University of California
Santa Barbara

Studying chemical and biological systems using high-throughput sequencing: analytical challenges and solutions

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Chemistry

by

Yuning Shen

Committee in charge:

Professor Irene Chen, Chair
Professor Ambuj Singh
Professor Joan-Emma Shea
Professor Mattanjah de Vries

March 2022

The Dissertation of Yuning Shen is approved.

Professor Ambuj Singh

Professor Joan-Emma Shea

Professor Mattanjah de Vries

Professor Irene Chen, Committee Chair

February 2022

Studying chemical and biological systems using high-throughput sequencing: analytical challenges and solutions

by

Yuning Shen

To my parents, they taught me to dream and walk.

Acknowledgements

First and foremost, I am deeply grateful to my advisor, Professor Irene A. Chen, for her guidance on my Ph.D. study. She opened the door for me when I had little background in biochemistry, microbiology, and computation. Without her patience and support, I could not build the knowledge and skills to explore these research areas. As a scientist and a scholar, she is an inspiring model with her broad knowledge and sharp insights. I can not express enough my thanks to you, Irene. I also want to thank Professor Ambuj K. Singh for his guidance on my research and on developing the mind and skills for computational research. Without his help, it would be impossible for me, an experimentalist who could barely code, to start my research in computation. His encouragement gave me the confidence to push my limit in my research and career.

My dissertation work would not be possible without my collaborators: Sam Verbanic, Abe Pressman, Evan Janzen, Nate Charest, Professor Yei-Chen Lai, Professor Juhee Lee, and Dr. John Deacon; I am honored to work with you and thank you! I must thank Professors Joan-Emma Shea and Professor Mattanjah de Vries for being on my committee and keeping my research in check. I want to thank all other members of Chen Lab and Dynamo Lab: Jen Mobberley, Celia Blanco, Huan Peng, Josh Kenchel, Ray Borg, Ranajay Saha, Alberto Vazquez-Salazar, Greg Campbell, Haraldur Hallgrimsson, Hongyuan You, Arlei Silva, Omid Askarischani, Sourav Medya, and many others. Thank you for being great mentors, colleagues, and friends.

Finally, I would like to express my gratitude to my family, Yaqi, and all my friends for their unwavering support and belief in me. They gave me courage and strength when I stumbled. It is never an easy road in graduate school and towards a career in science. I need to thank all the challenges and failures - I appreciate that I tried, failed, occasionally succeeded, and am still moving forward.

Curriculum Vitæ

Yuning Shen

Education

2016-2022	Ph.D. in Chemistry (Expected), University of California, Santa Barbara.
2018-2021	M.Sc. in Computer Science, University of California, Santa Barbara.
2012-2016	B.Sc. in Chemistry, Fudan University, Shanghai, China.

Teaching Experience

2021	CM145/CM245 (Molecular Biotechnology for Engineers) Teaching Assistant
2016	CHEM 1AL/1BL/1CL (General Chemistry Lab) Teaching Assistant

Publications

- **Shen, Y.**, Pressman, A., Janzen, E., & Chen, I. A. (2020) Kinetic sequencing (k-Seq) as a massively parallel assay for ribozyme kinetics: utility and critical parameters. *Nucleic Acids Research*, 49(12), e67–e67.
- Verbanic, S., **Shen, Y.**, Lee, J., Deacon, J. M., & Chen, I. A. (2020) Microbial predictors of healing and short-term effect of debridement on the microbiome of chronic wounds: the role of facultative anaerobes. *npj Bioflms & Microbiomes*, 6(1), 21.
- Ci, T., **Shen, Y.**, Cui, S., Liu, R., Yu, L., & Ding, J. (2017) Achieving high drug loading and sustained release of hydrophobic drugs in hydrogels through *in situ* crystallization. *Macromolecular Bioscience*, 17 (3).

Manuscripts submitted

- Janzen, E., **Shen, Y.**, Liu, Z., Blanco, C., Sutherland, J. D.,& Chen, I. A. (2021) Error minimization and specificity could emerge in a genetic code as by-products of prebiotic evolution. *Nature Communications*; *in revision.*
- Zhang, S., **Shen, Y.**, Chen, I. A., Lee, J. (2021) Bayesian Modeling of Interaction between Features in Sparse Multivariate Count Data with Application to Microbiome Study. *The Annals of Applied Statistics.*

Manuscripts in preparation

- **Shen, Y.**, Verbanic, S., Lee, J., Singh, A. K., & Chen, I. A., Correlations structures between bacteria and bacteriophages in the skin microbiome for patients with chronic wounds.
- Charest, N., **Shen, Y.**, Lai, Y., Chen, I. A., & Shea, J.-E., Discovering pathways through ribozyme fitness landscapes using information theoretic quantification of epistasis.

Abstract

Studying chemical and biological systems using high-throughput sequencing: analytical challenges and solutions

by

Yuning Shen

High-throughput sequencing (HTS) can identify unique DNA sequences and quantify their abundances from mixed DNA pools. HTS-based assays can profile complex biological or chemical systems with entities that can convert to unique DNA sequences. Computational models are also developed to analyze these HTS data at a larger scale. However, such data contain unique analytical challenges, including discrete counts, relative measurement, and small sample size. Careful assessments of these computational tools are required for robust interpretations of results.

In this dissertation, we investigated the computational challenges, proposed and assess the solutions for two applications of HTS-based assays. In the first work, we proposed k-Seq, a kinetic assay to measure the activity of self-aminoacylation ribozymes (catalytic RNA). Characterizing the kinetics for different molecules in a heterogeneous pool is challenging as their abundance and activities can vary in several orders of magnitude. We explored different designs of experiments and identified critical factors affecting the estimation of kinetic coefficients in the pseudo-first order kinetic model for these ribozymes. Using bootstrapping, we robustly quantified the uncertainty of estimation for individual sequences and determined the minimum sequencing counts required for reliable estimations. Combining the improved experimental design and new analytical tools, we robustly quantified the kinetics for 10^5 different ribozymes.

In the second work, we constructed the correlation networks between microorgan-

isms from metagenomic data and studied the structure of a human skin microbiome in patients with chronic wounds. We designed a variation of Gaussian graphical models to capture the direct correlations between the abundances of bacteria and viruses while accounting for the structure and limitations in the data. To minimize the discovery of false correlations from the small noisy dataset, we applied a two-step model selection to regularize the results. Lastly, we demonstrated the utility of the constructed correlation network in recovering the strong correlations between microbes, identifying potentially important microbes, and microbial clusters.

Contents

Chapter 1

Introduction

1.1 Background and motivations

Many chemical and biological systems are complex with interdependent components. These components can be interacting organisms and form a complex ecological system. For example, a microbial community contains different bacteria, fungi, and viruses that cohabitate in the same space and communicate through chemical and biological processes [1]. These components are closely related and need to be characterized simultaneously for a direct and comprehensive understanding of such systems. The components can also be a series of building blocks for biomolecules, such as a ribonucleic acid (RNA) molecule that contains sequential combination of four types of nucleotides (adenine, uracil, cytosine, and guanine). Studying the sequence-function relations is the first step to understand these molecules but requires characterizing a decent coverage of sequences in the astronomically large configuration space (e.g., a length 21 RNA can have $4^{21} \approx 4 \times 10^{12}$ possible sequences). In both examples, experimental studies need some frameworks for high-throughput analyses in order to understand the structure and the functions for these complex systems with different components or configurations.

Conducting high-throughput analyses for a complex system need a comprehensive consideration from the following three perspectives: the characterization techniques, the research design, and the computational methods for data analysis. The characterization techniques should quantify the property of interest (e.g. abundance) for a large number of entities in the system without the need to isolate each entity. Despite being high-throughput, the actual coverage of a measurement, that is the set of entities whose properties can be reliably determined, depends on the limit of detection, the accuracy of selected technique, and also the research design. The research design is critical to set the scope (the breadth and the depth) of the study given the frequently encountered trade-offs between the number of samples and the depth of measurement on each sample. If the original high-throughput characterization is only relative, one might design additional experiments for standardization or normalization. Moreover, complex systems are inherently heterogeneous that different entities can vary in magnitudes on the abundance or other properties and consequently, the quality of measurement for different entities varies. When applicable, an optimized research design can mitigate this effect and improve the measurements. Lastly, computational methods are important to analyze the high-throughput data from the measurements. A desired computational method must be efficient in handling the large amount of data from the high-throughput experiments. It should also account for aforementioned properties (e.g., heterogeneity and different measurement quality) of such study and other potential factors or constraints from the characterization techniques and the research design. And most importantly, all computational methods need to be evaluated with caution when applying to a specific characterization technique and research design.

As an emerging research area, researchers have been developing the experimental and computational methods for high-throughput characterization and meanwhile a deeper understanding is needed on the performance of those computational methods when ana-

lyzing data from a real-world experiments. In my Ph.D. study, I focused on developing and optimizing computational methods for real-world research problems encountered in lab. Specifically, we studied two applications of high-throughput sequencing (HTS), a high-throughput method to identify and quantify DNA molecules, on characterizing a chemical (ribozyme or catalytic RNA) and a biological system (microbiome). By evaluating current methods, we aimed to determine the important experimental factors affecting the computations and optimize the both experimental design and computational methods for improved characterization. During this process, we also aimed to understand the practical utilities and limitations of these computational methods given the research design and measurement noise from real-world datasets.

1.2 High-throughput sequencing (HTS) as quantitative assays

DNA sequencing is a method to determine the order of nucleotides (i.e., the sequence) for DNA molecules. Sanger sequencing is the first-generation DNA sequencing method developed in 1975 [2] and has led to several milestones in genomics, including the first draft of the human genome [3]. However, in Sanger sequencing, each DNA sequence needs to be separated for sequencing and the throughput is low even with 96-capillary systems [4]. The second-generation of DNA sequencing methods or called next-generation sequencing (NGS) were developed in 2000s, including Roche/454 pyrosequencing, Illumina sequencing, LifeTechnologies Ion Torrent etc. Starting from this generation, DNA sequencing methods are attributed as high-throughput sequencing (HTS) methods as they can generate reads at a large scale [5]. For example, the Illumina HiSeq 4000 can generate 1.5 Tb or 5 billion sequence reads in a single run with a cost of a few thousand

US dollars. The major limitations for the methods in this generation is the short reads (usually 100 $\sim$ 500 nucleotides, nt), and relative higher error rate ($10^{-2} \sim 10^{-4}$) [4], but these disadvantages can also be mitigated through read assembly and error correction in bioinformatic steps. The substantial improvement on the throughput and the affordability has expanded the applications of DNA sequencing in genomics and to many other research areas [4, 5, 6, 7, 8]. The third-generation of DNA sequencing is also being developed and commercialized. The DNA sequencing methods in this generation can produce extra long reads ($>$ 80 kilobase pairs, kb) and sequence on single molecules [9]. While the third-generation methods are gaining more attentions, the second-generation methods such as Illumina sequencing is still the mainstream in current research.

With the sufficient sequencing depth from the high-throughput, HTS not only can identify the content of sequences but also can use as a "sequence counter" to quantitate or semi-quantitate the abundance of different DNA sequences in samples. Entities (organisms or biomolecules) that can be converted to DNA molecules and prepared as DNA libraries can potentially be quantified using generic HTS platforms. Such versatility let HTS become one of the most widely applied techniques in many research areas and various HTS-assays have been developed for high-throughput characterization of biological/chemical systems. In metagenomics, 16s ribosomal RNA (rRNA, for bacteria/archea) or Internal Transcribed Spacer (ITS, for fungi) can survey specific regions of genomes from multiple microorganisms in the environment to identify and quantify their relative abundance using HTS [6, 7, 10]. In transcriptomics, RNA-seq has been developed to quantify the gene expression levels in different samples or conditions where the messenger RNA (mRNA) can be purified and converted to complementary DNA (cDNA) library for sequencing through reverse transcription [8, 11]. For *in vitro* studies in biochemistry, HTS-based assays have also been developed to quantify the abundance change of biomolecules such as enzymes, ribozymes (catalytic RNA), apatmers in binding

or reaction to quantify their biochemical activities in massive parallel [12, 13, 14, 15, 16].

1.3 General procedure for HTS-based assays

The general procedure for HTS-based assays contains following steps: 1) sample collection and DNA library preparation; 2) DNA sequencing; 3) quantify the abundance of entities from sequences; before passing to downstream analyses.

1.3.1 Sample collection and DNA library preparation

First, samples containing the entities of interest (e.g., biomolecules or organisms) are collected from environments or experiments, following the research design and the protocols for the subjects. The target type of entities need to be enriched in each sample and potential contaminants such as environmental DNA/RNA need to be removed. If the entities of interest are organisms, the target DNA or RNA molecules in cells or capsids also need to be extracted.

Next, a DNA library can be prepared from extracted DNA or RNA molecules in each sample. For RNA molecules, they often require to convert to complementary DNA molecules using reverse transcription. The DNA libraries can be prepared using either targeted or untargeted method. Targeted method can be applied to the entities of interest that are encoded by sequences with universal constant sequence regions and unique variable regions for identification, such as the 16s ribosomal RNA (rRNA) gene in bacteria/archea [10], Internal Transcribed Spacer (ITS) region in fungi [7], or constant-variable regions in some artificial biomolecule pools [17]. These constant-variable structure can be identified by specific sequencing primers (i.e., a short sequence complementary to the constant region) for targeted amplification using the polymerase chain reaction (PCR) to prepare DNA library. Targeted methods are often preferred as only the region of interest

is amplified, further removing the effects from potential contaminants. In contrast, when the entities do not contain such "fingerprint" regions to target on or additional information is desired (e.g., genes encoding proteins in bacterial genome), untargeted method can be used to prepare the DNA library. In metagenomics, Whole Genome Sequencing (WGS) can be applied to sequence the entire genomes of microorganisms by fragmenting the genome and prepare the DNA library using all the fragments.

After the DNA libraries are prepared for each sample, sample indices and adapters required for DNA sequencing are added to DNA molecules. These are short DNA fragments assist to distinguish DNA molecules from different samples and enable the binding and synthesis of DNA strains during the sequencing-by-synthesis step (see below). The total amount of library DNA can be measured using quantitative PCR (qPCR), fluorometry (e.g., QuBit) or other quantification methods [18]. DNA libraries from different samples can then be normalized by balancing their total amount and pooled together for multiplex DNA sequencing.

While the actual protocol depends on the research subjects, there are some common factors need to consider during sample collection and library preparation. A biased collection process might alter the composition of the entities in the system, making the sample less representative. The library preparation might also introduce biases through PCR amplification [19]. The sampling and sequencing depth is another important factor when surveying the system. Both insufficient sample collection and unbalanced DNA libraries when mixing the samples could lead to missing information for some important entities with lower abundance. Lastly, some systems might contain different categories of entities (e.g., bacteria vs. viruses) within the interest of the study and requires different sampling or preparation protocols.

1.3.2 Sequencing-by-synthesis using Illumina

One of the most commonly used HTS platforms is from Illumina, Inc, using sequencing-by-synthesis. Briefly, multiplexed DNA library is loaded to the flow cell, a glass slide with lanes, where each lane is coated by two types of short DNA oligos complementary to the adapters added to the sample DNA during library preparation. The loaded sample DNA hybridize with one of the oligos on the surface and a new sequence complementary to the sample DNA is synthesized from the oligo fragment. After removing the original sequence by denaturing, a repeated clonal amplification is performed using the surrounding oligos as primers, and billions of sequences are each amplified into clusters of cloned sequences. The reverse strains are then cleaved leaving the forward strains for sequencing. Sequencing starts when the first sequencing primers binds to the forward strains. In each cycle of sequencing-by-synthesis, four different types of nucleotides tagged with different colors of fluorescence are competing for the extension site and the one complementary to the template site is ligated. Each clonal cluster emits a characteristic color signals to be captured by a camera when the nucleotide is added. This sequencing-by-synthesis cycle is repeated until the sequencing for the entire forward reads is finished. All forwards reads are then cleaved leaving the reverse reads for a similar sequencing-by-synthesis process to determine the reverse read. Both forward and reverse reads are captured by the digital camera as sequential color signals at each location for a cluster on the flow cell, and the digital reads for the forward and reverse strains are generated by the sequencer.

1.3.3 Quantification using sequencing reads

Once the raw sequencing reads are obtained from the sequencer, some common steps for read preprocessing. Multiplexed reads are separated by their sample indices into separated files. Quality control steps are conducted to remove low quality reads or failed

in sequencing (e.g., too short/long). For paired-end sequencing, the forward and the reverse reads from each cluster are joined by finding an optimal overlapped matching region. The auxiliary indices, adapters, and primer regions are trimmed to recover the original sequences of DNA molecules. And additional QC steps can be added to finalize the read preprocessing step by removing those were failed in joining or trimming process.

The entities in the samples are identified based on the sequencing reads and the abundance of entities are quantified using the copy number of reads (or counts) in each sample. The strategy to identify the entities depends on the treatment of DNA sequencing errors, the type of entities, and area of research. The simplest strategy is to treat each unique sequence as an entity. This strategy might be suitable when the single-sequence resolution is desired (e.g., the actual sequences are close in the sequence space with $1 \sim 2$ nucleotides difference) and the sequencing error is either ignored or corrected [20, 21, 22]. For some other systems, small difference in the sequences is tolerated and reads with similar but slightly different sequences are all considered from a same entity. For example, 16s rRNA reads with 97% similarity are traditionally considered as from bacteria with same genus [23]. Such similar sequences can be grouped by classification (comparing to reference sequence databases) or clustering (without reference databases). Such methods are also called close- and open- Operational Taxonomic Unit (OTU) picking respectively in the context of metagenomics [24,25]. Once these entities are identified, their abundance can be quantified by directly counting the number of reads or by statistical models to estimate the coverage for each entity [26].

1.4 Analytical challenges in HTS-based assays

HTS-based assays are powerful tools to quantitatively profile complex chemical or biological systems that have changed the research landscapes in many domains. However,

using HTS for quantitation also introduces some unique challenges in data analysis that are usually not encountered in classic low-throughput measurements.

First, HTS-assays generate discrete counts rather than continuous signals compared to classic characterization methods. The relative abundance of entities are represented by the number of reads or counts detected in each sample. While the large counts can usually be approximated as continuous signals, the treatment for small counts and zeros is tricky. These small counts could be more susceptible to the measurement error and difficult to interpret (e.g., does zero mean missing or lower than the limit-of-detection). These small counts also do not follow the normal distribution that were frequently assumed in many common statistical tools. While one option is to filter out the entities associated with low or zero counts and focus on the more abundant entities, it potentially "waste" a large amount of sequencing resources and decreases the throughput of the method. Moreover, it is difficult to determine a proper threshold to consider a count to be "too small" and sometimes arbitrary [13]. More importantly, for some studies, smaller counts or excessive zeros reflect the sample heterogeneity under different conditions and is an inherent property of the system that should not be ignored. More comprehensive models have been proposed to explicitly account for such "zero-inflated" dataset [27, 28].

Second, the count data from HTS only represent the relative abundance of entities in each sample. For many studies, arbitrary amounts of samples are taken and different samples are further normalized to ensure the sequencing resources are allocated evenly. Thus, the HTS can not quantify the absolute amount of entities in the original environment but only the relative abundance represented by the number of reads. Lacking the knowledge of absolute abundance might cause difficulties in statistical analyses and interpreting the results. For example, a higher relative abundance does not necessarily mean a higher absolute abundance. The compositionality of HTS data needs to be resolved experimentally or computationally. Some systems like *in vitro* experiments con-

tains the entire population of entities in test tubes and thus the total amount of each sample can be measured experimentally. For example, the concentrations for DNA or RNA entities can be quantified using QuBit or qPCR [17,18] and the total amount can be obtained by multiplying the volume. Alternatively, a known amount of external standard or "spike-in" sequence can be added to each sample for calibration [17,29]. The sequence for the standard should be distant from the sequences for the entities of interest and the standard sequence goes through the same experimental process for quantification using HTS-assays. The abundance for entities of interests can be calculated from the known amount of the standard and the ratio between the relative abundance of target sequence and the sequence for the standard. However, these experimental methods might not always be available. For example, the total bacteria amount in samples collected from a human subject can not be directly quantified and applying external standards to human is restricted due to ethical reasons. In these cases, researchers might need to resort to special computational tools when analyzing compositional data. For example, compositional data has one less degree of freedom due to the constraints of fractions sum up to 1, and spurious correlations inferred between these fractions is problematic. A theoretical framework to treat compositional data in correlation analysis has been proposed to resolve such problem by using log transformations or latent variables [30, 31].

Lastly, small sample size, noisy data, and large data dimensions are also common challenges for HTS-based assays. Despite millions to billions of sequence reads can be generated from HTS, a large number of reads is required for each sample in order to cover the diverse range of entities with sufficient sequencing depth (i.e., average counts). Along with the labor and cost for sample collection and preparation, the number of samples in HTS-based assays is usually small. At the same time, the research questions for complex systems may contain a large number of variables (e.g., the abundance for entities) and the data from HTS for analysis is high-dimensional. Samples collected from

some complex systems in environment are often noisy. All these factors contribute to the difficulty in designing a HTS study and analyzing the data for meaningful interpretations. In practice, there are trade-offs between the sequencing depth of the data, the number of available samples, the scope of a study, and the complexity of the problem. With deeper sequencing for each sample, more counts can be obtained for each entity (i.e. less noisy from HTS), more lower-abundant entities can be potentially quantified, and more information could possible be obtained for a study with broader scope. But at the same time, less samples or conditions can be obtained given a total budget, limiting the power of analyses for such complex problems. Researchers need to balance these aspects and could use simulation to optimize for the design of the study. From another perspective, robust computational models are especially important for HTS studies which provide information about the uncertainty of estimation and focus on controlling the noise (false positive signals) during analyses.

1.5 Dissertation Overview

In this dissertation, I present my work on developing, evaluating, and improving the computational methods in two research topics using HTS-based assays: 1) measuring the kinetics for ribozymes using a massively parallel assay and 2) quantifying the correlations between microbes in human microbiome from metagenomic assays. While focusing on computational models, I particularly studied how computational models can be optimized for specific studies and, if applicable, use computational results to guide the research design. In each main part of my dissertation, I introduce the research backgrounds, computational models, and major challenges for each study. Combining the analyses on the real-world data and simulated data, I show the practical utility and limitations for these computational methods. And lastly, I discuss the problems solve by

current computational methods and those still remains challenging that requires future development.

In **Chapter 2**, I present an application of HTS in biochemistry to measure the reaction kinetics for a large number of catalytic RNA molecules (ribozymes) in parallel. I particularly focus on using an in-house developed method called k-Seq to quantify the kinetics for self-aminoacylating ribozymes using pseudo-first-order kinetics. The main challenge comes from the heterogeneity of the RNA pool and the difference in measurement quality for different ribozymes using HTS. As a solution, I show how we use bootstrapping, an resampling technique, to estimate the sampling noise for individual sequence and quantify the uncertainty and model identifiability for individual sequences. Compared to a classic method using point estimation and standard deviation from triplicated experiments, this method results in more accurate and robust estimations and more informational reports for the kinetic coefficients of each sequence. This work also determines some key design factors affecting the accuracy and coverage of HTS-based assays similar to k-Seq, providing some guidance on designing such studies.

In **Chapter 3**, I describe a work on quantifying correlations between bacteria and viruses in skin microbiome from HTS metagenomic data and studying the relations between microbes using inferred correlation network. Different from the paralleled assay in **Chapter 2**, microbiome is a complex community with interacting entities. Due to such complexity and the limitations from metagenomic pipelines, observed data are noisy and constrained (relative rather than absolute in abundance). To identify the strongest correlation signals for robust interpretations, a variation of Gaussian graphical model (GGM) called `CLR-cGGM` is used to account for above limitations in HTS data. A stringent two-step model selection procedure is applied to further regularize the inference results and minimize false correlations. This work also demonstrates the utility of correlation network to study the relations between different microbes and potentially provide some

plausible hypotheses for further studies on the structure and mechanisms of microbiome.

1.6 Permissions and Attributions

1. The content of **Chapter 2** and **Appendix A.1** is the result of a collaboration with Dr. Abe Pressman, Dr. Evan Janzen, and Prof. Irene Chen, and has previously appeared in the *Nucleic Acids Research* [12]. Dr. Abe Pressman conducted the experiments and collected the data for Method section 2.5.1 and 2.5.2. It is reproduced here with the permission of *Nucleic Acids Research*: `https://doi.org/10.1093/nar/gkab199`.

2. The content of **Chapter 3 and Appendix A.2** is the result of a collaboration with Dr. Samuel Verbanic, Prof. Ambuj Singh, Prof. Juhee Lee, and Prof. Irene Chen. Dr. Samuel Verbanic conducted the experiments, collected the data, and performed the bioinformatics for Method section 3.5.1 and curated the list of viral contaminants for Method section 3.5.2. It is presented here with the permission of the collaborators.

Chapter 2

k-Seq for robust characterizations of reaction kinetics for a heterogeneous RNA pool

2.1 Background

2.1.1 Measuring the chemical activities for biomolecules using HTS

Measuring the chemical activity of each sequence in a collection of biomolecules is important for determining the genotype-phenotype relationship and discovering novel functional sequences. Ideally, methods to accomplish this would: a) yield accurate activity measurements for individual sequences, b) achieve high throughput, such as through parallelization, to cover a large number of variants in sequence space, and c) be adaptable to different ribozymes (and deoxyribozymes). High-throughput sequencing (HTS) provides an opportunity to address these goals. The large amount of sequence data should allow high accuracy of count data for many sequences in a high throughput, parallelized format. In principle, as long as reacted and unreacted molecules can be separated from each other and prepared for the sequencing, HTS could be used to quantify the extent of reaction for multiple time points, substrate concentrations, or other experimental variables. Using sequencing as the assay would avoid the need to isolate and test each unique sequence.

HTS-based kinetic measurements have been proposed and demonstrated with nucleic acids, including catalytic DNA [13], catalytic RNA [14, 17, 32, 33], substrate RNA [15], RNA aptamers [34], and transcription factor (TF) binding DNA [35]. In these works, approximately $10^3 \sim 10^6$ unique sequences are measured. Similar approaches have also been developed for proteins, notably an assay of ligand binding affinities through mRNA display [36], 'deep mutational scanning' [16], in which the phenotype of fitness is assayed for many mutants by deep sequencing, and a large-scale measurement of dose-response curves [37].

Here we focus on kinetic sequencing (*k*-Seq), a model-based method recently reported for quantification of kinetics in a mixed pool of sequences [17]. Compared to methods conducting relative measurements [14,15] or requiring specialized instruments for *in situ* reactions [33,34], *k*-Seq quantifies the absolute values for kinetic coefficients using experiments in bulk solutions. A general schema of *k*-Seq is described as follows (Figure 2.1): an input pool is designed containing sequences of interest (e.g., candidate ribozymes). Aliquots of the input pool are reacted under different experimental conditions, such as different substrate concentrations or different time points. Then, reacted and unreacted molecules are separated. Each pool is converted to a DNA library and prepared for sequencing. Absolute measurement of reacted (or unreacted) quantities is also needed for normalization. Reads generated from HTS are subjected to quality control and de-replicated to generate a 'count' table of copies of each sequence detected in a sample. Count data are normalized to absolute abundance and fit to the appropriate kinetic model to estimate the rate constants and other coefficients of interest.

2.1.2 Challenges in HTS-based parallel assays

The development of *k*-Seq and related assays also raises questions about multiple issues potentially limit the practical applicability. Kinetic measurements require properly chosen experimental conditions (e.g. substrate concentrations or time points) for sufficient dynamic range. For a heterogeneous pool where sequences would prefer different optimal conditions, the conditions chosen will compromise for some sequences. For example, in a two time point (unreacted, reacted) experiment determining enzyme kinetics over 4096 RNA substrates by varying reaction time [15], authors had to choose the reaction time optimal for either highly active RNA substrates or less active ones. Previous work determining kinetic coefficients for ribozymes by *k*-Seq also showed the limited

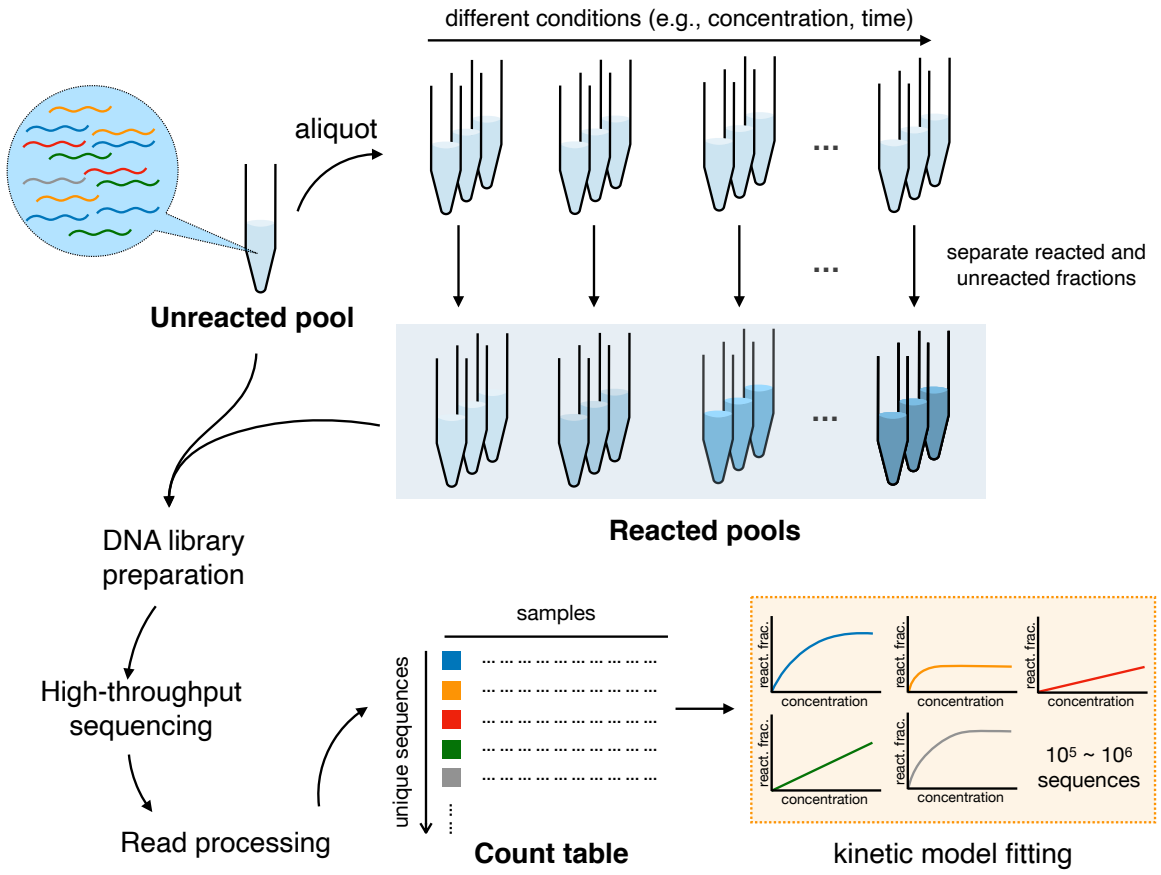


Figure 2.1: General scheme of *k*-Seq experiment and analysis. A heterogeneous input pool containing nucleic acids is reacted at different experimental conditions (e.g., different substrate concentrations or different reaction time). Reacted and unreacted molecules are separated and either (or both) of these fractions is prepared for high--throughput sequencing. The reads from DNA sequencing are processed to obtain a count table for each unique sequence across samples, normalized by a standard, and abundances across samples are fit into a kinetic model to estimate parameters (e.g., rate constants). react. frac. = reacted fraction.

characterization of less active ribozymes where the kinetic coefficients (rate constant k and maximum amplitude A) could not be independently estimated due to model identifiability problems [17]. While these problems could potentially be solved by expanding the number of experimental conditions, with HTS this solution becomes prohibitively expensive in time and resources. Therefore, it is important to rigorously understand how the choice of conditions would affect the estimation of kinetic coefficients and trustworthiness of kinetic measurements on each sequence.

Another consideration unique to HTS-based kinetic measurements derives from the sequencing errors. Sequencing error might misidentify a molecule as a nearby sequence variant, and subsequently change the quantification of both the true and incorrect sequences. This could be particularly problematic when one sequence is present in high abundance relative to others, and thus creates a relatively large number of misidentified reads that confounds quantitation of related sequences. This problem cannot be solved by increasing the number of replicates since the number of erroneous reads rises in proportion to the number of reads (a systematic bias rather than random noise).

A final concern is assessing the accuracy and precision of k-Seq measurements. Using discrete count data (number of reads of a particular sequence) as the approximation of a sequence's relative abundance in the sample introduces some complexity in assessing measurement accuracy, particularly at low counts where large stochastic variation exists. Earlier works limited the library size to thousands of sequences for high coverage of sequences [13, 15]. With the necessary extension to larger libraries ($> 10^5$ unique sequences), issues of high error rates associated with sequences with low counts are commonly reported [32, 33, 36]. While one may simply exclude counts lower than a cutoff value, it is not obvious how to choose such cutoffs, or estimate the uncertainty (e.g. confidence intervals) on fitted parameter values, given the experimental scenario of a low number of replicates but a high total number of counts.

To date, there is a relative lack of critical discussion addressing the theoretical and experimental effects of such variables on the outcome of high-throughput measurements. The purpose of the current work is to examine these issues and develop appropriate methodology to address them.

2.1.3 *k*-Seq on aminoacylation ribozymes

For each ribozyme sequence in a mixed pool, following reactions occur during the aminoacylation with substrate BYO (Biotinyl-Tyr(Me)-Oxazolone) [17]:

$$\text{Ribozyme}_i + \text{BYO} \xrightarrow{k_i} \text{BYO}{-}\text{Ribozyme}_i \tag{2.1}$$

$$\text{BYO} \xrightarrow{\text{degrade}} \text{BYO}^* \tag{2.2}$$

When BYO is present in vast excess compared to ribozymes, Reaction 2.1 can be approximated by the first-order kinetics. Due the degradation of BYO during the experiments (Reaction 2.2), we fixed the reaction time as 90 min and the concentration change of BYO due to degradation can be adjusted using a constant factor $\alpha = 0.479$, as measured in [17]. With these approximations, we apply a pseudo-first order kinetics with varying substrate concentrations to model the reaction:

$$f_{ij} = A_i(1 - e^{-\alpha t k_i c_j}) \tag{2.3}$$

f_{ij} is the reacted fraction for sequence i in sample j with initial BYO concentration c_j, $t = 90$ is the reaction time, and α is the degradation factor. By fitting the model to experimental data, we estimate two parameters for sequence i: A_i is the maximum amplitude of the reaction occurred and k_i is the rate constant.

2.1.4 Overview

In this work, we study and refine the HTS-based kinetic assay called k-Seq. Coupled with theoretical and simulation studies, we systematically characterize model identifiability, accuracy, and precision of kinetic estimation for ribozymes using HTS data. We discuss key factors when optimizing experimental design for k-Seq type experiments. With these knowledge, we demonstrate an k-Seq experiment on with improved experimental design and analytics on a mixed pool of variants based on ribozymes previously isolated from *in vitro* selection [17] and characterized by the pseudo-first-order kinetics.

2.2 Results

2.2.1 Model identifiability depends on kinetic coefficients, experimental conditions, and measurement error

To understand the factors affecting model identifiability and optimize conditions for experiments, we explored the effects of ribozyme activities (kinetic coefficients), experimental conditions, and measurement errors using the simulated reacted fraction dataset. We first evaluated model identifiability qualitatively. We selected sequences from 6 regions in the parameter space of $\log_{10} k \in [-1, 3]$, $\log_{10} A \in [-2, 0]$, and $kA > 0.1$ $\text{min}^{-1}\text{M}^{-1}$ (Figure A.1). For each sequence, we simulated the reacted fractions under a series of substrate concentrations and curve fits from repeated fitting or bootstrapping, given various measurement error levels (ϵ). As summarized in Table A.1, by examining the the distribution of fitted k_i and A_i (see Figure A.2 - A.7for examples), sequences with higher k, A values and lower ϵ are more likely to be separable.

To quantify the separability for each individual sequences, we proposed three metrics: ΔA or the range of A across repeated fittings (without resampling); σ_A, or the standard

deviation of A from bootstrapped samples; and γ, a measure of how noisy the separate estimation of k and A is compared to estimating the combined parameter kA. ΔA was able to identify sequences with numerically unstable fitting results which have small k, A values, but was not able to identify sequences whose fitting optima are sensitive to noise in the data. Almost all the sequences from the 6 selected regions each converged to a uniform optimum in repeated fitting, and the convergence was insensitive to noise. In contrast, bootstrapping results account for noise in the data and because the optima from fitting re-sampled data points were not always converged, this provides more comprehensive separability information than fitting convergence. By examining the distribution of each metric value for sequences in each region, both σ_A and γ reflected the trend observed in different regions: higher metric value corresponded to less separable parameters of a sequence (Figure A.8). In practice, we found γ aligned slightly better with human intuitions in evaluating the parameter separability in the experimental data of the mixed variant pool.

