Integrative Genome Assembly of Staphylococcus epidermidis and Staphylococcus hominis Strains

Cana L. Brown
University of South Alabama

Follow this and additional works at: https://jagworks.southalabama.edu/honors_college_theses

Part of the Other Medical Sciences Commons

Recommended Citation

This Undergraduate Thesis is brought to you for free and open access by the Honors College at JagWorks@USA. It has been accepted for inclusion in Undergraduate Theses by an authorized administrator of JagWorks@USA. For more information, please contact mduffy@southalabama.edu.
Integrative Genome Assembly of \textit{Staphylococcus epidermidis} and \textit{Staphylococcus hominis} Strains

By

Cana Brown

A thesis submitted in partial fulfillment of the requirements of the Honors College at the University of South Alabama and the Bachelor of Sciences in the Biomedical Sciences Department

University of South Alabama

Mobile

April 2021

Approved by:

Glen M. Borchert, Ph.D.
Mentor

Asma Hatoum-Aslan, Ph.D.
Committee Member

Mika Houserova, Ph.D.
Committee Member

Kathy J. Cooke, Ph.D.
Dean, Honors College
ACKNOWLEDGEMENTS

I would like to thank my research mentor, Dr. Glen Borchert, for the encouragement and
guidance he has given me through my time as an undergraduate researcher in his lab. I also
would like to express my thanks to Dr. Dominika Houserova, the post-graduate student in our lab
for teaching me and assisting me while I learned the techniques of genome sequencing and for
serving on my committee. Likewise, I would like to thank Dr. Asma Hatoum-Aslan for allowing
me to partake in her research by providing us with the reads to sequence and for serving on my
committee. I also would like to thank the University of South Alabama Honors College and the
Department of Pharmacology in the College of Medicine for providing the opportunity to
participate in this lab. Finally, I would like to thank my friends, family, and peers who have
encouraged me continuously in pursuit of my academic interests and to grow intellectually in so
many ways through my time at the University of South Alabama.
ABSTRACT

Due to rapid advances in sequencing technology, it is becoming increasingly easier to assemble unknown genomes from millions of short sequencing reads of nucleotides taken from the full genomic sequence (Hernandez, 2008). In this study, we used various computational programs to align reads from unsequenced strains of the bacteria Staphylococcus epidermidis and Staphylococcus hominis into a hitherto undefined, single contiguous genomic sequence. We used SPAdes to assist with template free assembly (Bankevich, 2012), BLAST to identify a suitable reference genome from closely related species (Madden, 2013), Bowtie2 to align our reads to the reference genome, SAMtools to sort and organize files (Li, 2009), RGAAT to incorporate variants into our reads and update our final genome (Liu, 2018), and Mauve to rapidly align the reads and provide a visual representation of the final genome (Darling, 2011). In all, we assembled ten bacterial genomes which have never been sequenced and assembled previously. Importantly, we have developed and validated a high throughput computational pipeline capable of quickly assembling full genomes from millions of individual reads. Excitingly this protocol can continue to be used as needed to sequence and assemble more bacterial genomes to provide a genetic basis for studying bacterial characteristics.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>iii</td>
</tr>
<tr>
<td>Abstract</td>
<td>iv</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>viii</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Aims</td>
<td>3</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>4</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>9</td>
</tr>
<tr>
<td>Conclusion</td>
<td>11</td>
</tr>
<tr>
<td>References</td>
<td>13</td>
</tr>
<tr>
<td>Appendix of Figures</td>
<td>16</td>
</tr>
<tr>
<td>Supplemental Data</td>
<td>20</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1. The central dogma of molecular biology. 16

Figure 2. Flow chart of genome sequencing protocol. 16

Figure 3. Mauve visualization of Genome 1. 17

Figure 4. Mauve visualization of Genome 3. 17

Figure 5. Mauve visualization of Genome 4. 17

Figure 6. Mauve visualization of Genome 5. 17

Figure 7. Mauve visualization of Genome 6. 18

Figure 8. Mauve visualization of Genome 7. 18

Figure 9. Mauve visualization of Genome 8. 18

Figure 10. Mauve visualization of Genome 9. 18

Figure 11. Mauve visualization of Genome 10. 19

Figure 12. Mauve visualization of Genome 11. 19
LIST OF TABLES

Table 1. *Genome statistics* 20

Table 2. *Command lines and instruction for protocol* 21
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>NGS</td>
<td>next generation sequencing</td>
</tr>
<tr>
<td>BLAST</td>
<td>“Basic Local Alignment Search Tool”</td>
</tr>
<tr>
<td>SAM</td>
<td>Sequence Alignment/Map</td>
</tr>
<tr>
<td>BAM</td>
<td>Binary Alignment/Map</td>
</tr>
<tr>
<td>RGAAT</td>
<td>Reference-based Genome Assembly and Annotation Tool</td>
</tr>
<tr>
<td>LCB</td>
<td>Locally Colinear Blocks</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>S. epidermidis</td>
<td>Staphylococcus epidermidis</td>
</tr>
<tr>
<td>S. hominis</td>
<td>Staphylococcus hominis</td>
</tr>
</tbody>
</table>
INTRODUCTION

Genetic Material

Every living organism is at its most basic level composed of a variety of genetic material that determines different aspects of that particular organism. The molecule that holds the instructions that formulate an organism’s genetic composition is deoxyribonucleic acid (DNA). DNA is composed of small building blocks made up of chemicals referred to at nucleotides. Each nucleotide has a phosphate group, a sugar group, and a nitrogen base. The type of nucleotide is determined by which nitrogen base is attached to the structure, which can either be adenine, guanine, thymine, or cytosine (Genome.gov). The human genome is composed of about 3 billion bases, however, bacterial genomes can be anywhere between 0.6 million to 9.4 million bases in size (Cole, 2014). For the functions of the organisms to be carried out, the DNA sequences must go through a process that is commonly referred to as the central dogma of biology. In this process, DNA is transcribed by proteins to create ribonucleic acids (RNA) sequences, which then are translated into proteins (Cobb, 2017). These translated proteins are comprised of small amino acids, of which there are 20 types (Genome.gov). The amino acid present in the protein is dependent on the codon in the genetic sequence. A codon is a grouping of 3 nucleotides that corresponds to a specific amino acid that has its place in the protein that is being translated (Saier, 2019).

Drug Resistance in Bacteria

One common problem faced in medicine today is the growing number of multi-drug-resistant bacteria. It has become a global health issue and because of this, many medications that used to be commonly used in hospitals have become largely ineffective (Suzuki, 2014). This
resistance to antimicrobial actions is due to genetic mutations that have occurred within the bacterial genomes, as well as the transmission of plasmids by horizontal gene transfer between different bacterial organisms. Due to these mutations and plasmid acquisitions, many bacteria are now tolerant to antibiotics as a result of new abilities such as inactivating drugs, decreasing drug access sites, causing the drug targets to no longer bind to the drug, and others (McManus, 1997). In the field of microbiology, NGS allows researchers to observe genomic features of various microbiota and how that affects phenotypic abilities, such as antibiotic resistance. Having a better understanding of the drug sensitivities of different strains of various species of bacteria can help researchers and physicians find more efficient medication options for treating bacteria. Genetic information can also help scientists gain knowledge of infection outbreaks, as it can allow specific strains to be traced back to a host of origin (Behjati, 2013).

**Next-Generation Sequencing**

On the internet, there is a GenBank database that has made the genomes of 2557 eukaryote genomes, 432 archaea genomes, and 7474 bacteria genomes. It even includes the genomes of some viruses, organelles, and plasmids (Goldman, 2016). However, despite the large number of genomes sequenced and made available, researchers regularly discover new strains of species that need to be sequenced and assembled. Through a process more commonly known as next-generation sequencing (NGS), an entire genome can be sequenced and assembled in a period as short as one day. Having the sequences of a genome readily available is important because knowing which proteins and segments of DNA are within a particular genome has many clinical and research benefits (Behjati, 2013). Functional characterization of the genes discovered through NGS allows scientists to further their research of newly discovered strains and to account for any newly arising, potentially drug-resistant mutants (Liao, 2015).
AIMS

**Overall Aim**

To assemble a series of genomes from sequenced reads provided by the lab of Dr. Asma Hatoum-Aslan using various next-generation sequencing techniques.

