, 63).

An alternative method for studying bacterial composition and diversity is whole-genome shotgun sequencing, which utilizes a sequencing technique with random primers to sequence overlapping regions in a genome (62, 63). In whole-genome shotgun sequencing, DNA sequences are extracted from all cells in a community and subsequently sheared into small fragments that are independently sequenced. The small fragments are then aligned to various genomic locations (64). Whole-genome shotgun sequencing can be used to sequence the genome over all microorganisms in a sample, whereas the domain of 16S rRNA is restricted to bacteria and archaea (65).

The major advantage to whole-genome shotgun sequencing is the capacity to identify taxa more accurately at the species level (63). Studies have shown that whole-genome shotgun sequencing can detect a more specific taxonomic and functional classification of sequencing reads (65, 66). Whole-genome shotgun sequencing also allows for the simultaneous study of archaea, viruses, virophages, and eukaryotes (66, 67).

Despite its benefits, whole-genome sequencing presents several challenges. The shotgun metagenomic sequencing is more expensive and usually not deep enough to detect rare species within the microbiome (62, 63). In addition, shotgun sequencing requires more extensive data analysis and sequencing a genome with high coverage in order to identify and understand the

taxa in bacteria (63). Moreover, the relative abundances obtained from shotgun sequencing vary significantly depending on the DNA extraction and sequencing protocol (62). When the genome database is limited, it may miss taxa due to unassigned sequences in the whole-genome shotgun sequencing approach (68). As the metagenomic dataset is relatively complicated and large, it may also be difficult to determine from which read the genome was derived (64).

16S rRNA sequencing has been the most commonly employed approach to study bacterial microbiomes (63). Previous studies revealed that 16S rRNA sequencing can profile bacterial communities in greater detail than shotgun sequencing. In addition, 16S rRNA sequencing can better identify species with low abundance compared with shotgun sequencing (62). Moreover, 16S rRNA sequencing is cost-effective, and allows data analysis to be performed on established pipelines (63). There is a large number of archived datasets available for reference in 16S rRNA sequencing (63). Studies have also shown that it may be more appropriate to use 16S rRNA sequencing for the analysis of a large number of samples, such as multiple patients and longitudinal studies (65).

1.5 Motivation and Rationale of the Thesis

Next-generation sequencing has been used to address a wide variety of biological investigations, for example, quantification of gene expression, microRNA profiling study, genome-wide association study of protein–DNA interactions, and polymorphism and mutation discovery (4, 5, 69-75). Next-generation sequencing applied to the clinical practice of medicine has laid the

foundation of personalized genome-based medicine, which is enhancing the accuracy and efficiency of disease diagnosis and treatment (*13*).

Due to the high capacity of parallel sequencing and much smaller reaction volume in next-generation sequencing, large-scale research projects have quickly come to depend entirely on next-generation sequencing because of the accompanying advantages in cost and sequencing capacity, and the practical aspects of implementation (*4, 5, 69, 74*).

However, the quantity of large number of short reads and quality of the data in next-generation sequencing has posed challenges for biostatistics and bioinformatics in terms of sequencing quality scoring (*5*). The quality and quantity of next-generation sequencing ultimately determines the comprehensiveness and accuracy of the downstream statistical analysis (*13*).

Qualitatively, the platforms of next-generation sequencing provide confidence scores for each individual base call, and the individual base calling error rates vary based on different sequencing platforms (*13*). According to a study on the chemistry of next-generation reactions, the initial portion of each sequencing read is generally more accurate than the latter portion (*13*). Quantitatively, the amount of sequencing data is assessed by the sequencing coverage, which refers to the average sequencing time for a base pair in the experiment (*13*).

Producing a large number of short reads at the same time leads to difficulty correctly and fully assembling the sequencing reads in next-generation sequencing (*69*). The quality of the sequencing data directly impacts the downstream genomic analysis (*69, 74, 76*). Before performing downstream genomic analysis, it is necessary to have quality assessment to ensure the sequencing data do not negatively impact the downstream genomic analysis (*69, 77*). The

traditional method that uses a reference to estimate the sequencing error rate depends on a well-established reference, which has limitations when sequencing new data, such as the microbiome or the genome of a new species (69, 78). A shadow regression approach that uses a reference-free method and is applicable to multiple datasets, including mRNA sequencing data and DNA sequencing data, has been proposed. (4, 69) However, the shadow regression approach depends on the underlying assumption that the number of sequencing reads with shadows increases linearly with the number of error-free sequences, which may not be appropriate for all types of sequencing data (69).

In Chapter 2, an empirical error rate estimation approach is proposed to model the nonlinear relationship between the number of sequences with shadows and the number of error-free read counts by cubic smoothing spline and robust smoothing spline approaches (69). The per-read sequence error rate is redefined, and provides more information according to the sequencing reads (69). In addition, a frequency-based simulation approach is developed based on the sequencing read counts, which can mimic the real sequencing data structure and reflect the nonlinear structure more realistically than Wang *et al.*'s simulation method (69). The study described in Chapter 2 has been published in BMC Bioinformatics (69).

Chapter 2 focuses on developing an empirical error rate estimation approach for substitution error rates, which are the sequencing error rates discussed in the majority of the literature. The shadows defined are sequences with up to two bases different from the error-free reads in the sample. In Chapter 3, the sequencing error rate estimation is extended to more general scenarios. The sequencing error rates are tested by extending the definition of shadows to be different numbers of bases in substitution, as well as error rates for deletion and insertion.

Chapter 4 describes statistical approaches to study the association between human microbiome and outcome of interest. The human microbiome has been associated with complex disorders. Advances in sequencing technology have allowed us to generate large amounts of microbiome data and enabled the quantification of the composition of the microbiome as high-dimensional data. Such a study on the human microbiome will potentially inform the discovery of species in the microbiome that are related to human health and disease. This chapter discusses approaches for analyzing microbiome data, from the initial steps of processing raw sequencing data to the downstream analysis, including assessing community-level diversities and their association with outcomes of interest.

Chapter 5 discusses the extension for the study of human microbiome cross-sectional data to longitudinal data, and provides a future direction regarding the analysis of the microbiome through a time series of longitudinal data, which can sufficiently reflect the overall performance of the composition of the human microbiome over time. A traditional longitudinal analysis of microbiome data using a general linear model with a fixed effect or a single summary about which indexes change for each sample has limitations, even in well-designed studies. Such limitations involve taking the correlation of repeated measures into account and detecting important effects on bacterial connections (79, 80). The analytical Bayesian network model for longitudinal data not only accounts for the correlation on repeated observations, but also performs more accurate estimation than conventional approaches. Chapter 5 provides a future direction regarding processing sparse compositional microbiome data and using a Bayesian network in a study of longitudinal time-series microbiome data.

Chapter 2

Empirical estimation of substitution sequencing error rates using smoothing splines

(Most of the studies in this chapter have been published online in BMC Bioinformatics, April 2016: Xuan Zhu, Jian Wang, Bo Peng, Sanjay Shete, “Empirical estimation of sequencing error rates using smoothing splines”. According to the journal policy, the author retains the right to include the published article in full or in part in a dissertation.)

2.1 Background

Complex diseases have been recognized associated with miscellaneous genetic variants. The advent of next generation sequencing technology has enabled the studies to investigate a diverse range of biological areas and provided unprecedented opportunities and insights to further understand multiple disease functions. Next generation sequencing has been used in biological investigations, such as identifying disease associated biomarkers, quantification of gene expression, microRNA profiling, discovery of transcription factor binding sites, and whole-genome sequencing (5, 69, 70, 73, 81-83). Next generation sequencing refers to be the massively paralleling high-throughput sequencing, which is capable of sequencing millions of short reads at the same time (69, 84, 85). Due to large numbers of parallel sequencing and much smaller reaction volumes, Next generation sequencing has a number of advantages over conventional Sanger sequencing, such as significantly driving down the per base sequencing cost and producing high-throughput sequencing (6, 19, 20, 69, 74, 83). Despite the advantages of next generation sequencing, large numbers of short-read sequences of this technique makes it difficult to correctly and fully assemble the sequences (69, 86-88). In addition, high throughput sequencing in next generation sequencing is at a cost of raw accuracy that the error rates in next generation sequencing are often higher than those in traditional Sanger sequencing, which has become a general and acute problem using next generation sequencing data that negatively impacts downstream genomic analysis (4, 25, 74, 83, 89, 90). Therefore, it is essential to assess the quality of next generation sequencing data before performing downstream genomic analysis (69, 77). Generally, the quality assessment methods include investigating sequence coverage,

paired-end and fragment-size distributions, measuring sequence error rates, and quality matrices (FastQC) (69, 77, 91, 92).

Currently, the percentage of reads mapped has been as a quality indicator assessing next generation sequencing data; nonetheless, it does not address the fundamental question about how much error is present from the sequencing reads in a sample (4). Alternative error rate estimation approaches have been developed either using or not using a reference genome. Bullard et al. (93) proposed a mismatch sequence counting approach under the assumption that the sequencing reads with 0 mismatch had no error and the sequencing reads with 1 or 2 mismatches were errors instead of polymorphisms. This mismatch sequence counting approach aligns the reads to a reference genome in order to acquire the number of uniquely mapped reads with 0, 1, or 2 mismatches, and then calculates the per-read error rate as the proportion of sequence reads containing errors (83). However, this approach requires a well-established reference, which may not be applicable to sequence a new genome (69, 91).

Another approach using error correction tools according to k -mer (77, 94-100) is capable to perform error rate estimation without requiring a reference genome and is less sensitive to errors due to polymorphisms (101). Nonetheless, these error correction tools based on k -mer approach calculates the frequencies of all distinct sub-strings in the sequence with length k , resulting in much computational cost and requiring a large amount of computer memory, which may be difficult to apply for large sequencing genomes (69, 102-104).

Wang et al. (83) developed a shadow regression approach to estimate the error rates for next generation sequencing data with the added advantage of not using a reference genome. Based on observations, the shadow regression approach estimates the error rates depending on the assumption that the number of shadows increases linearly with the number of error-free sequence

reads. In shadow regression approach, the per-read error rate referred to the rate of the number of sequences containing errors over the total number of sequence reads in a sample. The shadow regression approach employs a linear regression model to evaluate the sequencing error rates. Nonetheless, in practice, the linearity assumption may not be appropriate for all types of sequence data, which might lead to estimation bias when studying sequencing error rates and performing downstream genomic analysis (**4, 69**).

Consequently, in this study, we proposed the empirical error rate estimation approach to model the non-linear relationship between the number of sequence reads containing errors and the number of error-free sequences by employing cubic smoothing spline and robust smoothing spline approaches. We also redefined the sample per-read error rate which is capable to capture the sequence error rate according to the sequence reads and provide more information based on read counts. In addition, we developed a frequency-based simulation approach that is capable to mimic the real sequence data structure more realistically than the Wang et al.'s simulation approach, especially the non-linear relationship between the number of sequences containing errors and the number of error-free reads (**4, 69**). The simulation studies were investigated using next generation sequence data from the MicroArray Quality Control (MAQC) project (**105**), a mutation screening study (**106**), the Encyclopedia of DNA Elements (ENCODE) project (**107**), and bacteriophage PhiX 174RF1 DNA samples (**83**) and compared the performances between shadow regression approach and empirical error rate estimation approach. Under all simulation scenarios tested, the proposed empirical error rate approach provided a more accurate error rate estimation when the linearity assumption was not valid compared with shadow regression approach, and yielded similar error rate estimation when the linearity assumption hold. The real data analysis was also performed by applying the proposed empirical error rate estimation and

shadow regression approach for MAQC, mutation screening, ENCODE, and PhiX DNA samples (69).

2.2 Methods

2.2.1 Shadow Regression Overview

Wang et al. (83) made the important observation that the number of shadows due to sequencing errors increases linearly with the number of reads sequenced, whereas the number of true shadows is independent of the number of reads sequenced. In the current study, we employed and modified the definitions and notations from the Wang et al. (83) study.

Specifically, given a sequence t in a sample, the total number of reads can be given as $r_t = n_t + e_t$, where, r_t is the total number of reads with sequence t , n_t is the number of reads that are error free with sequence t , and e_t is the number of reads that contain sequencing errors with sequence t . In practice, one would not know the true error-free sequence reads. Wang et al. first converted a sequence file (i.e., fastq format with equal length for all reads) from a sample into a read counts file. The number of reads for each sequence was counted over the sample, and then the authors ranked all the sequences according to the read counts (see an example of read counts in Additional file 2). The top 1,000 reads with the highest frequencies in a sample were selected as the error-free sequences and used to calculate n_t and e_t in a different sequence t . The shadows of a given sequence t were defined as the reads differing from the error-free reads by up to two bases, which was deemed by Wang et al. (83) to be sufficiently similar to the given error-free

sequence t to estimate substitution errors. Finally, the top 1,000 reads with the highest frequencies in a sample were excluded from the assessment of the shadow counts.

The per-read error rate was defined as the proportion of reads containing sequencing errors over all the reads in a given sample (83), or $ER = \frac{\sum_t e_t}{\sum_t r_t} = \frac{\Delta e_t}{\Delta r_t}$. Based on the observation that the number of shadows due to the sequencing errors increases linearly with the number of reads, Wang et al. proposed a linear model as $s_t = \alpha + \beta n_t + \epsilon$, where s_t is the number of shadows of sequence t and ϵ is the independent error that follows approximately Gaussian distribution. Robust linear regression was used to estimate the coefficients α and β because shadows can come from legitimate error-free reads. In this situation, the per-read error rate can be estimated by the slope of the linear model as $ER = \frac{\Delta e_t}{\Delta r_t} = \frac{\Delta e_t}{\Delta n_t + \Delta e_t} = \frac{\beta}{1 + \beta}$, which was denoted as the shadow regression error rate (SRER) in the current study.

2.2.2 Empirical Per-Read Error Rate with Smoothing Splines

The error rates assessed by shadow linear regression were based on the assumption of a linear relationship between the error-free read counts and shadow counts. However, the real data examples in Wang et al.'s paper (e.g., Figure 3) show that for many types of sequencing data the relationship between error-free read counts and shadow counts does not follow a linear trend. In these situations, shadow regression based on a linear assumption might lead to a biased estimation of the sequence error rate. Therefore, in this study, we employed the cubic smoothing

spline and robust smoothing spline methods to model the relationship between the error-free read counts and the shadow counts; the piecewise curve resulting from these spline methods is capable of capturing relationships of widely varying form and tends to avoid erratic behavior near the extremes of the data (**108**). Based on the empirical read-shadow relationship, we proposed the error rate estimation as a function of error-free read counts, which we hypothesized would be more useful in practice than the shadow regression approach for estimating error rates.

We first used the cubic smoothing spline method (**109, 110**) to model the shadow counts (s_i) as a function of the error-free read counts (n_i). The cubic smoothing spline method fits a smooth curve to a set of observations using a cubic function (**110, 111**). Specifically, given a set of observations of error-free read counts and shadow counts, $(n_1, s_1), (n_2, s_2), \dots, (n_m, s_m)$, $n_1 < n_2 < \dots < n_m$, we modeled the relationship between n_i and s_i as a function $s_i = \mathcal{U}(n_i)$, with two continuous derivatives, where n_i is the number of error-free read counts with sequence i , s_i is the number of shadow counts with sequence i , $i = 1, \dots, m$, and m is the total number of sequences. Among all such functions with two continuous derivatives, the purpose of the smoothing spline is to find the estimated function minimizing the penalized residual sum of squares

$\sum_{i=1}^m (s_i - \hat{\mathcal{U}}(n_i))^2 + \lambda \int_{n_1}^{n_m} \hat{\mathcal{U}}''(n)^2 dn$, where λ is a smoothing parameter (**110**). We used the cubic smoothing spline implemented in the R function “smooth.spline” in the “stats” package (**93**). To improve the robustness of the spline, we used a robust smoothing spline approach to model the relationship between shadow counts (s_i) and error-free read counts (n_i). This approach uses an iterative re-weighted smoothing spline algorithm with the inverse of the absolute value of the residuals as the weights. We used the robust smoothing spline implemented in the R function “robustSmoothspline” in the package “aroma.light” (**112, 113**).

Based on the fitted model of shadow counts as a spline function of error-free read counts, we used the definition of the per-read error rate used by Wang et al. (83). Let n_1, n_2, \dots, n_m be the read counts of m sequences, where $n_1 < n_2, \dots, n_{m-1} < n_m$, and let $\hat{S}_1, \hat{S}_2, \dots, \hat{S}_m$ be the corresponding fitted values of shadow counts using the smoothing spline approaches described above. Given a sequence t with read count n_t and fitted shadow count \hat{S}_t , we defined the per-read error rate as

$$ER(n_t) = \frac{\sum_{j=1}^t \hat{S}_j}{\sum_{j=1}^t \hat{S}_j + \sum_{j=1}^t n_j}.$$

Estimation of error rates using this definition also assumes the increasing number of shadows with the increasing number of read counts in the presence of sequencing errors, but this definition allows error rates to vary by read counts, whereas the Wang et al. (83) approach assumes the error rate to be a constant irrespective of the number of reads. For the purpose of comparison with the Wang et al. (83) approach, we also defined a sample per-read empirical error rate. Given a sample with m sequences, we randomly sampled 1,000 numbers of read counts from n_1 to n_m and calculated the per-read error rates for each of these read counts using the fitted spline. The median of these per-read error rates was reported as a sample per-read empirical error rate (henceforth denoted as EER).

2.2.3 Next-Generation Sequencing Data

We used next-generation sequencing data from four projects--the MicroArray Quality Control (MAQC) project (mRNA-seq) (*105*), a mutation screening study (re-sequencing) (*106*), the Encyclopedia of DNA Elements (ENCODE) project (mRNA-seq) (*107*), and bacteriophage PhiX 174RF1 (PhiX) DNA samples (*83*)--to perform the simulations and demonstrate the accuracy of the proposed empirical error rate estimation approach. PhiX DNA data in FASTQ format were generated by the Center for Cancer Computational Biology at Dana-Farber Cancer Institute and provided by Wang et al. (*83*). The other three data sets are publicly available from the National Center for Biotechnology Information Sequence Read Archive (*114*) in FASTQ format with equal read lengths in each sample. All the data were converted to read counts using the shadow regression program provided by Wang et al. (*83*).

2.2.3.1 MAQC Brain Experiment 2 Data

The MAQC project was initiated to address concerns about the reliability of microarray technology (*105*). This project provided gene expression levels measured from two RNA samples, including a Universal Human Reference RNA from Stratagene and a Human Brain Reference RNA from Ambion, in four titration pools on seven microarray platforms with three expression methodologies. In this study, we used the sequence data from the MAQC brain experiment 2 (Sequence Read Archive [SRA]010153), including 14 samples on two flowcells (SRX016366 and SRX016368) run on the Illumina 1G Genome Analyzer with each sample containing ~12 million reads.

2.2.3.2 Mutation Screening Resequencing Data

The mutation screening study provided re-sequencing data (SRA010105) from 24 patients with X-linked mental retardation (XLMR) for mutations in 86 previously identified XLMR genes, using a method that combined a novel droplet-based multiplex PCR method and next-generation sequencing (*106*). An Illumina/Solexa Genome Analyzer II platform was used to perform the sequencing, and each sample contained ~12 million reads.

2.2.3.3 ENCODE Transcriptome Data

The ENCODE project used high-throughput approaches to provide a biologically more informative representation of the human genome (*107*). The ENCODE pilot phase included more than 200 experimental and computational data sets from 35 groups (*107*). In this study, we used the ENCODE human mRNA sequence data (SRA001150), including 5 samples of human cell line K562 (SRX000570) run on the Illumina 1G Genome Analyzer; each sample contained ~12 million reads.

2.2.3.4 PhiX DNA Data

The bacteriophage PhiX 174 is an icosahedral virus. It contains a closed circular single-stranded DNA molecule with 5,386 nucleotide bases (*115*). PhiX 174 was the first DNA-based genome for which the complete nucleotide sequence was successfully determined (*115-117*). In this study, we used two PhiX DNA samples provided by Wang et al., which were generated from the Center for Cancer Computational Biology at Dana-Farber (*83*). The PhiX DNA samples were sequenced using Illumina. One sample contained ~14.6 million reads, and the other contained ~25.7 million reads.

2.3 Simulation Approaches

We performed simulation studies to investigate the performance of the proposed EER and compare it with the SRER. We used two approaches to perform the simulation: (1) Wang et al.'s simulation approach and (2) a new frequency-based simulation approach described below.

2.3.1 Wang et al. Simulation Approach

In the Wang et al. study (*83*), the simulation was conducted based on a sample from the MAQC brain experiment 2 (SRR037440) using calibration. The authors considered reads uniquely mapped to the reference genome with no mismatches as error-free reads and then added substitution errors based on pre-specified base-specific error rates, which were estimated from the sample SRR037440 by counting the number of mismatches to the reference genome at each base. To mimic the Wang et al. simulation procedure, we used the same sample (SRR037440) to

retrieve the error-free reads and estimate the pre-specified base-specific error rates. In particular, we retrieved ~4.4 million perfectly matched reads from the SRR037440 sample, which originally had ~12 million reads. We used several approaches to estimate the pre-specified base-specific error rates. In our first approach, we aligned the reads to the reference genome, marked the mismatch locations, and then calculated base-specific error rates as a percentage of the mismatch nucleotides at each location (Additional file 3(C)). In our second approach, we located all unprocessed reads for each sequence, recorded locations and numbers of mismatch nucleotides, and calculated the base-specific error rates from the total number of mismatch nucleotides at each location (Additional file 3(D)). However, we found that neither of these approaches could obtain the same base-specific error rates shown in the Wang et al. study. Thus, we included the pre-specified base-specific error rates used in their study, which we extracted approximately from the figures in the article (83) (Additional file 3(E)). In addition, we used the base-specific error rates for sample SRR037440 obtained from the shadow regression software developed by Wang et al. (83) as another set of pre-specified base-specific error rates (Additional file 3(F)).

The Wang et al. simulation approach assumed an underlying linear relationship between the error-free read counts and shadow counts, no matter which base-specific error rate was used. Therefore, the data simulated by using the Wang et al. simulation approach would not reflect the non-linear relationships between error-free read counts and shadow counts (Additional file 3(A)). For example, in some samples, when the number of error-free read counts is low, the shadow counts may decrease as the error-free read counts increase. Therefore, we proposed a frequency-based simulation approach to generate the error-free read counts and shadow counts directly, which can mimic the real sequence counts of the data structure.

2.3.2 Frequency-based Simulation Approach

Given the counts of all sequence reads in a real sample, we followed the same procedure described in Wang et al. to obtain the error-free read counts (n_i) and shadow counts (s_i) for N sequence reads, $i = 1, 2, \dots, N$ and $n_1 < n_2, \dots, n_{N-1} < n_N$. We selected the top N reads with the highest frequencies to be the error-free reads and obtained the corresponding read counts (n_i); we then mapped all the rest of the sequences in the sample to the error-free reads to identify the corresponding shadow counts (s_i). We used $N=1,000$, as suggested in Wang et al. Based on the read counts and shadow counts in the real sample, we first constructed a frequency table for the error-free read counts (n_i) using pre-specified bin widths. Because in most of the samples we investigated, the error-free read counts became very sparse when they were large, we used unequal bin widths to avoid having almost empty bins. Within each bin of the error-free read counts, we then constructed a frequency table for the shadow counts (s_i) using pre-specified equal bin widths. Given M bins for error-free read counts and L bins for shadow counts within each of the error-free read count bins, the frequencies can be written as $p_j, j = 1, 2, \dots, M$, for error-free read counts and $q_{kj}, k = 1, 2, \dots, L$ and $j = 1, 2, \dots, M$, for shadow counts. To sample a pair of observations (n_{new}, s_{new}), we first sampled two independent random numbers, $U \sim \text{uniform}(0, 1)$ and $V \sim \text{uniform}(0, 1)$. If $U \in [\sum_{i=1}^j p_i, \sum_{i=1}^{j+1} p_i)$, we sampled $n_{new} \sim \text{uniform}(n_j, n_{j+1})$, where n_j and n_{j+1} are the endpoints for the corresponding bin j of read counts. Further, within the read counts bin j , if $V \in [\sum_{i=1}^k q_{ij}, \sum_{i=1}^{k+1} q_{ij})$, we sampled $s_{new} \sim \text{uniform}(s_k, s_{k+1})$, where s_k and s_{k+1} are the endpoints for the corresponding bin k of shadow counts. With this approach, we can generate more pairs of error-free read and shadow counts (i.e., $> 1,000$).

2.3.3 Comparisons of Different Simulation Approaches

We performed a simulation study to compare the Wang et al. simulation approach with our proposed frequency-based simulation approach using information from sample SRR037440 in Wang et al.'s study (83). Figure 1 shows the relationships between the shadow counts and error-free read counts for the original SRR037440 sample data (panel A), as well as the simulated data generated using the frequency-based simulation approach (panel B) and the Wang et al. simulation approach using different pre-specified base-specific error rates (as described in the Method; panels C-F). As expected, the data generated using the Wang et al. simulation approach showed that the shadow counts increase linearly as the read counts increase, which does not reflect the true pattern of the original data (panel A). Also, different pre-specified base-specific error rates had very similar performance. On the other hand, the data generated using the frequency-based simulation approach (panel B) showed a pattern very similar to that of the original data. We also plotted the fitted shadow linear regression lines, cubic smoothing spline curves, and robust smoothing spline curves in Figure 1 for the original and simulated data, respectively. It should be noted that the shadow linear regression lines and the smoothing spline curves performed similarly when the data were generated using the Wang et al. simulation approach because the approach generated data using a linear relationship. However, compared to the linear lines, the smoothing spline curves captured more features when both the original data and the data generated using the frequency-based simulation approach were used (panels A and B).