Using metric γ, we assessed the effects of experimental conditions and measurement error on parameter separability (Figure 2.2). Parameter separability depends on the true k and A values, choice of substrate concentrations, and the level of measurement error. Controlling the experimental design and measurement error, k and A were more separable for sequences with higher k and A value. Comparing the sequences along the kA = constant line, parameter separability appears to be more dependent on k than A, especially for lower measurement error cases (e.g. $\epsilon \leq 0.5$). To assess the effect of experimental conditions, we compared the case of adding one more replicate to each reaction (4 vs. 3 in the first implementation of k-Seq [17]) to adding a higher concentration of BYO (1250 μM) while maintaining the triplicates. As expected, despite having more samples, adding another replicate did not change the region of separability. However, adding a higher concentration of substrate shifted the boundary of separability

(right side of light color region) on kA values by a factor of ~ 10, effectively increasing the dynamic range (dark color region right to the light color region) of the k-Seq assay. Additionally, as shown in Figure 2.2B, the difficulty of separating k and A increased when the measurements were more noisy. Sequences that were separable at low measurement error (e.g. $kA \sim 10\ \text{min}^{-1}\text{M}^{-1}$ and $\epsilon < 0.2$) became not separable when the measurement error was large (e.g. $\epsilon = 1.0$).

Despite extending the range of BYO substrate might improve the model identifiability for sequences with lower k and A and estimating k and A separately is of general interests for kinetic measurements, achieving high substrate concentrations is experimentally challenging. Within the viable concentration range (Extended Substrate Range), we found the models for most of measured sequences were not identifiable (Figure A.8). For the purpose of analyzing accuracy and uncertainty for k-Seq over a wide range of k values (analyses below), we focus on the estimation of the combined parameter kA.

2.2.2 *k*-Seq variant pool read processing and quality controls

We conducted the k-Seq experiment on a multiplexed sample containing mixed pools of variants of ribozymes S-1A.1-a, S-1B.1-a, S-2.1-a., and S-3.1-a (Table 2.1), using the expanded experimental conditions evaluated above (BYO ranges from 2 to 1250 μM). A known amount of the 'spike-in' sequence was added to each reaction to aid absolute quantitation. After demultiplexing the reads, we obtained 39,151,684 paired-end reads in the unreacted sample and a mean of 13,057,929.1 (SD = 4,359,249.2) paired-end reads in reacted samples (Figure A.10A). Around 90 - 92 % of the reads were successfully joined in each sample (Figure A.10BC). Dereplication, removal of reads not having length = 21, and removal of the spike-in sequence reads (sequences within 2 edit distance to the spike-in sequence) yielded a count table of the number of reads for each unique sequence

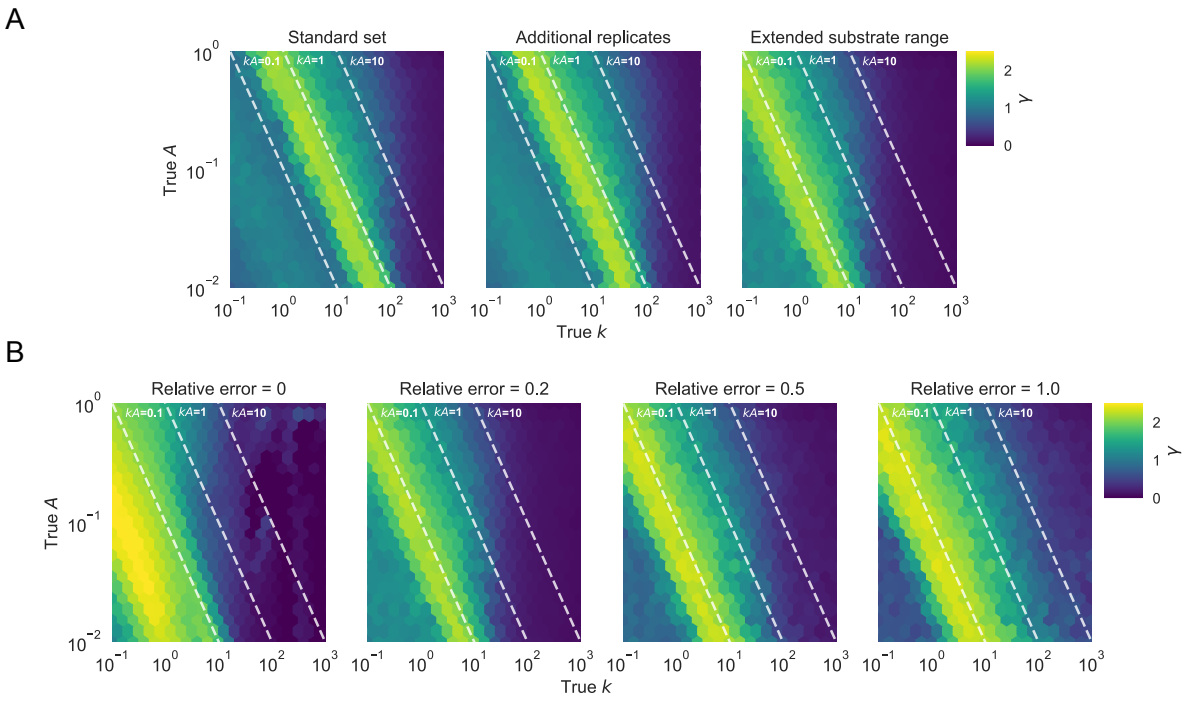


Figure 2.2: Effect of experimental factors on model identifiability to separately estimate k and A. Identifiability was evaluated using metric γ, based on the simulated effects of (A) choice of BYO samples (with relative error $\epsilon = 0.2$) and (B) relative error (using the BYO series of the extended substrate range). Reacted fractions for 10,201 (101^2) simulated sequences with true k, A in the parameter space shown in the figure were fit to the pseudo-first order model, and γ values for each sequence were calculated from 100 bootstrapped samples. Higher values of γ indicate that k and A are less separable. (A) Choosing a wider range of BYO concentration is more effective in improving the region of identifiable data compared to adding more replicates of the same BYO concentrations. (B) With higher measurement error, k and A become increasingly difficult to estimate separately.

detected in each sample. On average 87.9 % (SD = 1.1 %) of total reads were preserved in the samples (Figure A.10BC).

In principle, in order to calculate the reacted fraction with at least one non-zero value for fitting, a sequence must be detected in the unreacted sample (denominator) and in at least one of the reacted samples (numerator). Using this initial criterion, 764,756 valid unique sequences were considered to be analyzable for least-squares fitting to a pseudo-first order kinetic model, which comprised 77.9 % of total reads in the unreacted sample and an average of 87.7 % (SD = 0.6 %) of total reads among reacted samples (Figure A.10BC).

Ribozyme	Sequence (variant region)
S-2.1-a	ATTACCCTGGTCATCGAGTGA
S-1A.1-a	CTACTTCAAACAATCGGTCTG
S-1B.1-a	CCACACTTCAAGCAATCGGTC
S-3.1-a	AAGTTTGCTAATAGTCGCAAG

Table 2.1: Four wild-type sequences selected from [17] for the variant pool

2.2.3 Distribution of ribozyme mutants in the variant pool

For each wild-type ribozyme (S-1A.1-a, S-1B.1-a, S-2.1-a, S-3.1-a), a variant pool was chemically synthesized such that each position had the wild-type identity with 91% probability and with non-wild-type residues being equally probable (3% each). In theory, each variant pool contained roughly 14% wild-type sequences ($d = 0$) and 0.45% of each single mutant ($d = 1$), or a ratio of 0.033 for the abundance of each single mutant to the wild-type (see Text A.1 for calculation). We sequenced the unreacted pool and categorized each sequence read to four ribozyme families (1A.1, 1B.1, 2.1, 3.1) by the Hamming distance to the nearest wild-type. Sequencing results confirmed that the four variant pools followed the design (Figure 2.3A). The mixed variant pools contained at

least $\sim$ 1,000 reads per sequence for $d = 0$,1, and 2 (up to double mutants; Figure 2.3A, Table 1), and a mean of 39.7 reads per sequence for $d = 3$ (triple mutants). The unreacted pool showed good coverage of analyzable sequences for d= 0 to 3 (Table 2.2).

Sequencing error can erroneously assign reads of a ribozyme to related sequences that changes the apparent counts of sequences and potentially confounds the estimation for kinetic coefficients in k-Seq. It is critical to evaluate such effect when single nucleotide resolution is required. There are particularly two distinct effects arise from sequencing error. First, the number of reads observed for a given sequence is lower than the true number, due to erroneous reads that are assigned to other sequences. With constant error rate, this effect is expected to reduce counts proportionally across all sequences and normalized abundances should not impacted. Second, however, the number of reads observed for a given sequence will also be inflated by the contribution of erroneous reads arising from related sequences. It is a particularly acute problem for uneven pools where a small number of sequences (e.g., wild-type sequences) are highly represented and less abundance or less active sequences are closely related to them (e.g., resulting in estimation of parameters being biased toward those of the abundant sequence). The combination of two effects can changed the observed abundance for sequences. We calculated the expected fraction of reads for a sequence resulting from errors of its single-mutant neighbors in a variant pool, at different sequencing error rates (Figure 2.3B, Text A.2). In our variant pool with a 9% mutation rate, a sequencing error rate of 1% could cause more than 10% of reads for a mutant ($d \geq 1$) to be the result of sequencing error from its neighbors (Figure A.11). On the other hand, the most abundant sequences in the doped pool, wild-type sequences ($d = 0$) were least affected by this sequencing error effect. This problem can be mitigated by controlling the error rate. If the sequence length is small enough to be covered by paired-end sequencing, requiring absolute matching of the overlapped region between the paired-end reads of a single sequence during joining

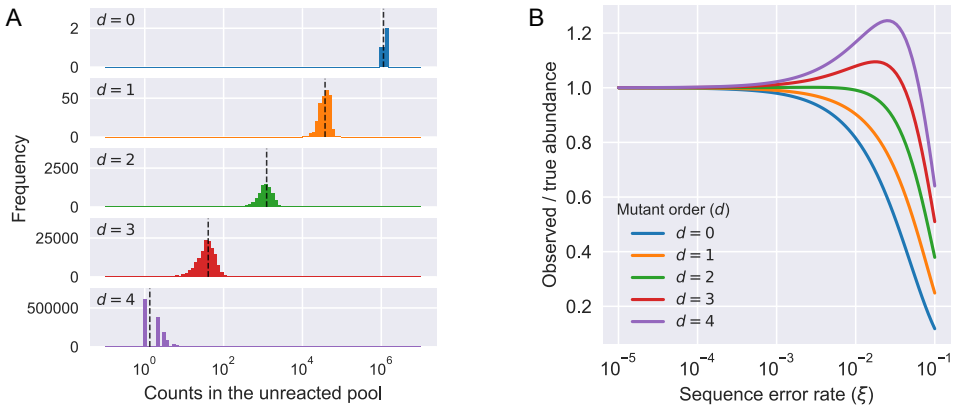


Figure 2.3: Distribution of mutants in the pool and the effect of sequencing error. (A) Relative abundance (counts) of sequences in the unreacted pool (four ribozyme families, total number of reads = 32,931,917), categorized by Hamming distance to its nearest family center. Observed abundance of different classes was similar to the expected number of counts (black dashed line). (B) The effect of different levels of sequencing error (ξ) to the expected observed abundance as the ratio to the true abundance for mutants with different orders (d) in a variant pool with 9% mutation rate. Due to the mixed effects of losing counts from being misidentified to a neighboring sequence and gaining counts from the misidentification of a neighboring sequence, the observed abundance for a sequence would either decrease ($d = 0, 1$) or first increase then decrease ($d = 2, 3, 4$) as the sequencing error increases. See Text A.2 for calculation details.

should result in a squared error rate (e.g., from 1% to 0.01%). In the scenario here, this reduces the fraction of spurious reads from neighboring sequences to be $< 0.5\%$ for up to quadruple mutants ($d \leq 4$) without significant loss of reads during joining (Figure A.11).

In the variant pool design, the peak centers are not only highly abundant but also highly reactive due to their selection from the previous study [17]. Consistent with this, the relative abundances of the peak centers increased in reacted pools compared to the input pool. This asymmetry affected the ability to assay low abundance sequences (e.g., triple mutants or sequences from Family 3.1), as their relative abundance decreased substantially in the reacted pools (Figure A.12).

Order of mutants (d)	# of unique sequences	# of analyzable sequences	Fraction of analyzable sequences	Mean counts in the unreacted pool (SD)	Expected counts in the unreacted pool
0	4	4	1.000	1,243,500 (151,356)	1,136,125
1	252	252	1.000	37,599.9 (10,607.1)	37,455
2	7,560	7,560	1.000	1,198.3 (454.8)	1,234
3	143,640	143,482	0.999	39.7 (18.9)	40.7
4	1,939,140	590,115	0.304	2.1 (1.3)	1.34
≥ 5	N/A	23,343	N/A	1.1 (1.3)	N/A

Table 2.2: Coverage of local sequence space in the variant pool containing four ribozyme families. N/A = not applicable. The calculation of expected counts in the unreacted pool does not include effects of sequencing error.

2.2.4 Quantify the total amount of sequences

While the relative abundance of each sequence in a particular reaction sample can be calculated by dividing read counts by the total reads in each sample, calculation of the reacted fraction of each sequence requires comparing the absolute quantity of each sequence in each reacted sample to the quantity of that sequence in the unreacted sample. This can be done by measuring the absolute RNA quantity in each sample. We compared two methods: 1) spiking in a sequence at a known concentration into each sample, providing a conversion between the number of sequence reads and absolute concentration in each sample; or 2) measurement of the total absolute RNA concentration of each sample by QuBit or qPCR. As shown in Figure 2.4, sample quantitation by both methods agreed well with each other, with both having comparable standard deviation among triplicates (Figure A.13). As the first method is disadvantageous in reducing the HTS reads available for ribozyme sequences, the second method is preferred and further analysis was done based on the second method for quantitation.

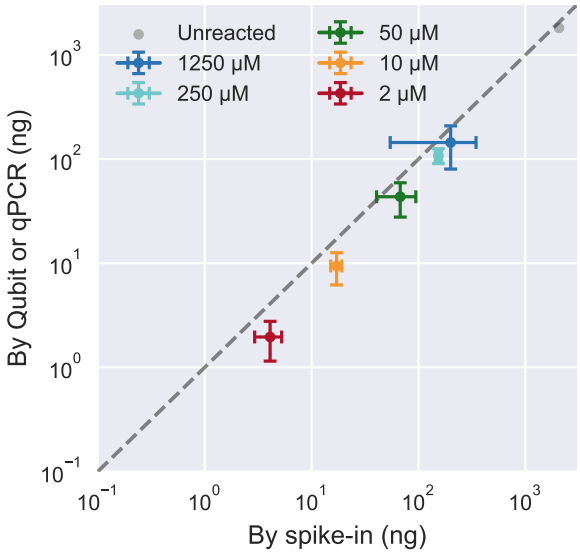


Figure 2.4: Distribution of γ (A) and σ_A (B) for sequences with in Hamming distance of 2 to the family centers from the variant pool k-Seq experiment. Example fitting results are shown for sequences within each score range (labels on the left) of γ (C) and σ_A (D). For explanation of (C) and (D), also see caption of Figure A.2. Sequences with low metric scores for both metrics showed good model identifiability while k and A cannot be separately estimated for those with high metric scores. k and a for most sequences up to double mutants could not be estimated separately.

2.2.5 Accuracy of k-Seq estimation

To evaluate the accuracy in estimating kinetic coefficients using k-Seq, we generated a simulated dataset from a heterogeneous pool containing sequences with known kinetic coefficients (k and A). To resemble the experimental pool, the ground truth values of coefficients were sampled from the point estimates of k and A in the mixed variant pool. Sequence counts and total amount of ribozymes were simulated for reacted sample with Extended Substrate Range in triplicates (see Method 2.5.7) and the data were fitted to the pseudo-first order kinetic model to estimate k and A for each analyzable sequences.

We expected that sequences having a low number of counts would show reduced estimation accuracy. To characterize this effect, we plotted the ratio of the point estimate of kA to its true kA against the average number of counts across the simulated samples (Figure 2.5A). We found that the error in estimation for sequences with high mean counts ($> 1,000$) was $< 10\%$. For a mean count around 100, the errors were roughly within 2-fold, but the error increased substantially as the mean count decreased below 100. Thus sequences with low mean counts (especially < 100), either from low abundance in the input pool or low abundance in the reacted pool (due to low activity), were susceptible to high error in estimation of kinetic parameters. Very high mean counts (e.g. $> 10,000$) would not substantially benefit the measurement, as other sources of experimental error would likely be greater (e.g. we added 10% simulated error in total DNA measurement during simulation) [17, 38]. Thus the results indicate that > 1000 mean count would be favorable for estimation, with > 100 counts being acceptable if a 2-fold error in estimation is tolerated.

While the above analysis provided the accuracy for point estimation, a different method is required to estimate the precision (uncertainty) for real data in which the ground truth is unknown. We therefore explored the accuracy of uncertainty estimation

using bootstrapping. Bootstrap resampling ($n = 100$) was used to estimate the 95% confidence intervals (CI-95) in two ways: first, using mean and standard deviation (SD) of estimated kA (mean $\pm$ 1.96 SD, assuming a normal distribution), and second, using the 2.5-percentile to 97.5-percentile confidence intervals (relax the normal distribution), for sequences in the simulated pool dataset that were analyzable (602,246 sequences in total). A sensible evaluation on these estimated CI-95 is to evaluate the fraction of sequences whose true kA value is included in estimated CI-95. If the estimation were correct, the CI-95 would include the true value for roughly 95% of sequences. We found that 96.5% of sequences included the true kA in estimated CI-95 from 2.5-to-97.5-percentiles and 96.4% from mean and standard deviation, indicating that bootstrapping gave an accurate CI estimation by either method for such sequences. The fractions of sequences with their true kA included in the CI-95 were relatively consistent regardless of their mean counts (Figure 2.5B) or true kA values (Figure A.14). For comparison, we also examined uncertainty estimation using the standard deviations estimated from triplicates (mean $\pm$ 1.96 SD) assuming a normal distribution. In our simulated pool dataset, 83.5% of sequences had the true kA value included in the CI-95 estimated from triplicates, indicating an underestimation of uncertainty using triplicates (Figure A.14).

2.2.6 Precision of *k*-Seq estimation

The precision of k-Seq measurement for the variant pool was evaluated in two ways. First, given the reasonable uncertainty estimated by the bootstrapping procedure, we calculated the fold-range (97.5-percentile divided by 2.5-percentile) of kA from bootstrapping ($n = 100$). While there was a slight tendency for sequences with higher kA to have higher estimation precision (lower fold-range; Figure A.15) in each order of mutants, the precision was more evidently dependent on the mean counts value for sequences

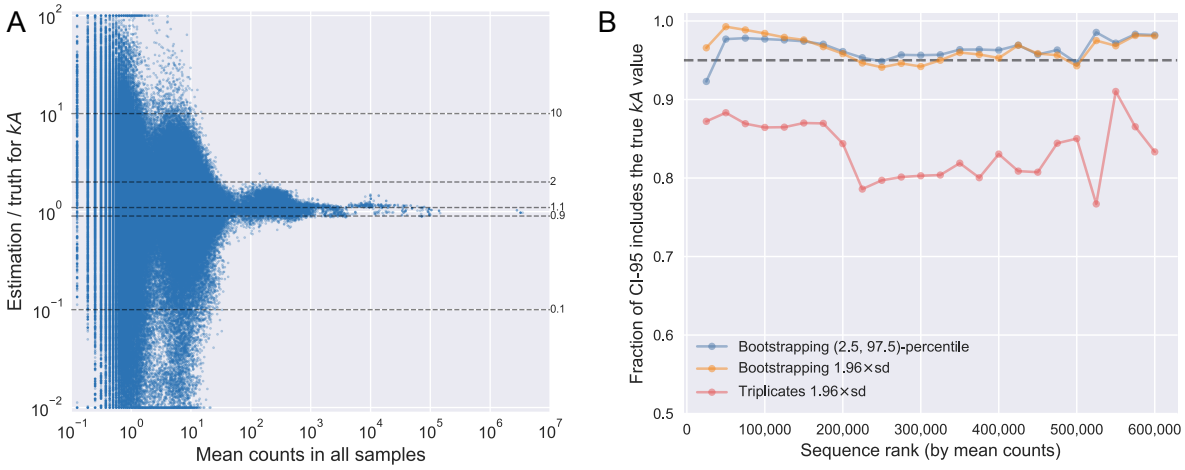


Figure 2.5: Accuracy of parameter estimation by *k*-Seq. (A) Dependence of accuracy (ratio of estimated kA to true kA) on mean counts across all simulated samples (including the unreacted pool sample). The dashed lines correspond to ratios as labeled. Ratios above 100-fold or below 0.01-fold are shown at the borders of the plot. (B) Fraction of sequences for which the CI-95, estimated using bootstrapping or using triplicates, includes the true kA values, for sequences with different mean counts across all samples. Sequences were ranked by mean counts (from highest to lowest) and binned in sets of 25,000 sequences. Each data point indicates the fraction of CI-95 that includes the true values in each bin.

(Figure 2.6). All wild-type sequences and most single and double mutants had 95% CI spanning less than one order of magnitude. Triple mutants seemed to have lower precision, which may be attributed to lower counts. Indeed, greater precision was seen with higher mean counts both within and across groups.

Precision as measured above includes variation among replicates done in the same experimental batch, but does not include variation between different k-Seq experiments. To understand the precision of estimates from independently designed and separately executed k-Seq experiments, we compared the results from the variant pool k-Seq reported here to a previously reported k-Seq assay from a selection pool [17]. 2513 unique sequences found had their 2.5-percentile for kA, estimated from bootstrapping, that was greater than the baseline kA of 0.124 $\text{min}^{-1}\text{M}^{-1}$ (measured in [17]) in both experiments. Point estimates of kA for these sequences from the two experiments were compared to each other (Figure 2.6B). For sequences with sufficient counts (e.g. mean counts of at least 1000 in both experiments, corresponding to 39 sequences), the results from two experiments were well correlated (Pearson's $r = 0.896$, P-value $= 1.20 \times 10^{-14}$, Spearman's $\rho = 0.864$, P-value$=1.38 \times 10^{-12}$), indicating good reproducibility of those measurements from different experiments. As expected, sequences with lower mean counts showed less correlation (for mean counts between 100 and 1000, Spearman's $\rho = 0.171$, P-value $=$ 0.130, 80 sequences), and those with mean counts < 100 showed weak to none correlation (Spearman s $\rho = 0.051$, P-value $= 1.3 \times 10^{-2}$).

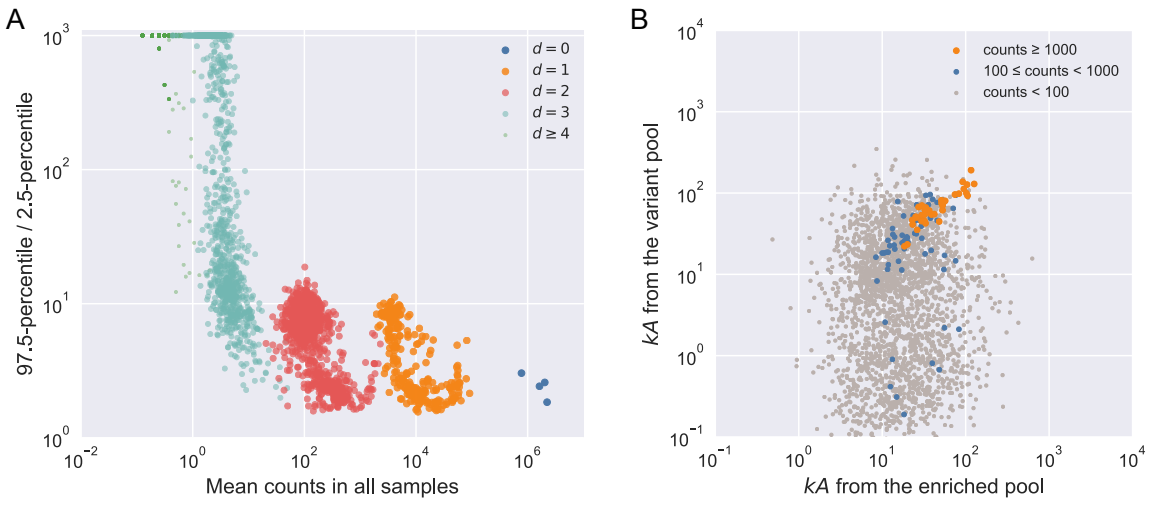


Figure 2.6: Precision of estimation by k-Seq. (A) Fold-range (97.5-percentile / 2.5-percentile) of kA estimation depended on the mean counts. Increasing mean counts increases precision, as shown by the relationship of fold-range with mean counts across different orders of mutants. For $d \geq 2$, only 1000 sequences were randomly selected for visualization. (B) Alignment between estimated kA from two independently conducted experiments (experiment from [17], and the k-Seq experiment reported here). Only sequences with 2.5-percentile higher than baseline catalytic coefficient ($kA = 0.124\text{min}^{-1}\text{M}^{-1}$, reported in [17]) were included. Each point represents a sequence whose color reflects the minimum of mean counts (between two experiments).

2.3 Discussion

2.3.1 Model identifiability for pseudo-first order model

A model is not identifiable when the optimal set of parameters fitting the model cannot be uniquely determined [39]. For the ribozymes exhibiting pseudo-first order kinetics studied here, the parameter k (rate constant) and A (maximum amplitude) cannot be separately estimated when the data collected do not show saturation behavior, i.e., the data fall into the initial linear region of the curve. While it is possible to adjust the substrate concentrations to mitigate the kA-separability problem for individual sequences, it is impossible in experiments to apply optimal conditions for each sequence due to pool heterogeneity and concentration required for slow reacting ribozymes might not be feasible. Thus, we previously used the combined parameter kA as the measure of chemical activity [17]. However, separate estimation of k and A is still an important goal. Therefore, we explored three metrics (σ_A, γ, and ΔA) to assess the model identifiability for each sequence. In general, higher values of k or A and lower noise level yield better separability. Of the metrics, γ, which measures the increased uncertainty in k considering A as an independent constant vs. a confounded parameter, showed good performance. However, it may not be appropriate to assign a strict cutoff value on k, A, kA, or γ, for identifiable models. Instead, we suggest that γ allows one to semi-quantitatively assess separability in combination with experimental intuition to determine a sequence or the parameter region in which the results can be reasonably reported as k and A separately vs. reported as the combined activity parameter kA.

To understand how to improve experimental design to increase the number of sequences in the separable region, we created a simulated reacted fraction dataset and mapped how separability depends on the true value of k and A, given a set of substrate concentrations and noise level. Extending the substrate concentration (to 1250 μM) ex-

pands the region of separable sequences by pushing the lower bound on k or A down by roughly one order of magnitude. In contrast, adding another replicate to each substrate concentration did not change the separability map much. Thus, we used the extended substrate range in the k-Seq experiment to provide a wider dynamic range. Even so, there was still a substantial fraction of sequences with non-separable kA, so in practice the choice of parameters to be reported depends on the goals of the experiment (e.g., for maximum exploitation of the data, all of k, A, and kA could be reported along with γ).

2.3.2 The accuracy and precision for k-Seq

Using DNA sequencing counts to quantify the abundance of sequences has two consequences that need consideration: sequencing error that misidentifies a sequence as a related sequence, and stochastic effects for sequences associated with lower counts.

Mis-identifying a sequence as a neighboring sequence can be particularly problematic for systems where sequences are close in the sequence space and single nucleotide resolution is expected (e.g. mutational analysis on a variant pool). If the pool is quite uneven, the bias from sequencing error can contribute a non-trivial portion of reads to less abundant neighbor sequences, causing the k-Seq measurement for these less abundant sequences to be effectively mixed with those of their neighboring sequences. While model-based sequencing error correction could be attempted [20, 21], we side-stepped this problem by taking advantage of paired-end reads in which the random region of the library was read completely in both directions. By enforcing absolute matching of paired-end reads during the joining process, the error rate of sequencing (approximately 1% per base) should be decreased to its square (0.01% per base). While this does decrease the number of sequence reads that pass quality control, the benefit is important as mis-identification from this level of sequencing error is essentially negligible for observed

abundances (Figure 2.3).

Low counts have a major impact on the accuracy of estimation of kinetic coefficients. In practice, we found that the average counts for a sequence across samples (i.e., mean counts) is a better guide for estimation accuracy compared to counts in the input pool (Figure 2.5, Figure A.14). We find that accuracy is good for sequences above a certain threshold of mean counts (e.g. 100 reads per sample), but decreases quickly below this. Meanwhile, the benefit from larger counts (e.g., 10,000 reads per sample) was marginal and other experimental factors likely contribute greater error.

Estimating uncertainty is important for k-Seq experiments, but replicates are likely to be limited due to the expense associated with HTS. Bootstrapping simulates virtual experiments by resampling data (in this case, the relative residuals from fitting) with replacement to its original size. Bootstrapping results can be used to estimate population characteristics, such as percentiles for kinetic coefficients. Indead, bootstrapping results reflected the true 95% uncertainty level more appropriately than the standard deviation estimated from triplicate experiments, as the latter tended to underestimate the uncertainty of estimation. As seen for the accuracy of low count sequences, precision also showed a steep drop-off when mean counts dropped below 100 (Figure 2.6A), while additional counts did not significantly improve precision. Using bootstrapping instead of replicates also provided resampling data that could be used for calculating statistics for kA-separability analysis. Although a disadvantage is computational expense, the number of bootstrap samples can be adjusted and computation can be implemented in parallel. As modern computational resources become cheaper and easier to access, bootstrapping becomes more affordable. Therefore, while experimental replicates are valuable for controlling for some sources of error, we suggest that bootstrapping analysis is an excellent method for properly estimating errors of parameter estimation.

2.3.3 Optimized experimental design for *k*-Seq experiments

To maximize coverage, or the number of sequences with estimated model parameters having acceptable accuracy and precision, it is desirable to maximize the number of sequences satisfying a minimal count requirement without spending excessive sequencing resources on abundant sequences. In the present experiment, we had approximately full coverage for single and double mutants in each family, for which the measurement precision may be considered reasonable (fold-range < 10; mean counts > 10) (Figure 2.6A). While HTS technology enabled the kinetic measurement for large pools with high richness (number of unique sequences), the practical coverage for k-Seq is affected by pool evenness, as highly uneven pools may have many sequences with insufficient counts for precise estimation. Such pools may result from enrichment after selections or from variant pool synthesis of ribozyme variants exploring many mutants of a given wild type. For enriched pools from selection experiments, the pool evenness usually decreases during the selection. For doped pools, evenness is tuned by the ratio of wild-type nucleotides at each position. In the analysis of BYO-aminoacylation ribozymes presented here, the designed variant pool was more even than the selection pool from which these ribozymes were derived (Figure A.16); thus k-Seq analysis of selected ribozymes may be improved by designing a new pool rather than directly analyzing the selected pool itself.

2.4 Summary

In this chapter, we developed a model analysis pipeline for k-Seq on a pseudo-first order kinetic system to measure the kinetic coefficients for up to 10^5 different ribozymes. To improve kinetic fitting, we proposed a bootstrapping process to estimate the joint distribution of fitted parameters observing the noise from the data. From bootstrapping results, we robustly quantify the estimation uncertainty and proposed a metric to semi-

quantify the model identifiability in fitting pseudo-first order models, for each individual sequences. Using this improved method, we further studied the critical parameters affecting the utility of k-Seq, and concluded that the pool composition played an important role on the coverage of well-fitted sequences. Lastly, both theoretical and experimental guidance were provided for future practitioners on the design of the pools for k-Seq type experiments and minimum sequencing depth required for desired fitting quality.

2.5 Methods

2.5.1 *k*-Seq experiment on mixed pool of ribozymes

We designed a mixed pool of ribozyme variants containing four active wild-type sequences (S-1A.1-a, S-1B.1-a, S-2.1-a, and S-3.1-a, see Table 2.1 for sequences) previously identified in [17] and their variants. Specifically, four DNA libraries were obtained from Keck Biotechnology Laboratory, with the sequence 5′-GATAATACGACTCACTATA-GGGAATGGATCCACATCTACGAATTC-[21 nt variable region]-TTCACTGCAGAC TTGACGAAGCTG-3′ (nucleotides upstream of the transcription start site are underlined). In each library, the central region corresponded to a 21-nucleotide variable sequence, based on the wild-type sequence, with partial randomization at each position (91% of the wild-type nucleotide and 3% of each base substitution at all 21 positions). RNA was transcribed using HiScribe T7 RNA polymerase (New England Biolabs) and purified by denaturing polyacrylamide gel electrophoresis (PAGE) as previously described (2). An equimolar mixture of these four RNA libraries (the variant pool) was prepared for the k-Seq experiment.

In the k-Seq experiment, reactions were carried out on pools of mixed ribozymes in triplicates at 2, 10, 50, 250, and 1250 μM BYO for 90 min, following the incubation, RNA

recovery and reverse-transcriptase protocols used in [17]. Briefly, in each 50 μL k-Seq reaction, 2 μg total RNA was reacted with BYO. The reactions were stopped by desalting and placed on ice. Reacted sequences were isolated by pull-down with Streptavidin MagneSphere paramagnetic beads (Promega) and eluted with formamide/EDTA. 10% of eluted RNA was taken to measure the total RNA amount using Qubit and qPCR (see below). A 'spike-in' RNA was added as an alternative quantification method (see below). RNA was prepared for sequencing by reverse transcription and PCR (RT-PCR), with primers complementary to the fixed sequences flanking the variable region. DNA from each of 15 samples was barcoded and pooled together in equal proportions. A reverse-transcribed unreacted sample was added at three times the total amount of DNA of one reacted sample to have similar total sequencing depth with each set of BYO concentration triplicates. Pooled DNA was sequenced on an Illumina NextSeq 500 with 150 bp paired-end run (Biological Nanostructures Laboratory, California NanoSystems Institute at UCSB), using a high output reagent kit expected to produce $>$ 400 million reads.

2.5.2 Quantitation of total amount of RNA in samples

We used two methods to quantify the absolute amount of RNA in each k-Seq samples. The first method measured the amount of RNA in the samples after elution using Qubit or qPCR. For reactions carried out at 250 and 1250 μM, 10% of the RNA recovered after elution was quantified with an Invitrogen Qubit 3.0 fluorometer. If the recovered RNA was below the limit of detection by Qubit, quantitation was done by reverse-transcription-qPCR using a Bio-Rad C1000 thermal cycler with CFX96 Real-Time PCR block.

The second method used an internal standard (a spike-in sequence) with known amount to normalize the data using sequencing results. A spike-in sequence different from

ribozyme sequences (5′-GATAATACGACTCACTATA-GGGAATGGATCCACATCT ACGAATTC-[AAAAACAAAAACAAAAACAAA]-TTCACTGCAGACTTGACGAAGC TG-3′, promoter underlined) was added to each sample before reverse transcription. 0.04, 0.2, 1, 2, and 2 μg of spike-in RNA was added to samples with 2, 10, 50, 250, 1250 μM BYO concentration respectively. 10 μg of spike-in RNA was added to the unreacted pool sample. The total RNA recovered (Q_j) in sample j was calculated as

$$Q_j = \frac{N_j - n_{sj}}{n_{sj}} \times q_{sj} \tag{2.4}$$

where N_j is the total number of reads in sample j, n_{sj} is the total reads of sequences within 2 edit distance (number of substitution, insertion, or deletion) of spike-in sequences in sample j, and q_{sj} is the quantity of spike-in sequence added to the sample.

2.5.3 Read pre-processing and quantitation of reacted fraction for ribozymes

`FASTQ` files of de-multiplexed paired-end Illumina reads were first processed through a customized shell script [40] to count the number of reads of each unique sequence in each sample. The forward and reverse reads were joined using `pandaSeq` [41] with the options `-a` to join the paired-end reads before trimming and `completely_miss_the_point:0` to enforce absolute matching in the overlapped variable region was required (any pairs with a disagreement between forward and reverse reads were discarded), thus minimizing sequencing errors. After joining, forward and reverse primers were also trimmed by `pandaSeq` using 'CTACGAATTC' (forward) and 'CTGCAGTGAA' (reverse) adapter sequences. Next, multiple lanes from Illumina sequencing for the same sample were combined and reads were de-replicated to unique sequences and counts.

The generated count files were analyzed using the `k-seq` python package developed

in house. We collected all detected sequences in unreacted and/or reacted samples and discarded those that were not 21 nucleotides long or within an edit distance of 2 from the spike-in sequence. The absolute amount (ng) for each sequence in samples were quantified using total RNA recovered Q_j and number of reads for sequence i in the sample (n_{ij}):

$$q_{ij} = \frac{n_{ij}}{N_j - n_{sj}} \times Q_j \tag{2.5}$$

The reacted fractions for sequence in reacted samples were further calculated as

$$f_{ij} = \frac{q_{ij}}{q_{0j}} \tag{2.6}$$

To be analyzable for fitting, a sequence needs at least one non-zero value among reacted samples as well as a non-zero count in the unreacted sample; non-analyzable sequences were discarded.

2.5.4 Estimation kinetic coefficients with uncertainty quantification

Kinetic coefficients k_i, A_j for sequence i were estimated using least-squares fitting on reacted fractions (f_{ij}) with different initial BYO concentrations (c_j). Least-squares fittings were performed using the `optimize.curve_fit` function from the `scipy` package in `python` with "trust region reflective" (`trf`) method. The initial values of k_i, A_i were uniformly sampled from (0, 1) for fittings. To ensure the convergence, the bounds [0, 1] were applied on A and [0, $+\infty$) on k. The tolerances for optimization termination (`ftol`, `xtol`, `gtol`) were kept as default (10^{-8}). Optimal k_i, A_i determined from all sample points for a sequence were reported as point estimates.

