7-2016

Sequencing and Comparative Analysis of *de novo* Genome Assemblies of *Streptomyces aureofaciens* ATCC 10762

Julien S. Gradnigo
University of Nebraska - Lincoln, jgradnigo@gmail.com

SEQUENCING AND COMPARATIVE ANALYSIS OF DE NOVO GENOME ASSEMBLIES OF *STREPTOMYCES AUREOFACIENS* ATCC 10762

by

Julien S. Gradnigo

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Biological Sciences

Under the Supervision of Professor Etsuko Moriyama

Lincoln, Nebraska

July, 2016

SEQUENCING AND COMPARATIVE ANALYSIS OF DE NOVO GENOME ASSEMBLIES OF *STREPTOMYCES AUREOFACIENS* ATCC 10762

Julien S. Gradnigo, M.S.

University of Nebraska, 2016

Advisor: Etsuko Moriyama

Streptomyces aureofaciens is a Gram-positive Actinomycete used for commercial antibiotic production. Although it has been the subject of many biochemical studies, no public genome resource was available prior to this project. To address this need, the genome of *S. aureofaciens* ATCC 10762 was sequenced using a combination of sequencing platforms (Illumina and 454-shotgun). Multiple *de novo* assembly methods (SGA, IDBA, Trinity, SOAPdenovo2, MIRA, Velvet and SPAdes) as well as combinations of these methods were assessed to determine which provided the most robust assembly. Combination strategies led to a consistent overestimation of the total genome size. Empirical data from targeted PCR of predicted gap regions provided a robust validation framework for our *de novo* assemblies. Overall, the best assembly was generated using SPAdes. The total length of this assembly was 9.47 Mb and the average G+C content was 71.15 %. We annotated this assembly using the NCBI Prokaryotic Genome Annotation Pipeline, revealing 8,073 total genes, including a total of 7,627 protein coding sequences. Additional functional analysis using the KEGG GENES database provided functional predictions for over 1,400 of these sequences whose functions were not initially inferred by NCBI. The information provided from multiple independent assemblies allowed us to close 200 scaffold gaps present in our first hybrid assembly. Comparative genomic and phylogenetic analyses suggested *S. aureofaciens*

ATCC 10762 may be more closely related to the genus *Kitasatospora* than to neighboring *Streptomyces* species. Our results highlight the need for, and the value of, multiple assemblies when attempting to produce high quality prokaryotic genome sequences.

ACKNOWLEDGEMENTS

I can think of no task more difficult than enumerating the myriad of people and organizations that have enabled me to complete this undertaking. Science is a collaborative endeavor, perhaps now more than ever. To begin, I must acknowledge the brilliant faculty, staff and students I have met in, or through, the UNL School of Biological Sciences who have assisted me in innumerable ways. This list begins with my advisor, Dr. Etsuko Moriyama, whose steadfast guidance provided the foundation for my graduate career. For their contributions and their friendship, I would also like to thank Dr. Hideaki Moriyama, Dr. Gregor Grass, Dr. Greg Somerville, Dr. Robert Norgren, Dr. Brian Couch, Dr. Jean-Jack Riethoven, Dr. Jeffrey French, Dr. Audrey Atkin, Dr. Steven Harris, Dr. Richard Wilson, Dr. Stephen Opiyo, Dr. Seong-Il Eyun, Dr. Adam Voshall, Dr. Davide Quaranta, Dr. Christophe Espírito Santo, Nicholas Johnson, Cate Anderson, The Nguyen, Renee Rodriguez Batman, Britta Osborne, Mindy Peck, Carla Tisdale, Tammy Kortum, Julie McManamey, Dr. Michael Huether, Dr. Richard Kemmy, Michael Oliver and Craig Johnson.

I would also like to acknowledge those who supported me outside of the lab, without whom my sanity would surely have been lost long ago. This especially includes: Jennie Catlett, Marvida Gonsoulin-Harris, Terry Harris, Rizwan Merchant, Richie Smith, Lexi Smith, Jason Galewski, Bree Galewski, Luke Ashworth-Sides, Scott Broussard, Danielle Ducrest, Michael Babeji, Jennifer Reed, Sean Carr, Erin Carr, Sam McCarthy, Justin Buchanan, Maya Khasin, Emily Wynn, Amy Ort, Adelle Burk, Josh Sammons, Rhiannon West, Felix Grewe, Ashley Atwell, Angelica ‘Licki’ Kallenberg, Dan Gates, and all my friends and family whom I have neglected to mention.

Last, but certainly not least, I gratefully acknowledge the institutions that have supported me in my time here, especially the University of Nebraska-Lincoln, Zoetis and the University of Nebraska Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Streptomyces aureofaciens is a Gram-positive Actinomycete bacterium identified in 1948 [1], from Plot 23 in Sanborn Field, a timothy hayfield at the University of Missouri [2]. Like many bacteria, *S. aureofaciens* produces compounds not required for immediate survival. These secondary metabolites often exhibit anti-microbial activity and include the common antibiotics tetracycline and chlortetracycline [2]. Although *S. aureofaciens* has been used for the commercial production of tetracycline antibiotics for some time [1,3] and has been the subject of numerous biochemical studies, no public genome assembly was published until very recently. Certain characteristics have been well-studied – for example, the *Streptomyces* are known to have high G+C genome content (estimated at 74%, overall), and fairly large genomes in the range of 9 – 12 Mbp [4,5]. Still, there remains a dearth of information with regard to the phylogenetic classification of many Actinomycetes, including the *S. aureofaciens* type strain.

Over the past 30 years, DNA sequencing methods have improved significantly**.** Emergent technologies like Sanger sequencing produced relatively little sequence data at great expense [6]; a decade later, the invention of the polymerase chain reaction (PCR) opened the door for molecular biologists to rapidly and specifically amplify DNA molecules [7]. However, sequencing entire genomes remained difficult until the arrival of next-generation sequencing platforms, such as those developed by Illumina (*e.g.*, the HiSeq and MiSeq platforms) and Roche (*e.g.*, 454 pyrosequencing). These technologies have contributed to a drastic decline in sequencing costs and an ever growing number of completed sequencing projects, including the human genome [8].

The advancement of sequencing technology was not itself sufficient to make this possible. Over the same time period, genome assembly algorithms have drastically improved to exploit more efficient processors, increased memory capacities and multi-core technologies that are now widely and cheaply available. The raw data (reads) produced by the aforementioned sequencing methods is very short relative to the length of a genome, typically around a few hundred base-pairs (bp). Assembly algorithms transform raw reads into longer, contiguous sequences (contigs) by identifying and joining overlapping regions between reads, which may then be further assembled into scaffolds (comprised of contigs and gap regions of an estimated size) or super-scaffolds (comprised of multiple scaffolds in a specific orientation). There are two assembly strategies: reference-guided assembly methods, which use information from prior assemblies of closely related taxa to minimize error and increase assembly accuracy, and *de novo* assembly methods, which utilize only sequence reads, without using another genome as a reference. While individual implementations differ, many modern assemblers make use of de Bruijn graphs for this purpose, breaking the individual reads into shorter pieces (k-mers) and representing their overlapping regions via a directed graph [9]. Additional information is also used, such as the approximate distance between a given pair of sequences (in the case of Illumina long-jump distance sequencing) or sequence information from both ends of the same DNA fragment (paired-end sequencing). In this way, large genomes composed of millions of nucleotide base pairs can be reconstructed from a large number of short reads.

Despite these algorithmic improvements, several factors continue to make it difficult to produce high-quality finished genomes. These difficulties are present at both

the sequencing and assembly levels, and affect both *de novo* and guided assemblers. At the sequencing level, GC-rich regions are more stable and less prone to denaturation, which can prove problematic during PCR amplification [10]; sequencers often have difficulty accurately sequencing repeat regions [11]; and even with low rates of sequencing error (*i.e.*, incorrect base-calling), larger sets of reads may contain hundreds of thousands of incorrectly called bases [12–14]. At the assembly level, repeat regions continue to pose a challenge [15]; short reads may leave segments of the genome uncovered [16], and suboptimal parameterization (*e.g.*, k-mer size or base quality score thresholds for base-clipping) contributes to erroneous, highly-fragmented assemblies [17].

Many of these challenges can be addressed by careful experimental planning – for example, ensuring sufficient sequencing coverage (defined as the average number of reads covering each base in the assembly), often 100X or more. In recent years, more complex computational approaches have evolved to take advantage of longer sequence reads. These methods are increasingly capable of integrating multiple sets of reads generated by differing sequencing platforms. These 'hybrid assemblers' can exploit the overlap information provided by long reads to build longer contigs and more complete scaffolds while using accurate, high-coverage short reads to more confidently infer the correct base at each position. Prior studies have indicated that the assembler SPAdes [18] consistently outperform many alternative assemblers, particularly when building hybrid assemblies [19–22].

The aims of my thesis were to: 1) thoroughly evaluate the performance of several assemblers for the *S. aureofaciens* ATCC 10762 genome, with a focus on comparing hybrid and non-hybrid assembly strategies; 2) evaluate the effectiveness of integrating

multiple existing assemblies into a single, meta-assembly; and 3) perform comparative genomics, functional, and phylogenetic analyses on *S. aureofaciens* and closely related species. We compared six non-hybrid assemblies generated with SOAPdenovo2, Trinity, IDBA, SGA, MIRA and SPAdes, and two hybrid assemblies generated with Velvet and SPAdes. Additionally, combination assemblies were generated using CISA. Overall, SPAdes, using hybrid data, produced the best assembly which we annotated. Phylogenetic and comparative genomic analyses were conducted to more clearly define the lineage of *S. aureofaciens* strain ATCC 10762. This strain was found to be more closely related to the genus currently known as *Kitasatospora* than to other *Streptomyces* species. Additional, functional analysis via the KEGG database provided additional information on over 1,400 sequences whose functions were not initially annotated from our hybrid SPAdes assembly. Our analyses showcase the utility of a hybrid assembly approach, emphasize the difficulty of proper phylogenetic placement and highlight shortcomings that may result from attempting to generate a meta-assembly.