**Specific Aim**

To develop an efficient protocol for assembling full genomes in an effective and timely manner.
**METHODS AND MATERIALS**

*SPAdes*

SPAdes is a computational biology program that helps assemble genomes of single-cell and multicell organisms. It uses multiple novel algorithmic solutions to improve from the previously available assemblers. At the time of its creation, it showed great improvements compared to the pre-existing programs, such as E + V – SC, Velvet, and SoapDeNovo. E + V – SC provided coverage of 93.8%, whereas SPAdes provided 96.1% coverage. The total time needed for SPAdes is about 3 hours. One aspect of SPAdes that helps it be more efficient is that it utilizes “read-pairs” to avoid misassemblies. Through a 4-step process, SPAdes creates contigs of DNA sequences to map out the reads to assist in assembling. In the first stage of SPAdes, the program constructs an assembly graph. In the second stage, the program adjusts the assembly graph based on the estimated distance between particular segments of nucleotides. In the third stage, the program creates a paired assembly graph. Then, the final stage consists of the final construction of the contigs of DNA sequences. These contigs that are output from SPAdes will be used to assist in template-free assembly (Bankevich, 2012).

*BLAST*

The National Center for Biotechnology Information (NCBI) has provided a program called the “Basic Local Alignment Search Tool,” also referred to as BLAST. BLAST uses heuristics and takes DNA sequences and aligns those sequences with known genomes in its database. It then provides a score that estimates the likelihood that the match occurred by chance so that the match is made with confidence. BLAST works through three main phases. First, there is a setup phase, where the system reads the input DNA, reviews the parameters, and looks
through the database. The second phase is a preliminary search. During this phase, the system looks for matches in the database. Then, any gapped extensions made from those matches are saved if they are given a high enough score. The final phase is a traceback through the saved extensions using more specific parameters to find the best matches. In our project, BLAST helped us to identify a suitable reference genome from closely related species so that the new genome could be assembled more accurately (Madden, 2013).

Bowtie2

The program Bowtie2 allows for genomic alignments at high speeds, with efficient sensitivity and accuracy by integrating the features of the “full-text minute index” with other powerful programming techniques. Its approach has two overarching stages. The first stage is an ungapped stage that works on finding the best seeds. The second stage works with gapped extensions to align the seeds. For every read that is processed by Bowtie2, there are 4 steps for the program. First, Bowtie2 notes parts that are referred to as “seed substrings” from the sequence and their reverse complements. Second, those substrings are aligned to the reference genome using the full-text minute index. Third, the alignments of the seeds are prioritized, and the index is used to calculate their positions within the reference genome. Lastly, the seeds are fully aligned by using SIMD-accelerated dynamic programming. Through these steps, we use the Bowtie2 program to align our reads to the reference genome at the beginning of our protocol and to align our new assembly to the original reads at the end of our protocol (Langmead, 2012).

SAMTools

The Sequence Alignment/Map (SAM) format is convenient for assemblies because it is a generic format for alignments that stores the read alignments along with the reference sequences
and it supports both single- and paired-end reads. SAMTools is a program used for a variety of purposes to sort, organize, or convert files of processed read alignments that are in the SAM format or the Binary Alignment/Map (BAM) format. In our project, we get our SAM files from the Bowtie2, convert them to sorted BAM files with SAMtools, and then use those BAM files to update our reference genome with RGAAT (Li, 2009).

**RGAAT**

The reference-based genome assembly and annotation tool (RGAAT) is a computational toolkit. It can detect sequence variants, which it then can use to update our reference genome to be more accurate to what our unknown genome may look like. RGAAT has three main modules. First, RGAAT can identify variants based on read alignment, then it creates coordinate conversion files based on the genome alignments that were found to be non-redundant, and lastly, it creates a final genome assembly and annotation as the output. In our project, RGAAT is used to incorporate variants into our reads and to update our reference genome (Liu, 2018).

**Mauve**

Mauve is a popular genome assembly tool used by many different researchers. Mauve maps out a new genome by using a reference genome and aligning the sequences using whole-genome alignment programs rather than local alignment programs (Darling, 2011). Mauve is a unique program because it is the first alignment program to integrate analysis of bigger evolutionary events with the more traditional multiple sequence alignment (Darling, 2004). The Mauve Contig Mover can be utilized to reorder the contigs being sequenced to match the reference genome. However, it is important to note that genome alignment algorithms are not faultless and there is always a potential for errors to be made in the process of genome assembly.
For us, Mauve rapidly aligns the reads and provides a visual representation of the final genome. Generally, the program outputs multiple options for alignments and we pick the one with the lowest number of Locally Colinear Blocks, referred to as LCBs (Darling, 2011).

**Protocol**

The first step in our protocol is to use SPAdes to provide contigs created from the original reads. With the command ‘spade.py’ we marked the forward reads with ‘-1,’ the reverse reads with ‘-2,’ and what we wanted the output directory to be titled with ‘-o.’ Once the program was finished, we went into the folder it created, found the contig file it had created, and renamed it to reflect which genome we were working with (Prjibelski, 2020). Next, we ran the first 10 contigs for that genome through the NCBI Nucleotide BLAST database and noted the accession numbers of the genomes that best matched each contig (Madden, 2013). Once we determined which genome matched best overall, we download the fasta sequence for that genome and used it as the ‘reference genome’ for that assembly. Once the reference genome was obtained, we then used Bowtie2 to build a “Bowtie index” using the command ‘bowtie2-build’ followed by the input of the fasta and the output name desired. Once the index was created, we used the ‘bowtie2’ command with the paired-end option ‘dovetail,’ which will consider mate alignment concordant if one of the mate’s alignment extends beyond the start of another so that the wrong mate starts upstream. The options we used with it were ‘-x’ for the reference genome, ‘-1’ for the “mate 1s,” ‘-2’ for the “mate 2s,” and ‘-S’ to title the file that will write the SAM alignments (Langmead, 2012). Once Bowtie2 finished running, we then used ‘SAMtools’ with ‘view’ and ‘-bS’ to convert the SAM formatted files into a BAM format. Then we used ‘SAMtools sort’ followed by the BAM file followed by ‘-o’ and the sorted output file name (Li, 2009). Once we had the sorted BAM file, we moved into the RGAAT-master directory. We used the command
line ‘RGAAT.pl’ to update our reference genome to include variants. We used ‘-g’ to note the genome file, ‘-b’ to note the sorted BAM file, and ‘-o’ to note the prefix of the file that would be the output (Liu, 2018). Once we download the updated genome for reference, we used Mauve to assemble and visualize the contigs. First, you go to ‘tools,’ then select ‘move contigs.’ Then you select where you want the alignments to be located, followed by uploading the reference genome and the original contigs after (Darling, 2004). Once it finishes its alignments, we picked the alignment with the least LCBs and renamed that newly arranged contig file to reflect the genome we were working on. We then used the Mauve visualization to find the contigs at the end of the alignment that did not align anywhere along with the reference. Taking those extra contigs, we ran each of them through the NCBI BLAST database to see if any were noted as plasmids (Madden, 2013). For each of the contigs that were noted to be a plasmid, we removed them and placed them into a separate file that was adequately titled to reflect that the file contigs plasmids from that respective genome. Finally, we ran the newly assembled genome through the Bowtie2 programs with the same commands, but this time we used an index file for the original contigs file, and we used ‘--un’ and ‘--un-conc’ which will note unpaired reads that failed to align to the final file and note paired-end reads that failed to align concordantly to the final file, respectively. We used the final report from this run of Bowtie2 to ascertain the number of reads and the overall alignment rate (Langmead, 2012).
RESULTS AND DISCUSSION

Overall, we were able to successfully assemble sequences of ten genomes. Nine of the genomes were *S. epidermidis* and the final genome was *S. hominis*. Genome 1 had a total of 406 contigs, composed of 1943991 reads. The reference genome used for it was CP043804 and the final Bowtie2 alignment rate for it was 85.41% (Figure 3). Genome 2 had 2786 contigs, and upon further examination, it was concluded that it had data contamination. Due to this, we did not complete the sequencing of this genome. Genome 3 had a total of 155 contigs, composed of 1222132 reads. The reference genome used for it was CP060248 and the final Bowtie2 alignment rate for it was 88.51% (Figure 4). Genome 4 had a total of 133 contigs, composed of 2903587 reads. The reference genome used for it was CP035643 and the final Bowtie2 alignment rate for it was 94.74% (Figure 5). Genome 5 had a total of 145 contigs, composed of 4033552 reads. The reference genome used for it was CP035643 and the final Bowtie2 alignment rate for it was 95.65% (Figure 6). Genome 6 had a total of 242 contigs, composed of 221 reads. The reference genome used for it was CP035643 and the final Bowtie2 alignment rate for it was 94.55% (Figure 7). Genome 7 had a total of 221 contigs, composed of 2572134 reads. The reference genome used for it was LR735437 and the final Bowtie2 alignment rate for it was 95.43% (Figure 8). Genome 8 had a total of 300 contigs, composed of 2531400 reads. The reference genome used for it was LR735432 and the final Bowtie2 alignment rate for it was 92.72% (Figure 9). Genome 9 had a total of 143 contigs, composed of 1803608 reads. The reference genome used for it was CP035643 and the final Bowtie2 alignment rate for it was 95.47% (Figure 10). Genome 10 had a total of 180 contigs, composed of 1206393 reads. The reference genome used for it was LR735437 and the final Bowtie2 alignment rate for it was 97.00% (Figure 11). Genome 11 had a total of 246 contigs, composed of 1829988 reads. The
reference genome used for it was CP054550 and the final Bowtie2 alignment rate for it was 95.33% (Figure 12).

The average number of contigs across all 10 assembled genomes was 217.1, with Genome 4 having the least at 133 and Genome 1 having the most at 406. The average number of reads across the genomes was 2201704.9, with Genome 10 having the least at 1206393 and Genome 5 having the most at 4033552. The average alignment rate for the genomes was 93.48%, with Genome 1 having the lowest at 85.41% and Genome having the highest at 97.00% (Table 1).
CONCLUSION

The ten genomes that we assembled in our job can now be used to observe genetic characteristics that lead to phenotypic traits in bacteria. The main field that will utilize this genetic information is the study of antibiotic resistance in bacteria. Having access to the genetic mapping of a bacteria allows researchers to gain a better understanding of genetic mutations that lead to the phenotypic change of antibiotic resistance. So far, there have already been a variety of mutations recognized as having a relation to giving a bacteria cell antibiotic resistance (Suzuki, 2014). They also have linked the transmission of plasmids as a potential way of acquiring resistance to different antibiotics. As the ability of bacterial cells to cause drug inactivation, decrease drug access sites, and prevent binding of the drugs has become more prominent, the topic of bacterial resistance to antibiotics has increased as well (McManus, 1997).

When researching to find new and important information regarding this resistance, typically one of the first steps is to perform phenotypic tests to see what the bacterial strains are resistant to and what they are susceptible to. After determining those characteristics, the genomes are analyzed and sequenced so that you can look at the specific alleles and nucleotide sequences that make up that particular strain’s genetic information. Once the genetic background of the strain is available, researchers can use that information to check for the presence of genes that are associated with resistance to the antibiotics being studied. Once particular genes are noted in various domains of the genome, researchers can confirm whether or not the mutations that allegedly cause the resistance are present. Researchers can also then identify gene mutations that will encode for particular proteins to be produced that contribute to antibiotic resistance (Bongiorno, 2010).
Overall, the final assembled genomes produced by this project assist researchers in further studies of these nine strains of *S. epidermidis* and one strain of *S. hominis*. With these genomes now fully sequenced and assembled, researchers can look at the genetic information within these cells and connect information from that to the cells’ phenotypic characteristics. With the ability to understand more of the mechanisms that can contribute to bacterial resistance, researchers can further their knowledge on how to efficiently eradicate these strains of *S. epidermidis* and *S. hominis*. Similarly, physicians can better choose medications and antibiotics that will effectively treat patients who have infections consisting of those strains of *S. epidermidis* and *S. hominis* (Behjati, 2013). It can even allow for scientists to trace outbreaks of these infections back to an origin so they can see how the infection may have spread throughout a population (Römling, 1995).
REFERENCES


Figure 1. The central dogma of molecular biology. The DNA in cells gets transcribed into RNA strands. Those RNA strands are then translated into proteins for the cells to use in a variety of biochemical processes.

Figure 2. Flow chart of genome sequencing protocol. Depicted above is a flow chart that shows the process of sequencing bacterial genomes using the protocol that our lab formulated.
Figure 3. Mauve visualization of Genome 1.

