Defining the Limits of Hi-C for Identifying Hosts of Plasmids and Antibiotic Resistance Genes in Simple Bacterial Communities

A Thesis
Presented in Partial Fulfillment of the Requirements for the
Degree of Master of Science
with a
Major in Bioinformatics and Computational Biology
in the
College of Graduate Studies
University of Idaho
by
Bethel J. Kohler

Major Professor: Celeste J. Brown, Ph.D.
Committee Members: Tanya Miura, Ph.D.; Steve Krone, Ph.D.; Stephen Lee, Ph.D.
Department Administrator: Eva Top, Ph.D.

December 2017

Authorization to Submit Thesis

This thesis of Bethel J. Kohler, submitted for the degree of Master of Science with a Major in Bioinformatics and Computational Biology and titled "Defining the Limits of Hi-C for Identifying Hosts of Plasmids and Antibiotic Resistance Genes in Simple Bacterial Communities," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: _____ Date: _____

Celeste J. Brown, Ph.D.

Committee Member: _____ Date: _____

Tanya Miura, Ph.D.

_____ Date: _____

Steve Krone, Ph.D.

_____ Date: _____

Stephen Lee, Ph.D.

Department Administrator: _____ Date: _____

Eva Top, Ph.D.

Abstract

The rapid spread of bacterial antibiotic resistance is a major public health threat that is making many of our known antibiotics ineffective at fighting pathogens. The horizontal transfer of antibiotic resistance genes (ARG) by mobile genetic elements such as plasmids plays a major role in this health crisis. To curb the spread of ARG, we need to identify the hosts and plasmids that carry them in the environment. Traditional metagenomic approaches have been of limited use in this because of their poor ability to link ARG to the chromosome or plasmid that carry them. Cultivation-based methods are also not effective due to the high abundance of unculturable bacteria in most ecosystems. The chromosomal conformation capture (Hi-C) approach has recently been proven useful for reconstructing individual genomes from mixed cell populations by physically linking DNA fragments that occupied the same cell prior to sequencing. However, the limits of Hi-C detection, especially in the areas of plasmid and antibiotic resistance research, have not been defined. This project tested the hypothesis that plasmid and bacterial carriers of ARG can be identified using the Hi-C method while also beginning to define its limits in these areas. Towards this end, a bioinformatics pipeline was constructed for Hi-C data analysis. Using this pipeline, the Hi-C clustering of metagenomic contigs by species can be done using a simple script implemented in R. A series of mock bacterial communities was designed to measure the limits of Hi-C detection for plasmids and ARG while mimicking realistic environmental and plasmid-transfer scenarios. The pipeline was shown to be effective for clustering the contigs from these communities by genome. Clustering based on Hi-C linkages also made it possible to identify the genetic context of ARG in bacterial communities, for instance, whether ARG were encoded on bacterial chromosomes or transmissible plasmids. Multiple plasmids could be accurately shown to be carried by the same species, the genome of a species present at as low as 5% of the community could be detected, and the same plasmid could be definitively determined to be present in multiple species when present in those species in equal amounts. This information will lead the way for future research applications using Hi-C to identify the bacterial carriers of ARG and plasmids in diverse environmental samples.

Acknowledgements

I would like to acknowledge the IBEST CRC and GRC for their technical support, and especially Dr. Samuel Hunter who provided his expertise during the creation of this bioinformatics pipeline. I would also like to acknowledge the University of Idaho Student Support Services/TRiO program and especially Reagen Iler who introduced me to the idea of graduate school and helped me to believe that it was for me. Thank you to my advisor, Dr. Celeste Brown for the bioinformatics support and guidance while finishing up this project. Thank you to the members of the Top lab who prepared me for and helped me with this research by teaching me so many lab skills and encouraging me to pursue graduate education. Also, thank you to Kenetta Nunn for all the help culturing *Lactobacillus*.

I would also like to acknowledge my fellow graduate students, Dr. Genevieve Metzger and Dr. Courtney Thompson, as well as Chad Manwaring. Without your support through the events of the past year, I would not have been able to complete this program.

I would also like to acknowledge our funding sources that made this project possible. This project was supported by NSF Science and Technology Center on evolution in action, DBI-0939454, and Institutional Development Award (IDeA) from the NIGMS of the NIH under grant number P30 GM103324. I was also supported by IBEST and the BCB graduate program through a fellowship.

Dedication

This thesis is dedicated to my dear siblings, and especially my little sisters: Mercy, Abigail, Glory, and Grace. You have always inspired me to become the best version of myself and to work to make your world a safer place.

Table of Contents

List of Figures

List of Tables

Chapter 1: Literature Review

The Spread of Antibiotic Resistance

The discovery of penicillin by Alexander Fleming in 1928 heralded the beginning of the antibiotic era. Considered one of the greatest advances in therapeutic medicine, these miracle drugs were mass-produced and simple infections no longer equaled a death sentence ("American Chemical Society", 2015). More drugs were developed and added to the market and it quickly became difficult to imagine a world without antibiotics, some of the most commonly prescribed drugs in modern medicine. However, as long ago as 1945, when Alexander Fleming gave his acceptance speech for the Nobel Prize that he won for his discovery, he warned of bacteria becoming resistant to the drug. At that time, penicillin resistance had already been identified in strains of *Streptococcus*. Resistances in more strains and to more antibiotics quickly appeared ("Centers for Disease Control", 2017) as bacteria adapted to the presence of these drugs.

The rapid spread of bacterial antibiotic resistance is a global public health issue that has left many of our known antibiotic drugs ineffective at fighting pathogens. In 2015, the White House recognized the significance of this issue, releasing a plan of action for combating the rise of antibiotic resistant bacteria and recognizing this plan as a necessary, lifesaving effort (White House, 2015). The term superbug was coined to refer to bacteria that can no longer be killed by some or all commonly used antibiotics. While these superbugs have created a strong push to discover new antibiotics (Maffioli et al., 2017), new resistances and superbugs emerge regularly and at a much faster rate than new antibiotic drugs are being developed. The World Health Organization warned in 2014 that we are very likely entering a "post-antibiotic" era where we will lack effective drug therapies for even the most common infections ("WHO", 2014).

The horizontal transfer of antibiotic resistance genes (ARG) by mobile genetic elements such as plasmids has played a major role in this health crisis (Lester et al., 2006). Plasmids are small, typically circular pieces of DNA that can be found in some bacterial cells in addition to the chromosome (Roth & Helinski, 1967). They often encode accessory genes that make it possible for their hosts to survive and even thrive in various,

more extreme living conditions (Kado, 1998). This includes surviving in the presence of antibiotics if a plasmid encodes for antibiotic resistance.

Some plasmids can transfer copies of themselves to neighboring cells. Once there, they may integrate into the chromosome or pick up more ARG before transferring new copies to other neighbors. This type of plasmid exchange is called conjugation. While not all plasmids are capable of conjugation, and some conjugating plasmids can only be transferred among specific species, some can be transferred to a broad host range (BHR). These BHR plasmids are especially important contributors to the rapid evolution of bacterial antibiotic resistance. One plasmid can often bestow multiple antibiotic resistances to a new species through just one of these conjugation events (Lester et al., 2006). "Super" plasmids, carrying multiple resistances (Oliva et al., 2017), are an especially big public threat if they have conjugative capabilities (Dang et al., 2017; Tang et al., 2017) and a broad host range.

Environmental Reservoirs of Antibiotic Resistance

To better understand the evolution of antibiotic resistance and better manage our antibiotic resources, a better understanding is needed of the direction of ARG movement in the environment (Berendonk et al., 2015). Environmental contamination with both antibiotics and ARG has become nearly ubiquitous (Berendonk et al., 2015). Due to their widespread presence, antibiotics have become contaminants of emerging concern (Dodder et al., 2014) that have been shown to get taken up by and be present in food sources such as vegetables (Kumar et al., 2005) and coastal mussels (Dodder et al., 2014). Antibiotic resistance determinants have also been found in most areas studied, including ecosystems as diverse as soil (Forsberg et al., 2012), manure (Zhu et al., 2013), water, waste effluent, and river sediment (Dang et al., 2017).

Our current health crisis is due to multiple factors. It is possibly due in part to naturally occurring environmental ARG becoming plasmid borne, transferring to pathogens, and then showing up in hospitals where they become medically relevant (Forsberg et al., 2012). Some environmental microorganisms naturally produce antibiotics (Waksman & Woodruff, 1940) and these get released into the environment. Some microorganisms thus naturally carry resistance genes (Forsberg et al., 2012; D'costa et al., 2007) to survive in their presence. However, our health crisis is also likely

due to human factors. Antibiotic resistant bacteria from clinics or farms can disseminate through the environment by use of vectors or due to waste management practices that allow for genetic dispersal to new locations by natural means, such as by waterways (Munir et al., 2011). Mixing bacteria from an anthropogenic source with environmental bacteria creates an ideal environment for genetic exchange and evolution due to plasmid-mediated horizontal gene transfer (Gotz & Smalla, 1997). Environmental contamination with commercially produced antibiotics also enriches for native bacteria that have evolved more resistances (Martinez, 2009). The fact that antibiotics can have significantly longer half-lives in matrices such as soil than they would in an aqueous solution (Du & Liu, 2012) means that contamination of some areas can have greater long term consequences. Compounding this problem, antibiotic waste is often disseminated in the environment alongside bacteria carrying resistance genes, such as when manure is spread on an agricultural field as fertilizer (Zhu et al., 2013; Heuer & Smalla, 2007, Udikovic-Kolic et al., 2014) or when treated waste water or biosolids are released into the environment (Munir et al., 2011).

Deciphering the most common origins and environmental reservoirs of ARG, before they show up in hospitals and become medically relevant, is fundamental to understanding the evolution of antibiotic resistance and managing antibiotic use. Gaining this understanding requires better methods for analyzing environmental bacterial samples. Whether a swab from a clinic or a gram of soil, these samples are comprised of an entire bacterial community. The community members could range from $2 – 5x10^4$ species in the case of something as complex as soil (Roesch et al., 2007). The major species present may or may not be known and some or all the members could be carrying unknown plasmids in addition to their chromosome. There has not been a good way to comprehensively look at these community members in detail, because, until lately, most bacterial research has been limited to using either culture-based methods or metagenomic approaches. Both of these methods have major limitations.

Previous Methods for Studying Antibiotic Resistance

Using the culture based method, bacterial isolates can be identified (Sengelov et al., 2003) and comprehensively studied. While bacterial culturing has allowed for the identification of many plasmids and ARG, and the fraction of culturable bacteria is

constantly expanding (Ferrari et al., 2008; Vartoukian et al., 2010), the method still has a very limited scope. Much of the microbial community is lost to analysis using this method because only a minor fraction of the microbiome can be cultured in a lab, approximately 1% (Vartoukian et al., 2010). This is a problem, especially in studies such as that done by Sengelov et al. (2003), where the level of antibiotic resistance in environmental samples is measured as the number of resistant bacterial isolates observed on a plate. Presumably, more bacteria carried the resistance but could not grow on the media provided. While the use of a control plate without antibiotics ensures that a researcher is calculating an accurate resistance ratio for culturable bacteria, this ratio of resistant to non-resistant bacteria could be different for the 99% of the microbiome not being analyzed.

The metagenomic approach bypasses the culturing problem by extracting total DNA directly from a sample and sequencing it. A larger fraction of the microbiome can then be studied. BLAST type searches against known databases, 16s rRNA identification, or PCR amplification (Munir et al., 2011) can be used on the extracted DNA. These can make it possible to identify the species, ARG, and even plasmid identifier genes present in such a dataset (Zhu et al., 2013). However, most extraction techniques fracture the DNA/genome making it difficult to attribute putative characteristics to specific bacterial taxa using these methods. Metagenomic assembly was developed to help with this.

Metagenomic assemblers are program that piece together overlapping, short DNA sequence reads into longer contiguous sequences (contigs). However, even after assembling metagenomic reads into longer contigs, any but the simplest bacterial communities can produce too many short contigs to identify which came from which species within the community. Improvements to assembly methods are made regularly, but high quality metagenomic assemblies can be difficult to achieve even when employing newer, more sophisticated methods (Howe et al., 2014). Highly complex bacterial communities such as those found in soil have proved especially difficult to assemble. When performing assemblies of soil metagenomes, Howe et al. (2014) found that only 10% of the sequences were sampled deeply enough to be assembled. This means that the majority of the genetic information in a complex sample like this is still lost to analysis. They also found that more than 60% of the predicted proteins in their

assemblies could not be annotated by comparison to known databases. This is because much of the environmental microbiome has never yet been characterized. It will continue to be difficult to characterize previously unknown bacterial species from fragmented assemblies.

The reliable identification of bacterial carriers of plasmids and replicon carriers of ARG using metagenomic approaches is also rare. This is because most reads and assembled contigs are too short to have both species identifier genes and ARG on the same stretch of DNA sequence and plasmid and chromosomal DNA should never be on the same stretch of DNA sequence. Determining whether ARG in a sample are located on a plasmid or chromosome is imperative if concerned about their mobility however. For instance, discovering which species a drug resistant plasmid has made its way into is relevant information when assessing an ARG's public health risk. While long read sequencing technologies such as those from PacBio and Oxford Nanopore have the potential to help with this issue (Koren & Phillippy, 2015), the methods are still in development and are cost prohibitive for many research projects. If cost was not an issue, long read sequencing could help with determining ARG location as it can aid in assembling whole replicons from simple metagenomic samples (Driscoll et al., 2017). However, plasmid and chromosomal DNA should still never be on either the same read or same contig using long read technology because they originated from different replicons. Long read technology thus would not link plasmids to the species they came from in the context of an environmental sample made up of many different species. Thus, neither culture based methods nor metagenomic approaches are sufficient for answering detailed questions about bacterial community members.

The Hi-C Method

The Hi-C genome conformation capture method is a relatively new approach which allows the study of chromosomal conformation in a cell's natural state by cross-linking DNA molecules in close physical proximity within the cell. Hi-C data thus reflects the spatial arrangement of DNA at the time of cross-linking. Hi-C's precursors, 3C and 4C, were originally developed for looking at chromosomal interactions within yeast and mammalian cells (Wit & Laat, 2012). The method has recently been applied to

prokaryotes and has proven useful for reconstructing individual species' genomes from bacterial communities (Beitel et al., 2014; Burton et al., 2014; Marbouty et al., 2014).

The Hi-C method is an elaborate DNA extraction method that produces DNA sequences where each end of the sequence originated from a different genetic location within the same cell. Each end of a sequenced Hi-C paired end read could thus come from different areas of the chromosome, from entirely different chromosomes in the case of eukaryotic cells, or plasmid and chromosome in the case of bacterial cells. To create Hi-C read sequences, a sample, such as a mixed bacterial community, is briefly treated with formaldehyde prior to cell lysis. The formaldehyde treatment is dilute and short enough to prevent significant DNA damage. It penetrates the intact cells and crosslinks DNA loci that are in close proximity within the cell (Orlando et al., 1997). The cells are then lysed and the recovered DNA is digested with a restriction enzyme. The digested ends are filled in and labeled with a biotin-labeled nucleotide. The still cross-linked restriction fragments are then diluted out and treated with ligase. As the DNA solution is dilute at this step, the ligase is more likely to ligate together two cross-linked restriction fragments than random DNA fragments within the solution. The crosslinks are then reversed on these now ligated restriction fragments. The DNA is purified and any unligated ends are digested to remove remaining biotin markers from sites that were not ligated. The DNA is purified again and sheared into fragments suitable for sequencing. Streptavidin is then used to isolate just the fragments with biotin labeled ligation junctions. These Hi-C, paired-end fragments, where each end came from a different restriction fragment, are then sequenced.

Figure 1.1 shows how these Hi-C read sequences are used to reconstruct individual genomes from a metagenomic dataset. The workflow for a Hi-C experiment involves splitting one sample to create two datasets. One dataset is the library of Hi-C reads as previously described. The other dataset is produced from a total DNA extraction of the same sample as the Hi-C reads were created from. The sequences from this total DNA extraction are shotgun sequenced. The resulting shotgun reads are overlapped and assembled into longer contigs (contiguous sequences) using a metagenomic assembler (Figure 1.1, Step 1). The Hi-C reads are then aligned to these contigs (Figure 1.1, Step 2). Each end of each Hi-C read pair is aligned separately, because each end of the read pair is

expected to align to different locations on a contig or to different contigs entirely. The number of Hi-C reads that link each contig pair is counted and stored in a matrix. Using this linkage data, the contigs are clustered by species (Figure 1.1, Step 3) since contigs from each species have more Hi-C read linkages among their own contigs than with contigs from other species.

The construction of not only chromosomes but also whole genomes including plasmids from a metagenomic sample could be revolutionary for the field of environmental microbiology. By identifying the genetic context of ARG in bacterial communities, this approach could determine if pathogenic bacteria or transmissible plasmids are carrying the ARG. It could make the characterization of environmental reservoirs of ARG more possible. This knowledge would greatly increase our ability to assess the risk associated with specific ARG and could inform decisions on waste management or the treatment of resistant infections in a medical scenario.

Previous Research and Project Objectives

Three proof of concept studies recently demonstrated that Hi-C could separate metagenomic datasets by species genomes when applied to microbial communities. Each of these studies was looking for an effective way to bin datasets of metagenomic contigs by species and demonstrated that Hi-C was a way to assemble more complete genomes from metagenomic samples than was previously possible. As Hi-C technology was previously used for looking at chromosomal interactions within cells, these studies also looked at 3D chromosomal and plasmid/chromosome interactions for some of the species in their samples.

Beitel et al. (2014) simulated the metagenomic assembly of a bacterial community of five species by breaking their reference genomes into read length fragments and assembling the synthesized reads into contigs. Of the five species, two were Gram positive, two were different strains of *E. coli*, and one contained two native plasmids. The group produced Hi-C data from a real cell culture of a mix of the five species and clustered the simulated assembled contigs by species. They successfully separated contigs by species although the two *E. coli* strains clustered with each other. However, it was not clear how many contigs were assigned to the wrong species using Hi-C linkages.

The group also observed the plasmids interacting definitively with the chromosome of the species that carried them.

Burton et al. (2014) applied Hi-C methodology to a community of yeast as well as to a large community of yeasts, nine bacterial species and an Archaea. Like Beitel et al. (2014), the research group used simulated metagenomic assemblies but showed that Hi-C could cluster by genome a metagenome consisting of both prokaryotes and eukaryotes. The group observed that only ~90% of the contigs in each simulated metagenomic assembly could be clustered using their Hi-C library (presumably because the remaining contigs did not contain the restriction site recognized by their restriction enzyme), but they did recover the genome of a bacterium containing two chromosomes and a plasmid. Again, the number of misassigned contigs was not made clear.

Marbouty et al. (2014) distinguished different species using Hi-C on a small bacterial community of three species, a larger community of yeast, and an environmental sample of river sediment which had been enriched for bacteria. Since this project was a proof of the Hi-C concept, for the bacterial and yeast communities grown in the lab, Hi-C reads were aligned to sections of reference genome (30 kb each) rather than to true assembled contigs. For the environmental sample, Marbouty et al. (2014) recovered 19 genomic clusters that each contained more than 1 Mb of genetic sequence. Eleven of these clusters could be identified by bacterial class and one appeared to carry plasmid identifiers.

