Title

Quantifying the extent of horizontal gene transfer in the genomes of Pink Berries

Author

Madejska, Ada Aleksandra

Publication Date

2024

Peer reviewed|Thesis/dissertation

University of California
Santa Barbara

Quantifying the extent of horizontal gene transfer in the genomes of Pink Berries

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

Doctor of Philosophy
in
Molecular, Cellular and Developmental Biology

by

Ada Aleksandra Madejska

Committee in charge:

Professor Boris Shraiman, Chair
Professor Kenneth S. Kosik
Professor David Low

September 2024

The Dissertation of Ada Aleksandra Madejska is approved.

Professor Kenneth S. Kosik

Professor David Low

Professor Boris Shraiman, Committee Chair

September 2024

Quantifying the extent of horizontal gene transfer in the genomes of Pink Berries

by

Ada Aleksandra Madejska

To my parents

Acknowledgements

First and foremost, I would like to thank my mentor, Boris Shraiman, for allowing me to work on these projects and for giving me the freedom and opportunities to keep learning new skills. I have learned a lot of useful skills throughout my time in his group. Thank you to my committee members: Kenneth Kosik and David Low for their constructive comments on my work.

I would also like to thank Otto Cordero, Lizzy Willbanks, and Alexander Petroff for sharing their data. It has been an interesting experience deep-diving into their datasets and untangling interesting results from them. Also, big thank-you to Nicolas Noll for his help with the 16S project and being very patient with me during our hours-long discussions. I appreciate all of your help and expertise.

Although working on one's research can feel very lonely, there have been many people that made my time at UCSB so much more enjoyable. I would like to thank my mentor and friend Arjun Rao for introducing me to the world of computational biology. Thank you to Lindsey Osimiri for being an excellent manager during my time at BigHat. Also big thank you to my labmates, Juliana Acosta-Uribe and Jemma Fendley. Our hangouts always left me feeling positive and ready to face the day.

Lastly, I would like to thank my friends and family. Thank you Carla Ladd and Julia Chung for your friendship. You made my time at UCSB so much better. Thank you to Yvonne Larkin for making sure that I'm okay and have fun once in a while. Thank you to Livien Bartkowska and Kasia Banaszek for their unconditional support. And, of course, thank you to my parents, Tomasz Madejski and Monika Madejska. Without their infinite support and love none of this would be possible.

Curriculum Vitæ
Ada Aleksandra Madejska

Education

2024 **Ph.D. in Molecular, Cellular and Developmental Biology,**
University of California, Santa Barbara.

2022 **M.A. in Molecular, Cellular and Developmental Biology,**
University of California, Santa Barbara.

2018 **B.S. in Molecular, Cellular and Developmental Biology with minor in Computer Science**
University of California, Santa Cruz

Professional Employment

2018 - Present **Graduate student researcher.** University of California Santa Barbara.

2023 **Data Science Intern**. Developed analyses and visualizations that related high-dimensional antibody sequence space to experimental metrics of antibody function and quality. BigHat, San Mateo, CA.

2021 **Santa Barbara Advanced School of Quantitative Biology Summer Research Course Participant**.
Participated in a creative, research-oriented course closely linked to the KITP program The Ecology and Evolution of Microbial Communities (ECOEVO21), KITP, Santa Barbara

2018-2023 **Teaching Assistant.** Systems Biology with MATLAB (graduate and undergraduate track), Introduction to Biology, Human Physiology, Cell Biology, Concepts in Biology, University Of California Santa Barbara.

2017-2018 **Undergraduate Research Assistant**. Tested and developed assessment tools for prediction of large genomics data using Python and R, UCSC Genomics Institute.

2016 **Summer Intern**.
Biotechnology Division of the University of Gdansk, Poland

Publications

- Rao A, Madejska A, Pfeil J, Paten B, Salama S, Haussler D.
  ProTECT - Prediction of T-Cell Epitopes for Cancer Therapy.
  Frontiers in Immunology, vol. 11, 2020, doi:10.3389/fimmu.2020.483296

- Toor J, Rao A, McShan A, Yarmarkovich M, Nerli S, Yamaguchi K, Madejska A, Nguyen S, Tripathi S, Maris J, Salama S, Haussler D, Sgouraki N. Genetic **A Recurrent Mutation in Anaplastic Lymphoma Kinase with Distinct Neoepitope Conformations.**
  Frontiers 2018 January 30 doi:https://doi.org/10.3389/fimmu.2018.00099

Abstract

Quantifying the impact of recombination on genome evolution within Pink Berries

by

Ada Aleksandra Madejska

Horizontal Gene Transfer (HGT) refers to sharing of small segments of genetic material from a donor to a recipient organism that does not have a parent-offspring relationship. Although not all transfers are successful, they are still abundant in natural bacterial communities. Moreover, the rate at which a gene is transferred differs widely across different genes and organisms. This makes it challenging, especially for healthcare professionals that study antimicrobial resistance and epidemics, to predict how the organism will evolve over time and over environmental fluctuations. Recombination creates more opportunities for a local adaptation by making it possible to acquire genes involved in antibiotic resistance, pathogenic determinants, etc. Many metagenomic studies have used natural biofilms in order to measure the rates of gene transfer across communities. Moreover, not only can one study genes responsible for protection against outside influence or metabolism but metagenomic studies of biofilms and microbial mats can provide invaluable insights into evolutionary processes within natural communities.

In the first chapter a comprehensive analysis of the Pink Berry metagenomic data is performed in context of extensive recombination of a quasi-sexual bacterial population. Evidence is presented showing that the populations are divided into clades with different evolutionary histories including a mixing layer where bacteria experience extensive recombination from other clades as well as with each other.

In the second part of this thesis I focus on self / non-self bacterial recognition. Toxin-antitoxin systems are important mechanisms for the bacteria to respond to intracellular

stress. Usually, the proteins that are a part of the contact dependent growth inhibition (CDI) secretion contain multiple distinct parts - the structure that acts as a delivery mechanism, the toxin, and the antitoxin. Here the preliminary findings of diversity in WapA and RhsC C-termini in purple sulfur bacteria is shown. The presence of large gaps in the alignment of the C-terminus region of the CDI proteins shows that the bacteria differ in their repertoire of toxins depending on their geographical location. These observations mark a promising starting point for studying CDI mechanisms in naturally occurring bacterial populations.

Lastly, the composition of bacterial communities grown in different conditions based on their 16S sequences is analyzed. In this chapter I have introduced a new tool for decreasing the error of 16S Nanopore sequences and identifying given samples. The accuracy of the pipeline using simulated PacBio and Nanopore datasets from CAMI2 is demonstrated and then it is used on experimental sediment data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Horizontal Gene Transfer

Horizontal Gene Transfer (HGT) refers to a sharing of a small piece of genetic material from a donor to a recipient organism that does not have a parent-offspring relationship. The history of HGT research dates back to the 1920s with experiments performed by Fredrik Griffith on what was then described as "genetic transformation" in *Streptococcus* bacteria[1]. In these experiments Griffith used two strains of bacteria (non virulent and virulent strains) and showed that even though a heat-killed virulent strain could not kill mice by themselves, when combined with a nonvirulent strain, the nonvirulent bacteria can become virulent. Given the novelty of these findings, most biologists and physicians found them hard to believe and the topic has not been revisited until twenty years later in 1944 by Oswald Avery where Avery along with Maclyn McCarty and Colin MacLeod continued Griffith's experiment and showed that bacteria, just like eukaryotic cells, contain genetic material [2]. They did so by consequently purifying the cellular components of the *Streptococcus* bacteria until they were left with the mysterious ”transforming principle” that they meticulously tested to reveal its identity. Joshua Lederberg, along with Edward Tatum, inspired by their results, started studying bacterial genetics and showed that gene exchange between bacteria is possible, coining the term “conjugation”[3] and earning a Nobel Prize for his research on bacterial HGT in 1958[4]. Although at the time the importance of HGT was not known, now it is regarded as an important evolutionary force in prokaryotes and a vital influence in creating genome plasticity.

Lederberg named one of the possible mechanisms through which bacteria exchange genetic information - conjugation - a transfer by direct contact. However, currently we are aware of several other means of genetic transfer. Transduction is a genetic transfer via phage while transformation is an uptake of DNA from the environment[5]. Moreover, other means of transfer are also known although have not been thoroughly studied yet. For example, outer membrane vesicles, packed with proteins, cell wall components, toxins, and chromosomal DNA, can be fused into bacterial or eukaryotic cells[6]. Another example is a gene transfer via “nanotubes”, an elongated extracellular structure, that allows a direct cell-to-cell contact and transport of cytoplasmic components[5, 7]. Indeed, there are many other mechanisms that help with information exchange and it is apparent that most likely more will be discovered in the future.

Transfer of genetic information between organisms is an important tool used for adaptation in bacteria. Changes in their environment, such as temperature increase/decrease, addition or extinction of predators or pray, availability of nutrients, etc. forces the organisms to quickly adapt. An ability to transfer and incorporate new pieces of DNA that may improve one’s survivability becomes a great advantage [8]. However, a successful uptake depends on many different factors. Although new pieces of DNA can bring advantageous traits, they can just as easily harm the organism. Therefore, the metabolic compatibility, compatible GC content, available gene expression systems, gene transfer mechanisms, and many other factors become crucial [9, 10, 11, 12]. The matter of conditions that have to be met for a successful transfer has been thoroughly researched. For example, Tock and Dryden [13] and other groups [14, 15] showed that if GC-content is too dissimilar between the bacteria, it is more likely that the transfer will be targeted by anti-HGT systems. Another example is the study of the gene copy numbers and how they influence the success of transfer since an increase in protein concentration inside the cell can negatively affect its fitness [16, 17].

It is fascinating that although there are many mechanisms preventing successful transfer, they are still abundant in the bacterial communities - a fact that the filed is learning even today. With the growing amount of data, the way we describe HGT and study it has been changing throughout the years. For example, a popular way of representing genetic relations between organisms is by using a phylogenetic tree where all organisms come from a common ancestor and each branch represents genetic changes that can only be passed to the next generation in a vertical manner [8]. However, when one studies bacteria, one need to keep in mind how HGT alters the genealogical relationships between strains in a lateral manner. This way, each fragment of DNA has its own genealogical history [18]. This observation sparked a discussion in the field of bacterial genomics, where some believe that rather than representing bacterial evolution as a bifurcating tree, a network-like graph gives a more comprehensive view on their relations[19].

Since the beginning of population genetics research, there has been an ever-growing interest in how HGT can alter the ecology and evolution of the complex bacterial communities. The focus still persists on understanding the important topic of bacterial acquisition of new antibiotic resistance and adaptation to new environments across the world. There are many questions that need to be addressed in order to gain a better understanding of HGT. What types of genes are most often transferred? How are they transferred and under what ecological pressures? Are there any genetic or physical barriers that slow down the transfer? At what rate does the transfer happen [20]? These questions are important especially if one wants to understand dynamics of naturally occurring communities that constantly experience new challenges to overcome in order to survive. Interestingly, the exact influence of HGT on such communities is not fully understood, especially since the rates, means, and conditions for transfer can differ widely between different species. The focus is slowly shifting from laboratory isolates to naturally occurring communities. The predominant amount of information that has been gathered about HGT is from laboratory isolates, with a skew towards model organisms such as *Escherichia coli*, *Streptococcus*, and *Pseudomonas*[21]. Recently, the number of studies exploring natural communities composed mostly of non-model organisms has been increasing, exploring dynamics of microbes of importance to healthcare and agriculture. For example, HGT has been characterized among communities of human gut bacteria across different hosts [22], the loss of microbial diversity in the gut has been linked to susceptibility to diseases[23], and the changes to the lung microbiome have been also linked to respiratory tract infections and a reservoir for antimicrobial resistance gene exchange[24]. Moreover, it has been suggested that anti-cancer drugs can influence the emergence and dissemination of resistance to antibiotics by inducing SOS responses and increasing HGT [25] and designing and maintaining diverse bacterial communities in waste treatment and fuel production has been crucial for production efficiency and product yield [26, 27, 28]. These and many more studies show the importance of focusing on how microbial communities interact and how recombination influences the fitness and evolution of the members of the communities. Undeniably, it will provide a more complete understanding on how HGT affects genome composition.

It has been investigated that HGT can have a positive effect on a community from enabling the coexistence of more competitors by allowing more genetic diversity and dynamic growth rates [29]. The phenomenon has been researched through modeling and simulations as well as *in vitro* and *in vivo* experiments. Currently, the wealth of genomic data available due to increased efforts and lowered costs of sequencing allows us to investigate the dynamic nature of genome diversification in naturally occurring bacterial communities.

1.2 Methods for quantitative analysis of HGT

There are many different types of information that one might want to extract by studying HGT in a community. For example, one can either i) acquire evidence that the exchange has happened, ii) identify the significantly different piece of DNA from the host genome (mosaic sequence), iii) find at what positions in the genome the breakpoints happened, or iv) calculate the rate at which the recombination happened [30]. A HGT study can happen on multiple scales: *in vitro*, *in vivo*, and *in situ* depending on the question one asks and the amount of detail one desires to describe [31]. In an *in vitro* setting, the environment and the organism of interest are very strictly controlled. *In vitro* experiments are ideal for studying details of specific mechanisms and rates of transfer of specific genes. Methods utilized in this type of study include but are not limited to single-cell fluorescence microscopy (which can track active genes of interest [32]), qPCR (a method that allows detection and quantification of the changes in gene expression [33]) , and flow cytometry (a technique for detection and enumeration of individual cells by utilizing their signature light scattering [34]). The advantages of using *in vitro* experiments is our ability to precisely control the experimental variables and therefore be able to answer precise mechanistic questions regarding specific proteins and structures.

Of course, the simplicity of the environmental conditions provided in the laboratory can also be a disadvantage since it struggles to capture possible population structures and long evolutionary timescales that could influence the mechanisms of HGT. Additionally, the choice of the organism that can be studied is limited to model organisms that can be cultured in the laboratory.

The studies performed *in vivo* strive to capture more of the nuances surrounding the HGT mechanisms by studying well-characterized organisms in their natural environments such as inside host animals or biofilms. For example, a study describing the transfer of specific plasmids in bacterial phylum *Enterobacteriaceae* found inside the mouse intestinal tract has discovered new insights of the mobility of those plasmids and highlighted particular plasmid (TP114) as a highly efficient DNA delivery system [35]. In other *in vivo* studies, the experimental design includes multiple bacterial species in a mixed population which makes tracking specific interactions very challenging. Because of that, those studies usually utilize engineered, easily tractable plasmids in order to capture the interactions. For examples, Jacquiod et al studied the spread of antimicrobial resistance genes in wastewater treatment plants using a broad-host range IncP-1 conjugative plasmid pKJK5 and described which phyla became recipient of this plasmid [36].

In order to observe the direct relevance of HGT to natural systems, one can perform analysis on the *in situ* scale. These kinds of studies gather genomic sequences from environmental samples directly in order to capture the whole complexity of the system. The advantages of this approach include the ability to capture HGT dynamics over large evolutionary timescales and to understand the extent of possible interactions between all species in the population [31]. However, it remains challenging to quantify specific biophysical processes involved in HGT on the community level, therefore, these studies usually are heavily based on bioinformatic approaches such as phylogenetics. Oftentimes, metagenomic sequencing is utilized to capture the gene content of microbial communities without having specific genes or plasmids in mind for quantification. For example, Munck et al used metagenomic sequencing in order to study the dissemination of antibiotic resistance genes between different wastewater treatment plants by screening for 15 different antibiotics [37].

A plethora of bioinformatics pipelines and methods has been created in order to identify and describe HGT events. Although it is impossible to list and describe them all, we will review some of the most popular ones below. First, building on the idea that strict clonality in the population would result in a single phylogenetic tree, some methods base their identification of horizontally transferred DNA fragments on building multiple phylogenetic trees (one for each DNA segment or gene) and comparing the structure to the whole genome tree. If the recombination is present then some of those trees will significantly differ from the main tree [30]. This method is mostly used for identifying recent transfer events in a small dataset with moderate divergence since any ancient events will not give enough signal for the trees to significantly differ [38].

Another method, which also stems from the idea of reconstructing the phylogeny and comparing trees, describes an analysis where the whole-genome dataset is filtered for conservative loci within housekeeping genes and those loci are then used for creating a phylogenetic tree since housekeeping genes are considered to rarely being involved in HGT and, therefore, most accurately depict vertical evolution of the population [39, 40]. A popular program that uses this approach is called ClonalFrameML. It creates a clonal tree for the population and then calculates the probability of a DNA site being recombined by utilizing Bayesian maximum likelihood calculations [41]. Although ClonalFrameML is a well established program in the bacterial genetics field, one of its biggest disadvantages is the lack of statistical power when the sequences one wants to analyze come from a pool of highly similar strains [42].

Linkage disequilibrium (LD) which describes a nonrandom association of alleles on the chromosome is usually used in the context of evolutionary biology and human genetics where chromosomes undergo meiosis and homologous chromosomes randomly swap their pieces with each other. When the alleles one studies are far apart from each other they are frequently crossed-over and show up in equal frequencies across gametes (linkage equilibrium). However, when alleles are close together they tend to stay together during the swapping, therefore having unequal frequencies in the gametes (linkage disequilibrium) [43, 44]. It is expected that when recombination is frequent in a population, the linkage between two alleles will decrease with the increase of the distance between them. Although LD analysis was primarily intended for analyzing human genetics, it is also performed on bacterial genomes in order to acquire evidence of recombination in the population [45, 46]. A popular measurement used for this analysis is r^2, described by Lewontin et.al [47].

Another approach for calculating recombination rate is based on the principle of population genetics. In this approach one calculates the ratio p/θ that represents the average recombination rate per mutation. Both of those variables are calculated by using the effective population size (N_e), per-site recombination rate per generation (r) and per-site mutation rate. They can be then used in the following equations: $p = 4N_e * r$ and $\theta = 4N_e * \mu$. Mcorr is an example of a program that utilizes this information and is useful for calculating recombination rates in metagenomic studies [48]. In short, the calculations in this program are split into two parts. First, pairwise 'correlation profiles' are created using synonymous substitutions. Those profiles contain information of conditional probability of observing a difference at a locus $i + l$ given that locus i is also different. This is repeated for all nucleotide distances l. Next,the program calculates a non-linear fit to $P(l)$ as a function of distance and, based on the fitted parameters, calculates mutation and recombination rates. If no recombination is present, the function will be a flat line, however, if there is recombination in the population, there will be a monotonic decrease in the $P(l)$. This way, Mcorr provides evidence of recombination in a population as well as the actual rate of transfer.

1.3 Biofilm and Pink Berries

In a natural setting, such as oceans, lakes, and marshes, bacteria live in communities, creating intricate interactions and structures, such as biofilm, that help them survive. Because of the close proximity to one another, these structures create an ideal environment for HGT [20]. Biofilms have been extensively studied, especially in the context of biofilm-mediated antibiotic resistance dissemination, but these studies have mostly utilized artificial biofilm setups [49]. An example is a 1996 study by Gold and Moellering who created an artificial biofilm system consisting of two bacterial species where they showed that a tetracycline resistance gene was transferred from one to the other [50]. Nowadays, many metagenomic studies use natural biofilms in order to study the genetic diversity of the community and the extent of their HGT by, for example, measuring the rates of gene transfer across communities. For example, a metagenomic study conducted using microbial mats from Shark Bay have found evidence of transfer of many genes involved in high-UV irradiation protection, response to changing salinity conditions, as well as adaptation to oxidative stress and heavy metals [51]. Another study used metagenomic assemblies of macroalgae and surrounding seawater microbial communities and showed that HGT was happening mostly between members of the same species and involved genes for nutrient transport and stress responses [52]. Moreover, not only can one study genes responsible for protection against outside influence or metabolism but metagenomic studies of biofilms and microbial mats can provide invaluable insights into evolutionary processes within natural communities [53].

Here, the focus is set on an interesting macroscopic bacterial aggregates called Pink

Berries. They are naturally occurring communities collected from microbial mats at the Little and Great Sippewissett Salt Marshes in Cape Cod, Massachiussetts. The bacterial communities of the berries are very diverse but they are dominated by two particular species that, due to their close physical proximity created by the agglomerates, participate in a sulfur cycle [54]. The bacteria that oxidizes sulfide to sulfate is a bacteria from a genus *Thiohalocapsa* (Sulfide-oxidizing bacteria, henceforth known as PSB) while the bacteria that reduces sulfate to sulfide is from a family *Desulfocapsaceae* (Sulfate-reducing bacteria (SRB)).

1.4 Self / Nonself Recognition in Pink Berries

Organisms can interact with each other in many different ways that can be beneficial to one or both parties. However, in the light of limited resources and space, bacteria developed many different tactics to out-compete one another. A classic example of a competition can be found in aerobic soil bacteria communities that need to eliminate other organisms in order to get more access to sunlight and oxygen. This strive for resources, and ultimately survival of the organism, demands a set of novel strategies that would impede the growth of other organisms. One of such fascinating strategies is called self / non-self recognition [55]. This strategy helps bacteria distinguish between its own species (self) and halt the growth of organisms that are not them by creating a set of toxin-antitoxin genes. A toxin (a small protein that negatively influences a cell) is produced by the organism along an immunity protein that immediately stops the toxin from harming the organism that produced it. However, if a toxin is inserted into a cell that does not have the immunity gene (therefore is not considered ”self”) it will be destroyed.

Contact dependent growth inhibition (CDI), first demonstrated by Aoki et al. [56], is a system that delivers a toxin from one bacteria to another upon contact and, if the recipient does not posses an immunity gene to neutralize it, the toxin inhibits its growth. The CDI machinery consists of multiple parts which number and complexity depends on the type of the system (discussed below). However, the mechanism can be summarized as containing three crucial elements: the delivery mechanism, the toxin, and the antitoxin [57, 58]. Interestingly, while the mechanism stays the same depending on the type of the CDI, the toxin-antitoxin pairs are much more variable, which helps the organism in creating new weapons against competition [59]. As of 2024, secretion system types III, IV, V, VI, and VII have been shown to use direct contact for toxin delivery while hundreds of toxin-antitoxin pairs have been identified.

Type III secretion system has been described to have a needle- or syringe-like structure that can be divided into three main parts: a base, a needle, and a translocon [60]. When the system senses a contact with a host cell, it will pierce the cell membrane and pass the toxin [61]. In type IV secretion system, somewhat related to bacterial conjugation systems, the proteins create a channel through which toxins travel [62, 63, 64]. It has an ability to pass the substrate directly into the host's cytoplasm. Type V distinguishes itself by having a minimalist number of components and being self-sufficient - they do not rely on the help of a dedicated secretion apparatus [65, 66]. Type VI system contains a sheathed needle that spans the inner and outer membrane of the cell [67, 68]. It has been studied in clinical samples of *Vibrio parahaemolyticus* in the context of describing fitness landscape and recombination patterns in naturally occurring populations. *V. parahaemolyticus* is a free living marine bacterium that is responsible for global spread of human gastroenteritis outbreaks [69].

Because of their mechanisms, the CDI systems have been explored in the context of human health. Studies have been conducted that show that the CDI mechanisms are not only used for toxin delivery but also for exchanging plasmids and, therefore, contributing to the spread of the antimicrobial resistance among harmful pathogens. Therefore, efforts have been put into engineering inhibitors for the machineries in order to decrease potential spread of new antimicrobial resistance genes [70, 71, 72, 73, 74].

1.5 Species identification in a bacterial community using 16S rRNA

The identity of a bacteria can be determined through various approaches such as its phenotype or full genome sequencing. However, the most popular method for identification nowadays that is fast and reliable is the sequencing of the 16S ribosomal RNA (rRNA) gene [75]. This gene is a part of the 30S subunit of the ribosome - an essential machinery that allows the cell to translate messenger RNA (mRNA) into a sequence of amino acids that will become a functional protein. The 16S rRNA gene is particularly useful for bacteria identification because of its variable regions. The gene is usually about 1,550 base pairs long in all bacteria, but there are nine regions in the sequence that vary between different species. Those regions tell us what species the sequence came from [76, 75, 75]. Its usefulness has been proven in not only scientific research but also industrial and clinical settings [77, 78, 79, 80, 81, 82].

The sequencing technologies has been developing rapidly throughout the years. Starting with Sanger sequencing, moving on to Illumina short-read sequencing, and now transitioning to long-read sequencing using Nanopore and PacBio platforms. The Illumina sequencing is still a major sequencing platform due to its amazing accuracy, however, one of its biggest drawbacks is the limited length of the sequenced reads (less than 500 bp). Therefore, in order to determine the sequence of a gene or a genome, one needs to assemble it back from the short pieces by overlapping them until there are no more gaps. However, such a task is not trivial and requires fast and accurate alignment algorithms which can struggle with accurate results especially in genomic regions of high repeatability. The appeal of the new sequencing technology, the long-read sequencing, is the theoretically unlimited length of the consecutive piece of DNA. This means that one can create a novel, fully sequenced genome without the need of using alignment algorithms and create more accurate representation of the regions containing tandem repeats. Currently, there are two ways long-read sequencing is performed. Pacific Biosciences (PacBio) utilizes data gathered from light emitted by a DNA polymerize. Oxford Nanopore Technologies (ONT), on the other hand, gather data of electric current changes. When a string of DNA passes through a special port, the different bases emit a slightly different current [83, 84]. However, because this technology is still new, the accuracy of the reads is much lower than Illumina's (10-15% error rate compared to 0.1%) [85]. Nevertheless, the unrestricted read length promised by the PacBio and Nanopore technologies is a promising alternative for sequencing 16S rRNA genes.

Chapter 2

Quantifying the extent of horizontal gene transfer in the genomes of Pink Berries

2.1 Introduction

"On the origin of species" by Charles Darwin contains only one illustration. It is a diagram depicting how common features between species can be explained by a vertical descent from a common ancestor - now widely known as a phylogenetic tree. In eukaryotes, a phylogenetic tree can be successfully applied to demonstrate how a genome changes and accumulates mutations - it depicts the chronological descent of relationships driven by evolution. The bifurcating tree diagram has been favored for centuries since it clearly shows how diverging species gradually split into separate lineages (known as cladogenesis) or how new species form without branching (known as anagenesis). The diagram works splendidly in depicting changes brought by both sexual and asexual reproduction of eukaryotes. However, when one tries to apply the same to bacteria, one will find it clear that because of an extensive horizontal gene transfer (HGT) between species the tree will change based on which genome fragments one looks at. HGT refers to sharing of small segments of genetic material from a donor to a recipient organism that does not have a parent-offspring relationship. Although not all transfers are successful, they are still abundant in natural bacterial communities. That fact is quite troublesome especially for population geneticists that strive to unravel evolutionary histories of bacteria. Rather than representing the genetic relations between them using a bifurcating tree [8], one needs to take into account the fact that bacteria can incorporate a new piece of DNA from an outside source, which ultimately alters the genealogical relationship between strains. An organism that horizontally acquired a new gene can no longer be simply placed on a phylogenetic tree since its genetic material is no longer considered clonal[18]. HGT ultimately leads to a network-like history of an organism rather than a tree-like scheme[19, 8].

Why do bacteria transfer genetic information between each other? The ever-changing environment forces organisms to quickly adapt to avoid extinction in the ongoing process of natural selection. Novel pieces of DNA, gained mostly by HGT, can provide them with necessary tools and new means of adaptation for surviving in new ecological niches[8]. However, one does not simply integrate a foreign piece of DNA into one's genome. Many different ways of genome exchange exist - mainly transduction, transformation, and conjugation, which refer to transfer via phage, uptake from the environment, and via direct contact respectively. A successful uptake and integration of a DNA piece depends on many different factors such as proximity, metabolic compatibility, adaptations to their abiotic environment, gene expression systems, gene-transfer mechanisms, among many others[9, 10, 11]. Even the matter of compatible GC content of the recipient and the donor genetic material or gene copy number can influence the success of integration and spread [13, 16, 17]. Given the overwhelming number of factors acting against HGT, the successful uptake is no mere accident and, in many bacterial species, HGT is a normal physiological process. For example, *Bacillus subtilis* has been used as a model organism for studying transduction due to less strict DNA sequence specificity [86]. Additionally, *Streptococcus pneumonia* has also been described as having a natural competence for uptake of foreign DNA[87]. The extent of the genes affected by HGT is vast, including all functional categories of genes and even rRNA operons and genes that have been known to define characteristics of a phylum[10, 88]. Moreover, the rate at which a gene is transferred differs widely across different genes and organisms[89, 8]. This makes it tricky, especially for healthcare professionals that study antimicrobial resistance and epidemics, to predict how the organism will evolve over time and over environmental fluctuations. Recombination creates more opportunities for a local adaptation by making it possible to acquire genes involved in antibiotic resistance, pathogenic determinants, etc [22, 48]. Therefore, it is important to understand how frequently genes are transferred and how they change the evolution of microbial populations, both synthetic and naturally occurring[90, 91, 92].

Recently, the number of studies exploring natural communities composed mostly of non-model organisms has been increasing, studying dynamics of microbes present in many different fields such as healthcare and agriculture, especially in the context of biofilms[22, 23, 24, 26, 27, 28]. Biofilms are versatile structures of matrix-enclosed agglomerations of unicellular organisms that are usually found in aquatic ecosystems[20, 93]. They are known to be hotspots for HGT since they provide a great environment for cells to live in a close proximity to each other at high density while also being protected from the harsh external conditions. Biofilms have been extensively studied, especially in the context of biofilm-mediated antibiotic resistance dissemination, but these studies have mostly utilized artificial biofilm setups[49]. Recently, there has been a shift from the lab-based cultures to the naturally occurring ones. Research in this field is especially important since microbes found in biofilms are experts at adaptive responses to various environmental changes. Many metagenomic studies have used natural biofilms in order to measure the rates of gene transfer across communities [50, 51, 52, 53]. Moreover, not only can one study genes responsible for protection against outside influence or metabolism but metagenomic studies of biofilms and microbial mats can provide invaluable insights into evolutionary processes within natural communities.

Pink berries, which are macroscopic, photosynthetic bacterial aggregates, create extensive microbial mats at the Little and Great Sippewissett Salt Marshes on Cape Cod, Massachusetts. They can form large aggregates, almost up to a centimeter in diameter[54]. Each berry contains hundreds of different bacterial species that together create complex metabolic interactions due to the proximity of the microbes that enhances cell-cell contact and allows for extensive genetic exchange. Although the berries can create a very diverse ecosystem, there are two species that dominate their space. The intense pink color of the berries come from a genus *Thiohalocapsa* (Sulfide-oxidizing bacteria, aka Purple Sulfur Bacteria (PSB)) which strongly depend on another widespread species of proteobacteria from the family *Desulfocapsaceae* (Sulfate Reducing Bacteria (SRB)). The pink berries are sulfur-cycling symbiotic consortia in which PSB and SRB physically accumulate along with a variety of other bacterial species. That close proximity allows PSB and SRB to engage in interspecies electron transfer and together create a sulfur cycle - PSB oxidizes S^{-2} to SO_4^{-2} (oxidizes sulfide to sulfate and stores elemental sulfur inside its cells) and SRB performs the reverse of that action which closes the cycle (reduces sulfate to sulfide).

In this paper a comprehensive analysis of the Pink Berry metagenomic data in context of extensive recombination of a quasi-sexual bacterial population is shown. The evidence that the populations are divided into clades with different evolutionary histories including a mixing layer where bacteria experience extensive recombination from other clades as well as with each other is presented.

2.2 Results

2.2.1 Collection and metagenomic sequencing of Pink Berries

Pink Berries agglomerates were collected from microbial mats at the Penzance Point, as well as Little and Great Sippewissett Salt Marshes in Cape Cod, Massachiussetts (Fig. 2.1a).The sequencing and assembly was performed by the Cordero and Wilbanks labs. Each berry was sequenced using metagenomic shotgun sequencing (Illumina's HiSeq Rapid 2x300 technology). Instead of sequencing from genetic material of singular cells of each species, the genomes from the full community were sequenced. Therefore, each "strain" is actually a mix of reads from multiple cells of the same species fromone berry. The co-assembly of the prepared reads into contigs was performed using MEGAHIT software [94] and contigs longer than 1 kb were chosen for binning (a total of 245,777 contigs longer than 1 kbp, with 4,911 contigs longer than 10 kbp). The sequenced reads were then mapped back to the constructed contigs using minimap2 [95]. In order to create the metagenomically-assembled genomes (MAGs) and reduce a chance for any potential binning bias, multiple different tools were used to bin the contigs (CONCOCT 1.0.0 [96], MaxBin 2.2.5[97], MetaBAT 2.12.1[98], and Vamb 1.0.1[99]). Lastly, the quality of the MAGs was assessed using CheckM software [100] and each MAG was taxonomically classified using GTDB toolkit [101].