Figure 1. Sample SRR037440 from the MAQC brain experiment 2 data set and corresponding simulated data using frequency-based and Wang et al. simulation approaches

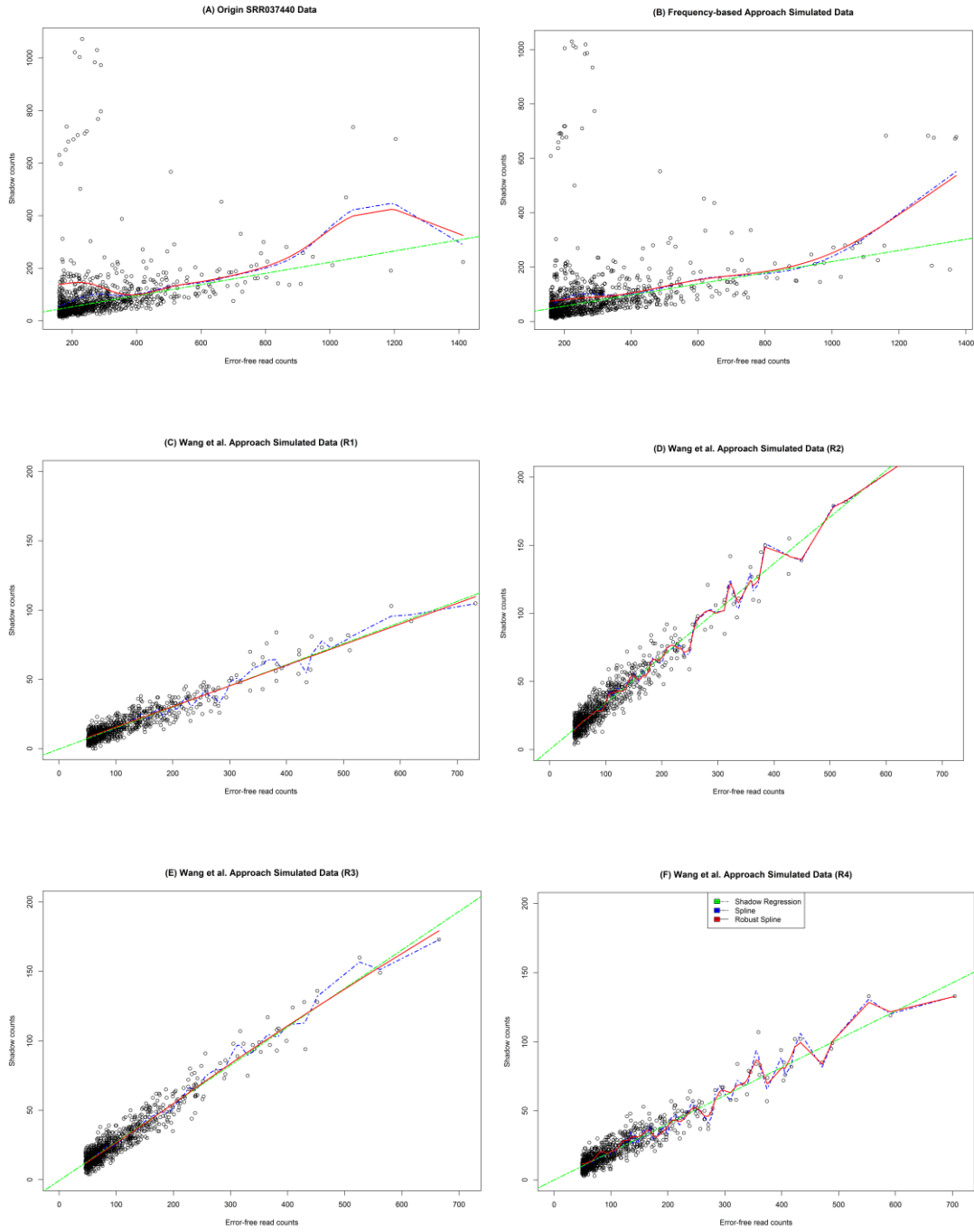


Figure 2. Read counts for the SRR032577 sample in the mutation screening re-sequencing study

CAAAGCAACTTATGGGACTTGGTTGGCTTCTGTTTG 22010
CCAGACTTCTCCTCAAGTTATGCAAATCTTATGTCA 17456
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAA 14515
ACTCATACAGGAGAGAAACCTTATGAATGCAGAGAC 13720
GAGACAGAGCTGTAGAGAAACAAAAAGAGAAAGATG 12759
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTAGAT 9832
GAACAAATGCTTTTCCAACCCATGAGTGCTAAGAGC 9192
CTCGCCATGATTATTTGACAAATAATGAGACTAGTA 7583
GTATTCATCAAGCTGACTGGATCCATTTGTCCGGGT 7357
GATGACAATCTCCAAGTGACATTTCACTGTGTTCCT 7039
AGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAA 6887
GGTATCTAATAGCAGGGAGGAGGAAGGCCTGTTGCT 6846
AGCAACATGCTATGAACAAAGACCCTTCTGTAACAA 6700
GGTGATGTCCTATCACACTAAACATCGATTGGAGTG 6399
AAGAAGCAGTTTTAGCTTCTTAGATCGGAAGAGCTC 6224
GATAATTGAGCTGTGCATGAATGTAACTCTTAAAA 6136
GGTAGTGGATCTTTCTGTCTATCAAGAACAGGCCTG 6022
GGTATGAGCTTGACTGAGAGGCCTCTCCTGACTATC 5894
ATCACAGAATCACAAAGCTGGTAAGGAGCCTCCCAA 5702
AGCAAACCTGAAATTTTCCCACTGAGTCATTCATG 5670
GCAGTTCTAGTCAAGCATACAATATCCAGAACCACC 5599
GTACATAGTAGGTGCTCGATGAATGGTTATTTCTC 5596
AATTTTATGACCAGGCAGTGTTAAAATTAATACTTA 5489
GTGCTCTTTCTTGAGAAGTCTTCCAGTTATTCAACT 5361
GTTACACAACACATGTCAGTTGCAGGGGCAGGAACT 5279
GATTCTCTTTGCCAGCTTAGATGGCTTCGGTTTCAG 5193

.....

We have also evaluated the error rates for the original sample SRR037440 and the simulated data generated using different simulation approaches and different error rate estimation approaches, including shadow linear regression (SRER), cubic smoothing spline (EER_CS), and robust smoothing spline (EER_RS) approaches. When the data sets were generated using the Wang et al. approach, the estimated error rates obtained using the different approaches were very similar and were close to the expected error rate, which was consistent with the results shown in Figure 1. For example, when the first pre-specified base-specific error rate (R1) was used, the expected per-read error rate was 0.1308 and the estimated error rates were 0.1313, 0.1328, and 0.1319 for SRER, EER_CS, and EER_RS, respectively. In the data set generated using the frequency-based simulation approach, the approaches using smoothing splines provided more accurate estimations for the per-read error rates than the shadow linear regression approach: the expected error rate was 0.2436, and the estimated error rates were 0.1614, 0.2157, and 0.2166 for SRER, EER_CS, and EER_RS, respectively.

Table 1. Comparison of the Wang et al. and proposed frequency-based simulation approaches using sample SRR037440*

Simulation Approaches		Expected ER	SRER	EER_CS	EER_RS
Original sample		-	0.1910	0.2297	0.2409
Wang et al.	R1	0.1308	0.1313	0.1328	0.1319
	R2	0.2563	0.2548	0.2560	0.2553
	R3	0.2135	0.2149	0.2173	0.2153
	R4	0.1702	0.1692	0.1704	0.1718
Frequency_based		0.2436	0.1614	0.2157	0.2166

* For the frequency-based simulation approach, we considered the top 1,000 reads of sample SRR037440 with the highest frequencies as the error-free reads and generated 1,000 pairs of error-free read and shadow counts.

Expected ER: expected error rate in simulation studies

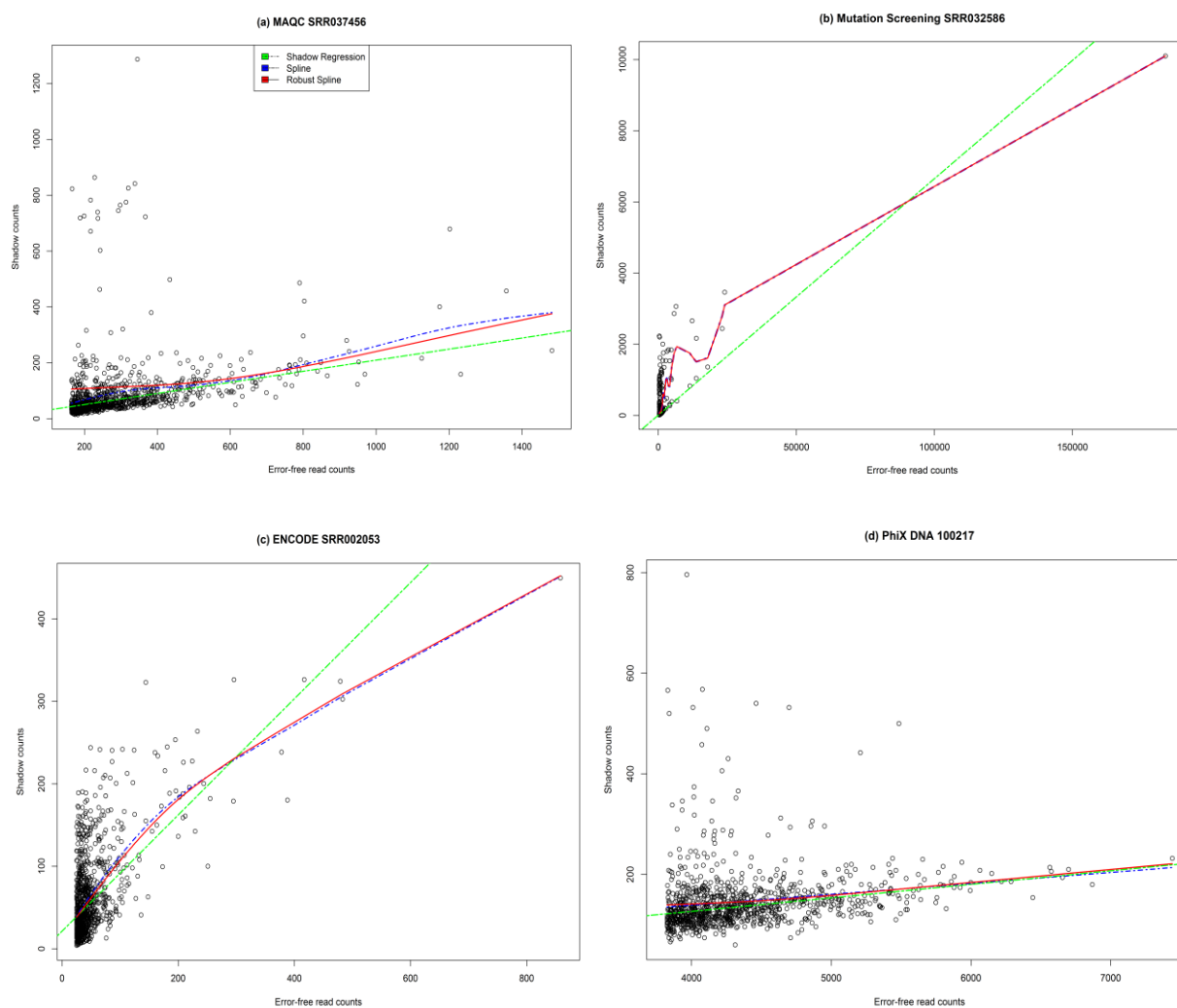
SRER: error rate estimated using shadow regression

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_RS: empirical error rate estimated using robust smoothing spline

R1, R2, R3 and R4: different pre-specified base-specific error rates used in the simulation as described in the Methods section

Figure 3. Sample sequencing data from MAQC, Mutation Screening Re-Sequencing, ENCODE, and PhiX DNA data sets



A comparison between the Wang et al. simulation approach and our proposed frequency-based simulation approach, using the SRR037440 sample, showed that unlike the frequency-based approach, the Wang et al. simulation approach does not mimic the observed non-linear data relationship between error-free read and shadow counts. Therefore, we applied only the proposed frequency-based simulation approach in the simulation studies. We generated data based on the information from samples from the four data sets described above. For each sample, the median EER was calculated based on 1,000 replicates and compared with the median SRER. For all the simulations, we considered the top 1,000 reads of the sample of interest with the highest frequencies as the error-free reads and then determined the shadow counts accordingly, which were then used to generate the frequency tables for the simulations. Based on the frequency tables, we generated 1,000 pairs of error-free reads and shadow counts for each replicate. For the simulation data sets, we also defined an expected per-read error rate as $\sum_i s_i / (\sum_i s_i + \sum_i n_i)$, where s_i and n_i are the shadow count and error-free read count, respectively, for sequence i , $i = 1, \dots, M$. Note that $M = 1,000$ in the simulation studies, as we generated 1,000 pairs of error-free read and shadow counts.

2.4 Results

2.4.1 Simulation Results

As shown in figure 1 and table 1, using the frequency-based simulation approach can better capture the relationship between error-free read and shadow counts, therefore, we used this simulation approach to perform further simulations based on next-generation sequencing data from the MAQC, mutation screening, ENCODE, and PhiX DNA sample data sets. We compared the performance of our proposed EER approach using the cubic or robust smoothing spline method (EER_CS or EER_RS, respectively) with that of the SRER approach.

2.4.1.1 Simulation Results for MAQC Data

Table 2 shows the median error rates (based on 1,000 replicates) obtained using the shadow linear regression approach and the proposed smoothing spline approaches, based on the next-generation sequencing data from the MAQC study. We have also reported the expected error rates (calculated as described above) and the estimation biases, which were calculated as the absolute differences between the estimated error rates and the expected error rates. For all 14 samples of MAQC data, both smoothing spline approaches provided more accurate estimations of the error rates with less bias than the shadow linear regression approach. Both smoothing

spline approaches performed very similarly. For example, for sample SRR037452, the median expected error rate in the simulation was 0.3305. Using SRER, the median estimated error rate was 0.2578, with a bias of 0.0727 compared to the expected error rate. In contrast, EER_CS and EER_RS had median estimated error rates of 0.3104 and 0.3096, respectively, with biases of 0.0201 and 0.0209, respectively. From these results, we can observe that the SRER approach underestimated the error rates, while the smoothing spline approaches provided more accurate estimated error rates.

Table 2. Median error rates in MAQC data using shadow linear regression and smoothing spline approaches*

Samples	Expected ER	SRER	SRER Bias	EER_CS	EER_CS Bias	EER_RS	EER_RS Bias
SRR037452	0.3305	0.2578	0.0727	0.3104	0.0201	0.3096	0.0209
SRR037453	0.1917	0.1584	0.0333	0.1824	0.0093	0.1818	0.0099
SRR037454	0.2354	0.1515	0.0839	0.2060	0.0294	0.2059	0.0295
SRR037455	0.1759	0.1448	0.0311	0.1675	0.0084	0.1668	0.0091
SRR037456	0.2312	0.1622	0.0690	0.2037	0.0275	0.2035	0.0277
SRR037457	0.1841	0.1480	0.0361	0.1777	0.0064	0.1771	0.0070
SRR037458	0.2653	0.2321	0.0332	0.2582	0.0071	0.2575	0.0078
SRR037459	0.2371	0.1943	0.0428	0.2202	0.0169	0.2203	0.0168
SRR037460	0.2530	0.2018	0.0512	0.2503	0.0027	0.2490	0.0040
SRR037461	0.2180	0.1704	0.0476	0.2105	0.0075	0.2104	0.0076
SRR037462	0.2443	0.1734	0.0709	0.2322	0.0121	0.2308	0.0135
SRR037463	0.2154	0.1654	0.0500	0.2023	0.0131	0.2045	0.0109
SRR037464	0.2624	0.1666	0.0958	0.2392	0.0232	0.2403	0.0221
SRR037465	0.2145	0.1742	0.0403	0.2038	0.0107	0.2037	0.0108

* Based on 1,000 replicates. The frequency-based simulation approach was applied. For each replicate, we considered the top 1,000 reads with the highest frequencies as the error-free reads and generated 1,000 pairs of error-free read counts and shadow counts.

Expected ER: expected error rate in simulation studies

SRER: error rate estimated using shadow regression

SRER Bias: the absolute value of the difference between SRER and Expected ER

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_CS Bias: the absolute value of the difference between EER_CS and Expected ER

EER_RS: empirical error rate estimated using robust smoothing spline

EER_RS Bias: the absolute value of the difference between EER_RS and Expected ER

2.4.1.2 Simulation Results for Mutation Screening Resequencing Data

Table 3 reports the median error rates obtained using the SRER approach and the two EER approaches, based on next-generation sequencing data from a mutation screening study. Similar to the results from the MAQC samples, all 24 mutation screening resequencing samples yielded error rates for the smoothing spline approaches that were more accurate than or comparable to the estimations of error rates for SRER. For example, in sample SRR032566, the median of the expected error rate in the simulation was 0.0734. Using SRER, the median of the estimated error rate was 0.0412, with a bias of 0.0323 compared with the expected error rate. Using the smoothing spline approaches, the median of the estimated error rate was 0.0705 for both approaches, with a very small bias of 0.0029.

Table 3. Median error rates in mutation screening data using shadow linear regression and smoothing spline approaches*

Samples	Expected ER	SRER	SRER Bias	EER_CS	EER_CS Bias	EER_RS	EER_RS Bias
SRR032565	0.1991	0.0493	0.1498	0.1080	0.0911	0.1082	0.0910
SRR032566	0.0734	0.0412	0.0323	0.0705	0.0029	0.0705	0.0029
SRR032567	0.2003	0.0542	0.1461	0.1111	0.0892	0.1111	0.0893
SRR032568	0.2040	0.0437	0.1603	0.1057	0.0984	0.1072	0.0968
SRR032569	0.1598	0.0509	0.1089	0.1018	0.0580	0.1014	0.0585
SRR032570	0.0985	0.0641	0.0345	0.0959	0.0026	0.0954	0.0032
SRR032571	0.1236	0.0575	0.0661	0.1406	0.0170	0.1406	0.0170
SRR032572	0.1495	0.0530	0.0965	0.1181	0.0314	0.1173	0.0323
SRR032573	0.1779	0.0912	0.0867	0.1518	0.0261	0.1506	0.0273
SRR032574	0.0986	0.0384	0.0602	0.0626	0.0361	0.0618	0.0368
SRR032575	0.1169	0.0839	0.0330	0.1228	0.0059	0.1227	0.0058
SRR032576	0.1554	0.0945	0.0609	0.1887	0.0333	0.1880	0.0326
SRR032577	0.1052	0.0694	0.0359	0.1054	0.0002	0.1055	0.0002
SRR032578	0.1143	0.0448	0.0695	0.1077	0.0067	0.1076	0.0067
SRR032580	0.0870	0.0619	0.0251	0.1169	0.0298	0.1163	0.0292
SRR032581	0.0770	0.0347	0.0424	0.0752	0.0018	0.0751	0.0019
SRR032582	0.0400	0.0041	0.0359	0.0309	0.0091	0.0310	0.0090
SRR032583	0.1224	0.0707	0.0517	0.1279	0.0055	0.1280	0.0056
SRR032584	0.1290	0.0540	0.0750	0.1052	0.0238	0.1032	0.0259
SRR032586	0.0445	0.0102	0.0343	0.0287	0.0158	0.0287	0.0159
SRR032587	0.1486	0.0786	0.0700	0.1562	0.0076	0.1552	0.0066
SRR032588	0.1240	0.0470	0.0770	0.1024	0.0216	0.1021	0.0220
SRR033543	0.1151	0.0587	0.0564	0.0828	0.0323	0.0832	0.0318
SRR033544	0.1267	0.0524	0.0743	0.1086	0.0181	0.1085	0.0182

*Based on 1,000 replicates. The frequency-based simulation approach was applied. For each replicate, we considered the top 1,000 reads with the highest frequencies as the error-free reads and generated 1,000 pairs of error-free read counts and shadow counts.

Expected ER: expected error rate in simulation studies

SRER: error rate estimated using shadow regression

SRER Bias: the absolute value of the difference between SRER and Expected ER

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_CS Bias: the absolute value of the difference between EER_CS and Expected ER

EER_RS: empirical error rate estimated using robust smoothing spline

EER_RS Bias: the absolute value of the difference between EER_RS and Expected ER

2.4.1.3 Simulation Results for ENCODE Transcriptome Data

Table 4 reports the median error rates obtained using different approaches based on next-generation sequencing data from the ENCODE study. The five samples from the ENCODE study had higher expected error rates than the samples in the MAQC and mutation screening studies, which might have been due to relatively large shadow counts that corresponded with smaller error-free read counts (Figure 3, C). In this situation, the smoothing spline approaches still performed better than shadow linear regression. For example, in sample SRR002056, the expected error rate was 0.3646, and the estimated error rates were 0.2906, 0.3270, and 0.3233 for SRER, EER_CS, and EER_RS, respectively, with biases of 0.0740, 0.0376, and 0.0413, respectively.

Table 4. Median error rates in ENCODE data using shadow linear regression and smoothing spline approaches*

Samples	Expected ER	SRER	SRER Bias	EER_CS	EER_CS Bias	EER_RS	EER_RS Bias
SRR002053	0.5548	0.4153	0.1395	0.4609	0.0939	0.4565	0.0983
SRR002056	0.3646	0.2906	0.0740	0.3270	0.0376	0.3233	0.0413
SRR002065	0.4578	0.3371	0.1207	0.3740	0.0838	0.3701	0.0877
SRR005092	0.6300	0.4047	0.2253	0.4797	0.1503	0.4727	0.1573
SRR005093	0.4839	0.3928	0.0911	0.4221	0.0618	0.4173	0.0666

***Based on 1,000 replicates. The frequency-based simulation approach was applied. For each replicate, we considered the top 1,000 reads with the highest frequencies as the error-free reads and generated 1,000 pairs of error-free read counts and shadow counts.**

Expected ER: expected error rate in simulation studies

SRER: error rate estimated using shadow regression

SRER Bias: the absolute value of the difference between SRER and Expected ER

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_CS Bias: the absolute value of the difference between EER_CS and Expected ER

EER_RS: empirical error rate estimated using robust smoothing spline

EER_RS Bias: the absolute value of the difference between EER_RS and Expected ER

2.4.1.4 Simulation Results for PhiX DNA Data

Table 5 reports the median error rates obtained using different approaches based on the next-generation sequencing data from two PhiX DNA samples. For both samples, the smoothing spline approaches provided more accurate estimations of the error rates than did shadow linear regression. For example, in sample 100217, the median of the expected error rate in the simulation was 0.0323. Using SRER, the median of the estimated error rate was 0.0250, with a bias of 0.0073 compared to the expected error rate. The median estimated error rate for EER_CS and EER_RS was 0.0315, resulting in a much smaller bias of 0.0008 for both approaches.

Table 5. MedianSimulation error rates in PhiX DNA data using shadow linear regression and smoothing spline approaches*

Samples	Expected ER	SRER	SRER Bias	EER_CS	EER_CS Bias	EER_RS	EER_RS Bias
100217	0.0323	0.0250	0.0073	0.0315	0.0008	0.0315	0.0008
100514	0.0152	0.0143	0.0009	0.0152	0.0000	0.0152	0.0000

*Based on 1,000 replicates. The frequency-based simulation approach was applied. For each replicate, we considered the top 1,000 reads with the highest frequencies as the error-free reads and generated 1,000 pairs of error-free read counts and shadow counts.

Expected ER: expected error rate in simulation studies

SRER: error rate estimated using shadow regression

SRER Bias: the absolute value of the difference between SRER and Expected ER

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_CS Bias: the absolute value of the difference between EER_CS and Expected ER

EER_RS: empirical error rate estimated using robust smoothing spline

EER_RS Bias: the absolute value of the difference between EER_RS and Expected ER

2.4.2 Next-Generation Sequencing Data Analysis Results

We applied our smoothing spline approaches to evaluate the error rates for the next-generation sequencing data from MAQC, mutation screening, ENCODE and PhiX DNA samples, and compared error rates using our smoothing spline approaches and shadow linear regression. The estimated error rates are reported in Tables 6, 7, 8, and 9, respectively, for samples from the MAQC, mutation screening, ENCODE, and PhiX DNA data sets. From the results we can observe that the smoothing spline approaches always provided relatively higher estimates of the error rates. For example, in sample SRR037454 from MAQC (Table 5), the estimated error rates were 0.1596, 0.2041, and 0.2196, respectively, for SRER, EER_CS, and EER_RS. This was expected given that shadow linear regression tended to underestimate the error rates in the simulation results.

Table 6. Error rates in real MAQC data using shadow linear regression and smoothing spline approaches

Samples	SRER	EER_CS	EER_RS
SRR037452	0.2695	0.3124	0.3362
SRR037453	0.1598	0.1819	0.1822
SRR037454	0.1596	0.2041	0.2196
SRR037455	0.1482	0.1694	0.1700
SRR037456	0.1657	0.2062	0.2162
SRR037457	0.1541	0.1796	0.1793
SRR037458	0.2386	0.2573	0.2575
SRR037459	0.1996	0.2216	0.2233
SRR037460	0.2027	0.2504	0.2647
SRR037461	0.1779	0.2058	0.2093
SRR037462	0.1858	0.2329	0.2319
SRR037463	0.1771	0.2019	0.2072
SRR037464	0.1850	0.2377	0.2448
SRR037465	0.1842	0.2019	0.2070

SRER: error rate estimated using shadow regression

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_RS: empirical error rate estimated using robust smoothing spline

Table 7. Error rates in real mutation screening data using shadow linear regression and smoothing spline approaches

Samples	SRER	EER_CS	EER_RS
SRR032565	0.0753	0.1167	0.1206
SRR032566	0.0584	0.0746	0.0745
SRR032567	0.0846	0.1420	0.1446
SRR032568	0.0686	0.1566	0.1566
SRR032569	0.0597	0.1015	0.1020
SRR032570	0.0691	0.0973	0.0954
SRR032571	0.0724	0.1400	0.1400
SRR032572	0.0818	0.1611	0.1593
SRR032573	0.1602	0.2756	0.2752
SRR032574	0.0557	0.1004	0.1053
SRR032575	0.0882	0.1191	0.1179
SRR032576	0.1101	0.1915	0.1899
SRR032577	0.0762	0.1282	0.1280
SRR032578	0.1365	0.1874	0.1873
SRR032580	0.0727	0.1262	0.1266
SRR032581	0.0981	0.0506	0.0511
SRR032582	0.0941	0.1689	0.1679
SRR032583	0.1141	0.2057	0.2042
SRR032584	0.0849	0.0963	0.0742
SRR032586	0.0623	0.3606	0.3621
SRR032587	0.0857	0.1524	0.1532
SRR032588	0.0701	0.1446	0.1440
SRR033543	0.0802	0.1084	0.1102
SRR033544	0.1175	0.1588	0.1586

SRER: error rate estimated using shadow regression

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_RS: empirical error rate estimated using robust smoothing spline

Table 8. Error rates in real ENCODE data using shadow linear regression and smoothing spline approaches

Samples	SRER	EER_CS	EER_RS
SRR002053	0.4134	0.4469	0.4453
SRR002056	0.3225	0.3020	0.3355
SRR002065	0.3842	0.3918	0.3913
SRR005092	0.4628	0.4884	0.4776
SRR005093	0.4090	0.4724	0.4013

SRER: error rate estimated using shadow regression

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_RS: empirical error rate estimated using robust smoothing spline

Table 9. Error rates in real PhiX DNA data using shadow linear regression and smoothing spline approaches

Samples	SRER	EER_CS	EER_RS
100217	0.0261	0.0317	0.0315
100514	0.0142	0.0155	0.0157

SRER: error rate estimated using shadow regression

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_RS: empirical error rate estimated using robust smoothing spline

2.5 Conclusion

In summary, we proposed an empirical error rate estimation approach in which cubic and robust smoothing splines were used to model the read-shadow relationship. The proposed approach does not assume a linear relationship between the error-free reads and shadows counts and provides more accurate estimations of error rates for next-generation, short-read sequencing data (69).

2.6 Discussion

Next generation sequencing has enabled high throughput capacity of parallel sequencing and significantly driven down the cost which makes it widely employed to study various biological phenomena. Nonetheless, due to the high volume of short reads produced in next generation sequencing, the raw accuracy may impact the downstream analysis for genetic studies.

Therefore, before performing downstream genomic analysis, it is essential to accurately assess the sequencing error rates. We have developed the empirical error rate estimation approach which provides more accurate estimation than other available approaches for assessing next generation sequencing data (69).

In this chapter, we first reviewed the shadow regression error rate estimation approach (83). The shadow regression approach is under the linearity assumption that the number of shadows increases linearly with the number of error-free read counts. The linearity assumption might be

valid if one could obtain and plot the true error-free sequence read counts and the corresponding shadow counts by using only the reads containing sequencing errors. However, in practice, such information is not identifiable and the non-linear relationship is not rare. Sequencing data is with noise. It is critical to estimate the sequencing error rate more accurately and robust. We proposed a sample-level error rate taking the median of error rates obtained from different sets of error-free read counts and corresponding fitted shadow read counts, which may provide more robust estimation (69).

