5-2016

Assessment of Next Generation Sequencing Technologies for *De novo* and Hybrid Assemblies of Challenging Bacterial Genomes

Sagar Mukund Utturkar
University of Tennessee - Knoxville, sutturka@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Sagar Mukund Utturkar entitled "Assessment of Next Generation Sequencing Technologies for *De novo* and Hybrid Assemblies of Challenging Bacterial Genomes." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Steven D. Brown, Major Professor

We have read this dissertation and recommend its acceptance:

Christopher W. Schadt, Mitchel J. Doktycz, Dale A. Pelletier, Gladys Alexandre

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Assessment of Next Generation Sequencing Technologies for *De novo* and Hybrid Assemblies of Challenging Bacterial Genomes

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Sagar Mukund Utturkar
May 2016

Dedicated to my beloved grandmother, late Mrs. Jayashree Gavankar,
For her prayers and unconditional love

ACKNOWLEDGEMENTS

I would like to thank the people who helped me during my research and dissertation. First and foremost, I would like to thank my advisor Dr. Steven D. Brown for insightful guidance, advice, support and encouragement throughout my Ph.D. curriculum. Dr. Brown showed immense patience during my early learning years, encouraged me for manuscript writing, presented me with right opportunities to work on collaborative projects, provided freedom to express my ideas and included me in interactions with top scientists. His goal-oriented research, dedicated nature and friendly adulation made him my role model for a scientist, mentor and a teacher. He transformed me into a better scientist and stronger person and I will be indebted to him all my life for his kindness.

I would also like to thank my distinguished committee members Dr. Mitch Doktycz, Dr. Dale Pelletier, Dr. Chris Schadt and Dr. Gladys Alexandre for their time, efforts, suggestions and critical review to move my research forward. I would like to express my special thanks to Dr. Mircea Podar for external help towards this dissertation research and collaborative research opportunities.

I would like to acknowledge the Genome Science and Technology program, University of Tennessee, Plant-Microbe Interfaces project and Oak Ridge National Laboratory for providing financial support and excellent work environment. I was never limited by the tools and resources required to perform productive research.

I have had the pleasure of working with amazing colleagues in Dawn Kilngeman, Charlotte Wilson, Kyle Sander, Miguel Rodriguez, Punita Manga, Chia-wei Wu, and Alex Dumitrache. I want to have a special mention of Dawn Klingeman for teaching me various wet-lab techniques and performing all the sequencing runs to make this research possible, Charlotte Wilson for sharing thoughts and laughs, and Miguel for being a nice friend and providing stimulating lab environment. I would also like to thank several people from computational biology group at Oak Ridge National Laboratory including Steve Moulton and Michael Galloway for providing technical help required during my research and Miriam Land for providing outstanding support with various computational tools, scripts and ideas.

I express my deepest gratitude for my parents, Mr. Mukund Utturkar and Mrs. Vidya Utturkar who were always besides me and provided freedom to pursue my dreams. I won't be where I am without them. I would like to thank my wife, Ketaki Bhide for her continuous support and staying strong during these years. I also want to thank my grandfather Mr. Vinayak Gavankar, my aunt Mrs. Varsha Agashe, and my in-laws for the encouragement and support. Finally, words will be limited to acknowledge the role of my friends in Knoxville, especially Snehal Joshi, Sarvesh Iyer and Snigdha Sewlikar, who are like my second family and never let me miss my home.

ABSTRACT

In past decade, tremendous progress has been made in DNA sequencing methodologies in terms of throughput, speed, read-lengths, along with a sharp decrease in per base cost. These technologies, commonly referred to as next-generation sequencing (NGS) are complimented by the development of hybrid assembly approaches which can utilize multiple NGS platforms. In the first part of my dissertation I performed systematic evaluations and optimizations of nine *de novo* and hybrid assembly protocols across four novel microbial genomes. While each had strengths and weaknesses, via optimization using multiple strategies I obtained dramatic improvements in overall assembly size and quality. To select the best assembly, I also proposed the novel rDNA operon validation approach to evaluate assembly accuracy. Additionally, I investigated the ability of third-generation PacBio sequencing platform and achieved automated finishing of *Clostridium autoethanogenum* without any accessory data. These complete genome sequences facilitated comparisons which revealed rDNA operons as a major limitation for short read technologies, and also enabled comparative and functional genomics analysis. To facilitate future assessment and algorithms developments of NGS technologies we publically released the sequence datasets for *C. autoethanogenum* which span three generations of sequencing technologies, containing six types of data from four NGS platforms. To assess limitations of NGS technologies, assessment of unassembled regions within Illumina and PacBio assemblies was performed using eight microbial genomes. This analysis confirmed rDNA operons as major breakpoints within Illumina assembly while gaps within PacBio assembly appears to be an unaccounted for event and assembly quality is cumulative effect of read-depth, read-quality, sample DNA quality and presence of phage DNA or mobile genetic elements. In a final collaborative study an enrichment protocol was applied for isolation of live endophytic bacteria from roots of the tree *Populus deltoides*. This protocol achieved a significant reduction in contaminating plant DNA and enabled use these samples for single-cell genomics analysis for the first time. Whole genome sequencing of selected single-cell genomes was performed, assembly and contamination removal optimized, and followed by the bioinformatics, phylogenetic and comparative genomics analyses to identify unique characteristics of these uncultured microorganisms.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 : INTRODUCTION

1.1 Background

Sanger and colleagues (Sanger, et al., 1977) and Maxam and Gilbert (Maxam and Gilbert, 1977) first developed methods to sequence DNA by chain termination and selective chemical cleavage of DNA, respectively. Maxam and Gilbert technique involves radioactive labelling at the 5' end of the DNA fragment, selective chemical cleavage at one or two bases (e.g. G, A+G, C, and C+ T) to generate series of labelled DNA fragments which are then separated by gel electrophoresis (Maxam and Gilbert, 1977). Fragment separation is visualized by autoradiography and base calling is performed by interpretation of banding pattern relative to chemical cleavage reactions. The technique developed by Sanger and colleagues, commonly referred to as Sanger sequencing became more popular because it required less handling of toxic chemicals and radioisotopes. This Sanger sequencing principle prevailed for the next 30 years and is still in use today with several modification including the use of fluorescent labelled nucleotides and reaction multiplexing. A growing demand for increased throughput and rapid technical advancements in laboratory automation, miniaturization and process parallelization, started a new revolution in sequencing technologies. These advancements enabled the use of Sanger technique to be used to complete the first human genome sequence in 2004 (International Human Genome Sequencing, 2004). However, the Human Genome Project required vast amounts of time and resources and further improvements for even faster, cheaper, and higher-throughput sequencing methods were seen as necessary. For this reason, the National Human Genome Research Institute (NHGRI) initiated a funding program with the goal of reducing the cost of human genome sequencing to US$1000 in ten years. This stimulated the development and commercialization of next-generation sequencing (NGS) technologies, which are characterized by the remarkable increase in the sequencing efficiency on the order of approximately 10^6 fold (Treangen and Salzberg, 2012). These new NGS technologies share three major characteristics. First, they rely on the preparation of NGS libraries in a cell free system instead of traditional bacterial cloning of DNA fragments. Second, these systems parallelize many millions of sequencing reactions on minuscule platforms, which dramatically increase sequencing throughput. Third, the base interrogation is performed directly through fluorescent or other forms of electrical or chemical signals without the need for the electrophoresis. All of this combined generates enormous number of sequencing reads at an unprecedented speed (van Dijk, et al., 2014).

The first NGS technology to be released in 2005 was the pyrosequencing method by 454 Life Sciences (now Roche) (Margulies, et al., 2005). During the next ten years several NGS platforms including 454, Illumina, SOLiD, Ion Torrent, Pacific Biosciences (PacBio) and Oxford Nanopore MinION sequencing were released and offered improvements in read length and output (van Dijk, et al., 2014). In general, the Illumina, 454, Ion Torrent and SOLiD are classified as second generation sequencing platforms, which are characterized by the shorter read lengths with high accuracy (Mavromatis, et al., 2012; Quail, et al., 2012) while PacBio and Nanopore are so called third generation sequencing platforms which generates significantly longer, but fewer and more error prone reads (Brown, et al., 2014; Koren and Phillippy, 2014; Madoui, et al., 2015). Particularly impressive increases in the sequencing throughput were achieved by the Illumina technology which currently offers the highest throughput per run and the lowest per-base

cost (Liu, et al., 2012). Performance comparisons of various NGS platforms and their recent advances are summarized in many publications (Brown, et al., 2014; Liu, et al., 2012; Quail, et al., 2012). Here, I provide brief overview of various sequencing platforms and discuss the most significant improvements.

Roche 454 pyrosequencing: (Figure 1.1)

The 454 pyrosequencing system is capable of sequencing 400-600 megabases of DNA per 10 hour run on the 454 GS FLX sequencing machine (Margulies, et al., 2005). The system employs emulsion bead-based sequencing in which nebulized and adapter ligated DNA fragments are captured in the beads in the water-in-oil emulsion and amplified by PCR (Metzker, 2010). Each DNA bound bead is placed into a ~ 29 um diameter well on a pico-titer plate and a mix of enzymes including DNA polymerase, ATP sulfurylase, and luciferase are also packed into the well. The sequencing principle is based on detecting the activity of DNA polymerase with another chemiluminescent enzyme. The adapter on the DNA fragment serves as a primer for the addition of deoxynucleoside triphosphates (dNTPs) by DNA polymerase to synthesize complementary DNA strand. The incorporation of dNTP releases pyrophosphate (PPi) which is converted into ATP by enzyme ATP sulfurylase. This ATP acts as a substrate for the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are proportional to the amount of ATP (Ronaghi, et al., 1998). The light produced by the luciferase catalyzed reaction is detected by the instrument control software for correct base-calling. Unincorporated nucleotides and ATP are degraded by the apyrase after each base cycle. The entire process is run in parallel where nucleotides are flowed in order, and multiple identical bases can be incorporated in single cycle. The addition of multiple bases is associated with higher signal intensity.

The 454 Genome Sequencer generates about 1,000,000 reads per sequencing run with an average read-length of up to 700 bp (Luo, et al., 2012; Utturkar, et al., 2014). The 454 supports the mate-pair library preparation protocol (which allows for large insert sizes; up to 8 kb), which provides added advantage for the downstream genome assembly process. The pros of the 454 sequencing include longer reads that are easier to map to the reference genome, and advantageous for *de novo* genome assemblies or metagenomics analysis. The run times were relatively fast as compared to technologies available in 2005. The cons include relatively low throughput, high reagent cost, and high error rates in homopolymer repeats. The inhibition of apyrase enzyme by Rp isomer of natural dATP nucleotide had a major impact on pyrosequencing read lengths (Gharizadeh et al., 2002). Another inherent problem is the difficulty in determining the number of incorporated nucleotides in homopolymeric region, due to the nonlinear light response associated with 5-6 nucleotides. Roche have recently pulled out of the sequencing business owing to the availability of cheaper and inherently higher throughput sequencing technologies, and the legacy 454 sequencing method is no longer being supported (van Dijk, et al., 2014).

Illumina (Figure 1.2)

The Illumina company is currently the most widely used NGS sequencing technology owing to lower per-base costs, streamlining and automation of library generation and instrument operations, and quantity of data generated (Mavromatis, et al., 2012). In

principle, the concept of Illumina sequencing is similar to initial Sanger sequencing method where DNA polymerase catalyzes the incorporation of fluorescently labelled dNTPs into a DNA template strand during sequential cycle of DNA synthesis. During each cycle, at the point of incorporation, the nucleotide is identified by fluorophore excitation. The standard Illumina sequencing workflow involves four basic steps, (i) library preparation – which involves random fragmentation of DNA, 5' and 3' adapter ligation, PCR amplification of adapter-ligated fragments and purification. (ii) Cluster generation – involves loading of the library into a flow cell where fragments are captured on a lawn of surface bound oligos complimentary to the library adapters. Each fragment is then amplified into distinct, clonal clusters through bridge amplification. (iii) Sequencing – Illumina uses a reversible terminator-based method that detects single bases as they are incorporated into DNA template strand (iv) Data analysis – where intensity values are converted to actual base-calls and sequence reads are generated for downstream analysis. This four step process delivers many, accurate sequence reads.

In early 2010, the release of the Illumina Genome Analyzer and the HiSeq 2000 instruments set the standard for high throughput massively parallel sequencing. In 2011, Illumina released a lower throughput instrument platform called the MiSeq that uses the same chemistry at scale more appropriate for smaller laboratories and clinical diagnostics labs (Quail, et al., 2012).The MiSeq instrument can generate up to 25 million reads in single sequencing run with 15 Gb data output and supports paired-end reads up to 300 bp. Recently Illumina released the HiSeq X Ten system, a set of ten HiSeq X sequencing machines with a massive capacity of 1.8 Tb of sequencing per run (http://www.illumina.com). This system is speculated to have broken the barrier of the US$1000 genome corresponding to original goal of the NHGRI funding program. It should be noted that US$1000 per genome cost assumes all HiSeq X machines running at full capacity. However, the massive cost (approximately US$10 million) of this system makes it available only to the large institutional users performing population-scale genome sequencing (van Dijk, et al., 2014). Another Illumina sequencing instruments include synthetic long read technology, previously known as Moleculo (McCoy, et al., 2014; Voskoboynik, et al., 2013), which relies on an advanced library preparation protocol to pool barcoded subsets of the genome, allowing construction of synthetic long reads. The resultant reads are of extremely high quality (> 99% accurate) but limited to approximately 18 kb in length (Koren and Phillippy, 2014). Pros for the Illumina sequencing includes highest throughput of all platforms, lowest per base cost and compatibility with almost all types of applications. The cons for the Illumina sequencing includes sample loading is technically challenging and requires expertise for accurate library construction. Another problem is the requirement for the sequence complexity where dilution/mixing with sheared PhiX is required for the low complexity samples such as amplicons to generate diversity (van Dijk, et al., 2014). Data from the amplicon libraries usually have lower yields and low quality. Mixing of low complexity samples with PhiX brings a nucleotide diversity and produce high-quality data. It also provides a quality and calibration control for cluster generation, sequencing and alignment (Illumina-Inc., 2014).

Pacific Biosciences Single Molecule Real-Time (SMRT) sequencing (Figure 1.3)
The PacBio SMRT sequencing technology utilizes a chip embedded with many "Zero-Mode Waveguide (ZMW)" well like structures. Inside each ZMW, a single active DNA polymerase enzyme, and a single stranded DNA template are immobilized to the bottom (Korlach, et al., 2010). The ZMW creates an illuminated visualization chamber that allows observations of single nucleotide incorporations by DNA polymerase (Eid, et al., 2009). Each nucleotide is attached with a specific fluorescent dye molecule that enables the detector to identify the base being incorporated by the DNA polymerase. This process is carried out in real-time as new strand of DNA is being synthesized using a template DNA strand. The PacBio technology is often called third generation sequencing platform owing to single molecule nature, which permits long read lengths, compared to the parallelized sequencing by synthesis of previous technologies. The original PacBio RS system with C1 chemistry generated mean read-lengths up to 1500 bp and yielded approximately 100 Mb of data per sequencing run (Quail, et al., 2012). Later, the RS-II platform was released with P4-C2 chemistry which improved average read lengths up to 5 kb and longest read reported in 2014 was 26 kb (Brown, et al., 2014). The most recent chemistry from PacBio is P5-C3 chemistry which provides average read lengths of approximately 8.5 kb and longest read lengths exceeding 30 kb. The P5 stands for the improved recombinant DNA polymerase enzyme proprietary to PacBio which have modified properties such as increased resistance to photodamage, reduced exonuclease activity, enhanced metal ion coordination, and reduced kinetic reaction rates. C3 stands for the various improvements to sequencing chemistry which include (i) proprietary hook molecules to facilitate isolation of polymerase-nucleic acid complex, and (ii) labelled phospholinked nucleotides instead of naturally occurring dNTPs to provide enhanced single-base identification. This P4-C5 chemistry combination achieves enhanced single molecule sequencing with increased yield, increased thermostability, increased accuracy, increased speed, and increased read-length (Kamtekar, et al., 2014; Korlach, 2014). These advances in chemistry and library preparation have boosted both the median read lengths up to 10 kb and some studies have reported longest reads beyond 50 kb (Berlin, et al., 2015; Lee, et al., 2014).

A major drawback associated with the PacBio technology is the high error rate (~ 15%), which initially made these long reads unsuitable for downstream applications such as *de novo* assembly by themselves, or for metagenomes. The optimal application required use of these longer reads in conjunction with higher accuracy short-read sequencing platforms and hybrid approaches (Koren, et al., 2013; Koren, et al., 2012). Even so, as these errors are randomly distributed over the entire length of the reads, with enough (> 100x) coverage of PacBio data, it is possible to perform self-correction and obtain longer contigs with accuracy up to 99%. The development of non-hybrid assembly approaches such as HGAP (Chin, et al., 2013) and PBcR self-correction approach (Koren, et al., 2012) have enabled utilization PacBio data without need of any accessory short sequencing reads. The latest algorithmic improvements and assembly approaches for PacBio data are summarized (Koren and Phillippy, 2014) and have enabled to obtained up to finished grade microbial genome assemblies without need for manual finishing (Brown, et al., 2014; Brown, et al., 2014). The nature of PacBio data, description of various files and data filtering procedures for downstream applications are summarized here (Utturkar, et al., 2015). Apart from the long read sequencing, the PacBio technology has the potential

to detect the DNA base modifications such as DNA methylation through direct detection of unamplified source material where kinetics of the base-addition are measured during the normal course of sequencing (Pacific-BioSciences, 2014; Roberts, et al., 2013). These kinetic measurements present characteristic patterns in response to a wide variety of base modifications. The pros of the PacBio sequencing platform include longer reads beyond 20 kb, ability to detect DNA base-modifications at no extra cost and ability to generate finished grade microbial genome assemblies. The cons of this technology include relatively high cost, high overall error rates, the lowest throughput of all platforms and these limit the range of applications.

Other sequencing platforms:
The 454, Illumina and PacBio are the primary sequencing platforms employed in current study for the *de novo* and hybrid assembly of challenging bacterial genomes and assessment of NGS platforms. Other sequencing platforms such as Ion Torrent Personal Genome Machine (Ion Torrent PGM), the Sequencing by Oligo Ligation Detection (SOLiD) and the Oxford Nanopore MinION (Nanopore) sequencing platforms are available and described in brief.

The Ion Torrent PGM
The Ion Torrent PGM uses a semiconductor based technology and does not rely on the optical detection of incorporated nucleotides with fluorescence and camera scanning (Koren and Phillippy, 2014). In brief, DNA fragments with specific adapter sequences are linked to and then clonally amplified by emulsion PCR on 3-micron diameter bead surface (Ion Sphere Particles). These beads are loaded into proton sensing wells that are fabricated on a silicon wafer and the sequencing reaction is primed from a specific location in adapter sequence. The bases are introduced sequentially, and if incorporated protons are released and detected signals are proportional to number of bases (Rothberg, et al., 2011). This results in higher speed, lower cost and smaller instrument size. The first PGM generated about 270 Mb of sequence data with read size up to 100 bp (van Dijk, et al., 2014). The latest upgrades to library preparation and sequencing chemistry have improved the sequencing output up to 2 GB per run with maximum and median read lengths up to 400 bp and 200 bp, respectively (https://www.thermofisher.com/order/catalog/product/4462921). The pros of the Ion Torrent platform includes semi-conductor based technology with no requirement for optical scanning and fluorescent nucleotides, fast run times and broad range of applications. The cons include high error rates, especially with homopolymer repeats, similar to those discussed above 454 platform. Unlike 454, where base-call accuracy decreases with length of homopolymer, the PGM tend to introduce homoplymeric errors following the “A” or “T” nucleotide flow-cycle which tend to produce over-calling or under-calling signals for the length of homopolymer region (Bragg, et al., 2013).

The SOLiD sequencing
The SOLiD sequencing is sequencing by ligation technology developed by Applied Biosystems (now Thermo Fisher Scientific). In brief, a sequencing primer is hybridized to adapter and its 5’ end is available for ligation to an oligonucleotide hybridizing to the adjacent sequence. A mixture of octamers compete for the ligation to the primer (with

bases 4 and 5 in these oligos are encoded by color labels). After color detection, the ligated octamer is cleaved between position 5 and 6, which removes the label and cycle is repeated. The first round detects the possible bases in position 4, 5, 9, 10, 14, 15 etc. The process is repeated with offset of one base using a shorter oligonucleotide primer which determines position 3, 4, 8, 9, 13, 14 etc. The cycle is repeated until the first base in the sequencing primer (position 0) is reached. This methods allows detection of each base twice and thereby features added accuracy (Metzker, 2010). The SOLiD system output varies from 8 GB to 24 GB per day based on instrument platform, however read lengths are comparatively smaller up to 75 bp. The pros of the SOLiD system includes second highest throughput system in market after Illumina, it is widely claimed to have the lowest error rates with 99.94% accuracy owing to double detection of each base. The cons include shortest read lengths of all platforms that are less well-suited for *de novo* genome assembly, and relatively long run times and (van Dijk, et al., 2014).

The Oxford Nanopore MinION

The Nanopore sequencing is most recently released third generation sequencing platform after PacBio. The MinION is a thumb drive size device which can be connected to a laptop, and measures deviations in electrical current as a single strand DNA is passed through a protein Nanopore (Schneider and Dekker, 2012). The technology is based on an array of nanopores embedded on a chip that detects consecutive 5-mers of a single stranded DNA molecule by electrical sensing (Cherf, et al., 2012). The library preparation is similar to other NGS platforms and requires DNA shearing, end repair, adaptor ligation and size selection. Finally, DNA is conditioned by the addition of a motor protein, libraries are mixed with buffer and a proprietary ‘fuel-mix’ and loaded directly into the sequencer (Mikheyev and Tin, 2014). As the sequencing is progress, base-calling takes place in real-time. During sequencing, two strands of the DNA molecule are linked by a hairpin and sequenced consecutively, and when two strands of the molecule are read successfully, a consensus is built to obtain more accurate (2D read) or called (1D read) when only forward strand is read (Madoui, et al., 2015). The size, robustness and affordability of the MinION make it a unique technology. The MinION sequencer is available as a small portable device which can be connected to a laptop and generate real-time results as sequencing in progress. Recent studies have reported the average read size from MinION as 5-5.5 kb (Ashton, et al., 2015; Quick, et al., 2015). However, as with other single molecule platforms MinION also suffers with low accuracy which is reported to be ~70% with R7 chemistry and ~80% for R7 2D sequences (Quick, et al., 2015). Despite of these higher error rates, the potential of Nanopore reads for microbial sequencing has been demonstrated by recent studies (Judge, et al., 2015; Madoui, et al., 2015). Previous computational methods are available which were developed for the long-reads generated by PacBio technology should in theory apply. However, updates to these tools would be necessary to handle the specific characteristics of the Nanopore data. To summarize, the size and portability are big plus for the Nanopore platform but technical improvements are necessary to handle the higher error rates for the best utilization of the Nanopore data.

Comparison of sequencing platforms:

Comparison of various parameters from different NGS platform is presented in Figure 1.4. Figure 1.4A displays a comparison of maximum read lengths obtained through 454,

Illumina, SOLiD, Ion Torrent and PacBio sequencing platforms. Orange bars indicate the maximum read lengths that can be obtained with these technologies today. Dark orange stands for the large instruments output from each platform, while light orange indicates output from bench-top versions. Illumina now produces reads of several hundreds of bases, while exceptionally long reads can be produced by the new PacBio RS-II platform. Figure 1.4B shows the maximum throughput of first commercially available sequencing instruments (blue bars), and current maximum throughput (dark orange bars). It should be noted that highest throughput and longest reads may not be obtained from the same instrument e.g. Illumina MiSeq generates the longest read lengths across all Illumina platforms while HiSeq X Ten generates maximum throughput. Figure 1.4C shows the evolution of the cost of sequencing a human genome from 2001 until today. There has been a dramatic decrease in sequencing costs owing to recent technological advances and computational improvements. Figure 1.4D shows comparison of time to complete the typical bacterial genome sequencing run. Figure 1.3D was created considering 400 bp run for 454 instrument, 150 bp paired-end run on Illumina MiSeq, 50 bp run for SOLiD's 5500W Series Genetic Analyzer and 200 bp run for Ion Torrent PGM. The PacBio run time rather than length was set to a 3 hours for bacterial genome sequencing. A more comprehensive comparison of these platforms in terms of purchase pricing, costs per run, and error rates is available (Liu, et al., 2012; Loman, et al., 2012; Quail, et al., 2012).

De novo and hybrid genome assembly methods and challenges

With rapidly falling costs genome sequencing is now a routine task even for a small-scale laboratories. The developments in NGS technologies have changed the course of biological studies in recent years (Mavromatis, et al., 2012). Increasingly, investigators have turned to rapid whole genome sequencing to trace the source of infectious disease outbreaks, to understand the source of pathogenesis, and to understand multidrug resistance among other questions (Illumina-Inc., 2015; Magoc, et al., 2013). Assembly of the DNA reads to correctly reconstruct genomes is an essential task to facilitate genomic studies. The process of sequence assembly dates back to 1980s when pioneering work of Esko Ukkonen revealed the fundamental difficulty of reconstructing a genome from sequenced fragments (Peltola, et al., 1984). Genome assembly complexity can range from trivial (when all repeats are shorter than read-lengths) to computationally intensive (requires trying an exponential number of arrangements of reads) to impossible (information contained within reads is insufficient to identify correct sequence reconstruction and continuity (especially in case of large eukaryotic genomes) (Nagarajan and Pop, 2013). The genome assemblers are based on one of the several different paradigms such as greedy, Overlap-Layout-Consensus (OLC), de Bruijn graph, and string graph (Nagarajan and Pop, 2013). The choice of approach depends upon the characteristics of the data being assembled. Most of the modern genome assemblers are based on the de Bruijn-graph based methods (successfully applied for short reads generated by most second generation sequencing platforms such as Illumina, Ion Torrent, and SOLiD) and OLC approaches (mostly used for the longer reads generated by 454, Sanger and third-generation sequencing platforms). A variety of assembly algorithms and quality evaluation methods for the *de novo* and hybrid assembly of various NGS data are available (Magoc, et al., 2013; Salzberg, et al., 2012; Utturkar, et al., 2014).

The de bruijn graph assemblers model the relationship between exact substrings (k-mer) of the length *k* derived from the input reads. The nodes in the graph represent the k-mers while the edges indicate that the adjacent k-mers overlap by exactly (k-1) letters (Compeau, et al., 2011). The reads are directly not modeled in this paradigm, but they indirectly represented as the paths through the de bruijn graphs. During the assembly process, the de bruijn graph structure is continuously refined with the new read information and graph patterns that are not consistent with reads are removed (Compeau, et al., 2011). The de bruijn graph building approach is based on exact matches of k-mer words and read accuracy plays an important role in building accurate graph structures. Therefore, use of error correction approaches is a crucial step for building accurate de bruijn graph based assemblies. This requirement prevents the applicability of de bruijn graph-based methods to longer reads, which tend to have high error rate and inaccurate base-calls. The OLC based assemblers start with identifying the read pairs that overlap sufficiently (as indicated by the custom cutoff value provided by the user) and then organizes this information into a graph. Every node of the graph represents the read and an edge represents an overlap between a read pair (Li, et al., 2012). The OLC graph structure takes into account the global relationships between the reads. The OLC paradigm was applied in the early assemblers such as Celera assembler designed to handle long reads generated by Sanger platform and dominated the assembly era until the emergence of short-read sequencing technologies. The use of OLC based assemblers is increased again with the emergence of long read producing third generation sequencing platforms such as Pacbio.

One of the first widely used short-read assembler was Velvet which made a mark by showing high-quality assemblies could be obtained by using ultra short-reads (~ 30 bp) and high coverage datasets (>100x) for small bacterial genomes (Koren and Phillippy, 2014; Zerbino and Birney, 2008). This approach was further extended for the assembly of large genomes by the ABySS software and for the first *de novo* assembly of mammalian genome entirely using short reads with the program SOAPdenovo. The SOAPdenovo was designed to have memory usage efficiency and also included robust error correction module and scaffolding modules. The concept of integration of two data types was popularized by the Euler (Pevzner, et al., 2001) assembler which used mate-pair reads along with standard paired-reads to efficiently resolve the repeat structures (Pevzner and Tang, 2001). This concept was further extended by modern assembler ALLPATHS-LG which proposed the approach of increasing read-lengths wherein mate-pair libraries are constructed such that they overlap with paired-end reads. This specific library preparation allowed overlapping mates to be stitched together into reads that are roughly twice the size produced by sequencing instruments. This approach is able to better resolve the repeat structures and thereby increases the size and accuracy of assembled contigs. ALLPATHS-LG represents the best joint design which increased interaction between experimental design and assembly approach, and arguably won the Assemblathon (Earl, et al., 2011) and GAGE (Salzberg, et al., 2012) competitions and generated the best results in another assembler evaluation study (Utturkar, et al., 2014). Another latest addition to assembly toolbox is the SPAdes assembly package (Bankevich, et al., 2012), which is designed for both standard isolate and single-cell genome assemblies. Based upon experience described later in this thesis, SPAdes is

recommended as the starting assembler for isolate or single-cell genome assemblies. More specifically, the SPAdes is constantly upgraded according to the newest sequencing platforms and chemistries, it is compatible with multiple platforms, integrates automatic read error correction and mismatch correction tool to reduce the rate of mismatch and short indel rates in final contigs. SPAdes also offers multilevel user control paradigm which allows the user to turn-on or turn-off specific features of the assembly pipeline as per the requirement which may be rewarding in terms of memory usage and computation time.

The assembly of long reads generated by third-generation technologies was challenging because associated high error rates interfere with the assembly process. Initially, these longer reads were insufficient to achieve high-quality genome assembly by themselves. The available solution was to perform correction of PacBio reads using more accurate second generation sequencing data followed by the assembly of corrected reads. The first PacBio read error correction program was PacBio Corrected Reads (PBcR) pipeline, which utilizes high-quality Illumina reads for error correction and achieves long PacBio reads with greater than 99.9 % base-call accuracy (Koren, et al., 2012). The error corrected PacBio reads are then assembled through Celera assembler as part of the default PBcR pipeline. However, the user can choose any other assembler as per the preference. Another approach was to utilize these longer reads to perform hybrid assemblies in which initial assembly was performed using high-quality second generation sequence data and longer reads are used for the scaffolding and gap filling (Bashir, et al., 2012; English, et al., 2012). In 2013, PacBio released their native assembly program called HGAP which was able to achieve self-correction of PacBio reads with >100x sequence coverage and uses hierarchical assembly process to obtain up to finished quality microbial genome assemblies (Chin, et al., 2013). The improved throughput from the PacBio RS-II platform and the random nature of its sequencing errors are important factors in the success of HGAP assembler. The HGAP protocol is integrated with a single round of quiver polishing which uses the raw PacBio data, underlying quality values and hidden Markov model-based probabilities of the basecall quality and generate an accurate consensus sequence. Later the PBcR pipeline was also updated to perform the self-correction of PacBio reads when greater than 50x sequence coverage is available. Owing to the utility of the long reads, most of the assembly programs including SPAdes and ALLPATHS-LG have provision to integrate the PacBio or Nanopore sequencing data to generate hybrid assemblies. Pilon is another program available for assembly polishing which utilizes high-quality Illumina reads to correct the assembly and obtain improved consensus sequence (Walker, et al., 2014).

Despite advances in assembly methods, the process of genome assembly is still a challenging task and assembly quality varies from sample to sample. Repetitive stretches of DNA are abundant in many bacterial genomes and pose one of the greatest technical challenges to *de novo* genome assembly (Treangen and Salzberg, 2012). In the case of bacteria, the rDNA gene operons are often the largest region of repetitive sequence and range in size between 5 and 7 kb (Treangen, et al., 2009). The challenge to the *de novo* assembly process is most exacerbated when repeat sequence regions are longer than the read lengths. The short read sequences generated by many second-generation

sequencing technologies such as Illumina often yield highly fragmented genome assemblies and achieve only high-quality draft status (Chain, et al., 2009). The relative values of the finished genome (Fraser, et al., 2002), technical challenges (Hurt, et al., 2012; Treangen and Salzberg, 2012) and what is missing from the finished versus draft genomes (Land, et al., 2014; Mavromatis, et al., 2012) have been summarized in multiple publications. Apart from the biological challenges associated with the process of genome assembly, certain engineering challenges determine the success of the modern assembly software. Modern assemblers should be able to handle and analyze the large datasets efficiently which requires efficient memory utilization and use of compressed graph structures (Nagarajan and Pop, 2013). Most of these open source assembly software are designed for the Linux based systems and upgrade to newer versions of the assembly software are highly recommended for optimal results.

Other applications of NGS technologies:

NGS sequencing technologies have revolutionized the field of genomics with applications in every area of research and industry. Initial success for sequencing microbial genomes, followed by sequencing the genomes, have been followed by those of more of complex eukaryotic organisms, and resequencing approaches to understand population genetic variation (e.g. the 1000 human genomes project). Other techniques based on NGS technologies include RNA-sequencing, single-cell genomics, metagenomics and Chip-Seq. These techniques have applications from industrial biotechnology, cancer genomics to personalized medicines. A short description for RNA sequencing and single-cell genomics applications is provided below:

RNA-Sequencing (RNA-seq)

RNA-seq is an approach to profile a complete set of transcripts in population of cells or tissues, and quantify their abundance in specific developmental stage or physiological condition using NGS technologies. Before the RNA-seq approach, initial transcriptomics studies were largely dependent on northern blots, quantitative PCR based methods, or hybridization-based microarray technologies which offered the ability to quantify the transcriptomes of diverse organisms. High-throughput NGS technologies have revolutionized the field of transcriptomics by adding the capability of RNA analysis through cDNA sequencing at massive scales individually or in parallel. This approach helped to eliminate several challenges associated with microarray technology including limited dynamic range of detection (Ozsolak and Milos, 2011; Wang, et al., 2009) or dependence on prior knowledge of a genome sequence. RNA-seq studies provide a progressively complete knowledge of quantitative and qualitative aspects of transcript dynamics and gene regulation. The recent developments in RNA-seq methods and computational approaches allow wide range of applications including transcription start site mapping, strand-specific measurements, gene fusion detection, small RNA characterization and detection of alternative splicing events (Ozsolak and Milos, 2011). A more recent direct RNA sequencing approach allows RNA quantification from very small amounts cellular material. Emergence of third generation sequencing technologies with longer read-lengths has enabled sequencing of individual full-length cDNA molecules representing entire transcripts. Paired-end sequencing approaches have enabled sequence information to be obtained from two points in a transcript with estimated

distance between reads. Longer reads allow better mapping of the reads to alternatively spliced junctions in eukaryotes, while paired-ends reads support better transcriptome assembly. On the contrary, some of the limitations of RNA-seq approach includes non-uniformity of coverage across transcripts, transcript-length bias because of multiple fragmentation or RNA/cDNA size selection during library preparation and read-mapping uncertainty (owing to sequence error rates, repetitive elements, incomplete genome sequences and inaccurate transcript annotations).

From bioinformatics perspective, some of the challenges associated with RNA-seq include accurate mapping of the reads to reference, or transcriptome assembly when reference is not available. There are several customized programs available for various steps in RNA-seq analysis and some commonly employed tools include Trinity for transcriptome assembly, TopHat for read mapping, HTSeq for read counting, and DeSeq for differential gene expression analysis (Wang, et al., 2009). These tools have specific flags or parameters which enable optimize the process of read mapping across splice junctions and identification of novel isoforms. Sequencing technologies and bioinformatics approaches are constantly advancing and promise to alleviate difficulties in RNA-seq analysis.

Single-cell genomics (SCG)

Single-cell genomics is a method to obtain sequence information from individual cells with optimized NGS technologies, to obtain higher resolution and better understanding of cellular function in the context of its microenvironment or understand the basic potential functions of uncultured organisms. The SCG relies on flow cytometry based Fluorescence-Activated Cell Sorting (FACS) or other methods to isolate single-cells (Kalisky and Quake, 2011). There are other methods available for cell sorting such as microfluidic fluorescence-activated droplet sorting (FADS) which combines the advantages of microtitre-plate screening and traditional FACS (Baret, et al., 2009), bead-based cell sorting, mechanical or optical micromanipulation (Ishii, et al., 2010) etc. Other devices include microfluidic based cell-sorters (Lecault, et al., 2012) which offer high-throughput and sensitive detection methods with efficient sorting (Shields, et al., 2015). Briefly, individual cells in liquid medium are labelled with fluorescent antibodies to specific membrane protein and passed through a path of multiple laser beams of different wavelengths. Optical detectors convert fluorescent light emitted from each cell into an electrical signal and based on the intensity of signals cell sorting is performed. A label free cell sorting is also available in which cells can be sorted based various physical properties such as size, granularity and optical properties (Blainey, 2013; Lasken, 2012). Only a few years ago, application of NGS technologies to SCG was limited because only a few femtograms of DNA content of single cells were insufficient for sequencing without PCR, which because of its gene-by-gene nature was impractical. This obstacle was overcome by the whole genome amplification method called Multiple-Displacement amplification (MDA), which generates micrograms of genomic template from single cell DNA in a linear reaction (Huang, et al., 2015). Some of the technical challenges associated with single-cell genomics include incomplete representation of the genome, uneven sequence coverage and contaminating sequences arising from host, reagents, human handling or cross-contamination which deteriorates the assembly quality. One of

the important applications of SCG involves studying uncultured majority of bacteria. Only a fraction of all microorganisms has been identified, and an even smaller fraction is grown in culture. The SCG approach provides a cultivation-independent method for obtaining genome sequences of microbes from uncultured candidate phyla or groups (Rinke, et al., 2013). Detail discussion of these challenges, bioinformatics solutions and methods, and application of SCG to study uncultured endophytic bacteria from *Populus deltoides* tree are discussed in chapter 6.

1.2 Statement of hypothesis

This dissertation research address two important areas associated with next-generation sequencing applications. First, I investigated variety of *de novo* and hybrid genome assembly methodologies for novel bacterial genomes without reference sequences and proposed rDNA operon evaluation approaches to validate the assembly accuracy. Second, I utilized third generation PacBio sequence data from *Clostridium autoethanogenum* strain JA1-1 (DSM 10061) to obtain a finished grade microbial genome assembly using RS-II data alone and without the need for manual finishing. We were one of the first to publish a complete microbial genome sequence using only the PacBio data. This also allowed us to compare second generation (Illumina/454) and third generation (PacBio) sequencing platforms to reveal the advantages and limitations associated with each platform and create a reference dataset useful for benchmarking new computational tools. This research addresses following specific hypothesis:

Hypothesis 1: *The combination of complementary libraries, sequencing technologies and optimization with hybrid assembly protocols can obtain dramatic improvements in assembly quality for bacterial genomes and these will vary from genome to genome.*

Previous research has developed a variety of *de novo* and hybrid assembly algorithms (Magoc, et al., 2013; van Dijk, et al., 2014) and various *in silico* evaluation matrices (Hunt, et al., 2013; Rahman and Pachter, 2013). However, there is no one best method for any single genome. It is important to determine the most appropriate NGS technology combinations, assembly protocols and parameter optimization to obtain an optimal genome assembly and to develop multiple specific evaluation criteria to validate the assembly accuracy. These comparisons provide a reliable and robust framework for others looking to improve existing draft genome sequences.

Hypothesis 2: *Longer read lengths generated by the PacBio platform are capable of generating high quality finished grade microbial genome sequences without need for manual finishing and will facilitate the comparative and functional genomics studies.*

Longer reads generated by the third generation sequencing (single molecule) technologies such as PacBio are useful to assemble majority of bacterial genomes in up to finished-grade quality despite associated higher error rates (Koren, et al., 2012). Various bioinformatic methods have been developed for the error correction of PacBio reads. The most efficient one described is recently developed HGAP protocol which could perform self-correction of PacBio reads and excludes the need for accessory sequencing data (Chin, et al., 2013). So, here I used only the PacBio sequence data, performed self-

correction and assembly of PacBio reads with HGAP protocol, and compared these data with Illumina/454 assemblies and reads to assess the potential of this new sequencing platform and to determine its advantages and limitations over second generation sequencing by synthesis sequencing technologies.

Hypothesis 3: *The breakpoints or gaps associated with Illumina technology are mostly associated with large repeats such as rDNA operons while gaps within PacBio assembly correspond to DNA regions that generate strong secondary structures or DNA hairpin-loops.*

Due to increased read-lengths of over 1 or 2 orders of magnitude, algorithm improvements and hybrid assembly approaches, the concept of one chromosome, one contig and automated finishing of microbial genomes is now a realistic and achievable task for many microbial genomes (Koren and Phillippy, 2014). The PacBio platform was predicted to be able to obtain finished genome assemblies for majority of bacterial genomes (Koren, et al., 2013) and this speculation is supported by increased number of finished genomes obtained using this technology (Eckweiler, et al., 2014; Harhay, et al., 2014; Kanda, et al., 2015; Mehnaz, et al., 2014; Nakano, et al., 2015; Satou, et al., 2014). However, at the same time there are some examples where microbial genomes could only be resolved into less than 10 contigs despite robust (>100x) PacBio sequence coverage and time-consuming manual finishing was necessary to obtain complete genomes (Bishnoi, et al., 2015; Dunitz, et al., 2014; Hoefler, et al., 2013; Okutani, et al., 2015; Shapiro, et al., 2015). There are few examples available where PacBio assemblies were compared with Illumina/454 assemblies and revealed rDNA operons as major breakpoints in short-read assemblies (Brown, et al., 2014). However, more examples of draft and finished genomes would be useful to confirm the nature of gaps within short-read assemblies. Therefore, eight microbial genomes were sequenced using Illumina Paired-End (PE) and PacBio RS-II platforms and I performed a comparison of draft and finished genome assemblies with the aim to elucidate the nature of gaps associated with Illumina and PacBio technology.

Hypothesis 4: *A sequencing dataset which span three generations of sequencing technologies, containing six types of data from four NGS platforms and originating from a single microorganism will facilitate algorithm developments that maximize the quality of past and future DNA sequencing efforts.*

The advancements in NGS technologies have led to the emergence of novel sequencing platforms such as 454, Illumina, SOLiD, Ion Torrent, PacBio with each having their own advantages and limitations for *de novo* genome sequencing (van Dijk, et al., 2014). Apart from these, there are several new sequencing platforms such as Nanopore which are in incipient stage or in the research and development pipeline. After release, these new platforms will require an assessment in terms of data quality, read-lengths and tool development for efficient utilization of data generated alone or in combination with other platforms. Here we describe a dataset that represents three generations of sequencing technologies, and contains six types of data from four NGS platforms; 454 GS FLX, Illumina MiSeq, Ion Torrent, and PacBio RS-II; and Sanger sequence data. The details

for the sequencing and library preparation protocols, data generation methods and descriptions for various data files are provided in a way to facilitate the broadest possible community and developer access.

Hypothesis 5: *Single-cell genomics analysis, which includes de novo single-cell genome assembly, binning and contamination removal, validation and genome completeness estimation and comparative genomics will help to identify the unique putative characteristics of uncultured endophytic bacteria isolated from root of the tree* Populus deltoides *via modified enrichment protocol.*

Endophytic bacteria that colonize the *Populus* trees are known to contribute towards nutrient acquisition and increases in both above and below ground biomass. Endophytic bacteria are embedded inside the plant roots and physical separation of live endophytes from plant roots is a challenging task. An enrichment protocol based on differential and density gradient centrifugation was developed to separate and isolate the live endophytic bacteria from plant roots. Further *in silico* characterization and validation of the previously uncultured endophytic bacteria was performed using phylogenetic and comparative genomics approaches.

1.3 Approach

Experiments to test above hypotheses can be categorized into three major parts. First, I performed optimization for nine *de novo* and hybrid assembly protocols to obtain improved genome assemblies for four bacterial genomes without reference sequences (Chapter 2). Second, I was able to obtain a complete genome sequence for *Clostridium autoethanogenum* using only the PacBio data and HGAP protocol without need for manual finishing. A comparison of PacBio and Illumina/454 assemblies was then performed to reveal the nature of gaps associated with Illumina technology (Chapter 3), followed by deposition of genomic data for *C. autoethanogenum* to public repositories which spans three generations of sequencing technologies, containing six types of data from four NGS platforms (Chapter 4), and further evaluation of gaps (unassembled regions) associated with Illumina and PacBio assemblies was performed to reveal the nature of the sequence gaps (Chapter 5). Finally, single-cell genomics analysis of uncultured endophytic bacteria was performed which includes *de novo* assembly, binning and contamination removal, genome completeness estimation and comparative genomics to identify the unique putative characteristics of each single-cell (chapter 6).

Chapter 2: Evaluation and validation of *de novo* and hybrid assembly techniques to derive high quality genome sequences.

There are a variety of *de novo* and hybrid assembly approaches available and benchmarking was performed using available finished genome sequences. The *in silico* assembly evaluation tools were limited to rank the assemblies rather than selection of best assembly. Therefore, my first step was evaluation of these assembly algorithms using combinations of sequencing technologies and complementary libraries generated for novel bacterial genomes without any reference sequence. Additionally, each assembly was validated using a PCR and Sanger sequencing approach to confirm the presence of

predicted rDNA operons and provides an additional evaluation criterion to validate the assembly accuracy. This chapter provides detailed comparisons and optimization approaches that I developed for various assembly protocols, and describes the PCR and Sanger sequencing approach to validate the rDNA operons and assembly accuracy.

Chapter 3: Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of *Clostridium autoethanogenum* and analysis of CRISPR systems in industrial relevant Clostridia.

The new generation of assembly methods were tested for their ability to utilize the longer reads generated by PacBio technology and to obtain high quality genome assemblies despite of high error rate associated with this technology. The genome of *C. autoethanogenum* was assessed as a complex bacterial genome based on genome features such as repeats, prophage, and nine copies of the rRNA gene operons. Here, we used only the PacBio data and HGAP protocol to obtain the complete genome sequence for *C. autoethanogenum* without need for manual finishing. Later we set out a comparison between draft and finished genome assemblies summary statistics, CGAL, QUAST and REAPR bioinformatics tools. Comparative genomic approaches were applied to against a close relative, *C. ljungdahlii,* to identify distinct features of the genome and relate them to known organismal differences from previously published physiological experiments.

Chapter 4: Sequence data for *Clostridium autoethanogenum* using three generations of sequencing technologies.

The whole genome sequencing for *C. autoethanogenum* was performed using 454 GS FLX, Illumina MiSeq, Ion Torrent, PacBio RS-II platforms and Sanger sequencing platforms. This includes two libraries on 454 (a sheared shotgun (average length 289 bp) and 3 kb paired end library), a paired-end library for Illumina, a single-end library for Ion Torrent, two SMRT cells on PacBio RS-II instrument (P4-C2 chemistry) and Sanger (ABI 3730) sequences for amplified PCR products. The data validation was performed by determining the basic quality statistics for the raw sequence data which includes sequence lengths distributions, GC-content, Ambiguous base-content, PHRED quality score distribution, nucleotide contributions (%A, %T, %G, %C), kmer distribution analysis and sequence read duplication levels. Secondly, to ensure sequences are correctly matching, mapping was performed against the model organism *C. autoethanogenum* DSM 10061 and its close relative *C. ljungdahlii* DSM 13528 (average nucleotide identity score over 99%) to avoid any bias. Mapping rates for each platform against each reference genomes were determined to illustrate the quality and usefulness of the datasets.

Chapter 5: Evaluation of unassembled DNA regions for Illumina and PacBio NGS platforms and microbial genome finishing.

Eight microbial genomes were sequenced using Illumina Paired-End (PE) and PacBio RS-II platforms. *De novo* and hybrid assemblies were performed using various assembly

algorithms followed by the manual finishing for two unfinished genomes. A comparison of draft and finished genome assemblies was performed to reveal the nature of gaps associated with Illumina sequencing. The gaps associated with PacBio sequencing were derived from a PCR and Sanger sequencing approach and investigated for specific properties such as genome coverage, annotations, and read quality.

Chapter 6: Enrichment of live bacterial endophytes from *Populus deltoides* for single-cell genomics.

Fine root samples from three one-year-old *Populus deltoides* saplings were harvested, rhizosphere and the rhizoplane soil and microorganisms of the roots were removed via rinsing and sonication, and roots were homogenized. Enrichment of endophytic microbial communities was performed using repeated differential and density gradient centrifugation. Total DNA extraction was performed from enriched and unenriched samples, and bacterial community composition was determined using 16S rDNA sequencing. Enriched bacterial samples were then subjected to flow cytometry, MDA amplification, 16S rRNA gene screening, and finally whole genome shotgun sequencing of selected cells was performed on 12 SAGs representing novel phylogenetic individuals and groups. Sequences were assembled using a variety of approaches, and the assemble contigs screened for potential contaminant sequences, and characterization of each SAG was performed using phylogenetic and comparative genomics approaches.

1.4 Significance

Recent advances in NGS technologies have enabled rapid and high-throughput sequencing at very low cost and microbial genome sequencing has become a routine technique used to study genomics aspects of bacteria, archaea and even microbial eukaryotes (Koren and Phillippy, 2014). Rigorous QC and generation of accurate and optimal genome assemblies of contigs and scaffolds is the first required step for subsequent genome analysis, thereby affecting the downstream accuracy and usefulness of all subsequent analyses. Several *de novo* and hybrid assembly algorithms and *in silico* assembly validation methods are available and often each claims specific advantages over others (Koren, et al., 2014; Magoc, et al., 2013). The selection of appropriate assembly program and validation of assembly accuracy remains a challenge for novices and experts alike. A systematic evaluation of nine leading *de novo* and hybrid assembly protocols presented in chapter 2 allowed me to select the optimal assembly algorithm based on available NGS data types and provided a rRNA operon validation method to select the most accurate assembly. Both assessments will aid others looking to improve exiting draft genome assemblies. This is evidenced by the fact that the paper describing this work has already been cited 20 times since publication approximately one year ago. Although this study used preliminary third generation sequencing data from the PacBio RS-I, the main focus was to generate optimal hybrid assemblies in combination with second generation sequencing platforms such as Illumina and 454. In later years, emergence of PacBio RS-II platform and improved sequencing chemistry (see details in introduction chapter) have enabled generation of substantially longer reads and improved algorithms to perform long-read assembly (Chin, et al., 2013). In chapter 3, I have presented an example of automated finishing of microbial genome using only the PacBio

RS-II data and assembly comparison to reveal the nature of breakpoints in Illumina/454 assemblies. This study asserts the advantages of PacBio sequencing technology for this application and demonstrates its superiority for automated genome finishing of even complex genomes. The first complete genome sequence for *Clostridium autoethanogenum* will facilitate the comparative and functional genomics analysis and strain improvement of this industrially relevant bacteria. Again this is evidence by 19 citations since this paper was published approximately 1 year ago. The fourth chapter describes the sequence dataset for *C. autoethanogenum* which span three generations of sequencing technologies, containing six types of data from four NGS platforms. This dataset will be useful for the scientific community to evaluate upcoming NGS platforms, enabling comparison of existing and novel bioinformatics approaches and will encourage interest in the development of innovative experimental and computational methods for NGS data. This dataset was published in April and has already been cited three times. In chapter five, we have presented both *in silico* (bioinformatics based) as well as laboratory methods for manual genome finishing of high complexity bacterial genomes which could not be finished using PacBio sequencing. We also evaluated intractable (unassembled) DNA regions from Illumina and PacBio technologies and revealed associated properties such as annotations, read-quality and read-depths. This genome finishing protocol can obtain substantial assembly quality improvements for genomes which have remained unfinished by PacBio technology and offers insights for sequencing companies and algorithm developers to make specific improvements. In chapter six, we have described a protocol for enrichment of endophytic bacteria from tree *Populus deltoides* and characterization of uncultured isolated using single-cell genomics. This protocol allowed enrichment of live endophytic bacteria away from the plant material and enabled single-cell and metagenomics analysis on natural root samples by greatly reducing the amount of contaminating plant and microbial DNA. This new protocol could be applied for the study of uncultured bacteria from different host-associated environments and shed light on genetic and symbiotic features. Chapters five and six of this dissertation are still being refined for publication as of this writing.

References

Abramovitch, R.B., Anderson, J.C. and Martin, G.B. (2006) Bacterial elicitation and evasion of plant innate immunity, *Nat. Rev. Mol. Cell Biol.*, **7**, 601-611.

Ashton, P.M.*, et al.* (2015) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island, *Nat. Biotechnol.*, **33**, 296-300.

Bankevich, A.*, et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J. Comput. Biol.*, **19**, 455-477.

Baret, J.C.*, et al.* (2009) Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity, *Lab Chip*, **9**, 1850-1858.

Bashir, A.*, et al.* (2012) A hybrid approach for the automated finishing of bacterial genomes, *Nat. Biotechnol.*, **30**, 701-707.

Berendsen, R.L., Pieterse, C.M. and Bakker, P.A. (2012) The rhizosphere microbiome and plant health, *Trends Plant Sci.*, **17**, 478-486.

Berlin, K.*, et al.* (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing, *Nat. Biotechnol.*

Bishnoi, U.*, et al.* (2015) Draft Genome Sequence of a Natural Root Isolate, Bacillus subtilis UD1022, a Potential Plant Growth-Promoting Biocontrol Agent, *Genome Announc*, **3**.

Blainey, P.C. (2013) The future is now: single-cell genomics of bacteria and archaea, *FEMS Microbiol. Rev.*, **37**, 407-427.

Bragg, L.M., et al. (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data, PLoS Comput Biol, 9, e1003031.

Brown, S.*, et al.* (2014) Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of *Clostridium autoethanogenum* and analysis of CRISPR systems in industrial relevant Clostridia, *Biotechnol. Biofuels*, **7**, 40.

Brown, S.D.*, et al.* (2014) Complete genome sequence of *Pelosinus* sp. strain UFO1 assembled using Single-Molecule Real-Time DNA sequencing technology, *Genome Announc*, **2**.

Chain, P.S.*, et al.* (2009) Genomics. Genome project standards in a new era of sequencing, *Science*, **326**, 236-237.

Cherf, G.M.*, et al.* (2012) Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision, *Nat. Biotechnol.*, **30**, 344-348.

Chin, C.S.*, et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, *Nat. Methods*, **10**, 563-569.

Dunitz, M.I.*, et al.* (2014) Draft Genome Sequences of Escherichia coli Strains Isolated from Septic Patients, *Genome Announc*, **2**.

Eckweiler, D.*, et al.* (2014) Complete genome sequence of highly adherent *Pseudomonas aeruginosa* small-colony variant SCV20265, *Genome Announc*, **2**.

Earl, D.*, et al.* (2011) Assemblathon 1: a competitive assessment of *de novo* short read assembly methods, *Genome Res.*, **21**, 2224-2241.

Eid, J.*, et al.* (2009) Real-time DNA sequencing from single polymerase molecules, *Science*, **323**, 133-138.

English, A.C.*, et al.* (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology, *PLoS One*, **7**, e47768.

Fraser, C.M.*, et al.* (2002) The value of complete microbial genome sequencing (you get what you pay for), *J. Bacteriol.*, **184**, 6403-6405.

Gharizadeh, B., et al. (2002) Long-read pyrosequencing using pure 2'-deoxyadenosine-5'-O'-(1-thiotriphosphate) Sp-isomer, Anal. Biochem., 301, 82-90.

Harhay, G.P.*, et al.* (2014) Complete closed genome sequences of three *Bibersteinia trehalosi* nasopharyngeal isolates from cattle with shipping fever, *Genome Announc*, **2**.

Hoefler, B.C., Konganti, K. and Straight, P.D. (2013) De novo Assembly of the Streptomyces sp. Strain Mg1 Genome Using PacBio Single-Molecule Sequencing, *Genome Announc*, **1**.

Huang, L.*, et al.* (2015) Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications, *Annu Rev Genomics Hum Genet*, **16**, 79-102.

Hunt, M.*, et al.* (2013) REAPR: a universal tool for genome assembly evaluation, *Genome Biol.*, **14**, R47.

Hurt, R.A.*, et al.* (2012) Sequencing intractable DNA to close microbial genomes, *PLoS One*, **7**, 7.

Illumina-Inc. (2015) Next generation sequencing aids researchers in the fight against the ebola virus.

Illumina-Inc. (2014) Using a PhiX Control for HiSeq® Sequencing Runs.

International Human Genome Sequencing, C. (2004) Finishing the euchromatic sequence of the human genome, *Nature*, **431**, 931-945.

Ishii, S., Tago, K. and Senoo, K. (2010) Single-cell analysis and isolation for microbiology and biotechnology: methods and applications, *Appl. Microbiol. Biotechnol.*, **86**, 1281-1292.

Judge, K.*, et al.* (2015) Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes, *J. Antimicrob. Chemother.*, **70**, 2775-2778.

Kalisky, T. and Quake, S.R. (2011) Single-cell genomics, *Nat. Methods*, **8**, 311-314.

Kamtekar, S., et al. (2014) Recombinant polymerases with increased phototolerance. Google Patents.

Kanda, K., Nakashima, K. and Nagano, Y. (2015) Complete Genome Sequence of Bacillus thuringiensis Serovar Tolworthi Strain Pasteur Institute Standard, *Genome Announc*, **3**.

Koren, S.*, et al.* (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing, *Genome Biol.*, **14**, R101.

Koren, S. and Phillippy, A.M. (2014) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly, *Curr. Opin. Microbiol.*, **23C**, 110-120.

Koren, S.*, et al.* (2012) Hybrid error correction and *de novo* assembly of single-molecule sequencing reads, *Nat. Biotechnol.*, **30**, 693-700.

Koren, S.*, et al.* (2014) Automated ensemble assembly and validation of microbial genomes, *BMC Bioinformatics*, **15**, 126.

Korlach, J. (2014) Phospholink nucleotides for sequencing applications. Google Patents.

Korlach, J.*, et al.* (2010) Real-time DNA sequencing from single polymerase molecules, *Methods Enzymol.*, **472**, 431-455.

Land, M.L.*, et al.* (2014) Quality scores for 32,000 genomes, *Stand Genomic Sci*, **9**, 20.

Lasken, R.S. (2012) Genomic sequencing of uncultured microorganisms from single cells, *Nat. Rev. Microbiol.*, **10**, 631-640.

Lecault, V.*, et al.* (2012) Microfluidic single cell analysis: from promise to practice, *Curr. Opin. Chem. Biol.*, **16**, 381-390.

Lee, H.*, et al.* (2014) Error correction and assembly complexity of single molecule sequencing reads, *bioRxiv*.

Li, Z.*, et al.* (2012) Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph, *Brief Funct Genomics*, **11**, 25-37.

Liu, L.*, et al.* (2012) Comparison of next-generation sequencing systems, *J. Biomed. Biotechnol.*, **2012**, 251364.

Loman, N.J.*, et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms, *Nat. Biotechnol.*, **30**, 434-439.

Lugtenberg, B. and Kamilova, F. (2009) Plant-growth-promoting rhizobacteria, *Annu. Rev. Microbiol.*, **63**, 541-556.

Luo, C.*, et al.* (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample, *PLoS One*, **7**, e30087.

Madoui, M.A.*, et al.* (2015) Genome assembly using Nanopore-guided long and error-free DNA reads, *BMC Genomics*, **16**, 327.

Magoc, T.*, et al.* (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms, *Bioinformatics*, **29**, 1718-1725.

Margulies, M.*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, **437**, 376-380.

Mavromatis, K.*, et al.* (2012) The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation, *PLoS One*, **7**, e48837.

Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA, *Proc Natl Acad Sci U S A*, **74**, 560-564.

McCoy, R.C.*, et al.* (2014) Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements, *PLoS One*, **9**, e106689.

Mehnaz, S., Bauer, J.S. and Gross, H. (2014) Complete genome sequence of the sugar cane endophyte *Pseudomonas aurantiaca* PB-St2, a disease-suppressive bacterium with antifungal activity toward the plant pathogen *Colletotrichum falcatum*, *Genome Announc*, **2**.

Metzker, M.L. (2010) Sequencing technologies - the next generation, *Nat. Rev. Genet.*, **11**, 31-46.

Mikheyev, A.S. and Tin, M.M. (2014) A first look at the Oxford Nanopore MinION sequencer, *Mol Ecol Resour*, **14**, 1097-1102.

Nagarajan, N. and Pop, M. (2013) Sequence assembly demystified, *Nat. Rev. Genet.*, **14**, 157-167.

Nakano, K.*, et al.* (2015) First Complete Genome Sequence of Clostridium sporogenes DSM 795T, a Nontoxigenic Surrogate for Clostridium botulinum, Determined Using PacBio Single-Molecule Real-Time Technology, *Genome Announc*, **3**.

Okutani, A.*, et al.* (2015) Draft Genome Sequences of Bacillus anthracis Strains Stored for Several Decades in Japan, *Genome Announc*, **3**.

Ozsolak, F. and Milos, P.M. (2011) RNA sequencing: advances, challenges and opportunities, *Nat. Rev. Genet.*, **12**, 87-98.

Pacific-BioSciences (2012) Detecting DNA Base Modifications.

Peltola, H., Soderlund, H. and Ukkonen, E. (1984) SEQAID: a DNA sequence assembling program based on a mathematical model, *Nucleic Acids Res.*, **12**, 307-321.

Pevzner, P.A. and Tang, H. (2001) Fragment assembly with double-barreled data, *Bioinformatics*, **17 Suppl 1**, S225-233.

Pevzner, P.A., Tang, H. and Waterman, M.S. (2001) An Eulerian path approach to DNA fragment assembly, *Proc Natl Acad Sci U S A*, **98**, 9748-9753.

Quail, M.A.*, et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, *BMC Genomics*, **13**, 341.

Quick, J., Quinlan, A.R. and Loman, N.J. (2015) Erratum: A reference bacterial genome dataset generated on the MinION(TM) portable single-molecule nanopore sequencer, *Gigascience*, **4**, 6.

Rahman, A. and Pachter, L. (2013) CGAL: computing genome assembly likelihoods, *Genome Biol.*, **14**, R8.

Rinke, C.*, et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter, *Nature*, **499**, 431-437.

Roberts, R.J., Carneiro, M.O. and Schatz, M.C. (2013) The advantages of SMRT sequencing, *Genome Biol.*, **14**, 405.

Ronaghi, M., Uhlen, M. and Nyren, P. (1998) A sequencing method based on real-time pyrophosphate, *Science*, **281**, 363, 365.

Rothberg, J.M.*, et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing, *Nature*, **475**, 348-352.

Salzberg, S.L.*, et al.* (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms, *Genome Res.*, **22**, 557-567.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors, *Proc Natl Acad Sci U S A*, **74**, 5463-5467.

Satou, K.*, et al.* (2014) Complete genome sequences of eight *Helicobacter pylori* strains with different virulence factor genotypes and methylation profiles, isolated from patients with diverse gastrointestinal diseases on Okinawa Island, Japan, determined using PacBio Single-Molecule Real-Time Technology, *Genome Announc*, **2**.

Schneider, G.F. and Dekker, C. (2012) DNA sequencing with nanopores, *Nat. Biotechnol.*, **30**, 326-328.

Shapiro, L.R.*, et al.* (2015) Draft Genome Sequence of Erwinia tracheiphila, an Economically Important Bacterial Pathogen of Cucurbits, *Genome Announc*, **3**.

Shields, C.W.t., Reyes, C.D. and Lopez, G.P. (2015) Microfluidic cell sorting: a review of the advances in the separation of cells from debulking to rare cell isolation, *Lab Chip*, **15**, 1230-1249.

Treangen, T.J.*, et al.* (2009) Genesis, effects and fates of repeats in prokaryotic genomes, *FEMS. Microbiol. Rev.*, **33**, 539-571.

Treangen, T.J. and Salzberg, S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions, *Nat. Rev. Genet.*, **13**, 36-46.

Utturkar, S.M.*, et al.* (2015) Sequence data for Clostridium autoethanogenum using three generations of sequencing technologies, *Sci Data*, **2**, 150014.

Utturkar, S.M.*, et al.* (2014) Evaluation and validation of *de novo* and hybrid assembly techniques to derive high quality genome sequences, *Bioinformatics*.

van Dijk, E.L.*, et al.* (2014) Ten years of next-generation sequencing technology, *Trends Genet.*, **30**, 418-426.

Voskoboynik, A.*, et al.* (2013) The genome sequence of the colonial chordate, Botryllus schlosseri, *Elife*, **2**, e00569.

Walker, B.J.*, et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, *PLoS One*, **9**, e112963.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nat. Rev. Genet.*, **10**, 57-63.

Appendix

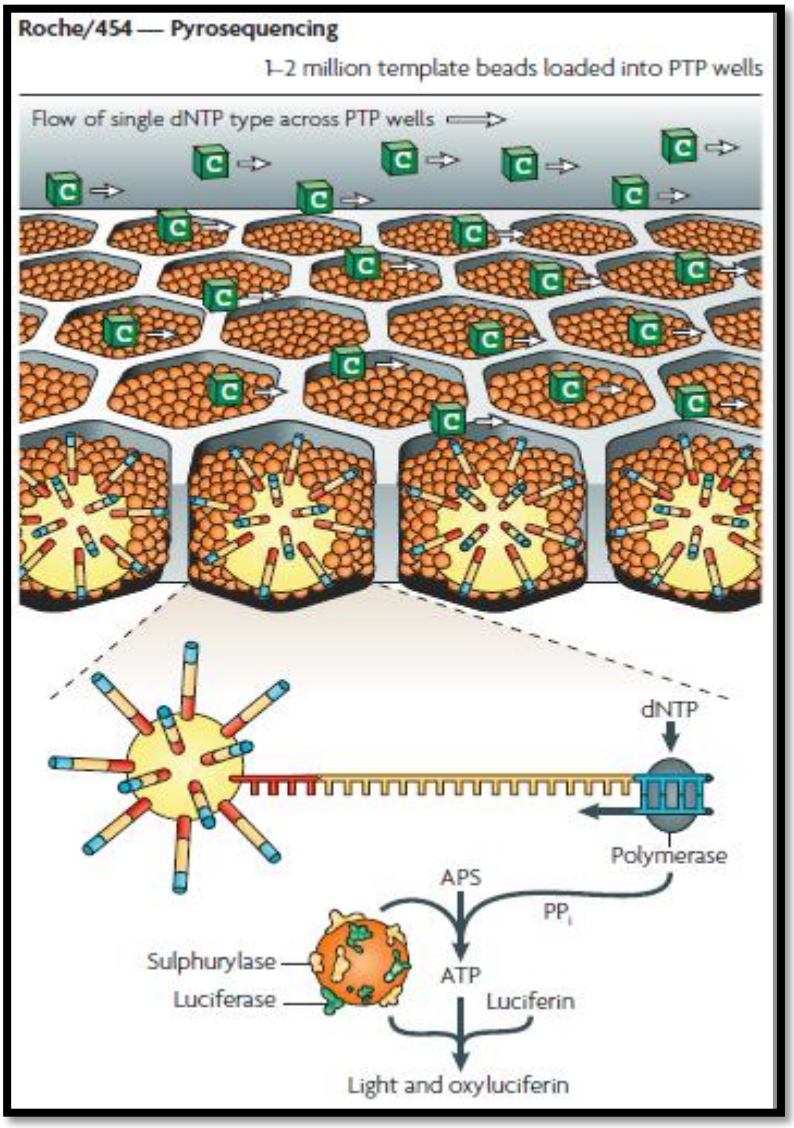

Figure 1.1: Overview of Roche 454 pyrosequencing method (Metzker, 2010).

