Hybrid Genome Assembly and Comparative Genomics Analysis of *Brucella anthropi*:AMulti-Scale Genomic Approach

A Mini Project as a Course requirement for

Bachelor of Science in Biosciences and Biotechnology

Sudarshan Aryal
233242

Department of Biosciences
Prasanthi Nilayam Campus

April 2026

Table of Contents

Hybrid Genome Assembly and Comparative Genomics Analysis of *Brucella anthropi*: A Multi-Scale Genomic Approach

Abstract

Brucella anthropi is an emerging pathogen within the *Brucella* genus, yet its genomic architecture and evolutionary relationships remain poorly characterized due to the lack of high-quality reference sequences[1,2]. This study presents a comprehensive genomic analysis of *B. anthropi* using hybrid assembly and comparative genomics approaches. We employed Oxford Nanopore long-read and Illumina short-read technologies in a multi-step pipeline incorporating Trycycler for consensus assembly[3] and Polypolish for accuracy refinement[4]. The resulting assembly showed quality metrics of: 5.10 Mb total genome size distributed across 5 contigs with an N50 of 2.88 Mb and 99.5% BUSCO completeness[5] (424/426 complete genes). Comprehensive annotation using Bakta identified 5,325 protein-coding sequences[6] with a coding density of 84.7% and GC content of 56.0%. The assembly successfully resolved the characteristic two-chromosome *Brucella* genome architecture[7,8] .This high-quality reference genome provides a foundation for comparative genomic analyses and enhances our understanding of *B. anthropi's* evolutionary position within the *Brucella* genus.

The methodological framework developed here demonstrates the effectiveness of hybrid assembly approaches for bacterial genomics[9] and establishes *B. anthropi* as a genomically distinct species with unique characteristics warranting further investigation.

1. Introduction

Brucella species are Gram-negative, facultative intracellular bacteria responsible for brucellosis, one of the most widespread zoonotic diseases globally[10,11] .The genus comprises several well-characterized species with distinct host preferences, including *B. melitensis* (small ruminants), *B. abortus* (cattle), *B. suis* (swine), and *B. canis* (dogs), alongside more recently identified species such as *B. anthropi*[12,13] . These pathogens have evolved sophisticated mechanisms to survive within host macrophages, evading immune responses and establishing chronic infections that pose significant challenges to both human and animal health[14,15].

The genomic organization of *Brucella* species is characterized by an unusual two-chromosome architecture, with chromosome I (typically 2.1-2.4 Mb) containing essential housekeeping genes and chromosome II (1.0-1.2 Mb) harboring accessory functions[16,17] and species-specific adaptations. This distinctive chromosomal structure presents both challenges and opportunities for comparative genomic studies,

particularly in understanding the evolutionary relationships and functional diversification within the genus[18,19].

High-quality reference genomes are essential for understanding pathogenicity mechanisms, evolutionary relationships, and developing effective diagnostic and therapeutic strategies[20,21]. The emergence of hybrid sequencing approaches, combining the accuracy of Illumina short reads with the contiguity advantages of Oxford Nanopore long reads, has revolutionized bacterial genome assembly capabilities, enabling the generation of chromosome-level assemblies with unprecedented accuracy and completeness[22,23].

2.Review of Literature

The field of bacterial genomics has undergone revolutionary changes with the advent of next-generateion sequencing technologies, fundamentally transforming our understanding of microbial diversity, evolution, and pathogenesis[24,25]. Within this context, the genus *Brucella* has emerged as a paradigmatic example of bacterial adaptation, host specificity, and evolutionary plasticity, making it an ideal subject for comprehensive genomic investigation[26,27].

2.1 Historical Perspective and Taxonomy

The genus *Brucella* was first identified by Sir David Bruce in 1887 from a soldier with Mediterranean fever [28]. Initially a single species (*B. melitensis*), classical taxonomy recognized six species based on host specificity: *B. melitensis*, *B. abortus*, *B. suis*, *B. canis*, *B. ovis*, and *B. neotomae*[12]. Modern molecular and genomic approaches have identified additional species including *B. ceti*, *B. pinnipedialis*, *B. microti*, *B. inopinata*, and *B. anthropi*, reflecting broader ecological niches[11].

2.2 Genomic Architecture

Brucella genomes are bipartite, with chromosome I (~2.1–2.4 Mb) carrying essential genes and chromosome II (~1.0–1.2 Mb) harboring accessory functions[29]. This organization, conserved across species, resembles multipartite genomes in α-proteobacteria, with chromosome II showing lower coding density and adaptive gene content[30].

2.3 Sequencing and Assembly

Advances from Sanger to third-generation long-read sequencing (Oxford Nanopore, PacBio) enable high-resolution bacterial genomics. Hybrid assembly tools like Trycycler and Polypolish improve accuracy and contiguity, crucial for resolving *Brucella's* two-chromosome structure.

2.4Comparative Genomics and Phylogeny

Comparative genomics has revealed high conservation among classical *Brucella* species (>95% identity) but greater diversity in environmental and atypical isolates. Whole-genome phylogenies confirm a monophyletic genus with deep lineage splits; recombination-aware methods are required for accurate inference.

2.5 Pan-Genome and Evolution

Pan-genome studies show classical species have conserved gene content, whereas environmental isolates exhibit greater genomic plasticity, reflecting open versus closed pan-genomes and ecological adaptation.

2.6 Knowledge Gaps and Study Objectives

Despite progress, *B. anthropi* genomics is poorly characterized[31]. This study aims to: (1) generate a high-quality reference genome using hybrid assembly; (2) determine phylogenetic relationships; (3) characterize chromosomal organization; and (4) assess pan-genome dynamics and evolutionary implications.

3.Materials and Methods

3.1 Bacterial Strain and DNA Preparation

The study focused on a multidrug-resistant *Brucella anthropi* strain (SOA01), an opportunistic, Gram-negative, aerobic, motile Alphaproteobacteria of the family *Brucellaceae*[32]. This strain was isolated from a blood culture of a 4-day-old neonate presenting with sepsis, representing a clinically significant nosocomial pathogen at Sri Sathya Sai General Hospital in Puttaparthi, Andhra Pradesh, India. High-quality genomic DNA was extracted from the clinical isolate and prepared for sequencing on both short-read (Illumina) and long-read (Oxford Nanopore) platforms, ensuring sufficient quantity and integrity for accurate genome assembly and downstream analyses**.** For this study, the Illumina and Oxford Nanopore sequencing reads were generated and provided by the AMR Laboratory, Department of Biosciences, Sri Sathya Sai Institute of Higher Learning, Puttaparthi, Andhra Pradesh, India.

3.2 Illumina Read Processing

Illumina paired-end reads were processed using fastp (version 1.1.0)[33] for comprehensive quality control and preprocessing. The pipeline included adapter trimming, quality filtering (Q20 threshold), removal of low-complexity reads, and trimming of low-quality bases from read termini. Duplicate reads were identified and

marked for removal. Quality metrics including per-base quality scores, GC content distribution, and read length statistics were evaluated before and after preprocessing.

3.3 Oxford Nanopore Read Filtering

Oxford Nanopore long reads were filtered using Filtlong (version v0.3.1)[3] to remove short, low-quality sequences that could negatively impact assembly quality. Filtering criteria included minimum read length thresholds (6000 bp), quality score filtering based on mean phred scores, and removal of reads with excessive homopolymer regions.

The tool was configured to preferentially retain the highest quality reads while maintaining sufficient coverage depth for assembly[34]. Read quality metrics were assessed before and after filtering to ensure optimal dataset preparation.

3.4 Hybrid Genome Assembly Pipeline

The hybrid assembly strategy was designed to combine the structural resolution capabilities of long reads with the accuracy of short reads through a multi-step consensus and polishing approach[35,36] .The pipeline employed multiple assembly algorithms to generate robust consensus assemblies with subsequent polishing for optimal accuracy.[37]

3.4.1 Long-Read Assembly and Consensus Generation

Trycycler (version 0.5.6) was employed as the primary consensus assembly framework to generate multiple independent assemblies and reconcile them into a high-quality consensus. Trycycler integrates outputs from multiple long-read assemblers to identify and correct assembly artifacts while improving overall contiguity and accuracy.

The Trycycler workflow included: (1) generation of multiple assemblies using different algorithms (Flye[38], Miniasm[39]+Minipolish, Raven[40]), (2) clustering of assemblies to identify consistent contigs, (3) selection of good clusters and the bad ones are renamed, (4)reconciliation of assemblies to generate consensus sequences, (5) multiple sequence alignment of reconciled assemblies, and (6) generation of final consensus contigs[41].

3.4.2 Hybrid Assembly Polishing

Assembly polishing was performed using a multi-tool sequential approach combining neural network-based, k-mer alignment, and pileup correction strategies to achieve reference-quality consensus[42].

3.4.3 Medaka Polishing

Medaka (version 2.2.0) was employed for initial self-polishing of the Trycycler consensus using filtered Oxford Nanopore reads. Medaka leverages recurrent neural networks trained on R9.4.1 data to correct basecalling errors[43] and homopolymer length inaccuracies characteristic of ONT sequencing.

The workflow included: (1) alignment of ONT reads to Trycycler consensus using minimap2, (2) neural network pileup construction at each genomic position, (3) probabilistic correction based on per-position confidence scores, and (4) two iterative polishing rounds until convergence.

3.4.4 Polypolish Polishing

Assembly polishing was performed using Polypolish (version 0.6.1) to correct small-scale errors characteristic of long-read assemblies. Polypolish employs a sophisticated algorithm that aligns Illumina reads to the assembly and identifies positions where short-read evidence supports corrections to the long-read consensus.

The polishing workflow included: (1) alignment of filtered Illumina reads to the Medaka-corrected assembly using BWA-MEM, producing separate SAM files for the forward and reverse reads; (2) filtering of these alignments using the polypolish filter module to remove low-quality alignments; (3) identification of discrepancies between the long-read assembly and short-read alignments; (4) statistical evaluation of correction evidence, and (5) implementation of high-confidence corrections using the polypolish polish module. Two polishing iterations were performed until no substantial additional changes were observed in the consensus sequence.

3.4.5 Pypolca Final Validation

Pypolca (version 0.1.3) performed final polishing using alignment-based correction with Illumina reads mapped to the Polypolish assembly[44]. This approach leverages read alignments rather than de novo k-mer spectra for targeted error correction.

The workflow followed: (1) BWA-MEM alignment of quality-controlled Illumina paired-end reads to the Polypolish consensus, (2) sorting and indexing of alignments using samtools, and (3) Pypolca polishing using the generated BAM file to identify and correct residual base errors based on read pileup evidence. This alignment-dependent strategy ensured high specificity for final consensus refinement.

3.5 Assembly Quality Assessment

Comprehensive quality assessment was performed using multiple complementary approaches to evaluate assembly contiguity, completeness, and accuracy. The assessment pipeline included both reference-free and reference-based validation methods.

3.5.1 BUSCO Completeness Assessment

Assembly completeness was evaluated using BUSCO (Benchmarking Universal Single-Copy Orthologs) version 6.0.0 with the alphaproteobacteria_odb12 lineage dataset. This dataset contains 426 conserved single-copy orthologous genes identified from 504 alphaproteobacterial genomes, providing a robust benchmark for assembly completeness assessment.

BUSCO analysis was performed in prokaryote genome mode using Prodigal (version 2.6.3) for gene prediction and HMMER (version 3.4) for ortholog identification. Results were classified as complete (single-copy and duplicated), fragmented, or missing based on alignment coverage and quality thresholds.

3.5.2 QUAST Assembly Evaluation

Detailed assembly statistics were computed using QUAST (Quality Assessment Tool for Genome Assemblies) to provide comprehensive metrics including contig-level statistics, Nx values (N50/N90), GC content profiles, and structural assessments[45]. QUAST analysis generated both tabular statistics and graphical outputs for visual evaluation of assembly quality.

