Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature: _____ _____

Robert A. Petit III Date

Macro-scale genomic studies of bacterial pathogens

By

Robert A. Petit III

Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Program in Population Biology, Ecology and Evolution

_____
Timothy D. Read
Advisor

_____
Karen N. Conneely
Committee Member

_____
Joanna B. Goldberg
Committee Member

_____
Levi T. Morran
Committee Member

_____
Hao Wu
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____
Date

Macro-scale genomic studies of bacterial pathogens.

By

Robert A. Petit III

M.S., Georgia Institute of Technology, 2011

B.S., Valdosta State University, 2010

Advisor: Timothy D. Read, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Graduate School of Biological and Biomedical Sciences

Population Biology, Ecology, and Evolution

2018

Abstract

Macro-scale genomic studies of bacterial pathogens.

By Robert A. Petit III

The low cost of genome sequencing has led to a significant increase in publicly available datasets of bacterial pathogens. Taking advantage of this data requires new strategies for using computational resources and bioinformatics, as well as applying traditional organism-specific knowledge. With this understanding, I used public datasets to investigate two important bacterial pathogens *Bacillus anthracis* and *Staphylococcus aureus*.

In my first research project, I focused on *Bacillus anthracis*, the etiologic agent of anthrax, which shares over 99% average nucleotide identity with *Bacillus cereus* Group (BCerG) bacteria. This closeness, coupled with sequencing error rates, can cause *B. cereus* to be falsely identified as *B. anthracis*. To address this issue, I developed a typing schema for fine-scale differentiation of these two species. I identified a set of 31-mers specific to *B. anthracis* and another set specific to all BCerG including *B. anthracis*. I determined the limits of detection of these k-mers on synthetic data and developed a model to predict the presence of true *B. anthracis* sequences. I then reanalyzed a New York subway metagenome dataset, which falsely identified evidence for *B. anthracis*. I found no evidence for anthrax but instead the presence of unsampled close relatives to *B. anthracis*.

My second project concerned *Staphylococcus aureus*, a major antibiotic-resistant pathogen responsible for a wide spectrum of hospital and community-associated infections. *S. aureus* was well represented in genome sequencing studies submitted to public repositories but there were no tools available to make use of this useful data. To fill this void, I developed Staphopia, an analysis pipeline, database and application programming interface focused on *S. aureus* and processed over 44,000 publicly available *S. aureus* genomes. I found patterns in antibiotic resistance between *S. aureus* sequence types and a bias towards sequencing clinically relevant methicillin-resistant *S. aureus* strains.

I conclude, with a discussion about future macro-scale comparative genomic studies consisting of tens of thousands of genomes. I also provide comments on the expected rewards and challenges associated with macro-scale studies. Overall, this body of work illustrates the importance of public datasets for bacterial pathogens and integrating organism specific knowledge into bacterial sequence analyses.

Macro-scale genomic studies of bacterial pathogens.

By

Robert A. Petit III

M.S., Georgia Institute of Technology, 2011

B.S., Valdosta State University, 2010

Advisor: Timothy D. Read, Ph.D.

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Graduate School of Biological and Biomedical Sciences

Population Biology, Ecology, and Evolution

2018

Acknowledgements

I have been fortunate to have met a lot of great people during my time at Emory. Some of you have been mentors, collaborators, therapists, or reasons to cut loose and have some fun!

To my advisor Tim Read, thank you very much for everything these past years. As a naive graduate student, I did not quite comprehend the investments an advisor must make when taking in a new lab member. Having gone through the process, I have a better understanding and thank you very much for the invitation to join your lab! Thank you very much for your guidance and patience over the years. Thank you very much for always looking out for myself and the rest of the lab! Thank you for letting me build stuff and molding me into a scientist.

To my mentor, Tauqeer Alam, thank you for the being such an awesome mentor. You truly exemplify what is means to be a mentor. During your postdoc in Tim's lab, you took the time to teach how to practice science. You were available for daily coffee hours to discuss work, science and life. You have always offered advice, even if it wasn't what I wanted to hear but needed to hear. Even as move into the next phase of my life, you still go out of your way to help me out. Thank you, Brother!

To current and former Read Lab members, when's the next happy hour? Honestly though thank you very much. Having now been a part of the lab for almost 8 years I've gotten to work with many great people. Thank you, Sandeep Joseph, for the fun back

and forth banter. Jessica Peterson, thank you for all your help! To the current lab members, Michelle Su, Jon Moller and Michelle Hargita, I can only hope that I can be as helpful to you as previous members have been to myself.

To the PBEE program, it has been a fun ride! Many thanks to my cohort, Erica, JR, Travis and Nate for the fun conversations about science, family and nonsense! My committee members Joanna, Karen, Levi and Hao, thank you very much for helping keeping tabs and making sure the ship was still going in the right direction! Bruce Levin, thanks for letting me disguise as TA for your microbiology class, when really, I was a student! Also, for introducing me to Véronique Perrot whose enthusiasm is a great reminder that this sequence analysis stuff is kind of fun! Finally, Nicole Gerardo, thank you very much for everything you have done for my family and for PBEE!

To my friends and family, thank you for still being there even though I really haven't been these last few years! To my in-laws, Steven and Tammy Thomas, many thanks for allowing us to live with you, although we outnumber you! To my brother, Bryan, thank you for the many conversations and the occasional one over drinks. To my sisters, Elizabeth and Alicia, thank you for doing so much and never expecting anything in return.

To my parents, Robert and Christi, thank you for the many sacrifices you made for Elizabeth, Alicia and myself. Times were tough. But, I wouldn't change a thing. You always put us first and never yourselves. While you will say I don't owe you anything,

“it’s what parents are supposed to do”. I can at least hope that you will let me pay it forward by doing the same for Robbie and Bryan. Thank you very much.

To my sons, Robbie and Bryan, you two always put me on an emotional rollercoaster. Most often it is the most exciting, happy and joyous feeling. Then sometimes its complete frustration, but I quickly feel the need to call Grandma D (my mom) to apologize doing the same to her when I was growing up! Thank you, boys, for brightening each my days. It is funny though, at this age you don’t quite comprehend what has been going on these last few years. Let’s just say, Daddy can play now!

To my wife, Shannon, I would like to extend my sincerest gratitude. Throughout this process you have had to make the most sacrifices. We have had two boys through this process and you still clear the path to keep going. This has been quite the emotional experience, but you were always there to listen. Your laugh has been the best escape from studies. Thank you for being so selfless to myself and the boys! Without you much of this would not have been possible.

Finally, yes there is a dissertation that follows. I wanted to thank the many people that have helped make my journey possible. Thank you!

Table of Contents

List of Boxes

List of Figures

List of Tables

Chapter 1: Introduction

Whole genome sequencing of bacterial pathogens has become a valuable investigative tool. This dissertation describes investigations of two species, *Bacillus anthracis* and *Staphylococcus aureus*. The goal of this introduction is to cover basic concepts of bacterial sequence analysis. I provide details on bacterial genome sequencing including sample collection and sequencing technologies. This is followed up with an overview of analysis techniques to extract the wealth of information available from genomic sequences. I also highlight how bacterial genome sequencing has changed in recent years. I conclude with a brief overview of the remaining chapters of this dissertation. Terminology associated with bacterial sequence analysis used throughout this dissertation has been summarized in **Box 1.1**.

Bacterial sequence analysis step by step

The first step in bacterial genome sequencing is to collect a DNA sample, which is assumed a single strain in pure culture (referred to here as a "genomic sample"). The alternative approach is to extract DNA directly from a complex environment such as a clinical specimen, without culture (a "metagenomic sample"). After extraction, a DNA library must be created for sequencing. During the library preparation, the DNA is fragmented into a desired length and PCR adapters are attached to the ends of the fragments (Head et al., 2014). A DNA sequencing instrument then takes the DNA library and determines the per-base identity. This combined process of creating a DNA library and sequencing is often referred to as a "*sequencing run*" (**Figure 1.1**). Once the

sequencing run is completed, fragments of sequenced DNA, or reads, are output in a standard format called FASTQ. The total amount of DNA sequenced, and lengths of reads are dependent on the sequencing instrument (**Table 1.1**). Illumina technology can produce billions of short reads (100-300bp), while PacBio and Nanopore produces 10s to 100s of thousands of long (>5kb) reads.

There are many different approaches to investigate the output of a sequencing run. These approaches fall into three general paradigms: *de novo* assembly, reference mapping and sequence decomposition (**Box 1.2**). I will provide a step by step description of a standard bacterial sequence analysis workflow building off these paradigms (**Figure 1.2**). I have also listed many bioinformatic tools for such a workflow in **Table 1.2**.

Sequence Quality Control

For each base sequenced, there is a probability that a sequenced base is an error (**Figure 1.2**). Depending on the technology, the per base error rate may be as a low as 0.1% or as high as 15% (**Table 1.1**). It is important quality control (QC) the sequenced DNA to achieve an acceptable level of sequencing quality. The base-calling error probability is often reported as the Phred quality score (Ewing & Green, 1998; Ewing et al., 1998). The value of the Phred quality score (Q) is determined by the equation: $Q = -10 \log_{10} P$, where P is the base-call error probability reported by the sequencer. This produces Q-scores that are logarithmically linked to error probabilities (**Table 1.3**). With this understanding, reads can be filtered or trimmed based on the Q-score (Bolger,

Lohse & Usadel, 2014; Bushnell, 2016). The next step is to correct any remaining erroneous reads. Using the context of the sequencing, probabilistic models can be applied to determining the likelihood a base is an error and correctable (Kelley, Schatz & Salzberg, 2010; Bankevich et al., 2012; Song, Florea & Langmead, 2014; Heo et al., 2016). The final QC step is to remove potential DNA "contaminating" sequences that have been introduced as part of the process of sequencing. Primers and controls used for the sequencing should be identified and removed from the reads (Bolger, Lohse & Usadel, 2014; Ondov et al., 2015; Bushnell, 2016). It is also important to filter out any potential biological contaminants such as human DNA (Haque et al., 2015; Bushnell, 2016; Lu & Salzberg, 2018). This can be done through read mapping or sequence decomposition. Each of these Sequence QC steps are important for improving downstream analysis.

Genome assembly

Sequencing technologies are not at a point in which a whole bacterial chromosome can be sequenced. During library prep, the bacterial chromosome must be split into many pieces. Once sequenced, these fragments must be reassembled together through the *de novo* assembly process. Assembly algorithms identify overlaps between millions of reads and merge these overlaps into longer contiguous sequences known as "contigs" or "assemblies" (Bankevich et al., 2012; Koren et al., 2017). Assemblies can be further improved by correcting local assembly errors (Walker et al., 2014). Metagenomic sequences must use an assembler that account for multiple organisms (Peng et al., 2012; Li et al., 2015; Nurk et al., 2017). Once completed, a "*draft*" assembly is produced that

may consist of 10s to 100s of contigs. In the past to “complete” a genome into a single contiguous sequence, required further sequencing using expensive Sanger sequencing. As a consequence, many assembled genomes have remained in draft state. The introduction of long read technology has made it possible to produce single contig hybrid assemblies (Wick et al., 2017a). Hybrid assemblies use short and long read sequencings to create high quality assemblies of a few contigs and potentially a single contig. This has made it more common to create single contig assemblies (Loman, Quick & Simpson, 2015; Rhoads & Au, 2015; Bayliss et al., 2017).

There are many statistics to determine the quality of the draft assembly (Bradnam et al., 2013). A few common statistics are total assembled size in base pairs, GC content, and the N50 statistic. The total assembled size is the sum of the contig lengths. The GC content is determined by dividing the total number of Gs and Cs in the assembled contigs by the total assembled size. The total assembled size and GC content present the biological quality of the assembly because it should be similar to the genome size and GC content of the sequenced bacteria. The N50 statistic is a technical representation of the assembly quality. The N50 is determined by sorting the contigs from longest to shortest contig, then adding contigs together until the total length is more than 50% of the expected genome size, the length of the final contig to break 50% is determined as the N50. The larger the N50, the better the quality of an assembly is assumed to be. Another approach is to visualize assemblies to identify problematic regions (E.g repeat regions), potential mis assemblies or compare multiple assemblies (Wick et al., 2015).

Genome annotation

An assembled genome alone is a blank canvas. Additional steps must be taken to identify genetic features such as antibiotic resistance and virulence factors. The process of scanning the assembly and labeling each relevant feature is called "*genome annotation*". A common form of genome annotation is to predict the coordinates and function of genes across the bacterial genome. Gene prediction relies on *ab initio* methods which account for characteristics of bacterial genomes (Delcher et al., 1999; Besemer, Lomsadze & Borodovsky, 2001; Hyatt et al., 2010). These *ab initio* methods identify regions between start and stop codons called open reading frames (ORFs). Then using probabilistic models that can account for promoter regions and codon usage bias, ORFs can accurately be predicted as genes. After the genes are predicted, functions are assigned to translated proteins through sequence homology or conserved protein domains (Quevillon et al., 2005; Eddy, 2009; Camacho et al., 2009). Genome annotation is not limited to gene prediction, CRISPRs and RNA features including ribosomal, transfer and non-coding RNAs can also be predicted (Laslett & Canback, 2004; Lagesen et al., 2007; Bland et al., 2007; Kolbe & Eddy, 2011).

Genotyping bacteria based on genome sequence

After sequencing a bacterial isolate, it is important to determine where it fits in the context of its species. This can be done by comparing its genetic relatedness to other members of the species to assign a subtype. Pulse-field gel electrophoresis (PFGE) has long been the standard molecular approach for subtyping bacterial strains (Tenover, Arbeit & Goering, 1997). PFGE uses DNA restriction patterns to '*DNA fingerprint*'

bacterial isolates. This is time consuming process that can give varying results depending on the protocol.

With the introduction of sequencing it became possible to develop alternative methods to subtype bacteria. One such method is multi-locus sequence typing (MLST) (Maiden et al., 1998). MLST selects 5-7 highly conserved genes within a bacterial species, based on the combination of alleles in these genes a sequence type (ST) is assigned. MLST has also been expanded to the core-genome (cgMLST, > 1500 loci) (Leopold et al., 2014) and the whole-genome (wgMLST, > 20k loci) (Sheppard, Jolley & Maiden, 2012). Each of these approaches offer a different phylogenetic resolution (Alikhan et al., 2018). Although MLST is limited to a few genes it provides a good overview of a species level resolution. In cases in which samples are all from the same ST, for example an outbreak, cgMLST or wgMLST are required to investigate diversity within STs.

Identifying Variation

Illuminating genetic variation between bacterial strains is an important step in bacterial sequence analysis. The process of identifying variants can be broken up into two broad steps. The first step is map the bacterial sequences to a reference genome (Li & Durbin, 2009b; Langmead & Salzberg, 2012). After mapping, single nucleotide polymorphisms (SNPs) and insertions and deletions (InDels) are determined (DePristo et al., 2011; Koboldt et al., 2012). There are a number of intermediate quality control steps that can be implemented to improve the accuracy of the variant calls (Van der Auwera et al., 2013). Many of the approaches were originally developed for human genetics and had to

be adapted to be used for bacterial sequences. Recently, alternative approaches specifically designed for haploid organisms have been introduced (Treangen et al., 2014; Gardner, Slezak & Hall, 2015).

Antimicrobial Resistance and Virulence Factors

Bacterial pathogens present a significant threat to public health (van Oosten et al., 2015; O'Neill, 2016). The success of a bacterial pathogen will depend on its armory of virulence factors and its ability to resist treatment through mutations and genes that confer antimicrobial resistance (AMR). Associations with virulence factors and AMR can be identified by the presence of genes or SNPs with known associations (Gupta et al., 2014; Inouye et al., 2014; Hunt et al., 2017). Often, AMR genes are carried on mobile genetic elements (MGE) (Stokes & Gillings, 2011), such as plasmids and transposable elements. Identification of these MGEs as a whole requires alternative approaches (Siguier et al., 2006; Carattoli et al., 2014; Antipov et al., 2016b; Rozov et al., 2017) A number of high quality databases have been specifically targeted towards antibiotic resistance and virulence factors (Zankari et al., 2012; Gupta et al., 2014; Joensen et al., 2014; Carattoli et al., 2014; Chen et al., 2016a; Jia et al., 2017; Lakin et al., 2017a).

Comparative genomic analyses

The following section discusses methods used for comparative genomics. Comparative genomics makes use of genomic features to better understand evolutionary relationships between organisms. Comparative genomics can be applied at the species

level or as high as the kingdom level. For the purpose of this dissertation, I have focused on species level analyses.

Phylogenetics

Phylogenetic analysis provides a view into the evolutionary history of an organism. It can also be used in outbreaks to identify recent evolutionary changes and transmission events (Alam et al., 2015a; Klinkenberg et al., 2017). Phylogenetic trees are generally constructed from multiple sequence alignments (MSA) of genes or SNPs from the core-genome. Creating the one true phylogeny for a large set of samples is a difficult problem due to the number of possible trees (Felsenstein, 1978). As an example, for 20 samples there are more than 10^{23} possible trees. This has required the use of heuristic methods to identify the most likely tree.

Advancements in computation during the late 90s and early 2000s have made maximum likelihood estimation (MLE) the standard for tree construction (Price, Dehal & Arkin, 2009; Stamatakis, 2014; Nguyen et al., 2015b). MLE uses DNA (or amino acid) substitution models to determine the probability that a phylogenetic tree is possible. Millions of potential phylogenetic trees are tested, and the most probable tree is selected. Support for this tree is determined through a process called "*bootstrapping*" (Felsenstein, 1985). Bootstrapping randomly sub selects portions of the input alignment to determine how often the MLE tree is recapitulated. Bootstrapping is important for assigning a level of confidence for each branch in a phylogeny.

An alternate approach to generating phylogenetic trees is to use Bayesian approaches. Bayesian algorithms require more information about samples but are useful for estimating divergence times and ancestral states (Pritchard, Stephens & Donnelly, 2000; Ronquist et al., 2012; Bouckaert et al., 2014). For bacterial sequences it is important to identify and mask recombination events to improve the accuracy of a tree (Didelot & Wilson, 2015). Recent reviews provide an assessment of commonly used tree construction programs (Nascimento, Reis & Yang, 2017; Lees et al., 2018).

Pan-genome

The complete set of genes within a bacterial species is termed the bacterial pan-genome (Medini et al., 2005). The pan-genome can be partitioned into a core genome (genes shared across all strains) and an accessory genome (genes found in at least one strain but not all). The core genome is predicted to be stable and the phylogeny of individual's core genes generally recapitulates the evolutionary history of the species. The accessory genome on the other hand, is much more variable in its origin and includes genes acquired through horizontal gene transfer (HGT) that promote adaptation to a local habitat. Determining the pan-genome for a species is a computationally difficult problem that worsens with sample size (Nguyen et al., 2015a). Due to this, in order to scale to 1,000s of genomes, viable pan-genome analysis approaches have used heuristic approaches (Zhao et al., 2012; Fouts et al., 2012; Sahl et al., 2014; Page et al., 2015b).

Genome wide association studies

Determining the genetic basis of a phenotype, such as antibiotic resistance or virulence, is important for bacteria. This can be done by performing laboratory manipulations or

by using sequenced genomes. By using genome sequences, many genetic markers can be tested across many genomes for associations to a phenotype. This type of study is called a genome-wide association study (GWAS). A standard GWAS, will often use single-nucleotide polymorphisms (SNPs) to test for associations with a phenotype. A simple form of GWAS is to independently test associations for each SNP through the use of regression (Purcell et al., 2007). The form of regression (logistic or linear) is dependent on how the phenotype is reported (categorical or continuous). After correcting for multiple tests using methods such as Bonferroni correction, significant associations can be determined. These significant associations can then be (and should be) validated in the lab. Recently algorithms specifically designed for bacterial samples that account for the strong population structure exhibited by bacteria have been developed (Feil & Spratt, 2001; Lees et al., 2016; Earle et al., 2016; Collins & Didelot, 2018).

A deluge of bacterial sequences

In this section I provide a brief history on DNA sequencing and highlight how bacterial genome sequencing has changed in recent years. I also discuss how this change has led to a deluge of data and an opportunity to conduct comparative genomic studies previously not possible.

A brief history of DNA sequencing technologies

Low throughput, or “first-generation”, DNA techniques were first developed in the 1970s (Wu, 1972; Jay et al., 1974; Sanger, Nicklen & Coulson, 1977). Over the next two

decades, the improvements in the technology allowed for the completed genome of a small bacteriophage φX174 (5kbp) in 1977, the slightly larger Epstein-Barr virus (180kbp) in 1984 and the first free-living organism, *Haemophilus influenzae* (1.8Mbp) in 1995 (Sanger et al., 1977; Baer et al., 1984; Fleischmann et al., 1995). "First-generation" technologies, such as Sanger, are extremely accurate but limited to a single reaction per capillary or tube (Sanger, Nicklen & Coulson, 1977). In the late 1990s, "second-generation", high-throughput DNA sequencing techniques were first introduced (Ronaghi et al., 1996; Brenner et al., 2000). These techniques were capable of producing thousands of reactions per tube. This allowed for millions of base pairs to be produced sequencing run. It would take until 2005 for technologies implementing these techniques to become commercially available. The "third generation" technologies introduced the sequencing of single DNA molecules (Eisenstein, 2012; Rhoads & Au, 2015) (**Table 1.1**). The advantages of these technologies are long reads (1- 1000 kb) but they have lower yield per dollar than Illumina and higher per base error rates. A recent review (Loman & Pallen, 2015) highlights each of the technologies by discussing milestones achieved over the course of twenty years following the release of the *Haemophilus influenzae* completed genome in 1995.

Affordable high-throughput sequencing

After the completion of the draft human genome in 2001, the National Human Genome Research Institute began recording the costs of sequencing ("DNA Sequencing Costs: Data"). At the turn of the century (2000) the price per megabase (1,000,000 bp) was $10,000. For a bacterium, such as *Staphylococcus aureus* with a genome size of

2.8Mbp, it would have cost almost $30,000 for 1x coverage of sequencing. After the introduction of massively-parallel sequencing technologies, the price of sequencing quickly fell from $1,000/Mbp in 2005 to $1/Mbp in 2009. In 2011, the cost per megabase dropped below $0.10, making the same *S. aureus* genome now $0.30 for 1x coverage. By 2017 the price per megabase was approximately $0.05. The rapid decline in sequencing costs has been driven by second-generation technologies. Third-generation technologies deliver long reads but due to a low output have a higher per base cost. Another noticeable shift has been the time required to sequence a genome going from weeks to days, and now in real-time (**Table 1.1**). However, the overall decrease in costs has created a flood of bacterial sequences in public sequence databases.

New opportunities in existing data

As of spring 2018, more than 400Tb of bacterial sequencing had been generated and made publicly available in the NCBI SRA database. In this data, there was 30,334 completed genomes, 138,427 assembled genomes and 733,815 sequenced bacterial samples. There were 19 human bacterial pathogens with more than 5,000 sequenced samples (**Table 1.4**). The rapid growth of bacterial sequencing has placed constant pressure on the field of bacterial genomics to adapt. This has created opportunities to produce scalable algorithms for analysis of 100s or 1000s of genomes (Treangen et al., 2014; Page et al., 2015b). It has also forced the field to seek solutions from other fields such as human genomics (DePristo et al., 2011) or search engine optimization (Ondov et al., 2015). An important opportunity is the ability to reuse existing data to conduct a

secondary analysis testing a different hypothesis. For example, the human microbiome project (Human Microbiome Project Consortium, 2012) alone has led to over 500 secondary analyses ("NIH Human Microbiome Project - Publications").

Outline for this dissertation

Sequencing has become a powerful tool for investigating bacterial pathogens. As a consequence, numerous datasets have been made publicly available. This has created an opportunity to not only reuse datasets in ways they were not originally created for but also combine multiple datasets to further investigate a bacterial pathogen. In this dissertation I have taken advantage of two such opportunities.

Chapter 2 demonstrates the use of sequence decomposition as a diagnostic tool for the fine-scale differentiation of *Bacillus anthracis*, the causative agent of anthrax, and *Bacillus cereus*, a common soil bacterium, in metagenomic sequences. The relatedness between *B. anthracis* and *B. cereus* at the chromosome level often makes the two indistinguishable to "generalist" approaches. This presents a significant biodefense problem in which *B. cereus* is commonly mistaken for *B. anthracis* in metagenomics sequencing. In this chapter, I describe a "specialist" approach that accounts for biological and technical nuances to more accurately distinguish *B.* anthracis from *B. cereus* within metagenomic sequences.

Chapter 3 examines *Staphylococcus aureus* from the perspective of 40,000+ genomes. *S. aureus* is an opportunistic pathogen responsible for hospital and community

associated infections in humans. The work from this chapter introduces Staphopia as a community resource focused on *S. aureus* genomics. I provide a global overview of current *S. aureus* sequencing efforts. Patterns in *S. aureus* evolution, such as methicillin-resistant *S. aureus* (MRSA), are explored in the global *S. aureus* population. I also present a novel method to rationally select publicly available *S. aureus* genomes for comparative genomic studies.

Chapter 4 provides a literature review describing barriers to horizontal gene transfer events in bacterial species. In order for an HGT event to become fixed in a new species it must overcome ecological processes that separate donor and recipient, genetic defense mechanisms that limit HGT, and evolutionary processes that eliminate novel DNA from genomes. I use *S. aureus* to demonstrate how barriers have influenced its recent evolutionary history.

Chapter 5, the final chapter, summarizes the findings and limitations of this dissertation. I also provide a discussion on the future direction of bacterial sequence analysis. I introduce macro-scale bacterial genomics, in which tens of thousands of genomes are leveraged for comparative genomics. I highlight the rewards associated with macro-scale studies. I also dive into the challenges, some of which I have already faced, of dealing with macro-scale studies. I conclude the discussion with examples of emerging macro-scale studies and a few final remarks.

Appendix

The following appendix contains boxes, tables and figures referenced in the text of this chapter.

Box 1.1: Bacterial sequence analysis terminology

Contig - A continuous sequence generated from the assembly of smaller DNA fragments

Coverage - The average number of times any given nucleotide in a genome has been sequenced, calculated by dividing the total sequence output by the sequenced organism's genome size. "10x" coverage means on average every base in a genome has been sequenced 10 times.

European Nucleotide Archive (ENA) - A mirror of SRA maintained by the European Bioinformatics Institute (EMBL-EBI)

FASTQ - The standard format for reporting sequenced reads and the corresponding per-base quality score (probability of error)

Metagenomic Sample - Genetic material sequenced directly from an environmental sample. Culturing bacteria is not required for metagenomic sequencing.

Paired-End Read - Two sequence reads sequenced from the opposite ends of the same fragment of genetic material. The genetic distance between the two reads is determined during library prep.

Sample - Bacteria to be sequenced either from culture or the environment

Scaffold - With the support of pair-end reads, contigs are overlapped with gaps of known distances

Sequence Read - A fragment of sequenced genetic material produced by a sequencer often referred to only as a "read".

Sequence Read Archive (SRA) - A public repository for storage for whole genome sequencing projects. SRA is maintained by the National Center for Biotechnology Information (NCBI)

Sequencing Run - The complete process of sequencing a sample. Steps include extracting and preparing the genetic material for sequencing then sequencing the genetic material.

Box 1.2: **Bacterial sequence analysis paradigms**

***De novo* assembly**

During the sequencing process the DNA is split into many small fragments. Reassembling these fragments into contiguous sequences without prior knowledge (i.e. a reference genome) is called *de novo* assembly. *De novo* assembly has two common approaches, greedy and graph algorithms (Nagarajan & Pop, 2013). Greedy algorithms require pairwise calculation between all reads, thus do not perform well on large number of reads common for bacterial sequencing. Graph algorithms decompose reads into smaller subsequences (nodes) and identify overlaps between subsequences (edges). Contiguous sequences are then generated by navigating connected nodes in the graph. Currently the most commonly used *de novo* assemblers have made use of De Bruijn graphs (Compeau, Pevzner & Tesler, 2011). The use of paired-end reads can improve the accuracy and quality of *de novo* assembly.

Reference mapping

Reference mapping is the process of aligning sequence reads to a reference sequence. Most often a completed genome is used as a reference sequence, but *de novo* assemblies, genes and proteins can also be used as a reference sequence. Using alignment tools (Li & Durbin, 2009b; Langmead & Salzberg, 2012), each sequence read is mapped to a region of the reference sequence in which the most similarity is shared. After each read is mapped an alignment file called Sequence Alignment/Map (SAM) (Li et al., 2009b) is generated. The SAM file contains information about the quality of the alignment and location of alignment. Reference mapping is a powerful tool to identify novel mutations (single nucleotide polymorphisms and insertions/deletions) in a sequenced sample. The use of paired-end reads can improve the accuracy of reference mapping. Another form of reference mapping is the well-known BLAST algorithm (Altschul et al., 1990; Camacho et al., 2009). BLAST databases hosted by NCBI allow users to map their sequence of interest to millions publicly available sequences.

Sequence decomposition

Alignments are computationally expensive and become time consuming as database sizes increase (Kemena & Notredame, 2009). It has become necessary to implement alignment-free methods with sequence decomposition, or k-mers. A k-mer is any substring of length k that is contained in a string. In the context of DNA sequences, it refers to all subsequences of length k contained in a sequence. For example, k=31, a string of DNA will be split into substrings of 31 nucleotides, or 31-mers, and a count for each 31-mer will be output (Marçais & Kingsford, 2011a; Deorowicz et al., 2015). Sequence decomposition has many applications in sequence analysis including sequence error correction (Song, Florea & Langmead, 2014; Sheikhizadeh & de Ridder, 2015), *de novo* assembly (Bankevich et al., 2012; Peng et al., 2012; Koren et al., 2017; Nurk et al., 2017) and taxonomic identification of metagenomic sequences (Wood & Salzberg, 2014; Koslicki & Falush, 2016; Breitwieser & Salzberg, 2018).

Figure 1.1: A basic workflow for a sequencing run

Image is reprinted from open access article Head et al. (Head et al., 2014).

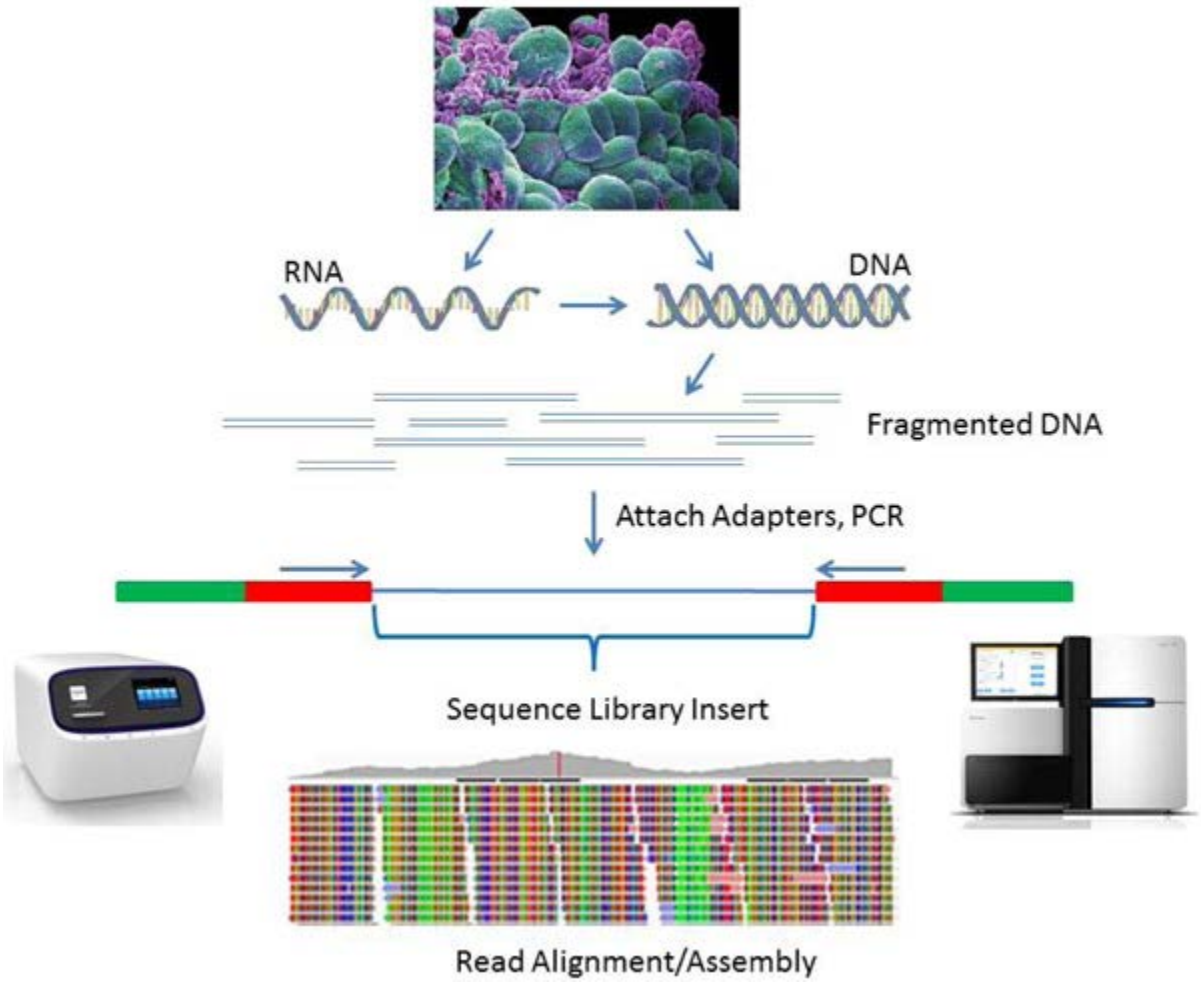


Figure 1.2: A broad representation of a standard workflow for bacterial sequence analysis

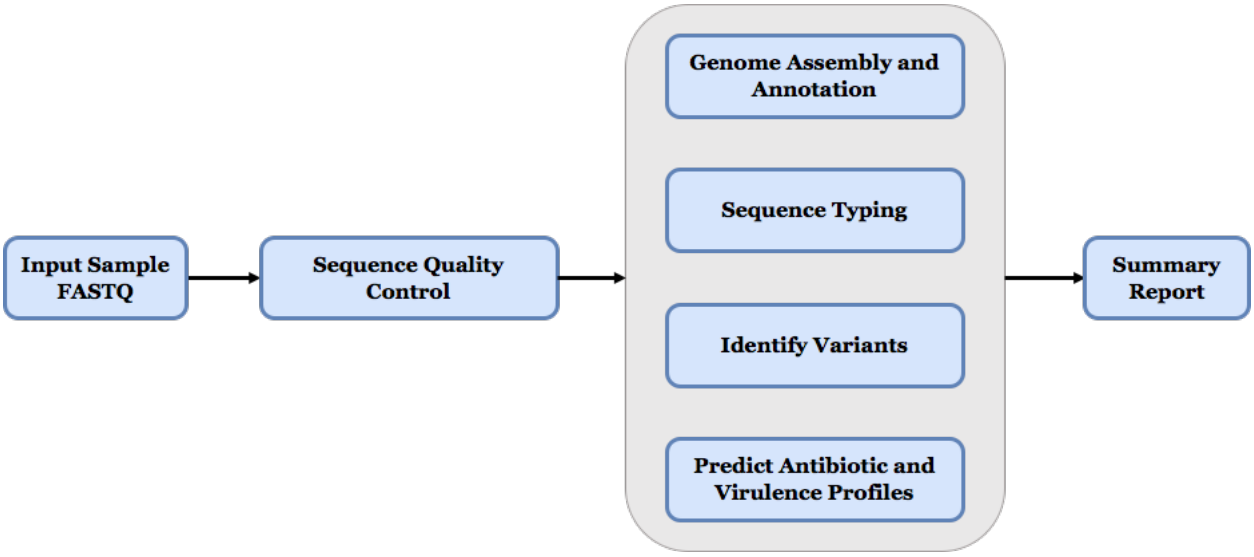


Figure 1.3: Visualization of per-base sequence quality

The quality of a sequencing run can be visualized using the FastQC tool (https://github.com/s-andrews/FastQC). Example reports available from FastQC are depicted below. Panel (A) presents a high quality sequencing run and (B) a low quality sequencing run. These visualizations can indicate where to trim reads based on the average Q-score.