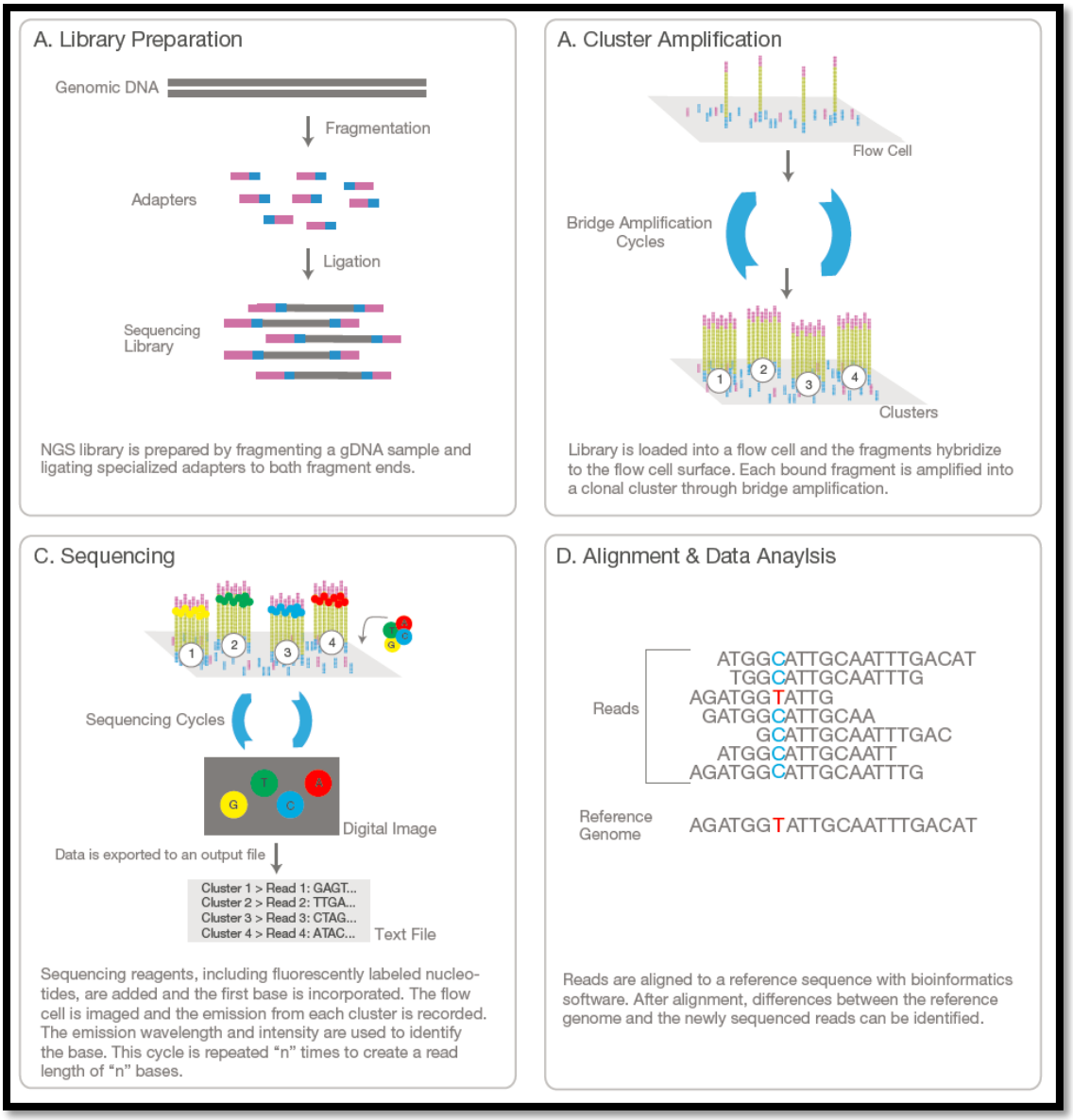

Figure 1.2: Four step workflow for Illumina Sequencing (Illumina Inc. 2015).

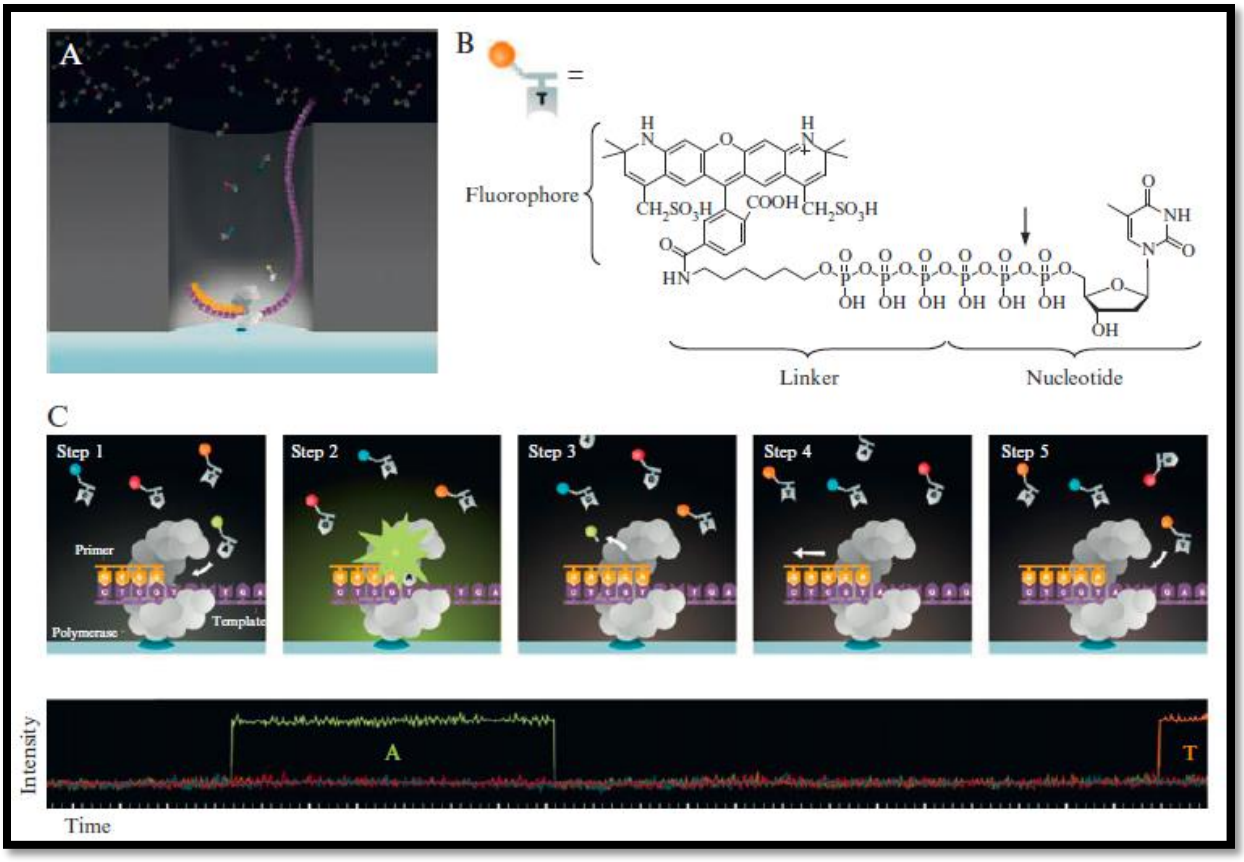

Figure 1.3 : Overview of PacBio sequencing principle(Korlach, et al., 2010).

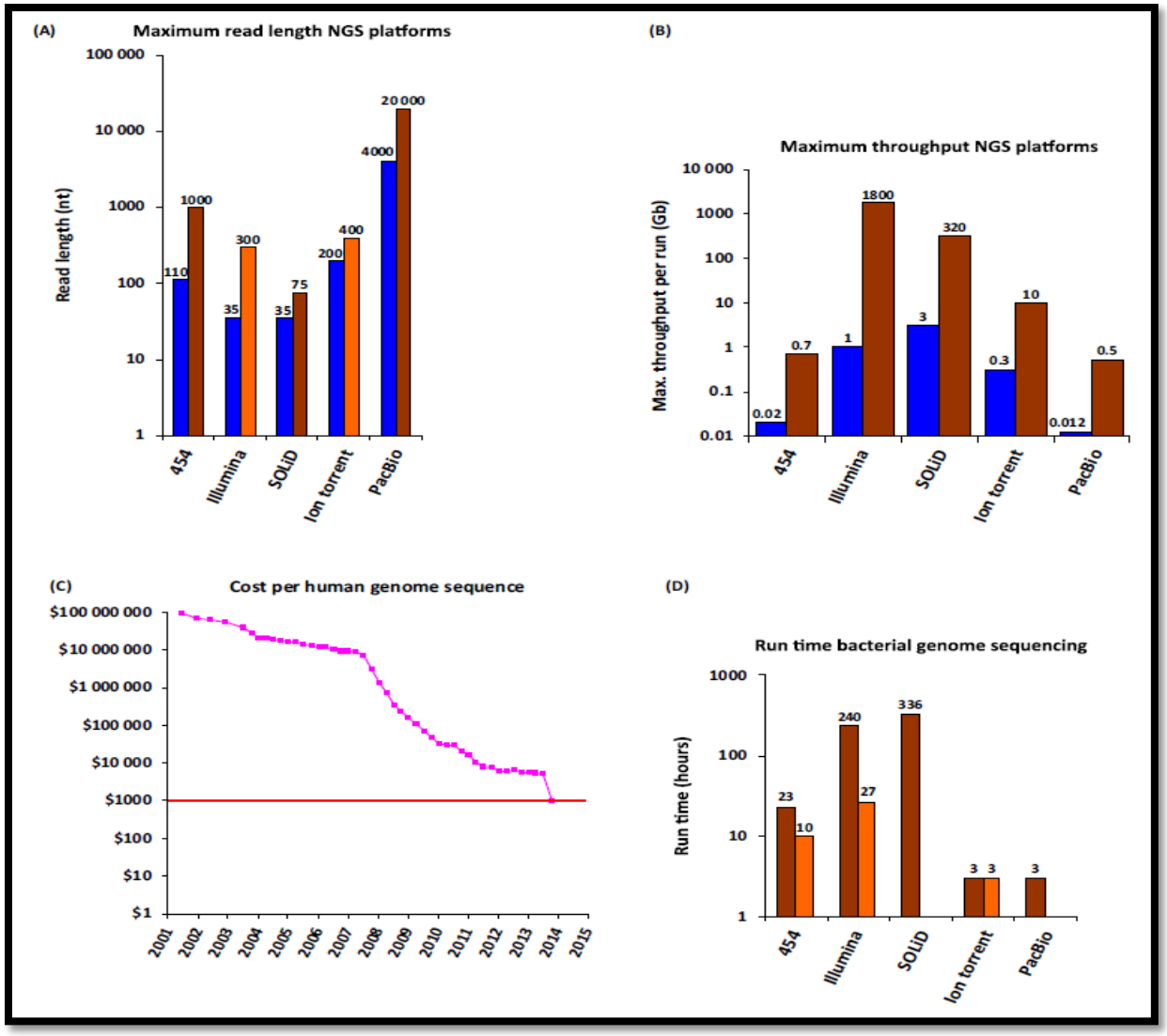

Figure 1.4: Comparison of sequencing platforms (van Dijk, et al., 2014)

CHAPTER 2 : EVALUATION AND VALIDATION OF *DE NOVO* AND HYBRID ASSEMBLY TECHNIQUES TO DERIVE HIGH QUALITY GENOME SEQUENCES

Disclosure: This chapter was published as:

Utturkar S.M., Klingeman D.M., Land M.L., Schadt C.W., Doktycz M.J., Pelletier D.A., and Brown S.D. (2014). Evaluation and validation of *de novo* and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*. 30:2709–2716.

Sagar Utturkar's contributions included leading the bioinformatics analysis; leading the *de novo* and hybrid assemblies and *in silico* evaluations; conceiving of, leading and performing the laboratory and *in silico* analysis of PCR and Sanger sequencing to validate assemblies; contributing to the initial design of the study; and leading the manuscript writing and preparation. Dr. Steven Brown, Dr. Chris Schadt, Dr. Dale Pelletier, Dr. Mitch Doktycz were responsible for the initial study design, decision to perform sequencing with multiple platforms and contributed towards manuscript preparation and corrections. Dawn Klingeman contributed towards genomic DNA isolations, library preparations and genome sequencing using Illumina and 454 platforms.

2.1 Abstract

Motivation: To assess the potential of different types of sequence data, combined with *de novo* and hybrid assembly approaches to improve existing draft genome sequences.

Results: Illumina, 454 and PacBio sequencing technologies were used to generate *de novo* and hybrid genome assemblies for four different bacteria, which were assessed for quality using summary statistics (e.g. number of contigs, N50) and *in silico* evaluation tools. Differences in predictions of multiple copies of rDNA operons for each respective bacterium were evaluated by PCR and Sanger sequencing and then the validated results were applied as an additional criterion to rank assemblies. In general, assemblies employing longer PacBio reads were better able to resolve repetitive regions. In this study, the combination of Illumina and PacBio sequence data assembled through the ALLPATHS-LG algorithm gave the best summary statistics and most accurate rDNA operon number predictions. This study will aid others looking to improve existing draft genome assemblies.

2.2 Introduction

The development and evolution of next-generation sequencing (NGS) platforms has dramatically changed biological studies in recent years (Mavromatis, et al., 2012). Assembly of DNA reads to correctly reconstruct genomes is an essential task to facilitate genomic studies, and a variety of assembly algorithms and methods for quality evaluation have been developed (Nagarajan and Pop, 2013). However, most sequenced genomes are incomplete due to technical difficulties, time and the expense leading to an increasing disparity in quality and usefulness between finished and draft genomes in databases (Chain, et al., 2009).

Due to their low cost, accuracy and high throughput, Illumina platforms have dominated the sequencing industry (Mavromatis, et al., 2012). Short read sequencing technologies have limited power to resolve large repetitive regions even within relatively small microbial genomes (Nagarajan and Pop, 2013). The so-called 'third generation' single-molecule sequencing technology developed by Pacific Biosciences (PacBio) has been compared to several NGS platforms (Quail, et al., 2012). Read lengths up to 14 kb have been reported for PacBio RS I chemistry (Nagarajan and Pop, 2013) and nearly 27 kb for RS II chemistry (Brown, et al., 2014).

Repetitive DNA such as ribosomal DNA (rDNA) operons present one of the greatest technical challenges during the assembly process, which is exacerbated when repeat sequence regions are longer than the read lengths (Treangen and Salzberg, 2012). In many cases where repetitive DNA is present, short read genome assemblies remain highly fragmented and often only achieve high-quality draft status (Chain, et al., 2009). The relative value of a finished genome (Fraser, et al., 2002), technical challenges (Hurt, et al., 2012; Treangen and Salzberg, 2012) and what is missing from finished versus draft quality genomes (Mavromatis, et al., 2012) have been discussed previously. Several strategies proposed and implemented for improving genome assemblies include the use of varying size fragment libraries, longer length reads, gap-closure software and post-processing to detect misassemblies (Treangen and Salzberg, 2012).

Recently, draft genome sequences for 41 bacteria isolated from the *Populus deltoides* rhizosphere and endosphere were obtained using an Illumina Hiseq2000 instrument and the genomes were represented by 187 contigs, on average (Brown, et al., 2012). An additional two genomes were unsuitable for publication at that time due to high contig numbers and 10 of the 43 genomes contained more than 280 contigs. The aim of the present study was to compare and select the most appropriate NGS technology combinations, assembly protocol and parameter optimization to improve the genome assemblies of the *Rhizobium* sp. strain CF080 and *Burkholderia* sp. strain BT03 that originally proved problematic, as well as two other strains, *Pseudomonas* sp. strain GM41 and *Pseudomonas* sp. strain GM30 of biological interest. In addition to a variety of *in silico* techniques for evaluation of genome assemblies, a PCR and Sanger sequencing strategy was used to validate rDNA operon predictions and further assess the assemblies.

2.3 Methods

DNA sequence data generation:

Illumina Paired-End (PE) sequencing has been described (Brown, et al., 2012). Illumina Mate-Pair (MP) libraries with an average insert size of 6 kb were prepared using the Nextera Mate-Pair Sample Preparation Kit following the manufacturer's protocols and sequencing was completed on a MiSeq instrument. Roche 454 libraries were prepared following the "Rapid Library Preparation" method according to manufacturer's recommendations for single end pyrosequencing using the Roche 454 GS FLX System and Titanium XLR70+ kit (Roche 454). PacBio sequencing data were generated at the Genome Sequencing and Analysis Core Resource at Duke University using the PacBio RS-I instrument, C2 chemistry and one SMRT cell per genome. Raw sequence data from all the platforms are available through the NCBI SRA database under accession number SRP010852.

Sequence data trimming, filtering, annotation and assembly:

Quality trimming and filtering of Illumina reads was performed as described previously (Brown, et al., 2012). The assemblers used for the *de novo* and hybrid assembly, their respective versions and assembly recipes are provided (Section 2.1 and Figure 2.1). The final assemblies were annotated by the Prodigal gene calling algorithm (Hyatt, et al., 2010) and Integrated Microbial Genomes (IMG) system (Markowitz, et al., 2012). The best hybrid assemblies for strain CF080, GM30, BT03 and GM41 were deposited at the NCBI Genbank database under accession numbers AKKC00000000, AKJP00000000, AKKD00000000 and AKJN00000000 respectively.

Assessment of genome assembly quality and rDNA analysis:

In silico assembly evaluations were performed using the CGAL (version 0.9.6) and REAPR (version 1.0.16) tools. rDNA operon predictions were performed using RNAmmer software (version 2.3.2) and alignments were created using Geneious software (version 6.1.5) (Auckland, New Zealand). PCR amplification and Sanger sequencing protocols are provided (Section 2.1, Figure 2.1, Table 2.1 and Table 2.2).

2.4 Results and Discussion

Sequencing details.

Illumina PE data were available (Brown, et al., 2012) and additional sequencing was performed using Roche 454, Illumina MP and PacBio RS-I platforms. The average read lengths and coverage values from each sequencing platform are summarized (Table 2.3) (All tables and figures are located in the appendix). Previously published draft genome assemblies generated from Illumina PE reads (Brown, et al., 2012) were improved using combined data from the different sequencing platforms and hybrid assembly protocols.

A non-hybrid assembly method HGAP has been developed which requires 80-100x of PacBio sequence coverage (Chin, et al., 2013) and several recent studies have shown that assembly of PacBio data alone generated the most complete and accurate *de novo* assemblies for several bacteria (Brown, et al., 2014; Koren, et al., 2013). In this study, *de novo* assembly of PacBio RS I data only with the HGAP method generated poor quality assemblies (highly fragmented with low N50 values and having smaller genome size than expected), which was likely due to the relatively low sequence coverage (18-32x). Hence, hybrid assemblies for these four strains were compared using summary statistics, assembly evaluation tools and rDNA content. The performance of each hybrid assembly algorithm is described below. However, for new PacBio sequence data generation one should aim for >100x coverage using the RS II Sequencing System, which can obtain better genome assemblies (Chin, et al., 2013).

In a recent example, a closed, high-quality genome sequence for *Clostridium autoethanogenum* DSM10061 was generated using only the latest single-molecule DNA sequencing technology and without the need for manual finishing (Brown, et al., 2014). Comparison of the PacBio assembly to assemblies based upon shorter read DNA technologies (454, Ion Torrent, and Illumina) showed they were confounded by the large number repeats and their size, which in the case of the rRNA gene operons were ~5 kb. The *C. autoethanogenum* PacBio sequence data cost ~US$ 1,500. A detailed cost-analysis for different sequence data types has been reported (Koren, et al., 2013). Longer reads, greater sequencing depth, the random nature of single molecule sequencing errors, and its cost and assembly performance suggests this technology will be increasingly used to produce finished microbial genomes (Koren, et al., 2013).

Assembly of data from Illumina PE

The initial assemblies of Illumina PE reads were mostly generated using CLC genomics workbench (CLC) (Brown, et al., 2012). We utilized the same dataset and alternative assembly algorithms such as Velvet (Zerbino and Birney, 2008), SOAP (Luo, et al., 2012), ABySS (Simpson, et al., 2009), MaSuRCA (Zimin, et al., 2013) and SPAdes (Bankevich, et al., 2012), which obtained improved assembly statistics. The SPAdes assembler generated the best summary statistics using Illumina PE reads with an exception of strain CF080. The ABySS assembler performed consistently for all four strains, as it generated similar statistics to the SPAdes assembler as well as generating the best assembly for strain CF080 using PE data. The performance of the MaSuRCA assembler was genome and data dependent as it generated poor assembly statistics for strain BT03 and GM30 while reasonable assembly statistics for strain CF080 and GM41 (Table 2.4).

Assembly of Illumina PE and MP data.
MP libraries are capable of resolving repetitive regions and structural variants while increasing the accuracy and size of assembled contigs (Ribeiro, et al., 2012). Short reads could be best assembled through de Bruijn Graph (DBG) assembly approach (Miller, et al., 2010). The PE-MP hybrid assemblies generated by DBG based ABySS, SOAP, Velvet and MaSuRCA were only slightly better than the previously published PE-only assemblies (Brown, et al., 2012) while greater improvements in summary statistics were obtained by SPAdes and ALLPATHS-LG assemblers (Table 2.5). In this study, the ALLPATHS-LG algorithm (Butler, et al., 2008) outperformed the SPAdes assemblies in terms of contig numbers and generated superior hybrid assemblies. The optimal performance of ALLPATH-LG can be attributed to a specific type of library requirement where PE and MP reads are designed to overlap each other and can be joined to yield roughly twice the read length of individual reads (Nagarajan and Pop, 2013). In recent years, the ALLPATHS-LG algorithm has arguably won the Assemblathon (Earl, et al., 2011) and GAGE (Salzberg, et al., 2012) competitions by employing this assembly approach.

Hybrid assembly of Illumina and Roche 454 data
Longer reads from the 454 platform could be best assembled through Overlap-Layout Consensus (OLC) approach (Miller, et al., 2008). The assembly of native, shotgun 454 reads through Newbler generated better summary statistics as compared to PE data alone (Table 2.5). One 454-Illumina hybrid assembly approach involved merging the 454-only assembly with Illumina reads by PHRAP (version 1.09) (de la Bastide and McCombie, 2007) or Minimus (version 3.0.1) (Sommer, et al., 2007) to extend contigs. In this study, PHRAP and Minimus merged assemblies often generated aberrant results (e.g., 1-2 Mb genome assemblies for 5-6 Mb *Pseudomonas* genomes) and contained a high number of singleton (non-assembled) sequences. Additionally, hybrid assembly is supported by the CLC, MaSuRCA and Celera (Miller, et al., 2008) assemblers. Hybrid assembly of Illumina and 454 reads was expected to exceed the 454 only assembly statistics based on earlier studies (Brown, et al., 2012). However, CLC did not substantially improve the assembly statistics. MaSuRCA hybrid assemblies with PE-MP-454 combination generated improved N50 values but contained high number of contigs as compared to 454 only assemblies of four strains (Table 2.4).

The Newbler software supports fasta/fastq input along with native 454 reads. However, when quality-trimmed Illumina reads or draft assembly of Illumina reads were used as additional input, Newbler failed to complete the assembly process. This was likely due to the large size of Illumina data or very long fasta sequences, respectively. Therefore, draft assemblies were cut into 1.5 kb pseudo reads with 300 bp overlap using fb_dice.pl script from the FragBlast module (http://www.clarkfrancis.com/codes/fb_dice.pl) and assembled together with native 454 reads using Newbler (Figure 2.2) (All tables and figures are located in the appendix), as described previously (Brown, et al., 2012), which alleviated failure issues and resulted in substantial improvements in N50 statistics and appropriate genome size estimates were maintained (Table 2.5). The *in silico* approach to generate 1.5 kb overlapping pseudo reads was influenced by the quality of initial draft assembly. Shredding of PE-MP hybrid assemblies (which had better summary statistics)

achieved better results as compared to shredding of PE only assemblies. Therefore, it appears that even when employing this shredding technique, generating the optimal draft genome assemblies from Illumina data prior to shredding is an important step towards successful hybrid assembly. Any misassembly in the initial assembly risks being propagated into the hybrid assembly.

To attain insight into the draft assembly generation, summary statistics of previously published draft assemblies of 43 bacterial isolates (Brown, et al., 2012) generated using four different assemblers are given (Table 2.6 and Table 2.7) and important parameters that influenced the assembly process are described below. Poor quality sequencing reads can adversely affect the assembly process (Salzberg, et al., 2012) and we observed that quality-based trimming of raw data gave approximately 15 fold improvements in N50 statistics. The assembly of PE Illumina reads by the ABySS and SPAdes assembler generated highest N50 statistics when compared to results from the Velvet, SOAP and CLC assemblers (Table 2.4, Table 2.6 and Table 2.7). Different Kmer values were tested (Chikhi and Medvedev, 2014) and optimal summary statistics were obtained at higher Kmer values, up to 60, and beyond this value summary statistics deteriorated (Table 2.6 and Table 2.7). The increase in raw read coverage up to 300x generated concomitant increases in N50 values while beyond 300x coverage the N50 statistics did not increase (Figure 2.1). Therefore, the quality and sequence coverage of raw reads, Kmer value and appropriate assembly algorithm selection are essential parameters for optimization of draft genome assemblies. We recommend using the ABySS assembler with Illumina PE data and ALLPATHS-LG or SPAdes assembler with Illumina PE-MP data for optimal results. Although we used N50 statistics for the initial shortlisting of assemblies, it should be noted that large N50 values are not always indicative of assembly quality and additional validation should be performed using various bioinformatics tools as described by (Koren, et al., 2014) and rDNA analysis approach described below.

Hybrid assembly of Illumina, 454 and PacBio data.
Single molecule sequencing technology currently produces the longest read lengths across all NGS platforms and the performance of PacBio RS sequencing system has been compared to other NGS platforms recently (Liu, et al., 2012; Quail, et al., 2012). The longer reads generated with the PacBio system have the potential to exceed the longest repeats in most bacterial genomes and greatly improve the genome assemblies (Koren, et al., 2013). However, PacBio sequencing technology has a high error rate, which has been reported as being 18% (Nagarajan and Pop, 2013). Different hybrid assembly protocols have been developed to overcome the high error rates associated with the single molecule sequencing technology and limitations of short read technologies (Bashir, et al., 2012; English, et al., 2012; Koren, et al., 2012; Ribeiro, et al., 2012). Various hybrid assembly protocols to improve earlier assemblies were pursued and results are described below.

<u>PacBio corrected Reads (PBcR) pipeline</u>
The higher error rate associated with PacBio technology obscures the read alignments and complicates the assembly process. Most genome assemblers are unable to handle this high error rate and hence error correction becomes necessary to unlock the full

potential of longer reads for *de novo* assembly. The PBcR pipeline uses higher fidelity Illumina and/or 454 reads to trim and correct the individual long-read sequences and generates hybrid consensus with > 99.99% base-call accuracy (Koren, et al., 2012). We employed 454 reads to correct errors in PacBio reads through the PBcR pipeline, which were then assembled via the Celera assembler (Miller, et al., 2008). The PBcR hybrid assembly statistics were similar to those generated with PE-MP and PE-454 combinations (Table 2.5). The PBcR assemblies contained few collapsed repeats as compared to other assemblies (Table 2.8), which is likely a product of longer, corrected reads. It should be noted that like HGAP, the PBcR pipeline is also capable of performing self-correction and non-hybrid assembly of PacBio reads when sufficient (~100x) coverage is available. However, due to the PacBio coverage limitation we could not perform the self-correction approach.

The AHA scaffolding method

The AHA scaffolding approach (Bashir, et al., 2012) is available through the SMRT analysis package (version 2.0, Pacific Biosciences) and it uses any previous assembly to which longer PacBio reads are aligned using the BLASR algorithm (Chaisson and Tesler, 2012) to create higher, ordered scaffolds. We used the best contig assembly generated through PE-MP-454 combination and error corrected PacBio reads as an input to AHA protocol. The resulting scaffolds were ranked second best after the ALLPATHS-LG (Table 2.5).

ALLPATHS-LG

The ALLPATHS-LG recipe uses a mixture of three data types, where Illumina PE and MP reads are assembled first using de Bruijn graph approach and then PacBio reads are incorporated to patch coverage gaps and resolve repeats (Maccallum, et al., 2009). The ALLPATHS-LG method requires all inputs in raw format and employs its own error correction pipeline. ALLPATHS-LG assemblies with PE-MP combination were found to be superior to the numerous other protocols compared here and consistent with earlier studies (Earl, et al., 2011; Salzberg, et al., 2012). Incorporation of PacBio reads with this method further improved the assembly results up to 'noncontiguous finished' quality (Table 2.5). However, incorporation of PacBio reads was memory intensive, the software crashed multiple times on a high memory (132 GB) server and it was unable to assemble the BT03 genome. This behaviour may be attributed to some combination of computational memory limitation; higher genome BT03 size (~11 Mb); and its content (the genome contained numerous phage and transposon sequences). Our datasets contained one MP library with ~6 kb insert sizes and achieved near-finished genome assemblies. Ribeiro et al. used multiple MP libraries with insert sizes ranging from 2-6 kb and were able to generate finished or near-finished assemblies for different bacterial genomes (Ribeiro, et al., 2012). Hence, inclusion of multiple MP libraries of varying length could be a possible path to further improve the assemblies in the future.

SPAdes

Recent GAGE-B comparisons identified SPAdes as one of the best algorithms for bacterial genome assemblies using Illumina data. Indeed, consistent with previous findings SPAdes performed well to assemble our four genomes using Illumina PE-MP

data. Recently SPAdes added support for the PacBio data which allowed a direct comparison of its performance to ALLPATHS-LG for PE-MP-PacBio combinations. The overall summary statistics generated by both assemblers were very similar but ALLPATHS-LG assemblies always contained lower contig numbers than SPAdes. Notably, SPAdes seamlessly assembled the PE-MP-PacBio combination for strain BT03 for which ALLPATHS-LG encountered crashing issues associated with memory limitation.

Gap filling by PBJelly algorithm

The PBJelly method (English, et al., 2012) aligns PacBio/454 reads to the scaffold assembly to extend the contigs and resolve the gaps. The PBJelly algorithm was applied to the best scaffolded assemblies generated by ALLPATHS-LG together with the PacBio reads. PBJelly was able to fill up (64%, 99% and 93%) gaps in BT03, CF080 and GM41 genomes, respectively (Table 2.9). Many microbial genomics analyses depend on the finished genomes and single unbroken contig is important for a wide range of disciplines (Koren, et al., 2013). Scaffolded assemblies are very helpful in the genome finishing process and are used to determine contig order and contig overlap (Nagarajan, et al., 2010; Swain, et al., 2012). Long range PacBio reads offer an attractive opportunity to reduce the number of gaps and resolve unidentified base-pairs (N's) in the scaffolds which reduces the overall cost of manual finishing.

Assembly Quality Assessments and Comparisons

Although the assembly metrics such as N50 and contig numbers are widely used for the assembly evaluation, they may not always correlate well with the actual quality of the assembly (Nagarajan and Pop, 2013) and several other bioinformatics approaches and metrics have been developed to assess assembly quality (Gurevich, et al., 2013; Hunt, et al., 2013; Koren, et al., 2014; Rahman and Pachter, 2013). The Computing Genome Assembly Likelihoods (CGAL) is one recent approach that incorporates genome coverage and assembly accuracy into the evaluation without need of reference sequence and combines them into a single metric score (Rahman and Pachter, 2013). The CGAL software ranked the SPAdes assemblies as highest, while ALLPATHS-LG and MaSuRCA assemblies have scores close to the SPAdes assemblies (Table 2.10). The REAPR genome assembly evaluation tool generates a positional error call metric, assesses potential collapsed repeats and single base-by-base scores (Hunt, et al., 2013). The REAPR evaluation generated the least number of error calls for the ALLPATHS-LG assemblies generated with Illumina only (PE-MP) data (Table 2.8). CGAL and REAPR both assigned high rankings to ALLPATHS-LG assemblies likely reflecting their higher accuracy and depth of coverage.

On the other hand, hybrid assemblies employing 454/PacBio reads which had better summary statistics were assigned with lower CGAL scores and a large number of error calls by REAPR (Table 2.8 and Table 2.10). These inconsistent scores by CGAL/REAPR are possibly due to the design limitation of these *in silico* evaluation tools which cannot currently use 454/PacBio reads during the evaluation. The 454/PacBio reads may have included data for repetitive regions that are not spanned by the Illumina reads and reported as errors based on evaluation by Illumina reads. To improve the consensus accuracy of PacBio assemblies we performed assembly polishing using the Quiver tool (Chin, et al., 2013). However, low coverage of PacBio reads may not have achieved the

required base-call quality and contributing towards low scores by *in silico* evaluation tools. REAPR detected fewer collapsed repeats in the assemblies employing PacBio reads (Table 2.8) and this suggests that the longer PacBio reads better resolved repetitive regions.

Reciprocal BLASTP analyses were conducted using proteins predicted from the draft and the best hybrid assemblies in order to gain insights into potential protein encoding differences (Table 2.11). The majority (87-98%) of proteins were unchanged by assembly improvements supporting the notion that for some studies draft quality genome sequences may be sufficient. However, a substantial number of proteins were longer after assembly improvement and a number of new proteins were predicted in most cases. The majority of newly predicted proteins were for hypothetical proteins, and others included genes with predicted regulatory functions or metabolic genes such as for a putative nitric oxide dioxygenase. The number of potential missing genes will be genome and assembly-specific, and this is difficult to assess in the absence of available finished reference genomes (Fraser, et al., 2002).

Assembly Validation

The CGAL and REAPR evaluation methods were only able to rank the assemblies based on number of errors, and verification of the error calls would require finished reference genome sequences, which were beyond the scope of the present study. Therefore, an additional level of verification was necessary to better assess assembly accuracy. Since genome assemblers are often confounded by large repetitive regions (e.g. 5-7 kb rDNA operons), (Treangen and Salzberg, 2012) accurate prediction of rDNA operon was selected as an additional criterion to assess the assembly accuracy and to gain insight into potential systematic issues.

Several copies of 5S, 16S, and 23S rDNA elements were predicted for strains CF080, GM41, GM30 and BT03 and in this study the complete rDNA operon is defined as an arrangement of 5S, 16S, and 23S rDNA elements in single operon structure on a single contig. rDNA genes were predicted by the RNAmmer program (Lagesen, et al., 2007) and predictions were tested using a PCR-based approach. Briefly, oligonucleotides were designed to bind to DNA regions that were 5’ and 3’ to the predicted rDNA operons and give amplified products of a predicted size. Additional internal oligonucleotides were designed to amplify and sequence end regions. Correct assembly of the rDNA operon was expected to generate a PCR product in the desired size range while an incorrectly assembled rDNA operon would fail to amplify or give unexpected sequence lengths. Measured and expected product sizes for positive PCR reactions for each rDNA operon in each strain are shown (Table 2.1), along with the length of DNA sequence that was verified by Sanger sequencing (Table 2.2). These presumptive positive results support this experimental approach, although the entire PCR product could be sequenced by primer-walking for increased assembly confidence.

<u>rDNA operons in Rhizobium sp. strain CF080</u>

Summary statistics and bioinformatics assessment suggested the ALLPATHS-LG assembly was optimal for strain CF080 (Table 2.5, Table 2.6, Table 2.8 and Table 2.10)

and three rDNA operons and their flanking chromosomal regions were predicted on three separate contigs (Figure. 2.2). The SPAdes assembly with PE-MP-PacBio combination also predicted three rDNA operons and a similar arrangement as in ALLPATHS-LG assemblies. Three copies of rDNA operons have been detected within 6 finished *Rhizobium* genomes sequences. The ~7 Mb ALLPATHS-LG genome assembly supported predictions for three rDNA operons, which were validated by PCR and Sanger sequencing. ABySS generated an assembly that was approximately 8 Mb in size and it supported predictions for six rDNA operon copies (Figure. 2.2). However, the ABySS assembly was unable to resolve regions of DNA that were 5' and 3' of different rDNA operons leading to their duplication within the assembly (Figure. 2.2). The rDNA operon duplication in the ABySS assembly accounts for a portion but not all of the higher genome size reported. Previous studies that employed the ABySS assembly method have also noted that ABySS assembler predicted larger genome sizes as compared to other methods (Haridas, et al., 2011; Salzberg, et al., 2012) but did not identify the specific reasons for these higher genome sizes. The Velvet and CLC algorithms were able to assemble only one complete rDNA operon in strain CF080 and were unable to predict flanking chromosomal regions; this is likely a contributing factor to these assemblies being more fragmented (Table 2.5). Hence, the ALLPATHS-LG assembly having the best summary statistics and accurate prediction of 3 copies of rDNA operons was selected as the best assembly for strain CF080. An analysis of rDNA operons in *Pseudomonas* sp. strains GM41 and GM30, and in *Burkholderia* sp. strain BT03 are presented (Figure 2.3 and Figure 2.4).

Comparison of Assembly Approaches

In this study, we examined a variety of *de novo* genome assembly methodologies for four novel bacterial isolates that do not have existing reference sequences. There are a large number of different assemblers and different parameters that one can employ for *de novo* studies. Numerous recent studies report continued assembly developments and comparisons, which reflects the importance of generating a high-quality, representative genome sequence (Bradnam, et al., 2013; Powers, et al., 2013). It has been shown that a number of assemblers perform well when a single metric is considered but few perform consistently across a set of quality metrics. In this study, in addition to a range of *in silico* methods we experimentally examined rDNA operons predictions from different assemblies, which provided an additional criterion for assembly quality assessment.

2.5 Conclusions

The ABySS and SPAdes software generated the best assembly statistics when only PE Illumina reads were used. The ABySS assembler performed well consistently for all four genomes and also correctly identified multiple copies of rDNA operons (Figure. 2.2). As expected, additional sequencing data from each NGS platform improved the assembly statistics (Table 2.5). Hybrid assemblies with PE-MP data combinations were superior as compared to PE-454 combinations. However, the superiority of the PE-MP combination can likely be attributed to the excellent performance of the ALLPATHS-LG and SPAdes algorithms. Inclusion of PacBio data resulted in substantial improvements in assembly statistics but success was dependent on the selection of assembly approach. The PBcR assembly statistics were comparable to that of the PE-454 combination. The AHA and

PBJelly methods facilitated scaffolding and gap filling, respectively and would be helpful during genome finishing. Among the eleven *de novo* and hybrid assembly protocols tested here, the ALLPATHS-LG assembler with the combination of PE-MP-PacBio data generated the best results and also provided the most accurate rDNA operons predictions, except in the case of the BT03 genome where computational resource limitations prevented evaluation. These results underscore the importance of comparing multiple appropriate algorithms and key parameters for genome assembly. Our results were consistent with earlier studies that demonstrated the advantage of including longer PacBio reads (Roberts, et al., 2013; Shin, et al., 2013) and our hybrid assembly results with PacBio data demonstrate the power of these longer reads to better resolve repetitive sequence regions. The evaluation framework described here should prove useful for others looking to improve existing draft genome sequences.

Our results showed that by using complementary libraries, sequencing technologies and appropriate hybrid assembly protocols, dramatic improvements in assembly quality for bacterial genomes could be obtained. The rDNA operon analysis through PCR and Sanger sequencing provided additional confidence for the assembly accuracy. The genomes for strains GM41 and GM30 were previously defined as “high-quality draft” (Brown, et al., 2012), using described criteria (Chain, et al., 2009) while previous assemblies for CF080 and BT03 consisted of 1,039 and 690 contigs respectively. The improved CF080 and BT03 genomes are now represented by 16 and 135 contigs, respectively. CF080 and GM41 assemblies can now be termed as "noncontiguous finished" where automated improvements have been performed and most of the gaps have been resolved (5 and 4 scaffolds respectively). The GM30 and BT03 can be termed as "improved high-quality draft".

References

Bankevich, A.*, et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J Comput Biol*, **19**, 455-477.

Bashir, A.*, et al.* (2012) A hybrid approach for the automated finishing of bacterial genomes, *Nat. Biotechnol.*, **30**, 701-707.

Bradnam, K.*, et al.* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species, *GigaScience*, **2**, 10.

Brown, S.*, et al.* (2014) Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of *Clostridium autoethanogenum* and analysis of CRISPR systems in industrial relevant Clostridia, *Biotechnology for Biofuels*, **7**, 40.

Brown, S.D.*, et al.* (2012) Draft genome sequence of *Rhizobium* sp. strain PDO1-076, a bacterium isolated from *Populus deltoides*, *J Bacteriol*, **194**, 2383-2384.

Brown, S.D.*, et al.* (2012) Twenty-one genome sequences from *Pseudomonas* species and 19 genome sequences from diverse bacteria isolated from the rhizosphere and endosphere of *Populus deltoides*, *J. Bacteriol.*, **194**, 5991-5993.

Butler, J.*, et al.* (2008) ALLPATHS: *de novo* assembly of whole-genome shotgun microreads, *Genome Res.*, **18**, 810-820.

Chain, P.S.*, et al.* (2009) Genomics. Genome project standards in a new era of sequencing, *Science*, **326**, 236-237.

Chaisson, M.J. and Tesler, G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory, *BMC Bioinformatics*, **13**, 238.

Chikhi, R. and Medvedev, P. (2014) Informed and automated k-mer size selection for genome assembly, *Bioinformatics*, **30**, 31-37.

Chin, C.S.*, et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, *Nat. Methods*, **10**, 563-569.

Chin, C.S.*, et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, *Nat Methods*, **10**, 563-569.

de la Bastide, M. and McCombie, W.R. (2007) Assembling genomic DNA sequences with PHRAP, *Curr. Protoc. Bioinformatics*, **Chapter 11**, Unit11 14.

Earl, D.*, et al.* (2011) Assemblathon 1: a competitive assessment of *de novo* short read assembly methods, *Genome Res.*, **21**, 2224-2241.

English, A.C.*, et al.* (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology, *PLoS One*, **7**, e47768.

Fraser, C.M.*, et al.* (2002) The value of complete microbial genome sequencing (you get what you pay for), *J. Bacteriol.*, **184**, 6403-6405.

Gurevich, A.*, et al.* (2013) QUAST: quality assessment tool for genome assemblies, *Bioinformatics*, **29**, 1072-1075.

Haridas, S.*, et al.* (2011) A biologist's guide to *de novo* genome assembly using next-generation sequence data: A test with fungal genomes, *J Microbiol Methods*, **86**, 368-375.

Hunt, M.*, et al.* (2013) REAPR: a universal tool for genome assembly evaluation, *Genome Biol.*, **14**, R47.

Hurt, R.A.*, et al.* (2012) Sequencing intractable DNA to close microbial genomes, *PLoS One*, **7**, 7.

Hyatt, D.*, et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, **11**, 119.

Koren, S.*, et al.* (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing, *Genome Biol.*, **14**, R101.

Koren, S.*, et al.* (2012) Hybrid error correction and *de novo* assembly of single-molecule sequencing reads, *Nat. Biotechnol.*, **30**, 693-700.

Koren, S.*, et al.* (2014) Automated ensemble assembly and validation of microbial genomes, *bioRxiv*.

Lagesen, K.*, et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes, *Nucleic Acids Res.*, **35**, 3100-3108.

Liu, L.*, et al.* (2012) Comparison of next-generation sequencing systems, *J. Biomed. Biotechnol.*, **2012**, 251364.

Luo, R.*, et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler, *Gigascience*, **1**, 18.

Maccallum, I.*, et al.* (2009) ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads, *Genome Biol.*, **10**, R103.

Markowitz, V.M.*, et al.* (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system, *Nucleic Acids Res.*, **40**, D115-122.

Mavromatis, K.*, et al.* (2012) The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation, *PLoS One*, **7**, e48837.

Miller, J.R.*, et al.* (2008) Aggressive assembly of pyrosequencing reads with mates, *Bioinformatics*, **24**, 2818-2824.

Miller, J.R., Koren, S. and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data, *Genomics*, **95**, 315-327.

Nagarajan, N.*, et al.* (2010) Finishing genomes with limited resources: lessons from an ensemble of microbial genomes, *BMC Genomics*, **11**, 242.

Nagarajan, N. and Pop, M. (2013) Sequence assembly demystified, *Nat. Rev. Genet.*, **14**, 157-167.

Nagarajan, N. and Pop, M. (2013) Sequence assembly demystified, *Nat Rev Genet*, **14**, 157-167.

Powers, J.*, et al.* (2013) Efficient and accurate whole genome assembly and methylome profiling of *E. coli*, *BMC Genomics*, **14**, 675.

Quail, M.A.*, et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, *BMC Genomics*, **13**, 341.

Rahman, A. and Pachter, L. (2013) CGAL: computing genome assembly likelihoods, *Genome Biol.*, **14**, R8.

Ribeiro, F.J.*, et al.* (2012) Finished bacterial genomes from shotgun sequence data, *Genome Res.*, **22**, 2270-2277.

Roberts, R.J., Carneiro, M.O. and Schatz, M.C. (2013) The advantages of SMRT sequencing, *Genome Biol.*, **14**, 405.

Salzberg, S.L.*, et al.* (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms, *Genome Res.*, **22**, 557-567.

Shin, S.C.*, et al.* (2013) Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes, *PLoS One*, **8**, e68824.

Simpson, J.T.*, et al.* (2009) ABySS: a parallel assembler for short read sequence data, *Genome Res.*, **19**, 1117-1123.

Sommer, D.D.*, et al.* (2007) Minimus: a fast, lightweight genome assembler, *BMC Bioinformatics*, **8**, 64.

Swain, M.T.*, et al.* (2012) A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs, *Nat. Protoc.*, **7**, 1260-1284.

Treangen, T.J. and Salzberg, S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions, *Nat. Rev. Genet.*, **13**, 36-46.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs, *Genome Res.*, **18**, 821-829.

Zimin, A.V.*, et al.* (2013) The MaSuRCA genome assembler, *Bioinformatics*, **29**, 2669-2677.

Appendix

Section 2.1: Supplementary methods. This section provides information about software versions used, assembly recipe, assembly evaluations and PCR and Sanger sequencing for rDNA analysis.

Genome Assembly

Software versions:

ABySS (version 1.3.2), Velvet (version 1.2.1), CLC Genomics Workbench (version 4.7), SOAPdenovo (version 1.05), ALLPATHS-LG (release 44849), Newbler (version 2.6), PHRAP (version 1.09), Minimus (version 3.0.1), PBJelly (version 12.9.14), PBcR pipeline (version 7.0), AHA and HGAP – SMRTanalysis (version 2.0), SPAdes (version 3.0.0), MaSuRCA (version 2.2.1)

Assembly Recipe:

The genome assembly recipes for Velvet, ABySS and SOAPdenovo and ALLPATHS-LG were followed from the GAGE protocols (Salzberg, et al., 2012). The genome assembly recipes for Spades and MaSuRCA were followed from the respective user manuals. The CLC Genomics Workbench and SMRT analysis assemblies were performed using default settings. The hybrid assemblies with Newbler were generated as described previously (Brown, et al., 2012).The PHRAP, Minimus, PBJelly and PBcR assemblies were performed as per the instructions in respective manuals and with default parameters.

Assessment of genome assembly quality and rDNA analysis

Assembly evaluation

The summary statistics for the assembly were calculated using (summrizeAssembly.py) script which is part of PBJelly software. The CGAL (version 0.9.6) and REAPR (version 1.0.16) assembly evaluations were performed as per the instructions in the respective manuals and with default parameters.

Prediction of rDNA operons

Individual rDNA (16S, 23S and 5S) sequences were predicted using RNAmmer software (version 2.3.2) (Lagesen, et al., 2007) and operon arrangements were determined manually using genomic positions. The 5' and 3' flanking chromosomal region of rDNA operon (1,000 bp on either side) were extracted (when available) using the custom Perl script. Alignments of rDNA sequences (including 5' and 3' flanking chromosomal regions) were performed using Geneious software (version 6.1.5) (Auckland, New Zealand).

PCR and Sanger sequencing

PCR primers for each predicted rDNA operon were designed using the Primer3 software (Untergasser, et al., 2012) and PCR reactions were carried out using Phusion high-fidelity PCR master mix with HF buffer (New England BioLabs) according to the manufacturer's instructions. The PCR amplification conditions include annealing for 30 seconds (at temperature stated in Table 2.1), an extension at 72^0 C for 1 minute (for products <2 kb) or 10 minute (for products < 2 kb) with 20 cycles. The PCR product purification was performed using QIAquick PCR purification kit as per the manufacturer's instruction. The

verification of PCR products was performed through Sanger sequencing by employing standard approach described previously (Brown, et al., 2011).

Table 2.1: PCR primers, annealing temperatures, expected and measured product lengths for each rDNA operon.

Strain	Operon ID	Forward Primer	Reverse Primer	Annealing Temp. (^{0}C)	Expected PCR Product (bp)	Measured PCR Product* (~Kb)
CF80	OP1_APLG_101	CGATGAAGCCTTAGCCTTGT	CTGGCCTGAAATCGACTGTT	65	6609	6.6
	OP2_APLG_102	CGCACAAGAATGGAAGGAAT	CATCTGAGGATTTGCGAGGT	65	6725	6.7
	OP3_APLG_01	GTTTTGACGGTTGTGCCTTC	GGCAGTTTCGAACTGTCCTT	65	6701	6.6
	Abyss_3178_1	AAGAAGGTTTATCCGGTTCG	CTCAAGACGCGGGAGAGTAG	67	606	0.6
	Abyss_3177_1	GATTCCCACGCGTTACTCAC	CGTCTGCGCTTGATTCAATA	65	701	0.7
	Abyss_3177_2	GGTCGGTCGGGAGCTCTAT	GATTCCCACGCGTTACTCAC	65	626	0.6
	Abyss_3179_1	CTCAAGACGCGGGAGAGTAG	GCGCTTTCCTCTTTGCTCT	65	607	0.6
	Abyss_3144_1	GAAAACAGTCTCCGGGAAAA	GATTCCCACGCGTTACTCAC	65	601	0.6
	Abyss_3142_1	CTCAAGACGCGGGAGAGTAG	GGGTTTCCAAAGTCATCGAA	67	714	0.7
GM41	OP1_APLG_Contig21	GTGGTCGATCGCACCTTTAT	ATGTCAGCATGCAAGTCTCG	67	6009	6
	OP2_APLG_Contig22	TAAGGAGTGGGCGGTTTATG	GCAAGTCGCACTCATGACAC	65	6109	6
	OP3_APLG_Contig11	GGTTGCCGAGGTTATTGAAG	TCCGAAGTAGGAAGCGAGAG	65	6035	6
	OP4_APLG_Contig12	CTGGGTATCCGCAACAATCT	GCCTCGAACCACGGTAGATA	65	6184	6.1
	OP1_APLG_Contig21	GTGGTCGATCGCACCTTTAT	ATTCCGATTAACGCTTGCAC	65	1205	1.2
	OP1_APLG_Contig21	GAAAGCATCTAAGCGGGAAA	GATGGGCCAATCACAAGAAG	65	800	0.8
	OP2_APLG_Contig22	GTGGGCGGTTTATGCTTCTA	ATTCCGATTAACGCTTGCAC	65	1216	1.2
	OP2_APLG_Contig22	GAAAGCATCTAAGCGGGAAA	GGTAGAGCAACAGGCCGTAA	65	1018	1

Table 2.1 continued …

Strain	Operon_ID	Forward Primer	Reverse Primer	Annealing Temp. (^{0}C)	Expected PCR Product (bp)	Measured PCR Product* (~Kb)
GM41	OP3_APLG_Contig11	GGTTGCCGAGGTTATTGAAG	GAAAGCATCTAAGCGGGAAA	65	601	0.6
	OP3_APLG_Contig11	ATTCCGATTAACGCTTGCAC	TCCGAAGTAGGAAGCGAGAG	65	1205	1.2
	OP4_APLG_Contig12	CTGGGTATCCGCAACAATCT	GAAAGCATCTAAGCGGGAAA	65	846	0.8
	OP4_APLG_Contig12	ATTCCGATTAACGCTTGCAC	GCCTCGAACCACGGTAGATA	65	1307	1.3
BT03	OP1_CLC_297	TCAACAGCCGATAAGTGTGG	GGGGTCTTGGTCTTGGGTAA	66	5469	5.5
	OP2_Abyss_10697	GTTTAGGGCGTGGACTACCA	TGCTTACGAGACGACATTGG	66	1404	1.3
	OP3_Abyss_10695	GTTTAGGGCGTGGACTACCA	CCTTTGATTGAATAGGCGAGT	65	1208	1.2
	OP4_Abyss_10696	GTTTAGGGCGTGGACTACCA	TGAAGACCACGTGCAAGTTC	65	1370	1.3
	OP5_Abyss_10833	CGCAGGCAAGTGTCTAGAAT	GTTTAGGGCGTGGACTACCA	65	1238	1.2
GM30	OP1_Abyss_2616	CTGCATATGCTGTGGATCGT	ACAACCCTTCCTCCCAACTT	65	1151	1.1
	OP2_CLC_107	ACAACCCTTCCTCCCAACTT	TGAGTTGCCTGATGATCTGC	66	1250	1.2
	OP4_APLG_31	GGGGTAACAGACGCATCAAT	CCTGGCAACTTTGAGGTTCT	66	1505	1.5
	OP5_APLG_0	CCTGGCAACTTTGAGGTTCT	CCGTATCCTGCGAAGATTGT	65	1420	1.4
	Hypothesized_OP5-OP1	CTGCATATGCTGTGGATCGT	CCGTATCCTGCGAAGATTGT	66	6100	6

*Approximate measure of PCR product length by gel electrophoresis and DNA ladder

Table 2.2: Verification of PCR products by Sanger sequencing.

Strain	Operon ID	Expected PCR Product Size (bp)	Sanger Verification* (bp)	
			3' end	5' end
CF80	OP1_APLG_101	6609	563	656
	OP2_APLG_102	6725	967	589
	OP3_APLG_01	6701	1270	941
GM41	OP1_APLG_Contig21	6009	1074	1292
	OP2_APLG_Contig22	6109	1051	1406
	OP3_APLG_Contig11	6035	1062	1010
	OP4_APLG_Contig12	6184	1026	1207
GM30	OP1_Abyss_2616	1151	NA	1044
	OP2_CLC_107	1250	NA	1130
	OP4_APLG_31	1505	1404	NA
	OP5_APLG_0	1420	1315	NA
	Hypothesized_OP5-OP1	6100	1032	1034
BT03	OP1_CLC_297	5469	317	995
	OP2_Abyss_10697	1404	1067	1027
	OP3_Abyss_10695	1208	1095	930
	OP4_Abyss_10696	1370	996	965
	OP5_Abyss_10833	1238	874	1029

*The ends of the PCR product were sequenced by Sanger method and number of bases with 100% identity on 3' and 5' end are shown.

Table 2.3: Summary of sequence data coverage.

NGS Technology	**Illumina PE**	**Illumina MP**	**Roche 454 SE**	**PacBio**
Avg. Read Length (bp)	100	150	565	5,456
BT03	240x*	24x	15x	18x
CF080	475x	41x	26x	20x
GM41	520x	46x	24x	32x
GM30	520x	36x	26x	NA

Note: *x defines raw read coverage value.

Table 2.4: Assembly summary information.

Strain	Library Type	No. of Contigs	Maximum Contig Size (kb)	N50 (kb)	Genome Size (Mb)	No. of scaffolds	Maximum Contig Size (kb)	N50 (kb)	Genome Size (Mb)	Software
CF080	454	71	1058	236	7.01	-	-	-	-	Newbler
	Pacbio-454	102	799	187	7.06	-	-	-	-	PBcR
	PE	1850	31	5	6.95	79	1098	321	7.07	SOAP
	PE	1426	692	312	7.96	1432	692	312	7.97	SPAdes
	PE	1039	335	75	7.54	897	631	383	7.56	CLC
	PE	856	76	13	6.96	56	1136	464	7.05	Velvet
	PE	270	302	69	7.04	263	308	79	7.04	MaSuRCA
	PE*	90	694	273	8.2	69	646	331	7.2	ABySS
	PE-454	57	1225	483	7.02	-	-	-	-	Newbler
	PE-MP	1372	1570	663	7.9	1369	2117	1894	7.95	SPAdes
	PE-MP	163	1413	597	7.12	103	4100	4100	7.21	MaSuRCA
	PE-MP*	40	1535	626	7.04	12	4813	4813	7.1	ALLPATHS-LG
	PE-MP-454	252	4095	4095	7.23	249	4095	4095	7.23	MaSuRCA
	PE-MP-454*	32	1341	615	7.01	-	-	-	-	Newbler
	PE-MP-454-Pacbio	-	-	-	-	6	4102	4102	7.04	AHA
	PE-MP-PacBio	25	2395	1779	7.04	23	2395	1844	7.04	SPAdes
	PE-MP-PacBio	**16**	**1885**	**671**	**7.04**	**5**	**4797**	**4797**	**7.05**	**ALLPATHS-LG**
GM41	454	112	236	89	6.61	-	-	-	-	Newbler
	Pacbio-454	80	371	140	6.79	-	-	-	-	PBcR
	PE	1162	41	9	6.56	95	475	152	6.62	SOAP
	PE	652	62	17	6.59	103	396	118	6.63	Velvet
	PE	212	271	85	6.64	204	271	86	6.64	MaSuRCA

Table 2.4 continued …

Strain	Library Type	No. of Contigs	Maximum Contig Size (kb)	N50 (kb)	Genome Size (Mb)	No. of scaffolds	Maximum Contig Size (kb)	N50 (kb)	Genome Size (Mb)	Software
GM41	PE	164	308	75	6.61	89	599	137	6.64	CLC
	PE	114	361	137	6.82	88	573	170	6.82	ABySS
	PE	101	436	165	6.64	96	679	183	6.64	SPAdes
	PE-454	96	345	143	6.63	-	-	-	-	Newbler
	PE-MP	157	621	279	6.7	117	2057	1560	6.71	MaSuRCA
	PE-MP	86	436	183	6.71	80	681	183	6.72	SPAdes
	PE-MP	62	415	107	6.65	5	3919	3919	6.72	ALLPATHS-LG
	PE-MP-454	696	1119	739	7.27	687	2486	1393	7.27	MaSuRCA
	PE-MP-454	66	345	159	6.62	-	-	-	-	Newbler
	PE-MP-454-Pacbio	-	-	-	-	17	1007	666	6.67	AHA
	PE-MP-PacBio	73	653	292	6.68	68	1070	292	6.69	SPAdes
	PE-MP-PacBio	**13**	**2562**	**1393**	**6.68**	**4**	**2835**	**2408**	**6.68**	**ALLPATHS-LG**
GM30	454	74	326	133	6.14	-	-	-	-	Newbler
	PE	1216	62	11	6.5	1192	62	11	6.61	MaSuRCA
	PE	1398	47	7	6.09	66	480	186	6.18	SOAP
	PE	773	83	13	6.11	138	313	96	6.15	Velvet
	PE	180	184	59	6.14	55	567	227	6.17	CLC
	PE	78	422	157	6.47	58	422	248	6.47	ABySS
	PE*	61	662	186	6.15	52	662	208	6.16	SPAdes
	PE-454	54	801	183	6.15	-	-	-	-	Newbler

Table 2.4 continued …

Strain	Library Type	No. of Contigs	Maximum Contig Size (kb)	N50 (kb)	Genome Size (Mb)	No. of scaffolds	Maximum Contig Size (kb)	N50 (kb)	Genome Size (Mb)	Software
GM30	PE-MP	570	214	62	6.38	403	2739	1607	6.41	MaSuRCA
	PE-MP	50	661	240	6.2	45	661	333	6.2	SPAdes
	PE-MP*	44	472	229	6.16	4	6208	6208	6.21	ALLPATHS-LG
	PE-MP-454	778	1986	448	6.7	765	2728	1149	6.71	MaSuRCA
	PE-MP-454*	**32**	**543**	**298**	**6.15**	**-**	**-**	**-**	**-**	**Newbler**
BT03	454	305	344	59	10.75	-	-	-	-	Newbler
	Pacbio-454	235	565	99	11.4	-	-	-	-	PBcR
	PE	3016	28	5	10.43	466	300	49	10.82	SOAP
	PE	2226	62	10	11.2	2201	63	11	11.21	MaSuRCA
	PE	1914	66	9	10.51	455	318	59	10.8	Velvet
	PE	690	155	29	10.64	422	295	63	10.77	CLC
	PE	475	243	47	11.55	403	243	52	11.55	ABySS
	PE*	397	363	80	10.82	386	363	85	10.83	SPAdes
	PE-454	315	344	70	10.82	-	-	-	-	Newbler
	PE-MP	806	240	59	10.95	457	1997	1161	11.04	MaSuRCA
	PE-MP	362	364	77	11.16	355	364	85	11.17	SPAdes
	PE-MP*	**135**	**562**	**177**	**10.91**	**22**	**2542**	**1282**	**11.11**	**ALLPATHS-LG**
	PE-MP-454	887	898	283	11.58	813	3314	2530	11.61	MaSuRCA
	PE-MP-454*	228	405	106	10.77	-	-	-	-	Newbler
	PE-MP-454-PacBio	-	-	-	-	55	1295	473	11.01	AHA

Table 2.4 continued …

Strain	Library Type	No. of Contigs	Maximum Contig Size (kb)	N50 (kb)	Genome Size (Mb)	No. of scaffolds	Maximum Contig Size (kb)	N50 (kb)	Genome Size (Mb)	Software
BT03	PE-MP-PacBio*	401	344	66	11.08	390	344	67	11.08	SPAdes
	PE-MP-PacBio	Program crashed: insufficient memory (132 GB)								ALLPATHS-LG

The best assembly for particular library type is shown by *
The best assembly for each strain is shown in bold.

Table 2.5: Summary of *de novo* and hybrid assembly results.

Strain	Library Type	No. of Contigs	Max. Contig Size (kb)	N50 (kb)	Genome Size (Mb)	No. of scaffolds	Max Scaffold Size (kb)	N50 (kb)	Genome Size (Mb)	Software
CF080	PE	1,039	335	75	7.54	897	631	383	7.56	CLC
	PE*	90	694	237	8.20	69	646	331	7.20	ABySS
	454	71	1,058	236	7.01	-	-	-	-	Newbler
	Pacbio-454	102	799	187	7.06	-	-	-	-	PBcR
	PE-454	57	1,225	483	7.02	-	-	-	-	Newbler
	PE-MP	163	1,413	597	7.12	103	4,100	4,100	7.21	MaSuRCA
	PE-MP*	40	1,535	626	7.04	12	4,813	4,813	7.10	APLG
	PE-MP-454	252	4,095	4,095	7.23	249	4,095	4,095	7.23	MaSuRCA
	PE-MP-454*	32	1,341	615	7.01	-	-	-	-	Newbler
	PE-MP-454-Pacbio	-	-	-	-	6	4,102	4,102	7.04	AHA
	PE-MP-PacBio	25	2,395	1,779	7.04	23	2,395	1,844	7.04	SPAdes
	PE-MP-PacBio	**16**	**1,885**	**671**	**7.04**	**5**	**4,797**	**4,797**	**7.05**	**APLG**
GM41	PE	164	308	75	6.61	89	599	137	6.64	CLC
	PE*	101	436	165	6.64	96	679	183	6.64	SPAdes
	454	112	236	89	6.61	-	-	-	-	Newbler
	Pacbio-454	80	371	140	6.79	-	-	-	-	PBcR
	PE-454	96	345	143	6.63	-	-	-	-	Newbler
	PE-MP	157	621	279	6.70	117	2,057	1,560	6.71	MaSuRCA
	PE-MP	86	436	183	6.71	80	681	183	6.72	SPAdes
	PE-MP*	62	415	107	6.65	5	3,919	3,919	6.72	APLG
	PE-MP-454	66	345	159	6.62	-	-	-	-	Newbler
	PE-MP-454-Pacbio	-	-	-	-	17	1,007	666	6.67	AHA

Table 2.5 continued …

Strain	Library Type	No. of Contigs	Max. Contig Size (kb)	N50 (kb)	Genome Size (Mb)	No. of scaffolds	Max Scaffold Size (kb)	N50 (kb)	Genome Size (Mb)	Software
GM41	PE-MP-PacBio	73	653	292	6.68	68	1,070	292	6.69	SPAdes
	PE-MP-PacBio*	**13**	**2,562**	**1,393**	**6.68**	**4**	**2,835**	**2,408**	**6.68**	**APLG**
GM30	PE	180	184	59	6.14	55	567	227	6.17	CLC
	PE*	61	662	186	6.15	52	662	208	6.16	SPAdes
	454	74	326	133	6.14	-	-	-	-	Newbler
	PE-454	54	801	183	6.15	-	-	-	-	Newbler
	PE-MP	50	661	240	6.20	45	661	333	6.20	SPAdes
	PE-MP*	44	472	229	6.16	4	6,208	6,208	6.21	APLG
	PE-MP-454	**32**	**543**	**298**	**6.15**	**-**	**-**	**-**	**-**	**Newbler**
BT03	PE	690	155	29	10.64	422	295	63	10.77	CLC
	PE*	397	363	80	10.82	386	363	85	10.83	SPAdes
	454	305	344	59	10.75	-	-	-	-	Newbler
	Pacbio-454	235	565	99	11.40	-	-	-	-	PBcR
	PE-454	315	344	70	10.82	-	-	-	-	Newbler
	PE-MP	806	240	59	10.95	457	1,997	1,161	11.04	MaSuRCA
	PE-MP	362	364	77	11.16	355	364	85	11.17	SPAdes
	PE-MP*	**135**	**562**	**177**	**10.91**	**22**	**2,542**	**1,282**	**11.11**	**APLG**

*defines the optimal assembly statistics for particular combination of library types as assembled by more than one assembler. The best assembly is shown in bold. Note: The hybrid assembly statistics which were worse than the PE assemblies are not included in above table. The complete table of *de novo* and hybrid assemblies is available through Table 2.4. ALLPATHS-LG is denoted with abbreviation APLG.

Table 2.6: Contig assembly statistics for 43 bacterial isolates using Velvet, ABySS, CLC Genomics workbench and SOAPdenovo software.

