BRADYBASE: A COMPREHENSIVE GENOMICS DATABASE FOR SOYBEAN ROOT-NODULATING *BRADYRHIZOBIUM* SPECIES

by

Menolin Sharma

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics & Computational Biology

Fall 2021

BRADYBASE: A COMPREHENSIVE GENOMICS DATABASE FOR SOYBEAN ROOT-NODULATING *BRADYRHIZOBIUM* SPECIES

by

Menolin Sharma

Approved: _____
Shawn W. Polson, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____
Cathy H. Wu, Ph.D.
Director of the center for Bioinformatics & Computational Biology (CBCB)

Approved: _____
Levi T. Thompson, Ph.D.
Dean of the College of Engineering

Approved: _____
Louis F. Rossi, Ph.D.
Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Shawn W. Polson who made this work possible. I would also like to thank my committee members Dr. K Eric Wommack and Dr. Jeffry J Fuhrmann whose insights and expertise guided me to complete this project and write this thesis. I am grateful to Barbra D Ferrell for her invaluable suggestions and advice through out my research and at the time of writing this thesis.

This project was supported by a grant from the National Science Foundation (1736030). My graduate research assistantship was supported through the University of Delaware College of Agriculture and Natural Resources. The use of the BIOMIX compute cluster was made possible through funding from Delaware INBRE (NIH NIGMS P20 GM103446).

Special thanks to my family and friends for encouraging and helping me during the process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The symbiotic relationship formed between soybean legumes and root-nodulating *Bradyrhizobium* spp. provides a sustainable and affordable source of nitrogen (N) to soybean plants via biological nitrogen fixation (BNF). During this process, *Bradyrhizobium* spp. form and perform nitrogen fixation inside root nodules with the help of nodulation (nod) and nitrogen fixation (nif and fix) genes located together in symbiosis islands in the bacterial genome. Commonly reported soybean root-nodulating *Bradyrhizobium* spp. in the US include *B. diazoefficiens*, *B. elkanii*, and *B. japonicum*, which are also used as commercial inoculants due to observed increases in soybean yield which varies based on phenotypic traits such as symbiotic effectiveness (nodulation and nitrogen fixation abilities) and competitiveness with indigenous *Bradyrhizobium* spp. Genomic analyses, including gene and symbiosis island composition, potentially provide a readily available and cost efficient approach to the prediction of phenotypic traits. However, only 21 complete genomes for these three species are available in NCBI GenBank and RefSeq, and no comprehensive resource is available to gather data about *Bradyrhizobium*, a genus critical to sustainable soybean production that will facilitate food security for a global population expected to reach 9.9 billion by 2050. Bradybase, an organism-specific database for soybean root-nodulating *Bradyrhizobium* spp., was developed as a publicly available web resource for bradyrhizobia researchers and stakeholders. Bradybase integrates genomes and phenomic data from NCBI RefSeq and the University of Delaware *Bradyrhizobium* Culture Collection (UDBCC), which consists of 352 *Bradyrhizobium* accessions (340 field isolates and 12 USDA reference strains) established to genotypically and phenotypically characterize the indigenous soybean root-nodulating *Bradyrhizobium* spp. in the state of Delaware. In this study,

21 UDBCC accessions were selected based on phenotypic diversity and assembled to complete or near complete genomes using long reads generated from Pacific Biosciences (PacBio) RSII Single Molecule Real Time (SMRT) technology. A novel pipeline corrected frameshifted genes, often a result of the high PacBio RSII error rate (13-15%), using reference RefSeq genes. Genome assembly and annotation highlighted the importance of the often overlooked manual assessment and correction of completely assembled microbial genomes, particularly those assembled from PacBio subreads alone, before depositing the genomes into large scale databases like NCBI. Bradybase presents a platform for the integration of tools, analyses, data, and collaboration forums specific to soybean-bradyrhizobia symbiosis research studies benefitting the research and agricultural communities.

Chapter 1

INTRODUCTION

1.1 Soybean

Soybean (*Glycine max L.*) is a leguminous species. It is one of the most important oil and protein sources in the world containing highest protein content among legumes (Liu, 1997). The average nutritional composition for moisture-free seeds is approximately 40% protein, 20% oil, 35% carbohydrate, and 5% ash including minerals, vitamins, and other components (Montgomery, 2003). It has been widely used for human consumption as well as biodiesel, cosmetic products, and feed for livestock, poultry, and fish. Soybean products are used to provide protein diet to humans often replacing animal protein sources. It plays a crucial role in maintaining food and nutritional security. The world population is expected to reach 9.9 billion by 2050 (United Nations, 2019), and a growth in soybean yield and production can help meet the food consumption demand.

1.2 Biological nitrogen-fixation in soybean

The high protein content of soybean demands large amounts of nitrogen (N) for growth and yield (80 kg N per Mg grain produced, (Salvagiotti et al., 2008)). N is required to synthesize enzymes, proteins, and chlorophyll molecules among others. Soybean can meet its N requirements through N fertilizers, indigenous soil resources, or biological nitrogen fixation (BNF) (Salvagiotti et al., 2008). Use of N fertilizers causes water, soil, and air pollution. It is linked to an increased nitrate concentration in drinking and surface waters, deterioration of soil health and structure, negative effects

on soil organisms, emission of gaseous N oxides contributing to global warming, and many other environmental issues (Savci, 2012). Also, application of N fertilizers for soybean production is expensive, requiring 10 billion USD annually (Rodrguez-Navarro et al., 2011). Soil N resources by themselves are available at low concentrations and unable to meet the soybean N demand for good yields.

BNF carried out by symbiotic soybean rhizobia, mainly *Bradyrhizobium* spp., is the most sustainable and cheapest source of N for soybeans (Rodrguez-Navarro et al., 2011). *Bradyrhizobium*-soybean symbiosis is an example of legume-rhizobial symbiosis. Its specificity is controlled by plant secreted flavonoids, nod factor receptors produced by the host plant, rhizobial exopolysaccharides, and host interactions with the bacteria (Wang et al., 2018). Under N-limiting conditions, legume roots secrete flavonoid compounds into the rhizosphere. It activates NodD proteins secreted by the bacteria. NodD proteins activate the expression of nod genes in the bacteria resulting in the release of nod factors. Nod factors bind to host nod factor receptors which result in the formation of infection thread which is converted to N fixing bacteroids in nodules.

Higher BNF is correlated to higher soybean productivity and seed yield (Ciampitti & Salvagiotti, 2018). On average, 50-60% of soybean N demand is met by BNF. Identification and utilization of high nitrogen fixing bradyrhizobial strains can increase BNF potential. Indigenous *Bradyrhizobium* strains are often reported to be either relatively ineffective or not present in sufficient numbers to meet soybean N demand (Chibeba et al., 2017). Host specificity expressed by soybean can also limit the development of soybean-bradyrhizobia symbiosis (Thuita et al., 2012). Commercial inoculants consisting of *Bradyrhizobium* spp. with higher symbiotic efficacy are therefore often used as inoculants in an effort to increase soybean productivity. Soybean response to these inoculants depend on indigenous rhizobial population as well as other soil physico-chemical properties such as temperature, salinity, pH, and N availability. Competition between existing and introduced strains of bradyrhizobia can lower the effectiveness of

these inoculants in increasing soybean productivity (McDermott & Graham, 1990).

1.3 Soybean *Bradyrhizobium*

Bradyrhizobium spp. are slow-growing (doubling time >8 h) Gram-negative Alphaproteobacteria (Jordan, 1982). They are aerobic and non-spore-forming. These short rod-shaped (0.5 to 0.9 μm by 1.2 to 3.0 μm) bacteria are motile by one polar or subpolar flagellum. They possess genomes with high GC content (62-66%) and larger sizes (7-10 Mbp) compared to other members of Nitrobacteraceae family which could be related to their lifestyle and metabolic diversity (Ormeo-Orrillo & Martnez-Romero, 2019). Each strain has 5.2% to 17.8% of the chromosome allocated as the genomic islands (GI) mobilome including symbiosis island. Symbiosis islands contain genes that carry out nodulation and nitrogen fixation in soybean or other host plants.

Independent studies have identified eight different species of *Bradyrhizobium* capable of nodulating soybean: *Bradyrhizobium daqingense*, *B. diazoefficiens*, *B. elkanii*, *B. huanghuaihaiense*, *B. liaoningense*, *B. ottawaense*, *B. yuanmingense*, and *B. diazoefficiens* (Tian et al., 2012). *B. japonicum*, *B. elkanii*, and *B. diazoefficiens* are mostly used to formulate commercial inoculants around the world and also the commonly reported soybean root-nodulating bradyrhizobia species in North America (Padukkage et al., 2021) (Joglekar et al., 2020). The symbiotic activity of bradyrhizobia depends on the symbiosis island present in its chromosome. The symbiosis island carries nod, nif, fix and Type-III secretion system (T3SS) genes which are responsible for nodulation and N fixation and that secrete effector proteins which can regulate symbiotic compatibility with soybean plants respectively (Keyser et al., 1992) (Arashida et al., 2021).

Effective symbiotic N fixation by bradyrhizobia in soybean depends on symbiosis island genes as well as host-strain compatibility, competitive ability of the strains for nodule occupancy, and tolerance to abiotic stresses (Keyser et al., 1992). In a study by Appunu et al. (2008), the symbiotic effectiveness of each of five *B. japonicum*

strains inoculants were found to vary among six soybean cultivars. Nodulation, plant growth, and seed yield were significantly ($P < 0.05$) affected by the host cultivar and inoculation treatments used. Abiotic stresses like salinity, pH, temperature, and nitrate concentration are also important factors contributing to symbiotic effectiveness (Keyser et al., 1992). Soybean plants were found to show deformation of root hairs when the concentration of NaCl in soil increased from 1% to 1.5%, with nodulation eliminated at 1.2% NaCl. Growth and multiplication of a *B. japonicum* strain declined rapidly as NaCl concentration increased from 0.2 to 0.8% (Tu, 1981). Similarly, decreasing pH (4.6 to 4.2) was found to decrease cell growth among five different *B. japonicum* and *B. diazoefficiens* strains. Tolerance to acidity varied among the strains. More tolerant strains were found to be better N fixers during symbiosis with soybeans (Taylor et al., 1990). High temperature too was shown to have a depressive effect in nodulation with some strains failing to nodulate soybean at temperatures >42C (Favre & Eaglesham, 1986). Excess N which can result especially due to high N fertilizer application to soil can also negatively affect BNF. The negative effect was found to occur due to reduced nodule formation, nitrogenase activity, and leghemoglobin concentrations in the presence of high nitrate concentrations (Du et al., 2020). Recent experiments have shown different rhizobia strains to possess genotypes for high/low temperature tolerance, drought-stress tolerance, and nitrate tolerance which could be interesting properties to screen for in bradyrhizobia (Ormeo-Orrillo & Martnez-Romero, 2019) (Rong Li et al., 2020).

Other than symbiotic activities, *Bradyrhizobium* species are also studied for denitrification activities to reduce N2O emissions from soybean fields which are regulated by nap, nir, nor and nos gene clusters (Sameshima-Saito et al., 2006). Hydrogen (H2) uptake (Hup) activity by these species can increase the efficiency of symbiotic N2 fixation and soybean yield due to high energy output from hydrogen oxidation. Different strains have been studied for the hydrogen uptake phenotypes (Hup+ , Hup-, and Hup host-regulated) which are controlled by hup gene clusters (van Berkum, 1990).

Some strains of *B.elkanii* produce a phytotoxin called rhizobitoxine (RT), an enol-ether amino acid (2-amino-4-(2-amino-3-hydropoxy)-trans-but-3-enoic acid). Though it is shown to increase nodulation in some legumes due to its ability to inhibit ethylene biosynthesis, it can cause foliar chlorosis in susceptible soybean cultivars. It is linked to reduced chlorophyll concentrations, shoot and nodule dry weight, leaf protein, and total nitrogen fixation in soybean plants (Robinson et al., 2020).

1.3.1 Genomic knowledge in Soybean *Bradyrhizobium*

Having an extensive genomic resource for soybean *Bradyrhizobium* spp. can accelerate in-depth research and analyses on these species. Whole genomes have been used to generate more resolved phylogeny using Average Nucleotide Identity (ANI) and Average Amino Acid Identity (AAI) analyses across shared genomic regions and protein sequences compared to 16S rRNA gene phylogeny alone (Avontuur et al., 2019). Studies are carried out to identify and analyze functionally important gene clusters including those required for nodulation, nitrogen fixation, photosynthesis, hydrogen uptake, and rhizobitoxine production. These studies provide not only insights on their lifestyles (free-living, photosynthetic, and symbiotic) but also generate hypotheses on evolution of these genes which can occur via vertical inheritance, gene duplication and reduction, or horizontal gene transfer. Similarly, comparative genomics studies among the soybean *Bradyrhizobium* spp. are gaining momentum to further investigate their adaptations to specific environmental conditions, competitiveness with indigenous strains, and nodulation and nitrogen fixation activities. For this purpose, complete genomes are essential as a genome sequence with hundreds of contigs can result in gene fragmentations on the contig boundaries making the genes unidentifiable in the sequence. In addition, complete genomes are necessary to explore chromosomal synteny and structural variants (Siqueira et al., 2014).

The commonly reported species of bradyrhizobia that nodulate different varieties of soybean in the United States are *B. japonicum*, *B. diazoefficiens*, and *B. elkanii*

(Joglekar et al., 2020). Despite the economic and environmental importance of these species, the available genomic knowledge about them is low. As of July 31st 2021, genomes from only 90 different accessions of *B. japonicum*, *B. diazoefficiens*, and *B. elkanii* are assembled and deposited in the Genbank. Most of the assemblies however are at a contig or scaffold level with only 21 accessions reported with complete genomes. Only two of the completely assembled accessions have plasmids.

1.4 Overview of genome sequencing technologies

Genome sequencing technologies have been rapidly evolving with numerous platforms now available that can produce reads of various size distributions (∼150 bp to ∼10 kbp) and accuracies (98.5-99.999%). These can be utilized for sequencing whole genomes of *Bradyrhizobium* spp. Single platform or a combination of sequencing technologies can be chosen based on desired accuracy for assembled genomes, genome lengths and complexities, and available budget for genome sequencing. Following sections provide an overview of available technologies and their applications.

1.4.1 Genome sequencing approaches

Rapid development in sequencing technology has brought down the costs of DNA sequencing with multiple sequencing platforms available today. Genomes have been sequenced using first-generation sequencing, second/next-generation sequencing (NGS), and third generation sequencing (TGS) technologies. First-generation sequencing includes Maxam-Gilbert and Sanger sequencing among which Sanger sequencing is most widely adopted. Sanger sequencing involves chain-termination PCR of template DNA with fluorescently labelled dideoxyribonucleotides (ddNTPs) followed by size separation using gel electrophoresis and fluorometric detection of terminating ddNTPs to construe the read sequence. It provides high accuracy of raw reads (up to 99.999%) but is low-throughput, costly, and produces read lengths of only up to ∼1000bp (Shendure & Ji, 2008). NGS technologies brought revolution in the field of genomics by providing

high-throughput DNA sequencing via massive parallelization. It includes 454, sequencing by oligonucleotide ligation and detection (SOLiD) system, IonTorrent and Illumina sequencing technologies among which Illumina is most sought-after. Illumina can produce raw reads with high accuracy of 98.5-99% which can be increased to >99.9% using base call metrics but produces short read lengths of only 150-300 bp (Slatko et al., 2018). The drawbacks of these short-read lengths include difficulty in resolving structural variants, repetitive elements, and homologous elements which often result in incomplete assemblies with numerous contigs. Also, GC-bias (low coverage of reads in the GC-poor or GC-rich regions) results in fragmented assemblies. TGS technologies including Pacific Biosciences Single Molecule Real Time (PacBio SMRT) and Oxford Nanopore Technologies (ONT) are now routinely used to provide more complete assemblies. PacBio SMRT uses zero-mode waveguide (ZMW) chambers where a template DNA (SMRTbell template) and a DNA polymerase are immobilized at the bottom of a well called zero-mode waveguide (ZMW) in a SMRT flow cell. Fluorescently labelled deoxynucleoside triphosphates (dNTPs) are incorporated during each synthesis reaction. Nucleotide incorporated in the growing chain is identified by sending a light pulse from the bottom of the well which excites the fluorophore of the dNTP during incorporation. PacBio provides two models RSII, and Sequel. Sequel systems provide higher throughput and less cost per base compared to RS II by providing up to 8 million ZMWs compared to 150,000 ZMWs in RS II (Slatko et al., 2018) (Logsdon et al., 2020). ONT reads the disruption in current characteristic to each nucleotide as a DNA strand enters and translocates through a nanopore. Processive enzymes bound to the long dsDNA molecules enable continuous passage through the pore. Due to longer read lengths, they can resolve repeats, and detect structural variants. They are also used for epigenetics and transcriptome sequencing. However, these platforms provide high raw reads error rates (13-15% in PacBio and 15% in ONT) (Tvedte et al., 2020). Recent improvements to PacBio reads include high-fidelity (HiFi) reads which are >99.9% accurate long reads. These are generated using circular consensus sequencing (CCS) mode in PacBio Sequel systems which involves multiple passes around circular template

sequence compared to single pass for commonly used continuous long reads (CLR).

The most successful sequencing platforms in recent years have been Illumina and long read sequencing technologies including SMRT from PacBio and GridION and MinION from ONT (Heather & Chain, 2016). ONT nanopore sequencers have been shown to produce raw data with higher error rate, especially at single nucleotide level, than their PacBio counterparts (Laver et al., 2015) (Weirather et al., 2017) (Lang et al., 2020) producing consensus sequences with lower accuracy than PacBio sequencing. Also, ONT has a more systematic error-profile compared to PacBio reads in which the errors are randomly distributed, which might not be overcome by increasing the coverage alone (Mantere et al., 2019). PacBio sequencing has therefore been used extensively to produce highly contiguous de novo assemblies and detect structural and epigenetic variation among others (Ardui et al., 2018).

1.4.2 Mitigating errors from PacBio sequencing

PacBio subreads contain a high error rate of 13-15% (Ardui et al., 2018). Insertions and deletions (indels) are the predominant errors with more than 90% of them occurring in homopolymer regions (Wenger et al., 2019). Though random distribution of the errors allows one to obtain a highly accurate consensus sequence during genome assembly, remaining indels can introduce frameshifts, result in shortened or extended open reading frames (ORFs), and barriers in single nucleotide analyses including single nucleotide variant (SNV) calling. Different strategies are employed to circumvent the errors: I) self-correction of longer subreads with shorter subreads or overlap information among them, II) correction of long reads with short highly accurate Illumina reads, III) correction and combination of contigs generated by short reads using long reads, and IV) hybrid assembly of genomes using short and long reads (Fu et al., 2019) (Mahmoud et al., 2019). Error correction strategies involving correction of longer reads by shorter overlapping reads are employed by hierarchical genome assembly pipeline (HGAP) built by PacBio (Chin et al., 2013).

1.5 Access to publicly available information about soybean bradyrhizobia

Soybean bradyrhizobia have been widely studied for their geographical distribution, host range, competitiveness, and symbiotic effectiveness. The quantity and availability of genotypic and phenotypic data for soybean bradyrhizobia is increasing. Large scale biological databases like National Center for Biotechnology Integration (NCBI), the European Molecular Biological Laboratory (EMBL), and the DNA Databank of Japan (DDBJ) continue to make the data available to researchers but do not consist of bioinformatic tools, analyses, data mining capabilities and other resources specific to soybean bradyrhizobia, including symbiosis islands, N fixation capacities for different *Bradyrhizobium* strains, and ITS gene phylogeny among others. A database dedicated to soybean root-nodulating bradyrhizobia can serve as a community resource, provide easier access and retrieval to available genotypic and phenotypic data, and encourage collaboration and networking among soybean bradyrhizobia researchers. Following sections provide an overview of biological databases, organism-specific databases, and available construction tools for organism-specific databases.

1.5.1 Biological databases

In response to the plethora of data generated by rapidly developing low-cost DNA sequencing technologies, several online repositories of biological data are now available to store and manage the staggering volume of data. Biological databases play crucial role for biological data analyses and discoveries. They offer biological data including genomic sequence data, gene transcriptome data, genetic variance data and more for a broad range of organisms, along with web services and data analysis tools. The modern leading source for public biological databases, software tools, and research in computational biology, National Center for Biotechnology Information (NCBI) was established in November 1988 with an aim to design, develop, implement, and manage automated systems for the collection, storage, retrieval, analysis, and disseminate human molecular biology, biochemistry and genetics. It stores molecular and

genomic data for both human and non-human species. Data is organized over 40 different integrated databases. GenBank and Reference Sequence (RefSeq) are two widely used nucleic acid repositories. GenBank (http://www.ncbi.nlm.nih.gov) contains all publicly available DNA and protein sequences and is managed by NCBI (Benson et al., 2013) in collaboration with the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ). Reference Sequence (RefSeq), also managed by NCBI, provides only non-redundant curated wild-type sequences derived from GenBank (Pruitt et al., 2012).

1.5.2 Organism-specific databases

The amount of data in the large-scale databases like GenBank is very large containing >1.8 billion sequences with >13 trillion bases as of August 2021 (GenBank and WGS Statistics, 2021) and it is expected to increase exponentially. The sheer volume and complexity of data makes data search, retrieval, aggregation, and visualization a convoluted process especially for an inexperienced user. It decreases the utility of these databases as data cannot be effectively retrieved and analyzed by the researchers. Also, these databases often lack background information on available data (e.g. experimental protocols, and environmental conditions), and specific data analyses and tools for the community. Online community databases are therefore constructed to serve a specific group of researchers working on one or more species. They host genotypic (genomes, annotations, transcriptome), metabolic (metabolic pathways, regulatory networks), and phenotypic (morphology, breeding data) data, along with analyses, tools, and outreach specific to one or a group of related species to serve the community of researchers (Spoor et al., 2019). One of the first community databases, FlyBase (http://flybase.org/) housing genomic and genetic data for *Drosophila melanogaster*, was developed in 1992 by the FlyBase Project (FlyBase Consortium, 1998). Now there are numerous organism-specific databases including the Saccharomyces Genome Database (SGD) for the model species *Saccharomyces cerevisiae*

(Cherry et al., 1998), Genome Database for Rosaceae (GDR) for Rosaceae species (Jung et al., 2008), and Ecocyc for *E. coli* species (Karp et al., 2014).

1.5.3 Generic Model Organism Database

(GMOD) project, Chado and Tripal To meet the increasing demands for organism-specific databases and reduce the time and expenses required to develop database schema, middleware and visualization softwares, Generic Model Organism Database (GMOD) project was initiated in 2000 (Stein et al., 2002). Initiated as a collaboration between four well-established model organism databases: FlyBase (for *Drosophila melanogaster*), SGD (for Saccharomyces cerevisiae), Mouse Genome Database (for *Mus musculus*) (Blake et al., 1999), and WormBase (for *Caenorhabditis elegans*) (Harris et al., 2003) (Stein et al., 2002), the GMOD project has been creating/maintaining several software tools and infrastructures for storage and visualization of biological data. At least 48 tools are listed as GMOD components as of March 30, 2021 (GMOD Components - GMOD, 2021). These tools are necessary to build an organism-specific database.

The data storage infrastructure developed by the project is Chado, a normalized generic relational database schema used by several organism-specific databases which is flexible enough to accommodate different biological data types including genomic features, genomic diversity data, expression data, stocks, genotypes, phenotypes, analyses, projects, literature, experimental protocols, and others. Tables in Chado are divided into different modules which can be utilized as per needs for a specific project and extended or customized if necessary (Mungall et al., 2007). The ability to customize and extend the Chado schema makes it a sustainable database schema for storing constantly evolving biological data types. Several bioinformatic tools from GMOD are compatible with Chado (Spoor et al., 2019).

Tripal is another tool developed by GMOD project publicly released in 2009 with a goal to provide a high-quality modular database infrastructure that can be easily

customized by specific research communities (Sanderson et al., 2013). It integrates with Drupal (http://drupal.org), a popular content management system (CMS) to provide web services and utilizes Chado as underlying database schema. Tripal provides various tools, modules and APIs to access the underlying data, and customize and extend the website by creating extension modules (Spoor et al., 2019). Tripal is widely adopted by different organism-specific databases with Tripal v3.3 reported to be used by at least 31 public sites housing biological data for different organisms as of October, 2020 (Staton et al., 2021).

1.6 Objectives

Soybean-rhizobia symbiosis has been studied for over 100 years with focus on increasing nodulation efficiency and screening highly efficient strains among others (Rong Li et al., 2020). Soybean seeds are inoculated with *Bradyrhizobium* strains to increase N-fixation resulting in higher soybean yield. The potential of an inoculant however depends on its competitiveness with indigenous strains which can limit their N-fixation ability (van Heerwaarden et al., 2018). Success of symbiotic interactions between soybean plants and indigenous or inoculated strains also depends on host specificity expressed by soybean varieties. It is therefore necessary to investigate the genetic diversity, geographical distribution, host compatibility and environmental conditions associated with the localization and dominance of the rhizobial strains in the soil (Shiro et al., 2013).

The University of Delaware *Bradyrhizobium* Culture Collection (UDBCC) was established with an aim to study the diversity of soybean bradyrhizobia in the state of Delaware and its impact in soybean agriculture. It consists of 352 (initially 382) soybean root-nodulating bradyrhizobia accessions which includes 340 isolates collected from 31 soybean farms spread across the state (Figure 1.1) and 12 reference USDA strains (USDA 31, 38, 46, 62, 76, 94, 110, 122, 123, 130, 135, and 138). Field isolates were collected from several soybean cultivars of growth stages V3 to R5.

