

DEVELOPING STATISTICAL AND ALGORITHMIC METHODS FOR
SHOTGUN METAGENOMICS AND TIME SERIES ANALYSIS

by

Li Charlie Xia

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(COMPUTATIONAL BIOLOGY AND BIOINFORMATICS)

May 2013

Copyright 2013

Li Charlie Xia

Dedications

To my wife Ge, my parents Songjuan and Xianhui, and to my professors and friends.

Acknowledgements

I am most grateful to my advisor, Prof. Fengzhu Sun. Without his insightful vision and continuous support, in particular in the GRAMMy project, I would not have come this far. I am also deeply impressed and influenced by his rigorous academic attitude, for instance, careful examination of every line of my manuscripts, which I will carry on these merits I learned from him to my future research and life.

I also would like to thank Prof. Jed Fuhrman, who has been a proactive advisor and collaborator on the experimental side. He always comes with challenging practical questions that have led my work onto the eLSA project. And because of his push-forward, the method has received a wide acceptance in the marine community.

I also would like to thank Prof. Ting Chen, who often brings creative ideas to my project through our joint group meetings. Prof. Chen is also an ardent researcher, bringing us to various branches of computational biology and bioinformatics through a number of journal clubs, which truly broaden my horizon.

I am also grateful to Profs. Andrew D. Smith, Liang Chen and Jay C.C. Kuo who served as my PhD guidance committee and outsider members. The questions and comments raised by them directed me to further improve my works.

I want to express my thanks to Profs. Sun and Chen's current and previous group members: Lin, Xiting, Joyce, Jing, Quan, Wangshu, Xuemei, Yang-ho, Kjong, Sungjie, Xiaolin, Tade and others, and Prof. Fuhrman group, Joshua, Jacob, Cheryl, Rohan and others I cannot enumerate. Thank you for your company through my PhD journey. Finally, I want to express my thanks to my family, other USC faculty members and friends for sharing with me a period of wonderful time in Los Angeles. Thank you all!

Table of Contents

Dedications	ii
Acknowledgements	iii
List of Tables	vii
List of Figures	ix
Abstract	xiii
Chapter 1: Introduction	1
1.1 High-throughput molecular biology	1
1.1.1 Molecular biology	1
1.1.2 Molecular microbial ecology	3
1.1.3 High-throughput era	4
1.2 Challenges of high-throughput data	6
1.2.1 Shotgun metagenomics	6
1.2.2 Molecular time series data	7
1.2.3 Computational approaches	8
1.3 Our approaches	9
1.3.1 GRAMMy for shotgun metagenomics	9
1.3.2 eLSA for molecular time series	10
Chapter 2: GRAMMy based on shotgun metagenomic reads	14
2.1 Background	14
2.2 The GRAMMy framework and algorithm	18
2.2.1 The GRAMMy framework	18
2.2.2 A finite mixture model	19
2.2.3 Estimation of GRA using Expectation Maximization (EM)	21
2.2.4 Read probability approximations	22
2.2.5 Standard errors for GRA estimates	23
2.2.6 Higher level taxonomic statistics	24
2.3 Materials and methods	25
2.3.1 Real read sets and reference genome sets	25
2.3.2 Read filtering and assignment procedures	26

2.3.3	Numerical error measures	27
2.3.4	Hierarchical biclustering	27
2.4	Simulation studies	28
2.4.1	Simulated read benchmarks	28
2.4.2	Artificial metagenomes with real reads	32
2.5	Application to real datasets	33
2.5.1	Meta-analysis of human gut metagenomes	33
2.5.2	The acid mine drainage data set	38
2.6	Discussion	39
Chapter 3: eLSA of molecular time series data		50
3.1	Background	50
3.2	Mathematical modeling and the eLSA algorithm	53
3.2.1	Local similarity analysis with replicates	53
3.2.2	Different ways of summarizing the replicate data	55
3.2.3	Bootstrap confidence interval for the LS score	56
3.2.4	Data normalization	57
3.2.5	Permutation test to evaluate the statistical significance	57
3.2.6	Computation complexity and implementation	58
3.2.7	The eLSA analysis pipeline	58
3.3	Materials and methods	60
3.3.1	Pearson's correlation coefficient-based analysis	60
3.3.2	False discovery rate (FDR) estimation	61
3.4	Simulation studies	62
3.4.1	Time-delayed association	62
3.4.2	Association within a subinterval	63
3.4.3	Different summarizing function	63
3.4.4	Running time comparison	64
3.5	Application to real data	65
3.5.1	Microbial community data analysis	65
3.5.2	Gene expression data analysis	68
3.6	Discussion	70
Chapter 4: Future work		79
4.1	Future work for GRAMMy	79
4.2	Future work for eLSA	82
4.3	Molecular microbial ecology analysis pipeline	83
Bibliography		87
Appendix A		
	Technical derivations for GRAMMy	97
A.1	Derivation of the GRAMMy EM algorithm	97
A.2	Derivation of the standard errors	101
A.3	Convergence of the GRAMMy EM algorithm	102

Appendix B	
GRAMMy simulated studies	104
B.1 Simulated read sets	104
B.2 Performance evaluation for simulations	106

List of Tables

1.1	Next generation sequencing technologies. Data from Metzker et al. [65]. Data with * are from the web.	11
1.2	Published shotgun metagenomics datasets. Datasets with * are datasets used in this study.	12
1.3	Published shotgun metagenomics tools. Softwares with * are compared in this study.	12
1.4	Ongoing and past microbial genome sequencing projects. Data from GOLD [58].	12
2.1	Comparison of estimation accuracy. A summary of Relative Root Mean Square Error (RRMSE) and Average Relative Error (AVGRE) measured from MEGAN-based, GAAS and GRAMMy ('map') estimates of simLC, simMC and simHC subsets of the FAMEs data. GRAMMy ('map') has the lowest error rate for both error measures across all the subsets.	33
2.2	Summary statistics for the metagenomic datasets. Median (Med.), minimum (Min.) and maximum (Max.) of mapped rate, ambiguity rate and estimated average genome length for the samples: two from U.S. adult human gut ('hg'), 13 from Japanese human gut ('jhg'), 18 from U.S. twin families human gut ('uhg') and 1 from acid mine drainage ('amd') are shown. Two reference genome sets, 'HGS', 'AMD', were used for human gut samples ('hg', 'jhg', 'uhg') and the acid mine drainage sample ('amd'), respectively.	35
3.1	Mean and standard error of the estimated LS score. The values are calculated based on 1000 simulations. 'se.' indicates standard error and 'na.' indicates not applicable.	64
3.2	Significant associations found in real datasets. Numbers of significant associations found by the extended Local Similarity Analysis (eLSA) and Pearson's Correlation Coefficient (PCC) by controlling both the p-value (P) and the q-value (Q). The p-values for eLSA were evaluated by permutations and p-values for PCC was calculated based on the t -distribution.	66

- 3.3 Top LS scores from the microbial community data. The 5 positive and 5 negative highest absolute LS Scores from associations uniquely found by eLSA in the microbial community dataset. The columns in succession are X (first factor), Y (second factor), LS (Local Similarity score), Xs (start of the best alignment in the first sequence), Ys (start of the best alignment in the second sequence), Len (alignment length), D (shift of the second sequence compared to the first sequence, -: X is ahead of Y, +: otherwise), P (p-value for the LS score, 0. stands for $P < 0.005$), PCC (Pearsons Correlation Coefficient), Ppcc (P-value for PCC), Q (q-value calculated for P, 0. stands for $Q < 0.005$), Qpcc (q-value for Ppcc). . . . 67
- 3.4 Top LS scores from the *C. elegans* gene-expression data. The 5 positive and 5 negative highest absolute LS Scores from the *C. elegans* gene expression dataset The columns in succession are X (first factor), Y (second factor), LS (Local Similarity score), lowCI (CI is lower bound), upCI (CI is upper bound), Xs (start of the best alignment in the first sequence), Ys (start of the best alignment in the second sequence), Len (alignment length), D (shift of the second sequence compared to the first sequence, -: X is ahead of Y, +: otherwise), P (p-value for the LS score, 0. stands for $P < 0.005$), PCC (Pearsons Correlation Coefficient), Ppcc (P-value for PCC), Q (q-value calculated for P, 0. stands for $Q < 0.005$), Qpcc (q-value for Ppcc). 71

List of Figures

1.1	Microbial community DNA sequencing. Figure from DeLong et al. [26].	13
2.1	The GRAMMy model. A schematic diagram of the finite mixture model underlies the GRAMMy framework for shotgun metagenomics. In the figure, ‘iid’ stands for “independent identically distributed”.	42
2.2	The GRAMMy flowchart. A typical flowchart of GRAMMy analysis pipeline employs ‘map’ and ‘ <i>k</i> -mer’ assignment.	43
2.3	Frequent species for human gut metagenomes. The 99 species occurring in at least 50% of the 33 human gut samples with a minimum relative abundance of 0.05% were selected. ‘gut_HGS_90’ indicates that the human gut (‘gut’) read sets were mapped to the reference genome set (‘HGS’) with a identity rate cut-off at 90% (‘90’).	44
2.4	Heatmap biclustering of human gut metagenomes. ‘gut_HGS_90’ indicates that the human gut (‘gut’) read sets were mapped to the reference genome set (‘HGS’) with a identity rate cut-off at 90% (‘90’). The bottom labels indicate human gut samples. The top right legend shows the color coding for columns indicating the sample age category and dataset origin. The bottom right legend shows color coding for rows indicating the top 4 most abundant phyla in human gut. The relative abundance for each sample is normalized by a rank transformation.	45
2.5	GRAMMy estimates of GRAs for the acid mine drainage data. Estimated relative abundance for each strain is shown as a percentage. The first two strains dominate the sample.	46
2.6	Running time comparison. GRAMMy is the fastest in all cases as compared to MEGAN and GAAS in processing time. The BLAT mapping time is excluded for all compared tools.	47
2.7	Frequent species for the human gut metagenomes. The 99 species occurring in at least 50% of the 33 human gut samples with a minimum relative abundance of 0.05% were selected. ‘gut_HGS_75’ indicates that the human gut (‘gut’) read sets were mapped to the reference genome set (‘HGS’) with a identity rate cut-off at 75% (‘75’).	48

2.8 Heatmap biclustering of the human gut metagenomes. ‘gut_HGS_90’ indicates that the human gut (‘gut’) read sets were mapped to the reference genome set (‘HGS’) with a identity rate cut-off at 90% (‘90’), while ‘gut_HGS_75’ indicates cut-off at 75%(‘75’). The bottom labels indicate human gut samples. The top right legend shows the color coding for columns indicating the sample age category and dataset origin. The bottom right legend shows color coding for rows indicating the top 4 most abundant phyla in human gut. (A) Heatmap clustering of the ‘gut’ samples, with strains of abundance $\geq 0.05\%$ in at least 50% of samples selected at 75% identity rate cut-off. (B) Heatmap clustering of the ‘gut’ samples, with strains of abundance $\geq 0.01\%$ in at least 50% of samples selected at 90% identity rate cut-off. (C) Heatmap clustering of the ‘gut’ samples, with strains of abundance $\geq 0.1\%$ in at least 50% of samples selected at 90% identity rate cut-off. 49

3.1 The eLSA pipeline. Users start with raw data (matrices of time series) as input and specify their requirements as parameters. The LSA tools subsequently *F*-transform and normalize the raw data and calculate Local Similarity (LS) scores and Pearson’s Correlation Coefficients. The tools then assess the statistical significance (P-values) of these correlation statistics using the permutation test and filter out insignificant results. Finally, the tools construct a partially directed association network from the significant associations. 73

3.2 Examples of simulated associations. (a) An example of simulated time-delayed association series with five replicates is shown, where X (red square) leads Y (blue circle) by three time units. The pattern is not significant by ordinary correlation analysis (PCC=-0.258, $P=0.272$); however, it is captured by local similarity analysis (LS=0.507, $P=0.006$). (b) An example of simulated subinterval association series with five replicates is shown, where X (red square) and Y (blue circle) are associated in the time interval from 6 to 15. The pattern is not significant by ordinary correlation analysis (PCC=0.258, $P=0.273$); however, it is captured by local similarity analysis (LS=0.428, $P=0.028$). 74

3.3 Typical association network from the microbial community data. Round- (brown), square- (blue) and triangle- (green) shaped nodes are bacteria, eukaryotes and environmental factors, respectively. Solid (red) edges are positively associated, while dashed (blue) edges are negatively associated. Arrow indicates the time-delay direction. 75

3.4	Examples of real data associations. (a) Shown are microbe group Bac675 (red square) and Bac609 (blue circle) ARISA abundance time series from the marine microbial community data analysis. Notice that there exists an almost regular yearly pattern where Bac609 leads Bac675 by one month in blooming time. (b) Shown are gene <i>32607</i> (red square) and <i>51986</i> (blue circle) expression level time series from <i>C. elegans</i> gene expression data analysis. Notice that <i>51986</i> leads <i>32607</i> in expression level change throughout the time course.	76
3.5	Node degree distribution of associations in <i>C. elegans</i> analysis. Shown is the node degree distribution of eLSA unique associations in <i>C. elegans</i> analysis. It shows a long-tail distribution with the maximum 189.	77
3.6	Translation initiation factor associations in <i>C. elegans</i> analysis. Shown is the association network of translation initiation factors learned from eLSA analysis. Solid (red) edges are positively associated. Edge labels are LS scores. The factors form a clique as expected.	77
3.7	Submission interface for the LSA web service. Upon submission, the job will perform eLSA analysis on the ‘CommonGenesData’ dataset (12 time spots and 4 replicates) with 200 permutations and 100 bootstraps within a delay limit of 3 units. In addition, by specification, it will use ‘simple’ averaging to summarize replicates and, by designating ‘none’, it will disregard the missing values.	78
4.1	A two-layer mixture model for taxonomic relative abundance estimation.	85
4.2	A molecular microbial ecology analysis pipeline integrating tools developed by our groups.	86
B.1	The convergence of GRAMMy. The estimation errors, as measured by different numerical methods: (A) Relative Root Mean Square Error (RRMSE) in percentage versus Read Number (RN) for different read lengths (RL). (B) Relative Root Mean Square Error (RRMSE), Average Relative Error (AVGRE), Maximum Relative Error (MAXRE), and Distance of Total Variation (DTV) versus Read Number for read length equal 100 bp. GRAMMy (map) was used.	111

B.2	Simulated read set benchmarks. Effects of different perturbations on GRAMMy's estimation: (A) Effects of sequencing errors: results from 'with sequencing error' and 'without sequencing error' read sets are labeled as 'w. Seq Err' and 'wo. Seq Err', respectively. (B) Effects of unknown genomes: results from estimation 'with unknown genomes' and 'without unknown genomes' read sets are labeled as 'w. Unknowns' and 'wo. Unknowns', respectively. (C) Effects of different genome relative abundance distributions: results from more concentrated abundance distribution and less concentrated read sets are labeled as 'steep' and 'flat', respectively. Relative Root Mean Square Error (RRMSE) as a percentage is plotted against Read Number. GRAMMy ('map') was used.	112
B.3	Performance comparison of different methods. The performance comparisons for different estimation methods: (A) MEGAN-based ('MEGAN'), GAAS ('GAAS') and GRAMMy ('map' and ' <i>k</i> -mer') on simulated read sets with sequencing errors at read length 100 bp and 400 bp. (B) 16S-based ('16S'), BLAT hit counting ('BLAT'), <i>rpoB</i> -based (' <i>rpoB</i> ') and GRAMMy ('map'). Relative Root Mean Square Error (RRMSE) as a percentage is plotted against Read Number (<i>RN</i>).	113
B.4	Estimation errors at different taxonomic levels. Average Relative Error (AVGRE) as a percentage is plotted against taxonomic level. The errors gradually decrease from strains to kingdom taxonomic levels.	114

Abstract

Recent developments in experimental molecular techniques, such as microarray, next generation sequencing technologies, have led molecular biology into a high-throughput era with emergent omics research areas, including metagenomics and transcriptomics. Massive-size omics datasets generated and being generated from the experimental laboratories put new challenges to computational biologists to develop fast and accurate quantitative analysis tools. We have developed two statistical and algorithmic methods, GRAMMy and eLSA, for metagenomics and microbial community time series analysis. GRAMMy provides a unified probabilistic framework for shotgun metagenomics, in which maximum likelihood method is employed to accurately compute Genome Relative Abundance of microbial communities using the Mixture Model theory (GRAMMy). We extended the Local Similarity Analysis technique (eLSA) to time series data with replicates, capturing statistically significant local and potentially time-delayed associations. Both methods are validated through simulation studies and their capability to reveal new biology is also demonstrated through applications to real datasets. We implemented GRAMMy and eLSA as C++ extensions to Python, with both superior computational efficiency and easy-to-integrate programming interfaces. GRAMMy and eLSA methods will be increasingly useful tools as new omics researches accelerating their pace.

Chapter 1

Introduction

1.1 High-throughput molecular biology

1.1.1 Molecular biology

On the frontier of modern sciences, molecular biology studies the sciences about biological molecules and their systems, not only individual molecules such as DNA, RNA and proteins, but also interactions between them, including the molecular mechanisms of replication, transcription, translation processes and their regulations [3]. One ultimate goal of molecular biology is to understand how biological systems operate from the perspective of bio-molecules. At the heart of molecular biology is the ‘central dogma’ [25], which describes the general principle regarding the flow of information between DNA, RNA and protein molecules. Since the mid twentieth century, we have witnessed many important discoveries along the line of studying the ‘central dogma’.

In 1953, James D. Watson and Francis Crick first discovered the double helix structure of the DNA molecule [114]. It was closely followed by a series of novel discoveries of other biological molecules, including the first X-ray structure of protein – hemoglobin, the first

tRNA molecule and many others [44, 48, 75]. Bio-molecules like these are the building bricks of the life machinery. Almost concurrently, essential enzymes of replication, transcription, translation processes were also discovered, to name a few, DNA polymerases, RNA polymerases and ribosomes [45, 46, 73]. Bio-molecules like these are the building tools of the life machinery. Bricks and tools are yet not all that needed to make real life, and we were still missing the code of execution, which is encoded in the genome of living organisms. Our curiosity has finally led to the sequencing of the first prokaryote genome *Haemophilus influenzae* in 1997 and the first two human genomes in 2001 [35, 52, 107]. Genomes were expected to be the architecture drafts of most organisms and to contain all the information needed to rebuild the same organism. Equipped with the bricks, tools and architecture drafts, it seems we were ready to understand and manipulate the life machinery.

However, the truth is that we are still far away from the truth. It turns out, there are various types of molecular regulations going on, coordinating DNA, RNA and proteins. Though the regulation mechanisms involved in prokaryotes may be straight-forward coded in their genomes, those of eukaryotes, particularly our human being, are extremely complicated. As the quest for the holy grail goes on, we gradually come to know that molecular regulations occur at different levels and stages, such as transcriptional, post-transcriptional and translational regulations, and operate through different mechanisms, including transcription factor control, alternative splicing, RNA silencing, RNA editing and many types of protein modifications [9, 33, 66, 90]. The possibilities of control explode in combinatorial number, which render the genome as deliberately encrypted as a

massive jigsaw puzzle. And even today, the jigsaw is not yet solved, making molecular biology still a fascinating field to work on, with occasional ground-breaking discoveries.

Moreover, molecular biology is not only a science but also a technology. The latter aspect may even be more related to this dissertation. In fact, numerous molecular technologies have been developed and they have permeated into every corner of modern biological laboratories. In the DNA sequencing technology direction alone, we have moved from Frederick Sanger's chain termination method [87] to the state-of-art high-throughput next generation sequencing (NGS) technologies [77], easily producing billions of base pairs of reads every day with one sequencing machine. Meanwhile, promising third generation technologies have already been proposed, and are making their way to daily labs. Certainly, there are other high-throughput molecular experimental technologies available for different purposes, such as PCR, Microarray, ChIP-chip, ChIP-seq and RNA-seq, some of which have already been widely adopted [4, 10, 69, 79, 112]. All these technologies, new or mature, are collectively bolstering the whole world of biological sciences into a new high-throughput data era.

1.1.2 Molecular microbial ecology

Molecular microbial ecology is a crossbred of ecology and molecular biology, which studies environmental microbial organisms using molecular approaches. Consequently, molecular microbial ecology is one among many fields that have been pushed forward greatly by the recent developments in high-throughput experimental molecular technologies. The biggest paradigm shift in the last 25 years is the transformation of main study subjects

from cultured microbial organisms to naturally occurring uncultured microbial organisms [14].

In 1977, Carl Woese showed the possibility that 16S rRNA could be used to derive evolutionary relationships [115]. In 1985, Norman Pace led the first direct analysis of rRNA sequences in the environment to describe the diversity of microorganisms without culturing [72]. Later, with the further development of PCR technology, Jo Handelsman and colleagues pioneered the idea that total DNA or RNA can be extracted from environmental samples, cloned into a suitable vectors and followed by the analysis using high throughput DNA sequencing of cloned DNA, or using direct sequencing of the original DNA or RNA (see Figure 1.1). Later, they further coined the name for this technique – ‘metagenomics’ [41].

Nowadays, metagenomics and other molecular experimental approaches are routinely applied to microbial communities, with habitats varying from open environmental bodies to sites within extreme surrounding, from human exterior to human interior, and with subjects including viruses, bacteria achaea to eukaryotes [15]. These studies help in discovering new proteins and enzymes, like the bacterial rhodopsin [8] and the Sep-tRNA synthetase [88]; in elucidating the mechanism of human diseases, such as obesity and Crohn’s disease [78, 101, 103]; and in evaluating the biogeochemical cycles of the earth, with direct application to biodiversity conservation and battling global climate changes.

1.1.3 High-throughput era

Meanwhile, at the beginning of this new century, with the aid of high-throughput experimental technologies, like microarray, ChIP-chip, NGS and many others, researchers open

the door of high-throughput studies for molecular biology. We can now not only look at one molecule at a time, but also thousands of molecules all together at a time, whether they are DNAs, RNAs or proteins. The high-throughput trend gives rise to waves of ‘omics’ datasets and jump-starts many new areas of research, such as transcriptomics, proteomics and metagenomics. There are many high-throughput molecular techniques present for molecular microbial ecology as well, such as terminal restriction fragment length polymorphism (TRFLP), automated rRNA intergenic spacer analysis (ARISA), 16/18S rRNA tag sequencing and shotgun metagenomics [5, 34, 91]. Thanks to these technologies, we can not only look at one organism at a time, but also thousands of organisms all together at one time, capturing their overall dynamics in one snapshot.

In this dissertation, we will focus on studying the high-throughput data from molecular microbial ecology, which are sequencing reads from the shotgun metagenomics and the operational taxonomic units (OTUs) time series obtained using different molecular technologies, including ARISA and TRFLP technologies. We will also study some transcriptomics data from microarray gene expression experiments of *C. elegans* [111]. Our major aims are developing new mathematical and computational analysis approaches for such kinds of data. Our new approaches show improved results upon previous studies and yield interesting biological findings. Our approaches are also applicable to similar forms of data from other experimental technologies.

1.2 Challenges of high-throughput data

Mathematical and computational techniques have been playing an important role in facilitating biological researches, giving birth to many useful tools such as BLAST, GENSCAN, etc [1, 16]. However, the specific aims of biological studies and the quantity and quality of molecular experimental data has been changing over the years. Taking the NGS technology (see Table 1.1) based metagenomics for example, we are now facing with not one single genome but a mixture of genomes, and we are handling millions of sequences all at once. The number of reads are three orders larger than the number of reads used to assemble first prokaryote genome back in 1997. Therefore, the new high-throughput data demand new computational approaches and softwares to be developed, which should be faster while remain accurate. In our cases, we would like to develop such methods specifically for shotgun metagenomics and molecular time series data.

1.2.1 Shotgun metagenomics

In metagenomics studies, two experimental approaches are widely used: first, the 16S/18S rRNA or its variable region sequencing; second, the whole genome shotgun sequencing (WGS), or briefly shotgun metagenomics. Shotgun metagenomics sample uncultured microorganisms, randomly shear DNA, and sequence many short reads.

Shotgun metagenomic reads are suitable for both functional analysis and taxonomical analysis, however, the scale of study was previously restricted by hefty sequencing costs. The recent adoption of NGS technologies in shotgun metagenomics studies, helped to

reduce the cost and increase the coverage. As a result, the most recent shotgun metagenomic data have a sharp increase in the number of samples and the average read number per sample, although they also suffer from a reduction in read length, due to the limitations of NGS technologies (see Table 1.2).

1.2.2 Molecular time series data

Molecular time series data from natural environmental samples is another important resource for studying temporal dynamics within microbial communities. The mature molecular technologies that characterizing the spacer length and restriction pattern of the variable regions of 16/18S rRNA, such as TRFLP and ARISA, have been existed for about a decade. There are time series from ARISA and TRFLP technologies describing hundreds of OTUs in several years of time [24, 92]. They already provide a good basis for studying temporal associations.

Riding the recent NGS trend, microbial molecular ecologists are also shifting into new sequencing technologies like Roche/454, which has the capability to generate millions of sequences each sample representing thousands of OTUs simultaneously. This new type of time series data have just started to appear since Gilbert et al. [37], yet are gaining more and more popularity [38]. For the new data from Gilbert et al. [38], we have more than 70 time spots and tens of thousands of OTUs which is two orders larger in number compared to ARISA and TRFLP technology. Though the data have the potential to reveal finer details of the intricate dynamics of microbial communities, currently, there are still a lack of computational approaches to smoothly handle and accurately analyze such time series, where computational biologists can come into play.

There are other time series data as well, from different sources, such as the microarray data from gene expression studies and RNA-seq data from transcriptomics [99, 111]. As compared to microbial ecology data, they are mostly shorter time series yet with an even larger number of factors (genes, mRNAs or transcripts, counterpart to OTUs) Though they are potentially useful for studying temporal associations in transcriptional regulations, there is also a similar lack of methods for analyzing these datasets, especially for discovering complicate time-dependent associations.

1.2.3 Computational approaches

Many computational approaches have been developed in the past to facilitate the analysis of new types of biological data. For instances, BLAST simplified the tedious manual homolog search and GENESCAN eased the manual prediction and curation of genes. In face of current high-throughput data, we also need new computational approaches for analyzing them. Fortunately, many computational tools have already been developed for shotgun metagenomics and molecular time series analysis.

Specifically for shotgun metagenomics, there are already several tools, which are summarized in Table 1.3. However, as we shall see, they have not yet fully met the challenges. We also notice the recent acceleration of microbial reference genomes sequencing projects (see Table 1.4) and the increase of read assignment ambiguities partially due to NGS technologies in the experimental part. Therefore, we are motivated to explicitly model the ambiguities and to utilize the increasingly available reference genomes. Thus we developed the GRAMMy method for the genome relative abundance estimation based on these improvements.

For the analysis aimed at finding time-dependent associations in relatively long time series in molecular biology and microbial ecology, there have been a few methods developed, including the previous local similarity analysis (LSA) approach from our lab [83]. Yet the recent increase of OTU number and the introduction of replicates in the molecular time series data have put new challenges as well as opened new opportunities for developing novel methods for finding more complicated time-dependent association patterns as well as the subsequent dynamics network analysis. Therefore, we are motivated to take advantage of these new features and extended the original LSA into a new eLSA pipeline with improved efficiency and wider applicability.