The uncertainty of estimation was assessed using bootstrap sampling of the relative

residuals. Let f_{ij} be the reacted fraction for sequence i in sample j, and $\hat{f}_{ij}$ be the fitted value from point estimation. For each sequence, we calculate the relative residual as $r_{ij} = \frac{f_{ij} - \hat{f}_{ij}}{\hat{f}_{ij}}$, where $j = 1, 2, \ldots$ corresponds to each reacted sample. Each bootstrapping process re-sampled the relative residuals for sequence i with replacement to the same sample size $(\hat{r}_{i1}, \hat{r}_{i2}, \ldots, \hat{r}_{iJ})$, then added the re-sampled f_{ij} (that is, $(1 + \hat{r}_{ij})\hat{f}_{ij}$) as bootstrapped data points. Estimation was performed on each bootstrapped datasets for which k_i, A_i and $k_i A_i$ values were recorded. Sample mean (μ), standard deviation (σ), median, and estimated 95% confidence interval (CI-95, as $\mu \pm 1.96\sigma$ or [2.5-percentile, 97.5-percentile]) on k_i, A_i and $k_i A_i$ were calculated from bootstrapped results for each sequence.

Bootstrapping was performed for 100 re-samples for each sequence for uncertainty estimation. To compare the performance of bootstrapping, we also applied the triplicates method, used previously [17], to the simulated pool dataset, with each replicate in a BYO concentration assigned to one of three series. Each of the simulated triplicate series was fitted separately to calculate the standard deviation of k_i, A_i and $k_i A_i$.

2.5.5 Model identifiability for different k and A

To quantify whether the pseudo-first order kinetic model for a sequence is identifiable [42] (i.e., the values of k_i and A_i can be separately estimated) given the BYO concentrations and experimental noise, we designed two metrics to estimate the model identifiability from bootstrapping results: σ_A and $\gamma = \log_{10} \frac{\sigma_k \mu_A}{\sigma_{kA}}$, where σ_A, σ_k, σ_{kA} are the standard deviations for A, k, and combined parameter kA in bootstrapped samples. σ_A represents the variance of estimated A during bootstrapping and the metric γ is the log-ratio of standard deviation of combined parameter kA over k adjusted by estimated constant A. If k and A are well-estimated independently, the ratio would be close to 1

and γ would be close to 0. For comparison, we also examined the fitting convergence through 20 independent fittings (no resampling) with initial values of k and A randomly sampled from Unif$[0, 10]$ $\text{min}^{-1}\text{M}^{-1}$ and Unif$[0, 1]$ respectively. The variance of the fitting results was evaluated by the range of fitted A values ($\Delta A = A_{\max} - A_{\min}$) and used as an extra candidate metric for model identifiability (higher range means less identifiable).

2.5.6 Simulated reacted fraction dataset

To study the model identifiability, we simulated a reacted fraction dataset containing 5×10^3 sequences with kinetic coefficients ($\log_{10} k$, $\log_{10} A$) sampled from a regular grid where $\log_{10} k \in [-1, 3]$ and $\log_{10} A \in [-2, 0]$. The reacted fraction for sequence i in sample j with BYO concentration c_j is calculated using the pseudo-first order model (Equation 2.3) with error:

$$f_{ij} = f_{ij}^0 + \text{err}_{ij} = A_i(1 - e^{-\alpha t k_i c_j}) + \text{err}_{ij}, \text{err}_{ij} \sim N(0, \epsilon f_{ij}^0) \tag{2.7}$$

The error term err_{ij} is parameterized by the relative error ϵ and the true reacted fraction f_{ij}^0. We chose $\epsilon = 0, 0.2, 0.5, 1.0$ to simulate different levels of measurement error. Negative values of f_{ij} after adding errors were reassigned to be zero. These simulated reacted fractions were used to estimate k and A for each sequence using least-squares fitting as described above.

To study whether sequencing effort would be best spent extending the substrate concentration range or performing additional replicates, we simulated reacted fraction data for three different sets of BYO concentrations: i) standard set: 2, 10, 50, and 250 μM with triplicates, 12 samples in total, as done previously analyzing a pool after *in vitro* selection [17]; ii) additional replicates: 2, 10, 50, and 250 μM with four replicates each, 16 samples in total; and iii) extended substrate range: 2, 10, 50, 250, and 1250 μM

with triplicates (15 samples, as done in the variant pool experiment reported here).

2.5.7 Simulation count data for RNA pools

Simulated k-Seq pool count data were constructed from a set of true p_{i0}, k_i, A_i values, where p_{i0} is the initial relative abundance for sequence i in the unreacted pool and k_i, A_i are point estimates from the k-Seq experiment on the mixed variant pool. We simulated $M = 10^6$ sequences with parameters (p_{i0}, k_i, A_i) sampled from above parameter set, and renormalized p_i0 to sum to 1. We used the pseudo-first order rate equation to calculate the reacted fraction f_{ij} for each sequence and obtained a new relative abundance for sequence i in sample j as $p_{ij} = \frac{p_{i0}f_{ij}}{\sum_{r=1}^{M} p_{r0}f_{rj}}$. We then used the multinomial distribution $\text{MultiNorm}(p_{1j}, p_{2j}, \ldots, p_{Mj}, N_j)$ (where $N_j = 40M$, yielding a similar mean count per sequence to that observed in the experimental pool), to model the process of sampling a given number of reads given p_{ij} during sequencing. The simulated number of counts for sequence i in sample j is drawn from this distribution. To simulate total RNA recovered, we sampled a value from a normal distribution with mean equal to the true RNA amount in the mixed pool reaction and 15% error (similar to standard deviation calculated in spike-in or direct RNA amount quantification), using $N(\mu_j, 0.15\mu_j)$, where $\mu_j = \sum_{i=1}^{M} p_{i0}f_{ij}$.

Chapter 3

microNet: construct correlation networks between microorganisms from metagenomic data

3.1 Background

3.1.1 Metagenomic study for human microbiome

Human microbiomes contain microorganisms colonizing on human body [43]. These microorganisms (or called microbes) in human microbiomes are abundant and diverse that 10^{13} microbial cells and $500 \sim 1000$ species were estimated on one's body [44,45,46]. The genetic and functional diversity of these microbes are even more substantial. For example, 3.3 million non-redundant genes were found in a study of human gut microbiome [47]. With these microbes being our close neighbors, human microbiomes have received great attention on understanding the composition, structure, functions, as well as how they are related to human health. Studies have showed that microbiomes are closely associated with many important health issues, including infections, inflammatory bowel diseases, metabolic disorders, cardiovascular diseases, etc [45,47,48]. The study on microbiomes is still at its early age and little is known or understood about microbiomes, especially the structure, functions, and mechanisms [1].

Microbiomes are complex microbial communities with many different microbes. High-throughput methods are required to simultaneously characterize different microbes and factors (e.g., metabolites) in a microbiome. Bioinformatic tools and statistical models are also important to identify and quantify entities in the samples, and to assess the relations between microbes and extract other system-level knowledge. HTS-based metagenomic methods are developed to profile the genomic content for microorganisms and quantify their relative abundance directly from the environmental samples. For example, 16s ribosomal RNA (16s rRNA) amplicon sequencing and Internal Transcribed Spacer (ITS) amplicon sequencing have been developed to profile bacteria/archea and fungi using some universal "fingerprint" genes. Whole metagenomic sequencing can be applied to profile the viral content (e.g., viruses, bacteriophages) in the microbiome [49]. These

metagenomic data contain DNA sequencing reads from microbes and bioinformatic tools are available to preprocess the sequencing reads and group them into Operational Taxonomic Units (OTUs) as proxies for original microbes (or microbial taxa) [25,49,50,51,52]. By counting the number of reads assigned to OTUs, the relative abundances of microbes are quantified and can be used for downstream statistical analysis.

Skin is the first barrier of human body against potential pathogens in the environment. Skin microbiome contains microbes and is closely related to the health of skin and wound infections [49, 53]. A specific interest in Chen lab is on the skin microbiome and chronic wounds (wounds that do not heal normally). Verbanic et. al. have collected samples from patients with chronic wounds and used 16s rRNA Amplicon sequencing and Whole Metagenomic Sequencing (WGS) to survey the bacterial and viral contents in those skin microbiome samples (see Method 3.5.1 for details) [49, 54]. Their work has showed the association between skin microbiome and the disease state that the abundance for some bacterial OTUs were significantly different in wound and skin samples. While the previous work showed the difference for individual OTUs under different conditions, it is important to further understand the potential relations between these bacterial and viral OTUs in the skin microbiome across the patients. Precisely, we are interested in quantifying the correlations between the microbial OTUs (bacterial and viral) to identify the most prominent signals and construct a correlation network to assess the structure of microbiome at the system level.

3.1.2 Construct the microbial networks from metagenomic data

Graphical models

A graph consists of a set of nodes (vertices) and edges connecting the nodes. Nodes can represent the entities of interests and edges can represent certain pairwise relation-

ships between the nodes. Graphs are common theoretical tools widely used to represent and model complex systems with network structures. In biology, graphs have been used to represent the knowledge of relations between organisms (e.g., a food web [55]), genes (e.g., gene regulatory network [56]), or proteins (e.g., protein-protein interaction network [57]). In addition to describing the knowledge, graphs can also be used to mathematically model the dependencies between some attributes (e.g., abundance) of entities (so called variables). For example, a probabilistic graphical model describes the statistical dependency between a set of random variables represented by nodes where the distribution of a random variable (node) depends on the values of nodes connecting to it (i.e., neighbors) [58, 59, 60]. Such graphical models provide a concise mathematical framework to describe the relations in complex biological data from high-throughput technologies and help to investigate the relations between different components of the system.

A microbial network is a graph describing the ecological relations between the microbes sharing the same niche (e.g., a microbiome). Determining the exact ecological relations (e.g., mutualism or competition) often requires thorough observational or experimental data to determine the mechanism of interactions. Alternatively, the associations between the abundance of two microbes in shared habitats provide an approximation on microbial interaction and help to construct the hypotheses on the potential interactions between microbes. These associations reflect the "co-occurrence" patterns in ecology and are usually quantified by calculating the pairwise correlations (e.g., Spearman's correlation or Pearson's correlation) [61, 62, 63, 64], ecological distance (e.g., Bray-Curtis distance) [62, 65, 66], or an ensemble of these measures [67] between two microbes. The statistical significance of each association or/and a threshold to accept an association as an edge need to be further determined in order to construct a parsimonious co-occurrence graph.

However, these co-occurrence methods discover the associations between the microbes

that can be the result of direct or indirect interactions. For example, in a minimal microbial community with three microbes (A, B, C) where microbe A and microbe C both interact with microbe B (e.g., positive correlation) but not each other, a positive Spearman's or Pearson's correlation can be calculated between microbe A and C. These correlations are 'marginal' correlations as they are determined from the marginal distribution of data over the abundance of microbe A and C, neglecting the abundance of other microbes. In a complex system like microbiome, microbes are expected to interact with many other microbes, thus, marginal correlations might lead to many edges from indirect interactions and are difficult to interpret. In comparison, a conditional correlation (or partial correlation) indicating direct relations between two variables while fix (i.e., conditioned on) the values of other variables. In the previous example, the conditional correlation between the abundance of microbe A and microbe C is zero, assuming no other confounding factors. The conditional correlations provide a more concise and interpretable way on the associations between microbes in a potentially interconnected system.

Gaussian Graphical Model

While it is impractical to determine the conditional correlations between each pair of variables by controlled experiments, Gaussian graphical model (GGM) provide a theoretical framework to estimate the conditional correlations from observational data. In Gaussian graphical model (GGM), the joint distribution of variables in the data is assumed to follow a multivariate Gaussian distribution:

$$\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma}) \tag{3.1}$$

$\mathbf{X} = [X_1, X_2, \ldots, X_p] \in \mathbb{R}^p$ is a vector of p random variables for the attribute of interests. The distribution of $\mathbf{X}$ follows the multivariate Gaussian, parameterized by a mean vector $\mu \in \mathbb{R}^p$, and a covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$. Inverse of the covariance matrix is called precision matrix ($\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$) and contains the information of conditional correlations between p random variables. The [marginal] correlations and the conditional correlations can be calculated from $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ respectively, by normalizing to the diagonal terms:

$$\rho(X_i, X_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \tag{3.2}$$

$$\rho(X_i, X_j | \{X\}_{\setminus X_i, X_j}) = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} \tag{3.3}$$

where $\sigma_{ij}, \sigma_{ii}, \sigma_{jj}$ are entries in the covariance matrix $\mathbf{\Sigma}$, $\omega_{ij}, \omega_{ii}, \omega_{jj}$ are the entries in the precision matrix $\mathbf{\Omega}$, and $\{X\}_{\setminus X_i, X_j}$ is the set of all variables excluding X_i and X_j. By fitting the GGM to the data, one can estimate for the precision matrix $\mathbf{\Omega}$ and thus the conditional correlations.

Gaussian Graphical LASSO

Gaussian Graphical LASSO (GLASSO) is an algorithm commonly used to estimate a sparse precision matrix $\mathbf{\Omega}$ from data [68]. It is a L1-penalized maximum likelihood estimation (MLE):

$$\mathcal{L}^{\circ} = \frac{N}{2} \log\det(\mathbf{\Omega}) - \frac{N}{2} \mathrm{tr}(\mathbf{\Omega}\hat{\mathbf{\Sigma}}) - \frac{Np\log(2\pi)}{2} \tag{3.4}$$

$$\mathcal{L} = \log\det(\mathbf{\Omega}) - \mathrm{tr}(\mathbf{\Omega}\hat{\mathbf{\Sigma}}) \tag{3.5}$$

$$\hat{\mathbf{\Omega}} = \arg\min_{\mathbf{\Omega}} \Big(-\mathcal{L} + \lambda ||\mathbf{\Sigma}||_1 \Big) \tag{3.6}$$

Equation 3.4 is the canonical form of log likelihood for multivariate Gaussian distributions where N is the number of samples, p is the dimension of the distribution (number of variables), $\det(\cdot)$ is the determinant of a matrix, and $\mathrm{tr}(\cdot)$ is the trace of a matrix. $\hat{\mathbf{\Sigma}}$ is the empirical covariance matrix that can be calculated from the data as $\hat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^\intercal$ and $\mathbf{x_i} = [x_{1i}, x_{2i}, \ldots, x_{pi}]$ is the data vector for sample i. Maximizing the log likelihood $\mathcal{L}^\circ$ is effectively same as maximizing $\mathcal{L}$ in Equation 3.5. For noisy "real-world" data, GLASSO applies L1 penalty (or called LASSO penalty) on the precision matrix $||\mathbf{\Omega}||_1 = |\sum_{i=1}^{p}\sum_{j=1}^{p}\omega_{ij}|$ for L1-penalized MLE in Equation 3.6. L1 penalty can enforce the entries in the precision matrix to be zero for a small set of conditional correlations that best describes the variance in the data. And the hyperparameter λ controls the sparsity of inferred precision matrix and thus the density of the network. The optimization problem in Equation 3.6 is convex and can be efficiently solved using block coordinate descent in GLASSO.

3.1.3 Design factors and challenges in applying GGMs to metagenomic data

Multi-domain compositional data

A well-known analytical challenge from metagenomic data is the compositionality. Due to the difficulty in measuring the total amount of microorganisms in a environment, most metagenomic data only provide the relative abundance of microbes detected in samples, represented as counts. Compositional data lack one degree of freedom and are defined on a simplex (all variables sum up to 1) when normalize to fractions. Many statistical tools are invalid for compositional data including correlation estimations. The decrease of the relative abundance of a variable increases the relative abundance of other variables, introducing false correlations between the variables. Log-ratio transforma-

tion has been proposed to tackle the compositional data and centered log-ratio (CLR) transformation has been applied to metagenomic data in the inference of microbial correlations [28, 30, 69, 70]. Let $\mathbf{c} = [c_1, c_2, \ldots, c_p]$ be the OTU count vector of a sample and the CLR transformation is defined as

$$x_i = \frac{\log c_i}{\sum_{k=1}^{p} \log c_k} \text{ for } i = 1, 2, \ldots, p \tag{3.7}$$

where the covariance matrix for $\mathbf{x} = [x_1, x_2, \ldots, x_p]$ has been proved to be a good approximate for the covariance matrix of the unobserved log true abundance for the microbes $\log \mathbf{z} = [\log z_1, \log z_2, \ldots, \log z_p]$ when p is large [69]. Consequently, inferring precision matrix over the CLR-transformed data is an alternative approach to construct the conditional correlation network between microbes. Another approach to tackle the compositionality is modeling the true abundance using latent variables. `PLNmodel` is a Poisson Log-Normal model where the log true abundance is modeled by a latent multivariate Gaussian distribution and observed counts are modeled by Poisson distributions [71]. Despite being able to directly model the log true abundance and the counts, the inclusion of latent variables requires using expectation-maximization (EM) for optimization and variational approximation for the conditional distribution of latent variables.

One of the emerging interests in the recent microbiome study is the inter-domain/kingdom relations between the bacteria, fungi, or viruses, using multi-domain metagenomic data [49, 54, 72]. Nevertheless, current metagenomic pipelines require different DNA library preparation protocols, sequencing, and bioinformatic pipelines for different domains of organisms. And the consequently, the multi-domain metagenomic data are compositional in each domain. A generalization of CLR transformation has been recently proposed for multi-domain compositional data [72]. The authors have showed that by applying CLR transformation to each domain and concatenating the CLR-transformed abundance ma-

trices, the covariance matrix of transformed data can still approximate the covariance matrix of true abundance of microbes with different domains.

Experimental variables as confounding factors

One important aspect of metagenomic studies is to compare microbiomes in different types of samples and experimental variables are part of research design for metagenomic studies. Microbes likely have different mean abundance in different types of samples and such systematic differences could be falsely attributed to correlations and confound the inference. The difference of sample mean can be accounted using conditional Gaussian models (cGGM) [73]:

$$\mathbf{X}_i \sim \mathcal{N}(\beta_0 + \beta_1 \mathbb{I}_{i1} + \beta_2 \mathbb{I}_{i2} + \cdots + \beta_K \mathbb{I}_{iK}, \mathbf{\Sigma}) \tag{3.8}$$

where $\mathbf{X}_i$ is the random vector for sample i and the mean of Gaussian is a linear function of effects of different sample types. $\beta_0 \in \mathbb{R}^p$ is the baseline mean abundance vector, $\beta_k \in \mathbb{R}^p$ $(k = 1, 2, \ldots, K)$ is the effect vector on mean abundance in sample type i compared to the baseline, and $\mathbb{I}_{ik}$ is an indicator function whose value is 1 if sample i belongs to sample type k and 0 otherwise. By setting off the mean effect from different sample types, the value in the covariance matrix $\mathbf{\Sigma}$ is not confounded by experimental variables.

Balancing the information and noise

Metagenomic data from HTS are usually noisy and have small number of samples. Given the extremely high dimensions (number of OTUs) in the data, it is particularity important to balance the information obtained and noise included in the inference when using GGM to infer the correlation networks (i.e., specificity vs. recall). Generally, the

more OTUs included for the inference and edges inferred from the model, the more useful information one might get but also potentially a greater number of false edges are likely included. Excessive number of false edges decrease the specificity of the inference and the reliability when interpreting inferred edges; they can also change the topology of inferred network with misleading results. To control the noise in the inferred network, one can A) clean OTUs before inference; B) tune the L1 penalty for a sparser graph in GLASSO; C) add post-hoc filtering for inferred edges.

A common practice in metagenomic analysis is cleaning the OTUs by aggregating OTUs with similar taxa and removing less abundant or prevalent OTUs. Depending on the bioinformatic pipeline, OTUs identified have certain taxonomic resolutions. For example, the sequences for the V1-V3 regions in the 16s RNA gene clustered at 97% similarity is commonly used to determine the genus of OTUs but can also possibly distinguish some OTUs at the species level (e.g., *Staphylococcus*) [74, 75]. Aggregating OTUs to corresponding taxonomic levels can reduce noise from bioinformatic steps and the number of variables. Moreover, less abundant OTUs or less prevalent OTUs that are only found in small number of samples can also be removed to focus the study on the major microbes that are more likely to play an important role in microbiome. Removing a subset of OTUs changes the relative abundance of remaining OTUs by a constant factor and its effect is removed after centered log-ratio transformation.

The L1 penalty in GLASSO is controlled by a hyperparameter λ: smaller the value of λ, weaker the L1 penalty, more the non-zero entries in the precision matrix, and thus denser the correlation network. However, determining a optimal or proper value for λ is not straightforward as the true density of the network is not known. In practice, a series of models with different λ are inferred and a "proper" graph is selected for downstream analysis by either practitioner's discretion or some data-driven criteria for model selection. Model selection criteria generally fall into three categories: 1) likelihood-based

criteria, including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or Extended Bayesian Information Criterion (EBIC) [71,76]. These proposed metrics are based on the likelihood of fitted distribution but penalized by model complexity to encourage a less complex (less edges) model. The model (λ) maximizing the metric is selected as the "best" model; 2) cross validation, a model selection method widely used in statistics and machine learning models. Cross validation fit the model to a subset of data and evaluate the goodness-of-fitting on the leave-out data. The model maximizing the goodness-of-fit in the leave-out set is selected; 3) stability-based criteria such as stARS proposed in [77]. stARS subsample a given number of copies of data and fit the model with different λ values. For each λ, the fraction of each edge inferred across subsamples can be calculated and a stability score of inference for each λ is calculated. The model yields a desired stability (suggested 0.9 by the authors) is selected. See Method 3.5.4 for details of above selection criteria.

Lastly, post-hoc edge filtering has also been proposed to remove edges with the absolute value of conditional correlation less than a threshold after graphs are inferred. The threshold can be determined using cross validation [78] or hypothesis testing [79]. They have showed that the original graph inferred from algorithms like GLASSO could be noisy and post-hoc filtering can help to control the False Positive Rate of discovered edges.

3.1.4 Structural analysis on microbial networks using graph algorithms

In addition to conditional correlations recovered by edges from the GGM model, network analysis has advantages on recovering other structural characteristics of microbiome using graph algorithms on the inferred network and assess their potential biological in-

terpretations. Two types of common structural characteristics in microbiome studies are central nodes and node clusters [80].

Node centralities measures how nodes connected on inferred graph and different centrality measures can reflect different perspectives of "connectedness" for the nodes. Nodes with high centralities might represent important members in the microbial community, such as "keystone" or "hub" taxa [80, 81]. Two types of centralities are often evaluated: degree (or normalized value, degree centrality) and betweenness centrality. Node degree measures the number of edges connected to a node. In microbial networks inferred from GGM, a node with high degree is a microbe whose abundance is correlated to many other microbes and could be hypothetically important in microbiome formation or active in microbial interactions. Betweenness centrality is the fraction of shortest paths (the continuous route from one node to another with least number of edges) between any two nodes in the graph that pass the node and nodes with high betweenness centrality is important in connecting different part of the graph. In microbial networks with high modularity (i.e., nodes form clusters), high betweenness centrality might indicate the importance in connecting different microbial groups of the microbiome.

Inferred network might contains clusters of nodes where nodes in each clusters are more closely connected than nodes from different clusters. One might hypothesize that microbes have preference in living with other microbes and such clusters in a microbial correlation network might indicating highly correlated subcommunities of microbes in a microbiome (if samples are collected from the same habitat, e.g., a longitudinal study) or a group of microbes that commonly co-occur in a type of microbiome (if samples are collected from multiple habitats, e.g., studies collect skin microbiome from multiple patients). Some algorithms have been proposed to discover node clusters on graph based on graph spectrum [82], modularity maximization [83, 84, 85], or random walk [86, 87]. Among these algorithms, Louvain is a greedy algorithm that iteratively merge nearby

nodes to find a clustering maximizing the modularity of the graph [83]. It is found favorable to recover larger communities [88] and has been applied to detect community structures of microbial networks [89, 90].

3.1.5 Overview

In this section, we present our studies on constructing the [conditional] correlation network in skin microbiome from a cohort of patients with chronic wounds. Using the bacterial and viral metagenomic data collected from clinical samples [49, 54], we present a `CLR-cGGM` model to account for the compositional data from multiple domains and potential confounding effects from experimental covariates (wound vs. skin samples). Given the small sample size of the study, we carefully control the balance between inferred edges and noise (false edges) using data-driven model selection criteria and post-hoc edge filtering. Lastly, we investigate the pairs of microbes detected with strong correlations and use graph algorithms to detect nodes potentially important to the microbiome and clusters of microbes identified in the clinical dataset.

3.2 Results

3.2.1 Data preprocessing and quality controls

Following the bioinformatic pipelines described in Section 3.5.1, we identified 22,753 bacterial OTUs (bOTU) and 20,098 viral OTUs (vOTU) across all the samples (including the control samples), and 22,229 bacterial OTUs and 19,953 viral OTUs from the clinical samples. On average, there are 52443.2 ± 60966.4 bacterial reads and 407034.9 ± 515703.6 viral reads from wound samples, 161515.2 ± 36875.7 bacterial reads and 325907.2 ± 708023.1 viral reads from skin samples. Table A.2 summarizes the detailed statistics for

reads recovered from samples. As described in Section 3.5.2, we removed the viral OTUs that are putative contaminants (582 OTUs) or are eukaryotic viruses (242 OTUs). Most of discarded reads were from putative contaminants for skin samples and eukaryotic viruses for wound samples (Figure A.18). On average, $14.0 \pm 11.0\%$ of viral reads in wound samples and $13.6 \pm 9.0\%$ of viral reads in skin samples were preserved.

OTUs were then aggregated according to their taxonomic annotations. bOTUs were aggregated at the genus level and the vOTUs were aggregated at the genus level of their hosts, except OTUs for *Staphylococcus spp.* that were aggregated at the species level. 11,962 bOTUs and 17,239 vOTUs were obtained from the taxonomic aggregation. However, most OTUs were found only in < 5 samples and having low relative abundance (Figure A.19A). To focus on a "core set" of microbiome with OTUs detected across different samples, we filtered bacterial and viral OTUs by their prevalence in samples: OTUs that were detected only in wound or skin samples, or detected in $< 80\%$ of patients (< 16) were removed; 99 of 11,962 bOTUs and 48 of 17,239 vOTUs passed this filter. Despite removing most of unique OTUs, the remaining OTUs were highly abundant in the samples (e.g., mean relative abundance $> 0.1\%$). $> 80\%$ of bacterial reads and $> 50\%$ of viral reads were preserved for most of the samples after removing less prevalent OTUs, except for the viral reads in sample 6C, 8C, and 13C (Figure A.19C). The major constituents ($> 90\%$) of the viral reads in these three samples were unannotated (except for one OTU identified as *Chryseobacterium* phage) and only detected in a small number of samples.

The filtered dataset contains 99 bOTUs and 48 vOTUs (Figure 3.1). There is a distinct difference between skin samples and wound samples. For bacterial OTUs, the skin samples are more diverse with more unique OTUs detected; for viral OTUs, the wound samples are more diverse with more unique OTUs. While most of samples contain at least 10^4 reads, sample 15A, 15B, 18A contain $< 10^3$ bacterial reads and sample 6C,

8C, 11C, 13C, 16C contains $< 10^3$ viral reads. Nevertheless, as all samples contain "prevalent" bacterial and viral OTUs, we did not remove them for downstream analyses.

3.2.2 Construct microbial correlation network using GGM

We proposed a conditional Gaussian Graphical model with centered log-ratio transformation (`CLR-cGGm`) to construct the (conditional) correlation networks between the 99 bacterial OTUs and 48 viral OTUs from the filtered dataset.

GGMs assume a multivariate Gaussian distribution for the observed data. To assess whether the data follow the normality assumption and compare different data transformation methods, we examined the marginal distribution of the relative abundance and CLR-transformed abundance for each OTU using Shapiro-Wilk normality test. Because the mean abundances of skin or wound samples for each OTU can be different, we also compared the data before and after removing the covariate effects by centering in each type of samples. The relative abundances of OTUs in general do not follow a normal distribution as most of the OTUs yielded small P-values to reject the null hypothesis (data is normally distributed), as indicated by the x axis of Figure 3.2A. Separately centering the relative abundance in skin and wound samples also did not improve the normality. In comparison, less OTUs have their marginal distributions rejected by Shapiro-Wilk test after the abundances are CLR-transformed and separately centered in skin and wound samples (Figure 3.2B). We also examined some example marginal distributions to confirm the normality (Figure A.20). Similar results were observed using d'Agostino-Pearson normality test (Figure A.21) and Kolmogorov-Smirnov normality test (Figure A.22). These results suggested an improved normality of the data when using CLR transformation compared with using relative abundance.

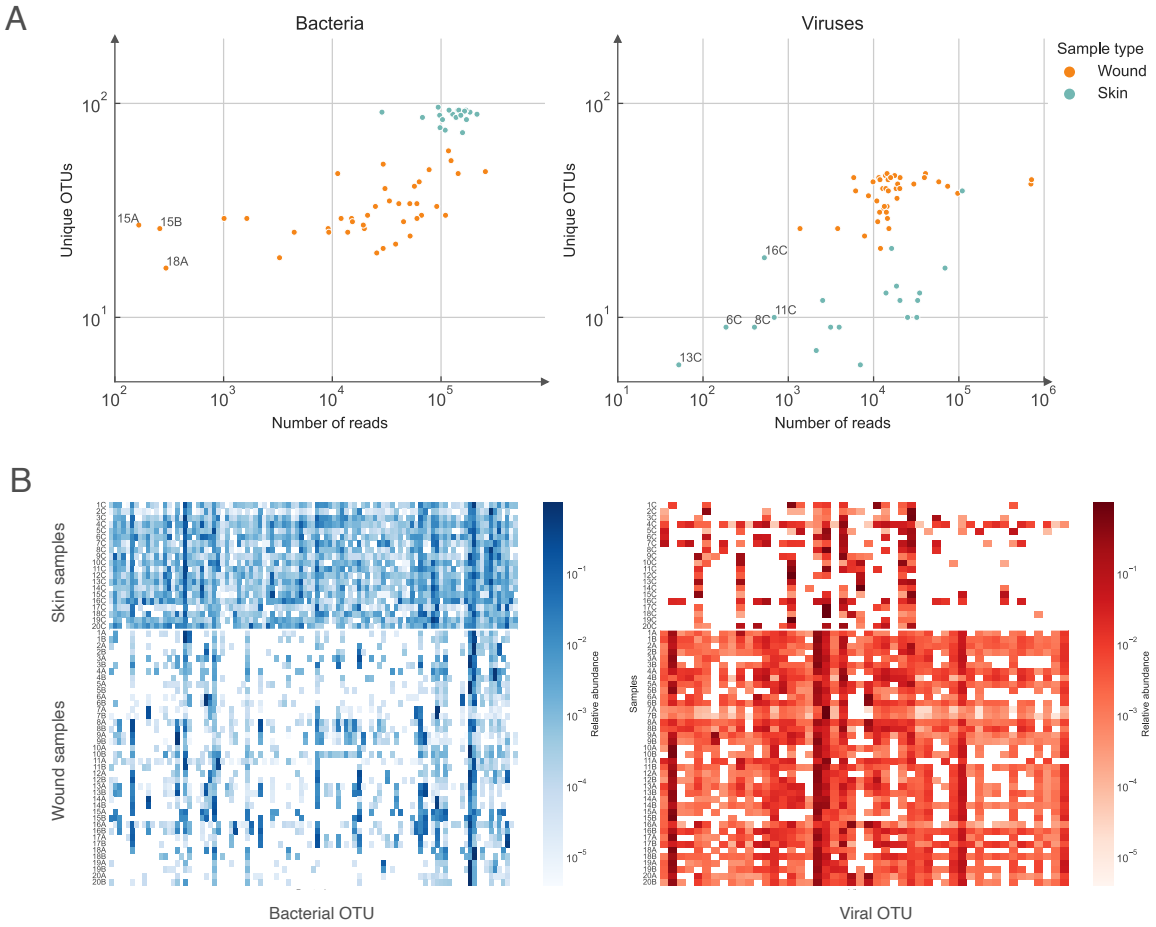


Figure 3.1: Overview of the skin microbiome dataset after OTU cleaning and filtering. (A) Scatter plots of the number of unique OTUs vs. the total number of reads for bacterial and viral content in each sample. Most of bacterial (51/60) and viral (43/60) samples contain at least 10^4 reads after filtering. Samples with $< 10^3$ reads are labeled in each figure. (B) Heatmaps of relative abundance of OTUs in each domain. Columns represent bacterial or viral OTUs and rows represents samples, ordered by sample types (skin vs. wound). There is a distinct difference between the distribution patterns of OTUs between wound and skin samples for both bacteria and viruses.

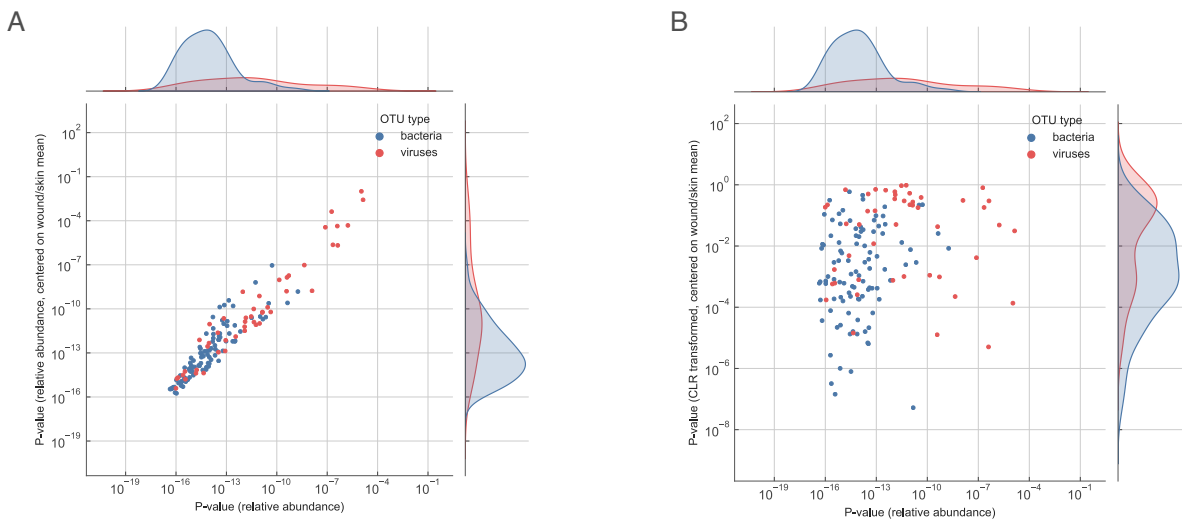


Figure 3.2: Comparisons of the normality for the marginal distributions on OTUs when using different data transformation methods. The P-values from the Shapiro-Wilk normality tests on the marginal distributions of each OTUs when using different data transformations that (A) relative abundance vs. relative abundance centered within wound and skin samples; (B) relative abundance vs. CLR transformed data and centered within wound and skin samples. A low P-value means the test reject the null hypothesis that the marginal distribution is normal.

To obtain a comprehensive view on the graph inference results with varying L1 penalty (λ), we inferred a series of models using `CLR-cGGM` with different λ values between 8 and 0.3 and examined how the graph changes across the results. As showed in Figure 3.3AB, when the value of λ decreases, the penalty relaxes and more edges are inferred. For each edge, the partial correlation also increases to plateau or slightly decreases when λ decreases. As the λ relaxes, edges detected in graphs with larger λ values still tend to have higher partial correlations than the edges inferred only in graphs with small λ values. This indicates that despite different number of edges are inferred with different λ, the strongest correlation signals from the model are consistent across different choices of λ. When comparing the edges with different domains, more intra-domain edges (bOTU-bOTU and vOTU-vOTU) edges were inferred than inter-domain edges (bOTU-vOTU)

for both the number of edges and the density in the type of edges. Despite lower number of edges, the vOTU-vOTU is more dense than bOTU-bOTU edges for most λ values (Figure 3.3B).

We also examined the connectivity of inferred graph. Transitivity and average clustering coefficients both quantify the local connectivity of a graph by comparing the number of triangles (three fully connected nodes with three edges) and triads (three nodes with two edges) in a graph. Higher values indicate the graph are more clustered locally. Both transitivity and average clustering coefficients of the graph decrease when λ changes from 8 to 1 and start to increase when λ changes from 1 to 0.3 (Figure 3.3C), suggesting the graph first expands to include more nodes by adding the edges between the unconnected node and a node in the connected graph, and then start to add more edges between the nodes in the connected graph. It is further confirmed by the size of the largest components and the number of unconnected nodes. There are only ≤ 3 connected components exist in the graph when relaxing λ and all nodes are included in a single main component when $\lambda < 1$ (Figure 3.3D).

In this analysis, we found that the edges with the strongest correlation signals relative consistent to the choice of λ values (except for too large λ values that only a few edges were inferred), however, the structure of graph may change greatly as the λ changes. A small λ value can result in a dense graph difficult for structural analysis.

3.2.3 Two-step model selection to control false discovered edges

In order to balance the signals (inferred edges) and noises from the inference, we applied a two-step model selection process on first choosing an λ and second apply post-hoc edge filtering to remove edges with weak correlations.