Two most abundant species present in the aggregates were *Thiohalocapsa* (sulfide oxidizing purple sulfur bacteria (PSB)) and *Desulfofustis* (sulfate reducing bacteria (SRB)) (Fig. 2.1d). Since the *Thiohalocapsa* MAG was very similar to the PB-PSB1 reference (assembled with long read PacBio data by Wilbanks lab (NCBI BioProject PRJNA215075)), the reference was used as a baseline instead of the contigs created by the MAG.

The PSB genome consists of one chromosome of length 7,950,631 base pairs. Out of 192 strains, strains with less than 80% of their biallelic bases called have been discarded, leaving 142 strains for further analysis. Only proper biallelic single nucleotide polymorphisms have been kept which we have assumed correspond to single mutational events in the history of its genomic locus. A total of 50,626 biallelic Single Nucleotide Polymorphisms (SNPs) have been selected. The SNP data has been labeled one of three different categories: wild type, mutant, or 'NaN' using majority calling. 'NaN' was given to any locus of a particular strain that did not have sufficient coverage to be informative. The majority of the biallelic loci have a mean depth of coverage of 25-35 reads (Fig. 2.2a). The depth of coverage for biallelic loci in individual berries varies. Some berries have a low mean depth of coverage of 2-3 reads while others have a much higher mean coverage of about 50-60 reads(Fig. 2.2c). Each biallelic site can have a different number of strains that contain a mutation at that particular site. A locus can be a singleton (only one strain has a mutation while the rest are either wild type or 'NaN'), doubleton (two strains have a mutation while the rest do not), etc. The distribution of the number of strains with a mutation at a particular biallelic loci is shown in Fig. 2.2b). The majority of the loci have a very small number of strains with mutations per site.

For the SRB dataset, the data was processed the same way as the PSB. One difference between the PSB and SRB genomes is that the SRB genome consists of 97 contigs with a total length of 4,710,529 base pairs. SRB dataset contains of more biallelic SNPs than the PSB dataset and the majority of the SNPs are located on five contigs (unitig_0, 67, 69, 70, and 87) which have more than 5,000 SNPs each. Overall, the SRB SNP matrix consists of 59,435 SNPs and 116 viable strains. The majority of the biallelic loci have a much lower mean depth of coverage compared to the PSB dataset (5-10 reads per site) (Fig. 2.3a). The depth of coverage for biallelic loci in individual berries also varies, with the majority of the strains having a mean depth of coverage of less than 10 reads per site. (Fig. 2.3c). As seen in the PSB dataset, the distribution of the number of strains with a mutation at a particular biallelic loci shows that the majority of the loci have a very small number of strains with mutations per site (Fig. 2.3b).

2.2.2 Creation of bifurcating trees based on biallelic alleles

Although it has been established that phylogenetic trees do not faithfully describe the evolutionary history of a bacterial population because of the high frequency of horizontal transfer of genomic fragments, a phylogeny reconstructed from the biallelic SNPs from our dataset can still be a source of information about the structure of the dataset and relationships between strains. If no recombination is present in the population (i.e. all mutations are vertically inherited), there will exist only one tree that will describe the evolutionary history of the strains. However, when recombination is present and a common occurrence, the structure of the tree will vary based on the regions of the chromosome used to build it. Keeping this in mind, a tree of the PSB strains was created (excluding singletons and strains with large amounts of missing data) and the resulting neighbor-joined tree revealed multiple clades that are clearly separated on the tree. While some of those structures are easily identifiable by eye, for clarity, we also plotted the relationship between strains as a heatmap of correlations. The rows and columns are ordered the same way as the strains on the tree and the scores represent the percentages of the shared SNPs between each pair of strains. The heatmap was annotated to show each tree clade. It is easier to see on the heatmap that the strains group themselves into multiple distinct clusters that are much more similar to each other than the background. Those clades include a clade of closely related strains exclusively from location E (which could indicate a possible recent population bottleneck), a highly diverged clade (top of the tree and top left of the heatmap) containing 9 bacterial strains (henceforth referred to as the “9-clade”), and a group of strains from the same geographical location F exhibiting

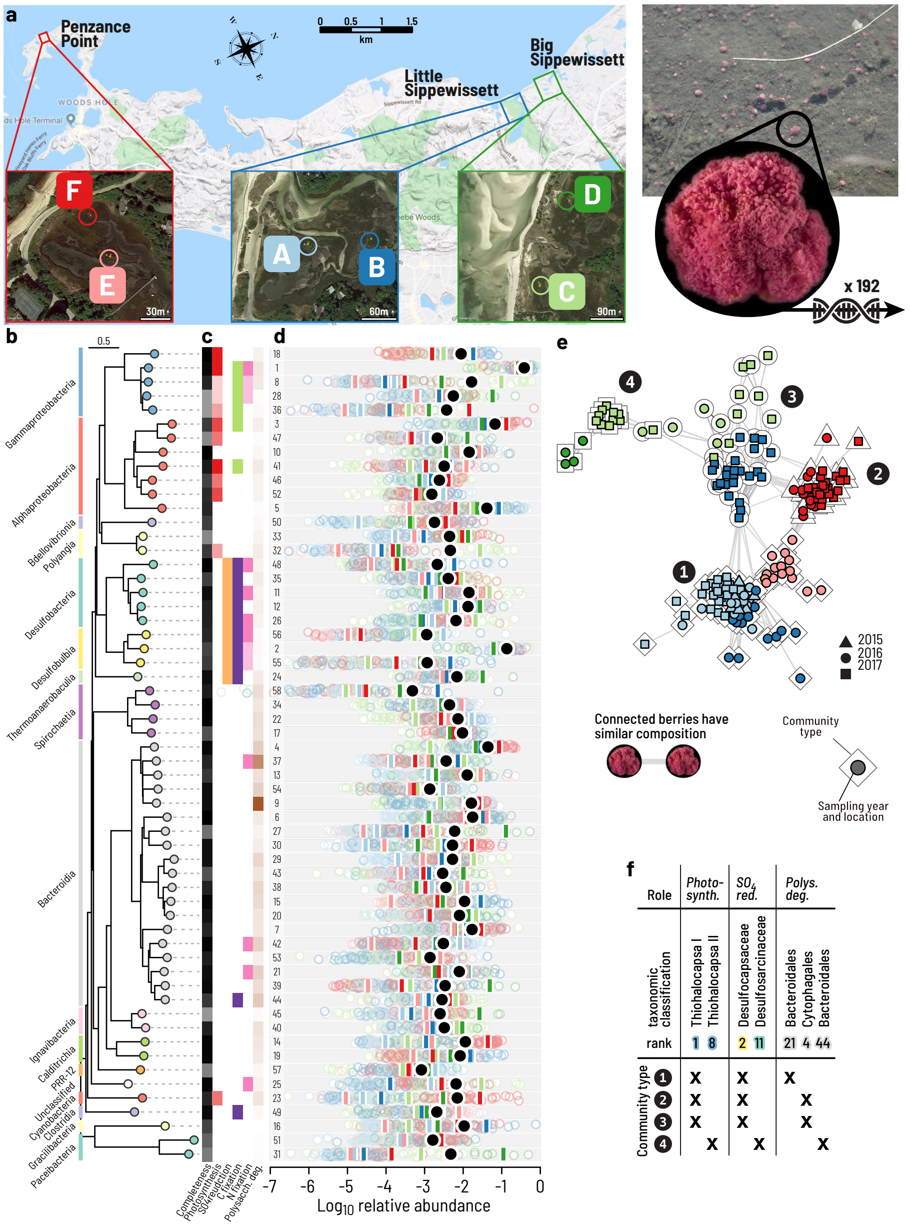

Figure 2.1: Pink Berry dataset description by Gabriel E. Leventhal, Jakob Russel, Shaul Pollak, Rachel Szabo, Tim N. Enke, Thomas Hackl, Boris Shraiman, Elizabeth Wilbanks, and Otto X. Cordero (unpublished). a) Geographical location of the collection sites and the image of the berries. b) Phylogenetic tree of different species found in each berry. c) Functional roles of each species in the berry. d) (colored circles) Relative abundance of a MAG in an individual berry. (black circles) Mean relative abundance across all berries. e) A network created based on the relative abundance of MAGs in each berry. f) Community types.

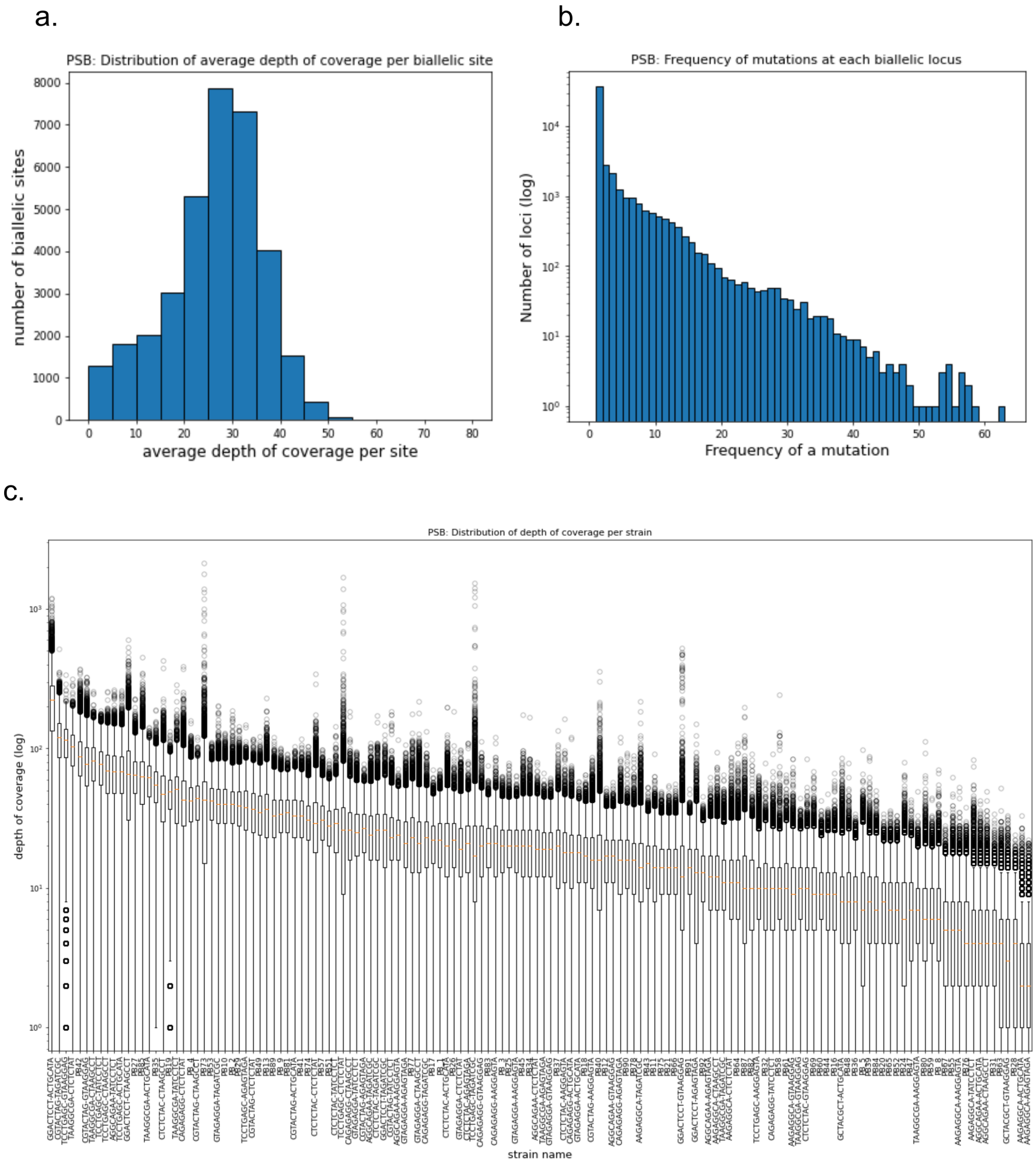

Figure 2.2: PSB dataset description. a) A distribution of average depths of coverage per a biallelic site. b) A distribution of number of strains with a mutation at a particular biallelic site. Most of the loci are singletons. c) A box plot of depth of coverage of the biallelic loci separate for each strain.

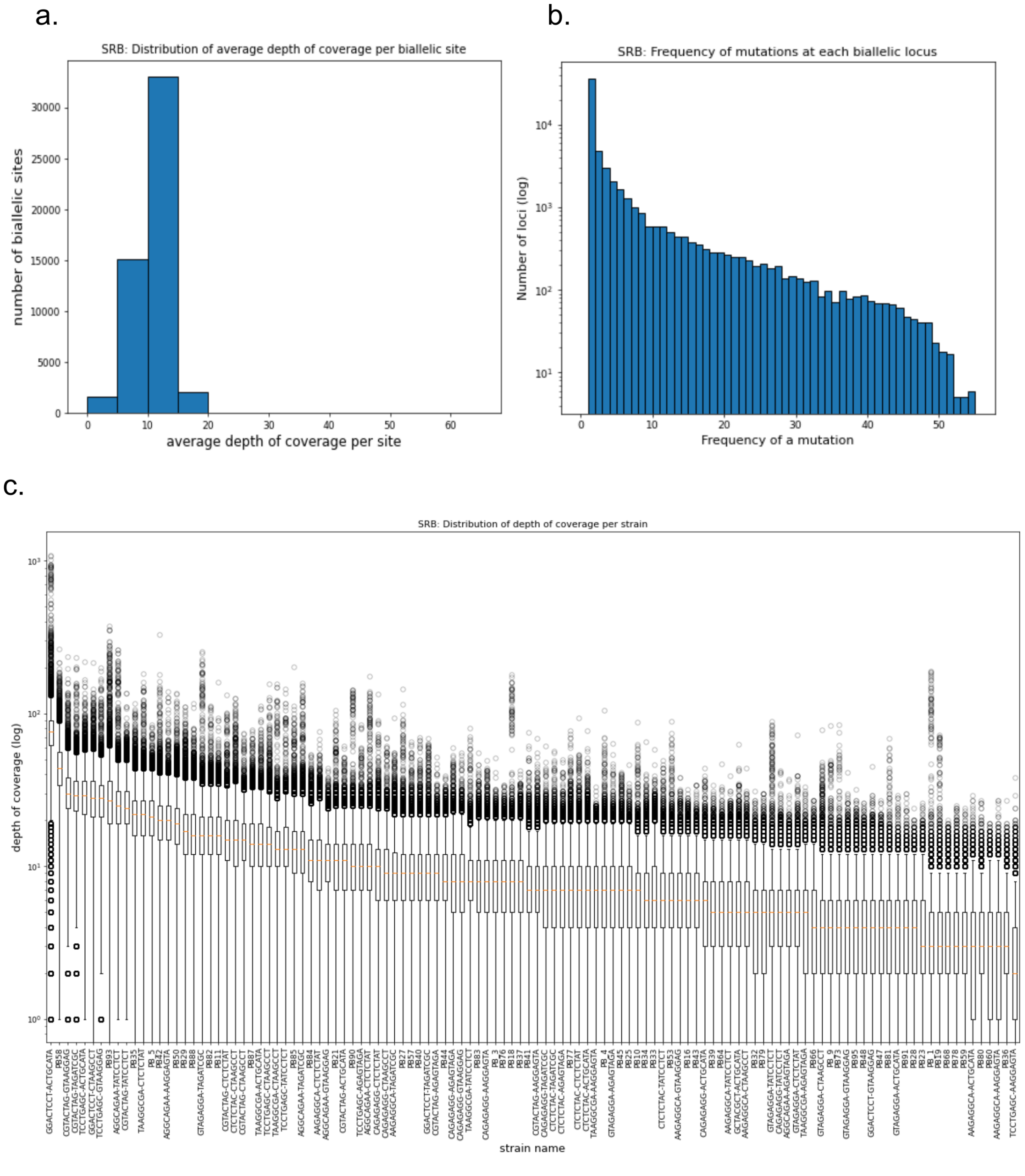

Figure 2.3: SRB dataset description. a) A distribution of average depths of coverage per a biallelic site. b) A distribution of number of strains with a mutation at a particular biallelic site. Most of the loci are singletons. c) A boxplot of depth of coverage of the biallelic loci separate for each strain.

a star-like structure (Fig 2.4). Of note is the observation that those strains create a larger cluster on the heatmap with similarities to the 9-clade and the strains preceding the ladder.

In line with expectations, after we created a tree of the SRB strains based on the whole SNP dataset using similar methods, it also revealed multiple interesting structures (Fig 2.5). The heatmap of the SRB strains also shows clusters sharing similarities between strains of the same clade. There are four highly similar clades and a region in the middle that exhibits similarities to both the top left, highly diverged cluster (the F clade) and middle cluster, consisting of strains from locations A and B (topAB). The second big cluster worth noting contains strains from the bottom right of the heatmap (basalAB and basal clades). For the rest of this study, we heavily focus on those structures.

2.2.3 SNP distribution across genome and phylogenetic trees

To begin our investigation into the impact of recombination, we performed a straightforward analysis of the overall biallelic SNP distribution across the PSB and SRB's genomes. As seen in Sakoparnig et al [102], when species have very low nucleotide divergence between each other, the effect of recombination can be nicely visualized by the SNP density across the chromosome. In order to see if we also see nonuniform distribution of SNPs along the genome, the SNP frequency was calculated per 1 kb window and plotted as a linear plot, one Mb per subplot. As anticipated, the resulting distribution was not evenly distributed across the chromosome (Fig 2.6A and B) and the distance between the consecutive SNPs varied from 0 base pairs (SNPs that are right next to each other) to over 5,000 base pairs. While we observe that the density of SNPs is usually low, there are few regions where the SNPs create a high SNP density islands. It is expected for bacteria that experience HGT and recombination of their genomes to have those kinds

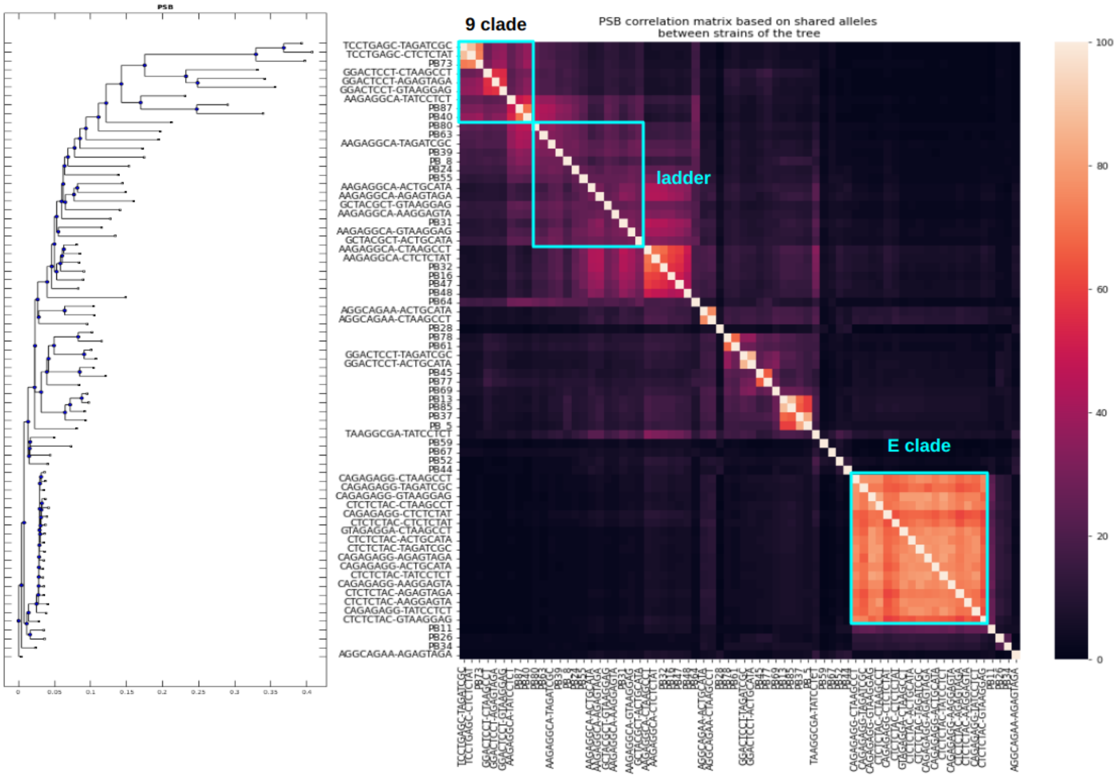

Figure 2.4: PSB tree (left) and correlation matrix (right) based on the collection of all non-singleton SNPs and probabilistic distance calculations, showing multiple separate clades.

of hotspots. A number of recent studies in different species of bacteria have identified this kind of behavior and argued that regions of high density almost certainly result from HGT events [102, 103]. The histogram of SNP densities per kilobase (Fig. 2.4C) shows that the distribution of SNP number per kilobase has an exponential tail. The most common number of SNPs is around 2 to 4 SNPs per kilobase while there are only a few regions where the number of SNPs per kb exceeds 40.

The density graph of the SRB dataset looks more stochastic. When we consider a full set of biallelic SNPs, the high SNP density regions are not as easily distinguishable as it was the case in PSB. This trend persists across all contigs. Fig 2.6D shows that the overall shape of the SRB SNP density distribution is similar to PSB, however, its mean

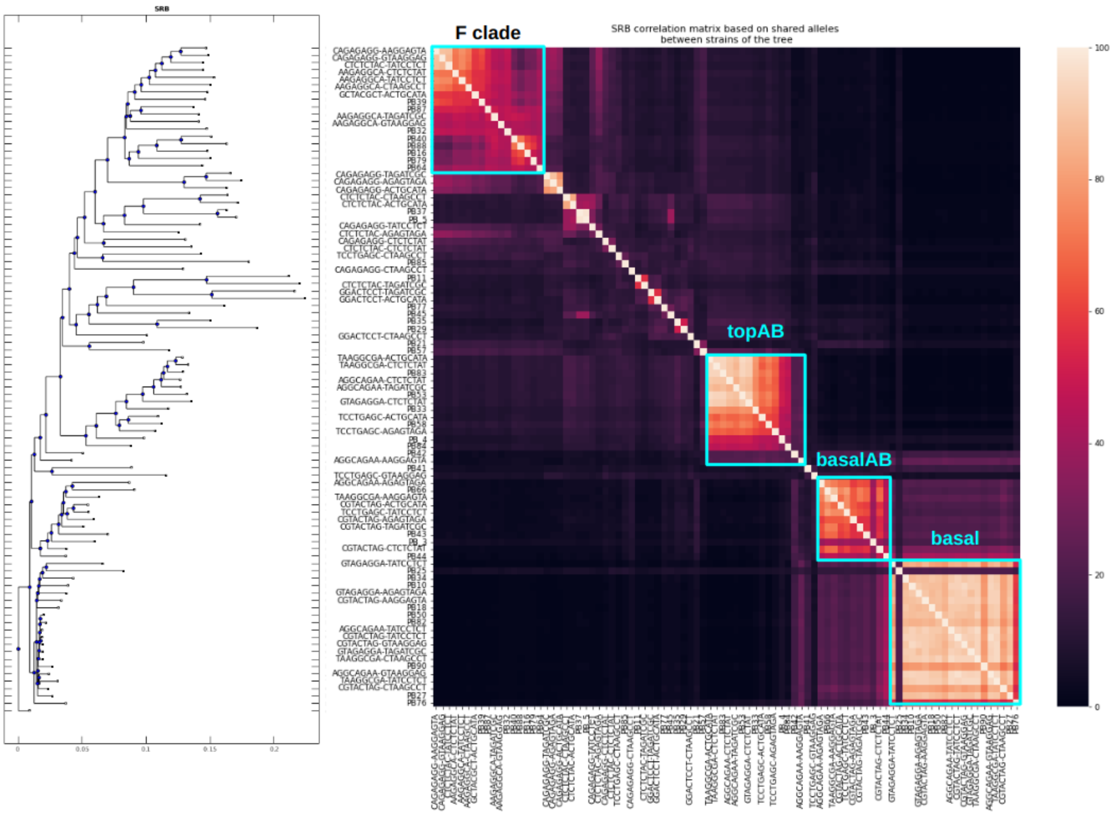

Figure 2.5: SRB tree (left) and correlation matrix (right) based on the collection of all non-singleton SNPs and probabilistic distance calculations, showing multiple separate clades.

and variance are significantly higher. We observe two distinct slopes in this distribution - an exponential on the high density side and a fast decaying “tail” for low density regions (high frequency SNPs).

Interestingly, when one focuses on bacteria that are very closely related to each other (i.e. create a separate clade), the SNP density graphs for strains of one clade shows high SNP density islands much clearer compared to the density graphs using the full dataset (Fig. 2.7).

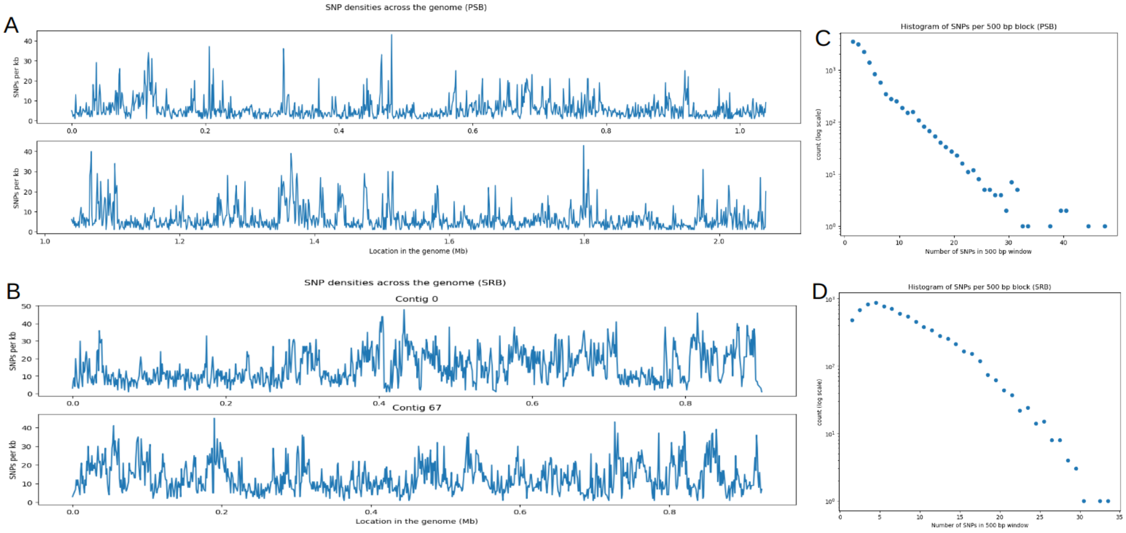

Figure 2.6: A) A sample window of the SNP density graph for PSB first 2 Mb with a window of 1 kb.B) A sample window of the SNP density graph for SRB first 1 Mb of contig 0 and 67.C) Histogram of SNPs per 1 kb block for PSB. Note that the y axis is in log scale. D) Histogram of SNPs per 1 kb block for PSB. Note that the y axis is in log scale.

2.2.4 Non-synonymous to synonymous ratio analysis

Knowing the location of possible transfer hotspots alone is not enough to let us know what is happening in each population. We can shine more light on how those mutations affect the population by calculating the ratio of variations that change the amino acid (non-synonymous changes) versus the ones that are silent (synonymous changes) - commonly known as d_N/d_S analysis. In this analysis, synonymous mutations are assumed to be regarded as neutral and the ratio of non-synonymous-to-synonymous changes evaluates the action of natural selection on mutations at non-synonymous sites. A ratio of $d_N/d_S < 1$ indicates purifying selection where the mutations are removed overtime while a ratio greater than 1 indicates selection for diversification. Under a neutral hypothesis of evolution, the mutations in the genome accumulate by genetic drift - changes happen due to random chance. In order to assess the possible effect of other, more systematic,

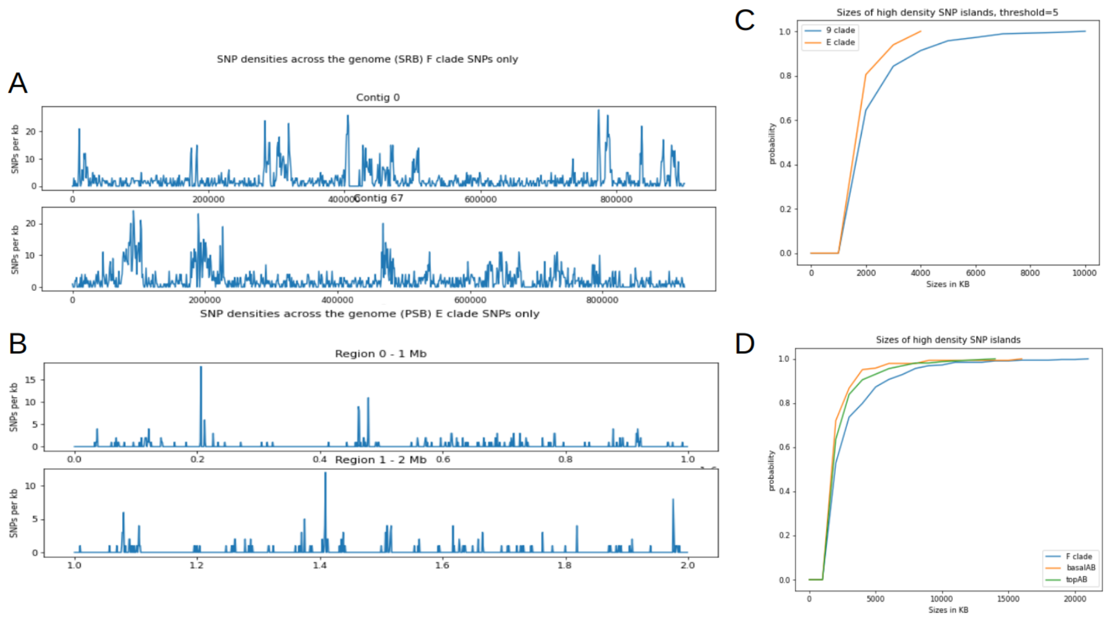

Figure 2.7: A sample window of the SNP density graph for PSB and SRB. Cumulative Distribution Functions for the lengths of the high density SNP islands given a threshold of 5 or more SNPs per 1 kb. (A) SRB, F clade SNP density graph. (B) SRB, E clade SNP density graph. (C) PSB CDF. (D) SRB CDF.

forces, such as selective pressure that favors one particular phenotype in order to increase organism's survival or reproductive success, we need to quantify the frequency of adaptations acting directly on the protein coding regions of the genome. The ratio of divergence at non-synonymous sites (d_N/d_S) was plotted as a function of d_S which serves here as a proxy for the average divergence across the genome (Fig. 2.8). By performing this analysis, we observe that the frequency of the non-synonymous mutations is lower than expected for a neutral model ($d_N/d_S = 1$). For both species we observe a consistent negative relationship between the non-synonymous ratio and the synonymous divergence which indicates that the populations are experiencing negative (purifying) selection and therefore, deleterious mutations are purged over time. Moreover, the d_N/d_S ratio rarely exceeds 1 for PSB strains and it is constantly below 1 for SRB strains for all pairs of strains, regardless of how closely they are related.