CHAPTER 2: MATERIALS AND METHODS

2.1 – Bacterial Culture and DNA Isolation

S. aureofaciens strain ATCC 10762 (lot 3856567) was purchased, lyophilized in a sealed glass ampule. It was hydrated with 5 ml of ISP Medium 1 (Tryptone Yeast Extract Broth) and used to inoculate 500 mL of WI FVM Seed Media (hereafter referred to as DM1) [23]. Bacteria were cultivated in 2L baffled flasks at 30°C, with 150 rpm aeration with a 2" throw[1] for 48 hours. This culture was used to make a master seed stock by aliquoting 4.5 mL into cryovials and storing at -80°C.

One vial of the master seed was thawed, and 2.5 mL used to inoculate 500 mL of DM1 media. This culture was grown in 2 L non-baffled flasks at 30°C, 150 rpm with a 2" throw for 9 days.[1] A 200 mL sample was taken for DNA extraction and refrigerated at 4°C. The isolation and purification of high molecular weight DNA from fresh *S. aureofaciens* cultures was completed by CTAB extraction [24]. Extracted genomic DNA was further evaluated for molecular weight integrity by agarose gel electrophoresis and nucleic acid fluorometric quantitation for construction of the DNA library

2.2 – DNA Sequencing

Illumina and 454-shotgun sequencing, and read quality filtering, were completed by Eurofins MWG Operon (Alabama, USA). Illumina MiSeq sequencing was done with long jumping distance sequencing (3-kb and 8-kb inserts), generating paired-end 150-bp reads; 454-shotgun sequencing was completed using the Roche 454 Genome Sequencer FLX platform. For quality filtering, very short (<30 bp) reads and Illumina adapter

[1] This distance describes the diameter of the orbital path produced by the shaking mechanism.

sequences were removed, and low quality bases were clipped out using Trimmomatic [25]. The raw reads have been deposited in the National Center for Biotechnology Information (NCBI) Short Read Archive; 454 reads are available under the accession number SRX1122678, and the 3-kb and 8-kb Illumina libraries are available under SRX1122692 and SRX1122693, respectively.

2.3 – *De novo* Genome Assembly

2.3.1 – Non-hybrid Genome Assembly

The Illumina reads were assembled using six methods: Iterative De Bruijn Graph Assembler (IDBA v. 1.1.1) [26], String Graph Assembler (SGA v. 0.10.13) [27], Trinity v. 2.0.6 [28], MIRA v. 4.0.2 [29], SOAPdenovo2 v. 2.04 [30] and SPAdes [18]. These assemblers are optimized for slightly different applications. Briefly, IDBA uses a range of k-values in an attempt to automatically identify the optimal k-mer length for building the de Bruijn graph; SGA eschews the de Bruijn method in favor of string graphs, with the goal of being extremely memory efficient; Trinity is a suite of three programs (*i.e.*, Inchworm, Chrysalis and Butterfly) designed to reconstruct transcripts from RNA-sequencing reads; MIRA is a memory-intensive, iterative assembler that also avoids de Bruijn graphs in favor of an overlap-layout-consensus approach; SOAPdenovo2 is primarily designed to handle larger genomes, like those of plants and animals; and SPAdes implements an iterative k-mer search strategy similar to that of IDBA, along with contig error-correction and assembly merging algorithms. Both the 3-kb and 8-kb Illumina libraries were provided as input. Excluding Trinity and MIRA, which do not implement scaffolding algorithms, each assembler generates a set of contigs and scaffolds.

Default k-mer selections were used for each assembly, requiring no additional parameter specification.

2.3.2 – Hybrid Genome Assembly

A hybrid assembly was performed by Eurofins MWG Operon (Alabama, USA) as follows. The quality filtered 454-shotgun reads were assembled with Newbler (GS Data Analysis Software package, 454 Life Sciences). The filtered Illumina reads were mapped to the resultant 454 contigs to infer the approximate insert size for each library, after which the paired-end Illumina reads and the 454 contigs were assembled using Velvet (v 1.2.10) [31] across a broad range of k-mer sizes. This assembly has been deposited in NCBI's GenBank under the accession GCA_001188955.1. It should be noted that this assembly (version 1) has been superseded by the assembly described below (GCA_001188955.2).

For the second hybrid assembly, SPAdes (v. 3.7.1) [18] was used to assemble all the quality filtered Illumina and 454 reads, including singletons, across a range of k-mers (the default behavior of SPAdes – this requires no specific k-mer arguments). The '—careful' option was used to reduce mismatches and short indels. This assembly has been published [32] and was deposited in NCBI's GenBank under the accession GCF_001188955.2.

2.3.3 – Integration of Multiple Assemblies

Because assemblies vary, the multiple combinations of contig sets were merged using the Contig Integrator for Sequencing and Assembly (CISA) [33]. CISA does not

implement its own *de novo* or guided assembly algorithm – rather, it attempts to identify and extend overlapping regions of pre-existing contigs. As such, use of CISA requires at least three separate assemblies. We generated four CISA datasets: set 1 consists of the IDBA, Trinity and Velvet contigs; set 2 includes all of set 1 with the addition of the SGA contigs; set 3 includes all of set 2, and the MIRA contigs, and set 4 additionally includes the SPAdes contigs.

2.4 – Comparative Analysis of Assemblies

2.4.1 – Contig and Scaffold Alignments

Pairwise alignments of contig sets were generated via nucleotide BLAST searches and MUMmer 3.0 [34]. MUMmer identifies and clusters matching sequence regions between the contig sets, then extends matches within these clusters using Smith-Waterman alignment techniques. We report the total percentage of aligned bases, indicating the total proportion of nucleotides from the first contig set that align to at least one match cluster in the second set. This is distinct from a measure of percent identity, which indicates the similarity of individual alignments between match clusters.

Assemblies were also compared using the Quality Assessment Tool for Genome Assemblies (v. 4.0) [35], which provides a number of summary statistics, including total length, G+C percentage, the N50 length (a commonly used statistical measure, defined as sequence length *N* such that half of the assembly is contained in contigs of *N* bp or greater) and L50 (the number of contigs equal to or longer than N50; in other words, the minimal number of contigs covering half of the assembly).

2.4.2 – PCR of Predicted Gap Regions

Regions that were predicted to contain short gaps (<100 bp, based on the Velvet assembly, GCA_001188955.1) were selectively amplified using the 'slowdown PCR' protocol which was designed to amplify GC-rich regions [10]. The difficulties of amplifying such templates are well documented [36–38]. The three hydrogen bonds formed between guanine and cytosine make GC-rich regions more stable than AT-rich regions, impeding DNA denaturation. Our initial PCR failed to amplify any templates. This led us to switch to the slowdown PCR protocol, which reduces the heating and cooling ramp rates, implements a progressively lowered annealing temperature over the length of the protocol and appends several annealing cycles at the end [10]. This method, combined with the addition of DMSO, was sufficient to facilitate template amplification.

Final reaction volumes were always 50 µL. Each reaction included: 25 µL DreamTaq PCR Master Mix (2X) from ThermoFisher Scientific (Waltham, MA); 1 µL each of forward and reverse primer (10 nmol concentration); and 1 µL template *S. aureofaciens* ATCC 10762 genomic DNA, and 19.5 µL nuclease-free water. Each set of reactions also included one replicate containing 2.5 µL (5% v/v) dimethyl sulfoxide (DMSO). Successfully amplified PCR products were isolated and purified with the QIAquick PCR Purification Kit from Qiagen (Valencia, CA) and sequenced in both directions by Eurofins MWG Operon (Alabama, USA). We targeted 34 regions in total, successfully amplifying 14. Of these 14, we were able to generate reliable sequence information for 9 regions. Manual sequence correction was performed as necessary according to the resultant chromatograms. All primer sets are listed in Appendix 1.

2.5 – Contaminant Filtering

The contig sets of both hybrid assemblies were screened for the presence of non-host DNA in the form of prophage and plasmid sequences. For prophage screening, we used the PHAge Search Tool (PHAST) webserver, available at http://phast.wishartlab.com [39].

Plasmid screening was conducted using two independent methods. First, contig sets were scanned using the PlasmidFinder webserver, available at https://cge.cbs.dtu.dk//services/PlasmidFinder/ [40]. Second, we manually attempted to identify plasmids by aligning both sets of contigs and scaffolds against all plasmid sequences available from NCBI as of 25 Apr 2016 using the BLASTN program (v. 2.2.30+), which is part of the standalone BLAST package (BLAST+) [41,42].

2.6 – Automated Annotation of Genomic Features

Genomic features (*e.g.*, coding sequences, rRNAs, tRNAs, *etc*) were annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP), the core of which is built on the gene prediction suite GeneMarkS+ (v. 2.6 rev. 440435 for the Velvet assembly, GCA_001188955.1; v. 3.1 for the SPAdes assembly, GCA_001188955.2) (Tatusova et. al., 2013).