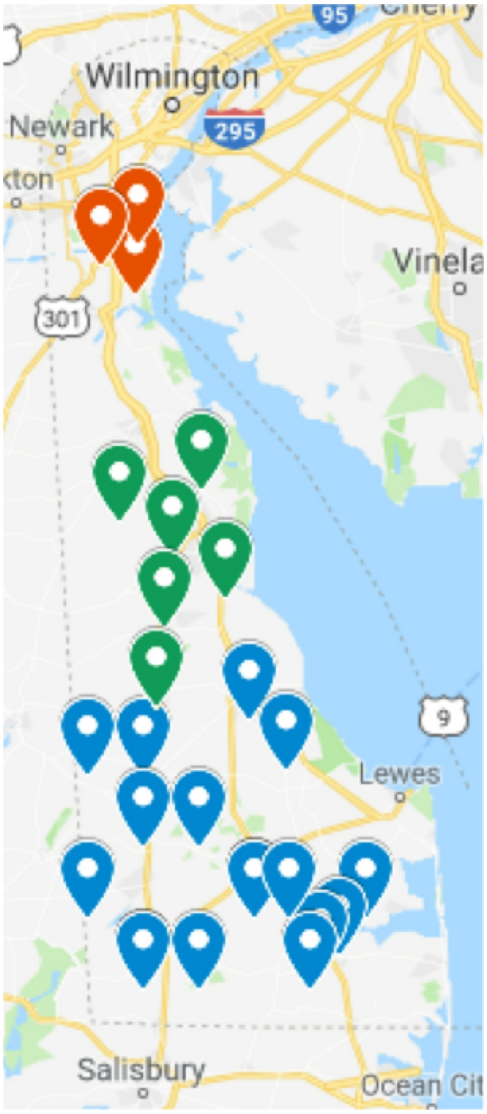


Figure 1.1: Sampling locations for field isolates of soybean root-nodulating *Bradyrhizobium* spp. in the state of Delaware. Colors represent different counties in Delaware. Blue: Sussex county, green: Kent county and red: New Castle county.

This thesis is focused on two aims. First, representative strains of *Bradyrhizobium* accessions from UDBCC are sequenced and assembled to complete or chromosome level assemblies. Second, genomic information available for reported soybean root-nodulating bradyrhizobia species are aggregated from NCBI RefSeq and housed in a database along with genotypic and phenotypic information available for UDBCC accessions. The database has been created using extension and customization of GMOD tools. The work presented herein increases publicly available genomic information about these species and provides a specific database for researchers working in soybean root-nodulating bradyrhizobia. This can advance soybean bradyrhizobia research by allowing better data access, analysis, and visualization among the research community and students.

Chapter 2

GENOME ASSEMBLY AND ANNOTATION

2.1 Abstract

Soybean is a source of oil and protein for humans and animals. Biological nitrogen fixation is a sustainable and environmentally friendly source of nitrogen (N) in these legume species which is carried out by soybean root-nodulating bradyrhizobia via symbiosis. Soybean seeds are inoculated with different strains of bradyrhizobia to increase the yield. University of Delaware *Bradyrhizobium* Culture Collection (UDBCC) contains 352 accessions including 340 soybean *Bradyrhizobium* field isolates collected from the state Delaware and 12 USDA reference strains. Single Molecule Real-Time (SMRT) sequencing using Pacific Biosciences (PacBio) RS II system was used for the genome sequencing of 21 *Bradyrhizobium* accession from UDBCC selected based on phenotypic diversity. PacBio sequencing has 13-15% error rate (80% insertions/deletions mostly in nucleotide homopolymer regions) which can result in frameshift errors in the assembled genomes. It compromises the quality of structural and functional annotation in the genomes, more significantly when the sequencing coverage for the genome is low. In this study, we have assembled genomes, and quantified and corrected putative frame shifted CDS. We have also compared the performance of two genome polishing tools (Arrow and Quiver) provided by PacBio for RS II reads to generate higher quality genomes. It highlighted the importance of the manual assessment and correction of completely assembled microbial genomes, especially those assembled from PacBio subreads alone, before depositing the genomes into large scale databases like NCBI.

2.2 Introduction

Symbiotic nitrogen fixation in soybeans is carried out by soybean root-nodulating *Bradyrhizobium* spp. which can provide 50-60% of soybean N demand on average. The nitrogen fixation efficiency from *Bradyrhizobium*-soybean symbiosis is known to be dependent on *Bradyrhizobium* strain (Rong Li et al., 2020). *Bradyrhizobium* spp.are usually characterized for the presence of symbiosis island and rhizobitoxine genes. Symbiosis island carries nodulation (nod), nitrogen fixation (nif, fix) and type III secretion system (T3SS) genes that carry out nodulation and nitrogen fixation, and secrete effector proteins which can regulate symbiotic compatibility with soybean plants (Keyser et al., 1992) (Arashida et al., 2021). Rhizobitoxine, produced by some species of soybean nodulating bradyrhizobia, has been linked to increased nodulation and foliar chlorosis in soybean plants (Yuhashi et al., 2000).

Frequently reported soybean root-nodulating *Bradyrhizobium* species in North America belong to *Bradyrhizobium japonicum*, *B. elkanii*, and *B. diazoefficiens* (Joglekar et al., 2020). These are also used to formulate commercial inoculants around the world (Padukkage et al., 2021). Symbiotic effectiveness of the inoculants in field soybeans depends on their competitive ability against indigenous *Bradyrhizobium* spp. in soil (McDermott & Graham, 1990). Insights about the availability and activity of the indigenous strains are obtained by conducting genetic diversity and geographical distribution studies on these species which can then help us improve inoculation techniques. In addition, more genomic information on these species along with phenotypic studies about their symbiotic effectiveness can help us establish genome to phenome relationships, and identify more effective inoculants.

To characterize indigenous soybean root-nodulating bradyrhizobia in the state of Delaware and increase genomic and phenotypic information about these species, the University of Delaware established a collection of soybean root-nodulating *Bradyrhizobium* spp. cultures. The University of Delaware *Bradyrhizobium* Culture Collection (UDBCC) consists of 352 *Bradyrhizobium* field isolates collected from 31 different farms

in Delaware (Figure 1.1) and 12 USDA reference strains: USDA 31, 38, 46, 62, 76, 94, 110, 122, 123, 130, 135, and 138 (Joglekar et al., 2020).

2.2.1 Genomic information on soybean root-nodulating bradyrhizobia

Despite the high agronomic importance and expanding research on these species, supportive genomic information is limited. Currently there are 126 different assemblies for 121 accessions of soybean bradyrhizobia species comprising *B. diazoefficiens*, *B. daqingense*, *B. elkanii*, *B. huanghuaihaiense*, *B. liaoningense*, *B. ottawaense*, *B. yuanmingense*, and *B. diazoefficiens* deposited to GenBank as of July 31st 2021. It includes 90 assemblies for the most commonly reported species: *B. diazoefficiens*, *B. elkanii* and *B. japonicum*, out of which only 21 are assembled to complete genomes. Complete genomes of *Bradyrhizobium* are necessary to perform genetic studies, evolutionary studies on rhizobia-host symbiosis, and analyses including comparative genomics, characterization of symbiosis islands, chromosomal synteny, and structural variants identification (Siqueira et al., 2014).

2.2.2 Single Molecule Real-Time sequencing for obtaining complete genomes

The Pacific Biosciences (PacBio) Single Molecule Real-Time (SMRT) sequencer is a single-molecule, long-read sequencing platform that enables the assembly of more complete genomes than traditional Sanger and next-generation technologies due to longer read lengths which can resolve complex repeats during an assembly (Adewale, 2020). The PacBio reads are however known to have a high error rate of 13-15% which can result in low accuracy of the assembled genomes (Liao et al., 2015). Insertions and deletions are the predominant errors with more than 90% of them occurring in homopolymer regions (Wenger et al., 2019). These insertions/deletions (indels) can introduce frameshifts, resulting in shortened or extended open reading frames (ORFs) and false-positive variant calls that alter the predicted identity and fidelity of proteins during annotation of the genome.

The objective of this study was to increase genomic knowledge on indigenous soybean-nodulating *Bradyrhizobium* species in the state of Delaware and reference USDA strains by sequencing and assembling UDBCC accessions selected based on phenotypic (serology, Fatty Acid Methyl Esters (FAME) analysis, and spontaneous production of virus-like particles (VLPs)) and genotypic (Internal Transcribed Sequence - Restriction Fragment Length Polymorphism (ITS-RFLP), 16S rRNA sequencing, Internal Transcribed Sequence (ITS) region (between the 16S rRNA and 23S rRNA genes) sequencing analyses. We utilized long-reads from PacBio SMRT sequencing to obtain complete genomes. We compared two genome polishing tools for their performance in producing higher quality genomes. We also analyzed the final de-novo assembled genomes for any presence of frameshifts introduced by the sequencing platform.

2.3 Methods

2.3.1 Selection of UDBCC accessions

Twenty-one accessions of bradyrhizobia from UDBCC were selected for genome sequencing. The accessions were selected to represent a broad representation of the collection based on genotypic and phenotypic analyses performed on each accession as described above. Field isolates were named with a letter representing the county name for the collection site in Delaware (New Castle [N], Kent [K], or Sussex [S]) followed by two digit code (01-12) denoting farm in the county and a letter (A-L) for each isolate from the farm. Accessions/isolates were termed with a suffix attached to their names according to their corresponding species determined as described by Joglekar et al. (2020). Bd specifies *Bradyrhizobium diazoefficiens*, Be specifies *Bradyrhizobium elkanii*, and Bj specifies *Bradyrhizobium japonicum*. The selected accessions included four USDA reference strains (USDA 31-Be, USDA 94-Be, USDA 135-Bj, USDA 123-Bj) and 17 soybean root-nodule isolates (Table A.1). Accessions USDA122-Bd, USDA76-Be, S06B-Bj and S10J-Bj were also chosen for sequencing as well in a different genomic analysis study in University of Delaware (Joglekar, 2021).

2.3.2 Isolation and sequencing of bradyrhizobia DNA

Bradyrhizobia cultures in 25% glycerol were stored at -80 C. Bacterial cultures were streaked on Modified Arabinose Gluconate (MAG) agar plates, and individual colonies were selected and used to inoculate MAG broth cultures grown at 30 C with shaking at 250 rpm. DNA was isolated and purified from 5-day old bradyrhizobia cultures using the All Prep PowerViral DNA/RNA isolation kit (Qiagen, Germantown, MD) following the manufacturers instructions. The quality and quantity of isolated DNA was determined using a Qubit fluorometer. Isolated DNA was run through 8% agarose gel and imaged to confirm the extraction. A total of 5-10 μg of the DNA was used to construct 20 kb SMRT-bell sequencing libraries and sequenced using the PacBio RS II sequencer at the University of Delaware Sequencing and Genotyping Center (UDSGC, Newark, DE).

2.3.3 Assembly of bradyrhizobia genomes

PacBio RS II subreads from each sample library were de novo assembled implementing Hierarchical Genome Assembly Process (HGAP) with HGAP3 in SMRT Analysis v2.3.0 (SMRT Analysis Release Notes (v2.3.0), 2015). The following parameters were used in HGAP3: minimum sub-read length: 1000 bp, minimum polymerase read length: 1000 bp, minimum polymerase read quality: 0.85, minimum seed read length: 10000 bp, genome size: 9.5 Mb, seed coverage: 25. The remaining parameters were set to default. For accessions with >15 HGAP3 contigs, an alternate assembly was performed using HGAP4 in SMRT Link v7.0.1 (SMRT Link Software Installation (v7.0.1), 2019) using these parameters: minimum sub-read length: variable (500-1000 bp), minimum polymerase read length: 1000 bp, minimum polymerase read quality: 0.85, minimum seed read length: variable (5000-10000 bp), genome size: 9.5 Mb, seed coverage: variable (25-30), aggressive option: true, and FALCON fcg overrides: pa_dbsplit_option = -x500 -s200, with all other parameters set to default values.

2.3.4 Identification of misassembled contigs

In each initial assembly with more than one contig, BLASTN (Camacho et al., 2009) analysis of each contig against remaining contigs was done to identify if it was misassembled or repeated. Contigs with more than 80% identity over 80% query length to a larger contig were identified as repeated contigs and removed. Misassembled contigs were identified when a contig produced BLASTN hits across different regions of a larger contig with each hit having >80% identity and a total query coverage of >80%. Such contig(s) were also removed.

2.3.5 Identification of putative contaminated contigs

All contigs were annotated with Prokka v1.14.6 (Seemann et al., 2014). Putative chromosomes were identified from lengths comparable to bradyrhizobia chromosomes (6.1-11.7 Mb) (Ormeo-Orrillo & Martnez-Romero, 2019). Putative plasmids were determined based on the presence of RepABC operons. Remaining contigs which did not belong to any of the putative chromosomes, plasmids, and repeated or mis-assembled ones categories were run through a BLASTN analysis against genomes from other isolates. Those producing best hits with >80% query coverage and >80% identity against other genomes were further analysed to identify if their origin could be attributed to cross-contamination from other accessions.

Due to suspected cross-contamination of accessions S07J-Be and S13E-Bd, the internal-transcribed spacer (ITS) regions from each accessions assembly was BLASTNed against the ITS sequences from the remaining UDBCC accessions. The contamination level of the S07J-Be and S13E-Bd genomes was assessed using CheckM v1.1.2 (Parks et al., 2015). Contigs from S07J-Be and S13E-Bd were then aligned (using progressiveMauve algorithm from Mauve v2.3.1) (Darling et al., 2010) and subjected to a BLASTN query against all assembled genomes to observe their similarity and positional homologies with other genomes. Contigs showing positional homology and BLASTN similarity (>80% identity over >80% query coverage) to genomes from other

accessions were identified as of putative cross contaminating origin. These contigs along with putative plasmids identified based on the presence of RepABC operon were removed from further analysis in the two isolates.

2.3.6 Genome circularization and polishing

All genomes were then circularized. For HGAP3 assembled genomes, presence of repeat regions at the ends of the contigs was observed using Gepard v1.40 dotplots (Krumsiek et al., 2007). A self-BLAST analysis of the contig against itself was performed to identify the repeated regions at each end. One of the repeated ends of the contig was then trimmed off. Genomes assembled using HGAP4 were circularized by the assembly pipeline itself. To ensure circularity of the resulting contig, bridgemapping analysis in SMRT Analysis v2.3.0 using RS.BridgeMapper.1 was performed for all genomes.

Two independent genome polishing analyses, Quiver and Arrow, were performed on each genome by implementing the PacBio variantCaller tool (PacificBiosciences/GenomicConsensus, 2012/2021). PacBio subreads were repeatedly realigned against the assembled genome using Blasr v1 or Pbmm2 v1 with minimum concordance threshold of 70%, minimum subread length: 1000 bp, minimum polymerase read length: 1000 bp, and other default values for each algorithm. Consensus and variant sequences were called to reach a PacBio concordance value of >99.999% (QV50) for all genomes in each of the analyses.

2.3.7 Genome completeness and analysis of missing single copy genes

Completeness and contamination of the polished genomes were estimated using CheckM. Truncated single copy genes not reported by CheckM were analysed for the presence of any indel in the gene sequences of both Quiver and Arrow polished genomes from K07G-Bd and K02K-Be accessions. Selected genomes represented genomes with low completeness among the Quiver and Arrow polished genomes. The Quiver polished

K07G-Bd genome showed only 91.5% completeness, less than the completeness level of 95% required to be a high quality reference genome (Parks et al., 2015), and the arrow polished K02K-Be genome had reported the lowest completeness (98.39%) among the Arrow polished genomes. Phylogenetic lineage conserved single copy genes reported as missing by CheckM from the assembled genomes were identified. Reference genes from *Bradyrhizobium* spp.for the identified genes were downloaded from the RefSeq database on November 1st, 2020. A nucleotide BLAST run of the reference genes against the assembled genome was carried out and the aligned nucleotides were compared in case of a positive hit to observe any insertions or deletions in the missing gene. A manual correction of the insertion/deletion in the homopolymeric region was done using the reference gene as a guide for the alteration. Corrected genomes were assessed again for completeness using CheckM.

2.3.8 Identification and correction of putative frameshifted ORFs

Local/in-house reference genes and protein databases were created by collecting all *Bradyrhizobium* spp.genes and proteins respectively from RefSeq on November 1st, 2020. A BLASTN run of each predicted ORF was performed against the reference genes database with a percentage identity cutoff of 80%. Differences between lengths of each predicted ORF in the genome and reference gene were calculated. Predicted ORFs with a length difference of more than 10% to the reference gene length were collected. Adjacently positioned ORFs having the same annotations were identified among the collected ORFs. A BLASTX analysis of the identified ORFs against the local protein database was carried out to identify if they were single genes split into two or more ORFs. They were binned as split ORFs if they matched adjacent regions of the same reference protein. A reciprocal BLAST run of all ORFs having a length difference of more than 10% to the reference gene length was performed to ensure any split ORF was reported only once. All reciprocal BLAST hits were identified as putative frameshifted genes.

An in-house script was used to automate reference gene-based correction of putative frameshifted ORFs in the genome (Figure 2.1). A BLASTN analysis of each predicted CDS against the local database containing all *Bradyrhizobium* genes in RefSeq was done with a BLASThomology cutoff of 80%. Length of the query CDS was compared to the reference gene length to identify only the query CDS with more than 10% length difference as putative framseshifted CDS. Any truncated CDS was extended along its truncated region using nucleotides from the genome sequence followed by a BLAST analysis against the same reference gene. BLASTalignment between the ORF and best hit reference gene was analyzed. Insertion or deletion in the predicted ORF in comparison to the reference gene in any homopolymeric region of either predicted ORF or reference gene was identified and replaced by a nucleotide in the equivalent position in the reference gene. Absence of indels in homopolymeric regions indicated the frameshift to be not an effect of insertion or deletions in the homopolymeric regions of the coding region and were not advanced through the pipeline. ORF was identified again in the altered putative frameshifted ORF using the getorf tool from EMBOSS v6.6. A BLAST analysis of the new ORF was run against the same reference gene and lengths were compared to ensure that the change in nucleotide restored the ORF length.

A resequencing analysis was performed using the corrected genome as the reference genome. PacBio subreads were aligned against the genome using Pbmm2 v1 and variant sequences were called using Arrow to avoid incorporating any insertions or deletions introduced into the genome via reference gene alignments but not present in the long reads. This was done with the assumption that any insertion/deletion present in the genome but not identified by previous rounds of resequencing will be retained due to better alignment of the PacBio subreads against the altered genome. Only the insertions and deletions still retained after resequencing analysis were permanently incorporated into the genome.

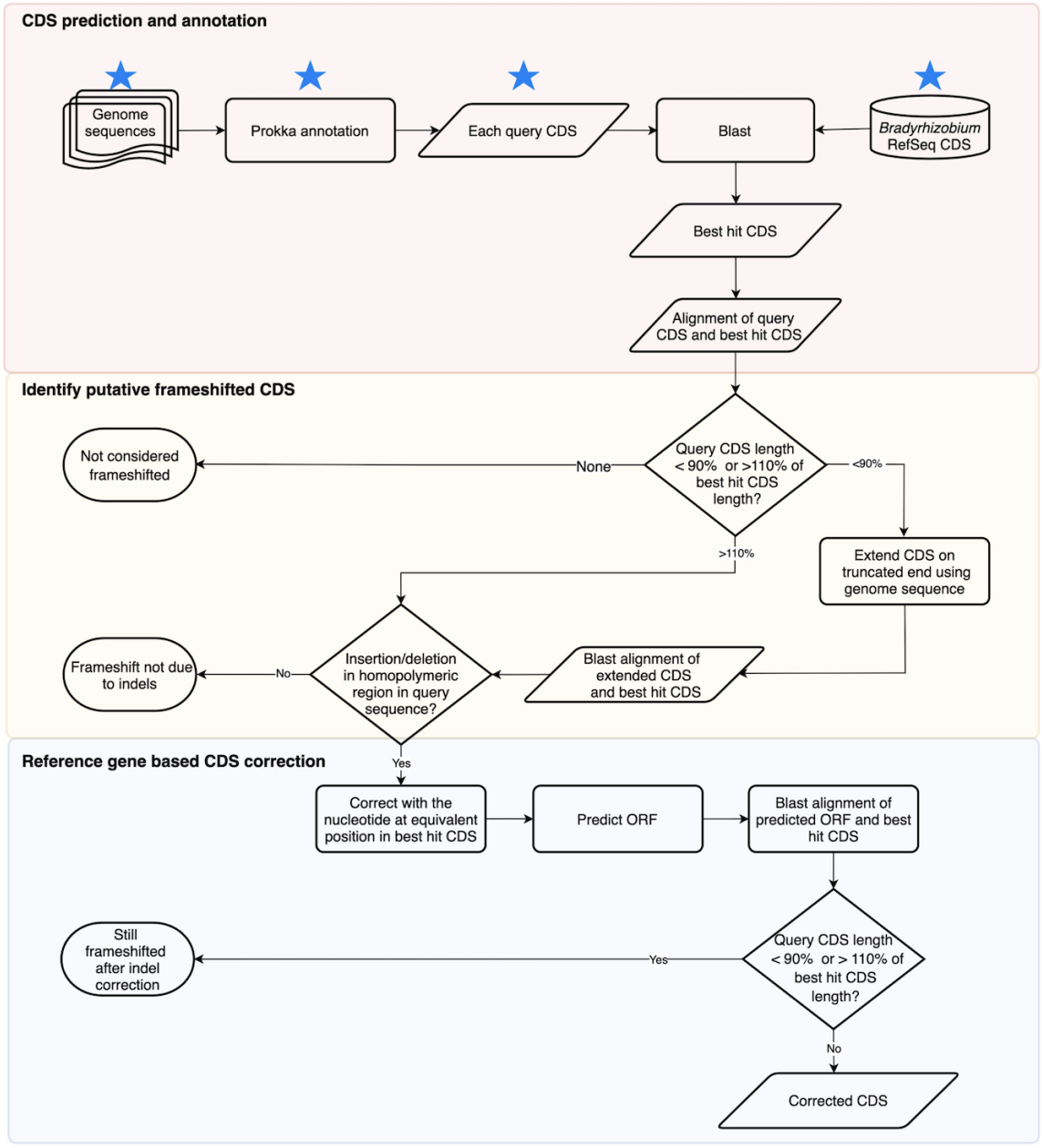


Figure 2.1: Flowchart for identification of putative frameshifted CDS in genome and their correction using *Bradyrhizobium* reference CDS collected from RefSeq in March 2021. A BLAST analysis of each predicted CDS is performed against the *Bradyrhizobium* RefSeq CDS database to get best hit reference CDS. Predicted CDS is classified as putative frameshifted if its length differs by more than 10% of the reference CDS length. Presence of insertion or deletion is identified in their BLAST alignment and correction is guided by the reference CDS sequence. Steps marked with blue stars are the processes carried out to generate inputs for the algorithm.

2.3.9 Comparison of genome polishing tools

Performance of the genome polishing tools Quiver and Arrow in terms of the quality of the genomes produced were compared using genome completeness and percentage of putative frameshifted CDS per total predicted CDS in each genome as quality metrics.

To investigate how completeness and putative frameshifted CDS in genomes could be affected by the subreads parameters in each genome polishing tool, polymerase read quality, mean subreads mapping concordance, and mean subreads fold coverage from the resequencing pipelines used to obtain final polished genomes as well as total predicted CDS, completeness, and percentage of putative frameshifted CDS were collected from each polished genome. Correlation analysis was performed between each pair of these variables and their differences obtained after using Arrow and Quiver genome polishing tools. Pearson correlation coefficient was calculated for each pair using R v3.6.2 and visualized using Corrplot v0.88. Dependent variables with positive/negative correlation coefficients at $p < 0.05$ to independent variables were selected for simple linear regression analyses. Also combinations of two or more independent variables with significant correlation coefficients were selected for multiple linear regression analyses. Variables having simple or multiple linear regression coefficients with statistically significant p-value (<0.05) were noted.

2.3.10 Genome annotation and mobilome analysis

Arrow polished bradyrhizobia genomes were annotated using Prokka v1.14.6 (Seemann, 2014). The parameters used were: Kingdom: Bacteria, Genus: *Bradyrhizobium*, and Genetic code/Translation table: 11 (Archaea, most Bacteria, most Virii, and some Mitochondria) in addition to default settings. A fasta file, containing all proteins collected from RefSeq in November 1st, 2020 belonging to the genus *Bradyrhizobium*, was created and provided as a supplementary fasta file to annotate the proteins in addition to core databases used by Prokka: ISfinder, NCBI Bacterial Anti Microbial

Resistance Reference Gene Database, and UniProtKB (SwissProt), and hmm database: HAMAP.

Assembled genomes were also uploaded to RAST v2 (Rapid Annotation using Subsystem Technology) server (Overbeek et al., 2014) and annotated using following parameters: genetic code: 11 (Archaea, most Bacteria, most Virii, and some Mitochondria); annotation scheme: RASTtk; and enabled fixframeshits and automatic error correction options. Putative symbiosis islands in the genomes were identified with IslandViewer 4 (Bertelli et al., 2017).

Plasmids were identified based on the presence of the repABC operon, encoding the repA, repB and repC genes responsible for plasmid segregation and replication (Cevallos et al., 2008); the tra operon required for conjugation; and the par operon required for plasmid partitioning. Assembled bacteriophages were identified using phage prediction tools PHASTER (PHAge Search Tool Enhanced Release) (Arndt et al., 2016) and PhiSpy (Akhter et al., 2012).

2.4 Results

2.4.1 Genome assembly

Bradyrhizobia genomes were assembled with either HGAP3 or HGAP4 with 48-157X coverage with total genome size 8.5 - 11.3 Mbp. For the isolates N03G-Bd, S13E-Bd, S14C-Bd, K03D-Be, S05J-Be, S07J-Be, S06K-Bj, S11L-Bj, and USDA reference strains USDA31-Be, and USDA135-Bd, HGAP3 produced more than 20 contigs. The HGAP4 pipeline was used to obtain <6 circular contigs in these isolates except for S07J-Be and S13E-Bd isolates which produced 14 and 11 contigs respectively.

2.4.2 Identification of misassembled and contaminating contigs

Following Prokka annotation, all contigs could be identified as either putative chromosome or plasmids for all isolates except S07J-Be and S13E-Bd.