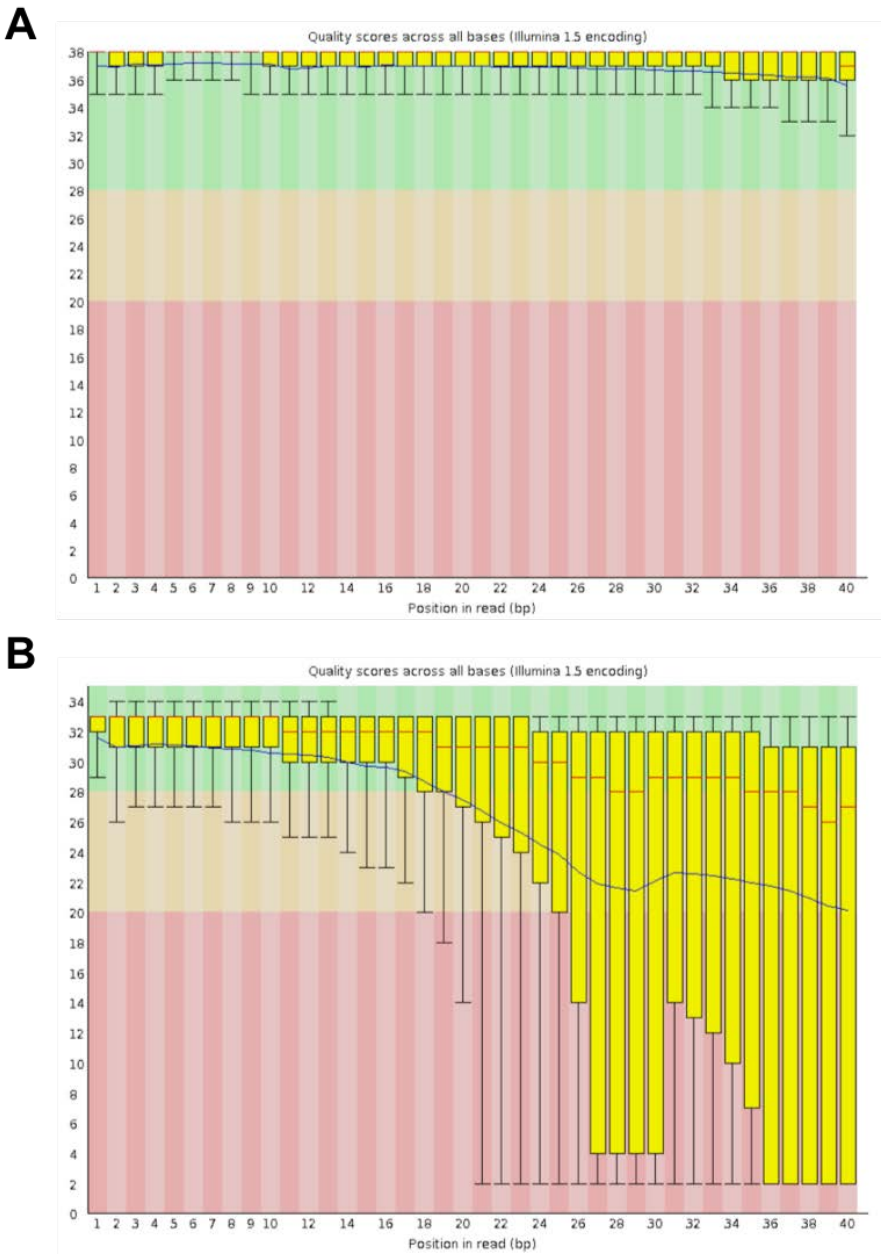


Table 1.1: Most commonly used sequencing technologies for bacteria

Generation	Make	Method	Runtime	Output	# of Reads	Read Length	Error Rate
Second	Illumina	Synthesis	1-6 days	15Gbp-1Tbp	billions	100-300bp	< 0.1%
Third	Oxford Nanopore	Nanopore	1min - 48 hours	<5Gb	10k-1M	> 1kb	5-15%
Third	Pacific Biosciences	Single Molecule	30min-20 hours	500Mb-10Gb	50k-500k	10-15kb	10-15%

Table 1.2: Bioinformatic tools for bacterial sequence analysis

Sequence Quality Control		
Tool	**Description**	**Link**
Ace	k-mer based error correction	https://github.com/sheikhizadeh/ACE/
BBDuk	Quality trimming and contaminant removal	https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/
BLESS2	*k*-mer based error correction	https://sourceforge.net/p/bless-ec/wiki/Home/
FastQC	Visualize the sequencing quality	https://github.com/s-andrews/FastQC
Hammer	*k*-mer based error correction	http://bix.ucsd.edu/projects/hammer/
Lighter	*k*-mer based error correction	https://github.com/mourisl/Lighter
Quake	Corrects correct substitution sequencing errors	http://www.cbcb.umd.edu/software/quake/
SPAdes	*k*-mer based error correction	http://cab.spbu.ru/software/spades/
Trimmomatic	Quality trimming and contaminant removal	https://github.com/timflutre/trimmomatic

Genome Assembly and Annotation		
Tool	**Description**	**Link**
Aragorn	tRNA and tmRNA prediction	http://mbio-serv2.mbioekol.lu.se/ARAGORN/
Bandage	Visualize the quality of an assembly	https://github.com/rrwick/Bandage
BLAST+	Sequence similarity search	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Canu	Long read assembler	https://github.com/marbl/canu
CRISPR Recognition Tool	CRISPR prediction	http://www.room220.com/crt/
GeneMarkS	Bacterial gene prediction	http://opal.biology.gatech.edu/GeneMark/
HMMER	Sequence homolog identification	http://hmmer.org/
IDBA-UD	Metagenomic *de novo* assembler	https://github.com/loneknightpy/idba
MEGAHIT	Metagenomic *de novo* assembler	https://github.com/voutcn/megahit
metaSPAdes	Metagenomic *de novo* assembler	http://cab.spbu.ru/software/meta-spades/
Pilon	Automatically improve draft assemblies	https://github.com/broadinstitute/pilon
Prodigal	Bacterial gene prediction	https://github.com/hyattpd/Prodigal
Prokka	A complete pipeline for prokaryotic genome annotation	https://github.com/tseemann/prokka

RNAmmer	rRNA prediction	http://www.cbs.dtu.dk/cgi-bin/sw_request?rnammer
Shovill	SPAdes wrapper for improved error correction and assemblies	https://github.com/tseemann/shovill
SPAdes	*De novo* assembler	http://cab.spbu.ru/software/spades/
Unicyler	Hybrid assembler for short and long reads	https://github.com/rrwick/Unicycler

Sequence Typing		
Tool	**Description**	**Link**
Ariba	Gene identification via local assembly	https://github.com/sanger-pathogens/ariba
BLAST+	Sequence similarity search	https://blast.ncbi.nlm.nih.gov/Blast.cgi
MentaLiST	k-mer based typing	https://github.com/WGS-TB/MentaLiST
SRST2	Gene identification via mapping	https://github.com/katholt/srst2

Identifying Variants		
Tool	**Description**	**Link**
bedtools	Collection of tools to analyze reference mapping formats	https://github.com/arq5x/bedtools2
Bowtie2	Sequence aligner	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
BWA	Sequence aligner	https://github.com/lh3/bwa
GATK	Collection tools for variant discovery	https://software.broadinstitute.org/gatk/
kSNP	*k*-mer based variant discovery	https://sourceforge.net/projects/ksnp/
ParSNP	SNP detection of closely related samples	https://github.com/marbl/parsnp
Picard Tools	Collection of tools to analyze reference mapping formats	https://github.com/broadinstitute/picard
Samtools	Tools for manipulating alignments	https://github.com/samtools/samtools
Snippy	Variant discovery pipeline for bacterial samples	https://github.com/tseemann/snippy

Antimicrobial Resistance and Virulence Factors		
Tool	**Description**	**Link**
ARG-ANNOT	Uses a curated antibiotic resistance database	http://en.mediterranee-infection.com/article.php?laref=283%26titre=arg-annot

Ariba	Gene identification via local assembly	https://github.com/sanger-pathogens/ariba
SRST2	Gene identification via mapping	https://github.com/katholt/srst2

K-mer analysis		
Tool	**Description**	**Link**
Braken	Computes abundance of species from metagenomic sequences	https://ccb.jhu.edu/software/bracken/
Jellyfish	*k*-mer counter	https://github.com/gmarcais/Jellyfish
KMC2	*k*-mer counter	http://sun.aei.polsl.pl/kmc
Kraken	Taxonomic classification of metagenomic sequences	https://github.com/DerrickWood/kraken
KrakenHLL	Metagenomics classifier with unique k-mer counting	https://github.com/fbreitwieser/krakenhll
mash	Fast genome and metagenome distance estimation	https://github.com/marbl/Mash
MetaPalette	Metagenomic profiling and phylogenetic distances	https://github.com/dkoslicki/MetaPalette

Pan-Genome		
Tool	**Description**	**Link**
LS-BSR	Compares all coding regions to generate pan-genome	https://github.com/jasonsahl/LS-BSR
panX	Pan-genome analysis with graphic interface	http://pangenome.tuebingen.mpg.de/
ParSNP	Core genome alignment of closely related samples	https://github.com/marbl/parsnp
PGAP-X	Pan-genome analysis with graphic interface	https://pgapx.ybzhao.com/
Roary	Large-scale prokaryote pan genome analysis	https://github.com/sanger-pathogens/Roary

Phylogenetics		
Tool	**Description**	**Link**
BEAST	Rooted, time measured phylogenies	http://beast.community/
ClonalFrameML	Identifies recombinant regions	https://github.com/xavierdidelot/ClonalFrameML
FastTree	Approximates a maximum likelihood phylogeny	http://www.microbesonline.org/fasttree/

IQ-Tree	Implements ultrafast bootstrap approximation	http://www.iqtree.org/
mashtree	Rapidly creates neighbor joined tree from mash output	https://github.com/lskatz/mashtree
MrBayes	Bayesian inference and model choice	https://github.com/NBISweden/MrBayes
PhyML	Accurate and highly tunable	http://www.atgc-montpellier.fr/phyml/
RAxML	Accurate and highly tunable	http://www.exelixis-lab.org/

Genome Wide Association Study (GWAS)		
Tool	**Description**	**Link**
BugWAS	Lineage effects and controls population structure	https://github.com/sgearle/bugwas
PLINK	Linear and logistic regression based GWAS	http://zzz.bwh.harvard.edu/plink/
ROADTRIPS	Allows partial or completely unknown population structure	http://www.stat.uchicago.edu/~mcpeek/software/ROADTRIPS/

Table 1.3: Probability of sequencing error for given Phred Quality scores

Phred Quality Score	Probability of Sequencing Error	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Table 1.4: Publicly available bacterial genome sequencing projects

As of May 2018, there were 733,815 sequenced samples, representing over 6,000 bacterial species, available from the European Nucleotide Archive. Of these, 19 bacterial species had more than 5,000 sequenced samples. These bacterial species, each associated with human health, represented 76% (562,849) of the sequenced bacterial samples.

Organism	**Sequenced Samples**
Salmonella enterica	144,551
Escherichia coli	70,407
Staphylococcus aureus	55,517
Streptococcus pneumoniae	51,958
Mycobacterium tuberculosis	44,269
Campylobacter jejuni	27,619
Streptococcus pyogenes	23,374
Listeria monocytogenes	20,133
Neisseria meningitidis	16,410
Klebsiella pneumoniae	15,274
Clostridioides difficile	12,903
Streptococcus agalactiae	8,106
Neisseria gonorrhoeae	8,043
Enterococcus faecium	7,046
Pseudomonas aeruginosa	6,412
Vibrio cholerae	6,220
Shigella sonnei	6,072
Campylobacter coli	5,668
Acinetobacter baumannii	5,026

Chapter 2: Fine-scale differentiation between *Bacillus anthracis* and *Bacillus cereus* group signatures in metagenome shotgun data

This work has been submitted to PeerJ for review.

Abstract

Background

It is possible to detect bacterial species in shotgun metagenome datasets through the presence of only a few sequence reads. However, false positive results can arise, as was the case in the initial findings of a recent New York City subway metagenome project. False positives are especially likely when species are separated by a small phylogenetic distance and both the pathogen and non-pathogen species are present in the same sample. *Bacillus anthracis,* the etiologic agent of anthrax, is a high-consequence pathogen that shares >99% average nucleotide identity with *Bacillus cereus* group (BCerG) genomes. Our goal was to create an analysis tool that used k-mers to detect *B. anthracis,* incorporating information about the coverage of BCerG in the metagenome sample.

Methods

Using public complete genome sequence datasets, we identified 31-mer signatures that differentiated *B. anthracis* from other members of the *B. cereus* group (BCerG), and from 31-mers conserved in all BCerG genomes (including *B. anthracis*), but *not* in other *Bacillus* strains. We also created a set of 31-mers for detecting the lethal factor gene, the key genetic diagnostic of the presence of anthrax-causing bacteria. We created synthetic sequence datasets based on existing genomes to test the accuracy of a k-mer based detection model.

Results

We found 239,503 *B. anthracis*-specific 31-mers (the *Ba31 set*), 10,183 BCerG 31-mers (the *BCerG31 set*), and 2,617 lethal factor k-mers (the *lef31* set). We showed that false positive *B. anthracis* k-mers - which arise from random sequencing errors - are observable at high genome coverages of *B. cereus*. We also showed that there is a “gray

zone" below ~0.18x coverage of the *B. anthracis* genome sequence, in which we cannot expect with high probability to identify lethal factor k-mers. We created a linear regression model to differentiate the presence of *B. anthracis*-like chromosomes from sequencing errors given the BCerG background coverage. We showed that while shotgun datasets from the New York City subway metagenome project had no matches to *lef31* kmers and hence were negative for *B. anthracis*, some samples showed evidence of strains very closely related to the pathogen.

Discussion

This work shows how extensive libraries of complete genomes can be used to create organism-specific signatures to help interpret metagenomes. We contrast "specialist" approaches to metagenome analysis such as this work to "generalist" software that seeks to classify all organisms present in the sample and note the more general utility of a k-mer filter approach when taxonomic boundaries lack clarity or high levels of precision are required.

Introduction

There is great interest in the use of shotgun metagenome data to detect pathogens in clinical and environmental samples. A large number of bioinformatic tools have been developed (McIntyre et al., 2017) that use different algorithmic approaches to rapidly parse and analyze sequence data files. Over the last 8-10 years, these data have been generated primarily by Illumina sequencing technology. Typically, sequences from metagenomic data files are matched against public reference databases, such as NCBI RefSeq. Consistency of matches across the tree of life is dependent therefore on the database entries being correctly labelled, having similar levels of representation across

species, and having species defined in a consistent manner. However, we are beginning to understand how the skewed representation of taxa contained in the database sometimes affects sampling accuracy (Nasko et al., 2018a). Furthermore, the classification of many bacterial species harks back to distinctions based on morphological, biochemical and virulence characteristics, made prior to the advent of DNA sequencing. Sometimes, unusually close species boundaries can confound metagenomic classifiers and result in false positive matches. In 2015, Afshinnekoo et al (Afshinnekoo et al., 2015) published initial findings from an extensive study of the New York Subway metagenome, which claimed that they had detected bacteria responsible for anthrax (*Bacillus anthracis*) and plague (*Yersinia pestis*). While these misidentifications were swiftly corrected by more targeted analyses (Mason, 2015), indistinct or fuzzy boundaries between species may yield many errors of this nature

B. anthracis, the pathogen that is the focus of this work, is a Gram-positive bacterium that forms tough endospores allowing it to survive dormant in the environment for years. The 5.2 (Mbp) main chromosome shares an ANI (average nucleotide identity) (Konstantinidis & Tiedje, 2005a) in excess of 99% with other members of the collection of species known as the '*Bacillus cereus* group' (*BCerG*)(Helgason et al., 2000). The most common species in this group are *B. cereus*, *B. thuringiensis* and *B. mycoides* (Helgason et al., 2000; Zwick et al., 2012). The recommended level of difference between bacterial species is an ANI of 95% (Konstantinidis & Tiedje, 2005a). While BCerG strains are mostly opportunistic pathogens of invertebrates and are commonly found in soil, *B. anthracis* kills mammals (Carlson et al., 2018). Spores are generally found at high titers in soils where animals have recently died from anthrax.

Phylogeographic analysis has shown that *B. anthracis* is probably native to Africa, with only recent transfer of a limited number of lineages to other continents (Keim & Wagner, 2009). For these reasons, it would be an unusual outcome to find spores in the New York subway.

What sets *B. anthracis* apart from other BCerG strains is the presence of two plasmids: pXO1 (181 kb), which carries the lethal toxin genes, and pXO2 (94 kb), which includes genes for a protective capsule. Without either of these plasmids, *B. anthracis* is considered attenuated in virulence and unable to cause classic anthrax (Dixon et al., 1999). Plasmids from other BCerG genomes may be very similar to pXO1 and pXO2 but lack the important virulence genes. Rarely, BCerG strains carry pXO1 and appear to cause anthrax-like disease (Hoffmaster et al., 2004; Hoffmann et al., 2017); pXO2-like plasmids are also quite common in BCerG and other *Bacillus* species(Pannucci et al., 2002; Cachat et al., 2008).

Shortly after the release of the NYC subway metagenome paper, we produced a blog post (Petit et al., 2015) that critically re-analyzed these data in the light of what was known about *B. anthracis* genomics. This work, and other critiques, led to reassessment of the data and revisions to the original manuscript. In this paper, we incorporate some of the results introduced informally on our blog and extend them to create a k-mer based approach - using recent public *B. anthracis* and BCerG data - to analyze in greater detail how to search for traces of *B. anthracis* in shotgun metagenome data. While elements of this method are necessarily specific to *B. anthracis* and the context of the BCerG group,

the general strategy has far broader utility and this work is a model for future “specialist” studies based on k-mer filtering.

Methods

Metagenome data and reference genome sequences

Shotgun metagenomic data from the “NYC” study SRP051511 (Afshinnekoo et al., 2015) were downloaded from the Sequence Read Archive (SRA) with sra-tools (v2.8.2, https://github.com/ncbi/sra-tools). Reference genomes for different taxonomic groups were downloaded from the NCBI Nucleotide database in April 2018 with the following queries:

> All BCerG genomes = ‘txid86661[Organism:exp] AND "complete genome"[Title] AND refseq[filter] AND 3000000:7000000[Sequence Length]’
>
> All *Bacillus* genomes = 'txid1386[Organism:exp] NOT txid86661[Organism:exp] "complete genome"[Title] AND 3000000:7000000[Sequence Length] AND refseq[filter]’

Bacillus anthracis genomes were included in the BCerG genome query. The lethal factor gene was extracted from completed pXO1 plasmids downloaded with the following query:

pXO1 plasmid = ‘pXO1[Title] AND 140000:200000[Sequence Length] ‘

Mapping metagenome data to *B. anthracis* plasmids and chromosomes

B. anthracis positive samples and control samples were mapped against reference pXO1 (CP009540) and pXO2 (NC_007323) plasmids and reference *B. anthracis* (CP009541) and *B. cereus* (NC_003909) completed genomes with BWA (v0.7.5a-r405, (Li & Durbin, 2009b)). The aligned reads in SAM format were converted to sorted BAM and indexed with samtools (v1.1, (Li et al., 2009b)). The per base coverage was extracted with genomeCoverageBed from bedtools (v2.16.2, (Quinlan & Hall, 2010)). Coverage across the plasmids and chromosomes was plotted for multiple sliding windows with a custom Rscript. Mapped reads were extracted and saved in FASTQ with bam2fastq (v1.1.0, https://gsl.hudsonalpha.org/information/software/bam2fastq) and FASTA format with fastq_to_fasta from FASTX Toolkit (v0.0.13.2, (Gordon & Hannon, 2010)). Scripts, parameters, and output are available at this site (Petit et al., 2015): https://github.com/Read-Lab-Confederation/nyc-subway-anthrax-study.

Custom 31-mer assay for *B. anthracis* and *Bacillus cereus* Group

In preliminary analysis we found four BCerG genomes misclassified in the NCBI Taxonomy database as not being part of the BCerG (see the Results section). To create a rational method to assign taxonomy to genomes for this study we used mash (v2.0, (Ondov et al., 2015)) to reclassify mislabelled *Bacillus* genomes as *B. anthracis*, non-*anthracis* BCerG, or non-BCerG. We identified *Bacillus anthracis* strain 2002013094 (NZ_CP009902) as the most distant (Mash distance 0.000687) *B. anthracis* member

from from *B. anthracis* str. Ames (NC_003997). We also identified *Bacillus cytotoxicus* NVH 391-98 (NC_009674) as the most distant (Mash distance 0.135333) BCerG member from *B. anthracis* str. Ames (NC_003997). We then determined the Mash distance of all *Bacillus* genomes from *B. anthracis* str. Ames. We used the Mash distance to reclassify each Bacillus genome as *B. anthracis* (Mash distance <= 0.000687), non-*anthracis* BCerG (Mash distance <= 0.135333), or non-BCerG (Mash distance > 0.135333). A phylogeny of all completed Bacillus genomes was created with mashtree (v0.32, https://github.com/lskatz/mashtree).

Sequence 31-mers were extracted and counted with Jellyfish (v2.2.3, (Marçais & Kingsford, 2011a)) and partitioned into two distinct sets characteristic of BCerG (BCerG31) and *B. anthracis* (Ba31) (**Figure 2.1**). The BCerG31 and Ba31 sets were initially comprised of 31-mers conserved within *every* member of BCerG (including *B. anthracis*) and those restricted to only *B. anthracis*, respectively. Any Ba31-mers found in non-*anthracis* BCerG members or non-BCerG genomes were filtered out. Likewise, any BCerG31-mers found in non-BCerG Bacillus genomes were filtered out. 31-mers found in rRNA were filtered out with a Jellyfish database created from the SILVA rRNA database (Quast et al., 2013). We further filtered the Ba31 and BCerG31 sets using the non-redundant nucleotide sequence database (NT v5, downloaded April 2017). We used BLASTN (v2.8.0, (Camacho et al., 2009)) to align Ba31 against non-*anthracis* BCerG sequences and BCerG31 against non-BCerG sequences. 31-mers with an exact match were filtered out.

Finding the limits for lethal factor-based detection of *B. anthracis*

We used *B. anthracis* whole genome shotgun sequencing projects to determine the limit of detection of lethal factor k-mers (lef31). We defined lef31 as the unique set of 31-mers identified in *lef* genes downloaded from the NCBI Nucleotide database (previously described) (**Figure 2.1**). *B. anthracis* projects were identified from the SRA with the following query:

> *B. anthracis* projects = 'genomic[Source] AND random[Selection] AND txid86661[Organism:exp] AND paired[Layout]) AND wgs[Strategy] AND "Illumina HiSeq"'

In this work we have assumed a 95% 'confidence limit' for detection of the lethal factor k-mers, so that detection is held to fail if fewer than 95% of a set of random subsamples are found to contain *at least one* lethal factor k-mer. The threshold is then obtained through computational experiment. For each project, we started at 0.2x *B. anthracis* genome coverage and extracted 100 random subsamples of sequences, using Jellyfish as before to determine if at least one lethal factor k-mer was present. We then continued this process, reducing the coverage until fewer than 95% of the subsamples contained at least one lethal factor k-mer. The previous coverage was then recorded as the limit of detection of the lethal toxin for a given sample.

Assessing Quality of *B. anthracis* and *B. cereus* Group specific 31-mers

We used ART (vMountRainier-2016-06-05, (Huang et al., 2012)) to simulate 100 bp reads with the built-in Illumina HiSeq 2000 error model for each non-*anthracis*

Bacillus genome. We simulated coverages ranging from 0.01x to 15x to determine if false positive Ba31 matches were uniform across non-*anthracis* BCerG members. We counted 31-mers for each simulated read set with Jellyfish as previously described. We then used the k-mer counts to determine the sensitivity and specificity of our *B. anthracis* and BCerG specific k-mers. We found the false positive Ba31 counts to be higher in non-*B. anthracis* genomes that were most closely related to *B. anthracis* (please see results section). A subset of non-*B. anthracis* BCerG genomes with a Mash distance less than 0.01 from *B. anthracis*, previously described, were selected as our model set. We further simulated coverages from 15x to 100x to match levels of coverage observed in the NYC dataset. We then applied linear regression, implemented in the R base stats package, on this subset to develop a predictive model with the observed Ba31 count as our dependent variable and the observed BCerG k-mer coverage as our independent variable.

Prediction of low coverage *B. anthracis* chromosome in shotgun sequencing datasets

We used ART to simulate metagenomic mixtures of *B. anthracis* str. Ames (NC_003997) and *B. cereus* strain JEM-2 (NZ_CP018935). *B. cereus* strain JEM-2 was selected because it was the closest non-*anthracis* BCerG member to *B. anthracis* str. Ames (Mash distance 0.00873073). We used coverages between 0-100x for *B. cereus* and coverages between 0-0.2x for *B. anthracis*. A python script (subsample-ba-lod.py) was created to simulate mixtures for each pairwise combination of *B. cereus* and *B. anthracis* coverages. For each mixture the *B. anthracis* and BCerG 31-mers were counted with Jellyfish as previously described. This process was repeated 20 times per

pairwise combination of coverages. The model was applied to determine what level of *B. anthracis* coverage was required to differentiate observed Ba31-mers from sequencing errors.

We counted *B. anthracis*, BCerG and lethal factor 31-mers for each sample in the NYC study. The model was applied to these counts to determine if observed *B. anthracis* k-mers exceeded the level expected due to sequencing errors.

We processed each of the subsampled mixtures and samples from the NYC study with KrakenHLL (v0.4.7, (Breitwieser & Salzberg, 2018)). The standard Kraken database (built April 2017) was used for this analysis. From the final Kraken report, the number of reads and unique k-mers identified for *B. anthracis* were extracted. We compared these results to our method.

Output, figures, job parameters and scripts from this study are available in a git repository hosted at: https://github.com/rpetit3/anthrax-metagenome-study.

Results

NY subway metagenome sequences map to core regions of *B. anthracis* and *B. cereus* chromosome and plasmids but not to lethal factor gene

In the original analysis of the subway metagenome (Afshinnekoo et al., 2015), two samples (P00134 (SRR1748707, SRR1748708), and P00497 (SRR1749083)) were reported to contain reads that mapped to *Bacillus anthracis* based on results obtained using the Metaphlan software (Segata et al., 2012). We found that 792,282 reads from

P00134 and 270,964 reads from P00497 mapped to the *B. anthracis* strain Sterne chromosome. The reads aligned along the entire length of the chromosome, forming a characteristic peak at the replication origin, a pattern often seen when other bacterial chromosomes have been recovered from metagenome samples (Brown et al., 2016). However, a similar number of reads from P00134 and P00497 (765,466 reads and 265,776 reads, respectively) mapped to the *B. cereus* 10987 chromosome. We also found that P00134 and P00497 reads mapped to the both the pXO1 and pXO2 plasmids in conserved "backbone" regions (Rasko et al., 2007) but that no read mapped to the mobile element containing the *lef* lethal factor gene. These results showed that the close taxonomic relationship of *B. anthracis* and BcerG made identification of the biothreat agent by mapping reads alone unreliable. In addition, the pXO1 and pXO2 plasmids were not reliable as positive markers for *B. anthracis* at low genome coverages (when the *lef* gene may not be sampled, see next section) because backbone sequences cross-matched against plasmids found in BCerG strains.

B. anthracis genome coverage below 0.18x is a "gray area" for detection, where lethal toxin genes may not be sampled

The best test for presence of virulent *B. anthracis* (or virulent *B. cereus* strains containing pXO1) is detection of the lethal factor gene (2,346 bp) (Bragg & Robertson, 1989). However, at low sequence coverage of the pathogen, it is not certain that reads from this gene will be present (given the 3:1 copy number ratio of pXO1 to *B. anthracis* chromosome (Read et al., 2002) the ratio of chromosome to *lef* is ~620:1). We identified 2,617 31-mers present in 36 *lef* gene sequences and called this set "lef31". To estimate the level coverage at which we would expect (with probability above some threshold

value, here 0.95) to observe lethal factor sequences, we randomly subsampled reads from 164 *B. anthracis* genome projects and tested for the presence of *at least one* lef31 match (**Figure 2.2**). This analysis showed that below ~0.18x-fold *B. anthracis* genome coverage (approximately 9,360 100 bp paired end reads), we would have a < 95% chance of sampling lethal factor *even if the lef gene were present*.

Conserved and specific 31-mer sets for *B. anthracis* and BCerG chromosomes

The results of the previous section showed that at low genome coverage, the presence of *B. anthracis* chromosomal markers was more reliable than those based on the lethal factor gene. In metagenomic samples, in which sequencing coverage is expected to be low for rare organisms, the most reliable way to detect *B. anthracis* was to use chromosomal genetic signatures that distinguished the species from close relatives. We identified 239,503 31-mers conserved in 48 *B. anthracis* reference genomes that were not also detected in the remainder of the *Bacillus* genus (331 genomes), rRNA sequences, or the BLAST non-redundant nucleotide database. We called this set “Ba31”.

We created a second set of 31-mers specific to and conserved in all BCerG genomes (including *B. anthracis*). Surprisingly, our initial analysis produced zero 31-mers specific to all 139 BCerG strains and no other *Bacillus*. Inspection of the whole genome phylogeny (**Figure 2.3**) showed that 4 genomes (NZ_CP007512, NZ_CP017016, NZ_CP020437, NZ_CP025122) that fell within the BCerG clade based on phylogeny had not been classified as BCerG in the NCBI Taxonomy hierarchy. After reclassifying these

strains as BCerG, we identified 10,183 BCerG specific 31-mers, which we called "BCerG31".

High background levels of *B. cereus* strains produce false positive *B. anthracis* specific k-mers due to random sequence errors

We defined 'coverage' of the k-mer sets as the sum of counts for k-mers detected divided by the number of k-mers in the k-mer set. Ba31 and BCerG k-mer coverage had a linear relationship with genome coverage (**Figure 2.4**). The coefficient was less than 1 (0.56 and 0.61 for Ba31 and BCerG31 respectively), because some portions of the chromosomes were not well sampled by the k-mers.

We simulated synthetic data representing subsamples of *B. anthracis* and *B. cereus* at different genome coverages using ART software with an error model based on Illumina short read data (Huang et al., 2012) (**Figure 2.5**). We found a strong linear relationship between Ba31 coverage and BCerG31 coverage within *B. anthracis* genome subsamples (Pearson's Correlation $r=0.99$, $p < 0.001$, **Figure 2.6**). As expected, the same relationship did not appear when we subsampled non-*B. anthracis* BCerG members. However, we did see a small number of Ba31 k-mers detected, which we suspected were due to random errors introduced by Illumina sequencing (**Figure 2.5**). The counts of false positive Ba31 k-mers scaled with the approximate genetic distance to *B. anthracis* (as measured by mash(Ondov et al., 2016)) (**Figure 2.7)**. We simulated synthetic data for a group of BCerG strains most closely related to *B. anthracis* (**Figure 2.3).** We developed a linear regression model to relate BCerG k-mer coverage and

sequencing errors based on this group (**Figure 2.5**). For every unit of BCerG31 k-mer coverage, we predicted 172 Ba31 false positive k-mer counts.

A "specialist" model to interpret patterns of *B. anthracis* genetic signatures in metagenome samples

In real metagenome samples *B. anthracis*, if present, may only account for a low proportion of the total reads and may also be mixed with higher proportions of closely related BCerG strains. We sought to use the k-mer sets developed in the previous sections and knowledge of the *lef* gray zone coverage and BCerG false positive rate to interpret both synthetic and real metagenome datasets. The logic for assignment is shown **Table 2.3** and **Figure 2.8.**

For our synthetic dataset we mixed low coverage *B. anthracis* with higher coverages of BCerG sequence data (see methods). We calculated the BCerG31 and Ba31 coverages for each mixture. Based on the BCerG sequence error model, we calculated the 99% count of Ba31 signatures predicted to be present by sequencing error under the assumption that there was no *B. anthracis* present and that all BCerG were drawn from the most closely related clade (**Figure 2.3**). We also reported whether the Ba31 coverage lay in or above the gray zone **(Table 2.1, Figure 2.9).** When *B. anthracis* was below 0.003x genome coverage (approximately 16,000 bp), we could not distinguish its presence from errors produced in the absence of *B. cereus*. As expected, we found that the level of BCerG coverage determined the lower limit to differentiate genuine Ba31 hits from sequencing errors. At 75x BCerG coverage the required *B. anthracis* coverage to

differentiate Ba31 matches from sequencing errors doubled to 0.006x. The threshold for accurate detection was further raised to 0.01x B. anthracis genome coverage at 100x BCerG coverage.

In contrast, when the samples were classified using KrakenHLL(Breitwieser & Salzberg, 2018), an accurate generalist program based on 31-mers, we found that all were predicted to contain *B. anthracis*, including negative controls (**Table 2.1**). The *B. anthracis* calls were made because of the sequence errors from the high coverage BCerG genomes.

Finally, we tested our model against the NYC dataset (**Table 2.2**). All 1,458 samples were negative for lef31, in line with the conclusion reached from re-analysis of the dataset that *B. anthracis* was absent from the NY subway (Mason, 2015). We found that 1,367 of the 1,458 samples had at least one BCerG31 k-mer match and, of these, 1,085 contained at least one Ba31 match. We identified 34 samples with Ba31 counts above the 99% threshold predicted by the BCerG coverage. These samples did not include the two (P00134 and P00497), previously flagged as *B. anthracis* positive (Afshinnekoo et al., 2015) (**Table 2.2**). KrakenHLL also classified each these 34 samples as positive for *B. anthracis*.

Discussion

In this work we have described a significant update to a *B. anthracis* specific 31-mer set that was introduced in earlier blog posts (Petit et al., 2015; Minot et al., 2015) and we

have shown how this set can be used to interpret *B. anthracis* specific signatures in Illumina metagenome samples. We chose to use k-mer-based signatures for the ease and speed of computation, with the length of 31 nt selected as it was identified as the shortest likely to be unique across bacteria datasets (Koslicki & Falush, 2016).

Some species present unusual challenges for metagenome identification. There is no consistently applied definition for the boundary that divides bacterial species based on DNA sequence identity and in some cases the presence or absence of mobile elements like plasmids and phages are required for speciation. *B. anthracis* is closely related to non-biothreat species and acquires its enhanced virulence from genes on mobile plasmids. Such species can be hard to model using "generalist" programs (such as Kraken) that attempt to classify every read in the dataset into one of thousands of taxonomic groups. We use a "specialist" approach aiming to solve a narrow problem that can be used to augment the predictions of generalist software. Specialist analyses can take advantage of unique features of the system and can also afford more effort in the curation of training data. In this case, we designed 31-mer signature sets based on comparison of hundreds of complete *Bacillus* genomes and we incorporated knowledge of false positive k-mers likely to be produced by close relatives of *B. anthracis*. We also used the fact that the presence of a specific gene (*lef*) was diagnostic for anthrax. In designing our k-mer sets we encountered some rare cases of taxonomic mis-assignment in public datasets and were able to take corrective action (**Figure 2.3**). Generalist programs also rely on the same taxonomy and reference sequence databases, but it is harder to detect small errors that lead to mis-assignments when done on a large scale (Nasko et al., 2018a). If we were to attempt approaches to specifically detect other

known *B. cereus* strains that contain pXO1(Hoffmaster et al., 2004; Klee et al., 2010), we would have to develop and test new k-mer sets based on their unique chromosomal SNPs.

Even when a specialized algorithm has been developed, judgement is still required in interpreting results. In the case of the *Bacillus* genomes in particular, DNA extraction biases may affect results in ways we cannot assess without empirical experiments. We can't tell what proportion of the DNA came from lysis-resistant spores and what proportion was from the more fragile vegetative state, and how this balance might vary between strains across environments. Similarly, using a different sequencing technology, such the Pacific Biosystems SMRT system, with a different error profile, would require recalibration of the model.

Our reanalysis of the NYC data (Afshinnekoo et al., 2015) showed that there was no direct evidence for the lethal factor k-mers in the metagenome samples. This confirms other work (Mason, 2015; Minot et al., 2015; McIntyre et al., 2017), and together with the low prior probability of encountering *B. anthracis* in New York City, suggests that the samples taken were all negative for anthrax. The two samples originally flagged as possibly positive (**Table 2.2**) fell under case 4 (**Table 2.3**), as did 1,049 out of the other 1,456 samples. There were 373 samples with no Ba31 k-mer matches. These are all most likely true negatives, although, as we showed in the synthetic dataset, high BCerG coverage can mask the signal of low coverage *B. anthracis* (**Table 2.1**). To get a true negative would theoretically involve sequencing every cell in the sample (assuming perfectly efficient DNA preparation), which is impossible currently for all but the

simplest communities. The limit of detection will be a complex calculation that involves the amount of DNA sequence generated and the complexity of the microbial community. Negative (and positive) calls ultimately have to be supported through sensitive detection assays such as PCR and/or culture.

We identified 34 samples above the BCerG thresholds for our model (**Table 2.2**). All the samples fell under case 3 except a single sample which fell under case 2 (**Table 2.3**). An outlier of case 3 samples, P00981, taken from a metal handrail on the A train route(Afshinnekoo et al., 2015), had high Ba31 counts (n=20,079). As we collect more genomes of *B. cereus* group we may see more Ba31 k-mers in BCerG genomes. These samples may contain members of yet unencountered lineages more closely related to *B. anthracis* than previously seen, or possibly the result of recent recombination between *B. anthracis* and *B. cereus* genomes (although the latter has not been reported). It important that these strains are isolated, sequenced and added to public databases to iteratively improve pathogen detection. The single case 2 sample, P00738 (**Table 2.2**), was also on a metal handrail from the A train route, although sampled 3 days earlier than P00981. This sample was possibly the most problematic because the Ba31 counts were in the gray zone, meaning there was not enough coverage to rule out *lef* being present. Most likely, this sample contained another near-*B. anthracis* strain, but case 2 samples should be a priority for retesting by culture and PCR methods.

Conclusions

If *B. anthracis*, or another BCerG strain containing pXO1, is present in a shotgun metagenome sample at high genome coverage, identification of *lef* k-mers is a strong signal for the likely presence of anthrax-causing bacteria. We showed that using a B. anthracis specific k-mer set alone to call the presence of *B. anthracis* produced many false positive calls because of common co-resident BCerG bacteria. We developed models to partition cases that contained evidence of possible low coverage *B. anthracis*, accounting for *B. cereus* coverage. However, in simulations, we showed that false negative results can arise when the BCerG coverage is high. Reanalysis of the NYC subway metagenome study confirmed the absence of *B. anthracis* containing *lef* but we found evidence in at least two samples of BCerG strains that contained what were considered *B. anthracis* specific sequences. Culturing strains such as these, genome sequencing and sharing to the public domain will help improve *B. anthracis* detection in metagenome shotgun samples.