2.7 – Circular Visualization of the Genome

A circular visualization of the genome assembly was generated using ClicO FS, a web-based implementation of the Circos plotting tool [43,44].

2.8 – Comparative Genomics

To identify homologous genes, we performed protein BLAST searches using the coding sequences from our SPAdes assembly. Identical homologs are defined by alignments with 100% query coverage and sequence identity (no gaps or mismatches). Non-identical homologs are defined by alignments with >95% sequence identity and query coverage.

2.9 – Phylogenetic Inference

Ortholog sets were aligned using MAFFT, v. 7.245 with the L-INS-I option [45–47]. Maximum likelihood phylogenies were inferred using RaxML,v. 8.2.4 [48], with the following options: '-f a', which performs a rapid bootstrap analysis and searches for the best-scoring tree in a single run; '-x', which enables rapid bootstrapping; and '-p' which is necessary for parsimony inferences. The '-x' and '-p' options were followed by random number seeds. The GTRGAMMA substitution model was used for both protein and nucleotide phylogenies, and 500 bootstrap replicates were sampled to assess branch support.

2.10 – Functional Annotation

Additional verification of the automated gene annotations was performed as necessary via local BLAST searches of the proteins made available by the Kyoto Encyclopedia of Genes and Genomes (KEGG) [49], specifically the KEGG GENES database.

Annotated genes were divided into two groups: genes with an associated function, and those annotated only as "hypothetical proteins". The latter group was further subdivided according to search results when queried against the NCBI CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) or the Pfam-A protein database, retrieved 15 Dec 2014 (using jackhmmer 3.1b1) [50]. Sequences that returned a hit from either (but not both) of these searches were labeled as "moderate confidence" with regard to their function. Sequence queries that produced no information via either method were classified into a "low confidence" group whose functions were weakly inferred according to the highest scoring subject sequence with an annotated function, when searched against the non-redundant protein database with BLASTP. HHpred, HHblits and jackhmmer were also used to annotate the moderate and low confidence sequence groups, as these methods apply hidden Markov models and are more sensitive than homology based methods like BLAST.

BLAST searches of the KEGG GENES database were performed in three iterations, with the aim of identifying the highest scoring subject sequence with an associated KEGG Orthology (KO) number. The first search was conducted using an E-value threshold of 10 and examined the top 100 BLAST hits per query. The second search used the same E-value threshold, but expanded to include the top 500 hits for each query. The third search reduced the E-value threshold to 1.0, and expanded the list of hits to a maximum of 10,000 per query.

CHAPTER 3 – RESULTS

3.1 – *S. aureofaciens* ATCC 10762 Genome Sequencing

After quality filtering, we obtained 2.46 Gb of Illumina sequences in 19.42 million short reads (3.90 million pairs and 12.84 million singletons) and 132.76 Mb of 454-shotgun sequence data in 209,530 reads with a mean length of 633 bp.

3.2 – Non-hybrid Assemblies

Six *de novo* assemblies of the *S. aureofaciens* ATCC 10762 genome were generated using only the Illumina short reads. Summary statistics are shown in Table 1. The non-hybrid assembly produced by SPAdes has the largest contig N50 (59,816 bp) and the fewest number of contigs overall (n = 574), with mean and maximum contig lengths of 16,131 and 412,063 bp, respectively. SPAdes also generated the assembly with the largest scaffold N50 (59,816 bp) and the fewest scaffolds (n = 393), with mean and maximum scaffold lengths of 23,722 and 685,539 bp, respectively. This assembly also includes all gap regions covered by PCR (n = 9). The SPAdes and SOAPdenovo2 assemblies exhibit the highest and lowest proportion of mapped Illumina reads, with 90.52% and 78.04%, respectively.

3.3 – Hybrid Assemblies

Two hybrid assemblies using Velvet and SPAdes were generated using both Illumina and 454 reads. Here also, the SPAdes assembly has the largest contig N50 length (228,235 bp, versus 46,576 bp from Velvet), but the scaffold N50 length for the

Velvet assembly is significantly larger (8,005,420 bp, versus 660,648 bp from SPAdes). The hybrid SPAdes assembly consists of 120 contigs in 60 scaffolds with respective N50 values of 228,235 bp and 660,648 bp (Table 1). Thus, this assembly is notably more contiguous than the non-hybrid SPAdes assembly which comprises 574 contigs (N50 = 59,816 bp) in 393 scaffolds (N50 = 155,320 bp). Notably, the total number of scaffolds produced by the non-hybrid SPAdes assembly (n = 393) and the hybrid velvet assembly (n = 389) are comparable. The hybrid SPAdes assembly also has a higher proportion of successfully mapped reads (90.56%) than the Velvet assembly (87.13%).

3.4 – Quality Assessment of Assemblies

We aligned the Velvet assembly against the IDBA and hybrid SPAdes assemblies to determine which gap regions could be closed on the Velvet scaffolds. We were able to close 109 gaps using IDBA, and 200 using the hybrid SPAdes assembled contigs. Additionally, we selected 34 regions predicted to have short gaps (<100bp) in scaffolds assembled using Velvet (Appendix 1). Among them, 14 targeted regions were successfully amplified. We observed no difference between PCR amplifications performed with and without the addition of DMSO. From these, we were able to sequence 9 regions (Appendix 2). Three of these sequences were of sufficient quality and did not require manual correction; the remaining 6 were corrected, using the provided chromatograms. These 9 sequences were used to evaluate our *de novo* assemblies; all 9 were correctly assembled (>50% query coverage and sequence identity when aligned via BLASTN) by every method except velvet (n = 0), SGA (n = 7) and SOAPdenovo2 (n = 8). This does not necessarily mean that the data from these sequences is missing within

these contig sets, but it does reflect the discontinuous nature of these assemblies (as SGA and SOAPdenovo2 have the largest number of contigs and the smallest N50 values of the assemblies evaluated). This adds to the evidence that these assemblers are performing poorly in this context. Predicted and actual gap sizes for all sequenced regions are shown in Appendix 2; 8 of the 9 regions have actual sequence lengths significantly longer than the predicted gap lengths. Only one sequence, spanning Velvet contigs 397 and 398, was shorter, with an actual length of 112 bp compared to a predicted length of 281 bp.

3.5 – Integration of Multiple Assemblies

CISA was used to merge different assemblies in four combinations (Table 3). Set 1, comprised of the IDBA, Trinity and Velvet assemblies, produced the assembly with the fewest number of contigs (n = 4,519) and the smallest total length (30,073,865 bp), but the largest N50 (18,974 bp). Set 4, which includes the assemblies from set 1 along with the SGA, MIRA and hybrid SPAdes assemblies, exhibits the largest total length (59,346,503 bp) and possesses the second-largest N50 (15,343 bp). The total lengths of the merged assemblies were notably and consistently larger (30 – 60 Mbp) than the total lengths of the assemblies produced by the corresponding individual methods. Individual assemblies, both hybrid and non-hybrid, provided a more consistent estimate of total *S. aureofaciens* genome length, in the range of 9.2 – 11.5 Mbp.

3.6 – Annotation of Genomic Features

The hybrid assembly generated by SPAdes was ultimately chosen as the best assembly owing to its high contiguity, low proportion of scaffold gaps and the superior

proportion of Illumina reads that map to it. We annotated this assembly in addition to the hybrid Velvet assembly provided by Eurofins. using the NCBI PGAP. Contaminant filtering of the hybrid SPAdes assembly resulted in the removal of four contigs that appeared to be of plasmid origin.

A significant difference in the number of genomic features was observed between the annotations of the two hybrid assemblies (Table 4). There are 1,393 more total genes and 205 more pseudogenes annotated within the annotation of SPAdes assembly. We identified 6,103 pairs of homologous coding sequences between the two annotations; of these, 5,270 are completely identical and 833 exhibit vary in length by at least 1 amino acid. We also observed sequences unique to both the Velvet (n = 21) and SPAdes (n = 192) annotations.

3.7 – Comparative Genomics

Presently, there are 5 other *S. aureofaciens* genomes available from NCBI (Table 5). We compared our SPAdes assembly of the *S. aureofaciens* ATCC 10762 genome to other publicly available *S. aureofaciens* genomes (Table 6). During the course of this work, five *S. aureofaciens* genomes were deposited in the NCBI Assembly database under accession numbers ASM71917v1, ASM97851v1, ASM71688v1, ASM72084v1 and ASM127066v1. These strains were designated as NRRL B-2657, NRRL 2209, NRRL B-1286, NRRL B-2183 and NRRL B-2658, respectively. We aligned our contigs from the hybrid SPAdes assembly against these assemblies using MUMmer (see section 2.4.1). Our assembled contigs are virtually identical to NRRL B-2657 / ASM71917v1 (99.75% total aligned bases) and only slightly divergent from NRRL 2209 /

ASM97851v1 (99.05% total aligned bases). Our assembly has a larger total length and N50, and a smaller number of contigs, compared to these assemblies. Interestingly, our assembled contigs differ significantly from NRRL B-1286 / ASM71688v1, NRRL B-2183 / ASM72084v1 and NRRL B-2658 / ASM127066v1 (83.69%, 10.31% and 9.92% total aligned bases, respectively). We also observe significant variation in the distribution of ATCC 10762 coding sequence homologs between these annotations, with NRRL B-2657 having the largest number of orthologous sequences (n = 7,483) while NRRL B-2183 and NRRL B-2658 have the fewest (n = 3,857 and 3,984, respectively), highlighting a potentially distant evolutionary relationship between the latter strains and ATCC 10762.