Over the years, there has been discussion in the field about how to interpret the d_N/d_S values calculated for individual genes within a closely related population [104, 105]. It is expected that d_N/d_S values will vary depending on the age of the variant since younger variants have not been exposed to selection for too long. Nevertheless, the non-synonymous-to-synonymous ratio calculated on individual genes can tell us useful information about which genes in the population are maintained over long evolutionary timescales. Since in the presence of recombination, regions that have been recently exchanged can be falsely identified as positively selected, we decided to only used genes that have a large number of mutations (a number that is larger than what one would expect for the length of a gene). Looking at a set of 20 highly mutated genes in the PSB dataset, we observe that half of them do indeed exhibit positive selection ($d_N/d_S > 1$). Those positively selected genes are involved in regulating recombination (XerD, CRISPR-associated nuclease Cas1, type-1 restriction enzyme,Group II intron-encoded protein LtrA, and recombination promoting nuclear RpnD), while others are involved in overall survival and well-being of the cell (Immunoglobulin A1 protease, Fe(3+) ions import ATP-binding protein FbpC, and glutamine amidotransferase).

2.2.5 Linkage Disequilibrium

In a simplified model, prokaryotes reproduce purely asexually. In other words, they possess the ability of inheriting all genetic material from a single parent, hence, mutations that happen closer to the root of the ancestral tree will be shared by all of the descendants of the branch. On the other hand, when recombination is present in the population, one can assess its frequency by looking at the statistics of linkage disequilibrium (LD) [106]. Linkage equilibrium describes a situation where alleles randomly and freely recombine with each other. LD, on the other hand, describes a nonrandom association between

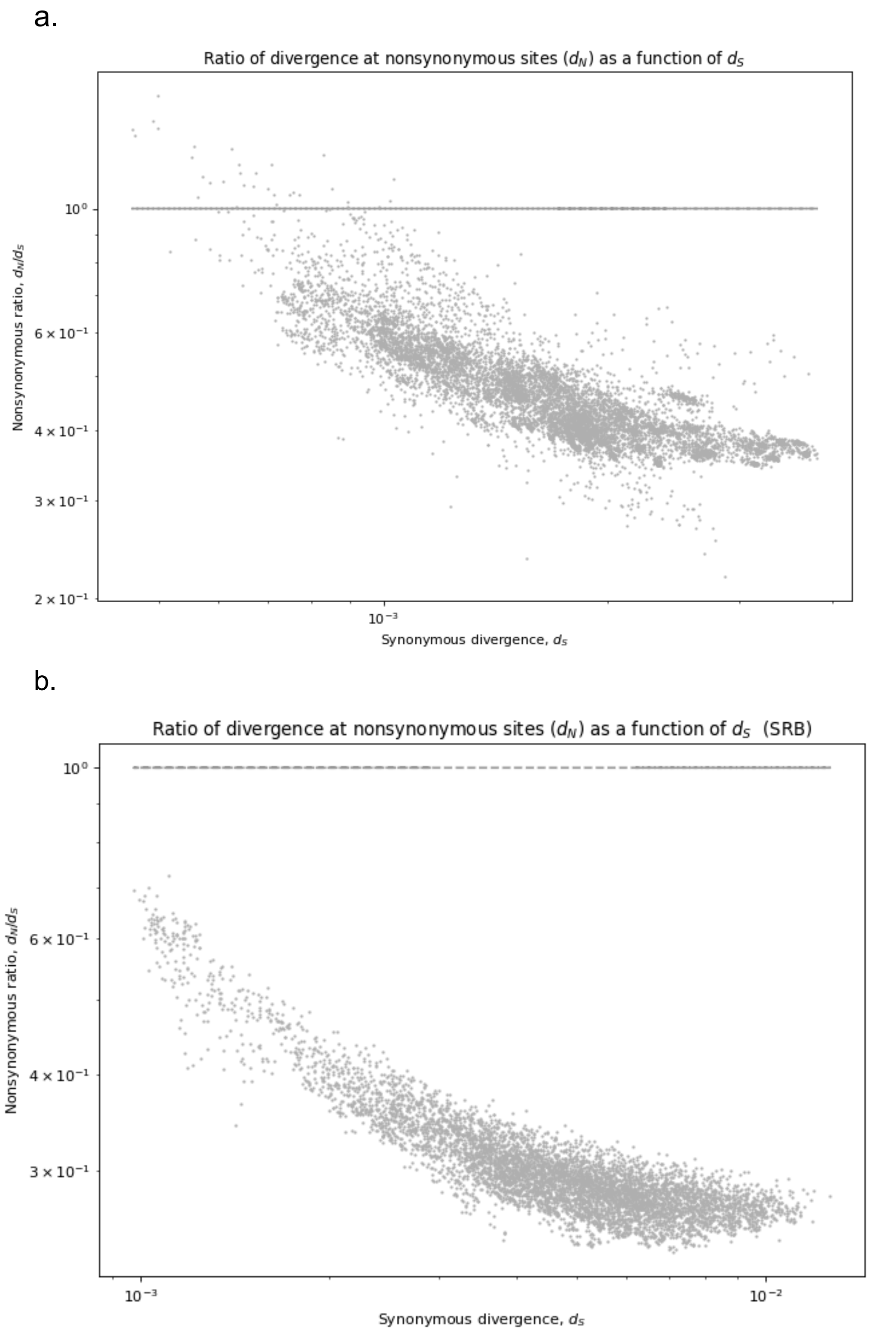

Figure 2.8: dN/dS plots for (a) PSB and (b) SRB

alleles at two loci on the chromosome. When a piece of DNA is transferred from one genome to another, the linkage between alleles of the transferred piece and what have not been transferred will be lost [43]. It is expected that when recombination is frequent in a population, the linkage between two alleles will decrease with the increase of the distance between them. To investigate the pattern of asexual inheritance and whether or not PSB and SRB follow this model, or are subjected to frequent recombination event, we used a method of average squared correlation of the pairs of genotypes at pair's position (r^2) described by Lewontin et.al.[47]. In short, Lewontin's method is a take on the comparison of consistency with a common phylogeny (commonly known as the four-gamete test in the literature on sexual species). Given that each pair consists of biallelic loci (either with 0 (wild type) or 1 (mutation)), there are only four possible genotypes: 00, 01, 10, and 11. The frequency of each genotype is then used to calculate the r^2 values using the equation listed in methods. If we assume a null hypothesis where PSB and SRB behave in a strict asexual manner, those tree-like mutations should pass the four gamete test and therefore follow an asexual pattern of inheritance[107, 43].

After visualizing the average r^2 score for distances between the members of the pair being between 1 and 1000 bp, in the PSB graph we see a sharp drop in linkage for very short distances that then becomes constant for distances above 200 bp. Therefore, we can conclude that sites that are physically close together on the chromosome are inherited together more often than sites further away from each other (also seen in [102] and [46]) (Fig 2.9). Similarly for the SRB the resulting graph shows a drop in the value of r^2 with distances over 200 bp.

This analysis shows us that segments that obey the asexual model of inheritance are relatively short (much shorter than a length of a gene) and that segments that are longer than 200-300 bp are not compatible with a tree which indicate a lot of HGT events in the history of the population. A sharp drop in LD like this can be explained by the presence of recombination in the population. Over time, correlations between allele frequencies are broken up by recombination and therefore, we observe a rapidly decaying linkage. Additionally, the short genomic distance across which alleles are linked could be due to the DNA fragments that are available for recombination being degraded via environmental factors or restriction-modification systems. Plotting both SRB and PSB distributions together shows that the distributions are comparable to each other. This is an interesting observation since it implies that the length of the pieces that are compatible with the asexual tree do not differ in those two species and might suggest that whatever is causing recombination is quite universal. Moreover, this distribution is consistent with a power-law behavior of $x^{-0.3}$, i.e. the distance distribution appears as a decreasing straight line on a log-log scale.

2.2.6 Creation of the SNP matrices and analysis

In Fig. 2.4 and 2.5, we showed that both PSB and SRB have tree structures that consist of multiple clades. Based on the heatmaps, we also know that some of those clades seem to interact more with each other while others tend to be more isolated. For example, the strains of the 9 clade and the ladder have more common alleles (i.e. their correlation scores are higher) than for the strains of the 9 clade and the E clade. In order to understand how exactly these clades interact with each other, we restructured the previously created tree as a matrix with individual strains as rows (organized in the same order as the tree) and individual loci as columns (sorted based on their genomic position). Instead of creating one matrix of 50,000 or 80,000 SNPs, we have created multiple smaller matrices that focus on individual clades in order to observe how the SNPs that are fixed in the clade in question are organized in the rest of the tree. A SNP was considered “fixed” if it was present in more than 50% of the strains in a particular

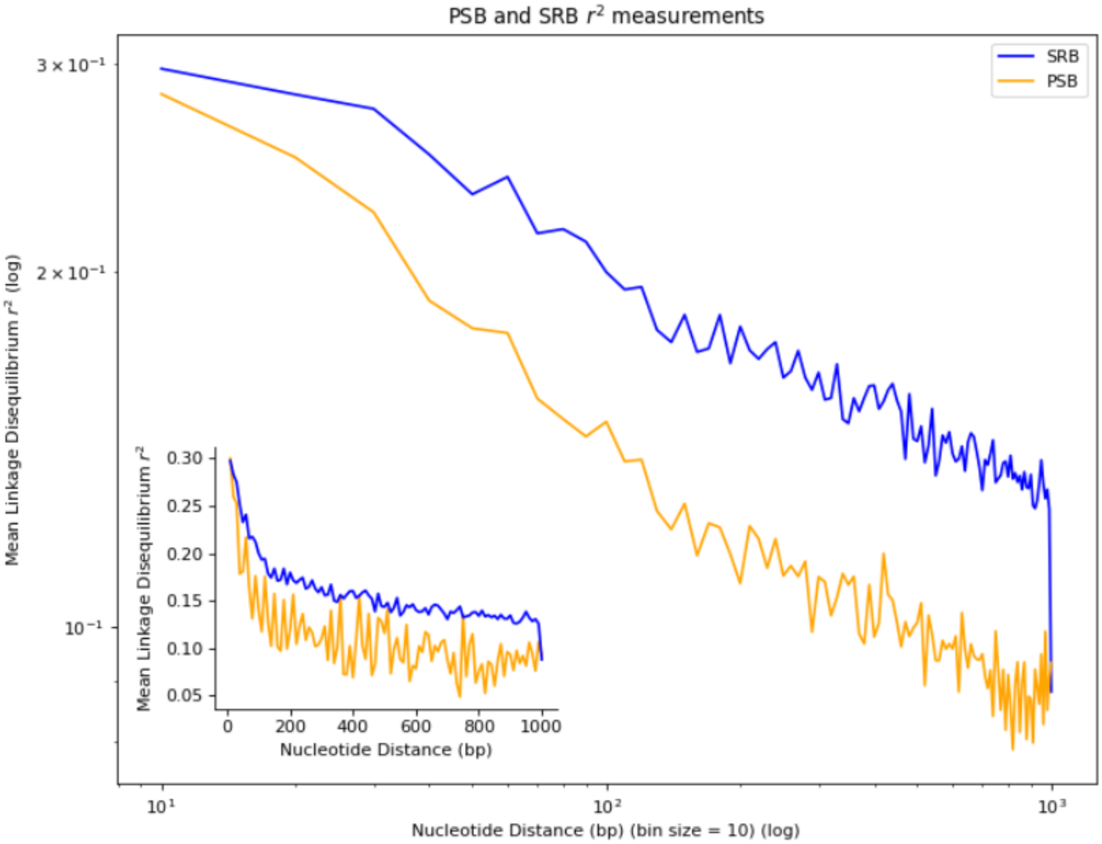

Figure 2.9: (main) Average r2 values for pairs of loci in the SRB and PSB dataset compared on a log-log scale with a different bin size (bin size = 10 bp).(inset) Average r2 values for pairs of loci in the PSB dataset and (yellow) average r2 values for pairs of loci in the SRB (blue) on a linear scale

clade.

Fig. 2.10 shows an example of a SNP matrix where the focus was on the loci fixed in the 9 clade which contains 3,776 fixed alleles. The dark green squares represent a mutation, white squares represent wild type, and light green squares represent missing data. Representing the dataset this way makes it easier to understand the relationships between the clades. For example, in Fig. 2.10 we observe that SNPs fixed in the 9 clade are rarely present in the E clade. Another interesting observation is the organisation of the mutated loci in the ladder. Although a lot of the mutations in that region show up

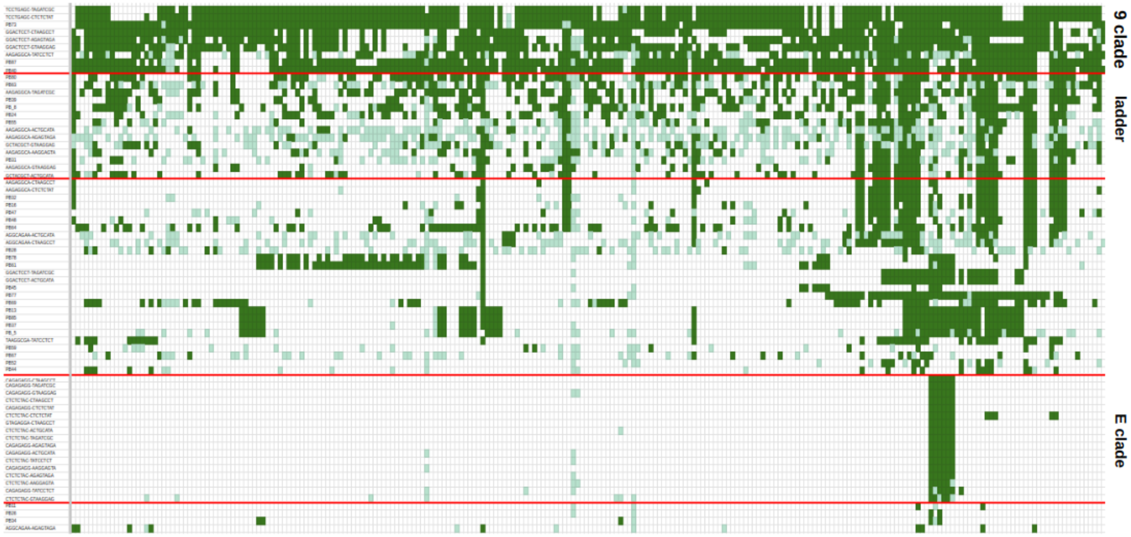

Figure 2.10: Part of a SNP matrix of fixed 9 clade SNPs. Rows represent strains in the order that they are observed on the tree and columns represent each biallelic locus (in order that they appear on the genome).

as a scatter of green squares, there are regions that create well-defined green blocks that indicate possible recent horizontal transfers.

The same methods were used to create SNP matrices for the SRB data. The SNP matrix shown in Fig. 2.11 combines alleles fixed in the three clades - the F, basal AB, and top AB. The loci fixed in the F clade are represented by green squares, loci fixed in the topAB are colored blue, and loci fixed in basalAB are colored orange. White squares represent wild type and light green squares represent missing information. The SRB matrix provides more evidence to the conclusions drawn from the heatmap in Fig. 2.5. The green mutations are rarely present in the basalAB or basal clades. The fixed loci of the topAB clade are also sparse in the basalAB or basal clades while being frequently observed in the F clade and the strains below the F clade. Those observations match what we have observed in the heatmap: F clade and topAB have more correlations between each other while basalAB and basal clade have higher values of correlation between

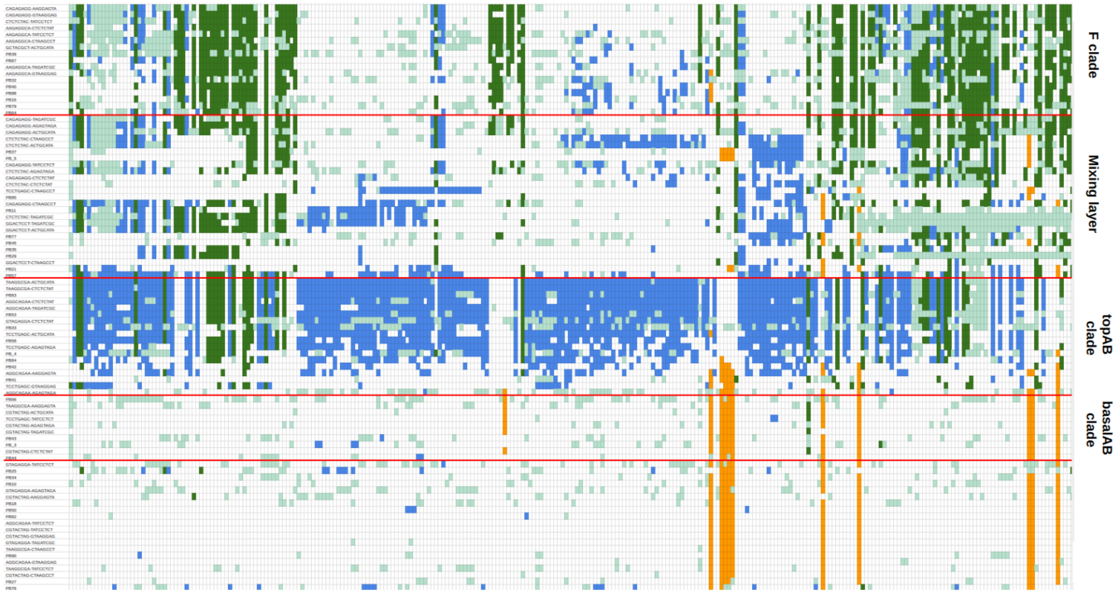

Figure 2.11: Part of a SNP matrix of fixed F, topAB, and basalAB clade SNPs. Rows represent strains in the order that they are observed on the tree and columns represent each biallelic locus (in order that they appear on the genome).

themselves. Additionally, similarly to what we have observed in the PSB matrix in Fig. 2.10, the fixed SNPs create well defined blocks of mutations in all clades, but especially in the clade between the F clade and topAB clade.

2.2.7 Observation of a genetic ”mixing layer” - like structure

Interestingly, the other strains that create a ladder-like structure (F strains that precede the 9 clade) have a monotonically decreasing number of the 9 clade fixed alleles which is not seen in the rest of the tree (Pearson’s correlation coefficient of -0.8353 and a P-value of 0.00019) (Fig. 2.12. This observation indicates that there have been many horizontal transfer events of the 9 clade genome pieces into the F strains closer to the base of the tree. This reservoir of fixed 9 clade alleles that is randomly incorporated to other strands is also observed in the heatmap in Fig.2.4. The percentage of shared SNPs is elevated for pairs of strains between the members of the 9 clade and the ladder compared to the background.

These observations suggest that the strains of the ladder actually represent a "mixing layer" - a set of recipient strains participating in an extensive horizontal transfer of genomic fragment originating from the 9 clade. In a sense, this pattern of variation can be compared to the pattern of recombination seen in eukaryotes, however, in contrast to the eukaryotic pattern of inheritance involving meiotic crossing-over and gene conversion where one would observe well-defined haploblocks, this SNP matrix shows that it is not the matter of crossovers between chromosomes but rather an exchange of information via small pieces of the genome that are received and incorporated into the chromosome.

In line with previous conclusions derived from our analysis of the PSB SNP matrix, we observed a mixing layer in the SRB tree as well. It is located in between the F clade and topAB clade. This mixing layer has a gradient of mutations that decreases from the F clade to the root for mutations identified as fixed in the F clade (green squares) and from the basal AB clade to the F clade for the mutations identified as fixed in the basal AB clade (yellow squares). Interestingly, the mixing layer does not show a clear gradient in any direction for the topAB fixed mutations. Instead, the mutations are more evenly distributed across the strains.

The observation of a "two-way" transfer into the mixing layer was investigated further by calculating the correlation between the fraction of alleles transferred from F clade to the mixing layer and the fraction of alleles transferred from the basal AB clade mutations to the mixing layer per strain. In order to do that, we calculated the frequency of fixed F clade alleles and the frequency of fixed basalAB alleles for each strain in the mixing layer. We plotted the results as a scatter plot where the frequency of the F clade alleles are on the x axis and the frequency of the basal AB clade alleles are on the y axis. Each point represents a strain from the mixing layer (Fig. 2.13). The Pearson's correlation coefficient is statistically significant (p-value < 0.01), implying that SRB is indeed experiencing a

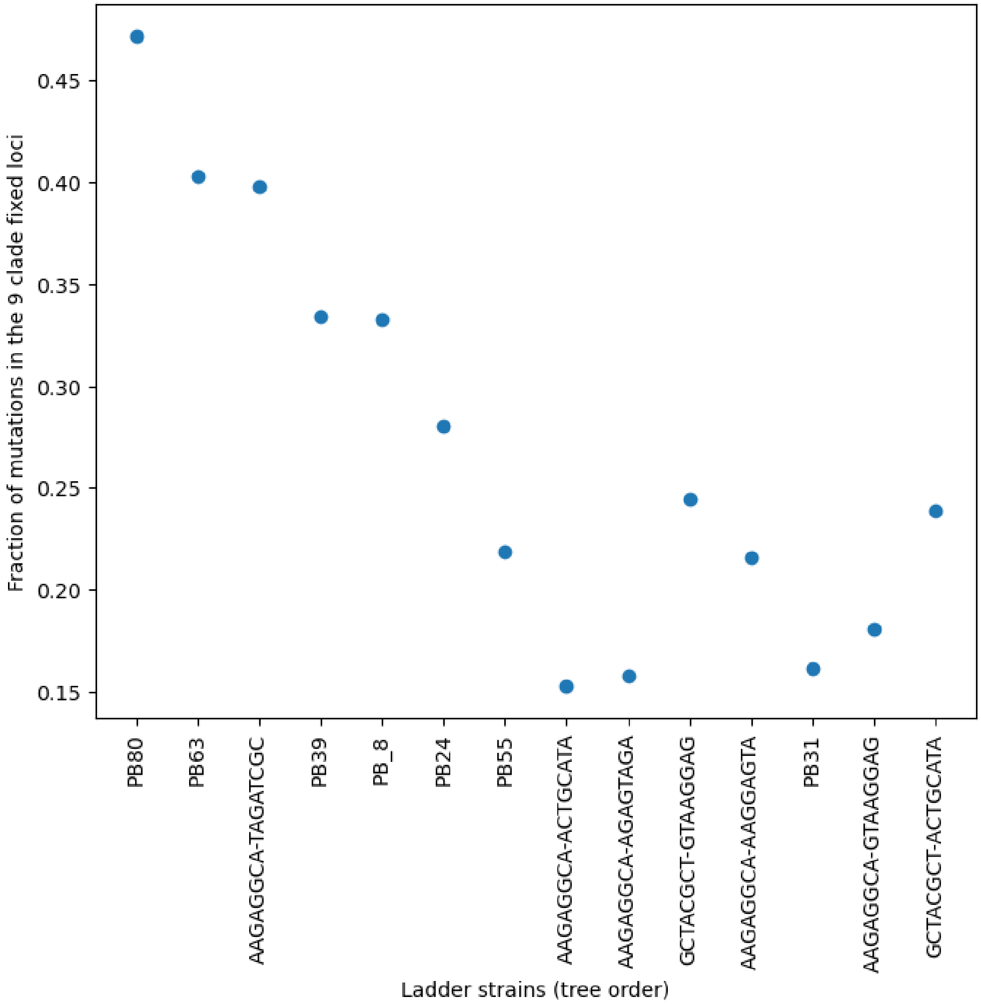

Figure 2.12: PSB frequency of 9 clade mutations in the ladder strains. The strains that create a ladder-like structure have a monotonically decreasing number of the 9 clade fixed alleles.

two-way exchange of information. Meaning, strains with more transfers from the F clade have less transfers from the basalAB clade and vice versa.

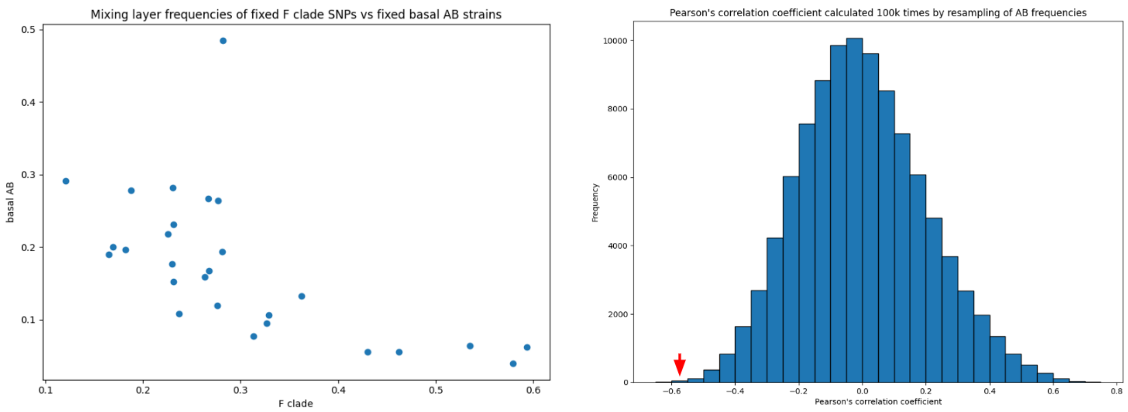

Figure 2.13: (Left) scatter plot of frequencies of F clade SNPs vs basal AB SNPs in the mixing layer. (Right) Pearson correlation coefficient calculated 100k times by resampling of AB frequencies.

2.2.8 Estimating recombination rate from one clade to another

The evidence presented above shows that both species are subjected to a lot of exchange of genome fragments. With this foundation established, we can now attempt to describe the patterns of HGT in this community. The bifurcating trees and heatmaps of SNP correlations have shown that each community is divided into smaller sub communities that share more alleles within each group than with the rest of the community. This is most evident in the PSB's E clade. This clade is clearly a very recent divergence since its alleles tend to be present through the whole clade, as if the ancestral strain that started the expansion of this clade already possessed those mutations and they were vertically inherited from then on. The same behavior can be observed in SRB's basal clade and basalAB clade. What those three clades have in common is the fact that they tend to keep those mutations and not share them with the rest of the community which suggests that those clades are recent expansions that did not have enough time to share their genome with the rest of the strains that tends to have a larger mix of different haplotypes. This notion is also supported by our measure of "age" of each clade. In the PSB, the E clade has the lowest average number of singletons (an average of 98.176 singletons per strain), while the other clades in PSB have a much higher average (for example, 598.22 for 9-clade). The SRB's clades have a more uniform number of singletons per strain (272.17 for F clade, 172.1 for topAB, 234.9 for basalAB, and 170.73 for basal clade).

However, the question of how each clade interacts with one another still persists. We have shown that recombination happens frequently in some subgroups while others tend to expand independently. Therefore, the next question is: can we describe how each clade interacts with the rest of the community and are those interactions similar between subgroups? In order to further quantify the results, we used our knowledge of the structure of these communities and calculated the rate of transfer of fixed alleles of one clade to another. We first calculated the relative "age" of each clade by counting the number of "private" alleles it contains. We define a private allele as any allele where the mutations only happen in the clade in question. This includes singletons but is not limited to them. The rest of the strains should have a WT allele or missing information at that position.

Once the age of the "acceptor" clade has been determined, we asked, among the alleles that are fixed in the donor clade, how many of them will also be found in the acceptor clade? In order for the fixed allele to be considered transferred, it had to have at least one mutation in any stain of the recipient clade. Lastly, this number was divided by the age of the recipient clade which determined the rate of transfer of an allele per de novo mutation. The results for both PSB and SRB are in the tables below.

The transfer rate from the 9 clade to the mixing layer is the highest out of all cal-

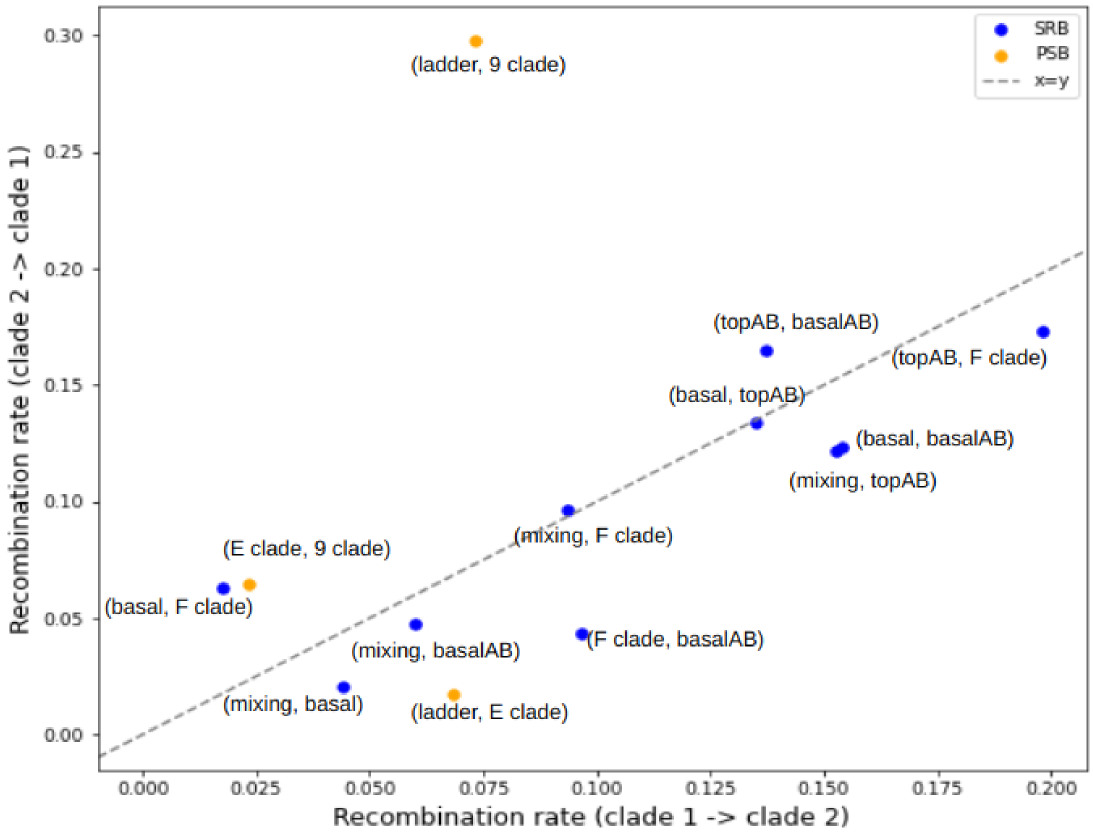

Figure 2.14: Recombination rates between clades. Each pair of clades is represented as a point on the scatter, labeled as (clade 1, clade 2). X axis represents the recombination rate from clade 1 to clade 2 while Y axis represents the recombination rate from clade 2 to clade 1. PSB clades are colored orange and SRB clades are colored blue.

culated rates and proves the conclusion we have drawn in the previous section - the 9 clade is the biggest contributor of transferred alleles into the mixing layer (Fig. 2.14). The rest of the transfer rates are very low which also enforces the conclusion drawn from the heatmap (Fig. 2.4). In the heatmap as well as Table 1, we see almost no correlation between the 9 clade and the E clade which enforces the conclusion that the transfer is minimal.

In the SRB, we see that the transfer is the highest between the F clade and topAB (0.198 one way and 0.173 the other) (Fig. 2.14. This is also apparent while looking at the SNP matrix in Fig. 2.11. The transfer to the mixing layer varies depending on which clade is the donor. The least amount of transfer comes from the clades closer to the root of the tree (basal clade and basalAB clade) while the most transfer comes from the F clade and topAB. One again, this information is also supported by both the SNP matrix and the heatmap.

2.2.9 Inference of haplotypes and HGT block sizes

In the previous sections, we have established that each tree has separate clades that have a unique set of fixed alleles resulting in a set of separate haplotypes that interact with each other and constantly share pieces of DNA between each other. Those clusters (F, basal AB, and top AB in SRB and 9 clade in PSB) send out small pieces of their genome and other strains incorporate those packages into their genome. Moreover, as proved above, the transfer rate from one clade to another is unique to that clade. To understand the features of those exchanges on a per strain basis, we first had to define which pieces of DNA come from which clade. This is an interesting problem since the blocks that can be distinguished by eye are not easily recognized computationally. Horizontally transferred regions will create a high degree of similarity for the strains or clades that participated in the transfer, creating well-defined blocks on the SNP matrix. However, our ability to efficiently determine the exact breakpoints of those regions is hindered by the noisiness of the blocks, with the majority including multiple missing data positions or possible reversal mutations. Therefore, the problem becomes: where does one transfer ends and another one starts? How probable is it for transfers to overlap? Is a data point labeled as wild type a wild type extending from a common ancestor or actually a reversal mutation from a mutated transferred block? In order to identify all of those transferred blocks and denoise them, we decided to utilize the information from all the strains that belong to a clade of interest and define it as a haplotype. In short, the problem can be simplified to the following question: given a profile for the haplotypes defined by the tree clades as well as a wild type haplotypes, and a new strain X with a sequence of characters S of the same length and the same notation as the profile matrices, with a substring of S (s) what is the probability that s corresponds to each haplotype?

