BV-BRC Test Report

A5. Service – Comprehensive Genome Analysis - Bacteria

Item to test	Comprehensive Genome Analysis Service using bacterial reads sets and contigs
URL	https://www.bv-brc.org/app/ComprehensiveGenomeAnalysis
Prerequisites	Bacterial read files and contig files in the workspace
References	https://www.bv-brc.org/docs/quick_references/services/comprehensive_genome_analysis_service.html https://www.bv-brc.org/docs/tutorial/comprehensive_genome_analysis/comprehensive_genome_analysis.html
Tester(s)	Rebecca Wattam, Maulik Shukla
Test date	21-Apr-2022 (follow-up from original test)
Test result	**Passed**

Overview

- Test the Comprehensive Genome Analysis service using exemplar bacterial datasets.
- Test input options, i.e., single end or paired end read files from workspace, sear sets using SRA accessions, or assembled contigs from workspace.
- Test assembly strategies, i.e., Auto, Unicycler, SPAdes, Canu, MetaSPAdes, and PlasmidSPAdes.
- For each job submitted, verify successful completion of the job, availability of the output files in the workspace, and quality of the assembly and annotations by comparing them to the same or similar public genome.
- Verify successful integration of the genome in BV-BRC by reviewing genome overview pages and other genome level tabs.
- Review the quality and accuracy of the comprehensive genome report by comparing the summary stats with those available on the genome overview page.

Test Data

Dataset	**Rational**	**Input Format**	**Input**
Buchnera aphidicola - SRR4240359	Workshop example	SRA accession, read files, fasta file	SRR4240359
Escherichia coli - SRR3584989	Workshop example	SRA accession, read files	SRR3584989
Mycobacterium tuberculosis H37Rv	Reference genome	FASTA file	Mycobacterium_tuberculosis_H37Rv.fasta
Escherichia coli MG1655	Reference genome	FASTA file	Escherichia_coli_MG1655.fasta
Brucella suis 1330	Reference	FASTA file	Brucella_suis_1330.fasta

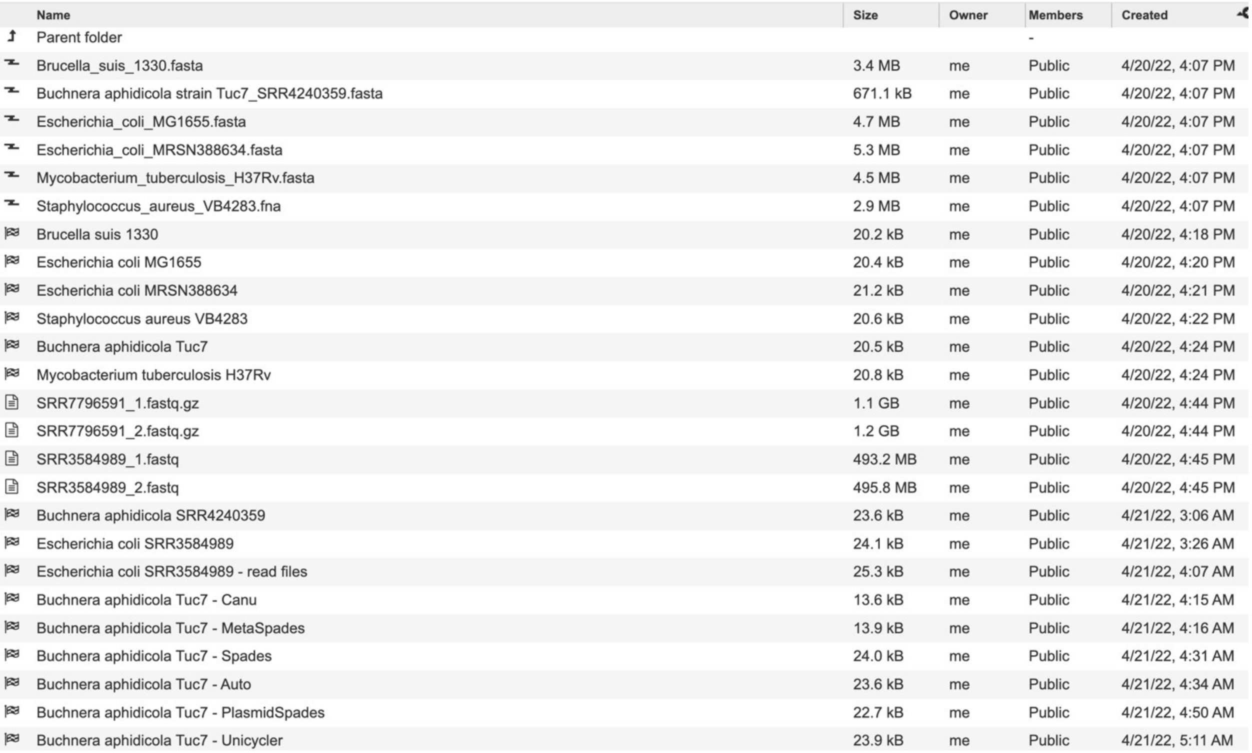
	genome		
Staphylococcus aureus VB4283	Workshop example	FASTA file	Staphylococcus_aureus_VB4283.fna