HGAP4 assembly of S07J-Be produced 14 total contigs. One circular 9.5 Mb contig represented the bacterial chromosome, and three other circular contigs of lengths 152, 222, and 295 kb were determined to be plasmids based on the presence of the repABC operon. Two rrn operons were identified, one in the circular 9.5 Mb contig and other in the linear 102 kb contig. BLASTN analysis of ITS regions from both rrn operons against the collection of ITS sequences from UDBCC accessions and NCBI nr nucleotide database was performed. ITS region in 9.5 Mb contig showed up to 100% identity with 100% query coverage to the ITS sequences from *B. elkanii* accessions while ITS region from 102 kb contig shared 100% query coverage and up to 99.75% identity with *B. diazoefficiens* and only <84% identity with *B. elkanii* accessions. S07J-Be contigs other than chromosome and plasmids were BLASTNed against contigs from all assembled accessions. They shared 99.9% identity with 100% query coverage to the S13E-Bd chromosome but <99% identity with only <5% query coverage with S07J-Be contigs. Similarly, mauve alignment of S07J-Be contigs against genomes from other accessions showed the circular 9.5 Mb contig shared similarity and positional homology to chromosomes from *B. elkanii* isolates, contigs identified as plasmids showed similarity to circular S13E-Bd contigs, and remaining contigs shared homology to chromosomes from *B. diazoefficiens* isolates. Since S07J-Be is an isolate of *B. elkanii* species, the sequenced sample could have been contaminated with S13E-Bd or any other genome from *B. diazoefficiens* species. CheckM analysis of all assembled contigs showed a completeness score of 100% with 91.67% contamination and 75.93% strain heterogeneity, which means 91.67% of single copy gene sets were found twice and belonged to different strains. However, the completeness was 100% with contamination being only 0.3% while analyzing only the circular 9.5 Mb contig which further supported our assumptions about the other contigs originating due to cross-containations.

Similarly, HGAP4 assembly of S13E-Bd produced 11 contigs including one 9.4 Mb circular contig identified as putative chromosome, five 152 - 484 kb circular contigs

identified as putative plasmids and five remaining contigs of sizes 44 kb - 5.4 Mb. Two rrn operons were identified, one in the circular 9.4 Mb contig and other in the linear 5.4 Mb contig. BLASTN analysis of ITS regions from both rrn operons against the collection of ITS sequences from UDBCC accessions and NCBI nr nucleotide database was performed. ITS region in 9.4 Mb contig showed up to 99.2% identity with 100% query coverage to the ITS sequences from *B. diazoefficiens* accessions. ITS region from 5.4 Mb contig shared up to 94.7% identity with 100% query coverage to the ITS sequences from *B. elkanii* but <90% identity with those from *B. diazoefficiens* accessions. Contigs that were neither putative chromosomes nor plasmids shared >98% identity with >98% query coverage to the chromosome from USDA31-Be isolate but <5% query coverage with S13E-Bd contigs during BLASTN analyses against contigs from all assembled accessions. Similarly, Mauve alignment of the S13E-Bd contigs against genomes from other isolates showed that the circular 9.4 Mb contig shared similarity and positional homology to chromosomes from *B. diazoefficiens* isolates, and contigs other than putative plasmids showed similarity to chromosomes from *B. elkanii* isolates. CheckM produced 94% completeness, 45.5% contamination and 34.38% strains heterogeneity during the analysis with all 11 contigs. It resulted in 94.7% completeness and 0.3% contamination during analysis with only the circular 9.4 Mb contig. All these results suggested a probable contamination by an accession from *B. elkanii* species.

Accessions S13E-Bd, S07J-Be, USDA 31-Be, USDA 94-Be, S05J-Be and K03D-Be went through DNA extraction at the same time for PacBio sequencing. It is probable for S07J-Be to be cross-contaminated with S13E-Bd and S13E-Bd with USDA 31-Be accession during the handling process. This supports our observations of contaminations in these accessions. Though new sequencing experiments under isolated culture conditions are required to obtain better assemblies and get further insights into the possible contamination, those identified to be of putative contamination origin as well as plasmids were removed from these assemblies as the source of plasmids was uncertain between the assembled and contaminating accessions. Only the circular contigs

containing dnaA gene with sizes equivalent to bradyrhizobia genomes were retained for further genomic analyses for isolates S07J-Be and S13E-Be.

2.4.3 Genome circularization and polishing

Chromosomes and plasmids from accessions N03G-Bd, S13E-Bd, S14C-Bd, K03D-Be, S05J-Be, S07J-Be, S06K-Bj, S11L-Bj and USDA reference strains USDA 135-Bj and USDA 31-Be were circularized by the HGAP4 assembly pipeline in SMRT analysis v7.0.1 during the assembly itself. Contigs from remaining HGAP3 assemblies required manual circularization. Dotplots from Gepard confirmed the presence of repeated ends in accessions K01E-Bd, K07G-Bd, K09F-Bd, N03B-Bd, K02K-Be, K03I-Be, S15H-Be, S04E-Bj, S15A-Bj, and USDA reference strains USDA 94-Be and USDA 123-Bj. One of the repeated ends was trimmed off after a self-BLAST analysis of each contig. RS.BridgeMapper.1 analysis in SMRT Analysis v2.3.0 showed all the resulting contigs to be circular.

Contigs consisting of genomes from 19 accessions and chromosomes from S07J-Be and S13E-Bd accessions were polished with Quiver to obtain 100% concordance value for all accessions except S13E-Bd, K07G-Bd, K03I-Be, and USDA31-Be whose concordance values fluctuated between 99.9998-99.9997% after each iteration of resequencing. All assembled contigs reached a concordance value of 100% when polished using Arrow.

2.4.4 Assessment of genome completeness

CheckM assessment of genome completeness and contamination for Quiver polished contigs showed most of the contigs to be 98-100% complete with <1.5% contamination except accessions S13E-Bd, K07G-Bd, K03I-Be, and USDA reference strain USDA31-Be having completeness 89-96%. Completeness scores of >98% for all genomes were achieved after genome polishing with Arrow (Figure 2.2).

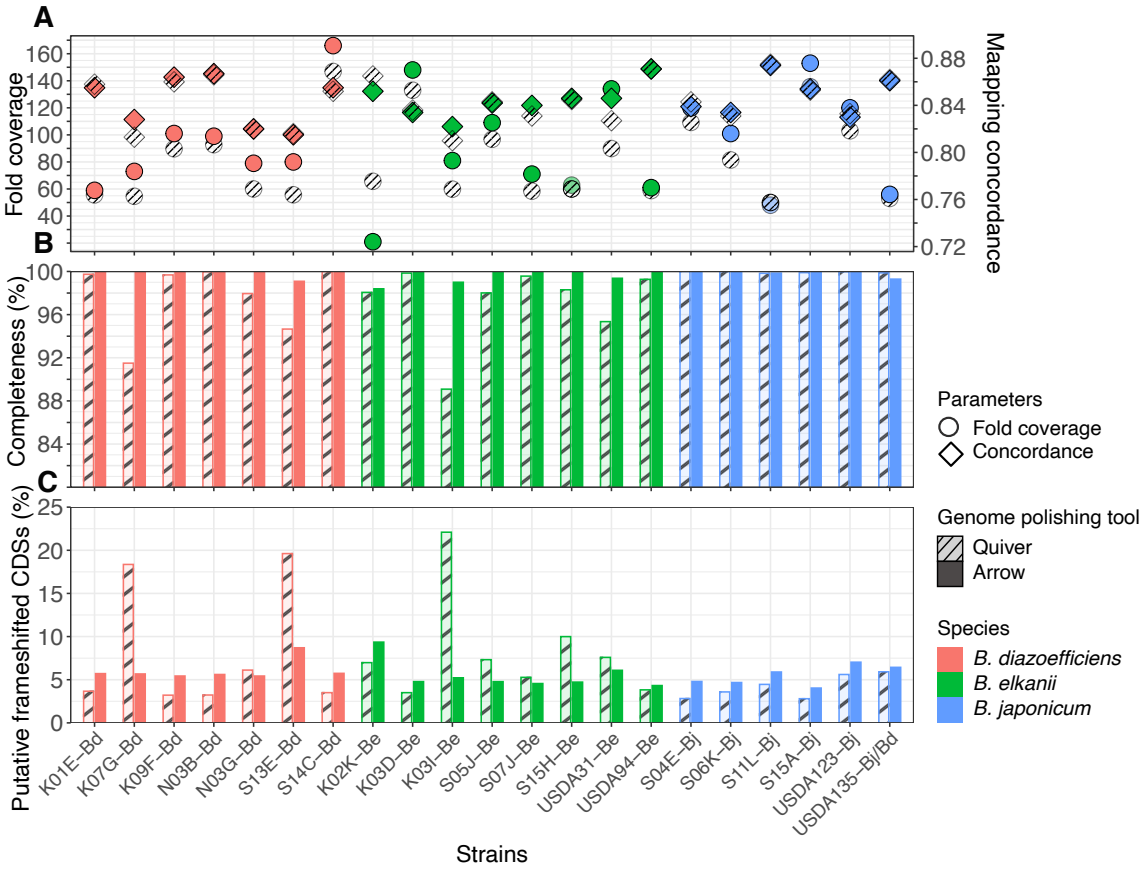


Figure 2.2: Genome analyses performed on de-novo assembled genomes. A. PacBio subreads coverage and mapping concordance for genomes from each strain. B. Completeness for each genome as measured by checkM. C. Putative frameshifted CDSs in percentage for genome from each strain. Using Arrow decreased the number of putative frameshifted CDSs significantly in isolates K07G-Bd, S13E-Bd and K03I-Be whose completeness had considerably increased after Arrow polishing compared to Quiver as shown in B. However, Arrow increased the number of putative frameshifted CDSs in isolates having comparable completeness scores using Quiver and Arrow genome polishing tools.

2.4.5 Identification of missing single copy genes

Following Quiver resequencing of the genomes, K07G-Bd genome reporting 91.5% completeness was assessed for missing phylogenetically conserved single copy genes. BLASTanalysis of reference genes for the 49 missing genes against the genomes showed all 49 of them to have gene sequences present in the genome with an insertion or deletion in homopolymeric regions. Manual correction of the homopolymeric regions decreased the number of missing genes to zero. For the Arrow polished K07G-Bd genome, only three single copy genes were reported to be missing. Two of the three genes were present during the assessment of completeness in the genome from the same isolate after Quiver polishing. All three of them were present in the genome with deletions in one or more homopolymeric regions within the coding region of the gene.

For the K02K-Be genome, 13 and 14 single copy genes were reported as missing by CheckM following genome polishing using Quiver and Arrow respectively. Genes missing in the Quiver polished genome were reported to be present in the Arrow polished genome and vice-versa. Ten out of 13 missing genes from the Arrow polished genome were present in the Quiver polished genome, and 11 out of 14 missing genes from the Quiver polished genome were present in the Arrow polished genome. All the genes could be identified in the genomes with one or more insertions/deletions in homopolymeric regions.

2.4.6 Analysis of putative frameshifted ORFs

All genomes were analyzed for the presence of frameshift errors due to indels in all predicted genes including the single copy genes. A three-case scenario was assumed. A gene could either be truncated, extended, or split into two or more fragments annotated as separate ORFs. Genes were identified as putatively frameshifted when their length differed by >10% compared to its reference gene obtained from RefSeq.

BLASTN analysis of all predicted ORFs from each genome against reference CDS sequences from RefSeq showed the number of total putative frameshifted genes to be 5-38% of total predicted ORFs after Quiver polishing. Following Arrow polishing, the number decreased to 4-13% in each genome (Figure 2.2).

2.4.7 Comparison of Genome polishing tools

The number of predicted CDS in each genome differs according to the genome polishing tools used. Percentage of putative frameshift genes from Quiver is correlated to mean subreads coverage and mean subreads mapping concordance reported by the pipeline with an adjusted-R2 value of 0.5 and p-value of 0.00081. Percentage of putative frameshift genes from Arrow however does not show any correlation to either mean subreads coverage or mean subreads mapping concordance within a p-value of 0.05. The percentage of putative frameshifted genes is negatively correlated to the completeness scores obtained from CheckM in both genome polishing tools Quiver and Arrow with correlation coefficients of -0.93 and -0.74 respectively with p-value <0.05. Mean subreads mapping concordance reported by each tool is positively correlated to polymerase read quality with a correlation value of 0.7 with p-value <0.05.

2.4.8 Reference gene based correction of ORFs

Due to higher completeness and lower percentages of putative frameshifted genes reported after genome polishing by Arrow compared to Quiver, further analyses were carried out in Arrow polished genomes. Putative frameshifted genes after Arrow polishing were corrected for any single nucleotide indels using a reference gene alignment based approach. 15-50% of the putative frameshifted genes showed insertions or deletions in a homopolymeric region when compared to reference genes (Table 2.1).

Table 2.1: Chromosome and plasmid sizes distribution in de-novo assembled genomes of *Bradyrhizobium* spp.

Species	Strains	Predicted CDS	Number of putative truncated/extended CDS	Percentage of putative truncated/extended CDS corrected using reference genes
B. diazoefficiens	K01E-Bd	8590	598	15.2
	K07G-Bd	9062	699	31.4
	K09F-Bd	8698	568	16.5
	N03B-Bd	8748	615	15.2
	N03G-Bd	9366	679	31.3
	S13E-Bd[1]	9229	1343	37.0
	S14C-Bd	8526	602	19.9
B. elkanii	K02K-Be	9194	1309	50.2
	K03D-Be	8892	493	18.0
	K03I-Be	8709	634	29.8
	S05J-Be	8650	513	18.5
	S07J-Be[1]	9126	567	24.1
	S15H-Be	9273	605	27.6
	USDA31-Be	9678	863	39.5
	USDA94-Be	8849	478	19.4
B. japonicum	S04E-Bj	9102	560	21.2
	S06K-Bj	9077	544	19.6
	S11L-Bj	9403	665	29.1
	S15A-Bj	7867	404	21.7
	USDA123-Bj	10642	1144	36.5
	USDA135-Bj	8760	1264	34.8

[1] Isolate putatively cross-contaminated during sequencing from which only full length chromosome was filtered out after removing contigs showing highest ITS sequence identity (>99%) and/or contig sequence identity (>98%) and homology to genomes from other species.

2.4.9 Genome annotation and mobilome analysis

The Arrow-polished genomes were annotated using Prokka and RAST. The percentage of ORFs annotated as hypothetical proteins decreased from 50-60% to 20-30% when using Prokka with an additional fasta file containing bradyrhizobia RefSeq proteins as an annotation reference. Number of insertion sequences varied from 0.5 to 15%. All the genomes had symbiosis islands with putative lengths 570-920 kb. The number of plasmids ranged from 0 to 4 in each isolate (Table 2.2), with sizes ranging from 81-641 kb. A circular bacteriophage of size 31 kb was assembled in USDA31-Be isolate as determined by both PHASTER and PhiSpy v4.2.15. Number of plasmids was unknown in S07J-Be and S13E-Bd due to putative contamination in the PacBio subreads obtained for the accessions.

2.5 Discussion

2.5.1 Long read sequencing can produce complete to near complete assemblies

De novo sequencing of bacterial genomes is important to gain scientific insights on genotype and phenotype relationships of an organism. Many sequencing platforms are available for high-throughput sequencing of bacterial genomes with continuous advancements in short read Illumina technology and long reads sequencing technologies such as PacBio SMRT and Oxford Nanopore Technology (ONT). Illumina sequencing produces high fidelity reads with low sequencing costs. De novo assemblies using short reads however are usually incomplete, low quality, and fragmented due to inability of the short reads to resolve repeat regions. Long repetitive regions including transposons, rRNA operons and hypothetical proteins have been linked to fragmented assemblies produced by Illumina assembly (Utturkar et al., 2017). Soybean bradyrhizobia genomes are found to possess multiple copies of each family of insertion sequences (∼26 total families), especially in highly reiterated sequence-possessing (HRS) strains (Iida et al., 2015) (Siguier et al., 2014), and some strains are reported to possess multiple

Table 2.2: Distribution of predicted and putative truncated/extended Coding Sequences (CDS) in each assembled UDBCC accession

Species	Strains	Assembly pipeline	Chromosome Length (Mb)	Plasmid size(s) (kb)
B. diazoefficiens	K01E-Bd	HGAP3	9.18	ND[1]
	K07G-Bd	HGAP3	9.50	296
	K09F-Bd	HGAP3	9.35	ND
	N03B-Bd	HGAP3	9.32	ND
	N03G-Bd	HGAP4	9.76	165 / 289
	S13E-Bd	HGAP4	9.39	Unknown[2]
	S14C-Bd	HGAP4	9.06	ND
B. elkanii	K02K-Be	HGAP3	9.38	ND
	K03D-Be	HGAP4	9.52	311 / 128
	K03I-Be	HGAP3	9.24	ND
	S05J-Be	HGAP4	9.27	374
	S07J-Be	HGAP4	9.56	Unknown
	S15H-Be	HGAP3	9.71	282
	USDA31-Be	HGAP4	10.00	295
	USDA94-Be	HGAP3	9.50	227
B. japonicum	S04E-Bj	HGAP3	9.71	ND
	S06K-Bj	HGAP4	9.67	ND
	S11L-Bj	HGAP4	9.77	641 / 266 / 161 / 132
	S15A-Bj	HGAP3	8.51	ND
	USDA123-Bj	HGAP3	10.89	187 / 123 / 81
	USDA135-Bj	HGAP4	8.86	3

[1] ND not determined

[2] Isolate putatively cross-contaminated during sequencing from which only full length chromosome was filtered out after removing contigs showing highest ITS sequence identity (>99%) and/or contig sequence identity (>98%) and homology to genomes from other species.

rrn operons with some having identical operons (Joglekar et al., 2020) which can create problems during short reads assembly. Long-read sequencing platform, due to its longer read length can resolve such repeat regions and produce more complete assemblies (Molina-Mora et al., 2020). Fifteen of the 21 complete assemblies from soybean-nodulating bradyrhizobia species in GenBank have used long-read sequencing technology while all the 69 fragmented assemblies involve the use of Illumina or 454 or other short-read sequencing technologies. We adopted SMRT sequencing by PacBio RS-II system to sequence the genomes from 21 accessions. We could assemble 19 accessions to complete genomes and two accessions to chromosome level.

2.5.2 Genome polishing tool can affect the quality of assembled genomes

SMRT sequencing by PacBio can yield reads of lengths of a few kilobases to more than a megabase with an average error rate of 13-15% in raw reads (Ardui et al., 2018). Such a high error rate can compromise the quality of consensus sequences assembled from these reads. Ensuring the quality of genomes is important for accurate downstream analyses on the genomes including comparative genomics, evolutionary studies and genome to phenome relationship inspections. Common strategies used to decrease error rates are either genome sequencing to achieve deep coverage by long reads or use of shorter Illumina reads for error correction of long reads (Zimin & Salzberg, 2020). Uniform sequence coverage however cannot be guaranteed and genome regions with low coverage might be prone to base-call errors despite high overall coverage. Also, both high coverage long read sequencing and additional Illumina sequencing can result in high per-genome costs and time for analysis.

Several long-read error correction tools including FLAS (Bao et al., 2019), and Long Read Multiple Aligner (LoRMA) (Salmela et al., 2017) are available that utilize overlapping information between long reads to self-correct the reads whose performances depend highly on sequencing depths (Zhang et al., 2020). Zhang et al. (2020) reported the significant decrease in genome fraction covered by corrected reads from

99.6% to 9.6% as sequencing depth decreased from 90x to 30x. FLAS also showed a loss in genome coverage with decreased sequencing depth though to a lesser extent. Hence, we used HGAP which has integrated self-correction of long reads, genome assembly, and genome polishing steps to provide more complete assemblies.

After genome assembly, genome resequencing and variant calling are additional crucial steps to decrease errors in the genome introduced by sequencing technologies. Error-corrected long reads are realigned to the consensus sequence with the rationale that previously unidentified sequencing errors might be identified and corrected. The process is iterated until no variants are called after the resequencing step. PacBio offers two genomic consensus or genome polishing algorithms for RS-II data with P6-C4 chemistry: Quiver and Arrow. Quiver utilizes raw pulse and base-call information obtained during SMRT sequencing to generate probabilities for true incorporations or spurious base calls using a model generated by PacBio in-house training for a particular SMRT sequencing chemistry (Chin et al., 2013). A maximum-likelihood consensus sequence is identified using a conditional random field approach. Arrow, on the other hand, uses hidden markov model (HMM) using the base caller quality value metric, per-read SNR (signal to noise ratio) and per-base pulse width metric to calculate likelihood parameters for the model. Quiver is being phased out and replaced by Arrow which does not require training and is easier to develop (PacificBiosciences/GenomicConsensus, 2012/2021). Quality of genomes produced after genome polishing measured as genome completeness score differed among the two tools which could be attributed to different algorithms used by these tools. Our results showed significant increase (5%) of completeness score in three genomes, a slight increase (1%) in 14 genomes, same score in four genomes, and a slight decrease (1%) in one genome after polishing the assembled genomes with Arrow compared to Quiver.

2.5.3 Performance of Quiver shows correlation to fold coverage and mapping concordance while Arrow does not

Quiver uses a conditional random field approach with parameters derived from an in-house training of SMRT sequencing data on a known template (Chin et al., 2013). Arrow uses a hidden Markov model with parameters adjusted according to the fixed covariates taken from each ZMW which makes it sensitive to the differences in the SMRT Sequencing process for each molecule (Hepler et al., 2016). Genome quality after genome polishing with Quiver was positively correlated to subreads mapping concordance and fold coverage while Arrow did not show any correlation.

Fold coverage and mapping concordance of subreads to the reference genome can be crucial in obtaining a higher quality genome. Mean fold coverage measures the mean depth of coverage for each base in the reference genome. Mean mapping concordance measures the mean agreement between subreads and reference genome against which subreads are realigned. Low concordance can decrease the accuracy of consensus base calls. It can be attributed to poor sample, instrument issues, consumable qualities, overloading of the sample or other errors during sequencing (Guide - Step-By-Step Run Performance Evaluation, 2020). A higher sequencing depth can help in averaging out the errors in each read producing more accurate consensus calls.

The importance of coverage in the quality of the assembled genome polished by Arrow genome polishing tool is highlighted by Ou et al. (2020) in which they showed size of contigs and gene space completeness of maize NC358 were positively correlated to sequencing depth. The improvement in the completeness score was however only minimal at a sequencing depth higher than 30x coverage. All of the UDBCC accessions sequenced in the work presented here had coverages ranging from 48-147x. This could be one reason why performance of Arrow did not show any correlation to fold coverage.

2.5.4 Genomes require manual inspection for residual sequencing errors before depositing into the biological databases

Developments in sequencing technologies have given rise to a plethora of microbial genomes in databases. Many deposited genomes, however, represent fragmented genomes and unresolved plasmids which might also contain misassemblies as quality assessments are not available for most of the deposited genomes. Some others, even though assembled to a complete genome using long reads, have not been followed up with manual inspections often resulting in a larger number of pseudogenes caused from frameshifts presumably originating in sequencing errors in the reads used (Smits, 2019).

PacBio subreads have a high error rate (13-15%) (Ardui et al., 2018). Insertions and deletions are the predominant errors with more than 90% of them occurring in homopolymer regions (Wenger et al., 2019). These indels can introduce frameshifts and premature stop codons in predicted open reading frames (ORFs). Utturkar et al. (2017) implemented additional rounds of corrections for four microbial genomes assembled from PacBio long reads and polished using Quiver with short reads from Illumina technology using Pilon which is a microbial genome polishing tool utilizing short reads (Walker et al., 2014). Three hundred fourteen modifications were suggested by Pilon with most of the base-call errors being insertions and deletions. About 85% of those corrections were found to be valid when checked with PCR and Sanger sequencing of 47 randomly chosen Pilon-corrected regions across different genomes. However, seven of those corrections were ruled-out due to no support from sanger sequencing. Similarly, Watson & Warr (2019) investigated the presence of indel errors in human genome assembled by Koren et al. (2017) using long reads generated by PacBio SMRT sequencing (P6-C4 chemistry) followed by multiple rounds of Quiver as performed in the work presented here. They aligned a sample of 40,949 transcripts downloaded from Ensembl against the genome and observed 845 protein coding transcripts to be disrupted by indel errors. These studies demonstrate that sequencing errors in PacBio long reads can introduce frameshift errors in predicted genes which can critically affect

the interpretation of translated regions.

In this work of assembly of 21 accessions of *Bradyrhizobium* spp., the percentage of putative frameshifted genes was found to between 5.1 and 14.5% in the genomes assembled with HGAP 3/4 and polished using Arrow. Since proteins with length cutoffs of 10% from either end usually retain function, only the genes with length difference of more than 10% with reference genes were identified as putative frameshifted (Lerat & Ochman, 2005). De Maio et al. (2019) performed similar analysis for the presence of artificially shortened proteins due to indels from long-read sequencing technologies by identifying proteins of length <90% of reference proteins length in Enterobacteriaceae genomes. Most of the putative frameshifted genes were truncated as observed in this study. When aligned with reference genes, 15 to 50% of the putative frameshifted genes (212 ORFs per genome) had insertions or deletions in homopolymer regions.

Manual refinements using reference genes have been carried out in some cases of genome assemblies. However, the refinements do not try to distinguish whether the errors observed are from sequencing technology or are truly pseudogenes. In this work, we applied a filtering step adding additional resequencing of the error corrected genome using Arrow to avoid falsely correcting pseudogenes present in the genome due to evolutionary reasons. The best method to check the validity of reference gene based error correction is however only via PCR and Sanger sequencing of the error-corrected regions.

2.5.5 Assembly of *Bradyrhizobium* spp.to complete and near complete genomes increased available genomic information

Research in soybean-bradyrhizobia symbiosis has been carried out for 100 years with a focus in several aspects of the symbiosis including host specificity of the *Bradyrhizobium* spp., screening for strains with high symbiotic nitrogen fixation efficiency and stress tolerance (Rong Li et al., 2020). Completely assembled genomes of

Bradyrhizobium spp.are indispensable to gain scientific insights on genotype and phenotype relationships of the organism. *Bradyrhizobium* spp.carry symbiosis islands in the chromosome which carry nodulation (nod) and nitrogen fixation (nif and fix) genes that determine symbiotic effectiveness (nodulation and nitrogen fixation activities) of the plant (Ormeo-Orrillo & Martnez-Romero, 2019).

As of July 31st 2021, there were 126 different assemblies for soybean bradyrhizobia in RefSeq. It included 90 assemblies for the most commonly reported species also used as commercial inoculants (*B. diazoefficiens*, *B. elkanii* and *B. japonicum*), 21 of which are reported as complete genomes with two accessions having 1 and 4 plasmids respectively. Incorporating the results of the assemblies reported here will increase the number of *B. diazoefficiens*, *B. elkanii* and *B. japonicum* accessions with complete genomes to 42, including 12 accessions with plasmids. The assembly also resulted in the assembly of a circular phage element infecting USDA31-Be isolate. Genomic characterization of the phage element can provide us more insights about *Bradyrhizobium* phages which can affect the N fixation ability of the host strain as well as community dynamics and evolution of soybean bradyrhizobia. Seventeen of the 21 assembled genomes are field isolates constituting indigenous soybean root-nodulating *Bradyrhizobium* spp.communities of Delaware. These were chosen based on different genotypic and phenotypic analyses to represent the broad community of the indigenous strains. Indigenous strains are known to limit the symbiotic effectiveness of commercial *Bradyrhizobium* spp.inoculants during soybean farming. This massive increase in the genomic repertoire of indigenous soybean bradyrhizobia can aid in optimizing biological nitrogen fixation and increasing soybean yields in the soybean farms of Delaware.