3.8 – Phylogenomic Analysis

Using 16S data from our *S. aureofaciens* annotation, we identified an additional set of 18 taxa, including *Streptacidiphilus* and *Kitasatospora* species, for further phylogenomic analyses. We identified orthologs of the 16S rRNA and *recA* genes (the latter having been selected for its known, high degree of conservation), and aligned sequences from this total set of taxa (n = 24) to reconstruct the maximum-likelihood phylogenies (Figures 2 – 3). In both trees, we observe *S. aureofaciens* strain ATCC 10762 clustering with *Kitasatospora* taxa, with large branch lengths between ATCC 10762 and *S. aureofaciens* strains NRRL B-2183 and NRRL B-2658, indicating greater than expected evolutionary distance (Figures 2 – 3).

3.9 – Functional Analysis

The NCBI annotation of our SPAdes assembly includes a large number of sequences of unknown function, annotated only as hypothetical proteins (n = 3,185). We examined the entire set of coding sequences from this assembly (n = 7,627). Our BLAST searches of the KEGG GENES database were able to associate some KEGG-described function, in the form of a KO number, with 5,783 sequences, including 1,786 sequences that were initially annotated at hypothetical proteins by NCBI. This represents 76% of the total CDS dataset and 56% of hypothetical proteins, respectively, from the hybrid SPAdes annotation.

Next, we identified a set of 72 sequences of interest, all annotated as hypothetical proteins by the NCBI pipeline. For these sequences, our combined searches of CDD, KEGG and Pfam were sufficient to infer function for 13 proteins with at least a moderate level of confidence (i.e., overlapping functional predictions endorsed by two or more independent search methods).

CHAPTER 4 – DISCUSSION

4.1 – *De novo* Assembly of *S. aureofaciens* ATCC 10762 Genome

As expected, we observed a substantial variation between the six non-hybrid assemblies. Overall, SOAPdenovo2 performed most poorly, producing an assembly with the largest number of contigs and the smallest contig N50 to which only 78% of reads could be mapped (Table 1). At the scaffold level, however, we observe that the SOAPdenovo2 assembly also has the largest total scaffold length and the largest scaffold N50. This highlights the danger of relying only on summary statistics to evaluate *de novo* genome assemblies, despite the widespread acceptance of this practice. Under a more comprehensive evaluation accounting for the percentage of mapped reads, coverage of known sequence regions (i.e., the gap regions sequenced by PCR) and a relatively small number of contigs and scaffolds, SPAdes clearly outperforms the competing non-hybrid assemblers.

The assemblies produced by CISA exhibit less overall variation than the set of non-hybrid assemblies, particularly with respect to the percentage of mapped reads (Table 3). The first three datasets are approximately equal by this metric, with 90.19, 90.21 and 90.24% of reads mapped, respectively. The fourth dataset is an exception, with only 88.89% of reads mapped. This suggests that merged assemblies based on the same data have a point of diminishing return, wherein relatively few new regions of the genome are covered with each successive addition. Additionally, we observed a consistent and significant overestimation of total genome size amongst all four CISA assemblies.

Both hybrid *de novo* assemblies are significantly more contiguous than those assemblies generated only from the Illumina reads (Tables 1 and 2). Even so, the two hybrid assemblies differ significantly from one another. Most notably, the Velvet assembly is comprised primarily of a single very large scaffold (8,005,420 bp) containing a large number (n = 310) of gap regions. These gaps represent 1.38% of the total bases in the assembly, or more than ten times the number of gap characters contained in all of the hybrid SPAdes assembled scaffolds. The SPAdes assembly generated 60 scaffolds with 57 total gap regions, representing 0.12% of all assembled bases. Why does the percentage of gaps present in scaffolds differ by more than an order of magnitude between these two assemblies, generated with the same input data? These assemblies represent notably different approaches, with significant implications. The pipeline implemented by Eurofins begins with assembly of the 454-shotgun reads by Newbler into contigs, onto which the paired-end Illumina reads are mapped. This allows them to infer the genome size and the insert sizes for each library, which are incorporated downstream as the 454 contigs and Illumina reads are assembled, then manually inspected. This results in an assembly with a deceptively high scaffold N50 of 8,005,420 bp, since the distribution of scaffold lengths is uneven, with the longest and second longest scaffold lengths equaling 8,005,420 bp and 52,293 bp, respectively. Excluding the longest scaffold, the remaining 388 scaffolds lengths sum to 1,451,044 bp or 15.3% of the total assembled scaffold length. A comparison of contig N50 lengths (Velvet: 46,576 bp; SPAdes: 228,235 bp) and the percentage of mapped reads (Velvet: 87.13%; SPAdes: 90.56%) implies that SPAdes is producing a better assembly (Table 2).

While both Velvet and SPAdes implement read error-correction algorithms [18,31], SPAdes performs a much larger number of functions overall, including: 1) iterative de Bruijn graph assembly using multiple k-mer sizes (similar to IDBA); 2) merging of these different assemblies, which facilitates better performance, particularly in cases where read coverage varies significantly and 3) contig error-correction, by aligning the original reads back to the contigs using the Burrows-Wheeler Aligner [51]. This allows SPAdes to take advantage of the information provided by very small k-mers (which are very sensitive, but not specific) and larger k-mers (which are specific, but not as sensitive). This is reflected across several metrics, including the percentage of mapped reads and the large number of Velvet scaffold gaps covered by the SPAdes contigs (n = 200).

The marked difference in the number of genomic features annotated within the two hybrid assemblies is difficult to interpret, as the core annotation software used by NCBI for this process (GeneMarkS+) underwent multiple, significant updates between the two submissions. Specifically, version 2.7 (released shortly after annotation of the Velvet contigs) implemented significant changes that improve the annotation of very short proteins (*e.g.*, leader peptides), and version 3.0 re-classified many partial proteins in the database as pseudogenes, affecting the annotation of proteins produced in the middle of contigs [52]. Even so, the NCBI PGAP process is necessarily conservative, as we observed during our efforts to gather additional information on sequences annotated only as hypothetical proteins.

4.2 - Comparative Genomics and Phylogenetic Analyses

The sequence of the 16S small ribosomal subunit is extremely conserved, and has long served as the gold standard for bacterial phylogenetic inference. However, low levels of 16S sequence diversity have been observed, which may make 16S-based phylogenetic analyses insufficient for confident inference of evolutionary relationships between closely related species, and there is no universal agreement on the level of 16S similarity required for definitive taxonomic classification [53–55]. The *recA* gene we selected, in combination with a large number of statistical replicates to assess clade support, provided a robust phylogenetic tree that implies evolutionary relatedness between the *S. aureofaciens* ATCC 10762 type strain and a number of *Kitasatospora* species. The distinguishing features of the *Streptomyces* and *Kitasatospora* genera have been debated for many years, with some proposing their union [56,57]. While our analysis is insufficient to make definitive claims about the relationship between these two genera, it does highlight continued need for robust bacterial classification schemes.

In addition to the evidence provided by our two individual gene phylogenies, the wide variation of homologous ATCC 10762 coding sequences observed within the other publicly available *S. aureofaciens* annotations suggests that two assemblies, NRRL B-2183 and NRRL B-2658, may be published under an incorrect taxonomic classification, as both assemblies only share approximately half of their coding sequences with *S. aureofaciens* ATCC 10762.

4.3 – Functional Analysis

As previously stated, additional functional analysis was necessary to infer the function of a large number of sequences whose functions were not predicted by the NCBI annotation pipeline. Our manual search of the KEGG GENES database produced a large quantity of additional information for these sequences without a large quantity of manual effort. This highlights the danger of relying on a single source for functional predictions and the value of integrating information from multiple database searches. It also highlights the need for intensive manual curation of gene annotations.

CHAPTER 5 – CONCLUSIONS

The aims of this project were: 1) to evaluate the performance of several *de novo* assembly strategies, particularly hybrid and non-hybrid approaches; 2) to assess the effectiveness of meta-assemblies, and 3) to better characterize *S. aureofaciens* ATCC 10762. We have shown that independent hybrid assemblies generated from the same input data can vary wildly, and that hybrid assembly approaches appear to outperform assembly strategies that rely on data generated only by a single sequencing platform. We have also shown that while merged assemblies generated with CISA may offer slightly more accurate representations of the genome than individual assemblies (according to the proportion of successfully mapped reads), they also significantly overestimate the actual genome size. Having thus selected the hybrid SPAdes assembly as the most robust, comparisons of this assembly with other, publicly available *S. aureofaciens* assemblies revealed significant genetic diversity, and phylogenetic and phylogenomic analyses support the notion that at least two of the publicly available *S. aureofaciens* assemblies may be taxonomically incorrect.

How should investigators robustly evaluate *de novo* genome assemblies? When is an assembly finished? In the absence of a proper reference genome (which may itself contain errors), these questions appear daunting. Ideally, a finished assembly should be exactly the same length as the biological molecule it represents. Manual assembly finishing remains a time- and labor-intensive task, but unfinished 'draft' genomes have enormous research value, even if all genes are not represented or contig order and orientation remain partially uncertain. Here, the draft genome annotation allowed us to perform the phylogenetic, phylogenomic and functional analyses that highlight

unexpected diversity among *S. aureofaciens* strains. Our analyses highlight the need for all investigators to have a realistic understanding of data quality and methodological limitations when assembling microbial genomes without a reference. In this regard, our work joins a growing body of literature[19,20,58,59] that asserts no single assembly strategy is objectively best across all contexts, and emphasizes the continued need for robust, empirical validation strategies. Future work must emphasize the development of such strategies and the importance of interleaving computational and empirical data, particularly for the purposes of functional and metabolic studies. We aim to conduct such studies for the purposes of more fully understanding *S. aureofaciens* and related species, given their enormous relevance to human health.