We tested the samples of next generation sequencing data presented in Wang et al. paper (4). The linearity assumption between the shadow read counts and the error-free read counts might hold in the PhiX DNA sequencing samples; however, it is not valid for the mutation screening re-sequencing samples and the mRNA sequencing data (MAQC and ENCODE) samples after investigation. Therefore, we proposed to employ the non-linear statistical model to estimate the relationship between the error-free read counts and shadow counts using smoothing spline approaches. The smoothness can be controlled and adjusted via tuning parameter in the penalty term, which can reduce the likelihood of overfitting the data (100). Because we desired a smoother interpolating function, the cubic smoothing spline approach was employed in our chapter instead of a linear smoothing spline approach. Moreover, in order to improve the estimation robustness, we also proposed to employ the robust smoothing spline approach, using a weight equal to the inverse of the absolute value of the residuals to iteratively estimate the re-weighted smoothing spline algorithm. In addition, to better investigate the performance of the proposed approach, we developed the frequency-based simulation approach, which can mimic the real sequence data structure more realistically than the Wang et al. simulation approach (4). Under the simulation studies tested, the shadow linear regression underestimated the sequencing

error rates, whereas the empirical error rate estimation approach provided more accurate estimation results. The superior performance of empirical error rate estimation approach also hold for the DNA sequencing data, in which the linear read-shadow relationship might be valid (69).

In practice, due to polymorphisms or duplications, the true genome sequence sample might differ from the reference genome. Therefore, in our chapter, we investigated two additional simulation scenarios: (1) as was assumed in Wang et al. (83), there is one polymorphism for every 1,000 base pairs in the sequence; and (2) there are two base-pair duplications for every 1,000 base pairs in the sequence, respectively. It is critical to note that these two assumptions are specific in simulation and are not necessary for shadow-based approaches. Specifically, we mapped the sample of SRR037440 to a reference and extracted the sequence reads that are uniquely mapped. We added one polymorphism or two base-pair duplications per 1,000 base pairs in the sequences and considered the resulting sequence reads as error-free sequence reads. Substitutional errors were then added according to the pre-specified per-base sequence error rates. After simulation studies tested in both scenarios, our proposed empirical error rate estimation approach provided more accurate estimation results. For example, in the simulation scenario under the first assumption that one polymorphism occurred per 1,000 base pairs, the expected error rate was 0.1977, whereas the estimated error rates were 0.1973 and 0.1989 using the cubic smoothing spline and robust smoothing spline, respectively. In the simulation scenario under the assumption that two base-pair duplications occurred per 1,000 base pairs, the expected error rate was 0.1728. The estimated error rates using the cubic smoothing spline and robust smoothing spline were 0.1722 and 0.1761, respectively. These simulation results showed that the proposed empirical error rate estimation approach is robust and not affected by polymorphisms or duplications as

well as the validity of the shadow-based approaches.

We applied the proposed empirical error rate estimation approach to real data analysis in studies of MAQC, mutation screening, ENCODE, and PhiX DNA, and compared the results with those obtained using the shadow linear regression approach. The error rates obtained using shadow regression approached were relatively lower compared with the ones estimated from the proposed approach. For example, in the samples of MAQC mRNA sequencing study, error rates yielded in our approach were between ~17% and ~32%, whereas error rates provided in shadow regression were ~15% to ~27%. As can be seen by even simple visual inspection, the lower error rates yielded by shadow regression approach may be attributed to the linearity assumption that is not valid in some real data sets. For the DNA sequence data where the linearity assumption might be valid, the proposed empirical error rate estimation approach provided similar estimations with those obtained using shadow linear regression (69).

To further investigate the performance of the proposed approach, simulation analysis was conducted. As the simulation approach proposed by Wang et al. (83) may not reflect the real sequence data structure, we developed a frequency-based simulation approach that captures the nonlinearity association between the number of reads containing errors and the number of error-free reads according to sequencing read counts from real data sets. These two simulation approaches were compared. From on the investigation, it showed that the data generated using the Wang et al. simulation approach followed the pattern that the shadow sequence read counts increased linearly with the error-free sequence read counts which was different from the real sequence data structure; compared with it, the data generated using the proposed frequency-based simulation approach mimics the real sequence data structure (69).

For the proposed approach, we adapted the original definition of per-read error rate that was the defined to be the proportion of reads containing sequencing errors among all the reads in a sample in Wang et al. (83). Instead of a fixed slope in the linear regression model, we re-defined the empirical per-read error rate as a function of error-free read counts that varies based on the basis of the sequence read. The proposed empirical per-read error rate could provide more information according to the sequence read counts compared to the original per-read error rate definition in Wang et al. (4) where a single fixed error rate is estimated for all the sequence reads in a given sample. We also defined a sample-level error rate that uses the median of the error rates estimated from different pair sets between the numbers of error-free read counts and corresponding fitted shadow counts (69).

In the shadow-based approaches, both the proposed error rate estimation approach and shadow linear regression approach can be impacted by outliers. Therefore, , we suggest to pre-process the next generation sequence data using standard statistical analysis approaches such as boxplot rule approach which is based on the upper and lower quartiles of data distribution, chi-square analysis (118), Dixon test (119), and Grubbs' test (120). Alternative approaches could be quantile regression (121) or Akima spline (122, 123) that might be more robust to outliers. In this chapter, we applied the proposed approaches to multiple next generation sequencing experiments, such as DNA sequencing, mRNA sequencing and re-sequencing data. We also showed that under the presence of polymorphisms and duplications, the shadow-bases approaches are valid. In addition, with the advantage of not requiring a reference genome in shadow-based approaches (i.e., shadow regression and our proposed approaches), it can be applicable to other types of sequencing studies, such as extensive polymorphisms, isoforms, or the microbiome. Specifically, the shadows (i.e., reads with errors) defined in the shadow-based

approaches are the sequence reads with up to two bases different from the error-free sequencing reads. The sequencing reads will not be counted as shadows with many differences, such as extensive polymorphisms, isoforms or microbiome data in the analysis (**69**).

Chapter 3

Extension Studies of Empirical Error Rate Estimation Approach for Substitution, Deletion, and Insertion Sequencing Errors

3.1 Introduction

Despite the growing number of sequenced genomes, our knowledge of mutation rate in insertion and deletions (indels) has only recently started to be explored (*124, 125*). In comparison, SNP has been deeply studied in most biological processes including genetic disease, species divergence, genetic evolution, and adaptation. Recent study showed that indels are associated with these processes and play an important role as well. Previous studies revealed that insertions radically increase frequency of ectopic recombination, which may lead to chromosome instability and genetic variation. Indels are mutational components of gene and pseudogene evolution and play an essential role in long-term evolution of genome size. Previous studies compared the ratio of nucleotide substitutions to the ratio of indels and showed that indels dominate the early divergence process (*126*). In addition, indels have been shown being associated with human disease, apoptosis mechanisms, phenotypic change, activity of surface antigen (*124, 127-136*). Recent study also found important association between indels and human cancer genes. For example, Yang *et al.* (2010) studied the somatic mutations in cancer

and found that one of the most significant characteristics of cancer related mutations is the high frequency of indels. (137)

Indels also happened to next generation sequencing data. To our knowledge, the majority existing sequencing error rate estimation approaches focused on substitution error rates. In shadow regression, the sequences containing errors were defined to be the sequences with up to two bases different from error free sequence reads for the short read next generation sequence data (4). However, the indel error rates also play an essential role that may negatively impact the quality of downstream genomic analysis. In genetics, sequencing deletion is a type of mutation that occurs when some genetic material is lost; whereas sequencing insertion occurs when extra genetic material inserted to a sequence (138, 139).

Conventional approaches (140-143) to study indels using a reference genome has the limitation when sequence a new genome as it may not be available to find a well-established reference for the new genome. It is critical to estimate the sequencing error rates more accurate and for more diverse sequencing error definitions including different bases in substitution, insertion, and deletion for next generation sequencing data. In this study, we extend the shadow definition by different bases in substitution and also study the indels without using reference based on the study in Wang et al. (4).

3.2 Method

In both previous chapter and shadow regression study to estimate short read sequencing error rates, the shadows were the sequences with up to two bases different from error-free reads for short reads sequencing data by substitution (4, 69). To better investigate the performance of the empirical error rate estimation approach (69) in more diverse scenarios, in this chapter, I adjusted the original shadow definition (4) to different bases of substitution, deletion, and insertion, respectively. Empirical error rate estimation approach was applied to multiple sequencing data in both simulation study and real data analysis, and compared with the estimation results obtained from shadow regression approach.

Shadows were the sequences containing errors. Different definitions of shadow would lead to different quality assessment results for next generation sequencing data. Based on Wang et al. (4), using shadows defined as sequence reads with up to two bases different from error-free reads provides enough estimation accuracy without excessive computational cost. In order to validate this statement, I adjusted the shadow definition to be only one base different, only two base different, only three bases different, and up to three bases different from error-free sequences in substitution, respectively.

As described in Chapter 2, both the empirical error rate estimation approach and shadow regression approach do not require using a reference genome. In the shadow-based approaches, the frequency of each sequence was first calculated according to the sequence read counts. Based on the frequency of each sequence read, the top 1000 sequences with the highest frequencies were selected as the error-free sequence reads (4, 69). All the possible sequences with only one

base different, only two bases different, only three bases different, up to two bases different, and up to three bases different from error-free reads in substitution were set out and generated as pseudo shadows, respectively. The corresponding pseudo shadows were then mapped to the remaining sequence reads that have not been selected as error-free sequences in the sample, to obtain the number of shadows of only one base different, only two bases different, only three bases different, up to two bases different, and up to three bases different from error-free reads in substitution for each sequence read. In this way, we can obtain the number of error-free sequences and corresponding shadow counts in different shadow definitions accordingly.

I applied the empirical error rate estimation approach using smoothing spline and robust smoothing spline data samples in DNA (sample 100217) and mRNA (sample SRR002053, SRR032586, SRR037456) and compared the results with those obtained from shadow regression approach to estimate the error rates in different substitution shadow definitions.

In addition to substitution, I also extend the study to deletion error rates. The shadows for deletion were computed differently from those by substitution. In the sequences data we were investigating, all the sequence reads are with the same lengths. Specifically, given a particular read in the sample, if one base is deleted from the original sequence, an extra base would be added to the end of the read to keep the length of number of bases in each sequence.

There are three approaches to count the deletion shadows in this study, which are left side deletion, right side deletion, and two-side deletion, respectively. In a forward direction sequence, suppose there is a base deleted from the original sequence, given that all the sequences are with the same lengths, the left side deletion is to add the extra base to the left end side of the read with

one of the four possible alleles (“A”, “T”, “G”, and “C”). Another way to estimate the deletion error rate is the right side deletion. In a forward direction sequence, the right side deletion error occurs when there is a base deleted from the original sequence, an extra base is added to the right end side of the read with one of the four possible alleles (“A”, “T”, “G”, and “C”). Furthermore, the third deletion error rate is two-side deletion. When one base is deleted from the original sequence, if the extra base is added to either left end side or right end side of the read, we called it two-side deletion. For example, given that the original sequence is ATGAC and all the sequences are with the same length, if the 3rd base was deleted, then the remaining sequence read would be ATAC. Since there were four possible alleles (“A”, “T”, “G”, and “C”), then the left side deletion is to add the extra base to the left side resulting the possible shadows as AATAC, TATAC, GATAC, and CATAC. The possible shadows in the right side deletion approach would then become ATACA, ATACT, ATACG, and ATACC. Moreover, for two side deletion, the possible shadows would become AATAC, TATAC, GATAC, CATAC, ATACA, ATACT, ATACG, and ATACC.

In addition to deletion and substitution errors, the third type of sequencing error refers to insertion. There are three types of sequencing insertion, which are left side insertion, right side insertion, and two-side insertion. In the data we investigated, all the sequences are with the same lengths. Suppose there is a base inserted into the original sequence, then an extra base will be removed from the end of the original sequence read. In a forward direction sequence, if the extra base is removed from the right end side, this insertion is defined as left side insertion. The base inserting to the last base in a forward sequence direction is not accounted for the left side insertion. The second type of insertion is right side insertion that supposing there is a base inserted to the original sequence, then an extra base will be removed from the left end side to

keep the sequence length same over all the sequences. In the right side insertion, the base inserting to the first base in a forward sequence direction is not accounted. In addition, there is also a two-side insertion approach, which refers to the scenario when a base inserted into the original sequence, given all the sequences are with the same lengths, an extra base removed from either left end side or right end side of the read.

I applied simulation study as in Chapter 2 and real data analysis using empirical error rate estimation approach and compared the results with those obtained from shadow regression approach with the use of different shadow definitions to investigate the performance of the proposed empirical approach.

3.3 Result

In order to better understand the error rate estimation performances when shadows varied, simulation studies were investigated for deletion and insertion error rates for different shadows of only one base different, only two bases different, and up to two bases different using empirical error rate estimation approach, and compared with the results from shadow regression approach.

Deletion error rates from left side, right side, and two sides in the simulation studies were investigated. Table 10 shows the median deletion error rates from simulation study based on the next-generation sequencing data samples, for shadows defining to be only one base different,

only two bases different, and up to two bases different from error-free sequences, (based on 1,000 replicates) obtained using the shadow regression approach and the proposed empirical error rate estimation approaches. The expected error rates was calculated as

$\sum_i s_i / (\sum_i s_i + \sum_i n_i)$, where s_i and n_i are the shadow count and error-free read count, respectively, for sequence i , $i = 1, \dots, M$, and $M = 1,000$ in the simulation studies, where we generated 1,000 pairs of error-free read and shadow counts (69). The estimation biases were calculated as the absolute differences between the estimated error rates and the expected error rates.

The empirical error rate estimation approaches provided more accurate error rate estimations with less bias than the shadow regression approach. Both empirical error rate estimation approaches using cubic smoothing spline and robust smoothing spline performed very similarly. For example, for sample SRR037452, the median expected left side deletion error rate with only one base different from error-free sequences in the simulation was 0.0453. Using SRER, the median estimated error rate was 0.0072, with a bias of 0.0382 compared to the expected error rate. In contrast, EER_CS and EER_RS had median estimated error rates of 0.0354 and 0.0351, respectively, with biases of 0.0099 and 0.0102, respectively. From these results, we can observe that the SRER approach underestimated the error rates, while the smoothing spline approaches provided more accurate estimated error rates.

Table 10 Simulation studies for deletion by testing shadows to be only 1 base, only 2 bases, and up to 2 bases deletion.

Shadow	Deletion	Per-read ER	SRR037452	SRR037456	SRR002053	SRR032586
only 1 deletion	left side	Expected ER	0.0453	0.0430	0.0622	0.0013
		SRER	0.0072	0.0042	0.0080	0.0001
		SRER Bias	0.0382	0.0387	0.0542	0.0012
		EER-CS	0.0354	0.0304	0.0337	0.0006
		EER-CS Bias	0.0099	0.0125	0.0285	0.0007
		EER-RS	0.0351	0.0306	0.0329	0.0006
		EER-RS Bias	0.0102	0.0124	0.0293	0.0007
only 2 deletion	left side	Expected ER	0.0250	0.0206	0.0423	0.0008
		SRER	0.0018	0.0006	0.0036	0.0000
		SRER Bias	0.0232	0.0200	0.0387	0.0008
		EER-CS	0.0178	0.0142	0.0197	0.0003
		EER-CS Bias	0.0072	0.0064	0.0226	0.0005
		EER-RS	0.0180	0.0146	0.0188	0.0003
		EER-RS Bias	0.0069	0.0060	0.0236	0.0005
up to 2 deletion	left side	Expected ER	0.0659	0.0581	0.1010	0.0020
		SRER	0.0113	0.0065	0.0141	0.0001
		SRER Bias	0.0545	0.0516	0.0869	0.0019
		EER-CS	0.0513	0.0424	0.0543	0.0008
		EER-CS Bias	0.0146	0.0157	0.0467	0.0011
		EER-RS	0.0508	0.0424	0.0515	0.0008
		EER-RS Bias	0.0150	0.0157	0.0496	0.0011
only 1 deletion	right side	Expected ER	0.0647	0.0640	0.1886	0.0032
		SRER	0.0124	0.0082	0.0246	0.0003
		SRER Bias	0.0523	0.0558	0.1640	0.0029
		EER-CS	0.0507	0.0473	0.1028	0.0022
		EER-CS Bias	0.0140	0.0167	0.0858	0.0009
		EER-RS	0.0503	0.0474	0.1006	0.0022
		EER-RS Bias	0.0144	0.0166	0.0880	0.0009
only 2 deletion	right side	Expected ER	0.1352	0.1303	0.2482	0.0043
		SRER	0.0205	0.0232	0.0425	0.0003
		SRER Bias	0.1147	0.1072	0.2058	0.0040
		EER-CS	0.1004	0.1024	0.1455	0.0028
		EER-CS Bias	0.0348	0.0280	0.1027	0.0015
		EER-RS	0.1011	0.1023	0.1427	0.0028
		EER-RS Bias	0.0341	0.0280	0.1055	0.0015

up to 2 deletion	right side	Expected ER	0.1808	0.1765	0.3615	0.0073
		SRER	0.0406	0.0374	0.0693	0.0008
		SRER Bias	0.1402	0.1391	0.2922	0.0065
		EER-CS	0.1419	0.1398	0.2222	0.0050
		EER-CS Bias	0.0389	0.0367	0.1394	0.0023
		EER-RS	0.1416	0.1391	0.2195	0.0050
		EER-RS Bias	0.0392	0.0374	0.1421	0.0023
only 1 deletion	two sides	Expected ER	0.0956	0.0924	0.2162	0.0037
		SRER	0.0172	0.0124	0.0246	0.0003
		SRER Bias	0.0784	0.0799	0.1916	0.0034
		EER-CS	0.0753	0.0680	0.1119	0.0023
		EER-CS Bias	0.0203	0.0244	0.1043	0.0014
		EER-RS	0.0745	0.0678	0.1098	0.0023
		EER-RS Bias	0.0211	0.0246	0.1064	0.0014
only 2 deletion	two sides	Expected ER	0.1974	0.1922	0.3041	0.0051
		SRER	0.0292	0.0261	0.0368	0.0003
		SRER Bias	0.1682	0.1661	0.2674	0.0047
		EER-CS	0.1472	0.1469	0.1613	0.0031
		EER-CS Bias	0.0502	0.0453	0.1428	0.0020
		EER-RS	0.1479	0.1466	0.1576	0.0031
		EER-RS Bias	0.0495	0.0457	0.1466	0.0020
up to 2 deletion	two sides	Expected ER	0.2576	0.2511	0.4177	0.0085
		SRER	0.0555	0.0433	0.0611	0.0008
		SRER Bias	0.2022	0.2079	0.3566	0.0077
		EER-CS	0.2016	0.1966	0.2411	0.0055
		EER-CS Bias	0.0560	0.0545	0.1766	0.0030
		EER-RS	0.2021	0.1956	0.2373	0.0056
		EER-RS Bias	0.0556	0.0556	0.1804	0.0030

***Based on 1,000 replicates. The frequency-based simulation approach was applied. For each replicate, we considered the top 1,000 reads with the highest frequencies as the error-free reads and generated 1,000 pairs of error-free read counts and shadow counts.**

Expected ER: expected error rate in simulation studies

SRER: error rate estimated using shadow regression

SRER Bias: the absolute value of the difference between SRER and Expected ER

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_CS Bias: the absolute value of the difference between EER_CS and Expected ER

EER_RS: empirical error rate estimated using robust smoothing spline

EER_RS Bias: the absolute value of the difference between EER_RS and Expected ER

Table 11 reports the median insertion error rates from simulation study for shadows defining to be only one base different, only two bases different, and up to two bases different from error-free sequences, obtained using the SRER approach and the two EER approaches, based on next-generation sequencing data samples. Insertion error rates from left side, right side, and two sides in the simulation studies were investigated. Similar to the simulation results from the deletion investigation, all samples in the insertion studies yielded error rates for the smoothing spline approaches that were more accurate than the estimations of error rates for SRER. For example, in sample SRR037452, the median of the left side insertion expected error rate with shadows defined as sequences of up to two bases different from error-free sequence reads in the simulation was 0.1930. Using SRER, the median of the estimated error rate was 0.0197, with a bias of 0.1733 compared with the expected error rate. Using the smoothing spline approaches, the median of the estimated error rates were 0.1689 for cubic smoothing spline with a very small bias of 0.0241, and 0.1783 for robust smoothing spline with a bias of 0.0147, respectively.

Table 11 Simulation studies for insertion by testing shadow to be only 1 base insertion, only 2 bases insertion, and up to 2 bases insertion.

Shadow	Insertion	Per-read ER	SRR037452	SRR037456	SRR002053	SRR032586
only 1 insertion	left side	Expected ER	0.1192	0.0886	0.1692	0.0028
		SRER	0.0125	0.0056	0.0204	0.0002
		SRER Bias	0.1067	0.0830	0.1487	0.0026
		EER-CS	0.1029	0.0581	0.0873	0.0022
		EER-CS Bias	0.0163	0.0305	0.0819	0.0006
		EER-RS	0.1068	0.0579	0.0854	0.0022
		EER-RS Bias	0.0124	0.0306	0.0838	0.0006
only 2 insertion	left side	Expected ER	0.1141	0.0856	0.1412	0.0027
		SRER	0.0122	0.0009	0.0109	0.0001
		SRER Bias	0.1019	0.0848	0.1303	0.0026
		EER-CS	0.0987	0.0490	0.0566	0.0015
		EER-CS Bias	0.0154	0.0366	0.0846	0.0012
		EER-RS	0.1091	0.0462	0.0612	0.0015
		EER-RS Bias	0.0051	0.0395	0.0799	0.0012
up to 2 insertion	left side	Expected ER	0.1930	0.1554	0.2682	0.0053
		SRER	0.0197	0.0083	0.0368	0.0003
		SRER Bias	0.1733	0.1471	0.2313	0.0050
		EER-CS	0.1689	0.1002	0.1452	0.0037
		EER-CS Bias	0.0241	0.0552	0.1230	0.0016
		EER-RS	0.1783	0.0962	0.1398	0.0037
		EER-RS Bias	0.0147	0.0593	0.1284	0.0016
only 1 insertion	right side	Expected ER	0.1532	0.1460	0.1501	0.0020
		SRER	0.0240	0.0182	0.0182	0.0002
		SRER Bias	0.1292	0.1278	0.1319	0.0018
		EER-CS	0.1217	0.1058	0.0805	0.0011
		EER-CS Bias	0.0315	0.0402	0.0695	0.0009
		EER-RS	0.1214	0.1056	0.0778	0.0011
		EER-RS Bias	0.0318	0.0404	0.0723	0.0009
only 2 insertion	right side	Expected ER	0.1573	0.1538	0.1892	0.0018
		SRER	0.0262	0.0043	0.0325	0.0001
		SRER Bias	0.1311	0.1495	0.1566	0.0016
		EER-CS	0.1217	0.1006	0.1188	0.0008
		EER-CS Bias	0.0356	0.0532	0.0704	0.0009
		EER-RS	0.1237	0.1011	0.1147	0.0008
		EER-RS Bias	0.0336	0.0527	0.0745	0.0009

up to 2 insertion	right side	Expected ER	0.2677	0.2537	0.2924	0.0035
		SRER	0.0554	0.0317	0.0494	0.0004
		SRER Bias	0.2123	0.2220	0.2430	0.0031
		EER-CS	0.2173	0.1841	0.1808	0.0019
		EER-CS Bias	0.0504	0.0696	0.1116	0.0016
		EER-RS	0.2170	0.1861	0.1770	0.0019
		EER-RS Bias	0.0507	0.0676	0.1154	0.0016
only 1 insertion	two sides	Expected ER	0.2245	0.2074	0.2385	0.0039
		SRER	0.0330	0.0237	0.0166	0.0003
		SRER Bias	0.1915	0.1838	0.2219	0.0036
		EER-CS	0.1737	0.1533	0.1082	0.0027
		EER-CS Bias	0.0509	0.0541	0.1303	0.0012
		EER-RS	0.1759	0.1532	0.1047	0.0027
		EER-RS Bias	0.0486	0.0542	0.1338	0.0012
only 2 insertion	two sides	Expected ER	0.2521	0.2316	0.3539	0.0053
		SRER	0.0590	0.0286	0.0349	0.0004
		SRER Bias	0.1930	0.2031	0.3190	0.0049
		EER-CS	0.2251	0.1706	0.1759	0.0030
		EER-CS Bias	0.0270	0.0610	0.1780	0.0023
		EER-RS	0.2310	0.1707	0.1711	0.0030
		EER-RS Bias	0.0211	0.0609	0.1828	0.0023
up to 2 insertion	two sides	Expected ER	0.3834	0.3577	0.4631	0.0089
		SRER	0.0873	0.0557	0.0498	0.0008
		SRER Bias	0.2961	0.3019	0.4133	0.0080
		EER-CS	0.3319	0.2783	0.2509	0.0056
		EER-CS Bias	0.0515	0.0794	0.2122	0.0032
		EER-RS	0.3321	0.2782	0.2467	0.0057
		EER-RS Bias	0.0513	0.0795	0.2164	0.0032

***Based on 1,000 replicates. The frequency-based simulation approach was applied. For each replicate, we considered the top 1,000 reads with the highest frequencies as the error-free reads and generated 1,000 pairs of error-free read counts and shadow counts.**

Expected ER: expected error rate in simulation studies

SRER: error rate estimated using shadow regression

SRER Bias: the absolute value of the difference between SRER and Expected ER

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_CS Bias: the absolute value of the difference between EER_CS and Expected ER

EER_RS: empirical error rate estimated using robust smoothing spline

EER_RS Bias: the absolute value of the difference between EER_RS and Expected ER

In order to know the error rates by varying shadow definitions, the real data analysis is performed. Table 12 shows the error rates in real data analysis obtained using shadow linear regression approach and empirical error rate approach in smoothing spline and robust smoothing spline when shadows are defined to be only one base different, only two bases different, only three bases different, up to two bases different, and up to three bases different from error-free reads in substitutions. The sample 100217 is a PhiX DNA sample, which were generated from the Center for Cancer Computational Biology at Dana-Farber and provided by Wang et al. (4). The PhiX 174 is an icosahedral virus containing a closed circular single-stranded DNA (144). The sample SRR002053 is a sample from ENCODE project transcriptome data with around 12 million reads. Sample SRR002053 is human mRNA sequence data from human cell line K562 running on Illumina 1G Genome Analyzer(145). The sample SRR037452 and sample SRR037456 are mRNA data from MAQC brain experiment 2, which contains about 12 million reads for each sample (145). The sample SRR032586 is mutation screening resequencing data from patient with X-linked mental retardation (XLMR) for mutations in 86 previously identified XLMR genes(145). Sample SRR032586 was applied to data analysis study after the outliers were removed in Table 12.

From the results in Table 12, the error rates of only one base different are generally greater than the ones of only two bases different. The error rates of only two bases different are generally greater than the ones of only three bases different. For example, in sample 100217, the estimated error rates by shadow regression were 0.0246, 0.0015, and 0.0004 for only one, only two, only three bases different in substitution, respectively. The error rates were 0.0289, 0.0030, and 0.0011 by using cubic smoothing spline, and 0.0289, 0.0030, and 0.0012 by using robust smoothing spline, for only one, only two, only three bases different in substitution, respectively.