While all these groups showed that in some instances Hi-C can assign a plasmid to its bacterial host in a metagenomic dataset, this ability was not explored in complex plasmid scenarios (Beitel et al., 2014; Burton et al., 2014; Marbouty et al., 2014). Moreover, simulated contigs were used except for the test on river water performed by Marbouty et al. (2014). When real contigs were used in that environmental scenario, many of the clusters did not contain enough genetic sequences for nearly a whole bacterial genome (1Mb). It would have been helpful to know how much DNA could be recovered in the other instances if a real metagenomic dataset were created. The number of cells used to build communities was also not given and none of these research groups published a working program for reproducing their cluster analysis of Hi-C data. So, while the research showed that species could be separated and plasmids could be detected

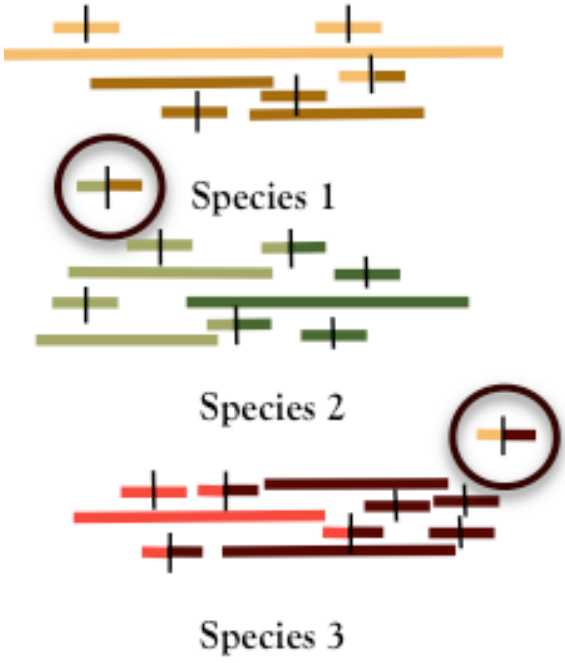
using Hi-C, it was not clear if this could be done easily or only in certain community circumstances. It was also not clear if the optimal cluster number could be determined when the number of species present was not already known. Scenarios where multiple species carried plasmids and the ways these plasmids interacted with the entire community were not tested. The placement of specific genes by Hi-C, necessary for the study of antibiotic resistance, was largely not discussed. Beitel et al. (2014) suggested that Hi-C may be useful for studying horizontal gene transfer but the previous published research has not classified Hi-C's capabilities or tested its limits of detection in several areas critical to plasmid and antibiotic resistance research in a metagenomic setting.

The purpose of this project was to test the limits of Hi-C when studying the genomes of simple microbial communities carrying plasmids and antibiotic resistance genes. Previously untested limits include the capacity of Hi-C to discriminate between closely related plasmids, to recover full genomes, to recover ARG and correctly assign them to species, to show if an ARG is found on a plasmid or chromosome, to correctly assign a plasmid to its host when the plasmid is present in multiple bacterial species and to detect a plasmid in a host when present at a low percentage of the community. To define these limits, the Hi-C method was applied to a series of 7 mock bacterial communities that were carefully chosen to test these critical parameters. Towards this end it became necessary to construct a bioinformatics pipeline capable of performing Hi-C data analysis.

Figure 1.1 Hi-C experiment workflow[1]

Step 1: Clean metagenomic reads and assemble into contigs

Step 2: Align each end of Hi-C reads individually to contigs



Step 3: Cluster contigs by species based on Hi-C linkages. Statistically there will be more Hi-C linkages within species than between species

[1] Sequence reads and contigs not shown to scale

Chapter 2: Methods

Growth of Bacterial Strains and Community Preparation

All strains used in this project were lab strains already present on site (Table 2.1). They were in the form of 20% glycerol stocks that had each been prepared from one bacterial colony. All stocks were stored at -70° C. To construct the bacterial communities, cultures of each Gram negative species were grown over night in Luria-Bertani (LB) broth (VWR). The only Gram positive organism, *Lactobacillus crispatus,* was grown in a richer MRS broth (Difco™ Lactobacilli MRS Broth, BD). The LB and MRS broths were prepared per their package specifications and autoclaved in glass bottles for sterilization. Culture and growth conditions for each species are listed in Table 2.1. Antibiotics were included in some cultures to ensure plasmid and ARG retention. Non-native plasmids, pB10 and pBP136, were carried by some strains and carried antibiotic resistance genes to infer plasmid presence and ensure that they were not lost during the culturing step.

Cell density was measured by visually counting cells in the cultures using a Petroff Hausser counting chamber. To impede cell division during the counting process, each bacterial culture was aliquoted into 2mL microcentrifuge tubes and centrifuged for 2 minutes at 13,000 g to pellet the cells. The supernatant was then removed and all the cell pellets for each species were combined and resuspended in 1x phosphate-buffered saline solution (PBS) to a final volume of 5mL. The PBS was used to hamper further cell division without causing cell death. A 100µL subsample of each 5mL culture was diluted 10x with more PBS and used for counting. The remainder of each culture was stored at 4°C to further prevent cell division while the cells were counted. Cell counts for each culture were performed in triplicate and averaged to calculate the number of cells in the culture (Tables 2.2 & 2.3). The counting chamber and cover slides were cleaned with 70% ethanol between each sample.

The communities were assembled on different days with Communities 1 & 2 grown and assembled first (Table 2.2) and the rest of the communities assembled on a later date (Table 2.3). To assemble the communities, differing amounts of each bacterial culture were added to 2mL microcentrifuge tubes to create the correct proportions of each

species based on cell density. The final volume of each community was approximately 1.3mL. Total cells present in each community varied but was kept within the order of 10^8-10^9.

Four replicates were assembled for each community. Immediately after community construction, a total DNA extraction was performed on one replicate of each. This was done using a GenElute™ Bacterial Genomic DNA extraction kit (Sigma). The optional Gram positive lysozyme extraction step included in the instructions was performed to ensure good cell lysis. The DNA was eluted into 200µL final volume of PCR grade water and stored at -20°C until sent for sequencing.

The remaining three replicates of each community were prepped for the Hi-C protocol by adding 37% formaldehyde directly to the PBS mixed culture to a final concentration of 1%. They were incubated for 20-30 minutes at room temperature and periodically swirled. The formaldehyde crosslinking reaction was quenched by the addition of glycine to a final concentration of 0.133M. The samples were swirled to mix and then incubated for another 20-30 minutes at room temperature. The samples were then spun down for 2 minutes at 13,000 g. The supernatant was removed and each pellet was rinsed with 800µL of PBS. This was to remove formaldehyde. The samples were spun down again for 2 minutes at 13,000 g. The supernatant was decanted and the cell pellets stored at -20°C.

Project Layout

The total DNA extractions from each community were sent to collaborators for shotgun sequencing. They were also sent 1 of the 3 Hi-C replicates for each community to complete the Hi-C prep using their proprietary Hi-C protocol and to perform the sequencing. These datasets were used for the development of the bioinformatics pipeline and to answer the research questions being asked of each community, and these are the results highlighted in Chapter 3. The remaining two replicates for each community were used for practice and development of the Hi-C method on site. This was done using previously published protocols (Beitel et al., 2014; Burton et al., 2014) that were combined and then modified to give the final protocol presented in Appendix A.

All samples were sequenced using Illumina Nextseq kits. The total DNA extractions were sequenced using paired end 150 kits for the shotgun reads. (This created

300bp of sequence per read.) The Hi-C samples were sequenced using paired end 75 kits. Longer reads are not needed for Hi-C libraries as only a short stretch of sequence is sufficient for aligning to the assembled contigs.

Bioinformatics Pipeline

An overview of the bioinformatics pipeline designed for Hi-C data analysis is shown in Figure 2.1. Details of the commands for running each program and the R script written to perform clustering of Hi-C data are presented in Appendix B.

Cleaning of Sequenced Reads

To remove superfluous reads from the metagenomic dataset, Super Deduper was used with default parameters. Super Deduper is an open source application (https://github.com/dstreett/Super-Deduper) and is designed to remove duplicate sequence reads introduced by PCR amplification prior to sequencing. These are present in most sequenced datasets and are known to hinder metagenomic assemblies (Petersen et al., 2015). Default parameters were used because files were in the form of paired end reads and that is the default mode for Super Deduper.

Flash2 was then used to merge overlapping paired-end reads. It is also an open source application (https://github.com/dstreett/FLASH2) and is a modified version of the original FLASH (Magoc & Salzbert, 2011). FLASH was designed to merge paired-end reads that were sequenced from DNA fragments shorter than the combined length of the reads. For example, when a 250bp sequence fragment is sequenced using a paired end 150 kit, each end is sequenced for 150bp. The resulting reads then overlap by 25bp. Merging these reads prior to metagenomic assembly results in a simpler dataset for the assembler and helps the assembly take less time and computational power. Flash2 was used with default parameters plus the additional options: -M 200 -O -C 70 -Q 20 (see Appendix B).

Sickle (Joshi & Fass, 2011; https://github.com/najoshi/sickle) was then used to quality trim both the merged and unmerged reads with a length threshold of 75 and a quality threshold of 20. Quality trimming of reads aids in assembly because reads produced from most modern sequencing technologies have progressively lower quality approaching the 3’-end and sometimes at the 5’-end as well. Trimming off these bases

increases the quality of the reads. Sickle will also discard reads based upon a length threshold if specified. This means that reads which have been quality trimmed to too short of a length get discarded. This capability was utilized because if a read is too short, it only makes for a larger dataset without adding much useful genetic information. These short, uninformative reads slow down assembly.

Metagenomic Assembly

The cleaned reads were overlapped and assembled into longer contiguous DNA sequences (contigs) using Spades 3.7.1 (Nurk et al., 2013) in meta mode with default parameters. Contigs longer than 500bp were kept for clustering.

Hi-C Read Cleaning

The Hi-C reads were cleaned (Figure 2.1) using HiCUP (Wingett et al., 2015), a set of scripts made available by Babraham Bioinformatics. As well as removing PCR duplicates, HiCUP sorts out valid Hi-C reads, which are those where each end of the paired end read aligns to different restriction fragments and where each end aligns to only one genetic location. In lieu of reference genomes, the assembled contigs were provided to HiCUP for its alignment step to demonstrate that this pipeline could be used with environmental samples where the species identities were not already known. One of the outputs of HiCUP is an alignment file with the extension ".bam". This file specifies to which contig(s) each end of every valid (Wingett et al., 2015; Servant et al., 2015) Hi-C read aligns. The bam file was turned into sam format using the "view" command from SAMtools (Li et al., 2009; https://github.com/samtools/samtools). The third column of this sam file was used as input for an R script written to do the clustering for this project.

Clustering Contigs Based on Hi-C Linkages

The first step in clustering was to build a matrix (Lajoie et al., 2015), the dimensions of which were equal to the number of contigs in the dataset. The matrix was filled in with counts of the number of Hi-C reads connecting each contig combination. The matrix was symmetrical. The matrix was turned into a correlation matrix and these values were used as distances in lieu of true Euclidean distances. Hierarchical clustering was performed using R's "hclust" function with the ward.D method. Each branch of the resulting dendrogram represents a contig. Visual analysis of the branching structure made the optimal number of clusters obvious in most cases and this number always

corresponded with the number of species present in that community. The dendrogram tree branches were cut at a height that yielded the desired amount of clusters. Alternatively, the optimal number of clusters could have been determined by creating a series of silhouette plots. However, this did not seem necessary in such simple communities as used here.

Contig Identification

The contigs within each cluster were then analyzed for species identity and the presence of any ARG and/or plasmid sequence. Since the species present in each community was already known, species identity was determined by aligning all contigs to a collection of reference genomes that represented the community composition. This was done using NCBI's blast tool with an e-value of 0.0001. In the case of multiple alignments, the match with the highest numerical value when the match percentage identity was multiplied by the length of the match was deemed to be the true identity of the contig. Each contig was identified by species or plasmid this way. The contigs were also compared to the Resfinder database to locate ARG in the community. The Resfinder database is a compilation of known ARG.

Visualization of Clustering

The hierarchically clustered contigs were further visualized using the R package "rgl". This made it possible to view each community's Hi-C interactions in 3-dimensional (3D) principle coordinate space. Principle coordinate analysis reduces the most important information within a large, many-dimensional dataset, such as this one, down to few enough dimensions to be visualized in a plot. While the 3D plots make comprehension of the data much easier, they do not show which contigs can be determined to be closely connected through the clustering algorithm. So while the 3D plots are easier to read, if one were trying to determine the genomic compartments of contigs whose species identity was not known, this would have to be determined using the hierarchical clustering pattern depicted in the dendrograms.

Table 2.1: Bacterial Species and Growth Conditions

Species (plasmid[1])	**Strain[2]**	**Media**	**Antibiotic**	**Temp** ℃
Escherichia coli (pB10)	MG1655	LB	Tc10[4]	37
Escherichia coli	MG1655	LB	-	37
Escherichia coli:Km	K12	LB	Km50[5]	37
Escherichia coli (pBP136)	BW25113	LB	Km50[5]	37
Acinetobacter baumannii (pB10)	ATCC17978[3]	LB	Tc10[4]	37
Acinetobacter baumannii	ATCC17978[3]	LB	-	37
Pseudomonas aeruginosa	PAO1	LB	-	37
Lactobacillus crispatus	MV-1A-US	MRS	-	37
Pseudomonas putida (pB10:Km)	UWC1	LB	Km50[5]	30

[1] *A. baumannii* also has two native plasmids: pAB1 & pAB2. They are not listed as they don't affect growth conditions.
[2] Some strains had been used in lab for some time. Mutations causing slight differences from reference genomes were thus expected.
[3] A variant of strain ATCC17978 used in our lab that no longer has *sul2* resistance gene
[4] Tetracycline added to a final concentration of 10ng/μL
[5] Kanamycin added to a final concentration of 50ng/μL

Table 2.2: Construction of Communities 1 and 2

Species (plasmid)	*Escherichia coli* (pB10)	*Escherichia coli*:Km	*Acinetobacter baumannii* (pAB1, pAB2)	*Pseudomonas aeruginosa*	*Lactobacillus crispatus*[1]	*Pseudomonas putida* (pB10:Km)	Total Cells in Community
Count Replicate							
1	207	173	39	25	116	57	
2	195	115	48	27	85	59	
3	182	131	41	32	110	59	
Average	194.7	139.7	42.7	28	103.7	58.3	
cells/mL	$2.4x10^9$	$1.8x10^9$	$8.5x10^9$	$5.6x10^9$	$1.3x10^8$	$7.3x10^8$	
Volume of culture used (mL)							
Comm1	0.8000		0.2280				$3.9x10^9$
Comm2		0.0743		0.0232	1.0000	0.1781	$5.2x10^8$

The calculation for cells/mL is average x dilution x 25 x 50,000 when counting 0.2mm squares and average x dilution x 16 x 25 x 50,000 for 0.05mm squares. Large squares were counted in all cases except for those counts shown in shading where smaller 0.05mm squares were used due to cell density.

[1] *L. crispatus* culture was so dilute as to not allow 10x dilution before counting. Dilution factor was removed from calculation

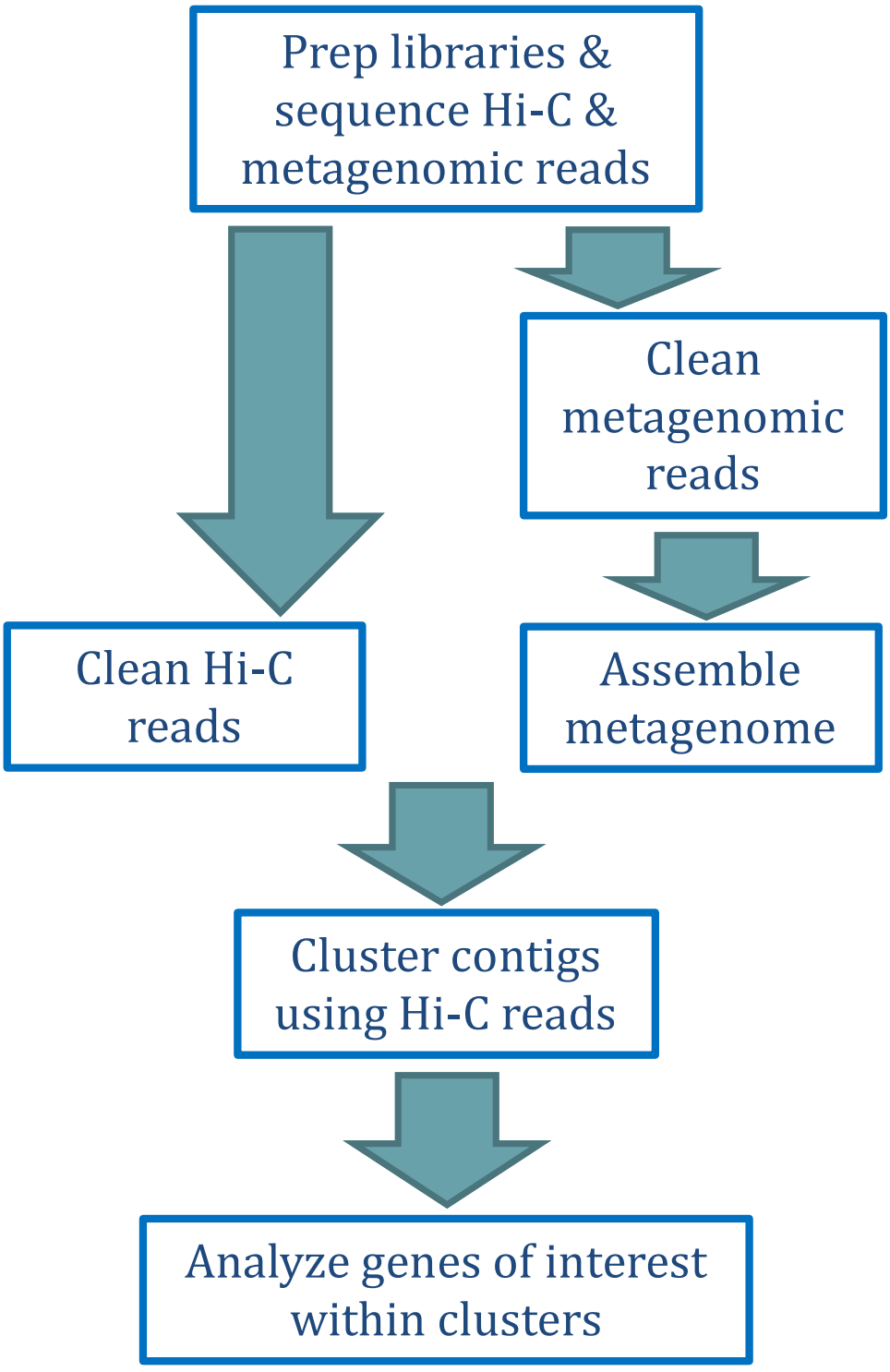
Table 2.3: Construction of Communities 3, 4a, 4b, 5a, and 5b

Species	*Escherichia coli*	*Escherichia coli*	*Escherichia coli*	*Acinetobacter baumannii*	*Pseudomonas aeruginosa*	*Lactobacillus crispatus*[1]	Total Cells in Community
Plasmid	(pB10)		(pBP136)	(pB10)			
Count Replicate							
1	30	29	22	36	83	72	
2	28	27	17	42	80	61	
3	26	20	30	44	84	75	
Average	28	25.3	23	40.7	82.3	69.3	
cells/mL	$5.6x10^9$	$5.1x10^9$	$4.6x10^9$	$8.1x10^9$	$1.7x10^{10}$	$8.7x10^8$	
Volume of culture used (mL)							
Comm3			0.1885	0.1066	0.0525	1.0000	$3.5x10^9$
Comm4a	0.1548			0.1066	0.0525	1.0000	$3.5x10^9$
Comm4b	0.0310			0.1920	0.0525	1.0000	$3.5x10^9$
Comm4c	0.0031			0.2111	0.0525	1.0000	$3.5x10^9$
Comm5a	0.0155	0.1540		0.1066	0.0525	1.0000	$3.5x10^9$
Comm5b	0.0015	0.1693		0.1066	0.0525	1.0000	$3.5x10^9$

The calculation for cells/mL is average x dilution x 25 x 50,000 when counting 0.2mm squares and average x dilution x 16 x 25 x 50,000 for 0.05mm squares. All cultures shown had been diluted 10x.