- All test datasets and corresponding job results are available in the following public workspace: https://www.bv-brc.org/workspace/BVBRC@patricbrc.org/BVBRC%20Tests/Comprehensive%20Genome%20Analysis

Test Results

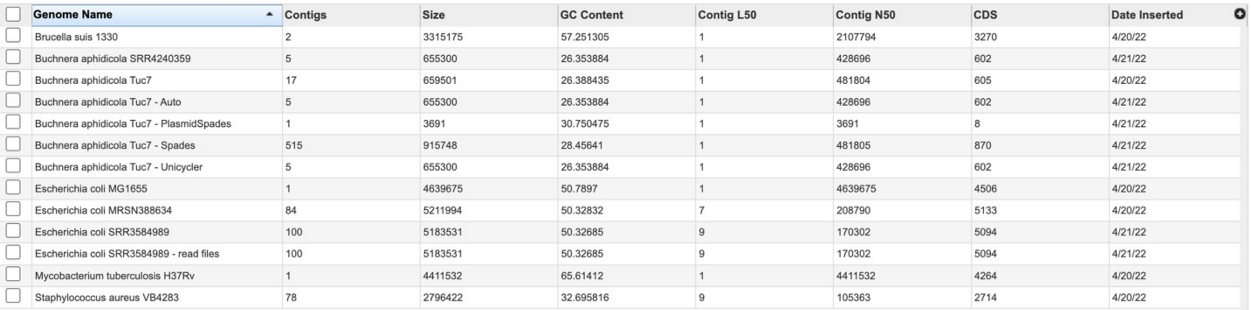
- All comprehensive genome analysis jobs completed successfully, without any errors.
- All jobs resulted in expected output files in corresponding job output directory, providing comprehensive analysis report in HTML format, assembled contigs in fasta format, and genome annotations in expected file formats.
- The genome report was informative and provided a concise summary of the assembly, annotations, circular genome view, subsystem pie chart, summary of specialty genes, list of AMR genes, phylogenetic tree of closest reference genomes, and list of references.
- For each of the genomes, the total length of the assembled contigs and number of features were as expected when compared to those in corresponding public genomes in PATRIC.
- All test datasets and corresponding job results are available in the following public workspace: https://www.bv-brc.org/workspace/BVBRC@patricbrc.org/BVBRC%20Tests/Comprehensive%20Genome%20Analysis
- Below are some sample of screenshots showing successful completion of the jobs, availability of the result files in the workspace, excerpts of the comprehensive genome repot.

- Input data files and completed analysis jobs in the workspace.



