MULTIDIMENSIONAL STATISTICS METRIC IN BIOLOGICAL DATA ANALYSIS

A Dissertation
Presented to
The Academic Faculty

By

Tzu-Hsueh Huang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Chemistry and Biochemistry

Georgia Institute of Technology

August 2017

MULTIDIMENSIONAL STATISTICS METRIC IN BIOLOGICAL DATA ANALYSIS

Approved by:

Dr. Robert M. Dickson, Advidor
School of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Joseph W. Perry
School of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Christoph J. Fahrni
School of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Jennifer E. Curtis
School of Physics
Georgia Institute of Technology

Dr. Fredrik O. Vannberg
School of Biology
Georgia Institute of Technology

Dr. Yih-Ling Tzeng
School of Medicine
Emory University

Date Approved: July 13, 2017

ACKNOWLEDGEMENTS

I wish to thank all the individuals who have helped me over the course of my Ph.D. Firstly, I would like to express my highest gratitude to my advisor Dr. Robert Dickson for giving me the freedom to explore the field that I am interested in and giving me guidance when I face difficulties. I could not have imagined having a better advisor for my Ph.D. study. I also would like to thank my committee members, Dr. Jennifer Curtis, Dr. Christoph Fahrni, Dr. Joseph Perry, Dr. Yih-Ling Tzeng, and Dr. Fredrik Vannberg, for their insightful comments and inspiring discussions over the years.

My sincere thanks also goes to Dr. Yih-Ling Tzeng for granting access to the laboratory over Emory University, giving stimulating advice and discussions. I would like to thank Dr. Fredrik Vannberg and his members Cai Huang and Dr. Peter Audano for their wealth knowledge of next-generation sequencing. Thanks to Dr. Eileen Burd and Prof. Colleen Kraft for their clinical experience on antibiotic resistance. Additionally, I would like to thank Dr. Phillip Rather for providing the clinical isolates that are used in this work. Also thanks to Dr. Niren Murthy and his member Xinghai Ning for synthesizing the maltohexaose-conjugated fluorescent dye. I also thanks Dr. Niren Murthy for investing his time to serve as my committee member for my candidacy exam and Dr. Wendy Kelly for sharing her laboratory space and resources.

I thank both current and former Dickson lab members, Dr. Saugata Sarkar, Dr. Chaoyang Fan, Dr. Amy Jablonski, Dr. Soonkyo Jung, Dr. Blake Fleischer, Dr. Jeff Petty, Dr. Daniel Mahoney, Jung-Cheng Hsiang, Aida Demissie, Joseph Richardson, Yen-Cheng Chen, Alexandra Mandl, Yi-Han Li, Baijei Peng, and Keith Creech, for their constant support and all the fun conversations. In particular, thanks to Joseph Richardson and Alexandra Mandl for their works on antibiotic resistant test and data analysis.

Finally, I would like to thank my family for their patience and encouragement. Special thanks to my husband, Denny Lie, for his continuous love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

From flow cytometry to next-generation sequencing (NGS), data analysis is the key to obtaining useful information from the massive, and complicated data. Traditional statistical methods might not be suitable for analyzing the new data due to the size and/or dimensionality. Moreover, novel analysis methods are needed to analyze these data with unprecedented data structure and its unique problems. This work focuses on developing new statistical methods to analyzing multidimensional and big data.

Probability binning - signature quadratic form (PB-sQF) is developed to analyze cytometric data. These type of data can have one million of data points spanned up to 30 dimensions. PB-sQF first compresses the data by calculating the signatures of the original data. Then, the differences between data can be obtained by calculating the distances between these signatures. Using PB-sQF, the test statistics (distances) between cytometric data of antibiotic-treated bacteria and the no-antibiotic control were calculated. Since only effective treatments induced cytometric signal changes, these distances can be used to select the correct treatment for multidrug-resistant bacteria. Because PB-sQF can objectively detect the minor changes in the cytometric signal, only one hour (instead of overnight) incubation is needed for the post-blood culture antibiotic susceptibility test (AST). With pre-blood culture, we have developed an experimental procedure to remove the high background from the blood cells and reduce the incubation time from 66 hours to 8 hours. Our method greatly reduces the test-to-result time can thus lower the mortality rate of patients and the mutation rates for bacteria.

PB-sQF can also be used to compare genome sequences similarities. By breaking the string sequence into pieces, the genome sequence can be viewed as a histogram data with repeated words/data points. PB-sQF can thus calculate the distances between genome sequences. These distances are then used to type unknown bacterial sequences, build phylogenetic trees, and perform outbreak analysis. Since NGS generates millions of short reads

instead of a complete sequence, this work also builds the short reads mapping (aligns the short reads to a reference sequence) scheme for distance-based reads mappers like PB-sQF and nearest-neighbor (NN). Like PB-sQF, NN computes the distances between histogram data. While most short reads aligners are built for short read length and exact unique maps, NN can perform error-tolerant, long read, multiple mapping locations alignments.

Copy number variation (CNV) is the differences in the number of a gene in different genome sequences. It is of great interests because of its associations with various diseases. While numbers of CNV detectors have been developed, references sequences with repeated regions are still a problem for current CNV detectors. Using the NN mapping results, we develop a new CNV detecting algorithm, copy number variation detections for mappings multiplicity (CNV-MM), specialized in estimating the copy numbers in both the unknown and reference sequences even when the reference sequences are highly repetitive. We show that using NN and CNV-MM, different sizes of CNV can be detected with various read length. We also show that the estimated copy numbers are accurate for both duplications and deletions from one copy to twenty copies. Eventually, we perform our method on real short reads data from multidrug-resistant *Acinetobacter baumannii* clinical isolates and show that copies of resistant-associated genes indeed increase relative to the sensitive strain.

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Motivation

With the advance in modern technology, scientists have to process ever-lengthening data with higher dimensions and greater variability. Often obscured within these data is information that has the potential to improve human health.[1–3] Due to data complexity, however, existing analyses are not suitable to extract underlying information.[3] For example, flow cytometry collects hundreds of thousands of data points in just a few minutes while simultaneously monitoring 17 fluorescence channels plus 2 scatter signals.[4, 5] With 30-parameter flow cytometry being available and 50-parameter on the way, new computational methods are required to scaled well with high-dimensionality.[6] On the other hand, a human genome which consists of 3 billion base pairs can now be sequenced within days with next-generation sequencing (NGS) by generating hundreds of millions of reads in a single run. With the size of the genome and/or the huge number of short reads, computationally efficient methods are needed to extract the relevant information from the sequences accurately.

Flow cytometry holds promise to accelerate antibiotic susceptibility determinations by rapidly measuring multichannel fluorescence and scatter from each cell within a large population. The accelerating emergence of multidrug-resistant bacteria and difficulty in quickly identifying appropriate treatment options are major threats to global public health.[7–10] The ability of bacteria to rapidly counter newly available antibiotics within only a few years of clinical introduction, has also produced super-bug infections that are essentially untreatable. Such rapid acquisition of resistance has also decreased both incentives and options for new antibiotic development, with only two new antibiotics having been approved since

2008.[9, 11] With 30% of hospital deaths attributable to sepsis, bacterial infections of the blood have become the 10th leading cause of hospital deaths in the US.[12, 13] Although rapidly tailored treatment to each individual patient can have a major impact on positive outcome,[14] currently, nearly 30% of patients receive inappropriate antimicrobial therapy. Such non-ideal treatment leads to 2-fold higher mortality rates than when correctly treated,[15] and also contributes to the increase in multidrug-resistance resulting from sublethal antibiotic exposure.[16, 17] While rapid initiation of appropriate treatment is crucial to positive patient outcome, only combined knowledge of the pathogen identity and its antibiotic sensitivity profile comprise actionable treatment information. Because bacterial load in sepsis patients is so low, $\sim$24-hr blood culture-based amplification is crucial to diagnosis treatment. Recent advances that employ mass spectrometry and genetic tests[18–22] enable identification of infectious agent within a few hours after positive blood culture. However, conventional antibiotic sensitivity tests (ASTs) typically require overnight subculturing, followed by an 18 to 24 hr AST, resulting in a 42$\sim$48 hr post blood culture delay in susceptibility data. Thus, improving AST time-to-result would have positive patient and public health outcomes.

NGS and genome sequence analysis have greatly improved our understanding in many different fields.[23, 24] Sequence analysis methods that compare sequences similarity of pathogens can track down the disease transmission and thus contribute to detection and control of outbreaks.[25, 26] Methods that perform sequence matching analysis can be applied to species typing or antibiotic resistant identification in clinical microbiology.[23, 27] Short reads mappers can detect variations between genome sequences through aligning the short reads to the reference sequence.[24, 28, 29] Different sequence analyses algorithms have been proposed for constructing phylogenetic trees,[30–33] typing unknown species,[34–36], mapping short reads, [28, 29, 37–40] and detecting sequence variations.[41–46] However, to our knowledge, there is no single method that can tackle all these different tasks. Also, although various short reads mappers have been proposed,[28, 29, 37–40] these meth-

ods still have difficulty in short reads mapping multiplicity,[28, 47] which is assigning reads to multiple possible mapping positions on repeated sequences. These problems obscure downstream copy number variations (CNVs) analysis relative to a reference genome, which is crucial to diagnosing many important medical conditions.[42, 43] Thus, the performance of CNVs analysis would be improved by properly detecting the repeated regions in genome sequences.

This work develops a new multidimensional statistics test, Probability Binning signature Quadratic Form (PB-sQF).[48] PB-sQF characterizes the multidimensional differences between data sets into a one-dimensional linear distance while accounting for the noise of the data. Since PB-sQF compresses the multidimensional data into a set of data signatures, it can efficiently compare the similarity between 2 multidimensional data sets to quantify the differences between flow cytometry data. With modifications, PB-sQF is applied in genome sequence analysis in building a phylogenetic tree, typing unknown species, and mapping short reads. Ultimately, PB-sQF can reveal new insights in different fields by quantifying biological relevant (dis)similarity between data sets.

1.2 Flow Cytometry

High-throughput flow cytometry has been an important technology for biological studies since it was created in 1965 by Mack Fulwyler.[49] Different from traditional lab techniques that record the bulk signals or average responses of the sample; flow cytometry obtains single cell information of hundreds to millions of individual cells while monitoring up to 30 parameters with the advance of fluorescence labeling.[4, 6, 50] Moreover, the data collection process is non-invasive and the cells sorted by the cytometry can be used for further analysis.[51, 52] Flow cytometry is thus widely used in immunology,[4, 6] cell-cycle analysis,[53–55] and cancer diagnosis.[56–58] With the ability to detect cell particles down to 0.2 μm, flow cytometry is also used in bacteria detection.[59, 60]

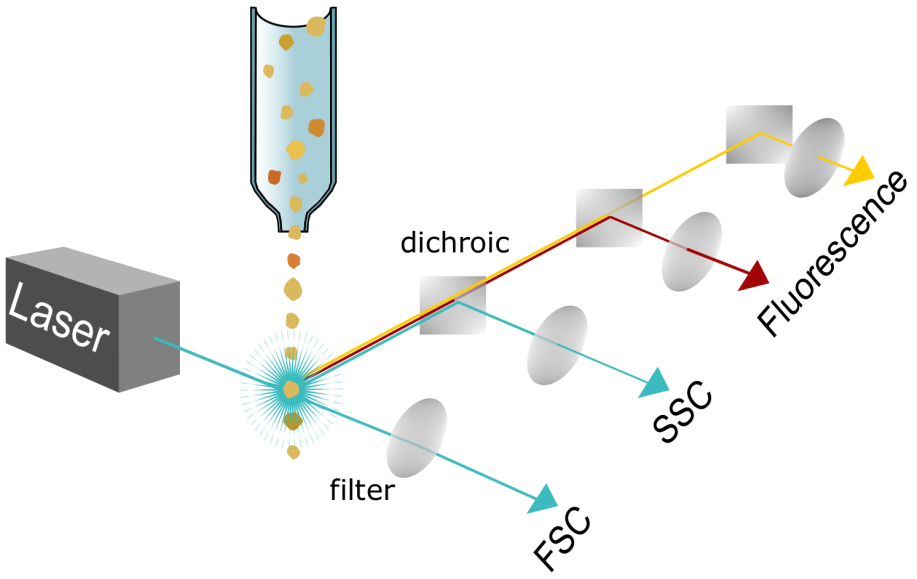

Figure 1.1: The schematic diagram of flow cytometry.

1.2.1 Flow Cytometry Principles

Using hydrodynamics focusing,[61] cells in the sample are carried by a sheath fluid and intercepted by the laser beam one cell at a time. The interaction between cells and incident laser light generate scattered light and fluorescence signals (if the cells are labeled with fluorescence probes). Photomultiplier tubes (PMTs) are used to record the signals. Scattered light propagates in all directions when the laser light hits a cell. Commercial flow cytometry collects the scattered light in both the forward and 90-degree angle (side scatter) directions. Forward scatter signal is proportional to the size of a cell while side scatter signal indicates the complexity or granularity of a cell. Fluorescence signals, depend on the labeling scheme, is relative to membrane permeability, protein abundance, and cellular identity. The multidimensional data is most often presented in histograms, scatter or pseudocolor plots. The schematic diagram of flow cytometry[62] is shown in Figure 1.1.

1.2.2 Flow Cytometry Bacterial Viability Test

The conventional antibiotic sensitivity tests (ASTs), which relies on monitoring bacterial growth, typically requires overnight subculturing resulting in a 48 to 72 hour delay in determining effective antibiotic treatment. Yet, antibiotic-induced bacterial morphology and physiology changes can be monitored by flow cytometry long before the detection of growth inhibition.[63] Previously described cytometry-based ASTs mostly rely on interpreting qualitative differences in live and dead cell populations using high-background fluorescent membrane integrity/potential sensors.[59, 64–73] For example, the LIVE/DEAD BacLight Bacterial Viability Kit from ThermoFisher contains 2 DNA chelating dyes: SYTO 9 and propidium iodide (PI). Fluorescence from both dyes is significantly enhanced upon binding to DNA. SYTO 9 is permeable to the cell membrane and thus stains both live and dead cells. On the other hand, PI only gains access to the cells when the membrane is permeabilized, an indicator of cell death. As a result, when running the sample through flow cytometry, the live cells are bright green (STYO 9 signal) while the dead cells are much dimmer in the SYTO 9 channel but exhibit bright red (PI signal). [59] Other than membrane integrity, membrane potential is also commonly used to assess cell viability. DiOC2(3) (3,3-diethyloxacarbocyanine iodide) is a membrane potential dye. It appears green in all cells. However, in healthy cells that maintain the membrane potential, self-association occurs due to high concentration of DiOC2(3). The fluorescence turns to red. By monitoring the red/green fluorescence ratios, one can assess cell viability. [74, 75] Such cytometric ASTs, however, suffer from high background, very large statistical variability, and insufficient changes from controls that have rendered comparisons unquantifiable. [68, 70, 76] Thus, even with a large number of observations probing antibiotic-induced changes within a given cell population, flow cytometric ASTs fail from the lack of accurate statistical metrics to quantify multidimensional changes relative to controls. This prompts the development of computational flow cytometry to compare the (dis)similarity between cytometric data.

1.2.3 Computational Flow Cytometry

Ideally, the dissimilarity between 2 distributions (sample vs. control) is quantified by various test statistics, yielding distances between measured distributions. One-dimensional test statistics are either too sensitive to provide meaningful analysis [77, 78] or require large numbers of events [78] and are usually not rigorously extensible to multidimensional data. Adaptive binning overcomes the dataset size and multidimensional extensibility issues to focus analysis on the most informative regions of the data; [79, 80] however, sample comparison is control-specific in probability binned chi-square (PB-χ^2) tests, making the test result unsuitable as a true (linear) metric for directly comparing multiple samples. As a consequence, it is difficult to adjust statistical tests for biological variability. Various multidimensional distance metrics are known, but necessary computational resources tend to scale with the number of bins raised to some large power. Scaling quadratically with number of bins, quadratic form (QF) distance statistics directly addresses the metric issue, providing a linear distance between any two multidimensional data sets, [81] but its reliance on fixed bin sizes can limit its extension to multidimensional cytometry datasets. Instead of comparing occurrences in fixed bins, "signature" QF (sQF) has been developed for image analysis to directly compare signatures, or the most important features, within images.[82, 83] By combining the adaptive probability binning with the signature QF distances, we have developed PB-sQF that can calculate linear distances between adaptively binned, multidimensional data sets.[48]

1.3 Next-Generation Sequencing (NGS)

Since it was published in 1977, Sanger sequencing[84] had been the standard DNA sequencing technique for nearly 40 years. Its reign has ended due to NGS in the mid-2000s. NGS has revolutionized a wide range of fields by greatly lowering the cost, reducing the sequencing time, and generating enormous data sets. Different from the Sanger sequenc-

ing which performs nucleotide addition, fluorescence signal detection, and reagent removal processes separately, NGS carries out all processes in repeated cycles which greatly reduces the sequencing time.

Although different sequencers applied different strategies for NGS, the general procedures are similar. The target sequence is first broken into small fragments, amplified by PCR into clusters, and then sequenced through cycling processes.[85, 86] In this massive parallel sequencing scheme, millions to billions of reactions on clusters/beads are running at the same time. Thereby enabling high-throughput sequencing.[85, 87, 88] Although the read length (the length of the output DNA fragments) is in general shorter and the error rates are higher than Sanger sequencing, NGS overcomes these limitations with high coverage and advanced computational algorithms and has become very useful in many different applications.

1.3.1 Sequence Similarity and Typing

Generating large volumes of data in short amounts of time at low cost has rendered NGS very useful in genomic research. Through whole-genome sequencing (WGS), scientists get nearly complete information on each genome and can compare one genome versus another. In clinical microbiology, WGS can contribute in pathogen species typing, resistance and virulence detection, and epidemiological analysis.[23, 24] For bacterial species typing, multi-locus sequence typing (MLST) is commonly used in genome sequence typing.[89] By comparing a set of housekeeping genes in the genome sequence to the MLST library, sequence type is assigned to the query sequence. With WGS available, instead of relying on data from PCR, one can map the individual reads back to the reference genes and perform MLST.[35, 36, 90] However, using assembled genomes requires de novo assembly to first be performed on the acquired short reads, limiting its application. Additionally, the diversity of bacterial genomes often makes it difficult to build a universal set of housekeeping genes to distinguish among strains, and there is often insufficient resolution among closely

related bacteria for gene-by-gene approaches.[91, 92] Further, with low sequencing depth, it is possible that one or more of the loci are missed, leading to false assignment.

Instead of relying on identical matches to a set of marker genes, the occurrence of short k-mers, nucleotides with length k within the reads, has proven to be very successful in short reads assembly[93–95] and mutation identification.[96] For genome sequence typing, KmerFinder compares the existence of unique k-mers within the unknown and library sequences without reliance on k-mer location.[97, 98] The unknown is assigned to the same species of the library sequence with which it shares the most k-mers. However, for unique assignments relying on matching binary (present or not) k-mer existence, much longer k-mers (k$>$16) must be used to avoid k-mer saturation in species with even moderate genome sizes. Utilizing the counts/frequency of each k-mer, the Ribosomal Database Project classifier utilizes Bayesian inference to predict the probability of an unknown sequence belonging to a certain genus by estimating the probabilities of k-mer existence in both the training and query sequences.[27, 99, 100] Concerns with Bayes-based approaches include its assumption of k-mer independence and the need for a training data set to build the prior probabilities. Instead of applying Bayesian inference, the nearest-neighbor approach (NN) treats each sequence as a k-dimensional data point with the k-mer occurrence frequency (i.e. counts) as the coordinate. Calculating Euclidean distances among all query and library k-mer abundances enables the query to be assigned to the genus of the library exhibiting the smallest paired distance and thus typing can be done without a training data set.[101] Unfortunately, the computational burden of these multidimensional distance approaches scales exponentially with k-mer length, as the number of dimensions increases with 4^k. Thus, using NN is computationally prohibitive with the long k-mers necessary to compare large genomes. This burden can be relieved by mapping the k-dimensional space on to a lower dimensional space by recognizing that most k-mers are either absent or correlated with each other, but a training data set is again required to calculate the correlation among k-mers.[102]

Taking a different approach, the extremely high information content encoded in even modestly sized genomes suggests that k-merized genetic information can be considered as probability distributions to be analyzed with multidimensional statistical metrics. PB-sQF treats the k-mer occurrences of entire k-merized genomes as 4k-dimensional histograms. This approach enables direct genome sequence analysis and comparison by reducing high dimensionality data to a single linear statistical distance. By calculating inter-genome distances, PB-sQF enables rapidly typing bacteria to the species level from the completed sequence or even from raw, unedited short reads data files.

1.3.2 Short Reads Mapping

After NGS generates millions to billions of reads from a target sequence, instead of piecing all the reads together through de novo assembly, one can align the reads to a reference sequence. This process is called short reads mapping or reads alignment and can lead to the discovery of differences between the target sequence and reference sequence. One of the very common mutations, single-nucleotide polymorphisms (SNPs), accumulates in the genome sequence and is used in outbreak analysis. Harris et al. analyzed the SNPs profiles between isolates and were able to track down the intercontinental spread and hospital transmissions of methicillin-resistant Staphylococcus aureus (MRSA) ST239.[25] Again, by comparing the SNP profiles, scientists can determine the transmission of all waves in the seventh cholera outbreaks since 1961, including the recent outbreak in Haiti in 2011.[103] Using MinION from the Oxford Nanopore, a real-time genomic surveillance was conducted during the Ebola outbreak in 2015.[26]

Numerous different short reads mapping methods have already proven quite successful. Based on indexing techniques, current alignment algorithms can be divided into two categories: hash table-based and BWT-based.[104, 105] For algorithms based on hash tables, including PatternHunter,[106] MAQ,[29] ZOOM,[107], mrFAST,[28] and MOSAIK,[37] the seed-and-extend scheme derived from BLAST was applied.[38] This type of algorithm

cuts the reads (or the genome) with k-mer subsequences, finds the exact matches on the genome (or the reads) as seeds and extends the seeds with dynamic programming algorithm by Smith and Waterman.[108] Bowtie,[39, 109] BWA,[40, 110, 111] SOAP2,[112] and BWT-SW[113] are methods based on BWT. In this category, the genome sequence is indexed by BWT to efficiently store the index. Backward search is then applied to find the exact matches as seeds. These seeds are extended again with dynamic programming. BWT-based methods are efficient because of the tries data structure, which concatenates exact repeats on the same path.

By calculating the distances between short reads and library reads, which are constructing by cutting the reference sequence into contigs with sizes equal to the short reads length, both PB-sQF and NN can map the short reads back to the reference sequence. Different from base-to-base comparisons which are very sensitive to any mismatch in a read, which can arise from sequencer errors and/or mutations/variations between genome sequences, k-mer based read-to-read comparisons are more relaxed in mismatches and thus have better error tolerance. Also, since pair-wise distances are calculated instead of reporting match or mismatch as in both hash table or BWT based methods, distance-based methods can easily report mapping multiplicity, which is crucial for CNVs detection as explained later.[28, 43, 47, 114]

1.3.3 Copy Number Variation Detection (CNV)

One of the major applications for short read mapping is CNVs detection. CNVs, which are the variations of the number of copies of specific genes from one individual to another, contribute to genome heterozygosity and have been found to be common in human genomes. [115, 116] CNVs have been linked to different diseases such as AIDS,[117] obesity,[118] cancer,[119] autism,[120] and Parkinson's disease.[121] CNVs have also been observed in bacteria where these variations are used to gain fitness for environmental adaption.[122] Greenblum et al. show that the intra-species CNVs have environment-related functions and

are associated with diseases.[123] Bacterial gene amplifications have also been found to be associated with the increasing antibiotic resistance.[124]

There are four strategies for detecting CNVs using NGS data. These include (i) Read-depth, (ii) paired-end mapping, (iii) split read, and (iv) de novo assembly.[41, 43, 125] Although these approaches have different strengths and limitations, read-depth based methods have become more and more popular because of increasing availability of high coverage data.[41] In read-depth based analysis, short reads are mapped to a reference genome and the number of reads per windows (read depths) are calculated. Assuming reads are randomly sampled from a target sequence, the read depth of each window would be constant throughout the sequence if the copy number of the target sequence is the same as in the reference genome. Any increase or decrease in read depth indicates duplications or deletions of the regions. Methods based on read-depth analysis can be further categorized into 3 groups depending on how many samples were used. Using only the sequenced target sample, CNVnator[42] applies the mean-shift approach[126] to find the CNV regions and calculates the absolute copy number by normalizing it with the average read depth. Methods using only a single sample to estimate the copy number suffer when there are repeated regions in the reference genome. In this case, most of the short reads algorithms perform poorly with mapping as they can't properly reconstruct the true read depth. ReadDepth[127], which also only analyzes the target sample, uses statistics to correct the read depth of regions with multiple alignments. For methods like CNV-seq,[128] RDXplorer,[129] and SegSeq,[45] which normalizes the read depth of the targeted genome with the read depth of the control/paired genome, canaccount for the repetitions in reference sequence. However, a second sequenced genome is required. Also, only the ratio of copy numbers between the target sequence and the reference sequence can be obtained. Methods analyzing multiple sequenced genome have also been proposed to better estimate the breakpoints (CNV boundaries) and true copy numbers.[130, 131] To cope with the repetitions in a reference sequence during the short reads mapping process, mrFAST was developed

to report multiple mapping locations.[28] However, mrFAST is tailored to perform short reads alignment with high accuracy (Illumina), short read length (36-mer) data and thus performs poorly on longer reads with higher error rates.[132, 133]

To apply the NGS data in CNV analysis, both the number of copies and the breakpoints of the variable region need to be accurately estimated. However, one of the disadvantages of read-depth based CNV detection methods is that they are relatively poor at the precisely determining the location of a CNV. This is because, to overcome the fluctuations in read depth (especially with low coverage data), most methods binned the read depth trajectory before analysis. In Alkan et al.[28], and RDXplorer[129], fixed/arbitrary window sizes are used. While in SegSeq,[45] CNVnator,[42] CNV-seq,[128] and ReadDepth,[127] the bin sizes are optimized by the statistical models to ensure the best resolution at a certain false discovery rate. No matter how the bin size is chosen, the resolution of breakpoints is determined by the bin size.

It has been shown that repetitions are very common in different genomes across all three domains in life.[134–138] As a result, it is crucial to develop a new method that is robust in mapping and detecting highly repeated regions. Using PB-sQF or NN, the short reads originating from one of the repetitive regions in the reference sequence would have short distances with the library reads from the repetitive regions. To fully exploit the mapping multiplicity, a CNV detection algorithm, copy number variation detection for mapping multiplicity (CNV-MM), is developed to calculate the absolute copy number and the breakpoint of repetitive regions. Since it is guarantee to find all the repetitive regions with distance-based short reads mapper, the number of repetitions of a region in the reference genome is obtained at the same time during the mapping process. As a result, although only the targeted sequence was sequenced and the absolute copy number is obtained, the ratio of the copy numbers between the targeted and reference sequences can also be calculated. Thus, the changes of copy number in the target genome can also be reported. The breakpoints can also be refined with the information from the reference sequence.

1.4 Organization of Thesis

Data analysis often boils down to the important question: How similar/different are any two data sets? Precisely for comparison of large, noisy datasets, this work developed a new multidimensional statistical test, PB-sQF. This work analyzes the flow cytometric data of antibiotic-induced changes in bacterial cells with PB-sQF and rapidly determines the effective treatment for multi-drug resistant pathogens. In genome sequence analysis, PB-sQF and a similar distance-based method NN were used to calculate distances between sequences. Using PB-sQF, the unknown bacterial species were typed from either the completed genome sequences or raw short reads data. A phylogenetic tree can be built from the pairwise PB-sQF distances for tracking outbreak infections. NN, on the other hand, can perform error-tolerant short reads mapping. Combining with CNV-MM, a CNV detection scheme that is robust with repeating regions in the reference sequence can be built.

The rest of this dissertation is organized as follows:

Chapter 2 develops the multidimensional statistical test, PB-sQF. This chapter explains how PB-sQF transforms original data into a set of signatures and how PB-sQF calculates a one-dimensional distance between multidimensional data sets. This chapter also explains the bootstrap method for confidence level construction, error propagation, and geometric quantiles calculation through quasi-newton minimization for characterizing flow cytometry data. Specific for genome sequence analysis, this chapter describes how to adapt PB-sQF for string comparison and how to construct the library reads for short reads mapping. This chapter also introduces NN and the how the CNV information was extracted from the read depth trajectory.

Chapter 3 proposes a new post-blood culture AST through characterizing the different between antibiotic-treated and no-antibiotic cytometric data by PB-sQF. This chapter examines a wide range of antibiotic-bacteria combinations and monitors antibiotic-induced changes with flow cytometry. First, the antibiotic-induced ROS was characterized with

ROS sensing dye. Then antibiotic-induced scatter patterns changes were monitored. PB-sQF was applied on the cytometric data to determine the effectiveness of an antibiotic toward the bacteria which reduce the time-to-results from 72 to 28 hrs.

Chapter 4 demonstrates a pre-blood culture AST (FAST) can be built with only 5-hour bacterial incubation time. This chapter searches for conditions to separate the bacteria cells from blood cells that generate high background signal in cytometric data. This chapter develops the work flow for pre-blood culture AST test which reduces the time-to-result from 72 to 8 hours.

Chapter 5 investigates the ability for PB-sQF to perform genome sequence analysis. This chapter demonstrates that PB-sQF can type bacteria at the species level using complete genome sequences. With pooled raw short reads raw data, PB-sQF can determine the genus of short reads data. With the distances between each sequence being calculated, PB-sQF can construct a phylogenetic tree to study the evolutionary relationships of known bacteria strains or perform outbreak analysis.

Chapter 6 studies short reads mapping with PB-sQF and NN. This chapter investigates the reliability of Euclidean distance-based mapping of the short reads back to a reference genome with different conditions being verified: read lengths, system errors, single-nucleotide polymorphism, insertions, and deletions. Both PB-sQF and NN have high error tolerance compared to Bowtie2,[109] SOAP2[139] and mrFAST.[28] This chapter shows that "unknown" sequences can be typed from the bacterial sequence library by mapping only a few as 10 short reads.

Chapter 7 applies NN on short reads mapping to address the challenge of mapping multiplicity. The alignment results were used in CNV detections. A CNV detection algorithm specialized in handling repeating regions in the reference sequence, CNV-MM, was developed. Using NN, read depth trajectory can still be recovered even with reads having higher error rate. CNV-MM then analyzed the alignment results with all valid mapping locations by comparing the read depth trajectories between the sequenced target genome and rebuilt

reference sequence (non-sequenced). With NN and CNV-MM, repeated regions in the reference sequence can be linked and one can distinguish the true CNVs from the repetitions in reference sequence.

CHAPTER 2

EXPERIMENTAL AND COMPUTATIONAL SECTION

2.1 Cytometric-based AST

This section describes the procedure of dyes preparation, bacteria culture, antibiotics incubation, and cytometric data acquisition

2.1.1 MH-IR786 and MH-hIR786 preparation

MH-IR786 was provided by Dr. N. Murthy in University of California, Berkeley and was used in the post-blood culture AST. The reduced form, MH-hIR786, was used in ROS detection.

MH-IR786 was prepared at a concentration of 1 mg/100 μL deionized water. The absorption of the MH-IR785 solution was then measured and the true MH-IR786 concentration was determined by the Beer's law with the extinction coefficient equals to 287,767 $M^{-1}cm^{-1}$ as reported by Nakayama et al.[140]

To prepare MH-hIR786, MH-IR786 solution was reduced. Sodium borohydride (Sigma) was dissolved in methanol (VWR, Batavia, IL) at 1 mg/mL and subsequently added, 10 μL at a time, until the MH-IR786 changed from dark green to yellow. The MH-hIR786 solution was then vacuum dried and resuspended in pH 6.0 acetate buffer (Fisher Scientific) at a final concentration of 1 mM. To ensure that the MH-hIR786 fluorescence can indeed be recovered by reacting with ROS, in vitro fluorescence recovery was tested by oxidizing hIR786 to IR786 with ROS generated from Fenton's reagent.[141, 142] First, the fluorescence baseline of 20 μL of MH-hIR786 in 2 mL of dimethyl sulfoxide (DMSO) (Fisher Scientific) was measured with a fluorimeter (QuataMaster, Photon Technology International). Then, Fenton's reaction was initiated by adding 60 μL of $FeSO_4$ (Mallinckrodt, St.

Louis, MO) at 3.5 mg/mL and 300 μL of H_2O_2 (VWR, Batavia, IL) at 200 nM to the MH-hIR786/DMSO solution. The fluorescence signal was measured immediately and 1-hour after the reaction.

2.1.2 AST by ROS detection

To show that antibiotic-induced ROS generation is correlated with cell death, bacteria were cultured overnight in an incubator shaker (MaxQ 4000, Thermal Fisher Scientific, Waltham, MA) in Luria-Bertani (LB) medium (Sigma-Aldrich, St. Louis, MO) at 37 °C and 225 rpm. Bacteria were then re-inoculated in 12 mL fresh LB medium in 50-mL tubes and incubated from $\sim$0.05 optical density (OD) to the mid-log phase. Bacteria in 1 mL of growth media were collected by centrifugation (Centrifuge 5417R, Eppendorf) at 13,400 rpm for 3 min and transferred into 12-well plates (Costar, New York, NY). Antibiotics and 20 μL of MH-hIR786 (provided by Dr. N. Murthy's lab) to achieve a final concentration 900 nM were added simultaneously. The MICs of different antibiotics were determined by standard microbroth dilution assays in advance. The 12-well plates were incubated at 37 °C for 1 hour (Isotemp standard incubator, Fisher Scientific, Waltham, MA). Bacteria were again collected by centrifugation and washed 3 times with phosphate-buffered saline (PBS) (Life Technologies, Carlsbad, CA) and resuspended in 1 mL PBS. The bacteria samples were maintained on ice until flow cytometry was performed. Bacteria samples were analyzed by a BD LSR II flow cytometer (Becton Dickinson, Franklin Lake, NY) equipped with a 14 mW, 488 nm solid-state coherent sapphire laser for the scatter signal, and a HeNe Laser (18 mW @ 633 nm) for IR786 fluorescence detection. Samples were gated by forward and side scatter, while a 750-810 nm bandpass filter was used to collect the IR786 fluorescence. Data were collected with FACSDiVa provided by BD. Further data analysis and display were carried out with Matlab 2013b (Math Works). For each data set, 100,000 bacterial detection events were collected.

2.1.3 Post-Blood Culture AST Procedure

For post-blood culture AST test, a similar sample preparation procedure was used as in the ROS sensing test. LB broth was used for incubating *E. coli* while cation-adjusted Mueller-Hinton Broth (CAMHB) was used for all other types of bacteria. For each well of the 12-well plate, 1 mL of fresh culture (OD $\sim$0.5) were spin down and resuspend to 480 μL or 500 μL of broth for samples incubating with or without 20 μL, 900nM of MH-IR786. For each well, antibiotic was prepared at in 500 μL, 2x higher of the designated concentration. The samples were incubated with antibiotics at their respective 1x, 1/4x and 1/16x MIC that was first determined by standard microbroth dilution assays. After 1-hr incubation, bacteria were pelleted, washed 3 times and resuspended in PBS for cytometric analyses. Fluorescence and scatter signals upon antibiotic challenge were monitored by flow cytometry. IR786 fluorescence, forward-scattered and side-scattered light were all collected for each of 100,000 measured bacterial cells per run, yielding 3-D histograms for each antibiotic concentration.

2.1.4 Pre-Blood Culture AST Procedure

To simulate blood from a patient with bacteremia, the clinical isolates were grown, diluted to the appropriate CFU/mL and added to the blood/saponin mixture to achieve the final diluted sample. Initial cultures for bacteria-laden blood samples were prepared using LB broth for incubating *E. coli*. For other bacteria, CAMHB was used. Bacteria were cultured overnight in an incubator shaker at 37°C and 225 rpm. The fresh bacterial culture was started from $\sim$0.05 OD by inoculating a 6 mL fresh growth medium with overnight culture. After the fresh culture reached mid-log phase, bacteria were diluted into $\sim$10 CFU/mL through a series of 10-fold dilutions and the concentration was confirmed by overnight plating from loading 100 μL of 1000 CFU/mL sample. The last 10-fold dilution was done by adding the 500 μL of 100 CFU/mL into 4500 μL of 10% human blood (ZenBio, Research Triangle Park, NC) in medium solution.

2.5% (w/v) of saponin (Sigma-Aldrich, St. Louis, MO) was prepared, sonicated (Branson 2510, Emerson, St. Louis, MO) for 20 minutes and spun down with a clinical centrifuge (Centrific Model 228, Fisher Scientific, Waltham, MA) for 4 minutes. The supernatant was collected to isolate the undissolved pellet. 500 μL of 2.5% saponin was then added to the 5 mL of $\leqq$10 CFU/mL, 10% human blood sample and incubated in an incubator shaker for 15 minutes at 37°CṪo ensure that the blood cells lysed completely, the sample was laid on the incubator floor, confined by the large flask clamps, and agitated at 300 rpm. The sample was flipped by hand every 5 minutes. After the saponin treatment, the bacteria were again pelleted and washed with 2 mL of PBS using a clinical centrifuge for 2 minutes. 2.5 mL of growth medium was then added to the tube without breaking the pellet and incubated for 2 hours in an incubator shaker at 37°C and 225 rpm.

After the 2-hour incubation, the pellet was removed by pipetting. The sample was mixed well and 500 μL of the sample was added to each well of one row of the 12-well plate (4 samples per row) that was loaded with 500 μL of growth medium with or without antibiotic at 2-fold of the specified concentrations. The plate was then incubated at 37°C for 3 hours. Bacteria were again collected by centrifugation and resuspended in 200 μL of PBS for flow cytometry detection. To ensure each clinically isolated strain was tested at its MIC, pure culture starting with 1000 CFU/mL was also tested for each experiment, confirming that the antibiotic concentrations we used indeed inhibited bacterial growth.

Bacterial growth inhibition was monitored by flow cytometry. Forward scatter and side scatter signals were recorded.

2.2 Probability Binning - signature Quadratic Form (PB-sQF) Overview

To create a linear statistical metric that readily scales to multiple dimensions, we have combined the adaptive probability binning with the signature QF distances. As described in Chapter 1, probability binning from PB-χ^2 allows one to represent the original data with many fewer signatures, while sQF calculates the linear distance between the signa-

tures.[79–83] Combining the best attributes of PB-χ^2 and sQF 2D image analysis, PB-sQF focuses binning on the high-density regions of the data, better facilitating similarity comparisons among multidimensional datasets while calculating the true distance between data sets.

PB-sQF is a multi-dimensional distance statistic that quantifies the (dis)similarity of any two distributions.[48] Probability binning codes were written to be equivalent to those described in published studies.[79, 80] Schematized in Figure 2.1 A, probability binning bins the data into approximately equal counts/bin by varying bin width. Because the data are adaptively binned, this procedure concentrates bins where the data are concentrated.[48] Thus, as shown in Figure 2.1 B and 2.1 C, the data are effectively represented by the set of centroids, or signatures,[82] which are the median data point (cytometric data) or the mean of the data (genome sequence) within each bin. Both the control and sample were binned in the same manner, and the centroids and weights of each bin were calculated. The PB-sQF matrix multiplication approach then yields calculated test statistics that reveal the overall Euclidean distance between the two sets of centroids through quadratic form calculation.[82, 83] The smaller the test statistic, the more similar the two distributions are.

2.3 Binning Procedure and Test Statistics Calculation

Binning is performed recursively by calculating the variance in each dimension, identifying the highest variance dimension and dividing data at the median (cytometric data) or mean (genome sequence) into two new bins along the high variance dimension (2.1 A). Data on the bin boundary line are randomly split between the two daughter bins. Each daughter bin is similarly split until the desired number of bins is obtained. This process enabled all the bins to have similar counts. The centroids, which are the signatures of the data, and weights of each bin are calculated for both the control and sample data. The test statistics are then calculated as described in sQF applications using these described centroids and weights[82, 83] The weight vector, which represents the probability of data points falling

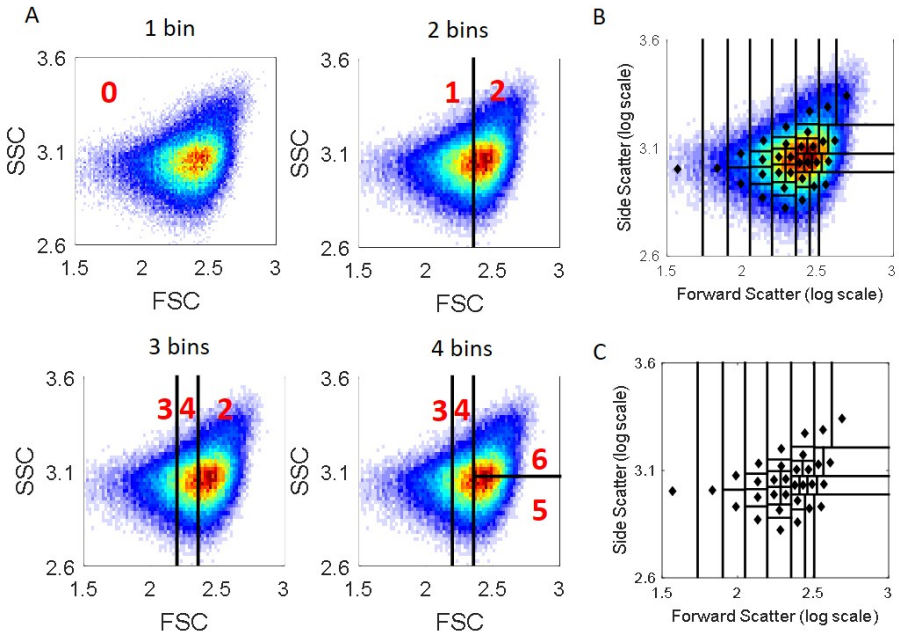

Figure 2.1: PB-sQF procedure. (A) Probability Binning. The mother bin is divided at the median of the of the highest variance dimension to reduce the within-bin variance. The raw data is taken as bin-0, labeled by the red number. After the first cut, it generates 2 bins: bin-1 and bin-2. The same procedure is applied on bin-1 to generate bin-3 and bin-4 (total 3 bins); on bin-2 to generate bin-5 and bin-6 (total 4 bins). The plots here are pseudocolor cytometric data with red indicating higher data counts. FSC: forward scatter. SSC: side scatter. (B) After binning, the median data point in each bin is taken as the centroid (black diamond). (C) The signatures of the data can be captured by the centroids, or the black dots. All inter-centroid distances are calculated through matrix multiplication, yielding dissimilarities that grow with increased difference between datasets.

into each bin, is defined as the number of counts per bin normalized by the total number of counts. The weight vector includes weights from the control and the sample as shown below:

$$Weight = (w_c^1, w_c^2, ..., w_c^N, -w_s^1, -w_s^2, ..., -w_s^N) \tag{2.1}$$

The subscripts indicate whether the weights were taken from the control (C) or the sample (S), and the superscripts indicate the bin for which the weights are calculated. The total number of bins is denoted by N. The negative sign in front of the sample weights is used to make sure that subtraction is carried out between the control and sample data in the later steps to measure the differences between the two.

The centroid vector was calculated from the geometric median (or Geometric quantile, see below) to represent multidimensional median:

$$Cent = (C_c^1, C_c^2, ..., C_c^N, C_s^1, C_s^2, ..., C_s^N) \tag{2.2}$$

The subscripts and superscripts were the same as in the weight vector. The centroids were then used to calculate the similarity matrix. We defined the matrix elements at i^{th} row and j^{th} column, A_{ij}, in the similarity matrix, A, as:

$$A_{ij} = 1_{ij} - \frac{L[Cent(i), Cent(j)]}{L_{max}} \tag{2.3}$$

in which the first term is the ij^{th} element in a 2Nx2N matrix of 1s. The second term is the dissimilarity matrix with the numerator denoting the Euclidean distance between centroids of bins i and j. The denominator is the normalized factor with L_{max} indicating the multidimensional maximum distance. For flow cytometry data, the size of each dimension, l, is the same. As a result, the maximum distance L_{max} of n dimensions is set as $\sqrt{n} \cdot l$. Even when the dynamic range of each dimension of the cytometry data is different, the maximum distance can be calculated accordingly. For genome sequence analysis, the maximum distance between centroids is $\sqrt{2 \cdot kmerlength}$. A detailed explanation is given in Section 2.4.1.

When the two centroids are exactly the same, the numerator is 0, meaning no dissimilarity exists. On the other hand, due to the normalization, when the distance between two centroids equals the dynamic range (the largest possible distance between two centroids), the dissimilarity is 1, representing a full dissimilarity. The similarity matrix, which is the logical opposite of the dissimilarity matrix, is 1 minus the dissimilarity as shown above. The diagonal elements of the similarity matrix are always 1, indicating that each centroid

is most similar to itself. The test statistics were calculated using the QF matrix operation:

$$D = \sqrt{Weight \cdot A \cdot Weight^T} \tag{2.4}$$

$Weight^T$ is the transpose of the Weight vector. Theoretically, D^2 can range from zero to two. The minimum occurs when the two distributions are exactly the same. The maximum two happens when all the centroids in each sample are the same while the centroids distances between two samples is L_{max}. The test statistics of antibiotic-treated data were then normalized by the 99% confidence level of its paired-control to yield a "fold distance" that can be compared across samples. In this study, we calculated 3-dimentional test with 128 bins.

2.4 PB-sQF for Cytometric Antibiotic Susceptibility Testing

Due to bio-variability, machine fluctuations, and/or different operating personnel, each flow cytometry data set is a different subsample from an unknown mother distribution. Thus, none of the cytometric data is completely identical among replicates even though they were all sampled from the same mother sample. In order to correlate the changes in PB-sQF test statistics with biologically relevant changes in antibiotic-induced cell damage, confidence levels and error bars were constructed. For a right-tailed statistics test, any data with test statistic that is greater than the confidence level is viewed as confidently different from the control data. Error bars, on the other hand, assess how confident each test statistics calculation is. The higher the confidence, the narrower the error bar. In the experiment, bacteria were treated at 1/16x, 1/4x and 1x of the minimum inhibition concentration (MIC) of each antibiotic. An effective antibiotic treatment will induce statistically significant changes in test statistics at 1x MIC (or even at 1/4x MIC) from the control-1/16x confidence level beyond error bars.

The following subsections describe how to construct the confidence levels and error

bars.

2.4.1 Confidence Level Estimation

The bootstrap method was used to determine the 99% confidence level. The post-blood culture flow data of the no-antibiotic control and the 1/16x data were treated as two mother distributions, and 70 daughter distributions with the corresponding sample size, ranging in bacterial counts from 4*(number of bins) to (8000 + minimum sample size) with step size 400, were sub-sampled from each mother distribution. For pre-blood culture cytometric data, since the collected bacterial counts varied and most of the time much fewer than 100,000 counts, the sample size ranges from 4*(number of bins) to (1/10 of sample size) with 20 steps instead. These distributions were then binned, and the PB-sQF protocol was followed to calculate the test statistics. For each sample size, test statistics were calculated between all 70 of the no-antibiotic control sub-distributions and all 70 of the 1/16x data sub-distributions. The distance measurements between all pairs of control-1/16x MIC daughter distributions yield a distribution of test statistics values, resulting from random sub-sampling from the mother distributions (biological variability). The 99% confidence level of all the test statistics at each sample size can be determined. The 99% confidence distances, which 99% of values are belows represents the confidence limit for a given sample size, decrease as the sub-sample size increases and can be fit by an equation similar to the standard error of the sample mean: $Conf(n) = a0 + \frac{a1}{\sqrt{n}}$, where $Conf$ is the 99% confidence level value, n is sample size, and a0 and a1 are fitting parameters. Here, a0 is the unknown confidence level of the population. According to central limit theorem, the test statistics distribution of the sub-distributions should approximate a Gaussian distribution at large sample size. Thus, the uncertainty (standard error) in estimating the population's confidence level should follow $\frac{a1}{\sqrt{n}}$. The observed confidence level then decreases with the inverse square root of the number of data points/bin. As sub-sample size increases, the deviation becomes smaller and the estimation converges to the population confidence level.

From the fitting, we can then estimate the 99% confidence level of the mother distribution with a sample size, n = 100,000 (for post-blood culture test).

2.4.2 Error Bar Determination

The error bars in the test statistics bar chart combine two different uncertainties. One results from biological variability while the other arises from the dispersion of data points in each bin (i.e. binning error). Biological variability is estimated by the standard deviation of the normalized test statistics of the triplicate data. Errors from the binning account for the uncertainty in accurately determining the centroid position within each bin. Median absolute deviation (MAD) is used to measure the dispersion of each bin since it is more robust toward outliers compared to standard deviation. The MAD is calculated as follows:

$$MAD = median[abs[X_i - centroid]] \tag{2.5}$$

That is, it is the median of the absolute distance between each data point, X_i, and the centroid, which is the multidimensional median of each bin.[143, 144] The MAD can then be used to estimate the standard deviation by:

$$\sigma_{perbin}^{MAD} = \frac{MAD}{\phi^{-1}(\frac{3}{4})} \tag{2.6}$$

where $\phi^{-}1$ is the inverse of the cumulative distribution function or the quantile function.[144] As a result, the standard deviation estimated from MAD is the MAD divided by the 75% quantile, which was determined by geometric quantile (which is a multidimensional quantile, see below). The uncertainty from each bin, σ_{perbin}^{MAD}, was then propagated to yield the final binning uncertainty for replica i, $\sigma_i^{binning}$.

The binning uncertainty from each replica was then pooled together to estimate the variance of the unknown population, where all triplicate data were presumably sampling

from this same unknown population,

$$\sigma^2_{binning} = \frac{\sum_{i=1}^{k}(n_i - 1)(\sigma_i^{binning})^2}{\sum_{i=1}^{k}(n_i - 1)} \tag{2.7}$$

in which k = 3, since we have triplicate data. n_i is the sample size of each replica, which will be close to 100,000 data points (n_i will be exactly 100,000 if no gate is applied before analysis). The errors from triplicate data and from the binning process were then propagated together to yield the final uncertainty.

$$\sigma^2 = \sigma^2_{Tri} + \sigma^2_{binning} \tag{2.8}$$

The error bars in the bar charts are one standard deviation above and below the test statistic value.

2.4.3 Geometric Quantiles

Geometric quantiles are applied first in determining the centroid of each bin and second in estimating the error from binning. The whole process uses the quasi-Newton method to solve an unconstrained minimization problem. The target function that we minimize here, as described by Chaudhuri[145] is:

$$f(\vec{Q}^{(m)}) = \sum_{i=1}^{n} [|\vec{X}_i - \vec{Q}^{(m)}| + \vec{u} \cdot (\vec{X}_i - \vec{Q}^{(m)})] \tag{2.9}$$

in which n is the number of data points in each bin; $\vec{X}_i$ is the data point; and $\vec{Q}^{(m)}$ is the quantile of the m_{th} iteration; $u = 2\alpha - 1$, where α is fractional quantile. For example, $\alpha = 0.5$ if we are calculating median (50% quantile). The target function then reduces to $f(\vec{Q}^{(m)}) = \sum_{i=1}^{n} |\vec{X}_i - \vec{Q}^{(m)}|$. The median is the Q that minimizes the sum of distances between each data point to the median. When u is non-zero, the second term in the target function takes the deviation from the median into account. Our initial guess, $\vec{Q}^{(0)}$, is the

1-D quantile in each dimension. $\vec{Q}^{(1)}$ is estimated using the following equations:

$$\vec{Q}^{(m+1)} = \vec{Q}^{(m)} + \vec{s}^{(m)} \tag{2.10}$$

$$\vec{s}^{(m)} = -\frac{\nabla f(\vec{Q}^{(m)})}{\nabla^2 f(\vec{Q}^{(m)})} \tag{2.11}$$

in which $\vec{s}^{(m)}$ is the increment determined by the first- and second-order derivative of the target function at the current iteration. For each step, we need to examine whether $f(\vec{Q}^{(m+1)}) < f(\vec{Q}^{(m)})$. If not, we need to choose a better $\vec{Q}^{(m+1)}$.[146]

The iteration stops when either (1) the iteration has been carried out 50 times or (2) the relative gradient in Q is smaller than the stopping criteria we set. The relative gradient is:

$$relgrad(Q) = \frac{\frac{f(\vec{Q}^{(m)}+\delta)-f(\vec{Q}^{(m)})}{f(\vec{Q}^{(m)})}}{\frac{\delta}{\vec{Q}^{(m)}}} \tag{2.12}$$

In this work, the iteration was stopped when $relgrad(Q)$ is smaller than 10^{-4}.

2.4.4 Convergence and Linearity

The quantitative PB-χ^2 methods[79, 80] from which we have adapted our binning methods have been a significant advance in analysis. Inter-sample differences, however, do not yield linear distances from calculated PB-χ^2 test statistics, suggesting that PB-χ^2 is not a statistical metric, and preventing all samples from being directly compared on the same distance axis. This was confirmed as the differences between the data and sample increased, a linear increase was not obtained in the test statistics value (Figure 2.2 A). Thus, test results between sub-distributions could not be directly aligned on the same scale, clouding direct comparisons to paired controls. Importantly, data with larger PB-χ^2 test statistics vs. the same control with the same binning pattern, greater test statistic differences indicate smaller similarity with the control. As the response curve for any given data set is unknown, this nonlinearity precludes knowing how different the two data sets actually are. Further,

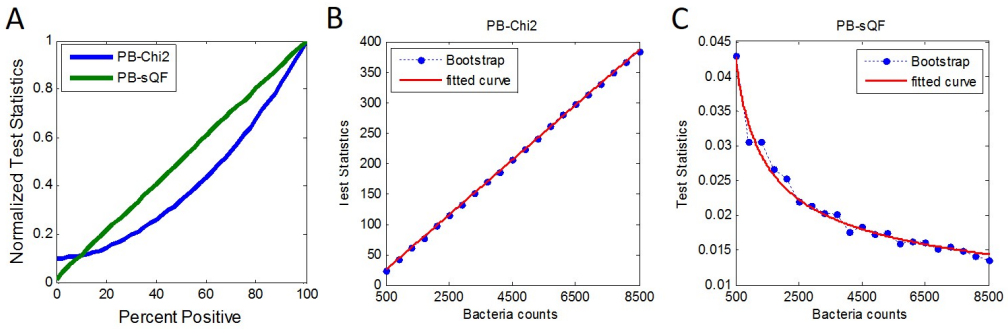

Figure 2.2: Linearity and convergence of PB-sQF and PB-χ^2. (A) The linearity of PB-sQF and PB-χ^2. PB-sQF showed a linear relation between the test results and the percent positive while PB-χ^2, although linearly increased, the linear relation could not be found. (B and C) The 99% confidence level of (B) PB-χ^2 and (C) PB-sQF. The confidence level of PB-χ^2 grew with bacteria counts while it reaches a limited value in PB-sQF and could be fitted with an equation derived from the standard deviation of the mean.

since all the data must be binned according to the control binning pattern, the response curves might possess different curvatures when different controls are used. Confounding comparisons among data from different days or machines, this control-dependent property might also contribute to error in test statistics when triplicate (or more) data with their own paired-controls are considered. Nonetheless, the bootstrap method can also be applied to PB-χ^2 and a 99% confidence level can be determined from fitting (Figure 2.2 B). This 99% confidence level can be used as a threshold and a similar bar chart regarding antibiotic susceptibility can be created. Although the fold distance is less meaningful here, a similar AST can still be built. PB-sQF directly circumvents these problems by combining the linear aspect of QF statistics with probability binning to better interpret the flow cytometry data, each with its own optimal binning.

For both the test of convergence and test of linearity, the azithromycin-treated *E. coli* data and the paired, no antibiotic control were used to perform the tests. The test of convergence was exactly the same as the determination of confidence level as described earlier. seventy sub-distributions of the no-antibiotic control and of the 1/16x MIC each ranging from 500 counts to 8500 counts were randomly generated from each mother distribution. 3-dimensional and 125 bins of PB-χ^2 and PB-sQF were then applied respectively to cal-

culate the test statistics between these sub-distributions. For the linearity test, the 1x MIC data were treated as 100% positive data while the no-antibiotic control data were viewed as 0%. Data points from the 100% positive data and the 0% positive data were then randomly selected and mixed into a spectrum of new fictional data set with sample size of 100000, ranging from 2.5% positive to 97.5% positive with 2.5% as the step increment. Test statistics were then calculated between the fictional data and the control data by either PB-χ^2 or PB-sQF. The test statistics from both PB-sQF and PB-χ^2 were normalized by their own largest distances (the test result between the 100% positive data and the control) for comparison purpose (Figure 2.2 A).

2.5 PB-sQF for Genome Sequence Analysis

Although PB-sQF was developed to calculate the distance between flow cytometry data, it can be used to analyze any type of multidimensional histogram data. To calculate the similarity between genome sequences, the sequence, which is a string data, is first converted into a numerical histogram. I also tailored PB-sQF to better quantify the differences between any two genome sequences. The pairwise PB-sQF distances are then used to type bacterial species and can be used to build a phylogenetic tree.

As in all other short reads mapping methods, the reference index must be constructed before reads alignments. To map the short reads to a reference sequence with PB-sQF, a library reads set must be built. The distances between short reads and library reads were then calculated to find the shortest distance. To reduce the library reads searching time, each test statistics calculation is used to update (narrow down) the new library reads search space. This is possible because of the linear distances between sequences when calculated with PB-sQF. The mapping results can then be imported to CNV-MM for CNVs analysis.

This section describes how to transform the genome sequence into a numerical histogram, what modifications were done to PB-sQF, how to construct the library reads for short reads mapping, how to reduce the library size for each run, and how to call CNVs

using the mapping results.

2.5.1 Adapting PB-sQF for Sequence Analysis

The genome sequences (both the assembled genome and short reads) were first k-merized using KAnalyze.[147] The k-mer set was then digitized as:

$$A \longrightarrow 1 \tag{2.13}$$

$$C \longrightarrow 2 \tag{2.14}$$

$$G \longrightarrow 3 \tag{2.15}$$

$$T \longrightarrow 4 \tag{2.16}$$

The occurrence distribution of letter sequences of all k-mers is then transformed into a k-dimensional histogram containing each unique k-mer and its corresponding k-mer counts. The unique k-mer letter sequences were then transformed into k-dimensional coordinates (Fig. 2.3) and can then be binned by probability binning as described in Section 2.2.

To obtain the centroid ("signature") of each bin after binning, the k-dimensional k-mer sequences within each bin were expanded into 4*k dimensional binary data points to avoid bias in distance calculation in the next step. With p representing the position of the nucleobase in the k-mer coordinate, the position in this new 4*k dimensional space where a particular nucleotide exists, p', is calculated as follows:

$$p' = (p-1) * 4 + X \tag{2.17}$$

in which the X is the digitized nucleobase. For example, a guanine (X = 3) in the 3^{rd} position of the original dimension (p = 3) will become an one at the 11^{th} position (p' = 11). So that $X = 1$, 2, 3 or 4 from the original dimension becomes $(1,0,0,0)$, $(0,1,0,0)$,

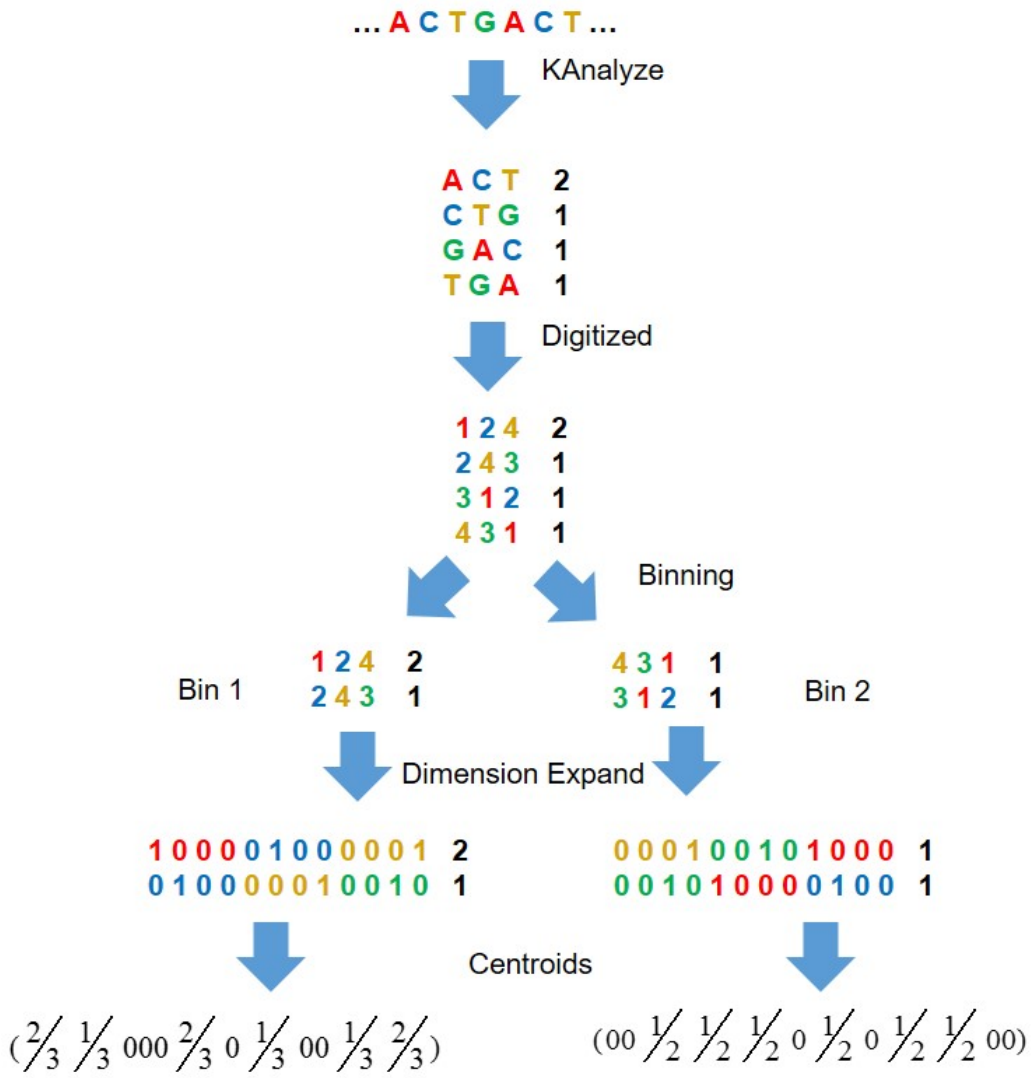

Figure 2.3: Modified PB-sQF procedure for genome sequence analysis. Converting raw sequence to centroid signatures. The letter sequence was divided into k-mers by KAnalyze and imported into MATLAB. k-mer sequences were digitized as points in k-dimensional space and binned by probability binning. Within each bin, the data was averaged to generate the centroids.

$(0, 0, 1, 0)$, or $(0, 0, 0, 1)$ in the new dimension as shown in the dimension expansion step in Figure 2.3. The average along each dimension of all the data points in each bin is then taken as the centroid of a bin (as schematized in Fig. 2.3). As a result, centroids are a measurement of how frequently a certain nucleotide appears in each of the k-dimensions, that represents how significant this nucleotide in each bin is. Different from the cytometric data where data points in each bin could form a broad distribution due to the wide dynamic

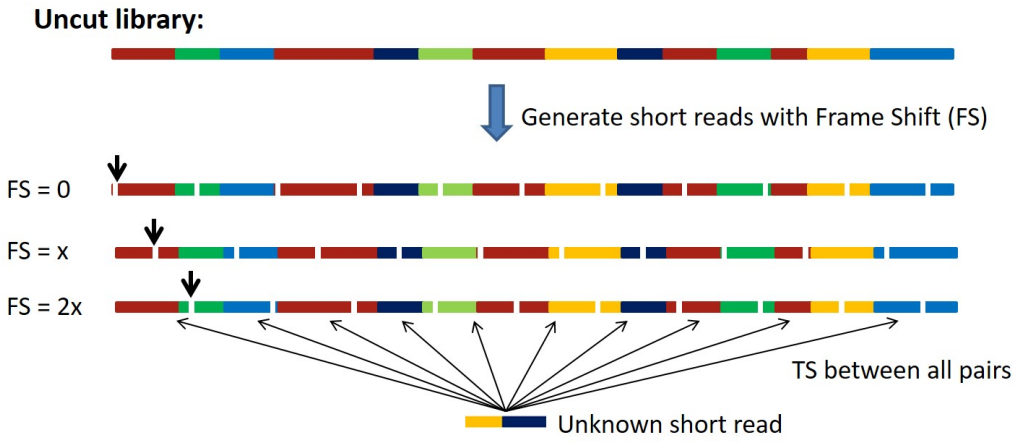

Figure 2.4: Illustration of short reads library construction. The complete library sequence is cut into to pieces at the reads length. A frame shift, x, is applied to generated more library reads. The test statistics are then calculated between the unknown reads and the library reads.

range, genome sequence only have 4 possible values (1, 2, 3, or 4). As a result, when calculating centroids, mean is calculated for genome sequence analysis while median is used to avoid outliers in the cytometric data.

After obtaining the centroids, the similarity matrix, A, can be calculated as Equation 2.3. The maximum distance, L_{max}, is $\sqrt{2 \cdot kmerlength}$. Since the largest distance between two centroids happens when they have totally different nucleotides and the Euclidean distance is the square root of the summation of difference squares over each dimension. As a result, for each k-dimension, the maximum distance is $1^2 + 1^2 = 2$ before taking the square root, and the maximum distance for k-mer data is 2*k-size before square root. The test statistics were then calculated using Equation 2.4.

2.5.2 Short Reads Library Construction

Bacterial library sequences were obtained from NCBI (`ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/`). To map the short reads onto the library strain, the library strain was chopped into similar length reads with varied frame shifts to ensure a complete reads library. Using 200-mer reads as an example, the first

library was the first 200-mer of the sequence. Shifting this 200-mer window every 200-mer would give the simplest and smallest library of a sequence. To get a larger library, the window could be shifted by 50, 20 or 10 bases instead as shown in Figure 2.4. The smaller the shift, the more library members one gets. 10-mer shifts were used for reads lengths equal to or larger than 50-mers. For 36-mers reads, 6-mer shifts were used. These reads were then k-merized, binned and saved as the reads library for each strain. To map the short reads to the library sequence, test statistics between the unknown reads and library reads. The unknown read is mapped to the library read with the smallest test statistic.

2.5.3 Simulated Reads Generation

Simulated short reads were generated from one of the library sequences and "unknown" reads were randomly generated from throughout the library sequence. To mimic real-world conditions, three different error sources were considered: uniform error, SNPs and insertion/deletion (indel). Both uniform errors and SNPs are single nucleotide mismatches where uniform errors are introduced by the sequencer while SNPs are from point mutations. The indel, on the other hand, can span several nucleotides. In the simulation, the indel rate is normally set at 2% unless indicated otherwise. The indel rate governs the possibility of the onset of a indel at each nucleotide. Once a indel starts, it's 50% chance to be either an insertion or a deletion. The length of a indel is set to either follow a geometric distribution or at a constant length. A geometric distribution was chosen since once an indel occurs, there are only 2 possible outcomes: stay in the indel or quit the indel. For each base, the probability to successfully quit the indel is set as p. The geometric distribution describes the number of failing attempts it takes before it successfully quit the indel. The number of failing attempts is the indel length. In the simulation, the probability for each base to successfully escape from the indel was set at 0.3.

Different from the single-end sequencing where each DNA fragment is only sequenced from one end, paired-end sequencing generates paired-reads that are sequenced from both

ends. Since the length of DNA fragments is determined by the experimental procedure, the distances between the two ends, i.e. the insert size, are known to be within a certain range. Reads that can be mapped to multiple locations due to the repeated regions can thus be mapped to the correct origins if their paired-reads are mapped concordantly. If the repeated region is too long, however, unique mapping locations might not be obtained for both reads.

Due to a new DNA strand is synthesized from the 5' to 3' end direction, paired-reads have opposite direction. In our simulation, the direction (forward or reverse) of read 1 in a read pair was chosen randomly and the read 2 was the opposite direction of read 1. The insert size followed a Gaussian distribution with mean equals 380 bps and the standard deviation is 50 bps.

2.5.4 Reduce Short Reads Search Space

Different from the whole sequence typing in which one strain is one library, in short reads mapping, there are more than "$\frac{GenomeSize}{ReadSize}$" reads library per strain (more because frame shift applied). The number of library reads greatly increases the calculation time. Since the test statistic calculated from PB-sQF is a true linear metric, the test statistics can be compared directly through a common control. The calculation time can thus be reduced by updating the search space with every test statistic that has been calculated.

To reduce calculation time, 50 control library reads were selected. The first control read was the first ReadLength-mer of the library sequence and the test statistics between the first control read and the rest of the library reads were calculated. To better represent the full dynamic range of library sequence, the test statistics results were sorted and the library read with the largest test statistics was chosen as the 50^{th} control read. The other 48 control reads were evenly distributed from the 1^{st} to 50^{th} control reads according to their test statistics results. This ensured that the 50 control reads were good representations of the complete library reads and can better help narrow down the search space in the next

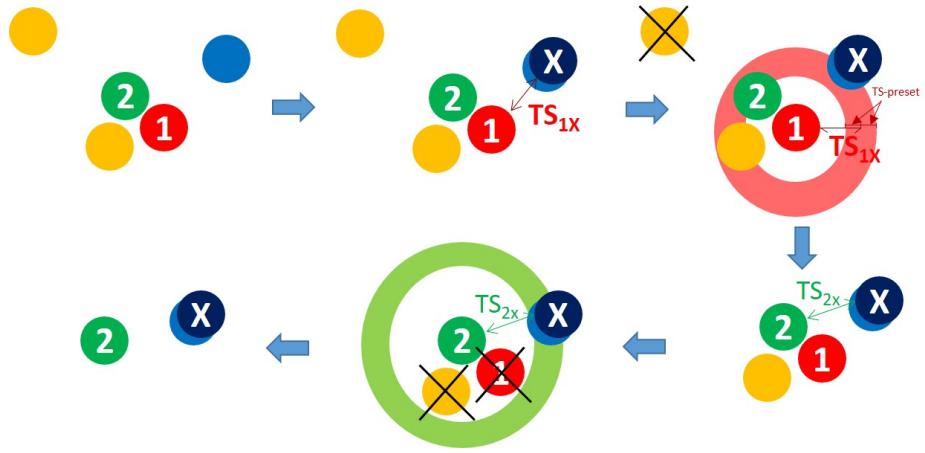

Figure 2.5: Reducing search space by linear metric distance. Numbered balls represent the control reads. The X marked ball is the unknown read. The yellow balls are the library reads which are not selected as control reads. The non-labeled light blue ball is the library read that serves as the mother read for unknown X. First, The test statistics between control read 1 and the unknown is calculated (TS_{1X}). An acceptable test statistics range is constructed as $TS_{1X} \pm TS_{preset}$ as shown in the red ring. Any library read and control read that doesn't lie within this range is excluded from further analysis since the probability for the unknown to map to that read is low. The same process is carried out again with control read 2.

step. After the control reads were selected, the test statistics between these 50 control reads and all the library reads were calculated. The library test statistics results were saved in a 50$\times$ "Total number of reads" matrix, M_{TS}, with the rows representing different control library reads and columns being all the library reads. The control library reads selection and test statistics calculation can be done in advance and only needs to be done it once for each reference sequence for a given read length. This process stops when either all 50 control reads are used or the updated library size is smaller than the number of remaining control reads.

When searching for the best-matched reads library, test statistics were calculated between the unknown read and control read 1. The test statistics result was then compared to the saved test statistics library (first row of M_{TS}). Since PB-sQF is a (linear) metric, the distances between the library reads and the unknown read can be inferred from the distances between the library reads and the control reads. Any library read that has a pre-

calculated test statistics much larger or smaller than the calculated value, the chance that the unknown originated from this library read was low and was thus excluded from further analysis. An example is given in Figure 2.5. Here, the unknown read X originates from the light blue ball (mother read) with some sequencing errors and/or mutations. The test statistics, TS_{1X}, between the unknown X and control read 1 is calculated. Since the unknown is derived from the mother read which is one of the library reads, the test statistics between the mother read (light blue ball) and control read 1 must be very similar to TS_{1X}. By setting a pre-defined test statistics range (TS_{preset}), any library reads that are within $TS_{1X} \pm TS_{preset}$ (the red ring in Fig. 2.5) are kept. Library reads that have test statistics with the control read 1 that are very different from TS_{1X} (library reads outside the red ring in Fig. 2.5) can thus be safely excluded without calculating the test statistics between those reads and the unknown therefore reduce the calculation time.

2.6 Nearest Neighbor (NN) Distances

NN, which is also a Euclidean distance-based sequence similarity method, has been applied in 16S genome sequence typing.[27, 101] However, probably due to high dimensionality at large k-mer size, NN has not been applied to whole genome sequence typing or short reads mapping. Nevertheless, when analyzing the k-mer frequency data, I discovered that NN can achieve high performance by analyzing 3-mer data with full frequency counts, treating as a probability distribution, and therefore circumvents the high dimensionality problem. In NN, each unique k-mer is treated as one dimension so each genome sequence is represented in a 4^k dimensional space with the coordinate being the k-mer counts (Fig. 2.6). The test statistics is the 4^k-d Euclidean distance between the coordinates. Since both PB-sQF and NN calculate the metric distance between 2 histograms, NN can be applied in a similar way as PB-sQF. The short reads library can be constructed by chopping the library sequence and the search space reduction method can also be applied. For short reads with 36, 50, 76, 100, 150, 200, 250 and 300-mer, the pre-defined test statistics range in search

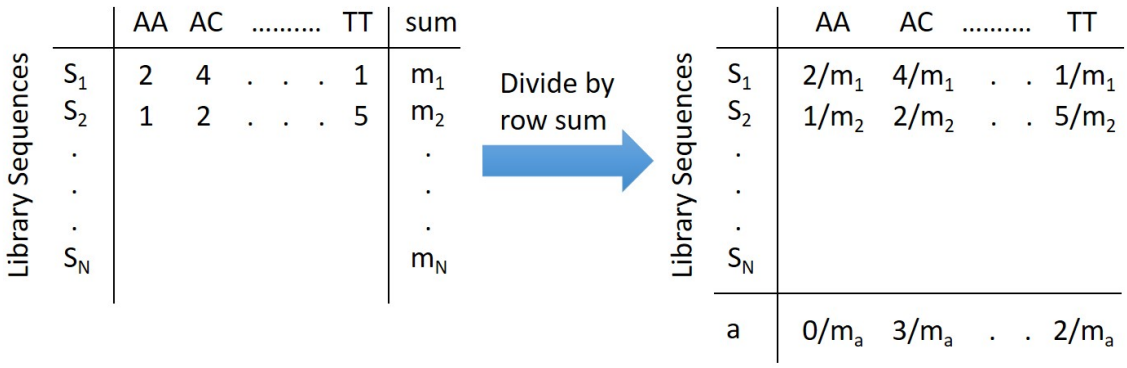

Figure 2.6: Nearest neighbor method A 2-mer example. Each unique 2-mer serves as an independent dimension. For each sequence, $S_n, n = 1 \sim N$, the k-mer counts and the sum of counts are obtained. The 4^k coordinate of each sequence is the k-mer frequency. To find the mother sequence of an unknown sequence, a, the sequence is k-merized and the 4^k coordinate is acquired for comparison. The Euclidean distances between the unknown and every library strain are calculated and the unknown is typed to the library strain with the smallest test statistic.

sapce reduction are set as 0.07, 0.06, 0.05, 0.045, 0.04, 0.035, 0.03 and 0.03.

2.6.1 Valid Short Reads Assignments

NN calculates test statistics between the query reads from the donor/target sequence and the library reads even when the library reads are not the correct mapping locations. As a result, when using NN to report multiple possible mapping locations, a threshold was set up to exclude the library reads that have low probability to be the correct mapping result. To set up the threshold, the genome sequence of *Syntrophomonas wolfei* subsp *wolfei* str *Goettingen* (Accession: NC-008346.1) was selected as the model sequence. For each read length, 10^6 library reads were randomly selected. Query reads were generated for each library read with frame shift and errors (1% of uniform error and 2% of indel rate with the indel length following the geometric distribution with a probability equal to 0.3). The frame shifts ranged from $\frac{FrameShift}{2}$ to $9 \times FrameShift + \frac{FrameShift}{2}$ with $1 \times FrameShift$ as the increment to either direction (Fig. 2.7 A). Reads originates beyond $9 \times FrameShift$ are not discussed here since the chance they are mistakenly mapped to the investigated library read should be low. Since NN performs read-to-read alignment and the library reads were

constructed with various frame shift, reads that were generated within $\pm\frac{FrameShift}{2}$ from a library read, ideally would be mapped back to the correct library read (Fig. 2.7 A, purple region). The test statistics calculated between the library reads and the corresponding query reads at each condition, due to errors and frame shifts, form a distribution (Fig. 2.7 B, purple). Test statistics between the library read and query reads that were from more than $\frac{FrameShift}{2}$ away and ideally were supposed to map to the neighbor library read were also calculated (Fig. 2.7 B). Left-tailed confidence levels were calculated since one wants to know the probability that a target read indeed originates from the library read. The default threshold is set at 95% confidence that a target read is mapped within 2 (for read length $<$ 100-mer) or 4 (for read length $>$ 100-mer) frame shift away from the true origin. Any library reads that have test statistics larger than this threshold were discarded.

For short reads mapping with multiple locations, the threshold for screening valid assignments can not be too strict in order to map all the reads to different copies of genes with some variations. At the same time, the threshold has to be narrow enough to reject all the incorrect mapping. The threshold, is thus set empirically for each read length.

2.6.2 SAM Format and MAPQ Score

The standardized short reads mapping results output, the Sequence Alignment/Map (SAM) format, has 11 mandatory fields including reads information such as the query names, query sequences and mapping positions.[148] NN can generate output following the SAM format, however, the CIGAR (Concise Idiosyncratic Gapped Alignment Report) field, which reports the match/mismatch and indel for each base, is unavailable (*) since as a read-to-read aligner, NN doesn't carry CIGAR information. The MAPQ (MAPping Quality) field records the mapping quality score, Q, for each mapped short read. The score Q quantifies the probability that a read alignment is wrong, P_{wrong}, and is defined as follows.[29]

$$Q = -10log_{10}P_{wrong} \tag{2.18}$$

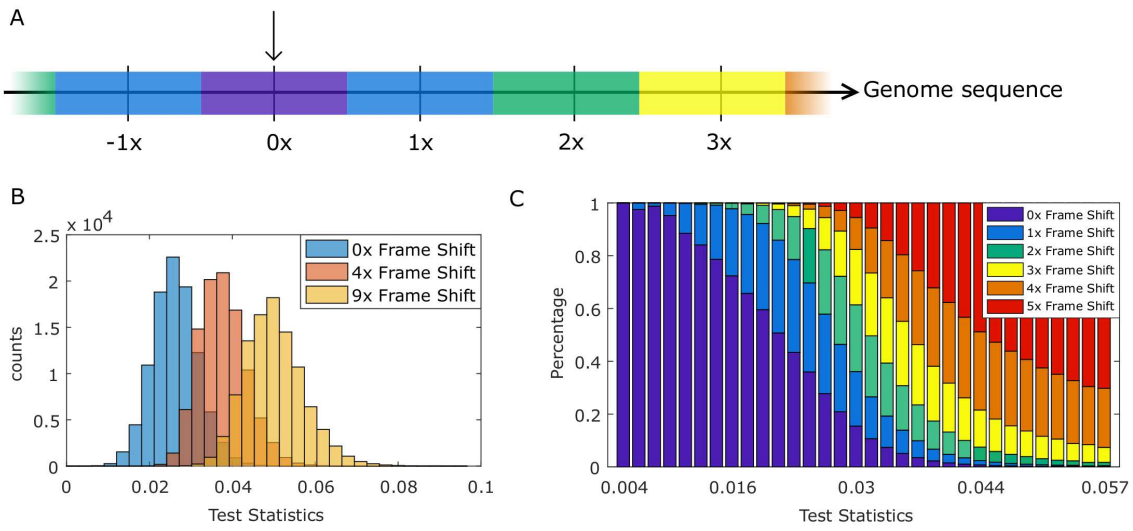

Figure 2.7: Confidence levels and incorrectly assigned probability. (A) Relation between simulated query reads and library read. The left-most index of the investigated library read is indicated by the arrow. This position is defined as 0x frame shift. The left-most index of neighbor library reads is shown as how many frame shifts away from this library read (x-axis). The shaded areas are where the query reads were generated at each condition. Query reads from the purple area, which is $\pm\frac{FrameShift}{2}$ away from the library read, were reads that ideally should map to the library read. Reads from the blue region were reads that ideally should map to library read at 1x or -1x frame shift. (B) Test statistics distributions between library reads and query reads from 0x, 4x, and 9x away from the correspond library reads as described in (A). (C) For a read with a certain test statistic value, the percentage of it originated from 0x, 1x, to 5x frame shift away from this library read.

Although all the short reads aligners use the above equation to calculate the mapping quality score, every aligner estimates P_{wrong} differently. Since neither the hash table-based aligners nor the BWT-based aligners actually calculate the distance between the query read and the reference sequence, none of these estimated P_{wrong} are statistically defined. Using our NN approach, however, P_{wrong} can be easily calculated using the test statistics distribution as described in subsection 2.6.1. Test statistics distributions were calculated between 10^6 pairs of library reads and the corresponding query reads that were generated 0x, 1x,..., 5x frame shifts away from the library reads. The test statistics from the 0x frame shift query reads were binned into 32 bins. The histogram counts for all six test statistics distributions in each bin were calculated. As shown in Figure 2.7 C, the probability that a read indeed

originated from the 0x frame shift decreases as the test statistic increases. The percent correct at each bin is thus $\frac{H_{0x}}{\sum_{i=0x}^{5x} H_i}$, with H being the histogram counts and the subscript indicates the test statistics distribution. The percent wrong, P_{wrong}, can, therefore, be calculated from $1 - P_{correct}$ and by equation 2.18, the MAPQ score for each alignment can be obtained.

2.7 Copy Number Variance (CNV) Detection

Since most of the major short reads aligners were developed for unique mapping for each read,[29, 40, 109, 112] most CNV detectors were built to analyze the read depth constructed only from the best mapping result causing them to have difficulty dealing with any repeated regions in the reference sequence.[41, 43, 125] Because current CNV detectors, consider only unique mappings and discard the multiply mapped reads, they will identify many false deletions due to the reads being removed.[42] When considering all mapping locations, without properly adjusting for the repeats in the reference sequence, a CNV detector can mistakenly assign the multiple copies that also exist in reference sequence as "true" CNV regions. Thereby resulting in incorrectly identify excess duplications.[28] Both scenarios produce high false positive rates. As a result, a new CNV detection algorithm, copy number variation detection for mapping multiplicity (CNV-MM), was developed to take advantage of and utilize the probability-based multiple location mapping data from NN. The procedure of CNV-MM will be explained in three parts: reconstructing read depths, segmentation, and copy number determination.

2.7.1 Building the Read Depth Trajectories

As shown in Figure 2.8, each donor read is first mapped onto the reference sequence. In this example, the donor read is similar to parts of gene A and gene A' with test statistics 0.1 and 0.2 respectively. After applying the threshold for valid assignments, only two assignments remain. To identify the CNV regions after rejecting invalid assignments, three different

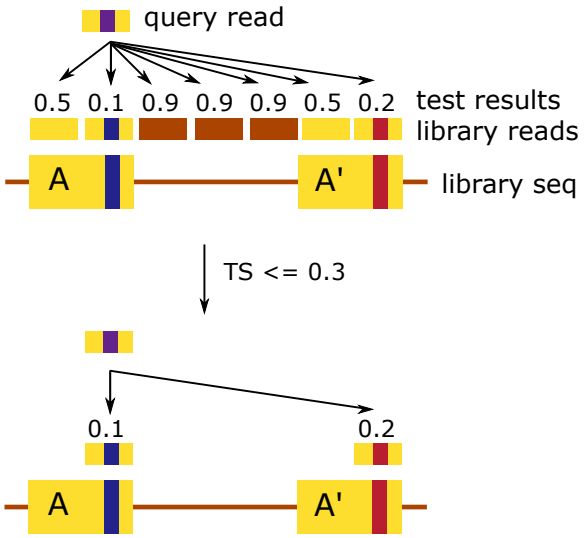

Figure 2.8: Model system with two copies of gene A. The test statistics are calculated between the query read and all the library reads (without frame shift). After threshold applied, only the valid alignments are kept. This model system is used for the rest of this section.

trajectories are constructed. For all three trajectories, read depth was assigned across the span of the read instead of only assigning at the left-most index and read depth contributed from each read was added together to form the final read depth trajectory. Even though the read depth was added across the whole read, the resolution is still determined by the mapping resolution, which is the frame shift of the reads library and the read length of the unknown reads.

Binary Trajectory

The first trajectory is the read depth generated using the binary approach. To build the binary trajectory, the query reads are either mapped (1) or not mapped (0) to a reference location.

The binary read depth at a given reference genome location, x, is the sum of all the read depths contributed from all the reads that are mapped to the given location.

$$ReadDepth_{Bi}(x) = \sum_{i=1}^{N_{ReadsMapped}(x)} RD_i^{Bi}(x) = N_{ReadsMapped}(x) \tag{2.19}$$

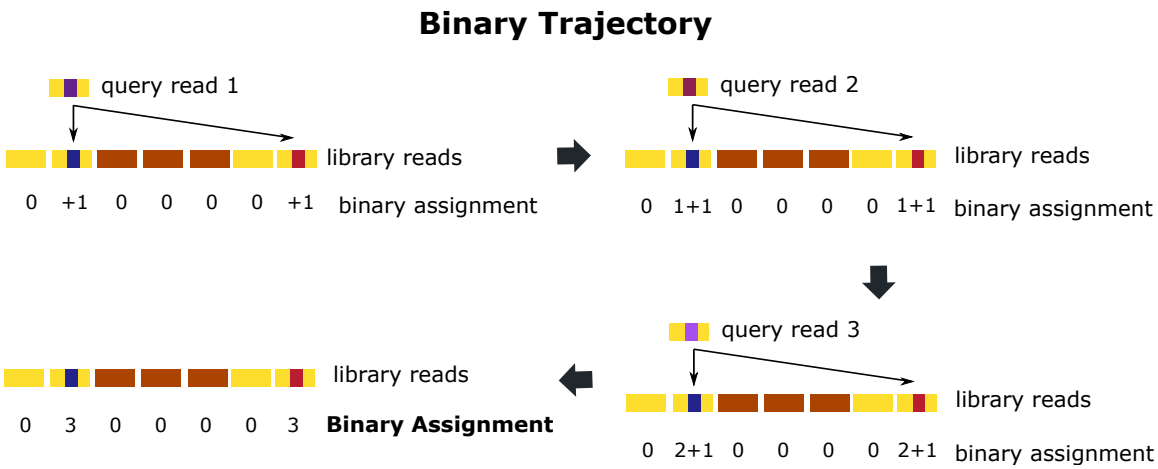

Figure 2.9: Building the binary trajectory. For each valid assignment of any given read, one (existence) is assigned to the mapped location on the reference genome. The binary trajectory is built by summing over all mapping results. The model system is the same as in Figure 2.8.

in which RD_i^{Bi} the binary read depth contributed from $read_i$, and it is one for all valid mapping locations of $read_i$. As a result, $\sum_{i=1}^{N_{ReadsMapped}(x)} RD_i^{Bi}(x)$ equals to the total number of reads mapped to the given reference index, $N_{ReadsMapped}(x)$.

As shown in Figure 2.9, the query read-1 maps to A and A'. Thus the binary assignments are one for both locations. For query read-2, it also maps to both A and A' so the read depths increase to two. The binary read depths keep increasing when more query reads are mapped to A and A'. The binary trajectory is obtained by summing over all valid mapping results from all query reads to each reference location, which is equivalent to the number of query reads mapped to a given genome position.

Since the multiple copies in the reference sequence only determine to where the query reads are mapped but not the read depth at each genome location, the binary read depth trajectory is only affected by the number of copies in the read donor (the query) sequence.

Average Trajectory

The second trajectory is the average number of assignments, $ReadDepth_{Avg}$, and is calculated at a given genome position, x, as follows:

$$ReadDepth_{Avg}(x) = \frac{\sum_{i}^{N_{ReadsMapped}(x)} TV_i(x)}{N_{ReadsMapped}(x)} \tag{2.20}$$

in which TV is the total number of valid assignments for each read,i, is that mapped to the given genome index. The numerator, $\sum_{i}^{N_{ReadsMapped}(x)} TV_i(x)$, is the total number of reference locations to which all donor reads mapped to the given genome index can be mapped. The denominator, $N_{ReadsMapped}(x)$, is the number of valid query reads mapped to a given genome location. When examined from the first nucleotide ($x = 1$) to the last ($x = SequenceSize$), $N_{ReadsMapped}(x)$ is equivalent to the binary trajectory. By normalizing the contributions from reads donor multiplicity, $ReadDepth_{Avg}$ represents the multiplicity of locations to which the reads mapped to this location are also mapped (i.e. the average total number of locations to which each read is mapped). The number of assignments reflects the copy number within the reference sequence and is not affected by the copy number of the reads donor (the query sequence). To avoid dividing by zero, the $ReadDepth_{Avg}(x)$ is set to zero when no read is aligned to the given reference index, x.

As illustrated in Figure 2.10, query read-1 can be mapped to two different locations, A and A'. Therefore, $TV_1 = 2$ for both A and A' regions. Query 2, which can also be mapped to both A and A', also have $TV_2 = 2$. Thus, at both A and A', the numerator, $\sum_{i}^{ReadsMapped} TV_i$, is $2 + 2 = 4$ while the denominator is two (total two reads mapped at either A or A'). Therefore, the average number of assignments at A and A' is two. The read depth two indicates that there are two locations in the reference genome that share similar sequence (two copies of gene A). As a result, when a third read is also mapped to A and A', the average number of assignments does not change.

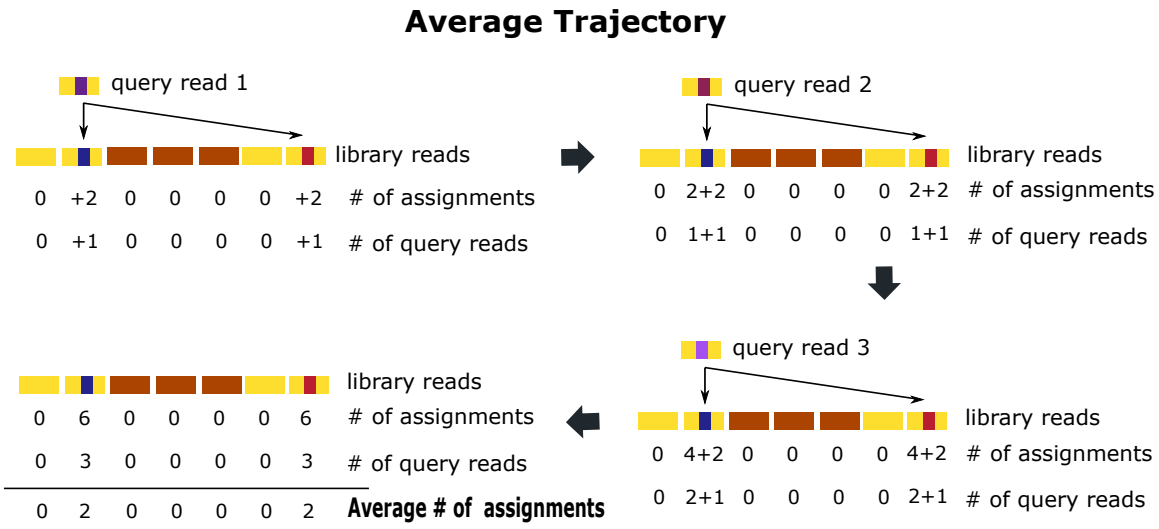

Figure 2.10: Building the average trajectory. At each reference genome location, the summation of the number of valid assignments for each query read mapped to the location is taken. This summation is then divided by the total number of reads mapped to the location, which is equal to the binary trajectory, to obtain the average number of assignments. The model system is the same as in Figure 2.8.

Test Statistics Trajectory

The third trajectory is the read depth recovered from the calculated test statistics (TS). Different from the binary and average trajectories, TS trajectory is the only trajectory that uses the test statistics information. In this trajectory, the sum of read depth of a query read is always one. When a read is mapped to multiple locations, the read depth is divided into all valid mapping locations with weights depending on the test statistics between the query read and library reads.

The TS read depth at a given reference genome location, x, is the sum of all the read depths contributed from all the reads that are mapped to the given location.

$$ReadDepth_{TS}(x) = \sum_{i=1}^{N_{ReadsMapped}(x)} RD_i^{TS}(x) \tag{2.21}$$

in which $RD_i^{TS}(x)$ the TS read depth contributed at x from $read_i$, and it is determined by the test statistics values of all valid mapping of $read_i$.

The read depth of the i^{th} assignment among total $N^{i}_{assignments}$ valid assignments of $read_i$ is calculated as follows:

$$RD_i^{TS}(x) = \frac{Weight_i}{\sum_{j=1}^{N^{i}_{assignments}} Weight_j} \tag{2.22}$$

There are different ways to determine the weights which is discussed in Appendix A and Appendix Figure A.1. Here, the weights are the inverse of test statistics as shown in Figure 2.11. As a result, the smaller the test statistic is, the larger the weight. For any assignments with TS = 0, 10^{-4} is used instead to avoid division by zero.The read depth at each location is the weight normalized by the sum of weights to which a query read can be confidently mapped. By summing over the contributions from all query reads, i, the final TS trajectory is obtained (Equation 2.21). Since the read depth is distributed to all valid assignments, which is proportional to the number of copies in the reference, and then summed over all mapped query reads, this trajectory is inversely proportional to the number of copies in the reference sequence while being directly proportional to the number of copies in the donor sequence

Since the weights are determined by the test statistics results, the TS trajectory can provide more information about the CNV regions then just the copy number. As shown in Figure 2.11, after the query read-1 is mapped, two cases are provided. The query read-2 in case 1 has the reversed weights than query read-1. Thus, the summation of read depths equals one for both assignments (gene A and A'). In case 2, the query read-2 has exactly the same weights as query read-1 which leans toward gene A. As a result, the final read depth also emphasizes on gene A. This shows that in case 2, although the query reads share similarity to gene A', they most likely originate from gene A. If the final analysis suggests a deletion in the reads donor sequence, there is a higher chance that gene A' was deleted. If the final analysis indicates a duplication in the reads donor, the duplication gene is most likely gene A. In the binary and average trajectories, no difference will be found between

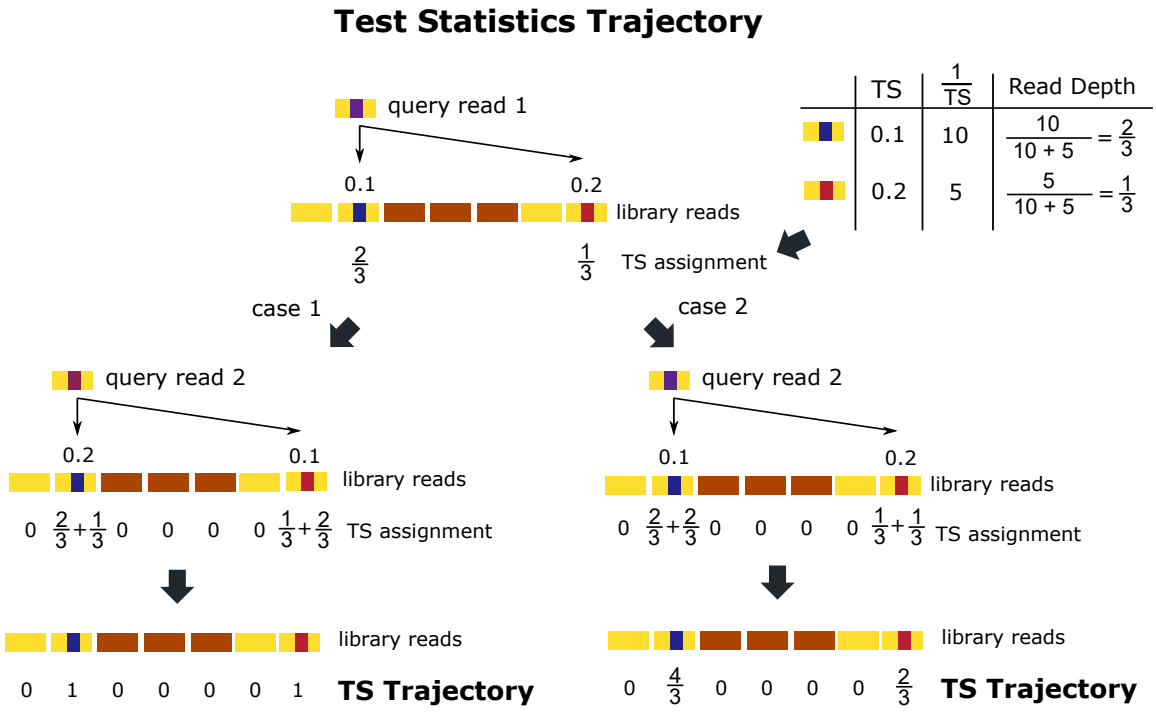

Figure 2.11: Building the test statistics trajectory. The read depth from each query read is divided into multiple mapping locations weighing by the test statistics. The final read depth is the sum of all divided read depth from each query read. Since the read depths depend on the test results, different distributions of test statistics will give different TS trajectory. The model system is the same as in Figure 2.8.

the two cases.

Evenly Distributed Trajectory

Although the test statistics trajectory provides more information about the multiple assignments, none of the short reads aligner that we know of calculates the test statistics information. In order to apply CNV-MM to any aligner of choice, the evenly distributed trajectory is used (Fig. 2.12). The read depth of the evenly distributed trajectory at a given reference genome index x is defined similar as in the binary and TS trajectories: the sum of the evenly distributed read depth at x contributed from each read, i.

$$ReadDepth_{Even}(x) = \sum_{i=1}^{N_{ReadsMapped}(x)} RD_i^{Even}(x) \tag{2.23}$$

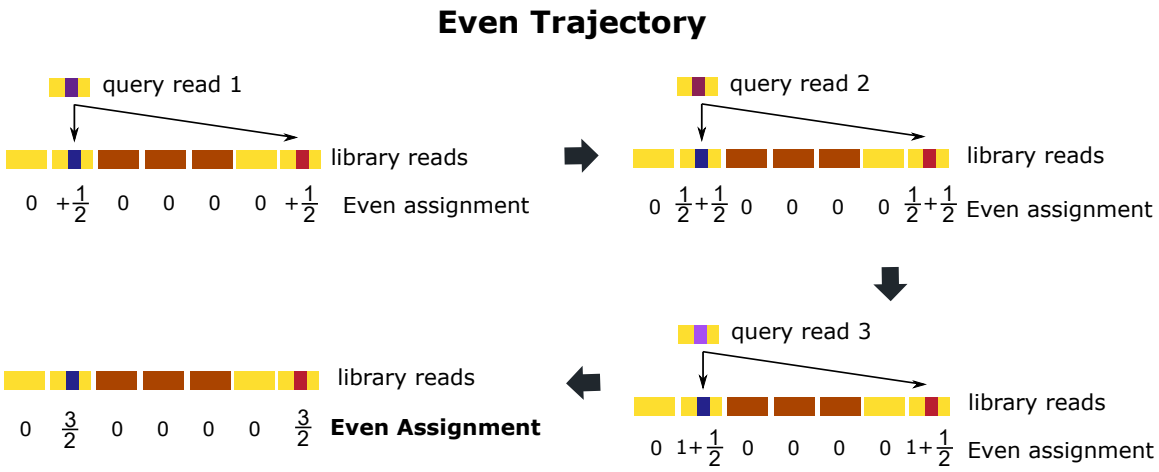

Figure 2.12: Building the test statistics trajectory. The read depth for a given query read is divided evenly to all the valid assignments. The final read depth is the summation of contributions from all query reads. The model system is the same as in Figure 2.8.

Instead of distributing the read depth to all valid assignments by their weights, the read depth, $RD_i^{Even}(x)$, is distributed evenly to all valid assignments:

$$RD_i^{Even}(x) = \frac{1}{N_{assignments}^i} \tag{2.24}$$

in which $N_{assignments}^i$ is the total number of valid assignments for query read i.

Although using the evenly distributed trajectory instead of the test statistics trajectory enables CNV-MM to be compatible with other short reads aligners, for CNV-MM to perform at its full capacity, the reads aligner needs to be able to handle the mapping multiplicity properly (discussed in Chapter 7).

Since the read depth of a given query read is evenly divided into all valid assignments and the final read depth is the summation of all query reads, the read depth of the evenly distributed trajectory is determined by the ratio of the read depths between the reads donor and the reference sequences. The relationship between the binary, average and evenly distributed trajectory is discussed next.

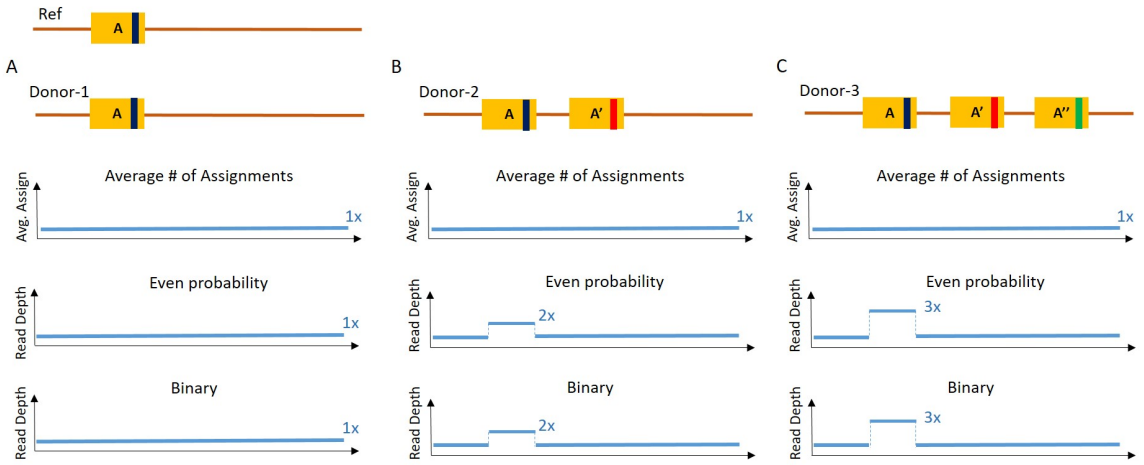

Figure 2.13: Read depth trajectories with different copy number in reads donor. The reference sequence is fixed to have one copy of gene A. The reads donor sequences, have (A) one copy of gene A, (B) two copies of gene A (and A'), and (C) three copies of gene A (A' and A").

2.7.2 Analyzing the Trajectories

To use these three trajectories (average, binary and the even trajectories) to identify CNVs, the behavior of all three trajectories were investigated with different scenarios of reads donors and reference sequences combinations. From the studies, the relationships between all three methods are deduced.

Relation between the Trajectories

First, the effects of different copy numbers of genes in the reads donor sequences are tested. As shown in Figure 2.13, the average number of assignments remains the same regardless of the copy number of gene A. This is because no matter how many short reads that are similar to those in gene A are generated from the target sequence, all the short reads had only one mapping location. As a result, the trajectory remains unchanged. The read depths for even and binary trajectories increase from 1x to 3x higher compared to the baseline. In both cases, since there is only one copy in the reference, all the extra short reads generated from gene A' and A" are all mapped to gene A on the reference genome.

To test the effects of multiple copies in the reference sequence, the number of copies of

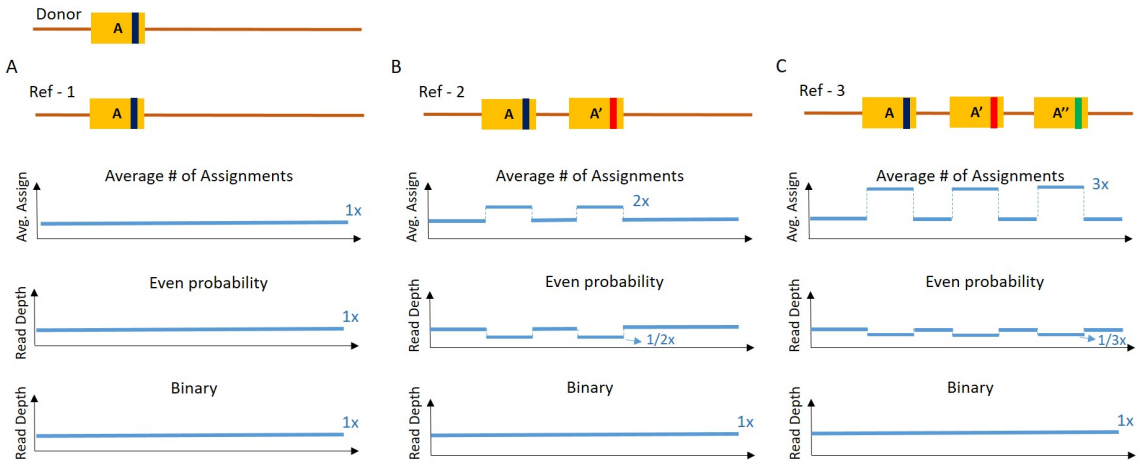

Figure 2.14: Read depth trajectories with different copy number in reference sequences. The reads donor sequence is fixed to have one copy of gene A. The reference sequences, have (A) one copy of gene A, (B) two copies of gene A (A and A'), and (C) three copies of gene A (A, A' and A").

gene A is fixed in the reads donor sequence and the copy number in the reference sequences varied. The average number of assignments increases linearly with the number of copies in the reference (Fig. 2.14). The evenly distributed read depth, on the other hand, decreases from 1x, 1/2x to 1/3x when the copy number of gene A in the reference increases from 1x, 2x to 3x. This is because as the copy number of the reference increases, the short reads generated from gene A in the reads donor sequence can be mapped to multiple locations. Since the sum of read depth contributed from one read in the evenly distributed trajectory is one and there is only one copy in the reads donor, the read depth is diluted $\frac{1}{NumberCopyinRef}$ at each gene A location. As for the binary read depth trajectory, the short reads from gene A are mapped to multiple locations on the reference genome as in the evenly distributed probability, however, since the binary approach assigns one to all mapping locations, the read depth is not diluted and remains as one.

As the average number of assignments is determined by the copy number in the reference sequence, the binary read depth trajectory is governed by the copy number in the donor sequence, and the evenly distributed read depth trajectory is affected by the ratio of

copy numbers in both sequences, these three trajectories follow this relation:

$$Trajectory^{Even} = \frac{Trajectory^{Binary}}{Trajectory^{Avg}} \tag{2.25}$$

The TS trajectory, although theoretically can not replace the evenly distributed trajectory in equation 2.25 since while the $RD_i^{TS}(x)$ is proportional to the number of copies in the read donor sequence and inversely proportional to the number of copies in the reference sequence, it is also determined by the weights of all valid assignments. The TS trajectory, however, resembles the evenly distributed trajectory in all the data we have tested so far. This is probably because, for most multiple copy regions, they share high similarity, so the test values are close. Thus, the weights at each assignment are close to evenly distributed. The equation 2.25, as a result, still holds for when replacing $Trajectory^{Even}$ with $Trajectory^{TS}$

Although the binary read depth trajectory alone can define the gene regions and estimate the number of copies of the donor sequence, it can not identify the true CNV regions since multiple copies can exist in both the donor and reference sequences. The average trajectory, on the other hand, provides the information of the copy number in the reference sequence. The evenly distributed read depth trajectory directly identifies the true CNV regions since the read depth only deviates from the baseline when there is a copy number difference between the reference and donor sequences. However, the deviations can be small. For example, if the reference sequence has ten copies while the donor sequence only has one, the ratio is $\frac{1}{10}$ of the baseline in the evenly distributed trajectory. This small deviation is very difficult to distinguish from the reads mapping Poison noise. However, the 10-fold higher number of assignments is very easy to identify in the average number of assignments trajectory. As a result, all three trajectories are used to help identify the true CNV regions.

Non-ideal conditions

It is clear that when the test values of all valid assignments are not similar, the equality of equation 2.25 might not stand when replacing $Trajectory^{Even}$ with $Trajectory^{TS}$. However, even when using $Trajectory^{Even}$, it sometimes does not equal the ratio between $Trajectory^{Binary}$ and $Trajectory^{Avg}$.

When repeated regions on the reference sequence share lower similarity, some of the query reads might only be mapped to a subset of CNV regions. This is because threshold is set to reject low-confidence mapping. While the threshold has to be relaxed to recognized repeated regions with small variations, it also has to be strict enough to remove the incorrect mapping. If repeated regions only partially resemble each other, then some of the query reads might not mapped to all the repeated regions at the same time.

As shown in Figure 2.15, this inconsistent query reads mapping makes the read depths at gene A as follows: $Trajectory^{Even} = \frac{3}{2}$, $Trajectory^{Binary} = 2$, and $Trajectory^{Avg} = \frac{3}{2}$. These read depths mean that there are $\frac{3}{2}$ of gene A in the reference sequence (from the $Trajectory^{Avg}$), two copes of gene A in the read donor sequence (from the $Trajectory^{Binary}$), and this is a true CNV region with the reads donor sequence have 1.5 times higher number of copy than the reference sequence (from the $Trajectory^{Even}$). However, these are not true.

Since related repeated regions can be grouped together using CNV-MM (See Section 2.6.4), instead of examining the read depth at each repeated region, the average read depth among related regions can be calculated. In the average trajectory, the group average of average number of assignments of gene A and A' is $\frac{7}{4}$. The average read depth for binary trajectory is $\frac{3}{2}$ and for the even trajectory (or the TS trajectory) is one. Since the read depth for each query read is always one, the average of even or TS read depth among valid assignments will always reflect the true ratio of copy numbers between the reference and read donor sequence. As a result, the group average TS read depth is the key to distinguish the true CNV regions from multiple mapping regions.