[1] *L. crispatus* culture was so dilute as to require counting of 0.2mm squares. The smaller 0.05mm squares were used for all other cultures.

Figure 2.1: Bioinformatics pipeline



Chapter 3: Experimental Plan and Results

The Hi-C method has been proven useful for reconstructing individual species' genomes from bacterial communities (Beitel et al., 2014; Burton et al., 2014; Marbouty et al., 2014). The previous published research has not classified Hi-C's capabilities or tested its limits of detection in several areas critical to plasmid and antibiotic resistance research in a metagenomic setting however. Untested limits include the capacity of Hi-C to discriminate between closely related plasmids, to separate species present at low percentages of a community, to recover ARG and correctly assign them by species, to show if an ARG is found on a plasmid or chromosome, and to correctly assign a plasmid when the plasmid is present in multiple bacterial species. The Hi-C method was applied to a series of seven bacterial communities that were carefully constructed to test these critical parameters. They were analyzed using the bioinformatics pipeline constructed to perform Hi-C cluster analysis.

Description of the Communities

The community composition of each of the seven assays was designed to test the limits of detection for the Hi-C method while also serving as low complexity datasets for pipeline development (Table 3.1). For example, Community 1 was composed of 50% *Escherichia coli* carrying the pB10 plasmid and 50% *Acinetobacter baumannii* carrying its native plasmids: pAB1 and pAB2. This first community was designed to be a very simple dataset on which to practice Hi-C genome clustering while the pipeline was being designed. Despite having only two species, it contained three plasmids, and these served to verify, early in the project, whether plasmids, including those of different sizes, could be detected using the developed bioinformatics pipeline.

The six remaining assays all contained *Pseudomonas aeruginosa* and *Lactobacillus crispatus.* These species served as background noise and introduced more complexity into the communities by increasing the total number of species to four. They were included to test if the questions being asked could be answered in communities slightly more complex than two species as in Community 1. The background species were each present at 25% of their communities in all cases. The addition of *L. crispatus*

also brought a Gram positive organism into the mix to ascertain whether the lab protocol worked equally well on Gram positive and Gram negative organisms.

In addition to *P. aeruginosa* and *L. crispatus*, Community 2 was composed of 25% *E. coli* with a kanamycin (Km) resistance gene carried on the chromosome and 25% *Pseudomonas putida* with the same Km resistance gene present on the plasmid pB10. This assay was designed to test if the same ARG could be detected on multiple replicons within one community or if having the same gene present in multiple locations would cause confusion during assembly. Since the *P. putida* shared some sequence similarity with *P. aeruginosa*, this assay also tested whether two related species were distinguishable using Hi-C.

Community 3 was composed of 25% *E. coli* carrying the IncP-1 plasmid, pBP136, and 25% *A. baumannii* carrying the IncP-1 plasmid, pB10, in addition to *P. aeruginosa* and *L. crispatus*. Coming from the same incompatibility group, these two plasmids bore moderate sequence identity over 35,097 nucleotides (76.3% identity). The assay was designed to test if these similar plasmids could be assembled and assigned to the species they came from or if sequence similarity would cause confusion during the assembly and genome clustering steps. Since *A. baumannii* contained two native plasmids, pAB1 and pAB2, in addition to pB10, it also tested whether as many as three different plasmids could be detected in a single species.

In addition to the two background species, Community 4a consisted of 25% *E. coli* and 25% *A. baumannii,* both carrying the plasmid pB10. This assay was designed to test if the same plasmid could be reliably assigned to multiple species within a community. The assay mimicked a possible plasmid transfer scenario, where a plasmid has spread to multiple species within a community. Community 4b took this concept a step further with *A. baumannii* carrying pB10 making up 45% of the community and *E. coli* with pB10 making up only 5% of the community. This assay tested whether a species would cluster individually even when present at only a small percentage of the total community. Many species present at small percentages of the total sample is more likely in most environmental settings than a few species present at high percentages of the community (Howe et al., 2014). Community 4b also tested whether the plasmid would show up in a species present at such a low percentage or whether the plasmid would have so many

more Hi-C reads linking it to the species present in a higher percentage that it would only show up in that cluster.

Community 5a mimicked a different possible plasmid transfer scenario. In addition to the background species, there was 25% *A. baumannii* with plasmid pB10, 2.5% *E. coli* with plasmid pB10, and 22.5% the same strain of *E. coli* without any plasmid. This community mimicked a realistic environmental scenario where the plasmid pB10 had been transferred from *A. baumannii* to *E. coli* via conjugation, but had only been acquired by 10% of the *E. coli* so far. The plasmid was expected to show up in the *A. baumannii* cluster. However, this assay tested whether it would also show up in the *E. coli* cluster at a level high enough above the background noise that it could be confidently determined to be present in both species. Community 5b took this scenario a step further and consisted of the background species, 25% *A. baumannii* with plasmid pB10, 0.25% *E. coli* with plasmid pB10, and 24.75% *E. coli* without any plasmid. It again mimicked a plasmid transfer scenario where the plasmid pB10 had been transferred from *A. baumannii* to *E. coli*, but where it had only been acquired by 1% of the *E. coli* so far. This assay was designed to further test the limits of plasmid detection using Hi-C.

Read Cleaning and Assembly Results

The sequenced metagenomic shotgun reads and the Hi-C reads were processed prior to further analysis. Cleaning metagenomic shotgun reads entailed three steps: removing duplicates introduced by PCR amplification, merging overlapping reads, and trimming off low quality read ends (Appendix B). The cleaning results for the metagenomic shotgun reads are given in Table 3.2.

Communities 1 and 2 both began the cleaning process with close to 75 million reads. The remaining communities were sequenced in a separate run and our collaborators sequenced neither the Hi-C nor the metagenomic reads as deeply as the first two communities. Of these later communities, some only had approximately 1/10 the number of reads as the first two communities, which was not sufficient to support good assemblies on these later communities. The fractured contigs combined with the lower number of Hi-C reads prevented informative downstream analysis as well. More of the same samples were thus sequenced in an additional run, creating two metagenomic and two Hi-C datasets for each of Communities 3, 4a, 4b, 5a, and 5b. The results were

concatenated to give one dataset for each sample and this total read count is given in Tables 3.2 and 3.3. Concatenating the data like this did not appear to cause any analysis problems when results were compared to those from Communities 1 and 2 which came from single sequence runs.

The metagenomic shotgun reads were sorted using Super Deduper to remove PCR duplicates introduced by the PCR amplification required prior to sequencing. Table 3.2 shows that the later communities had fewer starting reads compared to Communities 1 and 2. However Super Deduper removed a high percentage of duplicate reads in Communities 1 and 2 (20-30%). Communities 1 and 2 appear to have been over sequenced for their community complexity.

Flash2 was used to merge paired-end reads that were sequenced from DNA fragments shorter than the combined length of the reads. Communities 1 and 2 had a very high percentage of reads overlapped by Flash2 (~80%). This improved the speed and quality of their assemblies when compared to the rest of the samples where only ~30% of the reads overlapped.

The final number of cleaned reads after quality trimming by Sickle is shown in the last line of Table 3.2. While this number was close to 50 million for all samples, a higher percentage of these were longer, overlapped reads for Communities 1 and 2 and more of these were shorter, non-overlapped reads for the rest of the communities.

Spades -meta was the assembler used to overlap reads into longer contiguous sequences (contigs). The number of contigs produced by Spades -meta for each sample is shown in the first row of Table 3.3. Communities 1 and 2 both assembled into fewer contigs with a higher mean length when compared to the rest of the samples. Only contigs longer than 500bp were kept for clustering. The length of 500bp was chosen as an arbitrary cutoff point to remove contigs too short to provide meaningful results. The average number of contigs longer than 500bp was approximately 400 (Table 3.3). The total additive, assembled length of all contigs longer than 500bp is shown in Table 3.3. This value can be compared to the expected assembled length on the line below. Expected assembled length for each community was calculated by adding together the length of the reference genomes for all replicons present. For all samples these two values were similar enough to suggest that the sequence depth as well as cleaning and

assembly pipeline were adequate to recover most of the genetic information present in the sample.

The Hi-C reads were both cleaned and aligned to the assembled contigs by the program HiCUP. Three statistics were of the most interest during the cleaning of the Hi-C reads: the number of read pairs going into HiCUP, the number of valid Hi-C reads as determined by HiCUP (Wingett et al., 2015), and the number of those valid reads that were unique. Reads are valid if each end of the read aligned to different restriction fragments and each end aligned to only one genomic location. Valid reads are unique if they are not PCR duplicates created by the sequencing process. The numbers of raw, starting reads varied between samples but in all cases were greatly reduced during cleaning (Table 3.3). This was not a surprise because Hi-C datasets are inherently noisy (Beitel et al., 2014). For this project, it was observed that one quarter to one third of the sequenced read pairs in a reasonably good Hi-C dataset represented valid Hi-C associations as determined by HiCUP (Wingett et al., 2015). Removing PCR duplicates from these valid read pairs to get the number of both valid and unique Hi-C read pairs also decreased the variation in number of Hi-C reads among the samples despite the samples being sequenced to different depths. These results suggest that the Hi-C datasets can be over sequenced more easily than anticipated. The numbers of both raw and cleaned Hi-C reads are higher for Communities 4b and 5b when compared to the other samples (Table 3.3). These communities were sequenced slightly deeper to increase the chances of picking up community members that were present at lower concentrations than in the other communities.

Not all of the contigs over 500bp aligned to one of the reference genomes (Table 3.3). In addition, not all contigs were detected during the Hi-C clustering, meaning that some didn't have even one Hi-C read aligned to them. There was a high overlap in those contigs that did not align to a reference genome and those not detected by Hi-C. Also, those not detected by Hi-C were short contigs quite close to the 500bp cutoff mark. These small contigs were most likely either misassemblies (artifacts of the assembly process), or small repeats that neither the assembler nor Hi-C could assign uniquely to one genomic location. In some communities a few slightly longer contigs didn't align to a

reference genome but were still detected by Hi-C read alignments. When present they were always found in the *L. crispatus* genome cluster (1-4 contigs each time).

Description of Cluster Plots

Principle coordinate plots and hierarchical clustering were used to visualize how well Hi-C separated the contigs by species (Figure 3.1-3.7). In the 3 dimensional (3D), principle coordinate plots, each of the dots represents one contig. The dots are colored by true species identity as determined by alignment to reference genomes. These plots give a good visualization of the species separation and plasmid placement achieved by Hi-C for each community. The dendrogram following each 3D plot shows the hierarchical clustering that determined genome identity as determined by Hi-C. The dendrogram leaves each represent one contig and are also colored by true species identity as determined by alignment to reference genomes. The table following each dendrogram gives the same information in numerical form. Since hierarchical clustering cannot assign the same point to two clusters, in the case where the same plasmid was present in multiple locations, the dendrograms show that the plasmid was necessarily placed into only one genome cluster.

Clustering Results for Each Community

The two species in Community 1 separate quite dramatically on one axis (Figure 3.1.a). As the plasmids assembled to nearly full length in these samples it was quite easy to locate them within their species clusters. Plasmids pAB1 and pAB2 were one contig each and clustered neatly within the *A. baumannii* cluster because of plasmid-chromosome Hi-C linkages. Two contigs were identified as the larger plasmid, pB10, and both contigs can be seen clustering quite near the middle of the *E. coli* cluster. This assay verified that the bioinformatics pipeline worked effectively (Figures 3.1.b & 3.1.c), at least in simple communities, and that plasmids as well as their species of origin could be identified using the Hi-C method and our analysis pipeline.

As can be seen in the 3D depiction of Community 2 (Figure 3.2.a), good species separation was also achieved with four species. The two species included as noise, *P. aeruginosa* and *L. crispatus*, did not seem to affect the quality of clustering. They also demonstrated that the Hi-C protocol was effective for both Gram positive and Gram

negative organisms. Gram positive *L. crispatus* does appear to cluster less tightly than the Gram negative organisms. This could be due to *L. crispatus* having a smaller chromosome than the other species and them swamping the dataset. Despite their close relationship, *P. putida* and *P. aeruginosa* achieved good genomic separation and verified that our pipeline can separate species of the same genus. The orange *P. putida* contig seen in the purple *P. aeruginosa* cluster is a somewhat short contig (3379bp) that aligned to both the *P. putida* and *P. aeruginosa* reference genomes with very similar alignment scores (Fig 3.2.a). It had a very slightly higher score for the *P. putida* alignment and that is why the pipeline determined its identity to be *P. putida* even though hierarchical clustering (Figure 3.2.b) determined it to be part of the *P. aeruginosa* genome. It is not surprising that these two species have a stretch of DNA sequence in common since they are from the same genus. Of the two blue *E. coli* contigs in the orange *P. putida* cluster, one represents a small contig that aligned with very similar scores to both the *E. coli* and *P. putida* reference genomes. It was a short sequence that they happened to have in common. The other represents the only contig containing the Km resistance gene that was known to be on the *E. coli* chromosome as well as on plasmid pB10 in *P. putida*. This contig was the sixth longest contig assembled in this community so it was much longer than just this shared gene. Over half the length of this contig aligned to the *E. coli* reference genome. After that came the Km gene and approximately three quarters of the pBP10 plasmid. The assembler appeared to have merged plasmid pBP10 with a portion of the *E. coli* chromosome due to them having the Km gene in common. This contig along with the ARG was assigned to *P. putida* rather than *E. coli,* which made up most of the contig. As plasmids can have multiple copies within one cell, the higher copy number of pB10 versus *E. coli* chromosome appeared to have caused more Hi-C read linkages to pB10's host for this contig.

Community 3 was designed to test the limits of the method by including two plasmids that shared regions of nearly identical sequence. There were not major misassembly problems; however, the plasmid contigs were much more fragmented than in the other assays. While the species still showed good separation by hierarchical clustering, stretches of similar plasmid sequence made the separation of genomic clusters less visually obvious with principle coordinate analysis compared to the other

communities (Figure 3.3). The two species containing one of the plasmids, *A. baumannii* and *E. coli*, appeared together on the same plane in 3D space with a cloud of plasmid between them (Figure 3.3.a). For the most part, the pBP136 contigs were closer to *E. coli* and the pB10 contigs were closer to *A. baumannii* as they should be based on community composition. Hi-C appears to produce so many more plasmid-plasmid interactions than plasmid-chromosome interactions that in the case of similar plasmids the analysis may tend to give similar plasmids their own joint cluster between their hosts rather than clustering them within their host clusters as is seen otherwise. The majority of the contigs for both plasmids were determined to be in *A. baumannii* (Figure 3.3.c) as hierarchical clustering cannot assign a point to multiple clusters.

Community 4a again showed good species separation (Figure 3.4.a). The plasmids pAB1 (black) and pAB2 (gray) can just be seen hidden in the middle of the *A. baumannii* cluster. The white contigs seen in the *L. crispatus* cluster are two of the contigs mentioned earlier that did not align to any of the reference genomes but cluster quite definitively with *L. crispatus*. This assay caused no assembly problems so each plasmid assembled as one contig making their placement easier to determine. The question being asked with this assay was whether the plasmid pB10, which was present in both *A. baumannii* and *E. coli,* could be detected in both using Hi-C. Though the hierarchical clustering places pB10 within the *E. coli* genome cluster (Figure 3.4.b and 3.4.c), it is nearly halfway between the two species' clusters in the 3D plot (Figure 3.4.a).

Community 4b, which was a variation on 4a, looked quite similar other than the placement of pB10, which was now located somewhat closer to *A. baumannii* than *E. coli* (Figure 3.5.a). As expected, pB10 was determined to be part of the *A. baumannii* genome by the hierarchical clustering (Figure 3.5.b & 3.5.c) since *A. baumannii* was present at 45% of the community in this assay while *E. coli* was only present at 5%. What was somewhat surprising was that pB10 could still be seen so clearly to be between the two species rather than clustering in closer proximity to *A. baumannii*.

Community 5a is the community where only 10% of the *E. coli* contained pB10 while all *A. baumannii* carried it. Plasmid pB10 clustered much closer to *A. baumannii* than *E.coli* as expected (Figure 3.6.a). While the other species are positioned further back, *E. coli* and *A. baumannii* are again found on the same plane in the 3D plot with

pB10 placed between them. Observing a plasmid placement like this, rather than directly within the species cluster, seems to be the way to determine that a plasmid is present in both species to some extent rather than just in one species cluster as the hierarchical clustering data would make it appear (Figures 3.6.b & 3.6.c). *A. baumannii* was chosen for pB10's cluster placement, because pB10 had significantly more interactions there. In contrast to pB10, native *A. baumannii* plasmids pAB1 and pAB2 are clustered so deeply within the *A. bamannii* cluster that they cannot be seen in the 3D plot. Instead of being a tight cluster like *P. aeruginosa*, the *E. coli* cluster is distorted, perhaps from being stretched towards *A. baumannii* and pB10. This is presumably because its shared interactions with pB10 prevented it from forming a tight cluster only with itself. In Figure 3.6.a, the blue *E. coli* contig floating out in space is again a very short contig that bore sequence similarity with multiple species but had a slightly higher alignment score with *E. coli* than the other species.

The visual results for Community 5b, where the frequency of *E. coli* carrying pB10 was only 1% (Figure 3.7.a) are quite similar to 5a. The plasmid pB10 clustered a bit closer to *A. baumannii* in this assay and was still assigned to that genome cluster by the hierarchical clustering (Figure 3.7.b & 3.7.c). *A. baumannii* also clustered more tightly, probably because pB10 is now pulling it less towards *E. coli*. Native plasmids, pAB1 and pAB2, are again not visible because they are clustered so tightly within the *A. baumannii* cluster. *E. coli* appears to be even more elongated in 5b, with some of its contigs in closer proximity to pB10. Several contigs (white) that didn't align to the reference genome can again be seen within the *L. crispatus* cluster. While Figure 3.7.a might not suggest that pB10 is found in any species other than *A. baumannii*, pB10 is placed further outside of the cluster than is seen in communities where it was only located in one species, such as in Community 1 (Figure 3.1.a). Plasmid pB10 was only carried by *E. coli* in that assay and is placed within its species cluster rather than on the outskirts as in 5b.

ARG Placement

While each assay tested some specific question or questions about the ability of the Hi-C method to separate species and place plasmids, one of the overarching questions for the whole project was how well Hi-C could correctly assign ARG to the species and replicon that carried them. Table 3.4 lists the species and plasmids used in this study that

were known to carry some ARG. The ARG that each replicon carried were determined by comparing all reference genomes to the Resfinder database (Chapter 2). The ARG carried by each replicon are marked. All the ARG carried by plasmids shared 100% gene identity with those in the ResFinder database. Some ARG from the hosts had only 96.8 - 99.9% gene identity with ResFinder genes: fosA, aph(3')-IIb, and blaOXA-50 in *P. aeruginosa* and blaOXA-180 and blaADC-25 in *A. baumannii*. Most identities were above 99% but prior research in the lab has not confirmed whether any of these genes conferred functional resistances or not. The *A. baumannii* reference genome also had a sulphonamide resistance gene: sul2. This is not included in Table 3.4 because from previous research it was known to have been lost following the acquisition of pB10. Neither of our *A. baumannii* variants contained sul2 anymore. *P. putida* on the other hand showed an ARG, catA1, that was not in the reference genome. However, the strain used, *P. putida* UWC1, is an unsequenced derivative of *P. putida* KT2440 that has evolved spontaneous resistances and possibly picked up some catabolic activity from a plasmid (McClure et al., 1989). KT2440 was the closest reference genome available and is nearly the same but not an exact match. Given this information, it seemed quite likely that the local variant did carry this gene, which is why it was included in the table of known resistances. Since the catA1 ARG was present at 99.7% gene identity rather than 100%, the gene may not yet confer a functional resistance and that could be why it has not been documented.