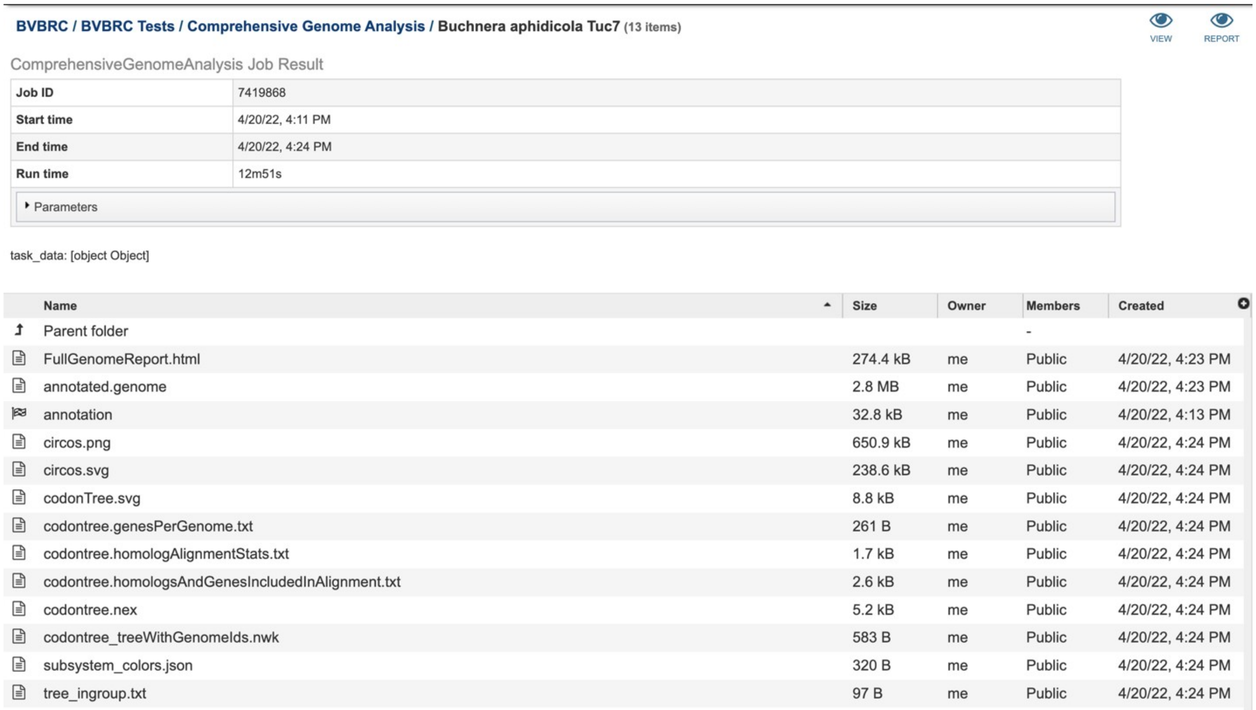
Name	Size	Owner	Members	Created
Parent folder			-	
Brucella_suis_1330.fasta	3.4 MB	me	Public	4/20/22, 4:07 PM
Buchnera aphidicola strain Tuc7_SRR4240359.fasta	671.1 kB	me	Public	4/20/22, 4:07 PM
Escherichia_coli_MG1655.fasta	4.7 MB	me	Public	4/20/22, 4:07 PM
Escherichia_coli_MRSN388634.fasta	5.3 MB	me	Public	4/20/22, 4:07 PM
Mycobacterium_tuberculosis_H37Rv.fasta	4.5 MB	me	Public	4/20/22, 4:07 PM
Staphylococcus_aureus_VB4283.fna	2.9 MB	me	Public	4/20/22, 4:07 PM
Brucella suis 1330	20.2 kB	me	Public	4/20/22, 4:18 PM
Escherichia coli MG1655	20.4 kB	me	Public	4/20/22, 4:20 PM
Escherichia coli MRSN388634	21.2 kB	me	Public	4/20/22, 4:21 PM
Staphylococcus aureus VB4283	20.6 kB	me	Public	4/20/22, 4:22 PM
Buchnera aphidicola Tuc7	20.5 kB	me	Public	4/20/22, 4:24 PM
Mycobacterium tuberculosis H37Rv	20.8 kB	me	Public	4/20/22, 4:24 PM
SRR7796591_1.fastq.gz	1.1 GB	me	Public	4/20/22, 4:44 PM
SRR7796591_2.fastq.gz	1.2 GB	me	Public	4/20/22, 4:44 PM
SRR3584989_1.fastq	493.2 MB	me	Public	4/20/22, 4:45 PM
SRR3584989_2.fastq	495.8 MB	me	Public	4/20/22, 4:45 PM
Buchnera aphidicola SRR4240359	23.6 kB	me	Public	4/21/22, 3:06 AM
Escherichia coli SRR3584989	24.1 kB	me	Public	4/21/22, 3:26 AM
Escherichia coli SRR3584989 - read files	25.3 kB	me	Public	4/21/22, 4:07 AM
Buchnera aphidicola Tuc7 - Canu	13.6 kB	me	Public	4/21/22, 4:15 AM
Buchnera aphidicola Tuc7 - MetaSpades	13.9 kB	me	Public	4/21/22, 4:16 AM
Buchnera aphidicola Tuc7 - Spades	24.0 kB	me	Public	4/21/22, 4:31 AM
Buchnera aphidicola Tuc7 - Auto	23.6 kB	me	Public	4/21/22, 4:34 AM
Buchnera aphidicola Tuc7 - PlasmidSpades	22.7 kB	me	Public	4/21/22, 4:50 AM
Buchnera aphidicola Tuc7 - Unicycler	23.9 kB	me	Public	4/21/22, 5:11 AM

- Genomes successfully assembled, annotated, and integrated into the database.