Nineteen of the genomes are assembled to complete genomes and two genomes to chromosomal level. These genomes can be utilized for comparative genomic analyses. A large number (77%) of RefSeq available genomes for commercial soybean inoculant species consist of hundreds of contigs of sizes ranging from few kb to Mb. These contigs can have fragmented genes which might sometimes be important for metabolism or

lifestyle of the organism. It limits the scope and accuracy of comparative genomic studies. Increasing the number of complete genomes will increase our understanding about the species.

In addition to symbiotic activities, *Bradyrhizobium* species are also studied for denitrification activities to reduce N_2O emissions from soybean fields which are regulated by nap, nir, nor and nos gene clusters (Sameshima-Saito et al., 2006). Hydrogen (H_2) oxidation activity by these species can increase the efficiency of symbiotic N_2 fixation and soybean yield due to high energy output from hydrogen oxidation. Different strains have been studied for the hydrogen uptake phenotypes (Hup+ , Hup-, and Hup host-regulated) which are controlled by hup gene clusters (van Berkum, 1990). Assembled genomes can be analyzed for these features as well to gain more insights on genome to phenome relationships.

Chapter 3

BRADYBASE

3.1 Abstract

Soybean root-nodulating *Bradyrhizobium* carry out biological nitrogen fixation (BNF) in soybeans which can meet 50-60% of soybean nitrogen (N) demand. They are also used as soybean field innoculants to increase soybean growth and yield. Soybean-bradyrhizobia symbiosis has been studied for more than a 100 years with research data available on symbiotic effectiveness, competitiveness, and host compatibility. Recent developments in sequencing technologies including next generation and third generation sequencing have increased the amount of genomic, transcriptomic, and genetic data. The deluge of data in current large scale databases makes storage and access of *Bradyrhizobium* specific genes, genomes, and other genotypic and phenotypic traits space and time consuming from existing databases. Also, they lack *Bradyrhizobium* spp. specific phenotypic and genotypic traits, analysis tools and results. An online community database dedicated to soybean root-nodulating bradyrhizobia species can benefit the community of researchers working in these species. In this work, we have developed Bradybase which presents a platform for the integration of tools, analyses, data, and collaboration forums specific to soybean-bradyrhizobia symbiosis research studies benefitting the research and agricultural communities.

3.2 Introduction

Soybean is one of the most important crops in the world, mostly used as an oil and protein source. Soybean seeds have 40% protein and 20% oil content (Montgomery, 2003). Biological nitrogen fixation (BNF) plays a significant role

to provide high demand of N by the crop by meeting 50-60% of soybean N demand (Rodrguez-Navarro et al., 2011). Symbiotic soybean rhizobia, mainly *Bradyrhizobium* spp., carry out BNF in soybean plants which is the most sustainable and cheapest source of N for soybeans. Studies have shown strains from eight different *Bradyrhizobium* species: *Bradyrhizobium diazoefficiens*, *B. daqingense*, *B. elkanii*, *B. huanghuaihaiense*, *B. liaoningense*, *B. ottawaense*, *B. yuanmingense*, and *B. diazoefficiens* (Jaiswal & Dakora, 2019) (Zhang et al., 2014) to nodulate soybean. *B. japonicum*, *B. elkanii*, and *B. diazoefficiens* are mostly used to formulate commercial inoculants around the world and the commonly reported soybean-nodulating bradyrhizobia species in North America (Padukkage et al., 2021) (Joglekar et al., 2020). Symbiotic nitrogen fixation in soybean by bradyrhizobia is an active area of research aimed towards reducing pollution from chemical N fertilizers by promoting affordable and sustainable soybean production using *Bradyrhizobium* inoculants for BNF (Gitonga et al., 2021). It involves selecting strains with high nitrogen fixation efficiency (Hungria & Mendes, 2015), increasing symbiotic nitrogen fixation efficiency of existing strains (Rong Li et al., 2020), and understanding their genetic diversity and geographical distribution (Shiro et al., 2013).

Recent developments in sequencing technologies including next generation and third generation sequencing have increased the amount of genomic, transcriptomic, and genetic data with more than 31 petabytes of DNA sequence deposited into the NCBI Sequence Read Archive (SRA) in the last decade (Spoor et al., 2020). The deluge of data makes just accessing and storing genes and genomes from desired species a time and space consuming task, let alone the manual processing, and transforming data from one tool to another which can be infused with human error. Online community databases are therefore developed to host genotypic (genomes, annotations, transcriptomes), metabolic (metabolic pathways, regulatory networks), and phenotypic (morphology, greenhouse experiments) data, along with analyses, tools, and outreach specific to one or a group of related species to serve the community of researchers

working on the species. *Saccharomyces* Genome Database (SGD) for *Saccharomyces cerevisiae* (Cherry et al., 1998), Genome Database for Rosaceae (GDR) for Rosaceae family (Jung et al., 2008), and Ecocyc for *E. coli* species (Karp et al., 2014) are some of the popular community databases. Many online community databases adopt Tripal, a free and open-source toolkit built by the Genome Model Organism Database (GMOD) project, for the construction of such online community databases (Spoor et al., 2019). Tripal uses the Chado schema to store data and integrates with Drupal (http://drupal.org), a popular content management system (CMS) which allows the integration of a variety of data analysis tools. Online community database for soybean root-nodulating bradyrhizobia A database dedicated to soybean root-nodulating bradyrhizobia species can benefit the community of researchers working in these species. As of July 31st 2021, 126 genome assemblies were hosted in NCBI along with their genes and annotations for *Bradyrhizobium* species: *B. diazoefficiens*, *B. daqingense*, *B. elkanii*, *B. huanghuaihaiense*, *B. liaoningense*, *B. ottawaense*, *B. yuanmingense*, and *B. diazoefficiens*. While genes and genome assemblies are available in large scale databases like National Center for Biotechnology Information (NCBI), the European Molecular Biological Laboratory (EMBL), and the DNA Databank of Japan (DDBJ), they do not provide background information, experimental protocols, and phenotypic data for the *Bradyrhizobium* spp. Even for the available information including taxonomy, genome assemblies, genes, proteins and annotations, aggregating and analyzing the information together is time consuming and error prone. A database for soybean-root nodulating bradyrhizobia can expedite this by aggregating data from multiple resources and providing necessary data analysis tools.

Here, we have developed Bradybase (http://bradybase.dbi.udel.edu) to address the need for a database focused on soybean root-nodulating *Bradyrhizobium* spp.. The Tripal-based web interface hosts the latest available genome assemblies, genes, nucleotide sequences, gene functional annotations, and gene ontology for the soybean root-nodulating *Bradyrhizobium* spp. Included are *Bradyrhizobium* spp. available in

NCBI and University of Delaware *Bradyrhizobium* Culture Collection (UDBCC), including available phenotypic features, taxonomy, and metadata for genome assemblies including sequencing technology, level of assembly, and number of contigs. Phylogenetic trees and a genome browser feature help to visualize the genomic and genotypic data. Bradybase can be extended to include field experimental results for symbiotic effectiveness of *Bradyrhizobium* spp. on soybean plants, experiment protocols, culture stocks information, comparative genomics information, bioinformatically predicted phages on each genome, and so on to provide a complete reference site for the soybean root-nodulating *Bradyrhizobium* species.

3.3 Methods

3.3.1 Architecture of bradybase database

Bradybase was built using Drupal (https://www.drupal.org/, v7.77), an open-source and extensively used content management system (CMS). Data is arranged in a normalized relational database schema, Chado (Mungall et al., 2007), which is widely adopted to manage biological information. The database is created in a PostgreSQL server v11.11 (https://www.postgresql.org/) (Figure 3.1). Core modules provided by Tripal v3.4, an open-source toolkit used for the construction of online genome databases (Spoor et al., 2019), were used on top of Drupal to manage and visualize data stored in the database. Tripal provides Application Programming Interfaces (APIs) to interact with the Chado database and uses Drupal-based PHP templates that allow extension and customization of web interfaces (Ficklin et al., 2011). Tripal and drupal extension modules were utilized depending on the data analyzed.

3.3.2 Data organization

Chado database schema is divided into modules that allows flexibility to store desired data using only selected tables from the schema with >200 tables. We used 8

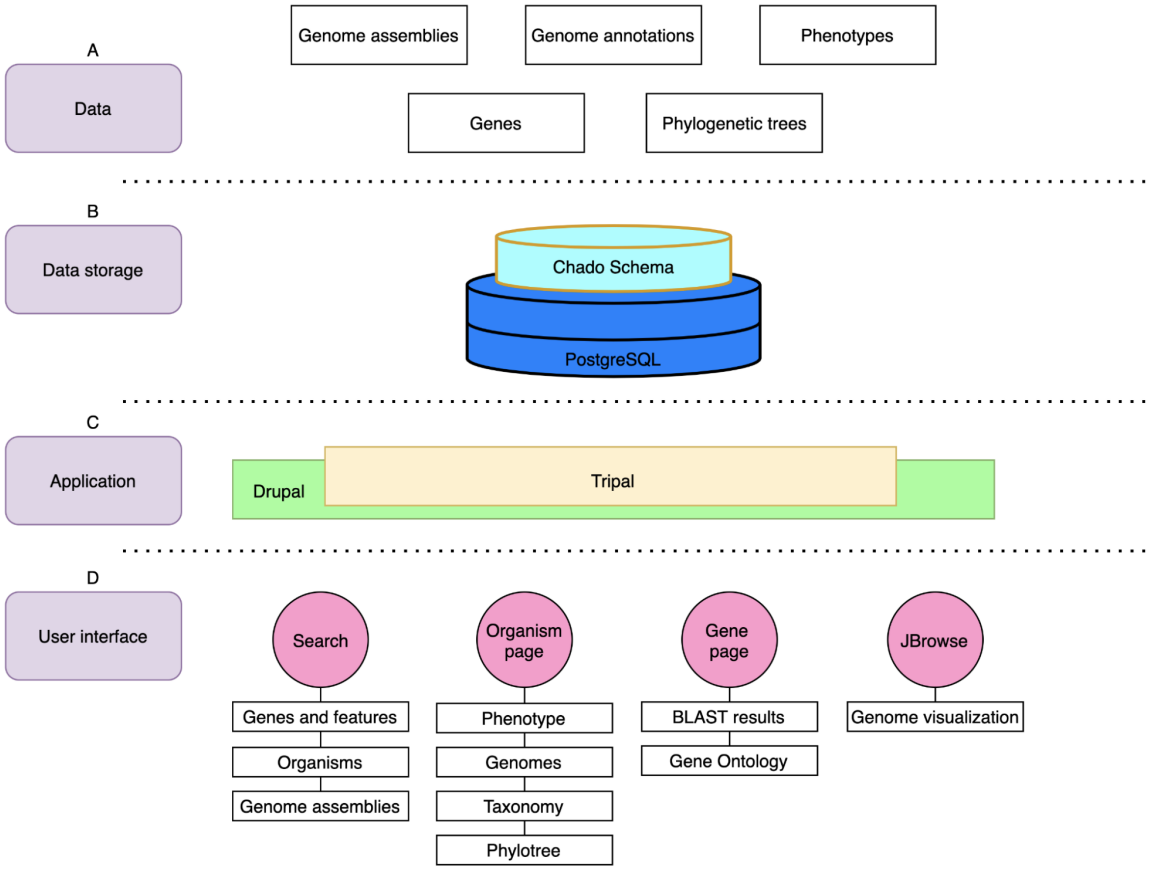


Figure 3.1: The architecture design of Bradybase. (A) Genotypic and phenotypic data were collected from NCBI database and inhouse UDBCC analyses. (B) Data storage implements PostgreSQL DBMS to store and manage data and data is arranged in Chado schema, a relational database scheme designed to store biological data. (C) The application tier uses algorithms to maintain integrity of the database and control display, creation and alteration of data. It uses the Drupal content management system including Tripal modules which are extensively used to manage biological information. (D) The user interface is built by Tripal which provides an API to interact with Chado database and Drupal-based PHP templates.

(general, organism, sequence, analysis, controlled vocabulary (cv), database, publication, and audit) out of >15 available modules.

3.3.3 Organisms

352 accessions including 340 isolates of UDBCC collected from 31 different farms in Delaware (Figure 1.1) and 12 USDA reference strains were added to the database (TableTable A.1). *Bradyrhizobium* species reported as soybean root-nodulating bradyrhizobia in different studies (Jaiswal & Dakora, 2019; Zhang et al., 2014) having genomes available in GenBank as of July 31, 2021, were also incorporated into the database. Genus and species name with strain identification, and abbreviation for each accession was added to organism table from the organism module of Chado schema. Taxonomy for each species was imported from NCBI Taxonomy database using Chado NCBI Taxonomy Loader in Tripal.

3.3.4 Phenotypic features

Phenotypic data available for 352 accessions from UDBCC were included in Bradybase. The collected data included Fatty Acid Methyl Ester (FAME) group, serogroup (from serology), and quantity of spontaneously induced virus-like particles per ml for each accession if available. FAME groups were determined by the unweighted pair group method with arithmetic averages (UPGMA) clustering of cellular fatty acid profiles of each accession as described by Joglekar et al.(2020). Serogroups were determined based on serological reactions of each accession against rabbit polyclonal antisera obtained from the U.S. Department of Agriculture (USDA, Beltsville, Maryland) or produced at the University of Delaware. Spontaneously induced VLPs were measured for a subset of 96 accessions and five additional strains of UDBCC (Richards and Fuhrmann, unpublished data). Selected accessions were individually grown in 15 ml Modified Arabinose Gluconate (MAG) broth for 7-10 days at 28 with shaking at 155 rpm. One milliliter was centrifuged for 15 min at 10,000 rcf. Two-hundred fifty

microliters were filtered through a 0.2 μm glass fiber filter (Whatman; Maidstone, UK). The filtrate was collected on a 0.02 μm Whatman Anodisc membrane filter and stained with 200 μl 2X (final concentration) SYBR Gold for 15 min and then washed with 500 μm SM buffer. Antifade (0.1% p-phenylenediamine solution) was added to the Anodisc to suppress photobleaching and was placed under 1000x epifluorescence microscopy. A 100-square grid was used to count the virus-like production (VLPs) of ten random fields of view of the Anodisc using Serif PhotoPlus X8 for each strain as VLPs per ml. The production levels of spontaneously induced virus-like particles were divided into five categories based on the VLPs per ml measurements: below detection (<1E7), low (1E7 - 1E8), medium (1E8 - 5E8), high (5E8 - 1E9), and very high (>1E9). These features were added to the organismprop table from the organism module of Chado schema.

3.3.5 Genotypic features

Results from genotypic analyses including Internal Transcribed Spacer Restriction Fragment Length Polymorphism (ITS-RFLP), 16S rRNA gene sequencing, and ITS sequencing for UDBCC accessions were added to the database (Table A.1). The internal transcribed spacer (ITS) region between the 16S rRNA and 23S rRNA genes can provide a higher resolution for discerning species and strain-level taxonomic relationships. Genome assemblies and annotations were also added to the database. In-house assemblies for 21 UDBCC accessions and 125 publically available assemblies mined from NCBI RefSeq as of July 31, 2021 (Table 3.1) were uploaded to the database. For NCBI RefSeq assemblies, genome annotations available in RefSeq were uploaded. For in-house assemblies, genome annotations generated by Prokka v1.14.6 (Seemann, 2014) in chapter 2 were uploaded. ITS-RFLP groups were added to the organismprop table while 16S rRNA gene and ITS sequences were added to the feature table from the feature module in Chado schema. Genome assemblies and metadata were stored in analysis and analysisprop tables from the analysis module of Chado schema.

3.3.6 Genomic data analyses

Available genomic features comprising genes, 16S sequences, and ITS sequences for each organism were analyzed further to provide functional annotation for each gene and build different phylogenetic trees for the organisms based on I) 16S rRNA gene, and II) ITS sequences during the construction of Bradybase. Results of the analyses were added to the database to provide additional information about the features.

3.3.6.1 BLASTX homology

BLASTX analysis of each gene was performed against the NCBI non-redundant proteins (nr) database from blast v5 databases with evalue filter of 0.001. The top ten BLAST hits (ordered by bit score) were uploaded into the database using Tripal Analysis BLAST module v3.1 (Tripal Analysis BLAST, 2016/2019). The results including BLASTX hit name, description, and accession were stored in the blast_hit_data table in Chado created by the Tripal Analysis BLAST module.

3.3.6.2 Functional annotation

Gene functional annotation was performed using InterProScan v5.51-85 (Zdobnov & Apweiler, 2001) against the InterPro (Binns et al., 2002) member databases TIGRFAM, SFLD, HAMAP, SMART, CDD, ProSiteProfiles, ProSitePatterns, SUPERFAMILY, PRINTS, PANTHER, PIRSF, Pfam, Coils and MobiDBLite. InterProScan results including InterPro hit and GO annotation for each gene were uploaded into Bradybase using the Tripal Analysis InterPro module (Tripal Analysis InterPro, 2016/2019), to analysisfeatureprop table from analysis module.

3.3.6.3 Phylogenetic trees for 16S rRNA genes

16S rRNA gene sequences of ∼1200 bp length were obtained from Sanger sequencing of 16S rRNA gene amplicons for 96 accessions from UDBCC including 12

USDA strains and 84 field isolates (Prasanna et al., 2020). Multiple sequence alignment was performed using MAFFT v7.450 (Katoh et al., 2002) (G-INS-I algorithm, default settings) and an approximate maximum likelihood phylogenetic tree was created using FastTree v2.1.11 (default settings) (Price et al., 2009) in Geneious v10.2.6 (https://www.geneious.com).

Similarly, A phylogenetic tree for all 16S rRNA sequences in the Bradybase was built. All the sequences were aligned using MAFFT (FTT-NS-i X2 algorithm, default settings). Sequences were trimmed to get equivalent aligned regions of 1200 bp sizes for the 16S rRNA sequences. The extracted sequences were realigned using MAFFT v7.450 (Q-INS-I algorithm, default settings) and an approximate maximum likelihood phylogenetic tree was built using FastTree v2.1.11 (default settings) in Geneious v10.2.6.

3.3.6.4 Phylogenetic trees for ITS sequences

ITS sequences of 900 bp size were amplified and Sanger sequenced from UDBCC accessions which included 12 USDA strains and 75 field isolates (Prasanna et al., 2020). Multiple sequence alignment was performed using MAFFT v7.450 (Katoh et al., 2002) (G-INS-I algorithm, default settings) and a phylogenetic tree was created using FastTree v2.1.11(default settings) (Price et al., 2009) in Geneious v10.2.6.

Similarly, a phylogenetic tree for all ITS sequences in the Bradybase was built. All the sequences were aligned using MAFFT (FTT-NS-i X2 algorithm, default settings). Sequences were trimmed to get equivalent aligned regions of 900 bp sizes for the ITS sequences. The extracted sequences were realigned using MAFFT v7.450 (G-INS-I algorithm, default settings) and a phylogenetic tree was built using FastTree v2.1.11 (default settings) in Geneious v10.2.6.

Newick files for all trees were uploaded into Bradybase and visualized using the Phylotree (Shank et al., 2018) module for Drupal.

3.3.7 Data visualization

3.3.7.1 Genome visualization

JBrowse v1.16 (Buels et al., 2016) was integrated into Bradybase for genome visualization. It was facilitated by Tripal Jbrowse Integration modules v3.0 (Tripal JBrowse Integration, 2015/2020), a package of Tripal extension modules. JBrowse instances were created for each genome assembly, each with four tracks including reference sequence, genes, coding sequences (CDS), 16S rRNA, and tRNA. Assembled genomes in fasta format were used as reference sequences using the JBrowse command prepare-refseqs.pl. Gene, CDS and rRNA tracks were created using gff3 files containing annotations for each genome, using the flatfile-to-json.pl command from JBrowse.

3.3.7.2 Organism page

For each accession, a page was created using Tripal's 'create tripal content type feature to display information including taxonomy, phenotypic analysis results (FAME, Serogroup, spontaneously induced VLPs), and genotypic analysis results (species from ITS sequencing, 16S rRNA gene sequencing, and ITS-RFLP). Links to other analyses for the organism within the database such as phylogenetic trees, genome assemblies, and JBrowse instances were added.

3.3.7.3 Gene page

For each gene, a gene page was created using Tripal's 'create tripal content type feature. The gene page was enabled to display an interactive viewer and a tabular list to visualize the top 10 BLASTX hits of the gene against the nr database from blast v5 databases using Tripal BLAST analysis module (Tripal Analysis BLAST, 2016/2019). Similar visualizations were also incorporated for the results of InterPro analysis of the gene using the Tripal InterPro analysis module (Tripal Analysis InterPro, 2016/2019).

3.3.7.4 Assembly page

For each genome assembly, a page was generated using Tripal's "create Tripal content type" functionality describing the metadata for the genome assembly including name of the accession, source of the assembly, level of assembly and genome-representation of accession by the assembly, sequencing technology used and coverage, scaffold/contig N50 sizes, number of chromosomes in the genome and submitter. For genome assemblies imported from NCBI RefSeq, metadata was also imported from RefSeq.

3.3.8 User accessibility

3.3.8.1 Organism search

A search page for *Bradyrhizobium* accessions in Bradybase was created for the users to search and access accession information from the database. The search interface was built with custom CSS styles added on top of the search interface provided by Mainlab Chado Search module (Jung et al., 2017). Functions were added to search for accessions by species name(s), accession name(s), level of genome assembly available for the accession(s), and phenotypic features: Serogroup, FAME group, and level of spontaneously induced virus-like particles (VLPs) if available.

3.3.8.2 Genes and features search

A search page for genes and other genomic features was created for the users to search and access contents from the database. The search interface was built with custom CSS styles added on top of the search interface provided by Mainlab Chado Search module (Jung et al., 2017). Functions were included to search genomic features such as CDS, gene, pseudogene, rRNA, tmRNA and tRNA according to *Bradyrhizobium* accessions, species names, and feature names. Additionally, functionality was added to search by gene functional annotations such as Gene Ontology (GO) term, BLAST description, and InterPro annotation term.

3.3.8.3 Genome assemblies and phylogenetic trees search

A search page for available genome assemblies in Bradybase was created using Views module (v3.24) for Drupal 7. A display of the list of available genome assemblies grouped by species with a search filter by accession or assembly name is created. Similarly, a search page displaying the list of available phylogenetic trees was also created using Views module (v3.24) for Drupal 7.

3.4 Results

3.4.1 Bradybase website

The homepage of the Bradybase website (Figure B.1) provides a brief description with a direct link to search accessions present in the database. In the main menu of the homepage, users can find a dropdown search menu which provides links to search interfaces in Bradybase, tools used in Bradybase which includes only JBrowse as of August 31, 2021, link to the list of UDBCC isolates with their genotypic and phenotypic characteristics and help menu which provides tutorials to use the website and brief descriptions of the website terminologies.

3.4.2 Organisms

Bradybase hosts 468 different accessions of soybean-root nodulating bradyrhizobia species consisting *B. daqingense*, *B. diazoefficiens*, *B. elkanii*, *B. huanghuaihaiense*, *B. liaoningense*, *B. ottawaense*, *B. yuanmingense*, and *B. diazoefficiens* along with their taxonomic information. It includes 340 field isolates from UDBCC (Table A.1), 19 USDA accessions and 109 other accessions available in GenBank. Taxonomic information is available for each of these accessions to the species level.

3.4.3 Phenotypic and genotypic features

FAME group, serogroup, measure of spontaneously induced VLPs per ml, ITS-RFLP group, 16S rRNA gene sequences and ITS sequences can be accessed for UDBCC accessions whenever available. Species inferred using each of these phenotypic and genotypic analyses can also be retrieved for these accessions.

A total of 146 assemblies including 125 RefSeq assemblies and 21 in-house assemblies of UDBCC accessions are available (TableTable B.2). Genome assemblies are available for 136 accessions. These assemblies include 41 complete genomes, including 19 complete genomes of UDBCC accessions. Gene annotations for all the assemblies are present.

3.4.4 Data visualization

3.4.4.1 Genome visualization

A JBrowse instance is available for each genome to visualize the genome sequence and its annotations (Figure B.6). Each instance consists of a reference sequence, genes, CDS and 16S rRNA tracks. Features can be searched with their names, and locations and clicked to get detailed information including their products, sequences, and location in the genome. JBrowse can be accessed for each assembly from the tools menu in the Bradybase homepage as well as each organism, gene, and genome assembly pages.

3.4.4.2 Organism page

A page is available for each accession in Bradybase. The page provides taxonomic information on the accession to species level. For UDBCC accessions, results from genotypic (ITS-RFLP, ITS sequencing, 16S rRNA gene sequencing) and phenotypic (FAME analysis, serology, spontaneously induced VLP production) analyses are included as well (Figure B.4). A cross reference to NCBI taxonomy database for the

accession or species is also provided in the page. If the accession has any assembled genome, a JBrowse link to visualize the genomes and annotations is available. All available genome assemblies with level of assembly attained for the accession are also displayed. Additionally, users can directly link out to the phylogenetic trees for the organism.

3.4.4.3 Gene page

Bradybase offers a page for each gene (Figure B.8). Users can access the sequence, sequence length, location coordinates in the parent contig, and transcript information for each gene. The page also displays top 10 hits of BLASTX analysis of gene against nr database from blast v5 databases and results from InterPro analysis. Each gene can be located and visualized in JBrowse using the JBrowse link from the page.

3.4.4.4 Assembly page

Assembly page (Figure B.9) provides metadata about each assembly including name of the accession, source of the assembly, level of assembly and genome-representation of accession by the assembly, sequencing technology used and coverage, scaffold/contig N50 sizes, number of chromosomes in the genome and submitter. Users can visualize the assembled genome using the JBrowse link. A cross reference to RefSeq is provided for the genomes obtained from NCBI RefSeq.