1.3 Our approaches

1.3.1 GRAMMy for shotgun metagenomics

In Chapter 2, we present the GRAMMy framework and tool we developed for the accurate genome relative abundance estimation, based on shotgun metagenomic reads [121]. Accurate estimation of microbial community composition based on metagenomic sequencing data is fundamental for subsequent metagenomics analysis. Prevalent estimation methods are mainly based on directly summarizing alignment results or its variants; often result in biased and/or unstable estimates. We have developed a unified probabilistic framework (named GRAMMy) by explicitly modeling read assignment ambiguities, genome size biases and read distributions along the genomes. Maximum likelihood method is employed to compute Genome Relative Abundance of microbial communities using the Mixture Model theory (GRAMMy). GRAMMy has been demonstrated to give estimates that

are accurate and robust across both simulated and real read benchmark datasets. We applied GRAMMy to a collection of 34 metagenomic read sets from four metagenomics projects and identified 99 frequent species (minimally 0.5% abundant in at least 50% of the datasets) in the human gut samples. Our results show substantial improvements over previous studies, such as adjusting the over-estimated abundance for *Bacteroides* species for human gut samples, by providing a new reference-based strategy for metagenomic sample comparisons. GRAMMy can be used flexibly with many read assignment tools (mapping, alignment or composition-based) even with low-sensitivity mapping results from huge short-read datasets. It will be increasingly useful as an accurate and robust tool for abundance estimation with the growing size of read sets and the expanding database of reference genomes. The GRAMMy software is freely available from the GRAMMy homepage, which can be accessed at <http://meta.usc.edu/softs/grammy>.

1.3.2 eLSA for molecular time series

In Chapter 3, we present the extended local similarity analysis (eLSA) technique for microbial community and other time series data with replicates [122]. The increasing availability of time series microbial community data from metagenomics and other molecular biological studies has enabled the analysis of large-scale microbial co-occurrence and association networks. Among the many analytical techniques available, the Local Similarity Analysis (LSA) method is unique in that it captures local and potentially time-delayed co-occurrence and association patterns in time series data that cannot otherwise be identified by ordinary correlation analysis. However LSA, as originally developed, does not consider time series data with replicates, which hinders the full exploitation

of available information. With replicates, it is possible to understand the variability of local similarity (LS) score and to obtain its confidence interval. We extended our LSA technique to time series data with replicates and termed it extended LSA, or eLSA. Simulations showed the capability of eLSA to capture subinterval and time-delayed associations. We implemented the eLSA technique into an easy-to-use analytic software package. The software pipeline integrates data normalization, statistical correlation calculation, statistical significance evaluation, and association network construction steps. We applied the eLSA technique to microbial community and gene expression datasets, where unique time-dependent associations were identified. The extended LSA analysis technique was demonstrated to reveal statistically significant local and potentially time-delayed association patterns in replicated time series data beyond that of ordinary correlation analysis. These statistically significant associations can provide insights to the real dynamics of biological systems. The newly designed eLSA software efficiently streamlines the analysis and is freely available from the eLSA homepage, which can be accessed at <http://meta.usc.edu/softs/lsa>.

Platform	RL	Days/Run	Gb/Run	Machine(\$)	HG Reseq(\$)
Sanger	800	24runs/d*	2Mb/day*	95,000*	70,000,000
454	330	.35	.45	500,000	1,000,000
Solexa	75	4	18	540,000	250,000
Solid	50	7	30	595,000	60,000
Helicos	32	8	37	999,000	48,000

Table 1.1: Next generation sequencing technologies. Data from Metzker et al. [65]. Data with * are from the web.

Dataset	Technology	RLxRN	Ref
Sargasso Sea	Sanger	818x25k	[108]
Acid Mine Drainage*	Sanger	737x103,462	[104]
Minnesota Soil	Sanger	700x150,000	[100]
3 samples Whale Falls	Sanger	700x100k ea.	[100]
2 samples Human Distal Gut*	Sanger	800x70,000 ea.	[39]
Lean and Obs Mice Gut	454 GS20	93x700k ea.	[103]
Termite Symbiosis	Sanger	750x106,000	[113]
44 samples Global Ocean (GOS)	Sanger	750x70k ea.	[85]
45 samples Nine biomes	454 GS20	100x200k ea.	[29]
31 Lean and Obs Twin Family *	454 GS20	200x1m ea.	[101]
124 samples Human Gut	Solexa	75x60m ea.	[78]

Table 1.2: Published shotgun metagenomics datasets. Datasets with * are datasets used in this study.

Name	Input	Sequence Assignment			Classify	Abundance	Ref
		Blast	Map	k-mer			
Tetra	Contigs	no	no	yes	yes	no	[98]
PhyloPythia	Contigs	no	no	yes	yes	no	[64]
MEGAN*	Reads	yes	no	no	some	indirect	[47]
CompostBin	Reads	no	no	yes	yes	no	[20]
GAAS*	Reads	yes	no	no	no	yes	[2]
Phymm	Reads	yes	no	yes	yes	no	[13]
GRAMMy*	Reads	yes	yes	yes	soft	yes	[121]

Table 1.3: Published shotgun metagenomics tools. Softwares with * are compared in this study.

Name	Since	# of G	Support
10,000 Microbial Genome	2009	10,000	CAS-BGI
Human Microbiome Project (HMP)	2007	hundreds	NIH-NHGRI
Genomic Encyclopedia of Bacteria and Archeae (GEBA)	2007	100	DOE-JGI
Marine Microbiome Init. (MMI)	2004	155	Moore, JCVI

Table 1.4: Ongoing and past microbial genome sequencing projects. Data from GOLD [58].

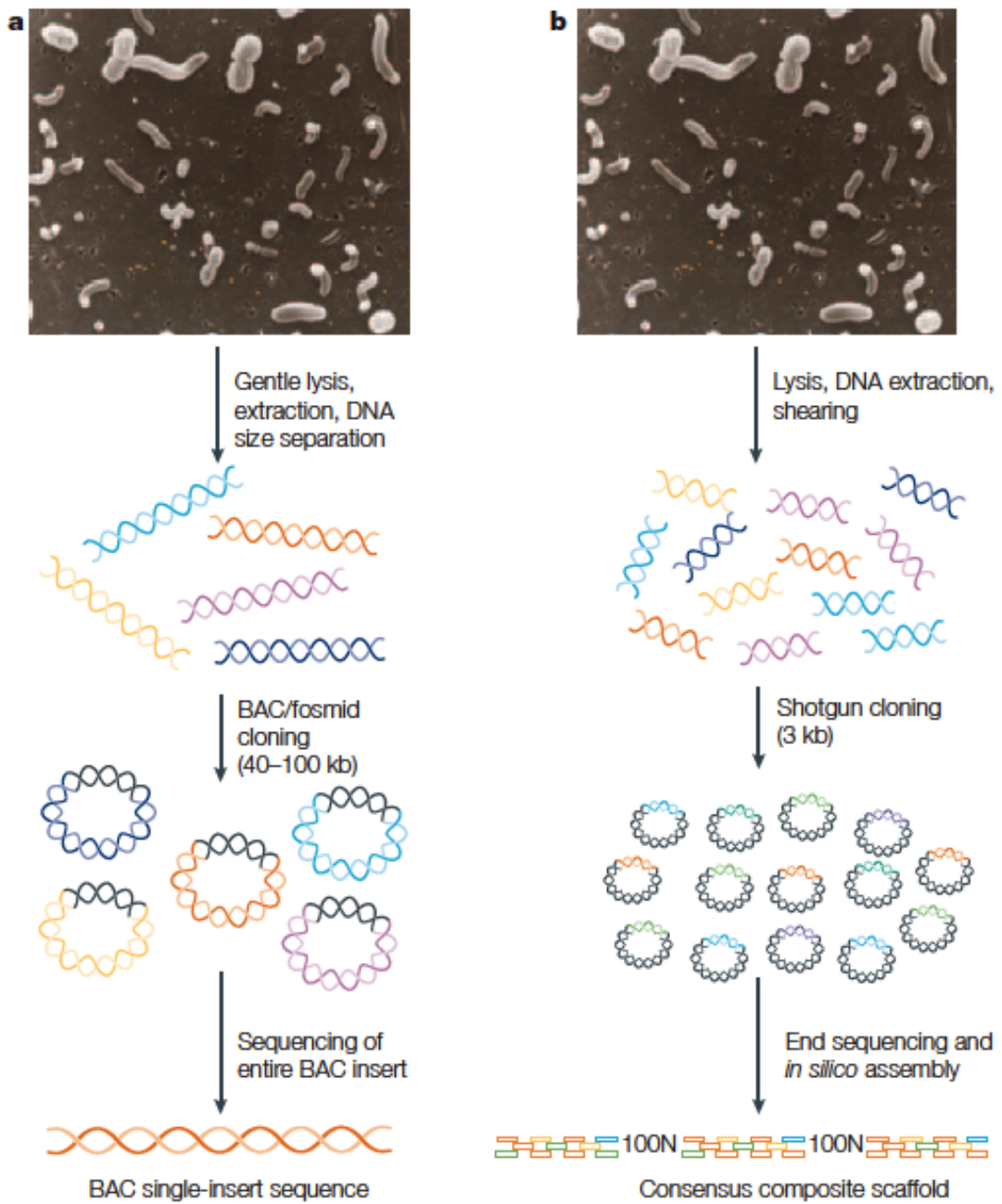


Figure 1.1: Microbial community DNA sequencing. Figure from Delong et al. [26].

Chapter 2

GRAMMy based on shotgun metagenomic reads

2.1 Background

Microbial organisms are ubiquitous dwellers of the earth's biosphere whose activities shape the earth's biogeochemistry. Through pathogenesis and symbiosis, they also play important roles in the health and metabolism of macro-organisms. For example, the human body is inhabited by trillions of microbes, affecting our digestive system, immune system, and physiology [102]. Thus, the knowledge of their presence and abundance in nature is of great relevance to ecology as well as to human well-being. To study microbes in natural environments, researchers frequently apply whole genome shotgun sequencing to uncultured samples to generate genomic sequence reads reflecting the structure of microbial communities [104, 108]. Using the sequencing data, investigators try to address basic community questions such as: who they are, how many they are, and what they do. As a consequence of the random sampling and sequencing scheme of the shotgun metagenomics approach, the presence and abundance information of metagenomes is preserved in raw reads although some studies have shown that biases in sampling can

occur, as is true for virtually all approaches [68]. However, the subsequent analysis of metagenomic data remains a challenging computational problem because of the mixed nature of metagenomes and the fact that we only sequence a small fraction of them.

Several computational methods have been developed to extract taxonomic information from metagenomic sequence reads. These existing methods can be separated into two classes: composition-based and alignment-based. In the composition-based approaches, similarity measures based on oligonucleotide composition, also known as k-mer frequencies, are used to classify metagenomic reads. For instances, TETRA, Compost-Bin, TACOA, and AbundanceBin are all reference-free methods and they cluster sequences with different binning strategies [20, 28, 98, 117]. PhyloPythia uses pre-trained composition-based classifiers to group sequences [64] and Phymm trains interpolated Markov model-based classifiers [13, 50]. However, none of these binning or classification approaches is designed to estimate the relative abundance of genomes for microbial communities (or the genome relative abundance (GRA)).

In the alignment-based approaches, alignment and mapping tools, such as BLAST, are commonly used to find similarity hits of the query reads to the references [1]. Some of them, such as Sort-ITEMS, use BLASTX for amino acid sequence similarity search [67]. However, we will only focus on similarity search based on nucleic acid sequence. The MEGAN software parses BLAST results and traces back the lowest common ancestor of ambiguously assigned reads to generate a phylogenetic distribution of the reads [47]. An intuitive way of estimating GRA based on MEGAN is using the normalized read distribution along the leaves of the phylogenetic tree, leaving out the reads assigned to multiple references. However, estimation of abundance levels by this method, which

discards reads with multiple origins, can be biased by many factors, including the variation of genome size [2, 11]. The latest Genome Abundance and Average Size (GAAS) tool weighs hits by their E-values and gives a direct estimation of genome relative abundance [2]. However, its accuracy and reliability are still hindered by the prevailing existence of read assignment ambiguities and the oversimplified estimation scheme.

In parallel with computational developments, significant improvements in sequencing technology have also been underway. Traditional metagenomic read sets are based on Sanger sequencing, which has an average read length at about 800 bp or above. At these lengths, taxonomic origin identification for the reads is relatively easy when the reference genomes are known. However, there was only limited availability of reference genomes as well as limited sequencing depth. Therefore, the relative abundance levels could not be accurately estimated, especially for complex communities in the past. Recent wide spread adoption of next generation sequencing (NGS) technologies in the metagenomics research community has led to the emergence of several massive, but short, read sets from Roche/454 (millions of 100 - 400 bp reads), Illumina/Solexa and ABI/SOLiD platforms (tens of millions of 50 - 100 bp reads) [78, 103].

The paradigm shift in sequencing technologies has impacted downstream analyses. Specifically, the identification of the origin of a read becomes more difficult for several reasons. First, a large number of short reads cannot be uniquely mapped to a specific location of one genome. Instead, they map to multiple locations of one or multiple genomes. These ambiguities are directly associated with the read length reduction in NGS technologies. Second, communities usually consist of many microbes with similar

genomes, different only in some parts, making it indeed impossible to determine the origin of a particular short read based solely on its sequence.

Despite these difficulties, NGS read sets have brought us richer abundance information of microbial communities than traditional datasets because of the significant increase in the number of reads. Along with the increase of read set size, efforts to assemble more reference genomes are ongoing [70, 76]. In addition, new experimental techniques, such as single-cell sequencing approaches are being developed to sequence reference genomes directly from environmental samples [94, 116]. Thus, in view of the constraints of current computational tools and the fast expanding sequencing capacities, we are motivated to develop a new method for accurate and reliable GRA estimation, one that can meet the challenges of short reads and the growing number of reference genomes.

In this chapter, we introduce GRAMMy, a unified Genome Relative Abundance (GRA) estimation framework using Mixture Model theory (MMy)-based modeling of shotgun metagenomic reads. Our GRAMMy framework is a reference-based method and utilizes the nucleic acid sequence similarity or composition. We first tested GRAMMy using our simulated reads as well as some synthetic communities with real reads from other studies (the FAMeS datasets) [63]. Compared to other reference-based methods, including GAAS and the abundance estimates from MEGAN, GRAMMy shows greatly improved accuracy in abundance estimations. Furthermore, with a reasonable sequencing depth, GRAMMy's estimates converge to the true abundance levels and remain stable. We then analyzed 34 real metagenomic read sets with GRAMMy, the results of which yielded interesting and new insights in biology. Finally, we packaged the GRAMMy

tools as a C++ extension to Python, which can be downloaded freely from GRAMMy's homepage (<http://meta.usc.edu/softs/grammy>).

2.2 The GRAMMy framework and algorithm

2.2.1 The GRAMMy framework

The GRAMMy framework is based on a mixture model for the short metagenomic reads and an Expectation Maximization (EM) algorithm, as outlined in the model schema and the analysis flowchart in Figures 2.1 and 2.2. GRAMMy accepts a set of shotgun reads, as well as some references (e.g. genomes, scaffolds or contigs) as inputs and subsequently performs the Maximum Likelihood Estimation (MLE) of the relative abundance levels. In the typical GRAMMy workflow, which is shown in Figure 2.2, the end user starts with the metagenomic read set and reference genome set and then chooses between mapping-based ('map') and k -mer composition-based (' k -mer') assignment options. In either option, after the assignment procedure, an intermediate matrix describing the probability that each read is assigned to one of the reference genomes is produced. This matrix, along with the read set and reference genome set, is fed forward to the EM algorithm module for estimation of the genome relative abundance levels. After the calculation, GRAMMy outputs the GRA estimates as a numerical vector, as well as the log-likelihood and standard errors for the estimates. If the taxonomy information for the input reference genomes is available, strain (genome) level GRA estimates can be combined to calculate high taxonomic level abundance, such as species and genus level estimates.

We implemented the computation-intensive core of GRAMMy in C++ with Standard Template Library (STL) for best performance and compatibility, and we integrated the typical workflow tools into a Python extension. Compared to other methods included in our study, we showed the superior accuracy and robustness of GRAMMy’s estimates, as detailed in the following sections. Other choices of read assignment schema, such as NGS mapping tools and Markov Model-based read assignment [82], can also be incorporated into GRAMMy, since they produce a reasonable read assignment probability matrix. The GRAMMy package is open source, and users are able to implement other workflow variants.

2.2.2 A finite mixture model

We developed a finite mixture model for the GRAMMy framework. Following Angly et al. we used genome relative abundance (GRA) as the relative abundance measure of mostly unicellular microbial organisms [2]. We describe the sampling and sequencing procedure as follows: First, randomly choose a genome g_j with probability π_j proportional to $a_j l_j$, where a_j is the abundance and l_j is the genome length. Second, randomly generate a read r_i from it. Without loss of generality, we further assume that for the given genome g_j we can reasonably approximate the generation of shotgun reads by some component distribution f_{g_j} such that the probability of generating a read r_i from g_j is $f_{g_j}(r_i)$. With a reasonable assumption of independence between the two sampling steps, the whole procedure is probabilistically equivalent to sampling from a mixture distribution M : $M = \sum_{j=1}^m \pi_j f_{g_j}$, with the mixing parameters denoted by $\pi = (\pi_1, \pi_2, \dots, \pi_m)$, $\sum_{j=1}^m \pi_j = 1$ and the component distributions denoted by $\mathbf{f} = (f_{g_1}, f_{g_2}, \dots, f_{g_m})$, where m is the number

of genomes. Subsequently, each read set, denoted by $\mathbf{R} = (r_1, r_2, \dots, r_n)$ can be regarded as a realized independent, identically distributed (iid) sample of size n from the mixture M . The relative abundance of known genomes is exactly a transformation of the mixing parameters π , which can be estimated based on the read set \mathbf{R} . A schematic view of the finite mixture model is shown in Figure 2.1. With the component distributions properly set up, we can find the maximum likelihood estimate (MLE) of the mixing parameters.

In many studies, our knowledge of the genomes present in the community is limited. Under these circumstances, we can define the mixture with the first $m - 1$ components for known genomes and the last m -th component for the collective of unknown genomes. Note that for the $m - 1$ known components, we suppose that their genome sequences $\mathbf{G} = (g_1, g_2, \dots, g_{m-1})$ and genome sizes $\mathbf{L} = (l_1, l_2, \dots, l_{m-1})$ are known. Therefore, the GRA for known genomes $\mathbf{a} = (a_1, a_2, \dots, a_{m-1})$ is the normalized abundance, where the relative abundance for the j -th known genome is $a_j = \frac{\# \text{ of } j\text{-th genome}}{\# \text{ known genomes}}$, where $\sum_{j=1}^{m-1} a_j = 1$. In the biological setting, we want to estimate vector \mathbf{a} , which is a measure of organism relative abundance. In the transformed mixture problem, \mathbf{a} is related to the mixing parameters π by:

$$a_j = \frac{\pi_j}{l_j \sum_{k=1}^{m-1} \frac{\pi_k}{l_k}}, \quad (2.1)$$

or the inverse:

$$\pi_j = (1 - \pi_m) \cdot \frac{a_j l_j}{\sum_{k=1}^{m-1} a_k l_k}, \quad (2.2)$$

for $j \in \{1, 2, \dots, m - 1\}$. The number of sampled reads is both proportional to the genome relative abundance and the length. Because the two factors are confounded, the missing

knowledge of the genome length l_j prohibits the estimation of a_j from the data. Since the effective genome length l_m for the unknown genomes is not available, we cannot estimate the relative abundance of the unknown component. However, the relative abundance of known genomes can still be estimated using our procedures.

2.2.3 Estimation of GRA using Expectation Maximization (EM)

To estimate the mixing parameters, we adopted the EM algorithm to calculate the maximum likelihood estimate (MLE). In the EM framework, we assume a ‘missing’ data matrix \mathbf{Z} , in which each entry z_{ij} is a random variable indicating whether r_i is from g_j or not. Then we can solve for the parameters by iteratively estimating π and \mathbf{Z} using Algorithm 2 (see derivations in Appendix A.1). Note that a variable with superscript (t) stands for its value at the t -th iteration, e.g., $p^{(t)}$ is the estimate of p at the t -th step. The EM at the t -th iteration is:

- **E-step**

Assuming $\pi^{(t)}$ known, $Z^{(t)}$ can be updated by the corresponding posterior probabilities:

$$z_{ij}^{(t)} = \frac{p(r_i | z_{ij} = 1; \mathbf{G})\pi_j^{(t)}}{\sum_{k=1}^m p(r_i | z_{ik} = 1; \mathbf{G})\pi_k^{(t)}}, \quad (2.3)$$

- **M-step**

Assuming $Z^{(t)}$ known, the new mixing parameters $\pi^{(t+1)}$ are updated by:

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^n z_{ij}^{(t)}}{n}. \quad (2.4)$$

When the MLE of π is found, using Equation 2.1, the MLE of \mathbf{a} can be calculated, thereby solving the original problem.

The space complexity of the EM algorithm is $O(c_1n)$ and the time complexity of the EM algorithm is $O(c_1c_2n)$, where c_1 the average number of associated genomes for one read and c_2 is the time cost related to the convergence criteria for EM. Since c_1 and c_2 are both constants not related to n , the algorithm is linear in space and time complexity with the read number n . Further, the concavity of the log-likelihood function can be shown and the EM algorithm is guaranteed to converge to global maximum, see Appendix A.3.

2.2.4 Read probability approximations

The probability $p(r_i|z_{ij} = 1; \mathbf{G})$ is assessed based on f_{g_j} . Ideally, it is the probability that r_i is generated when read being uniformly sampled from genome g_j . Let s_{ij} be the number of copies of read r_i in g_j . Then the probability is approximated by:

$$p(r_i|z_{ij} = 1; \mathbf{G}) \approx \frac{s_{ij}}{l_j} \quad (2.5)$$

However, due to sequencing errors and natural genetic variations, the s_{ij} 's are not readily observable. When the mapping or alignment results from BLAST, BLAT, or other mapping tools are available, the number of high quality hits of r_i on g_j can effectively be used as s_{ij} 's. To keep only these reliable and statistically significant hits, raw hits are filtered by E-value, alignment length and identity rate. We refer to the finite mixture model with the read probability from mapping and alignment results as 'map' in the remainder of the paper.

An alternative way to assess the read probabilities is by using k -mer composition. For the j -th genome, we calculate the fraction of a k -word w by $p_{wj} = \frac{\# \text{ of } w \text{ in } g_j}{l_j}$, the normalized frequency of the word w in genome g_j . For a read r_i , we define pseudo-likelihood for r_i by:

$$p(r_i | z_{ij} = 1; \mathbf{G}) = \prod_{w \in W_i} p_{wj}. \quad (2.6)$$

where W_i is the set of words formed by sliding windows of size k along. This probabilistic assignment captures the overall similarity between reads and genomes, an idea adopted in other composition-based studies such as in Sandberg et al. [86]. It is especially useful when a large number of reads do not have reliable hits with reference genomes. We will refer to the finite mixture model with the read probability from the multinomial k -mer composition as ‘ k -mer’ in the remainder of the paper.

2.2.5 Standard errors for GRA estimates

We also derived the asymptotic covariance matrix for the mixing parameters π using the asymptotic theory for MLE estimates. Because there are $m - 1$ independent parameters in π , we can choose them as $(\pi_1, \pi_2, \dots, \pi_{m-1})$ and denote by $\hat{\pi}$. Further, let $\hat{\pi}^*$ and \mathbf{a}^* be the MLE estimates for $\hat{\pi}$ and its corresponding GRA, respectively. Then, the asymptotic standard error for \mathbf{a}^* is approximately:

$$SE(a_j^*) = (\mathbf{Cov}(\mathbf{a}^*))_{jj} \approx ((\mathbf{I}_o^{-1}(\mathbf{a} | \mathbf{R}, \mathbf{G}))_{jj})^{\frac{1}{2}} \Big|_{\hat{\pi} = \hat{\pi}^*}, \quad (2.7)$$

for $j \in \{1, 2, \dots, m - 1\}$, where \mathbf{I}_o is the observed information matrix. See Appendix A.2 for details of derivation.

If only a small number (as compared to number of parameters) of reads are mapped, the conditions for the asymptotic to hold cannot be satisfied. We can alternatively use the bootstrap covariance estimator for the standard error of MLE:

$$SE(a_j^*) = (\mathbf{Cov}(\mathbf{a}^*))_{jj} \approx \left(\frac{1}{B-1} \sum_{b=1}^B (\mathbf{a}_{(b)}^* - \bar{\mathbf{a}}^*)(\mathbf{a}_{(b)}^* - \bar{\mathbf{a}}^*)^T \right)_{jj}, \quad (2.8)$$

for $j \in \{1, 2, \dots, m-1\}$, where $\bar{\mathbf{a}}^* = \frac{1}{B} \sum_{b=1}^B \mathbf{a}_{(b)}^*$ is the bootstrap mean estimator.

2.2.6 Higher level taxonomic statistics

Many downstream analyses can be carried out based on GRAMMy's estimates. For example, the average genome length \bar{l} is readily obtainable:

$$\bar{l} = \frac{1}{m-1} \sum_{j=1}^{m-1} a_j l_j \quad (2.9)$$

Subsequently, we can test the statistical significance of the median average genome length difference between two sample groups by Wilcoxon test (`wilcox.test` in R).

Since genome size bias has already been corrected, we can use GRAMMy estimates to calculate the relative abundance of a higher-level taxon by simple addition. For this purpose, we used the NCBI Taxonomy, which has the taxonomic assignments for all reference genomes we used here. To illustrate, for a specific taxonomic level h , the relative abundance of a i -th specific taxon $T_i^{(h)}$ is:

$$a_{T_i^{(h)}} = \sum_{j \in \{j: g_j \in T_i^{(h)}\}} a_j \quad (2.10)$$

and

$$SE(a_{T_i^{(h)}}) = \left(\frac{\sum_{j \in \{j: g_j \in T_i^{(h)}\}} SE(a_j)^2}{\#\{j : g_j \in T_i^{(h)}\}} \right)^{\frac{1}{2}}, \quad (2.11)$$

where h can be any one of the seven hierarchical levels in the taxonomy, from species to kingdom.

2.3 Materials and methods

2.3.1 Real read sets and reference genome sets

In preparing the real read sets, we downloaded the FAMEs data from JGI (<http://FAMEs.jgi-psf.org>), the ‘hg’ data from TraceDB (<ftp://ftp.ncbi.nih.gov/pub/TraceDB/>, NCBI project id: 16729), the ‘uhg’ data from Sequence Read Archive (<http://www.ncbi.nlm.nih.gov/Traces/sra/>, NCBI project id: 32089), the ‘jhg’ data from BGI (<http://gutmeta.genomics.org.cn/>) [51] and the ‘amd’ data from TraceDB (NCBI project id: 13696).