For each community, all contigs longer than 500bp were analyzed by Resfinder to identify which contigs carried ARG. This information was used to determine which species clusters these ARG had been assigned to as well as which replicon the ARG was recovered on, i.e. plasmid or chromosome. In all cases, all ARG were assigned to the expected species clusters and to the correct replicon within those clusters, i.e. plasmid or chromosome (Table 3.5). For example, in Community 1, all of the known pB10 ARG were found on a pB10 contig within the *E. coli*/pB10 cluster.

There are a few cases that should be noted, however. The first was in Community 2 where the same Km resistance gene was present on both pB10 and the *E. coli* chromosome. The gene was assigned to the *P. putida*/pB10 cluster. It did not show up at all in the *E. coli* cluster because, as previously mentioned, this gene caused problems at

the assembly level: a contig was assembled with both *E. coli* and pB10 sequence joined together by the Km gene. This contig was likely assigned to *P. putida* rather than *E. coli* because of pB10's multiple copy number. A higher copy number would mean a higher probability of Hi-C linkages linking it to *P. putida* than of *E. coli* intra-chromosomal linkages. The second case was in Community 3 where similar plasmids pBP136 and pB10 almost formed a small cluster of their own. For the most part both plasmids were assigned to *A. baumannii*. Even though contigs from both plasmids were found in the *A. baumannii* cluster, pB10 ARG were only found on pB10 contigs. The third case was in Community 4a where both *E. coli* and *A. baumannii* carried pB10 at the same percentages of the population. Plasmid pB10 was somewhat arbitrarily assigned to *E. coli* by the clustering algorithm and the ARG were assigned to that cluster along with it as was expected. In the rest of the communities, pB10 along with its ARG was always assigned to the species that carried it at the highest percentage of the community.

Table 3.1: Community Composition[1] of the Seven Assays Designed to Define the Limits of the Hi-C Method

Species (plasmid)	**1**	**2**	**3**	**4a**	**4b**	**5a**	**5b**
Escherichia coli (pB10)	50			25	5	2.5	0.25
Escherichia coli						22.5	24.75
Escherichia coli:Km		25					
Escherichia coli (pBP136)			25				
Acinetobacter baumannii[2] (pAB1, pAB2, pB10)			25	25	45	25	25
Acinetobacter baumannii[2] (pAB1, pAB2)	50						
Pseudomonas aeruginosa		25	25	25	25	25	25
Lactobacillus crispatus		25	25	25	25	25	25
Pseudomonas putida (pB10:Km)		25					

[1] All numbers given as percentage of the entire community

[2] Plasmids pAB1 and pAB2 are native to *A. baumannii*

Table 3.2: Cleaning Statistics for Shotgun Metagenomic Reads

	Comm1	**Comm2**	**Comm3**	**Comm4a**	**Comm4b**	**Comm5a**	**Comm5b**
Super Deduper							
Total read pairs	74,007,384	74,760,777	33,582,514	25,496,938	42,787,661	34,293,216	39,176,874
Duplicates	22,070,260	14,491,886	3,266,849	1,971,431	3,940,845	2,950,445	2,506,125
% removed	29.82	19.38	9.73	7.73	9.21	8.6	6.4
Flash2							
Total read pairs	51,937,124	60,268,891	30,315,665	23,525,507	38,846,816	31,342,771	36,670,749
Discarded	1215	4587	4054	2705	4442	3461	3191
% discarded	0.000	0.010	0.010	0.010	0.010	0.010	0.010
% combined	78.93	78.62	35.7	23.61	41.71	30.9	26.81
Sickle PE							
Input pairs	10,945,492	12,882,454	19,491,465	17,969,106	22,641,183	21,654,721	26,838,795
Kept pairs	9,884,934	11,298,829	18,011,150	16,773,246	21,114,161	20,173,992	24,156,297
Single records kept	674,953	1,075,249	1,276,021	1,044,691	1,311,393	1,273,564	2,279,565
Pairs discarded	385,605	508,376	204,294	151,169	215,629	207,165	402,933
% pairs discarded	3.52	3.95	1.05	0.84	0.95	0.96	1.50
Sickle SE							
Single records	40,990,417	47,381,850	10,820,146	5,553,696	16,201,191	9,684,589	9,828,763
Single reads kept	31,891,868	37,565,594	10,485,629	5,250,686	15,711,271	9,361,212	9,530,254
% kept	77.80	79.28	96.91	94.54	96.98	96.66	96.96
Assembly Reads[1]	52,336,689	61,238,501	47,783,950	39,841,869	59,250,986	50,982,760	60,122,413

[1] Total reads going into assembly rather than read pairs

Table 3.3: Assembly and Clustering Statistics

	Comm1	Comm2	Comm3	Comm4a	Comm4b	Comm5a	Comm5b
Spades Assembly							
Number of Contigs	549	1,452	3,462	3,179	3,308	2,911	3,094
1st quartile	63	69	58	58	57	57	59
Median length	88	106	73	69	66	67	71
Mean length	15,502	13,333	4,999	5,441	5,221	5,937	5,577
3rd quartile	211	2578	194	102	91	98	96
Max length	766,980	571,663	914,504	648,482	1,043,942	655,756	655,420
Contigs > 10kb	81	258	297	245	206	214	215
Contigs > 500bp	121	490	649	474	411	403	403
Total Assembled Length (Mb)	8.47	19l26	17.04	17.08	17.04	17.08	17.04
Total Expected Length (Mb)	8.72	19.00	17.27	17.29	17.29	17.29	17.29
HiCUP							
Raw Hi-C read pairs	16,087,327	21,535,332	22,048,451	22,683,831	34,390,863	27,350,848	82,124,574
Valid Hi-C pairs	3,029,349	9,044,708	6,833,465	6,661,693	12,212,963	7,990,584	28,363,039
Valid & Unique Hi-C Pairs	2,952,352	8,044,553	6,067,993	5,911,677	11,793,838	7,687,469	12,229,967
Ratio of Hi-C pairs to contigs	24,399.6	16,417.5	9,349.8	12,471.9	28,695.5	19,075.6	30,347.3
Contigs that align to references[1]	121	487	634	470	405	397	397
Contigs not detected by Hi-C[2]	2	16	24	12	17	12	7

[1] Alignments determined by NCBI blast
[2] Contigs that no Hi-C read pairs aligned to

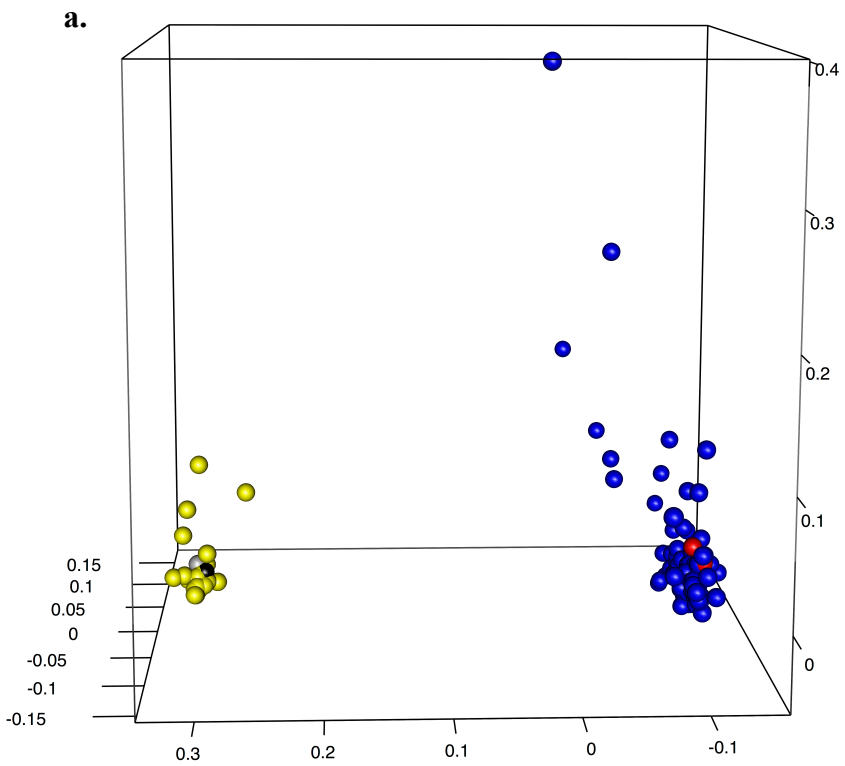
Table 3.4: Expected Antibiotic Resistance Genes by Replicon

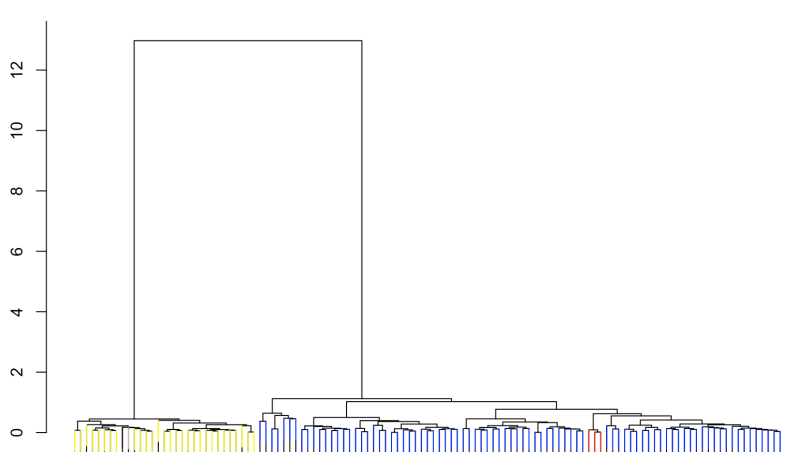
Expected Resistances[1]	blaPAO	catB7	fosA	blaOXA-50	aph(3')-IIb	aph(3') Ia	catA1	tet(A)	blaOXA-2	sul1	strB	strA	blaOXA-180	blaADC-25
A. baumannii													x	x
P. aeruginosa	x	x	x	x	x									
pB10								x	x	x	x	x		
pB10:Km						x		x	x	x	x	x		
E. coli:Km						x								
P. putida							x							

[1] Resistances determined by comparing reference genomes to Resfinder database (Zankari et al., 2012)

Table 3.5: Genomic Cluster Each Antibiotic Resistance Gene was Assigned to Using Hi-C

Resistance Genes	blaPAO	catB7	fosA	blaOXA-50	aph(3')-IIb	aph(3') Ia	catA1	tet(A)	blaOXA-2	sul1	strB	strA	blaOXA-180	blaADC-25
Comm1														
E. coli /pB10								x	x	x	x	x		
A. baumannii													x	x
Comm2														
P. aeruginosa	x	x	x	x	x									
P. putida /pB10						x	x	x	x	x	x	x		
Comm3														
P. aeruginosa	x	x	x	x	x									
A. baumannii/ pB10 /pBP136								x	x	x	x	x	x	x
Comm4a														
P. aeruginosa	x	x	x	x	x									
A. baumannii													x	x
E. coli /pB10								x	x	x	x	x		
Comm4b														
P. aeruginosa	x	x	x	x	x									
A. baumannii /pB10								x	x	x	x	x	x	x
Comm5a														
P. aeruginosa	x	x	x	x	x									
A. baumannii /pB10								x	x	x	x	x	x	x
Comm5b														
P. aeruginosa	x	x	x	x	x									
A. baumannii /pB10								x	x	x	x	x	x	x





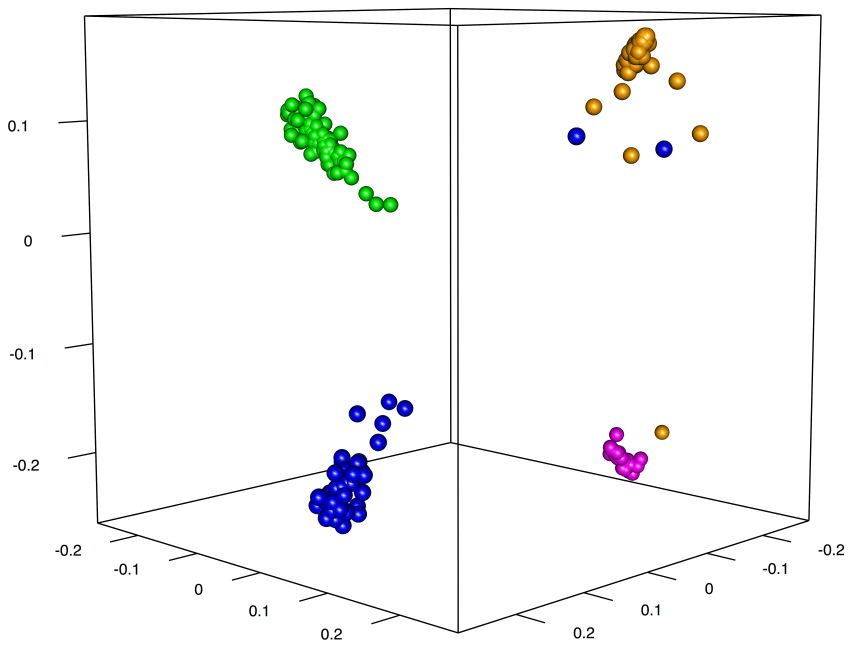
c.

		1	**2**
A. baumannii (50%)		29	0
pAB1		1	0
pAB2		1	0
E. coli (50%)		0	85
pB10		0	3

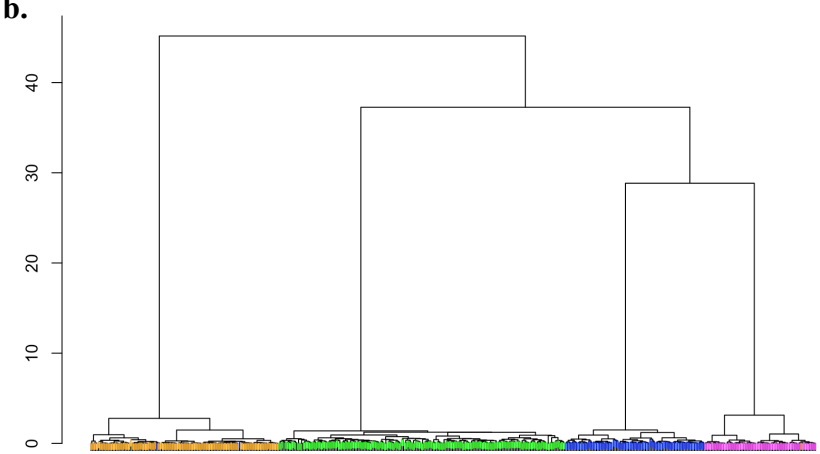
Figure 3.1: Contigs clustered based on Hi-C read linkages for Community 1, which had two plasmid containing bacterial species (**a.**) PCoA plot (**b.**) Hierarchical clustering dendrogram (**c.**) Number of contigs[1] in each cluster for each species[2]

[1] Length of contigs varies
[2] Species determined by alignment of contigs to reference genomes

a.



b.



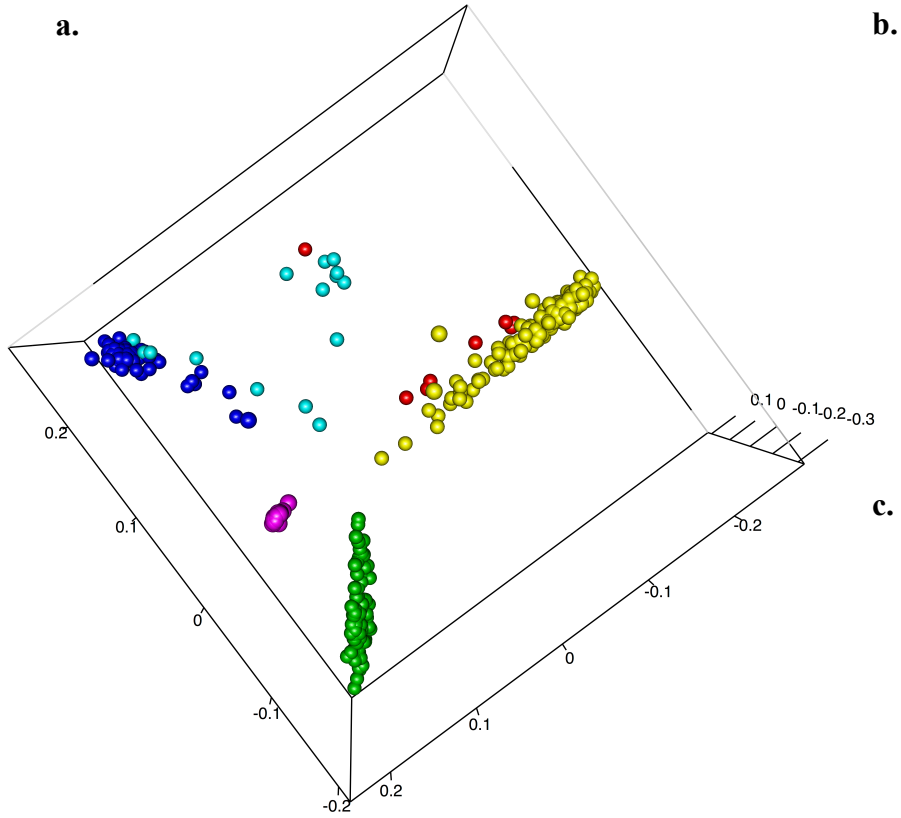
c.

		1	**2**	**3**	**4**
P. putida (25%)		1	0	0	121
pB10 [3,4]		0	0	0	0
P. aeruginosa (25%)		72	0	0	0
L. crispatus (25%)		0	184	0	0
E. coli [4] (25%)		0	0	91	2

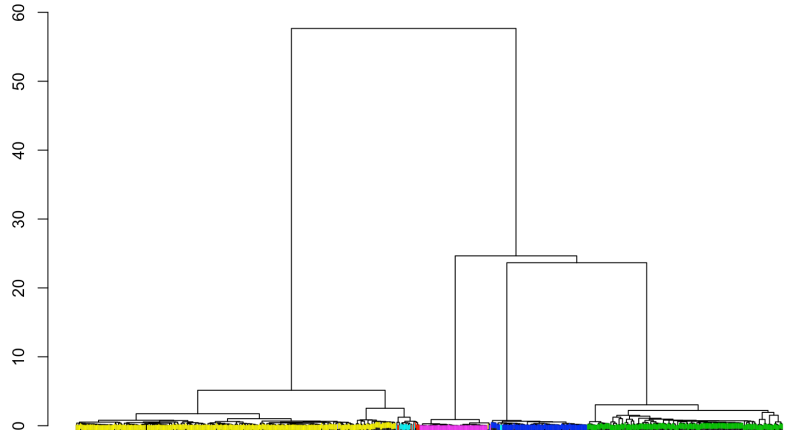
Figure 3.2: Contigs clustered based on Hi-C read linkages for Community 2, which had four bacterial species including two of the same genus and the same Km resistance gene on a plasmid and a chromosome (**a.**) PCoA plot (**b.**) Hierarchical clustering dendrogram (**c.**) Number of contigs[1] in each cluster for each species[2]

[1] Length of contigs varies
[2] Species determined by alignment of contigs to reference genomes
[3] Plasmid pB10 was combined with an *E. coli* contig in this analysis
[4] *E. coli* and pB10 carried the same Km resistance gene

a.



b.