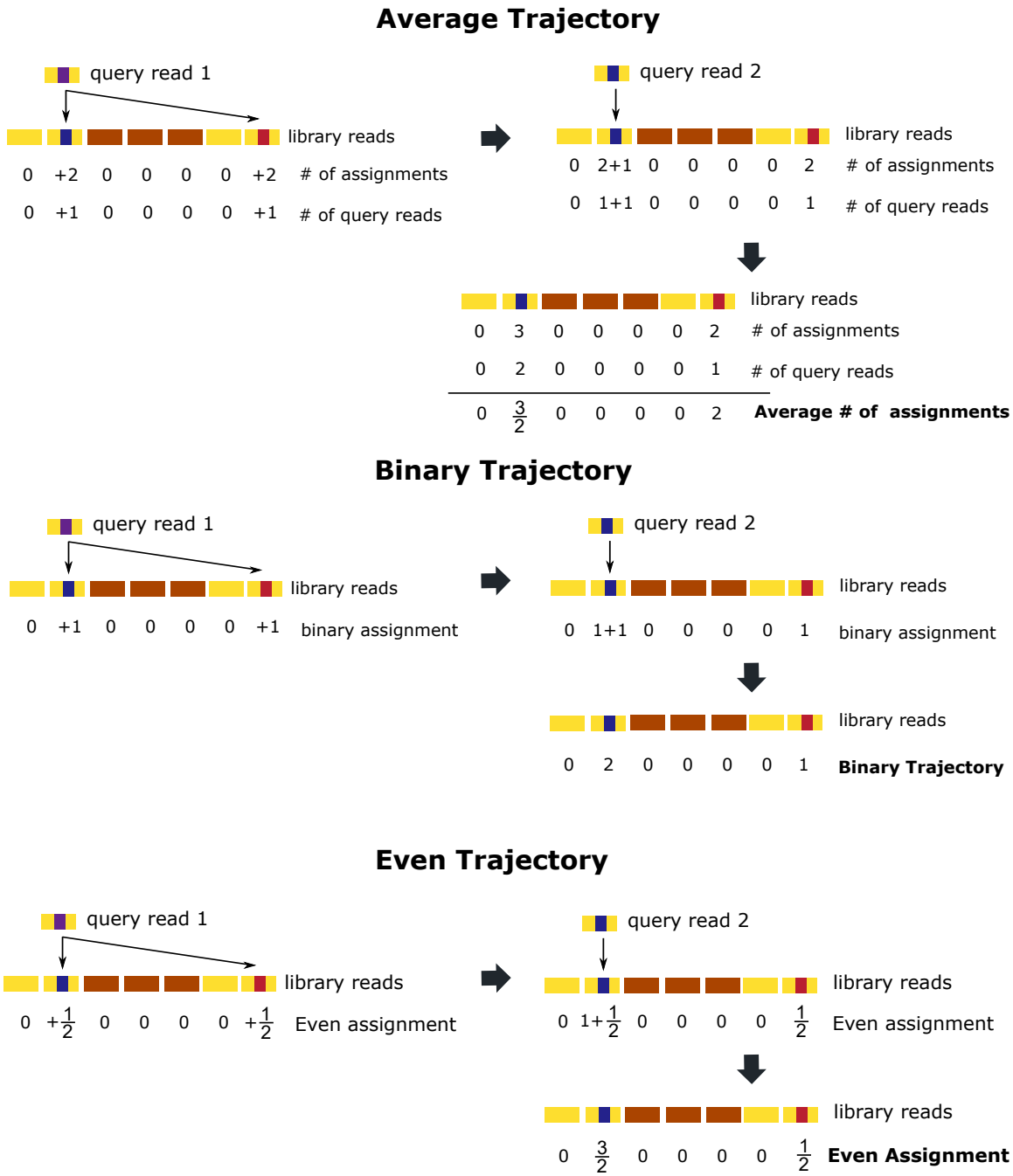

Figure 2.15: Inconsistent query reads mapped. When repeated regions on the reference sequence share lower similarity, query reads might be mapped to the reference sequence inconsistently. This invalidates equation 2.25

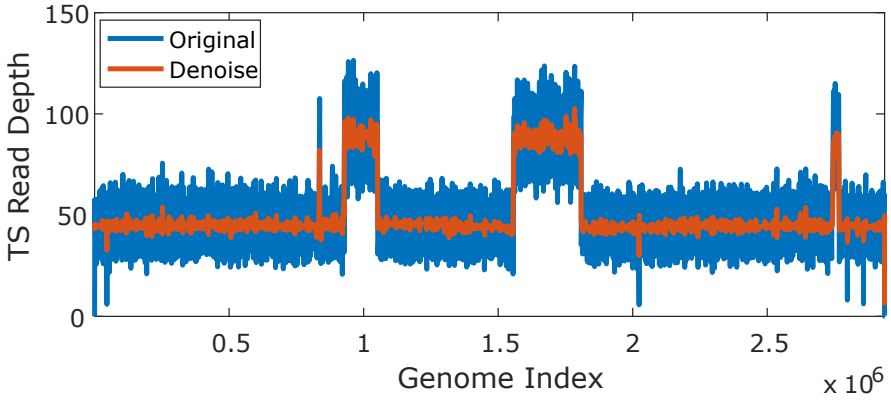

Figure 2.16: Wavelet Denoising of a TS trajectory. The blue lines and orange lines are the TS read depth trajectories before and after wavelet denoising respectively.

2.7.3 Segmentation: Finding the Copy Number Alternation Regions

After the binary and average number of assignments trajectories are constructed, possible CNV regions, which are regions with read depths deviate from the baseline (both duplications and deletions), are selected by segmenting the trajectories.

To detect the copy number varying regions, the trajectories are first denoised by wavelet transformation. Wavelet transforms have been used in signal denoising, edge detection,[149–152] and have been shown to be useful in copy number detection.[46, 153] In CNV-MM, the first and last ReadLength-bp of the trajectories are set to the mean read depth to prevent the edge effect from the mapping process. Then, discrete stationary wavelet transforms of the trajectories are performed with Haar wavelets. Haar wavelet is chosen since it is a step function which resembles the copy number transition. The level of decomposition is set as $log_2(SequenceSize)$. Each wavelet component was denoised by thresholded at three times of the mean read depth of each trajectory. The denoised read depth trajectory is reconstructed with the denoised wavelets. The denoised TS read depth trajectory is cleary smoother compared to the original trajectory (Fig. 2.16).

The mean ($AvgRD$) and standard deviation ($StdRD$) of the wavelet denoise trajectories are calculated. To find the positive peaks (peaks with read depth larger than the average read depth), simplified trajectories are constructed as follows:

$$Threshold = 2 \times AvgRD - StdRD \tag{2.26}$$

$$Tj^{simplified} = \begin{cases} Tj^{original}, & \text{if } Tj^{original} \geqq Threshold \\ 0, & \text{otherwise} \end{cases} \tag{2.27}$$

in which Tj is short for trajectory. The simplified trajectories are the trajectory with read depth equals to the original trajectories if the read depths are expected to be two times or higher than the mean read depth. To find all the peaks, the derivative of the simplified trajectories are taken to find the transition points. To reduce the false positive rate, only the transitions points that are larger than than the transition threshold are kept. Since this threshold is supposed to keep all transitions (peaks) with read depths equals or larger than $2 \times AvgRD$, but the $AvgRD$ changes from trajectories to trajectories, a dynamic threshold is needed.

To calculated the dynamic transition threshold, a 2^{nd} simplified trajectory is calculated. In this trajectory, all the peaks with read depths $\geqq 2 \times AvgRD + StdRD$ are set to $AvgRD$ and new set of transition points is calculated. Since any large peaks has been smoothed out, this new set of transition points only included peaks with $2 \times AvgRD$ and noises. As a result, when performing k-mean clustering with two clusters on the new set of transition points, cluster-1 will have transition points with the cluster mean $\sim 2 \times AvgRD$ and cluster-2 contains the transition points of noises. The transition threshold is then set and applied on the original set of transition points that were calculated from the 1^{st} simplified trajectory.

$$Threshold_{transitions} = Mean_{cluster2} + 0.7 \times (Mean_{cluster1} - Mean_{cluster2}) \tag{2.28}$$

$$LeftIndex = TransitionPoints > Threshold_{transitions} \tag{2.29}$$

$$RightIndex = TransitionPoints < -Threshold_{transitions} \tag{2.30}$$

The neighboring transitions that have opposite signs ($LeftIndex$ and $RightIndex$) are paired as the copy number alternation boundaries or potential gene breakpoints. The same process is applied to all three trajectories and the results are merged as the final potential CNV regions with positive peaks.

For the negative peaks (dips with read depth smaller than the average read depth), the simplified trajectories are constructed as follows:

$$Threshold = AvgRD - StdRD \tag{2.31}$$

$$Tj^{simplified} = \begin{cases} Tj^{original}, & \text{if } Tj^{original} < Threshold \\ AvgRD, & \text{otherwise} \end{cases} \tag{2.32}$$

The derivative of the simplified trajectories are again taken as the transitions points. The transition thresholds are set as the mean of the transition points. Only this time the $LeftIndex$ and $RightIndex$ are defined as follows:

$$LeftIndex = TransitionPoints < -Threshold_{transitions} \tag{2.33}$$

$$RightIndex = TransitionPoints > Threshold_{transitions} \tag{2.34}$$

As in the positive peaks, the same process is applied to all three trajectories and the estimated breakpoints of potential CNVs are merged as the final deletion candidates.

2.7.4 Grouping and Breakpoint Refinement

Since NN performs read-to-read alignments and the assignment threshold was set at five frame shifts (for $\geqq$ 100-bp reads), the valid assignments of each query read contain both library reads from repeated regions and the neighboring reads of each repeated region. The

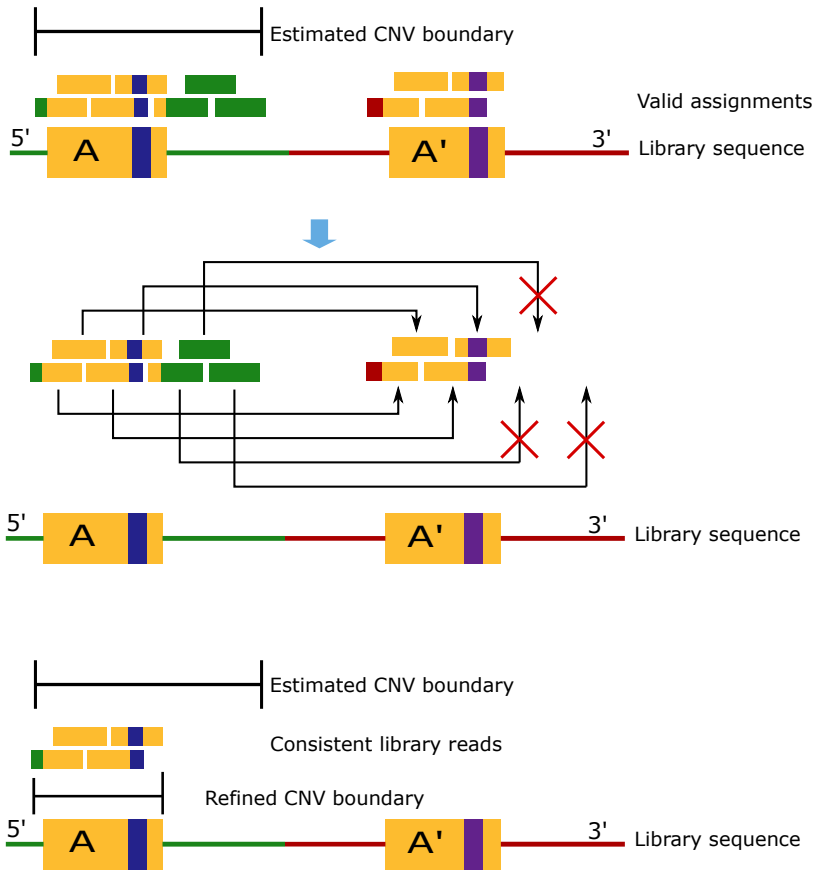

Figure 2.17: Breakpoint refinement. The estimated CNV boundary can be refined by only accepting library reads that have corresponding library reads in the repeated regions. Gene A and A' are two repeated regions in the library sequence. Green and red sequences are unique and different. Since some of the valid assignments from the estimated CNV boundary are not repeated in Gene A, they are excluded.

CNV detecting performance can thus be improved by incorporating the valid assignments information.

First, the valid assignments can be used to group related regions. Since all the repeated regions are presented at the same time as valid mapping results of a query read, these repeated regions are clearly related and the regions can be grouped together. Once the related regions are grouped, one can determined which region(s) within the group are deleted or duplicated in the reads donor sequence from the binary copy numbers.

Second, the breakpoints of the potential CNV regions can be refined using the information from the valid assignments. The estimated breakpoints of each gene come from the

segmentation process, which might not perfectly reflect the true gene boundary because of the noises in the read depth trajectory. The valid assignments contains reads from the repeated regions and their neighboring regions. Although the reads from the repeated region have corresponding reads in another region, the neighboring reads are unique to each region. Thus, by finding the corresponding reads in other regions for each valid assignment of a given region, one can re-define the breakpoints.

As shown in Figure 2.17, assuming gene A and gene A' are duplicated regions in the reference sequence and the sequence flanking gene A and A' at the 3' end are different (green and red sequences), The 3' portions are not part of the repeated regions. First, all valid assignments of the query reads that are assigned to this CNV regions are extracted. Since gene A and A' are similar repeated regions, even though the CNV boundary is for gene A, the corresponding library in gene A' are included. The library reads from the green sequence, however, only show up once since there is no similar sequence around gene A'. These inconsistent library reads are then excluded from valid assignments. The refined CNV boundary is then defined by the remaining, consistent valid assignments (Fig. 2.17. Thus, by identifying the repeated patterns of the library reads around the CNV regions, more precise boundaries can be drawn.

2.7.5 Copy Number Estimation and True CNVs Selection

After the groups and boundaries of the potential CNV regions are defined, the copy number of all three trajectories (average, binary and TS or even trajectory) for each candidate region is calculated by dividing the average read depth in the regions by $AvgRD$, the average read depth along the sequence. The average read depth is calculated by iterations from the original read depth trajectories, not the denoised trajectories. In each iteration, the average and standard deviation of the trajectory are obtained. A new average and standard deviation are then calculated by excluding all the data points that are $2 \times StdRD$ deviated (both larger or smaller) from the average. The iteration stops when the average is converged.

To determine the true CNV regions, the group average copy numbers are calculated. For duplication regions, any groups that satisfy $|CopyNumber_{Avg} - CopyNumber_{Avg} \times CopyNumber_{TS}| >= 0.9$ (which means that the expecting copies in the read donor sequence is different from the copies is the reference sequence) with $CopyNumber_{TS}$ deviated from one are (which means that the difference is real) considered to be true CNV regions. For deletions, the same criteria is applied on regions have multiple copies. For regions that have only one copy in the reference sequence, all the deletions are kept. Different from the other CNV detectors that do not work with mapping multiplicity or can not distinguish true CNV regions versus multiple copies in the reference sequence, CNV-MM greatly reduces the false positive rate.

In summary, CNV-MM uses both average and binary trajectories to identify the potential CNV regions. These regions are grouped and refined during breakpoint analysis. The copy numbers of all trajectories of these regions are calculated by dividing the read depth with the average read depth of each trajectory. The true CNV regions are gene regions that have an integer increment in the copy number of the reads donor sequence and the TS or evenly distributed trajectory has to deviate from one.

2.7.6 GC Content Correction for Real Reads

It has been shown that, instead of random sampling across the target sequence, reads are generated with GC-content bias in Illumina data due to the PCR process.[154–156] However, read depth based CNV detection is built on the assumption that reads are sampled evenly from the sequence. To remedy this, RDXplorer[129] and CNVnator[42] correct the GC-bias by normalizing the read depth with the deviation of the read depth at a given GC-content as follows.

$$RD^{i}_{corrected} = RD^{i}_{raw} \times \frac{RD_{overall}}{RD_{GC}} \tag{2.35}$$

in which $RD^{i}_{corrected}$ is the corrected read depth of bin i, RD^{i}_{raw} is the original read

depth of bin i, $RD_{overall}$ is the average/median read depth of the sequence, RD_{GC} is the average/median read depth of all bin with the same level of GC content as bin i.

Since GC content correction relies on calculating the GC content in each bin, the binning process will reduce the breakpoint resolution. To avoid the breakpoint resolution decrease, the GC-correction is performed after the segmentation. The trajectories are binned with the bin size equals the read length. The regions with estimated CNV segments are excluded to prevent read depth bias. The read depths and the GC contents of each bin are recorded. The median read depth of a given GC content is calculated as RD_{GC}. Since bins with GC content $<20\%$ and $>70\%$ is rare, we set the first GC-level as 0 to 20% of GC content and the last GC-level as 70% to 100% GC content. The rest of the bins are GC content ranges from 20% to 70% over increments of 2% GC content. The $RD_{overall}$ is the average read depth of all bins.

The average read depth for each estimated potential CNV region is calculated (RD^i_{raw}) and the corrected read depth ($RD^i_{corrected}$) is calculated following equation 2.35.

2.8 Availability of the Data

628 bacterial library sequences and 197 assembled bacterial genome sequences were obtained from NCBI (`ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/`). The accession numbers of the library sequences and the 197 bacterial sequences can be found in Table D.1 and Table D.2 respectively. The short reads files were downloaded from the sequence read archive (`http://www.ncbi.nlm.nih.gov/sra`). A full list of short reads files used in bacterial species typing is in Table Table D.3 The taxonomy used to assess the phylogenic tree was from the NCBI taxonomy database (`http://www.ncbi.nlm.nih.gov/taxonomy`).

CHAPTER 3

POST-BLOOD CULTURE ANTIBIOTIC SUSCEPTIBILITY TEST

3.1 Introduction

One of the bottlenecks in clinical microbiology is to determine effective antimicrobial treatments. For a patient suspected to have a bacterial infection, a blood sample is collected and incubates for $\sim$24 hours in blood culture. Once bacterial presence is confirmed, series of plates are streaked to isolate the infectious bacteria, taking to another 24 hours. Finally, microdilution or disk diffusion antibiotic susceptibility testing (AST) is performed, taking another 18 to 24 hours.

Flow cytometry can monitor the antibiotic-induced damages in bacterial cells much earlier than can growth-inhibition based methods. However, cytometric responses are hampered by biovariability and machine fluctuations. To compare cytometric data, appropriate multidimensional statistic tests are required. Existing statistical test either are not scalable to $\geqq$ one dimension, rely on data conforming to a specific underlying distribution, or scale disadvantageously with the number of bins needed. To address this need, I developed a new and rapid AST was developed by analyzing the cytometric data of the antibiotic-treated bacteria with PB-sQF. This chapter studies the bactericidal antibiotic-induced reactive oxygen species (ROS) and/or the scatter signals changes in both lab strain and multidrug resistant clinical isolates. A wide range of antibiotics was tested with lab strain *Escherichia coli* (*E. coli*). Other major pathogens were investigated including *Pseudomonas aeruginosa* (*P. aeruginosa*), *Acinetobacter nosocomialis* (*A. nosocomialis*), *Klebseilla pneumoniae* (*K. pneumoniae*) and *Staphylococcus aureus* (*S. aureus*). Multidrug-resistant clinical isolates were acquired from Emory University, and effective treatment can be selected by calculating the changes in the cytometric responses with PB-sQF.

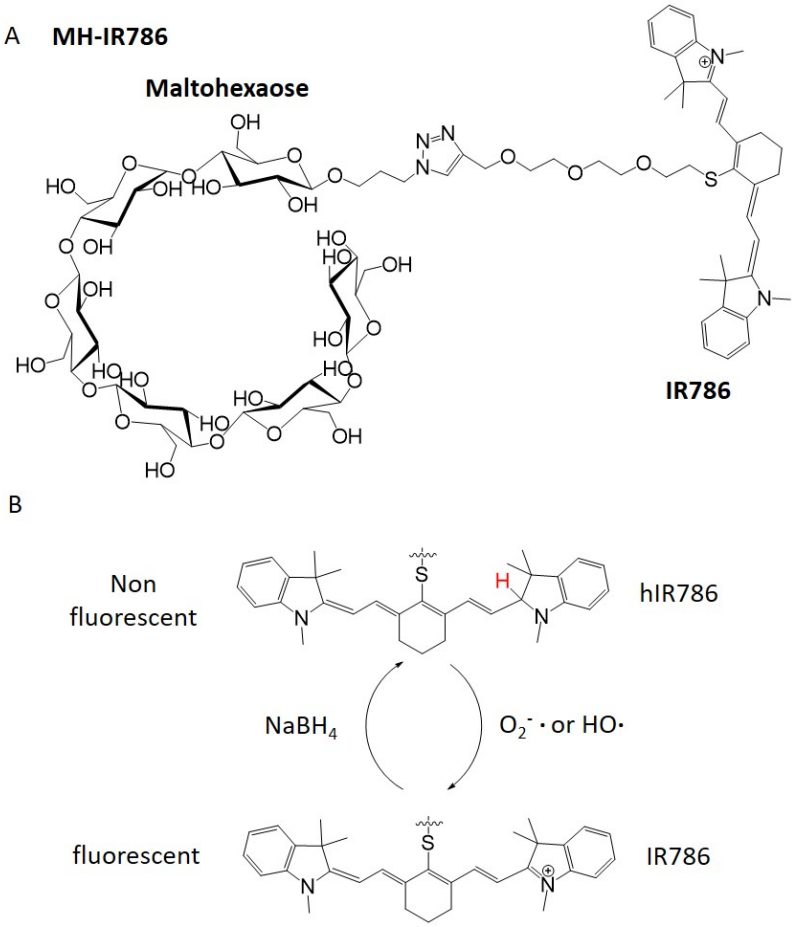

Figure 3.1: Maltohexaose-conjugated cyanine dye and hydrocyanine. (A) Maltohexaose-conjugated IR786 (MH-IR786). (B) Reversible reduction reaction. The fluorescent cyanine dye, IR786, is reduced to the non-fluorescent hydrocyanine (hIR786) by sodium borohydride. The fluorescence is recovered when the hIR786 is oxidized back to IR786 with ROS.

3.2 Sensing the Bactericidal Antibiotic-Induced ROS Generation

Recent studies reported that intracellular ROS generation is correlated with bactericidal antibiotics exposure.[17, 157, 158] Although it was later shown that ROS generation might not be crucial in cell death,[159–161] ROS production may still be a general mechanism by which antibiotic sensitivities can be more rapidly determined. Many ROS sensing dyes have been developed such as hydroxyphenyl fluorescein[162] and dihydroethidium[163]. However, they have been shown to suffer from autoxidation that produces non-zero background signal.

To specifically detect ROS generated inside the bacteria, maltohexaose-conjugated ROS sensing dye, provided by Dr. Murthy from UC Berkeley, was used. Maltohexaose (MH) belongs to the maltodextrin family and is a common source of glucose for bacteria. Utilizing the fact that bacterial uptake systems have broad substrate specificities, Ning, et al.[164] reported that bacteria can be selectively targeted through the maltodextrin uptake pathway. In contrast to other technologies, fluorescently labeled maltohexaose is selectively taken up by bacteria to reach mM intracellular concentrations, without detectable mammalian cell uptake, both in vitro and in vivo.[164] Such selective labeling offers a way to potentially identify bacterial populations prior to time-consuming subculturing and growth. Maltohexaose-conjugated dyes have been shown to stain bacteria exclusively since the maltodextrin metabolism pathway does not exist in mammalian cells.[164, 165] In this thesis work, Maltohexaose-conjugated IR786 (MH-IR786) was used (Figure 3.1 A).

To detect the antibiotic-induced ROS generation, MH-IR786 was first converted to Maltohexaose-conjugated hydro-IR786 (MH-hIR786) through a one-step reversible reduction reaction developed by Kundu et al.[166]. The reduction disrupts the conjugated system of the cyanine dye, making it nonfluorescent. The fluorescence is recovered upon reaction with ROS, which oxidizes the nonfluorescent dye to recover the original cyanine (Figure 3.1 B). Since maltohexaose dyes are actively brought into the bacterial cells, the combined changes in metabolic activity and ROS production may be indicative of near MIC antibiotic stress. Coupled with accurate statistical measures, general responses can be determined and behaviors quantified, possibly yielding a general mechanism by which antibiotic sensitivities can be rapidly determined.

3.2.1 MH-hIR786 Preparation and in vitro Fluorescence Recovery

MH-IR786 was prepared at a concentration of 1 mg/100 μL deionized water. Sodium borohydride (Sigma) was dissolved in methanol (VWR, Batavia, IL) at 1 mg/mL and subsequently added, 10 μL at a time, until the MH-IR786 changed from dark green to yellow.

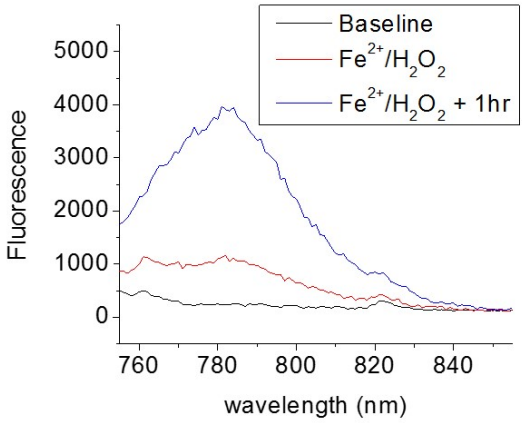

Figure 3.2: In vitro fluorescence recovery. MH-hIR786 was added to DMSO for fluorescence baseline detection (black line). Fluorescence recovery was monitored immediately (red line) and after an hour (blue line) after the addition of Fenton reagent.

The MH-hIR786 solution was then vacuum dried and resuspended in pH 6.0 acetate buffer (Fisher Scientific) at a final concentration of 1 mM. To ensure that the MH-hIR786 fluorescence can indeed be recovered by reacting with ROS, in vitro fluorescence recovery was tested by oxidizing hIR786 to IR786 with ROS generated from Fenton's reagent.[141, 142] First, the fluorescence baseline of 20 μL of MH-hIR786 in 2 mL of dimethyl sulfoxide (DMSO) (Fisher Scientific) was measured with a fluorimeter (QuataMaster, Photon Technology International). Then, Fenton's reaction was initiated by adding 60 μL of $FeSO_4$ (Mallinckrodt, St. Louis, MO) at 3.5 mg/mL and 300 μL of H_2O_2 (VWR, Batavia, IL) at 200 nM to the MH-hIR786/DMSO solution. The fluorescence signal was measured immediately and 1-hour after the reaction (Figure 3.2). Compared to the baseline, the fluorescence was clearly recovered upon ROS generation, and the fluorescence intensity was higher after 1-hour of reaction time.

3.2.2 Correlated ROS Production and Cell Death

To show that antibiotic-induced ROS generation is correlated with cell death, bacteria were cultured overnight in an incubator shaker (MaxQ 4000, Thermal Fisher Scientific, Waltham, MA) in Luria-Bertani (LB) medium (Sigma-Aldrich, St. Louis, MO) at 37 °C

and 225 rpm. Bacteria were then re-inoculated in 12 mL fresh LB medium in 50-mL tubes and incubated from $\sim$0.05 optical density to the mid-log phase. Bacteria in 1 mL of growth media were collected by centrifugation (Centrifuge 5417R, Eppendorf) at 13,400 rpm for 3 min and transferred into 12-well plates (Costar, New York, NY). Antibiotics and 20 μL of MH-hIR786 (provided by Dr. N. Murthy's lab) to achieve a final concentration 900 nM were added simultaneously. The MICs of different antibiotics were determined by standard microbroth dilution assays in advance. The 12-well plates were incubated at 37 °C for 1 hour (Isotemp standard incubator, Fisher Scientific, Waltham, MA). Bacteria were again collected by centrifugation and washed 3 times with phosphate-buffered saline (PBS) (Life Technologies, Carlsbad, CA) and resuspended in 1 mL PBS. The bacteria samples were maintained on ice until flow cytometry was performed. Bacteria samples were analyzed by a BD LSR II flow cytometer (Becton Dickinson, Franklin Lake, NY) equipped with a 14 mW, 488 nm solid-state coherent sapphire laser for the scatter signal, and a HeNe Laser (18 mW @ 633 nm) for IR786 fluorescence detection. Samples were gated by forward and side scatter, while a 750-810 nm bandpass filter was used to collect the IR786 fluorescence. Data were collected with FACSDiVa provided by BD. Further data analysis and display were carried out with Matlab 2013b (Math Works). For each data set, 100,000 bacterial detection events were collected.

When incubating lab strain *E. coli* strain ATCC33456 with 2x MIC (1x MIC is 100 μg/mL) of ampicillin, which is a bactericidal antibiotic, for an hour, the fluorescence signal recorded by flow cytometry was significantly higher (Figure 3.3, red curve) compared to the no-antibiotic control (Figure 3.3, grey curve). This shows that fluorescence signal indeed recovers upon ampicillin treatment. Since Fenton's reaction, which iron(II) reacts with hydrogen peroxide and superoxide, is responsible for the generation of ROS in the biological system,[167, 168] the addition of iron chelator, $2, 2'$-dipyridyl, should reduce the ROS generation as shown in the studies by Kohanski et al.[17, 157] Indeed, the fluorescence signal significantly decreased when the lab-strain *E. coli* was incubated with ampicillin,

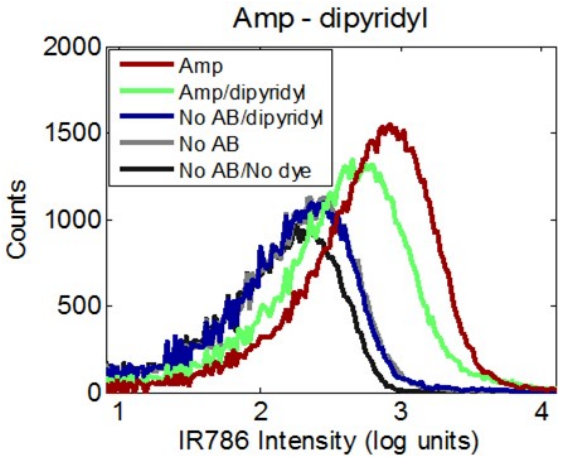

Figure 3.3: Fluorescence recovery with or without iron chelator. Flow cytometric data of lab-strain *E. coli* incubated without antibiotic, no dye (black); without antibiotic, with dye (grey, underneath the blue curve); without antibiotic, with dye, with dipyridyl (blue); with ampicillin with dye, with dipyridyl (green); with ampicillin, with dye (red).

MH-hIR786 and 500 μM of $2,2'$-dipyridyl (Figure 3.3, green curve) compared to the no chelating agent condition (Figure 3.3, red curve). On the other hand, the fluorescence signal remained very similar between the with and without $2,2'$-dipyridyl conditions when no antibiotic was added (Figure 3.3 blue curve and gray curve). This demonstrates that the fluorescence recovery of MH-hIR786 is indeed related to the antibiotic-induced ROS generation. Both the no-antibiotic, with MH-hIR786 conditions (with or without $2,2'$-dipyridyl, blue and grey curves in Fig. 3.3) have higher fluorescence signal than the no-antibiotic, no-dye control (Figure 3.3, black curve). This no-antibiotic ROS background is most likely because ROS are constantly generated from the aerobic respiration cycle.[157]

To confirm antibiotic-induced ROS production in different antibiotics, representatives of three major bactericidal antibiotics classes were tested: β-lactams (penicillin G, ampicillin, and cefotaxime), quinolones (ciprofloxacin and norfloxacin), aminoglycoside (kanamycin) and one bacteriostatic antibiotic (tetracycline). Intracellular ROS production upon 1-hour incubation with *E.coli* and each antibiotic around its MIC (Appendix Table B.1) were indicated by IR786 fluorescence recovered from ROS-reoxidized MH-hIR786, as monitored by flow cytometry (Figure 3.4 A to G). As expected, increased ROS production was

largely seen as MIC exposure is approached for bactericidal antibiotics, and no increase was perceptible for the bacteriostatic tetracycline. Similar to published data,[157] kanamycin only generated perceptible increases in ROS-induced fluorescence upon 2-hour incubation, but each antibiotic showed widely varying levels of antibiotic-induced ROS production, emphasizing the need for paired controls. Importantly, while the quinolone norfloxacin showed increased ROS production near its MIC, the more commonly used quinolone, ciprofloxacin exhibited no ROS-induced fluorescence increase. This observation may help reconcile observed ROS production with recent publications[159, 160] which showed that antibiotics retained their bactericidal ability under anaerobic conditions by interacting with their primary targets and that ROS generation was not necessary for antibiotic-induced cell death.

All experiments were performed in triplicate and compared to appropriate paired no-antibiotic controls. Test results were calculated between each antibiotic dataset and its paired-control, the antibiotic-free, MH-hIR786-labeled bacterial data. A 99% confidence level was determined by bootstrapping and was unique for each no-antibiotic control. Test results for each bacterium/antibiotic combination were normalized to its own 99% confidence distance, directly reporting on differences between antibiotic-treated samples vs. no-antibiotic paired controls. These individually normalized test results, or "fold distances" from each paired control, are then directly compared among all antibiotic exposure data, as described in Chapter 2. Fold distances from paired controls are plotted as averages of triplicate experiments with standard deviations representing biological variability and intra-bin data dispersion (Figure 3.4 H). Test results are considered statistically distinguishable beyond error bars, allowing identification of antibiotic-induced effects. Normalizing test results by the 99% confidence level of each paired-control, removes machine-to-machine and day-to-day variations, facilitating direct comparisons. The test results show that the flow cytometric data with 1x MIC of β-lactam antibiotics, kanamycin, and norfloxacin are statistically significantly different from their no-antibiotic paired controls (Figure 3.4 H).

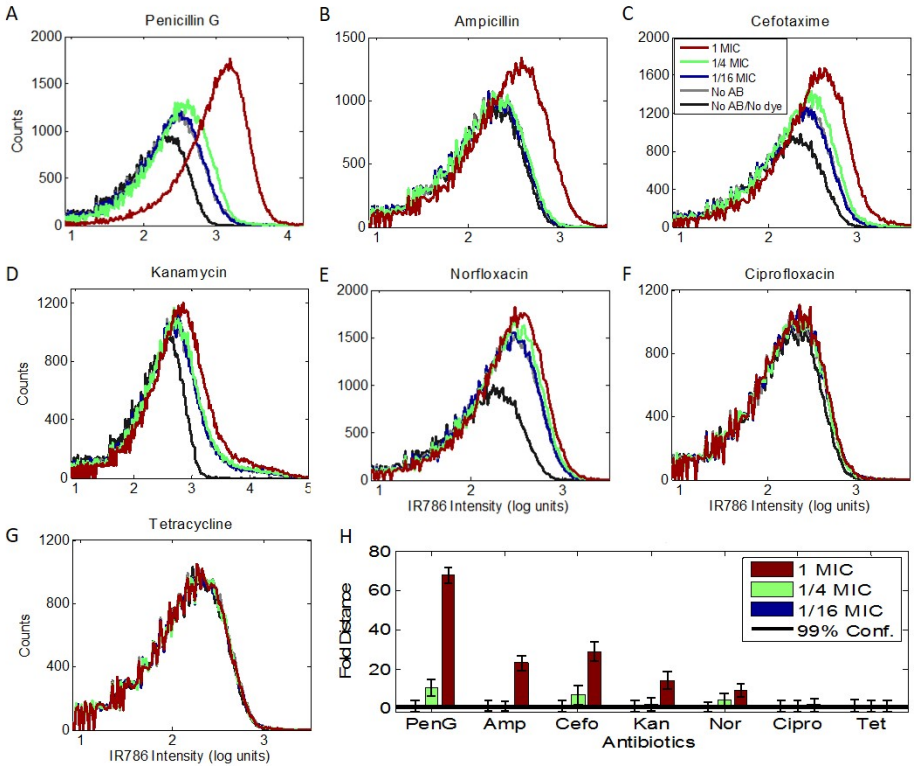

Figure 3.4: Antibiotic-induced ROS detection. Flow cytometric data for (A) penicillin G, (B) ampicillin, (C) cefotaxime, (D) kanamycin, (E) norfloxacin, (F) ciprofloxacin, and (G) tetracycline. (H) PB-sQF 1D test results, which coincide with the recovery of the fluorescence intensity from MH-IR786. All the antibiotics were incubated for 1hr with MH-hIR786 except that kanamycin was incubated for 2hr. Similar to the observation by Kohanski et al.,[157] bactericidal antibiotic-induced intracellular ROS generation. Among bactericidal antibiotics, Ciprofloxacin, which is a quinolone similar to norfloxacin, did not induce measurable ROS generation inside *E. coli*. The MIC can be found in Appendix Table B.1

3.2.3 Fluorescence Recovery by Antibiotic-Induced ROS Generation in Multidrug-Resistant *E. coli* clinical isolates

Although antibiotic-induced ROS-sensing may not generally work for all antibiotics, ROS generation appears correlated with the presence of lethal concentrations of some antibiotics, enabling sensitivity determinations for the subset of ROS-inducing antibiotics. As β-lactams stimulate the largest ROS response, ROS generation in the PenG-sensitive *E. coli* lab strain (ATCC 33456) and that in a highly multidrug resistant clinical isolate (Mu14S) were compared. ROS-induced fluorescence recovery was only observed in the PenG-

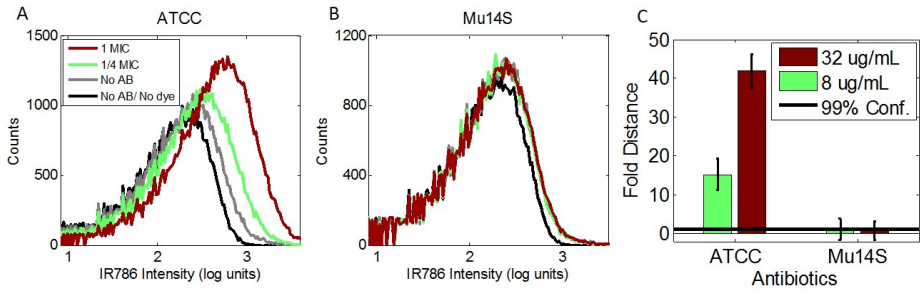

Figure 3.5: ROS-induced fluorescence recovery in resistant clinical isolate. Black curve: no PenG/no MH-hIR786. Grey curve: no PenG/MH-hIR786. Green curve: 1/4x MIC/MH-hIR786. Red curve: 1x MIC/MH-hIR786. The MIC, 32 μg/mL, was the concentration of the lab strain (ATCC) and was used for both strains. (A) Fluorescence recovered as expected in the lab strain (sensitive strain). (B) The fluorescence shows no significant recovery in the resistant strain at the sensitive strain's MIC. (C) PB-sQF quantification of data in A and B.

sensitive strain, indicating the ability to distinguish sensitive vs. resistant bacteria.

Antibiotic-induced ROS generation can be monitored by the fluorescence recovery from MH-hIR786, and appears to be a common, but not general response to even bactericidal stress. Even when present, the level of ROS generation varies widely with antibiotics, but even small changes from paired controls are quantifiable, showing distinct changes at or near the MIC. Unfortunately, while the norfloxacin MIC was obtainable with ROS sensing, the related bactericidal ciprofloxacin produced no discernable ROS generation. This shows that instead of ROS-mediated cell death, the traditional antibiotic targets, membrane synthesis, DNA replication and protein production, are the main cause of cell death.[159, 160]

In contrast to fluorescence viability and ROS-generation, label-free antibiotic-induced scatter signal changes appear to be universal throughout the antibiotics and bacteria we tested. Antibiotic-induced filamentation has been observed to from exposure to β-lactam,[66, 169] quinoline,[66, 170] and bacteriostatic[171] antibiotics. As a result, to build a general AST with flow cytometry, scatter changes were monitored instead.

3.3 Rapid AST Based on Cytometric 3D tests

Although fluorescence recovery from ROS re-oxidation of MH-hIR786 was a reliable guideline for AST with β-lactams, it is difficult to obtain full reduction of the MH-IR786 dyes to create MH-hIR86. Moreover, it did not work for bacteriostatic antibiotics and ciprofloxacin. On the other hand, the scatter signal changed significantly when *E. coli* were treated with either bactericidal or bacteriostatic antibiotics. Since forward scatter (FSC) reflects the size and morphology of cells and side scatter (SSC) represents the internal structure and granularity of cells,[172] it was expected that antibiotics-induced scatter changes were related to the morphology changes. MH-IR786 (instead of MH-hIR786) was added to label bacteria and gave an extra dimension to distinguish the resistant strain from the susceptible strains.

Antibiotic-induced scatter changes have been reported in different studies,[63–66] but like viability studies, remained difficult to quantify in complex flow cytometric datasets, clouding interpretations. By calculating the distance of the scatter patterns between the antibiotic-treated data and the no-antibiotic control, PB-sQF can reliably select the effective antibiotic treatment for the multi-drug resistant, clinical isolates within 4 hours post-blood culture.

3.3.1 Antibiotic-induced changes in susceptible *E. coli*

To assay dye uptake as a function of antibiotic exposure, a similar sample preparation procedure was used as in the ROS sensing test, except that instead of incubating with MH-hIR786, 900 nM of MH-IR786 was added. The samples were incubated with antibiotics at their respective 1x, 1/4x and 1/16x MIC that was first determined by standard micro-broth dilution assays. After 1-hr incubation, bacteria were pelleted, washed 3 times and resuspended in PBS for cytometric analyses. Three major bactericidal antibiotics classes, β-lactams (penicillin G and ampicillin), quinolones (ciprofloxacin and norfloxacin) and

aminoglycosides (kanamycin and gentamicin) as well as bacteriostatic antibiotics (tetracycline, erythromycin, and azithromycin) that target various biological processes were examined. Fluorescence and scatter signals upon antibiotic challenge were monitored by flow cytometry. IR786 fluorescence, forward-scattered and side-scattered light were all collected for each of 100,000 measured bacterial cells per run, yielding 3-D histograms for each antibiotic concentration.

Data of *E. coli* (ATCC 33456) treated for 1-hr with penicillin G, tetracycline, and kanamycin are shown in Figure 3.6 (Complete flow cytometry histograms with additional antibiotics can be found in Appendix Figs. B.1 and B.2). Log-scale units are chosen to represent the data. This is because dynamic range of the scatter signals ranges from 10 to 1000 and in the fluorescence channel ranges from 0 to 10^5. Linear scale simply can not properly display the data. The subsequent PB-sQF analyses, are also done in the log units since distances calculated in the linear scale will be dominated by the large data points. These large data points are orders of magnitude larger than the small data points thus any signal changes within data points with small value will become insignificant.

Upon penicillin G treatments at near MIC concentrations, both scatter and fluorescence signals significantly shift (Figs. 3.6 A and 3.6 D). Tetracycline, a bacteriostatic antibiotic targeting the 30S subunit of the bacterial ribosome, however, primarily altered only scatter signals relative to the no-antibiotic control (Fig. 3.6 B vs. 3.6 E). Conversely, kanamycin, another drug targeting the 30S ribosome, only induced very minor scatter changes, consistent with prior reports with aminoglycoside antibiotics,[63] but the fluorescence signal from MH-IR786 clearly increases upon 1x MIC exposure (Figs. 3.6 C and 3.6 F). Thus, multidimensional statistical metrics that combine both scatter and fluorescence are needed for generalizable, quantitative differentiation of population changes relative to paired controls.

Incorporating uncertainties arising from both biological variability and intrabin data dispersion into PB-sQF, all test results (Fig. 3.6 G) demonstrate statistically significant dis-

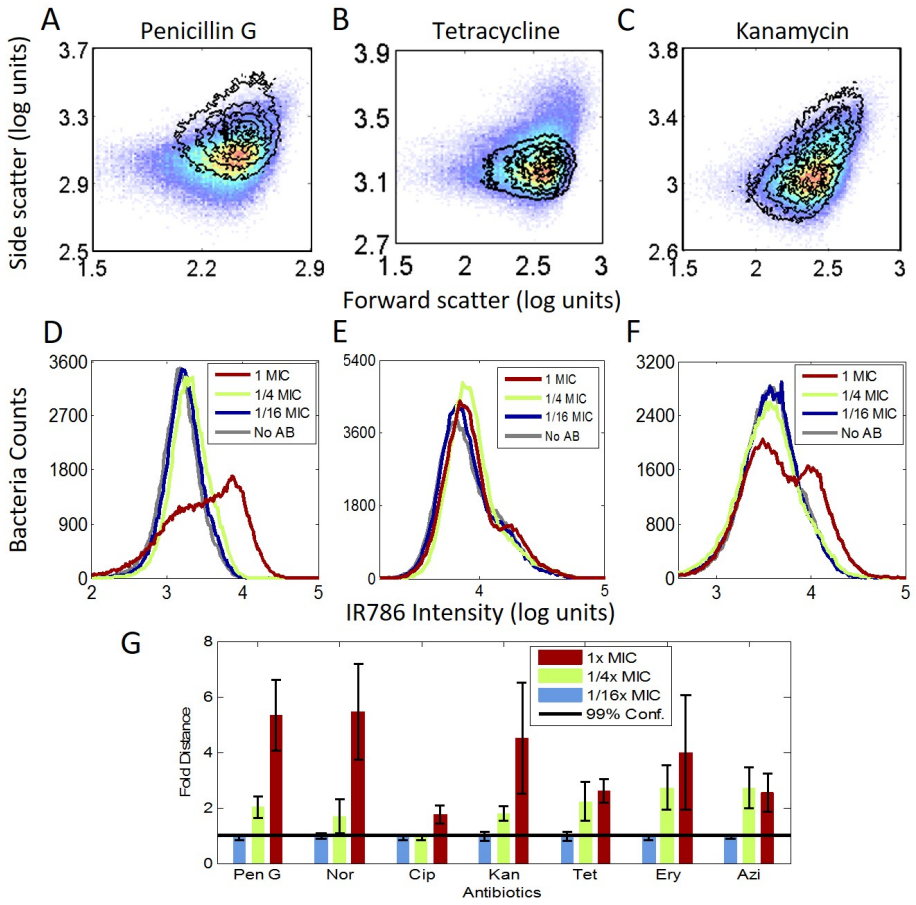

Figure 3.6: Antibiotic-induced signal changes. All data were collected in the presence of MH-IR786. (A to C) Scatter signal changes for different antibiotics. The pseudocolor plots are the no-antibiotic data. The overlay contour plots were data of the 1x MIC treatment. (A) Penicillin G (B) Tetracycline (C) Kanamycin. (D to F) Fluorescence signal changes from 1/16x MIC to 1x MIC and the no-antibiotic control. Grey curve: no antibiotic. Blue curve: 1/16x MIC. Green curve: 1/4x MIC. Red curve: 1x MIC. (D) Penicillin G (E) Tetracycline (F) Kanamycin. (G) The PB-sQF results of the 3D data. Black line: 99% confidence level from the test statistics between no-antibiotic control and 1/16x MIC data. All the data were normalized by the confidence level. Blue bar: 1/16x MIC. Green bar: 1/4x MIC. Red bar: 1x MIC.

tances of the 1x MIC data from that of the 1/16x MIC data, that is, beyond the 99% confidence level. Clear trends and transitions occur for all antibiotic/bacteria combinations with increasing antibiotic concentrations. The tested antibiotics target a wide range of processes (DNA replication, protein synthesis or cell wall synthesis), yet, when using our multidimensional statistical metric that reduces all differences to a single linear distance from its

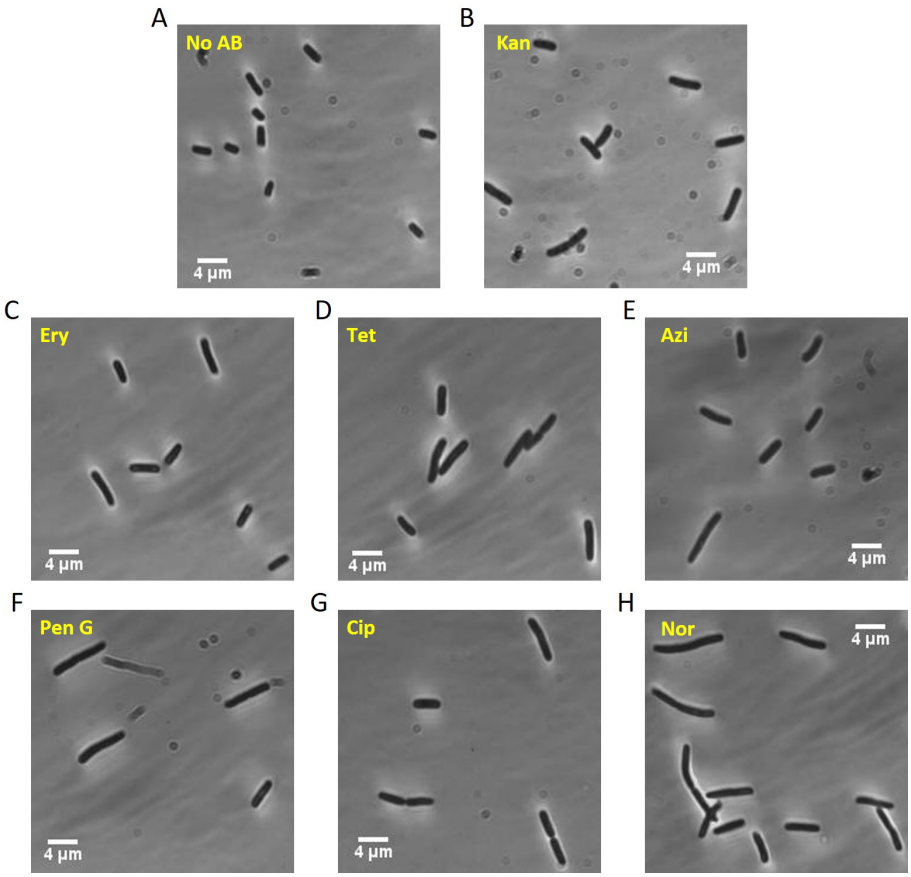

Figure 3.7: Morphology changes of bacteria treated with 1x MIC of different antibiotics. (A) non-antibiotic control. (B) Kanamycin. (C) Erythromycin. (D) Tetracycline. (E) Azithromycin. (F) Penicillin G. (G) Ciprofloxacin (H) Norfloxacin. In general, antibiotic-induced filamentation was observed compared to the non-antibiotic control.

paired control, all classes of antibiotics showed discernable, statistically significant changes in flow cytometry signals. Note that the 99% confidence level, which was determined by the bootstrap method of calculating the test statistics between the sub-sampling daughter distributions of the no-antibiotic control and the 1/16x MIC data at small sample size, accurately estimates the 99% confidence level distance between the two mother distributions. Thus, distances among all individually binned multidimensional histograms are reduced to single linear distances relative to paired controls using our PB-sQF distance metrics, enabling antibiotic sensitivities to be determined after only 1-hr exposures in comparison to overnight incubation in standard ASTs.

As maltohexaose conjugates are believed to be incorporated into bacteria via active up-

take processes,[164, 165] shifts in MH-targeted fluorescence signals likely indicate changes in bacterial physiological status. Forward and side scatter, however, provide label-free measurements that largely reflect cell size/morphology and internal cellular structure/granularity, respectively.[172] Indeed, images of bacteria at near-MIC antibiotic levels are often elongated relative to those without antibiotic present (Figure 3.7). Also, antibiotic-induced filamentation has been observed to result from exposure to β-lactam,[66, 169] quinoline,[66, 170] and bacteriostatic[171] antibiotics. Although it is not clear how these antibiotics, with different primary targets, uniformly induce changes in morphology and/or physiology, their changes in flow data from no-antibiotic controls appear to be generally correlated with antibiotic sensitivity levels

3.3.2 Cytometric susceptibility analysis of a resistant *E. coli* clinical isolate

To evaluate the antibiotic-induced changes in resistant strains, we examined a multi-drug resistant, clinically isolated *E. coli*, Mu14S. Mu14S was susceptible to gentamicin but was highly resistant to all other tested antibiotics. We examined both the laboratory strain (ATCC 33456) and Mu14S with penicillin G, tetracycline, and gentamicin in parallel. Consistent with the data shown in Figure 3.6, incubation with penicillin G at 1x MIC of ATCC 33456 strain clearly shifts its scatter distributions (Fig.3.8 A and Appendix Fig. B.3) while not affecting those of the resistant clinical strain Mu14S (Fig. 3.8 D and Appendix Fig. B.3). The 3D data (forward scatter, side scatter, and fluorescence) from both strains were quantified with PB-sQF (Fig. 3.8 G). The analysis confirmed that penicillin G was effective toward ATCC with both 1/4x and 1x MIC extending above the 1/16x no-antibiotic, 99% confidence level while all Mu14S results were below the 99% confidence level, indicating that PB-sQF registers no significant changes at these concentrations. On the other hand, while tetracycline-induced scatter changes in ATCC 33456 (Fig. 3.8 B), the scatter showed no shifts upon tetracycline exposure in Mu14S strain (Fig. 3.8 E and Appendix Fig. B.3). The PB-sQF results confirmed the ATCC 33456 sensitivity and Mu14S resistance toward

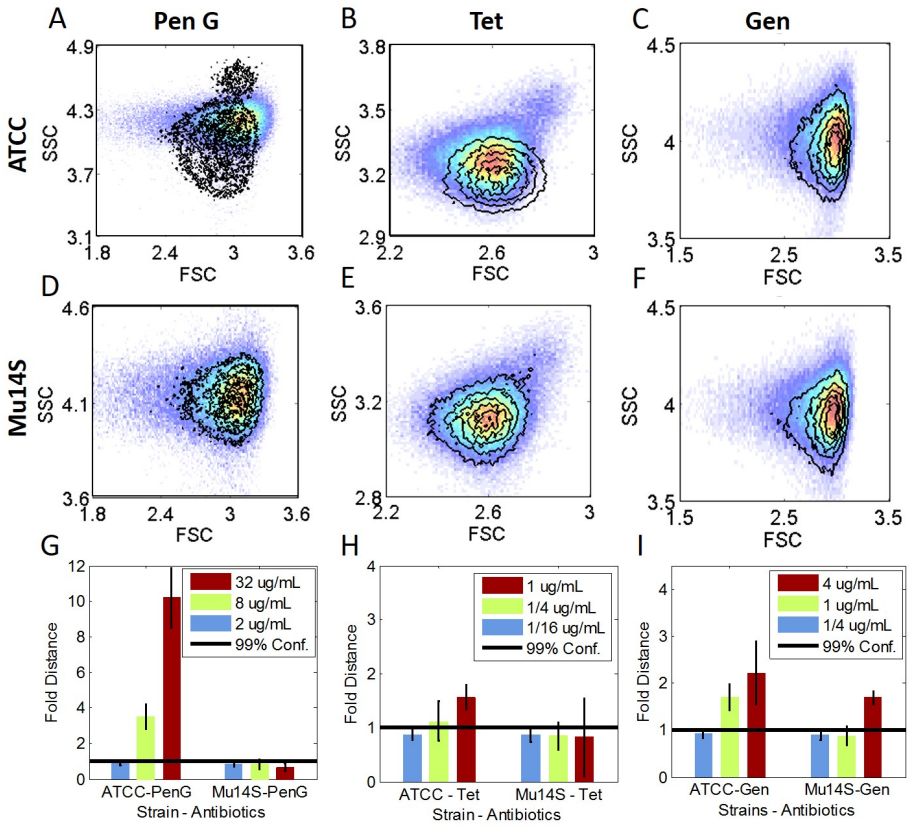

Figure 3.8: Signal changes induced by antibiotic treatments in *E. coli* with different susceptibility. All data were collected in the presence of MH-IR786. (A to F) Scatter signal changes. The pseudocolor plots are the no-antibiotic paired control, for each strain. The overlaid contour plots are the 1x MIC antibiotic concentration scatter data. (A to C) The lab strain *E. coli* (ATCC 33456). (D to F) The multi-drug clinical strain *E. coli* (Mu14S). (G to I) PB-sQF 3D test results. the first column (A, D and G) Penicillin G; Second Column (B, E and H) Tetracycline; Third column (C, F and I) Gentamicin. Penicillin G, and tetracycline was examined at the 1x, 1/4x and 1/16x of MIC of ATCC, (32 and 1 μg/mL, respectively). Gentamicin was applied at the MIC of Mu14S (4 μg/mL). FSC: forward scatter. SSC: side scatter.

tetracycline (Fig. 3.8 H).

The MICs for ATCC 33456 and Mu14S of gentamicin are 2 μg/mL and 4 μg/mL, respectively. Both strains were incubated with gentamicin at the MIC of Mu14S. Gentamicin-induced very little scatter shifts in either strain (Fig 3.8 C, F and Appendix Fig. B.3). However, our improved statistical metrics enable accurate quantification of these small differences (Fig. 3.8 I). Even with triplicate and centroid uncertainties, the test results registered significant changes from 1/4 μg/mL to 4 μg/mL for both strains, confirming the

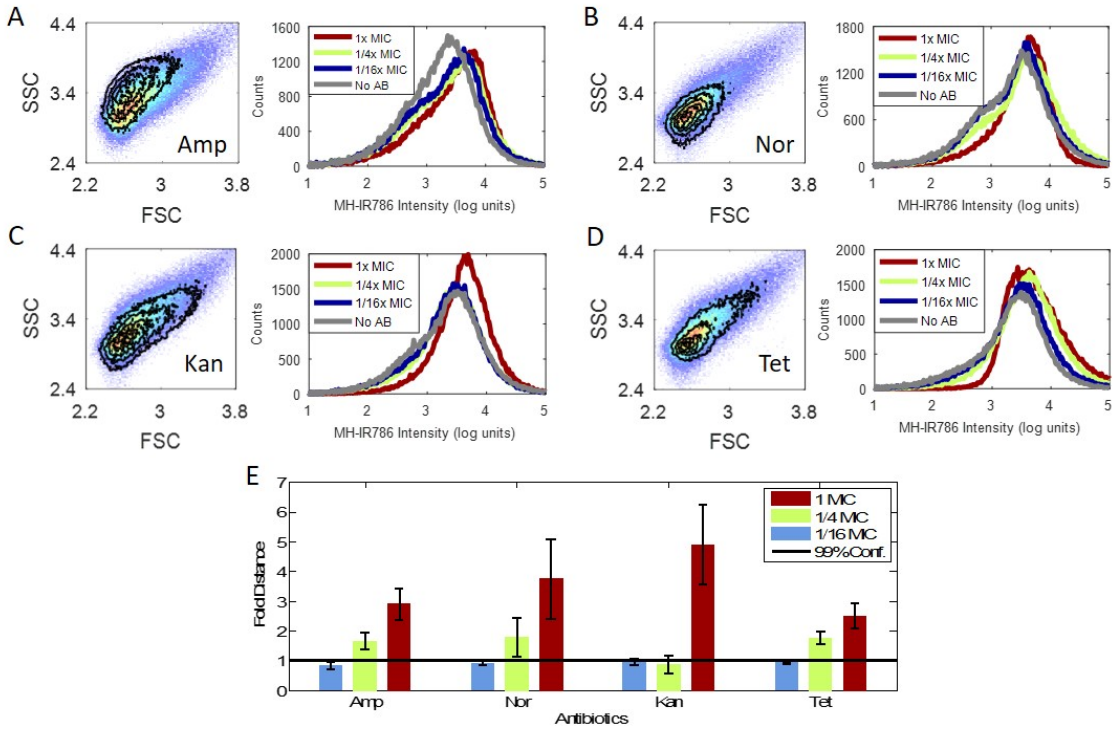

Figure 3.9: PB-sQF registered antibiotic-induced signal changes in *P. aeruginosa*. For each 2D scatter plot, pseudocolor plot is the no antibiotic control. The contour plots lay above is the 1x MIC scatter data. (A) Ampicilin (B) Norfloxacin (C) Kanamycin (D) Tetracycline. (E) The 3D PB-sQF test results for (A) to (D).

microbiological report, but with only 1-hr exposure.

3.3.3 Antibiotic-induced changes in susceptible *P. aeruginosa*

Previous studies have shown that *P. aeruginosa* strains are particularly difficult test cases with most antibiotics in which flow cytometry-based bacterial viability tests routinely fail.[70] For *P. aeruginosa*, this was explained by its outer membrane interaction with the dye propidium iodide (PI), yielding too high a background. Here, the same 3D PB-sQF was applied to *P. aeruginosa* treated with four different antibiotics. Using PB-sQF, *P. aeruginosa* exhibits readily distinguished sample-control distances analogous to those in *E. coli*, upon near MIC exposure to the same antibiotics (Fig. 3.9 and Appendix Fig. B.4). In the fluorescence data, it is clear that a single threshold is difficult to establish without any false positive or false negative counts since the control curve significantly overlaps the antibiotic-

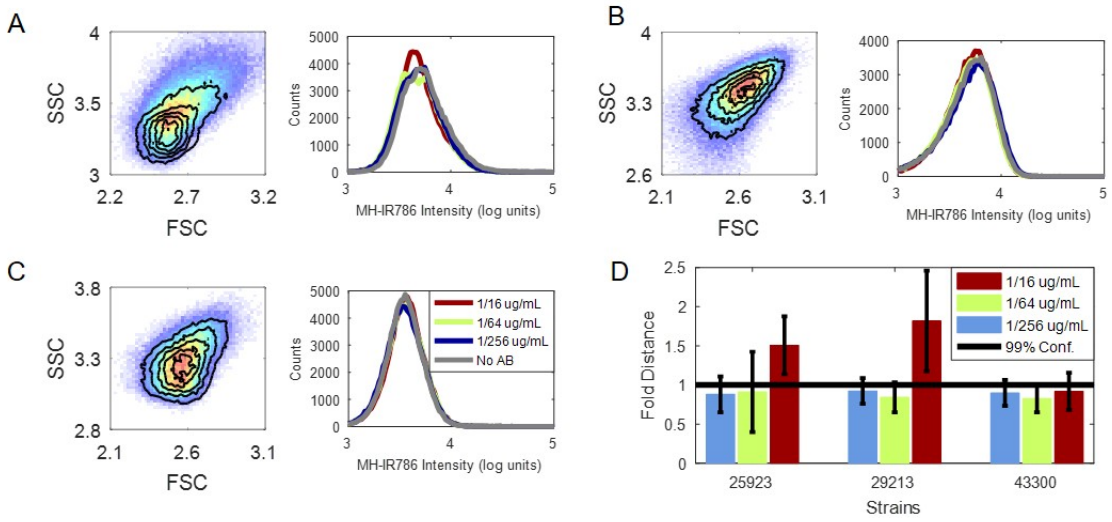

Figure 3.10: Penicillin G susceptibility for MRSA and MSSA strains Flow cytometric data of (A) ATCC 25923 (B) ATCC 29213 and (C) ATCC 43300 (MRSA). For 2D scatter histogram, the pseudocolor plots are the pair control, the no-antibiotic data, for each strain. The contour plots lay above are the highest antibiotic concentration scatter data. The fluorescence histograms share the same label as in (C). (D) 3D PB-sQF results for (A) to (C). The highest penicillin g concentration is 1/16 μg/mL, the MIC for strain ATCC 25923.

treated distributions. By directly comparing the whole data set, PB-sQF removes the need for artificial thresholds, enabling quantitative comparisons.

3.3.4 Cytometric susceptibility analysis of MRSA and MSSA

Methicillin-resistant *Staphylococcus aureus* (MRSA), which resists most clinically available β-lactam antibiotics, is a worldwide problem.[173] In 1992, MRSA comprised 35.9% of the total *Staphylococcus aureus* infections, but had increased to 64.4% by 2003,[174] with more recent reports suggesting that the rate has decreased and stabilized.[8, 175] To distinguish MRSA versus Methicillin-susceptible *Staphylococcus aureus* (MSSA), PB-sQF was applied to the cytometric data of one MRSA strain (ATCC 43300) and two MSSA strains (ATCC 25923 or ATCC 29213). For *S. aureus*, the bacteria were incubated in cation-adjusted Mueller-Hinton broth (CAMHB) with MH-IR786 and antibiotic (MIC at Table B.2) for 1 hr at 37 °C.

All three strains were incubated with penicillin g at 0.0625 μg/mL, the MIC of the

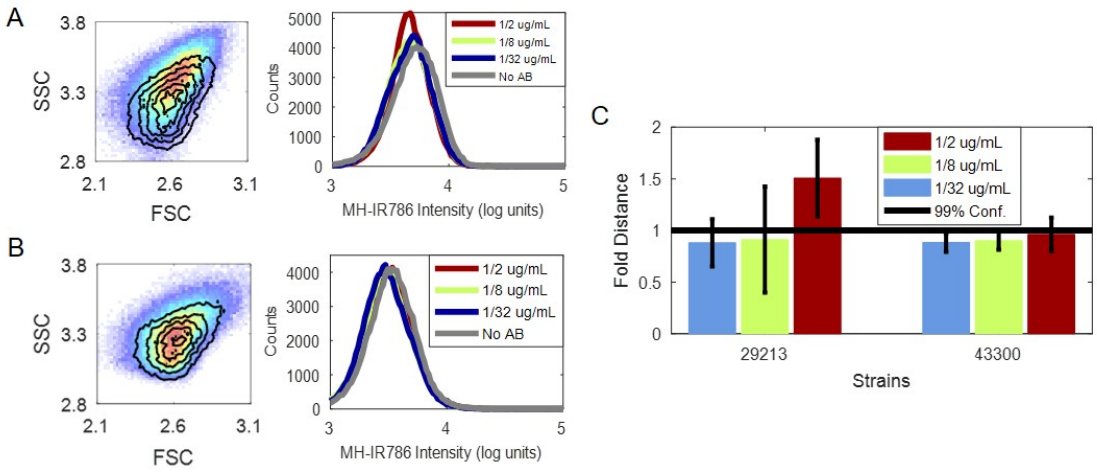

Figure 3.11: PB-sQF and select MRSA strain from MSSA strains. Flow cytometric data of (A) ATCC 29213 (B) ATCC 43300 (MRSA). For the 2D scatter histogram, the pseudocolor plots are the paired control, and the no-antibiotic data for each strain. The overlaid contour plots are the highest antibiotic concentration scatter data. (C) 3D PB-sQF results for (A) and (B). The highest oxacillin concentration is 1/2 μg/mL, the MIC for strain ATCC 29213.

penicillin-sensitive control strain ATCC 25923, for an hour (Figure 3.10). The ATCC 25923 exhibits clear scattered-light and fluorescence shifts (Figure 3.10 A) and thus shows statistically significant shifts at its own MIC (Figure 3.10 D). The cytometric signals of the MRSA strain (ATCC 43300) did not change significantly at this level of penicillin g since the MIC of ATCC 43300 is 16 μg/mL (Figure 3.10 C and D). The weak β-lactamase producing control strain ATCC 29213, however, shows significant signals shift at 0.0625 μg/mL while the MIC is 2 μg/mL (Figure 3.10 B and D). This is most likely because although penicillin g at 0.0625 μg/mL will not completely inhibit the growth, it already has a significant effect on the bacteria.

To gauge oxacillin resistance, ATCC strains 29213 and 43300 were treated with oxacillin at 0.5 μg/mL, the MIC of strain ATCC 29213. As expected, cytometric signals of ATCC 29213 change significantly at 1x MIC (Figure 3.11 A and C). On the other hand, oxacillin at 0.5 μg/mL does not significantly affect the flow data of the MRSA strain (43300), which has a MIC of 32 μg/mL (Figure 3.11 B). Correspondingly, no statistical distance was registered between the 1/16x MIC data and either the 1/4x or 1x MIC data (Figure 3.11 C),

thereby enabling discrimination of MSSA and MRSA within 2-hr processing times. Interestingly, label-free scattered light shifts relative to control are sufficient for distinguishing MRSA vs. MSSA, as oxacillin produced no consistent fluorescence shifts in either the weak β-lactamase producing control (29213) or MRSA. Thus, even subtle changes in the flow data are readily differentiated from paired controls after only 1-hr incubation time using our PB-sQF statistical distances, drastically reducing the required time for an AST-based MSSA/MRSA discrimination. The complete flow cytometric data can be found in Appendix Figure B.5.

3.4 Rapid AST Based on Cytometric Scatter Signals Changes

Although PB-sQF 3D test can select the effective treatments for *E. coli*, *P. aeruginosa*, and *S. aureus*, a label-free AST is more general and economical because a label-free approach does not need to consider the interactions between dyes and bacteria. Taking the maltohexaose-conjugated dye for example, although the maltohexaose transporter has been identified in a wide range of bacteria, not all bacteria produce it.[176] Also, the reproducibility of fluorescent signals is an issue. As shown in Appendix Figure B.6, the samples were prepared at the same time and the data were taken on the same machine, the fluorescent signal at 1/4x MIC, however, fluctuated.

As a result, here, PB-sQF was applied on the 2D scatter cytometric data of *K. pneumoniae* and *A. nosocomialis*. The PB-sQF 2D test results were related to the antibiotic concentrations and the MIC of each bacteria-antibiotic combination.

3.4.1 Cytometric susceptibility analysis of lab-strain *Klebsiella pneumoniae*

The same procedure was applied to *K. pneumoniae* strain ATCC 700603 with different antibiotics including bacteriostatic (azithromycin, erythromycin, and tetracycline) and bactericidal antibiotics: quinolone (ciprofloxacin), aminoglycoside (gentamicin) and β-lactams (ampicillin and cefotaxime) at each pre-determined MIC shown in Table B.3. Antibiotic-

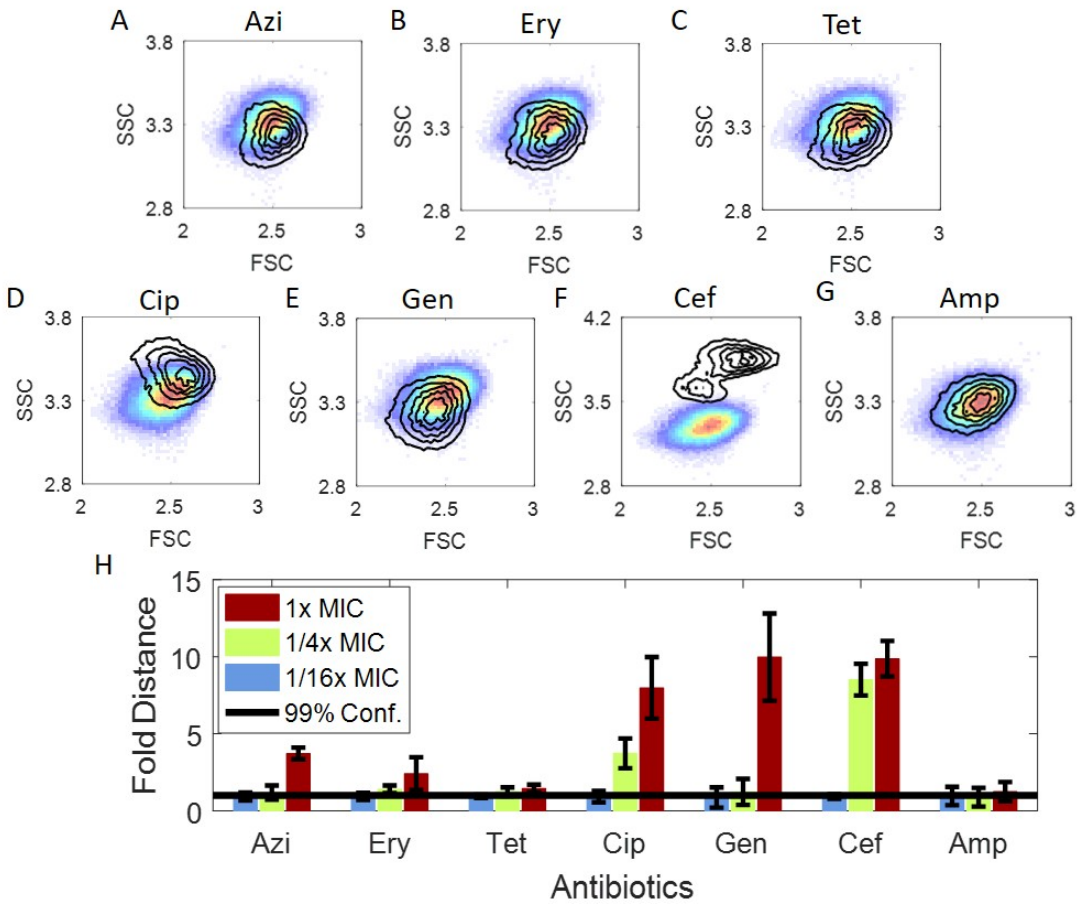

Figure 3.12: Antibiotic-induced scatter changes for *K. pneumoniae* ATCC 700603. (A to G) Scatter signal changes for different antibiotics. SSC: side scatter. FSC: forward scatter. The pseudocolor plots are the no-antibiotic data. The overlay contour plots were data of the 1x MIC treatment. (A) Azithromycin (B) Erythromycin (C) Tetracycline (D) Ciprofloxacin (E) Gentamicin (F) Cefotaxime (G) Ampicillin. (H) The PB-sQF results of the 2D data. Black line: 99% confidence level from the test statistics between no-antibiotic control and 1/16x MIC data. All the data were normalized by the confidence level. Blue bar: 1/16x MIC. Green bar: 1/4x MIC. Red bar: 1x MIC. The MIC of each concentration can be found in Appendix Table B.3. For ampicillin, 1x MIC was set at 80 μg/mL since the MIC is greater than 1024 μg/mL.

induced scatter shifts can be observed when *K. pneumoniae* were treated with effective antibiotic at 1x MIC (Figure 3.12). *K. pneumoniae* strain ATCC 700603 is highly resistant to ampicillin with the MIC larger than 1024 μg/mL. As a result, instead of incubating *K. pneumoniae* at its MIC, we treated the bacteria at 80 μg/mL of ampicillin, which is 10x of the sensitive breakpoint (8 μg/mL) determined by the Clinical & Laboratory Standards In-

stitute (CLSI).[177] As shown in Figure 3.12, the scatter patterns appear very similar with or without ampicillin treatment even at 80 μg/mL, which is a concentration beyond the resistant breakpoint (32 μg/mL). This demonstrated that *K. pneumoniae* is indeed resistant to ampicillin. The complete cytometric data can be found in Appendix Figure B.7 and Figure B.8.

PB-sQF was applied on all the triplicate, 2D scatter histogram to calculate the distance between the paired, no-antibiotic control and the antibiotic-treated data. All test results for tested antibiotics at 1x MIC show significant differences from the 99% confidence level other than ampicillin. Since ATCC 700603 is resistant to ampicillin, no scatter shift was observed in the cytometric data, and thus no statistically significant change was observed in fold distance. This shows that with PB-sQF, antibiotic resistance can be identified with only 1-hour incubation time.

3.4.2 *A. nosocomialis* Clinical Isolates Post-Blood culture AST

A. nosocomialis strain M2 was isolated in 1996 from Ohio[178] and the mutant strain M2-4B[179] has higher resistances to several different antibiotics (Gift from Dr. Philip Rather, Emory University). Both strains were incubated at their own MIC (Table B.3) determined from microdilution AST for each antibiotic for an hour, and then data were acquired with flow cytometry. The 2D scatter cytometric data, and PB-sQF results of strain M2 and M2-4B are shown in Figure 3.13 and Appendix Figure B.9 and B.10. For tetracycline and cefotaxime, the MICs are both 1 μg/mL for M2 and M2-4B. As a result, the scatter responses were very similar between the two strains with 1x MIC corresponding to $\sim$2-fold distance changes (Figure 3.13 A). For norfloxacin and ciprofloxacin, the MICs for M2-4B are 8-fold higher than are those of M2 with M2-4B being considered resistant to norfloxacin, while showing intermediate resistance to ciprofloxacin (according to the 2014 CLSI handbook).[177] The higher MIC (and thus resistant to antibiotics) for M2-4B is not only reflected in higher antibiotic concentrations being required to induced scatter signals

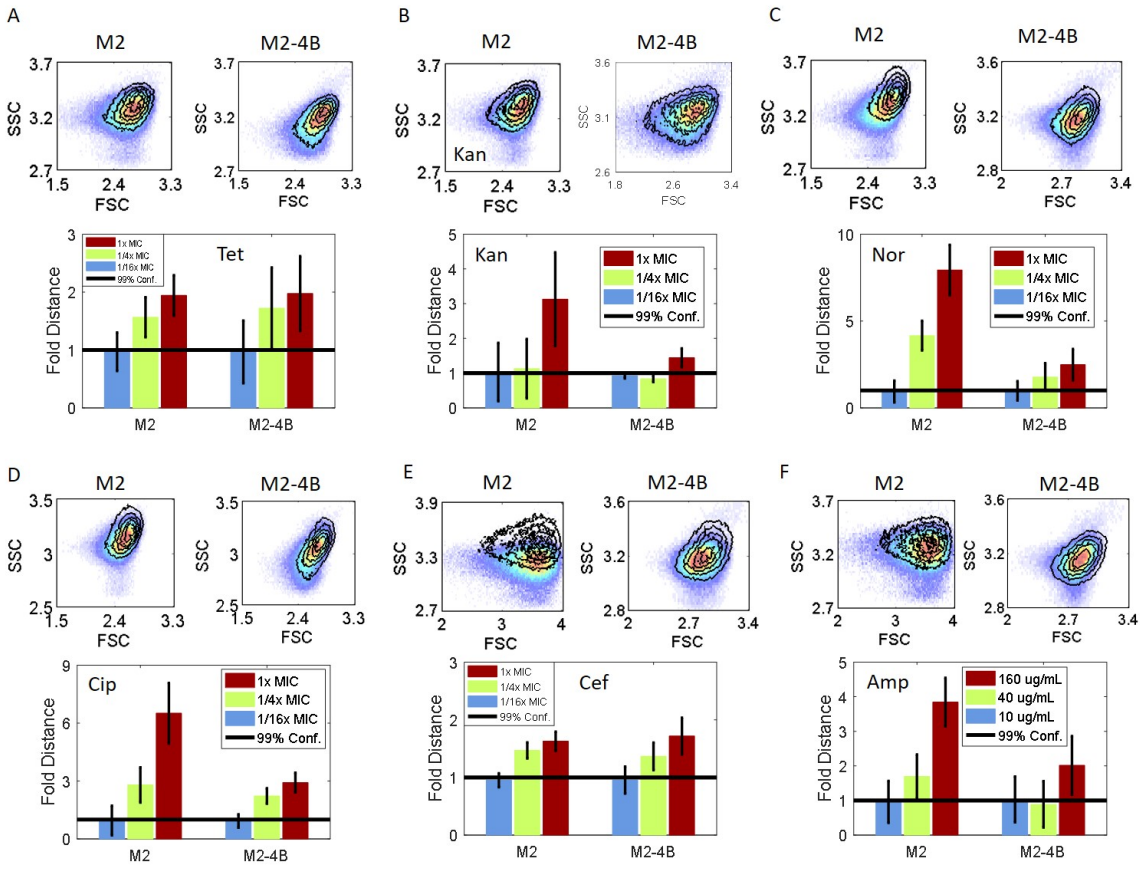

Figure 3.13: Cytometric data and PB-sQF results for *A. nosocomialis* strain M2 and M2-4B (A) Tetracycline (B) Kanamycin (C) Norfloxacin (D) Ciprofloxacin (E) Cefotaxime (F) Ampicillin. Each sub-figure contains 2D-scatter cytometric plots and the corresponding PB-sQF results. For the cytometric data, SSC: side scatter. FSC: forward scatter. The pseudocolor plots are the no-antibiotic data. The overlay contour plots were data of the 1x MIC treatment. For the PB-sQF results, Black line: 99% confidence level from the test statistics between no-antibiotic control and 1/16x MIC data. All the data were normalized by the confidence level. Blue bar: 1/16x MIC. Green bar: 1/4x MIC. Red bar: 1x MIC. The MIC of each concentration can be found in Appendix Table B.3. For ampicillin, 1x MIC was set at 160 μg/mL since the MIC is greater than 1024 μg/mL.

changes, but also smaller signal changes being observed in the M2-4B ($\sim$3 fold) compared to the M2 strain ($>$6 fold, Fig. 3.13 C and D). This shows that PB-sQF fold distance indeed correlates with the resistance profile of a bacteria strain.

Both M2 and M2-4B are highly resistant to ampicillin with MICs greater than 1024 μg/mL. However, when treated with 160 μg/mL, ten times higher than the sensitive break-point (16 μg/mL) according to the CLSI, both strains show ampicillin-induced scatter shifts

(Figure 3.13 F). This shows that although the MICs are higher than 1024 μg/mL for both M2 and M2-4B and thus both strains can survive and grow, ampicillin at 160 μg/mL still damages the bacteria cells. Nevertheless, M2-4B is still considered resistant to ampicillin since the antibiotic-induced scatter shifts was only observed at 160 μg/mL but not at 40 μg/mL. This means that the MIC of M2-4B is $\sim$160 μg/mL and is beyond the resistant breakpoint, 128 μg/mL, set by CLSI. On the other hand, with the just above threshold scatter shifts at 40 μg/mL, M2 is considered intermediate resistant to ampicillin since the MIC is larger than 16 μg/mL but lower than 128 μg/mL.

3.5 Conclusions

PB-sQF compresses the data by identifying the data signatures through adaptive binning. The adaptive binning provides advantages in calculating distance from multidimension of datasets, as calculations scale with the $(\#ofbins)^2$. PB-sQF then compares these signatures by reducing multidimensional differences to a linear distance between datasets. With rigorous uncertainties incorporated, subtle, but biologically relevant, antibiotic-induced changes become directly quantifiable, even as these antibiotics target widely varying biological pathways including DNA replication, protein synthesis, or cell wall synthesis.[169] Thus, scattered light and bacteria-targeted fluorescence, coupled with new statistical metrics appear to more generally enable rapid flow cytometry signal changes to gauge antibiotic resistance over a wide range of bacteria/antibiotic combinations. Even very difficult to discern combinations that fail with viability tests are readily distinguished with our statistical measures that are scalable to multiparameter measurements.

In the post-blood culture AST, since the bacterial inoculation concentration was high (Optical Density 0.5 $\sim 2 \times 10^8$ CFU/mL), the antibiotic susceptibility was assessed by the changes in morphology or physiological state of bacteria when exposed to antibiotics rather than monitoring bacterial growth as in the traditional AST. Thus, although for *E. coli*, *P. aeruginosa* and *K. pneumoniae* we can successfully determine ampicillin as an inappro-

priate treatment with no changes in the scatter signals, the weak β-lactamase producing control *S. aureus* strain ATCC 29213 shows a shift at a penicillin concentration below its MIC. Also, both *A. nosocomialis* strain M2 and M2-4B show statistically significant shifts at high ampicillin concentration (160 μg/mL) while the MICs for both strains are greater than 1024 μg/mL. For strain M2-4B, our approach still categorized it as a resistant strain since ampicillin-induced shifts happened only at a concentration beyond resistant breakpoint for penicillin-type antibiotics, 128 μg/mL. Strain M2, on the other hand, was a minor error as it was categorized as having intermediate resistance.

Our results indicate that antibiotic-induced damages occur before total growth inhibition thus scattered-light shifts are observed even at a antibiotic concentration lower than the MIC. The future direction for post-blood culture cytometric-based AST should focus on developing growth inhibition-based cytometry tests with low initial bacteria concentration. Since the starting bacteria concentration is low, the cytometric data for bacteria that are incubated with effective treatment at 1x MIC would be very closed to the background signals. The cytometric data for actively growing bacteria, on the other hand, will be dominated by bacterial signals. The sharp signal change (from zero to one) will help to resolve the problematic data.

Instead of growth-inhibition, another possible future direction is taking the PB-sQF fold distance into account. As shown in Figure 3.13, when the MIC for the M2-4B strain of a given antibiotic is more than 2-fold higher than the MIC for the M2 strain, the fold distance is smaller for M2-4B compared to M2. The same conditions might be applied to an ampicillin-susceptible strain versus the M2 strain: the fold distance might be much larger than $\sim$4 (the fold distance of M2 treated with 10x sensitive breakpoint) for a susceptible strain. Indeed, as a β-lactam antibiotic at a concentration as high as 160 μg/mL, the scatter patterns changes were not as significant as observed in the previous study.[48] The sizes of different bacteria, however, might complicate the analysis since the scattered-light signals are related to the size and granularity of a cell. Different size and species of quality control

strains might be requried to be done along with the tested strains.

With the potential to improve patient outcomes through shortening the window during which empiric antibiotic treatment is the only resource, flow-based antibiotic sensitivity determination suggests a $>$10-fold reduction in post blood culture time to result ($\sim$42 hrs to $\sim$4 hrs). As antibiotic susceptibilities and resistance proliferation are of great concern, these results suggest a path toward more effective and timely treatment. Since the time to treatment is a major determinant of positive patient outcome, the strong correlation of sample-control distance in combined scatter and fluorescence shifts with antibiotic sensitivity strongly suggests that general criteria can be established for developing robust flow cytometric based, rapid ASTs.

CHAPTER 4

PRE-BLOOD CULTURE AST

4.1 Introduction

Sepsis, a life-threatening immune response to blood infections (bacteremia), is the 10^{th} leading cause of hospital deaths in the US with a $\sim$30% mortality rate.[13, 180] Appropriate antibiotic treatment for bacteremia patients not only shortens hospitalizations and reduces antibiotic resistance proliferation, but it also lowers the incidence of septic shock and halves the fatality rate.[15, 17, 181–183] As sepsis can be caused by any of a number of bacteria, effective treatment relies on the combination of bacterial identification and sensitivity profile determinations. Unfortunately, adult sepsis patients often present $\leqq$100 bacterial cells (colony forming units, CFU) per mL blood, and pediatric patients exhibit $\sim$1000 CFU/mL.[184, 185] Because blood consists of $> 10^9$ cells/mL, bacteria populations must be amplified through $\sim$24-hr blood cultures to generate sufficient bacterial CFUs for diagnosis and further analyses. While pathogen identification has been hastened to just a few hours post positive blood culture,[21, 22, 186–188] antibiotic sensitivity tests (ASTs) still require an additional $\sim$36-44 hrs, after positive blood culture. In Chapter 3, a rapid flow cytometry-based AST was developed by calculating rigorous multidimensional statistical metrics[48] that matches the timescale of emerging post blood culture identification ($\sim$4hrs).[21, 22, 186–188] Unlike other rapid post-blood culture ASTs,[63–65, 67, 68, 71, 189] our phenotypic approach can drastically accelerate ASTs for sepsis-causing pathogens by removing the need for long blood culture-based amplification. Effective antibiotic treatments for all blood-stable pathogens investigated (multidrug-resistant *E. coli*, *K. pneumoniae*, and *A. nosocomialis* clinical isolates) were readily determined within what would correspond to 8 hours from initial blood draw. Results suggest that only 0.5mL of

adult blood or 0.05mL of pediatric blood is necessary per antibiotic. These methods should be readily adaptable to drastically improve patient outcomes by significantly reducing time to generate actionable treatment information the combination of pathogen identification and susceptibility profile.

In order to hasten treatment decisions that further improve patient outcomes and lower the incidence of antibiotic resistance, researchers have tried developing ASTs directly from raw blood to circumvent the initial $\sim$24 hr blood culture delay. Even though the $\sim 10^9$ mammalian blood cells/mL overwhelm any low-level bacteria signals, bacterial presence determinations within blood samples have been reported by flow cytometry,[190] microfluidics,[191–194] and PCR.[195, 196] While most of these detection schemes used genetic information from dead bacteria to detect presence, Hou et al. detected messenger RNA levels after pathogens were separated from blood in a microfluidic device.[194] Like other molecular diagnosis approaches, however, they can only target known RNA signatures for each strain and genetic indications of antibiotic resistance for each bacterium-antibiotic pair. A phenotype-detecting flow cytometry-based AST specific for *Y. pestis* was proposed that relies on post-growth recovery of bacteria from a gel matrix and viability dye detection.[197] Generalization, however, is problematic as careful bacterial recovery, significant post collection growth to reach $\sim 10^6$ CFU/mL, and user-dependent data gating are all needed to overcome the high scatter and fluorescence background. Additionally, viability dyes are known to produce false signals with various important bacteria/antibiotic pairs,[68, 70, 76] and gating is highly subject to variations in day-to-day instrument fluctuations, alignment, and parameters, limiting application of this approach.

Without blood culture-based growth, the highly disadvantageous bacteria:mammalian cell ratio, even in patients with bacteremia, demands that nearly all mammalian cell background be removed, without destroying the bacteria. Additionally, sufficient bacterial CFUs must be recovered to allow incubation with multiple antibiotics at various concentrations, suggesting that at least some amplification, or a higher volume of blood (at 100

CFU/mL), is needed. Because time is critical in ensuring appropriate treatment for patient survival[14] and reducing antibiotic resistance proliferation,[198] we avoid the need for lengthy blood culture by utilizing saponin to complex with cholesterols and induce hemolysis,[199, 200] without affecting bacterial growth or morphology.[201] This selective blood cell lysis enables even very small numbers of bacteria to be directly collected from the blood and enriched in blood-free growth medium for cytometric detection. Our robust statistics then enable quantification of very few bacterial counts, such that much shorter growth and antibiotic sensitivity times can be achieved.

4.2 Pre-Blood Culture AST Condition Search

To find the experimental conditions to remove blood cells, the detection limit of the flow cytometer was tested. Then, various conditions to separate the bacteria from blood samples, including serum separator tubes (SST) and saponin, were examined.

4.2.1 Varied Incubation Time with *E. coli* Only Samples

When taking flow cytometry data, background signals that come either from electronic noises or small particles in the solution always compete with signals of interest. Background noise obscures the events of interest when the signal is weak, as in the case of patients with sepsis. Since blood cells add more noise to the system, it is important to understand the detection limit of flow cytometric bacteria signals even before blood cells are added.

Samples with or without *E. coli* strain ATCC 33456 were incubated with different concentrations of penicillin g for either 1 hour or 3 hours. The flow cytometer recorded the noise signals even when there was no *E. coli* added. These background signals remained unchanged from 1/16x MIC to 1x MIC after 3 hours of incubation (Fig. 4.1 A). When the sample was spiked with 10^3 CFU/mL of *E. coli* and incubated for an hour with penicillin g, no discernible signal was observed in the no-antibiotic control and thus no scatter

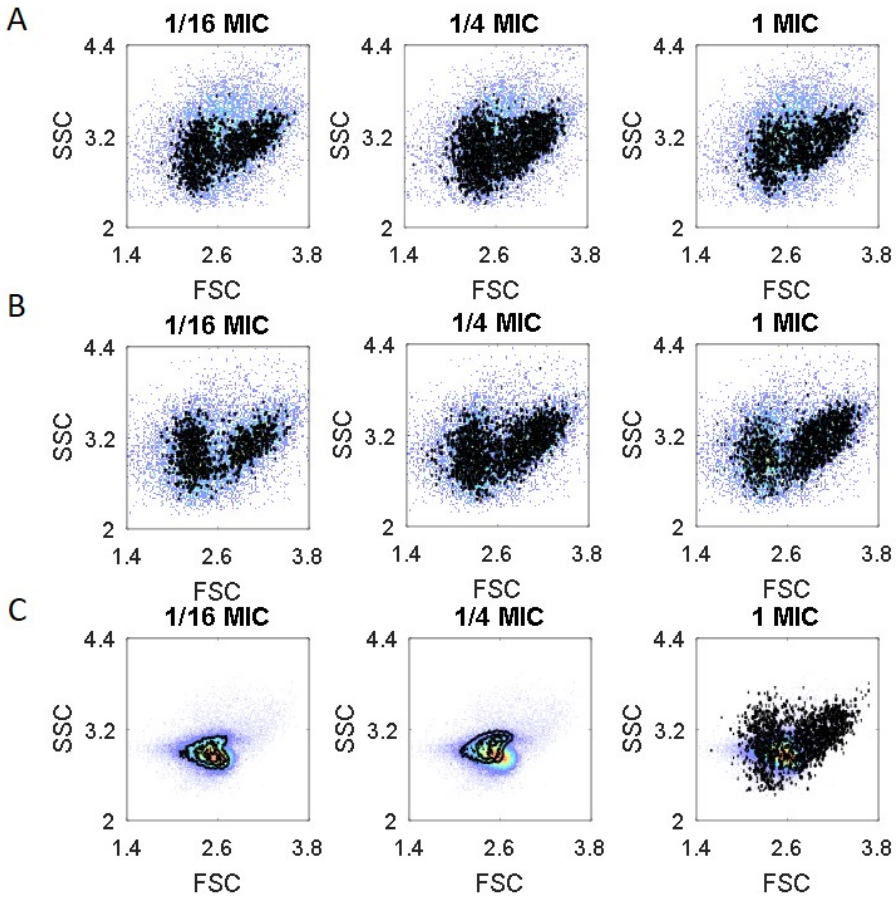

Figure 4.1: Detection limit for the flow cytometer. Flow cytometry data of (A) No *E. coli* control with 3 hours incubation. (B) 10^3 CFU/mL of *E. coli* spiked sample with 1 hour incubation. (C) 10^3 CFU/mL of *E. coli* spiked sample with 3 hours incubation. The black contours are the penicillin g- treated data with the penicillin g concentration labled on each figure. The psuedo-color plots are the no antibiotic controls. 1x MIC of penicillin g for *E. coli* strain ATCC 33456 is 32 μg/mL. FSC: forward scatter. SSC: side scatter.

signals change throughout the different concentration of penicillin g (Fig. 4.1 B). This is most likely because the bacteria concentration was too low to be observed with the flow cytometer. Since the doubling time of *E. coli* is about 20$\sim$30 minutes, the estimated *E. coli* concentration in the no-antibiotic treated sample should be around 4000 CFU/mL. On the other hand, when incubated for three hours, the final *E. coli* concentration should be around 6.4×10^4 CFU/mL. *E. coli* signals can be seen in the no-antibiotic control, and the 1/16x data. At 1/4x MIC, scattered-light signals start to shift while at 1x MIC, bacterial growth was inhibited. As a result, the 1x MIC data appears to be very similar to the no *E. coli* data

(Fig. 4.1 C). This shows that even when no blood cells exist, bacterial concentration amplification is necessary for the flow cytometer to detect sample with $\leqq$1000 of the bacterial counts.

4.2.2 Blood Cells Removal with Serum Separation Tube

Serum separator tubes (SST) are routinely used in the clinical lab to separate blood cells from serum for medical tests. When spinning down the blood sample in the SST, the blood cells penetrate into the gel layer at the bottom of the tube while the serum stays at the top. When the blood sample contains bacteria, it has been reported that the bacteria cells would be spun down on top of the gel layer thus separating from the blood cells.[197]

To examine whether SST can successfully separate bacteria from blood cells, three samples were tested including: (1) 10% human blood only, no *E. coli* control, (2) 10% blood spiked with 10^6 CFU/mL of *E. coli* (ATCC 33456) and (3) 10% blood sample spiked with 10^7 CFU/mL of *E. coli*. All three samples were loaded to the SSTs, inverted five times, waited for 30 minutes and spun down with a clinical centrifuge for ten minutes as the manufacturer (Becton Dickinson, Franklin Lake, NY) suggested. The supernatant (serum) was discarded, and 980 μL of LB broth was added to resuspend the bacteria. The solution was transferred to a 12-well plate that was loaded with 20 μL of MH-IR786 (final MH-IR786 concentration 900 nM) and incubated for 4.5 hours. Even though we have shown in Appendix Figure B.6 that the fluorescent signals from MH-IR786 fluctuated from data to data, MH-IR786 was used here since it has been shown that it can only be taken up by bacteria cells.[164] As a result, MH-IR786 should help distinguish the bacteria cells from blood cells. After the incubation, samples were collected and analyzed by flow cytometry.

As shown in Figure 4.2 A, the gel layer from the SST generated high cytometric background. As a result, even the blood only data is dominated by the signal from the SST generated background. This high noise also obscured the bacteria signals. The 10% blood only scatter data looks very similar to the *E. coli* spiked, 10% human blood data even when

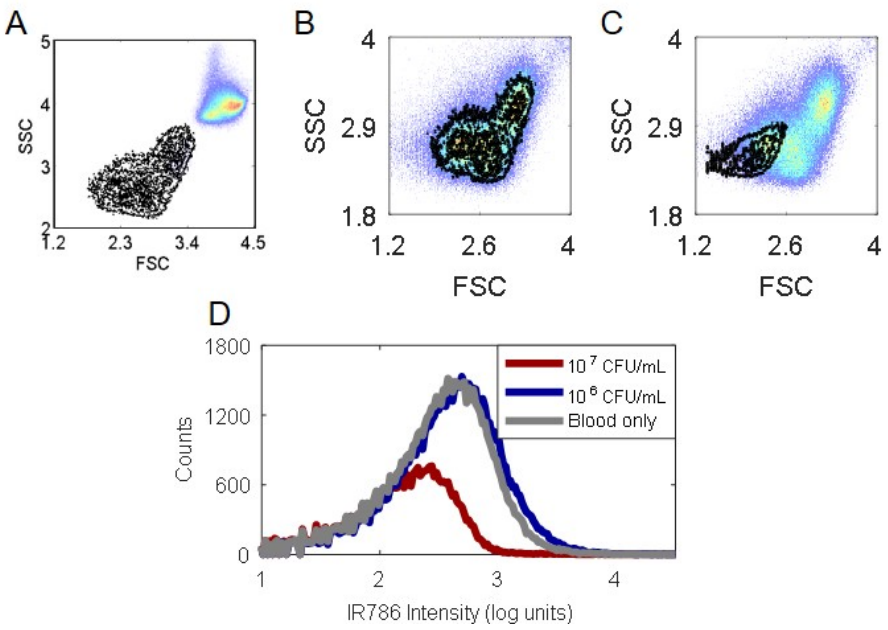

Figure 4.2: Failed attempts of *E. coli* separation using SST. (A) Blood only data. The black contour is the SST processed human blood after 4.5 hours of incubation. The pseudo-colo plot is the unprocessed human blood. For (B) to (C) The black contours are Flow cytometry data of (B) 10^6 CFU/mL of *E. coli* spiked human blood (C) 10^7 CFU/mL of *E. coli* spiked human blood. The psuedo-color plots are 10% human blood only. FSC: forward scatter. SSC: side scatter. (D) Cytometric data for IR786 fluorescence channel.

the inoculation concentration was as high as 10^6 CFU/mL and had been incubated for 4.5 hours (Fig. 4.2 B). A discernible scatter difference only appeared when the initial inoculated concentration was 10^7 CFU/mL (Fig. 4.2 C). The 1D fluorescence signal also only shows differences with the 10^7 CFU/mL spiked blood sample. The fluorescence signals of the 10^7 CFU/mL sample, however, was lower than the blood only sample (Fig. 4.2 D), which contradicts with the previous observation that mammalian cells do not uptake MH-IR786.[164] Or MH-IR786 is retained by the gel layer from the SST. With the high scatter backgrounds introduced by the gel layer of a SST, it would be difficult to detect any bacterial signal even after amplification. As a result, SST was excluded from further study.

4.2.3 Blood Cells Removal with Saponin

Although the blood cells were successfully removed by SST, the high background from the gel layers make it difficult to detect bacterial signal. Saponin, on the other hand, does not generate much background signal itself. Instead of removing the blood cells, it lyses them without affecting bacterial growth.[199–201] In this subsection, the hemolysis effect of saponin was characterized, the MH-IR786's ability to distinguish blood cells from bacteria was investigated and a pre-blood culture AST with sheep blood was demonstrated.

Effect of Saponin and MH-IR786 staining

To test saponin's lysis ability, 100 μL of 1% saponin was added to 100 μL of human blood with 800 μL of LB and incubated at 37 °C for 10 minutes. The lysed 10% blood solution was then spun down, and washed with PBS. The pellet was then resuspended in 980 μL of LB, loaded to a 12-well plate with 20 μL of 45 μL of MH-IR786 and incubated for an hour. With the saponin treatment, the scattered-light signal clearly shifted to the lower left corner with smaller side and forward scatter signal (Fig. 4.3 A). Since saponin lyses the blood cells, the smaller scattered-light signals show that the cells were indeed damaged and broken into debris. The fluorescence signal, probably because of increased accessibility of MH-IR786 into damaged cells, was higher when the blood cells were lysed (Fig. 4.3 A). As for *E. coli*, neither the scatter data nor the fluorescence signal changed with or without saponin treatment (Fig. 4.3 B) which is in consistent with the previous studies[199–201] showing that *E. coli* is not effected by saponin. When comparing the fluorescence intensities between MH-IR786 stained human blood cells and *E. coli*, the fluorescence intensity was not higher in the *E. coli* only data. Combining with the fact that the blood cells debris generated higher fluorescence intensity than did the no saponin sample, MH-IR786 is most likely not actively taken up by the healthy bacteria.

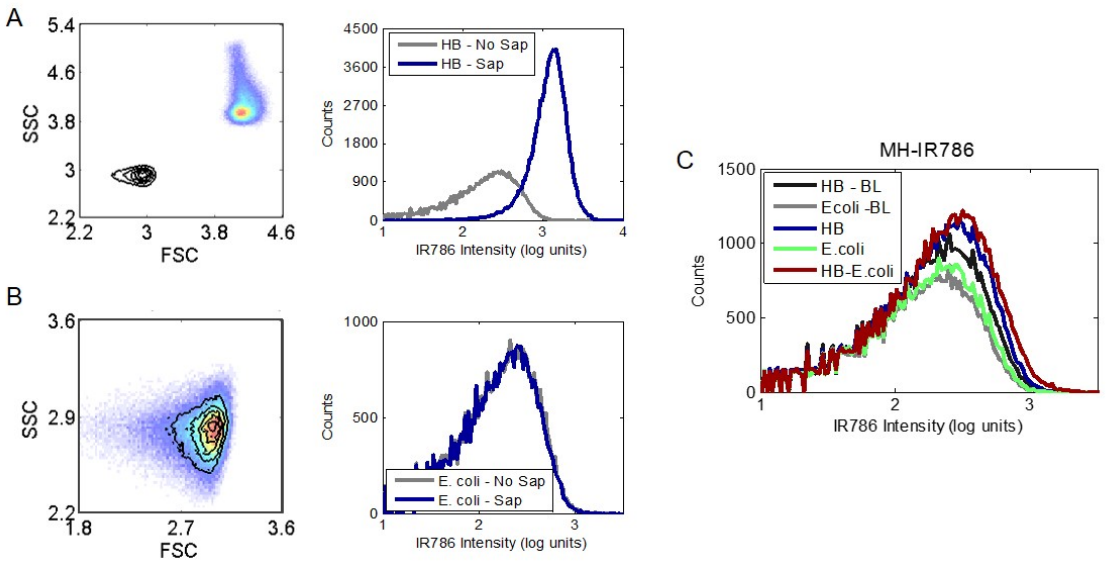

Figure 4.3: Saponin-treated human blood and *E. coli*. Flow cytometry data for (A) 10% human blood. (B) *E. coli*. For both (A) and (B) the black contours were the 1% Saponin treated data while the psuedo-color plots were without saponin treatment. (C) MH-IR786 fluorescence signal in *E. coli* and blood. HB: human blood. BL: blank (no dye).

Pre-Blood Culture AST with Sheep Blood

To search for the condition to separate bacteria from the blood cells, sheep blood was used as a substitute for human blood. In order to remove the blood cells that have a concentration that is 10^6 to 10^7 times higher than the bacteria, 100 μL of 10% (w/v) saponin was added to 1 mL of sheep blood that was spiked with either 100 μL of LB or 10^4 CFU/mL *E. coli* strain ATCC 33456 (final concentration $\sim 10^3$ CFU/mL) and incubated for 15 minutes at room temperature to lyse the blood cells. The same procedure was applied to 12 samples loaded in eppendorf tubes. These samples were then washed with PBS and resuspend in 500 μL LB broth. All 12 samples were added to the 12-well plate loaded with 480 μL of LB broth and/or penicillin g with 2x higher desired concentrations and 20 μL, 45 μM of MH-IR786. The plate was incubated for 5 hours. Each sample was then collected and analyzed by flow cytometry.

The cytometric data for blood only samples (100% blood) remained very similar to each other from no-antibiotic to 1x MIC of penicillin g (Figure 4.4 A) while clear growth

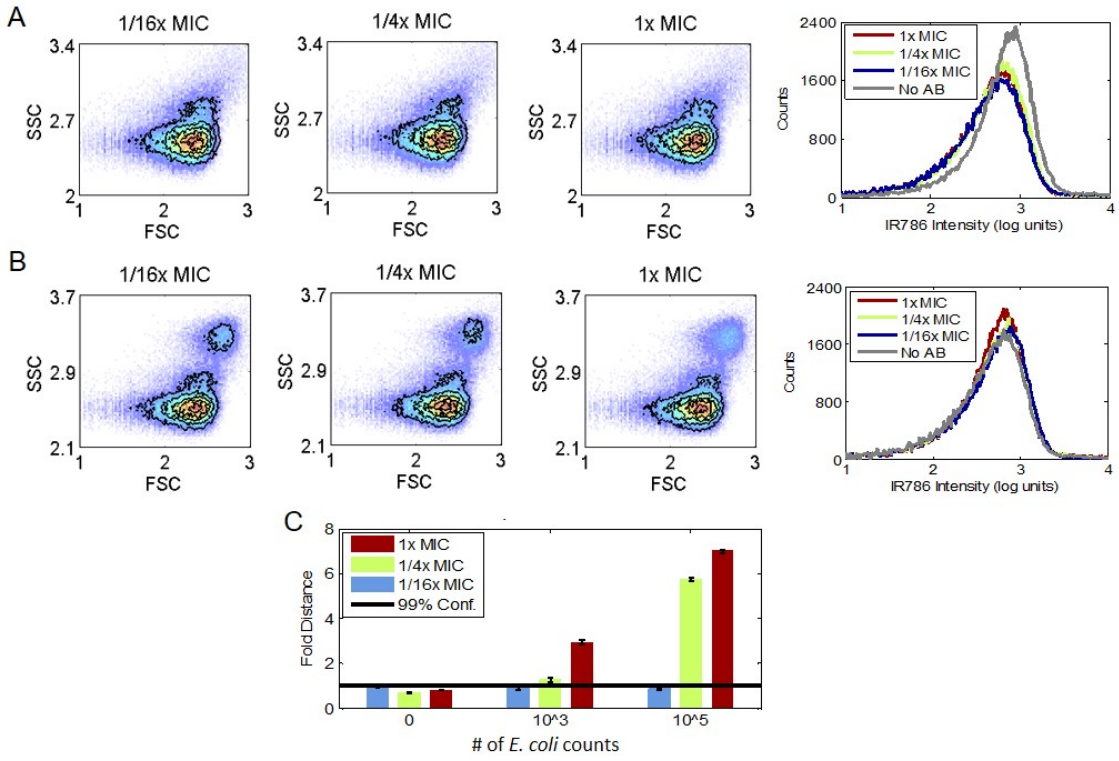

Figure 4.4: Pre-blood culture AST with sheep blood. Flow cytometry data for (A) 100% sheep blood only. (B) 1000 CFU/mL *E. coli* spiked blood sample. The black contours are the penicillin g-treated data with the penicillin g concentration labeled on each figure. The psuedo-color plots are the no antibiotic controls. 1x MIC of penicillin g is 32 μg/mL for *E. coli* strain ATCC 33456. FSC: forward scatter. SSC: side scatter. (C) PB-sQF results for (A), (B) and 10^5 CFU/mL spiked blood sample.

inhibition can be observed at 1x MIC in the 1000 CFU/mL of *E. coli* spiked sample (Figure 4.4 B). The differences in the 3D cytometric data are seen in PB-sQF results with the blood only data showing no statistically significant difference between each other while both 1000 CFU/mL and 10^5 CFU/mL show clear increment of distance increases from 1/16x MIC to 1x MIC (Figure 4.4 C).

4.2.4 Characterize the Killing Efficiency of Blood Cells over Bacteria

The same blood cell lysis condition that was found for the sheep blood was applied to the human blood sample. The lab-strain *E. coli*, however, were not recovered from the blood sample as they were in the sheep blood, but were killed by the human blood cells instead.

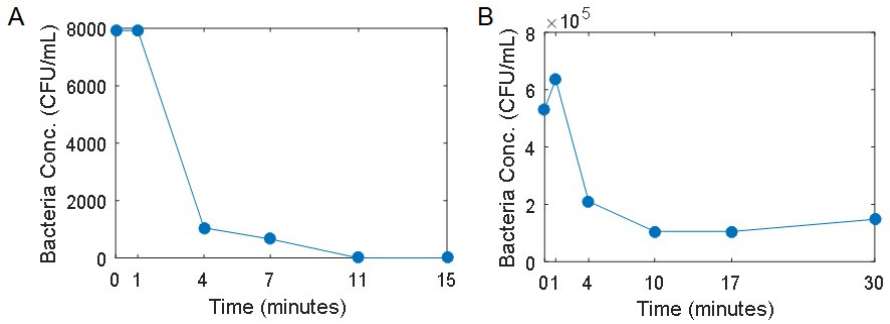

Figure 4.5: Human blood cells kill *E. coli*. (A) Lab strain ATCC 33456. (B) Clinical-isolate *E. coli* strain Mu14S.

To demonstrate this, 800 μL of 10^5 CFU/mL lab strain *E. coli* were incubated with 7200 μL of 10% human blood, and 800 μL of 1% saponin for 15 minutes. 1 mL of sample was taken out for overnight plating at 0, 1, 4, 7, 11 and 15 minutes of the incubation time. These 1 mL samples were diluted and plated overnight for colony counting. As shown in Figure 4.5 A, the colony counts kept decreasing from 1 minute of incubation time to 11 minutes when all of the *E. coli* were killed (no colonies detected). This shows that even with the present of saponin, 10^4 CFU/mL of the lab strain *E. coli* was readily killed by 10% human blood in 11 minutes. As a result, the same procedure, recovering 1000 CFU/mL of the lab strain *E. coli* from 100% of sheep blood, did not work in human blood.

Since bacteria isolated from blood should have a higher resistance to blood cells, the multidrug-resistant *E. coli* clinical isolate Mu14S was tested. Different from the lab strain *E. coli*, 10^5 CFU/mL of Mu14S were incubated with 10% human blood without saponin for 30 minutes. 1 mL of sample was taken out for over night plating at 0, 1, 4, 10, 17, and 30 minutes. The colony counts dropped for the first 10 minutes but remained stable and probably actively growing from 10 to 30 minutes. This shows that the clinical isolates indeed survives in human blood as the case in sepsis patients.