3.4.5 User accessibility

3.4.5.1 Organism search

Users can search through all the accessions of soybean root-nodulating bradyrhizobia in Bradybase based on different characteristics. Users can filter their searches on the species name (s), accession name(s), and phenotypic characteristics (serogroup, FAME group, ITS-RFLP group, and production of spontaneously induced virus-like

particles). Filters also include the source of the organism (UDBCC or RefSeq) and genome assembly levels (no assembly available, assembled to complete genome, chromosome, contig(s), and scaffolds) (Figure B.3). Resulting table can be downloaded in csv format.

3.4.5.2 Gene and features search

A search page for genomic features including CDS, gene, pseudogene, rRNA, tmRNA, and tRNA for *Bradyrhizobium* accessions in the Bradybase is created (Figure B.2). Users can filter their searches on gene/feature name(s), *Bradyrhizobium* species, accession name(s), type(s) of the genomic feature, GO annotation of a gene, BLASTN hits description, and InterPro annotation description. The output from the search can be customized to show only selected columns. The resulting table can be downloaded in the form of a csv file, and sequences for the resulting features can be downloaded in fasta format.

3.4.5.3 Genome assemblies and phylogenetic trees search

Genome assemblies for *Bradyrhizobium* spp. can be searched based on species or assembly name. Links to genome assembly page, organism page, JBrowse instance, and cross reference to NCBI RefSeq (for RefSeq assemblies) are available.

3.4.5.4 Phylogenetic trees search

Users can find a display of a list of available phylogenetic trees in the Bradybase. The list includes phylogenetic trees for 1) UDBCC accessions based on 16S rRNA gene sequences, 2) all Bradybase accessions based on 16S rRNA gene sequences, 3) UDBCC accessions based on 16S rRNA gene sequences, and 4) all Bradybase accessions based on ITS sequences. Users can filter for a species/accession using the search box and visualize the tree in either radial or linear pattern (Figure B.5).

3.5 Discussion

3.5.1 Comprehensive database for soybean root-nodulating *Bradyrhizobium* species can accelerate research

Soybean-bradyrhizobia symbiosis has been studied for over 100 years with active research in characterizing indigenous strains and identifying *Bradyrhizobium* spp. with higher symbiotic effectiveness. Knowledge gained is used to improve inoculants and inoculation techniques for soybean yield. Existing research studies are geared towards phenotypic observations for symbiotic effectiveness (nodulation and nitrogen fixation), rhizobiotoxin production, competitiveness with indigenous strains, and stress tolerant characteristics among the *Bradyrhizobium* spp. Bioinformatic analyses such as symbiosis island predictions, identification of bradyrhizobia lytic and lysogenic phages which affects their lifestyle and community dynamics, comparative genomics, and pathways predictions and analysis have been used to gain evolutionary insights and predict phenotypic traits such as N fixation efficiency, nodulation, competitiveness, and adaptability. Additional genomic studies include identification of genes involved in quorum sensing regulation, epigenetics, and understanding distribution of insertion sequences. In the present context, aggregating existing knowledge and running bioinformatics analyses require vast literature review, computational expertise, local space, and efforts. Having a database that can aggregate information available in large scale databases, and produced by different research communities can save time and space for individual research groups, and help the community to learn more about these species with minimized efforts. Sharing results from bioinformatics analyses on a web platform along with their methods and protocols can conserve resources, and help in the reproducibility of the research. Apart from the analyses, providing sampling site location, biological resource centers holding the culture stock along with their location and contact information, and other metadata on storage and collection for each strain can aid in dissemination of available strains. It can be used by researchers and agriculture practitioners to locate *Bradyrhizobium* strains based on phenotypic observations and acquire them for further

research or potential use as soybean inoculants.

Bradybase was therefore designed to benefit the community of researchers working in the soybean root-nodulating bradyrhizobia. No other databases specific to these species have been reported before. Bradybase integrates all the genome assemblies and annotations available in RefSeq for *Bradyrhizobium* species identified to nodulate soybean including *B. diazoefficiens*, *B. elkanii* and *B. japonicum* which are used as commercial inoculants for soybeans. It provides genes, genome visualizations, gene functional annotations, and cross links to external databases for entries from other databases. For each accession, available phenotypic and genotypic data, links to external databases, phylogeny, and other genomic analyses if available are integrated into a single page. This allows more user-friendly navigation and retrieval of genotypic and phenotypic information for each accession compared to existing large scale databases. The database can be extended to include more genotypic and phenotypic observations from literature or research groups, bioinformatic tools and/or analyses including BLAST, comparative genomics and pathways analysis to enable more in depth analyses of the genomes and features with reduced time and effort.

3.5.2 Better access and retrieval of data compared to large-scale databases

The amount of data in the large-scale databases like GenBank is substantial. GenBank contains 231 million sequences with 940 billion bases as of August 2021 (GenBank and WGS Statistics, 2021) and it is expected to increase exponentially. The sheer volume and complexity of data makes specific data search, retrieval, aggregation, and visualization for soybean-root nodulating *Bradyrhizobium* spp. in NCBI or other large scale databases a convoluted process especially for an inexperienced user, which might result in available data being unnoticed by the user. Bradybase therefore provides easier access to genomes, annotations, and visualization of these species.

For a particular use case of downloading all *nodD1* genes from Bd, Be and Bj species, in NCBI, users need to use the advanced search interface to build the

search. It involves the input of the names of *Bradyrhizobium* species/accessions, *nodD1* gene name and Boolean operators. Though users have the option to save the built search for future use, building the search from scratch each time after slight changes in search requirements is cumbersome and time consuming. It requires enough experience to retrieve desired sequences, which may result in error. Bradybase on other hand provides an easier selection menu for species/accession names and gene names through gene/feature search interface, providing faster results (Figure B.7).

Further, with ongoing efforts to include pre-computed pathway analyses, symbiosis islands, lytic and lysogenic phages, BLAST and synteny analysis tools, Bradybase will be able to provide users with specific analyses required for the species but absent in large scale databases.

3.5.3 Using Chado, Tripal, and other GMOD tools is sustainable and time efficient

Bradybase uses GMOD Chado schema, an open-source, generic, and highly normalized database schema that is supported by most of the GMOD tools (Jung et al., 2016). It has modular organization and developers can design custom tables and modules, reducing complexity while adding more flexibility in storing specific biological data (Mungall et al., 2007). Tripal, web front-end from the GMOD project was used to create the website using Drupal. Materialized views were created when necessary to speed up the queries. Both Tripal and Chado are constantly improved according to the increasing needs for biological data storage through a community-involved open process (Spoor et al., 2019) which ensures sustainability of these schema and tools. With out-of-the-box data loaders for the Chado schema and extension modules provided by Tripal, data loading and creating the online site was time efficient and less prone to error.

Similar to Bradybase, Tripal and Chado have been used together to create other organism-specific databases including the Banana Genome Hub (http://banana-genome.

cirad.fr/) (Droc et al., 2013), the *Medicago truncatula* genome database (http://medicago.jcvi.org/MTGD/?q=home) (Krishnakumar et al., 2015), and the *Arabidopsis* Information Portal (https://www.araport.org) (Krishnakumar et al., 2015). As of October 2020, 130 total installations of Tripal have been tracked by Drupal, and 30 databases have been reported to implement partial or whole Tripal software for different plants and animal species but none for bacterial species.

3.5.4 Bradybase enables sharing information among collaborators

With more analysis tools and customizations added to Bradybase in the future, researchers can share their in-house generated data in a searchable online format and use Bradybase as a platform to share ideas and collaborate. Currently, it features the in-house data on genotypic and phenotypic diversity of *Bradyrhizobium* in the state of Delaware generated by University of Delaware.

Chapter 4

CONCLUSION AND FUTURE DIRECTIONS

4.1 Conclusion

The high protein content (40% of dry seeds) and huge market for soybean products (214.36 billions USD by 2025) (Voora et al., 2020) makes soybean one of the most important crops in the world. Soybean-bradyrhizobia symbiosis has been an active area of research owing to its role in sustainably increasing soybean yield via biological nitrogen fixation (BNF). *Bradyrhizobium* strains are studied for their symbiotic effectiveness and highly efficient strains (especially from *B. diazoefficiens*, *B. elkanii* and *B. japonicum* species) are selected for inoculating soybean seeds to increase nitrogen fixation and yield. Despite the high agronomic importance, the number of publicly available complete genomes for these species is low (21 complete genomes as of July 31st, 2021). Complete genomes are necessary to perform genetic studies, conduct comparative genomics, understand the evolution of symbiotic associations, and establish genome to phenome relationships for phenotypic features such as high nodulation capacity and increased nitrogen fixation efficiency. Even for the available genomic data, aggregating and retrieving relevant data from a large scale database like NCBI which contains petabytes of data can be cumbersome and time consuming. Also, they lack support for collaboration and data sharing within the community of the researchers. Specific data types including symbiosis island, greenhouse experimental results on host ranges, symbiotic effectiveness (nodulation and nitrogen fixation capacities), and *Bradyrhizobium* phages are either missing or hard to aggregate in a reproducible manner.

The work presented herein addresses the above mentioned limitations. First, We assembled genomes for 21 accessions (17 field isolates and four USDA strains)

from University of Delaware Culture Collection (UDBCC) to 19 complete genomes and 2 chromosomes. This included five complete genomes and one chromosome for *B. diazoefficiens* accessions, seven complete genomes and one chromosome for *B. elkanii* accessions, and six complete genomes for *B. japonicum* accessions. As of July 31st, 2021, the RefSeq database consisted of only sixteen, four, and one complete genomes for *B. diazoefficiens*, *B. japonicum* and *B. elkanii* respectively. Four additional UDBCC genomes from these three species were previously assembled. Not only does these added genomes increase genomic repertoire available for *Bradyrhizobium* spp. but also amplifies the number of complete genomes for each of these important species which are commonly reported and also used as commercial inoculants. Seventeen accessions were previously unreported in RefSeq and out of remaining four completely assembled USDA reference strains, three (USDA 94-Be, USDA 123-Bj and USDA 135-Bj) were sequenced to only scaffold level, and one (USDA 31) was not sequenced before. This increased availability of genomic knowledge enhances our understanding of soybean bradyrhizobia and each of the commercially important species. Since the sequenced UDBCC accessions were selected to represent 352 UDBCC accessions based on genotypic and phenotypic analyses, it also boosts our knowledge on diversity of *Bradyrhizobium* spp. indigenous to Delaware. Genomic information of indigenous strains can help in improving inoculation techniques during soybean farming resulting in higher soybean yield.

During the work, we also assessed the limitations with Pacific Biosciences (PacBio) long read assembled genomes. The high PacBio subreads error rate of 13-15% (Ardui et al., 2018) which occurred mostly as indels in homopolymer regions resulted in frameshifted genes in the assembled genomes. Genomes contained 5-15% putative frameshifted genes upon comparison to RefSeq proteins, out of which 15-50% were due to single nucleotide indels in homopolymer regions. These putative frameshifted genes rates were observed even after multiple rounds of genome polishing with error-corrected PacBio subreads. We compared the performance of state-of-the-art genomic consensus

tool algorithms for PacBio RS-II data (Arrow and Quiver) to obtain high quality of de-novo assembled genomes. It was observed that while Quiver is positively correlated to mean coverage and mean mapping concordance of PacBio subreads, Arrow does not show significant correlation. Arrow showed an overall increase in genome completeness compared to Quiver which resulted in completeness of some assemblies below acceptable limits for reference genomes, especially when genome coverage was low. Results from chapter 2 suggested Arrow as a better genome genomic consensus tool algorithm for PacBio RS II data assembled genome compared to Quiver. However, genomes were still left with hundreds of putative frameshifted genes even after Arrow polishing of the genomes. This reflected the importance of manually inspecting any genome assembled and polished using only PacBio long reads for residual indel errors that could have been generated by the sequencing platform.

Next, we constructed Bradybase, which is the first known database specific to soybean root-nodulating bradyrhizobia. It stores information about 468 accessions of soybean root-nodulating bradyrhizobia, 142 of which have one or more genome assemblies with 40 of them assembled to complete genomes. It contains 761,714 genes and 724,048 Coding Sequences (CDS) with functional annotations available for each gene. Visualization is available for each gene, CDS and genome via JBrowse. The database also includes phenotypic and genotypic analysis results for UDBCC accessions and metadata for assemblies when available. This will help researchers easily access genomic information available in RefSeq and enables sharing genotypic and phenotypic data collected for UDBCC accessions among the research community.

4.1.1 Future recommendations

Assembled genomes were observed to contain 5-15% of putative frameshifted genes upon comparison with RefSeq CDS. All of these however cannot be categorized as errors from PacBio RS II sequencing technology since the genes could also be truncated due to the natural process of evolution, and existing errors in reference gene

sequences. The number of residual sequencing errors in the assembled genome can be decreased by sequencing the genome to a higher sequencing depth, or polishing long reads with short Illumina reads. Using sanger sequencing to establish authenticity for the source of remaining errors in the genome can enable us to assemble a high quality reference genome. These assembled genomes will be followed up with genomic analyses including symbiosis island identifications, pan genome analyses, and prophage identification. Symbiosis island characterization will help us understand nodulation and nitrogen fixation capabilities and develop genome to phenome relationships with symbiotic effectiveness measured via greenhouse studies for assembled UDBCC accessions. Pan genome analyses of these genomes along with existing completely assembled genomes for common soybean bradyrhizobia inoculant species can be performed to gain more insights on the evolution, environmental adaptability, and symbiotic effectiveness of the strains. Soybean bradyrhizobia prophages can alter their evolution, and community dynamics via horizontal gene transfer, and induction and lysis of the bacterial cells. Identification and characterization of these prophages can increase our understanding of their symbiotic activities.

Bradybase currently stores only genomes, genome annotations, gene functional annotations, phylogenetic trees, and tools for genome visualization. Tripal, a webkit tool used by Bradybase, offers many other modules to enable storage and visualization of genome synteny, comparative genomics, pathway analysis, and enable data sharing and communication with collaborators. These extension modules should be added to the website to increase the scope of Bradybase. Results and protocols from the undergoing pan genome studies as well as other genomic analyses including symbiosis island characterizations, and prophage predictions on UDBCC accessions can be presented in the website in a comprehensible manner. This can be achieved using open-source comparative genomics tools provided by the Generic Model Organism Database (GMOD) project such as Sybil, SynView, and comparative map viewer. Other genomic analyses including prophage predictions, and symbiosis island characterizations can be added

to the database. These genomic predictions can be complemented with outcomes from phenotypic observations from greenhouse studies which illustrate the effect of UDBCC accessions on nodulation, and nitrogen fixation activities, and soybean growth. Similar phenotypic and genotypic analysis results published by other soybean bradyrhizobia research groups can be mined from the literature and added to the database. This will allow users to explore and visualize pre-computed genomic analyses, and easily retrieve existing information on soybean nodulating bradyrhizobia avoiding the need to run the same analyses locally, and repeatedly extract relevant information from literature. It can considerably save time, resources, and effort required during the research.

BIBLIOGRAPHY

[1] Adewale, B. A. (2020). Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? African Journal of Laboratory Medicine, 9(1), 1340. https://doi.org/10.4102/ajlm.v9i1.1340

[2] Akhter, S., Aziz, R. K., & Edwards, R. A. (2012). PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Research*, 40(16), e126. https://doi.org/10.1093/nar/gks406

[3] Appunu, C., Sen, D., Singh, M., & Dhar, B. (2008). VARIATION IN SYMBIOTIC PERFORMANCE OF BRADYRHIZOBIUM JAPONICUM STRAINS AND SOYBEAN CULTIVARS UNDER FIELD CONDITIONS. Journal of Central European Agriculture, 9(1), 169174.

[4] Arashida, H., Odake, H., Sugawara, M., Noda, R., Kakizaki, K., Ohkubo, S., Mitsui, H., Sato, S., & Minamisawa, K. (2021). Evolution of rhizobial symbiosis islands through insertion sequence-mediated deletion and duplication. The ISME Journal, 110. https://doi.org/10.1038/s41396-021-01035-4

[5] Ardui, S., Ameur, A., Vermeesch, J. R., & Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: Applications and utilities for medical diagnostics. Nucleic Acids Research, 46(5), 21592168. https://doi.org/10.1093/nar/gky066

[6] Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: A better, faster version of the PHAST phage search tool. Nucleic Acids Research, 44(Web Server issue), W16W21. https://doi.org/10.1093/nar/gkw387

[7] Avontuur, J. R., Palmer, M., Beukes, C. W., Chan, W. Y., Coetzee, M. P. A., Blom, J., Stpkowski, T., Kyrpides, N. C., Woyke, T., Shapiro, N., Whitman, W. B., Venter, S. N., & Steenkamp, E. T. (2019). Genome-informed Bradyrhizobium taxonomy: Where to from here? Systematic and Applied Microbiology, 42(4), 427439. https://doi.org/10.1016/j.syapm.2019.03.006

[8] Bao, E., Xie, F., Song, C., & Song, D. (2019). FLAS: Fast and high-throughput algorithm for PacBio long-read self-correction. Bioinformatics, 35(20), 39533960. https://doi.org/10.1093/bioinformatics/btz206

[9] Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. Nucleic Acids Research, 41(Database issue), D36D42. https://doi.org/10.1093/nar/gks1195

[10] Bertelli, C., Laird, M. R., Williams, K. P., Lau, B. Y., Hoad, G., Winsor, G. L., & Brinkman, F. S. (2017). IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Research, 45(Web Server issue), W30W35. https://doi.org/10.1093/nar/gkx343

[11] Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Courcelle, E., Durbin, R., Falquet, L., Fleischmann, W., Grifth-Jones, S., Haft, D., Hermjakob, H., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Orchard, S., Sigrist, C. J. A. (2002). InterPro: An integrated documentation resource for protein families, domains and functional sites. . . SEPTEMBER, 3(3), 11.

[12] Blake, J. A., Richardson, J. E., Davisson, M. T., & Eppig, J. T. (1999). The Mouse Genome Database (MGD): Genetic and genomic information about the laboratory mouse. Nucleic Acids Research, 27(1), 9598. https://doi.org/10.1093/nar/27.1.95

[13] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics, 10(1), 421. https://doi.org/10.1186/1471-2105-10-421

[14] Cevallos, M. A., Cervantes-Rivera, R., & Gutirrez-Ros, R. M. (2008). The repABC plasmid family. Plasmid, 60(1), 1937. https://doi.org/10.1016/j.plasmid.2008.03.001

[15] Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., & Botstein, D. (1998). SGD: Saccharomyces Genome Database. Nucleic Acids Research, 26(1), 7379. https://doi.org/10.1093/nar/26.1.73

[16] Chibeba, A. M., Kyei-Boahen, S., Guimares, M. de F., Nogueira, M. A., & Hungria, M. (2017). Isolation, characterization and selection of indigenous Bradyrhizobium strains with outstanding symbiotic performance to increase soybean yields in Mozambique. Agriculture, Ecosystems & Environment, 246, 291305. https://doi.org/10.1016/j.agee.2017.06.017

[17] Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., & Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods, 10(6), 563569. https://doi.org/10.1038/nmeth.2474

[18] Ciampitti, I. A., & Salvagiotti, F. (2018). New Insights into Soybean Biological Nitrogen Fixation. Agronomy Journal, 110(4), 11851196. https://doi.org/10.2134/agronj2017.06.0348

[19] Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. PLOS ONE, 5(6), e11147. https://doi.org/10.1371/journal.pone.0011147

[20] De Maio, N., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., Wick, R., AbuOun, M., Stubberfield, E., Hoosdally, S. J., Crook, D. W., Peto, T. E. A., Sheppard, A. E., Bailey, M. J., Read, D. S., Anjum, M. F., Walker, A. S., & Stoesser, N. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microbial Genomics, 5(9). https://doi.org/10.1099/mgen.0.000294

[21] Droc, G., Larivire, D., Guignon, V., Yahiaoui, N., This, D., Garsmeur, O., Dereeper, A., Hamelin, C., Argout, X., Dufayard, J.-F., Lengelle, J., Baurens, F.-C., Cenci, A., Pitollat, B., DHont, A., Ruiz, M., Rouard, M., & Bocs, S. (2013). The banana genome hub. Database: The Journal of Biological Databases and Curation, 2013, bat035. https://doi.org/10.1093/database/bat035

[22] Du, M., Gao, Z., Li, X., & Liao, H. (2020). Excess nitrate induces nodule greening and reduces transcript and protein expression levels of soybean leghaemoglobins. Annals of Botany, 126(1), 6172. https://doi.org/10.1093/aob/mcaa002

[23] Favre, A. K. L., & Eaglesham, A. R. J. (1986). The effects of high temperatures on soybean nodulation and growth with different strains of bradyrhizobia. Canadian Journal of Microbiology, 32(1), 2227. https://doi.org/10.1139/m86-005

[24] Ficklin, S. P., Sanderson, L.-A., Cheng, C.-H., Staton, M. E., Lee, T., Cho, I.-H., Jung, S., Bett, K. E., & Main, D. (2011). Tripal: A construction toolkit for online genome databases. Database, 2011(bar044). https://doi.org/10.1093/database/bar044

[25] FlyBase Consortium. (1998). FlyBase: A Drosophila database. Nucleic Acids Research, 26(1), 8588. https://doi.org/10.1093/nar/26.1.85 Food and Agriculture Organization of the United Nations. (2020). OECD-FAO AGRICULTURAL OUTLOOK 2020-2029. FOOD & AGRICULTURE ORG.

[26] Fu, S., Wang, A., & Au, K. F. (2019). A comparative evaluation of hybrid error correction methods for error-prone long reads. Genome Biology, 20(1), 26. https://doi.org/10.1186/s13059-018-1605-z

[27] GenBank and WGS Statistics. (n.d.). Retrieved August 26, 2021, from https://www.ncbi.nlm.nih.gov/genbank/statistics/

[28] Gitonga, N. M., Njeru, E. M., Cheruiyot, R., & Maingi, J. M. (2021). Bradyrhizobium inoculation has a greater effect on soybean growth, production and yield quality in organic than conventional farming systems. Cogent Food & Agriculture, 7(1), 1935529. https://doi.org/10.1080/23311932.2021.1935529

[29] GMOD ComponentsGMOD. (n.d.). Retrieved September 3, 2021, from http://gmod.org/wiki/GMOD_Components

[30] GuideStep-By-Step Run Performance Evaluation. (2020). 101, 15.

[31] Harris, T. W., Lee, R., Schwarz, E., Bradnam, K., Lawson, D., Chen, W., Blasier, D., Kenny, E., Cunningham, F., Kishore, R., Chan, J., Muller, H.-M., Petcherski, A., Thorisson, G., Day, A., Bieri, T., Rogers, A., Chen, C.-K., Spieth, J., Stein, L. D. (2003). WormBase: A cross-species database for comparative genomics. Nucleic Acids Research, 31(1), 133137.

[32] Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 18. https://doi.org/10.1016/j.ygeno.2015.11.003

[33] Hepler, N. L., Delaney, N., Brown, M., Smith, M. L., Katzenstein, D., Paxinos, E. E., & Alexander, D. (n.d.). An Improved Circular Consensus Algorithm with an Application to Detect HIV-1 Drug-Resistance Associated Mutations (DRAMs). 1.

[34] Hungria, M., & Mendes, I. C. (2015). Nitrogen Fixation with Soybean: The Perfect Symbiosis? In Biological Nitrogen Fixation (pp. 10091024). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119053095.ch99

[35] Hymowitz, T. (1970). On the domestication of the soybean. Economic Botany, 24(4), 408421. https://doi.org/10.1007/BF02860745

[36] Hymowitz, T., & Shurtleff, W. R. (2005). Debunking Soybean Myths and Legends in the Historical and Popular Literature. Crop Science, 45(2), 473476. https://doi.org/10.2135/cropsci2005.0473

[37] Iida, T., Itakura, M., Anda, M., Sugawara, M., Isawa, T., Okubo, T., Sato, S., Chiba-Kakizaki, K., & Minamisawa, K. (2015). Symbiosis Island Shuffling with Abundant Insertion Sequences in the Genomes of Extra-Slow-Growing Strains of Soybean Bradyrhizobia. Applied and Environmental Microbiology, 81(12), 41434154. https://doi.org/10.1128/AEM.00741-15

[38] Jaiswal, S. K., & Dakora, F. D. (2019). Widespread Distribution of Highly Adapted Bradyrhizobium Species Nodulating Diverse Legumes in Africa. Frontiers in Microbiology, 10. https://doi.org/10.3389/fmicb.2019.00310

[39] Joglekar, P. (2021). THE IMPACT OF SPONTANEOUSLY PRODUCED LYSOGENIC PHAGES ON THE ECOLOGY AND BIOLOGY OF SOYBEAN BRADYRHIZOBIA [PhD dissertation]. University of Delaware.

[40] Joglekar, P., Mesa, C. P., Richards, V. A., Polson, S. W., Wommack, K. E., & Fuhrmann, J. J. (2020). Polyphasic analysis reveals correlation between phenotypic and genotypic analysis in soybean bradyrhizobia

(Bradyrhizobium spp.). Systematic and Applied Microbiology, 43(3), 126073. https://doi.org/10.1016/j.syapm.2020.126073

[41] Jordan, D. C. (1982). NOTES: Transfer of Rhizobium japonicum Buchanan 1980 to Bradyrhizobium gen. nov., a Genus of Slow-Growing, Root Nodule Bacteria from Leguminous Plants. International Journal of Systematic Bacteriology, 32(1), 136139. https://doi.org/10.1099/00207713-32-1-136

[42] Jung, S., Staton, M., Lee, T., Blenda, A., Svancara, R., Abbott, A., & Main, D. (2008). GDR (Genome Database for Rosaceae): Integrated web-database for Rosaceae genomics and genetics data. Nucleic Acids Research, 36(Database issue), D1034-1040. https://doi.org/10.1093/nar/gkm803

[43] Karp, P. D., Weaver, D., Paley, S., Fulcher, C., Kubo, A., Kothari, A., Krummenacker, M., Subhraveti, P., Weerasinghe, D., Gama-Castro, S., Huerta, A. M., Muiz-Rascado, L., Bonavides-Martinez, C., Weiss, V., Peralta-Gil, M., Santos-Zavaleta, A., Schrder, I., Mackie, A., Gunsalus, R., Paulsen, I. (2014). The EcoCyc Database. EcoSal Plus, 6(1), 10.1128/ecosalplus.ESP-00092013. https://doi.org/10.1128/ecosalplus.ESP-0009-2013

[44] Keyser, H. H. (n.d.). Potential for increasing biological nitrogen fixation in soybean. 17.