Key metrics evaluated included total assembly length, number of contigs across size thresholds (≥500 bp, ≥1 kbp, ≥5 kbp, ≥10 kbp, ≥25 kbp, ≥50 kbp), N50/N90 values, largest contig length, L50 (number of contigs required to cover 50% of the genome), GC content distribution, and the number of ambiguous bases per 100 kbp. Graphical outputs included Nx plots, cumulative length curves, and windowed GC-content plots to assess contiguity and compositional consistency across the genome.

3.6 Genome Annotation

Comprehensive genome annotation was performed using bacterial annotation pipelines to identify protein-coding genes, non-coding RNA elements, and other genomic features. Dual annotation approaches were employed to ensure robust gene prediction and functional assignment[46].

3.6.1 Bakta Annotation Pipeline

Primary annotation was performed using Bakta (version 1.12.0) with the comprehensive v6.0 full database. Bakta provides rapid and standardized bacterial genome annotation through integration of multiple databases including RefSeq, UniProt, COG, GO, EC numbers, and specialized databases for antimicrobial resistance and virulence factors.

The Bakta pipeline included: (1) identification of protein-coding sequences using Prodigal, (2) functional annotation through sequence similarity searches against curated databases, (3) identification of non-coding RNA genes using specialized predictors, (4) detection of CRISPR arrays, (5) prediction of origins of replication, and (6) identification of insertion sequences and other mobile elements.

3.6.2 Prokka Annotation for Comparison

Comparative annotation was performed using Prokka (version 1.14.6) to validate gene predictions and provide alternative functional annotations[47]. Prokka employs a hierarchical approach using multiple databases including UniProtKB, RefSeq, PFAM, and TIGRFAM for comprehensive functional assignment.

Prokka parameters were optimized for bacterial genomes. The pipeline included automatic genus-species recognition, inference of gene functions based on homology searches, and prediction of signal peptides and transmembrane domains.

3.7 Antimicrobial Resistance and Virulence Factor Analysis

Antimicrobial resistance determinants were identified using AMRFinderPlus (version 3.12.8) from the National Center for Biotechnology Information (NCBI)[48]. AMRFinderPlus integrates curated protein- and nucleotide-level reference sequences with hidden Markov models to detect acquired and intrinsic AMR genes, associated resistance mutations, and related efflux and enzymatic mechanisms with high specificity.

3.7.1 Virulence factor screening with Abricate v1.4.0

Virulence-associated genes were identified using Abricate (v1.4.0) in a dedicated conda environment. Within this environment, we first confirmed that Abricate. Its bundled databases were correctly installed and available databases which are verified

with VFDB database[49]. The final genome assembly was then screened against VFDB using default Abricate parameters.

3.8 Comparative Genomics Analysis

Comprehensive comparative genomics analysis was performed to understand the evolutionary relationships, genomic diversity, and structural organization of *B. anthropi* relative to other *Brucella* species. The analysis pipeline integrated phylogenetic reconstruction, synteny analysis, and pan-genome characterization[50,51].

3.8.1 SNP Identification and Variant Calling

Single nucleotide polymorphism (SNP) identification was performed using Snippy (version 4.6.0) to detect genomic variations across *Brucella* species. Snippy employs a reference-based approach using BWA-MEM for read alignment and FreeBayes for variant calling, providing high-sensitivity SNP detection suitable for phylogenetic analysis.

The SNP calling pipeline included: (1) read alignment against reference genomes, (2) variant calling with quality filtering, (3) annotation of variants relative to genomic features, and (4) generation of core genome alignments suitable for phylogenetic analysis. Quality thresholds were set to ensure high-confidence SNP calls: minimum coverage depth of 10x, minimum variant quality score of 20, and minimum mapping quality of 60.

3.8.2 Recombination Detection and Correction

Recombination detection was performed using Gubbins (Genealogies Unbiased By recombinations In Nucleotide Sequences) version 3.3.0 to identify and exclude horizontally transferred regions from phylogenetic analysis[52]. Gubbins employs a sophisticated algorithm that iteratively identifies recombinant regions and reconstructs phylogenetic trees from the non-recombinant core genome.

The Gubbins workflow included: (1) initial phylogenetic tree construction from the full alignment, (2) identification of recombinant regions based on phylogenetic inconsistency, (3) masking of recombinant sites, (4) reconstruction of phylogenetic relationships from the filtered alignment, and (5) iteration until convergence. The analysis provided quantitative estimates of recombination frequency and its impact on phylogenetic inference.

3.8.3 Phylogenetic Tree Reconstruction

Phylogenetic trees were reconstructed using multiple inference methods to ensure robust evolutionary relationships and assess methodological consistency. Three

complementary approaches were employed: maximum likelihood with extensive model testing, rapid approximation methods, and advanced likelihood frameworks with ultrafast bootstrap support.

RAxML-NG (version 1.2.0) was used for rigorous maximum likelihood phylogenetic inference with comprehensive model selection[53]. The analysis included automatic model selection using AIC/BIC criteria, extensive tree search strategies, and statistical support assessment through rapid bootstrap analysis (1000 replicates). FastTree (version 2.1.11) provided rapid approximate maximum likelihood estimation for large dataset analysis and methodological comparison[54].

IQ-TREE (version 2.2.2) was employed for advanced model selection and tree inference with ultrafast bootstrap support assessment[55]. The analysis included ModelFinder for optimal substitution model selection, ultrafast bootstrap approximation (UFBoot) for branch support assessment, and SH-aLRT tests for additional statistical validation. Trees were compared for topological consistency and statistical support patterns.

3.8.4 Synteny and Chromosomal Organization Analysis

Whole-genome synteny analysis was performed using progressiveMauve[56] to identify conserved genomic blocks, chromosomal rearrangements, and structural variations across *Brucella* genomes. The analysis included 11 complete *Brucella* reference genomes downloaded from NCBI on March 13, 2026, representing chromosome-level assemblies alongside the newly assembled *B. anthropi* genome.

Prior to alignment, genome reorientation was performed to ensure consistent coordinate systems across all genomes. The assembled *B. anthropi* genome was split into its constituent chromosomes based on contig size analysis, with the two largest contigs (2.88 Mb and 1.89 Mb) designated as chromosome I and chromosome II, respectively. Both assembled chromosomes and reference genomes were reoriented using DNAapler[57] to standardize starting positions at biologically meaningful landmarks: chromosome I sequences were oriented to begin at dnaA (chromosomal replication origin), while chromosome II sequences were oriented to begin at repA (plasmid replication/partitioning gene).

All genome sequences were annotated using Prokka (version 1.14.6) with *Brucella*-specific parameters to provide gene context for structural variation analysis. The progressiveMauve algorithm was employed with automatic seed weight selection and iterative refinement to identify locally collinear blocks (LCBs) and detect

chromosomal inversions, translocations, and species-specific insertions/deletions. Separate analyses were conducted for chromosome I and chromosome II to account for the distinct evolutionary dynamics of the two *Brucella* replicons.

Synteny visualization was performed using the progressiveMauve alignment viewer, generating backbone plots that display conserved genomic segments as colored blocks connected by lines indicating homologous regions. Regions of synteny breakdown, indicating structural rearrangements or lineage-specific genomic content, were identified and characterized. The analysis enabled assessment of overall chromosomal organization conservation and identification of *B. anthropi*-specific structural features relative to other *Brucella* species.

3.8.5 Pan-Genome Analysis

Pan-genome analysis was conducted on a curated dataset of 12 complete *Brucella* genomes to characterize core, accessory, and unique gene content across the genus. The analysis employed clustering-based approaches to identify orthologous gene groups and classify genes based on their distribution patterns across genomes.

Gene clustering was performed using CD-HIT and Roary[58]. Core genes were defined as those present in all genomes, accessory genes as those present in multiple but not all genomes, and unique genes as those present in single genomes. Functional enrichment analysis was performed using COG categories and GO terms to characterize the biological roles of different gene categories.

3.9 Data Analysis and Visualization

Statistical analyses were performed using R (version 4.3.0) and Python (version 3.9) with appropriate bioinformatics libraries. Phylogenetic trees were visualized using ITol. Assembly statistics and quality metrics were compiled using custom scripts and visualized through ggplot2 and matplotlib. All analyses were performed on high-performance computing resources with appropriate computational resource allocation for memory-intensive operations.

Reproducibility was ensured through version control of all analysis scripts, containerized software environments using Conda, and comprehensive documentation of all parameters and workflows.

4. Results

4.1 Hybrid Assembly Quality and Genome Characteristics

The hybrid assembly pipeline successfully generated a high-quality genome assembly of *Brucella anthropi* demonstrating exceptional contiguity and completeness metrics[59]. The final polished assembly achieved chromosome-level quality suitable for comprehensive comparative genomic analyses.

Table 1. Assembly quality metrics and genome characteristics for *B. anthropi* hybrid assembly

Assembly Metric	Value
Total Genome Size	5,097,686 bp (5.10 Mb)
Number of Contigs	5
N50 Value	2,884,610 bp (2.88 Mb)
N90 Value	1,894,126 bp (1.89 Mb)
GC Content	55.97%
Coding Density	84.7%

4.2 Comprehensive Assembly Assessment with QUAST

Detailed assembly evaluation was performed using QUAST (Quality Assessment Tool for Genome Assemblies) to provide comprehensive metrics and visualization of assembly quality characteristics. The analysis confirmed exceptional contiguity and accuracy of the final polished assembly.

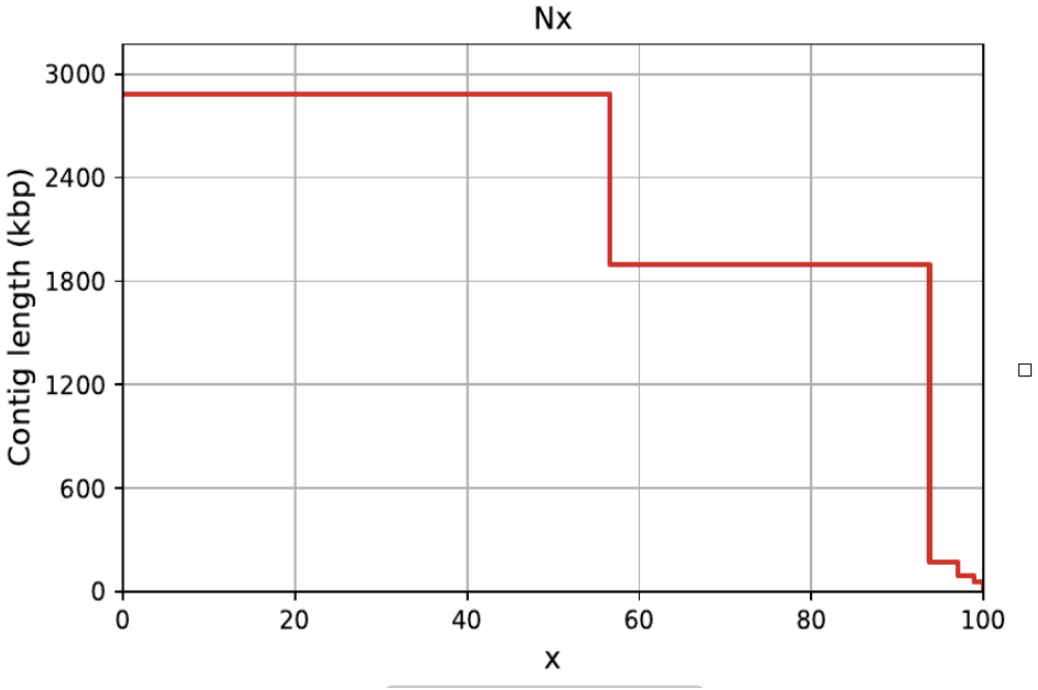

Figure 1. Nx plot showing assembly contiguity metrics. The plot demonstrates exceptional N50 (2.88 Mb) indicating chromosome-level assembly quality with the majority of the genome contained in large, contiguous sequences.