□	Genome Name ▲	Contigs	Size	GC Content	Contig L50	Contig N50	CDS	Date Inserted
□	Brucella suis 1330	2	3315175	57.251305	1	2107794	3270	4/20/22
□	Buchnera aphidicola SRR4240359	5	655300	26.353884	1	428696	602	4/21/22
□	Buchnera aphidicola Tuc7	17	659501	26.388435	1	481804	605	4/20/22
□	Buchnera aphidicola Tuc7 - Auto	5	655300	26.353884	1	428696	602	4/21/22
□	Buchnera aphidicola Tuc7 - PlasmidSpades	1	3691	30.750475	1	3691	8	4/21/22
□	Buchnera aphidicola Tuc7 - Spades	515	915748	28.45641	1	481805	870	4/21/22
□	Buchnera aphidicola Tuc7 - Unicycler	5	655300	26.353884	1	428696	602	4/21/22
□	Escherichia coli MG1655	1	4639675	50.7897	1	4639675	4506	4/20/22
□	Escherichia coli MRSN388634	84	5211994	50.32832	7	208790	5133	4/20/22
□	Escherichia coli SRR3584989	100	5183531	50.32685	9	170302	5094	4/21/22
□	Escherichia coli SRR3584989 - read files	100	5183531	50.32685	9	170302	5094	4/21/22
□	Mycobacterium tuberculosis H37Rv	1	4411532	65.61412	1	4411532	4264	4/20/22
□	Staphylococcus aureus VB4283	78	2796422	32.695816	9	105363	2714	4/20/22

- Output files in job result directory.



BVBRC / BVBRC Tests / Comprehensive Genome Analysis / Buchnera aphidicola Tuc7 (13 items)

ComprehensiveGenomeAnalysis Job Result

Job ID	7419868
Start time	4/20/22, 4:11 PM
End time	4/20/22, 4:24 PM
Run time	12m51s

▸ Parameters

task_data: [object Object]

	Name ▲	Size	Owner	Members	Created
↥	Parent folder			-	
🗎	FullGenomeReport.html	274.4 kB	me	Public	4/20/22, 4:23 PM
🗎	annotated.genome	2.8 MB	me	Public	4/20/22, 4:23 PM
🏁	annotation	32.8 kB	me	Public	4/20/22, 4:13 PM
🗎	circos.png	650.9 kB	me	Public	4/20/22, 4:24 PM
🗎	circos.svg	238.6 kB	me	Public	4/20/22, 4:24 PM
🗎	codonTree.svg	8.8 kB	me	Public	4/20/22, 4:24 PM
🗎	codontree.genesPerGenome.txt	261 B	me	Public	4/20/22, 4:24 PM
🗎	codontree.homologAlignmentStats.txt	1.7 kB	me	Public	4/20/22, 4:24 PM
🗎	codontree.homologsAndGenesIncludedInAlignment.txt	2.6 kB	me	Public	4/20/22, 4:24 PM
🗎	codontree.nex	5.2 kB	me	Public	4/20/22, 4:24 PM
🗎	codontree_treeWithGenomeIds.nwk	583 B	me	Public	4/20/22, 4:24 PM
🗎	subsystem_colors.json	320 B	me	Public	4/20/22, 4:24 PM
🗎	tree_ingroup.txt	97 B	me	Public	4/20/22, 4:24 PM

- Comprehensive genome analysis report.

Summary

An assembled genome for Buchnera aphidicola Tuc7 was submitted to the comprehensive genome analysis service at PATRIC[1]. Based on the annotation statistics and a comparison to other genomes in PATRIC within this same species, this genome appears to be of Good quality. Details of the analysis, including genes of interest (Specialty Genes), a functional categorization (Subsystems), and a phylogenetic tree (Phylogenetic Analysis) are provided below.

Genome Assembly

An assembled genome was submitted to the Comprehensive Genome Analysis service. This assembled genome had 17 contigs, with the total length of 659,501 bp and an average G+C content of 26.39% (**Table 1**).

Table 1. Assembly Details	
Contigs	17
GC Content	26.39
Plasmids	0
Contig L50	1
Genome Length	659,501 bp
Contig N50	481,804
Chromosomes	0

Genome Annotation

The Buchnera aphidicola Tuc7 genome was annotated using RAST tool kit (RASTtk)[2] and assigned a unique genome identifier of 9.1023. This genome is in the superkingdom Bacteria and was annotated using genetic code 11. The taxonomy of this genome is:

> cellular organisms > Bacteria > Proteobacteria > Gammaproteobacteria > Enterobacterales > Erwiniaceae > Buchnera > Buchnera aphidicola

This genome has 605 protein coding sequences (CDS), 32 transfer RNA (tRNA) genes, and 3 ribosomal RNA (rRNA) genes. The annotated features are summarized in **Table 2**.