[45] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research, 27(5), 722736. https://doi.org/10.1101/gr.215087.116

[46] Krishnakumar, V., Hanlon, M. R., Contrino, S., Ferlanti, E. S., Karamycheva, S., Kim, M., Rosen, B. D., Cheng, C.-Y., Moreira, W., Mock, S. A., Stubbs, J., Sullivan, J. M., Krampis, K., Miller, J. R., Micklem, G., Vaughn, M., & Town, C. D. (2015). Araport: The Arabidopsis Information Portal. Nucleic Acids Research, 43(D1), D1003D1009. https://doi.org/10.1093/nar/gku1200

[47] Krishnakumar, V., Kim, M., Rosen, B. D., Karamycheva, S., Bidwell, S. L., Tang, H., & Town, C. D. (2015). MTGD: The Medicago truncatula genome database. Plant & Cell Physiology, 56(1), e1. https://doi.org/10.1093/pcp/pcu179

[48] Krumsiek, J., Arnold, R., & Rattei, T. (2007). Gepard: A rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics, 23(8), 10261028. https://doi.org/10.1093/bioinformatics/btm039

[49] Lang, D., Zhang, S., Ren, P., Liang, F., Sun, Z., Meng, G., Tan, Y., Li, X., Lai, Q., Han, L., Wang, D., Hu, F., Wang, W., & Liu, S. (2020). Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. GigaScience, 9(12), giaa123. https://doi.org/10.1093/gigascience/giaa123

[50] Laver, T., Harrison, J., ONeill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. Biomolecular Detection and Quantification, 3, 18. https://doi.org/10.1016/j.bdq.2015.02.001

[51] Lerat, E., & Ochman, H. (2005). Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Research, 33(10), 31253132. https://doi.org/10.1093/nar/gki631

[52] Liao, Y.-C., Lin, S.-H., & Lin, H.-H. (2015). Completing bacterial genome assemblies: Strategy and performance comparisons. Scientific Reports, 5(1), 8747. https://doi.org/10.1038/srep08747

[53] Liu, K. (1997). Soybeans. Springer US. https://doi.org/10.1007/978-1-4615-1763-4

[54] Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. Nature Reviews Genetics, 21(10), 597614. https://doi.org/10.1038/s41576-020-0236-x

[55] Mahmoud, M., Zywicki, M., Twardowski, T., & Karlowski, W. M. (2019). Efficiency of PacBio long read correction by 2nd generation Illumina sequencing. Genomics, 111(1), 4349. https://doi.org/10.1016/j.ygeno.2017.12.011

[56] Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. Frontiers in Genetics, 10, 426. https://doi.org/10.3389/fgene.2019.00426

[57] McDermott, T., & Graham, P. (1990). Competitive Ability and Efficiency in Nodule Formation of Strains of Bradyrhizobium japonicum. Applied and Environmental Microbiology, 56(10), 30353039. https://doi.org/10.1128/aem.56.10.3035-3039.1990

[58] Molina-Mora, J. A., Campos-Snchez, R., Rodrguez, C., Shi, L., & Garca, F. (2020). High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers. Scientific Reports, 10(1), 1392. https://doi.org/10.1038/s41598-020-58319-6

[59] Montgomery, K. S. (2003). Soy Protein. The Journal of Perinatal Education, 12(3), 4245. https://doi.org/10.1624/105812403X106946

[60] Mungall, C. J., Emmert, D. B., & The FlyBase Consortium. (2007). A Chado case study: An ontology-based modular schema for representing genome-associated biological information. Bioinformatics, 23(13), i337i346. https://doi.org/10.1093/bioinformatics/btm189

[61] Ormeo-Orrillo, E., & Martnez-Romero, E. (2019). A Genomotaxonomy View of the Bradyrhizobium Genus. Frontiers in Microbiology, 10. https://doi.org/10.3389/fmicb.2019.01334

[62] Ou, S., Liu, J., Chougule, K. M., Fungtammasan, A., Seetharam, A. S., Stein, J. C., Llaca, V., Manchanda, N., Gilbert, A. M., Wei, S., Chin, C.-S., Hufnagel, D. E., Pedersen, S., Snodgrass, S. J., Fengler, K., Woodhouse, M., Walenz, B. P., Koren, S., Phillippy, A. M., Ware, D. (2020). Effect of sequence depth and length in long-read assembly of the maize inbred NC358. Nature Communications, 11(1), 2288. https://doi.org/10.1038/s41467-020-16037-7

[63] Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., & Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research, 42(D1), D206D214. https://doi.org/10.1093/nar/gkt1226

[64] PacificBiosciences/GenomicConsensus. (2021). [Python]. Pacific Biosciences. https://github.com/PacificBiosciences/GenomicConsensus (Original work published 2012)

[65] Padukkage, D., Geekiyanage, S., Reparaz, J. M., Bezus, R., Balatti, P. A., & Degrassi, G. (2021). Bradyrhizobium japonicum, B. elkanii and B. diazoefficiens Interact with Rice (Oryza sativa), Promote Growth and Increase Yield. Current Microbiology, 78(1), 417428. https://doi.org/10.1007/s00284-020-02249-z

[66] Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research, 25(7), 10431055. https://doi.org/10.1101/gr.186072.114

[67] Pruitt, K., Brown, G., Tatusova, T., & Maglott, D. (2012). The Reference Sequence (RefSeq) Database. In The NCBI Handbook [Internet]. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK21091/

[68] Robinson, K. O., Burton, J. W., Taliercio, E. W., Israel, D. W., & Carter Jr., T. E. (2020). Inheritance of rhizobitoxine-induced chlorosis in soybean. Crop Science, 60(6), 30273034. https://doi.org/10.1002/csc2.20193

[69] Rodrguez-Navarro, D. N., Margaret Oliver, I., Albareda Contreras, M., & Ruiz-Sainz, J. E. (2011). Soybean interactions with soil microbes, agronomical and molecular aspects. Agronomy for Sustainable Development, 31(1), 173190. https://doi.org/10.1051/agro/2010023

[70] Rong Li, Chen, H., Yang, Z., Yuan, S., & Zhou, X. (2020). Research status of soybean symbiosis nitrogen fixation. Oil Crop Science, 5(1), 610. https://doi.org/10.1016/j.ocsci.2020.03.005

[71] Salmela, L., Walve, R., Rivals, E., & Ukkonen, E. (2017). Accurate self-correction of errors in long reads using de Bruijn graphs. Bioinformatics, 33(6), 799806. https://doi.org/10.1093/bioinformatics/btw321

[72] Salvagiotti, F., Cassman, K. G., Specht, J. E., Walters, D. T., Weiss, A., & Dobermann, A. (2008). Nitrogen uptake, fixation and response to fertilizer N in soybeans: A review. Field Crops Research, 108(1), 113. https://doi.org/10.1016/j.fcr.2008.03.001

[73] Sameshima-Saito, R., Chiba, K., & Minamisawa, K. (2006). Correlation of Denitrifying Capability with the Existence of nap, nir, nor and nos Genes in Diverse Strains of Soybean Bradyrhizobia. Microbes and Environments, 21(3), 174184. https://doi.org/10.1264/jsme2.21.174

[74] Sanderson, L.-A., Ficklin, S. P., Cheng, C.-H., Jung, S., Feltus, F. A., Bett, K. E., & Main, D. (2013). Tripal v1.1: A standards-based toolkit for construction of online genetic and genomic databases. Database: The Journal of Biological Databases and Curation, 2013, bat075. https://doi.org/10.1093/database/bat075

[75] Savci, S. (2012). Investigation of Effect of Chemical Fertilizers on Environment. APCBEE Procedia, 1, 287292. https://doi.org/10.1016/j.apcbee.2012.03.047

[76] Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. Bioinformatics, 30(14), 20682069. https://doi.org/10.1093/bioinformatics/btu153

[77] Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechnology, 26(10), 11351145. https://doi.org/10.1038/nbt1486

[78] Shiro, S., Matsuura, S., Saiki, R., Sigua, G. C., Yamamoto, A., Umehara, Y., Hayashi, M., & Saeki, Y. (2013). Genetic Diversity and Geographical Distribution of Indigenous Soybean-Nodulating Bradyrhizobia in the United States. Applied and Environmental Microbiology, 79(12), 36103618. https://doi.org/10.1128/AEM.00236-13

[79] Siguier, P., Gourbeyre, E., & Chandler, M. (2014). Bacterial insertion sequences: Their genomic impact and diversity. FEMS Microbiology Reviews, 38(5), 865891. https://doi.org/10.1111/1574-6976.12067

[80] Siqueira, A. F., Ormeo-Orrillo, E., Souza, R. C., Rodrigues, E. P., Almeida, L. G. P., Barcellos, F. G., Batista, J. S. S., Nakatani, A. S., Martnez-Romero, E., Vasconcelos, A. T. R., & Hungria, M. (2014). Comparative genomics of Bradyrhizobium japonicum CPAC 15 and Bradyrhizobium diazoefficiens CPAC 7: Elite

model strains for understanding symbiotic performance with soybean. BMC Genomics, 15(1), 420. https://doi.org/10.1186/1471-2164-15-420

[81] Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next Generation Sequencing Technologies. Current Protocols in Molecular Biology, 122(1), e59. https://doi.org/10.1002/cpmb.59

[82] Smith, K. (2013). A Brief History of NCBIs Formation and Growth. In The NCBI Handbook [Internet]. 2nd edition. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK148949/

[83] Smits, T. H. M. (2019). The importance of genome sequence quality to microbial comparative genomics. BMC Genomics, 20(1), 662. https://doi.org/10.1186/s12864-019-6014-5

[84] SMRT Link Software Installation (v7.0.1). (n.d.). 21.

[85] Spoor, S., Cheng, C.-H., Sanderson, L.-A., Condon, B., Almsaeed, A., Chen, M., Bretaudeau, A., Rasche, H., Jung, S., Main, D., Bett, K., Staton, M., Wegrzyn, J. L., Feltus, F. A., & Ficklin, S. P. (2019). Tripal v3: An ontology-based toolkit for construction of FAIR biological community databases. Database, 2019(baz077). https://doi.org/10.1093/database/baz077

[86] Spoor, S., Wytko, C., Soto, B., Chen, M., Almsaeed, A., Condon, B., Herndon, N., Hough, H., Jung, S., Staton, M., Wegrzyn, J., Main, D., Feltus, F. A., & Ficklin, S. P. (2020). Tripal and Galaxy: Supporting reproducible scientific workflows for community biological databases. Database, 2020(baaa032). https://doi.org/10.1093/database/baaa032

[87] Staton, M., Cannon, E., Sanderson, L.-A., Wegrzyn, J., Anderson, T., Buehler, S., Cobo-Simn, I., Faaberg, K., Grau, E., Guignon, V., Gunoskey, J., Inderski, B., Jung, S., Lager, K., Main, D., Poelchau, M., Ramnath, R., Richter, P., West, J., & Ficklin, S. (2021). Tripal, a community update after 10 years of supporting open source, standards-based genetic, genomic and breeding databases. Briefings in Bioinformatics, bbab238. https://doi.org/10.1093/bib/bbab238

[88] Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A., & Lewis, S. (2002). The Generic Genome Browser: A Building Block for a Model Organism System Database. Genome Research, 12(10), 15991610. https://doi.org/10.1101/gr.403602

[89] Taylor, R. W., Williams, M. L., & Sistani, K. R. (n.d.). N 2 fixation by soybean-Bradyrhizobium combinations under acidity, low P and high AI stresses. 8.

[90] Thuita, M., Pypers, P., Herrmann, L., Okalebo, R. J., Othieno, C., Muema, E., & Lesueur, D. (2012). Commercial rhizobial inoculants significantly enhance growth

and nitrogen fixation of a promiscuous soybean variety in Kenyan soils. Biology and Fertility of Soils, 48(1), 8796. https://doi.org/10.1007/s00374-011-0611-z

[91] Tian, C. F., Zhou, Y. J., Zhang, Y. M., Li, Q. Q., Zhang, Y. Z., Li, D. F., Wang, S., Wang, J., Gilbert, L. B., Li, Y. R., & Chen, W. X. (2012). Comparative genomics of rhizobia nodulating soybean suggests extensive recruitment of lineage-specific genes in adaptations. Proceedings of the National Academy of Sciences, 109(22), 86298634. https://doi.org/10.1073/pnas.1120436109

[92] Tripal/tripal_blast. (2020). [PHP]. Tripal. https://github.com/tripal/tripal_blast (Original work published 2015)

[93] Tu, J. C. (1981). Effect of salinity on *Rhizobium*-root-hair interaction, nodulation and growth of soybean. Canadian Journal of Plant Science, 61(2), 231239. https://doi.org/10.4141/cjps81-035.

[94] Tvedte, E. S., Gasser, M., Sparklin, B. C., Michalski, J., Zhao, X., Bromley, R., Tallon, L. J., Sadzewicz, L., Rasko, D. A., & Hotopp, J. C. D. (2020). Comparison of long read sequencing technologies in resolving bacteria and fly genomes (p. 2020.07.21.213975). https://doi.org/10.1101/2020.07.21.213975

[95] United Nations. (2019). World Population Prospects 2019: Data Booklet. UN. https://doi.org/10.18356/3e9d869f-en USDA. (2020). Crop Production 2020 Summary. Crop Production, 125.

[96] Utturkar, S. M., Klingeman, D. M., Hurt, R. A. J., & Brown, S. D. (2017). A Case Study into Microbial Genome Assembly Gap Sequences and Finishing Strategies. Frontiers in Microbiology, 8. https://doi.org/10.3389/fmicb.2017.01272

[97] van Berkum, P. (1990). Evidence for a Third Uptake Hydrogenase Phenotype among the Soybean Bradyrhizobia. Applied and Environmental Microbiology, 56(12), 38353841.

[98] van Heerwaarden, J., Baijukya, F., Kyei-Boahen, S., Adjei-Nsiah, S., Ebanyat, P., Kamai, N., Wolde-meskel, E., Kanampiu, F., Vanlauwe, B., & Giller, K. (2018). Soyabean response to rhizobium inoculation across sub-Saharan Africa: Patterns of variation and the role of promiscuity. Agriculture, Ecosystems & Environment, 261, 211218. https://doi.org/10.1016/j.agee.2017.08.016

[99] Voora, V., Larrea, C., & Bermdez, S. (n.d.). Global Market Report: Soybeans. 20.

[100] Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE, 9(11), e112963. https://doi.org/10.1371/journal.pone.0112963

[101] Wang, Q., Liu, J., & Zhu, H. (2018). Genetic and Molecular Mechanisms Underlying Symbiotic Specificity in Legume-Rhizobium Interactions. Frontiers in Plant Science, 9. https://doi.org/10.3389/fpls.2018.00313

[102] Watson, M., & Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. Nature Biotechnology, 37(2), 124126. https://doi.org/10.1038/s41587-018-0004-z

[103] Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., & Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research, 6, 100. https://doi.org/10.12688/f1000research.10571.2

[104] Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Tpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., Hunkapiller, M. W. (2019). Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. BioRxiv, 519025. https://doi.org/10.1101/519025

[105] Yuhashi, K.-I., Ichikawa, N., Ezura, H., Akao, S., Minakawa, Y., Nukui, N., Yasuta, T., & Minamisawa, K. (2000). Rhizobitoxine Production by Bradyrhizobium elkanii Enhances Nodulation and Competitiveness on Macroptilium atropurpureum. Applied and Environmental Microbiology, 66(6), 26582663.

[106] Zdobnov, E. M., & Apweiler, R. (2001). InterProScan an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 17(9), 847848. https://doi.org/10.1093/bioinformatics/17.9.847

[107] Zhang, H., Jain, C., & Aluru, S. (2020). A comprehensive evaluation of long read error correction methods. BMC Genomics, 21(6), 889. https://doi.org/10.1186/s12864-020-07227-0

[108] Zhang, X. X., Guo, H. J., Wang, R., Sui, X. H., Zhang, Y. M., Wang, E. T., Tian, C. F., & Chen, W. X. (2014). Genetic Divergence of Bradyrhizobium Strains Nodulating Soybeans as Revealed by Multilocus Sequence Analysis of Genes Inside and Outside the Symbiosis Island. Applied and Environmental Microbiology, 80(10), 31813190. https://doi.org/10.1128/AEM.00044-14

[109] Zimin, A. V., & Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLOS Computational Biology, 16(6), e1007981. https://doi.org/10.1371/journal.pcbi.1007981

Appendix A

SUPPLEMENTARY FILES FOR CHAPTER 2

Table A.1: Accessions of University of Delaware Bradyrhizobia Culture Collection (UDBCC) with results from phenotypic and genotypic analyses. Selected accessions were chosen for sequencing.

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
K01A	1	94	1			Bj	
K01B	8	123	1	Bj	n/a	Bj	
K01C	2	6	1			Bj	
K01D	12	110-94-46w3	4			Bd	
K01E	12	110	4	Bd	Bd	Bd	Seq[6]
K01F	12	122-38-6	4			Bd	
K01G	12	122-110-38	4			Bd	
K01H	14	46-94	8			Be	
K01I	17	94-123	10			Be	
K01J	17	94-123	8			Be	
K01K	12	110	4			Bd	
K01L	15	94	11			Be	
K02A	5	94	4	Be	Be	Be	
K02B	15	94-122-123w	5	Bd	Bd	Bd	
K02C	2	38-123	3			Bj	
K02D	3	38-6	3	Bj	Bj	Bj	
K02E	3	38-6	1			Bj	
K02F	14	31	8	Be	Be	Be	
K02G	2	6-38	3			Bj	
K02H	12	122-110w	4			Bd	
K02I	14	76-94	8			Be	
K02J	17	94-123	8	Be	Be	Be	
K02K	17	94	8	Be	Be	Be	Seq
K02L	8	38w	3			Bj	
K03A	12	110	4			Bd	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
K03B	15	46	9			Be	
K03C	1	6-38	3	Bj	Bj	Bj	
K03D	15	94-123	8	Be	Be	Be	Seq
K03E	14	130	8	Be	Be	Be	
K03F	14	46-130	9			Be	
K03G	17	110-94-123	8			Be	
K03H	14	130	8			Be	
K03I	17	94-123	11	Be	Be	Be	Seq
K03J	15	94-123	8			Be	
K03K	12	110	4			Bd	
K03L	16	NR	8			Be	
K04A	15	31	8			Be	
K04B	12	122-38	4			Bd	
K04C	2	6-38	3	Bj	Bj	Bj	
K04D	12	122	4	Bd	Bd	Bd	
K04E	2	6-38	3			Bj	
K04F	14	46	9			Be	
K04G	14	46	9			Be	
K04H	15	94-123	8	Be	Be	Be	
K04I	14	46	9			Be	
K04J	2	6-38	3			Bj	
K04K	2	6-38	3			Bj	
K04L	12	122w	4			Bd	
K05A	12	110	5			Bd	
K05B	12	122w	4			Bd	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
K05C	15	94-123	8			Be	
K05D	12	122w	4			Bd	
K05E	15	94-123	9			Be	
K05F	12	110	5			Bd	
K05G	3	38-6	1			Bj	
K05H	2	6	3			Bj	
K05I	12	110	5			Bd	
K05J	12	110	5			Bd	
K05K	12	110	5			Bd	
K05L	12	46	4			Bd	
K06A	12	NR	5			Bd	
K06B	1	NR	4	Bd	Bd	Bd	
K06C	5	123	1			Bj	
K06D	5	123	3			Bj	
K06E	5	123	3			Bj	
K06F	17	94-123	9			Be	
K06G	12	110	4			Bd	
K06H	12	122w	4			Bd	
K06I	13	110w	4			Bd	
K06J	15	94-123	8			Be	
K06K	5	123	3			Bj	
K06L	13	NR	4	Bd	Bd	Bd	
K07A	3	38-6	1			Bj	
K07B	17	94-46-123	9			Be	
K07C	13	NR	4			Bd	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
K07D	12	110	4			Bd	
K07E	5	123	1			Bj	
K07F	12	94	4			Bd	
K07G	19	46	6	Bd	Bd	Bd	
K07H	12	122	4			Bd	
K07I	12	94-122-123w	4			Bd	
K07J	14	76	9			Be	
K07K	12	123-122	4			Bd	
K07L	15	31	8			Be	
K08A	12	122w	4			Bd	
K08B	17	31	9	Be	Be	Be	
K08C	9	94-123w-6w	1	Be	Be	Be	
K08D	1	6-38	3			Bj	
K08E	15	31	8			Be	
K08F	12	122-38	4	Bd	Bd	Bd	
K08G	14	46	8			Be	
K08H	15	31	8			Be	
K08I	13	NR	4	Bd	Bd	Bd	
K08J	17	31	8	Be	Be	Be	
K08K	15	94-123	8			Be	
K08L	12	110	4			Bd	
K09A	14	46	9			Be	
K09B	3	38-6	1			Bj	
K09C	12	NR	4	Bd	Bd	Bd	
K09D	12	123-122	4			Bd	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
K09E	12	110	4			Bd	
K09F	12	NR	4	Bd	Bd	Bd	Seq
K09G	15	31	8			Be	
K09H	14	46	8			Be	
K09I	12	110	4			Bd	
K09J	12	76-122w	4			Bd	
K09K	12	NR	4			Bd	
K09L	12	122w	4			Bd	
K10A	15	94-123	8			Be	
K10B	8	38-123	1			Bj	
K10C	5	123	2			Bj	
K10D	16	NR	8			Be	
K10E	3	38-6	1			Bj	
K10F	12	110	4			Bd	
K10G	15	94-123	8			Be	
K10H	1	6-38	1	Bj	Bj	Bj	
K10I	14	46	9			Be	
K10J	12	110	4			Bd	
K10K	12	110	4			Bd	
K10L	12	122-38	4	Bd	Bd	Bd	
N01A	16	46	11			Be	
N01B	13	46	4			Bd	
N01C	16	46-122	8			Be	
N01D	12	110-46	4	Bd	Bd	Bd	
N01E	5	123	3	Bj	Bj	Bj	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
N01F	16	46-123	6			Be	
N01G	12	122-38	4			Bd	
N01H	5	94-46-123	3	Bd	Bd	Bd	
N01I	12	123-122	5			Bd	
N01J	16	46	9			Be	
N01K	14	31	9	Be	Be	Be	
N01L	16	46	11	Be	Be	Be	
N02A	2	6-123	3			Bj	
N02B	12	62	4	Bd	Bd	Bd	
N02C	17	94-123	9			Be	
N02D	no product	6	3	Bj	Bj	Bj	
N02E	12	122-38	4			Bd	
N02F	12	NR	4			Bd	
N02G	12	62	4	Bd	Bd	Bd	
N02H	17	94-123	9			Be	
N02I	12	122	4			Bd	
N02J	12	122w	4			Bd	
N02K	12	122-38	4			Bd	
N02L	12	110-31	4			Bd	
N03A	12	122-38	4			Bd	
N03B	12	110	4	Bd	Bd	Bd	Seq
N03C	12	122-38	4			Bd	
N03D	2	6	3			Bj	
N03E	2	6	3			Bj	
N03F	1	6-38	3			Bj	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
N03G	1	NR	3	Bj	Bj	Bj	Seq
N03H	12	NR	4	Bd	Bd	Bd	
N03I	12	110	4			Bd	
N03J	15	94-123	8			Be	
N03K	12	NR	4			Bd	
N03L	14	76	8			Be	
S01A	12	122-38	4			Bd	
S01B	12	122-38	4			Bd	
S01C	14	46	8	Be	Be	Be	
S01D	12	110	4			Bd	
S01E	2	6-123	2			Bj	
S01F	2	6-123-110	3			Bj	
S01G	14	46	9			Be	
S01H	12	31-110	9	Be	Be	Be	
S01I	14	46	8			Be	
S01J	15	94-123	9			Be	
S01K	14	46	9			Be	
S01L	1	NR	1			Bj	
S02A	12	110	5			Bd	
S02B	8	94-123	3	Bj	Bj	Bj	
S02C	13	NR	4			Bd	
S02D	5	123	1	Bj	Bj	Bj	
S02E	12	NR	4			Bd	
S02F	8	123	3			Bj	
S02G	13	NR	4			Bd	

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
S02H	5	123w	3			Bj	
S02I	12	NR	4			Bd	
S02J	12	110	4	Bd	Bd	Bd	
S02K	13	NR	4			Bd	
S02L	13	NR	4			Bd	
S03A	8	123	1	Bj	n/a	Bj	
S03B	15	94-123	8			Be	
S03C	15	94-123w	9			Be	
S03D	12	122-123	4			Bd	
S03E	12	122-38	5			Bd	
S03F	12	122-38	4			Bd	
S03G	12	122	5	Bd	Bd	Bd	
S03H	15	94-123	11			Be	
S03I	12	122	5			Bd	
S03J	12	122-38	5			Bd	
S03K	14	46	9			Be	
S03L	14	76	9	Be	Be	Be	
S04A	8	123-6	1	Bj	Bj	Bj	
S04B	3	38-46	2			Bj	
S04C	12	31-122	4			Bd	
S04D	12	110w	4			Bd	
S04E	3	38-6	1	Bj	Bj	Bj	Seq
S04F	2	6	1			Bj	
S04G	12	122-38	5			Bd	
S04H	no product	NR	4			Bd	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
S04I	no product	NR	4			Bd	
S04J	15	94-123	11			Be	
S04K	12	122	3	Bd	Bd	Bd	
S04L	no product	NR	4			Bd	
S05A	14	31	8			Be	
S05B	14	31-76	8			Be	
S05C	14	46	9			Be	
S05D	14	31	9			Be	
S05E	14	46	8			Be	
S05F	14	31	8			Be	
S05G	12	123-122	5			Bd	
S05H	14	76	9			Be	
S05I	14	46	9			Be	
S05J	14	130	8	Be	Be	Be	Seq
S05K	14	130	9			Be	
S05L	15	94-123	9			Be	
S06A	17	94-123	11			Be	
S06B	8	123w	3	Bj	Bj	Bj	Seq
S06C	17	94-123	9			Be	
S06D	14	76	9			Be	
S06E	2	6	3	Bj	Bj	Bj	
S06F	3	38-6	1			Bj	
S06G	1	NR	3	Bj	Bj	Bj	
S06H	3	6-38	3	Bj	Bj	Bj	
S06J	3	6-38	3			Bj	