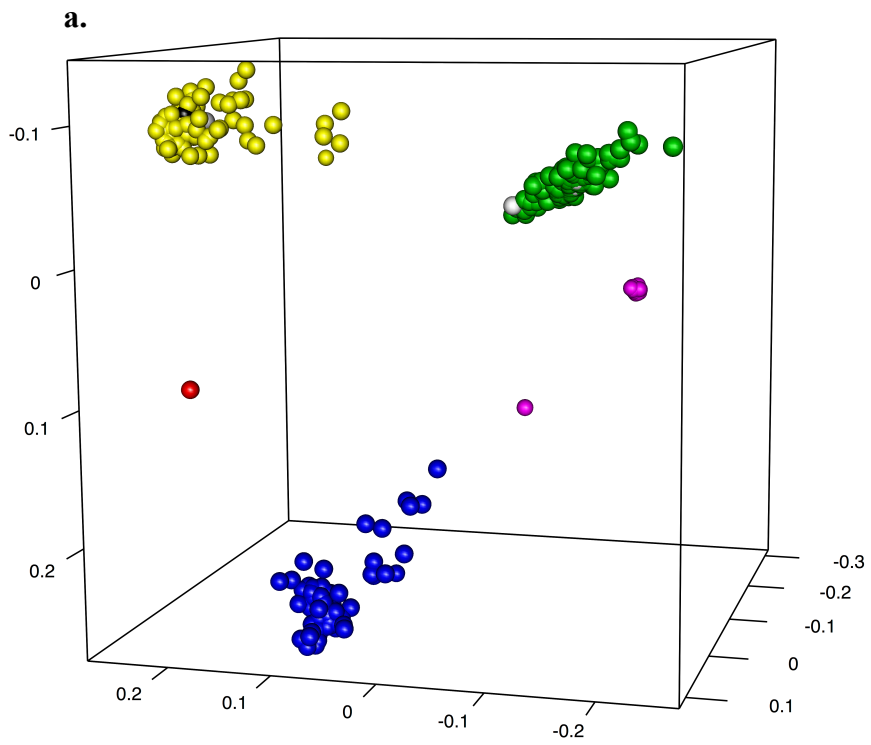


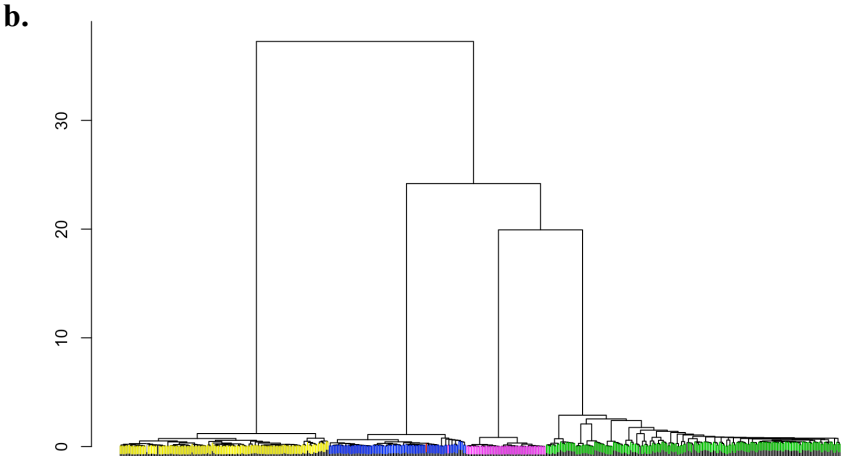
c.

		1	**2**	**3**	**4**
P. aeruginosa (25%)		61	0	0	0
A. baumannii (25%)		1	281	0	0
pAB1		0	1	0	0
pAB2		0	1	0	0
pB10		0	8	0	0
L. crispatus (25%)		0	0	173	0
E. coli (25%)		1	0	0	83
pBP136		0	13	0	2

Figure 3.3: Contigs clustered based on Hi-C read linkages for Community 3, which has four bacterial species and similar plasmids in different hosts (**a.**) PCoA plot (**b.**) Hierarchical clustering dendrogram (**c.**) Number of contigs[1] in each cluster for each species[2]

[1] Length of contigs varies
[2] Species determined by alignment of contigs to reference genomes



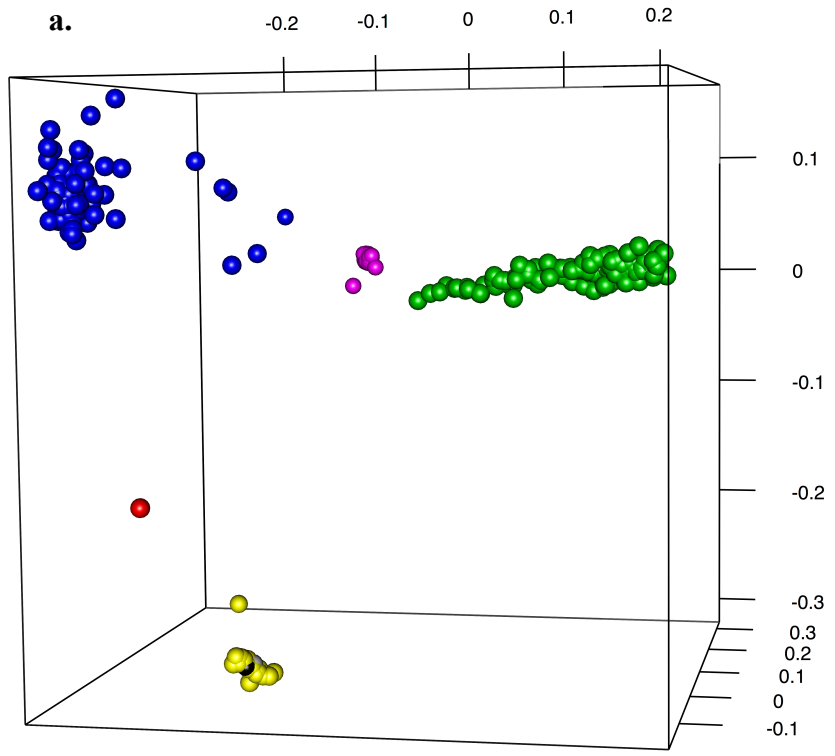


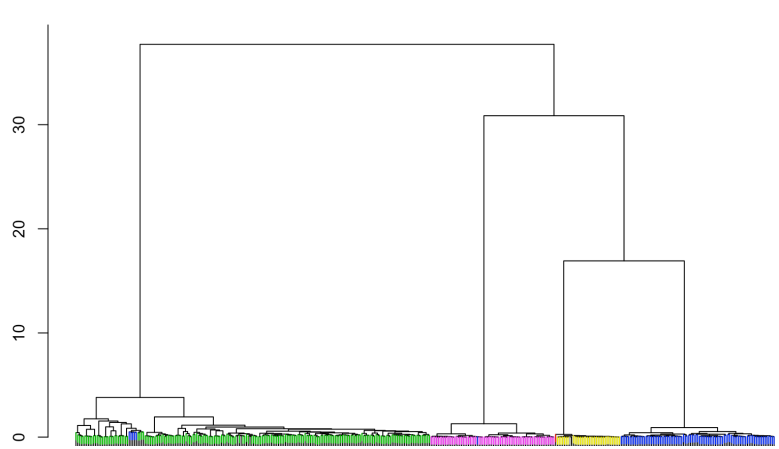
c.

		1	**2**	**3**	**4**
P. aeruginosa (25%)		51	0	1	0
A. baumannii (25%)		0	132	0	0
pAB1		0	1	0	0
pAB2		0	1	0	0
L. crispatus (25%)		0	0	0	186
E. coli (25%)		0	0	86	0
pB10 [3]		0	0	1	0

Figure 3.4: Contigs clustered based on Hi-C read linkages for Community 4a, which has four bacterial species and the same plasmid (pB10) in two different hosts that are present at equal quantities (**a.**) PCoA plot (**b.**) Hierarchical clustering dendrogram (**c.**) Number of contigs[1] in each cluster for each species[2]

[1] Length of contigs varies
[2] Species determined by alignment of contigs to reference genomes
[3] Plasmid pB10 was present in both *A. baumannii* and *E. coli*



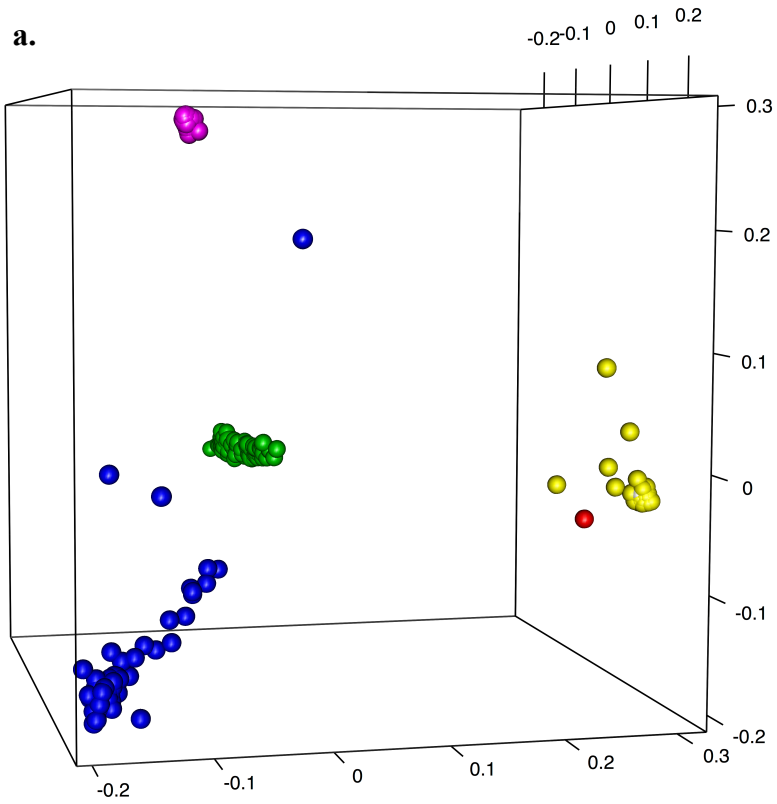


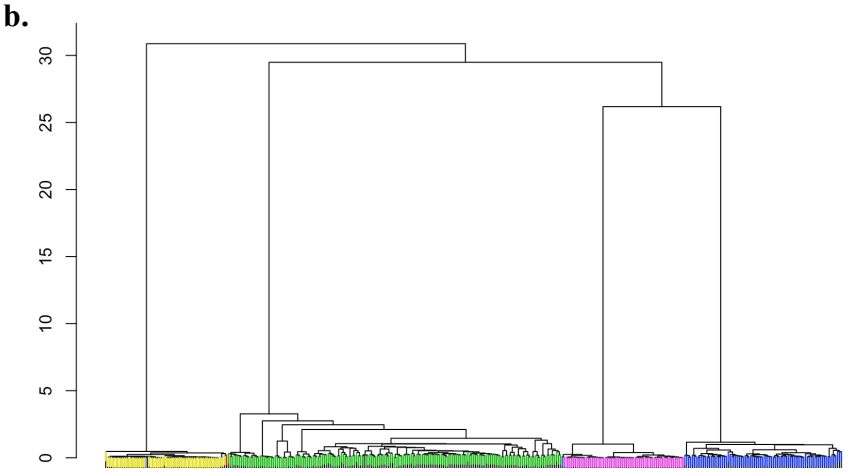
c.

		1	2	3	4
P. aeruginosa (25%)		0	0	69	0
A. baumannii (45%)		34	0	0	0
pAB1		1	0	0	0
pAB2		1	0	0	0
pB10 [3]		1	0	0	0
L. crispatus (25%)		0	0	0	192
E. coli (5%)		0	87	1	5

Figure 3.5: Contigs clustered based on Hi-C read linkages for Community 4b, which has four bacterial species and the same plasmid (pB10) in two different hosts which are present at different quantities (**a.**) PCoA plot (**b.**) Hierarchical clustering dendrogram (**c.**) Number of contigs[1] in each cluster for each species[2]

[1] Length of contigs varies
[2] Species determined by alignment of contigs to reference genomes
[3] Plasmid pB10 was present in both *A. baumannii* and *E. coli*





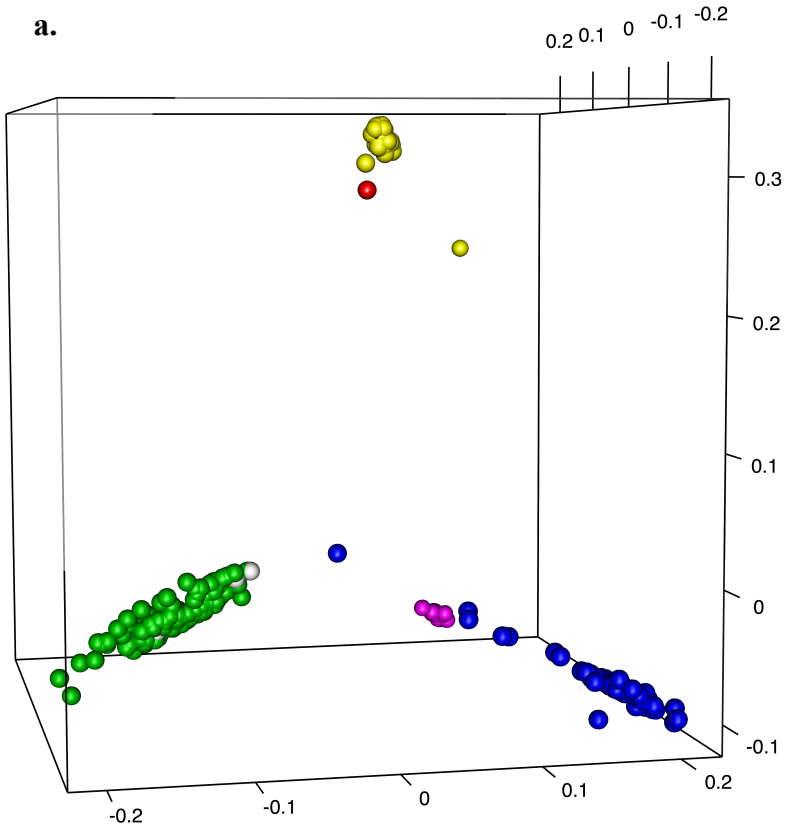
c.

		1	**2**	**3**	**4**
P. aeruginosa (25%)		64	0	0	0
A. baumannii (25%)		0	0	0	62
pAB1		0	0	0	1
pAB2		0	0	0	1
pB10 [3]		0	0	0	1
L. crispatus (25%)		0	0	176	0
E. coli (25%)		1	84	0	0

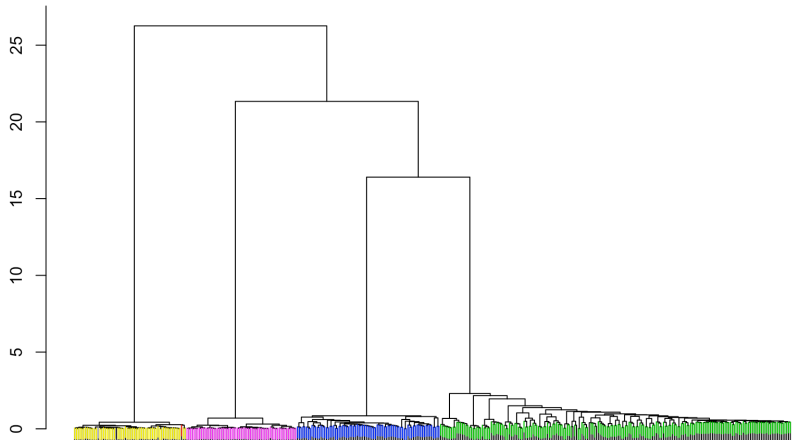
Figure 3.6: Contigs clustered based on Hi-C read linkages for Community 5a, which has four bacterial species and the same plasmid (pB10) in two different hosts but present in only 10% of one of the hosts (**a.**) PCoA plot (**b.**) Hierarchical clustering dendrogram (**c.**) Number of contigs[1] in each cluster for each species[2]

[1] Length of contigs varies
[2] Species determined by alignment of contigs to reference genomes
[3] 10% of the *E. coli* also carried pB10

a.



b.



c.

		1	**2**	**3**	**4**
P. aeruginosa (25%)		61	0	0	0
A. baumannii (25%)		0	0	0	59
pAB1		0	0	0	1
pAB2		0	0	0	1
pB10 [3]		0	0	0	1
L. crispatus (25%)		0	0	190	0
E. coli (25%)		0	79	0	0

Figure 3.7: Contigs clustered based on Hi-C read linkages for Community 5b, which has four bacterial species and the same plasmid (pB10) in two different hosts but present in only 1% of one of the hosts (**a.**) PCoA plot (**b.**) Hierarchical clustering dendrogram (**c.**) Number of contigs[1] in each cluster for each species[2]

[1] Length of contigs varies
[2] Species determined by alignment of contigs to reference genomes
[3] 1% of the *E. coli* also carried pB10

Chapter 4: Discussion and Future Directions

Seven simple bacterial communities were constructed, assayed using the Hi-C method and analyzed using our bioinformatics pipeline. For these simple communities, our analysis defines the limits of the Hi-C method in the areas of separating species, determining plasmid hosts, and placing ARG by replicon (i.e., plasmid or chromosome). Nearly all contigs clustered accurately by species. The plasmid host could be determined except when similar plasmids were found in the community. ARG were placed on the correct replicon except when they were found on multiple replicons in the community.

Species Separation Using Hi-C

Both Hi-C and the bioinformatics pipeline performed robustly when it came to accurately separating the contigs by species. There were a few contig misassignments, but several communities had perfect contig placement. Never more than six contigs in a community were misassigned, and these were never long contigs but short ones. In addition, these shorter contigs often had very similar alignment scores for multiple reference species. These "misassignments" were thus most likely reflective of sequence similarities between species rather than an incorrect Hi-C interaction or a problem with the clustering algorithm. The presence of these similar sequences in multiple genomic locations no doubt caused assembly difficulties that kept these shorter sequences from being included in a longer contig that would have clustered more easily. The placement of these genes may indicate the species that carried them in the highest quantity, especially if they were repeat sequences.

The angle at which some snapshots of the 3D plots were taken demonstrates another interesting pattern (e.g. Figure 3.5.a). Those species that assembled into few contigs like *P. aeruginosa* tended to cluster into a tighter sphere. Those that were more fragmented, like *L. crispatus*, tended to be visualized as a longer rod using principle coordinate analysis. These phenomena are likely because the clustering algorithm is based on how many Hi-C reads align to each contig. Longer contigs have more Hi-C read interactions because they have more length available for Hi-C interactions to occur. More Hi-C interactions corresponds to a shorter distance between contigs. They are then visualized as a tighter cluster of dots using principle coordinate analysis (*P. aeruginosa*,

Figure 3.5.a). Those with more, shorter contigs like *L. crispatus* have less Hi-C interactions between each contig pair and thus cluster less tightly. As Hi-C is known to show DNA interactions within the cell in 3D space (Le et al., 2013), it is also possible that this rod shape is at least partly due to *L. crispatus* having that cell morphology. If one end of the chromosome was more often found at one end of the cell than the other, there would be more Hi-C interactions between that section of the chromosome and itself than with the other end of the chromosome due to their proximity in 3D space.