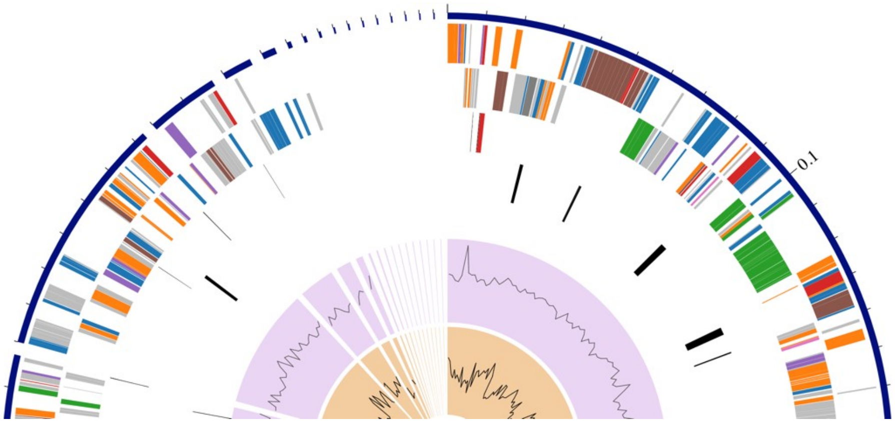
Table 2. Annotated Genome Features	
CDS	605
tRNA	32
rRNA	3
Partial CDS	0
Miscellaneous RNA	0
Repeat Regions	0
Job ID	annotation_4952
Job Started	April 20th 2022, 9:11:11pm
Job Completed	April 20th 2022, 9:13:53pm
Total Time	2 minutes and 42 seconds

The annotation included 26 hypothetical proteins and 579 proteins with functional assignments (**Table 3**). The proteins with functional assignments included 318 proteins with Enzyme Commission (EC) numbers[3], 267 with Gene Ontology (GO) assignments[4], and 241 proteins that were mapped to KEGG pathways[5]. PATRIC annotation includes two types of protein families[6], and this genome has 597 proteins that belong to the genus-specific protein families (PLFams) for , and 600 proteins that belong to the cross-genus protein families (PGFams).

Table 3. Protein Features	
Hypothetical proteins	26
Proteins with functional assignments	579
Proteins with EC number assignments	318
Proteins with GO assignments	267
Proteins with Pathway assignments	241
Proteins with PATRIC genus-specific family (PLfam) assignments	597
Proteins with PATRIC cross-genus family (PGfam) assignments	600

A circular graphical display of the distribution of the genome annotations is provided (**Figure 1**). This includes, from outer to inner rings, the contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to know virulence factors, GC content and GC skew. The colors of the CDS on the forward and reverse strand indicate the subsystem that these genes belong to (see Subsystems below).

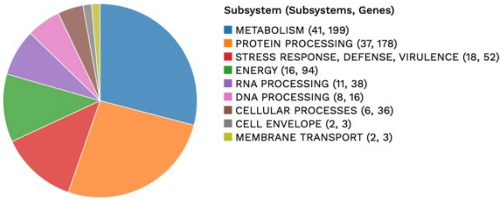
Figure 1



Subsystem Analysis

A subsystem is a set of proteins that together implement a specific biological process or structural complex[7] and PATRIC annotation includes an analysis of the subsystems unique to each genome. An overview of the subsystems for this genome is provided in **Figure 2**.

Figure 2



Specialty Genes

Many of the genes annotated in have homology to known transporters[8], virulence factors[9][10], drug targets[11][12], and antibiotic resistance genes[13]. The number of genes and the specific source database where homology was found is provided (**Table 4**).

Table 4. Specialty Genes

	Source	Genes
Antibiotic Resistance	CARD	1
Antibiotic Resistance	PATRIC	17
Drug Target	DrugBank	11
Drug Target	TTD	1
Transporter	TCDB	3
Virulence Factor	PATRIC_VF	3
Virulence Factor	Victors	4

Antimicrobial Resistance Genes

The Genome Annotation Service in PATRIC uses k-mer-based AMR genes detection method, which utilizes PATRIC's curated collection of representative AMR gene sequence variants[1] and assigns to each AMR gene functional annotation, broad mechanism of antibiotic resistance, drug class and, in some cases, specific antibiotic it confers resistance to. Please note, that the presence of AMR-related genes (even full length) in a given genome does not directl imply antibiotic resistant phenotype. It is important to consider specific AMR mechanisms and especially the absence/presence of SNP mutations conveying resistance. A summary of the AMR genes annotated in this genome and corresponding AMR mechanism is provided in **Table 5**.

