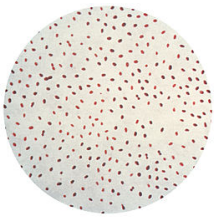
Introduction



Brucella is a genus of Gram-negative bacteria, named after David Bruce (1855–1931). They are small (0.5 to 0.7 by 0.6 to 1.5 μm), nonencapsulated, nonmotile, facultatively intracellular coccobacilli.

Brucella is the cause of brucellosis, which is a zoonosis transmitted by ingesting contaminated food (such as unpasteurized milk products), direct contact with an infected animal, or inhalation of aerosols. Transmission from human to human, for example through sexual intercourse or from mother to child, is exceedingly rare, but possible. Minimum infectious exposure is between 10 and 100 organisms.

The different species of *Brucella* are genetically very similar, although each has a slightly different host specificity. Hence, the NCBI taxonomy includes most *Brucella* species under *B. melitensis*.

The many names of brucellosis include (human disease/animal disease):

- Malta fever/Bang's disease
- Undulant fever/enzootic abortion
- Mediterranean fever/epizootic abortion
- Rock fever of Gibraltar/slinking of calves
- Gastric fever/ram epididymitis
- Contagious abortion/spontaneous abortion

Brucella neotomae was first isolated in 1957 from wood rats (Neotoma lepida) in North America and has been considered nonzoonotic. Brucella neotomae 5K33 (ATCC 23459) is the type strain and will be used for comparative analysis with other Brucella species.

Genome Assembly

Brucella neotomae 5K33 ran under job ID 2fee1c1b-f8bb-4781-bb20-f08fa66398e1 at PATRIC[1]. The assembly job started at 10/6/17, 2:53 PM and completed at 10/7/17, 4:41 PM, after 25h 47m 90011s. The auto assembly strategy was selected, and it runs BayesHammer [2] on short reads, followed by three assembly strategies that include Velvet [3], IDBA [4] and Spades [5], each of which is given an assembly score by ARAST, an in-house script. The minimum contig length was 120bp, and smaller contigs were not included in the assembly. The minimum contig coverage was 5, and contigs with less coverage were not included in the assembly. Also add reference to QUAST. ARAST ranked the Spades assembly best (**Table 1**).

The assembled genome has 11 contigs, with the total length of 3.33 Mbp and %GC of 57.3%.

Table 1. Assembly details for Brucella neotomae 5K33

Contigs	11	**GC Content**	57.3
Plasmids	0	**Contig L50**	1
Genome Length	3,329,623	**Contig N50**	1923503
Chromosomes	0		

Genome Annotation

The Genome Annotation Service in PATRIC [1] uses the RAST tool kit (RASTtk) [6] to provide annotation of genomic features. The job for Brucella neotomae 5K33 ran under job number 685fb8c1-348a-4cea-a298-496b037af1f8. The annotation job started at 10/9/17, 7:10 AM and completed at 10/9/17, 8:16 AM, after 1h 5m 3623s. The selected domain was Bacteria, the Taxonomy ID was 234.128. The Genetic code was 11. The taxonomy of Brucella neotomae 5K33 is:

> cellular organisms > Bacteria > Proteobacteria > Alphaproteobacteria > Rhizobiales > Brucellaceae > Brucella > Brucella neotomae > Brucella neotomae 5K33

The annotations includes 3434 CDS, 48 tRNAs and 3 rRNAs. All annotated genome features for this genome are summarized in **Table 2**.