Acknowledgements

Thanks to Sam Minot, Chris Greenfield, Chris Mason and his group for discussion relating to this project.

Funding

Partially supported by development funds to TDR from Emory University School of Medicine. Some analysis was performed using the the Seven Bridges NCI Cancer Genomics Cloud pilot, supported in part by the funds from the National Cancer

Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400008C.

Appendix

The following appendix contains figures and tables referenced in the text of this chapter.

Figure 2.1: Flowchart of strategy for primer design

We developed a strategy for selecting the Ba31 and BCerG31 (A) and lethal factor (B) k-mer sets. In A) the outgroup is determined by the k-mer set. For Ba31, the outgroup was comprised of all the non-*B. anthracis* genomes; for BCerG31, it consisted of all non-*B. cereus* group genomes.

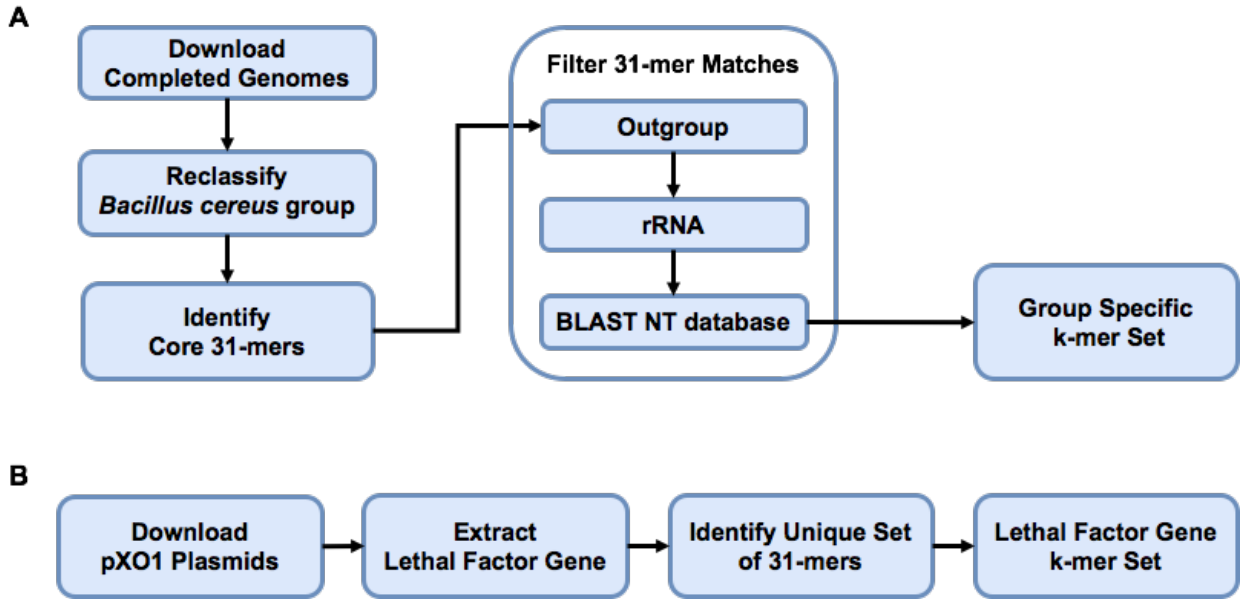


Figure 2.2: Limit of detection for lethal factor toxin k-mers (lef31).

164 *B. anthracis* sequencing projects were subsampled to different levels of genome coverage, with 100 random subsamples obtained for each coverage level. Our ability to detect the the lethal factor gene is assessed by considering the number of these subsamples for which we find at least one lef31 k-mer hit. Two thresholds - 95% and 100% - were employed and are shown as colored series below.

The figure thus shows the percentage of the *B. anthracis* sequencing projects for which 95% (or 100%) of the random subsamples contain at least one lef31 k-mer. Panel (A) shows results with respect to Ba31 k-mer coverage while panel (B) shows the corresponding results for BCerG coverage. The vertical dashed lines show the coverage limits for detection at the respective threshold levels.

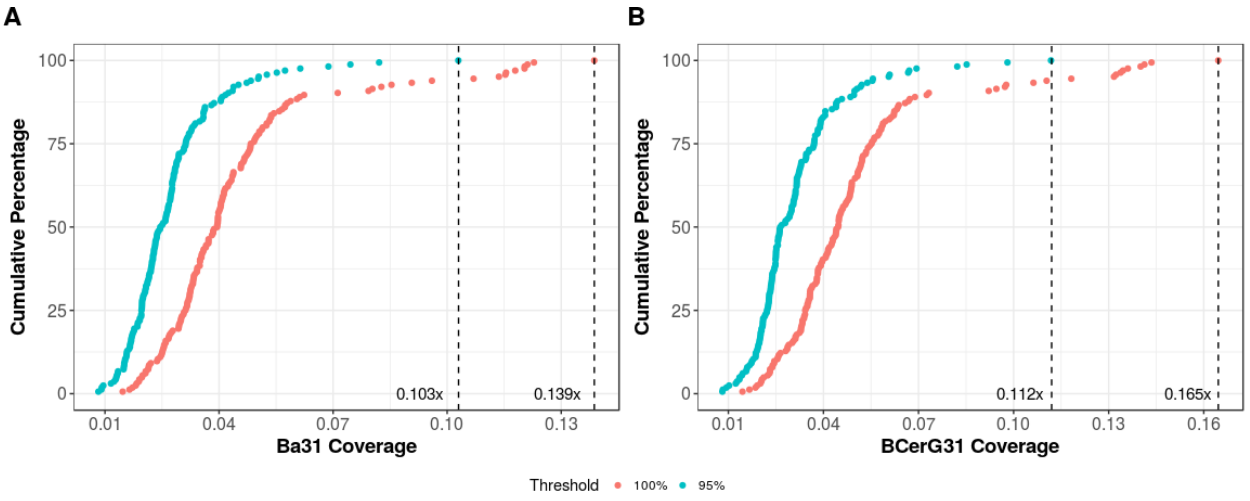


Figure 2.3: Unrooted phylogeny of BCerG genome assemblies used in the study after reclassifying BCerG strains

An unrooted phylogenetic representation of 140 BCerG genomes using Mashtree (v0.32, https://github.com/lskatz/mashtree). Genomes reclassified as BCerG members with mash (v2.0, (Ondov et al., 2015)) are indicated with stars. The clade colored blue are *B. cereus* genomes closely related to *B. anthracis* that were used to model false positive results (**Figure 2.5**).

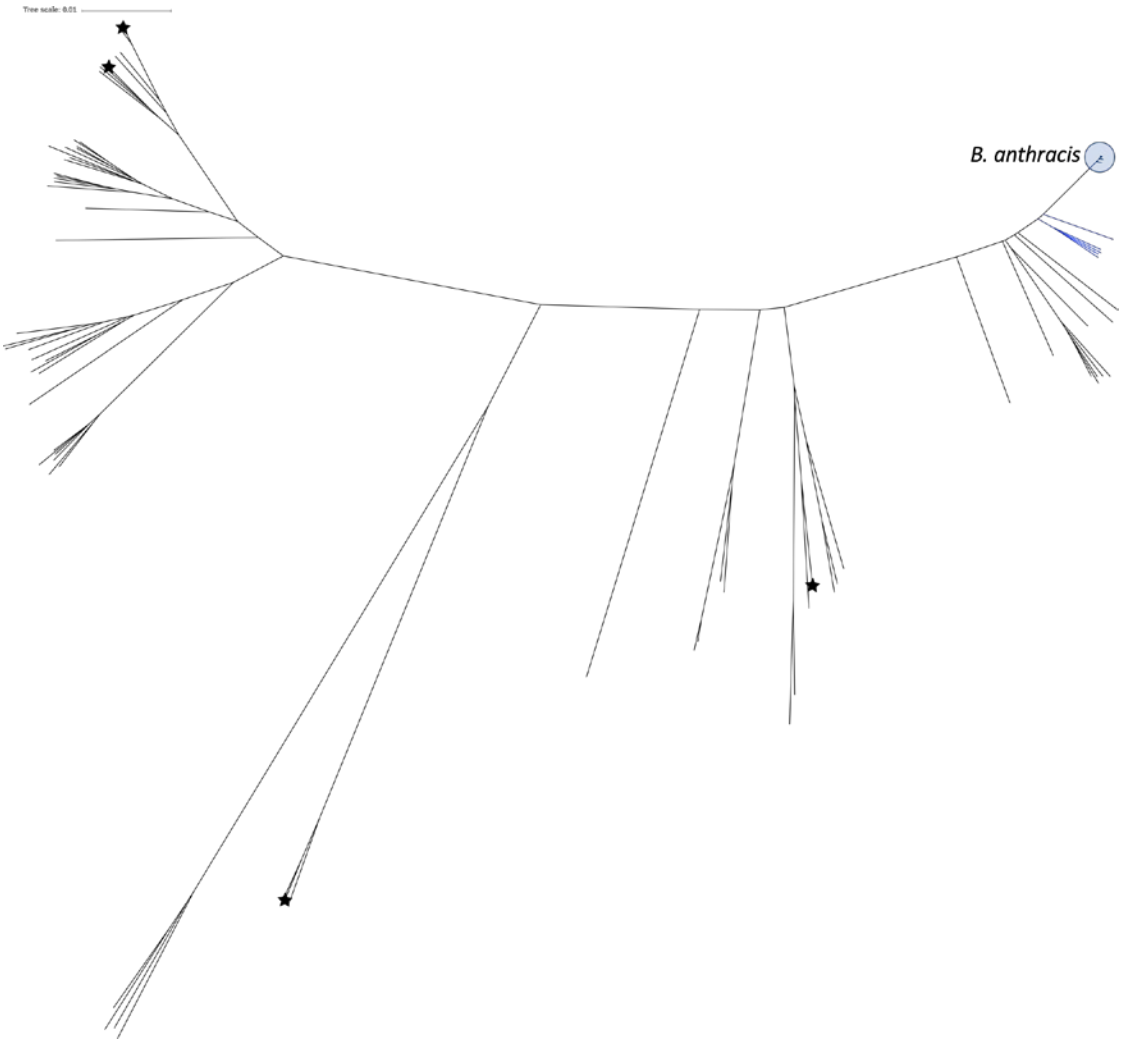


Figure 2.4: Ba31 and BCerG31 coverages have a linear relationship with genome coverage.

We created synthetic FASTQ files of *B. anthracis* (A) and BCerG (B) at different genome coverages and counted Ba31 and BCerG31 k-mers. A linear model with an intercept of 0 is displayed in each case.

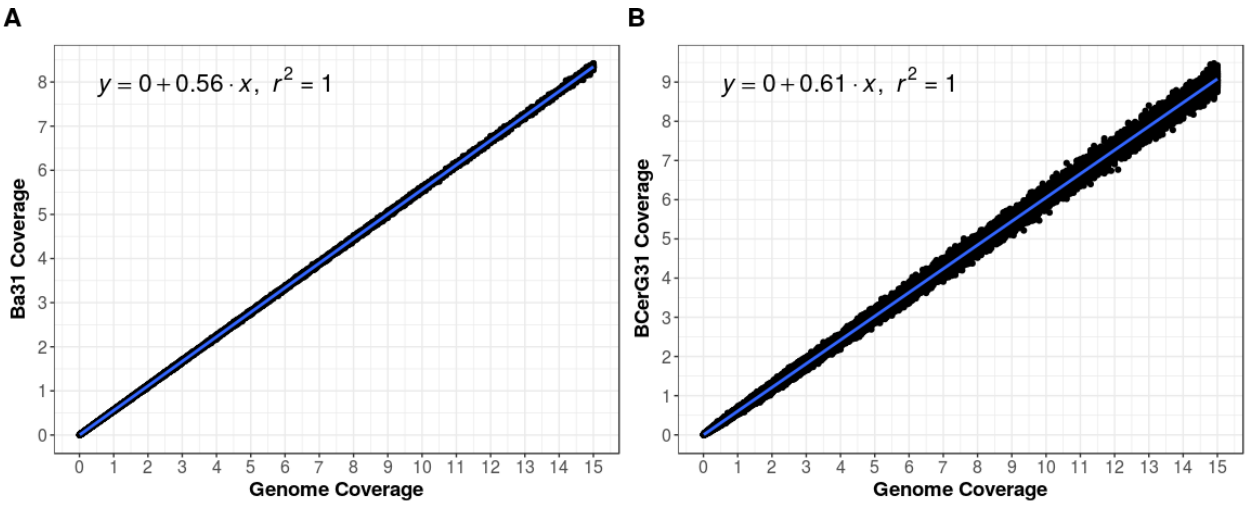


Figure 2.5: Linear regression model fit of BCerG coverage and false positive Ba31 counts.

We created random synthetic FASTQ files based on BCerG chromosomes from the clade closest to *B. anthracis* (blue in **Figure 2.3**) at different genome coverages and counted the false positive Ba31 k-mers. Shown is the fit of a linear regression model with an intercept of 0, with BCerG31 coverage as the independent variable and the Ba31 false positive count as the dependent variable. The solid line shows the predicted values from the model, and the dashed line reflects the upper 99% prediction interval for the parameters, which we use in the analyses above.

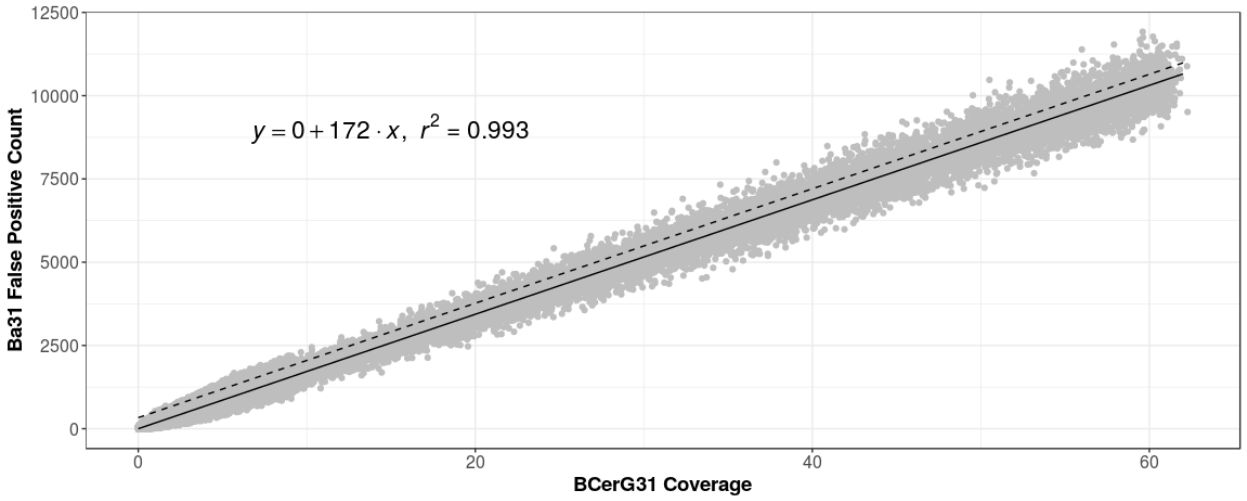


Figure 2.6: In *B. anthracis* genomes, Ba31 coverage is strongly correlated with BCerG31 coverage.

We created synthetic *B. anthracis* FASTQ files at different genome coverages and counted BCerG31 and Ba31 k-mers. A linear model with an intercept of 0 is displayed.

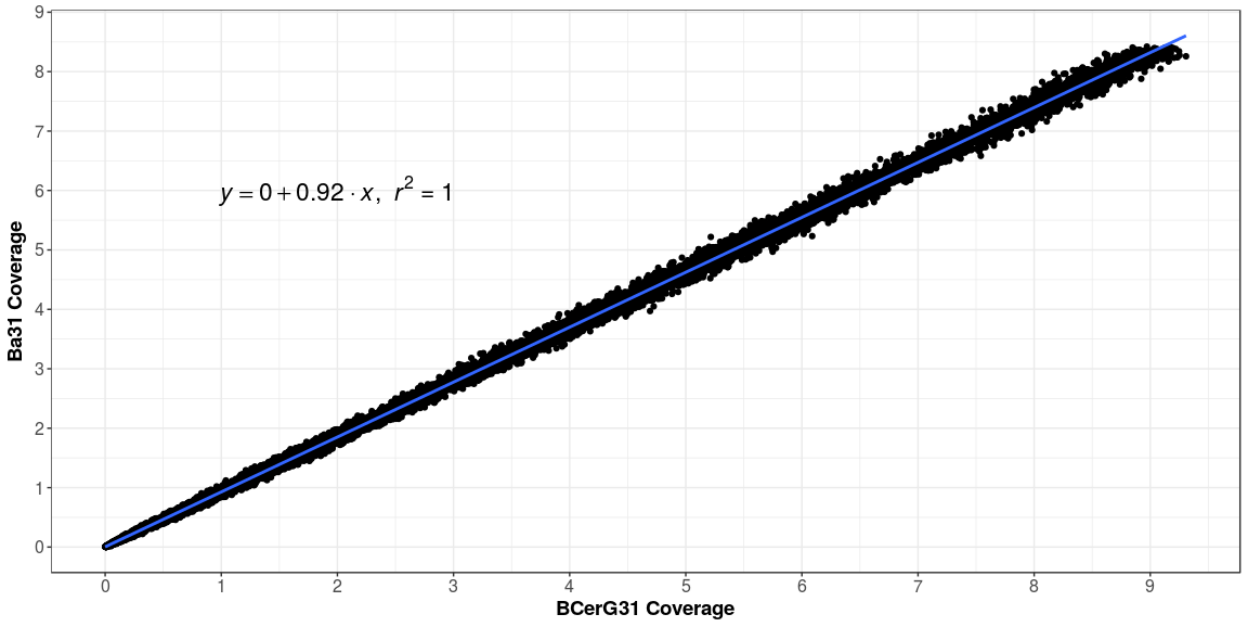


Figure 2.7: The genetic relatedness between *B. anthracis* and non-*B. anthracis* BCerG members affects Ba31 false positive matches.

Synthetic FASTQ files for all BCerG genomes shown in Figure 2 were created and the counts of Ba31 false positive k-mers were plotted against BCerG k-mer coverage. Dots are colored by the Mash distance (Ondov et al., 2016) from the *B. anthracis* str. Ames (NC_003997) genome.

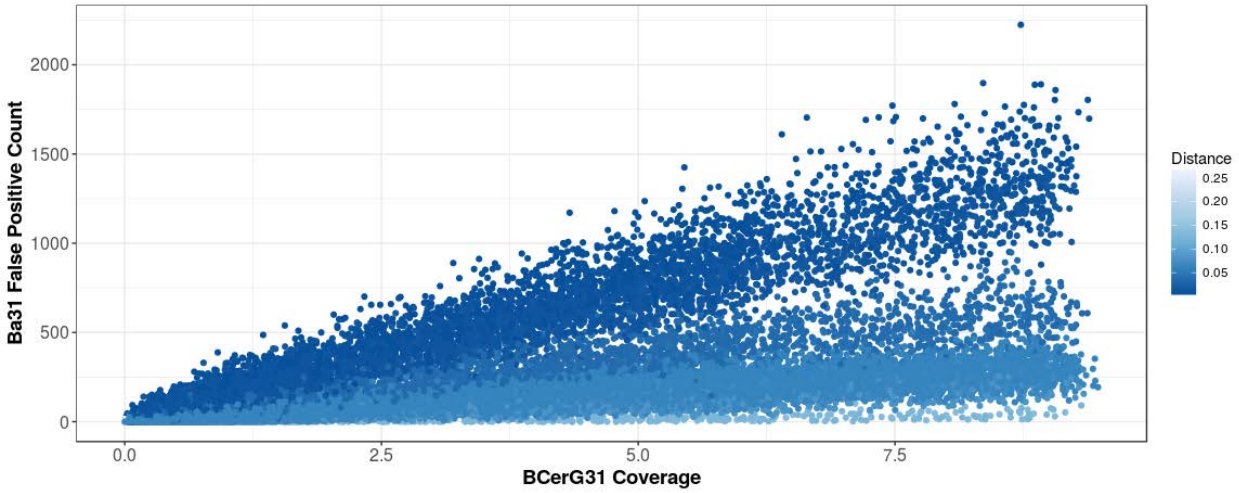


Figure 2.8: A flowchart of potential outcomes of *B. anthracis* detection, given matches to the Ba31 set in a shotgun metagenome dataset

This flowchart presents a visual representation of **Table 2.3**.

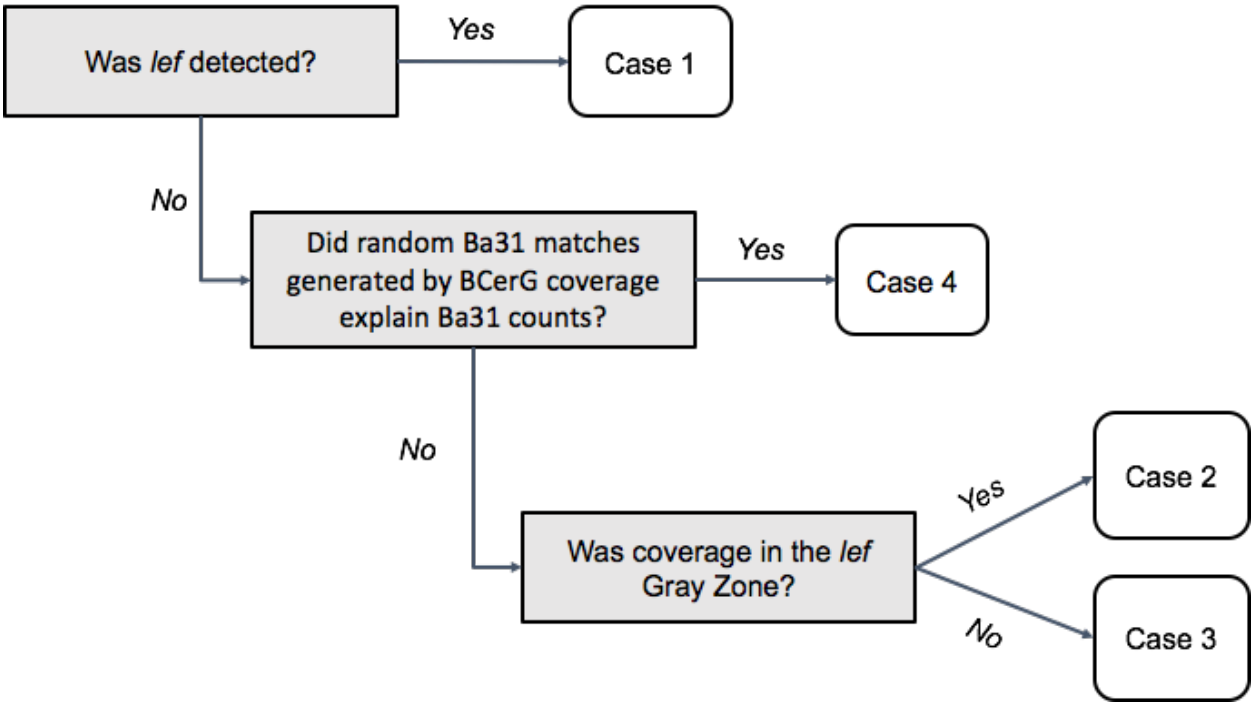


Figure 2.9: Limit of detection for *B. anthracis* k-mers (Ba31) in mixtures of low *B. anthracis* coverage and high *B. cereus* coverage.

We created artificial mixtures of *B. anthracis* and *B. cereus* to determine the limit of detection for *B. anthracis* k-mers (Ba31). Each panel represents a different coverage of *B. cereus* and the points are the different *B. anthracis* coverages. The points are colored red if Ba31 matches could not be differentiated from sequencing errors. The error model is indicated by the solid line and the 99% prediction interval by the dashed line. The first *B. anthracis* coverage value that exceeded the error model is determined as the limit of detection of Ba31.

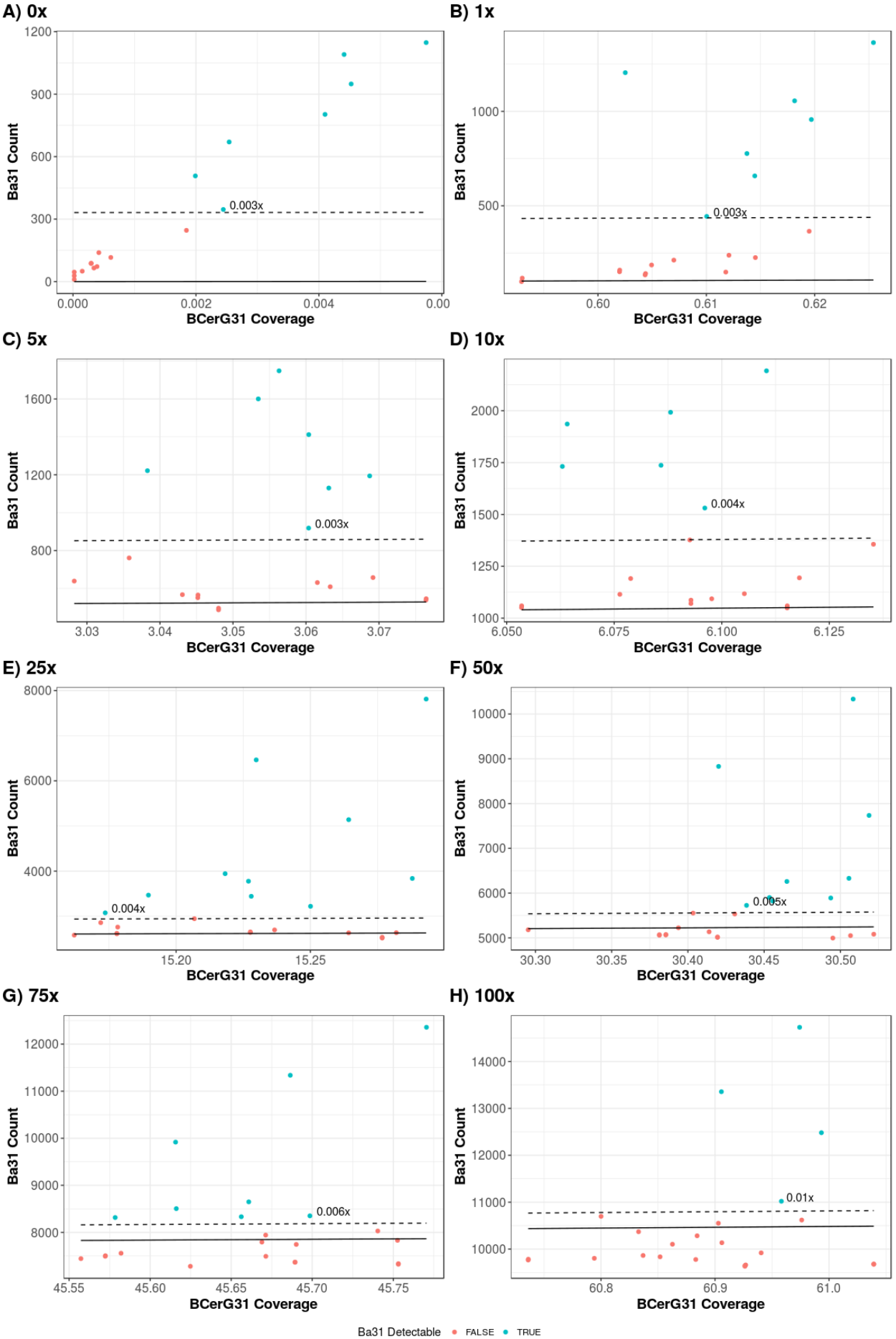

A) 0x
B) 1x
C) 5x
D) 10x
E) 25x
F) 50x
G) 75x
H) 100x
Ba31 Count
BCerG31 Coverage
0.003x
0.004x
0.005x
0.006x
0.01x
Ba31 Detectable
FALSE
TRUE

Table 2.1: Artificial mixtures of low coverage *B. anthracis* and high coverage *B. cereus*

This table shows some key results from more than 300 artificial mixtures of *B. anthracis* and *B. cereus* sequences created to test our specialized model (full table available at https://github.com/rpetit3/anthrax-metagenome-study/). The table includes three *B. anthracis* coverages for each *B. cereus* coverage. The *B. anthracis* simulated coverages represent the minimum *B. anthracis* coverage, the coverage at which *B. anthracis* was detectable, and the maximum *B. anthracis* coverage. The first two columns are the coverage in the artificial mixtures of *B. cereus* and *B. anthracis* genomes, respectively. The third column is the observed BCerG31 k-mer coverage. Columns 4-6 are the observed number of Ba31 k-mers, the expected number of Ba31 k-mers based on the BCerG31 coverage (see **Figure 2.4**) and the 99% prediction interval of the model. The seventh column summarizes whether the observed Ba31 is greater than the the 99% P.I. The eighth column is whether the Ba31 coverage is in the "gray zone" (< 0.18x coverage). "No" means the Ba31 exceeds the threshold (note it is possible for the Ba31 coverage to be at gray zone level but still have a positive match to a lef31k-mer). The final column shows whether KrakenHLL (Breitwieser & Salzberg, 2018) run on the sample predicted the presence of *B. anthracis*.

This table shows that false positives k-mers resulting from high BCerG coverage limit the detection of *B. anthracis* k-mers (Ba31) in mixed cultures. Below 0.006x (75x-fold *B. cereus*) and 0.01x (100x-fold *B. cereus*) *B. anthracis* genome coverages, true positive Ba31 matches cannot be differentiated from false positive matches. KrakenHLL predicted *B. anthracis* to be present even when it was not because of the background BCerG genomes coverage.

Artificial Genome Coverage		**Ba31 Count**			

B. cereus	***B. anthracis***	**BCerG31 Coverage**	**Observed**	**Model Fit**	**Model Upper 99 % P.I.[1]**	**Exceeds 99% P.I.[1]**	**lef31 Gray Zone**	**Kraken HLL**
0x	0.001x	0.00002	10	1	331	No	Yes	Yes
0x	0.003x	0.00245	346	1	332	Yes	Yes	Yes
0x	0.2x	0.123	25,396	21	352	Yes	No	Yes
1x	0x	0.593	99	102	433	No	Yes	Yes
1x	0.003x	0.610	444	104	437	Yes	Yes	Yes
1x	0.2x	0.727	25,627	125	456	Yes	No	Yes
5x	0x	3.048	487	524	855	No	Yes	Yes
5x	0.003x	3.060	919	526	857	Yes	Yes	Yes
5x	0.2x	3.155	25,502	542	874	Yes	No	Yes
10x	0x	6.115	1,050	1,051	1,382	No	Yes	Yes
10x	0.004x	6.100	1,531	1,048	1,379	Yes	Yes	Yes
10x	0.2x	6.450	26,346	1,074	1,405	Yes	No	Yes
25x	0x	15.277	2,516	2,625	2,957	No	Yes	Yes
25x	0.004x	15.174	3,075	2,608	2,939	Yes	Yes	Yes
25x	0.2x	15.339	27,536	2,636	2,967	Yes	No	Yes
50x	0x	30.381	5,058	5,221	5,552	No	Yes	Yes
50x	0.005x	30.438	5,726	5,231	5,562	Yes	Yes	Yes
50x	0.2x	30.595	29,766	5,257	5,589	Yes	No	Yes
75x	0x	45.753	7,323	4,530	8,194	No	Yes	Yes
75x	0.006x	45.699	8,351	7,853	8,184	Yes	Yes	Yes
75x	0.2x	45.859	31,971	7,881	8,212	Yes	No	Yes

100x	0x	60.926	9,633	10,470	10,801	No	Yes	Yes
100x	0.01x	60.958	11,020	10,475	10,807	Yes	Yes	Yes
100x	0.2x	61.093	33,761	10,498	10,830	Yes	No	Yes

[1] Prediction Interval

Table 2.2: Reanalysis of NYC subway metagenome sequencing

We counted Ba31, BCerG31 and lef31 k-mers in 1,458 NYC subway metagenomic samples (Afshinnekoo et al., 2015). The table is a breakdown of samples that were within the gray zone and/or had Ba31 matches that exceed the 99% prediction interval. Columns 2-8 display the same data types as columns 3-9 in **Table 2.1**. The additional *lef* column shows whether lef31 matches were identified or not. The final column provides the outcome case of the sample (**Table 2.3**). This table presents 4 samples excerpted from the complete results for all samples available at https://github.com/rpetit3/anthrax-metagenome-study/. There is one sample within the gray zone (P00738), two from the original study (P00134 and P00497) and an outlier of samples which exceed the 99% prediction interval (P00981).

		Ba31 Count							
Sample	**BCerG31 Coverage**	**Observed**	**Model Fit**	**Model Upper 99% P.I.[2]**	**Exceeds 99% P.I.[2]**	**Gray Zone**	**Kraken HLL**	**lef**	**Outcome Case**
P00134[1]	19.71	2,755	3,387	3,718	No	No	Yes	No	4
P00497[1]	4.05	953	696	1,027	No	No	Yes	No	4
P00981	1.32	20,079	226	558	Yes	No	Yes	No	3
P00738	0.002	396	1	331	Yes	Yes	Yes	No	2

[1] Samples previously identified as containing *B. anthracis*

[2] Prediction Interval

Table 2.3: Potential outcomes of *B. anthracis* detection, given matches to the Ba31 set in a shotgun metagenome dataset

This table discusses the interpretation of four cases when Ba31 k-mer matches are found in the dataset. Columns 1-3 are; lef31 match; whether Ba31 coverage is in the Gray Zone; and whether Ba31 coverage is above the 99% of the error model based on BCerG coverage.

Case	Lef31	Gray Zone	Exceeds 99% P.I.	Interpretation
1	yes	yes or no	yes or no	Evidence of lethal factor gene, could be *B. anthracis* or a *B. cereus* strain carrying the pXO1 plasmid.
2	no	yes	yes	Possible *B. anthracis* or closely related strain based on high Ba31 counts but genome coverage too low to guarantee seeing the *lef* gene. Requires more sequence coverage and/or validation by PCR or other methods.
3	no	no	yes	Ba31 matches exceed what is expected by the BCerG error model but are at a level of genome coverage at which lethal factor should have been detected. Most likely explanation is *B. anthracis* strain cured of pXO1 or unsequenced lineage closely related to *B. anthracis*.
4	no	yes or no	no	Most likely scenario is that BCerG background produced Ba31 k-mers through random errors but impossible to also rule out presence of low coverage *B. anthracis*

Chapter 3: *Staphylococcus aureus* viewed from the perspective of 40,000+ genomes

This work has been accepted for publication at PeerJ.

Abstract

Low-cost Illumina sequencing of clinically-important bacterial pathogens has generated thousands of publicly available genomic datasets. Analyzing these genomes and extracting relevant information for each pathogen and the associated clinical phenotypes requires not only resources and bioinformatic skills but organism-specific knowledge. In light of these issues, we created Staphopia, an analysis pipeline, database and Application Programming Interface, focused on *Staphylococcus aureus*, a common colonizer of humans and a major antibiotic-resistant pathogen responsible for a wide spectrum of hospital and community-associated infections.

Written in Python, Staphopia's analysis pipeline consists of submodules running open-source tools. It accepts raw FASTQ reads as an input, which undergo quality control filtration, error correction and reduction to a maximum of approximately 100x chromosome coverage. This reduction significantly reduces total runtime without detrimentally affecting the results. The pipeline performs *de novo* assembly-based and mapping-based analysis. Automated gene calling, and annotation is performed on the assembled contigs. Read-mapping is used to call variants (single nucleotide polymorphisms and insertion/deletions) against a reference *S. aureus* chromosome (Type strain, N315, ST5).

We ran the analysis pipeline on more than 43,000 *S. aureus* shotgun Illumina genome projects in the public ENA database in November 2017. We found that only a quarter of known multi-locus sequence types (STs) were represented but the top ten STs made up 70% of all genomes. MRSA (methicillin-resistant *S. aureus*) were 64% of all genomes.

Using the Staphopia database we selected 380 high quality genomes deposited with good metadata, each from a different multi-locus sequence type, as a non-redundant diversity set for studying *S. aureus* evolution. In addition to answering basic science questions, Staphopia could serve as a potential platform for rapid clinical diagnostics of *S. aureus* isolates in the future. The system could also be adapted as a template for other organism-specific databases.

Introduction

Staphylococcus aureus is a common and deadly bacterial pathogen that has been frequently investigated by whole genome sequencing over the last decade. It was the subject of arguably the first large scale bacterial genomic epidemiology study using Illumina sequencing technology (Harris et al., 2010). The cumulative number of Illumina shotgun genome projects deposited in public repositories [the National Center for Biotechnology Information Short Read Archive (NCBI SRA) and the European Nucleotide Archive (ENA)] had grown to almost 50,000 by March 2018 (**Figure 3.1**). *S. aureus* is therefore on the front edge of a cohort of bacterial species that are acquiring broad whole genome shotgun coverage, offering possibilities of new types of large scale analysis.