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
		Velvet						
Caulobacter sp.	AP07	41	1,689	500	22,408	3,150	4,566	5,320,299
Novosphingobium sp.	AP12	41	1,006	501	62,126	5,446	8,601	5,478,939
Rhizobium sp.	AP16	49	873	503	53,860	7,392	12,723	6,452,889
Sphingobium sp.	AP49	41	759	501	34,479	5,805	9,368	4,405,728
Brevibacillus sp.	BC25	49	136	571	230,556	46,140	78,080	6,275,075
Burkholderia sp.	BT03	41	1,914	500	66,459	5,500	9,249	10,526,374
Brevibacillus sp.	CF112	49	496	508	54,216	10,603	18,273	5,258,996
Rhizobium sp.	CF122	41	686	510	105,996	8,904	15,611	6,107,869
Flavobacterium sp.	CF136	57	1,186	501	45,195	4,971	7,928	5,895,148
Rhizobium sp.	CF142	41	925	502	63,630	7,994	13,626	7,394,842
Variovorax sp.	CF313	75	71	502	461,192	72,295	201,533	5,132,934
Chryseobacterium sp.	CF314	57	117	518	351,655	38,327	75,113	4,484,205
Acidovorax sp.	CF316	41	1,618	500	33,505	4,288	6,733	6,938,713
Polaromonas sp.	CF318	49	936	504	44,638	5,243	8,483	4,907,357
Herbaspirillum sp.	CF444	49	817	522	62,084	6,760	11,250	5,522,914
Rhizobium sp.	CF080	41	856	524	76,858	8,131	13,521	6,960,107
Pantoea sp.	GM01	49	456	510	105,947	11,563	20,872	5,272,558
Pseudomonas sp.	GM102	49	746	513	95,608	8,910	14,769	6,647,014
Pseudomonas sp.	GM16	49	705	500	51,311	9,230	15,315	6,507,489

Table 2.6 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Pseudomonas sp.	GM17	41	1,220	504	46,467	5,439	9,154	6,635,120
Pseudomonas sp.	GM18	49	613	510	65,690	10,226	17,520	6,268,774
Pseudomonas sp.	GM21	49	689	505	65,733	9,578	15,494	6,599,572
Pseudomonas sp.	GM24	57	636	506	66,811	10,263	17,348	6,527,289
Pseudomonas sp.	GM25	49	630	502	64,322	10,022	17,071	6,313,646
Pseudomonas sp.	GM30	49	773	500	83,996	7,905	13,514	6,110,842
Pseudomonas sp.	GM33	49	777	511	53,645	8,643	14,672	6,715,907
Pseudomonas sp.	GM41	49	652	523	62,096	10,122	17,113	6,599,607
Pseudomonas sp.	GM48	49	815	500	66,446	7,887	12,879	6,428,219
Pseudomonas sp.	GM49	49	565	506	71,300	11,643	18,354	6,578,159
Pseudomonas sp.	GM50	49	728	505	106,491	9,172	16,183	6,677,337
Pseudomonas sp.	GM55	49	652	515	86,970	9,929	17,642	6,473,642
Pseudomonas sp.	GM60	49	671	556	70,112	9,532	15,652	6,395,730
Pseudomonas sp.	GM67	49	680	522	84,859	9,533	14,869	6,482,653
Pseudomonas sp.	GM74	49	639	502	84,655	9,527	16,515	6,087,575
Pseudomonas sp.	GM78	41	885	512	58,818	8,166	13,765	7,226,657
Pseudomonas sp.	GM79	49	731	508	63,034	9,152	15,367	6,690,249
Pseudomonas sp.	GM80	49	813	501	133,427	8,316	13,328	6,761,036
Pseudomonas sp.	GM84	41	1,236	503	37,214	4,633	7,103	5,726,559
Rhizobium	PD01-76	49	752	500	55,347	7,318	13,136	5,503,410
Pantoea sp.	YR343	49	416	510	99,520	12,695	25,415	5,280,929
Herbaspirillum sp.	YR522	41	1,006	507	53,621	4,963	7,920	4,992,862
Phyllobacterium sp.	YR531	49	227	501	172,663	21,852	42,233	4,960,454

Table 2.6 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Bradyrhizobium sp.	YR681	31	1,586	502	35,063	4,844	7,646	7,682,095
		ABySS						
Caulobacter sp.	AP07	57	241	533	186,748	25,090	39,997	6,046,599
Novosphingobium sp.	AP12	57	141	574	284,966	44,887	92,739	6,329,017
Rhizobium sp.	AP16	57	59	1,025	767,234	123,351	270,199	7,277,701
Sphingobium sp.	AP49	57	50	770	516,401	95,447	168,966	4,772,372
Brevibacillus sp.	BC25	57	61	693	522,270	103,518	223,120	6,314,628
Burkholderia sp.	BT03	57	475	523	243,098	24,322	47,979	11,553,060
Brevibacillus sp.	CF112	57	122	552	230,851	44,528	99,678	5,432,476
Rhizobium sp.	CF122	57	93	824	304,677	70,620	157,170	6,567,690
Flavobacterium sp.	CF136	57	91	558	480,295	67,376	160,363	6,131,244
Rhizobium sp.	CF142	57	73	602	786,951	117,135	236,026	8,550,850
Variovorax sp.	CF313	63	55	612	346,595	94,529	208,316	5,199,120
Chryseobacterium sp.	CF314	49	95	517	355,809	48,580	93,280	4,615,129
Acidovorax sp.	CF316	31	205	613	154,112	37,669	67,051	7,722,173
Polaromonas sp.	CF318	57	87	519	337,727	62,559	108,102	5,442,614
Herbaspirillum sp.	CF444	57	69	1,136	504,862	85,220	193,959	5,880,205
Rhizobium sp.	CF080	57	90	526	694,002	92,218	237,871	8,299,640

Table 2.6 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Pantoea sp.	GM01	57	73	1,408	455,724	76,965	145,513	5,618,464
Pseudomonas sp.	GM102	57	117	504	496,363	59,188	112,063	6,924,991
Pseudomonas sp.	GM16	63	77	567	393,982	86,155	181,954	6,633,898
Pseudomonas sp.	GM17	57	137	995	523,299	51,147	100,814	7,007,117
Pseudomonas sp.	GM18	63	81	509	602,721	83,689	171,636	6,778,836
Pseudomonas sp.	GM21	57	126	506	439,915	55,093	104,974	6,941,718
Pseudomonas sp.	GM24	57	95	563	352,654	69,508	135,502	6,603,262
Pseudomonas sp.	GM25	57	56	514	911,574	117,469	241,133	6,578,263
Pseudomonas sp.	GM30	57	78	661	422,129	83,041	157,555	6,477,189
Pseudomonas sp.	GM33	57	136	522	452,080	51,991	97,545	7,070,778
Pseudomonas sp.	GM41	63	114	763	361,917	59,854	137,689	6,823,370
Pseudomonas sp.	GM48	57	151	511	338,250	44,126	92,451	6,663,082
Pseudomonas sp.	GM49	57	91	528	400,857	76,509	165,998	6,962,330
Pseudomonas sp.	GM50	63	89	581	425,673	78,275	163,367	6,966,496
Pseudomonas sp.	GM55	57	131	587	323,089	53,585	124,730	7,019,643
Pseudomonas sp.	GM60	57	128	588	307,522	52,835	97,658	6,762,823
Pseudomonas sp.	GM67	57	119	509	414,863	56,646	121,651	6,740,829

Table 2.6 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Pseudomonas sp.	GM74	57	133	612	415,216	48,820	118,235	6,493,077
Pseudomonas sp.	GM78	63	150	629	293,769	51,343	91,586	7,701,459
Pseudomonas sp.	GM79	57	83	545	666,065	86,544	178,557	7,183,187
Pseudomonas sp.	GM80	63	125	514	346,700	57,250	117,276	7,156,294
Pseudomonas sp.	GM84	57	137	779	236,705	44,320	84,710	6,071,779
Rhizobium	PD01-76	57	179	503	514,647	33,124	182,069	5,929,227
Pantoea sp.	YR343	63	60	1,634	474,976	89,993	205,970	5,399,574
Herbaspirillum sp.	YR522	57	106	531	361,407	52,373	108,348	5,551,529
Phyllobacterium sp.	YR531	57	36	565	972,540	153,378	311,456	5,521,624
Bradyrhizobium sp.	YR681	57	172	1,156	417,818	46,561	73,080	8,008,436
			CLC Genomics Workbench					
Caulobacter sp.	AP07	NA	327	508	189,850	17,177	30,563	5,617,011
Novosphingobium sp.	AP12	NA	187	528	306,606	30,009	54,713	5,611,617
Rhizobium sp.	AP16	NA	96	807	663,744	67,684	123,519	6,497,619
Sphingobium sp.	AP49	NA	103	516	337,149	43,509	89,526	4,481,471
Brevibacillus sp.	BC25	NA	374	512	703,074	17,299	141,920	6,469,833
Burkholderia sp.	BT03	NA	690	502	155,013	15,426	29,868	10,643,821
Brevibacillus sp.	CF112	NA	174	538	227,827	30,306	76,367	5,273,255

Table 2.6 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Rhizobium sp.	CF122	NA	130	507	295,274	47,248	117,778	6,142,299
Flavobacterium sp.	CF136	NA	437	509	179,457	13,613	45,171	5,948,936
Rhizobium sp.	CF142	NA	146	568	348,002	51,065	85,172	7,455,438
Variovorax sp.	CF313	NA	82	541	307,933	62,686	160,731	5,140,221
Chryseobacterium sp.	CF314	NA	119	518	336,918	37,686	80,113	4,484,672
Acidovorax sp.	CF316	NA	317	539	118,161	22,359	37,104	7,087,943
Polaromonas sp.	CF318	NA	159	555	216,842	31,502	61,518	5,008,816
Herbaspirillum sp.	CF444	NA	125	508	339,755	44,758	82,125	5,594,732
Rhizobium sp.	CF080	NA	1,039	500	335,740	7,258	75,530	7,540,864
Pantoea sp.	GM01	NA	102	587	259,669	52,162	91,591	5,320,548
Pseudomonas sp.	GM102	NA	159	517	192,171	41,870	88,165	6,657,346
Pseudomonas sp.	GM16	NA	127	583	278,410	51,580	122,673	6,550,699
Pseudomonas sp.	GM17	NA	279	504	230,552	24,326	44,590	6,786,856
Pseudomonas sp.	GM18	NA	139	598	240,258	45,304	106,048	6,297,187
Pseudomonas sp.	GM21	NA	212	527	185,369	31,189	57,752	6,612,109
Pseudomonas sp.	GM24	NA	399	502	107,081	16,335	32,656	6,517,673
Pseudomonas sp.	GM25	NA	91	535	365,551	69,787	137,130	6,350,607
Pseudomonas sp.	GM30	NA	180	527	184,453	34,116	59,627	6,140,967
Pseudomonas sp.	GM33	NA	205	518	207,002	32,816	61,913	6,727,223
Pseudomonas sp.	GM41	NA	164	538	308,020	40,338	75,073	6,615,479
Pseudomonas sp.	GM48	NA	201	500	171,623	32,062	59,542	6,444,521

Table 2.6 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Pseudomonas sp.	GM49	NA	345	530	143,192	19,101	31,212	6,589,890
Pseudomonas sp.	GM50	NA	155	535	378,902	43,175	68,220	6,692,143
Pseudomonas sp.	GM55	NA	165	501	299,293	39,335	77,637	6,490,356
Pseudomonas sp.	GM60	NA	181	618	226,607	35,493	62,804	6,424,244
Pseudomonas sp.	GM67	NA	183	545	293,542	35,531	68,050	6,502,113
Pseudomonas sp.	GM74	NA	181	505	294,898	33,728	75,201	6,104,807
Pseudomonas sp.	GM78	NA	235	611	204,409	31,011	57,174	7,287,561
Pseudomonas sp.	GM79	NA	128	518	274,135	52,407	96,213	6,708,073
Pseudomonas sp.	GM80	NA	284	504	188,591	23,899	39,805	6,787,457
Pseudomonas sp.	GM84	NA	387	541	107,668	15,040	24,795	5,820,528
Rhizobium	PD01-76	NA	256	508	295,679	21,661	102,115	5,545,093
Pantoea sp.	YR343	NA	128	729	269,790	41,516	94,033	5,314,049
Herbaspirillum sp.	YR522	NA	176	504	167,864	29,096	54,530	5,120,913
Phyllobacterium sp.	YR531	NA	42	606	811,600	118,998	257,533	4,997,930
Bradyrhizobium sp.	YR681	NA	351	510	167,815	22,313	37,719	7,831,714
		SOAPdenovo						
Caulobacter sp.	AP07	49	3,131	500	12,801	1,647	2,061	5,157,296
Novosphingobium sp.	AP12	49	1,855	500	18,613	2,946	4,431	5,465,077
Rhizobium sp.	AP16	57	1,612	502	53,869	4,008	6,456	6,460,422
Sphingobium sp.	AP49	57	1,562	500	45,324	2,814	3,914	4,394,732
Brevibacillus sp.	BC25	63	364	514	108,427	17,241	30,095	6,275,825
Burkholderia sp.	BT03	49	3,016	500	28,827	3,461	5,422	10,437,045

Table 2.6 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Brevibacillus sp.	CF112	57	985	505	40,083	5,315	8,405	5,235,183
Rhizobium sp.	CF122	57	1,291	500	41,365	4,733	7,671	6,110,473
Flavobacterium sp.	CF136	63	4040	500	7,713	1,216	1,388	4,912,125
Rhizobium sp.	CF142	49	1,746	504	33,605	4,236	6,594	7,396,623
Variovorax sp.	CF313	63	2200	500	15,038	2,222	3,062	4,888,628
Chryseobacterium sp.	CF314	63	663	501	45,749	6,709	11,007	4,447,853
Acidovorax sp.	CF316	49	3,024	500	14,940	2,255	3,089	6,817,862
Polaromonas sp.	CF318	57	1,727	500	22,892	2,843	4,236	4,909,738
Herbaspirillum sp.	CF444	57	1,447	504	37,444	3,828	6,100	5,539,077
Rhizobium sp.	CF080	57	1,850	501	31,770	3,761	5,825	6,958,517
Pantoea sp.	GM01	57	944	501	42,832	5,588	9,194	5,274,977
Pseudomonas sp.	GM102	57	1,332	501	39,591	4,963	7,991	6,610,435
Pseudomonas sp.	GM16	63	2019	500	21,966	3,172	4,517	6,403,496
Pseudomonas sp.	GM17	49	2,221	501	26,765	2,976	4,452	6,610,221
Pseudomonas sp.	GM18	57	1,184	515	36,600	5,281	8,800	6,252,562
Pseudomonas sp.	GM21	57	1,248	501	37,544	5,257	8,679	6,560,611
Pseudomonas sp.	GM24	63	4537	500	7,704	1,188	1,350	5,389,425
Pseudomonas sp.	GM25	57	1,330	502	37,980	4,734	7,516	6,296,252
Pseudomonas sp.	GM30	57	1,398	505	47,110	4,357	7,008	6,091,661
Pseudomonas sp.	GM33	57	1,338	507	44,498	4,981	8,295	6,664,267
Pseudomonas sp.	GM41	57	1,162	501	41,159	5,651	9,344	6,566,670
Pseudomonas sp.	GM48	57	1,308	500	38,599	4,897	7,975	6,405,061
Pseudomonas sp.	GM49	63	1,693	500	41,806	3,891	6,178	6,588,162

Table 2.6 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Pseudomonas sp.	GM50	57	1,306	501	38,695	5,096	8,094	6,655,832
Pseudomonas sp.	GM55	57	1,281	508	38,645	5,025	8,120	6,437,451
Pseudomonas sp.	GM60	57	1,178	503	38,629	5,428	8,729	6,394,249
Pseudomonas sp.	GM67	57	1,207	511	49,037	5,361	8,689	6,471,307
Pseudomonas sp.	GM74	57	1,137	502	40,592	5,320	8,626	6,048,706
Pseudomonas sp.	GM78	57	1,581	501	43,580	4,569	7,197	7,223,766
Pseudomonas sp.	GM79	57	1,218	501	55,375	5,478	8,820	6,672,318
Pseudomonas sp.	GM80	57	1,418	501	67,048	4,742	7,527	6,724,232
Pseudomonas sp.	GM84	49	1,567	504	25,203	3,658	5,595	5,732,251
Rhizobium	PD01-76	57	1,282	500	37,730	4,279	6,978	5,485,528
Pantoea sp.	YR343	57	994	501	40,615	5,318	9,229	5,285,700
Herbaspirillum sp.	YR522	57	1,975	500	26,915	2,521	3,592	4,979,313
Phyllobacterium sp.	YR531	63	566	516	85,043	8,811	14,859	4,986,995
Bradyrhizobium sp.	YR681	49	2,984	500	22,376	2,564	3,631	7,651,346

Table 2.7: Scaffolds assembly statistics for 43 bacterial isolates using Velvet, ABySS, CLC Genomics workbench and SOAPdenovo software.

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
		Velvet						
Caulobacter sp.	AP07	41	715	505	61,420	7,624	14,243	5,451,508
Novosphingobium sp.	AP12	31	326	508	205,255	17,082	42,008	5,568,836
Rhizobium sp.	AP16	57	98	554	534,177	66,724	159,968	6,538,929
Sphingobium sp.	AP49	49	44	533	735,458	102,405	298,727	4,505,816
Brevibacillus sp.	BC25	57	40	575	720,381	157,126	422,272	6,285,031
Burkholderia sp.	BT03	49	455	502	318,111	23,749	59,949	10,805,990
Brevibacillus sp.	CF112	45	141	500	351,379	37,521	103,237	5,290,448
Rhizobium sp.	CF122	49	76	527	357,645	81,155	214,828	6,167,779
Flavobacterium sp.	CF136	57	107	584	339,245	61,448	124,928	6,574,916
Rhizobium sp.	CF142	49	66	512	795,765	113,639	353,940	7,500,182
Variovorax sp.	CF313	49	431	509	214,704	13,908	32,365	5,994,309
Chryseobacterium sp.	CF314	57	82	518	472,758	54,694	93,808	4,484,913
Acidovorax sp.	CF316	41	138	543	490,281	51,646	114,740	7,127,163
Polaromonas sp.	CF318	41	49	648	489,649	102,508	212,676	5,022,876
Herbaspirillum sp.	CF444	49	151	501	500,836	36,971	85,272	5,582,656
Rhizobium sp.	CF080	49	56	500	1,136,365	126,003	464,592	7,056,194
Pantoea sp.	GM01	49	57	557	532,275	93,571	237,356	5,333,539
Pseudomonas sp.	GM102	61	120	541	355,536	55,946	120,132	6,713,517
Pseudomonas sp.	GM16	57	49	522	346,297	104,725	219,837	5,131,527
Pseudomonas sp.	GM17	49	253	536	176,633	26,686	55,840	6,751,582

Table 2.7 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Pseudomonas sp.	GM18	57	92	756	304,103	68,698	133,290	6,320,203
Pseudomonas sp.	GM21	57	171	532	279,583	38,860	91,322	6,645,128
Pseudomonas sp.	GM24	49	165	561	307,359	39,788	105,909	6,564,965
Pseudomonas sp.	GM25	49	67	550	832,179	94,655	208,441	6,341,879
Pseudomonas sp.	GM30	49	138	509	313,313	44,598	96,800	6,154,560
Pseudomonas sp.	GM33	57	169	551	330,332	40,119	89,433	6,780,123
Pseudomonas sp.	GM41	49	103	520	396,802	64,390	118,163	6,632,201
Pseudomonas sp.	GM48	57	117	544	360,072	55,457	108,942	6,488,496
Pseudomonas sp.	GM49	49	63	691	576,698	104,908	230,059	6,609,191
Pseudomonas sp.	GM50	57	119	520	321,253	56,566	118,200	6,731,363
Pseudomonas sp.	GM55	57	124	640	316,655	52,715	101,416	6,536,619
Pseudomonas sp.	GM60	57	143	647	242,126	45,165	86,177	6,458,561
Pseudomonas sp.	GM67	49	135	638	386,830	48,304	117,106	6,521,055
Pseudomonas sp.	GM74	49	134	578	293,548	45,698	119,963	6,123,559
Pseudomonas sp.	GM78	49	163	518	364,632	44,890	95,619	7,317,114
Pseudomonas sp.	GM79	57	83	508	384,398	81,287	195,050	6,746,859
Pseudomonas sp.	GM80	49	125	533	287,018	54,375	114,588	6,796,864
Pseudomonas sp.	GM84	49	174	537	186,538	33,605	64,520	5,847,216
Rhizobium	PD01-076	57	226	503	461,930	24,657	143,182	5,572,551
Pantoea sp.	YR343	41	45	620	933,359	118,342	304,427	5,325,373
Herbaspirillum sp.	YR522	41	239	502	218,167	21,261	42,572	5,081,315
Phyllobacterium sp.	YR531	61	30	878	1,427,269	166,053	506,356	4,981,589
Bradyrhizobium sp.	YR681	41	232	601	224,271	33,816	63,110	7,845,214

Table 2.7 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
		ABySS						
Caulobacter sp.	AP07	57	177	541	327,072	34,201	61,515	6,053,580
Novosphingobium sp.	AP12	57	110	574	532,529	57,567	156,490	6,332,376
Rhizobium sp.	AP16	63	42	1,034	914,279	157,954	470,169	6,634,053
Sphingobium sp.	AP49	57	39	770	516,401	122,405	194,467	4,773,782
Brevibacillus sp.	BC25	57	43	996	522,270	146,881	240,893	6,315,881
Burkholderia sp.	BT03	57	403	523	243,098	28,681	52,557	11,558,366
Brevibacillus sp.	CF112	57	110	552	354,019	49,401	104,397	5,434,086
Rhizobium sp.	CF122	57	75	912	369,775	87,612	181,428	6,570,886
Flavobacterium sp.	CF136	63	49	612	477,713	106,063	212,604	5,197,086
Rhizobium sp.	CF142	31	85	722	802,953	93,621	195,289	7,957,802
Variovorax sp.	CF313	57	43	600	476,691	120,728	242,932	5,191,313
Chryseobacterium sp.	CF314	63	81	848	354,388	57,692	103,405	4,673,029
Acidovorax sp.	CF316	57	150	555	308,895	50,610	84,313	7,591,440
Polaromonas sp.	CF318	57	52	1,446	358,225	104,774	194,399	5,448,271
Herbaspirillum sp.	CF444	57	38	1,136	964,863	155,053	313,542	5,892,008
Rhizobium sp.	CF080	63	69	660	646,070	104,360	331,111	7,200,836
Pantoea sp.	GM01	57	52	1,408	699,295	108,082	212,136	5,620,269
Pseudomonas sp.	GM102	57	89	547	496,363	77,851	173,802	6,928,704
Pseudomonas sp.	GM16	63	55	567	600,353	120,633	239,576	6,634,789
Pseudomonas sp.	GM17	57	86	1,403	1,021,488	81,548	182,112	7,013,096
Pseudomonas sp.	GM18	63	62	1,031	708,575	109,397	226,680	6,782,584
Pseudomonas sp.	GM21	57	104	506	439,915	66,771	152,458	6,944,215

Table 2.7 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Pseudomonas sp.	GM24	49	62	555	568,271	105,944	162,319	6,568,537
Pseudomonas sp.	GM25	57	41	514	1,345,329	160,465	276,562	6,579,071
Pseudomonas sp.	GM30	57	58	677	422,129	111,714	248,988	6,479,437
Pseudomonas sp.	GM33	57	110	522	452,080	64,316	121,447	7,074,795
Pseudomonas sp.	GM41	63	88	763	573,771	77,565	170,846	6,825,720
Pseudomonas sp.	GM48	57	99	767	442,406	67,357	113,454	6,668,311
Pseudomonas sp.	GM49	57	74	659	487,490	94,141	183,841	6,966,412
Pseudomonas sp.	GM50	63	66	581	520,774	105,635	206,676	6,971,878
Pseudomonas sp.	GM55	63	101	559	421,943	65,524	128,754	6,617,934
Pseudomonas sp.	GM60	57	104	588	311,833	65,051	122,013	6,765,284
Pseudomonas sp.	GM67	57	93	509	414,863	72,539	130,140	6,746,163
Pseudomonas sp.	GM74	57	90	917	573,721	72,201	171,253	6,498,094
Pseudomonas sp.	GM78	57	122	542	498,020	62,107	109,083	7,577,074
Pseudomonas sp.	GM79	57	62	557	735,936	115,914	258,612	7,186,691
Pseudomonas sp.	GM80	57	85	525	605,860	84,401	168,777	7,174,116
Pseudomonas sp.	GM84	57	100	1,331	326,764	60,788	111,956	6,078,810
Rhizobium	PD01-076	57	166	503	514,647	35,741	198,409	5,932,978
Pantoea sp.	YR343	57	51	1,235	636,236	108,670	330,803	5,542,158
Herbaspirillum sp.	YR522	63	74	602	361,356	70,238	128,011	5,197,606
Phyllobacterium sp.	YR531	31	33	4,502	694,207	155,883	278,734	5,144,144
Bradyrhizobium sp.	YR681	57	86	1,850	417,818	93,287	147,912	8,022,655
			CLC Genomics Workbench					
Caulobacter sp.	AP07	NA	167	500	290,781	33,991	89,015	5,676,427

Table 2.7 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Novosphingobium sp.	AP12	NA	99	513	500,282	56,968	152,392	5,639,869
Rhizobium sp.	AP16	NA	39	754	1,080,689	166,892	687,301	6,508,798
Sphingobium sp.	AP49	NA	618	500	734,378	7,910	205,795	4,888,534
Brevibacillus sp.	BC25	NA	501	500	703,074	13,121	350,312	6,573,456
Burkholderia sp.	BT03	NA	422	501	295,968	25,528	63,281	10,772,920
Brevibacillus sp.	CF112	NA	125	507	351,081	42,356	112,935	5,294,438
Rhizobium sp.	CF122	NA	218	500	492,692	28,623	208,861	6,239,840
Flavobacterium sp.	CF136	NA	35	563	556,591	146,446	304,122	5,125,612
Rhizobium sp.	CF142	NA	65	510	860,250	114,966	358,503	7,472,805
Variovorax sp.	CF313	NA	69	514	928,141	87,552	284,806	6,041,086
Chryseobacterium sp.	CF314	NA	97	501	353,678	46,276	103,298	4,488,768
Acidovorax sp.	CF316	NA	221	501	253,351	32,577	101,160	7,199,445
Polaromonas sp.	CF318	NA	76	500	597,150	66,497	346,983	5,053,736
Herbaspirillum sp.	CF444	NA	40	589	1,279,346	140,176	458,649	5,607,044
Rhizobium sp.	CF080	NA	895	500	885,602	8,459	433,018	7,571,170
Pantoea sp.	GM01	NA	60	557	530,090	88,835	199,845	5,330,074
Pseudomonas sp.	GM102	NA	86	500	448,747	77,674	174,958	6,679,934
Pseudomonas sp.	GM16	NA	71	508	667,065	92,217	253,931	6,547,404
Pseudomonas sp.	GM17	NA	115	534	542,669	59,337	165,870	6,823,802
Pseudomonas sp.	GM18	NA	41	529	543,685	154,118	259,667	6,318,832
Pseudomonas sp.	GM21	NA	122	532	361,937	54,391	119,518	6,635,702
Pseudomonas sp.	GM24	NA	52	517	667,079	126,117	291,362	6,558,064
Pseudomonas sp.	GM25	NA	45	617	768,371	140,895	323,935	6,340,264

Table 2.7 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Pseudomonas sp.	GM30	NA	55	960	567,708	112,243	227,402	6,173,340
Pseudomonas sp.	GM33	NA	114	521	452,669	59,234	138,719	6,752,674
Pseudomonas sp.	GM41	NA	89	506	599,076	74,627	137,048	6,641,788
Pseudomonas sp.	GM48	NA	106	533	400,793	61,138	124,555	6,480,638
Pseudomonas sp.	GM49	NA	2,004	500	699,496	4,176	139,187	8,369,038
Pseudomonas sp.	GM50	NA	73	502	715,550	92,180	242,150	6,729,172
Pseudomonas sp.	GM55	NA	96	544	511,865	67,948	181,053	6,522,991
Pseudomonas sp.	GM60	NA	87	612	467,119	74,134	165,635	6,449,652
Pseudomonas sp.	GM67	NA	96	703	398,455	67,974	124,788	6,525,512
Pseudomonas sp.	GM74	NA	118	521	517,300	51,894	104,993	6,123,490
Pseudomonas sp.	GM78	NA	128	596	613,841	57,051	114,116	7,302,512
Pseudomonas sp.	GM79	NA	40	500	892,527	168,234	367,233	6,729,350
Pseudomonas sp.	GM80	NA	84	525	794,443	80,950	195,939	6,799,840
Pseudomonas sp.	GM84	NA	132	729	217,999	44,329	81,577	5,851,422
Rhizobium	PD01-076	NA	172	512	606,181	32,352	231,197	5,564,522
Pantoea sp.	YR343	NA	42	656	727,795	126,934	306,992	5,331,240
Herbaspirillum sp.	YR522	NA	55	504	486,322	93,571	281,043	5,146,416
Phyllobacterium sp.	YR531	NA	26	606	1,536,862	192,467	508,475	5,004,140
Bradyrhizobium sp.	YR681	NA	66	517	1,197,401	119,591	239,572	7,893,028
		SOAPdenovo						
Caulobacter sp.	AP07	49	294	547	139,530	19,623	32,588	5,769,209
Novosphingobium sp.	AP12	49	114	534	608,242	49,620	142,179	5,656,737

Table 2.7 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Rhizobium sp.	AP16	57	49	512	1,002,627	133,766	410,074	6,554,516
Sphingobium sp.	AP49	57	68	509	565,819	66,862	143,086	4,546,647
Brevibacillus sp.	BC25	63	47	604	705,320	133,865	300,307	6,291,644
Burkholderia sp.	BT03	49	466	501	300,751	23,225	49,079	10,822,969
Brevibacillus sp.	CF112	57	121	519	351,309	43,491	98,941	5,262,392
Rhizobium sp.	CF122	57	102	502	366,514	60,026	184,136	6,122,693
Flavobacterium sp.	CF136	63	743	508	91318	7972	14326	5,923,309
Rhizobium sp.	CF142	49	86	512	514,173	87,473	253,494	7,522,708
Variovorax sp.	CF313	63	109	599	294550	46007	95849	5,014,732
Chryseobacterium sp.	CF314	57	85	512	234,465	51,861	91,752	4,408,217
Acidovorax sp.	CF316	49	206	522	190,678	35,253	57,575	7,262,170
Polaromonas sp.	CF318	57	79	721	503,964	63,774	131,032	5,038,176
Herbaspirillum sp.	CF444	57	51	582	605,110	109,359	279,386	5,577,321
Rhizobium sp.	CF080	57	79	501	1,098,065	89,611	321,890	7,079,234
Pantoea sp.	GM01	57	71	502	529,187	74,579	164,866	5,295,134
Pseudomonas sp.	GM102	57	97	501	438,882	68,627	154,277	6,656,850
Pseudomonas sp.	GM16	63	94	698	451371	69436	126158	6,526,971
Pseudomonas sp.	GM17	49	112	504	337,764	61,106	130,872	6,843,908
Pseudomonas sp.	GM18	57	63	557	476,862	100,137	282,922	6,308,637
Pseudomonas sp.	GM21	63	122	512	410,450	54,424	135,370	6,639,681
Pseudomonas sp.	GM24	63	720	506	76585	8959	13738	6,450,431
Pseudomonas sp.	GM25	57	44	617	955,687	144,238	309,879	6,346,476
Pseudomonas sp.	GM30	57	66	584	480,407	93,643	186,840	6,180,424

Table 2.7 continued …

Organism	Strain	KMER	Contigs	Minimum Contig Length (bp)	Maximum Contig Length (bp)	Average Length (bp)	N50 (bp)	Genome Size (bp)
Pseudomonas sp.	GM33	57	134	528	452,849	50,260	101,079	6,734,794
Pseudomonas sp.	GM41	57	95	566	475,934	69,700	152,241	6,621,510
Pseudomonas sp.	GM48	57	101	532	444,254	63,954	113,744	6,459,360
Pseudomonas sp.	GM49	57	602	500	353,875	11,676	102,643	7,028,751
Pseudomonas sp.	GM50	57	83	568	425,910	80,959	186,041	6,719,622
Pseudomonas sp.	GM55	57	107	553	499,085	60,865	129,162	6,512,544
Pseudomonas sp.	GM60	57	110	571	455,601	58,842	108,014	6,472,586
Pseudomonas sp.	GM67	57	102	600	790,372	64,154	140,928	6,543,690
Pseudomonas sp.	GM74	57	93	537	455,315	65,606	173,296	6,101,342
Pseudomonas sp.	GM78	57	124	525	613,332	58,693	119,064	7,277,951
Pseudomonas sp.	GM79	57	45	661	1,012,918	149,300	328,656	6,718,509
Pseudomonas sp.	GM80	57	87	561	382,391	77,846	135,651	6,772,559
Pseudomonas sp.	GM84	49	134	603	256,293	43,541	81,450	5,834,476
Rhizobium	PD01-076	49	168	503	607,718	32,884	211,768	5,524,448
Pantoea sp.	YR343	57	50	922	697,582	106,205	306,286	5,310,247
Herbaspirillum sp.	YR522	57	85	519	436,249	60,513	117,120	5,143,583
Phyllobacterium sp.	YR531	57	34	566	843,029	146,458	330,782	4,979,581
Bradyrhizobium sp.	YR681	49	110	564	372,490	72,282	133,556	7,951,029

Table 2.8: REAPR evaluation results for *Rhizobium* sp. strain CF080, *Burkholderia* sp. strain BT03, *Pseudomonas* sp. strain GM30 and *Pseudomonas* sp. strain GM41 assemblies.

Strain	Library type	Software	Genome Size (Mb)	No. of Contigs	N50 (kb)	No. of Contigs (Corrected)	N50 (Kb) (Corrected)	Errors	Error Free Bases	Collapse Repeats
Rhizobium sp. CF080	**PE-MP**	**APLG**	**7.04**	**40**	**626**	**40**	**626**	**8**	**98.1**	**8**
	PE	Soap	6.95	1850	6	1850	6	20	85.73	8
	PE-MP-454	Newbler	7.01	32	616	32	616	36	97.55	3
	PE	Velvet	6.96	856	14	856	14	36	92.3	13
	Pacbio-454	PBcR	7.06	102	188	102	188	49	85.43	8
	PE-MP-PacBio	SPAdes	7.04	22	1779	22	1779	59	86	5
	PE	Abyss*	8.2	90	238	108	193	60	74.61	1
	PE-MP-PacBio	APLG	7.05	16	671	16	671	74	86.47	3
	PE-MP-454	MaSuRCA	7.23	252	4095	252	4095	85	95	1
	PE-MP	MaSuRCA	7.12	163	597	163	597	143	95	6
	PE	MaSuRCA	7.04	270	69	270	69	219	94	6
	PE-MP	SPAdes	7.95	1372	663	1372	663	310	86	0
	PE	SPAdes	7.96	1426	312	1426	312	318	86	8
	PE	CLC*	7.54	1039	76	1368	76	432	89.59	12
Pseudomonas sp. GM30	**PE-MP**	**APLG**	**6.16**	**44**	**230**	**44**	**230**	**1**	**96.74**	**0**
	PE	Soap	6.09	1398	7	1398	7	19	87.89	2
	PE	Velvet	6.11	773	14	773	14	28	92.54	4
	PE-MP-454	Newbler	6.15	32	299	32	299	28	98.21	5
	PE	CLC	6.14	180	60	180	60	29	97.04	4
	PE	Abyss*	6.47	78	158	91	143	43	89.64	4
	PE	SPAdes	6.15	61	186	61	186	56	97.80	4

Table 2.8 continued …

Strain	Library type	Software	Genome Size (Mb)	No. of Contigs	N50 (kb)	No. of Contigs (Corrected)	N50 (Kb) (Corrected)	Errors	Error Free Bases	Collapse Repeats
Pseudomonas sp. GM30	PE-MP	SPAdes	6.20	50	240	50	240	59	97.34	0
	PE	MaSuRCA	6.50	1216	11	1216	11	91	86.95	2
	PE-MP-454	MaSuRCA	6.70	778	448	778	448	354	86.98	4
	PE-MP	MaSuRCA	6.38	570	62	570	62	528	89.37	5
Pseudomonas sp. GM41	**PE-MP**	**APLG**	**6.66**	**62**	**248**	**62**	**248**	**5**	**96.61**	**0**
	PE	Soap	6.56	1162	9	1162	9	130	88.9	4
	PE	Velvet	6.59	652	17	652	17	158	91.95	13
	PE	CLC	6.61	164	75	164	75	189	95.24	10
	PE-MP-454	Newbler	6.62	66	160	66	160	195	95.99	3
	PE	SPAdes	6.64	101	165	101	165	216	96	9
	PE-MP	SPAdes	6.71	86	1838	86	1838	219	95	1
	PE-MP-PacBio	APLG	6.68	13	1393	13	1393	236	93.58	2
	PE	Abyss*	6.82	114	138	193	59	239	91.89	7
	PE-MP-PacBio	SPAdes	6.67	70	291	70	291	242	93	2
	Pacbio-454	PBcR*	6.79	80	140	81	140	350	91.59	2
	PE	MaSuRCA	6.64	212	85	212	85	390	93	9
	PE-MP	MaSuRCA	6.7	157	279	157	279	419	93	12
	PE-MP-454	MaSuRCA*	7.27	686	739	696	739	421	86	3

Table 2.8 continued …

Strain	Library type	Software	Genome Size (Mb)	No. of Contigs	N50	No. of Contigs (Corrected)	N50 (Kb) (Corrected)	Errors	Error Free Bases	Collapse Repeats
Burkholderia sp. BT03	**PE-MP**	**APLG**	**10.91**	**135**	**177**	**135**	**177**	**10**	**93.2**	**25**
	PE-MP-454	Newbler	10.7	270	69	270	69	38	93	28
	PE	Soap	10.43	3016	5	3016	5	40	82.43	18
	PE	SPAdes	10.82	397	80	397	80	68	92	46
	PE-MP	SPAdes	11.16	362	77	362	77	77	89	7
	PE	CLC	10.64	690	30	690	30	82	89.92	46
	PE	Abyss*	11.55	475	48	508	45	106	82.94	11
	PE	Velvet	10.52	1914	9	1914	9	124	86.19	45
	PE	MaSuRCA	11.2	2226	10	2226	10	161	83	14
	PE-MP-454	MaSuRCA*	11.58	887	283	888	283	223	87	5
	PE-MP-PacBio	SPAdes	11.06	368	66	368	66	492	72	17
	Pacbio-454	PBcR*	11.4	235	100	236	100	1215	91.98	9
	PE-MP	MaSuRCA	10.95	806	59	806	59	1365	86	22

*Assemblies containing least number of errors in each genome are shown in bold.
[a]REAPR broke the assembly at erroronious regions and generated corrected contig numbers and corrected N50 value.
ALLPATHS-LG assembler is denoted with abbreviation APLG in current table.

Table 2.9: Summary of PBJelly gap filling results.

		BT03	**CF080**	**GM41**
[a]Input assembly statistics	No. of Gaps	96	7	5
	Total Gap Length (bp)	195,912	2,880	3,475
[b]PBJelly assembly statistics	No. of Gaps	26	2	3
	Total Gap Length (bp)	70,100	30	232

[a]Gap statistics for the best scaffold assembly; [b]Gap statistics after application of PBJelly algorithm

Table 2.10: CGAL (Rahman and Pachter, 2013) (version 0.9.6) evaluation results for *Rhizobium* sp. strain CF080, *Burkholderia* sp. strain BT03, *Pseudomonas* sp. strain GM30 and *Pseudomonas* sp. strain GM41 assemblies.

Data Types	Strain/Software	CF80	GM41	BT03	GM30
PE	Abyss	-1.37E+09	-1.28E+09	-1.15E+09	-1.28E+09
	Soap	-1.71E+09	-1.59E+09	-1.59E+09	-1.65E+09
	CLC	-1.40E+09	-1.30E+09	-1.21E+09	-1.31E+09
	Velvet	-1.54E+09	-1.37E+09	-1.36E+09	-1.42E+09
	SPAdes	-1.34E+09	-8.57E+07	-1.51E+08	-4.17E+07
	MaSuRCA	-1.48E+09	-9.56E+07	-1.74E+08	-4.45E+07
PE-MP	ALLPATHS-LG	-1.36E+09	-1.28E+09	-1.17E+09	-1.29E+09
	SPAdes	**-1.35E+09**	-8.47E+07	-1.49E+08	**-4.14E+07**
	MaSuRCA	-1.36E+09	-9.10E+07	-1.77E+08	-4.61E+07
PE-MP-454	Newbler	-1.37E+09	-1.32E+09	-1.23E+09	-1.32E+09
	MaSuRCA	-1.35E+09	-8.49E+07	-1.48E+08	-4.35E+07
PacBio	PacBiotoCA	-1.36E+09	-8.40E+07	-1.50E+08	NA
PE-MP-PacBio	ALLPATHS-LG	-1.35E+09	**-8.34E+07**	NA	NA
	SPAdes	-1.36E+09	-8.48E+07	**-1.48E+08**	NA

*The best CGAL scores are shown in bold.

Table 2.11: Comparison of ORFs predicted in draft and improved genome assemblies.

Strains	CF080	BT03	GM30	GM41
[a]Total ORFs	6,684	10,056	5,511	5,975
[b]No. of unchanged ORFs	5,819	9,385	5,424	5,881
No. of longer ORFs	786	413	77	71
No. of shorter ORFs	64	205	10	15
No. of new ORFs	15	53	0	8

[a]Total number of open reading frames predicted in improved genome assembly by Prodigal gene calling algorithm.
[b]Number of open reading frames in improved genome assemblies as compared to draft assemblies.

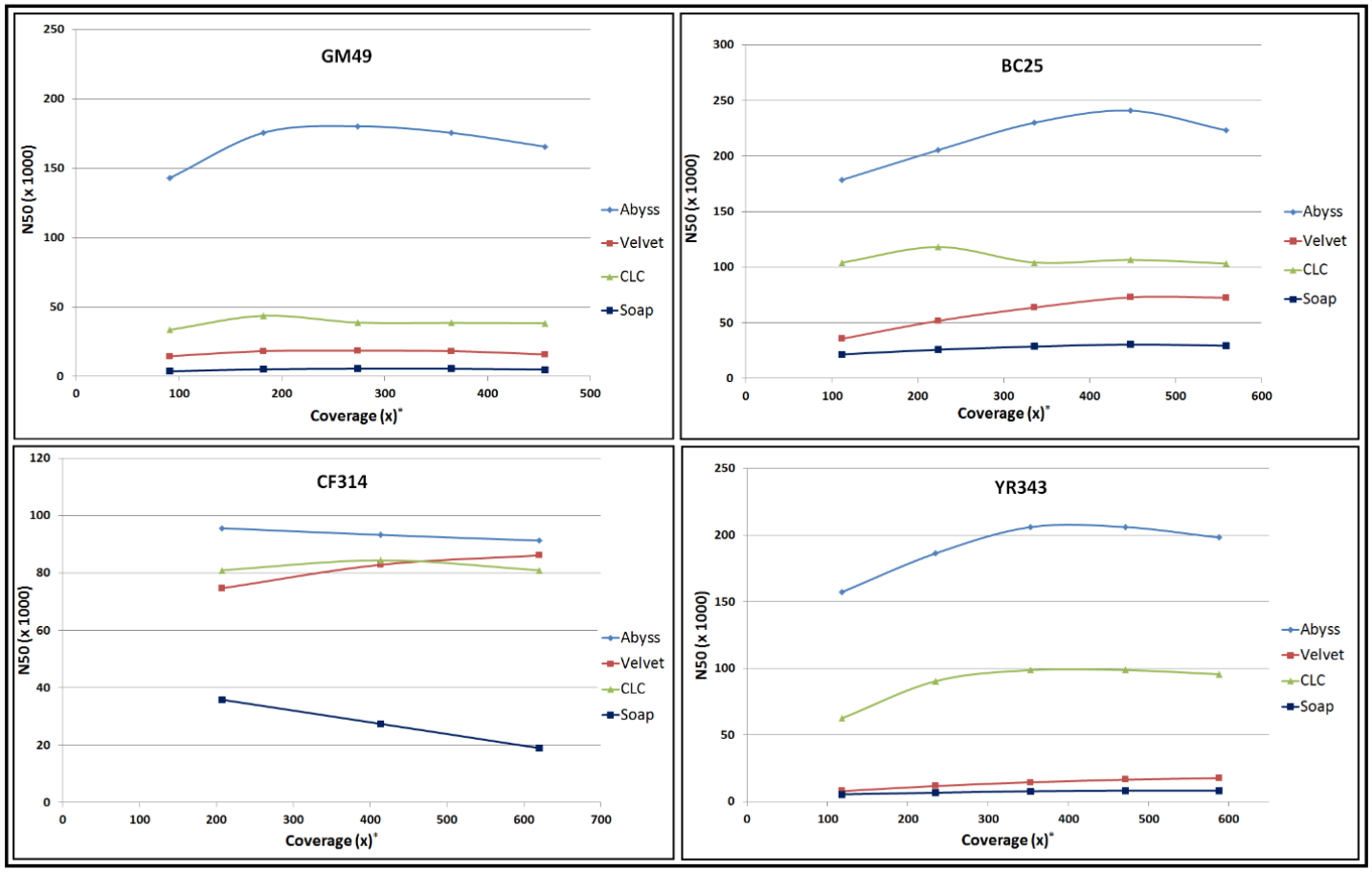

Figure 2.1: Coverage analysis.

The genome assemblies of four isolates were created at incremental raw read coverage levels using ABySS, Velvet, CLC and SOAP software. The optimal N50 values generated by each software at various coverage levels are shown.

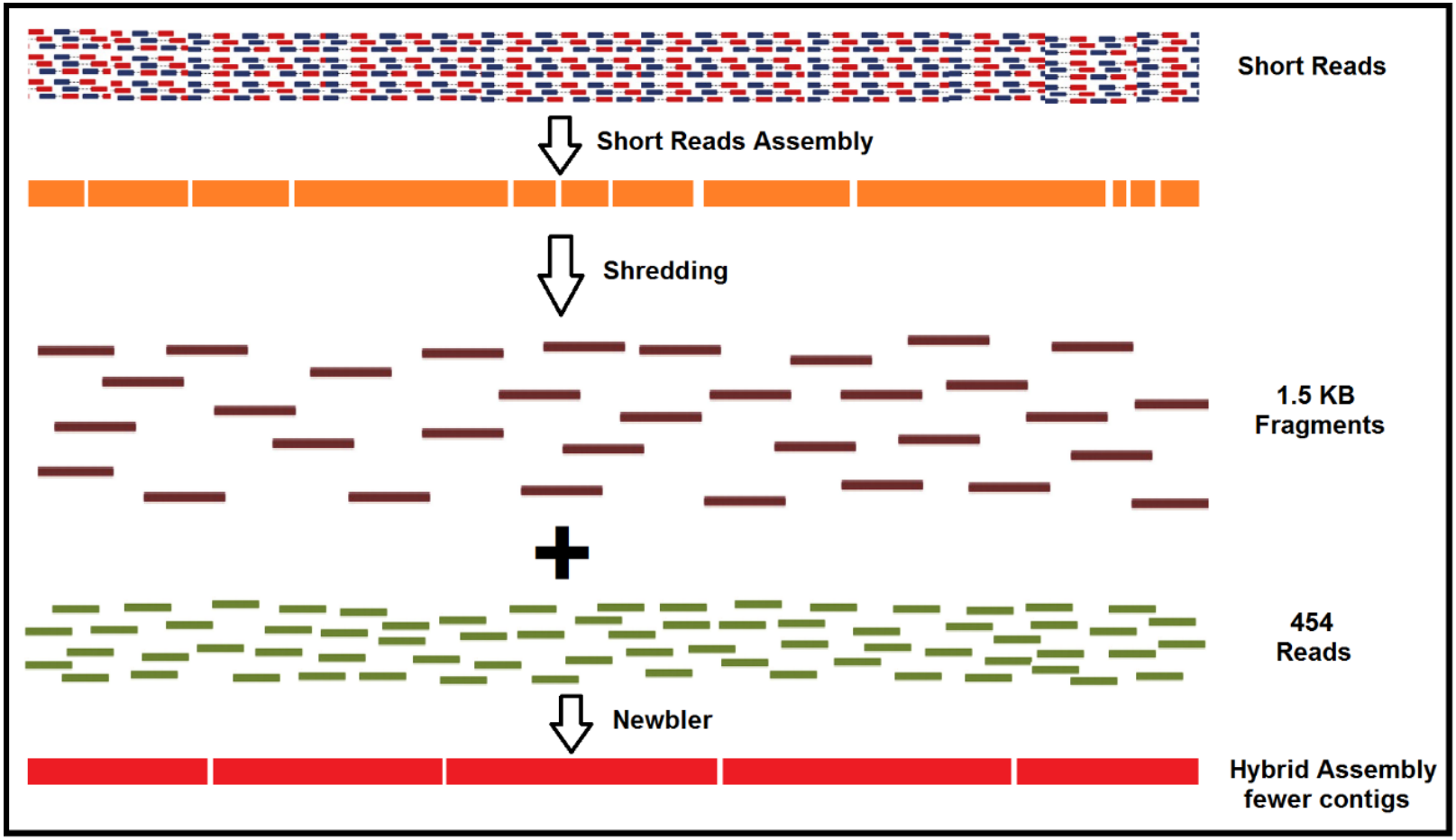

Figure 2.2: Overview of 454 and Illumina hybrid assembly.

Representation of shredding approach to generate 454 and Illumina hybrid assembly.

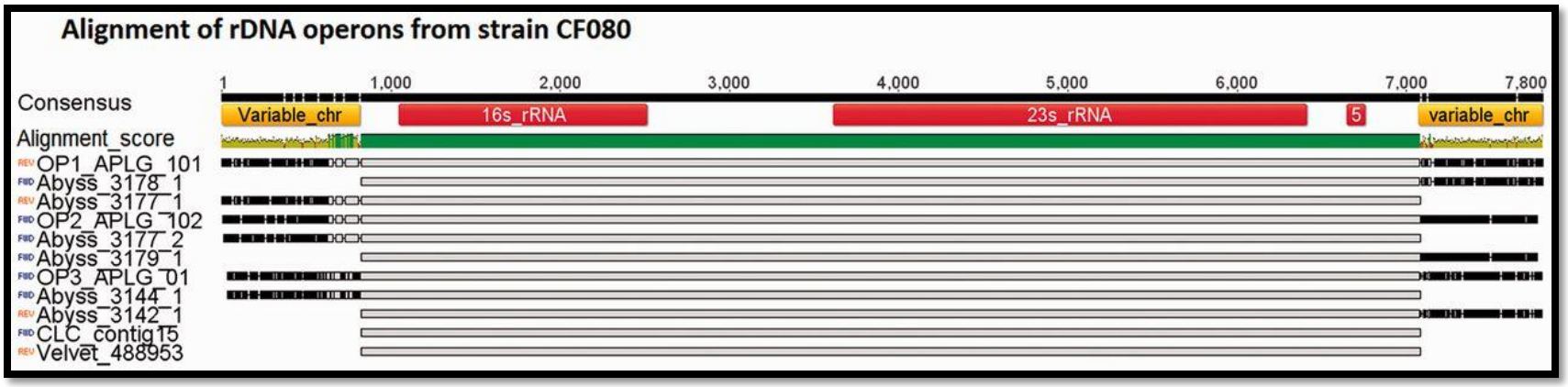

Figure 2.3: Alignment of predicted CF080 rDNA operons tested via PCR and Sanger sequencing.

The names of the operon denotes corresponding assembly algorithm (ALLPATHS-LG is displayed as APLG) and contig ID. The annotation and the genomic position are shown on the consensus sequence.

Figure 2.4: Alignment of predicted rDNA operons tested via PCR and Sanger sequencing.

Alignment of predicted rDNA operons in strain (b) GM41 (c) GM30 (d) BT03 is shown. Multiple copies of rDNA operon were detected by variability in the alignment identity and designated with prefix ‘OP’. The names of the operon denotes corresponding assembly algorithm (ALLPATHS-LG is displayed as APLG) and contig ID. The annotation and the genomic position are shown on the consensus sequence. The PCR primer sequences are denoted as green triangles and alignment identity is shown as “Alignment_score”. * The “Alignment_score” correspond to the color (Green – 100%, Red – Below 50%) and height of the graphic.

Explanation for Figure 2.4.

rDNA operons in Pseudomonas sp. strain GM41

Two to seven rDNA operons have been detected within 35 finished *Pseudomonas* genomes sequences available through IMG database (Markowitz, et al., 2012). For strain GM41, all four copies of rDNA operons from the ALLPATHS-LG assembly were verified by PCR and Sanger sequencing (b). The ALLPATHS-LG hybrid assembly predicted all four copies on separate contigs with their corresponding flanking chromosomal regions. SPAdes assembly predicted a one complete rDNA operon along with its flanking chromosomal regions. The CLC assembly was able to assemble a single complete rDNA operon but flanking chromosomal regions and multiple copies were missing. Velvet and ABySS were only able to assemble the individual (5S, 16S, and 23S) rDNA elements but the rDNA operon structure was incomplete. Hence for strain GM41 the ALLPATHS-LG assembly predicted multiple copies of the rDNA operons, which were supported by the PCR and Sanger sequencing, thus providing additional confidence in its quality.

rDNA operons in Pseudomonas sp. strain GM30

The ABySS, Velvet, SPAdes and CLC assemblies each supported predictions for only one partial rDNA operon (containing any two of the 5S, 16S, or 23S rDNA sequences) while the ALLPATHS-LG assembly predicted two partial rDNA operon copies (c). PCR and Sanger sequencing assessments identified four partial rDNA operons, that each had unique associated flanking DNA. Based on the operon arrangement it was hypothesized that contigs, labelled OP5_APLG_0 and OP1_Abyss_2616 could be joined, and tests based upon PCR and Sanger sequencing were able to join these contigs into one contiguous DNA sequence. However, a similar PCR strategy was unable to join the contigs labelled OP4_APLG_31 and OP2_CLC_107 (c). Hence, we were able to merge two partial operons into one complete rDNA operon. The poor assembly of complete rDNA operons in strain GM30 was attributed to a lack of PacBio data. This strain was characterized by the production of surfactant-like compound that may have interfered with the PacBio chemistry as sequencing failed in two attempts.

rDNA operons in Burkholderia sp. strain BT03

There are two to seven rDNA operons in 35 *Burkholderia* genomes available through IMG database (Markowitz, et al., 2012) and assemblies for strain BT03 supported one complete and four partial rDNA operons. One complete operon was identified by CLC and SPAdes while the remaining four were partial operons (identified by ABySS) (d). All five operons were confirmed by PCR and Sanger sequencing. The complete operon from CLC/SPAdes does not include the flanking chromosomal region while ABySS predicted operons include only 3' flanking chromosomal regions. It is possible that operon found in CLC/SPAdes assembly is the same as the one predicted from ABySS assembly. However, the exact arrangement could not be validated as 5' flanking chromosomal regions were missing from all our rDNA operon assemblies. The four copies of partial rDNA operons could prove useful if future manual finishing is to be undertaken. As mentioned previously, PacBio reads were unable to be incorporated into the ALLPATHS-LG assembly due to computational resource limitations.

CHAPTER 3 : COMPARISON OF SINGLE-MOLECULE SEQUENCING AND HYBRID APPROACHES FOR FINISHING THE GENOME OF *CLOSTRIDIUM AUTOETHANOGENUM* AND ANALYSIS OF CRISPR SYSTEMS IN INDUSTRIAL RELEVANT CLOSTRIDIA

Disclosure: This chapter was published as:

Brown S. D., Nagaraju S, Utturkar S. M., De Tissera S, Segovia S, Mitchell W, Land M. L., Dassanayake A, Kopke M. **(**2014). Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of *Clostridium autoethanogenum* and analysis of CRISPR systems in industrial relevant Clostridia. Biotechnology for Biofuels 7**:**40.

Sagar Utturkar's contributions include bioinformatics analysis, *de novo* and hybrid assemblies and *in silico* evaluations, OrthoMCL analysis, Circos figure creation, and assistance with PCR and Sanger sequencing. Sagar Utturkar also contributed towards the study design and manuscript preparation. Dr. Shilpa Nagaraju carried out the work around the CRISPR system and RT-PCR. Dr. S. De Tissera prepared genomic DNA. S. Segovia carried out the RNA-Seq experiment. Dr. Steven Brown, Dr. Wayne Mitchell, Dr. A. Dassanayake, and Dr. Michael Kopke contributed bioinformatics analysis. Dr. Steven Brown, Dr. Michael Kopke conceived and designed the study and prepared the manuscript.

3.1 Abstract

Background

Clostridium autoethanogenum strain JA1-1 (DSM 10061) is an Gram-positive, anaerobic, mesophilic, acetogenic bacterium capable of fermenting CO, CO_2 and H_2 (e.g. from syngas or waste gases) into biofuel ethanol and commodity chemicals such as 2,3-butanediol. The bacterium is currently being deployed for large-scale industrial applications and a draft genome sequence consisting of 100 contigs has been published recently.

Results

In this study, a closed, high-quality genome sequence for strain *C. autoethanogenum* DSM10061 was generated using only the latest single-molecule DNA sequencing technology and without the need for manual finishing. *C. autoethanogenum* strain DSM 10061 is assigned to the most complex genome classification based upon genome features such as repeats, prophage, nine copies of the rRNA gene operons and as an additional layer of complexity it has a low GC content of 31.1%. Illumina, 454, Illumina/454, IonTorrent/454 hybrid assemblies were generated and then compared to the draft and PacBio assemblies using summary statistics, CGAL, QUAST and REAPR bioinformatics tools and comparative genomic approaches. The results of this study indicated that assemblies based upon the shorter read DNA technologies were confounded by the large number repeats and their size, which in the case of the rRNA gene operons were ~5 kb. CRISPR loci (Clustered Regularly Interspaced Short Palindromic Repeats) are dynamic, hyper-variable, acquired bacterial defence islands that provide immunity against mobile genetic elements and are used for bacterial genotyping. CRISPR systems among biotechnologically relevant Clostridia were classified and related to plasmid content and prophages. While *C. autoethanogenum* contains an active CRISPR system, no such system is present in the closely related *Clostridium ljungdahlii* DSM 13528. There is a common prophage inserted into Arg-tRNA shared between the strains that suggests a common ancestor. However, *C. ljungdahlii* contains several additional putative prophages and it has more than double the amount of prophage DNA compared to *C. autoethanogenum*. Other differences include important metabolic genes for central metabolism (as an additional hydrogenase and the absence of a phosphoenolpyruvate synthase) and substrate utilization pathway (mannose and aromatics utilization) that might explain phenotypic differences between *C. autoethanogenum* and *C. ljungdahlii*. Among the Clostridia examined, seven strains contain CRISPR systems and only one of these contains plasmid DNA, while among the five strains that contain plasmid DNA only one has a CRISPR system. Potential associations between plasmid content and CRISPR systems may have implications for historical industrial scale Acetone-Butanol-Ethanol (ABE) fermentation failures and future large scale bacterial fermentations.

Conclusions

This study supports the notion that single molecule sequencing will be increasingly used to produce finished microbial genomes. The complete genome sequence for *C. autoethanogenum* strain DSM 10061 will facilitate future comparative genomics and functional genomics studies with this strain and developed strains for process

commercialisation. The complete genome will also support future comparisons between Clostridia and studies that examine the evolution of plasmids, bacteriophage and CRISPR systems.

3.2 Introduction

After completion of the first human genome sequence the development of next-generation DNA sequencing technologies led to remarkable increases in sequencing efficiency, in the order of approximately 100,000-fold (Treangen and Salzberg, 2012). Costs have dropped dramatically and computational methods have advanced along with sequencing technology leading to large increases in DNA sequencing output and in the number of available genome sequences (Chain, et al., 2009; Nagarajan and Pop, 2013). A variety of assembly algorithms and methods for quality evaluation have been developed (Brown, et al., 2011; Gurevich, et al., 2013; Hunt, et al., 2013; Kisand and Lettieri, 2013; Magoc, et al., 2013; Mavromatis, et al., 2012; Medini, et al., 2008; Nagarajan and Pop, 2013; Rahman and Pachter, 2013; Salzberg, et al., 2012; Vezzi, et al., 2012). However, the majority of sequenced genomes are incomplete due to technical difficulties, time, and expense leading to an increasing disparity between the number of finished and draft genomes in databases (Chain, et al., 2009; Koren, et al., 2013; Mavromatis, et al., 2012; Nagarajan and Pop, 2013; Treangen and Salzberg, 2012).

The PacBio sequencing system (Eid, et al., 2009) was the first long-read, single-molecule sequencer available and its performance has been compared to two short read sequencing platforms also released in 2011 (Quail, et al., 2012). The original RS system with C1 chemistry generated mean read lengths in the range of 1,500 bp and yielded approximately 100 Mb of sequence data per run, and reads in this range were useful in generating improved scaffolds for *de novo* assemblies. However, the original system was not optimal for *de novo* assembly applications (Quail, et al., 2012) and hybrid assembly approaches have been developed to overcome limitations in short read technologies and higher error rates associated with third generation technology (Bashir, et al., 2012; English, et al., 2012).

Repetitive stretches of DNA are abundant and are one of the main technical challenges that hinder accurate sequencing and genome assembly efforts (Treangen and Salzberg, 2012). In the case of bacteria, the rRNA gene operon is often the largest region of repetitive sequence and these range in size between 5 and 7 kb (Treangen, et al., 2009). Several years ago the longest PacBio RS reads were reported as being approximately 14 kb and these longer reads are useful in resolving repeats during genome assemblies (Nagarajan and Pop, 2013). The PacBio RS II system was released several years ago and it produces more and longer reads. In a recent study, the longest read before correction was 15,634 bp and the genomes of six bacteria were sequenced and assembled using single-molecule sequencing based on C2 chemistry (Koren, et al., 2013). Koren et al (2013) suggested that the majority of bacterial genomes could be assembled into finished-grade quality, i.e. without gaps, and with data derived from a single PacBio sequencing library per sample (Koren, et al., 2013). The combination of the longer reads, depth of coverage and random nature of sequencing errors facilitates *de novo* assemblies for microbial isolates (Chin, et al., 2013; Eid, et al., 2009; Koren, et al.,

2012). The advantages of single-molecule sequencing have been discussed (Roberts, et al., 2013). At the time the *C. autoethanogenum* genome sequence was determined, relatively few genomes sequences were available exclusively via single-molecule technology and only a handful represent finished genomes (Chin, et al., 2013; Hoefler, et al., 2013; Koren, et al., 2013; Koren, et al., 2012; Powers, et al., 2013; Rasko, et al., 2011).

In this study, a finished genome sequence for *Clostridium autoethanogenum* strain JA1-1 (DSM 10061) was generated using the latest PacBio RS II instrument. This represents one of the first *de novo* genomes finished into a single contiguous sequence using RS II data alone (i.e. without addition of other next-generation sequence data or manual finishing steps). To offer insights into this technology, the PacBio assembly was compared to assemblies based on 454 GS FLX Titanium and Illumina MiSeq data and an earlier draft genome sequence of 100 contigs for this strain obtained from 454 GS FLX Titanium and Ion Torrent data (Bruno-Barcena, et al., 2013).

C. autoethanogenum is an anaerobic, Gram-positive, mesophilic, acetogenic bacterium isolated using carbon monoxide (CO) (Abrini, et al., 1994). Other substrates include the greenhouse gas CO_2 plus H_2, pyruvate, xylose, arabinose, fructose, rhamnose, and L-glutamate. There is significant biotechnological interest in this organism as well as other acetogenic bacteria for their abilities to use gases containing CO, H_2 and CO_2 as the sole source of carbon and energy for the production of fuel and chemicals at scale. The ability to use these gases in fermentative processes enables acetogens to potentially provide a route to more sustainable fuel and chemical production from a range of feedstocks including biomass and municipal solid waste-derived syngas, reformed biogas and industrial waste gases derived for example from steel production facilities (Kopke, et al., 2010; Köpke, et al., 2011; Köpke, et al., 2011; Mohammadi, et al., 2011; Munasinghe and Khanal, 2010; Tirado-Acevedo, et al., 2010).

3.3 Methods

DNA sequence data generation

C. autoethanogenum strain JA1-1 was obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ) culture collection (DSM 10061). *C. autoethanogenum* strain JA1-1 was cultured in PETC medium as described (Kopke, et al., 2010). A single colony was purified and 16S rDNA sequence confirmed before genomic DNA was prepared. High molecular weight genomic DNA was prepared as described earlier (Kopke, et al., 2010), quantified with a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, DE) and quality was assessed with Agilent Bioanalyzer (Agilent, Santa Clara, CA).

Pyrosequencing was conducted using the Roche 454 GS FLX System (Roche 454) with the method of paired-end DNA library preparation and average insert sizes in the 3 kb range and Titanium chemistry, according to the manufacturer's instructions and described previously (Brown, et al., 2012; Mardis, 2008). Sequence data was also generated using an Illumina MiSeq instrument (Quail, et al., 2012) and a paired-end approach with an approximate insert library size of 500 bp and read lengths of 151 bp, as described

previously (Brown, et al., 2013) and according to the manufacturer's instructions. DNA for PacBio sequencing was sheared with G-tubes (Covaris, Inc., Woburn, Massachusetts), targeting 20 kb fragments. PacBio libraries were prepared with the DNA Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA) and library fragments above 4 kb were isolated using the Blue Pippin system (Sage Science, Inc., Beverly, MA). The average PacBio library insert size (including adapters) was ~19 kb and samples were sequenced using Magbead loading, C2 chemistry, Polymerase version P4, and software version 2.02. Raw next-generation sequence data available through the NCBI SRA database (accession SRX352885, SRX352888, SRP030033). Polymerase chain reactions (PCR) and Sanger sequencing were conducted using standard approaches as described previously (Brown, et al., 2011).

Sequence data trimming, filtering, annotation and assembly

The CLC Genomics Workbench (version 6.0.2) was used to trim and filter Illumina reads for quality sequence data and the subsequent Illumina assembly. The Newbler application (version 2.8) in the 454 GS FLX software package (Roche 454) was used to assemble reads generated from the GS FLX instrument and in combination with reads from the Illumina instrument, as described previously (Brown, et al., 2012). The consensus Illumina sequences were processed before inputting into the Newbler assembler by generating 1.5-kb overlapping fake reads using the fb_dice.pl script, which is part of the FragBlast module (http://www.clarkfrancis.com/codes/fb_dice.pl). The PacBio reads were assembled through SMRTanalysis v 2.0 (Pacific Biosciences) using HGAP protocol (Chin, et al., 2013). The DSM 10061 PacBio assembly was annotated using the Prodigal gene calling algorithm (Hyatt, et al., 2010) and deposited at in the NCBI Genbank database under accession number CP006763.

Assessment of genome assembly quality

The *in silico* evaluation of genome assemblies was performed using CGAL (version 0.9.6) (Rahman and Pachter, 2013), REAPR (version 1.0.16) (Hunt, et al., 2013), QUAST (version 2.2) (Gurevich, et al., 2013) and Circos (Krzywinski, et al., 2009). The genomic repeats were identified using Nucmer (Kurtz, et al., 2004) for genome complexity was determined based on count and length of the repeats as suggested earlier (Koren, et al., 2013). Gaps in the 454/Illumina hybrid and published draft assemblies were determined by performing multiple genome alignment through Mauve (version 2.3.1) (Darling, et al., 2010) with PacBio assembly was used as reference genome. The order of contigs in 454/Illumina hybrid assembly and alignment of Sanger sequences was determined using Geneious software (version 6.1.5) (Auckland, New Zealand).