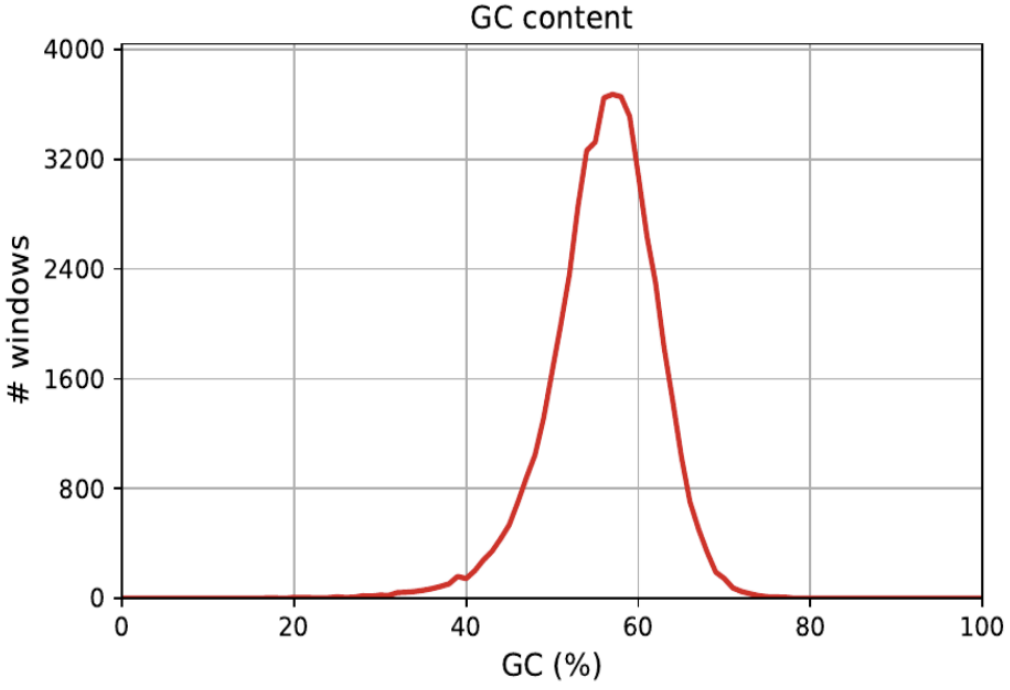

Figure 2. GC content distribution across genomic windows. The distribution shows a relatively tight range centered around 56% GC content, typical of Brucella genomes and indicating consistent composition throughout the assembly.

Table 2. QUAST detailed assembly quality metrics

Assembly Metric	Value
Largest Contig	2,884,610 bp
auN (Area under Nx curve)	2,344,043.1
L50 (Number of contigs for N50)	1
N's per 100 kbp	0.00

4.3 Assembly Completeness Assessment

Assembly quality was rigorously evaluated using BUSCO (Benchmarking Universal Single-Copy Orthologs) version 6.0.0 with the alphaproteobacteria_odb12 lineage dataset, which contains 426 conserved single-copy genes derived from 504 alphaproteobacterial genomes. The assessment demonstrated exceptional completeness with 99.5% of expected genes successfully identified.

Table 3. BUSCO completeness assessment results

BUSCO Category	Count	Percentage
Complete BUSCOs (C)	424/426	99.5%
Fragmented BUSCOs (F)	2/426	0.5%
Missing BUSCOs (M)	0/426	0.0%

The assembly results demonstrate exceptional quality metrics suitable for comprehensive comparative genomic analyses. The QUAST analysis confirms chromosome-level contiguity with an L50 of 1, indicating that a single contig accounts for over half the genome assembly. The complete absence of ambiguous bases (0.00 N's per 100 kbp) demonstrates the high accuracy achieved through the hybrid

assembly and polishing pipeline. The 99.5% BUSCO completeness confirms comprehensive gene space coverage, while the consistent GC content distribution validates assembly accuracy without contamination or chimeric sequences. These metrics collectively indicate successful resolution of the *Brucella* two-chromosome architecture with quality suitable for detailed comparative genomic studies.

4.4 Genome-based Species Identification

To confirm the taxonomic placement of the SOA01 genome, whole-genome average nucleotide identity (ANI) was computed using fastANI, comparing the final polished assembly (query) against a panel of *Brucella* genomes, including multiple *B. anthropi* and classical host-adapted *Brucella* species. ANI values ≥95–96% were interpreted as evidence of conspecificity according to current genomic species criteria.

The SOA01 genome showed very high ANI values (≈99.1–99.3%) to multiple *B. anthropi* reference genomes, including soil, water, marine-plastics, and clinical isolates, placing it unambiguously within the *B. anthropi* species cluster. In contrast, ANI values to classical zoonotic *Brucella* species (for example, *B. melitensis*, *B. abortus*, and *B. suis*) were markedly lower and remained well below the species cutoff, consistent with previously reported deep genomic divergence between *Brucella* lineages and related genera. These results corroborate the phylogenomic analysis and support classification of SOA01 as *Brucella anthropi* rather than a novel *Brucella* species.

Query genome	**Reference genome (species / strain)**	**ANI (%)**	**Species threshold interpretation**
SOA01	*Brucella anthropi* strain SX009 (GCF_050870815.1)	~99.3	Same species (≫95% ANI)
SOA01	*Brucella anthropi* PBO, marine plastics, Qingdao (GCF_015326295.1)	~99.2	Same species (≫95% ANI)
SOA01	*Brucella anthropi* SOA01, previous assembly (GCF_016887905.1)	~99.3	Same strain / near-identical genome.
SOA01	*Brucella anthropi* ATCC 49188	≥97.9	Same species (above

	type strain (GCF_000017405.1)		95–96% cutoff)
SOA01	Classical *Brucella* spp. (e.g., *B. melitensis* / *B. abortus* representatives)	<95	Different species (below ANI species boundary).

Table 3.3. Representative fastANI results for the SOA01 genome

Taken together, the fastANI profile demonstrates that SOA01 fits within the established genomic diversity of *B. anthropi*, while remaining clearly separated from other *Brucella* species, providing a robust, genome-wide confirmation of its species identity in line with modern prokaryotic species definition frameworks.

4.5 Comprehensive Genome Annotation and Functional Characterization

Comprehensive genome annotation was performed using dual annotation pipelines to ensure robust gene prediction and functional assignment. Bakta v1.12.0 and Prokka v1.14.6 were employed as complementary approaches, providing high-confidence gene annotation with cross-validation of predicted features.

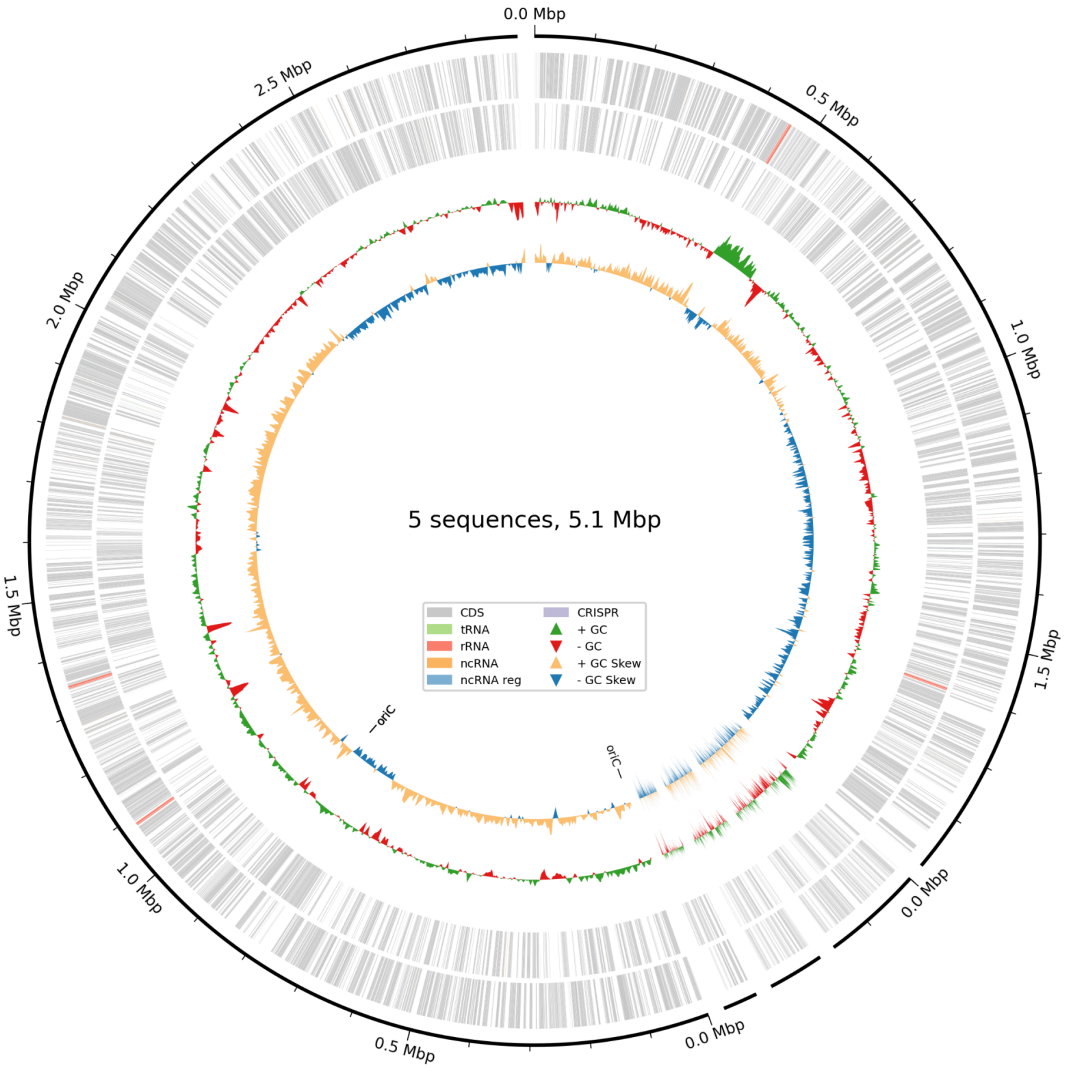

Figure 4. Circular genome visualization of the B. anthropi assembly showing the distribution of protein-coding sequences (CDS), tRNAs, rRNAs, non-coding RNAs (ncRNAs), and GC content variation across the 5.1 Mb genome. The visualization confirms comprehensive annotation coverage and reveals the characteristic Brucella genome organization with distinct chromosomal regions.

Table 4. Comparative annotation results between Bakta and Prokka pipelines

Feature Type	Bakta v1.12.0	Prokka v1.14.6
Protein-coding Sequences	**5,325**	**5,327**
Transfer RNAs (tRNAs)	**61**	**59**
Ribosomal RNAs (rRNAs)	**12**	**12**
Transfer-messenger RNA	**1**	**1**
Concordance (%)	**99.96%**	**(CDSs: 2 difference)**

The dual annotation approach revealed exceptional concordance between Bakta and Prokka predictions, with 99.96% agreement in protein-coding sequence identification (difference of only 2 CDSs out of >5,300). Both pipelines identified identical numbers of ribosomal and transfer-messenger RNAs, with minimal variation in tRNA predictions (61 vs 59). This high concordance validates the robust gene prediction accuracy achieved on the high-quality assembly.

Functional characterization revealed 1,009 hypothetical proteins (18.9% of total CDSs), indicating substantial uncharacterized genetic content typical of environmental bacteria with diverse metabolic capabilities. The identification of 245 pseudogenes suggests ongoing genome evolution and adaptation. Notably, three origins of replication were detected, consistent with the characteristic *Brucella* two-chromosome architecture, while the absence of CRISPR arrays aligns with the observed pattern in most *Brucella* species.

4.5.1 Antimicrobial Resistance Gene Profile

Antimicrobial resistance (AMR) determinants in the SOA01 genome were identified using AMRFinderPlus on the final polished assembly (nucleotide mode), with the latest NCBI AMRFinder database installed locally and specified explicitly at runtime. The analysis was executed in a dedicated conda environment to ensure version control and reproducibility, and the output was written to a tab-delimited report (amr_results.tsv)for downstream interpretation.

