UNIVERSITY OF CALIFORNIA SAN DIEGO

Analyzing Longitudinal Isolates in the Age of Affordable Sequencing

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Bioengineering

by

Jonathan Eugene Pekar

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Gert Cauwenberghs
Professor George Sakoulas

2018

The thesis of Jonathan Eugene Pekar is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2018

DEDICATION

To:

Family and friends, supporting me through all the trials and tribulations of graduate school.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I have encountered many valuable individuals during my time in graduate school, but there are a particular few I want to express gratitude for, as without them, none of this would have been possible.

Firstly, I would like to thank Bernhard Palsson, my advisor and source of support, who has helped inform my viewpoint on practical and theoretical elements of science which I had not previously considered. Within the Systems Biology Research Group (SBRG), Jonathan Monk has guided my thinking and foray into genetics, always encouraging me to both expand my scope and look deeper. All this said, I would like to thank the SBRG as a whole for the welcoming environment and insightful feedback.

In addition, I would like to thank the other members of my masters committee: Prof. Gert Cauwenberghs and Dr. George Sakoulas. Gert has provided mentorship since my first quarter at UCSD, and George has showcased how to further inform my perspective, but with a translational and medical point of view.

And lastly, the challenge of these past few years could not have been met without a few friends and family members by my side: Nafeesa, Dorothy, my parents, and my brother Eric. Each of you have played incredibly important roles in my life, and I am grateful for all the times you have listened to and supported me, especially when I was least sure of my own place here.

Chapter 3 in part is a reprint of material published in:

- **JE Pekar**, P Phaneuf, R Szubin, BO Palsson, A Feist, and JM Monk. 2018. “A Gapless, Unambiguous Genome Sequence for *E. coli* C - a Workhorse of Industrial Biology.” *ASM. Under review.* The thesis author was the primary author.

VITA

2014	Bachelor of the Arts in Biology, Williams College
2018	Master of Science in Bioengineering, University of California San Diego

ABSTRACT OF THE THESIS

Analyzing Longitudinal Isolates in the Age of Affordable Sequencing

by

Jonathan Eugene Pekar

Master of Science in Bioengineering

University of California San Diego, 2018

Professor Bernhard Ø. Palsson, Chair

Antimicrobial resistance (AMR) is becoming an increasingly important issue in healthcare, and a slowdown in discovery of new antibiotics necessitates an understanding of the development of this resistance. While there are various methods with which to study drug resistance, and various subtopics in this broad field, we have focused on understanding genetic mutations within bacterial pathogens. Here, we analyzed longitudinal isolates of bacterial pathogens within patients, as the patients are being treated with antibiotics. The first step in finding mutations is creating a high-quality reference genome. Reference genomes are often made with solely short-read DNA sequencing, but utilizing long-read DNA sequencing in tandem with short-read

sequencing can create a higher-quality genome. We employed long-read fragments to create a scaffold with which we aligned the short reads more accurately. Using the relatively new Oxford Nanopore MinION device, we began doing long-read sequencing in-house and found it to be an effective substitute for the more expensive PacBio long-read sequencing. We have slowly been developing a protocol for both wet and dry lab methods in acquiring and utilizing long-reads. For our longitudinal isolates, we used Illumina sequencing on all the samples and MinION sequencing on the base, or first, sample of each set of isolates. After creating a reference genome of the base strain, we compared short-read sequences from the subsequent isolates to the base strain. Collectively, these tools and methods allowed us to develop a pipeline for determining genetic mutations in longitudinal isolates of bacterial pathogens. This pipeline provides a starting point for understanding the pathogens phenotype and the relationship between antibiotics and bacterial evolution.

Chapter 1

Introduction to Long-read Sequencing

1.1 Importance of Genome Assembly

Genomics is a fundamental component in studying bacteria. The past decades have brought forth various sequencing technologies and reduced cost in sequencing, to the point where it is now commonplace and integral to science globally. In studying bacterial genomics, we seek to understand various elements of the organisms: evolution, antibiotic resistance, metabolic regulation, cellular signaling, etc. However, despite the rise of sequencing technologies and the drop in sequencing costs, assembling a high-quality genome remains a difficult and monetarily and computationally expensive task.

The reasons behind this are several-fold. Most obviously, genomes are magnitudes longer than the DNA fragments being analyzed [5]. Correctly assembling a genome involves a variety of computational tools and a high amount of computational resources, especially if one hopes to assemble genomes on a regular basis. The mathematics behind assembly revolves around creating a de Bruijn sequence of the fragments, which arranges based on overlaps between fragments [6]. While the fine-tuned details of the math behind assembly is beyond the scope of this thesis, it is simply important to note that the construction of a de Bruijn graph with DNA fragments can be done with fragments of various sizes and accuracies.

1.2 Moving Toward Closed Genomes

Genomics is dominated by short-read sequencing, particularly utilizing Illumina sequencing technologies. While the reads are typically only 150-250 base pairs (bp), they are very accurate and have a low cost per base [7]. However, these fragments are shorter than many of the repetitive regions found in genomes, including bacterial genomes [8]. Due to the commonality of repetitive regions in bacterial genomics, the assemblies generated by short reads contain may contiguous sequences, or contigs. For example, Bankevich et al. showed that with both a single-cell and multi-cell samples of *E. coli*, for every assembler tested, at least 195 contigs were created when assembling the genome [6], but they do not limit the ability to correctly order small portions of the genome. Therefore, we fail to get a closed genome, which is one where each chromosome is assembled as one sequence. For instance, we expect one chromosome for a sequence of *E. coli* [9]. If our assembly has one contig representing the one chromosome, the genome is closed; if there is more than one contig, it is not closed.

While a great deal of information and analysis can be done on assemblies that do not provide a closed genome, closing the genome is important for downstream genome analysis. For example, gaps in the genome, partly due to repetitive regions, can cause genes to be missing, especially if coverage is not high. Multicopy genes can also be collapsed into a single gene sequence, causing both a loss of genetic diversity and gene presence in the assembly (Ekblom and Wolf 2014). When only a fraction of the genes are available, the downstream analyses may become biased. Furthermore, repeat regions can be subject to spontaneous mutation rates and the loss and gain of repeats, such as with the spa gene (encoding the important virulence factor Staphylococcus Protein A) in S. aureus, which has 24-bp repeats [10, 11]. Repeat regions are important in understanding genetic regulation and antimicrobial resistance due to their place in plasmids [12–15]. Order-based genomic analyses are more feasible with a closed genome as well, allowing us to study gene regulation, operon structure, junction mutations, etc. more thoroughly [16]. Thus, it is clear that these factors are all important in understanding genetic mutations over time, and that tools that evaluate these mutations function better when the genome is closed.

1.3 Long Read Sequencing Today

Long-read sequencing can be used to close genomes and thereby address missed repeat regions and high contig counts. The long reads function as a scaffold for the short-reads and assist in making a closed genome [1]. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are the main sequencing platforms for long-read sequencing, and the platforms can provide reads of 5 kbp up to 20 kbp (and sometimes longer, but these are less common), and sometimes longer [17–19]. Due to the length of the DNA fragments, the reads can span across entire repeat regions, helping complete an assembly and reduce the number of contigs. Chin et al. demonstrated how, for various organisms including *E. coli*, assemblers were able to consistently produce fewer than 30 contigs, and often fewer than 10 contigs, when utilizing long reads [20]. However, long reads have a much higher per-base error rate than short reads (5-15% vs. <1%)

Table 1.1: Comparisons between PacBio and ONT MinION technologies, adapted from Birla [4].

Instrument	PacBio RS II	PacBio Sequel	ONT Minion
Average read length (kbp)	10 - 15	10 - 15	Up to 900 kb
Error rate (%)	10 - 15	10 - 15	5 - 15
Data output (Gb)	0.5 - 1	5 - 10	5 - 10
Read count	∼ 50,000	∼ 500,000	Up to 1,000,000
Price ($)	700,000	350,000	1,000
Run price ($)	400	∼ 850	500 - 1000

[20, 21]. While the accuracy is sufficient to produce an assembly, the assemblies utility for fine-grained genomic analyses is diminished. We will discuss the necessity of both long reads and short reads in Chapter 2, but first we will address the integration of long-read sequencing in the Systems Biology Research Group (SBRG).

1.4 Oxford Nanotechnologies MinION Sequencer in the lab

The PacBio platform has been an industry standard for several years now, but it is prohibitively expensive, as Illumina reads are substantially cheaper and are better suited for providing high accuracy [22]. While we have utilized this platform for a portion of our sequencing (Chapter 3.1: *E. coli C*), we have been seeking more affordable ways to implement these technologies in the SBRG (Table 1.1; note that the run price of the MinION is the cost of a flow cell and the sequencing kit, flow cells can be reused 1-2x, decreasing the price). Fortunately, in 2012, ONT released the MinION, a portable device for DNA and RNA sequencing [23]. The MinION is able to produce read lengths within the range above, but also ultra reads, with lengths greater than 100 kbp [24].

Furthermore the device can simply be plugged into a computer and used. Typically, DNA is extracted from samples of interest, prepared for the platform being utilized, and then shipped to a sequencing center. However, the MinION can be plugged directly into a computer and run in real time. Once the samples are inserted and the device begins to work, the results are produced in roughly 48 hours.

The output files of the MinION sequencer are in a FAST5 format, which is in fact an HDF5 data model [25]. This data model supports a variety of file types; our pipeline necessitates the extraction of the FASTQ files from the FAST5 files. The MinION has multiple types of basecalling (e.g., 1D, which utilizes information from only one strand, rather than two), and the type of basecalling allows for different computational tools [26]. However, the Albacore tool by ONT can work with each basecalling type and does not require specification, easing some of the burden of the individual doing the computational work [27]. Albacore takes FAST5 files and converts them into file types of interest (in our case, FASTQ) and demultiplexes samples if barcoding was used. If barcoding was used, a barcode was attached to up to 12 samples per each MinION sequencer, and when extracting the data, the files must be filtered by barcode, which Albacore is able to do.

Upon extraction of FASTQ files (and demultiplexing as needed), we are able to analyze the FASTQ files and use them for genome assembly. An initial analysis is critical, however, to ensure that a sufficient average read length and number of reads. Those metrics are likely sufficient, as our primary motive in this initial analysis is to determine coverage. The quality of the reads is already low, relative to Illumina-sequenced short reads, but we do still need to ensure enough coverage to provide the benefit of using long-reads to make the assembly. Assemblers are unable to produce a quality assembly if there is not a high enough coverage with the reads. If we find that the reads are too short (an average below 5 kbp) or the coverage is not high enough (below 30-50x), wet lab protocols must be adjusted to improve the quality of the sequencing run.

Chapter 2

Using Long-read Sequencing for Hybrid Assemblies

2.1 The Need for Hybrid Assembly

Long reads generated by MinIon technology have a much higher per-base error rate than Illumina short reads (5-15% vs. <1%) [20, 21]. While the accuracy is sufficient to produce an assembly (provided high enough coverage), the assemblies utility for fine-grained genomic analyses is diminished; we cannot determine point mutations, frameshifts, etc. with high confidence using these assemblies. Conversely, a short-read assembly can be better for picking up on point mutations due to the high accuracy of the reads, but other problems are introduced. For example, certain diseases, such as Huntingtons, are associated with short tandem repeats (STRs), and knowing the number of these repeats is critical in diagnosis and analysis of the disease [28]. With a short-read assembly, there is difficulty determining the number of repeats and what flanks a given region of the repeats [29]. In microbial genomics, antibiotic resistant genes and cassettes can be associated with STRs, so correctly integrating STRs in an assembly is critical in understanding AMR [30].