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
S06K	3	122	3	Bj	Bj	Bj	Seq
S06L	7	NR	3	Bj	Bj	Bj	
S07A	3	130	3			Bj	
S07B	3	38	1	Bj	Bj	Bj	
S07C	15	31-6	4			Be	
S07D	3	38-6	1			Bj	
S07E	3	38-122-31	1			Bj	
S07F	1	6-38	3			Bj	
S07G	3	6-38	3	Bj	Bj	Bj	
S07H	2	6	1	Bj	Bj	Bj	
S07I	12	122-38	4			Bd	
S07J	2	76-6	1	Be	Be	Be	Seq
S07K	1	6-38	3			Bj	
S07L	3	38-6	1			Bj	
S08A	14	130	8			Be	
S08B	1	6-38	3			Bj	
S08C	2	6-123	1			Bj	
S08E	14	76	9			Be	
S08F	2	76-6	8	Be	Be	Be	
S08G	14	130	9			Be	
S08H	3	130	3			Bj	
S08I	12	NR	5			Bd	
S08J	1	6-94	3			Bj	
S08K	12	122w	5			Bd	
S08L	no product	NR	4	Bd	Bd	Bd	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
S09A	8	123	3			Bj	
S09B	3	6-38	3			Bj	
S09C	3	38-6	3			Bj	
S09D	1	NR	3			Bj	
S09E	2	NR	3			Bj	
S09F	8	123-122-38	3			Bj	
S09G	8	94-123w	3			Bj	
S09H	8	122-123	3			Bj	
S09I	8	NR	3			Bj	
S09J	no product	NR	4	Bj	Bj	Bj	
S09K	3	38-6	1			Bj	
S09L	8	123-6w	3			Bj	
S10A	3	NR	3			Bj	
S10B	2	NR	3			Bj	
S10C	8	123	3			Bj	
S10D	8	6-123	3	Bj	n/a	Bj	
S10E	8	123-6	3			Bj	
S10F	8	123	1			Bj	
S10G	1	NR	3			Bj	
S10H	12	122-38	4			Bd	
S10I	8	NR	3	Bj	n/a	Bj	
S10J	8	123	6	Bj	Bj	Bj	Seq
S10K	1	NR	3			Bj	
S10L	8	123-6	3			Bj	
S11A	5	123w	3			Bj	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
S11B	8	123-6	3	Bj	n/a	Bj	
S11C	12	110	4			Bd	
S11D	5	123w	3			Bj	
S11E	8	NR	3	Bj	n/a	Bj	
S11F	no product	NR	3	Bj	n/a	Bj	
S11G	3	38-6	1			Bj	
S11H	8	123	3			Bj	
S11I	5	123w	3			Bj	
S11J	3	38-6	3			Bj	
S11K	no product	NR	4			Bd	
S11L	8	NR	3	Bj	n/a	Bj	Seq
S12A	14	46w	6	Be	Be	Be	
S12B	5	123w	1			Bj	
S12C	2	46-123	6	Be	Be	Be	
S12D	12	110-94	4			Bd	
S12E	5	123w	1			Bj	
S12F	8	123w	1			Bj	
S12G	5	123-6	3			Bj	
S12H	1	94-6-123w	3			Bj	
S12I	1	6-38	3			Bj	
S12J	8	123-6	3	Bj	n/a	Bj	
S12K	8	123-122	3	Bj	Bj	Bj	
S12L	3	NR	1			Bj	
S13A	no product	NR	4			Bd	
S13B	12	110	4			Bd	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
S13C	12	62	4			Bd	
S13D	2	6	3			Bj	
S13E	13	NR	4	Bd	Bd	Bd	Seq
S13F	3	38	1	Bj	Bj	Bj	
S13G	12	62	4			Bd	
S13H	13	110	4			Bd	
S13I	13	NR	4			Bd	
S13J	12	110	6			Bd	
S13K	12	122	7			Bd	
S13L	no product	NR	5			Bd	
S14A	12	110	4			Bd	
S14B	1	6w	3			Bj	
S14C	1	6	4	Bd	Bd	Bd	
S14D	17	94	11	Be	Be	Be	
S14E	12	NR	4			Bd	
S14F	12	110	4			Bd	
S14G	14	46w	8			Be	
S14H	14	76	9			Be	
S14I	10	NR	4	Bj	Bj	Bj	
S14J	8	NR	1			Bj	
S14K	1	NR	3			Bj	
S14L	12	110	4			Bd	
S15A	11	NR	4	Bj	Bj	Bj	Seq
S15B	14	46-110-123	11			Be	
S15C	2	6-110	3			Bj	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
S15D	2	123-6	1			Bj	
S15E	14	31-76-94	11			Be	
S15F	17	94-123	11			Be	
S15G	17	94-123	11			Be	
S15H	15	31	9	Be	Be	Be	Seq
S15I	17	94-123	11			Be	
S15J	14	76	8			Be	
S15K	2	6-38	3	Bj	Bj	Bj	
S15L	12	122-38	4			Bd	
S16A	3	38-6	1			Bj	
S16B	4	38	1	Bj	Bj	Bj	
S16C	12	110w	4			Bd	
S16D	18	94-123	9			Be	
S16E	1	6-38	3			Bj	
S16F	12	122-38	4			Bd	
S16G	4	38-6	1	Bj	Bj	Bj	
S16H	12	110	4			Bd	
S16I	3	38w	1			Bj	
S16J	12	110w	4			Bd	
S16K	17	94-123	8			Be	
S16L	12	110-31	4			Bd	
S17A	8	6w	1			Bj	
S17B	3	62	1			Bj	
S17C	8	123	3			Bj	
S17D	12	122w	4			Bd	

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
S17E	3	62-38	3			Bj	
S17F	8	123-6w	1			Bj	
S17G	8	123-6w	1			Bj	
S17H	12	94	4			Bd	
S17I	12	122	4			Bd	
S17J	8	123	3			Bj	
S17K	3	6-38	3			Bj	
S17L	5	123-6	3			Bj	
S18A	14	76	9			Be	
S18B	1	6-94	3			Bj	
S18C	15	94-123	6			Be	
S18D	15	31	8			Be	
S18E	1	6-38-94	6			Bj	
S18F	15	94-123	8			Be	
S18G	14	76	9			Be	
S18H	12	122	4			Bd	
S18I	15	94-123	9			Be	
S18J	14	46	9			Be	
S18K	no product	NR	4			Bd	
S18L	14	76-94	11			Be	
USDA110	12	110	4	Bd	Bd	Bd	Seq
USDA122	12	122	4	Bd	Bd	Bd	Seq
USDA123	5	123	1	Bj	Bj	Bj	Seq
USDA130	14	130	8	Be	Be	Be	
USDA135	6	135	4	Bj	Bj	Bj	Seq

Table A.1: continued

Culture	RFLP group[1]	Serogroup	FAME group[2]	16S sequencing[3]	ITS sequencing[4]	Consensus identity[5]	Note
USDA138	1	6	1	Bj	Bj	Bj	
USDA31	14	31	8	Be	Be	Be	Seq
USDA38	3	38	1	Bj	Bj	Bj	
USDA46	16	46	11	Be	Be	Be	
USDA62	12	62	4	Bd	Bd	Bd	
USDA76	14	76	11	Be	Be	Be	Seq
USDA94	15	94	8	Be	Be	Be	Seq

[1] Restriction fragment length polymorphism
[2] Fatty Acid Methyl Ester
[3] 16S rRNA sequencing
[4] Internal transcribed spacer sequencing
[5] Species identified for the accession as a consensus of results generated from listed phentoypic and genotypic methods
[6] Sequenced Bj: *Bradyrhizobium japonicum*
Bd: *Bradyrhizobium diazoefficiens*
Be: *Bradyrhizobium elkanii*
n/a: No yield from the test

Appendix B

SUPPLEMENTARY FILES FOR CHAPTER 3

Table B.1: Genome assemblies in Bradybase imported from RefSeq or UDBCC.

Name	Source	Accession
B. diazoefficiens 110spc4 Genome Assembly ASM435935v1	RefSeq	*B. diazoefficiens* 110spc4
B. diazoefficiens 113-2 Genome Assembly ASM1339030v1	RefSeq	*B. diazoefficiens* 113-2
B. diazoefficiens 172S4 Genome Assembly ASM1160462v1	RefSeq	*B. diazoefficiens* 172S4
B. diazoefficiens 182_5 Genome Assembly ASM1661253v1	RefSeq	*B. diazoefficiens* 182_5
B. diazoefficiens 36_1 Genome Assembly ASM1661688v1	RefSeq	*B. diazoefficiens* 36_1
B. diazoefficiens 38_8 Genome Assembly ASM1661623v1	RefSeq	*B. diazoefficiens* 38_8
B. diazoefficiens 41_2 Genome Assembly ASM1661642v1	RefSeq	*B. diazoefficiens* 41_2
B. diazoefficiens 65_7 Genome Assembly ASM1659985v1	RefSeq	*B. diazoefficiens* 65_7
B. diazoefficiens CCBAU 41267 Genome Assembly 41267	RefSeq	*B. diazoefficiens* CCBAU 41267
B. diazoefficiens F07S3 Genome Assembly ASM1416347v1	RefSeq	*B. diazoefficiens* F07S3
B. diazoefficiens H12S4 Genome Assembly ASM1416343v1	RefSeq	*B. diazoefficiens* H12S4
B. diazoefficiens HF08 Genome Assembly ASM1416345v1	RefSeq	*B. diazoefficiens* HF08
B. diazoefficiens HH15 Genome Assembly ASM1416341v1	RefSeq	*B. diazoefficiens* HH15
B. diazoefficiens Is-1 Genome Assembly ASM128058v1	RefSeq	*B. diazoefficiens* Is-1
B. diazoefficiens K01E Genome Assembly	UDBCC[1]	*B. diazoefficiens* K01E
B. diazoefficiens K07G Genome Assembly	UDBCC	*B. diazoefficiens* K07G
B. diazoefficiens K09F Genome Assembly	UDBCC	*B. diazoefficiens* K09F
B. diazoefficiens N03B Genome Assembly	UDBCC	*B. diazoefficiens* N03B
B. diazoefficiens NK6 Genome Assembly ASM154969v1	RefSeq	*B. diazoefficiens* NK6
B. diazoefficiens S13E Genome Assembly	UDBCC	*B. diazoefficiens* S13E
B. diazoefficiens S14C Genome Assembly	UDBCC	*B. diazoefficiens* S14C
B. diazoefficiens SEMIA 5080 Genome Assembly ASM64859v2	RefSeq	*B. diazoefficiens* SEMIA 5080
B. diazoefficiens SZCCT0113 Genome Assembly ASM1812964v1	RefSeq	*B. diazoefficiens* SZCCT0113
B. diazoefficiens SZCCT0122 Genome Assembly ASM1812970v1	RefSeq	*B. diazoefficiens* SZCCT0122
B. diazoefficiens SZCCT0126 Genome Assembly ASM1812975v1	RefSeq	*B. diazoefficiens* SZCCT0126

Table B.1: Continued

Name	Source	Accession
B. diazoefficiens SZCCT0130 Genome Assembly ASM1812978v1	RefSeq	*B. diazoefficiens* SZCCT0130
B. diazoefficiens SZCCT0132 Genome Assembly ASM1812980v1	RefSeq	*B. diazoefficiens* SZCCT0132
B. diazoefficiens SZCCT0137 Genome Assembly ASM1812992v1	RefSeq	*B. diazoefficiens* SZCCT0137
B. diazoefficiens SZCCT0138 Genome Assembly ASM1812996v1	RefSeq	*B. diazoefficiens* SZCCT0138
B. diazoefficiens SZCCT0139 Genome Assembly ASM1812993v1	RefSeq	*B. diazoefficiens* SZCCT0139
B. diazoefficiens SZCCT0235 Genome Assembly ASM1813030v1	RefSeq	*B. diazoefficiens* SZCCT0235
B. diazoefficiens SZCCT0287 Genome Assembly ASM1813048v1	RefSeq	*B. diazoefficiens* SZCCT0287
B. diazoefficiens SZCCT0340 Genome Assembly ASM1813058v1	RefSeq	*B. diazoefficiens* SZCCT0340
B. diazoefficiens SZCCT0341 Genome Assembly ASM1813063v1	RefSeq	*B. diazoefficiens* SZCCT0341
B. diazoefficiens SZCCT0406 Genome Assembly ASM1813089v1	RefSeq	*B. diazoefficiens* SZCCT0406
B. diazoefficiens SZCCT0423 Genome Assembly ASM1813101v1	RefSeq	*B. diazoefficiens* SZCCT0423
B. diazoefficiens SZCCT0435 Genome Assembly ASM1813114v1	RefSeq	*B. diazoefficiens* SZCCT0435
B. diazoefficiens SZCCT0440 Genome Assembly ASM1813116v1	RefSeq	*B. diazoefficiens* SZCCT0440
B. diazoefficiens SZCCT0449 Genome Assembly ASM1813118v1	RefSeq	*B. diazoefficiens* SZCCT0449
B. diazoefficiens USDA 110 Genome Assembly ASM1136v1	RefSeq	*B. diazoefficiens* USDA 110
B. diazoefficiens USDA 110 Genome Assembly 2 ASM164267v1	RefSeq	*B. diazoefficiens* USDA 110
B. diazoefficiens USDA 122 Genome Assembly ASM47302v1	RefSeq	*B. diazoefficiens* USDA 122
B. diazoefficiens USDA 122 Genome Assembly 2 ASM190831v1	RefSeq	*B. diazoefficiens* USDA 122
B. diazoefficiens XF7 Genome Assembly ASM318384v2	RefSeq	*B. diazoefficiens* XF7
B. diazoefficiens Y21 Genome Assembly ASM253199v1	RefSeq	*B. diazoefficiens* Y21
B. elkanii 587 Genome Assembly BelkAss1.0	RefSeq	*B. elkanii* 587
B. elkanii BLY3-8 Genome Assembly ASM171820v1	RefSeq	*B. elkanii* BLY3-8
B. elkanii BLY6-1 Genome Assembly ASM171818v1	RefSeq	*B. elkanii* BLY6-1
B. elkanii BR29 Genome Assembly ASM415295v1	RefSeq	*B. elkanii* BR29
B. elkanii CCBAU 05737 Genome Assembly 05737	RefSeq	*B. elkanii* CCBAU 05737

Table B.1: Continued

Name	Source	Accession
B. elkanii CCBAU 43297 Genome Assembly 43297	RefSeq	*B. elkanii* CCBAU 43297
B. elkanii K02K Genome Assembly	UDBCC	*B. elkanii* K02K
B. elkanii K03D Genome Assembly	UDBCC	*B. elkanii* K03D
B. elkanii K03I Genome Assembly	UDBCC	*B. elkanii* K03I
B. elkanii NBRC 14791 Genome Assembly ASM653966v1	RefSeq	*B. elkanii* NBRC 14791
B. elkanii S05J Genome Assembly	UDBCC	*B. elkanii* S05J
B. elkanii S07J Genome Assembly	UDBCC	*B. elkanii* S07J
B. elkanii S15H Genome Assembly	UDBCC	*B. elkanii* S15H
B. elkanii SEMIA 5019 Genome Assembly ASM1339274v1	RefSeq	*B. elkanii* SEMIA 5019
B. elkanii SEMIA 587 Genome Assembly ASM1339276v1	RefSeq	*B. elkanii* SEMIA 587
B. elkanii Semia 938 Genome Assembly ASM529809v1	RefSeq	*B. elkanii* Semia 938
B. elkanii SZCCT0424 Genome Assembly ASM1813104v1	RefSeq	*B. elkanii* SZCCT0424
B. elkanii TnphoA 33 Genome Assembly BE_33	RefSeq	*B. elkanii* TnphoA 33
B. elkanii UASWS1015 Genome Assembly UASWS1015 1.0	RefSeq	*B. elkanii* UASWS1015
B. elkanii USDA 130 Genome Assembly ASM1787670v1	RefSeq	*B. elkanii* USDA 130
B. elkanii USDA 406 Genome Assembly ASM1783412v1	RefSeq	*B. elkanii* USDA 406
B. elkanii USDA 61 Genome Assembly ASM1287105v1	RefSeq	*B. elkanii* USDA 61
B. elkanii USDA 76 Genome Assembly ASM37914v1	RefSeq	*B. elkanii* USDA 76
B. elkanii USDA 94 Genome Assembly	UDBCC	*B. elkanii* USDA 94
B. elkanii USDA 94 Genome Assembly ASM51922v1	RefSeq	*B. elkanii* USDA 94
B. elkanii USDA31 Genome Assembly	UDBCC	*B. elkanii* USDA31
B. elkanii WSM1741 Genome Assembly ASM47296v1	RefSeq	*B. elkanii* WSM1741
B. elkanii WSM2783 Genome Assembly ASM47286v1	RefSeq	*B. elkanii* WSM2783
B. huanghuaihaiense CGMCC 1.10948 Genome Assembly ASM783063v1	RefSeq	*B. huanghuaihaiense* CGMCC 1.10948
B. japonicum 22 Genome Assembly ASM48242v1	RefSeq	*B. japonicum* 22

Table B.1: Continued

Name	Source	Accession
B. japonicum 5038 Genome Assembly ASM1375273v1	RefSeq	*B. japonicum* 5038
B. japonicum 5873 Genome Assembly ASM986481v1	RefSeq	*B. japonicum* 5873
B. japonicum CCBAU 15354 Genome Assembly 15354	RefSeq	*B. japonicum* CCBAU 15354
B. japonicum CCBAU 15517 Genome Assembly 15517	RefSeq	*B. japonicum* CCBAU 15517
B. japonicum CCBAU 15618 Genome Assembly 15618	RefSeq	*B. japonicum* CCBAU 15618
B. japonicum CCBAU 25435 Genome Assembly 25435	RefSeq	*B. japonicum* CCBAU 25435
B. japonicum CCBAU 83623 Genome Assembly 83623	RefSeq	*B. japonicum* CCBAU 83623
B. japonicum E109 Genome Assembly ASM80731v1	RefSeq	*B. japonicum* E109
B. japonicum FN1 Genome Assembly ASM103818v1	RefSeq	*B. japonicum* FN1
B. japonicum in8p8 Genome Assembly ASM42684v1	RefSeq	*B. japonicum* in8p8
B. japonicum Is-34 Genome Assembly ASM77386v1	RefSeq	*B. japonicum* Is-34
B. japonicum is5 Genome Assembly ASM42130v1	RefSeq	*B. japonicum* is5
B. japonicum J5 Genome Assembly ASM188769v1	RefSeq	*B. japonicum* J5
B. japonicum N03G Genome Assembly	UDBCC	*B. japonicum* N03G
B. japonicum NBRC 14783 Genome Assembly ASM653964v1	RefSeq	*B. japonicum* NBRC 14783
B. japonicum S04E Genome Assembly	UDBCC	*B. japonicum* S04E
B. japonicum S06K Genome Assembly	UDBCC	*B. japonicum* S06K
B. japonicum S11L Genome Assembly	UDBCC	*B. japonicum* S11L
B. japonicum S15A Genome Assembly	UDBCC	*B. japonicum* S15A
B. japonicum SEMIA 5079 Genome Assembly ASM66193v1	RefSeq	*B. japonicum* SEMIA 5079
B. japonicum SZCCT0148 Genome Assembly ASM1812998v1	RefSeq	*B. japonicum* SZCCT0148
B. japonicum SZCCT0153 Genome Assembly ASM1812999v1	RefSeq	*B. japonicum* SZCCT0153
B. japonicum SZCCT0231 Genome Assembly ASM1813024v1	RefSeq	*B. japonicum* SZCCT0231
B. japonicum SZCCT0280 Genome Assembly ASM1813036v1	RefSeq	*B. japonicum* SZCCT0280
B. japonicum SZCCT0395 Genome Assembly ASM1813073v1	RefSeq	*B. japonicum* SZCCT0395

Table B.1: Continued

Name	Source	Accession
B. japonicum SZCCT0401 Genome Assembly ASM1813083v1	RefSeq	*B. japonicum* SZCCT0401
B. japonicum SZCCT0402 Genome Assembly ASM1813079v1	RefSeq	*B. japonicum* SZCCT0402
B. japonicum SZCCT0403 Genome Assembly ASM1813092v1	RefSeq	*B. japonicum* SZCCT0403
B. japonicum UBMA197 Genome Assembly ASM210893v1	RefSeq	*B. japonicum* UBMA197
B. japonicum USDA 123 Genome Assembly	UDBCC	*B. japonicum* USDA 123
B. japonicum USDA 123 Genome Assembly ASM48252v1	RefSeq	*B. japonicum* USDA 123
B. japonicum USDA 135 Genome Assembly	UDBCC	*B. japonicum* USDA 135
B. japonicum USDA 135 Genome Assembly ASM47294v1	RefSeq	*B. japonicum* USDA 135
B. japonicum USDA 300 Genome Assembly ASM1783198v1	RefSeq	*B. japonicum* USDA 300
B. japonicum USDA 38 Genome Assembly ASM47274v1	RefSeq	*B. japonicum* USDA 38
B. japonicum USDA 500 Genome Assembly ASM1783194v1	RefSeq	*B. japonicum* USDA 500
B. japonicum USDA 6 Genome Assembly 2 ASM28437v1	RefSeq	*B. japonicum* USDA 6
B. japonicum USDA 6 Genome Assembly ASM47298v1	RefSeq	*B. japonicum* USDA 6
B. liaoningense CCBAU 05525 Genome Assembly 05525	RefSeq	*B. liaoningense* CCBAU 05525
B. liaoningense CCBAU 83689 Genome Assembly 83689	RefSeq	*B. liaoningense* CCBAU 83689
B. liaoningense CCNWSX0360 Genome Assembly ASM159599v1	RefSeq	*B. liaoningense* CCNWSX0360
B. liaoningense SZCCT0008 Genome Assembly ASM1812946v1	RefSeq	*B. liaoningense* SZCCT0008
B. liaoningense SZCCT0133 Genome Assembly ASM1812985v1	RefSeq	*B. liaoningense* SZCCT0133
B. liaoningense SZCCT0154 Genome Assembly ASM1813003v1	RefSeq	*B. liaoningense* SZCCT0154
B. liaoningense SZCCT0233 Genome Assembly ASM1813027v1	RefSeq	*B. liaoningense* SZCCT0233
B. liaoningense SZCCT0285 Genome Assembly ASM1813044v1	RefSeq	*B. liaoningense* SZCCT0285
B. liaoningense SZCCT0293 Genome Assembly ASM1813052v1	RefSeq	*B. liaoningense* SZCCT0293
B. liaoningense SZCCT0337 Genome Assembly ASM1813056v1	RefSeq	*B. liaoningense* SZCCT0337
B. liaoningense SZCCT0342 Genome Assembly ASM1813059v1	RefSeq	*B. liaoningense* SZCCT0342
B. liaoningense SZCCT0347 Genome Assembly ASM1813068v1	RefSeq	*B. liaoningense* SZCCT0347

Table B.1: Continued

Name	Source	Accession
B. liaoningense SZCCT0396 Genome Assembly ASM1813072v1	RefSeq	*B. liaoningense* SZCCT0396
B. liaoningense SZCCT0397 Genome Assembly ASM1813085v1	RefSeq	*B. liaoningense* SZCCT0397
B. liaoningense SZCCT0399 Genome Assembly ASM1813078v1	RefSeq	*B. liaoningense* SZCCT0399
B. liaoningense SZCCT0400 Genome Assembly ASM1813082v1	RefSeq	*B. liaoningense* SZCCT0400
B. liaoningense SZCCT0420 Genome Assembly ASM1813095v1	RefSeq	*B. liaoningense* SZCCT0420
B. ottawaense GAS524 Genome Assembly IMG-taxon 2693430034 annotated assembly	RefSeq	*B. ottawaense* GAS524
B. ottawaense L2 Genome Assembly ASM253202v1	RefSeq	*B. ottawaense* L2
B. ottawaense OO99 Genome Assembly ASM227813v2	RefSeq	*B. ottawaense* OO99
B. ottawaense SZCCT0046 Genome Assembly ASM1813123v1	RefSeq	*B. ottawaense* SZCCT0046
B. ottawaense SZCCT0234 Genome Assembly ASM1813026v1	RefSeq	*B. ottawaense* SZCCT0234
B. ottawaense SZCCT0284 Genome Assembly ASM1813042v1	RefSeq	*B. ottawaense* SZCCT0284
B. ottawaense SZCCT0286 Genome Assembly ASM1813040v1	RefSeq	*B. ottawaense* SZCCT0286
B. ottawaense USDA 4 Genome Assembly ASM47272v1	RefSeq	*B. ottawaense* USDA 4
B. yuanmingense 3051 Genome Assembly ASM975837v1	RefSeq	*B. yuanmingense* 3051
B. yuanmingense BR3267 Genome Assembly ABySSBR3267	RefSeq	*B. yuanmingense* BR3267
B. yuanmingense CCBAU 05623 Genome Assembly 05623	RefSeq	*B. yuanmingense* CCBAU 05623
B. yuanmingense CCBAU 10071 Genome Assembly IMG-taxon 2617270741 annotated assembly	RefSeq	*B. yuanmingense* CCBAU 10071
B. yuanmingense CCBAU 25021 Genome Assembly 25021	RefSeq	*B. yuanmingense* CCBAU 25021
B. yuanmingense CCBAU 35157 Genome Assembly 35157	RefSeq	*B. yuanmingense* CCBAU 35157
B. yuanmingense CGMCC 1.3531 Genome Assembly ASM783057v1	RefSeq	*B. yuanmingense* CGMCC 1.3531
B. yuanmingense P10 130 Genome Assembly ASM402228v3	RefSeq	*B. yuanmingense* P10 130