S. aureus is a Gram-positive bacterium with a chromosome of ~2.8 Mbp. Plasmid content varies between strains. A multi-locus sequence typing (MLST) scheme that assigns each strain a ‘sequence type’ (ST) based on seven genes has proven a robust way of describing individual strain genotypes and membership of larger ‘clonal complexes’

(CCs) (Planet et al., 2016). The accumulated public *S. aureus* genome datasets present an opportunity for investigating basic questions about how genetic variations that cause antibiotic resistance evolve within populations and how long genes traded by horizontal gene transfer persist in populations. However, there has been a problem of access, as few public tools fill the niche of providing fine scale access to very large datasets from a pathogen species. For example, PATRIC (Wattam et al., 2014) and BIGSdb (Jolley & Maiden, 2010) web based analysis sites focus on high quality annotation and complete genome MLST (cgMLST), respectively, while Aureowiki (Fuchs et al., 2017) and PanX (Ding, Baumdicker & Neher, 2018) provide very detailed information on a smaller number of strains. In this study we describe the creation of Staphopia, an integrated analysis pipeline, database and Application Programming Interface (API) to analyze *S. aureus* genomes.

Materials & Methods

Staphopia Analysis Pipeline

The Staphopia Analysis Pipeline (StAP) processed FASTQ files from a single genome through quality control steps and bioinformatic analysis software. StAP (https://github.com/staphopia/staphopia-ap/) consisted of custom Python3 scripts and open source software organized by the the Nextflow (Di Tommaso et al., 2017) (v0.28.2) workflow management platform (**Figure 3.2**). When available we used BioConda (Grüning et al., 2017) to install the open source software. Summary statistics of the original input and subsequent downstream results files were collected at each step of the pipeline. For portability, StAP was wrapped in a Docker container. The version of the

pipeline used in this work was Docker Image Tag: 112017 (https://hub.docker.com/r/rpetit3/staphopia/).

The input to StAP was either single or paired end FASTQ file (or files) generated from Illumina technology. StAP contained an option that allowed FASTQ data to be pulled from the ENA based on the experiment accession number (ena-dl v0.1, https://github.com/rpetit3/ena-dl). A MD5 hash (md5sum) was generated from the input FASTQ data and cross-referenced against a list generated from processed genomes to prevent reanalysis of the same input. BBduk (Bushnell, 2016) (v37.66) was used to filter out adapters associated with Illumina sequencing and trim reads based on quality. Read errors were corrected using SPAdes (Bankevich et al., 2012) (v3.11.1). Based on the corrected reads, low quality reads were filtered out and the total dataset was subsampled to a maximum of 281 Mbases (100x coverage of the N315 reference chromosome (Kuroda et al., 2001)) with illumina-cleanup (v0.3, https://github.com/rpetit3/illumina-cleanup/). This file (or files, if paired end) we termed “processed FASTQ” or “pFASTQ”.

pFASTQ reads were assembled *de novo* using SPAdes (Bankevich et al., 2012) (v3.11.1). SPAdes also marked assembles as putative plasmids based on evidence such as relative read coverage (Antipov et al., 2016a). Summary statistics of the assembly are created using the assembly-summary script (https://github.com/rpetit3/assembly-summary). A BLAST+ nucleotide database was created from the assembled contigs to be used subsequently for sequence query matching. Open reading frames and their putative

functions were predicted and annotated using Prokka (Seemann, 2014) (v1.12) and its default database.

The *S. aureus* type strain N315 (Kuroda et al., 2001) chromosome (ST5 MRSA; accession NC_002745.2; length 2,814,816 bp) was used as a reference for calling consensus SNPs and indels in the pFASTQ reads using the GATK (McKenna et al., 2010) (v3.8.0) pipeline. GATK pipeline also incorporated BWA (Li & Durbin, 2009a) (v0.7.17), SamTools (Li et al., 2009a) (v1.6) and PicardTools (v2.14.1, http://broadinstitute.github.io/picard/) software. Identified variants were annotated using the vcf-annotator script (v0.4, https://github.com/rpetit3/vcf-annotator). Jellyfish (Marçais & Kingsford, 2011b) (v2.2.6) was used to count k-mers of length 31 base pairs (31-mers) in the pFASTQ file. We used Ariba (Hunt et al., 2017) (v2.10.2) to make antibiotic resistance and virulence predictions for paired-end reads only. Resistance phenotypes were predicted using the MegaRes reference database (Lakin et al., 2017b) and virulence using the Virulence Factor Database (Chen et al., 2016b) core dataset. Single-end reads

MLST was determined by two or three methods depending on the whether the pFASTQ was paired end. All methods used the *S. aureus* MLST allele sequence database downloaded from https://pubmlst.org/saureus/ (November 2017). Alleles for each of the seven loci were aligned against the assembled genome using BLAST+ (Camacho et al., 2009) (v2.7.1+). Alleles and sequence type (ST) were determined based on perfect matches (100% nucleotide identity with no indels). We also used the MentaLiST (Feijao et al., 2018) (v0.1.3) software to call MLST based on k-mer matching of the alleles to the

pFASTQ file. Unlike the BLAST+-based MLST method, MentaLiST did not require exact matches to alleles to predict a ST. If the pFASTQ was paired-end, Ariba (Hunt et al., 2017) (v2.10.2) also determined MLST alleles and ST. The default ST call for each genome was determined in the following order: agreement between each method, agreement between MentaLiST and Ariba, agreement between MentaLiST and BLAST+, agreement between Ariba and BLAST+, Ariba alone without a novel or uncertainty call, MentaLiST alone, and finally BLAST+ alone. Core genome MLST (cgMLST) was determined with MentaLiST using the *S. aureus* cgMLST scheme (Leopold et al., 2014) available at at http://www.cgmlst.org.

Evidence for SCC*mec* predictions were based on multiple approaches. The primary approach aligned SCCmec typing primers, downloaded from http://www.staphylococcus.net/, against the assembled genome using BLAST+ (v.2.7.1+, (Camacho et al., 2009)). Samples with a perfect match to at primer pairs for a given amplicon were assigned an SCCmec type following the Kondo et. al. algorithm (Kondo et al., 2007). Genes and proteins associated with SCCmec mere aligned against the assembled genome using BLASTN and TBLASTN. We also mapped the pFASTQ to each SCCmec cassette using BWA (Li & Durbin, 2009a) (v0.7.17). The overall cassette and *mec* region coverage statistics were determined as well as the per-base coverage determined for each cassette using genomeCoverageBed (Quinlan & Hall, 2010) (v2.26.0). The methods described above were based on the 11 SCC*mec* types currently listed in the http://www.sccmec.org (I - XI) and hence did not include recently described types XII and XIII (Wu et al., 2015; Kaya et al., 2018). We labelled a genome as “MRSA” only if each *mecA* typing primer (Kondo et al., 2007) had a perfect BLASTN

match on the *de novo* assembly, a predicted *mecA* gene ortholog had a BLAST score ratio of at least 95%, or Ariba (Hunt et al., 2017), as previously described, predicted reads in the paired-end pFASTQ file matching a *mecA* target.

Web Application, Relational Database and Application Programming Interface

We used Django (v2.0), a Python web framework, to develop a PostgreSQL (v10.1) backed relational database for storing the results from the analysis pipeline (**Figure 3.3**). A Django application was created for each module of the pipeline, automating the creation of database tables for the results. Python scripts building off Django were developed for insertion of results from each StAP module or the StAP as a whole. A web front-end was developed (staphopia.emory.edu) using the Bootstrap (v4.0) and jQuery (v3.2.1) web frameworks. We used the Django REST framework to develop an extensive application programming interface (API) that allowed users to create queries accessing multiple samples. We also developed an R package, Staphopia-R (https://github.com/staphopia/staphopia-r), to programmatically access the API. The API and its endpoints were documented to allow users to further develop their own packages in a language of their choice. The source code for our web application was made available at https://github.com/staphopia/staphopia-web/.

Processing Public Data

We used the Cancer Genomics Cloud (CGC) Platform, powered by Seven Bridges (http://www.cancergenomicscloud.org/), to process *S. aureus* genomes through StAP in November 2017. CGC allows users to create custom workflows based on Docker

containers, then execute these workflows on the Amazon Web Services (AWS) cloud platform. We obtained a list of publicly available *S. aureus* sequencing projects from the ENA web API using the following search term:

> *"tax_tree(1280) AND library_source=GENOMIC AND (library_strategy=OTHER OR library_strategy=WGS OR library_strategy=WGA) AND (library_selection=MNase OR library_selection=RANDOM OR library_selection=unspecified OR library_selection="size fractionation")".*

We only processed projects which used Illumina sequencing technology. CGC opened AWS r3.xlarge instances (30.5GB RAM, 4 processors) that downloaded FASTQ files from the ENA using ena-dl for each genome and ran the StAP pipeline. Results files were returned to the CGC, then uploaded into the Staphopia database server.

Metadata Collection

We used the ENA API to download and store any information linked to the 'Experiment', 'Study', 'Run' and 'BioSample' accessions into the database for each genome. We also determined each sample's publication status using three approaches.

The first approach identified existing links between SRA, a mirror of ENA, and PubMed using NCBI's Entrez Programming Utilities web API(*Entrez Programming Utilities Help*, 2010). For any links identified, we used the corresponding PubMed ID to extract information corresponding to the publication and stored them in the database.

The second approach searched for accessions within the text of scientific articles. We searched PubMed using the term, "*Staphylococcus aureus*", limited to the years between and including 2010 (the date of the first publicly available Illumina data upload), and 2017. The saved results, stored as XML, were then loaded into Paperpile, a subscription-based reference management tool, and the corresponding main-text PDFs were automatically downloaded. This process did not include supplementary information files, which required a manual operation. For those articles in which a PDF could not be automatically downloaded, attempts to manually acquire the PDF were made. Using the text search program 'mdfind', available on Apple OS X, each accession (BioSample, Experiment, Study and Run) in the Staphopia database was used as a separate query to search all the PDF files. Experiment accessions with a corresponding PubMed ID were then stored in the database. In cases where a Study, BioSample or Run accession was identified in PDF text, each associated Experiment accession was linked to the corresponding PubMed ID.

In the third a collection of PubMed articles with primary descriptions of *S. aureus* genome sequencing studies was manually curated. For these studies, the PDF and all available supplementary information were downloaded. The process of text-mining the articles and linking Experiment information to PubMed ID was repeated as described in the second approach.

Creating non-redundant *S. aureus* diversity set

Using available metadata, we selected a non-redundant diversity (NRD) set of genomes that were gold quality (please see results section), linked to a publication and each had a unique ST. When more than one strain from a ST was available, we randomly selected one individual giving priority to samples with collection date, site of isolation and location of isolation fields filled.

Using predicted variants against N315, we extracted a list of genes that had complete sequence coverage (ie “core” genes) but no predicted indels. We extracted the reference gene sequence and created an alternative gene sequence with SNPs predicted in each sample. The alternative gene sequences were split into 31-mers. Presence on these 31-mers in the pFASTQ file were cross-validated using the Jellyfish (Marçais & Kingsford, 2011b) tool. These reconstructed gene sequences or all genomes were stored in the database and made available through the API for rapid phylogenetic comparisons.

A set of 31-mer validated genes (please see results section) in which no more than 3 samples contained unvalidated 31-mers were selected for phylogenetic analysis. The set of validated genes were extracted and concatenated into a single sequence for each sample and saved in multi-FASTA and PHYLIP formats. A guide tree was generated with IQ-Tree (Nguyen et al., 2015b) (v8.2.11, -fast option) for identification recombination events with ClonalFrameML (Didelot & Wilson, 2015)(v1.11). A recombination free alignment was created with maskrc-svg (https://github.com/kwongj/maskrc-svg). We used IQ-Tree to generate the final maximum likelihood tree with the GTR model and bootstrap support. Bootstrap support

was generated from 1000 UFBoot2 (Hoang et al., 2018) (ultrafast bootstrap) replicates. We annotated the tree using iTOL (Letunic & Bork, 2016).

Results

Design of the Staphopia Analysis Pipeline and processing 43,000+ genomes

The Staphopia analysis pipeline (StAP; **Figure 3.2**) was written to automate processing of individual *S. aureus* genomes from Illumina shotgun data. The pipeline was designed as a series of modules running individual software packages, organized by the Nextflow (Di Tommaso et al., 2017) workflow language, which made it possible to run the entire pipeline or individual components as needed. The first step of the pipeline was to import single- or paired-end FASTQ files either as local files, or from the ENA database. We selected ENA over SRA due to ENA offering direct FASTQ downloads. Following quality-based trimming and down selection of the FASTQ to 281 Mbases (~100x coverage of the N315 reference chromosome (Kuroda et al., 2001), NC_002745.2), analyses were run on the raw processed FASTQ (pFASTQ) files directly, or on *de novo* genome assemblies constructed by the SPAdes program (see Methods for more details). We decided to down sample the input FASTQ files for two reasons: to manage the computational burden when running thousands of genome projects and also to achieve genome datasets with consistently sized pFASTQ input files. The threshold of ~100x coverage was chosen after preliminary studies showed that there was either small or no improvements in outcome for downstream assembly and remapping steps for input files > 100x but large increases in processing time and memory requirement. We created a

Postgres database to store results from the StAP analysis and a web front end and a web API for mining the data. An R package (Staphopia-R) was written for interacting with the API and was used for most analysis presented in the results.

In November 2017 there were 44,012 publicly-available shotgun sequencing projects with FASTQ files in ENA. Illumina technology was the dominant platform, accounting for 99% of samples (N=43,972). Eighty-one percent (N=35,580) of them had at least 281 Mbases sequence data. We processed all Illumina genomes in parallel through the StAP using cloud servers (please see Methods section). On parallel r3.xlarge instances with 30.5 Gb RAM and 4 processors, the mean time to process a genome was 52 minutes with an interquartile range of 47 to 56 minutes (**Figure 3.4**).

Sequence and assembly quality trends

We identified samples that were likely mixed-samples or not *S. aureus* whole genome shotgun projects and/or were of low technical quality and marked them to not be included in subsequent analysis. We removed genomes that failed to match to any known allele of the seven MLST loci (323 genomes), had a total assembly size that differed by more than 1Mb from a typical *S. aureus* chromosome (<1.8Mb or >3.8Mb; 764 genomes), or had a GC content differing more than 5% (<28% or > 38%; 467 genomes) of the expected 33% GC content. Failure to complete the StAP pipeline due to poor data quality, and coverages less than 20x were flagged in 101 and 142 genomes, respectively. In total, we removed 1,023 genome projects, leaving 42,949 for further analysis.

We placed genomes into an arbitrary ranking of 1-3 (“Bronze”, “Silver” and “Gold”) based on the pFASTQ coverage and average sequencing quality. Paired-end genomes that had read lengths exceeding 100bp, a coverage of 100x and an average per base quality score of at least 30 were given a Gold rank. The purpose of the Gold rank was to group together high-quality samples with near-identical coverage. Paired-end genomes with similar read length and quality cutoffs but a lower sequence coverage (between 50x and 100x) were classified as Silver. The remaining samples were given a rank of Bronze. Single-end reads were classified as Bronze no matter the read length, quality or coverage. More than 70% of the samples were of rank Gold (N=31,014). There were 5,931 Silver and 6,004 Bronze rank samples. Each year since 2012, the number of Gold ranked genomes have exceeded Silver and Bronze (**Figure 3.5**).

Changes in sequence quality and *de novo* genome assembly metrics over time reflected the development of Illumina technology. Mean per based quality scores increased from ~ 32 in 2010 to > 35 in 2012 and have stayed at that level since. The mean sequence read length rose in steps from < 50 in 2010 to ~ 150 bp in 2017. Assembly metrics such as N50 (Earl et al., 2011), and mean and maximum contig length have gradually increased since 2010. Bronze ranked genome projects had similar (or sometimes even higher) mean per read quality scores than Gold and Silver since 2011. However, Silver and Gold assembly metrics such as N50 and mean contig size were generally quite similar and higher than Bronze.

Genetic diversity measured by MLST

We obtained a view into the genetic diversity of the sequenced *S. aureus* genomes by *in silico* MLST using Ariba (Hunt et al., 2017), MentaLiST (Feijao et al., 2018) (both taking pFASTQ as input, but using different algorithms) and BLASTN against assembled contigs. A sequence type (ST) was assigned to 42,337 (98.6%) genomes. Of these, 41,226 (97.3%) calls were in agreement between MentaLiST, BLAST+ and (if paired-end) Ariba methods; 828 had agreement between two methods and a no-call on the other, and 189 were supported by one program with no-calls from the other two. Of the remaining 612 genomes not assigned to a known ST, 306 were predicted to be in a novel ST based on matches to known alleles of each of the 7 loci. The remaining 306 genomes had 1-6 known *S. aureus* MLST alleles.

The 42,337 genomes assigned to existing STs represented only 1,090 STs of 4,466 in the saureus.mlst.net database (November 2017). The abundance distribution was weighted toward common strains, with the top ten sequence types (STs 22, 8, 5, 239, 398, 30, 45, 15, 36, and 105) representing 70% (N=29,851) of the genomes (**Figure 3.6**).

The cgMLST (core genome MLST) set of 1861 loci (November 2017) were assigned to the genome set using MentaLiST. There were 38,677 distinct patterns, with only 1,850 patterns found in more than one sample, the remaining 36,827 patterns were represented by a single genome.

Antibiotic resistance genes

Treatment of *S. aureus* infections has been complicated by the evolution of strains

resistant to many commonly used antibiotics (Foster, 2017). In particular, methicillin-resistant *S. aureus* (MRSA), carrying the *mecA* gene encoding the PBP2a protein that confers resistance to beta-lactam antibiotics, has become a global problem. We designated a genome as MRSA if each *mecA* typing primer (Kondo et al., 2007) had a perfect BLASTN match on the *de novo* assemblies (26,743 strains), a predicted *mecA* gene ortholog had a BLASTN score ratio of at least 95% (26,430 strains), or the Ariba (Hunt et al., 2017) algorithm predicted reads in the paired-end pFASTQ file matching a *mecA* target in the MegaRes (Lakin et al., 2017b) database (27,120 strains). The number of genomes having at least one of these criteria (27,628) was 64% of the total number. Of these, 95% (26,340) of the samples had agreement between each of the criteria. The top five most common STs had a large portion of MRSA strains (**Figure 3.6**), which reflects the selection bias of the research community in investigating these significant hospital and community pathogen strains over other *S. aureus*.

The *mecA* gene is usually horizontally acquired as part of a mobile genetic element called “Staphylococcal Cassette Chromosome mec” (SCC*mec*) (Katayama, Ito & Hiramatsu, 2000). SCCmec elements have been classified into at least eleven classes that vary in composition of *mec* genes, *ccr* cassette recombinase genes and spacer regions (http://www.sccmec.org). Knowledge of the SCCmec type can be useful for high-level characterization of MRSA strain types (Kaya et al., 2018). We showed that ten of the eleven cassettes in the current schema map to at least one genome with highest coverage (an approximate method for assigning SCCmec type) (**Table 3.1**). Of the 26,462 (26,185 paired-end) genomes with at least 50% cassette coverage, 96%, 96% and 99% are MRSA based on primer BLASTN, protein BLASTN or MegaRes, respectively.

All type XI cassettes were *mecA* negative by primer BLASTN because these contained the *mecC* allele (García-Álvarez et al., 2011; Shore et al., 2011), which was sufficiently different to be outside the normal distance for a positive match. We found 53 genomes which matched to at least 50% of a SCCmec cassette but were not MRSA and had no reads mapping to the *mec* region of the cassette.

In addition to *mecA*, we found numerous other classes of non-core genes using the MegaRes (Lakin et al., 2017b) class designations (**Table 3.2**). We did not consider SNPs/indels in core genes associated with resistance for this analysis. The most common class of resistance genes were beta-lactamases found in 37,758 genomes. Following this, the most common were the genes putatively conferring fosfomycin, macrolide-lincosamide-streptogramin (MLS), and aminoglycosides resistance (24,205, 22,322, 17,968 genomes respectively). As with MRSA, the other common resistance genes were not distributed evenly among the top ST groups (**Figure 3.6**), reflecting sampling ascertainment bias and also possibly differences in geographic distribution and prevalence of healthcare-isolated strains in the most common genotypes.

Publication, metadata and strain geographic distribution

One challenge to using publicly available datasets through ENA or SRA is determining whether there is a published article describing the sequenced genome. We found through NCBI's Entrez Tools (eLink) that 6,712 genomes were linked to 48 publications in PubMed (March 2018). We attempted to add to the number by using text-mining methods to find *S. aureus* accession numbers in PDFs of *S. aureus* genome publications, ascertaining an additional 5,209 genomes in 30 publications. Therefore, of the 42,949

samples deposited between 2010 and 2017, only 28% (N=11,921) could be linked to a publication (**Figure 3.1**). Since many genomes have been deposited in the last 1-3 years, this reflected the often-significant time lag between depositing sequence data and final publication.

We noted that collection of metadata from public sequencing projects was another challenge. When submitting genome sequences to databases only a limited number of metadata fields are required, leading to the bulk of the information needing to be extracted manually from a publication, if it can be found. Only 40% (N=17,034) genomes had a collection date, 35% (N=14,983) had a geographic location and 35% (N=14,768) had isolate source metadata. Using the available geographic data to geocode the sites of collection, we found that strains were from five continents and at least 40 countries. There was a strong bias toward strains from Europe (N=7,314) and North America (N=5,882), reflecting where the funding for most of the early sequencing studies had originated.

A non-redundant *S. aureus* diversity set

The number of SNPs compared to the N315 reference strain varied from 6 to 141,893 within our collection of 42,949 genomes. The stepped pattern of the distribution (**Figure 3.7**) reflected the organization of *S. aureus* into clonal complexes. Apart from CC5 strains closely related to N315, the majority of *S. aureus* had ~50-50,000 SNPs and ~500-1500 indels called by the GATK pipeline (McKenna et al., 2010). There were a group of 240 most distant strains with > 55,000 SNP (**Figure 3.7**) that were found to

be closer to the sister species, *S. argenteus* (Holt et al., 2011) based on ANI imputed by mash (Ondov et al., 2016), although 230 of these were assigned a *S. aureus* ST.

Of the 6,904 *S. aureus* genomes of Gold rank linked to a publication we selected a group of 380 each having a distinct ST as a non-redundant diversity (NRD) set of genomes. Of the 2,756 annotated N315 genes (excluding RNAs), 1,113 genes had no indels when reads from each genome in the NRD dataset were mapped. Of these, 878 were "core" genes found in every genome. We reconstructed these genes for each of the NRD genomes starting with the N315 sequence and substituting predicted SNPs. These predicted sequences were then validated by decomposing into 31-mers and cross-checking whether each k-mer was present in pFASTQ files processed by Jellyfish (Marçais & Kingsford, 2011b). We concatenated the 878 genes for each member of the NRD set and created a tree based on the 44,377 variant SNP positions (**Figure 3.8**). The structure of the unrooted species tree resembles previous *S. aureus* phylogenies (Planet et al., 2016).

Discussion

The huge public library of genome sequence projects of *S. aureus* and other pathogens are a resource for microbiologists for testing genetic hypotheses in silico. Unfortunately, this has been a library of blank covers: most projects cannot be browsed to identify features such as ST, key SNPs and non-core genes. Staphopia makes the library searchable for a number of important attributes, and we have described example workflows in the results section.

We used three strategies for analysis of raw sequence data: mapping reads to a reference chromosome to identify variants; *de novo* genome assembly, and direct analysis of the reads. Each has its strengths and weaknesses. Reference mapping retains quality information about variant calls but is limited to regions of the core genome and accuracy is reduced as genetic distance increases between the query and the reference. *De novo* assembly allows for discovery of novel accessory genes and is reference independent but could be affected by genomic contamination and with Illumina short read data, and small portions of the sequence could be lost in gaps between contigs. Direct analysis of reads based on k-mer decomposition approaches allows examination of sequence independent of mapping and assembly algorithms but are susceptible to false results arising from contamination and random sequence error. Using different approaches to cross-validate wherever possible builds confidence and we showed that MLST and MRSA/MSSA identification were robust with different underlying data types collected.

There are many possible avenues for future extensions of the project. New tools for efficient direct querying of raw reads have recently become available (e.g BigSI (Bradley et al., 2017), and mash (Ondov et al., 2016)) and we plan to incorporate them in future iterations of the pipeline. Some of the principal improvements need to be in protein functional annotation. For speed and simplicity, we elected to map genes called from *de novo* assemblies against the included Prokka (Seemann, 2014) RefSeq database. This has the advantage of giving consistent proteins naming that can be linked to many functional annotation databases through UniProt cross-references. However, for fine resolution studies of sets of genomes from Staphopia, we recommend reprocessing with Roary (Page et al., 2015a) to incorporate paralog detection and to use more extensive

databases for homology matching. Even then, specific modules would need to be incorporated to improve naming of intrinsically hard to annotate protein families (e.g MSCRAMMs (microbial surface components recognizing adhesive matrix molecules) (Foster et al., 2014)).

A key problem highlighted in this study is the difficulty in tracing publications linked to public genome data and finding typical metadata on strains (date and place of isolation, body site). We were able here to link thousands of records to publications through searching text in PDFs. For this reason, we urge researchers publishing microbial genomes to quote the Project ID (i.e. the PRJN ID) of publicly submitted data in the full text of the publication. Extracting metadata from publications presented a more complicated process. Metadata is often available as spreadsheets, documents or PDFs which are not easily parsed. We believe that journals need to start to enforce machine readable standards for metadata associated with deposited strains. The routine usage of BioSample ID (https://www.ncbi.nlm.nih.gov/books/NBK169436/), which links strains to genomic information, would be a major step forward.

Staphopia was designed with Illumina shotgun data in mind but increased use of alternative sequencing technologies in the future may necessitate new development. “Long read” technologies (e.g. PacBio, Oxford Nanopore) tend to have assemblies with fewer gaps, higher per base errors and lower coverage. A “gold standard” PacBio assembly will have a different quality profile to Illumina technology data (which itself is also evolving). Another challenge for automated assembly of public data will be to identify projects sequenced with multiple technologies and assembled as hybrids (e.g. as

demonstrated by the Unicycler tool (Wick et al., 2017b)). To do this would mean altering the pipeline to perform hybrid assembly when experiments with multiple technologies are associated with a strain. Currently, within ENA (and SRA) a BioSample can be associated with multiple Experiments, but an Experiment can only be associated with a single BioSample. When a BioSample was linked to more than one Experiment, it was difficult to determine in an automated way if it is actually the same genomic DNA input to multiple experiments or, in rare cases, a mistaken assignment of a set of genetically non-identical isolates with the BioSample (e.g. all isolates from a study given the common strain name “USA300”). Because of this, Staphopia treated each ENA Experiment as a unique sample, rather than the BioSample.

It is unclear at this time whether the approach of processing of every public dataset will be sustainable as sequencing data production grows in the future. It would only be possible if storage and processing costs fall faster than the accumulation of new data, and multi-genome database queries may still be prohibitively slow. An alternative strategy to processing all strains, would be to filter the isolates for redundancy, by removing isolates that are less than *n* SNPs from any member of a canonical genome set. However, there is still information in deep sequencing studies that can be captured from distributions of reads and kmer distribution, even if the consensus sequences of the strains are identical. Plasmid copy number may differ between clones grown under different conditions and the distribution of reads across the genome can itself be used to infer relative growth rate (Brown et al., 2016). No two shotgun genome sequencing projects are identical, and all have some potential value, especially if they have strong supporting metadata.

Conclusions

- We analyzed 43,972 *S. aureus* public Illumina genome projects using the newly developed "Staphopia" analysis pipeline and database. 42,949 genomes were retained for subsequent analysis after filtering against low quality
- The data quality was high overall: 36,945 (86%) were from paired end projects with greater than 50-fold coverage and 30 average base quality ("Gold" and "Silver" quality)
- There has been a great concentration of effort on a sequencing a small number of sequence types: only 1,090 STs of 4,466 previously collected STs were recovered and 10 STs make up 70% of all genomes.
- 26,340 to 27,628 genomes were predicted MRSA depending on the criteria used for classification.
- We could link only 28% of the genomes to a PubMed referenced publication.
- We identified 380 non-redundant highly quality published genomes as a reference subset for diversity within the species.
- We identified 878 core genes that can be reliably used for rapid tree building based on SNPs compared to the reference N315 genome.

Funding

Funding was from Emory University, Amazon AWS in Education Grant Program, and NIH grants AI091827 and AI121860. The Seven Bridges NCI Cancer Genomics Cloud pilot was supported in part by the funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No.

HHSN261201400008C.

Acknowledgements

We would like to thank Tauqeer Alam, Jim Hogan, Santiago Castillo-Ramírez. Michelle Su, Michael Frisch and Erik Lehnert for their helpful suggestions. We would also like to acknowledge our gratitude to the many scientists and their funders who provided genome sequences to the public domain, ENA and SRA for storing and organizing the data, and the authors of the open source software tools and databases used in this work.

Links

Code for most analysis described in the results section -

https://github.com/staphopia/staphopia-paper

Staphopia - https://staphopia.emory.edu

R Package - https://github.com/staphopia/staphopia-r

StAP - https://github.com/staphopia/staphopia-ap

Web Package - https://github.com/staphopia/staphopia-web

Docker Image - https://hub.docker.com/r/rpetit3/staphopia/

NRD Dataset - https://doi.org/10.6084/m9.figshare.6263435

Appendix

The following appendix contains boxes, tables and figures referenced in the text of this chapter.

Figure 3.1: Cumulative submissions of *Staphylococcus aureus* genome projects 2010 - 2017 linked to publications.

There were 42,949 *S. aureus* genome projects investigated in this study. Of these samples, we have linked 11,921 to a publication.

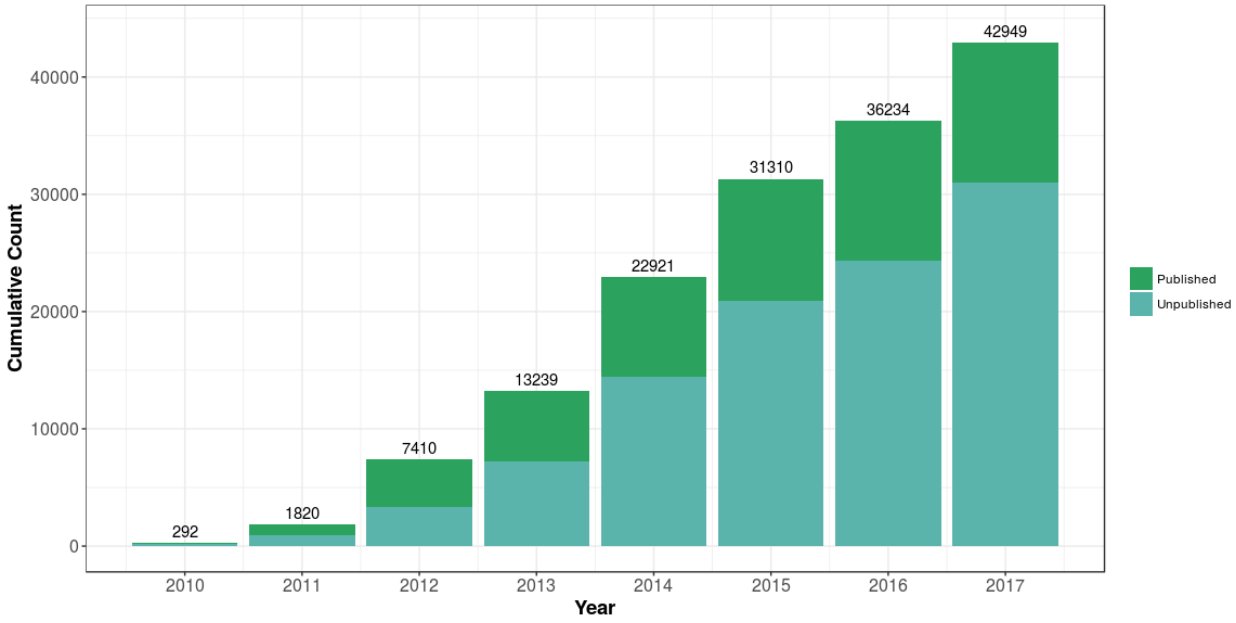


Figure 3.2: Staphopia Analysis Pipeline (StAP) Workflow

The diagram describes basic operations of the pipeline on a single genome input (FASTQ file) before uploading into the Postgres relational database. Details of the programs used are in the methods and https://github.com/staphopia/staphopia-ap. Green arrows indicate input from *de novo* assembled contigs, blue arrows were operations performed on pFASTQ files.

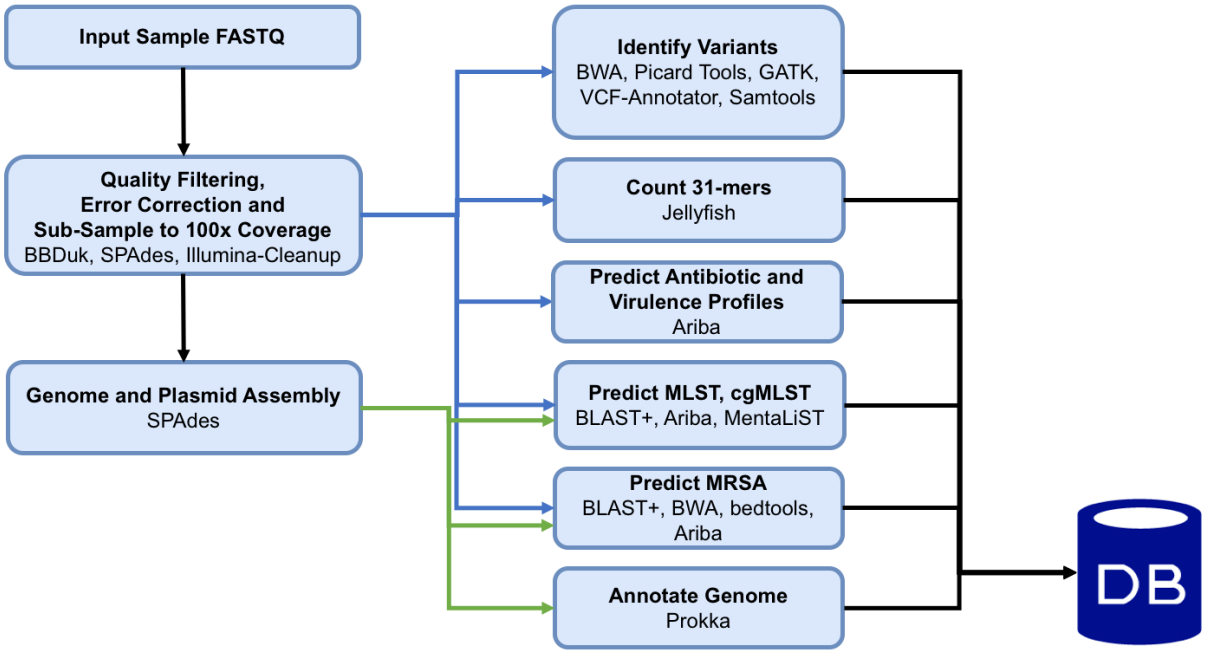


Figure 3.3: An overview of the Staphopia platform.

This figure provides an overview of Staphopia. Samples are processed by the analysis pipeline and stored in the database. Metadata collected from linked publications are also stored in the database. This information is then made available through a web front-end or a web application programming interface (API). An R package has also been developed to programmatically access the web API.

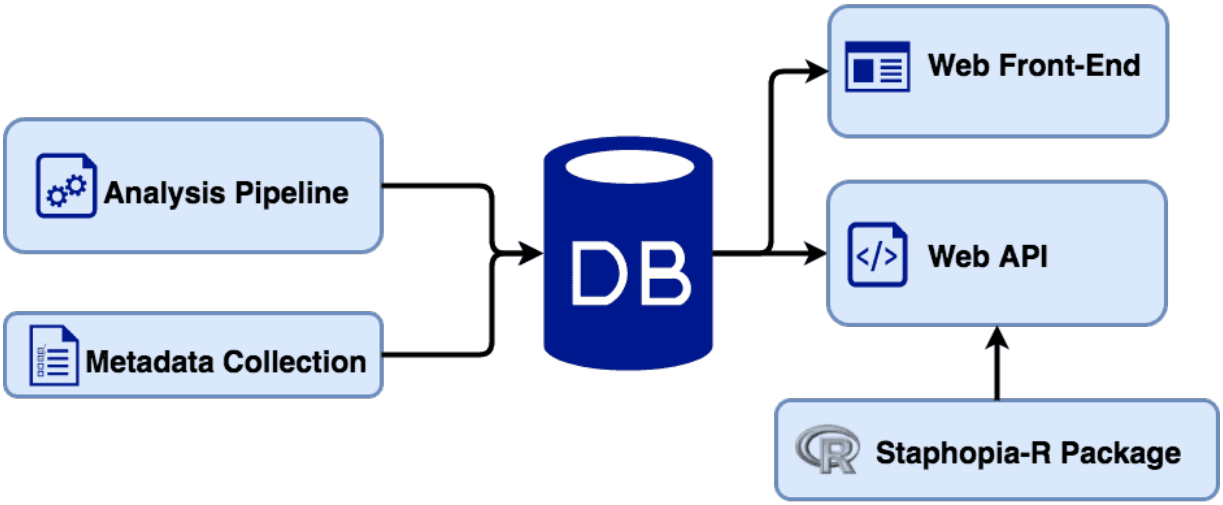


Figure 3.4: StAP run time using Cancer Genomics Cloud (CGC) platform.