Based on previous studies (*4, 146*), using shadows with up to two bases different provides accurate estimation without too much computational cost. Using shadow with up to three bases different may have higher possibility of identifying more errors, but it may also count polymorphism as sequence errors and take excessive computational cost. In addition, the error rates with up to two bases different from error-free reads in substitution are similar with the ones of up to three bases different. For example, in sample 100217, the estimated error rates were 0.0261 for up to two bases different, and 0.0251 for up to three bases different using shadow regression, and 0.0317 for up to two bases different, and 0.0322 for up to three bases different using cubic smoothing spline, and 0.0315 for up to two bases different and 0.0322 for up to three bases different using robust smoothing spline, respectively. Therefore, we confirmed with previous studies that using the shadow with up to two bases different is good enough without excessive computational cost.

Table 12. Real data application for substitution by varying shadow to be only 1 base different, only 2 bases different, only 3 bases different, up to 2 bases different, and up to 3 bases different.

Samples	Substitution ER Estimation	only 1	only2	only3	sub2	sub3
100217	SRER	0.0246	0.0015	0.0004	0.0261	0.0251
	EER-CS	0.0289	0.003	0.0011	0.0317	0.0322
	EER-RS	0.0289	0.003	0.0012	0.0315	0.0322
SRR002053	SRER	0.2832	0.2344	0.178	0.4134	0.269
	EER-CS	0.369	0.2945	0.2303	0.4469	0.4308
	EER-RS	0.3084	0.2868	0.2252	0.4453	0.4266
SRR032586	SRER	0.0707	0.0113	0.0054	0.0954	0.0696
	EER-CS	0.1199	0.0689	0.0659	0.1729	0.1353
	EER-RS	0.1193	0.0684	0.0663	0.1725	0.1386
SRR037456	SRER	0.1238	0.0548	0.0331	0.1657	0.1451
	EER-CS	0.1538	0.085	0.0624	0.2062	0.2413
	EER-RS	0.1605	0.1126	0.092	0.2162	0.2751

SRER: error rate estimated using shadow regression

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_RS: empirical error rate estimated using robust smoothing spline

Table 13 shows the error rates from real data analysis result in deletion. Different deletion error rates were investigated by defining the shadows to be only one base different, only two bases different, and up to two bases different in left side deletion, right side deletion, and two side deletion, respectively. For sample SRR037452, when the shadow is defined to be only one base different from error-free read in deletion, the error rates from left side deletion is 0.0121, 0.0386, and 0.0404 using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the shadow is defined to be only two bases deletion, the error rates of left side deletion for sample SRR037452 is 0.0037, 0.0185 and 0.0348 using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the shadow is defined to be up to two bases different from error-free reads in deletion, the error rates of left side deletion for sample SRR037452 is 0.0184, 0.0550 and 0.0782 using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. The right side deletion is the deletion adding the extra base to the right end side of the read when some base is deleted from the original sequence, given that all the sequences are with the same lengths. The right side deletion is relatively greater than the left side deletion in the samples studied. For sample SRR037452, when the shadow is defined to be only one base in deletion, the error rates of right side deletion is 0.0201, 0.0573, and 0.0687 by using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the shadow is defined to be only two bases different in deletion, the error rates of right side deletion for sample SRR037452 is 0.0318, 0.1118 and 0.1120 by using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the shadow is defined to be up to two bases deletion, the error rates of right side deletion for sample SRR037452 is 0.0680, 0.1559, and 0.1684 by using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. Another way of

deletion is two-side deletion. Suppose one or two bases are deleted from an original sequence read, given that all the sequences are with the same lengths, an extra base from one of the alleles (“A”, “T”, “G”, “C”) will be added to either left or right end side of the read. For sample SRR037452, the error rates for only one base in two-side deletion are 0.0288, 0.0829, and 0.0910 using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the deletion shadow is defined to be only two bases different from error-free reads, the error rates of two-side deletion for sample SRR037452 are 0.0439, 0.1635, and 0.1639 by using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the deletion shadow is defined to be up to two bases different from error-free sequences, the error rates of two-side deletion for sample SRR037452 are 0.0876, 0.2213, and 0.2269 by using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively.

Table 13. Real data application for deletion by varying shadow to be only 1 base deletion, only 2 bases deletion, and up to 2 bases deletion.

Shadow	Deletion	Per-read ER	SRR037452	SRR037456	SRR002053	SRR032586
only 1 deletion	left side	SRER	0.0121	0.0082	0.0307	0.0028
		EER-CS	0.0386	0.0352	0.0505	0.0136
		EER-RS	0.0404	0.0386	0.0519	0.0121
only 2 deletion	left side	SRER	0.0037	0.0023	0.0151	0.0002
		EER-CS	0.0185	0.0139	0.0294	-0.0277
		EER-RS	0.0348	0.0198	0.0381	-0.0223
up to 2 deletion	left side	SRER	0.0184	0.0124	0.0475	0.0030
		EER-CS	0.0550	0.0472	0.0761	0.0080
		EER-RS	0.0782	0.0587	0.0866	0.0078
only 1 deletion	right side	SRER	0.0201	0.0143	0.0674	0.0045
		EER-CS	0.0573	0.0557	0.1311	0.1224
		EER-RS	0.0687	0.0637	0.1332	0.1157
only 2 deletion	right side	SRER	0.0318	0.0365	0.1147	0.0035
		EER-CS	0.1118	0.1109	0.1834	0.0880
		EER-RS	0.1120	0.1126	0.1794	0.0835
up to 2 deletion	right side	SRER	0.0680	0.0596	0.1752	0.0140
		EER-CS	0.1559	0.1539	0.2670	0.1823
		EER-RS	0.1684	0.1634	0.2645	0.1778
only 1 deletion	two sides	SRER	0.0288	0.0217	0.0681	0.0052
		EER-CS	0.0829	0.0792	0.1434	0.1190
		EER-RS	0.0910	0.0861	0.1426	0.1181
only 2 deletion	two sides	SRER	0.0439	0.0482	0.0931	0.0045
		EER-CS	0.1635	0.1627	0.1991	0.0276
		EER-RS	0.1639	0.1623	0.1967	0.0266
up to 2 deletion	two sides	SRER	0.0876	0.0777	0.1558	0.0159
		EER-CS	0.2213	0.2158	0.2855	0.1406
		EER-RS	0.2269	0.2219	0.2816	0.1378

SRER: error rate estimated using shadow regression

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_RS: empirical error rate estimated using robust smoothing spline

Table 14 shows the insertion error rates from real data analysis by using shadow regression, cubic smoothing spline, robust smoothing spline and varying definition of insertion shadows to be only one base different, only two bases different, and up to two bases different from error-free sequences. The left side insertion is the insertion when some base is inserted from an original sequence, given that all the sequences are with the same length, an extra base is then removed from the right end side of the read. For sample SRR037452, when the shadow is defined to be only one base in insertion, the error rates of left side insertion is 0.0137, 0.1064, and 0.1267 by using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the insertion shadow is defined to be only two bases different from error-free reads, the error rates of left side insertion for sample SRR037452 is 0.0032, 0.0797, and 0.1367 by using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the shadow is defined to be up to two bases insertion, the error rates of left side insertion for sample SRR037452 is 0.0234, 0.1682, and 0.2209 using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. The right side insertion is the insertion when some base is inserted from an original sequence, given that all the sequences are with the same length, an extra base is removed from the left end side of the read. For sample SRR037452, when the shadow is defined to be only one base in insertion, the error rates of right side insertion is 0.0417, 0.1275, and 0.1472 using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the insertion shadow is defined to be only two bases difference, the error rates of right side insertion for sample SRR037452 is 0.0405, 0.1284, and 0.1534 using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the insertion shadow is defined to be up to two bases difference, the error rates of right side insertion for sample SRR037452 is 0.0847, 0.2254, and 0.2581 by using

shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. Another way of insertion is two-side insertion. Suppose one or two bases are inserted to a sequence read, given that all the sequences are with the same lengths, extra base(s) will be removed from either left or right end side of the read. For sample SRR037452, the error rates for only one base difference of two-side insertion approach are 0.0381, 0.1502, and 0.1631 by using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the shadow is defined to be only two bases difference, the error rates of two-side insertion for sample SRR037452 are 0.0045, 0.1008, and 0.1569 using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively. When the shadow is defined to be up to two bases difference, the error rates of two-side insertion for sample SRR037452 are 0.0463, 0.2222, and 0.2656 by using shadow regression, cubic smoothing spline, and robust smoothing spline, respectively.

Table 14. Real data application for insertion by varying shadow to be only 1 base insertion, only 2 bases insertion, and up to 2 bases insertion.

Shadow	Insertion	Per-read ER	SRR037452	SRR037456	SRR002053	SRR032586
only 1 insertion	left side	SRER	0.0137	0.0085	0.0539	0.0017
		EER-CS	0.1064	0.0604	0.1113	0.0332
		EER-RS	0.1267	0.0813	0.1123	0.0289
only 2 insertion	left side	SRER	0.0032	0.0035	0.0364	0.0024
		EER-CS	0.0797	0.0454	0.0884	-0.0022
		EER-RS	0.1367	0.0822	0.0407	-0.0004
up to 2 insertion	left side	SRER	0.0234	0.0137	0.0937	0.0039
		EER-CS	0.1682	0.0945	0.1774	0.0275
		EER-RS	0.2209	0.1484	0.1446	0.0279
only 1 insertion	right side	SRER	0.0417	0.0289	0.0538	0.0024
		EER-CS	0.1275	0.1157	0.1040	0.0235
		EER-RS	0.1472	0.1161	0.1045	0.0236
only 2 insertion	right side	SRER	0.0405	0.0267	0.0945	0.0007
		EER-CS	0.1284	0.1124	0.1494	0.0132
		EER-RS	0.1534	0.1389	0.1600	0.0144
up to 2 insertion	right side	SRER	0.0847	0.0592	0.1512	0.0056
		EER-CS	0.2254	0.2028	0.2202	0.0362
		EER-RS	0.2581	0.2218	0.2286	0.0393
only 1 insertion	two sides	SRER	0.0381	0.0248	0.0505	0.0025
		EER-CS	0.1502	0.1089	0.1237	0.0546
		EER-RS	0.1631	0.1238	0.1265	0.0470
only 2 insertion	two sides	SRER	0.0045	0.0033	0.0435	0.0027
		EER-CS	0.1008	0.0681	0.1131	0.0003
		EER-RS	0.1569	0.1052	-0.0045	0.0021
up to 2 insertion	two sides	SRER	0.0463	0.0349	0.0970	0.0043
		EER-CS	0.2222	0.1578	0.2050	0.0458
		EER-RS	0.2656	0.2021	0.1298	0.0449

SRER: error rate estimated using shadow regression

EER_CS: empirical error rate estimated using cubic smoothing spline

EER_RS: empirical error rate estimated using robust smoothing spline

3.4 Conclusion

The empirical error rate estimation approach for insertion and deletion do not require using a reference genome in estimating sequencing error rates and are also adjustable to various base error estimation for next generation sequencing data. We estimated the substitution error rates by varying different bases and confirm that using shadows of up to two base different is good enough for short read sequences without excessive computational cost. We performed simulation studies and investigated the performance of empirical error rate estimation approaches when defining shadows to be deletion and insertion with only one base different, only two bases different, and up to two bases different from error-free sequences for left side, right side, two sides approaches, respectively, and compared with the results using shadow regression approach. Under the simulation scenarios tested, the empirical error rate estimation approach using splines provided more accurate estimation for both insertion and deletion error rates with less bias than the shadow regression approach. The proposed empirical error rate approaches were also applied to real data analysis and compared with the shadow regression approach.

Chapter 4

Statistical Approaches for Processing and Analyzing

Microbiome Data

4. 1. Introduction

The human microbiome refers to the totality of all microbes consisting of prokaryotes (bacteria), archaea, fungi, and viruses in and on the human body (*147*). Microbiomes rarely live alone in nature as they function in a complex microbial community by processing energy and materials or constituting an extended human genome (*148, 149*). The human microbial partners carry out a number of metabolic reactions that are essential for human health, but are not encoded in the human genome (*150*)[150].

Advances in sequencing technology have allowed us to generate large amounts of microbiome data. Sequencing of the microbiome provides a great opportunity to assess the organismal and functional novelty to be discovered, which plays an essential role in human health and disease (*150*)[150]. The composition of the human microbiome varies as a result of one's health, diet, lifestyle, social interactions, early antibiotic therapy, genotype, and environmental chemicals

(*148, 151*). In addition, the microbiome produces beneficial compounds in our body such as vitamins and anti-inflammatories that our human genome cannot produce (*152*). The relationships between components of the microbiome are often referred to as commensal, where one partner benefits and the other is apparently unaffected, as opposed to mutualistic, where both partners experience increased fitness (*149*). Recent studies have shown the microbiome to be associated with complex diseases, including obesity (*51*), cardiovascular disease (*153*), type 2 diabetes (*51*), and colorectal cancer (*154*). Also, the gut microbiome was shown to modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders via circulating metabolites. Commensal bacteria were also shown to control the response of cancer to therapy by modulating the tumor microenvironment (*148*). This chapter discusses approaches for analyzing microbiome data, from the initial steps of processing raw sequencing data to the downstream analysis, including assessing community-level diversities and their association with outcomes of interest.

4.2 Methods in Human Microbiome Analysis

4.2.1 Next-generation Sequencing Methods for the Human Microbiome

Commonly used next-generation sequencing methods for analyzing the human microbiome include shotgun metagenomic sequencing and 16S ribosomal RNA (rRNA) sequencing (**49**)[49]. The shotgun metagenomic sequencing technique enables comprehensive sampling of all the genes over all organisms in a complex microbial sample; while the 16S rRNA sequencing approach focuses on identifying and comparing bacteria within a given sample (**49**)[49]. Over the last decade, 16S rRNA sequencing has become a popular technique to accurately identify bacterial isolates and discover novel bacteria in clinical studies of the microbiome (**155**). In particular, the 16S rRNA gene can be compared not only among all bacteria but also with the 16S rRNA gene of archeobacteria and the 18S rRNA gene of eukaryotes (**50**). The 16S rRNA gene can be used as a stable phylogenetic marker to study the lineage relationships in the sample (**156**). The 16S rRNA gene has been identified for a large number of strains and is universal in bacteria. According to previous studies (**50**), identifying the bacteria isolated by a sequencing technique has the potential to better identify strains that are poorly described, rarely isolated or biochemically aberrant. The 16S rRNA sequencing technique can identify novel pathogens and lead to the discovery of noncultured bacteria (**157, 158**). Most importantly, although microbiome data can be generated using multiple methods, 16S rRNA sequencing is more efficient and less expensive compared to other methods and, thus, is still the most popular approach for generating such data (**157, 158**). Therefore, we focus on data processing and analysis for microbiome data generated by the 16S rRNA sequencing approach.

4.2.1.1 Foregut Esophageal Adenocarcinoma Study

We use the dataset from a study of esophageal adenocarcinoma in the foregut (*159*) to illustrate data analysis and interpretation for a microbiome study. That study sought to uncover the associations between esophageal adenocarcinoma risk and the composition of the microbiome. The project studied the microbiome at multiple stages of disease development, from normal to inflammation, metaplasia, and neoplasia, and also used invasive endoscopic procedures to sample the microbiome deep inside the human body. Esophageal adenocarcinoma is a rare but aggressive cancer that has been increasing in incidence, especially among white males.

Esophageal adenocarcinoma develops in the distal esophagus in response to mucosal injury (*159*). Risk factors for esophageal adenocarcinoma include gastroesophageal reflux disease, cigarette smoking, obesity and low fruit or vegetable consumption, which together represent ~80% of the population attributable risk. The characteristics of the esophageal microbiome may facilitate the development of disease (*159*). The study evaluated 104 individuals (16 females and 88 males) and collected fecal and oral samples, as well as samples from the esophagus and stomach, for a total of 555 microbiome samples from different sites and different individuals.

The sequencing data for all the microbiome samples were generated using the Roche 454 GS FLX Titanium platform (*160*) and Illumina HiSeq 2000 platform (*161*). For the purpose of the illustration, we randomly selected 40 oral samples, among which 20 samples were from healthy subjects and the other 20 samples were from the esophageal adenocarcinoma patients, for which the sequencing data were generated using the Roche 454 GS FLX Titanium platform.

4.2.1.2 Processing Raw Microbiome Sequencing Data

Several pipelines are available to process raw microbiome data with millions of sequencing reads before assessing the association between microbial community profiles and outcomes of interest, among which the most popular pipelines are mothur (*162*) and Quantitative Insights into Microbial Ecology (QIIME, pronounced “chime”) (*148, 162, 163*). Mothur is an open-source, platform-independent, community-supported software that describes and compares microbiome communities (*162*). QIIME is a free and open-source analytic tool to process high-dimensional sequencing data (*163*). It has been shown that QIIME and mothur have similar statistical capabilities and user freedom; however, QIIME is more user-friendly than mothur as it is easier to understand and interpret the commands in QIIME than those in mothur (*164*). QIIME is available for analyzing whole metagenomic shotgun sequencing data, while mothur does not have this function (*164*). QIIME is preferred for analyzing large datasets due to the efficiency gain in computational time and the user-friendly interface. More importantly, the taxonomic summary table generated using QIIME is more easily adapted to the downstream analyses in statistical packages such as R (*164*). Therefore, in this study, we use QIIME to process the raw microbiome sequencing data and generate the operational taxonomic units (OTUs) for the downstream microbiome data analysis. QIIME can process raw sequencing data generated from different platforms, such as Illumina and 454 platforms (*163*).

We describe the data processing procedure for data generated using the 454 platform. Typically the raw sequencing data are available in fasta and fastq formats. We review the commands for data processing, quality checking and OTU extraction based on these two data formats. In addition to the sequencing read files, a metadata mapping file is required for the purpose of data

processing. The metadata mapping file usually contains the following information: the name of each sample, barcode sequence for each sample, primer sequence used to amplify the sample, and a descriptive column (for example, health status, sampling site, or which medications a patient was taking at the time of sampling) (*163*).

Figure 4 shows the basic procedure for processing the raw microbiome sequencing data.

Specifically, we need to perform the steps described below, including demultiplexing, removing barcodes and primers and quality control checks, before generating OTUs for downstream analysis. The commands for processing the raw microbiome sequencing data and extracting the OTUs using QIIME are provided in the Methods section.

- (a) Demultiplexing: If multiple samples are pooled in one single lane by using the sample index or barcode adapters, then the step of demultiplexing is needed to divide one file of sequencing reads into separate files, where one file corresponds to one sample (*163*).
- (b) Removing barcodes: High-throughput sequencing allows multiple different independent samples to be sequenced during a single run. In this situation, each sample is tagged with a unique sequence that is referred to as a barcode. The QIIME software utilizes barcodes to demultiplex and assign sequences to each sample (*165, 166*). However, the barcodes incorporated in the sequencing reads need to be removed because the reads containing barcodes can lead to a mismatching issue in the downstream procedures (*167*).
- (c) Removing primers: The primer is a starting point for DNA synthesis, which consists of a short nucleic acid sequence. For living organisms, primers are short strands of RNA that

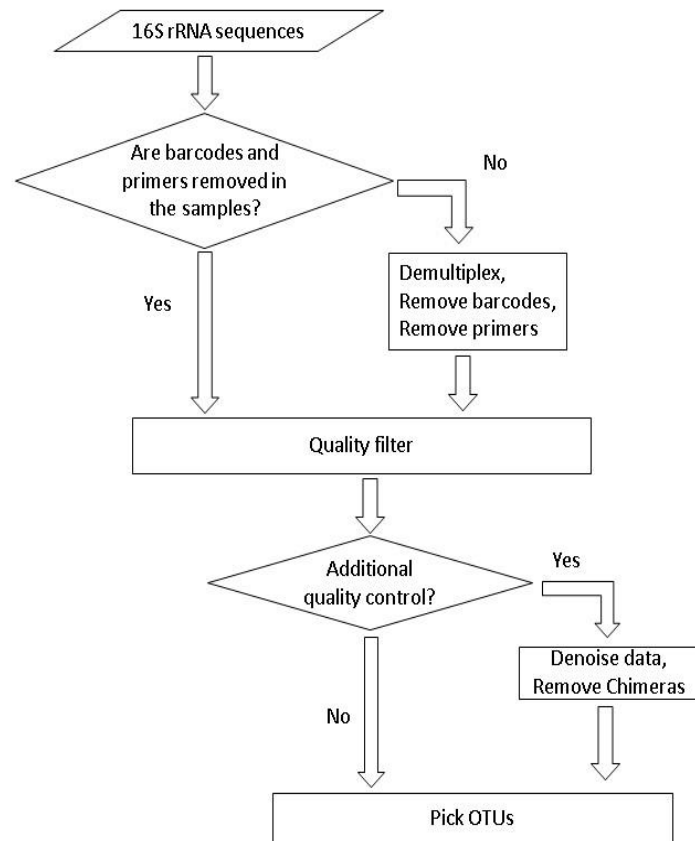
enzymes synthesize as a type of RNA polymerase before DNA replication occurs (**168**).

It is a necessary step for the primer synthesis as DNA polymerases can only attach new DNA nucleotides to an existing strand of nucleotides (**168**). However, to avoid a mismatching issue when converting data into OTUs, it is necessary to remove the primers. We note that for some sequencing data downloaded from dbGaP, such as the data from the foregut esophageal adenocarcinoma study used here, the barcodes and primers have been removed before the investigators uploaded the data. In this situation, we do not need to conduct these two steps of removing barcodes and primers.

(d) Quality filtering: The quality filtering step is used to remove any low-quality or ambiguous reads according to the quality scores provided in the sequencing. In the 454 sequencing pipeline, the quality score assigned to each called base is the flowgram values rounded to the nearest integer in the *Phred* format. The *Phred* score is the probability that each nucleotide was incorrectly read (**169**). Based on the Phred score and a pre-specified threshold parameter, we can remove sequences that do not satisfy the desired quality (**165**).

(e) Additional quality-control procedures: Sequencing platforms can produce characteristic sequencing errors, such as imprecise signals for longer homopolymer runs. The de-noising procedure removes such errors and thereby improves the accuracy of the OTU generation. In addition, the raw sequencing data may contain chimeras generated by abnormal amplification during the process of polymerase chain reaction (PCR); these chimeras also need to be removed before generating OTUs.

Figure 4. Flowchart of basic steps in handling raw microbiome sequencing data



4.3 Diversity Measures of the Microbiome

As stated previously, different species of microbes live as interacting communities, and the abundance and diversity of the microbiome is essential for maintaining human health (*148*). Measures of diversity, richness, and abundance are often used to summarize microbiome communities and compare them in different groups (*148*), among which the alpha (α) and beta (β) diversities are most commonly used. The alpha diversity measure shows the spatial patterns of biological diversity within a sample; thus, it evaluates the within-sample diversity (*170*). The beta diversity measure assesses the change in species composition between samples (*170*), evaluating the between-sample diversity. We describe several alpha and beta diversity measures commonly used in studies of the microbiome, with illustrations in the Methods section using the R package “phyloseq” (*171*), based on the data from the foregut esophageal adenocarcinoma study described previously.

4.3.1 The α Diversity of the Microbiome

On the level of the microbial community, the clustering of OTUs allows us to measure the within-sample diversity (α diversity) of the microbiome. The observed richness, which is the simplest measure, refers to the number of species within a sample (*172*), but does not incorporate information about their relative abundance. Here, we introduce several commonly used measures for assessing α diversity.

4.3.1.1 Shannon diversity index

The Shannon diversity index is a popular measure of species diversity that measures the entropy and uncertainty of the sampling outcome (56). The Shannon diversity index is calculated as follows: Suppose n is the total number of species, and p_i is the proportion of species i relative to the total number of species in a community (173), then the Shannon diversity index is $H = -\sum_{i=1}^n (p_i \times \ln p_i)$. Higher values of the Shannon diversity index represent more diversity within a community; a value of 0 means the community has only one species. The Shannon diversity index is nonparametric, which allows for the simultaneous measurement of a richness estimation from heterogeneous samples (58) and takes into account both the relative abundance and total number of species in a microbiome community (173-175).

4.3.1.2 Simpson's diversity index

The Shannon index emphasizes the richness component, while Simpson's index stresses the evenness component (57). Simpson's diversity index describes the probability that two randomly drawn reads from a sample are from the same taxon (56). Simpson's diversity index is given as $\lambda = \sum_{i=1}^n p_i^2$, where n is the total number of species, and p_i is the proportion of species i relative to the total number of species in a community (173). Simpson's diversity index is between 0 and 1, where 0 represents infinite diversity and 1 implies no diversity. The higher the value of Simpson's diversity index, the lower the diversity within the community.

For easier interpretation, the inverse Simpson's diversity index, calculated as $1/\lambda$, estimates the probability that two randomly chosen reads from a sample of the given community come from different taxa. A higher value for the inverse Simpson's diversity index represents greater diversity. A previous study showed that Simpson's diversity index has a small standard deviation, and for moderately large samples the bias is small (*173*). However, the inverse Simpson's diversity index is biased when estimating numerous species that have low abundance within a community (*56*).

4.3.1.3 Fisher's α diversity index

Fisher's α diversity index describes the relationship between the number of species and the number of individuals of the corresponding species by logarithmic distribution. Compared with the Shannon index and Simpson's index, Fisher's α diversity index is not influenced by the sample size and is less affected by the abundance of the most common species. Fisher's α index depends more on the number of species of intermediate abundance (*174, 176*).

Fisher's α index is used to iteratively fit the logarithmic-series distribution (*177, 178*). According to Fisher's theory, any population should have a constant value α for samples of any size taken from it under identical conditions (*179*). Suppose S is the total number of species in the sample, and n is the total number of individuals. The relative abundance is derived from the logarithmic-series formed as $\alpha x, \alpha x^2/2, \alpha x^3/3, \dots, \alpha x^n/n$, where the successive terms represent the

number of species predicted to have the corresponding $1, 2, 3, \dots, n$ individuals. The parameter x is estimated by an iterative solution of $S/N = (1 - x)/x(-\ln(1 - x))$, and the parameter α is the log-series index of diversity. The two parameters α and n summarize the distribution completely and are related by $S = \alpha \ln(1 + n/\alpha)$. The α index can also be obtained by $\alpha = n(1 - x)/x$ (84, 178, 179). A higher value of Fisher's α index represents more diversity within the sample. The estimation of Fisher's α index gives an unbiased estimation of diversity (179). However, this approach is computationally intensive when the sample size is large.

4.3.1.4 Chao1 α diversity index

Because the abundance in microbiome OTU data is often low for many species, the Chao1 α diversity index is defined as the number of species represented by just one or two individuals (58). The Chao1 α diversity index stresses the “singletons” (i.e., the number of species with only a single occurrence in the sample) and “doubletons” (i.e., the number of species with exactly two occurrences in the sample) and is calculated as follows: Suppose S_{obs} is the number of species for the sample, A_1 is the number of singletons, and A_2 is the number of doubletons, then

$Chao1_\alpha = S_{obs} + \frac{A_1^2}{2A_2}$. A higher estimated value represents more diversity within the sample.

The Chao1 α diversity index is particularly useful for microbiome datasets skewed to low abundance classes; while a drawback of the Chao1 α diversity index is that when the sample size is low, it underestimates the true value (180).

4.3.1.5 Abundance-based coverage estimator

The abundance-based coverage estimator (ACE) (*181, 182*) measures the α diversity by separating the species into two groups: an abundant species group and a rare species group.