REFERENCES

1. Duggar BM. Aureomycin; a product of the continuing search for new antibiotics. *Ann N Y Acad Sci*. 1948;51(Art. 2):177-181.

2. Nelson ML, Levy SB. The history of the tetracyclines. *Ann N Y Acad Sci*. 2011;1241(1):17-32. doi:10.1111/j.1749-6632.2011.06354.x.

3. Darken MA, Berenson H, Shirk RJ, Sjolander NO. Production of Tetracycline by Streptomyces aureofaciens in Synthetic Media. *Appl Microbiol*. 1960;8(1):46-51.

4. Wright F, Bibb MJ. Codon usage in the G+C-rich Streptomyces genome. *Gene*. 1992;113(1):55-65. doi:10.1016/0378-1119(92)90669-G.

5. Zhou Z, Gu J, Li Y-Q, Wang Y. Genome plasticity and systems evolution in Streptomyces. *BMC Bioinformatics*. 2012;13(10):1-17. doi:10.1186/1471-2105-13-S10-S8.

6. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463-5467.

7. Saiki RK, Gelfand DH, Stoffel S, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*. 1988;239(4839):487-491.

8. Collins FS, Morgan M, Patrinos A. The Human Genome Project: Lessons from Large-Scale Biology. *Science*. 2003;300(5617):286-290. doi:10.1126/science.1084564.

9. Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol*. 2011;29(11):987-991. doi:10.1038/nbt.2023.

10. Frey UH, Bachmann HS, Peters J, Siffert W. PCR-amplification of GC-rich regions: “slowdown PCR.” *Nat Protoc*. 2008;3(8):1312-1317. doi:10.1038/nprot.2008.112.

11. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 2011;13(1):36-46. doi:10.1038/nrg3117.

12. Jünemann S, Sedlazeck FJ, Prior K, et al. Updating benchtop sequencing performance comparison. *Nat Biotechnol*. 2013;31(4):294-296. doi:10.1038/nbt.2522.

13. Loman NJ, Misra RV, Dallman TJ, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*. 2012;30(6):562-562. doi:10.1038/nbt0612-562f.

14. Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*. 2011;12:451. doi:10.1186/1471-2105-12-451.

15. Zavodna M, Bagshaw A, Brauning R, Gemmell NJ. The Accuracy, Feasibility and Challenges of Sequencing Short Tandem Repeats Using Next-Generation Sequencing Platforms. *PLOS ONE*. 2014;9(12):e113862. doi:10.1371/journal.pone.0113862.

16. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 2011;27(21):2957-2963. doi:10.1093/bioinformatics/btr507.

17. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*. 2014;30(1):31-37. doi:10.1093/bioinformatics/btt310.

18. Bankevich A, Nurk S, Antipov D, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol*. 2012;19(5):455-477. doi:10.1089/cmb.2012.0021.

19. Salzberg SL, Phillippy AM, Zimin A, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*. 2011. doi:10.1101/gr.131383.111.

20. Magoc T, Pabinger S, Canzar S, et al. GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*. 2013;29(14):1718-1725. doi:10.1093/bioinformatics/btt273.

21. Jünemann S, Prior K, Albersmeier A, et al. GABenchToB: A Genome Assembly Benchmark Tuned on Bacteria and Benchtop Sequencers. *PLOS ONE*. 2014;9(9):e107014. doi:10.1371/journal.pone.0107014.

22. Scott D, Ely B. Comparison of Genome Sequencing Technology and Assembly Methods for the Analysis of a GC-Rich Bacterial Genome. *Curr Microbiol*. 2014;70(3):338-344. doi:10.1007/s00284-014-0721-6.

23. Laluce C, Molinari R. Selection of a chemically defined medium for submerged cultivation of Streptomyces aureofaciens with high extracellular caseinolytic activity. *Biotechnol Bioeng*. 1977;19(12):1863-1884. doi:10.1002/bit.260191210.

24. Tripathi G, Rawal SK. Simple and efficient protocol for isolation of high molecular weight DNA from Streptomyces aureofaciens. *Biotechnol Tech*. 1998;12(8):629-631. doi:10.1023/A:1008836214495.

25. Lohse M, Bolger AM, Nagel A, et al. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res*. 2012;40(W1):W622-W627. doi:10.1093/nar/gks540.

26. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. In: Berger B, ed. *Research in Computational Molecular Biology*. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2010:426-440. http://link.springer.com/chapter/10.1007/978-3-642-12683-3_28. Accessed June 23, 2016.

27. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res*. 2012;22(3):549-556. doi:10.1101/gr.126953.111.

28. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644-652. doi:10.1038/nbt.1883.

29. Chevreux. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. http://www.bioinfo.de/isb/gcb99/talks/chevreux/main.html. Published 1999. Accessed June 23, 2016.

30. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*. 2012;1:18. doi:10.1186/2047-217X-1-18.

31. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821-829. doi:10.1101/gr.074492.107.

32. Gradnigo JS, Somerville GA, Huether MJ, et al. Genome Sequence of Streptomyces aureofaciens ATCC Strain 10762. *Genome Announc*. 2016;4(3):e00615-e00616. doi:10.1128/genomeA.00615-16.

33. Lin S-H, Liao Y-C. CISA: Contig Integrator for Sequence Assembly of Bacterial Genomes. *PLOS ONE*. 2013;8(3):e60843. doi:10.1371/journal.pone.0060843.

34. Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes. *Genome Biol*. 2004;5:R12. doi:10.1186/gb-2004-5-2-r12.

35. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072-1075. doi:10.1093/bioinformatics/btt086.

36. Hubé F, Reverdiau P, Iochmann S, Gruel Y. Improved PCR method for amplification of GC-rich DNA sequences. *Mol Biotechnol*. 2005;31(1):81-84. doi:10.1385/MB:31:1:081.

37. Mammedov T, Pienaar E, Whitney S, et al. A Fundamental Study of the PCR Amplification of GC-Rich DNA Templates. *Comput Biol Chem*. 2008;32(6):452-457. doi:10.1016/j.compbiolchem.2008.07.021.

38. Strien J, Sanft J, Mall G. Enhancement of PCR amplification of moderate GC-containing and highly GC-rich DNA sequences. *Mol Biotechnol*. 2013;54(3):1048-1054. doi:10.1007/s12033-013-9660-x.

39. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: A Fast Phage Search Tool. *Nucleic Acids Res*. 2011;39(suppl 2):W347-W352. doi:10.1093/nar/gkr485.

40. Carattoli A, Zankari E, García-Fernández A, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother*. 2014;58(7):3895-3903. doi:10.1128/AAC.02412-14.

41. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2.

42. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421. doi:10.1186/1471-2105-10-421.

43. Krzywinski M, Schein J, Birol İ, et al. Circos: An information aesthetic for comparative genomics. *Genome Res*. 2009;19(9):1639-1645. doi:10.1101/gr.092759.109.

44. Cheong W-H, Tan Y-C, Yap S-J, Ng K-P. ClicO FS: an interactive web-based service of Circos. *Bioinformatics*. 2015;31(22):3685-3687. doi:10.1093/bioinformatics/btv433.

45. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30(14):3059-3066.

46. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 2005;33(2):511-518. doi:10.1093/nar/gki198.

47. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol*. 2013;30(4):772-780. doi:10.1093/molbev/mst010.

48. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312-1313. doi:10.1093/bioinformatics/btu033.

49. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*. 2014;42(D1):D199-D205. doi:10.1093/nar/gkt1076.

50. Punta M, Coggill PC, Eberhardt RY, et al. The Pfam protein families database. *Nucleic Acids Res*. 2012;40(Database issue):D290-D301. doi:10.1093/nar/gkr1065.

51. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324.

52. Tatusova T, DiCuccio M, Badretdin A, et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res*. June 2016:gkw569. doi:10.1093/nar/gkw569.

53. Clarridge JE. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clin Microbiol Rev*. 2004;17(4):840-862. doi:10.1128/CMR.17.4.840-862.2004.

54. Janda JM, Abbott SL. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. *J Clin Microbiol*. 2007;45(9):2761-2764. doi:10.1128/JCM.01228-07.

55. Sacchi CT, Alber D, Dull P, et al. High Level of Sequence Diversity in the 16S rRNA Genes of Haemophilus influenzae Isolates Is Useful for Molecular Subtyping. *J Clin Microbiol*. 2005;43(8):3734-3742. doi:10.1128/JCM.43.8.3734-3742.2005.

56. Wellington EMH, Stackebrandt E, Sanders D, Wolstrup J, Jorgensen NOG. Taxonomic Status of Kitasatosporia, and Proposed Unification with Streptomyces on the Basis of Phenotypic and 16S rRNA Analysis and Emendation of Streptomyces Waksman and Henrici 1943, 339AL. *Int J Syst Evol Microbiol*. 1992;42(1):156-160. doi:10.1099/00207713-42-1-156.

57. Hsiao N, Kirby R. Comparative genomics of Streptomyces avermitilis. *Antonie Van Leeuwenhoek*. 2007;93(1-2):1-25. doi:10.1007/s10482-007-9175-1.

58. Bradnam KR, Fass JN, Alexandrov A, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*. 2013;2:10. doi:10.1186/2047-217X-2-10.