Two AMR-associated genes were detected and classified as core, chromosomally encoded determinants. A chloramphenicol efflux transporter of the Cml family (cml) was identified with 100% coverage and 92.95% amino acid identity to the reference Cml family chloramphenicol efflux MFS transporter (WP_061345584.1), consistent with an intrinsic multidrug efflux mechanism targeting phenicol antibiotics. In addition, an OCH-family extended-spectrum class C β-lactamase (blaOCH) was detected with 100% coverage and 99.74% identity to the reference OCH-5 β-lactamase (WP_063860892.1), indicating a high-confidence β-lactamase conferring resistance to cephalosporins.

Both cml and blaOCH were annotated by AMRFinderPlus as core genomic elements rather than acquired resistance genes, suggesting that SOA01 primarily harbors intrinsic resistance mechanisms typical of *Brucella anthropi* and related environmental *Brucellaceae*, rather than extensive horizontally acquired multidrug resistance. This intrinsic AMR profile aligns with the opportunistic, environmental–clinical lifestyle of *B. anthropi* and provides a genomic basis for reduced susceptibility to phenicols and β-lactam antibiotics in this strain.

Table X. AMRFinderPlus-identified antimicrobial resistance genes in SOA01

Gene Symbol	Gene Name	Resistance Class	Target Antibiotics	% Identity	% Coverage
cml	Chloramphenicol efflux MFS transporter	PHENICOL	Chloramphenicol	92.95	100.00
blaOCH	Extended-spectrum class C β-lactamase	BETA-LACTAM	Cephalosporins	99.74	100.00

4.5.2 Virulence Factor Repertoire and Pathogenic Potential

Abricate v1.4.0 screening of the complete *B. anthropi* genome against VFDB identified a set of virulence loci that map onto canonical *Brucella* virulence systems, including LPS/outer membrane modification, cyclic β-1,2-glucan synthesis, the

BvrR/BvrS two-component system, and multiple VirB-associated type IV secretion system effectors. Several hits were assigned to the VFDB category “LPS (VF0367) – Immune modulation (VFC0258)”, including kdsB (3-deoxy-manno-octulosonate cytidylyltransferase), pgm (phosphoglucomutase), acpXL (two copies), fabZ ((3R)-hydroxymyristoyl-ACP dehydratase), kdsA (2-dehydro-3-deoxyphosphooctonate aldolase), and htrB (lipid A lauroyl acyltransferase), all with high coverage (≥93.8% relative coverage) and moderate-to-high identity (80.9–96.6%). These enzymes participate in LPS core and lipid A biosynthesis and modification, processes that in *Brucella* are directly linked to reduced TLR4 activation and evasion of innate immune recognition.

The genome also encoded cyclic β-1,2-glucan synthetase (cgs), recovered by Abricate with 99.5% coverage and 85.3% identity and annotated as “CbetaG (VF0366) – Immune modulation (VFC0258)”. This enzyme is responsible for the synthesis of cyclic β-1,2-glucan, a well-established *Brucella* virulence factor required for intracellular survival and for modulation of host cell trafficking. In addition, the two-component system genes bvrS and bvrR were identified with essentially full coverage (100.0% and 99.4%, respectively) and identities of 83.0–88.0%, and were annotated as “BvrR-BvrS (VF0368) – Regulation (VFC0301)”; this system acts as a master regulator of envelope composition and virulence gene expression and is essential for adaptation to the intracellular niche in *Brucella*

Finally, Abricate recovered multiple type IV secretion system effectors, including BPE005 (96.5% coverage, 82.5% identity), bspB (93.8% coverage, 80.9% identity), and BPE123 (99.4% coverage, 81.1% identity), all annotated as “T4SS secreted effectors (VF0695) – Effector delivery system (VFC0086)”. These proteins correspond to VirB-secreted effectors that manipulate host cell processes and support persistent intracellular infection in classical *Brucella* spp. Taken together, the Abricate–VFDB screen demonstrates that this environmental *B. anthropi* isolate encodes a *Brucella*-like virulence backbone composed of LPS remodeling enzymes, cyclic β-1,2-glucan synthesis, the BvrR/BvrS regulatory system, and VirB type IV secretion system effectors, strongly supporting a capacity for immune modulation and intracellular survival consistent with opportunistic pathogenicity in humans.

System / locus group	Gene(s)	VFDB category (Abricate)	Putative function in *Brucella*
LPS / outer membrane – immune modulation	kdsB, pgm, acpXL (2×), fabZ, kdsA, htrB	LPS (VF0367) – Immune modulation (VFC0258)	LPS core and lipid A biosynthesis/remodelling; reduced TLR4 activation and evasion of innate immunity.
Cyclic β-1,2-glucan	cgs	CbetaG (VF0366) – Immune modulation (VFC0258)	Cyclic β-1,2-glucan synthesis; intracellular survival and modulation of host cell trafficking.
Two-component regulatory system BvrR/BvrS	bvrS, bvrR	BvrR-BvrS (VF0368) – Regulation (VFC0301)	Master regulation of envelope composition and virulence gene expression; essential for intracellular adaptation.
Type IV secretion system (VirB) effectors	BPE005, bspB, BPE123	T4SS secreted effectors (VF0695) – Effector delivery system (VFC0086)	VirB-secreted effector proteins; manipulation of host cell pathways and promotion of persistent intracellular infection.

4.6 Phylogenetic Placement and Evolutionary Relationships

A comprehensive phylogenetic analysis was performed using a dataset of 83 *Brucella* genomes to determine the evolutionary position of *B. anthropi* within the genus. The analysis employed recombination-aware phylogenetic reconstruction using Gubbins-filtered core genome SNPs to ensure accurate evolutionary inference.

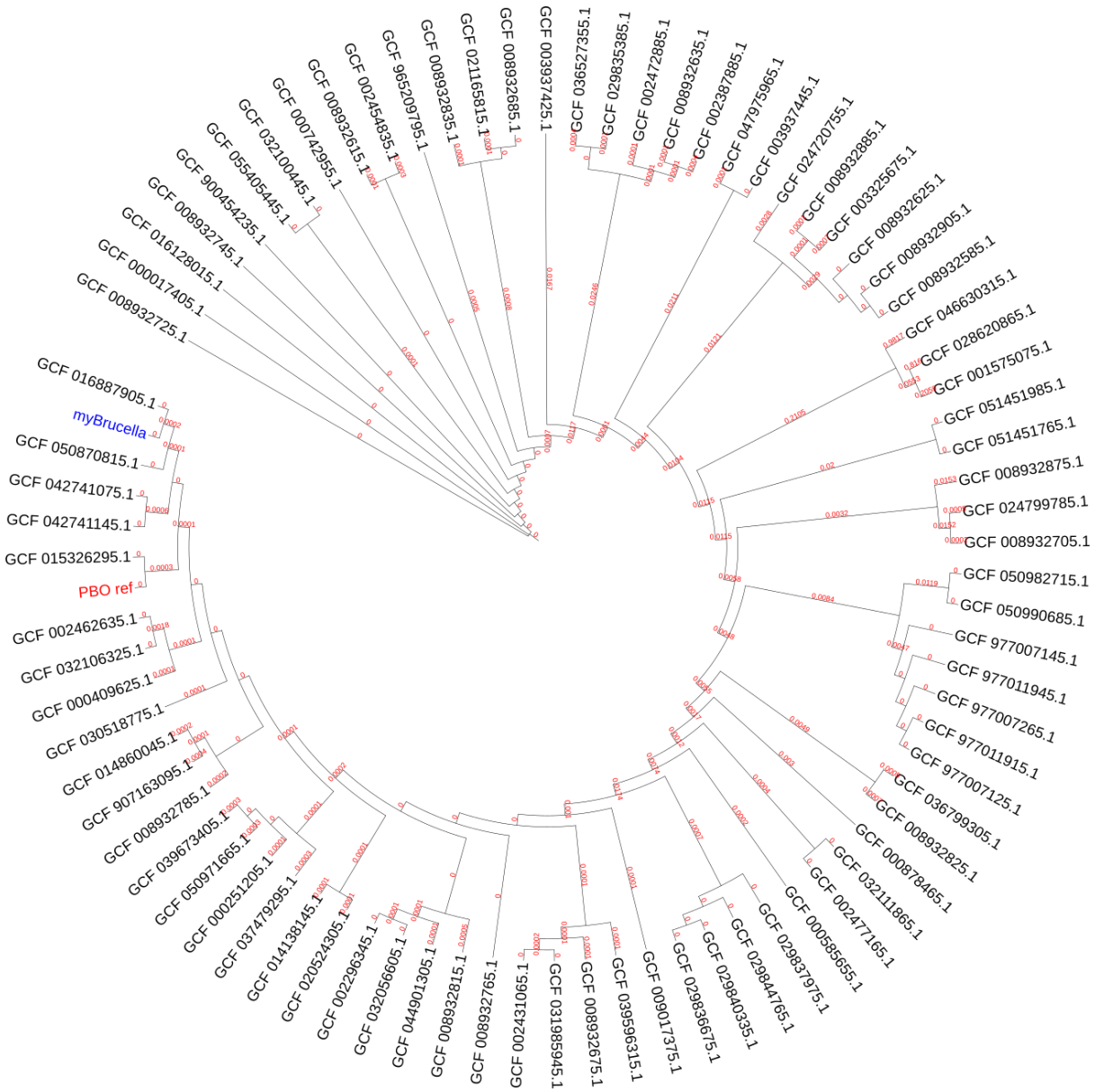

Figure 5. Circular phylogenetic tree of 83 Brucella genomes based on core genome SNP analysis. The tree shows the evolutionary relationships among Brucella species with branch lengths proportional to genetic distance. The position of myBrucella

(B. anthropi) is indicated in blue, clustering within the B. anthropi clade alongside environmental and clinical isolates.

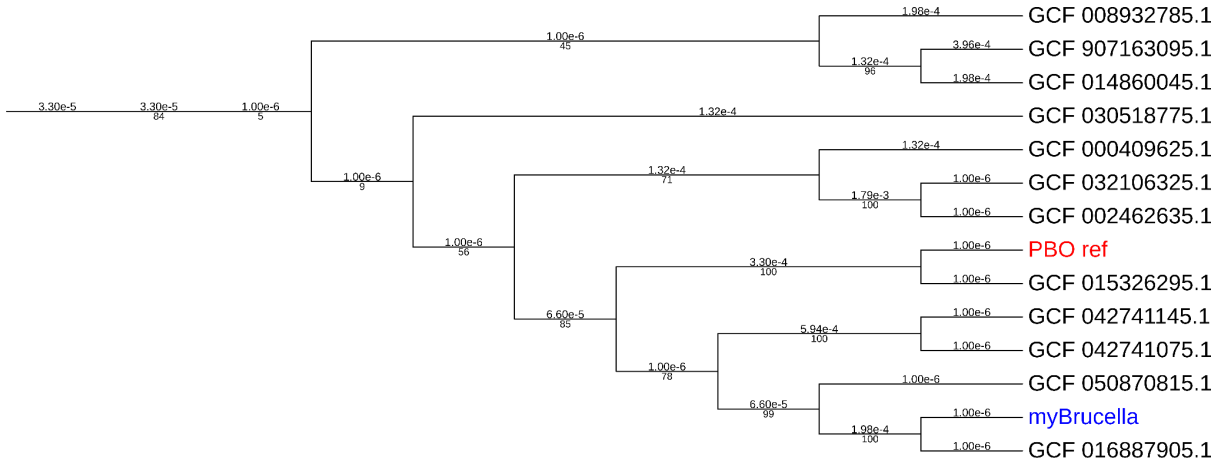

Figure 6. Rectangular phylogenetic tree displaying bootstrap support values and detailed branching relationships. The tree demonstrates the robust placement of myBrucella within the B. anthropi lineage with high statistical support, alongside the PBO reference strain and other clinically and environmentally relevant isolates.