Plasmid Placement

The plasmids in our assays also assembled well and could be accurately placed by species using Hi-C, except in the case of multiple species carrying similar plasmids. In Community 3, similar plasmids, pBP136 and pB10, resulted in much more fractured plasmid assemblies than were observed in the rest of the communities. In most of the assays, plasmids pAB1, pAB2, and pB10 assembled into 1 contig each and were easily placed by species. This showed that Hi-C could be quite useful for clustering unique, native plasmids with the species carrying them. In Community 3 however, the many contigs from these similar plasmids showed interactions with each other (due to sequence similarity) as well as with the species that carried them. Visually, this formed a plasmid cloud between the two species although pBP136 clearly associated more often with *E. coli* and pB10 associated more with *A. baumannii* (Figure 3.3.a). While the analysis of this community was still quite simple using this pipeline, it should be noted that a situation like this would be much more difficult to decipher in the context of an environmental sample with no *a priori* knowledge of the community members. Similar clustering patterns would need to be looked for and further investigated in these communities.

The other interesting case that came up when identifying the bacterial carriers of plasmids, occurred in Communities 4a, 4b, 5a, and 5b. In each of these assays the same plasmid, pB10, was present in two species but often in different percentages. Since it was the same plasmid in both species, it easily assembled into one contig in each assay. When pB10 was present in two species in equal percentages, it clustered directly between the two species (Figures 3.4.a). When pB10 was present at different percentages in the two species, it always clustered closer to the species where it was present at a higher

percentage (Figures 3.5.a, 3.6.a, & 3.7.a). This indicates that Hi-C could be a very useful tool for researching plasmid transfer scenarios. Using principle coordinate analysis, this plasmid placement was easy to observe in 3D space. With a more complex community of 10 or 100 species however, this 3D space would look much more cluttered and plasmid interactions such as these would be much more difficult to identify visually.

Hi-C still has the potential to be useful for linking plasmids to their hosts in larger communities if the data are looked at numerically instead of in principle coordinate space. Figures 4.1 through 4.4 show the number of Hi-C read pairs linking the pB10 plasmid contig to each species cluster in Communities 4a, 4b, 5a, and 5b. Figure 4.1 shows that pB10 had very similar amounts of Hi-C interactions with both *E. coli* and *A. baumannii* and that it had significantly more Hi-C interactions with them than it did with *L. crispatus* and *P. aeruginosa.* The fact that pB10 did have some interactions with species other than those it was found in was not a cause for concern as Hi-C datasets are known to be noisy (Beitel et al., 2014).

The numbers of Hi-C reads showing real interactions were also well above the background noise. Figure 4.2 shows how the Hi-C interactions of pB10 with *A. baumannii* increased proportionally and numerically as *A. baumannii* was present at a higher percentage of the community compared to Figure 4.1. The Hi-C interactions of plasmid pB10 with *E. coli,* which was now present at only 5% of the community, decreased to the level of the background noise, however. Presumably, if *E. coli* was present at only 5% of the community and carried no pB10, its Hi-C interactions with pB10 would have fallen further below the background noise of those species that were each 25% of the community. Similarly in Communities 5a and 5b, where pB10 was present in *E. coli* at 2.5% and 0.25% of the community respectively, pB10's interactions with *E. coli* fell to below the level of the background noise (Figures 4.3 & 4.4). The number of Hi-C read pairs linking plasmid pB10 with the *A. baumannii* chromosome were always significant. Plasmid pB10's Hi-C linkages with the other species appear to be a function of the number of contigs in each species cluster. For instance, the background noise is always highest in *L. crispatus*, which also happens to be the species with the most contigs in each of these communities.

ARG Placement

The Hi-C method worked quite well for accurately determining the placement of ARG within a bacterial community. ARG were placed on the correct replicons and within the correct species in all cases except for situations like Community 2, where the same ARG was present in multiple species. This community was included in the study as a control to determine whether or not the assembler/clustering method could perform robustly when the same gene was found in different locations. This caused assembly problems, as was expected. It is a major problem with metagenomic assemblies and a major limitation in downstream applications of those assemblies, like this method. The misassembled contig in Community 2 was easy to identify using the Resfinder blast search, as it was the only contig carrying the Km resistance gene. An alignment to the reference genomes showed that the Km resistance gene was connected to both *E. coli* and pB10 DNA. However, in the case of a more complex community where more than two replicons could potentially carry the same ARG and no known reference genomes were available, a misassembled contig like this would become increasingly impossible to deconvolute. It would have been easy to assume that the Km resistance gene was only present in *P. putida* if the decision was based on the Hi-C clustering alone or that it was only in *E. coli* if the decision was based solely on the NCBI database alignment. This looks to be a major limitation on Hi-C's usefulness with complex, environmental samples. This study does suggest however, that when an ARG is present in multiple locations, Hi-C is more likely to assign it to a species cluster where it is found on a plasmid than a cluster where it is found on a chromosome. This is likely due to plasmids being present at a higher copy number and thus having more Hi-C interactions linking them to their cluster.

Future Scenarios to Test

Other Community Assays

The results are promising in that it was easy to determine that a species carried a plasmid when the species was present in at least 25% of the community. This was true for pB10 in *A. baumannii* (Figures 4.2, 4.3, & 4.4). It was also easy to determine that multiple species carried a plasmid if both species were present at 25% of the community

and the entirety of their species carried it (Figure 4.1). The numbers of plasmid interactions with the plasmid carrying species are even high enough at this level that it seems likely that plasmid host could still be determined when species are present at less than 25% of the community (Figure 4.1). These results do suggest that a plasmid host would need to be present at more than 5% of the community to accurately determine plasmid location (Figure 4.2). Further testing would have to be done with plasmid carrying species present at different percentages of the community to determine Hi-C's absolute limits of detection when it comes to plasmid placement. It would also perhaps be helpful to repeat one of these assays with the same plasmid present in more than 2 community members to observe any differences that might arise in the clustering patterns.

In the 3D plots for the communities where plasmid pB10 was present in two species (Figures 3.4.a, 3.5.a, 3.6.a, & 3.7.a), pB10 appears to be placed between the host species. This between species placement was even found in scenarios where pB10 could not be seen in both species based solely on counts of Hi-C read interactions (Figures 4.2, 4.3, & 4.4). This 3D placement could still have been observed because the contigs from cells that pB10 was in had a greater strength of association with their species clusters than the background noise contigs had with their species clusters. This cannot be assumed for certain though and should be tested further before being used to determine definite plasmid placement in an environmental sample with no knowledge of actual plasmid placement. The possibility that pB10 was located between *A. baumannii* and *E. coli* because of similar DNA sequence or Hi-C interactions causing these particular species to cluster near each other should be ruled out. It would be helpful to set up a community similar to Community 4b but with different species to see if the same clustering pattern was observed.

In Communities 5a and 5b (Figures 3.6.a & 3.7.a), pB10 is seen slightly outside of the *A. baumannii* cluster. It cannot be known for certain if that is because of its valid associations with *E. coli* or because of the higher levels of background noise compared to Community 1 for instance. While the native *A. baumannii* plasmids, pAB1 and pAB2, do cluster directly within the *A. baumannii* cluster even in the case of background noise, they are smaller than pB10. They are thus expected to associate more with their species

and less with themselves than a larger replicon unit like pB10 which can form its own cluster more easily. Running an assay the same as Community 1 but with *P. aeruginosa* and *L. crispatus* as background noise could also serve as a control and show whether the large plasmid, pB10, would cluster directly within its species cluster as in Community 1 even with more background noise present.

Environmental Samples

This method and pipeline could be tested on more complex, environmental samples with a few adjustments. The included R code (Appendix B) was designed so that the only large input file is the list of contig interactions taken from the Hi-C cleaning results (HiCUP). This file will have as many lines as there are valid and unique Hi-C reads. While it can be a lengthy file, it is not extremely large as the lines themselves consist of nothing but contig names. The R script then uses this dataset to construct a symmetrical matrix, the dimensions of which are equal to the number of contigs. R can hold a matrix far larger than any produced in this study. If for some reason the dataset became many orders of magnitude larger and the R analysis could not be performed remotely on a computer with sufficient memory, the script could perhaps be modified to accommodate this using one of R's big data packages.

In lieu of reference genomes the contigs would have to be identified by blasting them against a much larger database. This could be done using NCBI-blast with its genomic or nucleotide database. In a more complex sample, it is quite likely that the number of clusters would not be visually obvious in either a dendrogram or 3D principle coordinate plot. In this situation, the silhouette function in R's "cluster" package could be used on the distance matrix as a way of determining the optimal number of clusters. When plotted over a range of values for *k*, the value which yields the highest average silhouette width can be assumed to be the optimal number of clusters.

Conclusions

Overall, the Hi-C method is a promising tool for linking ARG and replicon units within bacterial communities as well as distinguishing bacterial genomes from a mixed metagenomic dataset. The bioinformatics pipeline developed here performed quite robustly in the case of bacterial communities with only a few members as were used in

this study. The species not only achieved nearly perfect separation but also nearly the whole genome was retrieved in most cases. Further experimentation and statistical analysis could determine more precisely its limits of detection in the area of plasmid placement as well as how many species at a time this pipeline can separate clearly.

This research project did show that complications can arise when the same ARG is present in multiple species or when the same or similar plasmids are present in multiple species. These situations seem to cause difficulties during contig assembly. Comparison to the clustering patterns described here for these simulated situations could perhaps resolve some of the clustering ambiguity in simple environmental samples (i.e. river water). It looks like it would be difficult to impossible to determine with certainty the placement of shared ARG and plasmids in complex environmental samples such as soil (Howe et al., 2014) however. While ambiguity in the placement of shared genes is a major limitation of the method, being able to cluster most of the contigs by species would still be quite useful with any metagenomic samples. Even if only partial genomes for the major species could be recovered, as has been demonstrated previously (Marbouty et al., 2014), Hi-C can organize metagenomic datasets to a level which has been difficult up to this point. This pipeline provides a simple way to do the necessary cluster analysis. This method may prove most useful in lab or clinical scenarios however where there are bacterial communities with limited types of species. In such situations, Hi-C provides a sufficient level of resolution to determine the most likely bacterial carriers of specific genetic elements. The method could perhaps prove useful for tracking plasmid movement or informing the best antibiotic treatment strategies in the case of resistant infections. It could also hold promise as a method for tracking pathogen/ARG spread within hospitals as is being done in the Hospital Microbiome Project (Smith et al., 2013).

Figure 4.1: Hi-C read pairs linking pB10 to each species cluster in Community 4a, where pB10 was present in *A. baumannii* and *E. coli* (each present at 25% of the community)

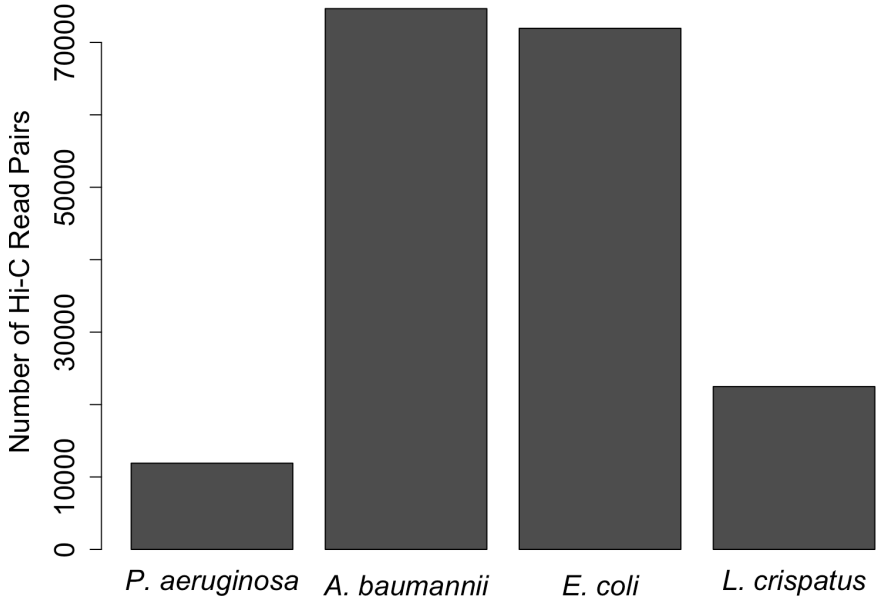


Figure 4.2: Hi-C read pairs linking pB10 to each species cluster in Community 4b, where pB10 was present in *A. baumannii* and *E. coli* (present at 45% and 5% of the community respectively)

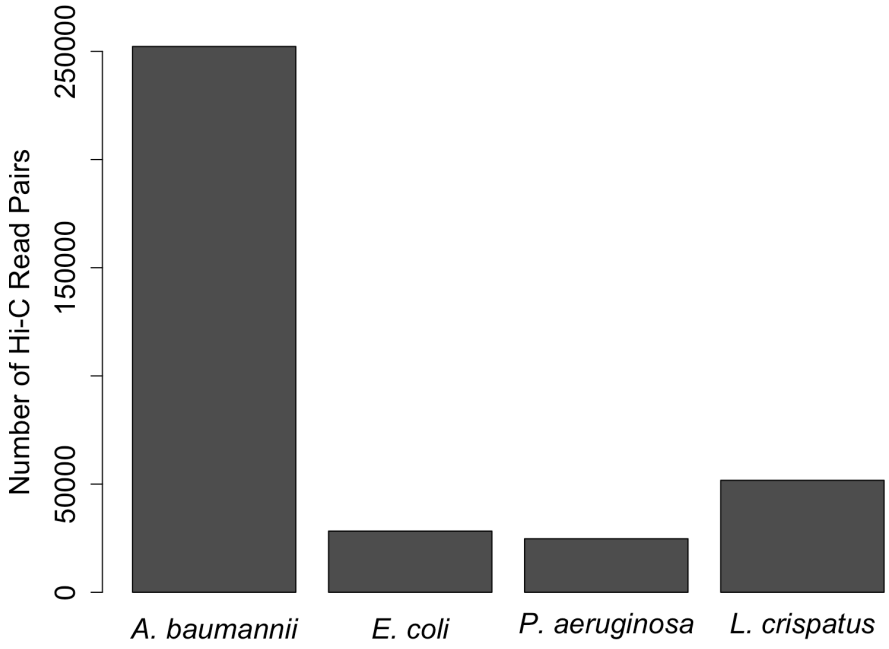


Figure 4.3: Hi-C read pairs linking pB10 to each species cluster in Community 5a, where pB10 was present in *A. baumannii* and *E. coli* (each present at 25% of the community) but only present in 10% of the *E. coli*

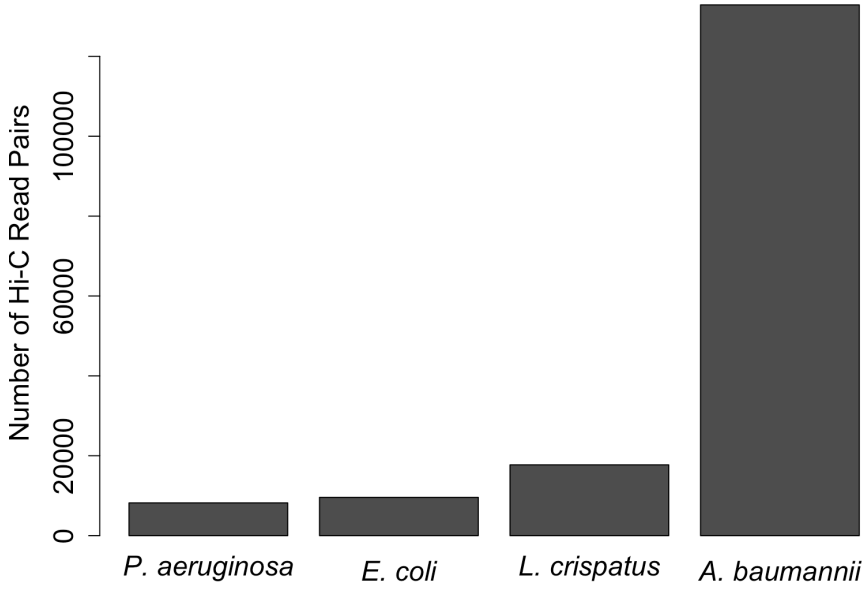
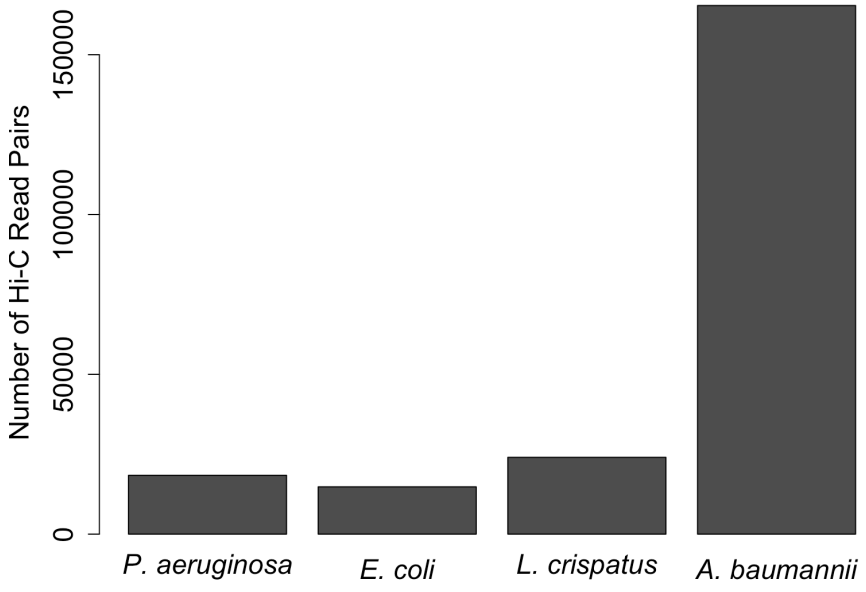


Figure 4.4: Hi-C read pairs linking pB10 to each species cluster in Community 5b, where pB10 was present in *A. baumannii* and *E. coli* (each present at 25% of the community) but only present in 1% of the *E. coli*



References

American Chemical Society International Historic Chemical Landmarks. Discovery and Development of Penicillin. (2015, November 5). Retrieved October 24, 2017, from http://www.acs.org/content/acs/en/education/whatischemistry/landmarks/flemingpenicillin.html

Beitel, C. W., Froenicke, L., Lang, J. M., Korf, I. F., Michelmore, R. W., Eisen, J. A., & Darling, A. E. (2014). Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ, 2*.

Berendonk, T. U., Manaia, C. M., Merlin, C., Fatta-Kassinos, D., Cytryn, E., Walsh, F., . . . Martinez, J. L. (2015). Tackling antibiotic resistance: the environmental framework. *Nature Reviews Microbiology, 13*(5), 310-317.

Burton, J. N., Liachko, I., Dunham, M. J., & Shendure, J. (2014). Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. *G3: Genes|Genomes|Genetics, 4*(7), 1339-1346.

Centers for Disease Control and Prevention Antibiotic / Antimicrobial Resistance. About Antimicrobial Resistance. (2017, September 19). Retrieved October 24, 2017, from http://www.cdc.gov/drugresistance/about.html

Dang, B., Mao, D., Xu, Y., & Luo, Y. (2017). Conjugative multi-resistant plasmids in Haihe River and their impacts on the abundance and spatial distribution of antibiotic resistance genes. *Water Research, 111*, 81-91.