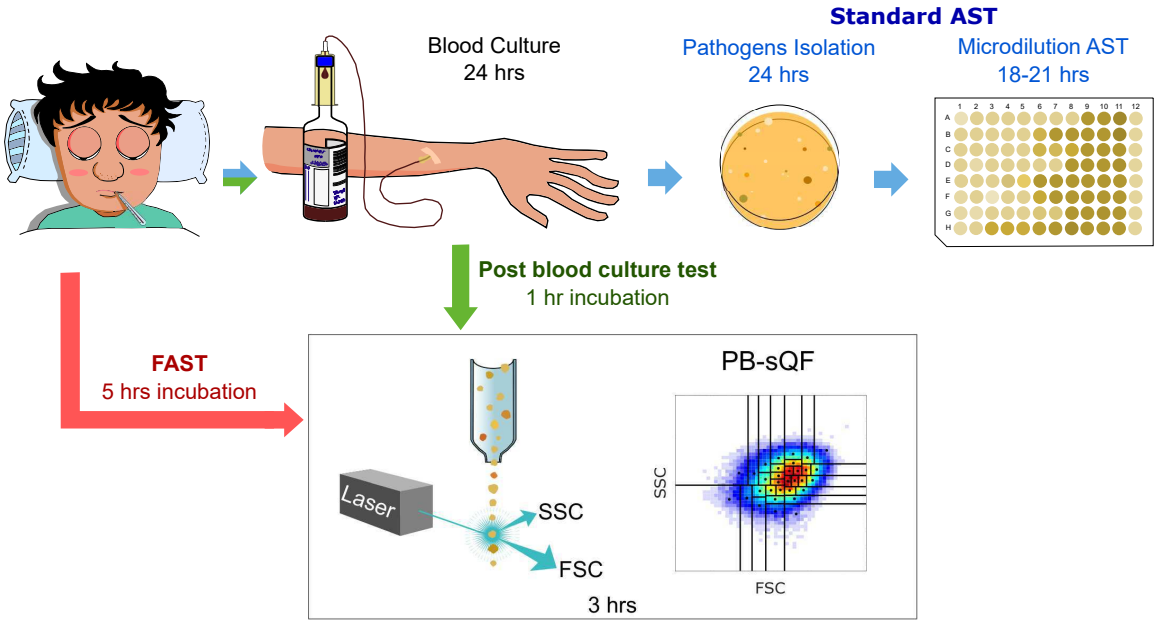

Figure 4.6: Antibiotic susceptibility test (AST) timelines. (Top, blue arrows) The standard clinical microbiology workflow requires >60 hours from initial blood draw. (Green arrows) Time line for the post-blood culture cytometric AST using PB-sQF distances.[48] (Red arrows) Time line from initial blood draw for Fast AST (i.e. FAST). FSC: forward scatter. SSC: side scatter.

4.3 FAST with Bacteria-Spiked Human Blood

Using the clinically isolated *E. coli* strain Mu14S, a pre-blood culture Fast AST(i.e. FAST, Fig. 4.6) was developed. To simulate blood from a patient with bacteremia, the clinical isolates were grown, diluted to the appropriate CFU/mL and added to the blood/saponin mixture to achieve the final diluted sample. Initial cultures for bacteria-laden blood samples were prepared using LB broth for incubating *E. coli*. For other bacteria, CAMHB was used. Bacteria were cultured overnight in an incubator shaker at 37°C and 225 rpm. The fresh bacterial culture was started from ∼0.05 OD by inoculating a 6 mL fresh growth medium with overnight culture. After the fresh culture reached mid-log phase, bacteria were diluted into ∼10 CFU/mL through a series of 10-fold dilutions and the concentration was confirmed by overnight plating from loading 100 μL of 1000 CFU/mL sample. The last 10-fold dilution was done by adding the 500 μL of 100 CFU/mL into 4500 μL of 10%

human blood (ZenBio, Research Triangle Park, NC) in medium solution.

2.5% (w/v) of saponin (Sigma-Aldrich, St. Louis, MO) was prepared, sonicated (Branson 2510, Emerson, St. Louis, MO) for 20 minutes and spun down with a clinical centrifuge (Centrific Model 228, Fisher Scientific, Waltham, MA) for 4 minutes. The supernatant was collected to isolate the undissolved pellet. 500 μL of 2.5% saponin was then added to the 5 mL of $\leqq$10 CFU/mL, 10% human blood sample and incubated in an incubator shaker for 15 minutes at 37°CṪo ensure that the blood cells lysed completely, the sample was laid on the incubator floor, confined by the large flask clamps, and agitated at 300 rpm. The sample was flipped by hand every 5 minutes. After the saponin treatment, the bacteria were again pelleted and washed with 2 mL of PBS using a clinical centrifuge for 2 minutes. 2.5 mL of growth medium was then added to the tube without breaking the pellet and incubated for 2 hours in an incubator shaker at 37°C and 225 rpm.

After the 2-hour incubation, the pellet was removed by pipetting. The sample was mixed well and 500 μL of the sample was added to each well of one row of the 12-well plate (4 samples per row) that was loaded with 500 μL of growth medium with or without antibiotic at 2-fold of the specified concentrations. The plate was then incubated at 37°C for 3 hours. Bacteria were again collected by centrifugation and resuspended in 200 μL of PBS for flow cytometry detection. To ensure each clinically isolated strain was tested at its MIC, pure culture starting with 1000 CFU/mL was also tested for each experiment, confirming that the antibiotic concentrations we used indeed inhibited bacterial growth.

Bacterial growth inhibition was monitored by flow cytometry, and the differences in the 2-D scatter histograms with and without antibiotic treatment was characterized with PBsQF statistics.[48] Since the FAST procedure lysed the blood cells and incubated the sample in growth medium, it greatly reduces the time-to-result compared to the standard AST (Fig. 4.6). Here, multidrug resistant, blood stable clinical isolates of common bacteremia-causing pathogens (*E. coli*, *K. pneumoniae*, *A. nosocomialis*), and *S. aureus* were obtained (Gift from Dr. Phillip Rather, Emory University), and FAST was applied to identify the

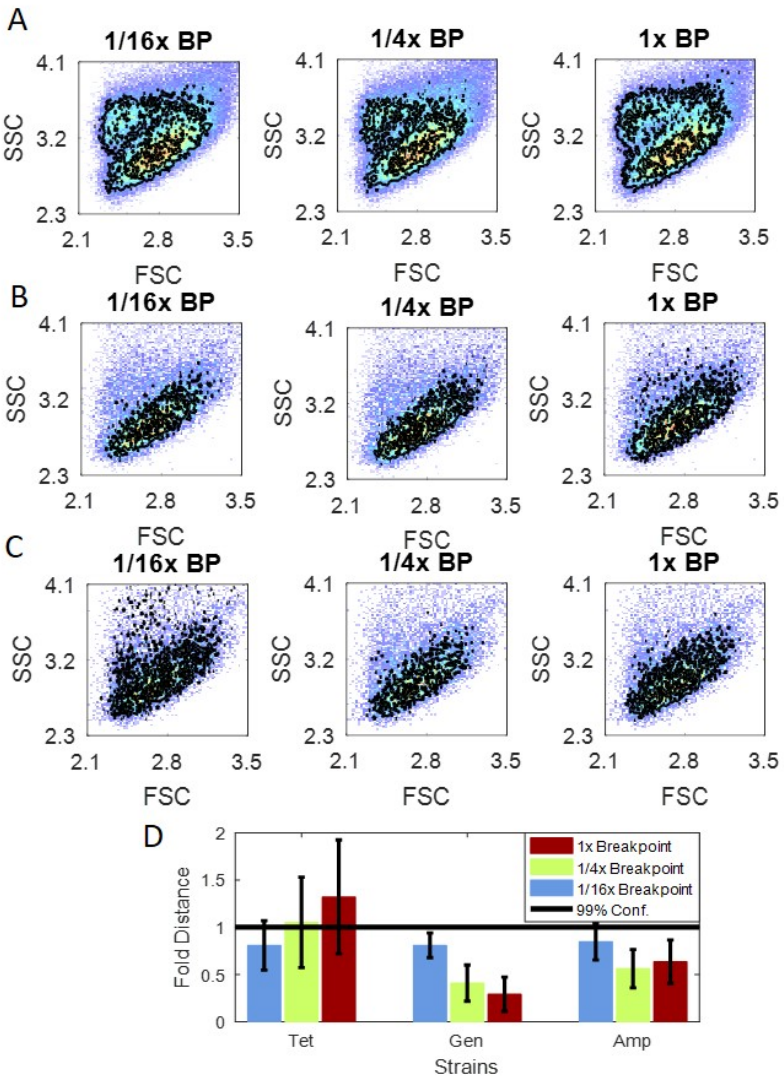

Figure 4.7: Antibiotic-treated 10% human blood only results. Cytometric data with (A) Ampicillin (B) Tetracycline (C) Gentamicin. The pseudo-color plots are the no-antibiotic controls and the black contour plots are the antibiotic-treated data with the antibiotic concentration indicated at each plot. (D) PB-sQF results for (A), (B), and (C). The resistant breakpoint of Enterobacteriaceae are 16 μg/mL for tetracycline and gentamicin. 32 μg/mL for ampicillin.

effective treatment.

4.3.1 FAST with Blood Only Sample

To show that the saponin procedure and the antibiotic treatments at the tested concentration do not significantly changes the scattered light histograms, controls of identical treatments

of 10% human blood samples without bacteria inoculation were tested (Fig. 4.7). When treated at the resistant breakpoint of tetracycline, gentamicin, and ampicillin of *Enterobacteriaceae*, the scatter signal remained unchanged from 1/16x, 1/4x to the 1x resistant breakpoint. Thus, diluting blood-stable bacteria-containing blood samples 1:9 (v:v) directly into saponin-containing growth medium provides a path to ASTs within 8 hours from initial blood draw, with excellent results matching independent (36-44 hr) MIC determinations from pure, overnight cultures ($\sim 10^8$ CFU/mL) that could only be initiated after ($\sim$24hr) positive blood culture.

4.3.2 FAST with Gram-Negative Bacteria

For gram-negative bacteria, tetracycline, gentamicin, and ampicillin were tested. The MICs of each clinical isolates were determined by microdilution first and are listed in Table 4.1. For antibiotics to which the clinical isolates were sensitive, the MIC was used to test the growth inhibition. For antibiotic to which the clinical isolates were resistant, the resistance breakpoints determined by the CLSI were used.

Table 4.1: MIC (μg/mL) for each antibiotic/bacteria combination. The MICs were determined from microdilution AST. S, I and R represents sensitive, intermediate and resistant according to the 2014 Clinical & Laboratory Standards Institute (CLSI) handbook.[177]

MIC (S/I/R)	Tetracycline	Gentamicin	Ampicillin
E. coli Mu890	1 (S)	8 (I)	$>$ 1024 (R)
E. coli Mu14S	$>$ 64 (R)	8 (I)	$>$ 1024 (R)
K. pneumoniae Mu670	2 (S)	4 (S)	$>$ 1024 (R)
K. pneumoniae Mu55	$>$ 64 (R)	1 (S)	$>$ 1024 (R)
A. nosocomialis M2	1 (S)	2 (S)	$>$ 1024 (R)

Multidrug-resistant E. coli isolates

Two clinically isolated, multi-drug resistant *E. coli* strains were tested, Mu890 and Mu14S. Following the hemolysis and growth procedure outlined above, flow cytometry was used

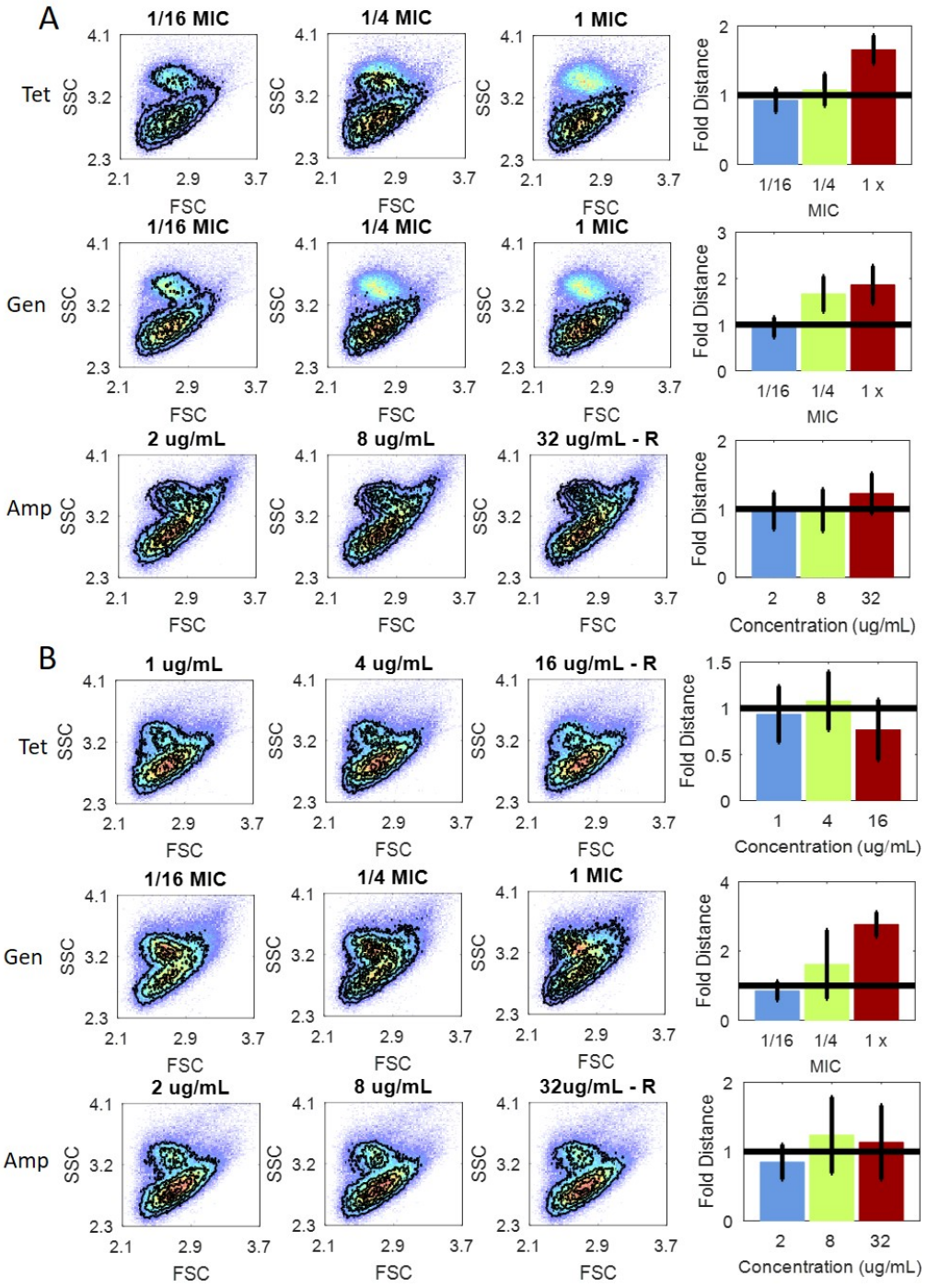

Figure 4.8: FAST antibiotic-induced scatter signals for *E. coli* strains Mu890 and Mu14S. (A) Mu890 antibiotic induced scatter histograms (black contours) overlaid on paired no-antibiotic control (color dots, red indicating highest occurrence) and PB-sQF results. (B) Mu14S antibiotic induced scatter histograms and PB-sQF results. For the PB-sQF results bar chart, the thick black lines correspond to each bacteria-antibiotic 99% confidence limit distance, and error bars represent one standard deviation above and below the mean from triplicate trials.

to collect forward and side scattered light signals. Statistical comparison of these histograms (Fig. 4.8) demonstrates that susceptibility testing is readily performed by immediate hemolysis of 0.5 mL blood/antibiotic, followed by 2-hrs preincubation and 3-hr AST. When treated with concentrations of either tetracycline or gentamicin below the sensitive (4 μg/mL for tetracycline) and intermediate resistance (8 μg/mL for gentamicin) breakpoints, the Mu890 signals disappeared, indicating effective growth inhibition (Fig. 4.8 A). When treated with 32 μg/mL ampicillin (the *Enterobacteriaceae* resistance breakpoint), the scatter signals became indistinguishable from those of the no-antibiotic control (Fig. 4.8 A). PB-sQF fold distance-based FAST beyond the 99% confidence levels match the much slower microscan AST data, demonstrating that tetracycline and gentamicin are indeed effective treatments for Mu890 (Fig. 4.8 A). The actual starting concentrations of Mu890 were confirmed with overnight plating to be 3, 3, and 5 CFU/mL for tetracycline, gentamicin, and ampicillin experiments, respectively. Since the *E. coli*/human blood was diluted 10-fold, the real concentrations before dilution corresponded to $\sim$30, 30 and 50 CFU/mL of whole blood.

Also matching its standard AST data, FAST shows that Mu14S is intermediate (1x MIC = 8 μg/mL) to gentamicin and when treated at the MIC, exhibits growth inhibition (Fig 4.8 B). When treated with tetracycline or ampicillin at each resistant breakpoint (16 μg/mL and 32 μg/mL), however, Mu14S signals remained statistically unchanged (Fig. 4.8 B). The PB-sQF fold distance average from triplicate data shows clear differences between the 1x MIC gentamicin data versus the paired-control (Fig. 4.8 B). This confirms that the gentamicin sensitivity of Mu14S observed after blood culture[48] can also be observed with FAST. Overnight plating confirms that initial Mu14S counts were 3, 2, 5 and CFU/mL for tetracycline, gentamicin and ampicillin data after 10-fold dilution of the blood/bacteria mixture, corresponding to FAST being performed on whole blood samples containing $\sim$30, $\sim$20 and $\sim$50 CFU/mL. Control experiments starting with 1000 CFU/mL of both Mu890 and Mu14S without human blood were incubated with antibiotics for 5 hours. The results

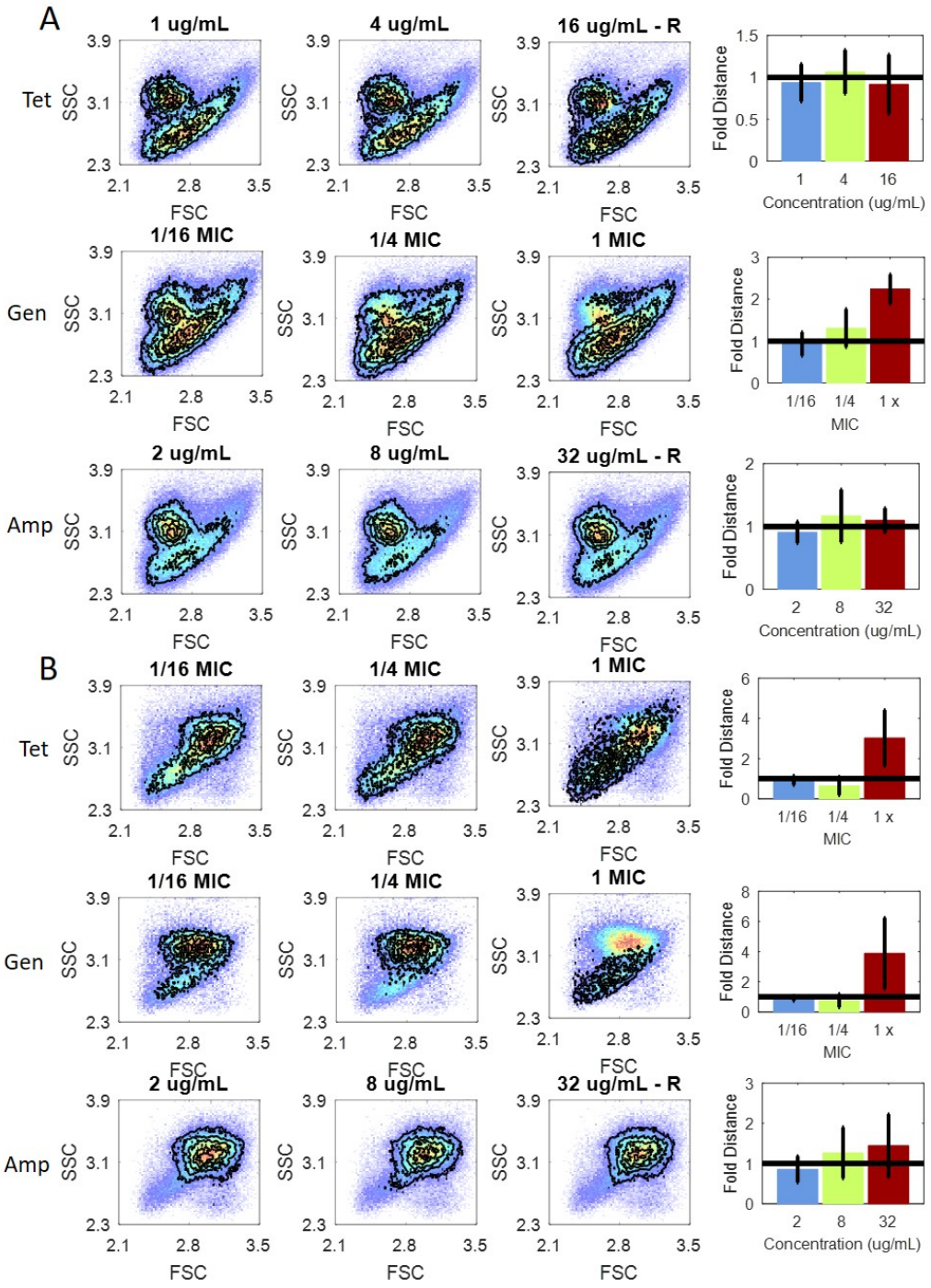

Figure 4.9: FAST antibiotic-induced scatter signal changes for Mu55 and Mu670 reveal different susceptibilities. (A) Mu55 antibiotic-induced scatter histograms (black contours) overlaid on paired no-antibiotic control (color dots, red indicating highest occurrence) and PB-sQF results. (B) Mu670 antibiotic-induced scatter histograms and PB-sQF results. For the PB-sQF results bar chart, the thick black lines correspond to each bacteria-antibiotic 99% confidence limit distance, and error bars represent one standard deviation above and below the mean from triplicate trials.

are consistent with the FAST results (Appendix Fig. C.1 and Fig. C.2).

Multidrug-resistant K. pneumoniae isolates

The same FAST procedure was applied to two clinically isolated, multidrug resistant *K. pneumoniae* strains, Mu55 and Mu670. Analogous to the *E. coli* data, *K. pneumoniae* growth inhibition is directly quantified with PB-sQF upon effective antibiotic treatment, and sensitivities are accurately determined. Importantly, when treated with antibiotics to which Mu55 or Mu679 were resistant, the scatter data (black contours) were not statistically different from each experiment's paired control (pseudo-color plots, Figure 4.9). PB-sQF confirms that tetracycline is effective toward Mu670 and gentamicin is an appropriate treatment for both Mu55 and Mu670. Post-dilution bacterial concentrations were confirmed by overnight plating to be $\sim$8 CFU/mL for Mu55 and $\sim$9 CFU/mL for Mu670, demonstrating that FAST can be readily completed within 8 hours of initial blood draw on blood samples exhibiting $\sim$100 CFU/mL. Again, control experiment starting with 1000 CFU/mL of both strains without human blood were incubated with antibiotics for 5 hours. This ensures that the antibiotics were indeed at the MIC (Appendix Fig. C.3 and Fig. C.4).

Multidrug resistant A. nosocomialis isolates

Also matching its standard AST data, FAST on clinically isolated *A. nosocomialis* strain M2 spiked in 10% human blood (Fig. 4.10) enabled its susceptibility profile to be quickly determined. As with other species, PB-sQF reveals that M2 is resistant to ampicillin but susceptible to both tetracycline and gentamicin when assayed within 8 hrs of initial simulated blood draw (Figure 4.10). Different from the $\sim 10^5$ CFU/mL *E. coli* and *K. pneumoniae* strains resulting from 2-hr pre-incubation plus 3-hr AST, the final *A. nosocomialis* concentrations were $\sim 10^4$ CFU/mL, as confirmed by plating. Even with an order of magnitude fewer bacterial counts, the clear growth inhibition was readily quantified with the same procedure. As initial $\sim$10 CFU/mL samples incubated for 5 hours are sufficient for

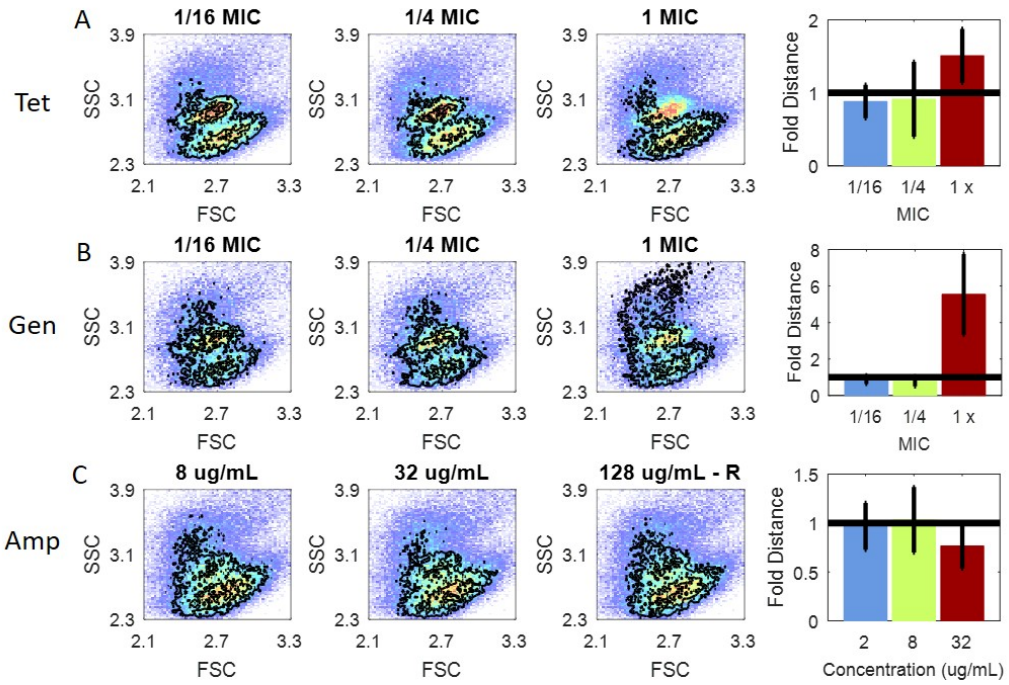

Figure 4.10: FAST antibiotic-induced scatter signal changes for *A. nosocomialis* strain M2. Flow cytometry data of antibiotic induced scatter histograms (black contours) overlaid on paired no-antibiotic control (color dots, red indicating highest occurrence) and PB-sQF results. (A) Tetracycline (B) Gentamicin (C) Ampicillin. For the PB-sQF results bar chart, the thick black lines correspond to each bacteria-antibiotic 99% confidence limit distance, and error bars represent one standard deviation above and below the mean from triplicate trials.

analysis, ∼10 doubling events are ideal to generate sufficient sample for FAST antibiotic panels to be performed on any bacteria. Pure culture starting from 1000 CFU/mL were incubated for 5 hours with antibiotics. The results, as shown in Appendix Figure C.5, show that the antibiotics were indeed treated at the MIC/breakpoint of M2.

4.3.3 FAST with Gram-Positive Bacteria

S. aureus clinical isolates, strain 95938 and strain NRS382, were obtained from Dr. Sarah Satola at the Georgia Emergin Infection Program of Emory University, to test the FAST procedure. Based on the clinical microbiology reports, both stains are hemolytic. The MICs were first determined by microdilution and are listed in Table 4.2.

Table 4.2: MIC (μg/mL) for each antibiotic/bacteria combination. The MICs were determined from microdilution AST. S, I and R represents sensitive, intermediate and resistant according to the 2014 Clinical & Laboratory Standards Institute (CLSI) handbook.[177]

MIC (S/I/R)	Vancomycin	Oxacillin	Gentamicin
S. aureus strain 95938	1 (S)	64 (R)	128 (R)
S. aureus strain NRS382	2 (S)	256 (R)	$\frac{1}{4}$ (S)

FAST with 5-hr Incubation

FAST was applied on *S. aureus* strain NRS382 with vancomycin, oxacillin, and gentamicin. NRS382 was treated at its MICs with vancomycin (2 μg/mL) and gentamicin $\frac{1}{4}\mu$g/mL). Since NR382 is resistant to oxacillin (MIC = 256 μg/mL), the resistant breakpoint, 4 μg/mL was used instead. As shown in Figure 4.11, when started at $\sim$20 CFU/mL with 10% human blood, no scattered-light signal change was observed and no colonies were recovered (with 1/10x dilution) after overnight plating of the samples. The pure culture control (Appendix Figure C.6), however, shows that the antibiotics were at the correct concentrations. These data indicates that with 5 hours incubation, *S. aureus* strain NRS382 was not recovered from the blood culture.

Growth Curve and Blood Stability

Since the FAST procedure with total of a 5 hours of incubation did not recover *S. aureus* strain NRS382, the growth curves were measured. Bacteria (both *S. aureus* strain NRS382 and strain 95938) were cultured overnight as described previously. The fresh bacterial culture was started from $\sim$0.05 OD by inoculating a 10 mL fresh growth medium (CAMHB) with overnight culture. The OD of both strains increased steadily over time (Figure 4.12). The doubling time, however, was around 80 minutes, much longer compared to $\sim$20-30 minutes for *E. coli*, *K. pneumoniae*, and *A. nosocomialis*. As a result, instead of ten or more doubling events, *S. aureus* strain NRS382 only went through less than four doubling

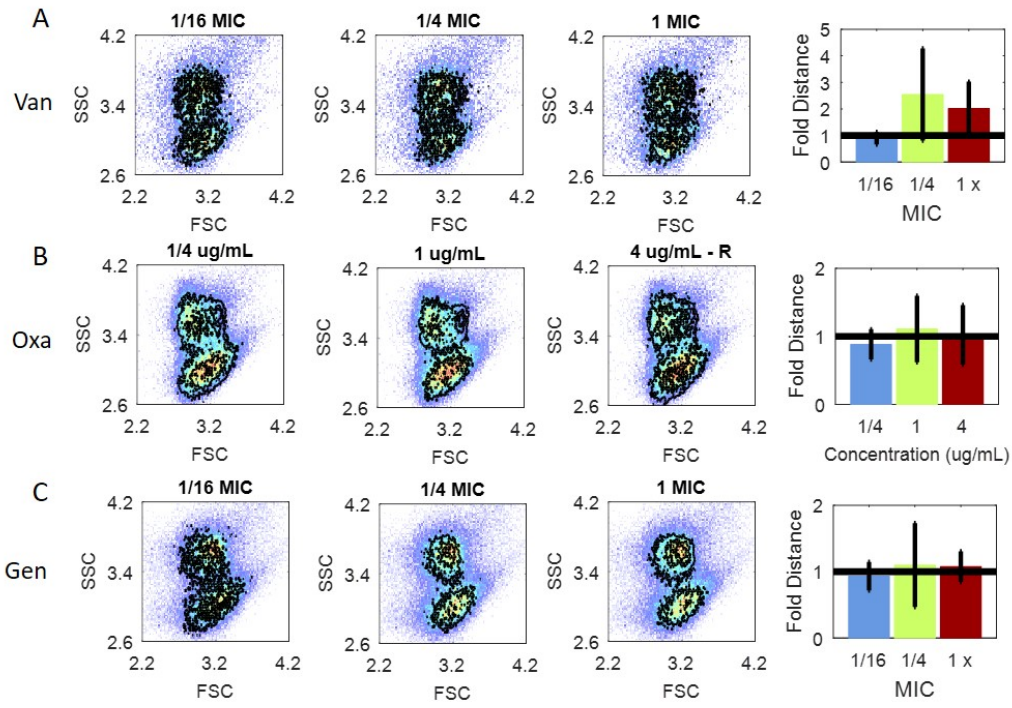

Figure 4.11: FAST antibiotic-induced scatter signal changes for *S. aureus* strain NRS382. Flow cytometry data of antibiotic induced scatter histograms (black contours) overlaid on paired no-antibiotic control (color dots, red indicating highest occurrence) and PB-sQF results. (A) Vancomycin (B) Oxacillin (C) Gentamicin. For the PB-sQF results bar chart, the thick black lines correspond to each bacteria-antibiotic 99% confidence limit distance, and error bars represent 1 standard deviation above and below the mean from triplicate trials.

events with 5 hours of incubation. With the starting bacterial concentration $\sim$ 20 CFU/mL, the end point concentration after incubation was $<$ 640 CFU/mL assuming no lose of bacteria during the washing step. Thus, no scattered-light shift was observed and no colony recovered (Fig. 4.11).

To ensure the blood culture did not kill *S. aureus* strain NRS382, the stability of bacteria in blood was tested. Mid-log phase fresh bacterial culture ($\sim 10^8$ CFU/mL) underwent 10-fold dilutions to 10^4 CFU/mL. The final solution was 9 mL of 10^4 CFU/mL of bacteria with 10% of human blood in CAMHB. 700 μL of 2.5% of saponin was then added and 1 mL sample was taken out and diluted 10-fold for plating ($t = 0$ data point) immediately. The bacteria counts were confirmed by plating at different incubation durations. At each

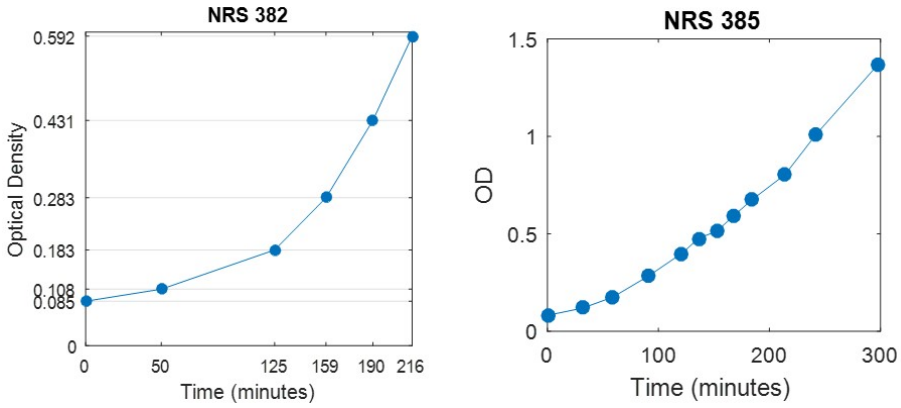

Figure 4.12: Growth curve for *S. aureus* strain 95938 and strain NRS382.

time point, the quantity of sample taken out and the amount of dilution vary depend on the expected growth/loss of the bacteria. As shown in Figure 4.13 A, both strains survived when incubated with 10% human blood and saponin in broth. The plate counts after 30 minutes of incubation were higher than the initial plate counts, suggesting a steady bacterial growth in the blood sample. The same process was repeated with *S. aureus* strain 95938 along with two more conditions: saponin only and blood only. The plate counts results show that even when the blood cells were not lysed by saponin, *S. aureus* strain 95938, which is a bloodstream isolate, still survived after incubated 30 minutes with 10% of blood. Also, when incubating strain 95938 with saponin without blood, no significant decrease in plate counts was observed. These data show that the *S. aureus* isolates are stable in blood and the saponin only lyses the blood cells.

Overnight Incubation

Since the doubling time for the clinically-isolated *S. aureus* strain 95938 are $\sim$ 80 minutes, a longer incubation time is necessary to reach the needed $\sim 10^4$ CFU/mL final bacterial concentration. 14 hours of total incubation time (3 hours of pre-incubation and 11 hours of AST) were tested. As shown in Figure 4.14, the cytometric data are dominated by the blood debris. Among the six no-antibiotic controls (triplicate data for each of two antibiotics), only two of them successfully recovered bacteria with plate counts around 2×10^6 and

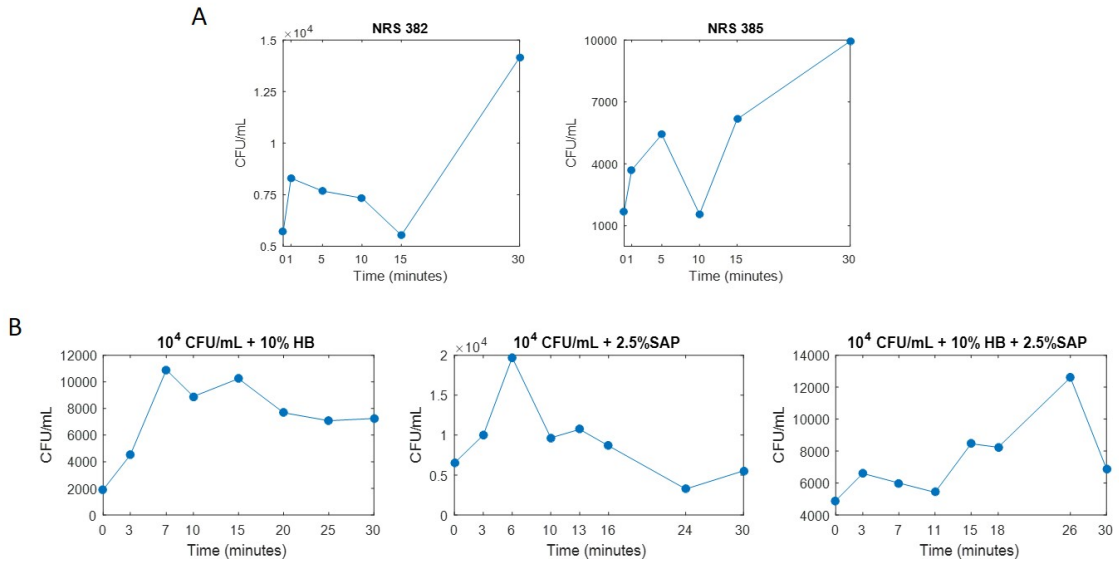

Figure 4.13: Overnight plate counts for 10% human blood incubated *S. aureus* strain 95938 and strain NRS382. (A) Strain 95938 and strain NRS382 with 10% human blood and saponin. (B) Strain 95938 with blood only (left), saponin only (middle), and both blood and saponin (right).

8×10^5 CFU/mL respectively. The rest had no colonies observed. This result contradicts the data shown in Figure 4.13, which indicates that the plate counts of *S. aureus* clinical isolates did not drop significantly after 30 minutes of incubation with blood. One possible explanation is that when the bacteria contact the blood sample, the competition between the killing time and doubling time begin. When measuring blood stability, since there were 10^4 CFU/mL of bacteria, part of the *S. aureus* have the chance reproduce before it was killed by the blood cells. In the pre-blood culture experiment, the bacteria count is as low as 10 CFU/mL. Unlike *E. coli*, *K. pneumoniae*, and *A. nosocomialis* clinical isolates which double every $\sim 20 - 30$ minutes, the doubling time of *S. aureus* strain 95938 is ~ 80 minutes. The probability that all ten CFU/mL of *S. aureus* survive before doubling is low.

To speed up the doubling time, different broth other than CAMHB can be considered. It is know that blood agar is useful to cultivate fastidious organisms such as *Streptococcus spp.*. *S. aureus* strain 95938 and NRS382 indeed grow faster on blood agar plates than LB agar plates. Also, the standard AST broth for *S. aureus* and *Streptococcus spp.* is CAMHB with 2$\sim$5% of lysed horse blood. As a result, with suitable culture medium, the doubling

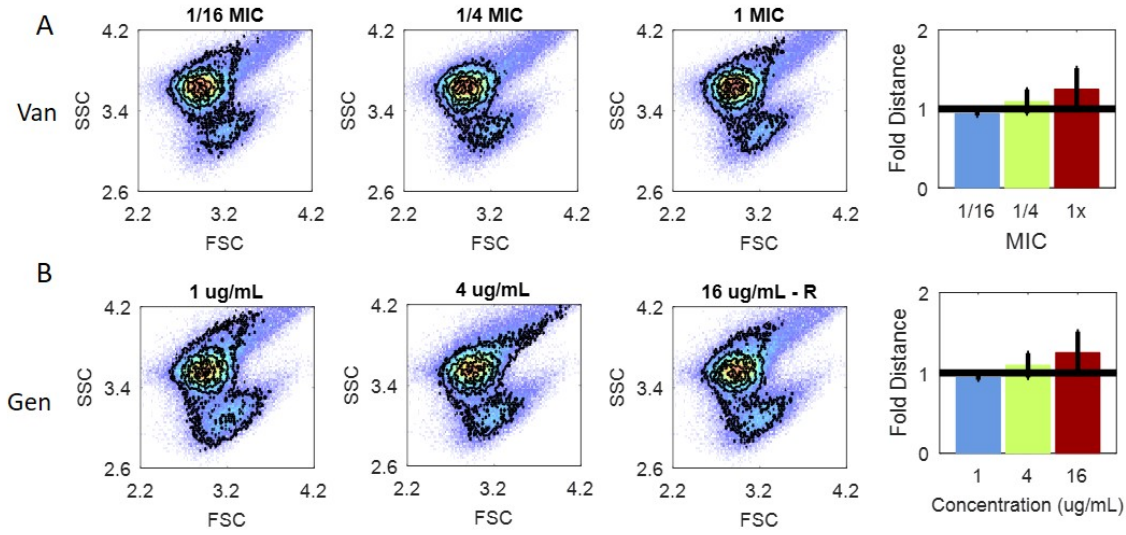

Figure 4.14: FAST antibiotic-induced scatter signal changes for *S. aureus* strain NRS382 14-hr culture. Flow cytometry data of antibiotic induced scatter histograms (black contours) overlaid on paired no-antibiotic control (color dots, red indicating highest occurrence) and PB-sQF results. (A) Vancomycin (B) Gentamicin. For the PB-sQF results bar chart, the thick black lines correspond to each bacteria-antibiotic 99% confidence limit distance, and error bars represent one standard deviation above and below the mean from triplicate trials.

time of *S. aureus* strain 95938 and NRS382 may decrease and the FAST procedure might be applied.

4.4 FAST with Different Conditions

Since PB-sQF utilizes the completed information of the cytometric data without gating, FAST is robust for different experimental conditions that affect the scatter patterns, such as biovariability between patients, different machines and/or cytometer parameters. In this section, *E. coli* strain Mu14S was spiked into the blood from another vendor and cytometric data of *E. coli* strain Mu890 were taken with different cytometer alignments. Nevertheless, without any modification in PB-sQF, FAST reveals the same susceptibility profiles as before (Fig. 4.8 and Fig. 4.9) for both strains regardless of the changes in scatter patterns.

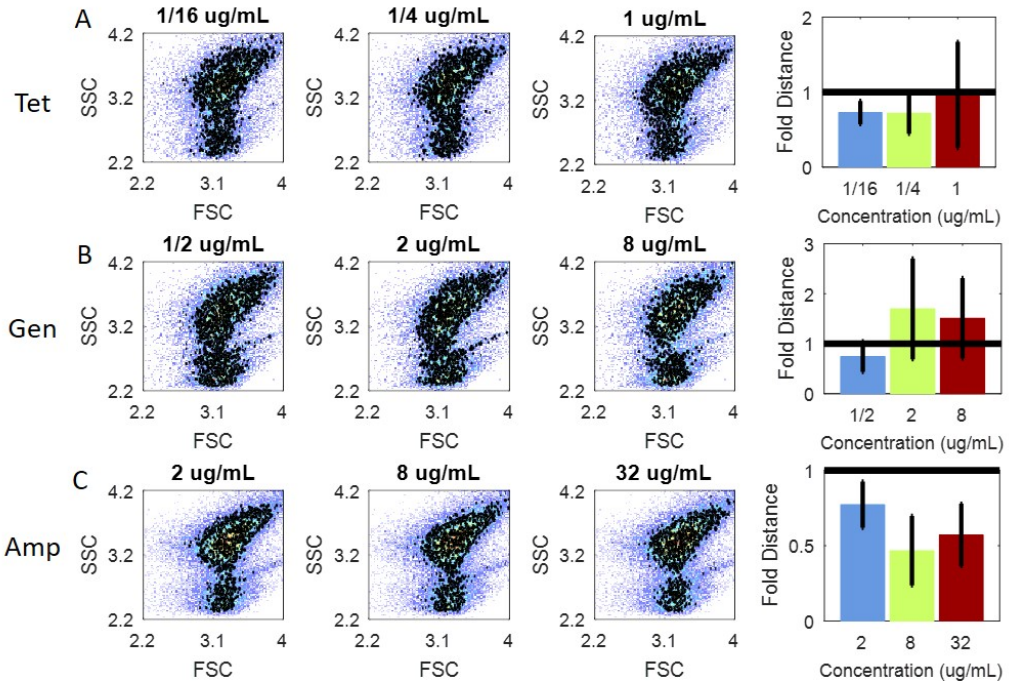

Figure 4.15: Antibiotic-treated 10% human blood from United States Biological. For all data, pseudocolor plot: no-antibiotic, paired control. Black contour: antibiotic-treated data. (A) Tetracycline at 1 μg/mL (MIC for Mu14S). (B) Gentamicin at 8 μg/mL (MIC for Mu14S). (C) Penicillin G at 32 μg/mL, the resistant breakpoint for penicillin group for *E. coli*. For the PB-sQF results, none of the antibiotics induce significant scatter signals shift for blood only data. All data were done in triplicate.

4.4.1 Blood Sample from Different Vendor

The blood samples used in this subsection were purchased from USBiology, Salem, MA. The same experimental procedure was applied to the the no-bacteria control or the *E. coli* spiked blood sample and the results are shown in Figure 4.15 and Figure 4.16.

FAST with Blood Cells Only

The scatter pattern (Fig. 4.15) of the USBiological blood sample taken by another machine, was distinctly different from the ZenBio blood sample (Fig. 4.7). But as in Figure 4.7, the scatter pattern stays unchanged throughout the 16-fold increment of antibiotic concentrations and the PB-sQF test results show no statistically significant change from the 99% confidence level. This shows that since the entire cytometric data was used in the PB-

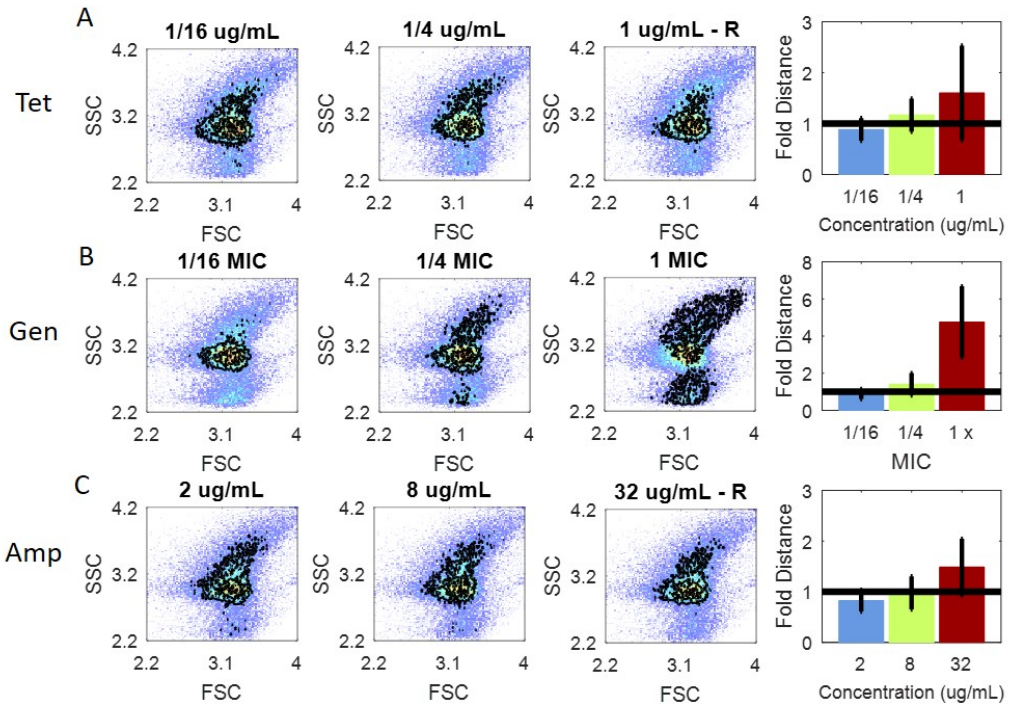

Figure 4.16: Bactericidal Antibiotic-induced scatter changes for *E. coli* strain Mu14S in USBiological Blood. For all data, pseudocolor plot: no-antibiotic, paired control. Black contour: antibiotic-treated data. (A) Tetracycline at 1 μg/mL. (B) Gentamicin at 8 μg/mL, the MIC for Mu14S. (C) Penicillin G at 32 μg/mL, the resistant breakpoint for penicillin group for *E. coli*. The starting *E. coli* concentration was around 100, 30 and 40 CFU/mL for tetracycline, gentamicin, and ampicillin treated data respectively. All data were done in triplicate.

sQF calculation and distances are referenced to paired controls, there is no need to adjust threshold and/or gate when the scatter patterns change from run to run.

FAST with Clinical Isolate E. coli strain Mu14S

Around 10 CFU/mL of *E. coli* Mu14S was spiked into 10% human blood purchased from USBiological, and FAST was applied on the sample. As in Fig. 4.8, the test results show that the Mu14S is resistant to tetracycline and penicillin g (an antibiotic belonging to the same group as ampicillin. When treated with 8 μg/mL of gentamicin, however, growth inhibition showed as the cytometric scatter signal resemble the blood only data (Fig. 4.16 B). The PB-sQF results also indicated that gentamicin is the effective treatment.

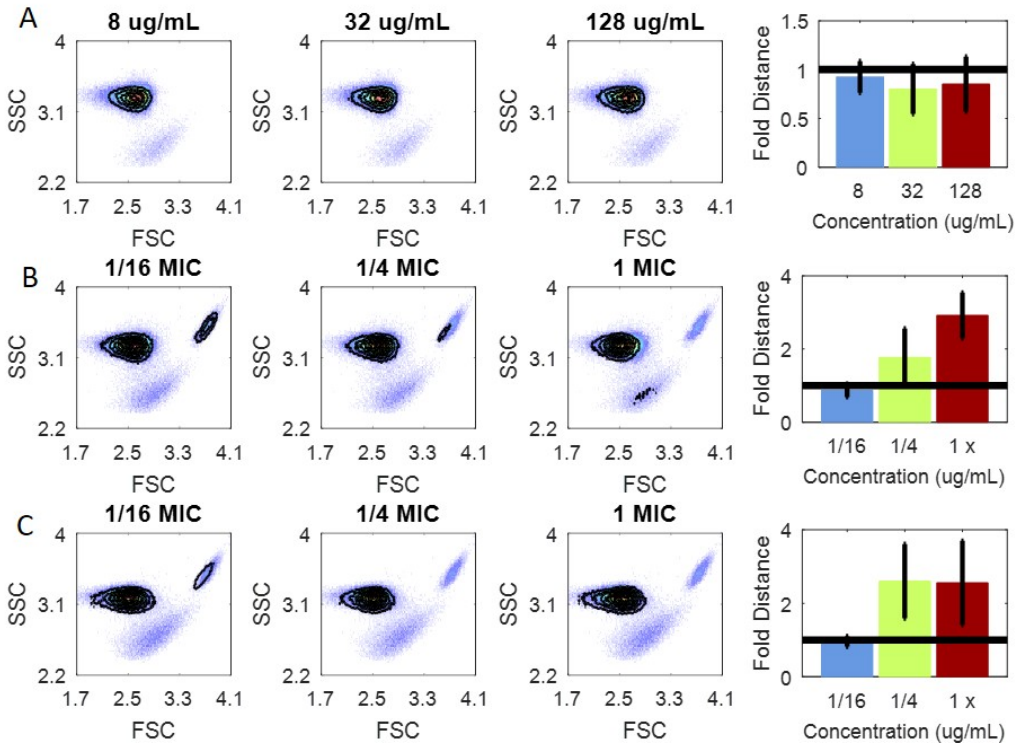

Figure 4.17: Flow cytometry data under different setting with 10% blood only or *E. coli* strain Mu890 in 10% human blood. (A) Ampicillin treated 10% human blood only sample at the *Acinetobacter* resistant breakpoint. (B)Tetracycline-treated data. 1x MIC is 2 μg/mL. (C) Gentamicin-treated data. 1xMIC is 8 μg/mL. All data were done in triplicate.

4.4.2 Different Flow Cytometric Settings

The scattered-light pattern is also influenced by the cytometer setting. Although the blood was purchased from the same vendor (ZenBio) and the data were taken by the same flow cytometer, the machine was often realigned under maintenance, causing the scatter pattern changes. The scatter signals of the blood-only sample treated with ampicillin from 1/16x to 1x of the resistance breakpoint of *Acinetobacter* (128 μg/mL) remained unchanged (Fig. 4.17 A), and so do the PB-sQF results. This demonstrated ampicillin at (128 μg/mL) still would not effect the scatter signals just as we have shown in Figure 4.7 A with the ampicillin at *Enterobacteriaceae* breakpoint (32 μg/mL) would not effect 10% human blood. This data suggests that the scattered-light pattern changes in Figure 4.10 is indeed the response from *Acinetobacter*.

As the scatter pattern of the blood debris was determined in Figure 4.17 A, the *E.coli* Mu890 scatter signal can be clearly identified in Figure 4.17 B and C. The relative position and shape of the blood debris and Mu890 signals were very different from Figure 4.8 A. The data nevertheless are consistent with each other. The growth inhibition from both tetracycline and gentamicin are clearly shown in Figure 4.17 B and C and PB-sQF data including triplicate errors indicated that both antibiotics are effective treatments. This data shows the advantage of PB-sQF over gate-based[202] or mean-based statistics[65, 66, 203]. By calculating distances of the whole data between control and antibiotic-treated samples, no adjustment is needed from one data set to another.

4.5 Conclusions

To rapidly determine appropriate treatments for gram-negative bacteria, we developed FAST to minimize time-to-result from initial blood draw. By selectively removing blood cells, FAST requires only a total of 8 hours to complete susceptibility testing. Using flow cytometry to acquire the entire distribution of bacterial responses to antibiotic exposure, PB-sQF statistical metrics directly quantify the differences between antibiotic-treated data and no-antibiotic paired controls. Consistent results are obtained even when data vary among different replicates, or if performed on different instruments. This procedure, without time-consuming overnight incubation and serial plating, reduces the time-to-result from $>$60 hours to $<$8 hours total time from initial blood draw, with identical susceptibility determinations. Since rapid identification of the correct antibiotic treatment is crucial in treating bacterial infections, FAST has the potential to greatly improve patient outcomes, while minimizing antibiotic resistance proliferation. Since CLSI breakpoints for the most common bacteremia-causing bacteria differ by $<$4-fold, adding an additional antibiotic concentration could rapidly provide susceptibility information without waiting for much slower, post blood culture bacterial identity determinations. As the majority of blood stream infections are caused by gram negative bacteria,[204–206] this approach offers a path to

drastically improved patient outcomes, while also allowing for subsequent confirmation from much slower post blood culture ASTs and identifications of both gram-positive and gram-negative pathogens.

To adapt the FAST procedure for gram-positive bacteria, it is important to decrease the doubling time of *S. aureus*. The doubling time might be able to reduced when the specific nutritious requirements for *S. aureus* and *Streptococcus* are met. Possible medium including the lysed blood supplement CAMHB and the Tryptic Soy broth.

CHAPTER 5
BACTERIAL GENOME SEQUENCE TYPING

5.1 Introduction

The previous chapters developed and applied PB-sQF on analyzing multidimensional flow cytometry data. Without making any assumption about the distributions of data sets, the general procedure of PB-sQF can be used to analyze different histogram-type data. Genomic information, which is a string data constitute of A, C, G, or T, can be viewed as a long string composed of recurring subsequences. These recurring substrings form a histogram, and PB-sQF can then be used to calculate the distance between sequences, thereby characterizing the genome similarity from one species to another.

In this chapter, a bacterial sequence library, including 628 genomes, was constructed. The library strains were then clustered based on genome similarity, and a phylogenetic tree was built. As a clear relation was observed between the genome similarity, and the strains or genuses of the bacteria, PB-sQF was applied to typed "unknown" bacterial species by calculating the distance between the "unknown" bacteria and the library strains. The bacterial typing can be done with both the assembled genome sequences and the pooled raw short reads data from the sequencer. Different PB-sQF schemes were investigated with the pooled short reads data. Ultimately, PB-sQF was applied to outbreak analysis to identify the outbreak strain.

5.2 PB-sQF in Analyzing K-mer Frequency

PB-sQF is a multi-dimensional statistic that quantifies the (dis)similarity of any two distributions.[48] To turn genome sequences into probability distributions, short sequence reads or, if available, the complete bacterial sequences were first k-merized by KAnalyze.[147]

The unique k-mer letter sequences were then digitized and transformed into k-dimensional coordinates. The k-dimensional data points, including the coordinates and k-mer counts, were treated as k-dimensional probability distributions and binned adaptively to the pre-designated number of bins as in the flow cytometric data. Different binning schemes are investigated in Section 5.8. Since all binning schemes give identical results at saturated conditions (5.10 and 5.11), the digitized scheme described in Chapter 2 was chosen to analyze the data, as its fewer binning dimensions greatly reduces computational load. After binning, the data points in each bin were expanded into 4k-dimensional data points (four possible bases at each position), and the centroids are the average of the data points within each bin. The PB-sQF test statistics are calculated by matrix multiplication of the similarity matrix, made of the Euclidean distances between centroids and the weights. Details about PB-sQF for sequence analysis can be found in Chapter 2.

5.3 Bacterial Genus Grouping: Binary vs. Full Data

628 complete genomes of 578 bacteria from 216 different genuses (Appendix Table D.1) were k-merized into 3-mer, 6-mer, or 9-mer libraries. Occurrences, either as binary presence (binary data) or the actual normalized k-mer occurrence probabilities (full data) of each unique k-mer within each data set were further adaptively binned in k-dimensions. Pairwise PB-sQF distances were calculated between all possible k-mer bacterial genome probability distributions. The calculated test statistics generated a symmetric 628x628 matrix in which each element corresponds to the statistical linear distance between the row and column bacterial genomes. Bacterial genuses that were present $\geqq$10 times in the library were aligned relative to a control strain. As distance is always positive, the strain exhibiting the largest dynamic range (*Anaeromyxobacter dehalogenans* strain 2CP-C) was chosen as the control strain. When aligning all bacteria relative to the largest dynamic range strain, the control strain and the most different strain had the same distance from the middle strain (same distance, different directions). Since the distance is always positive

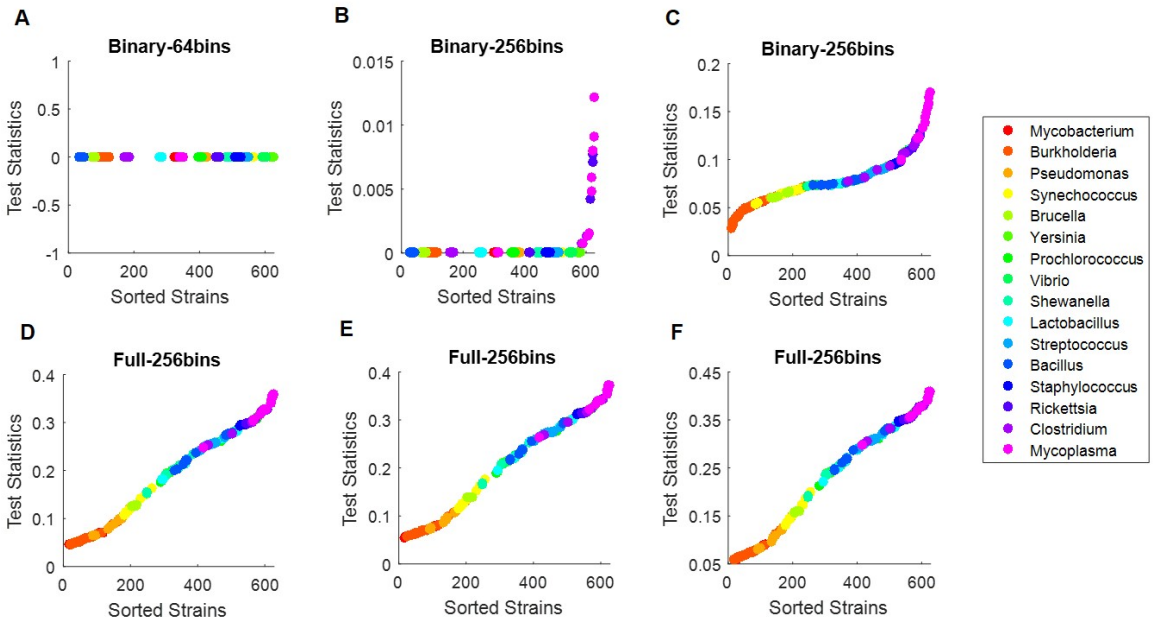

Figure 5.1: Test Statistics of selected bacteria lined up against *Anaeromyxobacter dehalogenans*. Top Row: Binary analysis of (A) 3-mers, (B) 6-mers, and (C) 9-mers. Only 9-mers showed distinguishability for binary analysis as shorter k-mers exhibit saturation. With 3-mers, no distance among library strains is observed, so bacteria are ordered alphabetically as in the library. Bottom Row: Full data of (D) 3-mers, (E) 6-mers, and (F) 9-mers. Independent of k-mer length, full data analysis yields nearly identical results with *Mycobacterium* being the closest to the control strain and mycoplasma being the most different strain.

and symmetric, aligning statistical distances from any other strain would lead to artificial overlapping distances of otherwise very different bacterial genomes (same distance, but different direction only distinguishable when properly aligned).

Aligning distances of all k-merized bacterial genomes (of a given k-mer length) relative to that from the control strain using binary k-mer presence alone, 3-mer and 6-mer test statistics show no or very little grouping ability (Fig. 5.1 A and B). This is because, with the size of bacteria genome ranging from 100 kbp to 10 Mbp, all the 3-mer (4^3 = 64 possible words) and most of the 6-mer (4^6 = 4096 possible words) were occupied among all of the bacteria in the library. Thus, there is no significant difference between the (saturated) probability distributions. One needs much higher dimensionality to separate genomes based on binary analysis of k-mer occurrences.[27] As a result, only 9-mer binary analysis (4^9

= 262,144 possible words) (Fig. 5.1 C) started to show a similar trend as in the full data (Fig. 5.1 F). Saturation issues become much more serious when comparing the longer genomes of more complex organisms, demanding that binary analyses utilize much higher dimensionalities.[207] Full data analysis, which considers the k-mer libraries as probability distributions with occurrences of each k-mer normalized to the total number of k-mer counts (thereby normalizing by genome size), enables much shorter k-mers to be used than that necessary with binary analysis. From 3-mer to 9-mer (Fig. 5.1, bottom row), bacteria from the same genus appear well-grouped, exhibiting similar dissimilarities from the control strain. The lower dimensionality afforded using shorter k-mer libraries, coupled with the excellent discrimination by treating genomes as probability distributions enables much faster and more direct comparisons with reduced computational demand. This becomes increasingly important as massive amounts of genomic data are rapidly generated and compared.

5.4 Bacterial Phylogenetic Tree

Since bacteria with common evolutionary histories are expected to have higher genome sequence similarity, the pairwise test statistics from PB-sQF can be used to build a phylogenetic tree.

5.4.1 Jaccard Index

To ensure the phylogenetic tree is built under the best PB-sQF performance, the clustering abilities of different test conditions were evaluated by Jaccard Index.[208] First, a hierarchical clustering algorithm was applied based on PB-sQF test distances using 3-mers, 6-mers, and 9-mers from 16 bins to 256 bins. Strains with paired-distances smaller than the threshold were clustered. The union of the strains satisfying the distance cutoff was taken. Thus, for a particular strain, the distances between this strain and all the other strains within the cluster might not all be smaller than the threshold. To group this strain to a particular clus-

ter, only one paired distance between this strain and a strain in the cluster had to be smaller than the threshold. To evaluate the clustering performance, a Jaccard index was calculated at different threshold values (Fig. 5.2).

As an external evaluation, the Jaccard index calculates the similarity between the hierarchical clustering and the (external) standard clustering, the latter of which was taken as bacteria clustered by their genus, as bacteria within the same genus are likely to share genomic similarity. True positives (TP) are when both the standard and PB-sQF-based clustering assign the two strains in the same cluster. False positives (FP) occur when the standard assigns the two strains to different clusters but our clustering method assigns them to the same cluster. False negatives (FN) occur when the standard assigns the two strains at the same cluster but PB-sQF clustering assigns them to different clusters. The Jaccard index utilizes the number of TP, FP, and FN occurrences and is defined as:

$$JaccardIndex = \frac{TP}{TP + FP + FN} \tag{5.1}$$

Independent of k-mer length, better clustering performance is achieved when more bins are used for both binary and full data (Fig. 5.2). Although k-mers longer than 9 bases are needed for discrimination in binary data, binary performance was not as good as the 3-mer and 6-mer analyses when full k-mer occurrence is used (Fig. 5.2 C, D and E). This results from the possible unique k-mers growing as 4^k, making 256 bins insufficient to capture the signatures of the 9-mer data. Indeed, for both 3-mer and 6-mer full data analyses, the performance did not significantly improve beyond 64 bins. A steadily improving Jaccard index is, however, observed in 9-mer analyses with more bins applied. For 3-mer libraries, the test performance becomes saturated at 64 bins, yielding identical results for 64, 128, and 256 bins (not shown), as only $4^3 = 64$ possible 3-mer sequences exist. Binary data using 3-mers (Fig. 5.2 A to C) showed no clustering and a Jaccard index equal to zero, while 9-mer binary analysis exhibited only limited grouping ability compared to the full

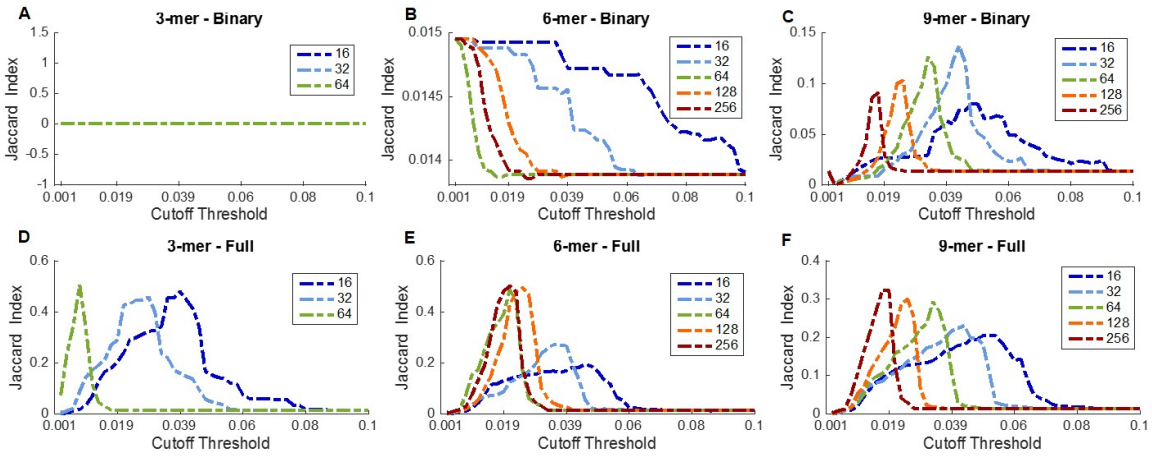

Figure 5.2: Hierarchical clustering results. The Jaccard index for different cutoff thresholds using (A) 3-mers, (B) 6-mers and (C) 9-mers. Independent of k-mer length, better clustering performance is achieved when more bins are used.

data results

A higher Jaccard index indicates greater similarity of the two clustering methods, meaning that it should be maximized at the best clustering conditions. The Jaccard Indices for different k-mers and bins are shown in Figure 5.2. Among the tested conditions, 3-mer, 64 bins with a test statistics cutoff threshold 0.0071 gave the highest Jaccard index and, therefore, the best clustering performance.

5.4.2 Phylogenic Tree

The pre-calculated 628x628 test statistics matrix of 3-mer 64 bins was loaded. To focus on the genome sequence similarity between strains, only the 1st chromosome was included for bacteria that have multiple chromosomes. The test statistics threshold was set at 0.0071 as it gives the best clustering result as shown in Subsection 5.4.1 and applied to include only the strains that have pairwise distances smaller or equal to the threshold to ensure good clustering performance. A total of 330 of strains were included, and the phylogenetic tree was generated by the in-built MATLAB function seqlinkage using the average distance method. The output was written into Newick format and loaded to iTOL (`http://itol.embl.de/`) to generate the figures.

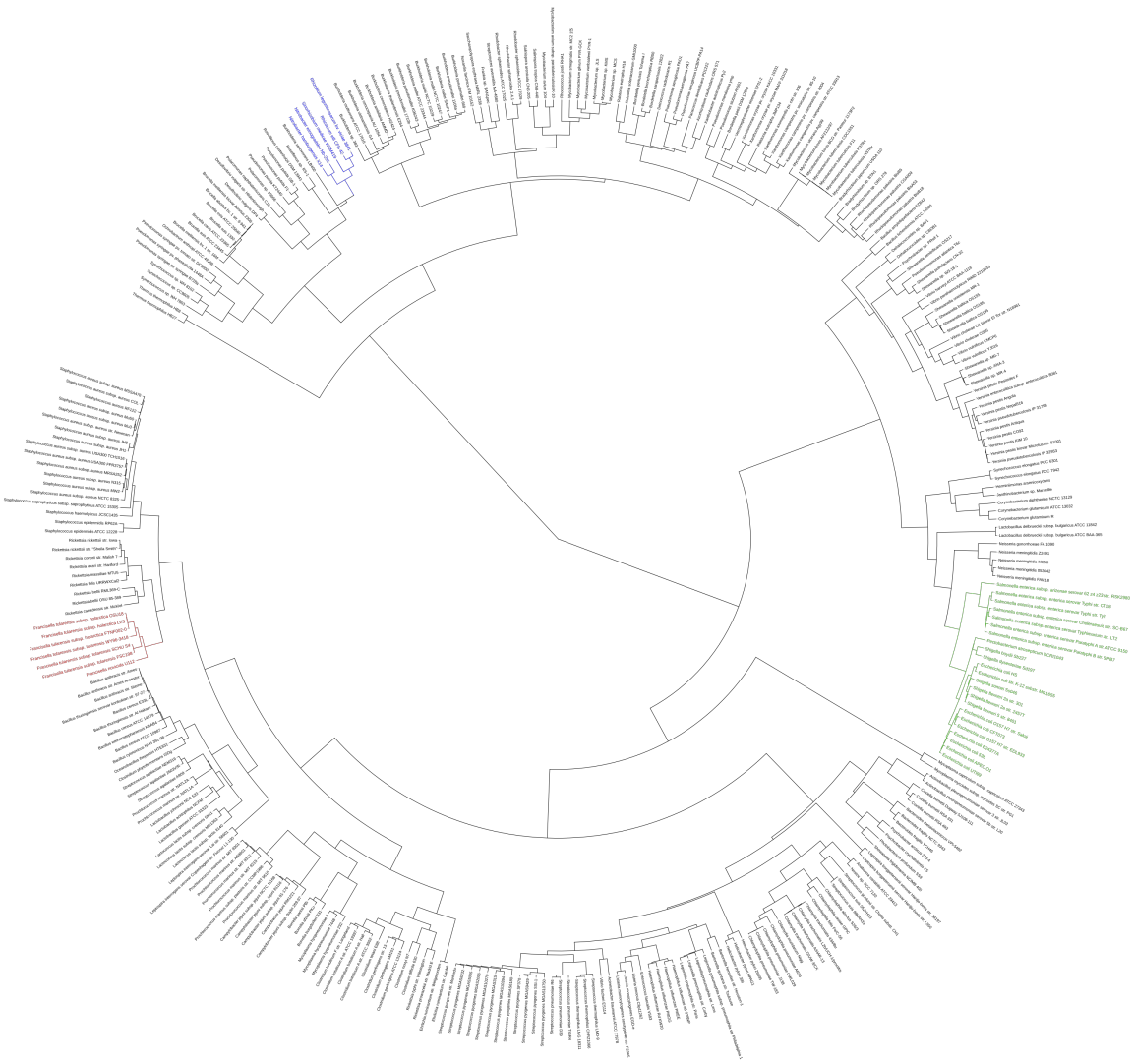

Figure 5.3: Bacterial phylogenetic Tree. Pairwise test statistics from 3-mer, 64 bins were used to build this phylogenetic tree. The red branch is the Francisella branch. The blue branch is one of the Rhizobiales order branch. And the green branch is the Enterobacteriaceae family branch. The taxomony was from the NCBI database. The tree was built in MATLAB and plotted in iTOL (`http://itol.embl.de/`).

Starting from the leaf nodes, bacteria from the same subspecies or strains, sharing higher genome similarity, tend to join together to form a node before further joining with bacteria from the same species (but different subspecies or strains). For example, *Francisella tularensis* has 2 subspecies: tularensis and holarctica. The 3 holarctica strains join together before forming a new node with the tularensis strains. The *Francisella tularensis*

branch is seen to connect with *Francisella novicida* U112 and form the *Francisella*-genus branch (Fig. 5.3, red). This shows that PB-sQF not only groups bacteria by genus (Fig. 5.1) but also identifies bacteria at the species and subspecies level. Further toward the root, bacteria sharing common higher taxonomic rank join together in a new node. For example, (Fig. 5.3, blue) *Rhizobium etli* CFN 42 and *Rhizobium leguminosarum* bv. *viciae* 3841 first merge into a genus-level node. Then, at larger distances, combine with *Sinorhizobium medicae* WSM419 to form the *Rhizobiaceae* family branch. This *Rhizobiaceae* family branch further merges with the *Nitrobacter*-branch since they are both belong to the *Rhizobiales* order. The same process can be observed for *Salmonella*, *Escherichia*, and *Shiegella* (Fig. 5.3, green). With bacteria first grouped together within their own strain/subspecies, species, and genus, the 3 branches eventually merge since they are all from the *Enterobacteriaceae* family. This shows that the different levels in the PB-sQF distance-based phylogenetic tree reflect the taxonomic rank of the bacteria, and the evolutionary history can be directly evaluated by whole genome PB-sQF

5.5 Bacterial Assembled Sequence Typing

The ability of PB-sQF to type unknown sequences to the correct library strains was evaluated for a total of 197 assembled genome sequences downloaded from the National Center for Biotechnology Information (NCBI) ftp site (`ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/`), covering 42 different genuses. While none of the "unknown" sequences was an identical match to those in our library (no repeated accession numbers), each of the unknown sequences belongs to one of the 216 different genuses in the 628 library strains. Of the 197 unknown sequences (Appendix Table D.2), 162 were of the same species (but different subspecies) as a genome in the library.

Using KAnalyze and PB-sQF, pairwise test statistics were calculated between each of the k-merized 162 unknowns that have corresponding library species and all of the library genomes. Unknowns were typed as being the same as the library species that yielded the

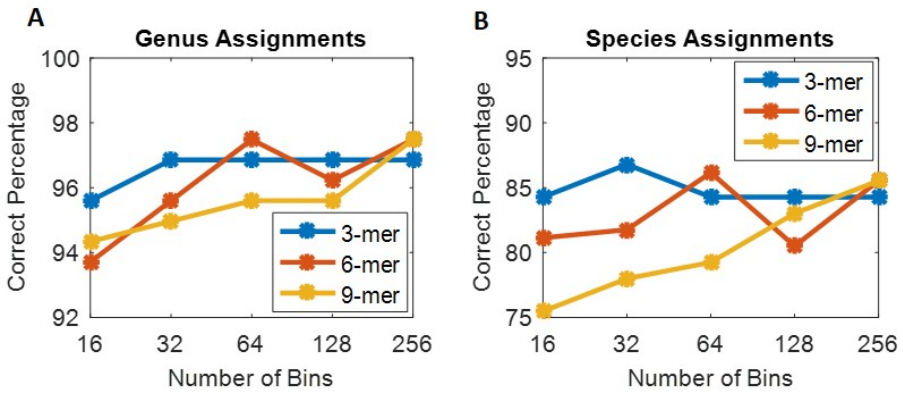

Figure 5.4: Assembled sequence typing. Percent correct assignments for 162 unknowns that have a corresponding library species. Percent correct of (A) genus assignments and (B) species assignments.

smallest test statistics value. As with clustering, 3-mers gave the best typing performance using the fewest bins, while both 6-mer and 9-mer analyses only improved at high bin numbers (Fig. 5.4 A and B). The species-typing accuracy was ∼85% and the genus-typing accuracy reached 97% under the conditions tested.

Since the test statistic calculated from PB-sQF is a statistical metric, the test value can be directly compared regardless of which library strains were used. As a result, PB-sQF can minimize false positive assignments by determining the confidence of typing and rejecting low confidence assignments. The test statistics from correct and incorrect assignments were sorted and, as shown in Figure 5.5 A and Appendix Figure D.1, the incorrect assignments tended to have larger test statistics (more dissimilar) compared to the correct assignments.

To include most of the correct assignments while rejecting a reasonable fraction of incorrect assignments, a test statistics threshold was set to include 95% of the correct assignments. By only accepting assignments with test statistics lower than this 95% threshold, the typing accuracy among the valid assignments increased from 85% to 92% for species-typing (Fig. 5.5 C) and from 97% to nearly 100% for the genus-typing (Fig. 5.5 B). The unassigned rates (percentage of strains having a test statistic exceeding the 95% threshold) are shown in Figure 5.5 D.

Test statistics between the library and all 197 unknowns, including the 35 strains that do

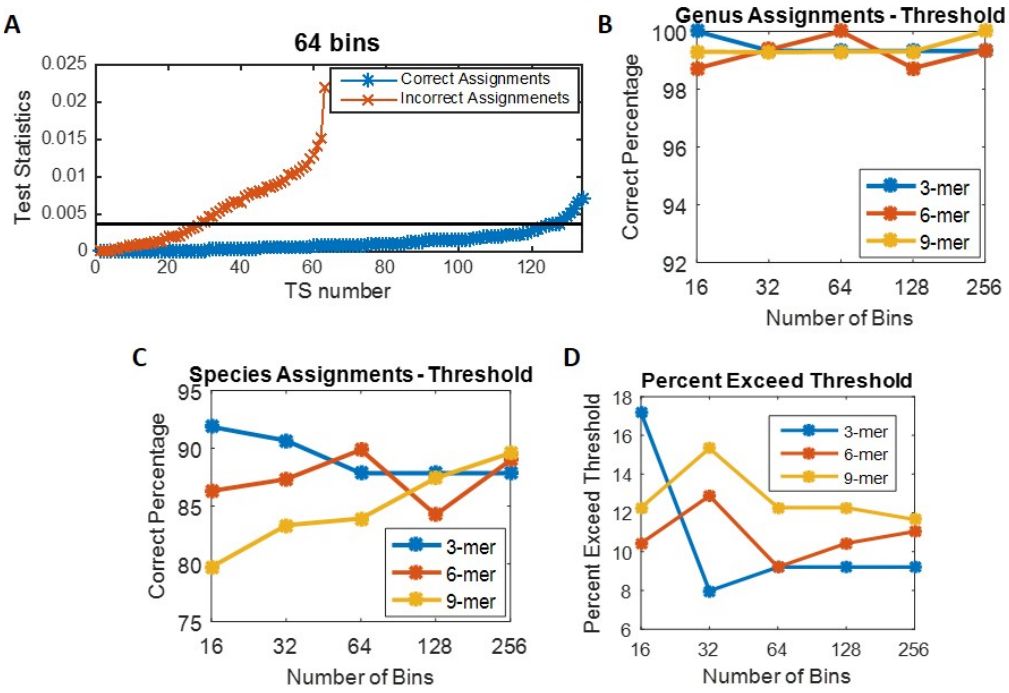

Figure 5.5: Assembled sequence typing with threshold. (A) Sorted test statistics (3-mer, 64 bins) from assembled bacterial sequences. The x-axis is the index of the sorted test statistics. The orange curve is the false assignments, and the blue curve is the correct assignments. The black line is the threshold determined from the 95% test statistics of the correct assignments. Percent correct assignments of those meeting the confidence level in genus identification (i.e. after applying empirical threshold) of (B) genus assignments and (C) species assignments, (D) Percent of unknowns that had test statistics exceeding the empirical threshold.

not have a matching species in the library were also calculated (Fig. 5.6). As expected, the species-typing accuracy was lower since there was no library species to which it could be typed. The genus-typing accuracy, on the other hand, was still as high as 92% and increased to nearly 100% after the test statistics threshold was applied. Note that the performance of the 3-mer analysis saturates at 64 bins, enabling much faster calculations due to reduced dimensionality (Fig. 5.6 C). Our results show that PB-sQF successfully screens out the low confidence assignments and increases the typing accuracy.

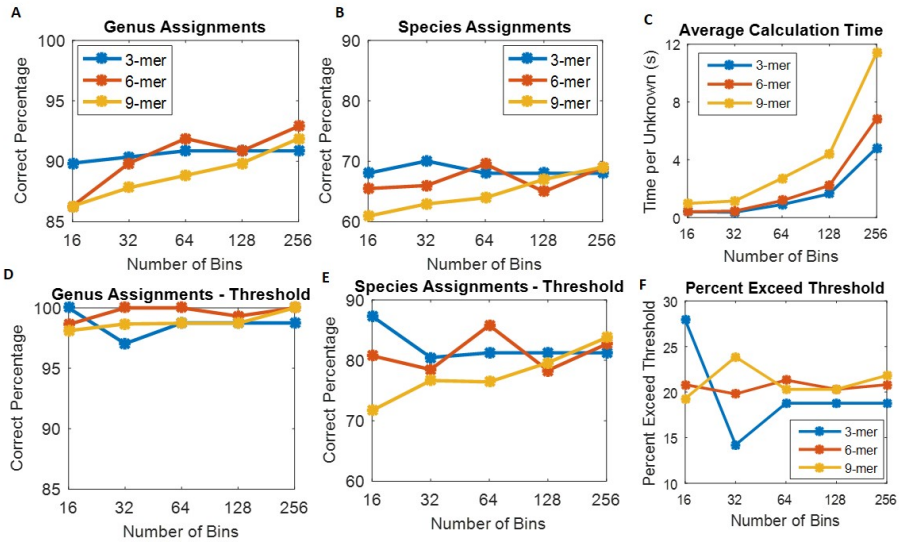

Figure 5.6: Assembled sequence typing All 197 strains. Percent correct assignments for 197 unknowns. Percent correct of (A) genus assignments and (B) species assignments. Percent correct after applied empirical threshold of (D) genus assignments and (E) species assignments, (F) Percent unknowns that had test statistics exceeding the empirical threshold, and are therefore classified as being "unassigned". (C) Averaged calculation time for binning and PB-sQF analysis for each unknown.

5.6 Bacterial Typing with Pooled Short Reads Data

Instead of completed sequences, the raw NGS data contains million of short reads that contain part of the genetic information. To apply PB-sQF on NGS output, raw sequencing data files were directly analyzed. 376 short reads accessions, covering 135 strains and 6 different sequencing platforms with both paired-end and single-end reads, were downloaded from the sequence read archive (`http://www.ncbi.nlm.nih.gov/sra`). The full list is in Appendix Table D.3. The raw short reads data from each accession were k-merized, and the k-mers from all reads were pooled together, regardless of the read direction. No corrections to or de novo sequencing of the raw data files was performed. PB-sQF was then applied, and the unknown short reads files were typed to the k-merized whole genome species as the library member that gave the smallest test statistic (distance).

The typing accuracy appears to depend on instrument error rate. Illumina sequencers have the smallest error rate ($\sim$0.1%), making both the genus and species typing accuracy

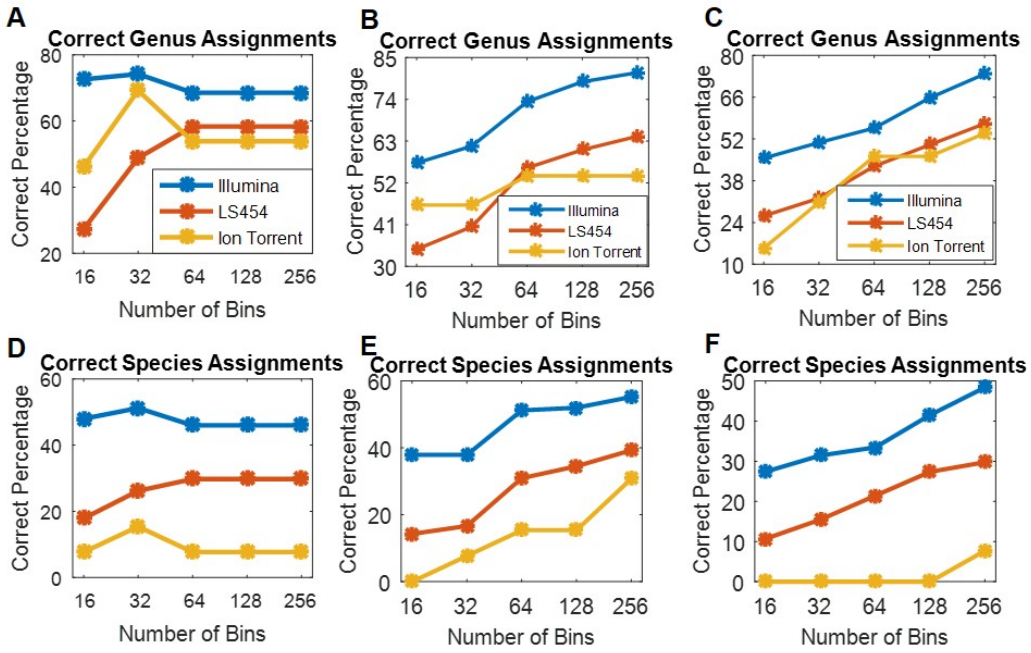

Figure 5.7: Pooled Sequence Typing for Illumina, LS454 and Ion Torrent (raw reads data files without threshold). Percent correct genus assignments of (A) 3-mers, (B) 6-mers, and (C) 9-mers. PB-sQF analyses of raw reads files compared to reconstructed whole genome libraries. Percent correct species assignments using (D) 3-mers, (E) 6-mers, and (F) 9-mers. The lower the error rate of a sequencer, the higher the typing accuracy. The legend in (E) is the same as in the other panels.

better than either LS454 or Ion Torrent (error rates $\sim$1%) data. With 6-mers and 256 bins, the genus typing accuracy reached 81%, and species typing accuracy was 55% with Illumina. Again, better accuracy was achieved with higher number of bins for both 6-mer and 9-mer (Fig. 5.7 B and C) analyses. Note that for 3-mer analyses, the test performance becomes saturated at 64 bins, yielding identical results for 64, 128, and 256 bins (not shown), as only $4^3 = 64$ possible 3-mer sequences exist.

As in the assembled sequence typing, an empirical threshold could be established by comparing the test statistics between the incorrect and correct assignments. Since the number of data from the 6 sequencing platforms was unevenly distributed with 246 datasets from Illumina, 84 from Roche LS454, 13 from Ion Torrent, 29 from PacBio, 3 from AB Solid and 1 from Oxford Nanopore, only the thresholds of Illumina and Roche LS454 were custom-determined with 95% and 90% thresholds respectively. An overall 95% threshold

determined from all 376 datasets was applied to the rest of the sequencers (Appendix Fig D.2 to D.4). By rejecting assignments with low confidence levels, typing accuracy consistently increased across different sequencers (Fig. 5.8) compared to without thresholding (Fig. 5.7). For example, with 6-mer, 256 bins, the genus-typing accuracy for Illumina increased from 81% to 88% and the species typing accuracy increased from 55% to 61% among the valid assignments (Fig. 5.7 and 5.8). Raw reads files from machines with higher error rates were less confidently typed, resulting in higher unassigned rates and a greater accuracy improvement when thresholds were applied. Indeed, compared to the data collected from Illumina, which has a reported error rate of 0.1%, the genus-typing accuracy increased upon thresholding from 64% to 84% and 54% to 75% for data collected from LS454 and Ion Torrent, instruments with modest (1%) error rates, respectively (6-mer, 256 bins in Fig. 5.7 and 5.8).[209, 210] The typing accuracy for data collected by PacBio, AB Solid and Oxford Nanopore, with the reported error rates are 16% (single pass), 0.06% (double- or triple-encoding) and 38.2% respectively,[209–211] can be found in Appendix Figure D.5. To address the accuracy properly, more short reads file from these 3 sequencers were downloaded, and the full list can be found in Appendix Table D.4.

Since PB-sQF calculates the dissimilarity between every library member and each unknown strain, instead of answering whether the unknown belongs to a certain strain, useful information can be drawn from the "incorrect" assignments. Indeed, all eight of the *Shigella* strains from the Illumina-generated data were typed as *Escherichia*. These assignments were classified as incorrect typing but *Shigella* species are sublineages of *Escherichia* so they are indeed closely related to each other.[212] Also, three out of four *Citrobacter* were typed as *Salmonella* - two species that have been shown to share high genetic similarity.[213] This demonstrates that PB-sQF can type the unknown to its most similar strain even if the unknown has never been discovered before. The accuracy of PB-sQF can potentially be improved by taking the 2^{nd} and/or 3^{rd} most similar library strains (or even all the test results) into account. An unknown is more likely to belong to a certain genus if

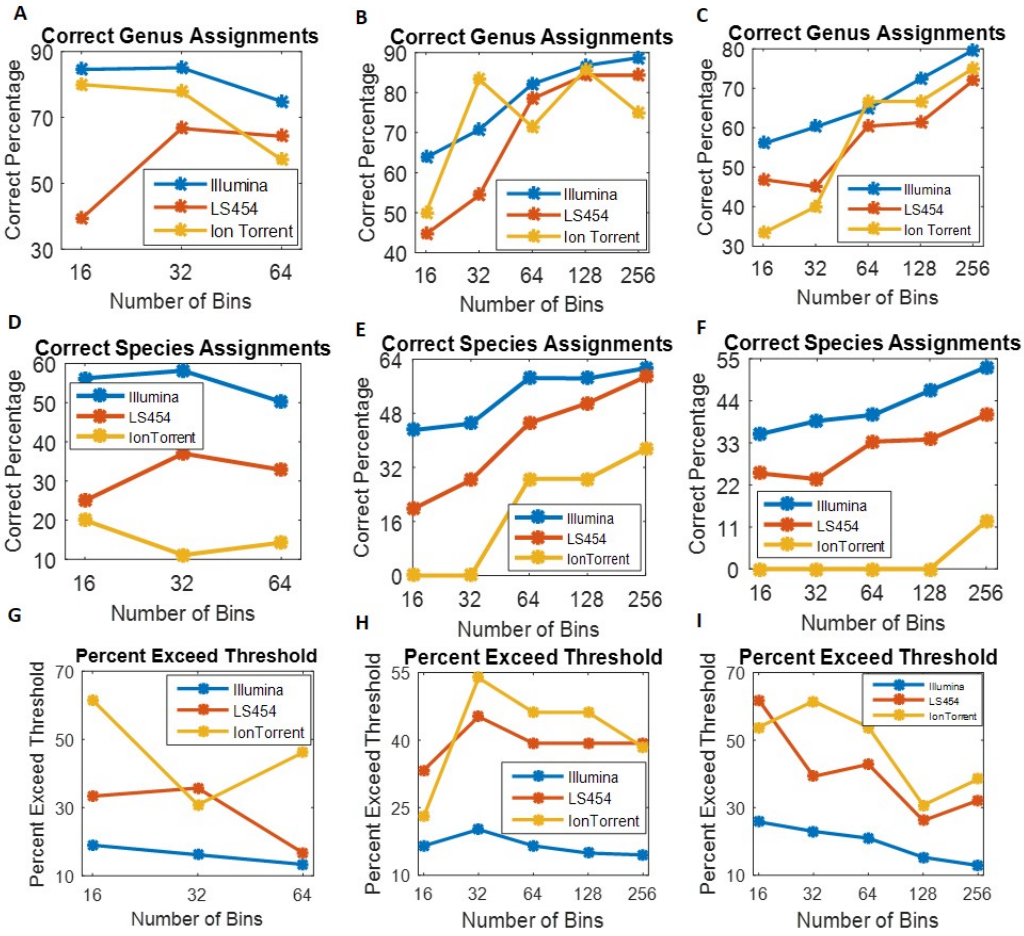

Figure 5.8: Pooled Sequence Typing for Illumina, LS454 and Ion Torrent data (threshold applied). Percent correct genus assignments of (A) 3-mer, (B) 6-mer, and (C) 9-mer full data analyses. PB-sQF analyses of raw reads files compared to reconstructed whole genome libraries. Percent correct species assignments using (D) 3-mers, (E) 6-mers, and (F) 9-mers. The lower the error rate of a sequencer, the higher the typing accuracy. Percent of datasets that had test statistics exceeding the threshold for (G) 3-mer, (H) 6-mer, and (I) 9-mer.

the top five matches all point to the same genus than if only the closest match assigns it to the genus. Also, the larger the difference between the test statistics of the 1^{st} and 2^{nd} most similar strain, the higher the possibility that the 1^{st} strain is the correct assignment. These issues will guide library construction and further understanding of the relation between the correct assignments and the test statistics values.

5.7 MRSA outbreak analysis with PB-sQF

In outbreak analysis, it is crucial to distinguish similarities and relationships among pathogens causing infections. As demonstrated by recent studies,[26, 214, 215] phylogenetic relationships among outbreak and non-outbreak strains can be deduced by base-to-base comparisons of assembled genomes, but require several steps of data preconditioning and subsequent analysis. In these studies, short reads were first mapped onto a reference strain and then variants, particularly single nucleotide polymorphism (SNP), were called. For both steps, filtering processes are needed to exclude the low score alignments and/or SNP detections. MLST studies were used to help identify strains based on similarity to seven housekeeping genes, but similarities in the rest of the genome may contain important information as new virulent strains emerge.

To avoid any possible biases from multiple data filtering, genome reconstruction, SNP detection, and MLST typing, PB-sQF kmer distance was used to analyze the published outbreak data. Fifty-seven methicillin-resistant Staphylococcus aureus (MRSA) raw short reads files that were processed and analyzed by Harris et al.[214] in a hospital-acquired outbreak were downloaded from the sequence reads archive (Appendix Table D.5 Table). Without any pre-processing, statistical distances were calculated between all k-merized raw, pooled short reads data files and a phylogenetic tree was built directly from the pairwise distances between all MRSA strains (Fig. 5.9). Using the k-merized information from the entire genomes, intergenome distances using k-merized raw reads files reveal no clear inter-patient relation as reported by Harris et al. Instead of transmitting from one patient to another, it was suspected that multiple infections resulted from an external source (the hospital).[214] As shown in Figure 5.9, most MRSA isolates from patients (P1 to P26) group separately from the isolates traced to the hospital personnel (green dots), and that the patient MRSA sequences do tend to group together. In contrast to the published findings, using the raw reads files suggests that P26, P19 and P21 may be linked to the hospital worker.

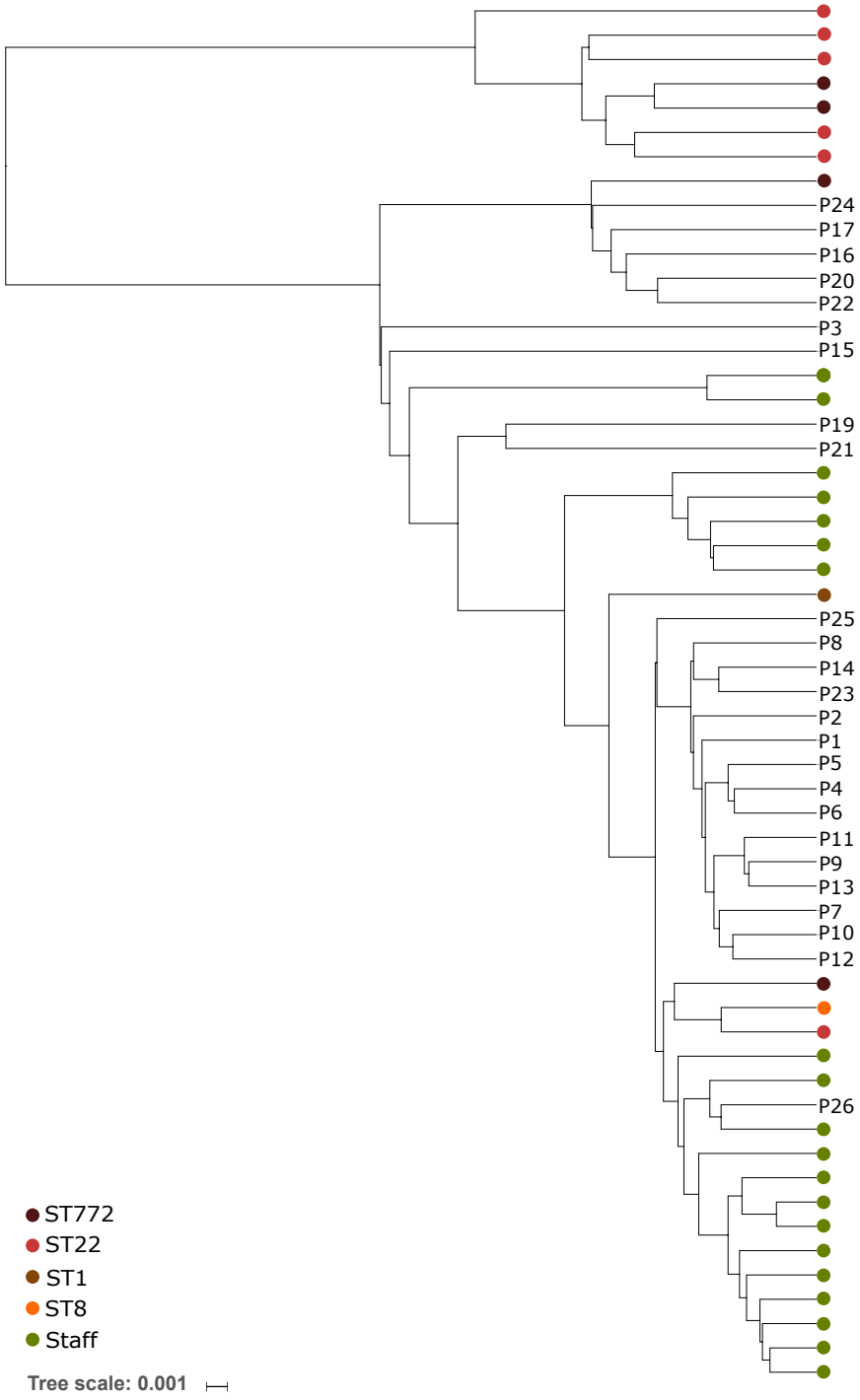

Figure 5.9: Phylogenetic tree of MRSA ST2371 outbreak. P1 to P26 represents the 1^{st} to 26^{th} patients with MRSA ST2371 (the outbreak strain) infection. Different colors represent different MRSA strains determined by MLST. All the data except the green dots, which are MRSA colonies (ST2371) collected from a hospital health care personnel, were sequenced from infected patients.

Although the groupings differ somewhat from those identified by Harris, et al., our PB-sQF approach with raw reads files enables fast screening by calculating the distances between pooled short reads data without individual read mapping, assembling, or SNPs calling, As with our bacterial genome clustering above, this interpretation would become even clearer as the entire assembled bacterial genomes became available for distance analysis.

For the non-outbreak strains determined by MLST by Harris et al.,[214] a distinct branch containing five ST22 and two ST772 did not appear to be related to the rest of the strains, other than to the root. By constructing pair-distance-based phylogenetic trees, PB-sQF can successfully exclude these seven isolates from the outbreak. By using the entire genome information instead of seven MLST housekeeping genes, however, five out of twelve "non-outbreak" isolate, have a closer relation with the outbreak strain. Because PB-sQF whole genome analyses examine regions both within and outside those in MLST analysis in an unbiased fashion, additional mutations and similarities giving rise to infectivity may be identified as being crucial to understanding the outbreak transmission path and guide effective treatment. This blind approach may have advantages if regions outside the pre-identified sequence areas are important in determining phenotype.

5.8 Alternative PB-sQF Modifications for Genome Sequence Analysis

In the first step of genome sequence analysis, the k-merized string data made from A, C, G, and T were first represented by integer values 1, 2, 3, and 4, respectively. These k-dimensional coordinates were then adaptively binned, dimensions expanded, and centroids and weights were calculated from the binning patterns. Since the binning procedure is performed prior to dimension expansion, A-T (1-4) rich dimensions, however, are prone to be divided first (larger variance). To investigate whether this significantly changes PB-sQF test results, the pooled short reads data were typed using multiple different permutations of the integer values assigned to A, C, G, and T, and with dimension expansion prior to binning.

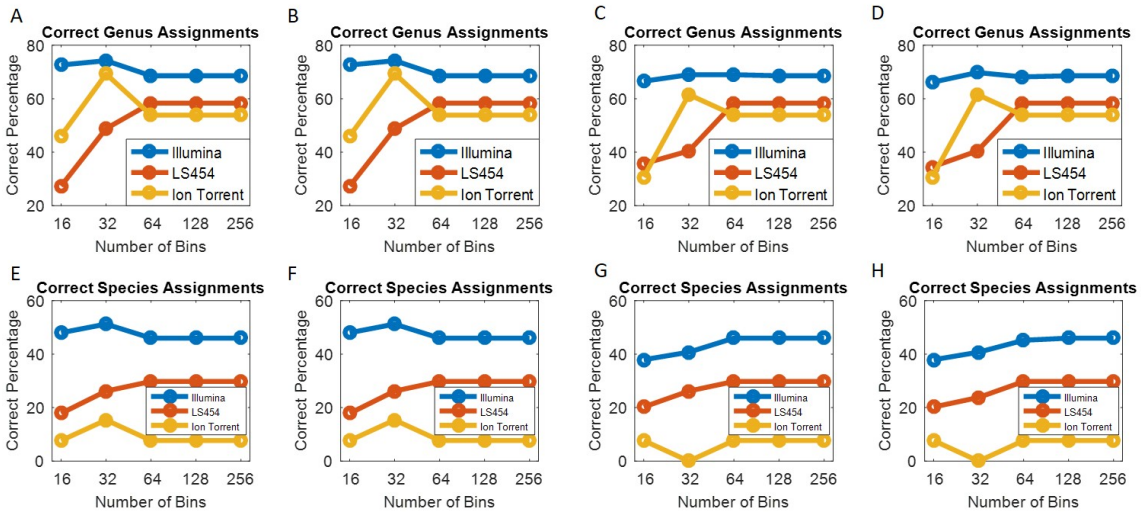

Figure 5.10: Pooled sequence typing for Illumina, LS454 and Ion Torrent (raw reads data files without threshold) with different digitized schemes. Percent correct genus assignments of 3-mers using (A) A-scheme, (B) G-scheme, (C) C-scheme, and (D) T-scheme. Percent correct species assignments of 3-mer using (E) A-scheme, (F) G-scheme (G) C-scheme and (H) T-scheme.

5.8.1 Different Digitization Schemes

To further gauge whether the k-dimensional binning process was influenced by the integer values assigned to A, C, G, and T, different numbers than used in the original scheme: 1, 2, 3, and 4 were investigated. The first is the A-scheme, which is the reverse order of the original scheme: A $\rightarrow$ 4, C $\rightarrow$ 3, G $\rightarrow$ 2, and T $\rightarrow$ 1, with A being the highest number. Reordering changes the absolute number of each nucleotide but the distance between each nucleotide remains the same. With this reverse order, all other permutations were tested. This includes the G-scheme, A $\rightarrow$ 2, C $\rightarrow$ 1, G $\rightarrow$ 4, and T $\rightarrow$ 3, the C-scheme: A $\rightarrow$ 1, C $\rightarrow$ 4, G $\rightarrow$ 3, and T $\rightarrow$ 2; and the T-scheme: A $\rightarrow$ 3, C $\rightarrow$ 2, G $\rightarrow$ 1, and T $\rightarrow$ 4.

When performing PB-sQF on the pooled short reads data with A-scheme using 3-mer, the test results (Fig. 5.10 A and E) remain the same as the results calculated from the original-scheme (Fig. 5.7 A and D). This is as expected since in this scheme, the relative distance between each nucleotide stays the same. Also, as observed in original-scheme, the test results saturated at 64 bins with 128 and 256 bins having the same results as 64 bins.

When using the G-scheme, the relative distance between C and G becomes 3 instead of 1. As a result, any G-C rich dimension would be divided first. However, as shown in Figure 5.10 B and F, the test results were exactly the same as the A-scheme or the original-scheme. When using the C-scheme and T-scheme, the test results were slightly different from the other schemes for 16 and 32 bins. This is probably because in these 2 schemes, the largest variance occurs with A-C or G-T rich dimension instead of A-T or C-G as before. The saturated accuracies (from 64 to 256 bins), however, were identical, regardless of digitization scheme used. This demonstrates that by always diving the dimension with largest variance, PB-sQF adjusts for the bias in numbering the nucleotides and evenly divides the data into bins.

5.8.2 Early Dimension Expansion

To remove any bias in the digitization process, the same dimension expansion was performed as described in Chapter 2.4.1 but before the binning process. As a result, each nucleotide at each new position is represented by a vector of length 1 in each of the 4k independent (expanded) dimensions, instead of having the value of 1-4 in the original k dimensions. More bins are needed to analyze the higher dimension binning data. Different from Figures 5.7 A and D, the pre-bin dimension expanded correct typing percentages were much lower when fewer bins were used (Fig. 5.11). The accuracy steadily improved as the number of bins increased. The typing accuracy, instead of saturating at 64 bins, maxed out at 512 bins. This maximum accuracy is the same as the saturation accuracies for the original-scheme data (Figs 5.7 A and D). This shows that, although expanding the dimension can correctly account for the independent occurrence of and distance corresponding to each nucleotide, this approach is much less efficient as much more extensive calculations are needed due to the much larger number of bins. Upon sufficiently high number of bins, the 4k-dimensional binning and distance accuracy limit is the same as the much lower k-dimensional nucleotide binning methods. As a result, the k-dimensional nucleotide binning

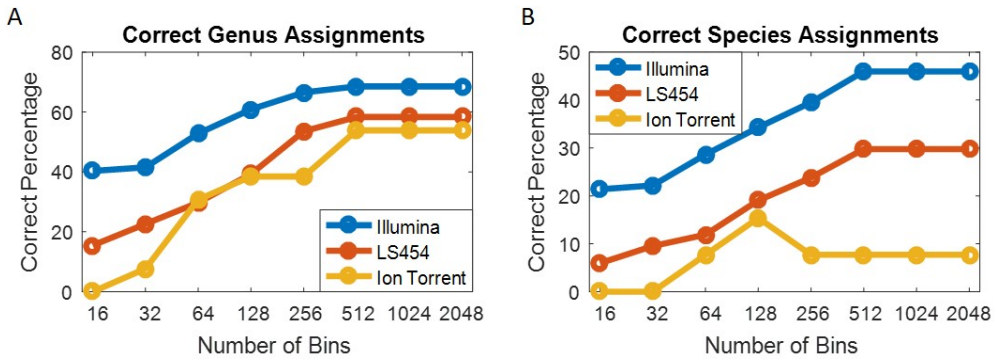

Figure 5.11: Pooled sequence typing for raw reads data files without threshold with early dimension expansion. (A) Percent correct genus assignments of 3-mers. (B) Percent correct species assignments of 3-mer.

is preferred.

Pre-binning dimension expansion for 3-mer data resulted in saturation requiring 512 bins, instead of 64 bins when dimension expansion was performed after binning. When performing dimension expansion after binning (i.e. A-T given integer values 1-4), $4^3 = (2^2)^3 = 64$ bins are required to account for all possible unique k-mers. In pre-binning dimension expansion, each mer in the original dimension is expanded into four possible positions. For each position, there are two possible outcomes, 0 or 1. However, since it is either A=(1,0,0,0), C=(0,1,0,0), G=(0,0,1,0) or T=(0,0,0,1), these four expanded positions are not independent but must consist of three 0 and one 1. As a result, knowing the outcomes of any of the 3 expanded positions guarantees to deduce the 4th. The total number of possible outcomes is thus 2^3 for each mer. Therefore, even though only four actual possibilities exist at each position, the dimension expansion prior to binning, increases the total number of bins needed. For 3-mer data, this results in $(2^3)^3 = 512$ total possibilities, thereby requiring 512 bins. While the saturated accuracies remain unchanged, many more bins are required ($(2^3)^3 = 512$ bins vs. $4^3 = (2^2)^3 = 64$ bins), resulting in longer calculation time. Moreover, when using higher k-mer, 9-mer for example, the number of bins will become unmanageable ($(2^3)^9 = 2^{27}$ bins vs. $(2^2)^9 = 2^{18}$ bins). Because the outcomes are indistinguishable, we analyze the data by assigning an integer value to each nucleotide

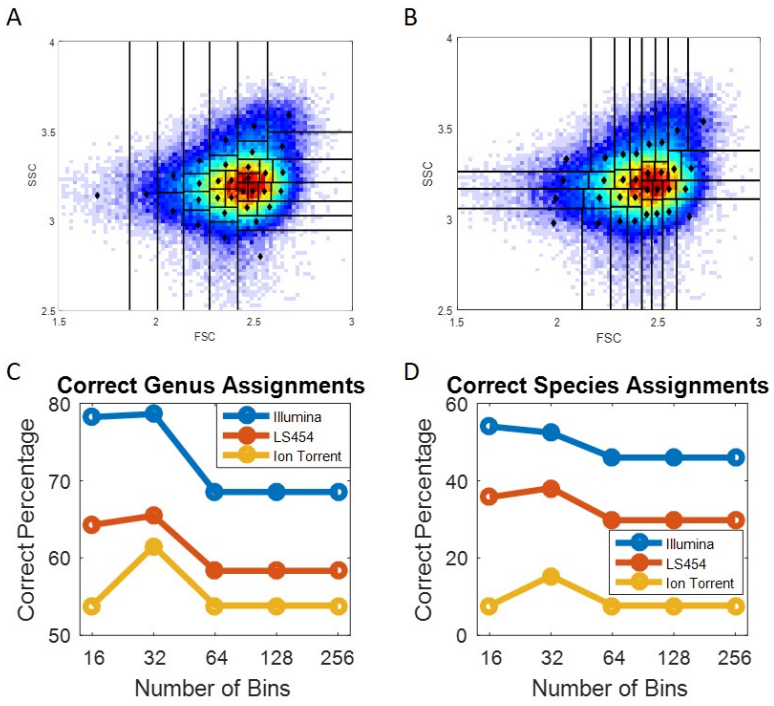

Figure 5.12: Demonstration of cycle-dimension PB-sQF Flow cytometry data binned with (A) largest variance dimension method or (B) cycle-dimension method. the dimension divided was cycled between the FSC and SSC domains. Percent correct of 3-mers pooled-short reads data of (C) genus assignments and (D) species assignments.

followed by dimension expansion for distance calculations.