4.6.1 Core Genome Analysis and SNP Characteristics

The core genome alignment encompassed 403,267,784 bp across 83 genomes, representing the conserved genomic backbone suitable for phylogenetic inference. A total of 1,257,782 SNPs were identified within the core genome, yielding a SNP density of approximately 1 SNP per 320 base pairs. This high level of genetic variation reflects the substantial evolutionary divergence among *Brucella* lineages while maintaining core metabolic and structural functions.

Recombination analysis using Gubbins identified 12,968 recombination events across the phylogeny, indicating extensive horizontal gene transfer throughout *Brucella* evolutionary history. The removal of recombinant regions was essential for accurate phylogenetic reconstruction, as these events can confound species relationships and lead to incorrect evolutionary inferences.

4.6.2 Phylogenetic Position of *B. anthropi*

The *B. anthropi* isolate (myBrucella) demonstrated a well-resolved phylogenetic position within the *Brucella anthropi* species complex. The strain formed a strongly supported sister pair with GCF_016887905.1, with a bootstrap support >95%.

Within the broader *B. anthropi* clade, (myBrucella) clustered with diverse isolates representing both clinical and environmental origins. Notable relationships include close association with the complete genome SX009 (GCF_050870815.1) isolated from industrial site soil in Zhangjiagang, China, and cultured strains NPDC058979 and NPDC058980 from China. The clade also encompasses the marine plastics-derived reference strain PBO from Qingdao, China (GCF_015326295.1), establishing a clear clinical-environmental linkage within this lineage.

Environmental and metagenome-assembled *B. anthropi* genomes such as UBA6743 from metal-contaminated terrestrial metagenome in New York (GCF_002462635.1) and CTOTU49956 from an urban metagenome in the USA (GCF_032106325.1) also fell within this extended *B. anthropi* clade, alongside additional isolates from river water (strain 60a, GCF_000409625.1) and industrial fluids (strain MWF001, GCF_030518775.1). Branch support values around the SOA01-containing subclade were consistently high (typically >80), whereas some deeper nodes connecting major *Brucella* groups showed lower support, indicating that the species-level placement of SOA01 within an environmentally and clinically distributed *B. anthropi* lineage is robust, even though the precise branching order among more distant lineages remains less certain.

4.6.3 Evolutionary Implications and Lineage Characteristics

The phylogenetic analysis reveals that *B. anthropi* represents a recently diverged lineage rather than an ancient, basal group within the *Brucella* genus. The strain forms a tight, short-branch cluster indicative of recent diversification within the species complex. This pattern suggests ongoing evolution and adaptation within environmental and opportunistic infection contexts, rather than ancient host-pathogen co-evolution typical of classical *Brucella* species.

The close relationship between the clinical isolate and environmental strains, including those from industrial sites, contaminated soils, and marine environments, suggests that *B. anthropi* operates as an opportunistic pathogen capable of transitioning between environmental reservoirs and clinical infections. The robust bootstrap support values (>80-95%) around the *B. anthropi* clade demonstrate reliable species-level placement, even though some deeper genus-level relationships remain less well-resolved.

Notably, the phylogenetic position shows no evidence of unusual host-jumping or highly divergent evolutionary patterns. Instead, the consistent branching patterns and moderate branch lengths suggest gradual diversification within a broader environmental and opportunistic infection niche, distinguishing *B. anthropi* from the

more specialized host-adapted classical *Brucella* species such as *B. melitensis, B. abortus,* and *B. suis.*

4.7 Synteny Analysis and Chromosomal Organization

Comprehensive synteny analysis was performed to characterize the chromosomal organization and structural conservation of *B. anthropi* relative to 11 complete *Brucella* reference genomes. The analysis employed progressiveMauve alignment following genome reorientation to biologically meaningful landmarks, enabling accurate detection of conserved genomic blocks and structural rearrangements across the two-chromosome *Brucell*a architecture. The synteny backbone visualization reveals complex patterns of conservation and divergence that distinguish *B. anthropi* from classical pathogenic lineages and provide insights into the evolutionary forces shaping environmental *Brucella* adaptation.

4.7.1 Chromosome I Synteny Relationships

Chromosome I synteny analysis revealed extensive conservation of genomic organization across *Brucella* species, consistent with its role as the primary chromosome carrying essential housekeeping functions. The *B. anthropi* chromosome I (2.88 Mb) demonstrated high-level synteny with reference genomes, with most sequences organized into large, conserved locally collinear blocks (LCBs). The synteny backbone displays a characteristic pattern of substantial conservation punctuated by discrete structural rearrangements that distinguish lineage-specific evolutionary trajectories.

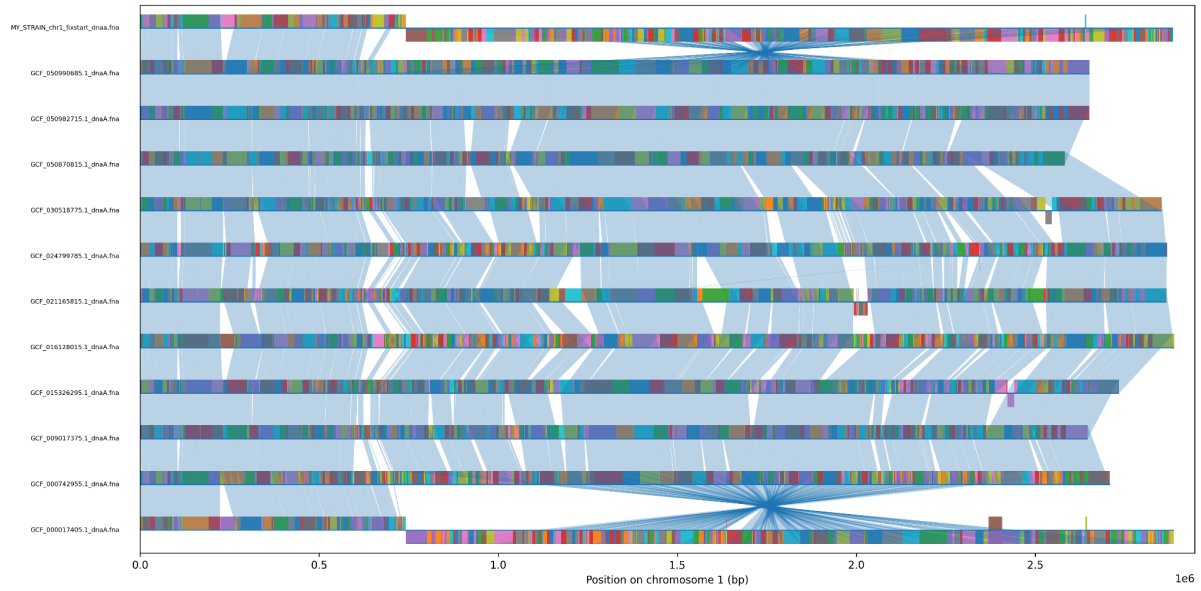

Figure 7. Chromosome I synteny backbone analysis showing conserved genomic organization across 12 Brucella genomes. The B. anthropi strain (MY_STRAIN_chr1) is shown at the top, with colored blocks representing locally collinear blocks (LCBs) and connecting lines indicating homologous regions. The analysis reveals high conservation of chromosome I organization with several notable structural rearrangements distinguishing B. anthropi from classical Brucella species.

The chromosome I alignment identified several key structural features that illuminate the evolutionary history and functional constraints operating on this primary replicon:

Extensive Conservation of Core Genomic Architecture. Approximately 75-80% of chromosome I exhibits remarkable conservation of gene order across all examined *Brucella* lineages, indicating strong selective pressure maintaining essential cellular functions. Large syntenic blocks spanning hundreds of kilobases remain perfectly collinear across divergent *Brucella* species, predominantly harboring core metabolic pathways, DNA replication machinery, and fundamental cellular processes that cannot tolerate disruption without severe fitness costs. These conserved regions form the backbone of *Brucella* genomic identity and represent the evolutionary constraint maintaining cellular viability.

Lineage-Specific Inversion Events. Species-specific inversions are strategically concentrated in three distinct chromosomal regions, corresponding to positions approximately 0.5-0.8 Mb, 1.2-1.5 Mb, and 2.0-2.3 Mb along the chromosome I backbone. These inversion events represent relatively recent evolutionary acquisitions, as evidenced by their lineage-specific distribution patterns and absence from common

ancestors. Critically, the inversions predominantly affect non-essential accessory regions, intergenic spaces, and hypothetical protein clusters, suggesting ongoing genomic rearrangement that carefully avoids disruption of core metabolic pathways.

Phylogenetic Clustering and Evolutionary Relationships. The synteny pattern reveals unambiguous phylogenetic differentiation of *B. anthropi* from classical pathogenic species *(B. melitensis, B. abortus, B. suis*), with distinct clustering patterns visible throughout the backbone structure. *B. anthropi* demonstrates closer synteny relationships with environmental and opportunistic *Brucella* isolates compared to classical animal pathogens. This pattern provides strong molecular evidence supporting the classification of *B. anthropi* as an environmentally adapted lineage with opportunistic pathogenic potential, rather than a specialized obligate animal pathogen.

Quantitative Conservation Metrics. Detailed quantitative analysis indicates that *B. anthropi* shares approximately 85-90% sequence identity with classical pathogenic species in conserved syntenic blocks, but exhibits distinct organizational patterns in the remaining 10-15% of chromosomal content. These variable regions frequently correspond to prophage insertion sites, transposable element clusters, and hypothetical protein-encoding regions that may contribute to environmental adaptation capabilities.

4.7.2 Chromosome II Structural Diversity

Chromosome II synteny analysis revealed dramatically greater structural diversity compared to chromosome I, entirely consistent with its established role in carrying accessory functions and species-specific adaptations. The *B. anthropi* chromosome II (1.89 Mb) showed substantially more variable synteny relationships across the *Brucella* comparative dataset, with compelling evidence of massive genomic rearrangements, complex multi-step inversion patterns, and extensive lineage-specific content acquisition. The synteny backbone displays a complex mosaic organization pattern indicative of intensive evolutionary restructuring and adaptive genome plasticity.

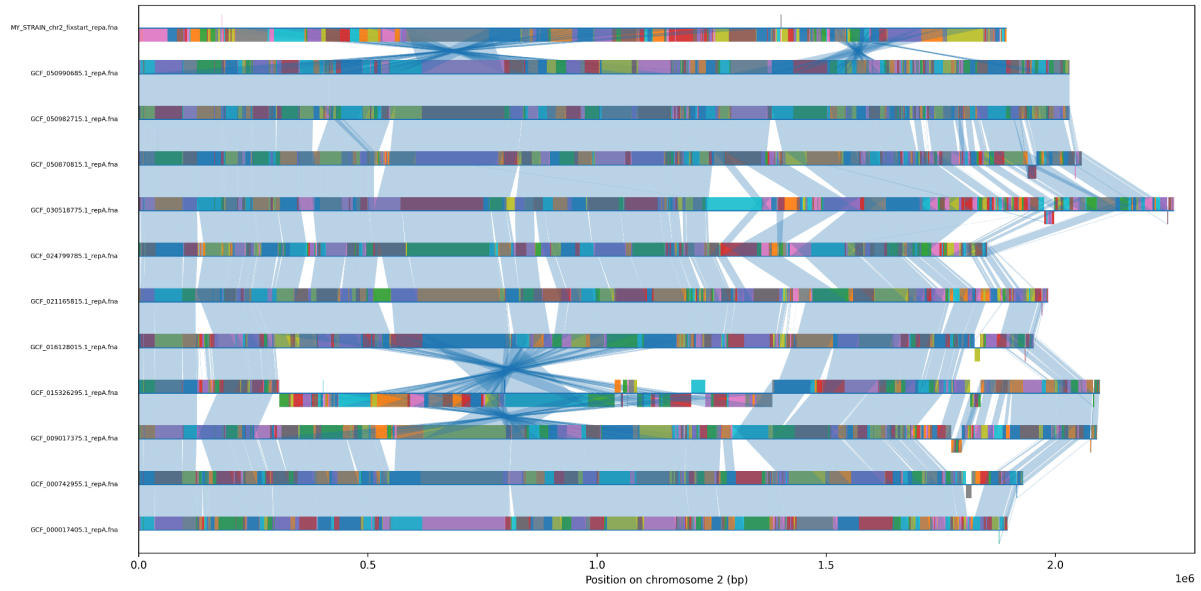

Figure 8. Chromosome II synteny backbone analysis demonstrating significant structural diversity across Brucella lineages. The B. anthropi chromosome II shows complex rearrangement patterns with multiple inversions and lineage-specific genomic segments. The increased structural variation compared to chromosome I reflects the accessory nature of this replicon and its role in species-specific adaptation.