In preparing the reference genome sets, we downloaded currently available complete and draft bacteria genomes from the NCBI Refseq (<http://ftp.ncbi.nih.gov/refseq/>), MetaHit (<http://www.metahit.eu/>), HMCJ (<http://metagenome.jp>), WUSTL Gordon Lab (<http://genome.wustl.edu/>) and JGI (<http://genome.jgi-psf.org/>). We manually curated genomes to remove redundancy and organized them into a NCBI Taxonomy (<http://www.ncbi.nlm.nih.gov/Taxonomy>) database. We used the genome information available from IMG/M (<http://img.jgi.doe.gov>), IMG/HMP (http://www.hmpdacc-resources.org/cgi-bin/img_hmp) and GOLD (<http://www.genomesonline.org>).

org) to group them by habitats [58, 61]. Finally, we obtained 388 human gastrointestinal tract genomes for a human gut reference genome set ('HGS').

2.3.2 Read filtering and assignment procedures

In the 'map' read probability backend, we used BLAT to map reads to reference genomes. We prefer BLAT to BLAST, as BLAT is tens of times faster in handling low-sensitivity similarity search for massive number of sequences than BLAST. Since we only kept alignment results with identity rate greater than 90%, the BLAT result should not differ much from what if BLAST was used. For the human gut and simulated data, we used similar filtering methods as by Turnbaugh et al. [101, 103] (E-value ≤ 0.0001 , aligned length more than 75% of its *RL* and identity $\geq 90\%$). In the '*k*-mer' read-probability backend, we used *k*-mer length $k = 6$. For GAAS and MEGAN, we used the same mapping results from BLAT, as a common starting point. We used GAAS's default filtering options (E-value ≤ 0.0001 , aligned length more than 80% of its *RL*, and identity $\geq 80\%$), as well as MEGAN's default options (min-score=35 for *RL* equal to 100 bp and min-score=50 for *RL* equal to 400 bp; top percent=5%, min support=2), for comparisons.

In evaluating the ribotype and protein marker based method, we used the E.coli 16S rRNA *rrsE* and ribosome protein *rpoB* genes to retrieve homolog sequences from the simulated reads, which were then filtered by options (E-value ≤ 0.0001 , aligned length more than 75% of its *RL* and identity $\geq 90\%$), according to Biers et al. [12]. Our validations have shown that variations of these parameters within a reasonable range had little effect on the results.

2.3.3 Numerical error measures

We use the following measures to evaluate the accuracy of the GRA estimate. Let the true GRA be \mathbf{t} and its estimate \mathbf{a} . The first measure $\text{RRMSE} = \sqrt{\frac{1}{m-1} \sum_{j=1}^{m-1} \left(\frac{|a_j - t_j|}{t_j} \right)^2}$ is the commonly used root mean square version of relative error [32]. We also included three other error measures: $\text{AVGRE} = \frac{1}{m-1} \sum_{j=1}^{m-1} \frac{|a_j - t_j|}{t_j}$ (the average relative error), $\text{MAXRE} = \max_j \left(\frac{|a_j - t_j|}{t_j} \right)$ (the maximum relative error), and $\text{DTV} = \frac{1}{2} \sum_{j=1}^{m-1} |a_j - t_j|$ (the Total Variation Distance [60]), which are all commonly used to evaluate the accuracy of an estimate.

2.3.4 Hierarchical biclustering

It is possible to use GRAMMy estimates for clustering analysis and statistical hypothesis testing. We clustered the samples based on the pairwise similarities (correlations) of their relative abundance distribution. Because of the long-tailed shape of the distribution, the signal-to-noise ratio is low for these less abundant genomes. Therefore, using the thresholds .05% for the minimum abundance and 50% for the minimum occurrence [78], we selected the estimates for these more abundant genomes (which are more reliable for clustering). We used rank transformation, which normalizes GRAs by taking their ranks and applying score transformation and R function `heatmap` for hierarchical clustering.

2.4 Simulation studies

2.4.1 Simulated read benchmarks

We first tested GRAMMy by using a series of simulated read sets. By using read sets generated from a collection of genomes included in the FAMeS study [63], we were able to assign the true relative abundance levels and confirm the estimation accuracies by analyzing the errors between the estimates and true values. The numerical error measure RRMSE (Relative Root Mean Square Error), which computes the root mean square average of relative errors, was used to assess the accuracy and robustness of estimates. The detailed discussion of the simulation studies is provided in the Appendix B and the results are presented in Figures B.1-B.4.

Figure B.1 shows that all the error measures decrease to zero as the number of reads increases. Figure B.2A shows that effect of sequencing errors on the GRA estimation accuracy and it shows that sequence errors have a significant effect on short reads ($\leq 100\text{bp}$) while the effect is minimal for long reads. Figure B.2B shows that missing reference genomes in the reference genome set does not significantly affect the estimation accuracy for the genomes in the reference data set even if 50% of the genomes in the community are unknown. Figure B.2C shows the effect of different abundance distribution on the estimation accuracy and it shows that such an effect is not significant although we do see a slight increase in the measurement errors for communities with uneven abundance distributions compared to that for the even abundance distributions. In summary, our simulations show that the GRAMMy estimates are accurate and stable across a range of anticipated scenarios.

Interestingly, a relatively small number of short reads is sufficient to obtain an accurate estimate of relative abundance of the genomes, thus eliminating the need for an excessively ‘deep sequencing’ scheme in certain richness assessing scenarios. As shown by all panels in Figure B.1, RRMSEs start to stabilize when the number of reads (RN) surpassed 10^5 , indicating the existence of a threshold for the number of reads needed to recover the community abundance structure. The trend also shows that a relatively small number of read sets could still provide substantial information for the abundance estimation, when the read assignment ambiguity is properly handled. However, the number of required reads depends on the number of organisms in the community and the distribution of relative abundances of the different organisms.

We also compared GRAMMy to other methods. With the objective of estimating the GRA of communities, we first benchmarked GRAMMy with GAAS. In addition, we included MEGAN, which produces a read profile that summarizes the number of reads assigned to their lowest common ancestors (LCA). We estimated the GRA based on MEGAN using the normalized percentages from the reads distributed on leaf taxon. In the benchmark, we used a series of simulated read sets generated from genomes randomly selected from the FAMeS study (see details in Appendix B). The same genomes used in read generation were also used as our reference genomes.

We then used BLAT to align the reads to the reference genomes and fed the output into GRAMMy, GAAS and MEGAN. The default options of GAAS and MEGAN were used in our study. Figure B.3A shows the results from the simulation read sets with read lengths (RLs) equal to 100 or 400 bp generated from MetaSim [80] using the ‘with sequencing errors’ option. We see that GRAMMy (‘map’) significantly outperformed

GAAS, MEGAN and GRAMMy (k -mer') in all settings. Among all the methods tested, GRAMMy ('map') is the only method with RRMSEs decreasing to zero as the number of reads increases.

To account for the poor performance of other methods, we can point to several possible reasons. For GAAS, assigning ambiguous hits based on their E-value weights is ad hoc and may reduce its accuracy because the E-value is only a statistical measure for the quality of the alignment. For MEGAN, its arbitrary cutoff at the top five percent hits and its non-probabilistic handling of ambiguous hits may reduce the accuracy of GRA estimates. In addition, for both MEGAN and GAAS, there is also the possibility of losing accuracy when changing from BLAST hits to BLAT hits. While it has been argued that BLAST alignment is the best way to assign reads [13], it is too computationally intensive for BLAST-aligning every read to references [78]. Instead, fast mapping tools like BLAT only keep a small number of high-similarity hits, while, at the same time, possibly reducing the accuracy of both GAAS and MEGAN. In contrast, the superior performance of GRAMMy ('map') shows that the probabilistic way of handling ambiguous hits could help to improve the estimation, which also gives GRAMMy an advantage over other methods when encountering incoming short read sets of very large sizes.

In conclusion, when the reference set is available, the GRAMMy framework based on mapping or alignment gave the best result for GRA estimation. Thus, the 'map' approach is generally the method of choice in most application settings. Only when assembled reference genomes are absent, GRAMMy (k -mer') is needed as a still viable solution for GRA estimation, since at RL equal to 400 bp its performance is comparable to the estimates from GAAS and MEGAN. However, the k -mer composition approach

has limited power to distinguish the different genomes, as the compositions of k -mers are usually heterogeneous across the genomes. In addition, there is no genome size bias correction if ‘ k -mer’ method is used without prior knowledge of genome lengths.

In addition to the above methods, relative abundance estimation based on ribotype (retrieving rRNA sequences and classifying into taxonomic bins, e.g. 16S rRNA), protein marker (similar to ribotype method except replacing rRNA by protein marker, e.g. *rpoB*) and hit counting (counting the total number of all hits in each taxonomic bin) has been used to estimate relative abundance [12, 18, 85]. We compared the 16S-based (adapted from Biers et al. [12]), *rpoB*-based and BLAT hit counting estimates to GRAMMy estimates using our simulated read set. Figure B.3B shows that GRAMMy outperformed all other methods in this controlled setting. All other methods show three obvious drawbacks: a persisting bias, significant variation and a strong dependence on the number of reads.

In fact, 16S rRNA and *rpoB* genes are only very small parts of genomes; therefore, even if the total number of reads is large, the reads covering these genes are barely about 1/1000 of all reads. If the total number of reads is small and there are not enough reads covering 16S rRNA genes, then the method is not viable as a result of its substantial instability. Even if the total number of reads increases, due to gene copy number and genome size variations, the estimates still do not converge to the true abundance values. Similar trends were also observed when BLAT mapping hit counts were directly normalized and used for abundance estimation. On the contrary, GRAMMy always produced much more accurate and reliable estimates.

For the estimates at different taxonomic levels, the estimation errors gradually decrease from the strain level to the kingdom level and are mostly small given a relatively large number of reads (see Figure B.4).

2.4.2 Artificial metagenomes with real reads

We further compared the estimates of GRAMMy with those of GAAS and MEGAN, using the third party FAMeS dataset [63]. The FAMeS data are comprised of three synthetic metagenomic read sets constructed by random sampling from real whole genome shotgun sequencing reads. These constructed read sets are labeled ‘simLC’, ‘simMC’ and ‘simHC’, according to different complexities of the communities. Each set is composed of approximately ten thousand Sanger reads from 113 microbial genomes. These artificially created metagenomes have considerably different abundance distributions, ranging from uniform-like in the ‘simLC’ set to steep power-law-like in the ‘simHC’ set, with the ‘simMC’ set in between. We ran GRAMMy (‘map’), MEGAN and GAAS on all three data sets.

The results, which are summarized in Table 2.1, show that the measured Relative Root Mean Square Error (RRMSE) and Average Relative Error (AVGRE) for GRAMMy (‘map’) are approximately 10-20%, while those for MEGAN-based estimates are approximately 40-50%, and those for GAAS are even larger. The benchmark further substantiates that GRAMMy (‘map’) yields the most accurate estimates for all these sets. Although the errors are not close to zero, the results are still respectable, considering that the overall sequencing depth is low in all these sets, which is, on average, less than a hundred reads per genome. The highest accuracies reachable are certainly affected by the limited

number of reads and the presence of sequencing errors in these read sets. Nonetheless, recent real read sets are frequently two to three orders of magnitude larger than the FAMeS data, making accurate GRA estimation more feasible.

	simLC		simMC		simHC	
	RRMSE	AVGRE	RRMSE	AVGRE	RRMSE	AVGRE
GRAMMy	20.0%	14.0%	25.6%	19.7%	21.6%	14.7%
MEGAN	48.6%	39.3%	50.0%	40.6%	50.2%	40.8%
GAAS	433.8%	152.5%	171.4%	111.6%	507.9%	165.8%

Table 2.1: Comparison of estimation accuracy. A summary of Relative Root Mean Square Error (RRMSE) and Average Relative Error (AVGRE) measured from MEGAN-based, GAAS and GRAMMy (‘map’) estimates of simLC, simMC and simHC subsets of the FAMeS data. GRAMMy (‘map’) has the lowest error rate for both error measures across all the subsets.

2.5 Application to real datasets

2.5.1 Meta-analysis of human gut metagenomes

The human gastrointestinal tract harbors the largest group of human symbiotic microbes. Several shotgun metagenomics studies on these communities have been published. With more than six hundred human-related bacteria reference genomes publicly available, we are well positioned to use these datasets to illustrate the practical uses of GRAMMy. We collected ‘gut’ data from three major human gut metagenome projects including two U.S. human distal guts (800 bp Sanger reads, about reads per sample, labeled ‘hg’), 18 U.S. adult samples from twin families (250 bp 454 reads, reads per sample, obese and lean, labeled ‘uhg’), and 13 Japanese gut samples (800 bp Sanger reads, reads per sample, weaned or unweaned infants and adults, labeled ‘jhg’) [39, 51, 101].

For the reference set for the 33 human gut samples, we used a comprehensive collection of human gut microbes (labeled ‘HGS’), containing 388 currently available human gastrointestinal microbial genomes from multiple sources (see Table S1A in Xia et al. [121]). BLAT was used to assign metagenomic reads to the ‘HGS’ set according to their alignment similarities, and the overall study was labeled using the combination of the read set name, the reference genome set name, and the cut-off identity rate, such as ‘hg_HGS_90’, ‘uhg_HGS_90’, ‘jhg_HGS_90’. The results with cut-off at ‘90 percent’ identity rate are summarized in Table 2.2 and that from both ‘75’ and ‘90’ are provided in Tables S2-3.

Table 2.2 gives the mapped rates and ambiguity rates for each data set. The mapped rate is the proportion of reads mapped at least once to the reference genomes. It can be seen from the table that 45-60%, in median, of human gut metagenomic reads were mapped to the references for all these studies. This value suggests that the reference genome set provides a good homolog resource for the human gut metagenomic reads, even though there are still several sets only showing less than 40% mapped rate.

Another quantity, ambiguity rate, is the proportion of reads that are mapped at least twice to the references. As we can see, about 21-65% of the reads were ambiguously mapped to the reference genome set across the human gut samples. While ‘uhg_HGS’ is a collection of 454 short reads, we also noticed that it has a comparable median ambiguity rate to the other two Sanger read sets. This indicates that at 250 bp, a 454 read is already as specific as a Sanger read. However, because of the ambiguities arising from the intrinsic composition of the communities, we still encountered a significant portion of reads having multiple hits regardless of their read lengths.

We estimated the relative abundances of reference genomes for these datasets and the results are summarized within Table S4 in Xia et al. [121]. Based on these estimates, we calculated the average genome lengths for these metagenomes. The medians of genome lengths range from 2.8 Mbp to 3.7 Mbp, as shown in Table 2.2 and Table S3 in Xia et al. [121]. These statistics show that the average genome lengths for the three human gut datasets are comparable. Indeed, there is no statistically significant difference in the medians of average genome length between ‘jhg’ and ‘ugh’ samples (*Wilcoxon’s* test, two-sided, $P=0.3539$). The test involving ‘hg’ set is not suitable since it only contains two samples.

Data Set(#Sets)	Mapped Rate(%)			Ambiguity Rate(%)			Average Genome Length(bp)		
	Med.	Min.	Max.	Med.	Min.	Max.	Med.	Min.	Max.
hg_HGS(2)	46.65	43.15	50.15	31.65	30.32	32.98	2890092	2660792	3119393
jhg_HGS(13)	59.61	35.99	76.92	45.11	22.53	65.71	3745629	2268438	5657331
uhg_HGS(18)	52.35	37.49	72.51	35.90	21.56	59.81	3619072	3047940	4752910
amd_AMD(1)	45.64	45.64	45.64	1.48	1.48	1.48	2163584	2163584	2163584

Table 2.2: Summary statistics for the metagenomic datasets. Median (Med.), minimum (Min.) and maximum (Max.) of mapped rate, ambiguity rate and estimated average genome length for the samples: two from U.S. adult human gut (‘hg’), 13 from Japanese human gut (‘jhg’), 18 from U.S. twin families human gut (‘uhg’) and 1 from acid mine drainage (‘amd’) are shown. Two reference genome sets, ‘HGS’, ‘AMD’, were used for human gut samples (‘hg’, ‘jhg’, ‘uhg’) and the acid mine drainage sample (‘amd’), respectively.

Next, we identified the most frequent species across all the metagenomes. In Figure 2.3, we show the 99 species with at least 0.05% of relative abundance in at least 50% of the metagenome samples in the order of their median relative abundance. Among the top ten most common species, there are eight from the *Firmicutes* phylum including members of *Faecalibacterium*, *Eubacterium* and *Ruminococcus* genera, and two from the *Bacteroides* genus of *Bacteroidetes* phylum. It shows the predominance of *Firmicutes*

and *Bacteroidetes* in the human gastrointestinal tract. In general, these frequent species display a long-tail distribution in relative abundance levels, meaning that most species are detected across many samples, though they are not highly abundant.

We also found that the abundance levels of some species are highly variable, while most others are relatively constant (see the quantile boxes and outliers in Figure 2.3). In choosing the minimum occurrence rate and minimum abundance threshold for a typical human gut read set (10^5 reads, 800 bp), the 0.05% of relative abundance roughly corresponded to a sequencing size of 40 Kbp from the genome. This size was 25-fold more than the size coverage per genome using 16S rRNA sequencing according to Qin et al. [78]. We used a different identity rate cut-off (75%) for parsing BLAT hits and similar frequent species results were obtained. They are shown in Figure 2.7.

We compared our results to the 75 non-redundant, frequent species identified in a recent study [78]. Although we used different datasets and methodologies, our study shows comparable results. For example, between the two identified sets, five of the top ten common species are shared and so are eleven of the top twenty. The criteria they used ($\geq 1\%$ genome coverage and $\geq 50\%$ presence), if converted, roughly correspond to 0.05% in minimum relative abundance levels in our study.

However, we had some improvements over their results. They used a smaller (195) reference genome set and did not consider the genome size bias and the ambiguous hits. Consequently, their result might have missed some of the top frequent species and misplaced some species into the top rankings. In fact, the *Bacteroides* species, with genome lengths ranging from 5 Mbp to 8 Mbp, well above the median average genome lengths of human gut samples, are constantly ranked higher in their ranking. In our results,

however, this bias is corrected, and the rankings are accordingly lowered, with some of their top 20 ranked *Bacteroides* species dropping out of the top 40.

Next, we used the GRA estimates for frequent species as the basis for hierarchically clustering all the human gut samples, as shown in Figure 2.4. It can be seen that most of the frequent species belong to *Firmicutes*, *Bacteroides* and *Actinobacteria* (see column color-coding). We also see that the unweaned infants (≤ 6 months) are all grouped closely together (see row color-coding), possibly indicating their distinct gut microbial communities in comparison to that for the weaned infant and adult samples. This phenomenon was noticed in the original paper [51], and our results further strengthened their claim by incorporating data from more human gut metagenomics studies. A close look at the top 20 most abundant strains revealed that the unweaned infants' community profiles were dominated by only a few strains from *Actinobacteria*. The lack of diversity of infant gastrointestinal tract has also been reported in other studies, for example, see Vaishampayan et al. [105]. The pattern might be related to the microbial colonization process of infant gastrointestinal tract; however, no clear explanation for this interesting phenomenon is available to date.

On the other hand, there is no clear-cut evidence showing that samples from the same dataset or Body Mass Index (BMI) category are grouped together, even though there is such a trend. Note that the clustering results depend on the criterion of identifying frequent species. These species were chosen as a trade-off between the number of frequent species required for resolution power and the number that would risk including too many unreliable estimates from less abundant species. The parameters we had chosen were based on Qin et al. [78]. We did the same analysis with a different identity rate cut-off

(75%) for BLAT hits and two different minimum relative abundance thresholds (0.01% and 0.1%) for frequent species selection. Similar results were obtained. They are shown in Figure 2.8.

2.5.2 The acid mine drainage data set

In samples from other environments where reference genomes are not well characterized, such as soil, ocean and some extreme environments, assemblages like contigs and draft genomes from the sample itself can be used in addition to available known genomes. Acid mine drainage sites are extreme environments where only a few species of specially adapted microbes can survive. We downloaded the raw read set (labeled ‘amd’), which contains 103,462 Sanger shotgun reads (~ 750 bp) from one environmental sample of a biofilm [3]. The genome sequences of coexisting species were partially assembled using the metagenomic reads, among which are two dominant ones: *Ferroplasma sp. Type II* and *Leptospirillum sp. Group II 5-way CG*. The genome assemblages are in the draft state, but we roughly know their genome sizes [104]. To study the community structure, we constructed an acid mine drainage reference genome set (‘AMD’) using the two draft genomes and other currently available bacterial genomes of acid mine habitats (Table S1B in Xia et al. [121]). We mapped the read set ‘amd’ to this reference genome set and subsequently labeled the result ‘amd_AMD’.

Out of the reads mapped to the references, only a slight portion of them ($\sim 2\%$) had multiple hits (Table 2.2). We then estimated the GRA for the acid mine drainage community using GRAMMy. Figure 2.5 shows the relative abundance of the six strains we included in the ‘AMD’ reference. It confirms that the community is dominated by

the two draft genomes (98% in total relative abundance) with only marginal fraction of the other acid mine strains. The dominance of the two strains is consistent with the results from the genomic study in the original work, even though their fluorescence in-situ hybridization (FISH) result only reveals the dominance of *Leptospirillum sp. Group II* species [104].

2.6 Discussion

We have developed the GRAMMy framework for estimating genome relative abundance with shotgun metagenomic reads. It has three unique features. First, it is unique in providing a rigorous probabilistic framework for estimating Genome Relative Abundance (GRA). The estimation can be easily extended to higher taxonomic levels by simply adding up the relative abundance of genomes affiliated with the specific higher-level taxon while maintaining the accuracy, since the estimated GRA is already properly normalized and corrected for genome size bias.

Second, GRAMMy provides users with a wide choice of mapping and alignment tools. Its ability to use the results from linear time NGS mapping tools helps to reduce the computation burden for analyzing current massive metagenomic read sets. The GRAMMy program currently supports tabular BLAST formats, however, the mapping results from other popular mapping tools, such as MAQ, Bowtie and PerM [22, 53, 55], can be easily adapted to the GRAMMy framework. The algorithm is also linear in time and space with the input data size and the current implementation is much faster than MEGAN and GAAS in handling large read sets, processing one million of reads in seconds (see Figure

2.6, the BLAT mapping time is excluded for all compared tools). In addition, GRAMMy is memory efficient and we have not encountered problems in processing read number in the order of 10^6 with hundreds of microbial genomes with our 12GB nodes. However, if memory bottleneck is reached, we can always divide the reads into sub-samples and use GRAMMy in a bootstrap fashion, because a certain number of reads can already provide substantial amount of abundance information as indicated by our simulations.

Third, the method is especially suitable for short read datasets due to its better handling of read assignment ambiguities. In typical cases of a short read set, there are 10% to 40% of reads having assignment ambiguities [47]. The source of assignment ambiguity can be sequencing errors, genetic variations, horizontal gene transfers or closely related genomes. By taking into account the information from the ambiguously assigned part of the read set, our study showed that we can improve the genome abundance estimation for metagenomic data.

In applying the GRAMMy framework to the real metagenomic datasets, we used two different identity rate cut-offs: 75% and 90%. While the results from 90% were shown, we also kept the 75% results in the supplementary files. Lowering the thresholds will certainly increase the mapped rate as well as the ambiguity rate, as shown in Table S2. However, in our analysis of human gut metagenomes, the average genome size estimates and abundance estimates were not significantly changed by using different cut-offs, as shown in Tables S3 and S4 in Xia et al. [121]. Still, in other applications, researchers have to trade off between ambiguity rate and mapped rate to obtain reasonable GRA estimates for their data.

There is also the practical question of how many genomes to be included as reference. This, however, is always the choice of users. As long as the read-to-genome associations found by mapping tools are reliable and the coverage rate is high (as in our simulations), GRAMMy can reliably estimate low abundance levels and the concern of over-fitting can be alleviated. In real data, the estimation accuracy of the GRA of the low-abundance genomes depend on the number of reads mapped to each genome and the reliability of the mappings. The estimated variance of the GRAs can give some ideas about the accuracy of the estimates.

In summary, with the experimental side of shotgun metagenomics accelerating its pace, the GRAMMy method we proposed has the potential to produce more accurate taxonomic abundance estimations for downstream computational analyses.

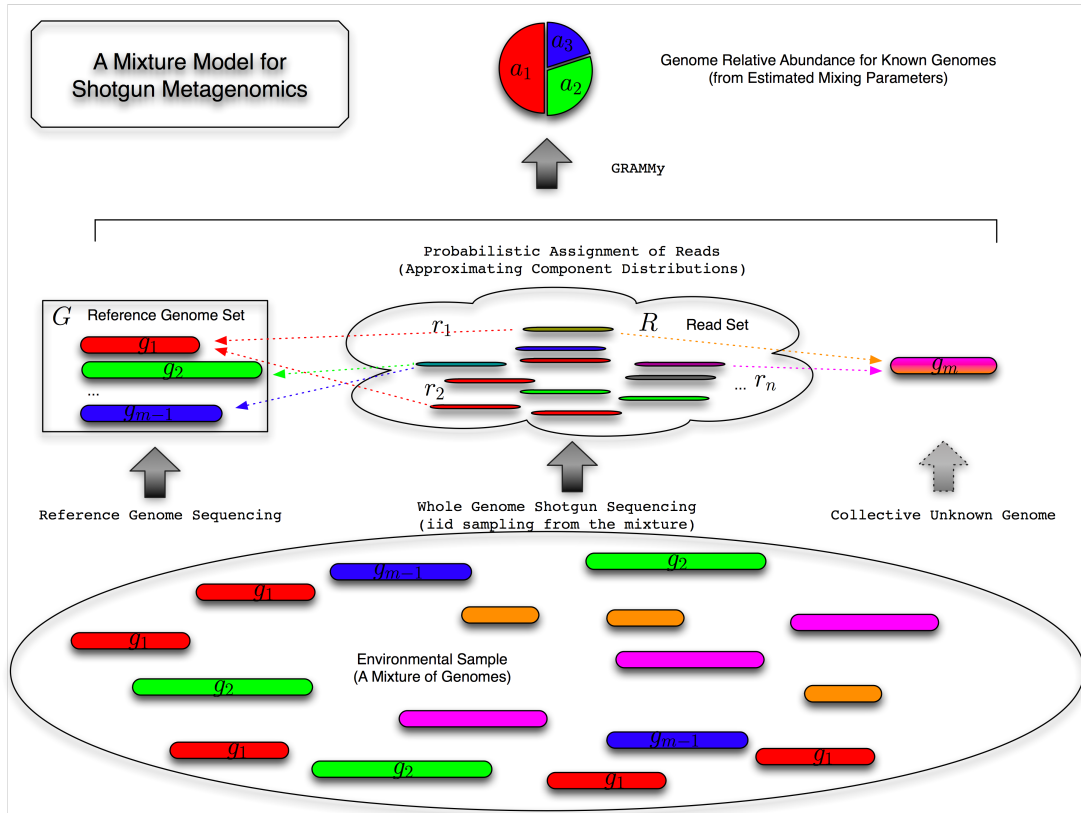


Figure 2.1: The GRAMMy model. A schematic diagram of the finite mixture model underlies the GRAMMy framework for shotgun metagenomics. In the figure, ‘iid’ stands for “independent identically distributed”.