59. Earl D, Bradnam K, John JS, et al. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res*. 2011;21(12):2224-2241. doi:10.1101/gr.126599.111.

Table 1. Summary statistics for non-hybrid assemblies.

Assembler	# contigs [# scaffolds]	Total length (bp)	Max length (bp)	Mean length (bp)	N50 length (bp)	# PCR Sequences Present	% Mapped Reads
IDBA	1,249 [1,382]	9,236,484 [9,073,474]	94,475 [220,877]	7,395 [6,565]	18,458 [26,759]	9 / 9	88.63
SGA	11,319 [5,264]	9,890,301 [10,722,212]	33,489 [35,102]	874 [2,037]	2,459 [4,822]	7 / 9	87.09
SOAPdenovo2	36,660 [19,305]	12,282,331 [18,990,823]	16,806 [5,538,220]	335 [984]	1,713 [1,678,425]	8 / 9	78.04
SPAdes	574 [393]	9,259,003 [9,322,718]	412,063 [685,539]	16,131 [23,722]	59,816 [155,320]	9 / 9	90.52
MIRA[1]	5,385	10,158,828	97,273	1,887	8,106	9 / 9	89.51[2]
Trinity[1]	2,559	11,511,866	48,849	4,499	10,612	9 / 9	89.60[2]

Statistics for scaffold assemblies are shown in brackets.
[1]These assemblers do not produce scaffolds.
[2]Reads were aligned against assembled contigs.

Table 2. Summary statistics for hybrid assemblies.

Assembler	# contigs [# scaffolds]	Total length (bp)	Max length (bp)	Mean length (bp)	N50 length (bp)	# PCR Sequences Present	% Mapped Reads
SPAdes	120 [60]	9,234,994 [9,244,380]	881,164 [1,746,076]	76,958 [154,073]	228,235 [660,648]	9 / 9	90.56
Velvet	711 [389]	9,325,515 [9,456,464]	309,247 [8,005,420]	13,116 [24,310]	46,576 [8,005,420]	0 / 9	87.13

Scaffold counts include singleton (unplaced) contigs.

Table 3. Summary statistics for merged assemblies generated by CISA.

Input assemblies[1]	# contigs	Total length (bp)	Max length (bp)	Mean length (bp)	N50 length (bp)	% Mapped Reads
1: IDBA + Trinity + Velvet	4,519	30,073,865	309,247	6,655	18,974	90.19
2: 1+ SGA	15,838	39,964,166	309,247	2,523	12,052	90.21
3: 2 + MIRA	21,072	50,111,509	309,247	2,378	10,784	90.24
4: 3 + SPAdes (hybrid)	21,192	59,346,503	881,164	2,800	15,343	88.89

[1]The contig sets merged by CISA.

Table 4. Comparison of annotated genomic features.[1]

Feature type	Number annotated	
	Velvet assembly	SPAdes assembly
Genes (total)	6,680	8,073
Protein coding genes	6,401	7,627
Pseudogenes[2]	144	349
Ribosomal RNA	37	22
Transfer RNA	74	72
Non-coding RNA	24	3

[1]NCBI RefSeq accession numbers: NZ_JPRF00000000.1 (Velvet assembly) and NZ_JPRF00000000.2. (SPAdes assembly).
[2]Includes incomplete sequences and entries with frameshifts and premature stop codons.

Table 5. Proportion of aligned bases and conserved coding sequences identified between *S. aureofaciens* ATCC 10762 and other *S. aureofaciens* assemblies.

Strain name	Accession No.	# contigs	# coding sequences	Total percentage of aligned bases	# conserve CDS
S. aureofaciens ATCC 10762	ASM118895v2	107	7,627	-	-
S. aureofaciens NRRL B-2657	ASM71917v1	279	7,587	99.75	7,483
S. aureofaciens NRRL 2209	ASM97851v1	989	7,395	99.05	7,302
S. aureofaciens NRRL B-1286	ASM71688v1	505	7,591	83.69	6,388
S. aureofaciens NRRL B-2183	ASM72084v1	167	7,367	10.31	3,857
S. aureofaciens NRRL B-2658	ASM127066v1	269	7,874	9.92	3,984

Table 6. Comparison of protein coding sequences annotated in Velvet and SPAdes assemblies of *S. aureofaciens* ATCC 10762.

Sequence category	Number annotated	
	Velvet assembly	SPAdes assembly
Identical CDS	5,270	
Non-identical CDS[1]	833	
Unique to annotation[2]	21	192

[1]These coding sequences differ in length between the two annotations.

[2]These coding sequences appear only in the indicated annotation.

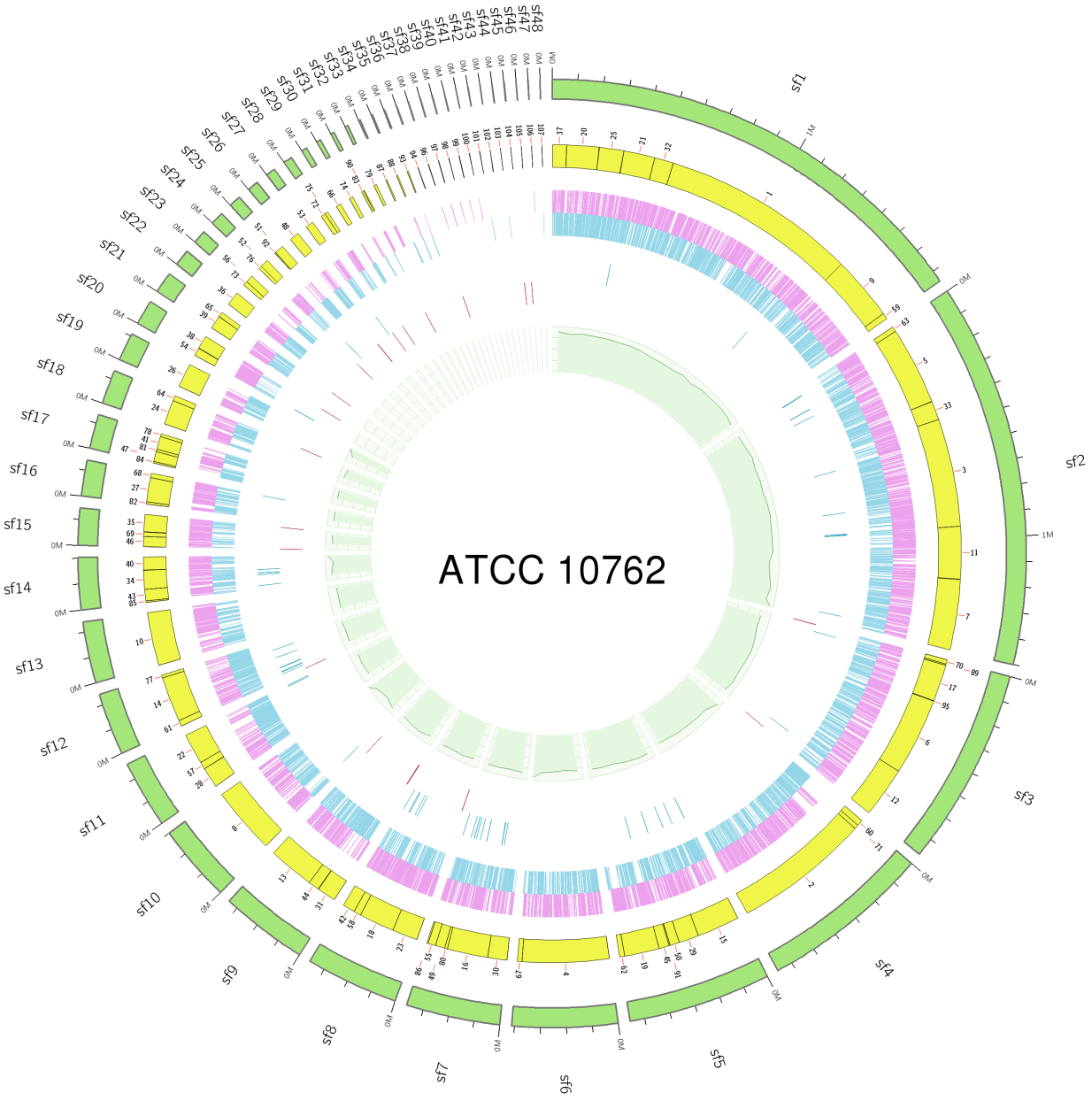

Figure 1. Circular genome plot of *S. aureofaciens* ATCC 10762. Scaffolds from the hybrid SPAdes assembly are plotted in descending order of scaffold length. From the outermost ring, the following elements are shown: scaffolds, contigs, forward strand CDS, reverse strand CDS, transfer RNAs, ribosomal RNAs and G+C content.