After we calculated the log likelihoods for each haplotype across all SNP matrix windows (described in methods), we had to determine the most probable sequence of haplotypes and positions where the switching between one and another happens (Fig. 2.15). Since loci that are closer together on the chromosome will tend to stay together (which we have established using the linkage disequilibrium analysis) we decided to set a dynamic penalty cost for switching. It depends on the real genomic distance between the consecutive windows and follows an exponential decay function. This makes our model more biologically relevant. In order to define the haplotype sequence, we use Bellman-Ford algorithm where the algorithm traverses the graph and looks for the path between the source to the sink, choosing the path where the cumulative weights of the path will be the lowest. This way the haplotype algorithm defines denoised blocks that are statistically supported based on our dataset and previous results.

After running all tree strains through the haplotyping algorithm, we have discovered

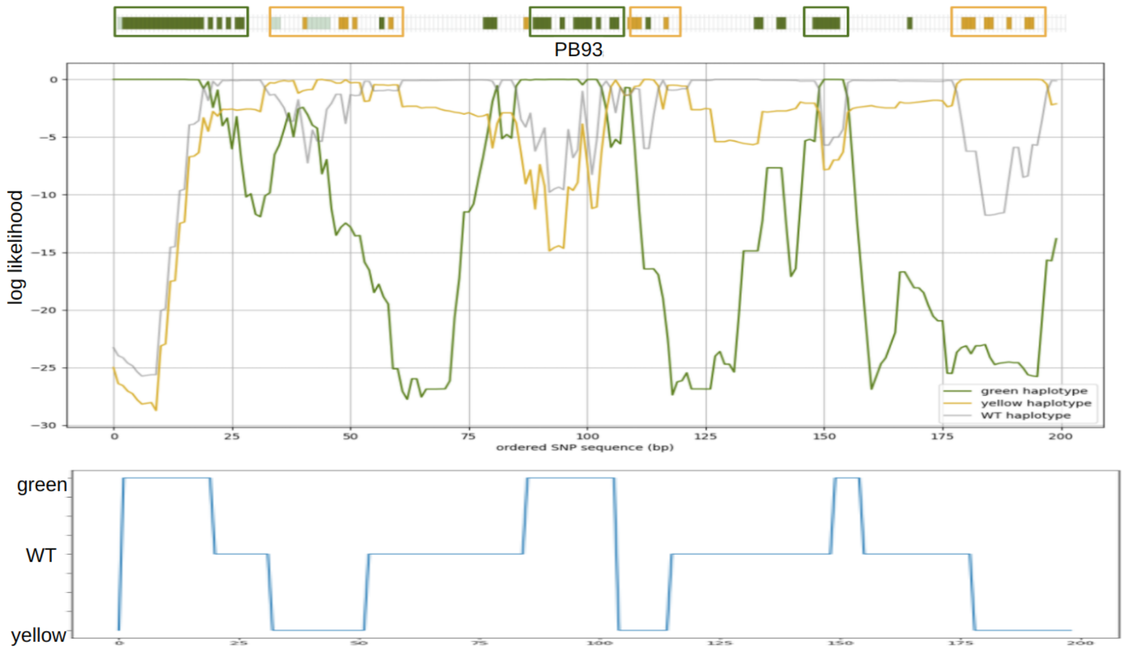

Figure 2.15: An example log likelihood graph of PB93 first 200 columns of the SNP matrix with the color-coded input array on top and the prediction using Bellman-Ford algorithm at the bottom. Each line represents log likelihoods for a different haplotype (green - F clade, yellow - basal AB, gray - WT). The closer the line is to 0, the more probable it is that this particular region came from the haplotype in question. Each level in the Bellman-Ford window represents the most likely haplotype for that part of the genome. Every time we switch a haplotype, we move to a different level on the graph.

that the mixing layer strains were always at the tails of the distribution, containing the largest amount of blocks transferred from the 9 clade or a composition of F clade, topAB, and basalAB (Fig. 2.16). When we look at the length distribution of those blocks in the mixing layer strains only we see an exponential decay in the length, with the majority of the transfers being very small. One can fit a line to this distribution and calculate the characteristic length which ended up being 8,686 bp for blocks transferred from the 9 clade, 2,824 bp for blocks transferred from F clade and basalAB clade only, and 5,156 bp for blocks transferred from F clade, topAB, and basalAB clades (Fig. 2.16) - much longer than what we have seen in the LD analysis. In order to understand why the characteristic

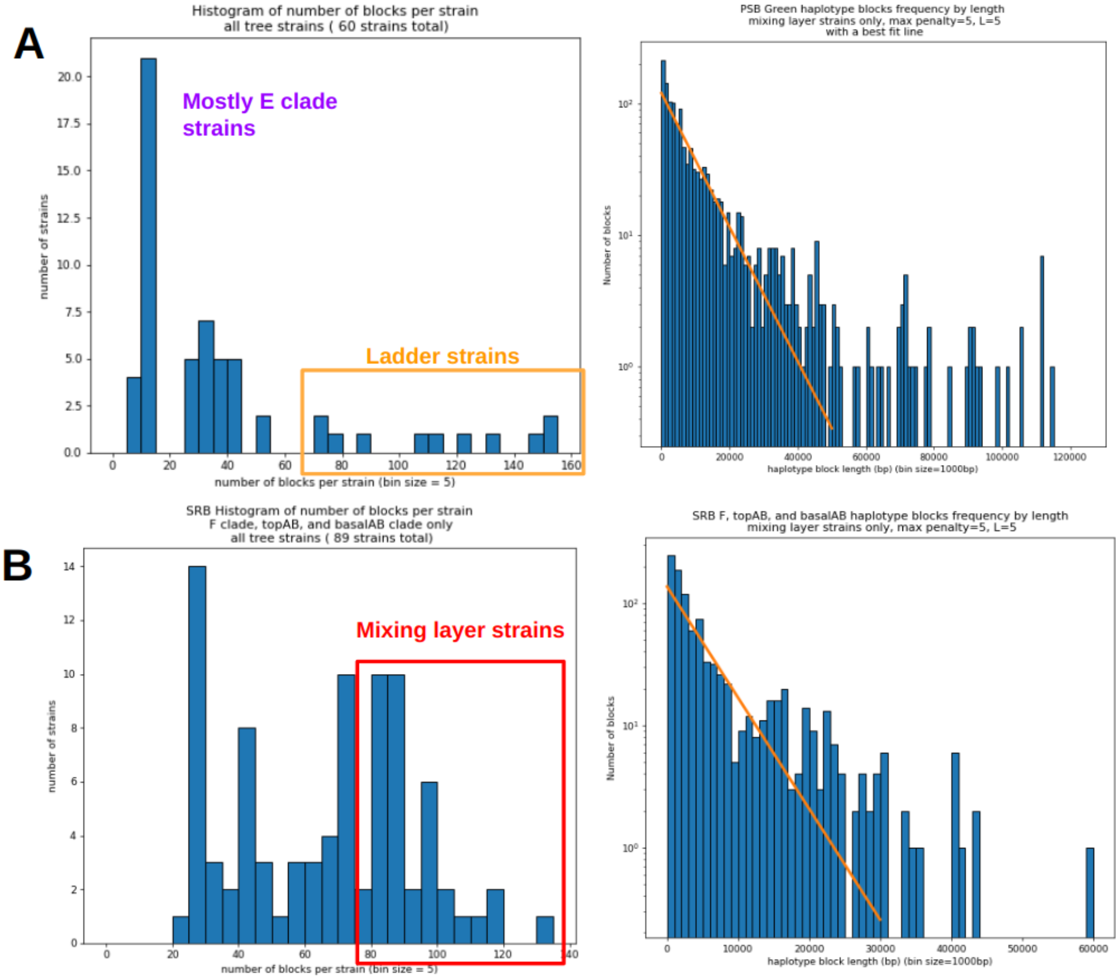

Figure 2.16: Distribution of number of transferred blocks per strain and their lengths. (A) Blocks transferred from the 9 clade (PSB). (B) Blocks transferred from the F clade, topAB clade, and basalAB clade (SRB).

lengths of the blocks and the characteristic length of LD does not match, it is important to remember that this section describes a different aspect of linkage. The identification of the blocks can only happen on larger distances since any very short blocks will most likely be labeled as noise rather than a signature of horizontal transfer. The previously described LD analysis uses pairwise correlations and, therefore, picks up more instances of linkage.

2.3 Discussion

In this study we have shown horizontal gene transfer characteristics of two closely symbiotic bacterial species: sulfide oxidizing bacteria and sulfur reducing bacteria. We note the extensive genetic exchange within each species and point out some universal characteristics of those exchanges, such as similar linkage disequilibrium characteristic lengths and transfer rate between different clades. Moreover, we show that the transfers seems to happen in a quasisexual manner and are most often detected in the genetic mixing layer.

We have shown that the SNP density distribution along the chromosome in both bacteria is not uniform. There are regions of high density that are more easily distinguished when one focuses on strains of just one clade. These regions, as also demonstrated by Sakoparing et al. [102]and Rosen et al. [46], are almost certainly involved in HGT transfer. What about the genes that are inside those SNP dense regions? A lot of those genes are involved in functions such as recombination, motility, antibiotic resistance, and biofilm formation. In addition, we noted that both bacteria exhibit negative (purifying) selection supported by the analysis of the ratio of non-synonymous and synonymous mutations happening in both genomes. Although both genomes experience negative selection as a whole, when we checked highly mutated genes of the PSB, we found a few genes that experience positive selection. They are involved in recombination (XerD, CRISPR-associated nuclease Cas1, type-1 restriction enzyme,Group II intron-encoded protein LtrA, and recombination promoting nuclear RpnD), while other are involved in overall survival of the cell (Immunoglobulin A1 protease, Fe(3+) ions import ATP-binding protein FbpC, and glutamine amidotransferase).

The phylogenetic trees and heatmaps show that each species exhibit a very well defined formation of clades that are sometimes well correlated with their geographical location (such as the E clade in PSB or the F clade in SRB) and can indicate a potential recent bottleneck event, but most of the time, there is a mix of different geographical locations within one clade. Based on the analysis of linkage disequilibrium measurement r^2, our bacteria shows signs of extensive HGT where loci are linked if the distance between them does not exceed 200 bp. This analysis also showed us that both SRB and PSB have a very similar pattern of r^2 distribution across different distances indicating that this behavior might be universal.

By visualizing our strains and their SNP content as a SNP matrix, we have found multiple blocks of SNPs with well defined edges that strongly indicate HGT events from one clade to another. Highly diverged clades (such as the F and basal AB clades in the SRB) also show how the transfer of their fixed alleles create a gradient from the clade into a mixing layer. Moreover, the visualization using the SNP matrices made it easier for us to spot blocks of missing data that indicate possible deletions. The significance of genes that are present in those regions still need to be assessed.

Moreover, we have shown that SRB's mixing layer is experiencing a two-way exchange of information from the F clade and the basal AB clade. By dividing the data into clades we were also able to determine transfer rates from each clade to the other clades. We have observed that the rates of transfer are unique and depend on the acceptor and donor clades but the rates reinforced the conclusions we drew from the correlation heatmaps.

Lastly, we have created a new algorithm for detecting haplotype blocks for highly similar strains. By establishing multiple haplotypes based on the structure of the tree we are able to probabilistically analyze each strain separately and determine the most probable sequence of haplotype blocks based not only on the sequenced data but also out knowledge of linked loci, making the results more biologically relevant. Using the results from this algorithm we were able to determine the presence of a positive correlation between the age of the strain or clade and the number of transferred blocks. This analysis gives us more insight into the relative strength of the recombination in these populations.

Since the beginning of population genetics research, there has been an ever-growing interest in how HGT can alter the ecology and evolution of the complex bacterial communities. The focus still persists on understanding the important topic of bacterial acquisition of new antibiotic resistance and adaptation to new environments across the world. There are many questions that need to be addressed in order to gain a better understanding of HGT. What types of genes are most often transferred? How are they transferred and under what ecological pressures? Are there any genetic or physical barriers that slow down the transfer? At what rate does the transfer happen? These questions are important especially if one wants to understand dynamics of naturally occurring communities that constantly experience new challenges to overcome in order to survive. Interestingly, the exact influence of HGT on such communities is not fully understood, especially since the rates, means, and conditions for transfer can differ widely between different species. In this chapter we have described the extent of the horizontal gene transfer in the PSB and SRB and showed that there is substantial evidence of transfer even between the strains of the same species.

2.4 Methods

2.4.1 Code availability

All scripts used for the analysis of the data can be found on github at github.com/adamadejska/pink_berry_scripts

2.4.2 The Dataset

The sequencing and assembly was performed by the Cordero and Wilbanks labs. Each berry was sequenced using metagenomic shotgun sequencing. The raw reads were trimmed, filtered, corrected for errors, and the adapters were cut off. The co-assembly of the prepared reads into contigs was performed using MEGAHIT software and contigs longer than 1 kb were chosen for binning. The sequenced reads were then mapped back to the constructed contigs using minimap2. In order to create the metagenomically-assembled genomes (MAGs), multiple different tools were used to bin the contigs to reduce potential bias. Lastly, the quality of the MAGs was assessed using CheckM software and each MAG was taxonomically classified using GTDB toolkit which identified each MAG using single-copy marker genes from the Genome Taxonomy Database.

Two most abundant species present in the aggregates were Thiohalocapsa (sulfide oxidizing purple sulfur bacteria (PSB)) and Desulfofustis (sulfate reducing bacteria (SRB)). Since the Thiohalocapsa MAG was very similar to the PB-PSB1 reference (assembled with long read PacBio data by Wilbanks lab), the reference was used as a baseline instead of the contigs created by the MAG. Additionally the single nucleotide variants were called by mapping both PSB and SRB to reference genomes. The SNVs were saved as a variant calling file (vcf) and shared with us for further analysis.

2.4.3 Creation of the dataset

For the PSB vcf, the strains have been sorted by the number of loci they have covered and out of 192 strains, strains with less than 80% of their biallelic bases called have been discarded, leaving 142 strains for further analysis. Only proper biallelic single nucleotide polymorphisms have been kept which we have assumed correspond to single mutational events in the history of its genomic locus. The SNP data has been labeled one of three different categories: wild type, mutant, or 'NaN' by majority calling. 'NaN' was given to any locus of a particular strain that did not have sufficient coverage to be informative (coverage less than 3 reads). The PSB data consists of one chromosome of length 7.9 Mb. There are 50,626 biallelic SNPs.

2.4.4 Creation of trees

A tree of the PSB strains was created based on the whole SNPs dataset (excluding singletons - genetic changes that happened in only one strain for a particular position in the genome). Additionally, we excluded any strains whose SNP matrix sequences contained more than 30% missing values. Since our dataset contains missing data, we calculate the distances between strains probabilistically using

$$D = \frac{\sum([v1_i - v2_i]^2)}{len(v)} \tag{2.1}$$

Where v1 and v2 are sequences of two strains (with mutations represented as 1, WT represented as 0, and missing sites (NaNs) represented as 0.5). The distances were calculated for each pair of strains, creating a triangular matrix of values that was then used to build a tree. To infer information about the structure of the population, a default MatLAB neighbor joining algorithm was used to build the actual relations. The trees were rooted to an "outgroup" that was entirely wild type.

2.4.5 Creation of the heatmap

The heatmap correlation figure was calculated using all non-singleton alleles and all strains present in the PSB tree. For each unique pair of strains, we counted the number of mutations they share. Any positions where both strains are wild type or where at least one strain contains missing information were ignored. The number of shared mutations was divided by the total number of acceptable sites and multiplied by 100. This information was then plotted as a heatmap using the seaborn python package.

2.4.6 SNP distribution across genome

The number of SNPs per 1 kb window was counted for PSB and SRB data separately and plotted as a line graph. For the SRB, PB93 strain was not included. The SNP-dense regions were defined as a region of length 1 kb with more than 20 SNPs for PSB and more than 30 SPNs for SRB. These regions were manually checked for presence of genes using the complementary gff file.

2.4.7 Nonsynonymous-to-Synonymous Ratio Analysis

The dN/dS analysis was performed in the following way for the PSB strains. First, we filtered the SNP loci based on their presence in open reading frames (ORFs). If a SNP locus was inside an ORF, we checked if the mutation is synonymous or nonsynonymous and saved this information. Next, we calculated synonymous and nonsynonymous opportunities for each gene in PSB. For each locus in an ORF, we substituted it with all three other possible nucleotides and checked if the change resulted in synonymous or nonsynonymous change. We added up all synonymous and nonsynonymous changes for each gene and used this information later to normalize the experimental counts. Lastly, for each pair of strains, we focused on the SNPs that were different between those strains and counted the number of synonymous and nonsynonymous mutations. For each ORF the SNPs were present in, we summed the number of synonymous and nonsynonymous opportunities and normalized our experimental counts to obtain the synonymous and nonsynonymous divergence (dS and dN).

The dN/dS values for the SRB were calculated in a similar manner but we had to take into the account that SRB's genome is divided into contigs so additional checks had to be added.

$$dN = \frac{\text{experimental number of nonsynonymous mutations}}{\text{number of nonsynonymous opportunities in genes affected by SNPs}} \tag{2.2}$$

$$dS = \frac{\text{experimental number of synonymous mutations}}{\text{number of synonymous opportunities in genes affected by SNPs}} \tag{2.3}$$

2.4.8 Linkage Disequilibrium and characteristic length analysis

Linkage disequilibrium was measured using the average squared-correlation of the genotypes at any pair of loci described in Lewontin, 1988. Each pair of biallelic SNPs have only four possible genotypes: 00, 01, 10, and 11. The frequencies of each genotype are represented as f_{00} , f_{01} , f_{10} , and f_{11}. The frequency of a mutation at position 1 is represented as $f_{1.}$ while the frequency of a wildtype is represented as $f_{0.}$. Similar frequencies are obtained for position 2 and represented as $f_{.1}$ and $f_{.0}$. The correlation is calculated as

$$r^2 = \frac{(f_{00}f_{11} - f_{01}f_{10})^2}{f_{1.}f_{0.}f_{.0}f_{.1}} \tag{2.4}$$

All SRB contigs were used for this analysis and the LD analysis was performed. The r^2 correlation was calculated for all pairs of loci below distance 1 kb from each other in the PSB dataset.

2.4.9 Creation of the SNP matrices & analysis

We restructured the previously created tree as a matrix with strains as rows (organized in the same order as the tree) and individual SNPs as columns (sorted based on their genomic position). Multiple matrices were created that varied in their SNP composition. One of the matrices focused on the SNPs that were nearly fixed in the 9 clade. In order to create this matrix, the SNPs were filtered based on their presence in the 9 clade. If a SNP was present in more than 50% of the 9 clade strains, then the SNP was included in the matrix, ultimately, the matrix consisted of 3776 SNPs. Additionally, strains that had over 40% of their SNPs classified as missing were discarded from the matrix as well as strains that had purely wild type alleles. The cells of the matrix were colored for easier interpretation of the results by eye. Mutations were colored dark green, wild type alleles remained white, and missing data was colored in light green.

A similar algorithm was used to restructure the SRB tree as a SNP matrix as described above. The SNPs were filtered based on their presence in either the F, basal AB clades or top AB. If a SNP was present in more than 50% of the F strains. basal AB or top AB strains, then the SNP was included in the matrix. Additionally, strains that had over 40% of the chosen SNPs classified as missing were discarded from the matrix as well as strains that had purely wild type alleles. The cells of the matrix were colored for easier interpretation of the results by eye. The SNPs fixed in the F clade were colored dark green and the SNPs fixed in the basal AB clade were colored orange, and SNPs fixed in the top AB clade were colored blue.

In order to check for any correlations between transfer of F clade and basal AB clade mutations to the mixing layer, we have scattered the mutation frequencies based on the F clade (x-axis) and basal AB clade (y-axis) for each strain of the mixing layer and calculated the Pearson's correlation coefficient of the scatter. To check for significance, we have performed a resampling analysis where the values of F clade transfer were shuffled and assigned to different values of basal AB transfers. We have recalculated the correlation coefficient 100 thousand times and plotted the distribution of all correlation coefficients.

2.4.10 Estimating transfer rate from one clade to another

The rate of transfer from one clade to another was calculated for all possible pairs of clades. Each clade was either an "acceptor" or a "donor". For the acceptor clade, we counted the number of loci unique to that clade. A locus was considered unique if it only contained mutations in the strains of the acceptor clade. The rest of the strains had to have a wild type or a missing allele at that position.

Next, we found all loci that are fixed in the donor clade. A locus was considered "fixed" if more than 50% of the strains of the clade had a mutation at that position. The alleles at the rest of the tree did not matter.

Finally, we counted how many of the donor loci have a mutation at at least one acceptor strain. If at least one acceptor clade strain contained a mutation at the fixed donor locus, it was considered horizontally transferred. The rate of transfer was calculated by dividing the number of transferred donor loci by the number of unique acceptor clade loci.

2.4.11 Inference of haplotypes

The creation of haplotype profiles. A 4 x m matrix is created for each haplotype where m is the number of SNPs in the SNP matrix. Each cell in the matrix represents the probability of observing a certain color at position j and is computed using equation 2.5.

$$P(color_i, j) = \frac{\text{sum of } color_i \text{ at position } j}{\text{number of strains}} \tag{2.5}$$

Computation of profile probabilities. The analysis is performed individually on each strain S chosen from the mixing layer of the SRB tree. S is an 1 x m array of integers from alphabet α = {-1, 0, 1, 2} where each number from α corresponds to a SNP from a particular haplotype. We traverse S with a small window s of size n (n in range 5 to 10 nucleotides). We calculate the probability that s came from each haplotype using Bayesian inference.

$$P(H|s) = \frac{P(s|H)P(H)}{P(s)} \tag{2.6}$$

$$P(H) = \frac{\text{number of haplotype alleles in the whole SNP matrix}}{\text{number of all alleles in the whole SNP matrix}} \tag{2.7}$$

$$P(s|H) = \prod_{i=1}^{n} P(allele_i, i) \text{ for a window of size } n \tag{2.8}$$

$$P(s) = \sum_{H} [P(s|H)P(H)] \tag{2.9}$$

By combining equations 2.6 - 2.9, we calculate the final log likelihood estimate for each window and each haplotype using equation 2.10.

$$log[P(H|s)] = log(\frac{\prod_{i=1}^{n} P(allele_i, i)P(H)}{\sum_H[\prod_i^n P(allele_i, i)P(H)]}) \tag{2.10}$$

At the end of these calculations, we obtain a graph of log likelihoods for each haplotype across the genome.

Inference of the most probable sequence of haplotypes We use the Bellman-Ford algorithm, a shortest path dynamic programming algorithm for the inference of the most probable sequence of haplotypes. In order to use this algorithm for our analysis, we re-imagined the log likelihoods graph as a directed, acyclic graph where the edges between nodes of the same haplotype have a weight of the absolute values of the log likelihood values while the weight of the edges that switch the haplotype have a weight of a penalty. The penalty cost follows an exponential decay function 2.11

$$p * e^{(\frac{-1}{L} * x)} \tag{2.11}$$

where x is the distance in kb, L is the characteristic length (L=5), and p is the maximum penalty (p=5 for the SRB, p=1 for PSB).

2.4.12 Transfer rate analysis

Given a sequence of most likely haplotypes per strain, we traverse each list and count how many blocks each strain acquired from all sources (F, topAB, and basal AB clades in case of SRB) or just one source (9 clade in case of PSB). The number of blocks for SRB dataset was calculated using the haplotypes sequences created using max penalty = 5 and L = 7 while the PSB blocks were calculated using the haplotypes sequences created using max penalty = 1 and L = 5. For the SRB dataset, the haplotyping was performed only on contigs that had more than 100 SNPs present in the SNP matrix in order to have enough information for meaningful haplotype inferences (total of 12 contigs: contig_0, contig_5, contig_6, contig_9, contig_14, contig_24, contig_28, contig_67, contig_69, contig_70, contig_79, and contig_87). A block was defined as a sequence from either source of any length. In the case of the PSB dataset, some of the strains that were closer to the 8 clade had a lot of SNPs that appeared scattered across the sequence. Because of that the Bellman-Ford algorithm sometimes identifies blocks of length 1. Upon closer manual inspection, those "blips" in the haplotyping were still blocks but those blocks were very short or introgressed by the wild type allele. We decided to include those very short blocks in the PSB analysis. In addition we counted the number of singletons and consider this number the "age" of the strain since its divergence from the root.

We have plotted the number of HGT blocks versus the number of singletons for each strain to see if there is any correlation between the two. After plotting the scatter plot, we calculated the best-fit line between the points and Pearson's correlation coefficient (coeff = 0.488 for SRB, coeff = 0.553 for PSB). To conclude if this correlation is statistically significant, we shuffled the HGT blocks data 100,000 times, assigning different HGT block numbers to different singleton counts, and recalculating the Pearson's correlation coefficient for each shuffle. The distribution of correlation coefficients is plotted and compared to the experimental data.

Each strain was processed using the haplotyping algorithm and the cleaned up haplotype blocks were used for the comparison. For each pair of strains, we have calculated the overlap between their transferred blocks. If a position p had the same color in each of the strains (aka. The transferred positions come from the same haplotype), we counted that as overlap. The number of those events were indicated as s1&s2 in the equation below. Otherwise, the position is counted as strain 1 or strain 2 exclusive (indicated as s1 only or s2 only). Positions where at least one strain from the pair had missing information

(NaN) were ignored. Positions where both strains had WT alleles were also ignored. The overlap was calculated using equation 2.12.

$$overlap = \frac{s1\&s2}{(s1\&s2) + \text{s1 only}} \tag{2.12}$$

Therefore, if the two strains are identical their calculated overlap will equal 1 and if there is no overlap at all, this value will be 0.

2.5 Permissions and Attributions

The content of Chapter 2 and is the result of a collaboration with Otto X. Cordero at the Massachusetts Institute of Technology and Elizabeth G. Wilbanks, at the University of California, Santa Barbara.

Chapter 3

Defense systems in Purple Sulfur Bacteria and Sulfate Reducing Bacteria

3.1 Introduction

In nature, bacteria live in complex environments, always competing with other organisms for limited environmental resources. There are many different ways an organism can interact with other organisms. Such interactions can be categorized as symbiosis, commensalism, or competition. In a symbiotic relationship both parties involved in the interaction benefit from it; for example, a bacteria can provide nitrogen fixing activity for a plant that acts as their host. In commensalism, one party benefits while the other does not get any benefits but also is not harmed in any way from the relationship. An example of commensalism would be when one bacteria produces a waste metabolite that is then used as an energy source to a different bacteria. Lastly, in a competition, both parties compete for limited resources and try to eliminate each other - neither organism benefits from the presence of the other. Many examples of competition between bacteria exist. For example, aerobic soil bacteria communities need to compete between each other for access to sunlight and oxygen. In order to ensure their own survival in a demanding environment, different species develop different strategies to impede the growth of other organisms. One of such strategies - and the main focus of this chapter - is self/nonself recognition[55]. It consists of creating toxins (small proteins that negatively influence cells) that other competitors are not immune to in order to halt their growth or eliminate them all together. Each bacteria produces its own unique set of toxins as well as a corresponding set of immunity proteins that ensure safety of the toxin-producing bacteria. Those immunity genes follow the toxin gene closely on the chromosome, ensuring concurrent transcription.

Contact dependent growth inhibition (CDI) is one of many examples of how bacteria compete in a shared environment. This system delivers a toxin from one bacteria to another upon contact and the toxin suppresses the growth of the target cell. Usually, the proteins that are a part of the CDI secretion contain multiple distinct parts - the structure that acts as a delivery mechanism, the toxin, and the antitoxin[57, 58]. The C-terminal region is highly variable and shares sequence identity with the toxic effector domains for CDI. In order to only eliminate competition and not itself, the system is equipped with an antitoxin that is expressed with the toxin[59]. This way bacteria that produced the toxin are immune to it. Each type differs in the mechanism it uses to pass the toxin to the target as well as their activity and primary purpose.

In this chapter, we discuss preliminary findings of diversity in WapA and RhsC C-termini in purple sulfur bacteria (PSB). We show that there is correlation between geographical location and the presence of different toxins. These observations mark a promising starting point for studying CDI mechanisms in naturally occurring bacterial populations. Additionally, we show evidence of clades having different defense genes based on the lowered depth of coverage of those ORFs when compared to the rest of the genes.

3.2 Results and Discussion

Pink berries, which are macroscopic, photosynthetic bacterial aggregates, create extensive microbial mats at the Little and Great Sippewissett Salt Marshes on Cape Cod, Massachiussetts. They can form large aggregates, almost up to a centimeter in diameter. Each berry contains hundreds of different bacterial species that together create complex metabolic interactions due to the proximity of the microbes that enhances cell-cell contact and allows for extensive genetic exchange. Although the berries can create a very diverse ecosystem, there are two species that dominate their space. The intense pink color of the berries come from a genus Thiohalocapsa (Sulfide-oxidizing bacteria, aka Purple Sulfur Bacteria (PSB)) which strongly depend on another widespread species of proteobacteria from the family Desulfocapsaceae (Sulfur Reducing Bacteria (SRB)). The pink berries are sulfur-cycling symbiotic consortia in which PSB and SRB physically accumulate along with a variety of other bacterial species. That close proximity allows PSB and SRB to engage in interspecies electron transfer and together create a sulfur cycle - PSB oxidizes S^{-2} to SO_4^{-2} (oxidizes sulfide to sulfate and stores elemental sulfur inside its cells) and SRB performs the reverse of that action which closes the cycle (reduces sulfate to sulfide)[54].

Although it has been established that phylogenetic trees do not faithfully describe the evolutionary history of a bacterial population because of the extensive of horizontal transfer of genomic fragments, a phylogeny reconstructed from the biallelic SNPs from our dataset can still be a source of information about the structure of the dataset and relationships between strains. If no recombination is present in the population (i.e. all mutations are vertically inherited), there exists only one tree that will describe the evolutionary history of the sample. However, when recombination is present and a common occurrence, the structure of the tree will vary based on the regions of the chromosome used to build it. Keeping this in mind, a tree of the PSB strains was created (excluding singletons and strains with large amounts of missing data) and the resulting neighbor-joined tree revealed multiple clades that are clearly separated on the tree. While some of those structures are easily identifiable by eye on the graph of the tree, for clarity, we also plotted it as a structured heatmap. The rows and columns are ordered the same way as the strains on the tree and the scores represent percentages of the shared SNPs between each pair of strains. The heatmap was annotated to show each clade. It is easier to see that the heatmap contains multiple distinct clusters that are much more similar to each other than the background. Those clades include a clade of closely related strains exclusively from location E (which could indicate a possible recent population bottleneck), a highly diverged clade (top of the tree and top left of the heatmap) containing 9 bacterial strains (what we call from now on a “9 clade”), and a group of strains from the same location F exhibiting a star-like structure (Fig.3.1). Of note is the observation that those strains create a larger cluster on the heatmap with similarities to the 9 clade.

PSB has 50,626 biallelic mutations and many of those mutations cluster into small regions of the genome creating high frequency regions. It is highly probable that those high SNP frequency regions have been horizontally transferred. In order to determine if any of those regions contain functionally interesting genes, we have checked which open reading frames (ORFs) have a high number of mutations compared to their length (Fig.3.2).