Overall run time statistics were available for 31,587 of the completed CGC jobs. Mean run time was 51 minutes (median 52 minutes). There were 983 jobs that took more than 80 minutes to complete.

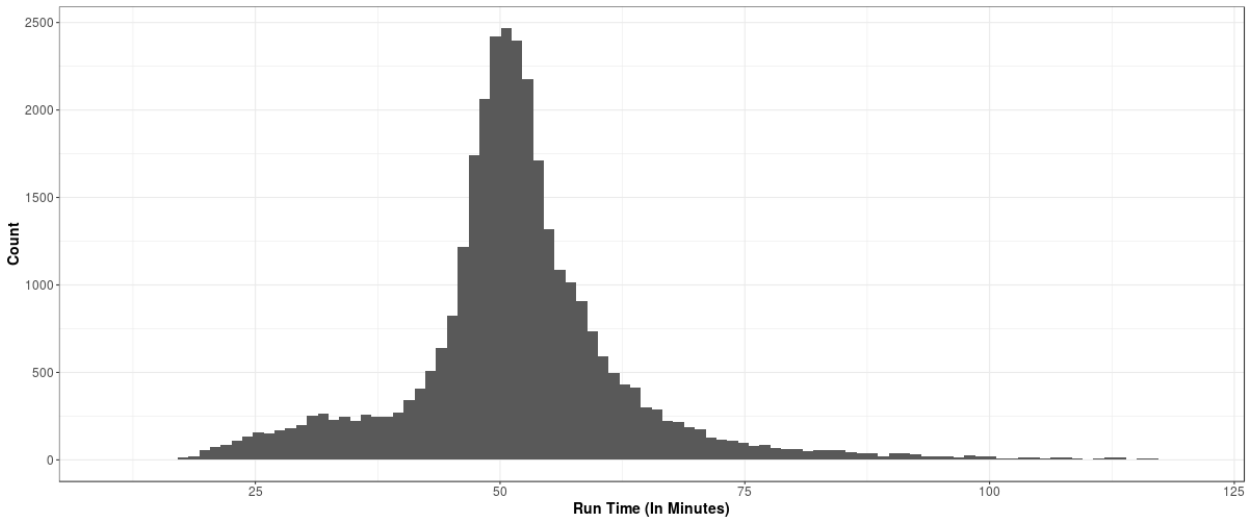


Figure 3.5: Sequencing quality ranks per year 2010-2017.

Genome projects were grouped into three increasing quality ranks: Bronze, Silver and Gold. The rank was based on coverage, read length and per-read quality (please see results section). The highest rank, Gold, represented 72% (N=31,014) of the available *S. aureus* genome projects. The remaining genomes were almost evenly split between Silver (N=5,931) and Bronze (N=6,004).

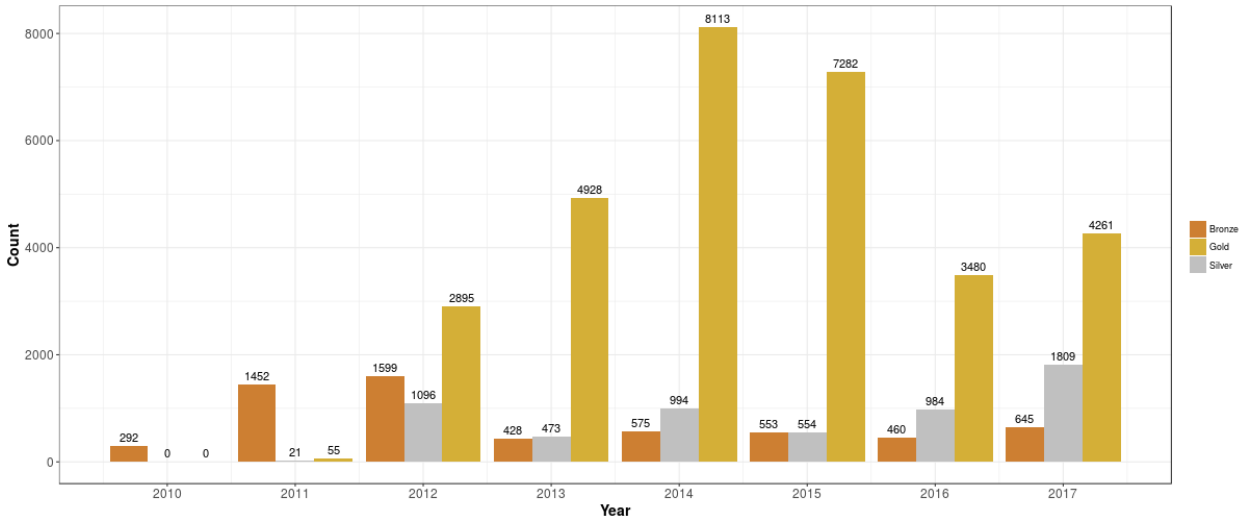


Figure 3.6: Resistance genes to methicillin (MRSA), aminoglycoside, fosfomycin, and macrolide-lincosamide-streptogramin (MLS) antibiotic in the top 10 STs.

The presence of resistance genes was predicted by Ariba (Hunt et al., 2017) using the reference MegaRes (Lakin et al., 2017b) database. The distribution MegaRes resistance classes for methicillin (A), aminoglycosides (B), fosfomycin(C) and macrolide-lincosamide-streptogramin (D) are presented for the top 10 sequence types (ST). The top 10 STs represent 70% (N=29,851) of the genomes analyzed in this study.

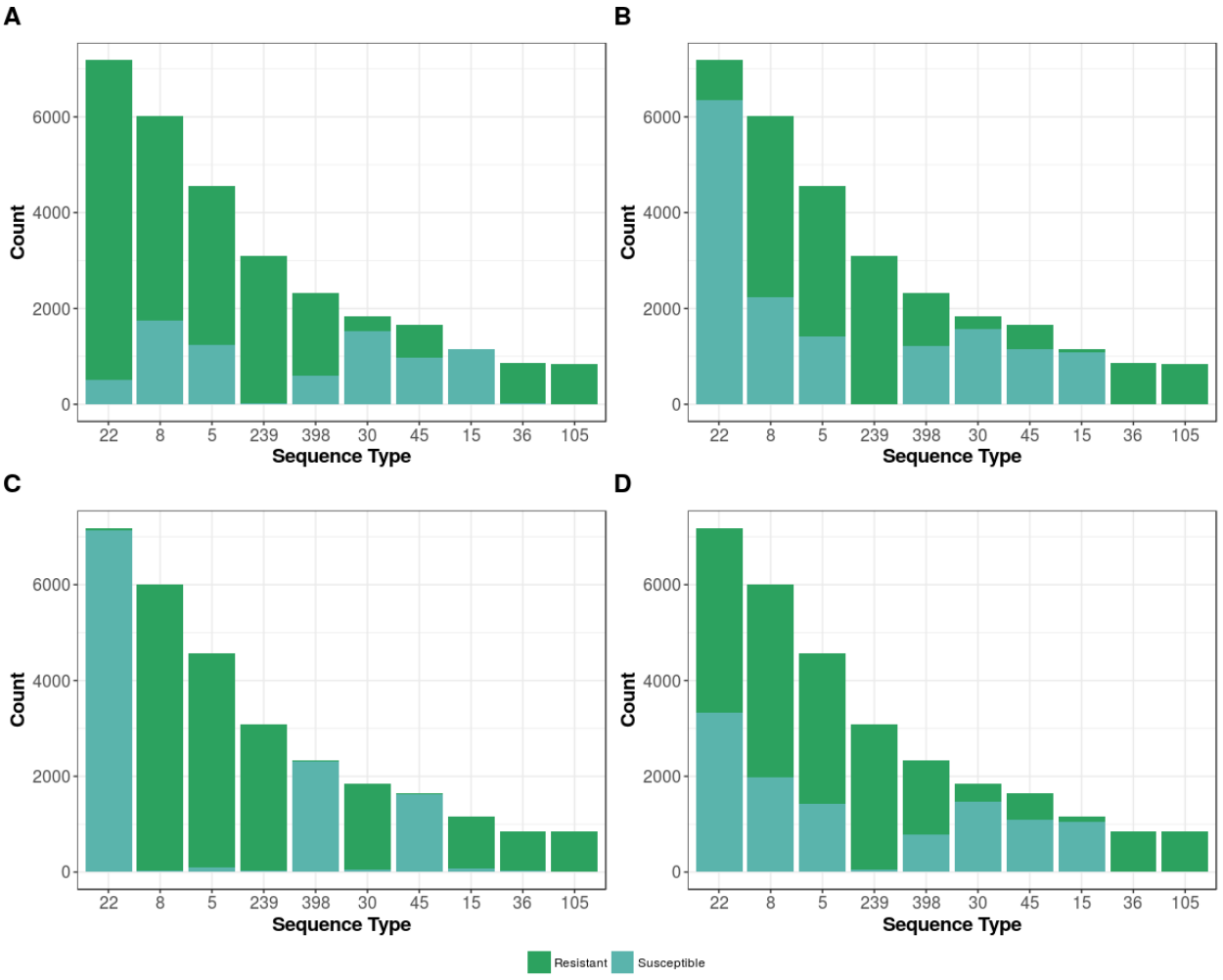


Figure 3.7: *S. aureus* SNP distance from reference *S. aureus* N315.

For each genome, the number of SNPs found by mapping reads to the N315 reference using GATK (McKenna et al., 2010) was plotted, with genomes ordered from least to most SNPs. 240 genomes with > 55,000 SNPs (dotted line) that had best matches to *S. argenteus* using mash (Ondov et al., 2016) were indicated by silver bars, the rest were *S. aureus* (gold).

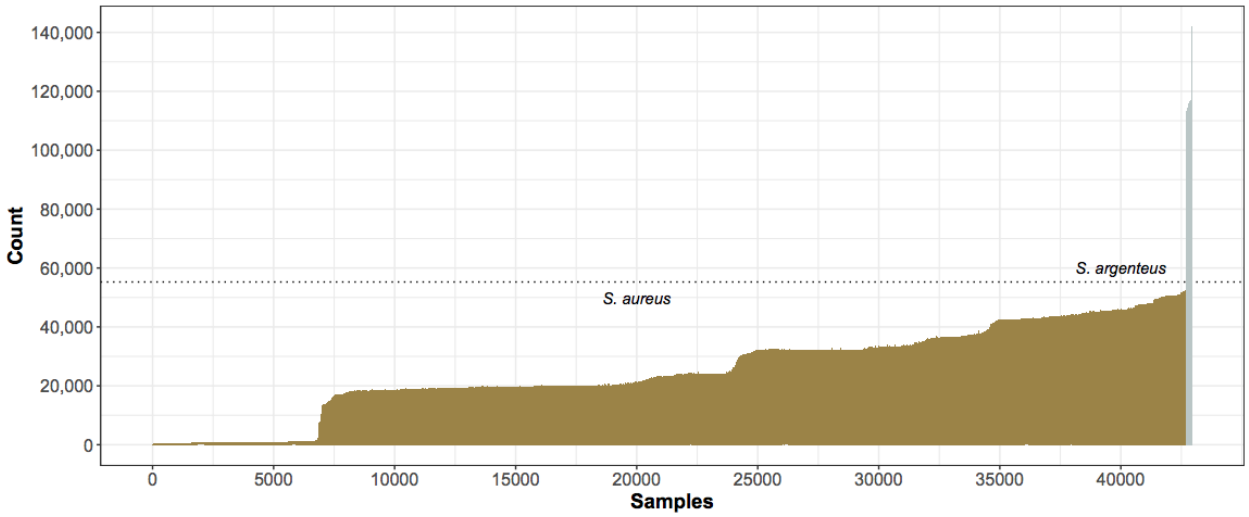


Figure 3.8: Unrooted phylogeny of the *S. aureus* Non-Redundant Diversity (NRD) dataset.

An unrooted phylogenetic representation of the 380 genome non-redundant set (one representative per ST, all published and gold rank) using IQ-Tree (Nguyen et al., 2015b). The putatively recombinant positions predicted using ClonalFrameML (Didelot & Wilson, 2015) were removed from the alignment. Clonal complexes containing the top ten most common STs are indicated with colored circles. The tree was built from 878 reconstructed core genes (please see Methods section) with 44,377 sites. Branches supported with probability > 0.9 are marked by red dots. The likelihood score for the tree was -1,890,510.

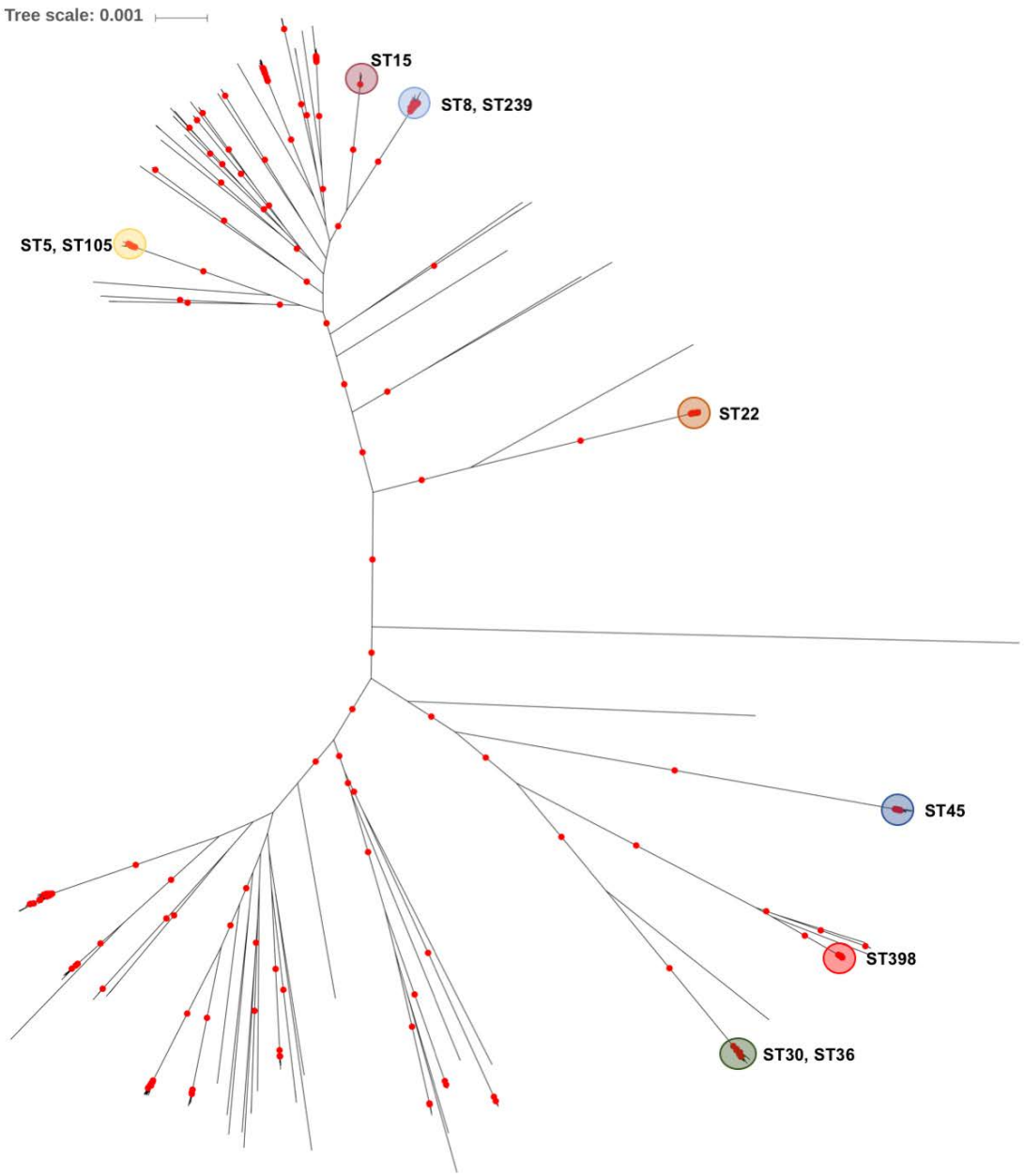

Tree scale: 0.001
ST15
ST8, ST239
ST5, ST105
ST22
ST45
ST398
ST30, ST36

Table 3.1: Predicted SCCmec cassette type representation.

There were 26,462 samples with reads mapped to at least 50% of a SCCmec cassette. The table is a breakdown of the SCCmec cassettes with the highest percent match for each sample.

SCCmec Type	**Count**
I	689
II	5,183
III	2,807
IV	14,526
V	1,684
VI	171
VII	19
VIII	468
IX	0
X	20
XI	895

Table 3.2: Antibiotic resistance classes predicted by non-core genes.

Number of genomes with genes of resistance classes predicted by Ariba using the reference MegaRes database naming scheme.

Antibiotic Resistance Class	**Count**
Aminocoumarin	46
Aminoglycoside	17,968
Beta-lactam	37,758
Fluoroquinolone	69
Fosfomycin	24,205
Fusidic Acid	346
Glycopeptide	5,777
Lipopeptide	44
Macrolide-Lincosamide-Streptogramin (MLS)	22,322
Multi-Drug Resistance	13,653
Phenicol	852
Rifampin	46
Sulfonamide	36
Tetracycline	8,638
Trimethoprim	6,605

Chapter 4: The influence of horizontal gene transfer barriers on *Staphylococcus aureus* and the potential of gene transfer networks to identify novel barriers.

Abstract

Horizontal gene transfer (HGT) is widespread among bacteria. Multiple barriers must be overcome in order for a gene to be transferred by HGT and become fixed in a new species. These barriers include evolutionary forces, genetic defense mechanisms, and ecological processes. In this review, I discuss the influence barriers of HGT have had on the recent evolutionary history of the human pathogen *Staphylococcus aureus*. I describe how DNA sequencing technologies can be used to monitor gene flow within microbial communities in the form of a gene transfer network. Finally, I provide a few examples of how gene transfer networks can be used to implement medical practices that aim to limit the spread of antibiotic resistance genes by introducing barriers to HGT.

Introduction

Horizontal gene transfer (HGT) is the exchange of genetic material from a donor genome to a recipient genome without reproduction. HGT events can occur across domains of life, for example between Archaea and Bacteria (Aravind et al., 1998; Nelson et al., 1999) or more rarely, from Bacteria to Eukarya (Dunning Hotopp et al., 2007; Moran et al., 2012). In particular, HGT events are a driving force in the evolution of bacteria (Koonin & Makarova, 2001; Jain et al., 2002; Beiko, Harlow & Ragan, 2005; Pál, Papp & Lercher, 2005; Chan et al., 2009; McDaniel et al., 2010; Chan, Beiko & Ragan, 2011). Because of HGT, an individual genome may contain genes identified as having originated from a different species (Lawrence & Ochman, 1998). As a consequence, this has led to much debate on how exactly to define a bacterial species (Wayne et al., 1987; Cohan, 2002; Konstantinidis & Tiedje, 2005b; Riley & Lizotte-Waniewski, 2009).

The focus of this review is gene exchange mediated by HGT between bacteria. There are three classical mechanisms of bacterial HGT that have now been studied at molecular levels for more than 50 years (**Box 4.1**). Almost all bacterial species can undergo HGT by at least one of these mechanisms (Nakamura et al., 2004). Therefore, the simplest hypothesis to gene transfer in bacteria can be formulated as this: a given bacterial gene can move into any other bacterial species, with the same probability as another gene moving to a new donor species. This simple hypothesis is refuted because there are barriers that reduce the successful transfer of a gene, and these barriers are more efficient at limiting gene flow between some species than others (Thomas & Nielsen, 2005). These barriers to HGT are summarized in **Box 4.2** and include what I define as

evolutionary forces, defense mechanisms against foreign DNA, and ecological processes.

For a gene to transfer successfully between a donor and recipient, three steps are needed. First the donor and recipient need to be co-resident in the same physical space. Second, DNA transfer has to have occurred. Finally, the new gene in the recipient cell needs to become fixed in the population. During each of these steps multiple barriers to gene flow (**Box 4.2**) must be overcome. Genetic defense mechanisms, such as restriction enzymes and clustered regularly interspaced short palindromic repeats (CRISPRs), destroy transferred DNA as it enters the recipient cell (Pinedo & Smets, 2005; Marraffini & Sontheimer, 2008). Once an immigrant gene has entered the recipient cell through transformation or transduction, it can integrate into the host genome either by site-specific recombination (which requires that the host has a target site) or through homologous recombination, (which requires that the host has a cognate with high nucleotide identity) (Matic, Rayssiguier & Radman, 1995). Therefore, in these cases, not all host species will be efficient recipients for transferred DNA. There also exist adaptive traits that allow plasmids and phage to circumvent commonly encountered host defenses (Read, Thomas & Wilkins, 1992; Maxwell, 2016). These examples would increase the chance that a gene will be successfully integrated while decreasing the chance of being identified as foreign DNA and degraded.

Following gene acquisition, evolutionary forces can act as a barrier to the long-term fixation of the immigrant gene in its new population. A newly acquired gene must survive genetic drift and selection. Assuming the neutral theory of evolution applies to bacterial populations (Rocha, 2018), the survival of a gene that is effectively neutral to

the recipient will be dependent on genetic drift, or random chance. Long-established population genetics theory states that the strength of drift, in removing genes from the population, is increased while the gene is rare in the population (Gillespie, 2010).

A gene that is deleterious to the recipient will be more likely than a neutral gene to be lost from the population. Conversely, a gene that offers a selective advantage to the recipient is more likely be maintained. Even if a gene offers a selective advantage, while it is rare in the population the force of selection may be weak in comparison to genetic drift and it may yet be lost through chance. Unsurprisingly, HGT events involving genes that confer strong selective advantages have been the most frequently reported in the early literature. Examples of these types of events include antibiotic resistance genes (Palmer, Kos & Gilmore, 2010), metabolic genes (Fournier & Gogarten, 2008), and virulence genes (Hacker et al., 1997).

Before genetic barriers and selection come into play, the donor and recipient in the HGT event must arrive in the same space. How does this occur? Of all the barriers, we know the least about how ecological processes limit gene flow between species. Understanding the role of ecological processes on microbial communities presents a complex set of problems.

In this review I discuss examples of how differential HGT barriers have shaped the evolutionary history of the human pathogen *Staphylococcus aureus*. First, I describe how a specific HGT event has overcome multiple barriers, and ultimately led to the global expansion of methicillin-resistant *S. aureus* (MRSA). I then contrast this with

another example in which barriers appear to be limiting the expansion of another key antibiotic resistance gene (*vanA*) into *S. aureus*. Finally, I open a discussion on the importance of monitoring gene flow within microbial communities and how emerging sequencing technologies can make this a possibility.

MRSA - a case where human action can break down barriers to HGT

Staphylococcus aureus, already introduced in Chapters 1 and 3, is estimated to colonize 20-50% of humans, both children and adults (Gorwitz et al., 2008). *S. aureus* most commonly colonizes the anterior nares as well as the skin of carriers (Kluytmans, van Belkum & Verbrugh, 1997). The bacterium is not limited to humans; it can also colonize pets and livestock (Price et al., 2012; Loeffler et al., 2013). *S. aureus* is primarily transmitted through skin-to-skin contact with a carrier (Miller & Diep, 2008) but it can also be acquired from environmental surfaces where it can survive for long periods of time (Dietze et al., 2001; Scott, Duty & Callahan, 2008). Carriers often remain asymptomatic, but as an opportunistic pathogen, *S. aureus* can cause a wide variety of diseases ranging from minor skin infections to life-threatening diseases. Treatment of *S. aureus* related infections has become challenging due to the emergence and evolution of strains resistant to almost all common antibiotics. More studies are still required to determine the correlation between increased resistance and pathogenicity (Watkins, David & Salata, 2012), but resistant strains have shown increased mortality rates due to prolonged and repeated infections (Cosgrove et al., 2003).

Methicillin-resistant *S. aureus* (MRSA) is the most widespread resistance phenotype of *S. aureus*. MRSA is the result of a horizontally transferred mobile genetic element named Staphylococcal Cassette Chromosome *mec* (SCC*mec*), harboring *mecA*, the gene responsible for methicillin resistance. The *mecA* gene is highly conserved in the resistome of other *S*taphylococcal species exhibiting methicillin resistance (Archer & Pennell, 1990). Based on sequence homology, *Staphylococcus sciuri* has been identified as a potential donor of *mecA* to early MRSA clones (Wu et al., 1996; Rolo et al., 2017). Although it is more commonly isolated from animals including pets, livestock and the soil, *S. sciuri* can colonize humans as well (Kloos, 1980; Kawano et al., 1996; Couto et al., 2000; Stepanović et al., 2001). Groups of close contact such as hospital personnel (Dulon et al., 2014), athletes (Benjamin, Nikore & Takagishi, 2007), prison inmates (Baillargeon et al., 2004), and military personnel (Zinderman et al., 2004) are often susceptible to MRSA outbreaks. In 2011, surveillance of 9 US states estimated 80,461 invasive MRSA cases occurred nationwide (Dantes et al., 2013). These MRSA cases included multiple diagnoses such as bloodstream infections (80%), skin infections (22%), pneumonia (16%), and osteomyelitis (13%) among others. Between 2005 and 2011, there was a large decrease (54%) of hospital associated MRSA (HA-MRSA) cases, but the number of community-acquired MRSA (CA-MRSA) cases remained relatively stable.

MRSA has played a significant role in *S. aureus* epidemics and has been described as having occurred in “waves of resistance” (Chambers & Deleo, 2009). The first epidemic wave was due to penicillin-resistant strains. After the introduction of methicillin, MRSA first emerged in the United Kingdom in 1961 (Patricia Jevons, 1961). However, ancestral

state reconstructions suggest methicillin resistance was already present in the *S. aureus* populations up to 14 years before methicillin was introduced (Harkins et al., 2017). These early strains of MRSA were later replaced by a second epidemic wave of resistance that occurred in the late 1960s. During this time MRSA strains were limited to healthcare facilities and often times would harbor additional resistance genes (Ito et al., 2001). Beginning in the 1990s, the third epidemic wave of MRSA introduced novel SCC*mec* types with lower fitness penalties (Turlej, Hryniewicz & Empel, 2011). These novel MRSA strains expanded their habitat into human communities (CA-MRSA) (Vandenesch et al., 2003) and livestock (LA-MRSA) (Price et al., 2012) and have gone on to spread around the globe (Ayliffe, 1997). SCC*mec* are commonly acquired and lost in subtypes of *S. aureus*, suggesting limited genetic barriers to the transfer of SCC*mec* *(Noto et al., 2008; Aanensen et al., 2016)*. When MRSA first emerged, both HGT and its role in the evolution of antibiotic resistance were not well understood. Also, measures to prevent the spread of antibiotic resistant pathogens, such as patient isolation or strict hospital hygiene practices, were not yet implemented. Had these measures been implemented, early cases of MRSA would have been susceptible to both the effects of genetic drift and ecological barriers. For example, an isolated patient has an increased chance of becoming a dead-end host for the pathogen. Compounded with better hospital hygiene practices, transmission between caretakers and other patients, an ecological barrier, is limited. Instead, the unintended over-prescription of beta-lactam antibiotics offered positive selection for SCC*mec* to be maintained in the *S. aureus* population. Although SCC*mec* was maintained in the population it still had multiple evolutionary forces and ecological barriers to overcome.

Several lines of research point to SCC*mec* conferring a selective disadvantage to *S. aureus* in the absence of beta-lactam antibiotics (Andersson & Levin, 1999; Ender et al., 2004). This presents an evolutionary barrier to the acquisition of SCC*mec* types in the community where antibiotic exposure is likely limited. As a consequence, CA-MRSA has selected for smaller SCC*mec* types which maintain methicillin resistance but confer fewer other resistances (Ma et al., 2002). Currently, eleven SCC*mec* types have been discovered and each type has multiple subtypes (Turlej, Hryniewicz & Empel, 2011). Another aspect of expansion into the community is transmission of CA-MRSA is expected to be difficult due to a lack of carrier contact (Uhlemann et al., 2014; Alam et al., 2015b). To circumvent this ecological barrier, CA-MRSA strains tend to exhibit elevated virulence levels in comparison to HA-MRSA, allowing increased infection rates among healthy individuals (Boyle-Vavra & Daum, 2007; Dantes et al., 2013; Stinear et al., 2014).

The history of MRSA is a standout example of successful horizontal transfer of antibiotic resistance. It has routinely continued to circumvent barriers to the spread of SCC*mec*. Although, SCC*mec* was already within *Staphylococcus* aureus, humans are likely responsible for the widespread expansion of MRSA.

VRSA - a case where barriers to HGT can have great public health consequences

Vancomycin, a glycopeptide antibiotic, is commonly used to treat MRSA infections (Pakyz et al., 2008). Shortly after vancomycin began to be used for large-scale

prescription for MRSA in the late 1980s, limited reports of *S. aureus* isolates showing decreased susceptibility to the drug emerged. The first published case of reduced susceptibility to vancomycin was reported in 1997 (Hiramatsu et al., 1997). Based on Clinical and Laboratory Standards Institute (CLSI) recommendations this isolate, which presented minimum inhibitory concentration (MIC) of no more that 8 μg/ml was labeled as vancomycin-intermediate *S. aureus* (VISA). Subsequent genetic studies showed that the evolution of VISA involves development of chromosomal mutations across more than 20 loci. VISA cells typically had a thickened cell wall phenotype (Sieradzki & Tomasz, 2003; Alam et al., 2014a). An alternative to pathway to vancomycin emerged later in the form of vancomycin-resistant *S. aureus* (VRSA). VRSA, of relevance to this review, involved acquisition via HGT of a *vanA* gene cluster, which conferred high-level resistance to vancomycin (MIC ≥ 16 μg/ml), from either *Enterococcus faecalis* or *E. faecium* (Périchon & Courvalin, 2009). Since the first case report of VRSA in 2002, only 18 cases of VRSA in total have been reported (Saha et al., 2008; Kos et al., 2012; Azimian et al., 2012; Melo-Cristino et al., 2013; Moravvej et al., 2013; Limbago et al., 2014a; Rossi et al., 2014). The low number of VRSA cases seems like an anomaly in the light of the number of MRSA cases reported each year around the world. What are the barriers that are limiting the evolution of VRSA?

The experience of MRSA suggests that genetic barriers to uptake and spread of genes within the *S. aureus* species may be limited (Noto et al., 2008). A few studies have shown instability of certain *vanA* plasmids in MRSA may act as a genetic barrier (Périchon & Courvalin, 2004, 2006). There may also be selective barriers against VRSA. Prolonged exposure to vancomycin offers positive selection for the transfer of *vanA*, but

there is evidence that expression of the *vanA* operon is detrimental to fitness (Foucault, Courvalin & Grillot-Courvalin, 2009). With acquisition being a rare event and previous cases having been isolated, VRSA thus far has been highly susceptible to being lost from the population from genetic drift. Susceptibility to genetic drift is strengthened by improved detection and aggressive control measures are now in place to reduce the spread of resistant bacteria. Protocols for antibiotic usage, such as duration and dosage amounts, are likely to play a role in limiting selection for resistance. These measures were not in place when MRSA first emerged.

Ecologically, *S. aureus*, *E. faecalis, E. faecium* occupy different body sites within the human microbiome. *E. faecalis* and *E. faecium* are commonly found in the intestines of healthy adults (Qin et al., 2010). *S. aureus* can colonize the intestines (Acton et al., 2008) but in healthy individuals it is likely to be outcompeted by normal flora (Vesterlund et al., 2006). These observations suggest HGT between *S. aureus* and *E. faecalis* or *E. faecium* is restricted by this ecological barrier. In these cases the ecological separation between *S. aureus* and *E. faecalis* or *E. faecium* have broken down, for example through the use of a catheter (Centers for Disease Control and Prevention (CDC), 2002) or having undergone amputation (Melo-Cristino et al., 2013). Most of the described VRSA cases have been limited to a single sequence type (ST5) often associated with HA-MRSA strains (Limbago et al., 2014b). This genetic background, is outcompeted by USA300 (ST8) is thought to have kept VRSA from expanding.

MRSA and VRSA are two almost opposite examples of the effects barriers to HGT can have on the evolution of a pathogen (**Table 4.1**). I believe that routinely sequencing

reported VRSA cases along with suspected *vanA* donors would allow for extensive comparative genomic studies. These studies can offer clues to the next path of evolution for VRSA.

Using high-throughput DNA sequencing to build gene transfer networks

How can we use gene transfer networks to predict antibiotic resistance genes in pathogens from the environmental reservoir (the "resistome") within a given time frame? Importantly, we need to build up a knowledge base of pathways of known HGT events. With this knowledge we can construct gene transfer networks. Similar to popular social networks, each member of a microbial community is interacting directly or indirectly with other members. A gene transfer network is the visual representation of HGT events between members of the community. From this network, clusters of bacteria frequently exchanging genes can be identified as well as the paths in which a specific bacterium can acquire a certain gene being transferred in the community (Eppstein, 1998; Girvan & Newman, 2002). This information could be useful to determine potential targets that may be most affected by HGT barriers. In order to construct a gene transfer network, high-throughput sequencing data must be generated for the community. With this sequencing data, the HGT events between donors and recipients can be predicted. Using the predicted HGT events, a directed network can be constructed to represent the HGT pathways within the community.

High-throughput sequencing is now routinely used to sequence microbial communities

(Mason et al., 2012; Aagaard et al., 2013; Afshinnekoo et al., 2015). We can determine the species composition and their relative abundance within these communities through 16S rRNA sequencing (Caporaso et al., 2011). The 16S sequences can be separated into operational taxonomic units (OTUs) based on sequence similarity. These OTUs are representative of the bacterial species present in the community. 16S sequencing alone is only useful to determine the taxonomic census of the community. Subjecting the community to random "shotgun" sequencing allows for functional analysis of the community (Thomas, Gilbert & Meyer, 2012; Lam et al., 2015). Metagenomic studies involve taking an environmental sample, extracting DNA and subjecting it to shotgun sequencing. The individual DNA reads can be assembled into longer contigs comprising DNA from the same species. From these longer fragments, the genome context of individual genes in a species can be ascertained. We are likely to have increasing information about genome context as long-read technologies become more commonly used for metagenomics.

Current limitations to metagenomic studies include over representation of abundant species in the sequencing data. Species that are abundant in the population will also represent a larger proportion of the sequenced DNA. This can be a major problem because rare species are often missed or undetectable in the sequencing data. Without these rare species, gaps will be introduced in the gene transfer networks. If a detected HGT event is acquired from a rare species, the event will be without an accurate donor label. HGT events in which the rare species was the recipient will also not be detected. These rare species may be very important members of the community, as they are expected to be potential keystone species (Sogin et al., 2006).

Recently DNA sequencing techniques have been applied to a single cell from the microbiome (Blainey, 2013; Shapiro, Biezuner & Linnarsson, 2013). Unlike metagenomic sequencing, whole genomes are ascertained, eliminating any ambiguity about gene origin. With the use of unique sample tags, the sequence reads will be easily associated to a particular cell and can then be treated the same way whole genome projects of a single organism are treated. Advanced cell sorting techniques allow for the selection of cells based on certain characteristics. Used in congruence with metagenomic sequencing, SSG can vastly increase the detection of rare species and the accuracy of OTU binning by generating reference genomes of the sample (Blainey & Quake, 2014). This would fill in missing HGT events associated with rare species.

Taken together metagenomics and SSG would only offer a “snapshot” of a community at a single time point. For investigating community composition and functional analysis this is sufficient. In order to really begin to get an idea of the underlying ecological interactions of the community multiple snapshots (longitudinal) would be required. Longitudinal sequencing of microbial communities would more accurately detect the donor and recipients of HGT events. As the time points decrease between sequencing, HGT events that lead to introduction of a gene that is quickly lost by selection and/or drift, which are effectively evolutionary “dark matter”, will be detectable. These events, while often not evolutionarily meaningful, may be suggestive of an ecological interaction, which is meaningful.

Based on sequence data we can infer HGT directly using comparative genomics.

Comparative genomics allows recent HGT events to be detected, possibly before gene loss, through selection, takes place (Sjöstrand et al., 2014; Whidden, Zeh & Beiko, 2014). For samples that have been sequenced longitudinally, detecting HGT events is simple. At the first time point, OTU binning will create a representation of the species composition and the functional overview of the community. The same is done for the second time point. The genes from each species are compared between the two time points. HGT events can be inferred based on presence or absence of a gene between the two time points. The potential donor and recipients for each event can also be determined. This process would continue until each of the time points is compared. Once complete, a list of HGT events and the corresponding donor and recipient will have been recorded.

Parametric methods take advantage of genomic signatures rather than alignments to known genes to detect HGT events. A given bacterial genome will contain patterns, or genomic signatures, that are reflective of background mutation and recombination rates. When a HGT event occurs, the transferred gene will retain the genomic signature of the donor. By analyzing short regions of the recipient genome, genomic signatures of the donor can be recognized. Parametric algorithms commonly use GC content, di- and tri- nucleotide usage, or protein structure features as genomic signatures (Lawrence & Ochman, 1998; Karlin, 1998; Mrázek & Karlin, 1999; Bohlin, Skjerve & Ussery, 2009; Azad & Lawrence, 2011; Retchless & Lawrence, 2012). Due to the process of amelioration (Lawrence & Ochman, 1997), parametric methods are only effective at identifying recent HGT events. Also, there must be a detectable difference between the genomic signatures to identify HGT events (**Figure 4.1**). As a consequence, HGT events

between closely related relatives would not be detected.