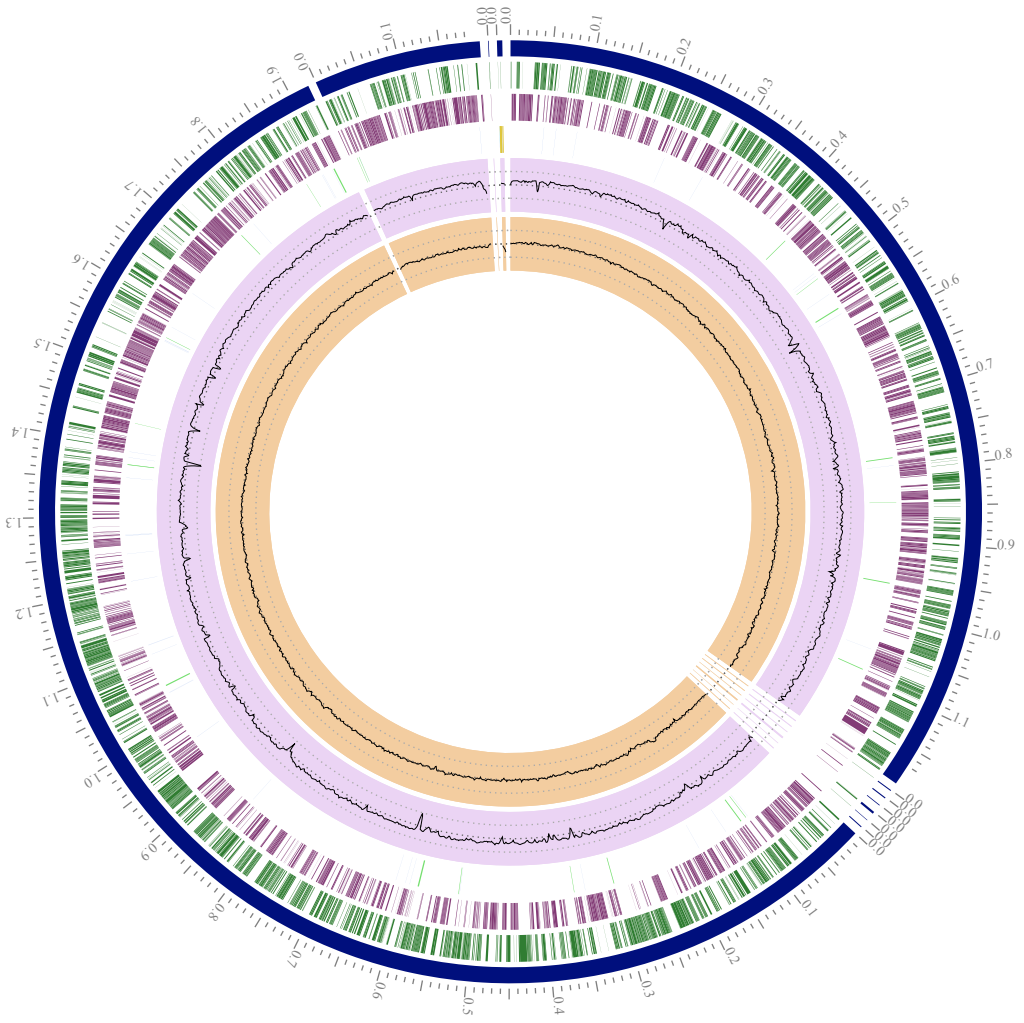
The functional annotation included 772 hypothetical proteins and 2662 proteins with functional assignments. Furthermore, 947 proteins were assigned EC numbers and 746 proteins were mapped to KEGG pathways. A breakdown of the proteins that have been annotated for this isolate is provided in **Table 3**.

Table 2. Annotated Genome Features	
	Count
CDS	3434
tRNA	48
pseudogene	23
rRNA	3
misc_RNA	1

Table 3. Protein Features	
	Count
Hypothetical proteins	772
Proteins with functional assignments	2662
Proteins with EC number assignments	947
Proteins with GO assignments	934
Proteins with Pathway assignments	746
Proteins with PATRIC genus-specific family (PLfam) assignments	3434
Proteins with PATRIC cross-genus family (PGfam) assignments	3434

A graphical display of the distribution of the CDS on each strand, the repeat_regions, GC content and GC skew are provided (**Figure 1**).