Furthermore, despite the developments in long-read technologies, public health and research laboratories primarily use Illumina sequencing due to their high accuracy and low cost [31]. Institutions will likely not utilize long-reads unless necessary to complete genomes of interest. As short reads are already available for many bacterial isolates, they are unlikely to be sequenced again with a long-read platform. However, much of the expense and inconvenience comes with requiring enough coverage to generate a long-read assembly. If that is desired, PacBio will typically be the platform of choice, as the platform can ensure high depth and higher accuracy than the ONT MinION [32]. A solution to this problem is to combine long-read and short-read technologies into what has become known as a hybrid assembly [1].

2.2 Benefits of Hybrid Assembly

Hybrid assembly can achieve the benefits of both long-read and short-read assemblies: a high-accuracy, low(er)-contig assembly. A hybrid assembly needs less long read depth compared to a long-read assembly, because it utilizes both long and short reads [1]. To reduce the long-read load, the hybrid assembly is completed using a short-read-first approach. The short-reads are first assembled into an accurate assembly graph (a de Bruijn graph), and then the long-reads are used as a scaffold to join the short-read-derived contigs together. A long-read-first approach is feasible as well, where the long-reads are assembled first and then error-corrected by the short reads. The latter approach requires a sufficient long-read depth to ensure that the initial assembly can be created [1]; coverage less than 25x has led to failed or erroneous assemblies. As Illumina sequencing is institutionally prevalent and would be supplemented by the long-reads,

Table 2.1: Read and assembly metrics for the hybrid assembly case studies (Chapter 3).

Sample	*E. coli* C	*S. aureus* LAC
Instrument	PacBio RS II	ONT MinION
Long reads	147015	187,271
Long reads mean	7229.6	4285.9
Long reads coverage	195.4	286.5
Short reads	3161393	3182905
Short reads coverage	145.9	113.7
Contigs in short-read assembly	165	38
Contings in hybrid assembly	1	2

we recommend the short-read-first approach to reduce cost in acquiring the necessary data for hybrid assembly.

The approach we suggest to minimize cost in attaining hybrid assemblies necessitates the ONT MinION. With low entry cost and low subsequent sequencing cost per sample, it can be utilized instead of the PacBio platform. However, learning and developing effective sample preparation protocols can be time-consuming and costly. Despite these initial issues, even the first sequenced samples have been effective in helping improve assemblies. While we strive for producing an average read length of 8-12 kb, we have been initially getting an average read length of 5-7 kb. However, we were still able to get high coverage because the MinION can generate many reads and then create low-contig assemblies. Therefore, we believe the cost and labor put into learning how to use the MinION are worthwhile.

2.3 Hybrid Assembly Protocol/Pipeline

We have assembled all the necessary steps for hybrid assembly with quality control and quality assurance into a detailed pipeline (Fig. 2.1). In step 1, we check to ensure the short reads and long reads are of sufficient quality and coverage to create an assembly. For short-reads, high-quality reads and high depth are critical and can be easily confirmed using FastQC. FastQC provides a quality control (QC) report which can locate problems from both the sequencer and

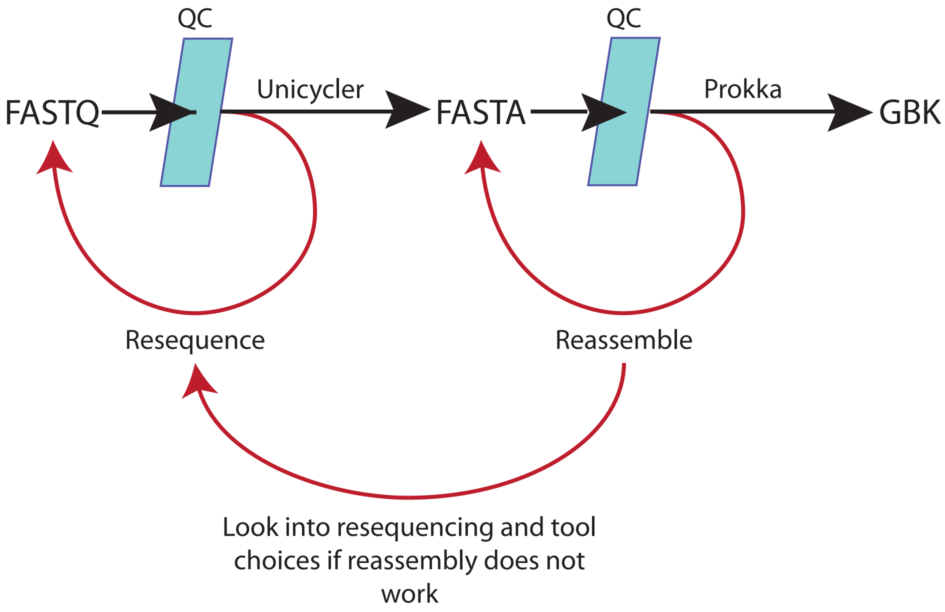
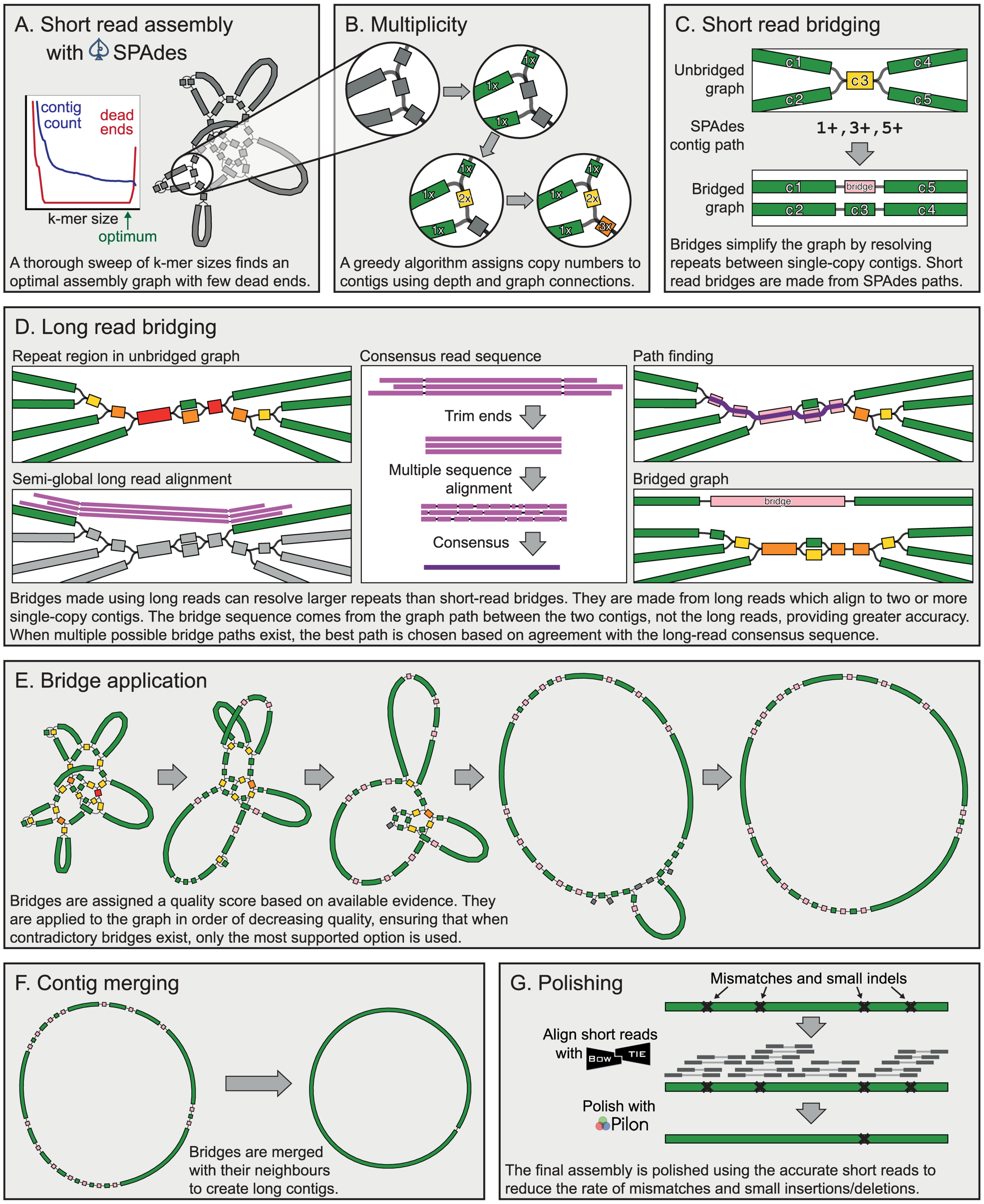

QC
QC
FASTQ
Unicycler
FASTA
Prokka
GBK
Resequence
Reassemble
Look into resequencing and tool choices if reassembly does not work

Figure 2.1: General pipeline to create hybrid assemblies after acquisition of FASTQ files.

the starting library material [33]. FastQC is meant for short reads, so for long reads we follow the protocol as above: determine the number of reads, average length, and coverage [34]. We are primarily looking for sufficient coverage and length and are less concerned with the quality of the reads. These reads will have lower quality than the short reads, but since they are being used to create a scaffold for the short reads, the low quality is acceptable.

In step 2, after confirming the short read coverage is at least 50x, the long read coverage is at least 15x, and the long read median length is at least 5-6 thousand bp, we utilize Unicycler to create the hybrid assembly [1]. The tool begins with the process of short-read assembly, using the St. Petersburg genome assembler (SPAdes), which sweeps through k-mer sizes to find an optimal de Bruijn graph from the short reads with as few dead ends as possible (Fig. 2.2) [6]. SPAdes then determines the multiplicity on contigs using depth and graph connections and uses repeat resolution (RR) in producing contigs from the assembly graph. Unicycler creates bridges from the graph by discriminating between the repeats and the single-copy contigs; if a path contains two or more anchor contigs, Unicycler creates a bridge from the path. Next, the tool relies on applying long reads to resolve the repeat regions and gaps in the short-read graph bridges and graph in general. If multiple bridge paths exist, Unicycler chooses the best path based on consistency with the long-read consensus sequence. The tool then ranks the bridges and assigns them a quality score. The most supported bridges are used if there are multiple and possibly contradictory bridges. When long reads allow for the creation of the appropriate bridge, contigs can then be merged, creating even longer contigs. Lastly, the short reads are aligned to the current assembly with Bowtie to find insertions and deletions (indels) and mismatches [35]. Pilon is then utilized to polish the draft assembly, including finding, fixing, and/or implementing single base differences, indels, and gap filling [36]. Unicycler then outputs a FASTA file with the assembly.

In step 3, we run QUality ASsessment Tool (QUAST) on the assembled genome file, which computes various metrics of the assembly, and then in step 4, we annotate the genome with Prokka [37,38]. The metrics showcase GC content, number of contigs, N50, etc., all of which inform how successful the assembly was. While a single-contig assembly is ideal, an assembly with fewer than five contigs has been sufficient for our purposes.


A. Short read assembly with SPAdes

A thorough sweep of k-mer sizes finds an optimal assembly graph with few dead ends.

B. Multiplicity

A greedy algorithm assigns copy numbers to contigs using depth and graph connections.

C. Short read bridging

Bridges simplify the graph by resolving repeats between single-copy contigs. Short read bridges are made from SPAdes paths.