We investigated the performance of several data-driven model selection criteria to

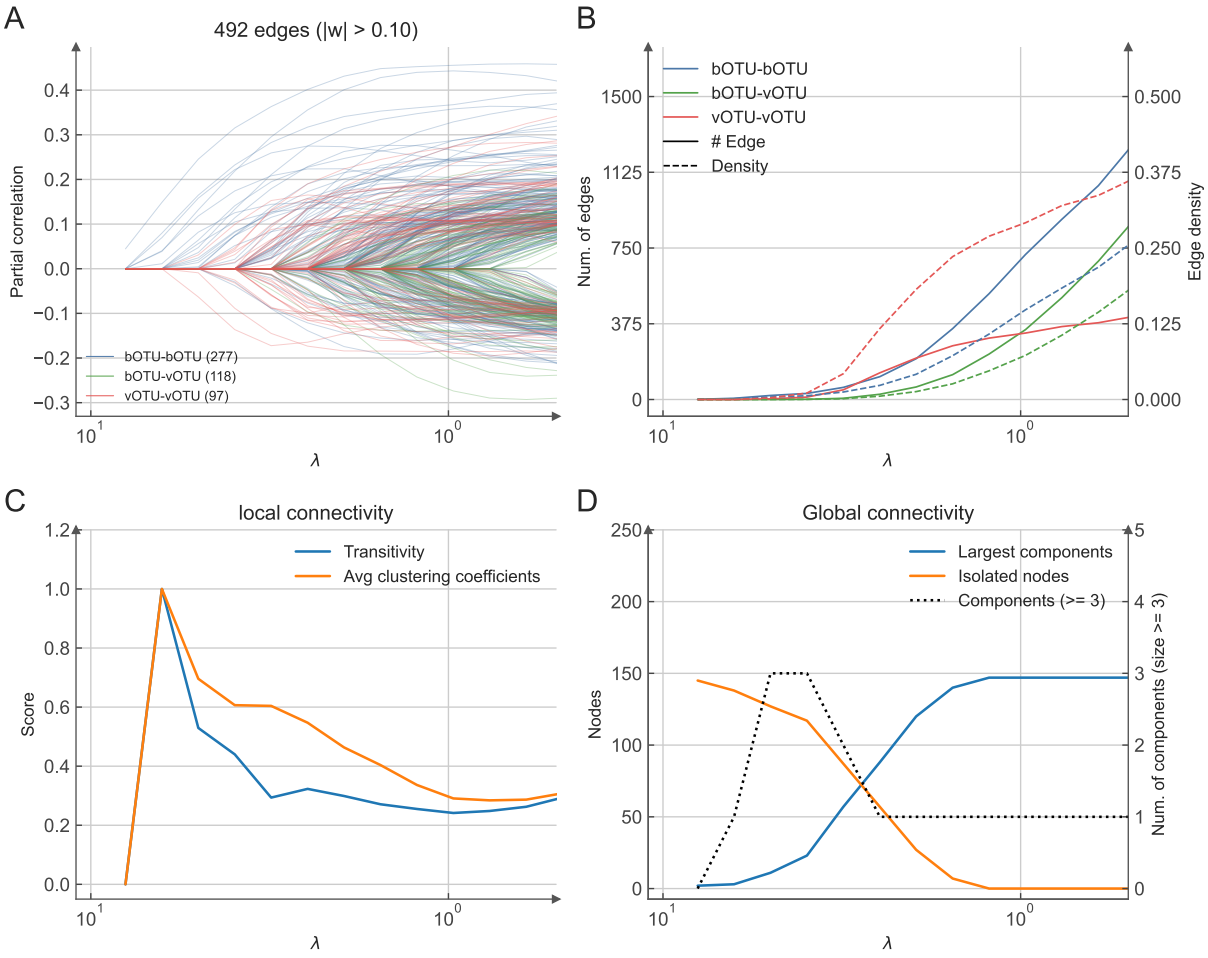


Figure 3.3: Characteristics of inferred graphs from `CLR-cGGM` with respect to different choices of L1 penalty weight (λ). (A) The change of edge weight for each edge with different λ values. 277 bOTU-bOTU edges, 118 bOTU-vOTU edges, and 97 vOTU-vOTU edges have absolute value > 0.10 for some choice of λ and are shown; (B) the number of edges (solid line) and the density of edges (dashed line) for intra–domain edges (bOTU-bOTU, vOTU-vOTU) and inter-domain (bOTU-vOTU) edges; (C) local connectivity evaluated using transitivity and average clustering coefficients for connected nodes; (D) the size of the largest components, the number of isolated (unconnected) nodes, as well as the number of components (dashed line).

determine λ, including likelihood-based criteria (AIC, BIC, EBIC), 5-fold stratified cross validation, and stARS, see Section 3.5.4 for details. Likelihood-based criteria penalize the log likelihood of a fitted model by model complexity (number of edges and other parameters) and select the λ maximizing the criteria. As showed in Figure 3.4A, extended Bayesian information criteria (EBIC-FD, EBIC-PLN) apply the strongest penalties on the number of edges and select larger λ values with sparser graphs. However, roughly $< 50\%$ OTUs are included in the graph and might only provide limited structural information about the network. On another extreme, AIC selected the most dense graph that might include too many false edges. In comparison, BIC provided the most suitable level of L1 penalty where most of nodes are connected in the graph but the graph is not too dense with many potential false edges. We also performed 5-fold cross validation (CV) for model selection stratified on sample types (wound and skin samples are each separated into 5 splits and combined to keep the ratio of sample types in each fold). The best model is the one with highest log likelihood on the holdout test sets averaged over 5 splits. Similar to BIC, cross validation selected a reasonable model with most of the nodes are connected but the graph is not too dense (Figure 3.4B). The stability-based method stARS is also examined with 100 repeats and using 0.9 as the stability threshold. stARS also yielded reasonable choice of λ (Figure 3.4C), however, it is the most computationally heavy method due to the large number of repeated inference.

The graph inferred with a selected λ value (e.g., using CV, BIC, or stARS) might still contain many edges with weak conditional correlations, which are more likely to be false edges. To further control the discovery of false edges, we applied a post-hoc edge filtering step to remove the edges with absolute values of correlations less than a threshold. The threshold τ is determined by performing multiple regressions on the filtered graph and selecting the τ values to minimize the mean sum-of-squared error in stratified 5-fold cross validation, following the process proposed in [78]. We examined this post-hoc edge

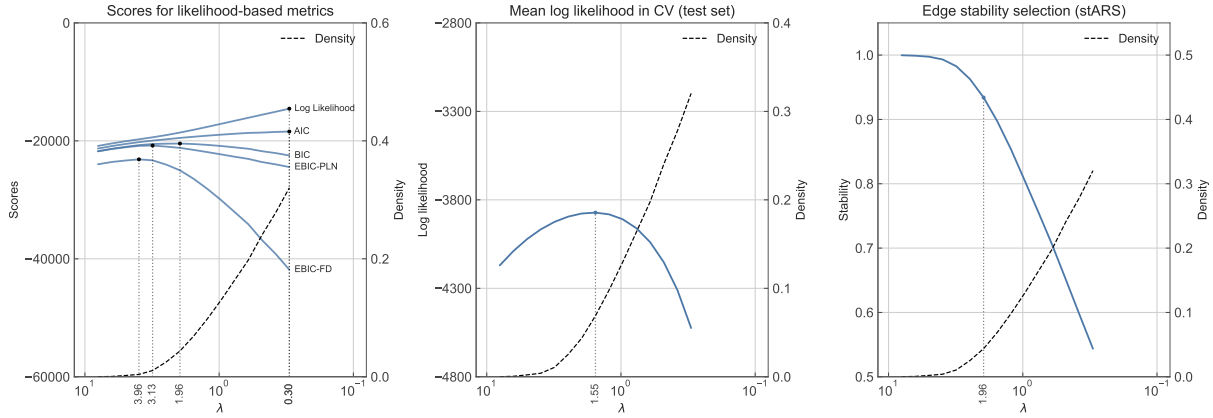


Figure 3.4: Model selection criteria on λ, including (A) likelihood-based criteria, (B) cross-validation, and (C) edge stability selection (stARS). The λ value selected by each criterion was marked on the curve. 5-fold cross validation was stratified on the sample types (wound or skin samples). The 100 repeat runs were performed for stARS selection and the selection stability threshold for stARS is 0.9.

filtering on graphs inferred with different λ values. As showed in Figure 3.5A, despite a smaller λ value leads to a denser graph, a higher value of τ is selected and thus more edges are removed. By applying this post-hoc edge filtering step, the densities of filtered graphs are kept under 0.05 even with largest λ (Figure 3.5B).

To understand to what extent this post-hoc edge filtering step changes the structure of interred graph, we compared the degree of nodes before and after the post-hoc edge filtering for selected λ values. While removing edges decreases the degrees, the order of node degrees before and after the post-hoc edge filtering remains well-aligned (Spearman correlations $r > 0.5$) for all selected λ values.

The local connectivity and the modularity of the graph before and after the edge filtering were also examined. With post-hoc edge filtering, the transitivity and average clustering coefficients do not increase when λ decreases below 1, indicating that some edges forming the triangles are likely to have lower correlation values and are removed, as showed in Figure 3.6D. And the modularity, a metric evaluating how modular a graph is, decreases slower as λ decreases when applying the post-hoc edge filtering. The boundary

between the clusters in the graph might be difficult to detect if too many edges are inferred with smaller λ values (thus, decreasing modularity) and the post-hoc edge filtering helps to keep the community structure clear (if exists) in the microbial network.

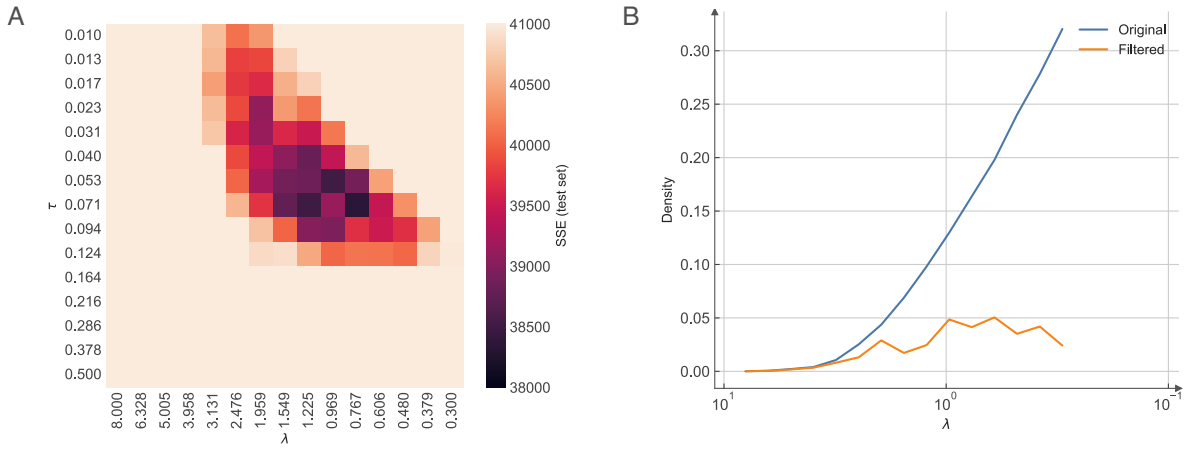


Figure 3.5: Post-hoc edge filtering for graphs inferred from `CLR-cGGM` model. (A) Heatmap of sum of squared errors (SSE) of multiple regression on the filtered graph using stratified 5-fold cross-validation, with different λ values for the graph inference and different τ for filtering. Higher thresholds are determined by CV when applying on a graph inferred with smaller λ values. (B) Effect of post-hoc edge filtering on the density of the graph inferred varying λ values. The threshold τ is determined by CV. Post-hoc edge filtering step controls the densities of filtered graphs to be less than 0.05 even with low λ values.

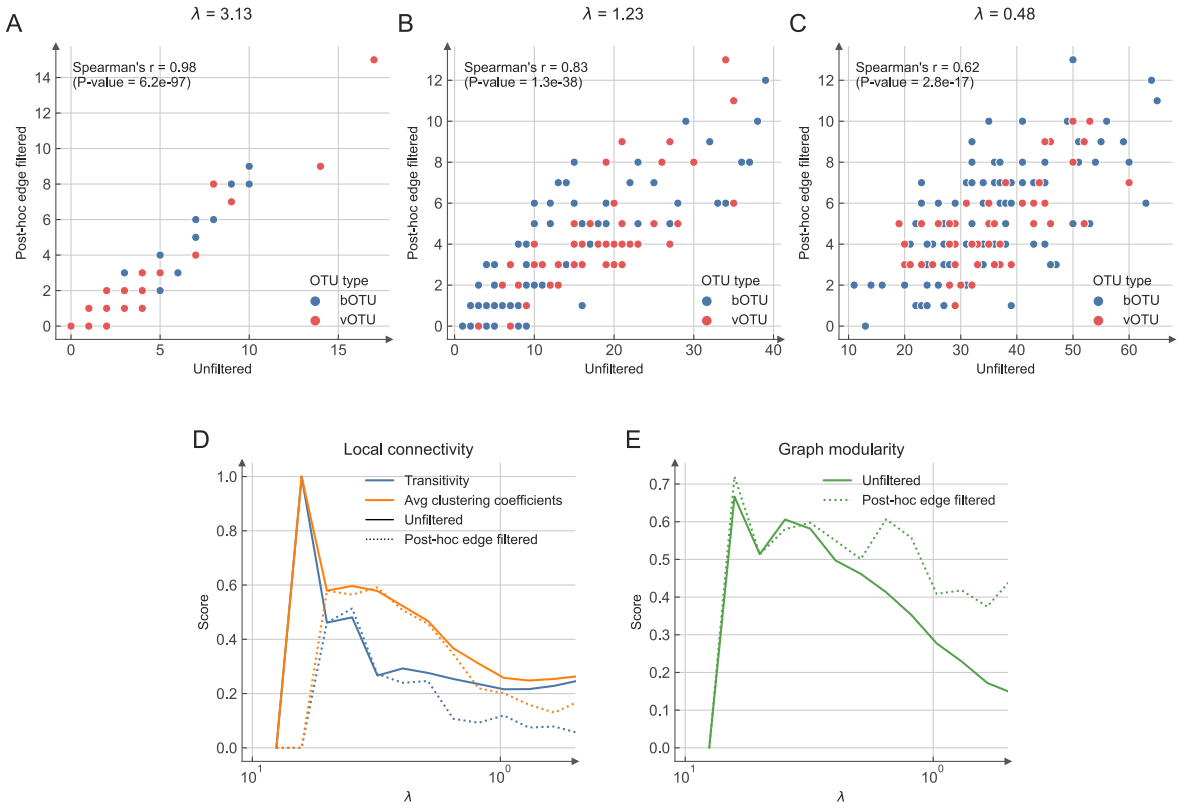


Figure 3.6: Effect of post-hoc edge filtering on node degrees (A - C), graph local connectivity (D), and graph modularity (E). (A) - (C) The node degrees of graph before and after the post-hoc edge filtering are compared. Despite lower node degrees for the filtered graph, good alignments are found for node degrees before and after the filtering. (D) Local connectivity does not increase as λ decreases below 1.0 after the filtering that less "local triangles" are formed. (E) The modularity of filtered graph decreases slower with post-hoc edge filtering, suggesting the community structure of the graph might be better preserved.

3.2.4 Key structures of the skin microbiome correlation network

The final skin microbiome correlation graph was inferred using `CLR-cGGM` with 5-fold stratified cross validation for λ selection and post-hoc edge filtering was applied with τ determined from 5-fold stratified cross validation.

The top edges with the strongest correlations in the network were examined. We verified that the relative abundance for each pair of the OTUs from the top edges also showed strong marginal correlations (Figure 3.7 and 3.8). In general, there are more strong positive correlations edges than negative edges that the top 15 edges are all positive (Figure 3.7) and the 15-th strongest negative edges is ranked 63 overall (Figure 3.8). Almost all strongest edges were intra-domain edges (i.e., bOTU-bOTU, vOTU-vOTU) rather than inter-domain edges (bOTU-vOTU), except for *Neisseria* phage-*Sphingomonadaceae* (Family), *Enterobacter* phage - *Staphylococcus aureus*. We noted that, based on the assumption of GGM, all zero counts were considered "below the limit of detection" and assigned a pseudo count of 0.1 for CLR transformation. If two OTUs both have zero count in a sample, they will have identical abundance and that might lead to a "inflated" positive correlation or "decreased" negative correlation (see data points at the diagonal in the plots). See Discussion 3.3.1 for more details.

We next investigated the nodes with high degree or betweenness centrality in the graph. Node degree counts the number of edges connecting to a node. We found the most connected bacterial OTUs in the networks include *Porphyromonas*, *Campylobacter*, *Peptoniphilus*, Helcococcus, and *Bacteroides*, and the most connected viral OTUs include *Staphylococcus aureus* phage, *Staphylococcus haemolyticus* phage, *Enterobacter* phage, and two unannotated viral OTUs (Figure 3.9). Interestingly, most connected OTUs are represented $> 0.1\%$ on average in the samples but are not the ones with highest relative abundance. Betweenness centrality measures the importance of a node in connecting different parts of a graph. We found the bacterial OTUs with the highest betweenness centrality were *Porphyromonas*, *Campylobacter*, *Bacteroides*, *Peptoniphilus*, and an unannotated OTU; and the viral OTUs with highest betweenness centrality were *Staphylococcus aureus* phage, *Enterobacter* phage, *Staphylococcus haemolyticus* phage, *Campylobacter* phage, and an unannotated vOTU. The OTUs that were found with

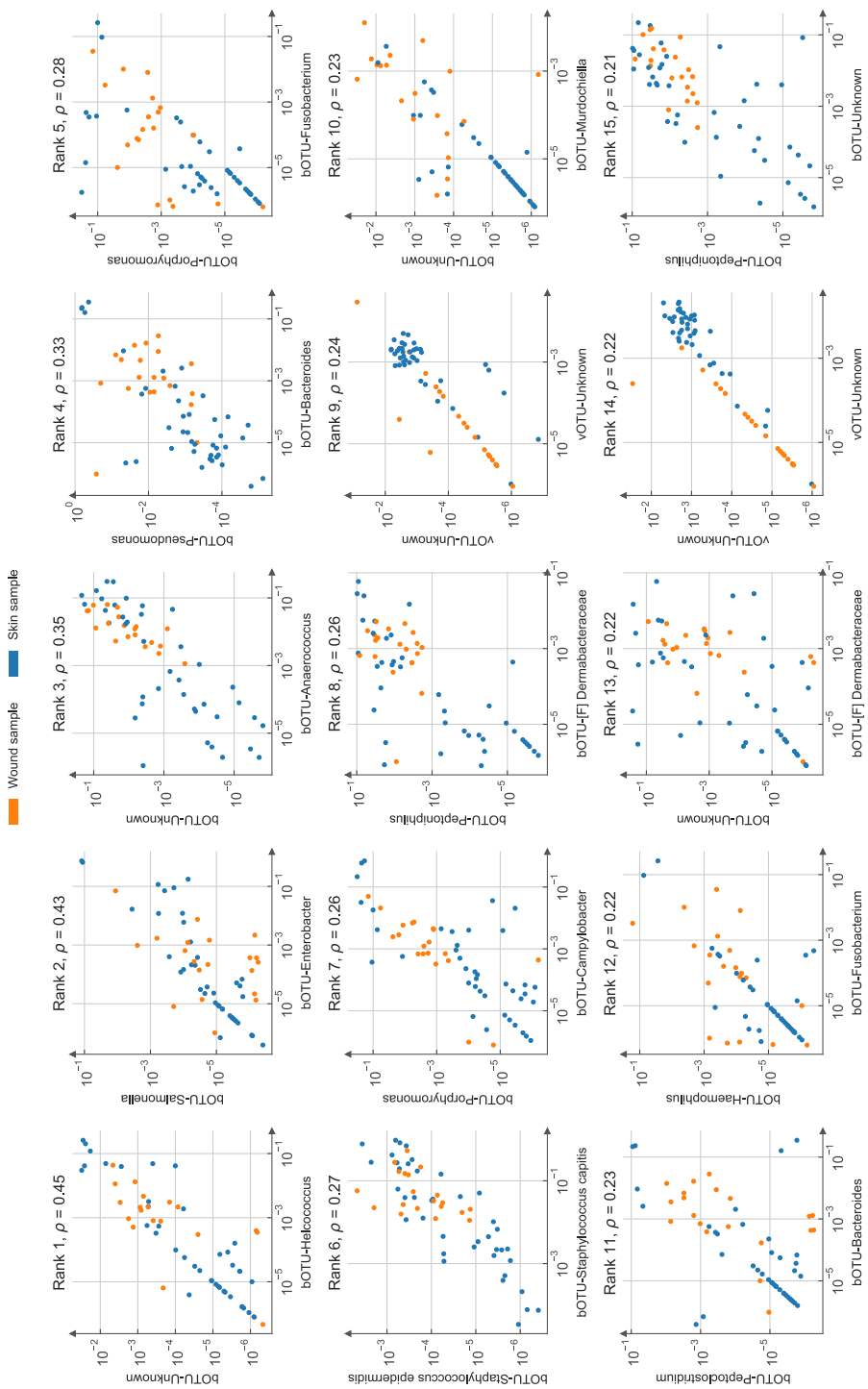


Figure 3.7: Marginal distributions of the relative abundance for each pair of the nodes in the top 15 edges with positive correlations. The rank of each edge and inferred conditional correlation are shown on top of each plot. For the purpose of visualization, pseudo count of 0.1 is added before calculating the relative abundance to avoid 0 on the logarithmic axes. Samples with zero counts for both OTUs are shown at the diagonal. OTUs without taxonomic annotations are labeled as "Unknown".

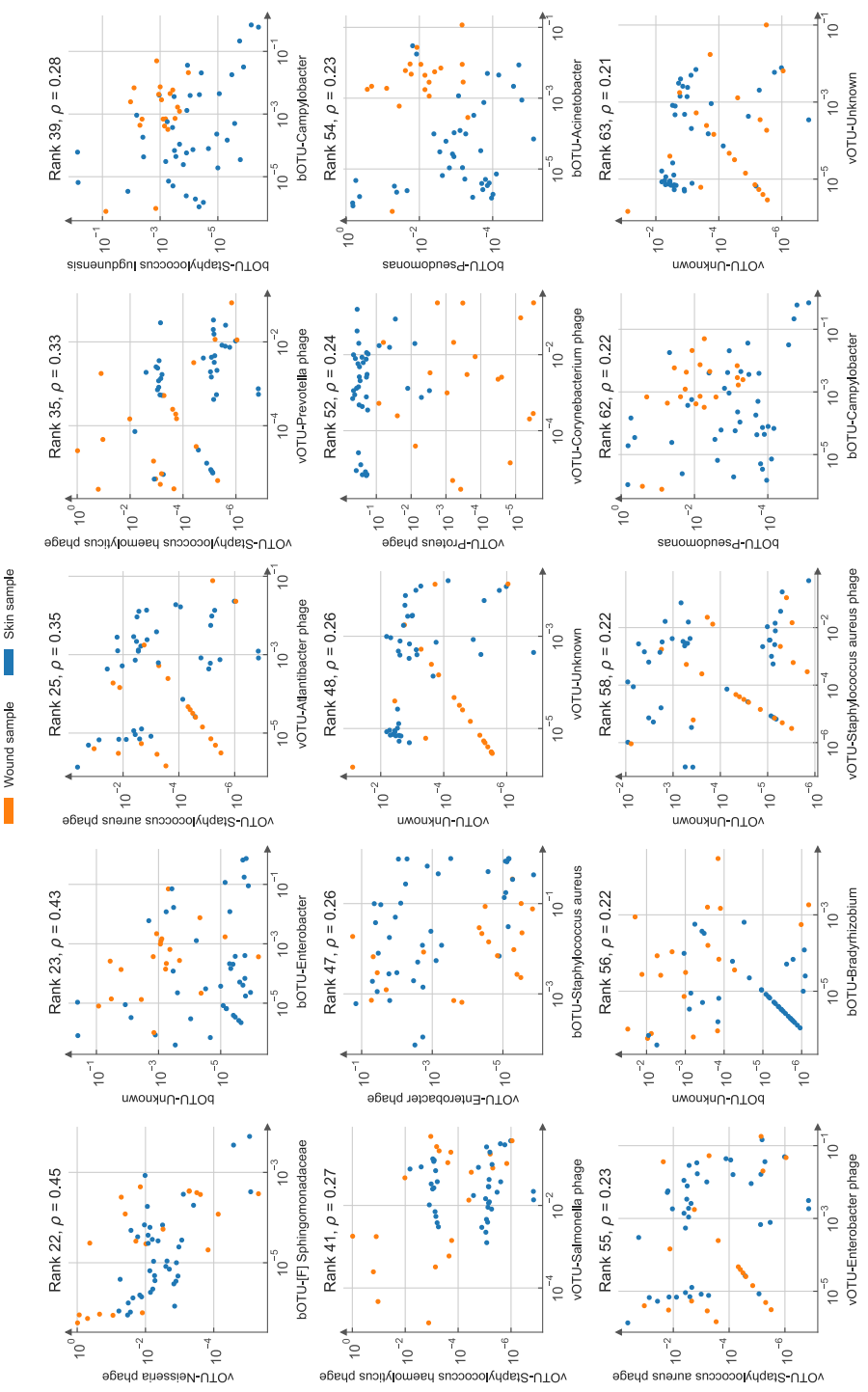


Figure 3.8: Marginal distributions of the relative abundance for each pair of the nodes in the top 15 edges with negative correlations. The rank of each edge and inferred conditional correlation are shown on top of each plot. For the purpose of visualization, pseudo count of 0.1 is added before calculating the relative abundance to avoid 0 on the logarithmic axes. Samples with zero counts for both OTUs are shown at the diagonal. OTUs without taxonomic annotations are labeled as "Unknown".

highest betweenness centrality (*Porphyromonas*, *Campylobacter*, *Staphylococcus aureus* phage, *Enterobacter* phage, *Staphylococcus haemolyticus* phage) were also found having high degrees, indicating their potential importance in the topological structure of the network.

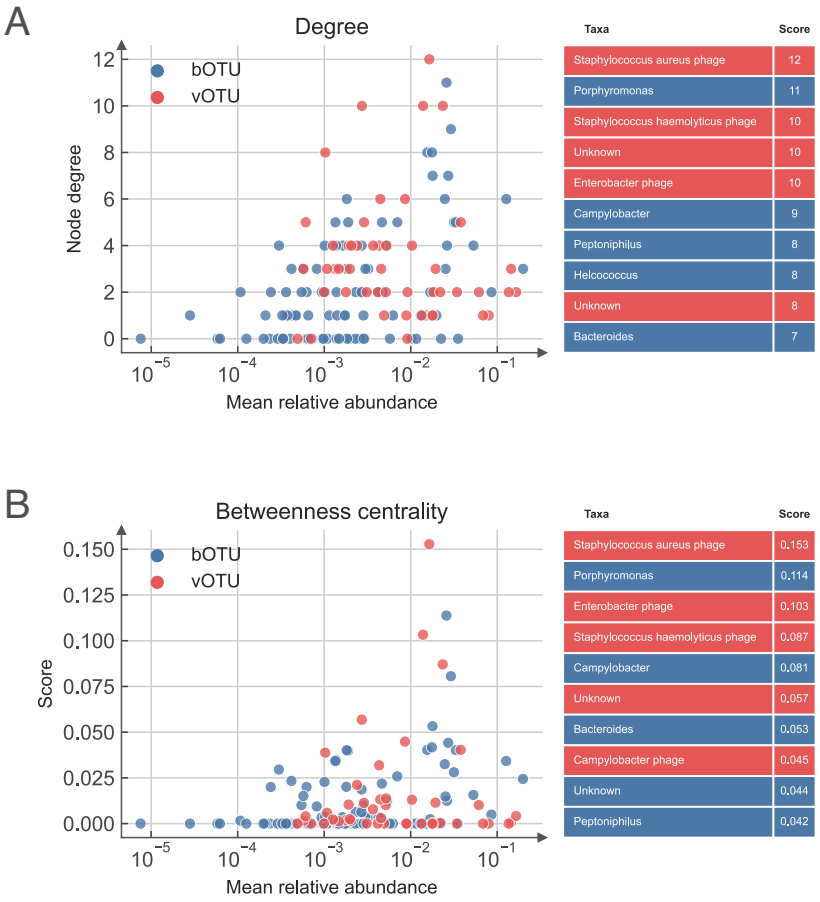


Figure 3.9: Top central nodes identified with highest (A) node degree and (B) betweenness centrality.

Lastly, we investigated the overall network structure to identify potential node clusters. Two types of node clustering schema using Louvain were applied: the first type of clusters were identified by perform clustering on the inferred graph with all edges for both positive and negative correlations (unweighted). The clusters identified are topological groups closely connected in the network, and we call these "Topological Groups". The

second type of clusters were identified by clustering on the graph with only the edges of positive correlations. Clusters identified in this schema contain mostly positive-correlated nodes, and we call them "Clusters" (Table A.3). A full list of OTUs and their classification in "Topological Groups" and "Clusters" are listed in Table A.4. As shown in Figure 3.10, 8 Topological Groups (Figure A.24) and 9 Clusters (Figure A.25) were identified by Louvain with the default setting of parameters. 109 OTUs are included in the main component of the graph. 36 OTUs are not connected by any other OTUs nodes, and two bOTU *Staphylococcus cohnii* and *Staphylococcus pettenkoferi* form an isolated pair.

OTUs in each "Clusters" are mostly positively correlated, we further investigated the relations between these "Clusters" for a higher level understanding of the network structure. As shown in Figure 3.11, among major clusters (Cluster 1 - 6), Cluster 1 is closely connected to Cluster 2 with negative correlations and to Cluster 6 with a mixture of positive and negative correlations. Cluster 2 is also closely connected to Cluster 6 with both positive and negative correlations. Cluster 3 is negatively corrected with Cluster 4 and 5 while Cluster 4 is positively correlated with Cluster 5 with one edge.

Confirmed with previous results, OTUs with high centrality scores were found important in the network structure. The bacterial OTUs of *Porphyromonas*, *Campylobacter*, and *Helcococcus* form the central structure for bacterial OTUs in Cluster 1 and also connect to Cluster 2, 5, 6, and 7. The bacterial OTU of *Peptoniphilus* seems to be the center of Cluster 6 and connects with the OTUs in Cluster 1 and Cluster 5. The viral OTUs of *Staphylococcus aureus* phage and *Staphylococcus haemolyticus* phage are part of Cluster 5 and negatively correlated with groups of viral OTUs in Cluster 2 and 3, as well as several isolated OTUs. The viral OTU of *Enterobacter* phage was found important to Cluster 2 and negatively correlated with Cluster 4 and 5.

The bacterial and viral OTUs in the networks are mostly separated but connected by some bridging OTUs. Cluster 1, 2, and 6 are the largest clusters containing mostly

bacterial OTUs; Cluster 3, 4 contain only viral OTUs and Cluster 5 contain both bacterial and viral OTUs. The frontier between bacterial and viral OTUs contains many edges with known host-viral relations. The viral OTUs either directly connect to its host (*Enterobacter* phage - *Enterobacter*, *Campylobacter* phage - *Camphylobacter*, *Streptococcus* phage - *Streptococcus*, and *Staphylococcus* phage - *Staphylococcus capitis*) or connect to a neighbor OTU of its host (*Pseudomonas* phage connects to *Bacteroides* which is a neighbor of *Pseudomonas*). Some of these viral OTUs at the frontier (e.g, *Enterobacter* phage, *Campylobacter* phage) are the highly central OTUs that connected to other viral OTUs.

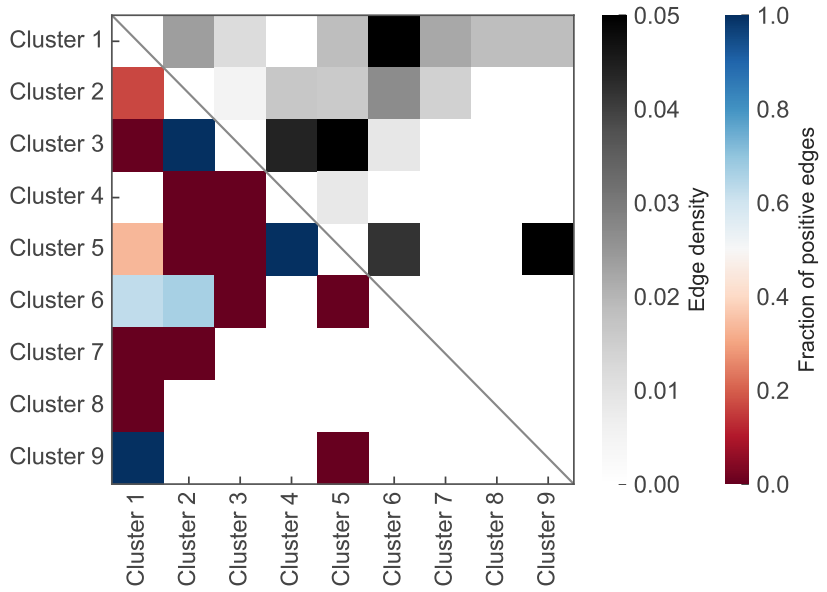


Figure 3.11: Top central nodes identified with highest (A) node degree and (B) betweenness centrality.

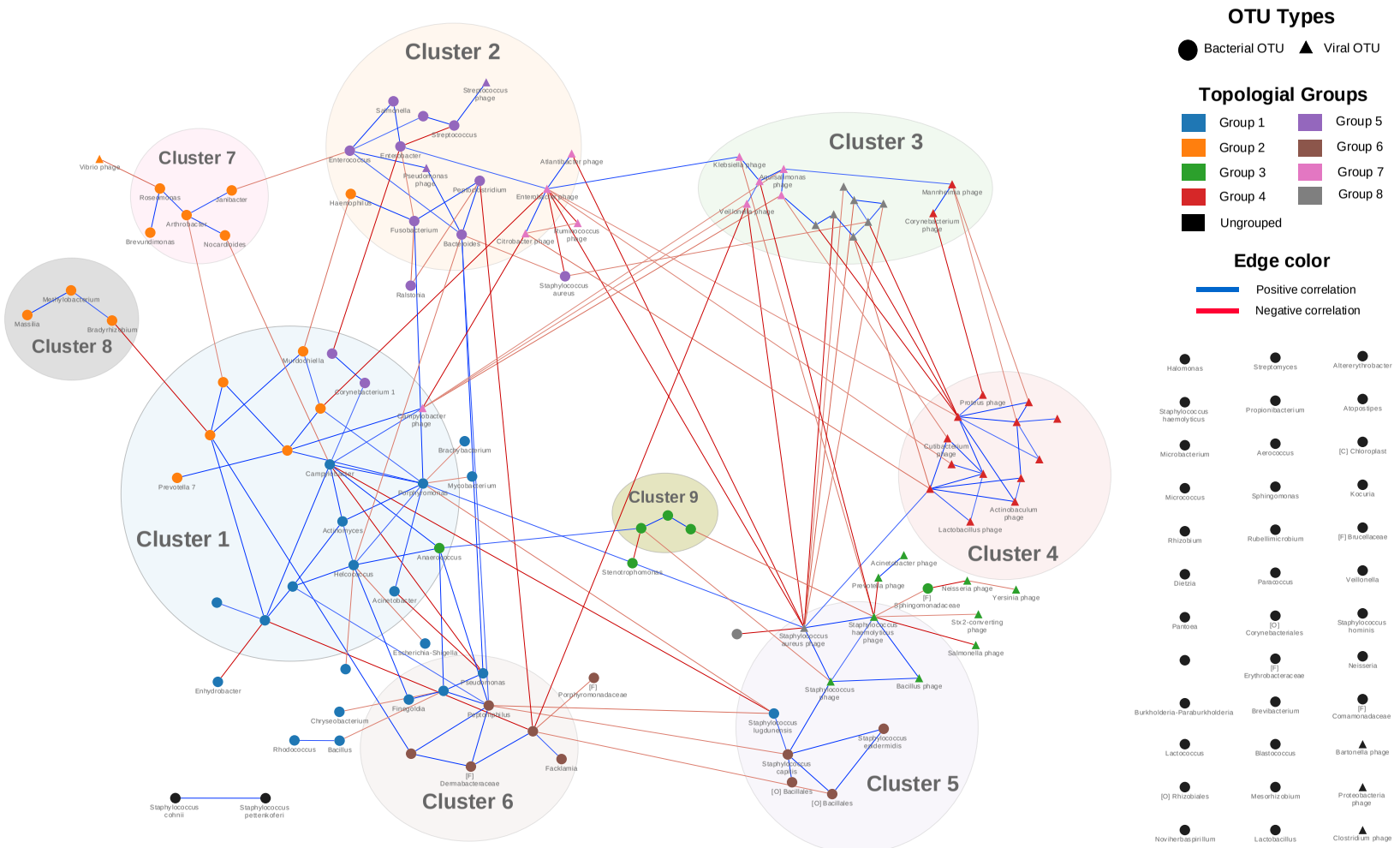


Figure 3.10: Microbial correlation network inferred using `CLR-cGGM`. The sparsity penalty λ and the post-hoc edge filtering threshold τ were each selected using cross validation. The OTUs are grouped in circles by their "Cluster" and colored by their "Topological Group". Colored nodes that are not in any circles represents for the OTUs do not belong to any Cluster but have assigned Topological Group (e.g, only has edges with negative correlations).