Given a prespecified threshold k , the abundant species are those with more than k taxa, and the rare species are those with less than or equal to k taxa. A value of $k = 10$ is suggested based on empirical evidence from previous studies (*182, 183*). The calculation of the ACE is described below (*182*).

Given that f_1 is the number of singleton species, f_k is the number of species with exactly k taxa, and S_{obs} is the total number of species in the sample, the number of rare species is described as $S_{rare} = \sum_{k=1}^{10} f_k$, and the number of abundant species is denoted as $S_{abund} = \sum_{k=11}^{S_{obs}} f_k$, assuming the value of 10 is the threshold. The total number of taxa in the rare species is $n_{rare} = \sum_{k=1}^{10} k f_k$, and $C_{ACE} = 1 - \frac{f_1}{n_{rare}}$ is the proportion of all non-singleton taxa in the rare species.

Let $\gamma_{ACE}^2 = \max \left[\frac{S_{rare}}{C_{ACE}} \frac{\sum_{k=1}^{10} k(k-1)f_k}{(N_{rare})(N_{rare}-1)} - 1, 0 \right]$ be the coefficient of variation. The ACE of the

species richness function is then $ACE = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{f_1}{C_{ACE}} \gamma_{ACE}^2$. A higher ACE value

implies more diversity within the sample. The ACE takes into account both the abundant and rare species. However, the ACE underestimates the true richness at low sample sizes (*180*).

4.3.2 Microbiome β Diversity

The β diversity measure represents the species community composition between two or more samples. β diversity also refers to the turnover of the community composition from place to place or from time to time (*184, 185*). The species turnover is the change in species composition resulting from species extinction and immigration (*186*)[186]. Various metrics can be used to measure the β diversity, such as the Manhattan index (*187*), Canberra index (*188, 189*), Jaccard distance (*190*), Kulczynski distance (*190*), binomial deviance (*191*), Morisita (*192*), Horn's index (*192*), Bray-Curtis (*59*), Cao dissimilarity index (*60*), and UniFrac (*61*).

The Manhattan index is the sum of absolute value of difference in the corresponding taxa counts between two samples, which is simple to utilize, whereas it may be sensitive to outliers and the lack of information about relative abundance. (*193*) The Canberra index overcomes this limitation and uses the proportion of difference to estimate diversity. Canberra distance is a weighted version of Manhattan distance (*194, 195*). However, according to Emran & Ye(2002) (*194*), when noise level increases, the performance in Canberra distance may drop as it is sensitive to random error. Another index is the Jaccard distance (*196*) measuring the dissimilarity between finite sample sets. Jaccard distance uses the size of the intersection divided by the size of the union of sample sets and measures the proportion of units not shared by the two observed samples. (*197*) The Kulczynski distance gives both species the same influence and improve the dominance of common absences from Jaccard distance. However, there is a limitation in Kulczynski distance that if the two samples are very disparate in size and if all the elements on the smaller sample appear in the larger, then the minimum value of the coefficient is 0.5 in which the coefficient's value is very unsatisfactory. Anderson & Millar (2004) (*198*)

proposed binomial deviance as a dissimilarity measure for the abundance data based on likelihood theory. This measure assumes the null hypothesis that two samples do not differ in composition and relative abundance of species. That is, for each species, they expect that half of the counts will fall into each transect. (198) Thus, a useful measure of dissimilarity between two transects is the sum of these deviances across all species. To take account of the fact that different species may be varying on different scales, a scale-invariance measure may be obtained by considering this quantity on a per-observation basis. The null hypothesis for this metric is to check if the two compared communities are equal, which can be quantified by using the chi-square test statistics. It should be able to handle variable sample sizes. The index does not have a fixed upper limit, but can vary among sites with no shared species. More details and discussions could be found in Anderson & Millar (2004) (198). Note that this binomial deviance dissimilarity is an improvement metric on the Bray-Curtis measure in terms of the likelihood theory. Here, the chi-square test statistic is equivalent to the dissimilarity measure known as the coefficient of divergence (Clark, 1952) (199). Hence, the binomial deviance has similar properties to the coefficient of divergence for use with species abundance data (198). Morisita proposed an index of similarity between communities. The Morisita metric is a statistical measure of dispersion of individuals in a population, and its extension (200) (Horn-Morisita variant) can be used to handle any abundance data. This is interpreted as the probability where two samples drawn randomly from two populations will belong to the same species, relative to the probability of randomly drawing two individuals of the same species alone (201). The Morisita's measure is useful as an empirical measure though the probability interpretation is only rigorous when particular species counts are very large. Horn's index is also an empirical measuring approach that utilizes different scaling factors based on Morisita metric. Horn index

improves from the Morisita's index that the upper limit is exactly 1, which makes it easier to interpret (201). Horn's index can also apply for clinical study and has been shown as a simple, useful, and aggregate scoring measure that enables combining disease severity and underlying comorbid conditions of the patient together (202). Specifically, we investigated the three widely used β diversity approaches in details, which are Bray-Curtis, Cao, and UniFrac.

4.3.2.1 Bray-Curtis dissimilarity index

The Bray-Curtis dissimilarity index measures the distance between two microbiome samples A and B by accounting for the abundance information (59). This approach does not require phylogenetic information. Given that n_{Aj} and n_{Bj} are the counts of the taxa in species j in samples A and B , and N_A and N_B are the total counts of all taxa in samples A and B , the Bray-Curtis dissimilarity index to measure the beta diversity between samples A and B is defined as (148)

$$\beta_{Bray-Curtis} = \sum_{j=1}^p \frac{|n_{Aj} - n_{Bj}|}{(N_A + N_B)},$$

where p is the number of taxa in the samples. The Bray-Curtis index varies between 0 and 1. If the two microbiome samples A and B are identical in composition, then the index is 0 (i.e., coincidence). If there are no species in common between the two samples, then the index is 1 (i.e., complementarity) (203).

One of the advantages of the Bray-Curtis index is the explicit meaning in the definition. The other advantage is the relative invariant characteristic that the relative values of a set of resemblances between two samples are not affected by a simple multiplicative scaling change (203). Also, the index is independent of joint absence, that is, the resemblance between two samples is not affected by the exclusion or inclusion of taxa that are absent (203). On the other hand, a common criticism of the Bray-Curtis index is its increasingly erratic behavior as values within samples become vanishingly sparse. When there are no taxa in both samples, the Bray-Curtis index does not work because the denominator is zero. Furthermore, even if the taxa in the two samples are from different species, the two samples, which are not empty but nearly so, are considered similar when using the Bray-Curtis index (203).

4.3.2.2 Cao dissimilarity index

To overcome the shortcomings of the traditional measures for β diversity, such as the Bray-Curtis index, Cao et al. (60) proposed a dissimilarity-similarity measure to respond to all of the types of variation with minimum bias in the weighting. The new dissimilarity measure, CY dissimilarity measure (CYd), is calculated as (60)

$$CYd = \frac{1}{N} \sum \left(\frac{(n_{Aj} + n_{Bj}) \log_{10} \left(\frac{n_{Aj} + n_{Bj}}{2} \right) - n_{Aj} \log_{10} n_{Bj} - n_{Bj} \log_{10} n_{Aj}}{n_{Aj} + n_{Bj}} \right),$$

where n_{Aj} is the counts of taxa for species j in sample A ; n_{Bj} is the counts of taxa for species j in sample B , and N is the total number of species present in the two samples.

In this formula, the numerator summand $(n_{Aj} + n_{Bj}) \log_{10} \left(\frac{n_{Aj} + n_{Bj}}{2} \right) - n_{Aj} \log_{10} n_{Bj} - n_{Bj} \log_{10} n_{Aj}$ is used to measure the dissimilarity between two samples. Caution should be taken when $n_{Aj} = 0$ or $n_{Bj} = 0$; a small constant value (e.g., 0.1) can be assigned to avoid a mathematical paradox. In this way, the measure can sensitively respond to all significant variations and weight them with minimum bias (60). The denominator $n_{Aj} + n_{Bj}$ is used to eliminate the effect of absolute abundance of species. The $1/N$ term is used to yield an average value and reduces the effect from the monotonicity of the total dissimilarity measure with respect to the species number. If the Cao dissimilarity index is 0, then the two samples are identical overall. This metric ranges from 0 to ∞ , with no fixed upper limit. A higher value of the Cao index means more dissimilarity between the samples. The Cao index (60) is considered to have a minimal bias in cases of high beta diversity and variable sampling intensity. This approach does not account for the phylogenetic information.

4.3.2.3 Unique fraction measures

The unique fraction (UniFrac) measures the difference between microbial samples by incorporating the phylogenetic information. The phylogenetic distance is measured between sets of taxa as the fraction of branch length in the phylogenetic tree that leads to descendants from either sample *A* or *B*, but not both (61). The UniFrac distance considers whether an OTU is present or absent in a sample but not the abundance information (204). To account for the corresponding phylogenetic information, we have a rooted phylogenetic tree for which n is the

total number of branches, h_k is the length for branch k , and p_{Ak} and p_{Bk} are the taxa proportions descending from branch k for samples A and B , respectively. The unweighted UniFrac metric is then defined as (148)

$$d^U = \sum_{k=1}^n \frac{h_k |I(p_{Ak} > 0) - I(p_{Bk} > 0)|}{\sum_{k=1}^n h_k}$$

where $I(\cdot)$ is the indicator function that is used to indicate the presence or absence of a species from branch k with branch length h_k .

To account for the abundance information of taxa sets, a weighted UniFrac was proposed by weighting the branch lengths with differences in abundance as (148)

$$d^W = \frac{\sum_{k=1}^n h_k |p_{Ak} - p_{Bk}|}{\sum_{k=1}^n h_k (p_{Ak} + p_{Bk})}$$

where p_{Ak} and p_{Bk} are the taxa proportions descending from branch k for samples A and B , respectively, and h_k is the branch length. The weighted UniFrac not only detects changes in the number of sequences present for each lineage, but also detects changes in which taxa are present (205). The weighted UniFrac can be normalized by the average distance for members in the two samples to the root, which corrects the unequal sampling effort or different evolutionary rates between taxa (204).

The advantage of the Unifrac measure is that it can be used to compare multiple communities simultaneously, to determine whether there is a significant difference between communities. The

UniFrac measures, which utilize phylogenetic information, are more powerful than the other distance measures by exploiting the degree of divergence between different sequences (61). Also, the UniFrac measures are applicable for integrating sequencing data from multiple communities or studies, which makes it suitable for large-scale comparisons between environments. When the sample size is large, UniFrac measures can be robust to similar samples (61). The unweighted UniFrac has been shown to have superior power compared with the weighted UniFrac (206). A drawback is that the Unifrac measures can be computationally expensive. Meanwhile, the UniFrac measures do not account for the evenness component, so they cannot address how much of the observed community is attributable to the phylogenetic tree (61, 207).

4.4 Association between Microbiome Diversity and Outcome

After assessing the microbiome community measures, such as α and β diversities, interest may lie in evaluating the potential association between these measures and outcomes of interest. Specifically, to assess the impact of a difference in the α diversity on the outcome, the α diversity can simply be used as a covariate in a regression model, such as a linear or logistic regression model (148).

To assess the association between the β diversity and outcomes, the most commonly used approach is the permutation-based nonparametric multivariate analysis of variance using distance matrices (PERMANOVA) (208). PERMANOVA is a distribution-free test that measures the overall difference for multiple responses based on permutation tests and partitions a symmetric distance matrix based on linear models. The PERMANOVA approach cannot incorporate other covariates in the test, and may be confounded by dispersion effects (148, 208).

4.4.1 Variable selection approach with sparse Dirichlet multinomial regression

One can also estimate the associated covariates for taxa counts using Chen & Li (2013) (209) approach, which uses the variable selection approach with sparse Dirichlet multinomial regression to estimate the association between the microbiome composition and environmental covariates. This approach not only accounts for the overdispersion in taxa counts, but also overcomes the limitation of losing power in multiple testing when the number of covariates is large. The penalized likelihood approach imposes a sparse group l_1 penalty to encourage both group level and within group sparsity. Given that there are q bacteria taxa with corresponding counts $Y = (Y_1, Y_2, \dots, Y_q)$ as random variables. Suppose the observed counts are $y =$

(y_1, y_2, \dots, y_q) , and $y_+ = \sum_{j=1}^q y_j$. The underlying species proportions are $\phi = (\phi_1, \phi_2, \dots, \phi_q)$.

The probability function is given as $f_M(y_1, y_2, \dots, y_q; \phi) = \binom{y_+}{y} \prod_{j=1}^q \phi_j^{y_j}$. The counts

marginally follow a Dirichlet-multinomial (DM) distribution as $f_{DM}(y_1, y_2, \dots, y_q; \gamma) =$

$\binom{y_+}{y} \frac{\Gamma(y_+ + 1) \Gamma(\gamma_+)}{\Gamma(y_+ + \gamma_+)} \prod_{j=1}^q \frac{\Gamma(\gamma_j + y_j)}{\Gamma(\gamma_j) \Gamma(y_j + 1)}$, where $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_q)$ are positive parameters, and $\gamma_+ =$

$\sum_{j=1}^q \gamma_j$. (209). To incorporate covariate effect vector $X = (x_1, x_2, \dots, x_p)$, it is under the assumption that the parameters γ_j in the DM model depend on covariates through a log-linear model that $\gamma_j(X) = \exp(\alpha_j + \sum_{k=1}^p \beta_{jk} x_k)$, where β_{jk} is the coefficient effect on the j th taxon of the k th covariate. (45, 209) The penalty term of variable selection for sparse Dirichlet-multinomial regression is $pl(\beta; Y, X, \lambda_1, \lambda_2) = -l(\beta; Y, X) + \lambda_1 \sum_{k=1}^p \|\beta_k\|_2 + \lambda_2 \sum_{k=1}^p \|\beta_k\|_1$, where $l(\beta; Y, X)$ is log-likelihood function, λ_1 and λ_2 are tuning parameters and $\|\beta_k\|_1 = \sum_{j=1}^q |\beta_{jk}|$ is the l_1 norm, $\|\beta_k\|_2 = \sqrt{\sum_{j=1}^q \beta_{jk}^2}$ is the group l_1 norm of coefficient vector β_k , respectively. (209)

This approach has limitation that it does not incorporate phylogenetic information. One potential future research topic for extension of this approach is how to incorporate the phylogenetic tree information into the DM regression model for more accurately estimation (45).

4.4.2 Microbiome Regression Based Kernel Association Test

Zhao et al. (2015) (210) proposed the microbiome regression-based kernel association test (MiRKAT), which extend the association study that directly regresses the outcome on both covariate information and semi-parametric kernel machine regression. (210). Compared to PERMANOVA, MiRKAT is more computationally efficient and can adjust for other covariates when assessing the association between outcomes of interest and microbiome profiles.

The kernel in MiRKAT is capable of incorporating phylogenetic distance information, which enables easy covariate adjustment and extension to alternative outcomes including survival outcomes into MiRKAT modeling. This approach tests association through p-value by applying a variance component score statistic. MiRKAT provides fast computation and also regresses outcome on the whole microbiome pairwise similarity profiles. The framework of MiRKAT approach is as the following. (210)

Suppose there are n samples, and for the i^{th} subject, y_i is the corresponding outcome. The abundances of all OTUs for individual i is $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{ip})'$, where p is the total number of OTUs. The covariate for individual i is $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{im})'$. The goal is to test the association between outcome and microbial similarity profiles by adjusting covariates \mathbf{X} . For continuous outcome variable, the linear kernel machine model is $y_i = \beta_0 + \beta' \mathbf{X}_i + f(\mathbf{Z}_i) + \varepsilon_i$; while for binary outcome variable, the logistic kernel machine model is $\text{logit}(P(y_i = 1)) = \beta_0 + \beta' \mathbf{X}_i + f(\mathbf{Z}_i)$, where β_0 is the intercept and $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_m]'$ is the coefficients of covariates. ε_i is the error term with mean 0 and variance σ^2 for continuous phenotypes. Under the kernel machine regression assumption, $f(\mathbf{Z}_i)$ is from a reproducing kernel Hilbert space, H_k , generated from a positive definite kernel function, $K(\cdot, \cdot)$, such that $f(\mathbf{Z}_i) = \sum_{i'=1}^n \alpha_{i'} K(\mathbf{Z}_i, \mathbf{Z}_{i'})$. Specifically, the kernel matrix can be constructed to be the transformation metrics of phylogenetic or taxonomic distance as $\mathbf{K} = -\frac{1}{2}(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})\mathbf{D}^2(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})$, where $\mathbf{D} = [d_{ij}]$ is the pairwise distance matrix such as Bray Curtis distance or weighted or unweighted UniFrac dissimilarity, \mathbf{I} is the identity matrix, $\mathbf{1}$ is the vector of ones. The estimation of coefficients β and $f(\mathbf{Z})$ is to maximize the penalized log-likelihood of $pl(f, \beta) = \sum_{i=1}^n \log L(f, \beta; y_i, x_i, z_i) - \frac{1}{2} \lambda \|$

$f \|_{H_k}^2 = \sum_{i=1}^n \log L(f, \beta; y_i, x_i, z_i) - \frac{1}{2} \lambda \alpha' K \alpha$. In order to test the association between outcome and microbiome profiles, MiRKAT approach uses variance component score test. The score statistic is $Q = \frac{1}{2\phi} (\mathbf{y} - \widehat{\mathbf{y}}_0)' \mathbf{K} (\mathbf{y} - \widehat{\mathbf{y}}_0)$, where $\widehat{\mathbf{y}}_0$ is the predicted mean of \mathbf{y} under null hypothesis that there is no association between microbiome profiles and outcome as $f(\mathbf{Z}) = 0$. ϕ is the dispersion parameter, where $\phi = \widehat{\sigma}_0^2$ and $\widehat{\sigma}_0^2$ is the estimated residual variance under null hypothesis for linear kernel machine regression, and $\phi = 1$ for logistic kernel machine regression. (210)

An essential advantage of the score test is that it only requires fitting null model and allows for fast, supervised, distance-based association testing under regression framework which permits controls for potential confounding effect. Better kernels chosen improves statistical power. Even if a poor kernel is chosen, the test is still statistically valid. Another advantage of MiRKAT enables to simultaneously consider multiple distance and dissimilarity metrics. An alternative approach for the MiRKAT is to implement a weighted combination of multiple kernels.

This approach can be implemented using the R package “MiRKAT.” (210)

4.5 Challenges in Analyzing Microbiome Data

This chapter covers relatively basic strategies for analyzing microbiome data, with the use of the community-level summary measures (e.g., α and β diversities). However, analyzing microbiome data can be challenging. First, there are challenges in data normalization and taxonomic

abundance estimation (*148*). For a single sample with high-throughput sequencing reads, the big data also present challenges for correct and full assembly and effective computing (*69*). Moreover, there are challenges to analyzing microbiome data. In particular, the relative abundances of species for one sample are usually quantified as a vector of the compositional data, which is calculated by dividing the observed count in the taxa or OTUs by the total reads in the sample, resulting in a vector of proportions residing in a nonnegative simplex. It is not straightforward to model the joint distribution of such data. The Dirichlet class of distributions is the most commonly used distribution to model such data (*148, 211, 212*). Alternatively, the compositional data can be transformed before conducting the analysis, which may include using square root transformation, additive log-ratio transformation or centered log-ratio transformation (*148, 213*). Shi et al. (*214*) presented linear regression models for compositional microbiome data analysis, using a set of linear constraints on the regression coefficients and estimating the regression coefficients by using a penalized estimation procedure.

However, these methods might not be able to address the complexity of microbiome data, especially the sparsity of the compositional data that involves a large number of zeros, because most OTUs have extremely low abundance and a large proportion of OTUs occur only once (possibly resulting from sequencing error) (*148*). Such a sparsity property increases the difficulty in analyzing the composition of the microbiome. For example, the commonly used statistical approaches, such as regression models and correlation measures, might not be directly applicable to such data. Moreover, the compositional data of the microbiome are intrinsically high dimensional, which further compounds the challenge for analyzing such data.

This chapter does not cover the analysis of longitudinal microbiome data, which are commonly observed in current studies of the human microbiome. Compared to cross-sectional analysis, longitudinal data analysis allows the relative abundances for a single individual to vary over time and utilizes dynamic modeling to understand the dependencies between species (214). Such longitudinal data are important when assessing microbial changes throughout treatment (e.g., during chemotherapy) and the impact of such changes on long-term outcomes of interest (e.g. overall survivorship). There are challenges and obstacles for such analyses (214): (i) the correlation between the abundances of species does not imply that these species directly interact, (ii) the constraint on the sum of relative abundance makes it difficult to perform parameter inference in longitudinal models, and (iii) errors due to experimental uncertainty or biased inferences of the interaction of species may occur for the longitudinal model. Fisher & Mehta (214) proposed the approach of learning interactions from microbial time series (LIMITS) for analyzing longitudinal microbiome data. The LIMITS approach uses a sparse linear regression method with bootstrap aggregation to infer the discrete time Lotka-Volterra model for microbial dynamics, which allows this approach to overcome the three aforementioned obstacles (214). Alternative methods for analyzing longitudinal microbiome data include using a Bayesian network (215) and a two-part mixed effects model (216).

In the next section, we provide commands for processing the raw microbiome sequencing data and extracting the OTUs, and demonstrate how to assess alpha diversities, beta diversities, and evaluate the association between diversities and outcomes of interest using a regression model, PERMANOVA and MiRKAT, based on the data from the foregut esophageal adenocarcinoma study described previously.

4.6. Processing Methods and Results

We provided commands in QIIME (*163*) for processing raw microbiome sequencing data and generating an OTU table. We used the “phyloseq” package in the R program (*171*) to assess the α and β diversities, the “vegan” package in R to conduct the PERMANOVA test, and “MiRKAT” package in R to conduct the MiRKAT test (*217*).

4.6.1 Processes to Convert Raw Data into the OTU Table

4.6.1.1 Demultiplexing data

1. For data in fasta format

The command `demultiplex_fasta.py` can demultiplex sequences from one input fasta file by using barcodes and/or data from the fasta labels provided in a metadata mapping file. Note that the information for the barcode is necessary for the purpose of demultiplexing.

Input data files: (1) read.fasta file, (2) metadata mapping file

Output files: directory of output files, including `seqs.fna`, `demultiplexed_sequences.log`, `seqs.qual`, and `seqs_not_assigned.fna`

```
demultiplex_fasta.py -f [read.fasta file] -m [metadata mapping file] -o [output directory  
name]
```

2. For data in fastq format

The function of `split_libraries_fastq.py` can demultiplex sequences from a data file in the fastq format in which the barcodes and sequences are stored in two separate fastq files (**163**).

Input data files: (1) read.fastq file, (2) barcode fastq file

Output files: directory of output files, including seqs.fna, split_libraries_log.txt, histograms.txt

```
split_libraries_fastq.py -i [read.fastq file] -b [barcode fastq file] -o [output directory name]
```

4.6.1.2 Removing barcodes

1. The barcodes can be removed when demultiplexing is conducted. Specifically, for the files in the fasta format, `demultiplex_fasta.py` removes the barcodes for the output sequences as a default.

2. For data in the fastq format, the barcode extraction command can be used to remove barcodes.

Input files: read.fastq file

Output files: directory of output files

```
extract_barcode.py -f [read.fastq file] -c [input type of barcode] --bc1_len [barcode length] -o [output directory name]
```

4.6.1.3 Removing primers

1. The primers can be removed by using a Perl script if the primers are known. This can be used for data in both fasta and fastq formats.

Input files: read file

Output files: new file with primer removed

```
use strict;
use warnings;
use Cwd;
my @fw_primers = ("AGAGTTTGATCCTGGCTCAG", "TACGGRAGGCAGCAG",
"GCCAGCAGCCGCGGTAA", "AGGATTAGATACCCT", "GAATTGACGGGGRCCC",
"GYAACGAGCGCAACCC");
my $dir = getcwd;
opendir(DIR, $dir) or die $!;
while (my $file = readdir(DIR)) {
    next if ($file =~ m/^\./);
    next if ($file =~ m/^\./);
    print $file;
    open(IN, $file) or die $!;
    my $outfile = "_dePrimer_".$file;
    open(OUT, ">$outfile");
    while(my $line = <IN>){
        foreach my $aprimer (@fw_primers){
            $line =~ s/$aprimer/\#/g;
        }
        $line =~ s/\#//g;
        print OUT $line;
    }
    close(OUT);
    close(IN);}
closedir(DIR);
exit 0;
```

2. The barcode extraction command can be used twice to remove both the barcode and primer (**163**).

Input files: read.fastq file, metadata file

Output files: directory of output files, including barcode file and read file(s)

```
extract_barcode.py -f [read file] -m [metadata file] -o [output directory name]
```

4.6.1.4 Performing quality filtering

1. For data in fastq format

Quality filtering in QIIME is performed as part of `split_libraries_fastq.py` (see Note 1). The following command can be executed to combine all the fastq files into one.

Input files: (1) directory of multiple fastq read files (2) parameter file

Output files: directory of output files, including `seqs.fna` file, which are used to select OTUs

```
multiple_split_libraries_fastq.py -i [input directory] -p [file path or name of parameter file] -o [output directory]
```

2. For data in fasta format

Quality filtering can be conducted by following the three steps described below.

Step 1: Clean the data to remove the incompatible lines and make the file compatible with QIIME (Note 2).

Input files: fasta read file(s)

Output files: directory of output fasta files

```
clean_fasta.py -f [read file name] -o [output file directory]
```

Step 2: Use `split_metadata.py` and `split_library.py` to transform from the previous combined fasta file format into another fasta format by parsing the sequences that meet user-defined quality thresholds and then renaming each read with the appropriate sample ID (*163, 218*).

Step 3: Concatenate all the fasta files using `cat` in linux.

Input files: multiple fasta read files

Output files: fasta file (.fna format)

```
cat file1 file2 ... file N > seq.fna
```

4.6.1.5 Additional quality control procedures

As described above, additional quality control procedures may be conducted. For example, the functions of `split_libraries.py` and `denoise_wrapper.py` can be used to de-noise the data (**163**).

Also, if chimeras are present in the sequencing read files, the function `parallel_identify_chimeric_seqs.py` can be used to identify chimeric sequences and apply the function of `filter_fasta.py` to filter the chimeras from the files (**163**).

4.6.1.6 Select OTUs

There are three approaches for picking OTU tables in QIIME. The first one is *de novo* OTU picking, which clusters reads against one another based on the similarity between reads (e.g., default is 97%) without utilizing sequencing information from external references (**163**). A drawback to the *de novo* approach is that there is no existing support for parallel computing in QIIME, which may slow down the process (**163**).

Input files: fasta file (`seqs.fna`, which is a post-`split_libraries` fasta file)

Output files: directory of output files, including a phylogenetic tree file and a biom-formatted OTU table

```
pick_de_novo_otus.py -i [input read file] -o [output directory name]
```

The second approach for selecting the OTU is closed-reference OTU picking, where reads are clustered against a reference. The reads that do not map to the reference are excluded from the downstream analysis. The reference database used by default in QIIME is constructed by clustering the full-length 16S rRNA sequences with 97% similarity in sequence fragments from the Greengenes database (55, 219, 220). The drawback of this closed-reference approach is that, by discarding reads that do not match the reference, the ability to detect novel diversity is lost (163).