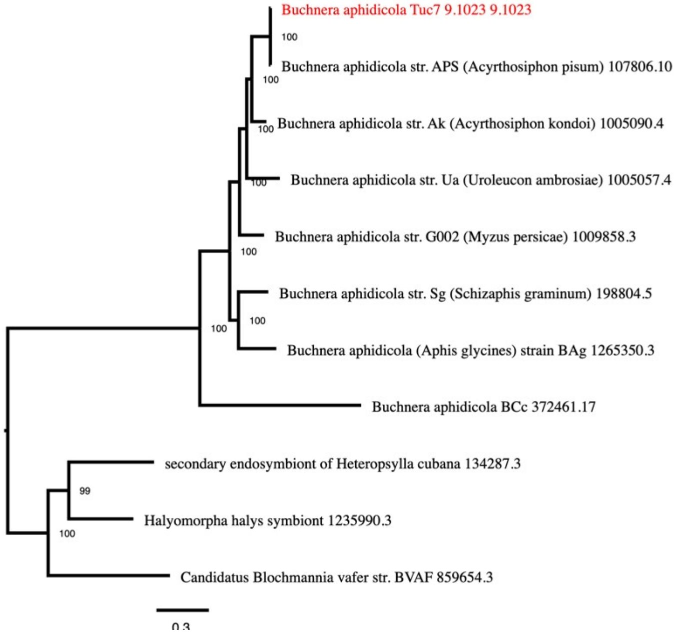
Table 5. Antimicrobial Resistance Genes

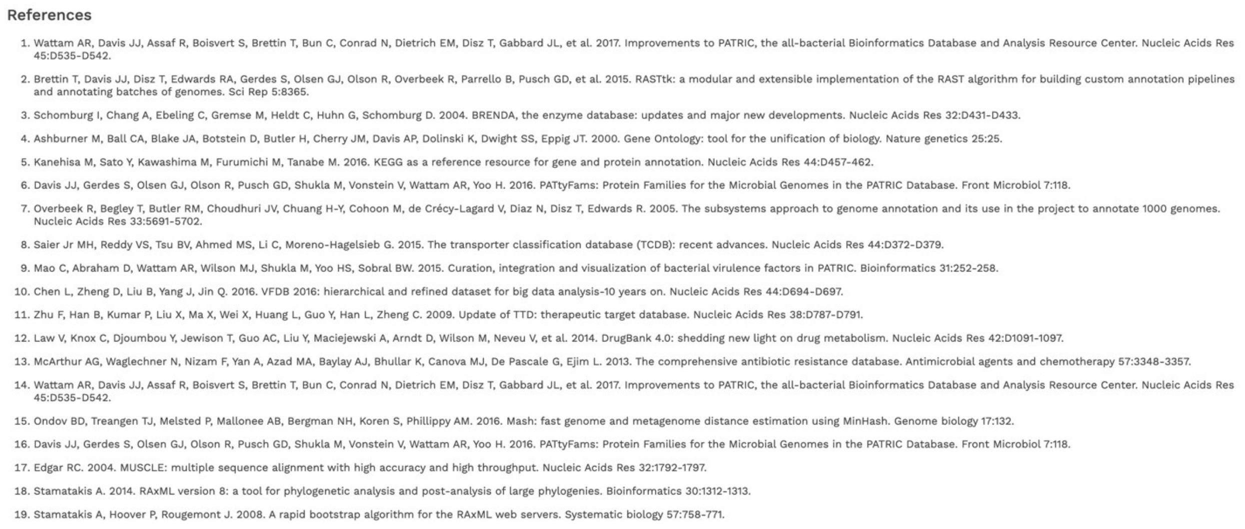
AMR Mechanism	Genes
Antibiotic target in susceptible species	Ddl, dxr, EF-G, EF-Tu, folA, Dfr, gyrA, gyrB, inhA, fabI, Iso-tRNA, MurA, rho, rpoB, rpoC, S10p, S12p
Regulator modulating expression of antibiotic resistance genes	H-NS

Phylogenetic Analysis

The National Center for Biotechnology Information (NCBI) staff manually select and categorize reference and representative genomes, which they consider to be of high quality and importance to the research community. PATRIC provides the reference and representative genomes, and includes them in the phylogenetic analysis that is part of the Comprehensive Genome Analysis report. The closest reference and representative genomes to were identified by Mash/MinHash[15]. PATRIC global protein families (PGFams)[6] were selected from these genomes to determine the phylogenetic placement of this genome. The protein sequences from these families were aligned with MUSCLE[17], and the nucleotides for each of those sequences were mapped to the protein alignment. The joint set of amino acid and nucleotide alignments were concatenated into a data matrix, and RaxML[18] was used to analyze this matrix, with fast bootstrapping[19] was used to generate the support values in the tree (**Figure 3**).