D. Long read bridging

Repeat region in unbridged graph
Semi-global long read alignment
Consensus read sequence
Trim ends
Multiple sequence alignment
Consensus
Path finding
Bridged graph
bridge

Bridges made using long reads can resolve larger repeats than short-read bridges. They are made from long reads which align to two or more single-copy contigs. The bridge sequence comes from the graph path between the two contigs, not the long reads, providing greater accuracy. When multiple possible bridge paths exist, the best path is chosen based on agreement with the long-read consensus sequence.

E. Bridge application

Bridges are assigned a quality score based on available evidence. They are applied to the graph in order of decreasing quality, ensuring that when contradictory bridges exist, only the most supported option is used.

F. Contig merging

Bridges are merged with their neighbours to create long contigs.

G. Polishing

The final assembly is polished using the accurate short reads to reduce the rate of mismatches and small insertions/deletions.

Figure 2.2: Main steps of the Unicycler pipeline adapted from Wick et al. [1].

Chapter 3

Hybrid Assembly Case Studies

3.1 *Escheria coli* C

Escherichia coli C, an oft-used industrial strain, was originally isolated by Ferdinand Hueppe from soured cow's milk and described in his 1884 publication in German (Lange, Takors, and Blombach 2017; Bruschi et al. 2012). The C strain was used extensively by Bertani and collaborators in studies of phage P2 and by many others as a strain lacking a Type I restriction-modification system (Wiman et al. 1970). The strain was originally termed NCTC 122 at the National Collection of Type Cultures, London, where its entry states that it was deposited by the Lister Institute, London, in 1920 as Escherichia coli (Bertani and Weigle 1953). It was recently featured in a publication comparing seven commonly used *E. coli* platform strains and shown to have high anaerobic growth rates and predicted to have high relative production potential for propanol, butanol and 3-Hydroxypropanoate in anaerobic conditions (Monk et al. 2016). A draft genome-sequence was deposited in NCBI in 2016, with accession number MNKV00000000 (Monk et al. 2016). This assembly has 4,180 ambiguous base calls and chromosomal gaps ranging from

Table 3.1: Metrics of the reference and updated *E. coli* C assemblies.

Sequence	Reference	Updated
Contigs	165	1
Length (bp)	4538245	4164184
N50 (bp)	119166	4164184
CDSs	4205	4285
tRNAs	76	87
rRNA	8	8
GC content (%)	50.74	50.77
Ambiguous base calls	4180	0

45 to 125,000 bp with a mean gap size of 7,920 bp. While this genome has already been beneficial in multi-strain reconstruction work (Monk et al. 2016), analyses reliant on a pristine and accurate reference genome (e.g., single nucleotide polymorphism studies, analyzing repeat regions, etc.) are hindered by ambiguities and gaps. We therefore sequenced this strain utilizing PacBio single-molecule and Illumina short read sequencing and assembled the reads with Unicycler (version 0.4.2) (Wick et al. 2017). This produced an unambiguous genome sequence with no assembly gaps and an updated genome annotation.

We obtained the *E. coli* C strain from DSMZ. This strain also goes by Sinshelmer C (DSMZ 4860, ATCC 13706, NCIB 10544). Genomic DNA was prepared for PacBio and Illumina sequencing. PacBio libraries were prepared according to standard library preparation using Pacific Biosciences SMRTbell Template Preparation Reagent Kits. Libraries were size-selected to ¿10 kb using a PippinHT (Sage Sciences) and then sequenced on a PacBio RS II sequencer at the UCSD IGM Genomics in La Jolla, CA. Illumina libraries were generated using the TruSeq DNA Sample Preparation Kit (Illumina Inc., USA). The libraries were sequenced using the Illumina MiSeq platform with a paired-end protocol and read lengths of 150 nt.

The updated C genome consists solely of a 4,614,215 bp chromosome composed of one contig, compared to a 4,538,245 bp chromosome in the previous C assembly composed of 165 contigs (Table 3.1, 3.1). This gapless assembly eliminates 4,180 ambiguous base calls compared to

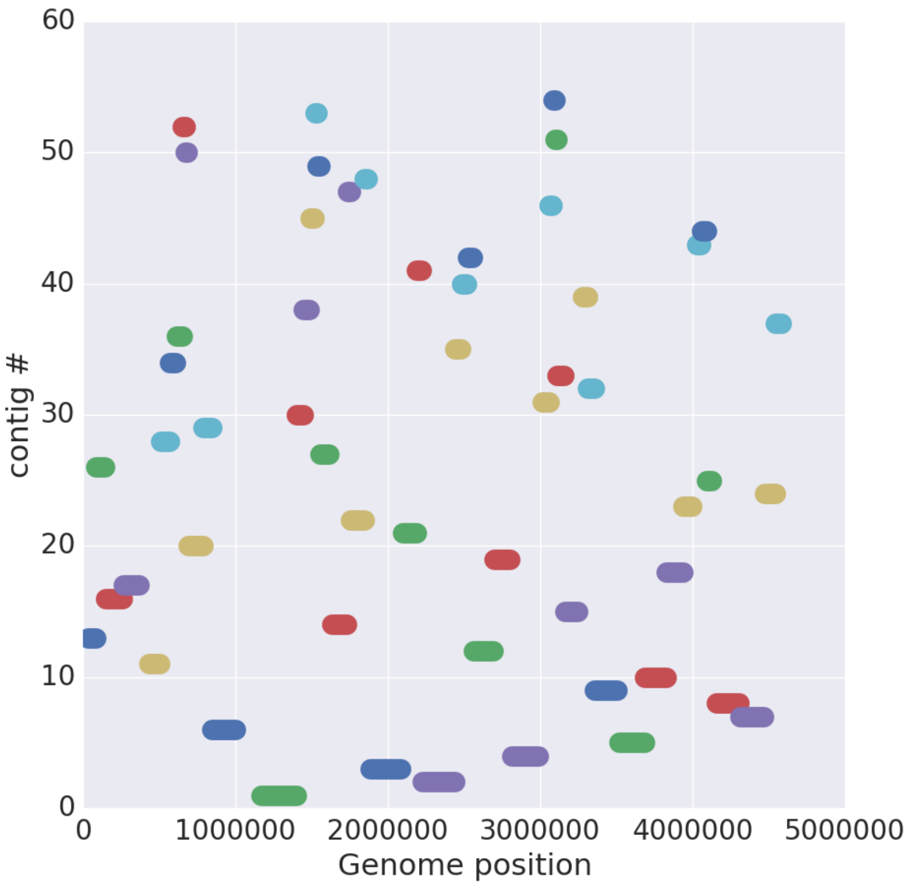


Figure 3.1: Contigs from 2016 *E. coli* C assembly mapped to updated assembly. Only showing the contigs that are at least 10,000 bp long.

the previous assembly . The final assembled genome was annotated using Prokka (version 1.12) (Seemann 2014). The updated genome has 4,284 annotated coding sequences (CDSs) compared with 4,205 CDSs in the previous reference annotation mentioned above.

This whole-genome project has been deposited in GenBank under the accession no. CP029371.

Table 3.2: Metrics of the reference and updated *S. aureus* LAC assemblies.

Sequence	Reference	Updated
Contigs	38	2
Length (bp)	2872137	2878171
N50 (bp)	265782	2875046
CDSs	2983	2652
tRNAs	56	70
rRNA	11	4
GC content (%)	32.62	32.78
Ambiguous base calls	0	0

3.2 *Staphylococcus aureus* LAC

Staphylococcus aureus contains various subgroups, each with many sequenced strains. Methicillin-resistant *S. aureus* (MRSA) represent a swath of *S. aureus* strains, and USA300 is a subgroup that accounts for some of the MRSA strains. USA300 is a methicillin-resistant clone that was isolated in September 2000, containing one circular chromosome and three plasmids. This strain is a source of many community-associated infections in the United States, Canada, and Europe. Many of its genes are clustered in novel allotypes of mobile genetic elements, and these mobile genetic elements tend to encode resistance and virulence features that are sources of the groups pathogenicity and fitness [39–41].

S. aureus Los Angeles County (LAC) is a clone used in many studies of USA300-0114, which is a particular strain of the USA300 subgroup [39, 42]. MRSA strains have been found to have between one and four plasmids [43]. The LAC strain is known to have two plasmids: pLAC01 and pLAC03, of lengths 27-28 kbp and 3 kbp, respectively [44].

The *S. aureus* LAC genome was previously submitted by Galac et al. (2018), with accession number NZ_PQBH00000000. The 2018 assembly by Galac et al. has 38 contigs, 2,983 genes, and is 2,872,137 bp long (Table 3.2).

While the LAC genome and genomes of similar strains (e.g., TCH1516, JE2) have already been useful in biofilm and phenotype screening work [42, 45], analyses reliant on an accurate

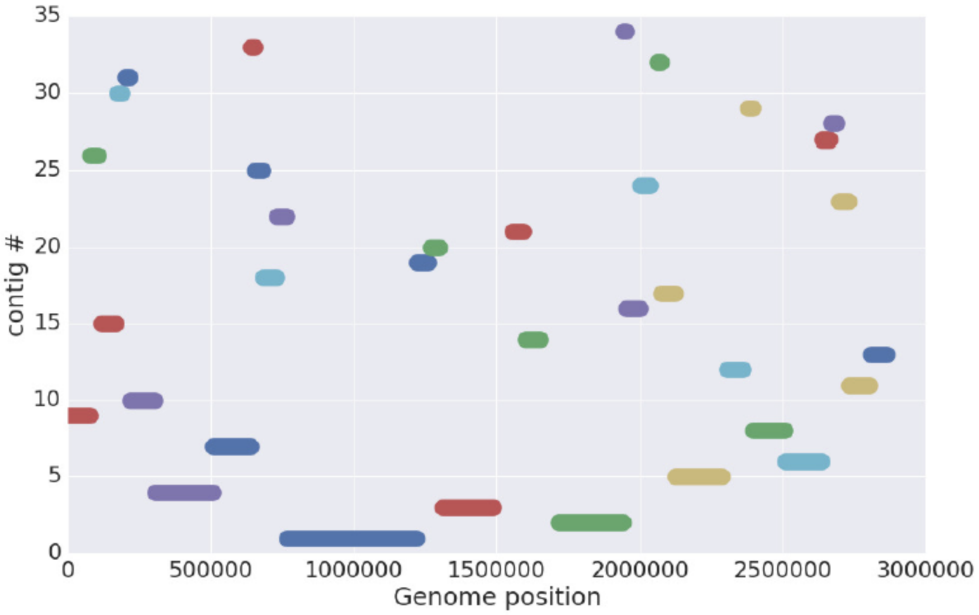


Figure 3.2: Contigs from the 2018 Galac et al. assembly mapped to updated assembly. Only showing the contigs that are at least 10,000 bp long.

reference genome with all its plasmids intact are hindered by the ambiguities and cured plasmid. We therefore sequenced this strain utilizing ONT MinION single-molecule and Illumina short-read sequencing and assembled the reads with Unicycler (version 0.4.2) [1]. This produced an unambiguous genome sequence with two contigs, no assembly gaps, and an updated genome annotation.

We obtained the *S. aureus* LAC strain from the Nizet group at UCSD, who has had it in storage for over five years. Genomic DNA was prepared for MinION and Illumina sequencing. The genomic DNA was isolated using a Zymo Quick-DNA Fungal/Bacterial Kit (catalog number D6007) and then sheared to 8 kb average size using a Covaris g-Tube (catalog number 520079). Libraries were prepared according to standard library preparation using ONTs Ligation Sequencing Kit 1D and then sequenced using the ONT MinION sequencer at the SBRG in La Jolla, CA. Illumina libraries were generated using the TruSeq DNA Sample Preparation Kit

(Illumina Inc., USA). The libraries were sequenced using the Illumina MiSeq platform with a paired-end protocol and read lengths of 150 nt.