3.3 Discussion

3.3.1 Treatments on the metagenomic data for GGM models

The the multi-domain compositional count data collected from presented metagenomic study on the chronic wound skin microbiome contain special statistical structures that prevent the direct use of GGM to infer the correlations between microbes. The study also contains environmental covariate (i.e., samples from wound vs. skin) that would confound with the inference if not treated properly. As a solution, we carefully prepared the dataset and designed the `CLR-cGGM` model for inference.

The raw dataset was cleaned using multiple bioinformatic steps. First, potential viral contaminants and viruses infecting human or other eukaryotic hosts were removed. As samples were collected from human skin, human viruses constituted a large fraction of viral reads (especially in wound samples) and the removal of these eukaryotic viral OTUs decreased the dimension (number of variables) of the problem and avoided potential noisy correlations from these viral OTUs. `Decontam` detected the putative contaminants by identifying OTUs that were statistical invariant across different samples [91] and removing these OTUs neither were likely to influence on the results. Despite a noticeable fraction of reads was removed, the remainder is more focused on the purpose of the study to understand the relations between bacteria and bacteriophages in the chronic wound skin microbiome. Second, OTUs were aggregated at genus level to reduce error and noise. The reads from HTS were clustered as OTUs and the actual resolution of OTUs dependents on many factors, including the region of genome sequenced, the quality of sequencing, and the bioinformatic pipelines. In this dataset, the V1-V3 region of 16s rRNA gene was sequenced and the OTUs were clustered at 97% sequence similarity which were more proper to represent taxa at the genus level for the majority of the genera. Specially, *Staphylococcus* species have been found that can be distinguished by

97% sequence similarity in the V1-V3 region [74, 75] and additional bioinformatic steps were added to annotate the species correctly. For the purpose of analysis, the viral OTUs were also aggregated to match the resolution of their hosts. Lastly, a prevalence filter is applied to remove OTUs only found in a small number of samples and to focus on a small "core set" of OTUs yet constituted the majority of cleaned OTU reads. Removing OTUs only detected in a few samples also helps remove the excessive zeros in the data. In our model, zeros were treated as "below the limit of detection" and a small pseudo counts 0.1 were added for analysis. OTUs with zero counts are treated as having same abundance in these samples and lead to "false" positive correlations as observed in some examples in Figure 3.7. While it is impractical to remove all OTUs with zero count, we chose to filter the OTUs by their prevalence across different sample sources (wound or skin, different patients). This filtered allowed us to drastically reduced number of OTUs, keep most of total reads in samples, and minimize the effect of "false" positive correlations from excessive zeros (despite not able to completely remove such effect).

In addition to data filtering and cleaning, the model we used `CLR-cGGM` was carefully designed to account for structures of the dataset. On top of classic GGM model, we applied two modifications for our dataset: 1) separate centered-log ratio transformation for multi-domain compositional data and 2) use conditional GGM to adjust the mean (μ) for data from different sample types. While the absolute quantitation of microorganisms is of researchers' interests [29], neither the absolute quantitation nor the total amounts were available for this dataset. Directly infer the correlations from the relative abundances can lead to spurious correlations due to constraint of all relative abundances needed sum up to a constant. Alternatively, it has been proved that the covariance matrix of log transformed relative abundances can approximate the covariance matrix for the log transformed true abundance [30, 69]. An extension of this approximate to multi-domain data has also been proposed [72] and we adopted this CLR transformation for our dataset

which contains compositional data from bacterial and viral domain. Previous studies on this dataset have showed that some bacterial OTUs had different abundances in wound and skin samples [54] and the pattern of relative abundances for bacterial and viral OTUs were different in these two types of samples (Figure 3.1B). Sample type is a confounding variable to the abundance of some OTUs and might be attributed as false correlations between these OTUs. A linear model is used to offset the difference of μ for two different type of samples to remove the effect of this confounding covariates. Combining above two adjustment, we proposed to use `CLR-cGGM` to infer the correlation network in our dataset.

3.3.2 Two-step model selection to regularize the edge discovery

The goal of the study is to infer a parsimonious representation of correlations between OTUs. Real-life biological data are usually noisy and having small sample sizes. Giving the larger number of possible node pairs ($p(p-1)/2$, p is the number of nodes), detecting real correlations is an imbalanced problem and susceptible to false discovery even with a small false positive rate. The inference of precision matrices needs to be carefully regularized in order to discover the signals from noisy data. An L1-penalized maximum likelihood estimation named Gaussian graphical LASSO (GLASSO) is applied to infer a sparse precision matrix with its density controlled by the parameter λ. Edges with small absolute correlation values are further removed using a post-hoc edge filtering step. Using this two-step model selection process, we applied a stringent regularization on edge discover without hurting the discovery of graph structural characteristics.

Selecting λ value is an active research area and the choice of selection criteria is still subjective and guided by practices [69, 71]. We explored three categories of commonly used selection criteria: likelihood-based metrics, cross validation, and stARS. Likelihood-

based metrics penalize the log likelihood of fitted Gaussian distribution with model complexity (number of parameters of the model, including the number of inferred edges). For example, AIC penalizes $\log \mathcal{L}$ by number of parameters k, BIC penalizes with additional weight $0.5 \log n$ on k, EBIC metrics are proposed for high-dimensional problems (e.g., a graph model) to further penalize the size of the model space (i.e., number of possible combinations to choose edges). As expected, AIC is not designed for high-dimensional problems and selected dense networks in our experiments. While EBIC-FD and EBIC-PLN are specifically designed for model selection on graphs, these two criteria selected very sparse networks with only 43 and 114 edges (< 1 edge per node on average) where most of the nodes are not connected. In comparison, BIC instead selected a graph that best align with our intuition. Compared to likelihood-based metrics, cross validation selects the optimal model from a more practical perspective and requires less theoretical assumptions. It selects the model with best goodness-of-fit on the holdout data. In practice, CV selected λ value aligned with our intuition with 1181 edges. Lastly, a popular stability-based method called stARS was also assessed. stARS selects the model based on the stability of inferred graph across data subsamples. However, repeated subsampling is computationally heavy and the selection of model still requires a "stability threshold" determined empirically. While several criteria (BIC, 5-fold CV, stARS) have showed suitable in our experiments, choosing which criterion to use is still subjective and depends on the problem.

While selecting the a criterion is relative subjective, we observed that the edge signals change smoothly and the order of edge weights is relatively stable with similar λ values. The model selection on λ can be compensated by setting a minimum edge weight threshold. We applied a post-hoc edge filtering and selected the edge weight threshold τ using cross validation [78]. We found this post-hoc edge filtering further regularizes the graph inference process. Higher threshold values were selected for models with smaller λ

and controlled the graph density to be < 0.05 for varying λ values. We further confirmed that removing these weak edges did not change the rank of node centrality significantly and maintained the structure of inferred graph. Despite that GGM learns the direct conditional correlations between variables, non-zero values can still be inferred for indirect correlations due to the noise in the data. Applying post-hoc edge filtering seems to remove such "indirect" edges with a lower mean clustering coefficients. Additionally, higher graph modularity was observed for the graph after the post-hoc edge filtering, indicating a cleaner graph structure was recovered.

3.3.3 Biological interpretation of key structural characteristics in the inferred graph

We investigated the key structural characteristics in the inferred graph, including edges with strongest correlations, nodes with high centralities, and node clusters.

Some noticeable pairs of OTUs recovered from the top edges with highest conditional correlations have shared biological properties. For example, both *Salmonella* and *Enterobacter* have been frequently found related to human gastrointestinal system and some species from these two genera have been reported in skin infections [92,93]. *Pseudomonas* and *Bacteroides* are both Gram-negative and some species such as *P. aeruginos* and *B. fragilis* are both known for antibiotic resistance and are important to skin or soft tissue infections [94, 95, 96, 97]. *Staphylococcus epidermidis* and *Staphylococcus capitis* are both considered to be common skin commensals but are also related to hospital-acquired or medical devices related infections [98, 99, 100]. The relations between species of *Porphyromonas* and *Campylobacter* are most commonly found in human oral microbiome and their co-infection has been reported in several clinical and lab studies [101, 102]. Interestingly, *P. gigivalis* has been report to antagonise the cytokines production induced by

C. rectus, indicating potential co-infection between these two species can suppress the host immune responses [101].

By identifying the nodes with high centralities, we found several bacterial and viral OTUs that have both high node degree centrality and betweenness centralities. These OTUs are important in connecting nodes in a cluster and across different clusters that are referred to as "hub" taxa [80]. These OTUs with valid taxonomic annotations include *Porphyromonas*, *Campylobacter*, *Staphylococcus aureus* phage, *Enterobacter* phage, *Staphylococcus haemolyticus* phage. Despite being not the most abundant ones in chronic wound skin microbiome samples, one might hypothesize that these taxa are more active in interacting with other microbes or important in the formation of different types of skin microbiome. However, these hypotheses can not be verified from current observational data and require further studies.

Lastly, at the system level, we found the bacterial and viral OTUs are in general organized in two different sections of graph but connected by some noticeable host-virus pairs can be verified by their taxonomic annotations. The inferred network aligns with our expectation as most of bOTU and vOTU are not found to interact based on their taxonomic annotations. And the ones found to be potentially interacts are found directly connected or close in the inferred network, providing an evidence on the potential use of the correlation network to understand the underlying relations between microbes. Another interesting structure found in the network is a cluster of taxonomic group of *Staphylococcus* species or bacteriophages infecting *Staphylococcus* species. Cluster 5 contains OTUs from *Staphylococcus lugdunesis*, *Staphylococcus capitis*, *Staphylococcus epidermidis*, *Staphylococcus* phage (not identified at the species level), *Staphylococcus aureus* phage, and *Staphylococcus haemolyticus* phage, and all these OTUs are positively correlated with each other.

3.4 Summary

In this section, we used Gaussian graphical models to construct the conditional correlations between the bacteria and viruses observed in a chronic wound skin microbiome metagenomic data. Considering the special structures in this HTS dataset, we proposed conditional Gaussian graphical model with centered log-ratio transformation `CLR-cGGM` to account for: 1) the compositionality of HTS data, 2) data from multiple life domains, each is compositional, and 3) experimental factors influence the mean abundance of OTUs in different samples. One major challenge in constructing the correlation network from noisy biological data is to balance the signal and noise detected by the model. We implemented a two-step model selection process to first select an optimal L1 penalty λ for a graph with desired density and a post-hoc edge filtering process to further remove edges with weak correlations. Both hyperparameter λ and τ are selected by a data-driven methods, K-fold cross validation. We showed this process regularize the inferred graph without losing much structural information. Lastly, we utilized graph analysis tools to identify the key characteristics in the correlation graph, including strongest correlations, OTUs with high centralities, OTUs clusters and other structural information. Some of these findings have meaningful biological interpretations and were further confirmed by previous studies. These findings support the potential utility of GGM and correlation networks to help guide the future studies on microbial interactions and the structure of microbiome.

3.5 Methods

3.5.1 Sample collection, sequencing, and bioinformatic pipelines for OTU picking

The sample collection, preparation for DNA sequencing, and bioinformatic pipelines for OTU picking for the bacterial and viral content of skin microbiome samples have been described in the previous work [49, 54]. Briefly, 20 outpatients were recruited at Ridley-Tree Center for Wound Management at Goleta Valley Cottage Hospital, with four types of chronic wounds (diabetic ulcers, venous wounds, arterial wounds, and pressure ulcers, 5 patient each). During their visit, samples from the surface of wounds before and after the debridement were collected, as well as a sample from healthy skin on the contralateral limb. Clinical swabs were placed into the dry, sterile collection tubes and store at 4 °C before being processed. Negative control samples from the wound center (WC) were collected by exposing the swab into the air in the room for the same duration as wound and skin swabs were collected. Negative control samples from the processing lab (CL) were obtained by exposing the swabs to air in the lab. A cell-based microbial mock community (Zymo) was also included as a positive control.

Samples with swab tips were resuspended in 500 µL 1x TE buffer and centrifuged to pallet cells. 250 µL supernatant was transferred for immediate viral-like particle (VLP) precipitation (viral fraction) and the remaining 250 µM supernatent, pelleted cells, and swab tips were used for whole-microbiome DNA extraction (bacterial fraction).

The bacterial fractions of 69 samples (60 clinical samples, 3 wound center negative controls, 3 processing lab negative controls, 3 mock communities) were processed as described in [54]. Cells were lysed extensively and the DNA content in each sample was extracted, purified, and quantified for 16S rRNA amplicon sequencing. In each

sample, 16S DNA library was prepared for PCR with custom adapter primers amplifying the V1-V3 regions of 16S rRNA gene. After PCR products were purified, each sample was indexed, normalized, and pooled for DNA sequencing on an Illumina MiSeq with PE300 V3 chemistry at UCSB's Biological Nanostructures Laboratory (BNL) sequencing core. Paired-end reads processed using `QIIME` [25]: reads were joined using `fastq-join` with the default settings [103] and the joined reads were then trimmed and filtered by `Trimmomatic` [104]. The open OTU picking pipeline was used with the default settings and the taxonomic annotations for OTUs were assigned using SILVA128 16S reference database clustered at 97% identity level [105].

As described in [49], the viral fractions from 66 samples (60 clinical samples, 2 wound center negative controls, 2 processing lab negative controls, 2 mock communities) were extracted. Free DNA was digested using DNase and viral-like particles were further purified, disrupted, and digested to extract the viral DNA. The viral DNA content in each sample was quantified using Qubit dsDNA HS kit. Samples with DNA concentration $>$ 0.2 ng/µL (43 out of 66 samples) were diluted to 0.2 ng/µL and samples with DNA content $<$0.2 ng/µL (23 out of 66 samples) were prepared using 'tagmentation' reaction described in [106]. All samples were indexed, normalized, and pooled for sequencing on an Illumina HiSeq 400 with PE150 V3 chemistry, using two lanes, at the UC Davis DNA Technologies Core. Reads were trimmed and quality controlled using `Trimmomatic` [104]. Trimmed reads were then joined using `PANDASeq` with the default settings [41]. All singletons (not joined) and joined pairs were classified against NCBI's Viral RefSeq database [107] and full IMG/VR database [108] using `Kraken2` [24] to identify viral OTUs. In each sample, OTU abundances were estimated using `Braken` with ideal read length of 150 bp [26]. The viral and host taxonomic annotations were abstracted from NCBI and IMG/VR with manual curation for standardization.

3.5.2 OTU cleaning, aggregation, and filtering

Putative viral contaminants were identified by `Decontam` using negative control samples as the reference and a threshold of 0.2 [91]. Additional contaminants were identified manually to exclude species, strains, or types known to be used in adjacent laboratory space [49]. Additionally, a list of viral OTUs infecting human or other eukaryotic hosts were curated manually based on the host annotation and were excluded from the analysis.

Bacterial and viral OTUs that were only detected in the negative control or mock communities samples were removed. Bacterial OTUs were aggregated at the genus level (except for *Staphylococcus spp.*) based on their taxonomic annotations. Bacterial OTUs belong to *Staphylococcus* were further annotated with their species using `blastn` and aggregated at the species level [54, 109]. Viral OTUs (bacteriophages) were also aggregated based on their host annotations. *Staphylococcus* phages were aggregated at the species level of their hosts and other viral OTUs were aggregated at genus level of their hosts. Bacterial or viral OTUs without annotations at the desired level were kept as separate OTUs. After aggregation, OTUs that were only detected in wound (pre- and post-debridement) or skin samples, were detected in $< 80\%$ of patients (16 patients) were removed from downstream analysis.

All OTU cleaning, aggregation, and filtering were conducted using `Pandas` (1.3.3) in Python (3.8).

3.5.3 Model implementation

`CLR-cGGM` was implemented in `R` (version 3.6.3) with customized scripts. A pseudo count of 0.1 was added to all counts before CLR transformation and the transformations were performed separately for each domain. The conditional GGM was implemented using `cglasso` (version 2.0.4). The CLR-transformed OTU tables from bacterial and

viral domains were concatenated and passed as argument `Y` and the covariate factors were passed as the argument `X` to `cglasso::datacggm` for inference. `cglasso` can also apply L1 penalty on the regression coefficients for the mean (argument `lambdas`) and we always set it to be 1e-6 for no penalty and the actually L1 penalty on the prevision matrix λ was passed to `rhos`, due to naming difference. The diagonal terms for the precision matrix were penalized and all other arguments were set as default.

3.5.4 Model selection (λ)

Let $\hat{\mathcal{L}}_\lambda$ be the log likelihood (Equation 3.5) for the multivariate Gaussian distribution with precision matrix ($\mathbf{\Omega}$) with given λ. Let k be the total number of parameters in the model, p be the number of variables, $|\mathcal{E}|$ be the number of inferred edges (non-zero entries in the upper triangle of $\mathbf{\Omega}$. Likelihood-based criteria used for model selection are calculated by following formulas:

- Akaike Information Criterion:

$$\text{AIC} = \hat{\mathcal{L}}_\lambda - k$$

- Bayesian Information Criterion:

$$\text{BIC} = \hat{\mathcal{L}}_\lambda - 0.5k\log(n)$$

- Extended Bayesian Information Criterion by Foygel and Dorton [76]:

$$\text{EBIC-FD} = \hat{\mathcal{L}}_\lambda - 0.5k\log(n) - 2k\log(p)$$

- Extended Bayesian Information Criterion used in [71]:

$$\text{EBIC-PLN} = \hat{\mathcal{L}}_\lambda - 0.5k\log(n) - 0.5|\mathcal{E}|\log n - 0.5\log\binom{p(p-1)/2}{|\mathcal{E}|}$$

  $\binom{n}{m}$ is the number of options to choose m elements from n elements.

In our implementation, AIC, BIC, EBIC-FD were calculated using `cglasso::AIC`, `cglasso::BIC` with default arguments suggested in `https://cran.r-project.org/web/packages/cglasso/cglasso.pdf` and multiplied the criteria scores by 0.5 to convert the values according to formulas above. EBIC-PLN is calculated directly using the formula above.

K-fold cross validation was implemented using customized `R` (3.6.3) scripts. K-fold cross validation splits the data into K random folds. For each K-1 data fold, a model is fitted and evaluated on the holdout test set by the log likelihood of fitted Gaussian. K experiments are repeated using each one of the K folds as the holdout set. The λ value maximizes the mean log likelihood on the test sets across K fitted models was selected. For K-fold stratified CV, splits are created by keeping the ratio of strata (different sample types) in each fold. In this work, we performed 5-fold CV stratified on the sample types.

Stability Approach to Regularization Selection (stARS) was also implemented in `R` (3.6.3), following Liu et. al. [77]. 80% of data are randomly sampled for a given number of times (100 in our experiments), and for each subsample, `CLR-cGGM` models are fitted with a series of λ values. For each λ, 100 graphs are inferred from sample repeats, and the fraction of graphs contain a edge between Node i and Node j is ξ_{ij}. The stability of a edge is calculated as $\xi_{ij}(1-\xi_{ij})$ and the stability of graphs inferred using a choice of λ is $1 - \frac{4\sum_{1\leq<j\leq p}\xi_{ij}}{p(p-1)}$. Given a desired stability for graph inference (usually 0.9, as recommended in [69, 71, 77]), the smallest λ value that has stability above the threshold is selected.

3.5.5 Post-hoc edge filtering

Post-hoc edge filtering was applied by removing the edges with weight (conditional correlation ω_{ij}) less than a threshold (τ). τ was determined using 5-fold stratified cross validation on multiple linear regression on the filtered graphs, following the method proposed in [78]. Let $G_{\lambda,\tau}$ be the filtered graph with edges inferred from `CLR-cGGM` with λ and edges with $|\omega_{ij}| < \tau$ were removed. A multiple regression on graph $G_{\lambda,\tau}$:

$$y_{ij} = \beta_{j0} + \beta_{j1}\mathbb{I}_{i1} + \beta_{j2}\mathbb{I}_{i2} + \cdots + \beta_{jK}\mathbb{I}_{iK} + \sum_{l \in \mathrm{Ne}(j)} b_{lj} y_{il} \tag{3.9}$$
$$\text{for } i = 1, 2, \ldots, N \text{ and } j = 1, 2, \ldots, p$$

where y_{ij} is the response variable that the CLR transformed abundance of OTU j in sample i, $\beta_{j0}, \beta_{j1}, \ldots, \beta_{jK}$ are the regression coefficients for K sample types for OTU j on the mean of the Gaussian, $\mathbb{I}_{i1}, \mathbb{I}_{i2}, \ldots, \mathbb{I}_{iK}$ are the indicator functions on sample i and sample type k that is 1 if sample i belongs to sample type k and is 0 if not. Ne(j) is a set of neighbor nodes connecting to OTU j in graph $G_{\lambda,\tau}$, b_{ij} is the regression coefficient for the abundance of neighbor node l (y_{il}). The multiple regression guided by the graph structure is performed on nodes and a τ value minimized the sum of squared errors (SSE) on the test sets in 5-fold stratified CV is selected. The method was implemented in a customized R (3.6.3) script.

3.5.6 Graph analysis

Analysis of inferred graph were performed in Python (3.8) using `NetworkX` for node degree and betweenness centrality. Edge weights were not included when calculating the centralities. The Louvain clustering was performed using `CDlib` (0.2.4). The final correlation network was visualized using `Cytoscape` (3.8.2) with Boundary Layout (`http:`

//www.rbvi.ucsf.edu/cytoscape/boundaryLayout/index.shtml).

Chapter 4

Conclusions

In this dissertation, I focused on analyzing the computational challenges in utilizing HTS data from experimental or clinical studies to understand complex chemical or biological systems, and developing solutions to more robustly quantify the biochemical or statistical properties of interests. Despite being exploratory, these works provided the knowledge on the practicality of these computational models on real-world datasets and critical factors to consider when analyzing similar data.

In the work of Chapter 2, we investigated the utility of a massively paralleled assay called k-Seq that can quantify the kinetic properties for 10^5 different ribozymes. Compared to conventional kinetic experiments, k-Seq can achieve unprecedented throughput but are compromised on the fitting quality for some of the sequences due to either their lower abundance in the pool and consequently the higher error rate in the measurement, or the experimental conditions not being optimal for these sequences. Thus, robust quantification on the fitting quality for individual sequences is particularly important in reporting the results with confidence from k-Seq type of studies. The proposed bootstrapping process takes the advantages of the sampling noise and estimates a joint distribution of fitting optima, providing richer information on the fitting quality for individual sequences. Using this method, we were able to study the effect of critical parameters on the utility of k-Seq, and how to improved the coverage of sequences with desired fitting quality. These theoretical and practical guidance can help the future practitioners to better design and optimize the k-Seq type experiments and maximize the benefits of using HTS for quantification.

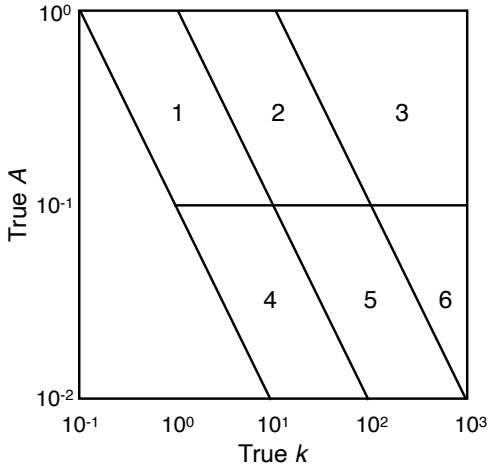
In the work of Chapter 3, we investigated the means of using Gaussian graphical models (GGM) to construct the correlation network between microbes, from a metagenomic dataset of our clinical study on the skin microbiome. We integrated three modifications on GGM to account for the special structure of data and proposed `CLR-cGGM` for the multi-domain compositional data with experimental covariates. Given the noisy clinical

data with small sample size, we applied a stringent two-step model selection process to regularize the results from the model. We were able to construct a sparse network representing the conditional correlations between the bacteria and viruses in the microbiome. We used graph analytical tools to identify key structural characteristics on the network and assessed the biological relevance for these structures.

Appendix A

Supplementary Information

A.1 Supplementary Information for Chapter 2



Region	k ($\mathrm{min^{-1}M^{-1}}$)	A	kA ($\mathrm{min^{-1}M^{-1}}$)
1	N/A	$10^{-1} < A < 1$	$10^{-1} < kA < 1$
2	N/A	$10^{-1} < A < 1$	$1 < kA < 10^{1}$
3	$k < 10^{3}$	$10^{-1} < A < 1$	$kA > 1$
4	N/A	$10^{-2} < A < 10^{-1}$	$10^{-1} < kA < 1$
5	N/A	$10^{-2} < A < 10^{-1}$	$1 < kA < 10^{1}$
6	$k < 10^{3}$	$10^{-2} < A < 10^{-1}$	$kA > 1$

Figure A.1: Illustration of the 6 regions selected to sample sequences and their fitting values from simulated reacted fraction dataset, with the boundary values for true k, A and kA indicated in the table (N/A = not applicable). Example sequences from these regions are in Figure A.2-A.7

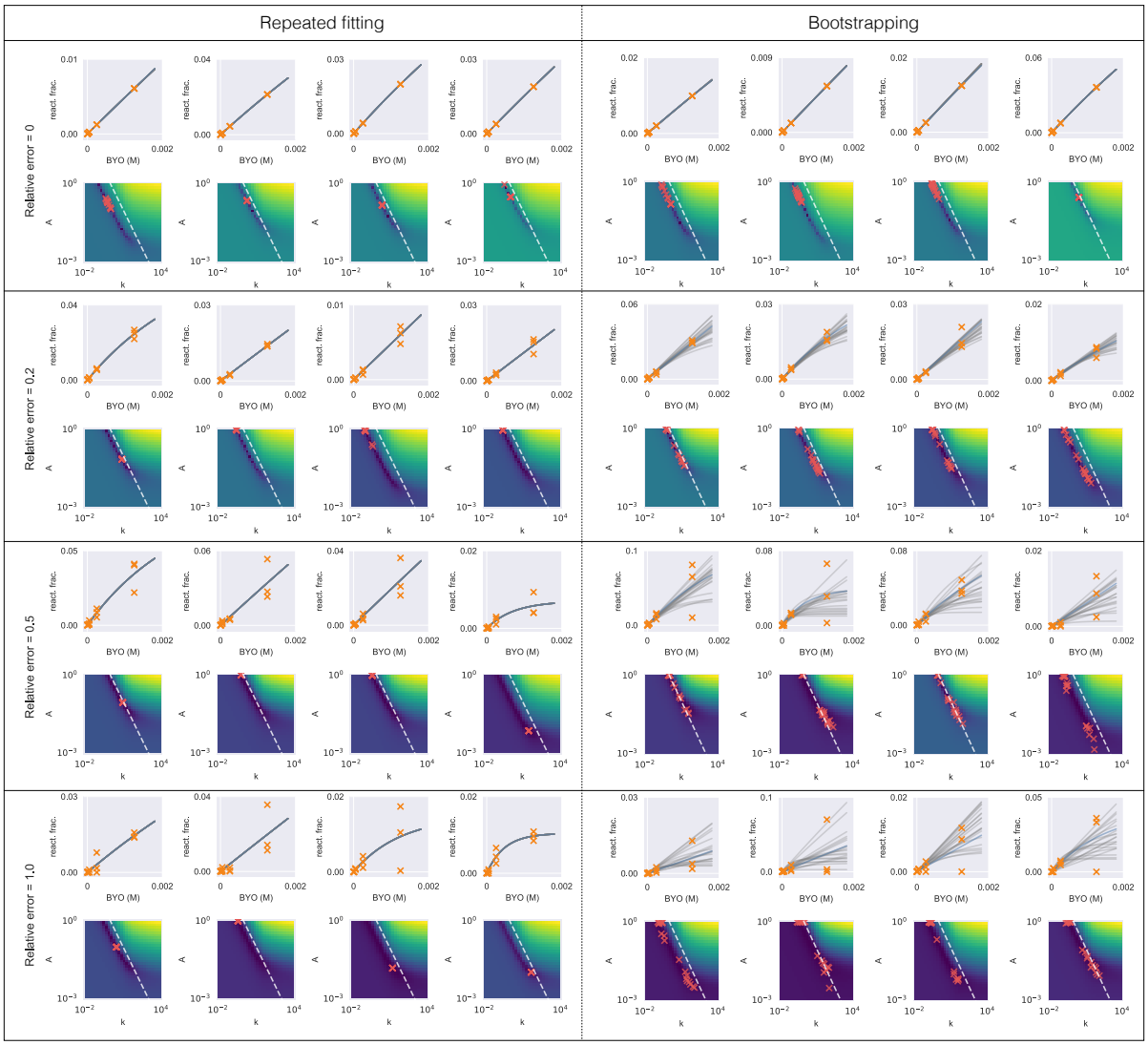


Figure A.2: Selected fitting results from Region 1 in simulated reacted fraction dataset with different relative error. Each curve plot shows the simulated reacted fraction (in triplicates) at various initial BYO concentrations (orange crosses), fitting curves from point estimation (blue line), and fitting curves from 20 repeated fitting or 20 bootstrapped samples (grey lines); fitted k, A values for the curves are shown in the corresponding heatmap (red crosses) under each curve plot. For visual guidance, background color of the heatmap indicates the relative values of mean squared error (normalized in each plot; blue to yellow is lower to higher error) over the parameter space given the data. The white dashed line marks $kA = 1\text{min}^{-1}\text{M}^{-1}$. An ideal fitting result would have converged fitting optima and is both numerically stable (from repeated fitting) and robust to noise (from bootstrapping). A large variance along the line of $kA =$ constant indicates the model is not identifiable, i.e., k and A cannot be separately estimated. (react. frac. = reacted fraction.)

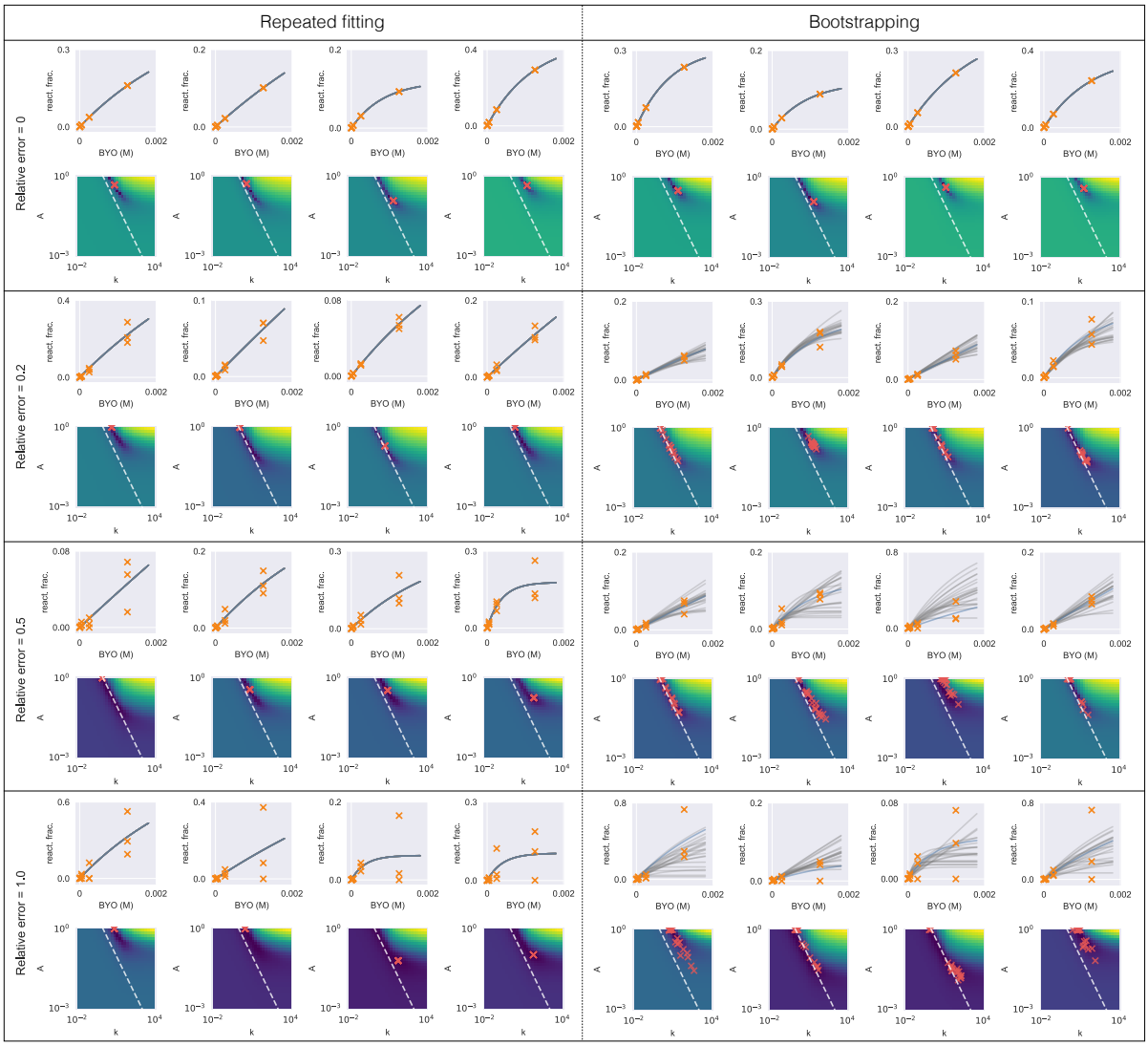


Figure A.3: Selected fitting results from Region 2 in simulated reacted fraction dataset with different relative error. Each curve plot shows the simulated reacted fraction (in triplicates) at various initial BYO concentrations (orange crosses), fitting curves from point estimation (blue line), and fitting curves from 20 repeated fitting or 20 bootstrapped samples (grey lines); fitted k, A values for the curves are shown in the corresponding heatmap (red crosses) under each curve plot. For visual guidance, background color of the heatmap indicates the relative values of mean squared error (normalized in each plot; blue to yellow is lower to higher error) over the parameter space given the data. The white dashed line marks $kA = 1\text{min}^{-1}\text{M}^{-1}$. An ideal fitting result would have converged fitting optima and is both numerically stable (from repeated fitting) and robust to noise (from bootstrapping). A large variance along the line of $kA =$ constant indicates the model is not identifiable, i.e., k and A cannot be separately estimated. (react. frac. = reacted fraction.)

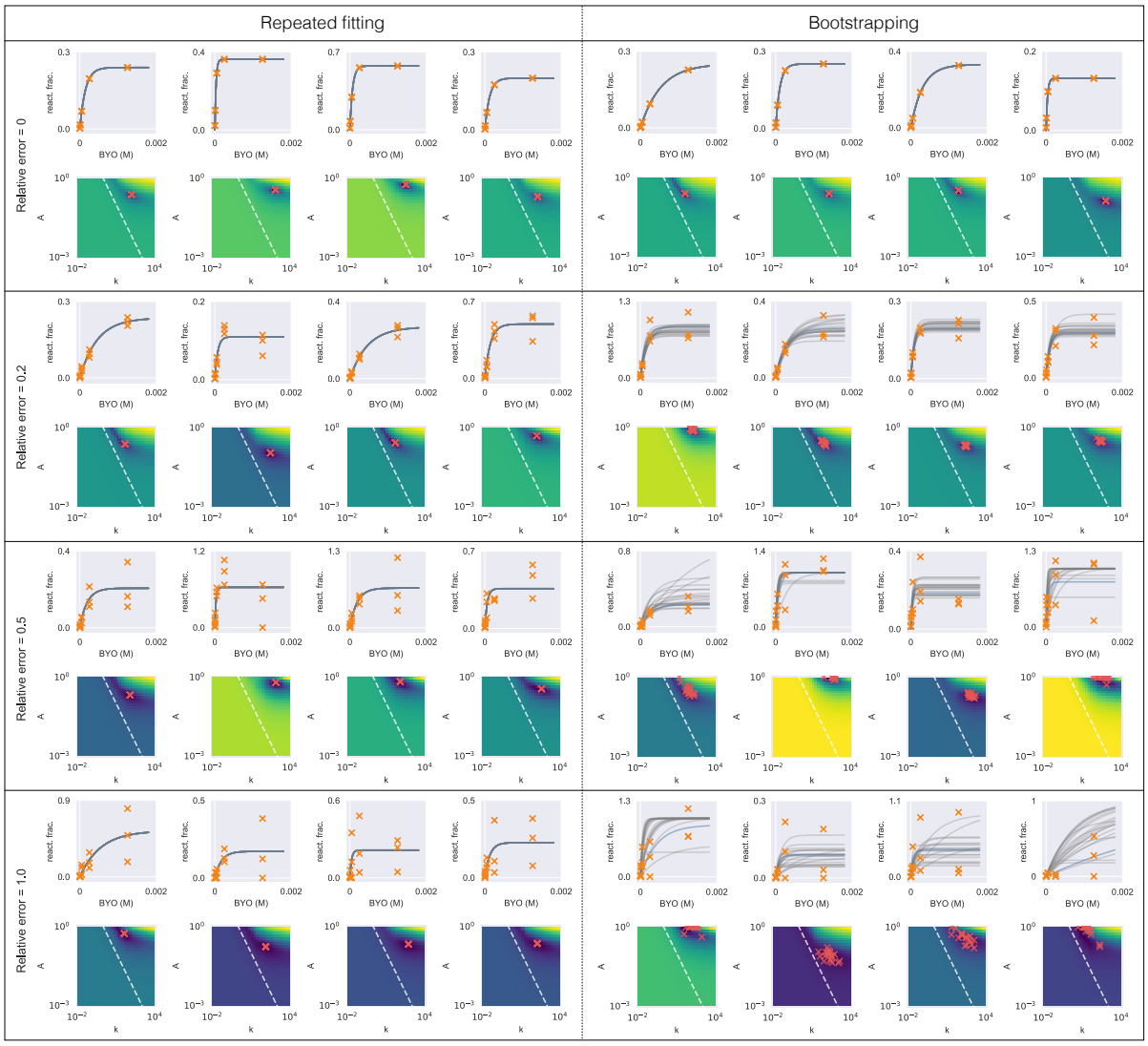


Figure A.4: Selected fitting results from Region 3 in simulated reacted fraction dataset with different relative error. Each curve plot shows the simulated reacted fraction (in triplicates) at various initial BYO concentrations (orange crosses), fitting curves from point estimation (blue line), and fitting curves from 20 repeated fitting or 20 bootstrapped samples (grey lines); fitted k, A values for the curves are shown in the corresponding heatmap (red crosses) under each curve plot. For visual guidance, background color of the heatmap indicates the relative values of mean squared error (normalized in each plot; blue to yellow is lower to higher error) over the parameter space given the data. The white dashed line marks $kA = 1\text{min}^{-1}\text{M}^{-1}$. An ideal fitting result would have converged fitting optima and is both numerically stable (from repeated fitting) and robust to noise (from bootstrapping). A large variance along the line of $kA =$ constant indicates the model is not identifiable, i.e., k and A cannot be separately estimated. (react. frac. = reacted fraction.)