Figure 1



Genome Quality

Based on the analysis of genome assembly and annotation statistics and comparing them to other closely related genomes available in PATRIC for the same species, the overall genome seems to be of [high/medium/low] quality. The reasons affecting the quality of genome are: [genome qc flags].

Specialty Genes

PATRIC [1] annotation also includes identification of proteins that have homology to known virulence factors, drug targets, and antibiotic resistance genes (Reference the sources). A breakdown of the distribution is provided (**Table 4**).

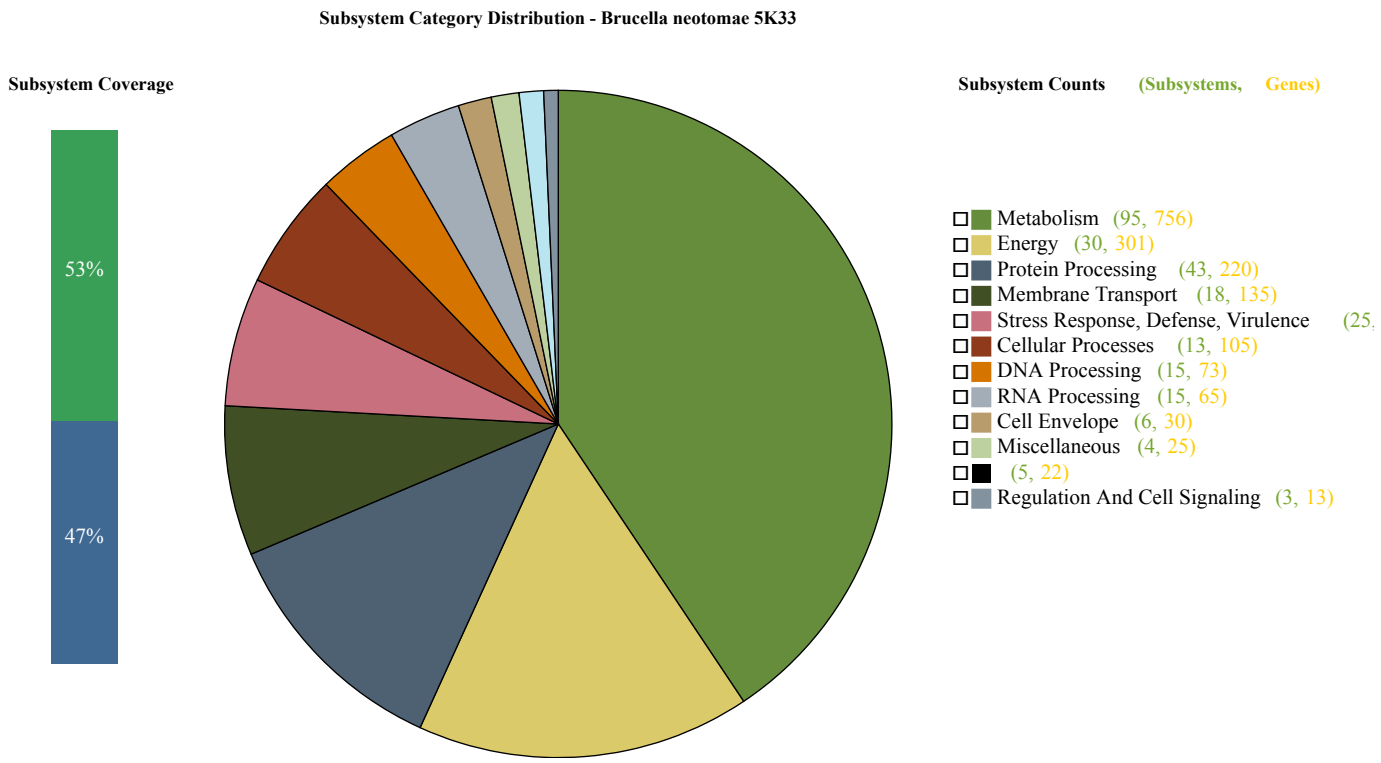
Table 4. Annotated Genome Features

	Source	Genes
Essential Gene	PATRIC	224
Virulence Factor	Victors	224
Human Homolog	Human	79
Virulence Factor	VFDB	44
Transporter	TCDB	25
Antibiotic Resistance	PATRIC	11
Drug Target	DrugBank	4
Antibiotic Resistance	CARD	3
Drug Target	TTD	1
Virulence Factor	PATRIC_VF	1

Subsystem Analysis

A subsystem is a set of protein functions that are related. Frequently, subsystems represent the collection of functional roles that make up a metabolic pathway, a complex (e.g., the ribosome), or a class of proteins (e.g., two-component signal-transduction proteins or AMR genes) [Overbeek]. An overview of the subsystems for this genome is provided in **Figure 2**.

Figure 2

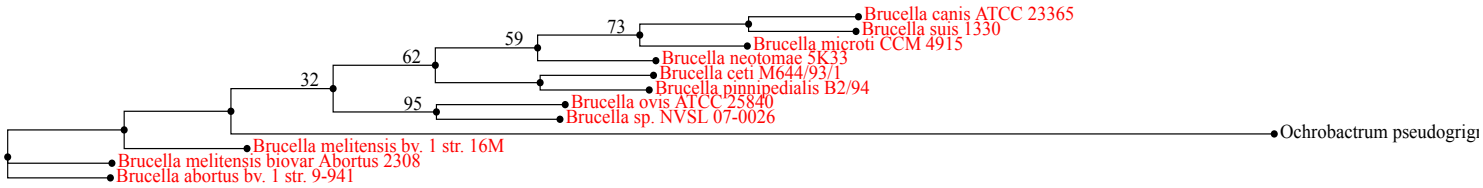


Phylogenetic Analysis

PATRIC [1] provides a phylogentic tree building service. The makes use of several third-party tools. These include, BLAST[1], MCL[2], Muscle[3], hmmbuild[4], hmmsearch[4], Gblocks[5], FastTree[6], and RAxML[7]. The pipeline begins with thegi set of ingroup genome protein files. These are filtered to remove duplicate species, resulting