Dodder, N. G., Maruya, K. A., Ferguson, P. L., Grace, R., Klosterhaus, S., Guardia, M. J., . . . Ramirez, J. (2014). Occurrence of contaminants of emerging concern in mussels (Mytilus spp.) along the California coast and the influence of land use, storm water discharge, and treated wastewater effluent. *Marine Pollution Bulletin, 81*(2), 340-346.

Driscoll, C. B., Otten, T. G., Brown, N. M., & Dreher, T. W. (2017). Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Standards in Genomic Sciences, 12*(1).

D'Costa, V. M., Griffiths, E., & Wright, G. D. (2007). Expanding the soil antibiotic resistome: exploring environmental diversity. *Current Opinion in Microbiology, 10*(5), 481-489.

Ferrari, B. C., Winsley, T., Gillings, M., & Binnerup, S. (2008). Cultivating previously uncultured soil bacteria using a soil substrate membrane system. *Nature Protocols, 3*(8), 1261-1269.

Forsberg, K. J., Reyes, A., Wang, B., Selleck, E. M., Sommer, M. O., & Dantas, G. (2012). The Shared Antibiotic Resistome of Soil Bacteria and Human Pathogens. *Science, 337*(6098), 1107-1111.

Gotz, A., & Smalla, K. (1997). Manure Enhances Plasmid Mobilization and Survival of Pseudomonas putida Introduced into Field Soil. *Applied and Environmental Microbiology*, *63*(5), 1980–1986.

Heuer, H., & Smalla, K. (2007). Manure and sulfadiazine synergistically increased bacterial antibiotic resistance in soil over at least two months. *Environmental Microbiology, 9*(3), 657-666.

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., & Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences, 111*(13), 4904-4909.

Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.

Kado, Clarence I. (1998). Origin and evolution of plasmids. *Antonie van Leeuwenhoek, 73*(1), 117-126.

Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology, 23*, 110-120.

Kumar, K., Gupta, S. C., Baidoo, S. K., Chander, Y., & Rosen, C. J. (2005). Antibiotic Uptake by Plants from Soil Fertilized with Animal Manure. *Journal of Environment Quality, 34*(6), 2082.

Lajoie, B. R., Dekker, J., & Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods, 72*, 65-75.

Le, T. B., Imakaev, M. V., Mirny, L. A., & Laub, M. T. (2013). High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science, 342*(6159), 731-734.

Lester, C. H., Frimodt-Moller, N., Sorensen, T. L., Monnet, D. L., & Hammerum, A. M. (2006). In Vivo Transfer of the vanA Resistance Gene from an Enterococcus faecium Isolate of Animal Origin to an E. faecium Isolate of Human Origin in the Intestines of Human Volunteers. *Antimicrobial Agents and Chemotherapy, 50*(2), 596-599.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Maffioli, S. I., Zhang, Y., Degen, D., Carzaniga, T., Gatto, G. D., Serina, S., . . . Ebright, R. H. (2017). Antibacterial Nucleoside-Analog Inhibitor of Bacterial RNA Polymerase. *Cell, 169*(7).

Magoc, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics, 27*(21), 2957-2963.

Marbouty, M., Cournac, A., Flot, J., Marie-Nelly, H., Mozziconacci, J., & Koszul, R. (2014). Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *ELife, 3*.

Martinez, J. L. (2009). Environmental pollution by antibiotics and by antibiotic resistance determinants. *Environmental Pollution, 157*(11), 2893-2902.

McClure, N. C., Weightman, A. J., & Fry, J. C. (1989). Survival of Pseudomonas putida UWC1 containing cloned catabolic genes in a model activated-sludge unit. *Applied and Environmental Microbiology*, *55*(10), 2627–2634.

Munir, M., Wong, K., & Xagoraraki, I. (2011). Release of antibiotic resistant bacteria and genes in the effluent and biosolids of five wastewater utilities in Michigan. *Water Research, 45*(2), 681-693.

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., . . . Pevzner, P. A. (2013). Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. *Lecture Notes in Computer Science Research in Computational Molecular Biology,* 158-170.

Oliva, M., Monno, R., Daddabbo, P., Pesole, G., Dionisi, A., Scrascia, M., . . . Pazzani, C. (2017). A novel group of IncQ1 plasmids conferring multidrug resistance. *Plasmid, 89*, 22-26.

Orlando, V., Strutt, H., & Paro, R. (1997). Analysis of Chromatin Structure byin VivoFormaldehyde Cross-Linking. *Methods, 11*(2), 205-214.

Petersen, K. R., Streett, D. A., Gerritsen, A. T., Hunter, S. S., & Settles, M. L. (2015). Super deduper, fast PCR duplicate detection in fastq files. *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics - BCB 15*.

Roesch, L. F., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K., Kent, A. D., . . . Triplett, E. W. (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*.

Roth, Thomas F., & Donald R. Helinski. (1967). Evidence for Circular DNA Forms of a Bacterial Plasmid. *Proceedings of the National Academy of Sciences of the United States of America, 58*(2), 650-657.

RStudio (2016). RStudio: Integrated development environment for R (Version 0.99.903) [Computer software]. Boston, MA. Available from http://www.rstudio.org/.

Sengeløv, G. (2003). Bacterial antibiotic resistance levels in Danish farmland as a result of treatment with pig manure slurry. *Environment International, 28*(7), 587-595.

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C., Vert, J., . . . Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology, 16*(1).

Smith, D., Alverdy, J., An, G., Coleman, M., Garcia-Houchins, S., Green, J., . . . Gilbert, J. (2013). The Hospital Microbiome Project: Meeting Report for the 1st Hospital Microbiome Project Workshop on sampling design and building science measurements, Chicago, USA, June 7th-8th 2012. *Standards in Genomic Sciences, 8*(1), 112-117.

Tang, Y., Dai, L., Sahin, O., Wu, Z., Liu, M., & Zhang, Q. (2017). Emergence of a plasmid-borne multidrug resistance gene cfr(C) in foodborne pathogen Campylobacter. *Journal of Antimicrobial Chemotherapy, 72*(6), 1581-1588.

Udikovic-Kolic, N., Wichmann, F., Broderick, N. A., & Handelsman, J. (2014). Bloom of resident antibiotic-resistant bacteria in soil following manure fertilization. *Proceedings of the National Academy of Sciences,111*(42), 15202-15207.

Vartoukian, S. R., Palmer, R. M., & Wade, W. G. (2010). Strategies for culture of ‘unculturable’ bacteria. *FEMS Microbiology Letters*.

Waksman, S.A., Woodruff, H.B., 1940. The soil as a source of microorganisms antagonistic to disease-producing bacteria. J. Bacteriol. 40, 581–600.

White House. (2015). Obama Administration Releases National Action Plan to Combat Antibiotic-Resistant Bacteria. whitehouse.gov. https://www.whitehouse.gov/the-press-office/2015/03/27/fact-sheet-obamaadministration-releases-national-action-plan-combat-ant (Accessed October 12, 2017).

WHO. (2014). Antimicrobial resistance: global report on surveillance 2014.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., & Andrews, S. (2015). HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research*.

Wit, E. D., & Laat, W. D. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes & Development, 26*(1), 11-24.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., . . . Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy, 67*(11), 2640-2644.

Zhu, Y., Johnson, T. A., Su, J., Qiao, M., Guo, G., Stedtfeld, R. D., . . . Tiedje, J. M. (2013). Diverse and abundant antibiotic resistance genes in Chinese swine farms. *Proceedings of the National Academy of Sciences, 110*(9), 3435-3440.

Appendix A

The Hi-C method has recently been proven useful for reconstructing individual genomes from mixed cell populations by physically linking DNA fragments that occupied the same cell, prior to sequencing. This project was designed to test whether plasmid and bacterial carriers of ARG can be identified using the chromosomal conformation capture (Hi-C) approach. Towards this end, a series of mock bacterial communities were designed to answer a series of questions including Hi-C's limits of detection for plasmids and ARG and whether we could develop a working version of the Hi-C protocol in house.

Hi-C Protocol

Three replicates were made of each bacterial community (Table 3.1). One replicate of each was prepared by our collaborators using a proprietary Hi-C protocol. The remaining two replicates for each community were used for practice and development of the Hi-C method on site. This was done using previously published protocols (Beitel et al., 2014; Burton et al., 2014) that were combined and then modified to give the following Hi-C wet lab protocol. During this development process two replicates each of samples 1 and 2 were used as test samples. While one set of these was sequenced lightly, the percentage of valid Hi-C reads in the libraries was so low that it was not deemed useful to sequence them to full depth. Two community replicates, 3.3 and 4a.3, were successfully prepped on site and sequenced by our collaborators however. The analysis of these samples is included for comparison to those prepped using the proprietary protocol. The Hi-C sequences were processed and analyzed using the same bioinformatics pipeline previously described (Chapter 2).

Formaldehyde Crosslinking

Samples were prepped for the Hi-C protocol by adding 37% formaldehyde directly to the PBS mixed culture to a final concentration of 1%. They were incubated for 20-30 minutes at room temperature and periodically swirled. The formaldehyde crosslinking reaction was quenched by the addition of glycine to a final concentration of 0.133M. The samples were swirled to mix and then incubated for another 20-30 minutes

at room temperature. The samples were then spun down for 2 minutes at 13,000 g. The supernatant was removed and each pellet was rinsed with 800µL of PBS. This was to remove formaldehyde. The samples were spun down again for 2 minutes at 13,000 g. The supernatant was decanted and the cell pellets stored at -20°C until the prep was completed.

Cell Lysis

Cross-linked cell pellets were thawed on ice. Five hundred µL of 0.1mm diameter zirconia-silica beads (BioSpec) and 1mL of chilled 1x TBS with 1% Triton-X and EDTA-free Protease Inhibitors (diluted per package specifications; Pierce) were added. The sample was then vortexed five times for 5min. each time with 1-2min. rests on ice in between. Each 2mL sample tube was put into a 15mL conical tube so that it hung inside the larger tube. A small hole was poked at the bottom of the 2mL tube with a hot needle so that the chromatin suspension could filter through the glass beads and into the larger tube. The assembly was spun at 3000 RPM for 3min. at 4°C so that the small tube dripped into the 15mL tube leaving the glass beads behind. The flowthrough was transferred to a fresh 2mL tube and spun at 13000 g for 10min. at 4°C to pellet chromatin and cellular debris. The supernatant was discarded and the pellet resuspended in 1mL of the same chilled 1x TBS with 1% Triton-X and EDTA-free Protease Inhibitors (pellet requires gentle mushing with pipette tip). The rinse was repeated using chilled 1x TBS (no additives). The sample was pelleted once more and resuspended in 500µL of 10mM Tris pH8. At this point the sample can be stored at 4°C and a subsample quantified with Qubit™ to check for a high enough DNA concentration (expecting 1-10ng/µL but can be a little higher).

Digest Chromatin

To 200µL of the chromatin suspension, 120µL of irradiated, sterile, deionized water, 20µL of the 4-cutter restriction enzyme Sau3AI, and 40µL of the associated NEBuffer 1.1 were added. The solution was incubated for 4 hours in a 37°C water bath. Four-cutter restriction enzymes recognize a specific sequence of 4 base pairs (bp) whereas other Hi-C protocols have used restriction enzymes that recognize a sequence of 6bp. Restriction sites of 4bp are much more common than those of 6bp so the 4-cutter,

Sau3AI, was chosen to maximize the number of cuts and increase the number of possible Hi-C interaction sites (Lajoie et al., 2015).

Fill in DNA Ends with Biotin

The digested ends were next filled in with biotin. Nine µL each of the nucleotides: dA, dT, and dG; all diluted to 1mM prior to addition, were added to 250µL of the digested chromatin. Twenty µL of 0.4mM biotinylated dCTP (Invitrogen) and 7.5µL of Klenow (NEB, DNA Polymerase I large fragment) were also added. No additional buffer was necessary because of that carried over from the last step. The solution was gently mixed and incubated for 45-60min. in a 37°C water bath. The enzymes were then deactivated by incubating at 70°C for 10-15min. in a heat block. It is important to not leave it for longer than that or the crosslinks will start reversing.

Ligation

At this point, DNA concentration was measured using a Qubit™ dsDNA BR Assay Kit kit. DNA concentration varies at this step but was approximately 3.5 ng/µL for samples 3.3 and 4a.3. Ligations are set up to have DNA concentrations of no higher than 0.5 ng/µL. Biotinylated chromatin was mixed with water to a final volume of 900µL and a final concentration of 0.4-0.5 ng/µL. To this dilution, 100µL of T4 DNA Ligase Buffer (NEB) and 6.25µL of T4 DNA Ligase enzyme (NEB) were added. If DNA concentrations are high enough this reaction can be multiplied for larger volumes. For samples 3.3 and 4a.3 the reaction was a little more than doubled resulting in final volumes of 2.1mL. If DNA concentrations are higher, ligase reactions up to 4-8mL final volume can be used. Final concentrations of ligase enzyme and buffer must be kept the same however. The ligase reaction was incubated for 4-6 hours at room temperature.

Reverse Crosslinks

The crosslinks were then reversed by the addition of proteinase K (25µL of 10mg/mL per 2mL total reaction). The reaction was incubated at 70°C overnight.

Purify DNA

The DNA was cleaned and concentrated using the Zymo DNA Clean & Concentrator™ -100 kit using 5 volumes of DNA Binding Buffer for each 1 volume of sample. In the case of larger ligation volumes, one recovery membrane should be used for each 2mL of ligation mixture. The solution was pushed through the membrane using a

sterile syringe and the DNA cleaned and recovered per the kit's directions. DNA was eluted into 300µL (2 x 150µL to increase recovery) of PCR grade water for each membrane used. Three µL of RNAse A (Thermo Scientific, 10mg/mL) were added for each 300µL of recovered DNA. This reaction was incubated for 30-45min. in a 37°C water bath.

Remove Unligated Biotinylated Ends

To remove unligated biotinylated ends, 33µL of NEBuffer 2.1 with BSA, 3µL of 10mM dATP, 3µL of 10mM dGTP, and 4µL of T4 DNA Polymerase (NEB) were added to the ~300µL of DNA. Increase amount of reagents added in proportion to final volume of recovered DNA if more than 300µL is available. The reaction was incubated for 10 minutes at room temperature and then 12°C for 1 hour in a PCR machine.

Clean DNA

The DNA was cleaned via DNA Clean & Concentrator^{TM} -5 Kit (Zymo). One membrane was used for each ~300µL of reaction to be concentrated. The DNA was eluted into 130µL of PCR grade water. Can elute into 65µL for each membrane if multiple cartridges are required. 4µL of each sample was used for Qubit^{TM} high sensitivity quantification. Final DNA concentrations were 0.48 ng/µL for sample 3.3 and 1.98 ng/ µL for 4a.3. The remaining 126µL of each sample were shipped to collaborators for a streptavidin pull down to enrich for fragments with a biotinylated ligation junction and Illumina library prep.

Clustering Results for Samples Prepped on Site

The Hi-C linkages from the samples prepped using the protocol modified on site did not cluster the contigs from Communities 3 and 4a as cleanly as the Hi-C linkages from the samples prepped by our collaborators did. While looking at hierarchical clustering dendrograms for those samples prepped by our collaborators was sufficient for determining the optimal number of species clusters for those communities, the optimal number was less obvious in the analysis of the communities prepped locally. As we wanted to compare the quality of clustering from our Hi-C reads to that from our collaborators' reads, the same numbers of clusters were used to analyze Communities 3.3

and 4a.3 as were used to analyze Communities 3 and 4a. This resulted in four species clusters for each community.

Community 3.3 was a replicate of Community 3 where the Hi-C prep was done by us using the proceeding protocol. Community 3.3's 3D plot (Figure A.1.a) shows the same overall trend as Community 3 (Figure 3.3.a) where the two plasmids, pBP136 and pB10, form a plasmid cloud between their two host species, *E. coli* and *A. baumannii*. Plasmid pBP136 is again closer to its carrier, *E. coli*, and pB10 is again closer to its carrier, *A. baumannii*. The plasmids are pulled towards each other by inter-plasmid Hi-C interactions as well as towards their respective hosts. Although the clustering trend is the same as Community 3, the clustering is not as clean. The species are clustered less tightly and there is less space between species clusters, thus not all contigs were assigned to the correct species (Figure A.1.b & A.1.c) as well as they were in Community 3 (Figure 3.3.c). As there were half as many valid and unique Hi-C read pairs to link the Community 3.3 contigs as there were for Community 3 (Tables A.1 & 3.3), looser clustering was expected. However, Community 3.3 (Table A.1) was sequenced as deeply as Community 3 (Table 3.3) even though it did not produce as many valid and unique Hi-C read pairs. This implies that Community 3.3 did not have as high quality of a Hi-C library since a smaller percentage of the read pairs could be classified as valid and unique Hi-C pairs.

Community 4a.3 (Figure A.2.a) shows the same plasmid placement as Community 4a (Figure 3.4.a) but the species are again clustered much less tightly and some contigs are thus misassigned (Figures A.2.b & A.2.c). The hierarchical clustering does not separate species genomes as clearly even though the amount of valid and unique Hi-C pairs performing the clustering is quite similar to Community 4a (Tables A.1 & 3.3). Community 4a.3 did not contain as high a percentage of PCR duplicates as Community 4a, implying that it could have been sequenced deeper to recover more unique valid reads. There is no telling whether this would have improved the clustering however. As Community 4a.3 already had nearly the same amount of valid and unique Hi-C read pairs as Community 4a, it appeared that the Hi-C library was not of as high quality in some way even though the valid reads for both communities met the same criteria of coming from different, non-contiguous restriction fragments.

ARG and Plasmid Placement

Community 3.3 (Table A.2) had the same results as Community 3 (Table 3.6) as far as ARG placement. ARG were accurately placed on plasmid contigs but the hosts of those plasmids were not always determined correctly by hierarchical clustering. Some of these plasmid contigs were misassigned because the similar plasmids, pB10 and pBP136, caused clustering confusion. In Community 4a, both *E. coli* and *A. baumannii* carried pB10 at the same percentages of the population. In the replicate prepped by our collaborators, pB10 was somewhat arbitrarily assigned to *E. coli* by the clustering algorithm and the ARG were assigned to that cluster along with it as was expected. In Community 4a.3, pB10 and thus its ARG were assigned to *A. baumannii* instead. As pB10 presumably had a 50/50 chance of being assigned to either one, this difference between the two replicates was not considered significant.

Discussion

Hi-C read pairs coming from the two Hi-C preps done using our modified protocol performed adequately at clustering. Species separation can be observed (Figures A.1.a & A.2.a) and the vast majority of the contigs were correctly assigned to their species cluster. A higher contig misassignment rate was observed in the samples prepped using our protocol however, with 31 and 20 contigs being misplaced in Communities 3.3 and 4a.3 respectively. While this is a somewhat high error rate, these misassignments were out of 625 and 462 contigs respectively (Table 3.3) making for an error rate of approximately 5% for both communities when based simply on number of contigs. However, the 5% error rate is somewhat misleading in that shorter contigs got misassigned while the much longer contigs, which made up the majority of the genetic sequence for the communities, were assigned to the correct species. A comparison of number of nucleotides misassigned to total nucleotides in the dataset would give a more accurate error rate that is much, much lower. The protocol was thus improved enough to perform adequately when it comes to separating a metagenomic dataset by species.