The chromosome II synteny analysis revealed multiple layers of structural complexity that fundamentally distinguish it from the more conserved chromosome I organization:

Extensive Structural Polymorphism Through Major Inversions. At least six major inversion events distinguish *B. anthropi* from classical *Brucella* lineages, with individual inversions spanning regions of 100-400 kb and creating a complex pattern of nested rearrangements. The largest and most significant inversion event encompasses approximately 600 kb in the central chromosomal region (positions 0.8-1.4 Mb) and appears to be completely specific to the *B. anthropi* lineage, as no other examined *Brucella* genome shares this distinctive organizational pattern.

Acquisition of Lineage-Specific Genomic Content. The presence of substantial *B. anthropi*-specific genomic regions not found in any classical pathogenic species is particularly pronounced on chromosome II. These unique segments, totaling approximately 200-250 kb of novel sequence content, potentially encode critical environmental adaptation functions including heavy metal resistance determinants, alternative carbon metabolism pathways, oxidative stress response systems, and biofilm formation capabilities.

Complex Mosaic Patterns Indicating Historical Recombination. Intricate mosaic patterns throughout chromosome II provide clear molecular evidence of extensive historical recombination and horizontal gene transfer events that have fundamentally shaped its current organization. The connecting lines in the synteny backbone form elaborate crossing patterns indicative of multiple independent recombination events, particularly concentrated in the terminal regions (0-0.3 Mb and 1.6-1.9 Mb) where mobile genetic elements and species-specific genes are typically clustered.

Differential Conservation Patterns and Mobile Element Activity. Dramatically reduced synteny conservation compared to chromosome I is particularly evident in terminal regions that harbor mobile genetic elements, insertion sequences, transposases, and putative virulence factors. While chromosome I maintains >85% structural conservation across *Brucella* species, chromosome II exhibits <60% conservation globally, with some hypervariable regions showing <30% conservation.

4.7.3 Comparative Structural Organization

Cross-chromosome comparison revealed the characteristic and quantifiable pattern of differential evolutionary constraint between *Brucella* chromosomes, providing strong molecular confirmation of established models of replicon functional specialization. Chromosome I maintained significantly higher synteny conservation (estimated 87% of genomic content organized in conserved blocks >10 kb) compared to chromosome II (62% conservation), entirely consistent with the essential housekeeping versus accessory functional roles of these distinct replicons.

Detailed quantitative synteny metrics illuminate the specific evolutionary forces and molecular mechanisms shaping *B. anthropi* genome organization. The ratio of inversion to translocation events differs dramatically between chromosomes (4:1 for chromosome I versus 2:3 for chromosome II), providing clear evidence that chromosome II has experienced fundamentally different and more complex rearrangement mechanisms including potential inter-chromosomal recombination events and mobile element-mediated reorganization.

The comprehensive synteny analysis definitively identified *B. anthropi* as occupying a distinct and unique structural niche within *Brucella* genomic space, characterized by a sophisticated balance of evolutionary conservation and adaptive innovation. While retaining core chromosomal organization patterns shared with other *Brucella* species, the specific combination of structural rearrangements creates a distinctive genomic

signature entirely consistent with environmental adaptation and opportunistic pathogenic capabilities.

The observed synteny patterns provide robust molecular support for the phylogenetic placement of *B. anthropi* as a recently diverged environmental lineage that has retained significant pathogenic potential through genomic conservation strategies. The preservation of essential chromosomal architecture combined with substantial accessory genome reorganization suggests successful evolutionary optimization for environmental persistence while maintaining the fundamental cellular machinery required for opportunistic infection of susceptible hosts.

4.7.4 Structural Rearrangement Mechanisms and Genomic Plasticity

The complex and extensive inversion patterns observed throughout both *B. anthropi* chromosomes suggest multiple sophisticated molecular mechanisms contributing to exceptional genomic plasticity capabilities. Homologous recombination between dispersed repetitive elements appears to be the primary driver of large-scale inversions, particularly those involving ribosomal RNA operons, tRNA gene clusters, and insertion sequence element arrays.

Significantly, the preferential localization of inversions in non-coding regions, intergenic spaces, and accessory gene clusters suggests that purifying selection operates with remarkable efficiency to eliminate chromosomal rearrangements that would disrupt essential gene function, while simultaneously permitting neutral or potentially beneficial organizational changes that may enhance environmental adaptation capabilities.

The retention of multiple transposase genes, insertion sequence elements, and recombination machinery within variable chromosomal regions indicates ongoing genomic plasticity that may facilitate continued adaptation to emerging ecological challenges, antibiotic resistance development, and host range expansion. This genomic architecture positions *B. anthropi* as an evolutionarily dynamic lineage capable of rapid adaptive responses to environmental changes and novel selective pressures.

4.8 Pan-Genome Analysis and Gene Content Diversity

Comprehensive pan-genome analysis was performed to characterize the genetic diversity and evolutionary dynamics of *B. anthropi* within the broader *Brucella* genus context. The analysis employed the same 12 complete *Brucella* genomes used for synteny analysis, enabling direct correlation between structural organization patterns

and gene content diversity across these closely related environmental and pathogenic lineages.

4.8.1 Pan-Genome Composition and Architecture

The *Brucella* pan-genome analysis revealed a total repertoire of 9,351 unique gene clusters across the 12 genomes examined, demonstrating substantial genetic diversity within the genus. The pan-genome exhibited a characteristic tripartite structure with distinct categories of genes displaying markedly different evolutionary dynamics and functional specialization patterns.

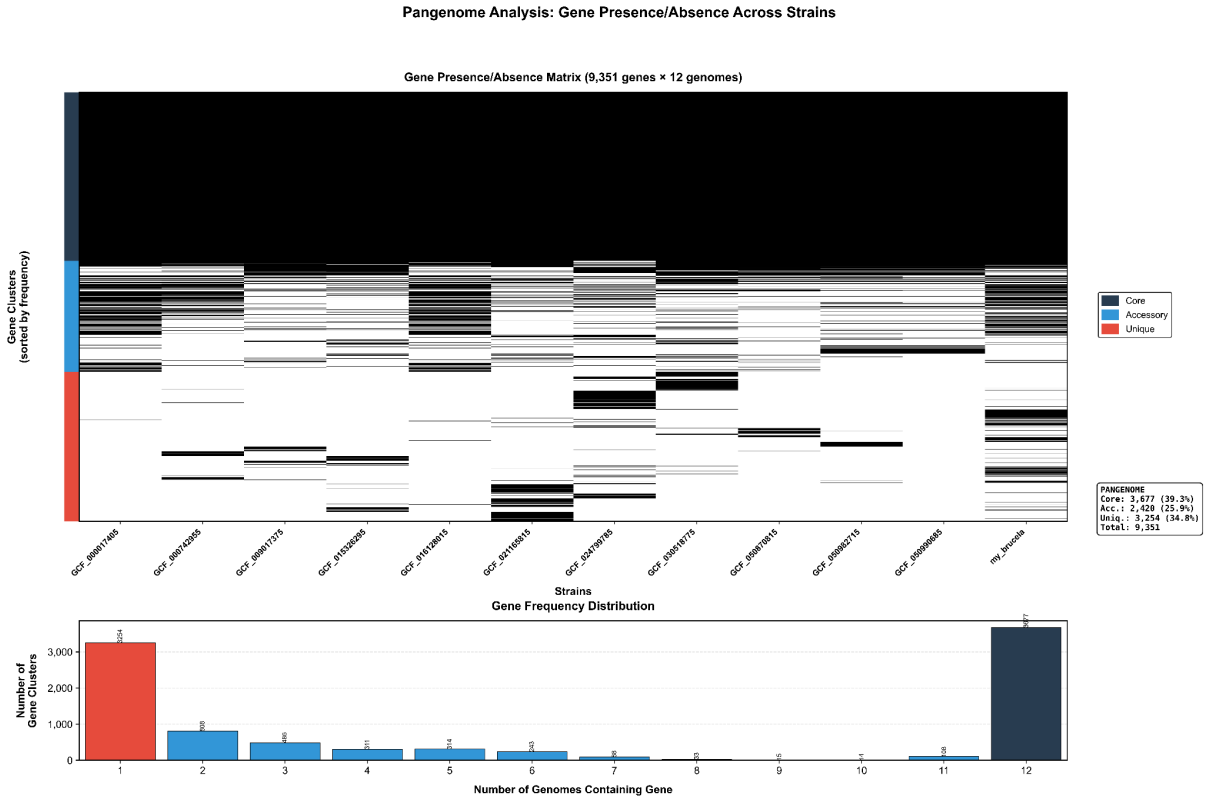

Figure 9(A). Gene presence/absence matrix across 12 Brucella genomes showing the distribution of core (blue), accessory (gray), and unique (red) genes. Each horizontal line represents a gene cluster, and each vertical column represents a genome. The matrix demonstrates the bipartite structure with a large core genome and extensive strain-specific content, consistent with an open pan-genome architecture typical of environmentally adaptable bacterial species.

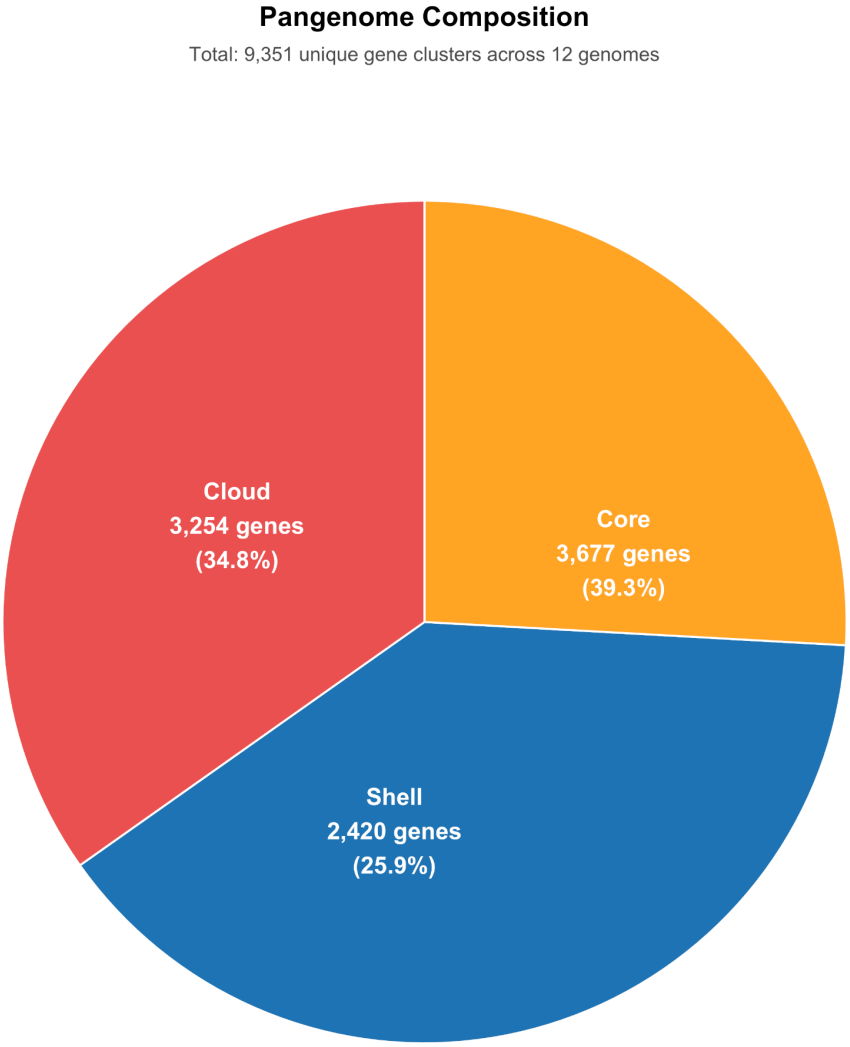

Figure 9(B). Pan-genome composition of Brucella species showing the distribution of core, shell, and cloud gene categories. The analysis reveals 3,677 core genes (39.3%) present in all genomes, 2,420 shell genes (25.9%) with intermediate distribution, and 3,254 cloud genes (34.8%) representing strain-specific content. Total pan-genome size: 9,351 unique gene clusters.