Phylogenetic approaches use the evolutionary history of an organism to detect HGT events. This approach can be separated into two categories, explicit and implicit. Explicit algorithms seek out discrepancies between gene and species trees. Statistically significant discrepancies suggest an HGT event likely occurred. Implicit algorithms use sequence similarity and time since divergence to test for HGT events. If a gene is likely to have been horizontally transferred the best alignment score is expected to be from the donor and not the recipient. Although, certain algorithms may not always return an accurate best hit (Koski & Golding, 2001). Explicit algorithm limitations include potential to be time consuming and sensitivity to the selection of algorithm parameters (Than et al., 2007; Roure, Baurain & Philippe, 2013). While implicit algorithms, are limited to recent transfers that have a detectable difference. A thorough review of parametric and phylogenetic methods is available from Ravenhall et al (Ravenhall et al.).

Once the HGT events have been accounted for, a directed network can be used to represent the gene transfer network. Two n x n matrices, where n is the number of species, can be constructed based on the HGT events. One binary matrix, or HGT event matrix, will represent whether or not an HGT event occurred between two species and whether the species was the donor or recipient. The other matrix, a weighted matrix, will represent the total number of HGT events that occurred between the two species. From these matrices a gene transfer network can be constructed. First, every species in the community is represented as a node in the network. Using the HGT event matrix,

connections can be made between two species (nodes) from the donor species to the recipient species. The total number of HGT events is then used to label the connection. This process would continue until all species are accounted for. Once complete the gene transfer network will offer patterns of ecological interactions (**Figure 4.2**). Gene transfer networks have been used to show a majority of recent HGT events occur among closely related species and that the functional distribution of these genes is not random (Popa et al., 2011).

Using gene transfer networks to predict and monitor future spread of antibiotic resistance

A potential use for gene transfer networks is in predicting the lifespan of antibiotics. For example, teixobactin is a newly discovered antibiotic from soil that inhibits cell wall synthesis (Ling et al., 2015). It has been suggested that because this antibiotic binds to a highly conserved motif of lipid II, any mutations in this motif are likely to be deleterious. As a consequence, resistance to this antibiotic should be rare and require a lot of time to emerge. There may exist an undiscovered gene, which confers resistance to teixobactin. Metagenomic sequencing and SSG could be used together to map out the community landscape. By mapping out the genomic landscape of the community targets for teixobactin resistant genes could be predicted. If a gene were to be discovered and validated, the community could be monitored and regularly sequenced. From this, gene transfer networks for this gene could be constructed. From these networks and known barriers of transfer, time until widespread acquisition, or antibiotic lifespan, could be predicted.

Recently, VRSA was isolated in Brazil from a patient receiving vancomycin for treatment of a MRSA infection (Rossi et al., 2014). This case of VRSA presented a novel agent of HGT; the *vanA* gene cluster was captured by a plasmid that was readily transferred to another *S. aureus*. Although the *vanA* gene has an *Enterococcal* origin, the plasmid was not transferred back into a laboratory strain of *E. faecalis*. *E. faecalis* was implicated as the origin of *vanA* in previous VRSA cases; this observation suggests an undiscovered set of intermediate donors. This case of VRSA is also more closely related to USA300 (CC8), a CA-MRSA clone that has seen intercontinental spread from United States (Glaser et al., 2016). Unlike previous cases, the novel plasmid did not affect i*n vitro* fitness of the recipient strain. Fortunately, this case of VRSA did not overcome the genetic drift barrier and is presently being considered an isolated event. If this plasmid were to be maintained in the population, its characteristics would likely allow it to overcome selective and ecological barriers in much of the same way as CA-MRSA. If this case of VRSA is a precursor of what may be to come, it will be of major public health concern. Fortunately, because VRSA is a recent event, we have the potential to prevent it from becoming an epidemic. With routine monitoring and sequencing of MRSA isolates and the gut microbiome of patients from this region of Brazil, gene transfer networks could be used to predict the likelihood of *S. aureus* acquiring a similar HGT event. More importantly, it would offer the potential to implement medical practices that aim to increase ecological barriers with suspect donors. One could possibly imagine a medical intervention such as administering a specific probiotic that could outcompete known to competitively exclude either the suspect donors or *S. aureus*.

Conclusions

HGT is essential to the the evolution and expansion of antibiotic resistance genes in bacterial pathogens. There are only small set of antibiotics classes currently in our arsenal to treat these pathogens. As a consequence, it is important to understand how barriers to HGT can be used to our advantage to prevent the evolution of resistant pathogens. Current infrastructure and costs do not allow for routine microbiome sequencing of patients, but as sequencing costs continue to fall, and technology improves, this may become a possibility in the near future. Until then we can continue to make efforts to disentangle the ecology of microbial communities through metagenomic sequencing and SSG, with the ultimate goal of implementing medical practices that are based on the ecology of the patient's microbiome.

Appendix

The following appendix contains boxes, tables and figures referenced in the text of this chapter.

Box 4.1: Classical mechanisms of horizontal gene transfer in prokaryotes

Transformation

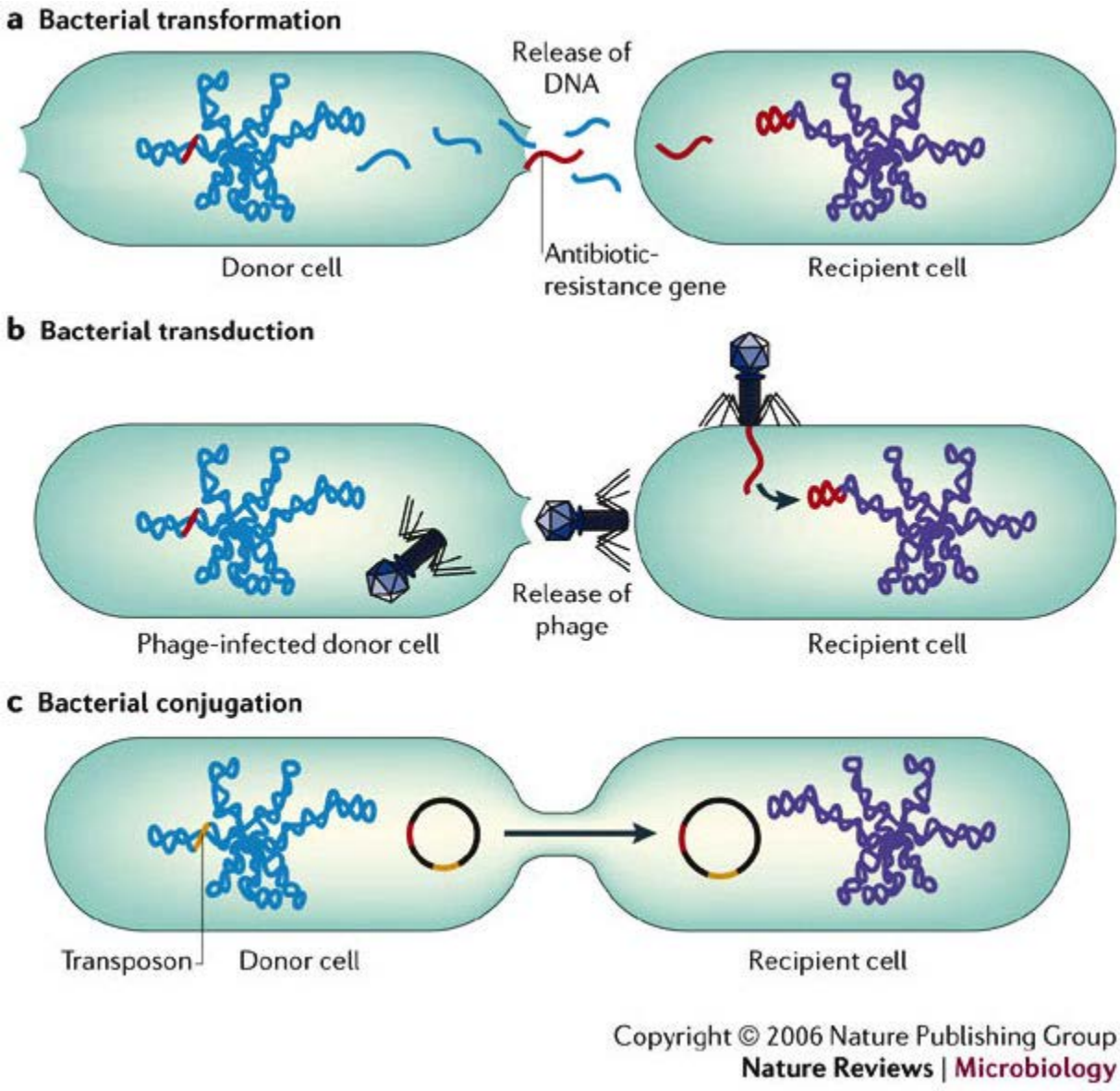
Originally discovered in 1928 (Griffith, 1928), transformation occurs when a bacterial cell acquires free-floating DNA from the environment. In order to uptake exogenous DNA bacterial cells must possess the required genes and be in a 'competent' state. Environmental cues, such as UV light, can switch a bacterium into a competent state (Michod, Wojciechowski & Hoelzer, 1988).

Transduction

Prophage excision from the host genome can, at a low rate result in packaging chromosome genes into the bacteriophage genome; these genes are then carried over to the next infected host. This process of horizontal gene transfer is referred to as transduction. Discovered in 1951 (Zinder & Lederberg, 1952), transduction takes two forms, generalized and specialized. In generalized transduction any gene present in the bacterial genome may by chance any packaged into the viral genome. In specialized transduction only, the genes on either side of a prophage may be excised.

Conjugation

Conjugation occurs when genetic material is transferred from one bacterial cell to another through a direct physical connection between the two bacterial cells through the action of an episomal genetic element. First discovered in 1946 (Tatum & Lederberg, 1947), a donor cell transfers a conjugative or mobile genetic element to a recipient cell in the form of a plasmid or transposon. Transferred plasmids vary in size and composition



Reprinted by permission from Springer Nature: (Yoko Furuya & Lowy, 2006).

Box 4.2: Barriers to horizontal gene transfer between bacteria

Evolutionary forces

Evolutionary forces can play a significant role in reducing the number of successful HGT events. Natural selection is expected to remove recipients of a deleterious gene from the population. While transfer events that offer a neutral or selective advantage will initially be highly susceptible to genetic drift (stochastic loss during replication). This is due to selection being weaker force than genetic drift when only a few cells contain a mutant (Gillespie, 2010). Genes that offer a selective advantage are more likely to spread though the population and avoid removal by drift.

Genetic mechanisms

Bacteria have many defense mechanisms in place to deal with foreign DNA. Bacteria can use restriction modification systems as protection for foreign DNA (Wilson & Murray, 1991). When certain unmethylated DNA fragments are detected, restriction endonucleases will cleave double stranded DNA at these points. The cleaved DNA is then further degraded by other endonucleases. Another defense mechanism for some bacteria, are clustered regularly interspaced short palindromic repeats (CRISPR), which can act as an adaptive immunity (Barrangou et al., 2007). CRISPRs retain DNA from previous bacteriophage infections, upon reinfection the bacteriophage DNA can be recognized and cleaved. For bacteria that undergo conjugation, surface exclusion (Achtman, Kennedy & Skurray, 1977) limits conjugative transfer if the recipient cell already contains the plasmid being transferred. Transferred genes have also been shown to be concentrated to ~1% of chromosomal regions (Oliveira et al., 2017) suggesting a link between recombination and horizontal gene transfer.

Ecological processes

The most complex limitations to HGT to understand are those that restrict the co-residence of bacterial species. In general, ecological processes might come in the form of habitat disturbance, climatic variables, resource variation, competition, or cooperation, predation (Radford, Robinson & Watson, 2009). A given microbial community will be a very complex network of these ecological processes. Collectively these processes are

driving gene exchange within the human microbiome (Smillie et al., 2011). There are also biogeographical limitations to the range of bacteria (Flores et al., 2011; Barberán et al., 2014). Adding to the complexity of ecological interactions, human interference can play significant roles in disrupting these interactions. The spread of a horizontally transferred antibiotic genes following the introduction of the drugs after WW2 is a classic example of how HGT “trade routes” allow delivery of genes with selective advantage to pathogens.

Figure 4.1: A bar graph depicting the counts of GC content for human bacterial pathogens with a completed genome in NCBI's Genome database.

In order for parametric methods to successfully detect HGT events, there must be a detectable difference in the genomic signature. This graph indicates a number of species have a similar GC content. As a genomic signature, GC content is only useful for two organisms that are significantly different.

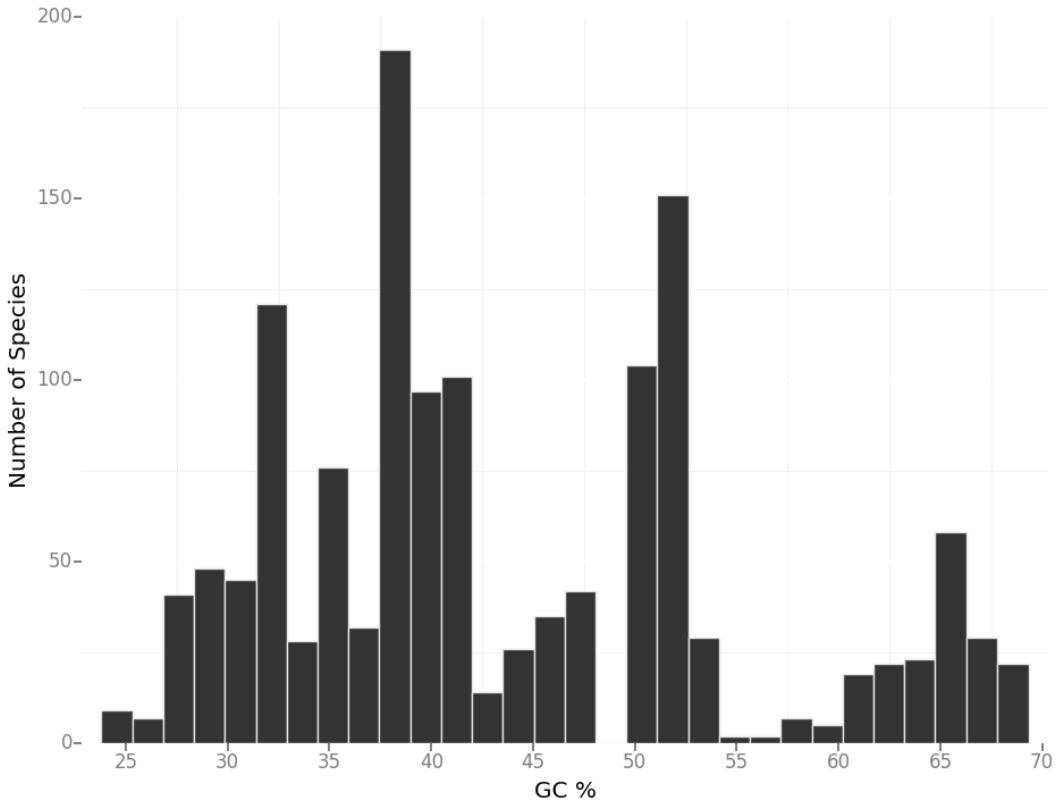


Figure 4.2: An example of a gene transfer network.
Each circle (node) of the network represents a bacterial species from a microbial community. The arrows (edges) represent recent HGT events. Annotations of these HGT events appear next to the arrows. Arrows that have numbers next to them, indicate the number of predicted HGT events between the two species. For example, the yellow circle (labeled as or numbered 109) represents *Peratoga mobilis* str. SJ95. There are two outgoing arrows and two incoming arrows. The outgoing arrows represent *P. mobilis* str. SJ95 donating a heavy metal ATPase gene to *Themoanaerobacter* str X514 and *Caldicellulosiruptor saccharolyticus* str. DSM8903. The incoming arrows represent cases when *P. moblis* str. SJ95 acted as a recipient of a heavy metal ATPase gene from *Themoanaerobacter* str X514 and a recombinase gene from *Clostridium thermocellum* str ATCC27405. Image is reprinted from open access article Popa et al (Popa et al., 2011).

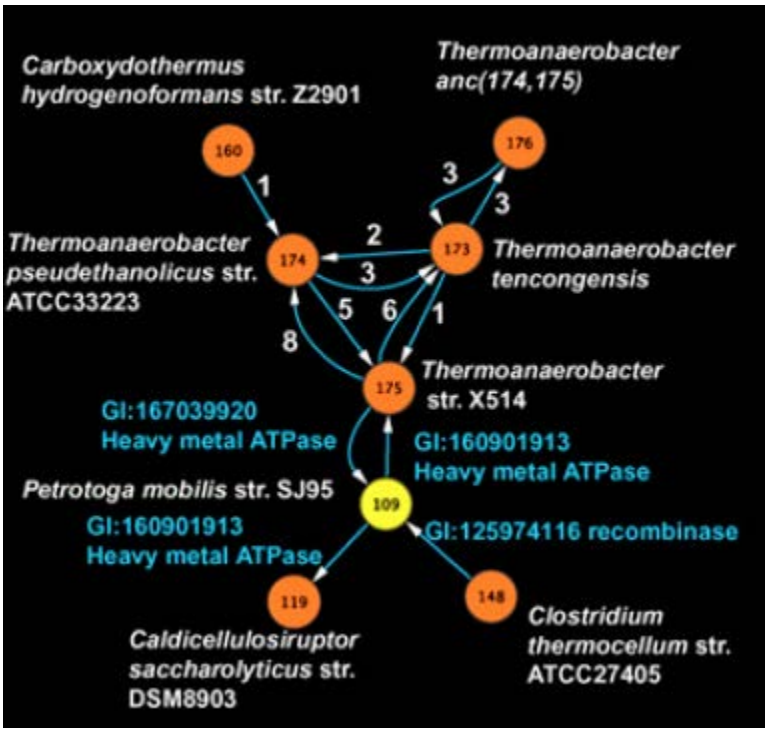


Table 4.1: The influence of barriers to HGT on MRSA and VRSA

	MRSA	**VRSA**
Genetic Mechanisms	• Limited, SCC*mec* may be regularly gained and lost in populations	• Plasmids harboring *vanA* were unstable • Novel plasmid was readily acquired
Evolutionary Forces	• Early hospital practices unintentionally selected for SCC*mec* • SCC*mec* types related to CA-MRSA are less detrimental to fitness	• Isolated cases susceptible to genetic drift • Expression of *vanA* operon may be detrimental to fitness • Novel *vanA* harboring plasmid does not reduce fitness
Ecological Processes	• Early hospital practices aided in transmission • Improved practices have decreased the spread of SCC*mec* within hospitals • Transmission in the community is limited	• Improved hospital practices reduce transmission • Donor *Enterococcal* species occupy different ecological niche • Limited to patients with extended hospital stays • A majority of past cases have an ecology similar to HA-MRSA, recent case has a CA-MRSA

Chapter 5: Summary and Future Directions

Summary

This dissertation has focused on the analysis of bacterial pathogens through the use of whole genome-sequencing. Genome sequencing has become cheap enough that it is now a common practice. The sheer amount of sequence data that is publicly available has created opportunities for secondary analyses. It has also created the opportunity to develop organism-specific bioinformatic approaches. I have taken advantage of such opportunities in two bacterial pathogens, *Bacillus anthracis* and *Staphylococcus aureus.*

In Chapter 2, I demonstrated a novel approach for differentiating low coverages of *B. anthracis* from *B. cereus* in metagenomic sequencing. I developed a model to account for the effect of sequencing errors and identified the limits of detection of *B. anthracis* and the lethal factor gene. Comparisons against "generalist" taxonomic classifiers showed my "specialist" approach was more accurate at differentiating *B. anthracis* from *B. cereus* in metagenomic sequencing. I applied my approach to the NYC dataset and identified a single sample as "Case 2". This sample would require either further sequencing or PCR validation to determine the presence of *B. anthracis* or a close *Bacillus cereus* Group (BCerG) lineage. However, it is unlikely that *B. anthracis* was present due to existing measures already in place and no reported cases of anthrax on the NYC subway. This was the first *B. anthracis* typing schema to take account for

sequencing errors, the genetic background and ability to detect the lethal factor in metagenomic sequences.

There is, however, still room for improvement in this approach. I purposely limited my dataset to high-quality completed genomes from the curated RefSeq database. There is now evidence that over time as more BCerG completed genomes have become available the misclassification of *B. anthracis* has decreased (Nasko et al., 2018b). Out of curiosity, I investigated a set of 3,400 uncurated assemblies from the Bacillus genus. I found only about 80k 31-mers remained specific to *B. anthracis*. Although these assemblies were incomplete and may still contain errors, they do provide an idea of what to expect in the future. As the phylogeny between *B. anthracis* and *B. cereus* is filled in, I suspect of many of the 240k *B. anthracis* specific 31-mers I identified to not be specific for *B. anthracis*. It may even become clear that there were no *B. anthracis* specific 31-mers due to the close relationship of BCerG members. An alternative approach is to redefine BCerG as single species and treat current members as strains of BCerG. This would require the use of 51-mers over 31-mers because 51-mers provide strain-level resolution as opposed to species-level (Koslicki & Falush, 2016). My approach could easily be adapted to 51-mers but would require significantly more computational resources due to the increased k-mer space of 51-mers (4^{51} possible k-mers).

In Chapter 3, I introduced Staphopia as a community resource focused on *Staphylococcus aureus* genomics. Through Staphopia, I provided analysis of over 40,000 *S. aureus* genomes. I explored patterns of evolution in *S. aureus* in the global *S.*

aureus population. I also presented a novel method to rationally select publicly available *S. aureus* genomes for comparative genomic studies. I developed a comprehensive analysis pipeline, a web platform, an application programming interface (API) and an R package for Staphopia.

The development of Staphopia has been a tremendous undertaking. I have generated over 50,000 lines of code, 10 TBs of analysis results, liked over 100 publications to sequencing projects and wrangled all of this into a usable resource for the community. This was a time-consuming process, causing much of the biological potential for Staphopia to remain unrealized. For example, Staphopia still lacks some analyses that are important for understanding the regulation of virulence, role of phage and mutation driven resistance in *S. aureus*. Algorithms to determine accessory gene regulator (agr) types (Shopsin et al., 2003), identification of *S. aureus* specific phage (Deghorain & Van Melderen, 2012), improved SCCmec typing (Enright et al., 2002), and antibiotic resistance associated with mutations (Gordon et al., 2014) are necessary for these analyses. There are also numerous studies that can be conducted using the data available in Staphopoia including sub-population divergence types, horizontal gene transfer events, and geographic patterns in evolution.

The long-term viability of Staphopia is ultimately dependent on future funding support. However, even with adequate funding it is an open question whether the analysis methods I have applied to 40,000 genomes will scale to the amount of date that will appear in the near future. Future data growth would require Staphopia to be migrated to the cloud. This is a simple procedure, but to properly host Staphopia it would cost 10s of

thousands of dollars per year. Without funding there are concerns for the eventual shutdown of Staphopia. As the sole developer of Staphopia, I have tried to develop Staphopia to prevent this. One of Staphopia's strong points is its portability. Staphopia's analysis pipeline is easily installed and run on small resources such as laptops to large-scale cloud platforms. The PostgreSQL-backed Staphopia database can readily be distributed to others. I have also chosen to use open-source frameworks where possible so that Staphopia is more accessible to other developers. Essentially, I have tried to develop Staphopia so someone other than myself can acquire the data and pick up where I left off. I think adopting a philosophy as I have for Staphopia will be necessary in the future where maintaining large databases permanently will not be plausible.

In Chapter 4, I provided a literature review which described barriers to horizontal gene transfer events. I used *S. aureus* to demonstrate how barriers have influenced its recent evolutionary history. I also described how of sequencing could be used to monitor gene flow and potential clinical applications of monitoring gene flow.

Future Directions: Macro-scale bacterial genomics

To date comparative genomic studies have been limited to 10s to 100s of bacterial genomes. As the deluge of sequencing continues, we will more commonly see comparative genomic studies consisting of 1,000s to 10,000s of genomes. For the remainder of the discussion I have defined these large scale studies as "macro-scale". Macro-scale studies will present us with a new set of rewards and challenges.

Rewards of macro-scale genomics

Statistical power

Statistical power is defined as the likelihood a study will detect an effect when there is one to be detected. In the context of bacteria, an example is to test if a mutation has an effect on a phenotype with statistical significance. Dependent on the strength, or size, of the effect, this can require a large number of samples to provide enough statistical power. Effect size is often times referred to as small (d=0.2), medium (d=0.5) or large (d=0.8) (Cohen, 1988). The power to detect an effect size with a given sample size can be calculated by estimating mean differences between two groups (Shein-Chung Chow, 2009). As an hypothetical example, a case control study of 50 cases and 50 controls has 80% power to detect a medium effect size. At 100,000 genomes 200 cases have 80% power to detect a small effect size. If the power is increased to 90%, the 100-sample study can now only detect large effect size while at 100,000 genomes small effects can still be detected. The sample size of macro-scale studies would provide statistical power to detect small effect sizes with minimal cases.

A better overview of a species

Comparative genomic studies often investigate a closely related set of samples. As a consequence, the total extent sequenced genetic variation within a species may be limited. One measure of sequenced diversity within a species is to use sequence types (ST). PubMLST (https://pubmlst.org/saureus/) currently lists over 4,500 different STs in *S. aureus*. In Chapter 3, I showed that of these STs only 25% were represented in current *S. aureus* sequencing efforts. This suggests that 75% of the *S. aureus* diversity

has yet to be sequenced. This is likely not the case as many of these unsequenced STs probably fall in existing clonal complexes. Another method to determine the sequenced diversity is to create a pan-genome across all the samples. As the total extent of genetic variation within a species is approached the size of the pan-genome will plateau. Macro-scale studies have the potential to offer a better view of a bacterial species as whole. Each sequenced genome will improve the overall view of a species. It will, however, be necessary to make an effort to sequence novel strains of a species.

Rational sampling

A direct product of broadening the view of a bacterial species is the concept of rational sampling. In the past comparative genomic studies have relied on what was available. This could involve including samples that may not be optimal for testing a hypothesis of interest. I showed in Chapter 3 that the current *S. aureus* sequencing effort has primarily focused on clinical isolates from a few sequence types and limited geographic locations. This is a classic case of oversampling a few groups within *S. aureus*. A consequence of this is that overrepresentation these groups can lead to biases in comparative genomic studies. As 100s of thousands of genomes become available, it will become possible to rationally sample the data to reduce these biases. This will shift the current trend of using all available genomes to using genomes that best test a hypothesis.

Challenges of macro-scale genomics

Imperfect data

Publicly available data is free and relatively easy to obtain, but the quality of sequencing and metadata is variable. Because macro-scale studies build off of multiple existing projects, setting standards that grade data quality is necessary. While sequencing data can be quantified, poor quality metadata is irretrievable. This is an important problem, because available metadata can dictate the ability to conduct comparative genomic studies, such as GWAS. The most prominent metadata issue I have had to deal with is missing data. Little information about a sample is required during submission to a public repository as a consequence it is instead reported in an associated publication. It can be difficult to determine if a sample is even linked to a publication and there is not a standard method for reporting results in publications. This requires each sample to be manually linked to a publication and the publication uniquely processed. Another prominent issue has been inconsistent naming. For example, for *S. aureus* sequencing projects, humans as a host are represented by different derivatives of the common name, human, and the scientific name *Homo sapiens*. Programmatically “human” is not the same as “Human”, likewise “*Homo sapiens*” is not the same as “*H. sapiens*”, so each of these differences must be discovered and accounted for. A controlled vocabulary for metadata fields, such as host, and would fix this type of issue. There are existing proposals for metadata standardization (Field et al., 2008; Hirwade, 2011; McQuilton et al., 2016), but their subsequent adoption and enforcement remains to be seen (Ten Hoopen et al., 2017).

Evolving sequencing technologies

Over the course of developing Staphopia, I have witnessed 3 shifts in sequencing technology usage. At the beginning Roche 454 pyrosequencing was the dominant technology for bacterial genome sequencing. It was quickly replaced by Illumina short-read sequencing. While Illumina still represents over 99% of *S. aureus* sequencing, usage of long-read technologies (PacBio and Nanopore) is becoming more common. Each of these technologies requires different approaches of bioinformatic analysis. Staphopia is based on Illumina data but will need to develop a completely different workflow for long-read projects. Long-read technologies require different algorithms for sequence error correction, assembly and variant identification. Although, I do foresee process of adapting to new technologies to be a "*temporary*" problem for sequencing bacterial isolates. Recently, a 2.2Mb read from the human genome was generated from an Oxford Nanopore MinION (Payne et al., 2018). Long-read technologies are improving but the ~10-15% per-base error rate is a significant limitation. Overcoming this error rate requires increased coverage or additional Illumina sequencing. If high-quality megabase length reads were to become the norm, it is likely long-read technologies would become the first choice for sequencing a bacterial isolate solely on the ability to assemble complete chromosomes.

Data management and distribution

When dealing with 1000s of genomes, data management becomes a significant issue. Without proper planning this can quickly become a logistical nightmare dealing with millions of files. Over the course of my graduate studies, I have put a lot of thought into

data management. The most basic step to managing data is to create a consistent directory structure that allows programmatic access. This is especially important to efficiently navigate through millions of directories. Another step that be taken is to make use of a database management system (DBMS) as I did in the case of Staphopia. In my case, I used PostgreSQL, a relational database, which aims to present data as connected tables of information. Usage of DBMS forces the user to really think about how best to organize the data and provides a starting point for distributing the data.

Generating scientific data and distributing it to the community is an important aspect of the scientific process. Distributing the amount of data produced by macro-scale studies will not fit with the current framework. In chapters 2 and 3 of this dissertation I produced over 20 TB of new data. I have taken multiple steps to ensure the reproducibility of my studies, but this does not negate a significant financial and time cost to do so. Currently, it is my burden as the researcher to maintain the data from my studies. This is not an optimal solution because most labs are not set up to act as data hosts which can ensure long-term data availability. I think it will become necessary for the funding agency, the host institute or even the journal publishing the research to ensure long-term data availability. As more macro-scale studies emerge, this problem will need to be addressed a solution determined.

Scalability

There has been an arms race between producing biological sequences and the ability to analyze the sequences. In other words, we are producing more sequences than we can

analyze. This boils down to two limitations. The first is computational and the other algorithmic. The current computational solution for macro-scale studies is to throw more compute power at it. This has become easier with the use of cloud platforms but with a price that funding may not allow. Eventually, and this leads to the second limitation, new algorithms will need to be developed for macro-scale genomics. There is a definite need for algorithms that can process 1,000s to 10,000s of genomes. However, I think the storage requirements of such analyses are often overlooked. Computational analysis is often done only once, but the results must be stored for the duration of the study and potentially for future distribution. I think there will be a need for better compression algorithms that can reduce the overall storage requirements of macro-scale studies. This will be critical for maintenance and distribution of the study results.

Emerging macro-scale genomic projects

Macro-scale genomic projects are now an intrinsic part of microbiology and will continue to grow in importance. Recently the first bacterial genome study to include over 100,000 samples, determined the global population structure of the Salmonella genus (Alikhan et al., 2018). The 100k Pathogen Genome Project (Weimer, 2017), a partnership between the Food and Drug Administration (FDA) and UC Davis, will sequence 100,000 bacterial isolates from foodborne illnesses. The NCTC 3000 project from Public Health England, will sequence and assemble completed chromosomes for 3,000 bacterial strains using PacBio long read technology. Another product of the FDA is GenomeTrakr (Safety & Nutrition). GenomeTrakr is a collaborative project consisting of 40 labs within the U.S. and 20 labs located outside the U.S. As of April 2017,

GenomeTrakr has sequenced more than 185,000 isolates and contributed more than 175 completed genomes.

There are also many macro-scale metagenomic projects that are being worked on. MetaSUB (http://metasub.org/) will include more than 15,000 metagenomic samples from over 60 subways and urban environments across the world. The American Gut Project (AGP) (McDonald et al., 2018) is a crowdsourced project in which more than 10,000 citizen-scientists have submitted fecal samples to the project. AGP is an extension of the Earth Microbiome Project (EMP) (Thompson et al., 2017) which currently has over 30,000 samples from many different biomes across the world. AGP and EMP are currently limited to 16S sequences but represent the first steps into macro-scale metagenomic projects that crowdsource sampling to more efficiently characterize numerous environments.

Final remarks

One thing not reflected in this dissertation is my interest in teaching. During my dissertation I was a teaching assistant in multiple classes and even developed an advanced programming course for graduate students. During these courses I tried to make bioinformatics more approachable to others outside the field. I often see bioinformatics distancing itself from other fields. It most likely boils down to the fact that in bioinformatics there are so many methods to do similar analyses and often little indication on which is most suitable. This can present an overwhelming learning curve for people getting started in sequence analysis. It is for this reason, I chose to focus on

how to conduct a bacterial sequence analysis step-by-step in the introduction. I made an effort to call out appropriate literature reviews that can be used to delve further into specific topics of sequence analysis. I have also highlighted emerging challenges and projects that will present future opportunities for those interested in bacterial sequence analysis. It is my hope that having read this, you may one day recommend this to someone looking to begin their adventure into bacterial sequence analysis.

Appendix: Other Published Work

Over the course of graduate school, I had the opportunity to take part in a number of projects. These projects provided valuable experiences in collaborative science and science for the public. These collaborations include members of the CDC, the Georgia Aquarium, University of Chicago and other Emory University colleagues. Many of these projects have led to 8 published works that fall into three groups: *Staphylococcus aureus* genomics, bioinformatic tool development, and Whale Shark genomics. Although, these were not directly related to my dissertation, the skills and knowledge gain have definitely indirectly influenced my dissertation work.

The first group of published works are related to *S. aureus* genomics. The first publication demonstrated within household microevolution and transmission of *S. aureus* USA300 (Alam et al., 2015a). The next publication used genomics to redefine the clinical definition of *S. aureus* USA500 (Frisch et al., 2018). The final publication demonstrated how USA300 persisted on multiple body sites after infection (Read et al., 2017b). These publications introduced me to a number or new bioinformatic approaches that were useful for my dissertation. In many of these works I worked alongside Tauqeer Alam. Working together has created a strong collaboration between us. It was also through this work that, Tim was able to introduce me to Santiago Castillo-Ramírez. Both Santiago and Tauqeer have been helpful in providing valuable feedback and improvements to this dissertation.

The second group, are more in line with my background in bioinformatics and involve the development of novel tools. The first used machine learning to classify *S. aureus* DNA within sequencing datasets (Hogan et al., 2013). Another publication also used machine learning to differentiate vancomycin susceptible from vancomycin resistant *S. aureus* samples using sequencing (Rishishwar et al., 2013). The final publication, developed strain-level resolution of *S. aureus* in metagenomic sequencing (Joseph et al., 2016). Similar to the previous group, these publications also introduced me to a number of new bioinformatic approaches as well as the thought process behind algorithm development. From these works I was introduced to Jim Hogan, a computer scientist from Queensland University of Technology, who I continued to work with for Chapter 2 of this dissertation.

The third group of published works were a valuable lesson in cross-discipline collaborations. Through a collaboration with Al Dove from the Georgia Aquarium we produced the first completed mitochondrial genome from the Whale Shark (Alam et al., 2014b). We also produced the first draft assembly of the Whale Shark genome (Read et al., 2017a). These publications were a huge effort, especially from a lab that only deals with microbial genomes. One experience from this collaboration that stuck out to me, was the interactions between science, the general public, and businesses. Through funding gained from business partnerships, the Georgia Aquarium allowed me to set up a server for others to analyze the data we produced. This was eventually integrated into an undergraduate biology course here at Emory.

Collectively, these publications broadly represent my research interests in using bioinformatics to improve public health. Most importantly, I think these opportunities provided me with valuable experiences I would have otherwise missed out on.

Bibliography

Aagaard K., Petrosino J., Keitel W., Watson M., Katancik J., Garcia N., Patel S., Cutting M., Madden T., Hamilton H., Harris E., Gevers D., Simone G., McInnes P., Versalovic J. 2013. The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* 27:1012–1022. DOI: 10.1096/fj.12-220806.

Aanensen DM., Feil EJ., Holden MTG., Dordel J., Yeats CA., Fedosejev A., Goater R., Castillo-Ramírez S., Corander J., Colijn C., Chlebowicz MA., Schouls L., Heck M., Pluister G., Ruimy R., Kahlmeter G., Åhman J., Matuschek E., Friedrich AW., Parkhill J., Bentley SD., Spratt BG., Grundmann H., European SRL Working Group. 2016. Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: a Population Snapshot of Invasive Staphylococcus aureus in Europe. *mBio* 7. DOI: 10.1128/mBio.00444-16.

Achtman M., Kennedy N., Skurray R. 1977. Cell--cell interactions in conjugating Escherichia coli: role of traT protein in surface exclusion. *Proceedings of the National Academy of Sciences of the United States of America* 74:5104–5108.

Acton DS., Tempelmans Plat-Sinnige MJ., van Wamel W., de Groot N., van Belkum A. 2008. Intestinal carriage of Staphylococcus aureus: how does its frequency compare with that of nasal carriage and what is its clinical impact? *European journal of clinical microbiology & infectious diseases: official publication of the European Society of Clinical Microbiology* 28:115–127. DOI: 10.1007/s10096-008-0602-7.