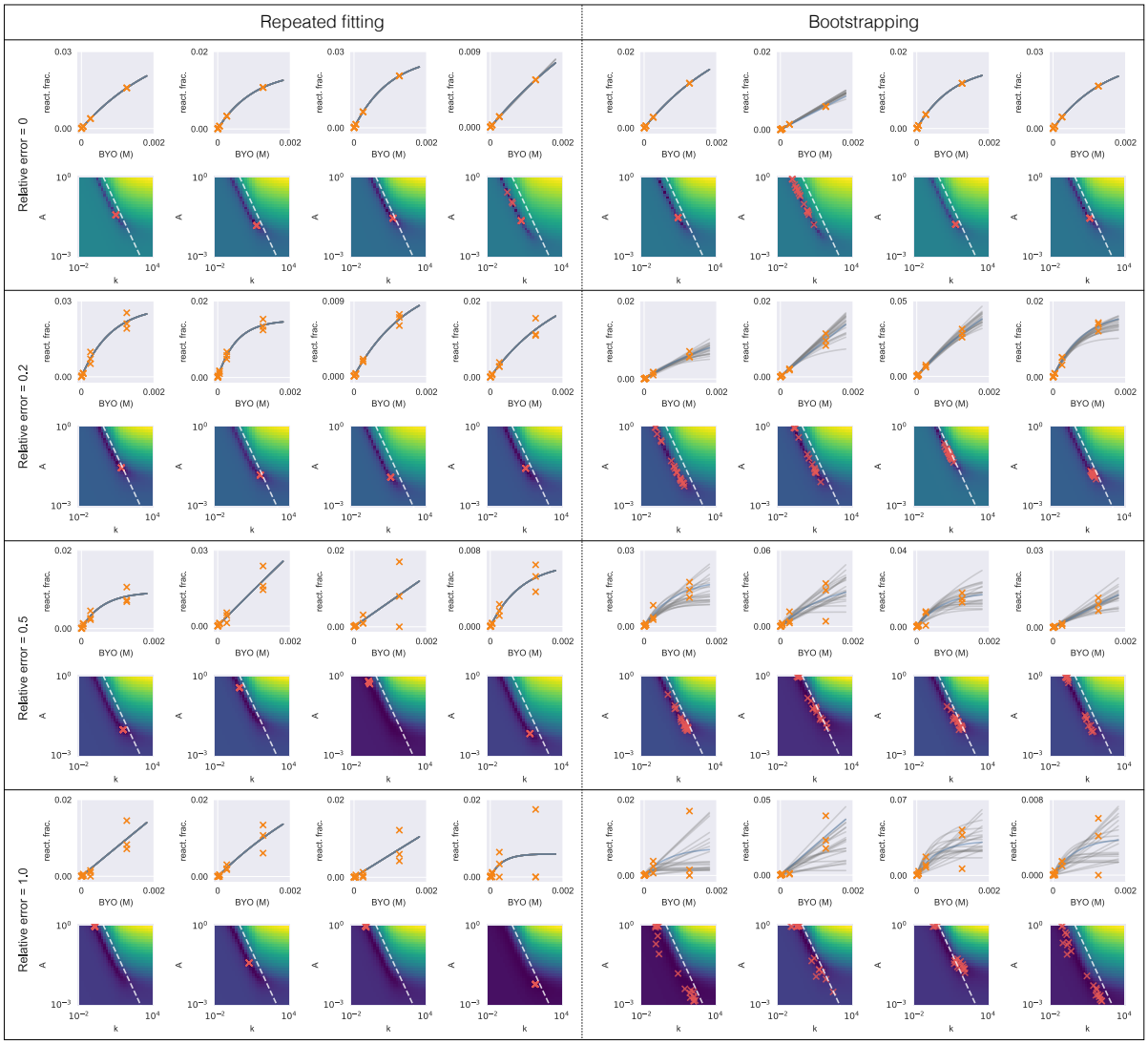


Figure A.5: Selected fitting results from Region 4 in simulated reacted fraction dataset with different relative error. Each curve plot shows the simulated reacted fraction (in triplicates) at various initial BYO concentrations (orange crosses), fitting curves from point estimation (blue line), and fitting curves from 20 repeated fitting or 20 bootstrapped samples (grey lines); fitted k, A values for the curves are shown in the corresponding heatmap (red crosses) under each curve plot. For visual guidance, background color of the heatmap indicates the relative values of mean squared error (normalized in each plot; blue to yellow is lower to higher error) over the parameter space given the data. The white dashed line marks $kA = 1\text{min}^{-1}\text{M}^{-1}$. An ideal fitting result would have converged fitting optima and is both numerically stable (from repeated fitting) and robust to noise (from bootstrapping). A large variance along the line of $kA =$ constant indicates the model is not identifiable, i.e., k and A cannot be separately estimated. (react. frac. = reacted fraction.)

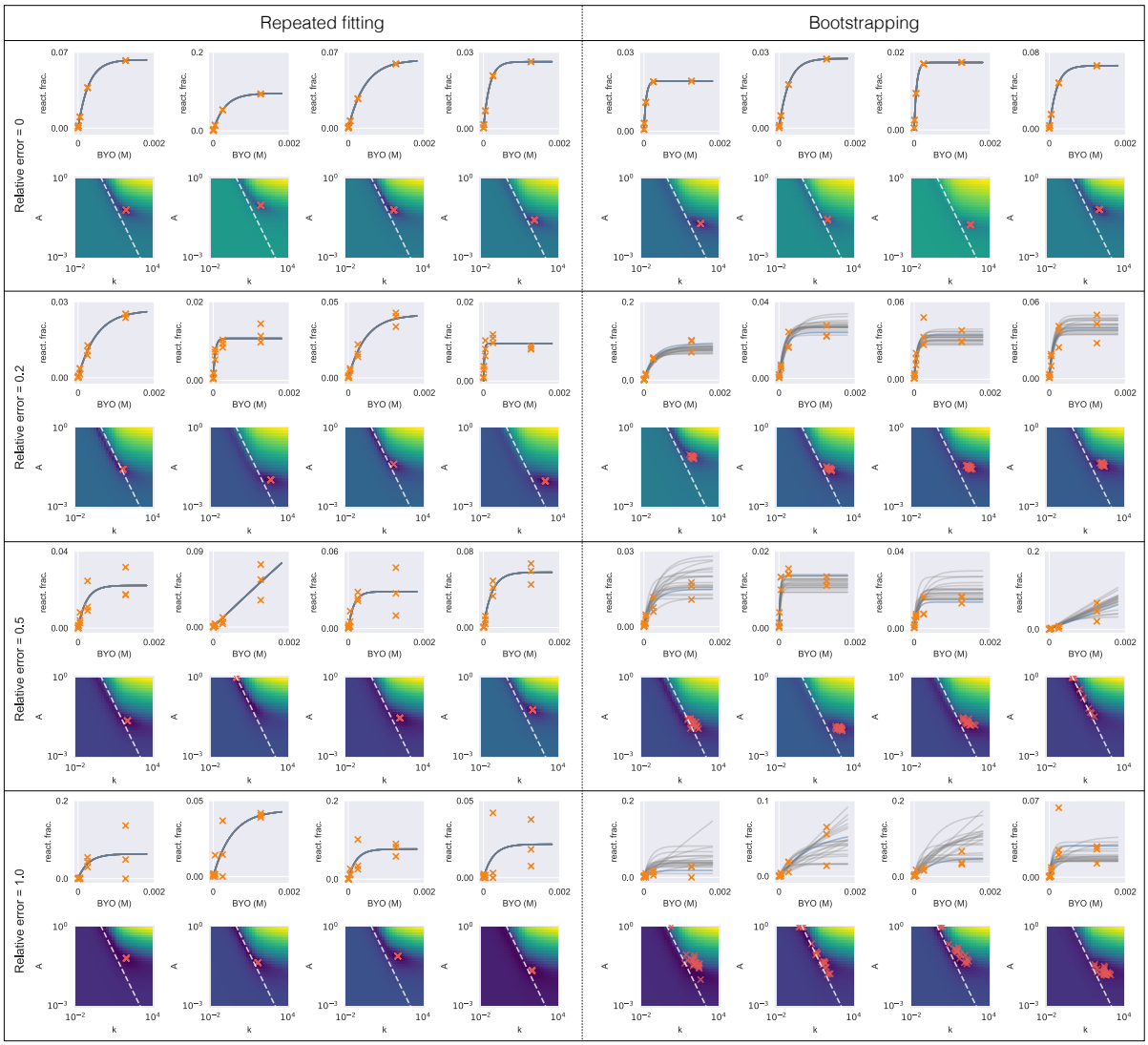


Figure A.6: Selected fitting results from Region 5 in simulated reacted fraction dataset with different relative error. Each curve plot shows the simulated reacted fraction (in triplicates) at various initial BYO concentrations (orange crosses), fitting curves from point estimation (blue line), and fitting curves from 20 repeated fitting or 20 bootstrapped samples (grey lines); fitted k, A values for the curves are shown in the corresponding heatmap (red crosses) under each curve plot. For visual guidance, background color of the heatmap indicates the relative values of mean squared error (normalized in each plot; blue to yellow is lower to higher error) over the parameter space given the data. The white dashed line marks $kA = 1\text{min}^{-1}\text{M}^{-1}$. An ideal fitting result would have converged fitting optima and is both numerically stable (from repeated fitting) and robust to noise (from bootstrapping). A large variance along the line of $kA =$ constant indicates the model is not identifiable, i.e., k and A cannot be separately estimated. (react. frac. = reacted fraction.)

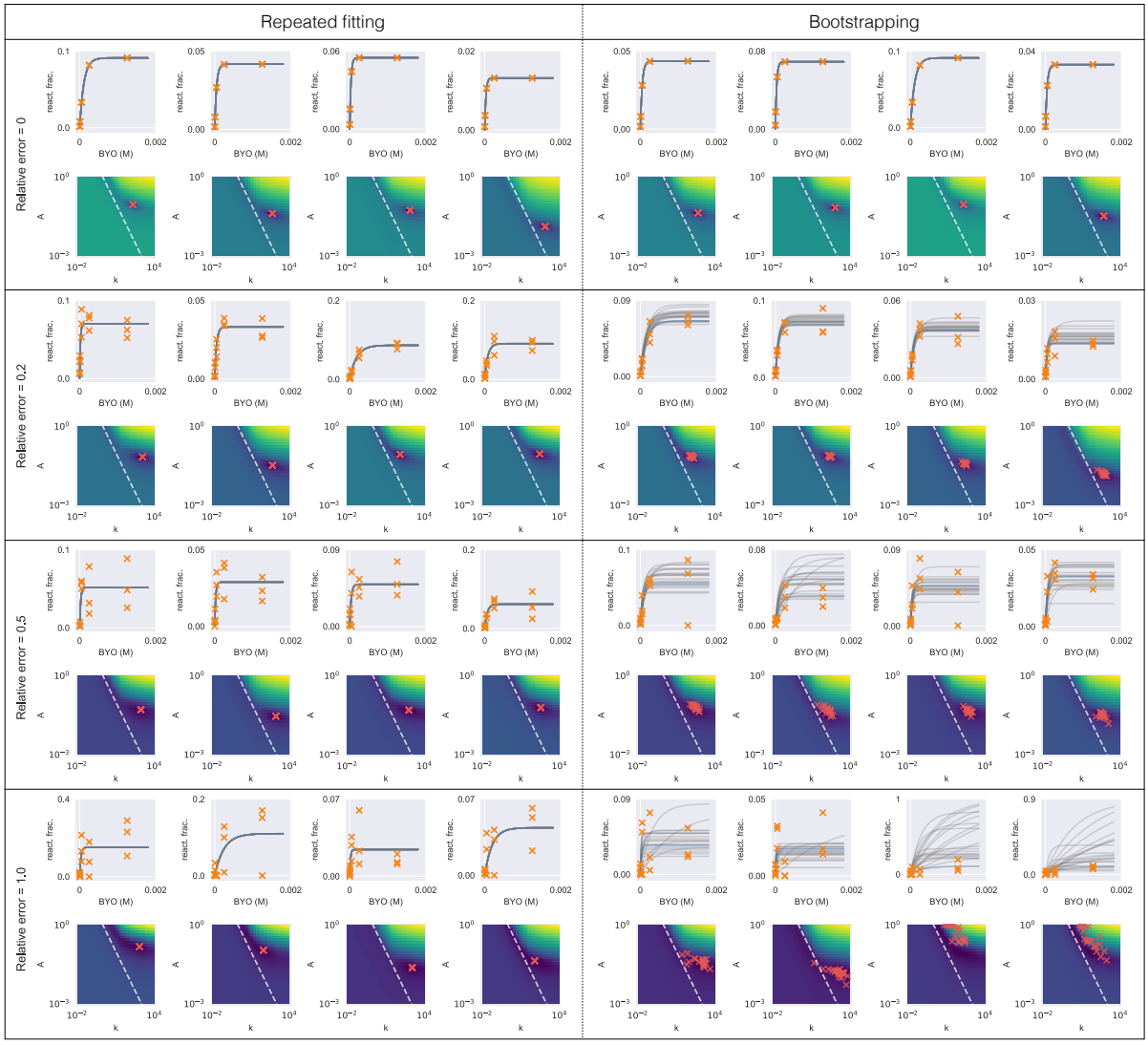


Figure A.7: Selected fitting results from Region 6 in simulated reacted fraction dataset with different relative error. Each curve plot shows the simulated reacted fraction (in triplicates) at various initial BYO concentrations (orange crosses), fitting curves from point estimation (blue line), and fitting curves from 20 repeated fitting or 20 bootstrapped samples (grey lines); fitted k, A values for the curves are shown in the corresponding heatmap (red crosses) under each curve plot. For visual guidance, background color of the heatmap indicates the relative values of mean squared error (normalized in each plot; blue to yellow is lower to higher error) over the parameter space given the data. The white dashed line marks $kA = 1\text{min}^{-1}\text{M}^{-1}$. An ideal fitting result would have converged fitting optima and is both numerically stable (from repeated fitting) and robust to noise (from bootstrapping). A large variance along the line of $kA =$ constant indicates the model is not identifiable, i.e., k and A cannot be separately estimated. (react. frac. = reacted fraction.)

Method	Relative error (ϵ)	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
Repeated fitting	0.0	N	Y	Y	N	Y	Y
	0.2	Y	Y	Y	Y	Y	Y
	0.5	Y	Y	Y	Y	Y	Y
	1.0	Y	Y	Y	Y	Y	Y
Bootstrapping	0.0	N	Y	Y	N	Y	Y
	0.2	N	N	Y	N	Y	Y
	0.5	N	N	N	N	N	N
	1.0	N	N	N	N	N	N

Table A.1: Summary of visual examination of model identifiability from repeated fitting (no resampling) and bootstrapping. 'Y' indicates regions where k and A appear that they can be separately estimated, 'N' indicates regions where they do not appear to be separately estimable. Results from repeated fitting account for the numeric effect from different initial values and results from bootstrapping also account for the effect of sample noise.

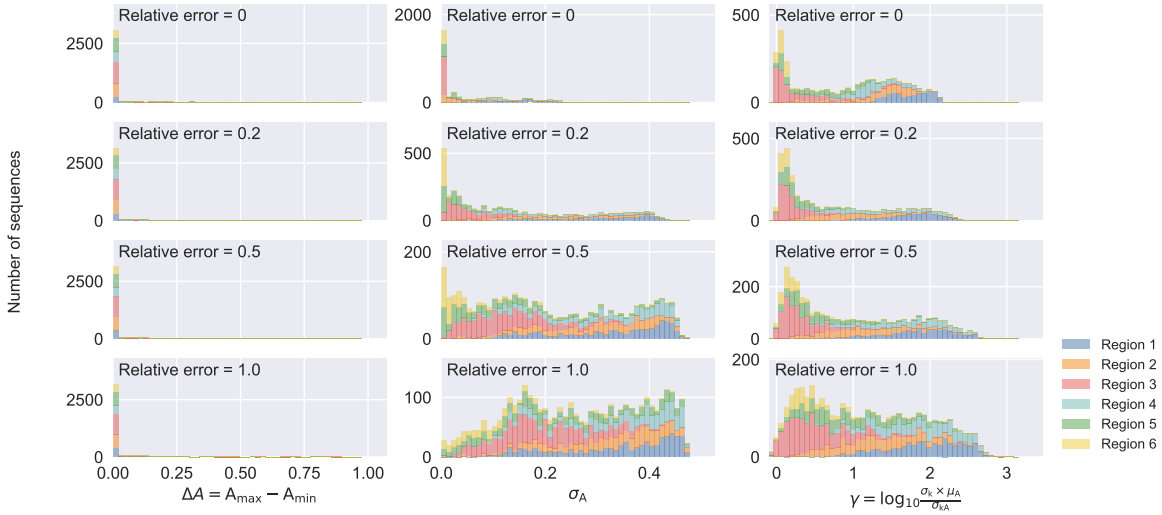


Figure A.8: Distribution of metric scores for sequences from 6 selected regions in the simulated reacted fraction dataset, with various noise level. Histogram bars are stacked for visibility. Both metrics σ_A and γ captured the trend of model identifiability as summarized in A.1. In contrast, ΔA failed to capture the difference in model identifiability between sequences with different regions and sample noise.

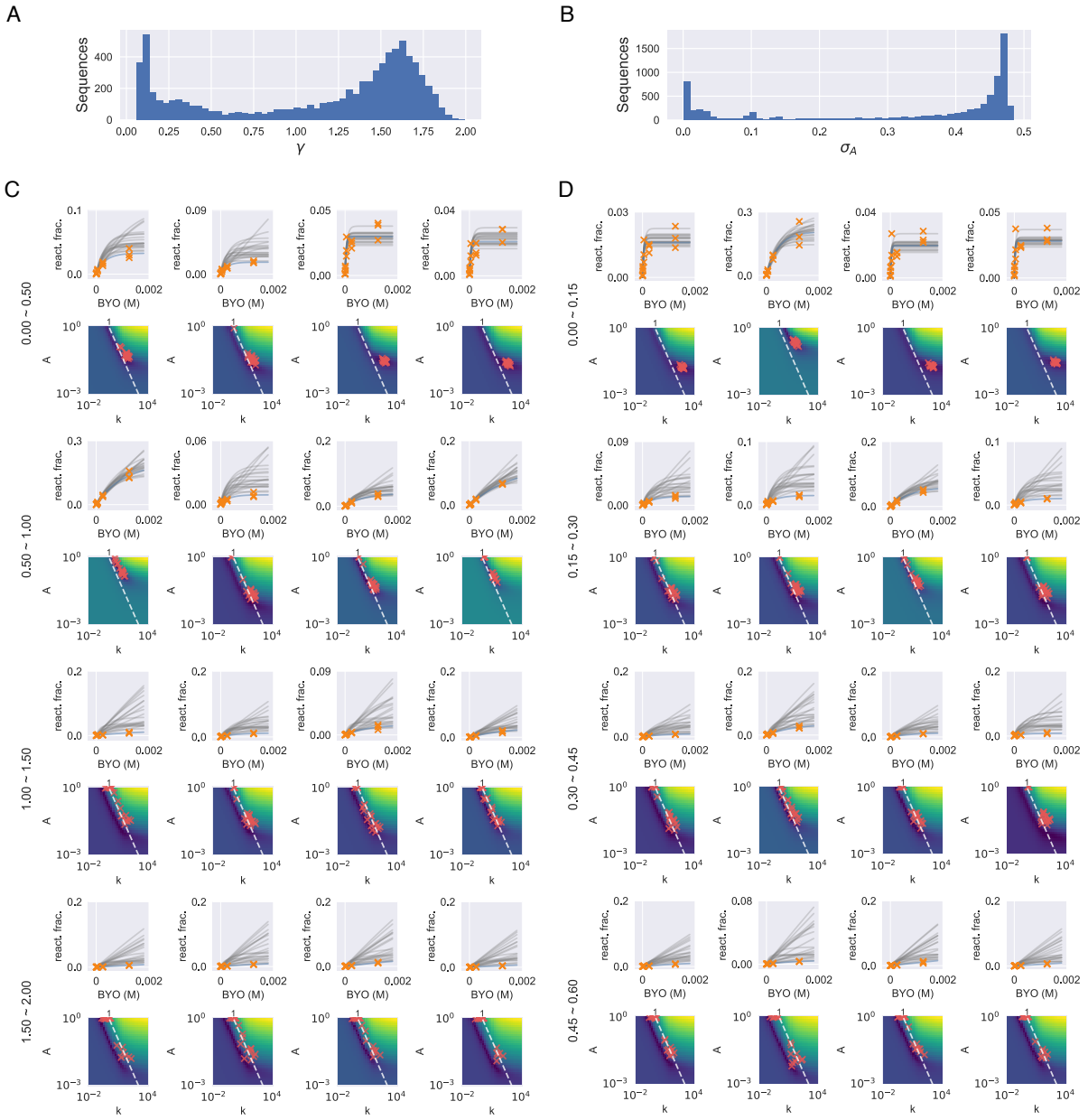


Figure A.9: Distribution of γ (A) and σ_A (B) for sequences with in Hamming distance of 2 to the family centers from the variant pool k-Seq experiment. Example fitting results are shown for sequences within each score range (labels on the left) of γ (C) and σ_A (D). For explanation of (C) and (D), also see caption of Figure A.2. Sequences with low metric scores for both metrics showed good model identifiability while k and A cannot be separately estimated for those with high metric scores. k and a for most sequences up to double mutants could not be estimated separately.

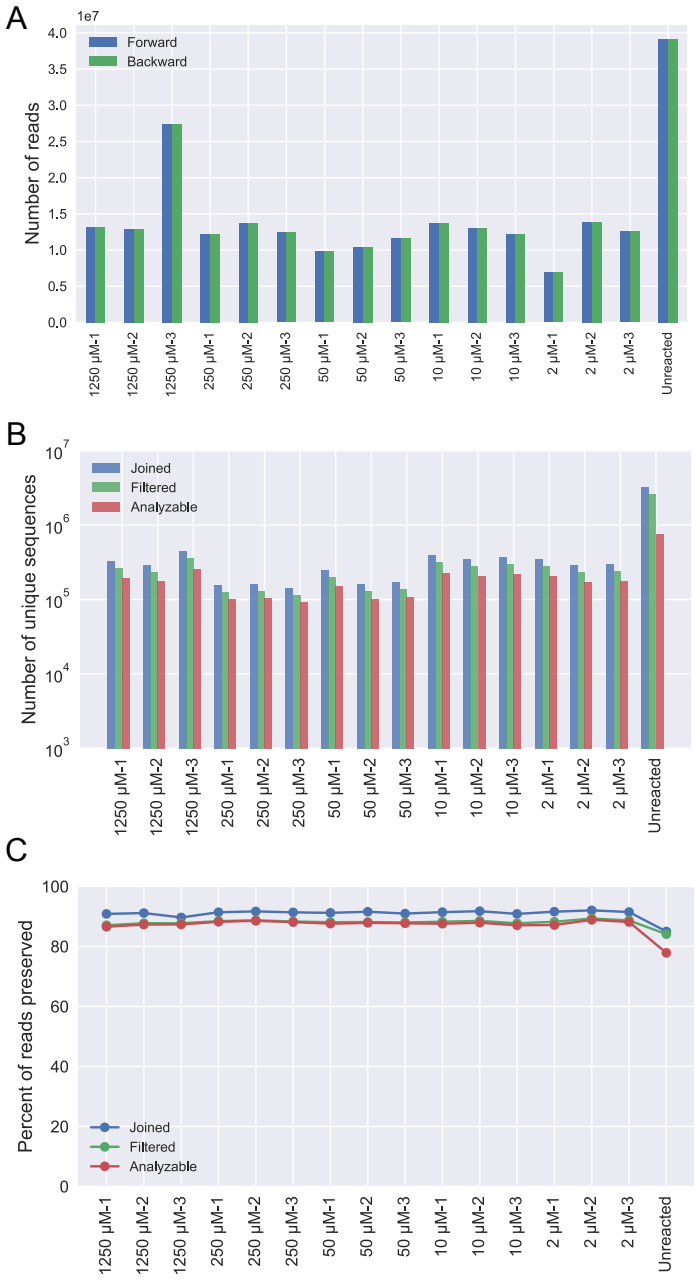


Figure A.10: Processing of sequence reads. (A) Number of raw paired-end reads in each sample. The input sample (unreacted) had 3x of other samples for total DNA input for sequencing. (B) Number of unique sequences and (C) percent of total reads retained after paired-end reads joining, filtering (removal of spike-in sequence and sequences that are not 21 nt long), and checking for analyzability (has non-zero counts in the input pool and in at least one of the reacted samples)

Text A.1: Theoretical design for the variant pool. We defined a sequence with d substitutions (Hamming distance) compared to the wild-type sequence as a d-th order mutant. For each variant pool (a single family) with randomized length $L = 21$ residues and mutation rate η, the fraction of a d-th order mutant in the pool is

$$p_d = (1-\eta)^{L-d}\left(\frac{\eta}{3}\right)^d \tag{A.1}$$

The value of η maximizing the fraction of a single d-th order mutant satisfies $\frac{dp_d}{d\eta} = 0$, thus

$$\eta_d^0 = \arg\max_{\eta} p_d = d/L \tag{A.2}$$

and the maximum fraction for a d-th order mutant in the variant pool is

$$p_{d,\max} = \left(1-\frac{d}{L}\right)^{L-d}\left(\frac{d}{3L}\right)^d \tag{A.3}$$

We used above equation to determine the optimal η maximizing the fraction of a given order of mutant in the pool. For a sample with N total reads, the expected counts for a d-th order mutant is $p_d N$.

Text A.2: Effects of sequencing error in the variant pool. From Equation A.1, the relative abundance ratio between a d-th mutant and a $(d+1)$-th mutant is

$$\phi = \frac{p_d}{p_{d+1}} = \frac{3(1-\eta)}{\eta} \tag{A.4}$$

Assuming a constant sequencing error rate ξ per nucleotide and considering the effect of sequencing error only by substitution, the probability that a sequence will be misidentified as one of its one-mutation neighbors is

$$(1-\xi)^{L-1}\frac{\xi}{3} \tag{A.5}$$

A d-th order mutant has $3L$ neighbors consisting of d $(d-1)$-th mutants (i.e., one of the d mutated nucleotides reverted to the wild type), $2d$ d-th mutants (i.e., one of the d mutated nucleotides changed to one of other two possible mutations), and $3(L-d)$ $(d+1)$-th mutants (i.e., one of the $L-d$ wild type nucleotides mutated). Assuming the real abundance for this d-th order mutant is 1, the $(d-1)$-th mutant is ϕ, and $(d+1)$-th mutant is $1/\phi$. The expected observed abundance for a d-th mutant, in a variant pool with mutation rate η and sequencing error rate ξ is

$$\rho(d,\xi,\eta) = (d\phi + 2d + 3(L-d)/\phi)(1-\xi)^{L-1}\frac{\xi}{3} + (1-\xi)^L \tag{A.6}$$

The fraction of abundance that originates from its neighbors due to sequencing error is

$$1 - \frac{(1-\xi)^L}{\rho(d,\xi,\eta)} \tag{A.7}$$

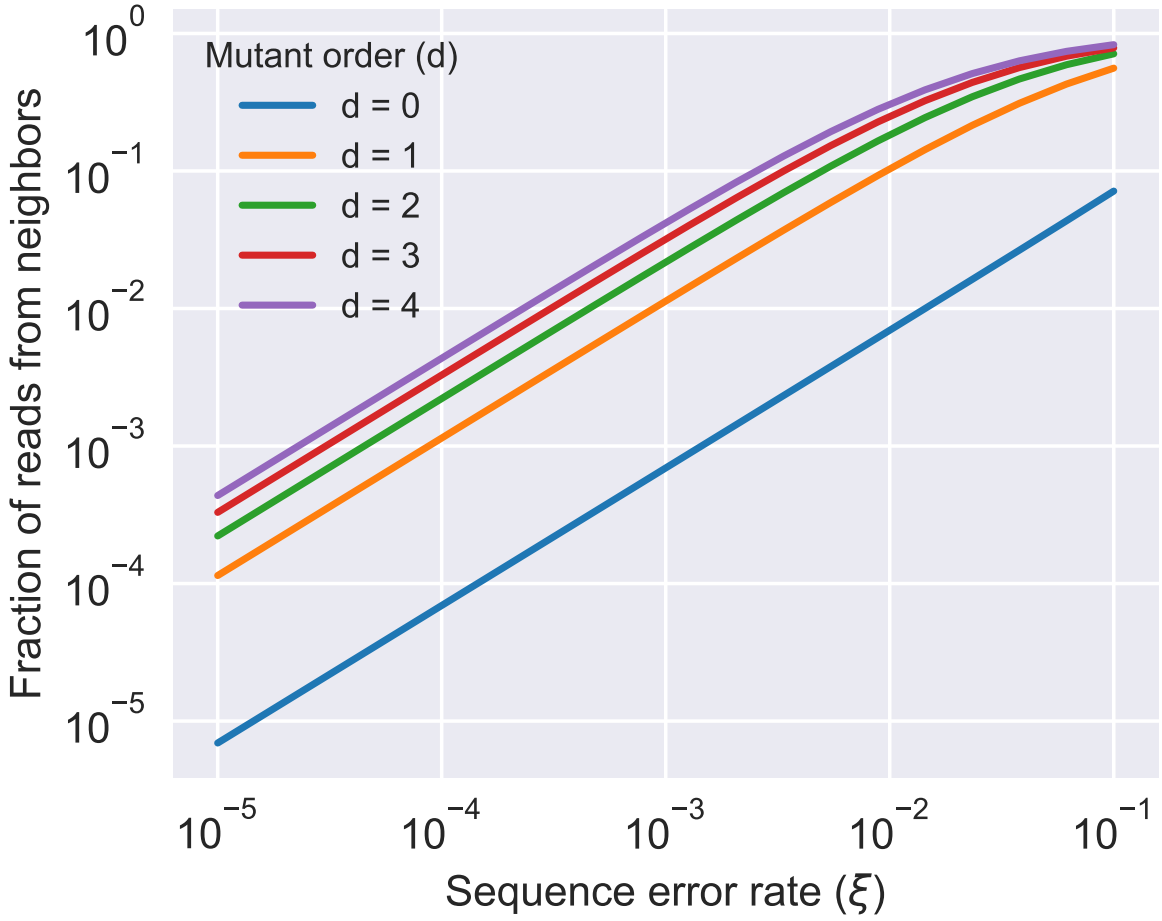


Figure A.11: Expected error from single-mutation neighbor sequences due to sequencing errors, for different orders of mutants (d) and different rates of sequencing error (ξ). Family centers ($d = 0$) are the most abundant sequences and would be least affected by sequencing error. With decreased error rate, the fraction of reads resulting from erroneous reads of neighboring sequences is decreased for each order of mutants. The mutation rate in synthesizing the variant pool is 9%.

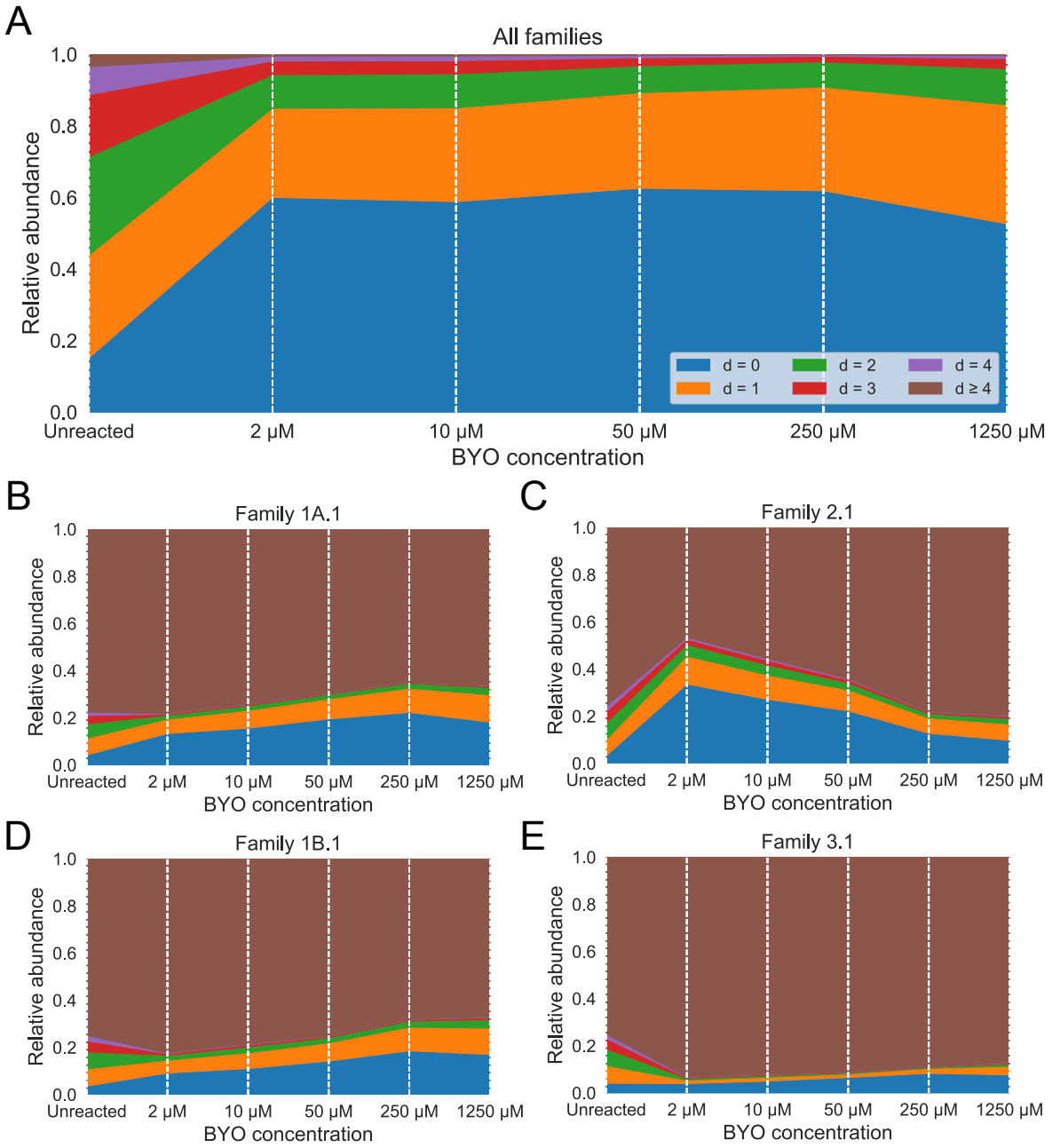


Figure A.12: Change in relative abundance for mutants in the mixed variant pool (A) and in each individual family (B - E). Most reads were within triple mutants in the unreacted pool ($\sim$ 90%) and reacted pool ($\sim$ 95%). Activities of ribozymes affected the abundance of sequences in the reacted pool that the abundance for more active sequences (e.g., family centers) might increase after the reaction, affecting the quantitation for lower abundance, lower active sequences. In Figure B-E, sequences from other families are also classified as $d \geq 4$.