5.8.3 Cycle-Dimension PB-sQF

Instead of dividing the data set at the largest variance dimension, one can partition the data set into each dimension by cycling through all the dimensions recursively.[216] That is: $1^{st} \rightarrow 2^{nd} \rightarrow 3^{rd} \rightarrow ... \rightarrow 1^{st} \rightarrow ...$, until the assigned number of bins is reached. All dimensions are thus equally divided. The binnning patterns of the largest variance dimension method and the cycle-dimension method were compared using the flow cytometry data (Fig. 5.12 A and B). It is s clear to see that the binning patterns resemble each other with the centroids located at the data area.

The cycle-dimension PB-sQF was applied to the pooled-short reads data from Illumina, LS454 and Ion Torrent. All the assignments between the two binning methods, except the

incorrect → correct assignments, stay the same as correct or incorrect assignments. This shows that only a small portion of data were behaved differently under the two binning methods. The accuracies, when a fewer number of bins were used, are in general higher than the original-binning-method (Fig. 5.12 C and D). This is most likely because, with a fewer number of bins, the original-binning-method indeed bias the 1-4 difference than others. As a result, without enough of bins, the centroids are not a good representative of the original data and thus poorer test performance. With cycle-dimension, since it evenly rotates through all the dimensions regardless of the variance, it better represents the data set with fewer bins. The test results, however, also saturated at 64 bins and above with 3-mer data and the saturation accuracies are the same as the original binning-method (Fig. 5.7 A and D).

The higher accuracies that we have observed at fewer number of bins with cycle-dimension are not expected to be seen in the flow cytometry data. This is because the assignments that are incorrect in the original-binning-method but correct in the cycle-dimension PB-sQF are having a different first division dimension. The first division in the original-binning-method deviated from the library strain because of the errors in the reads influence the variance in each dimension. Since there are only four possible values in each dimension, the variance of each dimension is very similar to each other. Even a single data point changes (ex. from (1,3,3) to (3,3,1)) can change the largest variance dimension and thus change the binning pattern. The cytometry data, however, have much more value available, the highest variance dimension is less effected by single data point changes.

5.9 Conclusions

With the potential to apply to different genome analysis problems through its generality, PB-sQF readily enables direct comparisons among whole genomes to build true linear statistical distances separating them, without having to rely on comparing much shorter, limited information ribosomal sequences. This chapter has shown the potential of PB-sQF

for typing unknown sequences, and it may find application in outbreak analysis,[23, 25, 217] for example, by tracing pathogen origin and evolution. The independence of genome size, the high correlation between test statistics and correct assignments, the general approach toward different strains, the ease to expand libraries, the full use of whole-genomic information, and the construction of intrinsic confidence levels suggests that PB-sQF can be used to tackle a wide array of genome analysis challenges.

CHAPTER 6

ERROR TOLERANT SHORT READS MAPPING

6.1 Introduction

In Chapter 5, PB-sQF has been applied to bacterial genome sequence typing with both assembled sequences and pooled raw short reads data. Instead of analyzing whole genome sequences, short reads mapping, which assigns each read to a reference genome, has many important applications in the field of genomic from typing the unknown reads data to investigating expression levels. [35, 218]

In this Chapter, read-by-read typing was demonstrated with or without read errors. Then, the mapping accuracy was investigated. A 3-species mixed genome reads data was studied as a mini-metagenomic system. In the end, error-tolerant PB-sQF short reads mapping was compared with other methods.

6.2 Short Reads Mapping Overview

Short reads mapping was performed as described in Chapter 2.4. In brief, a read library is built by chopping the randomly selected library sequence into ReadLength-bp pieces. To increase the number of reads in the library and ensure a complete library, frame shifts, with a length depending on the given read length, are applied. Simulated reads were generated from the selected library sequence, which is called the "mother sequence". To perform search space reduction, the test statistics between the selected 50 control sequences and all other library reads are pre-calculated and saved. The pre-defined test statistics range, which is the TS_{preset} described in Chapter 2 to reduce the search space, was set at 0.05. As a result, for each test statistics calculation between the unknown reads and the control reads, only the library reads that have library-control distances within 0.05 from the unknown-

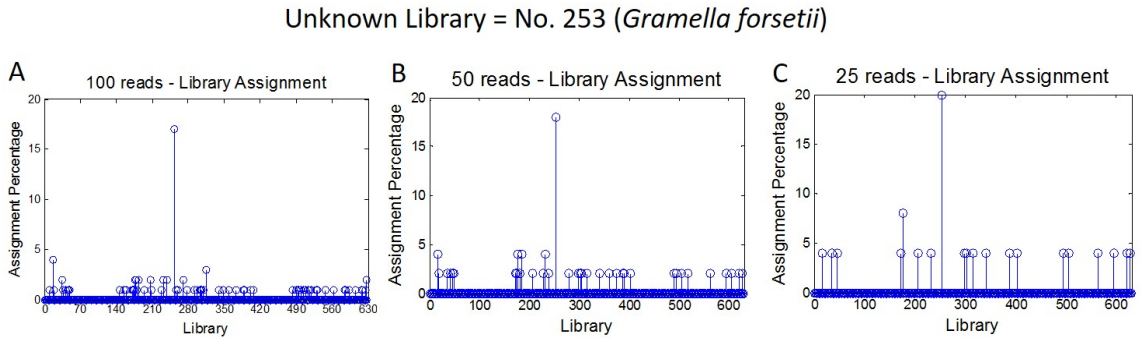

Figure 6.1: Read-by-read typing results with different number of reads. Percent reads mapped to each library when (A) 100, (B) 50, or (C) 25 reads were mapped.

control distance were kept. With 200-bp reads mapped to a bacterial genome which has $\sim 2Mbps$, the library reads size was reduced from $\sim 2 \times 10^5$ (10^4 per frame with 20 frame shift) to $\sim 10^3$ of library reads after 50 control iterations. The test statistics between the unknown reads and the 10^3 library reads were then calculated. The simulated reads are each mapped to the library read that gives the smallest test statistic value. Details of reads library construction, search space reduction, and simulated reads generation can be found in Chapter 2.4.

6.3 Read-by-Read typing

Instead of k-merizing and pre-binning the whole complete library sequence as in bacteria typing described in Chapter 5, bacterial genome typing can be done with read-by-read mapping. To demonstrate short reads typing, test statistics were calculated between all the randomly generated reads from the unknown bacterium and the 628 pre-binned library strains with no frame shift. Since the goal is to type the bacteria instead of correctly aligning all the short reads, the basic reads library without frame shift is applied to reduce calculation time. For each read, the library read that gave the smallest test statistics result was assigned as the mother sequence, that is the library sequence from which the simulated reads were generated from. Due to short reads sequence similarity and lack of frame shift, not every individual read could be assigned back to the correct mother sequence, even when the

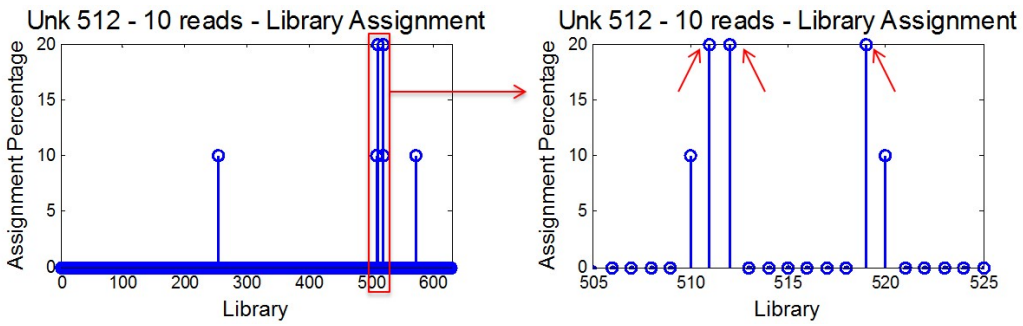

Figure 6.2: Reads-by-reads typing results for *S. aureus* strain MRSA252. Due to genome sequence similarity, several reads were mapped to different *S. aureus* strains.

simulated reads were generated without error. However, noise (incorrect assignments) was "randomly" distributed across all the 627 incorrect library strains. As a result, within each unknown, the most common mother strain assignment was then taken as the most likely true mother strain. For example, *Gramella forsetii* was randomly chosen as the mother strain and 100, 50, and 25 single-end reads of 200-bp long were randomly generated, respectively, from the complete sequence. Independent of the number of reads generated, $\sim$18% of the reads were correctly assigned as *Gramella forsetti*. Because mis-assignments were spread over much of the remaining library members (Figure 6.1 A to C), 25 short reads were more than sufficient to identify the correct mother strain out of all 628 possible candidates. Since the calculation time is linearly proportional to the number of reads, typing the unknown strain with only 25 reads instead of 100 reads reduces calculation time four fold. Real short reads data could have more than a million reads in a single run. Being able to identify the unknown strain with only 25 reads will greatly reduce the calculation time.

Because species within the 628 strains share genetic similarity, incorrect library assignments were not actually randomly distributed but were concentrated on species of the same genus. When using only ten reads, for example, this genetic similarity occasionally caused difficulty in species identification (Fig. 6.2). For example, when the randomly chosen "unknown" bacterium was the strain 512, *Staphylococcus aureus subsp. aureus* MRSA252, each of three different *Staphylococcus aureus* genomes were all assigned as the correct sequence for two out of the ten reads. To secure the correct library identification, a larger

Table 6.1: Read-by-read typing with different mother sequences Number of candidates

Mother Strain	# of Candidates	Incorrect Assignment
Staphylococcus aureus	3	—
Xanthomonas axonopodis	1	—
Streptococcus pyogenes	9	—
AYWB phytoplasma	1	—
Burkholderia sp.	2	—
Salinispora tropica	9	—
Chlamydophila pneumonia JI138	1	*C. pneumonia* TW183

library with finer frame shifts could be used, or more unknown reads should be processed so that the consensus library assignment would better stand out from the noise. Either option, however, would significantly increase calculation time. To solve this problem, finer frame shifts were used but only on the strains that were identified most often within Poisson counting noise. In this particular case, finer frame shifts were used just with the three different *Straphylococcus aureus* strains. The "true" library sequence was determined to be the genome with the most assignments among the selected candidate genomes. As shown in Table 6.1, out of seven runs, four randomly selected "unknown" strains also required this secondary selection to be applied. The only mis-assigned run: the reads generated from *Chlamydophila pneumoniae* strain J138 was assigned to *C. pneumonia* strain TW183, which belongs to the same species.

6.4 Short Reads Mapping with Read Errors

Single nucleotide polymorphisms (SNP) and insertions/deletions (indels) are common throughout bacterial genomes. Whether from genetic variations or instrumental errors, it is highly possible that the unknown reads would contain one or more SNPs and/or indels. Current short reads mappers, however, are built on algorithms for exact matches. To account for sequence mismatches, modifications are applied. For early hash table-based methods, like MAQ[29] and ZOOM[107], specially designed seed templates were used to ensure

that the mapping locations with limiting numbers of mismatches are found. For early BWT-based methods like Bowtie[39, 109] and BWA[40], different mismatch scenarios of a read were examined to find the mapping location(s) on the reference sequence. To map longer reads with more mismatches, most recent reads mappers, including Bowtie2,[109] BWA-MEM,[111] MOSAIK,[37] CUSHAW3,[219] GASSST,[220] and GEM,[221] take on the seed-and-extension approach regardless of their exact matching method (hash table or BWT). With the seed-and-extension approach, it is assumed that for most reads, short exact matches (seeds) can be found in the reference sequence. From these anchored seeds, dynamic programming such as the Smith-Waterman algorithm[108] is then used to extend the seeds with different constraints and filtering to prune the search space.

Instead of mapping the short read base-by-base, PB-sQF transforms the sequence information into signatures, allowing comparisons of statistical distances between probability distributions. Using the 200-bp reads as an example, since there are (200 - k + 1) k-mer counts out of a 200-bp read, for one SNP or one indel, it only alters k of the (200 k + 1) counts. As a result, PB-sQF is an innately mismatch (error) tolerant method, as only a very small percentage of k-mers are affected, leading to only minor variances in to the signature-based, adaptively binned library.

6.4.1 PB-sQF Mapping Accuracy with SNPs and Indels

To test the reliability of PB-sQF mapping in the presence of read errors, short reads with two SNPs or with both SNPs and indels were generated. The separation between any two SNPs or between one SNP and one indel was constrained to be fewer than 21 bases. This constraint of having two errors within 21 bases was chosen to mimic methods like MAQ[29] and Bowtie[39, 109], which only perform short reads mapping at the high-quality end of a read (28-mer). All 50 simulated short reads were taken from one of the *Streptococcus pneumoniae* strain. As expected, the performance was poor for 3-mer with binary data. However, with full counts data and PB-sQF distances, up to 98% mapping accuracy for one

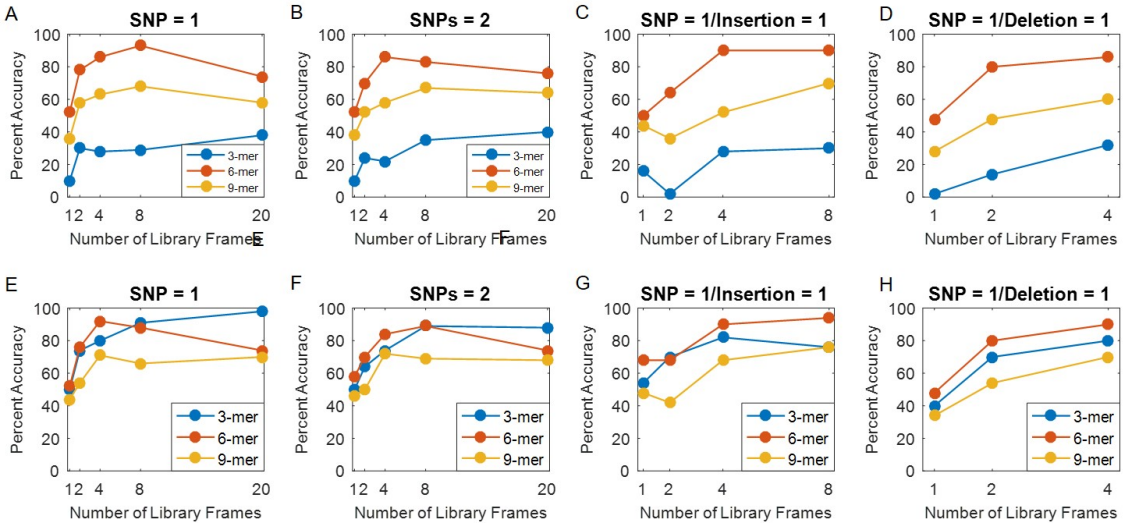

Figure 6.3: Mapping accuracy for short reads with assigned errors. Top Row: Binary analysis of 200-mer reads with (A) 1 SNP, (B) 2 SNPs, (C) 1 SNP with 1 insertion, and (D) 1 SNP and 2 deletions. Bottom Row: Full data of (E) 1 SNP, (F) 2 SNPs, (G) 1 SNP with 1 insertion, and (H) 1 SNP and 2 deletions. Since the read is 200-mer long, 1, 2, 4, 8, and 20 library frames represent 200-, 100-, 50-, 25-, and 10-mer frame shift of the reads library.

SNP and almost 90% accuracy is achieved for two SNPs or indels. Also, as indicated by the Jaccard analysis in Figure 5.2, 9-mer analysis yields poorer performance than do 3-mer and 6-mer libraries for analyzing the full data. For most cases, the smaller the frame shift (the larger the library), the better the mapping accuracy. Overall, with full data, mapping accuracy can reach 90% when enough library frames are used (Fig. 6.3). This shows that PB-sQF is indeed robust with short reads mapping with closely placed errors.

6.4.2 Read-by-Read Typing with Read Errors

To assess bacterial typing robustness in the presence of SNPs and indels, seven strains were randomly selected as mother sequences. Ten 200-bp reads, each incorporating a single randomly placed SNP were generated for each selected strain. Library candidates were searched without frame shift. After the library candidates had narrowed down from 628 total strains, frame shifts of 10 (20 times as many libraries, each consisting of 200-mer reads, with starting points shifted by ten bases) were used to select the mother strain out of the library candidates and to map the reads to the mother strains. As shown in Table

Table 6.2: PB-sQF typing with read errors

Library Strain	# of Candidates	% Mapping Accuracy
Staphylococcus aureus	9	100
Shigella flexneri	2	100
Treponema denticola	1	90
Paracoccus denitrificans	9	90
Streptococcus sanguinis	1	100
Vibrio vulnificus	9	100
Staphylococcus aureus	9	90

6.2, all the mother strains were accurately assigned with a 90 to 100% accuracy in reads mapping. This shows that PB-sQF can perform read-by-read typing even when reads have mismatch(es) from the library strain.

6.5 Valid assignments and Metagenomic Application

To improve the mapping accuracy, confidence levels can be established to reject the uncertain assignments as in the bacterial typing in Chapter 5. The mapping accuracy among the valid assignments can thereby increase. The 0% to 99% confidence levels are the left-tailed test of the test statistics distributions between the library reads and the neighboring reads within 30-mer frame shifts as described in Chapter 2. For each mapped query read, a confidence level can be assigned by comparing the test statistics between query reads and assigned library reads to the pre-calculated test statistic at each confidence level.

When building the confidence levels and generating simulated query reads, read errors were incorporated. Instead of fixing the number and types of errors as in Section 6.2, a more realistic way to introduce errors in reads is by setting an error rate for each mer. In this section, errors, including a 2% uniform error rate, a 0.09% SNP rate, and a 0.01% indel rate for each mer as implemented in [40], was introduced to the simulated reads. The indel length was determined by the Poisson distribution with lambda equals 5.

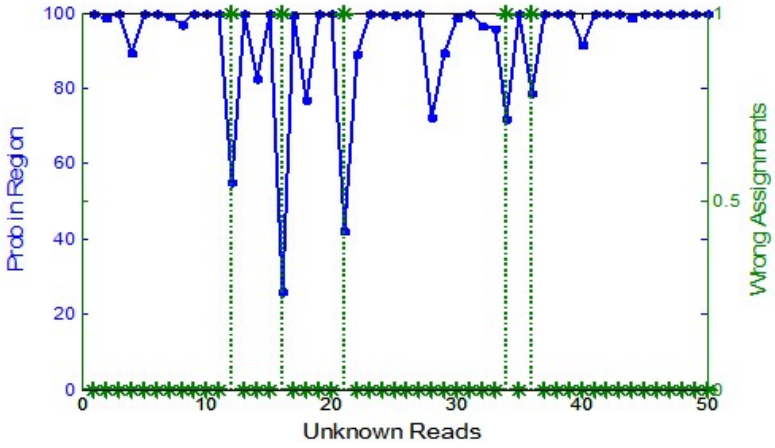

Figure 6.4: Mapping probability and mapping accuracy. The blue curve is the probability that the unknown reads originate from the assigned region and the axis is on the left: probability in the region. The green stem plot, using the axis at the right, indicates the assignments for the unknown reads are correct (0) or wrong (1).

6.5.1 Mapping Performance with Simulated Reads Errors

Fifty single-end, 200-mer long, simulated short reads with errors from 10 different bacteria randomly selected from the 628 library strains were generated and mapped to the library genomes. For each read, the top ten best-matched library reads were kept and the left-tailed probabilities (1 - confidence level), P^{left}, were calculated. In these ten best-matched library reads, regions were constructed with library reads that were separated by fewer than 100 bases. The probability that an unknown read originates from a region is calculated as one minus the product of the left-tailed probabilities for all the library reads in the region, $1 - \prod_{i=1}^{n} P_i^{left}$, where n is the number of library reads in a region. As a result, the more library reads that belong to the same region and the smaller the left-tailed percentage, the higher the probability that the unknown read was from the region. For reads that have multiple possible mapping regions, only the largest mapping probability region was considered. As shown in Figure 6.4, the accuracies of short reads assignments are related to the calculated probability with the wrong assignments mainly occurring with low-confidence assignments.

The mapping accuracies of these ten simulations are shown in Table 6.3. The correct

Table 6.3: Mapping Performance

Library Strain	% Correct	% Assigned	% Valid Correct
Mycoplasma genitalium strain G37	94	88	97.8
Pectobacterium atrosepticum strain SCRI1043	78	70	100
Bartonella henselae strain Houston-1	94	88	97.7
Pelobacter propionicus strain DSM 2379	86	78	100
Helicobacter pylori strain J99	92	82	100
Bacillus licheniformis strain ATCC 14580	90	78	100
Burkholderia sp. strain 383 chromosome 2	72	80	85
Flavobacterium johnsoniae strain UW101	90	88	100
Prochlorococcus marinus strain NATL2A	96	90	100
Propionibacterium acnes strain KPA171202	88	80	100

mapping percentages ranged from 72% to 96%. To improve the mapping accuracy, a probability threshold was set at 70%. By only accepting assignments with a probability larger than 70%, the mapping accuracies among mapped reads, which were around 80% of the total reads, are close to 100% for most strains. This shows that the low probability assignments are indeed correlated with the wrong assignments and can be excluded efficiently with the probability threshold.

6.5.2 Metagenomic Read-by-Read Mapping

The PB-sQF short reads mapping scheme was modified to perform metagenomic short reads mapping. In the metagenomic reads file, each read could come from different species. To correctly map each read, one has to identify the species for each read first. To test the ability of PB-sQF to process metagenomic data, a mini metagenomic data were simulated. Total of one hundred 200-bp long short reads were generated from the assembled genomes of *Acaryochloris marina* strain MBIC11017, *Acholeplasma laidlawii* strain PG-8A, and *Acidiphilium cellulolyticus* strain 11B.

To apply the search space reduction, the test statistics between the control library reads and all the library reads from all three strains were calculated and saved in advance. The

test statistics between query reads and control library reads were calculated. For each query read, the number of library reads candidates was reduced as every test statistic was calculated. The final library reads candidates, comprised of reads from all three library strains. Each query read was then assigned to the library strain with the highest assignment percentage, which is the number of candidates from the specific strain for each query read divided by the total number of library reads for that library strain. Once the library strain was decided for each read, the library reads candidates that were from other library strains were discarded. The rest of the procedure is the same as previously described. The test statistics between the query read and all of the library reads candidates were calculated. The top ten best-matched library reads were kept and grouped into regions. The left-tailed probability of query reads originating from each region was calculated, and the query read was assigned to the region with the largest probability.

Out of the 100 error-free short reads, PB-sQF correctly typed 94 of them back to the correct library strain, and 91 out of 100 reads were mapped to the correct library reads. Since it was impossible for the six wrongly typed reads to be mapped correctly, more accurately, only 3 out of the 94 correctly typed reads were mapped incorrectly. The mapping accuracy and probability are shown in Figure 6.5 A. There was a clear correlation between the left-tailed probability and mapping accuracy. Once again, by only accepting mapping with the probability larger than 70%, the mapping accuracy was increased to 96% among the valid reads mapping.

When errors were applied, the typing accuracy remained similar with 97 out of 100 being typed to the correct strains. Short reads mapping, on the other hand, was much more uncertain as shown in Figure 6.5 B. With errors, the mapping accuracy is 83% and increases to 87% when the low confidence assignments are rejected. These results demonstrate the potential of PB-sQF on short reads mapping in metagenomic data. To map the real metagenmoic data, however, the pre-calculated test statistics between the control reads and all library reads will need to be stored separately or a more efficient memory management will

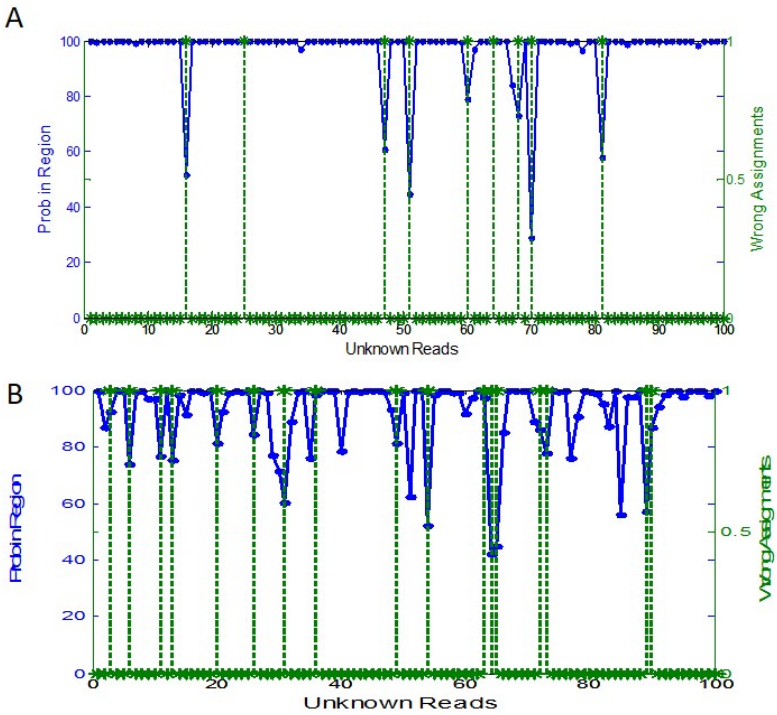

Figure 6.5: Metagenomic short reads mapping Mapping results of (A) reads without error, (B) reads with a 2% uniform error, a 0.09% SNP rate, and a 0.01% indel rate. The blue curve is the probability that the query read originated from the assigned region and the axis is on the left. The green stem plot, using the axis at the right, indicates the assignments for the particular query read is correct (0) or wrong (1).

need to be applied.

6.6 Error-tolerant Mapping and Existing Methods

Since nearest neighbor (NN) is also an Euclidean distance based similarity test, the short reads mapping procedure developed for PB-sQF can also be applied with NN. Although NN distance has been used in comparing 16S RNA sequences,[27, 101] NN has not yet been used on short reads mapping. In this section, simulated reads with different error rates were generated, and short reads mapping accuracies were compared. Different reads-mapping methods including BWA[40, 110], Bowtie2[39, 109] and SOAP2[112] were tested along with PB-sQF and Nearest Neighbor (NN).

Different from mer-based mapping, PB-sQF and NN are read-based mapping. As a

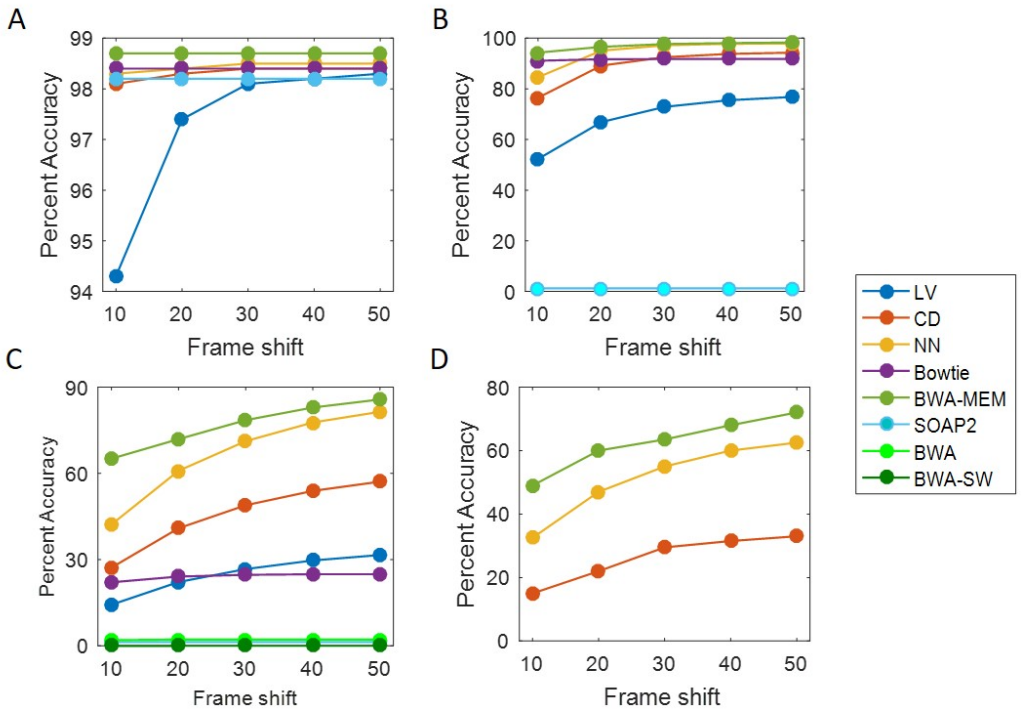

Figure 6.6: Reads errors and reads-mapping accuracies. (A) No error applied. (B) 1% uniform error rate and 2% indel rate. The indel length was determined by the geometric distribution with probability set as 0.3. (C) Similar as (B), but the indel length was fixed at 16 bps long. (D) Similar as (B), but the uniform error rate was set as 13%. Bowtie analysis was performed with Bowtie2. LV: standard PB-sQF divided each dimension at the Largest Variance. CD: modified PB-sQF where the divided dimension was determined by Cycle Dimension.

result, the mapping resolution is determined by the frame shift applied in the reads library. When the library reads were constructed with 10-mer frame shifts, the mapping results only possess number follows "10*n frame shift", where n is a integer number. Also, since both PB-sQF and NN are distance-based, all the neighbor library reads, which are similar to the correct match, will also have low (dissimilarity-based) test statistics. To fairly compare the mapping performance, instead of only considering the exact match of the assigned index and the leftmost index of the query reads, the accuracies were calculated based on frame shift. As a result, the mapping accuracy is determined as the percentage of reads that are assigned within 10, 20, 30 40 and 50-bases from the reads origins.

When no error was applied on the 10^4 paired-end, 200-bp long simulated reads gen-

erated from *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen, all methods show very good reads-mapping accuracies (Fig. 6.6 A). For PB-sQF and NN, the paired short reads were chopped into 3-mers, and in PB-sQF, 16 bins were used. To improve the mapping accuracy, a modification had been made on PB-sQF. Up until now, when dividing data into bins, "N" data points that lay on the mean (the dividing line) were randomly assigned to each bin to achieve an approximately equal number of data points per bin. Here, to reduce the variation originating from the random process and increase the mapping accuracy, the first "N" data points were assigned to one bin and the rest to another. Although the accuracy has improved, PB-sQF using the largest-variance (LV) approach (dividing each bin at the largest variance dimension) was lower especially when "within 10-bases (or frame shift)" mapping accuracy. The mapping accuracy was eventually more than 98% when the within 50-mer apart assignments were considered. The cycle-dimension version of PB-sQF, which divides the bin by cycling through all the dimensions as described in Chapter 5, increased the accuracy from 94.3% to 98.1% for the 10-mer-apart assignments. The accuracy improvement has also been observed in the pooled short reads data typing where the cycle-dimension PB-sQF had better accuracy when using 3-mer and 16 bins. The improvement most likely comes from the less biased binning process.

With 1% of uniform error and 2% indel rates and the geometric probability set at 0.3, the mapping accuracy of SOAP2[112] (with the default setting) quickly dropped to $\sim$1.2% (Fig. 6.6 B). Since the 0.3 geometric probability gives an average of 3-mer long indels, the SOAP2 typing accuracy was tested by relaxing the mapping criteria to accept 20-mer long gap and 50 mismatches; however, the mapping accuracy was similar (1.4%) to the default setting (data not shown). To test the performance of each method with longer indel gaps, the reads error rates were set the same but the indel length was fixed at 16-bp long, and the results are in Figure 6.6 C. BWA-MEM[111] and NN both had the best mapping accuracies followed by cycle-dimension PB-sQF. Again, SOAP2[112] had poor performance. Along with BWA-MEM[111], different versions of BWA including BWA[40] and BWA-ST[110]

were also tested, but both results were far from ideal. With these errors, the mapping accuracy for Bowtie2 also dropped below 30%. Lastly, the mapping performance of BWA-MEM, NN, and cycle-dimension PB-sQF at higher uniform error rate (13%) was tested with the indel rate stayed at 2% and 0.3 of geometric probability. The accuracies of all three methods dropped significantly, but the trend stayed the same: BWA-MEM had the best performance with NN slightly behind, and cycle-dimension PB-sQF followed.

6.7 Conclusions

Different from the assembled genome or pooled short read typing, PB-sQF can perform read-by-read typing through short reads mapping. The mapping accuracy can be improved by only accepting the valid assignments which have test statistics values lower than the threshold. Although the read-by-read calculation time increases linearly with the number of iterations, it can be significantly improved with pre-calculated control reads and thus opens the door for using PB-sQF in metagenomic short reads mapping.

To type the unknowns in a real metagenomic data, however, a better memory management is needed. This is because the size of the matrix of the pre-calculated test statistics increases when the number of library sequence increases. One possible solution is performing several levels of library candidates search where the densities of library reads are different at each level. The first level search is the crude search with the lowest density of library reads. With the crude search, not all the library reads need to be called at once. The library reads can be narrowed down as in the regular search space reduction. After the search space is reduced from the results of the first level search, only a subset of the second level library reads that are close to the selected first level library reads are called. With levels library reads search, some library reads will never have to be called. This greatly reduces the memory requirement. The accuracy, however, will decrease due to the missing library reads in low library reads density levels.

Although PB-sQF is great for handling complex and high dimensionality data, when

analyzing the genome sequence similarity, the similar method, NN, has a better mapping accuracy. It is because when analyzing the k-mer counts frequency, 3-mer and 6-mer are sufficient to distinguish one sequence from another. With only 64 or 4096 unique k-mers, there is no need for data compression while the optimized accuracy is reached when all k-mer are used, as in the NN case. With high error tolerance and assigning true distance between query and reference sequence, NN is a good candidate for copy number variations detection where an error-tolerant, multiple-mapping short reads mapping method is needed. I will discuss more about NN and copy number variation in the next chapter.

CHAPTER 7

COPY NUMBER VARIATIONS DETECTION

It has been shown in Chapter 6 that NN can perform error tolerant short reads mapping. Different analyses can be built on this short reads mapping such as genomic structural variation, which is the differences between individuals' chromosomes. One important structural variance is copy number variations (CNVs), which are related to different diseases such as HIV,[117] obesity,[118] cancer,[119] autism,[120] and Parkinson's disease.[121] It has also been reported that the copy numbers of resistant-related genes have increased in multidrug-resistant bacteria.[124] In these studies, a reference genome of the unknown sequence (the reads donor, or the query sequence) are usually known, but one is interested in differences in the sequenced genome from the reference genome. To study CNVs, all the short reads are first mapped to the reference genome and the differences between the reads and reference can then be studied. Short reads mapping is thus an important first step in studying genome sequence variations.

One of the major approaches in detecting CNVs is through read depth analysis.[41, 43, 222] Read depth is the average number of times that a given nucleotide is sampled in the raw short reads sequences. Assuming reads were generated evenly across the query genome, the read depth would be constant throughout the whole sequence. However, if a particular gene has multiple copies in the query genome and only one copy in the reference, the reads would appear to have a higher read depth when mapped back to the reference genome. To build the correct read depth trajectory, it is important that the short reads mapping algorithm can account for multi-reads, which are reads that can be mapped to multiple reference locations.[43, 47, 223–225]

Since most of the short reads aligners were not developed for mapping multiplicity, the subsequent CNV detectors are built to analyze read depth from uniquely mapped reads.

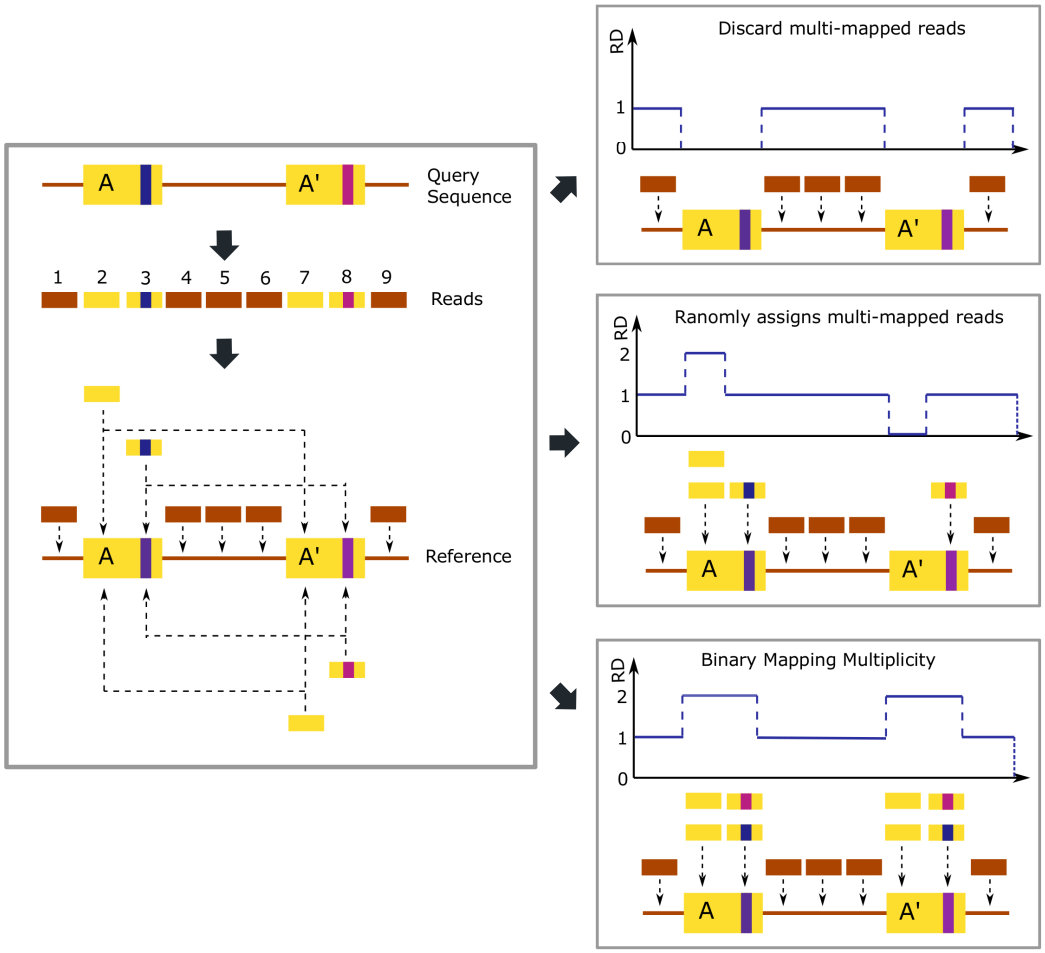

Figure 7.1: Mappability and read depth. In this example, the query sequence resembles the reference sequence except for small variation in gene A and A'. As a result, no CNV exists in this example. Reads-2, 3, 7, and 8 can be mapped to multiple locations. If the multiple mapped reads are discarded, false deletions are detected (top right). If these reads are randomly mapped to one location, one of the possible scenarios is that false duplication(s) and deletion(s) might be detected (middle right). When assigning these reads to all possible mapping locations, the copy numbers in the query sequence are obtained (bottom right). But without knowing the copy number of the reference sequence, gene A and A' might not be true CNVs.

Repetitive DNA, however, is common across all three kingdoms of life. Repeated regions comprised ∼50% of human genomes[226] and it has been found that in bacteria *Orientia tsutsugamushi* ∼40% of the genome are repeated regions.[227] When programs encounter multiple possible locations, reads aligners may (1) discard those reads or (2) randomly map them to one of the locations. For methods that discard all multi-reads, information about those repeated regions are lost (Fig. 7.1, upper right). False deletions might be detected

or important biological variants might be missed.[47, 228] Randomly choosing a mapping location, although have a better chance to recover the true read depth, might give false duplication or deletions (Fig. 7.1, middle right). As a result, methods like this tend to have a higher false CNV discovery rate.[42]

mrFast, on the other hand, was developed to map short reads to multiple positions.[28] By assigning read depth of one to all the mapping positions, the absolute copy numbers in the query sequence can be obtained (Fig. 7.1, bottom right). However, lacking the knowledge of the copy numbers in the reference sequence, true CNV regions can not be defined.

The advantage of distance-based genome sequence analysis is that mapping multiplicity can be easily handled. Instead of reporting the number of mismatches or gaps, NN reports the distances between the query read and library reads. As a result, multiple mapping locations of each query read would be those library reads that give confident test statistics. In this Chapter, the short reads mapping performance with different reads conditions are studied for NN and other aligners. A CNV detector that can directly report true CNV regions and specialize in mapping multiplicity, copy number variation detection for mapping multiplicity (CNV-MM), is developed and tested with both simulated data and real short reads data.

7.1 Reviews of Short Reads Aligners

There are several different short reads aligners. Current reads aligners can be divided into two categories based on the indexing method: hash table-based and BWT based.[104, 105] Both algorithms find the exact matches on the genome as seeds and extend the seeds with dynamic programming such as the Smith-Waterman algorithm.[108].

7.1.1 Dynamic Programming

Dynamic programming matches two strings one character at a time with a pre-assigned score for matches, mismatches, and gaps. The Smith-Waterman algorithm[108] is one of the most widely used methods to find local alignment between two sequences. An example is given in Figure 7.2, where string $TACGTAT$ is aligned to string $AACGATGA$.

To find the best alignment, the scoring system is first determined. In this example, it gives $+3$ for a match, -3 for a mismatch, and -2 for a gap, which is a result due to either deletions or insertions. Then the scoring matrix is built and and initialized by setting the first row and column as zero (Fig. 7.2 A. Different initialization will be discussed later). Then the alignment scores are calculated for each element starting from (1^{st} row, 1^{st} column), where the row and column numbers start from zero. The alignment score for each element at row i and column j, $M(i,j)$, is calculated as follows:

$$M(i,j) = \max \begin{cases} M(i-1,j-1)+3, & \text{match} \\ M(i-1,j-1)-3, & \text{mismatch} \\ M(i-1,j)-2, & \text{gap in } sequence_{left} \\ M(i,j-1)-2, & \text{gap in } sequence_{top} \\ 0, & \text{avoid negative value} \end{cases} \tag{7.1}$$

in which $sequence_{top}$ is the sequence placed on the top of the matrix ($TACGTAT$) and $sequence_{left}$ is the sequence placed at the left of the matrix ($AACGATGA$). $M(i,j)$ is thus the maximum values of all possible scenarios. In Figure 7.2 B, the first alignment, $M(1,1)$, has a score 0 since both mismatch or gaps would give negative value and thus the score is set to 0. For $M(1,2)$, it is an exact match with $A \longrightarrow A$. As a result, $0+3=3$ gives the highest score. For $M(1,3)$, the highest score is a gap in the $sequence_{top}$ thus $M(1,3) = 3-2 = 1$. The same rules are applied to all the element calculations and the direction for which the maximum score was given of each element is calculated and

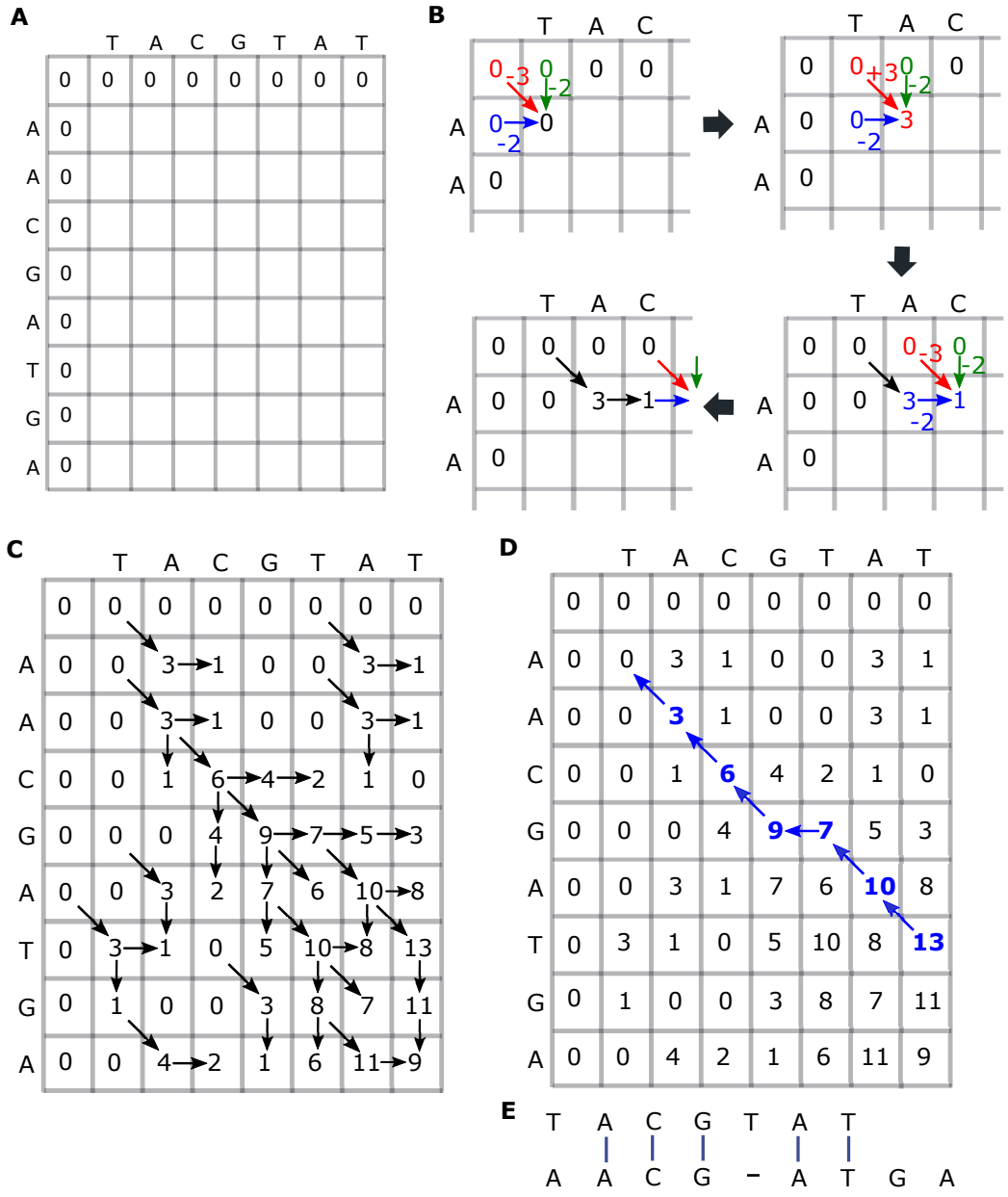

Figure 7.2: Example of Smith-Waterman alignment. (A) Initialization (B) Step-by-step score calculations. The red arrows calculate the match/mismatch scores, and the blue and green arrows calculate the gap penalties in the $sequence_{left}$ and $sequence_{top}$, respectively The black arrows are the sources of maximum scores. (C) Final score matrix. (D) Backtrack best alignment. (E) Alignment result.

recorded (black arrows in Fig. 7.2 C). The final scoring matrix is shown in Figure 7.2 C. To find the best alignment, the highest score element is selected (13 in Fig. 7.2 D) and the trace back through the source of the score is constructed (reverse directions of the black arrows in Fig. 7.2 C). In this example, the query sequence $TACGTAT$ is aligned from the

2^{nd} character to the end onto the reference sequence with a gap in the query sequence (Fig. 7.2 E).

By initializing the matrix to zero, all the characters in the string are equal, and an alignment can start at any position in either string. On the other hand, if the initialization is $0, -1, -2, ..., -9$ in the first column, these penalties prevent a new alignment to start in the middle of the $sequence_{left}$. As a result, by setting the initialization differently, one can guide the alignment to the desired behavior.

In this example, a constant gap penalty is used for simplicity. Affine gaps, however, are more frequently used. Affine gap separates the penalties from opening a new gap and extending an existing gap. It is calculated as follows.

$$Penalty_{gap} = -(g + k \cdot s) \tag{7.2}$$

in which g is the gap opening penalty, s is the gap insertion penalty, and k is the length of the gap. As a result, when using the affine gap to calculate the score matrix for each element, one needs to consider whether extending the existing gap in the current direction is more affordable than opening a gap in another direction.

Before the hash table-based methods such as BLAST[38] had been developed, dynamic programming was used in sequence alignments. When the sequences are long, however, this process can be slow. Nowadays, reads mapping algorithms find exact matches to anchor the query reads to the reference first to narrow down the search space and then extend the alignments that may contain gaps and mismatches with dynamic programming. As seen in Figure 7.2 C, there are many alignments between $TACGTAT$ and $AACGATGA$ other than the best alignment. Most reads mapping algorithms prune the dynamic programming process to focus the search for finding the best match only to reduce the calculation time.

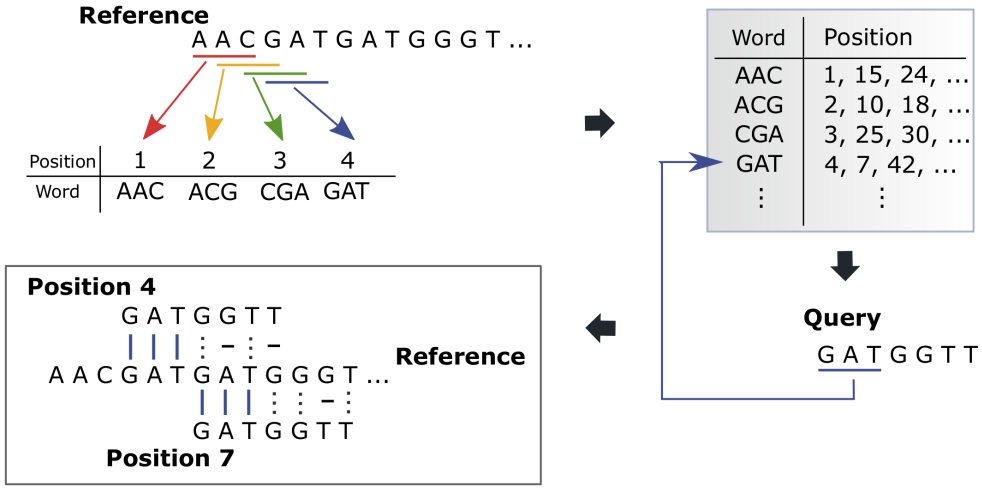

Figure 7.3: Hash table-based short reads mapping. In this illustration, the reference sequence is hashed into 3-mers (seeds). The locations of each seed in the reference are recorded in the hash table. For each query read, the first 3-mer is searched through the table. Here, the query read, GATGGTT, can be mapped to position 4, 7, 42, ..., and more. Taking position 4 and 7 as an example, once the query read is anchored to the possible position, the mapping is completed by extensions. The blue "|" represents the mapped seed. The dotted straight lines are matches via extension. The "-" indicates mismatches. Since position 7 has fewer mismatches compared to position 4, the query read is mapped to position 7.

7.1.2 Hash-based Algorithm

For algorithms based on hash tables, including PatternHunter,[106] MAQ,[29] ZOOM,[107], mrFAST,[28] and MOSAIK,[37] the words-search scheme derived from BLAST is applied.[38] This type of algorithm indexes the genome (or the reads) as k-mer subsequences and stores the information in a hash table (lookup table). When mapping the short reads to the reference sequence, the reads are also broken into k-mer subsequences for finding exact matches in the table. The process is shown in Figure 7.3

Two hash-table based reads aligners are compared in this study: MAQ[29] and mrFAST[28] since they are widely used in conjunction with CNV detections. Details of these two methods are discussed.

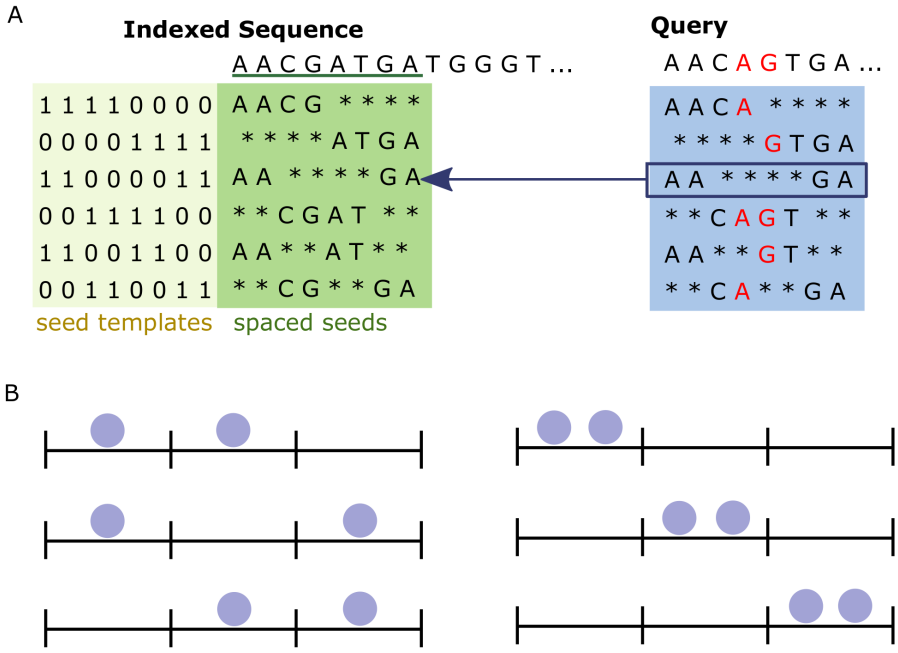

Figure 7.4: Spaced seeds and pigeon hole principle. (A) Spaced seeds indexing and mapping. The mutations nucleotide in the query read is labeled in red. (B) Pigeon hole principle. Using black lines to define three holes (seeds) and balls to represent pigeons (mismatches), all the possible arrangements of pigeons (mismatches) in holes (seeds) are listed. This shows that $k + 1$ seeds can identify sequence with k mismatches.

MAQ

Instead of indexing the reference sequence, MAQ[29] indexes the reads and maps the reference sequence to the reads' hash table. MAQ uses six spaced seeds to index the first 28-bp of reads instead of using the continuous k-mer as shown in Figure 7.3. An example of spaced seeds with 8-bp reads is, 11110000, 00001111, 00111100, 11000011, 11001100, and 00110011, where 1 means the nucleotide of a read is indexed and must be an exact match between reads and the reference sequence while 0 can be either a match or mismatch. An example of spaced reads indexing is shown in Figure 7.4 A.

The spaced seeds, which uses the pigeon hole principle, guarantee to find mapping with up to two mismatches. As shown in Figure 7.4 B, k mismatches can not fill up all $k + 1$ seeds. As a result, by specially designing the spaced seed, MAQ can map reads with two

mismatches back to the reference sequence (Fig. 7.4 A). In MAQ, reads are only mapped by the 28-bp spaced seeds. When there are multiple mapping results with the same mapping quality for a given read, the read is randomly mapped to one of the equivalent locations.

mrFAST

Designed for 36-bp read lengths, mrFAST[28] uses three 12-bp long contiguous seeds to cover the first, middle, and last part of the reads. After the seeds are mapped, an extension method similar to the Smith-Waterman algorithm[108] is applied to extend the map. By mapping a read three times with different seed regions each time, mrFAST records more mapping locations than other aligners. [28] Since mrFAST uses three seeds for 36-bp read lengths, reads with two mismatches are guaranteed to be mapped back to the reference sequence with the pigeon hole principle (Fig. 7.4 B).

When using reads longer than 36-bp, default maximum allowed mismatches are 4% of the read length and the maximum allowed indels are 4+4 (two indels of length 4-bp). Any reads that carry more than 4% of mismatches, 3 indels, or longer than 4-bp indel are not mappable using mrFAST. The size of k-mer seeds also changes when the read length increases. The default size of a seed is $floor(\frac{ReadLength}{Mismatches_{Max}})$. However, the maximum seed size is set at 14-bp. These settings show that although mrFAST can detect multiple mapping locations using multiple seeds, it is not robust for longer and error-prone reads.

7.1.3 Burrows-Wheeler Transform

In this category, the genome sequence is indexed by Burrows-Wheeler transform (BWT), which can be built from suffix array or the Burrows-Wheeler matrix (BWM).[229] As shown in Figure 7.5 A, string T: $acaaca$ is a reference sequence, and its characters are colored-coded. The color represents the rank of a letter, which is related to the number of times a letter show up. In Figure 7.5, the first occurrence of a letter is coded red, the second is green, the third is blue and the fourth is purple. The $ sign denotes a terminator

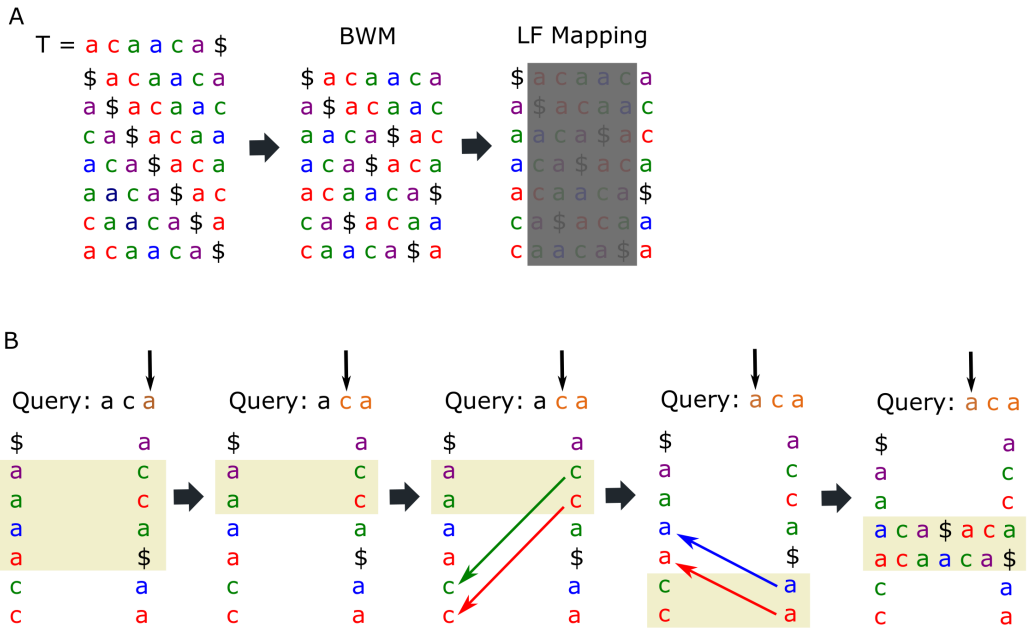

Figure 7.5: BWT indexing and mapping. (A) Building BWM and the LF mapping. The string T is the reference sequence, $ is the string terminator, and the color denotes the rank of each alphabet (B) Backward search. To map the query reads to the string, the LF mapping is used recursively to narrow down the search range.

of a sequence, and it is lexicographically prior to all characters. To build the BWM, all the character rotations of T are written down, and each row is ordered lexicographically. The last column is the BWT of T. The last-first (LF) mapping states that the rank (or the color here) of the alphabet of the last column to be the same as in the first column. As shown in Figure 7.5 A, the letter a shows up in green, blue and then red in both the first and last columns. The same applies to character c.

To find matches between reference sequence and query reads, backward matching is applied.[39] As shown in Figure 7.5 B, for a query sequence: aca, the search starts from the last a. The rows that contain a in the first column are selected (row 1 to 4, start with row 0). The next character is c, so the rows narrowing down to row 1 and 2 since they contain c in the last column. Because the BWM is constructed by permutating string T, first column a with last column c means that c is followed by a. Now, since the alphabet rank (color) is the same in the first and last column as governed by LM mapping, these green and red c in

the last column are the same green and red c in row 5 and 6 of the first column. The same process is then repeated. Since the next character in the query is a, the rows (within rows 5 and 6) with a in the last column are selected. The rank of a (color of a) in the last column again guides the search to the rows in the first column that have the same rank.

In this example, when the query is complete, the search ends at row 3 and 4. This means that the query sequence, aca, occurs in T twice, aca and $acaaca$. If the range of row is empty before the query is complete, it means that the query read does not exist in the string. The BWT is efficient because, for each character it maps, the search space is reduced. Also, exact repeats are searched at the same time (i.e. aca and $acaaca$).

Bowtie,[39, 109] and Burrows-Wheeler Aligner (BWA),[40, 110] are both reads aligners based on BWT. Details of both methods are discussed next.

Bowtie and Bowtie2

Bowtie[39] uses BWT backward search to find exact matches. For query reads that run out of rows before the reads are completed during the backward search (which means that the query reads do not appear in the reference sequence), the algorithm chooses one of the mapped nucleotides, changes it into a different nucleotide and resumes the exact match procedure till it finds a match. This scheme, however, is not suitable for longer reads with more mismatches or gaps in the reads. As a result, in Bowtie2, instead of trying to map the whole reads with BWT, it uses BWT to map part of the short reads to the genome as seeds and extends the seed with dynamic programming. The default seeding process accepts zero mismatch and the seed length is 20-bp. As a result, reads that have error(s) in the 20-bp seeds will not be mapped in the default setting.

In addition to the reads that are dropped during the seeding process, Bowtie2[109] also rejects invalid reads, reads with bad mapping scores. The mapping score is based on the base quality that is provided by the sequencer. When a mismatch occurs at a base that has low quality, the mismatch is most likely due to sequencer error and thus is given smaller

mismatch penalty compared to a mismatch that occurs at a high-quality base. The mapping score for each alignment is the summation of all mismatch penalties in a read. The quality threshold is determined by the read length, the longer the reads, the more mismatches are expected so the lower (more negative) the threshold. If reads mapping scores are lower than the threshold, these reads are categorized as invalid.

BWA, BWA-SW, and BWA-MEM

BWA[40] also uses BWT[229] to find exact matches. To find the inexact match, a maximum edited distance (Number of actions one needs to edit the query read into the reference sequence) is set. The reads can be mapped to positions on the reference with the allowed maximum differences, which is set as 4% of the read length (default). This method, however, does not perform well with longer read lengths. As a result, BWA-SW[110] was proposed. As with other aligners, BWA-SW uses BWT to find seeds and uses the Smith-Waterman algorithm[108] to extend the seeds. To speed up the process, the seeds are filtered and merged before the extension to narrow down the number of searches. To further reduce the calculation time, the dynamic programing process is pruned so the search focuses on the top z-best nodes.

Heng[111] further tuned the algorithm toward longer reads and higher error rates. The result, BWA-MEM[111], can map reads as long as $\sim$10 Mbp. In BWA-MEM, a bidirectional BWT is performed and maximal exact matches (MEM) are found and served as seeds. Since the true alignments might contain mismatches, when the MEM is longer than 28-bp, the seeds are re-seeded by the middle part of the original seeds. These seeds are again filtered and merged as in BWA-SW[110] before they are extended by dynamic programing. A more relaxed scoring system is used (smaller mismatch and gap penalties) to ensure that reads with errors can be mapped.

7.2 Mapping Robustness and Mapping Multiplicity

In this section, the mapping performance of NN is compared with Bowtie2,[109] BWA-MEM,[111] MAQ[29] and mrFAST[28] at different error rates, read length, the linearity of mapping locations and number of copies in the reference sequence. A short reads aligner that is suitable for the subsequent CNV detections should be robust against read errors while identifying all repeated regions in a reference sequence.

7.2.1 Mapping with Repeated Regions in the Reference Sequence

The bacterial strain *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen (NC007759.1), which is 2937839 bps long, was randomly selected as the model system. A randomly selected 812 bp-long region of genome index from 1542312 to 1543124, gene A, of the original sequence was repeated 2, 4, 6, 10, 15, 20, 25, 30, and 40 times and inserted back to the original sequence to generate nine pseudo-sequences. Error-free, paired-end short reads were generated with 5x coverage, which is defined as $\frac{N_{reads} \times ReadLength}{SequenceLength}$, in which N_{reads} is the number of reads. Read lengths of 36-, 50-, 76-, 100-, 150-, 200-, 250-, and 300-bp were tested. The simulated reads were mapped to the nine pseudo-sequences by NN, Bowtie2, BWA-MEM, MAQ, and mrFAST. The numbers of valid assignments for reads that were mapped to gene A (1542312 to 1543124) were counted. An ideal short reads mapper should align the reads to all repeated regions. Thus, the number of valid assignments should increase linearly with the number of repeats increases in the reference sequence.

BWA-MEM by default can produce multiple alignments for a query sequence. Since the repeated regions have different flanked sequence, the $-a$ option is used. The $-a$ (all) option in BWA-MEM map all unpaired paired-end reads. The mapping results for BWA-MEM, however, is far from linear (Fig. 7.6 C and Appendix Fig. E.3). The linearity is better when mrFAST (default setting) is used, which was built for mapping multiplicity

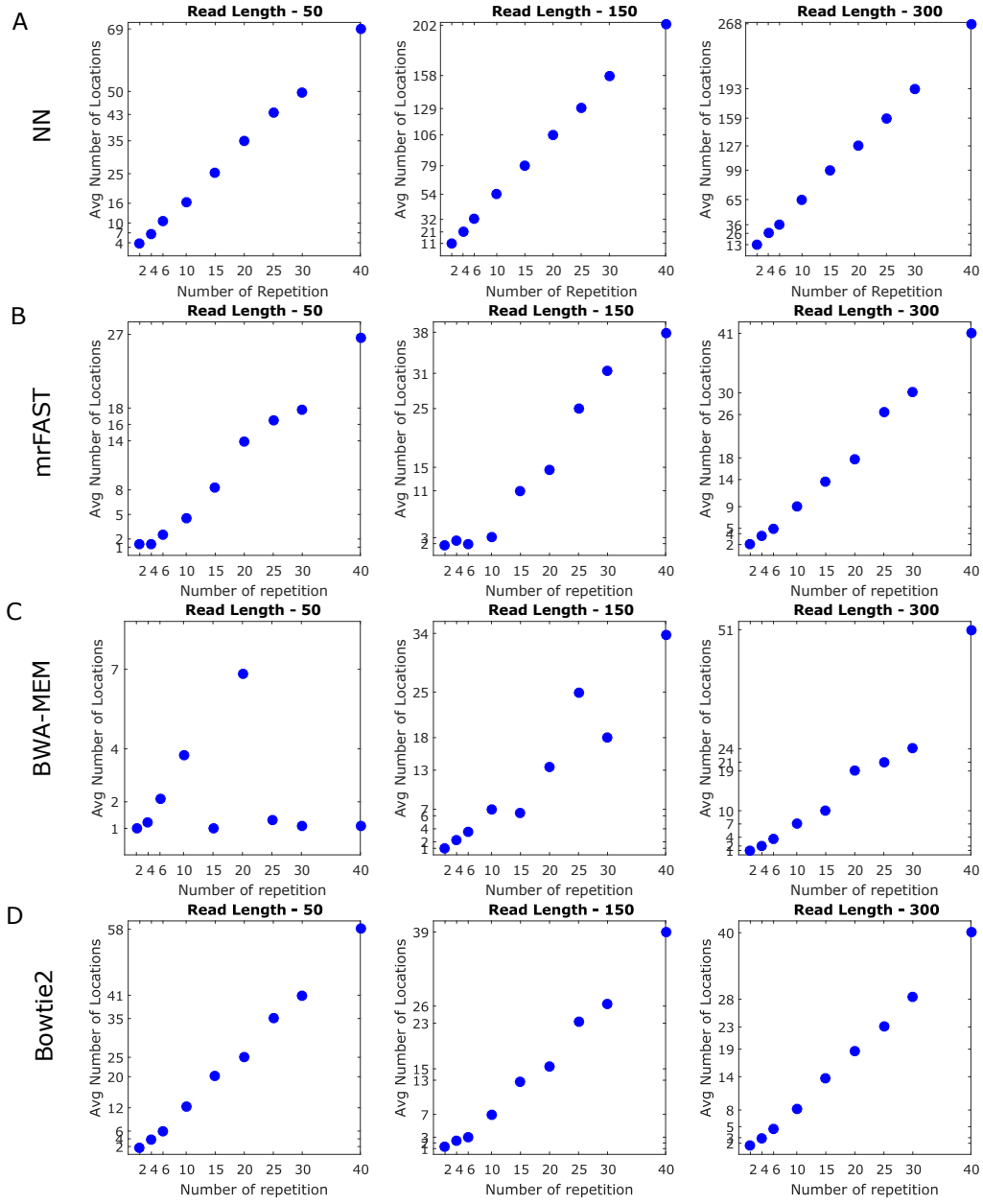

Figure 7.6: Mapping Linearity with different aligners and read length. Mapping linearity results from (A) NN, (B) mrFast, (C) BWA-MEM, and (D) Bowtie2. For NN, the average numbers of locations are not the same as the numbers of repetitions in the reference sequence. This is because in NN, the neighboring library reads of the exact match are also counted as valid reads. Since the copy number is the read depth normalized by the average read depth, this higher baseline will not influence the subsequent copy numbers determination. Results with other read lengths are shown in Appendix Fig. E.1, E.2, E.3, E.4, and E.5.

(Fig. 7.6 B and Appendix Fig. E.1). In Bowtie2, the $-a$ option is also used. But the $-a$ (all) option in Bowtie2 directs Bowtie2 to find all mapping hits instead of the default which stops when any one of the hits is found. As a result, Bowtie2 can really map multi-reads. Bowtie2 ($-a$) has good linearity with the total number of alignments increasing linearly with the number of repetitions in the reference sequence (Fig. 7.6 D and Appendix Fig. E.5). NN has the best linearity across different read length (Fig. 7.6 A and Appendix Fig. E.4). This result shows that both NN and Bowtie2 can be used in mapping reads to the highly repeated regions withing the reference sequence. MAQ, which randomly selects one mapping result can not report all the possible mapping locations. Moreover, it can not map short reads longer than 100-bp. Thus, MAQ is not suitable for mapping multiplicity (Appendix Fig. E.2).

Although BWA-MEM and Bowtie2 are both BWT-based methods, the linearities are very different. This might be explained by the differences in implementing the dynamic programming process. In Bowtie2, the extension process searches all the possible mapping results without pruning.[109] In BWA-MEM, however, dynamic programming is restricted to the top ten best nodes to reduce the calculation time.[111] As a result, valid read assignments might be lost during the pruning process.

Although the numbers of valid assignments from NN are linear with the numbers of repeats in the reference sequence, they are not the same as the number of repeats. As shown in Figure 7.6 A, the numbers of valid assignments (NN) are 4, 8, 12, for 2-, 4- and 6- times repeated reference (50-bp). This is because NN performs read-to-read mapping with thresholds. Therefore, it also picks up neighboring reads other than the exact match. Because the copy numbers are calculated by normalized the read depth with average read depth, this behavior will not influence the copy number estimation.

7.2.2 Mapping with Different Read Lengths

In the previous subsection, the aligners ability to find all mapping locations was tested. Here, the mapping accuracies for different read lengths are calculated. 10^4 of paired-end short reads with 1% uniform error, and a 2% indel rate with indel length determined by a geometric distribution of escape probability equals 0.3 were generated from the entire *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen genome and mapped back to itself. As in Chapter 6, the read mapping accuracies are calculated at each frame shift, since NN maps short reads back to library reads with frame shifts. For the details of uniform errors, indels, and geometric distribution-based indel length, please see Subsection 2.5.3.

The mapping accuracies of all the simulated reads with errors are relatively low for mrFAST (Fig. 7.7 D, 1^{st} column), especially when the reads are long. This is expected since when the reads get longer, the number of mismatches and indels accumulated. Since the probability for each base to mutate or start a indel is fixed, the longer the reads, the higher the expectation value for the number of errors. mrFAST, however, was developed for 36-mer Illumina short reads which have relatively low error rates. Even though the maximum number of mismatches are 4% of the read lengths, the maximum allowed indels are fixed to two 4-bp indels. NN, Bowtie2, and BWA-MEM all have better mapping accuracy with longer reads, with BWA-MEM having the best accuracy, NN next, and Bowtie2 last (Fig. 7.7 A to C, 1^{st} column).

As described in Chapter 6, reads mapping thresholds can be applied to NN to exclude the low confident assignments. This greatly increases the mapping accuracy among the valid assignments (assignments that pass the thresholds). For other methods, thresholds or seedings are applied by default, and more reads are unmapped when the errors are higher. As a result, the mapping accuracies among the valid assignments can also be calculated with BWA-MEM, Bowtie2, and mrFAST. The accuracies greatly improve for mrFAST (Fig. 7.7 D, 2^{nd} column) since mrFAST strictly map reads with lower than 4%-read-length mismatches and two 4-bp indels (default). As a result, most reads, especially the longer

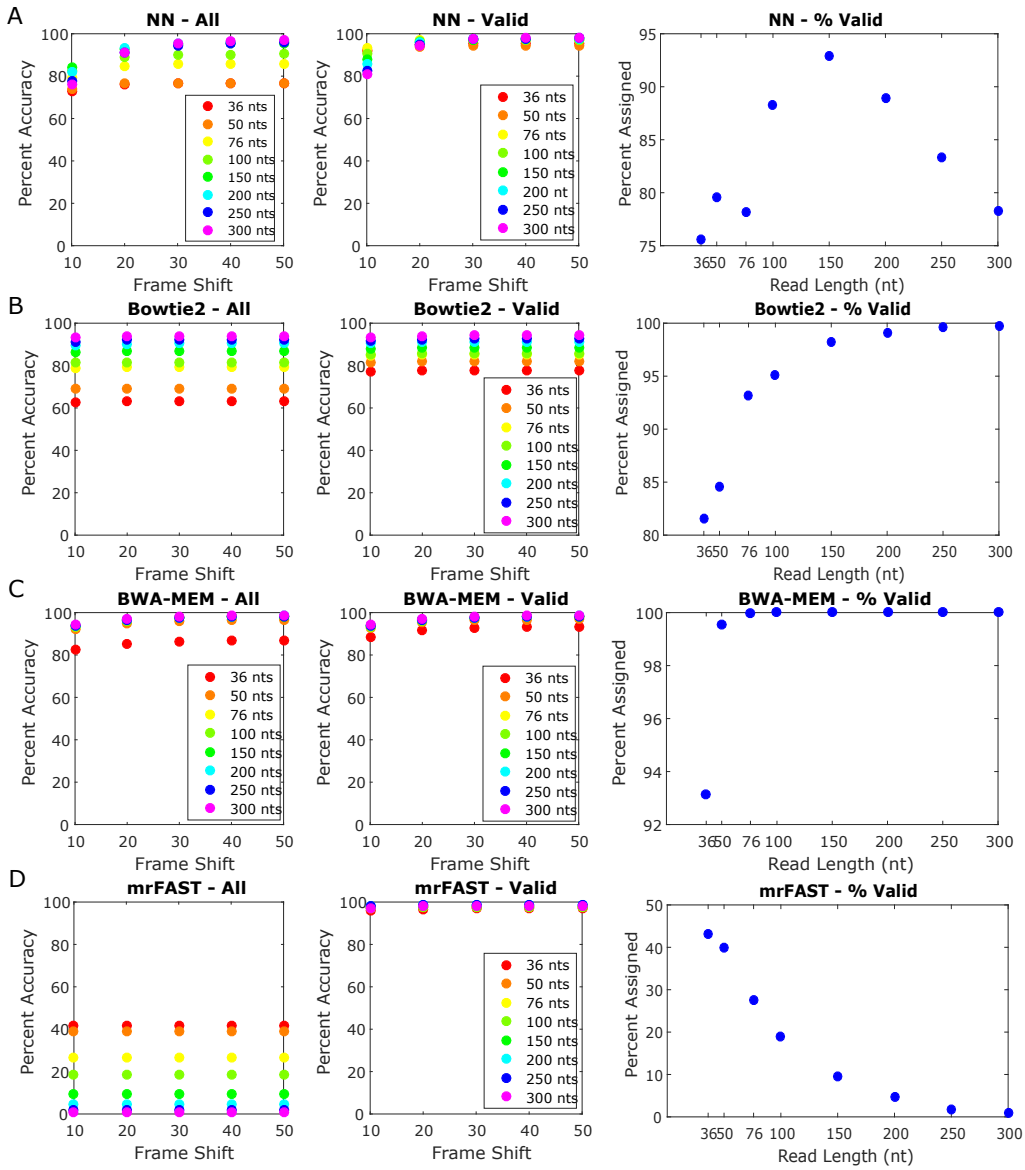

Figure 7.7: Mapping accuracies with different aligners and read lengths. Mapping results of (A) NN, (B) Bowtie2, (C) BWA-MEM, and (D) mrFAST. First column: mapping accuracies of all reads. Second column: mapping accuracies for all valid reads. The mapping accuracies are calculated as the percentage of assignments that are within 10-, 20-, 30-, 40- and 50-bp from the correct origins (for 36-bp reads, it is 6-, 12-, 18-, 24-, and 30-bp) since NN maps the short reads back to the library reads generated with 10-bp (6-bp for 36-bp reads) frame shift. As a result, the reads alignments resolution for NN is 10-bp. Third column: valid reads percentage. nt: nucleotide.

reads, easily exceed that threshold. The percent assigned is thus low for mrFAST (Fig. 7.7 D, 3^{rd} column). In BWA-MEM, the differences between the accuracies with all assignments or among the valid assignments are less significant (Fig. 7.7 C, 1^{st} and 2^{nd} columns). The percentages of assigned reads increase as the read lengths increase in both BWA-MEM and Bowtie2.

For NN, however, there is no trend in the percent assigned rate as in other methods (7.7 A, 3^{rd} column). This is because for other methods, the thresholds are based on number/score of mismatches (and indels), which is a function of read length. The thresholds in NN are designed for every read length to achieve at least $\sim 90\%$ accuracy of valid assignments for being within a certain frame shift (Fig. 7.7, A, 2^{nd} column). With thresholds specifically set to each read length, NN can better reject the incorrect assignments and increase the mapping accuracies while maintaining a decent mapping percentage. Since these thresholds for each read length balance the accuracies and mappability, they are set throughout all the analysis in our study.

7.2.3 Mapping with Different Read Errors

To examine the ability for each aligner to map reads with errors, 10^4 of 300-bp long, paired-end short reads with different error rates were generated and mapped to *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen. The types of errors examined here are uniform error, indel, and indel length.

Uniform Error Rates

Mapping efficiency as a function of uniform error rates was tested for 1%, 3%, 5%, 10% and 15% uniform error. Along with the changing uniform errors, a constant 2% indel rate with indel length determined by a geometric distribution was fixed throughout all test conditions. For mrFAST, since the read length is 300, it has poor mapping accuracy among all the reads (Fig. 7.8 D, 1^{st} column). The mapping accuracies among mapped reads are

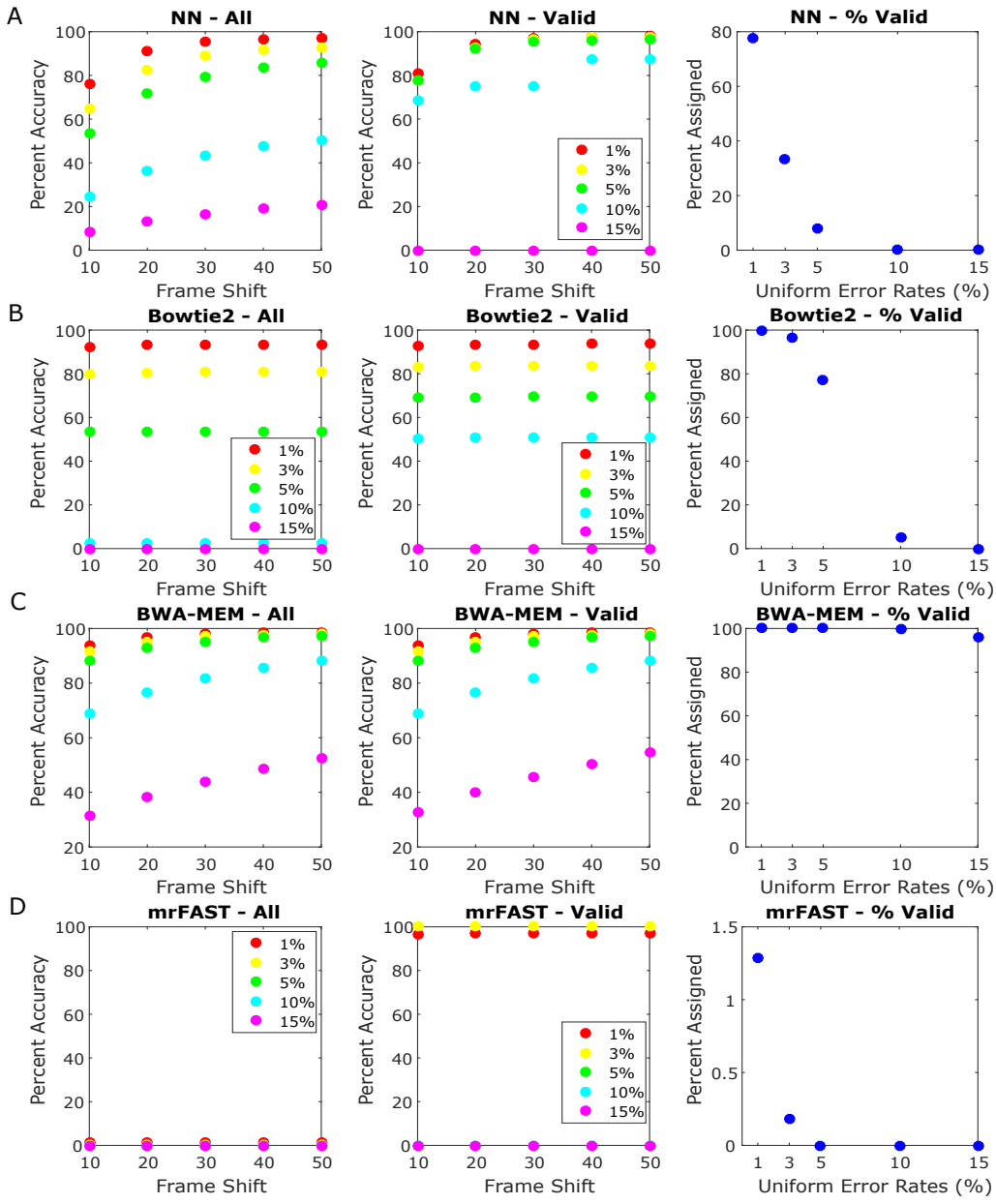

Figure 7.8: Mapping accuracies with different aligners and uniform error rates. Mapping results of (A) NN, (B) Bowtie2, (C) BWA-MEM, and (D) mrFAST. For (A) and (C), the legend is the same as (B) and (D). First column: mapping accuracies of all reads. Second column: mapping accuracies for all valid reads. The mapping accuracies are calculated as the percentage of assignments that are within 10-, 20-, 30-, 40- and 50-bp from the correct origins (for 36-bp reads, it is 6-, 12-, 18-, 24-, and 30-bp) since NN maps the short reads back to the library reads generated with 10-bp (6-bp for 36-bp reads) frame shift. As a result, the reads alignments resolution for NN is 10-bp. Third column: valid reads percentage.

close to 100% with 1% and 3% uniform errors rates (Fig. 7.8 D, 2^{nd} column). However, no reads were mapped with 5% or higher uniform error rates (Fig. 7.8 D, 3^{rd} column).

Again, BWA-MEM has the highest mapping accuracy with NN also have good mapping performance and then Bowtie2 (Fig. 7.8 A to C, 1^{st} column). Due to the default threshold in Bowtie2, reads with uniform errors higher than 10% were mostly not mapped (Fig.7.8 B, 3^{rd} column). This threshold does successfully increase the accuracies from $\sim$ 0% to $\sim$ 50% for reads with 10% of uniform errors (Fig. 7.8 B, 3^{rd} column). For NN, the mapping accuracy also drops with 10% of uniform errors (Fig. 7.8 A, 1^{st} column), however, the accuracy is still $\sim$ 50% with 50-mer shift, and it increases to almost 90% after the threshold for 300-bp reads was applied (Fig. 7.8 D, 2^{nd} column). This threshold, however, excludes all the reads with 15% uniform errors (Fig. 7.8 A, 3^{rd} column). For BWA-MEM, the increasing uniform error rates does not influence the mappability of the simulated reads (Fig. 7.8 C).

The results show the different strategies that are used by different aligners. For NN without threshold, it resembles BWA-MEM which maps as many reads as it can to achieve high error-tolerance and higher accuracy in the expense of precision . For NN with threshold applied, it implements a more cautious mapping scheme that resembles Bowtie2 and mrFAST. In this type of mapping setting, both the accuracy and precision are high but the mappability decreases. By switching on or off the threshold in NN, analysis with different focuses can be performed.

Indel Rates

Various indel rates of 2%, 5%, 10%, 15% and 20% with the indel length determined by a geometric distribution are tested. Along with the assigned indel rates, a constant 1% uniform error rate was applied. As shown in Figure 7.9 1^{st} column, indel rates have a larger impact on the mapping accuracy than do the uniform error rates. The percent accuracies drop significantly for all methods with high indel rates. mrFAST, again, has low mapping

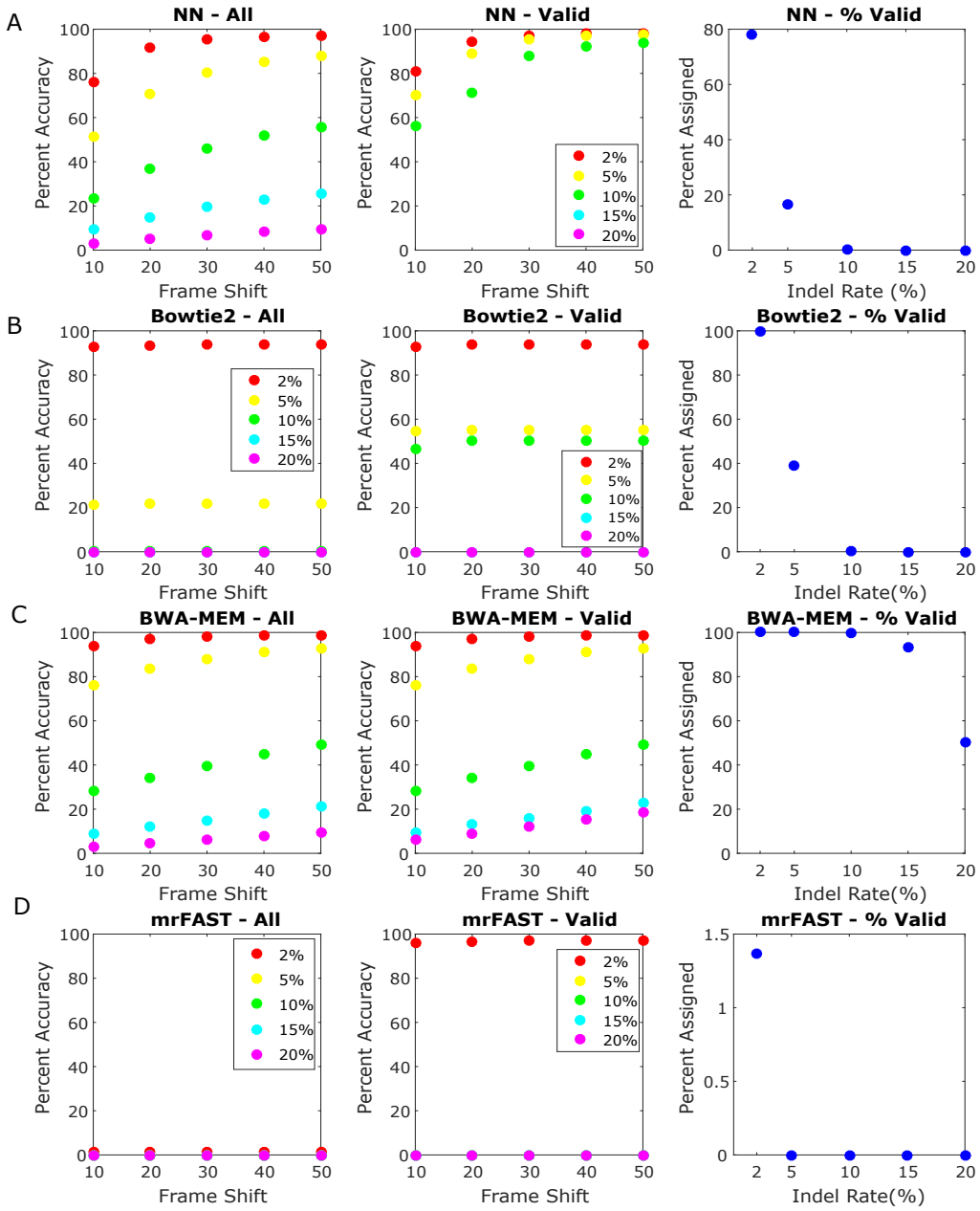

Figure 7.9: Mapping accuracies with different aligners and indel rates. Mapping results of (A) NN, (B) Bowtie2, (C) BWA-MEM, and (D) mrFAST. For (A) and (C), the legend is the same as (B) and (D). First column: mapping accuracies of all reads. Second column: mapping accuracies for all valid reads. The mapping accuracies are calculated as the percentage of assignments that are within 10-, 20-, 30-, 40- and 50-bp from the correct origins (for 36-bp reads, it is 6-, 12-, 18-, 24-, and 30-bp) since NN maps the short reads back to the library reads generated with 10-bp (6-bp for 36-bp reads) frame shift. As a result, the reads alignments resolution for NN is 10-bp. Third column: valid reads percentage.

accuracy and very few reads were mapped (Fig. 7.9 D, 1^{st} and 3^{rd} columns). Among all the reads, Bowtie2 only has good mapping accuracy with 2% of indel rates and the mapping accuracy drop significantly starting with 5% indel rates (Fig. 7.9 B, 1^{st} column). Both NN and BWA-MEM have good mapping accuracy with 2% and 5% indel rates (Fig. 7.9 A and C, 1^{st} column). NN even has a higher mapping accuracy for reads with 10% and 15% of indel rates (Fig. 7.9 A, 1^{st} column).

After applying the distance threshold, NN mapping accuracies increase from $\sim$ 50%, $\sim$ 20%, and $\sim$ 10% to $\sim$ 90% for reads with 10%, 15% and 20% of indel rates (Fig. 7.9 A, 2^{nd} column). BWA-MEM still maps most of the reads except reads with 20% indel rates and the accuracy among the mapped reads does not increase significantly (Fig. 7.9 B, 2^{nd} and 3^{rd} columns). For Bowtie2, the thresholds depend on the mismatch penalties, but they do not efficiently reject the uncorrected mapping results. Although the mapping accuracies increase, they are still only around 50% for reads with 5% and 10% indel rate (Fig. 7.9, 2^{nd} column). This result shows that NN is better in handling frequent indels in the reads than any other aligner. NN also has the benefit that the threshold is determined for each read length and can successfully exclude the non-confident mapping to increase the mapping accuracy.

Indel Lengths

Indel lengths were varied from 3, 5, 10, 15, to 20 nucleotides (nts) and mapping accuracy was tested. Along with the varying indel length, the uniform error rate and indel rates were set at 1% and 2% , respectively, for all conditions. BWA-MEM, again mapped nearly all the reads back to the reference sequence, regardless of the confidence of mappings (Fig. 7.10 C, 1^{st} and 2^{nd} columns). The mapping accuracies among valid assignments increases significantly for both NN and Bowtie2 (Fig. 7.10 A and B, 2^{nd} column), with NN having better accuracies both with and without threshold (Fig. 7.10 A, 1^{st} and 2^{nd} columns). mrFAST, again, does not have good performance with longer read length and errors (Fig.

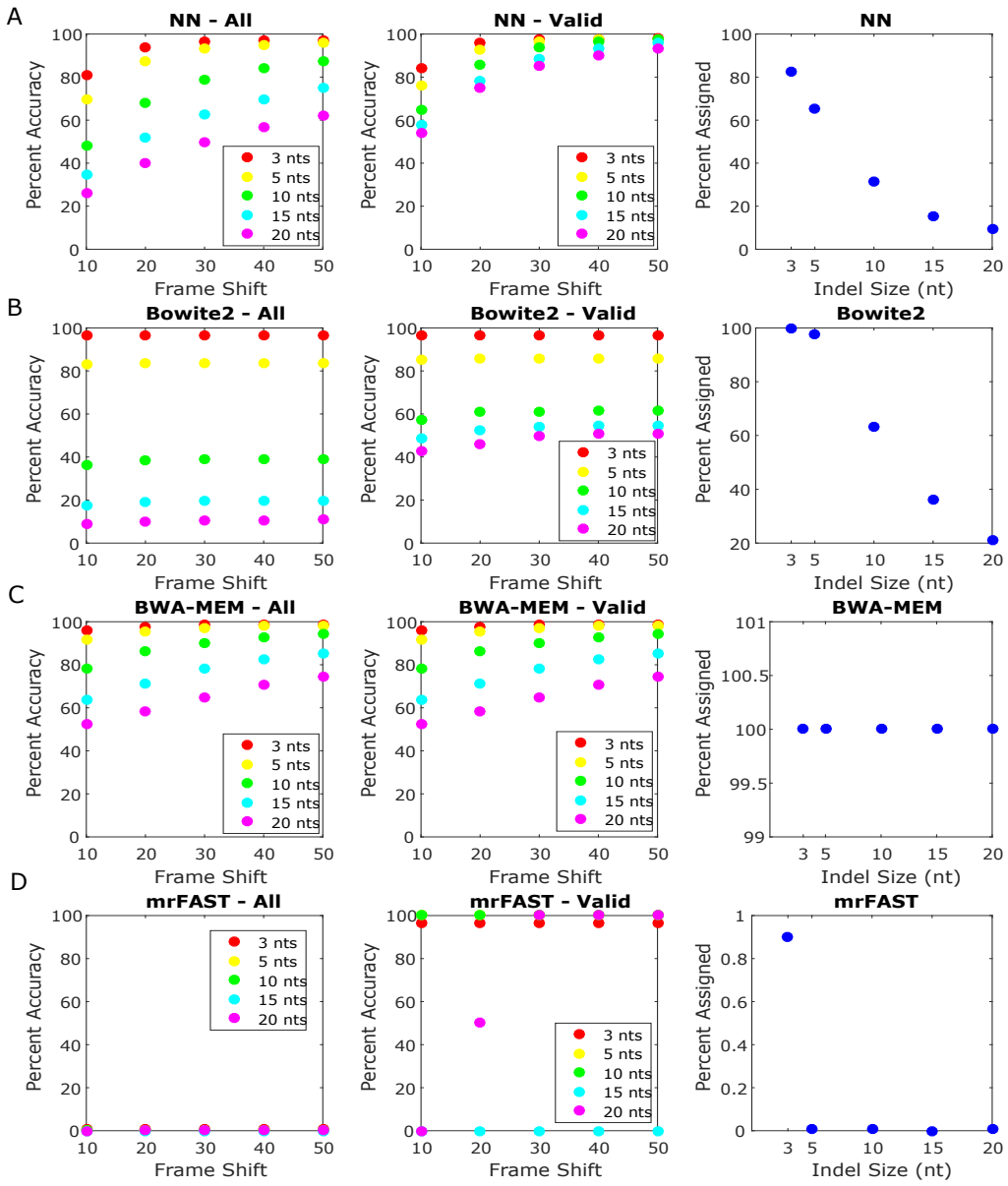

Figure 7.10: Mapping accuracies with different aligners and indel length. Mapping results of (A) NN, (B) Bowtie2, (C) BWA-MEM, and (D) mrFAST. First column: mapping accuracies of all reads. Second column: mapping accuracies for all valid reads. The mapping accuracies are calculated as the percentage of assignments that are within 10-, 20-, 30-, 40- and 50-bp from the correct origins (for 36-bp reads, it is 6-, 12-, 18-, 24-, and 30-bp) since NN maps the short reads back to the library reads generated with 10-bp (6-bp for 36-bp reads) frame shift. As a result, the reads alignments resolution for NN is 10-bp. Third column: valid reads percentage. nt: nucleotide.

7.10 D).

From the analysis of mapping linearity (Fig. 7.6) and mapping accuracies under different read lengths (Fig. 7.7) and errors (Fig. 7.8 to Fig. 7.10), NN is the best short reads aligner that can successfully find all the mapping locations while maintaining high mapping accuracy. mrFAST, and Bowtie2, although they have a nearly linear relationship between the number of locations mapped and the number of repeated region in the reference sequence (Fig. 7.6), do not perform well under reads errors (Fig. 7.8 to Fig. 7.10). BWA-MEM, on the other hand, has very high mapping accuracy even with high error rates (Fig. 7.8 to Fig. 7.10) but can not find all mapping locations (Fig. 7.6). As a result, NN is the most suitable aligner to recover the read depth for the subsequent CNV detection where mapping multiplicity is important.

7.3 Copy Number Variations Detection – Simulated Reads

CNV-MM takes the reads mapping results from NN and reports the CNVs in the query sequence. To test the reliability of CNV-MM, simulated reads were generated from *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen or pseudo-sequences with different conditions derived from it. These conditions include different similarities of repeated genes, various read lengths and CNV sizes, and varying the copy numbers in the reference and/or query sequences.

7.3.1 CNV Detection with Different Region Similarities

In the short reads mapping process, thresholds are set to exclude the low confidence mapping results. As a result, reads with less similarity to the library read will not be mapped. To test the mappability in association with sequence similarity, the genome sequence from 1542312 to 1543124 of *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen was taken as the mother gene. A pseudo-sequence (Seq-2) was generated by repeating the mother gene six times and was randomly inserted into the original sequence. Multiple, randomly placed

Table 7.1: Repeated regions. The repeated regions are listed using the index of the repeated genome (Seq-2). There are two 100% similarity regions. Region 4 is the original region while region 5 is an identical repeat.

Regions	5'-end	3'-end	Similarity
1	286300	287112	90%
2	373543	374355	92%
3	819059	819871	94%
4	1544748	1545560	100%
5	1607635	1608447	96%
6	1859340	1860152	98%
7	2684987	2685799	100%

point mutations were introduced to these six repeated genes, so they are 90%, 92%, 94%, 96%, 98% and 100% similar to the mother gene, respectively. paired-end, 100-bp reads with 50x coverage were generated from *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen without error. These error-free reads allow us to focus the analysis on the mismatches due to different levels of similarities. The six repeated regions are listed in Table 7.1.

The reconstructed trajectories are shown in Figure 7.11 A to C. Instead of the read depth, the trajectories have been dividing by their own average read depth. As a result, the copy number is shown on the y-axis. The blue curves are trajectories built directly from the mapping results. Wavelet denoising, segmentation, and CNV detection were then applied on the blue curves via CNV-MM, and the results were used to build the orange curves. Thus, the highly overlapping results between the blue curves and orange curves indicate that most of the peaks are recovered in the process. Since the average copy number of each peak is plotted in the orange curve, the maximum copy numbers appear to be smaller in the orange curves compared to the blue curves, which is clearly shown is the zoom-in plots. The mapped trajectories (blue curves) are spiky while the extracted trajectories are smooth since they are plotted using the average copy number of a region (Fig. 7.11 D to F). Therefore when zooms out, the copy numbers in the mapped trajectories seem to be higher than the extracted copy numbers. For details about trajectories building, denoising,

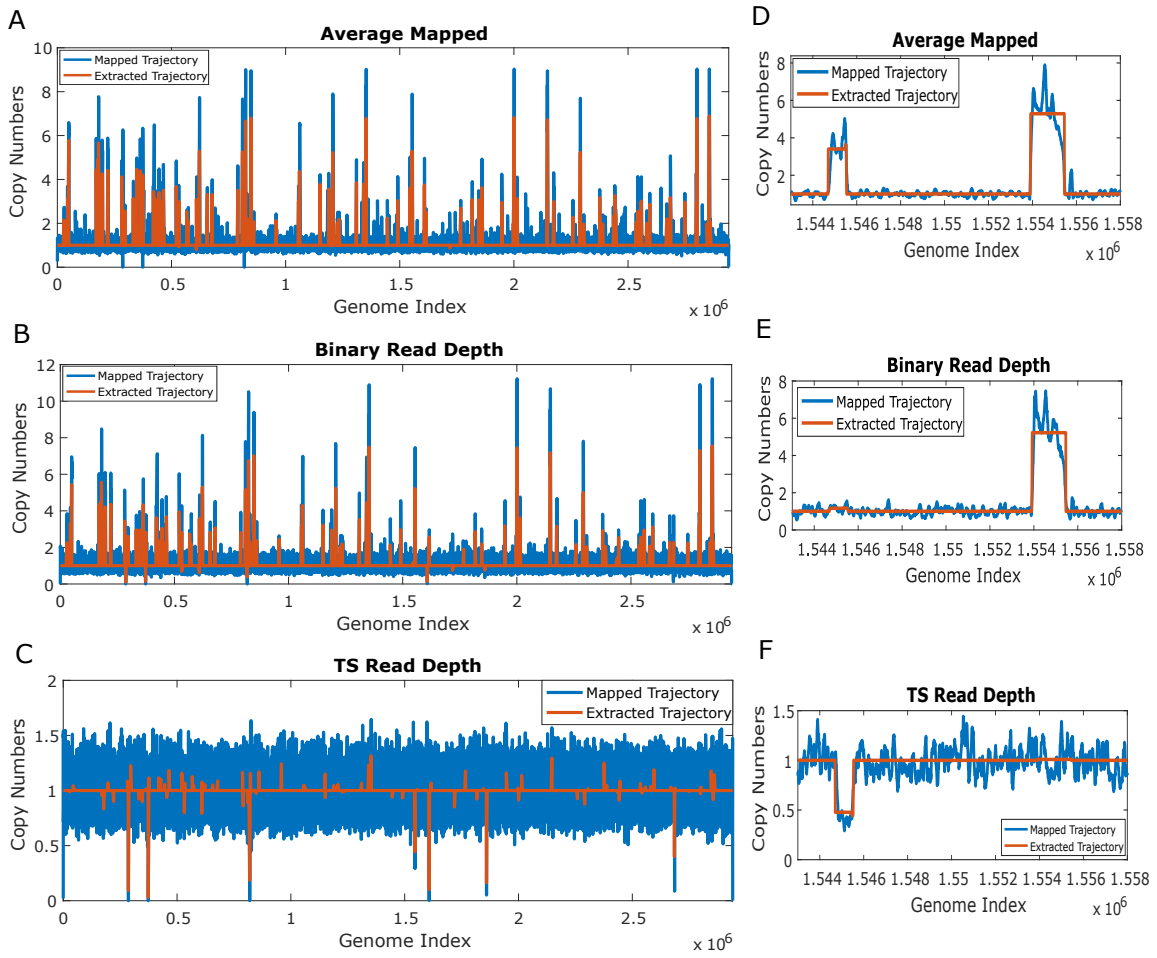

Figure 7.11: Mapping trajectories for 100-bp reads mapped seven repeated regions with different similarities. (A) and (D) Average number of assignments trajectory. (B) and (E) Binary trajectory. (C) and (F) Test statistics trajectory. (A) to (C) are the whole trajectories while (D) to (F) are the trajectories zoom in to (1543000, 1558000). The blue lines are the read depth obtained directly from the mapping results. The orange lines are read depth extracted by CNV-MM.

segmentation, and CNV detections, please see Chapter 2.

The binary and average trajectories reflect the absolute copy numbers in the reads donor sequence (original sequence) and in the reference sequence (Seq-2), respectively. As a result, there are many peaks in both trajectories. These peaks, however, are not necessarily true CNVs since they might have the same copy numbers between the reads donor sequence and the reference sequence. As shown in Figure 7.11 D to F, the peak at 1554000-1556000 has on average five copies in both the reference and query sequence (Fig. 7.11 D and E).

Although high copy number, this peak is not a CNV region since the copy numbers are the same in both trajectories. Also, in the TS trajectory, which measures the ratio of the copy numbers in the query and reference sequence, clearly shows that the ratio is one (Fig. 7.11 F).

The TS trajectory is shown in Figure 7.11 C and F. As expected, the repeated regions have lower read depth in the TS trajectory since the reads donor sequence (original sequence) only has one copy of the repeated genes which is shown in Figure 7.11 E, the copy number of the 1544748-1545560 region is one in the binary trajectory. The read depth from this one copy is thus distributed to all seven repeated regions in the reference sequence (Seq-2).

Table 7.2: CNVs with different similarity detected by CNV-MM. The CNV regions are listed using the index of the repeated sequence (Seq-2). "map" represents the indices determined by CNV-MM and "CN" means "copy number". The numbers in the parentheses are the number of bases from the real CNV insertion points (negative: toward 5'-end.) Complete list is available in Appendix Table E.1.

5'-end (map.)	3'-end (map.)	Ref CN	Query CN	TS CN	Size (bp)	Similarity
286208 (-92)	287168 (+56)	2.39	0.105	0.0929	960	90%
373440 (-103)	374400 (+45)	2.47	0.202	0.0583	960	92%
818944 (-115)	819968 (+97)	2.98	0.391	0.185	1024	94%
1544704 (-44)	1545536 (-24)	3.45	1.18	0.489	832	100%
1607616 (-19)	1608448 (+1)	3.63	0.576	0.105	832	96%
1859328 (-12)	1860160 (+8)	3.61	0.758	0.167	832	98%
2684928 (-59)	2685952 (+153)	3.17	1	0.4	1024	100%

The estimated breakpoints and copy numbers are listed in Table 7.2 and Appendix Table E.1. The breakpoints of these seven genes are close to the true gene size (812 bp). As described in Chapter 2 (Subsection 2.7.4), related regions in a reference sequence can be grouped together simply by taking a union of all the mapping results of the multi-reads. Since CNV-MM identifies these seven regions are related, meaningful observations can be drawn. There are seven regions in the reference are grouped but the estimated copy number

in the reference sequence is not seven. This is because of these regions are not 100% similar to each other. As a result, most of the reads can only be mapped to a subset of all seven regions. The copy numbers in the query sequence (the binary trajectory), on the other hand, varied depending on the similarities of the regions (Table 7.2, Query CN). For regions with 100% similarity, the estimated query copy number is around one. This means that the query sequence has one copy of region 1544704-1545536 and region 2684928-2685952. Ideally, all seven regions should have a query copy number equal to one, due to sequence dissimilarity, however, the query copy numbers are around 0.76, 0.58, 0.39, 0.20, and 0.11 for regions with 98%, 96%, 94%, 92% and 90% similarity, respectively. This means that with the current threshold setting, only $\sim$76%, 58%, 39%, 20% and 11% of the total reads are mapped to each region with different similarities. These percentage can be used to predict which region(s) is deleted in the query sequence. Since the there is clearly a deletion in the query sequence (although the complicated similarity issue clouded the information as how many region(s) is deleted), the regions with the smallest query copy number must have a higher probability to be deleted.

Without the grouping information, these seven regions would be treated as individual CNV regions and less information can be extracted. Taking the region 286208-287168 for example, without grouping, one only knows that there are $\sim$2 copies (2.39) of this region in the reference while one is the current region but the other is not known. From the query copy number (0.105) and TS copy number (0.0929), it is known that there is a deletion in the query sequence, however, one can not tell it is this 286208-287168 is deleted or the related region that is somewhere in the reference sequence is deleted.

This data helps us understand the non-integer and inconsistent value in copy numbers of other simulated data or real data. First, when the estimated copy numbers in the reference sequence are not consistent with the number of regions grouped, it might be due to the regions, although they share some similarity to be grouped, are different enough, so the test statistics between these regions and some reads are lower than the thresh-

Table 7.3: Regions repeated in Seq-2. The repeated regions are listed using the index of the repeated genome (Seq-2). For each region length, there are two copies: the original and the duplication. "ori." represents the original sequence, and "dpc." represents the duplicated sequence.

Regions	5'-end (ori.)	3'-end (ori.)	5'-end (dpc.)	3'-end (dpc.)	Size
1	3142001	3167001	45711	70711	25 kb
2	952139	1077139	2048741	2173741	125 kb
3	1583501	1833501	2536587	2786587	250 kb
4	859203	861703	3197904	3200404	2.5 kb
5	2935863	2936113	3260410	3260660	250 bp

old. Second, as discussed in Chapter 2, these inconsistent mappings render the equation $CopyNumber_{TS} = \frac{CopyNumber_{Binary}}{CopyNumber_{Avg}}$ invalid. In this situation, the group average TS test statistics can be used to distinguish the real CNVs from others.

7.3.2 CNV Detection with Different CNV sizes

To test the ability of CNV-MM to detect CNVs of different sizes, five genome regions from *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen were randomly selected and inserted back to the original sequence to construct the pseudo-sequence. This pseudo-sequence is now the new Seq-2. The length of these five regions are 250 bps, 2500 bps, 25 kbps, 125 kbps, and 250 kbps and the positions of the repeated regions are listed in Table 7.3 (in Seq-2 index) and Table 7.5 (in original sequence index). When taking the original sequence as the reference sequence, the reads were generated from the pseudo-sequence and mapped to the reference sequence and vice versa. Reads were generated with 1% uniform error rate, and 2% indel rate with the indel length follows a geometric distribution.

Deletions - Seq-2 as the Reference

The trajectories reconstructed from 200-bp reads generated from the original sequence with 50x coverage mapped by NN using Seq-2 as the reference sequence are shown in Figure

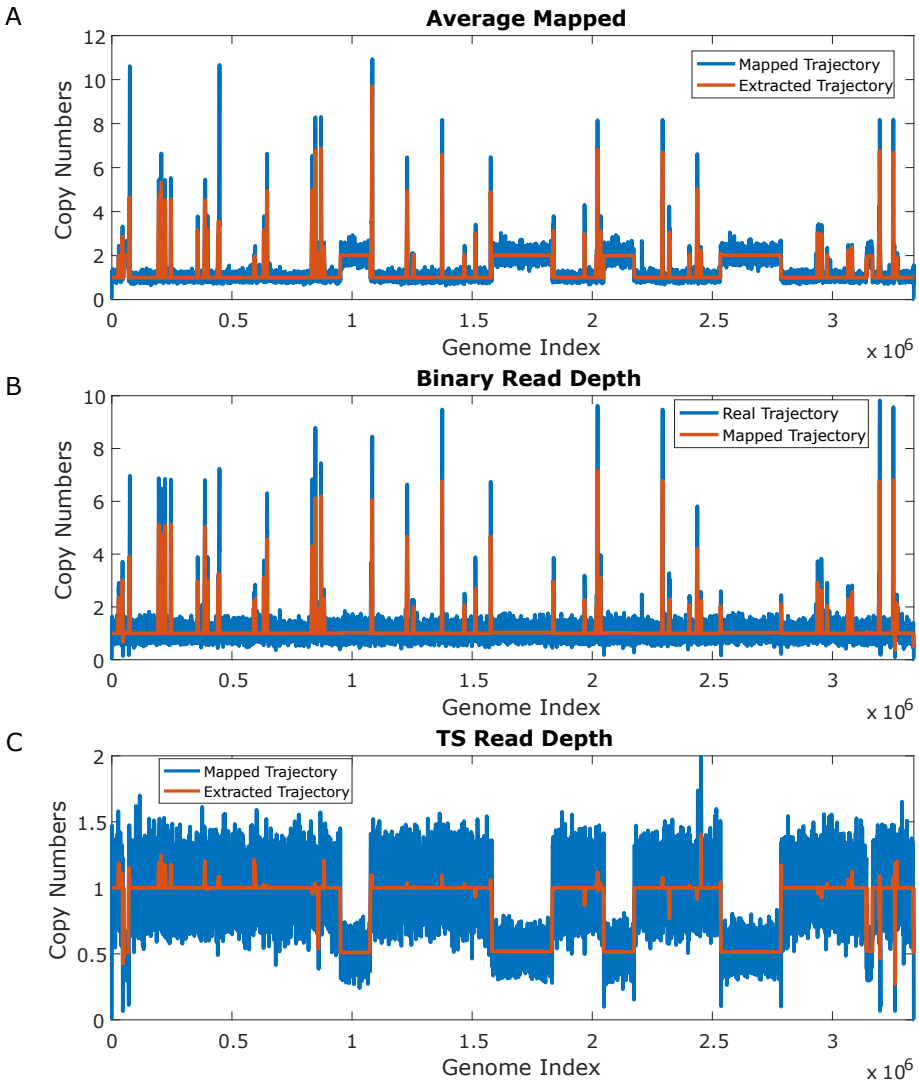

Figure 7.12: Mapping trajectories for 200-bp reads mapped to Seq-2. (A) Average number of assignments trajectory. (B) Binary trajectory. (C) Test statistics trajectory. The blue lines are the read depth obtained directly from the mapping results. The orange lines are read depth extracted by CNV-MM.

7.12. The repeated region indices are shown in Table 7.3. Since Seq-2 is the reference sequence, the average trajectory (Fig. 7.12 A) shows 2x copy numbers for the repeated regions. In the binary trajectory (Fig. 7.12 B), only one copy is shown in the corresponding regions. The TS trajectory (Fig. 7.12 C), as a result, show that the copy number ratio is 0.5.

The detailed mapping results are shown in Table 7.4. The analysis was done with the grouping function of CNV-MM turned on. As a result, regions sharing sequence similar-

Table 7.4: Deletions detected by CNV-MM The repeated regions are listed using the index of the repeated genome (Seq-2). "map" represents the indices determined by CNV-MM and "CN" means "copy number". For TS copy number, the copy number ratio is measured instead. The numbers in the parentheses are the number of bases from the real CNV insertions (negative: toward 5'-end.) "–" means the difference is not applicable since it is a subset of an existing region.

Group	5'-end (map.)	3'-end (map.)	Ref CN	Query CN	TS CN	Size (bp)
1	1583616 (+115)	1834496 (+995)	2.01	1.03	0.52	250880
1	2535424 (-1163)	2786688 (-101)	2.02	1.03	0.52	251264
2	952320 (+181)	1077248 (+109)	2.01	1.02	0.51	124928
2	2048000 (-741)	2174976 (+1235)	2	1.02	0.51	126976
3	45568 (-143)	70784 (+73)	2	1.03	0.52	25216
3	3143680 (–)	3167232 (+231)	2	1.03	0.52	23552
4	3142144 (+143)	3142656 (–)	1.86	0.93	0.51	512
5	859136 (-67)	861184 (-519)	2.23	1.14	0.51	2048
5	3197824 (-80)	3200512 (+108)	2.13	1.06	0.50	2688
6	3260288 (-122)	3260800 (+140)	1.55	0.20	0.13	512
7	1967616	1969600	3.06	2.25	0.81	1984
7	2320128	2322048	3.06	2.22	0.74	1920
7	3193728	3195712	3.10	2.25	0.74	1984
8	3338240	3338743	0.98	0.49	0.50	503

ity are grouped. Group 1 to Group 6 are repeated sequences from 2500-bp to 250-kb as designed in Table 7.3. The copy number in the reference sequence calculated from the average trajectory (Fig. 7.12 A) is two and the copy number in the query calculated from the binary trajectory (Fig. 7.12 B) is one. The copy number ratio between the two sequences are calculated from the TS trajectory (Fig. 7.12 C) shows that the original sequence (query sequence) indeed only has half of the copy numbers present in those regions of the reference sequences.

The regions in group 3 have been divided into two subregions including group 3 and group 4. This is because of fluctuations in trajectories in these regions and CNV-MM identifies them as separate events. Group 6 is the deletion of the 250-bp repeated regions while CNV-MM does not detect the deletion at the original sequence (2935863-2936113).

This is because the breakpoint (CNV region boundaries) resolution depends on the read length since NN performs read-to-read mapping. When using 200-bp reads, the 250-bp repeated region is right at the detection limit and thus the contrast of read depth at 2935863-2936113 is reduced. Although group 6 should suffer the same read depth dilution and therefore renders it harder to detect, the edge effect makes this inserted sequence stand out. The edge effect happens at all inserted sequence, where the sequences flanking the 5' and 3' ends of the inserted sequence are different from those flanking the original sequence. Therefore, the reads generated from the edges of the original sequence can not be mapped to the edge of the inserted sequence and it appears as extra deletion for at the edge. This edge effect is not significant when the region is long. When the region length is 250-bp, however, it becomes prominent. Groups 7 and 8 are false deletions detected by CNV-MM due to uneven read depth, which might be a result of uneven reads sampling. The full list of detected CNV regions is shown in Appendix Table E.13.

Duplications - the Original Sequence as the Reference

We shown that CNV-MM can detect deletions with a wide range of CNV sizes. To examine whether CNV-MM can detect duplications with different CNV sizes, Seq-2 (the two times repeated sequence) was used as a reference sequence while the original sequence served as the read donor sequence instead. Again, the reads are paired-end, 200-bp long and the coverage is 50x. The regions that were duplicated in the original sequence to create the Seq-2 are listed in Table 7.5.

All three trajectories are shown in Figure 7.13. Since the reference sequence (original sequence) now has only one copy while the reads donor (Seq-2) has two, the binary trajectory shows two copies at the "repeated" regions. The TS trajectory also clearly shows that there are CNV regions with copy numbers two times higher than in the reference sequence. Again, as discussed previously, the inconsistent heights between the mapped trajectories (blue curves) and the extracted trajectories (orange curves) are mainly because the average

Table 7.5: Regions in the original sequence that are repeated in Seq-2. The regions that served as the mother regions for the repeated regions in Seq-2 are listed using the index of the original genome.

Regions	5'-end	3'-end	Size
1	2742001	2767001	25 kb
2	927139	1052139	125 kb
3	1558501	1808501	250 kb
4	834203	836703	2.5 kb
5	2535863	2536113	250 bp

Table 7.6: Duplications detected by CNV-MM The repeated regions are listed using the index of the original genome. "map" represents the indices determined by CNV-MM and "CN" means "copy number". The numbers in the parentheses are the number of bases from the real CNV insertions (negative: toward 5'-end.)

Group	5'-end (map.)	3'-end (map.)	Ref CN	Query CN	TS CN	Size (bp)
1	834304 (+101)	836608 (-95)	1.06	2.07	2.00	2304
2	927232 (+93)	1052032 (-107)	1.02	2.06	2.03	124800
3	1558528 (+27)	1808384 (-117)	1.01	2.02	1.99	249856
4	2742272 (+271)	2766848 (-153)	1.07	2.02	1.93	24576

copy number of a given peak is plotted in the extracted trajectories. The higher peaks in the mapped trajectories are due to the fluctuated read depth as shown in Figure 7.11 D to F.

The full list of detected CNV duplication regions is shown in Table 7.6. Again, all the 2-fold copy numbers are captured except for the 250-bp region as was also true in the deletion case. The estimated copy numbers in the reference sequence is one and in the reads donor is two. The copy number ratio, determined by the TS trajectory, indicates that the copy numbers in these regions indeed increase as suggested by the binary and average trajectories. The breakpoints (boundaries) of these regions are more accurate compared to those determined in the deletion case with the estimated breakpoints mostly being within 200-bp away from the true breakpoints. There is no false discovery.

Since the absolute copy numbers are estimated by the binary and average trajectories,

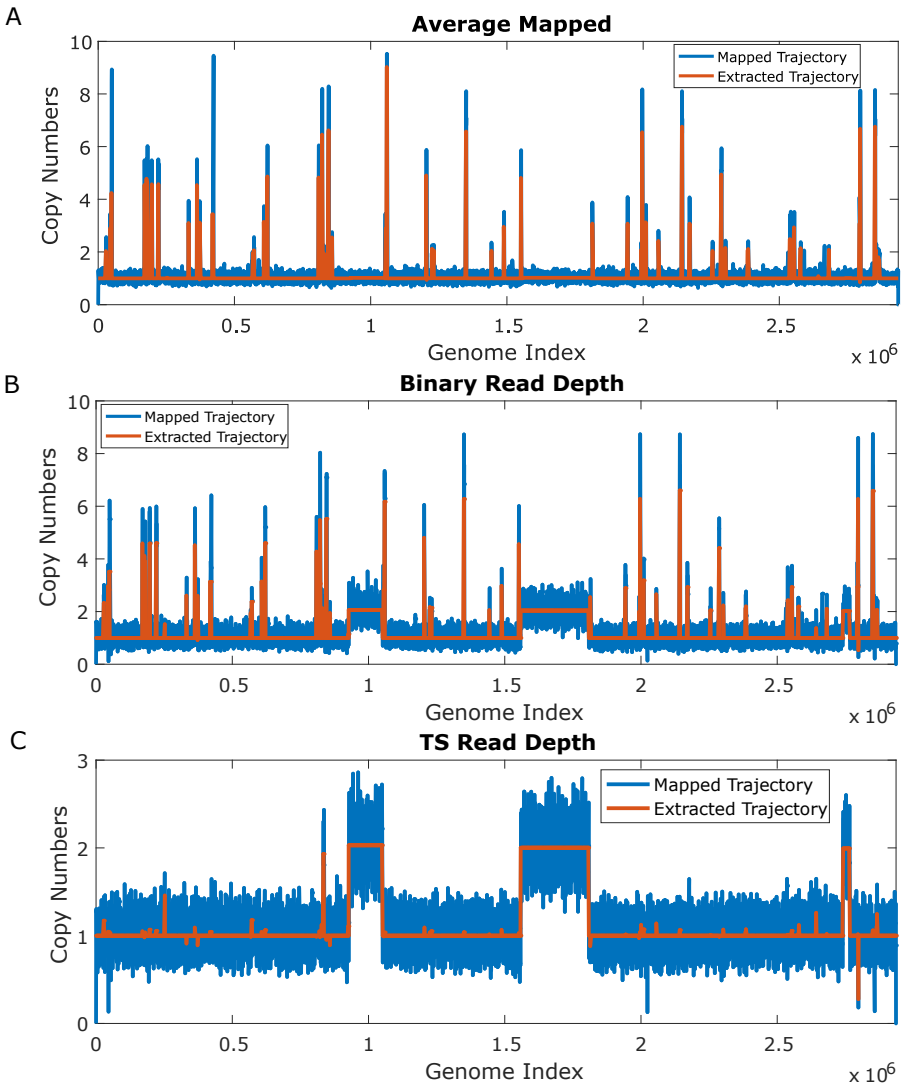

Figure 7.13: Mapping trajectories for 200-bp reads mapped to the original sequence. (A) Average number of assignments trajectory. (B) Binary trajectory. (C) Test statistics trajectory. The blue lines are the read depth obtained directly from the mapping results. The orange lines are read depth extracted by CNV-MM.

it seems like it is not necessary to calculate the TS trajectory. This is, however, not true. First, the TS trajectory is used to determine the CNV boundaries. When using binary and average trajectories alone, CNV regions longer than 2500 bp tend to be break into pieces due to the fluctuating read depth (trajectories). Because only the true CNV regions are shown in the TS trajectory, the TS trajectory is relatively clean. As a result, the wavelet denoised trajectory is also cleaner and has a reduced tendency to break regions apart. This

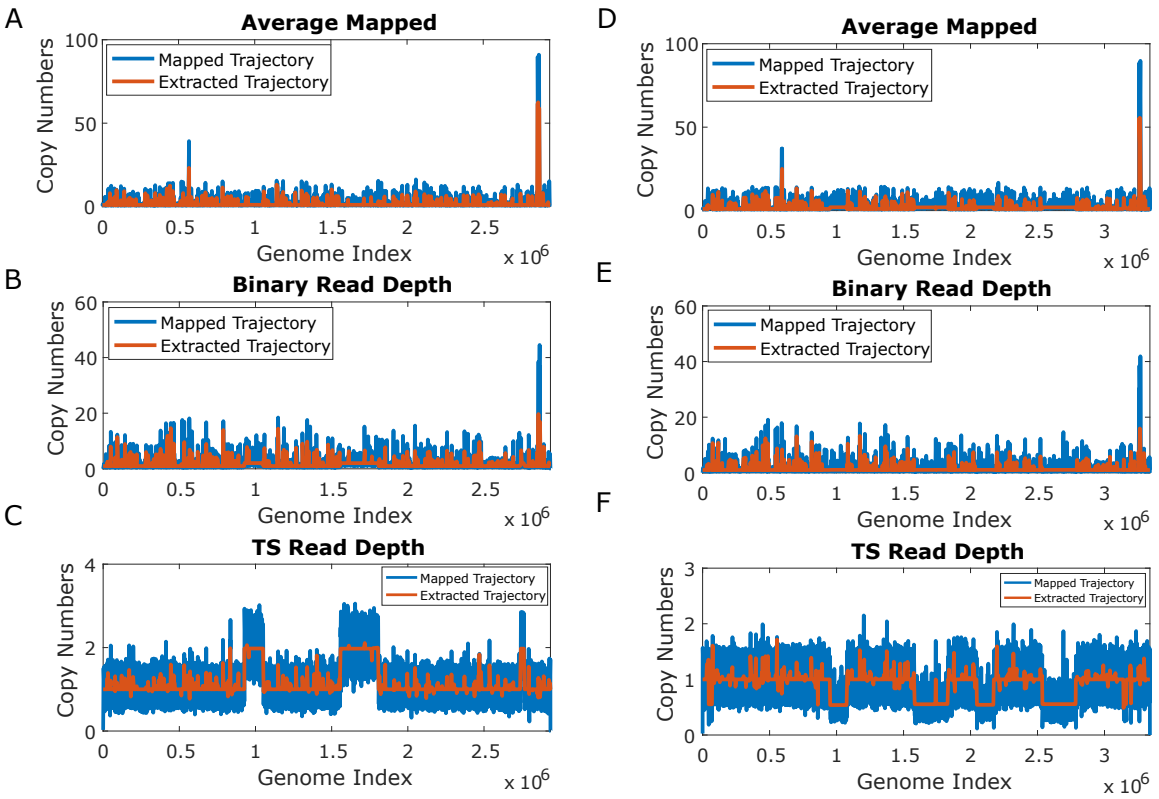

Figure 7.14: Mapping trajectories for 36-bp reads. For (A) to (C), the reference sequence is the original sequence, and the reads donor sequence is the repeated sequence (Seq-2). (A) Average number of assignments trajectory. (B) Binary trajectory. (C) Test statistics trajectory. For (D) to (E), the reference sequence is the repeated sequence, and the reads donor sequence is the original sequence. (D) Average number of assignments trajectory. (E) Binary trajectory. (F) Test statistics trajectory. The blue lines are the read depth obtained directly from the mapping results. The orange lines are read depth extracted by CNV-MM.

greatly helps to identify long CNV regions. Also, there are 157 duplicated regions and 31 deletion regions detected when all three trajectories are considered. The TS copy number greatly reduce the false positive rate by excluding regions with a TS copy number within the Poison noise from the average TS read depth. And when the regions have inconsistent copy numbers due to the differences in similarities as discussed in Subsection 7.3.1, the average TS copy number can help exclude the false positive detections.

7.3.3 CNVs Detection with Various Read Lengths

The same system used in the CNV length study is used to investigate CNV detection with the NN mapping results from different read lengths. 50x coverage of paired-end reads of length 36, 50, 76, 100, 150, 200, 250, to 300 bp were generated from both the original sequence and the repeated sequence (Seq-2). Each sequence takes turns to be the reads donor or reference sequence to detect various deletions and duplications as in Subsection 7.3.2.

When building the trajectories from 300-bp to 36-bp reads, all three trajectories become noisier with shorter read length. The trajectories from 36-bp reads are shown in Figure 7.14, 200-bp results are in Figures 7.12 and 7.13, while the trajectories from other read lengths are shown in Appendix Fig. E.6 to Appendix Fig. E.11. Read depth noise becomes high when the read length is shorter than 100-bp. Noise is especially high for 36-bp and 50-bp reads trajectories.

For 36-bp binary and average trajectories, two highly repeated regions appear around reference genome index 60000 and 2850000. These two peaks do not show up in any other trajectories with longer read length. This is probably because there are multiple short repeated units in both regions (these two regions are related). These two high copy number regions, however, have copy numbers equal to one in the TS trajectory (Fig. 7.14 C and F.) despite the ratio between the binary and average copy numbers being $\sim$0.33 for the 2850000 bp region with the copy numbers in the average trajectory being $\sim$90 and in the binary trajectory being $\sim$30 (Fig. 7.14 A, B, D, and E.). Since the TS copy numbers are inconsistent with the $\frac{CopyNumber_{Binary}}{CopyNumber_{Avg}}$ ratio, and these highly repeated peaks disappear with longer read length, it is highly possible that these regions are only partially similar to each other. As a result, the TS copy number should be used as the guideline to exclude these regions as being true CNVs.

A partial list of CNV results focused on the inserted regions is shown in Table 7.7 (Full list in Appendix Table E.2). The 2500 bp to 250 kb duplicated regions are detected without

Table 7.7: Duplications detected by CNV-MM from 36-bp read depth trajectories. The repeated regions are listed using the index of the original genome. "map" represents the indices determined by CNV-MM and "CN" means "copy number". The numbers in the parentheses are the number of bases from the real CNV insertions (negative: toward 5'-end.)

Group	5'-end (map.)	3'-end (map.)	Ref CN	Query CN	TS CN	Size (bp)
1	1558528 (+27)	1808512 (+11)	1.06	2.04	1.97	249984
2	927104 (-35)	1052160 (+21)	1.08	2.06	1.98	125056
3	2742016 (+15)	2766976 (-25)	1.08	2.04	1.98	24960
4	834048 (-155)	836672 (-31)	1.05	2.01	1.92	2624

grouping and the breakpoints accuracy is even better than the breakpoints extracted from the 200-bp read depth trajectories (Table 7.6) since the resolution is depending on the read length and frame shift. There are, however, much more false positive CNV detections (Full list in Appendix Table E.2). The same trend can be found in the CNV results from 50-bp to 300-bp. (Appendix Table E.2 to Appendix Table E.17.)

When using the repeated sequence (Seq-2) as the reference sequence, the deleted regions from 2500 bp to 250 kb are detected (Table 7.8). However, two regions are broken down into pieces. The 1583501-1833501 bp region is divided into two regions, 1583104-1740288 and 1740321-1833472 while the 3142001-3166976 into 4 regions, which cover from 3142144 to 3166976 (Fig. 7.14 F).

For CNV detections with read lengths shorter than 100 bp, it is not recommended to turn on the grouping function. This is because when the read length is short, the probability that different regions share similar k-mer distribution is high. As a result, the NN test statistics value is similar. This makes grouping difficult to exclude unrelated regions which obscures the subsequent analysis when the group average copy numbers are calculated.

7.3.4 CNV Detection with Multiple Copies of Duplications or Deletions

Since NN can identify all the repeated regions in the reference sequence as shown in Subsection 7.2.1, it is expected that CNVs from complicated regions (multiple copies in both

Table 7.8: Deletions detected by CNV-MM from 36-bp read depth trajectories. The repeated regions are listed using the index of the repeated genome (Seq-2). "map" represents the indices determined by CNV-MM and "CN" means "copy number". The numbers in the parentheses are the number of bases from the real CNV insertions (negative: toward 5'-end.) "–" means the difference is not applicable since it is a subset of an existing region.

Group	5'-end (map.)	3'-end (map.)	Ref CN	Query CN	TS CN	Size (bp)
1	1583104 (-397)	1740288 (–)	1.93	1.05	0.55	157184
2	1740321 (–)	1833472 (-29)	1.96	1.06	0.56	93151
3	2536448 (-139)	2786304 (-283)	1.94	1.06	0.56	249856
4	952320 (+181)	1077248 (+109)	2	1.06	0.54	124928
5	2048000 (-741)	2173952 (+211)	1.99	1.06	0.54	125932
6	45696 (-15)	70656 (-55)	1.97	1.06	0.552	24960
7	3142144 (+143)	3142656 (–)	1.97	0.98	0.49	512
8	3142912 (–)	3143072 (–)	2.19	1.11	0.57	160
9	3145728 (–)	3166208 (–)	1.91	1.03	0.55	20480
10	3166720 (–)	3166976 (0)	1.64	0.92	0.56	256
11	860160 (+957)	861698 (-5)	1.98	1.05	0.55	1536
12	3197889 (-15)	3200512 (+108)	1.93	1.06	0.57	2623

the reference and unknown sequences) can be detected and quantified by using NN and CNV-MM. Five sequences were generated from *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen with varying copies of two sets of randomly chosen genes (1022 bp and 8177 bp). The 1022 bp gene was extracted from 2538089-2539111, and the 8177 bp gene was extracted from 835334-843511 of the original sequence. All the copies were randomly inserted in the simulated sequences with 100% similarity. The copy numbers are designed as shown in Table 7.9. Paired-end, 300-bp simulated reads with 1% uniform error rate and 2% indel rate with the indel length determined by the geometric distribution (p = 0.3) were generated from each sequence with 50x coverage.

Sequence-1 Maps to Sequence-1

We have shown that CNV-MM can estimate the absolute copy numbers for repeats (not necessary CNVs). We also show that CNV-MM can detect copy number differences from both deletions and duplications. It is interesting to see when the query sequence is exactly

Table 7.9: Copy numbers in five simulated sequences. The copy numbers listed in the table include the original copy. For example, sequence-5 has one copy of gene-1 indicating that it is the original copy. Gene-1022 represents the 1022-bp-long gene while gene-8177 represents the 8177-bp-long gene.

Sequence	Gene-1022	Gene-8177
1	20	2
2	16	4
3	8	6
4	2	10
5	1	12

the same as the reference sequence and carry high copy numbers, whether CNV-MM can (1) have no false discovery and (2) correctly estimate the copy number in both sequences.

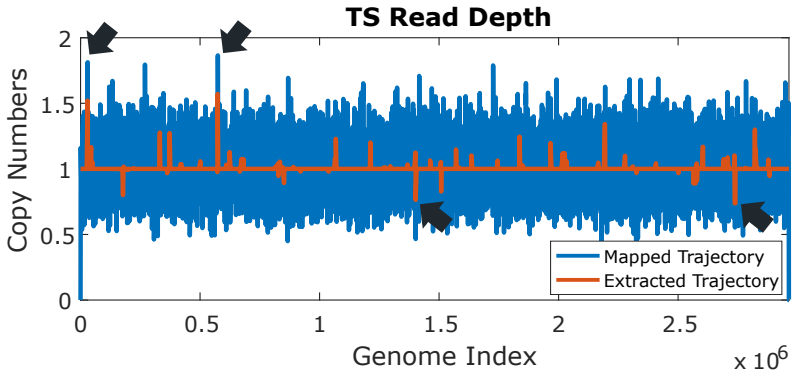

Figure 7.15: TS trajectories for Sequence-1 mapped to Sequence-1. The blue lines are the read depth obtained directly from the mapping results. The orange lines are read depth extracted by CNV-MM. Black arrows indicate false discovery.

Figure 7.15 shows the TS trajectory of sequence-1 maps to sequence-1. As indicated by the black arrows, there are four false discoveries. Two grouped $\sim$1000-bp duplications from 2.23 copies to 3.35 copies, and two unrelated deletions from 0.98 to 0.71 copy and 0.89 to 0.67 copy. None of the detected CNVs are originated from the inserted sequences. As in previous studies, the false discovery is most likely due to the uneven reads sampling.

When mapping the reads generated from sequence-1 to sequence-1, it is expected that both the binary and average trajectories show 20 copies of gene-1022 and two copies of

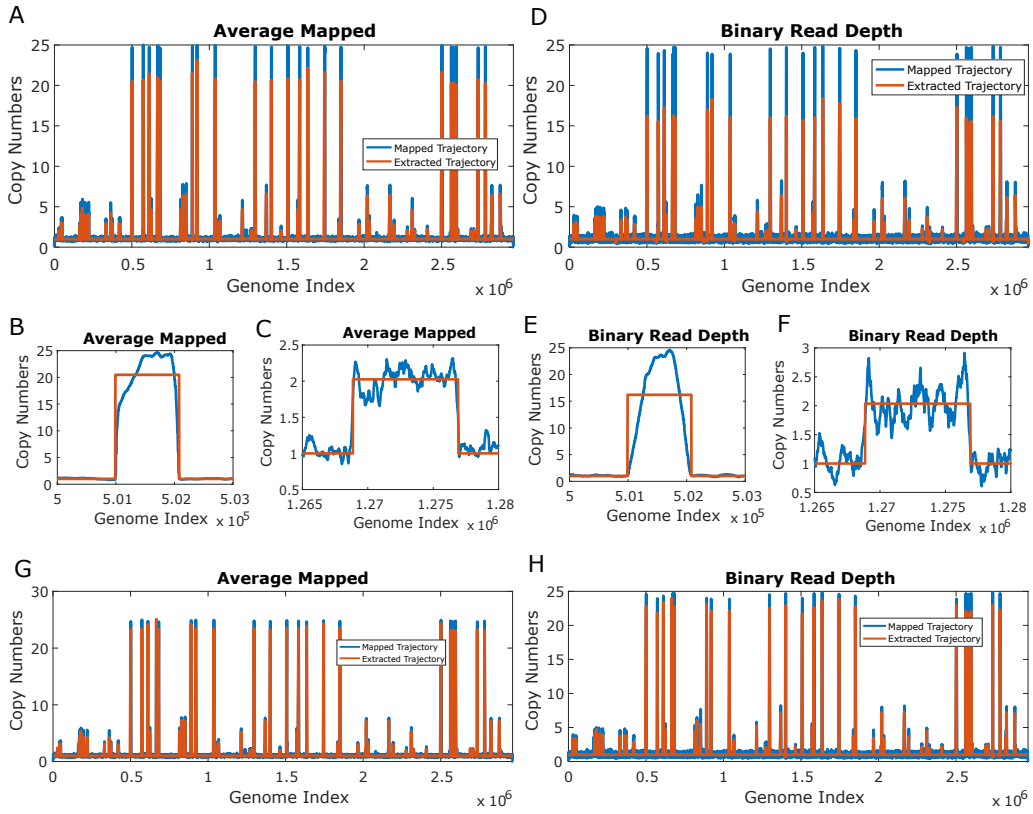

Figure 7.16: Mapping trajectories for Sequence-1 mapped to Sequence-1. (A) to (C) Average number of assignments trajectory. (D) to (E) Binary trajectory before boundary shift. (A) and (B) are the whole trajectories. Before boundary shift applied, the estimated copy numbers (orange curves) are inconsistent to the real copy numbers (blue curves). (C) and (E) zoom in to 500000-503000 for gene-1022. (D) and (F) zoom in to 1265000-1280000 for gene-8177. (G) Average trajectory after boundary shift. (H) Binary trajectory after boundary shift. After boundary shift, the orange curves better overlap with the blue curves. The blue lines are the read depth obtained directly from the mapping results. The orange lines are read depth extracted by CNV-MM.

gene-8711 (Although none of these are identified as CNV regions). The estimated average copy number is indeed around 20 (Fig. 7.16 A, orange curve). The extracted binary trajectory (Fig. 7.16 D, orange curve), however, does not recover the mapped copy number. This is because the binary (and TS) trajectories depend on the number of reads mapped to a nucleotide. As a result, instead of following a step function, the read depth for a CNV region increases gradually at the boundaries (Fig. 7.16 E). The average trajectory, on the other hand, does not depending onhow many reads are mapped to a given nucleotide. Therefore, the average number of mapping locations per read is the same. As a result, the trajectory

is less affected by the number of reads mapped, and the boundaries follow a step function (Fig. 7.16 B). For gene-8177 regions, the edge effect is not as severe as in gene-1022 regions due to the larger size of the gene (Fig. 7.16 D and F). As a result, the copy numbers are accurately calculated from average read depth of the regions defined by the estimated breakpoints.

The gradually increasing read depth in the binary and TS trajectories decrease the average copy number of a given region, even though the breakpoints determined from CNV-MM are mostly within 100-bp from the real boundaries (Table 7.10). To adjust for the edge effect, instead of calculating the read depth for the whole gene region determined by the estimated breakpoints, the read depth is calculated from $Breakpoint^{5'} + ReadLength$ to $Breakpoint^{3'} - ReadLength$. The read depths calculated with a boundary shift better estimate the real read depth for the gene-1022 regions with the blue and orange peaks more consistent with each other (Fig. 7.16 G and H). This shows that although any region longer than the read length can potentially be detected, the true copy number, however, can only be accurately estimated if the region is longer than at least two times the read length. For regions that are shorter than two times the read length, there might be a omission.

The copy numbers calculated from trajectories with boundary shift (Fig. 7.16 G and H), however, are 23.47 for the reference sequence and 22.65 in the reads donor sequence (which are both sequence-1) (Table 7.12). This implies that there are 22$\sim$24 copies instead of 20 in sequence-1. Since the analysis is done with grouping turns on, there are indeed 22 regions (not include gene-8177 and its replica) grouped together. These 22 regions include the 19 inserted sequences, the original sequence, and two unexpected sequences. After searching through the original sequence, we found that these two unexpected sequences are naturally-occured, repeated regions of gene-1022. There is one more sequence on *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen is partially related to gene-1022, however, was not picked up in this analysis. Nevertheless, it does seem to contribute to the estimated copy numbers for both the reference and query sequence. Due to this naturally-

Table 7.10: Repeated regions in sequence-1 and sequence-1 mapping results. The regions are presented in the sequence-1 index. The first two columns are the repeated regions breakpoints. The last two columns are the mapping results from mapping reads from sequence-1 mapped to sequence-1. These regions are not defined as true CNVs since they have the TS copy numbers close to one. The numbers in the parenthesis are the distances from the real breakpoints (negative: toward 5'-end.). CN: copy number. The last two rows are regions grouped by CNV-MM and are the intrinsic duplications of region 2562619-2563640.

Group	5'-end	3'-end	Size	5'-end (map.)	3'-end (map.)	TS CN
1	501025	502047	1022	500992 (-33)	502080 (+33)	1.06
1	572651	573672	1022	572608 (-43)	573696 (+24)	0.98
1	611764	612785	1022	611840 (+76)	612832 (+47)	1.02
1	666140	667161	1022	666112 (-28)	667184 (+23)	0.98
1	672278	673309	1022	672248 (-30)	673320 (+11)	1.07
1	681588	682609	1022	681560 (-28)	682656 (+47)	1.08
2[b]	841467	849644	8177	842752 (+1285)[a]	849664 (+20)	1.01
1	890283	891304	1022	890368 (+85)	891360 (+56)	1.02
1	920232	921254	1022	920192 (-40)	921280 (+26)	0.98
1	1037552	1038573	1022	1037520 (-32)	1038592 (+19)	0.98
2	1268768	1276945	8177	1268864 (+96)	1276928 (-17)	1.01
1	1296140	1297161	1022	1296120 (-20)	1297216 (+55)	0.96
1	1400546	1401568	1022	1400512 (-34)	1401600 (+32)	1.12
1	1583249	1584270	1022	1583230 (-19)	1584320 (+50)	1.01
1	1635137	1636158	1022	1635200 (+63)	1636170 (+12)	1.10
1	1744425	1745446	1022	1744640 (+215)	1745504 (+58)	1.06
1	1849434	1850455	1022	1849408 (-26)	1850496 (+41)	1.03
1	2501532	2502553	1022	2501632 (+100)	2502592 (+39)	0.97
1[b]	2562619	2563640	1022	2562560 (-59)	2563648 (+8)	1.00
1	2578005	2579026	1022	2577952 (-53)	2579040 (+14)	1.04
1	2735647	2736668	1022	2735616 (-31)	2736704 (+6)	1.12
1	2782519	2783540	1022	2782464 (-55)	2783584 (+44)	1.01
1[c]	—	—	—	1507616	1508672	1.05
1[c]	—	—	—	2595328	2596448	0.96

a: The region was divided into two subregions. The other one is 841728-842240. b. These are the original genes that exist in *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen, while the rest are randomly inserted duplications. c:Pre-existing duplications of gene-1022 (2562619-2563640) in *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen

occurred, repeated regions of gene-1022, the estimated copy numbers increase for all five simulated sequences.

All Mapping Results

Table 7.11: Test results for gene-8177 Rows: reference sequences. Columns: reads donor sequences. CPs: copies. Numbers of copies indicated in the first row and column are the number of inserted copies.

Ref\Reads	Seq-1	2 CPs	Seq-2	4 CPs	Seq-3	6 CPs	Seq-4	10 CPs	Seq-5	12 CPs
Seq 1	2.04[a]	2.04[b]	2.04	4.02	2.08	6.35	2.07	10.69	2.07	12.58
2 CPs	1.01[c]	7488[d]	2.02	8192	3.04	8064	5.24	8064	6.18	8064
Seq 2	4.01	2.05	4.05	4.02	4.16	6.40	4.09	10.66	4.09	12.60
4 CPs	0.53	8192	1.04	8176	1.59	8192	2.69	8208	3.20	8208
Seq 3	6.01	2.01	6.05	3.95	6.22	6.35	6.14	10.62	6.14	12.56
6 CPs	0.35	8576	0.68	8224	1.06	8160	1.80	8240	2.13	8232
Seq 4	10.10	2.03	10.14	4.03	10.42	6.40	10.27	10.69	10.26	12.62
10 CPs	0.21	8525	0.42	8218	0.64	8290	1.08	8163	1.28	8250
Seq 5	12.02	2.03	12.07	4.01	12.26	6.15	12.25	10.56	12.27	12.53
12 CPs	0.18	8037	0.35	8223	0.52	8731	0.90	8365	1.06	8156

a. Estimated average copy number in the reference b. Estimated average copy number in the reads donor sequence. c. Estimated average copy number ratio. d. Average gene size (bp). All the numbers in different cells follow the same structure.

The five sequences serve both as the reference and the reads donor sequences. The results of gene-8177 are shown in Table 7.11. There are 2, 4, 6, 10, and 12 copies of gene-8177 for sequence 1, 2, 3, 4, and 5. Using CNV-MM and group the related sequence together, the average estimated copy numbers are calculated. For Table 7.11, the four numbers in each cell are: the average copy number in the reference sequence (upper left), the average copy number in the reads donor sequence (upper right), the copy number ratio (lower left) and the average region length (lower right). As shown in the table, the estimated copy numbers indeed reflect the real copy number in each sequence. Both duplications and deletions can be detected. The estimated regions sizes are close to the true region size: 8177 bp.

As mentioned in the previous subsection, the real copy numbers in all five pseudo-sequences become complicated since the original sequence of the selected gene-1022 intrinsically has three extra repeats with different similarities. The CNV-MM results are listed

Table 7.12: Test results for gene-1022 Rows: reference sequences. Columns: reads donor sequences. For each reference sequence, two sets of conditions are presented and separated by dashed line. The first two rows in each reference sequence are the average results of all regions in a group. The last two rows in each reference sequence are the average results of only the inserted regions are included. "—" represents the results are not significantly different from the first condition. Numbers of copies indicated in the first row and column are the number of inserted copies.

Ref\Reads	Seq-1	20 CPs	Seq-2	16 CPs	Seq-3	8 CPs	Seq-4	2 CPs	Seq-5	1 CP
	23.57[a]	22.65[b]	22.07	16.04	22.50	8.36	21.2	3.60	21.3	2.79
Seq 1	1.03[c]	1041[d]	0.78	1389	0.40	1311	0.17	1484	0.14	1341
20 CPs	—	—	—	—	—	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—
	19.24	20.41	19.25	19.27	16.24	7.15	18.16	3.67	18.04	2.82
Seq 2	1.13	1052	1.06	1028	0.51	1573	0.21	1384	0.16	1267
16 CPs	—	—	—	—	19.54	9.56	—	—	—	—
	—	—	—	—	0.52	1076	—	—	—	—
	9.30	14.69	10.65	17.55	9.34	8.21	8.90	5.39	8.65	4.47
Seq 3	1.62	1259	1.76	1063	0.90	1250	0.72	1310	0.63	1283
8 CPs	10.67	20.82	—	—	10.90	9.76	10.57	4.13	10.10	3.04
	2.02	1040	—	—	0.93	982	0.40	1079	0.32	1063
	3.62	9.8	3.36	8.59	3.66	6.22	4.28	4.96	4.07	3.48
Seq 4	2.64	1524	2.40	1851	1.73	1414	1.14	992	0.89	1040
2 CPs	4.22	19.4	4.12	16.81	4.19	8.74	—	—	—	—
	4.94	1046	4.36	1024	2.19	1040	—	—	—	—
	3.20	7.06	2.92	6.08	3.24	4.93	3.25	3.48	3.11	2.76
Seq 5	2.33	1613	2.06	2099	1.56	1664	1.11	1024	0.89	1024
1 CP	3.27	15.47	3.22	13.65	3.40	7.72	—	—	—	—
	5.11	1024	4.48	1024	2.37	1024	—	—	—	—

a. Estimated average copy number in the reference b. Estimated average copy number in the reads donor sequence. c. Estimated average copy number ratio. d. Average gene size.

in Table 7.12. Two sets of results are presented for each reference sequence. The first one is the average results for all grouped regions, including the intrinsic repeated regions. The second is the average results for only the inserted repeats. Because the intrinsic repeated regions only partially resemble the inserted sequence, the average estimated copy numbers are generally smaller than inserted-sequence-only copy numbers. Although it is possible to separate the mapping regions of the intrinsic repeated regions from the inserted repeats, the reads generated from the intrinsic repeats are still included. As a result, the estimated

copies, 21 23, 16-19, 8 10, 4 5, and 2 3 for sequence 1 to 5, tend to be higher than the expected values (the diagonal cells in Table 7.12). Despite the extra copies provided by the intrinsic repeats, the estimated copy numbers maintain internal consistencies over different reference-query sequences pairs except when using sequence-5 as the reference sequence, where the number of copies in the query sequences tend to be underestimated. Full list of detected CNVs are listed in Appendix Tables E.18 to E.42.

7.4 Comparing CNV-MM with CNVnator

One of the most widely use CNVs detectors is CNVnator.[42, 116, 230] CNVnator takes the MAQ mapping results, which randomly chooses a mapping location for a multi-read (read has multiple mapping locations). As explained in Figure 7.1, although the correct read depth might be reconstructed when map multi-reads to random positions, false duplications, and deletions can occur. To compare the performance of CNV-MM and CNVnator, the mapping results from NN (in SAM format, see Chapter 2) are analyzed by both methods. To mimic the MAQ output, all the multi-reads are given zero quality scores (please see Chapter 2 for details about quality scores) for CNVnator to recognize the multiple aligned reads. Instead of mapping reads to all possible locations as mrFAST and extract the absolute copy numbers of the query sequence (while the output entries might not be real CNV regions), the output of CNVnator is expected to be the copy number of "real" CNV regions.

As in the similarity test in Section 7.3.1, the genome sequence from 1542312 to 1543124 of *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen was taken as the mother gene (gene-A, 812bp) and inserted at 1543144 (to 1543956) to generate pseudo-sequence-2 with 100% similarity. To generate pseudo-sequence-3, an extra piece of the gene-A is inserted at 1545653 (to 1546465, also 100% similarity). Simulated reads with 1% uniform error rate, and 2% indel rate with the indel length follows a geometric distribution were generated from Seq-2 (two copies of gene-A) and were mapped to Seq-3 (three copies of gene-A).

The reconstructed trajectories are shown in Figure 7.17. When using CNV-MM, not

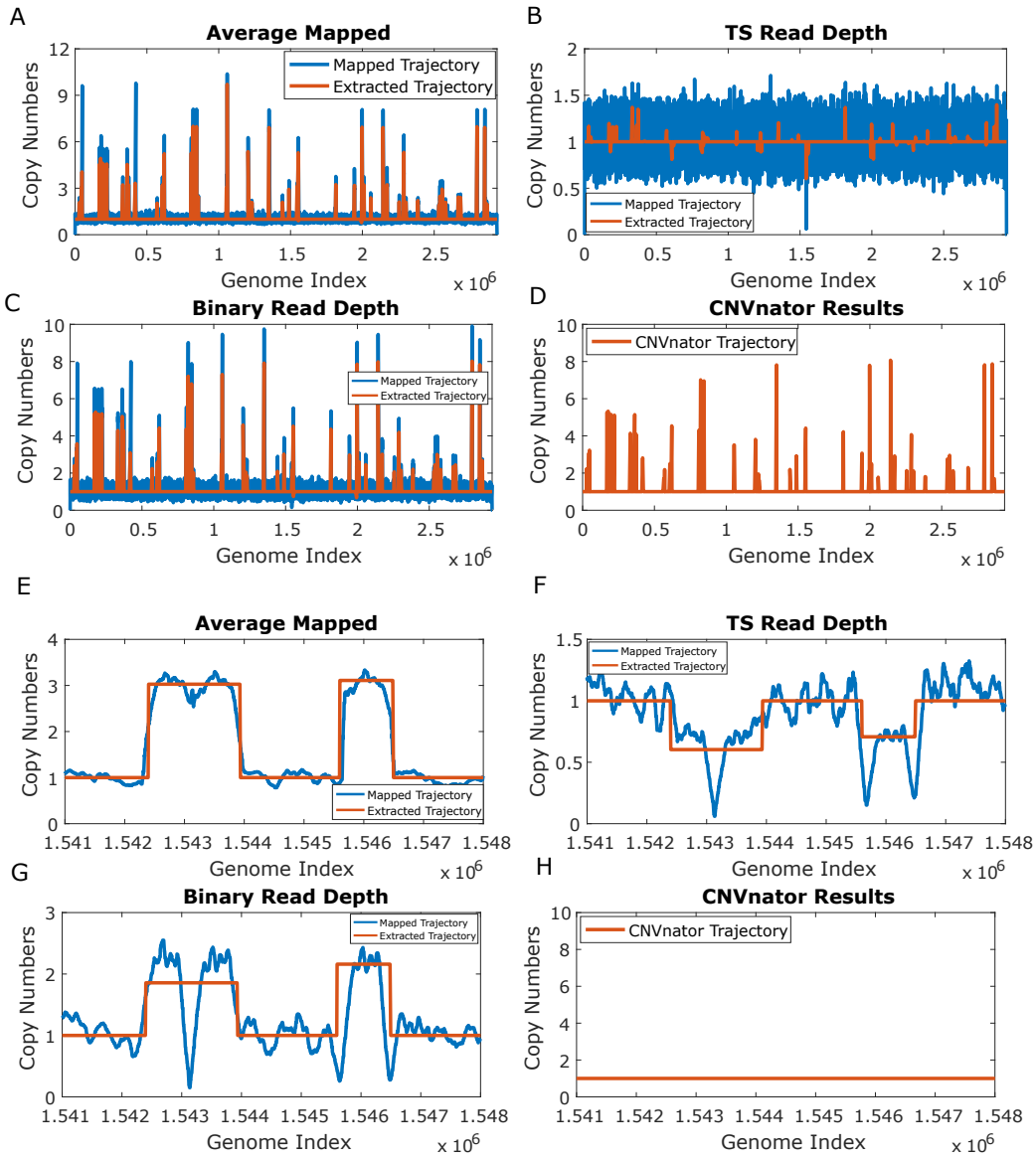

Figure 7.17: Mapping trajectories for CNV-MM and CNVnator. (A) to (D) the whole trajectories while (E) to (G) are trajectories zoomed into 1541000-1548000. (A) to (C) and (E) to (G) are CNV-MM results. (A) and (E) are average number of assignments trajectories. (B) and (F) are test statistics trajectories. (C) and (G) are binary trajectories. The blue lines are the read depth obtained directly from the mapping results. The orange lines are read depth extracted by CNV-MM. (D) and (H) are trajectories reconstructed by the results of CNVnator.

only the absolute copy numbers of the reference (Fig. 7.17A) and query sequence (Fig. 7.17 C) are obtained, but also the identification of true CNV regions with the TS trajectory (Fig. 7.17 B). Although "real" CNV regions should be reported with CNVnator, the trajectories

using the CNVnator output shows that there are many of false discoveries. Trajectories zoom into 1541000-1548000 are shown in Figure 7.17 E to H. This region is where the three copies of gene A located in Seq-3 (reference sequence). While CNV-MM successfully identifies the CNV regions (Fig. 7.17 E to G), CNVnator misses this only true CNV genes (Fig. 7.17 H).

Although CNV-MM does detect gene-A from 1542312 to 1543124, and from 1543144 to 1543956, these two copies, which are only 20-bp apart, are merged (Fig. 7.17 E to G). This contributes to a lower estimated query copy number (Table 7.13) for this merged region. Nevertheless, with the grouping function on, the three gene-A are identified as related. The group average copy numbers can, therefore, be calculated and are 3.07 for the reference sequence, 2.01 for the query sequence and 0.66 for the TS ratio. Since the reference sequence is Seq-3 (three copies), and the query sequence is Seq-2 (two copies), CNV-MM accurately determines the copy numbers of the CNV region. As for false discovery rates, CNVnator reports 53 false discoveries while there are only 11 false discoveries in

Table 7.13: CNVs report from CNV-MM CNV size is in base pairs (bps). Negative group numbers indicate deletions. "map" represents the indices determined by CNV-MM and "CN" means "copy number". For TS copy number, the copy number ratio is measured instead.

Group	5'-end (map.)	3'-end (map.)	Ref CN	Query CN	TS CN	Size (bp)
1	331712	333568	3.22	4.29	1.37	1856
1	372672	374528	3.22	4.29	1.35	1856
1	1815296	1817088	3.24	4.36	1.37	1792
2	821760	823296	7	7.21	1.1	1536
2	845120	846592	6.97	6.8	1.04	1472
2	1349760	1351328	6.92	7.91	1.2	1568
2	1997952	1999488	6.94	7.86	1.19	1536
2	2143744	2145280	6.92	7.98	1.19	1536
2	2798048	2799488	6.97	8.02	1.21	1440
2	2852128	2853632	6.93	7.83	1.16	1504
-1	1542400	1543936	3.03	1.86	0.603	1536
-1	1545600	1546496	3.11	2.16	0.707	896
-2	1994240	1994496	0.777	0.674	0.832	256

CNV-MM. Our data shows that by properly accounting for mapping multiplicity and calculating three read depth trajectories, CNV-MM can better detect CNVs than the popular algorithm CNVnator. A full list of CNV reports of CNVnator can be found in Appendix Table E.43.

7.5 Copy Number Detection – Real Reads

A. baumannii is one of the major pathogens around the world and has becom resistant to many different antibiotics.[231, 232] Resistance has been acquired through the incorporation of mobile genetic elements including insertion sequences (IS), resistance islands, and plasmids.[233, 234] *A. baumannii* is naturally resistant to β-lactam antibiotics due to the intrinsic β-lactamase and OXA-51 carbapenemase (bla_{OXA-51}).[234–236] Several IS, including ISAba1, ISAba2, ISAba3, ISAba4, and ISAba10 have been shown to be associated with the expression of the OXA-51-like enzymes.[234, 235, 237–240] Although β-lactam sensitive strains, such as *A. baumannii* strain ATCC 17978, have both the IS and OXA-51-like enzymes encoded in the chromosome, the expression levels of the OXA-51-like enzymes are not high enough to provide resistance to carbapenem. The expression levels of OXA-51-like enzyme is significantly increased when the IS is transposed to be upstream of bla_{OXA-51}. This upstream IS provides an extra promoter for gene bla_{OXA-51} and thus increases the expression level of OXA-51-like enzymes. [234, 235, 238, 240]

To further test CNV-MM, paired-end, 251-bp real short reads data, SRR2558867, from multidrug-resistant, *A. baumannii* clinical isolate was downloaded from the sequence reads archive (`http://www.ncbi.nlm.nih.gov/sra`). The reads were aligned to two reference sequences, *A. baumannii* strain ATCC 17978 (sensitive strain) and *A. baumannii* strain MDR-ZJ06 (multidrug-resistant strain), downloaded from NCBI (`ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/`).

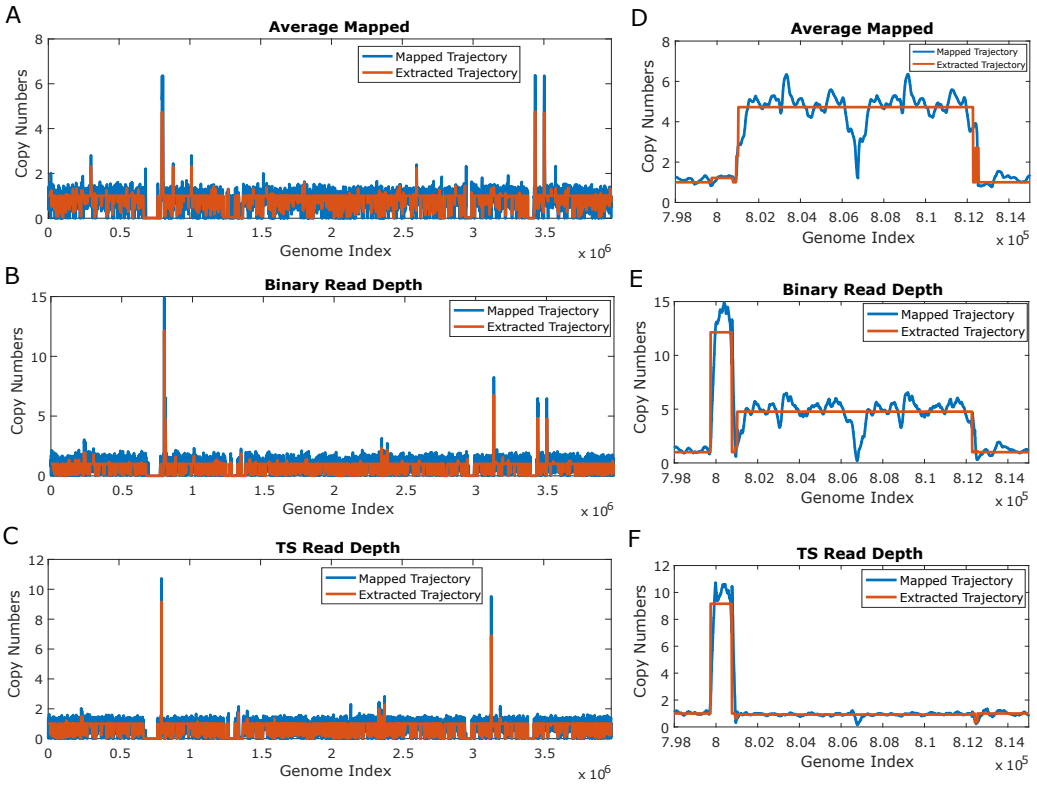

Figure 7.18: Mapping trajectories for SRR2558867 mapped to *A. baumannii* strain ATCC 17978. (A) to (C) the whole trajectories. (D) to (F) zoom in to 798000 - 815000. (A) and (D) Are average number of assignments trajectories. (B) and (E) are binary trajectories. (C) and (F) are test statistics trajectories. The blue lines are the read depth obtained directly from the mapping results. The orange lines are read depth extracted by CNV-MM.

7.5.1 *A. baumannii* strain ATCC 17978 as Reference Sequence

The coverage, calculated from the sequence size of *A. baumannii* strain ATCC 17978, is 135x. The real coverage, however, is 74x since the test results are too large (less confident) for some of the reads, either due to sequencing errors or unique sequence from the reads donor sequence. The reconstructed trajectories corrected for GC-content (details in Chapter 2) are shown in Figure 7.18.

There are two major duplication peaks: 799744-800768 and 3128448-3128704. The sequences corresponding to these two regions were exported and search with BLAST.[38, 241] The latter is a hypothetical protein with unknown function. The 799744-800768 is the ISAba1 insertion sequence, including the two putative transposases and OXA-95 (belongs

to the OXA-51 cluster) promoters (both -35 and -10) derived from ISAba1. In the trajectories, it is estimated to have about 1.27 copy of the ISAba1 in ATCC 17878, $\sim$14 copies in the reads donor sequence, and the ratio is about $\sim$10 times higher in the reads donor sequence (Fig. 7.18 D to F). The sequence between 800953-812288 encodes the 23S, and 16S ribosomal RNA and both sequences have $\sim$4 5 copies. Thus, the ratio is one in the TS trajectory. Full list of deletions and duplications can be found in Appendix Tables E.44 and E.44.

Although the ISAba1 insertion is $\sim$10 times higher in the short reads sequence, the bla_{OXA-51} gene (c1765468-c1766292), which is more than 950 kb downstream on the complemented strand for ATCC 17978, does not have a higher number of copies. Both the reference and the query sequences have one copy of bla_{OXA-51}. Since the read depth-based method can only detect the copy number differences but not the transposition of genes, it is unknown that whether the extra ISAba1 copies are indeed inserted upstream of the bla_{OXA-51} gene in the reads donor sequence and thus increase the expression levels of OXA-51-like enzymes. Although the higher copy numbers of ISAba1, ares very encouraging, we still need more analysis such as de novo assembly to verify whether the high copy numbers of ISAba1 directly contribute to the high antibiotic resistance.

7.5.2 *A. baumannii* strain MDR-ZJ06 as Reference Sequence

The same short reads file is mapped to another reference sequence, *A. baumannii* strain MDR-ZJ06.[242]. Again, some of the reads are not mappable under the current threshold. As a result, the real coverage is 96x instead of 135x. Compared to *A. baumannii* strain ATCC 17978, *A. baumannii* strain MDR-ZJ06 is a closer strain to the short reads sequence since the read depth baselines are much smoother with far fewer deletions (Fig. 7.19 A to C). Although there are several multiple copies regions in both the average and binary trajectories, most of them share the same copy numbers in both trajectories. The major copy number differences shown in (Fig. 7.19 C) is peak 3212032 - 3212288, where the reference

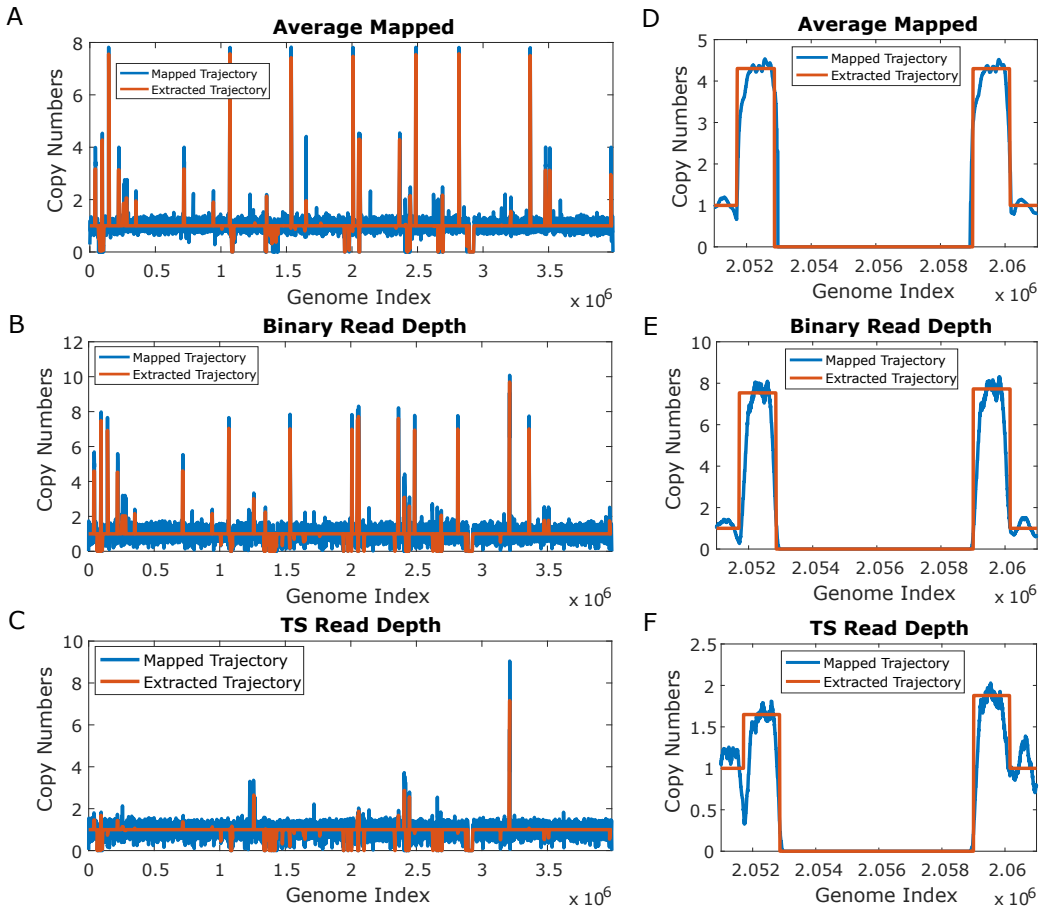

Figure 7.19: Mapping trajectories for SRR2558867 mapped to *A. baumannii* strain MRD-Zj06. (A) to (C) the whole trajectories. (D) to (F) zoom in to 2051000 - 2061000. (A) and (D) are average number of assignments trajectories. (B) and (E) are binary trajectories. (C) and (F) are test statistics trajectories. The blue lines are the read depth obtained directly from the mapping results. The orange lines are read depth extracted by CNV-MM.

has 1.56 copies and the reads donor sequence has around 9.69 copies. By mapping these regions with BLAST, this peak corresponds to biofilm-associated protein. The peaks (TS trajectory) around 1250000 and 2410000 are several hypothetical proteins.

There are several deletions and duplications detected by CNV-MM (Appendix Table E.47 and E.46). When determining the CNV regions with grouping, two groups of duplications and one groups of deletions are found. The first duplication group (39424-44800, 218112-223744, and 716544-721920) has $\sim$3 copies in the reference and $\sim$4.5 copies in the reads donor sequence. This group encodes 16S, 23S, 5S, and tRNA. The deletion group (including regions: 3474944-3480576, 3507328-3512960, and 3974144-3980544)

has $\sim$3 copies in the reference and 1$\sim$2 copies in the reads donor. The complementary strand of those regions also encode rRNA and tRNA. The second duplication group (94464-95712, 2051712-2052960, 2058975-2060160, and 2363392-2364647) encodes the IS4 family transposase (ISAba1 is part of the IS4 family) open read frame (ORF) 1 and 2. However, the reference sequence has four copies while the reads donor has 5$\sim$6 copies. Since duplication group 1 and deletion group 1 encode the same gene but on different DNA strands and the copy numbers compensate each other, the CNV differences observed here might be because of the different distribution of the same amount of genes on the complementary strains.

For duplication group 2, the 2051712-2052960 and 2058975-2060160 regions flank a deletion region as shown in Figure 7.19 D to F. Since no reads is mapped to the region, the read depth is zero for all three trajectoreis. This deletion region is the Tn2009 transposon that encodes the DEAD/DEAH box helicase-like protein, ATPase, and the β-lactamase OXA-23.[242] Since there are at least eight distantly related OXA-type carbapenemases,[236, 243] it is possible that OXA-23 (reference OXA type) and OXA-51 (known query sequence OXA type from mapping to ATCC 17978) do not share enough of sequence similarity to be mapped by NN and thus a deletion is shown.

7.6 Conclusions

As most of the aligners can not properly account for mapping multiplicity without having to know what sequence to look for, the subsequent CNV detectors are not developed for handling repeated genes in the reference sequence. In this chapter, it has been shown that NN can detect all mapping locations while maintaining high mapping accuracies with different types of reads errors.

To use the mapping results from NN to detect CNVs, CNV-MM is developed. By calculating the read depth trajectories differently, CNV-MM can estimate the copy numbers in the reference and the query sequence from single short reads file. An independent

copy number ratio can also be calculated to confirm the CNV events. We also demonstrate that CNV-MM can be applied to different read length from 36-bp to 300-bp, although it performs better with read longer than 100-bp. It does not have a theoretical upper limit for the CNV size, and it can detect the CNVs as short as the read length. It is robust for a wide range of copy number pairs in the reference and query sequence. Although more simulations have to be done to determine the detection limit for copy number calculations. The NN-CNV-MM method is also applied to real short reads data, and it shows that the multidrug-resistant clinical *A. baumannii* isolates have 9$\sim$10 times higher copy of carbapenemase-associated insertion sequence (ISAba1). Although more analyses are needed to confirm the correlation between the high copies of ISAba1 and higher MIC, this analysis demonstrates one of the possible applications for CNV-MM.

CHAPTER 8
CONCLUSIONS AND OUTLOOK

Data analysis is an integral part of scientific research. From mean and standard deviation to clustering and correlation, different information can be extracted from a given data set. With advancing technology, the size and the complexity of data set have grown. To adapt to the ever-changing data structures, new quantitative methods are needed.

The data acquired from flow cytometry contains hundreds of thousands of data points with high dimensionality. Although methods have been developed to quantify the differences among cytometric data, PB-sQF, by calculating the Euclidean distance between the adaptively binned signatures, is the only one that calculates the linear distance while being scalable with multi-dimensions. Using PB-sQF to characterize the differences between the cytometric data of antibiotic-treated bacteria and the control, we have confirmed the relation between ROS generation and bactericidal antibiotics. A rapid post-blood culture AST (four hours processing time) has been developed based on the antibiotic-induced scattered light changes. With only one hour incubation with antibiotics at the MIC, clear cytometric signal changes are observed for the antibiotic-treated, susceptible bacteria. PB-sQF shows significant changes in test statistics for susceptible bacteria while the resistant strains have constant test results. In Chapter 4, a pre-blood culture fast AST, (FAST), has been developed using saponin to remove the blood cells, which enable bacteria to grow in optimal conditions thereby reducing the incubation time from more than two overnight cultures (blood culture and AST) to five hours (including AST). PB-sQF can again select the effective treatments by analyzing the scattered-light pattern changes for each bacteria-antibiotic combination.

Although we have shown that rapid ASTs can be built from using PB-sQF to analyze cytometric data, we have encountered some difficulties with gram-positive bacteria includ-

ing the weak β-lactamase generating *S. aureus* in post-blood culture and slow growing *S. aureus* and *Streptococcus spp.* clinical isolates for pre-blood culture. For the post-blood culture test, the antibiotic-induced scattered-light shift was recorded. The MIC, however, is defined as the minimum antibiotic concentration that "inhibits bacterial growth". As a result, measuring the scattered-light shift is not the same as growth inhibition. As a result, consistent results should be obtained if growth inhibition is measured instead. For the pre-blood culture, different broth other than CAMHB should be used. In the CLSI guidelines, CAMHB with 2$\sim$5% lysed horse blood is the standard AST broth for both *S. aureus* and *Streptococcus spp.*. Since the microdilution AST depends on the turbidity due to bacterial growth, both bacteria must have decent growth rates with CAMHB with 2$\sim$5% lysed horse blood. Being able to handle gram-positive bacteria will greatly increase the generality of our method.

With a few modifications, PB-sQF can also be used in genome sequence analysis. By converting the character string data into histogram data with coordinates and counts, PB-sQF can calculate the distance between two genome sequences. This distance allows one to perform sequence typing, build a phylogenetic tree and track down outbreak strains. The complete sequence, however, is not always obtainable since the raw data from NGS is millions of unassembled short reads. These short reads are mapped by short reads aligners. Current short reads aligners, however, have difficulties mapping long reads with high error rates, and assigning reads to multiple mapping locations. Since PB-sQF is a distance-based aligner, long reads, high error rates and mapping multiplicity can easily be solved by setting a test threshold. Although PB-sQF has better error tolerance than most of the tested short reads aligner, NN, which is also a Euclidean distance-based statistic, has even better high error tolerance and is computationally faster. Since NN has only been used in sequence typing with 16S RNA, we adapted the short reads mapping procedure developed for PB-sQF to use NN. We have shown that while NN is robust against reads errors, it also finds all the mapping locations in a reference sequence with repeated genes.

Properly accounting for mapping multiplicity is important in finding CNV between sequences. Current CNV detectors, however, only build the read depth trajectory with uniquely mapped reads. To fully utilize the NN mapping results, we have developed CNV-MM to calculate read depth from all valid assignments of all the reads. By calculating the read depth differently, read depth trajectories proportional to the copy number in the reference sequence, the copy number of the query sequence, and the copy number ratio can all be obtained. The NN-CNV-MM methods can analyze data with different read length, CNV sizes, and copy numbers of genes. By applying NN-CNV-MM on real short reads data from multidrug-resistant *A. baumannii* clinical isolates, we found that the insertion sequence, ISAba1, which is associated with the carbapenem-resistant gene bla_{OXA} by increasing the expression levels of carbapenemase, increased $\sim$10 times over that in the sensitive strains.

While we have shown that NN-CNV-MM is great in CNV detections for references with repeated genes, its performance can be improved. First, CNV-MM should adapt a more sophisticated segmentation process such as the hidden Markov model, mean-shift algorithm or shifting level model. The current procedure for segmentation is to determine the copy number transition points by the derivative of a read depth trajectory. This method, however, might encounter some problems when the coverage is low even with wavelet denoising. For bacterial data, the coverage is normally not a problem since the bacterial genome is small. To adapt our method to human genomes, however, coverage is an important issue. Second, the NN mapping is time consuming even though we have narrowed down the search space by pre-calculating the distances between the library reads and control reads. Since this pruning search space process, which only has 50 test statistics calculations, is slower than the test statistics calculations between the unknown reads with all the candidate library reads, which could have 20 times more calculation, a more efficient memory management can greatly speed up the searching process. Third, both NN and CNV-MM are written in MATLAB, to make the codes more universal, memory efficient, and faster, the algorithms should be transferred to open source software like C++.

Despite there being possible improvements for both the fast AST with PB-sQF and the CNV detections with NN-CNV-MM, this dissertation has shown the potential of using Euclidean distance-based statistics to extract useful information from two very distinct types of data. While the improvements can greatly increase the impact of our methods in both fields, these distance-based methods can be applied to other types of data and can potentially be useful in many more fields.

Appendices

APPENDIX A

WEIGHTS OF TS TRAJECTORY

In the TS trajectory, a read depth from a single read is divided into several valid alignments by the weights defined in equation 2.22. The read depth of alignment-1 is $\frac{Weight_i}{\sum_{j=1}^{N^i_{assignments}} Weight_j}$. Ideally, a read depth should be distributed as follows. When the test statistic is small, this alignment has a higher probability to be the major read contributer. Therefore, it should have a larger share of the read depth. On the other hand, the read depth of an alignment should be small when another alignment has a small test statistics. Also, a read depth should be evenly distributed for all alignments when the test statistics are the same.

Instead of the inverse of the test statistics, the weights determined by $\sqrt{2}-TS$ was also considered since the maximum test statistic value is $\sqrt{2}$ as discussed in Section 2.2 (Fig. A.1 A). To understand the differences between the two weights, two alignments with test statistics ranging from 0 to 2 were generated. In Figure A.1 B, the weights were calculated from $2-TS$ and the read depth for alignment-1 is $\frac{\sqrt{2}-TS_1}{(\sqrt{2}-TS_1)+(\sqrt{2}-TS_2)}$. As a result, the read depths for alignment-1 are zeros when the test statistics of alignment-1 equals to two and the read depths goes to one when the test statistics of the alignment-2 are two. However, it does not follow the idea behaviors as mentioned above.

When weights are calculated from the inverse of test statistics, the read depth for alignment-1 is $\frac{\frac{1}{TS_1}}{(\frac{1}{TS_1})+(\frac{1}{TS_2})}$. Since the test statistics are set as 10^{-4} when zero occurs to avoid division by zero, the read depth of alignment-1 is close zero when the test statistics of alignment-2 is zero and the read depth of alignment-1 is close to one when the test statistics of alignment-1 is zero. The read depth is $\frac{1}{2}$ when test statistics of both mapping results are zero. The read depth follows the ideal behavior when the weights are depended on the inverse of test statistics and thus was taken in this study (Fig. A.1 C).

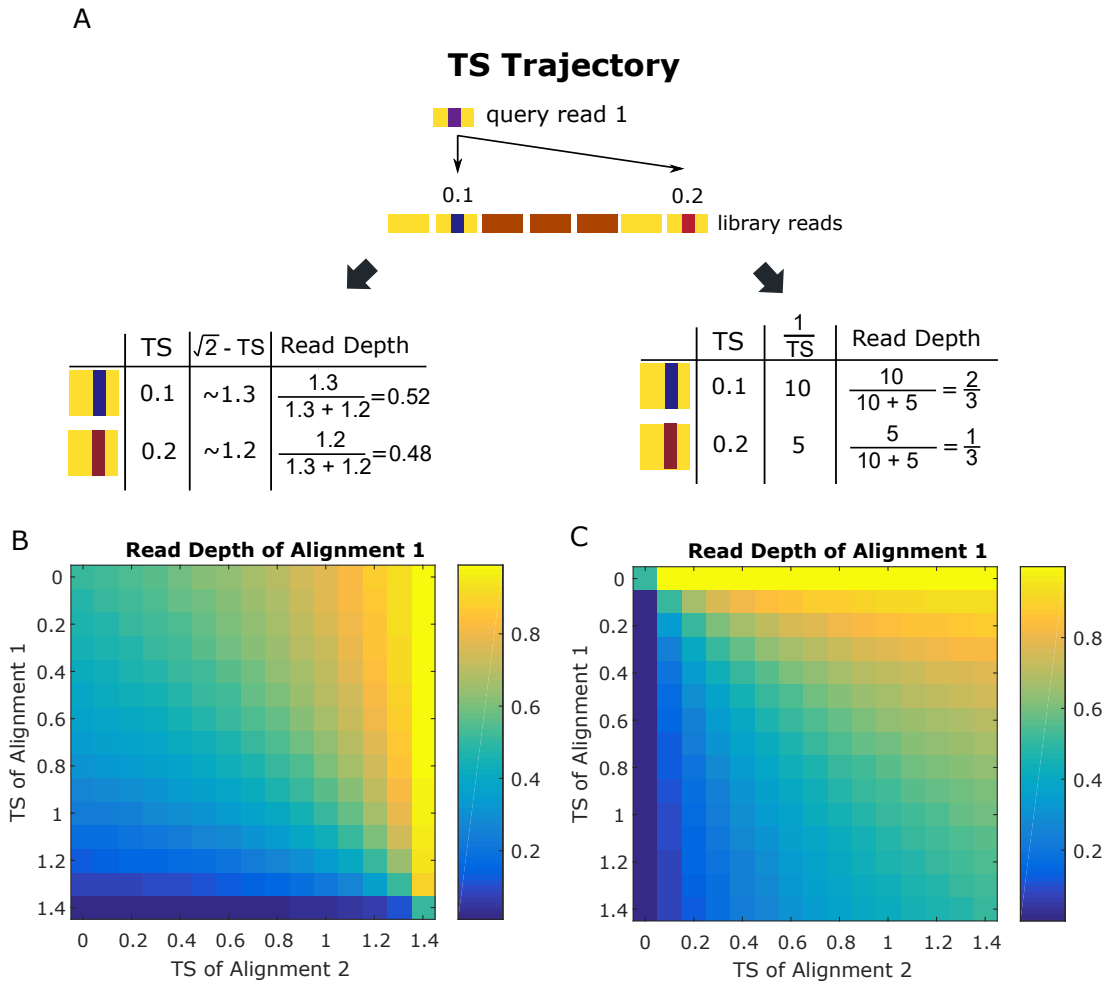

Figure A.1: Read depth with different test statistics weights. (A) the $\sqrt{2} - TS$ weights (left) and inverse test statistics weights (right). (B) Read depth of mapping result-1 with weights calculated as $\sqrt{2} - TS$. (C) Read depth of mapping result-2 with weights were calculated as the inverse of test statistics.

APPENDIX B

SUPPORTING INFORMATION FOR CHAPTER 3

This chapter contains the complete cytometry data for Chapter 3, including data of lab-strain and *E. coli* clinical isolate, lab-strain *P. aeruginosa*, lab-strain *K. pneumoniae*, *A. nosocomialis* clinical isolate and *S. aureus* (MSSA and MRSA).

All the flow cytometry data presented here were labeled with MH-IR786. For the scatter 2D plots, the pseudocolor plots are the paired-control, the no-antibiotic data, for each antibiotic-strain. The contours are the antibiotic-treated data with the antibiotic concentration indicated otherwise.

B.1 MIC Tables

Table B.1: MIC (μg/mL) for *E. coli* and *P. aeruginosa*.

MIC	Pen G	Amp	Nor	Cip	Kan	Tet	Ery	Azi	Gen
E. coli ATCC 33456	32	100	0.125	0.016	8	1	150	8	2
E. coli Mu14S	> 5000	—	—	—	—	> 8	—	—	4
P. aeruginosa ATCC 27853	—	512	2	—	1024	16	—	—	—

Table B.2: *S. aureus* MIC (μg/mL) for each antibiotic-strain combination.

MIC	Pen G	Oxa	Van	Azi
S. aureus ATCC 25923	0.0625	0.25	0.2	0.1
S. aureus ATCC 29213	2-4	0.5	1	1
S. aureus ATCC 43300	8-16	32	2	> 256

Table B.3: MIC (μg/mL) for textitK. pneumoniae and *A. nosocomialis*.

MIC	Amp	Nor	Cip	Kan	Tet	Ery	Azi	Gen	Cef
K. pneumoniae ATCC 700603	> 2000	—	0.5	—	16	256	64	8	8
A. nosocomialis strain M2	> 1024	4	$\frac{1}{4} - \frac{1}{2}$	8	1	—	4-8	4	16
A. nosocomialis strain M2-4B	> 1024	32	2-4	256	1	—	2	2	32

B.2 Full Cytometric Data

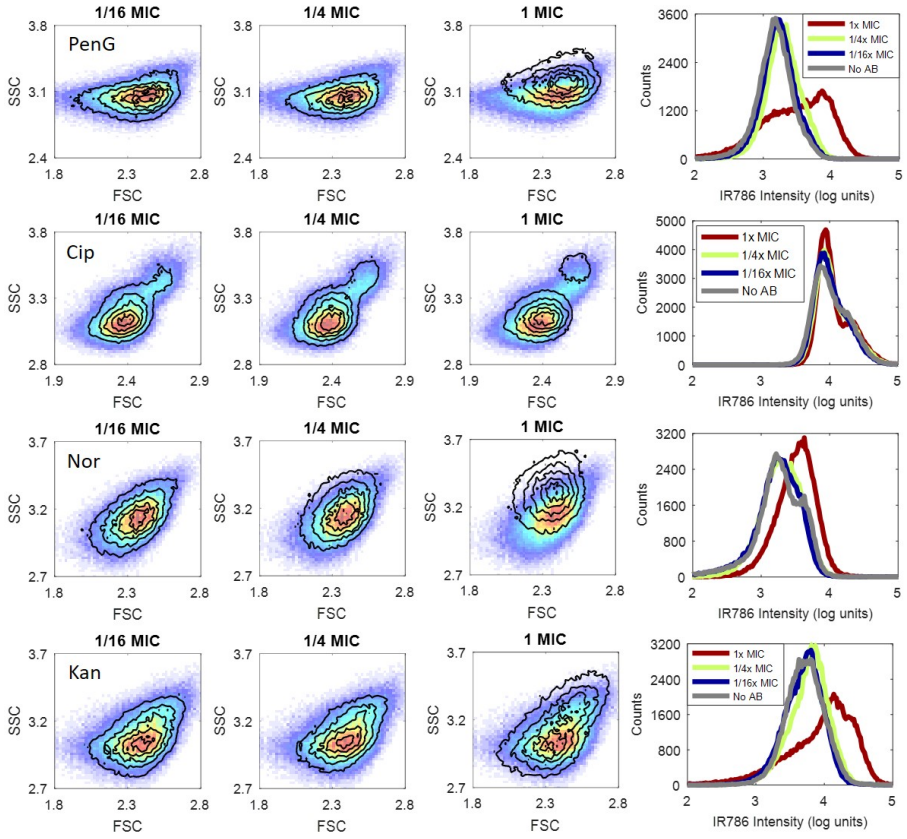

Figure B.1: Antibiotic-induced flow cytometry signal changes at different antibiotic concentrations for *E. coli* (ATCC 33456). The contours are the antibiotic-treated data from 1/16x MIC, 1/4x MIC, to 1x MIC as indicated at the top of each column. From the top to the bottom rows are data of penicillin g, ciprofloxacin, norfloxacin, and kanamycin. The right column contains the corresponding fluorescence data. Scattered light histograms correspond to the concentrations labeling the blue, green, and red curves in the fluorescence histograms.

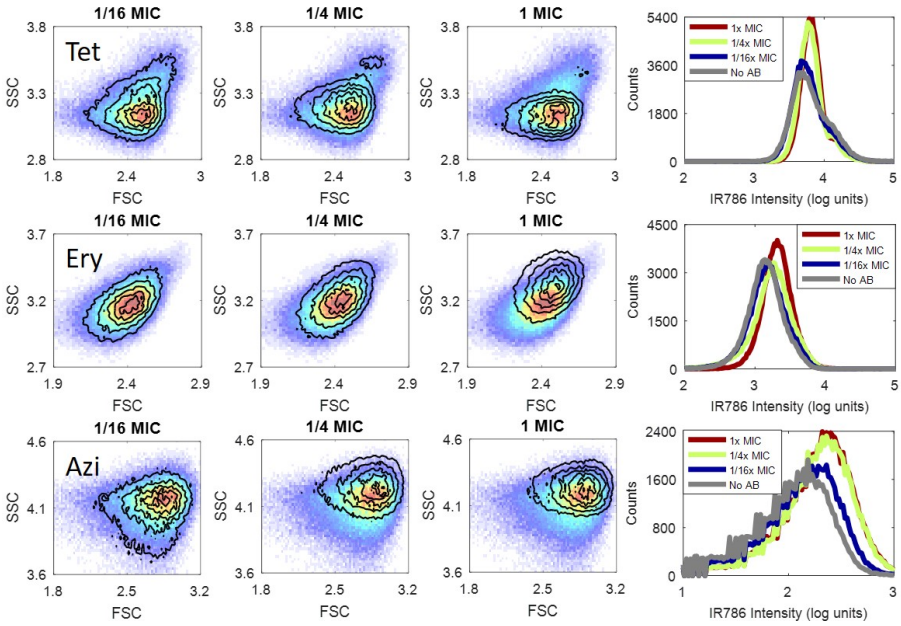

Figure B.2: Flow cytometry data for bacteriostatic antibiotics. Analogous to data in Appendix Figure B.1, from the top to the bottom rows are data of *E. coli* (ATCC 33456) exposed to tetracycline, erythromycin and azithromycin. Both bactericidal and bacteriostatic antibiotics give gradually increasing scattered light and fluorescence signal shifts from 1/16x MIC to 1x MIC.

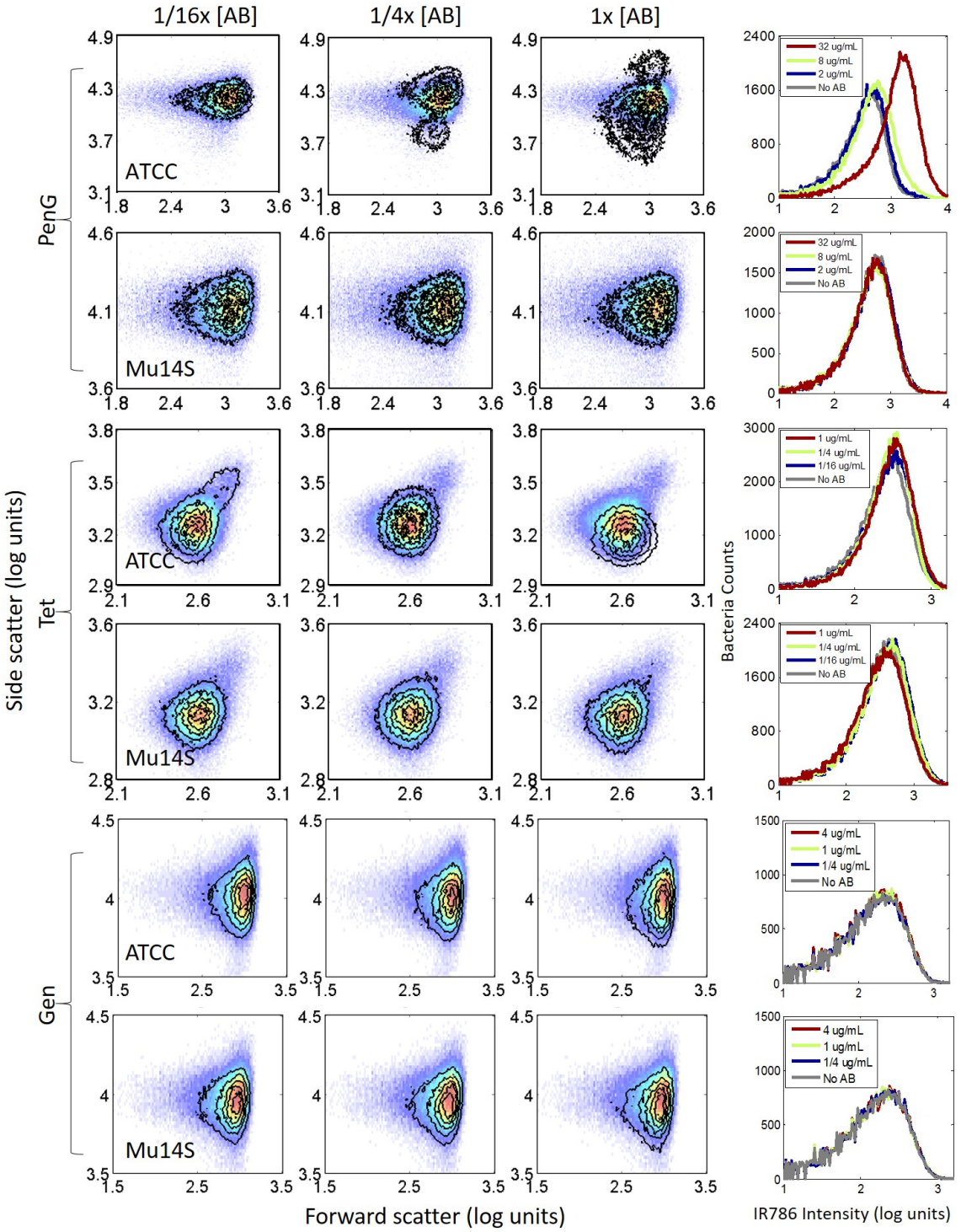

Figure B.3: Flow cytometry data for lab strain *E. coli* (ATCC 33456) and clinically isolated resistant strain *E. coli* (Mu14S). The contours represent the antibiotic-treated data and the colored dot plots are the no-antibiotic control data. From left to right, the antibiotic concentrations correspond to those indicated by the blue, green, and red curves, respectively in the fluorescence histograms of column 4. For PenG and Tet data, both strains were treated at the MIC of ATCC. For Gen data, both strains were treated at the MIC of Mu14S.

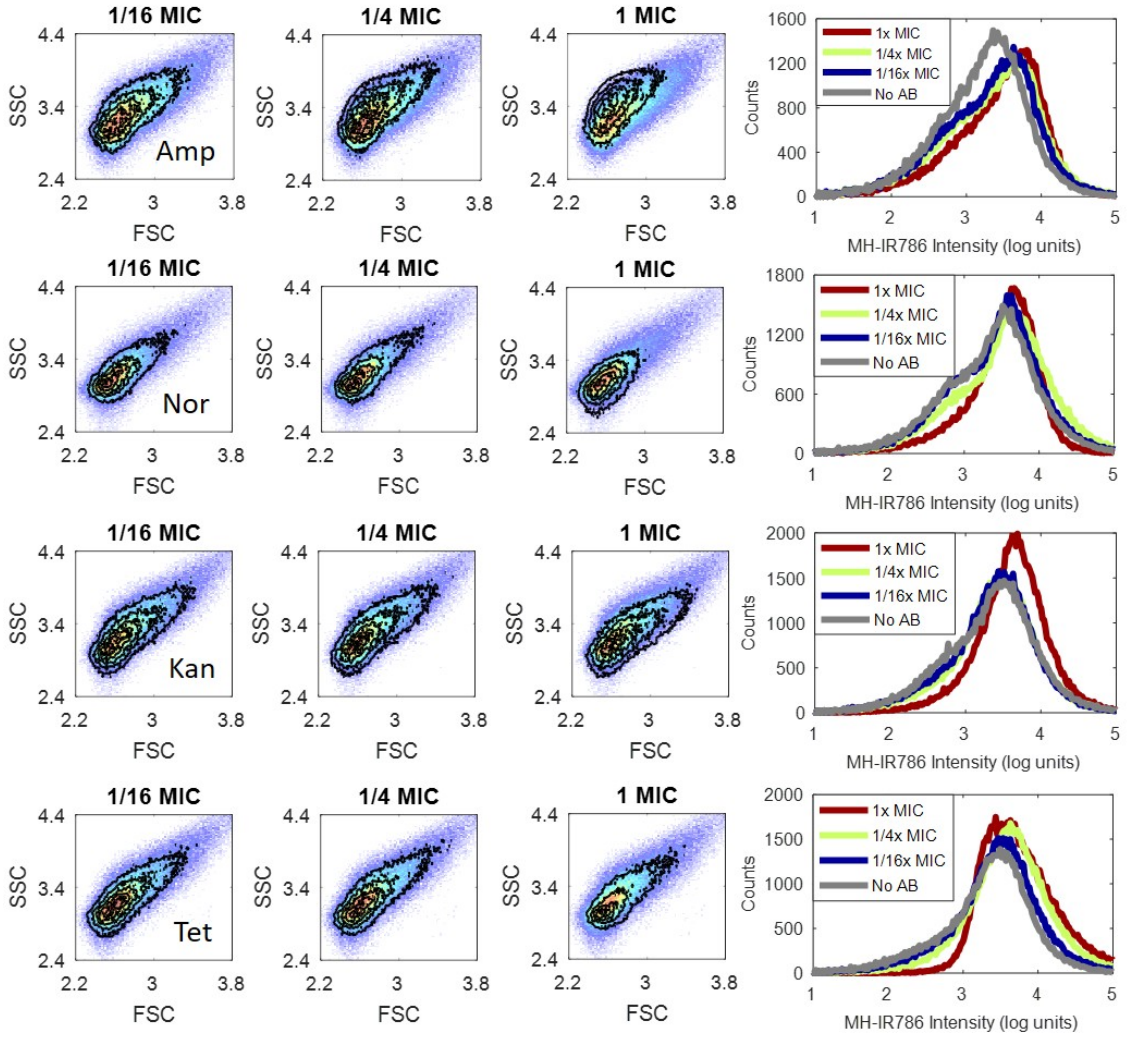

Figure B.4: Antibiotic-induced scatter signal changes in *P. aeruginosa*. Scatter plots of *P. aeruginosa* treated with antibiotic from 1/16x MIC to 1x MIC. Actual antibiotic concentrations again correspond to those indicated in the fluorescence histograms for blue, green, and red curves. Scatter changes were most prominent at 1x MIC. The top to the bottom rows show data with ampicillin, norfloxacin, kanamycin, and tetracycline.

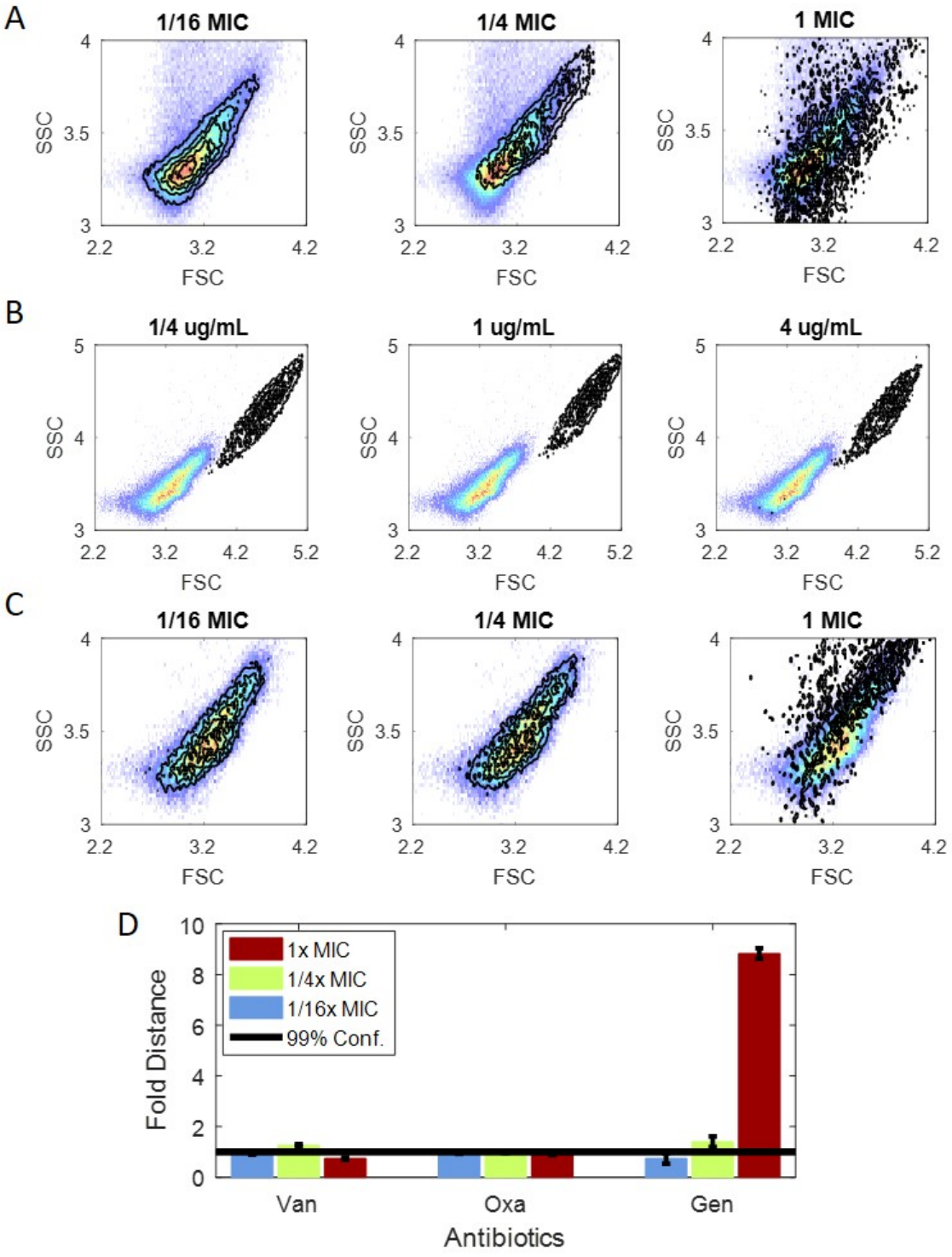

Figure B.5: Flow Cytometry data of MRSA and MSSA. Top 3 rows: penicillin g treated *S. aureus* strain 25923, strain 29213, and strain 43300 (MRSA). Bottom 2 rows: oxacillin-treated *S. aureus* strain 25923, strain 29213, and strain 43300 (MRSA)

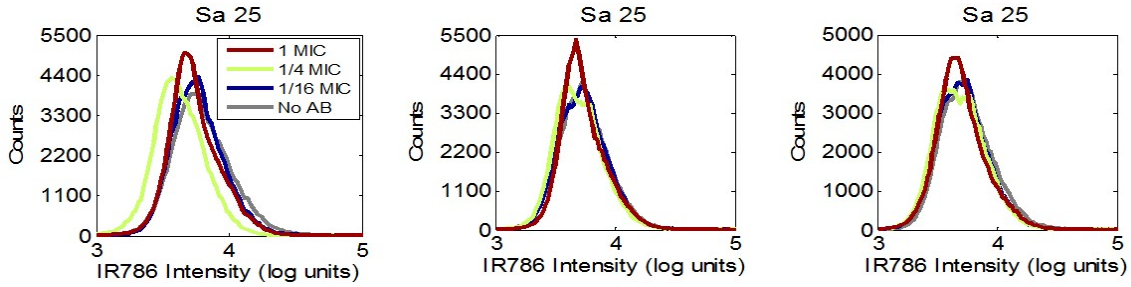

Figure B.6: Triplicates cytometric data for penicillin-treated *S. aureus* strain ATCC 25923. The data were prepared at the same time and taken on the same machine. The fluorescence signals, however, fluctuated. 1x MIC is 1/16 μg/mL.

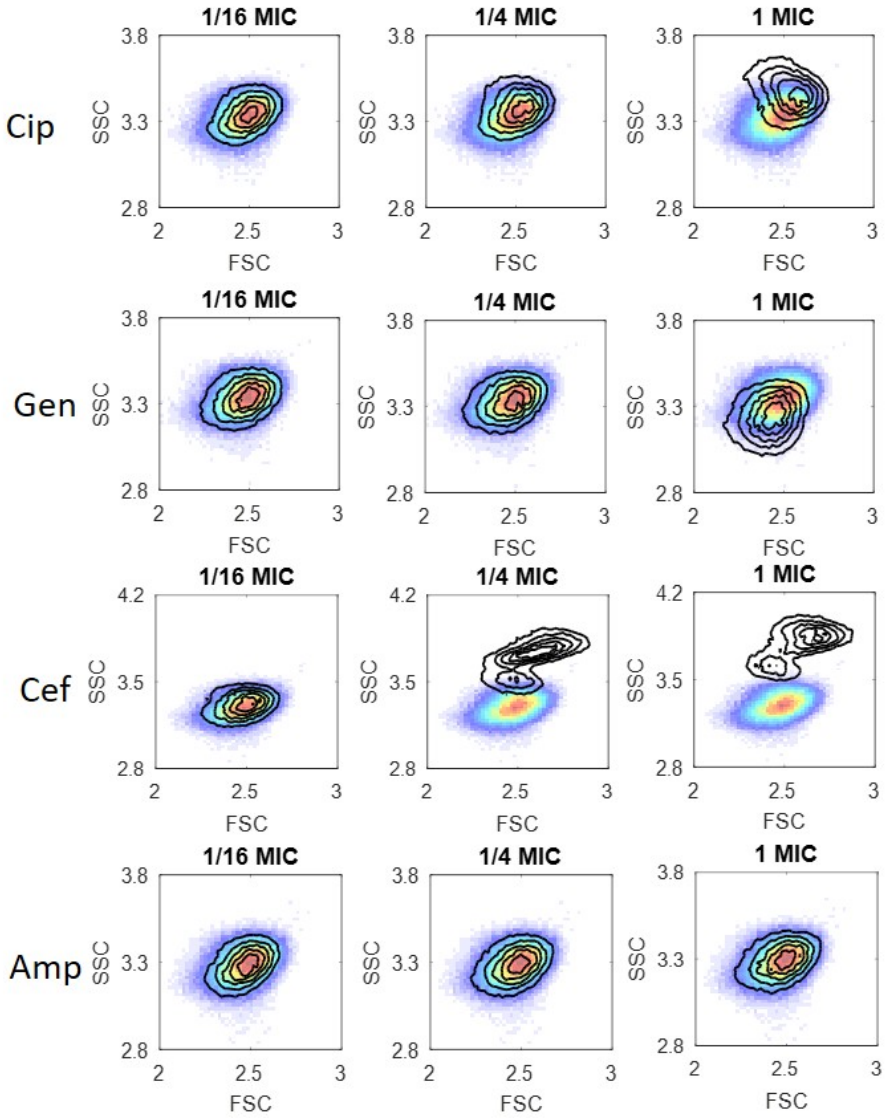

Figure B.7: Bactericidal Antibiotic-induced scatter changes for *K. pneumoniae*. Scatter plots of *K. pneumoniae* treated with antibiotic from 1/16x MIC to 1x MIC. The top to the bottom rows show data with ciprofloxacin, gentamicin, cefotaxime, and ampicillin.

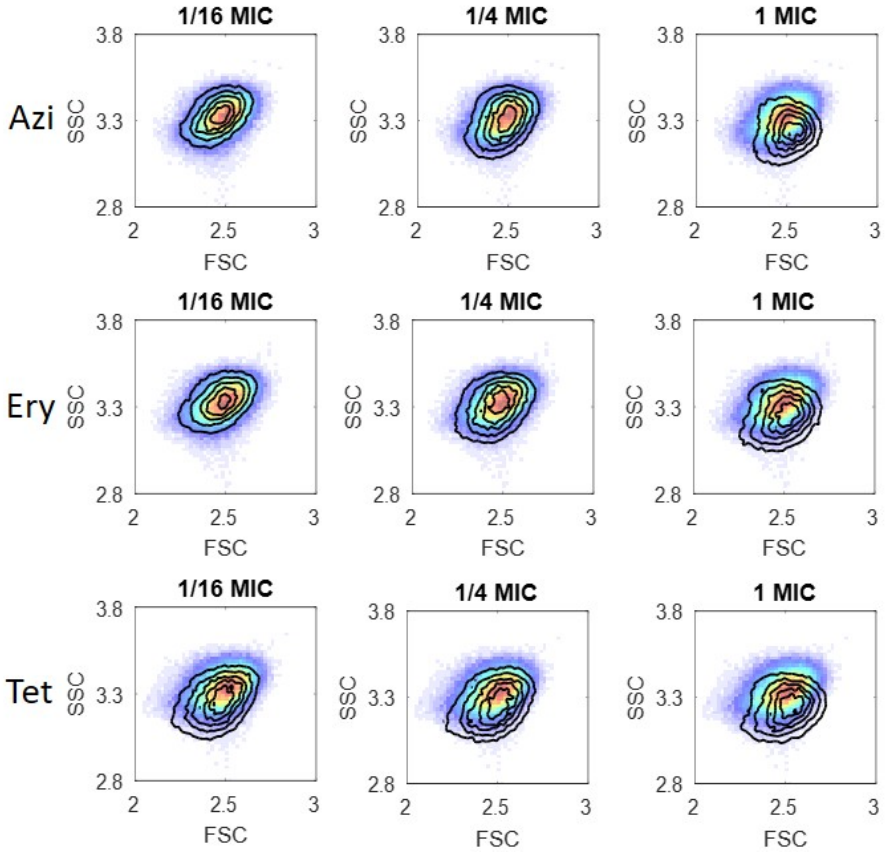

Figure B.8: Bacteriostatic Antibiotic-induced scatter changes for *K. pneumoniae*. Analogous to data in Appendix Figure B.7, from the top to the bottom rows are data of *K. pneumoniae* (ATCC 700603) exposed to azithromycin, erythromycin and tetracycline. Both bactericidal and bacteriostatic antibiotics give gradually increasing scattered light shifts from 1/16x MIC to 1x MIC.

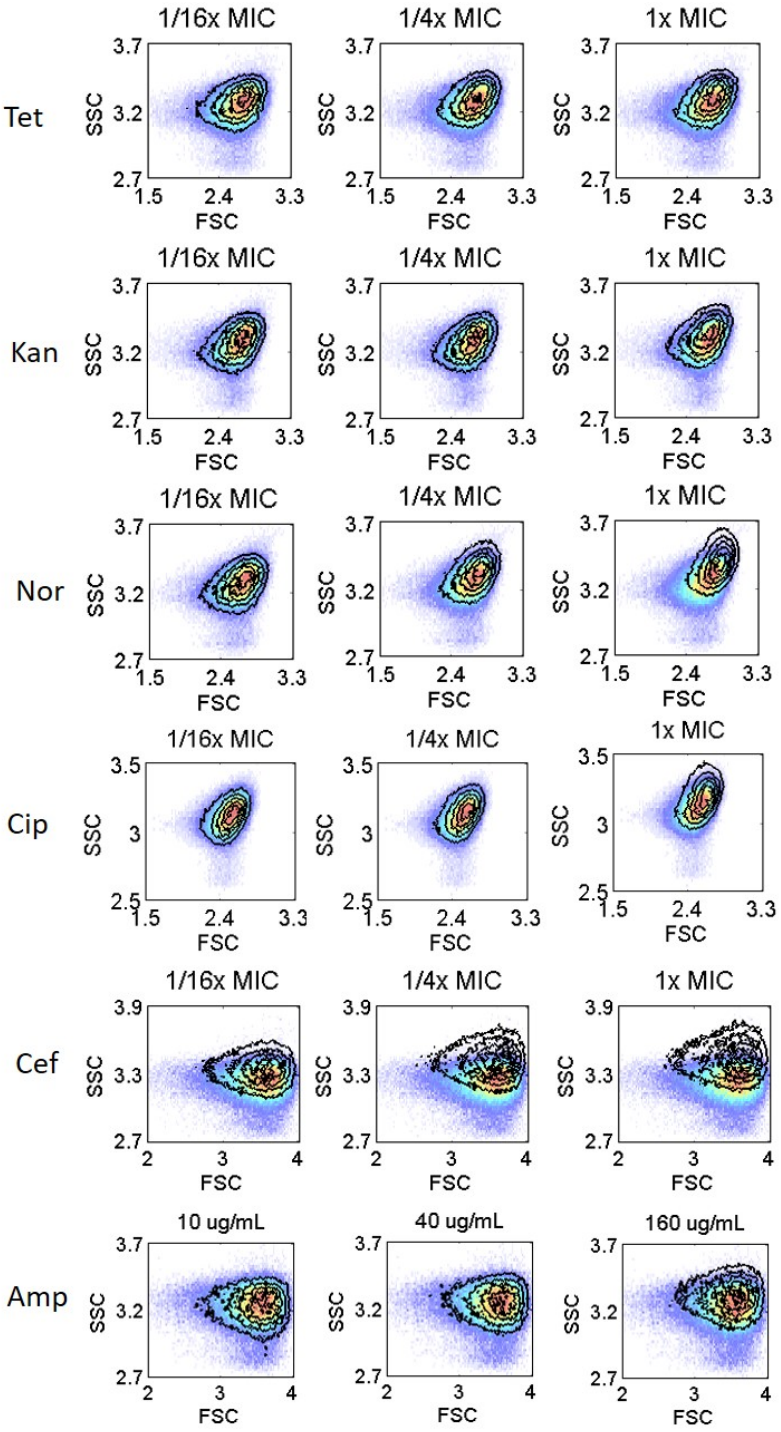

Figure B.9: Antibiotic-induced scatter changes for *A. nosocomialis* strain M2. Scatter plots of *A. nosocomialis* strain M2 treated with antibiotic from 1/16x MIC to 1x MIC or at clinical breakpoints. The top to the bottom rows show data with tetracycline, kanamycin, norfloxacin, ciprofloxacin, cefotaxime and ampicillin. Since M2 is resistant to ampicillin with MIC greater than 1024 μg/mL, the highest ampicillin concentration was set at 10x of the sensitive breakpoint, 160 μg/mL.

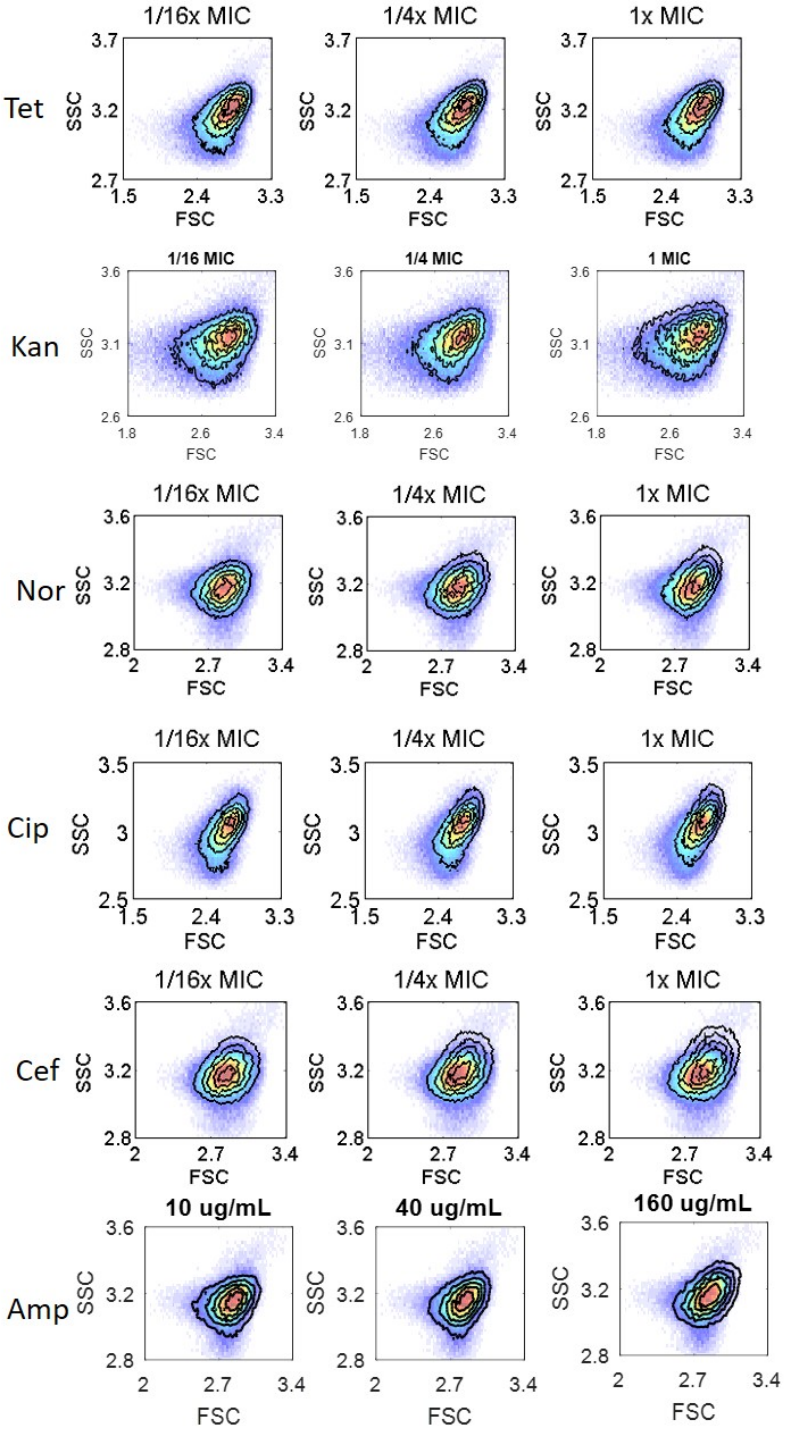

Figure B.10: Antibiotic-induced scatter changes for *A. nosocomialis* strain M2-4B. Scatter plots of *A. nosocomialis* strain M2-4B treated with antibiotic from 1/16x MIC to 1x MIC or at clinical breakpoints. The top to the bottom rows show data with tetracycline, kanamycin, norfloxacin, ciprofloxacin, cefotaxime and ampicillin. Since M2-4B is resistant to ampicillin with MIC greater than 1024 μg/mL, the highest ampicillin concentration was set at 10x of the sensitive breakpoint, 160 μg/mL.

APPENDIX C

SUPPORTING INFORMATION FOR CHAPTER 4

This chapter contains the pure culture cytometry data for strains tested in Chapter 4, including *E. coli* strain Mu890 and Mu14S, *K. pneumoniae* strain Mu55 and Mu670, and *A. nosocomialis* strain M2.

All the flow cytometry data presented here were label-free. For the scatter 2D plots, the pseudocolor plots are the paired-control, the no-antibiotic data, for each antibiotic-strain. The contours are the antibiotic-treated data with the antibiotic concentration indicated otherwise.

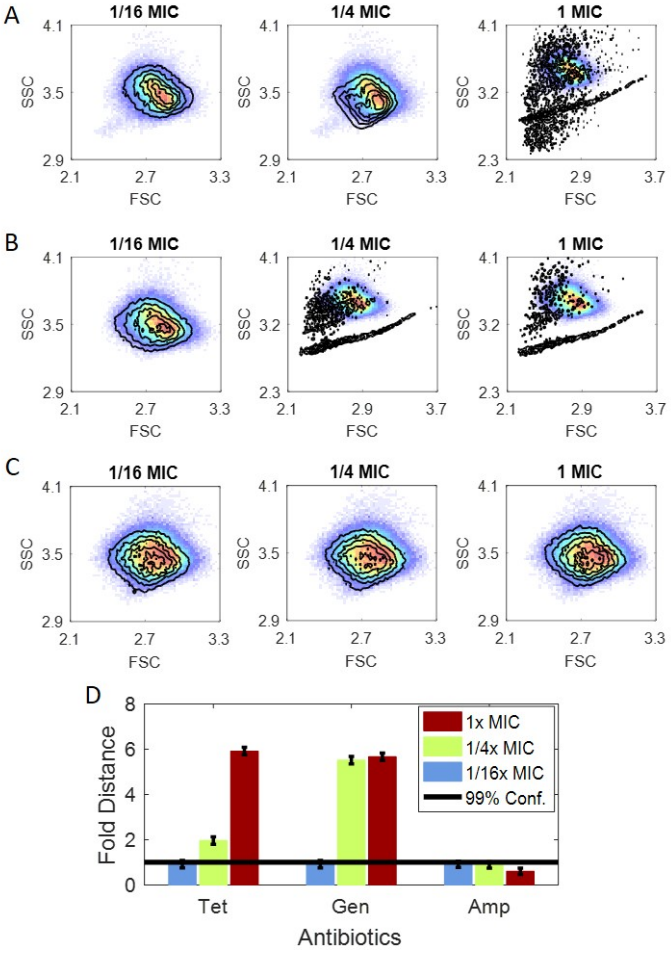

Figure C.1: Bactericidal Antibiotic-induced scatter changes for *E. coli* strain Mu890. For all data, pseudocolor plot: no-antibiotic, paired control. Black contour: antibiotic-treated data. Mu890 pure culture started from around 1000 CFU/mL and incubated for 5 hours, complementary to Figure 4.8 A. (A) Tetracycline (B) Gentamicin (C) Ampicillin. The 1x MIC for tetracycline is 2 μg/mL and 8 μg/mL for gentamicin. For ampicillin, 1x MIC was set as 32 μg/mL, the resistant breakpoint for Enterobacteriaceae. (D) PB-sQF 2D test results. The error bar only include the binning error since data was only done once.

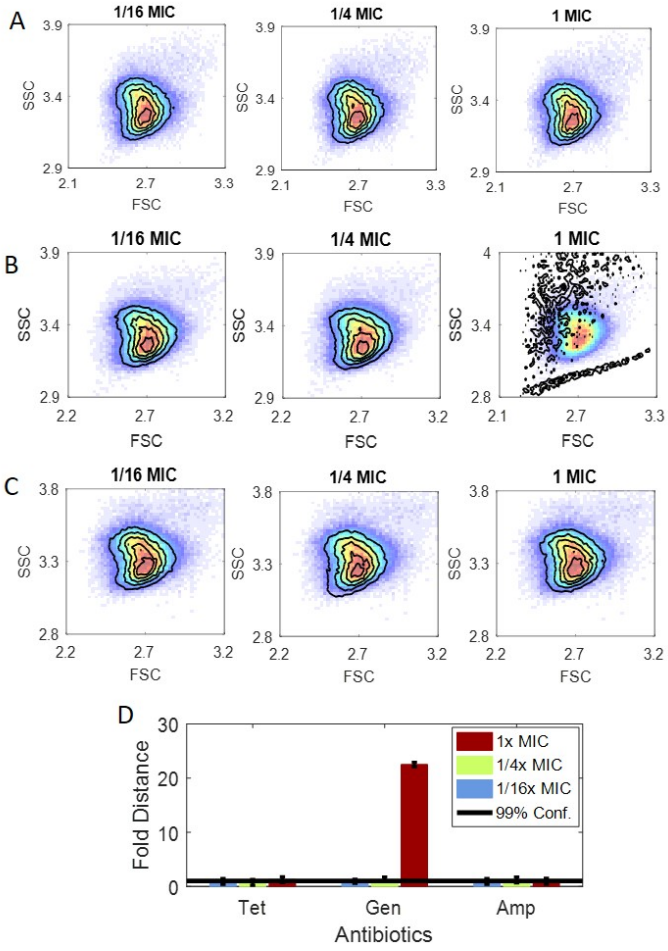

Figure C.2: Bactericidal Antibiotic-induced scatter changes for *E. coli* strain Mu14S. For all data, pseudocolor plot: no-antibiotic, paired control. Black contour: antibiotic-treated data. Mu14S pure culture started from around 1000 CFU/mL and incubated for 5 hours, complementary to Figure 4.8 B. (A) Tetracycline (B) Gentamicin (C) Ampicillin. The 1x MIC for gentamicin is 8 μg/mL. For ampicillin, 1x MIC was set as 32 μg/mL; while for tetracycline, 1x MIC was set as 16 μg/mL. Both are the resistant breakpoint for Enterobacteriaceae. (D) PB-sQF 2D test results. The error bar only include the binning error since data was only done once.

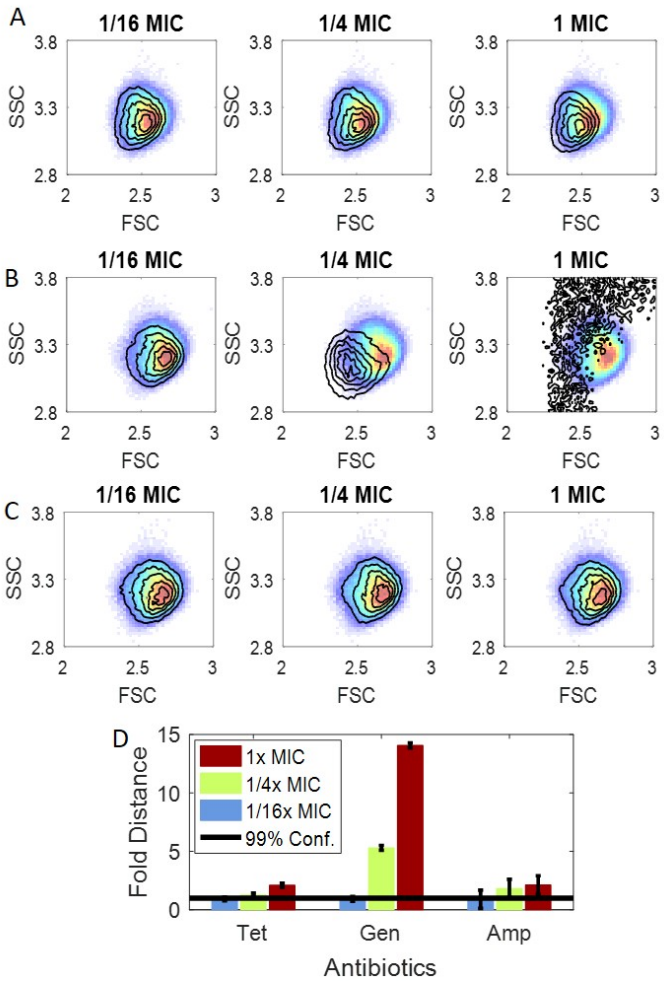

Figure C.3: Bactericidal Antibiotic-induced scatter changes for *K. pneumoniae* strain Mu55. For all data, pseudocolor plot: no-antibiotic, paired control. Black contour: antibiotic-treated data. Mu55 pure culture started from around 1000 CFU/mL and incubated for 5 hours as complementary to Fig. 4.9 A. (A) Tetracycline (B) Gentamicin (C) Ampicillin. The 1x MIC for gentamicin is 1 μg/mL. For ampicillin, 1x MIC was set as 32 μg/mL; while for tetracycline, 1x MIC was set as 16 μg/mL. Both are the resistant breakpoint for Enterobacteriaceae. (D) PB-sQF 2D test results. The error bar only include the binning error since data was only done once.

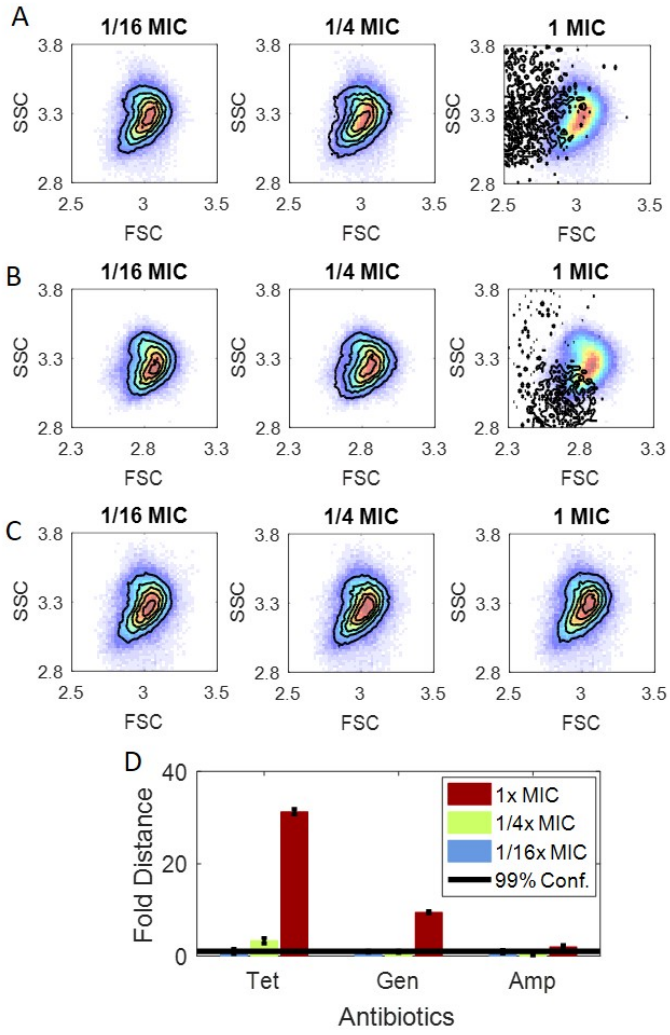

Figure C.4: Bactericidal Antibiotic-induced scatter changes for *K. pneumoniae* strain Mu670. For all data, pseudocolor plot: no-antibiotic, paired control. Black contour: antibiotic-treated data. Mu670 pure culture started from around 1000 CFU/mL and incubated for 5 hours as complementary to Fig. 4.9 B. (A) Tetracycline (B) Gentamicin (C) Ampicillin. The 1x MIC for tetracycline is 2 μg/mL and for gentamicin is 4 μg/mL. For ampicillin, 1x MIC was set as 32 μg/mL, the resistant breakpoint for Enterobacteriaceae. (D) PB-sQF 2D test results. The error bar only include the binning error since data was only done once.

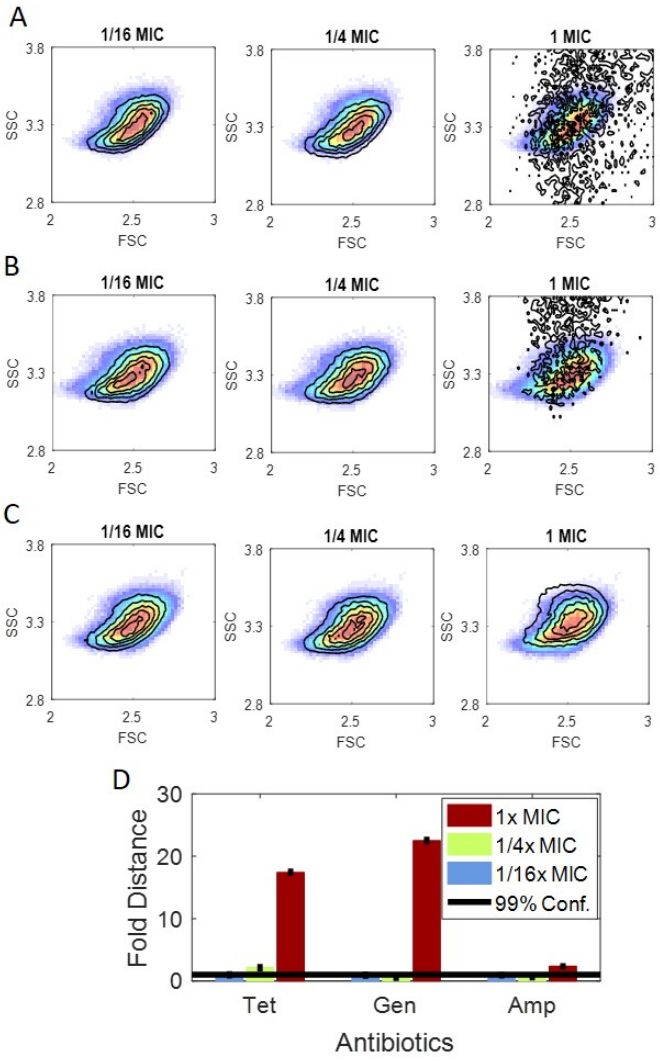

Figure C.5: Bactericidal Antibiotic-induced scatter changes for *A. nosocomialis* strain M2. For all data, pseudocolor plot: no-antibiotic, paired control. Black contour: antibiotic-treated data. M2 pure culture started from around 1000 CFU/mL and incubated for 5 hours as complementary to Fig. 4.10. (A) Tetracycline (B) Gentamicin (C) Ampicillin. The 1x MIC for tetracycline is 1/4 μg/mL and for gentamicin is 2 μg/mL. For ampicillin, 1x MIC was set as 128 μg/mL, the resistant breakpoint for *Acinetobacter*. (D) PB-sQF 2D test results. The error bar only include the binning error since data was only done once.

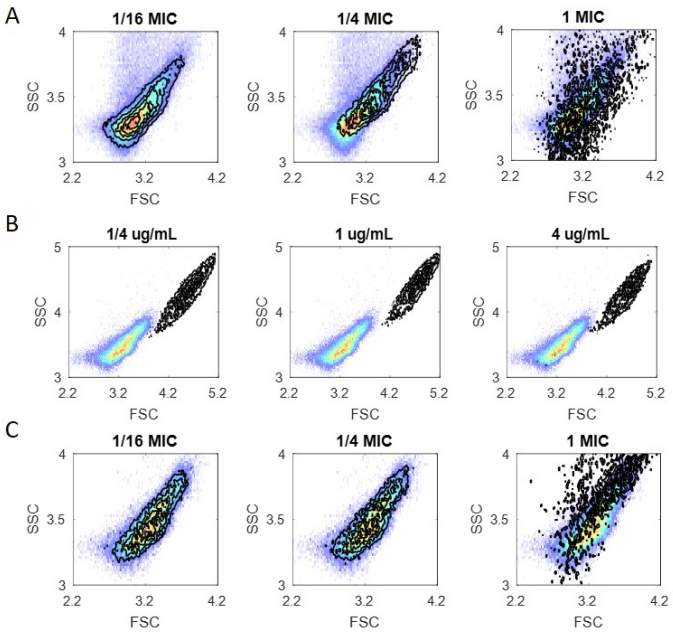

Figure C.6: Bactericidal Antibiotic-induced scatter changes for *S. aureus* strain NRS382. For all data, pseudocolor plot: no-antibiotic, paired control. Black contour: antibiotic-treated data. NRS382 pure culture started from around 1000 CFU/mL and incubated for 5 hours as complementary to Fig. 4.11. (A) Vancomycin (B) Oxacillin (C) Gentamicin. The 1x MIC for vancomycin is 2 μg/mL and for gentamicin is 1/4 μg/mL. For oxacillin, it was set as 4 μg/mL, the resistant breakpoint for *S. aureus*.

APPENDIX D

SUPPORTING INFORMATION FOR CHAPTER 5

This chapter contains the complete list of the library strains, assembled "unknown" sequences, and the raw short reads files. The threshold construction for bacterial typing, the typing results for all the assembled sequences, and the results for pooled-short reads typings from different sequencers were also included.

D.1 Downloaded Sequence Lists

Table D.1: Assembled library sequence from NCBI.

Library List
gi—158333233—ref—NC˙009925.1— Acaryochloris marina MBIC11017 chromosome
gi—162446888—ref—NC˙010163.1— Acholeplasma laidlawii PG-8A chromosome
gi—148259021—ref—NC˙009484.1— Acidiphilium cryptum JF-5 chromosome
gi—117927211—ref—NC˙008578.1— Acidothermus cellulolyticus 11B chromosome
gi—120608714—ref—NC˙008752.1— Acidovorax citrulli AAC00-1 chromosome
gi—121592436—ref—NC˙008782.1— Acidovorax sp. JS42 chromosome
gi—126640115—ref—NC˙009085.1— Acinetobacter baumannii ATCC 17978 chromosome
gi—50083297—ref—NC˙005966.1— Acinetobacter sp. ADP1 chromosome
gi—165975457—ref—NC˙010278.1— Actinobacillus pleuropneumoniae serovar 3 str. JL03 chromosome
gi—126207488—ref—NC˙009053.1— Actinobacillus pleuropneumoniae serovar 5b str. L20 chromosome
gi—152977688—ref—NC˙009655.1— Actinobacillus succinogenes 130Z chromosome
gi—117617447—ref—NC˙008570.1— Aeromonas hydrophila subsp. hydrophila ATCC 7966 chromosome
gi—145297124—ref—NC˙009348.1— Aeromonas salmonicida subsp. salmonicida A449
gi—159185562—ref—NC˙003063.2— Agrobacterium fabrum str. C58 chromosome linear
gi—110832861—ref—NC˙008260.1— Alcanivorax borkumensis SK2 chromosome
gi—114319166—ref—NC˙008340.1— Alkalilimnicola ehrlichii MLHE-1 chromosome
gi—150387853—ref—NC˙009633.1— Alkaliphilus metalliredigens QYMF chromosome
gi—158319059—ref—NC˙009922.1— Alkaliphilus oremlandii OhILAs chromosome
gi—75906225—ref—NC˙007413.1— Anabaena variabilis ATCC 29413 chromosome
gi—86156430—ref—NC˙007760.1— Anaeromyxobacter dehalogenans 2CP-C chromosome
gi—153002879—ref—NC˙009675.1— Anaeromyxobacter sp. Fw109-5 chromosome
gi—56416370—ref—NC˙004842.2— Anaplasma marginale str. St. Maries chromosome
gi—88606690—ref—NC˙007797.1— Anaplasma phagocytophilum HZ
gi—15282445—ref—NC˙000918.1— Aquifex aeolicus VF5
gi—157736271—ref—NC˙009850.1— Arcobacter butzleri RM4018 chromosome
gi—56475432—ref—NC˙006513.1— Aromatoleum aromaticum EbN1 chromosome
gi—119960487—ref—NC˙008711.1— Arthrobacter aurescens TC1
gi—116668568—ref—NC˙008541.1— Arthrobacter sp. FB24
gi—85057280—ref—NC˙007716.1— Aster yellows witches'-broom phytoplasma AYWB
gi—119896292—ref—NC˙008702.1— Azoarcus sp. BH72 chromosome
gi—158421624—ref—NC˙009937.1— Azorhizobium caulinodans ORS 571 chromosome
gi—154684518—ref—NC˙009725.1— Bacillus amyloliquefaciens FZB42
gi—50196905—ref—NC˙007530.2— Bacillus anthracis str. 'Ames Ancestor' chromosome
gi—30260195—ref—NC˙003997.3— Bacillus anthracis str. Ames chromosome
gi—49183039—ref—NC˙005945.1— Bacillus anthracis str. Sterne chromosome
gi—42779081—ref—NC˙003909.8— Bacillus cereus ATCC 10987
gi—30018278—ref—NC˙004722.1— Bacillus cereus ATCC 14579
gi—52140164—ref—NC˙006274.1— Bacillus cereus E33L chromosome
gi—56961782—ref—NC˙006582.1— Bacillus clausii KSM-K16
gi—152973854—ref—NC˙009674.1— Bacillus cytotoxicus NVH 391-98 chromosome
gi—57596592—ref—NC˙002570.2— Bacillus halodurans C-125 chromosome
gi—163119169—ref—NC˙006270.3— Bacillus licheniformis ATCC 14580 chromosome
gi—157690798—ref—NC˙009848.1— Bacillus pumilus SAFR-032 chromosome
gi—255767013—ref—NC˙000964.3— Bacillus subtilis subsp. subtilis str. 168 chromosome
gi—49476684—ref—NC˙005957.1— Bacillus thuringiensis serovar konkukian str. 97-27 chromosome
gi—118475778—ref—NC˙008600.1— Bacillus thuringiensis str. Al Hakam chromosome
gi—163938013—ref—NC˙010184.1— Bacillus weihenstephanensis KBAB4 chromosome
gi—60679597—ref—NC˙003228.3— Bacteroides fragilis NCTC 9343 chromosome
gi—53711291—ref—NC˙006347.1— Bacteroides fragilis YCH46 chromosome
gi—29345410—ref—NC˙004663.1— Bacteroides thetaiotaomicron VPI-5482 chromosome
gi—150002608—ref—NC˙009614.1— Bacteroides vulgatus ATCC 8482 chromosome
gi—121601635—ref—NC˙008783.1— Bartonella bacilliformis KC583
gi—49474831—ref—NC˙005956.1— Bartonella henselae str. Houston-1 chromosome
gi—49473688—ref—NC˙005955.1— Bartonella quintana str. Toulouse
gi—163867306—ref—NC˙010161.1— Bartonella tribocorum CIP 105476 chromosome

Table D.1 Assembled library sequence from NCBI. Continued

Library List
gi—94676460—ref—NC˙007984.1— Baumannia cicadellinicola str. Hc (Homalodisca coagulata)
gi—42521650—ref—NC˙005363.1— Bdellovibrio bacteriovorus HD100
gi—119025018—ref—NC˙008618.1— Bifidobacterium adolescentis ATCC 15703 chromosome
gi—58036264—ref—NC˙004307.2— Bifidobacterium longum NCC2705 chromosome
gi—33598993—ref—NC˙002927.3— Bordetella bronchiseptica RB50 chromosome
gi—33594723—ref—NC˙002928.3— Bordetella parapertussis 12822 chromosome
gi—33591275—ref—NC˙002929.2— Bordetella pertussis Tohama I chromosome
gi—163854304—ref—NC˙010170.1— Bordetella petrii DSM 12804 chromosome
gi—111114823—ref—NC˙008277.1— Borrelia afzelii PKo
gi—15594346—ref—NC˙001318.1— Borrelia burgdorferi B31 chromosome
gi—51598263—ref—NC˙006156.1— Borrelia garinii PBi chromosome linear
gi—27375111—ref—NC˙004463.1— Bradyrhizobium japonicum USDA 110 chromosome
gi—148251626—ref—NC˙009485.1— Bradyrhizobium sp. BTAi1 chromosome
gi—146337175—ref—NC˙009445.1— Bradyrhizobium sp. ORS 278 chromosome
gi—62288991—ref—NC˙006932.1— Brucella abortus bv. 1 str. 9-941 chromosome I
gi—62316961—ref—NC˙006933.1— Brucella abortus bv. 1 str. 9-941 chromosome II
gi—161617991—ref—NC˙010103.1— Brucella canis ATCC 23365 chromosome I
gi—161620094—ref—NC˙010104.1— Brucella canis ATCC 23365 chromosome II
gi—82698932—ref—NC˙007618.1— Brucella melitensis biovar Abortus 2308 chromosome I
gi—83268957—ref—NC˙007624.1— Brucella melitensis biovar Abortus 2308 chromosome II
gi—17986284—ref—NC˙003317.1— Brucella melitensis bv. 1 str. 16M chromosome I
gi—17988344—ref—NC˙003318.1— Brucella melitensis bv. 1 str. 16M chromosome II
gi—148557829—ref—NC˙009504.1— Brucella ovis ATCC 25840 chromosome II
gi—148558820—ref—NC˙009505.1— Brucella ovis ATCC 25840 chromosome I
gi—56968325—ref—NC˙004310.3— Brucella suis 1330 chromosome I
gi—56968493—ref—NC˙004311.2— Brucella suis 1330 chromosome II
gi—163844199—ref—NC˙010167.1— Brucella suis ATCC 23445 chromosome II
gi—163842277—ref—NC˙010169.1— Brucella suis ATCC 23445 chromosome I
gi—15616630—ref—NC˙002528.1— Buchnera aphidicola str. APS (Acyrthosiphon pisum) chromosome
gi—27904513—ref—NC˙004545.1— Buchnera aphidicola str. Bp (Baizongia pistaciae) chromosome
gi—116514950—ref—NC˙008513.1— Buchnera aphidicola str. Cc (Cinara cedri)
gi—21672294—ref—NC˙004061.1— Buchnera aphidicola str. Sg (Schizaphis graminum) chromosome
gi—115350056—ref—NC˙008390.1— Burkholderia ambifaria AMMD chromosome 1
gi—115357970—ref—NC˙008391.1— Burkholderia ambifaria AMMD chromosome 2
gi—115360317—ref—NC˙008392.1— Burkholderia ambifaria AMMD chromosome 3
gi—107021562—ref—NC˙008060.1— Burkholderia cenocepacia AU 1054 chromosome 1
gi—107025343—ref—NC˙008061.1— Burkholderia cenocepacia AU 1054 chromosome 2
gi—107028231—ref—NC˙008062.1— Burkholderia cenocepacia AU 1054 chromosome 3
gi—116688024—ref—NC˙008542.1— Burkholderia cenocepacia HI2424 chromosome 1
gi—116691273—ref—NC˙008543.1— Burkholderia cenocepacia HI2424 chromosome 2
gi—116686245—ref—NC˙008544.1— Burkholderia cenocepacia HI2424 chromosome 3
gi—53723370—ref—NC˙006348.1— Burkholderia mallei ATCC 23344 chromosome 1
gi—77358719—ref—NC˙006349.2— Burkholderia mallei ATCC 23344 chromosome 2
gi—124381141—ref—NC˙008835.1— Burkholderia mallei NCTC 10229 chromosome II
gi—124383319—ref—NC˙008836.1— Burkholderia mallei NCTC 10229 chromosome I
gi—126445587—ref—NC˙009079.1— Burkholderia mallei NCTC 10247 chromosome II
gi—126447966—ref—NC˙009080.1— Burkholderia mallei NCTC 10247 chromosome I
gi—121596444—ref—NC˙008784.1— Burkholderia mallei SAVP1 chromosome II
gi—121598179—ref—NC˙008785.1— Burkholderia mallei SAVP1 chromosome I
gi—161523180—ref—NC˙010084.1— Burkholderia multivorans ATCC 17616 chromosome 1
gi—161519706—ref—NC˙010086.1— Burkholderia multivorans ATCC 17616 chromosome 2
gi—161522356—ref—NC˙010087.1— Burkholderia multivorans ATCC 17616 chromosome 3
gi—126451443—ref—NC˙009076.1— Burkholderia pseudomallei 1106a chromosome I
gi—126455463—ref—NC˙009078.1— Burkholderia pseudomallei 1106a chromosome II
gi—76808520—ref—NC˙007434.1— Burkholderia pseudomallei 1710b chromosome I

Table D.1 Assembled library sequence from NCBI. Continued

Library List
gi—76817237—ref—NC˙007435.1— Burkholderia pseudomallei 1710b chromosome II
gi—126438353—ref—NC˙009074.1— Burkholderia pseudomallei 668 chromosome I
gi—126442307—ref—NC˙009075.1— Burkholderia pseudomallei 668 chromosome chromosome II
gi—53717639—ref—NC˙006350.1— Burkholderia pseudomallei K96243 chromosome 1
gi—53721039—ref—NC˙006351.1— Burkholderia pseudomallei K96243 chromosome 2
gi—78059643—ref—NC˙007509.1— Burkholderia sp. 383 chromosome 3
gi—78064658—ref—NC˙007510.1— Burkholderia sp. 383 chromosome 1
gi—78060853—ref—NC˙007511.1— Burkholderia sp. 383 chromosome 2
gi—83716035—ref—NC˙007650.1— Burkholderia thailandensis E264 chromosome II
gi—83718394—ref—NC˙007651.1— Burkholderia thailandensis E264 chromosome I
gi—134290884—ref—NC˙009254.1— Burkholderia vietnamiensis G4 chromosome 3
gi—134292031—ref—NC˙009255.1— Burkholderia vietnamiensis G4 chromosome 2
gi—134294128—ref—NC˙009256.1— Burkholderia vietnamiensis G4 chromosome 1
gi—91781384—ref—NC˙007951.1— Burkholderia xenovorans LB400 chromosome 1
gi—91777110—ref—NC˙007952.1— Burkholderia xenovorans LB400 chromosome 2
gi—91780071—ref—NC˙007953.1— Burkholderia xenovorans LB400 chromosome 3
gi—146295085—ref—NC˙009437.1— Caldicellulosiruptor saccharolyticus DSM 8903 chromosome
gi—157163852—ref—NC˙009802.1— Campylobacter concisus 13826
gi—154173617—ref—NC˙009715.1— Campylobacter curvus 525.92 chromosome
gi—118474057—ref—NC˙008599.1— Campylobacter fetus subsp. fetus 82-40 chromosome
gi—154147866—ref—NC˙009714.1— Campylobacter hominis ATCC BAA-381
gi—57236892—ref—NC˙003912.7— Campylobacter jejuni RM1221
gi—153950938—ref—NC˙009707.1— Campylobacter jejuni subsp. doylei 269.97 chromosome
gi—121612099—ref—NC˙008787.1— Campylobacter jejuni subsp. jejuni 81-176 chromosome
gi—157414322—ref—NC˙009839.1— Campylobacter jejuni subsp. jejuni 81116
gi—15791399—ref—NC˙002163.1— Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819 chromosome
gi—33519483—ref—NC˙005061.1— Candidatus Blochmannia floridanus chromosome
gi—71891793—ref—NC˙007292.1— Candidatus Blochmannia pennsylvanicus str. BPEN chromosome
gi—94967031—ref—NC˙008009.1— Candidatus Koribacter versatilis Ellin345 chromosome
gi—71082709—ref—NC˙007205.1— Candidatus Pelagibacter ubique HTCC1062 chromosome
gi—46445634—ref—NC˙005861.1— Candidatus Protochlamydia amoebophila UWE25 chromosome
gi—118602060—ref—NC˙008610.1— Candidatus Ruthia magnifica str. Cm (Calyptogena magnifica)
gi—116619145—ref—NC˙008536.1— Candidatus Solibacter usitatus Ellin6076 chromosome
gi—161833634—ref—NC˙010118.1— Candidatus Sulcia muelleri GWSS
gi—148244169—ref—NC˙009465.1— Candidatus Vesicomyosocius okutanii HA
gi—78042616—ref—NC˙007503.1— Carboxydothermus hydrogenoformans Z-2901 chromosome
gi—16124256—ref—NC˙002696.2— Caulobacter crescentus CB15 chromosome
gi—110632362—ref—NC˙008254.1— Chelativorans sp. BNC1 chromosome
gi—29337300—ref—NC˙002620.2— Chlamydia muridarum Nigg
gi—166153973—ref—NC˙010287.1— Chlamydia trachomatis 434/Bu chromosome
gi—76788711—ref—NC˙007429.1— Chlamydia trachomatis A/HAR-13
gi—15604717—ref—NC˙000117.1— Chlamydia trachomatis D/UW-3/CX
gi—352951305—ref—NC˙010280.2— Chlamydia trachomatis L2b/UCH-1/proctitis chromosome
gi—62184647—ref—NC˙004552.2— Chlamydophila abortus S26/3
gi—29839769—ref—NC˙003361.3— Chlamydophila caviae GPIC chromosome
gi—89897807—ref—NC˙007899.1— Chlamydophila felis Fe/C-56
gi—58021288—ref—NC˙002179.2— Chlamydophila pneumoniae AR39
gi—15617929—ref—NC˙000922.1— Chlamydophila pneumoniae CWL029 chromosome
gi—15835535—ref—NC˙002491.1— Chlamydophila pneumoniae J138 chromosome
gi—33241335—ref—NC˙005043.1— Chlamydophila pneumoniae TW-183
gi—78187984—ref—NC˙007514.1— Chlorobium chlorochromatii CaD3 chromosome
gi—78185892—ref—NC˙007512.1— Chlorobium luteolum DSM 273 chromosome
gi—119355857—ref—NC˙008639.1— Chlorobium phaeobacteroides DSM 266 chromosome
gi—145218822—ref—NC˙009337.1— Chlorobium phaeovibrioides DSM 265 chromosome
gi—21672841—ref—NC˙002932.3— Chlorobium tepidum TLS chromosome

Table D.1 Assembled library sequence from NCBI. Continued

Library List
gi—163845603—ref—NC˙010175.1— Chloroflexus aurantiacus J-10-fl chromosome
gi—34495455—ref—NC˙005085.1— Chromobacterium violaceum ATCC 12472 chromosome
gi—92112136—ref—NC˙007963.1— Chromohalobacter salexigens DSM 3043 chromosome
gi—157144296—ref—NC˙009792.1— Citrobacter koseri ATCC BAA-895 chromosome
gi—148271178—ref—NC˙009480.1— Clavibacter michiganensis subsp. michiganensis NCPPB 382 chromosome
gi—15893298—ref—NC˙003030.1— Clostridium acetobutylicum ATCC 824 chromosome
gi—150014892—ref—NC˙009617.1— Clostridium beijerinckii NCIMB 8052 chromosome
gi—153930785—ref—NC˙009697.1— Clostridium botulinum A str. ATCC 19397 chromosome
gi—148378011—ref—NC˙009495.1— Clostridium botulinum A str. ATCC 3502 chromosome
gi—153934468—ref—NC˙009698.1— Clostridium botulinum A str. Hall chromosome
gi—153937894—ref—NC˙009699.1— Clostridium botulinum F str. Langeland chromosome
gi—126697566—ref—NC˙009089.1— Clostridium difficile 630
gi—153952670—ref—NC˙009706.1— Clostridium kluyveri DSM 555 chromosome
gi—118442852—ref—NC˙008593.1— Clostridium novyi NT chromosome
gi—110798562—ref—NC˙008261.1— Clostridium perfringens ATCC 13124 chromosome
gi—110801439—ref—NC˙008262.1— Clostridium perfringens SM101 chromosome
gi—18308982—ref—NC˙003366.1— Clostridium perfringens str. 13 chromosome
gi—160878162—ref—NC˙010001.1— Clostridium phytofermentans ISDg chromosome
gi—28209834—ref—NC˙004557.1— Clostridium tetani E88 chromosome
gi—125972525—ref—NC˙009012.1— Clostridium thermocellum ATCC 27405 chromosome
gi—71277742—ref—NC˙003910.7— Colwellia psychrerythraea 34H chromosome
gi—38232642—ref—NC˙002935.2— Corynebacterium diphtheriae NCTC 13129 chromosome
gi—25026556—ref—NC˙004369.1— Corynebacterium efficiens YS-314 chromosome
gi—58036263—ref—NC˙003450.3— Corynebacterium glutamicum ATCC 13032
gi—145294042—ref—NC˙009342.1— Corynebacterium glutamicum R chromosome
gi—68535062—ref—NC˙007164.1— Corynebacterium jeikeium K411 chromosome
gi—154705721—ref—NC˙009727.1— Coxiella burnetii Dugway 5J108-111 chromosome
gi—161829703—ref—NC˙010117.1— Coxiella burnetii RSA 331 chromosome
gi—77358712—ref—NC˙002971.3— Coxiella burnetii RSA 493 chromosome
gi—156932229—ref—NC˙009778.1— Cronobacter sakazakii ATCC BAA-894 chromosome
gi—94308945—ref—NC˙007973.1— Cupriavidus metallidurans CH34 chromosome
gi—110636427—ref—NC˙008255.1— Cytophaga hutchinsonii ATCC 33406 chromosome
gi—71905642—ref—NC˙007298.1— Dechloromonas aromatica RCB
gi—57233530—ref—NC˙002936.3— Dehalococcoides ethenogenes 195
gi—147668652—ref—NC˙009455.1— Dehalococcoides sp. BAV1 chromosome
gi—73747956—ref—NC˙007356.1— Dehalococcoides sp. CBDB1 chromosome
gi—94984109—ref—NC˙008025.1— Deinococcus geothermalis DSM 11300
gi—15805042—ref—NC˙001263.1— Deinococcus radiodurans R1 chromosome 1
gi—15807672—ref—NC˙001264.1— Deinococcus radiodurans R1 chromosome 2
gi—160895450—ref—NC˙010002.1— Delftia acidovorans SPH-1 chromosome
gi—89892746—ref—NC˙007907.1— Desulfitobacterium hafniense Y51 chromosome
gi—158520017—ref—NC˙009943.1— Desulfococcus oleovorans Hxd3 chromosome
gi—51243852—ref—NC˙006138.1— Desulfotalea psychrophila LSv54
gi—134297881—ref—NC˙009253.1— Desulfotomaculum reducens MI-1 chromosome
gi—78355047—ref—NC˙007519.1— Desulfovibrio alaskensis G20 chromosome
gi—120601051—ref—NC˙008751.1— Desulfovibrio vulgaris DP4 chromosome
gi—46562128—ref—NC˙002937.3— Desulfovibrio vulgaris str. Hildenborough chromosome
gi—146328629—ref—NC˙009446.1— Dichelobacter nodosus VCS1703A chromosome
gi—159042556—ref—NC˙009952.1— Dinoroseobacter shibae DFL 12 chromosome
gi—73666633—ref—NC˙007354.1— Ehrlichia canis str. Jake chromosome
gi—88657561—ref—NC˙007799.1— Ehrlichia chaffeensis str. Arkansas
gi—58616727—ref—NC˙006831.1— Ehrlichia ruminantium str. Gardel
gi—57238731—ref—NC˙005295.2— Ehrlichia ruminantium str. Welgevonden chromosome
gi—146309667—ref—NC˙009436.1— Enterobacter sp. 638
gi—29374661—ref—NC˙004668.1— Enterococcus faecalis V583 chromosome

Table D.1 Assembled library sequence from NCBI. Continued

Library List
gi—85372828—ref—NC˙007722.1— Erythrobacter litoralis HTCC2594 chromosome
gi—110640213—ref—NC˙008253.1— Escherichia coli 536
gi—117622295—ref—NC˙008563.1— Escherichia coli APEC O1 chromosome
gi—26245917—ref—NC˙004431.1— Escherichia coli CFT073 chromosome
gi—157154711—ref—NC˙009801.1— Escherichia coli E24377A chromosome
gi—157159467—ref—NC˙009800.1— Escherichia coli HS
gi—16445223—ref—NC˙002655.2— Escherichia coli O157:H7 str. EDL933 chromosome
gi—15829254—ref—NC˙002695.1— Escherichia coli O157:H7 str. Sakai chromosome
gi—91209055—ref—NC˙007946.1— Escherichia coli UTI89 chromosome
gi—49175990—ref—NC˙000913.2— Escherichia coli str. K-12 substr. MG1655
gi—154248705—ref—NC˙009718.1— Fervidobacterium nodosum Rt17-B1 chromosome
gi—146297766—ref—NC˙009441.1— Flavobacterium johnsoniae UW101 chromosome
i—511542232—ref—NC˙009613.3— Flavobacterium psychrophilum JIP02/86 complete genome
gi—118496615—ref—NC˙008601.1— Francisella novicida U112 chromosome
gi—156501369—ref—NC˙009749.1— Francisella tularensis subsp. holarctica FTNF002-00 chromosome
gi—89255449—ref—NC˙007880.1— Francisella tularensis subsp. holarctica LVS chromosome
gi—115313981—ref—NC˙008369.1— Francisella tularensis subsp. holarctica OSU18 chromosome
gi—110669657—ref—NC˙008245.1— Francisella tularensis subsp. tularensis FSC198 chromosome
gi—255961454—ref—NC˙006570.2— Francisella tularensis subsp. tularensis SCHU S4 chromosome
gi—134301169—ref—NC˙009257.1— Francisella tularensis subsp. tularensis WY96-3418 chromosome
gi—111219505—ref—NC˙008278.1— Frankia alni ACN14a chromosome
gi—86738724—ref—NC˙007777.1— Frankia sp. CcI3 chromosome
gi—158311867—ref—NC˙009921.1— Frankia sp. EAN1pec chromosome
gi—19703352—ref—NC˙003454.1— Fusobacterium nucleatum subsp. nucleatum ATCC 25586 chromosome
gi—56418535—ref—NC˙006510.1— Geobacillus kaustophilus HTA426 chromosome
gi—138893679—ref—NC˙009328.1— Geobacillus thermodenitrificans NG80-2 chromosome
gi—78221228—ref—NC˙007517.1— Geobacter metallireducens GS-15 chromosome
gi—400756305—ref—NC˙002939.5— Geobacter sulfurreducens PCA chromosome
gi—148262085—ref—NC˙009483.1— Geobacter uraniireducens Rf4 chromosome
gi—37519569—ref—NC˙005125.1— Gloeobacter violaceus PCC 7421 chromosome
gi—162145846—ref—NC˙010125.1— Gluconacetobacter diazotrophicus PAl 5 chromosome
gi—58038491—ref—NC˙006677.1— Gluconobacter oxydans 621H chromosome
gi—120434372—ref—NC˙008571.1— Gramella forsetii KT0803 chromosome
gi—114326664—ref—NC˙008343.1— Granulibacter bethesdensis CGDNIH1 chromosome
gi—33151282—ref—NC˙002940.2— Haemophilus ducreyi 35000HP chromosome
gi—162960935—ref—NC˙007146.2— Haemophilus influenzae 86-028NP chromosome
gi—148825133—ref—NC˙009566.1— Haemophilus influenzae PittEE chromosome
gi—148826757—ref—NC˙009567.1— Haemophilus influenzae PittGG chromosome
gi—16271976—ref—NC˙000907.1— Haemophilus influenzae Rd KW20 chromosome
gi—113460149—ref—NC˙008309.1— Haemophilus somnus 129PT chromosome
gi—83642913—ref—NC˙007645.1— Hahella chejuensis KCTC 2396 chromosome
gi—121996810—ref—NC˙008789.1— Halorhodospira halophila SL1 chromosome
gi—109946640—ref—NC˙008229.1— Helicobacter acinonychis str. Sheeba chromosome
gi—32265499—ref—NC˙004917.1— Helicobacter hepaticus ATCC 51449 chromosome
gi—15644634—ref—NC˙000915.1— Helicobacter pylori 26695 chromosome
gi—108562424—ref—NC˙008086.1— Helicobacter pylori HPAG1 chromosome
gi—15611071—ref—NC˙000921.1— Helicobacter pylori J99 chromosome
gi—134093294—ref—NC˙009138.1— Herminiimonas arsenicoxydans chromosome
gi—159896533—ref—NC˙009972.1— Herpetosiphon aurantiacus DSM 785 chromosome
gi—114797051—ref—NC˙008358.1— Hyphomonas neptunium ATCC 15444 chromosome
gi—56459112—ref—NC˙006512.1— Idiomarina loihiensis L2TR chromosome
gi—89052491—ref—NC˙007802.1— Jannaschia sp. CCS1 chromosome
gi—152979768—ref—NC˙009659.1— Janthinobacterium sp. Marseille chromosome
gi—255961475—ref—NC˙009664.2— Kineococcus radiotolerans SRS30216 chromosome
gi—152968582—ref—NC˙009648.1— Klebsiella pneumoniae subsp. pneumoniae MGH 78578 chromosome

Table D.1 Assembled library sequence from NCBI. Continued

Library List
gi—159162017—ref—NC˙006814.3— Lactobacillus acidophilus NCFM chromosome
gi—116332681—ref—NC˙008497.1— Lactobacillus brevis ATCC 367
gi—116493574—ref—NC˙008526.1— Lactobacillus casei ATCC 334 chromosome
gi—104773257—ref—NC˙008054.1— Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 chromosome
gi—116513228—ref—NC˙008529.1— Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365 chromosome
gi—116628683—ref—NC˙008530.1— Lactobacillus gasseri ATCC 33323 chromosome
gi—161506634—ref—NC˙010080.1— Lactobacillus helveticus DPC 4571
gi—42518084—ref—NC˙005362.1— Lactobacillus johnsonii NCC 533
gi—380031102—ref—NC˙004567.2— Lactobacillus plantarum WCFS1
gi—148543243—ref—NC˙009513.1— Lactobacillus reuteri DSM 20016 chromosome
gi—81427616—ref—NC˙007576.1— Lactobacillus sakei subsp. sakei 23K chromosome
gi—90960990—ref—NC˙007929.1— Lactobacillus salivarius UCC118 chromosome
gi—125622882—ref—NC˙009004.1— Lactococcus lactis subsp. cremoris MG1363 chromosome
gi—116510843—ref—NC˙008527.1— Lactococcus lactis subsp. cremoris SK11
gi—15671982—ref—NC˙002662.1— Lactococcus lactis subsp. lactis Il1403 chromosome
gi—94986445—ref—NC˙008011.1— Lawsonia intracellularis PHE/MN1-00 chromosome
gi—295815281—ref—NC˙009494.2— Legionella pneumophila str. Corby chromosome
gi—54292964—ref—NC˙006369.1— Legionella pneumophila str. Lens
gi—54295983—ref—NC˙006368.1— Legionella pneumophila str. Paris
gi—52840256—ref—NC˙002942.5— Legionella pneumophila subsp. pneumophila str. Philadelphia 1 chromosome
gi—50953925—ref—NC˙006087.1— Leifsonia xyli subsp. xyli str. CTCB07 chromosome
gi—116329799—ref—NC˙008510.1— Leptospira borgpetersenii serovar Hardjo-bovis str. JB197 chromosome 1
gi—116332445—ref—NC˙008511.1— Leptospira borgpetersenii serovar Hardjo-bovis JB197 chromosome chromosome 2
gi—116326852—ref—NC˙008508.1— Leptospira borgpetersenii serovar Hardjo-bovis str. L550 chromosome 1
gi—116329556—ref—NC˙008509.1— Leptospira borgpetersenii serovar Hardjo-bovis L550 chromosome chromosome 2
gi—45655914—ref—NC˙005823.1— Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130 chromosome I
gi—45655585—ref—NC˙005824.1— Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130 chromosome II
gi—294827553—ref—NC˙004342.2— Leptospira interrogans serovar Lai str. 56601 chromosome I
gi—294653513—ref—NC˙004343.2— Leptospira interrogans serovar Lai str. 56601 chromosome II
gi—116617174—ref—NC˙008531.1— Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293 chromosome
gi—16799079—ref—NC˙003212.1— Listeria innocua Clip11262
gi—16802048—ref—NC˙003210.1— Listeria monocytogenes EGD-e
gi—85700163—ref—NC˙002973.6— Listeria monocytogenes serotype 4b str. F2365 chromosome
gi—116871422—ref—NC˙008555.1— Listeria welshimeri serovar 6b str. SLCC5334 chromosome
gi—117923318—ref—NC˙008576.1— Magnetococcus marinus MC-1 chromosome
gi—83309099—ref—NC˙007626.1— Magnetospirillum magneticum AMB-1 chromosome
gi—52424055—ref—NC˙006300.1— Mannheimia succiniciproducens MBEL55E chromosome
gi—114568554—ref—NC˙008347.1— Maricaulis maris MCS10 chromosome
gi—120552944—ref—NC˙008740.1— Marinobacter aquaeolei VT8 chromosome
gi—152994043—ref—NC˙009654.1— Marinomonas sp. MWYL1 chromosome
gi—50364815—ref—NC˙006055.1— Mesoplasma florum L1 chromosome
gi—57165207—ref—NC˙002678.2— Mesorhizobium loti MAFF303099 chromosome
gi—124265193—ref—NC˙008825.1— Methylibium petroleiphilum PM1 chromosome
gi—91774356—ref—NC˙007947.1— Methylobacillus flagellatus KT
gi—163849457—ref—NC˙010172.1— Methylobacterium extorquens PA1 chromosome
gi—77128441—ref—NC˙002977.6— Methylococcus capsulatus str. Bath chromosome
gi—166362741—ref—NC˙010296.1— Microcystis aeruginosa NIES-843 chromosome
gi—83588874—ref—NC˙007644.1— Moorella thermoacetica ATCC 39073 chromosome
gi—118462219—ref—NC˙008595.1— Mycobacterium avium 104 chromosome
gi—41406098—ref—NC˙002944.2— Mycobacterium avium subsp. paratuberculosis K-10
gi—31791177—ref—NC˙002945.3— Mycobacterium bovis AF2122/97 chromosome
gi—121635883—ref—NC˙008769.1— Mycobacterium bovis BCG str. Pasteur 1173P2 chromosome
gi—145220606—ref—NC˙009338.1— Mycobacterium gilvum PYR-GCK chromosome
gi—15826865—ref—NC˙002677.1— Mycobacterium leprae TN chromosome
gi—118467340—ref—NC˙008596.1— Mycobacterium smegmatis str. MC2 155 chromosome

Table D.1 Assembled library sequence from NCBI. Continued

Library List
gi—126432613—ref—NC˙009077.1— Mycobacterium sp. JLS chromosome
gi—119866057—ref—NC˙008705.1— Mycobacterium sp. KMS chromosome
gi—108796981—ref—NC˙008146.1— Mycobacterium sp. MCS chromosome
gi—50953765—ref—NC˙002755.2— Mycobacterium tuberculosis CDC1551 chromosome
gi—148821191—ref—NC˙009565.1— Mycobacterium tuberculosis F11 chromosome
gi—148659757—ref—NC˙009525.1— Mycobacterium tuberculosis H37Ra chromosome
i—448814763—ref—NC˙000962.3— Mycobacterium tuberculosis H37Rv complete genome
gi—118615919—ref—NC˙008611.1— Mycobacterium ulcerans Agy99 chromosome
gi—120401028—ref—NC˙008726.1— Mycobacterium vanbaalenii PYR-1 chromosome
gi—148377268—ref—NC˙009497.1— Mycoplasma agalactiae PG2 chromosome
gi—83319253—ref—NC˙007633.1— Mycoplasma capricolum subsp. capricolum ATCC 27343 chromosome
gi—294660180—ref—NC˙004829.2— Mycoplasma gallisepticum str. R(low) chromosome
gi—108885074—ref—NC˙000908.2— Mycoplasma genitalium G37
gi—54019969—ref—NC˙006360.1— Mycoplasma hyopneumoniae 232 chromosome
gi—72080342—ref—NC˙007332.1— Mycoplasma hyopneumoniae 7448 chromosome
gi—71893359—ref—NC˙007295.1— Mycoplasma hyopneumoniae J chromosome
gi—47458835—ref—NC˙006908.1— Mycoplasma mobile 163K
gi—127763381—ref—NC˙005364.2— Mycoplasma mycoides subsp. mycoides SC str. PG1 chromosome
gi—26553452—ref—NC˙004432.1— Mycoplasma penetrans HF-2
gi—13507739—ref—NC˙000912.1— Mycoplasma pneumoniae M129 chromosome
gi—15828471—ref—NC˙002771.1— Mycoplasma pulmonis UAB CTIP
gi—71894025—ref—NC˙007294.1— Mycoplasma synoviae 53
gi—108756767—ref—NC˙008095.1— Myxococcus xanthus DK 1622 chromosome
gi—59800473—ref—NC˙002946.2— Neisseria gonorrhoeae FA 1090 chromosome
gi—161869018—ref—NC˙010120.1— Neisseria meningitidis 053442 chromosome
gi—121633901—ref—NC˙008767.1— Neisseria meningitidis FAM18 chromosome
gi—77358697—ref—NC˙003112.2— Neisseria meningitidis MC58 chromosome
gi—15793034—ref—NC˙003116.1— Neisseria meningitidis Z2491 chromosome
gi—88607955—ref—NC˙007798.1— Neorickettsia sennetsu str. Miyayama chromosome
gi—152989753—ref—NC˙009662.1— Nitratiruptor sp. SB155-2
gi—92115633—ref—NC˙007964.1— Nitrobacter hamburgensis X14 chromosome
gi—75674199—ref—NC˙007406.1— Nitrobacter winogradskyi Nb-255 chromosome
gi—77163561—ref—NC˙007484.1— Nitrosococcus oceani ATCC 19707 chromosome
gi—30248031—ref—NC˙004757.1— Nitrosomonas europaea ATCC 19718 chromosome
gi—114330036—ref—NC˙008344.1— Nitrosomonas eutropha C91 chromosome
gi—82701135—ref—NC˙007614.1— Nitrosospira multiformis ATCC 25196 chromosome
gi—54021964—ref—NC˙006361.1— Nocardia farcinica IFM 10152 chromosome
gi—119714272—ref—NC˙008699.1— Nocardioides sp. JS614 chromosome
gi—17227497—ref—NC˙003272.1— Nostoc sp. PCC 7120 chromosome
gi—87198026—ref—NC˙007794.1— Novosphingobium aromaticivorans DSM 12444 chromosome
gi—23097455—ref—NC˙004193.1— Oceanobacillus iheyensis HTE831 chromosome
gi—153007346—ref—NC˙009667.1— Ochrobactrum anthropi ATCC 49188 chromosome 1
gi—153010078—ref—NC˙009668.1— Ochrobactrum anthropi ATCC 49188 chromosome 2
gi—116490126—ref—NC˙008528.1— Oenococcus oeni PSU-1
gi—255961248—ref—NC˙005303.2— Onion yellows phytoplasma OY-M
gi—148283997—ref—NC˙009488.1— Orientia tsutsugamushi str. Boryong
gi—150006674—ref—NC˙009615.1— Parabacteroides distasonis ATCC 8503 chromosome
gi—119382757—ref—NC˙008686.1— Paracoccus denitrificans PD1222 chromosome 1
gi—119385557—ref—NC˙008687.1— Paracoccus denitrificans PD1222 chromosome 2
gi—154250456—ref—NC˙009719.1— Parvibaculum lavamentivorans DS-1 chromosome
gi—15601865—ref—NC˙002663.1— Pasteurella multocida subsp. multocida str. Pm70 chromosome
gi—50118965—ref—NC˙004547.2— Pectobacterium atrosepticum SCRI1043 chromosome
gi—116491818—ref—NC˙008525.1— Pediococcus pentosaceus ATCC 25745
gi—90960985—ref—NC˙007498.2— Pelobacter carbinolicus DSM 2380 chromosome
gi—118578449—ref—NC˙008609.1— Pelobacter propionicus DSM 2379 chromosome

Table D.1 Assembled library sequence from NCBI. Continued

Library List
gi—147676335—ref—NC˙009454.1— Pelotomaculum thermopropionicum SI chromosome
gi—160901491—ref—NC˙010003.1— Petrotoga mobilis SJ95 chromosome
gi—54307237—ref—NC˙006370.1— Photobacterium profundum SS9 chromosome 1
gi—54301680—ref—NC˙006371.1— Photobacterium profundum SS9 chromosome 2
gi—37524032—ref—NC˙005126.1— Photorhabdus luminescens subsp. laumondii TTO1
gi—121602919—ref—NC˙008781.1— Polaromonas naphthalenivorans CJ2 chromosome
gi—91785913—ref—NC˙007948.1— Polaromonas sp. JS666 chromosome
gi—145588189—ref—NC˙009379.1— Polynucleobacter necessarius subsp. asymbioticus QLW-P1DMWA-1 chromosome
gi—34539880—ref—NC˙002950.2— Porphyromonas gingivalis W83 chromosome
gi—123967536—ref—NC˙008816.1— Prochlorococcus marinus str. AS9601
gi—159902540—ref—NC˙009976.1— Prochlorococcus marinus str. MIT 9211
gi—157412338—ref—NC˙009840.1— Prochlorococcus marinus str. MIT 9215 chromosome
gi—126695337—ref—NC˙009091.1— Prochlorococcus marinus str. MIT 9301
gi—124021714—ref—NC˙008820.1— Prochlorococcus marinus str. MIT 9303 chromosome
gi—78778385—ref—NC˙007577.1— Prochlorococcus marinus str. MIT 9312
gi—33862273—ref—NC˙005071.1— Prochlorococcus marinus str. MIT 9313 chromosome
gi—123965234—ref—NC˙008817.1— Prochlorococcus marinus str. MIT 9515
gi—124024712—ref—NC˙008819.1— Prochlorococcus marinus str. NATL1A
gi—162958048—ref—NC˙007335.2— Prochlorococcus marinus str. NATL2A chromosome
gi—33239452—ref—NC˙005042.1— Prochlorococcus marinus subsp. marinus str. CCMP1375 chromosome
gi—33860560—ref—NC˙005072.1— Prochlorococcus marinus subsp. pastoris str. CCMP1986 chromosome
gi—50841496—ref—NC˙006085.1— Propionibacterium acnes KPA171202 chromosome
gi—109896332—ref—NC˙008228.1— Pseudoalteromonas atlantica T6c chromosome
gi—77358982—ref—NC˙007481.1— Pseudoalteromonas haloplanktis TAC125 chromosome I
gi—77361923—ref—NC˙007482.1— Pseudoalteromonas haloplanktis TAC125 chromosome II
gi—152983466—ref—NC˙009656.1— Pseudomonas aeruginosa PA7 chromosome
gi—110645304—ref—NC˙002516.2— Pseudomonas aeruginosa PAO1 chromosome
gi—116048575—ref—NC˙008463.1— Pseudomonas aeruginosa UCBPP-PA14 chromosome
gi—104779316—ref—NC˙008027.1— Pseudomonas entomophila L48 chromosome
gi—255961261—ref—NC˙007492.2— Pseudomonas fluorescens Pf0-1 chromosome
gi—146305042—ref—NC˙009439.1— Pseudomonas mendocina ymp chromosome
gi—70728250—ref—NC˙004129.6— Pseudomonas protegens Pf-5 chromosome
gi—148545259—ref—NC˙009512.1— Pseudomonas putida F1 chromosome
gi—167031021—ref—NC˙010322.1— Pseudomonas putida GB-1 chromosome
gi—26986745—ref—NC˙002947.3— Pseudomonas putida KT2440 chromosome
gi—146280397—ref—NC˙009434.1— Pseudomonas stutzeri A1501 chromosome
gi—71733195—ref—NC˙005773.3— Pseudomonas syringae pv. phaseolicola 1448A chromosome
gi—66043271—ref—NC˙007005.1— Pseudomonas syringae pv. syringae B728a chromosome
gi—28867243—ref—NC˙004578.1— Pseudomonas syringae pv. tomato str. DC3000 chromosome
gi—71064581—ref—NC˙007204.1— Psychrobacter arcticus 273-4 chromosome
gi—93004831—ref—NC˙007969.1— Psychrobacter cryohalolentis K5 chromosome
gi—148651817—ref—NC˙009524.1— Psychrobacter sp. PRwf-1 chromosome
gi—119943794—ref—NC˙008709.1— Psychromonas ingrahamii 37 chromosome
gi—113866031—ref—NC˙008313.1— Ralstonia eutropha H16 chromosome 1
gi—116693960—ref—NC˙008314.1— Ralstonia eutropha H16 chromosome 2
gi—73539706—ref—NC˙007347.1— Ralstonia eutropha JMP134 chromosome 1
gi—73537298—ref—NC˙007348.1— Ralstonia eutropha JMP134 chromosome 2
gi—17544719—ref—NC˙003295.1— Ralstonia solanacearum GMI1000 chromosome
gi—163838769—ref—NC˙010168.1— Renibacterium salmoninarum ATCC 33209 chromosome
gi—86355669—ref—NC˙007761.1— Rhizobium etli CFN 42 chromosome
gi—116249766—ref—NC˙008380.1— Rhizobium leguminosarum bv. viciae 3841 chromosome
gi—77461965—ref—NC˙007493.1— Rhodobacter sphaeroides 2.4.1 chromosome 1
gi—77464988—ref—NC˙007494.1— Rhodobacter sphaeroides 2.4.1 chromosome 2
gi—146276058—ref—NC˙009428.1— Rhodobacter sphaeroides ATCC 17025 chromosome
gi—126460778—ref—NC˙009049.1— Rhodobacter sphaeroides ATCC 17029 chromosome 1

Table D.1 Assembled library sequence from NCBI. Continued

Library List
gi—126463752—ref—NC˙009050.1— Rhodobacter sphaeroides ATCC 17029 chromosome 2
gi—111017022—ref—NC˙008268.1— Rhodococcus jostii RHA1 chromosome
gi—89898822—ref—NC˙007908.1— Rhodoferax ferrireducens T118 chromosome
gi—32470666—ref—NC˙005027.1— Rhodopirellula baltica SH 1 chromosome
gi—115522030—ref—NC˙008435.1— Rhodopseudomonas palustris BisA53 chromosome
gi—90421528—ref—NC˙007925.1— Rhodopseudomonas palustris BisB18 chromosome
gi—91974482—ref—NC˙007958.1— Rhodopseudomonas palustris BisB5 chromosome
gi—39933080—ref—NC˙005296.1— Rhodopseudomonas palustris CGA009 chromosome
gi—86747127—ref—NC˙007778.1— Rhodopseudomonas palustris HaA2 chromosome
gi—83591340—ref—NC˙007643.1— Rhodospirillum rubrum ATCC 11170 chromosome
gi—157825125—ref—NC˙009881.1— Rickettsia akari str. Hartford chromosome
gi—157826385—ref—NC˙009883.1— Rickettsia bellii OSU 85-389 chromosome
gi—91204815—ref—NC˙007940.1— Rickettsia bellii RML369-C chromosome
gi—157803189—ref—NC˙009879.1— Rickettsia canadensis str. McKiel
gi—15891923—ref—NC˙003103.1— Rickettsia conorii str. Malish 7
gi—67458392—ref—NC˙007109.1— Rickettsia felis URRWXCal2 chromosome
gi—157964072—ref—NC˙009900.1— Rickettsia massiliae MTU5 chromosome
gi—15603881—ref—NC˙000963.1— Rickettsia prowazekii str. Madrid E chromosome
gi—157827862—ref—NC˙009882.1— Rickettsia rickettsii str. 'Sheila Smith' chromosome
gi—319717301—ref—NC˙010263.2— Rickettsia rickettsii str. Iowa chromosome
gi—51473215—ref—NC˙006142.1— Rickettsia typhi str. Wilmington
gi—156740028—ref—NC˙009767.1— Roseiflexus castenholzii DSM 13941 chromosome
gi—148654188—ref—NC˙009523.1— Roseiflexus sp. RS-1 chromosome
gi—110677421—ref—NC˙008209.1— Roseobacter denitrificans OCh 114 chromosome
gi—108802856—ref—NC˙008148.1— Rubrobacter xylanophilus DSM 9941 chromosome
gi—56694928—ref—NC˙003911.11— Ruegeria pomeroyi DSS-3 chromosome
gi—99079841—ref—NC˙008044.1— Ruegeria sp. TM1040 chromosome
gi—90019649—ref—NC˙007912.1— Saccharophagus degradans 2-40 chromosome
gi—134096620—ref—NC˙009142.1— Saccharopolyspora erythraea NRRL 2338 chromosome
gi—83814055—ref—NC˙007677.1— Salinibacter ruber DSM 13855 chromosome
gi—159035674—ref—NC˙009953.1— Salinispora arenicola CNS-205 chromosome
gi—145592566—ref—NC˙009380.1— Salinispora tropica CNB-440 chromosome
gi—161501984—ref—NC˙010067.1— Salmonella enterica subsp. arizonae serovar 62:z4
gi—62178570—ref—NC˙006905.1— Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67 chromosome
gi—56412276—ref—NC˙006511.1— Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150 chromosome
gi—161612313—ref—NC˙010102.1— Salmonella enterica subsp. enterica serovar Paratyphi B str. SPB7 chromosome
gi—16758993—ref—NC˙003198.1— Salmonella enterica subsp. enterica serovar Typhi str. CT18 chromosome
gi—29140543—ref—NC˙004631.1— Salmonella enterica subsp. enterica serovar Typhi str. Ty2 chromosome
gi—16763390—ref—NC˙003197.1— Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 chromosome
gi—157368249—ref—NC˙009832.1— Serratia proteamaculans 568 chromosome
gi—119773142—ref—NC˙008700.1— Shewanella amazonensis SB2B chromosome
gi—126172257—ref—NC˙009052.1— Shewanella baltica OS155 chromosome
gi—152998555—ref—NC˙009665.1— Shewanella baltica OS185 chromosome
gi—160873126—ref—NC˙009997.1— Shewanella baltica OS195 chromosome
gi—91791369—ref—NC˙007954.1— Shewanella denitrificans OS217
gi—114561188—ref—NC˙008345.1— Shewanella frigidimarina NCIMB 400 chromosome
gi—127510935—ref—NC˙009092.1— Shewanella loihica PV-4 chromosome
gi—414561716—ref—NC˙004347.2— Shewanella oneidensis MR-1 chromosome
gi—157959830—ref—NC˙009901.1— Shewanella pealeana ATCC 700345 chromosome
gi—146291111—ref—NC˙009438.1— Shewanella putrefaciens CN-32 chromosome
gi—157373141—ref—NC˙009831.1— Shewanella sediminis HAW-EB3 chromosome
gi—117918459—ref—NC˙008577.1— Shewanella sp. ANA-3 chromosome 1
gi—113968346—ref—NC˙008321.1— Shewanella sp. MR-4 chromosome
gi—114045513—ref—NC˙008322.1— Shewanella sp. MR-7 chromosome
gi—120596833—ref—NC˙008750.1— Shewanella sp. W3-18-1 chromosome

Table D.1 Assembled library sequence from NCBI. Continued

Library List
gi—82542618—ref—NC˙007613.1— Shigella boydii Sb227 chromosome
gi—82775382—ref—NC˙007606.1— Shigella dysenteriae Sd197
gi—30061571—ref—NC˙004741.1— Shigella flexneri 2a str. 2457T
gi—344915202—ref—NC˙004337.2— Shigella flexneri 2a str. 301 chromosome
gi—110804074—ref—NC˙008258.1— Shigella flexneri 5 str. 8401 chromosome
gi—74310614—ref—NC˙007384.1— Shigella sonnei Ss046 chromosome
gi—150395228—ref—NC˙009636.1— Sinorhizobium medicae WSM419 chromosome
gi—15963753—ref—NC˙003047.1— Sinorhizobium meliloti 1021 chromosome
gi—85057978—ref—NC˙007712.1— Sodalis glossinidius str. 'morsitans' chromosome
gi—162448269—ref—NC˙010162.1— Sorangium cellulosum 'So ce 56' chromosome
gi—148552929—ref—NC˙009511.1— Sphingomonas wittichii RW1 chromosome
gi—103485498—ref—NC˙008048.1— Sphingopyxis alaskensis RB2256 chromosome
gi—82749777—ref—NC˙007622.1— Staphylococcus aureus RF122
gi—57650036—ref—NC˙002951.2— Staphylococcus aureus subsp. aureus COL chromosome
gi—150392480—ref—NC˙009632.1— Staphylococcus aureus subsp. aureus JH1 chromosome
gi—148266447—ref—NC˙009487.1— Staphylococcus aureus subsp. aureus JH9 chromosome
gi—49482253—ref—NC˙002952.2— Staphylococcus aureus subsp. aureus MRSA252 chromosome
gi—49484912—ref—NC˙002953.3— Staphylococcus aureus subsp. aureus MSSA476 chromosome
gi—21281729—ref—NC˙003923.1— Staphylococcus aureus subsp. aureus MW2
gi—156978331—ref—NC˙009782.1— Staphylococcus aureus subsp. aureus Mu3
gi—57634611—ref—NC˙002758.2— Staphylococcus aureus subsp. aureus Mu50 chromosome
gi—29165615—ref—NC˙002745.2— Staphylococcus aureus subsp. aureus N315 chromosome
gi—88193823—ref—NC˙007795.1— Staphylococcus aureus subsp. aureus NCTC 8325 chromosome
gi—87159884—ref—NC˙007793.1— Staphylococcus aureus subsp. aureus USA300˙FPR3757 chromosome
gi—161508266—ref—NC˙010079.1— Staphylococcus aureus subsp. aureus USA300˙TCH1516 chromosome
gi—151220212—ref—NC˙009641.1— Staphylococcus aureus subsp. aureus str. Newman chromosome
gi—27466918—ref—NC˙004461.1— Staphylococcus epidermidis ATCC 12228 chromosome
gi—57865352—ref—NC˙002976.3— Staphylococcus epidermidis RP62A
gi—70725001—ref—NC˙007168.1— Staphylococcus haemolyticus JCSC1435 chromosome
gi—73661309—ref—NC˙007350.1— Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305
gi—22536185—ref—NC˙004116.1— Streptococcus agalactiae 2603V/R chromosome
gi—76786714—ref—NC˙007432.1— Streptococcus agalactiae A909 chromosome
gi—25010075—ref—NC˙004368.1— Streptococcus agalactiae NEM316
gi—157149651—ref—NC˙009785.1— Streptococcus gordonii str. Challis substr. CH1 chromosome
gi—347750429—ref—NC˙004350.2— Streptococcus mutans UA159 chromosome
gi—116515308—ref—NC˙008533.1— Streptococcus pneumoniae D39 chromosome
gi—15902044—ref—NC˙003098.1— Streptococcus pneumoniae R6
gi—194172857—ref—NC˙003028.3— Streptococcus pneumoniae TIGR4 chromosome
gi—94989509—ref—NC˙008022.1— Streptococcus pyogenes MGAS10270
gi—50913346—ref—NC˙006086.1— Streptococcus pyogenes MGAS10394 chromosome
gi—94993396—ref—NC˙008024.1— Streptococcus pyogenes MGAS10750 chromosome
gi—94991497—ref—NC˙008023.1— Streptococcus pyogenes MGAS2096 chromosome
gi—21909536—ref—NC˙004070.1— Streptococcus pyogenes MGAS315 chromosome
gi—71909814—ref—NC˙007297.1— Streptococcus pyogenes MGAS5005 chromosome
gi—71902667—ref—NC˙007296.1— Streptococcus pyogenes MGAS6180 chromosome
gi—19745201—ref—NC˙003485.1— Streptococcus pyogenes MGAS8232 chromosome
gi—94987631—ref—NC˙008021.1— Streptococcus pyogenes MGAS9429 chromosome
gi—15674250—ref—NC˙002737.1— Streptococcus pyogenes SF370 chromosome
gi—28894912—ref—NC˙004606.1— Streptococcus pyogenes SSI-1 chromosome
gi—139472888—ref—NC˙009332.1— Streptococcus pyogenes str. Manfredo chromosome
gi—125716887—ref—NC˙009009.1— Streptococcus sanguinis SK36 chromosome
gi—146317663—ref—NC˙009442.1— Streptococcus suis 05ZYH33 chromosome
gi—146319850—ref—NC˙009443.1— Streptococcus suis 98HAH33
gi—55821993—ref—NC˙006449.1— Streptococcus thermophilus CNRZ1066 chromosome
gi—116626972—ref—NC˙008532.1— Streptococcus thermophilus LMD-9

Table D.1 Assembled library sequence from NCBI. Continued

Library List
gi—55820103—ref—NC˙006448.1— Streptococcus thermophilus LMG 18311 chromosome
gi—162960844—ref—NC˙003155.4— Streptomyces avermitilis MA-4680
gi—32141095—ref—NC˙003888.3— Streptomyces coelicolor A3(2) chromosome
gi—78776201—ref—NC˙007575.1— Sulfurimonas denitrificans DSM 1251 chromosome
gi—152991597—ref—NC˙009663.1— Sulfurovum sp. NBC37-1 chromosome
gi—51891138—ref—NC˙006177.1— Symbiobacterium thermophilum IAM 14863 chromosome
gi—56750010—ref—NC˙006576.1— Synechococcus elongatus PCC 6301 chromosome
gi—81298811—ref—NC˙007604.1— Synechococcus elongatus PCC 7942 chromosome
gi—113952711—ref—NC˙008319.1— Synechococcus sp. CC9311
gi—78211558—ref—NC˙007516.1— Synechococcus sp. CC9605
gi—78183584—ref—NC˙007513.1— Synechococcus sp. CC9902 chromosome
gi—86607503—ref—NC˙007776.1— Synechococcus sp. JA-2-3B'a(2-13) chromosome
gi—86604733—ref—NC˙007775.1— Synechococcus sp. JA-3-3Ab chromosome
gi—148241099—ref—NC˙009482.1— Synechococcus sp. RCC307 chromosome
gi—148238336—ref—NC˙009481.1— Synechococcus sp. WH 7803 chromosome
gi—33864539—ref—NC˙005070.1— Synechococcus sp. WH 8102
gi—16329170—ref—NC˙000911.1— Synechocystis sp. PCC 6803 chromosome
gi—116747452—ref—NC˙008554.1— Syntrophobacter fumaroxidans MPOB chromosome
gi—114565576—ref—NC˙008346.1— Syntrophomonas wolfei subsp. wolfei str. Goettingen chromosome
gi—85857845—ref—NC˙007759.1— Syntrophus aciditrophicus SB chromosome
gi—167036431—ref—NC˙010321.1— Thermoanaerobacter pseudethanolicus ATCC 33223 chromosome
gi—167038675—ref—NC˙010320.1— Thermoanaerobacter sp. X514 chromosome
gi—20806542—ref—NC˙003869.1— Thermoanaerobacter tengcongensis MB4 chromosome
gi—72160406—ref—NC˙007333.1— Thermobifida fusca YX chromosome
gi—150019913—ref—NC˙009616.1— Thermosipho melanesiensis BI429 chromosome
gi—22297544—ref—NC˙004113.1— Thermosynechococcus elongatus BP-1 chromosome
gi—157362870—ref—NC˙009828.1— Thermotoga lettingae TMO chromosome
gi—15642775—ref—NC˙000853.1— Thermotoga maritima MSB8 chromosome
gi—148269145—ref—NC˙009486.1— Thermotoga petrophila RKU-1 chromosome
gi—46198308—ref—NC˙005835.1— Thermus thermophilus HB27
gi—55979969—ref—NC˙006461.1— Thermus thermophilus HB8 chromosome
gi—74316018—ref—NC˙007404.1— Thiobacillus denitrificans ATCC 25259 chromosome
gi—118139508—ref—NC˙007520.2— Thiomicrospira crunogena XCL-2 chromosome
gi—42516522—ref—NC˙002967.9— Treponema denticola ATCC 35405 chromosome
gi—15638995—ref—NC˙000919.1— Treponema pallidum subsp. pallidum str. Nichols chromosome
gi—113473942—ref—NC˙008312.1— Trichodesmium erythraeum IMS101 chromosome
gi—28572175—ref—NC˙004551.1— Tropheryma whipplei TW08/27
gi—32447382—ref—NC˙004572.3— Tropheryma whipplei str. Twist
gi—13357558—ref—NC˙002162.1— Ureaplasma parvum serovar 3 str. ATCC 700970 chromosome
gi—121607004—ref—NC˙008786.1— Verminephrobacter eiseniae EF01-2 chromosome
gi—15640032—ref—NC˙002505.1— Vibrio cholerae O1 biovar El Tor str. N16961 chromosome I
gi—15600771—ref—NC˙002506.1— Vibrio cholerae O1 biovar El Tor str. N16961 chromosome II
gi—147671401—ref—NC˙009456.1— Vibrio cholerae O395 chromosome 1
gi—147673035—ref—NC˙009457.1— Vibrio cholerae O395 chromosome 2
gi—172087630—ref—NC˙006840.2— Vibrio fischeri ES114 chromosome I
gi—172087787—ref—NC˙006841.2— Vibrio fischeri ES114 chromosome II
gi—156972381—ref—NC˙009783.1— Vibrio harveyi ATCC BAA-1116 chromosome I
gi—156975952—ref—NC˙009784.1— Vibrio harveyi ATCC BAA-1116 chromosome II
gi—28896774—ref—NC˙004603.1— Vibrio parahaemolyticus RIMD 2210633 chromosome 1
gi—28899855—ref—NC˙004605.1— Vibrio parahaemolyticus RIMD 2210633 chromosome 2
gi—326423644—ref—NC˙004459.3— Vibrio vulnificus CMCP6 chromosome I
gi—326424156—ref—NC˙004460.2— Vibrio vulnificus CMCP6 chromosome II
gi—37678184—ref—NC˙005139.1— Vibrio vulnificus YJ016 chromosome I
gi—37675660—ref—NC˙005140.1— Vibrio vulnificus YJ016 chromosome II
gi—32490749—ref—NC˙004344.2— Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis chromosome

Table D.1 Assembled library sequence from NCBI. Continued

Library List
gi—42519920—ref—NC˙002978.6— Wolbachia endosymbiont of Drosophila melanogaster
gi—58584261—ref—NC˙006833.1— Wolbachia endosymbiont strain TRS of Brugia malayi
gi—34556458—ref—NC˙005090.1— Wolinella succinogenes DSM 1740 chromosome
gi—154243958—ref—NC˙009720.1— Xanthobacter autotrophicus Py2 chromosome
gi—21240774—ref—NC˙003919.1— Xanthomonas axonopodis pv. citri str. 306 chromosome
gi—66766352—ref—NC˙007086.1— Xanthomonas campestris pv. campestris str. 8004 chromosome
gi—21229478—ref—NC˙003902.1— Xanthomonas campestris pv. campestris str. ATCC 33913 chromosome
gi—78045556—ref—NC˙007508.1— Xanthomonas campestris pv. vesicatoria str. 85-10 chromosome
gi—58579623—ref—NC˙006834.1— Xanthomonas oryzae pv. oryzae KACC 10331 chromosome
gi—84621657—ref—NC˙007705.1— Xanthomonas oryzae pv. oryzae MAFF 311018 chromosome
gi—57014152—ref—NC˙002488.3— Xylella fastidiosa 9a5c chromosome
gi—28197945—ref—NC˙004556.1— Xylella fastidiosa Temecula1 chromosome
gi—123440403—ref—NC˙008800.1— Yersinia enterocolitica subsp. enterocolitica 8081 chromosome
gi—162418099—ref—NC˙010159.1— Yersinia pestis Angola chromosome
gi—108805998—ref—NC˙008150.1— Yersinia pestis Antiqua chromosome
gi—16120353—ref—NC˙003143.1— Yersinia pestis CO92 chromosome
gi—22123922—ref—NC˙004088.1— Yersinia pestis KIM 10 chromosome
gi—108810166—ref—NC˙008149.1— Yersinia pestis Nepal516 chromosome
gi—145597324—ref—NC˙009381.1— Yersinia pestis Pestoides F chromosome
gi—45439865—ref—NC˙005810.1— Yersinia pestis biovar Microtus str. 91001 chromosome
gi—153946813—ref—NC˙009708.1— Yersinia pseudotuberculosis IP 31758 chromosome
gi—51594359—ref—NC˙006155.1— Yersinia pseudotuberculosis IP 32953 chromosome
gi—283856168—ref—NC˙006526.2— Zymomonas mobilis subsp. mobilis ZM4 chromosome

Table D.2: Assembled unknown sequence from NCBI.

Unknown Assembled Sequences
gi—215481761—ref—NC˙011595.1— Acinetobacter baumannii AB307-0294
gi—375133618—ref—NC˙016603.1— Acinetobacter calcoaceticus PHEA-2 chromosome
gi—384141246—ref—NC˙017171.1— Acinetobacter baumannii MDR-ZJ06 chromosome
gi—523529121—ref—NC˙021733.1— Acinetobacter baumannii BJAB0715
gi—217957581—ref—NC˙011658.1— Bacillus cereus AH187 chromosome
gi—222093774—ref—NC˙011969.1— Bacillus cereus Q1 chromosome
gi—229599883—ref—NC˙012659.1— Bacillus anthracis str. A0248
gi—316994385—ref—NC˙013791.2— Bacillus pseudofirmus OF4 chromosome
gi—384157612—ref—NC˙017188.1— Bacillus amyloliquefaciens TA208 chromosome
gi—449086670—ref—NC˙020238.1— Bacillus thuringiensis serovar kurstaki str. HD73
gi—530612796—ref—NC˙022081.1— Bacillus amyloliquefaciens subsp. plantarum UCMB5113
gi—375356399—ref—NC˙016776.1— Bacteroides fragilis 638R
gi—403529933—ref—NC˙018533.1— Bartonella quintana RM-11 chromosome
gi—384200575—ref—NC˙017221.1— Bifidobacterium longum subsp. longum KACC 91563 chromosome
gi—386866198—ref—NC˙017834.1— Bifidobacterium animalis subsp. animalis ATCC 25527 chromosome
gi—384202563—ref—NC˙017223.1— Bordetella pertussis CS chromosome
gi—408414082—ref—NC˙018518.1— Bordetella pertussis 18323
gi—412337338—ref—NC˙019382.1— Bordetella bronchiseptica 253
gi—189022234—ref—NC˙010740.1— Brucella abortus S19 chromosome 2
gi—376274175—ref—NC˙016778.1— Brucella canis HSK A52141 chromosome 1
gi—384210366—ref—NC˙017246.1— Brucella melitensis M5-90 chromosome chromosome I
gi—170731356—ref—NC˙010508.1— Burkholderia cenocepacia MC0-3 chromosome 1
gi—172059067—ref—NC˙010551.1— Burkholderia ambifaria MC40-6 chromosome 1
gi—312794749—ref—NC˙014722.1— Burkholderia rhizoxinica HKI 454 chromosome
gi—386860126—ref—NC˙017831.1— Burkholderia pseudomallei 1026b chromosome 1
gi—488601775—ref—NC˙021173.1— Burkholderia thailandensis MSMB121 chromosome 1
gi—312134082—ref—NC˙014657.1— Caldicellulosiruptor owensensis OL chromosome
gi—384447320—ref—NC˙017279.1— Campylobacter jejuni subsp. jejuni IA3902 chromosome
i—543945414—ref—NC˙022351.1— Campylobacter jejuni subsp. jejuni 00-2538 genome
gi—293977746—ref—NC˙014004.1— Candidatus Sulcia muelleri DMIN chromosome
gi—330812975—ref—NC˙015380.1— Candidatus Pelagibacter sp. IMCC9063 chromosome
gi—440509586—ref—NC˙020075.1— Candidatus Blochmannia chromaiodes str. 640 chromosome
gi—527324386—ref—NC˙021894.1— Candidatus Carsonella ruddii DC
gi—237802433—ref—NC˙012686.1— Chlamydia trachomatis B/Jali20/OT chromosome
gi—376282008—ref—NC˙016798.1— Chlamydia trachomatis A2497
gi—407453476—ref—NC˙018619.1— Chlamydia psittaci 84/55 chromosome
i—478459453—ref—NC˙020968.1— Chlamydia trachomatis D/SotonD6 high quality draft genome sequence
gi—568111252—ref—NC˙023060.1— Chlamydia trachomatis C/TW-3
gi—330443755—ref—NC˙015408.1— Chlamydophila pecorum E58 chromosome
gi—384450095—ref—NC˙017287.1— Chlamydophila psittaci 6BC chromosome
gi—170754211—ref—NC˙010516.1— Clostridium botulinum B1 str. Okra chromosome
gi—226947222—ref—NC˙012563.1— Clostridium botulinum A2 str. Kyoto chromosome
gi—337735209—ref—NC˙015687.1— Clostridium acetobutylicum DSM 1731 chromosome
gi—479133135—ref—NC˙017175.1— Clostridium difficile M68
gi—383843666—ref—NC˙017178.1— Clostridium difficile complete genome
gi—384460459—ref—NC˙017297.1— Clostridium botulinum F str. 230613 chromosome
gi—550916528—ref—NC˙022571.1— Clostridium saccharobutylicum DSM 13864
gi—376244562—ref—NC˙016786.1— Corynebacterium diphtheriae HC01 chromosome
gi—511052542—ref—NC˙021351.1— Corynebacterium glutamicum SCgG1
gi—530314182—ref—NC˙022040.1— Corynebacterium glutamicum MB001
gi—374298386—ref—NC˙016629.1— Desulfovibrio africanus str. Walvis Bay chromosome
gi—387151873—ref—NC˙017310.1— Desulfovibrio vulgaris RCH1 chromosome
gi—218556939—ref—NC˙011742.1— Escherichia coli S88 chromosome
gi—218687878—ref—NC˙011745.1— Escherichia coli ED1a chromosome
gi—218703261—ref—NC˙011751.1— Escherichia coli UMN026 chromosome

Table D.2 Assembled unknown sequence from NCBI. Continued

Unknown Assembled Sequences
gi—238899406—ref—NC˙012759.1— Escherichia coli BW2952 chromosome
gi—254791136—ref—NC˙013008.1— Escherichia coli O157:H7 str. TW14359 chromosome
gi—260842239—ref—NC˙013353.1— Escherichia coli O103:H2 str. 12009
gi—387605479—ref—NC˙017626.1— Escherichia coli 042
gi—443615330—ref—NC˙020163.1— Escherichia coli APEC O78
gi—471332236—ref—NC˙020518.1— Escherichia coli str. K-12 substr. MDS42 DNA
gi—187930913—ref—NC˙010677.1— Francisella tularensis subsp. mediasiatica FSC147 chromosome
gi—379716390—ref—NC˙016933.1— Francisella tularensis TIGB03 chromosome
gi—387885754—ref—NC˙017909.1— Francisella noatunensis subsp. orientalis str. Toba 04 chromosome
gi—423049750—ref—NC˙019537.1— Francisella tularensis subsp. holarctica F92 chromosome
gi—312193897—ref—NC˙014666.1— Frankia sp. EuI1c chromosome
gi—336176139—ref—NC˙015656.1— Frankia symbiont of Datisca glomerata chromosome
gi—384898367—ref—NC˙017365.1— Helicobacter pylori F30
gi—385230889—ref—NC˙017381.1— Helicobacter pylori 2018 chromosome
gi—386745526—ref—NC˙017733.1— Helicobacter pylori HUP-B14 chromosome
gi—386748726—ref—NC˙017737.1— Helicobacter cetorum MIT 00-7128 chromosome
gi—425788638—ref—NC˙019560.1— Helicobacter pylori Aklavik117 chromosome
gi—526465356—ref—NC˙021882.1— Helicobacter pylori UM298
gi—238892256—ref—NC˙012731.1— Klebsiella pneumoniae NTUH-K2044 chromosome
gi—402778297—ref—NC˙018522.1— Klebsiella pneumoniae subsp. pneumoniae 1084 chromosome
gi—184152655—ref—NC˙010609.1— Lactobacillus reuteri JCM 1112
gi—191636824—ref—NC˙010999.1— Lactobacillus casei BL23 chromosome
gi—268318562—ref—NC˙013504.1— Lactobacillus johnsonii FI9785 chromosome
gi—315037230—ref—NC˙014724.1— Lactobacillus amylovorus GRL 1112 chromosome
gi—313122775—ref—NC˙014727.1— Lactobacillus delbrueckii subsp. bulgaricus ND02 chromosome
gi—347524522—ref—NC˙015975.1— Lactobacillus ruminis ATCC 27782 chromosome
gi—385812838—ref—NC˙017467.1— Lactobacillus helveticus H10 chromosome
gi—385839818—ref—NC˙017481.1— Lactobacillus salivarius CECT 5713 chromosome
gi—448819523—ref—NC˙020229.1— Lactobacillus plantarum ZJ316
gi—472405735—ref—NC˙020819.1— Lactobacillus brevis KB290 DNA
gi—529087732—ref—NC˙021181.2— Lactobacillus acidophilus La-14
gi—501145339—ref—NC˙021224.1— Lactobacillus plantarum subsp. plantarum P-8
gi—512590512—ref—NC˙021494.1— Lactobacillus reuteri I5007
gi—523512490—ref—NC˙021721.1— Lactobacillus casei LOCK919
gi—560151351—ref—NC˙022909.1— Lactobacillus johnsonii N6.2
gi—459284225—ref—NC˙020450.1— Lactococcus lactis subsp. lactis IO-1 DNA
gi—378775961—ref—NC˙016811.1— Legionella pneumophila subsp. pneumophila ATCC 43290 chromosome
gi—221229343—ref—NC˙011896.1— Mycobacterium leprae Br4923 chromosome
gi—224988383—ref—NC˙012207.1— Mycobacterium bovis BCG str. Tokyo 172 chromosome
gi—315441696—ref—NC˙014814.1— Mycobacterium gilvum Spyr1 chromosome
gi—340625033—ref—NC˙015848.1— Mycobacterium canettii CIPT 140010059 chromosome
gi—385989534—ref—NC˙017522.1— Mycobacterium tuberculosis CCDC5180 chromosome
gi—386003090—ref—NC˙017528.1— Mycobacterium tuberculosis RGTB423 chromosome
gi—387873410—ref—NC˙017904.1— Mycobacterium sp. MOTT36Y chromosome
gi—499074415—ref—NC˙021200.1— Mycobacterium avium subsp. paratuberculosis MAP4
gi—544161316—ref—NC˙022350.1— Mycobacterium tuberculosis str. Haarlem
gi—291319937—ref—NC˙013948.1— Mycoplasma agalactiae chromosome
gi—339320528—ref—NC˙015725.1— Mycoplasma bovis Hubei-1 chromosome
gi—385325086—ref—NC˙017502.1— Mycoplasma gallisepticum str. R(high) chromosome
gi—385326614—ref—NC˙017504.1— Mycoplasma pneumoniae FH chromosome
gi—401771165—ref—NC˙018413.1— Mycoplasma gallisepticum NC08˙2008.031-4-3P chromosome
gi—402550799—ref—NC˙018495.1— Mycoplasma genitalium M2321 chromosome
gi—479052799—ref—NC˙020076.1— Mycoplasma pneumoniae M129-B7
gi—479183780—ref—NC˙021025.1— Mycoplasma mycoides subsp. mycoides SC str. Gladysdale MU clone SC5
gi—525903163—ref—NC˙021831.1— Mycoplasma hyopneumoniae 7422

Table D.2 Assembled unknown sequence from NCBI. Continued

Unknown Assembled Sequences
gi—194097589—ref—NC˙011035.1— Neisseria gonorrhoeae NCCP11945 chromosome
gi—254804028—ref—NC˙013016.1— Neisseria meningitidis alpha14 chromosome
gi—385323172—ref—NC˙017501.1— Neisseria meningitidis 8013
gi—385854193—ref—NC˙017517.1— Neisseria meningitidis M01-240355 chromosome
gi—330806657—ref—NC˙015379.1— Pseudomonas brassicacearum subsp. brassicacearum NFM421 chromosome
gi—330500914—ref—NC˙015410.1— Pseudomonas mendocina NK-01 chromosome
gi—339492077—ref—NC˙015740.1— Pseudomonas stutzeri ATCC 17588 = LMG 11199 chromosome
gi—426406915—ref—NC˙019670.1— Pseudomonas sp. UW4 chromosome
gi—431799958—ref—NC˙019905.1— Pseudomonas putida HB3267
gi—558672313—ref—NC˙022808.1— Pseudomonas aeruginosa PA1
gi—568136993—ref—NC˙023064.1— Pseudomonas sp. TKP
gi—568179884—ref—NC˙023075.1— Pseudomonas monteilii SB3078
gi—568306739—ref—NC˙023149.1— Pseudomonas aeruginosa SCV20265
gi—221368938—ref—NC˙011958.1— Rhodobacter sphaeroides KD131 chromosome 2
gi—294675557—ref—NC˙014034.1— Rhodobacter capsulatus SB 1003 chromosome
gi—378722019—ref—NC˙016909.1— Rickettsia rickettsii str. Arizona chromosome
gi—379017167—ref—NC˙016914.1— Rickettsia rickettsii str. Hino chromosome
gi—379022404—ref—NC˙016929.1— Rickettsia canadensis str. CA410 chromosome
gi—379713087—ref—NC˙016931.1— Rickettsia massiliae str. AZT80 chromosome
gi—383489123—ref—NC˙017051.1— Rickettsia prowazekii str. Dachau chromosome
gi—383500935—ref—NC˙017058.1— Rickettsia australis str. Cutlack chromosome
gi—383842824—ref—NC˙017062.1— Rickettsia typhi str. B9991CWPP chromosome
gi—478693373—ref—NC˙020993.1— Rickettsia prowazekii str. Breinl
gi—198241740—ref—NC˙011205.1— Salmonella enterica subsp. enterica serovar Dublin str. CT˙02021853 chromosome
gi—339998036—ref—NC˙015761.1— Salmonella bongori NCTC 12419
gi—386589256—ref—NC˙017623.1— Salmonella enterica subsp. enterica serovar Heidelberg str. B182 chromosome
gi—488652559—ref—NC˙021176.1— Salmonella enterica subsp. enterica serovar Typhi str. Ty21a
gi—525855729—ref—NC˙021818.1— Salmonella enterica subsp. enterica Serovar Cubana str. CFSAN002050
gi—549722728—ref—NC˙022544.1— Salmonella enterica subsp. enterica serovar Typhimurium str. DT2 chromosome
gi—167621941—ref—NC˙010334.1— Shewanella halifaxensis HAW-EB4 chromosome
gi—217971216—ref—NC˙011663.1— Shewanella baltica OS223 chromosome
gi—386311792—ref—NC˙017566.1— Shewanella putrefaciens 200 chromosome
gi—386322495—ref—NC˙017571.1— Shewanella baltica BA175 chromosome
gi—187730020—ref—NC˙010658.1— Shigella boydii CDC 3083-94 chromosome
gi—377520096—ref—NC˙016822.1— Shigella sonnei 53G
gi—384541581—ref—NC˙017328.1— Shigella flexneri 2002017 chromosome
gi—560154719—ref—NC˙022912.1— Shigella dysenteriae 1617
gi—224475494—ref—NC˙012121.1— Staphylococcus carnosus subsp. carnosus TM300 chromosome
gi—379794527—ref—NC˙016941.1— Staphylococcus aureus subsp. aureus MSHR1132
gi—387141638—ref—NC˙017331.1— Staphylococcus aureus subsp. aureus TW20
gi—387149188—ref—NC˙017340.1— Staphylococcus aureus 04-02981 chromosome
gi—385782956—ref—NC˙017353.1— Staphylococcus lugdunensis N920143
gi—470192280—ref—NC˙020532.1— Staphylococcus aureus subsp. aureus ST228 complete genome
gi—537459744—ref—NC˙022226.1— Staphylococcus aureus subsp. aureus CN1
gi—194396645—ref—NC˙011072.1— Streptococcus pneumoniae G54 chromosome
gi—209558587—ref—NC˙011375.1— Streptococcus pyogenes NZ131 chromosome
gi—221230948—ref—NC˙011900.1— Streptococcus pneumoniae ATCC 700669
gi—222152201—ref—NC˙012004.1— Streptococcus uberis 0140J chromosome
gi—225860012—ref—NC˙012469.1— Streptococcus pneumoniae Taiwan19F-14 chromosome
gi—253752822—ref—NC˙012925.1— Streptococcus suis P1/7
gi—290579526—ref—NC˙013928.1— Streptococcus mutans NN2025
gi—383479207—ref—NC˙017040.1— Streptococcus pyogenes MGAS15252 chromosome
gi—386343608—ref—NC˙017581.1— Streptococcus thermophilus JIM 8232
gi—386587281—ref—NC˙017622.1— Streptococcus suis A7 chromosome
gi—410593712—ref—NC˙019048.1— Streptococcus agalactiae SA20-06 chromosome

Table D.2 Assembled unknown sequence from NCBI. Continued

Unknown Assembled Sequences
i—479134147—ref—NC˙021003.1— Streptococcus pneumoniae SPN032672 draft genome
i—512538239—ref—NC˙021485.1— Streptococcus agalactiae 09mas018883 complete genome
gi—538369494—ref—NC˙022244.1— Streptococcus anginosus C1051
gi—556587607—ref—NC˙022665.1— Streptococcus suis T15
gi—170076636—ref—NC˙010475.1— Synechococcus sp. PCC 7002 chromosome
gi—427711179—ref—NC˙019680.1— Synechococcus sp. PCC 6312 chromosome
gi—383489963—ref—NC˙017052.1— Synechocystis sp. PCC 6803 substr. PCC-N
gi—451813329—ref—NC˙020286.1— Synechocystis sp. PCC 6803
gi—307723218—ref—NC˙014538.1— Thermoanaerobacter sp. X513 chromosome
gi—320114857—ref—NC˙014964.1— Thermoanaerobacter brockii subsp. finnii Ako-1 chromosome
gi—197333880—ref—NC˙011184.1— Vibrio fischeri MJ11 chromosome I
gi—197336667—ref—NC˙011186.1— Vibrio fischeri MJ11 chromosome II
gi—294510242—ref—NC˙011744.2— Vibrio splendidus LGP32 chromosome 2
gi—294514841—ref—NC˙011753.2— Vibrio splendidus LGP32 chromosome 1
gi—320154846—ref—NC˙014965.1— Vibrio vulnificus MO6-24/O chromosome I
gi—320157827—ref—NC˙014966.1— Vibrio vulnificus MO6-24/O chromosome II
gi—360034408—ref—NC˙016445.1— Vibrio cholerae O1 str. 2010EL-1786 chromosome 1
gi—360037214—ref—NC˙016446.1— Vibrio cholerae O1 str. 2010EL-1786 chromosome 2
gi—375129161—ref—NC˙016602.1— Vibrio furnissii NCTC 11218 chromosome 1
gi—375132168—ref—NC˙016628.1— Vibrio furnissii NCTC 11218 chromosome 2
gi—433656322—ref—NC˙019955.1— Vibrio parahaemolyticus BB22OP chromosome 1
gi—433659170—ref—NC˙019971.1— Vibrio parahaemolyticus BB22OP chromosome 2
gi—379009272—ref—NC˙016893.1— Wigglesworthia glossinidia endosymbiont of Glossina morsitans morsitans (Yale colo
gi—188574270—ref—NC˙010717.1— Xanthomonas oryzae pv. oryzae PXO99A chromosome
gi—285016821—ref—NC˙013722.1— Xanthomonas albilineans GPE PC73 chromosome
gi—346722940—ref—NC˙016010.1— Xanthomonas axonopodis pv. citrumelo F1 chromosome
gi—384425691—ref—NC˙017271.1— Xanthomonas campestris pv. raphani 756C chromosome
gi—471265562—ref—NC˙020815.1— Xanthomonas citri subsp. citri Aw12879
gi—186893344—ref—NC˙010634.1— Yersinia pseudotuberculosis PB1/+ chromosome
gi—294502110—ref—NC˙014029.1— Yersinia pestis Z176003 chromosome
gi—384120592—ref—NC˙017154.1— Yersinia pestis D106004 chromosome
gi—384412706—ref—NC˙017265.1— Yersinia pestis biovar Medievalis str. Harbin 35 chromosome

Table D.3: "Unknown" short read files

Accession	Organism
ERR193649	Acinetobacter baumannii ATCC 17978
ERR776855	Acinetobacter sp. ADP1
ERR788913	Acinetobacter sp. ADP1
SRR006330	Acinetobacter sp. ADP1
SRR006332	Acinetobacter sp. ADP1
SRR006465	Acinetobacter sp. ADP1
ERR200084	Actinobacillus pleuropneumoniae
ERR271099	Actinobacillus pleuropneumoniae
ERR271132	Actinobacillus pleuropneumoniae
SRR191908	Actinobacillus succinogenes 130Z
DRR015722	Aeromonas hydrophila subsp. hydrophila
DRR015723	Aeromonas hydrophila subsp. hydrophila
SRR253101	Alkaliphilus oremlandii OhILAs
SRR000278	Anaeromyxobacter dehalogenans 2CP-1
SRR000279	Anaeromyxobacter dehalogenans 2CP-1
SRR422133	Anaplasma marginale str. St. Maries
SRR422132	Anaplasma phagocytophilum str. HZ
SRR1776687	Bacillus anthracis Ames BBF
ERR760543	Bacillus cereus ATCC 10987
DRR000002	Bacillus subtilis subsp. subtilis str. 168
DRR000852	Bacillus subtilis subsp. subtilis str. 168
DRR008448	Bacillus subtilis subsp. subtilis str. 168
ERR055715	Bacillus thuringiensis
SRR253092	Bacillus weihenstephanensis KBAB4
SRR1173438	Bartonella bacilliformis str. Heidi Mejia
SRR1173496	Bartonella bacilliformis str. Heidi Mejia
SRR445757	Bartonella henselae str. Zeus
SRR445766	Bartonella quintana JK 19
SRR2088903	Bifidobacterium longum
ERR225614	Bordetella bronchiseptica
ERR380650	Bordetella parapertussis
ERR370327	Bordetella pertussis
SRR1772332	Borrelia burgdorferi B31
ERR418017	Brucella canis
ERR485956	Brucella canis
ERR554818	Brucella canis
SRR642809	Brucella canis
SRR960778	Brucella canis 96-7258
SRR011104	Brucella melitensis bv. 1 str. 16M
SRR2146152	Brucella melitensis bv. 1 str. 16M
SRR2146907	Burkholderia ambifaria AMMD
SRR2146908	Burkholderia ambifaria AMMD
ERR406386	Burkholderia pseudomallei K96243
SRR1614021	Burkholderia pseudomallei K96243
SRR1614022	Burkholderia pseudomallei K96243
SRR1614023	Burkholderia pseudomallei K96243
SRR1146477	Burkholderia thailandensis E264
SRR1146481	Burkholderia thailandensis E264
SRR1146490	Burkholderia thailandensis E264
SRR1146496	Burkholderia thailandensis E264
SRR1146508	Burkholderia thailandensis E264
SRR1146510	Burkholderia thailandensis E264
SRR253095	Caldicellulosiruptor saccharolyticus DSM 8903
SRR942779	Campylobacter jejuni subsp. jejuni 81-176-55
SRR942780	Campylobacter jejuni subsp. jejuni 81-176-55-O1

Table D.3 "Unknown" short read files. Continued

Accession	Organism
SRR942781	Campylobacter jejuni subsp. jejuni 81-176-55-O3
SRR942782	Campylobacter jejuni subsp. jejuni 81-176-55-T1
SRR942783	Campylobacter jejuni subsp. jejuni 81-176-55-T3
SRR064701	Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819
SRR437910	Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819
SRR437911	Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819
SRR437914	Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819
SRR437931	Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819
SRR497623	Campylobacter jejuni subsp. jejuni NCTC 11168-PO
SRR824843	Caulobacter crescentus CB15
SRR824846	Caulobacter crescentus CB15
SRR824849	Caulobacter crescentus CB15
SRR824857	Caulobacter crescentus CB15
SRR824865	Caulobacter crescentus CB15
SRR834582	Caulobacter crescentus CB15
ERR386226	Chlamydia muridarum str. Nigg
SRR1736658	Chlamydia muridarum str. Nigg CM972
SRR125859	Chlamydia trachomatis D/UW-3/CX
SRR125860	Chlamydia trachomatis D/UW-3/CX
ERR386222	Chlamydia trachomatis L2b/UCH-1/proctitis
ERR386223	Chlamydia trachomatis L2b/UCH-1/proctitis
ERR768074	Chromobacterium violaceum
SRR769602	Chromobacterium violaceum
SRR1981078	Chromohalobacter salexigens DSM-3043
SRR1981208	Chromohalobacter salexigens DSM-3043
ERR163867	Citrobacter koseri
ERR772459	Citrobacter koseri
SRR755330	Citrobacter koseri ADL-328
SRR1609114	Citrobacter koseri
SRR1656096	Citrobacter koseri
SRR1656442	Citrobacter koseri
DRR018799	Clostridium botulinum
ERR022312	Clostridium botulinum A5(B') str. H04402 065
SRR036930	Clostridium botulinum 5311a
SRR764973	Clostridium botulinum str. LANGELAN F
SRR770049	Clostridium botulinum CFSAN002367
ERR171255	Clostridium perfringens
SRR065411	Clostridium perfringens WAL-14572
SRR096826	Clostridium perfringens F262
SRR1655401	Clostridium perfringens
SRR2089300	Clostridium perfringens
ERR599177	Clostridium phytofermentans
SRR253100	Clostridium phytofermentans ISDg
SRR400549	Clostridium thermocellum ATCC 27405
SRR891397	Clostridium thermocellum ATCC 27405
SRR891400	Clostridium thermocellum ATCC 27405
SRR891797	Clostridium thermocellum ATCC 27405
SRR089499	Corynebacterium efficiens YS-314
ERR845240	Coxiella burnetii RSA 493
SRR253103	Delftia acidovorans SPH-1
SRR253097	Desulfococcus oleovorans Hxd3
SRR402910	Desulfovibrio vulgaris str. Hildenborough
SRR1291430	Desulfovibrio vulgaris str. Hildenborough
SRR1291434	Desulfovibrio vulgaris str. Hildenborough
ERR506948	Dichelobacter nodosus

Table D.3 "Unknown" short read files. Continued

Accession	Organism
ERR506983	Dichelobacter nodosus
ERR506989	Dichelobacter nodosus
ERR314470	Enterobacter sp.
ERR387211	Enterobacter sp.
ERR502553	Enterobacter sp.
SRR2127604	Enterobacter sp. BIDMC92
SRR172995	Enterococcus faecalis V583
SRR182361	Enterococcus faecalis V583
SRR638571	Enterococcus faecalis V583
SRR248516	Erythrobacter litoralis HTCC2594
ERR351257	Escherichia coli 536
ERR305884	Escherichia coli APEC O1
ERR305901	Escherichia coli APEC O1
SRR1021212	Escherichia coli O157:H7 str. EDL933
SRR1509640	Escherichia coli O157:H7 str. EDL933
SRR1509643	Escherichia coli O157:H7 str. EDL933
SRR1509803	Escherichia coli O157:H7 str. EDL933
SRR1783841	Escherichia coli O157:H7 str. EDL933
SRR1795985	Escherichia coli O157:H7 str. EDL933
ERR687900	Escherichia coli UTI89
ERR687901	Escherichia coli UTI89
SRR000868	Escherichia coli UTI89
SRR000871	Escherichia coli UTI89
ERR376619	Escherichia coli str. K-12 substr. MG1655
ERR376625	Escherichia coli str. K-12 substr. MG1655
SRR1635255	Escherichia coli str. K-12 substr. MG1655
SRR253105	Fervidobacterium nodosum Rt17-B1
SRR000311	Francisella tularensis subsp. holarctica OSU18
SRR292171	Francisella tularensis subsp. holarctica OSU18
SRR999318	Francisella tularensis subsp. tularensis str. SCHU S4 substr. FSC237
SRR999323	Francisella tularensis subsp. tularensis str. SCHU S4 substr. SL
SRR1061349	Francisella tularensis subsp. tularensis str. SCHU S4 substr. SL
SRR1714340	Francisella tularensis subsp. tularensis str. SCHU S4 substr. NR-28534
SRR1284499	Francisella tularensis subsp. tularensis WY96-3418
SRR1284500	Francisella tularensis subsp. tularensis WY96-3418
SRR1284501	Francisella tularensis subsp. tularensis WY96-3418
SRR896553	Frankia sp. CcI3
ERR125051	Haemophilus influenzae
ERR125090	Haemophilus influenzae
ERR658012	Haemophilus influenzae Rd KW20
ERR716321	Haemophilus influenzae
SRR065202	Haemophilus influenzae Rd KW20
SRR065206	Haemophilus influenzae 86-028NP
SRR253108	Halorhodospira halophila SL1
SRR1980752	Helicobacter pylori J99
SRR1980757	Helicobacter pylori J99
SRR1981186	Helicobacter pylori J99
SRR1981237	Helicobacter pylori J99
SRR1981622	Helicobacter pylori J99
DRR003232	Klebsiella pneumoniae subsp. pneumoniae JCM 1662
ERR706867	Klebsiella pneumoniae subsp. pneumoniae MGH 78578
ERR706873	Klebsiella pneumoniae subsp. pneumoniae MGH 78578
ERR718767	Klebsiella pneumoniae subsp. pneumoniae
SRR515628	Klebsiella pneumoniae subsp. pneumoniae KPNIH15
SRR770033	Klebsiella pneumoniae subsp. pneumoniae WGLW2

Table D.3 "Unknown" short read files. Continued

Accession	Organism
SRR1166990	Klebsiella pneumoniae subsp. pneumoniae
SRR1510962	Klebsiella pneumoniae subsp. pneumoniae KPR0928
SRR1510963	Klebsiella pneumoniae subsp. pneumoniae KPR0928
SRR1510964	Klebsiella pneumoniae subsp. pneumoniae KPR0928
ERR570198	Lactobacillus casei
ERR256994	Lactobacillus delbrueckii subsp. bulgaricus
ERR387526	Lactobacillus delbrueckii subsp. bulgaricus
ERR433470	Lactobacillus delbrueckii subsp. indicus
ERR204049	Lactobacillus gasseri
ERR570066	Lactobacillus gasseri MV-22
ERR570280	Lactobacillus gasseri
ERR204044	Lactobacillus helveticus
ERR387534	Lactobacillus helveticus
SRR077393	Lactobacillus helveticus DSM 20075 = CGMCC 1.1877
ERR570285	Lactobacillus plantarum
SRR010987	Lactobacillus plantarum subsp. plantarum ATCC 14917 = JCM 1149 = CGMCC 1.2437
SRR010988	Lactobacillus plantarum subsp. plantarum ATCC 14917 = JCM 1149 = CGMCC 1.2437
SRR1552613	Lactobacillus plantarum ATCC 801
ERR022445	Lactobacillus reuteri ATCC 53608
ERR256993	Lactobacillus reuteri
SRR011135	Lactobacillus reuteri CF48-3A
SRR1151170	Lactobacillus reuteri DSM 20016
DRR003258	Lactobacillus sakei subsp. sakei DSM 20017 = JCM 1157
ERR387465	Lactobacillus sakei subsp. sakei
SRR1151267	Lactobacillus sakei subsp. sakei DSM 20017 = JCM 1157
ERR570279	Lactobacillus salivarius
SRR010995	Lactobacillus salivarius DSM 20555 = ATCC 11741
SRR010996	Lactobacillus salivarius DSM 20555 = ATCC 11741
SRR1151172	Lactobacillus salivarius DSM 20555 = ATCC 11741
SRR1656220	Lactobacillus salivarius
DRR003259	Lactococcus lactis subsp. lactis JCM 5805 = NBRC 100933
ERR387536	Lactococcus lactis subsp. lactis
ERR440991	Lactococcus lactis subsp. lactis
SRR088758	Lactococcus lactis subsp. lactis KF147
ERR340950	Legionella pneumophila
ERR351253	Legionella pneumophila str. Paris
ERR351261	Legionella pneumophila str. Corby
ERR351262	Legionella pneumophila str. Lens
ERR485161	Legionella pneumophila
SRR801743	Legionella pneumophila subsp. pneumophila str. Philadelphia 1
SRR801793	Legionella pneumophila subsp. pneumophila str. Philadelphia 1
SRR801840	Legionella pneumophila subsp. pneumophila str. Philadelphia 1
SRR714504	Leptospira interrogans serovar Copenhageni str. Fiocruz LV4174
SRR717627	Leptospira interrogans serovar Copenhageni str. Fiocruz LV2750
SRR717876	Leptospira interrogans serovar Copenhageni str. Fiocruz LV4113
ERR760535	Listeria monocytogenes EGD-e
SRR1031054	Listeria monocytogenes serotype 4b str. NCTC 11994
SRR1031055	Listeria monocytogenes serotype 4b str. NCTC 11994
ERR760535	Listeria monocytogenes EGD-e
SRR2352237	Listeria monocytogenes
SRR000215	Methylobacterium extorquens PA1
SRR000217	Methylobacterium extorquens PA1
SRR190860	Methylobacterium extorquens DSM 13060
SRR1046370	Methylobacterium extorquens AM1
ERR037949	Mycobacterium avium subsp. paratuberculosis

Table D.3 "Unknown" short read files. Continued

Accession	Organism
SRR060191	Mycobacterium avium subsp. paratuberculosis K-10
SRR1793723	Mycobacterium avium subsp. paratuberculosis
SRR799346	Mycobacterium leprae 3125609
ERR274521	Mycobacterium smegmatis
ERR550505	Mycobacterium smegmatis
SRR071425	Mycobacterium smegmatis str. MC2 155
SRR453241	Mycobacterium smegmatis JS623
ERR760549	Mycobacterium tuberculosis H37Ra
SRR024229	Mycobacterium tuberculosis F11
SRR974839	Mycobacterium tuberculosis F11
SRR974842	Mycobacterium tuberculosis H37Rv
SRR1949885	Mycobacterium tuberculosis H37Rv
ERR339034	Mycoplasma agalactiae
SRR006331	Mycoplasma agalactiae PG2
ERR486835	Mycoplasma genitalium G37
ERR486841	Mycoplasma genitalium
ERR713979	Mycoplasma hyopneumoniae 232
ERR736802	Mycoplasma hyopneumoniae 232
SRR631043	Mycoplasma pneumoniae M129
SRR643250	Mycoplasma pneumoniae M129
SRR2135833	Mycoplasma pneumoniae
SRR2135852	Mycoplasma pneumoniae
ERR191802	Neisseria gonorrhoeae
ERR355927	Neisseria gonorrhoeae
SRR004146	Neisseria gonorrhoeae SK-92-679
SRR016778	Neisseria gonorrhoeae F62
ERR051677	Neisseria meningitidis
ERR484778	Neisseria meningitidis
ERR636419	Neisseria meningitidis MC58
SRR057353	Neisseria meningitidis K1207
SRR1425912	Nitrosococcus oceani C-27
SRR1020892	Nostoc sp. PCC 7120
ERR841688	Ochrobactrum anthropi
SRR253117	Parvibaculum lavamentivorans DS-1
SRR253118	Petrotoga mobilis SJ95
SRR001351	Porphyromonas gingivalis W83
SRR001352	Porphyromonas gingivalis W83
SRR413299	Porphyromonas gingivalis W50
SRR248518	Prochlorococcus marinus str. MIT 9211
SRR253119	Prochlorococcus marinus str. NATL2A
SRR1805320	Prochlorococcus marinus
ERR246369	Pseudomonas aeruginosa
SRR396638	Pseudomonas aeruginosa MPAO1/P1
SRR1103537	Pseudomonas aeruginosa PAO1-GFP
SRR1374997	Pseudomonas aeruginosa PAO1
SRR2099465	Pseudomonas aeruginosa PAO1
DRR001171	Pseudomonas fluorescens Pf0-1
DRR001172	Pseudomonas fluorescens Pf0-1
SRR567996	Pseudomonas fluorescens
SRR949275	Pseudomonas fluorescens BBc6R8
DRR017738	Pseudomonas putida JCM 9802
SRR253120	Pseudomonas putida F1
SRR924720	Pseudomonas putida LF54
ERR005143	Pseudomonas syringae pv. syringae B728a
SRR020199	Pseudomonas syringae pv. syringae FF5

Table D.3 "Unknown" short read files. Continued

Accession	Organism
SRR1039777	Pseudomonas syringae pv. tomato
SRR1039794	Pseudomonas syringae pv. tomato
ERR726246	Rhizobium leguminosarum
SRR004795	Rhizobium leguminosarum bv. trifolii WSM2304
ERR760546	Rhodobacter sphaeroides 2.4.1
SRR387291	Rhodobacter sphaeroides 2.4.1
SRR522245	Rhodobacter sphaeroides 2.4.1
SRR620446	Rhodobacter sphaeroides 2.4.1
ERR039479	Rhodopseudomonas palustris BisB5
SRR031640	Rhodopseudomonas palustris DX-1
SRR1791643	Rhodopseudomonas palustris
SRR1791673	Rhodopseudomonas palustris
SRR949058	Saccharopolyspora erythraea D
ERR212582	Salmonella enterica subsp. enterica serovar Choleraesuis
SRR955200	Salmonella enterica subsp. enterica serovar Choleraesuis str. ATCC 10708
SRR1586583	Salmonella enterica subsp. enterica serovar Choleraesuis Var. Kunzendorf str
ERR235289	Salmonella enterica subsp. enterica serovar Typhimurium
ERR744244	Salmonella enterica subsp. enterica serovar Typhimurium
SRR1176802	Salmonella enterica subsp. enterica serovar Typhimurium str. LT2-4
SRR072372	Shewanella amazonensis SB2B
SRR1798623	Shewanella amazonensis SB2B
SRR058889	Shewanella baltica OS678
SRR253122	Shewanella baltica OS195
DRR003249	Shewanella putrefaciens JCM 20190 = NBRC 3908
SRR253124	Shewanella putrefaciens CN-32
SRR1801183	Shewanella putrefaciens HRCR-6
ERR017663	Shigella sonnei
ERR017664	Shigella sonnei
ERR212316	Shigella dysenteriae
ERR852667	Shigella dysenteriae
SRR364091	Shigella dysenteriae 225-75
SRR765082	Shigella dysenteriae
ERR126957	Shigella flexneri
ERR590906	Shigella flexneri 2a
ERR591308	Shigella flexneri 5
SRR041664	Shigella flexneri 2a str. 2457T
SRR1165139	Sinorhizobium meliloti 1021
ERR142616	Staphylococcus aureus subsp. aureus TW20
ERR580966	Staphylococcus aureus subsp. aureus COL
SRR292151	Staphylococcus aureus subsp. aureus USA300˙TCH1516
SRR528763	Staphylococcus aureus strain Newman
SRR578343	Staphylococcus aureus subsp. aureus USA300˙FPR3757
SRR1955495	Staphylococcus aureus
SRR1955594	Staphylococcus aureus
SRR1955595	Staphylococcus aureus
SRR1955861	Staphylococcus aureus
DRR017649	Staphylococcus epidermidis JCM 2414
ERR234787	Staphylococcus epidermidis
ERR387262	Staphylococcus epidermidis
SRR014815	Staphylococcus epidermidis W23144
SRR071338	Staphylococcus epidermidis VCU120
SRR1609104	Staphylococcus epidermidis
SRR1656424	Staphylococcus epidermidis
ERR085220	Staphylococcus haemolyticus
SRR1656451	Staphylococcus haemolyticus

Table D.3 "Unknown" short read files. Continued

Accession	Organism
ERR204108	Streptococcus pneumoniae
ERR204135	Streptococcus pneumoniae
ERR654515	Streptococcus pneumoniae
ERR716254	Streptococcus pneumoniae
SRR068305	Streptococcus pneumoniae GA17457
SRR1408840	Streptococcus pneumoniae 13856
ERR046238	Streptococcus pyogenes
ERR144738	Streptococcus pyogenes
ERR662589	Streptococcus pyogenes
SRR004693	Streptococcus pyogenes AA216
SRR090459	Streptococcus pyogenes ATCC 10782
SRR1147086	Streptococcus pyogenes
SRR1655199	Streptococcus thermophilus
SRR1770414	Streptomyces avermitilis MA-4680 = NBRC 14893
ERR588636	Synechococcus sp. WH 8103
SRR038529	Synechococcus sp. CB0205
SRR1798191	Synechococcus sp. PEB5 55AY5-B PE B5
DRR001143	Synechocystis sp. PCC 6803 PCC-N strain
SRR253126	Thermoanaerobacter sp. X514
SRR516571	Thermoanaerobacter sp. X514
SRR253128	Thermosipho melanesiensis BI429
SRR896531	Thermotoga maritima MSB8
SRR000332	Treponema pallidum subsp. pallidum str. Nichols
SRR029224	Treponema pallidum subsp. pallidum str. Nichols
SRR364930	Vibrio cholerae O1 biovar El Tor str. N16961
SRR1124794	Vibrio cholerae O1 biovar El Tor
SRR1199311	Vibrio cholerae O1 biovar El Tor
DRR017980	Vibrio parahaemolyticus RIMD 2210633
DRR017981	Vibrio parahaemolyticus RIMD 2210633
DRR017982	Vibrio parahaemolyticus RIMD 2210633
DRR017983	Vibrio parahaemolyticus RIMD 2210633
ERR175748	Wolbachia endosymbiont of Drosophila simulans
SRR2064179	Wolbachia endosymbiont of Brugia malayi
SRR1998069	Wolinella succinogenes
SRR1207369	Xanthomonas oryzae pv. oryzicola
SRR1592663	Xanthomonas oryzae ATCC 35933
ERR015575	Yersinia enterocolitica
SRR2149856	Yersinia enterocolitica
SRR2149857	Yersinia enterocolitica
SRR2149858	Yersinia enterocolitica
ERR245863	Yersinia pseudotuberculosis
ERR752453	Yersinia pseudotuberculosis
SRR2148416	Yersinia pseudotuberculosis
SRR2148805	Yersinia pseudotuberculosis
SRR017901	Zymomonas mobilis subsp. mobilis ZM4 = ATCC 31821
SRR191898	Zymomonas mobilis subsp. mobilis ATCC 29191

Table D.4: PacBio, ABSolid and Oxford Nanopore Short Reads Data.

Accession	Accession	Accession	Accession	Accession
ERR845240	DRR021341	ERR234280	SRR001354	SRR1635255
SRR328412	ERR557006	ERR557023	SRR656856	SRR656873
SRR546564	ERR202400	SRR1950255	SRR1950259	SRR2034255
ERR421272	ERR421274	ERR421275	SRR538068	SRR538825
SRR542334	SRR542433	SRR653010	SRR653053	SRR653609
SRR653610	SRR3169819	ERR029927	SRR350980	SRR350983
SRR578343	SRR1014692	SRR1014693	SRR1014695	SRR1014696
SRR1014697	SRR353676	SRR353678	SRR364894	SRR908310
SRR908311	SRR909270	SRR909545	SRR909586	SRR909667
SRR035442	SRR035443	SRR035444	SRR035445	ERR701171
ERR776851	ERR776853	ERR776854	ERR776855	SRR2671867
SRR2671868	ERR977574	ERR637419	ERR701174	ERR764952
ERR968962	ERR968963	ERR968968	ERR968971	ERR968974
ERR1309549	SRR1596423	ERR701176	SRR3473969	SRR1177097
SRR1177844	SRR1178887	SRR3893881	ERR581145	SRR3951708
ERR1046620	SRR2146908	SRR1206479	SRR2863235	SRR2148792
SRR1614023	ERR849508	ERR1354173	ERR1366099	ERR768074
SRR1981078	ERR772459	SRR1609114	SRR4095613	SRR1003149
SRR1003209	SRR1004230	SRR1004237	SRR1509640	SRR1509803
SRR1795985	SRR1284501	SRR1714340	ERR526296	ERR526297
SRR386136	SRR1980752	SRR1980757	ERR768078	SRR1980767
ERR706867	SRR1510964	SRR1556927	SRR3756808	SRR1950323
ERR550505	SRR1757045	SRR1776953	SRR2048552	SRR631043
SRR643250	SRR631043	SRR643250	ERR1466806	SRR1223220
SRR2063176	SRR3723122	ERR1109366	ERR841688	SRR387291
SRR386671	SRR386717	SRR2063211	SRR1177090	SRR1178078
SRR1178840	SRR1425899	SRR1425900	SRR1425901	ERR852667
ERR159975	SRR1179003	SRR1955594	SRR1609103	SRR1609104
ERR1562473	ERR654515	ERR1517154	DRR015080	SRR364930
DRR017980	SRR1207369	SRR2149858	ERR752453	

Table D.5: MRSA outbreak strains.

Accession	STtype	Patients
ERR070033	ST2371	P8
ERR070034	ST2371	P9
ERR070035	ST8	NonOutBreak
ERR070036	ST2371	P10
ERR070037	ST22	NonOutBreak
ERR070038	ST2371	P13
ERR070039	ST2371	P11
ERR070040	ST2371	P12
ERR070041	ST1	NonOutBreak
ERR070042	ST2371	P2
ERR070043	ST2371	P3
ERR070044	ST2371	P4
ERR070045	ST2371	P1
ERR070046	ST2371	P5
ERR070047	ST2371	P6
ERR070048	ST2371	P7
ERR072246	ST2371	P14
ERR072247	ST2371	P23
ERR072248	ST772	NonOutBreak
ERR108054	ST2371	P15
ERR124429	ST2371	P16
ERR124430	ST2371	P17
ERR124431	ST2371	P20
ERR124432	ST2371	P19
ERR124433	ST2371	P21
ERR124434	ST2371	P22
ERR124435	ST2371	P24
ERR124436	ST772	NonOutBreak
ERR128707	ST2371	P26
ERR128708	ST2371	P25
ERR128709	ST2371	Staff
ERR128710	ST2371	Staff
ERR128711	ST2371	Staff
ERR128712	ST2371	Staff
ERR128713	ST2371	Staff
ERR128714	ST2371	Staff
ERR128715	ST2371	Staff
ERR128716	ST2371	Staff
ERR128717	ST2371	Staff
ERR128718	ST2371	Staff
ERR128719	ST2371	Staff
ERR128720	ST2371	Staff
ERR131800	ST772	NonOutBreak
ERR131801	ST22	NonOutBreak
ERR131802	ST22	NonOutBreak
ERR131804	ST772	NonOutBreak
ERR131805	ST22	NonOutBreak
ERR131806	ST22	NonOutBreak
ERR131807	ST22	NonOutBreak
ERR131808	ST2371	Staff
ERR131809	ST2371	Staff
ERR131810	ST2371	Staff
ERR131811	ST2371	Staff
ERR131812	ST2371	Staff
ERR131813	ST2371	Staff
ERR131814	ST2371	Staff
ERR131815	ST2371	Staff

D.2 Thresholds Constructing and Typing Accuracies

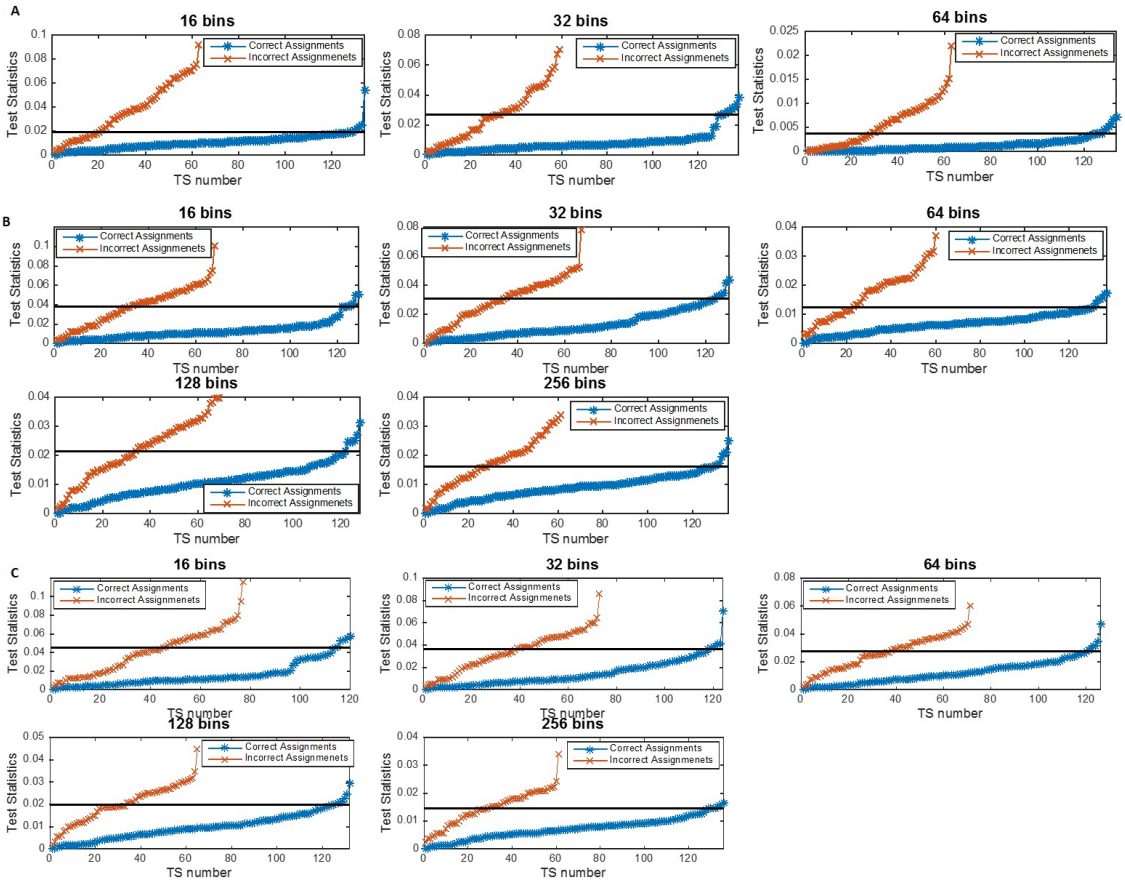

Figure D.1: Sorted test statistics from assembled bacterial sequences. (A) 3-mer. (B) 6-mer. (C) 9-mer. The orange curve is the false assignments and the blue curve is the correct assignments. The black line is the threshold determined from the 95% test statistics of the correct assignments.

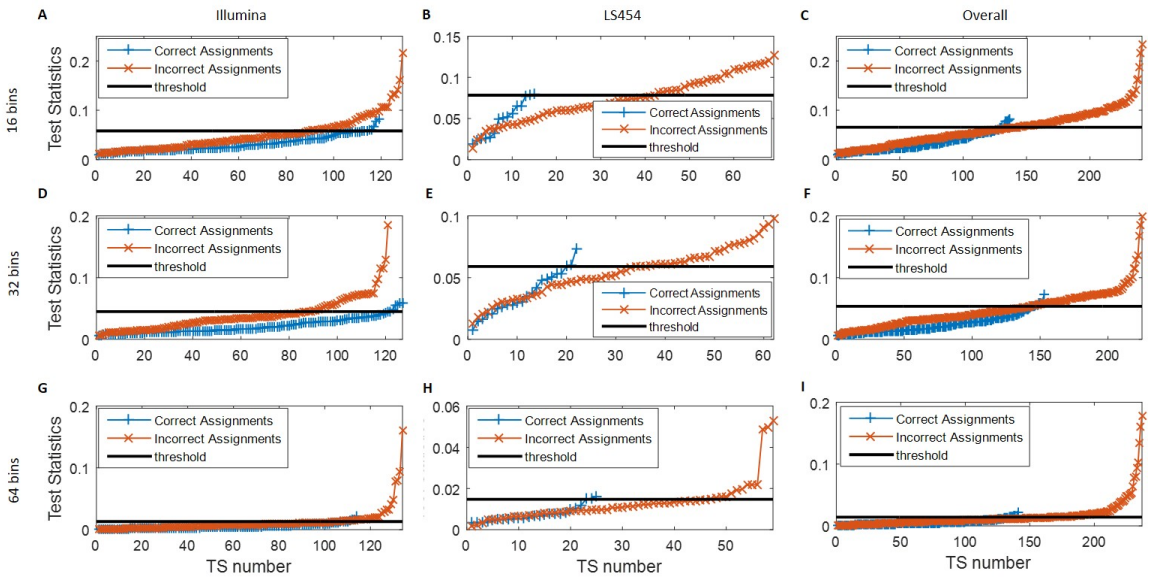

Figure D.2: Sorted test statistics (3-mer) from pooled short reads data (raw reads data files). (A) 16 bins, (D) 32 bins, and (G) 64 bins for Illumina. (B) 16 bins, (E) 32 bins, and (H) 64 bins for LS454. (C) 16 bins, (F) 32 bins, and (I) 64 bins for all data. The orange curve is the false assignments and the blue curve is the correct assignments. The black line is the threshold determined from the 95% test statistics of the correct assignments for Illumina and Overall data. For LS454, 90% was used.

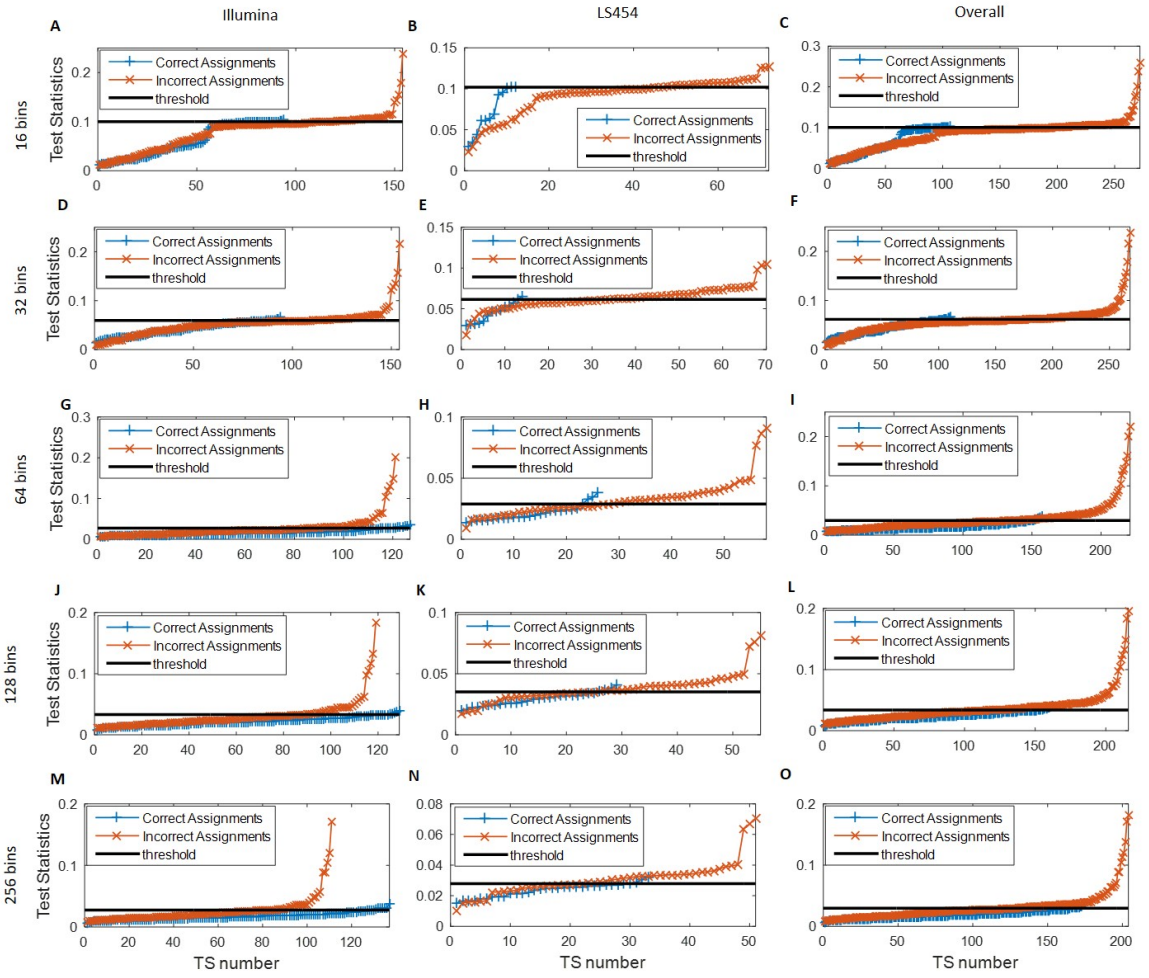

Figure D.3: Sorted test statistics (6-mer) from pooled short reads data (raw reads data files). (A) 16 bins, (D) 32 bins, (G) 64 bins, (I) 128 bins, and (M) 256 bins for Illumina. (B) 16 bins, (E) 32 bins, (H) 64 bins, (K) 128 bins, and (N) 256 bins for LS454. (C) 16 bins, (F) 32 bins, (I) 64 bins, (L) 128 bins, and (O) 256 bins for all data. The orange curve is the false assignments and the blue curve is the correct assignments. The black line is the threshold determined from the 95% test statistics of the correct assignments for Illumina and Overall data. For LS454, 90% was used.

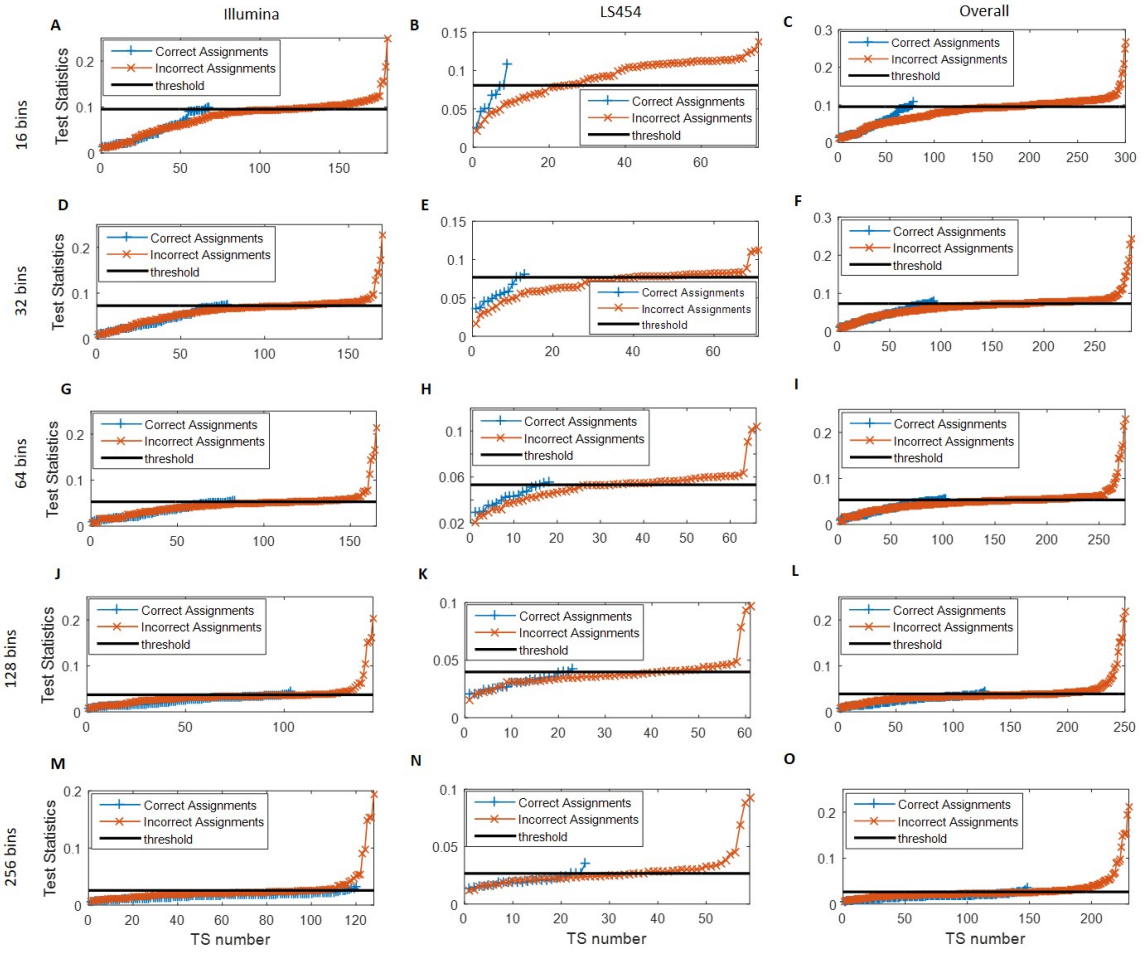

Figure D.4: Sorted test statistics (9-mer) from pooled short reads data (raw reads data files) (A) 16 bins, (D) 32 bins, (G) 64 bins, (J) 128 bins, and (M) 256 bins for Illumina. (B) 16 bins, (E) 32 bins, (H) 64 bins, (K) 128 bins, and (N) 256 bins for LS454. (C) 16 bins, (F) 32 bins, (I) 64 bins, (L) 128 bins, and (O) 256 bins for all data. The orange curve is the false assignments and the blue curve is the correct assignments. The black line is the threshold determined from the 95% test statistics of the correct assignments for Illumina and Overall data. For LS454, 90% was used.

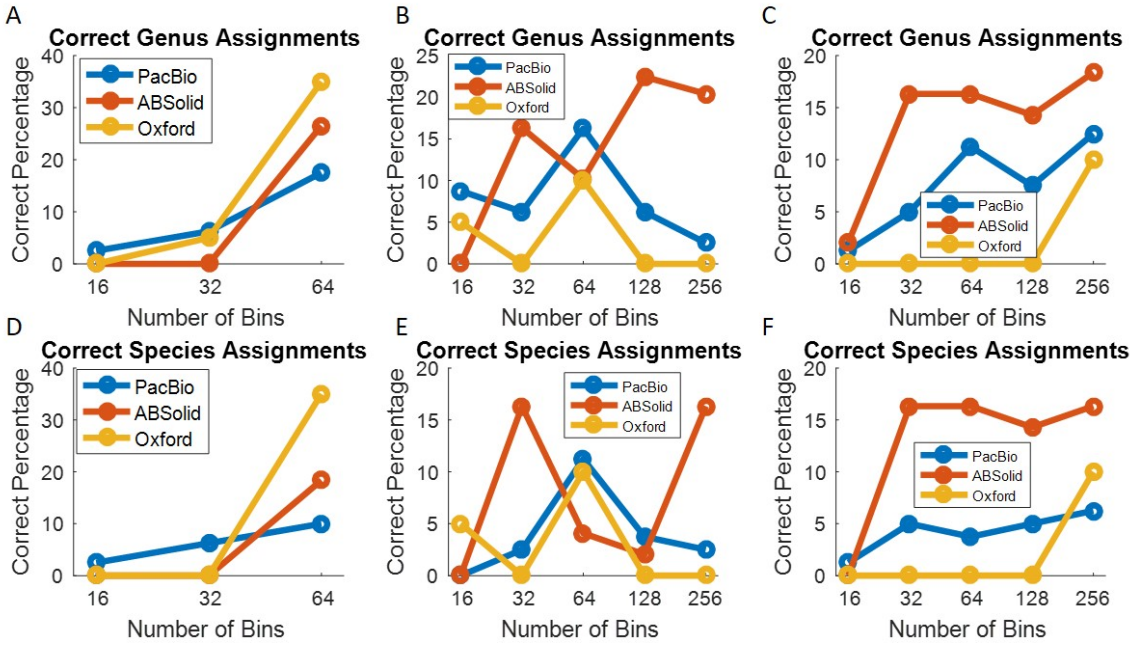

Figure D.5: Pooled Sequence Typing for PacBio, AB Solid and Oxford Nanopore (raw reads data files without threshold). Percent correct genus assignments of (A) 3-mers, (B) 6-mers, and (C) 9-mers. PB-sQF analyses of raw reads files compared to reconstructed whole genome libraries. Percent correct species assignments using (D) 3-mers, (E) 6-mers, and (F) 9-mers. Again, there are 80, 49 and 20 total number of data files for PacBio, AB Solid and Oxford Nanopore respectively.

APPENDIX E

SUPPORTING INFORMATION FOR CHAPTER 7

E.1 Aligners and Reads Lengths

In this section, the linearity at different read lengths for mrFAST, MAQ, BWA-MEM, NN, and Bowtie2 are reported.

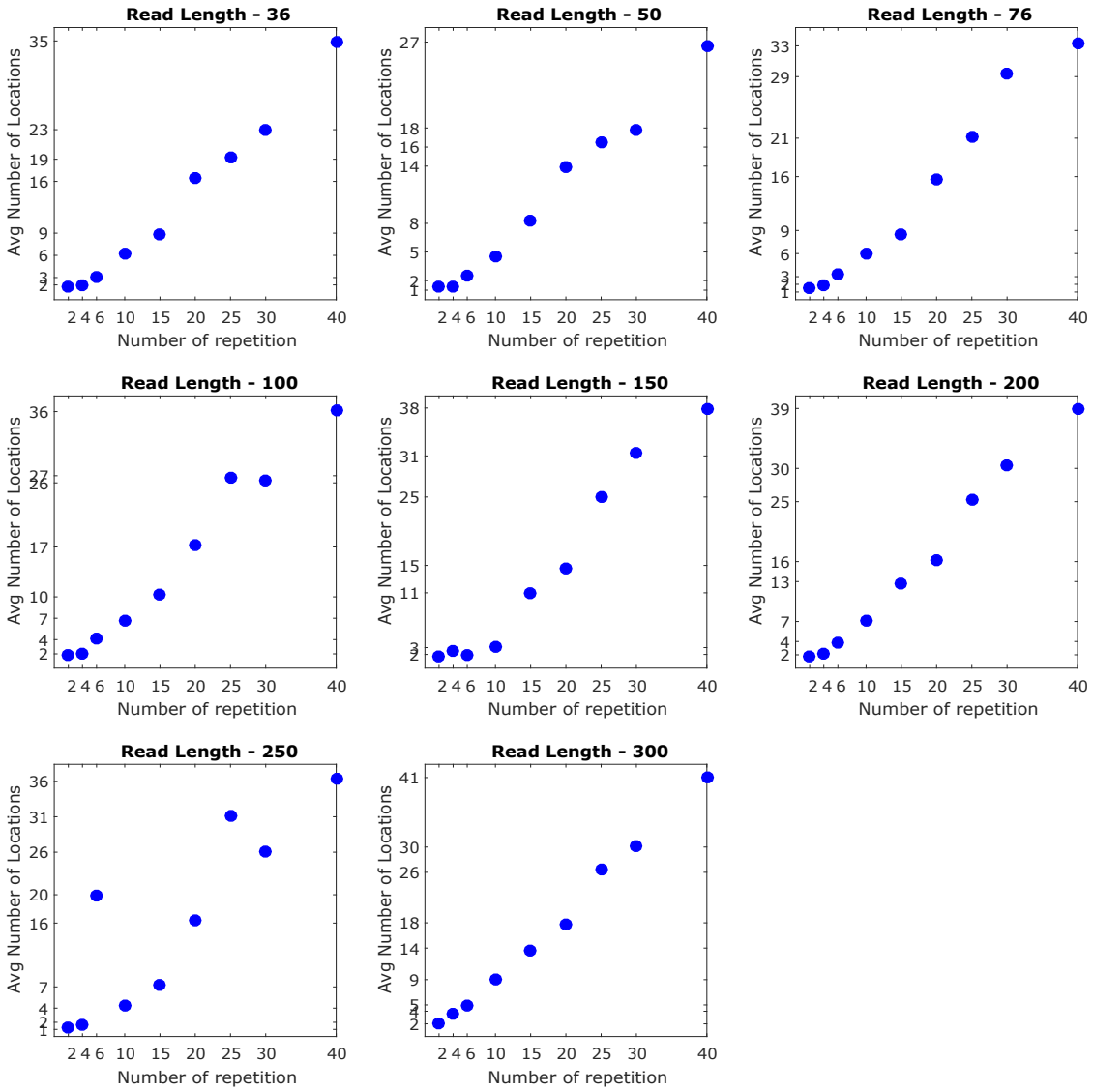

Figure E.1: Mapping Linearity using mrFAST with different read length.

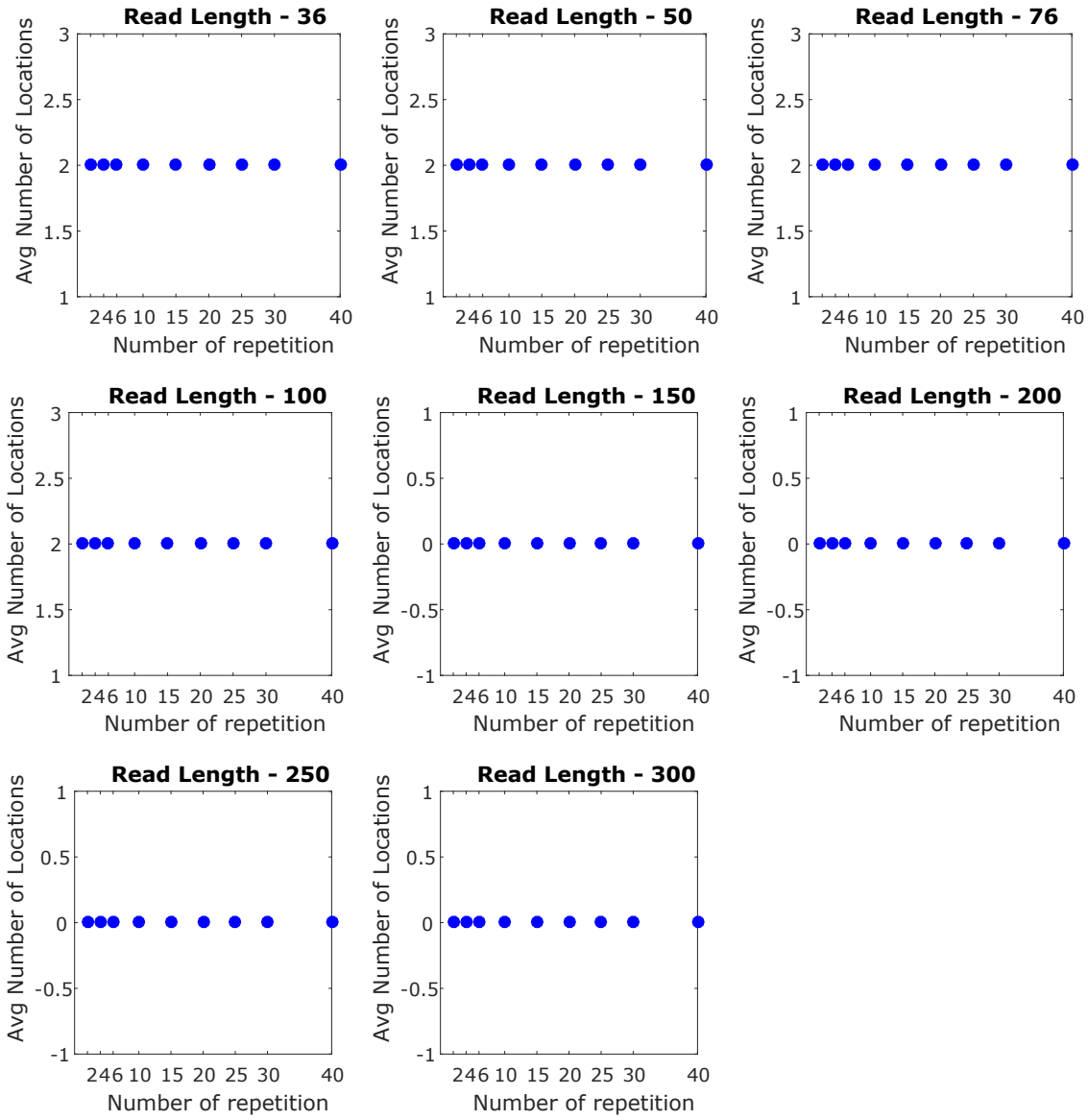

Figure E.2: Mapping Linearity using MAQ with different read length.

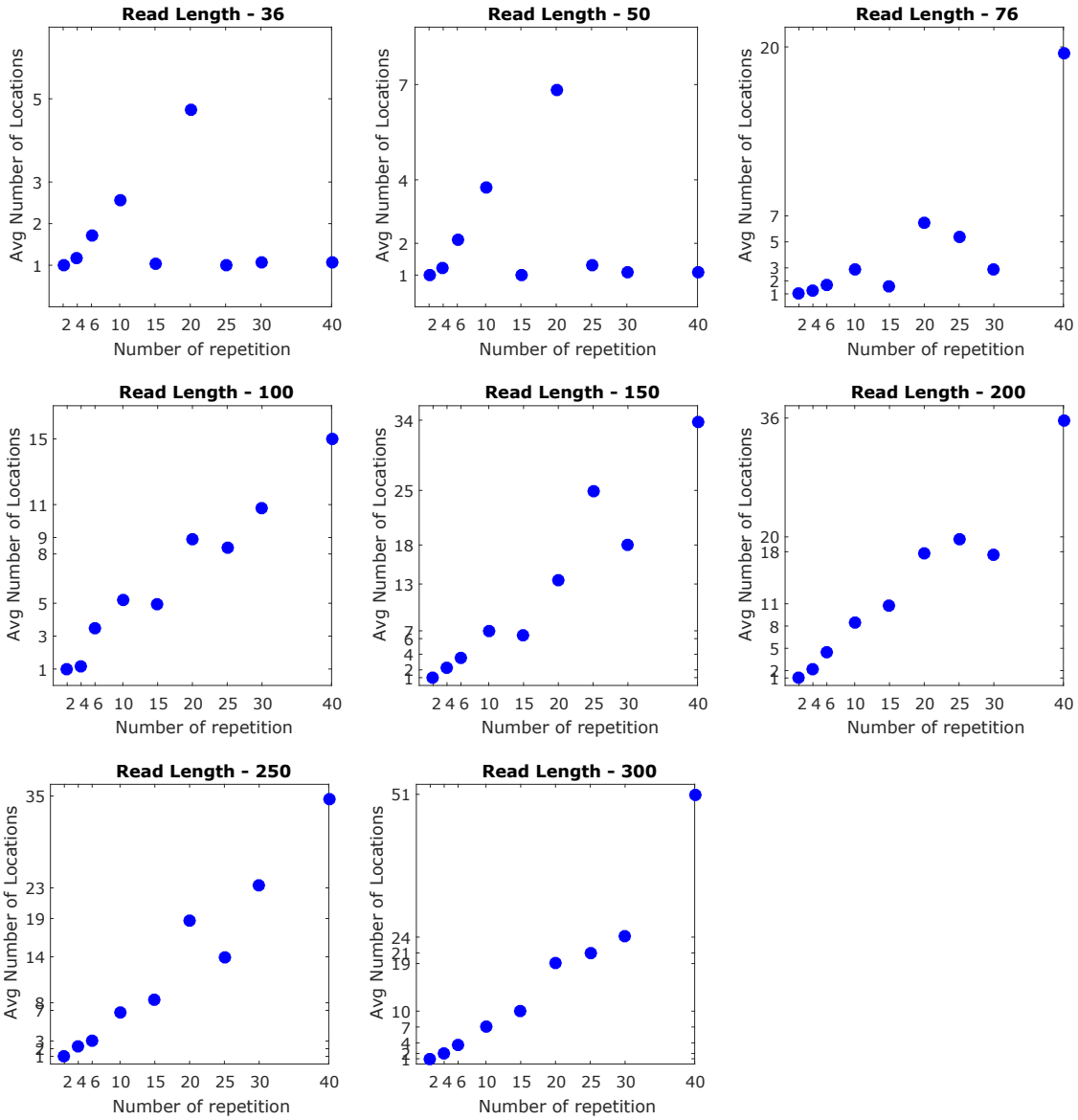

Figure E.3: Mapping Linearity using BWA-MEM with different read length.

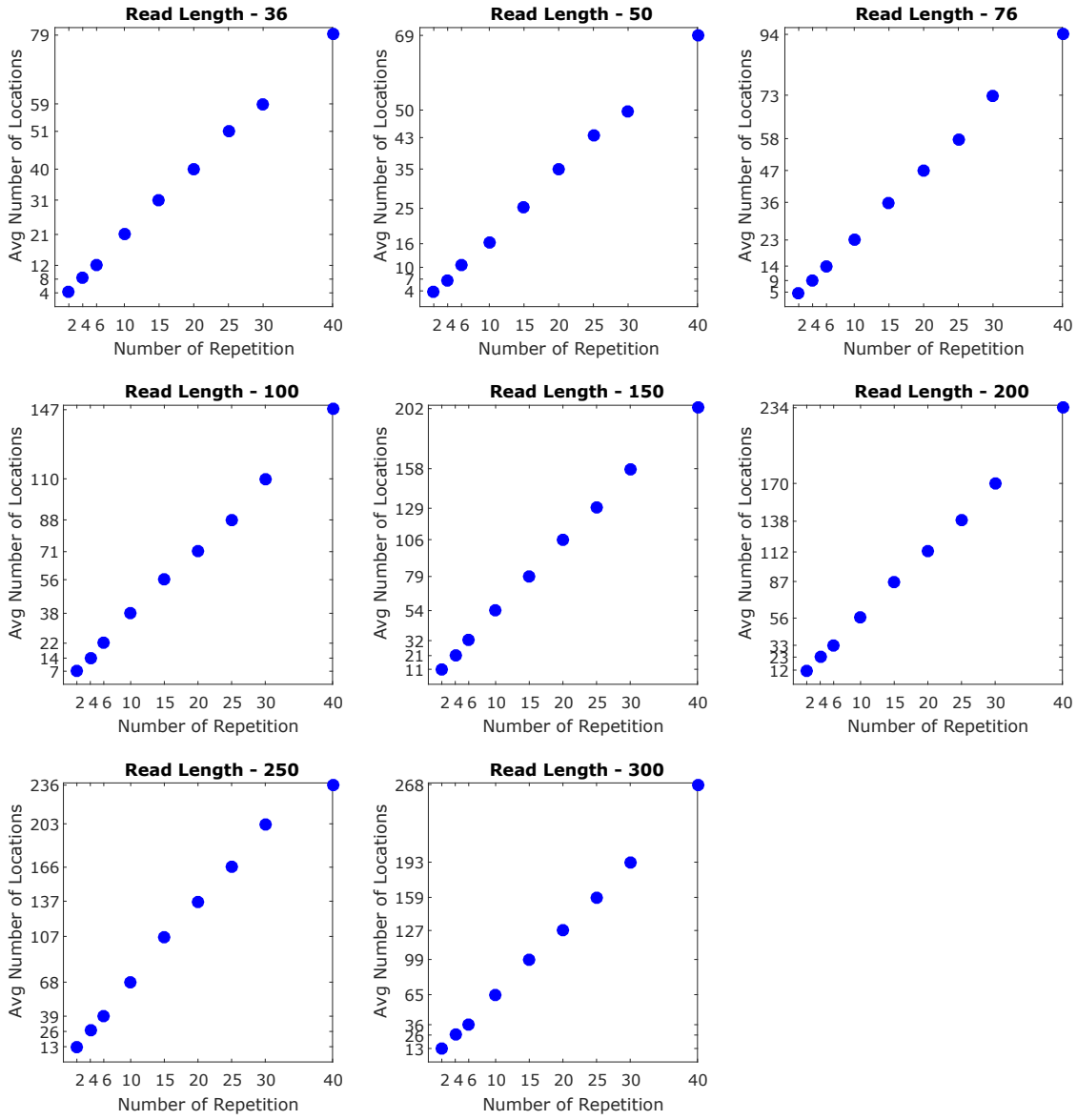

Figure E.4: Mapping Linearity using NN with different read length.

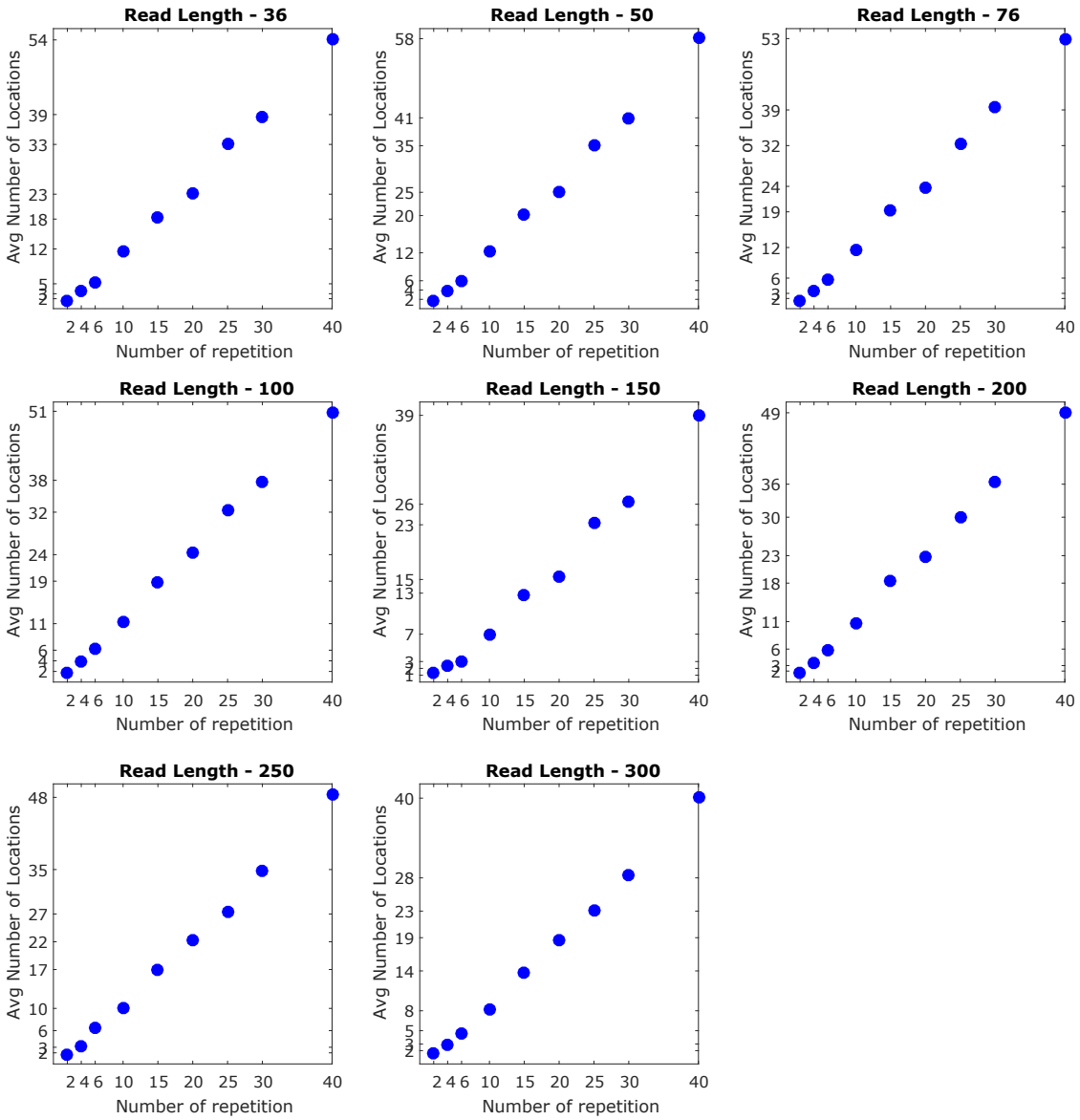

Figure E.5: Mapping Linearity using Bowtie2 with different read length.

E.2 CNVs Detection with Different Similarity

In the section, the detected CNV lists is reported with the grouping function is on, and the CNV regions are detected by rejecting the regions that have estimated query copy numbers within 0.9 of the estimated reference copy numbers ($|CopyNumber_{Avg} - CopyNumber_{Avg} \times CopyNumber_{TS}| >= 0.9$). In all the tables, CNV size is presented in base pairs (bps) and the breakpoints are listed in the reference (Seq-2) genome index. The negative group numbers represent deletions and there is no false duplications detected.

Table E.1: CNV detection results for similarity test. CNV size is in base pairs (bps).

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
286208	287168	2.39	0.105	0.0929	-1	960
373248	374400	2.22	0.284	0.162	-1	1152
818944	819968	2.98	0.391	0.185	-1	1024
1544768	1545536	3.62	1.18	0.445	-1	768
1607616	1608448	3.63	0.576	0.105	-1	832
1859328	1860160	3.61	0.758	0.167	-1	832
2684928	2685952	3.17	1	0.4	-1	1024
608768	609024	0.806	0.618	0.792	-2	256
532096	532224	0.928	0.69	0.799	-3	128
1718656	1718784	0.871	0.749	0.85	-4	128
287105	287168	0.861	0.351	0.405	-5	63

E.3 CNVs Detection with Different Read Lengths

In the section, the detected CNV lists and the read depth trajectories from all read lengths are reported. For each read length, two sets of data are reported. One is using Seq-2 (repeated sequence) as the query sequence and the original sequence as the reference sequence and vice versa.

For all the lists reported here, the grouping function is off, and the CNV regions are detected by rejecting the regions that have estimated query copy numbers within 0.9 of the estimated reference copy numbers ($|CopyNumber_{Avg} - CopyNumber_{Avg} \times CopyNumber_{TS}| >= 0.9$). To reduce false discoveries, regions with TS read depth within Poisson noise of the average TS read depth are also rejected. In all the tables, CNV size is presented in base pairs (bps) and the breakpoints are listed in the reference genome index. For group numbers, positive groups represent duplications while negative groups represent deletions. As a result, $group_{-n} \neq group_n$, in which n is the group number.

E.3.1 CNV Lists

Table E.2: CNV detection results of 36-bp reads with the original as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
39360	39432	8.06	5.11	1.22	4	72
49616	49920	8.08	8.01	1.21	9	304
50048	50144	8.94	9.25	1.52	10	96
60928	60992	6.94	5.71	1.35	13	64
78784	78920	5.89	4.54	1.19	17	136
89920	90000	4.3	4.21	1.39	19	80
106112	106192	4.47	3.43	1.21	26	80
114384	114432	5.31	2.3	1.2	29	48
114816	114864	3.66	2.66	1.29	30	48
121344	121548	3.85	3.31	1.33	31	204
170112	172160	4.43	5.28	1.46	39	2048
183296	183344	6.35	3.77	1.23	43	48
196512	198656	4.26	5.02	1.3	47	2144
377040	377154	7.07	6.48	1.34	73	114
387216	387296	4.55	5.01	1.44	74	80
393216	393280	7.18	6.8	1.31	75	64
397632	397952	5.45	4.94	1.26	78	320
412176	412272	7.78	7.45	1.4	79	96
423232	423344	9.99	10.9	1.26	83	112
443952	444080	9.8	10.1	1.36	90	128
444096	444208	7.38	6.62	1.25	91	112
462880	463168	9.36	9.12	1.34	93	288
463200	463328	6.27	3.71	1.16	94	128
519232	519328	10.6	10.3	1.27	97	96
519392	519488	7.69	7.81	1.24	98	96
530304	530464	7.62	7.85	1.32	102	160
532416	532528	6.48	7.77	1.46	104	112
554816	554992	3.48	2.85	1.33	108	176
564918	564960	26	9.16	1.23	113	42
565507	565578	18.5	6.35	1.21	121	71
565668	565712	34	11.8	1.25	123	44
567204	567252	29.4	9.02	1.18	139	48
583424	583552	5.98	5.41	1.29	155	128
586432	586528	3.82	2.62	1.24	156	96
620672	622208	5.03	6.14	1.32	158	1536
651488	651536	6.23	5.05	1.37	161	48
651712	651776	4.03	3.73	1.39	162	64
676736	676928	8.56	8.95	1.4	165	192
684832	684912	4.92	4.46	1.24	167	80
689718	689824	6.31	6.38	1.29	168	106
701632	701760	4.68	4.76	1.29	170	128
742688	742880	4.2	4.11	1.29	178	192
759744	759800	5.26	3.44	1.24	185	56
759968	760032	7.16	2.58	1.17	186	64
785856	786112	5.83	7.69	1.63	187	256
788976	789096	10.4	9.86	1.23	188	120
789120	789312	8.17	8.22	1.32	189	192
821728	823264	6.21	5.96	1.21	192	1536
834304	836672	1.05	2.08	1.98	195	2368
888312	888448	5.11	5.1	1.4	208	136
895872	895968	5.32	4.3	1.26	209	96
927168	1052032	1.08	2.07	1.98	212	124864
940448	940528	6.19	6.9	1.82	213	80
944152	944280	5.44	6.88	2.05	214	128
955904	956096	5.49	6.67	1.76	215	192

Table E.2 CNV detection results of 36-bp reads with the original as reference sequence. Continued

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
956112	956288	5.41	6.35	1.87	216	176
1005952	1006016	7.78	8.51	1.89	217	64
1011184	1011232	5.31	7.07	1.95	218	48
1011328	1011368	8.23	6.92	1.68	219	40
1018816	1018912	3.55	6.88	2.29	220	96
1025184	1025248	6.36	4.94	2.06	221	64
1025280	1052160	1.03	2.01	1.98	222	26880
1076064	1076176	4.72	3.46	1.22	232	112
1148416	1148560	9.13	9.65	1.35	240	144
1181664	1181778	7.02	4.8	1.22	244	114
1203808	1205248	5.07	6.12	1.27	248	1440
1205249	1205344	3.37	3.67	1.34	249	95
1253616	1253664	4.31	4.49	1.34	262	48
1266208	1266328	7.22	6.78	1.38	264	120
1272368	1272448	7.39	4.64	1.22	266	80
1307984	1308096	7.53	7.1	1.35	269	112
1331872	1331988	6.64	3.99	1.18	271	116
1339264	1339360	10.8	11.3	1.27	273	96
1339408	1339584	7.8	9.79	1.53	274	176
1358944	1359040	7.29	7.64	1.43	277	96
1399888	1399968	8.61	6.54	1.19	281	80
1402336	1402448	4.82	5.97	1.57	283	112
1415552	1415616	2.33	2.62	1.44	287	64
1418272	1418496	3.12	3.77	1.39	288	224
1477888	1477952	5.32	4.92	1.25	298	64
1504560	1504640	6.98	6.17	1.47	304	80
1504800	1504928	8.49	5.5	1.17	305	128
1533856	1533920	3.96	3.28	1.32	309	64
1545552	1545632	5.38	4.27	1.22	312	80
1551472	1553024	4.99	5.9	1.26	314	1552
1553184	1553248	8.39	5.53	1.17	315	64
1558528	1808384	1.06	2.04	1.97	318	249856
1613120	1613312	4.19	6.37	2.05	319	192
1644416	1644480	3.49	4.95	2.14	320	64
1690432	1690496	3.85	4.38	2.13	321	64
1715248	1715328	7.08	10	2.12	322	80
1717440	1717632	5.02	6.63	2.11	323	192
1730400	1730448	6.83	5.53	1.79	324	48
1739808	1739856	6.65	5.54	1.76	325	48
1739904	1739970	8.78	7.47	1.9	326	66
1748864	1748944	3.06	5.44	2.1	327	80
1763968	1764096	5.6	4.31	1.77	328	128
1774720	1794048	1.05	2.04	1.99	329	19328
1794080	1808384	1.05	2.06	2	330	14304
1808896	1808936	4.54	2.58	1.22	331	40
1836976	1837024	3.92	2.99	1.33	341	48
1842560	1842624	7.25	5.32	1.27	343	64
1851904	1852032	5.27	5.32	1.33	345	128
1859376	1859442	5.53	4.62	1.3	348	66
1862896	1862992	6.81	5.46	1.21	349	96
1905888	1905960	7.62	7.46	1.41	350	72
1930688	1930752	6.54	6.13	1.46	356	64
1958464	1958528	3.98	3.39	1.29	360	64
1996288	1997824	6.22	6.21	1.38	365	1536
2047232	2047368	6.3	6.41	1.38	382	136

Table E.2 CNV detection results of 36-bp reads with the original as reference sequence. Continued

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
2056176	2056224	11.9	5.75	1.26	386	48
2056240	2056352	10.7	4.58	1.19	387	112
2057496	2057544	12.2	4.42	1.17	397	48
2086620	2086687	7.14	5.39	1.28	407	67
2088032	2088072	4.66	3.57	1.29	408	40
2130624	2130816	4.97	5.98	1.36	413	192
2131384	2131488	7.37	4.98	1.17	414	104
2176464	2176560	6.02	4.56	1.25	419	96
2248944	2249008	9.06	5.06	1.2	426	64
2257648	2257984	6.54	6.53	1.3	429	336
2257912	2257968	9.59	8.88	1.22	430	56
2286592	2288192	4.83	5.55	1.36	433	1600
2348160	2348208	4.11	4.06	1.33	440	48
2371760	2371928	7.4	6.41	1.33	443	168
2394752	2394832	6.6	6.64	1.51	447	80
2404992	2405096	4.91	3.78	1.21	449	104
2433184	2433312	6.22	4.64	1.26	453	128
2437656	2437760	5.84	4.04	1.2	456	104
2441472	2441792	5.17	4.85	1.19	458	320
2452864	2452952	6.47	6.31	1.2	460	88
2464512	2464768	5.65	5.82	1.34	463	256
2469952	2470080	5.09	6.31	1.6	464	128
2479520	2479680	5.35	4.58	1.27	469	160
2483264	2483328	6.42	5.89	1.19	473	64
2535680	2536096	1.66	2.28	1.58	480	416
2640960	2641024	3.93	3.53	1.27	496	64
2692512	2692560	9.22	7.42	1.4	507	48
2721993	2722064	7.39	5.77	1.28	508	71
2732096	2732192	7.09	6.42	1.34	510	96
2742016	2766848	1.09	2.05	1.98	511	24832
2743104	2743168	6.63	5.97	1.61	512	64
2743968	2744080	7.3	6.85	1.78	513	112
2754944	2758144	1.05	1.93	1.95	514	3200
2758208	2766848	1.03	2.02	1.97	515	8640
2793488	2793600	5.61	4.91	1.32	520	112
2856720	2856834	60.1	17.2	1.22	537	114
2856784	2856896	65.1	19.2	1.19	538	112
2857176	2857232	60.1	14.1	1.18	542	56
2857248	2857307	63.6	14.9	1.32	543	59
2857504	2857560	67.5	21.7	1.3	547	56
2857627	2857692	60.5	21.9	1.18	549	65
2857840	2857890	73	23.9	1.22	552	50
2858700	2858742	84	26.9	1.3	563	42
2858826	2858880	66.6	19	1.18	565	54
2859024	2859072	66.5	17.8	1.45	568	48
2859288	2859336	66.7	17.6	1.19	572	48
2860416	2860464	70.1	18.2	1.21	586	48
2860740	2860794	70.5	23.1	1.28	590	54
2861592	2861652	69	23.4	1.23	599	60
2861800	2861844	83.4	26.9	1.22	602	44
2862060	2862114	62.3	12.1	1.17	606	54
2862528	2862576	67.6	21.6	1.28	611	48
2863128	2863168	73.3	17.6	1.18	618	40
2863248	2863368	62.5	20.3	1.17	620	120
2863446	2863566	60.8	17	1.18	622	120
2864232	2864288	64	15.1	1.17	628	56

Table E.2 CNV detection results of 36-bp reads with the original as reference sequence. Continued

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
2865019	2865144	54.4	16.3	1.18	636	125
2865162	2865210	69.3	12.4	1.17	637	48
2865354	2865408	75.1	29.1	1.17	640	54
2865424	2865474	76.4	22.5	1.2	641	50
2865558	2865606	72.7	17.7	1.22	643	48
2865824	2865864	70.8	14.3	1.17	646	40
2865888	2866072	47	10.2	1.19	647	184
2866144	2866215	51.2	16.9	1.23	649	71
2867400	2867448	63.6	12.4	1.17	663	48
2867532	2867580	67.1	21	1.32	665	48
2868000	2868048	80.3	25	1.27	668	48
2892176	2892288	6.9	6.36	1.31	673	112
2900992	2901032	4.59	2.2	1.33	675	40
181376	181504	0.901	0.707	0.778	-8	128
431360	431424	0.942	0.718	0.789	-13	64
1073408	1073472	0.891	1.08	1.23	-26	64
1391360	1391424	1.03	0.823	0.825	-32	64
1545472	1545536	0.91	1.11	1.2	-36	64
2188416	2188480	0.848	0.642	0.778	-51	64
2689152	2689248	0.882	0.649	0.737	-61	96
2857770	2857824	68.5	15.2	0.833	-67	54
2858166	2858216	71.6	17.5	0.818	-68	50
2899392	2899456	0.819	0.634	0.787	-69	64
2901856	2901944	8.17	2.39	0.809	-70	88

Table E.3: CNV detection results of 36-bp reads with Seq-2 as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
17760	17824	4.64	3.42	1.31	1	64
43520	43712	2.57	3.05	1.35	6	192
74816	74928	9.87	10.4	1.48	26	112
74929	75138	7.63	8.63	1.41	27	209
180592	180640	4.71	2.06	1.19	50	48
217806	217848	6.4	3.02	1.24	59	42
221504	223520	4.32	4.31	1.22	61	2016
223904	223952	4.89	2.62	1.21	63	48
230328	230368	5.41	3.39	1.36	67	40
240492	240567	4.99	2.8	1.28	68	75
356672	358592	2.79	3.63	1.42	81	1920
397632	399552	2.82	3.61	1.38	85	1920
412224	412288	5.18	4.9	1.39	89	64
437184	437280	8.48	7.3	1.34	95	96
448032	448352	8.39	8.84	1.39	99	320
464720	464768	5.2	2.4	1.18	104	48
468960	469092	9.91	9.3	1.3	105	132
477804	477888	8.62	7.48	1.28	107	84
487896	488000	10.5	11.3	1.38	109	104
488048	488160	7.86	7.81	1.29	110	112
544240	544320	11.6	10.1	1.18	115	80
544384	544512	6.8	6.12	1.17	116	128
555312	555392	9.49	8.98	1.32	119	80
555361	555472	6.76	6.78	1.36	120	111
557312	557384	11.5	9.54	1.24	121	72
557408	557536	6.72	6.61	1.26	122	128
589920	589962	27.4	11.4	1.25	134	42
590856	590916	23.5	9.8	1.39	145	60
591204	591252	28.7	9.28	1.21	149	48
591264	591318	27.4	9.68	1.27	150	54
592051	592122	18.6	5.84	1.18	161	71
592408	592458	27.3	10.2	1.26	165	50
592476	592528	28	8.48	1.25	166	52
592872	592928	25.9	10.8	1.2	168	56
593550	593600	28.5	9.27	1.18	176	50
593616	593687	19.8	6	1.17	177	71
593731	593802	21.6	8.08	1.17	179	71
750744	750784	4.79	2.61	1.2	203	40
814112	814336	8.27	6.96	1.19	218	224
892176	892320	4.14	3.24	1.27	230	144
920896	920976	7.1	4.97	1.21	237	80
1113248	1113408	4.21	3.05	1.22	297	160
1173432	1173568	9.74	9.8	1.34	305	136
1206656	1206784	6.47	4.36	1.25	309	128
1278616	1278688	4.18	3.68	1.24	325	72
1291392	1291440	5.8	2.53	1.2	328	48
1297368	1297440	7.69	4.46	1.28	329	72
1364272	1364352	11.7	11.2	1.22	337	80
1364416	1364576	7.86	8.71	1.37	338	160
1364481	1364576	6.6	6.6	1.24	339	95
1374752	1376304	6.22	7.44	1.51	340	1552
1383936	1384032	7.99	6.71	1.22	342	96
1427264	1427304	5.81	3.16	1.31	347	40
1427328	1427456	5.26	4.52	1.18	348	128
1459032	1459072	5.37	3.04	1.22	354	40

Table E.3 CNV detection results of 36-bp reads with the Seq-2 as reference sequence. Continued

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
1484672	1484752	5.28	2.36	1.19	360	80
1502880	1502944	5.64	3.96	1.17	366	64
1740248	1740320	9.02	8.38	1.34	429	72
1876896	1877024	5.91	4.67	1.2	480	128
1930896	1930963	9.12	7.95	1.39	486	67
2021280	2022848	6.22	7.4	1.28	501	1568
2024320	2024384	6.66	3.37	1.16	503	64
2024616	2024880	5.99	3.44	1.22	506	264
2197248	2197376	6.49	5.66	1.39	560	128
2205976	2206016	13.2	6.08	1.18	563	40
2206110	2206152	12.2	6.26	1.36	565	42
2206153	2206224	8.75	4.11	1.19	566	71
2206304	2206352	12.4	7.2	1.29	568	48
2206368	2206416	8.33	3.29	1.2	569	48
2221440	2221504	7.72	4.49	1.22	582	64
2236624	2236672	9.68	5.68	1.24	583	48
2292096	2293632	6.24	7.52	1.19	592	1536
2320032	2322048	2.85	3.68	1.47	596	2016
2375136	2375200	6.47	2.36	1.19	604	64
2459776	2459840	4.2	2.6	1.22	617	64
2498160	2498208	5.29	4.14	1.25	620	48
2521760	2521871	5.98	4.7	1.19	623	111
2521872	2521928	11.9	8.35	1.18	624	56
2693336	2693408	8.82	9.43	1.45	666	72
3196352	3197888	6.26	7.54	1.19	786	1536
3252992	3254528	6.25	7.46	1.21	794	1536
3258816	3258870	70.4	22.4	1.16	801	54
3259408	3259458	64.2	16.6	1.22	808	50
3259812	3259992	48.8	12.5	1.19	812	180
3260008	3260058	62.4	16.8	1.3	813	50
3261448	3261492	78.8	22.4	1.3	827	44
3261774	3261824	54.2	11.4	1.18	830	50
3261904	3261960	64.7	14	1.19	832	56
3262104	3262152	65.3	14	1.19	835	48
3262302	3262356	65.4	23.7	1.25	837	54
3262640	3262680	73.4	18.6	1.27	842	40
3263104	3263144	75.5	19.6	1.18	848	40
3263560	3263610	66.7	17.9	1.38	855	50
3264414	3264468	66.6	23	1.22	865	54
3264469	3264600	49.6	13.9	1.22	866	131
3264876	3264930	70.6	24	1.17	870	54
3264931	3264996	52.6	12.5	1.16	871	65
3265480	3265520	70.6	18.5	1.22	878	40
3265680	3265722	75.8	22.8	1.2	880	42
3267054	3267104	69.3	13.4	1.22	896	50
3267114	3267168	62	17	1.29	897	54
3267367	3267432	55.1	17.5	1.2	900	65
3268312	3268352	75.2	21.4	1.19	911	40
3268768	3268884	62.8	17.4	1.25	915	116
3269034	3269072	76.3	15.9	1.22	918	38
3269232	3269357	53	12.2	1.25	921	125
3269296	3269336	71	14.7	1.29	922	40
3269941	3270000	62.4	21.1	1.36	930	59
3270024	3270072	74.8	21.7	1.17	931	48
3270084	3270138	65.7	19.1	1.3	932	54

Table E.3 CNV detection results of 36-bp reads with the Seq-2 as reference sequence. Continued

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
3270273	3270330	59.2	21.3	1.2	935	57
3270348	3270402	63.7	13.7	1.33	936	54
3270750	3270800	76.5	26.6	1.24	939	50
3294924	3295040	7.31	5.57	1.25	943	116
45696	70656	1.97	1.06	0.552	-1	24960
860160	861696	1.98	1.05	0.545	-3	1536
952192	952256	2.9	1.45	0.465	-4	64
952320	1077248	2	1.06	0.538	-5	124928
1083776	1083839	6.24	3.09	0.797	-6	63
1564448	1564576	5.09	2.09	0.792	-7	128
1570552	1570624	6.42	3.59	0.837	-8	72
1583104	1740288	1.93	1.05	0.553	-9	157184
1740321	1833472	1.96	1.06	0.557	-10	93151
1883392	1883504	6.63	2.15	0.815	-11	112
2040704	2040768	6.76	2.23	0.809	-12	64
2048000	2173952	1.99	1.06	0.543	-13	125952
2217312	2217472	5.9	2.85	0.831	-14	160
2415420	2415552	5.71	2	0.819	-15	132
2536448	2786304	1.94	1.06	0.555	-16	249856
3142144	3142656	1.97	0.981	0.486	-17	512
3143088	3143168	8.18	3.9	0.668	-18	80
3145728	3166208	1.91	1.03	0.554	-19	20480
3166720	3166976	1.64	0.92	0.564	-20	256
3197952	3200512	1.93	1.07	0.567	-22	2560
3334728	3334784	8.54	1.81	0.715	-23	56

Table E.4: CNV detection results of 50-bp reads with the original as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
49616	49920	6.69	7.41	1.24	10	304
78592	78720	6.04	4.43	1.17	16	128
143264	143360	4.49	4.1	1.21	28	96
177504	178400	4.95	6.11	1.23	30	896
281696	281920	7.55	6.78	1.16	40	224
346672	346880	6.7	5.36	1.19	46	208
362336	364320	4.17	5.07	1.33	47	1984
377056	377152	7.13	7.22	1.4	51	96
397632	397952	4.41	4.27	1.37	55	320
412160	412272	5.85	4.82	1.25	56	112
423024	423104	8.39	8.08	1.17	59	80
443950	444192	7.82	7.77	1.34	63	242
452800	452960	5.76	5.43	1.38	64	160
462896	463168	8.07	7.93	1.19	65	272
519230	519504	7.59	8.68	1.41	69	274
530160	530240	8.56	7.09	1.24	71	80
530304	530480	6.86	6.99	1.33	72	176
532320	532544	6.73	6.25	1.23	73	224
583480	583552	7.18	7.6	1.62	84	72
651488	651552	3.42	3.68	1.5	89	64
676576	676704	7.82	6.75	1.2	92	128
676784	676928	7.3	7.2	1.26	93	144
684832	684912	3.38	3.38	1.4	95	80
689720	689840	5.28	4.74	1.26	96	120
737460	737580	7.83	7.19	1.28	100	120
785856	786112	4.53	5.78	1.56	108	256
788976	789088	8.07	7.6	1.22	109	112
789152	789312	6.99	6.59	1.19	110	160
834560	836608	1.06	1.94	1.86	116	2048
845109	846592	6.1	6.87	1.17	118	1483
888384	888448	2.17	2.96	1.48	126	64
927232	1052096	1.04	2	1.95	130	124864
1095712	1095808	3.98	3.59	1.39	138	96
1148416	1148544	7.76	7.38	1.25	141	128
1181664	1181776	4.87	4.36	1.4	146	112
1182912	1183011	5.8	4.37	1.22	147	99
1317152	1317216	5.89	4.05	1.27	164	64
1339248	1339584	7.83	9.39	1.44	166	336
1358848	1358912	5.14	4.31	1.34	168	64
1358944	1359040	6.01	6.71	1.38	169	96
1363392	1363504	5.43	4.26	1.28	170	112
1399872	1399968	6.46	4.53	1.19	173	96
1476736	1476848	4.96	5.45	1.47	184	112
1477888	1477952	4.01	4.3	1.39	185	64
1504544	1504640	6.12	4.46	1.2	191	96
1551488	1553024	4.64	4.94	1.33	199	1536
1558528	1688576	1.05	2.04	1.95	202	130048
1611584	1611680	3.07	6.48	2.22	203	96
1613504	1613600	2.48	5.76	2.56	204	96
1644544	1684608	1.05	2.06	1.97	205	40064
1684736	1688576	1.01	1.93	1.93	206	3840
1688640	1724544	1.06	2.01	1.95	207	35904
1724672	1808384	1.04	2	1.94	208	83712
1739904	1739968	6.66	5.86	1.98	209	64
1792512	1802112	1.02	2.01	1.98	210	9600

Table E.4 CNV detection results of 50-bp reads with the original as reference sequence. Continued

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
1802176	1808384	1.03	1.93	1.91	211	6208
1810896	1811040	5.73	4.12	1.2	212	144
1822224	1822304	5.7	4.4	1.21	215	80
1834944	1835008	4.47	4.31	1.32	218	64
1842464	1842630	6.03	5.03	1.27	220	166
1851904	1851968	6.13	5.74	1.35	222	64
1859392	1859456	3.3	2.99	1.35	226	64
1862912	1862992	6.45	4.23	1.2	227	80
1905888	1905960	5.56	4.88	1.4	228	72
1925776	1925888	5.61	3.57	1.17	232	112
1930688	1930752	6.74	4.13	1.3	233	64
1932896	1932960	4.47	2.98	1.23	234	64
1958448	1958528	5.79	3.47	1.2	237	80
1996288	1997824	6.2	7.17	1.26	241	1536
2047232	2047368	5.26	5.43	1.41	247	136
2047296	2047360	7.56	8.58	1.73	248	64
2070176	2070272	4.49	3.15	1.22	251	96
2130624	2130832	6.82	8.11	1.27	259	208
2131380	2131460	6.98	5.03	1.17	260	80
2176472	2176560	5.75	4.32	1.35	265	88
2239200	2239264	4.02	3.62	1.36	270	64
2257648	2257984	4.81	4.32	1.29	274	336
2315904	2316032	4.74	3.34	1.21	280	128
2320912	2321024	4.57	3.33	1.2	281	112
2352480	2352544	2.14	3.02	1.58	285	64
2371744	2371936	5.61	4.74	1.23	286	192
2394752	2394848	5.52	5.12	1.42	292	96
2423680	2423808	2.78	3.33	1.38	295	128
2433216	2433312	6.12	4.08	1.21	297	96
2452864	2452952	6.05	5.68	1.49	302	88
2456192	2456280	7.81	8.8	1.64	303	88
2464512	2464768	4.41	4.9	1.44	306	256
2469952	2470176	4.63	5.43	1.46	307	224
2640640	2640768	4.66	4.14	1.3	336	128
2692256	2692560	4.41	3.41	1.22	344	304
2722000	2722072	6.9	5.15	1.32	345	72
2732096	2732208	5.77	5.76	1.47	347	112
2742272	2766848	1.07	2	1.95	348	24576
2796350	2797824	6.05	7.22	1.41	353	1474
2797825	2797888	4.16	4.24	1.3	354	63
2850496	2852032	6.22	7.16	1.25	360	1536
2892160	2892288	5.4	5.69	1.45	368	128
2923790	2923851	6.86	3.71	1.24	370	61
65408	65472	0.927	1.01	1.17	-1	64
302592	302656	0.952	1.19	1.26	-4	64
838048	838144	4.02	1.35	0.746	-11	96
869808	869872	5.15	1.6	0.794	-13	64
885632	885720	5.07	2.71	0.791	-14	88
1546624	1546688	1.25	0.773	0.773	-23	64
2797888	2797952	1.31	0.578	0.539	-33	64

Table E.5: CNV detection results of 50-bp reads with Seq-2 as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
75008	75152	7.4	7.96	1.22	31	144
195105	197120	4.25	5.51	1.35	47	2015
202496	203392	4.82	6.15	1.31	48	896
204928	205216	4.08	5.25	1.28	49	288
221504	223520	4.24	5.6	1.35	51	2016
245552	247552	4.28	5.53	1.24	54	2000
306688	306944	6.71	6.1	1.19	55	256
371841	371904	3.43	2.06	1.35	62	63
387328	389344	4.16	5.37	1.41	64	2016
402048	402160	6.91	5.32	1.22	68	112
412224	412288	3.65	3.74	1.37	69	64
437180	437280	6.75	5.06	1.3	73	100
448032	448576	6.84	7.9	1.21	77	544
451792	451920	6.1	4.45	1.32	78	128
468952	469216	7.76	7.53	1.31	80	264
477792	477952	5.49	5.43	1.46	82	160
487888	488000	8.09	7.58	1.2	83	112
488048	488160	6.63	6.81	1.31	84	112
544224	544512	7.36	8.18	1.41	88	288
555153	555248	8.13	6.53	1.19	90	95
555312	555472	7.44	6.79	1.16	91	160
608448	608560	6.14	4.72	1.18	103	112
680384	680640	5.31	4.4	1.18	110	256
708640	708704	3.25	3.27	1.35	113	64
709824	709920	3.51	3.5	1.33	114	96
714720	714832	6.14	5.26	1.26	115	112
810848	811136	5.05	5.5	1.33	126	288
813968	814080	7.42	7.41	1.26	127	112
814160	814336	6.6	5.79	1.21	128	176
846720	848256	6.07	7.25	1.22	131	1536
870112	871648	5.93	7.06	1.39	139	1536
920880	920992	3.52	3.18	1.33	147	112
1083840	1085080	6.54	7.21	1.22	207	1240
1120720	1120800	5.11	3.88	1.27	213	80
1173420	1173568	7.76	7.37	1.17	215	148
1291200	1291328	5.68	6.31	1.26	234	128
1297360	1297440	7.26	4.45	1.22	236	80
1332992	1333088	6.53	5.73	1.27	239	96
1364224	1364352	6.4	5.66	1.2	244	128
1364416	1364608	6.27	7.35	1.38	245	192
1374753	1376312	6.16	7.41	1.17	246	1559
1383936	1384032	6.59	6.52	1.44	248	96
1424880	1424960	7.39	5.28	1.25	252	80
1434848	1434931	6.88	3.99	1.19	256	83
1576448	1578048	4.65	4.74	1.21	277	1600
1876896	1877024	5.51	4.56	1.18	467	128
2021280	2022848	6.09	7.49	1.22	487	1568
2197248	2197376	5.98	5.06	1.23	540	128
2280629	2280832	6.91	8.1	1.28	550	203
2281376	2281472	6.01	4.83	1.35	551	96
2292096	2293632	6.17	7.43	1.17	553	1536
2293184	2293248	8.06	10.2	1.31	554	64
2521728	2521855	4.31	3.7	1.27	580	127
2521856	2521930	7.91	6.21	1.25	581	74

Table E.5 CNV detection results of 50-bp reads with Seq-2 as reference sequence. Continued

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
2794752	2794848	6.47	4.55	1.21	742	96
3040640	3040768	5.62	4.69	1.21	783	128
3122000	3122080	6.79	4.34	1.21	791	80
3132096	3132224	6.28	4.69	1.22	793	128
3196352	3197888	5.97	7.45	1.43	816	1536
3252992	3254544	6.14	7.43	1.21	822	1552
3262528	3262656	4.03	2.9	1.26	829	128
3294928	3295040	7.09	5.37	1.17	837	112
45568	70656	1.92	1.03	0.546	-1	25088
82784	82880	4.52	1.53	0.715	-2	96
117824	117992	5.88	4.02	0.839	-3	168
401792	401920	5.3	2.08	0.788	-4	128
859136	861696	2.05	1.09	0.523	-5	2560
952320	1077248	1.93	1.06	0.548	-6	124928
1432480	1432576	4.44	1.64	0.754	-7	96
1583616	1740288	1.85	1.06	0.578	-8	156672
1740352	1833472	1.84	1.05	0.578	-9	93120
2048000	2173760	1.92	1.06	0.55	-11	125760
2536576	2693376	1.85	1.06	0.579	-13	156800
2693440	2785280	1.84	1.05	0.579	-14	91840
3142144	3166976	1.93	1.03	0.537	-15	24832
3197889	3200512	2.04	1.07	0.516	-16	2623

Table E.6: CNV detection results of 76-bp reads with the original as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
443952	444160	4.5	4.02	1.25	21	208
620672	622208	5.15	5.75	1.2	35	1536
834496	836672	1.04	2.11	2.03	46	2176
927232	1052096	1.03	2.05	2.01	50	124864
1203808	1205248	5.28	5.89	1.23	59	1440
1339264	1339584	4.78	5.25	1.33	70	320
1551488	1553024	5.17	5.78	1.18	74	1536
1558528	1808384	1.02	2.04	2	76	249856
2469984	2470080	2.74	3.21	1.39	103	96
2535936	2536064	1.14	2.24	2.04	105	128
2742016	2753664	1.04	2.05	1.99	118	11648
2753792	2766848	1.01	2.01	1.98	119	13056
2796352	2797904	6.58	7.05	1.2	121	1552
2797888	2797952	1.52	0.463	0.451	-9	64

Table E.7: CNV detection results of 76-bp reads with Seq-2 as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
195072	197120	4.33	5	1.21	18	2048
202496	203392	4.99	5.85	1.21	19	896
205296	206336	5.5	6.23	1.18	21	1040
371680	371840	3.89	2.99	1.24	27	160
589824	590096	7.08	8.73	1.31	41	272
645632	647200	4.99	5.78	1.3	47	1568
1228800	1230336	5.08	5.77	1.22	91	1536
1364256	1364544	5.05	5.67	1.28	101	288
1576448	1577984	5.13	5.88	1.23	106	1536
3252992	3254528	6.49	6.58	1.17	272	1536
45568	70656	1.93	1.04	0.545	-1	25088
859136	861696	1.93	1.01	0.525	-2	2560
952320	1077248	1.96	1.03	0.539	-3	124928
1583616	1833984	1.95	1.02	0.533	-4	250368
2048640	2050048	1.92	0.915	0.479	-5	1408
2052544	2052864	3.38	1.32	0.447	-6	320
2060800	2062208	2.07	1.06	0.531	-7	1408
2062336	2065280	1.92	0.965	0.51	-8	2944
2065728	2076928	1.95	1.01	0.529	-9	11200
2077184	2080128	2.15	1.18	0.567	-10	2944
2080768	2081024	2.38	1.25	0.56	-11	256
2081280	2091008	1.94	1.02	0.532	-12	9728
2092032	2095360	2.12	1.13	0.537	-13	3328
2106624	2124160	1.98	1.05	0.542	-14	17536
2124288	2126848	1.95	1.02	0.535	-15	2560
2168960	2170624	2.06	1.03	0.527	-16	1664
2170880	2171392	2.09	1.12	0.542	-17	512
2171904	2172416	2.05	1.06	0.537	-18	512
2172672	2173440	2.11	1.07	0.532	-19	768
2536448	2786624	1.95	1.02	0.533	-21	250176
3142144	3166976	1.93	1.03	0.545	-22	24832
3197889	3200448	1.95	1.04	0.543	-23	2559

Table E.8: CNV detection results of 100-bp reads with the original as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

519232	519424	3.82	3.78	1.25	23	192
834304	836608	1.04	2.17	2.06	37	2304
927232	1052160	1.02	2.04	2	41	124928
1558528	1808384	1.03	2.05	2	61	249856
2742016	2766848	1.03	2.07	2.02	92	24832
2796352	2797888	6.85	7.65	1.18	94	1536
376960	377024	0.84	0.55	0.704	-2	64
1450368	1450496	0.843	0.669	0.798	-5	128

Table E.9: CNV detection results of 100-bp reads with Seq-2 as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

195072	197120	4.5	5.43	1.23	7	2048
202496	203392	5.12	6.13	1.2	8	896
221504	223520	4.48	5.43	1.22	11	2016
387328	389376	4.43	5.39	1.27	16	2048
544256	544448	3.62	4.01	1.39	24	192
3196352	3197824	6.75	7.57	1.18	177	1472
45568	70656	1.99	1.02	0.512	-1	25088
859136	861696	2.07	1.14	0.547	-2	2560
952320	1076991	2.02	1.04	0.515	-3	124671
1076992	1077120	1.74	0.68	0.419	-4	128
1583104	1835008	1.98	1.02	0.522	-6	251904
2048512	2173952	2.02	1.04	0.514	-7	125440
2536448	2786688	1.99	1.03	0.52	-8	250240
3142144	3166976	2	1.02	0.511	-9	24832
3197889	3200512	2.06	1.12	0.535	-10	2623
3260416	3260672	1.94	0.694	0.35	-11	256
3338752	3338848	0.986	0.239	0.249	-12	96

Table E.10: CNV detection results of 150-bp reads with the original as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

834304	836608	1.05	2.23	2.13	29	2304
927232	1051904	1.01	2.03	2	32	124672
1551360	1553024	5.6	5.96	1.16	45	1664
1558528	1808384	1.01	2.04	2.01	46	249856
2742016	2766848	1.01	2.06	2.05	71	24832
2797888	2797952	3.13	0.48	0.333	-5	64

Table E.11: CNV detection results of 150-bp reads with Seq-2 as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

45568	70784	2	0.979	0.491	-1	25216
859136	862208	2	0.971	0.521	-2	3072
952064	1077248	2.02	1.02	0.505	-3	125184
1583616	1833984	2	1.02	0.513	-4	250368
2048640	2048768	1.14	0.277	0.276	-6	128
2048769	2173952	2.03	1.02	0.504	-7	125183
2536448	2786688	2	1.02	0.512	-8	250240
3142144	3166976	2	0.982	0.492	-9	24832
3197824	3200512	2.18	0.971	0.452	-10	2688
3260416	3260672	1.61	0.563	0.337	-11	256

Table E.12: CNV detection results of 200-bp reads with the original as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

834304	836608	1.06	2.07	1.99	22	2304
927232	1052032	1.02	2.06	2.03	25	124800
1558528	1808384	1.02	2.04	2	37	249856
2742272	2766848	1.01	2.02	2	59	24576
2797888	2797952	3.73	0.537	0.278	-4	64

Table E.13: CNV detection results of 200-bp reads with Seq-2 as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

195072	197120	4.63	5.51	1.21	12	2048
204928	205216	3.64	3.07	1.25	14	288
221440	223520	4.65	5.52	1.2	16	2080
245504	247552	4.64	5.51	1.2	17	2048
387328	389376	4.59	5.33	1.21	19	2048
3196352	3197920	7.06	7.82	1.19	101	1568
45568	70784	2	1.03	0.519	-1	25216
859136	861184	2.23	1.14	0.511	-2	2048
861440	861568	1.99	1.02	0.561	-3	128
952320	1077248	2.01	1.02	0.509	-4	124928
1583616	1834496	2.01	1.03	0.517	-5	250880
2048000	2174976	2	1.02	0.513	-6	126976
2535424	2786688	2.02	1.03	0.516	-7	251264
3142144	3142656	1.86	0.927	0.512	-8	512
3143680	3167232	2	1.03	0.52	-9	23552
3197824	3200512	2.13	1.06	0.499	-10	2688
3260288	3260800	1.55	0.201	0.128	-11	512
3338240	3338743	0.98	0.485	0.504	-12	503

Table E.14: CNV detection results of 250-bp reads with the original as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

834304	836608	1.08	2.21	2.07	24	2304
927232	960000	1.02	2.04	2	27	32768
960256	1007104	1.02	2.04	2	28	46848
1007616	1051648	0.998	1.98	1.99	29	44032
1203776	1205248	5.53	5.94	1.2	35	1472
1551488	1552896	5.61	5.88	1.2	41	1408
1558528	1805568	1.01	2.05	2.02	42	247040
2742272	2749440	1.01	2.08	2.06	64	7168
2750464	2754560	1.01	2	1.99	65	4096
2755584	2766848	1.01	2.05	2.02	66	11264
45568	45824	0.874	0.297	0.356	-1	256
2797888	2798080	1.91	0.424	0.392	-2	192

Table E.15: CNV detection results of 250-bp reads with Seq-2 as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

195072	197120	4.21	5.28	1.29	10	2048
221504	223488	4.19	5.26	1.28	14	1984
245536	247552	4.2	5.26	1.27	15	2016
387328	389376	4.21	5.02	1.24	18	2048
1576448	1577984	5.4	5.69	1.2	55	1536
45568	46592	2.04	1	0.515	-1	1024
47616	48640	2.27	1.27	0.558	-2	1024
49152	70912	2.03	1	0.501	-3	21760
861184	861696	2.03	1.14	0.609	-4	512
952320	1077248	2	1.01	0.506	-5	124928
1583616	1584896	2.1	1.16	0.556	-6	1280
1585152	1835008	2	1.02	0.518	-7	249856
2048640	2049024	1.52	0.515	0.295	-9	384
2049024	2173952	2	1.01	0.508	-10	124928
2535424	2785280	2	1.02	0.518	-11	249856
2786560	2786688	1.63	0.412	0.293	-12	128
3142144	3166720	2.01	1.02	0.514	-14	2688
3260416	3260800	1.5	0.271	0.197	-15	384
3338496	3338695	0.921	0.265	0.257	-16	199

Table E.16: CNV detection results of 300-bp reads with the original as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

220544	222592	4.22	4.97	1.22	10	2048
620672	622208	5.06	5.32	1.23	17	1536
834304	836608	1.03	2.23	2.15	22	2304
927232	978944	1.03	2.04	2.01	24	51712
979456	1051904	1.01	2.03	2.01	25	72448
1203840	1205248	5.25	5.55	1.19	30	1408
1551488	1552960	5.16	5.51	1.25	34	1472
1559808	1623040	1.02	2.07	2.03	35	63232
1623296	1808384	1.01	2.01	1.99	36	185088
2742272	2753536	1.03	1.99	1.95	58	11264
2753792	2766848	0.99	2.03	2.05	59	13056
45568	45824	0.904	0.304	0.306	-1	256
2849792	2850304	0.802	0.5	0.582	-2	512
2857728	2857984	0.928	0.359	0.389	-3	256
2935552	2935808	0.785	0.319	0.415	-4	256

Table E.17: CNV detection results of 300-bp reads with Seq-2 as reference sequence. CNV size is in base pairs (bps). The list is generated by excluding regions within Poisson Noise of the average TS read depth and without applying the grouping function.

195072	197120	4.19	5.05	1.26	9	2048
221504	223488	4.17	5.07	1.28	12	1984
245568	247552	4.17	5	1.24	13	1984
387328	389312	4.17	4.88	1.23	15	1984
3196416	3197888	7.16	7.45	1.2	94	1472
45568	46080	1.66	0.551	0.354	-1	512
46081	70912	2.01	1.03	0.518	-2	24831
859392	859648	2.12	1.08	0.524	-3	256
860160	861696	1.99	1.04	0.533	-4	1536
875008	875520	1.13	0.751	0.67	-5	512
952320	1077248	2.03	1.02	0.509	-6	124928
1091584	1092096	0.965	0.556	0.591	-7	512
1583616	1833472	2.01	1.02	0.526	-8	249856
2048000	2048511	1	0.823	0.834	-9	511
2048512	2049024	1.44	0.411	0.28	-10	512
2049024	2172928	2.03	1.02	0.51	-11	123904
2172928	2173952	1.9	0.753	0.401	-12	1024
2535424	2785280	2.01	1.03	0.525	-13	249856
2786304	2786816	1.86	0.59	0.32	-14	512
3142144	3167232	2.01	1.03	0.521	-15	25088
3197696	3200512	2.07	1.08	0.528	-16	2816
3260416	3260672	1.1	0.029	0.0229	-17	256
3260673	3260928	1.21	0.702	0.594	-18	255
3338240	3338645	1.09	0.579	0.484	-19	405

E.3.2 Trajectories

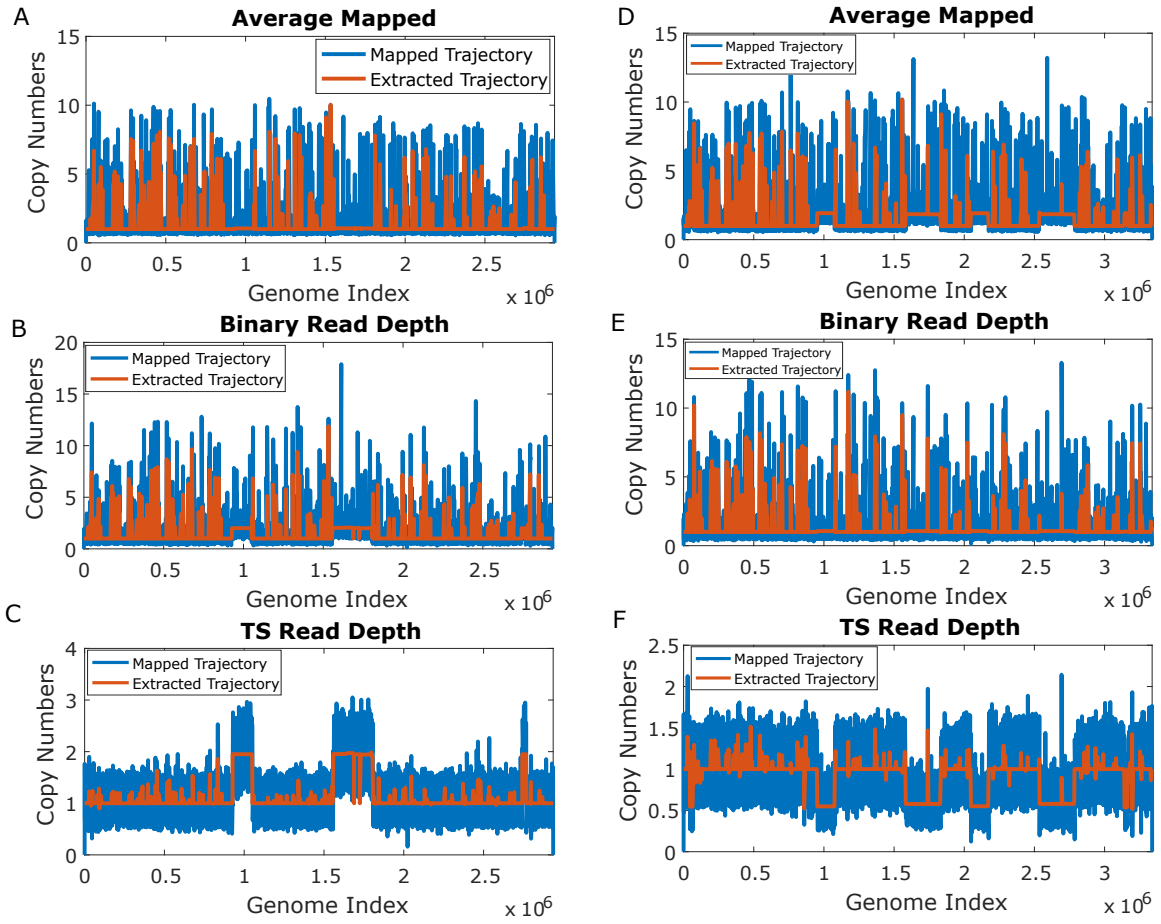

Figure E.6: The read depth trajectories reconstructed from 50-bp reads. (A) to (C) reference: original sequence. Query: Seq-2. (D) to (F) reference: Seq-2. Query: original sequence. (A) and (D) are the average number of assignments trajectories. (B) and (E) are the binary read depth trajectories. (C) and (F) are the test statistics trajectories.

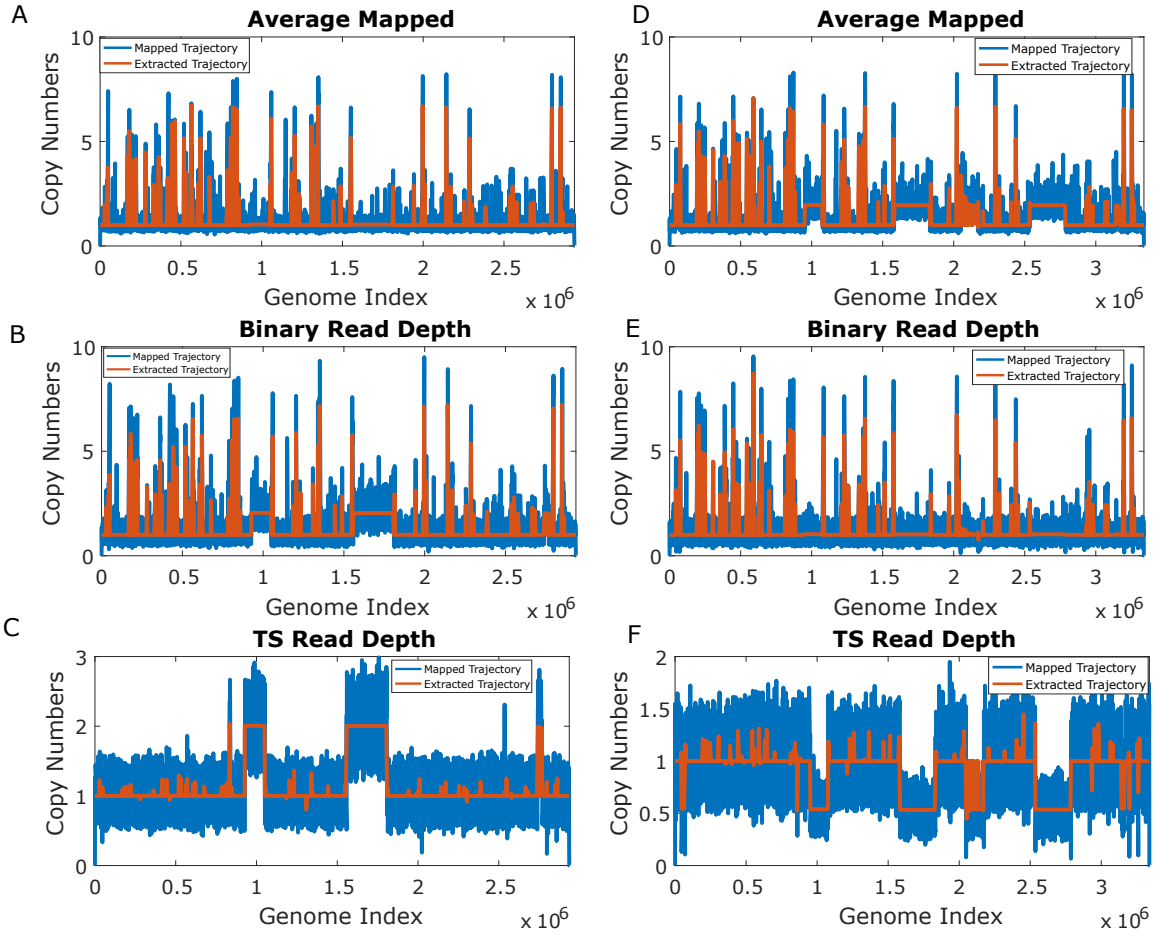

Figure E.7: The read depth trajectories reconstructed from 76-bp reads. (A) to (C) reference: original sequence. Query: Seq-2. (D) to (F) reference: Seq-2. Query: original sequence. (A) and (D) are the average number of assignments trajectories. (B) and (E) are the binary read depth trajectories. (C) and (F) are the test statistics trajectories.

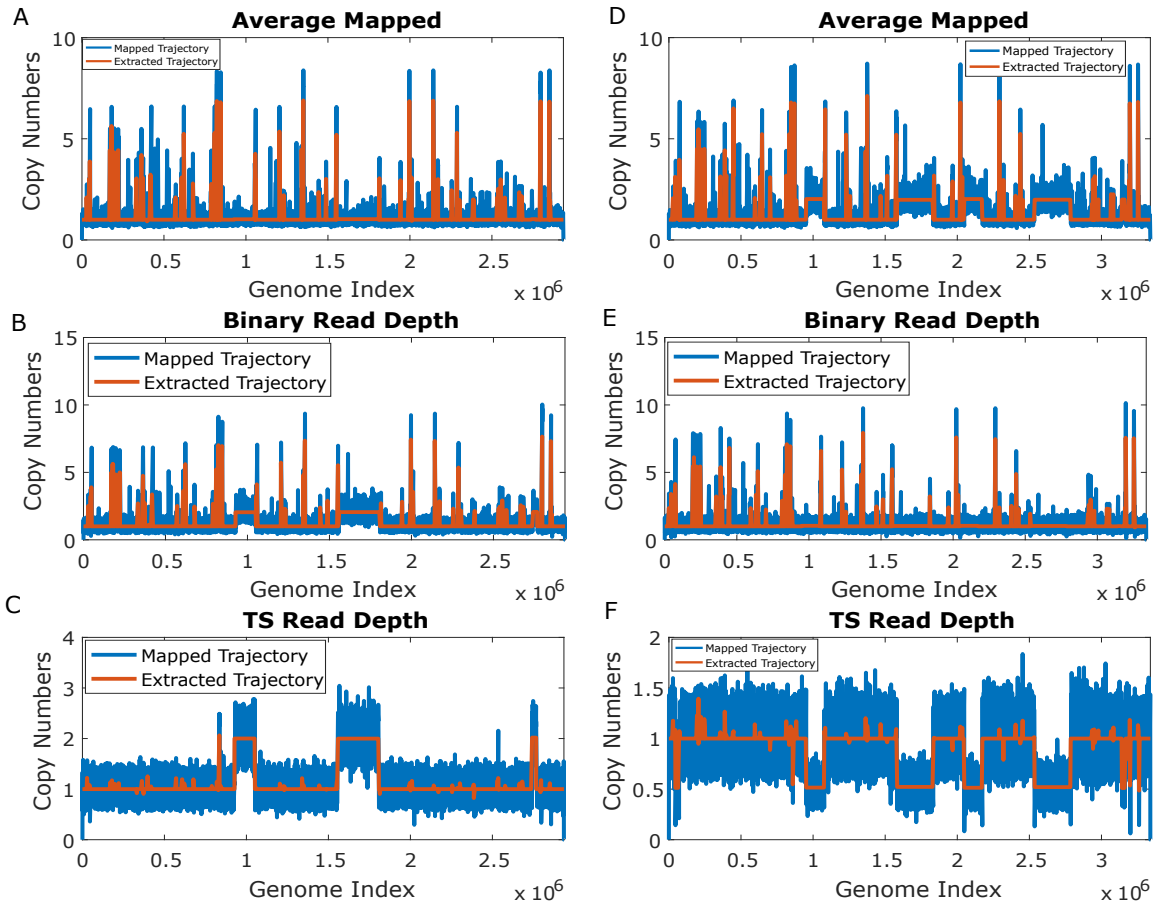

Figure E.8: The read depth trajectories reconstructed from 100-bp reads. (A) to (C) reference: original sequence. Query: Seq-2. (D) to (F) reference: Seq-2. Query: original sequence. (A) and (D) are the average number of assignments trajectories. (B) and (E) are the binary read depth trajectories. (C) and (F) are the test statistics trajectories.

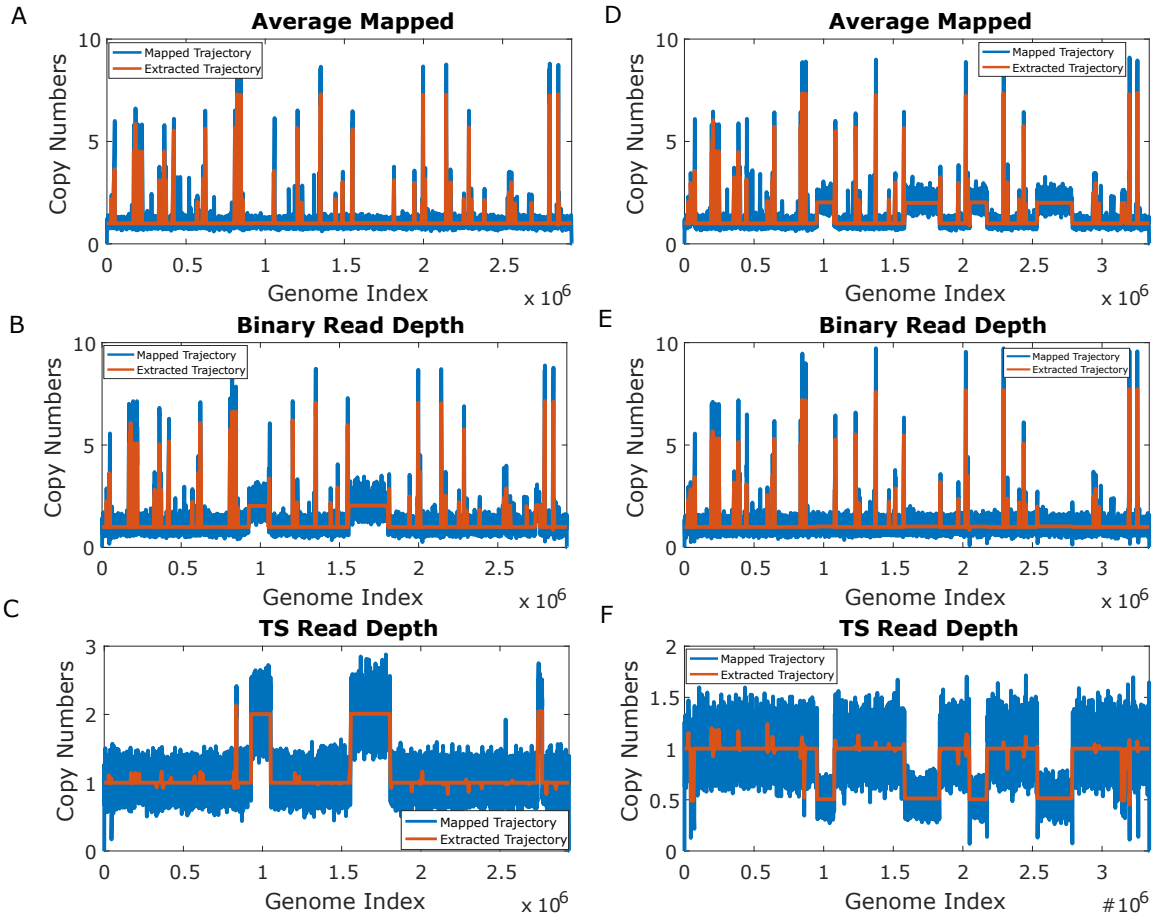

Figure E.9: The read depth trajectories reconstructed from 150-bp reads. (A) to (C) reference: original sequence. Query: Seq-2. (D) to (F) reference: Seq-2. Query: original sequence. (A) and (D) are the average number of assignments trajectories. (B) and (E) are the binary read depth trajectories. (C) and (F) are the test statistics trajectories.

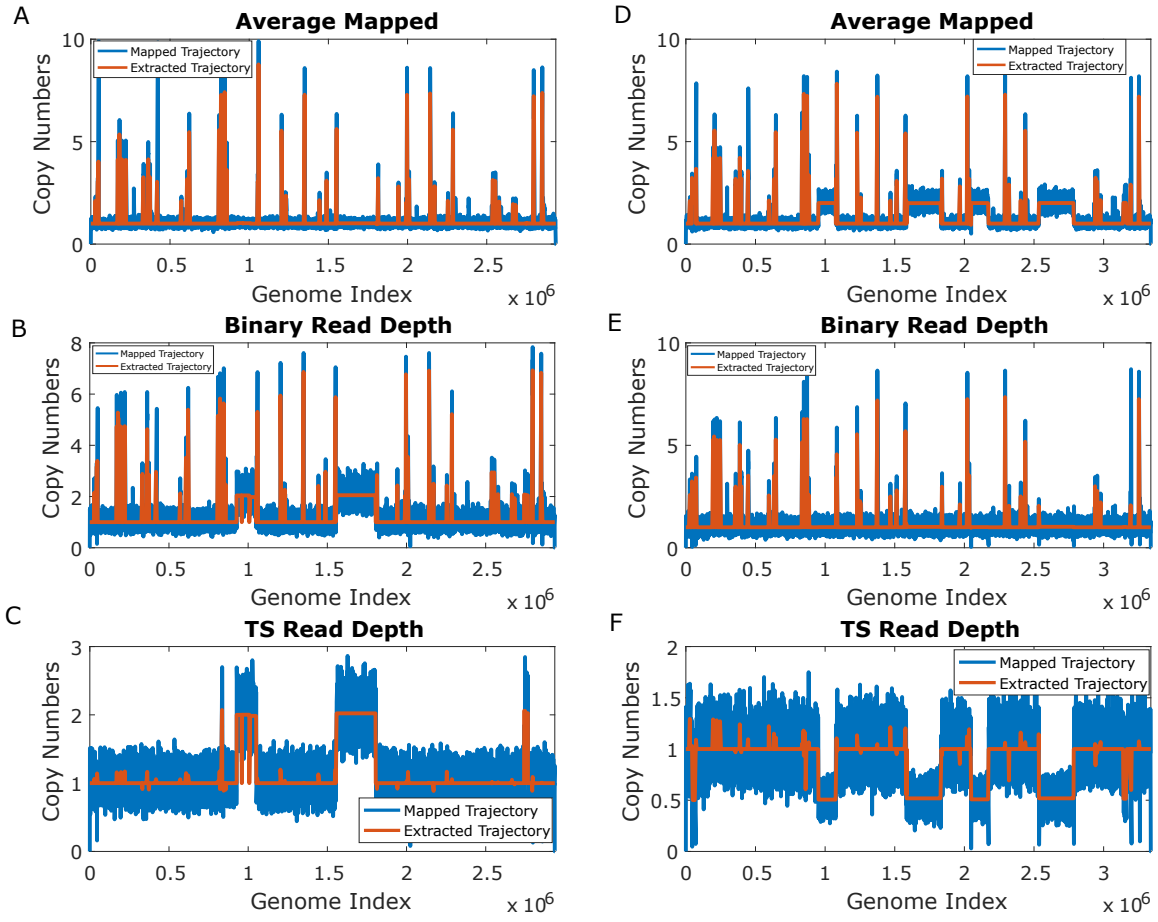

Figure E.10: The read depth trajectories reconstructed from 250-bp reads. (A) to (C) reference: original sequence. Query: Seq-2. (D) to (F) reference: Seq-2. Query: original sequence. (A) and (D) are the average number of assignments trajectories. (B) and (E) are the binary read depth trajectories. (C) and (F) are the test statistics trajectories.

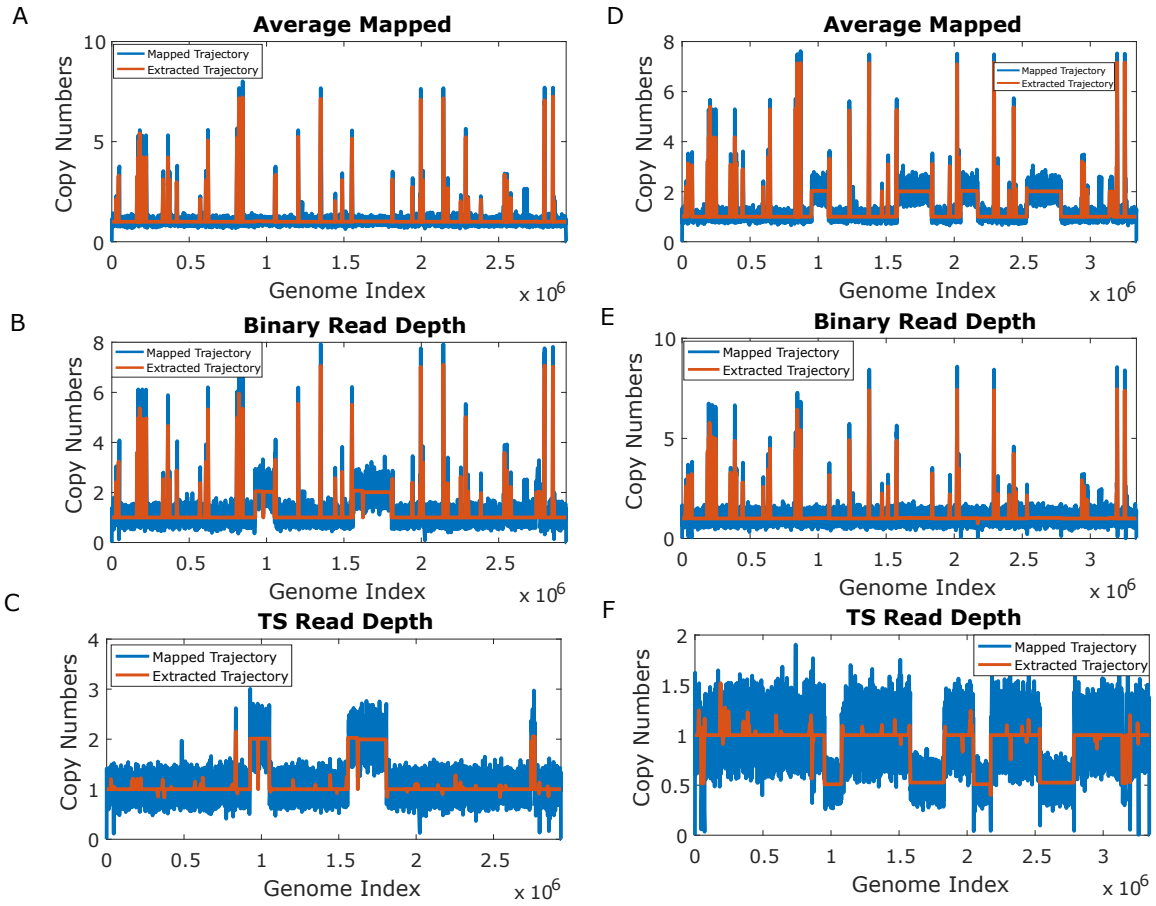

Figure E.11: The read depth trajectories reconstructed from 300-bp reads. (A) to (C) reference: original sequence. Query: Seq-2. (D) to (F) reference: Seq-2. Query: original sequence. (A) and (D) are the average number of assignments trajectories. (B) and (E) are the binary read depth trajectories. (C) and (F) are the test statistics trajectories.

E.4 CNV Detection of Different Number of Copies

In the section, the detected CNV lists from all the sequence pairs are reported. In all the tests, the grouping function is on, and the CNV regions are detected by rejecting the regions that have estimated query copy numbers within 0.9 of the estimated reference copy numbers ($|CopyNumber_{Avg} - CopyNumber_{Avg} \times CopyNumber_{TS}| >= 0.9$). To insure small differences are detected, the Poisson noise in the TS trajectory is not considered here. In all the tables, CNV size is presented in base pairs (bps) and the breakpoints are listed in the reference genome index. For group numbers, positive groups represent duplications while negative groups represent deletions. As a result, $group_{-n} \neq group_n$, in which n is the group number.

E.4.1 Query:Sequence-1

Table E.18: CNV detection for sequence-1 maps to sequence-1. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
29184	30208	2.26	3.37	1.52	14	1024
573665	574848	2.21	3.33	1.57	14	1183
2578368	2578752	24	23.1	1.06	21	384
2736896	2740224	0.975	0.705	0.738	-1	3328
1401856	1402880	0.885	0.67	0.763	-3	1024
2735104	2735616	0.96	0.787	0.831	-4	512
2324480	2324992	0.994	0.596	0.604	-5	512
2543616	2544128	1.02	0.607	0.649	-6	512
298496	299008	0.895	0.593	0.664	-7	512
866304	866816	1.08	0.625	0.587	-8	512

Table E.19: CNV detection for sequence-1 maps to sequence-2. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
28288	30208	2.2	3.35	1.53	3	1920
582144	583936	2.18	3.4	1.6	3	1792
340928	342720	3.2	4.28	1.4	6	1792
382912	384704	3.21	4.17	1.39	6	1792
1850432	1852160	3.21	4.29	1.38	6	1728
334272	335360	19.1	20.2	1.12	9	1088
353920	354880	19.8	20.9	1.18	9	960
552624	553744	19.3	20.2	1.09	9	1120
662016	663104	19	20.5	1.15	9	1088
769632	770731	19	20.2	1.15	9	1099
770048	770432	19.7	20.7	1.14	9	384
782840	783920	19.2	20.3	1.1	9	1080
885184	886112	20	21	1.1	9	928
950910	952016	19.1	20.6	1.2	9	1106
1216064	1216896	19.9	20.9	1.1	9	832
1313600	1314688	19.2	20.7	1.15	9	1088
1522944	1524000	19.3	19.8	1.11	9	1056
1780720	1781824	19.2	20.7	1.13	9	1104
1805056	1806080	19	20.1	1.09	9	1024
2124160	2125250	19.1	20.7	1.13	9	1090
2575872	2576960	19.1	20.8	1.2	9	1088
2591232	2592320	19.2	19.1	1.06	9	1088
2693568	2694672	19.1	20.1	1.12	9	1104
2914336	2915328	18.9	20.5	1.13	9	992
250112	258816	4.04	1.99	0.518	-1	8704
849664	857856	4.02	2.04	0.525	-1	8192
1259136	1267712	4.04	2.02	0.52	-1	8576
1513472	1521920	4.02	2.04	0.524	-1	8448
306688	307200	0.898	0.62	0.671	-2	512
2556928	2557440	1.01	0.645	0.65	-3	512
2794496	2795008	0.886	0.525	0.572	-4	512
553472	553984	9.94	4.02	0.508	-5	512
874496	875008	1.05	0.635	0.591	-6	512
2608512	2608896	1.08	0.42	0.409	-7	384
1408512	1409024	0.852	0.344	0.4	-8	512
1045504	1046016	1.12	0.47	0.469	-9	512
782336	782976	3.68	1.28	0.609	-10	640
928256	928768	1.11	0.469	0.413	-11	512
885888	886272	10.8	4.04	0.45	-12	384
680192	680704	0.967	0.407	0.437	-13	512
1304064	1304576	1.09	0.514	0.493	-14	512
675072	675584	0.908	0.435	0.445	-15	512
661888	662272	9.27	4.31	0.448	-16	384
1649152	1649664	0.796	0.28	0.339	-17	512
884736	885248	7.49	3.28	0.46	-18	512

Table E.20: CNV detection for sequence-1 maps to sequence-3. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
28160	30208	2.2	3.52	1.63	3	2048
571776	573696	2.2	3.49	1.64	3	1920
85728	86816	10.7	20.1	2.04	6	1088
779520	780544	10.4	19.9	2.02	6	1024
831936	833280	7.24	5.9	0.967	6	1344
855296	856864	7.09	5.6	0.994	6	1568
1368160	1369600	7.21	6.76	0.993	6	1440
1452416	1453504	10.7	20.4	2.05	6	1088
1507616	1508608	10.7	19.5	1.95	6	992
1733696	1734784	10.7	19.8	1.94	6	1088
2033152	2034432	7.3	6.82	0.992	6	1280
2178880	2181359	8.22	10.4	1.29	6	2479
2180352	2181312	11	20.5	1.97	6	960
2388736	2389760	10.5	20.5	2.2	6	1024
2576896	2577984	10.7	20.2	2	6	1088
2592320	2593280	10.8	19.8	2.04	6	960
2764960	2766051	10.7	20.2	2.01	6	1091
2836224	2837632	7.25	6.78	0.992	6	1408
2898496	2899968	7.18	6.74	0.996	6	1472
332800	334592	3.21	4.34	1.46	8	1792
373696	375552	3.19	4.26	1.45	8	1856
1850432	1852224	3.18	4.35	1.46	8	1792
695552	704512	5.9	1.95	0.345	-1	8960
845568	853760	6.03	2.04	0.347	-1	8192
1353472	1361920	6.06	2.03	0.347	-1	8448
1544960	1553664	6.05	2.01	0.356	-1	8704
1628160	1636864	6.01	1.99	0.339	-1	8704
2859520	2867968	6.02	2.05	0.363	-1	8448
1508608	1509376	3.12	1.8	0.598	-2	768
2593280	2593792	3.46	2.16	0.754	-2	512
779264	779648	2.73	1.24	0.605	-3	384
2557952	2558464	1.04	0.627	0.649	-4	512
1591296	1591808	1.13	0.524	0.49	-5	512
86656	87040	4.63	2.57	0.59	-6	384
2859008	2859519	0.93	0.725	0.752	-7	511
1039360	1039872	1.09	0.482	0.491	-8	512
1400320	1400832	0.909	0.358	0.384	-9	512
923136	923648	1.11	0.46	0.412	-10	512
2609536	2609920	1	0.378	0.405	-11	384
1288832	1289216	1.04	0.427	0.421	-12	384
669056	669440	0.952	0.353	0.355	-13	384
2795520	2795904	0.827	0.444	0.534	-14	384
1650176	1650688	0.809	0.296	0.342	-15	512

Table E.21: CNV detection for sequence-1 maps to sequence-4. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
687872	688940	4.21	19.2	4.95	1	1068
1521920	1524224	3.58	9.21	2.48	1	2304
1522945	1524224	3.24	3.77	1.25	1	1279
2588160	2589184	4.22	19.6	4.94	1	1024
2591744	2593024	3.25	4.04	1.27	1	1280
2603520	2605952	3.56	8.93	2.39	1	2432
2604545	2605824	3.28	3.84	1.23	1	1279
110336	118784	10.1	2.05	0.207	-1	8448
844288	852480	10	2.02	0.214	-1	8192
1084672	1093376	10.1	2	0.207	-1	8704
1175296	1183744	10.1	2.05	0.207	-1	8448
1456896	1465344	10.1	2.04	0.21	-1	8448
1860608	1869312	10.1	2	0.207	-1	8704
2030080	2038528	10.1	2.04	0.206	-1	8448
2641408	2650112	10.1	2	0.207	-1	8704
2732032	2740480	10.1	2.05	0.216	-1	8448
2958848	2967552	10.1	2.02	0.219	-1	8704
28672	29184	2.32	0.971	0.428	-2	512
29185	29440	2.16	1.63	0.785	-2	255
579072	580096	2.36	0.963	0.415	-2	1024
1859584	1860607	1.03	0.799	0.791	-3	1023
1597440	1597952	1.14	0.597	0.548	-5	512
306688	307200	0.89	0.614	0.672	-6	512
2569216	2569728	1	0.631	0.67	-7	512
2620672	2621184	1.05	0.481	0.505	-8	512
869376	869888	1.05	0.636	0.593	-9	512
1407488	1408000	0.849	0.328	0.395	-10	512
1038336	1038848	1.03	0.456	0.479	-11	512
922112	922624	1.08	0.464	0.436	-12	512
1304064	1304576	1.13	0.54	0.485	-13	512
2776320	2776832	0.917	0.311	0.366	-14	512
676096	676608	0.931	0.412	0.445	-15	512
670976	671488	0.912	0.445	0.471	-16	512
2822144	2822528	0.82	0.391	0.468	-17	384
1648128	1648640	0.812	0.288	0.331	-18	512

Table E.22: CNV detection for sequence-1 maps to sequence-5. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
1537280	1539584	3.16	7.18	2.32	1	2304
2628096	2629120	3.27	15.5	5.11	1	1024
2631616	2632704	3.23	3	0.953	1	1088
2643456	2645760	3.16	6.8	2.28	1	2304
2644417	2645760	3.18	2.86	0.957	1	1343
28160	30208	2.18	3.98	1.86	3	2048
595200	597248	2.17	4.12	1.94	3	2048
1999872	2001792	2.96	3.73	1.26	4	1920
2243584	2245632	2.96	3.85	1.33	4	2048
2883584	2885632	2.94	3.97	1.44	4	2048
348160	349824	3.23	4.52	1.48	6	1664
388992	390784	3.17	4.45	1.53	6	1792
1862656	1864448	3.18	4.38	1.45	6	1792
288768	297216	12.1	2.03	0.173	-1	8448
320512	328960	12.1	2.01	0.173	-1	8448
415872	424448	12.1	2.01	0.172	-1	8576
791040	799488	12.1	2.03	0.18	-1	8448
868096	876032	12	2.02	0.174	-1	7936
1014400	1022976	12.1	2.03	0.173	-1	8576
1487360	1496064	12	1.98	0.173	-1	8704
1963008	1971456	12	2.02	0.178	-1	8448
2047232	2055680	12.1	2.03	0.179	-1	8448
2176896	2185472	12.1	2.01	0.177	-1	8576
2354432	2362880	12	2.02	0.178	-1	8448
2412544	2421248	12.1	2	0.171	-1	8704
2799872	2801664	0.863	0.625	0.729	-2	1792
2609152	2609664	1.05	0.725	0.715	-3	512
2362881	2363392	0.941	0.665	0.702	-4	511
306688	307200	0.908	0.62	0.657	-5	512
892928	893440	1.03	0.588	0.583	-6	512
1422848	1423360	0.898	0.362	0.385	-7	512
1319424	1319936	1.12	0.553	0.533	-8	512
692480	692992	0.997	0.428	0.43	-9	512
2845696	2846080	0.892	0.449	0.478	-10	384
687360	687872	0.893	0.404	0.431	-11	512
1663488	1664000	0.78	0.291	0.36	-12	512

E.4.2 Query:Sequence-2

Table E.23: CNV detection for sequence-2 maps to sequence-1. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
841472	849664	2.04	4.02	2.02	1	8192
1268736	1276928	2.04	4.02	2.02	1	8192
28416	30464	1.9	2.82	1.55	4	2048
571904	573055	5.71	2.12	0.625	4	1151
573697	574976	2.12	3.54	1.73	4	1279
1966080	1968000	2.91	3.59	1.25	5	1920
2193536	2195456	2.85	3.71	1.43	5	1920
2821312	2823168	2.87	3.7	1.39	5	1856
500992	502272	23.4	17.3	0.798	-1	1280
572672	573952	23.9	15.6	0.714	-1	1280
611584	612832	22.1	17.4	0.843	-1	1248
666304	667392	24.5	16.4	0.703	-1	1088
672000	673408	21.8	15.8	0.769	-1	1408
681472	682752	22.7	17.7	0.839	-1	1280
890112	891648	22.4	14.5	0.688	-1	1536
920064	921344	22.4	17.2	0.803	-1	1280
1037312	1038848	22.5	14.8	0.69	-1	1536
1295872	1297408	22.2	15	0.747	-1	1536
1400320	1401856	22.3	15.2	0.782	-1	1536
1507584	1508672	23	17.9	0.842	-1	1088
1583104	1584512	22.9	16.4	0.768	-1	1408
1634816	1636352	21.8	14.7	0.72	-1	1536
1744128	1745536	21.5	15.5	0.769	-1	1408
1849344	1850624	23	17.7	0.835	-1	1280
2501376	2502656	22.5	17.3	0.799	-1	1280
2562560	2563648	22.8	18.5	0.864	-1	1088
2577952	2579040	23	18.2	0.89	-1	1088
2595072	2596608	21.6	14.6	0.713	-1	1536
2735488	2736768	22.3	17.6	0.847	-1	1280
2781184	2783744	11	7.38	0.73	-1	2560
1306368	1307136	1.18	0.767	0.657	-2	768
543232	543744	0.949	0.483	0.494	-3	512
1791488	1792000	1.08	0.459	0.472	-4	512
326016	326400	0.814	0.291	0.324	-5	384
1506048	1506560	0.916	0.332	0.375	-6	512

Table E.24: CNV detection for sequence-2 maps to sequence-2. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
334272	335360	19	19	1.09	9	1088
353856	354880	19.4	19.5	1.11	9	1024
552960	553728	20	19.8	1.03	9	768
662016	663104	18.9	19.2	1.06	9	1088
769664	770736	19.1	19.1	1.1	9	1072
782912	783920	19.4	19.3	1.04	9	1008
885184	886096	19.9	20.2	1.05	9	912
950976	952016	19.4	19.5	1.1	9	1040
1215904	1217024	19.1	19.4	1.04	9	1120
1313664	1314560	19.3	19.4	1.05	9	896
1522944	1524000	19.1	18.6	1.04	9	1056
1780928	1781824	19.9	20.2	1.05	9	896
1805056	1806144	18.9	18.9	1.05	9	1088
2124160	2125248	19	19.4	1.05	9	1088
2575872	2576960	19	19.3	1.07	9	1088
2591232	2592320	19	18	0.979	9	1088
2693568	2694672	18.9	18.9	1.09	9	1104
2914336	2915424	19	19.4	1.05	9	1088
552640	552959	13.7	9.88	1.15	21	319

Table E.25: CNV detection for sequence-2 maps to sequence-3. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
28416	30208	2.12	3.3	1.57	4	1792
571904	573696	2.13	3.2	1.56	4	1792
332800	334592	3.18	4.23	1.4	6	1792
373760	375552	3.16	4.26	1.43	6	1792
1850432	1852288	3.12	4.26	1.46	6	1856
85728	86816	10.6	17.4	1.76	9	1088
779536	780672	10.5	17.5	1.79	9	1136
1452429	1453504	10.6	17.8	1.78	9	1075
1507648	1508672	10.7	17.3	1.69	9	1024
1733696	1734784	10.6	17.2	1.73	9	1088
2180544	2181376	11.1	18.1	1.7	9	832
2388736	2389856	10.5	17.9	1.86	9	1120
2576896	2577984	10.6	17.5	1.77	9	1088
2592288	2593344	10.6	17.3	1.77	9	1056
2764960	2766080	10.5	17.5	1.75	9	1120
164352	165120	2.93	3.79	1.34	15	768
695552	704000	6.09	3.98	0.686	-1	8448
845568	853760	6.08	4	0.675	-1	8192
1353216	1361920	6.01	3.88	0.684	-1	8704
1544960	1553408	6.07	3.98	0.709	-1	8448
1628416	1636864	6.08	3.98	0.672	-1	8448
2859520	2868224	5.95	3.87	0.689	-1	8704
2388480	2388864	2.95	1.23	0.541	-2	384
1452032	1452544	2.44	1.09	0.538	-3	512
779264	779648	2.48	1.04	0.561	-4	384
543232	543744	0.933	0.479	0.493	-5	512
2181248	2181632	3.27	1.24	0.509	-6	384
327040	327424	0.782	0.29	0.317	-7	384
778496	779008	0.917	0.395	0.457	-8	512
1506048	1506560	0.845	0.332	0.384	-9	512

Table E.26: CNV detection for sequence-2 maps to sequence-4. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
687872	688928	4.12	16.7	4.34	2	1056
1521920	1524224	3.44	8.29	2.27	2	2304
1522945	1524224	3.05	3.8	1.3	2	1279
2588160	2589184	4.12	17	4.39	2	1024
2589185	2592896	2.27	2.3	1.01	2	3711
2603520	2605824	3.45	8.33	2.27	2	2304
2604545	2605824	3.04	3.82	1.25	2	1279
110336	118784	10.2	4.02	0.408	-1	8448
844544	852672	10.2	4.05	0.409	-1	8128
1084928	1093632	9.9	3.87	0.416	-1	8704
1175168	1183744	10.2	3.99	0.404	-1	8576
1456896	1465344	10.1	3.98	0.418	-1	8448
1860608	1869312	10.1	3.93	0.404	-1	8704
2030080	2038784	10.1	3.95	0.402	-1	8704
2641664	2650112	10.1	3.98	0.409	-1	8448
2732032	2740480	10.1	4.01	0.436	-1	8448
2958848	2967424	10.1	3.98	0.429	-1	8576
28672	29440	1.92	0.936	0.495	-2	768
578560	580096	2.04	0.926	0.488	-2	1536
879744	880640	0.98	0.607	0.616	-3	896
1520384	1521152	0.82	0.35	0.414	-4	768
1519616	1520128	1.06	0.837	0.814	-5	512
550400	550912	0.947	0.505	0.504	-6	512
2639360	2639872	1.22	0.706	0.613	-7	512
1172992	1173504	1.09	0.63	0.624	-8	512
334208	334592	0.78	0.294	0.326	-9	384
1802752	1803264	1.05	0.445	0.457	-10	512
778624	779008	0.936	0.298	0.343	-11	384

Table E.27: CNV detection for sequence-2 maps to sequence-5. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
1537280	1539584	3.03	6.33	2.06	3	2304
2628096	2629120	3.22	13.6	4.48	3	1024
2629248	2632768	2.27	1.99	0.91	3	3520
2643456	2645760	3.04	5.96	1.98	3	2304
2644417	2645760	3.02	2.47	0.854	3	1343
28416	30208	2.09	4.09	1.93	4	1792
595456	596992	2.05	2.78	1.39	4	1536
596736	597248	2.14	9.3	4.52	4	512
1999872	2001792	2.95	3.78	1.34	5	1920
2243584	2245632	2.96	3.85	1.36	5	2048
2883648	2885632	2.92	3.88	1.46	5	1984
348160	349824	3.08	3.84	1.29	7	1664
389120	390912	3.05	3.87	1.37	7	1792
1862656	1864448	3.05	3.77	1.28	7	1792
288768	297472	12.1	3.92	0.337	-1	8704
320512	329216	12.1	3.9	0.34	-1	8704
415744	424448	12	3.88	0.34	-1	8704
791040	799488	12.1	3.99	0.36	-1	8448
868032	876160	12.1	4.02	0.345	-1	8128
1014528	1022976	12.1	3.98	0.341	-1	8448
1487616	1496064	12.1	3.97	0.345	-1	8448
1963008	1971456	12.1	3.98	0.348	-1	8448
2047232	2055680	12.1	3.99	0.355	-1	8448
2177008	2185232	12.1	4.01	0.356	-1	8224
2354176	2362880	12.1	3.9	0.344	-1	8704
2412544	2421248	12.1	3.93	0.338	-1	8704
1972224	1973760	0.912	0.638	0.679	-2	1536
1538304	1539072	3.01	2.03	0.707	-3	768
2644480	2644992	2.85	2.01	0.74	-3	512
903296	904192	0.962	0.61	0.63	-4	896
1535744	1536512	0.929	0.419	0.485	-5	768
2055681	2056192	0.966	0.62	0.71	-7	511
1783808	1784320	1.04	0.628	0.597	-8	512
1818112	1818624	1.03	0.413	0.453	-9	512
802048	802560	0.999	0.393	0.395	-10	512
675200	675584	0.846	0.284	0.316	-11	384

E.4.3 Query:Sequence-3

Table E.28: CNV detection for sequence-3 maps to sequence-1. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
841472	849408	2.09	6.34	3.08	1	7936
1268736	1276928	2.09	6.36	3.09	1	8192
1966080	1968000	2.9	3.63	1.27	4	1920
2193536	2195456	2.86	3.78	1.4	4	1920
2821312	2823168	2.83	3.77	1.4	4	1856
2781184	2783744	11.1	3.88	0.521	-1	2560
500992	502272	23.8	8.99	0.412	-2	1280
572416	573952	22.6	7.3	0.356	-2	1536
611584	612864	22.1	8.73	0.436	-2	1280
665600	667392	18	5.89	0.379	-2	1792
672128	673536	23.1	8.31	0.396	-2	1408
681472	682752	23	8.97	0.419	-2	1280
889856	891392	19	6.51	0.354	-2	1536
920064	921344	22.6	8.54	0.379	-2	1280
1037312	1038720	22.2	7.94	0.361	-2	1408
1295872	1297408	22.4	7.48	0.366	-2	1536
1400320	1401856	22.4	7.66	0.391	-2	1536
1507648	1508480	22.8	8.79	0.475	-2	832
1583104	1584384	22.7	8.83	0.405	-2	1280
1634816	1636352	21.7	7.42	0.368	-2	1536
1744384	1745536	23.3	9.66	0.446	-2	1152
1849344	1850624	23.3	8.97	0.426	-2	1280
2501376	2502656	22.7	8.62	0.388	-2	1280
2562688	2563648	24.3	9.85	0.419	-2	960
2578048	2578816	23	9.18	0.465	-2	768
2595328	2596608	23.6	8.96	0.401	-2	1280
2735616	2736896	23.9	8.94	0.417	-2	1280
1267712	1268736	1.04	0.721	0.693	-3	1024
15360	15872	1.05	0.634	0.622	-4	512
2752512	2753024	0.858	0.411	0.463	-5	512
85504	86016	0.912	0.376	0.404	-6	512
776192	776704	0.935	0.366	0.402	-7	512
1544960	1545472	1.03	0.506	0.487	-8	512
2374272	2374656	0.82	0.363	0.43	-9	384

Table E.29: CNV detection for sequence-3 maps to sequence-2. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
250368	258560	4.16	6.38	1.6	1	8192
849664	857856	4.16	6.43	1.58	1	8192
1259264	1267456	4.16	6.38	1.6	1	8192
1513472	1521664	4.16	6.4	1.59	1	8192
334080	335872	15.2	5.94	0.448	-1	1792
353280	355072	15	6.01	0.421	-1	1792
552448	553984	18.8	7.33	0.411	-1	1536
661888	663296	18.9	8.31	0.473	-1	1408
769536	771072	17.9	7.15	0.417	-1	1536
782592	784896	11.4	4.31	0.449	-1	2304
884736	886272	18.3	7.4	0.406	-1	1536
950784	952064	18.8	8.77	0.495	-1	1280
1215744	1217024	18.4	8.62	0.48	-1	1280
1313408	1314816	18.5	8.28	0.473	-1	1408
1522944	1525248	9.8	5.15	0.791	-1	2304
1780736	1782016	20	8.81	0.476	-1	1280
1804800	1806336	18.3	7.25	0.408	-1	1536
2124032	2125312	18.7	8.75	0.478	-1	1280
2575872	2576960	19.4	9.6	0.549	-1	1088
2579456	2580736	3.23	2.85	0.912	-1	1280
2591232	2593664	9.59	4.92	0.764	-1	2432
2693376	2694912	18.8	7.42	0.424	-1	1536
2914304	2915584	19.6	8.97	0.481	-1	1280
1521664	1522176	1.11	0.705	0.678	-2	512
15360	15872	1.05	0.635	0.621	-3	512
1259008	1259392	2.13	1.03	0.56	-4	384
85504	86016	0.913	0.376	0.404	-5	512
250112	250496	2.15	0.799	0.413	-6	384
1635328	1635840	0.872	0.414	0.433	-7	512
2764800	2765184	0.831	0.288	0.327	-8	384

Table E.30: CNV detection for sequence-3 maps to sequence-3. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
332800	334592	3.16	3.89	1.31	7	1792
373696	375552	3.14	3.89	1.31	7	1856
1850432	1852288	3.15	3.9	1.32	7	1856

Table E.31: CNV detection for sequence-3 maps to sequence-4. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
687872	688928	4.2	8.58	2.1	3	1056
1521920	1524224	3.48	5.62	1.69	3	2304
2588160	2589184	4.18	8.9	2.28	3	1024
2591744	2593024	3.23	4.04	1.3	3	1280
2604545	2605952	3.18	3.93	1.27	3	1407
830912	832448	6.94	6.38	1.07	6	1536
854304	855840	7.02	5.77	0.973	6	1536
1375360	1376768	6.99	7.95	1.23	6	1408
2046336	2047872	6.93	7.78	1.23	6	1536
2192192	2193664	6.97	7.91	1.21	6	1472
2862848	2864128	7.07	7.87	1.17	6	1280
2916992	2918400	7.01	7.85	1.34	6	1408
2603520	2604544	4.16	8.28	2.09	10	1024
110336	118784	10.4	6.39	0.623	-1	8448
844528	852480	10.4	6.43	0.626	-1	7952
1084416	1093376	10.1	6.01	0.62	-1	8960
1175040	1183616	10.4	6.3	0.614	-1	8576
1456640	1465344	10.4	6.19	0.626	-1	8704
1860608	1869312	10.4	6.22	0.622	-1	8704
2030080	2038656	10.4	6.34	0.621	-1	8576
2641664	2650112	10.4	6.31	0.63	-1	8448
2732032	2740480	10.4	6.35	0.658	-1	8448
2958848	2967296	10.4	6.36	0.666	-1	8448
28672	29184	2.15	1.19	0.574	-2	512
85504	86016	0.906	0.383	0.415	-3	512
1634304	1634816	0.793	0.399	0.451	-4	512
779264	779776	0.953	0.347	0.374	-5	512

Table E.32: CNV detection for sequence-3 maps to sequence-5. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
1537408	1539584	3.13	4.35	1.4	3	2176
2628096	2629120	3.4	7.73	2.37	3	1024
2631680	2632832	3.29	3.23	1.02	3	1152
2643456	2645760	3.12	4.39	1.45	3	2304
1999872	2001728	2.97	3.76	1.27	4	1856
2243648	2245632	2.93	3.93	1.41	4	1984
2883584	2885632	2.94	3.98	1.45	4	2048
348032	349824	3.13	3.89	1.27	7	1792
388992	390784	3.12	3.87	1.33	7	1792
1862656	1864448	3.11	3.86	1.29	7	1792
28672	30208	2.15	3.14	1.55	8	1536
595456	597248	2.16	2.94	1.45	8	1792
288768	297472	12.4	6.24	0.518	-1	8704
319488	329216	11.2	5.55	0.527	-1	9728
415744	424704	12.2	6	0.505	-1	8960
791040	799616	12.4	6.31	0.534	-1	8576
868032	876032	12.4	6.4	0.528	-1	8000
1014272	1022976	12.4	6.25	0.518	-1	8704
1487616	1496064	12.5	6.28	0.524	-1	8448
1963008	1971456	12.4	6.31	0.534	-1	8448
2046976	2056192	11.9	5.87	0.517	-1	9216
2176768	2185344	12.4	6.27	0.531	-1	8576
2354176	2362880	12.4	6.18	0.526	-1	8704
2412544	2421248	12.4	6.24	0.517	-1	8704
15360	15872	1.06	0.636	0.622	-2	512
85504	86016	0.898	0.348	0.386	-3	512
1649664	1650176	0.872	0.405	0.441	-4	512
802816	803328	0.869	0.341	0.372	-5	512

E.4.4 Query:Sequence-4

Table E.33: CNV detection for sequence-4 maps to sequence-1. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
841472	849408	2.07	10.7	5.23	1	7936
1268736	1276928	2.07	10.7	5.25	1	8192
827872	829376	7.14	6.65	1.07	7	1504
851264	852736	7.13	6.21	1.06	7	1472
1368160	1369600	7.02	7.92	1.2	7	1440
2019808	2021312	7.03	7.83	1.26	7	1504
2165600	2167168	7.04	7.88	1.19	7	1568
2823936	2825472	6.98	7.83	1.29	7	1536
2878080	2879488	7.01	7.94	1.24	7	1408
500736	502272	22.2	3.58	0.171	-2	1536
572416	573824	22.2	3.86	0.195	-2	1408
611328	612864	19.2	3.25	0.199	-2	1536
666112	667648	19.1	3.04	0.195	-2	1536
672000	673536	22.2	3.63	0.17	-2	1536
681472	682752	22.8	4.27	0.193	-2	1280
889856	891904	16.4	2.44	0.229	-2	2048
920064	921344	22.6	4.12	0.189	-2	1280
1037312	1038592	22	3.85	0.176	-2	1280
1296000	1297408	22.9	3.88	0.173	-2	1408
1400320	1401856	22.3	3.58	0.177	-2	1536
1583104	1584384	22.6	4.26	0.192	-2	1280
1634816	1636352	21.9	3.52	0.169	-2	1536
1743872	1745920	16.4	2.45	0.212	-2	2048
1849344	1850752	23.1	3.77	0.181	-2	1408
2501376	2502656	22.6	4.19	0.193	-2	1280
2562560	2563584	22.9	4.34	0.195	-2	1024
2595328	2596864	19.8	3.3	0.201	-2	1536
2735488	2737152	19.3	3.09	0.187	-2	1664
2782208	2783744	22.2	3.57	0.162	-2	1536
2699264	2699776	1	0.452	0.467	-3	512
2919936	2920448	0.858	0.382	0.411	-4	512
2011648	2012160	1.06	0.487	0.435	-5	512
665600	666111	0.997	0.735	0.728	-6	511
685568	686080	0.849	0.383	0.444	-7	512

Table E.34: CNV detection for sequence-4 maps to sequence-2. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
250368	258560	4.09	10.6	2.69	1	8192
849664	857856	4.1	10.7	2.67	1	8192
1259232	1267456	4.09	10.7	2.73	1	8224
1513472	1521696	4.1	10.7	2.68	1	8224
836096	837568	7.19	6.72	1.04	9	1472
859392	860960	7.05	6.27	1.11	9	1568
1376384	1377792	7.11	7.98	1.26	9	1408
2033152	2034624	7.15	7.94	1.18	9	1472
2179968	2181440	7.14	7.93	1.18	9	1472
2835200	2836736	7.09	7.92	1.23	9	1536
2889344	2890752	7.02	7.92	1.34	9	1408
1955840	1956864	0.987	0.599	0.59	-2	1024
334080	335872	15	2.76	0.224	-3	1792
353790	355072	19	3.94	0.21	-3	1282
552448	553728	18.3	3.85	0.218	-3	1280
662016	663552	15.8	2.96	0.208	-3	1536
769536	770816	18.6	4.05	0.239	-3	1280
782848	784128	19.1	3.82	0.214	-3	1280
884736	886272	18.2	3.35	0.186	-3	1536
950784	952064	18.5	4.09	0.217	-3	1280
1215488	1217280	16.1	2.82	0.187	-3	1792
1313536	1314816	18.8	4.04	0.22	-3	1280
1523200	1523712	18.7	4.29	0.257	-3	512
1780480	1782016	18.3	3.45	0.19	-3	1536
1805056	1806336	19	3.88	0.211	-3	1280
2124032	2125312	18.6	4.01	0.215	-3	1280
2576000	2576896	19.9	4.35	0.213	-3	896
2591232	2592256	18.6	4.2	0.248	-3	1024
2693376	2694912	18.4	3.43	0.192	-3	1536
2914304	2915584	19	3.99	0.21	-3	1280
2189312	2190336	0.882	0.541	0.651	-4	1024
335360	335872	0.913	0.62	0.652	-5	512
258560	259072	1.21	0.718	0.594	-6	512
2712576	2713088	0.974	0.453	0.468	-7	512
2932224	2932736	0.855	0.375	0.423	-8	512
2024960	2025472	0.999	0.462	0.444	-9	512
691712	692224	0.853	0.412	0.451	-10	512

Table E.35: CNV detection for sequence-4 maps to sequence-3. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
695696	703936	6.15	10.6	1.79	1	8240
845536	853760	6.15	10.6	1.75	1	8224
1353600	1361872	6.14	10.6	1.82	1	8272
1545168	1553408	6.13	10.6	1.85	1	8240
1628512	1636752	6.16	10.6	1.75	1	8240
2859584	2867808	6.13	10.7	1.85	1	8224
332800	334592	3.15	4.41	1.45	6	1792
373760	375552	3.15	4.39	1.47	6	1792
1850432	1852288	3.13	4.33	1.44	6	1856
181312	182272	5.59	5.36	1.14	8	960
621696	623168	5.21	5.69	1.36	8	1472
819200	820736	5.38	5	1.09	8	1536
1214016	1215488	5.39	5.64	1.19	8	1472
1579072	1580544	5.39	5.55	1.17	8	1472
2324480	2326016	5.29	5	1.09	8	1536
85504	87040	10.2	3.33	0.34	-1	1536
779264	780800	10	3.42	0.351	-1	1536
831936	833280	7.1	6.34	1.02	-1	1344
855328	856864	7.01	5.99	1.09	-1	1536
1368160	1369600	7.08	7.19	1.08	-1	1440
1452032	1453568	8.84	3.03	0.351	-1	1536
1507840	1508352	10.1	4.09	0.461	-1	512
1733632	1734912	10.5	3.94	0.388	-1	1280
2033152	2034432	7.14	7.32	1.1	-1	1280
2178880	2181632	8.5	4.88	0.688	-1	2752
2388480	2390016	10.1	3.36	0.345	-1	1536
2576896	2577920	10.6	4.06	0.381	-1	1024
2764800	2766336	10.2	3.35	0.337	-1	1536
2836224	2837504	7.05	7.26	1.13	-1	1280
2898528	2899968	7.03	7.19	1.11	-1	1440
372736	373760	1.1	0.697	0.61	-2	1024
1628160	1628672	2.49	1.19	0.539	-3	512
2712576	2713088	0.981	0.444	0.464	-4	512
2940416	2940928	0.92	0.386	0.417	-5	512
2024960	2025472	1.02	0.484	0.446	-6	512
680448	680960	0.861	0.388	0.449	-7	512

Table E.36: CNV detection for sequence-4 maps to sequence-4. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
830944	832448	7.14	7.2	1.11	4	1504
854304	855840	7.12	6.62	1.12	4	1536
1375296	1376768	6.99	8.53	1.31	4	1472
2046464	2047872	7.06	8.68	1.35	4	1408
2192192	2193664	7.04	8.57	1.29	4	1472
2862848	2864128	7.05	8.81	1.32	4	1280
2916992	2918400	7.02	8.6	1.41	4	1408
384512	385024	0.771	0.549	0.714	-1	512

Table E.37: CNV detection for sequence-4 maps to sequence-5. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
1999872	2001792	2.88	3.63	1.31	4	1920
2243648	2245632	2.86	3.76	1.36	4	1984
2883584	2885632	2.87	3.74	1.39	4	2048
348160	349824	3.16	3.9	1.28	7	1664
389120	390784	3.12	3.92	1.35	7	1664
1862720	1864448	3.14	3.92	1.31	7	1728
854432	855968	7.12	6.29	1.04	10	1536
877824	879360	7.14	6.05	1.04	10	1536
1390720	1392128	7.12	7.62	1.19	10	1408
2061696	2063232	7.05	7.46	1.2	10	1536
2215680	2217248	7.06	7.48	1.14	10	1568
2886272	2887680	7.02	7.46	1.16	10	1408
2940416	2941984	7.02	7.47	1.18	10	1568
2045440	2045952	1.01	0.482	0.468	-2	512
704000	704512	0.849	0.358	0.421	-3	512

E.4.5 Query:Sequence-5

Table E.38: CNV detection for sequence-5 maps to sequence-1. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
841472	849408	2.07	12.6	6.18	1	7936
1268736	1276928	2.07	12.6	6.19	1	8192
827840	829184	6.89	6.34	1.06	7	1344
851232	852736	7	5.86	1.01	7	1504
1368160	1369600	6.97	7.49	1.15	7	1440
2019776	2021120	6.85	7.48	1.2	7	1344
2165600	2167168	7.03	7.44	1.13	7	1568
2823936	2825472	7	7.44	1.16	7	1536
2878080	2879488	6.9	7.48	1.21	7	1408
1507584	1509952	10.2	2.66	0.6	-1	2368
2563840	2567424	2.35	2.07	0.926	-1	3584
2572288	2574336	2.18	2.05	0.977	-1	2048
2578112	2580480	8.63	2.63	0.666	-1	2368
500992	502272	22.3	2.98	0.148	-2	1280
572544	573696	21.8	2.98	0.143	-2	1152
611584	612864	21.2	2.98	0.15	-2	1280
665984	667392	21.8	2.71	0.131	-2	1408
672128	673792	17.9	2.21	0.187	-2	1664
681472	682752	21.6	2.98	0.15	-2	1280
889856	891392	18	2.26	0.154	-2	1536
920064	921344	21.5	2.85	0.134	-2	1280
1037312	1038592	20.8	2.67	0.132	-2	1280
1295872	1297408	21.4	2.48	0.126	-2	1536
1400320	1401856	21.2	2.58	0.132	-2	1536
1583104	1584384	21.6	2.89	0.138	-2	1280
1635072	1636352	21.9	3.04	0.15	-2	1280
1744128	1745664	21	2.51	0.131	-2	1536
1849344	1850624	22	3.02	0.144	-2	1280
2501376	2502656	21.5	2.89	0.138	-2	1280
2562752	2563584	23	3.44	0.159	-2	832
2595328	2596608	22.1	2.91	0.138	-2	1280
2735616	2736896	22.1	3.02	0.148	-2	1280
2782208	2783744	21.2	2.39	0.114	-2	1536
1584640	1585152	1.07	0.774	0.762	-3	512
772608	773120	0.973	0.48	0.493	-4	512
288768	289280	1.11	0.463	0.41	-5	512
2962944	2963456	0.897	0.452	0.49	-6	512

Table E.39: CNV detection for sequence-5 maps to sequence-2. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
250336	258560	4.09	12.6	3.19	1	8224
849664	857856	4.09	12.6	3.19	1	8192
1259232	1267456	4.09	12.6	3.22	1	8224
1513472	1521664	4.1	12.6	3.19	1	8192
1522688	1525248	9.03	2.61	0.635	-1	2560
2577152	2580736	2.35	2.18	0.962	-1	3584
2585344	2587392	2.12	2.07	1.01	-1	2048
2591232	2593792	8.29	2.7	0.65	-1	2560
334208	335360	18.1	2.91	0.178	-2	1152
353664	354944	17.8	2.83	0.172	-2	1280
552448	553728	17.5	2.72	0.16	-2	1280
661888	663296	18.1	2.66	0.158	-2	1408
769536	770816	18	2.76	0.161	-2	1280
782720	784128	18.2	2.65	0.15	-2	1408
884992	886272	18.6	2.86	0.158	-2	1280
950784	952064	17.9	2.82	0.171	-2	1280
1215744	1217024	17.6	2.83	0.16	-2	1280
1313536	1314816	18.3	2.89	0.17	-2	1280
1780736	1781800	18.3	3.18	0.178	-2	1064
1804800	1806336	17.7	2.37	0.135	-2	1536
2124032	2125312	17.8	2.86	0.162	-2	1280
2576000	2576896	18.8	3.35	0.194	-2	896
2693376	2694656	17.5	2.58	0.161	-2	1280
2914304	2915584	18.4	2.91	0.162	-2	1280
876544	877056	1.01	0.555	0.53	-3	512
296960	297472	1.11	0.474	0.426	-4	512
1473024	1473536	0.844	0.428	0.515	-5	512
2368512	2369024	0.963	0.39	0.39	-6	512
998016	998400	1.06	0.358	0.347	-7	384

Table E.40: CNV detection for sequence-5 maps to sequence-3. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
695712	703967	6.14	12.5	2.11	1	8255
845568	853760	6.15	12.6	2.08	1	8192
1353616	1361856	6.14	12.6	2.14	1	8240
1545168	1553408	6.12	12.6	2.21	1	8240
1628512	1636752	6.15	12.6	2.08	1	8240
2859584	2867808	6.12	12.6	2.17	1	8224
28288	30208	2.16	2.85	1.39	4	1920
571776	573440	2.11	2.94	1.46	4	1664
332800	334592	3.18	4.3	1.4	6	1792
373760	375552	3.19	4.28	1.4	6	1792
1850432	1852288	3.16	4.2	1.35	6	1856
85728	86912	10.3	2.89	0.294	-1	1184
779392	780800	9.98	2.58	0.274	-1	1408
832000	833280	7	5.79	0.99	-1	1280
855360	856704	6.87	5.3	1.06	-1	1344
1368128	1369600	7.08	6.41	0.958	-1	1472
1452288	1453568	9.83	2.82	0.31	-1	1280
1507840	1509376	6.24	2.3	0.501	-1	1536
1733632	1734912	10.2	2.73	0.285	-1	1280
2033152	2034432	7.16	6.42	0.948	-1	1280
2178880	2181632	8.33	4.09	0.627	-1	2752
2388480	2390016	9.66	2.39	0.259	-1	1536
2576896	2577920	10.1	3.03	0.321	-1	1024
2764800	2766336	9.79	2.37	0.251	-1	1536
2836224	2837504	7.06	6.41	0.964	-1	1280
2898560	2899968	7.15	6.45	0.953	-1	1408
2867712	2868224	2	0.947	0.633	-2	512
2983424	2983936	0.97	0.418	0.38	-3	512

Table E.41: CNV detection for sequence-5 maps to sequence-4. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
110464	118720	10.3	12.7	1.25	1	8256
844528	852736	10.3	12.7	1.25	1	8208
1084928	1093200	10.2	12.6	1.29	1	8272
1175328	1183584	10.3	12.7	1.25	1	8256
1456896	1465168	10.2	12.6	1.29	1	8272
1860832	1869072	10.3	12.6	1.27	1	8240
2030264	2038528	10.3	12.6	1.25	1	8264
2641664	2649904	10.3	12.6	1.27	1	8240
2732160	2740400	10.2	12.6	1.33	1	8240
2959040	2967296	10.2	12.6	1.33	1	8256
830912	832256	6.96	6.54	1.04	5	1344
854272	855840	6.99	6.13	1.03	5	1568
1375360	1376768	7.04	8.04	1.21	5	1408
2046400	2047744	6.97	7.94	1.24	5	1344
2192160	2193664	6.99	7.87	1.18	5	1504
2862848	2864128	7.01	7.94	1.21	5	1280
2916928	2918400	6.98	7.85	1.23	5	1472
579072	579584	2.23	1.34	0.611	-1	512
2967040	2967552	4.59	2.33	0.567	-2	512
1868928	1869312	3.84	1.31	0.469	-3	384
296960	297472	1.09	0.466	0.419	-4	512
688640	689152	2.71	1.29	0.486	-5	512
1175168	1175552	5.93	2.44	0.476	-6	384
1480192	1480704	0.884	0.456	0.504	-7	512
2380800	2381312	0.983	0.361	0.377	-8	512
1464960	1465344	5.38	2.12	0.389	-9	384
687616	688128	2.15	0.756	0.397	-10	512
2641408	2641792	3.7	1.08	0.541	-11	384
3010048	3010496	0.851	0.368	0.402	-12	448
1456640	1456960	1.68	0.544	0.558	-13	320

Table E.42: CNV detection for sequence-5 maps to sequence-5. CNV size is in base pairs (bps). Positive groups represent duplications. Negative groups represent deletions.

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
1999872	2001792	2.97	3.46	1.2	4	1920
2243648	2245632	2.93	3.57	1.33	4	1984
2883648	2885632	2.9	3.63	1.38	4	1984
348160	349952	3.14	3.9	1.28	6	1792
389120	390784	3.1	4.01	1.38	6	1664
1862720	1864512	3.11	3.96	1.33	6	1792
894976	895488	1.03	0.549	0.524	-1	512
3025408	3025856	0.813	0.34	0.332	-2	448

E.5 Comparing wiht CNVnator

In the section, the detected CNV list from CNVnator is reported.

Table E.43: CNVs report from CNV-MM The table is taken directly from the CNVnator output except the first column indicating all CNVs are duplications is deleted. Coordinate: CNV breakpoints. RD: read depth. P-val: p-value. q0: Quality score. For details of the column, please see CNVnator[42].

Coordinate	CNV-Size	RD	P-val1	P-val2	P-val3	P-val4	q0
28301-30450	2150	2.23555	0	2.23253e-29	497428	17.3433	1
45951-47400	1450	2.68503	0.000626263	7.65628e+06	1	1	1
177551-178300	750	5.32745	2.7384e-05	8.95024e-191	1	1	1
180351-181250	900	4.51578	1.68581e-06	6.4933e-145	1	1	1
417701-423050	5350	2.81295	0	9701.95	0	1.07686e+06	1
565351-568850	3500	1.69867	0	2.4299e-05	0.000121186	2672.92	1
570651-572900	2250	2.18276	0	1.29493e-15	123129	541.452	1
809051-810400	1350	4.29401	1.88886e-09	3.38017e-154	1	1	1
827751-828800	1050	2.12186	0.00024832	1.7156e-06	1	1	1
1053451-1059950	6500	3.50688	0	8.32469e+06	0	5.0256e+07	1
1203551-1205250	1700	3.81133	1.17778e-05	3.35185e+08	1	1	1
1223901-1224600	700	1.95137	0.0218964	4.29084e-08	1	1	1
1228551-1232100	3550	1.63153	1.97532e-09	198.572	0.0420266	2.14677e+06	1
1443151-1444800	1650	2.194	0	4.02213e-09	1	1	1
1488201-1490550	2350	2.92859	0	8.47598e+08	9576.86	9.76731e-05	1
1553201-1554600	1400	4.41865	1.36605e-09	4.39256e-16	1	1	1
1815351-1817050	1700	4.21932	0	8.2999e-147	1	1	1
1944301-1946150	1850	3.06671	0	1.59164e-88	1	1	1
1994601-1996400	1800	2.11166	0	2.77406e-12	1	1	1
2002151-2003800	1650	2.05565	8.46123e-08	666.624	1	1	1
2011051-2012900	1850	2.50232	0	2.87152e-16	1	1	1
2057851-2059800	1950	1.77522	0.00491184	0.0724911	1	1	1
2171551-2173600	2050	2.91904	0	3.34486e-15	1	1	1
2255951-2257700	1750	2.13136	0	2.51779e-20	1	1	1
2288351-2289750	1400	4.0658	0	7.22966e-59	1	1	1
2303301-2305200	1900	2.11804	0	3.64459e-27	1	1	1
2384151-2388900	4750	1.80368	0	4.01408e+07	0	1.08221e+07	1
2539801-2540750	950	2.7549	4.86506e-08	1.3771e-64	1	1	1
2540901-2544500	3600	2.47723	0	2.68849e-27	5.84439e-06	2.7884e-07	1
2549101-2551350	2250	2.15837	0	8.71375e-14	7574.44	0.688413	1
2555201-2557500	2300	2.95321	0	5.50771e+06	4719.09	8.11376e-13	1
2579201-2581000	1800	2.12205	0	2.77825e-09	1	1	1
2683551-2684600	1050	2.29531	0.00650178	5643.56	1	1	1
2795301-2797400	2100	2.91666	0	1.14939e-10	1.62959e+06	8.91547e-15	1
2798051-2799450	1400	7.81974	0	3.34122e-236	1	1	1
2856401-2869450	13050	1.69711	0	1.40966e-11	0	1.82638e-08	1

E.6 CNV Detection of *A. baumannii* Clinical Isolate

In the section, the detected CNV lists of real short read data SRR25588867 using either *A. baumannii* strain ATCC 17978 or strain MDR-ZJ06 is reported. In all the tests, the grouping function is on, and the CNV regions are detected by rejecting the regions that have estimated query copy numbers within 0.9 of the estimated reference copy numbers ($|CopyNumber_{Avg} - CopyNumber_{Avg} \times CopyNumber_{TS}| >= 0.9$). To insure small differences are detected, the Poisson noise in the TS trajectory is not considered here. In all the tables, CNV size is presented in base pairs (bps) and the breakpoints are listed in the reference genome index. Deletions and duplications are listed in separated tables.

E.6.1 Reference: *A. baumannii* strain ATCC 17978

Table E.44: Duplication in SRR2558867 when using *A. baumannii* strain ATCC 17978 as the reference sequence. CNV size is in base pairs (bps)

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
799744	800768	1.27	13.9	10.1	2	1024
2375168	2376192	1.02	2.43	2.39	4	1024

Table E.45: Deletions in SRR2558867 when using *A. baumannii* strain ATCC 17978 as the reference sequence. CNV size is in base pairs (bps)

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
1651968	1654784	0.628	0.346	0.503	1	2816
1685504	1688538	0.558	0.267	0.509	2	3034
1634304	1637888	0.525	0.213	0.362	3	3584
1637463	1638400	0.684	0	0	3	937
2494810	2498047	0.48	0.189	0.352	4	3237
1579008	1582080	0.799	0.337	0.37	5	3072
2193177	2195456	0.608	0.331	0.541	6	2279
2771584	2774016	0.472	0.178	0.372	7	2432
88064	90112	0.533	0.147	0.257	8	2048
1333505	1335295	0.67	0.472	0.831	9	1790
535296	537600	0.469	0.145	0.29	10	2304
1650176	1651712	0.726	0.316	0.442	11	1536
1568768	1570432	0.629	0.402	0.542	12	1664
1493389	1495040	0.689	0.27	0.35	13	1651
1507841	1509376	0.521	0.262	0.544	14	1535
847872	849920	0.434	0.14	0.311	15	2048
1642474	1644543	0.651	0.228	0.398	16	2069
1575680	1577472	0.394	0.153	0.372	17	1792
1573120	1574912	0.486	0.199	0.419	18	1792
1441792	1443328	0.567	0.282	0.459	19	1536
141312	143360	0.42	0.0878	0.242	20	2048
97307	98816	0.629	0.239	0.406	21	1509
3598080	3601408	0.275	0.113	0.183	22	3328
879361	880640	0.737	0.493	0.715	23	1279
855552	857600	0.446	0.124	0.236	24	2048
1261824	1263616	0.483	0.12	0.181	25	1792
2113537	2114754	0.746	0.632	0.782	26	1217
112640	114176	0.488	0.236	0.454	27	1536
3757056	3758592	0.671	0.237	0.366	28	1536
2073345	2074584	0.862	0.45	0.459	29	1239
1274347	1318912	0.0388	0.00429	0.00718	30	44565
1056000	1062912	0.216	0.0445	0.0885	31	6912
91649	92858	0.619	0.279	0.48	32	1209
2668544	2669965	0.7	0.576	0.753	33	1421
2363136	2374656	0.089	0.0156	0.0277	34	11520
2140928	2142208	0.609	0.191	0.343	35	1280
2153984	2156543	0.518	0.138	0.276	37	2559
2767104	2768409	0.66	0.251	0.471	38	1305
2076032	2077952	0.477	0.216	0.389	39	1920
954368	956160	0.809	0.362	0.374	41	1792
2752512	2753536	0.431	0.278	0.555	42	1024
1998848	1999872	0.525	0.128	0.321	43	1024
1558510	1559552	0.487	0.145	0.332	44	1042
2078208	2080105	0.489	0.153	0.332	45	1897
1678336	1679530	0.532	0.44	0.648	46	1194
563072	566272	0.276	0.0489	0.121	47	3200
247296	248832	0.733	0.15	0.222	48	1536
2498048	2500096	0.369	0.0891	0.236	49	2048
1121280	1122304	0.446	0.186	0.375	50	1024
2148352	2150400	0.518	0.075	0.16	51	2048
1987584	1988864	0.354	0.127	0.346	54	1280
3434524	3435776	0.49	0.117	0.28	56	1252
1607680	1608704	0.7	0.507	0.602	57	1024
3342807	3344083	0.475	0.191	0.461	58	1276
2758656	2759680	0.515	0.266	0.602	60	1024

Table E.45 Deletions in SRR2558867 when using *A. baumannii* strain ATCC 17978 as the reference sequence. Continued

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
2493423	2494909	0.477	0.217	0.445	61	1486
3748864	3750143	0.9	0.346	0.412	62	1279
2596864	2597888	0.669	0.278	0.392	64	1024
1691648	1693184	0.453	0.113	0.243	66	1536
3755521	3756544	0.545	0.382	0.607	67	1023
57344	72448	0.0683	0.00837	0.0191	68	15104
3750144	3754368	0.0769	0.0124	0.0293	70	4224
2054657	2055680	0.69	0.234	0.299	71	1023
1488891	1489920	0.648	0.39	0.477	72	1029
1195008	1196288	0.35	0.14	0.386	73	1280
1062913	1063936	0.674	0.44	0.538	74	1023
1554176	1555456	0.381	0.0501	0.231	76	1280
1201152	1205761	0.218	0.0231	0.0432	77	4609
1110784	1112063	0.701	0.463	0.621	79	1279
2351104	2353024	0.278	0.0517	0.154	80	1920
1103872	1104896	0.443	0.237	0.541	81	1024
929536	933376	0.161	0.027	0.064	83	3840
3463168	3464192	0.283	0.033	0.112	84	1024
2744320	2745344	0.807	0.441	0.506	85	1024
850944	851967	0.57	0.38	0.651	87	1023
1483776	1484800	0.686	0.287	0.366	89	1024
3603456	3604479	0.657	0.0328	0.0366	90	1023
3064320	3070464	0.0182	6.79e-06	1.66e-05	91	6144
223232	225280	0.239	0.0479	0.0606	93	2048
82432	83456	0.536	0.0567	0.16	94	1024
561152	562176	0.445	0.116	0.261	95	1024
3607552	3608576	0.386	0.074	0.162	99	1024
2150337	2151424	0.762	0.131	0.276	101	1087
1680896	1681920	0.338	0.0458	0.153	102	1024
1501952	1505536	0.0548	0.00213	0.0059	115	3584
96256	97280	0.556	0.17	0.198	116	1024
1495296	1496320	0.432	0.0936	0.267	119	1024
2026496	2028032	0.337	0.068	0.214	120	1536
163072	164096	0.526	0.203	0.52	121	1024
664576	665600	0.286	0.0257	0.0832	122	1024
1841587	1845248	0.095	0.0206	0.0545	124	3661
1645305	1647616	0.171	0.00778	0.0297	127	2311
2024448	2025472	0.264	0.00917	0.046	128	1024
1481216	1482240	0.388	0.0923	0.278	129	1024
1594368	1595391	0.58	0.379	0.56	133	1023
1546752	1547776	0.857	0.222	0.266	134	1024
2769408	2770437	0.321	0.0954	0.241	138	1029
1236480	1260032	0.00901	0.00136	0.00279	140	23552
3758847	3760128	0.502	0.0625	0.0976	152	1281
2674688	2676224	0.238	0.000686	0.00344	153	1536
1592313	1593856	0.438	0.0798	0.212	158	1543
1585408	1587200	0.18	0.00231	0.00896	159	1792
2170624	2174592	0.12	0.00616	0.0254	163	3968
3909248	3910656	0.44	0.137	0.194	168	1408
2166784	2169856	0.103	0.00804	0.0202	194	3072
17408	18432	1.2	0.0486	0.0165	209	1024
1562624	1563647	0.53	0.238	0.403	212	1023
564224	566272	0.214	0.0109	0.0417	216	2048
2400768	2401792	0.252	0.00611	0.0153	218	1024
1688064	1689600	0.245	0.033	0.075	220	1536

Table E.45 Deletions in SRR2558867 when using *A. baumannii* strain ATCC 17978 as the reference sequence. Continued

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
2060097	2061312	0.318	0.118	0.246	225	1215
1838575	1841152	0.049	0.00534	0.0198	226	2577
1570688	1571839	0.374	0.0705	0.243	232	1151
854017	855197	0.59	0.149	0.211	234	1180
3885056	3886403	0.333	0.00891	0.0262	236	1347
2093056	2094605	0.522	0.0739	0.137	238	1549
821248	826880	0.0701	0.0159	0.0337	251	5632
3903488	3907584	0.0542	0.00584	0.0131	259	4096
2156019	2157568	0.151	0.0152	0.0431	262	1549
674048	687104	0.0266	2.51e-05	6.16e-05	279	13056
2944491	2950656	0.0396	0.000637	0.00112	279	6165
3184640	3186176	0.293	0.0419	0.0898	287	1536
2030080	2031103	0.398	0.207	0.481	291	1023
1597440	1598463	0.328	0.103	0.354	294	1023
2074496	2075648	0.348	0.125	0.227	301	1152
2134400	2137600	0.0616	0.00294	0.00855	306	3200
871168	879360	0.0811	0.0118	0.0153	322	8192
2067456	2072832	0.0301	0.00251	0.0088	327	5376
2016256	2017279	0.352	0.0511	0.0757	351	1023
863232	864768	0.222	0.0415	0.126	353	1536
2329600	2334209	0.0314	0.000262	0.00079	370	4609
1506301	1507840	0.25	0.0492	0.145	378	1539
638208	639232	0.307	0.0202	0.0561	387	1024
2062336	2063360	0.344	0.13	0.337	392	1024
2446848	2447872	0.364	0.0207	0.0599	396	1024
2314752	2316288	0.232	0.0645	0.186	397	1536
1345536	1375751	0.00875	0.000788	0.00194	415	30215
99072	106752	0.0333	0.0034	0.00785	422	7680
1828864	1837824	0.0347	0.00337	0.00778	426	8960
1597952	1598976	0.372	0.0211	0.0627	429	1024
2174848	2176512	0.101	0.00106	0.00417	456	1664
124416	125952	0.163	0.0131	0.0484	466	1536
2004887	2006016	0.211	0.00829	0.0221	472	1129
1053696	1054720	0.249	0.00306	0.00767	490	1024
3008512	3012608	0.0773	0.0202	0.031	522	4096
1061888	1062912	0.312	0.0945	0.237	562	1024
2129463	2132224	0.113	0.0176	0.026	593	2761
2360064	2362379	0.105	0.0153	0.0517	607	2315
3229817	3233408	0.0762	0.00113	0.00142	681	3591
2107392	2113536	0.0444	0.00391	0.00963	691	6144
511488	514560	0.214	0.00618	0.00532	746	3072

E.6.2 Reference: *A. baumannii* strain MDR-ZJ06

Table E.46: Deletions in SRR2558867 when using *A. baumannii* strain MDR-ZJ06 as the reference sequence. CNV size is in base pairs (bps)

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
3474944	3480576	3.12	1.77	0.524	1	5632
3507328	3512960	3.12	1.77	0.604	1	5632
3974144	3980544	2.96	1.75	0.545	1	6400
70656	76715	0.704	0.517	0.748	2	6059
1095363	1098245	0.465	0.254	0.568	3	2882
1640448	1642752	0.804	0.536	0.696	4	2304
66223	68511	0.565	0.311	0.475	6	2288
1421251	1423079	0.779	0.614	0.786	7	1828
1393152	1395712	0.516	0.231	0.423	9	2560
1407488	1409536	0.577	0.422	0.681	10	2048
1402880	1405440	0.625	0.316	0.53	11	2560
1430528	1432185	0.689	0.376	0.651	12	1657
1427438	1429635	0.704	0.451	0.588	13	2197
1375232	1376805	0.615	0.414	0.805	14	1573
1092864	1094067	0.809	0.97	1.16	15	1203
1423189	1424384	0.67	0.57	0.831	16	1195
1433600	1435648	0.668	0.422	0.604	17	2048
2419107	2420736	0.277	0.174	0.247	18	1629
1418179	1419264	0.641	0.522	0.843	19	1085
1424896	1426503	0.738	0.615	0.737	20	1607
1432205	1433561	0.726	0.251	0.554	21	1356
1649791	1651456	1.96	0.586	0.54	22	1665
95616	108032	0.0595	0.0183	0.0515	23	12416
1384448	1385549	0.753	0.285	0.574	24	1101
1414133	1415659	0.425	0.194	0.484	27	1526
1101824	1102848	0.549	0.334	0.716	29	1024
2429952	2438173	0.0877	0.0338	0.0732	30	8221
68608	69632	0.495	0.296	0.762	32	1024
3141632	3143168	1.03	0.519	0.514	33	1536
1412608	1413632	0.523	0.463	0.699	35	1024
1371136	1372160	0.761	0.37	0.495	37	1024
76800	94592	0.0367	0.00492	0.0128	39	17792
1008128	1009152	0.869	0.347	0.449	41	1024
2441507	2444032	0.123	0.011	0.0234	42	2525
1567232	1568237	0.868	0.687	0.82	43	1005
1762304	1763328	1.1	0.846	0.691	45	1024
1477632	1478656	1.01	0.324	0.317	49	1024
1534976	1536000	0.907	0.174	0.196	50	1024
2876416	2878407	0.218	0.0108	0.0148	57	1991
2426368	2428352	0.157	0.00154	0.00482	125	1984
2686976	2690304	0.133	0.0164	0.0164	173	3328
2439680	2440704	1.33	0.168	0.0786	186	1024

Table E.47: Duplications in SRR2558867 when using *A. baumannii* strain MDR-ZJ06 as the reference sequence. CNV size is in base pairs (bps)

5' end	3' end	CN in Ref	CN in query	CN ratio	Group	Size
39424	44800	3.17	4.59	1.46	2	5376
218112	223744	3.14	4.52	1.43	2	5632
716544	721920	3.17	4.6	1.37	2	5376
2406656	2410240	1.04	3.09	2.87	3	3584
2410624	2414080	1.05	2.72	2.54	4	3456
3211648	3212544	1.56	9.69	7.16	5	896
1259256	1260416	1.11	3.02	2.65	7	1160
94464	95712	4.28	7.48	1.68	8	1248
2051712	2052960	4.3	7.54	1.65	8	1248
2058975	2060160	4.3	7.72	1.88	8	1185
2363392	2364647	4.29	7.61	1.68	8	1255
2446976	2447936	1.01	2.58	2.57	10	960
2404608	2405504	0.988	2.02	2.04	12	896

REFERENCES

[1] Muin J. Khoury and John P. A. Ioannidis. “Big data meets public health”. In: *Science* 346.6213 (2014), pp. 1054–1055.

[2] David Lazer et al. “The Parable of Google Flu: Traps in Big Data Analysis”. In: *Science* 343.6176 (2014), pp. 1203–1205.

[3] R. Bellazzi. “Big Data and Biomedical Informatics: A Challenging Opportunity”. In: *IMIA Yearbook* 1 (2014), pp. 8–13.

[4] Stephen P. Perfetto, Pratip K. Chattopadhyay, and Mario Roederer. “Seventeen-colour flow cytometry: unravelling the immune system”. In: *Nat Rev Immunol* 4.8 (2004), pp. 648–655.

[5] Pratip K. Chattopadhyay et al. “Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry”. In: *Nat Med* 12.8 (2006), pp. 972–977.

[6] Yvan Saeys, Sofie Van Gassen, and Bart N. Lambrecht. “Computational flow cytometry: helping to make sense of high-dimensional immunology data”. In: *Nat Rev Immunol* 16.7 (2016), pp. 449–462.

[7] Cesar A. Arias and Barbara E. Murray. “Antibiotic-Resistant Bugs in the 21st Century A Clinical Super-Challenge”. In: *New England Journal of Medicine* 360.5 (2009), pp. 439–443.

[8] Ronald N. Master et al. “Recent trends in resistance to cell envelopeactive antibacterial agents among key bacterial pathogens”. In: *Annals of the New York Academy of Sciences* 1277.1 (2013), pp. 1–7.

[9] Helen W. Boucher et al. “10 ’20 ProgressDevelopment of New Drugs Active Against Gram-Negative Bacilli: An Update From the Infectious Diseases Society of America”. In: *Clinical Infectious Diseases* (2013).

[10] R. Klevens et al. “INvasive methicillin-resistant staphylococcus aureus infections in the united states”. In: *JAMA* 298.15 (2007), pp. 1763–1771.

[11] Helen W. Boucher et al. “Bad Bugs, No Drugs: No ESKAPE! An Update from the Infectious Diseases Society of America”. In: *Clinical Infectious Diseases* 48.1 (2009), pp. 1–12.

[12] Helen W. Boucher and G. Ralph Corey. “Epidemiology of Methicillin-Resistant Staphylococcus aureus”. In: *Clinical Infectious Diseases* 46.Supplement 5 (2008), S344–S349.

[13] Alan C Heffner et al. “Etiology of illness in patients with severe sepsis admitted to the hospital from the emergency department”. In: *Clinical Infectious Diseases* 50.6 (2010), pp. 814–20.

[14] G V Doern et al. “Clinical impact of rapid in vitro susceptibility testing and bacterial identification”. In: *Journal of Clinical Microbiology* 32.7 (1994), pp. 1757–1762.

[15] E. H. Ibrahim et al. “The influence of inadequate antimicrobial treatment of bloodstream infections on patient outcomes in the ICU setting”. In: *Chest* 118.1 (2000), pp. 146–55.

[16] Y. H. Chen, S. L. Nyeo, and C. Y. Yeh. “Model for the distributions of k-mers in DNA sequences”. In: *Phys Rev E* 72 (2005).

[17] Michael A. Kohanski, Daniel J. Dwyer, and James J. Collins. “How antibiotics kill bacteria: from targets to networks”. In: *Nat Rev Micro* 8.6 (2010), pp. 423–435.

[18] Amy Fothergill et al. “Rapid Identification of Bacteria and Yeasts from Positive-Blood-Culture Bottles by Using a Lysis-Filtration Method and Matrix-Assisted Laser Desorption IonizationTime of Flight Mass Spectrum Analysis with the SARAMIS Database”. In: *Journal of Clinical Microbiology* 51.3 (2013), pp. 805–809.

[19] Remco P. H. Peters et al. “Faster Identification of Pathogens in Positive Blood Cultures by Fluorescence In Situ Hybridization in Routine Practice”. In: *Journal of Clinical Microbiology* 44.1 (2006), pp. 119–123.

[20] Kim B. Barken, Janus A. J. Haagensen, and Tim Tolker-Nielsen. “Advances in nucleic acid-based diagnostics of bacterial infections”. In: *Clinica Chimica Acta* 384.12 (2007), pp. 1–11.

[21] M. Christner et al. “Rapid identification of bacteria from positive blood culture bottles by use of matrix-assisted laser desorption-ionization time of flight mass spectrometry fingerprinting”. In: *J Clin Microbiol* 48.5 (2010), pp. 1584–91.

[22] A. K. Barczak et al. “RNA signatures allow rapid identification of pathogens and antibiotic susceptibilities”. In: *Proc Natl Acad Sci U S A* 109.16 (2012), pp. 6217–22.

[23] X. Didelot et al. “Transforming clinical microbiology with bacterial genome sequencing”. In: *Nat Rev Genet* 13.9 (2012), pp. 601–12.

[24] J. C. Kwong et al. “Whole genome sequencing in clinical and public health microbiology”. In: *Pathology* 47.3 (2015), pp. 199–210.

[25] Simon R. Harris et al. “Evolution of MRSA During Hospital Transmission and Intercontinental Spread”. In: *Science* 327.5964 (2010), pp. 469–474.

[26] Joshua Quick et al. “Real-time, portable genome sequencing for Ebola surveillance”. In: *Nature* 530.7589 (2016), pp. 228–232.

[27] Hilde Vinje et al. “Comparing K-mer based methods for improved classification of 16S sequences”. In: *BMC Bioinformatics* 16.1 (2015), p. 205.

[28] C. Alkan et al. “Personalized copy number and segmental duplication maps using next-generation sequencing”. In: *Nat Genet* 41.10 (2009), pp. 1061–7.

[29] Heng Li, Jue Ruan, and Richard Durbin. “Mapping short DNA sequencing reads and calling variants using mapping quality scores”. In: *Genome Research* 18.11 (2008), pp. 1851–1858.

[30] B. E. Blaisdell. “A measure of the similarity of sets of sequences not requiring sequence alignment.” In: *Proceedings of the National Academy of Sciences* 83.14 (1986), pp. 5155–5159.

[31] G. E. Sims et al. “Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions”. In: *Proceedings of the National Academy of Sciences* 106.8 (2009), pp. 2677–2682.

[32] B. Haubold. “Alignment-free phylogenetics and population genetics”. In: *Briefings in Bioinformatics* 15.3 (2013), pp. 407–418.

[33] K. R. Patil and A. C. McHardy. “Alignment-Free Genome Tree Inference by Learning Group-Specific Distance Metrics”. In: *Genome Biology and Evolution* 5.8 (2013), pp. 1470–1484.

[34] M. C. Maiden et al. “MLST revisited: the gene-by-gene approach to bacterial genomics”. In: *Nat Rev Microbiol* 11.10 (2013), pp. 728–36.

[35] Michael Inouye et al. “Short read sequence typing (SRST): multi-locus sequence types from short reads”. In: *BMC Genomics* 13.1 (2012), p. 338.

[36] K. A. Jolley et al. “Ribosomal Multi-Locus Sequence Typing: universal characterisation of bacteria from domain to strain”. In: *Microbiol* 158 (2012).

[37] Wan-Ping Lee et al. “MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping”. In: *PLoS ONE* 9.3 (2014), e90581.

[38] S. F. Altschul et al. “Basic local alignment search tool”. In: *J Mol Biol* 215.3 (1990), pp. 403–10.

[39] Ben Langmead et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome Biology* 10.3 (2009), R25.

[40] H. Li and R. Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25.14 (2009), pp. 1754–60.

[41] M. Zhao et al. “Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives”. In: *BMC Bioinformatics* 14 Suppl 11 (2013), S1.

[42] A. Abyzov et al. “CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing”. In: *Genome Research* 21.6 (2011), pp. 974–984.

[43] S. M. Teo et al. “Statistical challenges associated with detecting copy number variations with next-generation sequencing”. In: *Bioinformatics* 28.21 (2012), pp. 2711–8.

[44] F. Hormozdiari et al. “Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes”. In: *Genome Research* 19.7 (2009), pp. 1270–1278.

[45] Derek Y Chiang et al. “High-resolution mapping of copy-number alterations with massively parallel sequencing”. In: *Nature Methods* 6.1 (2008), pp. 99–103.

[46] S. Ivakhno et al. “CNAseg–a novel framework for identification of copy number changes in cancer from second-generation sequencing data”. In: *Bioinformatics* 26.24 (2010), pp. 3051–3058.

[47] Derek Y. Chiang and Steven A. McCarroll. “Mapping duplicated sequences”. In: *Nat Biotech* 27.11 (2009), pp. 1001–1002.

[48] Tzu-Hsueh Huang et al. “Rapid Cytometric Antibiotic Susceptibility Testing Utilizing Adaptive Multidimensional Statistical Metrics”. In: *Analytical Chemistry* 87.3 (2015), pp. 1941–1949.

[49] M. J. Fulwyler. “Electronic separation of biological cells by volume”. In: *Science* 150.3698 (1965), pp. 910–1.

[50] A. Pierzchalski, A. Mittag, and A. Tarnok. “Introduction A: recent advances in cytometry instrumentation, probes, and methods–review”. In: *Methods Cell Biol* 102 (2011), pp. 1–21.

[51] H. R. Hulett et al. “Cell sorting: automated separation of mammalian cells as a function of intracellular fluorescence”. In: *Science* 166.3906 (1969), pp. 747–9.

[52] J. Paul Robinson and Mario Roederer. “Flow cytometry strikes gold”. In: *Science* 350.6262 (2015), pp. 739–740.

[53] D. W. Galbraith et al. “Rapid flow cytometric analysis of the cell cycle in intact plant tissues”. In: *Science* 220.4601 (1983), pp. 1049–51.

[54] Jong Ah Kim et al. “Role of cell cycle on the cellular uptake and dilution of nanoparticles in a cell population”. In: *Nat Nano* 7.1 (2012), pp. 62–68.

[55] Thomas Blasi et al. “Label-free cell cycle analysis for high-throughput imaging flow cytometry”. In: *Nature Communications* 7 (2016), p. 10256.

[56] Barthel Barlogie et al. “Flow Cytometry in Clinical Cancer Research”. In: *Cancer Research* 43.9 (1983), pp. 3982–3997.

[57] Eric Tran et al. “Cancer Immunotherapy Based on Mutation-Specific CD4+ T Cells in a Patient with Epithelial Cancer”. In: *Science* 344.6184 (2014), pp. 641–645.

[58] V. Denes et al. “Metastasis blood test by flow cytometry: in vivo cancer spheroids and the role of hypoxia”. In: *Int J Cancer* 136.7 (2015), pp. 1528–36.

[59] Michael Berney et al. “Assessment and Interpretation of Bacterial Viability by Using the LIVE/DEAD BacLight Kit in Combination with Flow Cytometry”. In: *Applied and Environmental Microbiology* 73.10 (2007), pp. 3283–3290.

[60] Dan A. Buzatu et al. “An Integrated Flow Cytometry-Based System for Real-Time, High Sensitivity Bacterial Detection and Identification”. In: *PLoS ONE* 9.4 (2014), e94254.

[61] M. L. Shuler, R. Aris, and H. M. Tsuchiya. “Hydrodynamic Focusing and Electronic Cell-Sizing Techniques”. In: *Applied Microbiology* 24.3 (1972), pp. 384–388.

[62] Semrock Filters for Flow Cytometry. `www.google.com`. 2016.

[63] V. A. Gant et al. “The application of flow cytometry to the study of bacterial responses to antibiotics”. In: *J Med Microbiol* 39.2 (1993), pp. 147–54.

[64] D. J. Mason et al. “Rapid estimation of bacterial antibiotic susceptibility with flow cytometry”. In: *Journal of Microscopy* 176.1 (1994), pp. 8–16.

[65] Mette Walberg, Peter Gaustad, and Harald B. Steen. “Rapid flow cytometric assessment of mecillinam and ampicillin bacterial susceptibility”. In: *Journal of Antimicrobial Chemotherapy* 37.6 (1996), pp. 1063–1075.

[66] Mette Walberg, Peter Gaustad, and Harald B. Steen. “Rapid assessment of ceftazidime, ciprofloxacin, and gentamicin susceptibility in exponentially-growing E. coli cells by means of flow cytometry”. In: *Cytometry* 27.2 (1997), pp. 169–178.

[67] M. T. E. Suller and D. Lloyd. “Fluorescence monitoring of antibiotic-induced bacterial damage using flow cytometry”. In: *Cytometry* 35.3 (1999), pp. 235–241.

[68] Fiona C. Mortimer, David J. Mason, and Vanya A. Gant. “Flow Cytometric Monitoring of Antibiotic-Induced Injury in Escherichia coli Using Cell-Impermeant Fluorescent Probes”. In: *Antimicrobial Agents and Chemotherapy* 44.3 (2000), pp. 676–681.

[69] Mette Walberg and Harald B. Steent. “flow cytometric monitoring of bacterial susceptibility to antibiotics”. In: *Methods in Cell Biology*. Ed. by Harry A. Crissman J. Paul Robinson Zbigniew Darzynkiewicz. Vol. Volume 64, Part B. Academic Press, 2001, pp. 553–566. ISBN: 0091-679X.

[70] Christian Gauthier, Yves St-Pierre, and Richard Villemur. “Rapid antimicrobial susceptibility testing of urinary tract isolates and samples by flow cytometry”. In: *Journal of Medical Microbiology* 51.3 (2002), pp. 192–200.

[71] P. Assuno et al. “Application of flow cytometry for the determination of minimal inhibitory concentration of several antibacterial agents on Mycoplasma hyopneumoniae”. In: *Journal of Applied Microbiology* 102.4 (2007), pp. 1132–1137.

[72] I. Faria-Ramos et al. “A novel flow cytometric assay for rapid detection of extended-spectrum beta-lactamases”. In: *Clinical Microbiology and Infection* 19.1 (2013), E8–E15.

[73] S Nuding and T. L Zabel. “Detection, Identification and Susceptibility Testing of Bacteria by Flow Cytometry”. In: *J Bacteriol Parasitol* S5 (2013), p. 005.

[74] Deirdre Kennedy, Ultan P. Cronin, and Martin G. Wilkinson. “Responses of Escherichia coli, Listeria monocytogenes, and Staphylococcus aureus to Simulated Food Processing Treatments, Determined Using Fluorescence-Activated Cell Sorting and Plate Counting”. In: *Applied and Environmental Microbiology* 77.13 (2011), pp. 4657–4668.

[75] David J. Novo et al. “Multiparameter Flow Cytometric Analysis of Antibiotic Effects on Membrane Potential, Membrane Permeability, and Bacterial Counts of

Staphylococcus aureus andMicrococcus luteus". In: *Antimicrobial Agents and Chemotherapy* 44.4 (2000), pp. 827–834.

[76] Howard M. Shapiro. "Multiparameter flow cytometry of bacteria: Implications for diagnostics and therapeutics". In: *Cytometry* 43.3 (2001), pp. 223–226.

[77] I T Young. "Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources". In: *Journal of Histochemistry & Cytochemistry* 25.7 (1977), pp. 935–41.

[78] Christopher Cox et al. "Comparison of frequency distributions in flow cytometry". In: *Cytometry* 9.4 (1988), pp. 291–298.

[79] Mario Roederer et al. "Probability binning comparison: a metric for quantitating multivariate distribution differences". In: *Cytometry* 45.1 (2001), pp. 47–55.

[80] Mario Roederer et al. "Probability binning comparison: A metric for quantitating univariate distribution differences". In: *Cytometry* 45.1 (2001), pp. 37–46.

[81] Tytus Bernas et al. "Quadratic form: A robust metric for quantitative comparison of flow cytometric histograms". In: *Cytometry Part A* 73A.8 (2008), pp. 715–726.

[82] Christian Beecks, Merih Seran Uysal, and Thomas Seidl. "Signature quadratic form distances for content-based similarity". In: *Proceedings of the seventeen ACM international conference on Multimedia - 09*. Association for Computing Machinery (ACM), 2009.

[83] Christian Beecks, Merih Seran Uysal, and Thomas Seidl. "Signature Quadratic Form Distance". In: *Proceedings of the ACM International Conference on Image and Video Retrieval - 10*. Association for Computing Machinery (ACM), 2010.

[84] F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain-terminating inhibitors". In: *Proc Natl Acad Sci U S A* 74.12 (1977), pp. 5463–7.

[85] Sara Goodwin, John D. McPherson, and W. Richard McCombie. "Coming of age: ten years of next-generation sequencing technologies". In: *Nat Rev Genet* 17.6 (2016), pp. 333–351.

[86] H.P.J. Buermans and J.T. den Dunnen. "Next generation sequencing technology: Advances and applications". In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1842.10 (2014), pp. 1932–1941.

[87] E. R. Mardis. "Next-generation sequencing platforms". In: *Annu Rev Anal Chem (Palo Alto Calif)* 6 (2013), pp. 287–303.

[88] M. L. Metzker. “Sequencing technologies - the next generation”. In: *Nat Rev Genet* 11.1 (2010), pp. 31–46.

[89] M. C. Maiden et al. “Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms”. In: *Proc Natl Acad Sci U S A* 95.6 (1998), pp. 3140–5.

[90] M. V. Larsen et al. “Multilocus sequence typing of total-genome-sequenced bacteria”. In: *J Clin Microbiol* 50.4 (2012), pp. 1355–61.

[91] K. A. Jolley and M. C. Maiden. “Using multilocus sequence typing to study bacterial variation: prospects in the genomic era”. In: *Future Microbiol* 9.5 (2014), pp. 623–30.

[92] M. Achtman. “Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens”. In: *Annu Rev Microbiol* 62 (2008), pp. 53–70.

[93] Pavel A. Pevzner, Haixu Tang, and Michael S. Waterman. “An Eulerian path approach to DNA fragment assembly”. In: *Proceedings of the National Academy of Sciences* 98.17 (2001), pp. 9748–9753.

[94] D. Zerbino and E. Birney. “Velvet: Algorithms for de novo short read assembly using de Bruijn graphs”. In: *Genome Res* 18 (2008).

[95] Phillip E. C. Compeau, Pavel A. Pevzner, and Glenn Tesler. “How to apply de Bruijn graphs to genome assembly”. In: *Nat Biotech* 29.11 (2011), pp. 987–991.

[96] Karl J. V. Nordstrom et al. “Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers”. In: *Nat Biotech* 31.4 (2013), pp. 325–330.

[97] Henrik Hasman et al. “Rapid Whole-Genome Sequencing for Detection and Characterization of Microorganisms Directly from Clinical Samples”. In: *Journal of Clinical Microbiology* 52.1 (2014), pp. 139–146.

[98] Mette V. Larsen et al. “Benchmarking of Methods for Genomic Taxonomy”. In: *Journal of Clinical Microbiology* 52.5 (2014), pp. 1529–1539.

[99] K. L. Liu and T. T. Wong. “Naive Bayesian Classifiers with Multinomial Models for rRNA Taxonomic Assignment”. In: *IEEE/ACM Trans Comput Biol Bioinform* (2013).

[100] Q. Wang et al. “Nave Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy”. In: *Appl Enviromental Microbiol* 73 (2007).

[101] B. L. Maidak et al. “The Ribosomal Database Project”. In: *Nucleic Acids Res* 22 (1994).

[102] S. Wold, H. Martens, and H. Wold. “The Multivariate Calibration Problem in Chemistry solved by the PLS Method”. In: *Lect Notes Math* 973 (1983).

[103] Ankur Mutreja et al. “Evidence for several waves of global transmission in the seventh cholera pandemic”. In: *Nature* 477.7365 (2011), pp. 462–465.

[104] Cole Trapnell and Steven L. Salzberg. “How to map billions of short reads onto genomes”. In: *Nat Biotech* 27.5 (2009), pp. 455–457.

[105] Ayat Hatem et al. “Benchmarking short sequence mapping tools”. In: *BMC Bioinformatics* 14.1 (2013), p. 184.

[106] B. Ma, J. Tromp, and M. Li. “PatternHunter: faster and more sensitive homology search”. In: *Bioinformatics* 18.3 (2002), pp. 440–5.

[107] H. Lin et al. “ZOOM! Zillions of oligos mapped”. In: *Bioinformatics* 24.21 (2008), pp. 2431–7.

[108] T. F. Smith and M. S. Waterman. “Identification of common molecular subsequences”. In: *J Mol Biol* 147.1 (1981), pp. 195–7.

[109] Ben Langmead and Steven L. Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nat Meth* 9.4 (2012), pp. 357–359.

[110] Heng Li and Richard Durbin. “Fast and accurate long-read alignment with BurrowsWheeler transform”. In: *Bioinformatics* 26.5 (2010), pp. 589–595.

[111] Li Heng. “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN]”. In: ().

[112] Q. Li et al. “SOAP2: an improved ultrafast tool for short read alignment”. In: *Bioinformatics* 25 (2009).

[113] T. W. Lam et al. “Compressed indexing and local alignment of DNA”. In: *Bioinformatics* 24.6 (2008), pp. 791–7.

[114] J. A. Bailey et al. “Recent segmental duplications in the human genome”. In: *Science* 297 (2002).

[115] Andy W Pang et al. “Towards a comprehensive structural variation map of an individual human genome”. In: *Genome Biology* 11.5 (2010), R52.

[116] Ryan E. Mills et al. “Mapping copy number variation by population-scale genome sequencing”. In: *Nature* 470.7332 (2011), pp. 59–65.

[117] E. Gonzalez et al. “The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility”. In: *Science* 307.5714 (2005), pp. 1434–40.

[118] Mario Falchi et al. “Low copy number of the salivary amylase gene predisposes to obesity”. In: *Nat Genet* 46.5 (2014), pp. 492–497.

[119] Kristen M. Turner et al. “Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity”. In: *Nature* 543.7643 (2017), pp. 122–125.

[120] J. Sebat et al. “Strong Association of De Novo Copy Number Mutations with Autism”. In: *Science* 316.5823 (2007), pp. 445–449.

[121] A. B. Singleton. “Alpha-Synuclein Locus Triplication Causes Parkinson's Disease”. In: *Science* 302.5646 (2003), pp. 841–841.

[122] Marit S Bratlie et al. “Gene duplications in prokaryotes can be associated with environmental adaptation”. In: *BMC Genomics* 11.1 (2010), p. 588.

[123] Sharon Greenblum, Rogan Carr, and Elhanan Borenstein. “Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species”. In: *Cell* 160.4 (2015), pp. 583–594.

[124] Linus Sandegren and Dan I. Andersson. “Bacterial gene amplification: implications for the evolution of antibiotic resistance”. In: *Nature Reviews Microbiology* 7.8 (2009), pp. 578–588.

[125] Lorenzo Tattini, Romina DAurizio, and Alberto Magi. “Detection of Genomic Structural Variants from Next-Generation Sequencing Data”. In: *Frontiers in Bioengineering and Biotechnology* 3 (2015), p. 92.

[126] L. y. Wang et al. “MSB: A mean-shift-based approach for the analysis of structural variation in the genome”. In: *Genome Research* 19.1 (2008), pp. 106–117.

[127] Christopher A. Miller et al. “ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads”. In: *PLoS ONE* 6.1 (2011). Ed. by Stein Aerts, e16327.

[128] Chao Xie and Martti T Tammi. “CNV-seq, a new method to detect copy number variation using high-throughput sequencing”. In: *BMC Bioinformatics* 10.1 (2009), p. 80.

[129] S. Yoon et al. “Sensitive and accurate detection of copy number variants using read depth of coverage”. In: *Genome Research* 19.9 (2009), pp. 1586–1592.

[130] A. Magi et al. “Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm”. In: *Nucleic Acids Research* 39.10 (2011), e65–e65.

[131] G. Klambauer et al. “cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate”. In: *Nucleic Acids Research* 40.9 (2012), e69–e69.

[132] M. P. Mahmud, J. Wiedenhoeft, and A. Schliep. “Indel-tolerant read mapping with trinucleotide frequencies using cache-oblivious kd-trees”. In: *Bioinformatics* 28.18 (2012), pp. i325–i332.

[133] M. Ruffalo, T. LaFramboise, and M. Koyuturk. “Comparative analysis of algorithms for next-generation sequencing read alignment”. In: *Bioinformatics* 27.20 (2011), pp. 2790–6.

[134] Bernhard Haubold and Thomas Wiehe. In: *BMC Bioinformatics* 7.1 (2006), p. 541.

[135] Ekaterina Avershina and Knut Rudi. “Dominant short repeated sequences in bacterial genomes”. In: *Genomics* 105.3 (2015), pp. 175–181.

[136] A. P. Jason de Koning et al. “Repetitive Elements May Comprise Over Two-Thirds of the Human Genome”. In: *PLoS Genetics* 7.12 (2011). Ed. by Gregory P. Copenhaver, e1002384.

[137] N. Delihas. “Impact of Small Repeat Sequences on Bacterial Genome Evolution”. In: *Genome Biology and Evolution* 3.0 (2011), pp. 959–973.

[138] Todd J. Treangen and Steven L. Salzberg. “Repetitive DNA and next-generation sequencing: computational challenges and solutions”. In: *Nature Reviews Genetics* (2011).

[139] R. Li et al. “SOAP: short oligonucleotide alignment program”. In: *Bioinformatics* 24 (2008).

[140] Akira Nakayama et al. “Quantitation of brown adipose tissue perfusion in transgenic mice using near-infrared fluorescence imaging”. In: *Molecular imaging* 2.1 (2003), p. 15353500200303103.

[141] JH FentonH. “Oxidationoftartaricacidinpresenceof iron”. In: *Journal of the Chemical Society, Transactions* 65 (1894), pp. 899–910.

[142] Maan Hayyan, Mohd Ali Hashim, and Inas M. AlNashef. “Superoxide Ion: Generation and Chemical Implications”. In: *Chemical Reviews* 116.5 (2016), pp. 3029–3085.

[143] *Zeitschrift fr Astronomie und verwandte Wissenschaften*. J. G. Cotta, 1816.

[144] D. Ruppert. *Statistics and Data Analysis for Financial Engineering*. Springer, 2010. ISBN: 9781441977878.

[145] Probal Chaudhuri. “On a Geometric Notion of Quantiles for Multivariate Data”. In: *Journal of the American Statistical Association* 91.434 (1996), pp. 862–872.

[146] John E Dennis Jr and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. Vol. 16. Siam, 1996. ISBN: 1611971209.

[147] Peter Audano and Fredrik Vannberg. “KAnalyze: A Fast Versatile Pipelined K-mer Toolkit”. In: *Bioinformatics* (2014).

[148] H. Li et al. “The Sequence Alignment/Map format and SAMtools”. In: *Bioinformatics* 25 (2009).

[149] Yansun Xu et al. “Wavelet transform domain filters: a spatially selective noise filtration technique”. In: *IEEE Transactions on Image Processing* 3.6 (1994), pp. 747–758.

[150] Quan Pan et al. “Two denoising methods by wavelet transform”. In: *IEEE Transactions on Signal Processing* 47.12 (1999), pp. 3401–3406.

[151] S. Mallat and S. Zhong. “Characterization of signals from multiscale edges”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14.7 (1992), pp. 710–732.

[152] Lei Zhang and Paul Bao. “Edge detection by scale multiplication in wavelet domain”. In: *Pattern Recognition Letters* 23.14 (2002), pp. 1771 –1784.

[153] Carson Holt et al. “WaveCNV: allele-specific copy number alterations in primary tumors and xenograft models from next-generation sequencing”. In: *Bioinformatics* 30.6 (2014), pp. 768–774.

[154] D. R. Bentley et al. “Accurate whole human genome sequencing using reversible terminator chemistry”. In: *Nature* 456 (2008).

[155] André E Minoche, Juliane C Dohm, and Heinz Himmelbauer. “Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems”. In: *Genome Biology* 12.11 (2011), R112.

[156] Y. Benjamini and T. P. Speed. “Summarizing and correcting the GC content bias in high-throughput sequencing”. In: *Nucleic Acids Research* 40.10 (2012), e72–e72.

[157] Michael A. Kohanski et al. “A Common Mechanism of Cellular Death Induced by Bactericidal Antibiotics”. In: *Cell* 130.5 (2007), pp. 797–810.

[158] M. A. Kohanski et al. “Mistranslation of membrane proteins and two-component system activation trigger antibiotic-mediated cell death”. In: *Cell* 135.4 (2008), pp. 679–90.

[159] Iris Keren et al. “Killing by Bactericidal Antibiotics Does Not Depend on Reactive Oxygen Species”. In: *Science* 339.6124 (2013), pp. 1213–1216.

[160] Yuanyuan Liu and James A. Imlay. “Cell Death from Antibiotics Without the Involvement of Reactive Oxygen Species”. In: *Science* 339.6124 (2013), pp. 1210–1213.

[161] Heleen Van Acker and Tom Coenye. “The Role of Reactive Oxygen Species in Antibiotic-Mediated Killing of Bacteria”. In: *Trends in Microbiology* (2017).

[162] K. Setsukinai et al. “Development of novel fluorescence probes that can reliably detect reactive oxygen species and distinguish specific species”. In: *J Biol Chem* 278.5 (2003), pp. 3170–5.

[163] Kristine M. Robinson et al. “Selective fluorescent imaging of superoxide in vivo using ethidium-based probes”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.41 (2006), pp. 15038–15043.

[164] Xinghai Ning et al. “Maltodextrin-based imaging probes detect bacteria in vivo with high sensitivity and specificity”. In: *Nature Materials* 10.8 (2011), pp. 602–607.

[165] W. Boos and H. Shuman. “Maltose/maltodextrin system of Escherichia coli: Transport, metabolism, and regulation”. In: *Microbiology and Molecular Biology Reviews* 62.1 (1998), pp. 204–+.

[166] K. Kundu et al. “Hydrocyanines: a class of fluorescent sensors that can image reactive oxygen species in cell culture, tissue, and in vivo”. In: *Angew Chem Int Ed Engl* 48.2 (2009), pp. 299–303.

[167] J. A. Imlay. “Pathways of oxidative damage”. In: *Annu Rev Microbiol* 57 (2003), pp. 395–418.

[168] J. A. Imlay. “Cellular defenses against superoxide and hydrogen peroxide”. In: *Annu Rev Biochem* 77 (2008), pp. 755–76.

[169] Hak Suk Chung et al. “Rapid -lactam-induced lysis requires successful assembly of the cell division machinery”. In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21872–21877.

[170] H. J. Wickens et al. “Flow Cytometric Investigation of Filamentation, Membrane Patency, and Membrane Potential in Escherichia coli following Ciprofloxacin Exposure”. In: *Antimicrobial Agents and Chemotherapy* 44.3 (2000), pp. 682–687.

[171] Christina Steel, Qian Wan, and Xiao-Hong Nancy Xu. “Single Live Cell Imaging of Chromosomes in Chloramphenicol-Induced Filamentous Pseudomonas aeruginosa”. In: *Biochemistry* 43.1 (2003), pp. 175–182.

[172] Susann Mller and Gerhard Nebe-von Caron. “Functional single-cell analyses: flow cytometry and cell sorting of microbial populations and communities”. In: *FEMS Microbiology Reviews* 34.4 (2010), pp. 554–587.

[173] Gian Maria Rossolini et al. “Update on the antibiotic resistance crisis”. In: *Current Opinion in Pharmacology* 18 (2014), pp. 56–60.

[174] R. M. Klevens et al. “Changes in the Epidemiology of Methicillin-Resistant Staphylococcus aureus in Intensive Care Units in US Hospitals, 1992-2003”. In: *Clinical Infectious Diseases* 42.3 (2006), pp. 389–391.

[175] Deron C. Burton. “Methicillin-resistant Staphylococcus aureus central line-associated bloodstream infections in US intensive care units, 1997-2007”. In: *JAMA* 301.7 (2009), p. 727.

[176] R. Dippel and W. Boos. “The Maltodextrin System of Escherichia coli: Metabolism and Transport”. In: *Journal of Bacteriology* 187.24 (2005), pp. 8322–8331.

[177] CLSI. *M100-S24 Performance Standards for Antimicrobial Susceptibility Testing; Twenty-fourth Informational Supplement*. Jan. 2014.

[178] M. D. Carruthers et al. “Draft Genome Sequence of the Clinical Isolate Acinetobacter nosocomialis Strain M2”. In: *Genome Announc* 1.6 (2013).

[179] S. D. Saroj et al. “Novel mechanism for fluoroquinolone resistance in Acinetobacter baumannii”. In: *Antimicrob Agents Chemother* 56.9 (2012), pp. 4955–7.

[180] Ron Daniels. “Surviving the first hours in sepsis: getting the basics right (an intensivist’s perspective)”. In: *Journal of Antimicrobial Chemotherapy* 66.suppl 2 (2011), pp. ii11–ii23.

[181] B. E. Kreger, D. E. Craven, and W. R. McCabe. “Gram-negative bacteremia. IV. Re-evaluation of clinical features and treatment in 612 patients”. In: *Am J Med* 68.3 (1980), pp. 344–55.

[182] Marin H. Kollef. “Broad-Spectrum Antimicrobials and the Treatment of Serious Bacterial Infections: Getting It Right Up Front”. In: *Clinical Infectious Diseases* 47.Supplement 1 (2008), S3–S13.

[183] A. Fraser et al. “Benefit of appropriate empirical antibiotic treatment: thirty-day mortality and duration of hospital stay”. In: *Am J Med* 119.11 (2006), pp. 970–6.

[184] Dale E. Dietzman, Gerald W. Fischer, and Fritz D. Schoenknecht. “Neonatal escherichia coli septicemiabacterial counts in blood”. In: *The Journal of Pediatrics* 85.1 (1974), pp. 128–130.

[185] P. Yagupsky and F. S. Nolte. “Quantitative aspects of septicemia”. In: *Clinical Microbiology Reviews* 3.3 (1990), pp. 269–279.

[186] S. Sauer and M. Kliem. “Mass spectrometry tools for the classification and identification of bacteria”. In: *Nat Rev Microbiol* 8.1 (2010), pp. 74–82.

[187] James R. Carey et al. “Rapid Identification of Bacteria with a Disposable Colorimetric Sensing Array”. In: *Journal of the American Chemical Society* 133.19 (2011), pp. 7571–7576.

[188] A. Huletsky et al. “New Real-Time PCR Assay for Rapid Detection of Methicillin-Resistant Staphylococcus aureus Directly from Specimens Containing a Mixture of Staphylococci”. In: *Journal of Clinical Microbiology* 42.5 (2004), pp. 1875–1884.

[189] M.A.C. Broeren et al. “Antimicrobial susceptibility testing in 90 min by bacterial cell count monitoring”. In: *Clinical Microbiology and Infection* 19.3 (2013), pp. 286–291.

[190] J. D. Mansour et al. “Detection of Escherichia coli in blood using flow cytometry”. In: *Cytometry* 6.3 (1985), pp. 186–90.

[191] W. G. Pitt et al. “Rapid separation of bacteria from blood-review and outlook”. In: *Biotechnol Prog* 32.4 (2016), pp. 823–39.

[192] Dong-Ku Kang et al. “Rapid detection of single bacteria in unprocessed blood using Integrated Comprehensive Droplet Digital Detection”. In: *Nature Communications* 5 (2014), p. 5427.

[193] S. Zelenin et al. “Microfluidic-based isolation of bacteria from whole blood for sepsis diagnostics”. In: *Biotechnol Lett* 37.4 (2015), pp. 825–30.

[194] H. W. Hou et al. “Direct detection and drug-resistance profiling of bacteremias using inertial microfluidics”. In: *Lab Chip* 15.10 (2015), pp. 2297–307.

[195] T. Gosiewski et al. “Comparison of methods for isolation of bacterial and fungal DNA from human blood”. In: *Curr Microbiol* 68.2 (2014), pp. 149–55.

[196] Tomasz Gosiewski et al. “A novel, nested, multiplex, real-time PCR for detection of bacteria and fungi in blood”. In: *BMC Microbiology* 14.1 (2014), p. 144.

[197] Raphael Ber et al. “Enrichment of Yersinia pestis from Blood Cultures Enables Rapid Antimicrobial Susceptibility Determination by Flow Cytometry”. In: *The Genus Yersinia*. Ed. by RobertD Perry and JacquelineD Fetherston. Vol. 603. Advances In Experimental Medicine And Biology. Springer New York, 2007. Chap. 31, pp. 339–350. ISBN: 978-0-387-72123-1.

[198] Michael A. Kohanski, Mark A. DePristo, and James J. Collins. “Sublethal Antibiotic Treatment Leads to Multidrug Resistance via Radical-Induced Mutagenesis”. In: *Molecular Cell* 37.3 (2010), pp. 311–320.

[199] H. Gogelein and A. Huby. “Interaction of saponin and digitonin with black lipid membranes and lipid monolayers”. In: *Biochim Biophys Acta* 773.1 (1984), pp. 32–8.

[200] G. Francis et al. “The biological action of saponins in animal systems: a review”. In: *Br J Nutr* 88.6 (2002), pp. 587–605.

[201] Michal Arabski et al. “Effects of Saponins against Clinical E. coli Strains and Eukaryotic Cell Line”. In: *Journal of Biomedicine and Biotechnology* 2012 (2012), p. 6.

[202] Nabin K. Shrestha et al. “Rapid Differentiation of Methicillin-Resistant and Methicillin-Susceptible Staphylococcus aureus by Flow Cytometry after Brief Antibiotic Exposure”. In: *Journal of Clinical Microbiology* 49.6 (2011), pp. 2116–2120.

[203] José V Ordóñez and Natalie M Wehman. “Rapid flow cytometric antibiotic susceptibility assay for Staphylococcus aureus”. In: *Cytometry Part A* 14.7 (1993), pp. 811–818.

[204] C. H. Kao et al. “Isolated pathogens and clinical outcomes of adult bacteremia in the emergency department: a retrospective study in a tertiary Referral Center”. In: *J Microbiol Immunol Infect* 44.3 (2011), pp. 215–21.

[205] Jean-Louis Vincent. “International Study of the Prevalence and Outcomes of Infection in Intensive Care Units”. In: *JAMA* 302.21 (2009), p. 2323.

[206] Jacqueline Deen et al. “Community-acquired bacterial bloodstream infections in developing countries in south and southeast Asia: a systematic review”. In: *The Lancet Infectious Diseases* 12.6 (2012), pp. 480–487.

[207] Wentian Li, Jan Freudenberg, and Pedro Miramontes. “Diminishing return for increased Mappability with longer sequencing reads: implications of the k-mer distributions in the human genome”. In: *BMC bioinformatics* 15.1 (2014), pp. 1–12.

[208] P. Jaccard. “Distribution de la flore alpine dans le bassin des Dranses et dans quelques rgions voisines”. In: *Bulletin de la Socit Vaudoise des Sciences Naturelles* 37 (1901), pp. 241–272.

[209] Travis C. Glenn. “Field guide to next-generation DNA sequencers”. In: *Molecular Ecology Resources* 11.5 (2011), pp. 759–769.

[210] F. Ozsolak. “Third-generation sequencing techniques and applications to drug discovery”. In: *Expert Opin Drug Discov* 7.3 (2012), pp. 231–43.

[211] T. Laver et al. “Assessing the performance of the Oxford Nanopore Technologies MinION”. In: *Biomolecular Detection and Quantification* 3 (2015), pp. 1–8.

[212] G. M. Pupo, R. Lan, and P. R. Reeves. “Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics”. In: *Proc Natl Acad Sci U S A* 97 (2000).

[213] Gabriela Delgado et al. “Genetic Characterization of Atypical Citrobacter freundii”. In: *PLoS ONE* 8.9 (2013), e74120.

[214] Simon R Harris et al. “Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study”. In: *The Lancet Infectious Diseases* 13.2 (2013), pp. 130–136.

[215] E. S. Snitkin et al. “Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing”. In: *Science Translational Medicine* 4.148 (2012), 148ra116–148ra116.

[216] Jon Louis Bentley. “Multidimensional binary search trees used for associative searching”. In: *Communications of the ACM* 18.9 (1975), pp. 509–517.

[217] C. A. Gilchrist et al. “Whole-genome sequencing in outbreak analysis”. In: *Clin Microbiol Rev* 28.3 (2015), pp. 541–63.

[218] Axel Visel et al. “ChIP-seq accurately predicts tissue-specific activity of enhancers”. In: *Nature* 457.7231 (2009), pp. 854–858.

[219] Yongchao Liu, Bernt Popp, and Bertil Schmidt. “CUSHAW3: Sensitive and Accurate Base-Space and Color-Space Short-Read Alignment with Hybrid Seeding”. In: *PLoS ONE* 9.1 (2014). Ed. by Oliver Hofmann, e86869.

[220] Guillaume Rizk and Dominique Lavenier. “GASSST: global alignment short sequence search tool”. In: *Bioinformatics* 26.20 (2010), pp. 2534–2540.

[221] Santiago Marco-Sola et al. “The GEM mapper: fast, accurate and versatile alignment by filtration”. In: *Nature Methods* 9.12 (2012), pp. 1185–1188.

[222] Paul Medvedev, Monica Stanciu, and Michael Brudno. “Computational methods for discovering structural variation with next-generation sequencing”. In: *Nature Methods* 6.11s (2009), S13–S20.

[223] T. J. Treangen and S. L. Salzberg. “Repetitive DNA and next-generation sequencing: computational challenges and solutions”. In: *Nat Rev Genet* 13 (2012).

[224] Mehdi Pirooznia, Fernando S Goes, and Peter P Zandi. “Whole-genome CNV analysis: advances in computational approaches”. In: *Frontiers in genetics* 6 (2015).

[225] Can Alkan, Bradley P Coe, and Evan E Eichler. “Genome structural variation discovery and genotyping”. In: *Nature Reviews Genetics* 12.5 (2011), pp. 363–376.

[226] Mark A Batzer and Prescott L Deininger. “Alu repeats and human genomic diversity”. In: *Nature reviews genetics* 3.5 (2002), pp. 370–379.

[227] Nam-Hyuk Cho et al. “The Orientia tsutsugamushi genome reveals massive proliferation of conjugative type IV secretion system and host–cell interaction genes”. In: *Proceedings of the National Academy of Sciences* 104.19 (2007), pp. 7981–7986.

[228] Budd A Tucker et al. “Exome sequencing and analysis of induced pluripotent stem cells identify the cilia-related gene male germ cell-associated kinase (MAK) as a cause of retinitis pigmentosa”. In: *Proceedings of the National Academy of Sciences* 108.34 (2011), E569–E576.

[229] M. Burrows and D. J. Wheeler. *A Block Sorting Lossless Data Compression Algorithm. Technical Report 124*. Palo Alto, CA: Digital Equipment Corporation, 1994.

[230] 1000 Genomes Project Consortium et al. “A map of human genome variation from population-scale sequencing”. In: *Nature* 467.7319 (2010), pp. 1061–1073.

[231] Federico Perez et al. “Global challenge of multidrug-resistant Acinetobacter baumannii”. In: *Antimicrobial agents and chemotherapy* 51.10 (2007), pp. 3471–3484.

[232] Pilar Villalón et al. “Epidemiology of the Acinetobacter-derived cephalosporinase, carbapenem-hydrolysing oxacillinase and metallo-β-lactamase genes, and of common insertion sequences, in epidemic clones of Acinetobacter baumannii from Spain”. In: *Journal of Antimicrobial Chemotherapy* 68.3 (2012), pp. 550–553.

[233] Anton Y Peleg, Harald Seifert, and David L Paterson. “Acinetobacter baumannii: emergence of a successful pathogen”. In: *Clinical microbiology reviews* 21.3 (2008), pp. 538–582.

[234] Mariana Pagano, Andreza Francisco Martins, and Afonso Luis Barth. “Mobile genetic elements related to carbapenem resistance in Acinetobacter baumannii”. In: *brazilian journal of microbiology* 47.4 (2016), pp. 785–792.

[235] Han-Yueh Kuo et al. “Insertion sequence transposition determines imipenem resistance in Acinetobacter baumannii”. In: *Microbial Drug Resistance* 20.5 (2014), pp. 410–415.

[236] Benjamin A Evans and Sebastian GB Amyes. “OXA β-lactamases”. In: *Clinical microbiology reviews* 27.2 (2014), pp. 241–263.

[237] Kai-Chih Chang et al. “Transcriptome profiling in imipenem-selected Acinetobacter baumannii”. In: *BMC genomics* 15.1 (2014), p. 815.

[238] Jane F Turton et al. “The role of ISAba1 in expression of OXA carbapenemase genes in Acinetobacter baumannii”. In: *FEMS microbiology letters* 258.1 (2006), pp. 72–77.

[239] Yangsoon Lee et al. “A novel insertion sequence, ISAba10, inserted into ISAba1 adjacent to the blaOXA-23 gene and disrupting the outer membrane protein gene carO in Acinetobacter baumannii”. In: *Antimicrobial agents and chemotherapy* 55.1 (2011), pp. 361–363.

[240] Jennifer K Mak et al. “Antibiotic resistance determinants in nosocomial strains of multidrug-resistant Acinetobacter baumannii”. In: *Journal of antimicrobial chemotherapy* 63.1 (2008), pp. 47–54.

[241] Zheng Zhang et al. “A greedy algorithm for aligning DNA sequences”. In: *Journal of Computational biology* 7.1-2 (2000), pp. 203–214.

[242] Hua Zhou et al. “Genomic analysis of the multidrug-resistant Acinetobacter baumannii strain MDR-ZJ06 widely spread in China”. In: *Antimicrobial agents and chemotherapy* 55.10 (2011), pp. 4506–4512.

[243] Jan Walther-Rasmussen and Niels Høiby. “OXA-type carbapenemases”. In: *Journal of Antimicrobial Chemotherapy* 57.3 (2006), pp. 373–383.