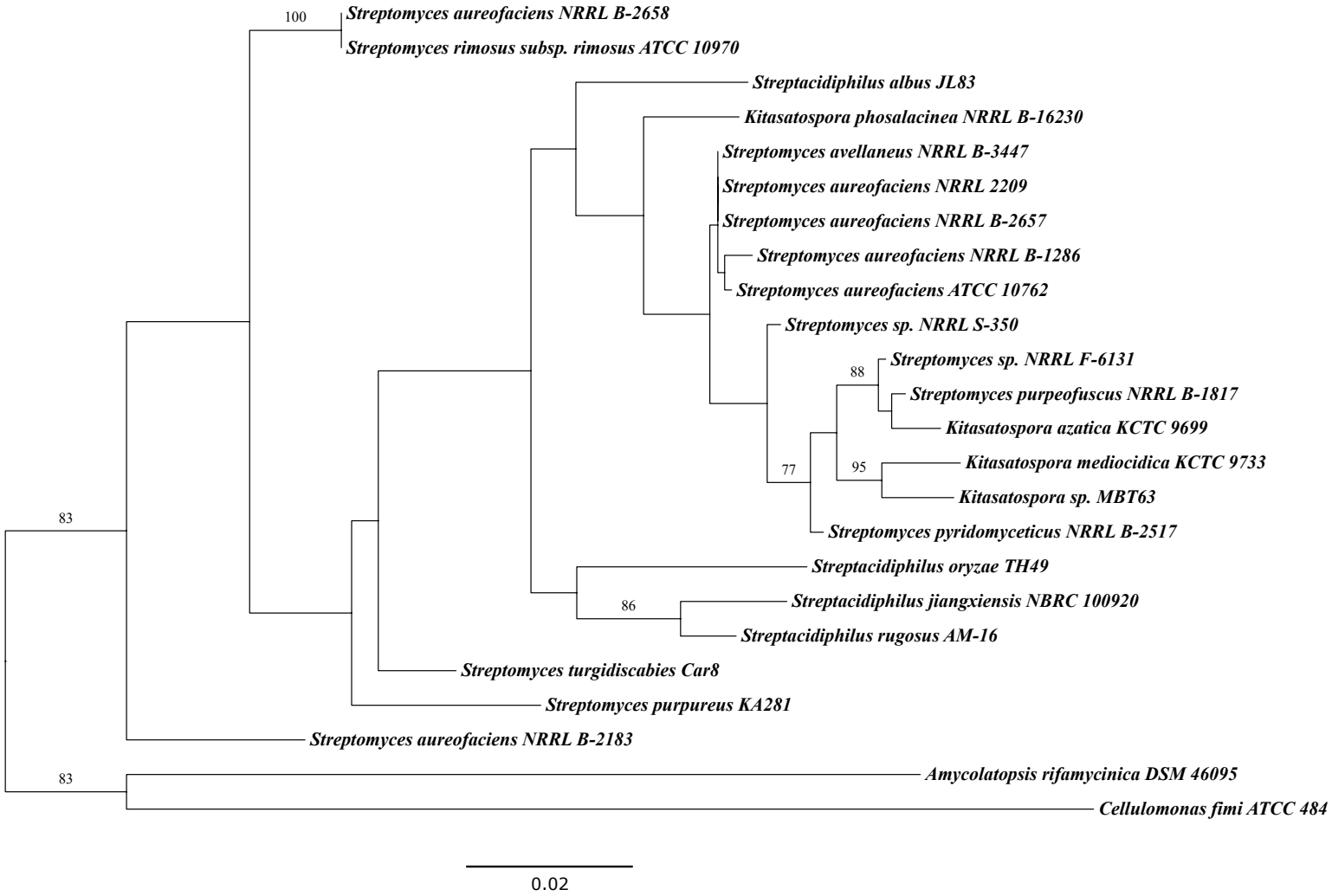

Figure 2. Maximum likelihood phylogeny of *S. aureofaciens* ATCC 10762 and neighboring species inferred from 16S rRNA gene sequence. The phylogeny was reconstructed with RAxML as decribed in Materials and Methods. Bootstrap values for well supported clades (≥ 70%) are shown. The scale bar indicates the number of nucleotide substitutions per site.

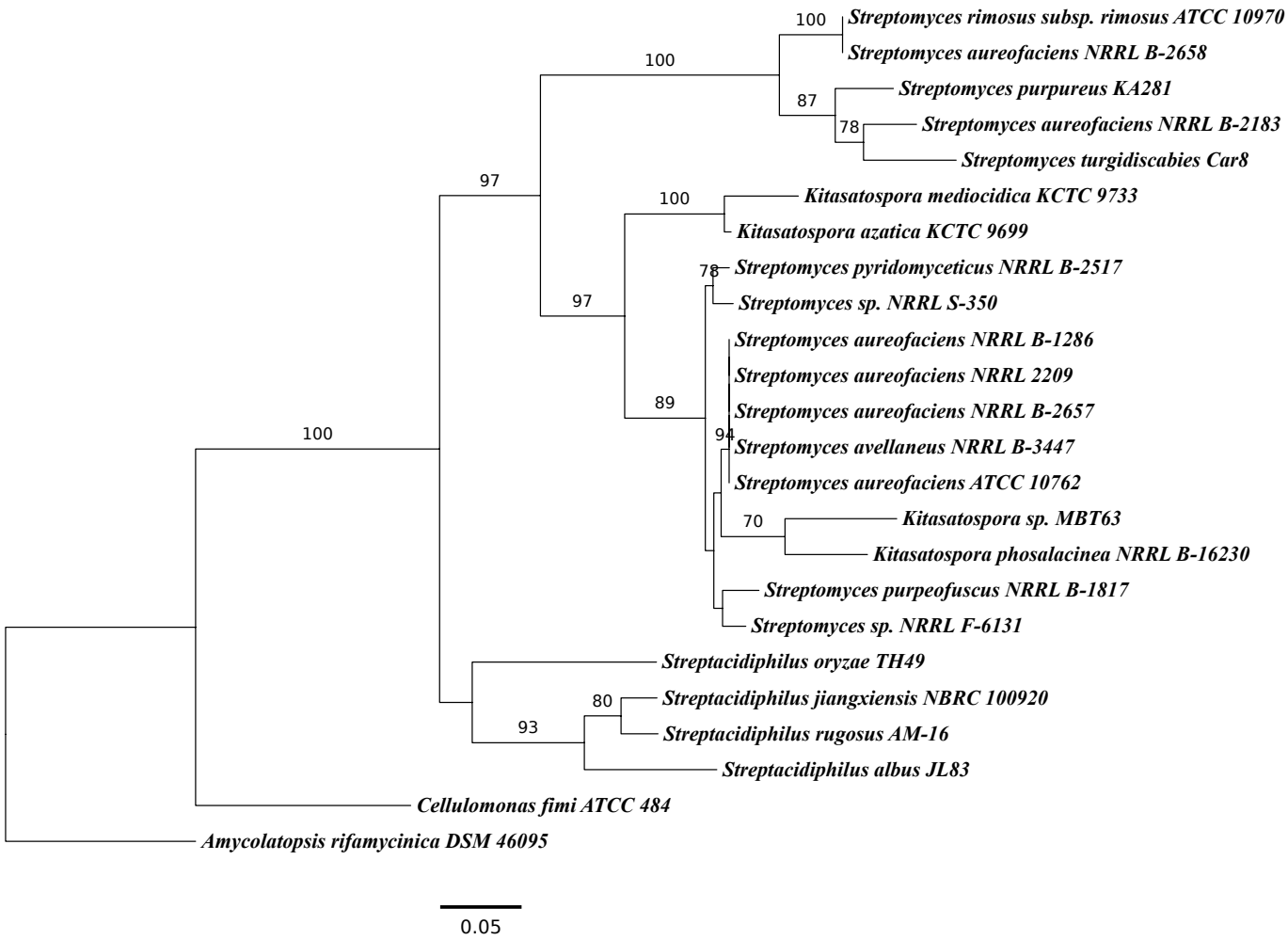

Figure 3. Maximum likelihood phylogeny of *S. aureofaciens* ATCC 10762 and neighboring species inferred from *recA* gene sequence. The phylogeny was reconstructed with RAxML as decribed in Materials and Methods. Bootstrap values for well supported clades (≥ 70%) are shown. The scale bar indicates the number of nucleotide substitutions per site.

APPENDICES

Appendix 1: List of PCR primers for sequencing of gap regions found in the Velvet scaffolds.