Figure 4. Mauve visualization of Genome 3.

Figure 5. Mauve visualization of Genome 4.

Figure 6. Mauve visualization of Genome 5.
Figure 7. Mauve visualization of Genome 6.

Figure 8. Mauve visualization of Genome 7.

Figure 9. Mauve visualization of Genome 8.

Figure 10. Mauve visualization of Genome 9.
Figure 11. Mauve visualization of Genome 10.

Figure 12. Mauve visualization of Genome 11.
**SUPPLEMENTAL DATA**

**Table 1. Genome Statistics**

<table>
<thead>
<tr>
<th>Genome</th>
<th>Number of Contigs</th>
<th># of Reads</th>
<th>Final Bowtie2 Alignment Rate</th>
<th>Reference Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>406</td>
<td>1943991</td>
<td>85.41%</td>
<td>CP043804</td>
</tr>
<tr>
<td>3</td>
<td>155</td>
<td>1222132</td>
<td>88.51%</td>
<td>CP060248</td>
</tr>
<tr>
<td>4</td>
<td>133</td>
<td>2903587</td>
<td>94.74%</td>
<td>CP035643</td>
</tr>
<tr>
<td>5</td>
<td>145</td>
<td>4033552</td>
<td>95.65%</td>
<td>CP035643</td>
</tr>
<tr>
<td>6</td>
<td>242</td>
<td>1970264</td>
<td>94.55%</td>
<td>CP035643</td>
</tr>
<tr>
<td>7</td>
<td>221</td>
<td>2572134</td>
<td>95.43%</td>
<td>LR735437</td>
</tr>
<tr>
<td>8</td>
<td>300</td>
<td>2531400</td>
<td>92.72%</td>
<td>LR735432</td>
</tr>
<tr>
<td>9</td>
<td>143</td>
<td>1803608</td>
<td>95.47%</td>
<td>CP035643</td>
</tr>
<tr>
<td>10</td>
<td>180</td>
<td>1206393</td>
<td>97.00%</td>
<td>LR735437</td>
</tr>
<tr>
<td>11</td>
<td>246</td>
<td>1829988</td>
<td>95.33%</td>
<td>CP054550</td>
</tr>
<tr>
<td><strong>Total Average</strong></td>
<td><strong>217.1</strong></td>
<td><strong>2201704.9</strong></td>
<td><strong>93.48%</strong></td>
<td></td>
</tr>
</tbody>
</table>


Table 2. Command lines and instruction for protocol

<table>
<thead>
<tr>
<th>Program</th>
<th>Command Line/Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST</td>
<td>Copy and Paste first 5 Nodes into blast and note the top 10 or so matches for each. Then determine which is the best match and download the fasta sequences for that genome.</td>
</tr>
<tr>
<td>Bowtie2</td>
<td>bowtie2-build [input reference genome] [database name for reference genome]</td>
</tr>
<tr>
<td>Samtools</td>
<td>samtools view -bS [allreads versus reference output].sam &gt; [allreads versus reference output].bam</td>
</tr>
<tr>
<td>Samtools</td>
<td>samtools sort [allreads versus reference output].bam -o [allreads versus reference output].sorted.bam</td>
</tr>
<tr>
<td>RGAAT (note: must be in RGAAT directory)</td>
<td>perl RGAAT.pl -g [reference genome sequence].fasta -b [allreads versus reference output].sorted.bam -o [genome with variants incorporated]</td>
</tr>
<tr>
<td>Mauve</td>
<td>Go to &quot;tools,&quot; then &quot;move contigs.&quot; Select where you want the alignments outputted, then upload the reference genome with variants incorporated first, then upload the contigs for that genome to be assembled.</td>
</tr>
<tr>
<td>Mauve</td>
<td>Open the reordered contigs for your genome and find in the alignment visualization where the leftover contigs begin, then move those contigs into a separate text file.</td>
</tr>
<tr>
<td>BLAST</td>
<td>Run the extra contigs in the NCBI database and move any contigs that are not plasmids back to the end of the main file.</td>
</tr>
<tr>
<td>Bowtie2</td>
<td>bowtie2-build [input original contigs] [output database for original contigs]</td>
</tr>
</tbody>
</table>