The updated LAC genome consists of a 2,875,046 bp chromosome, compared to a 2,872,137 bp chromosome in the Galac et al. LAC assembly (Table 3.2, Fig. 3.2). Both of these assemblies contain the pLAC01 plasmid and pLAC03 plasmids. The final assembled genome was annotated using Prokka (version 1.12) [37]. The updated genome has 2,652 annotated coding sequences (CDSs) compared with 2,712 CDSs in the Galac et al. reference sequence.

Acknowledgements

Chapter 3 in part is a reprint of material published in:

- **JE Pekar**, P Phaneuf, R Szubin, BO Palsson, A Feist, and JM Monk. 2018. "A Gapless, Unambiguous Genome Sequence for *E. coli* C - a Workhorse of Industrial Biology." *ASM. Under review.* The thesis author was the primary author.

Chapter 4

Longitudinal Isolates: Analyzing Samples over Time

4.1 Utilizing the Hybrid Assembly

As bacteria become increasingly resistant to antibiotics, the risks surrounding chronic infections increase. Whole genome sequencing has provided ways to investigate the AMR and evolution of bacterial pathogens. Bacterial diversity and evolution has been considered limited relative to viruses, when exposed to various treatments and host environments. However, recent studies have showcased that short-term mutation rates of bacteria are of the order of 10-̂7 to 10-̂5 substitutions per site per year rather than the long-believed rates of 10-̂10 to 10-̂9 substitutions per site per year [46–48]. While this is still lower than the evolution rates of RNA viruses (10-̂4 to 10-̂3), bacteria have larger genomes and therefore could experience evolutionary rates similar to or higher than these viruses. As our understanding of within-host evolution and antibiotic resistance improves, we increasingly need to utilize new means of tracking and analyzing bacterial evolution.

Due to these high mutation rates, more traditional methods of analyzing a genome, such as multi-locus sequence typing (MLST), must be replaced with means of finer-grained analyses. Whole-genome sequencing (WGS) allows for such studies and has already been utilized to better understand bacterial evolution. If the reference genome is known, we can compare the sequenced samples to it and search for differences. However, this methodology is limited in helping understand the mutations between the pathogen in the host and the reference genome, rather than a clearer understanding of mutations within the patient over time. As these mutations can occur in various parts of the genome, including junctions, indels, and operons, having a closed genome becomes important as well. Mutations in regions that are either under- or over-represented in a genome that is not closed could also be under- or over-represented. The mutations and DNA fragments generally would map better to a closed genome, where we are more confident about the sequence and location of repeat regions, indels, etc.

4.2 Protocol/Pipeline

When analyzing longitudinal isolates, it is best to create a new reference genome from the first sample in the sequence (Fig. 4.1.C). Therefore, rather than having to compare the samples to each other and a foreign reference genome, one can limit the comparisons to samples from the host. To create the reference genome, we utilize long- and short-read platforms (MinION and Illumina, respectively) in order to acquire the DNA fragments needed to perform a hybrid assembly (Fig. 4.1.A, Fig. 4.1.C).

The ONT MinION sequencer was used to acquire long read data for all of the base strains of the longitudinal isolates (Table 4.1). While long read output was variable, in terms of both

Table 4.1: Read and assembly metrics for the longitudinal isolate base strains.

Sample	96.0	96.2	D592	Sparrow1
Long reads (bp)	45435	505219	53345	76719
Long reads mean (bp)	5549.3	5254.8	6781.4	8969.3
Long reads coverage	90.0	948.2	129.2	244.0
Short reads (bp)	2632150	2040088	1918711	198573
Short reads coverage	94.9	73.6	68.5	10.6
Contigs in short-read assembly	36	44	34	246
Contigs in hybrid assembly	2	3	2	5

mean length and read count, we were able to consistently acquire coverage greater than 50x. The hybrid assemblies were improvements over the short-read assemblies, decreasing the contig count.

The remainder of the samples only require short-read sequencing. We align the short reads to the reference genome and then undertake a comparative genome analysis, where we determine which positions in the genome differ between the various samples (Fig. 4.1.D). Typically, some of the differences are higher confidence than others, and so we trust that those differences are the ones that should be analyzed first. However, this method does not merely find point mutations; we are also able to find frame shifts, indels, and new junctions. To reiterate, this method requires several steps: aligning the reads to the reference genome, identifying genetic variation between the samples, and annotating these differences. We used breseq, a computational pipeline that automates the listed steps and is designed primarily for haploid microbial-sized organisms (¡20 Mb) [49]. The output comes in several formats: an HTML format, a Genome Diff flat file format that can be viewed with a text editor, and in community formats that can be used with various genome viewers (e.g., the Integrative Genomics Viewer from the Broad Institute).

After breseq is run and the initial results are available, we examine the mutations present among samples, where in the reference genome they occurred, etc. If certain regions are not annotated or are of particular interest and necessitate more detail, we BLAST the region against the NCBI database to see what matches occur.

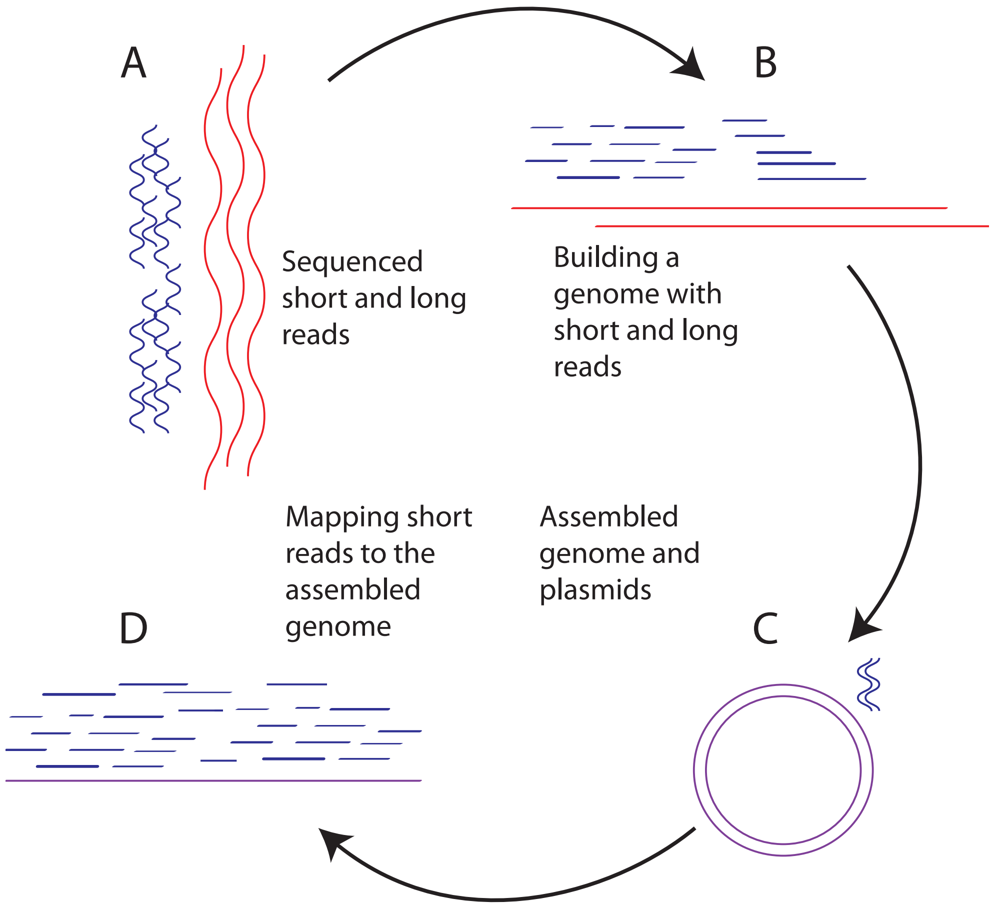


Figure 4.1: The general pipeline for a longitudinal analysis. **A.** Obtain short and long reads of the first sample in the sequence (e.g., using Illumina and the MinION, respectively). **B, C.** Build a de novo hybrid assembly of the first sample in the sequence (e.g., Unicycler). **D.** Align the short reads of the subsequent samples to the base assembly (e.g., breseq).

Chapter 5

Longitudinal Isolate Case Studies

5.1 Patient 96: *Staphylococcus aureus*

Our first case study is patient 96, a 50-to-60 year old male with a methicillin-resistant Staphylococcus aureus (MRSA) infection. He had a dedicated central line for nutrition following short gut 2/2 Crohns. He had multiple hospitalizations for recurrent and relapse bacteremias over the past twenty years. The patient has had multiple outpatient parenteral antimicrobial suppressive therapies, and he developed antibiotic tolerance during oxacillin, daptomycin, and dalbavancin therapies.

We have eight samples from the patient, between the years of 2011 and 2017 (Table 5.1, Fig. 5.1). The clinicians believed that during this time, the particular strain of MRSA changed. We set out to determine the strains of MRSA present in the patient and how they evolved over time.

Upon receiving the samples, we used the Illumina sequencing platform to acquire short read data for each of them. In addition, we utilized the MinION sequencer to acquire long read

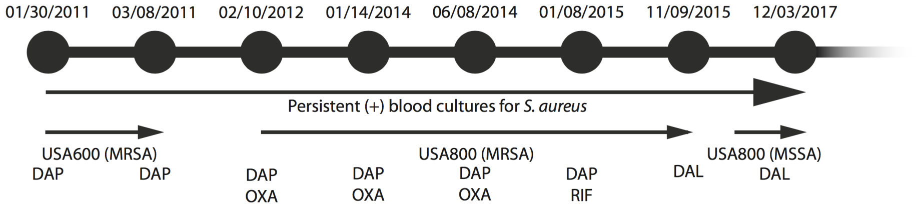


Figure 5.1: Timeline of case study of MRSA bacteremia. All isolates were subject to analysis.

.

Table 5.1: Samples from patient 96 and their collection dates.

Sample	Date collected
96.0	1/30/2011
96.1	3/8/2011
96.2	2/10/2012
96.3	1/14/2014
96.4	6/8/2014
96.5	1/8/2015
96.6	11/9/2015
96.7	12/3/2017

Table 5.2: Assembly metrics for three of the samples from patient 96.

Sample	Type of assembly	# contigs	Total length	N50 (bp)	GC content (%)
96.0	Hybrid	2	2,820,784	2,796,131	32.96
96.2	Hybrid	3	2,763,882	2,736,680	32.76
96.7	Short-read	48	2,740,521	194,037	32.70

data for samples 96.0 and 96.2.

We used a bidirectional BLAST with the samples assembled genome files to determine the similarity between the 8 samples. Using a percentage identify cutoff of 99, we BLASTed the sequence of each of each sample against the sequences of the genes in sample 96.0 (Fig. 5.2). Sample 96.0 had approximately 2600-2700 genes, and the gene content of sample 96.1 was very similar, whereas the gene content of samples 96.2 thought 96.7 differed by at least 1000 genes, suggesting that samples 96.0 and 96.1 were a different strain than the rest of the samples.