Figure 3

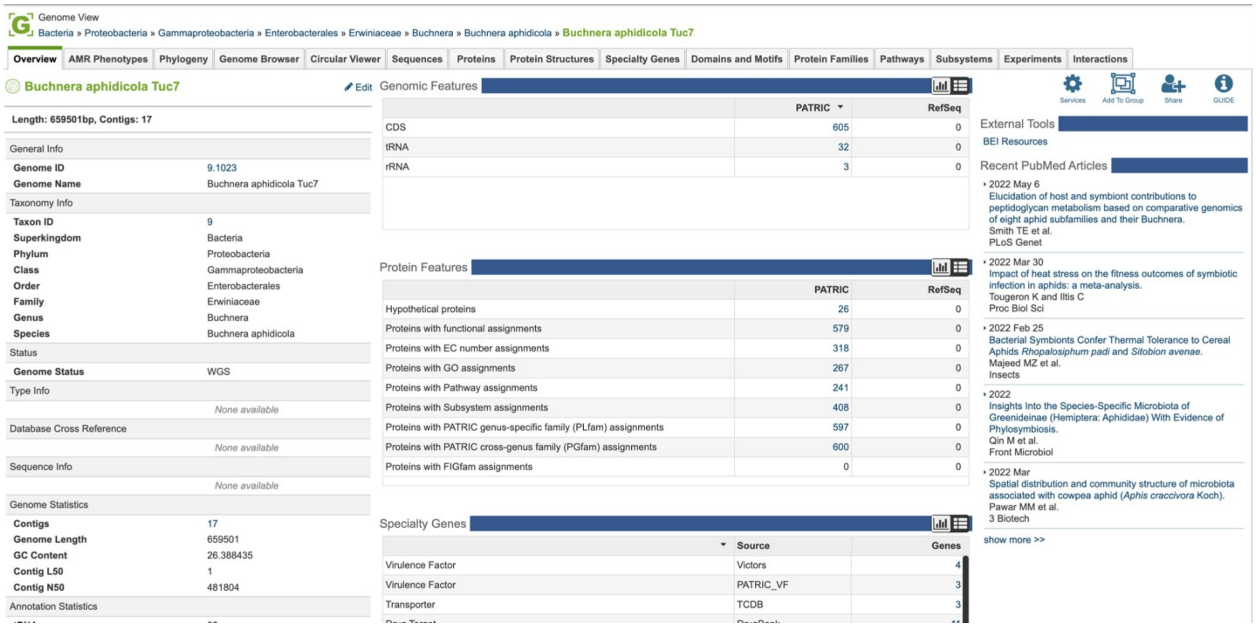

Buchnera aphidicola Tuc7 9.1023 9.1023
100
Buchnera aphidicola str. APS (Acyrthosiphon pisum) 107806.10
100
Buchnera aphidicola str. Ak (Acyrthosiphon kondoi) 1005090.4
100
Buchnera aphidicola str. Ua (Uroleucon ambrosiae) 1005057.4
100
Buchnera aphidicola str. G002 (Myzus persicae) 1009858.3
100
Buchnera aphidicola str. Sg (Schizaphis graminum) 198804.5
100
100
Buchnera aphidicola (Aphis glycines) strain BAg 1265350.3
Buchnera aphidicola BCc 372461.17
secondary endosymbiont of Heteropsylla cubana 134287.3
99
Halyomorpha halys symbiont 1235990.3
100
Candidatus Blochmannia vafer str. BVAF 859654.3
0.3