Summary CRISPR analysis

Disclaimer: *The CRISPR analysis was performed by the Michael Köpke and colleagues at LanzaTech Ltd. New Zealand. A summary of this analysis is provided and a more complete description is available in online version of this manuscript.*

The genome of *C. autoethanogenum* (NC_022592) and genome sequences of *C. acetobutylicum* ATCC824 (NC_003030), DSM1731 (NC_015687) and EA2018 (NC_017295), *C. beijerinckii* NCIMB8052 (NC_009617), *C. saccharobutylicum*

(NC_022571), *C. saccharoperbutylacetonicum* (NC_020291), *C. cellulolyticum* H10 (NC_011898), *C. cellulovorans* 743B (NC_014393), *C. thermocellum* ATCC27405 (NC_009012) and DSM1313 (NC_017304), *C. phytofermentans* ISDg (NC_010001), *C. ljungdahlii* DSM13528 (NC_014328) and *C. carboxidivorans* (ACVI01000000; ADEK01000000) were retrieved from NCBI Genbank. The genome sequences for these organisms and plasmid contents (is any) were analysed for for CRISPR repeats using PILER algorithm (Edgar, 2007) and CRISPRdb (Grissa, et al., 2007). Analysis of prophage regions was performed through PHAST (Zhou, et al., 2011), Phage_Finder (Fouts, 2006). Phylogenetic analysis was performed using multiple sequence alignment of 16S rRNA and *cas1* genes using Geneious software. Reverse Transcriptase PCR (RT-PCR) was performed to study the expression and operon structure of *cas* genes and the expression CRISPR arrays. RNA-Seq was performed from *C. autoethanogenum* growing in continuous culture in a 1.5L continued-stirred tank reactor (CSTR) with steel mill waste gas (composition: 42% CO, 36% N_2, 20% CO_2, and 2% H_2; collected from a New Zealand Steel site in Glenbrook, New Zealand) as the sole energy and carbon source as described previously (Wang, et al., 2013).

3.4 Results and Discussion

Sequencing Output and Assembly Statistics for *C. autoethanogenum* DSM 10061.
Sequencing statistics show that for each platform a large number of raw reads were attained that resulted in high degrees of genome coverage (Table 3.1) (All tables and figures are located in the appendix section). Raw Illumina data were trimmed and filtered before assembly, but in the case of the 454 and PacBio assembler's raw instrument output files were used. Bruno-Barcena et al. used a combination of 454 GS FLX Titanium and Ion Torrent Personal Genome Machine (PGM) data to generate a genome reported as 4.5 Mb for *C. autoethanogenum* DSM 10061 (Bruno-Barcena, et al., 2013). The number of 454 reads (452,052) and genome coverage (39x) from the earlier study was similar to this one (Table 3.1), although addition of the PGM reads resulted in 905,738 raw reads being used to generate the preliminary assembly by Newbler (version 2.6). The Genbank record (ASZX00000000.1) for strain DSM 10061 draft genome is reported as 4,323,309 bp.

In this study, Newbler (version 2.8) was used to assemble new 454 paired-end reads from a 3-kb insert length library (Table 3.1) into a draft genome sequence that consisted of 32 contigs (Table 3.2). The lower number of contigs (32 vs 100) from the new 454 only assembly compared to the draft version (Bruno-Barcena, et al., 2013) is likely due to differences in library types (paired-end verses shotgun) and software versions. Assembly of Illumina only data was conducted using the SPAdes (Bankevich, et al., 2012), Velvet (Zerbino and Birney, 2008), Abyss (Simpson, et al., 2009) and the CLC Genomics Workbench (CLC Bio) assemblers and the best results were obtained by the Velvet assembler (Table 3.2). Previously, we have assembled genome sequences for a range of bacteria using a combination of 454 and Illumina technologies, whereby initial Illumina consensus sequences were shredded into 1.5-kb overlapped fake reads and assembled together with the 454 data (Brown, et al., 2012; Brown, et al., 2012; Brown, et al., 2012; Elkins, et al., 2010; Utturkar, et al., 2013; Utturkar, et al., 2013). The best genome

assembly obtained for strain DSM 10061 using second generation sequencing technologies employed such a hybrid approach, which is reflected in the lowest number of contigs, the largest single contig and highest N50 value (Table 3.2). Preliminary studies using the *Clostridium ljungdahlii* DSM 13528 genome as a reference and a PCR/Sanger sequencing strategy showed contigs could be joined by such an approach (Figure 3.1). As manual finishing is time consuming the potential of PacBio data to generate finished microbial genome sequences was assessed.

Remarkably, one PacBio library preparation and two SMRT cells produced sufficient sequence such that it could be assembled into one contiguous DNA fragment that represented the DSM 10061 genome. The PacBio genome assembly is a similar size to the other assemblies (Table 3.1 and Table 3.2) and genome completeness was confirmed by sequence wrap-around. This is one of the first *de novo* sequenced genomes that we are aware that has been closed without manual finishing or additional data, despite the complexity of the *C. autoethanogenum* genome.

A comparison of the 454/Illumina hybrid assembly to the PacBio assembly showed there were small regions of overlap in the hybrid assembly that weakly joined contigs, and were supported by PCR and Sanger data, but there was insufficient support for the Newbler software to join them (Figure 3.1). PCR and Sanger data joined small gaps between contigs (e.g. Figure 3.1) in line with predictions using *C. ljungdahlii* DSM 13528 as a reference but in other examples much larger products were obtained compared to the predicted PCR product sizes (Figure 3.1C). Other challenges involved using a related but different species or strain from manual finishing included instances of software not being able to design PCR primers, not obtaining PCR products, and instances of obtaining multiple PCR products of different sizes and/or DNA smears.

Assembly Quality Assessments and Comparisons. The complexity of the *C. autoethanogenum* DSM 10061 genome sequence was assessed and it is classified as a class III genome, according to previously described criteria for repeat sequence content and type (Koren, et al., 2013). Class III genomes are defined as containing repeats that can include rRNA gene operons, many mid-scale repeats, such as insertion sequences and simple sequence repeats, and large phage-mediated repeats, duplications, or large tandem arrays that are considerably larger than the rRNA gene operon.

PacBio sequencing technology has a high error rate, which has been reported as being approximately 18% (Nagarajan and Pop, 2013). Due to the random nature of the error (Eid, et al., 2009), it is however possible to get a highly accurate consensus sequence when there is high coverage (Chin, et al., 2013; Koren, et al., 2013; Koren, et al., 2012). For genomes such as *C. autoethanogenu*m with extreme GC contents (31.1 mol% GC content) and long homonucleotide stretches this provides an advantage over other sequencing technologies.

Beyond simple metrics, such as contig number, N50 and largest contig size, several bioinformatics approaches have been developed to assess assembly quality. The computing genome assembly likelihoods (CGAL) method is one recent approach that

assesses uniformity of read coverage for assemblies and also evaluates the read errors, library insert size distribution and the degree of unassembled data (Rahman and Pachter, 2013). At present, CGAL is only able to utilize Illumina reads for its assembly assessment and using Illumina reads it ranked the assemblies in the order of best to worst as Illumina only, Illumina/454 hybrid, 454, published draft, to PacBio, respectively (Table 3.3). The CGAL likelihood principle is based on the possibility that a read is produced from every single location in the assembly. Regions of repetitive DNA were to be sequenced by longer reads, which were at times not resolved by the Illumina reads (Figure 3.2) (All tables and figures are located in the appendix section) and this may have contributed to the lower CGAL scores for assemblies that contained longer reads and no Illumina data. QUAST (Gurevich, et al., 2013), which used the PacBio assembly as the reference, ranked the Illumina/454 hybrid, 454, published draft, and Illumina only assemblies in the order of best to worst, respectively and additional details are provided (Table 3.4).

The REAPR tool for genome assembly evaluation (Hunt, et al., 2013) detected no collapsed repeats in the PacBio assembly and five in the hybrid assembly and four in each of the other assemblies (Table 3.5). The fragment coverage distribution (FCD) error detected by REAPR in PacBio assembly was at location 3872494-3873407 (913 bp). This region contains an rRNA gene operon and had very low Illumina coverage (40x as compared to the average of 127x). Hence, REAPR reported an error (based on Illumina reads only). Even 454 coverage was low in this region (19x as compared to average of 46x). However, there was 108x PacBio reads covering this (913 bp) region and for the first 392 bp there was also high quality Sanger sequence support indicating it is unlikely that there is an issue for the PacBio assembly in this region. The hybrid and PacBio assemblies contained the fewest warnings (83 and 96, respectively), followed by the Illumina assembly (182) and then published draft assembly contained the most (190).

A multiple genome alignment was conducted by aligning contigs from the different assemblies to the PacBio reference assembly to identify conserved regions and to evaluate gaps in the different DSM 10061 assemblies. Regions with no or partial 454 or Illumina contig coverage predominantly contained predicted rRNA gene operons and other duplicated genes (Figure 3.2 and Table 3.6). While the draft genome sequence for strain DSM 10061 predicts one copy of the 16S rRNA gene (Bruno-Barcena, et al., 2013), nine rRNA clusters were predicted using the DSM 10061 PacBio assembly, which is the same number of rRNA operons as in the closely related *C. ljungdahlii* DSM 13528 (Kopke, et al., 2010). Based on findings in this study and earlier ones (Koren, et al., 2013; Nagarajan and Pop, 2013; Treangen and Salzberg, 2012), the large number of DSM 10061 rRNA clusters and their repetitive nature confounded assembly of the shorter reads.

The latest PacBio RS II SMRT cells are designed to select for larger read lengths when long insert libraries (10-20 kb) are being prepared, however preferential loading of smaller fragments can still occur and this limits sequence output. In this study, smaller fragments were removed from the PacBio library by size exclusion leading to longer read lengths and greater amounts of sequence data than otherwise might have been attained. The long reads produced by the new PacBio RS II system, combined with sequence depth

meant that the principal regions of complexity could be resolved using one library preparation and two SMRT cells to generate a complete genome sequence. The application of long, single-molecule sequencing data will lead to a greater number of finished genomes and quality improvements in microbial genome databases (Koren, et al., 2013), however the application of the newest version of this technology requires more evaluation before its full potential can be assessed for complex genomes.

General Features of the *C. autoethanogenum* Genome, its Metabolism and Comparison to *C. ljungdahlii*.

The finished genome of *C. autoethanogenum* DSM 10061 consists of one chromosome of 4,352,205 bp in size with a GC content of 31.1 mol% and consists of 89 RNA genes (Table 3.7). Of the 4,161 genes predicted for this strain, 4,042 are protein-coding genes (CDSs) and 18 are pseudogenes. The distribution of genes into COG functional categories is presented (Table 3.8). The previously published draft DSM 10061 genome annotation included 4,135 predicted coding sequences (Bruno-Barcena, et al., 2013) and the related finished *C. ljungdahlii* DSM 13528 genome which is 277,860 bp larger in size contained 4,184 protein coding genes (Kopke, et al., 2010). Predicted gene content differences reflect the use of different gene calling algorithms, that draft sequences can split genes in two and genotypic differences. The methodology, accuracy, and specificity of the Prodigal gene prediction algorithm used in study has been described previously (Hyatt, et al., 2010).

Phenotypic and metabolic differences have been reported for *C. autoethanogenum* and *C. ljungdahlii* (Abrini, et al., 1994; Cotter, et al., 2009; Cotter, et al., 2009; Tanner, et al., 1993; Tirado-Acevedo, et al., 2011). The two are indistinguishable at the 16S rRNA gene level (Stackebrandt, et al., 1999) and have high scores for similarity based on *in silico* average nucleotide identity comparisons across the genomes (0.9977 ANIb) (Bruno-Barcena, et al., 2013). To evaluate potential coding sequence differences between the two organisms OrthoMCL (Chen, et al., 2006), a genome-scale algorithm for grouping orthologous protein sequences, was used to compare all the *C. autoethanogenum* proteins to those in *C. ljungdahlii* and for the reciprocal evaluation. Putative paralogs were identified along with putative orthologs. Proteins without orthologs or paralogs were identified using the default settings. This analysis revealed that over 10 % of the proteome is unique to each bacterium when comparing *C. autoethanogenum* (427 proteins out of 4,134) and *C. ljungdahlii* (447 out of 4,198). The 427 proteins unique genes to DSM 10061 (as listed by OrthoMCL) were searched against entire *C. ljungdahlii* proteome using BLASTP and an e-value similarity criteria of 1e-5 to identify proteins with truly unique function and no homolog, which reduced the number of dissimilar or unique proteins to 221 (Table 3.9). From the proteins identified as unique to each bacterium, the majority were proteins with hypothetical functions or proteins related to particular phage, transposon or CRISPR sequences, but proteins with key functions in the metabolism were also identified that could explain different phenotypes . These differences are discussed below.

Disclaimer: *Results below were derived from analysis of complete genome sequence and other experiments by the Dr. Michael Köpke and colleagues at LanzaTech Ltd. New Zealand. A summary of this analysis is provided and a more complete description is available in online version of this manuscript.*

The Wood-Ljungdahl pathway plays a key role in acetogen metabolism and the genes encoding for the enzymes of this pathway were found to be co-localized in one large cluster (CAETHG_ 1606-1621). As in *C. ljungdahlii*, two additional monofunctional carbon monoxide dehydrogenases (CAETHG_3005 and CAETHG_3899) are encoded in the genome of *C. autoethanogenum* that may also be involved in utilization of CO and CO_2. The genome of *C. autoethanogenum* encodes for six hydrogenases, one [NiFe] hydrogenase and five [FeFe] hydrogenases. Interestingly, the iron-only hydrogenases from *C. autoethanogenum* was missing from the *C. ljungdahlii*. This unique [FeFe] hydrogenase is in an operon with two genes for NuoF-like oxidoreductases (CAETHG_1575-78). The presence of an additional hydrogenase enzyme complex could represent a significant advantage for *C. autoethanogenum* during autotrophic growth on CO, CO_2 and H_2 containing gases. Additionally, preliminary RNA-Seq experiments show that this cluster is highly expressed under such conditions underlining the importance of this enzyme. Other features of *C. autoethanogenum* include two pyruvate:ferredoxin oxidoreductases (PFOR) that catalyze the conversion of acetyl-CoA into pyruvate, incomplete TCA cycle to succinate and 3-oxogluterate, A PTS system and other respective genes with possible role in heterotrophic growth on range of C5 and C6 sugars, and some extra genes involved in mannose metabolism and aromatic compound degradation. Other differences between *C. autoethanogenum* and *C. ljungdahlii* include several additional alcohol dehydrogenases, variations in the sporulation program with several unique proteins and regulators present in *C. autoethanogenum*, and differences in defense systems such as restriction/methylation systems and a CRISPR system. Despite the geographical separation of the isolates, the overall degree of similarity between *C. ljungdahlii* and *C. autoethanogenum* suggests a common ancestor.

***C. autoethanogenum* CRISPR system.**

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) are prokaryotic DNA loci that carry the memory of past bacterial infections of phages and plasmids to provide immunity against mobile genetic elements (Bhaya, et al., 2011; Sorek, et al., 2008). In the last decade, several studies have unravelled CRISPR defence molecular details and mechanisms of action (Bhaya, et al., 2011; Haft, et al., 2005; Makarova, et al., 2011). Briefly, CRISPR loci are composed of arrays of 24-47 bp partially palindromic, highly conserved repeats separated by variable spacers specific to the infecting DNA. CRISPR-associated (*cas*) genes are involved in spacer acquisition, expression and interference to phage or plasmid. *cas* gene operons are classified into three types, several subtypes and can target either DNA or RNA or both (Bhaya, et al., 2011). CRISPR and *cas* gene operons are proposed to be transferred between distinctly related strains by horizontal gene transfer and/or by transposons (Horvath, et al., 2009), and the later can be identified by the presence of insertion elements and transposase/mutase in its vicinity. Thus, CRISPR appear to be dynamic heritable defence systems in bacteria against

plasmids and phages that are ever fast-evolving and play important roles in the co-evolution of both bacteria and phage.

The genome of *C. autoethanogenum* is found to contain eight *cas* genes of Type-I B, all predicted to be in one operon on the antisense strand with a predicted transcription terminator at the end of *cas2* gene and it is flanked by three CRISPR arrays with a total of 93 30-bp-repeats (consensus 5'-GTTGAACCTCAACATGAGATGTATTTAAAT-3') and 90 spacers of 35-38 bp. In addition to the three CRISPR arrays flanking the *cas* genes, a putative extra CRISPR array was identified in the genome, consisting of three 55-bp-repeats and two 16-bp-spacer. Expression of *cas* genes and CRISPR arrays along with their leader sequence were studied by Reverse Transcriptase PCR (RT-PCR) and RNA-Seq during logarithmic growth under autotrophic conditions. All eight predicted *cas* genes appear to be co-expressed and from a single operon. Preliminary RNA-Seq data showed expression of all CRISPR RNAs (crRNAs), with different abundances. CRISPR spacer sequences in *C. autoethanogenum* were analyzed to identify potential target DNA sequence. A comparison of regions of DNA from putative *C. autoethanogenum* processing crRNAs from all three arrays identified the sequence 5′-ATTTAAAT-3′.

Identification and Classification of CRISPR systems in industrial relevant Clostridia.

The presence of a CRISPR system in *C. autoethanogenum* compared to *C. ljungdahlii* could provide an advantage in industrial fermentations. The *C. autoethanogenum* CRISPR system was compared to those from other industrial relevant Clostridia strains to better understand their characteristics and their potential physiological and applied roles. CRISPR systems from 14 *Clostridium* species were examined for the first time, and CRISPR elements were identified only in 8 of the 14 *Clostridium* species analysed by PILER (Edgar, 2007) and CRISPRdb (Grissa, et al., 2007). From the ABE fermentation-Clostridia examined, most lacked the CRISPR system and that might be one of the reasons why the ABE fermentation process was historically found to be prone to phage infections (Jones, et al., 2000). From the three acetogenic strains investigated only *C. autoethanogenum* had a CRISPR system

In all *Clostridium* species that harbour CRISPR arrays, *cas* genes were identified. A more detailed account on *cas* genes in different Clostridia is provided in online version of this manuscript. The *C. autoethanogenum* CRISPR repeat DNA was not found in any of the other *Clostridium* species included in this study. A search for organisms with repeats similar to *C. autoethanogenum* in CRISPRdb database resulted in *Clostridium novyi*, *Eubacterium limosum*, along with a few *Clostridium botulinum* substrains. The *cas* gene operon architecture, the arrangement of arrays on the chromosome and the presence of two hypothetical genes separating arrays 2 and 3 in *C. autoethanogenum* and *C. novyi* are strikingly alike, suggesting a common lineage of these two CRISPR-*cas* systems. This observation was further strengthened by the phylogenetic classification placing *C. autoethanogenum cas1* gene together with *cas1* genes from *C. novyi*.

Comparison of strains with/without CRISPR system to plasmids and prophages content.

Correlation between the presence of CRISPR and the occurrence of prophage or plasmids has been reported (Nozawa, et al., 2011). The putative prophage content and presence of CRISPR system was analysed, but no general trend was observed. The timeline for prophage infection (before or after acquisition of CRISPR system) could not be determined. Similarly, specific role of CRISPR in driving plasmid and phage evolution could not be determined.

3.5 Conclusion

A comparative genomic analysis revealed short-read technologies were unable to overcome *C. autoethanogenum* DSM 10061 repeat regions largely associated with nine copies of the rRNA gene operons. A previous study suggested that long single-molecule reads are sufficient to assemble most known microbial genomes based on a bioinformatics analysis of 2,267 complete genomes for bacteria and archaea and sequencing results for six bacteria (Koren, et al., 2013). The genome sequence of *C. autoethanogenum* DSM 10061 is classified as within the most complex class of bacterial genomes and a complete genome sequence was generated for it using long single-molecule reads and without the need for manual finishing. The relatively low cost to generate the PacBio data (~US$1,500) and the outcome of this study support the assertion this technology will be valuable in future studies where a complete genome sequence is important and for complex genomes that contain large repeat elements. Since the publication of our original report, there are more than hundred complete genome sequences have been obtained using only the PacBio data and complete list is available on Pacific Biosciences website under the scientific publications list.

Clostridia are known for their substrate and metabolic flexibility, which makes them attractive biocatalysts for biofuel and biorefinery applications (Tracy, et al., 2012). Acetogenic Clostridia such as *C. autoethanogenum* are of interest due to their abilities to ferment abundant syngas or waste gases to useful products (Tracy, et al., 2012). The *C. autoethanogenum* genome sequence will facilitate strain development for biofuels and biochemicals production and comparative genomics in the future. A comparison between *C. autoethanogenum* and *C. ljungdahlii* identified distinct differences, notably the presence of a CRISPR system, an additional *C. autoethanogenum* hydrogenase, and several differences in central metabolism, although the two bacteria likely descent from a common ancestor. Comparative genomic analysis and characterization of CRISPR, plasmid content and prophage among Clostridia with biotechnological interest was performed. Notably, the classic ABE fermentation strains *C. acetobutylicum* and *C. beijeinckii* are reported to be prone to bacteriophage infections (Tracy, et al., 2012) and all lack a CRISPR system and only one of the analysed 14 strains contain both a plasmid and a CRISPR system. From the acetogenic *Clostridium* strains sequenced to date, only *C. autoethanogenum* possesses a CRISPR system. Further consideration of Clostridia CRISPR systems may be informative for bioprocess development strategies and for ecological studies.

References

Abrini, J., Naveau, H. and Nyns, E.J. (1994) *Clostridium autoethanogenum*, sp. nov., an anaerobic bacterium that produces ethanol from carbon monoxide, *Arch. Microbiol.*, **161**, 345-351.

Bankevich, A.*, et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing., *J. Comput. Biol.*, **19**, 455-477.

Bao, G.*, et al.* (2011) Complete genome sequence of *Clostridium acetobutylicum* DSM 1731, a solvent-producing strain with multireplicon genome architecture, *J. Bacteriol.*, **193**, 5007-5008.

Bashir, A.*, et al.* (2012) A hybrid approach for the automated finishing of bacterial genomes, *Nat. Biotechnol.*, **30**, 701-707.

Bhaya, D., Davison, M. and Barrangou, R. (2011) CRISPR-Cas systems in Bacteria and Archaea: Versatile small RNAs for adaptive defense and regulation, *Ann. Rev. Genetics*, **45**, 273-297.

Brouns, S.J.J.*, et al.* (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes, *Science*, **321**, 960-964.

Brown, S.D.*, et al.* (2011) Genome sequence of the mercury-methylating strain *Desulfovibrio desulfuricans* ND132, *J. Bacteriol.*, **193**, 2078-2079.

Brown, S.D.*, et al.* (2011) Mutant alcohol dehydrogenase leads to improved ethanol tolerance in *Clostridium thermocellum*, *Proc. Natl. Acad. Sci. USA*, **108**, 13752–13757.

Brown, S.D.*, et al.* (2013) Draft Genome Sequences for Three Mercury-Methylating, Sulfate-Reducing Bacteria, *Genome Announcements*, **1**, e00618-00613.

Brown, S.D.*, et al.* (2012) Draft genome sequence of *Rhizobium* sp. strain PDO1-076, a bacterium isolated from *Populus deltoides*, *J Bacteriol*, **194**, 2383-2384.

Brown, S.D.*, et al.* (2012) Draft genome sequences for *Clostridium thermocellum* wild-type strain YS and derived cellulose adhesion-defective mutant strain AD2, *J Bacteriol*, **194**, 3290-3291.

Brown, S.D.*, et al.* (2012) Draft genome sequences for two metal-reducing *Pelosinus fermentans* strains isolated from a Cr(VI)-contaminated site and for type strain R7, *J. Bacteriol.*, **194**, 5147-5148.

Bruno-Barcena, J.M., Chinn, M.S. and Grunden, A.M. (2013) Genome Sequence of the Autotrophic Acetogen *Clostridium autoethanogenum* JA1-1 Strain DSM 10061, a

Producer of Ethanol from Carbon Monoxide, *Genome Announcements*, **1**, (4):e00628-00613.

Carte, J.*, et al.* (2008) Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes, *Genes Dev.*, **22**, 3489-3496.

Chain, P.S.*, et al.* (2009) Genomics. Genome project standards in a new era of sequencing, *Science*, **326**, 236-237.

Chen, F.*, et al.* (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups, *Nucleic Acids Res.*, **34**, D363-368.

Chin, C.*, et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, *Nat Methods*, **10**, 563 - 569.

Cotter, J.L., Chinn, M.S. and Grunden, A.M. (2009) Ethanol and acetate production by *Clostridium ljungdahlii* and *Clostridium autoethanogenum* using resting cells, *Bioproc. Biosyst.Eng.*, **32**, 369-380.

Cotter, J.L., Chinn, M.S. and Grunden, A.M. (2009) Influence of process parameters on growth of *Clostridium ljungdahlii* and *Clostridium autoethanogenum* on synthesis gas, *Enz. and Microbial Technol.*, **44**, 281-288.

Darling, A.E., Mau, B. and Perna, N.T. (2010) progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement, *PLoS One*, **5**, e11147.

del Cerro, C.*, et al.* (2013) Genome sequence of the butanol hyperproducer *Clostridium saccharoperbutylacetonicum* N1-4, *Genome Announcements*, **1**.

Edgar, R. (2007) PILER-CR: Fast and accurate identification of CRISPR repeats, *BMC Bioinformatics*, **8**, 18.

Eid, J.*, et al.* (2009) Real-time DNA sequencing from single polymerase molecules, *Science*, **323**, 133 - 138.

Elkins, J.G.*, et al.* (2010) Complete genome sequence of the cellulolytic thermophile *Caldicellulosiruptor obsidiansis* OB47T, *J. Bacteriol.*, **192**, 6099-6100.

English, A.C.*, et al.* (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology, *PLoS One*, **7**, e47768.

Feinberg, L.*, et al.* (2011) Complete genome sequence of the cellulolytic thermophile *Clostridium thermocellum* DSM1313, *J. Bacteriol.*, **193**, 2906-2907.

Fouts, D.E. (2006) Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences, *Nucleic Acids Res.*, **34**, 5839-5851.

Grissa, I., Vergnaud, G. and Pourcel, C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats, *BMC Bioinformatics*, **8**, 172.

Gurevich, A.*, et al.* (2013) QUAST: quality assessment tool for genome assemblies, *Bioinformatics*, **29**, 1072-1075.

Haft, D.H.*, et al.* (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes, *PloS Comp. Biol.*, **1**, 474-483.

Hale, C.R.*, et al.* (2012) Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs, *Mol. Cell*, **45**, 292-302.

Hoefler, B.C., Konganti, K. and Straight, P.D. (2013) De novo Assembly of the *Streptomyces* sp. Strain Mg1 Genome Using PacBio Single-Molecule Sequencing, *Genome Announcements*, **1**, 1:e00535-00513.

Horvath, P.*, et al.* (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes, *Int. J. Food Microbiol.*, **131**, 62-70.

Hu, S.*, et al.* (2011) Comparative genomic and transcriptomic analysis revealed genetic characteristics related to solvent formation and xylose utilization in *Clostridium acetobutylicum* EA 2018, *BMC Genomics*, **12**, 93.

Hunt, M.*, et al.* (2013) REAPR: a universal tool for genome assembly evaluation, *Genome Biology*, **14**, R47.

Hyatt, D.*, et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, **11**, 119.

Jones, D.T.*, et al.* (2000) Bacteriophage infections in the industrial acetone butanol (AB) fermentation process, *J. Mol. Microbiol. Biotechnol.*, **2**, 21-26.

Kisand, V. and Lettieri, T. (2013) Genome sequencing of bacteria: sequencing, *de novo* assembly and rapid analysis using open source tools, *BMC Genomics*, **14**, 211.

Kopke, M.*, et al.* (2010) *Clostridium ljungdahlii* represents a microbial production platform based on syngas, *Proc. Natl Acad. Sci. USA*, **107**, 13087-13092.

Köpke, M.*, et al.* (2011) Fermentative production of ethanol from carbon monoxide, *Curr. Opin. Biotechnol.*, **22**, 320-325.

Köpke, M.*, et al.* (2011) 2,3-butanediol production by acetogenic bacteria, an alternative route to chemical synthesis, using industrial waste gas, *Appl. Environ. Microbiol.* , **77**, 5467-5475.

Köpke, M., Straub, M. and Dürre, P. (2013) *Clostridium difficile* is an autotrophic bacterial pathogen, *PLoS One*, **8**.

Koren, S.*, et al.* (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing, *Genome Biology*, **14**, R101.

Koren, S.*, et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads, *Nat Biotechnol*, **30**, 693 - 700.

Krzywinski, M.*, et al.* (2009) Circos: An information aesthetic for comparative genomics, *Genome Res.*, **19**, 1639-1645.

Kurtz, S.*, et al.* (2004) Versatile and open software for comparing large genomes, *Genome biology*, **5**, R12.

Lillestøl, R.K.*, et al.* (2006) A putative viral defence mechanism in archaeal cells, *Archaea*, **2**, 59–72.

Lillestol, R.K.*, et al.* (2009) CRISPR families of the crenarchaeal genus *Sulfolobus*: bidirectional transcription and dynamic properties, *Molecular Microbiology*, **72**, 259-272.

Magoc, T.*, et al.* (2013) GAGE-B: An Evaluation of Genome Assemblers for Bacterial Organisms, *Bioinformatics*, **29**, 1718-1725.

Makarova, K.S.*, et al.* (2011) Evolution and classification of the CRISPR-Cas systems, *Nature Rev. Microbiol.*, **9**, 467-477.

Mardis, E.R. (2008) Next-Generation DNA Sequencing Methods, *Annual Review of Genomics and Human Genetics*, **9**, 387-402.

Marraffini, L.A. and Sontheimer, E.J. (2008) CRISPR Interference Limits Horizontal Gene Transfer in *Staphylococci* by Targeting DNA, *Science*, **322**, 1843-1845.

Mavromatis, K.*, et al.* (2012) The Fast Changing Landscape of Sequencing Technologies and Their Impact on Microbial Genome Assemblies and Annotation, *PLoS ONE*, **7**, e48837.

Medini, D.*, et al.* (2008) Microbiology in the post-genomic era, *Nat Rev Micro*, **6**, 419-430.

Mohammadi, M.*, et al.* (2011) Bioconversion of synthesis gas to second generation biofuels: A review, *Renewable & Sustainable Energy Reviews*, **15**, 4255-4273.

Munasinghe, P.C. and Khanal, S.K. (2010) Biomass-derived syngas fermentation into biofuels: Opportunities and challenges, *Biores. Technol.*, **101**, 5013-5022.

Nagarajan, N. and Pop, M. (2013) Sequence assembly demystified, *Nat. Rev. Genet.*, **14**, 157-167.

Nölling, J.*, et al.* (2001) Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*, *J. Bacteriol.*, **183**, 4823-4838.

Nozawa, T.*, et al.* (2011) CRISPR Inhibition of Prophage Acquisition in *Streptococcus pyogenes*, *PLoS One*, **6**.

Paul, D.*, et al.* (2010) Genome sequence of the solvent-producing bacterium *Clostridium carboxidivorans* strain P7^{T}, *J. Bacteriol.*, **192**, 5554-5555.

Poehlein, A.*, et al.* (2013) Complete Genome Sequence of the Solvent Producer *Clostridium saccharobutylicum* NCP262 (DSM 13864), *Genome Announcements*, **1**.

Pougach, K.*, et al.* (2010) Transcription, processing and function of CRISPR cassettes in *Escherichia coli*, *Mol. Microbiol.*, **77**, 1367-1379.

Powers, J.G.*, et al.* (2013) Efficient and accurate whole genome assembly and methylome profiling of *E. coli*, *BMC Genomics*, **14**, 675.

Quail, M.*, et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, *BMC Genomics*, **13**, 341.

Rahman, A. and Pachter, L. (2013) CGAL: computing genome assembly likelihoods, *Genome Biology*, **14**, R8.

Rasko, D.A.*, et al.* (2011) Origins of the *E. coli* strain causing an outbreak of hemolytic–uremic syndrome in Germany, *New Eng. J. Med.*, **365**, 709-717.

Richter, H.*, et al.* (2012) Characterization of CRISPR RNA processing in *Clostridium thermocellum* and *Methanococcus maripaludis*, *Nucleic Acids Research*, **40**, 9887-9896.

Roberts, R., Carneiro, M. and Schatz, M. (2013) The advantages of SMRT sequencing, *Genome Biology*, **14**, 405.

Salzberg, S.L.*, et al.* (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms, *Genome Res*, **22**, 557-567.

Simpson, J.T.*, et al.* (2009) ABySS: A parallel assembler for short read sequence data, *Genome Res.*, **19**, 1117-1123.

Sorek, R., Kunin, V. and Hugenholtz, P. (2008) CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea, *Nat. Rev.Microbiol.*, **6**, 181-186.

Stackebrandt, E.*, et al.* (1999) Phylogenetic basis for a taxonomic dissection of the genus *Clostridium*, *FEMS Immunology & Medical Microbiology*, **24**, 253-258.

Tamaru, Y.*, et al.* (2010) Genome sequence of the cellulosome-producing mesophilic organism *Clostridium cellulovorans* 743B, *J. Bacteriol.*, **192**, 901-902.

Tanner, R.S., Miller, L.M. and Yang, D. (1993) *Clostridium ljungdahlii* sp. nov., an acetogenic species in Clostridial rRNA homology group I, *Int. J. Syst. Bacteriol.*, **43**, 232-236.

Tirado-Acevedo, O., Chinn, M.S. and Grunden, A.M. (2010) Production of Biofuels from Synthesis Gas Using Microbial Catalysts. In Laskin, A.I., Sariaslani, S. and Gadd, G.M. (eds), *Advances in Applied Microbiology, Vol 70*. pp. 57-92.

Tirado-Acevedo, O.*, et al.* (2011) Influence of carbon source preadaptation on *Clostridium ljungdahlii* growth and product formation, *J. Bioprocess. Biotechniq.*, **S2**, 001.

Tracy, B.P.*, et al.* (2012) Clostridia: the importance of their exceptional substrate and metabolite diversity for biofuel and biorefinery applications, *Curr. Opin. Biotechnol.*, **23**, 364-381.

Treangen, T.J.*, et al.* (2009) Genesis, effects and fates of repeats in prokaryotic genomes, *FEMS Microbiology Reviews*, **33**, 539-571.

Treangen, T.J. and Salzberg, S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions, *Nat Rev Genet*, **13**, 146-146.

Utturkar, S.M.*, et al.* (2013) Draft Genome Sequence for *Caulobacter* sp. Strain OR37, a Bacterium Tolerant to Heavy Metals, *Genome Announc*, **1**, e00322-00313

Utturkar, S.M.*, et al.* (2013) Draft Genome Sequence for *Ralstonia* sp. Strain OR214, a Bacterium with Potential for Bioremediation, *Genome Announc*, **1**, e00321-00313.

Vezzi, F., Narzisi, G. and Mishra, B. (2012) Feature-by-feature – evaluating *de novo* sequence assembly, *PLoS One*, **7**, e31002.

Wang, S.*, et al.* (2013) NADP-specific electron-bifurcating [FeFe]-hydrogenase in a functional complex with formate dehydrogenase in *Clostridium autoethanogenum* grown on CO, *J. Bacteriol.*, **195**, 4373-4386.

Wilson, C.M.*, et al.* (2013) Global transcriptome analysis of *Clostridium thermocellum* ATCC 27405 during growth on dilute acid pretreated Populus and switchgrass, *Biotechnol. Biofuels*, **6**, 179.

Wood, H.G. (1991) Life with CO or CO_2 and H_2 as a source of carbon and energy, *FASEB J.*, **5**, 156-163.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs, *Genome Res.*, **18**, 821-829.

Zhou, Y.*, et al.* (2011) PHAST: A Fast Phage Search Tool, *Nucleic Acids Research*, **39**, W347-W352.

Appendix

Table 3.1: Sequencing statistics.

	Number of Reads	Total Bases	Mean Read Length (bp)	Longest Read (bp)	Coverage (x)
454-3kb PE	511,515	202,048,425	395	945	46x
Illumina PE	3,689,644	553,446,600	151	151	127x
PacBio	122,933	782,530,012	6,366	26,777	179x

Table 3.2: Assembly statistics for *C. autoethanogenum* strain DSM 10061.

	# Contigs	Largest Contig (bp)	Contig N50 (bp)	Genome Size (Mb)	Scaffolds	Largest Scaffold (bp)	Scaffold N50 (bp)	Assembler
454/Ion Torrent *	100	436,795	115,901	4.32	NA	NA	NA	Newbler 2.6
Illumina only	57	460,940	255,482	4.3	53	769,812	328,660	Velvet 1.2
454 only	32	134,546	330,116	4.3	13	1,137,876	898,466	Newbler 2.8
Illumina/ 454 Hybrid	22	1,137,625	687,076	4.3	13	1,137,625	899,926	Newbler 2.8
PacBio	1	4,352,205	4,352,267	4.3	1	4,352,267	4,352,267	SMRT 2.0

*Previously published as a 4.5 Mb draft genome (Bruno-Barcena, et al., 2013), but present in Genbank (ASZX00000000.1) as 4,323,309 bp.

Table 3.3: CGAL scores for *C. autoethanogenum* DSM 10061 assemblies.

Assembly	**CGAL Score**	**CGAL Score (formatted)**
Illumina_Only	-49339432.8	-4.93E+07
454_Hybrid	-52049311.37	-5.20E+07
454_Only	-52511662.82	-5.25E+07
Draft	-54157668.31	-5.42E+07
PacBio	-56209788.73	-5.62E+07

Table 3.4: QUAST analysis of *C. autoethanogenum* DSM 10061 assemblies.

Assembly	**Illumina_Only**	**NCBI Draft**	**454Only**	**454Hybrid**
# contigs (>= 0 bp)	57	100	32	22
# contigs (>= 1000 bp)	47	96	30	21
Total length (>= 0 bp)	4311676	4323309	4305482	4308316
Total length (>= 1000 bp)	4303892	4319422	4303912	4307500
# contigs	57	100	32	22
Total length	4311676	4323309	4305482	4308316
Largest contig	460940	436795	639527	1137625
Reference length	4352205	4352205	4352205	4352205
GC (%)	30.92	30.97	30.91	30.92
Reference GC (%)	31.09	31.09	31.09	31.09
N50	255482	115901	330116	687076
NG50	255482	115901	330116	687076
N75	114708	65006	110889	224907
NG75	114708	64087	110889	224907
L50	7	12	5	3
LG50	7	12	5	3
L75	12	23	11	6
LG75	12	24	11	6
# misassemblies	3	0	0	0
Misassembled contigs length	362096	0	0	0
# local misassemblies	3	0	0	0
# unaligned contigs	2 + 0 part	21 + 1 part	1 + 0 part	1 + 0 part
Unaligned contigs length	11033	39942	5499	5499
Genome fraction (%)	98.764	98.407	98.784	98.853
Duplication ratio	1.001	1	1	1

Table 3.4 continued …

Assembly	**Illumina_Only**	**NCBI Draft**	**454Only**	**454Hybrid**
# N's per 100 kb	0	0	0.12	0
# mismatches per 100 kb	1.26	0.16	1.37	0.98
# indels per 100 kb	5.65	6.07	6.65	6.16
Largest alignment	460756	436795	639307	1137445
NA50	246708	115901	330116	687032
NGA50	246708	115901	330116	687032
NA75	112457	65006	110889	224907
NGA75	112457	64087	110889	224907
LA50	7	12	5	3
LGA50	7	12	5	3
LA75	13	23	11	6
LGA75	13	24	11	6

Table 3.5: REAPR analysis of *C. autoethanogenum* DSM 10061 assemblies.

Assembly	Total Length	Gaps	Total Gap Length	Original Cotigs	Original N50	Corrected Contigs	Corrected N50	Detected Errors	FCD Errors	Low Coverage Error	Error Free Bases (%)	Warnings and Notes
Illumina_ only	4311676	1	512	57	255482	57	255482	2	1	1	97.23	182 warnings: Low score regions: 0 Links: 95 Soft clip: 2 Collapsed repeats: 4 Low read coverage: 0 Low perfect coverage: 81 Wrong read orientation: 0
NCBI_Draft	4323309	0	0	100	115901	100	115901	0	0	0	97.02	190 warnings: Low score regions: 0 Links: 112 Soft clip: 2 Collapsed repeats: 4 Low read coverage: 0 Low perfect coverage: 70 Wrong read orientation: 2

Table 3.5 continued …

Assembly	Total Length	Gaps	Total Gap Length	Original Cotigs	Original N50	Corrected Contigs	Corrected N50	Detected Errors	FCD Errors	Low Coverage Error	Error Free Bases (%)	Warnings and Notes
454_Hybrid	4308316	0	0	22	687076	22	687076	2	2	0	98.6	83 warnings: Low score regions: 0 Links: 29 Soft clip: 4 Collapsed repeats: 5 Low read coverage: 0 Low perfect coverage: 45 Wrong read orientation: 0
Pacbio	4352267	0	0	1	4352267	1	4352267	1	1	0	98.44	96 warnings: Low score regions: 1 Links: 0 Soft clip: 0 Collapsed repeats: 0 Low read coverage: 1 Low perfect coverage: 94 Wrong read orientation: 0 FCD Error Location for PacBio: 3872494-3873407 - (913 bp) Coverage @ this region Illumina - 40x 454 - 19x Pacbio 108x Sanger_Coverage - first 392 bp

Table 3.6: Regions of low sequence coverage.

Locus tag	Start[a]	End[a]	Product Description	PacBio Coverage (x[b])	454 Coverage (x)	Illumina Coverage (x)	454 Hybrid Contig Coverage[c]	Draft Assembly Contig Coverage
CAETHG_0145	156117	156914	Methionine synthase	87	26	62	Complete	Partial
CAETHG_0152	161167	161292	hypothetical protein	94	16	55	Complete	Partial
CAETHG_0153	161313	161963	dihydropteroate synthase DHPS	93	22	46	Complete	Partial
CAETHG_0433	472649	474331	transcriptional regulator, PucR family	110	25	57	Complete	Partial
CAETHG_0601	661798	663339	citrate lyase, alpha subunit	109	25	64	Partial	Partial
CAETHG_0602	663332	664234	citrate lyase, beta subunit	111	29	65	None	None
CAETHG_0603	664234	664530	Citrate lyase acyl carrier protein	107	29	63	None	None
CAETHG_0604	664553	665587	citrate lyase ligase	109	23	63	None	Partial
CAETHG_0605	665628	666806	malic protein NAD-binding protein	101	27	69	None	None
Intergenic	827340	827520	NA	106	30	53	None	None
CAETHG_0774	832108	833028	SufBD protein	109	23	65	Complete	Partial
CAETHG_0814	873533	874333	hypothetical protein	106	23	69	Complete	None
CAETHG_0815	874375	874953	hypothetical protein	102	23	55	Complete	None
rRNA	885055	887942	23s_rRNA	87	77	147	None	None
rRNA	888206	889703	16s_rRNA	102	56	165	None	None
CAETHG_0871	940541	941353	3-dehydroquinate dehydratase	109	27	59	Complete	Partial
CAETHG_1038	1116305	1121431	cell wall binding repeat 2-containing protein	127	27	69	Partial	None
CAETHG_1052	1136476	1138017	citrate lyase, alpha subunit	107	22	53	Partial	None
CAETHG_1053	1138010	1138912	citrate lyase, beta subunit	106	29	75	Complete	None
CAETHG_1054	1138912	1139208	Citrate lyase acyl carrier protein	109	37	70	Complete	None
CAETHG_1055	1139370	1140533	malic protein NAD-binding protein	107	27	51	Partial	Partial

Table 3.6 continued …

Locus tag	Start[a]	End[a]	Product Description	PacBio Coverage (x[b])	454 Coverage (x)	Illumina Coverage (x)	454 Hybrid Contig Coverage[c]	Draft Assembly Contig Coverage
Intergenic	1148600	1148780	NA	131	16	63	Complete	None
CAETHG_1100	1186843	1187643	hypothetical protein	118	23	68	Complete	None
CAETHG_1101	1187685	1188263	hypothetical protein	105	28	59	Complete	None
CAETHG_1630	1752229	1753149	SufBD protein	118	26	79	Complete	Partial
CAETHG_1634	1755642	1756505	modD protein	115	22	69	Complete	Partial
CAETHG_1708	1841018	1841572	Lumazine-binding	132	23	66	Complete	Complete
CAETHG_1816	1956238	1956534	microcompartments protein	138	35	76	Complete	Partial
CAETHG_1817	1956609	1956899	microcompartments protein	139	19	81	Complete	None
CAETHG_1818	1956948	1957598	Propanediol utilization protein	144	24	74	Complete	None
CAETHG_1819	1957600	1959153	acetaldehyde dehydrogenase (acetylating)	153	25	67	Complete	None
CAETHG_1826	1963196	1964038	ethanolamine utilization protein EutJ family protein	161	34	73	Complete	Partial
CAETHG_1827	1964020	1964790	hypothetical protein	162	22	68	Complete	Partial
CAETHG_1949	2079078	2080271	hypothetical protein	161	30	79	Complete	Partial
CAETHG_1963	2095013	2096206	hypothetical protein	128	36	97	Complete	Partial
tRNA	2113813	2113886	tRNA_Met	128	15	61	None	Complete
rRNA	2114155	2117042	23s_rRNA	122	81	161	None	None
rRNA	2117334	2118831	16s_rRNA	118	66	128	None	None
tRNA	2135117	2135189	tRNA_Met	132	22	64	Complete	None
tRNA	2135201	2135286	tRNA_Leu	133	16	59	Complete	None
tRNA	2135301	2135374	tRNA_Met	133	17	57	Complete	None
tRNA	2135394	2136466	tRNA_Met	139	35	74	Complete	None
tRNA	2135478	2135563	tRNA_Leu	140	30	62	Complete	None

Table 3.6 continued …

Locus tag	Start[a]	End[a]	Product Description	PacBio Coverage (x[b])	454 Coverage (x)	Illumina Coverage (x)	454 Hybrid Contig Coverage[c]	Draft Assembly Contig Coverage
CAETHG_2076	2220169	2221506	sigma54 specific transcriptional regulator, Fis family	122	32	85	Partial	Partial
CAETHG_2077	2221658	2221885	transcriptional regulator, Fis family	126	21	92	Partial	None
CAETHG_2078	2222014	2222994	putative sigma54 specific transcriptional regulator	135	30	77	Partial	Partial
rRNA	2271738	2273235	16s_rRNA	165	10	26	None	None
rRNA	2273527	2276414	23s_rRNA	158	10	26	None	None
tRNA	2276744	2276817	tRNA_Met	153	28	70	None	Complete
rRNA	2355334	2356831	16s_rRNA	145	11	24	None	None
rRNA	2357123	2360010	23s_rRNA	136	13	23	None	None
tRNA	2360340	2360412	tRNA_Lys	122	15	65	Complete	Partial
rRNA	2372238	2373735	16s_rRNA	128	13	21	None	None
rRNA	2374027	2376914	23s_rRNA	126	14	19	None	None
rRNA	2392702	2394199	16s_rRNA	134	12	20	None	None
rRNA	2394596	2397483	23s_rRNA	142	11	21	None	None
CAETHG_2238	2397706	2397882	hypothetical protein	138	23	57	Partial	Complete
CAETHG_2268	2424703	2425503	Integrase catalytic region	115	26	61	Complete	None
CAETHG_2269	2425545	2426123	hypothetical protein	124	26	56	Complete	None
Intergenic	2666300	2666515	NA	145	25	69	Complete	None
Intergenic	2710650	2710840	NA	124	36	71	Complete	None
CAETHG_2526	2714747	2715550	hypothetical protein	133	28	74	Complete	Partial
Intergenic	2769840	2769880	NA	124	23	67	Complete	None
CAETHG_2620	2822788	2823741	transposase IS66	124	30	59	Partial	Complete
CAETHG_2621	2823723	2824328	Transposase IS66	127	30	52	Partial	Partial

Table 3.6 continued …

Locus tag	Start[a]	End[a]	Product Description	PacBio Coverage (x[b])	454 Coverage (x)	Illumina Coverage (x)	454 Hybrid Contig Coverage[c]	Draft Assembly Contig Coverage
rRNA	2935186	2936683	16s_rRNA	127	14	27	None	None
tRNA	2936973	2937045	tRNA_Ala	125	19	51	None	None
tRNA	2937053	2937126	tRNA_Ile	125	26	58	None	None
rRNA	2937443	2940330	23s_rRNA	117	14	28	None	None
rRNA	2966992	2968489	16s_rRNA	126	11	20	None	None
tRNA	2968779	2968851	tRNA_Ala	132	20	50	None	None
tRNA	2968859	2968932	tRNA_Ile	131	23	70	None	None
rRNA	2969222	2972109	23s_rRNA	128	10	19	None	None
CAETHG_2843	3078642	3079445	dihydropteroate synthase DHPS	152	30	66	Complete	Partial
CAETHG_2844	3079499	3080131	hypothetical protein	148	32	71	Complete	Partial
CAETHG_2848	3085939	3086742	dihydropteroate synthase DHPS	146	27	66	Complete	Partial
CAETHG_2849	3086796	3087428	hypothetical protein	139	31	75	Complete	Partial
CAETHG_3037	3301321	3302088	MCP methyltransferase, CheR-type	149	23	65	Complete	Partial
CAETHG_3075	3342748	3343524	transposase IS66	112	39	74	Complete	Partial
CAETHG_3281	3537107	3537880	hypothetical protein	109	27	55	Complete	Partial
CAETHG_3282	3537862	3538704	ethanolamine utilization protein	107	30	62	Complete	None
CAETHG_3283	3538721	3539026	microcompartments protein	103	20	65	Complete	None
CAETHG_3284	3539020	3539286	Ethanolamine utilization protein EutN/carboxysome structural protein Ccml	106	25	55	Complete	None
CAETHG_3285	3539304	3539975	Ethanolamine utilization EutQ family protein	110	29	63	Complete	None
CAETHG_3286	3540008	3540784	microcompartments protein	106	30	61	Complete	None

Table 3.6 continued …

Locus tag	Start[a]	End[a]	Product Description	PacBio Coverage (x[b])	454 Coverage (x)	Illumina Coverage (x)	454 Hybrid Contig Coverage[c]	Draft Assembly Contig Coverage
CAETHG_3287	3540833	3542350	acetaldehyde dehydrogenase (acetylating)	111	27	61	Complete	Partial
Intergenic	3848150	3848350	NA	126	34	39	Complete	None
rRNA	3872016	3873511	16s_rRNA	98	10	18	None	None
rRNA	3873937	3876824	23s_rRNA	107	14	21	None	None
CAETHG_4028	4315106	4316413	VanW family protein	98	24	66	Complete	Partial
CAETHG_4029	4316730	4319132	Collagen triple helix repeat-containing protein	94	13	38	Complete	Partial
CAETHG_4035	4325792	4326292	VanW family protein	78	21	54	Complete	Partial

[a]The genomic regions which were not assembled in 454/Draft assembly are listed above.
[b]The 'x' coverage defines the raw read coverage averaged over given coordinates.
[c]'Complete/partial' contig coverage defines whether the region was completely/partially assembled while 'None' defines that this region is missing in the respective assembly. Missing regions in either 454/Draft assembly are shown in bold.

Table 3.7: General genome statistics for DSM 10061 PacBio assembly.

Attribute	Value	% of Total
Genome size (bp)	4,352,205	100%
DNA coding region(bp)	3,679,866	84.6%
DNA G+C content (bp)	1,352,824	31.1%
DNA scaffolds	1	100.0%
CRISPR Count	3	
Insertion Sequences	4	
Riboswitches	27	
Cobalamin	7	
FMN	3	
SAM	12	
TPP	5	
Total genes	4,161	100.0%
Protein coding genes	4,042	97.1%
Pseudo genes	18	0.4%
RNA genes	101	2.4%
rRNA genes	27	0.6%
5S rRNA	9	0.2%
16S rRNA	9	0.2%
23S rRNA	9	0.2%
tRNA genes	67	1.6%
Other RNA genes (inc. tmRNA, RNaseP, SRP RNA, and 6S)	7	0.2%
Genes with function prediction	3,283	78.9%
Genes assigned to COGs	2,722	65.4%
Genes with Pfam domains	3,136	75.4%
Genes with signal peptides	242	5.8%
Genes with transmembrane helices	1,092	26.2%

The total is based on either the size of the genome in base pairs or the protein coding genes in the annotated genome.

Table 3.8: Number of genes associated with COG functional categories for DSM 10061 PacBio assembly.

Code	Value	%	Description
J	239	5.9	Translation, Ribosomal Structure and Biogenesis
K	451	11.2	Transcription
L	221	5.5	DNA Replication, Recombination and Repair
B	6	0.1	Chromatin structure and dynamics
Cellular processes			
D	146	3.6	Cell Division and Chromosome Partitioning
V	155	3.8	Defense mechanisms
T	342	8.5	Signal Transduction Mechanisms
M	381	9.4	Cell Envelope Biogenesis, Outer Membrane
N	193	4.8	Cell Motility and Secretion
U	75	1.9	Intracellular trafficking and secretion
O	234	5.8	Posttranslational Modification, Protein Turnover, Chaperones
Metabolism			
C	458	11.3	Energy production and Conversion
G	341	8.4	Carbohydrate Transport and Metabolism
E	584	14.5	Amino Acid Transport and Metabolism
F	155	3.8	Nucleotide Transport and Metabolism
H	366	9.1	Coenzyme Metabolism
I	93	2.3	Lipid Metabolism
P	311	7.7	Inorganic Ion Transport and Metabolism
Q	226	5.6	Secondary metabolites biosynthesis, transport and catabolism
Poorly characterized			
R	737	18.3	General Function Prediction Only
S	305	7.6	Function Unknown

Table 3.9: OrthoMCL analysis of *C. autoethanogenum* and *C. ljungdahlii*.

The 427 unique genes in DSM10061 (as listed by OrthoMCL) and were searched against entire *C. ljungdahlii* genome using BLASTP. Genes that does not have any hit to *C. ljungdahlii* with e-value cut-off 1e-5 were selected. This reduces the unique genes number from 427 to 221. Most gene encode hypotheticals.

Gene	Product
CAETHG_0177	stage V sporulation protein AD
CAETHG_0195	hypothetical protein
CAETHG_0270	hypothetical protein
CAETHG_0279	peptidase M29 aminopeptidase II
CAETHG_0280	AroM family protein
CAETHG_0283	Oligopeptide transporter OPT superfamily protein
CAETHG_0289	Sporulation stage 0, Spo0E-like regulatory phosphatase
CAETHG_0294	ABC transporter ATP-binding protein
CAETHG_0336	hypothetical protein
CAETHG_0453	hypothetical protein
CAETHG_0516	hypothetical protein
CAETHG_0524	hypothetical protein
CAETHG_0528	hypothetical protein
CAETHG_0549	hypothetical protein
CAETHG_0626	hypothetical protein
CAETHG_0688	hypothetical protein
CAETHG_0689	hypothetical protein
CAETHG_0693	hypothetical protein
CAETHG_0701	hypothetical protein
CAETHG_0703	hypothetical protein
CAETHG_0704	hypothetical protein
CAETHG_0705	hypothetical protein
CAETHG_0763	hypothetical protein
CAETHG_0793	hypothetical protein
CAETHG_0798	hypothetical protein
CAETHG_0809	hypothetical protein
CAETHG_0927	hypothetical protein
CAETHG_0945	hypothetical protein
CAETHG_0952	hypothetical protein
CAETHG_0953	Na(+)/H(+) antiporter nhaA
CAETHG_0954	transposase IS200-family protein
CAETHG_0960	HxlR family transcriptional regulator
CAETHG_1011	hypothetical protein
CAETHG_1017	hypothetical protein
CAETHG_1018	hypothetical protein

Table 3.9 continued …

Gene	Product
CAETHG_1019	hypothetical protein
CAETHG_1020	hypothetical protein
CAETHG_1022	hypothetical protein
CAETHG_1023	hypothetical protein
CAETHG_1024	hypothetical protein
CAETHG_1087	transglutaminase domain-containing protein
CAETHG_1093	hypothetical protein
CAETHG_1094	hypothetical protein
CAETHG_1095	hypothetical protein
CAETHG_1096	hypothetical protein
CAETHG_1106	hypothetical protein
CAETHG_1107	hypothetical protein
CAETHG_1157	hypothetical protein
CAETHG_1158	hypothetical protein
CAETHG_1213	hypothetical protein
CAETHG_1233	hypothetical protein
CAETHG_1378	hypothetical protein
CAETHG_1394	CRISPR associated protein Cas2
CAETHG_1395	CRISPR-associated protein Cas1
CAETHG_1396	Dna2/Cas4, domain of unknown function DUF83
CAETHG_1397	CRISPR-associated helicase Cas3
CAETHG_1398	CRISPR-associated protein Cas5, Hmari subtype
CAETHG_1399	CRISPR-associated protein TM1801
CAETHG_1400	hypothetical protein
CAETHG_1401	CRISPR-associated protein TM1814
CAETHG_1404	Abortive infection protein
CAETHG_1405	hypothetical protein
CAETHG_1434	hypothetical protein
CAETHG_1511	hypothetical protein
CAETHG_1636	hypothetical protein
CAETHG_1637	hypothetical protein
CAETHG_1640	Abortive infection protein
CAETHG_1642	hypothetical protein
CAETHG_1643	hypothetical protein
CAETHG_1645	hypothetical protein
CAETHG_1646	hypothetical protein
CAETHG_1647	hypothetical protein
CAETHG_1650	protein of unknown function DUF1156

Table 3.9 continued …

Gene	Product
CAETHG_1651	Protein of unknown function DUF3780
CAETHG_1652	Fn3 associated repeat
CAETHG_1653	hypothetical protein
CAETHG_1659	hypothetical protein
CAETHG_1660	hypothetical protein
CAETHG_1661	hypothetical protein
CAETHG_1662	hypothetical protein
CAETHG_1663	hypothetical protein
CAETHG_1664	hypothetical protein
CAETHG_1682	hypothetical protein
CAETHG_1696	type IV pilus assembly PilZ
CAETHG_1700	hypothetical protein
CAETHG_1706	hypothetical protein
CAETHG_1710	hypothetical protein
CAETHG_1711	peptidase M28
CAETHG_1723	membrane protein of unknown function UCP033111
CAETHG_1752	hypothetical protein
CAETHG_1803	hypothetical protein
CAETHG_1852	hypothetical protein
CAETHG_1853	hypothetical protein
CAETHG_1922	hypothetical protein
CAETHG_2012	hypothetical protein
CAETHG_2061	hypothetical protein
CAETHG_2155	hypothetical protein
CAETHG_2164	hypothetical protein
CAETHG_2338	hypothetical protein
CAETHG_2388	Benzoate membrane transport protein
CAETHG_2390	molybdopterin binding domain-containing protein
CAETHG_2391	2-oxopent-4-enoate hydratase
CAETHG_2393	hypothetical protein
CAETHG_2513	Carboxyvinyl-carboxyphosphonate phosphorylmutase
CAETHG_2556	Zinc finger, YgiT-type
CAETHG_2560	Conserved hypothetical protein CHP00245
CAETHG_2561	transposase IS200-family protein
CAETHG_2603	hypothetical protein
CAETHG_2605	hypothetical protein
CAETHG_2608	WxcM-like domain-containing protein
CAETHG_2612	hypothetical protein

Table 3.9 continued …

Gene	Product
CAETHG_2614	hypothetical protein
CAETHG_2647	hypothetical protein
CAETHG_2648	hypothetical protein
CAETHG_2649	Pilus assembly protein PilO
CAETHG_2650	hypothetical protein
CAETHG_2651	hypothetical protein
CAETHG_2652	hypothetical protein
CAETHG_2653	Type 4 fimbrial biogenesis protein PilX, N-terminal domain
CAETHG_2668	hypothetical protein
CAETHG_2672	putative esterase
CAETHG_2676	hypothetical protein
CAETHG_2702	hypothetical protein
CAETHG_2736	hypothetical protein
CAETHG_2856	methyl-accepting chemotaxis sensory transducer
CAETHG_2857	4HB MCP domain
CAETHG_2901	sporulation protein YqfD
CAETHG_2944	hypothetical protein
CAETHG_3216	hypothetical protein
CAETHG_3366	hypothetical protein
CAETHG_3380	hypothetical protein
CAETHG_3435	hypothetical protein
CAETHG_3458	hypothetical protein
CAETHG_3484	Protein of unknown function DUF3793
CAETHG_3517	protein of unknown function DUF1254
CAETHG_3518	hypothetical protein
CAETHG_3522	hypothetical protein
CAETHG_3523	hypothetical protein
CAETHG_3530	hypothetical protein
CAETHG_3531	hypothetical protein
CAETHG_3532	hypothetical protein
CAETHG_3533	hypothetical protein
CAETHG_3534	peptidase A24A prepilin type IV
CAETHG_3535	hypothetical protein
CAETHG_3536	His-Xaa-Ser system radical SAM maturase HxsC
CAETHG_3538	hypothetical protein
CAETHG_3539	hypothetical protein
CAETHG_3540	hypothetical protein
CAETHG_3541	hypothetical protein

Table 3.9 continued …

Gene	Product
CAETHG_3542	metallophosphoesterase
CAETHG_3543	hypothetical protein
CAETHG_3544	AAA-ATPase-like protein
CAETHG_3546	hypothetical protein
CAETHG_3549	hypothetical protein
CAETHG_3550	Resolvase domain-containing protein
CAETHG_3551	protein of unknown function DUF891
CAETHG_3558	hypothetical protein
CAETHG_3561	hypothetical protein
CAETHG_3562	Immunity protein Imm6
CAETHG_3586	protein of unknown function DUF35, rubredoxin-like zinc ribbon
CAETHG_3598	hypothetical protein
CAETHG_3667	hypothetical protein
CAETHG_3671	4Fe-4S ferredoxin iron-sulfur binding domain-containing protein
CAETHG_3672	hypothetical protein
CAETHG_3673	hypothetical protein
CAETHG_3687	hypothetical protein
CAETHG_3745	Spore germination protein
CAETHG_3753	hypothetical protein
CAETHG_3754	hypothetical protein
CAETHG_3756	hypothetical protein
CAETHG_3757	restriction endonuclease
CAETHG_3758	hypothetical protein
CAETHG_3759	hypothetical protein
CAETHG_3761	hypothetical protein
CAETHG_3762	hypothetical protein
CAETHG_3764	hypothetical protein
CAETHG_3765	hypothetical protein
CAETHG_3768	hypothetical protein
CAETHG_3769	Excinuclease ABC C subunit domain protein
CAETHG_3770	hypothetical protein
CAETHG_3771	protein of unknown function DUF4236
CAETHG_3772	hypothetical protein
CAETHG_3774	hypothetical protein
CAETHG_3775	hypothetical protein
CAETHG_3776	hypothetical protein
CAETHG_3779	hypothetical protein
CAETHG_3782	phage major capsid protein, HK97 family

Table 3.9 continued …

Gene	Product
CAETHG_3783	putative phage DNA packaging-like protein
CAETHG_3784	head-tail joining family protein
CAETHG_3785	hypothetical protein
CAETHG_3786	hypothetical protein
CAETHG_3787	hypothetical protein
CAETHG_3788	XkdM protein, phage-like element PBSX
CAETHG_3789	hypothetical protein
CAETHG_3791	hypothetical protein
CAETHG_3793	Protein of unknown function, DUF2577
CAETHG_3794	Phage-like element PBSX protein, XkdS
CAETHG_3795	Baseplate J family protein
CAETHG_3796	Protein of unknown function DUF2313
CAETHG_3797	hypothetical protein
CAETHG_3798	hypothetical protein
CAETHG_3800	hypothetical protein
CAETHG_3802	PemK family protein
CAETHG_3811	hypothetical protein
CAETHG_3973	hypothetical protein
CAETHG_3974	hypothetical protein
CAETHG_3985	Arsenical pump membrane protein
CAETHG_3986	hypothetical protein
CAETHG_3987	protein of unknown function DUF3794
CAETHG_3988	hypothetical protein
CAETHG_3991	hypothetical protein
CAETHG_3997	Cupin 2 conserved barrel domain protein
CAETHG_4003	Periplasmic binding protein domain
CAETHG_4004	deoxyribose-phosphate aldolase/phospho-2-dehydro-3-deoxyheptonate aldolase
CAETHG_4005	hypothetical protein
CAETHG_4010	Glycosyltransferase, capsule biosynthesis protein
CAETHG_4016	hypothetical protein
CAETHG_4021	hypothetical protein
CAETHG_4022	hypothetical protein
CAETHG_4052	hypothetical protein
CAETHG_4060	protein of unknown function DUF1540

Figure 3.1: Examples of preliminary PCR and Sanger sequencing studies to close DSM 10061 genome compared to PacBio assembly.

Small regions of overlap in the hybrid assembly weakly joined contigs, and were supported by PCR and Sanger data, but had insufficient support for the Newbler assembly to join contigs (A), PCR and Sanger data joined small gaps between contigs in line with predictions using *C. ljungdahlii* DSM 13528 as a reference (B), and in other examples much larger products were obtained compared to the predicted PCR product sizes (C)

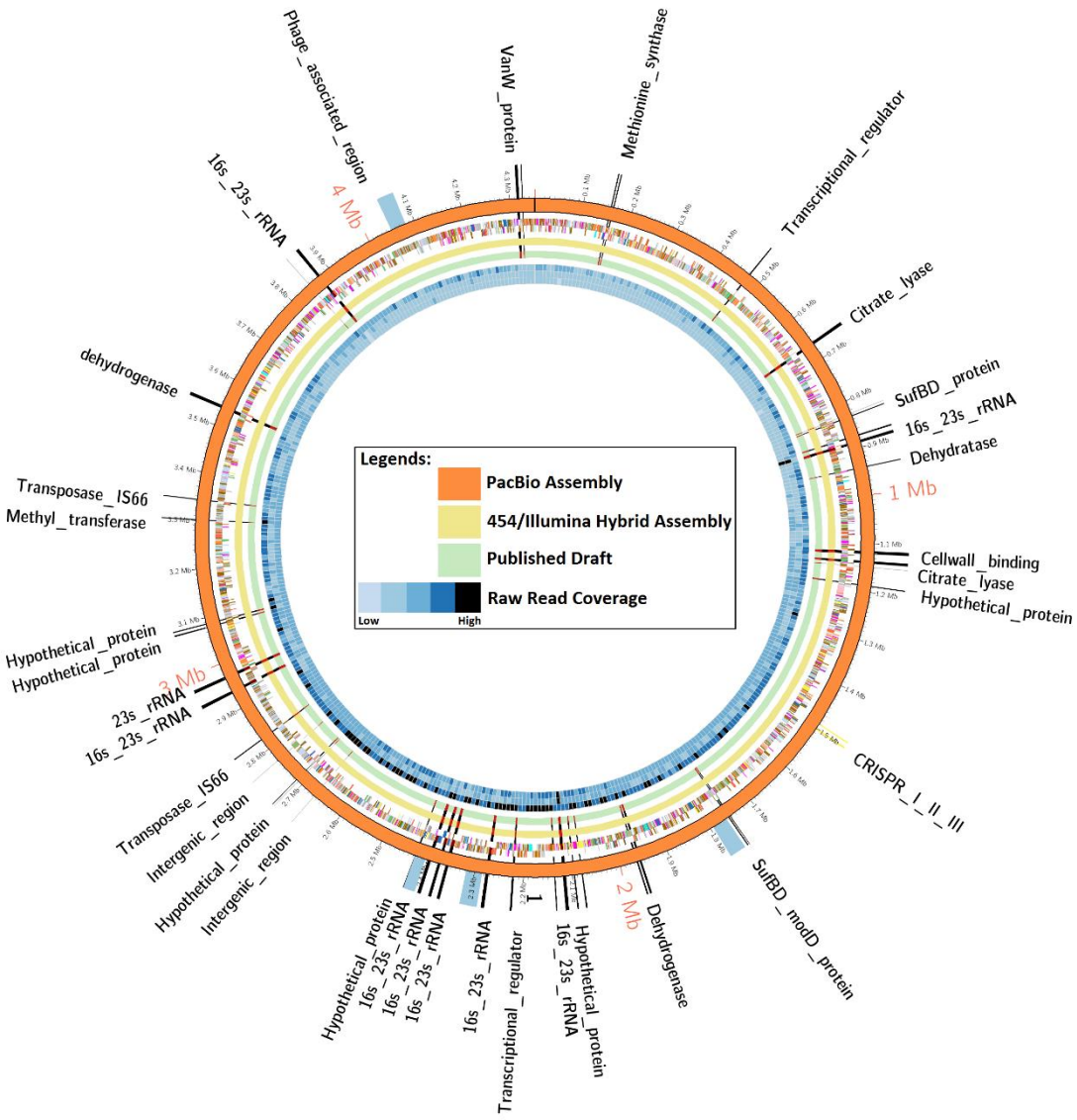

Figure 3.2: Comparison of DSM10061 genome assemblies.