[1] University of Delaware Bradyrhizobium Culture Collection

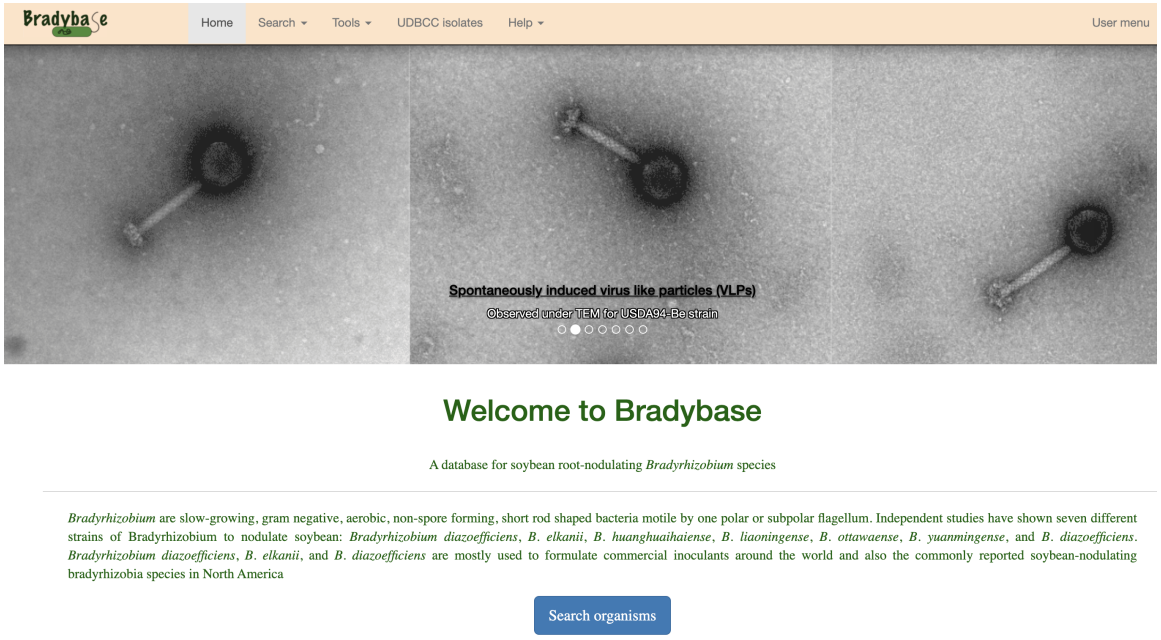
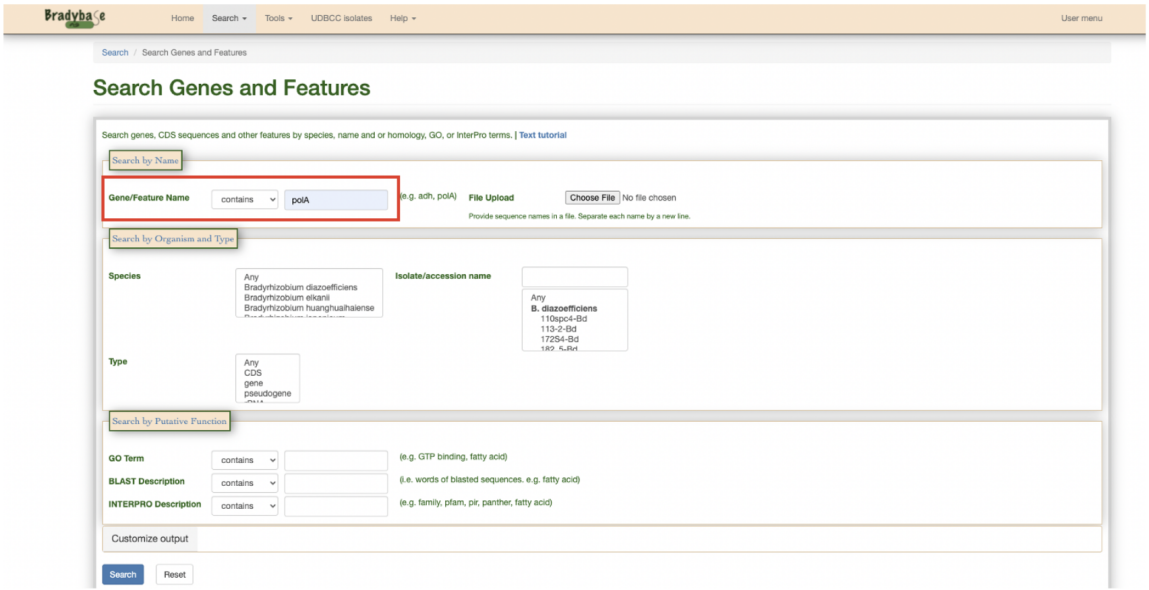


Figure B.1: Homepage of Bradybase. The menu bar at the top provides links to data search pages and tools used in Bradybase. Also included is a link to University of Delaware *Bradyrhizobium* Culture Collection (UDBCC) accessions information that can be directly downloaded for reference.

A



B

#	Species	Organism	Name	Length	Type
1	Bradyrhizobium diazoefficiens	172S4-Bd	polA		gene
2	Bradyrhizobium diazoefficiens	CCBAU 41267-Bd	polA		gene
3	Bradyrhizobium diazoefficiens	113-2-Bd	polA		gene
4	Bradyrhizobium diazoefficiens	110spc4-Bd	polA		gene
5	Bradyrhizobium diazoefficiens	H12S4-Bd	polA		gene
6	Bradyrhizobium diazoefficiens	F07S3-Bd	polA		gene
7	Bradyrhizobium diazoefficiens	HH15-Bd	polA		gene
8	Bradyrhizobium diazoefficiens	HF08-Bd	polA		gene
9	Bradyrhizobium diazoefficiens	USDA 110-Bd	polA		gene
10	Bradyrhizobium diazoefficiens	Is-1-Bd	polA		gene
11	Bradyrhizobium diazoefficiens	SEMIA 5080-Bd	polA		gene
12	Bradyrhizobium diazoefficiens	Y21-Bd	polA		gene
13	Bradyrhizobium diazoefficiens	NK6-Bd	polA		gene
14	Bradyrhizobium diazoefficiens	XF7-Bd	polA		gene
15	Bradyrhizobium diazoefficiens	USDA 122-Bd	polA		gene
16	Bradyrhizobium diazoefficiens	USDA 110-Bd	polA		gene
17	Bradyrhizobium japonicum	CCBAU 25435-Bj	polA		gene
18	Bradyrhizobium japonicum	CCBAU 15618-Bj	polA		gene
19	Bradyrhizobium japonicum	CCBAU 15517-Bj	polA		gene
20	Bradyrhizobium japonicum	CCBAU 83623-Bj	polA		gene

Figure B.2: Genes and features search page of Bradybase. Genomic features (CDS, gene, pseudogene, rRNA, tRNA, and tmRNA) can be searched by name(s), species(s), isolate/accession name(s), type(s) of feature, and gene functional annotation (Gene Ontology (GO), BLAST hit description and InterPro matches description). A) Using the search tool for genes named polA (DNA Polymerase A) from all accessions in the Bradybase. B) Results for the search. The result table can be downloaded in csv form from the Table link. All the polA residues can be downloaded from Fasta link. The fasta identifier of each residue includes the gene ID and name of the accession it belongs to.

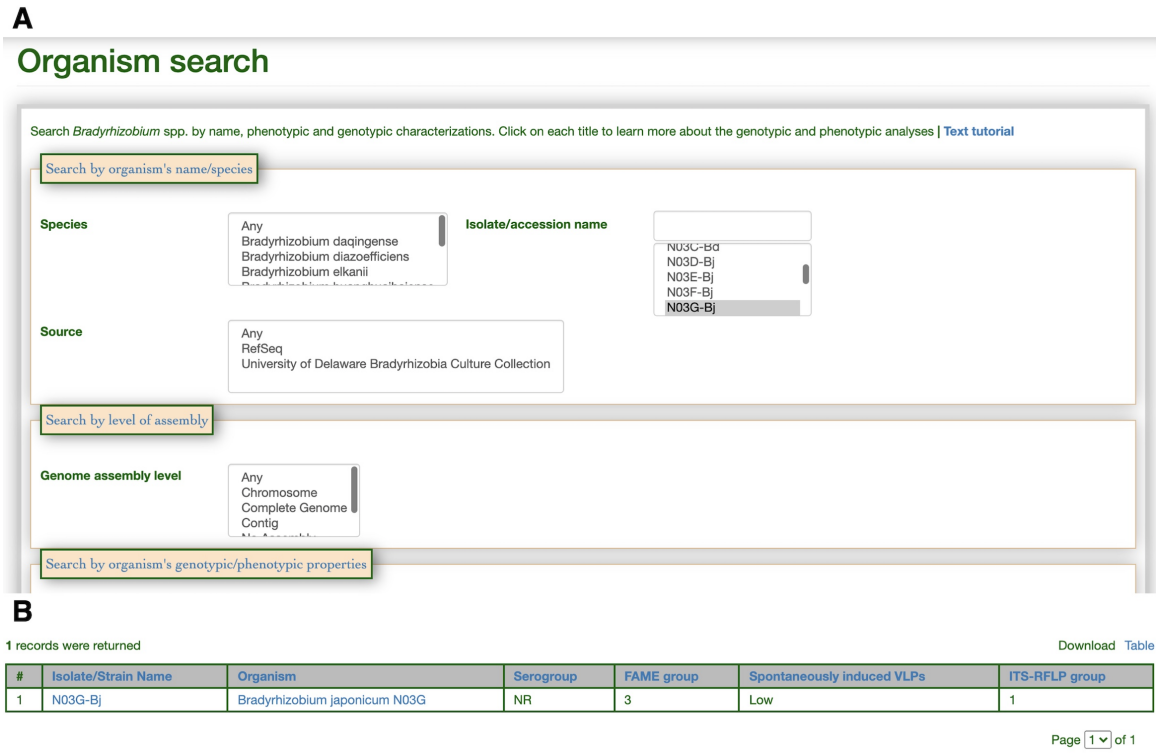


Figure B.3: Organism search interface of Bradybase. Bradybase accessions can be searched according to species name(s), isolate/accession name(s), source(s) of the accession, genome assembly level(s) if available, serogroup(s), FAME group(s), and level of production of spontaneously induced virus-like particles. A) Search tool used to search accession N03G-Bj B) Results for the search. Results can be downloaded in csv format using Table link. Users can access the page for N03G-Bj accession using either isolate/accession name or organism link.

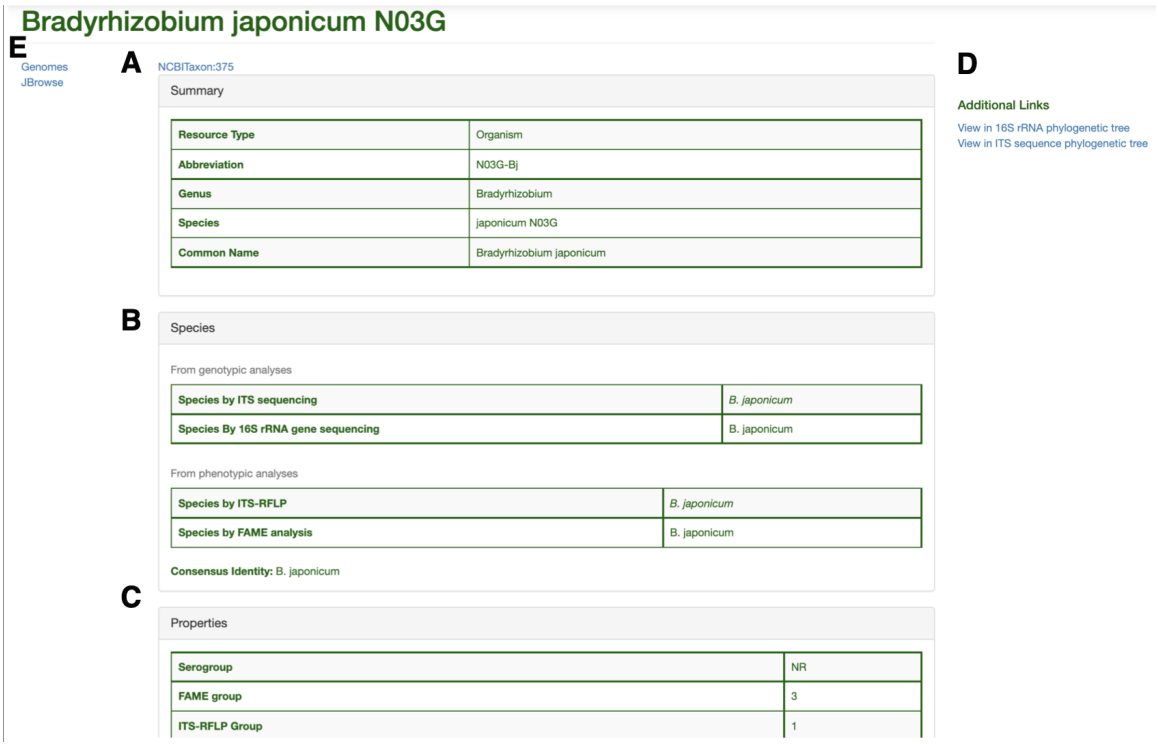


Figure B.4: Organism page for *Bradyrhizobium japonicum* N03G. A) Taxonomy for the species B) Species identified from various genotypic and phenotypic analyses as listed C) Phenotypic and genotypic properties for the organism. D) Additional links provide links to different phylogenetic trees for the organism. E) Genomes lists all available genome assemblies for the organism in Bradybase and JBrowse links the organism to JBrowse page to visualize its genome.

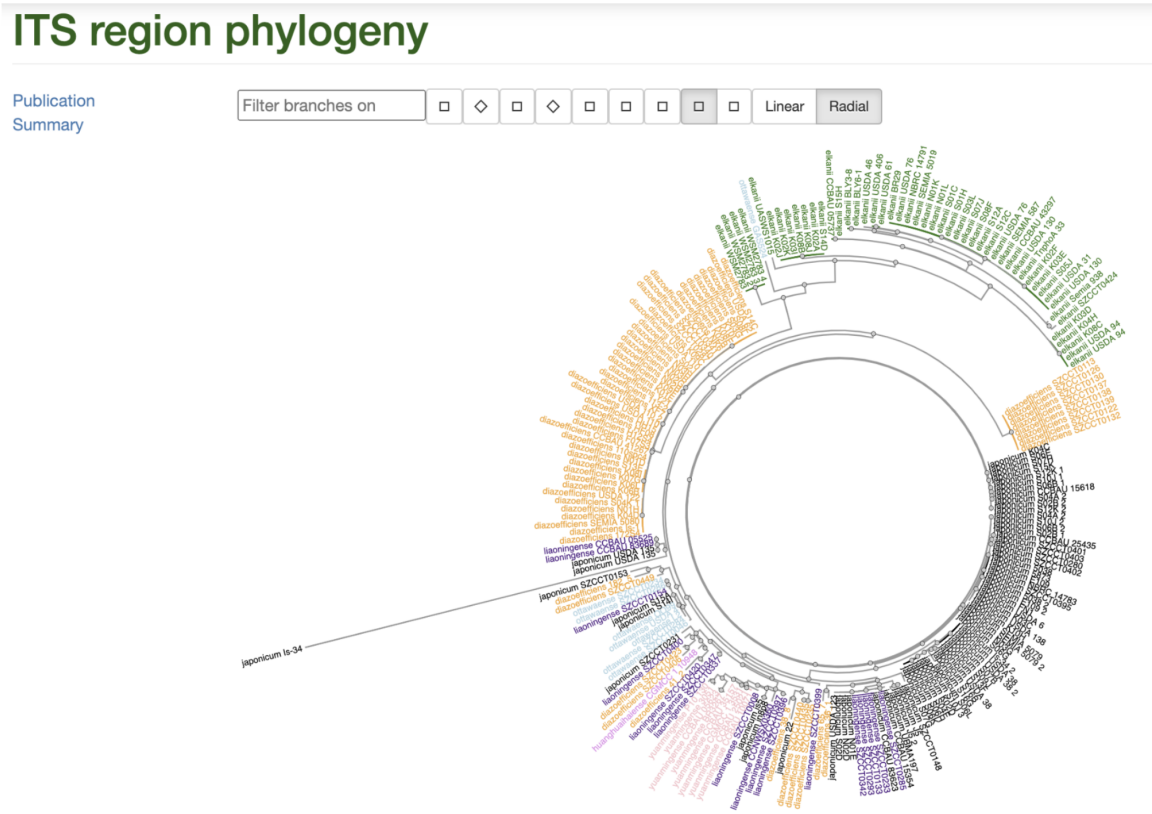


Figure B.5: An instance of phylogenetic tree for Bradybase accessions. It shows a phylogenetic tree built with available ITS sequences in Bradybase. ITS sequences are available for all RefSeq imported accessions and UDBCC accessions which underwent ITS sequencing or whole genome assembly. Color represents each soybean root-nodulating *Bradyrhizobium* spp. Orange: *Bradyrhizobium diazoefficiens*, green: *B. elkanii*, black: *B. japonicum*, violet: *B. huanghuaihaiense*, purple: *B. liaoningense*, light blue: *B. ottwaense*, and pink: *B. yuanmingense*.

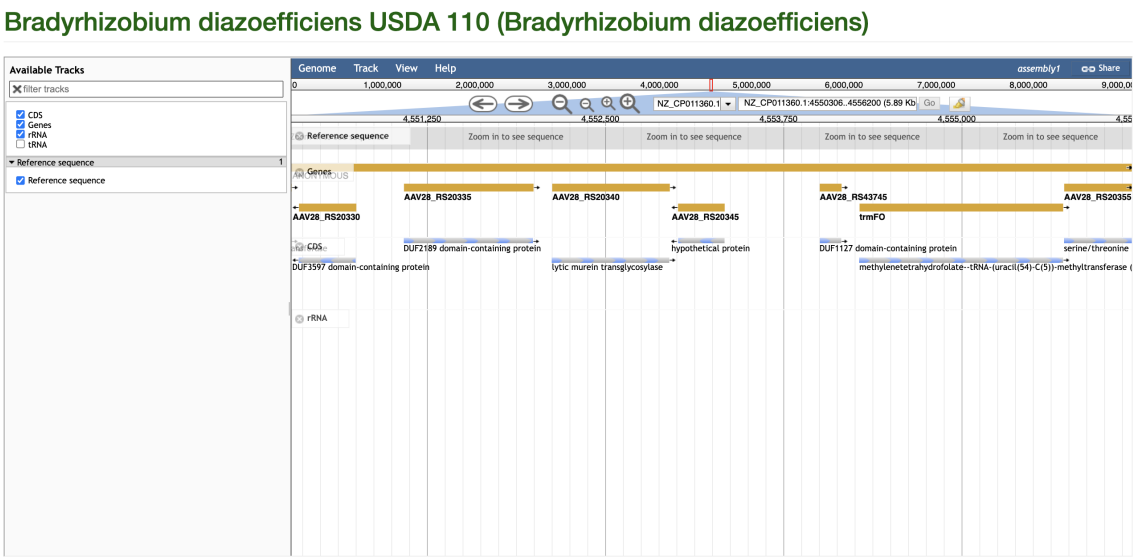


Figure B.6: A Jbrowse instance for genome from *Bradyrhizobium diazoefficiens* USDA 110 accession. All available tracks: reference sequence, Genes, Coding Sequence (CDS), and mRNA are displayed.

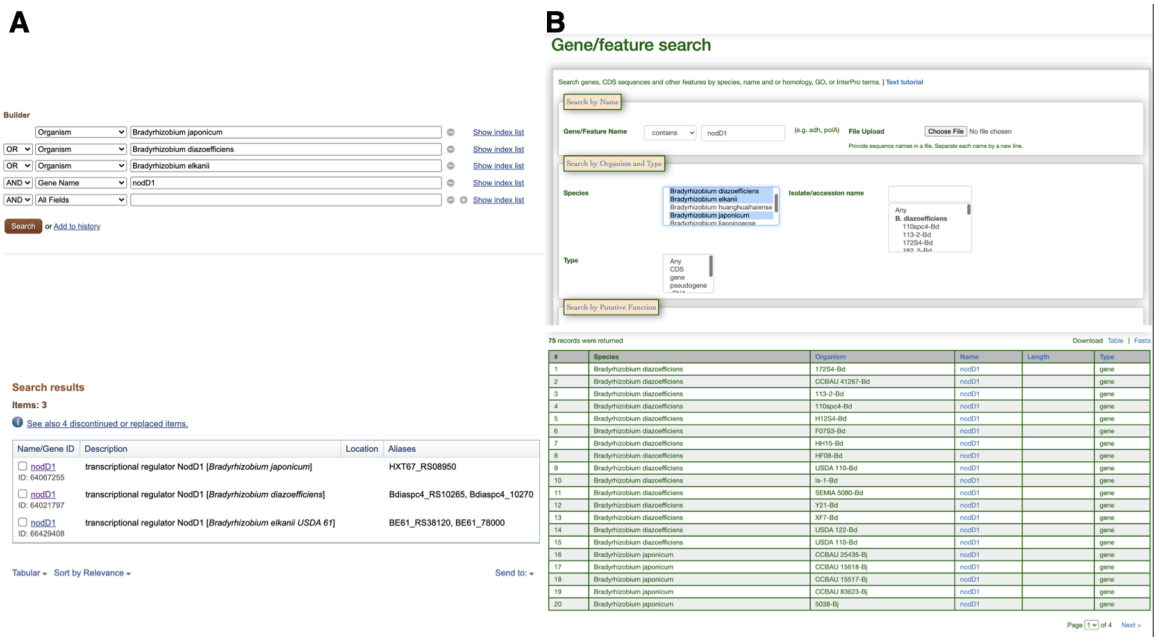


Figure B.7: Comparison of gene search interfaces between A) NCBI and B) Bradybase to search nodD1 genes from three species: *B. diazoefficiens*, *B. elkanii* and *B. japonicum*. A. Query is built by providing names of the three species, Boolean operator AND and name of the *nodD1* gene. The result provides a total of three *nodD1* gene representatives, one from each species. B. Search interface in Bradybase requires users to input the gene name in the Gene/feature name box and select the three species from the species menu. The result gives all *nodD1* genes from each accession belonging to the species, all of which can be downloaded for further analyses.

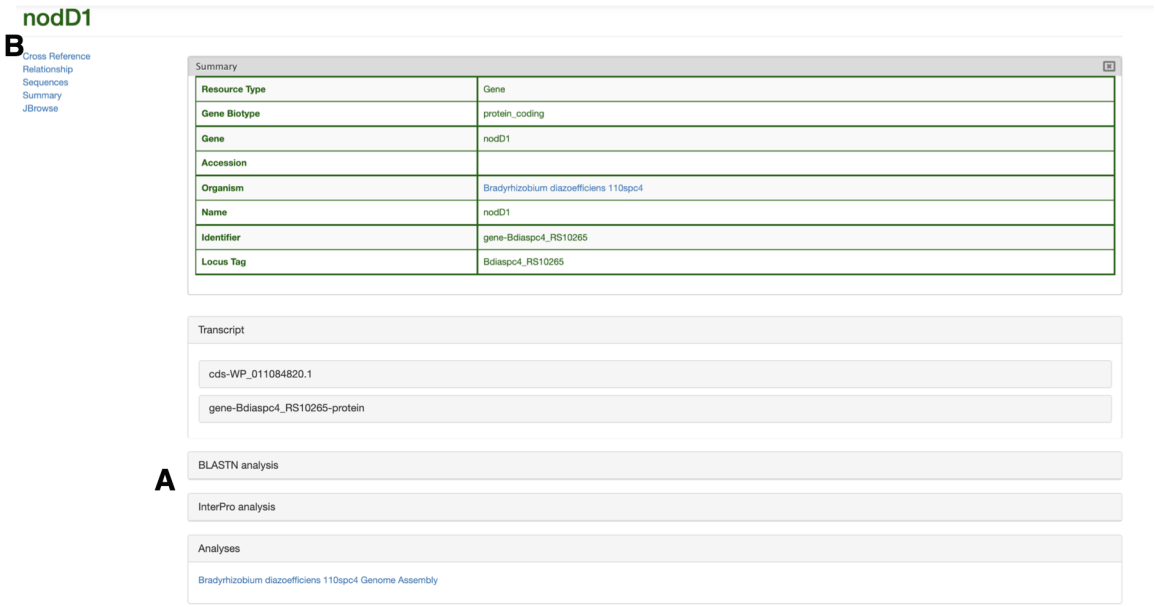


Figure B.8: Gene page for nodD1 gene from *Bradyrhizobium diazoefficiens* 110spc4' accession. Each page contains its transcript information, and links to its source organism and genome assembly. A) Homology and functional annotations generated using BLASTN and InterPro analyses are also incorporated. B) Additional links to retrieve its sequence, and locate the gene in the genome using JBrowse are provided.

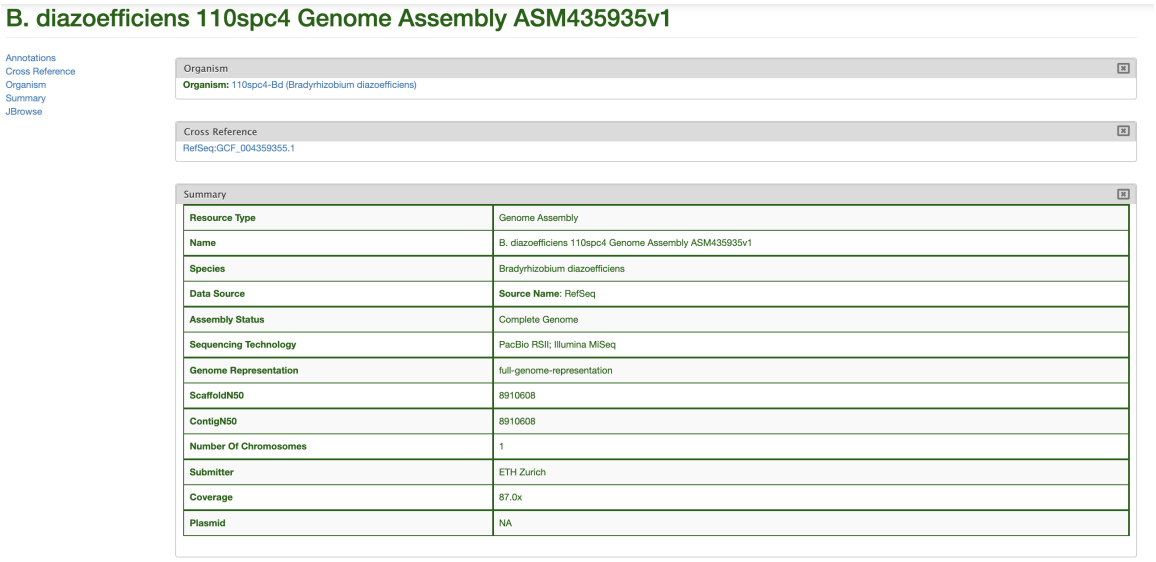


Figure B.9: Page for genome assembly of *Bradyrhizobium diazoefficiens* 110spc4 accession. Metadata such as accession name, assembly level and sequencing technology used for the assembly are included. JBrowse link and RefSeq cross-reference allow genome visualization, and assembly files download and further analyses from RefSeq.