Figure 3.2 shows all known genes in the PSB dataset, indicated as blue dots. For each gene we have counted the number of mutations in that gene and plotted it against the length of the gene. We would expect that longer ORFs have a higher random chance of accumulating more mutations than shorter ORFs. The red line shows the best fit line for this dataset and the green lines represent one standard deviation from the mean. The genes present within the green boundaries would be considered as having an expected

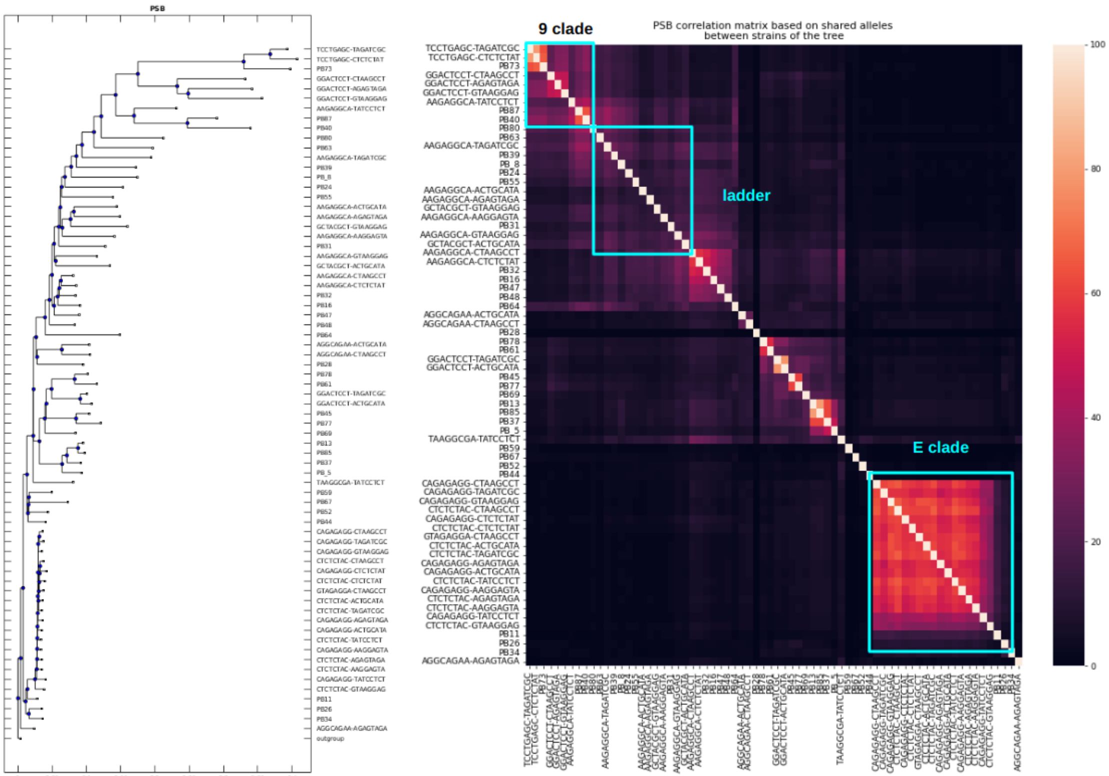

Figure 3.1: PSB tree (left) and correlation matrix (right) based on the collection of all non-singleton SNPs and probabilistic distance calculations, showing multiple separate clades.

number of mutations given their length and assumption of a model where mutations are truly random and uniform. The genes that have an abnormal number of mutations given their length have been labeled. As seen in the figure, five of those genes are from the wapA and rhsC families. Since those genes are known to play a part in the self/nonself recognition system, we have decided to investigate them further.

Further inspection of the PSB genome, based on our general feature file, tells us that there are 6 WapA and 5 RhsC ORFs in total in the PSB genome. The set of WapA and RhsC ORFs we have chosen are followed by a set of hypothetical genes (blue blocks labeled "hp") (Fig.3.3). After extracting the DNA sequence of those ORFs and using BLAST on their amino acid sequences, we found that a lot of them were identified as

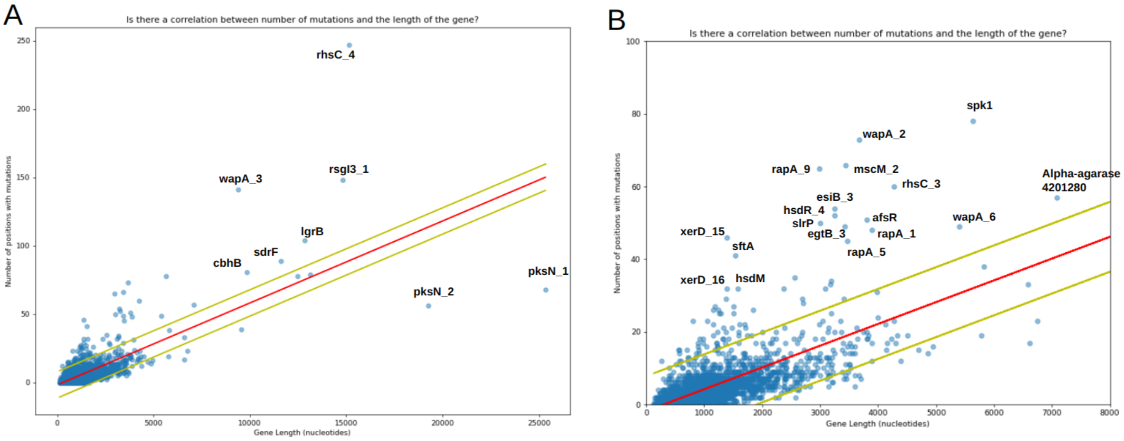

Figure 3.2: Length of the gene vs number of mutations it contains. Right plot is a zoomed in version of the left plot. The red line indicates a best-fit line, showing correlation between the number of mutations and the length of the gene. The two yellow-green lines indicate one standard deviation from the best-fit line. Genes with an abnormal number of mutations based on their length have been labeled.

transposases and Ig-like domains. Since both WapA and RhsC are proteins that encode toxins used for growth inhibition of competitors, the antitoxin has to be transcribed at the same time to prevent autoinhibition. The Ig-like domains might be those antitoxins.

Next, we further inspected the mutations found in the CDI genes. It is apparent that each strain will have a unique set of mutations found in its chromosome. We have also established above that the strains are closely related to each other and create multiple clades when represented as a phylogenetic tree, showing relations between strains based on their composition of mutations. Moreover, the clades seem to also be influenced by the geographical location of the strains (seen for example in the E clade). In PSB, we see two very distinct clades - the 9 clade and the E clade. We have examined each strain and counted the number of mutations contained for each CDI protein of interest. We found it surprising that the strains with the largest number of mutations for a specific CDI gene were all from the same clade. For example, strains that contained the most mutations in the WapA 6 ORF were mostly from the geographical location C (12 out

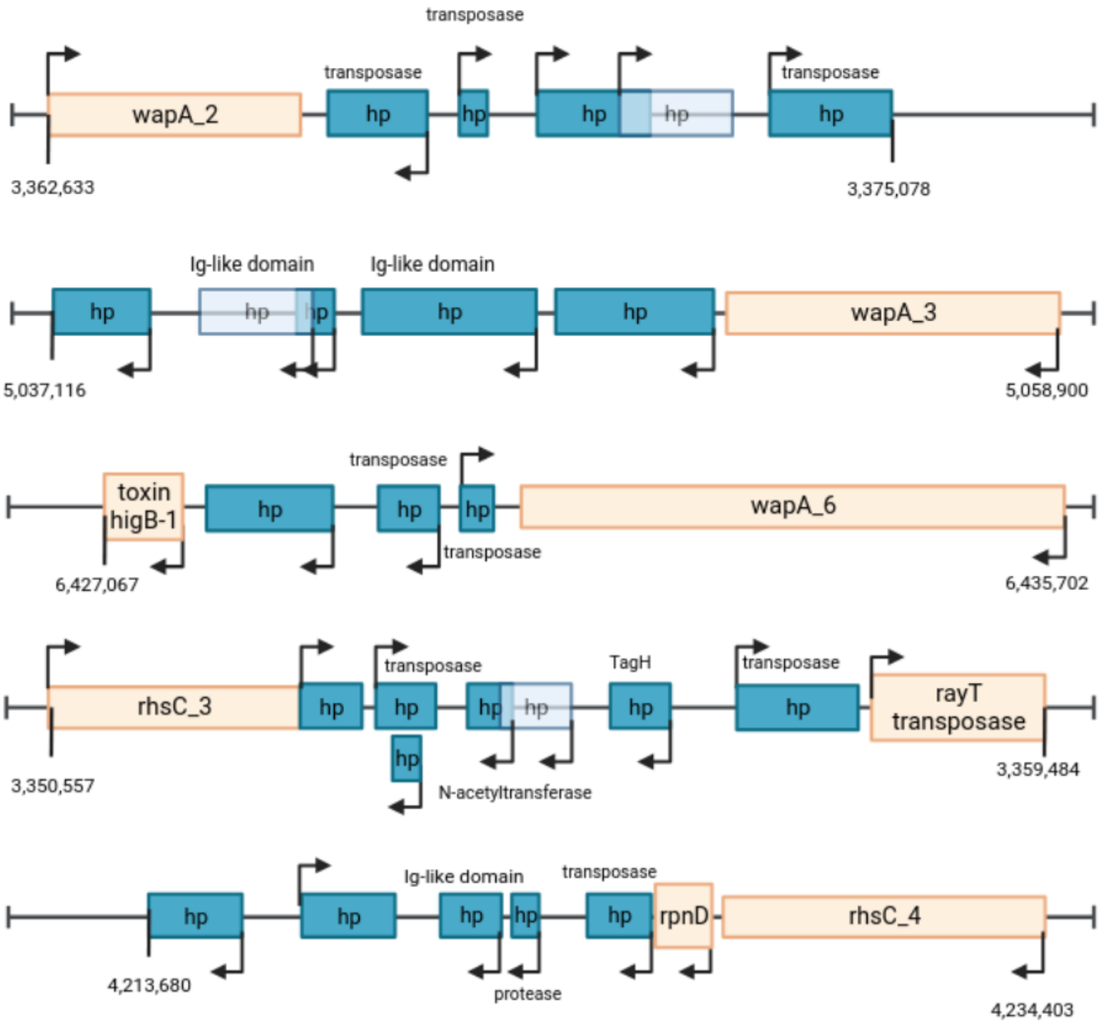

Figure 3.3: A diagram of CDI ORFs and ORFs that follow them on the chromosome. Many of them are labeled as “hypothetical proteins” (hp) in our data files, however, when one runs the sequences through BLAST, a lot of those unknown proteins are recognized as Ig-like domains or transposases.

of 16 strains containing a high number of mutations) (Supplemental Fig. 1). Similarly, strains from the E clade were the ones that contained the most mutations for WapA 2 ORF (16 out of 19 strains containing a high number of mutations) (Supplemental Fig. 2). The trend was seen in all 5 genes - the mutations seem to be fixed in a particular clade and then randomly scattered in the rest of the berries if present at all (Supplemental Fig.

1-5). WapA 2 ORF's mutations are almost exclusively present in the E clade. WapA 3 has mutations that are accumulated in the majority of the 9 clade strains. WapA 6 accumulates mutation in another geographical location - C. This is an interesting observation, especially the fact that a lot of the clades are separated by geography as well, making it look like the berries have different toxins or mechanisms of defense against their cousins from a different pond.

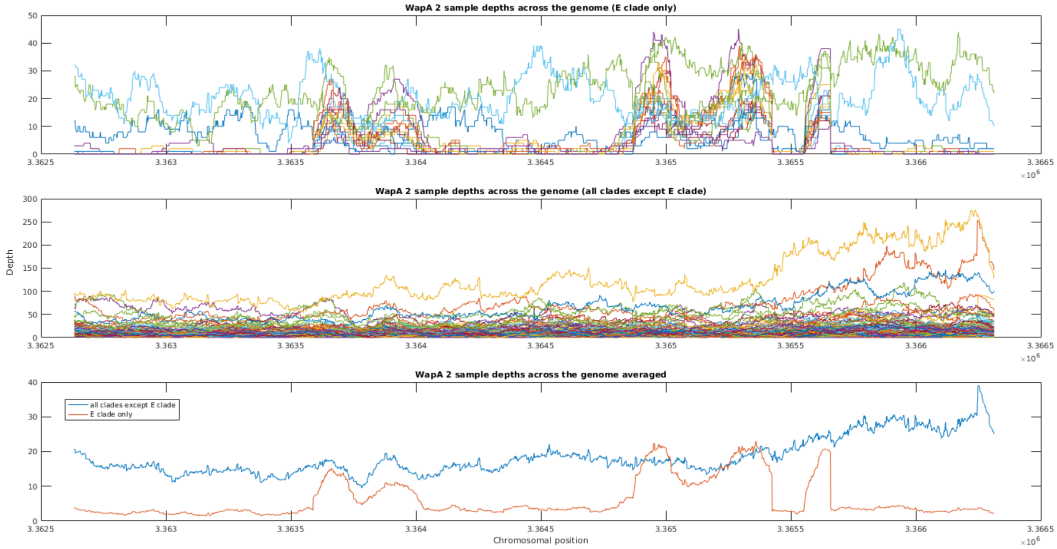

Figure 3.4: Depth of coverage for open reading frame of WapA 2 gene. Top panel shows individual depth of coverage for each strain in the E clade. Middle panel shows individual depth of coverage for each strain in the PRB tree (except for the E clade). Bottom panel shows the average depth of coverage for E clade strains (orange line) and all other strains (blue line).

We have been wondering why we see such a high number of mutations in only specific clades. After examining how the mutations are distributed on a SNP matrix, we have also noticed a high number of missing loci (loci with not enough coverage for a confident call). Intrigued, we checked the full depth of coverage of each ORF for each berry. We have separated the strains into two groups. Group one contained strains that had a high number of mutations in an ORF in question (usually the whole clade) and group two contained strains with a low number of mutations in the ORF in question (usually the rest of the tree). We plotted three graphs - the top graph shows the depth of coverage across the ORF for each separate strain in the group with high mutation numbers, the middle graph shows the depth of coverage across the ORF for each separate strain in the group with low mutation numbers, and the bottom graph shows the average depth of coverage for each group. After examining the graphs for each gene, it is apparent that there are differences in the depth of coverage between the two groups. For example, in WapA 2 ORF, the E clade's gene is almost completely deleted (the depth of coverage is mostly close to 0) while the rest of the tree has a very steady depth of coverage of about 20 across the gene (Fig.3.4). This kind of signal is also present in the other genes.

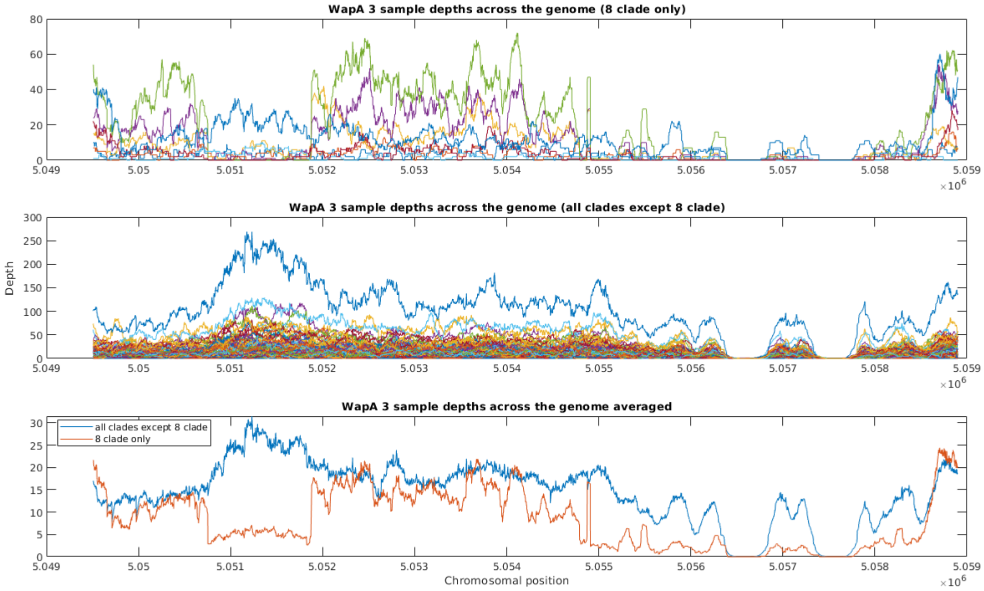

Figure 3.5: Depth of coverage for open reading frame of WapA 3 gene. Top panel shows individual depth of coverage for each strain in the 9 clade. Middle panel shows individual depth of coverage for each strain in the PRB tree (except for the 9 clade). Bottom panel shows the average depth of coverage for 9 clade strains (orange line) and all other strains (blue line).

Figure 3.5 shows that WapA 3 lacks a specific chunk of its gene in the 9 clade which is close to the C terminus of the gene (since wapA 3 is on the negative strand). This could indicate a deletion for this gene in the 9 clade. However, another explanation could be that this specific region contains a different DNA fragment that did not align to the reference. It is a plausible explanation given the fact that Illumina sequencing used for collecting this dataset can only sequence small DNA fragments at a time (around 500 bp in length). It is possible that this region contains a different toxin and the alignment algorithm was not able to align this divergent sequence to the reference.

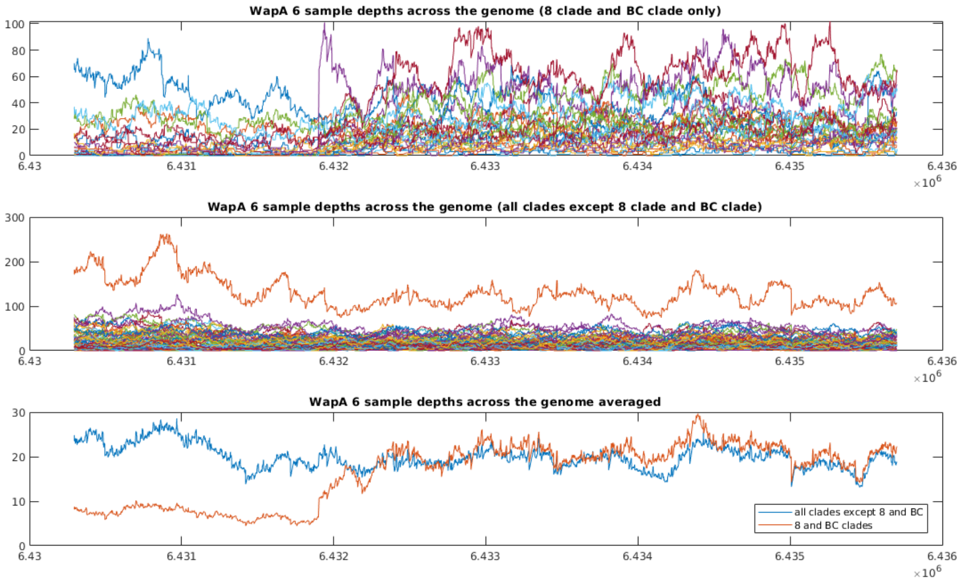

Figure 3.6: Depth of coverage for open reading frame of WapA 6 gene. Top panel shows individual depth of coverage for each strain in the 9 clade and BC clade. Middle panel shows individual depth of coverage for each strain in the PRB tree (except for the 9 clade and BC clade). Bottom panel shows the average depth of coverage for 9 clade and BC clade strains (orange line) and all other strains (blue line).

The same situation is seen in WapA 6 where 9 clade and BC clade completely lack the C-terminus part of the protein (WapA 6 is also on the negative strand) (Fig.3.6). This could suggest that WapA 6 in the 9 clade and BC clade underwent a deletion that completely got rid of the toxin or that the genome was recombined and WapA 6 now has a different toxin attached to it that did not align to the reference. Just as we have seen in Fig.3.5 with WapA 3.

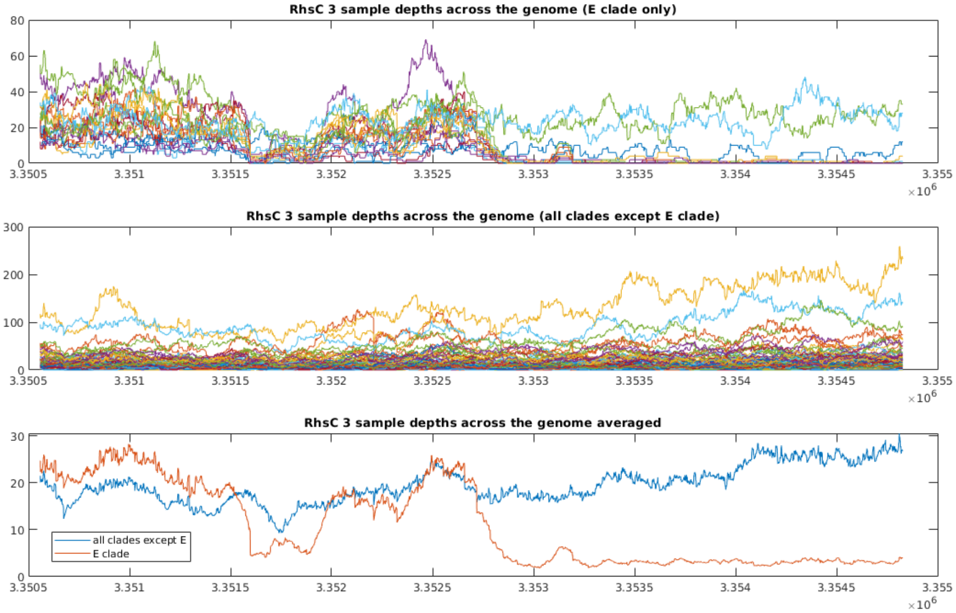

Figure 3.7: Depth of coverage for open reading frame of RhsC 3 gene. Top panel shows individual depth of coverage for each strain in the E clade. Middle panel shows individual depth of coverage for each strain in the PRB tree (except for the E clade). Bottom panel shows the average depth of coverage for E clade strains (orange line) and all other strains (blue line).

Amusingly, the same behavior is also seen in RhsC 3 (Fig.3.7). Just as with WapA 3 and WapA 6, the C-terminus of the gene is gone in the E clade. The only exception from this set of genes is RhsC 4 where the depth of coverage is more or less the same regardless of the clade we look at. As previously mentioned, there are six WapA and five RhsC ORFs in the PSB genome. We have checked the average depth of coverage for the rest of the genes (WapA 1, WapA 4, WapA 5, RhsC 1, RhsC 2, and RhsC 5) to check if we will see C-terminus deletions in those ORFs as well even if they do not have an abnormally high number of mutations (Fig.3.11, 3.12). All of those ORFs show a very steady average distribution per clade. E clade, BC clade, and the basal clade have high average depth of coverage for all of those six ORFs. Interestingly, the F clade and the 9 clade have a very low coverage for the majority of these ORFs (except for WapA 4 and RhsC 5 where each clade has a comparable depth of coverage). For the rest, the average depth of coverage for the 9 clade and F clade is kept around 10, which could indicate a possible deletion of the whole gene.

3.2.1 Movement of Genes involved in defense against phages

Transduction - a process where a phage transfers a piece of DNA from one cell to another - is one of the possible ways the DNA can be horizontally transferred. Of course, this process needs to be regulated and bacteria developed multiple different defense mechanisms against the transfer. One of the most famous ones is the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) that matches viral sequences and protects the cells against known viruses. However, CRISPR is only one of many protection mechanisms, which include mechanisms such as Septu (discovered in *Bacillus* in 2018), DarTG toxin-antitoxin system, the Defense-associated Reverse Transcriptases (DRT), and the Defense Island System Associated with Restriction-Modification (DISARM) among many others. It is apparent that more systems will be discovered in the future.

Since exchanging new tools for phage defense is advantageous to bacteria, in this section we focused on identifying the defense systems used by PSB and SRB and found that ORFs associated with phage defense have a higher probability of having lower depth of coverage and, therefore, being more mobile. We started our analysis by running both genomes through DefenseFinder - a tool for detecting prokaryotic antiviral systems from a genome sequence. DefenseFinder uses a collaborative knowledge base of defense systems curated through thorough literature search that so far includes over 150 systems.

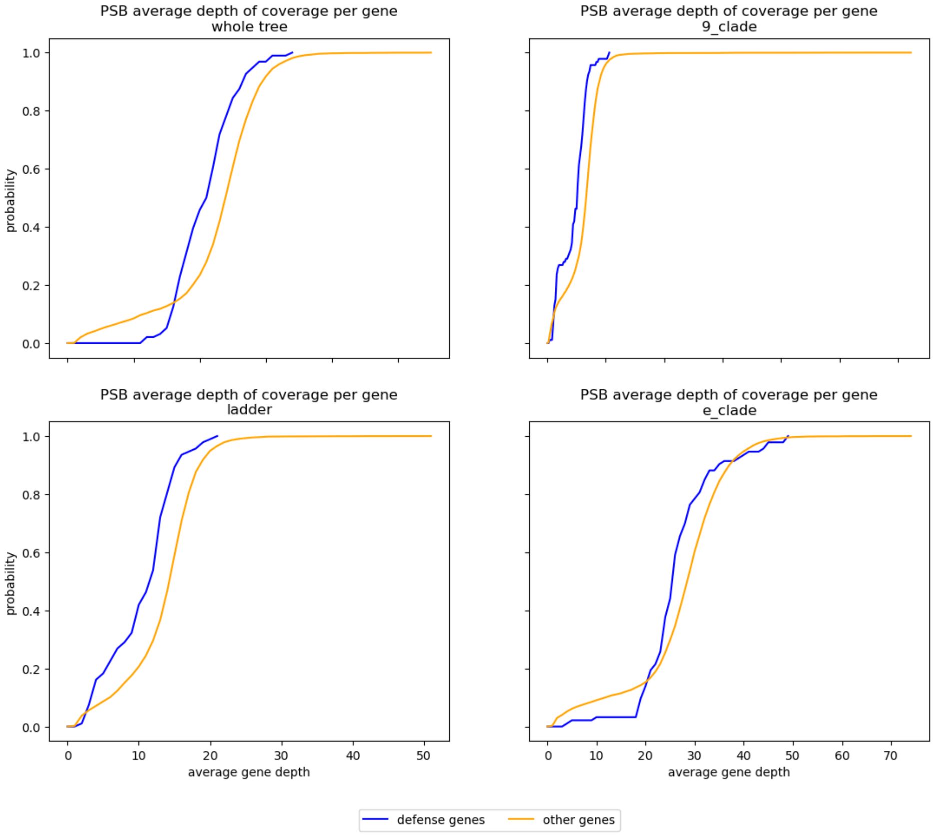

Figure 3.8: PSB defense genes vs other genes average depth of coverage per gene cumulative distribution function. Top left plot shows the distribution of coverages using all strains of the tree. The rest of the plots show the average coverage of depth per gene using strains from specific clades.

DefenseFinder found 25 regions on the PSB chromosome with different defense systems such as CRISPR, Septu, DarTG, DISARM, Restriction-Modification (RM) systems, among others. For SRB, DefenseFinder found less defense regions (14) which included RMs, CRISPR, BREX, PD-Lambda, and SoFic. After visualizing each region as a SNP matrix, we have noticed that for many ORFs there were certain clades that mostly showed

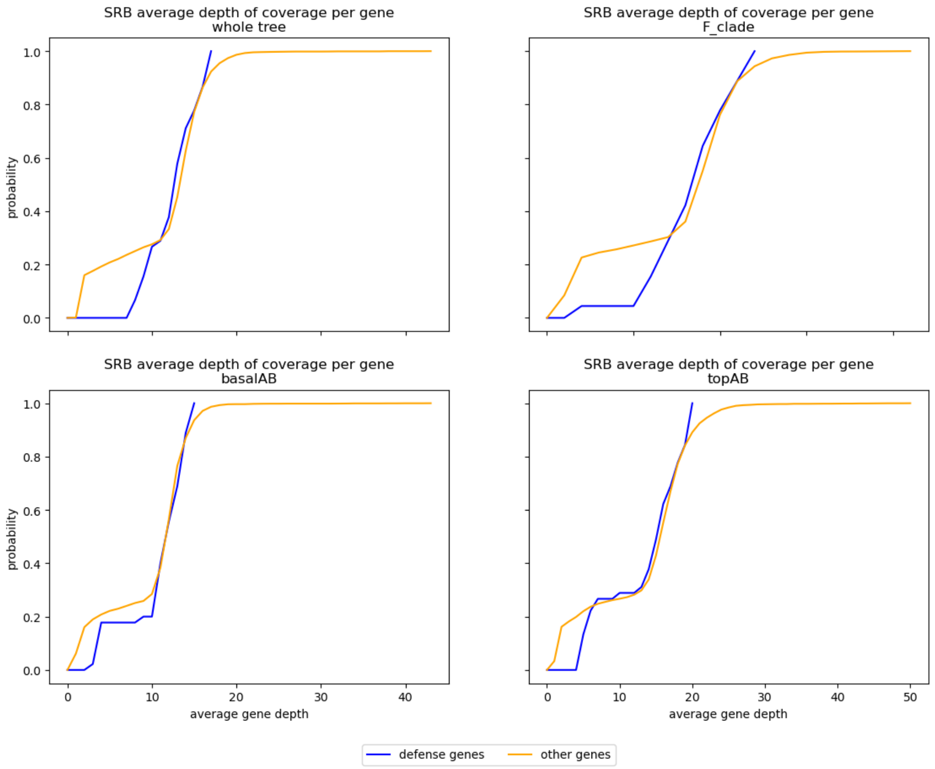

Figure 3.9: SRB defense genes vs other genes average depth of coverage per gene cumulative distribution function. Top left plot shows the distribution of coverages using all strains of the tree. The rest of the plots show the average coverage of depth per gene using strains from specific clades.

missing data. Intrigued, we plotted the average depth of coverage per gene as a cumulative distribution function (CDF) and compared it to the average depth of coverage of all the other genes (Fig. 3.8). Immediately, even when one compares the distributions calculated using all strains of the tree together, one notices differences in the distributions. For PRB, the defense genes have lower depth of coverage (for example, 50% of the defense genes have depth of coverage less than 18 while 50% of the rest of the genes have depth of coverage of less than 22). The Kolmogorov-Smirnov two sample test also enforces this conclusion, showing that those two distributions are statistically different (statistic=0.3, P-value=4.681e-8). Similar observations were made when examining the average depth of coverage for individual clades. The CDF of defense genes in the 9 clade, PSB mixing layer, and E clade show that defense genes have lower depth of coverage when compared to the rest of the genes. This observation is supported again by the Kolmogorov-Smirnov test with P-values of 2.336e-12, 8.881e-11, and 1.674e-5 respectively.

A similar analysis for the set of defense genes in the SRB genomes shows that the defense genes do not have lower depth of coverage compared to the rest of the genes. The CDFs of defense genes versus other genes do differ (also confirmed by the Kolmogorov-Smirnov test), however, in actuality, the difference lies within the very low coverage regions for which the non-defense genes are more prevalent.

The difference in behavior of the defense genes between PSB and SRB is quite fascinating. It has been shown in other studies that different prokaryotes contain different numbers of defense systems (varying between 0 and over 50). Additionally, the number of systems seems to depend on many different factors such as the size of the genome or the number of prophages present in the bacterial genome. In our case, it is apparent that the size of the genome positively correlates with the number of defense systems since PSB's genome is twice as big as SRB's and PSB contains roughly twice as many defense regions. However, we have not seen a positive correlation between the number of prophages and the defense systems. In order to look for prophage traces in the genomes, we have utilized PHASTEST - a web tool designed for rapid identification and visualization of prophage sequences within bacterial genomes. PHASTEST found a single phage region in the PSB genome (Enterobacter phage Arya of length 17.6 Kb) while the SRB genome contained two regions (Enterobacter phage iAA91 of length 25 Kb and Escherichia phage SH2026Stx1 of length 18.5 Kb).

3.3 Future Directions

The data presented above is only preliminary and further data analysis and experiments are necessary. The presence of large gaps in the alignment of the C-terminus region of the CDI proteins shows us that there is a possibility that the bacteria differ in their repertoire of toxins depending on their geographical location. However, we cannot state for sure what is happening at those low coverage regions without additional data, further computational analysis, and verification using live cells. There is a possibility of this region being simply deleted. In order to determine what exactly is happening, further steps need to take place.