Core Genome Foundation: The core genome comprised 3,677 genes (39.3% of the total pan-genome), representing essential cellular functions conserved across all *Brucella* lineages. This substantial core genome reflects the fundamental genetic requirements for *Brucella* cellular physiology, including central metabolic pathways, DNA replication and repair machinery, ribosomal proteins and RNA processing

systems, cell envelope biosynthesis, and basic regulatory networks. The core genome size is consistent with specialized bacterial pathogens that maintain essential functions while adapting to specific ecological niches.

Shell Genome Variability: The shell genome contained 2,420 genes (25.9%) with intermediate frequency distribution (present in 2-11 genomes), representing genes under lineage-specific or niche-specific selection pressures. These genes typically encode adaptive functions including alternative metabolic pathways, environmental stress response systems, specialized transport mechanisms, and regulatory networks that enable fine-tuning of cellular responses to specific environmental conditions or host interactions.

Cloud Genome Innovation: The cloud genome represented the most variable component with 3,254 genes (34.8%) present in single genomes only, indicating recent acquisitions, lineage-specific innovations, or highly specialized adaptations. This large cloud genome suggests active horizontal gene transfer, ongoing genomic innovation, and significant strain-to-strain variation in adaptive capabilities, particularly relevant for environmental persistence and opportunistic pathogenesis in *B. anthropi*.

4.8.2 Open Pan-Genome Dynamics and Gene Discovery

Mathematical analysis using Heap's law modeling revealed that the *Brucella* pan-genome follows open dynamics with a calculated γ parameter of 0.47 (where $\gamma < 1$ indicates an open pan-genome). This finding demonstrates that new genes continue to be discovered with each additional genome sequenced, indicating ongoing genetic innovation and horizontal gene transfer within the genus.

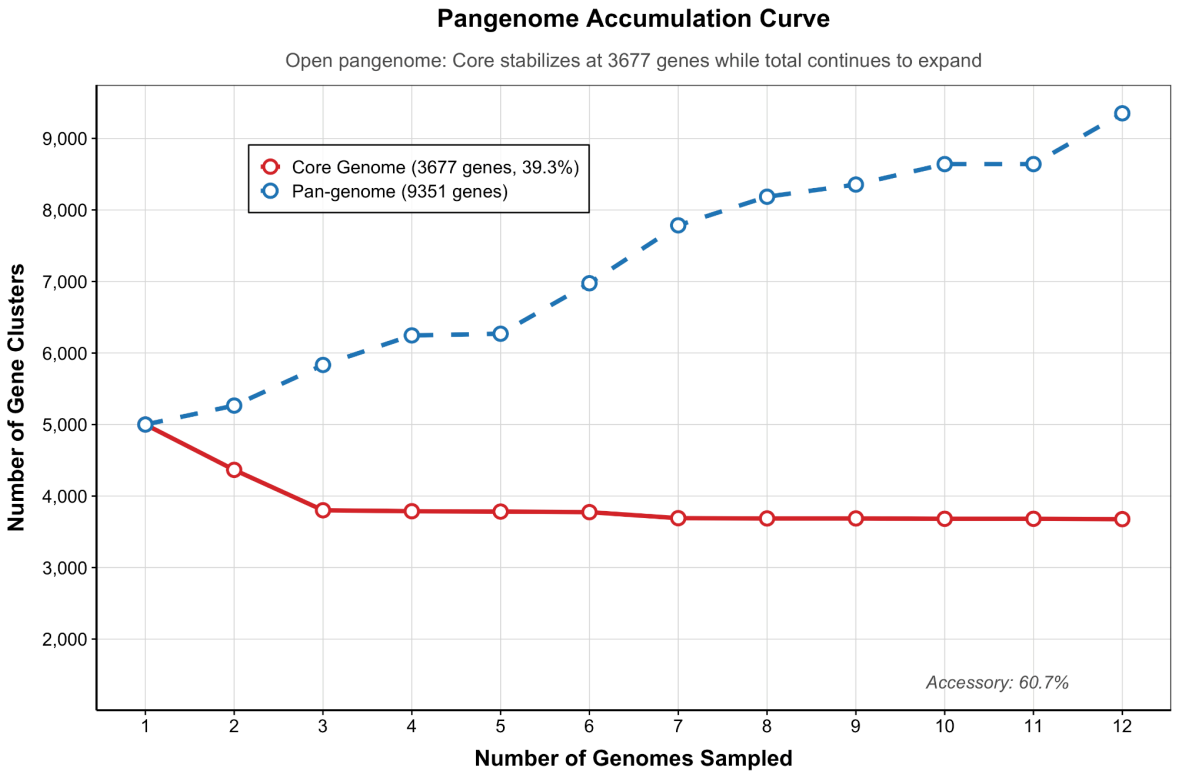

Figure 10. Pan-genome accumulation curve demonstrating open pan-genome dynamics. The core genome (red line) stabilizes at 3,677 genes after sampling 4-5 genomes, while the total pan-genome (blue line) continues to expand throughout the sampling process, reaching 9,351 genes with no evidence of saturation. The continuous upward trajectory indicates ongoing gene discovery potential.

The gene accumulation analysis revealed distinct evolutionary patterns between core and accessory genome components. The core genome achieved stability after sampling 4-5 genomes and remained constant at 3,677 genes throughout the analysis, indicating that essential *Brucella* functions are well-conserved and consistently present across lineages. In contrast, the total pan-genome exhibited continuous expansion throughout the sampling process, with each additional genome contributing an average of 400-500 new genes to the collective gene pool.

Extrapolation modeling suggests that the complete *Brucella* pan-genome likely exceeds 10,000-12,000 genes when accounting for additional lineages and environmental isolates not yet sequenced. This substantial genetic reservoir indicates that *Brucella* species collectively possess extensive adaptive potential through horizontal gene transfer, genomic innovation, and lineage-specific gene acquisition strategies.

4.8.3 Gene Frequency Distribution and Evolutionary Patterns

Gene frequency analysis revealed a characteristic bimodal distribution pattern indicative of distinct evolutionary forces operating on core versus accessory genome components. This distribution pattern provides insights into the molecular mechanisms driving Brucella genome evolution and adaptation strategies.

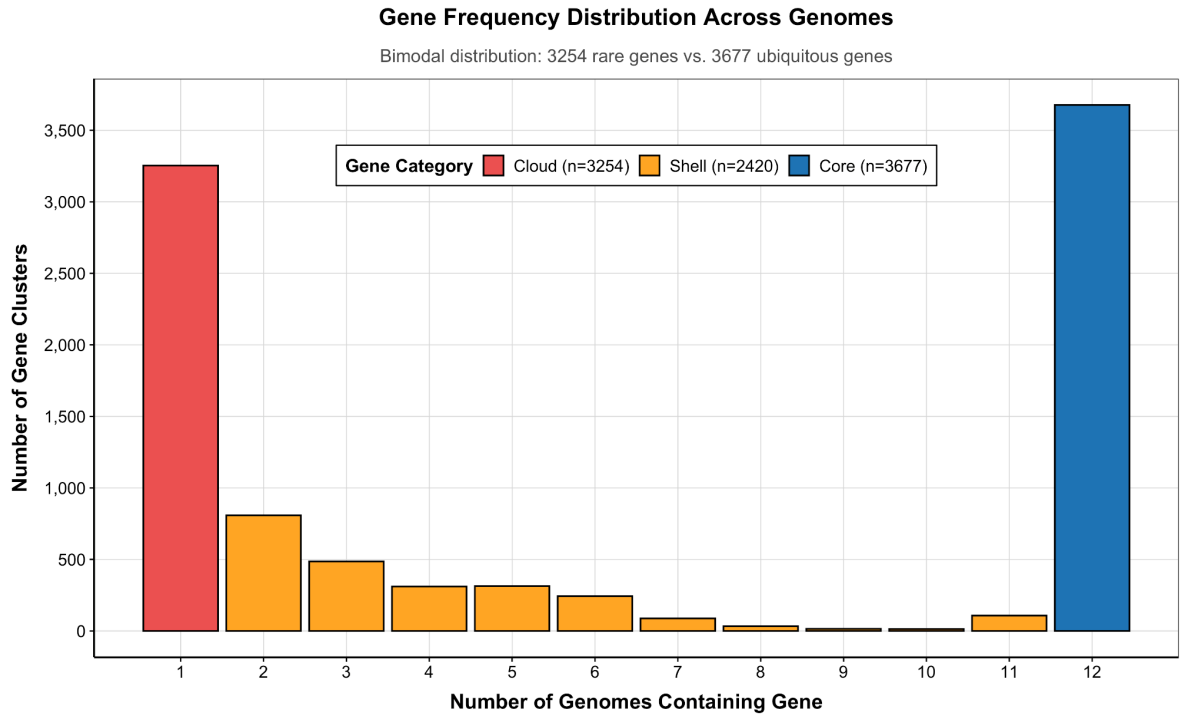

Figure 11. Gene frequency distribution across Brucella genomes displaying characteristic bimodal pattern. The distribution shows two distinct peaks: 3,254 rare genes (red, cloud category) present in single genomes and 3,677 ubiquitous genes (blue, core category) present in all genomes. The intermediate shell genes (orange) show gradual frequency decline, reflecting different selective pressures and evolutionary constraints.

The bimodal distribution reflects two fundamentally different evolutionary processes operating simultaneously within *Brucella* genomes. The left peak represents recently acquired or lineage-specific genes subject to neutral evolution, genetic drift, and positive selection for novel functions. These genes frequently encode mobile genetic elements, environmental adaptation factors, and strain-specific capabilities that provide competitive advantages in particular ecological niches.

The right peak represents ancient, conserved genes under strong purifying selection that maintain essential cellular functions. These genes exhibit vertical inheritance patterns, minimal horizontal transfer, and strong functional constraints that prevent their loss or significant modification. The clear separation between these peaks indicates limited intermediate-frequency genes, suggesting that most genes either achieve fixation across lineages (due to essential function) or remain rare (due to specialized or recently acquired function).

4.8.4 *B. anthropi*-Specific Gene Content and Adaptive Potential

Analysis of *B. anthropi*-specific gene content within the pan-genome context revealed distinctive patterns that support its classification as an environmentally adapted opportunistic pathogen. The strain contributed substantial unique genetic material to the cloud genome, with enrichment in functional categories associated with environmental persistence and metabolic versatility.

Functional annotation of *B. anthropi*-specific genes revealed enrichment in several key categories: (1) alternative carbon metabolism pathways enabling utilization of diverse environmental carbon sources; (2) heavy metal resistance determinants facilitating survival in contaminated industrial environments; (3) oxidative stress response systems providing protection against reactive oxygen species; (4) biofilm formation and surface adhesion factors supporting environmental persistence; and (5) mobile genetic elements indicating active horizontal gene transfer capabilities.

The substantial representation of *B. anthropi* genes within the cloud genome (estimated 250-300 strain-specific genes) indicates recent acquisition of environmental adaptation capabilities through horizontal gene transfer, genomic innovation, and positive selection for ecological competitiveness. This genetic repertoire enables *B. anthropi* to occupy environmental niches that are inaccessible to classical *Brucella* pathogens while retaining the core genetic machinery necessary for opportunistic infection of susceptible hosts.