Input files: fasta file (seqs.fna, which is a post-split_libraries fasta file)

Output files: directory of output files, including a phylogenetic tree file and a biom-formatted OTU table

```
pick_closed_reference_otus.py -i [input read file] -o [output directory name]
```

The third approach is the open-reference OTU picking approach. It combines the *de novo* and closed-reference OTU picking approaches. The reads are first clustered against a reference, and any reads that do not match the reference are clustered according to the *de novo* approach. All reads in the sample are clustered by the open-reference approach (163). For illustration, we used the open-reference OTU picking approach for the foregut esophageal adenocarcinoma dataset.

Input files: fasta file (seqs.fna, which is a post-split_libraries fasta file)

Output files: directory of output files, including a phylogenetic tree file and a biom-formatted OTU table


```
pick_open_reference_otus.py -i [input read file] -o [output directory name]
```

4.6.1.7 Summary of the OTU table

After selecting the OTU, we can summarize the generated OTU table in terms of the number of samples, the number of observations and a summary of the total counts observed for each sample (*163*). The OTU picking procedure may exclude some samples with low-quality scores in the OTU table file.

Input files: the file generated in the previous OTU picking step, called

“otu_table_mc2_w_tax_no_pynast_failures.biom”

Output files: summary of OTU picking result

```
biom summarize-table -i otu_table_mc2_w_tax_no_pynast_failures.biom
```

4.6.2 Import OTUs into R for Downstream Analysis

There are multiple ways to import the OTUs generated by QIIME into R (*221*) for the downstream analysis, as described below.

4.6.2.1 Read .biom file into R by using “compositional” library and QIIME

1. Convert .biom format to .txt file in QIIME (*163*).

Input files: biom format file

Output files: file in the tab-delimited table format

```
biom convert -i otu_table_mc2_w_tax_no_pynast_failures.biom -o  
table.from_biom_w_taxonomy_2.txt --to-tsv
```

2. Use R library to read the .txt file and convert to OTUs (**222**).

Input files: file in the tab-delimited table format from previous biom conversion step

Output files: OTU object of compositions in the framework of Aitchison Simplex by `acomp()` or
compositions as elements of simplex embedded in D-dimensional real space by `rcomp()`

```
install.packages("compositions")  
library(compositions)  
  
sample<-read.table("table.from_biom_w_taxonomy_2.txt ")  
counts_sample<-sample[,-1]  
t_ct_sample<-t(counts_sample)  
closure_t_sample<-clo(t_ct_sample)  
comp_sample<-acomp(closure_t_sample)  
  
or  
  
real_sample<-rcomp(closure_t_sample)
```

4.6.2.2 Read .biom file and phylogenetic tree information into R using the “phyloseq” package (171)

1. Install “phyloseq” package

```
source("https://bioconductor.org/biocLite.R")  
biocLite("phyloseq")
```

or

```
install.packages("devtools")  
library("devtools")  
install_github("phyloseq", "joey711")  
library("phyloseq")
```

2. Import .biom file

Input files: (1) biom file (2) tree file

Output files: OTU object

```
biom_sample<-import_biom("otu_table_mc2_w_tax_no_pynast_failures.biom",  
treefilename="rep_set.tre")  
otu_sample<-otu_table(biom_sample)
```

3. Import phylogenetic tree file into R

Input files: tree file in

Output files: phylo class object

```
tree_sample<-read_tree("rep_set.tre")
```

4.6.3 Computations of the Diversities

For the calculations of the diversities, the input object “biom_sample”, which is necessary for all scripts presented below, is obtained using the `import_biom` command in the “phyloseq” package described previously.

4.6.3.1 Alpha Diversity

```
## summarize alpha diversity  
> alpha_biom_sample<-estimate_richness(biom_sample,measures=c("Observed", "Chao1", "ACE",  
"Shannon", "Simpson", "InvSimpson", "Fisher"))
```

Table 15. Alpha diversity for foregut esophageal adenocarcinoma dataset

	Observed richness	Chao1	ACE	Shannon	Simpson	InvSimpson	Fisher
SRR1023240	818	1167.20	1300.73	5.05	0.98	50.09	250.41
SRR1023425	904	989.52	1084.98	3.73	0.94	16.51	190.43
SRR1023460	1144	1468.38	1615.59	3.59	0.89	8.87	244.86
...							

Table 15 reports α diversities for microbiome samples from the foregut esophageal adenocarcinoma study using various alpha diversity measures. In Table 15, we denote the Shannon diversity index as Shannon, Simpson's index as Simpson, inverse Simpson's diversity index as InvSimpson, Fisher's α diversity index as Fisher, Chao1 α diversity index as Chao1, and the abundance-based coverage estimator as ACE. The observed richness measure, Chao1 measure, and the ACE measure are generally higher than the Shannon index, InvSimpson, and Fisher's alpha index. For example, the α diversity for sample SRR1023425 is 904 using the observed richness, 989.52 using the Chao1 index, and 1084.98 in the ACE measure, respectively; while the α diversities are respectively 3.73, 16.51, and 190.43 when using the Shannon, InvSimpson, and Fisher diversity measures. The different α diversity measures assess different aspects of the community. For instance, a Shannon index implies higher uncertainty in correctly predicting the identity of the next species chosen at random for the given sample (*175*). Simpson's index emphasizes the evenness component in the sample (*57*). For example, for sample SRR1023240, Simpson's index is 0.98, which means that within the sample, the probability that two randomly selected taxa belong to the same species is as high as 0.98. We used the coefficient of variation to compare the different alpha diversity measures (see Note 3).

4.6.3.2 Plot Richness Graph

```
> ## plot alpha diversity
> plot_richness(biom_sample, measures=c("Observed", "Chao1", "ACE", "Shannon", "Simpson",
"InvSimpson", "Fisher"))
```

Figure 5. Alpha diversity plots for the foregut esophageal adenocarcinoma dataset

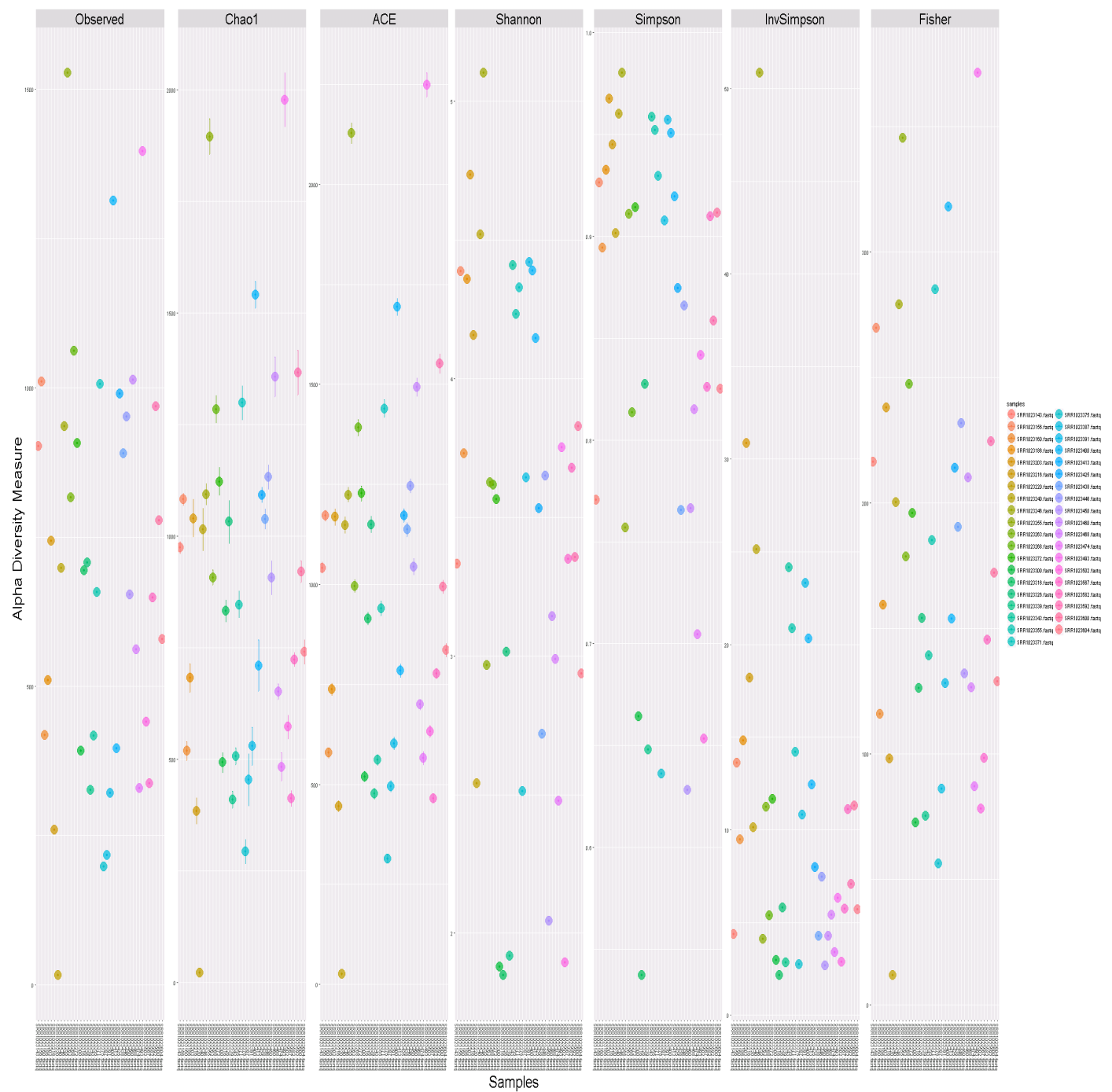


Figure 5 shows the richness of each sample using different alpha diversity measures, including the observed richness, Chao1, ACE, Shannon, Simpson's, inverse Simpson's, and Fisher index. The alpha diversity is relatively more evenly distributed when using the observed richness, Chao1, and Shannon index compared to those obtained using inverse Simpson's and Simpson's index.

4.6.3.2 Beta Diversity

4.6.3.2.1 Bray-Curtis Dissimilarity Index

```
> bray_dist <- distance(biom_sample, "bray")
```


Table 16. Bray-Curtis dissimilarity index using the foregut esophageal adenocarcinoma dataset

	SRR1023240	SRR1023582	SRR1023263	SRR1023425	SRR1023460
SRR1023582	0.81				
SRR1023263	0.81	0.83			
SRR1023425	0.89	0.91	0.72		
SRR1023460	0.89	0.88	0.79	0.68	
SRR1023343	0.82	0.86	0.88	0.91	0.91
...					

Table 16 shows the Bray-Curtis dissimilarity index for the foregut esophageal adenocarcinoma dataset. The larger the difference between two samples, the larger is the Bray-Curtis dissimilarity index. From Table 16, we observe that the difference between samples is relatively high, generally greater than 0.5. For example, the Bray-Curtis dissimilarity index between samples SRR1023425 and SRR1023582 is 0.91. The Bray-Curtis dissimilarity index between samples SRR1023425 and SRR1023263 is 0.72. Thus, sample SRR1023425 is more similar to sample SRR1023263 compared with sample SRR1023582.

4.6.3.2.2 Cao Dissimilarity Index

```
> cao_dist <- distance(biom_sample, "cao")
```

Table 17. Cao dissimilarity index using the foregut esophageal adenocarcinoma dataset

	SRR1023240	SRR1023582	SRR1023263	SRR1023425	SRR1023460
SRR1023582	2.28				
SRR1023263	2.26	2.41			
SRR1023425	2.31	2.45	2.34		
SRR1023460	2.21	2.35	2.28	2.33	
SRR1023343	2.17	2.34	2.35	2.34	2.28
...					

Table 17 shows the Cao dissimilarity index for the foregut esophageal adenocarcinoma dataset. A higher Cao dissimilarity index indicates more dissimilarity between samples. For example, the index between samples SRR1023582 and SRR1023425 is 2.45, while the index between samples SRR1023582 and SRR1023460 is 2.35, which implies that samples SRR1023460 and SRR1023582 are more similar compared with samples SRR1023425 and SRR1023582.

4.6.3.2.3 UniFrac Measures

4.6.3.2.3.1 Unweighted UniFrac measure

```
> unweighted_unifrac<-distance(biom_sample,"uunifrac")
```

Table 18. Unweighted UniFrac measure for the foregut esophageal adenocarcinoma dataset

	SRR1023240	SRR1023582	SRR1023263	SRR1023425	SRR1023460
SRR1023582	0.80				
SRR1023263	0.82	0.83			
SRR1023425	0.86	0.88	0.79		
SRR1023460	0.82	0.85	0.81	0.83	
SRR1023343	0.78	0.79	0.83	0.85	0.84
...					

Table 18 lists the results of the unweighted UniFrac measure for the foregut esophageal adenocarcinoma dataset. A higher value of the UniFrac measure indicates more dissimilarity between the samples. For example, the unweighted UniFrac dissimilarity between samples SRR1023240 and SRR1023582 is 0.80. The dissimilarity between SRR1023460 and SRR1023240 is 0.82, which implies that sample SRR1023240 is more similar to sample SRR1023460 compared with sample SRR1023582.

4.6.3.2.3.2 Weighted UniFrac measure

```
> weighted_unifrac<-distance(biom_sample,"wunifrac")
```

Table 19. Weighted UniFrac measure for the foregut esophageal adenocarcinoma dataset

	SRR1023240	SRR1023582	SRR1023263	SRR1023425	SRR1023460
SRR1023582	0.37				
SRR1023263	0.41	0.44			
SRR1023425	0.54	0.58	0.43		
SRR1023460	0.50	0.51	0.47	0.41	
SRR1023343	0.41	0.35	0.38	0.42	0.42
...					

The dissimilarity values in Table 19 are lower overall than those in Table 18, which were obtained using the unweighted UniFrac measure. Nonetheless, the patterns are similar. For example, in Table 19, the weighted UniFrac dissimilarity between sample SRR1023240 and sample SRR1023582 is 0.37, and the weighted UniFrac dissimilarity between sample SRR1023263 and SRR1023240 is 0.41.

4.6.3.2.3.3 Normalized unweighted UniFrac measure

```
> unweighted_unif_norm<-UniFrac(biom_sample, weighted=FALSE, normalized=TRUE)
```


Table 20. Normalized unweighted UniFrac measure for the foregut esophageal adenocarcinoma dataset

	SRR1023240	SRR1023582	SRR1023263	SRR1023425	SRR1023460
SRR1023582	0.80				
SRR1023263	0.82	0.83			
SRR1023425	0.86	0.88	0.79		
SRR1023460	0.82	0.85	0.81	0.83	
SRR1023343	0.78	0.79	0.83	0.85	0.84
...					

Table 20 shows the results for the normalized unweighted UniFrac dissimilarity, which is 0.82 between sample SRR1023460 and sample SRR1023240, and 0.80 between sample SRR1023240 and sample SRR1023582.

4.6.3.2.3.4 Normalized weighted UniFrac measure

```
> weighted_unif_norm<-UniFrac(biom_sample, weighted=TRUE, normalized=TRUE)
```

Table 21. Normalized weighted UniFrac measure for the foregut esophageal adenocarcinoma dataset

	SRR1023240	SRR1023582	SRR1023263	SRR1023425	SRR1023460
SRR1023582	0.34				
SRR1023263	0.36	0.38			
SRR1023425	0.42	0.44	0.31		
SRR1023460	0.40	0.41	0.36	0.27	
SRR1023343	0.34	0.29	0.30	0.29	0.30
...					

Table 21 provides the results of the normalized weighted UniFrac measure for the foregut esophageal adenocarcinoma dataset. The distance between samples SRR1023263 and SRR1023240 is 0.36. The distance between samples SRR1023263 and SRR1023582 is 0.38, which implies that sample SRR1023263 is more similar to sample SRR1023240 compared with sample SRR1023582.

4.6.4 Association Tests

4.6.4.1 Regression Models for Alpha Diversity

```
>sample_data<-read.table("book_chapter_samples_2.txt")
>colnames(sample_data)<-c("Run","disease_affected_status","Sex","BioSample","sample_name","SRA_sample","dbGap_sample_id")
>sample_sex<-(sample_data$Sex == "male")**2
>disease_status<-(sample_data$disease_affected_status== "Yes")**2
>alpha_biom_sample<-estimate_richness(biom_sample,measures=c("Observed", "Chao1", "ACE", "Shannon", "Simpson", "InvSimpson", "Fisher"))
>fit<-glm(disease_status~alpha_biom_sample$Chao1,family="binomial")
>summary(fit)
```

To study the association between outcomes and α diversity, we applied logistic regression for each of the α diversity measures using the foregut esophageal adenocarcinoma dataset. We used the esophageal adenocarcinoma status as the outcome variable. The input information is available in the file book_chapter_samples_2.txt. The association results show that the P -value is 0.18 for the observed richness approach, 0.33 for the Chao1 index, 0.33 for the ACE approach,

0.36 for the Shannon index, 0.98 for Simpson's index, 0.97 for the inverse Simpson's index, and 0.52 for the Fisher index approach. Therefore, at the nominal significance level of 0.05, there is no association between esophageal adenocarcinoma status and α diversity.

4.6.4.2 Permutation-based Nonparametric MANOVA (PERMANOVA)

The PERMANOVA test can be applied by using R package “vegan” (223) as described here.

Similar to the association test for alpha diversity, we used the esophageal adenocarcinoma status as the outcome variable. We used the Bray-Curtis measure as the dissimilarity metric for the illustration below. Other beta diversity measures can also be used by modifying the argument in the “distance” function.

```
>sample_data<-read.table("book_chapter_samples_2.txt")
>colnames(sample_data)<-
c("Run","disease_affected_status","Sex","BioSample","sample_name","SRA_sample","dbGap_sample_
id")
>sample_sex<-(sample_data$Sex == "male")**2
>disease_status<-(sample_data$disease_affected_status== "Yes")**2
>bray_dist <- distance(biom_sample, "bray")
>adonis(bray_dist~disease_status )
```

Given 20 oral samples from healthy subjects and 20 oral samples from the patients with esophageal adenocarcinoma, the PERMANOVA test gives a *P*-value of 0.15 for the association between esophageal adenocarcinoma status and the Bray-Curtis dissimilarity measure. The *P*-values for the other beta diversity measures were 0.07 for Cao beta diversity, and 0.17 and 0.04 for the unweighted and weighted UniFrac measures, respectively.

4.6.4.3 Microbiome Regression-based Kernel Association Test (MiRKAT)

The R packages of “MiRKAT” (217), “phyloseq” (171), and “ape” (224) are needed to perform the MiRKAT test. In addition to the information about OTUs (otu_table_mc2_w_tax_no_pynast_failures.biom), to conduct this test, information is needed on the phylogenetic tree (rep_set.tre file) and phenotypes (book_chapter_samples_2.txt file). Given all these files, the MiRKAT test, using the UniFrac measure in the kernel matrix, can be conducted by running the following script:

```
>##import biom file into phyloseq library
>biom_sample<-import_biom("otu_table_mc2_w_tax_no_pynast_failures.biom",
treefilename="rep_set.tre")
>otu_sample<-otu_table(biom_sample)
>### read tree file
>tree_sample<-read_tree("rep_set.tre")
>##Prepare the data
>sample_data<-read.table("book_chapter_samples_2.txt")
>colnames(sample_data)<-c("Run","disease_affected_status","Sex","BioSample","sample_name","SR
A_sample","dbGap_sample_id")
>sample_sex<-(sample_data$Sex == "male")**2
>disease_status<-(sample_data$disease_affected_status== "Yes")**2
>#transpose otu sample
>trans_otu_sample<-t(otu_sample)
>#root the tree
>tr <- root(tree_sample,1,resolve.root=TRUE)
>#Create the UniFrac Distances
>otu.tab.rff <- Rarefy(trans_otu_sample)$otu.tab.rff
>unifrac <- GUniFrac(otu.tab.rff, tr, alpha=c(0, 0.5, 1))$unifrac
>D.weighted = unifrac[, "d_1"]
>D.BC= as.matrix(vegdist(otu.tab.rff , method="bray"))
># Convert Distances to kernel matrices
>K.weighted = D2K(D.weighted)
>K.BC = D2K(D.BC)
># Testing using a single kernel
>MiRKAT(y = disease_status, Ks = K.weighted, X = sample_sex,out_type = "D",method =
"permutation", nperm=10000)
```

Given 20 oral samples from healthy subjects and 20 samples from the patients with esophageal adenocarcinoma, the MiRKAT test gives a P -value of 0.04. The MiRKAT test accounts for the covariates, which may have confounding effects, and incorporates the phylogenetic tree information; therefore, the MiRKAT approach can be robust compared to the PERMANOVA approach.

4.7. Notes

1. Details on the quality filtering strategy in QIIME, including motivation for the default parameter settings, are provided by Bokulich et al. (225). We note that the threshold parameter may require modification for quality scores, if necessary, by the creation of a parameter file. For example, the threshold parameter can be defined as given below in the parameter file.

```
split_libraries_fastq:phred_offset      33
```

2. It needs to install the PrimerProspector, a pipeline of programs for designing and analyzing PCR primers (226), to make the file compatible with QIIME.

3. To compare the different alpha diversity measures, we calculated the coefficient of variation, which is the ratio of the standard deviation to the mean, for each of the measures using the foregut esophageal adenocarcinoma dataset. The coefficient of variation shows the variability of

each measure over all the samples. The results show that the variability is 0.43 for the observed approach, 0.43 for the Chao1 index, 0.43 for the ACE approach, 0.2 for the Shannon index, 0.1 for Simpson's index, 0.82 for inverse Simpson's index, and 0.41 for the Fisher index approach. Among all the approaches, the inverse Simpson's index has the highest variability, while the Shannon index and Simpson's index have relatively lower variability. Based on the coefficient of variation, the Shannon index and Simpson's index have less variability than the other measures.

Chapter 5

Conclusions and Future Directions

This chapter summarizes the thesis, discusses its findings and contributions, points out limitations of the current work, and also outlines directions for future research. Depending on the genomic data for the human and microbiome, the aim of this thesis is to explore the statistical applications of frequentist methods to the sequencing data. However, many extensions of this research deserve further consideration under the Bayesian statistical point of view. The chapter is divided into two sections. Section 5.1 summarizes the thesis and provides conclusions. Section 5.2 discusses future work.

5.1 Conclusions

In this dissertation, we investigated and developed novel statistical approaches for assessing human and microbiome genetic data. First, we introduced the definition of sequencing data and summarized the well-known sequencing techniques and approaches, e.g., Sanger sequencing, next-generation sequencing, DNA sequencing, and RNA sequencing approaches. Detailed discussions were provided as the background of our proposed methods. We also gave a brief introduction to the definitions of 16S rRNA and provided background information for the human microbiome, which may be associated with human disease.

In Chapter 2, we introduced the sequencing error rates in next-generation sequencing short read data and reviewed the traditional error rate estimation approaches, including the shadow regression estimation method. When estimating sequencing error rates, the shadow regression estimation approach is widely used, but operates under the linearity assumption that the number of shadow counts increases linearly with the number of error-free read counts, which may not be appropriate for all types of sequencing data. To obtain a unified model, we developed the empirical error rate estimation approach by using a nonlinear statistical model of cubic smoothing splines or robust smoothing splines to estimate the shadow counts in next-generation sequencing data. Our proposed empirical error rate estimation approach is applicable for multiple datasets, including DNA sequencing data and mRNA sequencing data. Under all the simulation scenarios tested, our proposed empirical error rate estimation approach is less biased than shadow regression estimation. We also reproduced the results of the Wang et al. simulation study (4). We note that the data generated using the simulation approach of Wang et al. may not reflect the real sequencing data structure, especially the nonlinear relationship. Therefore, the

frequency-based simulation approach was proposed in this thesis, aiming to mimic the real sequencing data structure and increase computational efficiency. To make our method expandable to specific sequencing reads, we made some effort to re-define the sequencing per-base error rate that enables the expansion of the error rate estimation to specific sequencing reads instead of a fixed value for all the sequences in the sample. The adjusted per-base error rate definition provides more information according to the sequencing read compared with the original per-base error rate definition in the shadow regression approach.

In Chapter 3, we expanded the error rate estimation to more diverse definitions of shadow. The shadow error rate investigated in Chapter 2 focused on error rates with substitution, which are the error rates studied in the majority of the literature that discusses sequencing errors.

Specifically, the shadows studied in Chapter 2 are sequences with up to two base differences from error-free sequences, which were recommended in a previous study (4). In Chapter 3, we varied the definition of shadows to be sequences that are different from error-free sequencing reads by only one base, only two bases, and up to two bases, respectively. We also extended the estimation to deletion and insertion error rates, where the difference is only one base, only two bases, and up to two bases from those of the error-free sequences. In addition, we confirmed with the comments from a previous study (4) that using shadows that differ from the error-free sequences by up to two bases are good enough without excessive additional computational cost.

In Chapter 4, we studied the association of the human microbiome with complex diseases and reviewed multiple approaches for processing and analyzing human microbiome data. We

provided detailed programming scripts about processing raw microbiome data into downstream analysis OTU format, calculating various alpha and beta diversities to estimate within-sample and between-sample distances, and analyzing the statistical association between the distance metrics of the microbiome and outcomes of interest. We performed real data analysis using data from the foregut microbiome in esophageal adenocarcinoma to illustrate the statistical approaches.

5.2 Future Directions

The human microbiome interacts with the environment dynamically and lives on surfaces with which we interact daily (42). The composition of the microbiome in the human body varies over time according to a person's health, diet, lifestyle, social interactions, early antibiotic therapy, genotype, and environmental chemicals (45, 48). A cross-sectional study, in which only a single response is available, ignores the unstable and changeable characteristic of the human microbiome and does not sufficiently reflect the overall performance of the composition of the human microbiome over time. In contrast to cross-sectional data, longitudinal data comprise repeated observation measures over time (227).

Heuristically, a longitudinal study enables each microbiome sample to act as its own control, which allows for a reduction of the ever-present heterogeneity (227). Longitudinal analysis of time-series data helps to capture the outcome changes and progression of the microbiome over time (79). Traditional longitudinal analysis, such as a simple regression of a dependent variable on a time measure, a general linear model with a fixed effect, or analyzing a single summary subject-level number that indexes changes for each subject, has limitations in detecting important signals and effects, and loses information on bacterial connections as the individual taxon is treated as a separate outcome, even in well-designed and conducted studies (79, 80). The analytical model for longitudinal data should take into account the correlation on repeated observations in order to obtain valid statistical inference on coefficient estimates. Ignoring bacterial correlations may lead to inefficient parameter estimates and inconsistent estimates of precision. (227)

A Bayesian network is a graphical model of joint multivariate probability that describes the statistical dependence and conditional independence between variables according to the data (228). It has the ability to capture complicated stochastic processes and has been shown to be a clear methodology for learning from noisy observations (228). In addition, a Bayesian network has the capability of describing interaction relationships between random variables. The established networks, working as a framework of relationships, can describe the dependence structure between multiple interacting quantities (228). Another widely-known advantage of a Bayesian network is the capacity to indicate causal influence. A Bayesian network has strictly mathematically defined probabilities and conditional independence statements, which implies the direct causal influence based on the network (228). Even though several assumptions may not be necessary in gene expression or taxa count data, the connections for some causality may be indicative based using the network to learn from the observational data (228).