in a distinct-species subset of the ingroup genomes. This is done to reduce biasing the homolog sets with overrepresented species. BLAST searches are used to find bi-directional best hit protein pairs between genomes, and these bidirectional best hit pairs are clustered using MCL. Clusters containing members from at least half of the distinct genomes are chosen as initial, or seed, homolog sets. These seed sets are expanded to include members from all ingroup and outgroup taxa using tools from the HMMer suite. A hidden Markov Model (HMM) is built from each seed set using hmmbuild. These HMMs are used to search each genome, with hmmsearch, to find the best match from each genome for each homolog set model. The final, expanded, homolog sets are created from the hmmsearch results. Homolog sets representing fewer than 80% of ingroup genomes are removed. The remaining sets are aligned using Muscle, and the alignments are trimmed using GBlocks. The trimmed alignments are concatenated and this concatenated alignment is used to build the main tree with either RAxML or FastTree.

Comparison to Reference Genomes in the Same Genus

Figure 3



Unique Protein Families

In **Table 5** we present protein families that the isolate has which all reference genomes lack.

Table 5. Core Protein Families Missing in the Genome

Protein Family ID	Description
PGF_00025847	Nitrite reductase accessory protein NirV
PGF_00194700	hypothetical protein
PGF_00458082	Transposase
PGF_00716243	Transposase
PGF_00734221	Xanthine and CO dehydrogenases maturation factor, XdhC/CoxF family
PGF_01030857	hypothetical protein
PGF_01030864	hypothetical protein
PGF_03067324	Putative protease
PGF_06264218	hypothetical protein