The orange colored ring represents the PacBio assembly. The next inner ring represents the genes encoded on positive and negative strands respectively and color coded by COG categories. The 454/Illumina hybrid assembly and published draft assembly are represented as yellow and green circles, respectively. Next, three rings represents the raw read coverage from PacBio, 454 and Illumina technology, respectively. The gaps in the 454/Illumina hybrid assembly and published draft assembly as compared to PacBio assembly are highlighted by red colors. The key genes in the gap regions are shown by black markers while intergenic regions are shown by grey markers. The phage region and CRISPR repeats are highlighted on PacBio assembly by blue and yellow color, respectively. Additional detail is provided in Table 3.7.

CHAPTER 4 : SEQUENCE DATA FOR *CLOSTRIDIUM AUTOETHANOGENUM* USING THREE GENERATIONS OF SEQUENCING TECHNOLOGIES

Disclosure: This chapter was published as:

Utturkar S.M., Klingeman D. M., Bruno-Barcena J.M., Chinn M.S., Grunden A.M., Köpke M., Brown S. D. **(**2015). Sequence data for *Clostridium autoethanogenum* using three generations of sequencing technologies. Scientific Data. 2**:**150014.

Sagar Utturkar's contributions include bioinformatics data analysis, data deposition. Sagar Utturkar, Dr. Steven Brown and Dr. Michael Kopke conceived and designed the study and prepared the manuscript. Dawn Klingeman performed genomic DNA isolations, library preparations, and 454 and Illumina sequencing and also contributed to manuscript preparation. Dr. D.M. Bruno-Barcena J.M., Dr. Chinn M.S., Dr. Grunden A.M provided the 454 and Ion-torrent sequence data and contributed towards manuscript preparation.

4.1 Abstract

During the past decade, DNA sequencing output has been mostly dominated by the second generation sequencing platforms which are characterized by low cost, high throughput and shorter read lengths e.g. Illumina. The emergence and development of so called third generation sequencing platforms such as PacBio has permitted exceptionally long reads (over 20 kb) to be generated. Due to read length increases, algorithm improvements and hybrid assembly approaches, the concept of one chromosome, one contig and automated finishing of microbial genomes is now a realistic and achievable task for many microbial laboratories. In this paper, we describe high quality sequence datasets which span three generations of sequencing technologies, containing six types of data from four NGS platforms and originating from a single microorganism, *Clostridium autoethanogenum*. The dataset reported here will be useful for the scientific community to evaluate upcoming NGS platforms, enabling comparison of existing and novel bioinformatics approaches and will encourage interest in the development of innovative experimental and computational methods for NGS data.

4.2 Introduction

It has been a decade since the release of the initial Next Generation Sequencing (NGS) platform by 454 Life Sciences (now Roche) in 2005 (Margulies, et al., 2005). During these ten years several NGS platforms including 454, Illumina, SOLiD, Ion Torrent and Pacific Biosciences (PacBio) have been released and improved (van Dijk, et al., 2014). Currently, Illumina offers the highest throughput and the lowest per base cost (Liu, et al., 2012), while PacBio is the leader in so-called third generation sequencing technologies and offers read lengths of over 20 kb (Brown, et al., 2014). A performance comparison of various NGS platforms and recent advances are summarised (Liu, et al., 2012; Quail, et al., 2012; van Dijk, et al., 2014). In general, the second generation sequencing platforms are characterized by shorter read lengths while third generation platforms generate significantly longer, but fewer and more error prone reads.

The majority of published draft genomes have been sequenced using second generation sequencing technologies (Illumina and 454) and this data is readily available (Koren, et al., 2013). Since its introduction, the PacBio sequencing platform has become more widely used due to the utility of its longer read lengths (Roberts, et al., 2013) and range of applications (Kim, et al., 2014). A limitation for earlier versions of PacBio technology for producing accurate genome assemblies was high error rates (> 15%) and low sequence output (100 Mb) (Koren, et al., 2012). To address this, efficient algorithms were developed (Chin, et al., 2013; Koren, et al., 2012), which require either >100x PacBio sequence coverage or accurate Illumina reads for error correction. Therefore, development of hybrid approaches which utilize previous sequencing data and also provide an option to employ long-read data remains as the major scientific focus area. An evaluation of various hybrid assembly strategies was recently published in mid-2014 (Utturkar, et al., 2014) and within a short time frame the field continued to progress with the release of newer hybrid algorithms (Chengxi Ye, 2014; Hackl, et al., 2014; Lee, et al., 2014; Salmela and Rivals, 2014; Walker, et al., 2014) and updates to existing ones (English, et al., 2014; Prjibelski, et al., 2014). Generally, the hybrid sequencing strategies are more affordable and scalable especially for small-size laboratories than using the

PacBio sequencing alone (Rhoads and Au, 2015).This underlines the requirement and utility of hybrid approaches to the scientific community. The long-read PacBio platform was speculated to be increasingly used to produce finished microbial genome assemblies (Brown, et al., 2014; Koren, et al., 2013), supported by several recent examples (Brown, et al., 2014; Eckweiler, et al., 2014; Harhay, et al., 2014; Mehnaz, et al., 2014; Satou, et al., 2014) and the utility of long-read sequencing for microbial genomes has been reviewed recently (Koren and Phillippy, 2014). PacBio has the ability to detect DNA base modifications such as 4-methylcytosine (4-mC), 5-methylcytosine (5-mC) or 6-methyladenine (6-mA) (Davis, et al., 2013). This methylome information can be useful to understand biological processes such as gene expression and for optimizing transformation protocols (Lesiak, et al., 2014; Mermelstein and Papoutsakis, 1993; Pyne, et al., 2013).

Examples of former NGS platforms include Helicos Biosciences (Pushkarev, et al., 2009), and upcoming platforms include examples such Qiagen-intelligent Biosystems (Ju, et al., 2006), Oxford Nanopore (Clarke, et al., 2009), and Quantum Biosystems (BusinessWire, 2014) platforms. Oxford Nanopore has released its portable sequencer MinION, and a recent publication describes the nature of data produced (Quick, et al., 2014). Many of these newer platforms are still in the initial development stages and especially for customized methods for alignment, consensus, variant calling, *de novo* assembly and scaffolding. During the maturation of these upcoming platforms, evaluations and assessments for sequence data error rates, accuracy, length, output, cost and performance will be critical, as will the development and assessment of bioinformatics tools. Therefore, datasets which contain high-quality data from various generations of sequencing platforms for a single microorganism will be useful for others to test, compare and contrast existing and novel experimental and computational advances and benchmark automated bioinformatics pipelines.

To facilitate further assessments and tool development for current and future NGS technologies, we report and describe in detail the methods, data and quality measurements for five sequencing technologies used to sequence the biofuel producing *C. autoethanogenum* genome. This dataset represents three generations of sequencing technologies, and contains six types of data from four NGS platforms; 454 GS FLX, Illumina MiSeq, Ion Torrent, and PacBio RS-II; and Sanger sequence data. The PacBio data alone was sufficient to obtain the complete genome assembly of *C. autoethanogenum*. Several datasets were initially released into the NCBI Sequence Read Archive (SRA) with the finished *C. autoethanogenum* genome (Brown, et al., 2014). At present the NCBI SRA supports deposition of PacBio fastq files, but not the raw files required by certain software. The earlier study showed that assemblies utilizing shorter read DNA technologies were confounded by the nine copies of the 5 kb rRNA gene operons and other repetitive sequences. Raw Ion Torrent and 454 shotgun sequence data for the draft genome sequence were not been previously released (Bruno-Barcena, et al., 2013), nor were *C. autoethanogenum* DNA methylation data.

4.3 Methods

Microorganism and genomic DNA preparation

Clostridium autoethanogenum strain JA1-1 (DSMZ 10061) was obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ).

In order to prepare genomic DNA for 454 paired-end (PE), Illumina PE and PacBio sequencing the strain was cultured in PETC medium as described (Kopke, et al., 2010). A single JA1-1 colony was purified and its 16S rDNA sequence confirmed before genomic DNA was prepared for Illumina and PacBio sequencing (Kopke, et al., 2010). Genomic DNA for 454 paired-end, Illumina PE and PacBio sequencing was prepared as described previously (Brown, et al., 2014). Genomic DNA for 454 shotgun and Ion Torrent shotgun sequencing was prepared using the UltraClean Microbial DNA Isolation kit (catalog# 12224-250) from MoBio Laboratories, Inc. (Carlsbad, CA). Prior to library preparation DNA quality was assessed by Nanodrop analysis (Thermo Scientific) and visualization on an agarose gel. Quality samples have an A260/280 ratio above 1.8, and appear on a gel as a single high molecular weight band. The quantity was determined by Qubit broad range double stranded DNA assay (Life Technologies).

Illumina TruSeq Library Preparation and Sequencing

Illumina TruSeq libraries were prepared as described in the manufacturer's protocols (Part #15005180 RevA) following the low throughput protocol. In short, 3 µg of DNA was sheared to a size between approximately 200 bp and 1,000 bp by nebulization (using nitrogen as the carrier gas) for 1 min at 30 PSI. Sheared DNA was purified on a QIAquick Spin column (Qiagen). The quantity of sheared material was accessed with a broad range double stranded DNA assay from Qubit (Life Technologies) and visualized on an Agilent Bioanalyzer DNA 7500 chip (Agilent). One microgram of sheared DNA was used in the end repair reaction, and subsequently cleaned up by Agencourt AMPure XP bead purification (Beckman Coulter). The ends of the DNA were modified by adenylation of the 3' ends and Illumina adapters were then ligated to the DNA. The DNA was cleaned up using Agencourt AMPure XP beads, and samples were then run for 2 hours at 120 Volts on a 2% agarose gel containing SYBR Gold (Life Technologies). Ligation products were then purified from the sample by excising a band from the gel from approximately 350-450 bp. The DNA from the gel slice was then purified using a MinElute Gel Extraction kit (Qiagen) for each library/band. The DNA fragments were enriched by performing 10 cycles of amplification [98° C-30 sec, 10 cycles of: 98° C for 10 seconds, 60° C for 30 seconds, 72° C for 30 seconds, followed by a final extension at 72° C for 5 minutes. Amplified products were then cleaned up using Agencourt AMPure XP beads. Final libraries were validated by Qubit (Life Technologies) and visualized by Agilent Bioanalyzer for appearance and size determination. Samples were normalized using the Illumina's Library dilution calculator to a 10 nM stock, and subsequently run on an Illumina MiSeq Instrument (M02014R).

454 Shotgun Library Preparation and Sequencing

The 454 shotgun library was prepared using Roche's GS FLX Titanium Rapid Library Preparation Kit and was run on the Titanium platform according to manufacturer's specifications. Briefly, DNA was fragmented under gas pressure and the ends repaired.

Adapters were ligated onto the fragments and then small fragments were selected out of the library. The library was then assessed for quality and concentration (including size length assessment and contaminating fragments of inappropriate size) using an Agilent Bioanalyzer 2100 prior to running on the 454 instrument.

454 3 kb Library Preparation and Sequencing

A 454 3 kb paired end library was prepared following the manufacturer's instructions (Roche- Paired End Library Preparation Method Manual – 3 kb Span GS FLX Titanium Series- Oct 2009) and in detail (Yang, et al., 2012). Five micrograms of high quality, high molecular weight DNA was sheared to an average fragment size of 3 kb using a HydroShear apparatus (Genomic Solutions). The sheared material was then purified using Angencourt AMPure XP magnetic beads (Beckman Coulter). A portion of the sheared DNA was run on an Agilent Bioanalyzer 2100 to verify the size of the fragments. The fragment ends were polished and purified. The circularization adapters were appended and the product was again purified. Size selection of the material was completed followed by a fill in reaction and circularization. The sample was sheared by nebulization, purified, and checked for size on an Agilent Bioanalyzer 2100. The fragment ends were again polished and purified. The library was immobilized on Dynal M270 Streptavidin beads (Life Technologies) and the library adapters were ligated and gaps were filled. The library was amplified and a final purification step yielded a single stranded paired end library. The final library was amplified using emulsion PCR (emPCR); the products were purified, and then sequenced on a Roche 454 GS FLX system using Titanium chemistry according to the manufacturer's instructions (Roche).

SMRTbell Library Preparation and PacBio Sequencing

Ten micrograms of DNA was sheared using G-tubes (Covaris, Inc., Woburn, MA, USA), targeting 20 kb fragments. SMRTbell libraries were prepared with the DNA Template Kit 1.0 (Pacific Biosciences, Menlo Park, CA, USA) and library fragments above 4 kb were isolated using the BluePippin system (Sage Science, Inc., Beverly, MA, USA). The average SMRTbell library insert size (including adapters) was approximately 19 kb. Sequencing primers were annealed to the SMRTbell template and samples were sequenced on PacBio RS II system (2013) using Magbead loading, C2 chemistry, Polymerase version P4, and SMRT analysis software version 2.2. DNA base modifications analysis was performed by "RS Modification and Motif Analysis" workflow with default settings. Detailed information about detection of DNA base modifications workflow is available as online documentation (Pacific-BioSciences, 2014).

Ion Torrent Library Preparation and Sequencing

Genomic libraries were prepared separately for each genomic sample from 100 ng of DNA. DNA was fragmented with Ion Shear™ Plus Reagents, Ion Torrent specific adapters Ion Xpress™ P1 (5' - CCTCTCTATGGGCAGTCGGTGAT -3') and Ion Xpress™ Barcode X Adapters (5'- CCATCTCATCCCTGCGTGTCTCCGACTCAG-3') were ligated to DNA using DNA ligase (Life Technologies, Grand Island, NY). The Ion Xpress™ Barcode X Adapters contain a 10 bp sequence, Ion Xpress™ Barcode (Life Technologies, Grand Island, NY) unique to each of the samples. Ligated DNA was nick repaired using Nick Repair Polymerase ((Life Technologies, Grand Island, NY) and purified with

Agencourt® AMPure® XP Reagent (Beckman Coulter, Indianapolis, IN). The ligated and nick repaired DNA was size-selected individually with the E-GelR SizeSelect™ Agarose Gel (Life Technologies, Grand Island, NY). The size selected libraries were amplified using PlatinumR PCR SuperMix High Fidelity and Library Amplification Primer Mix ((Life Technologies, Grand Island, NY). The thermal profile for the amplification of each sample had an initial denaturing step at 94° C for 5 minutes, followed by a cycling of denaturing of 95° C for 15 seconds, annealing at 58° C for 15 seconds and a 1 minute extension at 70° C (5 cycles) and a final hold at 4° C. Each sample was again purified individually using Agencourt® AMPure® XP Reagent (Beckman Coulter, Indianapolis, IN) and standardized prior to pooling. Template-Positive Ion OneTouch™ 200 Ion Sphere™ Particles were prepared from the library pool using the Ion OneTouch™ DL system (Life Technologies, Grand Island, NY, Invitrogen division). Prepared template was sequenced on an Ion Torrent PGM instrument (Microbiome Core Facility, Chapel Hill NC) using the Ion PGM 300 Sequencing reagents and protocols ((Life Technologies, Grand Island, NY). Initial data analysis, base pair calling and trimming of each sequence was performed on an Ion Torrent browser to yield high quality reads.

4.4 Results

Data Records

Raw data from each sequencing platform was submitted to the Sequence Read Archive (SRA) at NCBI under Project ID SRP030033 [Data Citation 1]. Raw data deposited at SRA is organized by the type of sequencing platforms and corresponding accessions and file sizes are provided in Table 4.1.

Illumina sequencing instruments generate raw image files which are automatically processed through instrument control software to output sequence data in fastq format. More details about different types of data files generated by the instrument and fastq conversion steps are described in online documentation (Illumina-Inc., 2011). The 150 bp paired-end (PE) Illumina reads in fastq format were deposited to SRA with run ID SRR989790. The fastq is standard file format which can be directly used to perform several downstream applications such as *de novo* assembly or mapping to a reference genome. The 454 Pyrosequencing and Ion Torrent instrument generates the sequencing data in Standard Flowgram Format (SFF). The SRA deposition for 454 shotgun, 454 3kb PE and Ion Torrent data was made in SFF format under run ID SRR1748017, SRR989497 and SRR1748018, respectively. For validation purpose, quality statistics were determined for each short-read dataset using CLC Genomics Workbench (CLC) software version 7.5.1 and complete report is available online at external link (http://www.nature.com/article-assets/npg/sdata/2015/sdata201514/extref/sdata201514-s2.pdf)

The PacBio sequencing was performed using two SMRT cells. Each SMRT cell generates metadata.xml file which contains information about run conditions and barcodes. Three bax.h5 files containing base calls and quality information of actual sequencing data and one bas.h5 file that acts as a pointer to consolidate three bax.h5 files (Kim, et al., 2014). A typical raw read from PacBio sequencing is composed of DNA insert with both ends flanked by the adapter sequences (Kim, et al., 2014). During downstream processing

through SMRT Analysis software, the adapter sequences are removed and subreads are created which contains only the DNA sequence of interest. The PacBio filtered subreads were deposited at SRA in fastq format under run ID SRR1740585. Additionally, all the primary analysis data in the original formats as provided by the PacBio RS-II instrument is now made available on external server (Table 4.1). Methylation in bacteria generally occurs at specific sequence motifs that are recognized by methyltransferases. Genome wide analysis of DNA base modifications was performed and a high level summary of the motifs discovered is provided in Table 4.2. Additionally, "motifs_and_modifications.gff" file is provided at external link, which shows all of the sites in the genome that are methylated, all the sites with one of the discovered motifs and the overlap between the methylation and the motifs as detected by SMRT analysis software version 2.2. Prior to PacBio sequencing, a manual finishing strategy for *C. autoethanogenum* generated high-quality Sanger sequence data and it is available to download on external server (Table 4.1).

Raw reads represent the actual output from sequencing instruments. However, quality based trimming of Illumina and 454 data is recommended and often yields better results with downstream applications such as *de novo* assembly (Salzberg, et al., 2012; Utturkar, et al., 2014). On the other hand, PacBio raw read filtering to generate subreads is a necessary step to remove adapter sequences (Kim, et al., 2014). Quality based trimming of Illumina and 454 data was performed using CLC software while PacBio filtering and mapping was performed using SMRT analysis version 2.2. The post-filter summary statistics for Illumina, 454 and Ion Torrent datasets are listed in Table 3 and for PacBio dataset in Table 4. The Illumina and PacBio datasets were sequenced to sufficient high coverage (>100x) for *de novo* genome assembly while 454 and Ion Torrent dataset have coverage (<50x) which is sufficient for hybrid assembly application. See the Technical Validation section for details on quality statistics and filtering parameters used.

Technical validation

DNA and Sample Preparation

All samples were required to pass a quantity and quality assessment using a Qubit (Life Technologies), Nanodrop (ThermoFisher) and gel electrophoresis. Samples were required to have readings indicative of pure DNA and of sufficient quantity to move forward with library preparations. DNA was visualized by gel electrophoresis and was required to be high molecular weight DNA without shearing or RNA contamination.

Each sequencing library preparation method includes specific technical validation to determine quality and quantity of the final libraries to ensure high quality output from the various sequencing platforms. This technical validation typically involves assessment of the final libraries with a Qubit assay (Life Technologies) to determine quantity and visualization of the final libraries on an Agilent Bioanalyzer chip to determine quality.

Quality Determination and Analysis

To assess the quality of the libraries sequenced, we determined basic quality statistics for Illumina, 454 and Ion Torrent datasets using CLC software. This includes the calculation of sequence lengths distribution, GC-content, Ambiguous base-content, PHRED quality score distribution, nucleotide contributions, kmer distribution analysis and

sequence duplication levels. The quality statistics are calculated for every read, averaged for each dataset and provided in complete quality report (http://www.nature.com/article-assets/npg/sdata/2015/sdata201514/extref/sdata201514-s2.pdf). More than 95% of the Illumina, 454 and Ion Torrent reads have PHRED score above 20 (Figure 4.1) with a very low percentage of ambiguous bases and sequence duplication levels detected (See section 2.3 and 4.2 for each dataset - http://www.nature.com/article-assets/npg/sdata/2015/sdata201514/extref/sdata201514-s2.pdf). Quality based trimming of these short-read datasets was performed at a stringent cut-off value of 0.02. More details about the trimming algorithm used by CLC and an example can be found in online documentation(CLC, 2015). After quality trimming, only a few reads were discarded and minor changes in average read lengths were observed (Table 3). The PacBio data was processed through SMRT analysis software version 2.2. Filtering conditions applied were read quality score > 0.8, read length >500 bp, subread length >500 bp. In addition, adapter sequences were removed and ends of the reads were removed when found outside of the high-quality region (Kim, et al., 2014; Pacific-Biosciences, 2014). PacBio data retained 72% of the bases after filtering. The PacBio data by itself was sufficient to generate finished genome sequence. The complete genome sequence of *C. autoethanogenum* strain DSM10061 and *de novo* and hybrid assembly comparison using QUAST, REAPR , CGAL and Mauve tools have been described previously (Brown, et al., 2014). The Sanger sequencing data were found to be in agreement with the finished genome sequence of strain DSM10061 and provide additional validation for the high quality of PacBio dataset (Brown, et al., 2014).

To further ensure that the sequences matched with the model organism of interest, we mapped the post-filtering reads from each dataset to the model organism of interest. We used *C. autoethanogenum* DSM 10061 genome from NC_022592.1 [Data citation 3] and *C. ljungdahlii* DSM 13528 from NC_014328.1 [Data citation 4] at the NCBI Genbank as reference sequences. Since a finished genome sequence for *C. autoethanogenum* was obtained using the PacBio reads from the current dataset, we used another independent reference *C. ljungdahlii* DSM 13528 to avoid any bias. These two genomes have an average nucleotide identity score over 99%. Illumina and 454 reads were mapped to reference using the bowtie2 algorithm (Langmead and Salzberg, 2012) while PacBio reads were mapped using the BLASR algorithm (Chaisson and Tesler, 2012) from the SMRT Analysis software. The Illumina and 454 datasets have mapping rates above 90% with *C. ljungdahlii* and above 97% with the finished genome of *C. autoethanogenum*. Ion Torrent data have a comparatively lower mapping rate, 86% with *C. ljungdahlii* and 91% with *C. autoethanogenum*. For the PacBio dataset, plots showing the distributions of mapped subread concordances and coverage are shown in Figure 4.2 and provide an estimate of read agreement with reference genomes. Therefore, the data quality statistics, trimming reports and mapping results articulate the high quality of the datasets described in this manuscript.

Usage Notes

The five NGS datasets described can be downloaded from the SRA with accession numbers provided in Table 4.1. Detailed instructions for downloading each dataset from NCBI SRA and md5 checksum values are provided at (http://www.nature.com/article-

assets/npg/sdata/2015/sdata201514/extref/sdata201514-s2.pdf). The fastq/SFF formatted files from second generation sequencing data are sufficient to use for any downstream analysis using most third-party tools. On the other hand, original data formats are necessary for analysing the PacBio data through SMRT analysis software or other algorithms. Currently the SRA allows depositions of fastq formatted PacBio reads only. Therefore, all the primary analysis data in original formats as generated by the PacBio RS II instrument (*.metadata.xml, *.bas.h5, *.bax.h5 files) is available on external server (Table 4.1). The sequence IDs provided in primary analysis files are different than those available through SRA because SRA uses internal naming convention which changes existing sequence IDs. The sequence IDs in original format contain information about run and the naming convention is described in detail here (Kim, et al., 2014). Sanger data are posted at external server (Table 4.1).

Some of the datasets described here were initially released with the manuscripts describing the draft (Bruno-Barcena, et al., 2013) and finished genome of *C. autoethanogenum* (Brown, et al., 2014), with primary focus on genomic features and characteristics of this microorganism. Previous manuscripts did not include Ion torrent/454 shotgun data release and detailed quality evaluation and usage instructions were not provided. In addition, DNA modification data for *C. autoethanogenum* from the PacBio is provided, identifying three m6A adenosine methylation patterns CAAAAA'R, GWTAA'T, SNNGCAA'T. The "motifs_and_modifications.gff" file is a text file which can be opened in most of the graphical sequence viewer software. This data descriptor in Scientific Data provides an opportunity to present the collection of these five different datasets which are originated from a single microorganism and spans three generations of sequencing technologies. Here we provide the detailed characteristics for each dataset and appropriate instructions to download and use the data. Since sequencing technologies are rapidly evolving, this legacy dataset can be used as a benchmark to compare the data from newer NGS technologies and will encourage the development of new and existing hybrid algorithms.

References

Brown, S.*, et al.* (2014) Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of *Clostridium autoethanogenum* and analysis of CRISPR systems in industrial relevant Clostridia, *Biotechnol. Biofuels*, **7**, 40.

Brown, S.D.*, et al.* (2014) Complete genome sequence of *Pelosinus* sp. strain UFO1 assembled using Single-Molecule Real-Time DNA sequencing technology, *Genome Announc*, **2**.

Bruno-Barcena, J.M., Chinn, M.S. and Grunden, A.M. (2013) Genome sequence of the autotrophic acetogen *Clostridium autoethanogenum* JA1-1 strain DSM 10061, a producer of ethanol from carbon monoxide, *Genome Announc*, **1**.

BusinessWire (2014) Quantum Biosystems Demonstrates First Reads Using Quantum Single Molecule Sequencing.

Chaisson, M.J. and Tesler, G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory, *BMC Bioinformatics*, **13**, 238.

Chengxi Ye, C.H., Sergey Koren, Jue Ruan, Zhanshan (Sam)Ma, James A. Yorke, Aleksey Zimin (2014) DBG2OLC: Efficient Assembly of Large Genomes Using the Compressed Overlap Graph, *bioRxiv*.

Chin, C.S.*, et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, *Nat. Methods*, **10**, 563-569.

Clarke, J.*, et al.* (2009) Continuous base identification for single-molecule nanopore DNA sequencing, *Nat Nanotechnol*, **4**, 265-270.

CLCbio (2015) CLC Genomics Workbenach Manual - Trimming using the Trim tool.

Davis, B.M., Chao, M.C. and Waldor, M.K. (2013) Entering the era of bacterial epigenomics with single molecule real time DNA sequencing, *Curr. Opin. Microbiol.*, **16**, 192-198.

Eckweiler, D.*, et al.* (2014) Complete genome sequence of highly adherent *Pseudomonas aeruginosa* small-colony variant SCV20265, *Genome Announc*, **2**.

English, A.C., Salerno, W.J. and Reid, J.G. (2014) PBHoney: identifying genomic variants via long-read discordance and interrupted mapping, *BMC Bioinformatics*, **15**, 180.

Hackl, T.*, et al.* (2014) proovread: large-scale high-accuracy PacBio correction through iterative short read consensus, *Bioinformatics*, **30**, 3004-3011.

Harhay, G.P.*, et al.* (2014) Complete closed genome sequences of three *Bibersteinia trehalosi* nasopharyngeal isolates from cattle with shipping fever, *Genome Announc*, **2**.

Illumina-Inc. (2011) CASAVA v1.8.2 User Guide.

Ju, J.*, et al.* (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators, *Proc Natl Acad Sci U S A*, **103**, 19635-19640.

Kim, K.E.*, et al.* (2014) Long-read, whole-genome shotgun sequence data for five model organisms, *Scientific Data*, **1**.

Kopke, M.*, et al.* (2010) *Clostridium ljungdahlii* represents a microbial production platform based on syngas, *Proc Natl Acad Sci U S A*, **107**, 13087-13092.

Koren, S.*, et al.* (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing, *Genome Biol.*, **14**, R101.

Koren, S. and Phillippy, A.M. (2014) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly, *Curr. Opin. Microbiol.*, **23C**, 110-120.

Koren, S.*, et al.* (2012) Hybrid error correction and *de novo* assembly of single-molecule sequencing reads, *Nat. Biotechnol.*, **30**, 693-700.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2, *Nat. Methods*, **9**, 357-359.

Lee, H.*, et al.* (2014) Error correction and assembly complexity of single molecule sequencing reads, *bioRxiv*.

Lesiak, J.M., Liebl, W. and Ehrenreich, A. (2014) Development of an *in vivo* methylation system for the solventogen *Clostridium saccharobutylicum* NCP 262 and analysis of two endonuclease mutants, *J. Biotechnol.*, **188C**, 97-99.

Liu, L.*, et al.* (2012) Comparison of next-generation sequencing systems, *J. Biomed. Biotechnol.*, **2012**, 251364.

Margulies, M.*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, **437**, 376-380.

Mehnaz, S., Bauer, J.S. and Gross, H. (2014) Complete genome sequence of the sugar cane endophyte *Pseudomonas aurantiaca* PB-St2, a disease-suppressive bacterium with antifungal activity toward the plant pathogen *Colletotrichum falcatum*, *Genome Announc*, **2**.

Mermelstein, L.D. and Papoutsakis, E.T. (1993) *In vivo* methylation in *Escherichia coli* by the *Bacillus subtilis* phage phi 3T I methyltransferase to protect plasmids from restriction upon transformation of *Clostridium acetobutylicum* ATCC 824, *Appl. Environ. Microbiol.*, **59**, 1077-1081.

Pacific-BioSciences (2012) Detecting DNA Base Modifications.

Pacific-Biosciences (2014) Statistics Output Guide.

Prjibelski, A.D.*, et al.* (2014) ExSPAnder: a universal repeat resolver for DNA fragment assembly, *Bioinformatics*, **30**, i293-301.

Pushkarev, D., Neff, N.F. and Quake, S.R. (2009) Single-molecule sequencing of an individual human genome, *Nat. Biotechnol.*, **27**, 847-850.

Pyne, M.E.*, et al.* (2013) Development of an electrotransformation protocol for genetic manipulation of *Clostridium pasteurianum*, *Biotechnol Biofuels*, **6**, 50.

Quail, M.A.*, et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, *BMC Genomics*, **13**, 341.

Quick, J., Quinlan, A.R. and Loman, N.J. (2014) A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer, *Gigascience*, **3**, 22.

Rhoads, A. and Au, K.F. (2015) PacBio Sequencing and Its Applications, Genomics Proteomics *Bioinformatics*.

Roberts, R.J., Carneiro, M.O. and Schatz, M.C. (2013) The advantages of SMRT sequencing, *Genome Biol.*, **14**, 405.

Salmela, L. and Rivals, E. (2014) LoRDEC: accurate and efficient long read error correction, *Bioinformatics*.

Salzberg, S.L.*, et al.* (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms, *Genome Res.*, **22**, 557-567.

Satou, K.*, et al.* (2014) Complete genome sequences of eight *Helicobacter pylori* strains with different virulence factor genotypes and methylation profiles, isolated from patients with diverse gastrointestinal diseases on Okinawa Island, Japan, determined using PacBio Single-Molecule Real-Time Technology, *Genome Announc*, **2**.

Utturkar, S.M.*, et al.* (2014) Evaluation and validation of *de novo* and hybrid assembly techniques to derive high quality genome sequences, *Bioinformatics*.

van Dijk, E.L.*, et al.* (2014) Ten years of next-generation sequencing technology, *Trends Genet.*, **30**, 418-426.

Walker, B.J.*, et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, *PLoS One*, **9**, e112963.

Yang, S., Klingeman, D.M. and Brown, S.D. (2012) Ethanol-Tolerant Gene Identification in Clostridium thermocellum Using Pyro-Resequencing for Metabolic Engineering. In, *Microbial Metabolic Engineering*. pp. 111-136.

Appendix

Table 4.1: Summary of datasets accessions.

Datasets described in this manuscript, which can be accessed using the accession numbers provided.

Sequencing Platform	**Data type**	**SRA Accession/ Dryad doi**	**Size**
Accession linking all SRA data for this project		SRP030033	-
Roche 454 shotgun	Raw data in SFF format	SRR1748017	1.5 Gb
Roche 454 3 kb	Raw data in SFF format	SRR989497	1.4 Gb
Illumina	Raw data in fastq format	SRR989790	(669x2) Mb†
Ion Torrent	Raw data in SFF format	SRR1748018	858 Mb
PacBio RS II	Filtered subreads in fastq format	SRR1740585	1.2 Gb
Dryad doi linking all depositions for this project		doi: 10.5061/dryad.6fm1p	-
PacBio RS II	Raw PacBio data in tar.gz format	doi:10.5061/dryad.6fm1p/4	8.5 Gb
PacBio RS II	DNA methylation motifs in gff format	doi:10.5061/dryad.6fm1p/2	1.99 Mb
Sanger Sequencing	Chromatogram files in ABI format	doi:10.5061/dryad.6fm1p/1	4.39 Mb

†There are two files for Illumina data corresponding read_1 and read_2 for Illumina data. Detailed instructions for downloading data from SRA are provided in supplementary information.

Table 4.2: Summary of DNA methylation motif patterns discovered across the *C. autoethanogenum* genome.

Motif	Modified Position	Modification Type	% Motifs Detected	# of Motifs Detected	# of Motifs In Genome	Mean Modification QV	Mean Motif Coverage
CAAAA**A**R	6	m6A	95.44	4190	4390	68.4	56.8
GWTA**A**T	5	m6A	93.87	7975	8496	78.5	58.1
SNNGCA**A**T	7	m6A	85.27	3242	3802	75.9	57.8

Modified base within each motif is shown in bold.

Table 4.3: Summary of quality trimming statistics for Illumina, 454 and Ion Torrent data.

Sequencing Platform	Type	No. of reads	Average length	No. of reads after Trim	Average length after Trim	Total Trimmed bases	Fold Coverage
Roche 454	Singletons*	128,856	275	128,806	261	33,631,416	46x
	Paired end reads	764,756	151	764,744	144	110,124,864	
	Shotgun Data	462,052	289	458,340	249	114,126,660	26x
Ion Torrent	Single end reads	453,686	215	419,010	188	78,773,880	18x
Illumina	Paired end reads	3,689,644	150	3,682,655	149	549,756,956	126x

*The singleton sequences are generated from 454 3 kb sequencing run.

Table 4.4: Post-filter quality statistics for PacBio data.

Sequencing Platform	Type	No. of filtered subreads	N50 filtered subread length	Maximum filtered subread length	Total filtered bases	Fold Coverage
PacBio RSII	Single end reads	94,408	9,196	26,777	631,598,400	145x

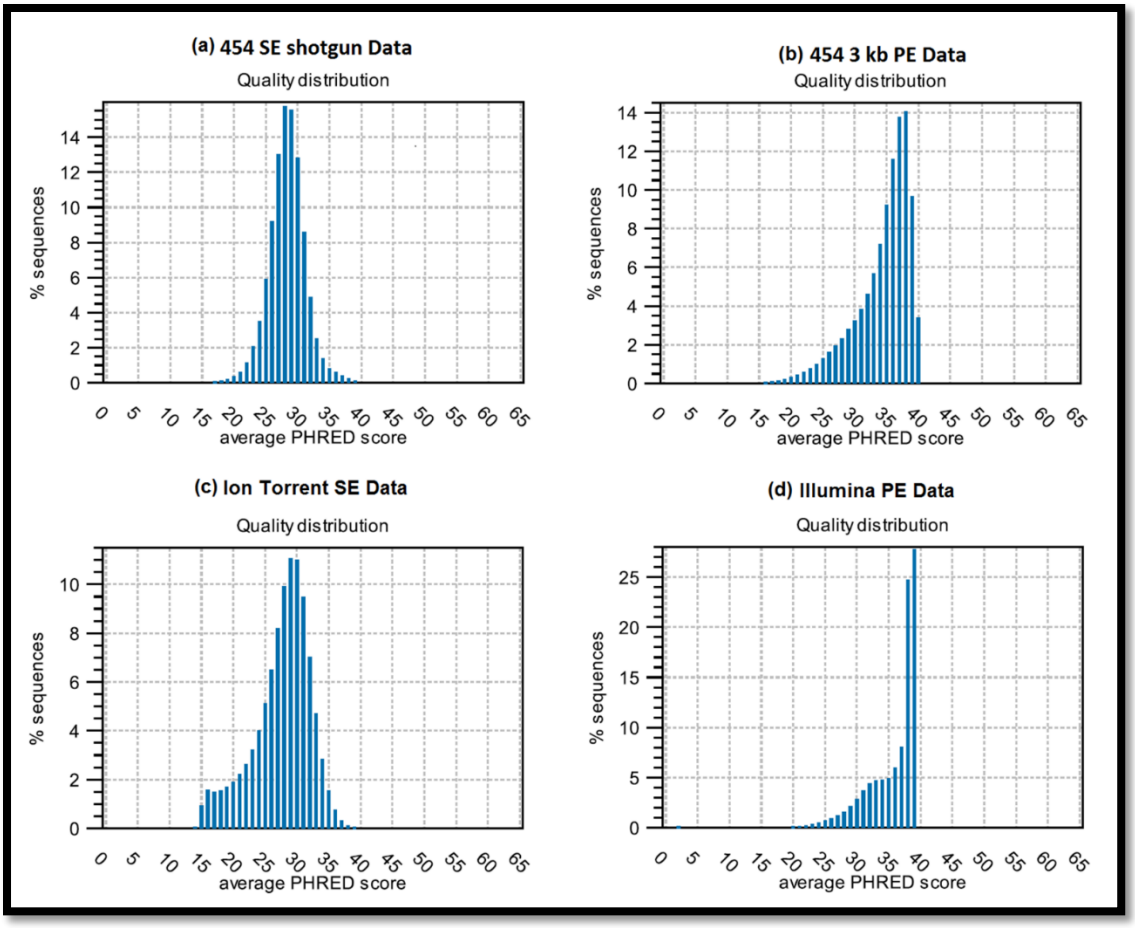

Figure 4.1: PHRED quality score distribution.

The distribution of average PHRED quality score is plotted on X-axes and percentage of sequences on Y-axes for (a) 454 single end shotgun data (b) 454 3 kb paired end data (c) Ion Torrent single end data and (d) Illumina paired end data. Quality distribution shows that more than 95% reads from each dataset have average PHRED scores above 20.

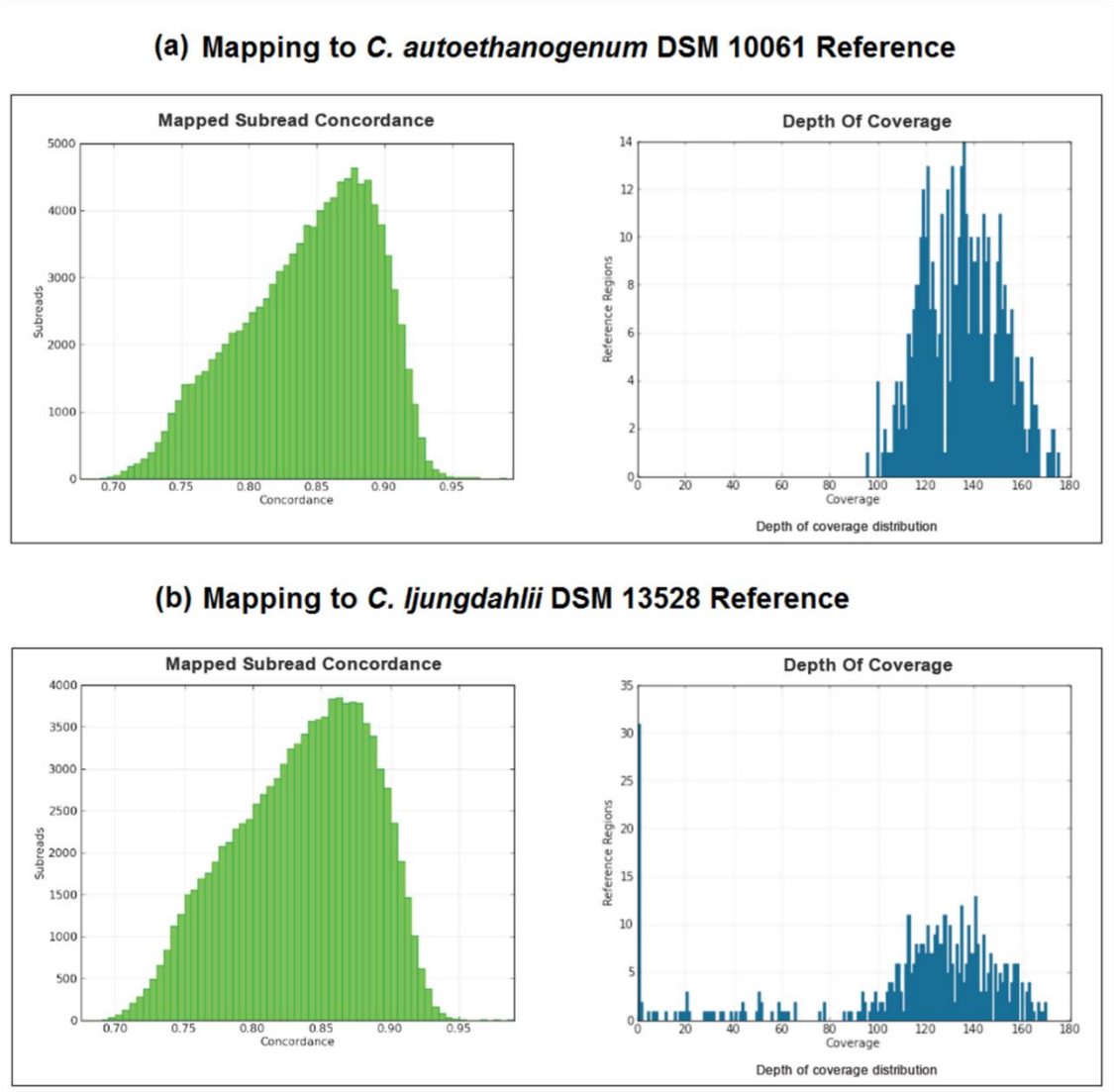

Figure 4.2: Mapped subread concordance and coverage.

The distribution of mapped subread concordances and mapped subread coverages are plotted with (a) *C. autoethanogenum* DSM 10061 finished genome and (b) *C. ljungdahlii* DSM 13528 as reference. These graphs suggest good agreement between reads and reference genomes.

CHAPTER 5 : EVALUATION OF UNASSEMBLED DNA REGIONS FROM ILLUMINA AND PACBIO SEQUENICNG PLATFORMS AND MICROBIAL GENOME FINISHING

5.1 Abstract

Development of next generation sequencing (NGS) technologies has revolutionized genomics research by providing high-throughput, low-cost sequencing methods. Despite extensive sequencing and assembly advances, there are several examples of microbial genomes that remain unfinished by PacBio and Illumina platforms even at the coverage levels estimated in range of 50x to 296x. The aim of the present study was to reveal and characterize regions of DNA which remained unassembled by either by individual or both technologies. We sequenced genomes of eight microorganisms using a combination of Illumina paired-end (PE) and PacBio RS-II platforms. *De novo* and hybrid assemblies were performed with only Illumina, only PacBio and (Illumina + PacBio) data combinations using SPAdes, ABySS and SMRTanalysis software. Complete genome assemblies generated by PacBio data were compared with Illumina draft assemblies to reveal genomic regions which were unassembled by Illumina technology. Two genomes, which could not be automatically finished using either NGS data were manually finished using bioinformatics and PCR/Sanger sequencing approaches to analyze unassembled regions from PacBio sequencing. Analysis of unassembled regions revealed that short reads from Illumina technology were unable to resolve many of the repetitive rRNA operon elements. The unassembled regions through PacBio sequencing appear to be an unaccounted for event and assembly quality and final contig number is the cumulative effect of read-depth, read-quality, sample DNA quality and biological features of the respective genome such presence of infecting phage DNA or mobile genetic elements. In general, the PacBio sequencing generated better assembly statistics as compared to both Illumina and hybrid assemblies. A complete description and importance of post-assembly polishing steps and manual genome finishing approaches is provided and it should be extendible for other studies looking to improve existing genome assemblies. The systematic evaluation of the unassembled DNA from NGS technologies will also be useful for the sequencing companies and algorithm developers to design improved strategies for sequencing and data analysis.

5.2 Introduction

Since the release of first Next-Generation Sequencing (NGS) platform by 454 Life sciences (Margulies, et al., 2005), there is a remarkable increase in sequencing efficiency, throughput and read lengths (Koren and Phillippy, 2014). Sequencing costs have dropped dramatically and whole genome sequencing is within reach even for the small-scale laboratories on relatively modest budgets. During the past decade, the sequencing industry was largely dominated by the second generation, sequencing by synthesis platforms such as Illumina which are characterized by the low-cost, high-throughput, and short reads with high accuracy (van Dijk, et al., 2014). However the short sequencing reads generated have limited power to resolve large repetitive regions even within small microbial genomes (Nagarajan and Pop, 2013). On the other hand, so-called third-generation, single-molecule sequencing platforms such as Pacific Biosciences (PacBio) (Roberts, et al., 2013) are characterized by the longer reads with median read length over 4-5 kb and longest reads well beyond 20 kb (Brown, et al., 2014; Koren and Phillippy, 2014; Utturkar, et al., 2015). A detailed performance comparison between various NGS platforms and recent advances have been summarized for various applications (Liu, et al., 2012; Quail, et al., 2012; Rhoads and Au, 2015; van Dijk, et al.,

2014). However, these comparisons did not perform in-depth analysis of unassembled DNA regions from Illumina or PacBio assemblies.

The short read technologies are generally able to resolve the microbial genomes up to the high-quality draft standard which is sufficient for many applications such understanding gene-coding potential, strain typing or pan-genome analysis (Koren and Phillippy, 2014). However, draft genomes contain fragmented genome assemblies which might contain misassembled regions, incorrect gene calls and other artifacts. Additionally, there is great risk of false negative error when making statements about the absence of any metabolic function within pathways. The reason for fragmented assemblies are often attributed towards repetitive DNA regions which are abundant in microbial genomes and present the greatest technical challenge to assembly process especially when the repetitive region is longer than the read lengths (Treangen and Salzberg, 2012). The rRNA operons are considered as longest repetitive regions within microbial genomes and size ranges from 5 kb to 7 kb although often not arrayed in tandem repeats as in eukaryotic genomes. For example, *Saccharomyces cerevisiae* has an estimated 100 copies of the rRNA genes repeated end after end on chromosome II. Although long since considered “complete” by many measures, they are obviously not fully resolved. The longer reads from PacBio platform have an ability to span through some large repetitive regions and greatly aid the assembly process to generate up to finished quality microbial genomes when sufficient coverage (> 100x) is available (Chin, et al., 2013; Koren, et al., 2013). The finished genome sequences are of relatively higher value (Fraser, et al., 2002), represent more accurate genomic information and often desirable for model organisms or industrially important microbes to support downstream applications. A relative value of PacBio reads for automated finishing of microbial genomes was demonstrated by a recent example of *Clostridium autoethanogenum* as part of our lab group work – the most complex (class III) bacterial genome based on type and content of repeat sequences (Koren, et al., 2013), where complete circular genome sequence was obtained using only the PacBio data without the need for manual finishing (Brown, et al., 2014). In the same study, a comparison of draft (Illumina/454/Ion Torrent) and finished assemblies (PacBio) revealed rRNA operons as the major contributors to the fragmented assembly of short read data (Brown, et al., 2014). However, there are a few examples available where short reads from multiple libraries and platforms were able to achieve finished microbial genome assemblies (Ikegami, et al., 2015; Ribeiro, et al., 2012). Therefore, more examples of the draft and finished genomes comparison would be useful to assess the nature of assembly gaps associated with short-read technologies.

The PacBio sequencing platform was predicted to be able to obtain finished genome assemblies for the majority of bacterial genomes (Koren, et al., 2013) and this has demonstrated by increased number of finished genomes obtained using this technology (Brown, et al., 2014; Eckweiler, et al., 2014; Harhay, et al., 2014; Kanda, et al., 2015; Mehnaz, et al., 2014; Nakano, et al., 2015; Satou, et al., 2014). However, at the same time there were few examples where PacBio alone generated finished (circularized) genome assemblies. More recently many Pacbio alone genomes are only be resolve into 10 or fewer contigs despite high sequence coverage, and manual finishing is often necessary to obtain complete genome sequences (Bishnoi, et al., 2015; Dunitz, et al.,

2014; Hoefler, et al., 2013; Okutani, et al., 2015; Shapiro, et al., 2015). In most cases, these unassembled regions or gaps within the PacBio assembly were not investigated, and the nature of these gaps or reason for assembly failures remains unknown or unverified as of their publication. A systematic comparison of the draft and finished assemblies of multiple microbial genomes may be helpful to reveal the features and properties of these unassembled regions from PacBio and Illumina platforms.

In the present study, eight bacterial genomes were sequenced using Illumina Paired-End (PE) and PacBio RS-II platforms. *De novo* and hybrid genome assemblies were created using individual and/or combinations of data from Illumina and PacBio platforms using various assembly software algorithms and parameters. A manual genome finishing step was performed for several selected genomes where automated finishing with PacBio could not be achieved. A comparison of draft and finished genome assemblies of eight microbial genomes was performed to confirm the nature of gaps associated with Illumina assembly. Gaps sequences for PacBio assemblies were revealed by manual finishing and further investigated for specific properties such as associated annotations, read-lengths, and read coverage. In summary, this study offers insights into the nature of gaps associated with Illumina and PacBio assemblies of microbial genomes and describes the bioinformatics and PCR/Sanger sequencing based genome finishing approaches which could potentially be extended for many unfinished bacterial and archaeal genomes.

5.3 Methods

Whole genome sequencing

Whole genome sequencing for eight microorganisms (*Clostridium pasteurianum* ATCC 6013, *Clostridium autoethanogenum* DSM 10061, *Clostridium paradoxum* JW/YL-7T, *Pelosinus fermentans* UFO1, *Pelosinus fermentans* JBW45, *Halomonas* sp. KO116 and *Bacteroides cellulosolvens* DSM 2933) was performed using Illumina MiSeq (Illumina, San Diego, CA, USA) (Quail, et al., 2012) and PacBio RS-II (Pacific Biosciences, Menlo Park, CA, USA) (Korlach, et al., 2010) platforms. PacBio sequencing for *Clostridium thermocellum* AD2 was performed at Joint Genome Institute (JGI) (http://jgi.doe.gov/). Illumina paired-end library preparation, PacBio SMRTbell library preparation, and sequencing were performed as described previously (Utturkar, et al., 2015).

Data quality control, genome assembly, and annotation

Quality based trimming of raw Illumina data was performed using CLC genomics workbench software (CLC) as described previously (Utturkar, et al., 2014). Adapter trimming of raw PacBio data was performed through SMRT analysis software to obtain filtered subreads as described previously (Utturkar, et al., 2015). *De novo* genome assembly of Illumina data was performed using SPAdes version 3.5.0 (Bankevich, et al., 2012) and ABySS version 1.5.2 (Simpson, et al., 2009) with optimized kmers as described previously (Utturkar, et al., 2014). The hybrid assembly of Illumina and PacBio data was performed using SPAdes hybrid assembler version 3.5.0 with default parameters. Long read data from PacBio was assembled using SMRT Analysis software and HGAP protocol (Chin, et al., 2013). The specific versions of SMRT Analysis software used for each genome are provided in the results section. The HGAP parameter of “Target Coverage” was updated to 15X as recommended by PacBio (Pacific-Biosciences, 2014).

The assembly summary statistics were determined using Quast software version 2.3 (Gurevich, et al., 2013). PacBio only assemblies were polished using an additional round of quiver correction (Chin, et al., 2013) and Pilon software version 1.13 (Walker, et al., 2014). Quiver and Pilon algorithms use PacBio and Illumina reads, respectively, in order to perform the assembly basecall correction and derive an accurate consensus sequence. DNA base modification analysis was performed through SMRT analysis software to determine the complete methylation profile (Pacific-BioSciences, 2014). Gene-calling and genome annotation was performed through the Prodigal algorithm and microbial genome annotation pipeline at Oak Ridge National Laboratory as described previously (Hyatt, et al., 2010; Woo, et al., 2014).

Manual genome finishing

Manual genome finishing was performed using bioinformatics and a PCR/Sanger sequencing approaches. During bioinformatics finishing steps, the contigs from the draft and hybrid genome assemblies were mapped to the PacBio only assemblies using Geneious software version 8.1.6 (Biomatters, Auckland, New Zealand) (Kearse, et al., 2012). Mapping results were manually checked to identify a possible sequence extension for the reference contig ends. After contigs extension, the “Super-assembly” workflow from Geneious software was applied to determine the possible overlap between contigs. Contig extensions and contig overlaps detected by mapping and super-assembly were verified using a PCR and Sanger sequencing approach, as described previously (Utturkar, et al., 2014). Briefly, oligonucleotide primers were designed to each flanking contig overlap end and validation occurred when PCR amplified products of the predicted size. In the case of large PCR products (> 3 kb), an additional set of internal oligonucleotide primers were designed to amplify the end regions. PCR reactions were performed using a Phusion High-Fidelity PCR Kit (New England Biolabs, Ipswich, MA) following the manufacturer’s protocol. PCR product purification was performed using MinElute PCR purification kit (Qiagen) following the manufacturer’s protocol. Sanger sequencing of purified PCR products was performed at Molecular Biology Research Facility, University of Tennessee, Knoxville using ABI 3730 Genetic Analyzer Instrument (Life Technologies). Sanger reads were quality trimmed and aligned to reference sequences using Geneious software to validate consensus accuracy and contiguity.

Obtaining circular genomes

Single contigs assemblies derived from the HGAP protocol often have the overlapping ends representing the potential circular genome assembly. The circular nature of contigs was confirmed via a dot-plot and alignment approach as described in PacBio training protocol (Pacific-Biosciences, 2015). Additionally, singleton sequences and deg.fasta files generated during HGAP assembly were tested *in silico* to account for the possibility of non-chromosomal DNA such as plasmid or DNA-phage elements. Whenever the assembly generated less than 5 contigs, each separate contig was tested for circularity to identify non-chromosomal DNA. If more than one circular contigs were detected then presence of plasmid DNA was analyzed by searching for annotated plasmid genes such as “RepA – plasmid replication protein”.

Assembly comparisons and nature of DNA gaps

Gaps or breakpoints within Illumina assemblies compared to PacBio assemblies were determined by mapping of draft contigs against the PacBio assembly using the "Map to Reference" module from the Geneious software. The mapping was manually reviewed to identify the genomic coordinates and annotations associated with Illumina gaps. Sequences for gaps derived through manual finishing and PCR/Sanger sequencing approach represented an unassembled DNA in PacBio assembly. PacBio gap sequences were submitted to the mfold web server (Zuker, 2003) to determine DNA folding properties and secondary structures. Default DNA folding parameters in mfold were modified to mimic the PCR conditions (folding temperature of 55^0 C, $[Na^+]$ concentration of 50 mM and $[Mg^{++}]$ concentration of 2.5 mM). Additionally, reciprocal BLASTP analyses were performed to gain insights into potential protein coding differences from the draft and finished genome assemblies as described previously (Utturkar, et al., 2014).

5.4 Results and Discussion

Sequencing and assembly details

Raw sequence data for eight microbial genomes was output as de-multiplexed fastq files (Illumina-Inc., 2011) and in the SMRT sequencing data format (Kim, et al., 2014) by Illumina and PacBio RS-II platforms, respectively. Quality trimming procedures removed low-quality bases and/or adapter sequences from the raw reads. Post trimming statistics for Illumina data and post filtering statistics for PacBio data such as number of reads, average read lengths and genome coverage and total bases are summarized in Table 5.1 and Table 5.2, respectively. Illumina sequence coverage for each genome is greater than 200x which is sufficient to derive high-quality genome assemblies (Haridas, et al., 2011; Utturkar, et al., 2014). The PacBio sequence coverage for each genome was greater 100x except for the isolates of *Pelosinus* sp. UFO1 and *B. cellulosolvens* DSM 2933.

The assembly summary statistics for *de novo* and hybrid assemblies are described in Table 5.3. In this study, the SPAdes software generated superior assembly statistics as compared to ABySS software in terms of fewer contigs, longest contig lengths, and better N50 lengths. However, the performance of these assemblers was data and genome specific and ABySS performed better for three of the eight genomes. Therefore based on current data it is difficult to weigh one short-reads assembler over other and it is recommended to try multiple assembly programs to select the optimal assembly and use our rRNA analysis approach (Utturkar, et al., 2014). The hybrid assemblies generated using a combination of Illumina and PacBio data have better statistics as compared to draft assemblies and these results are consistent with an earlier study (Brown, et al., 2014). The PacBio only assemblies generated with HGAP protocol always generated the best assembly statistics among all the tested protocols in this study. In fact, four of the eight genomes were assembled as complete circular chromosomes using only the PacBio data and HGAP assembler.

It is worth mentioning that use of the latest version of assembly software has a significant impact on overall assembly quality. For example, the *B. cellulosolvens* genome assembled through HGAP protocol from SMRT analysis version 2.0 generated 12

contigs, while the HGAP method from SMRT analysis version 2.2 generated a 3 contig assembly using the same data. Similarly, for *C. pasteurianum* ATCC 6013, HGAP protocol obtained 12 contigs with SMRT analysis version 2.0 while version 2.2 generated a 2 contig assembly. SMRT analysis 2.2 version updated the HGAP.3 protocol and contains significant performance updates in terms of speed, use of computation time and space, and additional parameters for sequence filtering and chimera detection (Pacific-BioSciences, 2013; Pacific-BioSciences, 2014). Newer softwaere versions are often associated with significant algorithm improvements or more relevant default parameters which can obtain better assemblies. Therefore, it is recommended to review the software version changes in release notes and keep the assembly toolbox updated.

The HGAP protocol was also sensitive enough to assemble circular plasmid DNA sequences as separate elements from chromosomal DNA. For example, the HGAP protocol generated three contigs in the assembly for *Halomonas* sp. KO116 genome (a halophilic gamma-proteobacteria) which initially appeared to be a fragmented genome assembly with near-finished status. However, when each contig was analyzed separately for circularity, it revealed the presence of a circular chromosome and two circular megaplasmid sequences (O'Dell, et al., 2015).

Manual genome finishing

The *Clostridium thermocellum* AD2, *Bacteroides cellulosolvens* DSM 2933, *Clostridium pasteurianum* ATCC 6013 and *Clostridium paradoxum* JW/YL-7T genomes could not be automatically assembled into single contig using HGAP protocol and manual finishing was necessary. The *C. paradoxum* genome was reported to contain multiple 16S rRNA genes with heterogeneous intervening sequences (15 different sequences in variable region I of 16S rRNA) (Rainey, et al., 1996) which could be the possible reason for incomplete assembly. The genome finishing for *C. paradoxum* was out of scope for current study due to time constraints. Hence the *C. paradoxum* genome was submitted to NCBI with near-finished status and utilized only for the assessment of gaps present within Illumina assembly. The analysis of the 2 contig assembly for strain ATCC 6013 revealed a possible phage integration and large sequence duplication which may have prevented complete assembly. Meanwhile, a complete genome sequence for this genome was reported by another group using a manual finishing approach (Rotta, et al., 2015) and hence manual finishing was not performed. Sanger sequence data for gaps within ATCC 6013 genome was not available publically and this genome was utilized only for the assessment of Illumina assembly gaps. Manual finishing was performed for the *C. thermocellum* AD2 and *B. cellulosolvens* DSM 2933 genomes and utilized for assessment of gaps present within PacBio assemblies. Details of the manual genome finishing approaches are provided below.

The best assembly for strain AD2 using PacBio only data contained 10 contigs. Mapping of the draft assemblies generated by SPAdes and ABySS to the 10 contig reference sequence permitted the ends of 7 contigs to be extended. Super-assembly of all 10 contigs (including extended ends) generated 4 super-contigs. Ends of the four supercontigs could not be extended further by re-mapping of draft contigs or raw reads. The longest super-contig (AD2_SC1) was of size 2.06 Mb and derived from the assembly

of three extended contigs. The super-assembly of AD2_SC1 and consensus accuracy was validated by verification of the overlaps (AD2_overlap1 and AD2_Overlap2) between three child contigs (Figure 5.1). Analysis of the SPAdes hybrid assembly revealed the presence of the longest contig (AD2_HC1), which was 2.27 Mb. Super-assembly of these two longest contigs (AD2_SC1 and AD2_HC1) derived a consensus sequence of size 3.5 Mb with ~780 kb overlapping sequence (Figure 5.1). The consensus of contig overlap was manually reviewed to verify that sequence and annotations are matching and a few mismatches were manually corrected based on raw reads mapping. The expected genome size for the AD2 was similar to derived consensus length (3.5 Mb). However, dot plot and other circularity tests did not obtain any evidence for the circular genome. Additional PCR amplification reactions were performed to extend the ends of the 3.5 Mb consensus sequence. PCR obtained a clean band at ~1 kb location and Sanger sequencing revealed sequence for this gap region (AD2_Gap1) comprising 1,069 bp. This gap was consisted of a repetitive transposon DNA with annotation “transposon mutator type CDS”. Gap closure obtained the 3,554,860 bp circular genome sequence for AD2. Overview of bioinformatics and manual finishing process for AD2 is shown in Figure 5.1. The sequences obtained through manual finishing (AD2_overlap1, AD2_Overlap2 and AD2_Gap1) constituted the gaps present within the PacBio assemblies and further characterization is described in later section.

The best assembly for *B. cellulosolvens* DSM 2933 using PacBio only data contained 3 contigs. Scaffolding using AHA protocol (Bashir, et al., 2012) could not achieve better assembly results. Mapping of the draft or hybrid assembly to 3 contig reference could not extend any contig ends. However, super-assembly of 3 contigs detected a 6.7 kb overlap (BC_overlap1) between contigs BC_C1 and BC_C3. The overlap and resulting consensus sequence was validated by PCR and Sanger sequencing. The remaining contig (BC_C2) could not be assembled together using bioinformatics approaches and suggested the presence of an unknown gap. This gap (BC_Gap1) was resolved by PCR and Sanger sequencing to uncover the 406 bp gap sequence between contigs BC_C2 and BC_C3. The BC_Gap1 region was most difficult region to resolve by PCR and perhaps for the multiple reasons. First of all, the orientation and order of adjoining contigs was unknown, and other genomes from same lineage had low (below 90%) average nucleotide identity scores. Therefore, multiple combinations of forward/reverse primers need to be tested to determine the correct order and orientation. Even after determination of the correct contig order and orientation based on PCR products, the amplification of this gap sequence required multiple rounds of PCR and several optimizations. The PCR optimizations included the use of strand-displacing DNA polymerase for uncoiling of double-stranded DNA and use of nucleotide analogue 7-deaza-2′-dGTP to break any DNA hairpin structures and with conditions as described previously (Hurt, et al., 2012). The closure of BC_Gap1 resulted in a single contig assembly for the DSM 2933 genome. However, further finishing approaches to obtain circular chromosome were unsuccessful and genome was deposited at Genbank with near-finished status. An example of bioinformatics and manual finishing of *B. cellulosolvens* is provided in Figure 5.2. The sequences obtained through manual finishing (BC_overlap1 and BC_Gap1) constituted the gaps present within the PacBio assemblies and further characterization is described in later section.

The bioinformatics finishing approach described in this manuscript was successfully applied to two near-finished genomes from PacBio data and was able to obtain finished genome sequences. This approach derives its advantages by using multiple assembly software. The SPAdes, ABySS and ALLPATHS-LG assemblers are listed as consistent performers in several assembly comparison reports (Bradnam, et al., 2013; Liao, et al., 2015; Magoc, et al., 2013; Salzberg, et al., 2012; Utturkar, et al., 2014), each having their own specific advantages. These assemblers may perform better than other in certain aspects of the genome assembly. For example, the ALLPATHS-LG is optimized to take advantage of paired-end and mate-pair library type (Maccallum, et al., 2009), the ABySS achieves better assembly contiguity (Utturkar, et al., 2014) while SPAdes generates consistent assembly results with the integration of multiple platforms and also generates accurate consensus sequence (Liao, et al., 2015; Magoc, et al., 2013). The extension of the PacBio contig ends achieved through mapping of draft contigs was a crucial step for the successful super-assembly. Several other genome finishing approaches defined in the literature (Galardini, et al., 2011; Nagarajan, et al., 2010; Ramos, et al., 2013; Swain, et al., 2012) are primarily targeted for the assemblies generated with short reads and the same steps may not be applicable for near-finished genome assemblies (< 10 contigs). The AHA scaffolding approach could not obtain improvement in near-finished assemblies of four unfinished genomes described in this study (data not shown). The HGAP algorithm currently the most efficient algorithm for native PacBio assembly and generated finished circular genomes in many cases (Brown, et al., 2014; De Leon, et al., 2015; Harhay, et al., 2014; Kanda, et al., 2015; Mehnaz, et al., 2014; Nakano, et al., 2015; O'Dell, et al., 2015; Satou, et al., 2014). A few gaps remained in near-finished assemblies generated by HGAP method may represent large repetitive regions which are not feasible for automated finishing at this time and manual inspection is necessary to avoid misassembly. Therefore, other scaffolding (Bashir, et al., 2012; Bosi, et al., 2015) or gap-filling (Boetzer and Pirovano, 2012; English, et al., 2012; Kosugi, et al., 2015; Paulino, et al., 2015) approaches are also speculated to have limited utility for near-finished genomes. Moreover, the scaffolded assemblies are often associated with unresolved base-calls represented as “N”s in the assembly and may not be used as direct evidence for contigs joining. On the contrary, the current finishing approach is tailor-made for near-finished genome which includes only a few steps to obtain direct evidence for contigs overlap and further confirmation using PCR/Sanger sequencing method. This finishing approach should be extendible to any unfinished genome, but due care is necessary when overlapping contigs are of small sizes or might be mapping to more than one location in the genome.

Assembly Polishing

In its early stages, the PacBio sequencing platform was criticized for the high error rates (~15%) associated with this technology (Koren, et al., 2012). In later stages, the high error rate was overcome by improved throughput from PacBio RS-II platform and development of HGAP algorithm which corrects random errors using high sequence coverage to generate high quality genome assemblies (Chin, et al., 2013). However even with the random error profiles, the PacBio sequencing chemistry/platform does not guarantee uniform or average coverage across the entire genome and some regions might be underrepresented i.e. have less than average sequence coverage. These

underrepresented sequences have high probability to contain base-call errors which may lead to spurious overlaps and a misassembly. Therefore, assembly polishing is a crucial step for PacBio data to obtain accurate consensus sequence and facilitate downstream finishing process. The HGAP protocol is integrated with a single round of quiver polishing which uses the raw PacBio data, underlying quality values and hidden Markov model based probabilities for the basecall quality and generate improved consensus sequence (Chin, et al., 2013).