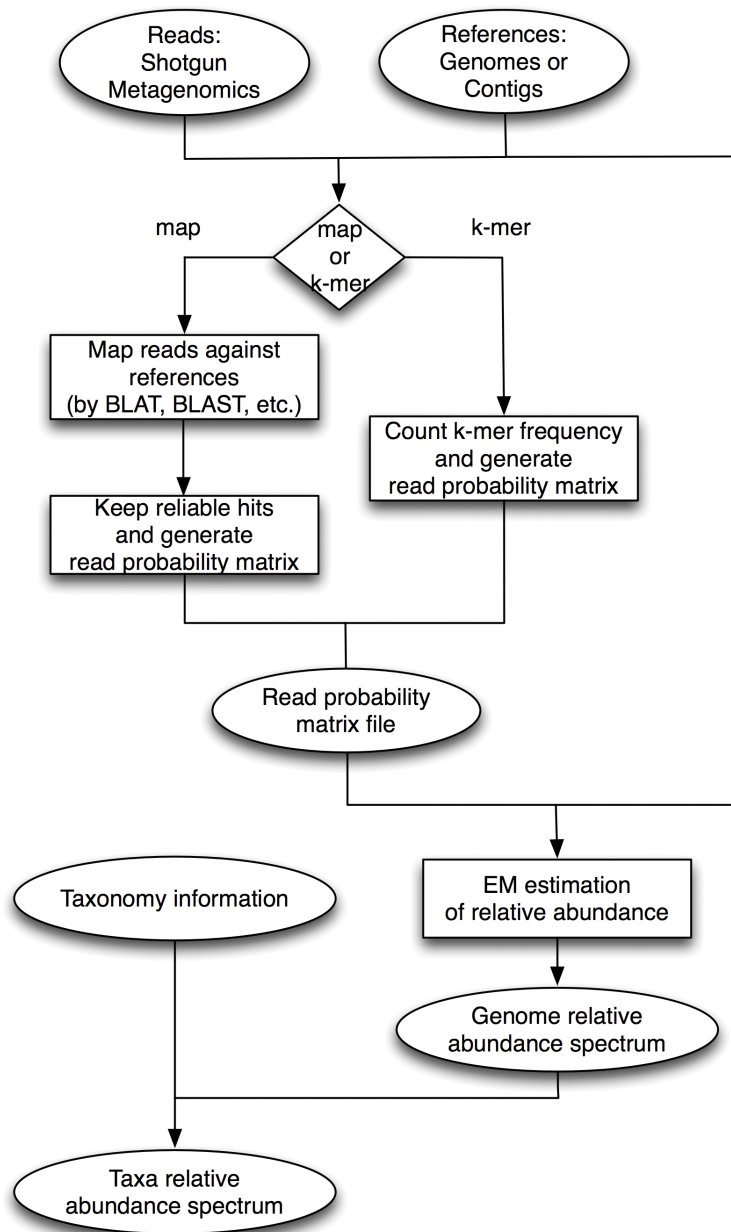


Figure 2.2: The GRAMMy flowchart. A typical flowchart of GRAMMy analysis pipeline employs ‘map’ and ‘*k*-mer’ assignment.

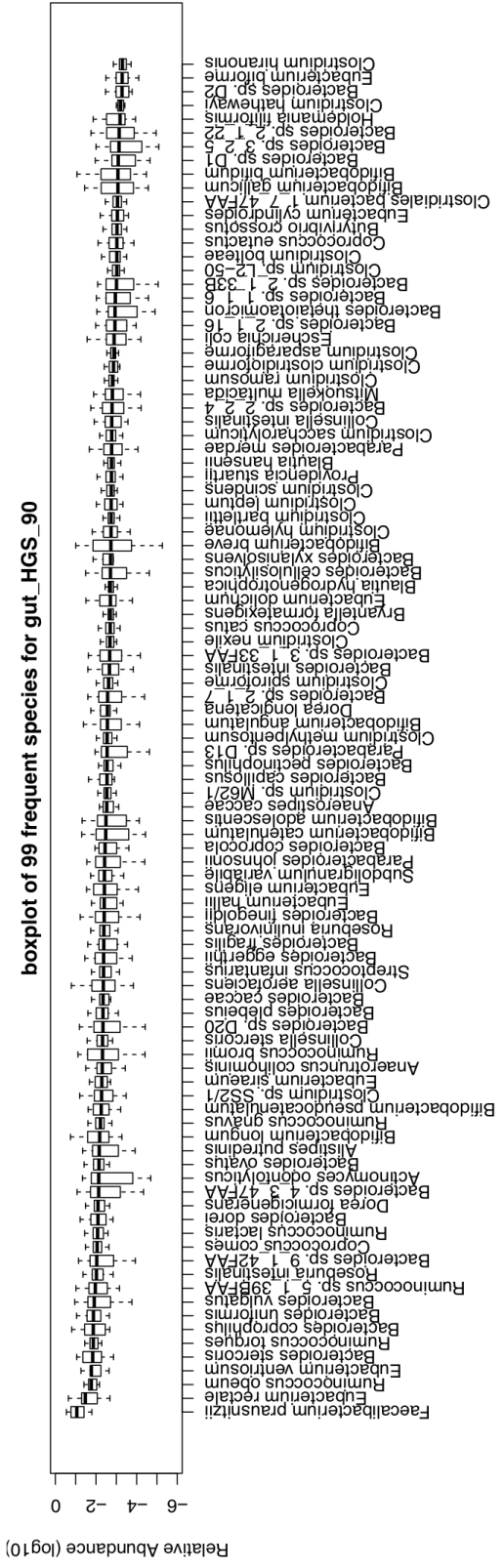


Figure 2.3: Frequent species for human gut metagenomes. The 99 species occurring in at least 50% of the 33 human gut samples with a minimum relative abundance of 0.05% were selected. 'gut_HGS_90' indicates that the human gut ('gut') read sets were mapped to the reference genome set ('HGS') with an identity rate cut-off at 90% ('90').

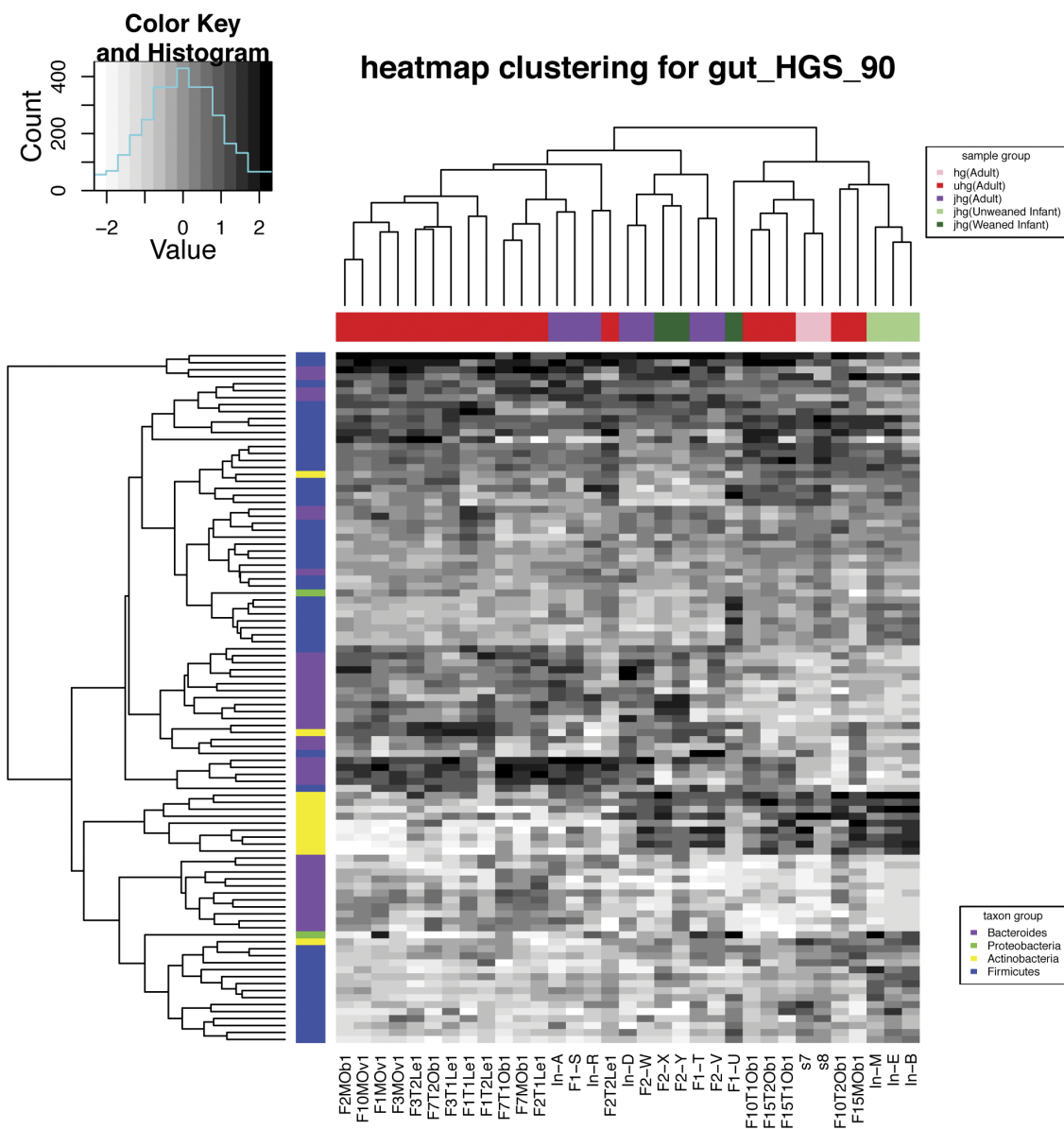


Figure 2.4: Heatmap biclustering of human gut metagenomes. ‘gut_HGS_90’ indicates that the human gut (‘gut’) read sets were mapped to the reference genome set (‘HGS’) with a identity rate cut-off at 90% (‘90’). The bottom labels indicate human gut samples. The top right legend shows the color coding for columns indicating the sample age category and dataset origin. The bottom right legend shows color coding for rows indicating the top 4 most abundant phyla in human gut. The relative abundance for each sample is normalized by a rank transformation.

GRAMMy estimation for amd (amd_AMD_90), top 20, against AMD group, strain level, from sample 5waycg

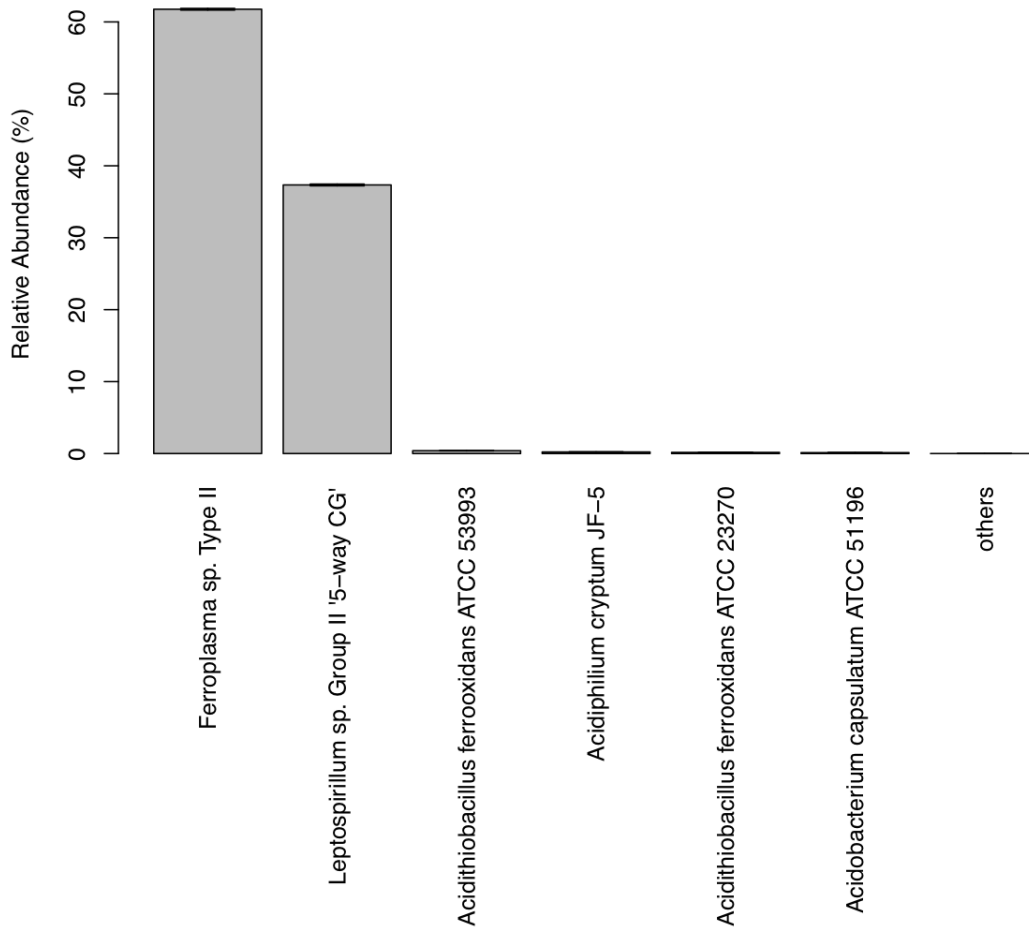


Figure 2.5: GRAMMy estimates of GRAs for the acid mine drainage data. Estimated relative abundance for each strain is shown as a percentage. The first two strains dominate the sample.

Running time comparison

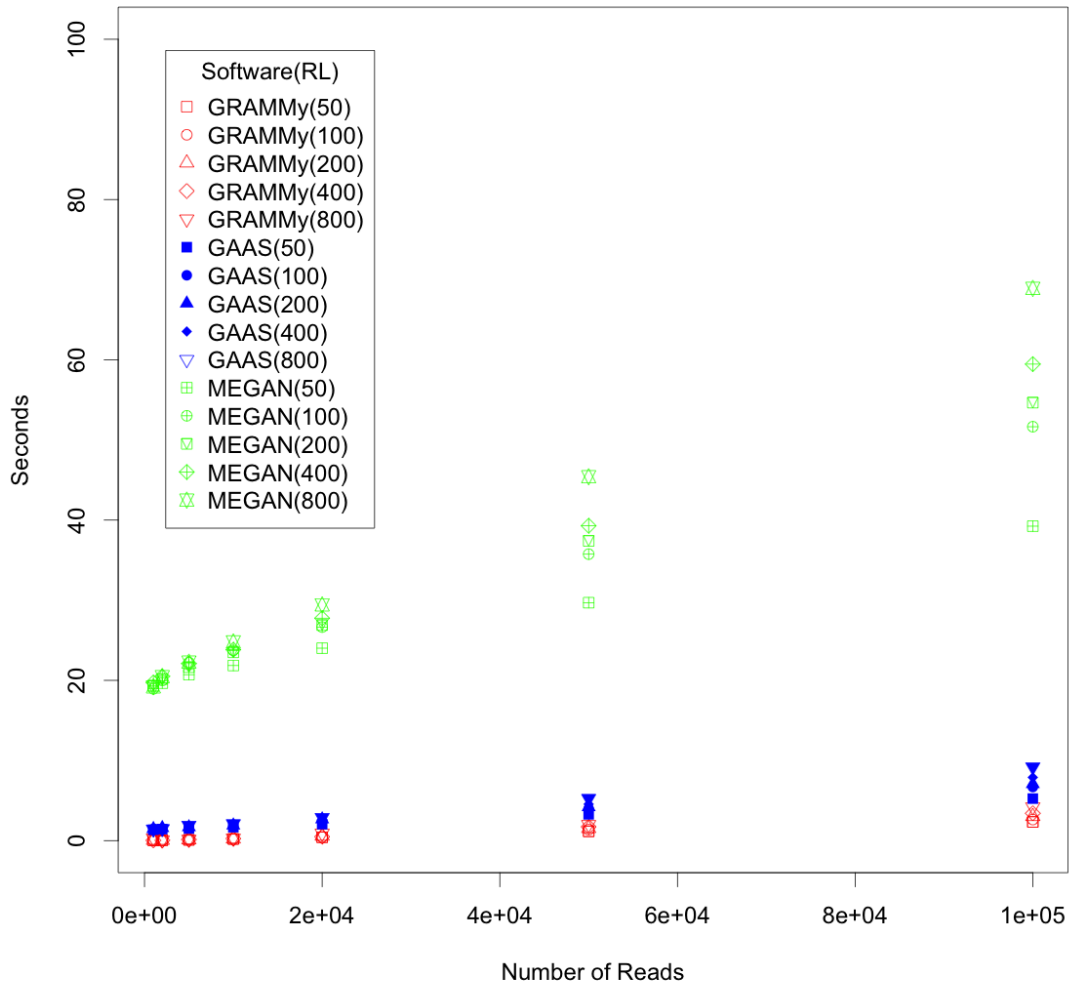


Figure 2.6: Running time comparison. GRAMMy is the fastest in all cases as compared to MEGAN and GAAS in processing time. The BLAT mapping time is excluded for all compared tools.

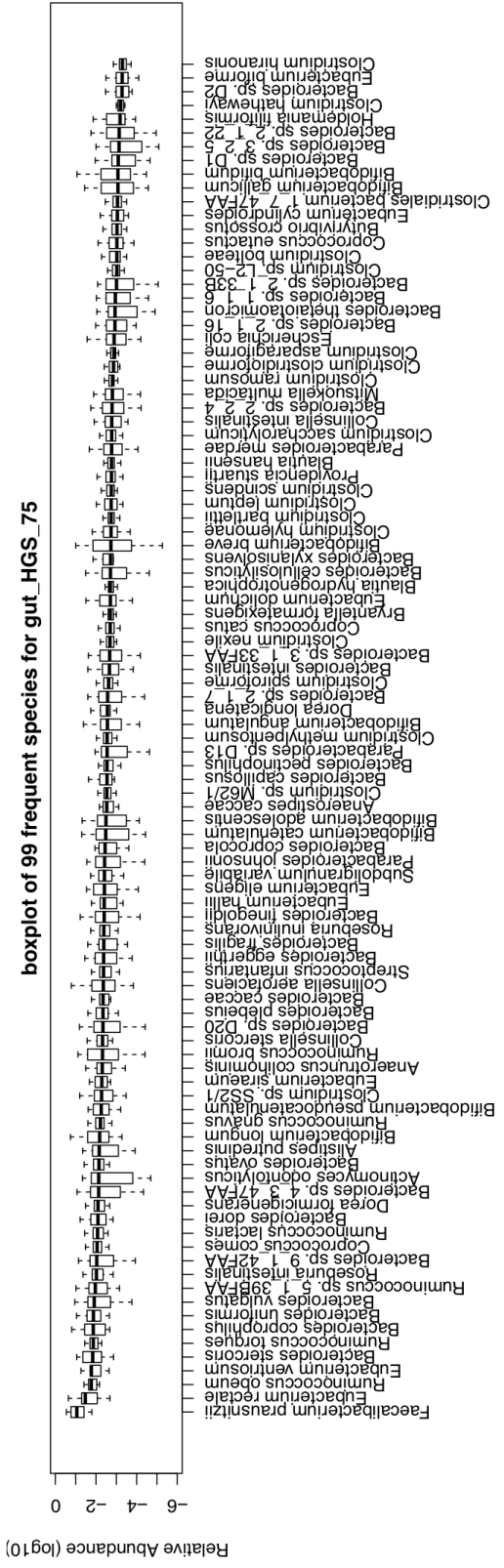


Figure 2.7: Frequent species for the human gut metagenomes. The 99 species occurring in at least 50% of the 33 human gut samples with a minimum relative abundance of 0.05% were selected. 'gut_HGS_75' indicates that the human gut ('gut') read sets were mapped to the reference genome set ('HGS') with a identity rate cut-off at 75% ('75').

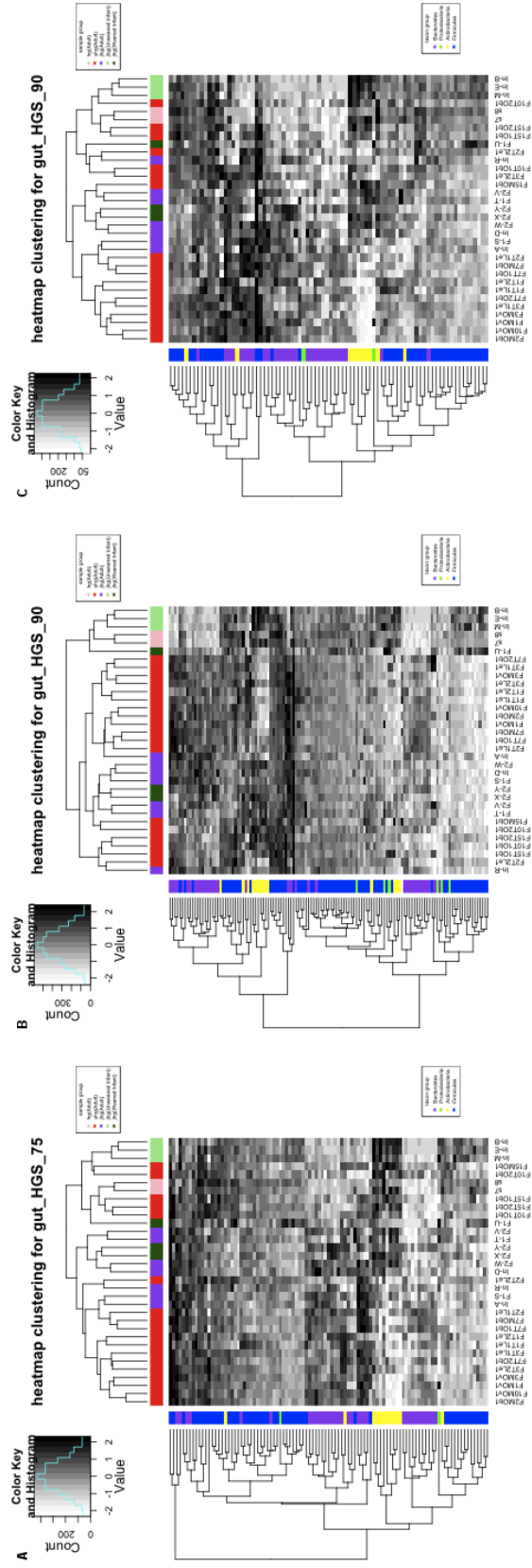


Figure 2.8: Heatmap biclustering of the human gut metagenomes. ‘gut_HGS_90’ indicates that the human gut (‘gut’) read sets were mapped to the reference genome set (‘HGS’) with a identity rate cut-off at 90% (‘90’), while ‘gut_HGS_75’ indicates cut-off at 75% (‘75’). The bottom labels indicate human gut samples. The top right legend shows the color coding for columns indicating the sample age category and dataset origin. The bottom right legend shows color coding for rows indicating the top 4 most abundant phyla in human gut. (A) Heatmap clustering of the ‘gut’ samples, with strains of abundance $\geq 0.05\%$ in at least 50% of samples selected at 75% identity rate cut-off. (B) Heatmap clustering of the ‘gut’ samples, with strains of abundance $\geq 0.01\%$ in at least 50% of samples selected at 90% identity rate cut-off. (C) Heatmap clustering of the ‘gut’ samples, with strains of abundance $\geq 0.1\%$ in at least 50% of samples selected at 90% identity rate cut-off.

Chapter 3

eLSA of molecular time series data

3.1 Background

In recent years, advances in microbial molecular technologies, such as next generation sequencing and molecular profiling, have enabled researchers to spatially and temporally characterize natural microbial communities without laboratory cultivation [36]. However, to reveal existing symbiotic relationships and microbe-environment interactions, it is necessary to mine and analyze temporal and spatial co-occurrence association patterns of organisms within these new datasets [19, 92]. Time series data, in particular, are receiving increased attention, since not only ordinary associations, but also other local and potentially time-delayed associations can be inferred from these datasets. Here local association indicates that the association only occurs in a subinterval of the time of interest, and time-delayed association indicates that there is a time lag for the response of one organism to the change in another organism. The rapid accrual of time series data is not limited to the microbial ecology field. Progress in high-throughput low-cost experimental technologies has also brought such changes to gene transcription and translation

studies. Thus, while the subjects may vary, the association network we build from local and potentially time-delayed association patterns will likely pave the way to a better understanding of these systems.

To analyze microbial community and other data under various conditions, researchers typically use techniques such as Pearson’s Correlation Coefficient (PCC), principal component analysis (PCA), multi-dimensional scaling (MDS), discriminant function analysis (DFA) and canonical correlation analysis (CCA) [34, 93, 106, 123, 124]. Although these analytic methods yield interesting patterns, they generally analyze the data throughout the whole time interval of interest without considering potential local and time-delayed associations. We are specifically interested in discovering local and potentially time-delayed associations. Such associations have been shown to play important roles in understanding gene expression dynamics and the association of organisms in microbial communities [57, 74, 83, 110].

To understand local and time-delayed associations, we originally designed a Local Similarity Analysis (LSA) for time series data measured typically at successive and equal time intervals without replicates [83]. Studies adopting the original LSA technique have shown interesting and novel discoveries for microbial community datasets. To name a few, Paver et al. [57] successfully applied LSA to study glycolate-utilizing bacterial and phytoplankton associations, while Shade et al. [89] used LSA to discover bacterial association dynamics during lake mixing.

Since biological experiments are often associated with many potential sources of noise, repeated measurements (replicates) are usually carried out in order to better assess inherent uncertainties of the quantities of interest [54]. Furthermore, data emerging from

such experiments are typically analyzed by mean effect or by the development of profiles where variability is not properly accounted for [71]. Temporal and spatial data with replicates are being generated in Dr. Cardon’s laboratory and others. The lack of support for replicated data in the original LSA program has prevented its application to these new datasets. With replicates, it is possible to evaluate the variation of and to give a bootstrap confidence interval for the local similarity (LS) score as defined in Ruan et al. [83]. Furthermore, the original LSA is restricted by the low computing efficiency of the R language, as well as poor handling of missing values. In order to improve upon these issues and make the technique more accessible to the scientific community, we developed an extended LSA technique, named eLSA, and implemented it as a C++ extension to Python.

Briefly, given time series data of two factors and a user-constrained delay limit, eLSA finds the configuration of the data that yields the highest local similarity (LS) score, which is a type of similarity metric. For example, within a delay limit of two units, the first time spot of one series might be aligned to the third time spot of the other series, thus maximizing their LS. For a dataset of many factors, eLSA is applied to each pairwise combination of factors in the dataset. Candidate associations are then evaluated statistically by a permutation test, which calculates the p-value which is the proportion of scores exceeding the original LS score after shuffling the first series and re-evaluating the LS score many times, and by the false discovery rate (FDR q-value), which is used to correct multiple comparisons. Researchers can use eLSA to detect undirected associations, i.e., association patterns without time delays, and directed associations, where the change of one factor may temporally lead or follow another factor.

The organization of this chapter is as follows. In the first two sections, we describe the eLSA algorithm for calculating LS score with replicates, data normalization, estimation of confidence interval for the LS score, and testing the statistical significance of a LS score. We then describe briefly the implemented analysis pipeline of eLSA. In the next two sections, we first show the efficacy of eLSA by simulations, and finally apply the pipeline to analyze a microbiological dataset and a gene expression dataset. The chapter concludes with some discussion and conclusions.

3.2 Mathematical modeling and the eLSA algorithm

3.2.1 Local similarity analysis with replicates

The original LSA method considers only data without replicates. In this paper, we extend the Local Similarity Analysis (LSA) method [83] to samples with replicates. Suppose that the time series data for factors X and Y with replicates are measured simultaneously. We denote them as $X = X_{[1:n][1:m]}$ and $Y = Y_{[1:n][1:m]}$, where n is the number of samples (time points) and m is the number of replicates. Let $X_{i[1:m]}$ and $Y_{i[1:m]}$, or, in more abbreviated form, X_i and Y_j , be the vectors containing the m replicates from the i -th time spot of X and the j -th time spot of Y , respectively. To formulate the algorithm, we suppose each sample have m replicates and let F be some summarizing function for the repeated measurements. Thus, we extend the original LSA dynamic programming algorithm to data with replicates as in Algorithm 1.

Algorithm 1 Extended Local Similarity Analysis (eLSA)

Require: factor time series X and Y , delay limit D , summarizing function F ;

Variables: score matrix P , scoring function S_{XY} .

for i, j in $\{1, 2, \dots, n\}^2$ **do**

$$P_{0,i} = 0, P_{j,0} = 0 \text{ and } N_{0,i} = 0, N_{j,0} = 0$$

end for

for i, j in $\{1, 2, \dots, n\}^2$ with $|a - b| \leq D$ **do**

$$P_{i+1,j+1} = \max\{0, P_{i,j} + S_{XY}[F(X_{i[1:m]}), F(Y_{j[1:m]})]\} \text{ and}$$

$$N_{i+1,j+1} = \max\{0, N_{i,j} + S_{XY}[F(X_{i[1:m]}), F(Y_{j[1:m]})]\}$$

end for

$$P_{max}(X, Y) = \max_{1 \leq i, j \leq n} P_{i,j} \text{ and } N_{max}(X, Y) = \max_{1 \leq i, j \leq n} N_{i,j}$$

$$S_{max}(X, Y) = \frac{\max[P_{max}(X, Y), N_{max}(X, Y)]}{n} \text{ and}$$

$$S_{sgn}(X, Y) = \text{sgn}[P_{max}(X, Y) - N_{max}(X, Y)]$$

return $S_{max}(X, Y)$ and $S_{sgn}(X, Y)$

The $S_{max}(X, Y)$ obtained is the maximum local similarity score possible for all configurations of m -replicated time series X and Y within time-delay D . In this extended algorithm, the scalars x_i 's and y_i 's from the non-replicated series in Ruan et al. [83] are replaced by vector functions $F(X_i)$'s and $F(X_j)$'s to handle data with replicates. Alternatively, we can also consider $F(X_i)$'s and $F(X_j)$'s as the same input data for the original algorithm in Ruan et al. [83], except that they are F -transformed data. In addition, this extended LSA framework easily accommodates the original version of LSA without replicates ($m = 1$) as a special case.

3.2.2 Different ways of summarizing the replicate data

Notice that the only additional component we introduced in the eLSA algorithm is the function F . Many reports have suggested different possible forms for F , and several computational methods have been proposed for summarizing the additional information available from replicates, including the simple average method (abbreviated as ‘simple’) and the Standard Deviation (SD)-weighted average method (abbreviated as ‘SD’), and the multivariate correlation coefficient method [59, 125, 127]. However, the result of the multivariate correlation coefficient method from Zhu et al. [127] can be shown to be the same as the ‘simple’ method. Therefore, in eLSA, we used the first two methods. We also propose the use of median in place of average and Median Absolute Deviation (MAD) in place of SD when robust statistics are needed to handle outliers [31]. The corresponding methods are named simple median method (abbreviated as ‘Med’) and MAD-weighted median method (abbreviated as ‘MAD’), respectively.

The ‘simple’ method is, in spirit, to take the mean profiles to represent the replicated series. In practice, we take F to be the simple average of repeated measurements: $F(X_i) = \bar{X}_i$. The ‘SD’ method, on the other hand, takes the standard deviation of the replicates into account. Here we take F to be the replicate average divided by its standard deviation (SD): $F(X_i) = \frac{\bar{X}_i}{\sigma_{X_i}}$. Importantly, this method utilizes the variability information available, and, as such, it is claimed to be better than the ‘simple’ method in estimating the true correlation [125]. However, in order for the ‘SD’ method to be effective, a relatively large number of replicates, m , are needed, e.g., $m \geq 5$. For a small number of replicates, the ‘SD’ method may not work well since the standard deviation

may not be reliably estimated. Further, if we replace average with median and SD with MAD, we obtain the ‘Med’ method: $F(X_i) = \text{Median}(X_i)$ and the ‘MAD’ method: $F(X_i) = \frac{\text{Median}(X_i)}{\text{MAD}(X_i)}$, where $\text{MAD}(X_i) = \text{Median}(|X_i - \text{Median}(X_i)|)$. The two transformations have similar properties as their corresponding average and SD versions, but they are more robust.

3.2.3 Bootstrap confidence interval for the LS score

With replicate data, researchers can study the variation of quantities of interest and to give their confidence intervals. Due to the complexity of calculating the LS score, the probability distribution of the LS score is hard to study theoretically. Thus, we resort to bootstrap to give a bootstrap confidence interval (CI) for the LS score. Bootstrap is a resampling method for studying the variation of an estimated quantity based on available sample data [31]. In this study, we use bootstrap to estimate a confidence interval for the LS score. For a given type I error α , the $1 - \alpha$ confidence interval is the estimated range that covers the true value with probability $1 - \alpha$. Thus, for a given number, B , of bootstraps, we construct the bootstrap sample set $\{(\tilde{X}^{(1)}, \tilde{Y}^{(1)}), (\tilde{X}^{(2)}, \tilde{Y}^{(2)}), \dots, (\tilde{X}^{(B)}, \tilde{Y}^{(B)})\}$, where each $\tilde{X}_i^{(k)}$ and $\tilde{Y}_j^{(k)}$ are samples with replacement from X_i and Y_j , respectively. The rest of the calculation is the same as that used for the original data, and we obtain $\tilde{S}_{max}^{(k)} = S_{max}(\tilde{X}^{(k)}, \tilde{Y}^{(k)})$. Without the loss of generality, we suppose that these values are sorted in ascending order: $\tilde{S}_{max}^{(1)} \leq \tilde{S}_{max}^{(2)} \leq \dots \leq \tilde{S}_{max}^{(B)}$. Then, a $1 - \alpha$ bootstrap CI of S_{max} can be estimated by $[\tilde{S}_{max}^{(\lfloor \frac{\alpha}{2} B \rfloor)}, \tilde{S}_{max}^{(\lfloor (1 - \frac{\alpha}{2}) B \rfloor)}]$, as suggested by Efron et al. [31].

3.2.4 Data normalization

eLSA analyses require the series of factors X and Y to be normally distributed, but this may not be the case in the real dataset. Therefore, through normalization, the normality of the data can be enforced. To accommodate possible nonlinear associations and the variation of scales within the raw data, we apply the following approach [95] to normalize the raw dataset before any LS score calculations. We use $F(X_i)$ to denote the F -transformed data of the i -th time spot of a variable X . First, we take

$$R_k = \text{rank of } F(X_k) \text{ in } \{F(X_1), F(X_2), \dots, F(X_n)\}. \quad (3.1)$$

Then, we take

$$Z_k = \Phi^{-1}\left(\frac{R_k}{n+1}\right), \quad (3.2)$$

where Φ is the cumulative distribution function of the standard normal distribution. We will take $Z = Z_{[1:n]}$ obtained through the above procedure as the normalization of $\{F(X_1), F(X_2), \dots, F(X_n)\}$. Therefore, the normalization steps are taken after the F -transformation.

3.2.5 Permutation test to evaluate the statistical significance

It is important to evaluate the statistical significance of the LS score measured by the p-value, the probability of observing a LS score no smaller than the observed score when two factors are not associated locally or globally. To achieve this objective, permutation test is used. To perform the test, we fix Y and reshuffle all the columns of X for each

permutation. For a fixed number of permutations L , suppose $\{X^{(1)}, X^{(2)}, \dots, X^{(L)}\}$ is the permuted set of X ; then the p-value P_L is obtained using

$$P_{boot} = Prob[\tilde{S} \geq S_{max}(X, Y)] \approx \frac{1}{B} \sum_{k=1}^B I[\tilde{S}_{max}(\tilde{X}^{(k)}, \tilde{Y}^{(k)}) \geq S_{max}(X, Y)], \quad (3.3)$$

where $I(\cdot)$ is the indicator function. With large enough number of permutations, we can evaluate the p-value to any desired accuracy.

3.2.6 Computation complexity and implementation

For a single pair of time series, the time complexity for calculating the LS score using the dynamic programming algorithm is $O(n)$, where n is the number of time points. The estimation of the bootstrap confidence interval for the LS score using B bootstraps will need $O(Bn)$ calculations. The estimation of statistical significance for each pair of factors using L permutations will need $O(Ln)$ calculations. Thus, the number of calculations for a full analysis of each pair of factors will be $O(BLn)$. If there are a total of T factors, there are a total of $\frac{T(T-1)}{2}$ pairs of factors that need to be compared. Thus, the number of calculations for a full analysis of T factors will be in the order of $O(T^2BLn)$, which can be computationally intensive.

3.2.7 The eLSA analysis pipeline

In this subsection, we briefly describe the eLSA analysis pipeline implemented into the eLSA software package, as shown in Figure 3.1.

- ***F*-transformation and data normalization:**

The eLSA tool accepts a matrix file where each row is a time series for one factor. It fills up missing data by a user-specified method. Zero to third order spline-based methods and the nearest neighbor method as implemented in the Scipy (<http://www.scipy.org>) interpolation module are available. It then transforms the data by the user-specified F function and normalizes the F -transformed data by the normal score transformation following Li et al. [56].

- **Local similarity scoring:**

Local similarity analysis calculates the highest similarity score between any pair of factors. Users can specify parameters, including, for example, the maximum shifts allowed. Local Similarity score is calculated using the eLSA dynamic programming algorithm.

- **Permutation test:**

The statistical significance, the p-value, of LS score is evaluated using a permutation test. Briefly, eLSA randomly shuffles the components of the original time series and recalculates the LS score for the pairs. The p-value is approximated by the fraction of permutation scores that are larger (in absolute value) than the original score. Confidence interval for a given LS score is also found by bootstrapping from the replicated data. Finally, users can obtain significant eLSA association results by the combined use of p-value and FDR q-value thresholds as their filtering criteria.

- **Association network construction:**

Using only the significant associations, users can construct a partially directed association network. Generally, for two factors X and Y , if the time interval $[s_1, t_1]$ in X and $[s_2, t_2]$ in Y have the highest LS and $s_1 < s_2$, we can infer that X leads Y ; in other words, X possibly activates Y . In network visualization software (e.g., Cytoscape [23]), one can use arrows to directionally indicate these lead patterns (i.e., X to Y , if X leads Y ; otherwise undirected, if no direction is inferred). One can also use lines to indicate association types (solid, if X is positively associated with Y ; otherwise dashed). Following these rules, one can build a partially directed association network based on eLSA result.

In summary, the internal support for replicates and the use of CI estimates are the two major methodological enhancements to LSA. The eLSA software, however, also incorporates other new features, such as faster permutation and false discovery rate evaluations and more options to handle missing values. Other implementation details are available from the software documentation.

3.3 Materials and methods

3.3.1 Pearson’s correlation coefficient-based analysis

The application of Pearson’s Correlation Coefficient (PCC) requires taking the profile means, i.e. \bar{X}_i and \bar{Y}_i . Then the PCC between X and Y is defined as:

$$r(X, Y) = \frac{\sum_{1 \leq j \leq m} (\bar{X}_j - \bar{X})(\bar{Y}_j - \bar{Y})}{\sqrt{\sum_{1 \leq j \leq m} (\bar{X}_j - \bar{X})^2 \sum_{1 \leq j \leq m} (\bar{Y}_j - \bar{Y})^2}}$$

where $\bar{X}_j = \frac{1}{m} \sum_{k=1}^m X_{jk}$, $\bar{Y}_j = \frac{1}{m} \sum_{k=1}^m Y_{jk}$, $\bar{X} = \frac{1}{n} \sum_{j=1}^n \bar{X}_j$ and $\bar{Y} = \frac{1}{n} \sum_{j=1}^n \bar{Y}_j$ are the means of X and Y , respectively. The statistical significance of r is tested by the fact that $t = r \sqrt{\frac{n-2}{1-r^2}}$ follows a t -distribution (degree of freedom: $v = n - 2$, mean: 0 and variance $\frac{v}{v-2}$) when $m = 1$. For a pair of non-replicated series where $m = 1$, PCC is a straightforward and powerful method to test and identify linear relationship between two bivariate normally distributed random variables. It is widely adopted in the literature but with limitations. Specifically, when the real relationships are more complex, for example, the association between the two factors only occurs in a subinterval of the region of interest or the change of one factor has a time-delay in response to the change of another factor. Several methods, including the original LSA method, have been proposed to overcome such difficulties [6, 83]. We also include PCC analysis in conjunction of our eLSA analysis in the software pipeline.

3.3.2 False discovery rate (FDR) estimation

In most biological studies, a large number of factors need to be considered. If there are T factors, there will be $\frac{T(T-1)}{2}$ eLSA pairwise calculations, representing its quadratic growth in T . In order to avoid many falsely declared associated pairs of factors, we need to correct for multiple testing. Many methods have been developed to correct for multiple testing and here we use the method by Storey et al. [95] to address this issue. In particular, we report the q-value, Q , for each pair of factors. The q-value for a pair of factors is the proportion of false positives incurred when that particular pair of factors is declared significant.

3.4 Simulation studies

We generated simulated data to show the efficacy of eLSA in capturing time-dependent association patterns, such as time-delayed associations and associations within a subinterval. We also studied the difference between the eLSA inference using the simple average (referred to as ‘simple’) method, the SD-weighted average method (referred to as ‘SD’), the median (referred to as ‘Med’) method, and the MAD (referred to as ‘MAD’) method.

3.4.1 Time-delayed association

In this case, X and Y are assumed to be positively correlated with a time delay D . For a particular example with $D = 3$, we assume that (X_{j+3}, Y_j) 's follows a bivariate normal distribution with mean $\mu = \mathbf{0}$ and covariance matrix $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. for $j = 1, 2, \dots, 20$, where $\rho = 0.8$. X_j 's are assumed to be standard normal for $j = 1, 2, 3$. The generated (X_j, Y_j) 's are further perturbed m times by a measurement disturbance $\epsilon_{ij} \sim N(0, 0.01)$ to obtain the m -replicated series. A pair of simulated series is shown in Figure 3.2a for a typical simulation with $m = 5$.

We see that the two series closely follows each other if we shift the Y series three units toward right. In this particular example, the PCC is -0.258 ($P=0.272$) while the LS score using ‘simple’ averaging method is 0.507 with a p-value of 0.006. We did 1000 bootstraps and the 95% bootstrap confidence interval for this particular example is (0.448, 0.549). Therefore, this time-delayed association is only found significant by the eLSA analysis.

3.4.2 Association within a subinterval

In this case, we assume X and Y are positively associated within a subinterval and not associated in other regions. In our simulation, we generate 20 time spots of the two series by sampling (X_j, Y_j) from a bivariate-normal distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ where $\rho = 0.8$ for $6 \leq j \leq 15$, and $\rho = 0$ for $j \leq 5$ or $16 \leq j \leq 20$. The generated (X_j, Y_j) 's are further perturbed m times by a measurement disturbance $\epsilon_{ij} \sim N(0, 0.01)$ to obtain the m -replicated series. One generated series are shown in Figure 3.2b for a typical simulation with $m = 5$.

We can see the two series mostly closely follow each other within the intended subinterval $6 \leq j \leq 15$. In this particular example, the PCC is 0.258 ($P=0.272$) while the LS score using ‘simple’ averaging method is 0.428 with a p-value of 0.028. We did 1000 bootstraps and the 95% bootstrap confidence interval is (0.404,0.446). This pattern is again uniquely captured by the eLSA analysis. In real applications, there are many other possibilities that two factors are associated without a significant Pearson or Spearman’s correlation coefficient. The eLSA can capture these associations as long as their LS score can be maximized through dynamically enumerating their configurations.

3.4.3 Different summarizing function

To see the effect of replicates, we also let $m = \{1, 10, 15, 20\}$ in the time-delayed simulation and did the same analysis as above with 1000 simulations. The results are summarized in Table 3.1. It can be seen from the table that the results using ‘simple’ and ‘Med’ are similar with mean LS scores ranging from 0.490 to 0.498 and standard errors ranging from 0.078 to 0.091. On the other hand, if the noise in the replicates is not normally

distributed, the ‘Med’ method should be more robust. On the other hand, the mean LS scores using ‘SD’ and ‘MAD’ are generally lower than that using the ‘simple’ and ‘Med’ methods. This maybe caused by the extra variation introduced when estimating the standard deviation or maximum absolute deviation from the data.

<i>F</i> -function	<i>m</i> =1		<i>m</i> =5		<i>m</i> =10		<i>m</i> =15		<i>m</i> =20	
	mean	se.	mean	se.	mean	se.	mean	se.	mean	se.
‘simple’	.495	.078	.495	.085	.491	.088	.493	.076	.496	.091
‘SD’	na.	na.	.332	.127	.391	.124	.412	.119	.435	.109
‘Med’	.495	.078	.490	.090	.490	.090	.490	.083	.498	.083
‘MAD’	na.	na.	.494	.115	.302	.128	.325	.129	.371	.119

Table 3.1: Mean and standard error of the estimated LS score. The values are calculated based on 1000 simulations. ‘se.’ indicates standard error and ‘na.’ indicates not applicable.

3.4.4 Running time comparison

We benchmarked the running time performance of the new eLSA implementation and the old R script. For a dataset of 72 time series each with 35 time points, we tried eLSA analysis with 100 bootstraps, 1000 permutations and a delay limit of 3. It took the old script 20462 seconds to finish the computation while the new C++ program used 2054 seconds, which is about 9 times faster. Meanwhile, the new implementation also reduces the memory consumption and increases input/output efficiency. The benchmark is carried out on a “Dell, PE1950, Xeon E5420, 2.5GHz, 12010MB RAM” computing node.

3.5 Application to real data

3.5.1 Microbial community data analysis

As an immediate application, we applied the eLSA pipeline to a set of real microbial community time series data. This San Pedro Ocean Time Series (SPOTs) dataset, originally reported in Steele et al. [92] and Countway et al. [24], was collected following a biological feature (i.e. the chlorophyll maximum depth) off the coast of Southern California. The bacterial community was analyzed using the ARISA [34] technique and the protistan community was analyzed using the T-RFLP [109] technique. The dataset is composed of monthly sampled data from September 2000 to March 2004, including 40 time points without replicates. We analyzed the dataset with a delay limit of 3 months and 1000 permutations to evaluate the statistical significance of the LSA score. In this dataset, the factor names, including the operational taxonomic units and environmental factors, are previously defined by Steele et al. [92].

First, we compared the performance of Pearson’s correlation coefficient (PCC) and eLSA analysis in identifying potential local and time-delayed associations. Restricting the significance threshold for the q-value $Q \leq 0.01$ and the p-value $P \leq 0.01$, 1643 pairs of significant associations with eLSA were identified, and among them only 293 (~18%) were discovered by PCC (see Table 3.2). Therefore, most significant associations found by eLSA would have been missed by PCC analysis in this case. The results are similar if we use less stringent criteria, i.e., $Q \leq 0.05$ and $P \leq 0.05$, where only 658 out of 2804 (~23%) eLSA significant associations were also found by PCC. We need to point out that, PCC also found some associations that were missed by eLSA. For example, with q-value

$Q \leq 0.01$ and the p-value $P \leq 0.01$, PCC found 3237 significant associations and only 293 of them were found to be significant using eLSA. Therefore, eLSA is not a substitute but a complimentary approach to PCC, which specializes in finding local and possibly time-delayed associations. For a thorough analysis of a dataset, one should apply both approaches, which is why we also integrated PCC analysis into our software pipeline.

Dataset	# of factors	$P \leq 0.01$ and $Q \leq 0.01$			$P \leq 0.05$ and $Q \leq 0.05$		
		eLSA	PCC	both	eLSA	PCC	both
Microbial	515	1643	3237	293	2804	4242	658
<i>C. elegans</i>	446	42532	56605	39114	57991	71799	54201

Table 3.2: Significant associations found in real datasets. Numbers of significant associations found by the extended Local Similarity Analysis (eLSA) and Pearson’s Correlation Coefficient (PCC) by controlling both the p-value (P) and the q-value (Q). The p-values for eLSA were evaluated by permutations and p-values for PCC was calculated based on the t -distribution.

If we look at the top five positive and negative absolute highest LS scores from the unique associations ($|D| \leq 1$) found by eLSA ($Q \leq 0.05$ and $P \leq 0.05$, see Table 3.3), we can see most of them are time-dependent associations, either time-shifted or within a subinterval. The majority of these are, in any case, beyond the capacity of PCC. In addition, eLSA provides more information about its findings. For example, in the table, Bac609 and Bac675 factors are associated with a shift of one and Euk97 and boxy (oxygen) factors are best associated within a time interval of length 21 starting at time point 15 with no delay. This kind of additional information is not easily obtainable from the PCC analysis but very important for further functional analysis. For instance, we construct an association network using all above unique eLSA associations, as shown in Figure 3.3. The obtained network obviously reveals some interesting dynamics of

the microbial community, such as the domination of positive directed associations, the existence of environmental factors as hubs that are associated with many other factors, (e.g. nutrients such as NO_2 , PO_4 , SiO_3 and oxygen), and the existence of some highly connected clusters formed by certain bacteria or eukaryote groups.

X	Y	LS	Xs	Ys	Len	D	P	PCC	Ppcc	Q	Qpcc
Euk239	Euk269	.82	1	1	40	0	0	.09	.59	.02	1.
Bac609	Bac675	.77	1	2	39	-1	0	.14	.41	0.	1.
Euk381	Euk462	.77	1	1	40	0	0	.44	0.	.02	.11
Euk583	Euk989	.68	2	1	39	1	0	.30	.06	.02	.73
Euk229	Euk339	.57	1	2	39	-1	0	.05	.77	.02	1.
Euk97	boxy	-.62	15	15	21	0	0	-.42	.01	0.	.17
Euk98	boxy	-.62	15	15	21	0	0	-.42	.01	0.	.17
Euk109	boxy	-.62	15	15	21	0	0	-.42	.01	0.	.17
Euk112	boxy	-.62	15	15	21	0	0	-.42	.01	0.	.17
Euk116	boxy	-.62	15	15	21	0	0	-.42	.01	0.	.17

Table 3.3: Top LS scores from the microbial community data. The 5 positive and 5 negative highest absolute LS Scores from associations uniquely found by eLSA in the microbial community dataset. The columns in succession are X (first factor), Y (second factor), LS (Local Similarity score), Xs (start of the best alignment in the first sequence), Ys (start of the best alignment in the second sequence), Len (alignment length), D (shift of the second sequence compared to the first sequence, -: X is ahead of Y, +: otherwise), P (p-value for the LS score, 0. stands for $P < 0.005$), PCC (Pearsons Correlation Coefficient), Ppcc (P-value for PCC), Q (q-value calculated for P, 0. stands for $Q < 0.005$), Qpcc (q-value for Ppcc).

Taking a closer look at one of the topmost ranked association: Bac609 and Bac675 (see Table 3.3), we found that they are closely following each other with a time shift of one month, where Bac609 precedes Bac675. Further inspection suggests a yearly pattern that recurs with near regularity for this association, such that Bac609 blooms in early springtime each year (time spots 6, 18 and 29 are February, January and March, respectively), and Bac675 blooms one month later (see Figure 3.4a). From the binning definition in Steele et al. [92], Bac609 is a *Bacteroidetes* group bacterium while Bac675 is

an undefined bacterium. Since these microbial groups are uncultured, this association as well as many others uniquely identified by eLSA provides new insight into their ecological role in the ocean surface waters. Notice there is an unexpected abundance jump at time spot 35 of the Bac675 series. The reason for this outlier however is unknown to us. While such prominent time-delayed associations as the Bac609 and Bac675 are easily visible, we must caution that time-dependent associations could also be too subtle to be viewed directly. Thus, statistical significance can provide a much more reliable guideline.

3.5.2 Gene expression data analysis

Although LSA had its roots grounded in microbial community analysis, the technique can be readily applied to other biological time series data, such as replicated gene expression time series data from microarray and RNA-Seq experiments [7, 96, 97]. Here we show an example of applying eLSA to the dauer exit gene expression profile time series data of 446 genes from a *C. elegans* study. The result of the original study suggests that the 446 genes under investigation have similar kinetics in both the dauer exit and the L1 starvation time course [111]. Here we use the dauer exit time series data consisting of 12 hourly time spots, each with four replicates. We analyzed the dataset with a delay limit of 3 hours and with 1000 permutations and 100 bootstraps.

The results are summarized in Table 3.2. Comparing the *C. elegans* results to those of the microbial community, we see that gene-gene associations in this network are much denser, since a smaller number of genes end up with a much larger, rather than smaller, number of eLSA significant associations (e.g. 2804 versus 57991 for $Q \leq 0.05$ and $P \leq 0.05$, see Table 3.2). Also different is that about 93% of these associations are found by PCC

analysis as well. The high congruence between PCC and eLSA analysis may be due to the fact that about 90% of the eLSA findings are without delays, which thus are also amenable to PCC analysis.

Because these genes do not change expression level in both dauer exit and L1 starvation conditions, they are considered as common feeding response genes [111]. However, it is not clear whether they are correlated with each other in expression profiles under the dauer exit condition. To study this, we combined all eLSA and PCC significant associations with $Q \leq 0.05$ and $P \leq 0.05$, and found the average degree of the resulting association network is around 169, while that of previous microbial community data is around 12. Such high average degree for *C. elegans* genes shows the high similarity of their expression profiles, which also reflects their intimate functional coordination along the process. Therefore, our result suggests those feeding response genes are likely to be co-expressed under the dauer exit condition.

We next analyzed the unique eLSA associations. These associations form a dense association network themselves with a long-tailed degree distribution, as shown in Figure 3.5. While the degree distribution peaks at five, the most highly connected gene *48941* has a degree of 189. We also looked at the top 5 positive and 5 negative highest absolute LS scores unique associations by eLSA. Because replicates are available for this dataset, we are able to obtain the bootstrap confidence intervals for the LS score and they are given in Table 3.4. Interestingly, we found most of the top LS associations involve high degree nodes, such as genes *48941*(189), *29494*(129), *29504*(128), *27993*(116), *436287*(106), *32607*(58), and *51986*(52) (degree in parenthesis). These high degree nodes could be regulation hubs in the feeding response pathway. Here we show an example of

time-delayed association of gene *32607* and gene *51986* in Figure 3.4b. In the figure, gene *51986* leads gene *32607* in expression profile change.

We also analyzed all the eLSA associations together, including both unique and non-unique eLSA findings. Though most of the genes are still hypothetical protein coding genes, we do find a group of eukaryotic initiation factors: *30080*(eIF-3E), *33683*(eIF-3K), *21358*(eIF-3D), *33525*(eIF-4E), *32503*(eIF-1A) and *23975*(eIF-2B) in the 446 selected genes. This is as expected because both L1 starvation recovery and dauer exit will increase translation activities and result in high expression level of these genes. In addition, in the translation process, these factors work closely together to form different translation related complexes [49], so their expression levels should be highly correlated with each other. Actually, if we check the associations found by eLSA, we do see these factors form a clique together with all edges being positive associations and statistically significant (see Figure 3.6). The coherence of the eLSA finding and our biological knowledge shows that eLSA associations do reveal true associations within the biological system. However, as the majority of genes are still hypothetical, a thorough examination for true functional discoveries will require biological experiments.

3.6 Discussion

The eLSA technique extends LSA to time series data with replicates. This will help investigators better utilize the available information from their sample replicates and assist them in more effective and reliable hypothesis generation of time-dependent associations. In addition, a bootstrap framework is developed to estimate the confidence interval for the

X	Y	LS	lowCI	upCI	Xs	Ys	Len	D	P	PCC	Ppcc	Q	Qpcc
48087	27993	.53	.41	.61	1	2	11	-1	0.	.56	.06	0.	.01
32607	51986	.52	.41	.61	2	1	10	1	.01	.51	.09	0.	.01
29504	48087	.52	.40	.61	2	1	11	1	0.	.41	.18	0.	.03
23193	27993	.51	.41	.59	1	2	11	-1	0.	.48	.11	0.	.02
29494	30208	.51	.39	.61	2	1	11	1	0.	.58	.05	0.	.01
27993	53694	-.55	-.62	-.44	2	1	11	1	0.	-.53	.08	0.	.01
436287	53694	-.54	-.62	-.44	2	1	11	1	.01	-.55	.06	0.	.01
48941	53694	-.52	-.61	-.42	2	1	11	1	0.	-.38	.22	0.	.03
29494	22857	-.52	-.61	-.41	2	1	11	1	0.	-.49	.10	0.	.02
29494	436727	-.52	-.61	-.40	2	1	11	1	.01	-.55	.06	0.	.01

Table 3.4: Top LS scores from the *C. elegans* gene-expression data. The 5 positive and 5 negative highest absolute LS Scores from the *C. elegans* gene expression dataset The columns in succession are X (first factor), Y (second factor), LS (Local Similarity score), lowCI (CI is lower bound), upCI (CI is upper bound), Xs (start of the best alignment in the first sequence), Ys (start of the best alignment in the second sequence), Len (alignment length), D (shift of the second sequence compared to the first sequence, -: X is ahead of Y, +: otherwise), P (p-value for the LS score, 0. stands for $P < 0.005$), PCC (Pearsons Correlation Coefficient), Ppcc (P-value for PCC), Q (q-value calculated for P, 0. stands for $Q < 0.005$), Qpcc (q-value for Ppcc).

LS score. We also provided flexible missing value options and integrated efficient multiple testing control methods for the new eLSA technique. Using the microbial community and gene expression datasets, we demonstrated that eLSA uniquely captures additional time-dependent associations, including local and time-delayed association patterns, when compared to ordinary correlation methods, such as PCC. In this chapter, we described the applications of our method with the time series data. Actually, the eLSA can be applied to any type of data with some gradients, including the response to different levels of treatments, temperature, humidity, or spatial distributions.

Currently, we use permutation test to assess the statistical significance of LS scores and bootstrap re-sampling to estimate the confidence interval of LS score. Both the permutation test and bootstrap methods are time consuming if high precise determination of statistical significance or confidence interval is desired. Theoretical developments on

the distribution of the LS score are needed to eliminate or mitigate the computational burden required for these processes, and would be interesting topics for future studies. There is also a minimum sample number requirement for eLSA analysis. We suggest the sample number to be greater than $5+D$, where D is the desired delay limit, since shifting and trimming by eLSA will further reduce the effective sample number and result in lower statistical power.

Finally, we implemented the eLSA technique and analysis pipeline into an Open Source C++ extension to Python with many new features. Specifically, the pipeline streamlines data normalization, local similarity scoring, permutation testing and network construction. As shown in Figure 3.7, we also provide a Galaxy web framework-based version [40] of the eLSA pipeline. This eLSA service features customized workflow, history and data sharing. In addition, we integrated Cytoscape [23] Java Web Start technology so that the association network generated by eLSA can be immediately visualized. Based on these efforts, we anticipate that our novel eLSA methodology, as implemented by the newly developed pipeline software, will significantly assist researchers requiring systematic discovery of time-dependent associations. More information about the software and web services is available from the eLSA homepage at <http://meta.usc.edu/softs/lisa>.

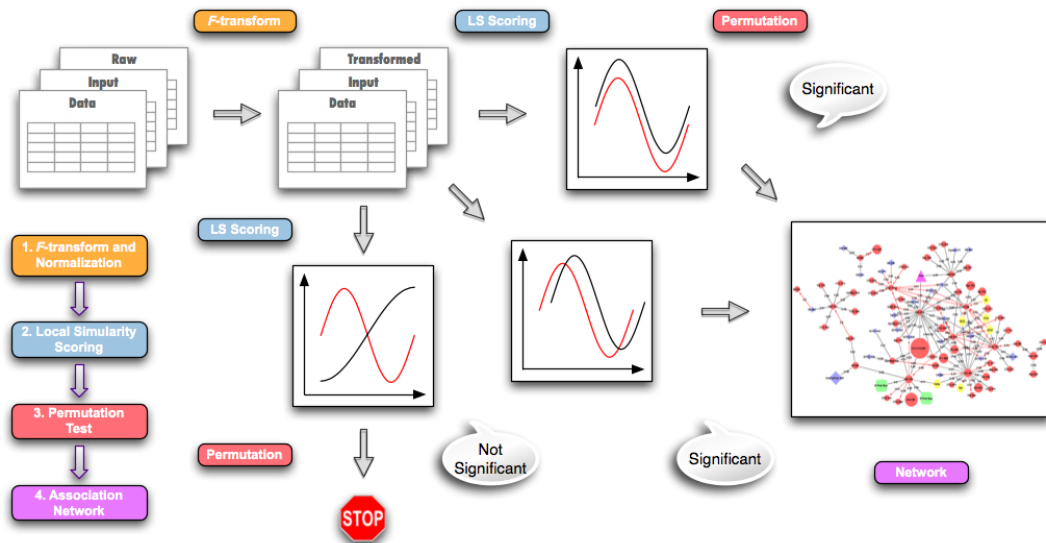


Figure 3.1: The eLSA pipeline. Users start with raw data (matrices of time series) as input and specify their requirements as parameters. The LSA tools subsequently F -transform and normalize the raw data and calculate Local Similarity (LS) scores and Pearson's Correlation Coefficients. The tools then assess the statistical significance (P-values) of these correlation statistics using the permutation test and filter out insignificant results. Finally, the tools construct a partially directed association network from the significant associations.

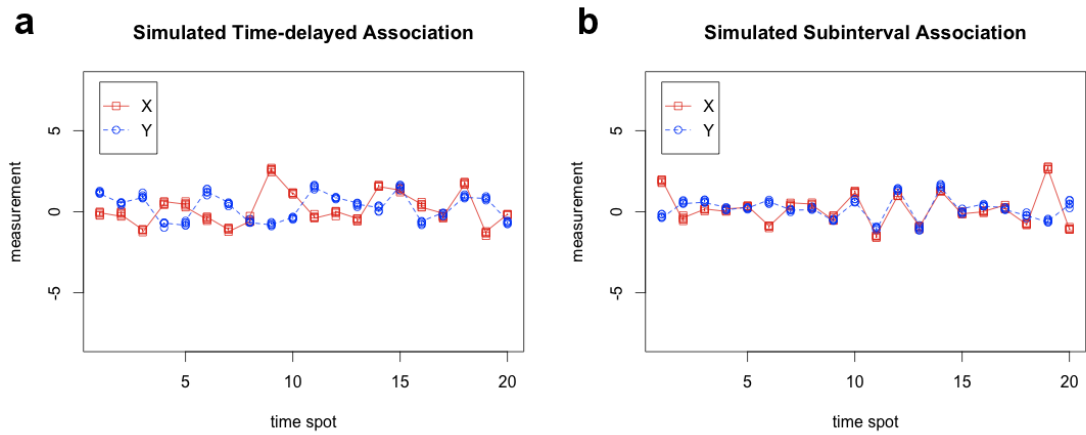


Figure 3.2: Examples of simulated associations. (a) An example of simulated time-delayed association series with five replicates is shown, where X (red square) leads Y (blue circle) by three time units. The pattern is not significant by ordinary correlation analysis ($PCC=-0.258$, $P=0.272$); however, it is captured by local similarity analysis ($LS=0.507$, $P=0.006$). (b) An example of simulated subinterval association series with five replicates is shown, where X (red square) and Y (blue circle) are associated in the time interval from 6 to 15. The pattern is not significant by ordinary correlation analysis ($PCC=0.258$, $P=0.273$); however, it is captured by local similarity analysis ($LS=0.428$, $P=0.028$).

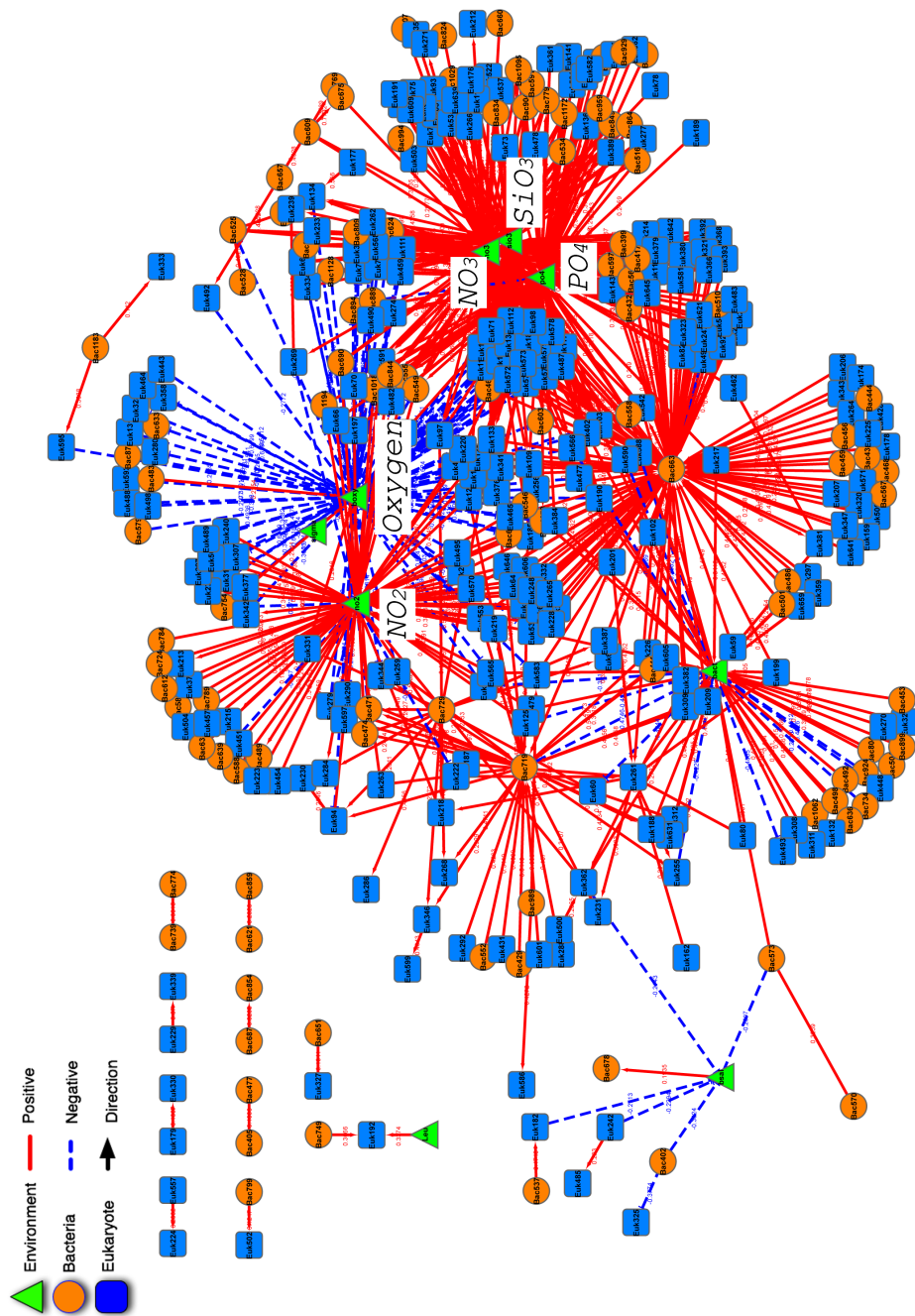


Figure 3.3: Typical association network from the microbial community data. Round- (brown), square- (blue) and triangle- (green) shaped nodes are bacteria, eukaryotes and environmental factors, respectively. Solid (red) edges are positively associated, while dashed (blue) edges are negatively associated. Arrow indicates the time-delay direction.

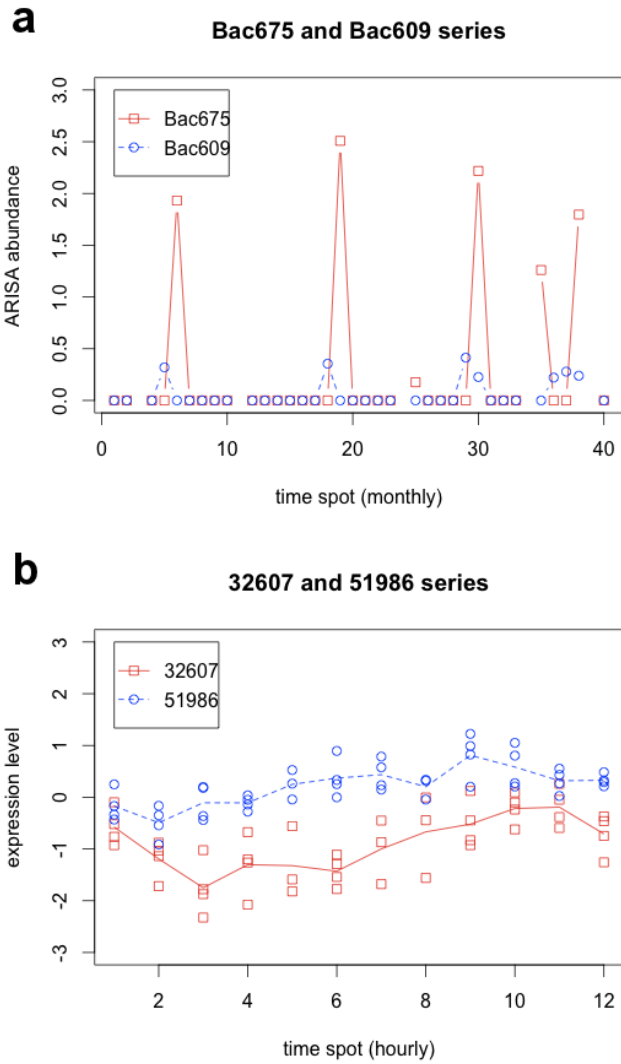


Figure 3.4: Examples of real data associations. (a) Shown are microbe group Bac675 (red square) and Bac609 (blue circle) ARISA abundance time series from the marine microbial community data analysis. Notice that there exists an almost regular yearly pattern where Bac609 leads Bac675 by one month in blooming time. (b) Shown are gene *32607* (red square) and *51986* (blue circle) expression level time series from *C. elegans* gene expression data analysis. Notice that *51986* leads *32607* in expression level change throughout the time course.

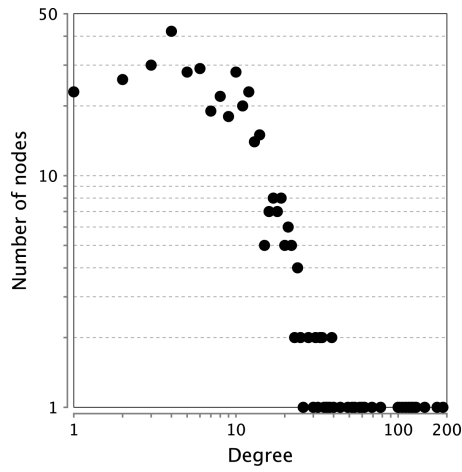


Figure 3.5: Node degree distribution of associations in *C. elegans* analysis. Shown is the node degree distribution of eLSA unique associations in *C. elegans* analysis. It shows a long-tail distribution with the maximum 189.

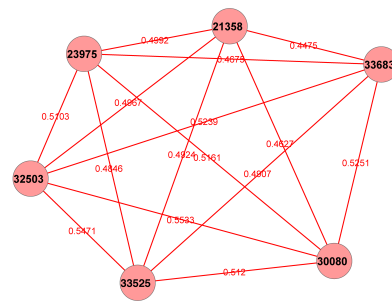






Figure 3.6: Translation initiation factor associations in *C. elegans* analysis. Shown is the association network of translation initiation factors learned from eLSA analysis. Solid (red) edges are positively associated. Edge labels are LS scores. The factors form a clique as expected.

LSA Compute

LSA Compute Input:
 
 See \bf{LSA Compute Input Format} Below

DELAYLIMIT:
 
 Maximum Delay Possible, default is NO delay

PERMUNUM:
 
 Number of Permutations, default is 1000

BOOTNUM:
 
 Number of Permutations, default is 100

REPNUM:

 Number of Replicates, must be provided and valid with your data

SPOTNUM:

 Number of Timespots, must be provided and valid with your data

TRANSFUNC:
 simple
 SD
 Method to Summarize Replicates, default is simple averaging

FILLMETHOD:
 none
 zero
 linear
 slinear
 quadratic
 cubic
 nearest
 Method to Fill Missing Data, default is none (filling ZEROS)

Figure 3.7: Submission interface for the LSA web service. Upon submission, the job will perform eLSA analysis on the ‘CommonGenesData’ dataset (12 time spots and 4 replicates) with 200 permutations and 100 bootstraps within a delay limit of 3 units. In addition, by specification, it will use ‘simple’ averaging to summarize replicates and, by designating ‘none’, it will disregard the missing values.

Chapter 4

Future work

4.1 Future work for GRAMMy

In Chapter 2, we described the GRAMMy framework and tool for shotgun metagenomics. We have mentioned several potential directions worth exploring in the discussion there and we will elaborate on them below.

First, there is the possibility to transfer the same methodology to the question of abundance estimation of functional gene groups based on shotgun metagenomic reads. Like the taxonomic relative abundance, relative abundance of functional gene groups is currently estimated using direct counting [29] and Lowest Common Ancestor (LCA) approaches [47]. These methods do not take into account the read assignment ambiguities, and are susceptible to the biases caused by gene length variation of functional gene groups. To improve upon current methods, in particular resolving the two issues raised above, we can directly apply our GRAMMy method to model the problem as a mixing parameter estimation problem for a mixture of functional gene groups: $M : M = \sum_{j=1}^m \pi_j f_{g_j}$. Now,

g_j is the j -th functional group of the m groups and f_{g_j} is the probability a read can be generated from g_j .

The remainder of the method then follows the same procedures as our previous derivation, except that, l_{g_j} – the effective length of g_j and f_{g_j} – the probability of a read is from g_j , have to be approximated and estimated differently. In an *ad hoc* manner, we can take the l_{g_j} to be the average of all known genes belonging to that functional group. We can also calculate f_{g_j} based on the normalized effective map counts or the k -mer composition distance between the read and the functional group, similar to what we did for the genome relative abundance estimation. Finally, the lower level functional group estimates can also be summed up to give estimates for higher level functional groups. We can first see how far this direct application can take us with simulated data and improve upon these preliminary results with more realistic modeling.

Second, for the original taxonomic relative abundance problem, we can also borrow ideas from functional group relative abundance estimation scheme proposed above, in the sense that, metagenomic sampling may be better described by a two layer mixture model, including a functional layer and a taxonomic layer. This motivation is rooted to the facts that genes form homolog (mostly functional similar) groups, in which genes from the same functional group are expected to be similar in sequence characteristics; and that there is a prevalence of Horizontal Gene Transfers (HGT) in the microbial world [81], which migrates the genes across different microbe organisms. Therefore, we can expect the same homolog groups to appear in different microbes because of HGTs, as well as the evolution descendance. As a result, the read ambiguity may be better resolved at the

homolog group level so that the taxonomic abundance can be more reliably estimated based on homolog profiles.

In detail, suppose each read set \mathbf{R} is a sample from the mixture of reads from \mathbf{C} functional groups. Then the finite mixture model for the sample space of read set \mathbf{R} is:

$$\Omega_{\mathbf{R}} = \beta_1 C_1 + \beta_2 C_2 + \cdots + \beta_l C_l, \quad (4.1)$$

where \mathbf{C} are functional groups which are well separated either by their k -mer composition or sequence similarity. A reasonable assumption is that the groups follow a multivariate gaussian distribution. With such an assumption, the mixing parameters can be inferred from unsupervised learning for gaussian mixtures. Then, we consider a functional group C_j is shared by multiple genomes, where the mixing model is:

$$C_j = \gamma_{j1} G_1 + \gamma_{j2} G_2 + \cdots + \gamma_{jm} G_m. \quad (4.2)$$

Then, if the genome relative abundance is directly estimated (as in the GRAMMy), we have:

$$\Omega_{\mathbf{R}} = \alpha_1 G_1 + \alpha_2 G_2 + \cdots + \alpha_m G_m. \quad (4.3)$$

From above, we can find:

$$\alpha_i = \sum_{j=1}^l \beta_j \gamma_{ji} \quad (4.4)$$

If the missing parameters in the two-layer model can be more reliably estimated (using the EM procedure or others), we can subsequently find more reliable genome relative

abundance for the original problem. However, the gain by including an additional layer has to be checked by implementation and simulations.

Third, there can be improvements from the technical aspects in computation. Currently, the GRAMMy is designed and implemented as a sequential algorithm, whose future use may be limited by the available memory to a single core. However, we can parallelize the EM algorithm to avoid such bottleneck of memory usage, as well as improve the computation efficiency. The real technical difficulty lies in storing the missing value matrix \mathbf{Z} , of which each row is a read and each column is a genome. In the near future, the number of reads can be hundreds of millions while the number of genomes can also be tens of thousands. Thus the matrix would seize up memory size in the order of 10^4 GB using full matrix storage and still tens of GB using sparse matrix storage (as currently implemented in GRAMMy). In fact, we can follow the work of Chen et al. [21] to partition the missing value matrix \mathbf{Z} into stripes of rows or columns and store them separately across different cores. Since the EM algorithm is mostly column or row sums, it can be coordinated parallel with message passing interface. In this way, the memory bottleneck is circumvented and if the acceleration from parallelization outweighs the overhead of data communication, there is also a speed up in computation time.

4.2 Future work for eLSA

In Chapter 3, we described the extended LSA technique for replicated time series analysis. Here, we will elaborate the potential improvement directions for eLSA. Like GRAMMy,

an immediate technical difficulty that eLSA face is the ever increasing number of factors. Currently the algorithm utilizes permutation test strategy for p-value evaluation. However the computation is time-consuming for a large number of pairwise computations.

We can parallelize the current algorithm. Using the coarse-grained parallelization, we can allocate pairwise jobs by partitioning the pairwise matrix. For example, if there are T factors and the time cost for one pairwise computation is one unit, for the sequential eLSA, it would take $\frac{T(T-1)}{2}$ units of time to complete. However if we have multiple cores, say m^2 cores, we can divide the jobs into this number of chunks, where the pairwise computations between factor sets indexed by $\{\frac{T}{m}i + 1, \frac{T}{m}i + 2, \dots, \frac{T}{m}(i + 1)\}$ and $\{\frac{T}{m}j + 1, \frac{T}{m}j + 2, \dots, \frac{T}{m}(j + 1)\}$ are carried out on the $(mi + j)$ -th core. We can expect a speed up of m^2 using this simple parallelization.

4.3 Molecular microbial ecology analysis pipeline

There are several methodology developments from our group addressing different issues in molecular microbial ecology high-throughput data analysis including dynamic binning, OTU calling for 16S RNA sequencing, local similarity analysis, eLSA and GRAMMy [42, 43, 83, 84, 118–122, 126]. We can bring them all together and forge an integrated analysis pipeline, see Figure 4.2.

As shown in the figure, there are three basic types of molecular microbial ecology experimental data, including shotgun metagenomic reads, 16S RNA sequencing reads and data from non-sequencing molecular technologies, such as ARISA and TRFLP. Our tools like GRAMMy, Crop and Dynamic Binning can be used to preprocess data and

prepare a taxonomic relative abundance profile for each sample, either in the form of OTUs or genomes, for subsequent analysis. Then, essential ecological analysis such as richness (alpha diversity) can be assessed from the output of above tools. Further, for data that are time series, we can pass them to association analysis tools like eLSA, which can in addition identify significant time-dependent associations. The resulting associations can then be fed into cytoscape to generate an association network for subsequent network based analysis.

We can further integrate our pipeline with other state-of-art and widely accepted software platforms, such as Galaxy, Cytoscape, QIIME [17, 23, 40], which will help bring our tools to more audience in molecular biology and microbial ecology community and assist investigators in their practical studies.

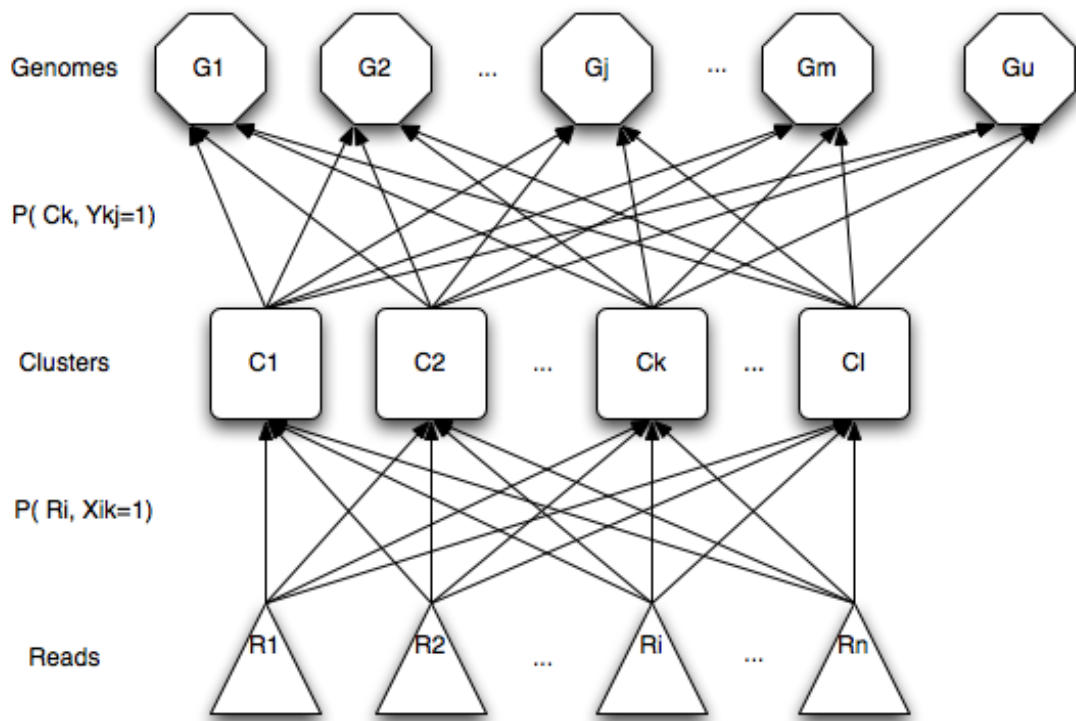


Figure 4.1: A two-layer mixture model for taxonomic relative abundance estimation.

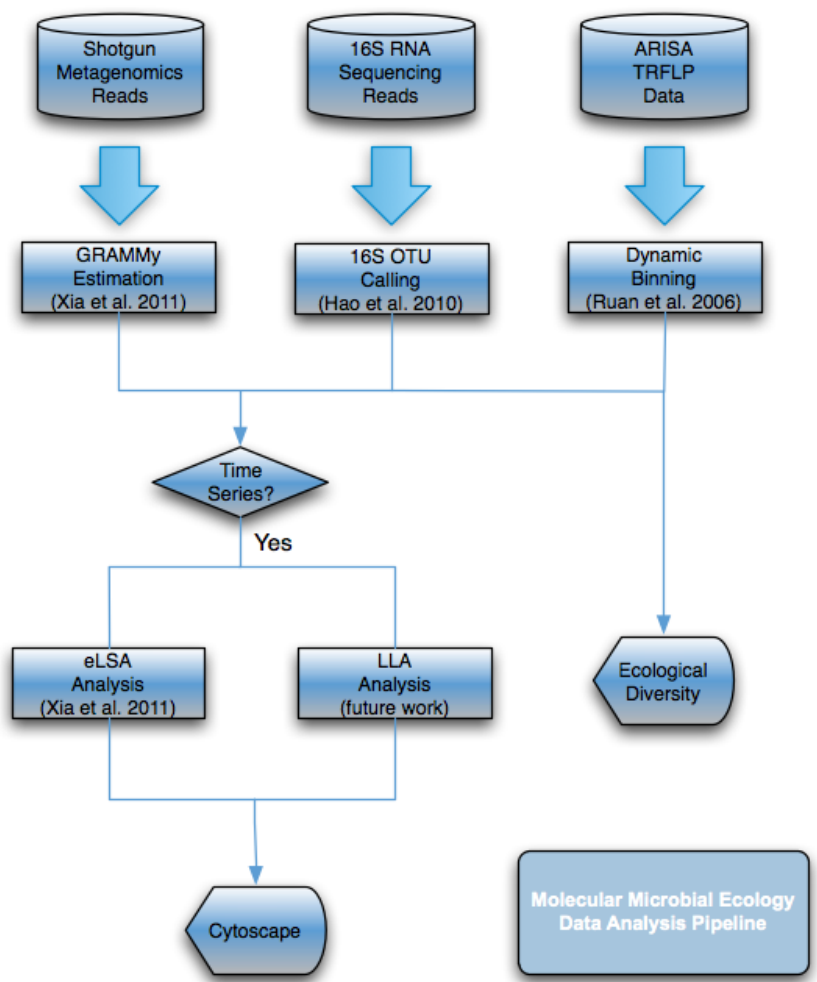


Figure 4.2: A molecular microbial ecology analysis pipeline integrating tools developed by our groups.

Bibliography

- [1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.
- [2] F. E. Angly, D. Willner, A. Prieto-Davo, R. A. Edwards, R. Schmieder, R. Vega-Thurber, D. A. Antonopoulos, K. Barott, M. T. Cottrell, C. Desnues, et al. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol*, 5(12):e1000593, 2009.
- [3] W. T. Astbury. Molecular biology or ultrastructural biology? *Nature*, 190:1124, 1961.
- [4] L. H. Augenlicht and D. Kobrin. Cloning and screening of sequences expressed in a mouse colon tumor. *Cancer Res*, 42(3):1088–93, 1982.
- [5] E. Avaniss-Aghajani, K. Jones, D. Chapman, and C. Brunk. A molecular technique for identification of bacteria using small subunit ribosomal RNA sequences. *Biotechniques*, 17(1):144–6, 148–9, 1994.
- [6] R. Balasubramaniyan, E. Hullermeier, N. Weskamp, and J. Kamper. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, 21(7):1069–77, 2005.
- [7] Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–503, 2004.
- [8] O. Beja, L. Aravind, E. V. Koonin, M. T. Suzuki, A. Hadd, L. P. Nguyen, S. B. Jovanovich, C. M. Gates, R. A. Feldman, J. L. Spudich, et al. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, 289(5486):1902–6, 2000.
- [9] S. M. Berget, C. Moore, and P. A. Sharp. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A*, 74(8):3171–5, 1977.
- [10] B. E. Bernstein, M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, D. J. Huebert, S. McMahon, E. K. Karlsson, r. Kulbokas, E. J., T. R. Gingeras, et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–81, 2005.
- [11] B. Beszteri, B. Temperton, S. Frickenhaus, and S. J. Giovannoni. Average genome size: a potential source of bias in comparative metagenomics. *ISME J*, 4(8):1075–7, 2010.

- [12] E. J. Biers, S. Sun, and E. C. Howard. Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol*, 75(7):2221–9, 2009.
- [13] A. Brady and S. L. Salzberg. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods*, 6(9):673–6, 2009.
- [14] F. J. d. Bruijn. *Handbook of molecular microbial ecology I: metagenomics and complementary approaches*. John Wiley & Sons, Inc., Singapore, 2011.
- [15] F. J. d. Bruijn. *Handbook of molecular microbial ecology II: metagenomics in different habitats*. John Wiley & Sons, Inc., Singapore, 2011.
- [16] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268(1):78–94, 1997.
- [17] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–6, 2010.
- [18] R. J. Case, Y. Boucher, I. Dahllorf, C. Holmstrom, W. F. Doolittle, and S. Kjelleberg. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol*, 73(1):278–88, 2007.
- [19] S. Chaffron, H. Rehrauer, J. Pernthaler, and C. von Mering. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res*, 20(7):947–59, 2010.
- [20] S. Chatterji, I. Yamazaki, Z. Bai, and J. Eisen. Compostbin: A DNA composition-based algorithm for binning environmental shotgun reads. *Research in Computational Molecular Biology, Proceedings*, 4955:17–28, 2008.
- [21] C. M. Chen, S. Y. Lee, and Z. H. Cho. Parallelization of the em algorithm for 3-d pet image reconstruction. *IEEE transactions on medical imaging*, 10(4):513–22, 1991.
- [22] Y. Chen, T. Souaiaia, and T. Chen. PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics*, 25(19):2514–21, 2009.
- [23] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, et al. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*, 2(10):2366–82, 2007.
- [24] P. Countway, P. Vigil, A. Schnetzer, S. Moorthi, and D. Caron. Seasonal analysis of protistan community structure and diversity at the USC Microbial Observatory (San Pedro Channel, North Pacific Ocean). *Limnol Oceanogr*, 55(6):2381–2396, 2010.

- [25] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561, 1970.
- [26] E. F. DeLong. Microbial community genomics in the ocean. *Nat Rev Microbiol*, 3(6):459–69, 2005.
- [27] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via EM algorithm. *J R Stat Soc B*, 39(1):1–38, 1977.
- [28] N. N. Diaz, L. Krause, A. Goesmann, K. Niehaus, and T. W. Nattkemper. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, 10:56, 2009.
- [29] E. A. Dinsdale, R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, et al. Functional metagenomic profiling of nine biomes. *Nature*, 452(7187):629–32, 2008.
- [30] B. Efron and D. V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65(3):457–482, 1978.
- [31] B. Efron and R. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall/CRC, Boca Raton ; London, 1998.
- [32] R. M. Engeman, R. T. Sugihara, L. F. Pank, and W. E. Dusenberry. A comparison of plotless density estimators using Monte Carlo simulation. *Ecology*, 75(6):1769–1779, 1994.
- [33] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391(6669):806–11, 1998.
- [34] M. M. Fisher and E. W. Triplett. Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol*, 65(10):4630–6, 1999.
- [35] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, et al. Whole-genome random sequencing and assembly of *haemophilus influenzae* rd. *Science*, 269(5223):496–512, 1995.
- [36] J. A. Fuhrman. Microbial community structure and its functional implications. *Nature*, 459(7244):193–9, 2009.
- [37] J. A. Gilbert, D. Field, P. Swift, L. Newbold, A. Oliver, T. Smyth, P. J. Somerfield, S. Huse, and I. Joint. The seasonal structure of microbial communities in the western english channel. *Environ Microbiol*, 11(12):3132–9, 2009.
- [38] J. A. Gilbert, J. A. Steele, J. G. Caporaso, L. Steinbruck, J. Reeder, B. Temperton, S. Huse, A. C. McHardy, R. Knight, I. Joint, et al. Defining seasonal marine microbial community dynamics. *ISME J*, 2011.

- [39] S. R. Gill, M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778):1355–9, 2006.
- [40] J. Goecks, A. Nekrutenko, and J. Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.
- [41] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*, 5(10):R245–9, 1998.
- [42] X. Hao, R. Jiang, and T. Chen. Clustering 16s rrna for otu prediction: a method of unsupervised bayesian clustering. *Bioinformatics*, 27(5):611–8, 2011.
- [43] P. He and L. Xia. Oligonucleotide profiling for discriminating bacteria in bacterial communities. *Combinatorial Chemistry & High Throughput Screening*, 10(4):247–255, 2007.
- [44] R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir. Structure of a ribonucleic acid. *Science*, 147:1462–5, 1965.
- [45] U. Hubscher, G. Maga, and S. Spadari. Eukaryotic DNA polymerases. *Annu Rev Biochem*, 71:133–63, 2002.
- [46] J. Hurwitz. The discovery of RNA polymerase. *J Biol Chem*, 280(52):42477–85, 2005.
- [47] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. MEGAN analysis of metagenomic data. *Genome Res*, 17(3):377–86, 2007.
- [48] H. F. Judson. *The eighth day of creation : makers of the revolution in biology*. Simon and Schuster, New York, 1979.
- [49] L. D. Kapp and J. R. Lorsch. The molecular mechanics of eukaryotic translation. *Annu Rev Biochem*, 73:657–704, 2004.
- [50] D. R. Kelley and S. L. Salzberg. Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics*, 11:544, 2010.
- [51] K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V. K. Sharma, T. P. Srivastava, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res*, 14(4):169–81, 2007.
- [52] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

- [53] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.
- [54] M. L. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A*, 97(18):9834–9, 2000.
- [55] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–8, 2008.
- [56] K. C. Li. Genome-wide coexpression dynamics: theory and application. *Proc Natl Acad Sci U S A*, 99(26):16875–80, 2002.
- [57] X. Li, S. Rao, W. Jiang, C. Li, Y. Xiao, Z. Guo, Q. Zhang, L. Wang, L. Du, J. Li, et al. Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. *BMC Bioinformatics*, 7:26, 2006.
- [58] K. Liolios, I. M. Chen, K. Mavromatis, N. Tavernarakis, P. Hugenholtz, V. M. Markowitz, and N. C. Kyrpides. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*, 38(Database issue):D346–54, 2010.
- [59] R. C. Littell, J. Pendergast, and R. Natarajan. Modelling covariance structure in the analysis of repeated measures data. *Stat Med*, 19(13):1793–819, 2000.
- [60] J. Liu. *Monte Carlo strategies in scientific computing*. Springer, New York, 2008.
- [61] V. M. Markowitz, I. M. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, I. Anderson, A. Lykidis, K. Mavromatis, et al. The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res*, 38(Database issue):D382–90, 2010.
- [62] P. A. Marquet, R. A. Quinones, S. Abades, F. Labra, M. Tognelli, M. Arim, and M. Rivadeneira. Scaling and power-laws in ecological systems. *J Exp Biol*, 208(Pt 9):1749–69, 2005.
- [63] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods*, 4(6):495–500, 2007.
- [64] A. C. McHardy, H. G. Martin, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods*, 4(1):63–72, 2007.
- [65] M. L. Metzker. Applications of next-generation sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46, 2010.

- [66] P. J. Mitchell and R. Tjian. Transcriptional regulation in mammalian cells by sequence-specific dna binding proteins. *Science*, 245(4916):371–8, 1989.
- [67] M. Monzoorul Haque, T. S. Ghosh, D. Komanduri, and S. S. Mande. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 25(14):1722–30, 2009.
- [68] J. L. Morgan, A. E. Darling, and J. A. Eisen. Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One*, 5(4):e10209, 2010.
- [69] K. B. Mullis. *Dancing naked in the mind field*. Pantheon Books, New York, 1st edition, 1998.
- [70] K. E. Nelson, G. M. Weinstock, S. K. Highlander, K. C. Worley, H. H. Creasy, J. R. Wortman, D. B. Rusch, M. Mitreva, E. Sodergren, A. T. Chinwalla, et al. A catalog of reference genomes from the human microbiome. *Science*, 328(5981):994–9, 2010.
- [71] T. T. Nguyen, R. R. Almon, D. C. DuBois, W. J. Jusko, and I. P. Androulakis. Importance of replication in analyzing time-series gene expression data: corticosteroid dynamics and circadian patterns in rat liver. *BMC Bioinformatics*, 11:279, 2010.
- [72] N. Pace, D. Stahl, D. Lane, and G. Olsen. The analysis of natural microbial populations by ribosomal RNA sequences. *Adv Microb Ecol*, 9:1–55, 1986.
- [73] G. E. Palade. A small particulate component of the cytoplasm. *J Biophys Biochem Cytol*, 1(1):59–68, 1955.
- [74] S. F. Paver and A. D. Kent. Temporal patterns in glycolate-utilizing bacterial community composition correlate with phytoplankton population dynamics in humic lakes. *Microb Ecol*, 60(2):406–18, 2010.
- [75] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. North. Structure of haemoglobin: a three-dimensional fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature*, 185(4711):416–22, 1960.
- [76] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, et al. The NIH human microbiome project. *Genome Res*, 19(12):2317–23, 2009.
- [77] E. Pettersson, J. Lundeberg, and A. Ahmadian. Generations of sequencing technologies. *Genomics*, 93(2):105–11, 2009.
- [78] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.
- [79] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, et al. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–9, 2000.

- [80] D. C. Richter, F. Ott, A. F. Auch, R. Schmid, and D. H. Huson. Metasim: a sequencing simulator for genomics and metagenomics. *PLoS One*, 3(10):e3373, 2008.
- [81] M. C. Rivera and J. A. Lake. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 431(7005):152–5, 2004.
- [82] G. L. Rosen, B. A. Sokhansanj, R. Polikar, M. A. Bruns, J. Russell, E. Garbarine, S. Essinger, and N. Yok. Signal processing for metagenomics: extracting information from the soup. *Curr Genomics*, 10(7):493–510, 2009.
- [83] Q. Ruan, D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman, and F. Sun. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics*, 22(20):2532–8, 2006.
- [84] Q. Ruan, J. A. Steele, M. S. Schwalbach, J. A. Fuhrman, and F. Sun. A dynamic programming algorithm for binning microbial community profiles. *Bioinformatics*, 22(12):1508–14, 2006.
- [85] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, et al. The Sorcerer II Global Ocean Sampling expedition: northwest atlantic through eastern tropical Pacific. *PLoS Biol*, 5(3):e77, 2007.
- [86] R. Sandberg, G. Winberg, C. I. Branden, A. Kaske, I. Ernberg, and J. Coster. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res*, 11(8):1404–9, 2001.
- [87] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–7, 1977.
- [88] A. Sauerwald, W. Zhu, T. A. Major, H. Roy, S. Palioura, D. Jahn, W. B. Whitman, r. Yates, J. R., M. Ibba, and D. Soll. Rna-dependent cysteine biosynthesis in archaea. *Science*, 307(5717):1969–72, 2005.
- [89] A. Shade, C. Y. Chiu, and K. D. McMahon. Differential bacterial dynamics promote emergent community robustness to lake mixing: an epilimnion to hypolimnion transplant experiment. *Environ Microbiol*, 12(2):455–66, 2010.
- [90] L. Simpson and J. Shaw. RNA editing and the mitochondrial cryptogenes of kinetoplastid protozoa. *Cell*, 57(3):355–66, 1989.
- [91] M. L. Sogin, H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A*, 103(32):12115–20, 2006.
- [92] J. A. Steele, P. D. Countway, L. Xia, P. D. Vigil, J. M. Beman, D. Y. Kim, C. E. Chow, R. Sachdeva, A. C. Jones, M. S. Schwalbach, et al. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J*, 2011.

- [93] R. Stepanauskas, M. A. Moran, B. A. Bergamaschi, and J. T. Hollibaugh. Covariance of bacterioplankton composition and environmental variables in a temperate delta system. *Aquat Microb Ecol*, 31(1):85–98, 2003.
- [94] R. Stepanauskas and M. E. Sieracki. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci U S A*, 104(21):9052–7, 2007.
- [95] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445, 2003.
- [96] Y. C. Tai and T. P. Speed. A multivariate empirical bayes statistic for replicated microarray time course data. *Ann Statist*, 34(5):2387–2412, 2006.
- [97] Y. C. Tai and T. P. Speed. On gene ranking using replicated microarray time course data. *Biometrics*, 65(1):40–51, 2009.
- [98] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glockner. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics*, 5:163, 2004.
- [99] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–5, 2010.
- [100] S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, et al. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–7, 2005.
- [101] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, et al. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–4, 2009.
- [102] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project. *Nature*, 449(7164):804–10, 2007.
- [103] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–31, 2006.
- [104] G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, 2004.
- [105] P. A. Vaishampayan, J. V. Kuehl, J. L. Froula, J. L. Morgan, H. Ochman, and M. P. Francino. Comparative metagenomics and population dynamics of the gut microbiota in mother and infant. *Genome Biol Evol*, 2:53–66, 2010.

- [106] B. A. S. Van Mooy, A. H. Devol, and R. G. Keil. Relationship between bacterial community structure, light, and carbon cycling in the eastern subarctic North Pacific. *Limnol Oceanogr*, 49(4):1056–1062, 2004.
- [107] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.
- [108] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004.
- [109] P. Vigil, P. Countway, J. Rose, D. Lonsdale, C. Gobler, and D. Caron. Rapid shifts in dominant taxa among microbial eukaryotes in estuarine ecosystems. *Aquat Microb Ecol*, 54(1):83–100, 2008.
- [110] G. Wang, L. Yin, Y. Zhao, and K. Mao. Efficiently mining time-delayed gene expression patterns. *IEEE Trans Syst Man Cybern B Cybern*, 40(2):400–11, 2010.
- [111] J. Wang and S. K. Kim. Global analysis of dauer gene expression in *Caenorhabditis elegans*. *Development*, 130(8):1621–34, 2003.
- [112] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, 2009.
- [113] F. Warnecke, P. Luginbuhl, N. Ivanova, M. Ghassemian, T. H. Richardson, J. T. Stege, M. Cayouette, A. C. McHardy, G. Djordjevic, N. Aboushadi, et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450(7169):560–5, 2007.
- [114] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953.
- [115] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, 74(11):5088–90, 1977.
- [116] T. Woyke, G. Xie, A. Copeland, J. M. Gonzalez, C. Han, H. Kiss, J. H. Saw, P. Senin, C. Yang, S. Chatterji, et al. Assembling the marine metagenome, one cell at a time. *PLoS One*, 4(4):e5299, 2009.
- [117] Y. W. Wu and Y. Ye. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J Comput Biol*, 18(3):523–34, 2011.
- [118] L. Xia and C. Zhou. Phase transition in sequence unique reconstruction. *Journal of Systems Science and Complexity*, 20(1):18–29, 2007.
- [119] L. C. Xia. Efficient statistical significance approximation for local association analysis of high-throughput time series data. Master’s thesis, University of Southern California, 2012.

- [120] L. C. Xia, D. Ai, J. A. Cram, J. A. Fuhrman, and F. Sun. Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics*, 29(2):230–237, 2012.
- [121] L. C. Xia, J. A. Cram, T. Chen, J. A. Fuhrman, and F. Sun. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One*, 6(12):p.e27992, 2011.
- [122] L. C. Xia, J. A. Steele, J. A. Cram, Z. G. Cardon, S. L. Simmons, J. J. Vallino, J. A. Fuhrman, and F. Sun. Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Systems Biology*, 5(Suppl 2):S15, 2011.
- [123] A. C. Yannarell and E. W. Triplett. Within- and between-lake variability in the composition of bacterioplankton communities: investigations using multiple spatial scales. *Appl Environ Microbiol*, 70(1):214–23, 2004.
- [124] A. C. Yannarell and E. W. Triplett. Geographic and environmental sources of variation in lake bacterial community composition. *Appl Environ Microbiol*, 71(1):227–39, 2005.
- [125] J. Yao, C. Chang, M. L. Salmi, Y. S. Hung, A. Loraine, and S. J. Roux. Genome-scale cluster analysis of replicated microarrays using shrinkage correlation coefficient. *BMC Bioinformatics*, 9:288, 2008.
- [126] S. Zhang, L. Y.J., L. Xia, and Q. Pan. Pplook: an automated data mining tool for protein-protein interaction. *BMC Bioinformatics*, 11(1):326, 2010.
- [127] D. Zhu, Y. Li, and H. Li. Multivariate correlation estimator for inferring functional relationships from replicated genome-wide data. *Bioinformatics*, 23(17):2298–305, 2007.

Appendix A

Technical derivations for GRAMMy

A.1 Derivation of the GRAMMy EM algorithm

Many estimation methods have been developed for estimating components' mixing parameters for finite mixture models, among which are the Expectation Maximization (EM) algorithm based approaches [27]. The EM based solutions have been proved to be accurate and robust in many cases. Many acceleration methods, like Aitken's, Quasi-Newton and Conjugated Gradient, exist to improve its convergence rate for large size problems. Thus, we adopted the EM based estimation as our solution to the MLE estimation in the transformed mixture problem. In the EM framework, we further assume a 'missing' data matrix \mathbf{Z} , in which each entry z_{ij} is a random variable indicating whether the read r_i is from the genome g_j . The model then can be solved by estimating π and \mathbf{Z} iteratively using Algorithm 2.

Algorithm 2 Genome Relative Abundance estimation by Finite Mixture Model(GRAMMY)

Require: read set \mathbf{R} , reference genomes \mathbf{G} , genome lengths \mathbf{L} as inputs.

Variables: missing indices \mathbf{Z} , reads probability \mathbf{f} , mixing parameters π .

if backend is ‘map’ **then**

 estimate \mathbf{f} by mapping procedures by Equation 2.5.

end if

if backend is ‘ k -mer’ **then**

 estimate \mathbf{f} by k -mer compositions by Equation 2.6.

end if

Mixing parameters $\pi \leftarrow Initialize()$ by moment estimates.

repeat

$\pi' \leftarrow \pi$

 E-step: $\mathbf{Z} \leftarrow Prob(\mathbf{Z}|\pi, \mathbf{R}, \mathbf{G})$ as in Equation 2.3.

 M-step: $\pi \leftarrow MLE(\pi|\mathbf{Z}, \mathbf{R}, \mathbf{G})$ as in Equation 2.4.

until π', π converged

Convert $(\pi_1, \pi_2, \dots, \pi_{m-1})$ to relative abundance \mathbf{a} by Equation 2.1.

return \mathbf{a}

We will describe the details of the algorithm in following subsections. Note: a variable with a superscript (t) stands for its value at the t -th iteration in EM, *e.g.* $\pi^{(t)}$ is the estimate of π at the t -th step. The t -th iteration in EM is:

- **E-step**

Assuming that mixing parameters $\pi^{(t)}$ are known, the ‘missing’ indicator entries in $\mathbf{Z}^{(t)}$ can be updated using their corresponding posterior probabilities or:

$$\begin{aligned}
z_{ij}^{(t)} &= p(z_{ij} = 1 | r_i; \pi^{(t)}, \mathbf{G}) \\
&= \frac{p(z_{ij} = 1, r_i | \pi^{(t)}, \mathbf{G})}{p(r_i | \pi^{(t)}, \mathbf{G})} \\
&= \frac{p(r_i | z_{ij} = 1; \pi^{(t)}, \mathbf{G}) p(z_{ij} = 1 | \pi^{(t)}, \mathbf{G})}{\sum_{k=1}^m p(r_i | z_{ik} = 1; \mathbf{a}^{(t)}, \mathbf{G}) p(z_{ik} = 1 | \pi^{(t)}, \mathbf{G})} \\
&= \frac{p(r_i | z_{ij} = 1; \mathbf{G}) \pi_j^{(t)}}{\sum_{k=1}^m p(r_i | z_{ik} = 1; \mathbf{G}) \pi_k^{(t)}}. \tag{A.1}
\end{aligned}$$

Notice that we used $p(r_i | z_{ij} = 1; \mathbf{G}, \pi^{(t)}) = p(r_i | z_{ij} = 1; \mathbf{G})$ because of the independence of the two sampling steps in our mixture model and that the read probability $p(r_i | z_{ij} = 1; \mathbf{G})$ can be accessed from $f_{g_j}(r_i | \mathbf{G})$, which is to be approximated using different methods later. Obviously, the update of $\mathbf{Z}^{(t)}$ depends solely on the updating value of $\pi^{(t)}$. This is how we obtain Equation 2.3 in Chapter 2.

- **M-step:**

Now, assuming ‘missing’ data $\mathbf{Z}^{(t)}$ are known, we calculate new mixing parameters $\pi^{(t+1)}$ that maximize the conditional expectation of the full data log likelihood function $Q(\pi | \pi^{(t)})$ of both the ‘missing’ and the known data, *i.e.*, we update them using:

$$\pi^{(t+1)} = \arg \max_{\pi} Q(\pi | \pi^{(t)}),$$

where

$$\begin{aligned}
Q(\pi|\pi^{(t)}) &= E(\log L(\mathbf{R}, \mathbf{Z}|\pi, \mathbf{G})|\mathbf{R}, \pi^{(t)}) \\
&= E(\log \prod_{i=1}^n \prod_{j=1}^m (p(z_{ij} = 1|\pi, \mathbf{G})p(r_i|z_{ij} = 1; \pi, \mathbf{G}))^{z_{ij}}|\mathbf{R}, \pi^{(t)}) \\
&= E(\sum_{i=1}^n \sum_{j=1}^m z_{ij}(\log p(z_{ij} = 1|\pi, \mathbf{G}) + \log p(r_i|z_{ij} = 1; \mathbf{G}))|\mathbf{R}, \pi^{(t)}) \\
&= \sum_{i=1}^n \sum_{j=1}^m p(z_{ij} = 1|\pi^{(t)}, \mathbf{G})(\log p(z_{ij} = 1|\pi, \mathbf{G}) + \log p(r_i|z_{ij} = 1; \mathbf{G})) \\
&= \sum_{i=1}^n \sum_{j=1}^m \pi_j^{(t)}(\log \pi_j + \log p(r_i|z_{ij} = 1; \mathbf{G})).
\end{aligned}$$

and

$$\log L(\mathbf{R}, \mathbf{Z}|\pi, \mathbf{G}) = \sum_{i=1}^n \sum_{j=1}^m z_{ij}(\log p(z_{ij} = 1|\pi, \mathbf{G}) + \log p(r_i|z_{ij} = 1; \mathbf{G}))$$

is the model log likelihood function for the complete data (\mathbf{Z}, \mathbf{R}) . The exact form of the maximum likelihood estimator (MLE) for $Q(\pi|\pi^{(t)})$ can be found, and it can be expressed using a simple closed form in $\pi^{(t+1)}$:

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^n z_{ij}^{(t)}}{n}. \tag{A.2}$$

This is how we obtain Equation 2.4 in Chapter 2.