The 3D PCoA plots for Communities 3.3 and 4a.3 (Figures A.1.a & A.2.a) are more visually ambiguous than those from the samples prepped using our collaborator's proprietary protocol. Not having clarity in these PCoA plots could be a big drawback in

samples such as 5b where a clear 3D plot is the easiest way to see that a plasmid may be present in more than one species. The higher resolution offered by our collaborator's protocol was thus quite helpful for determining plasmid placement (Figure 3.3.a vs. Figure A.1.a; Figure 3.4.a vs. Figure A.2.a).

Table A.1: Hi-C Cleaning and Clustering Statistics for Communities 3.3[1] and 4a.3[1]

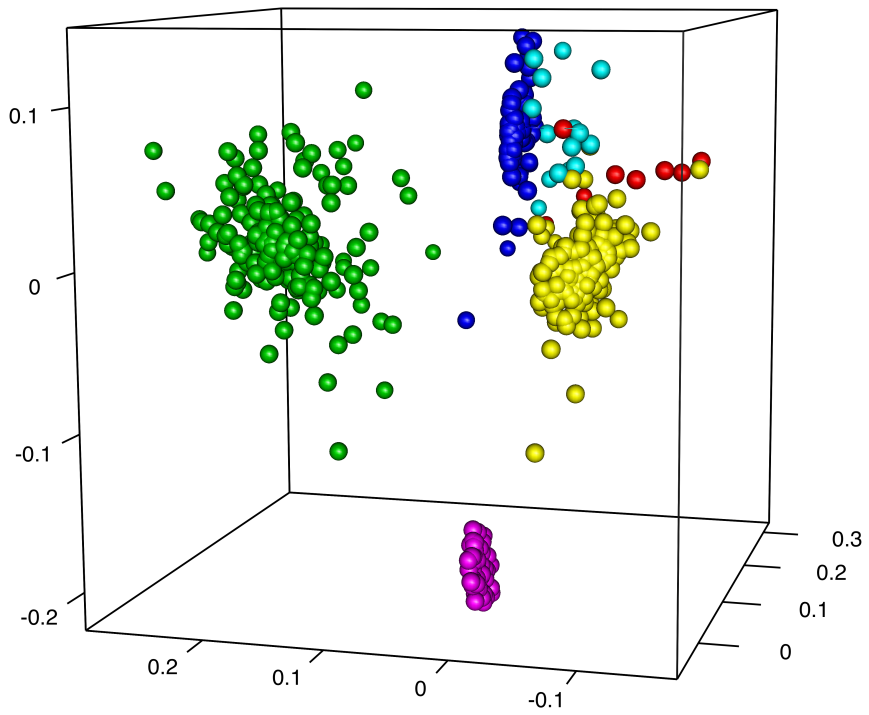
	Comm3.3	**Comm4a.3**
HiCUP		
Raw Hi-C read pairs	22,434,054	23,187,814
Valid Hi-C pairs	3,388,148	5,664,470
Valid & Unique Hi-C Pairs	2,920,341	5,565,686
Ratio of Hi-C pairs to contigs	4,499.8	11,742.0
Contigs that align to references[2]	634	470
Contigs not picked up by Hi-C[3]	24	12

[1] Clustering same set of contigs as for Communities 3 & 4
[2] Alignments determined by NCBI blast
[3] Contigs that no Hi-C read pairs aligned to

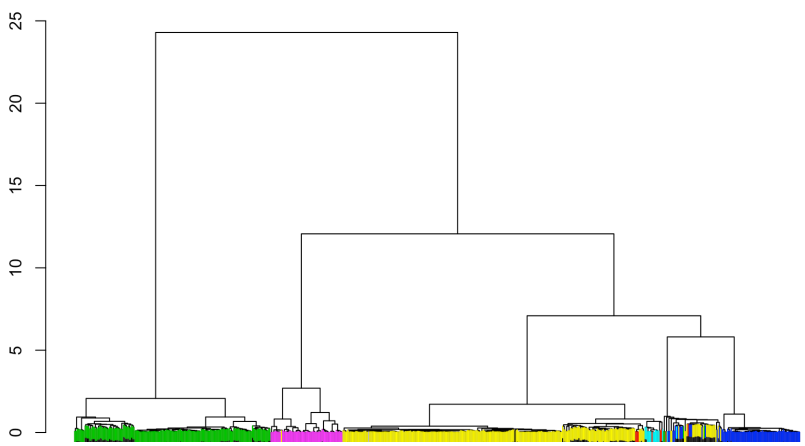
Table A.2: Genomic Cluster Each Antibiotic Resistance Gene was Assigned to Using Hi-C for Communities 3.3 and 4a.3

Resistance Genes	blaPAO	catB7	fosA	blaOXA-50	aph(3')-IIb	aph(3') Ia	catA1	tet(A)	blaOXA-2	sul1	strB	strA	blaOXA-180	blaADC-25
Comm3.3														
P. aeruginosa	x	x	x	x	x									
A. baumannii/ pB10 /pBP136								x	x	x	x	x	x	x
Comm4a.3														
P. aeruginosa	x	x	x	x	x									
A. baumannii /pB10								x	x	x	x	x	x	x

a.



b.

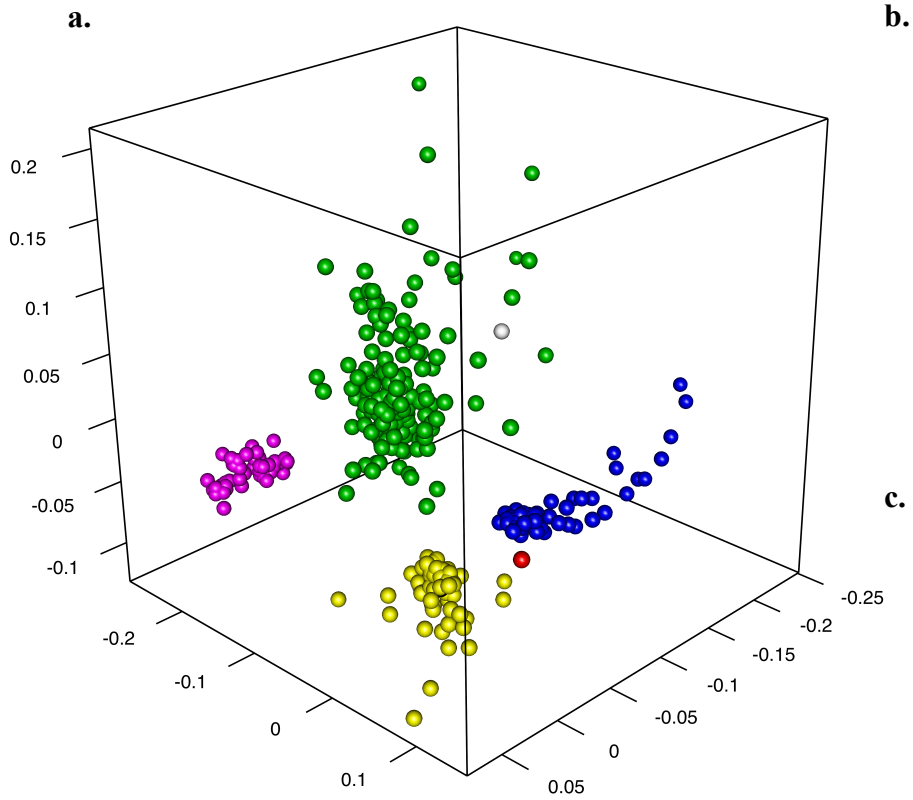


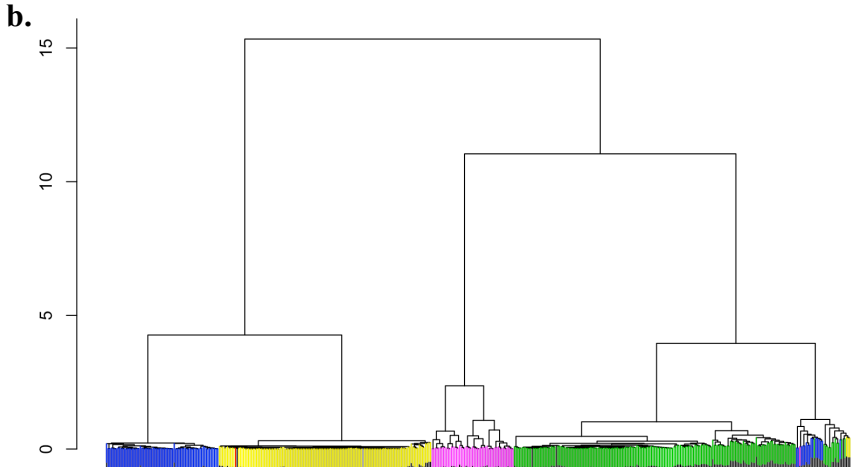
c.

		1	2	3	4
P. aeruginosa (25%)		61	0	0	0
A. baumannii (25%)		1	255	0	26
pAB1		0	1	0	0
pAB2		0	1	0	0
pB10		0	6	0	2
L. crispatus (25%)		0	0	169	4
E. coli (25%)		0	0	0	84
pBP136		0	13	0	2

Figure A.1: Contigs clustered based on Hi-C read linkages for Community 3.3, which has four bacterial species and similar plasmids in different hosts (**a.**) PCoA plot (**b.**) Hierarchical clustering dendrogram (**c.**) Number of contigs[1] in each cluster for each species[2]

[1] Length of contigs varies
[2] Species determined by alignment of contigs to reference genomes





c.

		1	**2**	**3**	**4**
P. aeruginosa (25%)		51	0	0	1
A. baumannii (25%)		0	129	0	3
pAB1		0	1	0	0
pAB2		0	1	0	0
pB10 [3]		0	1	0	0
L. crispatus (25%)		0	0	0	186
E. coli (25%)		0	0	70	16

Figure A.2: Contigs clustered based on Hi-C read linkages for Community 4a.3, which has four bacterial species and the same plasmid (pB10) in two different hosts which are present at equal quantities (**a.**) PCoA plot (**b.**) Hierarchical clustering dendrogram (**c.**) Number of contigs[1] in each cluster for each species[2]

[1] Length of contigs varies
[2] Species determined by alignment of contigs to reference genomes
[3] PB10 was present in both *A. baumannii* and *E. coli*

Appendix B

Cleaning Metagenomic Shotgun Reads

##Duplicate reads were removed using Super Deduper, which can be installed from: https://github.com/dstreett/Super-Deduper. Our paired end reads came as two files: one for forward reads and one for reverse reads. Because of this file format, default parameters were used via the following command:

>super_deduper -1 shotgun_forward_R1.fastq.gz -2 shotgun_reverse_R2.fastq.gz

##Flash2 was used on the output from Super Deduper to merge overlapping paired end reads. Flash2 can be installed from: https://github.com/dstreett/FLASH2.

##Options:
-M 200 to set the maximum overlap length to 200 (overkill since our reads were only 150bp in each direction)
-O checks for read overlaps at both ends of the reads
-Q 20 sets the quality score (reads will be cut off if they fall below)
-C 70 sets the percentage of the read that will be cut off if it falls below the quality score (quality score is used to break ties in the case of multiple overlaps)

> flash2 -M 200 -O -Q 20 -C 70 output_nodup_PE1.fastq output_nodup_PE2.fastq

##Sickle, which can be installed from: https://github.com/najoshi/sickle, was used to trim the reads based on quality. It was used on the three read files output from Flash2. The two, unmerged paired end files were quality trimmed using sickle pe. The output is three files: two for the forward and reverse reads and one for single reads. The single reads were created when the quality of one read in the pair was too low. The merged, extended reads output from Flash2 were quality trimmed using sickle se with the same options as sickle pe.

##Options:
-n removes all reads with an N in them because this denotes low quality
--length-threshold 75 discards reads that were trimmed to a length shorter than 75bp
--qual-threshold 20 sets the quality threshold for trimming
--qual-type sanger is used for reads processed using CASAVA 1.8 or higher as was the case with our modern, Illumina reads

>sickle pe -n --length-threshold 75 --qual-threshold 20 --qual-type sanger -f out.notCombined_1.fastq -r out.notCombined_2.fastq -o cleaned_PE1.fastq -p cleaned_PE2.fastq -s cleaned_SE1.fastq

>sickle se -n --length-threshold 75 --qual-threshold 20 --qual-type sanger --fastq-file out.extendedFrags.fastq --output-file cleaned_SE2.fastq

##The cleaned, unpaired reads output from sickle pe and those output from sickle se were concatenated into one file. The two separate files were then discarded.

>cat cleaned_SE1.fastq cleaned_SE2.fastq > cleaned_SE.fastq

Assembly into Contigs

##Cleaned metagenomic reads were assembled into contigs using Spades in meta mode. Spades can be installed from: http://cab.spbu.ru/software/spades/. It was used in default meta mode. The default k-mer sizes are 21, 33, and 55. Different k-mer sizes were tested but did not improve assemblies for our data. (Number of threads or amount of memory used can be changed from defaults depending on size of dataset and computational power available.)

>spades.py --meta -1 cleaned_PE1.fastq -2 cleaned_PE2.fastq -s cleaned_SE.fastq -o spades_output/

##Contigs longer than 500bp were saved as one file called: longcontigs.fasta

Species Identification

##If reference genomes are available, the species identity of contigs longer than 500bp can be determined by an alignment to the reference genomes using NCBI-blast.

##Turn file of contigs longer than 500bp into a blast database.

> makeblastdb -dbtype nucl -in longcontigs.fasta -hash_index

##Blast concatenated file of all references against contig blast database

> blastn -query all_ref.fasta -db longcontigs.fasta -outfmt 6 -evalue 0.0001 -out outfile -max_hsps 1

##Save columns 1, 2, and 4 of output as future input for R script.

> awk '{print$1,$2,$4}' outfile > outfile1

Cleaning Hi-C Reads

##Hi-C reads were cleaned using HiCUP which can be installed from: https://www.bioinformatics.babraham.ac.uk/projects/hicup/. As using the scripts takes several steps, watching or working alongside their detailed instructional video is highly recommended: https://www.youtube.com/watch?v=i6imVs66aew.

##Digest references. Provide sequence recognized by restriction enzyme used during Hi-C prep (GATC for enzyme Sau3AI). Sample can be named using --genome option:

>hicup_v0.5.8/hicup_digester -z -re1 ^GATC --genome Comm1 longcontigs.fasta

##Make Bowtie indices (must have Bowtie installed) that HiCUP will use, again have option to name samples:
bowtie2-build -f longcontigs.fasta Comm1

##Must create output folder prior to running HiCUP and designate the output folder's location as well as the Hi-C read file names in configuration file template provided with HiCUP: hicup_example.conf.
##HiCUP runs off the information provided in hicup_example.conf using the command:
>hicup_v0.5.8/hicup --config hicup_example.conf

##One of the files output by HiCUP has the extension ".hicup.bam". Convert it into sam format using samtools.

>samtools view _____.hicup.bam > hicup.sam

##Only the third column of hicup.sam, which lists the names of the contigs that each Hi-C read aligned to, is needed for the next step. The following short Python script was used to extract it in a concise format:

```
import sys

with open (sys.argv[1],'r') as infile:
        for line in infile:
                column2 = line.split()[2]
                print(column2.split('l')[0])
```

##This script was saved as an executable and applied using the command:

>python2.7 Matrix.py hicup.sam > interactions.csv

##The file: interactions.csv, was used as input for the R script which clustered the contigs based on Hi-C linkage frequencies.

##The other input required by the R script is the number of valid and unique Hi-C reads as determined by HiCUP. This number can be found near the end of the .html file found in the HiCUP output folder.

##Optionally, the output from the NCBI-blast alignment to reference genomes, outfile1, can be used as an input to color the contig cluster plots by species identity.

Clustering and Visualization in R

##This script was run in RStudio ("RStudio", 2016)
install.packages("sparcl")

```
install.packages("rgl")
install.packages("cluster")


##set up variables
a = #number of valid and unique Hi-C reads, enter as a number, no commas
b = 2*a
c = b-1
Interactions <- read.csv("interactions.csv", header = FALSE)


##build symmetrical matrix of Hi-C interactions
even_indexes <- seq(2,b,2)
odd_indexes <- seq(1,c,2)
contig1 <- data.frame(x=Interactions[odd_indexes,1])
contig2 <- data.frame(x=Interactions[even_indexes,1])
m <- cbind.data.frame(contig1,contig2,deparse.level = 2)
colnames(m, do.NULL = TRUE, prefix = "col")
colnames(m) <- c("contig","partner")
x <- with(m, table(contig, partner))
y <- t(x)
u <- unique(c(rownames(x),colnames(x)))
imat <- matrix(0,ncol=length(u),nrow=length(u),dimnames=list(u,u))
i1 <- as.matrix(expand.grid(rownames(x),colnames(x)))
i2 <- as.matrix(expand.grid(rownames(y),colnames(y)))
imat[i1] <- x[i1]
imat[i2] <- imat[i2] + y[i2]


##perform clustering
##uses correlations between variables "as distance"
dd <- as.dist((1 - cor(imat))/2)
hc <- hclust(dd,method = "ward.D")
plot(hc) #to see a dendrogram of clustered variables
rect.hclust(hc, k = 4) #k = number of clusters as determined visually or by silhouette pltos
dataset <- cutree(hc, k = 4)
##to save dataset of the cluster each contig was assigned to:
#write.table(dataset, "file name", quote = FALSE, sep="\t")


##color dendrogram by species identity as determined by Hi-C
library(sparcl)
ColorDendrogram(hc, as.numeric(dataset$V1), main = 'Title')
```

```
##color dendrogram by species as determined by blasting contigs against reference
genomes (need outfile1)
identity = read.table('outfile1')
identity[,5] = (identity[,3]/100)*identity[,4]
identity = identity[,c(1:2,5)]
identity = identity[order(identity[,2], identity[,3], decreasing = T),]
identity = subset(identity, !duplicated(identity[,2]))
identity$V2 = sub("l.*", "", identity$V2)
dataset = as.data.frame(dataset)
dataset = merge(dataset, identity[,-3], by.x = "row.names", by.y = "V2", all.x = T)
dataset[is.na(dataset$V1),]
ColorDendrogram(hc, as.numeric(dataset$V1))
##view summary of clustering (how many contigs for each species)
table(dataset$V1)
##view summary of which clusters species got assigned to
table(dataset$V1, dataset$dataset)


##view clusters in 2D (colored by species as determined by reference genomes)
loc <- cmdscale(dd)
plot(x = loc[,1], y = loc[,2])
text(x = loc[,1], y = loc[,2], rownames(loc), cex = 0.5, col = as.numeric(dataset$V1), asp
= 1)


##view clusters in 3D (colored by species as determined by reference genomes)
##manually open xquartz prior to running this section if using a Mac
library(rgl)
open3d()
loc2 = cmdscale(dd, k = 3)
plot3d(x = loc2[,1], y = loc2[,2], z = loc2[,3], col=as.numeric(dataset$V1),
size=1,type='s')
##add legend
legend3d("topright", legend = paste(c('vector','of','species','names')), pch = 16,
      col = c('vector','of','color','names'), cex=1.2, inset=c(0.05))
##to save snapshot of 3D plot
snapshot3d(filename = 'filename', fmt ='png')
##to save snapshot in high def, save as pdf using rgl.postscript and convert to png later
rgl.postscript("filename", fmt = 'pdf', drawText = T)


##manually set colors (adjust vector accordingly for more than 5 species/plasmids)
col.pal <- palette()
col.pal[1:5] <- c('yellow','magenta','green','red','blue')
palette(col.pal)
##only run next command after done creating plots
```

```
palette("default")


##to determine optimal number of cluster by silhouette plot, run this command using a
##range of values instead of just 4
##chose number of clusters that gives the maximum average silhoutte width
library(cluster)
plot(silhouette(cutree(hc,4), dd))
```