Region name	Forward primer (5' – 3')	Reverse primer (5' – 3')	Amplified	Sequenced?
contig_392_393	GAAGGTGTCGTGGTCCATCT	CTGGAGAAGTCGGACGGTTC		
contig_397_398	ACTGGGCGCAGATCCTCT	CTCGTGCATGAAGCGTTGC	x	x
contig_398_399	ACTGGGCGCAGATCCTCT	CTCGTGCATGAAGCGTTGC	x	x
contig_399_400	CCGAGATCACCGTCATGGTC	GGTAGCAGTCGTCGATCCG		
contig_401_402	CCGATAGTTCCGCCTGTACG	CACCGGATGAGCCTGTTGTA		
contig_404_405	GCTGAGATGGAACTCGCAGA	CAACTCTGCCGGGCGTC	x	x
contig_411_412	ACGCTTCGGTCTCGGG	TTCGGCGTGCCTGTTTATCG		
contig_416_417	AAGAACGCGAACCGCCA	GGCGGTCACCGAACCG	x	
contig_428_429	ACACCGTCTTGGCGATCTG	CCAACGATCGATCAGGAGCA	x	x
contig_437_438	ATGGAACCGCGCTTGAGG	CGGCCTCGCCTACACC		
contig_444_445	GAACGGGAACGGCTGGAG	GTTCTCGGTGGAGGTGCC		
contig_452_453	AAGGGATCGTCCCAGGTCA	GACGATCACGTCGCTCATCA		
contig_456_457	GCGGGCGGCTCGTATAAC	GACGGTCGAACTACGCTTCC		
contig_460_461	CAGTTCGTCCCACTCCTCGG	CGGACAAGCCGACCACAC		
contig_462_463	TCCTGGACACTGACGCACA	AATCGCCCGGAGTTTCGAG		
contig_466_467	GCAGTCCCGACGACCAGAG	CGAGGATCAGCGGCGTCT		
contig_470_471	CGACGTAGCCGAGCGTG	CGCAGGCCGCTGTCA		
contig_487_488	AGTTGCACTCTACGGGGTGA	CAAGTATTCGTGCAGACACGG		
contig_512_513	CGGGCCAAGGGGTTAGTTAC	GCCTTCGGGCTCACCTT		
contig_523_524	CTGCTCGACACCGCCC	CGAGCAGCCATTCGACCG		
contig_533_534	GGCGAATGTCCACCGAGC	CCCTCGTAGCGGTCGAACA	x	
contig_536_537	CCACCAGCAGCCAGTTCA	GTGGTGATCGTGGACGAGG		
contig_554_555	CGCTGGCGACCGAGAAC	CGCCGTACCGGAGCAC		
contig_557_558	GACTGCTCGCCGAAGCC	CCCGGGTCAACTCGCCTT	x	x
contig_561_562	TGGAGTTCGGCTACGAGACC	CAGGCGCTCATGCTCGAAG		
contig_604_605	TACGGGAGTTGGGTGGAGAG	CCAACTACGCCTACGAGCG	x	x
contig_631_632	CCCCTGTGATCCCGTGAAG	CGATCATGGTGAACTCCGGC		
contig_634_635	GACCCTCAGGCGGTAAGG	GGCACCCTGGTCGTTCC	x	
contig_635_636	GTAGGTCGGAAGCTCGACGG	CCAGGAGACGATCGAGGACG		
contig_636_637	AGGAGACCGTCCAGGTCC	TGTCCTCCTTCGGGGTCAG	x	
contig_641_642	GAGGTCCTTGAAGGGGTGC	GTCACCTGGGAGCGGTTC	x	x
contig_644_645	CCAGTACTCCATTTGCCGC	TTCCACGCCAAGCACGAC	x	
contig_651_652	AGCGAAACACGGAGACATAGA	GGGATTCGACGGTGTACGA	x	x
contig_699_700	TTCGCATGCGGTTGGAGAT	GGTGGTCCCTATCAGCGTG	x	x

Appendix 2 – Lengths of gap regions predicted by Velvet compared to the actual sequences.

Region name	Predicted gap length (bp)	Actual length (bp)	G+C content (%)
contig_397_398*	281	112	85.05
contig_398_399*	148	235	80.85
contig_404_405	10	142	82.98
contig_428_429	10	101	80.20
contig_557_558*	10	234	82.70
contig_604_605	10	64	85.94
contig_641_642*	10	61	81.97
contig_651_652*	10	108	57.41
contig_699_700*	10	50	70.00

*Sequences were manually corrected according to the corresponding chromatogram.

Appendix 3 – Full-text PDF of Gradnigo et. al., 2016.

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

Genome Sequence of *Streptomyces aureofaciens* ATCC Strain 10762

Julien S. Gradnigo,[a] Greg A. Somerville,[b] Michael J. Huether,[c] Richard J. Kemmy,[c] Craig M. Johnson,[c] Michael G. Oliver,[c] Etsuko N. Moriyama[a,d]

School of Biological Sciences, University of Nebraska–Lincoln, Lincoln, Nebraska, USA[a]; School of Veterinary Medicine and Biomedical Sciences, University of Nebraska–Lincoln, Lincoln, Nebraska, USA[b]; Zoetis, Lincoln, Nebraska, USA[c]; Center for Plant Science Innovation, University of Nebraska–Lincoln, Lincoln, Nebraska, USA[d]

***Streptomyces aureofaciens* is a Gram-positive actinomycete that produces the antibiotics tetracycline and chlortetracycline. Here, we report the assembly and initial annotation of the draft genome sequence of *S. aureofaciens* ATCC strain 10762.**

Received 9 May 2016 **Accepted** 11 May 2016 **Published** 23 June 2016

Citation Gradnigo JS, Somerville GA, Huether MJ, Kemmy RJ, Johnson CM, Oliver MG, Moriyama EN. 2016. Genome sequence of *Streptomyces aureofaciens* ATCC strain 10762. Genome Announc 4(3):e00615-16. doi:10.1128/genomeA.00615-16.

Address correspondence to Etsuko N. Moriyama, emoriyama2@unl.edu.

Streptomyces aureofaciens was first identified in 1948 (1), from Plot 23 in Sanborn Field, a timothy hayfield at the University of Missouri (2). Although it has been used for the commercial production of tetracycline antibiotics (1, 3) and has been the subject of numerous biochemical studies, a genome assembly of *S. aureofaciens* was not publicly available. Here, we report the draft assembly of the whole-genome sequence of the *S. aureofaciens* ATCC strain 10762 and its initial annotation.

S. aureofaciens ATCC strain 10762 was cultivated in a chemically defined medium [4] and grown for 9 days at 30°C with 150-rpm aeration. High-molecular-weight DNA was prepared from fresh cultures by cetyltrimethylammonium bromide extraction (5). Genomic DNA was evaluated for molecular weight integrity by agarose gel electrophoresis and nucleic acid fluorometric quantitation for construction of the DNA library.

Genome sequencing and read quality filtering were done by Eurofins MWG Operon (Alabama, USA). Illumina MiSeq sequencing was done with long-jumping-distance sequencing (3-kb and 8-kb inserts), generating paired-end 150-bp reads. After removing very short (<30 bp) reads, adapter-trimming, and quality-clipping using Trimmomatic (6), 2.46 Gb of sequence information in 19.42 million reads (3.90 million pairs and 12.84 million singletons) were obtained. Shotgun sequencing on the Roche 454 Genome Sequencer FLX platform produced 132.76 Mb of sequence data in 209,530 reads with a mean length of 633 bp after trimming with Trimmomatic.

Genome assembly was done using SPAdes version 3.7.1 with the "--careful" option (to reduce mismatches and short indels) (7) with Illumina (both paired-end and singleton) and 454 reads, resulting in 120 contigs (total length: 9,234,994 bp; maximum length: 881,164 bp; N_{50}: 228,235 bp) in 60 scaffolds (total length: 9,244,380 bp; maximum length: 1,746,076 bp; N_{50}: 660,648 bp). The average G+C content was 71.2%.

Genome annotation was done using the NCBI Prokaryotic Genome Annotation Pipeline version 3.1 (8). In total, 13 contigs were removed from the annotation because they were <200 bp or were of non–*S. aureofaciens* origin (e.g., plasmid). From the remaining 107 contigs, a total of 8,083 genes were annotated. This includes 7,541 protein-coding genes, 22 rRNA genes (5S, 16S, and 23S), 72 tRNA genes, and 445 potential pseudogenes.

Four related *S. aureofaciens* strains were recently deposited in the NCBI Assembly database (ASM71917v1, ASM97851v1, ASM71688v1, and ASM127066v1; these strains were designated NRRL B-2657, NRRL 2209, NRRL B-1286, and NRRL B-2658, respectively). Our assembled contigs are virtually identical to ASM71917v1 (99.75% total aligned bases using MUMmer version 3.23 [9]) and only slightly divergent from ASM97851v1 (99.05% total aligned bases). Our assembly has a greater total length and N_{50} and a smaller number of contigs compared to these assemblies. Interestingly, our assembled contigs differ significantly from ASM71688v1 and ASM127066v1 (83.69% and 9.92% total aligned bases, respectively).

Nucleotide sequence accession numbers. This whole-genome shotgun project of *S. aureofaciens* ATCC strain 10762 has been deposited in DDBJ/ENA/GenBank under the accession number JPRF00000000. The version described in this paper is the second version, JPRF02000000.

FUNDING INFORMATION

This work, including the efforts of Julien S. Gradnigo, Greg A. Somerville, and Etsuko N. Moriyama, was funded by Zoetis.

This research was supported by a contract from Zoetis, Inc.

REFERENCES

1. **Duggar BM.** 1948. Aureomycin; a product of the continuing search for new antibiotics. Ann N Y Acad Sci 51:177–181. http://dx.doi.org/10.1111/j.1749-6632.1948.tb27262.x.
2. **Nelson ML, Levy SB.** 2011. The history of the tetracyclines. Ann N Y Acad Sci 1241:17–32. http://dx.doi.org/10.1111/j.1749-6632.2011.06354.x.
3. **Darken MA, Berenson H, Shirk RJ, Sjolander NO.** 1960. Production of tetracycline by *Streptomyces aureofaciens* in synthetic media. Appl Microbiol 8:46–51.
4. **Laluce C, Molinari R.** 1977. Selection of a chemically defined medium for submerged cultivation of *Streptomyces aureofaciens* with high extracellular caseinolytic activity. Biotechnol Bioeng 19:1863–1884. http://dx.doi.org/10.1002/bit.260191210.
5. **Tripathi G, Rawal SK.** 1998. Simple and efficient protocol for isolation of

high molecular weight DNA from *Streptomyces aureofaciens*. Biotechnol Tech 12:629–631. http://dx.doi.org/10.1023/A:1008836214495.

6. **Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B.** 2012. *RobiNA*: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res 40:W622–W627. http://dx.doi.org/10.1093/nar/gks540.
7. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA.** 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. http://dx.doi.org/10.1089/cmb.2012.0021.
8. **Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Ciufo S, Li W.** 2013. Prokaryotic Genome Annotation Pipeline. The NCBI Handbook [internet], 2nd ed. NCBI, Bethesda, MD. http://www.ncbi.nlm.nih.gov/books/NBK174280.
9. **Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL.** 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12. http://dx.doi.org/10.1186/gb-2004-5-2-r12.