While samples 96.2 through 96.7 all had similar amounts of genes match to sample 96.0, sample 96.7 was acquired approximately two years after sample 96.6 (Fig. 5.1). Due to this, we decided to BLAST each sample against 96.2 to discover whether sample 96.7 might be a different

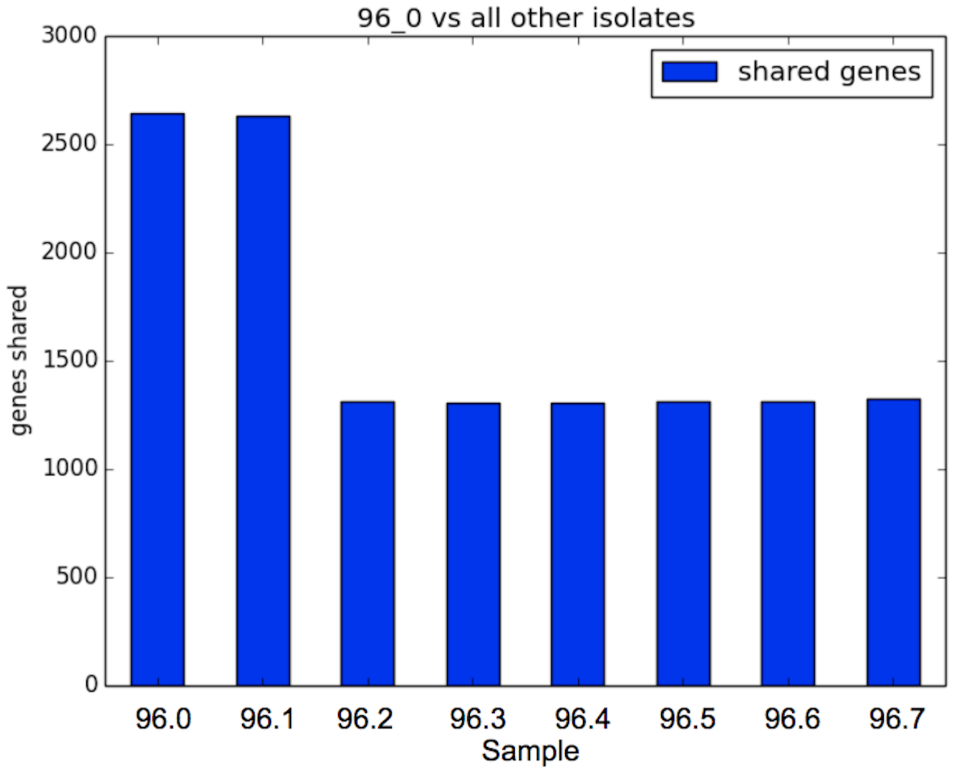


Figure 5.2: Bidirectional BLAST of all the samples against sample 96.0.

strain and to uncover anything we might have missed with the previous bidirectional BLAST (Fig. 5.3. This BLAST analysis showed that 96.7 appeared to lack 50-100 genes that samples 96.2 thought 96.6 have, suggesting that it could potentially be a different strain as well.

After being advised by the clinicians that the patient initially was infected with S. aureus USA600 and then was infected/recolonized with S. aureus USA800, we confirmed that the first two samples (96.0 - 96.1) are the same strain, the following five samples a different strain (96.2 - 96.6), and the last sample (96.7) a third strain. Upon BLASTing with NCBI, sample 96.0 best matches a USA600 isolated (strain AR466, accession number CP029080), sample 96.2 best matches the USA800 isolate (strain ECT-R 2, accession number FR714927), and sample 96.7 is a USA800 isolate as well, but it a MSSA isolate, rather than a MRSA one (S. aureus SA564,

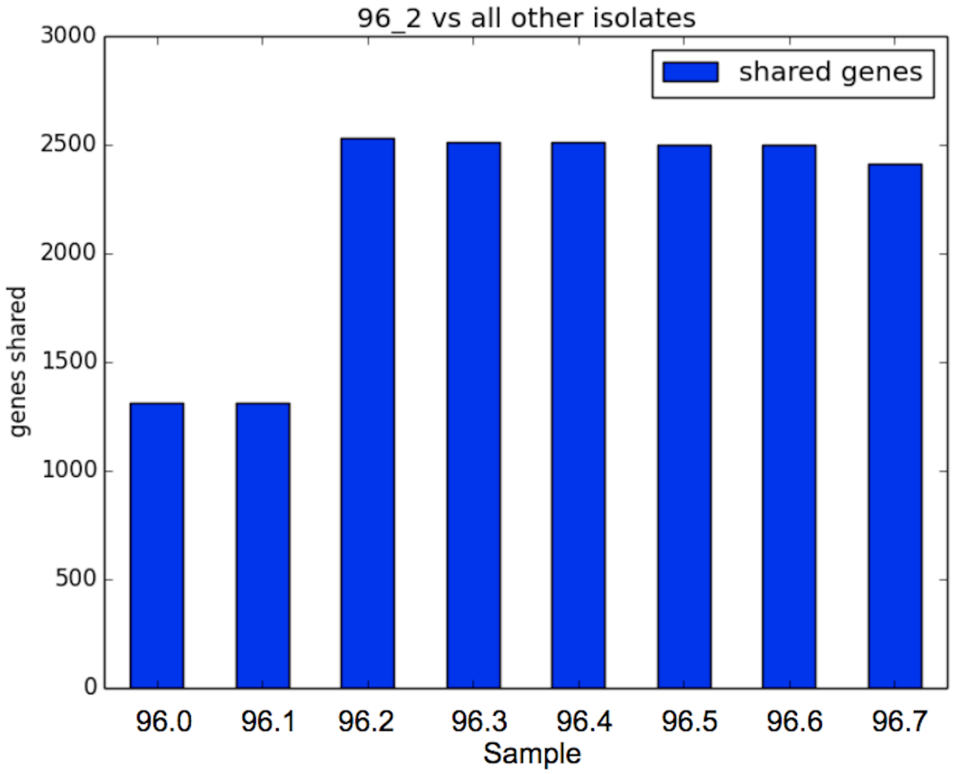


Figure 5.3: Bidirectional BLAST of all the samples against sample 96.2.

accession number CP010890).

We performed a hybrid assembly for samples 96.0 and 96.2, and a short-read assembly for sample 96.7. The assembly for 96.0 has 2 contigs, an N50 of 2,796,131 bp, a total length of 2,820,784 bp, and a GC content of 32.96%. The assembly for 96.2 has 3 contigs, an N50 2,736,680 bp, a total length of 2,763,882 bp, and a GC content of 32.76%. The assembly for 96.7 has 48 contigs, an N50 of 194,037 bp, a total length of 2,740,521 bp, and a GC content of 32.70

Next, we wanted to study the longitudinal evolution of each specific strain. We started by comparing 96.1 to 96.0 using breseq. We were only able to find one mutation with high confidence: a coding mutation in dapE, a putative succinyl-diaminopimelate desuccinylase. There were more mutations in S. aureus USA800, as there were more samples and the bacteria was present in the

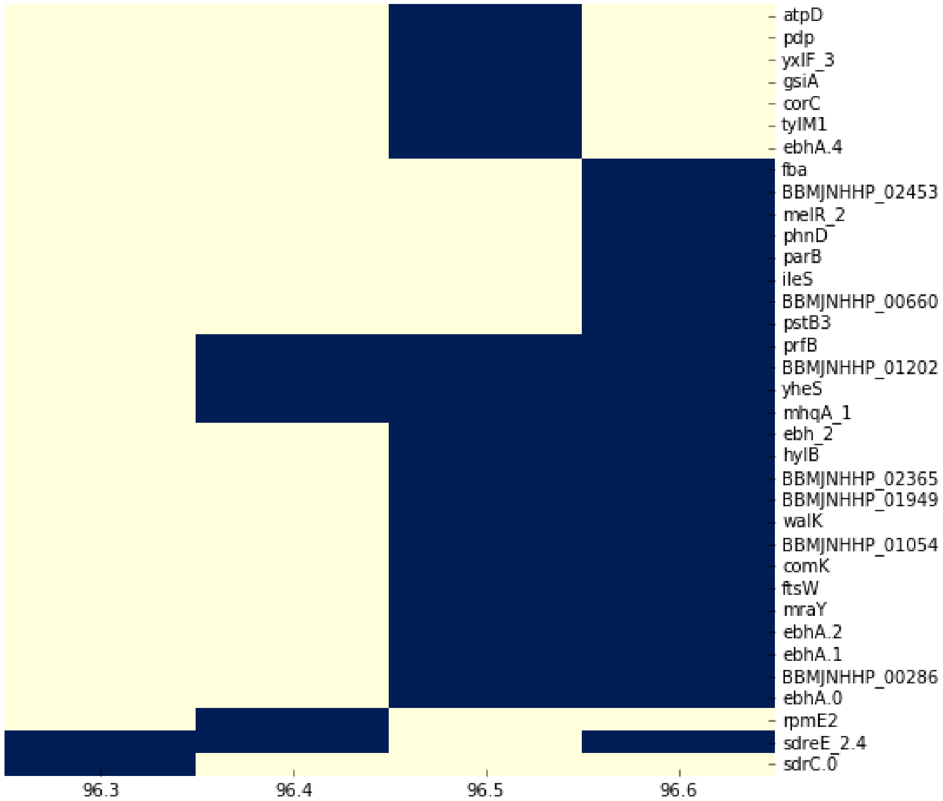


Figure 5.4: Subset of mutations in the 96.3 to 96.6 samples from the breseq results. All genes that did not have mutations for each sample of 96.3 through 96.6 were included. Blue indicates a mutation has occurred.

host for a greater period of time (Table 5.3). Many of these mutations were fixed by sample 96.5, including ebhA (extracellular matrix-binding protein) and walK (sensor protein kinase) (Fig. 5.4). Ebh is a large protein with many repeats found in staphylococci, responsible for tolerance to osmotic shock and binding sugars [50]. The walK gene codes for a sensory kinase, and point mutations in the gene are known to be responsible for conferring vancomycin resistance [51]. Furthermore, mutations in the gene cause decreased expression of genes associated with cell wall metabolism, a thickened cell wall, and decreased autolytic activity.

Table 5.3: Mutations from Figure 5.4, excluding the genes that only had a mutation in one sample. All the mutations listed occurred in 2-3 samples from samples 96.3 to 96.6.

Mutation	Annotation	Gene	Description
A→T	intergenic (–393/+40)	prfB ← / ← secA_1	Peptide chain release factor 2/Protein translocase subunit SecA
C→T	R243H (CGC→CAC)	yheS ←	putative ABC transporter ATPbinding protein YheS
T→G	V136G (GTC→GGC)	mhqA_1 →	Putative ringcleaving dioxygenase MhqA
T→C	intergenic (+15/+304)	ebh_2 → / ← mtlD	Extracellular matrixbinding protein ebh/Mannitol1phosphate 5dehydrogenase
A→G	intergenic (–273/ –135)	hylB ← / → adhR	Hyaluronate lyase/HTHtype transcriptional regulator AdhR
A→G	S221P (TCC→CCC)	walK ←	Sensor protein kinase WalK
A→G	K115E (AAA→GAA)	comK →	Competence transcription factor
A→T	L109I (TTA→ATA)	ftsW ←	putative peptidoglycan glycosyltransferase FtsW
C→A	intergenic (–269/+23)	mraY ← / ← pbpB	PhosphoNacetylmuramoylpentapeptide transferase/Penicillinbinding protein 2B
G→T	V2404L (GTG→TTG)	ebhA.2 →	Extracellular matrixbinding protein EbhA
G→A	R1313H (CGT→CAT)	ebhA.1 →	Extracellular matrixbinding protein EbhA
C→T	R552C (CGC→TGC)	ebhA.0 →	Extracellular matrixbinding protein EbhA
C→T	T1124T (ACG→ACA)	sdrE_2 ←	Serineaspartate repeatcontaining protein E

5.2 D strains: *Staphylococcus aureus*

A 50-something male patient presented with fever, cough, and dysuria, and had a prior history of diabetes mellitus, emphysema, tracheomalacia, hypothyroidism, idiopathic thrombocytopenia, osteoporosis, and left foot osteomyelitis requiring below knee amputation [2]. He experienced prolonged and persistent MRSA bacteremia, and after 21 days of positive S. aureus USA100 blood cultures, he was placed on a combination therapy of daptomycin and nafcillin. The bacteremia was then resolved within 48 hours. However, on day 28 of his hospitalization, he was diagnosed with sacral osteomyelitis and was treated for nosocomial Pseudomonas aeruginosa pneumonia.