Running additional rounds of quiver polishing might be beneficial to further improve the consensus accuracy. Therefore, finished genome assemblies in current study were polished using multiple rounds of quiver. If Illumina data is available, the Pilon software offers further opportunity to correct the PacBio consensus using high-quality Illumina reads. The quiver polished assemblies were further corrected with Illumina reads using the Pilon software. The modifications suggested by quiver/Pilon were mostly of type insertions/deletions resulting in frameshift mutation corrections. Open Reading Frame (ORF) or gene prediction was performed before and after the Pilon correction using Prodigal gene prediction algorithm (Hyatt, et al., 2010). The impact of Pilon correction on gene-calls was analyzed by reciprocal BLASTP analyses as described previously (Utturkar, et al., 2014). A summary of modifications suggested by Pilon and associated changes in the gene calls are summarized in Table 5.4. In most cases, Pilon basecall corrections resulted in improved gene calling accuracy i.e. a substantial number of proteins were longer (previously split genes were joined together to represent a single longer gene or basecall corrections updated six-frame translation results to generate longer ORFs) and a number of new proteins were predicted. However, it should be noted that these results are obtained from standalone *in silico* searches with Prodigal gene prediction algorithm and additional *in vitro* validation using RNA-sequencing or proteomic studies would be required to validate the gene models.

Another important aspect of assembly polishing is the removal of overlapping ends from circular assemblies. Circular assemblies generated through HGAP protocol often have overlapping ends (which represent sequence duplication) and one of the ends needs to be trimmed off from the final linear assembly. In this study, the removal of overlapping ends was performed manually through alignment and read mapping approach. However, a recent development in this area includes a software called circlator (Hunt, et al., 2015) which performs automated assembly circularization and produces a linear representation of circular sequences. We tested this software with three genomes and it was able to correctly trim-off the overlapping end and generated accurate consensus sequences (data not shown). We recommended to use circlator software for genome circularization followed by a careful manual inspection of trimmed region.

The impact of assembly improvement on gene-calling was tested by evaluation of protein coding differences between draft and final genome assemblies for each microorganism. Reciprocal BLASTP analysis was performed to determine the number of new ORFs, and quantify longer and shorter ORFs (Table 5.5). A substantial number of proteins (ranging from 9 to 342) were longer in the final genome assemblies and a number of new proteins (ranging from 1 to 20) were also predicted for each isolate. The majority of the newly

predicted proteins were hypothetical proteins and others included metabolic or regulatory functions such as glycoside hydrolase, transcriptional regulators, and putative type III restriction protein. Consistent with earlier studies, the majority of proteins (92-98%) remained unchanged within draft and finished assemblies. This result supports the notion that draft quality genomes are sufficient for certain applications such as resequencing, phylogenomics or SNP calling.

Unassembled DNA regions in Illumina only assembly

The unassembled DNA or assembly breakpoints in Illumina assemblies were revealed by mapping against the complete genome assemblies. Short reads from Illumina technology have limited power to resolve longer repetitive regions (Treangen and Salzberg, 2012; Utturkar, et al., 2014) and rRNA operons are considered as most difficult regions to assemble (Brown, et al., 2014). Our comparison results were consistent with previous findings (Brown, et al., 2014). For example, seven of the eight genomes analyzed in this study have at least 50% of the total rRNA operons missing (unassembled) from the Illumina assembly. While from remaining 50% of rRNA operons, most could only be assembled partially (i.e. missing one of the 5S, 16S or 23S elements). The actual number of rRNA operons present in each genome and number of rRNA operons missing or with partial coverage in Illumina assembly are described in Table 5.6. Other regions that contributed to fragmented Illumina assemblies included transposon sequences, ABC-type transporters (which number in the double digits for most genomes), RNA-directed DNA polymerases (which have long sequences and share high homology), as well hypothetical proteins. A complete table describing the details for the annotations, coordinates and locus tags associated with the gaps within the Illumina assembly of each genome are provided in Table 5.7.

Unassembled DNA regions in PacBio only assembly

Manual inspection of PacBio assemblies revealed the presence of two overlaps and one gap sequence in AD2 genome, and one overlap and one gap sequence in DSM 2933 genome. These overlap/gap sequences represent the unassembled DNA regions in PacBio assembly which could not be resolved using current software. Indeed, certain DNA regions are much more difficult to resolve by sequencing because of GC rich sequences, ability to form hairpin structures, homopolymeric stretches and repeat contents (Hurt, et al., 2012). These gaps present within current PacBio assemblies were resolved by using specialized PCR amplification protocol which include the use of strand-displacing *Pfu* DNA polymerase, ramped PCR extension cycle, nucleotide analogue 7-deaza-2′-dGTP as described previously (Hurt, et al., 2012). For the AD2 and DSM 2933 genomes, we were able to obtain the high quality Sanger sequence data for the five PacBio gaps (three from AD2 and two from DAM 2933) which allowed further investigation of these unassembled regions. The basic properties such as genome coordinates, length, PacBio read coverage, % GC and corresponding annotations were determined for these gap sequences and described in Table 5.8. Four of five PacBio gaps were associated with lower than recommended coverage (>100x) for the native PacBio assembly through HGAP protocol. The AD2 genome was of particular interest because despite 296x average PacBio read coverage, two PacBio gaps had only 36x and 82x sequence coverage. The second region of interest was a gap sequence in DSM 2933

genome (BC_Gap1) had only 4x PacBio read coverage compared to 48x coverage on average in other regions.

The current PCR optimizations applied for the amplification of PacBio gaps in our study were adapted from a previous study by Hurt et al (Hurt, et al., 2012), where these PCR optimizations could obtain successful amplification through GC rich secondary stem loop structures. Therefore, we hypothesized that these PacBio gap sequences with low coverage might have the ability to form strong hairpin loop structures that prevent the DNA polymerase enzyme from being able to unwind and extend through the DNA region. For further investigation, Sanger derived Pacbio gap sequences (three from AD2 and two from DSM 2933) were analyzed using mfold web server to determine minimum free energies (ΔG) and their abilities to form DNA hairpins and secondary structures associated. For comparison, ten random sequences were selected from the AD2 and DSM 2933 genomes and similar analyses were performed using mfold web server analysis (Table 5.9). This *in silico* analysis detected formation of small stem-loop structures in PacBio gap sequences but there was no evidence for large secondary loops which might interfere with DNA polymerase and result in low sequence coverage. Secondly, there was no significant difference observed between minimum free energies and secondary structures of PacBio gaps and randomly selected regions. Based on our data, low sequence coverage regions within PacBio assemblies appears to be a random event or as yet unaccounted for event. There is a possibility that sequencing depth for these regions is affected by the overall sample quality and/or availability of high molecular weight DNA. In terms of assembly, it is likely that the HGAP software could not obtain a sufficient number of reads to support the automatic closure of these gaps and generated fragmented assemblies. Additionally, many of these gap sequences were corresponding to repetitive DNA elements such as “Transposon-related proteins”. Lower accuracy of individual PacBio reads could also be the contributing factor for assembly fragmentation or errors (Koren, et al., 2012).

To assess the assemblies in greater detail biological aspects were considered. The complete genome sequence of strain JBW45 was characterized by the presence of active transposon element which interfered with the genome circularization (De Leon, et al., 2015). The contig terminal regions of *B. cellulosolvens* DSM 2933 were also characterized by the presence of transposon-related genes and speculated to interfere with the genome closure (Dassa, et al., 2015). Further analysis of our two contigs assembly for *C. pasteurianum* ATCC 6013 revealed that contig 2 corresponds to an excised phage product (unpublished results). In the case of KO116, the presence of megaplasmids could have been easily confused as a near-finished assembly without careful analysis. In another example, a small 5.5 kb *C. autoethanogenum* plasmid was apparently absent from the PacBio complete genome assembly, which may have resulted in the size exclusion method used to obtain high molecular weight DNA for SRMT cell library preparation. The draft assembly created with legacy 454 and Illumina data for the same strain was later found to contain a 5.5 kb contig and within the draft assembly there was evidence for the presence of several plasmid related proteins such as “"plasmid recombination protein” and “COG5655 plasmid rolling circle replication initiator protein and truncated derivatives”. Later, the presence of plasmid DNA was confirmed by

extraction and separation on the gel by collaborators, although it has not been reported in the literature at this time. It is possible that plasmids can lost and can vary between laboratories. Therefore, biological features of the specific genome such as presence of phage DNA, active mobile genetic elements or the presence of plasmids can be reasons for apparently unfinished PacBio assemblies.

To summarize, unresolved regions in the PacBio assembly appear to result from the cumulative effect of low read-depth, quality of the reads spanning these regions, overall quality of sample DNA and presence of repetitive DNA or transposon elements. Investigation of these gap regions with bioinformatics approaches and a certain level of manual inspection is recommended as it might be able to achieve the closure of these gaps and obtain the finished genome sequences. Currently, the cost for the PacBio sequencing is high as compared to Illumina and small scale project may rely on a limited number of SMRT cells per genome. Further, the instrument's sheer physical size and high error rate combine to limit PacBio applications currently and additional developments will be required for widespread adaptation in applications outside of de novo genome generation. A recent announcement from PacBio includes the release of the Sequel system (Pacific-BioSciences, 2015), which provides higher throughput, greater scalability, and seven times the sequencing output of PacBio RS-II system without significant change in library preparation protocol. The Nanopore sequencing has also been released in beta form and may prove useful also for identification or detection of important genomic features such as bacterial antibiotic resistance island (Ashton, et al., 2015) and antimicrobial resistance genes (Judge, et al., 2015). In the future, lower costs, shorter timelines and improved sequencing chemistry for third generation sequencing platforms will help to obtain higher sequence coverage, improved read-lengths and is anticipated to generate even greater number of finished genomes.

5.5 Conclusion

Illumina is the most widely used sequencing platform and can obtain high draft-quality genome assemblies for microbial genomes. The single-molecule sequencing method from PacBio is currently one of the best methods available to obtain finished grade microbial genome assemblies in an automated fashion. However, there are certain more difficult genomes which cannot be readily sequenced or assembled solely using Illumina and/or PacBio platforms. Comparison of Illumina and PacBio assemblies of eight microbial genomes revealed that the gaps in the Illumina assemblies were mostly associated with repetitive rRNA operons, phages and a similar features. However, there was no specific trend observed related to PacBio gaps and appears to be an unaccounted event based on current data. A manual genome finishing approach is proposed at present, which uses a combination of bioinformatics tools and PCR/Sanger sequencing based validation to successfully obtain up to finished quality genome assemblies. This approach could be extendible to any near-finished genomes.

References

Ashton, P.M., et al. (2015) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island, Nat. Biotechnol., 33, 296-300.

Bankevich, A., et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol., 19, 455-477.

Bashir, A., et al. (2012) A hybrid approach for the automated finishing of bacterial genomes, Nat. Biotechnol., 30, 701-707.

Bishnoi, U., et al. (2015) Draft Genome Sequence of a Natural Root Isolate, Bacillus subtilis UD1022, a Potential Plant Growth-Promoting Biocontrol Agent, Genome Announc, 3.

Boetzer, M. and Pirovano, W. (2012) Toward almost closed genomes with GapFiller, Genome Biol, 13, R56.

Bosi, E., et al. (2015) MeDuSa: a multi-draft based scaffolder, Bioinformatics, 31, 2443-2451.

Bradnam, K., et al. (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species, GigaScience, 2, 10.

Brown, S., et al. (2014) Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of Clostridium autoethanogenum and analysis of CRISPR systems in industrial relevant Clostridia, Biotechnol. Biofuels, 7, 40.

Brown, S.D., et al. (2014) Complete genome sequence of Pelosinus sp. strain UFO1 assembled using Single-Molecule Real-Time DNA sequencing technology, Genome Announc, 2.

Chin, C.S., et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, Nat. Methods, 10, 563-569.

Dassa, B., et al. (2015) Near-Complete Genome Sequence of the Cellulolytic Bacterium Bacteroides (Pseudobacteroides) cellulosolvens ATCC 35603, Genome Announc, 3.

De Leon, K.B., et al. (2015) Complete Genome Sequence of Pelosinus fermentans JBW45, a Member of a Remarkably Competitive Group of Negativicutes in the Firmicutes Phylum, Genome Announc, 3.

Dunitz, M.I., et al. (2014) Draft Genome Sequences of Escherichia coli Strains Isolated from Septic Patients, Genome Announc, 2.

Eckweiler, D., et al. (2014) Complete genome sequence of highly adherent Pseudomonas aeruginosa small-colony variant SCV20265, Genome Announc, 2.

English, A.C., et al. (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology, PLoS One, 7, e47768.

Fraser, C.M., et al. (2002) The value of complete microbial genome sequencing (you get what you pay for), J. Bacteriol., 184, 6403-6405.

Galardini, M., et al. (2011) CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes, Source Code Biol Med, 6, 11.

Gurevich, A., et al. (2013) QUAST: quality assessment tool for genome assemblies, Bioinformatics, 29, 1072-1075.

Harhay, G.P., et al. (2014) Complete closed genome sequences of three Bibersteinia trehalosi nasopharyngeal isolates from cattle with shipping fever, Genome Announc, 2.

Haridas, S., et al. (2011) A biologist's guide to de novo genome assembly using next-generation sequence data: A test with fungal genomes, J. Microbiol. Methods, 86, 368-375.

Hoefler, B.C., Konganti, K. and Straight, P.D. (2013) De novo Assembly of the Streptomyces sp. Strain Mg1 Genome Using PacBio Single-Molecule Sequencing, Genome Announc, 1.

Hunt, M., et al. (2015) Circlator: automated circularization of genome assemblies using long sequencing reads, bioRxiv.

Hurt, R.A., et al. (2012) Sequencing intractable DNA to close microbial genomes, PLoS One, 7, 7.

Hyatt, D., et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinformatics, 11, 119.

Ikegami, T., et al. (2015) Hybrid De novo Genome Assembly Using MiSeq and SOLiD Short Read Data, PLoS One, 10, e0126289.

Illumina-Inc. (2011) CASAVA v1.8.2 User Guide.

Judge, K., et al. (2015) Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes, J. Antimicrob. Chemother., 70, 2775-2778.

Kanda, K., Nakashima, K. and Nagano, Y. (2015) Complete Genome Sequence of Bacillus thuringiensis Serovar Tolworthi Strain Pasteur Institute Standard, Genome Announc, 3.

Kearse, M., et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data, Bioinformatics, 28, 1647-1649.

Kim, K.E., et al. (2014) Long-read, whole-genome shotgun sequence data for five model organisms, Scientific Data, 1.

Koren, S., et al. (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing, Genome Biol., 14, R101.

Koren, S. and Phillippy, A.M. (2014) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly, Curr. Opin. Microbiol., 23C, 110-120.

Koren, S., et al. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads, Nat. Biotechnol., 30, 693-700.

Korlach, J., et al. (2010) Real-time DNA sequencing from single polymerase molecules, Methods Enzymol., 472, 431-455.

Kosugi, S., Hirakawa, H. and Tabata, S. (2015) GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments, Bioinformatics.

Liao, Y.C., Lin, S.H. and Lin, H.H. (2015) Completing bacterial genome assemblies: strategy and performance comparisons, Sci Rep, 5, 8747.

Liu, L., et al. (2012) Comparison of next-generation sequencing systems, J. Biomed. Biotechnol., 2012, 251364.

Maccallum, I., et al. (2009) ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads, Genome Biol., 10, R103.

Magoc, T., et al. (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms, Bioinformatics, 29, 1718-1725.

Margulies, M., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors, Nature, 437, 376-380.

Mehnaz, S., Bauer, J.S. and Gross, H. (2014) Complete genome sequence of the sugar cane endophyte Pseudomonas aurantiaca PB-St2, a disease-suppressive bacterium

with antifungal activity toward the plant pathogen Colletotrichum falcatum, Genome Announc, 2.

Nagarajan, N., et al. (2010) Finishing genomes with limited resources: lessons from an ensemble of microbial genomes, BMC Genomics, 11, 242.

Nagarajan, N. and Pop, M. (2013) Sequence assembly demystified, Nat. Rev. Genet., 14, 157-167.

Nakano, K., et al. (2015) First Complete Genome Sequence of Clostridium sporogenes DSM 795T, a Nontoxigenic Surrogate for Clostridium botulinum, Determined Using PacBio Single-Molecule Real-Time Technology, Genome Announc, 3.

O'Dell, K.B., et al. (2015) Genome Sequence of Halomonas sp. Strain KO116, an Ionic Liquid-Tolerant Marine Bacterium Isolated from a Lignin-Enriched Seawater Microcosm, Genome Announc, 3.

Okutani, A., et al. (2015) Draft Genome Sequences of Bacillus anthracis Strains Stored for Several Decades in Japan, Genome Announc, 3.

Pacific-BioSciences (2013) SMRT Analysis Release Notes v2.1.

Pacific-Biosciences (2014) HGAP in SMRT Analysis.

Pacific-BioSciences (2014) SMRT Analysis Release Notes v2.2.0.

Pacific-Biosciences (2015) Circularizing and trimming.

Pacific-BioSciences (2015) Sequel™ System Offers Significantly Higher Throughput, Reducing Project Costs and Timelines.

Paulino, D., et al. (2015) Sealer: a scalable gap-closing application for finishing draft genomes, BMC Bioinformatics, 16, 230.

Quail, M.A., et al. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC Genomics, 13, 341.

Rainey, F.A., et al. (1996) Clostridium paradoxum DSM 7308T contains multiple 16S rRNA genes with heterogeneous intervening sequences, Microbiology, 142 (Pt 8), 2087-2095.

Ramos, R.T., et al. (2013) Graphical contig analyzer for all sequencing platforms (G4ALL): a new stand-alone tool for finishing and draft generation of bacterial genomes, Bioinformation, 9, 599-604.

Rhoads, A. and Au, K.F. (2015) PacBio Sequencing and Its Applications, Genomics Proteomics Bioinformatics.

Ribeiro, F.J., et al. (2012) Finished bacterial genomes from shotgun sequence data, Genome Res., 22, 2270-2277.

Roberts, R.J., Carneiro, M.O. and Schatz, M.C. (2013) The advantages of SMRT sequencing, Genome Biol., 14, 405.

Rotta, C., et al. (2015) Closed Genome Sequence of Clostridium pasteurianum ATCC 6013, Genome Announc, 3.

Salzberg, S.L., et al. (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms, Genome Res., 22, 557-567.

Satou, K., et al. (2014) Complete genome sequences of eight Helicobacter pylori strains with different virulence factor genotypes and methylation profiles, isolated from patients with diverse gastrointestinal diseases on Okinawa Island, Japan, determined using PacBio Single-Molecule Real-Time Technology, Genome Announc, 2.

Shapiro, L.R., et al. (2015) Draft Genome Sequence of Erwinia tracheiphila, an Economically Important Bacterial Pathogen of Cucurbits, Genome Announc, 3.

Simpson, J.T., et al. (2009) ABySS: a parallel assembler for short read sequence data, Genome Res., 19, 1117-1123.

Swain, M.T., et al. (2012) A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs, Nat. Protoc., 7, 1260-1284.

Treangen, T.J. and Salzberg, S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions, Nat. Rev. Genet., 13, 36-46.

Utturkar, S.M., et al. (2015) Sequence data for Clostridium autoethanogenum using three generations of sequencing technologies, Sci Data, 2, 150014.

Utturkar, S.M., et al. (2014) Evaluation and validation of de novo and hybrid assembly techniques to derive high quality genome sequences, Bioinformatics.

van Dijk, E.L., et al. (2014) Ten years of next-generation sequencing technology, Trends Genet., 30, 418-426.

Walker, B.J., et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, PLoS One, 9, e112963.

Woo, H.L., et al. (2014) Draft Genome Sequence of the Lignin-Degrading Burkholderia sp. Strain LIG30, Isolated from Wet Tropical Forest Soil, Genome Announc, 2.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Res., 31, 3406-3415.

Appendix

Table 5.1: Data summary statistics for Illumina sequencing.

Organism	Number of Reads	Mean Read Length After Trim (bp)	Total Bases	Coverage
Clostridium thermocellum AD2	22,031,042	96	2,122,690,897	597x
Halomonas sp. KO116	9,277,426	228	2,115,253,128	450x
Pelosinus sp. UFO1	17,883,813	259	4,631,907,567	905x
Pelosinus sp. JBW45	34,276,660	93	3,187,729,380	592x
Clostridium paradoxum JW/YL-7T	18,423,215	255	4,697,919,825	2434x
Bacteroides cellulosolvens ATCC 35603	16,708,471	100	1,670,847,100	242x
Clostridium pasteurianum ATCC 6013	10,221,462	145	1,482,111,990	340x

Table 5.2: Data summary statistics for PacBio sequencing.

Organism	Number of SMRT cells	Number of Reads	Mean Read Length (bp)	Total Bases	Longest Read (bp)	Coverage
Clostridium thermocellum AD2	4	445,834	2,364	1,054,379,633	25,849	296x
Halomonas sp. KO116	2	199,363	6,743	1,344,484,059	36,120	286x
Pelosinus sp. UFO1	3	106,197	4,677	496,733,292	23,938	97x
Pelosinus sp. JBW45	2	202,124	6,658	1,345,758,432	35,018	250x
Clostridium paradoxum JW/YL-7T	3	140,177	6,028	604,691,408	23,588	313x
Bacteroides cellulosolvens ATCC 35603	4	80,397	4,162	334,636,538	20,966	48x
Clostridium pasteurianum ATCC 6013	7	186,225	3,637	677,451,123	28,542	155x

Table 5.3: Assembly summary statistics for *de novo* and hybrid assemblies.

Organism	NGS Technology	No. of contigs	Maximum Contig Size (kb)	N50 (kb)	Genome Size (Mb)	Software
Clostridium thermocellum AD2	Illumina	102	331	116	3.48	SPAdes*
		107	282	84	3.54	ABySS
	Illumina + PacBio	14	2270	2270	3.57	SPAdes
	PacBio only	10	982	891	3.49	SMRTanalysis v 2.2
	PacBio only	**1**	**3554**	**3554**	**3.55**	**Manual Finishing**
Halomonas sp. KO116	Illumina	110	373	194	5.13	SPAdes*
		120	315	115	5.19	ABySS
	Illumina + PacBio	30	4654	4654	5.19	SPAdes
	PacBio only	**1**	**4649**	**4649**	**4.65**	**SMRTanalysis v 2.2**
Pelosinus fermentans UFO1	Illumina	175	1025	637	5.13	SPAdes
		131	169	78	5.03	ABySS*
	Illumina + PacBio	147	4498	4498	5.19	SPAdes
	PacBio only	**1**	**5115**	**5115**	**5.12**	**SMRTanalysis v 2.1**
Pelosinus fermentans JBW45	Illumina	70	477	244	5.3	SPAdes*
		114	318	110	5.4	ABySS
	Illumina + PacBio	1	5381	5381	5.38	SPAdes
	PacBio only	**1**	**5381**	**5381**	**5.38**	**SMRTanalysis v 2.2**
Clostridium paradoxum JW/YL-7T	Illumina	661	293	121	2.23	SPAdes
		43	235	74	1.84	ABySS*
	Illumina + PacBio	612	1061	323	2.26	SPAdes
	PacBio only	**3**	**1855**	**1855**	**1.93**	**SMRTanalysis v 2.2**

Table 5.3 continued…

Organism	NGS Technology	No. of contigs	Maximum Contig Size (kb)	N50 (kb)	Genome Size (Mb)	Software
Bacteroides cellulosolvens DSM 2933	Illumina	194	1143	271	6.81	SPAdes
		172	358	130	6.99	ABySS*
	Illumina + PacBio	122	3522	3522	6.91	SPAdes
	PacBio only	12	2261	1340	6.94	SMRTanalysis v 2.0
	PacBio only	3	6349	6349	6.88	SMRTanalysis v 2.2
	PacBio only	**1**	**6878**	**6878**	**6.87**	**Manual Finishing**
Clostridium pasteurianum ATCC 6013	Illumina	6	4108	4108	4.36	SPAdes*
		101	207	73	4.35	ABySS
	Illumina + PacBio	9	4022	4022	4.36	SPAdes
	PacBio only	**2**	**4374**	**4374**	**4.39**	**SMRTanalysis v 2.2**
Clostridium autoethanogenum DSM 10061	Illumina	53	462	251	4.3	SPAdes*
		61	399	196	4.39	ABySS
	Illumina + PacBio	3	435	435	4.36	SPAdes
	PacBio only	**1**	**435**	**435**	**4.35**	**SMRTanalysis v 2.0**

The best assembly for each genome on shown in bold. The best draft assembly achieved with only the Illumina data are marked with *.

Table 5.4: Number of modifications suggested by Pilon and impact on number of protein coding genes.

Organism	No. of predicted ORFs in original assembly	No. of changes suggested by Pilon	No. of predicted ORFs in corrected assembly	No. of new ORFs in corrected assembly	No. of longer ORFs in corrected assembly	No. of shorter ORFs in corrected assembly
C. thermocellum AD2	3072	59	3077	5	15	0
Halomonas sp. KO116	4527	111	4192	23	11	29
Pelosinus sp. UFO1	4793	26	4790	1	5	1
Pelosinus sp. JBW45	4829	118	4771	3	62	7
B. cellulosolvens ATCC 35603	5897	542	5744	12	203	68

Note: Pilon was run only for the single contig genome assemblies.

Table 5.5: Comparison of Open Reading Frames (ORFs) predicted in draft and finished genome assemblies.

Organism	Total ORFs	No. of unchanged ORFs	No. of longer ORFs	No. of shorter ORFs	No. of new ORFs
C. thermocellum AD2	3224	2987	183	34	20
Halomonas sp. KO116	4500	4222	184	93	1
Pelosinus sp. UFO1	4811	4720	63	23	5
Pelosinus sp. JBW45	4800	4703	71	7	19
C. paradoxum JW/YL-7T	1963	1897	32	27	7
B. cellulosolvens ATCC 35603	6184	5760	342	68	14
C. pasteurianum ATCC 6013	4062	4034	9	18	1

Table 5.6: Comparison of Open Reading Frames (ORFs) predicted in draft and finished genome assemblies.

Organism	[a]Total ORFs	[b]No. of unchanged ORFs	No. of longer ORFs	No. of shorter ORFs	No. of new ORFs
C. thermocellum AD2	3224	2987	183	34	20
Halomonas sp. KO116	4500	4222	184	93	1
Pelosinus sp. UFO1	4811	4720	63	23	5
Pelosinus sp. JBW45	4800	4703	71	7	19
C. paradoxum JW/YL-7T	1963	1897	32	27	7
B. cellulosolvens ATCC 35603	6184	5760	342	68	14
C. pasteurianum ATCC 6013	4062	4034	9	18	1

[a]Total number of open reading frames predicted in improved genome assembly by Prodigal gene calling algorithm.

[b]Number of open reading frames in improved genome assemblies as compared with draft assemblies.

Table 5.7: Annotations, coordinates and locus tags associated with the gap regions within the Illumina assembly.

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. thermocellum AD2	16s rRNA	rRNA	17,295	18,810	1,516	forward	NA	None
	Transposase DDE domain CDS	CDS	181,219	181,857	639	reverse	AD2_0168	None
	Integrase catalytic region CDS	CDS	384,056	384,898	843	reverse	AD2_0340	None
	hypothetical protein CDS	CDS	384,915	385,208	294	reverse	AD2_0341	None
	Hedgehog/intein hint domain-containing protein CDS	CDS	386,029	386,166	138	forward	AD2_0342	None
	hypothetical protein CDS	CDS	388,401	388,583	183	forward	AD2_0345	Partial
	transposase IS200-family protein CDS	CDS	412,322	412,798	477	forward	AD2_0358	None
	hypothetical protein CDS	CDS	412,978	414,012	1,035	forward	AD2_0359	None
	Primase 1 CDS	CDS	418,685	418,912	228	forward	AD2_0366	Partial
	Hedgehog/intein hint domain-containing protein CDS	CDS	419,309	421,156	1,848	forward	AD2_0367	Partial
	transposase mutator type CDS	CDS	423,910	425,034	1,125	reverse	AD2_0371	None
	hypothetical protein CDS	CDS	425,619	425,966	348	forward	AD2_0372	None
	hypothetical protein CDS	CDS	426,109	426,888	780	forward	AD2_0373	None
	Hedgehog/intein hint domain-containing protein CDS	CDS	426,890	427,756	867	forward	AD2_0374	Partial
	Integrase catalytic region CDS	CDS	435,886	436,728	843	reverse	AD2_0386	None
	hypothetical protein CDS	CDS	436,745	437,038	294	reverse	AD2_0387	None
	transposase mutator type CDS	CDS	437,092	438,078	987	reverse	AD2_0388	None
	hypothetical protein CDS	CDS	438,651	438,998	348	forward	AD2_0389	None
	Hedgehog/intein hint domain-containing protein CDS	CDS	439,141	440,814	1,674	forward	AD2_0390	Partial
	DNA polymerase beta domain protein region CDS	CDS	443,200	444,954	1,755	forward	AD2_0394	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. thermocellum AD2	Integrase catalytic region CDS	CDS	447,525	448,472	948	reverse	AD2_0397	None
	hypothetical protein CDS	CDS	448,660	448,878	219	forward	AD2_0398	None
	hypothetical protein CDS	CDS	452,698	452,991	294	forward	AD2_0404	None
	HTH-like domain CDS	CDS	453,008	453,304	297	forward	AD2_0405	None
	Integrase catalytic region CDS	CDS	453,267	453,842	576	forward	AD2_0406	None
	Primase 1 CDS	CDS	455,883	456,233	351	forward	AD2_0409	Partial
	Hedgehog/intein hint domain-containing protein CDS	CDS	456,629	457,981	1,353	forward	AD2_0410	None
	transposase IS200-family protein CDS	CDS	535,371	535,847	477	forward	AD2_0477	None
	transposase mutator type CDS	CDS	558,593	559,816	1,224	reverse	AD2_0488	None
	Dockerin type 1 protein CDS	CDS	566,496	568,709	2,214	forward	AD2_0497	Partial
	Transposase DDE domain CDS	CDS	569,284	570,231	948	forward	AD2_0498	None
	hypothetical protein CDS	CDS	611,185	614,154	2,970	forward	AD2_0534	Partial
	hypothetical protein CDS	CDS	616,015	618,717	2,703	forward	AD2_0539	Partial
	hypothetical protein CDS	CDS	620,340	621,002	663	forward	AD2_0541	Partial
	Ig domain protein group 2 domain protein CDS	CDS	647,643	654,545	6,903	forward	AD2_0567	Partial
	transposase IS3/IS911 family protein CDS	CDS	766,094	766,387	294	forward	AD2_0676	None
	Integrase catalytic region CDS	CDS	766,404	767,246	843	forward	AD2_0677	None
	hypothetical protein CDS	CDS	845,603	846,043	441	forward	AD2_0744	None
	transposase IS204/IS1001/IS1096/IS1165 family protein CDS	CDS	846,114	846,974	861	forward	AD2_0745	None
	transposase mutator type CDS	CDS	920,732	921,955	1,224	reverse	AD2_0817	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. thermocellum AD2	RHS repeat-associated core domain-containing protein CDS	CDS	925,515	930,167	4,653	forward	AD2_0822	Partial
	methyl-accepting chemotaxis sensory transducer CDS	CDS	962,174	964,045	1,872	forward	AD2_0854	Partial
	transposase IS200-family protein CDS	CDS	1,097,451	1,097,927	477	forward	AD2_0959	None
	hypothetical protein CDS	CDS	1,180,972	1,181,265	294	forward	AD2_1037	None
	HTH-like domain CDS	CDS	1,181,282	1,181,572	291	forward	AD2_1038	None
	hypothetical protein CDS	CDS	1,181,624	1,182,847	1,224	forward	AD2_1039	None
	hypothetical protein CDS	CDS	1,183,190	1,183,645	456	forward	AD2_1040	None
	hypothetical protein CDS	CDS	1,219,332	1,220,279	948	reverse	AD2_1067	None
	hypothetical protein CDS	CDS	1,285,393	1,286,235	843	reverse	AD2_1114	None
	hypothetical protein CDS	CDS	1,286,252	1,286,545	294	reverse	AD2_1115	None
	hypothetical protein CDS	CDS	1,311,384	1,312,331	948	reverse	AD2_1136	None
	hypothetical protein CDS	CDS	1,365,807	1,367,030	1,224	forward	AD2_1183	None
	hypothetical protein CDS	CDS	1,371,176	1,372,018	843	forward	AD2_1187	None
	hypothetical protein CDS	CDS	1,388,062	1,389,285	1,224	reverse	AD2_1202	Partial
	Transposase DDE domain CDS	CDS	1,437,226	1,438,173	948	reverse	AD2_1243	Partial
	Cellulose 1,4-beta-cellobiosidase CDS	CDS	1,438,308	1,440,995	2,688	forward	AD2_1244	Partial
	Cellulose 1,4-beta-cellobiosidase., Cellulase CDS	CDS	1,441,522	1,445,214	3,693	forward	AD2_1245	Partial
	transposase mutator type CDS	CDS	1,475,383	1,476,603	1,221	forward	AD2_1271	None
	transposase IS200-family protein CDS	CDS	1,546,506	1,546,982	477	forward	AD2_1344	None
	transposase IS200-family protein CDS	CDS	1,547,238	1,547,714	477	forward	AD2_1345	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. thermocellum AD2	hypothetical protein CDS	CDS	1,623,091	1,623,912	822	forward	AD2_1405	Partial
	Integrase catalytic region CDS	CDS	1,624,101	1,624,943	843	reverse	AD2_1406	Partial
	Integrase catalytic region CDS	CDS	1,639,610	1,640,761	1,152	reverse	AD2_1419	None
	hypothetical protein CDS	CDS	1,640,825	1,641,265	441	forward	AD2_1420	None
	lipolytic protein G-D-S-L family CDS	CDS	1,878,799	1,880,385	1,587	reverse	AD2_1636	Partial
	hypothetical protein CDS	CDS	1,882,795	1,884,018	1,224	reverse	AD2_1639	None
	transposase IS200-family protein CDS	CDS	1,934,591	1,935,067	477	reverse	AD2_1684	None
	Dockerin type 1 protein CDS	CDS	2,011,202	2,017,237	6,036	reverse	AD2_1759	Partial
	hypothetical protein CDS	CDS	2,017,274	2,018,347	1,074	reverse	AD2_1760	None
	transposase IS3/IS911 family protein CDS	CDS	2,018,537	2,018,830	294	forward	AD2_1761	None
	Integrase catalytic region CDS	CDS	2,018,847	2,019,689	843	forward	AD2_1762	None
	hypothetical protein CDS	CDS	2,019,921	2,020,940	1,020	reverse	AD2_1763	Partial
	hypothetical protein CDS	CDS	2,073,500	2,074,723	1,224	forward	AD2_1813	None
	MutS2 protein CDS	CDS	2,125,828	2,128,209	2,382	reverse	AD2_1860	None
	transposase mutator type CDS	CDS	2,276,498	2,277,721	1,224	forward	AD2_1997	None
	protein of unknown function DUF1910 CDS	CDS	2,278,024	2,279,292	1,269	reverse	AD2_1998	None
	hypothetical protein CDS	CDS	2,279,319	2,279,663	345	reverse	AD2_1999	Partial
	protein of unknown function DUF1910 CDS	CDS	2,280,170	2,281,411	1,242	reverse	AD2_2000	Partial
	hypothetical protein CDS	CDS	2,281,438	2,281,782	345	reverse	AD2_2001	None
	protein of unknown function DUF1910 CDS	CDS	2,282,271	2,283,509	1,239	reverse	AD2_2002	None
	APHP domain protein CDS	CDS	2,333,792	2,358,892	25,101	reverse	AD2_2042	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. thermocellum AD2	transposase IS200-family protein CDS	CDS	2,518,940	2,519,416	477	reverse	AD2_2178	None
	RHS repeat-associated core domain-containing protein CDS	CDS	2,531,816	2,537,569	5,754	reverse	AD2_2188	Partial
	Intergenic	Intergenic	2,560,779	2,561,100	321	forward	Intergenic	None
	hypothetical protein CDS	CDS	2,599,985	2,600,461	477	reverse	AD2_2235	None
	hypothetical protein CDS	CDS	2,612,725	2,613,201	477	reverse	AD2_2246	None
	hypothetical protein CDS	CDS	2,640,799	2,642,022	1,224	forward	AD2_2273	None
	hypothetical protein CDS	CDS	2,642,208	2,642,456	249	reverse	AD2_2274	None
	hypothetical protein CDS	CDS	2,642,469	2,642,702	234	reverse	AD2_2275	Partial
	hypothetical protein CDS	CDS	2,642,781	2,643,035	255	reverse	AD2_2276	Partial
	RHS repeat-associated core domain-containing protein CDS	CDS	2,643,002	2,648,299	5,298	reverse	AD2_2277	Partial
	transposase mutator type CDS	CDS	2,648,656	2,649,876	1,221	reverse	AD2_2278	None
	Ankyrin repeat-containing domain-containing protein CDS	CDS	2,650,048	2,650,977	930	reverse	AD2_2279	Partial
	RHS repeat-associated core domain-containing protein CDS	CDS	2,650,974	2,656,778	5,805	reverse	AD2_2280	Partial
	copper amine oxidase-like domain-containing protein CDS	CDS	2,660,647	2,661,447	801	reverse	AD2_2283	None
	copper amine oxidase-like domain-containing protein CDS	CDS	2,661,521	2,662,324	804	reverse	AD2_2284	None
	copper amine oxidase-like domain-containing protein CDS	CDS	2,662,503	2,663,291	789	reverse	AD2_2285	Partial
	hypothetical protein CDS	CDS	2,663,538	2,664,050	513	reverse	AD2_2286	Partial
	hypothetical protein CDS	CDS	2,664,285	2,664,386	102	reverse	AD2_2287	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. thermocellum AD2	transposase IS3/IS911 family protein CDS	CDS	2,664,819	2,665,112	294	forward	AD2_2288	None
	Integrase catalytic region CDS	CDS	2,665,129	2,665,971	843	forward	AD2_2289	None
	hypothetical protein CDS	CDS	2,671,221	2,672,444	1,224	reverse	AD2_2295	None
	hypothetical protein CDS	CDS	2,799,485	2,799,961	477	reverse	AD2_2401	None
	tRNA-Asn1	tRNA	2,801,273	2,801,348	76	reverse	RNA_47	None
	5s rRNA	rRNA	2,801,355	2,801,470	116	reverse	NA	None
	23s rRNA	rRNA	2,801,562	2,804,467	2,906	reverse	NA	None
	hypothetical protein CDS	CDS	2,803,013	2,803,216	204	reverse	AD2_2403	None
	hypothetical protein CDS	CDS	2,803,431	2,803,904	474	forward	AD2_2404	None
	tRNA-Ala3	tRNA	2,804,669	2,804,744	76	reverse	RNA_46	None
	16s rRNA	rRNA	2,804,857	2,806,372	1,516	reverse	NA	None
	transposase IS200-family protein CDS	CDS	2,823,142	2,823,618	477	reverse	AD2_2421	None
	cellulosome anchoring protein cohesin region CDS	CDS	2,845,581	2,846,978	1,398	reverse	AD2_2438	Partial
	cellulosome anchoring protein cohesin region CDS	CDS	2,847,160	2,850,627	3,468	reverse	AD2_2439	Partial
	5s rRNA	rRNA	2,865,860	2,865,975	116	reverse	NA	None
	23s rRNA	rRNA	2,866,067	2,868,972	2,906	reverse	NA	None
	hypothetical protein CDS	CDS	2,867,518	2,867,721	204	reverse	AD2_2453	None
	hypothetical protein CDS	CDS	2,867,936	2,868,409	474	forward	AD2_2454	None
	16s rRNA	rRNA	2,869,261	2,870,776	1,516	reverse	NA	Partial
	hypothetical protein CDS	CDS	2,878,018	2,878,458	441	reverse	AD2_2462	None
	Mu-like prophage protein Com CDS	CDS	2,878,885	2,879,022	138	reverse	AD2_2463	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. thermocellum AD2	hypothetical protein CDS	CDS	2,879,015	2,879,404	390	reverse	AD2_2464	None
	hypothetical protein CDS	CDS	2,879,518	2,880,075	558	reverse	AD2_2465	None
	RHS repeat-associated core domain-containing protein CDS	CDS	2,880,075	2,888,996	8,922	reverse	AD2_2466	Partial
	transposase mutator type CDS	CDS	2,889,367	2,890,590	1,224	reverse	AD2_2467	None
	hypothetical protein CDS	CDS	2,973,461	2,974,831	1,371	reverse	AD2_2542	Partial
	Endo-1,4-beta-xylanase., Cellulase CDS	CDS	2,978,258	2,979,631	1,374	forward	AD2_2546	Partial
	Transposase DDE domain CDS	CDS	2,984,981	2,985,928	948	reverse	AD2_2552	None
	hypothetical protein CDS	CDS	2,995,957	2,996,250	294	forward	AD2_2561	None
	Integrase catalytic region CDS	CDS	2,996,267	2,997,109	843	forward	AD2_2562	None
	Integrase catalytic region CDS	CDS	2,997,939	2,998,781	843	reverse	AD2_2563	None
	hypothetical protein CDS	CDS	2,998,798	2,999,091	294	reverse	AD2_2564	None
	hypothetical protein CDS	CDS	2,999,129	2,999,596	468	reverse	AD2_2565	None
	5s rRNA	rRNA	3,054,261	3,054,376	116	reverse	NA	None
	23s rRNA	rRNA	3,054,468	3,057,373	2,906	reverse	NA	None
	hypothetical protein CDS	CDS	3,055,919	3,056,122	204	reverse	AD2_2629	None
	Protein of unknown function DUF4323 CDS	CDS	3,056,337	3,056,810	474	forward	AD2_2630	None
	tRNA-Ile1	tRNA	3,057,577	3,057,653	77	reverse	RNA_40	None
	16s rRNA	rRNA	3,057,767	3,059,282	1,516	reverse	NA	None
	Glucan endo-1,3-beta-D-glucosidase CDS	CDS	3,119,560	3,123,966	4,407	reverse	AD2_2675	Partial
	transposase mutator type CDS	CDS	3,390,151	3,391,374	1,224	forward	AD2_2924	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. thermocellum AD2	S-layer domain-containing protein CDS	CDS	3,451,451	3,454,726	3,276	reverse	AD2_2985	Partial
Halomonas sp. KO116	AraC family transcriptional regulator CDS	CDS	196,773	197,675	903	reverse	KO116_RS00955	Partial
	4-hydroxybenzoate 3-monooxygenase CDS	CDS	197,822	199,006	1,185	forward	KO116_RS00960	Partial
	integrase CDS	CDS	414,144	414,413	270	reverse	KO116_RS02010	Partial
	reverse transcriptase CDS	CDS	414,713	416,224	1,512	reverse	KO116_RS02015	None
	transcriptional regulator CDS	CDS	418,263	419,279	1,017	forward	KO116_RS02035	Partial
	Pseudogene	gene	487,461	488,761	1,301	forward	KO116_RS02335	None
	transposase CDS	CDS	500,841	501,380	540	reverse	KO116_RS02385	Partial
	hypothetical protein CDS	CDS	501,451	501,804	354	reverse	KO116_RS02390	Partial
	Pseudogene	gene	817,063	818,117	1,055	reverse	KO116_RS03825	Partial
	transposase CDS	CDS	923,998	924,537	540	reverse	KO116_RS04330	Partial
	hypothetical protein CDS	CDS	924,608	925,012	405	reverse	KO116_RS04335	Partial
	transposase CDS	CDS	947,472	948,833	1,362	reverse	KO116_RS04450	Partial
	16S rRNA	rRNA	975,057	976,595	1,539	forward	KO116_RS04550	Partial
	23S rRNA	rRNA	977,312	980,223	2,912	forward	KO116_RS04565	None
	hypothetical protein CDS	CDS	1,002,014	1,003,063	1,050	reverse	KO116_RS04680	Partial
	transcriptional regulator CDS	CDS	1,031,763	1,032,779	1,017	reverse	KO116_RS04835	Partial
	transcriptional regulator CDS	CDS	1,036,212	1,037,228	1,017	reverse	KO116_RS04865	Partial
	transposase CDS	CDS	1,281,483	1,282,844	1,362	reverse	KO116_RS05940	Partial
	transposase CDS	CDS	1,403,081	1,404,286	1,206	reverse	KO116_RS06500	Partial
	hypothetical protein CDS	CDS	1,743,188	1,744,237	1,050	forward	KO116_RS08050	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Halomonas sp. KO116	transcriptional regulator CDS	CDS	1,758,108	1,759,124	1,017	forward	KO116_RS08110	Partial
	16S rRNA	rRNA	1,880,907	1,882,446	1,540	forward	KO116_RS08670	Partial
	23S rRNA	rRNA	1,882,739	1,885,651	2,913	forward	KO116_RS08675	None
	tRNA-Ile	tRNA	2,207,212	2,207,288	77	forward	KO116_RS10225	None
	tRNA-Ala	tRNA	2,207,386	2,207,461	76	forward	KO116_RS10230	None
	23S rRNA	rRNA	2,207,722	2,210,635	2,914	forward	KO116_RS10235	None
	transposase CDS	CDS	2,620,342	2,620,659	318	forward	KO116_RS12165	Partial
	transposase CDS	CDS	2,620,659	2,621,558	900	forward	KO116_RS12170	Partial
	transposase CDS	CDS	2,674,027	2,674,926	900	reverse	KO116_RS12400	Partial
	transposase CDS	CDS	2,674,926	2,675,243	318	reverse	KO116_RS12405	Partial
	hypothetical protein CDS	CDS	2,990,252	2,991,301	1,050	forward	KO116_RS13820	Partial
	hypothetical protein CDS	CDS	3,183,627	3,184,910	1,284	reverse	KO116_RS14675	Partial
	hypothetical protein CDS	CDS	3,184,955	3,185,635	681	reverse	KO116_RS14680	None
	transcriptional regulator CDS	CDS	3,186,026	3,187,045	1,020	forward	KO116_RS14685	Partial
	transposase CDS	CDS	3,244,583	3,245,884	1,302	reverse	KO116_RS14945	None
	transposase CDS	CDS	3,390,351	3,391,985	1,635	reverse	KO116_RS15610	None
	transposase CDS	CDS	3,392,031	3,392,399	369	reverse	KO116_RS15615	None
	hypothetical protein CDS	CDS	3,392,399	3,392,716	318	reverse	KO116_RS15620	Partial
	amino acid adenylation protein CDS	CDS	3,668,246	3,682,060	13,815	reverse	KO116_RS16840	Partial
	transcriptional regulator CDS	CDS	3,735,629	3,736,645	1,017	forward	KO116_RS17020	Partial
	hypothetical protein CDS	CDS	4,252,675	4,253,193	519	forward	KO116_RS19440	Partial
	16S rRNA	rRNA	4,299,609	4,301,148	1,540	reverse	KO116_RS19660	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Halomonas sp. KO116	transposase CDS	CDS	4,445,339	4,446,544	1,206	forward	KO116_RS20365	Partial
	23S rRNA	rRNA	4,562,221	4,565,132	2,912	reverse	KO116_RS20950	None
	16S rRNA	rRNA	4,565,367	4,566,906	1,540	reverse	KO116_RS20955	Partial
	23S rRNA	rRNA	4,598,659	4,601,569	2,911	reverse	KO116_RS21100	None
	16S rRNA	rRNA	4,601,862	4,603,401	1,540	reverse	KO116_RS21105	Partial
Pelosinus sp. UFO1	16S rRNA	rRNA	1	1,581	1,581	forward	UFO1_RS00005	None
	tRNA-Ile	tRNA	1,657	1,733	77	forward	UFO1_RS00010	None
	tRNA-Ala	tRNA	1,797	1,872	76	forward	UFO1_RS00015	None
	23S rRNA	rRNA	2,147	5,079	2,933	forward	UFO1_RS00020	None
	5S rRNA	rRNA	5,179	5,295	117	forward	UFO1_RS00025	None
	16S rRNA	rRNA	5,620	7,208	1,589	forward	UFO1_RS00030	None
	23S rRNA	rRNA	7,488	10,420	2,933	forward	UFO1_RS00035	None
	5S rRNA	rRNA	10,511	10,627	117	forward	UFO1_RS00040	None
	hypothetical protein CDS	CDS	89653	93099	>3447	forward	UFO1_RS00405	Partial
	hypothetical protein CDS	CDS	95575	98829	>3255	forward	UFO1_RS00410	Partial
	Intergenic	NA	194,451	194,575	124	NA	NA	None
	hypothetical protein CDS	CDS	352,424	357,820	5,397	forward	UFO1_RS01470	Partial
	16S rRNA	rRNA	603,644	605,232	1,589	forward	UFO1_RS02550	None
	tRNA-Ile	tRNA	605,346	605,422	77	forward	UFO1_RS02555	None
	tRNA-Ala	tRNA	605,486	605,561	76	forward	UFO1_RS02560	None
	23S rRNA	rRNA	605,829	608,761	2,933	forward	UFO1_RS02565	None
	5S rRNA	rRNA	608,998	609,114	117	forward	UFO1_RS02570	None
	16S rRNA	rRNA	609,439	611,025	1,587	forward	UFO1_RS02575	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Pelosinus sp. UFO1	23S rRNA	rRNA	611,308	614,240	2,933	forward	UFO1_RS02580	None
	5S rRNA	rRNA	614,331	614,447	117	forward	UFO1_RS02585	None
	tRNA-Thr	tRNA	650,590	650,665	76	forward	UFO1_RS02785	None
	tRNA-Met	tRNA	650,739	650,815	77	forward	UFO1_RS02790	None
	tRNA-Thr	tRNA	650,895	650,970	76	forward	UFO1_RS02795	None
	tRNA-Tyr	tRNA	650,984	651,068	85	forward	UFO1_RS02800	None
	tRNA-Met	tRNA	651,219	651,294	76	forward	UFO1_RS02805	None
	tRNA-Thr	tRNA	651,342	651,417	76	forward	UFO1_RS02810	None
	tRNA-Met	tRNA	651,483	651,559	77	forward	UFO1_RS02815	Partial
	hypothetical protein CDS	CDS	1238971	1240611	>1641	reverse	UFO1_RS05515	Partial
	tRNA-Val	tRNA	1,638,031	1,638,106	76	forward	UFO1_RS07400	None
	tRNA-Asp	tRNA	1,638,127	1,638,202	76	forward	UFO1_RS07405	None
	tRNA-Gly	tRNA	1,638,206	1,638,280	75	forward	UFO1_RS07410	None
	16S rRNA	rRNA	1,638,511	1,640,197	1,687	forward	UFO1_RS07415	None
	tRNA-Ala	tRNA	1,640,371	1,640,446	76	forward	UFO1_RS07420	None
	23S rRNA	rRNA	1,640,650	1,643,582	2,933	forward	UFO1_RS07425	None
	5S rRNA	rRNA	1,643,788	1,643,904	117	forward	UFO1_RS07430	None
	tRNA-Asn	tRNA	1,643,909	1,643,984	76	forward	UFO1_RS07435	Partial
	16S rRNA	rRNA	1,649,041	1,650,727	1,687	forward	UFO1_RS07470	None
	23S rRNA	rRNA	1,651,217	1,654,152	2,936	forward	UFO1_RS07475	Partial
	hypothetical protein CDS	CDS	1,804,399	1,805,661	1,263	forward	UFO1_RS08205	Partial
	hypothetical protein CDS	CDS	1,805,784	1,806,803	1,020	forward	UFO1_RS08210	Partial
	hypothetical protein CDS	CDS	1,806,839	1,807,183	345	forward	UFO1_RS08215	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Pelosinus sp. UFO1	hypothetical protein CDS	CDS	1,807,138	1,807,497	360	forward	UFO1_RS08220	Partial
	tRNA-Phe	tRNA	1,820,776	1,820,851	76	forward	UFO1_RS08290	None
	tRNA-Gly	tRNA	1,820,855	1,820,929	75	forward	UFO1_RS08295	Partial
	16S rRNA	rRNA	2,074,283	2,075,871	1,589	forward	UFO1_RS09515	None
	tRNA-Ile	tRNA	2,075,985	2,076,061	77	forward	UFO1_RS09520	None
	tRNA-Ala	tRNA	2,076,125	2,076,200	76	forward	UFO1_RS09525	None
	23S rRNA	rRNA	2,076,468	2,079,400	2,933	forward	UFO1_RS09530	None
	5S rRNA	rRNA	2,715,923	2,716,039	117	reverse	UFO1_RS12955	None
	23S rRNA	rRNA	2,716,139	2,719,071	2,933	reverse	UFO1_RS12960	None
	16S rRNA	rRNA	2,719,352	2,721,038	1,687	reverse	UFO1_RS12965	None
	5S rRNA	rRNA	2,738,345	2,738,461	117	reverse	UFO1_RS13030	Partial
	23S rRNA	rRNA	2,738,553	2,741,485	2,933	reverse	UFO1_RS13035	None
	16S rRNA	rRNA	2,741,910	2,743,597	1,688	reverse	UFO1_RS13040	None
	hypothetical protein CDS	CDS	3,061,021	3,065,874	4,854	reverse	UFO1_RS14530	Partial
	hypothetical protein CDS	CDS	3,915,825	3,916,448	624	reverse	UFO1_RS18535	None
	hypothetical protein CDS	CDS	3,916,655	3,917,275	621	forward	UFO1_RS18540	None
	hypothetical protein CDS	CDS	3,917,565	3,918,332	768	forward	UFO1_RS18545	None
	ammonia monooxygenase CDS	CDS	3,918,465	3,919,526	1,062	forward	UFO1_RS18550	None
	hypothetical protein CDS	CDS	3,919,551	3920462	>912	forward	UFO1_RS18555	None
	3-ketoacyl-ACP reductase CDS	CDS	3,920,643	3,921,389	747	reverse	UFO1_RS18560	None
	hypothetical protein CDS	CDS	3,921,416	3,921,889	474	reverse	UFO1_RS18565	None
	glycosyl hydrolase CDS	CDS	3,922,004	3,923,185	1,182	reverse	UFO1_RS18570	None
	sodium:solute symporter CDS	CDS	3923309	3,924,670	>1362	reverse	UFO1_RS18575	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Pelosinus sp. UFO1	hypothetical protein CDS	CDS	3,924,808	3,925,011	204	reverse	UFO1_RS18580	None
	chemotaxis protein CDS	CDS	3,925,731	3,927,692	1,962	reverse	UFO1_RS18585	None
	membrane protein CDS	CDS	3,927,921	3,929,198	1,278	reverse	UFO1_RS18590	None
	alcohol dehydrogenase CDS	CDS	3,929,292	3,930,452	1,161	reverse	UFO1_RS18595	None
	phosphoglycerate dehydrogenase CDS	CDS	3,930,649	3,931,626	978	reverse	UFO1_RS18600	None
	dihydrodipicolinate synthase CDS	CDS	3,931,764	3,932,639	876	reverse	UFO1_RS18605	None
	pdxA CDS	CDS	3,932,684	3,933,685	1,002	reverse	UFO1_RS18610	None
	type III effector CDS	CDS	3,933,863	3,935,143	1,281	reverse	UFO1_RS18615	None
	Fis family transcriptional regulator CDS	CDS	3,935,290	3,937,194	1,905	reverse	UFO1_RS18620	None
	glucose-6-phosphate isomerase CDS	CDS	3,937,549	3,939,033	1,485	reverse	UFO1_RS18625	None
	fructose-bisphosphate aldolase CDS	CDS	3,939,135	3,940,064	930	reverse	UFO1_RS18630	None
	transporter CDS	CDS	3,940,476	3,941,225	750	reverse	UFO1_RS18635	None
	hypothetical protein CDS	CDS	3,941,765	3,942,316	552	forward	UFO1_RS18640	None
	membrane protein CDS	CDS	3,942,450	3,942,986	537	reverse	UFO1_RS18645	None
	pyridoxamine kinase CDS	CDS	3,943,018	3,943,872	855	reverse	UFO1_RS18650	None
	hypothetical protein CDS	CDS	3,944,275	3,944,520	246	reverse	UFO1_RS18655	None
	multidrug transporter AcrB CDS	CDS	3,944,532	3,947,624	3,093	reverse	UFO1_RS18660	None
	RND transporter CDS	CDS	3,947,621	3,948,733	1,113	reverse	UFO1_RS18665	None
	arabinose isomerase CDS	CDS	3,949,283	3,950,788	1,506	reverse	UFO1_RS18670	None
	ribulokinase CDS	CDS	3,950,846	3,952,522	1,677	reverse	UFO1_RS18675	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Pelosinus sp. UFO1	transcriptional regulator CDS	CDS	3,952,842	3,954,044	1,203	reverse	UFO1_RS18680	None
	hypothetical protein CDS	CDS	3,954,290	3,955,312	1,023	reverse	UFO1_RS18685	None
	galactose-1-phosphate uridylyltransferase CDS	CDS	3,955,539	3,957,032	1,494	reverse	UFO1_RS18690	None
	UDP-glucose 4-epimerase CDS	CDS	3,957,045	3,958,037	993	reverse	UFO1_RS18695	None
	membrane protein CDS	CDS	3,958,324	3,958,668	345	forward	UFO1_RS18700	None
	RNA-binding protein CDS	CDS	3,958,744	3,959,094	351	reverse	UFO1_RS18705	None
	signal peptidase CDS	CDS	3,959,671	3,960,237	567	forward	UFO1_RS18710	None
	hypothetical protein CDS	CDS	3,960,463	3,961,092	630	forward	UFO1_RS18715	None
	siderophore-interacting protein CDS	CDS	3,961,182	3,961,601	420	reverse	UFO1_RS18720	None
	glycoside hydrolase CDS	CDS	3,961,748	3,962,308	561	reverse	UFO1_RS18725	None
	hypothetical protein CDS	CDS	3,962,336	3,962,800	465	reverse	UFO1_RS18730	None
	metal transporter CDS	CDS	3,962,887	3,966,075	3,189	reverse	UFO1_RS18735	None
	RND transporter CDS	CDS	3,966,087	3,967,367	1,281	reverse	UFO1_RS18740	None
	transporter CDS	CDS	3,967,396	3,968,694	1,299	reverse	UFO1_RS18745	None
	histidine kinase CDS	CDS	3,969,332	3,970,507	1,176	reverse	UFO1_RS18750	None
	PhoP family transcriptional regulator CDS	CDS	3,970,500	3,971,195	696	reverse	UFO1_RS18755	None
	LemA family protein CDS	CDS	3,971,281	3,971,838	558	reverse	UFO1_RS18760	None
	hypothetical protein CDS	CDS	3,971,987	3,972,175	189	reverse	UFO1_RS18765	None
	hypothetical protein CDS	CDS	3,972,477	3,973,154	678	reverse	UFO1_RS18770	None
	glycoside hydrolase CDS	CDS	3,973,526	3,974,203	678	reverse	UFO1_RS18775	None
	hypothetical protein CDS	CDS	3,975,248	3,976,930	1,683	reverse	UFO1_RS18780	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Pelosinus sp. UFO1	hypothetical protein CDS	CDS	3,977,091	3,977,843	753	reverse	UFO1_RS18785	None
	hypothetical protein CDS	CDS	3,978,115	3,978,357	243	reverse	UFO1_RS18790	None
	hypothetical protein CDS	CDS	3,978,412	3,978,621	210	reverse	UFO1_RS18795	None
	cytochrome oxidase biogenesis protein Surf12C CDS	CDS	3,978,676	3,979,185	510	reverse	UFO1_RS18800	None
	hypothetical protein CDS	CDS	3,979,452	3,979,661	210	reverse	UFO1_RS18805	None
	hypothetical protein CDS	CDS	3,980,274	3,981,023	750	reverse	UFO1_RS18810	None
	sirohydrochlorin cobaltochelatase CDS	CDS	3,981,098	3,981,934	837	reverse	UFO1_RS18815	None
	isoprenylcysteine carboxyl methyltransferase CDS	CDS	3,982,483	3,983,145	663	reverse	UFO1_RS18820	None
	diacylglyceryl transferase CDS	CDS	3,983,138	3,983,899	762	reverse	UFO1_RS18825	None
	hypothetical protein CDS	CDS	3,984,123	3,984,557	435	reverse	UFO1_RS18830	None
	hypothetical protein CDS	CDS	3,984,642	3,984,836	195	reverse	UFO1_RS18835	None
	hypothetical protein CDS	CDS	3,984,886	3,985,146	261	reverse	UFO1_RS18840	None
	hypothetical protein CDS	CDS	4,384,875	4,386,146	1,272	reverse	UFO1_RS20640	Partial
	Intergenic	NA	4,387,031	4,387,433	402	NA	NA	None
	hypothetical protein CDS	CDS	4,408,553	4,410,127	1,575	reverse	UFO1_RS20765	Partial
	23S rRNA	rRNA	4,711,705	4,714,637	2,933	reverse	UFO1_RS22115	None
	16S rRNA	rRNA	4,714,917	4,716,505	1,589	reverse	UFO1_RS22120	None
	23S rRNA	rRNA	4,717,318	4,720,250	2,933	reverse	UFO1_RS22135	None
	16S rRNA	rRNA	4,720,530	4,722,118	1,589	reverse	UFO1_RS22140	None
	5S rRNA	rRNA	4,722,445	4,722,561	117	reverse	UFO1_RS22145	None
	23S rRNA	rRNA	4,722,658	4,725,593	2,936	reverse	UFO1_RS22150	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Pelosinus sp. UFO1	16S rRNA	rRNA	4,725,873	4,727,461	1,589	reverse	UFO1_RS22155	None
	tRNA-Lys	tRNA	4,822,498	4,822,573	76	forward	UFO1_RS22580	Partial
	tRNA-Glu	tRNA	4,822,581	4,822,655	75	forward	UFO1_RS22585	None
	tRNA-Val	tRNA	4,822,661	4,822,736	76	forward	UFO1_RS22590	None
	tRNA-Asp	tRNA	4,822,758	4,822,833	76	forward	UFO1_RS22595	None
	tRNA-Phe	tRNA	4,822,840	4,822,915	76	forward	UFO1_RS22600	None
	tRNA-Lys	tRNA	4,822,922	4,822,997	76	forward	UFO1_RS22605	None
	tRNA-Glu	tRNA	4,823,004	4,823,078	75	forward	UFO1_RS22610	None
	tRNA-Val	tRNA	4,823,085	4,823,160	76	forward	UFO1_RS22615	Partial
	hypothetical protein CDS	CDS	4,885,930	4,886,256	327	reverse	UFO1_RS22925	Partial
	16S rRNA	rRNA	5,100,726	5,102,314	1,589	forward	UFO1_RS23875	None
	tRNA-Ala	tRNA	5,102,487	5,102,562	76	forward	UFO1_RS23880	None
	23S rRNA	rRNA	5,102,798	5,105,730	2,933	forward	UFO1_RS23885	Partial
	5S rRNA	rRNA	5,105,830	5,105,946	117	forward	UFO1_RS23890	None
	16S rRNA	rRNA	5,106,273	5,107,861	1,589	forward	UFO1_RS23895	None
	tRNA-Ile	tRNA	5,107,937	5,108,013	77	forward	UFO1_RS23900	None
	tRNA-Ala	tRNA	5,108,077	5,108,152	76	forward	UFO1_RS23905	None
	23S rRNA	rRNA	5,108,427	5,111,359	2,933	forward	UFO1_RS23910	None
	5S rRNA	rRNA	5,111,459	5,111,575	117	forward	UFO1_RS23915	None
	16S rRNA	rRNA	5,111,900	5,113,487	1,588	forward	UFO1_RS23920	None
Pelosinus sp. JBW45	16S rRNA	rRNA	25,872	27,539	1,668	forward	JBW_RS00115	Partial
	23S rRNA	rRNA	27,953	30,885	2,933	forward	JBW_RS00120	None
	16S rRNA	rRNA	31,821	33,388	1,568	forward	JBW_RS00135	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Pelosinus sp. JBW45	23S rRNA	rRNA	33,862	36,794	2,933	forward	JBW_RS00140	None
	transposase CDS	CDS	500,260	501,789	1,530	forward	JBW_RS01940	None
	transposase CDS	CDS	978,834	979,529	696	reverse	JBW_RS04100	Partial
	transposase CDS	CDS	979,727	980,413	687	reverse	JBW_RS04105	None
	hypothetical protein CDS	CDS	1037234	1042591	>5358	forward	JBW_RS04310	Partial
	RNA-directed DNA polymerase CDS	CDS	1,164,275	1,165,519	1,245	forward	JBW_RS04725	None
	reverse transcriptase CDS	CDS	1,212,068	1212979	>912	forward	JBW_RS04940	None
	RNA-directed DNA polymerase CDS	CDS	1,213,528	1,214,772	1,245	forward	JBW_RS04945	None
	transposase CDS	CDS	1,674,322	1,675,017	696	reverse	JBW_RS07000	None
	transcriptional regulator CDS	CDS	1,675,215	1,675,901	687	reverse	JBW_RS07005	None
	transposase CDS	CDS	1,749,676	1,750,362	687	forward	JBW_RS07325	None
	transposase CDS	CDS	1,750,560	1,751,255	696	forward	JBW_RS07330	None
	transposase CDS	CDS	1,890,115	1,891,644	1,530	reverse	JBW_RS08040	None
	aldehyde dehydrogenase CDS	CDS	2,077,671	2,079,089	1,419	forward	JBW_RS08875	Partial
	transcriptional regulator CDS	CDS	2,149,515	2,150,201	687	forward	JBW_RS09195	None
	transposase CDS	CDS	2,150,399	2,151,094	696	forward	JBW_RS09200	None
	16S rRNA	rRNA	2,255,610	2,257,277	1,668	forward	JBW_RS09695	Partial
	5S rRNA	rRNA	2,260,814	2,260,930	117	forward	JBW_RS09705	None
	tRNA-Asn	tRNA	2,260,934	2,261,009	76	forward	JBW_RS09710	Partial
	transposase CDS	CDS	2,349,240	2,350,769	1,530	forward	JBW_RS10175	None
	transposase CDS	CDS	2,460,504	2,462,033	1,530	reverse	JBW_RS10730	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Pelosinus sp. JBW45	ABC transporter substrate-binding protein CDS	CDS	2,489,238	2,490,182	945	forward	JBW_RS10880	None
	transposase CDS	CDS	2,491,926	2,492,504	579	forward	JBW_RS10890	None
	hypothetical protein CDS	CDS	2,498,560	2498769	>210	forward	JBW_RS10915	Partial
	hypothetical protein CDS	CDS	2,498,896	2,499,354	459	reverse	JBW_RS10920	Partial
	transposase CDS	CDS	2,503,618	2,504,904	1,287	forward	JBW_RS10945	None
	transcriptional regulator CDS	CDS	2,725,492	2,726,178	687	forward	JBW_RS12110	Partial
	transposase CDS	CDS	2,726,376	2,727,071	696	forward	JBW_RS12115	Partial
	tRNA-Asn	tRNA	2,772,268	2,772,343	76	reverse	JBW_RS12345	Partial
	5S rRNA	rRNA	2,772,347	2,772,463	117	reverse	JBW_RS12350	None
	23S rRNA	rRNA	2,772,658	2,775,591	2,934	reverse	JBW_RS12355	None
	tRNA-Ala	tRNA	2,775,868	2,775,943	76	reverse	JBW_RS12360	None
	tRNA-Ile	tRNA	2,776,030	2,776,106	77	reverse	JBW_RS12365	None
	16S rRNA	rRNA	2,776,200	2,777,859	1,660	reverse	JBW_RS12370	Partial
	citrate lyase subunit alpha CDS	CDS	2,926,312	2,927,853	1,542	reverse	JBW_RS12965	Partial
	citrate lyase subunit beta CDS	CDS	2,927,856	2,928,737	882	reverse	JBW_RS12970	None
	hypothetical protein CDS	CDS	3,192,565	3,193,893	1,329	forward	JBW_RS14130	Partial
	23S rRNA	rRNA	3,211,881	3,214,813	2,933	reverse	JBW_RS14240	None
	16S rRNA	rRNA	3,215,227	3,216,894	1,668	reverse	JBW_RS14245	None
	RNA-directed DNA polymerase CDS	CDS	3,406,693	3,407,937	1,245	forward	JBW_RS15180	None
	reverse transcriptase CDS	CDS	3408014	3,408,871	>858	forward	JBW_RS15185	None
	hypothetical protein CDS	CDS	3,489,540	3,489,998	459	reverse	JBW_RS15510	Partial
	hypothetical protein CDS	CDS	3,490,896	3,491,447	552	reverse	JBW_RS15520	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Pelosinus sp. JBW45	Intergenic	NA	3,777,045	3,777,309	264	NA	NA	None
	RNA-directed DNA polymerase CDS	CDS	3,804,681	3,805,925	1,245	reverse	JBW_RS16955	None
	transposase CDS	CDS	3,807,729	3808016	>288	reverse	JBW_RS16965	Partial
	transposase CDS	CDS	3,808,064	3,808,924	861	reverse	JBW_RS16970	Partial
	filamentous hemagglutinin CDS	CDS	3811898	3820369	>8472	reverse	JBW_RS16990	Partial
	transposase CDS	CDS	3,955,836	3,956,597	762	reverse	JBW_RS17540	Partial
	transposase CDS	CDS	4,204,267	4,204,950	684	reverse	JBW_RS18595	Partial
	ribonuclease J CDS	CDS	4,263,412	4,265,085	1,674	reverse	JBW_RS18855	Partial
	23S rRNA	rRNA	4,372,018	4,374,951	2,934	reverse	JBW_RS19475	None
	tRNA-Ala	tRNA	4,375,334	4,375,409	76	reverse	JBW_RS19480	None
	tRNA-Ile	tRNA	4,375,553	4,375,629	77	reverse	JBW_RS19485	None
	16S rRNA	rRNA	4,375,723	4,377,290	1,568	reverse	JBW_RS19490	None
	5S rRNA	rRNA	4,377,711	4,377,827	117	reverse	JBW_RS19495	None
	23S rRNA	rRNA	4,377,923	4,380,856	2,934	reverse	JBW_RS19500	None
	tRNA-Ala	tRNA	4,381,239	4,381,314	76	reverse	JBW_RS19505	None
	tRNA-Ile	tRNA	4,381,458	4,381,534	77	reverse	JBW_RS19510	None
	16S rRNA	rRNA	4,381,628	4,383,195	1,568	reverse	JBW_RS19515	None
	transposase CDS	CDS	4,472,346	4,473,875	1,530	forward	JBW_RS19925	None
	transposase CDS	CDS	4,601,886	4,602,581	696	reverse	JBW_RS20430	None
	transposase CDS	CDS	4,602,779	4,603,465	687	reverse	JBW_RS20435	None
	hypothetical protein CDS	CDS	4,628,169	4,628,627	459	reverse	JBW_RS20535	Partial
	ankyrin CDS	CDS	4,628,726	4,629,670	945	reverse	JBW_RS20540	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
Pelosinus sp. JBW45	hypothetical protein CDS	CDS	4,632,518	4,633,069	552	reverse	JBW_RS20565	Partial
	23S rRNA	rRNA	5,034,508	5,037,440	2,933	reverse	JBW_RS22250	None
	16S rRNA	rRNA	5,037,849	5,039,416	1,568	reverse	JBW_RS22255	Partial
	5S rRNA	rRNA	5,039,837	5,039,953	117	reverse	JBW_RS22260	None
C. paradoxum JW/YL-7T	Type II secretion system F domain-containing protein CDS	CDS	2	316	315	reverse	PD_0001	None
	hypothetical protein CDS	CDS	409	561	153	reverse	PD_0002	None
	twitching motility protein CDS	CDS	586	1,635	1,050	reverse	PD_0003	None
	type II secretion system protein E CDS	CDS	1,646	3,325	1,680	reverse	PD_0004	None
	cytosol aminopeptidase CDS	CDS	3,454	4,959	1,506	forward	PD_0005	None
	Protein of unknown function DUF2508 CDS	CDS	5,076	5,312	237	forward	PD_0006	None
	sigmaK-factor processing regulatory BofA CDS	CDS	5,324	5,605	282	forward	PD_0007	None
	hypothetical protein CDS	CDS	5,778	6,191	414	forward	PD_0008	None
	50S ribosomal protein L25 CDS	CDS	6,273	6,833	561	forward	PD_0009	None
	16s rRNA	rRNA	7,080	8,583	1,504	forward	NA	None
	tRNA-Ala1	tRNA	8,640	8,715	76	forward	PD_RNA_1	None
	23s rRNA	rRNA	8,773	11,688	2,916	forward	NA	None
	5s rRNA	rRNA	11,749	11,864	116	forward	NA	None
	16s rRNA	rRNA	12,607	14,110	1,504	forward	NA	None
	tRNA-Ala2	tRNA	14,167	14,242	76	forward	PD_RNA_2	None
	23s rRNA	rRNA	14,300	17,215	2,916	forward	NA	None
	5s rRNA	rRNA	17,276	17,391	116	forward	NA	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. paradoxum JW/YL-7T	transposase, IS605 OrfB family CDS	CDS	39,076	39,996	921	forward	PD_0034	None
	tRNA-Gly1	tRNA	81,910	81,983	74	forward	PD_RNA_4	None
	16s rRNA	rRNA	82,124	83,728	1,605	forward	NA	None
	23s rRNA	rRNA	83,872	86,787	2,916	forward	NA	None
	5s rRNA	rRNA	86,879	86,994	116	forward	NA	None
	hypothetical protein CDS	CDS	86,993	87,346	354	forward	PD_0093	None
	tRNA-Asn1	tRNA	86,999	87,073	75	forward	PD_RNA_5	None
	tRNA-Leu1	tRNA	87,081	87,169	89	forward	PD_RNA_6	None
	tRNA-Met1	tRNA	87,176	87,251	76	forward	PD_RNA_7	None
	tRNA-Glu1	tRNA	87,254	87,328	75	forward	PD_RNA_8	None
	tRNA-Val1	tRNA	87,337	87,412	76	forward	PD_RNA_9	None
	tRNA-Asp1	tRNA	87,420	87,496	77	forward	PD_RNA_10	None
	tRNA-Thr2	tRNA	87,508	87,583	76	forward	PD_RNA_11	None
	tRNA-Tyr1	tRNA	87,589	87,673	85	forward	PD_RNA_12	None
	tRNA-Met2	tRNA	87,684	87,760	77	forward	PD_RNA_13	None
	tRNA-Trp1	tRNA	87,765	87,840	76	forward	PD_RNA_14	None
	tRNA-Pro1	tRNA	87,868	87,944	77	forward	PD_RNA_15	None
	tRNA-Ile1	tRNA	87,953	88,029	77	forward	PD_RNA_16	None
	tRNA-Gly2	tRNA	88,044	88,117	74	forward	PD_RNA_17	None
	tRNA-Arg1	tRNA	88,122	88,198	77	forward	PD_RNA_18	None
	tRNA-Gln1	tRNA	88,205	88,280	76	forward	PD_RNA_19	None
	tRNA-Lys1	tRNA	88,290	88,365	76	forward	PD_RNA_20	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. paradoxum JW/YL-7T	tRNA-Ser1	tRNA	88,371	88,459	89	forward	PD_RNA_21	None
	tRNA-Ser2	tRNA	88,464	88,554	91	forward	PD_RNA_22	None
	tRNA-Pro2	tRNA	88,561	88,637	77	forward	PD_RNA_23	None
	tRNA-Ile2	tRNA	88,646	88,722	77	forward	PD_RNA_24	None
	tRNA-Met3	tRNA	88,733	88,809	77	forward	PD_RNA_25	None
	tRNA-Phe1	tRNA	88,813	88,888	76	forward	PD_RNA_26	None
	tRNA-Met4	tRNA	88,894	88,970	77	forward	PD_RNA_27	None
	16s rRNA	rRNA	89,042	90,645	1,604	forward	NA	None
	tRNA-Ala3	tRNA	90,702	90,777	76	forward	PD_RNA_28	None
	23s rRNA	rRNA	90,835	93,750	2,916	forward	NA	None
	5s rRNA	rRNA	93,842	93,957	116	forward	NA	None
	hypothetical protein CDS	CDS	93,956	94,309	354	forward	PD_0094	None
	tRNA-Asn2	tRNA	93,962	94,036	75	forward	PD_RNA_29	None
	tRNA-Leu2	tRNA	94,044	94,132	89	forward	PD_RNA_30	None
	tRNA-Met5	tRNA	94,139	94,214	76	forward	PD_RNA_31	None
	tRNA-Glu2	tRNA	94,217	94,291	75	forward	PD_RNA_32	None
	tRNA-Val2	tRNA	94,300	94,375	76	forward	PD_RNA_33	None
	tRNA-Asp2	tRNA	94,383	94,459	77	forward	PD_RNA_34	None
	tRNA-Thr3	tRNA	94,471	94,546	76	forward	PD_RNA_35	None
	tRNA-Leu3	tRNA	94,553	94,635	83	forward	PD_RNA_36	None
	tRNA-Gly3	tRNA	94,729	94,803	75	forward	PD_RNA_37	None
	tRNA-Arg2	tRNA	94,808	94,884	77	forward	PD_RNA_38	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. paradoxum JW/YL-7T	transposase, IS605 OrfB family CDS	CDS	104,007	104,927	921	forward	PD_0105	None
	transposase, IS605 OrfB family CDS	CDS	164,930	165,850	921	forward	PD_0166	None
	tRNA-Gly4	tRNA	410,356	410,429	74	forward	PD_RNA_39	None
	16s rRNA	rRNA	410,570	412,173	1,604	forward	NA	None
	tRNA-Ala4	tRNA	412,230	412,305	76	forward	PD_RNA_40	None
	23s rRNA	rRNA	412,363	415,278	2,916	forward	NA	None
	5s rRNA	rRNA	415,339	415,454	116	forward	NA	None
	transposase, IS605 OrfB family CDS	CDS	600,340	601,260	921	forward	PD_0587	None
	transposase, IS605 OrfB family CDS	CDS	625,494	625,880	387	forward	PD_0614	None
	transposase, IS605 OrfB family CDS	CDS	679,962	680,759	798	forward	PD_0672	None
	transposase, IS605 OrfB family CDS	CDS	718,593	719,390	798	forward	PD_0716	None
	transposase, IS605 OrfB family CDS	CDS	1,102,009	1,102,395	387	reverse	PD_1143	None
	Transposase, helix-turn-helix domain-containing protein CDS	CDS	1,102,701	1,102,928	228	reverse	PD_1144	None
	transposase, IS605 OrfB family CDS	CDS	1,170,443	1,171,240	798	reverse	PD_1218	None
	hypothetical protein CDS	CDS	1,175,234	1,175,839	606	forward	PD_1222	None
	hypothetical protein CDS	CDS	1,175,872	1,176,612	741	forward	PD_1223	None
	transposase, IS605 OrfB family CDS	CDS	1,454,296	1,455,480	1,185	reverse	PD_1500	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. paradoxum JW/YL-7T	transposase, IS605 OrfB family CDS	CDS	1,475,979	1,477,163	1,185	reverse	PD_1525	None
	5s rRNA	rRNA	1,525,627	1,525,742	116	reverse	NA	None
	23s rRNA	rRNA	1,525,780	1,528,694	2,915	reverse	NA	None
	16s rRNA	rRNA	1,528,838	1,530,450	1,613	reverse	NA	None
	transposase, IS605 OrfB family CDS	CDS	1,657,029	1,657,949	921	reverse	PD_1693	None
	5s rRNA	rRNA	1,690,223	1,690,338	116	reverse	NA	None
	23s rRNA	rRNA	1,690,376	1,693,291	2,916	reverse	NA	None
	16s rRNA	rRNA	1,693,435	1,695,036	1,602	reverse	NA	None
	hypothetical protein CDS	CDS	1,733,752	1,735,260	1,509	reverse	PD_1766	None
	16s rRNA	rRNA	1,811,390	1,812,893	1,504	forward	NA	None
	tRNA-Ala5	tRNA	1,812,950	1,813,025	76	forward	PD_RNA_44	None
	23s rRNA	rRNA	1,813,083	1,815,998	2,916	forward	NA	None
	5s rRNA	rRNA	1,816,056	1,816,171	116	forward	NA	None
	16s rRNA	rRNA	1,838,274	1,839,876	1,603	forward	NA	None
	23s rRNA	rRNA	1,840,020	1,842,935	2,916	forward	NA	None
	5s rRNA	rRNA	1,843,029	1,843,144	116	forward	NA	None
	hypothetical protein CDS	CDS	1,843,143	1,843,496	354	forward	PD_1869	None
	tRNA-Asn3	tRNA	1,843,149	1,843,223	75	forward	PD_RNA_46	None
	tRNA-Leu5	tRNA	1,843,231	1,843,319	89	forward	PD_RNA_47	None
	tRNA-Met6	tRNA	1,843,326	1,843,401	76	forward	PD_RNA_48	None
	tRNA-Glu3	tRNA	1,843,404	1,843,478	75	forward	PD_RNA_49	None
	tRNA-Val3	tRNA	1,843,487	1,843,562	76	forward	PD_RNA_50	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. paradoxum JW/YL-7T	tRNA-Asp3	tRNA	1,843,570	1,843,646	77	forward	PD_RNA_51	None
	tRNA-Thr4	tRNA	1,843,658	1,843,733	76	forward	PD_RNA_52	None
	tRNA-Tyr2	tRNA	1,843,739	1,843,823	85	forward	PD_RNA_53	None
	tRNA-Leu6	tRNA	1,843,838	1,843,920	83	forward	PD_RNA_54	None
	tRNA-Gly6	tRNA	1,843,926	1,844,000	75	forward	PD_RNA_55	None
	tRNA-Gly7	tRNA	1,844,016	1,844,089	74	forward	PD_RNA_56	None
	tRNA-Arg3	tRNA	1,844,094	1,844,170	77	forward	PD_RNA_57	None
	tRNA-Gln2	tRNA	1,844,177	1,844,252	76	forward	PD_RNA_58	None
	tRNA-Lys2	tRNA	1,844,262	1,844,337	76	forward	PD_RNA_59	None
	tRNA-Ser5	tRNA	1,844,343	1,844,431	89	forward	PD_RNA_60	None
	tRNA-Phe2	tRNA	1,844,440	1,844,515	76	forward	PD_RNA_61	None
	tRNA-Met7	tRNA	1,844,521	1,844,597	77	forward	PD_RNA_62	None
	tRNA-Met8	tRNA	1,844,603	1,844,679	77	forward	PD_RNA_63	None
	tRNA-Pro3	tRNA	1,844,689	1,844,765	77	forward	PD_RNA_64	None
	tRNA-His1	tRNA	1,844,777	1,844,853	77	forward	PD_RNA_65	None
	tRNA-Lys3	tRNA	1,844,858	1,844,933	76	forward	PD_RNA_66	None
	tRNA-Cys1	tRNA	1,844,943	1,845,016	74	forward	PD_RNA_67	None
	tRNA-Val4	tRNA	1,845,021	1,845,096	76	forward	PD_RNA_68	None
	tRNA-Asn4	tRNA	1,845,163	1,845,237	75	forward	PD_RNA_69	None
	tRNA-Glu4	tRNA	1,845,242	1,845,316	75	forward	PD_RNA_70	None
	tRNA-Val5	tRNA	1,845,325	1,845,400	76	forward	PD_RNA_71	None
	tRNA-Asp4	tRNA	1,845,408	1,845,484	77	forward	PD_RNA_72	None
	tRNA-Thr5	tRNA	1,845,496	1,845,571	76	forward	PD_RNA_73	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. paradoxum JW/YL-7T	tRNA-Tyr3	tRNA	1,845,577	1,845,661	85	forward	PD_RNA_74	None
	tRNA-Leu7	tRNA	1,845,675	1,845,757	83	forward	PD_RNA_75	None
	tRNA-Gly8	tRNA	1,845,763	1,845,837	75	forward	PD_RNA_76	None
	tRNA-Gly9	tRNA	1,845,853	1,845,926	74	forward	PD_RNA_77	None
	tRNA-Arg4	tRNA	1,845,931	1,846,007	77	forward	PD_RNA_78	None
	tRNA-Gln3	tRNA	1,846,014	1,846,089	76	forward	PD_RNA_79	None
	tRNA-Lys4	tRNA	1,846,099	1,846,174	76	forward	PD_RNA_80	None
	tRNA-Ser6	tRNA	1,846,180	1,846,268	89	forward	PD_RNA_81	None
	tRNA-Phe3	tRNA	1,846,277	1,846,352	76	forward	PD_RNA_82	None
	tRNA-Met9	tRNA	1,846,358	1,846,434	77	forward	PD_RNA_83	None
	tRNA-Met10	tRNA	1,846,440	1,846,516	77	forward	PD_RNA_84	None
	tRNA-Pro4	tRNA	1,846,526	1,846,602	77	forward	PD_RNA_85	None
	tRNA-His2	tRNA	1,846,614	1,846,690	77	forward	PD_RNA_86	None
	tRNA-Cys2	tRNA	1,846,742	1,846,815	74	forward	PD_RNA_87	None
	tRNA-Arg5	tRNA	1,846,833	1,846,909	77	forward	PD_RNA_88	None
	hypothetical protein CDS	CDS	251	454	204	reverse	PD_1877	None
	diguanylate cyclase CDS	CDS	514	1,173	660	reverse	PD_1878	None
	16s rRNA	rRNA	1,649	3,250	1,602	forward	NA	None
	23s rRNA	rRNA	3,394	6,309	2,916	forward	NA	None
	5s rRNA	rRNA	6,347	6,462	116	forward	NA	None
	transposase, IS605 OrfB family CDS	CDS	35,726	36,544	819	forward	PD_1909	None
	16s rRNA	rRNA	56,395	57,998	1,604	forward	NA	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
C. paradoxum JW/YL-7T	tRNA-Ala6	tRNA	58,055	58,130	76	forward	PD_RNA_1	None
	23s rRNA	rRNA	58,188	61,103	2,916	forward	NA	None
	5s rRNA	rRNA	61,141	61,256	116	forward	NA	None
	Phosphoglycerate mutase CDS	CDS	396	680	285	reverse	PD_1932	None
	hypothetical protein CDS	CDS	888	1,019	132	reverse	PD_1933	None
	cobalamin-5-phosphate synthase CobS CDS	CDS	1,373	1,525	153	reverse	PD_1934	None
	cobalamin-5-phosphate synthase CobS CDS	CDS	1,580	1,777	198	reverse	PD_1935	None
	cobalbumin biosynthesis protein CDS	CDS	1,961	2,341	381	reverse	PD_1936	None
	5s rRNA	rRNA	2,434	2,549	116	reverse	NA	None
	23s rRNA	rRNA	2,587	5,501	2,915	reverse	NA	None
	tRNA-Ala7	tRNA	5,559	5,634	76	reverse	PD_RNA_2	None
	16s rRNA	rRNA	5,691	7,294	1,604	reverse	NA	None
	tRNA-Gly10	tRNA	7,435	7,508	74	reverse	PD_RNA_1	None
	etfA; electron transfer flavoprotein subunit alpha CDS	CDS	7,662	7,808	147	reverse	PD_1937	None
	Electron transfer flavoprotein alpha/beta-subunit CDS	CDS	8,050	8,439	390	reverse	PD_1938	None
B. cellulosolvens ATCC 35603	hypothetical protein CDS	CDS	246,037	247,821	1,785	forward	Bccel_0248	None
	Bacteriophage portal protein, SPP1 Gp6-like protein CDS	CDS	247,836	249,182	1,347	forward	Bccel_0249	None
	phage head morphogenesis protein, SPP1 gp7 family CDS	CDS	249,182	250,207	1,026	forward	Bccel_0250	None
	cellulosome anchoring protein cohesin region CDS	CDS	335,896	338,145	2,250	forward	Bccel_0335	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
B. cellulosolvens ATCC 35603	16s rRNA	rRNA	592,524	594,023	1,500	forward	NA	Partial
	23s rRNA	rRNA	594,647	597,693	3,047	forward	NA	Partial
	hypothetical protein CDS	CDS	704,457	706,079	1,623	reverse	Bccel_0635	Partial
	transposase IS66 CDS	CDS	710,753	712,384	1,632	reverse	Bccel_0641	Partial
	IS66 Orf2 family protein CDS	CDS	712,422	712,775	354	reverse	Bccel_0642	Partial
	hypothetical protein CDS	CDS	712,772	713,125	354	reverse	Bccel_0643	None
	Transposase DDE domain CDS	CDS	1,063,263	1,064,165	903	reverse	Bccel_0929	None
	hypothetical protein CDS	CDS	1,079,010	1,079,192	183	reverse	Bccel_0956	None
	hypothetical protein CDS	CDS	1,079,248	1,079,514	267	reverse	Bccel_0957	None
	hypothetical protein CDS	CDS	1,079,638	1,080,819	1,182	reverse	Bccel_0958	None
	Zonular occludens toxin CDS	CDS	1,080,941	1,081,696	756	reverse	Bccel_0959	None
	hypothetical protein CDS	CDS	1,081,710	1,081,970	261	reverse	Bccel_0960	None
	hypothetical protein CDS	CDS	1,081,963	1,083,273	1,311	reverse	Bccel_0961	None
	hypothetical protein CDS	CDS	1,083,335	1,083,583	249	reverse	Bccel_0962	None
	transposase IS111A/IS1328/IS1533 CDS	CDS	1,084,005	1,085,285	1,281	reverse	Bccel_0963	None
	hypothetical protein CDS	CDS	1,085,428	1,085,535	108	reverse	Bccel_0964	None
	hypothetical protein CDS	CDS	1,085,547	1,085,702	156	reverse	Bccel_0965	None
	hypothetical protein CDS	CDS	1,085,707	1,086,012	306	reverse	Bccel_0966	None
	hypothetical protein CDS	CDS	1,086,015	1,086,200	186	reverse	Bccel_0967	None
	hypothetical protein CDS	CDS	1,086,393	1,086,569	177	forward	Bccel_0968	None
	hypothetical protein CDS	CDS	1,086,580	1,086,915	336	forward	Bccel_0969	None
	hypothetical protein CDS	CDS	1,087,432	1,087,644	213	forward	Bccel_0970	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
B. cellulosolvens ATCC 35603	Intergenic	NA	1,302,342	1,302,729	387	NA	NA	None
	hypothetical protein CDS	CDS	1,397,628	1,399,250	1,623	reverse	Bccel_1252	Partial
	5s rRNA	rRNA	1,405,751	1,405,866	116	reverse	NA	None
	23s rRNA	rRNA	1,406,025	1,409,071	3,047	reverse	NA	None
	16s rRNA	rRNA	1,409,798	1,411,298	1,501	reverse	NA	None
	hypothetical protein CDS	CDS	1,443,604	1,445,226	1,623	forward	Bccel_1282	Partial
	hypothetical protein CDS	CDS	1,554,227	1,555,849	1,623	forward	Bccel_1369	Partial
	transposase IS66 CDS	CDS	1,566,670	1,567,308	639	reverse	Bccel_1380	Partial
	hypothetical protein CDS	CDS	1,569,058	1,570,455	1,398	reverse	Bccel_1383	Partial
	tRNA-Met2	tRNA	1,658,230	1,658,303	74	forward	NA	Partial
	tRNA-Phe1	tRNA	1,658,355	1,658,430	76	forward	NA	None
	tRNA-Tyr1	tRNA	1,658,434	1,658,518	85	forward	NA	Partial
	Fibronectin type III domain protein CDS	CDS	1,693,002	1,699,823	6,822	reverse	Bccel_1468	Partial
	Phage-like element PBSX protein, XkdS CDS	CDS	1,845,072	1,845,485	414	reverse	Bccel_1579	Partial
	Protein of unknown function, DUF2577 CDS	CDS	1,845,488	1,845,778	291	reverse	Bccel_1580	None
	hypothetical protein CDS	CDS	1,845,794	1,846,765	972	reverse	Bccel_1581	None
	Peptidoglycan-binding lysin domain-containing protein CDS	CDS	1,846,758	1,847,471	714	reverse	Bccel_1582	None
	phage tape measure protein CDS	CDS	1,847,443	1,849,470	2,028	reverse	Bccel_1583	None
	XkdN-like protein CDS	CDS	1,849,689	1,850,099	411	reverse	Bccel_1584	None
	XkdM protein, phage-like element PBSX CDS	CDS	1,850,179	1,850,634	456	reverse	Bccel_1585	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
B. cellulosolvens ATCC 35603	hypothetical protein CDS	CDS	1,850,650	1,851,978	1,329	reverse	Bccel_1586	None
	hypothetical protein CDS	CDS	1,851,982	1,852,158	177	reverse	Bccel_1587	None
	hypothetical protein CDS	CDS	1,852,164	1,852,598	435	reverse	Bccel_1588	Partial
	hypothetical protein CDS	CDS	1,856,017	1,856,955	939	reverse	Bccel_1596	Partial
	Protein of unknown function DUF4355 CDS	CDS	1,856,988	1,857,428	441	reverse	Bccel_1597	None
	hypothetical protein CDS	CDS	1,857,401	1,857,532	132	reverse	Bccel_1598	None
	phage head morphogenesis protein, SPP1 gp7 family CDS	CDS	1,857,652	1,858,683	1,032	reverse	Bccel_1599	Partial
	23s rRNA	rRNA	2,021,025	2,024,073	3,049	reverse	NA	None
	16s rRNA	rRNA	2,024,672	2,026,172	1,501	reverse	NA	None
	PKD domain containing protein CDS	CDS	2,338,945	2,373,633	34,689	reverse	Bccel_1985	Partial
	protein of unknown function DUF11 CDS	CDS	2,455,355	2,462,293	6,939	reverse	Bccel_2050	Partial
	Transposase DDE domain CDS	CDS	2,580,590	2,581,492	903	forward	Bccel_2154	Partial
	Transposase DDE domain CDS	CDS	2,628,132	2,629,034	903	forward	Bccel_2215	Partial
	hypothetical protein CDS	CDS	2,704,380	2,706,002	1,623	forward	Bccel_2279	Partial
	tRNA-Asn3	tRNA	2,857,876	2,857,948	73	reverse	RNA_64	None
	5s rRNA	rRNA	2,857,954	2,858,069	116	reverse	NA	None
	23s rRNA	rRNA	2,858,318	2,861,362	3,045	reverse	NA	None
	tRNA-Ala5	tRNA	2,861,787	2,861,862	76	reverse	RNA_63	None
	16s rRNA	rRNA	2,861,984	2,863,483	1,500	reverse	NA	None
	protein of unknown function DUF4347 CDS	CDS	2,968,552	2,976,708	8,157	reverse	Bccel_2494	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
B. cellulosolvens ATCC 35603	23s rRNA	rRNA	3,008,041	3,011,088	3,048	reverse	NA	Partial
	tRNA-Ala4	tRNA	3,011,512	3,011,587	76	reverse	RNA_58	None
	16s rRNA	rRNA	3,011,710	3,013,209	1,500	reverse	NA	None
	tRNA-Gly5	tRNA	3,085,783	3,085,854	72	reverse	RNA_54	Partial
	tRNA-Phe2	tRNA	3,085,858	3,085,933	76	reverse	RNA_53	None
	tRNA-Asp2	tRNA	3,085,941	3,086,017	77	reverse	RNA_52	Partial
	transposase IS66 CDS	CDS	3,337,299	3,338,930	1,632	reverse	Bccel_2830	Partial
	IS66 Orf2 family protein CDS	CDS	3,338,968	3,339,321	354	reverse	Bccel_2831	None
	Intergenic	NA	3,371,572	3,371,832	260	NA	NA	None
	transposase IS66 CDS	CDS	3,379,041	3,380,477	1,437	forward	Bccel_2864	Partial
	transposase CDS	CDS	3,750,055	3,751,233	1,179	reverse	Bccel_3181	None
	integrase family protein CDS	CDS	3,751,220	3,751,705	486	reverse	Bccel_3182	None
	hypothetical protein CDS	CDS	3,751,842	3,752,135	294	forward	Bccel_3183	None
	Integrase catalytic region CDS	CDS	3,752,183	3,752,836	654	forward	Bccel_3184	None
	RNA-directed DNA polymerase (Reverse transcriptase) CDS	CDS	3,779,006	3,780,823	1,818	forward	Bccel_3206	None
	hypothetical protein CDS	CDS	3,781,129	3,781,908	780	forward	Bccel_3207	Partial
	23s rRNA	rRNA	4,033,857	4,036,904	3,048	reverse	NA	None
	16s rRNA	rRNA	4,037,580	4,039,080	1,501	reverse	NA	Partial
	hypothetical protein CDS	CDS	4,052,439	4,052,831	393	forward	Bccel_3444	Partial
	hypothetical protein CDS	CDS	4,052,948	4,053,358	411	forward	Bccel_3445	None
	transposase IS66 CDS	CDS	4,053,375	4,053,782	408	forward	Bccel_3446	Partial
	cellulosome anchoring protein cohesin region CDS	CDS	4,063,363	4,065,333	1,971	forward	Bccel_3452	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
B. cellulosolvens ATCC 35603	IS66 Orf2 family protein CDS	CDS	4,077,014	4,077,256	243	forward	Bccel_3465	None
	integrase family protein CDS	CDS	4,077,375	4,078,628	1,254	forward	Bccel_3466	None
	integrase family protein CDS	CDS	4,078,625	4,079,608	984	forward	Bccel_3467	None
	integrase family protein CDS	CDS	4,079,605	4,080,633	1,029	forward	Bccel_3468	None
	transposase IS66 CDS	CDS	4,080,928	4,082,559	1,632	forward	Bccel_3469	Partial
	hypothetical protein CDS	CDS	4,820,100	4,821,497	1,398	reverse	Bccel_4117	None
	hypothetical protein CDS	CDS	4,824,443	4,826,065	1,623	forward	Bccel_4120	Partial
	Transposase DDE domain CDS	CDS	5,211,748	5,212,650	903	forward	Bccel_4474	Partial
	hypothetical protein CDS	CDS	5,264,212	5,265,357	1,146	forward	Bccel_4506	Partial
	hypothetical protein CDS	CDS	5,295,124	5,296,746	1,623	reverse	Bccel_4533	None
	23s rRNA	rRNA	5,829,929	5,832,976	3,048	reverse	NA	None
	16s rRNA	rRNA	5,833,528	5,835,028	1,501	reverse	NA	Partial
	Intergenic	NA	5,878,924	5,879,070	146	NA	NA	None
	hypothetical protein CDS	CDS	5,972,371	5,973,768	1,398	reverse	Bccel_5121	Partial
	hypothetical protein CDS	CDS	5,987,673	5,988,575	903	forward	Bccel_5135	Partial
	hypothetical protein CDS	CDS	6,024,906	6,026,240	1,335	reverse	Bccel_5163	Partial
	Transposase DDE domain CDS	CDS	6,026,357	6,027,259	903	reverse	Bccel_5164	None
	hypothetical protein CDS	CDS	6,027,375	6,027,662	288	reverse	Bccel_5165	Partial
	hypothetical protein CDS	CDS	6,164,401	6,166,023	1,623	forward	Bccel_5274	Partial
	hypothetical protein CDS	CDS	6,199,755	6,200,657	903	forward	Bccel_5306	Partial
	transposase IS66 CDS	CDS	6,256,509	6,257,945	1,437	reverse	Bccel_5355	Partial
	RNA-directed DNA polymerase (Reverse transcriptase) CDS	CDS	6,259,130	6,260,947	1,818	reverse	Bccel_5357	Partial