First, a good starting point for determining what is happening at the C-terminus of different clades would be to take a look at the raw sequencing data. The data used for this analysis has been sequenced using Illumina technology. This means that the reads used for aligning are less than 500 bp in length. Alas, this makes it impossible to track back the raw reads of the CDI genes since the genes' lengths range from 4,000 bp to 18,000 bp. Therefore, the best course of action would be to look into long-read sequencing data which is capable of sequencing full genes in a continuous fashion. This way we could identify the CDI genes by their N-terminus part and take a close look at the variations in the C-terminus. Furthermore, we could take the sequences of the toxins and determine their structure and function. There are many different types of toxins that can be used by the CDI system such as adhesins, iron-scavenger proteins, lipases, pore-forming toxins, nucleases, and RNases[61, 108]. Most of the toxins that have been studied are nucleases, however, they do show a lot of variation in terms of cleavage sites, or the cofactors they need in order to activate[109, 110]. RNases are less common, mostly studied in Yersinia Kristensenii[111]. It would be fascinating to check how the toxins change depending on the geographical location the berry was collected from.

There is a wide range of directions one can take after that. One inspiration comes from the research done by Jackson et al. With a larger dataset, consisting of long-read sequences, one could perform a comprehensive phylogenetic analysis of the community. The CDI genes could be placed on a phylogenetic tree based on their sequence similarity. If the strains contain varying C-terminus sequences, they will most likely assemble on a tree as separate clades. One could then compare and contrast those clades[58].

In another example, one could study the set of toxins and immunity genes pairs seen for each CDI gene. Similarly to research conducted by Koskiniemi et al. After determining where the immunity genes are, a set of plasmids with different sets of immunity genes and toxins could be made and expressed in E.coli. The cultures can be then plated and incubated. This would be a simple way of determining which colonies survive, therefore, which immunity genes protect against which toxins[59].

Lastly, since most of the toxins have been identified as nucleases, one could also test this with this set of toxins, as seen in Koskiniemi et al[59]. Each toxin could be expressed in an E.coli and stained with DAPI (4',6-diamidino-2-phenylindole). This can be done with and without the immunity genes. By staining the cells using DAPI we will be able to visualize the DNA in the cells and check if it is degraded.

The functional and structural diversity of the bacterial CDI effectors is still a very active area of research. With so many types of structures that inject the toxins as well as many diverse types of toxins and immunity genes, it is a gold mine for learning about how bacteria fight the competition in a natural environment. It would not be surprising if additional types of systems are identified in the future. Of course, this knowledge can also be applied to biotechnology as well as health care. It is a very interesting option for dealing with antibiotic resistance. Studies have been conducted that explore the usefulness of CDI systems as an alternative method for eliminating infections. For example, it has been shown that T4SS is used by pathogenic bacteria for exchanging plasmids and contributing to the spread of the antimicrobial resistance. Therefore, efforts have been put into engineering inhibitors for the T4SS machinery[70, 71, 72, 73, 74]. Another example would be its potential use for delivering CRISPR-Cas9 into cells that have a strain-specific antibacterial activity[112, 113, 114].

3.4 Methods

3.4.1 Code availability

All scripts used for the analysis of the data can be found on github at github.com/adamadejska/pink_berry_scripts

3.4.2 The Dataset

The sequencing and assembly was performed by the Cordero and Wilbanks labs. Each berry was sequenced using metagenomic shotgun sequencing. The raw reads were trimmed, filtered, corrected for errors, and the adapters were cut off. The co-assembly of the prepared reads into contigs was performed using MEGAHIT software and contigs longer than 1 kb were chosen for binning. The sequenced reads were then mapped back to the constructed contigs using minimap2. In order to create the metagenomically-assembled genomes (MAGs), multiple different tools were used to bin the contigs to reduce potential bias. Lastly, the quality of the MAGs was assessed using CheckM software and each MAG was taxonomically classified using GTDB toolkit which identified each MAG using single-copy marker genes from the Genome Taxonomy Database. Two most abundant species present in the aggregates were Thiohalocapsa (sulfide-oxidizing purple sulfur bacteria (PSB)) and Desulfofustis (sulfate reducing bacteria (SRB)). Since the Thiohalocapsa MAG was very similar to the PB-PSB1 reference (assembled with long read PacBio data by Wilbanks lab), the reference was used as a baseline instead of the contigs created by the MAG. Additionally the single nucleotide variants were called by mapping both PSB and SRB to reference genomes. The SNVs were saved as a variant calling file (vcf) and shared with us for further analysis.

3.4.3 Creation of the dataset

For the PSB vcf, the strains have been sorted by the number of loci they have covered and out of 192 strains, strains with less than 80% of their biallelic bases called have been discarded, leaving 142 strains for further analysis. Only proper biallelic single nucleotide polymorphisms have been kept which we have assumed correspond to single mutational events in the history of its genomic locus. The SNP data has been labeled one of three different categories: wild type, mutant, or 'NaN' by majority calling. 'NaN' was given to any locus of a particular strain that did not have sufficient coverage to be informative (coverage less than 3 reads). The PSB data consists of one chromosome of length 7.9 Mb. There are 50,626 biallelic SNPs.

3.4.4 Creation of trees

A tree of the PSB strains was created based on the whole SNPs dataset (excluding singletons - genetic changes that happened in only one strain for a particular position in the genome). Additionally, we excluded any strains whose SNP matrix sequences contained more than 30% missing values. Since our dataset contains missing data, we calculate the distances between strains probabilistically using

$$D = \frac{\sum([v1_i - v2_i]^2)}{len(v)} \tag{3.1}$$

Where v1 and v2 are sequences of two strains (with mutations represented as 1, WT represented as 0, and missing sites (NaNs) represented as 0.5). The distances were calculated for each pair of strains, creating a triangular matrix of values that was then used to build a tree. To infer information about the structure of the population, a default MatLAB neighbor joining algorithm was used to build the actual relations. The trees were rooted to an "outgroup" that was entirely wild type.

3.4.5 Creation of the heatmap

The heatmap correlation figure was calculated using all non-singleton alleles and all strains present in the PSB tree. For each unique pair of strains, we counted the number of mutations they share. Any positions where both strains are wild type or where at least one strain contains missing information were ignored. The number of shared mutations was divided by the total number of acceptable sites and multiplied by 100. This information was then plotted as a heatmap using the seaborn python package.

3.4.6 Analysis of the CDI genes

We have explored the number of mutations per berry per ORF by taking all biallelic PSB alleles and counting the number of those alleles that are within the specified window (the start and end positions of each ORF). We performed this calculation for each sample separately and then plotted berries that had at least one mutated allele per ORF as a color coded bar graph where each color represents a corresponding geographical location of the sample. The depth of coverage analysis was performed by utilizing the depth of coverage information from the original bcf file. The depth of coverage was extracted for each position of the ORF in question separately for each group (clade vs the rest of the tree or each clade separately). The average coverage was calculated for each position in the ORF and that information was plotted.

3.4.7 Analysis of genes involved in phage defense

The defense genes have been identified using the DefenseFinder web tool by uploading the PSB and SRB DNA sequences. The ORFs that were identified as defense genes has been used to calculate the average depth of coverage per game. The rest of the genes listed in the PSB and SRB gff files have been used to calculate the average depth of coverage per "other" gene. The cumulative distribution functions were plotted using custom scripts.

3.5 Permissions and Attributions

The content of Chapter 3 and is the result of a collaboration with Otto X. Cordero at the Massachusetts Institute of Technology and Elizabeth G. Wilbanks, at the University of California, Santa Barbara.

3.6 Supplementary material

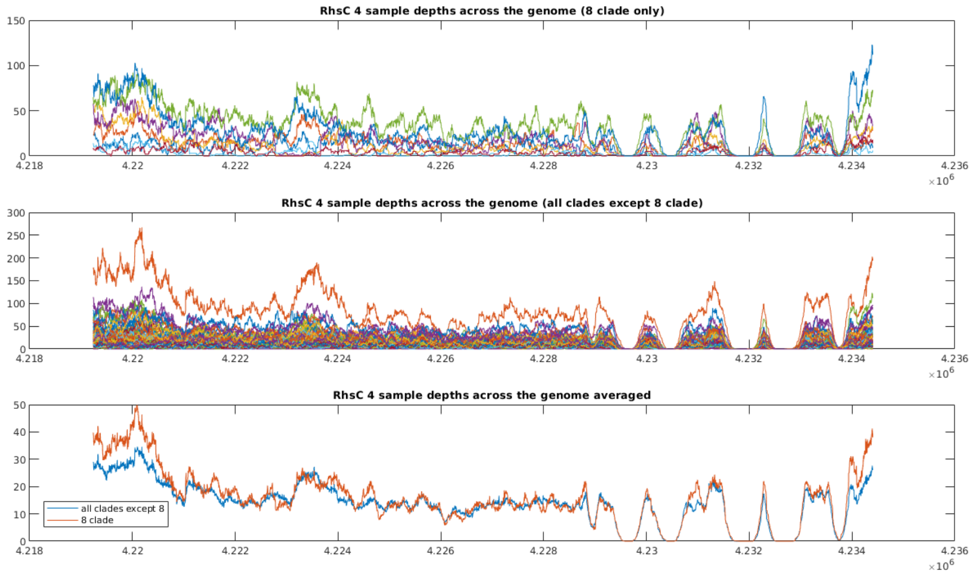

Figure 3.10: Depth of coverage for open reading frame of RhsC 4 gene. Top panel shows individual depth of coverage for each strain in the E clade. Middle panel shows individual depth of coverage for each strain in the PRB tree (except for the E clade). Bottom panel shows the average depth of coverage for E clade strains (orange line) and all other strains (blue line).

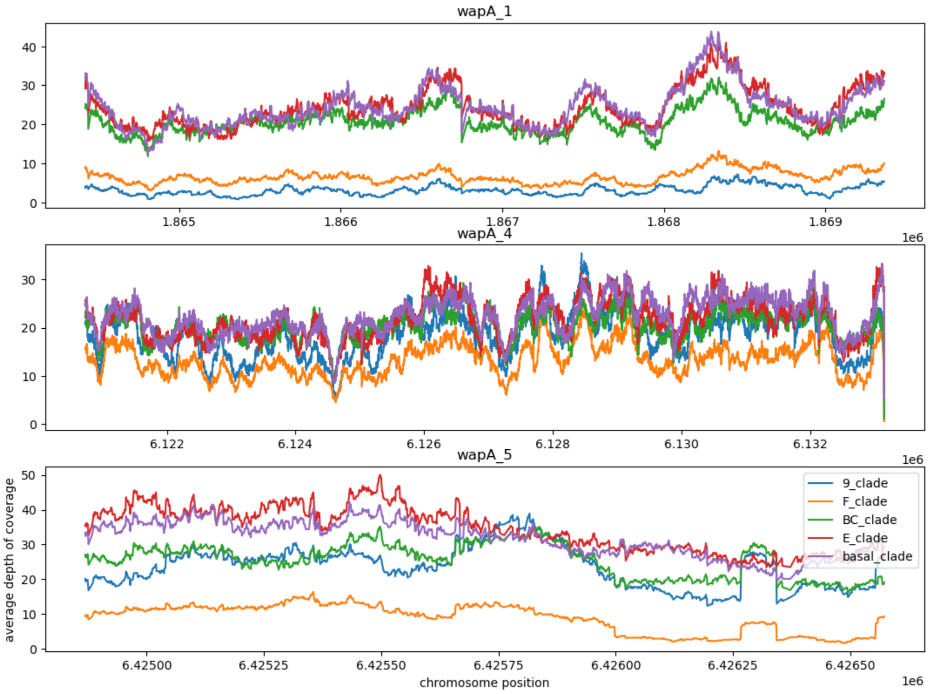

Figure 3.11: Depth of coverage per clade for the rest of WapA ORFs. Top panel shows average depth of coverage for each clade for the WapA 1 ORF. Middle panel shows average depth of coverage for each clade for the WapA 4 ORF. Bottom panel shows average depth of coverage for each clade for the WapA 5 ORF.

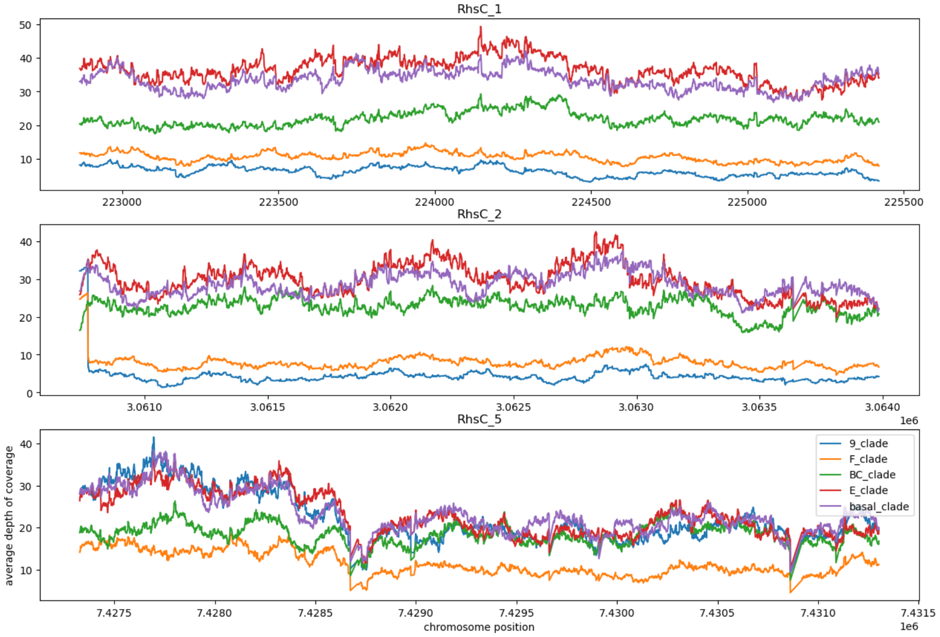

Figure 3.12: Depth of coverage per clade for the rest of RhsC ORFs. Top panel shows average depth of coverage for each clade for the RhsC 1 ORF. Middle panel shows average depth of coverage for each clade for the RhsC 2 ORF. Bottom panel shows average depth of coverage for each clade for the RhsC 5 ORF.

Chapter 4

Algorithm for de-noising Nanopore 16S sequences

4.1 Introduction

Organisms create communities containing a variety of species in order to survive and thrive. For example, the human gut microbiome is a diverse microbial community that changes its composition based on the varying environmental factors [22]. It has been shown that the composition of its community changes depending on the nutrients it has access to (via diet [115]) or the presence or absence of other organisms [116, 117]. Another example of a complex microbial community is the stratified structure of a microbial mat where microbes live in different layers of the mat based on their metabolic requirements [118]. Many studies focused on how different environmental factors change the composition of the microbial mats community. For instance, Bordenave et al. showed that the changes in the community composition depend on the different time periods following a petroleum contamination [119].

All of those different types of communities constantly change the composition of the species involved. Those adjustments are crucial for responding to constantly varying environmental conditions and maintaining their diversity in order to survive. In bacterial communities, their taxonomic and metabolic diversity is constrained by the availability of resources, such as oxygen or light, as well as environmental factors, such as UV radiation and temperature among many others [120, 121]. Those varying conditions have a profound effect on the community, such as the change in their metabolic rates or the strength of interactions between members of the population [122, 123, 124]. To assess such diversity, a substantial amount of research has been done on surveying the naturally occurring microbial communities that inhabit different environments, such as intertidal microbial mats [125, 126, 127, 128]. Additionally, this approach can also be applied to clinical research [75, 129]. For example, the loss of microbial diversity in the gut has been linked to susceptibility to diseases [23] and the changes to the lung microbiome have been linked to respiratory tract infections [24].

In order to study a community, one needs to first determine the composition of the community in question. What are the most abundant species? Does one species dominate the space or are all species more evenly distributed? These and many other questions are answered through a field called metagenomics which focuses on studying the structure and function of DNA extracted from sample of mixed (often unculturable) organisms. High-throughput sequencing methods allows us to identify bacteria and even recover their whole genome sequences [130]. However, if the sole purpose of the study is to identify the taxonomic profile of a sample, one would only need to use the sequence of bacteria's 16S ribosomal RNA (rRNA) gene [75]. The 16S rRNA is a component of the 30S subunit of the bacterial ribosome which is essential for reading the messenger RNA (mRNA) sequence and translating it into a string of amino acids which later get folded into a functional protein. The 16S rRNA sequence is about 1,550 bp long and consists of nine variable regions which are separated by highly conserved regions [76, 75]. The sequence of those variable regions is used to distinguish between different taxa and allows for a more precise identification, compared to more phenotypic methods, of even understudied strains [75]. It has been utilized for identification for clinical purposes and many more [77, 78, 79, 80, 81, 82].

There are many sequencing technologies available for geneticists, such as Sanger sequencing, Illumina, and Nanopore. Illumina sequencing, known as the second generation sequencing, is still the most popular sequencing platform. It has a relatively high accuracy but it is limited by the length of each sequenced read - capped at about 500 bp. This means that when sequencing the whole genome or even just single genes, the sequenced pieces need to be assembled back together into longer sequences based on the similarities between overlapping segments. Of course, such a feat is not a trivial task and many alignment algorithms have been developed and are under constant development, demanding faster and more accurate software. On the other hand, the latest sequencing technology, long-read sequencing can create lower accuracy but much longer consecutive stretches of DNA sequence. There are two main companies that specialize in this type of sequencing - Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). PacBio uses light data from a DNA polymerase which is measured in real time and used to determine the given sequence. On the other hand, Nanopore uses data of changes in an electric current detected when a DNA sequence passes through a designed pore [83, 84]. However, both PacBio and Nanopore have their own drawbacks, mainly the lower throughput and higher error rates (10-15% compared to Illumina's 0.1%) [85]. Longer reads deliver hope of making genome alignments easier and faster solving one of the impossible-to-solve problems using Illumina sequencing - the alignment of tandem repeats [131]. Nevertheless, the unrestricted read length of long-read sequencing is an attractive alternative for sequencing genes such as 16S rRNA.

Since the popularization of long-read sequencing for 16S rRNA profiling, multiple pipelines have been published that attempt to reduce the sequencing error and identify the species present in the given experimental dataset. The first one, published in 2021, was NanoClust [132] which uses clustering to identify species. The second one was Emu [76], published in 2022, which uses an expectation-maximization algorithm to identify taxa and generate taxonomic abundance profiles. Both pipelines have their own drawbacks. One of the biggest drawbacks of NanoClust is its use of HDBscan for clustering and the assumption that reads from different species are divergent enough that one species will cluster far away from each other. As we will see, often an environmental sample containing a variety of species will contain clusters that overlap, especially if reads come from the same genus. On the other hand, the main drawback of Emu is its reliance on a database. If a species is not included in a specified database, Emu will only be able to find the closest neighbor that is present in the database or label such read as "unclassified"

In this chapter I present a different approach to the problem that gives users working with 16S Nanopore sequences more options for identifying their samples. We demonstrate the accuracy of the pipeline using simulated PacBio and Nanopore datasets from CAMI2 and then use it on experimental sediment data.

4.2 Results

4.2.1 The Experimental Dataset

My analysis will use sequencing data from Alex Petroff's lab at Clark University. Samples of water-saturated sediment were collected in Carpinteria, CA, and cultured in a thin film between glass slides in order to create a quasi-2D system. The slides were cultured in a 12-hour day/night cycle in an attempt to establish a stable carbon cycle. The closed ecosystems were also cultivated under different concentrations of dextrose and peptone. The oxygen production was measured via fluorescence and it has been observed that, at first, the oxygen concentration decreased but after a few hours it relaxed into stable metabolic dynamics. Each community was sequenced either when its oxygen concentration initially decreased or when it stabilized. The description of each experiment can be found in Table 4.1.

Each community from the experiment described above was sequenced using 16S ONT sequencing. The reads were demultiplexed and partitioned into separate files according to the sequenced barcode (barcode 1 - 24). Based on the notes on each experiment, only 14 experiments were performed. The larger amount of barcodes comes from the fact that we were not able to tell exactly which experiment they came from. Therefore, new barcodes were called and those unidentified reads were collected there.

After a careful look at each of our files, it is apparent that the number of 16S reads vastly differ between each file, ranging from 40 to 1,000,000 sequences (Fig. 4.1). Additionally, the usual sequence length of 16S is around 1,500 bp, however, the sequences in each file span a large range of lengths (starting from 100 bp to almost 5,000 bp, with peaks spaced by 1.5 kbp) which hinders successful alignment (especially for the shorter reads) (Fig. 4.2).

4.2.2 Algorithm overview

In this section we will briefly explain the motivation behind the approach as well as some of the details on how the pipeline works. For those interested, the pipeline is available on github. The approach is based on the notion that Nanopore sequences that are similar to each other are most likely from the same species or genus of bacteria. Once

Barcode	Experiment type	Description
1	Flow-through column	DNA from the effluent. 5 mM dextrose + peptone in the medium flowing through the chamber. Started from community in barcode 8.
2	Flow-through column	DNA from sediment in the chamber. 5 mM dextrose + peptone in the medium flowing through the chamber. Started from community in barcode 8.
3	Thin mat	0.5 mM dextrose + peptone. Started from community B (not sequenced).
4	Thin mat	5 mM dextrose + peptone. Started from community B (not sequenced).
5	Thin mat	1.58 mM dextrose + peptone. Started from community B (not sequenced).
6	Thin mat	Material from stock mats
7	Thin mat	1.58 mM dextrose + peptone.Started from community in barcode 6.
8	Thin mat	natural sediment that was inoculated into the flow-through chamber and thin mat chamber.
9	Thin mat	0.5 mM dextrose + peptone. DNA extracted at rebound. Started from community in barcode 8.
10	Thin mat	50 mM dextrose + peptone. DNA extracted at oxygen minimum. Started from community A (not sequenced).
11	Thin mat	No sugar or peptone. DNA extracted at oxygen minimum. Started from community in barcode 8.
12	Thin mat	No sugar or peptone. DNA extracted at oxygen rebound. Started from community in barcode 8.
13	Thin mat	1.58 mM peptone, no dextrose. Started from community in barcode 6.
14	Thin mat	0.5 mM dextrose + peptone. DNA extracted at oxygen minimum.Started from community A (not sequenced).

Table 4.1: Metadata on the type of experiment and the conditions at which the samples were collected.

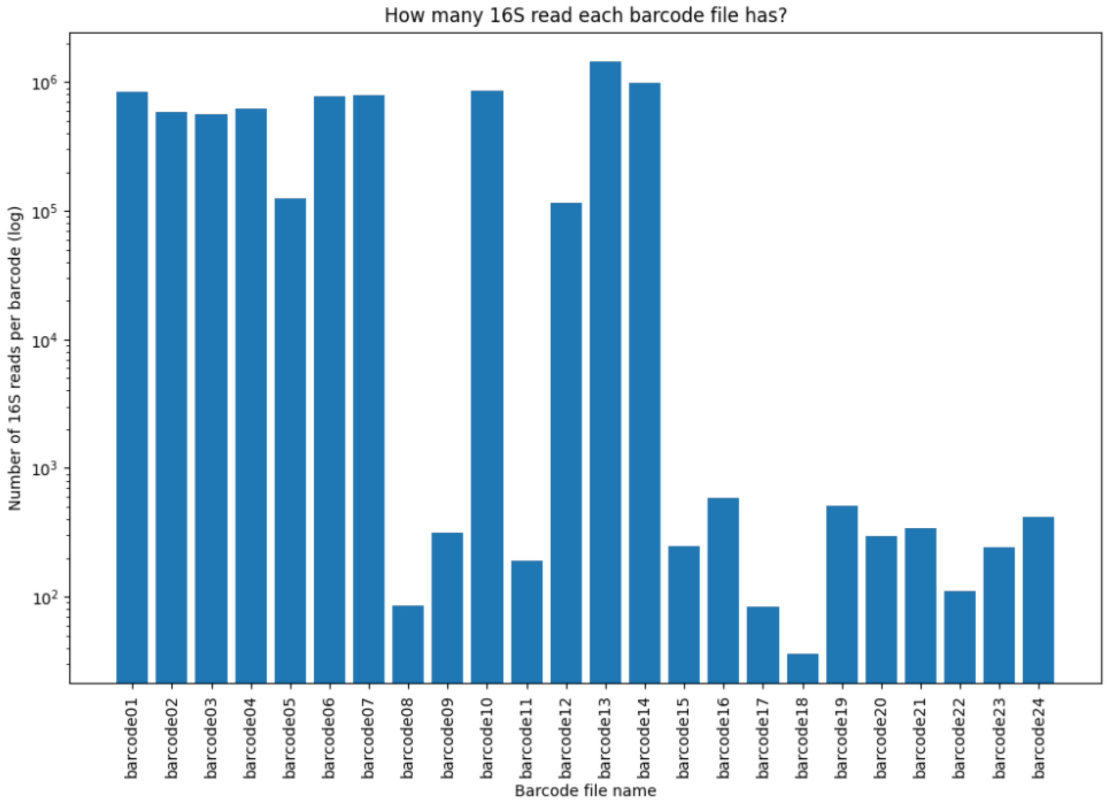

Figure 4.1: Bar graph of the number of reads per barcode showing high variability between barcodes.

we create clusters of very similar sequences, we align them together and create a consensus of each cluster, which ultimately lowers the error caused by long read sequencing. In other words, each sequence alone has a high error rate caused by not only single nucleotide changes but also short insertions and deletions. However, each Nanopore read has a unique set of errors, meaning, once we group similar sequences together we should be able to correct them based on the majority alleles.

In the world of machine learning, the task of figuring out which pieces of text are similar to each other is a well researched topic [133, 134, 135]. Most methods focus on first vectorizing the string (representing a sentence or a word as a vector of numbers unique to that string rather than a sequence of letters) and then calculating similarity

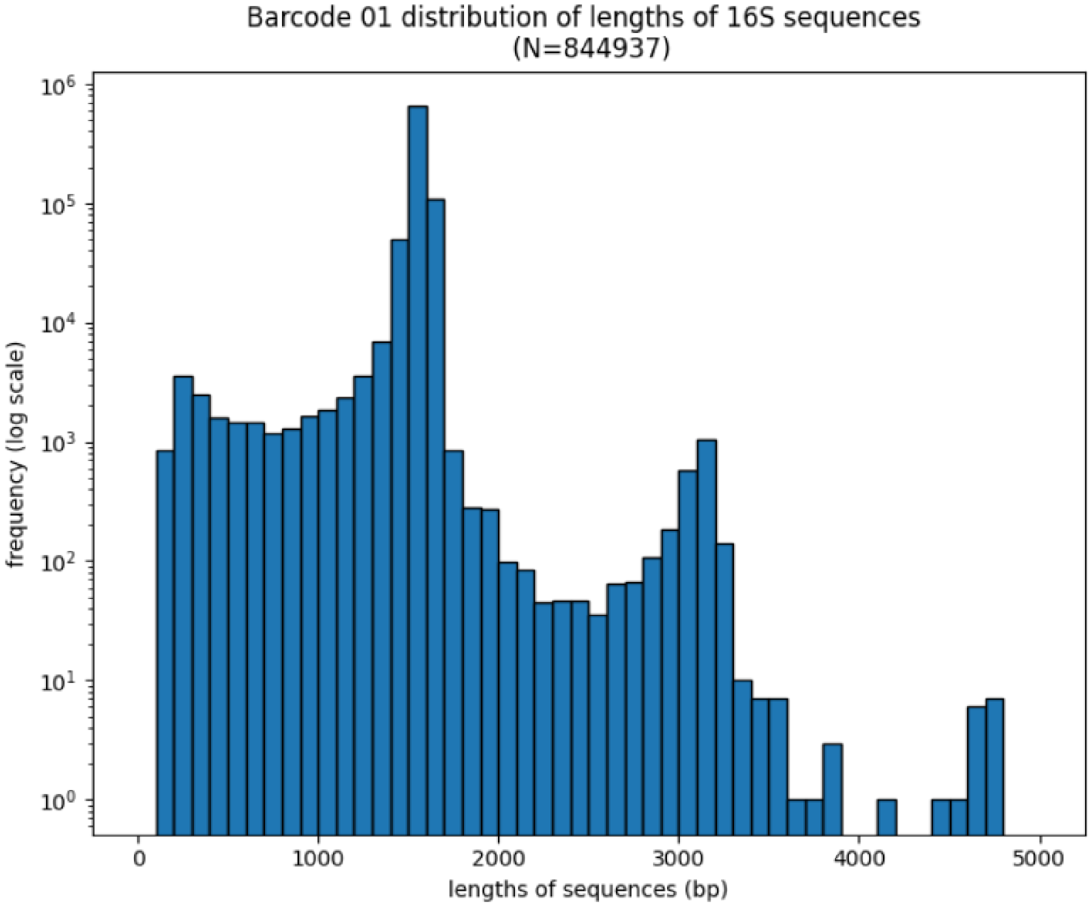

Figure 4.2: Bar graph of the number of reads per per length in barcode 01. The length of reads varies between 100 and 5000 base pairs. Note that the secondary peak at length = 3,000 bp is a concatenation of two 16S sequences (2*1,500).

based on those multidimensional vectors. Some of the more basic approaches include calculating cosine similarity [136](aka. figuring out if the two vectors in question point in a similar direction in their multidimensional space), Euclidean distance (the length of the line between two points in a multidimensional space) or Manhattan distance [137] (where the distance is defined by the sum of the absolute differences between two points (traveling along a grid rather than “as the crow flies’)).

In our algorithms, we have decided to calculate the kmer frequency matrix for each sequence in order to vectorize them. A vector of kmer frequencies is a simple way of

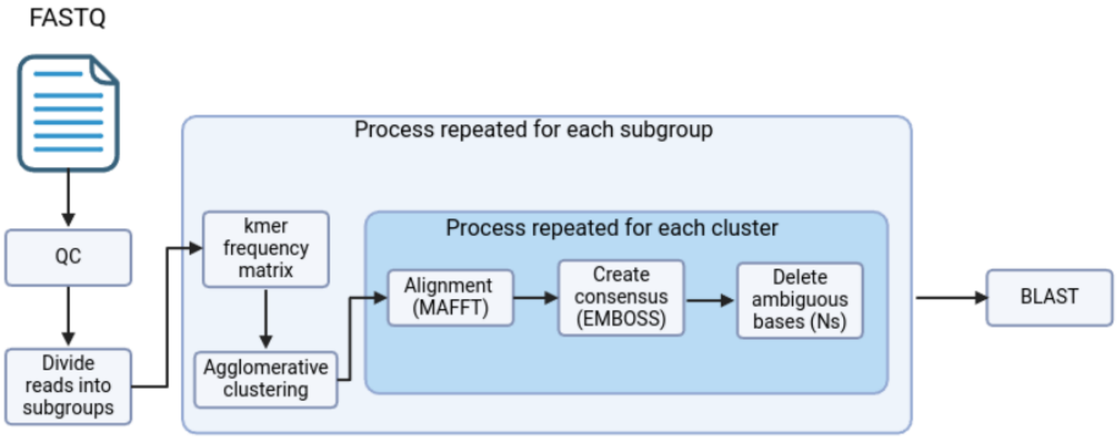
FASTQ
QC
Divide reads into subgroups
Process repeated for each subgroup
kmer frequency matrix
Agglomerative clustering
Process repeated for each cluster
Alignment (MAFFT)
Create consensus (EMBOSS)
Delete ambiguous bases (Ns)
BLAST

Figure 4.3: The schematic of the flow of the pipeline showing the processor repeated for each cluster (dark blue area) and the processes repeated for each thread (light blue area).

representing a string in a multidimensional space. A kmer is a word that consists of k characters. In our case, the only letters that a kmer can consist of are A, T, C, and G - the DNA bases. Therefore, a set of kmers where, for example, k is 2 is AT, TA, TC, CT, CA, etc. In our pipeline we use k=5 which gives us 1,024 unique kmers (4^5) and therefore 1,024 dimensions for each vectorized sequence. The kmers are counted using a sliding window, meaning, rather than dividing each sequence into segments of k nucleotides, the kmers overlap. This gives us more confidence in the similarities in case of short deletions or insertions.

Next, we used the information from the vectorized strings to cluster similar sequences. Again, many different clustering algorithms are available but not all of them will be useful for the task. The main purpose of the clustering algorithm is to make small clusters of very similar sequences. The emphasis is on the size of the clusters. If the clusters are too large, the alignment and consensus algorithms will not be able to agree on an alignment and the overwhelming amount of sequence data will result in the majority or all of the information being lost. Small clusters of 2 to 10 sequences have a better chance of preserving information and output longer consensus sequences (demonstrated in Fig. 4.4). During the testing process, we used mock datasets to ensure the accuracy of the pipeline (details below). We tested different cluster sizes to check what is the optimal number of sequences per cluster that retains most information. We plotted the distribution of lengths of consensus sequences as a function of the size of the cluster (Fig. 4.4). We observe a clear anti-correlation between the cluster size and consensus sequence length.

Because of the cluster size restrictions, we need a clustering algorithm that will allow for defining the size or number of clusters that need to be created. By that logic, clustering using HDBSCAN [138], affinity propagation [139], mean-shift[140], OPTICS, or BIRCH[141] will not work since no such parameter can be specified. The viable options at this point are k-means clustering[142] or agglomerative clustering[143]. Since agglomerative clustering is slightly better computationally with dealing with a very large number of clusters and a very large number of samples, it is the one that was chosen for this pipeline.

How does agglomerative clustering work? In short, agglomerative clustering is a type of hierarchical clustering which builds clusters based on a tree structure. Each sample is a leaf on a tree and the tree branches connect each leaf to another based on their similarity. This branching is repeated until all branches are connected to a single root. Agglomerative clustering recursively merges pairs based on linkage distance. It creates clusters based on the number of clusters specified as well as the specified linkage criterion (which tells it what kind of distance metric to use between sets of observations). The most popular is ward linkage which uses Euclidean distance to minimize the variance of the clusters being merged. In order to create many small clusters, the number of clusters need to be set to at least half of the number of samples (i.e. if we want to cluster 1,000 sequences, the number of clusters specified need to be at least 500).

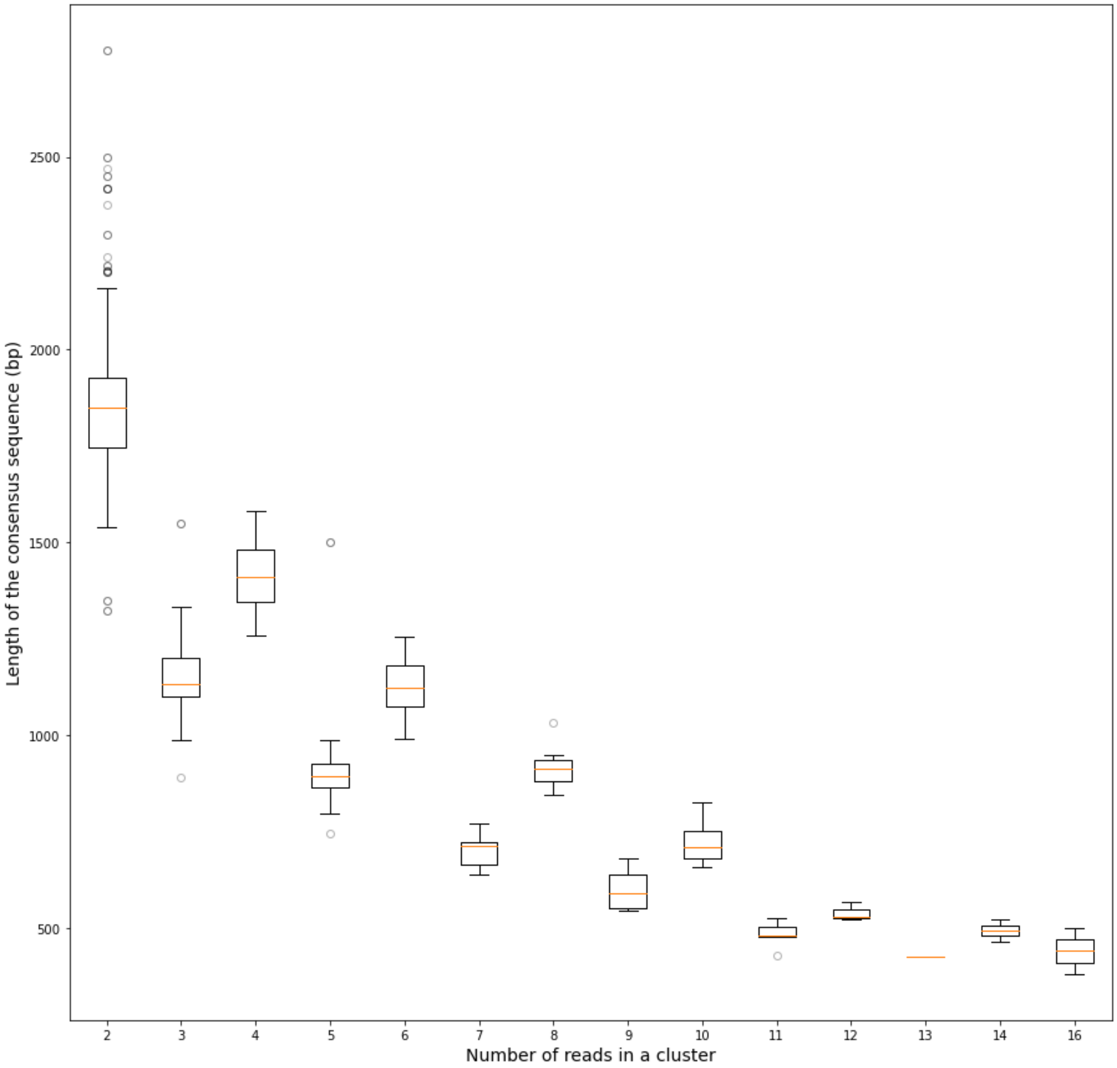

Figure 4.4: Box plot of the lengths of consensus sequences as a function of the size of the cluster. The larger the cluster, the shorter the consensus sequence.

After all the clusters are created and filtered based on their size thresholds, we use a set of off-the-shelf programs to create consensus sequences. We align the sequences using MAFFT[144] and then create consensus using EMBOSS[145]. We get rid of any ambiguous bases and use BLAST to find the most similar hits[146]. Lastly, we have used the up-to-date version of BLAST+ (v2.15) with a prokaryotic database (nt_prok) which includes a total of 8,520,872 bacterial sequences. The BLAST file is then parsed for clarity and for each consensus sequence, the first unambiguous hit is reported. The full flow of the pipeline can be seen below (Fig. 4.3).

4.2.3 Testing on simulated datasets

In order to test the accuracy of the pipeline, we used multiple simulated PacBio and Nanopore datasets from the Critical Assessment of Metagenome Interpretation's second round of challenges (CAMI2)[147]. CAMI2, hosted by Microbiome Community Of Special Interest (COSI), consists of a variety of different datasets of simulated PacBio and Nanopore reads. The sets include a marine dataset and a Toy Human Microbiome Project Dataset among many more. The datasets are available online at data.cami-challenge.org and the file of anonymous reads as well as the corresponding mapping file can be obtained in order to verify the results.

For the testing of the pipeline, we focused on three datasets - the marine dataset, airways, skin and urogenital tract dataset, as well as gastrointestinal tract and oral cavity dataset. Overall, each dataset had a wide distribution of read lengths which were filtered by our pipeline in order to retain only high quality reads (based on the similarity of the length to the 16S gene). Reads between the length of 1,300 bp and 1,700 bp were used in the analysis. The marine dataset has a very wide ensemble of species represented in each file. The frequencies of the reads for each species range from 1 to 30%. In the human microbiome datasets, the distribution of species is very skewed towards only a couple of species (usually taking up 40% or more of all reads).

Although we mostly care about identifying the most abundant species, we used the entire dataset regardless of how many reads there were per species. This way we are able to make an educated decision when analyzing the experimental dataset which hits are more likely to be true positives and we will have a greater understanding of how pure clusters are. This is especially important due to the fact that the data can include species of the same genus which have a very high level of similarity in their 16S sequences (s.a. *Aliivibrio salmonicida* and *Aliivibrio wodanis* which have 99% similarity for their 16S rRNA sequences).

The test was performed on all marine dataset files (10 files). Each file was run through the pipeline - the reads were first filtered based on their length, then they were clustered and small clusters of 2 to 10 reads each were passed to the alignment and consensus defining step. Lastly the reads were BLASTed and the first unambiguous hit was recorded. We then compared our findings to the true identity of the filtered reads, noting the frequency of the true species as well as the true and false positive hits from the pipeline output. We have analyzed the results on two levels: a species level and genus level in order to assess the accuracy versus sensitivity of the results. First, the results of the frequency analysis on the species level can be seen in Figure 4.5. Overall frequency of the identified consensus sequences resembled the true frequency of the original subset. The figure shows a very clear positive correlation between true and experimental frequencies with a positive slope of 0.962.

How many of the reads we found are true positives and how many are false positives? About 80% of the hits in the output file were true positive hits on the species level and this behavior was consistent across all 10 files. The reads that were incorrectly identified usually had a smaller number of reads per hit (less than 20) . The true positive species usually created larger clusters of reads - each species contained between 10 and 110 reads. Based on those results, it is prominent that the more reads a species was identified by, the greater the chance it is a true positive hit, however, it would not be possible to completely discern which consensus sequences belong to false positive hits based on the number of consensus sequences per species because of the large number of the true positive hits that also consisted of a small number of hits. It could, however, be used as one of the criteria for the level of confidence that a species is a true positive hit.

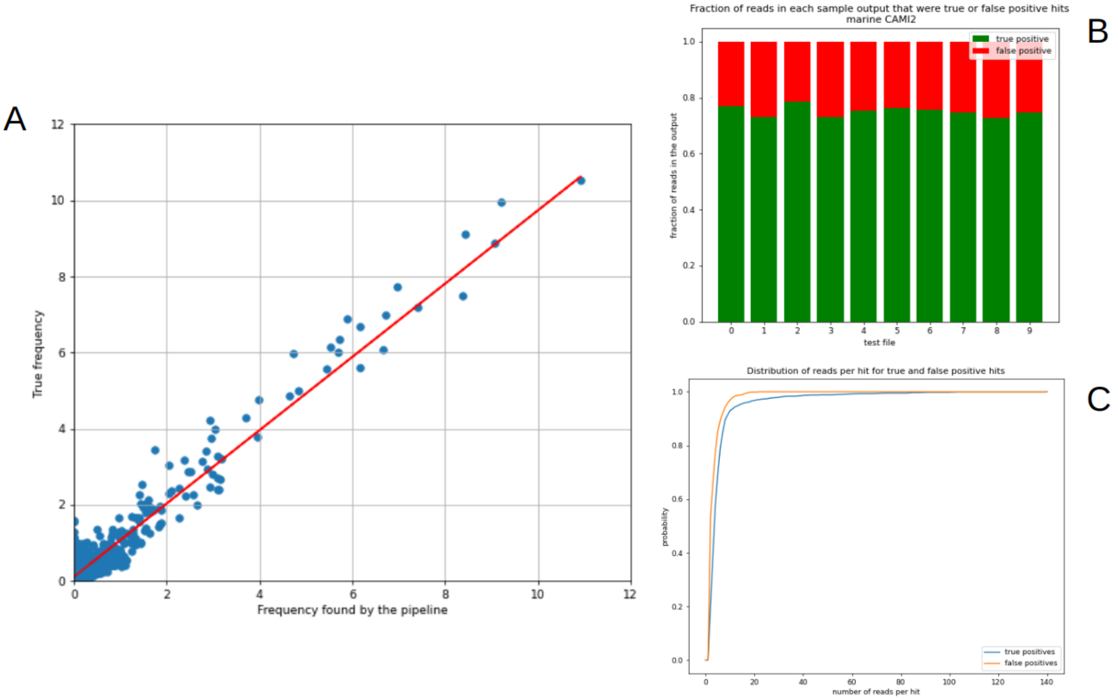

Figure 4.5: CAMI2 Marine simulated datasets: true frequency of found species vs frequency found by the pipeline. The true frequency of reads found in the original dataset are always very close to the frequency of reads found in the pipeline's output. The analysis of true vs false positive hits of the CAMI2 marine dataset. Fraction of reads in each sample output that were true or false positive hits (B). Distribution of reads per species for true and false positive hits (C).

Another approach for establishing a filter for the false positive hits could be the percent identity measure of the BLAST output. Identity percentage measures the percentage of the nucleotides that are the same between the two sequences. We wanted to check if the false positive hits could have an overall lower percent identity scores compared to the true positives. We have plotted the scores separately for false and true positive hits as a cumulative distribution function. This graph shows that although the curves are different from each other (Kolmogorov-Smirnov two sample test P-value of 2.034e-15) , the false positive hits can have percent identity scores over 90%. Therefore, similarly to the conclusions of the number of reads per hit, using percent identity will not be a plausible way of filtering false positive hits.

A quick look at the identity of the false positive hits showed us that most of them (between 60% to 75% depending on the file) although did not match any hit from the mapping file on a species level, they did have the same identity on the genus level. For example, a false positive hit *Vibrio atlanticus* shares the same genus with a true positive species *Vibrio anguillarum*. Of course, the trade-off of looking at the hits on the genus level instead of the species level is the loss of specificity. One needs to be aware of that and careful while interpreting the results. Nevertheless, when one assesses the results of the marine test datasets on the genus level, one can see a great improvement in the ratio of the true positive hits to the false positive hits (Fig. 4.6). Now over 90% of the reads agree with the mapping file and the majority of the false positive hits have less than 10 reads per genus. The true frequency of each genus in the original file versus the output file frequency per genus is also still very similar. The slope of the best fit line in that case is 1.044.

The same test was performed for the human microbiome dataset (air-skin-urogenital samples (6 files in total) and gastro-oral samples (6 files in total)). This dataset was much bigger than the marine samples and the simulated reads were only identifiable on the genus level rather than species level. As stated above, the majority of the reads represent bacteria from a single genus. For example, for the air-skin-urogenital sample

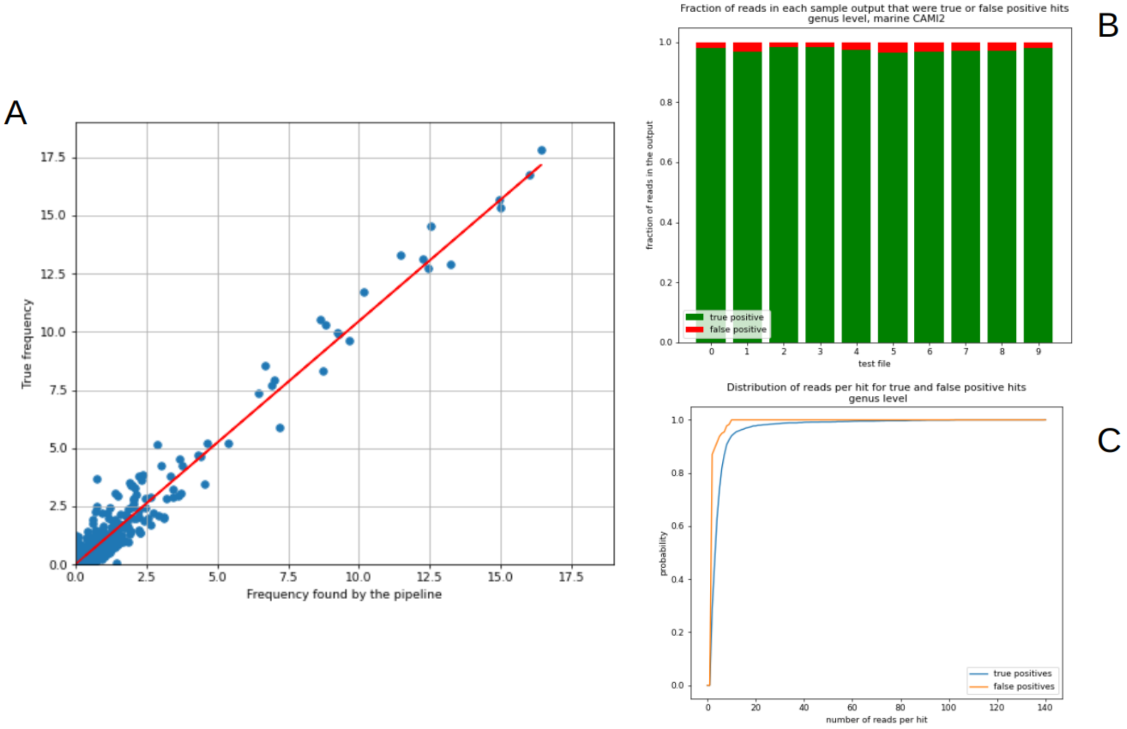

Figure 4.6: CAMI2 Marine simulated datasets: true frequency of found genera vs frequency found by the pipeline. The true frequency of reads found in the original dataset are always very close to the frequency of reads found in the pipeline's output. The analysis of true vs false positive hits of the CAMI2 marine dataset. Fraction of reads in each sample output that were true or false positive hits (B). Distribution of reads per genus for true and false positive hits (C).

file 0, the leading genera were *Pseudomonas* (45% of reads) and *Herbaspirillum* (42% of reads) while the rest were mostly from *Lactobacillus* and *Sphingobin*. In sample file 1, the majority of the reads (64.4%) belong to the genus Staphylococcus while the rest of the frequencies vary between 0.5 and 14% (Fig. 4.7 top). Similarly to the results from the marine dataset, we see that the sequences found by the pipeline have similar genus sequences to the frequency of the sequences in the original file. The slope of the best-fit line between true positive results from the pipeline and the original dataset shows a nearly 1:1 ratio with a slope of 1.0764 (Fig 4.8). It is worth noting, however, that there are a few outliers from the best-fit line. This might have been caused by the aforementioned

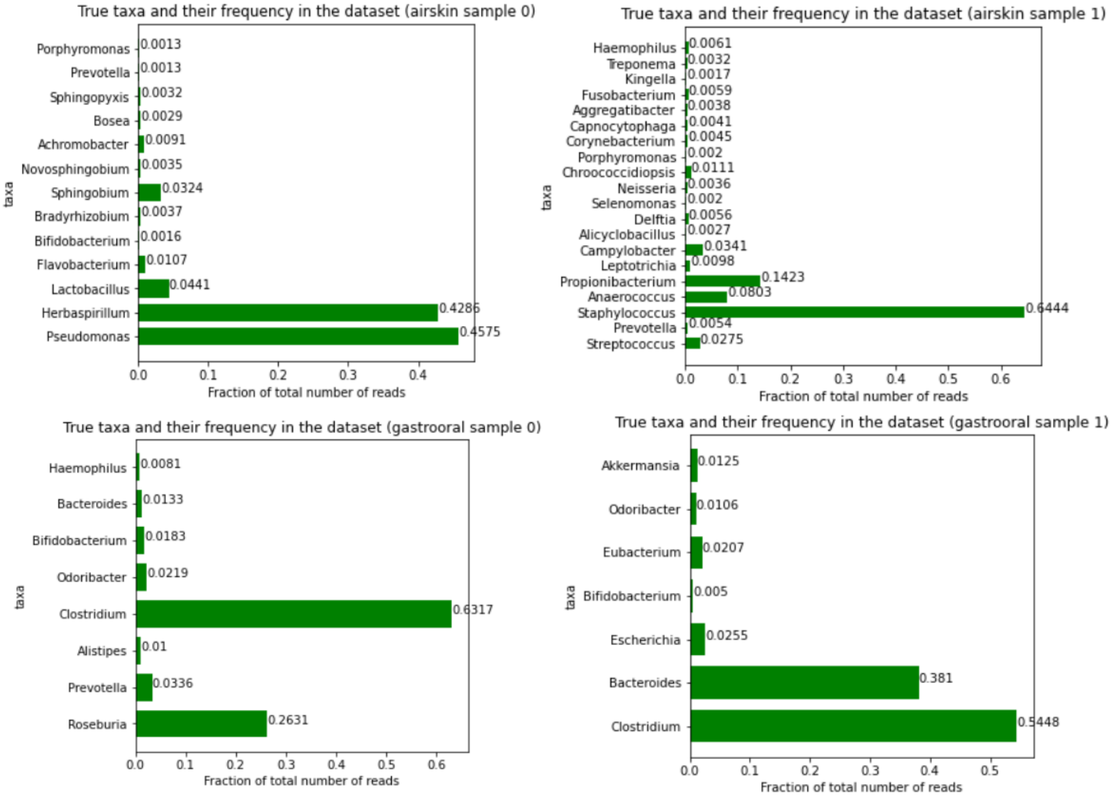

Figure 4.7: True read frequencies for species in the Human Microbiome dataset: air-skin-urogenital samples 0 and 1 (top) and gastro-oral samples 0 and 1 (bottom). In all samples only one or two species represent the majority of the reads. The rest of the species are represented by a very small number of reads.

inconsistent distribution of genera. Additionally, looking at the ratio of true positive and false positive hits in the pipeline output files, we see that in the majority (10 out of 12 files) about 90% of the consensus reads have a true positive identification. There are two files that performed much worse (with true and false positives having about a 50:50 ratio) for unknown reasons.

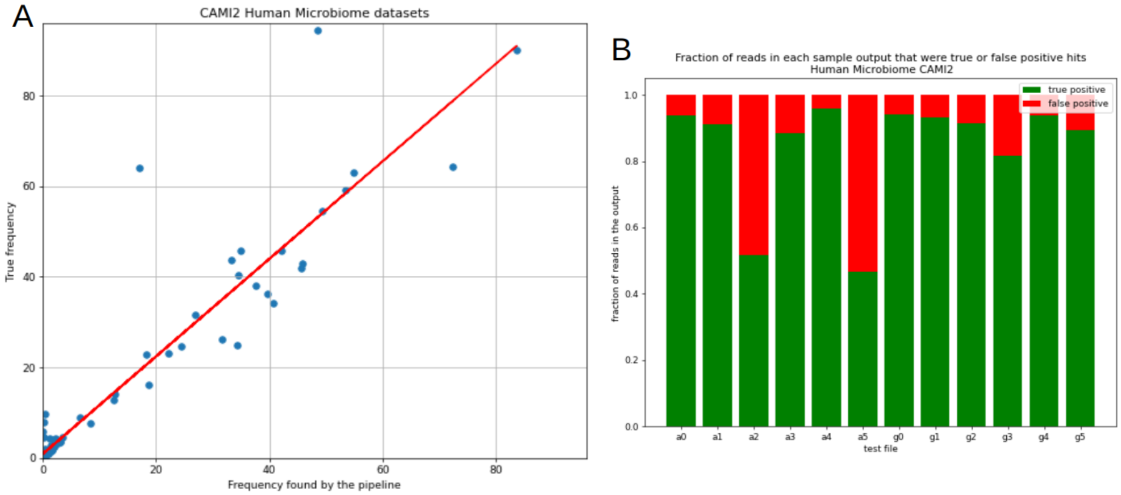

Figure 4.8: CAMI2 Human Microbiome simulated datasets: true frequency of found species vs frequency found by the pipeline. The true frequency of reads found in the original dataset are close to the frequency of reads found in the pipeline's output. The analysis of true vs false positive hits of the CAMI2 marine dataset. Fraction of reads in each sample output that were true or false positive hits (B)

4.2.4 The dynamics of taxonomic diversity in water-saturated sediment ecosystem

Given the results from the test datasets, we can now move on to analyzing the experimental datasets. As stated above, 14 FASTQ files, each corresponding to a different experiment, were run through the pipeline. The sequences were filtered based on their length, leaving only the reads with length between 1,500 bp and 1,700 bp, consensus sequences were made using small clusters, and the results were BLASTed against the nt_prok database. It is important to note that, for the sake of time, 5,000 random reads from each barcode were BLASTed instead of the full set of consensus sequences, however, the smaller random sample of reads does not change the overall frequency of particular species (Fig. 4.9). To ensure that, we created two sets of random consensus sequences per barcode and ran them through BLAST. Each time, we have observed that the frequen-

Barcode	Number of consensus sequences BLASTed	Number of reads with unambiguous hits
1	5,000	5,000
2	5,000	4,508
3	5,000	3,117
4	5,000	3,384
5	5,000	2,844
6	5,000	888
7	5,000	3,833
8	20	14
9	44	32
10	5,000	4,740
11	31	24
12	5,000	4,551
13	5,000	3,502
14	5,000	4,332

Table 4.2: Barcode files and the number of reads per barcode BLASTed and found with unambiguous hits.

cies of species were very similar to each other regardless of the random set of consensus sequences, proving reproducibility of the results. The number of reads BLASTed and the number of sequences with unambiguous hits are listed in Table 4.2. Based on the samples with sufficient number of reads (i.e. all excluding barcodes 8,9 and 11), we observe significant differences between different conditions in frequencies of the main bacterial species (Fig. 4.10). It is the most noticeable in barcodes 13 and 7 and barcodes 1 and 2. However, in its current state, it is impossible to draw any concrete conclusions regarding the environmental and metabolic correlates of the observed compositional variation. Future work should involve more conditions and temporal analysis of each experiment in order to identify systematic and reproducible changes in the community. The present analysis does however establish the feasibility and provides preliminary data for such a study.

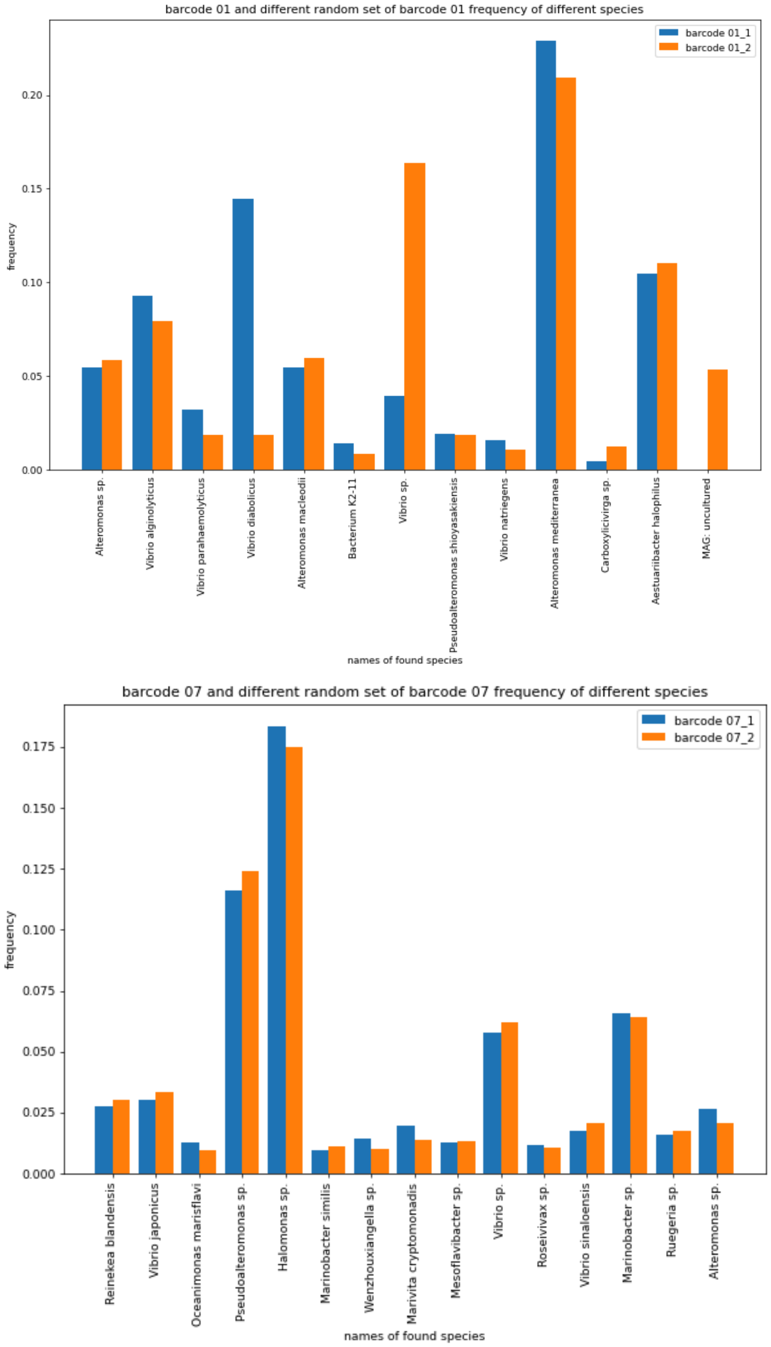

Figure 4.9: Frequency of species found in different random samples.

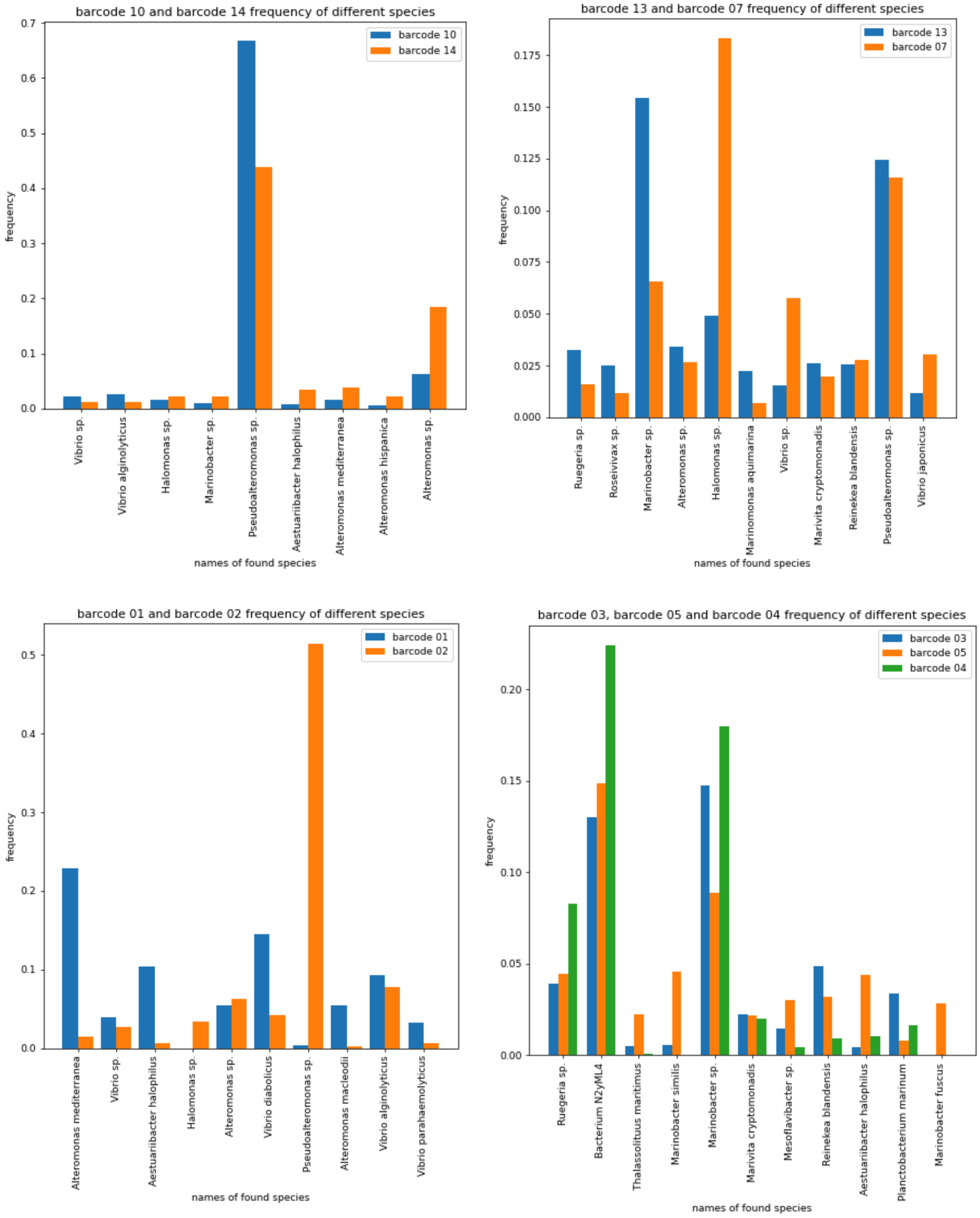

Figure 4.10: Frequency of species found in different experiments.

4.3 Discussion

In bacterial communities, their taxonomic diversity constantly changes based on the availability of resources, presence of absence of other organisms as well as environmental factors, such as UV radiation and temperature. Studies of how different factors change community composition are important, especially from the healthcare and biotechnology perspectives. Metagenomic sequencing approaches are often used to deterimne the taxonomy of the organisms in a sample, however, because of disadvantages of sequencing using short read sequencing and the high similarity of sequence between species, the results might be lacking details of the actual diversity of the sample. The approach using long read sequencing has become an intriguing new solution for identifying those kinds of species. Multiple pipelines have been developed over the last few years that address the issue of high-error long-read sequencing. However, each has their own disadvantages.

In this chapter we have introduced a new tool for decreasing the error of 16S Nanopore sequences and identifying the species present in each sample. This approach is fast and straightforward, making it a solution that can be utilized even by people not familiar with programming. It requires minimal set up and utilizes well-documented software. Moreover, in this analysis we have used the NCBI nt_prok database, however, the choice of the database can be changed based on the user's needs. We demonstrate the accuracy of the pipeline using simulated PacBio and Nanopore datasets from CAMI2 and then use it on experimental sediment data.

Using the mock datasets, we have showed that the pipeline performs well with balanced bacterial populations and identifies genera with the most reads in the sample when the organisms are unbalanced and dominated by one or two species. We also demonstrated that the frequency of species or genera found by the pipeline closely resembled the true frequency of the species or genera from the testing set. This way, one can not only identify species present in the sample but they will be able to determine how the organisms are distributed in the sample. Additionally to the testing datasets, we used the pipeline on our experimental reads from fourteen different experiments that aimed to address the question of how the community composition changes given different starting conditions? However, the results in their current state are inconclusive and further work should be done in order to gather more data across different conditions and time points.

There are still many ways this analysis can be improved and many additional questions it can answer. As stated above, studies have shown that the environment the community grows in has a big impact on its composition. The experimental setup described and analyzed above could be expanded to other environmental conditions such as change in salinity or temperature.

4.4 Methods

4.4.1 Collection of bacterial samples and experimental design

In the Petroff lab at Clark University, samples of water-saturated sediment were collected in Carpinteria, CA, and cultured in a thin film between glass slides in order to create a quasi-2D system. The slides were cultured in a 12-hour day/night cycle in hopes of creating a stable carbon cycle. The closed ecosystems were also cultivated under different concentrations of dextrose and peptone. The oxygen production was measured via fluorescence and it has been observed that, at first, the oxygen concentration decreased but after a few hours it relaxed into stable metabolic dynamics. Each community was sequenced either when its oxygen concentration initially decreased or when it stabilized. The description of each experiment can be found in Table 4.1.

Each community from the experiment described above was sequenced using 16S ONT sequencing. The reads were demultiplexed and partitioned into separate files according to the sequenced barcode (barcode 1 - 24). Based on the notes on each experiment, only 14 experiments were performed. The larger amount of barcodes comes from the fact that we were not able to tell exactly which experiment they came from. Therefore, new barcodes were called and those unidentified reads were collected there.

4.4.2 Testing the algorithm using mock datasets

The algorithm was tested on CAMI2 datasets: the 2nd CAMI Challenge Marine Dataset and the 2nd CAMI Toy Human Microbiome Project Dataset. The FASTQ files containing the mock datasets were downloaded from cami-challenge.org website along with the mapping files containing species information for each sequence. The FASTQ files were run through the pipeline and through BLAST+ v2.15 using the nt_prok database.

The results from the BLAST output file were analyzed using custom scripts. This analysis was performed separately for each dataset.

4.4.3 Analysis of the experimental dataset

The 24 FASTQ files containing the sequences were downloaded from a private repository. Only files named barcode 01 to barcode 14 were used in the analysis since they corresponded to the conducted experiments. Each file was run through the pipeline and 5,000 randomly selected consensus sequences were used for further analysis. The random samples were run through BLAST+ v.2.15 using nt_prok database and the results were analyzed using custom scripts. The scripts can be found in this github repository.

4.5 Permissions and Attributions

The content of Chapter 4 is the result of a collaboration with Alexander Petroff at Clark University, MA.

Chapter 5

Concluding Remarks

5.1 Summary

The work presented in this dissertation describes the extensive horizontal gene transfer in Pink Berries, promising preliminary findings of changes is defense systems in Purple Sulfur Bacteria, and a new tool for de-noising 16S sequences sequenced using Nanopore technology.

In chapter 2, we show a comprehensive analysis of the Pink Berry metagenomic data in the context of extensive recombination of a quasi-sexual bacterial population. We present evidence that the populations are divided into clades with different evolutionary histories including a genetic mixing layer where bacteria experience extensive recombination from other clades as well as with each other.

In chapter 3, we continue our investigation of the Pink Berries in the context of self / non-self recognition and defense tactics against phages . We show preliminary findings of diversity in WapA and RhsC C-termini in purple sulfur bacteria. The presence of large gaps in the alignment of the C-terminus region of the CDI proteins shows us that there is a possibility that the bacteria differ in their repertoire of toxins depending on their geographical location. These observations mark a promising starting point for studying CDI mechanisms in naturally occurring bacterial populations. Additionally, we show evidence of clades having different defense genes based on the lowered depth of coverage of those ORFs when compared to the rest of the genes.

Lastly, in chapter 4, we introduce a new tool for decreasing the error of 16S Nanopore sequences and identifying given samples. We demonstrate the accuracy of the pipeline using simulated PacBio and Nanopore datasets from CAMI2 and then use it on experimental sediment data. We use the pipeline on our experimental reads from fourteen different experiments that aimed to address the question of how the community composition changes given different starting conditions. However, the results in their current state are inconclusive and further work should be done in order to gather more data across different conditions and time points.

5.2 Outlook

It is not a surprise that bacteria experience extensive HGT between different members of the same species. However, the details of the transfer such as the identity of the genes, breakpoints, or rates of transfer differ vastly between species. In order to understand how and why transfer happens in naturally occurring communities one needs to study those details and compare them between different communities. Moreover, studying specific genes, such as toxin-antitoxin pairs and CDI mechanisms and how they differ between different locations can enrich our understanding on those proteins. With so many types of structures that inject the toxins as well as many diverse types of toxins and immunity genes, it is a gold mine for learning about how bacteria fight the competition in a natural environment. Lastly, with new sequencing technologies, such as log-read sequencing, there is a need for new algorithms and programs that will clean and analyze the datasets. The pipeline described above is a new and easy-to-use tool that give researchers more options for analyzing their 16S rRNA log-read sequences. The field of metagenomics, especially the studies of microbial communities and their compositions, make it possible for us to understand the organisms in their natural and often chaotic environment. It is an exciting field that helps us understand life in the context of community instead of isolated individuals in a petri dish. In conclusion, this dissertation shows the importance of statistics and bioinformatics in order to understand microbial genomics.

Bibliography

[1] Griffith F. The Significance of Pneumococcal Types. The Journal of Hygiene. 1928;27(2):113-59. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2167760/.

[2] Avery OT, MacLeod CM, McCarty M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES;79(2):137-58. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2135445/.

[3] Lederberg J, Tatum EL. Gene Recombination in Escherichia Coli;158(4016):558-8. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/158558a0.

[4] Amábile-Cuevas CF, Chicurel ME. Horizontal Gene Transfer;81(4):332-41. Publisher: Sigma Xi, The Scientific Research Society. Available from: https://www.jstor.org/stable/29774969.

[5] Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria;20(4):206-18. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41579-021-00650-4.

[6] Bitto NJ, Chapman R, Pidot S, Costin A, Lo C, Choi J, et al. Bacterial membrane vesicles transport their DNA cargo into host cells;7(1):7072. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41598-017-07288-4.

[7] Dubey GP, Ben-Yehuda S. Intercellular Nanotubes Mediate Bacterial Communication;144(4):590-600. Publisher: Elsevier. Available from: https://www.cell.com/cell/abstract/S0092-8674(11)00016-X.

[8] Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution;3(9):679-87. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nrmicro1204.

[9] Cordero OX, Hogeweg P. The impact of long-distance horizontal gene transfer on prokaryotic genome size;106(51):21748-53. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/full/10.1073/pnas.0907584106.

[10] Gogarten JP, Doolittle WF, Lawrence JG. Prokaryotic evolution in light of gene transfer;19(12):2226-38.

[11] Bhatia RP, Kirit HA, Lewis CM Jr, Sankaranarayanan K, Bollback JP. Evolutionary barriers to horizontal gene transfer in macrophage-associated Salmonella;7(4):227-39. Available from: https://doi.org/10.1093/evlett/qrad020.

[12] Gomes ALC, Johns NI, Yang A, Velez-Cortes F, Smillie CS, Smith MB, et al. Genome and sequence determinants governing the expression of horizontally acquired DNA in bacteria;14(9):2347-57. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7608860/.

[13] Tock MR, Dryden DTF. The biology of restriction and anti-restriction;8(4):466-72.

[14] Navarre WW. The Impact of Gene Silencing on Horizontal Gene Transfer and Bacterial Evolution;69:157-86.

[15] Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms;8(5):317-27. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nrmicro2315.

[16] Papp B, Pál C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast;424(6945):194-7. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature01771.

[17] Park C, Zhang J. High Expression Hampers Horizontal Gene Transfer;4(4):523-32. Available from: https://doi.org/10.1093/gbe/evs030.

[18] Sigwart J. Coalescent Theory: An Introduction;58(1):162-5. Available from: https://doi.org/10.1093/schbul/syp004.

[19] Zhaxybayeva O, Gogarten JP. Cladogenesis, coalescence and the evolution of the three domains of life;20(4):182-7. Publisher: Elsevier. Available from: https://www.cell.com/trends/genetics/abstract/S0168-9525(04)00042-3.

[20] Brito IL. Examining horizontal gene transfer in microbial communities;19(7):442-53. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41579-021-00534-7.

[21] Woods LC, Gorrell RJ, Taylor F, Connallon T, Kwok T, McDonald MJ. Horizontal gene transfer potentiates adaptation by reducing selective constraints on the spread of genetic variation;117(43):26868-75. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/full/10.1073/pnas.2005331117.

[22] Garud NR, Good BH, Hallatschek O, Pollard KS. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts;17(1):e3000102. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000102.

[23] Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease;19(1):55-71. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41579-020-0433-9.

[24] Chu VT, Tsitsiklis A, Mick E, Ambroggio L, Kalantar KL, Glascock A, et al. The antibiotic resistance reservoir of the lung microbiome expands with age in a population of critically ill patients;15(1):92. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-023-44353-1.

[25] Papanicolas LE, Gordon DL, Wesselingh SL, Rogers GB. Not Just Antibiotics: Is Cancer Chemotherapy Driving Antimicrobial Resistance?;26(5):393-400. Publisher: Elsevier. Available from: https://www.cell.com/trends/microbiology/abstract/S0966-842X(17)30236-6.

[26] Lee SY, Kim HU, Chae TU, Cho JS, Kim JW, Shin JH, et al. A comprehensive metabolic map for production of bio-based chemicals;2(1):18-33. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41929-018-0212-4.

[27] Yuan SF, Alper HS. Metabolic engineering of microbial cell factories for production of nutraceuticals;18(1):46. Available from: https://doi.org/10.1186/s12934-019-1096-y.

[28] Logan BE, Rabaey K. Conversion of Wastes into Bioelectricity and Chemicals by Using Microbial Electrochemical Technologies;337(6095):686-90. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/10.1126/science.1217412.

[29] Blazanin M, Turner PE. Community context matters for bacteria-phage ecology and evolution;15(11):3119-28.

[30] Shikov AE, Malovichko YV, Nizhnikov AA, Antonets KS. Current Methods for Recombination Detection in Bacteria;23(11):6257.

[31] Moralez J, Szenkiel K, Hamilton K, Pruden A, Lopatkin AJ. Quantitative analysis of horizontal gene transfer in complex systems;62:103-9.

[32] Slager J, Kjos M, Attaiech L, Veening JW. Antibiotic-induced replication stress triggers bacterial competence by increasing gene dosage near the origin;157(2):395-406.

[33] Wan Z, Varshavsky J, Teegala S, McLawrence J, Goddard N. Measuring the Rate of Conjugal Plasmid Transfer in a Bacterial Population Using Quantitative PCR;101(1):237-44. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3127179/.

[34] Sørensen SJ, Sørensen AH, Hansen LH, Oregaard G, Veal D. Direct detection and quantification of horizontal gene transfer by using flow cytometry and gfp as a reporter gene;47(2):129-33.

[35] Neil K, Allard N, Grenier F, Burrus V, Rodrigue S. Highly efficient gene transfer in the mouse gut microbiota is enabled by the Incl2 conjugative plasmid TP114;3(1):1-9. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s42003-020-01253-0.

[36] Jacquiod S, Brejnrod A, Morberg SM, Abu Al-Soud W, Sørensen SJ, Riber L. Deciphering conjugative plasmid permissiveness in wastewater microbiomes;26(13):3556-71.

[37] Munck C, Albertsen M, Telke A, Ellabaan M, Nielsen PH, Sommer MOA. Limited dissemination of the wastewater treatment plant core resistome;6(1):8452. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ncomms9452.

[38] Martin DP, Lemey P, Posada D. Analysing recombination in nucleotide sequences;11(6):943-55.

[39] Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations;34(5):1167-82.

[40] Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins;43(3):e15.

[41] Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes;11(2):e1004041. Publisher: Public Library of Science. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004041.

[42] Hanage WP. Not So Simple After All: Bacteria, Their Population Genetics, and Recombination;8(7):a018069.

[43] Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future;9(6):477-85. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nrg2361.

[44] Nordborg M, Tavaré S. Linkage disequilibrium: what history has to tell us;18(2):83-90. Available from: https://www.sciencedirect.com/science/article/pii/S016895250202557X.

[45] Good BH. Linkage disequilibrium between rare mutations;220(4):iyac004. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8982034/.

[46] Rosen MJ, Davison M, Bhaya D, Fisher DS. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche;348(6238):1019-23. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/10.1126/science.aaa4456.

[47] Lewontin RC. On measures of gametic disequilibrium.;120(3):849-52. Available from: https://doi.org/10.1093/genetics/120.3.849.

[48] Lin M, Kussell E. Inferring bacterial recombination rates from large-scale sequencing datasets;16(2):199-204. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41592-018-0293-7.

[49] Graham EB, Knelman JE, Schindlbacher A, Siciliano S, Breulmann M, Yannarell A, et al. Microbes as Engines of Ecosystem Function: When Does Community Structure Enhance Predictions of Ecosystem Processes?;7. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2016.00214/full.

[50] Gold Howard S , Moellering Robert C . Antimicrobial-Drug Resistance;335(19):1445-53. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJM199611073351907. Available from: https://www.nejm.org/doi/full/10.1056/NEJM199611073351907.

[51] Wong HL, White RA, Visscher PT, Charlesworth JC, Vázquez-Campos X, Burns BP. Disentangling the drivers of functional complexity at the metagenomic level in Shark Bay microbial mat microbiomes;12(11):2619-39. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6194083/.

[52] Song W, Wemheuer B, Steinberg PD, Marzinelli EM, Thomas T. Contribution of horizontal gene transfer to the functionality of microbial biofilm on a

macroalgae;15(3):807-17. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41396-020-00815-8.

[53] Wilpiszeski RL, Aufrecht JA, Retterer ST, Sullivan MB, Graham DE, Pierce EM, et al. Soil Aggregate Microbial Communities: Towards Understanding Microbiome Interactions at Biologically Relevant Scales;85(14):e00324-19. Publisher: American Society for Microbiology. Available from: https://journals.asm.org/doi/10.1128/aem.00324-19.

[54] Wilbanks EG, Jaekel U, Salman V, Humphrey PT, Eisen JA, Facciotti MT, et al. Microscale sulfur cycling in the phototrophic pink berry consortia of the Sippewissett Salt Marsh;16(11):3398-415. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4262008/.

[55] Cuthbert BJ, Hayes CS, Goulding CW. Functional and Structural Diversity of Bacterial Contact-Dependent Growth Inhibition Effectors;9. Publisher: Frontiers. Available from: https://www.frontiersin.org/articles/10.3389/fmolb.2022.866854.

[56] Aoki SK, Pamma R, Hernday AD, Bickham JE, Braaten BA, Low DA. Contact-dependent inhibition of growth in Escherichia coli;309(5738):1245-8.

[57] Hill CW, Sandt CH, Vlazny DA. Rhs elements of Escherichia coli: a family of genetic composites each encoding a large mosaic protein;12(6):865-71.

[58] Jackson AP, Thomas GH, Parkhill J, Thomson NR. Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement;10(1):584. Available from: https://doi.org/10.1186/1471-2164-10-584.

[59] Koskiniemi S, Lamoureux JG, Nikolakakis KC, t'Kint de Roodenbeke C, Kaplan MD, Low DA, et al. Rhs proteins from diverse bacteria mediate intercellular competition;110(17):7032-7. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/full/10.1073/pnas.1300627110.

[60] Abrusci P, McDowell MA, Lea SM, Johnson S. Building a secreting nanomachine: a structural overview of the T3SS;25(100):111-7.

[61] Spitz O, Erenburg IN, Beer T, Kanonenberg K, Holland IB, Schmitt L. Type I Secretion Systems—One Mechanism for All?;7(2):10.1128/microbiolspec.psib-00032018. Publisher: American Society for Microbiology. Available from: https://journals.asm.org/doi/10.1128/microbiolspec.psib-0003-2018.

[62] Costa TRD, Patkowski JB, Macé K, Christie PJ, Waksman G. Structural and functional diversity of type IV secretion systems;22(3):170-85.

[63] Wallden K, Rivera-Calzada A, Waksman G. Type IV secretion systems: versatility and diversity in function;12(9):1203-12. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3070162/.

[64] Cascales E, Christie PJ. The versatile bacterial type IV secretion systems;1(2):137-49.

[65] Meuskens I, Saragliadis A, Leo JC, Linke D. Type V Secretion Systems: An Overview of Passenger Domain Functions;10. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2019.01163/full.

[66] Pohlner J, Halter R, Beyreuther K, Meyer TF. Gene structure and extracellular secretion of Neisseria gonorrhoeae IgA protease;325(6103):458-62.

[67] Coulthurst S. The Type VI secretion system: a versatile bacterial weapon;165(5):503-15.

[68] Russell AB, Peterson SB, Mougous JD. Type VI secretion system effectors: poisons with a purpose;12(2):137-48.

[69] Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, et al. Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus;32(6):1396-410.

[70] Cabezón E, de la Cruz F, Arechaga I. Conjugation Inhibitors and Their Potential Use to Prevent Dissemination of Antibiotic Resistance Genes in Bacteria;8:2329. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5723004/.

[71] Boudaher E, Shaffer CL. Inhibiting bacterial secretion systems in the fight against antibiotic resistance;10(5):682-92. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6677025/.

[72] Paschos A, den Hartigh A, Smith MA, Atluri VL, Sivanesan D, Tsolis RM, et al. An in vivo high-throughput screening approach targeting the type IV secretion system component VirB8 identified inhibitors of Brucella abortus 2308 proliferation;79(3):1033-43.

[73] Shaffer CL, Good JAD, Kumar S, Krishnan KS, Gaddy JA, Loh JT, et al. Peptidomimetic Small Molecules Disrupt Type IV Secretion System Activity in Diverse Bacterial Pathogens;7(2):e00221-16.

[74] Álvarez Rodríguez I, Arana L, Ugarte-Uribe B, Gómez-Rubio E, Martín-Santamaría S, Garbisu C, et al. Type IV Coupling Proteins as Potential Targets to Control the Dissemination of Antibiotic Resistance;7:201. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7434980/.

[75] Clarridge JE. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases;17(4):840-62. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC523561/.

[76] Curry KD, Wang Q, Nute MG, Tyshaieva A, Reeves E, Soriano S, et al. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data;19(7):845-53. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41592-022-01520-4.

[77] Salipante SJ, Sengupta DJ, Rosenthal C, Costa G, Spangler J, Sims EH, et al. Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections;8(5):e65226.

[78] Jenkins C, Ling CL, Ciesielczuk HL, Lockwood J, Hopkins S, McHugh TD, et al. Detection and identification of bacteria in clinical samples by 16S rRNA gene sequencing: comparison of two different approaches in clinical practice;61(4):483-8. Publisher: Microbiology Society,. Available from: https://www.microbiologyresearch.org/content/journal/jmm/10.1099/jmm.0.030387-0.

[79] Huang Y, Xiao Z, Cao Y, Gao F, Fu Y, Zou M, et al. Rapid microbiological diagnosis based on 16S rRNA gene sequencing: A comparison of bacterial composition in diabetic foot infections and contralateral intact skin;13. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2022.1021955/full.

[80] Do TT, Delaney S, Walsh F. 16S rRNA gene based bacterial community structure of wastewater treatment plant effluents;366(3):fnz017.

[81] Tallei TE, Fatimawali, Yelnetty A, Kusumawaty D, Effendi Y, Park MN, et al. Predictive Microbial Community and Functional Gene Expression Profiles in Pineapple Peel Fermentation Using 16S rRNA Gene Sequences;8(5):194. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/2311-5637/8/5/194.

[82] Timke M, Wang-Lieu NQ, Altendorf K, Lipski A. Community Structure and Diversity of Biofilms from a Beer Bottling Plant as Revealed Using 16S rRNA Gene Clone Libraries;71(10):6446-52. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1266004/.

[83] Mitsuhashi S, Kryukov K, Nakagawa S, Takeuchi JS, Shiraishi Y, Asano K, et al. A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer;7(1):5657.

[84] Nakagawa S, Inoue S, Kryukov K, Yamagishi J, Ohno A, Hayashida K, et al. Rapid sequencing-based diagnosis of infectious bacterial species from meningitis patients

in Zambia;8(11):e01087. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6831930/.

[85] Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in functional genomics;61(5):316-26. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/dgd.12608. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/dgd.12608.

[86] Maier B. Competence and Transformation in Bacillus subtilis;37(1):57-76. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/1467-3045/37/1/5.

[87] Gontier N. Historical and Epistemological Perspectives on What Horizontal Gene Transfer Mechanisms Contribute to Our Understanding of Evolution. In: Gontier N, editor. Reticulate Evolution: Symbiogenesis, Lateral Gene Transfer, Hybridization and Infectious Heredity. Springer International Publishing;. p. 121-78. Available from: https://doi.org/10.1007/978-3-319-16345-1_5.

[88] Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE. Whole-Genome Analysis of Photosynthetic Prokaryotes;298(5598):1616-20. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/abs/10.1126/science.1075558.

[89] Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: The complexity hypothesis;96(7):3801-6. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/full/10.1073/pnas.96.7.3801.

[90] Townsend JP, Bøhn T, Nielsen KM. Assessing the Probability of Detection of Horizontal Gene Transfer Events in Bacterial Populations;3:27. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3282476/.

[91] Townsend JP, Nielsen KM, Fisher DS, Hartl DL. Horizontal acquisition of divergent chromosomal DNA in bacteria: effects of mutator phenotypes.;164(1):13-21. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1462543/.

[92] Nielsen KM, Townsend JP. Monitoring and modeling horizontal gene transfer;22(9):1110-4. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nbt1006.

[93] Michaelis C, Grohmann E. Horizontal Gene Transfer of Antibiotic Resistance Genes in Biofilms;12(2):328. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9952180/.

[94] Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph;31(10):1674-6.

[95] Li H. Minimap2: pairwise alignment for nucleotide sequences;34(18):3094-100. Available from: https://doi.org/10.1093/bioinformatics/bty191.

[96] Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition;11(11):1144-6. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nmeth.3103.

[97] Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets;32(4):605-7. Available from: https://doi.org/10.1093/bioinformatics/btv638.

[98] Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies;7:e7359.

[99] Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, et al. Improved metagenome binning and assembly using deep variational autoencoders;39(5):555-60.

[100] Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes;25(7):1043-55.

[101] Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database;36(6):1925-7.

[102] Sakoparnig T, Field C, van Nimwegen E. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species;10:e65366. Publisher: eLife Sciences Publications, Ltd. Available from: https://doi.org/10.7554/eLife.65366.

[103] Oliveira PH, Touchon M, Cury J, Rocha EPC. The chromosomal organization of horizontal gene transfer in bacteria;8(1):841. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-017-00808-w.

[104] Wilson DJ, The CRyPTIC Consortium. GenomegaMap: Within-Species Genome-Wide dN/dS Estimation from over 10,000 Genomes;37(8):2450-60. Available from: https://doi.org/10.1093/molbev/msaa069.

[105] Kryazhimskiy S, Plotkin JB. The Population Genetics of dN/dS;4(12):e1000304. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000304.

[106] Lewontin RC, Kojima Ki. The Evolutionary Dynamics of Complex Polymorphisms,,;14(4):458-72. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1558-5646.1960.tb03113.x. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.1960.tb03113.x.

[107] Hudson RR, Kaplan NL. STATISTICAL PROPERTIES OF THE NUMBER OF RECOMBINATION EVENTS IN THE HISTORY OF A SAMPLE OF DNA SEQUENCES;111(1):147-64. Available from: https://doi.org/10.1093/genetics/111.1.147.

[108] Batot G, Michalska K, Ekberg G, Irimpan EM, Joachimiak G, Jedrzejczak R, et al. The CDI toxin of Yersinia kristensenii is a novel bacterial member of the RNase A superfamily;45(9):5013-25. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5435912/.

[109] Allen JP, Ozer EA, Minasov G, Shuvalova L, Kiryukhina O, Satchell KJF, et al. A comparative genomics approach identifies contact-dependent growth inhibition as a virulence determinant;117(12):6811-21. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7104216/.

[110] Poole SJ, Diner EJ, Aoki SK, Braaten BA, t'Kint de Roodenbeke C, Low DA, et al. Identification of functional toxin/immunity genes linked to contact-dependent growth inhibition (CDI) and rearrangement hotspot (Rhs) systems;7(8):e1002217.

[111] Cuthbert BJ, Burley KH, Goulding CW. Introducing the new bacterial branch of the RNase A superfamily;15(1):9-12. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5786019/.

[112] Hamilton TA, Pellegrino GM, Therrien JA, Ham DT, Bartlett PC, Karas BJ, et al. Efficient inter-species conjugative transfer of a CRISPR nuclease for targeted bacterial killing;10(1):4544. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-019-12448-3.

[113] Vrancianu CO, Popa LI, Bleotu C, Chifiriuc MC. Targeting Plasmids to Limit Acquisition and Transmission of Antimicrobial Resistance;11. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2020.00761/full.

[114] Reuter A, Hilpert C, Dedieu-Berne A, Lematre S, Gueguen E, Launay G, et al. Targeted-antibacterial-plasmids (TAPs) combining conjugation and CRISPR/Cas systems achieve strain-specific antibacterial activity. 2021;49(6):3584-98.

[115] David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome;505(7484):559-63.

Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature12820.

[116] Seedorf H, Griffin NW, Ridaura VK, Reyes A, Cheng J, Rey FE, et al. Bacteria from diverse habitats colonize and compete in the mouse gut;159(2):253-66.

[117] Rakoff-Nahoum S, Foster KR, Comstock LE. The evolution of cooperation within the gut microbiota;533(7602):255-9.

[118] van Gemerden H. Microbial mats: A joint venture;113(1):3-25. Available from: https://www.sciencedirect.com/science/article/pii/002532279390146M.

[119] Bordenave S, Goñi-Urriza MS, Caumette P, Duran R. Effects of Heavy Fuel Oil on the Bacterial Community Structure of a Pristine Microbial Mat;73(19):6089-97. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2075027/.

[120] Grilli J, Adorisio M, Suweis S, Barabás G, Banavar JR, Allesina S, et al. Feasibility and coexistence of large ecological communities;8(1):14389. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ncomms14389.

[121] Song C, Rohr RP, Saavedra S. A guideline to study the feasibility domain of multi-trophic and changing ecological communities;450:30-6.

[122] Gillooly JF, Brown JH, West GB, Savage VM, Charnov EL. Effects of size and temperature on metabolic rate;293(5538):2248-51.

[123] Walther GR, Post E, Convey P, Menzel A, Parmesan C, Beebee TJC, et al. Ecological responses to recent climate change;416(6879):389-95. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/416389a.

[124] Tylianakis JM, Laliberté E, Nielsen A, Bascompte J. Conservation of species interaction networks;143(10):2270-9. Available from: https://www.sciencedirect.com/science/article/pii/S0006320709005126.

[125] Armitage DW, Gallagher KL, Youngblut ND, Buckley DH, Zinder SH. Millimeter-scale patterns of phylogenetic and trait diversity in a salt marsh microbial mat;3. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2012.00293/full.

[126] Ramos VMC, Castelo-Branco R, Leão PN, Martins J, Carvalhal-Gomes S, Sobrinho da Silva F, et al. Cyanobacterial Diversity in Microbial Mats from the Hypersaline Lagoon System of Araruama, Brazil: An In-depth Polyphasic Study;8. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2017.01233/full.

[127] Goh F, Allen MA, Leuko S, Kawaguchi T, Decho AW, Burns BP, et al. Determining the specific microbial populations and their spatial distribution within the stromatolite ecosystem of Shark Bay;3(4):383-96. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ismej2008114.

[128] Santoyo G. Unveiling Taxonomic Diversity and Functional Composition Differences of Microbial Mat Communities Through Comparative Metagenomics;38(7):639-48. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01490451.2021.1926600. Available from: https://doi.org/10.1080/01490451.2021.1926600.

[129] Xiao Y, Yu D. Tumor microenvironment as a therapeutic target in cancer;221:107753.

[130] Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis;35(9):833-44. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nbt.3935.

[131] Ichikawa K, Kawahara R, Asano T, Morishita S. A landscape of complex tandem repeats within individual human genomes;14(1):5530. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-023-41262-1.

[132] Rodríguez-Pérez H, Ciuffreda L, Flores C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data;37(11):1600-1. Available from: https://doi.org/10.1093/bioinformatics/btaa900.

[133] Deng Y, Mauri M, Vallet M, Staudinger M, Allen RJ, Pohnert G. Dynamic Diatom-Bacteria Consortia in Synthetic Plankton Communities;88(22):e01619-22. Publisher: American Society for Microbiology. Available from: https://journals.asm.org/doi/10.1128/aem.01619-22.

[134] Uysal AK, Gunal S. The impact of preprocessing on text classification;50(1):104-12. Available from: https://www.sciencedirect.com/science/article/pii/S0306457313000964.

[135] Mirończuk MM, Protasiewicz J. A recent overview of the state-of-the-art elements of text classification;106:36-54. Available from: https://www.sciencedirect.com/science/article/pii/S095741741830215X.

[136] Lahitani AR, Permanasari AE, Setiawan NA. Cosine similarity to determine similarity measure: Study case in online essay assessment. In: 2016 4th International Conference on Cyber and IT Service Management;. p. 1-6. Available from: https://ieeexplore.ieee.org/document/7577578.

[137] Suwanda R, Syahputra Z, Zamzami EM. Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K;1566(1):012058. Available from: https://iopscience.iop.org/article/10.1088/1742-6596/1566/1/012058.

[138] McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering;2(11):205. Available from: https://joss.theoj.org/papers/10.21105/joss.00205.

[139] Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points;315(5814):972-6. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/10.1126/science.1136800.

[140] Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis;24(5):603-19. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. Available from: https://ieeexplore.ieee.org/document/1000236.

[141] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases;25(2):103-14. Available from: https://doi.org/10.1145/235968.233324.

[142] Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data;622:178-210. Available from: https://www.sciencedirect.com/science/article/pii/S0020025522014633.

[143] Müllner D. Modern hierarchical, agglomerative clustering algorithms. Available from: https://www.semanticscholar.org/paper/Modern-hierarchical%2C-agglomerative-clustering-M%C3%BCllner/3d28a83e86dc5ab0334ae014b1218f0ad856f76b.

[144] Katoh K, Misawa K, Kuma Ki, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform;30(14):3059-66.

[145] Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, et al. Search and sequence analysis tools services from EMBL-EBI in 2022;50:W276-9.

[146] Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements;41:W29-33. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3692093/.

[147] Meyer F, Fritz A, Deng ZL, Koslicki D, Lesker TR, Gurevich A, et al. Critical Assessment of Metagenome Interpretation: the second round of challenges;19(4):429-40. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41592-022-01431-4.