Afshinnekoo E., Meydan C., Chowdhury S., Jaroudi D., Boyer C., Bernstein N., Maritz JM., Reeves D., Gandara J., Chhangawala S., Ahsanuddin S., Simmons A., Nessel T., Sundaresh B., Pereira E., Jorgensen E., Kolokotronis S-O., Kirchberger N., Garcia I., Gandara D., Dhanraj S., Nawrin T., Saletore Y., Alexander N., Vijay P., Hénaff EM., Zumbo P., Walsh M., O'Mullan GD., Tighe S., Dudley JT., Dunaif A., Ennis S., O'Halloran E., Magalhaes TR., Boone B., Jones AL., Muth TR., Paolantonio KS., Alter E., Schadt EE., Garbarino J., Prill RJ., Carlton JM., Levy S.,

Mason CE. 2015. Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Systems*. DOI: 10.1016/j.cels.2015.01.001.

Alam MT., Petit RA 3rd., Crispell EK., Thornton TA., Conneely KN., Jiang Y., Satola SW., Read TD. 2014a. Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. *Genome biology and evolution* 6:1174–1185. DOI: 10.1093/gbe/evu092.

Alam MT., Petit RA 3rd., Read TD., Dove ADM. 2014b. The complete mitochondrial genome sequence of the world's largest fish, the whale shark (Rhincodon typus), and its comparison with those of related shark species. *Gene* 539:44–49. DOI: 10.1016/j.gene.2014.01.064.

Alam MT., Read TD., Petit RA 3rd., Boyle-Vavra S., Miller LG., Eells SJ., Daum RS., David MZ. 2015a. Transmission and microevolution of USA300 MRSA in U.S. households: evidence from whole-genome sequencing. *mBio* 6:e00054. DOI: 10.1128/mBio.00054-15.

Alikhan N-F., Zhou Z., Sergeant MJ., Achtman M. 2018. A genomic overview of the population structure of Salmonella. *PLoS genetics* 14:e1007261. DOI: 10.1371/journal.pgen.1007261.

Altschul SF., Gish W., Miller W., Myers EW., Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* 215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Andersson DI., Levin BR. 1999. The biological cost of antibiotic resistance. *Current opinion in microbiology* 2:489–493.

Antipov D., Hartwick N., Shen M., Raiko M., Lapidus A., Pevzner PA. 2016b. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics* 32:3380–3387. DOI: 10.1093/bioinformatics/btw493.

Aravind L., Tatusov RL., Wolf YI., Walker DR., Koonin EV. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends in genetics: TIG* 14:442–444.

Archer GL., Pennell E. 1990. Detection of methicillin resistance in staphylococci by using a DNA probe. *Antimicrobial agents and chemotherapy* 34:1720–1724.

Ayliffe GA. 1997. The progressive intercontinental spread of methicillin-resistant Staphylococcus aureus. *Clinical infectious diseases: an official publication of the*

Infectious Diseases Society of America 24 Suppl 1:S74–9. DOI: 10.1093/clinids/24.Supplement_1.S74.

Azad RK., Lawrence JG. 2011. Towards more robust methods of alien gene detection. *Nucleic acids research* 39:e56. DOI: 10.1093/nar/gkr059.

Azimian A., Havaei SA., Fazeli H., Naderi M., Ghazvini K., Samiee SM., Soleimani M., Peerayeh SN. 2012. Genetic characterization of a vancomycin-resistant Staphylococcus aureus isolate from the respiratory tract of a patient in a university hospital in northeastern Iran. *Journal of clinical microbiology* 50:3581–3585. DOI: 10.1128/JCM.01727-12.

Baer R., Bankier AT., Biggin MD., Deininger PL., Farrell PJ., Gibson TJ., Hatfull G., Hudson GS., Satchwell SC., Séguin C., Tuffnell PS., Barrell BG. 1984. DNA sequence and expression of the B95-8 Epstein—Barr virus genome. *Nature* 310:207. DOI: 10.1038/310207a0.

Baillargeon J., Kelley MF., Leach CT., Baillargeon G., Pollock BH. 2004. Methicillin-resistant Staphylococcus aureus infection in the Texas prison system. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* 38:e92–5. DOI: 10.1086/383146.

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. *Journal of computational biology: a journal of computational molecular cell biology* 19:455–477. DOI: 10.1089/cmb.2012.0021.

Barberán A., Ramirez KS., Leff JW., Bradford MA., Wall DH., Fierer N. 2014. Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecology letters* 17:794–802. DOI: 10.1111/ele.12282.

Barrangou R., Fremaux C., Deveau H., Richards M., Boyaval P., Moineau S., Romero DA., Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315:1709–1712. DOI: 10.1126/science.1138140.

Bayliss SC., Hunt VL., Yokoyama M., Thorpe HA., Feil EJ. 2017. The use of Oxford Nanopore native barcoding for complete genome assembly. *GigaScience*. DOI: 10.1093/gigascience/gix001.

Beiko RG., Harlow TJ., Ragan MA. 2005. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America* 102:14332–14337. DOI: 10.1073/pnas.0504068102.

Benjamin HJ., Nikore V., Takagishi J. 2007. Practical management: community-associated methicillin-resistant Staphylococcus aureus (CA-MRSA): the latest sports epidemic. *Clinical journal of sport medicine: official journal of the Canadian Academy of Sport Medicine* 17:393–397. DOI: 10.1097/JSM.0b013e31814be92b.

Besemer J., Lomsadze A., Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic acids research* 29:2607–2618.

Blainey PC. 2013. The future is now: single-cell genomics of bacteria and archaea. *FEMS microbiology reviews* 37:407–427. DOI: 10.1111/1574-6976.12015.

Blainey PC., Quake SR. 2014. Dissecting genomic diversity, one cell at a time. *Nature methods* 11:19–21.

Bland C., Ramsey TL., Sabree F., Lowe M., Brown K., Kyrpides NC., Hugenholtz P. 2007. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC bioinformatics* 8:209. DOI: 10.1186/1471-2105-8-209.

Bohlin J., Skjerve E., Ussery D. 2009. Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering. *BMC genomics* 10:487. DOI: 10.1186/1471-2164-10-487.

Bolger AM., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.

Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C-H., Xie D., Suchard MA., Rambaut A., Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology* 10:e1003537. DOI: 10.1371/journal.pcbi.1003537.

Boyle-Vavra S., Daum RS. 2007. Community-acquired methicillin-resistant Staphylococcus aureus: the role of Panton-Valentine leukocidin. *Laboratory investigation; a journal of technical methods and pathology* 87:3–9. DOI:

10.1038/labinvest.3700501.

Bradley P., den Bakker H., Rocha E., McVean G., Iqbal Z. 2017. Real-time search of all bacterial and viral genomic data. *bioRxiv*:234955. DOI: 10.1101/234955.

Bradnam KR., Fass JN., Alexandrov A., Baranay P., Bechner M., Birol I., Boisvert S., Chapman JA., Chapuis G., Chikhi R., Chitsaz H., Chou W-C., Corbeil J., Del Fabbro C., Docking TR., Durbin R., Earl D., Emrich S., Fedotov P., Fonseca NA., Ganapathy G., Gibbs RA., Gnerre S., Godzaridis É., Goldstein S., Haimel M., Hall G., Haussler D., Hiatt JB., Ho IY., Howard J., Hunt M., Jackman SD., Jaffe DB., Jarvis ED., Jiang H., Kazakov S., Kersey PJ., Kitzman JO., Knight JR., Koren S., Lam T-W., Lavenier D., Laviolette F., Li Y., Li Z., Liu B., Liu Y., Luo R., MacCallum I., MacManes MD., Maillet N., Melnikov S., Naquin D., Ning Z., Otto TD., Paten B., Paulo OS., Phillippy AM., Pina-Martins F., Place M., Przybylski D., Qin X., Qu C., Ribeiro FJ., Richards S., Rokhsar DS., Ruby JG., Scalabrin S., Schatz MC., Schwartz DC., Sergushichev A., Sharpe T., Shaw TI., Shendure J., Shi Y., Simpson JT., Song H., Tsarev F., Vezzi F., Vicedomini R., Vieira BM., Wang J., Worley KC., Yin S., Yiu S-M., Yuan J., Zhang G., Zhang H., Zhou S., Korf IF. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2:1–31. DOI: 10.1186/2047-217X-2-10.

Bragg TS., Robertson DL. 1989. Nucleotide sequence and analysis of the lethal factor gene (lef) from Bacillus anthracis. *Gene* 81:45–54.

Breitwieser FP., Salzberg SL. 2018. KrakenHLL: Confident and fast metagenomics classification using unique k-mer counts. *bioRxiv*:262956. DOI: 10.1101/262956.

Brenner S., Johnson M., Bridgham J., Golda G., Lloyd DH., Johnson D., Luo S., McCurdy S., Foy M., Ewan M., Roth R., George D., Eletr S., Albrecht G., Vermaas E., Williams SR., Moon K., Burcham T., Pallas M., DuBridge RB., Kirchner J., Fearon K., Mao J., Corcoran K. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology* 18:630–634. DOI: 10.1038/76469.

Brown CT., Olm MR., Thomas BC., Banfield JF. 2016. Measurement of bacterial replication rates in microbial communities. *Nature biotechnology* 34:1256–1263. DOI: 10.1038/nbt.3704.

Bushnell B. 2016. BBMap short read aligner. *University of California, Berkeley*,

California. URL http://sourceforge. net/projects/bbmap.

Cachat E., Barker M., Read TD., Priest FG. 2008. A Bacillus thuringiensis strain producing a polyglutamate capsule resembling that of Bacillus anthracis. *FEMS microbiology letters* 285:220–226.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009. BLAST+: architecture and applications. *BMC bioinformatics* 10:421. DOI: 10.1186/1471-2105-10-421.

Caporaso JG., Lauber CL., Walters WA., Berg-Lyons D., Lozupone CA., Turnbaugh PJ., Fierer N., Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America* 108 Suppl 1:4516–4522. DOI: 10.1073/pnas.1000080107.

Carattoli A., Zankari E., García-Fernández A., Voldby Larsen M., Lund O., Villa L., Møller Aarestrup F., Hasman H. 2014. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial agents and chemotherapy* 58:3895–3903. DOI: 10.1128/AAC.02412-14.

Carlson CJ., Getz WM., Kausrud KL., Cizauskas CA., Blackburn JK., Bustos Carrillo FA., Colwell R., Easterday WR., Ganz HH., Kamath PL., Økstad OA., Turner WC., Kolstø A-B., Stenseth NC. 2018. Spores and soil from six sides: interdisciplinarity and the environmental biology of anthrax (Bacillus anthracis) : The environmental biology of Bacillus anthracis. *Biological Reviews* 424:329. DOI: 10.1111/brv.12420.

Centers for Disease Control and Prevention (CDC). 2002. Staphylococcus aureus resistant to vancomycin--United States, 2002. *MMWR. Morbidity and mortality weekly report* 51:565–567.

Chambers HF., Deleo FR. 2009. Waves of resistance: Staphylococcus aureus in the antibiotic era. *Nature reviews. Microbiology* 7:629–641. DOI: 10.1038/nrmicro2200.

Chan CX., Beiko RG., Darling AE., Ragan MA. 2009. Lateral transfer of genes and gene fragments in prokaryotes. *Genome biology and evolution* 1:429–438. DOI: 10.1093/gbe/evp044.

Chan CX., Beiko RG., Ragan MA. 2011. Lateral transfer of genes and gene fragments in Staphylococcus extends beyond mobile elements. *Journal of bacteriology*

193:3964–3977. DOI: 10.1128/JB.01524-10.

Chen L., Zheng D., Liu B., Yang J., Jin Q. 2016a. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. *Nucleic acids research* 44:D694–D697. DOI: 10.1093/nar/gkv1239.

Cohan FM. 2002. What are bacterial species? *Annual review of microbiology* 56:457–487. DOI: 10.1146/annurev.micro.56.012302.160634.

Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Collins C., Didelot X. 2018. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS computational biology* 14:e1005958. DOI: 10.1371/journal.pcbi.1005958.

Compeau PEC., Pevzner PA., Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nature biotechnology* 29:987–991. DOI: 10.1038/nbt.2023.

Cosgrove SE., Sakoulas G., Perencevich EN., Schwaber MJ., Karchmer AW., Carmeli Y. 2003. Comparison of mortality associated with methicillin-resistant and methicillin-susceptible Staphylococcus aureus bacteremia: a meta-analysis. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* 36:53–59. DOI: 10.1086/345476.

Couto I., Sanches IS., Sá-Leão R., de Lencastre H. 2000. Molecular characterization of Staphylococcus sciuri strains isolated from humans. *Journal of clinical microbiology* 38:1136–1143.

Dantes R., Mu Y., Belflower R., Aragon D., Dumyati G., Harrison LH., Lessa FC., Lynfield R., Nadle J., Petit S., Ray SM., Schaffner W., Townes J., Fridkin S., Emerging Infections Program–Active Bacterial Core Surveillance MRSA Surveillance Investigators. 2013. National burden of invasive methicillin-resistant Staphylococcus aureus infections, United States, 2011. *JAMA internal medicine* 173:1970–1978. DOI: 10.1001/jamainternmed.2013.10423.

Deghorain M., Van Melderen L. 2012. The Staphylococci phages family: an overview. *Viruses* 4:3316–3335.

Delcher AL., Harmon D., Kasif S., White O., Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. *Nucleic acids research* 27:4636–4641.

Deorowicz S., Kokot M., Grabowski S., Debudaj-Grabysz A. 2015. KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics* 31:1569–1576. DOI:

10.1093/bioinformatics/btv022.

DePristo MA., Banks E., Poplin R., Garimella KV., Maguire JR., Hartl C., Philippakis AA., del Angel G., Rivas MA., Hanna M., McKenna A., Fennell TJ., Kernytsky AM., Sivachenko AY., Cibulskis K., Gabriel SB., Altshuler D., Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 43:491–498. DOI: 10.1038/ng.806.

Didelot X., Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS computational biology* 11:e1004041. DOI: 10.1371/journal.pcbi.1004041.

Dietze B., Rath A., Wendt C., Martiny H. 2001. Survival of MRSA on sterile goods packaging. *The Journal of hospital infection* 49:255–261. DOI: 10.1053/jhin.2001.1094.

Ding W., Baumdicker F., Neher RA. 2018. panX: pan-genome analysis and exploration. *Nucleic acids research* 46:e5. DOI: 10.1093/nar/gkx977.

Di Tommaso P., Chatzou M., Floden EW., Barja PP., Palumbo E., Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nature biotechnology* 35:316–319. DOI: 10.1038/nbt.3820.

Dixon TC., Meselson M., Guillemin J., Hanna PC. 1999. Anthrax. *The New England journal of medicine* 341:815–826. DOI: 10.1056/NEJM199909093411107.

DNA Sequencing Costs: Data. *Available at https://www.genome.gov/sequencingcostsdata/* (accessed April 29, 2018).

Dulon M., Peters C., Schablon A., Nienhaus A. 2014. MRSA carriage among healthcare workers in non-outbreak settings in Europe and the United States: a systematic review. *BMC infectious diseases* 14:363. DOI: 10.1186/1471-2334-14-363.

Dunning Hotopp JC., Clark ME., Oliveira DCSG., Foster JM., Fischer P., Muñoz Torres MC., Giebel JD., Kumar N., Ishmael N., Wang S., Ingram J., Nene RV., Shepard J., Tomkins J., Richards S., Spiro DJ., Ghedin E., Slatko BE., Tettelin H., Werren JH. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 317:1753–1756. DOI: 10.1126/science.1142490.

Earl D., Bradnam K., St John J., Darling A., Lin D., Fass J., Yu HOK., Buffalo V., Zerbino DR., Diekhans M., Nguyen N., Ariyaratne PN., Sung W-K., Ning Z., Haimel M., Simpson JT., Fonseca NA., Birol İ., Docking TR., Ho IY., Rokhsar DS., Chikhi

R., Lavenier D., Chapuis G., Naquin D., Maillet N., Schatz MC., Kelley DR., Phillippy AM., Koren S., Yang S-P., Wu W., Chou W-C., Srivastava A., Shaw TI., Ruby JG., Skewes-Cox P., Betegon M., Dimon MT., Solovyev V., Seledtsov I., Kosarev P., Vorobyev D., Ramirez-Gonzalez R., Leggett R., MacLean D., Xia F., Luo R., Li Z., Xie Y., Liu B., Gnerre S., MacCallum I., Przybylski D., Ribeiro FJ., Yin S., Sharpe T., Hall G., Kersey PJ., Durbin R., Jackman SD., Chapman JA., Huang X., DeRisi JL., Caccamo M., Li Y., Jaffe DB., Green RE., Haussler D., Korf I., Paten B. 2011. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research* 21:2224–2241. DOI: 10.1101/gr.126599.111.

Earle SG., Wu C-H., Charlesworth J., Stoesser N., Claire Gordon N., Walker TM., Spencer CCA., Iqbal Z., Clifton DA., Hopkins KL., Woodford N., Grace Smith E., Ismail N., Llewelyn MJ., Peto TE., Crook DW., McVean G., Sarah Walker A., Wilson DJ. 2016. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology* 1:16041. DOI: 10.1038/nmicrobiol.2016.41.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome informatics. International Conference on Genome Informatics* 23:205–211.

Eisenstein M. 2012. Oxford Nanopore announcement sets sequencing sector abuzz. *Nature biotechnology* 30:295–296. DOI: 10.1038/nbt0412-295.

Ender M., McCallum N., Adhikari R., Berger-Bächi B. 2004. Fitness cost of SCCmec and methicillin resistance levels in Staphylococcus aureus. *Antimicrobial agents and chemotherapy* 48:2295–2297. DOI: 10.1128/AAC.48.6.2295-2297.2004.

Enright MC., Robinson DA., Randle G., Feil EJ., Grundmann H., Spratt BG. 2002. The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). *Proceedings of the National Academy of Sciences of the United States of America* 99:7687–7692. DOI: 10.1073/pnas.122108599.

Entrez Programming Utilities Help 2010. National Center for Biotechnology Information (US).

Eppstein D. 1998. Finding the k Shortest Paths. *SIAM Journal on Computing* 28:652–673. DOI: 10.1137/S0097539795290477.

Ewing B., Green P. 1998. Base-calling of automated sequencer traces using phred. II.

Error probabilities. *Genome research* 8:186–194. DOI: 10.1101/gr.8.3.186.

Ewing B., Hillier L., Wendl MC., Green P. 1998. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome research* 8:175–185. DOI: 10.1101/gr.8.3.175.

Feijao P., Yao H-T., Fornika D., Gardy J., Hsiao W., Chauve C., Chindelevitch L. 2018. MentaLiST - A fast MLST caller for large MLST schemes. *Microbial genomics*. DOI: 10.1099/mgen.0.000146.

Feil EJ., Spratt BG. 2001. Recombination and the population structures of bacterial pathogens. *Annual review of microbiology* 55:561–590. DOI: 10.1146/annurev.micro.55.1.561.

Felsenstein J. 1978. The Number of Evolutionary Trees. *Systematic zoology* 27:27–33. DOI: 10.2307/2412810.

Felsenstein J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution; international journal of organic evolution* 39:783–791. DOI: 10.2307/2408678.

Field D., Garrity G., Gray T., Morrison N., Selengut J., Sterk P., Tatusova T., Thomson N., Allen MJ., Angiuoli SV., Ashburner M., Axelrod N., Baldauf S., Ballard S., Boore J., Cochrane G., Cole J., Dawyndt P., De Vos P., DePamphilis C., Edwards R., Faruque N., Feldman R., Gilbert J., Gilna P., Glöckner FO., Goldstein P., Guralnick R., Haft D., Hancock D., Hermjakob H., Hertz-Fowler C., Hugenholtz P., Joint I., Kagan L., Kane M., Kennedy J., Kowalchuk G., Kottmann R., Kolker E., Kravitz S., Kyrpides N., Leebens-Mack J., Lewis SE., Li K., Lister AL., Lord P., Maltsev N., Markowitz V., Martiny J., Methe B., Mizrachi I., Moxon R., Nelson K., Parkhill J., Proctor L., White O., Sansone S-A., Spiers A., Stevens R., Swift P., Taylor C., Tateno Y., Tett A., Turner S., Ussery D., Vaughan B., Ward N., Whetzel T., San Gil I., Wilson G., Wipat A. 2008. The minimum information about a genome sequence (MIGS) specification. *Nature biotechnology* 26:541–547. DOI: 10.1038/nbt1360.

Fleischmann RD., Adams MD., White O., Clayton RA., Kirkness EF., Kerlavage AR., Bult CJ., Tomb JF., Dougherty BA., Merrick JM. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269:496–512. DOI: 10.1126/science.7542800.

Flores GE., Bates ST., Knights D., Lauber CL., Stombaugh J., Knight R., Fierer N. 2011.

Microbial biogeography of public restroom surfaces. *PloS one* 6:e28132. DOI: 10.1371/journal.pone.0028132.

Foster TJ. 2017. Antibiotic resistance in Staphylococcus aureus. Current status and future prospects. *FEMS microbiology reviews*. DOI: 10.1093/femsre/fux007.

Foster TJ., Geoghegan JA., Ganesh VK., Höök M. 2014. Adhesion, invasion and evasion: the many functions of the surface proteins of Staphylococcus aureus. *Nature reviews. Microbiology* 12:49–62. DOI: 10.1038/nrmicro3161.

Foucault M-L., Courvalin P., Grillot-Courvalin C. 2009. Fitness cost of VanA-type vancomycin resistance in methicillin-resistant Staphylococcus aureus. *Antimicrobial agents and chemotherapy* 53:2354–2359. DOI: 10.1128/AAC.01702-08.

Fournier GP., Gogarten JP. 2008. Evolution of acetoclastic methanogenesis in Methanosarcina via horizontal gene transfer from cellulolytic Clostridia. *Journal of bacteriology* 190:1124–1127. DOI: 10.1128/JB.01382-07.

Fouts DE., Brinkac L., Beck E., Inman J., Sutton G. 2012. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic acids research* 40:e172. DOI: 10.1093/nar/gks757.

Frisch MB., Castillo-Ramírez S., Petit RA., Farley MM., Ray SM., Albrecht VS., Limbago BM., Hernandez J., See I., Satola SW., Read TD. 2018. Invasive Methicillin-Resistant Staphylococcus aureus USA500 Strains from the U.S. Emerging Infections Program Constitute Three Geographically Distinct Lineages. *mSphere* 3:e00571–17. DOI: 10.1128/mSphere.00571-17.

Fuchs S., Mehlan H., Bernhardt J., Hennig A., Michalik S., Surmann K., Pané-Farré J., Giese A., Weiss S., Backert L., Herbig A., Nieselt K., Hecker M., Völker U., Mäder U. 2017. AureoWiki–The repository of the Staphylococcus aureus research and annotation community. *International journal of medical microbiology: IJMM*. DOI: 10.1016/j.ijmm.2017.11.011.

García-Álvarez L., Holden MTG., Lindsay H., Webb CR., Brown DFJ., Curran MD., Walpole E., Brooks K., Pickard DJ., Teale C., Parkhill J., Bentley SD., Edwards GF., Girvan EK., Kearns AM., Pichon B., Hill RLR., Larsen AR., Skov RL., Peacock SJ., Maskell DJ., Holmes MA. 2011. Meticillin-resistant Staphylococcus aureus with a

novel mecA homologue in human and bovine populations in the UK and Denmark: a descriptive study. *The Lancet infectious diseases* 11:595–603. DOI: 10.1016/S1473-3099(11)70126-8.

Gardner SN., Slezak T., Hall BG. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* . DOI: 10.1093/bioinformatics/btv271.

Gillespie JH. 2010. *Population genetics: a concise guide*. JHU Press.

Girvan M., Newman MEJ. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* 99:7821–7826. DOI: 10.1073/pnas.122653799.

Glaser P., Martins-Simões P., Villain A., Barbier M., Tristan A., Bouchier C., Ma L., Bes M., Laurent F., Guillemot D., Wirth T., Vandenesch F. 2016. Demography and Intercontinental Spread of the USA300 Community-Acquired Methicillin-Resistant Staphylococcus aureus Lineage. *mBio* 7:e02183–15. DOI: 10.1128/mBio.02183-15.

Gordon A., Hannon GJ. 2010. Fastx-toolkit. *Computer program distributed by the author, website http://hannonlab. cshl. edu/fastx_toolkit/index. html [accessed 2014--2015]*.

Gordon NC., Price JR., Cole K., Everitt R., Morgan M., Finney J., Kearns AM., Pichon B., Young B., Wilson DJ., Llewelyn MJ., Paul J., Peto TEA., Crook DW., Walker AS., Golubchik T. 2014. Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing. *Journal of clinical microbiology* 52:1182–1191. DOI: 10.1128/JCM.03117-13.

Gorwitz RJ., Kruszon-Moran D., McAllister SK., McQuillan G., McDougal LK., Fosheim GE., Jensen BJ., Killgore G., Tenover FC., Kuehnert MJ. 2008. Changes in the prevalence of nasal colonization with Staphylococcus aureus in the United States, 2001-2004. *The Journal of infectious diseases* 197:1226–1234. DOI: 10.1086/533494.

Griffith F. 1928. The Significance of Pneumococcal Types. *The Journal of hygiene* 27:113–159.

Grüning B., Dale R., Sjödin A., Rowe J., Chapman BA., Tomkins-Tinch CH., Valieris R., The Bioconda Team., Köster J. 2017. Bioconda: A sustainable and comprehensive software distribution for the life sciences. *bioRxiv*:207092. DOI: 10.1101/207092.

Gupta SK., Padmanabhan BR., Diene SM., Lopez-Rojas R., Kempf M., Landraud L., Rolain J-M. 2014. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial agents and chemotherapy* 58:212–220. DOI: 10.1128/AAC.01310-13.

Hacker J., Blum-Oehler G., Mühldorfer I., Tschäpe H. 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Molecular microbiology* 23:1089–1097. DOI: 10.1046/j.1365-2958.1997.3101672.x.

Haque MM., Bose T., Dutta A., Reddy CVSK., Mande SS. 2015. CS-SCORE: Rapid identification and removal of human genome contaminants from metagenomic datasets. *Genomics* 106:116–121. DOI: 10.1016/j.ygeno.2015.04.005.

Harkins CP., Pichon B., Doumith M., Parkhill J., Westh H., Tomasz A., de Lencastre H., Bentley SD., Kearns AM., Holden MTG. 2017. Methicillin-resistant Staphylococcus aureus emerged long before the introduction of methicillin into clinical practice. *Genome biology* 18:130. DOI: 10.1186/s13059-017-1252-9.

Harris SR., Feil EJ., Holden MTG., Quail MA., Nickerson EK., Chantratita N., Gardete S., Tavares A., Day N., Lindsay JA., Edgeworth JD., de Lencastre H., Parkhill J., Peacock SJ., Bentley SD. 2010. Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science* 327:469–474. DOI: 10.1126/science.1182395.

Head SR., Komori HK., LaMere SA., Whisenant T., Van Nieuwerburgh F., Salomon DR., Ordoukhanian P. 2014. Library construction for next-generation sequencing: overviews and challenges. *BioTechniques* 56:61–4, 66, 68, passim. DOI: 10.2144/000114133.

Helgason E., Okstad OA., Caugant DA., Johansen HA., Fouet A., Mock M., Hegna I., Kolstø AB. 2000. Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis--one species on the basis of genetic evidence. *Applied and environmental microbiology* 66:2627–2630.

Heo Y., Ramachandran A., Hwu W-M., Ma J., Chen D. 2016. BLESS 2: accurate, memory-efficient and fast error correction method. *Bioinformatics* 32:2369–2371. DOI: 10.1093/bioinformatics/btw146.

Hiramatsu K., Hanaki H., Ino T., Yabuta K., Oguri T., Tenover FC. 1997. Methicillin-resistant Staphylococcus aureus clinical strain with reduced vancomycin susceptibility. *The Journal of antimicrobial chemotherapy* 40:135–136.

Hirwade MA. 2011. A study of metadata standards. *Library Hi Tech News* 28:18–25. DOI: 10.1108/07419051111184052.

Hoang DT., Chernomor O., von Haeseler A., Minh BQ., Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular biology and evolution* 35:518–522. DOI: 10.1093/molbev/msx281.

Hoffmann C., Zimmermann F., Biek R., Kuehl H., Nowak K., Mundry R., Agbor A., Angedakin S., Arandjelovic M., Blankenburg A., Brazolla G., Corogenes K., Couacy-Hymann E., Deschner T., Dieguez P., Dierks K., Düx A., Dupke S., Eshuis H., Formenty P., Yuh YG., Goedmakers A., Gogarten JF., Granjon A-C., McGraw S., Grunow R., Hart J., Jones S., Junker J., Kiang J., Langergraber K., Lapuente J., Lee K., Leendertz SA., Léguillon F., Leinert V., Löhrich T., Marrocoli S., Mätz-Rensing K., Meier A., Merkel K., Metzger S., Murai M., Niedorf S., De Nys H., Sachse A., van Schijndel J., Thiesen U., Ton E., Wu D., Wieler LH., Boesch C., Klee SR., Wittig RM., Calvignac-Spencer S., Leendertz FH. 2017. Persistent anthrax as a major driver of wildlife mortality in a tropical rainforest. *Nature* 548:82–86. DOI: 10.1038/nature23309.

Hoffmaster AR., Ravel J., Rasko DA., Chapman GD., Chute MD., Marston CK., De BK., Sacchi CT., Fitzgerald C., Mayer LW., Maiden MCJ., Priest FG., Barker M., Jiang L., Cer RZ., Rilstone J., Peterson SN., Weyant RS., Galloway DR., Read TD., Popovic T., Fraser-Liggett CM. 2004. Identification of anthrax toxin genes in a Bacillus cereus associated with an illness resembling inhalation anthrax. *Proceedings of the National Academy of Sciences of the United States of America* 101:8449–8454. DOI: 10.1073/pnas.0402414101.

Hogan JM., Holland P., Holloway AP., Petit RA III., Read TD. 2013. Read Classification for Next Generation Sequencing. *ESANN 2013 proceedings: European Symposium on Artificial Neural Networks, Computational Intelligence*.

Holt DC., Holden MTG., Tong SYC., Castillo-Ramirez S., Clarke L., Quail MA., Currie BJ., Parkhill J., Bentley SD., Feil EJ., Giffard PM. 2011. A Very Early-Branching Staphylococcus aureus Lineage Lacking the Carotenoid Pigment Staphyloxanthin. *Genome biology and evolution* 3:881–895. DOI: 10.1093/gbe/evr078.

Huang W., Li L., Myers JR., Marth GT. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* 28:593–594. DOI: 10.1093/bioinformatics/btr708.

Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–214. DOI: 10.1038/nature11234.

Hunt M., Mather AE., Sánchez-Busó L., Page AJ., Parkhill J., Keane JA., Harris SR. 2017. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial genomics* 3:e000131. DOI: 10.1099/mgen.0.000131.

Hyatt D., Chen G-L., LoCascio P., Land M., Larimer F., Hauser L. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* 11:119. DOI: 10.1186/1471-2105-11-119.

Inouye M., Dashnow H., Raven L-A., Schultz MB., Pope BJ., Tomita T., Zobel J., Holt KE. 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome medicine* 6:90. DOI: 10.1186/s13073-014-0090-6.

Ito T., Katayama Y., Asada K., Mori N., Tsutsumimoto K., Tiensasitorn C., Hiramatsu K. 2001. Structural comparison of three types of staphylococcal cassette chromosome mec integrated in the chromosome in methicillin-resistant Staphylococcus aureus. *Antimicrobial agents and chemotherapy* 45:1323–1336. DOI: 10.1128/AAC.45.5.1323-1336.2001.

Jain R., Rivera MC., Moore JE., Lake JA. 2002. Horizontal gene transfer in microbial genome evolution. *Theoretical population biology* 61:489–495.

Jay E., Bambara R., Padmanabhan R., Wu R. 1974. DNA sequence analysis: a general, simple and rapid method for sequencing large oligodeoxyribonucleotide fragments by mapping. *Nucleic acids research* 1:331–353.

Jia B., Raphenya AR., Alcock B., Waglechner N., Guo P., Tsang KK., Lago BA., Dave BM., Pereira S., Sharma AN., Doshi S., Courtot M., Lo R., Williams LE., Frye JG., Elsayegh T., Sardar D., Westman EL., Pawlowski AC., Johnson TA., Brinkman FSL., Wright GD., McArthur AG. 2017. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic acids research* 45:D566–D573. DOI: 10.1093/nar/gkw1004.

Joensen KG., Scheutz F., Lund O., Hasman H., Kaas RS., Nielsen EM., Aarestrup FM. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *Journal of clinical microbiology* 52:1501–1510. DOI: 10.1128/JCM.03617-13.

Jolley KA., Maiden MCJ. 2010. BIGSdb: Scalable analysis of bacterial genome variation

at the population level. *BMC bioinformatics* 11:595. DOI: 10.1186/1471-2105-11-595.

Joseph SJ., Li B., Petit RA Iii., Qin ZS., Darrow L., Read TD. 2016. The single-species metagenome: subtyping Staphylococcus aureus core genome sequences from shotgun metagenomic data. *PeerJ* 4:e2571. DOI: 10.7717/peerj.2571.

Karlin S. 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. *Current opinion in microbiology* 1:598–610.

Katayama Y., Ito T., Hiramatsu K. 2000. A new class of genetic element, staphylococcus cassette chromosome mec, encodes methicillin resistance in Staphylococcus aureus. *Antimicrobial agents and chemotherapy* 44:1549–1555.

Kawano J., Shimizu A., Saitoh Y., Yagi M., Saito T., Okamoto R. 1996. Isolation of methicillin-resistant coagulase-negative staphylococci from chickens. *Journal of clinical microbiology* 34:2072–2077.

Kaya H., Hasman H., Larsen J., Stegger M., Johannesen TB., Allesøe RL., Lemvigh CK., Aarestrup FM., Lund O., Larsen AR. 2018. SCCmecFinder, a Web-Based Tool for Typing of Staphylococcal Cassette ChromosomemecinStaphylococcus aureusUsing Whole-Genome Sequence Data. *mSphere* 3. DOI: 10.1128/mSphere.00612-17.

Keim PS., Wagner DM. 2009. Humans and evolutionary and ecological forces shaped the phylogeography of recently emerged diseases. *Nature reviews. Microbiology* 7:813–821. DOI: 10.1038/nrmicro2219.

Kelley DR., Schatz MC., Salzberg SL. 2010. Quake: quality-aware detection and correction of sequencing errors. *Genome biology* 11:R116. DOI: 10.1186/gb-2010-11-11-r116.

Kemena C., Notredame C. 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* 25:2455–2465. DOI: 10.1093/bioinformatics/btp452.

Klee SR., Brzuszkiewicz EB., Nattermann H., Brüggemann H., Dupke S., Wollherr A., Franz T., Pauli G., Appel B., Liebl W., Couacy-Hymann E., Boesch C., Meyer F-D., Leendertz FH., Ellerbrok H., Gottschalk G., Grunow R., Liesegang H. 2010. The genome of a Bacillus isolate causing anthrax in chimpanzees combines chromosomal properties of B. cereus with B. anthracis virulence plasmids. *PloS one* 5:e10986. DOI: 10.1371/journal.pone.0010986.

Klinkenberg D., Backer JA., Didelot X., Colijn C., Wallinga J. 2017. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS computational biology* 13:e1005495. DOI: 10.1371/journal.pcbi.1005495.

Kloos WE. 1980. Natural Populations of the Genus Staphylococcus. *Annual review of microbiology* 34:559–592. DOI: 10.1146/annurev.mi.34.100180.003015.

Kluytmans J., van Belkum A., Verbrugh H. 1997. Nasal carriage of Staphylococcus aureus: epidemiology, underlying mechanisms, and associated risks. *Clinical microbiology reviews* 10:505–520.

Koboldt DC., Zhang Q., Larson DE., Shen D., McLellan MD., Lin L., Miller CA., Mardis ER., Ding L., Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* 22:568–576. DOI: 10.1101/gr.129684.111.

Kolbe DL., Eddy SR. 2011. Fast filtering for RNA homology search. *Bioinformatics* 27:3102–3109. DOI: 10.1093/bioinformatics/btr545.

Kondo Y., Ito T., Ma XX., Watanabe S., Kreiswirth BN., Etienne J., Hiramatsu K. 2007. Combination of multiplex PCRs for staphylococcal cassette chromosome mec type assignment: rapid identification system for mec, ccr, and major differences in junkyard regions. *Antimicrobial agents and chemotherapy* 51:264–274. DOI: 10.1128/AAC.00165-06.

Konstantinidis KT., Tiedje JM. 2005a. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America* 102:2567–2572. DOI: 10.1073/pnas.0409727102.

Koonin EV., Makarova KS. 2001. Horizontal gene transfer in prokaryotes: quantification and classification 1. *Annual reviews in control*.

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 research*. DOI: 10.1101/gr.215087.116.

Kos VN., Desjardins CA., Griggs A., Cerqueira G., Van Tonder A., Holden MTG., Godfrey P., Palmer KL., Bodi K., Mongodin EF., Wortman J., Feldgarden M., Lawley T., Gill SR., Haas BJ., Birren B., Gilmore MS. 2012. Comparative genomics of vancomycin-resistant Staphylococcus aureus strains and their positions within the clade most commonly associated with Methicillin-resistant S. aureus hospital-acquired

infection in the United States. *mBio* 3. DOI: 10.1128/mBio.00112-12.