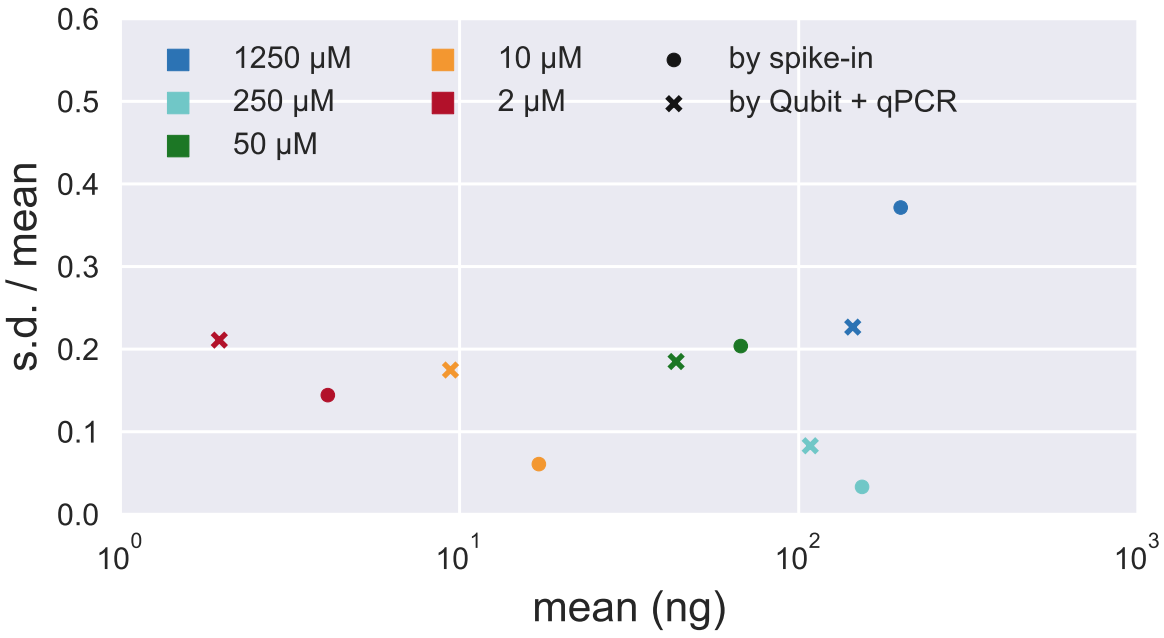


Figure A.13: Relative standard deviation vs mean of total RNA measured in each reacted sample triplicates using spike-in sequence or qPCR + Qubit as the absolute quantification methods. The mean relative standard deviation for spike-in method is 0.162 and for qPCR + Qubit method is 0.176.

Figure A.14: Fraction estimated CI-95 using bootstrap or triplicates includes true kA values for sequences with different true kA values. Sequences were ordered by true kA values (from large to small) and each data point represented the fraction of CI-95 include the truth in each 25,000 consecutive sequences. While results from bootstrapping consistently includes $\sim 95\%$ of truth in the CI-95, results from triplicates underestimate the uncertainty (over-confidence), especially for sequences with high kA values.

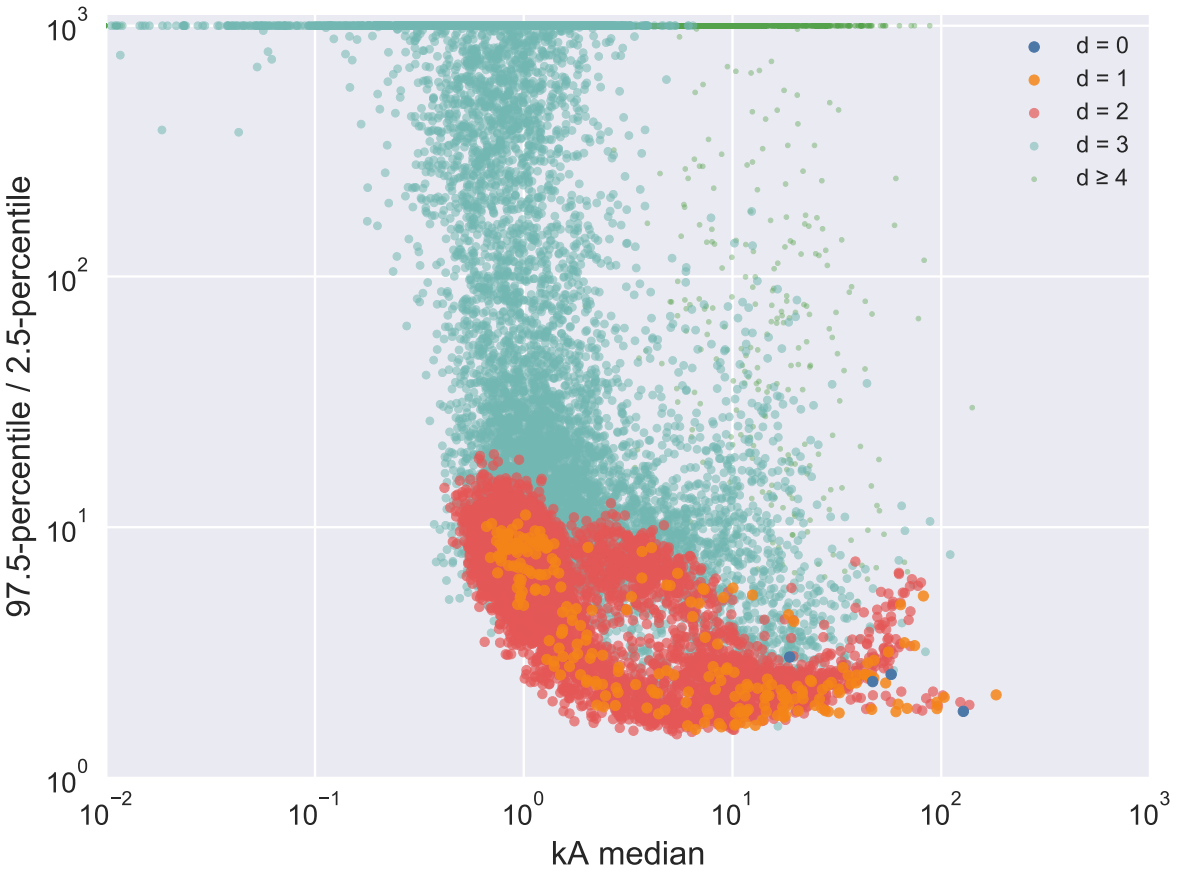


Figure A.15: The relation of fold range (97.5-percentile / 2.5-percentile) did not have a strong dependence on median kA. A decrease then plateaued trend was observed as median kA increases. Colored by sequences' hamming distance to the nearest peak center

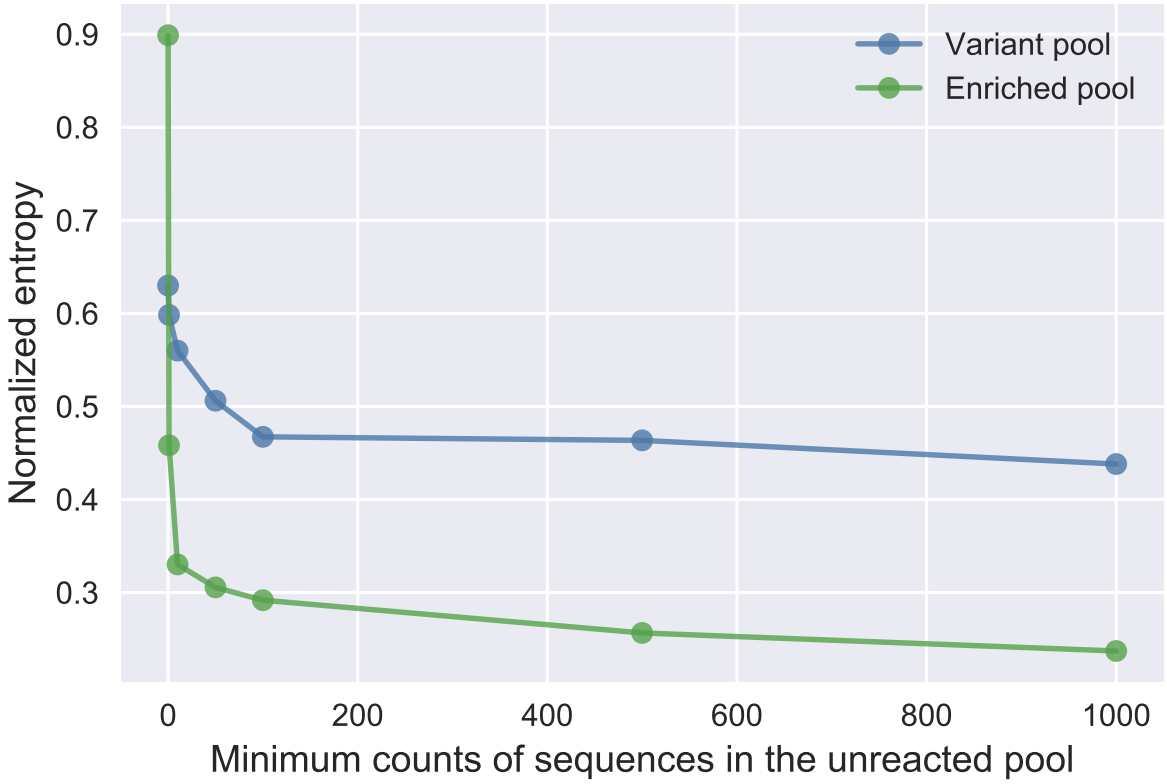


Figure A.16: Pool evenness for enriched pool and variant pool. Pool evenness is evaluated by entropy efficiency ($-\sum_{i=1}^{N} p(i) \log p(i) / \log(N)$) for sequences with various minimum count thresholds in the unreacted samples. As designed, variant pool is more even (higher entropy efficiency) than the enriched pool.

A.2 Supplementary Information for Chapter 3

	Bacterial OTUs				Viral OTUs		
Sample type	Microbial Standard	Negative Control	Skin	Wound	Negative Control	Skin	Wound
# of samples	3	6	20	40	4	20	40
Mean	158182.7	21481.5	161515.2	52443.2	111400.2	325907.2	407034.9
SD	141681.3	18280.3	36875.7	60966.4	106651.3	708023.1	515703.6
Min	70528	1714	108253	195	7446	17803	23394
25%	76454.5	7813	143464.2	11979.8	57111.8	55435.2	202312.2
50%	82381	19705.5	157753	35438.5	89459.5	142193	242536
75%	202010	32018.8	177004.8	68553.5	143748	267070.5	312258
Max	321639	47792	269516	322811	259236	3266230	2597506

Table A.2: Summary of statistics for the raw reads recovered in bacterial and viral bioinformatic pipelines after OTU picking, categorized by different sample types. SD = standard deviation.

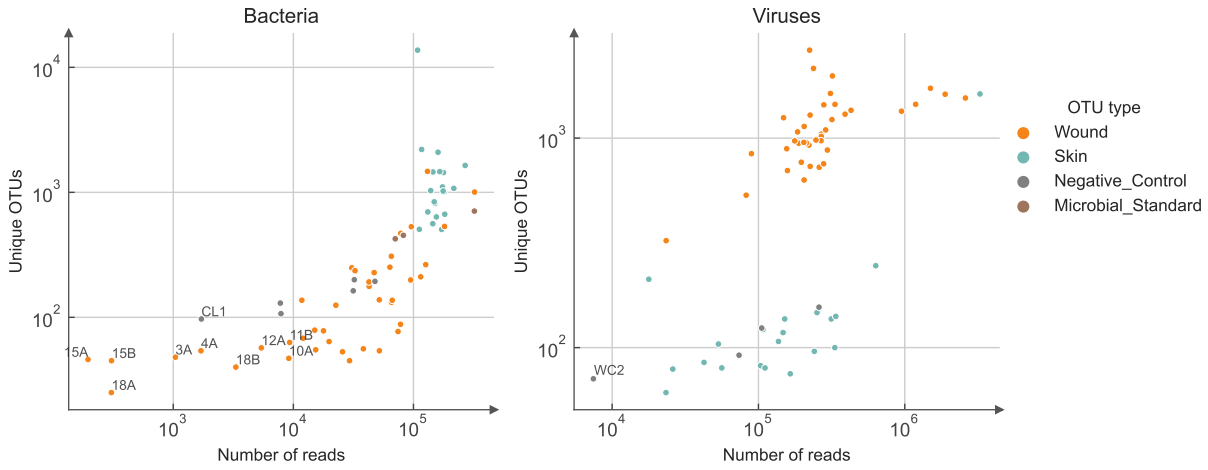


Figure A.17: Number of unique OTUs vs. total reads in each sample from the output of OTU picking pipelines.

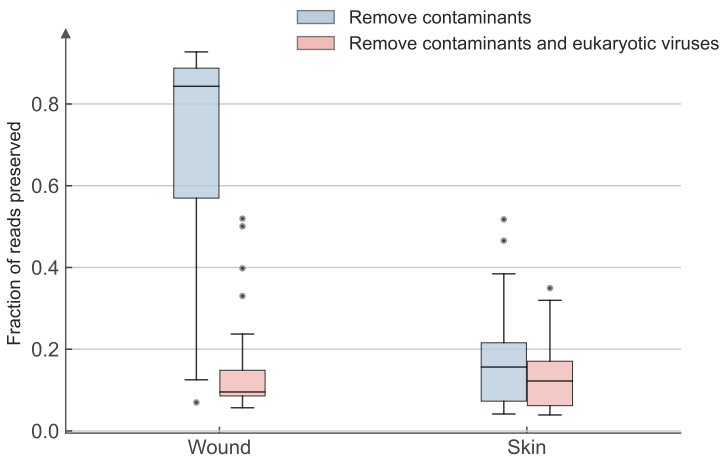


Figure A.18: Fraction of reads preserved after removing putative viral contaminants and eukaryotic viruses.

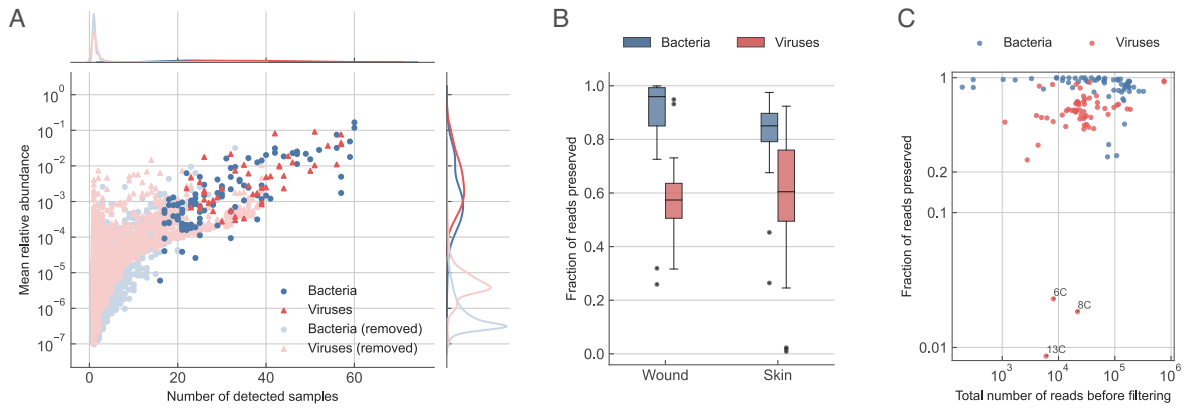


Figure A.19: Prevalence filter for aggregated OTUs. (A) Most of the bacterial and viral OTUs were only detected in a small number of samples and and low relative abundance in the samples. By applying the prevalence filter (removing the OTUs only detected in skin or wound samples, or OTUs only detected in $< 80\%$ (16) of patients, most of the low abundant and non-prevlend OTUs were removed. (B) Removing these OTUs does not change the total number of reads for most of samples ($> 50\%$ of reads preserved), (C) excepting for the viral reads in sample 6C, 8C, and 13C .

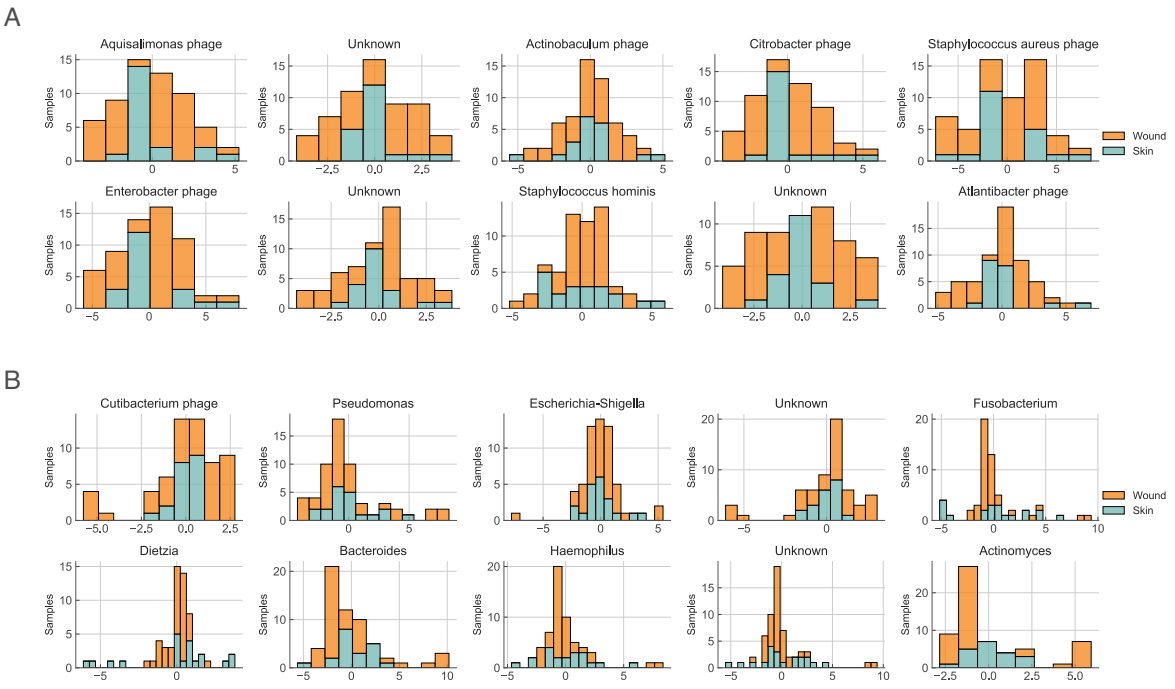


Figure A.20: Selected marginal distribution for CLR transformed abundance and centered within wound and skin samples. (A) top OTU dimensions with highest p-values from Shapiro-Wilk test (most normal) and (B) top OTU dimensions with lowest p-values from Shapiro-Wilk test (least normal)

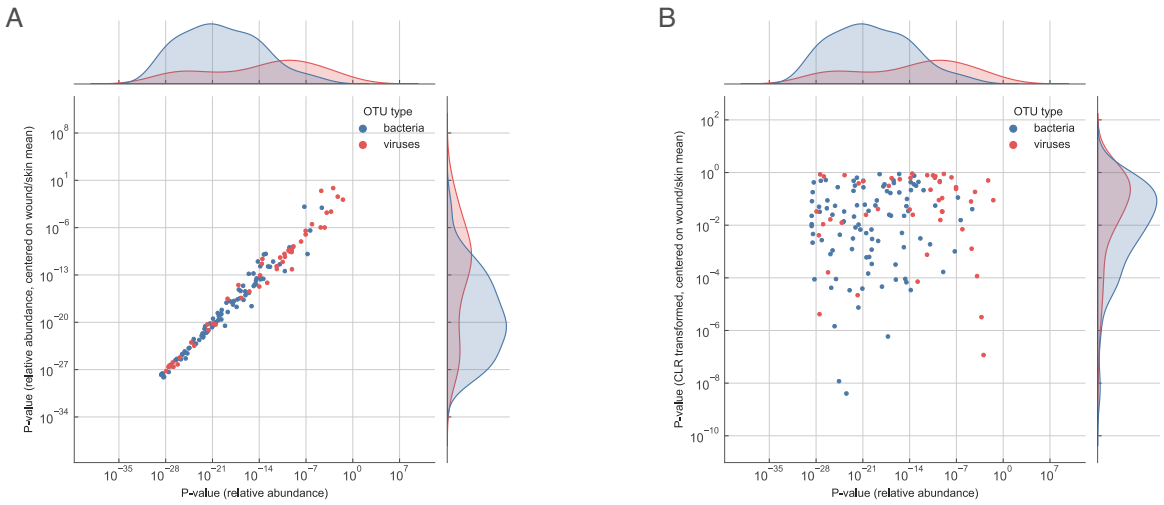


Figure A.21: Compare the normality of the marginal distributions of the dataset. The p-values from the d'Agostino-Pearson normality test for transformed data with different methods were plotted against each other as (A) relative abundance vs. relative abundance with wound and skin samples each centered on the group mean; (B) relative abundance vs. centered log transformed data with wound and skin samples each centered on the group mean. A low p-value means the test reject the null hypothesis that the marginal distribution is normal.

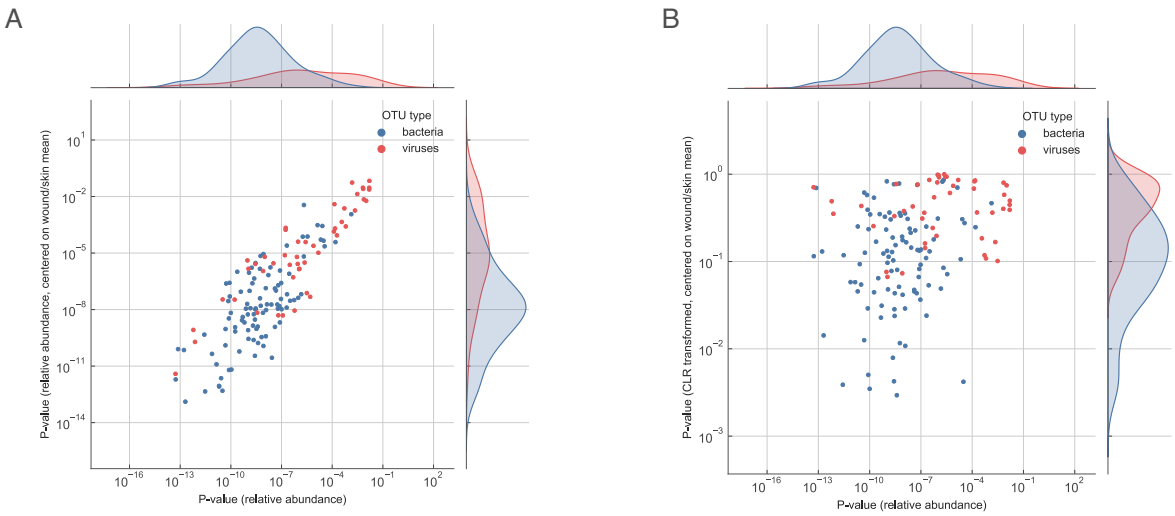


Figure A.22: Compare the normality of the marginal distributions of the dataset. The p-values from the Kolmogorov-Smirnov normality test for transformed data with different methods were plotted against each other as (A) relative abundance vs. relative abundance with wound and skin samples each centered on the group mean; (B) relative abundance vs. centered log transformed data with wound and skin samples each centered on the group mean. A low p-value means the test reject the null hypothesis that the marginal distribution is normal.

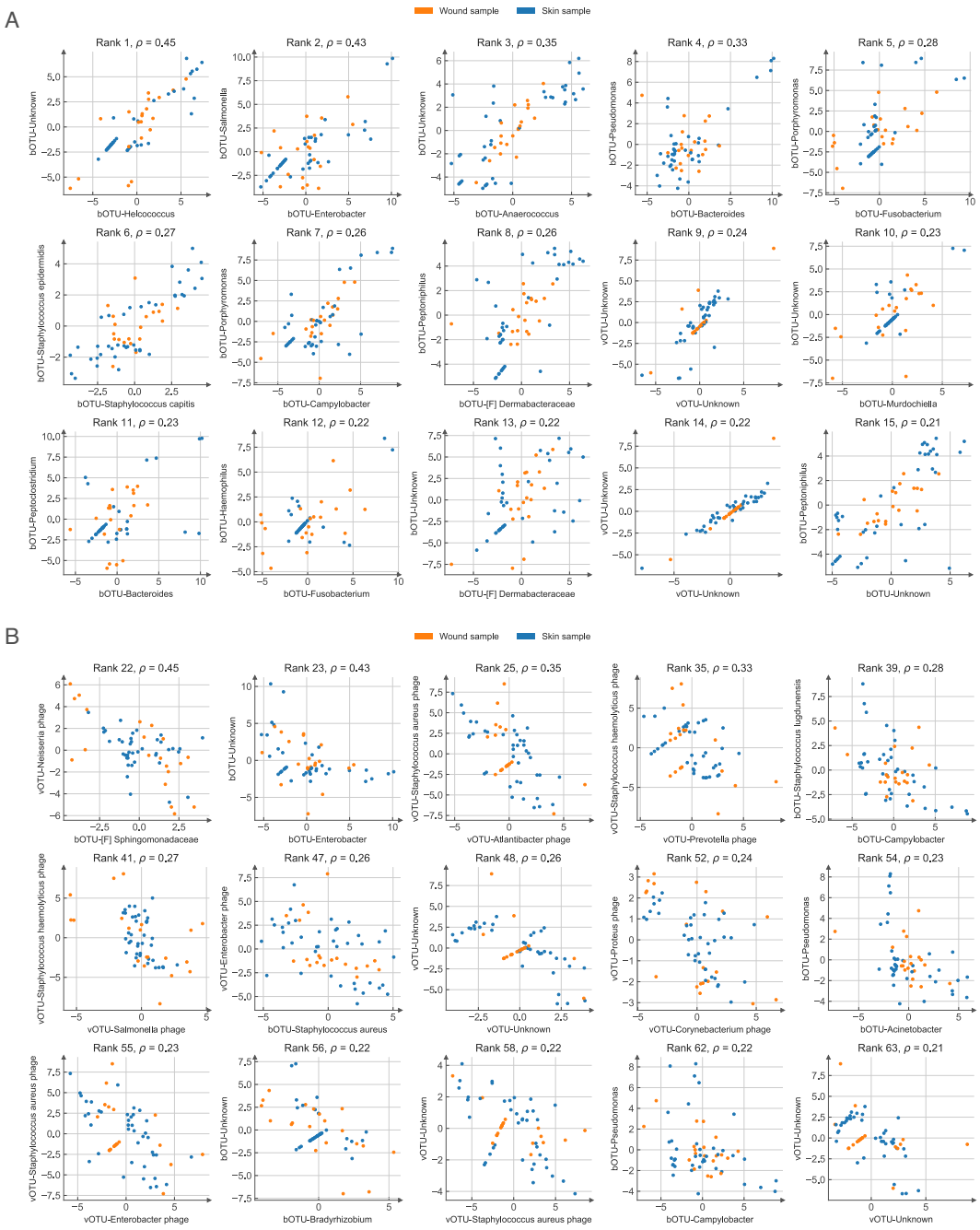


Figure A.23: Marginal distributions of CLR transformed and centered (in each sample type) abundance for each pair of the nodes from the top (A) positive edges and (B) negative edges.

Cluster	Cluster size	Edge density	Fraction of positive correlations
Cluster 0	18	0.169935	1.000
Cluster 1	14	0.175824	0.875
Cluster 2	14	0.153846	1.000
Cluster 3	13	0.243590	1.000
Cluster 4	9	0.305556	1.000
Cluster 5	8	0.285714	1.000
Cluster 6	5	0.400000	1.000
Cluster 7	3	0.666667	1.000
Cluster 8	3	0.666667	1.000

Table A.3: Basic statistics of Clusters identified from the skin microbiome dataset. Cluster size is the number of nodes in the cluster.

Table A.4: Table of OTUs in the filtered dataset with two groupings. Group is the topological group identified from Louvain using both positive and negative edges; Cluster is the group identified from Louvain using only positive edges, OTUs in the sample Cluster are mostly positively correlated.

OTU ID	OTU type	Mean rel. abun.	Sample detected	Phylum	Class	Order	Family	Genus	Group	Cluster
FJ900888.1.1397	bOTU	0.053062	59	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Group 1	Cluster 6
ACFU01000050.379.1869	bOTU	0.029133	46	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	Campylobacter	Group 1	Cluster 1
New.ReferenceOTU855	bOTU	0.026262	47	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus lugdunensis	Group 1	Cluster 5
FJ470579.1.1493	bOTU	0.025857	40	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas	Group 1	Cluster 1
New.ReferenceOTU1087	bOTU	0.024755	49	Unknown	Unknown	Unknown	Unknown	Unknown	Group 1	Cluster 6
EU777788.1.1403	bOTU	0.019742	59	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia-Shigella	Group 1	Other
JN713234.1.1491	bOTU	0.015520	39	Firmicutes	Clostridia	Clostridiales	Family XI	Helcococcus	Group 1	Cluster 1
JN866570.1.1464	bOTU	0.014945	38	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Enhydrobacter	Group 1	Other
GU247457.1.1249	bOTU	0.004784	34	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	Group 1	Cluster 1
HG917228.1.1519	bOTU	0.002816	21	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	Group 1	Other
New.ReferenceOTU371	bOTU	0.002699	30	Unknown	Unknown	Unknown	Unknown	Unknown	Group 1	Cluster 1
New.ReferenceOTU314	bOTU	0.001821	24	Unknown	Unknown	Unknown	Unknown	Unknown	Group 1	Cluster 1
GU940714.1.1403	bOTU	0.001629	30	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces	Group 1	Cluster 1
DQ337009.1.1477	bOTU	0.001137	20	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Chryseobacterium	Group 1	Other
JX398975.1.1483	bOTU	0.000653	25	Actinobacteria	Actinobacteria	Micrococcales	Dermabacteraceae	Brachybacterium	Group 1	Other
New.ReferenceOTU617	bOTU	0.000471	20	Unknown	Unknown	Unknown	Unknown	Unknown	Group 1	Other
EU234319.1.1490	bOTU	0.000357	24	Actinobacteria	Actinobacteria	Corynebacteriales	Mycobacteriaceae	Mycobacterium	Group 1	Other
EU773843.1.1375	bOTU	0.000328	22	Actinobacteria	Actinobacteria	Corynebacteriales	Nocardiaceae	Rhodococcus	Group 1	Other
New.ReferenceOTU143	bOTU	0.000209	23	Unknown	Unknown	Unknown	Unknown	Unknown	Group 1	Cluster 1
KF098956.1.1345	bOTU	0.000108	32	Firmicutes	Clostridia	Clostridiales	Family XI	Finegoldia	Group 1	Cluster 6
v1000755642	vOTU	0.017888	56	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio	Group 2	Other
HK555708.1.1441	bOTU	0.006227	32	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Brevundimonas	Group 2	Cluster 7
JQ458561.1.1388	bOTU	0.004668	26	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus	Group 2	Cluster 2
JF417721.1.1485	bOTU	0.002708	24	Actinobacteria	Actinobacteria	Propionibacteriales	Nocardioidaceae	Nocardioides	Group 2	Cluster 7
AB089480.1.1490	bOTU	0.002315	19	Actinobacteria	Actinobacteria	Micrococcales	Intrasporangiaceae	Janibacter	Group 2	Cluster 7
New.ReferenceOTU1110	bOTU	0.001890	26	Unknown	Unknown	Unknown	Unknown	Unknown	Group 2	Cluster 1
HQ323460.1.1460	bOTU	0.001804	18	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	Roseomonas	Group 2	Cluster 7
KF712542.1.1494	bOTU	0.001688	24	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	Massilia	Group 2	Cluster 8

Continued on next page

OTU ID	OTU type	Mean rel. abun.	Sample detected	Phylum	Class	Order	Family	Genus	Group	Cluster
KT365979.1.1460	bOTU	0.001367	18	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Arthrobacter	Group 2	Cluster 7
New.ReferenceOTU58	bOTU	0.001347	23	Unknown	Unknown	Unknown	Unknown	Unknown	Group 2	Cluster 1
LN555083.1.1419	bOTU	0.000626	21	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Bradyrhizobium	Group 2	Cluster 8
HM587324.1.1489	bOTU	0.000575	21	Firmicutes	Clostridia	Clostridiales	Family XI	Murdochiella	Group 2	Cluster 1
AAAA02020712.626.2096	bOTU	0.000546	21	Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Methylobacterium	Group 2	Cluster 8
New.ReferenceOTU329	bOTU	0.000418	25	Unknown	Unknown	Unknown	Unknown	Unknown	Group 2	Cluster 1
JQ465848.1.1395	bOTU	0.000375	20	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella 7	Group 2	Cluster 1
New.ReferenceOTU780	bOTU	0.000299	21	Unknown	Unknown	Unknown	Unknown	Unknown	Group 2	Cluster 1
v1000527313	vOTU	0.079518	57	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Salmonella	Group 3	Other
v1000751529	vOTU	0.068450	57	Unknown	Unknown	Unknown	Unknown	Unknown	Group 3	Other
v1000344231	vOTU	0.061238	54	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	Neisseria	Group 3	Other
v1000752872	vOTU	0.037888	32	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus	Group 3	Cluster 5
D14146.1.1375	bOTU	0.031583	52	Firmicutes	Clostridia	Clostridiales	Family XI	Anaerococcus	Group 3	Cluster 1
v1000748918	vOTU	0.023483	26	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus haemolyticus	Group 3	Cluster 5
v1000748834	vOTU	0.018284	51	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	Group 3	Cluster 5
v379329	vOTU	0.013381	50	Unknown	Unknown	Unknown	Unknown	Unknown	Group 3	Other
v1000681193	vOTU	0.005145	36	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	Group 3	Other
v1000588946	vOTU	0.004918	44	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	Group 3	Other
JF345182.1.1505	bOTU	0.001738	40	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Stenotrophomonas	Group 3	Other
New.ReferenceOTU987	bOTU	0.001009	24	Unknown	Unknown	Unknown	Unknown	Unknown	Group 3	Cluster 9
New.ReferenceOTU1116	bOTU	0.000965	28	Unknown	Unknown	Unknown	Unknown	Unknown	Group 3	Cluster 9
New.ReferenceOTU119	bOTU	0.000933	19	Unknown	Unknown	Unknown	Unknown	Unknown	Group 3	Cluster 9
New.ReferenceOTU749	bOTU	0.000241	29	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Unknown	Group 3	Other
v1000757850	vOTU	0.165265	51	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Proteus	Group 4	Cluster 4
v1000377947	vOTU	0.144631	42	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinobaculum	Group 4	Cluster 4
v1000319712	vOTU	0.034216	45	Unknown	Unknown	Unknown	Unknown	Unknown	Group 4	Cluster 4
v1000295641	vOTU	0.021935	46	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium	Group 4	Cluster 3
v1000754595	vOTU	0.010341	36	Unknown	Unknown	Unknown	Unknown	Unknown	Group 4	Cluster 3
v1000350480	vOTU	0.009148	41	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	Group 4	Cluster 4
v1000741140	vOTU	0.008889	40	Unknown	Unknown	Unknown	Unknown	Unknown	Group 4	Cluster 4
v1000142855	vOTU	0.004544	40	Unknown	Unknown	Unknown	Unknown	Unknown	Group 4	Cluster 4

Continued on next page

OTU ID	OTU type	Mean rel. abun.	Sample detected	Phylum	Class	Order	Family	Genus	Group	Cluster
v1000094829	vOTU	0.004445	39	Unknown	Unknown	Unknown	Unknown	Unknown	Group 4	Cluster 4
v1000573340	vOTU	0.002722	37	Unknown	Unknown	Unknown	Unknown	Unknown	Group 4	Cluster 4
v1000069610	vOTU	0.001780	39	Unknown	Unknown	Unknown	Unknown	Unknown	Group 4	Cluster 4
v1000487836	vOTU	0.001473	39	Unknown	Unknown	Unknown	Unknown	Unknown	Group 4	Cluster 4
v1000733923	vOTU	0.001028	32	Unknown	Unknown	Unknown	Unknown	Unknown	Group 4	Cluster 4
v1000741139	vOTU	0.000610	35	Unknown	Unknown	Unknown	Unknown	Unknown	Group 4	Cluster 4
v1000583763	vOTU	0.000572	33	Unknown	Unknown	Unknown	Unknown	Unknown	Group 4	Cluster 4
JF188243.1.1374	bOTU	0.198745	60	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus aureus	Group 5	Other
v1000756557	vOTU	0.135812	57	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Group 5	Cluster 2
EU465496.1.1354	bOTU	0.085704	56	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium 1	Group 5	Cluster 1
HG972968.1.1404	bOTU	0.033241	42	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Enterobacter	Group 5	Cluster 2
DQ798826.1.1408	bOTU	0.025253	47	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Group 5	Cluster 2
DQ807591.1.1392	bOTU	0.017883	36	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Group 5	Cluster 2
v1000748917	vOTU	0.017549	43	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Group 5	Cluster 2
FJ671786.1.1387	bOTU	0.016789	31	Unknown	Unknown	Unknown	Unknown	Unknown	Group 5	Cluster 1
JQ475328.1.1395	bOTU	0.006988	22	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium	Group 5	Cluster 2
ACTE01000019.540.2030	bOTU	0.005334	25	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Peptoclostridium	Group 5	Cluster 2
EU459226.1.1333	bOTU	0.004643	34	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus	Group 5	Cluster 2
LFCI01000011.1.1253	bOTU	0.004244	29	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Salmonella	Group 5	Cluster 2
AY191849.1.1359	bOTU	0.001386	41	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Ralstonia	Group 5	Other
New.ReferenceOTU833	bOTU	0.001027	23	Unknown	Unknown	Unknown	Unknown	Unknown	Group 5	Cluster 2
FJ557822.1.1408	bOTU	0.126985	60	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus capitis	Group 6	Cluster 5
New.ReferenceOTU411	bOTU	0.027142	42	Unknown	Unknown	Unknown	Unknown	Unknown	Group 6	Cluster 6
FJ557977.1.1384	bOTU	0.017609	50	Firmicutes	Clostridia	Clostridiales	Family XI	Peptoniphilus	Group 6	Cluster 6
DQ532350.1.1524	bOTU	0.003219	39	Firmicutes	Bacilli	Bacillales	Unknown	Unknown	Group 6	Cluster 5
New.ReferenceOTU240	bOTU	0.002986	35	Actinobacteria	Actinobacteria	Micrococcales	Dermabacteraceae	Unknown	Group 6	Cluster 6
New.ReferenceOTU789	bOTU	0.000820	33	Unknown	Unknown	Unknown	Unknown	Unknown	Group 6	Cluster 6
New.ReferenceOTU772	bOTU	0.000457	19	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Facklamia	Group 6	Cluster 6
GQ448359.1.1396	bOTU	0.000421	20	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Unknown	Group 6	Other
FJ957846.1.1474	bOTU	0.000362	38	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus epidermidis	Group 6	Cluster 5

Continued on next page

OTU ID	OTU type	Mean rel. abun.	Sample de-tected	Phylum	Class	Order	Family	Genus	Group	Cluster
FJ957668.1.1482	bOTU	0.000028	24	Firmicutes	Bacilli	Bacillales	Unknown	Unknown	Group 6	Cluster 5
v1000528030	vOTU	0.019351	43	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella	Group 7	Cluster 3
v1000528681	vOTU	0.013882	33	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Enterobacter	Group 7	Cluster 2
v1000754260	vOTU	0.008591	23	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	Campylobacter	Group 7	Cluster 1
v1000122390	vOTU	0.005185	25	Unknown	Unknown	Unknown	Unknown	Unknown	Group 7	Cluster 3
v1000297408	vOTU	0.004320	22	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Veillonella	Group 7	Cluster 3
v1000589052	vOTU	0.004153	30	Unknown	Unknown	Unknown	Unknown	Unknown	Group 7	Cluster 2
v1000582339	vOTU	0.003671	26	Unknown	Unknown	Unknown	Unknown	Unknown	Group 7	Cluster 3
v1000317293	vOTU	0.003102	38	Unknown	Unknown	Unknown	Unknown	Unknown	Group 7	Other
v1000583756	vOTU	0.002394	29	Unknown	Unknown	Unknown	Unknown	Unknown	Group 7	Cluster 3
v1000759990	vOTU	0.001918	27	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Citrobacter	Group 7	Cluster 2
v1000748726	vOTU	0.016393	35	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus aureus	Group 8	Cluster 5
v1000540970	vOTU	0.002883	26	Unknown	Unknown	Unknown	Unknown	Unknown	Group 8	Cluster 3
New.ReferenceOTU818	bOTU	0.002836	30	Unknown	Unknown	Unknown	Unknown	Unknown	Group 8	Other
v1000132826	vOTU	0.002055	28	Unknown	Unknown	Unknown	Unknown	Unknown	Group 8	Cluster 3
v1000530252	vOTU	0.001977	26	Unknown	Unknown	Unknown	Unknown	Unknown	Group 8	Cluster 3
v1000348340	vOTU	0.001341	25	Unknown	Unknown	Unknown	Unknown	Unknown	Group 8	Cluster 3
v1000588167	vOTU	0.001259	23	Unknown	Unknown	Unknown	Unknown	Unknown	Group 8	Cluster 3
v1000528358	vOTU	0.001078	25	Unknown	Unknown	Unknown	Unknown	Unknown	Group 8	Cluster 3
v1000739801	vOTU	0.000992	28	Unknown	Unknown	Unknown	Unknown	Unknown	Group 8	Cluster 3
HK557047.45.1575	bOTU	0.035230	56	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus hominis	Other	Other
AB971806.1.1489	bOTU	0.022365	44	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Propionibacterium	Other	Other
New.ReferenceOTU601	bOTU	0.013008	47	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus pettenkoferi	Other	Other
HQ323458.1.1504	bOTU	0.011559	33	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Micrococcus	Other	Other
FJ674767.1.1391	bOTU	0.010162	32	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Kocuria	Other	Other
v1000748048	vOTU	0.009105	44	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bartonellaceae	Bartonella	Other	Other
JX668718.1.1298	bOTU	0.005732	57	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Rhizobium	Other	Other
KF100049.1.1347	bOTU	0.002941	36	Actinobacteria	Actinobacteria	Micrococcales	Brevibacteriaceae	Brevibacterium	Other	Other

Continued on next page

OTU ID	OTU type	Mean rel. abun.	Sample de-tected	Phylum	Class	Order	Family	Genus	Group	Cluster
JN178817.1.1488	bOTU	0.002857	30	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas	Other	Other
New.ReferenceOTU251	bOTU	0.002595	32	Unknown	Unknown	Unknown	Unknown	Unknown	Other	Other
KF924610.1.1424	bOTU	0.002293	29	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Paracoccus	Other	Other
GU731299.1.1390	bOTU	0.001971	57	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Unknown	Other	Other
EF510348.1.1506	bOTU	0.001834	38	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus haemolyticus	Other	Other
JX047062.1.1485	bOTU	0.001483	17	Actinobacteria	Actinobacteria	Streptomycetales	Streptomycetaceae	Streptomyces	Other	Other
AF467407.1.1549	bOTU	0.001412	39	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus cohnii	Other	Other
KJ808213.1.1451	bOTU	0.001390	35	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	Mesorhizobium	Other	Other
GALZ01007484.194.1638	bOTU	0.001084	23	Cyanobacteria	Chloroplast	Unknown	Unknown	Unknown	Other	Other
FJ865214.1.1481	bOTU	0.000980	25	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Microbacterium	Other	Other
AY422712.1.1545	bOTU	0.000942	20	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Aerococcus	Other	Other
GU940889.1.1356	bOTU	0.000744	18	Actinobacteria	Actinobacteria	Frankiales	Geodermatophilaceae	Blastococcus	Other	Other
KF842207.1.1408	bOTU	0.000723	21	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	Other	Other
v1000753739	vOTU	0.000703	33	Unknown	Unknown	Unknown	Unknown	Unknown	Other	Other
New.ReferenceOTU967	bOTU	0.000686	26	Actinobacteria	Actinobacteria	Corynebacteriales	Unknown	Unknown	Other	Other
FJ671979.1.1378	bOTU	0.000668	17	Actinobacteria	Actinobacteria	Corynebacteriales	Dietziaceae	Dietzia	Other	Other
HQ697766.1.1452	bOTU	0.000650	17	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Rubellimicrobium	Other	Other
ACOR01000003.1451.2935	bOTU	0.000617	32	Proteobacteria	Alphaproteobacteria	Rhizobiales	Brucellaceae	Unknown	Other	Other
v1000079907	vOTU	0.000492	30	Unknown	Unknown	Unknown	Unknown	Unknown	Other	Other
JQ800847.1.1452	bOTU	0.000403	22	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Erythrobacteraceae	Altererythrobacter	Other	Other
EU773827.1.1406	bOTU	0.000335	21	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus	Other	Other
HQ143300.1.1507	bOTU	0.000333	18	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Pantoea	Other	Other
EU328097.1.1450	bOTU	0.000312	25	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Erythrobacteraceae	Unknown	Other	Other
KF859635.1.1495	bOTU	0.000291	23	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	Halomonas	Other	Other
JF514555.1.1530	bOTU	0.000233	25	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia-Paraburkholderia	Other	Other
KF842111.1.1439	bOTU	0.000199	21	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Veillonella	Other	Other
GZ793541.182.1466	bOTU	0.000126	17	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	Neisseria	Other	Other
New.CleanUp.ReferenceOTU3225	bOTU	0.000062	21	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	Noviherbaspirillum	Other	Other
GQ135081.1.1362	bOTU	0.000058	17	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Atopostipes	Other	Other
JN177850.1.1490	bOTU	0.000008	16	Proteobacteria	Alphaproteobacteria	Rhizobiales	Unknown	Unknown	Other	Other

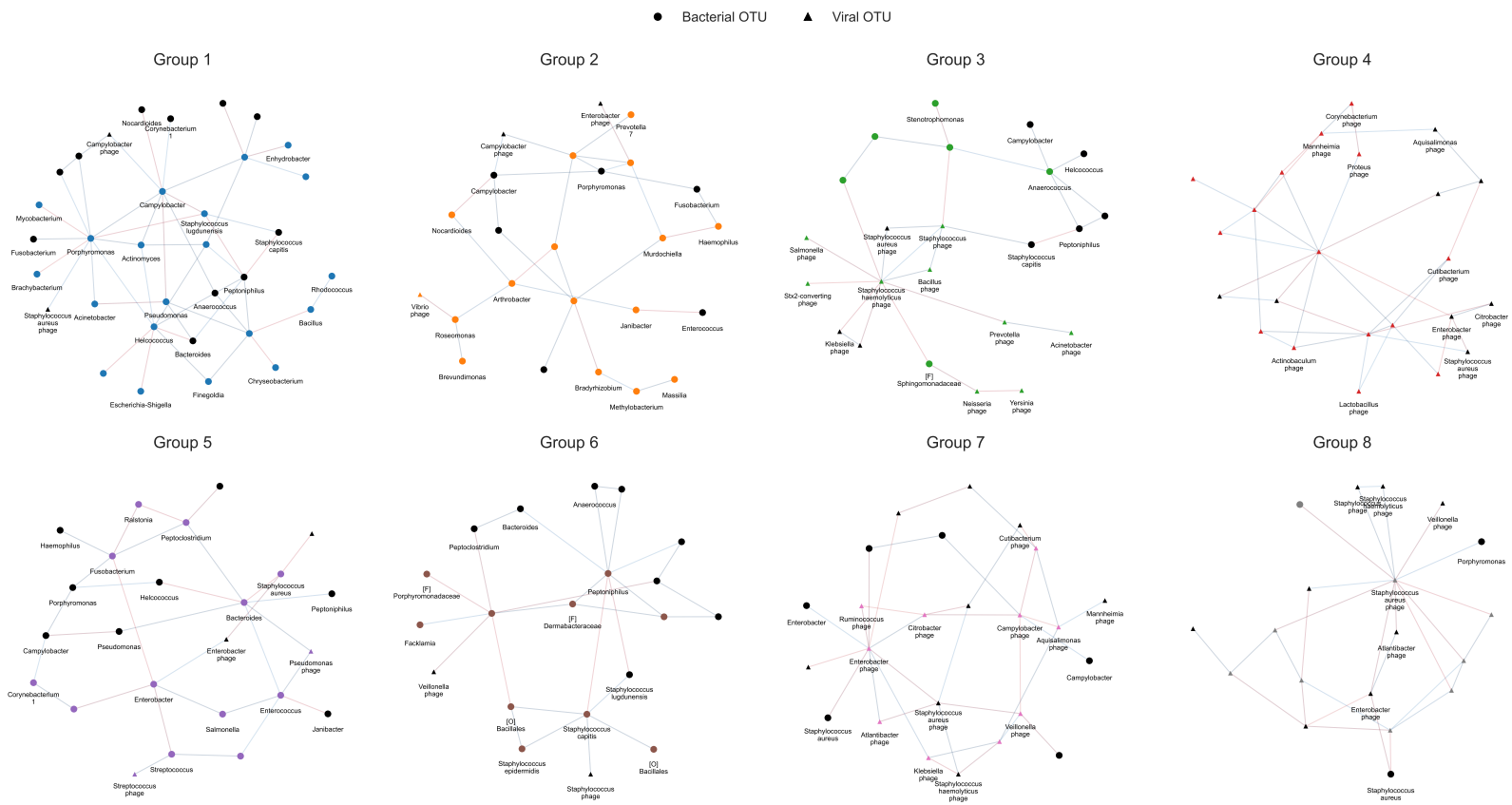


Figure A.24: "Topological Groups" identified using Louvain algorithm. All nodes belong to the Group (colored) and adjacent neighbors (black) were shown.

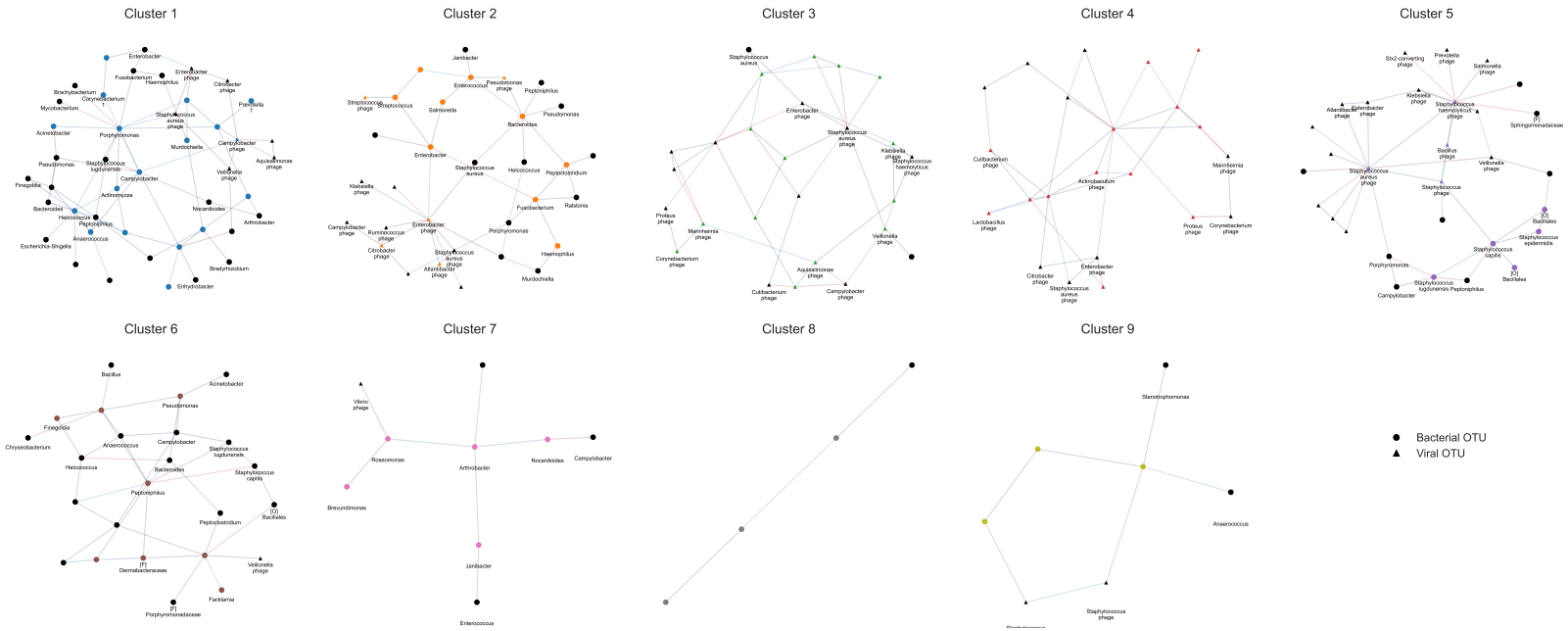


Figure A.25: "Clusters" identified using Louvain algorithm on the graph with edges of positive correlations. All nodes belong to the Cluster (colored) and adjacent neighbors (black) were shown.

Bibliography

[1] J. A. Gilbert and S. V. Lynch, *Community ecology as a framework for human microbiome research*, *Nature Medicine* **25** (June, 2019) 884–889.

[2] F. Sanger, S. Nicklen, and A. R. Coulson, *DNA sequencing with chain-terminating inhibitors*, *Proceedings of the National Academy of Sciences* **74** (Dec., 1977) 5463–5467.

[3] 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, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint,

S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, *The Sequence of the Human Genome*, *Science* **291** (Feb., 2001) 1304–1351.

[4] M. Kircher and J. Kelso, *High-throughput DNA sequencing – concepts and limitations*, *BioEssays* **32** (2010), no. 6 524–536.

[5] J. A. Reuter, D. V. Spacek, and M. P. Snyder, *High-Throughput Sequencing Technologies*, *Molecular Cell* **58** (May, 2015) 586–597.

[6] J. S. Johnson, D. J. Spakowicz, B.-Y. Hong, L. M. Petersen, P. Demkowicz, L. Chen, S. R. Leopold, B. M. Hanson, H. O. Agresta, M. Gerstein, E. Sodergren, and G. M. Weinstock, *Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis*, *Nature Communications* **10** (Nov., 2019) 5029.

[7] K. A. Seifert, *Progress towards DNA barcoding of fungi*, *Molecular Ecology Resources* **9** (2009), no. s1 83–89.

[8] R. Stark, M. Grzelak, and J. Hadfield, *RNA sequencing: the teenage years*, *Nature Reviews Genetics* **20** (Nov., 2019) 631–656.

[9] E. L. van Dijk, Y. Jaszczyszyn, D. Naquin, and C. Thermes, *The Third Revolution in Sequencing Technology*, *Trends in Genetics* **34** (Sept., 2018) 666–681.

[10] J. M. Janda and S. L. Abbott, *16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls*, *Journal of Clinical Microbiology* **45** (Sept., 2007) 2761–2764.

[11] 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** (May, 2010) 511–515.

[12] Y. Shen, A. Pressman, E. Janzen, and I. A. Chen, *Kinetic sequencing (k-Seq) as a massively parallel assay for ribozyme kinetics: utility and critical parameters*, *Nucleic Acids Research* **49** (July, 2021) e67–e67.

[13] V. Dhamodharan, S. Kobori, and Y. Yokobayashi, *Large Scale Mutational and Kinetic Analysis of a Self-Hydrolyzing Deoxyribozyme*, *ACS Chemical Biology* **12** (Dec., 2017) 2940–2945.

[14] Y. Yokobayashi, *High-Throughput Analysis and Engineering of Ribozymes and Deoxyribozymes by Sequencing*, *Accounts of Chemical Research* **53** (Dec., 2020) 2903–2912.

[15] C. N. Niland, E. Jankowsky, and M. E. Harris, *Optimization of High-Throughput Sequencing Kinetics for determining enzymatic rate constants of thousands of RNA substrates*, *Analytical biochemistry* **510** (Oct., 2016) 1–10.

[16] D. M. Fowler and S. Fields, *Deep mutational scanning: a new style of protein science*, *Nature Methods* **11** (Aug., 2014) 801–807.

[17] A. D. Pressman, Z. Liu, E. Janzen, C. Blanco, U. F. Müller, G. F. Joyce, R. Pascal, and I. A. Chen, *Mapping a Systematic Ribozyme Fitness Landscape Reveals a Frustrated Evolutionary Network for Self-Aminoacylating RNA*, *Journal of the American Chemical Society* **141** (Apr., 2019) 6213–6223.

[18] J. D. Robin, A. T. Ludlow, R. LaRanger, W. E. Wright, and J. W. Shay, *Comparison of DNA Quantification Methods for Next Generation Sequencing*, *Scientific Reports* **6** (Apr., 2016) 24067.

[19] D. Aird, M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke, *Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries*, *Genome Biology* **12** (Feb., 2011) R18.

[20] R. Xulvi-Brunet, G. W. Campbell, S. Rajamani, J. I. Jiménez, and I. A. Chen, *Computational analysis of fitness landscapes and evolutionary networks from in vitro evolution experiments*, *Methods* **106** (Aug., 2016) 86–96.

[21] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, *DADA2: High-resolution sample inference from Illumina amplicon data*, *Nature Methods* **13** (July, 2016) 581–583.

[22] T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale, *Counting absolute numbers of molecules using unique molecular identifiers*, *Nature Methods* **9** (Jan., 2012) 72–74.

[23] N.-P. Nguyen, T. Warnow, M. Pop, and B. White, *A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity*, *npj Biofilms and Microbiomes* **2** (Apr., 2016) 1–8.

[24] D. E. Wood, J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2*, *Genome Biology* **20** (Nov., 2019) 257.

[25] J. Kuczynski, J. Stombaugh, W. A. Walters, A. González, J. G. Caporaso, and R. Knight, *Using QIIME to Analyze 16S rRNA Gene Sequences from Microbial Communities*, *Current Protocols in Bioinformatics* **36** (Dec., 2011).

[26] J. Lu, F. P. Breitwieser, P. Thielen, and S. L. Salzberg, *Bracken: estimating species abundance in metagenomics data*, *PeerJ Computer Science* **3** (Jan., 2017) e104.

[27] S. Jiang, G. Xiao, A. Y. Koh, Y. Chen, B. Yao, Q. Li, and X. Zhan, *HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity*, *Frontiers in Genetics* **11** (June, 2020) 445.

[28] M. J. Ha, J. Kim, J. Galloway-Peña, K.-A. Do, and C. B. Peterson, *Compositional zero-inflated network estimation for microbiome data*, *BMC Bioinformatics* **21** (Dec., 2020) 581.

[29] A. Tkacz, M. Hortala, and P. S. Poole, *Absolute quantitation of microbiota abundance in environmental samples*, *Microbiome* **6** (June, 2018) 110.

[30] J. Aitchison, *A new approach to null correlations of proportions*, *Journal of the International Association for Mathematical Geology* **13** (Apr., 1981) 175–189.

[31] J. Chiquet, M. Mariadassou, and S. Robin, *The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances*, *Frontiers in Ecology and Evolution* **9** (Mar., 2021) 588292.

[32] S. Kobori and Y. Yokobayashi, *High-Throughput Mutational Analysis of a Twister Ribozyme*, *Angewandte Chemie International Edition* **55** (July, 2016) 10354–10357.

[33] J. O. L. Andreasson, A. Savinov, S. M. Block, and W. J. Greenleaf, *Comprehensive sequence-to-function mapping of cofactor-dependent RNA catalysis in the glmS ribozyme*, *Nature Communications* **11** (Apr., 2020) 1663.

[34] J. M. Tome, A. Ozer, J. M. Pagano, D. Gheba, G. P. Schroth, and J. T. Lis, *Comprehensive analysis of RNA-protein interactions by high-throughput sequencing–RNA affinity profiling*, *Nature Methods* **11** (June, 2014) 683–688.

[35] D. D. Le, T. C. Shimko, A. K. Aditham, A. M. Keys, S. A. Longwell, Y. Orenstein, and P. M. Fordyce, *Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding*, *Proceedings of the National Academy of Sciences* **115** (Apr., 2018) E3702–E3711.

[36] F. Jalali-Yazdi, L. Huong-Lai, T. T. Takahashi, and R. W. Roberts, *High-Throughput Measurement of Binding Kinetics by mRNA Display and Next-Generation Sequencing*, *Angewandte Chemie International Edition* **55** (Feb., 2016) 4007–4010.

[37] D. S. Tack, P. D. Tonner, A. Pressman, N. D. Olson, S. F. Levy, E. F. Romantseva, N. Alperovich, O. Vasilyeva, and D. Ross, *The genotype-phenotype landscape of an allosteric protein*, *Molecular Systems Biology* **17** (Mar., 2021).

[38] A. Pressman, J. E. Moretti, G. W. Campbell, U. F. Müller, and I. A. Chen, *Analysis of in vitro evolution reveals the underlying distribution of catalytic activity among random sequences*, *Nucleic Acids Research* **45** (Aug., 2017) 8167–8179.

[39] K. E. Hines, T. R. Middendorf, and R. W. Aldrich, *Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach*, *The Journal of General Physiology* **143** (Mar., 2014) 401–416.

[40] C. Blanco, S. Verbanic, B. Seelig, and I. A. Chen, *EasyDIVER: A Pipeline for Assembling and Counting High-Throughput Sequencing Data from In Vitro Evolution of Nucleic Acids or Peptides*, *Journal of Molecular Evolution* **88** (Aug., 2020) 477–481.

[41] A. P. Masella, A. K. Bartram, J. M. Truszkowski, D. G. Brown, and J. D. Neufeld, *PANDAseq: paired-end assembler for illumina sequences*, *BMC Bioinformatics* **13** (Feb., 2012) 31.

[42] A. C. Daly, D. Gavaghan, J. Cooper, and S. Tavener, *Inference-based assessment of parameter identifiability in nonlinear biological models*, *Journal of The Royal Society Interface* **15** (July, 2018) 20180318.

[43] L. K. Ursell, J. L. Metcalf, L. W. Parfrey, and R. Knight, *Defining the Human Microbiome*, *Nutrition reviews* **70** (Aug., 2012) S38–S44.

[44] R. Sender, S. Fuchs, and R. Milo, *Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans*, *Cell* **164** (Jan., 2016) 337–340.

[45] J. A. Gilbert, M. J. Blaser, J. G. Caporaso, J. K. Jansson, S. V. Lynch, and R. Knight, *Current understanding of the human microbiome*, *Nature Medicine* **24** (Apr., 2018) 392–400.

[46] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, *The Human Microbiome Project*, *Nature* **449** (Oct., 2007) 804–810.

[47] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, and J. Wang, *A human gut microbial gene catalog established by metagenomic sequencing*, *Nature* **464** (Mar., 2010) 59–65.

[48] V. B. Young, *The role of the microbiome in human health and disease: an introduction for clinicians*, *BMJ* (Mar., 2017) j831.

[49] S. Verbanic, J. M. Deacon, and I. A. Chen, *The chronic wound virome: phage diversity and associations with wounds and healing outcomes*, *medRxiv* (Jan., 2022) 2022.01.05.22268807.

[50] E. Bellemain, T. Carlsen, C. Brochmann, E. Coissac, P. Taberlet, and H. Kauserud, *ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases*, *BMC Microbiology* **10** (July, 2010) 189.

[51] P. J. McMurdie and S. Holmes, *phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data*, *PLOS ONE* **8** (Apr., 2013) e61217.

[52] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber, *Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities*, *Applied and Environmental Microbiology* **75** (Dec., 2009) 7537–7541.

[53] A. L. Byrd, Y. Belkaid, and J. A. Segre, *The human skin microbiome*, *Nature Reviews Microbiology* **16** (Mar., 2018) 143–155.

[54] S. Verbanic, Y. Shen, J. Lee, J. M. Deacon, and I. A. Chen, *Microbial predictors of healing and short-term effect of debridement on the microbiome of chronic wounds*, *npj Biofilms and Microbiomes* **6** (Dec., 2020) 21.

[55] J. M. Montoya, S. L. Pimm, and R. V. Solé, *Ecological networks and their fragility*, *Nature* **442** (July, 2006) 259–264.

[56] F. Emmert-Streib, M. Dehmer, and B. Haibe-Kains, *Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks*, *Frontiers in cell and developmental biology* **2** (Aug., 2014) 38–38. Publisher: Frontiers Media S.A.

[57] C. L. Tucker, J. F. Gera, and P. Uetz, *Towards an understanding of complex protein networks*, *Trends in Cell Biology* **11** (Mar., 2001) 102–106.

[58] N. Friedman, *Inferring Cellular Networks Using Probabilistic Graphical Models*, *Science* **303** (Feb., 2004) 799–805.

[59] E. M. Airoldi, *Getting Started in Probabilistic Graphical Models*, *PLOS Computational Biology* **3** (Dec., 2007) e252.

[60] C. Glymour, K. Zhang, and P. Spirtes, *Review of Causal Discovery Methods Based on Graphical Models*, *Frontiers in Genetics* **10** (2019).

[61] B. Ma, H. Wang, M. Dsouza, J. Lou, Y. He, Z. Dai, P. C. Brookes, J. Xu, and J. A. Gilbert, *Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China*, *The ISME Journal* **10** (Aug., 2016) 1891–1901.

[62] D. Berry and S. Widder, *Deciphering microbial interactions and detecting keystone species with co-occurrence networks*, *Frontiers in Microbiology* **5** (2014).

[63] A. Barberán, S. T. Bates, E. O. Casamayor, and N. Fierer, *Using network analysis to explore co-occurrence patterns in soil microbial communities*, *The ISME Journal* **6** (Feb., 2012) 343–351.

[64] R. J. Williams, A. Howe, and K. S. Hofmockel, *Demonstrating microbial co-occurrence pattern analyses within and between ecosystems*, *Frontiers in Microbiology* **5** (2014).

[65] M. B. Araújo, A. Rozenfeld, C. Rahbek, and P. A. Marquet, *Using species co-occurrence networks to assess the impacts of climate change*, *Ecography* **34** (2011), no. 6 897–908.

[66] P. J. Auster, B. X. Semmens, and K. Barber, *Pattern in the Co-occurrence of Fishes Inhabiting the Coral Reefs of Bonaire, Netherlands Antilles*, *Environmental Biology of Fishes* **74** (Oct., 2005) 187–194.

[67] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower, *Microbial co-occurrence relationships in the human microbiome*, *PLoS computational biology* **8** (2012), no. 7 e1002606.

[68] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, *Biostatistics* **9** (July, 2008) 432–441.

[69] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau, *Sparse and compositionally robust inference of microbial ecological networks*, *PLoS computational biology* **11** (May, 2015) e1004226.

[70] S. He and M. Deng, *Direct interaction network and differential network inference from compositional data via lasso penalized D-trace loss*, *PLOS ONE* **14** (July, 2019) e0207731.

[71] J. Chiquet, S. Robin, and M. Mariadassou, *Variational Inference for sparse network reconstruction from count data*, in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 1162–1171, PMLR, June, 2019.

[72] L. Tipton, C. L. Müller, Z. D. Kurtz, L. Huang, E. Kleerup, A. Morris, R. Bonneau, and E. Ghedin, *Fungi stabilize connectivity in the lung and skin microbial ecosystems*, *Microbiome* **6** (Dec., 2018) 12.

[73] J. Yin and H. Li, *A sparse conditional Gaussian graphical model for analysis of genetical genomics data*, *The Annals of Applied Statistics* **5** (Dec., 2011).

[74] H. K. Allen, D. O. Bayles, T. Looft, J. Trachsel, B. E. Bass, D. P. Alt, S. M. D. Bearson, T. Nicholson, and T. A. Casey, *Pipeline for amplifying and analyzing amplicons of the V1–V3 region of the 16S rRNA gene*, *BMC Research Notes* **9** (Aug., 2016) 380.

[75] S. Conlan, H. H. Kong, and J. A. Segre, *Species-Level Analysis of DNA Sequence Data from the NIH Human Microbiome Project*, *PLOS ONE* **7** (Oct., 2012) e47075.

[76] R. Foygel and M. Drton, *Extended Bayesian Information Criteria for Gaussian Graphical Models*, *arXiv:1011.6640 [math, stat]* (Nov., 2010).

[77] H. Liu, K. Roeder, and L. Wasserman, *Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models*, *arXiv:1006.3316 [stat]* (June, 2010).

[78] G. Lafit, F. Tuerlinckx, I. Myin-Germeys, and E. Ceulemans, *A Partial Correlation Screening Approach for Controlling the False Positive Rate in Sparse Gaussian Graphical Models*, *Scientific Reports* **9** (Dec., 2019) 17759.

[79] M. J. Ha and W. Sun, *Partial correlation matrix estimation using ridge penalty followed by thresholding and re-estimation*, *Biometrics* **70** (2014), no. 3 762–770.

[80] Z. Liu, A. Ma, E. Mathé, M. Merling, Q. Ma, and B. Liu, *Network analyses in microbiome based on high-throughput multi-omics data*, *Briefings in Bioinformatics* **22** (Mar., 2021) 1639–1655.

[81] M. Layeghifard, D. M. Hwang, and D. S. Guttman, *Disentangling Interactions in the Microbiome: A Network Perspective*, *Trends in Microbiology* **25** (Mar., 2017) 217–228.

[82] Jianbo Shi and J. Malik, *Normalized cuts and image segmentation*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22** (Aug., 2000) 888–905.

[83] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large networks*, *Journal of Statistical Mechanics: Theory and Experiment* **2008** (Oct., 2008) P10008.

[84] V. A. Traag, L. Waltman, and N. J. van Eck, *From Louvain to Leiden: guaranteeing well-connected communities*, *Scientific Reports* **9** (Mar., 2019) 5233.

[85] A. Clauset, M. E. J. Newman, and C. Moore, *Finding community structure in very large networks*, *Physical Review E* **70** (Dec., 2004) 066111.

[86] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, *An efficient algorithm for large-scale detection of protein families*, *Nucleic Acids Research* **30** (Apr., 2002) 1575–1584.

[87] M. Rosvall and C. T. Bergstrom, *Maps of random walks on complex networks reveal community structure*, *Proceedings of the National Academy of Sciences* **105** (Jan., 2008) 1118–1123.

[88] S. Emmons, S. Kobourov, M. Gallant, and K. Börner, *Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale*, *PLOS ONE* **11** (July, 2016) e0159161.

[89] M. A. Jackson, M. J. Bonder, Z. Kuncheva, J. Zierer, J. Fu, A. Kurilshikov, C. Wijmenga, A. Zhernakova, J. T. Bell, T. D. Spector, and C. J. Steves, *Detection of stable community structures within gut microbiota co-occurrence networks from different human populations*, *PeerJ* **6** (Feb., 2018) e4303.

[90] J. L. Riera and L. Baldo, *Microbial co-occurrence networks of gut microbiota reveal community conservation and diet-associated shifts in cichlid fishes*, *Animal Microbiome* **2** (Sept., 2020) 36.

[91] N. M. Davis, D. M. Proctor, S. P. Holmes, D. A. Relman, and B. J. Callahan, *Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data*, *Microbiome* **6** (Dec., 2018) 226.

[92] M. Sfeir, P. Youssef, and J. E. Mokhbat, *Salmonella typhi sternal wound infection*, *American Journal of Infection Control* **41** (Dec., 2013) e123–e124.

[93] X. Li, Y. Chen, W. Gao, W. Ouyang, J. Wei, and Z. Wen, *Epidemiology and Outcomes of Complicated Skin and Soft Tissue Infections among Inpatients in Southern China from 2008 to 2013*, *PLOS ONE* **11** (Feb., 2016) e0149960.

[94] V. Ki and C. Rotstein, *Bacterial skin and soft tissue infections in adults: A review of their epidemiology, pathogenesis, diagnosis, treatment and site of care*, *The Canadian Journal of Infectious Diseases & Medical Microbiology* **19** (Mar., 2008) 173–184.

[95] J. D. Williams, *Activity of Imipenem Against Pseudomonas and Bacteroides Species*, *Reviews of Infectious Diseases* **7** (July, 1985) S411–S416.

[96] D. C. Wu, W. W. Chan, A. I. Metelitsa, L. Fiorillo, and A. N. Lin, *Pseudomonas Skin Infection*, *American Journal of Clinical Dermatology* **12** (June, 2011) 157–169.

[97] C. Erridge, A. Pridmore, A. Eley, J. Stewart, and I. R. Poxton, *Lipopolysaccharides of Bacteroides fragilis, Chlamydia trachomatis and Pseudomonas aeruginosa signal via toll-like receptor 2*, *Journal of Medical Microbiology* **53** (Aug., 2004) 735–740.

[98] M. Otto, *Staphylococcus epidermidis – the "accidental" pathogen*, *Nature reviews. Microbiology* **7** (Aug., 2009) 555–567.

[99] D. Cameron, J.-H. Jiang, K. Hassan, L. Elbourne, K. Tuck, I. Paulsen, and A. Peleg, *Insights on virulence from the complete genome of Staphylococcus capitis*, *Frontiers in Microbiology* **6** (2015).

[100] M. Sabaté Brescó, L. G. Harris, K. Thompson, B. Stanic, M. Morgenstern, L. O'Mahony, R. G. Richards, and T. F. Moriarty, *Pathogenic Mechanisms and Host Interactions in Staphylococcus epidermidis Device-Related Infection*, *Frontiers in Microbiology* **8** (Aug., 2017) 1401.

[101] N. Bostanci, R. P. Allaker, G. N. Belibasakis, M. Rangarajan, M. A. Curtis, F. J. Hughes, and I. J. McKay, *Porphyromonas gingivalis antagonises Campylobacter rectus induced cytokine production by human monocytes*, *Cytokine* **39** (Aug., 2007) 147–156.

[102] J. Y. W. Lam, A. K. L. Wu, D. C. Ngai, J. L. L. Teng, E. S. Y. Wong, S. K. P. Lau, R. A. Lee, and P. C. Y. Woo, *Three Cases of Severe Invasive Infections Caused by Campylobacter rectus and First Report of Fatal C. rectus Infection*, *Journal of Clinical Microbiology* **49** (Apr., 2011) 1687–1691.

[103] E. Aronesty, *Comparison of Sequencing Utility Programs*, *The Open Bioinformatics Journal* **7** (Jan., 2013) 1–8.

[104] A. M. Bolger, M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data*, *Bioinformatics* **30** (Aug., 2014) 2114–2120.

[105] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools*, *Nucleic Acids Research* **41** (Jan., 2013) D590–D596.

[106] M. Yazdi, M. Bouzari, and E. Ghaemi, *Isolation and Characterization of a Lytic Bacteriophage (vB_pmis-TH) and Its Application in Combination with Ampicillin against Planktonic and Biofilm Forms of Proteus mirabilis Isolated from Urinary Tract Infection*, *Microbial Physiology* **28** (2018), no. 1 37–46.

[107] J. R. Brister, D. Ako-adjei, Y. Bao, and O. Blinkova, *NCBI Viral Genomes Resource*, *Nucleic Acids Research* **43** (Jan., 2015) D571–D577.

[108] D. Paez-Espino, I.-M. A. Chen, K. Palaniappan, A. Ratner, K. Chu, E. Szeto, M. Pillay, J. Huang, V. M. Markowitz, T. Nielsen, M. Huntemann, T. B. K Reddy, G. A. Pavlopoulos, M. B. Sullivan, B. J. Campbell, F. Chen, K. McMahon, S. J. Hallam, V. Denef, R. Cavicchioli, S. M. Caffrey, W. R. Streit, J. Webster, K. M. Handley, G. H. Salekdeh, N. Tsesmetzis, J. C. Setubal, P. B. Pope, W.-T. Liu, A. R. Rivers, N. N. Ivanova, and N. C. Kyrpides, *IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses*, *Nucleic Acids Research* **45** (Jan., 2017) D457–D465.

[109] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *Basic local alignment search tool*, *Journal of Molecular Biology* **215** (Oct., 1990) 403–410.