The two samples extracted from the patient are from day 1 and day 21, and they are, respectively, sample D592 and D712 (Fig. 5.5) [2]. We did short-read sequencing on both samples, and used the MinION platform to get long reads from D592. We then performed a hybrid assembly on D592 and used breseq to compare D712 to D592.

The D592 sample has 2 contigs, an N50 of 2,820,177 bp, a total length of 2,847,444 bp, and a GC content of 32.85% (Table 5.4).

There were many mutations, with most on annotated proteins, and some of the annotated

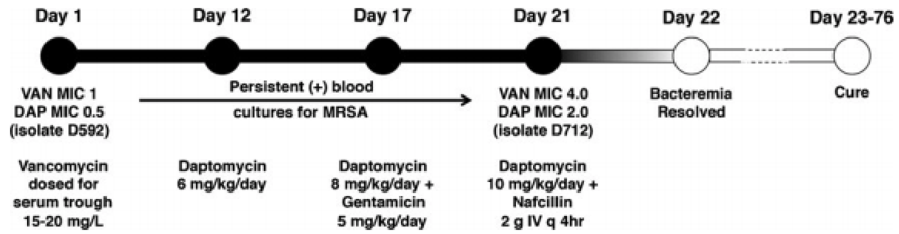


Figure 5.5: Timeline of case study of MRSA bacteremia. Isolate D592(daptomycin-susceptible) and D712 (daptomycin-nonsusceptible) were subject to analysis in the original study and in this thesis. Adapted from Sakoulas et al. (2013) [2].

Table 5.4: Assembly metrics for the D592 strain.

Sample	Type of assembly	# contigs	Total length	N50 (bp)	GC content (%)
D592	Hybrid	2	2,847,444	2,820,177	32.85

Table 5.5: Mutations in strain D712 from the breseq results. Mutations selected were ones that were not entirely hypothetical proteins.

Mutation	Annotation	Gene	Description
C→T	intergenic (+161/ −240)	ydeN → / → KJJPFECL_00163	Putative hydrolase YdeN/tRNAMet
T→C	I29I (ATT→ATC)	srrB →	Sensor protein SrrB
A→G	L341S (TTA→TCA)	mprF ←	Phosphatidylglycerol lysyltransferase
G→A	A588V (GCA→GTA)	tkt ←	Transketolase
T→A	I334F (ATC→TTC)	patA_2 ←	Peptidoglycan Oacetyltransferase
G→A	R24* (CGA→TGA)	bceA_1 ←	Bacitracin export ATPbinding protein BceA
C→T	intergenic (−165/+146)	pth ← / ← rplY	PeptidyltRNA hydrolase/50S ribosomal protein L25
T→C	E341G (GAA→GGA)	pflB ←	Formate acetyltransferase
C→T	A151V (GCT→GTT)	ysdC_2 →	Putative aminopeptidase YsdC
A→C	D282A (GAT→GCT)	yfmJ →	Putative NADPdependent oxidoreductase YfmJ

mutations were of particular interest (Table 5.5). Mutations of interest included coding mutations of srrB (sensor protein) and patA_2 (peptidoglycan O-acetyltransferase). The SrrB protein, coded for by srrB, is acts in the global regulation of staphylococcal virulence factors, in response to environmental oxygen levels [52,53]. Two-component regulatory systems affiliated with virulence and antibiotic resistance have become targets of interest for antimicrobial therapy against Gram-positive pathogens [54], showcasing that breseq can be used as a starting point for genetic, and perhaps mechanistic, investigation.

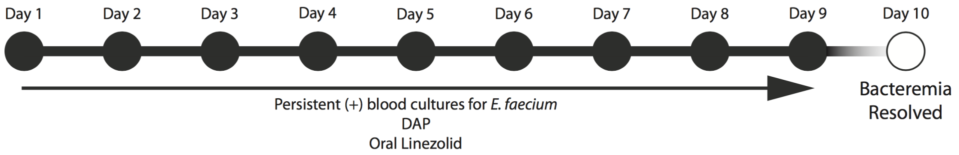


Figure 5.6: Timeline of case study of *Enterococcus faecium.* All nine isolates were subject to analysis.

5.3 Sparrow: *Enterococcus faecium*

A patient of unknown age and sex presented with Enterococcus faecium infection, and they were treated with daptomycin and oral linezolid. They were treated over the course of nine days, and the infection was resolved after the ninth day. There were nine samples total: Sparrow1 through Sparrow9 (Fig. 5.6).

We performed short-read sequencing of all nine samples at a local UCSD sequencing center, and then MinION long-read sequencing of Sparrow1 in our laboratory. We then performed a hybrid assembly on Sparrow1 and ran breseq on the remainder of the samples, comparing them to Sparrow1.

The Sparrow1 assembly has 5 contigs, an N50 of 2,870,749 bp, a total length of 3,157,023 bp, and a GC content of 37.77%.

The eight longitudinal isolates have numerous mutations, especially on hypothetical proteins, over the short course of treatment, and some of the annotated mutations were of particular interest (Table 5.6, Fig. 5.7). Mutations of interest included coding mutations of pbpX_1 (penicillin-binding protein), glpQ_1 (glycerophosphodiester phosphodiesterase), sdrD (Serine-aspartate repeat-containing protein D; a marginal prediction, but this is actually a silent mutation, so the amino acid stays the same), and bca (C protein alpha-antigen). The pbpX genes code for PBP2x proteins, which are responsible for binding to penicillin, and mutations in these

Table 5.6: Assembly metrics for the base Sparrow strain.

Sample	Type of assembly	# contigs	Total length	N50 (bp)	GC content (%)
Sparrow1	Hybrid	5	3,157,023	2,870,749	37.77

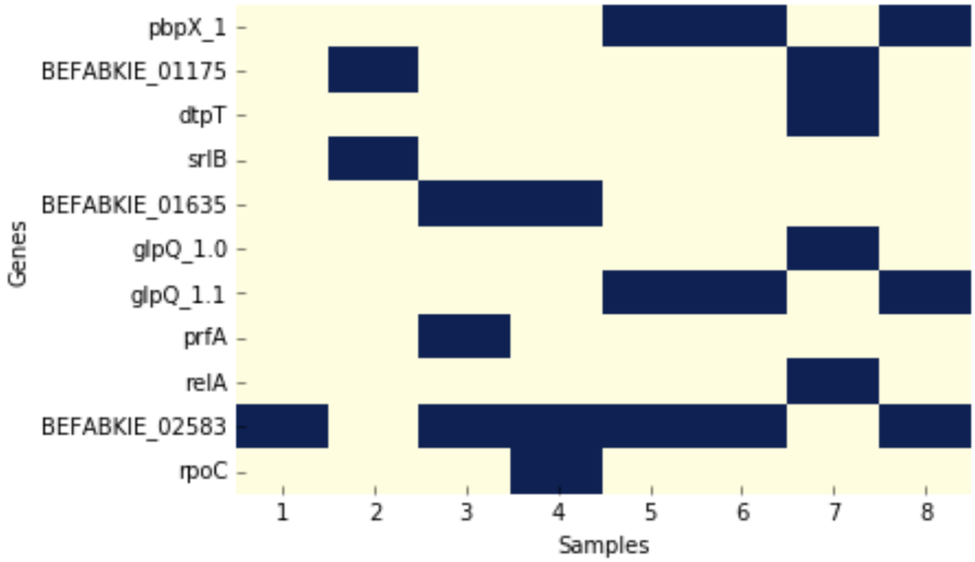


Figure 5.7: Subset of mutations of the *E. faecium* samples from the breseq results. All genes that did not have mutations for each sample of samples 1 through 8 were included.

can cause them proteins to have low affinitiy for penicillin [55, 56]. Interestingly, pbpX is also involved in a two-component transducing system, like srrB above [56]. The glpQ gene is involved in fatty acid and phospholipid metabolism. The gene tends to be down-regulated in the presence of certain bacteriocins, but its mechanism and exact role in resistance to bacteriocins has yet to be determined [57]. The bca gene is a virulence determinant and confers protective immunity, as well as immune system evasion, and while present in E. faecium, it is more commonly studied in streptococci [58–61].

Despite the samples only spanning nine consecutive days, there are similar amounts of mutations compared to the USA800 set of isolates from patient 96, which were collected over several years. This is due to E. faecium being a much more volatile and easily mutated organism, whereas MRSA has a lower mutation rate.

Table 5.7: Mutations in the Sparrow2 to Sparrow9 samples from the breseq results. Mutations were selected based on them being fixed or occurring in the later isolates, with a preference for genes with known products.

Mutation	Annotation	Gene	Description
A→C	T539P (ACA→CCA)	pbpX_1 →	Penicillinbinding protein 2X
(T)5→6	intergenic (+132/+429)	dtpT → / ← BEFABKIE_01284	Di/tripeptide transporter/hypothetical protein
A→G	K335E (AAA→GAA)	fhs1 →	Formatetetrahydrofolate ligase 1
+45 bp	intergenic (–18/ –348)	srlB ← / → BEFABKIE_01534	PTS system glucitol/sorbitolspecific EIIA component/hypothetical protein
T→A	E184V (GAA→GTA)	BEFABKIE_01635 ←	hypothetical protein
A→G	H29R (CAT→CGT)	glpQ_1.0 →	Glycerophosphodiester phosphodiesterase
T→A	L93H (CTT→CAT)	glpQ_1.1 →	Glycerophosphodiester phosphodiesterase
G→A	N151N (AAC→AAT)	prfA ←	Peptide chain release factor 1
T→A	K560* (AAA→TAA)	relA ←	GTP pyrophosphokinase
(A)5→6	intergenic (+6/+103)	BEFABKIE_02583 → / ← lysS	hypothetical protein/LysinetRNA ligase
T→C	E775G (GAA→GGA)	rpoC ←	DNAdirected RNA polymerase subunit beta’

Table 5.8: Mutations in the Sparrow2 to Sparrow9 samples from the breseq marginal predictions. Mutations were selected based on them being fixed or occurring in the later isolates, with a preference for genes with known products.

Mutation	Annotation	Gene	Description
A→G	L745L (TTA→CTA)	sdrD	Serineaspartate repeatcontaining protein D
C→T	intergenic (–3310/+117)	BEFABKIE_02580/BEFABKIE_02581	tRNAVal/tRNAAla
1 bp	intergenic (+127/+434)	dtpT/BEFABKIE_01284	Di/tripeptide transporter/hypothetical protein
C→T	intergenic (–521/+220)	polC_1/exoA	DNA polymerase III PolCtype/Exodeoxyribonuclease
C→T	G776G (GGG→GGA)	bca.0	C protein alphaantigen
C→T	E901E (GAG→GAA)	bca.1	C protein alphaantigen
C→T	E1010E (GAG→GAA)	bca.2	C protein alphaantigen

Chapter 6

Scale and Scope: Analyzing Longitudinal Isolates in the Clinic

The large-scale generation of genetic data holds the potential to increase our understanding of microbial evolution and drug resistance. With the price of sequencing decreasing, more versatile sequencers being created, and better computational tools with which to analyze data, there are more opportunities and capabilities to analyze and circumvent antibiotic resistance. The MinION serves as a starting point of fast and efficient sequencing; hybrid assemblies allow for more accurate analyses of sequencing data; and longitudinal analyses of bacteria provides insight into their evolution during infection of the host and treatment.