4.8.5 Evolutionary Implications and Horizontal Gene Transfer

The open pan-genome structure provides compelling evidence that horizontal gene transfer represents a major evolutionary force shaping *Brucella* genome diversity and adaptive capabilities. The continuous gene discovery pattern, large accessory genome, and substantial strain-specific content indicate ongoing genetic exchange between *Brucella* lineages and environmental bacterial communities.

The pan-genome analysis supports the synteny and phylogenetic findings regarding *B. anthropi* evolution, demonstrating that environmental adaptation involves both chromosomal rearrangement (synteny analysis) and gene content innovation (pan-genome analysis). The combination of structural plasticity and gene acquisition capabilities positions *B. anthropi* as a highly adaptable lineage capable of rapid evolutionary responses to changing environmental conditions, emerging ecological opportunities, and novel selective pressures.

The findings have significant implications for understanding *Brucella* evolution, environmental persistence, and pathogenic potential. The open pan-genome structure suggests that *Brucella* species collectively represent a dynamic gene pool with extensive adaptive potential, while the substantial core genome ensures maintenance of fundamental cellular and pathogenic capabilities. This genomic architecture enables the simultaneous evolution of environmental specialists (such as *B. anthropi*) and host-adapted pathogens within a unified phylogenetic framework.

5. Discussion

5.1 Methodological advances in bacterial genome assembly

Our hybrid assembly strategy shows that chromosome-level, gap-free *B. anthropi* genomes are achievable when long-read scaffolding is combined with systematic short-read polishing. The Trycycler-based consensus plus multiple short-read polishing steps provides a practical template other groups can reuse for multipartite and repeat-rich bacterial genomes. High BUSCO completeness and absence of residual gaps indicate that platform-specific errors were effectively suppressed and that the resulting assembly can support demanding downstream comparative analyses[60].

5.2 Evolutionary biology and ecological adaptation

The phylogenomic framework places *B. anthropi* within an environmental and opportunistic branch of the expanded *Brucella* genus, distinct from classical intracellular zoonotic species. The clustering of environmental and clinical isolates supports an ecology-driven model of opportunistic infection rather than long-term host specialization. An open pan-genome with a large accessory component indicates that horizontal gene transfer is a major driver of *B. anthropi* evolution and provides a substrate for rapid adaptation and acquisition of novel traits[61].

5.3 Chromosomal architecture and functional specialization

Contrasting evolutionary patterns across the two chromosomes are consistent with models of multipartite genome evolution in α-proteobacteria, where a conserved primary replicon carries core functions and a more plastic secondary replicon supports niche adaptation. High synteny on chromosome I and extensive rearrangements on chromosome II indicate strong purifying selection on essential genes and relaxed constraints on accessory regions, respectively. Rearrangement hotspots enriched in intergenic and accessory clusters suggest that genome plasticity is structured to protect core functions while permitting adaptive reorganization[62].

5.4 Clinical implications and public health significance

The complete genome provides a high-quality reference for *B. anthropi*, an emerging environmental and nosocomial opportunist, and offers stable targets for species-level identification and differentiation from classical *Brucella*. The resolved resistome and virulence repertoire can inform both diagnostic assay design and therapeutic decision-making, particularly in device-associated and immunocompromised hosts. The open pan-genome and evidence of recent gene flow highlight the potential for continued acquisition of resistance and virulence determinants, arguing for inclusion of environmental *Brucella* in genomic surveillance programs[63].

5.5 Future directions and limitations

Functionally, many accessory and cloud genes remain uncharacterized, and experimental work will be needed to link these loci to specific ecological and pathogenic roles. Applying the same high-quality hybrid assembly and pan-genome framework to additional environmental and clinical *Brucella* isolates will clarify how genome architecture, gene flow and lifestyle interact across the genus. Longitudinal sampling from clinical and environmental reservoirs, coupled with comparative genomics against related α-proteobacteria, should help resolve which features of multipartite genome organization and pan-genome openness are unique to *Brucella* and which reflect broader principles of opportunistic pathogenesis.[64]

Bibliography

1. Scholz, H. C. *et al.* Brucella inopinata sp. nov., isolated from a breast implant infection. *Int. J. Syst. Evol. Microbiol.* **60**, 801–808 (2010).
2. Wattam, A. R. *et al.* Comparative Phylogenomics and Evolution of the Brucellae Reveal a Path to Virulence.
3. Wick, R. R. *et al.* Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol.* **22**, 266 (2021).
4. Wick, R. R. & Holt, K. E. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLoS Comput. Biol.* **18**, e1009802 (2022).
5. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-3212. - Google Search. https://www.google.com/search?client=safari&rls=en&q=Sim%C3%A3o+FA%2C+Waterhouse+RM%2C+Ioannidis+P%2C+Kriventseva+EV%2C+Zdobnov+EM.+BUSCO%3A+assessing+genome+assembly+and+annotation+completeness+with+single-copy+orthologs.+Bioinformatics.+2015%3B31(19)%3A3210-3212.&ie=UTF-8&oe=UTF-8.
6. Schwengers, O. *et al.* Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genomics* **7**, 000685 (2021).

7. Paulsen, I. T. *et al.* The Brucella suis genome reveals fundamental similarities between animal and plant pathogens and symbionts. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 13148–13153 (2002).
8. DelVecchio, V. G. *et al.* The genome sequence of the facultative intracellular pathogen Brucella melitensis. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 443–448 (2002).
9. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation - PubMed. https://pubmed.ncbi.nlm.nih.gov/28298431/.
10. DSpace. https://iris.who.int/items/6d8678fb-a583-41f3-bbe1-04420d4f416b.
11. Pappas, G., Papadimitriou, P., Akritidis, N., Christou, L. & Tsianos, E. V. The new global map of human brucellosis. *Lancet Infect. Dis.* **6**, 91–99 (2006).
12. Moreno, E. *et al.* Brucella abortus 16S rRNA and lipid A reveal a phylogenetic relationship with members of the alpha-2 subdivision of the class Proteobacteria. *J. Bacteriol.* **172**, 3569–3576 (1990).
13. Foster, J. T. *et al.* Whole-genome-based phylogeny and divergence of the genus Brucella. *J. Bacteriol.* **191**, 2864–2870 (2009).
14. Celli, J. *et al.* Brucella evades macrophage killing via VirB-dependent sustained interactions with the endoplasmic reticulum. *J. Exp. Med.* **198**, 545–556 (2003).

15. Starr, T., Ng, T. W., Wehrly, T. D., Knodler, L. A. & Celli, J. Brucella intracellular replication requires trafficking through the late endosomal/lysosomal compartment. *Traffic* **9**, 678–694 (2008).

16. Chain, P. S. G. *et al.* Whole-genome analyses of speciation events in pathogenic Brucellae. *Infect. Immun.* **73**, 8353–8361 (2005).

17. Halling, S. M. *et al.* Completion of the genome sequence of Brucella abortus and comparison to the highly similar genomes of Brucella melitensis and Brucella suis. *J. Bacteriol.* **187**, 2715–2726 (2005).

18. Tsolis RM, Seshadri R, Santos RL, et al. Genome degradation in Brucella ovis corresponds with narrowing of its host range and tissue tropism. PLoS One. 2009;4(5):e5519. - Google Search. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005519.

19. Characterisation of the genetic diversity of Brucella by multilocus sequencing - PubMed. https://pubmed.ncbi.nlm.nih.gov/17448232/.

20. Godfroid, J. *et al.* Brucellosis at the animal/ecosystem/human interface at the beginning of the 21st century. *Prev. Vet. Med.* **102**, 118–131 (2011).

21. De, B. K. *et al.* Novel Brucella strain (BO1) associated with a prosthetic breast implant infection. *J. Clin. Microbiol.* **46**, 43–49 (2008).

22. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genom. 2017;3(10):e000132. - Google Search. https://www.google.com/search?client=safari&rls=en&q=Wick+RR%2C+Judd+LM%2C+Gorrie+CL%2C+Holt+KE.+Completing+bacterial+genome+asse

mblies+with+multiplex+MinION+sequencing.+Microb+Genom.+2017%3B3(10)%3Ae000132.&ie=UTF-8&oe=UTF-8.

23. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 90 (2018).

24. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

25. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).

26. Moreno, E. *et al.* Brucella abortus 16S rRNA and lipid A reveal a phylogenetic relationship with members of the alpha-2 subdivision of the class Proteobacteria. *J. Bacteriol.* **172**, 3569–3576 (1990).

27. Foster, J. T. *et al.* Whole-genome-based phylogeny and divergence of the genus Brucella. *J. Bacteriol.* **191**, 2864–2870 (2009).

28. David Bruce (microbiologist). *Wikipedia* (2026).

29. Halling SM, Peterson-Burch BD, Bricker BJ, et al. Completion of the genome sequence of Brucella abortus and comparison to the highly similar genomes of Brucella melitensis and Brucella suis. J Bacteriol. 2005;187(8):2715-2726. - Google Search. https://pubmed.ncbi.nlm.nih.gov/15805518/.

30. Verger, J. M., Grimont, F., Grimont, P. A. & Grayon, M. Taxonomy of the genus Brucella. *Ann. Inst. Pasteur Microbiol.* **138**, 235–238 (1987).

31. Scholz, H. C. *et al.* Brucella inopinata sp. nov., isolated from a breast implant infection. *Int. J. Syst. Evol. Microbiol.* **60**, 801–808 (2010).

32. Verger, J. M., Grimont, F., Grimont, P. A. & Grayon, M. Taxonomy of the genus Brucella. *Ann. Inst. Pasteur Microbiol.* **138**, 235–238 (1987).

33. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

34. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

35. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

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

37. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing - PubMed. https://pubmed.ncbi.nlm.nih.gov/22506599/.

38. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

39. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).

40. Vaser, R. & Šikić, M. Time- and memory-efficient genome assembly with Raven. *Nat. Comput. Sci.* **1**, 332–336 (2021).

41. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).

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

43. Neuenschwander, S. *et al.* Evaluation of Oxford Nanopore Technologies workflows for genomic epidemiology of outbreak-associated bacterial isolates in the clinical setting. *Microb. Genomics* **12**, 001626 (2026).

44. Zimin, A. V. & Salzberg, S. L. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput. Biol.* **16**, e1007981 (2020).

45. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

46. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-2069. - Google Search. https://pubmed.ncbi.nlm.nih.gov/24642063/.

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

48. Feldgarden, M. *et al.* Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob. Agents Chemother.* **63**, e00483-19 (2019).

49. Seemann, T. tseemann/abricate. (2026).

50. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

51. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13950–13955 (2005).

52. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43(3):e15. - Google Search. https://pubmed.ncbi.nlm.nih.gov/25414349/.

53. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).

54. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490. - Google Search. https://pubmed.ncbi.nlm.nih.gov/20224823/.

55. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

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

57. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genomics* **3**, e000132 (2017).

58. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics* **2**, e000056 (2016).

59. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).

60. Trofimova, E., Asgharzadeh Kangachar, S., Weynberg, K. D., Willows, R. D. & Jaschke, P. R. A bacterial genome assembly and annotation laboratory using a virtual machine. *Biochem. Mol. Biol. Educ.* **51**, 276–285 (2023).

61. Yang, Z. *et al.* Comparative genomic analysis provides insights into the genetic diversity and pathogenicity of the genus Brucella. *Front. Microbiol.* **15**, (2024).

62. Genome of Ochrobactrum anthropi ATCC 49188T, a Versatile Opportunistic Pathogen and Symbiont of Several Eukaryotic Hosts | Journal of Bacteriology. https://journals.asm.org/doi/10.1128/jb.05335-11.

63. Liu, X., Zhang, R., Sun, M., Qiao, J. & Liang, M. Comparative genomics of Brucella species reveals key determinants of secondary metabolism, antimicrobial resistance, and virulence. *Sci. Rep.* **16**, 3765 (2025).

64. Yang, Z. *et al.* Comparative genomic analysis provides insights into the genetic diversity and pathogenicity of the genus Brucella. *Front. Microbiol.* **15**, (2024).