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
B. cellulosolvens ATCC 35603	RNA-directed DNA polymerase (Reverse transcriptase) CDS	CDS	6,261,626	6,263,443	1,818	reverse	Bccel_5358	Partial
	hypothetical protein CDS	CDS	6,267,826	6,268,728	903	forward	Bccel_5365	Partial
	hypothetical protein CDS	CDS	6,325,473	6,327,095	1,623	reverse	Bccel_5403	Partial
	transposase IS66 CDS	CDS	6,456,703	6,458,334	1,632	forward	Bccel_5525	Partial
	RNA-directed DNA polymerase (Reverse transcriptase) CDS	CDS	6,459,530	6,461,347	1,818	reverse	Bccel_5527	Partial
	16s rRNA	rRNA	6,495,494	6,496,994	1,501	forward	NA	Partial
	23s rRNA	rRNA	6,497,625	6,500,672	3,048	forward	NA	None
	5s rRNA	rRNA	6,501,018	6,501,133	116	forward	NA	None
	hypothetical protein CDS	CDS	6,751,076	6,751,666	591	reverse	Bccel_5760	None
	hypothetical protein CDS	CDS	6,751,746	6,752,984	1,239	reverse	Bccel_5761	Partial
	Protein of unknown function DUF2493 CDS	CDS	6,753,231	6,753,575	345	reverse	Bccel_5762	None
	Protein of unknown function DUF2493 CDS	CDS	6,755,324	6,755,668	345	forward	Bccel_5765	None
	hypothetical protein CDS	CDS	6,755,915	6,757,153	1,239	forward	Bccel_5766	None
	signal peptide	sig. peptide	6,757,233	6,757,313	81	forward	Bccel_5767	None
	hypothetical protein CDS	CDS	6,757,927	6,758,880	954	forward	Bccel_5768	None
	hypothetical protein CDS	CDS	6,758,918	6,759,613	696	forward	Bccel_5769	None
	hypothetical protein CDS	CDS	6,759,720	6,760,100	381	forward	Bccel_5770	None
	hypothetical protein CDS	CDS	6,760,223	6,760,549	327	reverse	Bccel_5771	None
	hypothetical protein CDS	CDS	6,760,605	6,760,727	123	reverse	Bccel_5772	None
	hypothetical protein CDS	CDS	6,761,414	6,762,049	636	forward	Bccel_5773	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
B. cellulosolvens ATCC 35603	hypothetical protein CDS	CDS	6,771,401	6,771,772	372	reverse	Bccel_5782	Partial
	hypothetical protein CDS	CDS	6,771,888	6,772,424	537	reverse	Bccel_5783	None
	hypothetical protein CDS	CDS	6,772,639	6,772,812	174	forward	Bccel_5784	None
	transposase IS116/IS110/IS902 family protein CDS	CDS	6,772,845	6,773,411	567	reverse	Bccel_5785	None
	transposase IS111A/IS1328/IS1533 CDS	CDS	6,773,411	6,774,082	672	reverse	Bccel_5786	None
	Protein of unknown function DUF2493 CDS	CDS	6,774,330	6,774,674	345	reverse	Bccel_5787	None
	hypothetical protein CDS	CDS	6,774,863	6,775,033	171	reverse	Bccel_5788	Partial
	hypothetical protein CDS	CDS	6,776,070	6,776,240	171	forward	Bccel_5789	None
	Protein of unknown function DUF2493 CDS	CDS	6,776,429	6,776,773	345	forward	Bccel_5790	None
	transposase IS111A/IS1328/IS1533 CDS	CDS	6,777,021	6,778,259	1,239	forward	Bccel_5791	Partial
	hypothetical protein CDS	CDS	6,779,730	6,780,554	825	forward	Bccel_5795	None
	hypothetical protein CDS	CDS	6,780,532	6,781,149	618	forward	Bccel_5796	None
	hypothetical protein CDS	CDS	6,787,573	6,788,163	591	reverse	Bccel_5803	None
	transposase IS111A/IS1328/IS1533 CDS	CDS	6,788,242	6,789,480	1,239	reverse	Bccel_5804	Partial
	hypothetical protein CDS	CDS	6,789,726	6,790,070	345	reverse	Bccel_5805	None
	protein of unknown function DUF1813 HSP20-like protein CDS	CDS	6,790,259	6,790,429	171	reverse	Bccel_5806	None
	protein of unknown function DUF1813 HSP20-like protein CDS	CDS	6,791,465	6,791,635	171	forward	Bccel_5807	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
B. cellulosolvens ATCC 35603	hypothetical protein CDS	CDS	6,791,824	6,792,168	345	forward	Bccel_5808	None
	transposase IS111A/IS1328/IS1533 CDS	CDS	6,792,412	6,793,650	1,239	forward	Bccel_5809	Partial
	hypothetical protein CDS	CDS	6,806,676	6,807,068	393	reverse	Bccel_5821	None
	hypothetical protein CDS	CDS	6,807,141	6,808,379	1,239	reverse	Bccel_5822	None
	hypothetical protein CDS	CDS	6,808,624	6,808,968	345	reverse	Bccel_5823	None
	protein of unknown function DUF1813 HSP20-like protein CDS	CDS	6,809,147	6,809,317	171	reverse	Bccel_5824	Partial
	protein of unknown function DUF1813 HSP20-like protein CDS	CDS	6,810,353	6,810,523	171	forward	Bccel_5825	None
	hypothetical protein CDS	CDS	6,810,702	6,811,046	345	forward	Bccel_5826	None
	hypothetical protein CDS	CDS	6,811,291	6,812,529	1,239	forward	Bccel_5827	Partial
	hypothetical protein CDS	CDS	6,812,602	6,812,994	393	forward	Bccel_5828	Partial
	hypothetical protein CDS	CDS	6,813,134	6,813,925	792	forward	Bccel_5829	None
	hypothetical protein CDS	CDS	6,813,952	6,814,239	288	forward	Bccel_5830	None
	hypothetical protein CDS	CDS	6,814,630	6,814,794	165	forward	Bccel_5831	None
	RDD domain containing protein CDS	CDS	6,814,901	6,815,482	582	forward	Bccel_5832	None
	hypothetical protein CDS	CDS	6,815,915	6,816,298	384	forward	Bccel_5833	None
	hypothetical protein CDS	CDS	6,816,681	6,817,343	663	reverse	Bccel_5834	None
	hypothetical protein CDS	CDS	6,817,315	6,817,689	375	reverse	Bccel_5835	Partial
	hypothetical protein CDS	CDS	6,830,314	6,831,552	1,239	reverse	Bccel_5851	Partial
	Protein of unknown function DUF2493 CDS	CDS	6,831,881	6,832,225	345	reverse	Bccel_5852	None

Table 5.7 continued…

Organism	Annotation	Feature	Start	End	Length	Orientation	Locus Tag	Assembly Coverage
B. cellulosolvens ATCC 35603	protein of unknown function DUF1813 HSP20-like protein CDS	CDS	6,832,404	6,832,574	171	reverse	Bccel_5853	Partial
	hypothetical protein CDS	CDS	6,833,562	6,833,732	171	forward	Bccel_5854	Partial
	hypothetical protein CDS	CDS	6,833,921	6,834,217	297	forward	Bccel_5855	Partial
	hypothetical protein CDS	CDS	6,834,593	6,835,831	1,239	forward	Bccel_5856	None
	hypothetical protein CDS	CDS	6,835,925	6,836,371	447	forward	Bccel_5857	None
	hypothetical protein CDS	CDS	6,836,440	6,836,892	453	forward	Bccel_5858	None
	hypothetical protein CDS	CDS	6,836,962	6,837,699	738	forward	Bccel_5859	None
	hypothetical protein CDS	CDS	6,837,833	6,838,156	324	forward	Bccel_5860	None
	hypothetical protein CDS	CDS	6,849,398	6,850,015	618	reverse	Bccel_5870	Partial
	hypothetical protein CDS	CDS	6,850,114	6,851,352	1,239	reverse	Bccel_5871	Partial
	transposase mutator type CDS	CDS	6,871,849	6,872,940	1,092	reverse	Bccel_5891	None
	integral membrane sensor hybrid histidine kinase CDS	CDS	6,873,159	6,875,603	2,445	reverse	Bccel_5892	None
	hypothetical protein CDS	CDS	6,875,600	6,875,875	276	reverse	Bccel_5893	None
	hypothetical protein CDS	CDS	6,877,356	6,877,523	168	reverse	Bccel_5894	None
	ABC-type glycine betaine transport, periplasmic subunit CDS	CDS	6,877,520	6,877,792	273	reverse	Bccel_5895	None
	hypothetical protein CDS	CDS	6,877,795	6,877,977	183	reverse	Bccel_5896	None
	ABC-type transporter, integral membrane subunit CDS	CDS	6,877,998	6,878,759	762	reverse	Bccel_5897	None
C. pasteurianum ATCC 6013	Contigs obtained from Illumina assembly were overlapping and no gaps were detected							

Table 5.8: Characteristics of unassembled DNA regions from PacBio technology.

Organism	Region Name	Start	Stop	Length	PacBio read coverage	% GC	Corresponding Annotation	ΔG (kcal/mol)
Clostridium thermocellum AD2	AD2_Overlap1	3502	5535	2033	36x	39.4	Membrane protein insertase	-20.41
	AD2_Overlap2	180557	182612	2055	116x	35.1	Transposase DDE domain	-15.44
	AD2_Gap1	558824	559892	1068	82x	39	Transposase mutator type	-12.92
Bacteroides cellulosolvens DSM 2933	BC_Overlap1	6343204	6349991	6788	36x	32.5	Transposase Tn3 family protein	-30.6
	BC_Gap1	6389652	6390057	405	4x	35.5	RNA-binding protein	-2.63

Table 5.9: Characteristics of randomly selected DNA regions from PacBio technology

Genome	Region_Name	Start	End	Length	% GC	ΔG (kcal/mol)	ΔH (kcal/mol)	ΔS (cal/(K·mol))
AD2	AD2_Random1	165211	166279	1069	34.5	-10.34	-449.7	-1338.9
AD2	AD2_Random2	950959	952027	1069	40.6	-4.43	-283.1	-849.2
AD2	AD2_Random3	1051495	1052563	1069	42.3	-16.47	-632.9	-1878.5
AD2	AD2_Random4	1677510	1678578	1069	32.6	-6.55	-210.1	-620.2
AD2	AD2_Random5	2216493	2217561	1069	42.8	-15.01	-590.9	-1754.9
AD2	AD2_Random6	2875296	2876364	1069	35.1	-3.2	-232.7	-699.3
AD2	AD2_Random7	3099308	3100376	1069	37.9	-9.97	-429.3	-1277.8
AD2	AD2_Random8	1835679	1836747	1069	35.17	-8.76	-397.7	-1185.2
AD2	AD2_Random9	2576564	2577632	1069	35.9	-7.94	-228.7	-672.7
AD2	AD2_Random10	3265470	3266538	1069	36.2	-7.56	-387.9	-1159
BC	BC_Random1	426289	426689	401	31.9	-0.09	-52	-158.1
BC	BC_Random2	557464	557864	401	37.4	-0.59	-96	-290.7
BC	BC_Random3	1372444	1372844	401	30.7	-1.82	-136.6	-410.7
BC	BC_Random4	1629296	1629696	401	39.4	-2.36	-140.9	-422.1
BC	BC_Random5	2124598	2124998	401	41.6	-2.36	-141	-422.4
BC	BC_Random6	2688359	2688759	401	28.9	-1.14	-62.2	-186
BC	BC_Random7	3123425	3123825	401	43.4	-7.51	-267.9	-793.5

Table 5.9 continued…

Genome	Region_Name	Start	End	Length	% GC	ΔG (kcal/mol)	ΔH (kcal/mol)	ΔS (cal/(K·mol))
BC	BC_Random8	3592492	3592892	401	36.4	-0.36	-33.9	-102.2
BC	BC_Random9	4069796	4070196	401	37.9	-1.08	-204.1	-618.6
BC	BC_Random10	4812535	4812935	401	41.9	-3.47	-181.4	-542.2

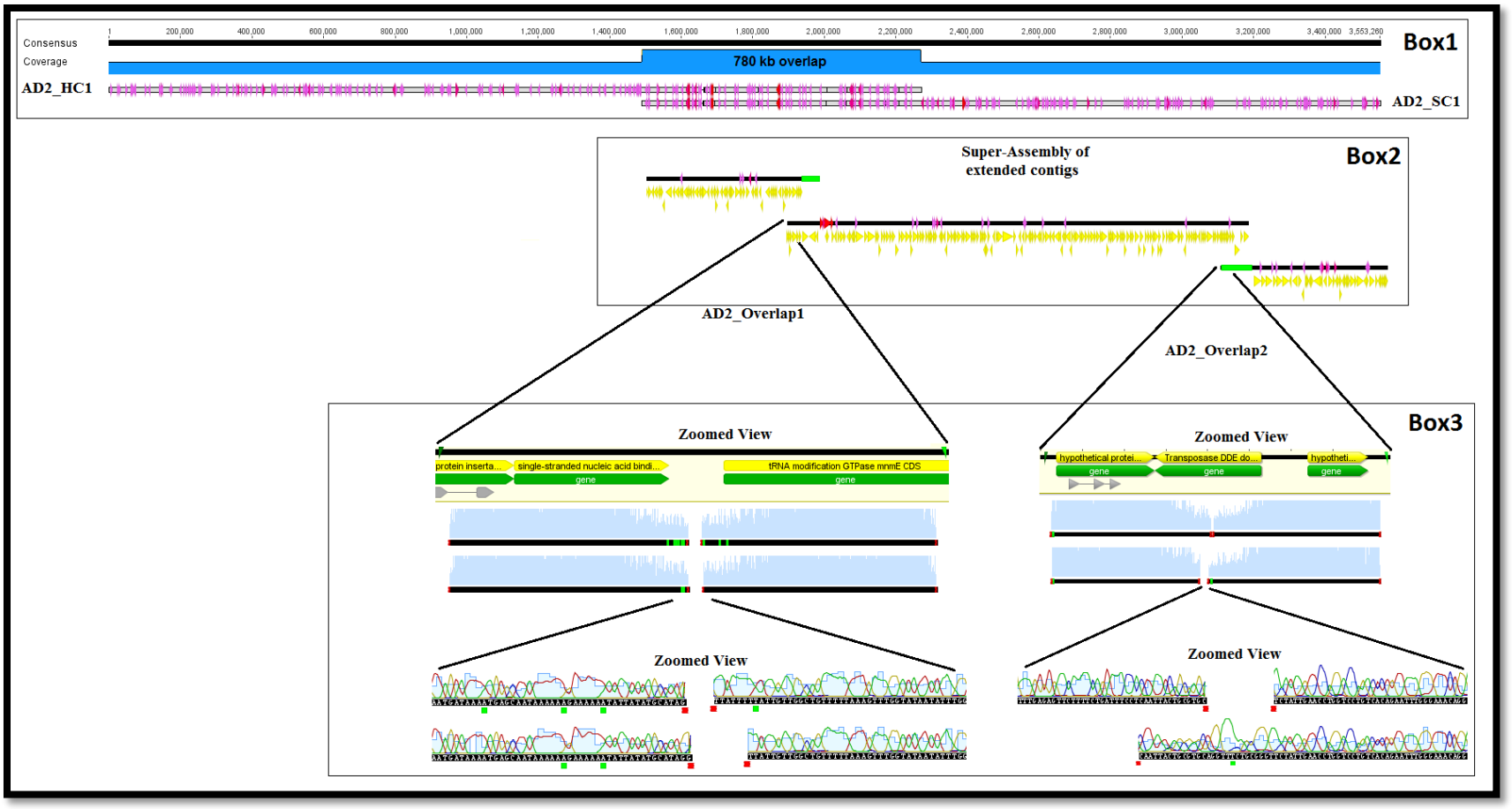

Figure 5.1: Example of manual genome finishing for AD2 genome.

Box1 shows a ~780 kb overlap between hybrid assembly contigs (AD2_HC1) and longest contig from super-assembly (AD2_SC1). Box2 describes super-assembly of three small overlapping contigs. Box3 shows an overview super-contig assembly verification by PCR and Sanger sequencing.

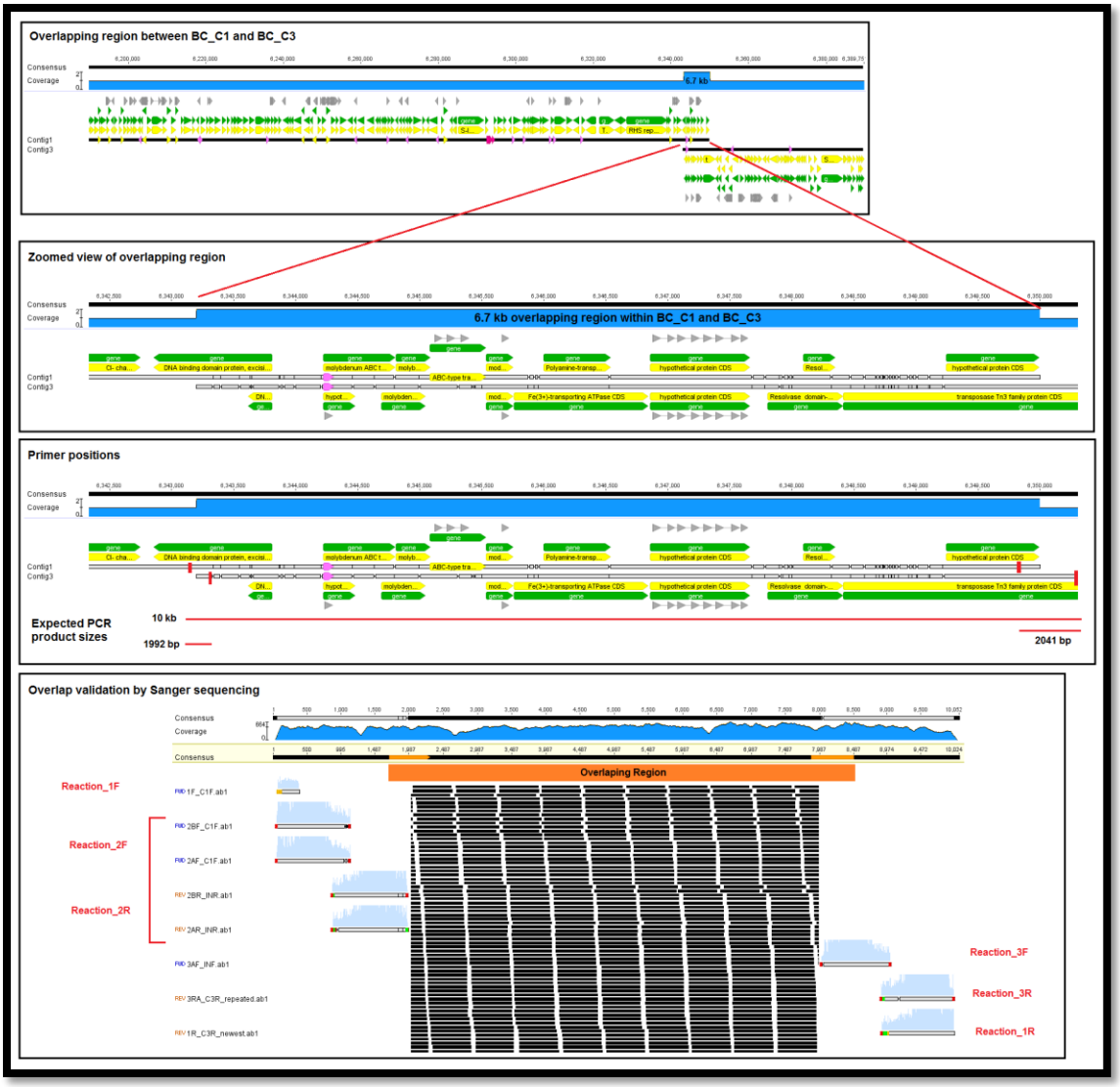

Figure 5.2: Validation of overlapping contigs from *B. cellulosolvens* DSM 2933 genome.

First box shows overview of an overlap between contigs BC_C1 and BC_C3. Next two boxes show the zoomed view, primer positions and expected PCR products. Last box shows an overview of overlap validation by PCR and Sanger sequencing and Illumina read mapping.

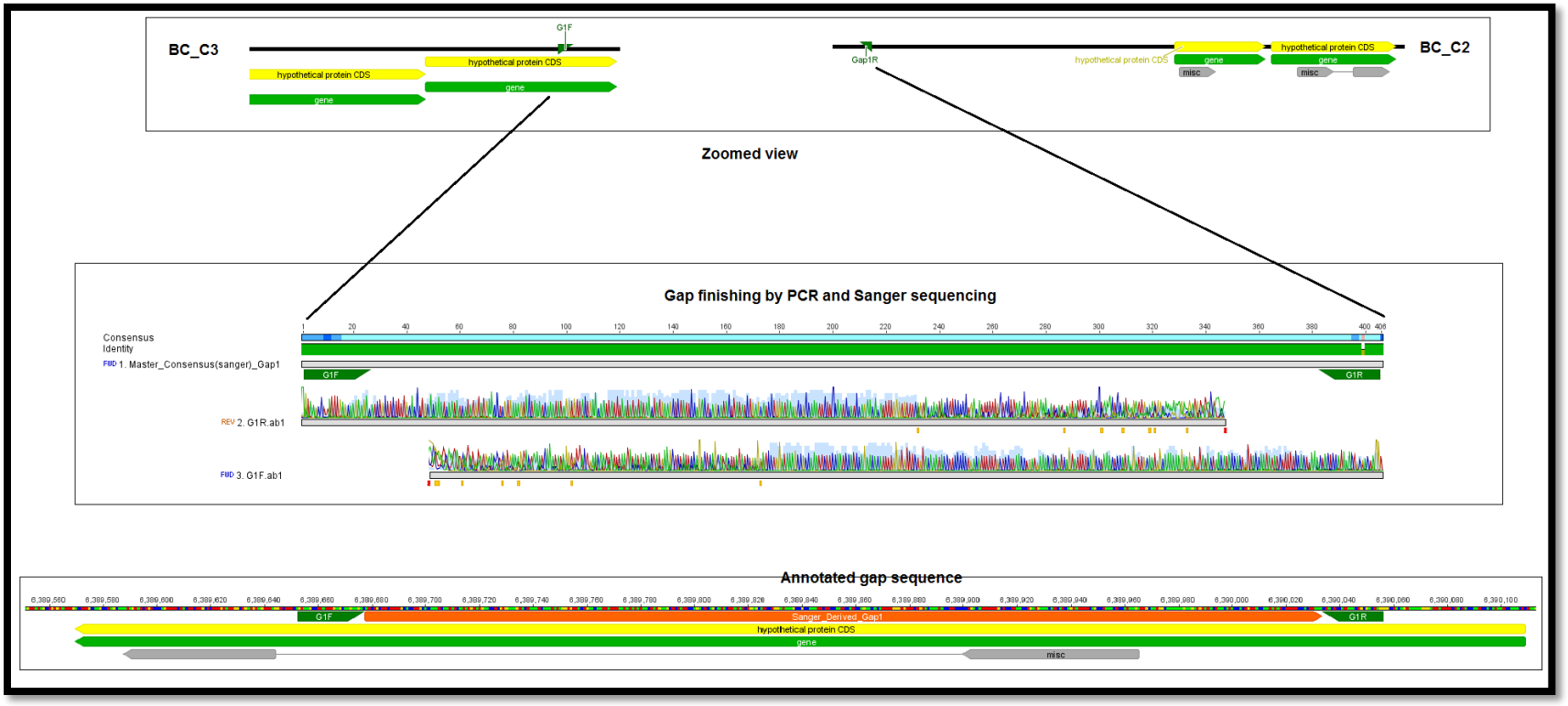

Figure 5.3: Overview of manual finishing of gap (BC_Gap1) from *B. cellulosolvens* DSM 2933 genome.

First box shows the unknown gap (BC_Gap1) between contigs (BC_C3 and BC_C2) and location of PCR primers. Next box shows the zoomed view of extended sequence for BC_Gap1 and verification by PCR and Sanger sequencing. Last box shows the post finishing annotation of BC_Gap1 sequence.

CHAPTER 6 : ENRICHMENT OF LIVE BACTERIAL ENDOPHYTES FROM *POPULUS DELTOIDES* FOR SINGLE-CELL GENOMICS

Disclosure:

Sagar Utturkar's contributions include bioinformatics analysis, de novo assemblies, contamination removal procedures, genome submissions and comparative phylogenetic and functional comparative analysis. Sagar also contributed towards designing the single-cell data analysis approach and development and writing of thesis manuscript which is under preparation for Nature Methods.

6.1 Abstract

Bacterial endophytes that colonize *Populus* trees have been shown to contribute to nutrient acquisition, prime immunity responses and as a result either directly or indirectly increase both above- and below-ground biomass. Because endophytes are found within plant material, a method for the physical separation of live endophytes from roots was developed for application for both Single Cell Genomics (SCG) and metagenomic studies. Root samples from three one-year-old *Populus deltoides* saplings were harvested from the Oak Ridge National Laboratory campus. The rhizosphere and rhizoplane of roots were removed by washes and sonication, and the roots were homogenized. Endophytic bacterial communities were enriched using differential and density gradient centrifugation. Total DNA was extracted from enriched and unenriched samples, and the endophytic bacterial community composition was determined by 16S rRNA gene amplification and sequencing. Our enrichment protocol reduced the number of contaminating chloroplast DNA reads by approximately by 10 fold and significantly increased the relative abundance of reads of Actinobacteria, Planctomycetia, and Alpha- and Gammaproteobacteria classes. Live bacterial enrichments inoculated onto agar plates for isolation or sorted by flow cytometry for single-cell genomics. Twelve single-cell genomes were selected for whole genome amplification depending on abundance of OTU in *Populus* rhizosphere, ability to form associations with plant and representing rare and uncultured phyla (from NCBI database). Single-cell genomics analysis including assembly, contamination removal and completeness estimation was performed. Comparative genomic analysis of each single-amplified genome (SAG) was performed to reveal the unique characteristics such as presence of biotin biosynthesis gene cluster in Armatimonadetes SAG, urease gene cluster in Planctomycetes SAG and, distinguished features such as iron scavenging genes in Acidobacteria SAG and putative ability to degrade complex plant material in Verrucomicrobia SAG. In conclusion, the current protocols allowed enrichment of endophytic bacteria away from the plant material and enabled single-cell genomics analysis on natural root samples by greatly reducing the amount of contaminating plant DNA which might otherwise mask such organisms in a background of 'contaminant' host data. These analyses will shed light on the genetic functions that contribute towards the various types of symbiotic relationship with *Populus* trees and potentially other species.

6.2 Introduction

Endophytic bacterial communities:

The soil surrounding the roots of plants accommodates an abundance of microorganisms due to the presence of nutrient rich plant derived exudates. The interface between plant root and soil constitute the rhizosphere (Rout and Callaway, 2012) and inside of the root tissues constitute the endosphere environment (Turner, et al., 2013). A microbiome in these root-associated environments is comprised of bacteria, fungi and to a lesser extent archaea which are virtually absent from the endosphere (Shakya, et al., 2013). Each of these may have potentially beneficial, neutral or detrimental effects on plant growth and development. Microorganisms associated with roots, within both the rhizosphere and endosphere, have been shown to positively contribute to plant growth. These organisms can promote plant growth by fixing atmospheric nitrogen, solubilizing inorganic phosphorus, increase the availability of nitrogen sources, producing plant auxins,

decreasing ethylene stress, suppressing pathogens, and inducing systemic resistance (Abramovitch, et al., 2006; Berendsen, et al., 2012; Bulgarelli, et al., 2013; Lugtenberg and Kamilova, 2009). Within this rhizosphere, bacterial concentrations can be as high as 10^9 cells/g of soils. A phylogenetically distinct portion of the soil and rhizosphere populations is able to cross into the root and comprise the bacterial endosphere (Shakya, et al., 2013). Endophyte populations can be as high as 10^8 cells/g of root material (Bulgarelli, et al., 2013), but most often are several orders of magnitude less at 10^4 of 10^5 cell/g of root. Because of the close association between endophytic bacterial communities and trees, metagenomes have been difficult to obtain due to the prevalence of contaminating plant DNA. Further, certain endophytic groups have been difficult to isolate and culture in a laboratory settings. Culture independent methods can provide information about yet to date uncultured endophytes and their phylogenetic and functional diversity.

The uncultured majority:

Microorganisms are the most diverse and abundant life forms on earth and yet our understanding of these microbes has been largely limited to the species which can be grown in culture. The widespread presence of numerous uncultured bacteria has been revealed through cultivation-independent approaches such as survey of molecular marker genes (e.g. 16S rRNA) or through metagenomics (Gilbert and Dupont, 2011; Rajendhran and Gunasekaran, 2011). However, conventional approaches to bring these bacteria to pure culture are limited by inability to mimic the required nutrients and microenvironment conditions. Some of the modern approaches previous applied for cultivating these difficult to isolate organisms includes the use of microfluidics chips to run several experiments in parallel (Seshadri, et al., 2003) or the recent iChip design to cultivate microbes in their natural environment (Ling, et al., 2015) separated from other potentially much faster growing organisms, that can overwhelm slower-growers. Despite a few successes achieved through above approaches, the large majority of these microorganisms have not been obtained in pure culture. An alternative culture-independent approach was to bypass the culturing altogether and instead learning from DNA by direct sequencing of uncultured microbes. This approach was termed single-cell genomics (Raghunathan, et al., 2005; Zhang, et al., 2006).

Single-cell sequencing:

Single-cell sequencing provides direct access to DNA sequence information from individual cells and thus access to genomic information of uncultured bacteria which can reveal novel insights into the functional potential, lifestyle and ecology of these understudied organisms. Single-cell sequencing can be advantageous over metagenomics sequencing for targeted recovery of genomes of uncultured varieties. With metagenomics it was generally not possible to assemble the genomes of the individual species, except for some of the most abundant in the community (Gilbert and Dupont, 2011) until recent advances in binning and assembly techniques (Wu, et al., 2015) to assemble complete genomes from metagenomes (Albertsen, et al., 2013; Narasingarao, et al., 2012; Smits, et al., 2014). In particular, natural populations with high degree of genomic heterogeneity are more accessible through single-cell genomics (SCG) (Rinke, et al., 2013). For example, two 16S rRNA genes of differing sequence were retrieved from single-cell belonging to bacillus cluster; while with metagenomics it would have been

impossible to differentiate the origin (Lasken, 2012). With SCG approach it is also possible to link plasmids (Dean, et al., 2001), viruses (Yoon, et al., 2011), archaea and bacteria (Rinke, et al., 2013) to the correct host organisms. The power of SCG approach was demonstrated by a recent study in which 200 single-cells were isolated from different habitats, including Nevada hot spring sediments and water from near hydrothermal vents in Pacific ocean; The researchers sequenced the genome of each cell and classified the cells into more than 20 new archaeal and bacterial lineages without any cultivated representatives (Rinke, et al., 2013). In general, the sequence information of uncultivated microbes provides information about putative genes, associated cellular functions and pathways, which might prove crucial information to develop appropriate culturing conditions. For example, efforts to culture *Coxiella burnetii* – the causative agent for Q-fever got a major boost when its genome sequence was available; Scientists were able to find differentially expressed genes when bacteria were growing successfully inside host cell and when they were struggling to grow alone, which provides hint for addition of certain amino-acids and peptides to the growth medium (Lok, 2015; Seshadri, et al., 2003).

Only a few years ago, DNA sequencing from single-cells was not feasible because of two major challenges. First, efficient sorting of individual single-cells from various microenvironments was a difficult task. Some of the early single-cell isolation techniques include dilution-to-extinction (serially dilute a sample solution until single cell remains) (Button, et al., 1993) and mechanical or optical micromanipulations (Brehm-Stecher and Johnson, 2004). Another important technological development was the application of flow-cytometry based Fluorescent-Activated Cell Sorting (FACS) method which allowed for high-throughput single-cell sorting at the rate of > 10^4 cells/second (Ishii, et al., 2010). Further the development of microfluidic based cell sorting devices allowed for mechanical capturing, incubation, release and compartmentalization of single-cells for further analyses (Arakawa, et al., 2011). Most recent advances include cell sorting devices designed on microchips to provide complete lab-on-a-chip setup for cellular isolation, analysis, amplification, culturing and/or experimental processing (Shields, et al., 2015). Briefly, cells of interest are (i) tagged with fluorescent labels such as nucleic-acid staining dye SYTO 9, specific antibodies, and DNA probes or (ii) captured using magnetic/polystyrene beads with specialized surface-binding capacity. The fluorescent-labelled cells are passed through path of multiple laser beams of different wavelength or separated and light emitted from each cell is converted into electric signals by optical detectors. Alternatively cells captured using specific beads are separated using an external magnetic, electrical or optical fields. Another approach for cell-sorting is label-free sorting which relies on physical differences in the cell properties such as size, shape, density, elasticity or magnetic susceptibility.

Most bacterial single-cells contain a few femtograms of DNA, which is highly insufficient for current NGS technologies (Lasken, 2012). Whole genome amplification from single-cell was first achieved by pioneering technique Multiple Displacement Amplification (MDA) which includes randomly primed PCR (Telenius, et al., 1992; Zhang, et al., 1992) and uses a highly processive φ29 DNA polymerase with strong strand displacement activity (Blanco and Salas, 1984). Under isothermal conditions, MDA extends random

primers to produce branched structures, which are extended by other primers and eventually form multibranched structures (Huang, et al., 2015). The resulting reaction allows enormous amplification of any DNA template (e.g. single bacterial genome can be amplified > 1 billion fold) (Raghunathan, et al., 2005).

Some of the technical challenges associated with SCG include occurrence of random amplification bias where different regions of the genome are under-represented in each MDA reaction and chimera formation occurs during DNA branching process (Lasken, 2007; Raghunathan, et al., 2005). Despite several approaches to limit the level of contamination (Lasken, 2012), a background amplification of contaminating DNA from samples and reagents themselves presents a challenge. The genome coverage for single-cell data-sets is highly variable and poses a challenge to traditional genome assembly methods which uses single coverage cutoffs and thus prevents assembly of significant proportion of data (Nurk, et al., 2013). Assembly problems have been overcome to certain extent by designing single-cell specific assemblers which can deal with non-uniform coverage and elevated levels of chimerism (Bankevich, et al., 2012; Peng, et al., 2012). Several studies have presented comprehensive strategies for choosing appropriate sequencing methods, experimental design and bioinformatics approaches (Lok, 2015). However, in practice the percentage of a given genome recovered varies greatly depending on the type of microorganism, the cell collection procedures, damage to DNA template during cell lysis and handling, and amplification bias (Lasken, 2012). A summary of recent advances in the field of SCG includes improvements in the ability to isolate single-cells by flow cytometry, micromanipulations, microfluidics, and various alternatives for whole genome amplification with wide range of applications are available (Blainey, 2013; Kalisky and Quake, 2011; Lasken, 2012; Macaulay and Voet, 2014; Stepanauskas, 2012).

In summary, application of shotgun metagenomics sequencing to interrogate endophytic samples is difficult due to presence of large amounts of host DNA relative to microbial DNA in the sample. Single-cell genomics analysis offers a complementary, cultivation independent, approach to obtain genomes of the uncultured candidate organisms from endophytic samples. In this study, we describe a method of enriching live endophytes from *Populus deltoides* roots, upstream from cultivation, isolation which in turn achieves reduction in host plant DNA and facilitates single-cell genomics analysis. The bioinformatics analysis steps ranging from genome assembly, contamination removal to comparative phylogenomics approaches are described. Comparative functional analysis helped to reveal the unique characteristics of the single-amplified uncultured endophytic bacteria as compared to their close relatives.