6.1 Putting together the Pipeline

The past two decades have seen a dramatic decrease in the price of sequencing, exceeding even Moores law since 2007 (Figure 6.1) [3].With the advent of new technologies, sequencing

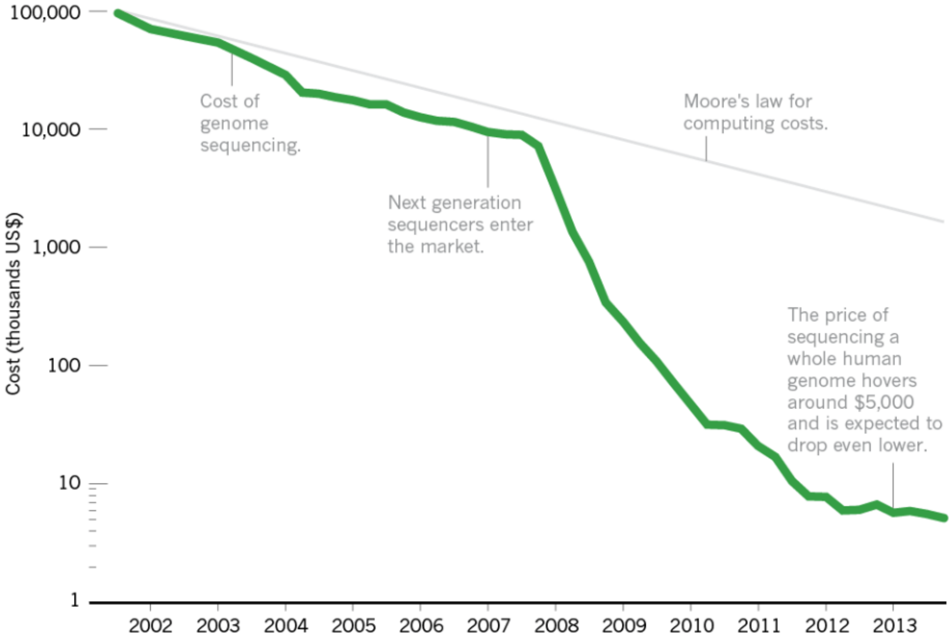


Figure 6.1: Cost of sequencing between 2002 and 2013, adapted from Check Hayden [3].

organisms has become efficient and commonplace. Older genomic techniques, such such as multi-locus sequence typing (MLST), pulsed-field gel electrophoresis (PFGE), variable-number tandem repeats (VNTR), and multi-locus enzyme electrophoresis (MLEE) have low resolutions relative to whole genome sequencing (WGS) and subsequent analysis techniques, such as variant calling [62–64]. As such, WGS has also become the first step in screening for genetic mutations of bacterial pathogens.

Short-read sequencing is already common in clinics and research facilities worldwide. However, its limitations prevent more complete genetic analyses, and to get a larger picture of the genetic landscape of an organism, long-read sequencing is necessary, whether through a third party or in-house. If in-house and more immediate (and affordable) methods are desired, the ONT MinION is a good solution due to its lower price point, ease of use (despite the learning curve), and ability to generate large amounts of data.

Regardless of which short- and long-sequencing platforms are used, once both data types are acquired, they can be utilized to create a hybrid assembly. The assembly tends to have fewer contigs than a short-read assembly, while still maintaining precision and accuracy due to the short reads. The long reads enable the contigs to be better bridged together and help resolve the location of repeat regions in the genome.

In analyzing longitudinal isolates from a patient, our goal was to develop a means to understand how the bacteria was evolving *in vivo*, under the stress of treatment. To do so, we first needed to short-read sequence all the samples. However, comparing to an external reference strain can be problematic, as that strain is only a close match, and perhaps not indicative of the current genomic state of the strain within the patient, even before treatment. However, by using the first strain in a set of longitudinal samples, we can understand how the researched strain is evolving over time. To best do this, we use a long-read sequencing platform for the base strain in order to be able to make a hybrid assembly. After applying quality control to the assembly, we compare the later samples against the base strain. Fortunately, computational tools such as breseq exist, easing the burden on the bioinformatician to initiate an analysis. Breseq can find point mutations, indels, junctions, etc., and it is made for a haploid microorganism [49].

However, this pipeline is simply the beginning of an analysis, rather than the end. Upon finding mutations, both high- and low-confidence, the organism can be examined in various ways, both computationally experimentally in a wet lab setting. This includes finding correlations between mutation occurrence and drug treatment, modeling the organism computationally and studying drug treatment and evolution of the organism in simpler settings, relative to in vivo.

6.2 Longitudinal Isolates in the Future

As the amount of sequencing data increases, there will be more opportunities for analyzing and understanding pathogens, both in and outside of hospitals. The increasing application of WGS to the clinic has helped elucidate the genomic epidemiology of various organisms. We have showcased a pipeline to aid understanding of short- and long-term within-host evolution. There is growing evidence that short-term evolution is actually greater than long-term when in patients, and can result in a more genomically diverse population [46,65,66]. Furthermore, WGS is able to help track the spread of pathogens in complex healthcare pathways within one or more hospitals in a particular region [67,68]. Whole genome sequencing has the potential to trace human-to-human transmission as well, and it has already been useful in understanding intercontinental spread and global geographic phenotypes [69, 70]. By researching pathogen evolution and spread from a small scale (within-host evolution and human-to-human transmission) to a global scale, we can better elucidate and track the emergence of new phenotypes, aiding in both immediate treatment development and future prevention efforts [71]. Analyzing longitudinal isolates can move from simply understanding evolutionary dynamics of a pathogen within a patient to understanding a pathogen historically and as it spreads and even becomes what might be currently classified as a different strain. Our pipeline can serve as an integral tool in these developments.

The MinION is already being applied more readily in research institutions, hospitals, and even in space [72, 73]. With its low price point and efficiency of use, it has become more widespread and has the potential for large-scale application. With large clinical and healthcare burdens tied to antibiotic resistance and pathogenic evolution in general, and the rate of drug development decreasing while the rate of important genes are found, urgent action is necessary (Figure 6.2) [74, 75].

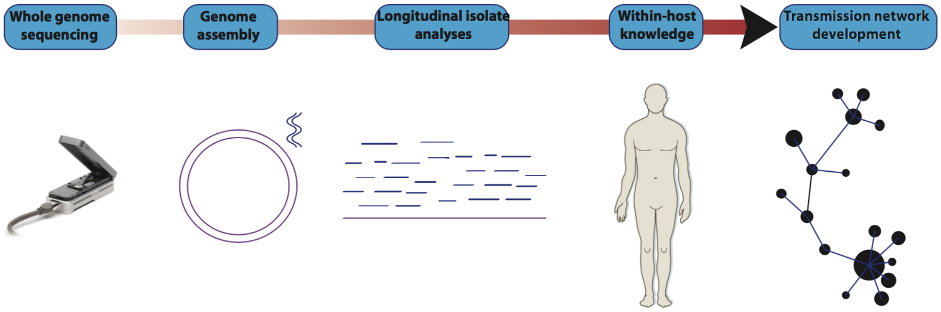

Whole genome sequencing
Genome assembly
Longitudinal isolate analyses
Within-host knowledge
Transmission network development

Figure 6.2: Workflow for moving from WGS to generating transmission networks.

These burdens are likely to increase over time, but fortunately, the cost to study pathogens is likely to continue decreasing [76]. Using the MinION and WGS generally is one of the various tools that can be utilized in combating the human-pathogen arms race. By tracking more pathogens in more areas globally, we can move toward understanding the small- and large-scale patterns of the pathogens. Implementing WGS, hybrid assembly, and longitudinal isolate analysis at a large scale is a costly, but reasonable, step in achieving these goals. Furthermore, as the precise understanding of AMR burdens have various uncertainties, it is critical to carefully model future scenarios [77]. Understanding the evolution of antibiotic resistant pathogens can aid the development of population-based disease transmission surveillance networks (Figure 6.2).

The benefits and importance of affordable sequencing, hybrid assemblies, and longitudinal isolate analyses are not limited to single-patient studies. While these studies are an important first step in implementing such techniques and pipelines, the long-term opportunities for them lie in better prediction, treatment, and prevention of pathogens and pathogen spread. By understanding the evolution and resistance patterns of organisms within-host, between-hosts, geographically, and historically, better and more accurate action can be taken to tackle disease burdens locally and globally.

Bibliography

[1] Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13: e1005595.

[2] Sakoulas G, Okumura CY, Thienphrapa W, Olson J, Nonejuie P, Dam Q, Dhand A, Pogliano J, Yeaman MR, Hensler ME, Bayer AS, Nizet V (2014) Nafcillin enhances innate immune-mediated killing of methicillin-resistant staphylococcus aureus. J Mol Med 92: 139–149.

[3] Check Hayden E (2014) Technology: The $1,000 genome. Nature News 507: 294.

[4] Birla B (2017). PacBio vs. oxford nanopore sequencing. https://blog.genohub.com/2017/06/16/pacbio-vs-oxford-nanopore-sequencing/. Accessed: 2018-7-18.

[5] Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20: 265–272.

[6] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19: 455–477.

[7] Werner JJ, Zhou D, Caporaso JG, Knight R, Angenent LT (2012) Comparison of illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. ISME J 6: 1273–1276.

[8] Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 34: D32–6.

[9] Kohara Y, Akiyama K, Isono K (1987) The physical map of the whole e. coli chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. Cell 50: 495–508.

[10] Harmsen D, Claus H, Witte W, Rothgänger J, Claus H, Turnwald D, Vogel U (2003) Typing of methicillin-resistant staphylococcus aureus in a university hospital setting by using novel software for spa repeat determination and database management. J Clin Microbiol 41: 5442–5448.

[11] Shopsin B, Gomez M, Montgomery SO, Smith DH, Waddington M, Dodge DE, Bost DA, Riehman M, Naidich S, Kreiswirth BN (1999) Evaluation of protein a gene polymorphic region DNA sequencing for typing of staphylococcus aureus strains. J Clin Microbiol 37: 3556–3563.

[12] Stalker DM, Kolter R, Helinski DR (1979) Nucleotide sequence of the region of an origin of replication of the antibiotic resistance plasmid R6K. Proc Natl Acad Sci U S A 76: 1150–1154.

[13] Uchiyama H, Weisblum B (1985) N-methyl transferase of streptomyces erythraeus that confers resistance to the macrolidelincosamide-streptogramin B antibiotics: amino acid sequence and its homology to cognate r-factor enzymes from pathogenic bacilli and cocci. Gene 38: 103–110.

[14] Horie N, Aiba H, Oguro K, Hojo H, Takeishi K (1995) Functional analysis and DNA polymorphism of the tandemly repeated sequences in the 5'-terminal regulatory region of the human gene for thymidylate synthase. Cell Struct Funct 20: 191–197.

[15] Pelham HR (1982) A regulatory upstream promoter element in the drosophila hsp 70 heat-shock gene. Cell 30: 517–528.

[16] Nagarajan N, Cook C, Di Bonaventura M, Ge H, Richards A, Bishop-Lilly KA, DeSalle R, Read TD, Pop M (2010) Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. BMC Genomics 11: 242.

[17] Oikonomopoulos S, Wang YC, Djambazian H, Badescu D, Ragoussis J (2016) Benchmarking of the oxford nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. Sci Rep 6: 31602.

[18] Jain M, Olsen HE, Paten B, Akeson M (2016) Erratum to: The oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17: 256.