The Bayesian network consists of two components that describe the joint probability distribution of random variables. The first component is a directed or undirected acyclic graph for which the vertices represent random variables. The second component includes the probabilities that describe the conditional distribution of each variable given its parents in the graph. Under the Markov assumption, each variable is independent of its descendants given the corresponding parents in the graph (228). There are two commonly used approaches to describe the conditional distributions, depending on the data types of the variables. If the variables are discrete, the general representation involves constructing a table with the probability of each joint assignment to the corresponding parents specified. Specifying the conditional distribution of discrete variables is flexible and can describe any discrete conditional distribution in all possible densities, at the cost of an exponential increase in the number of free parameters. (228) If the

variables are continuous, the conditional independence relationship usually will be described with a linear Gaussian conditional density, where the variable is normally distributed with a mean linearly associated with the values of its parents, and the variance is independent of its parents. A joint multivariate Gaussian distribution will be attained under the assumption that all variables have linear Gaussian distribution in the network (228). In addition, the general approach to learn a Bayesian network is to use a scoring function that estimates the possible networks based on the training data and then to select the one with the optimal score (228).

Multivariate Gaussian graphical models (GGMs) are widely used and considered as the standard approach to model the conditional independence relationship in multivariate Gaussian data (229). Multivariate GGMs have been successfully implemented in various areas, including health sciences, genomics, and economics. However, traditional multivariate GGMs may have limitations due to the lack of efficiency and accuracy, and loss of statistical power in handling high-dimensional data (229). Ni et al. developed the novel matrix-variate directed acyclic graph approach and extended it to model both directed and undirected graphs and paired it with an efficient Markov chain Monte Carlo (MCMC) algorithm (229). This matrix-variate graph can be a mixture of directed and undirected graphs (hybrid mGGM), i.e., when the prior ordering information is known, it may be represented as a directed graph; whereas when the prior ordering information is missing, it may be represented as an undirected graph. The hybrid mGGM approach enables accounting for the uncertainty of the graphical structure and providing regularized and sparse estimators under a Bayesian framework. The uncertainty consideration is essential when handling high-dimensional complicated datasets, as there may be several graphs that equally explain the observed data and which are well under the limited number of samples. (229) Based on previous simulation studies, the hybrid mGGM approach performs better than the

traditional alternative naïve directed acyclic graph approach, as well as the state-of-the-art Bayesian approach (230) and the matrix-variate Gaussian non-Bayesian graphical model approach in terms of the true positive rate, false discovery rate, and Mathews correlation coefficient (229, 231).

In future studies, we will expand the cross-sectional analysis of the microbiome to an analysis of longitudinal microbiome data to study the specific relationships and general trend in microbiome taxa with time series information. The hybrid mGGM approach can be used to study the longitudinal microbiome data. If the prior ordering information is not clear, an undirected graph can be used to specify the microbiome taxa, while a directed graph can be used to specify the time. This approach will overcome the limitations in previous longitudinal studies of the microbiome, which ignored the interaction association between taxa, and will be able to predict future changes in the composition of the microbiome over time. A Bayesian network has been successfully utilized to study time-series data in other areas, including clinical research and ecological prediction. So far, there is little in the literature that proposes using a Bayesian network to analyze longitudinal microbiomic data. The byproducts of using a Bayesian network are the estimated relationships of how one taxon influences another taxon over time (232).

Before fitting the hybrid mGGM approach to a longitudinal dataset of the microbiome, there is a need to preprocess the high-dimensional raw microbiome data into an appropriate format. There are mainly two challenges involving the microbiome data before constructing a Bayesian network. The first challenge involves high percentage of zero counts in the OTU table. One essential problem in microbiomic data analysis is that there are many zeros in taxa counts, which will not allow for the direct use of standard statistical models. One approach to solve this issue is to combine the OTUs into the genera level that is a higher taxonomic level in the microbiome,

and then remove the genera with zero taxa counts in more than 90% of the sample. After the zero-checking stage, the remaining zero counts may be replaced by a pseudo value to better facilitate the estimation (usually 0.05 or 0.5). This replacement will not affect the estimation, which was confirmed by a previous study, especially when the replaced number is very small and the number of read counts is large enough ($>>0.05$ or 0.5). These read counts will be converted into compositions, and a log-ratio transformation will then be applied to these compositional data (213, 214, 233, 234). Then the hybrid mGGM approach can be applied to the log-transformed compositional data to study the specific relationship structure and general trends in the microbiome genera over time.

A simulation analysis may be applied to further study the performance of the hybrid mGGM approach on longitudinal microbiome data. One possible approach was introduced by using Chen and Li's approach (235). This simulation approach uses a Dirichlet multinomial distribution to model the OTU counts, which takes the overdispersion of the OTU table into account. The maximum likelihood method is used to estimate the mean proportions and overdispersion parameters in the Dirichlet distribution of the OTU table. The taxa counts will then be generated using a multinomial distribution. The outcome of interest will be generated using a normal distribution if it is continuous; otherwise, if the outcome is binary, it will be generated using a Bernoulli distribution with the mean following a pre-specified distribution. (235)

One of the major goals when using the hybrid mGGM approach to construct the Bayesian network is to integrate the taxa count data in the microbiome and capture the common structure of the genera-level taxa over time to find how the microbiome genera interacted with each other, which may be associated with the outcome. Another goal is to utilize the Bayesian network in a

clinical study, which may help to make predictions for future therapeutic strategies or disease prevention. For example, if the outcome of interest is a fatal or disease event with some microbiome genera available in high proportions, it would be feasible to inhibit both these genera and the associated genera in order to cure or prevent the disease. Previous simulation studies showed that the hybrid mGGM approach demonstrated superior performance compared to that of the conventional Bayesian network approaches in terms of increased computational efficiency, estimation accuracy, and statistical power (229). A Bayesian network also accounts for microbiome taxa interaction and is capable of predicting future changes in the microbiome's composition over time.

References

- [1] Goldstein, D. B. (2009) Common genetic variation and human traits, *N Engl J Med* 360, 1696-1698.
- [2] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009) Finding the missing heritability of complex diseases, *Nature* 461, 747-753.
- [3] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X. G., and Mortazavi, A. (2016) A survey of best practices for RNA-seq data analysis (vol 17, 13, 2016), *Genome Biol* 17.
- [4] Wang, X. V., Blades, N., Ding, J., Sultana, R., and Parmigiani, G. (2012) Estimation of sequencing error rates in short reads, *Bmc Bioinformatics* 13.
- [5] Shendure, J., and Ji, H. (2008) Next-generation DNA sequencing, *Nat Biotechnol* 26, 1135-1145.
- [6] Ross, J. S., and Cronin, M. (2011) Whole Cancer Genome Sequencing by Next-Generation Methods, *Am J Clin Pathol* 136, 527-539.
- [7] Kobayashi, S., Boggon, T. J., Dayaram, T., Janne, P. A., Kocher, O., Meyerson, M., Johnson, B. E., Eck, M. J., Tenen, D. G., and Halmos, B. (2005) EGFR mutation and resistance of non-small-cell lung cancer to gefitinib, *N Engl J Med* 352, 786-792.
- [8] Shearer, A. E., DeLuca, A. P., Hildebrand, M. S., Taylor, K. R., Gurrola, J., Scherer, S., Scheetz, T. E., and Smith, R. J. H. (2010) Comprehensive genetic testing for hereditary

- hearing loss using massively parallel sequencing, *P Natl Acad Sci USA* 107, 21104-21109.
- [9] Sanger, F., Nicklen, S., and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors, *Proc Natl Acad Sci U S A* 74, 5463-5467.
- [10] Olena Morozova, M. H., Marco A. Marra. (2009) Applications of New Sequencing Technologies for Transcriptome Analysis, *Annu. Rev. Genomics Hum. Genet.*
- [11] Altimari, A., de Biase, D., De Maglio, G., Gruppioni, E., Capizzi, E., Degiovanni, A., D'Errico, A., Pession, A., Pizzolitto, S., Fiorentino, M., and Tallini, G. (2013) 454 next generation-sequencing outperforms allele-specific PCR, Sanger sequencing, and pyrosequencing for routine KRAS mutation analysis of formalin-fixed, paraffin-embedded samples, *Onco Targets Ther* 6, 1057-1064.
- [12] Ladouceur, M., Dastani, Z., Aulchenko, Y. S., Greenwood, C. M., and Richards, J. B. (2012) The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals, *PLoS Genet* 8, e1002496.
- [13] Rizzo, J. M., and Buck, M. J. (2012) Key principles and clinical applications of "next-generation" DNA sequencing, *Cancer Prev Res (Phila)* 5, 887-900.
- [14] Craig DW, P. J., Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, Homer N, Huentelman MJ. (2008) Identification of genetic variants using bar-coded multiplexed sequencing, *Nat Methods*.
- [15] Park, P. J. (2009) ChIP-Seq: advantages and challenges of a maturing technology, *Nat Rev Genet.*
- [16] Alkan C, K. J., Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE. (2009) Personalized

- copy number and segmental duplication maps using next-generation sequencing, *Nat Genet*.
- [17] Joshua S Bloom, Z. K., Leonid Kruglyak, Mona Singh, Amy A Caudy. (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays, *BMC Genomics*.
- [18] Mardis, E. R. (2008) Next-generation DNA sequencing methods, *Annu Rev Genomics Hum Genet* 9, 387-402.
- [19] Sultan M, S. M., Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome, *Science*.
- [20] Paul L. Auer, R. W. D. (2010) Statistical Design and Analysis of RNA Sequencing Data, *GENETICS*.
- [21] Noonan, J. P., Hofreiter, M., Smith, D., Priest, J. R., Rohland, N., Rabeder, G., Krause, J., Detter, J. C., Paabo, S., and Rubin, E. M. (2005) Genomic sequencing of Pleistocene cave bears, *Science* 309, 597-599.
- [22] Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., Macphee, R. D., Buigues, B., Tikhonov, A., Huson, D. H., Tomsho, L. P., Auch, A., Rampp, M., Miller, W., and Schuster, S. C. (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA, *Science* 311, 392-394.
- [23] Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., Du, L., Egholm, M., Rothberg, J. M., Paunovic, M., and Paabo, S. (2006) Analysis of one million base pairs of Neanderthal DNA, *Nature* 444, 330-336.

- [24] Noonan, J. P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Paabo, S., Pritchard, J. K., and Rubin, E. M. (2006) Sequencing and analysis of Neanderthal genomic DNA, *Science* 314, 1113-1118.
- [25] Mark A DePristo, E. B., Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, Mark J Daly. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nature Genetics*.
- [26] Mortazavi, A., Williams, Brian A., McCue, Kenneth, Schaeffer, Lorian, Wold, Barbara. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature Methods*.
- [27] Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nat Rev Genet* 10, 57-63.
- [28] Goldman, D., and Domschke, K. (2014) Making sense of deep sequencing, *Int J Neuropsychopharmacol* 17, 1717-1725.
- [29] Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays, *Genome Res* 18, 1509-1517.
- [30] Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J.,

- and Grimmond, S. M. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing, *Nature Methods* 5, 613-619.
- [31] Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., Jones, S. J. M., and Marra, M. A. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing, *Biotechniques* 45, 81-+.
- [32] Ugrappa Nagalakshmi, Z. W., Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, Michael Snyder. (2008) The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing, *Science*.
- [33] Andrea L. Eveland, D. R. M., Karen E. Koch. (2008) Transcript Profiling by 3#-Untranslated Region Sequencing Resolves Expression of Gene Families, *Plant Physiology*.
- [34] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing, *Science* 320, 1344-1349.
- [35] Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J., and Bahler, J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution, *Nature* 453, 1239-U1239.
- [36] Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature Methods* 5, 621-628.
- [37] Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis, *Cell* 133, 523-536.

- [38] Nicole Cloonan, Q. X. G. J. F. D. F. T. D. T. P. T. G. K. S. M. G. (2009) RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data *Bioinformatics*.
- [39] Eloë-Fadrosch, E. A., and Rasko, D. A. (2013) The human microbiome: from symbiosis to pathogenesis, *Annu Rev Med* 64, 145-163.
- [40] Bromberg, J. S., Fricke, W. F., Brinkman, C. C., Simon, T., and Mongodin, E. F. (2015) Microbiota-implications for immunity and transplantation, *Nat Rev Nephrol* 11, 342-353.
- [41] Johnson, C. L., and Versalovic, J. (2012) The human microbiome and its potential importance to pediatrics, *Pediatrics* 129, 950-960.
- [42] Luke K Ursell, J. L. M., Laura Wegener Parfrey, and Rob Knight. (2012) Defining the Human Microbiome, *Nutr Rev*.
- [43] Hooper, L. V., and Gordon, J. I. (2001) Commensal host-bacterial relationships in the gut, *Science* 292, 1115-1118.
- [44] Maslowski, K. M., and Mackay, C. R. (2011) Diet, gut microbiota and immune responses, *Nat Immunol* 12, 5-9.
- [45] Li, H. (2015) Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis, *Annual Review of Statistics and Its Application*.
- [46] Backhed F, L. R., Sonnenburg JL, Peterson DA, Gordon JI. (2005) Host-bacterial mutualism in the human intestine., *Science*.
- [47] http://hmpdacc.org/micro_analysis/microbiome_analyses.phpProject, N. H. M. About HMP Metagenomic Sequencing & Analysis.
- [48] Sandrine P Claus, H. G., Sandrine Ellero-Simatos. (2016) The gut microbiota: a major player in the toxicity of environmental pollutants?, *npj Biofilms and Microbiomes*.

- [49] <http://www.illumina.com/areas-of-interest/microbiology/human-microbiome-analysis.html> Introduction to Human Microbiome Analysis.
- [50] Clarridge, J. E. (2004) Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases, *Clin Microbiol Rev* 17, 840-862.
- [51] Hartstra, A. V., Bouter, K. E. C., Backhed, F., and Nieuwdorp, M. (2015) Insights Into the Role of the Microbiome in Obesity and Type 2 Diabetes, *Diabetes Care* 38, 159-165.
- [52] Janda, J. M., and Abbott, S. L. (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls, *J Clin Microbiol* 45, 2761-2764.
- [53] Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., and Tiedje, J. M. (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis, *Nucleic Acids Res* 42, D633-642.
- [54] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glockner, F. O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, *Nucleic Acids Res* 41, D590-596.
- [55] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB, *Appl Environ Microb* 72, 5069-5072.
- [56] Li, K., Bihan, M., Yooseph, S., and Methe, B. A. (2012) Analyses of the Microbial Diversity across the Human Microbiome, *Plos One* 7.

- [57] Nagendra, H. (2002) Opposite trends in response for the Shannon and Simpson indices of landscape diversity, *Appl Geogr* 22, 175-186.
- [58] Basualdo, C. V. (2011) Choosing the best non-parametric richness estimator for benthic macroinvertebrates databases, *Rev. Soc. Entomol. Argent* 70(1-2), 27-38.
- [59] Anderson, M. J., Ellingsen, K. E., and McArdle, B. H. (2006) Multivariate dispersion as a measure of beta diversity, *Ecol Lett* 9, 683-693.
- [60] Cao, Y., Williams, W. P., and Bark, A. W. (1997) Similarity measure bias in river benthic Aufwuchs community analysis, *Water Environ Res* 69, 95-106.
- [61] Lozupone, C., and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities, *Appl Environ Microb* 71, 8228-8235.
- [62] Shah, N., Tang, H., Doak, T. G., and Ye, Y. (2011) Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics, *Pac Symp Biocomput*, 165-176.
- [63] Ranjan, R., Rani, A., Metwally, A., McGee, H. S., and Perkins, D. L. (2016) Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing, *Biochem Biophys Res Commun* 469, 967-977.
- [64] Sharpton, T. J. (2014) An introduction to the analysis of shotgun metagenomic data, *Front Plant Sci* 5, 209.
- [65] Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A. L., Madsen, K. L., and Wong, G. K. S. (2016) Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics, *Front Microbiol* 7.

- [66] Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., and Huttenhower, C. (2015) Sequencing and beyond: integrating molecular 'omics' for microbial community profiling, *Nat Rev Microbiol* 13, 360-372.
- [67] Norman, J. M., Handley, S. A., and Virgin, H. W. (2014) Kingdom-agnostic metagenomics and the importance of complete characterization of enteric microbial communities, *Gastroenterology* 146, 1459-1469.
- [68] Poretsky, R., Rodriguez, R. L., Luo, C., Tsementzi, D., and Konstantinidis, K. T. (2014) Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics, *Plos One* 9, e93827.
- [69] Zhu, X., Wang, J., Peng, B., and Shete, S. (2016) Empirical estimation of sequencing error rates using smoothing splines, *Bmc Bioinformatics* 17.
- [70] van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014) Ten years of next-generation sequencing technology, *Trends Genet* 30, 418-426.
- [71] Schlotterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014) Sequencing pools of individuals-mining genome-wide polymorphism data without big funding, *Nature Reviews Genetics* 15, 749-763.
- [72] Mardis, E. R. (2013) Next-Generation Sequencing Platforms, *Annu Rev Anal Chem* 6, 287-303.
- [73] Yang, Y., Xie, B., and Yan, J. (2014) Application of next-generation sequencing technology in forensic science, *Genomics Proteomics Bioinformatics* 12, 190-197.
- [74] Metzker, M. L. (2010) Sequencing technologies - the next generation, *Nat Rev Genet* 11, 31-46.

- [75] Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., Dangl, J. L., and Jones, C. D. (2007) Extending assembly of short DNA sequences to handle error, *Bioinformatics* 23, 2942-2944.
- [76] Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F., and Hackermuller, J. (2009) Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures, *Plos Comput Biol* 5.
- [77] Simpson, J. T. (2014) Exploring genome characteristics and sequence quality without a reference, *Bioinformatics* 30, 1228-1235.
- [78] Trivedil, U. H., Cezard, T., Bridgett, S., Montazam, A., Nichols, J., Blaxter, M., and Gharbi, K. (2014) Quality control of next-generation sequencing data without a reference, *Front Genet* 5.
- [79] Michael J. McGeachie, J. E. S., Travis Gibson, George M. Weinstock, Yang-Yu Liu, Diane R. Gold, Scott T. Weiss, Augusto Litonjua. (2016) Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks, *Nature Scientific Reports*.
- [80] Joseph J. Locascio, Alireza A. (2011) An Overview of Longitudinal Data Analysis Methods for Neurological Research, *Dement Geriatr Cogn Disord Extra*
- [81] Schlotterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014) Sequencing pools of individuals - mining genome-wide polymorphism data without big funding, *Nature reviews. Genetics* 15, 749-763.
- [82] Mardis, E. R. (2013) Next-generation sequencing platforms, *Annual review of analytical chemistry* 6, 287-303.
- [83] Wang, X. V., Blades, N., Ding, J., Sultana, R., and Parmigiani, G. (2012) Estimation of sequencing error rates in short reads, *BMC Bioinformatics* 13, 185.

- [84] Magurran, A. E. (2004) *Measuring Biological Diversity*, Blackwell Publishing, Oxford, UK.
- [85] Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., and Gordon, J. I. (2009) A core gut microbiome in obese and lean twins, *Nature* 457, 480-484.
- [86] Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., Dangel, J. L., and Jones, C. D. (2007) Extending assembly of short DNA sequences to handle error, *Bioinformatics*.
- [87] Sundquist, A. e. a. (2007) Whole-genome sequencing and assembly with high-throuput, short-read technologies, *PLoS One*.
- [88] Hert, D. G., Fredlake, C. P., and Barron, A. E. (2008) Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods, *Electrophoresis* 29, 4618-4626.
- [89] Brown, T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012) A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data, *Cornell University Library*.
- [90] Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F., and Hackermuller, J. (2009) Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures, *PLoS One*.
- [91] Trivedi, U. H., Cezard, T., Bridgett, S., Montazam, A., Nichols, J., Blaxter, M., and Gharbi, K. (2014) Quality control of next-generation sequencing data without a reference, *Frontiers in genetics* 5, 111.

- [92] <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Andrews>, S. FastQC a quality control tool for high throughput sequence data.
- [93] Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, *BMC Bioinformatics* 11, 94.
- [94] Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads, *Genome Res* 18, 810-820.
- [95] Schroder, J., Schroder, H., Puglisi, S. J., Sinha, R., and Schmidt, B. (2009) SHREC: a short-read error correction method, *Bioinformatics* 25, 2157-2163.
- [96] Kelley, D. R., Schatz, M. C., and Salzberg, S. L. (2010) Quake: quality-aware detection and correction of sequencing errors, *Genome Biol* 11, R116.
- [97] Salmela, L. (2010) Correction of sequencing errors in a mixed set of reads, *Bioinformatics* 26, 1284-1290.
- [98] Schroder, J., Bailey, J., Conway, T., and Zobel, J. (2010) Reference-free validation of short read data, *PLoS One* 5, e12681.
- [99] Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010) De novo assembly of human genomes with massively parallel short read sequencing, *Genome Res* 20, 265-272.
- [100] Keele, L. J. (2008) *Semiparametric Regression for the Social Sciences*, John Wiley & Sons Ltd. , Chichester, England.
- [101] Schroder, J., Bailey, J., Conway, T., and Zobel, J. (2010) Reference Free Validation of Short Read Data, *PLoS One*.

- [102] Gunewardena, S. S. (2013) Optimum-time, Optimum-space, Algorithms for k-mer Analysis of Whole Genome Sequences, *Bioinformatics and Comparative Genomes*.
- [103] Melsted, P., and Pritchard, J. K. (2011) Efficient counting of k-mers in DNA sequences using a bloom filter, *BMC Bioinformatics*.
- [104] Heo, Y., Wu, X. L., Chen, D., Ma, J., and Hwu, W. M. (2014) BLESS: bloom filter-based error correction solution for high-throughput sequencing reads, *Bioinformatics* 30, 1354-1362.
- [105] Shi, L., and Reid, L. H., and Jones, W. D., and Shippy, R., and Warrington, J. A., and Baker, S. C., and Collins, P. J., and de Longueville, F., and Kawasaki, E. S., and Lee, K. Y., and Luo, Y., and Sun, Y. A., and Willey, J. C., and Setterquist, R. A., and Fischer, G. M., and Tong, W., and Dragan, Y. P., and Dix, D. J., and Frueh, F. W., and Goodsaid, F. M., and Herman, D., and Jensen, R. V., and Johnson, C. D., and Lobenhofer, E. K., and Puri, R. K., and Schrf, U., and Thierry-Mieg, J., and Wang, C., and Wilson, M., and Wolber, P. K., and Zhang, L., and Amur, S., and Bao, W., and Barbacioru, C. C., and Lucas, A. B., and Bertholet, V., and Boysen, C., and Bromley, B., and Brown, D., and Brunner, A., and Canales, R., and Cao, X. M., and Cebula, T. A., and Chen, J. J., and Cheng, J., and Chu, T. M., and Chudin, E., and Corson, J., and Corton, J. C., and Croner, L. J., and Davies, C., and Davison, T. S., and Delenstarr, G., and Deng, X., and Dorris, D., and Eklund, A. C., and Fan, X. H., and Fang, H., and Fulmer-Smentek, S., and Fuscoe, J. C., and Gallagher, K., and Ge, W., and Guo, L., and Guo, X., and Hager, J., and Haje, P. K., and Han, J., and Han, T., and Harbottle, H. C., and Harris, S. C., and Hatchwell, E., and Hauser, C. A., and Hester, S., and Hong, H., and Hurban, P., and Jackson, S. A., and Ji, H., and Knight, C. R., and Kuo, W. P., and LeClerc, J. E., and

- Levy, S., and Li, Q. Z., and Liu, C., and Liu, Y., and Lombardi, M. J., and Ma, Y., and Magnuson, S. R., and Maqsodi, B., and McDaniel, T., and Mei, N., and Myklebost, O., and Ning, B., and Novoradovskaya, N., and Orr, M. S., and Osborn, T. W., and Papallo, A., and Patterson, T. A., and Perkins, R. G., and Peters, E. H., and Peterson, R., and Philips, K. L., and Pine, P. S., and Pusttai, L., and Qian, F., and Ren, H., and Rosen, M., and Rosenzweig, B. A., and Samaha, R. R., and Schena, M., and Schroth, G. P., and Shchegrova, S., and Smith, D. D., and Staedtler, F., and Su, Z., and Sun, H., and Szallasi, Z., and Tezak, Z., and Thierry-Mieg, D., and Thompson, K. L., and Tikhonova, I., and Turpaz, Y., and Vallanat, B., and Van, C., and Walker, S. J., and Wang, S. J., and Wang, Y., and Wolfinger, R., and Wong, A., and Wu, J., and Xiao, C., and Xie, Q., and Xu, J., and Yang, W., and Zhang, L., and Zhong, S., and Zong, Y., and Slikker, W., Jr. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements, *Nature biotechnology* 24, 1151-1161.
- [106] Hu, H., Wrogemann, K., Kalscheuer, V., Tzschach, A., Richard, H., Haas, S. A., Menzel, C., Bienek, M., Froyen, G., Raynaud, M., Van Bokhoven, H., Chelly, J., Ropers, H., and Chen, W. (2009) Mutation screening in 86 known X-linked mental retardation genes by droplet-based multiplex PCR and massive parallel sequencing, *Hugo J* 3, 41-49.
- [107] Birney, E., and Stamatoyannopoulos, J. A., and Dutta, A., and Guigo, R., and Gingeras, T. R., and Margulies, E. H., and Weng, Z., and Snyder, M., and Dermitzakis, E. T., and Thurman, R. E., and Kuehn, M. S., and Taylor, C. M., and Neph, S., and Koch, C. M., and Asthana, S., and Malhotra, A., and Adzhubei, I., and Greenbaum, J. A., and Andrews, R. M., and Flicek, P., and Boyle, P. J., and Cao, H., and Carter, N. P., and Clelland, G. K., and Davis, S., and Day, N., and Dhami, P., and Dillon, S. C., and

Dorschner, M. O., and Fiegler, H., and Giresi, P. G., and Goldy, J., and Hawrylycz, M.,
 and Haydock, A., and Humbert, R., and James, K. D., and Johnson, B. E., and Johnson,
 E. M., and Frum, T. T., and Rosenzweig, E. R., and Karnani, N., and Lee, K., and
 Lefebvre, G. C., and Navas, P. A., and Neri, F., and Parker, S. C., and Sabo, P. J., and
 Sandstrom, R., and Shafer, A., and Vetrie, D., and Weaver, M., and Wilcox, S., and Yu,
 M., and Collins, F. S., and Dekker, J., and Lieb, J. D., and Tullius, T. D., and Crawford,
 G. E., and Sunyaev, S., and Noble, W. S., and Dunham, I., and Denoeud, F., and
 Reymond, A., and Kapranov, P., and Rozowsky, J., and Zheng, D., and Castelo, R., and
 Frankish, A., and Harrow, J., and Ghosh, S., and Sandelin, A., and Hofacker, I. L., and
 Baertsch, R., and Keefe, D., and Dike, S., and Cheng, J., and Hirsch, H. A., and Sekinger,
 E. A., and Lagarde, J., and Abril, J. F., and Shahab, A., and Flamm, C., and Fried, C., and
 Hackermuller, J., and Hertel, J., and Lindemeyer, M., and Missal, K., and Tanzer, A., and
 Washietl, S., and Korbel, J., and Emanuelsson, O., and Pedersen, J. S., and Holroyd, N.,
 and Taylor, R., and Swarbreck, D., and Matthews, N., and Dickson, M. C., and Thomas,
 D. J., and Weirauch, M. T., and Gilbert, J., and Drenkow, J., and Bell, I., and Zhao, X.,
 and Srinivasan, K. G., and Sung, W. K., and Ooi, H. S., and Chiu, K. P., and Foissac, S.,
 and Alioto, T., and Brent, M., and Pachter, L., and Tress, M. L., and Valencia, A., and
 Choo, S. W., and Choo, C. Y., and Ucla, C., and Manzano, C., and Wyss, C., and
 Cheung, E., and Clark, T. G., and Brown, J. B., and Ganesh, M., and Patel, S., and
 Tammana, H., and Chrast, J., and Henrichsen, C. N., and Kai, C., and Kawai, J., and
 Nagalakshmi, U., and Wu, J., and Lian, Z., and Lian, J., and Newburger, P., and Zhang,
 X., and Bickel, P., and Mattick, J. S., and Carninci, P., and Hayashizaki, Y., and
 Weissman, S., and Hubbard, T., and Myers, R. M., and Rogers, J., and Stadler, P. F., and