Koski LB., Golding GB. 2001. The closest BLAST hit is often not the nearest neighbor. *Journal of molecular evolution* 52:540–542. DOI: 10.1007/s002390010184.

Koslicki D., Falush D. 2016. MetaPalette: a k-mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation. *mSystems* 1. DOI: 10.1128/mSystems.00020-16.

Kuroda M., Ohta T., Uchiyama I., Baba T., Yuzawa H., Kobayashi I., Cui L., Oguchi A., Aoki K., Nagai Y., Lian J., Ito T., Kanamori M., Matsumaru H., Maruyama A., Murakami H., Hosoyama A., Mizutani-Ui Y., Takahashi NK., Sawano T., Inoue R., Kaito C., Sekimizu K., Hirakawa H., Kuhara S., Goto S., Yabuzaki J., Kanehisa M., Yamashita A., Oshima K., Furuya K., Yoshino C., Shiba T., Hattori M., Ogasawara N., Hayashi H., Hiramatsu K. 2001. Whole genome sequencing of meticillin-resistant Staphylococcus aureus. *The Lancet* 357:1225–1240.

Lagesen K., Hallin P., Rødland EA., Staerfeldt H-H., Rognes T., Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research* 35:3100–3108. DOI: 10.1093/nar/gkm160.

Lakin SM., Dean C., Noyes NR., Dettenwanger A., Ross AS., Doster E., Rovira P., Abdo Z., Jones KL., Ruiz J., Belk KE., Morley PS., Boucher C. 2017a. MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic acids research* 45:D574–D580. DOI: 10.1093/nar/gkw1009.

Lakin SM., Dean C., Noyes NR., Dettenwanger A., Ross AS., Doster E., Rovira P., Abdo Z., Jones KL., Ruiz J., Belk KE., Morley PS., Boucher C. 2017b. MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic acids research* 45:D574–D580. DOI: 10.1093/nar/gkw1009.

Lam KN., Cheng J., Engel K., Neufeld JD., Charles TC. 2015. Current and future resources for functional metagenomics. *Frontiers in microbiology* 6:1196. DOI: 10.3389/fmicb.2015.01196.

Langmead B., Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* 9:357–359. DOI: 10.1038/nmeth.1923.

Laslett D., Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic acids research* 32:11–16. DOI: 10.1093/nar/gkh152.

Lawrence JG., Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *Journal of molecular evolution* 44:383–397.

Lawrence JG., Ochman H. 1998. Molecular archaeology of the Escherichia coli genome. *Proceedings of the National Academy of Sciences of the United States of America* 95:9413–9417.

Lees JA., Kendall M., Parkhill J., Colijn C., Bentley SD., Harris SR. 2018. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome Open Research* 3:33. DOI: 10.12688/wellcomeopenres.14265.1.

Lees JA., Vehkala M., Välimäki N., Harris SR., Chewapreecha C., Croucher NJ., Marttinen P., Honkela A., Parkhill J., Bentley SD., Corander J. 2016. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *bioRxiv*:038463. DOI: 10.1101/038463.

Leopold SR., Goering RV., Witten A., Harmsen D., Mellmann A. 2014. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. *Journal of clinical microbiology* 52:2365–2370. DOI: 10.1128/JCM.00262-14.

Letunic I., Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research* 44:W242–5. DOI: 10.1093/nar/gkw290.

Li H., Durbin R. 2009a. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760. DOI: 10.1093/bioinformatics/btp324.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. DOI: 10.1093/bioinformatics/btp352.

Li D., Liu C-M., Luo R., Sadakane K., Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31:1674–1676. DOI: 10.1093/bioinformatics/btv033.

Limbago BM., Kallen AJ., Zhu W., Eggers P., McDougal LK., Albrecht VS. 2014a. Report of the 13th vancomycin-resistant Staphylococcus aureus isolate from the United States. *Journal of clinical microbiology* 52:998–1002. DOI: 10.1128/JCM.02187-

13.

Ling LL., Schneider T., Peoples AJ., Spoering AL., Engels I., Conlon BP., Mueller A., Schäberle TF., Hughes DE., Epstein S., Jones M., Lazarides L., Steadman VA., Cohen DR., Felix CR., Fetterman KA., Millett WP., Nitti AG., Zullo AM., Chen C., Lewis K. 2015. A new antibiotic kills pathogens without detectable resistance. *Nature*. DOI: 10.1038/nature14098.

Loeffler A., McCarthy A., Lloyd DH., Musilová E., Pfeiffer DU., Lindsay JA. 2013. Whole-genome comparison of meticillin-resistant Staphylococcus aureus CC22 SCCmecIV from people and their in-contact pets. *Veterinary dermatology* 24:538–e128. DOI: 10.1111/vde.12062.

Loman NJ., Pallen MJ. 2015. Twenty years of bacterial genome sequencing. *Nature reviews. Microbiology* 13:787–794. DOI: 10.1038/nrmicro3565.

Loman NJ., Quick J., Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods* 12:733–735. DOI: 10.1038/nmeth.3444.

Lu J., Salzberg S. 2018. Removing Contaminants from Metagenomic Databases. *bioRxiv*:261859. DOI: 10.1101/261859.

Maiden MC., Bygraves JA., Feil E., Morelli G., Russell JE., Urwin R., Zhang Q., Zhou J., Zurth K., Caugant DA., Feavers IM., Achtman M., Spratt BG. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* 95:3140–3145.

Ma XX., Ito T., Tiensasitorn C., Jamklang M., Chongtrakool P., Boyle-Vavra S., Daum RS., Hiramatsu K. 2002. Novel type of staphylococcal cassette chromosome mec identified in community-acquired methicillin-resistant Staphylococcus aureus strains. *Antimicrobial agents and chemotherapy* 46:1147–1152. DOI: 10.1128/AAC.46.4.1147-1152.2002.

Marçais G., Kingsford C. 2011a. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764–770. DOI: 10.1093/bioinformatics/btr011.

Marraffini LA., Sontheimer EJ. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* 322:1843–1845. DOI:

10.1126/science.1165771.

Mason C. 2015.The long road from Data to Wisdom, and from DNA to Pathogen. *Available at https://www.microbe.net/2015/02/17/the-long-road-from-data-to-wisdom-and-from-dna-to-pathogen/* (accessed December 18, 2017).

Mason OU., Hazen TC., Borglin S., Chain PSG., Dubinsky EA., Fortney JL., Han J., Holman H-YN., Hultman J., Lamendella R., Mackelprang R., Malfatti S., Tom LM., Tringe SG., Woyke T., Zhou J., Rubin EM., Jansson JK. 2012. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *The ISME journal* 6:1715–1727. DOI: 10.1038/ismej.2012.59.

Matic I., Rayssiguier C., Radman M. 1995. Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. *Cell* 80:507–515.

Maxwell KL. 2016. Phages Fight Back: Inactivation of the CRISPR-Cas Bacterial Immune System by Anti-CRISPR Proteins. *PLoS pathogens* 12:e1005282. DOI: 10.1371/journal.ppat.1005282.

McDaniel LD., Young E., Delaney J., Ruhnau F., Ritchie KB., Paul JH. 2010. High frequency of horizontal gene transfer in the oceans. *Science* 330:50. DOI: 10.1126/science.1192243.

McDonald D., Hyde E., Debelius JW., Morton JT., Gonzalez A., Ackermann G., Aksenov AA., Behsaz B., Brennan C., Chen Y., DeRight Goldasich L., Dorrestein PC., Dunn RR., Fahimipour AK., Gaffney J., Gilbert JA., Gogul G., Green JL., Hugenholtz P., Humphrey G., Huttenhower C., Jackson MA., Janssen S., Jeste DV., Jiang L., Kelley ST., Knights D., Kosciolek T., Ladau J., Leach J., Marotz C., Meleshko D., Melnik AV., Metcalf JL., Mohimani H., Montassier E., Navas-Molina J., Nguyen TT., Peddada S., Pevzner P., Pollard KS., Rahnavard G., Robbins-Pianka A., Sangwan N., Shorenstein J., Smarr L., Song SJ., Spector T., Swafford AD., Thackray VG., Thompson LR., Tripathi A., Vázquez-Baeza Y., Vrbanac A., Wischmeyer P., Wolfe E., Zhu Q., American Gut Consortium., Knight R. 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* 3. DOI: 10.1128/mSystems.00031-18.

McIntyre A., Ounit R., Afshinnekoo E., Prill R., Henaff E., Alexander N., Minot S., Danko D., Foox J., Ahsanuddin S., Tighe S., Hasan NA., Subramanian P., Moffat K.,

Levy S., Lonardi S., Greenfield N., Colwell R., Rosen G., Mason CE. 2017. Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers. *bioRxiv*:156919. DOI: 10.1101/156919.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20:1297–1303. DOI: 10.1101/gr.107524.110.

McQuilton P., Gonzalez-Beltran A., Rocca-Serra P., Thurston M., Lister A., Maguire E., Sansone S-A. 2016. BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database: the journal of biological databases and curation* 2016. DOI: 10.1093/database/baw075.

Medini D., Donati C., Tettelin H., Masignani V., Rappuoli R. 2005. The microbial pan-genome. *Current opinion in genetics & development* 15:589–594. DOI: 10.1016/j.gde.2005.09.006.

Melo-Cristino J., Resina C., Manuel V., Lito L., Ramirez M. 2013. First case of infection with vancomycin-resistant Staphylococcus aureus in Europe. *The Lancet* 382:205. DOI: 10.1016/S0140-6736(13)61219-2.

Michod RE., Wojciechowski MF., Hoelzer MA. 1988. DNA repair and the evolution of transformation in the bacterium Bacillus subtilis. *Genetics* 118:31–39.

Miller LG., Diep BA. 2008. Clinical practice: colonization, fomites, and virulence: rethinking the pathogenesis of community-associated methicillin-resistant Staphylococcus aureus infection. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* 46:752–760. DOI: 10.1086/526773.

Minot SS., Greenfield N., Afshinnekoo E., Mason CE. 2015.Anthrax Marker Panel. *Available at https://science.onecodex.com/bacillus-anthracis-panel/* (accessed December 19, 2017).

Moran Y., Fredman D., Szczesny P., Grynberg M., Technau U. 2012. Recurrent horizontal transfer of bacterial toxin genes to eukaryotes. *Molecular biology and evolution* 29:2223–2230. DOI: 10.1093/molbev/mss089.

Moravvej Z., Estaji F., Askari E., Solhjou K., Naderi Nasab M., Saadat S. 2013. Update on the global number of vancomycin-resistant Staphylococcus aureus (VRSA)

strains. *International journal of antimicrobial agents* 42:370–371. DOI: 10.1016/j.ijantimicag.2013.06.004.

Mrázek J., Karlin S. 1999. Detecting alien genes in bacterial genomes. *Annals of the New York Academy of Sciences* 870:314–329.

Nagarajan N., Pop M. 2013. Sequence assembly demystified. *Nature reviews. Genetics* 14:157–167. DOI: 10.1038/nrg3367.

Nakamura Y., Itoh T., Matsuda H., Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature genetics* 36:760–766. DOI: 10.1038/ng1381.

Nascimento FF., Reis MD., Yang Z. 2017. A biologist's guide to Bayesian phylogenetic analysis. *Nature ecology & evolution* 1:1446–1454. DOI: 10.1038/s41559-017-0280-x.

Nasko DJ., Koren S., Phillippy AM., Treangen TJ. 2018a. RefSeq database growth influences the accuracy of k-mer-based species identification. *bioRxiv*:304972. DOI: 10.1101/304972.

Nasko DJ., Koren S., Phillippy AM., Treangen TJ. 2018b. RefSeq database growth influences the accuracy of k-mer-based species identification. *bioRxiv*:304972. DOI: 10.1101/304972.

Nelson KE., Clayton RA., Gill SR., Gwinn ML., Dodson RJ., Haft DH., Hickey EK., Peterson JD., Nelson WC., Ketchum KA., McDonald L., Utterback TR., Malek JA., Linher KD., Garrett MM., Stewart AM., Cotton MD., Pratt MS., Phillips CA., Richardson D., Heidelberg J., Sutton GG., Fleischmann RD., Eisen JA., White O., Salzberg SL., Smith HO., Venter JC., Fraser CM. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* 399:323–329. DOI: 10.1038/20601.

Nguyen N., Hickey G., Zerbino DR., Raney B., Earl D., Armstrong J., Kent WJ., Haussler D., Paten B. 2015a. Building a pan-genome reference for a population. *Journal of computational biology: a journal of computational molecular cell biology* 22:387–401. DOI: 10.1089/cmb.2014.0146.

Nguyen L-T., Schmidt HA., von Haeseler A., Minh BQ. 2015b. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* 32:268–274. DOI: 10.1093/molbev/msu300.

NIH Human Microbiome Project - Publications. *Available at https://hmpdacc.org/hmp/publications.php* (accessed May 29, 2018).

Noto MJ., Kreiswirth BN., Monk AB., Archer GL. 2008. Gene acquisition at the insertion site for SCCmec, the genomic island conferring methicillin resistance in Staphylococcus aureus. *Journal of bacteriology* 190:1276–1283. DOI: 10.1128/JB.01128-07.

Nurk S., Meleshko D., Korobeynikov A., Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome research* 27:824–834. DOI: 10.1101/gr.213959.116.

Oliveira PH., Touchon M., Cury J., Rocha EPC. 2017. The chromosomal organization of horizontal gene transfer in bacteria. *Nature communications* 8:841. DOI: 10.1038/s41467-017-00808-w.

Ondov BD., Treangen TJ., Melsted P., Mallonee AB., Bergman NH., Koren S., Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome biology* 17:132. DOI: 10.1186/s13059-016-0997-x.

O'Neill J. 2016. Tackling drug-resistant infections globally: Final report and recommendations. The review on antimicrobial resistance. *London: HM Government and the Wellcome Trust*.

van Oosten M., Hahn M., Crane LMA., Pleijhuis RG., Francis KP., van Dijl JM., van Dam GM. 2015. Targeted imaging of bacterial infections: advances, hurdles and hopes. *FEMS microbiology reviews* 39:892–916. DOI: 10.1093/femsre/fuv029.

Page AJ., Cummins CA., Hunt M., Wong VK., Reuter S., Holden MTG., Fookes M., Falush D., Keane JA., Parkhill J. 2015a. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* . DOI: 10.1093/bioinformatics/btv421.

Pakyz AL., MacDougall C., Oinonen M., Polk RE. 2008. Trends in antibacterial use in US academic health centers: 2002 to 2006. *Archives of internal medicine* 168:2254–2260. DOI: 10.1001/archinte.168.20.2254.

Palmer KL., Kos VN., Gilmore MS. 2010. Horizontal gene transfer and the genomics of enterococcal antibiotic resistance. *Current opinion in microbiology* 13:632–639. DOI: 10.1016/j.mib.2010.08.004.

Pál C., Papp B., Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature genetics* 37:1372–1375. DOI: 10.1038/ng1686.

Pannucci J., Okinaka RT., Williams E., Sabin R., Ticknor LO., Kuske CR. 2002. DNA sequence conservation between the Bacillus anthracis pXO2 plasmid and genomic sequence from closely related bacteria. *BMC genomics* 3:34.

Patricia Jevons M. 1961. “Celbenin” - resistant Staphylococci. *British medical journal* 1:124.

Payne A., Holmes N., Rakyan V., Loose M. 2018. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv*:312256. DOI: 10.1101/312256.

Peng Y., Leung HCM., Yiu SM., Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–1428. DOI: 10.1093/bioinformatics/bts174.

Périchon B., Courvalin P. 2004. Heterologous expression of the enterococcal vanA operon in methicillin-resistant Staphylococcus aureus. *Antimicrobial agents and chemotherapy* 48:4281–4285. DOI: 10.1128/AAC.48.11.4281-4285.2004.

Périchon B., Courvalin P. 2006. Synergism between beta-lactams and glycopeptides against VanA-type methicillin-resistant Staphylococcus aureus and heterologous expression of the vanA operon. *Antimicrobial agents and chemotherapy* 50:3622–3630. DOI: 10.1128/AAC.00410-06.

Périchon B., Courvalin P. 2009. VanA-type vancomycin-resistant Staphylococcus aureus. *Antimicrobial agents and chemotherapy* 53:4580–4587. DOI: 10.1128/AAC.00346-09.

Petit RA III., Ezewudo M., Joseph SJ., Read TD. 2015.Searching for anthrax in the New York City subway metagenome. *Available at http://dx.doi.org/10.5281/zenodo.17158* (accessed December 18, 2017). DOI: 10.5281/zenodo.17158.

Pinedo CA., Smets BF. 2005. Conjugal TOL transfer from Pseudomonas putida to Pseudomonas aeruginosa: effects of restriction proficiency, toxicant exposure, cell density ratios, and conjugation detection method on observed transfer efficiencies. *Applied and environmental microbiology* 71:51–57.

Planet PJ., Narechania A., Chen L., Mathema B., Boundy S., Archer G., Kreiswirth B. 2016. Architecture of a Species: Phylogenomics of Staphylococcus aureus. *Trends in microbiology*. DOI: 10.1016/j.tim.2016.09.009.

Popa O., Hazkani-Covo E., Landan G., Martin W., Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome research* 21:599–609. DOI: 10.1101/gr.115592.110.

Price MN., Dehal PS., Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution* 26:1641–1650. DOI: 10.1093/molbev/msp077.

Price LB., Stegger M., Hasman H., Aziz M., Larsen J., Andersen PS., Pearson T., Waters AE., Foster JT., Schupp J., Gillece J., Driebe E., Liu CM., Springer B., Zdovc I., Battisti A., Franco A., Zmudzki J., Schwarz S., Butaye P., Jouy E., Pomba C., Porrero MC., Ruimy R., Smith TC., Robinson DA., Weese JS., Arriola CS., Yu F., Laurent F., Keim P., Skov R., Aarestrup FM. 2012. Staphylococcus aureus CC398: host adaptation and emergence of methicillin resistance in livestock. *mBio* 3. DOI: 10.1128/mBio.00305-11.

Pritchard JK., Stephens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira MAR., Bender D., Maller J., Sklar P., de Bakker PIW., Daly MJ., Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81:559–575. DOI: 10.1086/519795.

Qin J., Li R., Raes J., Arumugam M., Burgdorf KS., Manichanh C., Nielsen T., Pons N., Levenez F., Yamada T., Mende DR., Li J., Xu J., Li S., Li D., Cao J., Wang B., Liang H., Zheng H., Xie Y., Tap J., Lepage P., Bertalan M., Batto J-M., Hansen T., Le Paslier D., Linneberg A., Bjørn Nielsen H., Pelletier E., Renault P., Sicheritz-Ponten T., Turner K., Zhu H., Yu C., Li S., Jian M., Zhou Y., Li Y., Zhang X., Li S., Qin N., Yang H., Wang J., Brunak S., Doré J., Guarner F., Kristiansen K., Pedersen O., Parkhill J., Weissenbach J., Antolin M., Artiguenave F., Blottiere H., Borruel N., Bruls T., Casellas F., Chervaux C., Cultrone A., Delorme C., Denariaz G., Dervyn R., Forte M., Friss C., van de Guchte M., Guedon E., Haimet F., Jamet A., Juste C., Kaci G., Kleerebezem M., Knol J., Kristensen M., Layec S., Le Roux K., Leclerc M., Maguin E., Minardi RM., Oozeer R., Rescigno M., Sanchez N., Tims S., Torrejon T., Varela E., de Vos W., Winogradsky Y., Zoetendal E., Bork P., Dusko Ehrlich S., Wang J. 2010. A human gut microbial gene catalogue established by metagenomic

sequencing. *Nature* 464:59–65. DOI: 10.1038/nature08821.

Quast C., Pruesse E., Yilmaz P., Gerken J., Schweer T., Yarza P., Peplies J., Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research* 41:D590–6. DOI: 10.1093/nar/gks1219.

Quevillon E., Silventoinen V., Pillai S., Harte N., Mulder N., Apweiler R., Lopez R. 2005. InterProScan: protein domains identifier. *Nucleic acids research* 33:W116–20. DOI: 10.1093/nar/gki442.

Quinlan AR., Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842. DOI: 10.1093/bioinformatics/btq033.

Radford JQ., Robinson D., Watson J. 2009. Ecological processes: a key element in strategies for nature conservation. *Ecological*.

Rasko DA., Rosovitz MJ., Økstad OA., Fouts DE., Jiang L., Cer RZ., Kolstø A-B., Gill SR., Ravel J. 2007. Complete sequence analysis of novel plasmids from emetic and periodontal Bacillus cereus isolates reveals a common evolutionary history among the B. cereus-group plasmids, including Bacillus anthracis pXO1. *Journal of bacteriology* 189:52–64. DOI: 10.1128/JB.01313-06.

Ravenhall M., Škunca N., Lassalle F., Dessimoz C. Inferring horizontal gene transfer.

Read TD., Petit RA 3rd., Joseph SJ., Alam MT., Weil MR., Ahmad M., Bhimani R., Vuong JS., Haase CP., Webb DH., Tan M., Dove ADM. 2017a. Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: Rhincodon typus Smith 1828. *BMC genomics* 18:532. DOI: 10.1186/s12864-017-3926-9.

Read TD., Petit RA., Yin Z., Montgomery T., McNulty MC., David MZ. 2017b. USA300 MRSA lineages persist on multiple body sites following infection. *bioRxiv*:192096. DOI: 10.1101/192096.

Read TD., Salzberg SL., Pop M., Shumway M., Umayam L., Jiang L., Holtzapple E., Busch JD., Smith KL., Schupp JM., Solomon D., Keim P., Fraser CM. 2002. Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. *Science* 296:2028–2033. DOI: 10.1126/science.1071837.

Read TD., Thomas AT., Wilkins BM. 1992. Evasion of type I and type II DNA restriction systems by IncI1 plasmid CoIIb-P9 during transfer by bacterial conjugation.

Molecular microbiology 6:1933–1941.

Retchless AC., Lawrence JG. 2012. Ecological adaptation in bacteria: speciation driven by codon selection. *Molecular biology and evolution* 29:3669–3683. DOI: 10.1093/molbev/mss171.

Rhoads A., Au KF. 2015. PacBio Sequencing and Its Applications. *Genomics, proteomics & bioinformatics* 13:278–289. DOI: 10.1016/j.gpb.2015.08.002.

Riley MA., Lizotte-Waniewski M. 2009. Population genomics and the bacterial species concept. *Methods in molecular biology* 532:367–377. DOI: 10.1007/978-1-60327-853-9_21.

Rishishwar L., Petit RA 3rd., Kraft CS., Jordan IK. 2013. Genome sequence-based discriminator for vancomycin-intermediate Staphylococcus aureus. *Journal of bacteriology* 196:940–948. DOI: 10.1128/JB.01410-13.

Rocha EPC. 2018. Neutral theory, microbial practice: challenges in bacterial population genetics. *Molecular biology and evolution*. DOI: 10.1093/molbev/msy078.

Rolo J., Worning P., Nielsen JB., Sobral R., Bowden R., Bouchami O., Damborg P., Guardabassi L., Perreten V., Westh H., Tomasz A., de Lencastre H., Miragaia M. 2017. Evidence for the evolutionary steps leading to mecA-mediated β-lactam resistance in staphylococci. *PLoS genetics* 13:e1006674. DOI: 10.1371/journal.pgen.1006674.

Ronaghi M., Karamohamed S., Pettersson B., Uhlén M., Nyrén P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* 242:84–89. DOI: 10.1006/abio.1996.0432.

Ronquist F., Teslenko M., van der Mark P., Ayres DL., Darling A., Höhna S., Larget B., Liu L., Suchard MA., Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* 61:539–542. DOI: 10.1093/sysbio/sys029.

Rossi F., Diaz L., Wollam A., Panesso D., Zhou Y., Rincon S., Narechania A., Xing G., Di Gioia TSR., Doi A., Tran TT., Reyes J., Munita JM., Carvajal LP., Hernandez-Roldan A., Brandão D., van der Heijden IM., Murray BE., Planet PJ., Weinstock GM., Arias CA. 2014. Transferable vancomycin resistance in a community-associated MRSA lineage. *The New England journal of medicine* 370:1524–1531. DOI: 10.1056/NEJMoa1303359.

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular biology and evolution* 30:197–214. DOI: 10.1093/molbev/mss208.

Rozov R., Brown Kav A., Bogumil D., Shterzer N., Halperin E., Mizrahi I., Shamir R. 2017. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* 33:475–482. DOI: 10.1093/bioinformatics/btw651.

Safety CFF., Nutrition A.Whole Genome Sequencing (WGS) Program - GenomeTrakr Network. *Available at https://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/ucm363134.htm* (accessed June 7, 2018).

Saha B., Singh AK., Ghosh A., Bal M. 2008. Identification and characterization of a vancomycin-resistant Staphylococcus aureus isolated from Kolkata (South Asia). *Journal of medical microbiology* 57:72–79. DOI: 10.1099/jmm.0.47144-0.

Sahl JW., Caporaso JG., Rasko DA., Keim P. 2014. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *PeerJ* 2:e332. DOI: 10.7717/peerj.332.

Sanger F., Air GM., Barrell BG., Brown NL., Coulson AR., Fiddes JC., Hutchison CA III., Slocombe PM., Smith M. 1977. Nucleotide sequence of bacteriophage φX174 DNA. *Nature* 265:687. DOI: 10.1038/265687a0.

Sanger F., Nicklen S., Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74:5463–5467.

Scott E., Duty S., Callahan M. 2008. A pilot study to isolate Staphylococcus aureus and methicillin-resistant S aureus from environmental surfaces in the home. *American journal of infection control* 36:458–460. DOI: 10.1016/j.ajic.2007.10.012.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. DOI: 10.1093/bioinformatics/btu153.

Segata N., Waldron L., Ballarini A., Narasimhan V., Jousson O., Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* 9:811–814. DOI: 10.1038/nmeth.2066.

Shapiro E., Biezuner T., Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics* 14:618–630.

DOI: 10.1038/nrg3542.

Sheikhizadeh S., de Ridder D. 2015. ACE: Accurate Correction of Errors using K-mer tries. *Bioinformatics* . DOI: 10.1093/bioinformatics/btv332.

Shein-Chung Chow JSAHW. 2009. *Sample size calculations in clinical research*. DOI: 10.1002/sim.3468.

Sheppard SK., Jolley KA., Maiden MCJ. 2012. A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter. *Genes* 3:261–277. DOI: 10.3390/genes3020261.

Shopsin B., Mathema B., Alcabes P., Said-Salim B., Lina G., Matsuka A., Martinez J., Kreiswirth BN. 2003. Prevalence of agr Specificity Groups among Staphylococcus aureus Strains Colonizing Children and Their Guardians. *Journal of clinical microbiology* 41:456–459. DOI: 10.1128/JCM.41.1.456-459.2003.

Shore AC., Deasy EC., Slickers P., Brennan G., O'Connell B., Monecke S., Ehricht R., Coleman DC. 2011. Detection of staphylococcal cassette chromosome mec type XI carrying highly divergent mecA, mecI, mecR1, blaZ, and ccr genes in human clinical isolates of clonal complex 130 methicillin-resistant Staphylococcus aureus. *Antimicrobial agents and chemotherapy* 55:3765–3773. DOI: 10.1128/AAC.00187-11.

Sieradzki K., Tomasz A. 2003. Alterations of cell wall structure and metabolism accompany reduced susceptibility to vancomycin in an isogenic series of clinical isolates of Staphylococcus aureus. *Journal of bacteriology* 185:7103–7110.

Siguier P., Perochon J., Lestrade L., Mahillon J., Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research* 34:D32–6. DOI: 10.1093/nar/gkj014.

Sjöstrand J., Tofigh A., Daubin V., Arvestad L., Sennblad B., Lagergren J. 2014. A Bayesian method for analyzing lateral gene transfer. *Systematic biology* 63:409–420. DOI: 10.1093/sysbio/syu007.

Smillie CS., Smith MB., Friedman J., Cordero OX., David LA., Alm EJ. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480:241–244. DOI: 10.1038/nature10571.

Sogin ML., Morrison HG., Huber JA., Welch DM., Huse SM., Neal PR., Arrieta JM., Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare

biosphere." *Proceedings of the National Academy of Sciences* 103:12115–12120. DOI: 10.1073/pnas.0605127103.

Song L., Florea L., Langmead B. 2014. Lighter: fast and memory-efficient sequencing error correction without counting. *Genome biology* 15:509. DOI: 10.1186/s13059-014-0509-9.

Stamatakis A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* . DOI: 10.1093/bioinformatics/btu033.

Stepanović S., Dimitrijević V., Vuković D., Dakić I., Savić B., Svabic-Vlahović M. 2001. Staphylococcus sciuri as a part of skin, nasal and oral flora in healthy dogs. *Veterinary microbiology* 82:177–185.

Stinear TP., Holt KE., Chua K., Stepnell J., Tuck KL., Coombs G., Harrison PF., Seemann T., Howden BP. 2014. Adaptive Change Inferred from Genomic Population Analysis of the ST93 Epidemic Clone of Community-Associated Methicillin Resistant Staphylococcus aureus. *Genome biology and evolution*. DOI: 10.1093/gbe/evu022.

Stokes HW., Gillings MR. 2011. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS microbiology reviews* 35:790–819. DOI: 10.1111/j.1574-6976.2011.00273.x.

Tatum EL., Lederberg J. 1947. Gene Recombination in the Bacterium Escherichia coli. *Journal of bacteriology* 53:673–684.

Ten Hoopen P., Finn RD., Bongo LA., Corre E., Fosso B., Meyer F., Mitchell A., Pelletier E., Pesole G., Santamaria M., Willassen NP., Cochrane G. 2017. The metagenomic data life-cycle: standards and best practices. *GigaScience* 6:1–11. DOI: 10.1093/gigascience/gix047.

Tenover FC., Arbeit RD., Goering RV. 1997. How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists. Molecular Typing Working Group of the Society for Healthcare Epidemiology of America. *Infection control and hospital epidemiology: the official journal of the Society of Hospital Epidemiologists of America* 18:426–439.

Than C., Ruths D., Innan H., Nakhleh L. 2007. Confounding factors in HGT detection:

statistical error, coalescent effects, and multiple solutions. *Journal of computational biology: a journal of computational molecular cell biology* 14:517–535. DOI: 10.1089/cmb.2007.A010.

Thomas T., Gilbert J., Meyer F. 2012. Metagenomics - a guide from sampling to data analysis. *Microbial informatics and experimentation* 2:3. DOI: 10.1186/2042-5783-2-3.

Thomas CM., Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews. Microbiology* 3:711–721. DOI: 10.1038/nrmicro1234.

Thompson LR., Sanders JG., McDonald D., Amir A., Ladau J., Locey KJ., Prill RJ., Tripathi A., Gibbons SM., Ackermann G., Navas-Molina JA., Janssen S., Kopylova E., Vázquez-Baeza Y., González A., Morton JT., Mirarab S., Zech Xu Z., Jiang L., Haroon MF., Kanbar J., Zhu Q., Jin Song S., Kosciolek T., Bokulich NA., Lefler J., Brislawn CJ., Humphrey G., Owens SM., Hampton-Marcell J., Berg-Lyons D., McKenzie V., Fierer N., Fuhrman JA., Clauset A., Stevens RL., Shade A., Pollard KS., Goodwin KD., Jansson JK., Gilbert JA., Knight R., Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551:457–463. DOI: 10.1038/nature24621.

Treangen TJ., Ondov BD., Koren S., Phillippy AM. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome biology* 15:524. DOI: 10.1186/s13059-014-0524-x.

Turlej A., Hryniewicz W., Empel J. 2011. Staphylococcal cassette chromosome mec (Sccmec) classification and typing methods: an overview. *Polish journal of microbiology / Polskie Towarzystwo Mikrobiologow = The Polish Society of Microbiologists* 60:95–103.

Uhlemann A-C., Dordel J., Knox JR., Raven KE., Parkhill J., Holden MTG., Peacock SJ., Lowy FD. 2014. Molecular tracing of the emergence, diversification, and transmission of S. aureus sequence type 8 in a New York community. *Proceedings of the National Academy of Sciences of the United States of America*. DOI: 10.1073/pnas.1401006111.

Vandenesch F., Naimi T., Enright MC., Lina G., Nimmo GR., Heffernan H., Liassine N., Bes M., Greenland T., Reverdy M-E., Etienne J. 2003. Community-acquired

methicillin-resistant Staphylococcus aureus carrying Panton-Valentine leukocidin genes: worldwide emergence. *Emerging infectious diseases* 9:978–984. DOI: 10.3201/eid0908.030089.

Van der Auwera GA., Carneiro MO., Hartl C., Poplin R., Del Angel G., Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J., Banks E., Garimella KV., Altshuler D., Gabriel S., DePristo MA. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 43:11.10.1–33. DOI: 10.1002/0471250953.bi1110s43.

Vesterlund S., Karp M., Salminen S., Ouwehand AC. 2006. Staphylococcus aureus adheres to human intestinal mucus but can be displaced by certain lactic acid bacteria. *Microbiology* 152:1819–1826. DOI: 10.1099/mic.0.28522-0.

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. DOI: 10.1371/journal.pone.0112963.

Watkins RR., David MZ., Salata RA. 2012. Current concepts on the virulence mechanisms of meticillin-resistant Staphylococcus aureus. *Journal of medical microbiology* 61:1179–1193. DOI: 10.1099/jmm.0.043513-0.

Wattam AR., Abraham D., Dalay O., Disz TL., Driscoll T., Gabbard JL., Gillespie JJ., Gough R., Hix D., Kenyon R., Machi D., Mao C., Nordberg EK., Olson R., Overbeek R., Pusch GD., Shukla M., Schulman J., Stevens RL., Sullivan DE., Vonstein V., Warren A., Will R., Wilson MJC., Yoo HS., Zhang C., Zhang Y., Sobral BW. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic acids research* 42:D581–91. DOI: 10.1093/nar/gkt1099.

Wayne LG., Brenner DJ., Colwell RR., Grimont PAD., Kandler O., Krichevsky MI., Moore LH., Moore WEC., Murray RGE., Stackebrandt E., Starr MP., Truper HG. 1987. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International journal of systematic bacteriology* 37:463–464. DOI: 10.1099/00207713-37-4-463.

Weimer BC. 2017. 100K Pathogen Genome Project. *Genome announcements* 5. DOI: 10.1128/genomeA.00594-17.

Whidden C., Zeh N., Beiko RG. 2014. Supertrees Based on the Subtree Prune-and-Regraft Distance. *Systematic biology* 63:566–581. DOI: 10.1093/sysbio/syu023.

Wick RR., Judd LM., Gorrie CL., Holt KE. 2017a. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology* 13:e1005595. DOI: 10.1371/journal.pcbi.1005595.

Wick RR., Schultz MB., Zobel J., Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 31:3350–3352. DOI: 10.1093/bioinformatics/btv383.

Wilson GG., Murray NE. 1991. Restriction and modification systems. *Annual review of genetics* 25:585–627. DOI: 10.1146/annurev.ge.25.120191.003101.

Wood DE., Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 15:R46. DOI: 10.1186/gb-2014-15-3-r46.

Wu R. 1972. Nucleotide Sequence Analysis of DNA. *Nature: New biology* 236:198. DOI: 10.1038/newbio236198a0.

Wu Z., Li F., Liu D., Xue H., Zhao X. 2015. Novel Type XII Staphylococcal Cassette Chromosome mec Harboring a New Cassette Chromosome Recombinase, CcrC2. *Antimicrobial agents and chemotherapy* 59:7597–7601. DOI: 10.1128/AAC.01692-15.

Wu S., Piscitelli C., de Lencastre H., Tomasz A. 1996. Tracking the evolutionary origin of the methicillin resistance gene: cloning and sequencing of a homologue of mecA from a methicillin susceptible strain of Staphylococcus sciuri. *Microbial drug resistance* 2:435–441. DOI: 10.1089/mdr.1996.2.435.

Yoko Furuya E., Lowy FD. 2006. Antimicrobial-resistant bacteria in the community setting. *Nature reviews. Microbiology* 4:36–45. DOI: 10.1038/nrmicro1325.

Zankari E., Hasman H., Cosentino S., Vestergaard M., Rasmussen S., Lund O., Aarestrup FM., Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. *The Journal of antimicrobial chemotherapy* 67:2640–2644. DOI: 10.1093/jac/dks261.

Zhao Y., Wu J., Yang J., Sun S., Xiao J., Yu J. 2012. PGAP: pan-genomes analysis pipeline. *Bioinformatics* 28:416–418. DOI: 10.1093/bioinformatics/btr655.

Zinder ND., Lederberg J. 1952. Genetic exchange in Salmonella. *Journal of bacteriology* 64:679–699.

Zinderman CE., Conner B., Malakooti MA., LaMar JE., Armstrong A., Bohnker BK. 2004. Community-acquired methicillin-resistant Staphylococcus aureus among military recruits. *Emerging infectious diseases* 10:941–944. DOI: 10.3201/eid1005.030604.

Zwick ME., Joseph SJ., Didelot X., Chen PE., Bishop-Lilly KA., Stewart AC., Willner K., Nolan N., Lentz S., Thomason MK., Sozhamannan S., Mateczun AJ., Du L., Read TD. 2012. Genomic characterization of the Bacillus cereus sensu lato species: backdrop to the evolution of Bacillus anthracis. *Genome research* 22:1512–1524. DOI: 10.1101/gr.134437.111.