When the MLE of π is found, using the one-to-one relation in Equation 2.1, the MLE of \mathbf{a} can be also found, thus we can solve the original biological problem.

A.2 Derivation of the standard errors

Using the asymptotic theory for MLE estimates, we can derive the asymptotic covariance matrix for the mixing parameters π . Remember, there are $m - 1$ independent parameters in π we are estimating and let us choose these to be $(\pi_1, \pi_2, \dots, \pi_{m-1})$ and denoted by $\hat{\pi}$. Let $\hat{\pi}^*$ and \mathbf{a}^* be the MLE estimates for $\hat{\pi}$ and its corresponding GRA vector. We can derive the observed information matrix \mathbf{I}_o ,

$$\mathbf{I}_o(\hat{\pi}|\mathbf{R}, \mathbf{G}) = -\frac{\partial^2 \log L(\mathbf{R}|\hat{\pi}, \mathbf{G})}{\partial \hat{\pi} \partial \hat{\pi}^T},$$

where:

$$L(\mathbf{R}|\hat{\pi}, \mathbf{G}) = \sum_{i=1}^n \log \left(\sum_{j=1}^{m-1} \pi_j f_{g_j}(r_i|\mathbf{G}) + (1 - \sum_{j=1}^{m-1} \pi_j) f_{g_m}(r_i|\mathbf{G}) \right)$$

is the log likelihood function of the observed data \mathbf{R} . Therefore, we write each entry of \mathbf{I}_o as:

$$\mathbf{I}_o(\hat{\pi}|\mathbf{R}, \mathbf{G})_{kl} = \sum_{i=1}^n \frac{(f_{g_k}(r_i|\mathbf{G}) - f_{g_m}(r_i|\mathbf{G}))(f_{g_l}(r_i|\mathbf{G}) - f_{g_m}(r_i|\mathbf{G}))}{(\sum_{j=1}^{m-1} \pi_j f_{g_j}(r_i|\mathbf{G}) + (1 - \sum_{j=1}^{m-1} \pi_j) f_{g_m}(r_i|\mathbf{G}))^2},$$

for $k, l \in \{1, 2, \dots, m - 1\}$. Because the GRA vector \mathbf{a} is a rank preserving transformation of $\hat{\pi}$, we can subsequently write the observed information matrix $\mathbf{I}_o(\mathbf{a}|\mathbf{R}, \mathbf{G})$ with regard to the parameterization of \mathbf{a} as:

$$\mathbf{I}_o(\mathbf{a}|\mathbf{R}, \mathbf{G}) = \nabla_{\mathbf{a}}(\hat{\pi})^T \mathbf{I}_o(\hat{\pi}|\mathbf{R}, \mathbf{G}) \nabla_{\mathbf{a}}(\hat{\pi}),$$

and the asymptotic standard error for our MLE estimate a_j^* as:

$$SE(a_j^*) = (\mathbf{Cov}(\mathbf{a}^*))_{jj} \approx ((\mathbf{I}_o^{-1}(\mathbf{a}|\mathbf{R}, \mathbf{G}))_{jj})^{\frac{1}{2}} \Big|_{\hat{\pi}=\hat{\pi}^*}, \quad (\text{A.3})$$

for $j \in \{1, 2, \dots, m-1\}$, considering $\hat{\pi}$ as the natural parameter set and \mathbf{a} as another parameter set, and that the asymptotic variance matrix can be effectively calculated by taking the inverse of the observed information matrix [30] and the standard error is the square root of variance entries on the diagonal. This is how we arrive at Equation 2.7 as our standard errors for our GRA estimates.

However, when the number of reads as compared to number of parameters is small or the majority of reads fails to be mapped, the asymptotic condition is not met and the application of previous result is not valid. However, we can still use the bootstrap estimator for covariance to estimate the standard error of our MLE using the empirical distribution:

$$SE(a_j^*) = (\mathbf{Cov}(\mathbf{a}^*))_{jj} \approx \left(\frac{1}{B-1} \sum_{b=1}^B (\mathbf{a}_{(b)}^* - \bar{\mathbf{a}}^*)(\mathbf{a}_{(b)}^* - \bar{\mathbf{a}}^*)^T \right)_{jj}, \quad (\text{A.4})$$

where $\bar{\mathbf{a}}^* = \frac{1}{B} \sum_{b=1}^B \mathbf{a}_{(b)}^*$ is the bootstrap mean estimator of the samples' MLEs, which is exactly our Equation 2.8 in Chapter 2.

A.3 Convergence of the GRAMMy EM algorithm

Because the EM method is greedy, it may not converge to the global maximum of the objective function. However, in this case, we shall show the observed data log likelihood

function $L(\mathbf{R}|\hat{\pi}, \mathbf{G})$ is concave with regard to $\hat{\pi}$. Thus, any local maximum the EM converge to is the global maximum.

Proposition 1. $L(\mathbf{R}|\hat{\pi}, \mathbf{G})$ is concave.

Proof. Since the sum of concave functions is still concave, proving the concavity of the log likelihood function of single observation suffices. Taking the second-order derivatives of the summands of $L(\mathbf{R}|\hat{\pi}, \mathbf{G})$, we have

$$\frac{\partial^2 \log L(r_i|\hat{\pi}, \mathbf{G})}{\partial \hat{\pi} \partial \hat{\pi}^T} = -\frac{(f_{g_k}(r_i|\mathbf{G}) - f_{g_m}(r_i|\mathbf{G}))(f_{g_l}(r_i|\mathbf{G}) - f_{g_m}(r_i|\mathbf{G}))}{(\sum_{j=1}^{m-1} \pi_j f_{g_j}(r_i|\mathbf{G}) + (1 - \sum_{j=1}^{m-1} \pi_j) f_{g_m}(r_i|\mathbf{G}))^2}. \quad (\text{A.5})$$

If consider the Hessian matrix \mathbf{H} where the (k, l) -th element is Equation A.5, we can write \mathbf{H} as $\mathbf{H} = -d\mathbf{v}^t\mathbf{v}$, where

$$\mathbf{v} = (f_{g_1}(r_i|\mathbf{G}) - f_{g_m}(r_i|\mathbf{G}), \dots, f_{g_{m-1}}(r_i|\mathbf{G}) - f_{g_m}(r_i|\mathbf{G}))$$

is a vector and

$$d = \frac{1}{(\sum_{j=1}^{m-1} \pi_j f_{g_j}(r_i|\mathbf{G}) + (1 - \sum_{j=1}^{m-1} \pi_j) f_{g_m}(r_i|\mathbf{G}))^2}$$

is a scalar. Notice $d \geq 0$, therefore \mathbf{H} is negative semi-definite because for any vector $\mathbf{u} = (u_1, \dots, u_{m-1})$, we have $\mathbf{u}\mathbf{H}\mathbf{u}^t = -d(\mathbf{u}\mathbf{v}^t)(\mathbf{u}\mathbf{v}^t)^t = -d(\mathbf{u}\mathbf{v}^t)^2 \leq 0$. Thus, the concavity of the log likelihood function $L(\mathbf{R}|\hat{\pi}, \mathbf{G})$ is proved. \square

Appendix B

GRAMMy simulated studies

B.1 Simulated read sets

To evaluate the performance of the GRAMMy, we generated a series of simulated read sets using MetaSim [80], which is a tool specialized to simulate large shotgun metagenomic read sets from input reference genomes and has full-fledged simulating options, such as sequencing error models, population variations and read length distributions.

In our simulation study, we randomly chose ten microbial genomes from the collection of genomes given by the FAMeS study [63]. We then generated an artificial GRA vector from the power-law (*Zipf's*) distribution [62]: $f(k; \alpha, N) = \frac{1/k^\alpha}{\sum_{n=1}^N 1/n^\alpha}$ with $\alpha = 2$.

Both the reference genomes and the vector of relative abundance were provided to MetaSim, with its population sampling option on, to generate a series of read sets, with RL (read length) in $\{50, 100, 200, 400, 800\}$ bp, RN (read number) in $\{1000, 2000, 5000, 10000, 20000, 50000, 100000\}$, and SE (sequencing error mode) in either ‘with’ and ‘without sequencing errors’. For each parameter triplet (RL, RN, SE) , we generated ten replicates.

To simulate the ‘with sequencing errors’ scenario, the sequencing errors were introduced into read sets by enabling the ‘454’ or ‘Sanger’ error mode option in MetaSim to mimic the reads generating behavior of the Roche/454 ($RL = 50\text{-}400$ bp) and Sanger platforms ($RL = 800\text{bp}$). The read length distribution option was also on to generate reads with normally distributed lengths for the two platforms. These options were conservative because MetaSim was originally published in 2008 and the technologies have been greatly improved since then.

While generating the read replicates, we also permuted the order of all the components in the GRA vector so that every genome had the chance to become either a major or a minor member in the read sets. This permutation procedure reduces the artifact introduced by manually choosing genomes and their abundance levels, as a consequence, the robustness of estimation could be assessed by measuring the standard deviations of all replicates’ estimates. The series of replicated simulated read sets obtained above was then extended with additional non-replicated read sets with RN in $\{200000, 500000, 1000000\}$ and RL, SE the same as above for larger scale benchmarks.

To evaluate the estimation with different community structures, we randomly generated another GRA vector from the same power law distribution with larger variations in component abundances. We then repeated the above read generation process for all parameter triplets (RL, RN, SE) using this new GRA vector without replicates. This produced a new independent series of read sets with significant differences in microbial community structure from the previous one. We labeled the new series ‘steep’ since its GRA only had a few dominant species and the previous one ‘flat’ since its GRA was more evenly distributed.

B.2 Performance evaluation for simulations

We first used the same set of genomes used in read generation as our reference genomes. The alignment program BLAT was used to align the reads to the references and the output was fed into GRAMMy, GAAS and MEGAN. Then, we used different numerical error measures (see “Materials and Methods” in Chapter 2) and their standard deviations to assess the quality of the estimations.

In Figure B.1, we plotted the measured errors (with deviation bars) against the read number (RN) to show the convergence of the GRA estimates to their true values. It can be seen from Figure B.1A that, as RN increases, the Relative Root Mean Square Error (RRMSE) diminishes to almost zero with decreased variation for all RLs , which indicates, regardless of read lengths, the GRAMMy (‘map’) accurately converge to their true values and become stable once the read number ensures a high coverage. For instance, when 10^5 reads are available, the RRMSE is less than 2% and its standard deviation is marginal for all RLs .

In Figure B.1B, in addition to RRMSE, we measured the Average Relative Error (AVGRE), the Maximum Relative Error (MAXRE), the Distance of Total Variation (DTV) and their standard deviations for the read sets with a RL equal to 100 bp. According to the plot, all four measures converge to zero and stabilize. This pattern is similar using other read lengths. From Figures B.1A and B.1B, we concluded that the GRAMMy estimation is accurate and robust for different read lengths and error measures.

To further study the performance of GRAMMy within the limitations of partially available reference genomes and current sequencing technologies, we next added more

perturbations to the simulation study, such as sequencing errors, unknown genomes. We also applied a different abundance distribution to evaluate the effects from the complexity of a community. The results from these studies were summarized in a series of RRMSE-versus- RN plots in Figure B.2.

As we can see from Figure B.2A, sequencing errors do affect the estimation accuracy for short reads since the estimation accuracy for read sets ‘with sequencing errors’ is lower than that for ‘without sequencing errors’, particularly at $RLs \leq 200$ bp. However, for a reasonably large number of reads, a scale routinely achieved in recent metagenomic read sets, the estimates are close to the true values, as in the worst case here, the limiting RRMSE is about 20% for the shortest read length ($RL=50$ bp). We can also infer from the plot that, developments from sequencing technologies, such as increased read length and reduced error rates, can help to improve the estimates. For example, at RN equal to 10^5 and ‘with sequencing errors’, when the RL is increased from 50 to 200 bp, it helps to reduce the RRMSEs from 20% to 10% approximately. Moreover, when sequencing errors are negligible, 50 bp reads are as informative as any longer ones in the purpose of abundance estimation using our framework.

In reality, inaccuracies in the GRA estimation can also arise from the limited knowledge of reference genomes. In the next simulation, we masked out 50% of the reference genomes and repeated the estimations. As Figure B.2B indicates, a partial reference genome set does not substantially affect the accuracy of estimates, despite that they become less robust at a low sequencing depth. In fact, at a sufficient high coverage (RN

equal to 10^6), the estimates for read sets ‘with unknowns’ also converge and is comparably accurate to that of ‘without unknowns’. Even if 80% of the reference genomes were masked out, the estimation still had good convergence, as our study indicates.

Another factor that may affect the estimation is the community’s natural complexity. To study this, we prepared two communities which are different from each other in their shape of GRA distribution. In these read sets, the GRA of the ‘flat’ sets is more spread among all genomes while that of the ‘steep’ sets is more concentrated on a few genomes. From the estimations, as shown in Figure B.2C, we do not observe significant effects resulting from different complexities, though there are some decrease in accuracy for the ‘steep’ sets, which may be related to a less coverage of minority genomes.

We also compared GRAMMy to other methods. With the objective of estimating the GRA of communities, we first benchmarked GRAMMy with GAAS. In addition, we included MEGAN, which produces a read profile that summarizes the number of reads assigned to their lowest common ancestors (LCA). We estimated the GRA based on MEGAN using the normalized percentages from the reads distributed on leaf taxon. The default options of GAAS and MEGAN were used in our study. Figure B.3A shows the results from the simulation read sets with read length (RL) equal to 100 or 400 bp generated from MetaSim using the with sequencing errors option. We see that GRAMMy (‘map’) significantly outperformed GAAS, MEGAN and GRAMMy (‘ k -mer’) in all settings. Among all the methods tested, GRAMMy (‘map’) is the only method with RRMSEs decreasing to zero as the number of reads increases.

In addition to the above methods, We compared the 16S-based, *rpoB*-based and BLAT hit counting estimates to GRAMMy estimates using our simulated read set. Figure B.3B

shows that GRAMMy outperformed all other methods in this controlled setting. All other methods show three obvious drawbacks: a persisting bias, significant variation and a strong dependence on the number of reads.

Finally, we evaluated the computation time and the error propagation to higher taxonomic levels using our simulated data set. The time and space complexity of our algorithm are shown to be $O(c_1c_2n)$ and $O(c_1n)$, respectively, where n is the size of the read set, c_1 (related to associated genomes each read) and c_2 (related to EM convergence criteria) are two constants.

We benchmarked GRAMMy with MEGAN and GAAS for running time with different *RLs* and *RNs*, see Figure 6. The mapping or alignment time is excluded for all compared tools. We see GRAMMy is consistently faster than the other two in processing the same read set and it scales as expected. In addition, as shown in Figure B.4, the errors gradually reduce from lower to higher taxonomic levels. And the error is consistently small when the *RN* is large. All the simulations are carried out on our “Dell, PE1950, Xeon E5420, 2.5GHz, 12010MB RAM” computing nodes.

In conclusion, our simulations showed GRAMMy estimates are accurate and stable across a range of anticipated settings. Furthermore, it is superior in speed as compared to other available tools. An interesting observation is, when the purpose is to estimate the abundance of a predefined set of reference genomes, an excessively ‘deep sequencing’ scheme is not necessary. As shown in the subfigures of Figure B.1 -B.3, the RRMSEs start to stabilize when the *RN* passes over 10^4 reads, which indicates there may be a threshold for read number that is needed to recover the community abundance structure. This trend also represents that, when the reads ambiguity are properly handled, a read set

of relatively smaller number can still provide substantial information for the abundance estimation. Even though the specific threshold value may differ in real settings, it can be predicted using pre-study simulations and is informative for a more economical design of the actual sequencing depth.

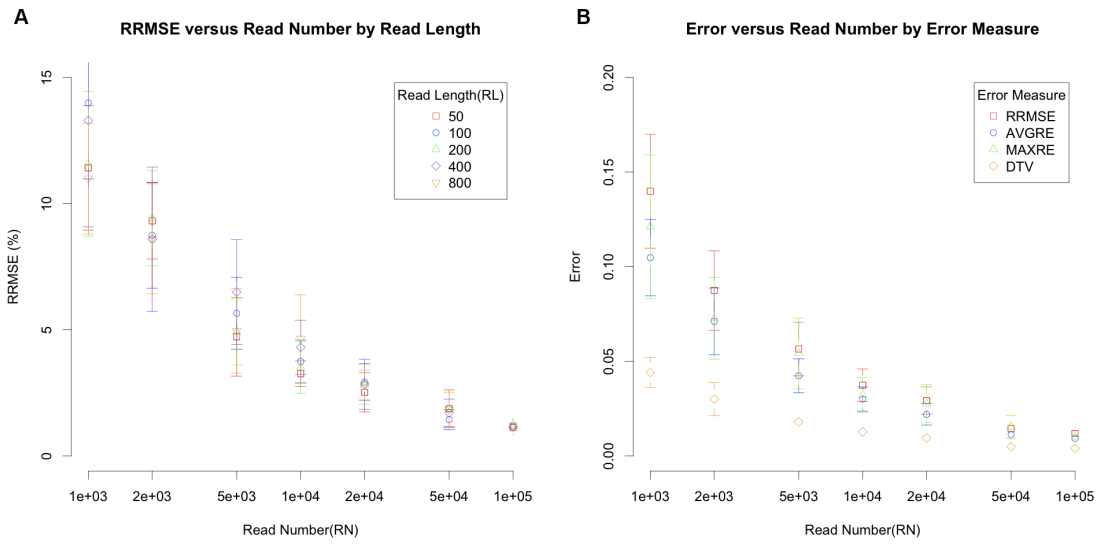


Figure B.1: The convergence of GRAMMy. The estimation errors, as measured by different numerical methods: (A) Relative Root Mean Square Error (RRMSE) in percentage versus Read Number (RN) for different read lengths (RL). (B) Relative Root Mean Square Error (RRMSE), Average Relative Error (AVGRE), Maximum Relative Error (MAXRE), and Distance of Total Variation (DTV) versus Read Number for read length equal 100 bp. GRAMMy (map) was used.

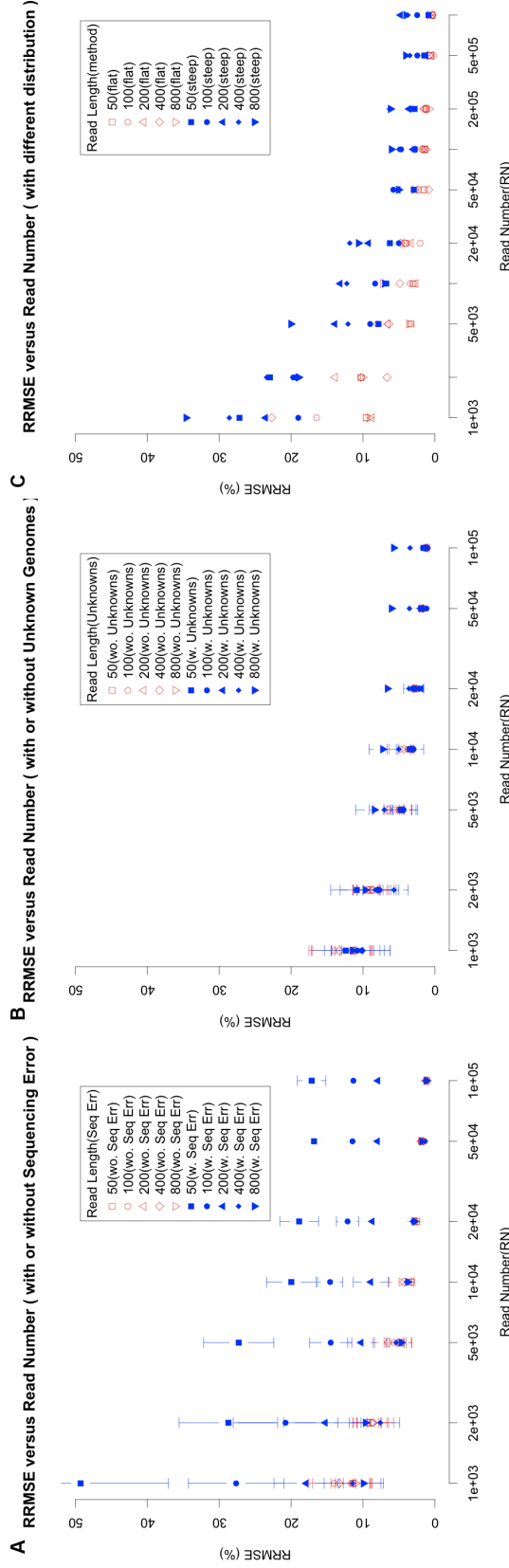


Figure B.2: Simulated read set benchmarks. Effects of different perturbations on GRAMMy's estimation: (A) Effects of sequencing errors: results from 'with sequencing error' and 'without sequencing error' read sets are labeled as 'w. Seq Err' and 'wo. Seq Err', respectively. (B) Effects of unknown genomes: results from estimation 'with unknown genomes' and 'without unknown genomes' read sets are labeled as 'w. Unknowns' and 'wo. Unknowns', respectively. (C) Effects of different genome relative abundance distributions: results from more concentrated abundance distribution and less concentrated read sets are labeled as 'steep' and 'flat', respectively. Relative Root Mean Square Error (RRMSE) as a percentage is plotted against Read Number. GRAMMy ('map') was used.

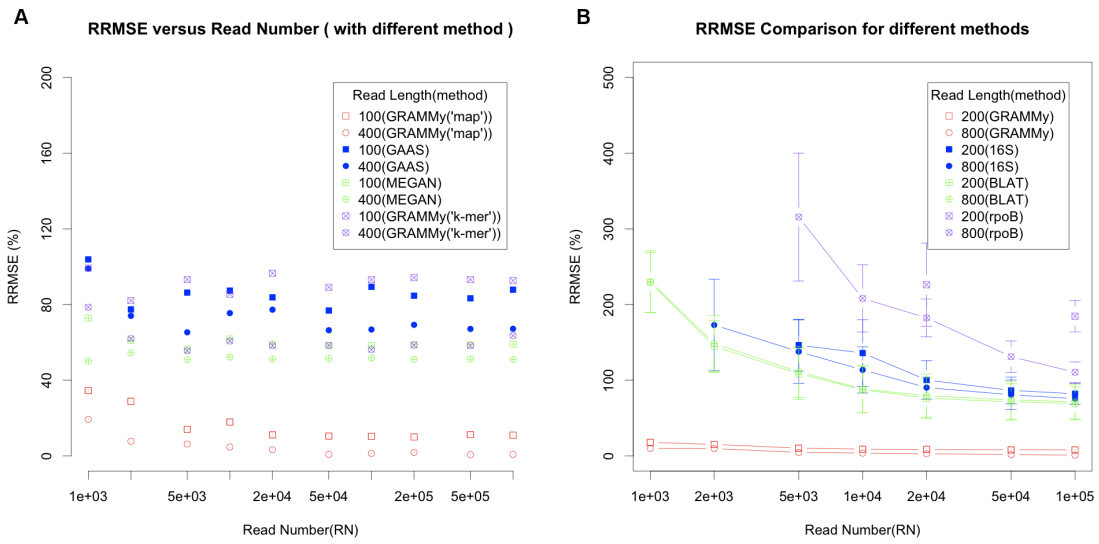


Figure B.3: Performance comparison of different methods. The performance comparisons for different estimation methods: (A) MEGAN-based ('MEGAN'), GAAS ('GAAS') and GRAMMy ('map' and '*k*-mer') on simulated read sets with sequencing errors at read length 100 bp and 400 bp. (B) 16S-based ('16S'), BLAT hit counting ('BLAT'), *rpoB*-based ('*rpoB*') and GRAMMy ('map'). Relative Root Mean Square Error (RRMSE) as a percentage is plotted against Read Number (*RN*).

AVGRE vs taxonomic level for GRAMMy

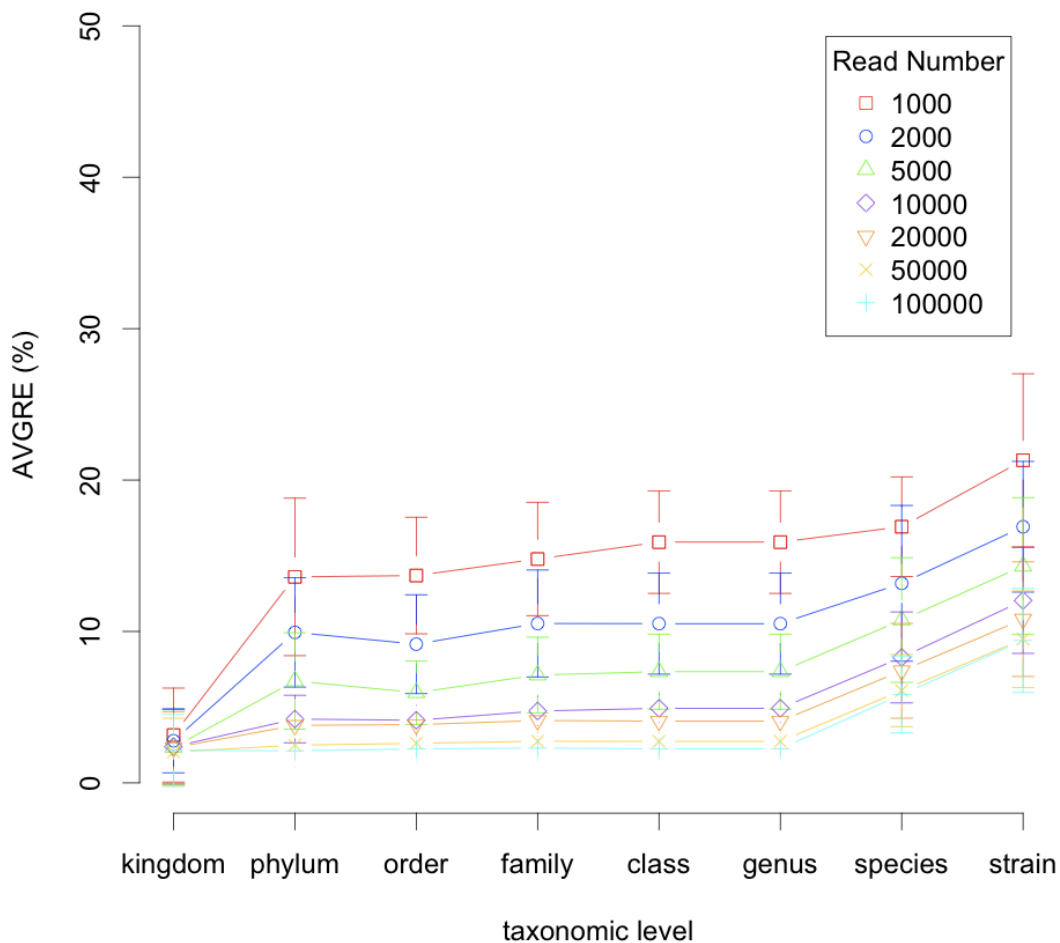


Figure B.4: Estimation errors at different taxonomic levels. Average Relative Error (AVGRE) as a percentage is plotted against taxonomic level. The errors gradually decrease from strains to kingdom taxonomic levels.