Lowe, T. M., and Wei, C. L., and Ruan, Y., and Struhl, K., and Gerstein, M., and
 Antonarakis, S. E., and Fu, Y., and Green, E. D., and Karaoz, U., and Siepel, A., and
 Taylor, J., and Liefer, L. A., and Wetterstrand, K. A., and Good, P. J., and Feingold, E.
 A., and Guyer, M. S., and Cooper, G. M., and Asimenos, G., and Dewey, C. N., and Hou,
 M., and Nikolaev, S., and Montoya-Burgos, J. I., and Loytynoja, A., and Whelan, S., and
 Pardi, F., and Massingham, T., and Huang, H., and Zhang, N. R., and Holmes, I., and
 Mullikin, J. C., and Ureta-Vidal, A., and Paten, B., and Seringhaus, M., and Church, D.,
 and Rosenbloom, K., and Kent, W. J., and Stone, E. A., and Program, N. C. S., and
 Baylor College of Medicine Human Genome Sequencing, C., and Washington University
 Genome Sequencing, C., and Broad, I., and Children's Hospital Oakland Research, I., and
 Batzoglou, S., and Goldman, N., and Hardison, R. C., and Haussler, D., and Miller, W.,
 and Sidow, A., and Trinklein, N. D., and Zhang, Z. D., and Barrera, L., and Stuart, R.,
 and King, D. C., and Ameer, A., and Enroth, S., and Bieda, M. C., and Kim, J., and
 Bhinge, A. A., and Jiang, N., and Liu, J., and Yao, F., and Vega, V. B., and Lee, C. W.,
 and Ng, P., and Shahab, A., and Yang, A., and Moqtaderi, Z., and Zhu, Z., and Xu, X.,
 and Squazzo, S., and Oberley, M. J., and Inman, D., and Singer, M. A., and Richmond, T.
 A., and Munn, K. J., and Rada-Iglesias, A., and Wallerman, O., and Komorowski, J., and
 Fowler, J. C., and Couttet, P., and Bruce, A. W., and Dovey, O. M., and Ellis, P. D., and
 Langford, C. F., and Nix, D. A., and Euskirchen, G., and Hartman, S., and Urban, A. E.,
 and Kraus, P., and Van Calcar, S., and Heintzman, N., and Kim, T. H., and Wang, K.,
 and Qu, C., and Hon, G., and Luna, R., and Glass, C. K., and Rosenfeld, M. G., and
 Aldred, S. F., and Cooper, S. J., and Halees, A., and Lin, J. M., and Shulha, H. P., and
 Zhang, X., and Xu, M., and Haidar, J. N., and Yu, Y., and Ruan, Y., and Iyer, V. R., and

Green, R. D., and Wadelius, C., and Farnham, P. J., and Ren, B., and Harte, R. A., and Hinrichs, A. S., and Trumbower, H., and Clawson, H., and Hillman-Jackson, J., and Zweig, A. S., and Smith, K., and Thakkapallayil, A., and Barber, G., and Kuhn, R. M., and Karolchik, D., and Armengol, L., and Bird, C. P., and de Bakker, P. I., and Kern, A. D., and Lopez-Bigas, N., and Martin, J. D., and Stranger, B. E., and Woodroffe, A., and Davydov, E., and Dimas, A., and Eyraas, E., and Hallgrimsdottir, I. B., and Huppert, J., and Zody, M. C., and Abecasis, G. R., and Estivill, X., and Bouffard, G. G., and Guan, X., and Hansen, N. F., and Idol, J. R., and Maduro, V. V., and Maskeri, B., and McDowell, J. C., and Park, M., and Thomas, P. J., and Young, A. C., and Blakesley, R. W., and Muzny, D. M., and Sodergren, E., and Wheeler, D. A., and Worley, K. C., and Jiang, H., and Weinstock, G. M., and Gibbs, R. A., and Graves, T., and Fulton, R., and Mardis, E. R., and Wilson, R. K., and Clamp, M., and Cuff, J., and Gnerre, S., and Jaffe, D. B., and Chang, J. L., and Lindblad-Toh, K., and Lander, E. S., and Koriabine, M., and Nefedov, M., and Osoegawa, K., and Yoshinaga, Y., and Zhu, B., and de Jong, P. J.

(2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature* 447, 799-816.

- [108] Fox, J. (2000) *Nonparametric Simple Regression : Smoothing Scatterplots*, In Quantitative Applications in the Social Sciences. Thousand Oaks, Calif : Sage
- [109] Pollock, D. S. G., and Green, R. C. (1999) *Handbook of Time Series Analysis, Signal Processing, and Dynamics (Signal Processing and its Applications)*, Academic Press.
- [110] Reinsch, C. H. (1967) Smoothing by Spline Functions, *Numerische Mathematik*.
- [111] Pollock, D. S. G. (1999) *A Handbook of Time-series Analysis, Signal Processing and Dynamics*, Academic, San Diego, CA.

- [112] Bengtsson, H. (2004) aroma - An R Object-oriented Microarray Analysis environment, Centre for Mathematical Sciences, Lund University.
- [113] Bengtsson, H., and Hossjer, O. (2006) Methodological study of affine transformations of gene expression data with proposed robust non-parametric multi-dimensional normalization method, *BMC Bioinformatics* 7, 100.
- [114] Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database, C. (2011) The sequence read archive, *Nucleic acids research* 39, D19-21.
- [115] McKenna, R., Xia, D., Willingmann, P., Ilag, L. L., Krishnaswamy, S., Rossmann, M. G., Olson, N. H., Baker, T. S., and Incardona, N. L. (2014) Atomic structure of single-stranded DNA bacteriophage Φ X174 and its functional implications, *Nature*.
- [116] Sanger, F., Air, G., Barrell, B., Brown, N., Coulson, A., Fiddes, C., Hutchison, C., Slocombe, P., and Smith, M. (1977) Nucleotide sequence of bacteriophage phi X174 DNA, *Nature*.
- [117] Shaw, D., Walker, J., Northrop, F., Barrell, B., Godson, G., and Fiddes, J. (1978) Gene K, a new overlapping gene in bacteriophage G4, *Nature*.
- [118] Dixon, W. J. (1950) Analysis of extreme values, *The Annals of Mathematical Statistics*.
- [119] Dixon, W. J. (1951) Ratios involving extreme values, *The Annals of Mathematical Statistics*.
- [120] Grubbs, F. E. (1950) Sample criteria for testing outlying observations, *The Annals of Mathematical Statistics*.
- [121] Wei, Y., Pere, A., Koenker, R., and He, X. (2006) Quantile regression methods for reference growth charts, *Statistics In Medicine*.

- [122] Akima, H. (1970) A new method of interpolation and smooth curve fitting based on local procedures, *Journal of the Association for Computing Machinery*.
- [123] Knott, G. D. (2000) *Interpolating Cubic Splines*, Springer-Science+Business Media, LLC, New York.
- [124] Chen JQ, W. Y., Yang H, Bergelson J, Kreitman M, Tian D. (2009) Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria, *Mol Biol Evol*.
- [125] Denver DR, M. K., Lynch M, Thomas WK. (2004) High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome, *Nature*.
- [126] Britten RJ, R. L., Williams J, Cameron RA. (2003) Majority of divergence between closely related DNA samples is due to indels, *Proc Natl Acad Sci USA*.
- [127] Sun X, Z. Y., Yang S, Chen JQ, Hohn B, Tian D. (2008) Insertion DNA promotes ectopic recombination during meiosis in *Arabidopsis*, *Mol Biol Evol*.
- [128] AM, N. (2001) Complex patterns of plastid 16S rRNA gene evolution in nonphotosynthetic green algae, *Mol Evol*.
- [129] Zhang Z, G. M. (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes, *Nucleic Acids Res*.
- [130] Petrov DA, S. T., Johnston JS, Hartl DL, Shaw KL. (2000) Evidence for DNA loss as a determinant of genome size. *Science, Science*
- [131] TR, G. (2004) Insertion-deletion biases and the evolution of genome size, *Gene*.
- [132] Zoghbi HY, O. H. (2000) Glutamine repeats and neurodegeneration, *Annu Rev Neurosci*.
- [133] Kondrashov AS, R. I. (2004) Context of deletions and insertions in human coding sequences, *Hum Mutat*.

- [134] Ferro P, d. E. R., Pfeffer U. (2001) Are there CAG repeat expansion-related disorders outside the central nervous system?, *Brain Res Bull.*
- [135] Burch CL, D. R., Stein DC. (1997) Antigenic variation in *Neisseria gonorrhoeae*: production of multiple lipooligosaccharides, *J Bacteriol.*
- [136] Rocha EP, B. A. (2002) Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution, *Nucleic Acids Res.*
- [137] Yang H, Z. Y., Peng C, Chen JQ, Tian D. (2010) Important role of indels in somatic mutations of human cancer genes, *BMC Med Genet.*
- [138] Lee, H., Doak, T. G., Popodi, E., Foster, P. L., and Tang, H. (2016) Insertion sequence-caused large-scale rearrangements in the genome of *Escherichia coli*, *Nucleic Acids Res* 44, 7109-7119.
- [139] <https://ghr.nlm.nih.gov/primer/mutationsanddisorders/structuralchangesReference>, G. H. Can changes in the structure of chromosomes affect health and development?, U.S. National Library of Medicine.
- [140] Korbelt, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z. T., Tanzer, A., Saunders, A. C. E., Chi, J. X., Yang, F. T., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M., and Snyder, M. (2007) Paired-end mapping reveals extensive structural variation in the human genome, *Science* 318, 420-426.
- [141] Chen, J., Kim, Y. C., Jung, Y. C., Xuan, Z. Y., Dworkin, G., Zhang, Y. M., Zhang, M. Q., and Wang, S. M. (2008) Scanning the human genome at kilobase resolution, *Genome Res* 18, 751-762.

- [142] Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tuzun, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. (2008) Mapping and sequencing of structural variation from eight human genomes, *Nature* 453, 56-64.
- [143] Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z. D., Snyder, M., and Gerstein, M. B. (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data, *Genome Biol* 10.
- [144] Robert McKenna, D. X., Peter Willingmann, Leodevico L. Ilag, S. Krishnaswamy, Michael G. Rossmann, Norman H. Olson, Timothy S. Baker, Nino L. Incardona. (1992) Atomic structure of single-stranded DNA bacteriophage Φ X174 and its functional implications, *Nature*.
- [145] Xuan Zhu, J. W., Bo Peng, Sanjay Shete. (2016) Empirical estimation of sequencing error rates using smoothing splines, *BMC Bioinformatics*.
- [146] Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011) Sequence-specific error profile of Illumina sequencers, *Nucleic Acids Res* 39, e90.

- [147] Ursell, L. K., Metcalf, J. L., Parfrey, L. W., and Knight, R. (2012) Defining the human microbiome, *Nutr Rev* 70 Suppl 1, S38-44.
- [148] Li, H. (2015) Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis, *Annual Review of Statistics and Its Application* 2, 73-94.
- [149] Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., and Gordon, J. I. (2005) Host-bacterial mutualism in the human intestine, *Science* 307, 1915-1920.
- [150] http://hmpdacc.org/micro_analysis/microbiome_analyses.phpAbout HMP Metagenomic Sequencing & Analysis
- [151] Sandrine P Claus, H. G., Sandrine Ellero-Simatos. (2016) The gut microbiota: a major player in the toxicity of environmental pollutants?, *npj Biofilms and Microbiomes* 2.
- [152] [https://www.nih.gov/news-events/news-releases/nih-human-microbiome-project-defines-normal-bacterial-makeup-body\(2012\)](https://www.nih.gov/news-events/news-releases/nih-human-microbiome-project-defines-normal-bacterial-makeup-body(2012)) NIH Human Microbiome Project defines normal bacterial makeup of the body
- [153] Tang, W. H. W., and Hazen, S. L. (2014) The contributory role of gut microbiota in cardiovascular disease, *J Clin Invest* 124, 4204-4211.
- [154] Dulal, S., and Keku, T. O. (2014) Gut Microbiome and Colorectal Adenomas, *Cancer J* 20, 225-231.
- [155] Woo, P. C. Y., Lau, S. K. P., Teng, J. L. L., Tse, H., and Yuen, K. Y. (2008) Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories, *Clin Microbiol Infect* 14, 908-934.
- [156] Hamady, M., and Knight, R. (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges, *Genome Res* 19, 1141-1152.

- [157] Fiona Stewart, E. Y. (2014) Addressing Challenges in Microbiome DNA Analysis, *NEB UK Expressions*.
- [158] Brooks, J. P. (2016) Challenges for case-control studies with microbiome data, *Ann Epidemiol* 26, 336-341 e331.
- [159] Yang, L., Chaudhary, N., Baghdadi, J., and Pei, Z. (2014) Microbiome in reflux disorders and esophageal adenocarcinoma, *Cancer J* 20, 207-210.
- [160] Gilles, A., Meglecz, E., Pech, N., Ferreira, S., Malausa, T., and Martin, J. F. (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing, *BMC Genomics* 12, 245.
- [161] Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J. A., Smith, G., and Knight, R. (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms, *ISME J* 6, 1621-1624.
- [162] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities, *Appl Environ Microbiol* 75, 7537-7541.
- [163] Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J.,

- Yatsuneneko, T., Zaneveld, J., and Knight, R. (2010) QIIME allows analysis of high-throughput community sequencing data, *Nat Methods* 7, 335-336.
- [164] Erica Plummer, J. T., Dieter M. Bulach, Suzanne M. Garland, Sepehr N Tabrizi. (2015) A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data, *Journal of Proteomics & Bioinformatics* 8, 283-291.
- [165] Navas-Molina, J. A., Peralta-Sanchez, J. M., Gonzalez, A., McMurdie, P. J., Vazquez-Baeza, Y., Xu, Z. J., Ursell, L. K., Lauber, C., Zhou, H. W., Song, S. J., Huntley, J., Ackermann, G. L., Berg-Lyons, D., Holmes, S., Caporaso, J. G., and Knight, R. (2013) Advancing Our Understanding of the Human Microbiome Using QIIME, *Method Enzymol* 531, 371-444.
- [166] Mir, K., Neuhaus, K., Bossert, M., and Schober, S. (2013) Short Barcodes for Next Generation Sequencing, *Plos One* 8.
- [167] Edgar, R. C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads, *Nature Methods* 10, 996–998.
- [168] [http://www.nature.com/scitable/definition/primer-305\(2014\)](http://www.nature.com/scitable/definition/primer-305(2014)) Scitable by nature education.
- [169] De Beuf, K., De Schrijver, J., Thas, O., Van Criekinge, W., Irizarry, R. A., and Clement, L. (2012) Improved base-calling and quality scores for 454 sequencing based on a Hurdle Poisson model, *Bmc Bioinformatics* 13.
- [170] Si, X. F., Baselga, A., Leprieur, F., Song, X., and Ding, P. (2016) Selective extinction drives taxonomic and functional alpha and beta diversities in island bird assemblages, *J Anim Ecol* 85, 409-418.

- [171] McMurdie, P. J., and Holmes, S. (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data, *Plos One* 8.
- [172] Hill, M. O. (1973) Diversity and Evenness: A Unifying Notation and Its Consequences, *Ecology* 54, 427-432.
- [173] Lande, R. (1996) Statistics and partitioning of species diversity, and similarity among multiple communities, *Oikos* 76, 5-13.
- [174] Sandra D. Williamson, K. B. (2013) Species richness and diversity of a terrestrial insular environment: serpentine of the Barberton Greenstone Belt, South Africa, *International Journal of Biodiversity and Conservation* 5(5), 296-310.
- [175] Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., Meiners, T., Muller, C., Obermaier, E., Prati, D., Socher, S. A., Sonnemann, I., Waschke, N., Wubet, T., Wurst, S., and Rillig, M. C. (2014) Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories, *Ecol Evol* 4, 3514-3524.
- [176] Saucedo-Garcia, A., Anaya, A. L., Espinosa-Garcia, F. J., and Gonzalez, M. C. (2014) Diversity and Communities of Foliar Endophytic Fungi from Different Agroecosystems of *Coffea arabica* L. in Two Regions of Veracruz, Mexico, *Plos One* 9.
- [177] Williams, V. L., Witkowski, E. T. F., and Balkwill, K. (2005) Application of diversity indices to appraise plant availability in the traditional medicinal markets of Johannesburg, South Africa, *Biodivers Conserv* 14, 2971-3001.
- [178] Colwell, R. K. (2009) Biodiversity: concepts, patterns, and measurement, In *The Princeton Guide to Ecology* (Levin, S. A., Ed.), pp 257-263, Princeton Univ. Press, Princeton, NJ.

- [179] Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population, *J Anim Ecol* 12, 42-58.
- [180] Hughes, J. B., Hellmann, J. J., Ricketts, T. H., and Bohannan, B. J. (2001) Counting the uncountable: statistical approaches to estimating microbial diversity, *Appl Environ Microbiol* 67, 4399-4406.
- [181] Chao, A., Ma, M. C., and Yang, M. C. K. (1993) Stopping Rules and Estimation for Recapture Debugging with Unequal Failure Rates, *Biometrika* 80, 193-201.
- [182] Gotelli, N. J. a. R. K. C. (2010) Estimating species richness, In *Frontiers in measuring biodiversity* (Anne E. Magurran, B. J. M., Ed.), pp 39-54, Oxford University New York.
- [183] Chao, A., Chazdon, R. L., Colwell, R. K., and Shen, T. J. (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data, *Ecol Lett* 8, 148-159.
- [184] Soininen, J. (2010) Species Turnover along Abiotic and Biotic Gradients: Patterns in Space Equal Patterns in Time?, *Bioscience* 60, 433-439.
- [185] Koleff, P., Gaston, K. J., and Lennon, J. J. (2003) Measuring beta diversity for presence-absence data, *J Anim Ecol* 72, 367-382.
- [186] http://biology-forums.com/definitions/index.php/Species_turnoverIn *Biology-forums.com*.
- [187] Shirkhorshidi, A. S., Aghabozorgi, S., and Wah, T. Y. (2015) A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data, *Plos One* 10, e0144059.
- [188] Emran, S. M., and Ye, N. (2002) Robustness of Chi-square and Canberra distance metrics for computer intrusion detection, *Qual Reliab Eng Int* 18, 19-28.

- [189] Giuseppe Jurman, S. R., Roberto Visintainer, Cesare Furlanello. (2009) Canberra distance on ranked lists, *Advances in Ranking–NIPS 09 Workshop*, 22-27.
- [190] Hennig, C., and Hausdorf, B. (2006) Design of dissimilarity measures: A new dissimilarity between species distribution areas, In *Stud Class Data Anal* (Batagelj, V., Bock, H.-H., Ferligoj, A., and Žiberna, A., Eds.), pp 29-37, Springer Berlin Heidelberg, Berlin, Heidelberg.
- [191] Anderson, M. J., and Millar, R. B. (2004) Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern New Zealand, *J Exp Mar Biol Ecol* 305, 191-221.
- [192] Horn, H. S. (1966) Measurement of Overlap in Comparative Ecological Studies, *Am Nat* 100, 419-&.
- [193] Ali Seyed Shirkhorshidi, S. A., Teh Ying Wah. (2015) A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data, *PLOS ONE*.
- [194] Syed Masum Emran, N. Y. (2002) Robustness of Chi-Square and Canberrra distance metreics for computer intrusion detection, *Quality and Reliability Engineering International*.
- [195] Giuseppe Jurman, S. R., Roberto Visintainer, Cesare Furlanello. (2009) Canberra distance on ranked lists, *Advances in Ranking–NIPS 09 Workshop*.
- [196] Christian Hennig, B. H. (2006) Design of Dissimilarity Measures: A New Dissimilarity Between Species Distribution Areas In *Data Science and Classification* (Prof. Dr. Vladimir Batagelj, P. D. H.-H. B., Prof. Dr. Anuška Ferligoj, Aleš Žiberna, Ed.), pp 57-203, Springer Berlin Heidelberg.

- [197] Tulloss, R. E. (1997) Assessment of Similarity Indices for Undesirable Properties and a new Tripartite Similarity Index Based on Cost Functions, *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*.
- [198] Anderson, M. J. a. M., R.B. . (2004) Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern New Zealand. , *Journal of Experimental Marine Biology and Ecology* 191–221.
- [199] Clark, P. J. (1952) An Extension of the Coefficient of Divergence for Use with Multiple Characters, *American Society of Ichthyologists and Herpetologists (ASIH)* 2, 61-64.
- [200] Horn, H. S. (1966) Measurement of "Overlap" in comparative ecological studies. , *The American Naturalist*, 419-424.
- [201] Horn, H. S. (1966) Measurement of "Overlap" in Comparative Ecological Studies, *The American Naturalist* 100.
- [202] V. Arora a, S. K., S.S. Ghantoji, H.L. DuPont, K.W. Garey. (2011) High Horn's index score predicts poor outcomes in patients with Clostridium difficile infection, *Journal of Hospital Infection* 79.
- [203] Clarke, K. R., Somerfield, P. J., and Chapman, M. G. (2006) On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages, *J Exp Mar Biol Ecol* 330, 55-80.
- [204] Fukuyama, J., McMurdie, P. J., Dethlefsen, L., Relman, D. A., and Holmes, S. (2012) Comparisons of distance methods for combining covariates and abundances in microbiome studies, *Pac Symp Biocomput*, 213-224.
- [205] Lozupone, C. A., and Knight, R. (2007) Global patterns in bacterial diversity, *P Natl Acad Sci USA* 104, 11436-11440.

- [206] Schloss, P. D. (2008) Evaluating different approaches that test whether microbial communities have the same structure, *Isme Journal* 2, 265-275.
- [207] Ives, A. R., and Helmus, M. R. (2010) Phylogenetic metrics of community similarity, *Am Nat* 176, E128-142.
- [208] McArdle, B. H., and Anderson, M. J. (2001) Fitting multivariate models to community data: A comment on distance-based redundancy analysis, *Ecology* 82, 290-297.
- [209] Jun Chen, H. L. (2013) Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis, *Ann. Appl. Stat.*
- [210] Ni Zhao, J. C., Ian M. Carroll, Tamar Ringel-Kulka, Michael P. Epstein, Hua Zhou, Jin J. Zhou, Yehuda Ringel, Hongzhe Li, Michael C. Wu. (2015) Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test, *The American Society of Human Genetics*.
- [211] Sceaaly, J. L., and Welsh, A. H. (2011) Regression for compositional data by using distributions defined on the hypersphere, *J R Stat Soc B* 73, 351-375.
- [212] Kent, J. T. (1982) The Fisher-Bingham Distribution on the Sphere, *J Roy Stat Soc B Met* 44, 71-80.
- [213] Aitchison, J. (1982) The Statistical-Analysis of Compositional Data, *J Roy Stat Soc B Met* 44, 139-177.
- [214] Shi, P. X., Zhang, A. R., and Li, H. Z. (2016) Regression Analysis for Microbiome Compositional Data, *Ann Appl Stat* 10, 1019-1040.
- [215] Fisher, C. K., and Mehta, P. (2014) Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression, *Plos One* 9.

- [216] Chen, E. Z., and Li, H. Z. (2016) A two-part mixed-effects model for analyzing longitudinal microbiome compositional data, *Bioinformatics* 32, 2611-2617.
- [217] Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. Z., and Wu, M. C. (2015) Testing in Microbiome-Profilng Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test, *Am J Hum Genet* 96, 797-807.
- [218] Gevers, D., Knight, R., Petrosino, J. F., Huang, K., McGuire, A. L., Birren, B. W., Nelson, K. E., White, O., Methe, B. A., and Huttenhower, C. (2012) The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome, *Plos Biol* 10.
- [219] <http://dx.doi.org/10.1101/088666>Edgar, R. C. (2016) Filtering of high-frequency cross-talk in 16S amplicon reads.
- [220] McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., and Hugenholtz, P. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea, *Isme Journal* 6, 610-618.
- [221] <https://www.R-project.org/Team>, R. C. (2015) R: A language and environment for statistical computing, In *R Foundation for Statistical Computing*, Vienna, Austria.
- [222] K. Gerald van den Boogaart, R. T., Matevz Bren. (2014) compositions: Compositional Data Analysis, In *R package version 1.40-1*.
- [223] Jari Oksanen, F. G. B., Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens. (2016) vegan: Community Ecology Package, (Wagner, E. S. a. H., Ed.).

- [224] Paradis, E., Claude, J., and Strimmer, K. (2004) APE: Analyses of Phylogenetics and Evolution in R language, *Bioinformatics* 20, 289-290.
- [225] Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A., and Caporaso, J. G. (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing, *Nature Methods* 10, 57-U11.
- [226] Walters, W. A., Caporaso, J. G., Lauber, C. L., Berg-Lyons, D., Fierer, N., and Knight, R. (2011) PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers, *Bioinformatics* 27, 1159-1161.
- [227] Scott L. Zeger, K. Y. L. (1992) An overview of methods for the analysis of longitudinal data, *Statistics in Medicine*.
- [228] Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000) Using Bayesian networks to analyze expression data, *J Comput Biol* 7, 601-620.
- [229] Yang Ni, F. C. S. V. B. (2016) Sparse Multi-dimensional Graphical Models: A Unified Bayesian Framework, *Journal of the American Statistical Association*.
- [230] Dobra, A., Lenkoski, A., and Rodriguez, A. (2011) Bayesian Inference for General Gaussian Graphical Models With Application to Multivariate Lattice Data, *J Am Stat Assoc* 106, 1418-1433.
- [231] Leng, C. L., and Tang, C. Y. (2012) Sparse Matrix Graphical Models, *Journal of the American Statistical Association* 107, 1187-1200.
- [232] McGeachie, M. J., Sordillo, J. E., Gibson, T., Weinstock, G. M., Liu, Y. Y., Gold, D. R., Weiss, S. T., and Litonjua, A. (2016) Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks, *Sci Rep-Uk* 6.

- [233] Xia, F., Chen, J., Fung, W. K., and Li, H. Z. (2013) A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis, *Biometrics* 69, 1053-1063.
- [234] Kurtz, Z. D., Muller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015) Sparse and Compositionally Robust Inference of Microbial Ecological Networks, *Plos Comput Biol* 11.
- [235] Chen, J., and Li, H. Z. (2013) Kernel Methods for Regression Analysis of Microbiome Compositional Data, *Springer P Math Stat* 55, 191-201.

VITA

Xuan Zhu was born in March 1989 in China. She graduated from The Ohio State University, Columbus, with a major in actuarial science and a minor in statistics. In January 2013, she officially started her Ph.D. training in the Biostatistics, Bioinformatics and Systems Biology Program at The University of Texas MD Anderson Cancer Center UT Health Graduate School of Biomedical Sciences. She conducted her thesis research under the supervision of Dr. Sanjay S. Shete, a professor in the Department of Biostatistics at The University of Texas MD Anderson Cancer Center. She expects to finish her Ph.D. in May 2017.