[19] English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA (2012) Mind the gap: upgrading genomes with pacific biosciences RS long-read sequencing technology. PLoS One 7: e47768.

[20] Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10: 563–569.

[21] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27: 722–736.

[22] Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13: 341.

[23] Maitra RD, Kim J, Dunbar WB (2012) Recent advances in nanopore sequencing. Electrophoresis 33: 3418–3428.

[24] Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36: 338–345.

[25] Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J (2015) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol 33: 296–300.

[26] Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, Lemainque A, Wincker P, Aury JM (2015) Genome assembly using nanopore-guided long and error-free DNA reads. BMC Genomics 16: 327.

[27] Wick R. Basecalling-comparison.

[28] Mirkin SM (2007) Expandable DNA repeats and human disease. Nature 447: 932–940.

[29] Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. Genome Res 18: 324–330.

[30] Sandegren L, Andersson DI (2009) Bacterial gene amplification: implications for the evolution of antibiotic resistance. Nat Rev Microbiol 7: 578–588.

[31] Kwong JC, McCallum N, Sintchenko V, Howden BP (2015) Whole genome sequencing in clinical and public health microbiology. Pathology 47: 199–210.

[32] Karlsson E, Lärkeryd A, Sjödin A, Forsman M, Stenberg P (2015) Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci Rep 5: 11996.

[33] Andrews S, Others (2010). FastQC: a quality control tool for high throughput sequence data.

[34] Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863–864.

[35] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9: 357–359.

[36] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963.

[37] Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30: 2068–2069.

[38] Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072–1075.

[39] Diep BA, Gill SR, Chang RF, Phan TH, Chen JH, Davidson MG, Lin F, Lin J, Carleton HA, Mongodin EF, Sensabaugh GF, Perdreau-Remington F (2006) Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant staphylococcus aureus. Lancet 367: 731–739.

[40] Tenover FC, Goering RV (2009) Methicillin-resistant staphylococcus aureus strain USA300: origin and epidemiology. J Antimicrob Chemother 64: 441–446.

[41] Enright MC (2006) Genome of an epidemic community-acquired MRSA. Lancet 367: 705–706.

[42] Fey PD, Endres JL, Yajjala VK, Widhelm TJ, Boissy RJ, Bose JL, Bayles KW (2013) A genetic resource for rapid and comprehensive phenotype screening of nonessential staphylococcus aureus genes. MBio 4: e00537–12.

[43] Shahkarami F, Rashki A, Rashki Ghalehnoo Z (2014) Microbial susceptibility and plasmid profiles of Methicillin-Resistant staphylococcus aureus and Methicillin-Susceptible s. aureus. Jundishapur J Microbiol 7: e16984.

[44] Kennedy AD, Porcella SF, Martens C, Whitney AR, Braughton KR, Chen L, Craig CT, Tenover FC, Kreiswirth BN, Musser JM, DeLeo FR (2010) Complete nucleotide sequence analysis of plasmids in strains of staphylococcus aureus clone USA300 reveals a high level of identity among isolates with closely related core genome sequences. J Clin Microbiol 48: 4504–4511.

[45] McCarthy H, Waters EM, Bose JL, Foster S, Bayles KW, O'Neill E, Fey PD, O'Gara JP (2016) The major autolysin is redundant for staphylococcus aureus USA300 LAC JE2 virulence in a murine device-related infection model. FEMS Microbiol Lett 363.

[46] Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ (2016) Within-host evolution of bacterial pathogens. Nat Rev Microbiol 14: 150–162.

[47] Biek R, Pybus OG, Lloyd-Smith JO, Didelot X (2015) Measurably evolving pathogens in the genomic era. Trends Ecol Evol 30: 306–313.

[48] Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148: 1667–1686.

[49] Deatherage DE, Barrick JE (2014) Identification of mutations in Laboratory-Evolved microbes from Next-Generation sequencing data using breseq. In: Sun L, Shou W, editors, Engineering and Analyzing Multicellular Systems: Methods and Protocols, New York, NY: Springer New York. pp. 165–188.

[50] Kuroda M, Tanaka Y, Aoki R, Shu D, Tsumoto K, Ohta T (2008) Staphylococcus aureus giant protein ebh is involved in tolerance to transient hyperosmotic pressure. Biochem Biophys Res Commun 374: 237–241.

[51] Hu J, Zhang X, Liu X, Chen C, Sun B (2015) Mechanism of reduced vancomycin susceptibility conferred by walk mutation in community-acquired methicillin-resistant staphylococcus aureus strain MW2. Antimicrob Agents Chemother 59: 1352–1355.

[52] Yarwood JM, McCormick JK, Schlievert PM (2001) Identification of a novel Two-Component regulatory system that acts in global regulation of virulence factors ofstaphylococcus aureus. J Bacteriol 183: 1113–1123.

[53] Throup JP, Koretke KK, Bryant AP, Ingraham KA, Chalker AF, Ge Y, Marra A, Wallis NG, Brown JR, Holmes DJ, Others (2000) A genomic analysis of two-component signal transduction in streptococcus pneumoniae. Mol Microbiol 35: 566–576.

[54] Stephenson K, Hoch JA (2002) Virulence- and antibiotic resistance-associated two-component signal transduction systems of gram-positive pathogenic bacteria as targets for antimicrobial therapy. Pharmacol Ther 93: 293–305.

[55] Sibold C, Henrichsen J, König A, Martin C, Chalkley L, Hakenbeck R (1994) Mosaic pbpx genes of major clones of penicillin-resistant streptococcus pneumoniae have evolved from pbpx genes of a penicillin-sensitive streptococcus oralis. Mol Microbiol 12: 1013–1023.

[56] Guenzi E, Gasc AM, Sicard MA, Hakenbeck R (1994) A two-component signal-transducing system is involved in competence and penicillin susceptibility in laboratory mutants of streptococcus pneumoniae. Mol Microbiol 12: 505–515.

[57] Calvez S, Prevost H, Drider D (2008) Relative expression of genes involved in the resistance/sensitivity of enterococcus faecalis JH2-2 to recombinant divercin RV41. Biotechnol Lett 30: 1795–1800.

[58] Shankar V, Baghdayan AS, Huycke MM, Lindahl G, Gilmore MS (1999) Infection-Derived enterococcus faecalisstrains are enriched in esp, a gene encoding a novel surface protein. Infect Immun 67: 193–200.

[59] Leavis H, Top J, Shankar N, Borgen K, Bonten M, van Embden J, Willems RJL (2004) A novel putative enterococcal pathogenicity island linked to the esp virulence gene of enterococcus faecium and associated with epidemicity. J Bacteriol 186: 672–682.

[60] Michel JL, Madoff LC, Olson K, Kling DE, Kasper DL, Ausubel FM (1992) Large, identical, tandem repeating units in the C protein alpha antigen gene, bca, of group B streptococci. Proc Natl Acad Sci U S A 89: 10060–10064.

[61] Li J, Kasper DL, Ausubel FM, Rosner B, Michel JL (1997) Inactivation of the alpha C protein antigen gene, bca, by a novel shuttle/suicide vector results in attenuation of virulence and immunity in group B streptococcus. Proc Natl Acad Sci U S A 94: 13251–13256.

[62] Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW (2012) Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 13: 601–612.

[63] Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, Holden MTG, Dougan G, Bentley SD, Parkhill J, Peacock SJ (2012) Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 8: e1002824.

[64] Wilson DJ (2012) Insights from genomics into bacterial pathogen populations. PLoS Pathog 8: e1002874.

[65] Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, Falush D, Suerbaum S (2011) Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci U S A 108: 5033–5038.

[66] Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG, Iqbal Z, Rimmer AJ, Cule M, Ip CLC, Didelot X, Harding RM, Donnelly P, Peto TE, Crook DW, Bowden R, Wilson DJ (2012) Evolutionary dynamics of staphylococcus aureus during progression from carriage to disease. Proc Natl Acad Sci U S A 109: 4550–4555.

[67] Snitkin ES, Zelazny AM, Thomas PJ, Stock F, NISC Comparative Sequencing Program Group, Henderson DK, Palmore TN, Segre JA (2012) Tracking a hospital outbreak of carbapenem-resistant klebsiella pneumoniae with whole-genome sequencing. Sci Transl Med 4: 148ra116.

[68] McAdam PR, Templeton KE, Edwards GF, Holden MTG, Feil EJ, Aanensen DM, Bargawi HJA, Spratt BG, Bentley SD, Parkhill J, Enright MC, Holmes A, Girvan EK, Godfrey PA, Feldgarden M, Kearns AM, Rambaut A, Robinson DA, Fitzgerald JR (2012) Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant staphylococcus aureus. Proc Natl Acad Sci U S A 109: 9107–9112.

[69] Unemo M, Golparian D, Sánchez-Busó L, Grad Y, Jacobsson S, Ohnishi M, Lahra MM, Limnios A, Sikora AE, Wi T, Harris SR (2016) The novel 2016 WHO neisseria gonorrhoeae reference strains for global quality assurance of laboratory investigations: phenotypic, genetic and reference genome characterization. J Antimicrob Chemother 71: 3096–3108.

[70] McGinnis J, Laplante J, Shudt M, George KS (2016) Next generation sequencing for whole genome analysis and surveillance of influenza a viruses. J Clin Virol 79: 44–50.

[71] Hampson A, Barr I, Cox N, Donis RO, Siddhivinayak H, Jernigan D, Katz J, McCauley J, Motta F, Odagiri T, Tam JS, Waddell A, Webby R, Ziegler T, Zhang W (2017) Improving the selection and development of influenza vaccine viruses - report of a WHO informal consultation on improving influenza vaccine virus selection, hong kong SAR, china, 18-20 november 2015. Vaccine 35: 1104–1109.

[72] Johnson SS, Zaikova E, Goerlitz DS, Bai Y, Tighe SW (2017) Real-Time DNA sequencing in the antarctic dry valleys using the oxford nanopore sequencer. J Biomol Tech 28: 2–7.

[73] Votintseva AA, Bradley P, Pankhurst L, Del Ojo Elias C, Loose M, Nilgiriwala K, Chatterjee A, Smith EG, Sanderson N, Walker TM, Morgan MR, Wyllie DH, Walker AS, Peto TEA, Crook DW, Iqbal Z (2017) Same-Day diagnostic and surveillance data for tuberculosis via Whole-Genome sequencing of direct respiratory samples. J Clin Microbiol 55: 1285–1298.

[74] Stewardson AJ, Allignol A, Beyersmann J, Graves N, Schumacher M, Meyer R, Tacconelli E, De Angelis G, Farina C, Pezzoli F, Bertrand X, Gbaguidi-Haore H, Edgeworth J, Tosas O, Martinez JA, Ayala-Blanco MP, Pan A, Zoncada A, Marwick CA, Nathwani D, Seifert H, Hos N, Hagel S, Pletz M, Harbarth S, TIMBER Study Group (2016) The health and economic burden of bloodstream infections caused by antimicrobial-susceptible and non-susceptible enterobacteriaceae and staphylococcus aureus in european hospitals, 2010 and 2011: a multicentre retrospective cohort study. Euro Surveill 21.

[75] De Kraker MEA, Davey PG, Grundmann H, Group BS, Others (2011) Mortality and hospital stay associated with resistant staphylococcus aureus and escherichia coli bacteremia: estimating the burden of antibiotic resistance in europe. PLoS Med 8: e1001104.

[76] Rehm HL (2013) Disease-targeted sequencing: a cornerstone in the clinic. Nat Rev Genet 14: 295–300.

[77] de Kraker MEA, Stewardson AJ, Harbarth S (2016) Will 10 million people die a year due to antimicrobial resistance by 2050? PLoS Med 13: e1002184.