6.3 Methods

Disclaimer: This project was collaborative effort. Development of enrichment protocol and sample preparation and single-cell genomics analysis was performed by Dr. Nathan Cude. Data quality control and QIIME analysis of 16S data was performed by Dr. Michael Robeson. Library preparation and whole genome amplification was performed by Dawn Klingeman. Single-cell genomics data analysis including genome assembly, contamination removal and comparative genomics analysis was performed by Sagar

Utturkar. Dr. Dale Pelletier, Dr. Chris Schadt, Dr. Mircea Podar and Dr. Steven Brown designed and conceive the initial study, guided and contributed for manuscript preparation and corrections.

Root harvesting

Three one-year-old (above-ground) *Populus deltoides* saplings were harvested from a field on the Oak Ridge National Laboratory campus (35°55'20.2"N, 84°19'24.4"W). Whole root samples were collected from each tree, and roots ≤5 mm in diameter were separated and utilized for microbial enrichment. Total root weights used for enrichments were ~10 g (wet weight). Roots were cut into 1-2 cm long pieces and placed into a 300 ml sterile flask with 40 ml of autoclaved Milli-Q water. The flasks were shaken at 200 rpm for one min and the liquid was poured through sterile miracloth (EMD Millipore, Billerica, MA) and collected in a 50 ml conical tube. 100 ml of sterile Milli-Q water was added to the flasks containing the roots and the flask was placed in a water bath sonicator at 40 kHz (Branson 2510, Danbury, CT) for 5 min to remove the rhizosphere and rhizoplane soil and organisms. The liquid was then poured through sterile miracloth and collected in a 50 ml conical tube. The two washes were pooled for each tree and represented the rhizosphere samples. The roots were washed with sterile Mill-Q four more times and the liquid was discarded. An ethanol and UV (15 min) sterilized grinder (Braun KSM2, Kronberg, Germany) was used to homogenize the root samples in 40 ml of sterile Milli-Q water. The homogenate was poured through sterile miracloth and collected in a 50 ml conical tube. This homogenate thus largely represents the endosphere sample but might contain strongly adhered bacteria from the rhizoplane.

Differential and density centrifugation for microbial enrichment

Microbes were enriched using an adaptation of a previously described method (Ikeda, et al., 2009; Ikeda, et al., 2010). Prior to the enrichment, 1 ml of the rhizosphere and endosphere samples were saved as an unenriched control for sequencing. The endosphere homogenates and the rhizosphere samples were centrifuged at 500 × *g* for 5 min at 10°C (Beckman Coulter SPINCHRON R, Brea, CA). The supernatants were transferred to new conical tubes and centrifuged at 5500× *g* for 20 min at 10°C (Sorvall Evolution RC, Carlsbad, CA). The supernatants were discarded and the pellet was resuspended in 40 ml BCE buffer (50 mM Tris–HCl [pH7.5] and 1%Triton X-100). The suspension was filtered through a layer of sterile miracloth and transferred to a sterile 50 ml Oak Ridge tube (Nalgene, Rochester, NY). The suspensions were centrifuged at 10,000× *g* for 10 min at 10°C. The supernatants were discarded and the pellet was resuspended in 40 ml BCE buffer and filtered through a layer of sterile miracloth. The filtrate was centrifuged again at 10,000× *g* for 10 min at 10°C. The supernatant was discarded and the pellet was resuspended in 6 ml of 50 mM Tris-HCl (pH 7.5). The suspension was overlaid on 4 ml Histodenz (Sigma-Aldrich, St. Louis, MO) solution (8 g Histodenz dissolved in 10 ml of 50 mM Tris-HCl [pH 7.5]) in 10 ml Ultra-Clear centrifuge tubes (Beckman, Palo Alto, CA). The density centrifugation was run at 10,000× *g* for 40 min at 10°C (Beckman Coulter Optima LE-80K, Brea, CA). The microbial fraction (~1 ml) was visible as a white band at the Histodenz-water interface. The microbial fraction was collected and washed by centrifugation at 10,000× *g* for 3 min, removal of the supernatant, and the resuspending the pellet in 1 ml 50 mM Tris-HCl (pH 7.5). Half of the

sample was pelleted by centrifugation and stored at -20°C for DNA extraction. Glycerol at a final concentration of 25% v/v was added to the other half of the sample and it was stored at -80°C for single-cell sorting.

DNA extraction for microbiome sequencing

DNA for the enriched and unenriched rhizosphere samples was extracted using the PowerSoil DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA) using the provided protocol. DNA for the enriched and unenriched endosphere samples was extracted using the PowerPlant Pro DNA Isolation Kit with phenolic removal protocol (MO BIO Laboratories, Carlsbad, CA) using the provided protocol.

Sequencing, quality control, and analysis of paired end Illumina data

Libraries were prepared for the enriched endosphere samples. Paired-end sequencing of the V4 region of the bacterial rRNA was performed on the Illumina MiSeq platform (San Diego, CA) using the protocol of Lundberg *et al*. (Lundberg, et al., 2013). Sequence processing and quality control were performed through a combination of the UPARSE and QIIME pipelines (Caporaso, et al., 2010; Edgar, 2013). Cutadapt (Martin, 2011) was used in paired-end mode to trim sequencing primers from the forward and reverse reads. If either the forward or reverse primer was not detected within the read pairs, the pair was discarded. Paired-ends were merged using the fastq_mergepairs option of usearch (v.7.0.1001) (Edgar, 2013). An in-house python script was used to remove unused barcodes of paired-end sequences that did not survive merging. The QIIME (v1.7) (Caporaso, et al., 2010) script, split_libraries_fastq.py, was used to demultiplex the sequence data with the quality filter set to zero. Quality control processing was carried out via the UPARSE pipeline (e.g. -fastq_maxee 0.5) (Edgar, 2013) including *de novo* and reference-based (with -minh 1.5) chimera detection. The resulting OTU table was converted to BIOM format (McDonald, et al., 2012). Taxonomy was assigned using the RDP classifier (Wang, et al., 2007) against the updated May 2013 (v13_8) Greengenes database (DeSantis, et al., 2006; McDonald, et al., 2012; Werner, et al., 2012) via QIIME. Low read count OTUs were removed using the command QIIME command filter_otus_from_otu_table.py --min_count_fraction 0.00005. A phylogeny was constructed using FastTree (Price, et al., 2010) from a masked PyNAST (Caporaso, et al., 2010) alignment. The resulting phylogeny was manually rooted to Archaea via Dendroscope (v3) (Huson and Scornavacca, 2012). Finally, various diversity metrics were calculated via QIIME script core_diversity_analyses.py.

Single-cell sorting, multiple displacement amplification, and 16S rRNA Sanger sequencing

The enriched samples were stained with 5 µM Syto 9 nucleic acid stain (Life Technologies, Grand Island, NY). The stained samples were sorted on a Cytopeia Influx cell sorter (BD, Franklin Lakes, NJ) according to a previously published method (Campbell, et al., 2013). A flow cytometry plot was generated from forward scatter and green fluorescence. Ten gates were chosen from different positions on the plot. Single cells from enriched rhizosphere and endosphere samples from one tree were sorted into twenty 96-well plates (ten plates from the rhizosphere and ten plates from the endosphere; one plate each per gate).

The single-cell sorted plates were stored at -80°C prior to whole genome amplification by multiple displacement amplification (MDA) as published previously (Campbell, et al., 2013). Briefly, cells were lysed by 3 µL of a buffer of 0.13 M KOH, 3.3 mM EDTA pH 8.0 and 27.7 mM DTT, and heated to 95°C for 30 s. The reactions were immediately placed on ice for 10 min, and then neutralized by the addition of a buffer of 0.13 M HCl, 0.42 M Tris pH 7.0, 0.18 M Tris pH 8.0. The MDA was performed by adding 11µl to each well of a reaction solution of 90.9 µM random hexamers with two protective, phosporothioate bonds on the 3′ end (Integrated DNA Technologies, Coralville, IA, USA), 1.09 mM dNTPs (Roche Indianapolis, IN, USA), 1.8× phi29 DNA polymerase buffer (New England BioLabs, Ipswich, MA, USA), 4 mM DTT (Roche) and ~100 U phi29 DNA polymerase enzyme (purified in house). The MDA was performed in a thermocycler at 30°C for 10 h followed by inactivation at 80°C for 20 min. Plates were stored at -20°C.

For 16S rRNA sequencing of amplified DNA, 1 µl of the MDA was diluted into150 µl of PCR grade water. The remainder of the MDA was stored at -20°C. Universal 16S rRNA primers 27f (5'-AGAGTTTGATCMTGGCTCAG-3') and 1492r (5'-TACGGYTACCTTGTTACGACTT-3') were used to PCR amplify (in 50 µl reactions: 1x Pfu buffer, 200 µM dNTPs, 2 mM MgCl2, 5 µg bovine serum albumin, 300 µM forward and reverse primers, 0.2 µl Pfu polymerase, 37.90 µl dH2O, and 1 µl 1:150 MDA product) the majority of the 16S rRNA sequences. Conditions for the PCR were 94°C for 2 min, followed by 30 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 2 min, with a final extension at 72°C for 5 min. Positive amplifications were identified by gel electrophoresis (1.5% agarose w/v). Positive PCR products were purified with PCR filtration plates (Millipore, Billerica, MA). The purified 16S rRNA products were sequenced by fluorescent dye-terminator cycle Sanger sequencing at the University of Tennessee, Molecular Biology Resource Facility. Chromatograms were automatically trimmed using DNA Baser software (Heracle BioSoft, Pitești, Romania). Phylogenetic identifications were acquired using RDP classifier (Wang, et al., 2007) and NCBI BLASTN.

Whole genome amplification of single-cells

Twelve single-cell genomes were selected for whole genome amplification based on 16S rRNA assignment. Nextera XT sequencing libraries (Illumina, La Jolla, CA) were prepared according to the manufacturer's recommendations (Part # 15031942 Rev. E) stopping after library validation. In short, samples were fragmented, barcodes were appended, and samples were amplified. Libraries were cleaned using AMPure XP beads (Beckman Coulter, Indianapolis). Final libraries were validated on an Agilent Bioanalyzer (Agilent, Santa Clara, CA) using a DNA7500 chip and concentration was determined on a Qubit (Life Technologies) with the broad range double stranded DNA assay (life Technologies, Grand Island NY). Libraries were prepared for sequencing following the manufactures recommended protocols. The library was denatured with 0.2N sodium hydroxide and then diluted to the final sequencing concentration (19pM). Libraries were loaded into the sequencing cassette (v3) and a paired-end (2x300) run was competed on an Illumina MiSeq Instrument to obtain Single Amplified Genomes (SAGs). Later analysis discovered cross-contamination within two single-cells and hence new libraries were prepared followed by the second round whole genome sequencing for the Acidobacteria and Aramatimonadetes single cells.

Single-cell assembly, annotation and quality control analysis

Demultiplexed Illumina reads from the MiSeq software output were pre-processed using two separate approaches: (a) Khmer digital normalization (C. Titus Brown, 2012) and (b) Regular assembly protocol (Utturkar, et al., 2014). The Khmer digital normalization is a routinely applied method to single-cell data in order to decrease the memory and time requirements for *de novo* assembly without significant impact on the assembly contents. The Khmer protocol removes the redundant sequence reads, decreases sampling variation, removes the majority of errors and substantially reduces the size of the sequence data (C. Titus Brown, 2012). On the other hand, the regular assembly protocol utilized the complete set of raw reads without any data reduction. During regular assembly protocol, the quality trimming and filtering of raw sequence reads was performed for each SAG using CLC genomics workbench (CLC) (version 7.5.2) at quality cut-off value 0.02 (CLC, 2015). *De novo* genome assembly for each dataset (Khmer normalized and CLC trimmed) was performed using four assembly software - IDBA-UD (version 1.1.1) (Peng, et al., 2012), SPAdes (version 3.1.0) (Bankevich, et al., 2012), Velvet-sc (version 0.7.62) (Chitsaz, et al., 2011) and CLC, each ran with default options.

Single-cell sequence data is often found to be contaminated with organisms other than the target population and contamination removal is a necessary step (Beall, et al., 2014; Rinke, et al., 2013). A number of filtering operations was performed on assembled data to search for contaminated contigs. A nucleotide BLAST search was performed against NCBI non-redundant database and any contigs that matched (over half the contig length) with eukaryotic organisms were discarded. GC contents were determined for each contig and any outliers which were outside +/- 10% GC content range of target organism were discarded. Cross-contamination between samples was analyzed by conservative searching of all assemblies against each other using BLASTN. Sequence regions that have more than 99.5% identity over at least 5000 bp with another single-cell were removed from the smaller contigs. All discarded contigs were manually verified to identify any false positives. The initial annotation of the screened single-cell genomes was performed using the annotation pipeline at Oak Ridge National Laboratory (Hyatt, et al., 2010) and any contigs that did not contain protein coding regions were discarded.

The quality of the screened assemblies was verified using Kmer frequency analysis (outliers of the main cloud using settings: fragment window 1000 bp, fragment step 200 bp, oligomer size 4, minimum variation 10) before and after contamination removal. In second step, the 16S rRNA gene identified in each screened assembly was searched against the RDP database using “Sequence Match” and “Classifier” tools (Cole, et al., 2014). After contamination removal and quality analysis, assemblies for each SAG were submitted to the Integrated Microbial Genomes Expert Review (IMG-ER) system (Markowitz, et al., 2012) for gene prediction and annotation. Genome statistics and comparative analysis was performed using various IMG-ER tools (Chen, et al., 2013). The abundance profile tool was employed to create functional profiles (containing COG categories and Pfam clans) for each of the SAGs and their corresponding draft/finished genomes. The complete list of description/annotation for the Pfam clans (https://img.jgi.doe.gov/cgi-bin/er/main.cgi?section=FindFunctions&page=pfamListClans) and the COG categories

(https://img.jgi.doe.gov/cgi-bin/er/main.cgi?section=FindFunctions&page=cogid2cat) is available at the IMG website.

Single-cell genome completeness estimation. Genome size and assembly completeness was estimated using a previously published quality matrix (Land, et al., 2014) which assigns (a) Essential score – based on the presence of a set of essential genes containing 102 conserved Pfam-A domain found in nearly all bacteria and archaea (b) tRNA score – based on the presence of at least one tRNA coding for all of the 20 standard amino acids (c) rRNA score – based on the presence of a full-length 5S, 16S, and 23S rRNA; and (d) quality score – based on sequence quality (function of number of contigs and number of non-standard bases in the assembly).

Data sharing information. The five assembled single-cell genomes are available on IMG website with IDs 2626541630 (*Zavarzinella* R9F7), 2626541631 (*Zavarzinella* E9H3), 2626541628 (*Acidobacteria*), 2626541627 (*Verrucomicrobia*), 2626541629 (*Armatimonadetes*).

6.4 Results

Enrichment and analysis of endophytic bacteria

Approximately 10^7 – 10^8 cells were enriched from the rhizosphere and endosphere samples using the current method (data not shown). On average 33.67 ± 7.07 ng of DNA was isolated from the enrichments. By contrast, unenriched extractions yielded an average of 605.25 ± 469.84 ng of DNA. The 16S rRNA phylotyping performed on the three enriched and three unenriched endosphere samples demonstrated that *Proteobacteria* dominated the endosphere of these saplings. These data showed similar read percent abundance at the phylum level, though significant differences exist (Figure 6.1). Phyla that were significantly increased in read abundance percentage in the average enrichment of the three trees were the Actinobacteria and the Planctomycetia ($P<0.01$; false discovery rate (FDR) corrected). The *Proteobacteria* showed different enrichment profiles at the class level. Alpha- and Gammaproteobacteria were significantly increased in read abundance percentage ($P<0.1$, FDR corrected). Betaproteobacteria showed no significant difference, while Deltapreoteobacteria were significantly decreased read abundance percentage ($P<0.01$, FDR corrected). Contaminating chloroplast reads from the roots were also significantly decreased in the enrichment by approximately 10 fold (~7% to ≤0.7% of all reads; $P<0.01$, FDR corrected).

Single-cell sorting, MDA amplification, and sequencing

For single-cell sorting, the endosphere and rhizosphere enrichments from one tree were chosen, and cells from each sample were sorted into ten 96-well plates from 10 different gates on the cytometry plot. After MDA whole genome amplification and 16S rRNA gene PCR amplification, there were 169 positive 16S rRNA gene amplifications (86 from the endosphere and 83 from the rhizosphere) based on agarose gel observations. PCR investigations of wells that did not produce bacterial 16S rRNA gene signals suggested that a further 179 wells may have contained fungal cells (data not shown). Of the 169 positive 16S rRNA signals, 115 were successfully sequenced by the Sanger method. RDP Classifier (Wang, et al., 2007) and the NCBI reference RNA database were used to

assign phylogeny to the amplified signals. Sorted cells represented multiple phyla including Acidobacteria, Actinobacteria, Armatimonadetes (formally OP10), Bacteroidetes, Firmicutes, Planctomycetes, *Proteobacteria*, and Verrucomicrobia. Several 16S rRNA sequences appeared to represent members of the human microbiome, implying skin contamination. These sequences correspond to *Corynebacterium* spp., *Propionibacterium acnes*, and *Staphylococcus epidermidis.* It is unclear where this contamination originated as care was taken during the harvest and preparation of the samples. OTUs of these sequences were present in the 16S rRNA gene phylotyping data, though at low abundances (data not shown). Regardless, novel 16S rRNA sequences (<97% identity to sequenced relatives) from multiple phyla were present in the sorted cells. Twelve single-cell genomes were selected for whole genome amplification depending on abundance of OTU in *Populus* rhizosphere, ability to form associations with plant and representing rare and uncultured phyla (from NCBI database). The 16S rRNA gene sequences from these single-cells analyzed by the BLAST search and revealed greater than 99% identity to *Zavarzinella* sp. (9 SAGs), *Armatimodetes* sp., *Acidobacteria* sp., and *Verrucomicrobia* sp. The MDA amplification and whole genome sequencing of 12 selected single-cells was performed using Illumina MiSeq instrument which generated 300 bp paired-end reads.

Genome assembly and contamination screening of single-cell amplified genomes. *De novo* genome assembly of single-cells was performed using two data pre-processing approaches (khmer digital normalization and regular assembly) and four assembly software (SPAdes, Velvet-sc, IDBA-UD and CLC) as described in methods section. Independent of applied pre-processing approach, the IDBA-UD assembler always generated the best assembly results in terms of N50 statistics and total length assembled. It is worth mentioning that although khmer normalization have become prevalent step during single cell assembly, the Khmer authors have prepared a blog about application of Khmer protocol (http://ivory.idyll.org/blog/why-you-shouldnt-use-diginorm.html) which clearly suggests that normalization steps are not necessary when comparable results are obtained through regular assembly protocol. Our data generated comparable statistics with both khmer and regular assembly protocols. Therefore, the IDBA-UD assemblies generated with regular assembly protocols were used for further downstream analysis. Contamination screening was performed as described in methods section. The assembly kmer frequency distribution graph before contamination removal showed presence of two distinct clouds. After contamination removal steps the majority of the second cloud (belonging to outlier contaminants sequences) disappeared. Additionally, the 16S rRNA sequence derived from each assembly was found to be matching with target organism of interest using RDP database, “Sequence Match” and “Classifier” tools. Detailed assembly statistics for each SAG after contamination removal are presented in Table 6.1.

Genome completeness analysis: The genome completeness for each SAG was measured as described in the methods section. The essential score provides a completeness estimation based on the presence of a set of essential genes containing 102 conserved Pfam-A domains found in nearly all bacteria and archaea (Land, et al., 2014). The tRNA score, rRNA score and sequence quality score provides additional criteria for completeness estimation. The *Zavarzinella* and *Armatimodetes* SAGs are

estimated to represent more than 60% of the complete genome, *Verrucomicrobia* SAGs estimated to represent 57% of the complete genome while *Acidobacteria* SAG represented the least 39% of the complete genome. These results are in accordance with a recent study which estimated genome completeness of 201 SAG from uncultivated archaeal and bacterial cells in the range of less than 10% to greater than 90% and mean of 40% (Rinke, et al., 2013). Low scores for tRNA and sequence quality can be attributed towards fragmented and incomplete assemblies. The individual quality scores for completeness estimation of each SAG are presented in Table 6.2.

Functional characterization of single-cells: The COG and Pfam functional profiles for SAGs were created as described in methods section. The functional profile assigned gene counts to each COG and Pfam categories. Additional filtering was applied to identify the COG and Pfam categories that are shared, unique to SAGs, missing from SAGs as compared to finished/draft genome. The putative functional characteristics for individual SAGs are described below.

1. SAG of phylum Armatimonadetes

The Armatimonadetes SAG was compared against the complete genomes of two other Armatimonadetes members, *Fimbriimonas ginsengisoli Gsoil* 348 (IMG ID 2585427636) (Hu, et al., 2014) and *Chthonomonas calidirosea T49*, DSM 23976 (IMG ID 2524614646) (Lee, et al., 2014).

The main difference between the Armatimonadetes SAG and the finished genomes was the presence of genes related to biotin (vitamin B7) biosynthesis in Armatimonadetes SAG. The biotin biosynthesis starts with the metabolite malonyl-ACP which is converted to pimeloyl-ACP through a series of reactions. Alternatively, some bacteria derive pimeloyl-CoA from pimelate (Lin and Cronan, 2011). The pimeloyl-ACP/pimeloyl-CoA acts as a precursor and conversion to biotin takes place through four reaction step. Interestingly, the genes involved in the final four steps (8-amino-7-oxononanoate synthase, adenosylmethionine-8-amino-7-oxononanoate transaminase, dethiobiotin synthase, and biotin synthase) were present only in the Armatimonadetes SAG and missing from the finished genomes. However, some intermediate genes involved in conversion of malonyl-ACP or pimelate to precursor molecules pimeloyl-ACP/pimeloyl-CoA were missing from the SAG genome (Figure 6.2).

The Armatimonadetes SAG contains 21 σ-70-like proteins and has a high σ-factor to genome size (σ/Mb) ratio similar to *Chthonomonas calidirosea* strain *T49*. Central metabolism appears to proceed via standard glycolysis, tricarboxylic acid cycle although some key genes were missing. The presence of genes related to oxidative phosphorylation supports possible aerobic respiration phenotype. This SAG also contains genes for extracellular nitrate/nitrite transporter and assimilatory nitrite reductase components (*nirB, nirD*) which are involved in nitrogen metabolism. We also identified the genes encoding for cyanase (Ga0064453_13372) and carbonic anhydrase (Ga0064453_11265, Ga0064453_12094) which have possible role in environmental cyanate tolerance. Additionally, genes involved in twin-arginine translocation (Tat) pathway such as “Tat pathway signal sequence” and “translocase protein, tatA/E family”, and 61 ABC-transporter related genes were identified.

2. SAG of phylum Planctomycetes

Two SAGs (E9H3 and R9F7) of phylum Planctomycetes corresponding to *Zavarzinella* spp. based on 16S rRNA assignment. These SAGs were compared against the draft genome of cultured isolate *Zavarzinella formosa* strain A10^T (IMG ID 2548877000) (Guo, et al., 2012).

The key distinction between the *Zavarzinella* SAGs and *Zavarzinella formosa* strain A10^T was the presence of the urease system as a unique feature of SAG E9H3. The urease gene cluster (including urease alpha, beta and gamma subunits, urease accessory proteins UreF, UreG, UreH) was detected as part of the operon on contig Ga0068558_10012 in SAG E9H3. Other accessory genes encoding for urea binding protein and urea ABC transporter were also detected on the same contig and as part of the operon (Figure 6.3). Active ureases have a nickel containing active site to catalyze the hydrolysis of urea to ammonia and carbamate (Lv, et al., 2011). We also identified the genes related to COG0378 with predicted function "$Ni2^+$-binding GTPase involved in regulation of expression and maturation of urease and hydrogenase" in both E9H3 and R9F7 SAGs and were missing from strain A10^T. Secondly, SAG E9H3 also contained the gene related to "Hydrogenase/urease accessory protein HupE" which is implicated as secondary transporter for nickel or cobalt (Zhang, et al., 2009). Additionally, the glutamate and arginine decarboxylase genes were present in SAG E9H3 while genes for FoF1-type ATPase were present in SAG R9F7.

Most of the genes involved in glycolysis, citric acid cycle, pentose phosphate pathway and pyruvate metabolism were identified in both SAGs and *Zavarzinella formosa* strain A10^T which suggest a common route for central metabolism. The IMG phenotype prediction tool (Chen, et al., 2013) have predicted the aerobic phenotype for the SAG E9H3 based on presence of the genes "cytochrome bd-I ubiquinol oxidase" (Ga0068558_1004513, Ga0068558_1004514) which are known to be involved in ubiquinol oxidation. Several other genes involved in oxidative phosphorylation were detected in both SAGs. Interestingly, the cytochrome-bd complex genes were detected only in E9H3 but were missing from strain A10^T and R9F7. Furthermore, several genes involved in antibiotic resistance (multidrug resistance and transport, Beta-lactamase class A, antimicrobial peptide transport), genes encoding for pilus assembly proteins, and genes related to various non-specific "glycosyl hydrolase" families were identified in both SAGS. A gene encoding for putative pectate lyase was found only in R9F7 SAG.

3. SAG of phylum Verrucomicrobia

The SAG of Verrucomicrobia E1D9 was compared against the draft genome sequence of its relative *Chthoniobacter flavus* Ellin428 (Kant, et al., 2011). Most of the genes involved in glycolysis pathway, several genes involved in citric acid cycle and pentose phosphate pathway were identified in this SAG suggesting traditional route for carbon metabolism. Although, majority of the members of phylum Verrucomicrobia exhibit aerobic phenotype, majority of the genes involved in oxidative phosphorylation were missing from current SAG. A putative catalase gene (Ga0068556_101862) was present in SAG which was similar to one present in strain Ellin428. Based on Pfam functional profile, 43 protein coding genes related to various "Glycosyl hydrolase" families were identified which includes 7 genes corresponding to "Cellulase (glycosyl hydrolase family 5)", 15 genes corresponding to "Glycosyl hydrolases family 16". Thirteen "Glycosyl

hydrolase" genes were found only in the Verrucomicrobia SAG. Additionally, a few genes involved in multidrug transport and resistance were detected in SAG.

4. SAG of phylum Acidobacteria

The SAG of phylum Acidobacteria had maximum BLASTP hits against soil bacterium *Candidatus Koribacter versatalis* (*Koribacter*) (IMG ID 2606217699) and *Candidatus Solibacter usitatus* (*Solibacter*) (IMG ID 639633060) (Ward, et al., 2009). The average nucleotide identity (ANI) (Richter and Rossello-Mora, 2009) of *Solibacter* was 72.66% while that of *Koribacter* was 61%. Therefore, the *Solibacter* genome was selected for comparative analysis.

The Acidcobacteria SAG was the least complete genome (39%) of all and many genes related to central carbon metabolism and energy metabolism were missing. Only a few intermediate genes encoding for key enzymes in TCA cycle such as ("Succinyl-CoA synthetase", "Citrate synthase", and "Succinate dehydrogenase") were observed. However, other important genes detected in this SAG include genes related to ABC transporters, genes corresponding to various "Glycosyl hydrolase" families, putative enzymes such as polysaccharide lyase and Pectate lyase, genes related to nitrate/nitrite transport system and genes involved in dissimilatory nitrite reduction. The ABC transporter genes include iron and urea transporters, and different type of secondary porters for polysaccharides, Na^{+}, antimicrobial peptides etc. Putative genes related to macrolide exporters were detected and also found to be abundant in *Solibacter* genome. The genes related to "Cellulase (glycosyl hydrolase family 5" suggest possible ability to degrade cellulose substrates. The nitrite reduction genes include putative respiratory nitrite reductase (NrfH) precursor and ABC type transporters for nitrate. Additionally, several genes related to ferrous (Fe^{+2}), hemin and siderophore transport system and putative TonB receptors were detected which could possibly be involved in iron scavenging.

6.5 Discussion

Single-cell characterization:

Comparisons between SAGs and corresponding finished/draft genomes revealed the presence of several unique genes and functional characteristics of each SAGs which allowed for the prediction of putative roles for these bacteria in plant vicinity. The final four steps in biotin biosynthesis pathway are known to be conserved among biotin-producing organisms (Rodionov, et al., 2002) and also found as unique characteristic of Armatimonadetes SAG suggesting biotin autotroph phenotype. The high abundance σ-factors are predicted to coordinate transcriptional regulation of functionally related but dispersed genes (Lee, et al., 2014) and likely to be involved in transcription regulatory mechanism. The Tat pathway is known to be involved in translocation of folded proteins across lipid bilayer membranes (Lee, et al., 2006) and likely to be serving similar function in Armatimonadetes. The carbonic anhydrases gene is involved in rapid inter-conversion of carbon dioxide and water to bicarbonate, which is an important intermediate for cyanate degradation reaction (Guilloton, et al., 1993; Smith and Ferry, 2000). While the cyanase gene catalyzes the conversion of cyanate and bicarbonate to produce ammonia and carbon dioxide, and confers ability to tolerate environmental cyanate (Sung and Fuchs, 1988). The presence of both carbonic anhydrases and cyanase gene suggests that

Armatimonadetes could have adapted the cynate tolerance mechanism. The acid tolerance or pH homeostasis in bacteria is maintained through various systems such as F1F0-ATPase proton pump (Cotter and Hill, 2003), arginine and/or glutamate decarboxylase system (Richard and Foster, 2004; Richard and Foster, 2003) and urease system (Stingl, et al., 2002; Wilson, et al., 2014). The presence of these genes in Planctomycetes SAGs suggests an ability of pH tolerance and regulation. The pilus assembly related genes in Planctomycetes SAGs might serve to function in cell-to-cell or surface attachment, as observed in case of *Z. formosa* strain A10^{T}A (Kulichevskaya, et al., 2009). Acidobacteria 16S rRNA gene sequence have been detected and sometimes found to dominate the iron-rich mine environments (Blothe, et al., 2008; Kleinsteuber, et al., 2008) and there is growing evidence in literature that Acidobacteria play an important role in iron redox reactions (Mondani, et al., 2011). The presence of genes related to ferrous (Fe^{+2}), hemin and siderophore transport suggest possible ability of Acidobacteria SAG to use siderophores produced by other microorganisms. Bacteria that can scavenge iron via excreted siderophores have potential advantage in soil (Ward, et al., 2009). The presence of genes such as “Cellulase (glycosyl hydrolase family 5”, polysaccharide lyase and pectate lyase might provide ability to degrade complex plant material. A catalase gene was detected in both Verrucomicrobia SAG and *C. flavus* Ellin428. However, biochemical testing revealed that strain Ellin428 is catalase negative (Sangwan, et al., 2004), and same could be true for Verrucomicrobia SAG.

In past few years, there have been a large increase in the number of single-cell genomics studies, and many taxonomic groups have received their first genomes (Rinke, et al., 2013). Large scale studies have been carried out including the Microbial Earth Project, which aims to generate a comprehensive genome catalogue of all archaeal and bacterial type strains (http://www.microbial-earth.org), the Human Microbiome Project (Turnbaugh, et al., 2007) with goal of sequencing uncultivated bacteria from human microbiome to understand the microbial components of human genetic and metabolic landscape and their contribution to normal physiology and disease predisposition. DNA amplification methods that were originally developed for bacteria are becoming reliable enough for use with diploid cells (McConnell, et al., 2013; Shapiro, et al., 2013) and sequencing of single eukaryotic cells is also improving. The *in silico* analysis methods for single-cell genomics data including *de novo* genome assembly, contamination screening and comparative genomics are under constant improvement with exciting era of single-cell biology ahead.

6.6 Conclusion

Physical separation and isolation of endophytic bacteria associated with plant materials is a challenging task. Our modified enrichment protocol based on differential and density gradient centrifugation was able to achieve a significant reduction in contaminating plant DNA and enriched the endophytic bacteria. This protocol enabled to perform single-cell genomics analysis of enriched bacterial samples which allowed to select, amplify and analyze the genomes of the previously uncultured bacteria of interest. These samples could also be analyzed by the metagenomics approach to perform bacterial community analysis and functional characterization. Bioinformatics and comparative genomics analysis revealed the unique characteristics of these SAGs as compared to their close relative bacteria. The unique characteristics include the presence of biotin biosynthesis

gene cluster in Armatimonadetes SAG, urease gene cluster in Planctomycetes SAG and, distinguished features such as iron scavenging genes in Acidobacteria SAG and putative ability to degrade complex plant material in Verrucomicrobia SAG. This genomic information could facilitate future efforts to culture these bacteria. For example, addition of urea to the growth medium for Planctomycetes might help bacteria to thrive and improve the survival rate. Ultimately the primary focus of this study is to provide a proof-of-concept for the modified enrichment protocol for separation and isolation of live endophytic bacteria sample and further analysis by single-cell genomics method.

References

Abramovitch, R.B., Anderson, J.C. and Martin, G.B. (2006) Bacterial elicitation and evasion of plant innate immunity, *Nat. Rev. Mol. Cell Biol.*, **7**, 601-611.

Albertsen, M.*, et al.* (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes, *Nat. Biotechnol.*, **31**, 533-538.

Arakawa, T.*, et al.* (2011) High-throughput single-cell manipulation system for a large number of target cells, *Biomicrofluidics*, **5**, 14114.

Bankevich, A.*, et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J. Comput. Biol.*, **19**, 455-477.

Beall, C.J.*, et al.* (2014) Single cell genomics of uncultured, health-associated Tannerella BU063 (Oral Taxon 286) and comparison to the closely related pathogen Tannerella forsythia, *PLoS One*, **9**, e89398.

Berendsen, R.L., Pieterse, C.M. and Bakker, P.A. (2012) The rhizosphere microbiome and plant health, *Trends Plant Sci.*, **17**, 478-486.

Blainey, P.C. (2013) The future is now: single-cell genomics of bacteria and archaea, *FEMS Microbiol. Rev.*, **37**, 407-427.

Blanco, L. and Salas, M. (1984) Characterization and purification of a phage phi 29-encoded DNA polymerase required for the initiation of replication, *Proc Natl Acad Sci U S A*, **81**, 5325-5329.

Blothe, M.*, et al.* (2008) pH gradient-induced heterogeneity of Fe(III)-reducing microorganisms in coal mining-associated lake sediments, *Appl. Environ. Microbiol.*, **74**, 1019-1029.

Brehm-Stecher, B.F. and Johnson, E.A. (2004) Single-cell microbiology: tools, technologies, and applications, *Microbiol. Mol. Biol. Rev.*, **68**, 538-559, table of contents.

Bulgarelli, D.*, et al.* (2013) Structure and functions of the bacterial microbiota of plants, *Annu. Rev. Plant Biol.*, **64**, 807-838.

Button, D.K.*, et al.* (1993) Viability and isolation of marine bacteria by dilution culture: theory, procedures, and initial results, *Appl. Environ. Microbiol.*, **59**, 881-891.

C. Titus Brown, A.H., Qingpeng Zhang, Alexis B. Pyrkosz, Timothy H. Brom (2012) A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data, *arXivorg*.

Campbell, A.G.*, et al.* (2013) Multiple single-cell genomes provide insight into functions of uncultured Deltaproteobacteria in the human oral cavity, *PLoS One*, **8**, e59361.

Caporaso, J.G.*, et al.* (2010) PyNAST: a flexible tool for aligning sequences to a template alignment, *Bioinformatics*, **26**, 266-267.

Caporaso, J.G.*, et al.* (2010) QIIME allows analysis of high-throughput community sequencing data, *Nat. Methods*, **7**, 335-336.

Chen, I.M.*, et al.* (2013) Improving microbial genome annotations in an integrated database context, *PLoS One*, **8**, e54859.

Chitsaz, H.*, et al.* (2011) Efficient de novo assembly of single-cell bacterial genomes from short-read data sets, *Nat. Biotechnol.*, **29**, 915-921.

CLC (2015) CLC Genomics Workbenach Manual - Trimming using the Trim tool.

Cole, J.R.*, et al.* (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis, *Nucleic Acids Res.*, **42**, D633-642.

Cotter, P.D. and Hill, C. (2003) Surviving the acid test: responses of gram-positive bacteria to low pH, *Microbiol. Mol. Biol. Rev.*, **67**, 429-453, table of contents.

Dean, F.B.*, et al.* (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification, *Genome Res.*, **11**, 1095-1099.

DeSantis, T.Z.*, et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB, *Appl. Environ. Microbiol.*, **72**, 5069-5072.

Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads, *Nat. Methods*, **10**, 996-998.

Gilbert, J.A. and Dupont, C.L. (2011) Microbial metagenomics: beyond the genome, *Ann Rev Mar Sci*, **3**, 347-371.

Guilloton, M.B.*, et al.* (1993) A physiological role for cyanate-induced carbonic anhydrase in Escherichia coli, *J. Bacteriol.*, **175**, 1443-1451.

Guo, M.*, et al.* (2012) Genome sequences of three species in the family Planctomycetaceae, *J. Bacteriol.*, **194**, 3740-3741.

Hu, Z.Y.*, et al.* (2014) The first complete genome sequence of the class Fimbriimonadia in the phylum Armatimonadetes, *PLoS One*, **9**, e100794.

Huang, L.*, et al.* (2015) Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications, *Annu Rev Genomics Hum Genet*, **16**, 79-102.

Huson, D.H. and Scornavacca, C. (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks, *Syst. Biol.*, **61**, 1061-1067.

Hyatt, D.*, et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, **11**, 119.

Ikeda, S.*, et al.* (2009) Development of a bacterial cell enrichment method and its application to the community analysis in soybean stems, *Microb. Ecol.*, **58**, 703-714.

Ikeda, S.*, et al.* (2010) Community- and genome-based views of plant-associated bacteria: plant-bacterial interactions in soybean and rice, *Plant Cell Physiol.*, **51**, 1398-1410.

Ishii, S., Tago, K. and Senoo, K. (2010) Single-cell analysis and isolation for microbiology and biotechnology: methods and applications, *Appl. Microbiol. Biotechnol.*, **86**, 1281-1292.

Kalisky, T. and Quake, S.R. (2011) Single-cell genomics, *Nat. Methods*, **8**, 311-314.

Kant, R.*, et al.* (2011) Genome sequence of Chthoniobacter flavus Ellin428, an aerobic heterotrophic soil bacterium, *J. Bacteriol.*, **193**, 2902-2903.

Kleinsteuber, S.*, et al.* (2008) Diversity and in situ quantification of Acidobacteria subdivision 1 in an acidic mining lake, *FEMS Microbiol. Ecol.*, **63**, 107-117.

Kulichevskaya, I.S.*, et al.* (2009) Zavarzinella formosa gen. nov., sp. nov., a novel stalked, Gemmata-like planctomycete from a Siberian peat bog, *Int. J. Syst. Evol. Microbiol.*, **59**, 357-364.

Land, M.L.*, et al.* (2014) Quality scores for 32,000 genomes, *Stand Genomic Sci*, **9**, 20.

Lasken, R.S. (2007) Single-cell genomic sequencing using Multiple Displacement Amplification, *Curr. Opin. Microbiol.*, **10**, 510-516.

Lasken, R.S. (2012) Genomic sequencing of uncultured microorganisms from single cells, *Nat. Rev. Microbiol.*, **10**, 631-640.

Lee, K.C.*, et al.* (2014) Genomic analysis of Chthonomonas calidirosea, the first sequenced isolate of the phylum Armatimonadetes, *ISME J*, **8**, 1522-1533.

Lee, P.A., Tullman-Ercek, D. and Georgiou, G. (2006) The bacterial twin-arginine translocation pathway, *Annu. Rev. Microbiol.*, **60**, 373-395.

Lin, S. and Cronan, J.E. (2011) Closing in on complete pathways of biotin biosynthesis, *Mol Biosyst*, **7**, 1811-1821.

Ling, L.L.*, et al.* (2015) A new antibiotic kills pathogens without detectable resistance, *Nature*, **517**, 455-459.

Lok, C. (2015) Mining the microbial dark matter, *Nature*, **522**, 270-273.

Lugtenberg, B. and Kamilova, F. (2009) Plant-growth-promoting rhizobacteria, *Annu. Rev. Microbiol.*, **63**, 541-556.

Lundberg, D.S.*, et al.* (2013) Practical innovations for high-throughput amplicon sequencing, *Nat. Methods*, **10**, 999-1002.

Lv, J.*, et al.* (2011) Structural and functional role of nickel ions in urease by molecular dynamics simulation, *J Biol Inorg Chem*, **16**, 125-135.

Macaulay, I.C. and Voet, T. (2014) Single cell genomics: advances and future perspectives, *PLoS Genet.*, **10**, e1004126.

Markowitz, V.M.*, et al.* (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system, *Nucleic Acids Res.*, **40**, D115-122.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet.journal*, **17**, 10--12.

McConnell, M.J.*, et al.* (2013) Mosaic copy number variation in human neurons, *Science*, **342**, 632-637.

McDonald, D.*, et al.* (2012) The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome, *Gigascience*, **1**, 7.

McDonald, D.*, et al.* (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea, *ISME J*, **6**, 610-618.

Mondani, L.*, et al.* (2011) Influence of uranium on bacterial communities: a comparison of natural uranium-rich soils with controls, *PLoS One*, **6**, e25771.

Narasingarao, P.*, et al.* (2012) De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities, *ISME J*, **6**, 81-93.

Nurk, S.*, et al.* (2013) Assembling single-cell genomes and mini-metagenomes from chimeric MDA products, *J. Comput. Biol.*, **20**, 714-737.

Peng, Y.*, et al.* (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, *Bioinformatics*, **28**, 1420-1428.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments, *PLoS One*, **5**, e9490.

Raghunathan, A.*, et al.* (2005) Genomic DNA amplification from a single bacterium, *Appl. Environ. Microbiol.*, **71**, 3342-3347.

Rajendhran, J. and Gunasekaran, P. (2011) Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond, *Microbiol. Res.*, **166**, 99-110.

Richard, H. and Foster, J.W. (2004) Escherichia coli glutamate- and arginine-dependent acid resistance systems increase internal pH and reverse transmembrane potential, *J. Bacteriol.*, **186**, 6032-6041.

Richard, H.T. and Foster, J.W. (2003) Acid resistance in Escherichia coli, *Adv. Appl. Microbiol.*, **52**, 167-186.

Richter, M. and Rossello-Mora, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition, *Proc Natl Acad Sci U S A*, **106**, 19126-19131.

Rinke, C.*, et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter, *Nature*, **499**, 431-437.

Rodionov, D.A., Mironov, A.A. and Gelfand, M.S. (2002) Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea, *Genome Res.*, **12**, 1507-1516.

Rout, M.E. and Callaway, R.M. (2012) Interactions between exotic invasive plants and soil microbes in the rhizosphere suggest that 'everything is not everywhere', *Ann. Bot.*, **110**, 213-222.

Sangwan, P.*, et al.* (2004) Chthoniobacter flavus gen. nov., sp. nov., the first pure-culture representative of subdivision two, Spartobacteria classis nov., of the phylum Verrucomicrobia, *Appl. Environ. Microbiol.*, **70**, 5875-5881.

Seshadri, R.*, et al.* (2003) Complete genome sequence of the Q-fever pathogen Coxiella burnetii, *Proc Natl Acad Sci U S A*, **100**, 5455-5460.

Shakya, M.*, et al.* (2013) A multifactor analysis of fungal and bacterial community structure in the root microbiome of mature *Populus deltoides* trees, *PLoS One*, **8**, e76382.

Shapiro, E., Biezuner, T. and Linnarsson, S. (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science, *Nat. Rev. Genet.*, **14**, 618-630.

Shields, C.W.t., Reyes, C.D. and Lopez, G.P. (2015) Microfluidic cell sorting: a review of the advances in the separation of cells from debulking to rare cell isolation, *Lab Chip*, **15**, 1230-1249.

Smith, K.S. and Ferry, J.G. (2000) Prokaryotic carbonic anhydrases, *FEMS Microbiol. Rev.*, **24**, 335-366.

Smits, S.L.*, et al.* (2014) Assembly of viral genomes from metagenomes, *Front Microbiol*, **5**, 714.

Stepanauskas, R. (2012) Single cell genomics: an individual look at microbes, *Curr. Opin. Microbiol.*, **15**, 613-620.

Stingl, K., Altendorf, K. and Bakker, E.P. (2002) Acid survival of Helicobacter pylori: how does urease activity trigger cytoplasmic pH homeostasis?, *Trends Microbiol.*, **10**, 70-74.

Sung, Y.C. and Fuchs, J.A. (1988) Characterization of the cyn operon in Escherichia coli K12, *J. Biol. Chem.*, **263**, 14769-14775.

Telenius, H.*, et al.* (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer, *Genomics*, **13**, 718-725.

Turnbaugh, P.J.*, et al.* (2007) The human microbiome project, *Nature*, **449**, 804-810.
Turner, T.R., James, E.K. and Poole, P.S. (2013) The plant microbiome, *Genome Biol*, **14**, 209.

Utturkar, S.M.*, et al.* (2014) Evaluation and validation of *de novo* and hybrid assembly techniques to derive high quality genome sequences, *Bioinformatics*.

Wang, Q.*, et al.* (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, *Appl. Environ. Microbiol.*, **73**, 5261-5267.

Ward, N.L.*, et al.* (2009) Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils, *Appl. Environ. Microbiol.*, **75**, 2046-2056.

Werner, J.J.*, et al.* (2012) Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys, *ISME J*, **6**, 94-103.

Wilson, C.M.*, et al.* (2014) Lactobacillus reuteri 100-23 modulates urea hydrolysis in the murine stomach, *Appl. Environ. Microbiol.*, **80**, 6104-6113.

Wu, Y.W., Simmons, B.A. and Singer, S.W. (2015) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets, *Bioinformatics*.

Yoon, H.S.*, et al.* (2011) Single-cell genomics reveals organismal interactions in uncultivated marine protists, *Science*, **332**, 714-717.

Zhang, K.*, et al.* (2006) Sequencing genomes from single cells by polymerase cloning, *Nat. Biotechnol.*, **24**, 680-686.

Zhang, L.*, et al.* (1992) Whole genome amplification from a single cell: implications for genetic analysis, *Proc Natl Acad Sci U S A*, **89**, 5847-5851.

Zhang, Y.*, et al.* (2009) Comparative genomic analyses of nickel, cobalt and vitamin B12 utilization, *BMC Genomics*, **10**, 78.

Appendix

Table 6.1: Post contamination removal assembly statistics for each SAG.

SAG ID	Contigs (> 500 bp)	Total Length (bp)	Average Contig Size (bp)	Maximum Contig Size (bp)	N50 Contig Length (bp)
Acidobacteria	1,434	3,091,572	2,156	67,197	3,598
Armatimonodetes	433	2,392,594	5,526	61,689	10,465
Verrucomicrobia	1,258	3,687,054	2,931	61,282	6,500
Zavarzinella_R9_F7	2,213	7,216,666	3,261	79,486	7,334
Zavarzinella_E9_H3	1,303	6,229,561	4,781	110,550	15,019

Table 6.2: Genome completeness estimation scores for each SAG.

SAGs	Sequence Quality Score	tRNA Score	rRNA Score	Essential score
Zavarzinella_R9_F7	0.25	1.00	0.30	0.71
Zavarzinella_E9_H3	0.32	0.90	0.90	0.67
Armatimonodetes	0.36	0.10	0.70	0.64
Verrucomicrobia	0.23	0.40	0.50	0.57
Acidobacteria	0.18	0.10	0.90	0.39

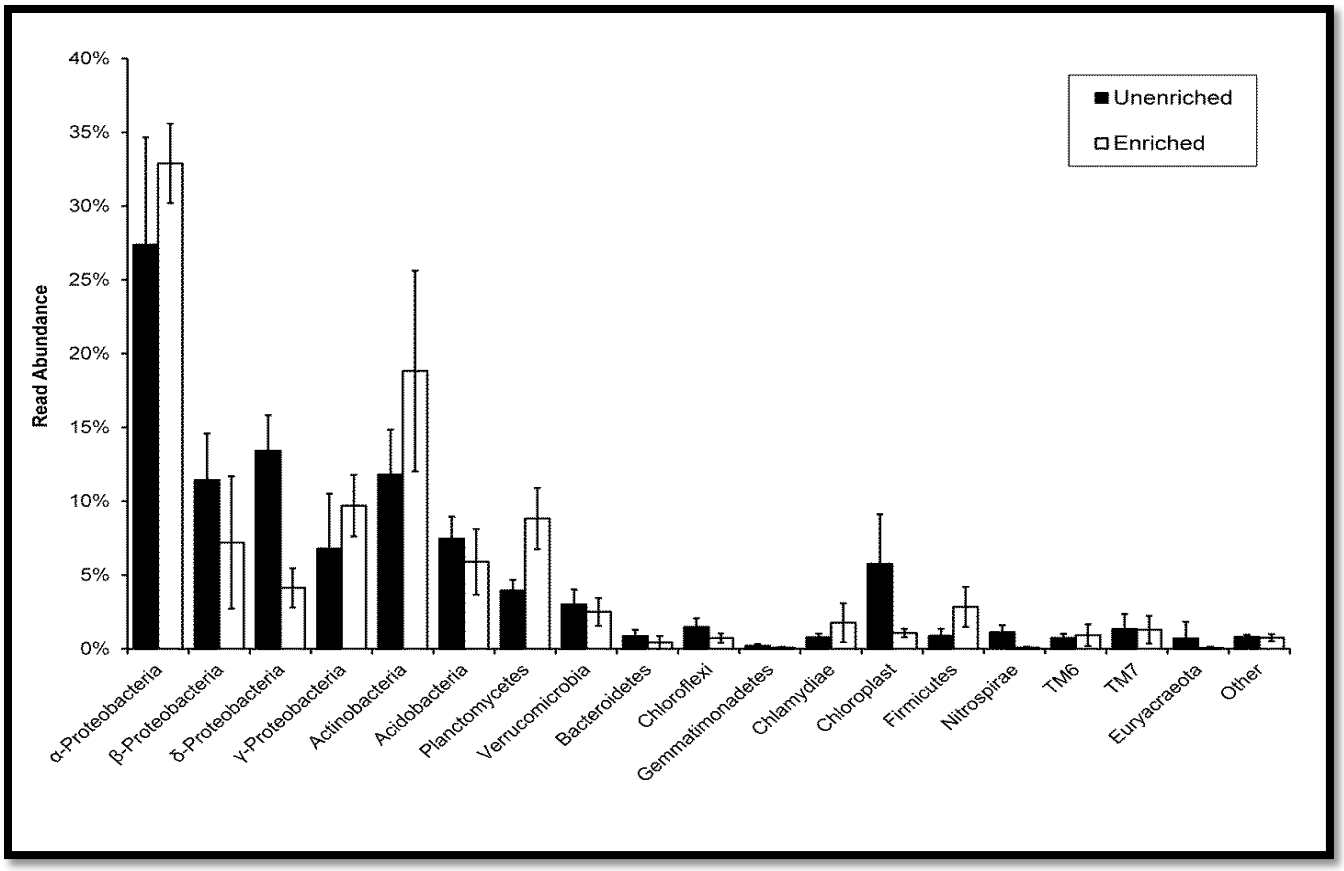

Figure 6.1: Read abundance percentages of enriched and unenriched samples at phylum level.

Dark colored bars indicate unenriched samples while white colored bars indicate enriched samples. Each bar is marked with corresponding error bar.

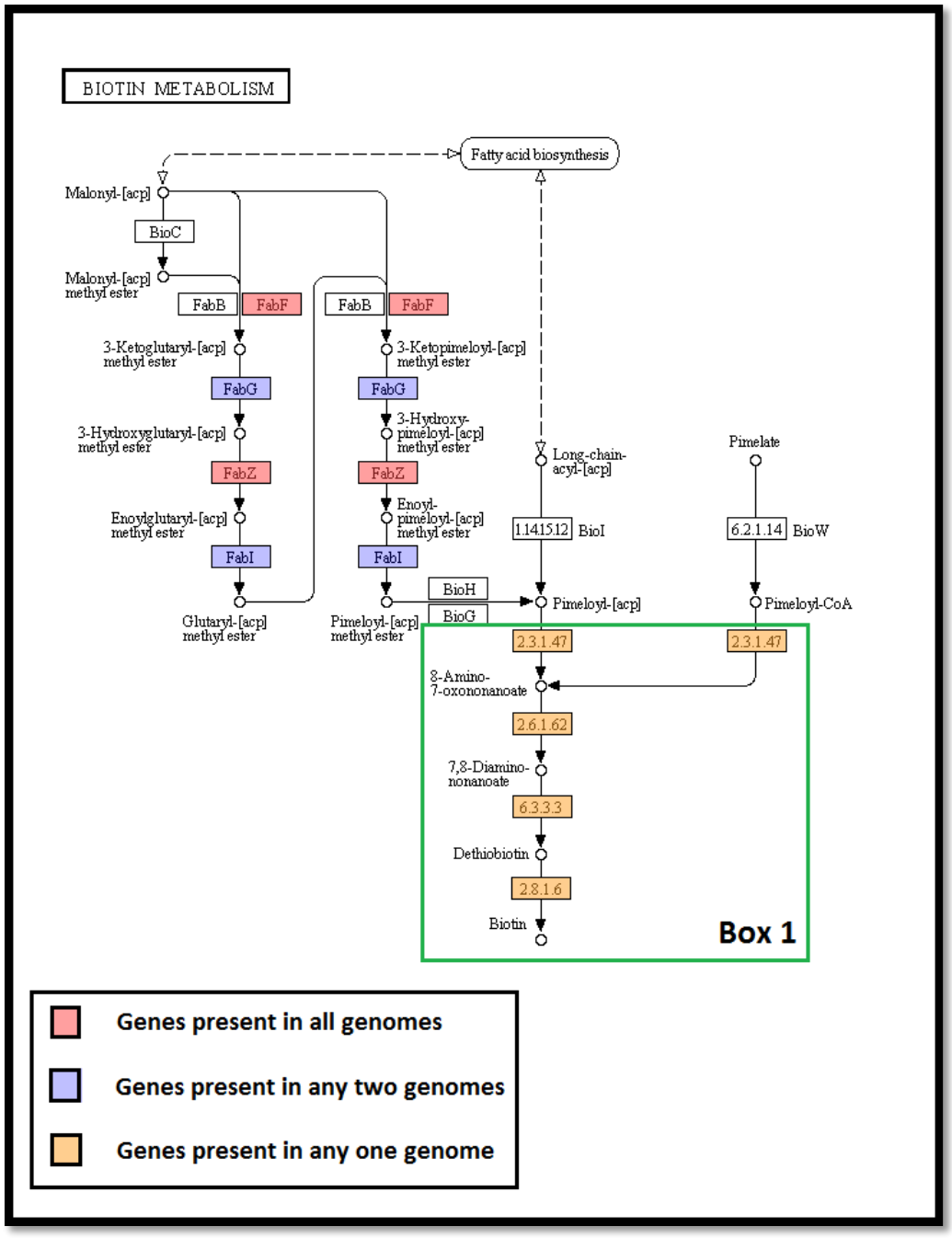

Figure 6.2: Biotin metabolism pathway in Armatimonadetes SAG and corresponding complete genomes of *Fimbriimonas ginsengisoli Gsoil* 348 and *Chthonomonas calidirosea T49*. Box 1 represents the genes present only in Armatimonadetes SAG.

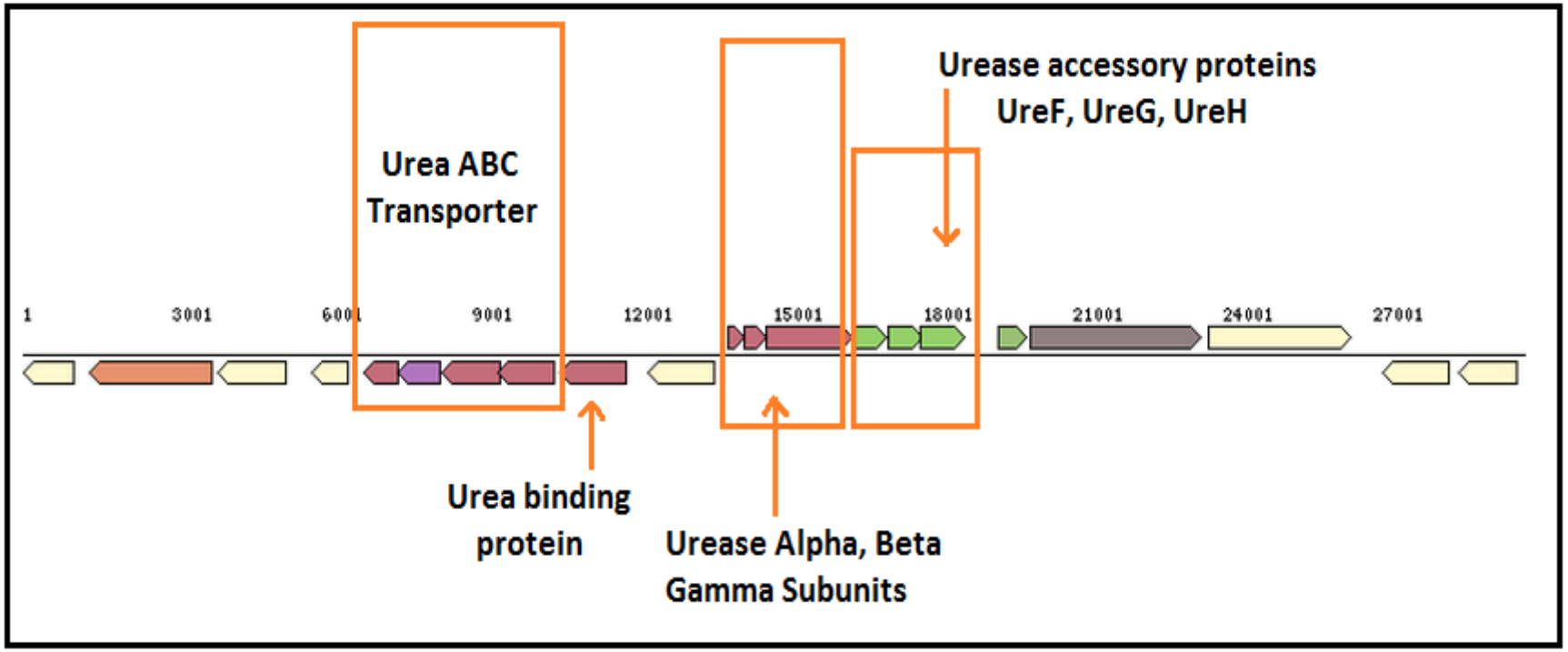

Figure 6.3: Urease gene cluster in Planctomycetes SAG E9_H3.

CHAPTER 7 : CONCLUSION

7.1 Conclusions

This research was undertaken with the goal of improving the genome assemblies of novel microorganisms without any reference sequences to achieve better downstream analysis results. At the same time, the field of NGS was experiencing a rapid change in terms of the emergence of third generation sequencing platforms which offers exceptionally long read-lengths and simultaneous development of new generation of de novo and hybrid assembly algorithms to leverage advantages from each sequencing platform. However, an assessment of the utility of third generation sequencing platforms and systematic comparisons to newer hybrid assembly algorithms was necessary to reveal the reliability and pros/cons of each method.

For the benchmarking purpose, I tested the performance of nine de novo and hybrid algorithms to generate optimal genome assemblies for four novel microorganisms. Our results showed that by using complementary libraries, multiple sequencing platforms, and appropriate assembly algorithms, dramatic improvements could be obtained in overall genome assembly quality of bacterial genomes. Additionally, we proposed the rDNA operons analysis method for assembly validation using PCR and Sanger sequencing approach. Previously proposed in silico assembly evaluation methods were only able to rank the assemblies while rDNA operon analysis provides a measure for accuracy and allows to select optimal assembly. This dissertation research provides in-depth analysis of de novo and hybrid assembly methods, description of assembly protocols, and recommendations for optimal assembly algorithms depending on availability of sequencing data. These protocols should be extendible for others looking to improve existing draft genome assemblies.

Although PacBio sequencing platform was able to obtain substantial improvements in assembly quality, it was criticized for high error rates and use of accessory sequencing data from other platforms was necessary. In later years, PacBio launched upgrades to the sequencing platform and chemistry which obtained substantial improvements in sequencing output and read-lengths. Subsequently, new de novo assembly protocol – HGAP was proposed which relies solely on PacBio data and claimed to overcome the high-error rates with greater sequence coverage. We sequenced the genome of *C. autoethanogenum* using PacBio RS-II and various second-generation sequencing platforms. By application of HGAP protocol, we were able to obtain complete genome sequence for *C. autoethanogenum* using only the PacBio data and without the need for manual finishing. Secondly, the de novo and hybrid assemblies were generated and comparison of the draft and finished assemblies was performed using assembly statistics, and in silico bioinformatics tools. Our results indicated that assemblies based on short read sequencing technologies were confounded by the large repetitive DNA elements, especially rDNA operons. The complete genome sequence also enabled the comparative genomics analysis to reveal the unique feature of this industrially important microorganism which includes the presence of CRISPR system, an additional hydrogenase enzyme, mannose and aromatic substrate utilization pathways and other metabolic differences. This research is one of the first to show the utility of PacBio sequencing platform to obtain automated finishing of microbial genomes and increased use of this technology to obtain finished genome sequences is speculated. We

communicated the public release of sequencing data which span three generations of sequencing technologies, containing six types of data from four NGS platforms and will facilitate the assessment and tool development for current and future NGS technologies. The results above were further supported by subsequent analysis of eight microbial genomes to reveal the unassembled regions within Illumina and PacBio assemblies. Assembly comparison for eight microbial genomes confirmed that rDNA operons are major contributors towards the breakpoints within short-read assemblies. Although a specific common factor or trend linking the PacBio gaps present within different genomes was not determined, it appears to be a cumulative effect of low read-depth, read-quality, and sample variations such as presence of phage DNA or mobile genetic elements. This analysis suggests that the single-molecule sequencing method from PacBio is currently one of the best methods available to obtain finished grade microbial genome assemblies in an automated fashion. However, there are certain difficult genomes which could not be resolved in an automated fashion and manual finishing may be necessary. Some initial bioinformatics steps described as part of our manual finishing approach are easy to apply, extendible and worth trying for any near-finished genome assemblies. This analysis of unassembled DNA regions from Illumina and PacBio assemblies will be useful for sequencing companies and algorithm developers to achieve further technical improvements.

Bacterial endophytes that colonize *Populus* trees play an important role in nutrient acquisition and increased biomass. Endophytes are usually embedded within plant material and physical separation is a difficult task. A differential and density gradient centrifugation based enrichment protocol was developed which reduced contaminating plant DNA and prepared the samples for single-cell genomics analysis. Whole genome sequencing of five selected rare and uncultured bacteria isolated by single-cell genomics was performed. I performed an in-depth bioinformatics analysis of these single-amplified bacteria starting from genome assembly, contamination removal to comparative genomics analysis. I was able to identify the unique characteristics of these uncultured bacteria such as biotin biosynthesis, the presence of urease gene cluster and unique glycoside hydrolase enzymes. This analysis shows that enrichment protocol could achieve a significant reduction in contaminating plant DNA and prepared the samples for further analysis by single-cell genomics or metagenomics to provide the proof of concept. During five years of the Ph.D. program, I have experienced a major leap forward in sequencing technologies from the second generation to third generation platform. During early years of my Ph.D. (2010-2012), sequencing technologies have generally favored low-cost sequencing at the expense of read-length (of few hundred base-pairs). Indeed, low-cost sequencing technologies have accelerated the microbial genomics research and enabled projects on a scale previously unimaginable. Such an example I have seen is the Plant-Microbe Interfaces (PMI) project, where (16S, metagenomics, whole-genome and single-cell) sequencing methods were applied to gain insights into the diversity and functioning of mutually beneficial interactions between plants and microbes in the rhizosphere. The parallel whole-genome sequencing of 42 bacterial isolated from PMI projects using short-read technologies resulted in fragmented genome assemblies. No doubt, these assemblies were useful for several downstream analyses such as genomic characterization, comparative genomics and pan-genome surveys. However, the

accuracy of these genomes could be improved by the inclusion of finished genomes. I have seen a similar scenario in public databases where average quality of genomes was lowered (percentage of finished genomes was below 35% in 2011), which limited quality of the downstream analyses performed. Sequencing with complementary libraries and various platforms, and selection of appropriate assembly software allowed to generate higher quality hybrid assemblies. Hybrid methods obtained dramatic improvements over draft-quality genome assemblies, but the generation of finished genomes was still a tedious process which required several rounds of PCR and manual finishing steps.

In contrast, so-called third-generation sequencing technologies can now produce reads which are tens of kilobases in length. Like a jigsaw puzzle with large pieces, a genome sequence is easier to resolve with longer reads. There were concomitant improvements in the area of bioinformatics to develop cutting-edge assembly algorithms for efficient utilization of long-reads information. We applied the third-generation PacBio technology for whole-genome sequencing of the genome *C. autoethanogenum*, and I found the results were quite astonishing. We were one of the first to obtain complete microbial genome sequence using only the PacBio technology and without the need for manual finishing. These results were important because it eliminates the need for tedious manual finishing process, potentially there is no need for accessory sequencing data when greater than 100x PacBio coverage is available and newer polishing algorithms can generate consensus sequence which is up to 99.9% accurate. The potential of the long read platform was quickly realized with increased number of finished microbial genomes deposited at public databases and higher sequencing cost was justified with lower manual costs and an improved accuracy for downstream applications. There were subsequent updates to sequencing platforms and chemistry, which include more efficient enzymes and reagents to obtain longer reads, and increased throughput with massively parallel sequencing (e.g. new SEQUEL system by PacBio). Alternative sequencing platforms such as Nanopore have the potential to generate uninterrupted read lengths of hundreds of kilobases, but further improvements are necessary for error-handling and data analyses software. The current long-read platforms promise automated finishing for most microbial genomes for under $1000 per genome and anticipated to improve the quality of reference databases and facilitate new studies of chromosomal structure and variation. From my perspective, the sequencing platforms, chemistries, read-length and computational approached will keep improving and assembling contiguous chromosomes (even for small eukaryotic genomes) would be a trivial task in near future.

In summary, the field of NGS technologies have experienced rapid advances in last few years e.g. emergence of third generation sequencing platforms (PacBio and Nanopore), increased read-length and subsequent developments in the field of bioinformatics. The application of these new sequencing platforms is not only limited to de novo genome sequencing but also extended to RNA-sequencing, metagenomics and epigenomics applications. Even greater advances in sequencing are expected in terms of higher-throughput systems which will offer reduced sequencing costs and timelines, generate huge datasets and bioinformatics field will continue to grow with development of new powerful and automated algorithms, data storage solutions and super-computing facilities.

VITA

Sagar Utturkar was born to Vidya Utturkar and Mukund Utturkar in Mumbai, India. He completed Bachelor of Science degree in biotechnology from University of Mumbai and Master of Science degree in bioinformatics from Nottingham Trent University, UK. Later, he worked as a bioinformatics domain expert at Persistent Systems Ltd, India. After 2.5 years of work experience, he decided to come to United States to pursue a PhD degree in bioinformatics and attended Genome Science and Technology Program at the University of Tennessee, Knoxville.