References

1. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL, et al. 2017. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res 45:D535-D542.
2. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, et al. 2015. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep 5:8365.
3. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. 2004. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 32:D431-D433.
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT. 2000. Gene Ontology: tool for the unification of biology. Nature genetics 25:25.
5. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457-462.
6. Davis JJ, Gerdes S, Olsen GJ, Olson R, Pusch GD, Shukla M, Vonstein V, Wattam AR, Yoo H. 2016. PATtyFams: Protein Families for the Microbial Genomes in the PATRIC Database. Front Microbiol 7:118.
7. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33:5691-5702.
8. Saier Jr MH, Reddy VS, Tsu BV, Ahmed MS, Li C, Moreno-Hagelsieb G. 2015. The transporter classification database (TCDB): recent advances. Nucleic Acids Res 44:D372-D379.
9. Mao C, Abraham D, Wattam AR, Wilson MJ, Shukla M, Yoo HS, Sobral BW. 2015. Curation, integration and visualization of bacterial virulence factors in PATRIC. Bioinformatics 31:252-258.
10. Chen L, Zheng D, Liu B, Yang J, Jin Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis-10 years on. Nucleic Acids Res 44:D694-D697.
11. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, Huang L, Guo Y, Han L, Zheng C. 2009. Update of TTD: therapeutic target database. Nucleic Acids Res 38:D787-D791.
12. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, et al. 2014. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 42:D1091-1097.
13. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L. 2013. The comprehensive antibiotic resistance database. Antimicrobial agents and chemotherapy 57:3348-3357.
14. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL, et al. 2017. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res 45:D535-D542.
15. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome biology 17:132.
16. Davis JJ, Gerdes S, Olsen GJ, Olson R, Pusch GD, Shukla M, Vonstein V, Wattam AR, Yoo H. 2016. PATtyFams: Protein Families for the Microbial Genomes in the PATRIC Database. Front Microbiol 7:118.
17. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.
18. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.
19. Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. Systematic biology 57:758-771.

- Fully integrated genome on the genome overview page.



Bacteria » Proteobacteria » Gammaproteobacteria » Enterobacterales » Erwiniaceae » Buchnera » Buchnera aphidicola » Buchnera aphidicola Tuc7

Buchnera aphidicola Tuc7

Length: 659501bp, Contigs: 17

General Info	
Genome ID	9.1023
Genome Name	Buchnera aphidicola Tuc7
Taxonomy Info	
Taxon ID	9
Superkingdom	Bacteria
Phylum	Proteobacteria
Class	Gammaproteobacteria
Order	Enterobacterales
Family	Erwiniaceae
Genus	Buchnera
Species	Buchnera aphidicola
Status	
Genome Status	WGS
Type Info	
	None available
Database Cross Reference	
	None available
Sequence Info	
	None available
Genome Statistics	
Contigs	17
Genome Length	659501
GC Content	26.388435
Contig L50	1
Contig N50	481804
Annotation Statistics	

Genomic Features

	PATRIC	RefSeq
CDS	605	0
tRNA	32	0
rRNA	3	0

Protein Features

	PATRIC	RefSeq
Hypothetical proteins	26	0
Proteins with functional assignments	579	0
Proteins with EC number assignments	318	0
Proteins with GO assignments	267	0
Proteins with Pathway assignments	241	0
Proteins with Subsystem assignments	408	0
Proteins with PATRIC genus-specific family (PLfam) assignments	597	0
Proteins with PATRIC cross-genus family (PGfam) assignments	600	0
Proteins with FIGfam assignments	0	0

Specialty Genes

	Source	Genes
Virulence Factor	Victors	4
Virulence Factor	PATRIC_VF	3
Transporter	TCDB	3

External Tools

BEI Resources

Recent PubMed Articles

- 2022 May 6
  Elucidation of host and symbiont contributions to peptidoglycan metabolism based on comparative genomics of eight aphid subfamilies and their Buchnera.
  Smith TE et al.
  PLoS Genet
- 2022 Mar 30
  Impact of heat stress on the fitness outcomes of symbiotic infection in aphids: a meta-analysis.
  Tougeron K and Iltis C
  Proc Biol Sci
- 2022 Feb 25
  Bacterial Symbionts Confer Thermal Tolerance to Cereal Aphids *Rhopalosiphum padi* and *Sitobion avenae*.
  Majeed MZ et al.
  Insects
- 2022
  Insights Into the Species-Specific Microbiota of Greenideinae (Hemiptera: Aphididae) With Evidence of Phylosymbiosis.
  Qin M et al.
  Front Microbiol
- 2022 Mar
  Spatial distribution and community structure of microbiota associated with cowpea aphid (*Aphis craccivora* Koch).
  Pawar MM et al.
  3 Biotech

References

- Comprehensive Genome Analysis Service
- Comprehensive Genome Analysis Service Tutorial
- Genome Assembly Service Quick Reference Guide
- Genome Annotation Service Quick Reference Guide
- Phylogenetic Tree Building Service Quick Reference Guide
- Genome Annotation Protocol
- *RASTtk:* A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes