2012-04-18

Species Identification and Strain Attribution with Unassembled Sequencing Data

Owen Eric Francis
Brigham Young University - Provo

Species Identification and Strain Attribution with Unassembled Sequencing Data

Owen Francis

A selected project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Dr. G. Bruce Schaalje, Chair
Dr. Del T. Scott
Dr. Keith A. Crandall

Department of Statistics

Brigham Young University

June 2012

ABSTRACT

Species Identification and Strain Attribution with Unassembled Sequencing Data

Owen Francis
Department of Statistics, BYU
Master of Science

Emerging sequencing approaches have revolutionized the way we can collect DNA sequence data for applications in bioforensics and biosurveillance. In this research, we present an approach to construct a database of known biological agents and use this database to develop a statistical framework to analyze raw reads from next-generation sequence data for species identification and strain attribution. Our method capitalizes on a Bayesian statistical framework that accommodates information on sequence quality, mapping quality and provides posterior probabilities of matches to a known database of target genomes. Importantly, our approach also incorporates the possibility that multiple species can be present in the sample or that the target strain is not even contained within the reference database. Furthermore, our approach can accurately discriminate between very closely related strains of the same species with very little coverage of the genome and without the need for genome assembly - a time consuming and labor intensive step. We demonstrate our approach using genomic data from a variety of known bacterial agents of bioterrorism and agents impacting human health.

Keywords: Next-generation sequencing, bioforensics, biosurveillance, Bayesian mixture model

CONTENTS

Contents . . . iii

1 Introduction . . . 1

2 Literature Review . . . 3

2.1 Current Methods of Bacterial Identification . . . 3

2.2 DNA Sequencing Technologies . . . 4

2.3 Sequence Alignment Algorithms . . . 5

2.4 Expectation Maximization Algorithm . . . 6

3 Strain Attribution Model . . . 8

3.1 Mapping the reads to the genome database . . . 8

3.2 Bayesian reassignment method . . . 12

4 Methods . . . 19

4.1 Simulation Study 1 . . . 19

4.2 Simulation Study 2 . . . 20

4.3 Applications . . . 21

5 Results . . . 23

5.1 Simulation Study 1 . . . 23

5.2 Simulation Study 2 . . . 23

5.3 *E. coli* O104:H4 . . . 24

5.4 *B. anthracis* A2012 . . . 27

5.5 Combination of read sets 27

6 Conclusion 29

Bibliography 30

CHAPTER 1

INTRODUCTION

The accurate and rapid identification of species and strains is an essential component of biosurveillance from both human health and biodefense perspectives (Vaidyanathan 2011). For example, misidentification was among the issues that resulted in a three week delay in accurate diagnosis of the recent outbreak of hemorrhagic *Escherichia coli* being due to strain O104:H4. This resulted in over 2,400 infections and 23 deaths across 13 countries in Europe (Turner 2011). The most accurate genealogical information, necessary for species identification and strain attribution, comes from the most refined level of biological data – genomic DNA sequences (Eppinger et al. 2011). Advances in DNA sequencing technologies allows for the rapid collection of extraordinary amounts of such data, yet robust approaches to analyze these new kinds of data are just developing, from both statistical and algorithmic perspectives.

Next-generation sequencing approaches have revolutionized the way we collect DNA sequence data, including for applications in bioforensics and biosurveillance. However, these sequencing technologies produce errors with rates that vary by approach and sample. Such errors are typically less important for species identification given the relatively larger genetic divergences between species than between individuals within species. But for strain attribution, sequencing error has the potential to swamp out the genealogical signal in a data set. Furthermore, current approaches for next-generation sequencing have read lengths between 25-800 base pairs. To obtain a whole genome from a sample, the sequenced reads requires a computationally difficult and lengthy process, called assembly, and a large number of reads to achieve an adequate coverage level. Alternatively, one can map the reads back to a database of known genomes. However, most of the sequence data from closely related species or strains matchs locations in multiple genomes. Assembly and mapping approaches are further complicated by the fact that data collection at a crime scene or hospital might include additional environmental sources of DNA in the biological sample (naturally occurring bacterial and viral species), necessitating the need for highly sensitive and refined computational models for both species and strain attribution.

Here we develop and describe an approach to analyze next-generation sequence data for species identification and strain attribution that capitalizes on a Bayesian statistical framework that accommodates information on sequence quality, mapping quality, and provides posterior probabilities of matches to a reference database of known genomes. We show that our approach can discriminate between closely related strains of the same species with less than 1X coverage of the genome. We use simulation studies and applications to demonstrate our approach using genomic data from a variety of known bacterial agents of bioterrorism and agents impacting human health.

CHAPTER 2

LITERATURE REVIEW

2.1 Current Methods of Bacterial Identification

Identifying the exact strain of a bacterial species may seem inconsequential, but when the distinction is between the beneficial strains of *E. coli*, commonly found in the lower intestines of warm-blooded organisms, and the pathogenic strains such as the O104:H4 strain, which caused over 2,400 illnesses and 23 deaths during the summer of 2011, incorrect identification can have dire consequences. Unfortunately, many current methods for identifying bacteria are unable to differentiate among closely related species or strains.

The most common method for identifying bacterial strains involves the use of molecular assays and other bacteriological tests (Eppinger et al. 2011). In some cases, the results of several tests must be combined to differentiate between closely related species; however, even this approach has had limited success differentiating between *Bacillus anthracis*, the causative agent of anthrax, and its benign relative *Bacillus cereus* (Pilo and Frey 2011). Many of the problems associated with identification methods based on phenotypic attributes of bacteria can be alleviated by examining the most refined level of biological data – genomic DNA sequences (Eppinger et al. 2011).

The identification of bacterial strains using genetics has been common for several years, but efforts so far have focused on a specific segment of genetic material (the 16S ribosomal RNA), that is highly conserved within various species of bacteria. However, these methods are incapable of differentiating between strains within a species. Recently, DNA-based identification methods have moved toward the use of multiple genomic elements to differentiate between strains of a bacterial species; however, the current approach requires well-defined phylogenies, which renders the approach impractical for rapid identification (Boykin et al. 2011).

The ideal identification method would incorporate information from every gene, instead of a few selected genes, and require little phylogenetic information of the candidate genomes. However, an approach using every gene is complicated by the fact that bacteria have a remarkable ability to rapidly acquire new genetic code through horizontal gene transfer (HGT). HGT occurs in bacteria through the transfer of plasmids, DNA molecules that are separate from the chromosomal DNA. HGT is more common between bacteria of the same species or closely related species, but some bacteria have shown the ability for HGT with distantly related species (Brown et al. 2003). HGT is a serious problem because transferred genetic elements may contain genes that give bacteria greater disease causing capabilities or genes that give the bacteria higher medication resistance (Eppinger et al. 2011). Identifying when genes have been acquired can be important because determining if a specific pathogen has acquired virulence factors or medication resistances could profoundly affect the optimal containment or treatment strategies.

2.2 DNA Sequencing Technologies

DNA sequencing describes any of several different methods for determining the order of the nucleotide bases in a DNA molecule. Modern DNA sequencing began in 1977 with the chain-termination method of Sanger (Shendure and Ji 2008). Sanger sequencing can produce accurate sequences up to 1,000 nucleotides long; however, it is time consuming and relatively expensive (Shendure and Ji 2008).

The "next-generation" of sequencing technology began in 2005 with the 454 Life Sciences' Genome Sequencer, followed rapidly by the llumina Genome Analyzer, the SOLiD platform, the Polonator, and HeliScope's Single Molecule Sequencer technology (Shendure and Ji 2008). These technologies were significantly better than Sanger sequencing in speed and cost, but initially suffered from shorter sequence length and lower accuracy. No next-generation sequencers achieved the quality or length of Sanger sequencing. Third-generation sequencing technologies began with the 2010 release of Pacific Biosciences PacBio RS platform and Ion Torrent's ion semiconductor sequencing technology (Rusk 2011).

Each of the next- and third-generation DNA sequencers rely on different technologies, and vary widely in terms of sequence quality and length. In general, technologies that produce longer sequences tend to have higher error rates. There is no universally superior technology because of trade-off. The decision among current technologies is based on the specific challenges and goals of each application (Shendure and Ji 2008).

2.3 Sequence Alignment Algorithms

Current sequencing technologies can produce up to millions of sequences in a single run. Finding the regions of the genome from which these reads came involves aligning the reads to the genome, a computationally expensive and time consuming process. Finding optimal alignments for millions of reads is nearly impossible. To alleviate this problem, efficient algorithms have been developed to rapidly search the genome for possible matches. These algorithms are not guaranteed to find the optimal alignment for every read; however, the gains in efficiency compensate for the occasional non-optimal alignments (Ruffalo et al. 2011).

Many genomes contain highly repetitive regions where identical sequences occur several times in succession. When sequenced reads originate from these regions, the alignment algorithms can identify every possible location that the read could have come from, but no unique position can be identified as the origin of the read. The alignment algorithms handle such reads in different ways. Some algorithms discard these reads, effectively ignoring the data contained in these reads. Others randomly choose one matching location and assign the read, reporting only the randomly selected location. According to Clement et al. (2009), the most efficient method is to report every matching location and provide a measure of how strong the match is at each location. The Genomic Next-generation Universal MAPper

(GNUMAP) algorithm and reports the posterior probability that each read came from each location to which it matched.

2.4 Expectation Maximization Algorithm

The Expectation Maximization (EM) algorithm was formally introduced by Dempster et al. (1977); however, special cases of the EM algorithm had been proposed many times by previous authors. The 1977 paper provided a general form of the algorithm which extended it beyond the exponential family. The original paper also included a convergence analysis; however, the proof used was flawed and a corrected proof was published by Wu (1983).

Maximizing the likelihood function for a specific statistical model is usually a straightforward process, but it becomes more complicated when the likelihood is dependent on a set of observed data $\mathbf{Y} \in \mathcal{Y}$, a set of unobserved data $\mathbf{X} \in \mathcal{X}$, and a vector of unknown parameters $\boldsymbol{\theta}$. In these cases, the maximum likelihood estimates can be obtained using the marginal likelihood of the observed data. The likelihood becomes

$$L(\boldsymbol{\theta}; \mathbf{Y}) = f(\mathbf{Y}|\boldsymbol{\theta}) = \sum_{\mathbf{X} \in \mathcal{X}} f(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta}), \tag{2.1}$$

where $L(\cdot)$ is the likelihood function and $f(\cdot)$ is the joint density function. Define the function

$$Q(\boldsymbol{\theta}'|\boldsymbol{\theta}) = E(\log L(\boldsymbol{\theta}'|\mathbf{X})|\mathbf{Y}, \boldsymbol{\theta}), \tag{2.2}$$

which is assumed to exist for all pairs $(\boldsymbol{\theta}', \boldsymbol{\theta})$, and let $\boldsymbol{\theta}^{(p)}$ denote the current estimate of $\boldsymbol{\theta}$ after p iterations of the algorithm. The EM algorithm consists of two steps: the expectation step, or E-step, and the maximization step, or M-step.

- E-step: Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})$
- M-step: Choose $\boldsymbol{\theta}^{(p+1)}$ to be the value of $\boldsymbol{\theta}$ which maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})$

These steps are repeated until the estimates converge to stable values across iterations.

The heuristic idea is that direct maximization of $\log L(\boldsymbol{\theta}'|\mathbf{X})$ is made intractable by the presence of unobserved data, but the EM algorithm calculates $E(\log L(\boldsymbol{\theta}|\mathbf{X}))$ given the observed data $\mathbf{Y}$ and the current parameter estimates $\boldsymbol{\theta}^{(p)}$, and calculates $\boldsymbol{\theta}^{(p+1)}$ as the maximum likelihood estimates for the current expected value of $\log L(\boldsymbol{\theta}|\mathbf{X})$. By iteratively calculating the expected value of $\log L(\boldsymbol{\theta}|\mathbf{X})$ and updating the parameter estimates, the algorithm converges to the maximum likelihood estimates (Dempster et al. 1977).

Unfortunately, the EM algorithm is not guaranteed to converge to the maximum likelihood estimates if the likelihood is multimodal. In these cases the algorithm may converge to a local, instead of global, maximum of the likelihood function. Several methods such as random restart and simulated annealing have been used to prevent local convergence (Hastie et al. 2009).

The EM algorithm was extended to a Bayesian framework where the maximum a posteriori (MAP) estimates are calculated instead of the maximum likelihood estimates. Given the log of the prior density denoted $G(\boldsymbol{\theta})$, the E- and M-steps now become

- E-step: Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})$
- M-step: Choose $\boldsymbol{\theta}^{(p+1)}$ to be the value of $\boldsymbol{\theta}$ which maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) + G(\boldsymbol{\theta})$.

If the prior distribution is from the conjugate family for $f(\mathbf{X}|\boldsymbol{\theta})$, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) + G(\boldsymbol{\theta})$ will have the same functional form as $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})$ alone. This implies that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) + G(\boldsymbol{\theta})$ can be maximized in the same manner as $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})$, requiring only a slight adjustment to the maximization formulas (Dempster et al. 1977).

CHAPTER 3

STRAIN ATTRIBUTION MODEL

Next-generation sequencing technologies are powerful tools for the identification and determination of the species or strain of origin for a biological sample. These technologies sequence millions of small DNA fragments, or reads, that can be used to accurately identify their source genome. Our strategy for attributing sequencing reads to the correct organism within a database uses a Bayesian mixture model that reassigns reads that map to multiple genomes. A flowchart of our data analysis pipeline is given in Figure 3.1.

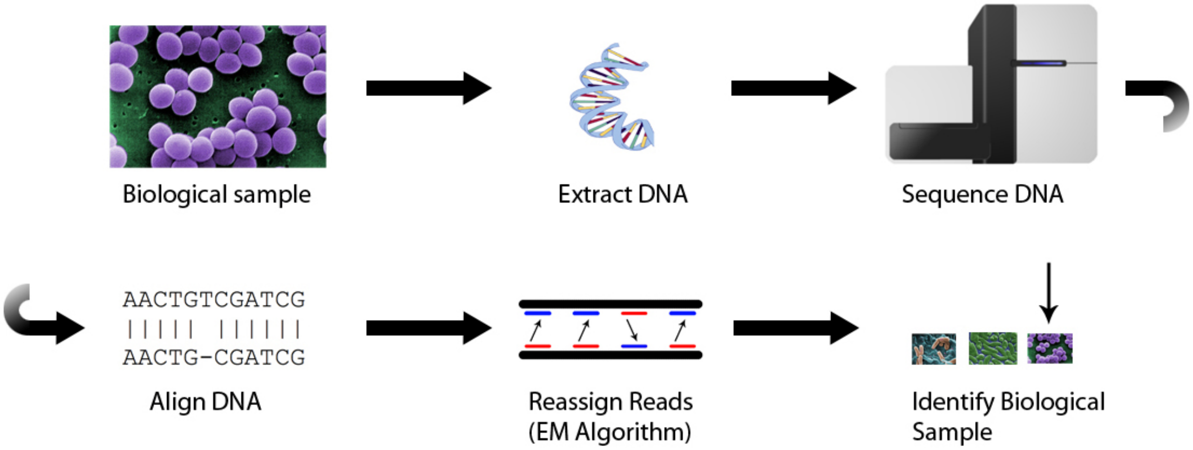

Figure 3.1: First, a biological sample is obtained. From this biological sample, DNA is extracted and sequenced. The sequenced reads are aligned to a database of known genomes. The non-unique reads are reassigned to the most likely genome via the EM algorithm. The results of the EM algorithm are summarized to identify the genome which is the most likely source of the set of reads.

3.1 Mapping the reads to the genome database

We have constructed a database containing 170 genomes (totaling 610 million base pairs of sequence) of bacterial agents of bioterrorism, agents impacting human health, as well as benign bacterial species that may resemble harmful species or may often be present in biological samples. The genomes were obtained from GenBank, and were chosen based on their similarity to eight potential bacterial agents of bioterrorism identified by the Centers for Disease Control CDC. Given a set of reads, to determine which of these genomes are potential sources of the reads, they must be compared to the database of genomes. This is done using sequence alignment algorithms, of which several have been developed.

Alignment Strategy Comparison

One of the biggest issues that alignment algorithms face is the assignment of reads which match multiple genomic locations. These reads, henceforth denoted as non-unique reads, can not be deterministically assigned to any single location in the genomes. Because the information provided by these reads is unclear, several strategies for dealing with non-unique reads have been implemented. The most common strategies are to delete non-unique reads, only report the locations with the closest matching sequence, or to randomly assign each non-unique reads to one of the locations to which it matched. For the purpose of species identification and strain attribution, we believe that all of these strategies loose valuable information. Here, the different strategies for non-unique reads are explained further in the context of a particularly difficult strain attribution example.

1. **Deletion**: Deleting reads with multiple alignments greatly reduces the power to identify the correct genome as it often disposes of a vast majority of the reads. Using this approach with reads originating from the MG1655 substrain of the *E. coli* K-12 strain, the identification method assigned only 15% of the reads to the correct genome and 18% of the reads to the W3110 substrain of the *E. coli* K-12 strain. The remaining unique reads were assigned to a variety of incorrect genomes. In cases where there are many closely related species and strains, almost all of the reads map to multiple genomes. The reads that map uniquely are usually low quality reads with many errors. Many of the unique reads align to incorrect strains that happen to have mutations at

the same positions as the sequencing errors. Therefore, this method often discards the most valuable data, and retains data that do not align to the correct genome.

2. **Best Match**: Reporting only the best match for each non-unique read is advantageous as it does not discard any of the reads, but it does disregard a considerable amount of information that is contained in the non-unique reads. The power to identify a single strain is increased with this method over the deletion approach, but it can result in the identification of the wrong strain. Using this approach with the same reads as in the previous *E. coli* K-12 MG1655 example, the identification method was again unable to identify the correct strain, assigning 90% of the reads to the W3110 substrain, and only 3% to the correct substrain.

3. **Assign All Matches**: If the reads are assigned to the genomes without reassignment, the correct genome often has the highest read probability, but many other genomes also have relatively high read probabilities. Not all of the genomes that have matches to the reads are contained in the sample, but without reassignment, the read probabilities are not transferred from the less likely genomes to the more likely genomes. Without reassignment, the MG1655 substrain received 8.9% of the read probability, but the W3110 substrain received a comparable 8.8% of the read probability. This method leaves a lot of uncertainty about the identification, and is not able to reliably identify the correct strain if there are multiple, unrelated, species in the sample.

4. **Random Assignment**: If a read is randomly reassigned to one of the genomes for which it had a match, the results are similar to using only the unique reads because the method forces an artificial uniqueness on all of the reads. However this method assigns many of the reads to incorrect genomes. This results in relatively low read probabilities spread across many genomes. In the *E. coli* K-12 MG1655 substrain example, if the reads were randomly reassigned to the matching genomes regardless of alignment quality, 5.14% of the reads were assigned to the W3110 substrain, and 5.02%

of the reads were assigned to the correct strain. Additionally, the BW2952 strain of *E. coli* received 4.78% of the read probability, and the DH10B substrain of the K-12 strain received 4.77% of the read probability. If the reads were reassigned only to the genomes with the best matches, the reads were spread across fewer genomes. However, the method still failed to find the correct genome with 12.54%, 12.16%, and 11.44% of the reads assigned to the W3110, MG1655, and DH10B substrains of the K-12 strain respectively and 11.39% to the BW2952 strain.

5. **Probabilistic Assignment**: If the reads are reassigned using all matches together with the quality scores for each match, none of the data are discarded, leading to accurate identifications of the correct genome. In the *E. coli* K-12 MG1655 substrain example, 99% of the reads were assigned to the correct genome, with 0.71% of the reads assigned to the *E. coli* BL21(DE3) strain, and 0.17% assigned to the W3110 substrain.

We believe that probabilistic assignment is the best alignment strategy for species identification and strain attribution. Therefore, we have chosen the GNUMAP algorithm because it implements a probabilistic assignment method.

GNUMAP aligns each read to the genomes by first creating a hash table of all the genomes in the database. The hash table contains a list of sequences and every location in the genome where the sequences occur. The specifics of how the hash table is created is covered in the paper by Clement et al. (2009). To align the sequences, the algorithm then takes segments of each read and compares them with the hash table to find regions of the genomes that could be the source of the read. The reads are then compared to the full genomic sequences at the identified regions, and the alignments are scored using a probabilistic Needleman-Wunsch algorithm which incorporates the quality information provided for each nucleotide in the FASTQ sequence format. This allows for GNUMAP to penalize reads more for mismatches where the base call is reliable, and less for mismatches where the base call is less certain. The sequence is also aligned to itself to establish maximum alignment score with which to compare the other alignments. Any alignment that does not pass a user-defined threshold is discarded. For this project, the threshold is set at 90% of the maximum alignment score. The scores for each recorded alignment are then converted to posterior probabilities. Given the alignment scores $S_1, S_2, \ldots, S_n$, the posterior probability assigned to the lth alignment, P_l, is computed as $P_l = \frac{\exp S_l}{\sum_{i=1}^{n} \exp S_i}$ (Clement et al. 2009). The posterior probability is interpreted as the probability that each alignment is the true source of the read, and is discussed in further detail in section 3.2.

3.2 Bayesian reassignment method

Using the GNUMAP probabilistic read alignment algorithm, the non-unique reads are probabilistically or partially assigned to two or more genomes in the database. However, each read can only have one true genomic source, so to identify the species present in the sample, the non-unique reads probabilities must be reassigned to the correct genome of origin. To properly reassign the reads, we have formulated a Bayesian mixture model that integrates information contained within the read (mapping probability) with information obtained by borrowing strength across all reads from the sample (e.g. proportions of unique reads or imbalances in non-unique probabilities across all reads). This approach is superior to a naive mapping approach that merely assigns reads based on information contained solely in the read as it helps to overcome mistakes in mapping caused by sequencing errors or low quality bases.

The Bayesian mixture model assumes that reads are drawn from a small subset of the genomes in the database and that each read is drawn from only one of those genomes. Parameters in the model represent the proportions of reads that originate from each genome as well as the proportion of the non-unique reads that are incorrectly assigned to each genome due to sequence similarity. Using a Bayesian missing data mixture model formulation (where the genome of origin is the missing data) the model re-weighs the read assignment probabilities using the mapping qualities and the parameters of the model. In the reassignment process, the parameters are designed to penalize the value of non-unique reads in the presence of unique reads and re-weight the non-unique reads based on overall mapping proportions when no reads map uniquely. For example, as illustrated in Figure 3.2, consider a case with 6 reads that map to 3 genomes, where 3 of the reads map uniquely and 3 reads are non-unique. In this situation, 5 reads come from the first genome, and 1 read comes from the third genome, and none of the reads originate from the second genome. However, due to genomic similarity half of the reads map to either two or three different genomes and therefore must be reassigned. In addition, we have a sequencing error in Read 2 that leads it to map to Genome 3 with higher mapping probability, even though the true source is Genome 1. Using only the original GNUMAP alignment for the reads, estimates of the genome proportions are be given by (0.51, 0.10, 0.39) for Genomes 1, 2, and 3 respectively, clearly showing a bias due to sequence similarity. However, after applying our Bayesian reassignment algorithm, the model correctly reassigns the non-unique reads to Genome 1, and correctly estimates the proportion of reads for the genomes as (0.83, 0.00, 0.17).

Bayesian reassignment mixture model

Once the reads are probabilistically or partially assigned to two or more genomes in the database, we need to reassign the reads to their true genomic source so the species or strains present in the sample can be distinguished or identified. Here we describe in detail the Bayesian mixture modeling approach for reassigning the reads to the correct genome. This approach combines information contained within the read (mapping probability) with information obtained by borrowing strength across all reads from the sample (e.g. proportions of unique reads or imbalances in non-unique probabilities across all reads). This approach is superior to a naive mapping approach which assigns reads based on information contained solely in the read as it helps to overcome mistakes in mapping caused by sequencing errors or low quality bases.

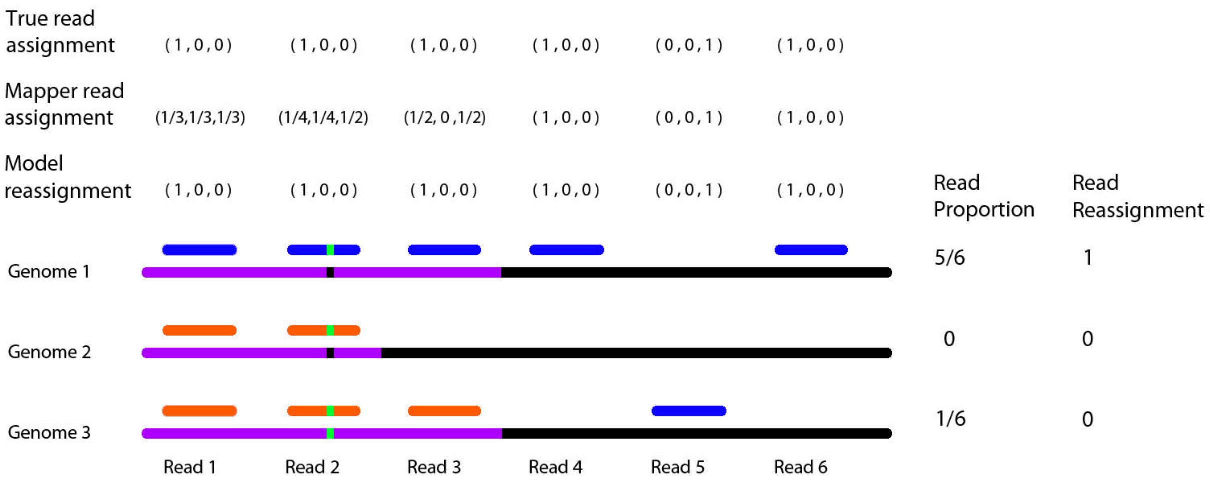

Figure 3.2: Illustration of read reassignment. Blue represents an alignment to the correct genome, and red represents an alignment to an incorrect genome. Reads 1-4 and 6 originated from Genome 1, and read 5 from Genome 3. The purple sections of the genomes represent regions of genome similarity between the species. Read 1 aligns to all 3 genomes, while read 3 aligns to genomes 1 and 3. Read 2 aligns to all 3 genomes, but there is a sequencing error which corresponds to a difference in genome 3, giving it a higher mapper read assignment. Using only the original GNUMAP alignment for the reads, estimates of the genome proportions would be given by (0.51, 0.10, 0.39) for Genomes 1, 2, and 3 respectively, clearly showing a bias due to sequence similarity. After applying the read reassignment model, the ambiguous reads are correctly assigned to genome 1, and the correct proportion of reads for each genome is correctly determined to be (0.83, 0.00, 0.17).

To formally describe our model, let $i = 1, \ldots, R$ index the reads and let $j = 1, \ldots, G$ index the genomes in the database. We let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iG}) = \{x_{ij}\}$ be a set of genome indicators for read i where $x_{ij} = 1$ if the read originated from the jth genome and $x_{ij} = 0$ if the read did not come from genome j. Note that by assumption one and only one element in the vector $\mathbf{x}_i$ can be equal to 1 (i.e. each read has only one source genome). We assume that $\mathbf{x}_i$ follows a multinomial distribution, with probability of success $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_G) = \{\pi_j\}$ where π_j is the proportion of the reads that originated from the jth genome.

For reads that align to only one genome we directly observe the genome indicator $\mathbf{x}_i$ for the ith read. However, due to the similarities in some genomes and the tolerance for closely matching sequences, some reads align to multiple genomes. In the case of these reads, the genome indicator $\mathbf{x}_i$ is unobserved or missing data. For the non-unique reads, what is observed are partial mapping qualities for each of the genomes. These mapping probabilities are provided as the posterior probabilities, which are scaled mapping quality or relative likelihood scores obtained from the GNUMAP algorithm. Specifically, for the ith read we denote these mapping scores by $\mathbf{q}_i = (q_{i1}, q_{i2}, \ldots, q_{iG}) = q_{ij}$. For unique reads, the q_{ij} values are equal to the x_{ij} values. For non-unique reads, these represent the uncertainty in mapping and need to be rescaled–or equivalently–these reads need to be reassigned to the correct genome of origin. To do this, we define a second set of parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_G) = \{\theta_j\}$ where θ_j is a reassignment parameter that represents the proportion of the non-unique reads that need to be reassigned to the jth genome.

In order to simplify the notation in the likelihood function, we define y_i as the uniqueness indicator for read i, namely letting $y_i = 1$ if read i is unique and $y_i = 0$ otherwise. Under the modeling assumptions above, the complete data likelihood of the parameters $(\boldsymbol{\pi}, \boldsymbol{\theta})$ given the observed data (reads, y_i, unique $\mathbf{x}_i$) and the missing data (non-unique $\mathbf{x}_i$) is given by:

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\theta} | \mathbf{x}_i, \mathbf{q}_i, \mathbf{y}) \propto \prod_{i=1}^{R} \prod_{j=1}^{G} \left[\boldsymbol{\pi}_j \boldsymbol{\theta}_j^{1-y_i} q_{ij} \right]^{x_{ij}} \tag{3.1}$$

Although the reassigned reads (estimated $\mathbf{x}_i$) and reassignment parameters (estimated $\boldsymbol{\theta}$) are very informative, the quantity of interest from the modeling steps are the estimates for the genome read proportions (estimated $\boldsymbol{\pi}$). These probabilities identify the single or multiple organisms from the database that are present in the samples, based on the proportion of the reads that are assigned to each genome after the reads are reassigned. In the case where only one bacteria is present in the sample, we identify the most likely genome as $\hat{j} = \text{argmax}(\pi_j)$. In cases where multiple bacteria are present in a sample, the values of π_j are expected to roughly correspond to the proportion of reads from each of the bacteria.

Bayesian Prior Distributions

Both $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are assumed to follow Dirichlet distributions, which are,

$$f(\boldsymbol{\pi}, \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{G} \alpha_j)}{\prod_{j=1}^{G} \Gamma(\alpha_j)} \prod_{j=1}^{G} \pi_j^{\alpha_j - 1} \tag{3.2}$$

$$f(\boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{\Gamma(\sum_{j=1}^{G} \beta_j)}{\prod_{j=1}^{G} \Gamma(\beta_j)} \prod_{j=1}^{G} \theta_j^{\beta_j - 1}. \tag{3.3}$$

If $\alpha_j = 1$ for all $j = 1, \ldots, G$, this is equivalent to adding one unique read for each of the G genomes, and $\alpha_j = n$ would be the equivalent of adding n unique reads to the jth genome. Similarly, $\beta_j = n$ is the equivalent of adding n reads worth of non-unique reads to the jth genome. However, the prior information for $\boldsymbol{\theta}$ does not behave like true non-unique reads because it is not subject to reassignment. Prior information assigned to each genome is always associated with that genome, but its effect is diminished as the number of reads increases. This can be seen clearly in the maximization formulas in equations 3.5 and 3.6. The prior information stabilizes the algorithm by preventing the estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ from converging to the boundaries of the Dirichlet distribution. Inclusion of prior information biases the results, possibly even leading to the identification of the wrong genome. However, this only happens in rare circumstances, and it requires initially favoring some genomes above others. To avoid this, all of the j genomes receive the same values for its priors. That is $\alpha_j = \alpha_{j'}$ and $\beta_j = \beta_{j'}$ for $j \neq j'$.

Model estimation via the EM algorithm

Computing the maximum likelihood estimates for $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ is complicated by the fact that x_{ij} is unobserved. Therefore, we use the EM algorithm to simplify the calculation of the estimates. To start the EM algorithm, initial estimates of the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are proposed, usually $\pi_j = \theta_j = \frac{1}{G}$, $\forall j$. In the E-step, the expected value of x_{ij} is computed for each combination of $i = 1, \ldots, R$ and $j = 1, \ldots, G$ based on the estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, as well as the observed data $\mathbf{q}_i$ and $\mathbf{y}$. In the E-step, the expected value of x_{ij} is calculated as,

$$E(x_{ij}) = \frac{q_{ij}(\pi_j\theta_j)^{1-y_i}}{\sum_{k=1}^{G} q_{ik}(\pi_k\theta_k)^{1-y_i}}. \tag{3.4}$$

Next, the M-step calculates the new estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ given $\mathbf{q}_i$, $\mathbf{y}$ and the current expected values of $\mathbf{x}_i$. The formulas for estimating $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, provide the Bayesian maximum a posteriori (MAP) estimates; however, if the prior information $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is set equal to the zero vector, these equations provide the maximum likelihood estimates.

$$\tilde{\pi}_j = \frac{\sum_{i=1}^{R} x_{ij} + \alpha_j}{N + \sum_{k=1}^{G} \alpha_k} \tag{3.5}$$

$$\tilde{\theta}_j = \frac{\sum_{i=1}^{R}(1-y_i)^{x_{ij}} + \beta_j}{\sum_{i=1}^{R}(1-y_i) + \sum_{k=1}^{G} \beta_k}. \tag{3.6}$$

The E-step is then repeated using the updated estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, followed again by the M-step. These steps are repeated until the expected value of x_{ij} and the estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ converge to stable values across iterations.

Likelihood interpretation of the $\mathbf{q}_i$ *mapping probabilities*

It is interesting to note that the q_{ij} mapping probabilities can be interpreted as proportional likelihood values, namely the likelihood that read i is aligned to the jth genome and scaled so that the likelihood values sum to 1 across all genomes. Specifically, the alignment score derived from GNUMAPs Probabilistic Needleman-Wunsch (PNW) algorithm gives the log-likelihood of alignment based on a continuous product negative multinomial distribution. The GNUMAP PNW default alignment scoring function: +1 for a base match, -2 for a transition, -3 for a transversion, -4 gap penalty, which is equivalent to having negative multinomial likelihood "success" probabilities/parameters of: 0.96 for a match, 0.01 transition, 0.005 transversion, and 0.001 for gap. Although these parameters can be user-specified, these default values have been carefully selected to represent the mutation rates found in nature plus some noise for sequencing error. Note that by fixing these values in the alignment, we are effectively fixing the conditional likelihood of the reads given the genomes and scoring/alignment function. Therefore, it not necessary to include the raw read data in Equation 3.1 above, because the $\mathbf{q}_i$'s are sufficient statistics for the reads and can be used in place of the reads in the likelihood.

Homogeneity of the reassignment parameter ($\boldsymbol{\theta}$)

In this formulation of the model, there is a single θ parameter for each genome. If there is a single species in the biological sample from which the reads were extracted, then this is the correct formulation of the model. The single genome present in the sample should have a θ value of 1 and the genomes not present should have θ values of 0, which is consistent with the observed performance of the model. However, if there are multiple species in the sample, this becomes a simplifying assumption. The definition could be relaxed to allow θ to be a function of the region from which the read originated, $\theta_j(z)$. In unique regions of the genome, regions that are not shared by other species in the database, $\theta_j(z)$ would equal 1 because reads from that region would not be found in other species, and any alignment to another species would be due to alignment error. Therefore, all reads from these regions would be reassigned to the genome from which it originated. If the read originated from a region that was shared by multiple species within the database, the value of $\theta_j(z)$ would be divided between the k genomes that shared that region such that $\theta_j(z) = \frac{1}{k}$. However, relaxing this assumption requires the genomic locations to which the reads aligned and information on which regions of each genome are shared with other genomes. This information is available, but we are currently using the average value of $\theta_j(z)$ across all regions of the each genome. This approach is computationally more straightforward, and doesn't appear to hinder the performance of the model.

CHAPTER 4

METHODS

The performance of the method was examined through simulation studies and applications of actual sequencing reads. The read sets used were gathered from the 2011 *E. coli* outbreak in Europe and 2001 anthrax attacks (Read et al. 2002). An additional read set was created by combining read sets from 3 species of bacteria. This provided information on how well the method performs in real situations, and under conditions in which the method might fail.

4.1 Simulation Study 1

The first simulation study was based on the *B. anthracis* CDC 684 genome. We simulated reads from the genome, and aligned them to the database of known genomes. We then applied the reassignment method to the results of the alignment and identified the most likely genome of origin. This process was iteratively repeated to provide a distribution of results.

To create the reads for the simulation studies, 200 base pair segments were sampled from various locations within the target genome by randomly drawing numbers from a discrete uniform distribution. These numbers correspond to the starting positions in the genome from which the reads were generated. The full reads were generated by taking the 200 nucleotides following each of the starting positions. As these 200 base pair segments were sampled, errors were introduced into the reads by giving each nucleotide a chance of being changed to a different nucleotide based on a Bernoulli random draw.

For this simulation study, the error rate was set at 5%. This error rate was chosen based on the findings of a study by Margulies et al. (2005), which observed an error rate of about 3.3% using the Roche 454 sequencing technology. The coverage was set at 1X, meaning that the number of base pairs in the reads are approximately equal to the number of base pairs in the genome. The error rate and coverage level were chosen to provide a circumstance which is similar to, or worse than what would be expected in real situations.

The *B. anthracis* CDC 684 genome is approximately 5.3 million base pairs long (5.3 megabase), therefore 26,524 of the 200 base pair reads were generated to provided approximately 1X coverage of the genome. The simulated reads were then aligned to a database containing the gnomes of 6 different strains of *B. anthracis*, which are CDC 684, Ames, Ames Ancestor, Sterne, A0248, and CI (RefSeq IDs NC_012581, NC_003997, NC_007530, NC_005945, NC_012659, and NC_014335 respectively). Our reassignment method was then applied to the results of the alignment to determine which strain is the most likely source of the reads. This process was repeated on 100 randomly generated read sets.

4.2 Simulation Study 2

For the second simulation study, a target genome was simulated by selecting a 50,000 base pair segment of the *B. anthracis* CDC 684 genome. A second simulated genome was then created from the first simulated genome by inverting a 5,000 nucleotide long segment near the middle of the genome. Due to the behavior of the alignment algorithm, this provides two sites where the differences between these two genomes can be detected, and therefore, only the reads that overlap one of the boundaries of the inverted region were informative for detecting the strain of origin.

Two hundred base pair reads were simulated from the target genome using the same process described above, and miscalled bases were randomly introduced at rates of 1%, 2%, 3%, 4%, and 5%. These error rates were selected to resemble the standard range of error rates that have been observed in high throughput sequencing technologies. The reads were generated in sets of 250, 500, 1,000 and 2,500 reads yielding 1, 2, 4, and 10X coverage of the genome respectively. These levels of coverage were chosen because the definitive method of genome identification, full assembly, requires around 30X coverage of the genome to give an accurate assembly. The simulated reads were then aligned to a database containing the 2 simulated gnomes. Our reassignment method was then applied to the results of the alignment to determine which of the two genomes is the most likely source of the reads. One thousand read sets were simulated at each of the 20 error rate and read coverage combinations.

4.3 Applications

The identification method was applied to 3 read sets from actual genetic sequencing runs.

The first involved the recent outbreak of *E. coli* O104:H4 in Europe which resulted in a number of deaths that may have been prevented by an early identification of the affecting parasite. We obtained 92,370 reads from the first sequencing run of this strain by Beijing Genome Institute (BGI) using the Ion Torrent technology. The analysis of these reads was done in two stages. First, we aligned the reads to a database containing more than 170 genomes including the genomes of 29 strains of *E. coli.* However, the correct strain was not contained in the database. This corresponded to the observed situation, where the threat was unidentifiable because no reference genome had been assembled. In the second stage, the partial genome assembled by BGI was added to the genome database, and the reads were realigned to the expanded database. This corresponded to a situation where an illness is caused by a bacteria with an assembled genome.

The second application involved *B. anthracis*, more commonly known as "anthrax," a previously used weapon for bioterrorism. Specifically, we examined the anthrax attacks of 2001, where anthrax sent through the U.S. Postal Service infected 22 people. Of these 22 individuals, 5 individuals died and at least 7 of the survivors had confirmed cutaneous anthrax disease. We obtained 120 reads from white powder which was contained in the letter to Rep. Tom Daschle during the 2001 terrorist attacks (Read et al. 2002). We remapped the reads to our database and apply our reassignment algorithm to determine the anthrax strain that was most likely used in the attacks.

The third application involves read sets from the Sequencing Read Archive (SRA), a database of raw sequencing data maintained by the National Center for Biotechnology Information (NCBI). From the SRA database, we have downloaded read sets originating from *E. coli* K-12 MG1655, *Francisella tularensis* holarctica OSU18, and *Yersinia pestis* KIM D27. We have combined these read sets into one artificial read set. This resulted in 490,416 reads with 29.3%, 5.8%, and 64.9% of the reads originating from the *E.coli*, *F. tularensis*, and *Y. pestis* strains respectively. This corresponds to a read set with multiple species where the relative concentrations of each of the species is known. We mapped the reads to our full database and applied our reassignment algorithm. The goal is to recover the 3 species used and proportion of reads which came from each species.

Through these simulation studies and applications of real data, we show that our method is highly accurate at identifying the correct genome of origin and is capable of recovering the correct proportion of reads in a sample with multiple species.

CHAPTER 5

RESULTS

5.1 Simulation Study 1

All 100 replicates in simulation study 1 correctly identified the *B. anthracis* CDC 684 strain as the source of the reads. The average value of $\tilde{\pi}_j$ for the *B. anthracis* CDC 684 genome was 99.05%, and the minimum value of $\tilde{\pi}_j$ (see equation 3.5) for *B. anthracis* CDC 684 over the 100 replicates was 98.89%. Thus the method appears capable of consistently differentiating between closely related strains under substandard conditions.

5.2 Simulation Study 2

The results of simulation study 2 are given in Figure 5.1. The example with 1% read errors and 10X coverage had near perfect accuracy, successfully identifying the correct genome 99.7% of the time. However, the accuracy rate reduced as the base calling error rate increased and as the coverage decreased. On average, the accuracy was reduced by 5.8%, 7.9%, 9.4%, and 12.9% as the error rate increased from 1% to 2%, 3%, 4%, and 5%, respectively.

Greater coverage increases the probability that reads cover the two sites where the difference between the genomes is detectable. Thus, the accuracy of identifying the correct genome should decrease as the the coverage decreases. The accuracy was reduced by an

Table 5.1: Summary statistics for simulation study 1. The amount of reads assigned to the correct genome ranged between 98.89% and 99.25%.

Minimum	98.89%
1st Quartile	99.00%
Median/Mean	99.05%
3rd Quartile	99.10%
Maximim	99.25%

average of 9.4%, 17.7%, and 25.0% as the coverage decreased from 10X to 4X, 2X, and 1X, respectively. It also appears that increasing the number of reads reduces the impact of the changes in the error rate, suggesting that even poor quality reads can reliably identify the correct genome provided that the coverage is high enough.

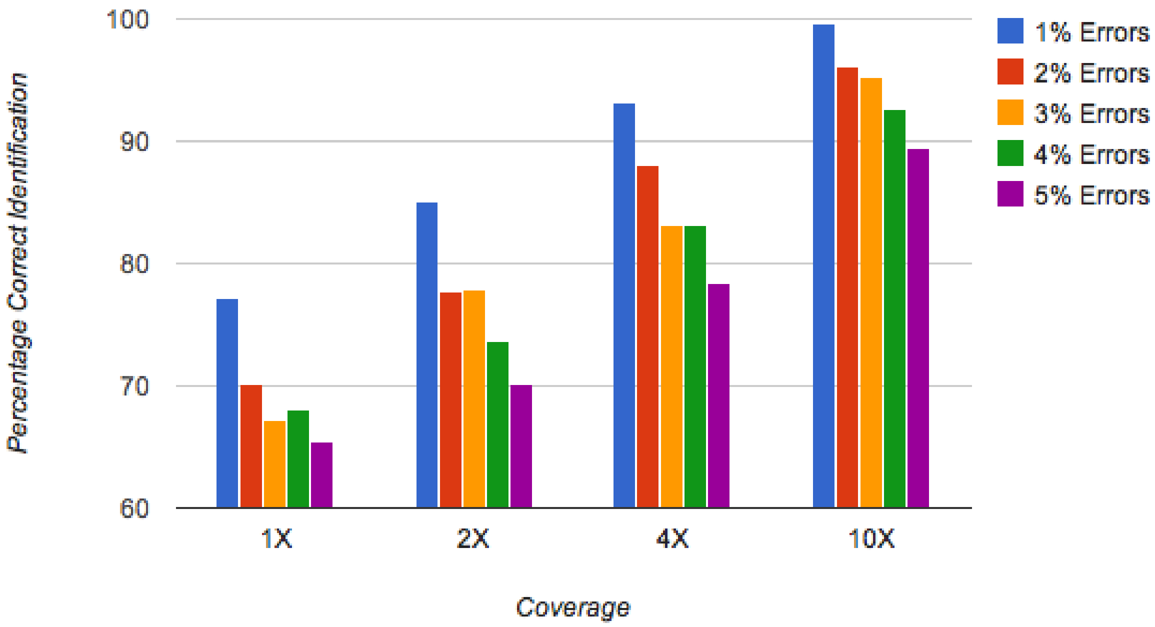

Figure 5.1: Identification rates from simulation 2, based on 1,000 replicates at each coverage and error combination. Identification of the correct genome generally increases as the error rate decreases. Increasing coverage leads to a greater probability of identifying the correct genome.

5.3 *E. coli* O104:H4

To motivate the need for the reassignment method, we first mapped the original 92,370 reads to our database of 170 genomes which did not contain the correct genome. This mapping resulted in 62,694 (67.9%) of the reads aligning to at least one genome in the database. Most of these reads, 62,583 (99.8%) mapped to at least one strain of *E. coli*, thus clearly identifying the species of origin. However, 60,272 (96.1%) of the reads aligned non-uniquely

Table 5.2: Results from *E. coli* O104:H4 when the correct genome of origin was not present in the database. Percentages of reads assigned to each genome are given before and after reassignment. In the absence of the correct genome, the method does not converge on one genome.

E. coli strain	Before Reassignment	After Reassignment
55989	9.5 %	89.5%
O103:H2 12009	3.2 %	1.1 %
B7A	3.2 %	0.4 %
O26:H11 11368	3.1 %	0.4 %
E24377A	3.1 %	0%
E22	3.1 %	0.5 %
SE11	3.0 %	0%
IAI1	3.0%	0%
E110019	2.9 %	0%
O111:H-11128	2.9 %	0%
B171	2.8 %	0.9 %
HS	2.0 %	0%
All others	58.2%	7.2%

to multiple *E. coli* strains. Before applying our reassignment method, the 55989 strain received the largest proportion of the aligned reads (9.5%), followed by the O103:H2 strain (3.2%), the B7A, O26:H11, E24377A, and the E22 strains (3.1%), then the SE11 and IAI1 strains (3.0%).

After application of our Bayesian reassignment model, 89.5% of the reads were reassigned to the 55989 strain of *E. coli*, which has been identified by Turner (2011) as being the most closely related fully sequenced genome to the correct O104:H4 strain. The genomes with the next highest read proportions were the O157:H7 strain (3.2%), and the O103:H2 strain (1.1%). Therefore, our approach did not completely converge on one genome, as it shouldn't because the origin strain was not present in the database. However, it is clear that our approach definitively identifies the closest fully sequenced neighboring strain with high confidence.

We then added to our database the partially assembled contiguous sequences (contigs) available from BGI Turner (2011) for the O104:H4 strain, and remapped the reads. After

Table 5.3: Results from *E. coli* O104:H4 when the correct genome of origin was present in the database. Percentages of reads assigned to each genome are given before and after reassignment. With the correct genome present, the method identifies the O104:H4 strain as the most likely.

E. coli strain	Before Reassignment	After Reassignment
O104:H4	12.0%	98.7%
55989	7.0%	0.8 %
O103:H2	2.8 %	0%
E24377A	2.8 %	0%
B7A	2.8%	0%
O26:H11	2.7%	0%
SE11	2.7%	0%
IAI1	2.7%	0%
All Others	64.5%	0.5

alignment, an additional 1,202 reads aligned uniquely to the O104:H4 strain, resulting in 63,896 of the 92,370 reads aligning to at least one genome, and 98.0% of the 63,896 reads aligning to multiple genomes. In the non-reassigned GNUMAP read alignment, the O104:H4 strain received the largest proportion of the aligned reads (12.0%), followed closely by the 55989 strain (7.0%), the O103:H2, E24377A, and the B7A strains (2.8%), then the O26:H11, SE11, and IAI1 strains (2.7%). After Bayesian reassignment, 98.7% of the read probability was assigned to the O104:H4 strain, 0.8% to the 55989 strain, and 0.1% to the O157:H7 strain. All other strains were reassigned less than 0.1% of the reads. The read probability (π_j, eq. 3.5) is close to what we would expect when the correct genome is present (i.e. ~99%). Despite the fact that the O104:H4 genome was not fully assembled, the reassignment method was able to assign a vast majority of the reads to the correct genome. Therefore, based on this knowledge we are confident that our approach is a highly sensitive procedure for identifying the exact strain of origin or its closest related neighbor in the database. In addition, we note that whether database includes the genome or not, we were able to identify the correct or nearest strain using only 63,896 mapped reads, which we estimate to total only 1.3X coverage of the species of origin.

Table 5.4: Percentages if reads assigned to genomes before and after reassignment in the *B. anthracis* A2012 example. Before reassignment the reads are shared between many strains, but after reassignment our method identifies the A2012 strain with only 120 reads.

Anthrax strain	Before Reassignment	After Reassignment
A2012	8.8%	98.3%
A0389 BAK	6.5%	0.8%
Australia 94	5.6%	0%
A0193 BAQ	5.3%	0%
All others	73.8%	0.9%

5.4 *B. anthracis* A2012

In the Daschle whit powder example, we observed that before reassignment, 8.8% of the reads were assigned to the A2012 anthrax strain, 6.5% of the reads were assigned to the A0389 BAK anthrax strain, 5.6% of the reads were assigned to the Australia 94 anthrax strain, and 5.3% of the reads were assigned to the A0193 BAQ anthrax strain. After reassignment, 98.3% of the reads were assigned/reassigned to the A2012 strain, which according to previous research (Read et al. 2002), is the correct genome of origin for these reads. This result illustrates the utility of our model, which (within a few hours) can identify the correct genome of origin without the need to perform comparative genomics (Read et al. 2002).

5.5 Combination of read sets

When the reads were aligned to the genome database, only 198,083 of the 490,416 reads had suitable matches to the genomic sequences. Of the 198,083 reads, 72.2%, 3.5%, and 24.3% originated from the *E.coli*, *F. tularensis*, and *Y. pestis* strains, respectively. This resulted in mapping proportions which were different from the true proportions 29.3%, 5.8%, and 64.9% for *E. coli*, *F. tularensis*, and *Y. pestis*, respectively. This was because relatively few of the reads, 24.4% and 15.1%, from the *F. tularensis*, and *Y. pestis* strains respectively were matched by GNUAMP to the genome database; however, most reads from the *E. coli* strain were matched to the genome database (99.5%). Before reassigning the reads 6.4%

Table 5.5: Percentages if reads assigned to genomes before and after reassignment in the read set combination example. The correct species in the sample is not apparent until after reassignment.

Strain	Before Reassignment	After Reassignment
E. coli K-12 MG1655	6.4%	63.6%
E. coli K-12 W3110	6.4%	0%
E. coli K-12 DH10B	6.0%	0%
E. coli BW2952	6.0%	0%
E. coli BL21(DE3)	2.4%	0.2 %
E. coli ATCC 8739	2.3%	0%
E. coli B REL606	2.3%	0%
E. coli 53638	2.2%	0%
E. coli 101-1	2.1%	0%
E. coli HS	1.9%	0%
Y. pestis KIM D27	1.6%	32.1%
F. tularensis OSU18	0.2%	3.1%
All others	60.2%	1.0%

of the read probability was assigned to the *E. coli* MG1655 strain and another 6.4% for the *E. coli* W3110 strain. In all, 10 *E. coli* strains had higher read probability than the 1.6% originally assigned to the *Y. Pestis* KIM D27 strain. The *F. tularensis* holarctica OSU18 strain received only 0.2% of the read probability, with 74 other genomes receiving higher read probabilities before reassignment. Once the reads were reassigned 63.6%, 3.1%, and 32.1% of the read probability was assigned to the *E. coli* K-12 MG1655, *F. tularensis* holarctica OSU18, and *Y. pestis* KIM D27 strains respectively. These read probabilities are much closer to the mapping proportions of 69.9%, 3.8%, and 26.3%. Therefore, our method is able to recover the mapping proportions, but is unable to recover the true proportion of reads in the sample if there is differential mapping success for each species.

CHAPTER 6

CONCLUSION

In this paper, we have presented an accurate and sophisticated computational approach for species attribution and strain identification. Our approach relies on the construction of a genome database containing multiple strains or species that are possible source genomes for the sample and the utilizes a probabilistic mapping approach to align the reads to the genome. Reads that map to multiple genomes are then reassigned to the most likely source genome using a Bayesian statistical framework that accommodates information on sequence quality and mapping quality. While a Bayesian approach to the reassignment algorithm has been developed, it was not implemented in this project. In the future we will examine the effect of the prior information on the reassignment algorithm with a prior sensitivity analysis.

We show in multiple real data examples that our method is highly accurate in identifying the source genome or genomes for a biological sample. We show that in many cases, we can identify the source species or strain with only a small number of reads that represent only fractional coverage of the genome. In addition, we show that our approach is able to accurately identify the proper genome of origin, even when several closely related strains or substrains are present within the database. We believe that our approach will play an important role in the fields of pathology, bioforensics, and biosurveillance by allowing government officials and health professionals to quickly and efficiently identify the source of an outbreak to almost the lab level, or to quickly and precisely identify the best treatment for the specific bacteria used.

BIBLIOGRAPHY

Boykin, L. M., Armstrong, K. F., Kubatko, L., and De Barro, P. (2011), "Species Delimitation and Global Biosecurity," *Evolutionary Bioinformatics*, 8, 1–37.

Brown, E. W., Mammel, M. K., LeClerc, J. E., and Cebula, T. A. (2003), "Limited boundaries for extensive horizontal gene transfer among Salmonella pathogens," *PNAS*, 100, 15676–15681.

Clement, N. L., Snell, Q., Clement, M. J., Hollenhorst, P. C., Purwar, J., Graves, B. J., Cairns, B. R., and Johnson, W. E. (2009), "The GNUMAP algorithm: unbiassed probabilistic mapping of oligonucleotides from next-generation sequencing," *Bioinformatics*, 26, 38–45.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1–38.

Eppinger, M., Mammel, M. K., Leclerc, J. E., Ravel, J., and Cebula, T. A. (2011), "Genomic anatomy of Escherichia coli O157:H7 outbreaks," *PNAS*, 108, 20142–20147.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning* (2nd ed.), Springer.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon,

J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005), "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, 437, 376–380.

Pilo, P., and Frey, J. (2011), "Bacillus anthracis: Molecular taxonomy, population genetics, phylogeny and patho-evolution," *Infection, Genetics and Evolution*, 11, 1218–1224.

Read, T. D., Salzberg, S. L., Pop, M., Shumway, M., Umayam, L., Jiang, L., Holtzapple, E., Busch, J. D., Smith, K. L., Schupp, J. M., Solomon, D., Keim, P., and Fraser, C. M. (2002), "Comparative Genome Sequencing for Discovery of Novel Polymorphisms in Bacillus anthracis," *Science*, 296, 2028–2033.

Ruffalo, M., LaFramboise, T., and Koyutürk, M. (2011), "Comparative analysis of algorithms for next-generation sequencing read alignment," *Bioinformatics*, 27, 2790–2796.

Rusk, N. (2011), "Torrents of sequence," *Nature Methods*, 8, 44.

Shendure, J., and Ji, H. (2008), "Next-generation DNA sequencing," *Nature Biotechnology*, 26, 1135–1145.

Turner, M. (2011), "Microbe outbreak panics Europe," *Nature News*, 474, 137.

Vaidyanathan, G. (2011), "Better biosurveillance could halt disease spread," *Nature News*, 477, 392.

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103.