AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology®

EPIDEMIOLOGY

Core Genome Multilocus Sequence Typing Scheme for Improved Characterization and Epidemiological Surveillance of Pathogenic *Brucella*

Mostafa Y. Abdel-Glil,[a,b,c] Prasad Thomas,[d] Christian Brandt,[c] Falk Melzer,[a] Anbazhagan Subbaiyan,[d] Pallab Chaudhuri,[d] Dag Harmsen,[e] Keith A. Jolley,[f] Anna Janowicz,[g] Giuliano Garofolo,[g] Heinrich Neubauer,[a] Mathias W. Pletz[c]

[a]Institute of Bacterial Infections and Zoonoses, Friedrich-Loeffler-Institut, Jena, Germany
[b]Faculty of Veterinary Medicine, Zagazig University, Zagazig, Sharkia Province, Egypt
[c]Institute for Infectious Diseases and Infection Control, Jena University Hospital – Friedrich Schiller University, Jena, Germany
[d]Division of Bacteriology and Mycology, ICAR-Indian Veterinary Research Institute, Izatnagar, Uttar Pradesh, India
[e]Department of Periodontology and Operative Dentistry, University Hospital Muenster, Muenster, Germany
[f]Department of Zoology, University of Oxford, Oxford, United Kingdom
[g]WOAH and National Reference Laboratory for Brucellosis, Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "G. Caporale", Teramo, Italy

ABSTRACT Brucellosis poses a significant burden to human and animal health worldwide. Robust and harmonized molecular epidemiological approaches and population studies that include routine disease screening are needed to efficiently track the origin and spread of *Brucella* strains. Core genome multilocus sequence typing (cgMLST) is a powerful genotyping system commonly used to delineate pathogen transmission routes for disease surveillance and control. Except for *Brucella melitensis*, cgMLST schemes for *Brucella* species are currently not established. Here, we describe a novel cgMLST scheme that covers multiple *Brucella* species. We first determined the phylogenetic breadth of the genus using 612 *Brucella* genomes. We selected 1,764 genes that were particularly well conserved and typeable in at least 98% of these genomes. We tested the new scheme on 600 genomes and found high agreement with the whole-genome-based single nucleotide polymorphism (SNP) analysis. Next, we applied the scheme to reanalyze the genome of *Brucella* strains from epidemiologically linked outbreaks. We demonstrated the applicability of the new scheme for high-resolution typing required in outbreak investigations as previously reported with whole-genome SNP methods. We also used the novel scheme to define the global population structure of the genus using 1,322 *Brucella* genomes. Finally, we demonstrated the possibility of tracing distribution of *Brucella* strains by performing cluster analysis of cgMLST profiles and found nearly identical cgMLST profiles in different countries. Our results show that sequencing depth of more than 40-fold is optimal for allele calling with this scheme. In summary, this study describes a novel *Brucella*-wide cgMLST scheme that is applicable in *Brucella* molecular epidemiology and helps in accurately tracking and thus controlling the sources of infection. The scheme is publicly accessible and should represent a valuable resource for laboratories with limited computational resources and bioinformatics expertise.

KEYWORDS *Brucella*, genomic typing, cgMLST, whole-genome typing, epidemiology, core genome MLST

Brucella belongs to the *Brucellaceae* family, and members of that genus are Gram-negative facultative intracellular bacterial pathogens (1). Brucellae are highly infectious and cause brucellosis, a zoonosis reported globally (2). Typically, Brucellae are animal pathogens but can also infect humans as an incidental host, and about 500,000 new human infections occur annually (3, 4). *Brucella* species are listed as a category B biological

Editor Daniel J. Diekema, University of Iowa College of Medicine

Copyright © 2022 Abdel-Glil et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Mostafa Y. Abdel-Glil, mostafa.abdelglil@fli.de.

The authors declare a conflict of interest. D. Harmsen is a co-founder of Ridom GmbH (Münster, Germany). The other authors declare that there are no conflicts of interest.

[This article was published on 19 July 2022 with an incomplete affiliation for Anna Janowicz and Giuliano Garofolo. The missing information was added in the current version, posted on 28 July 2022.]

Received 25 February 2022
Returned for modification 23 April 2022
Accepted 28 June 2022
Published 19 July 2022

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

warfare agent by the Centers for Disease Control and Prevention (CDC) due to their ability to undergo aerosolization (5, 6). The genus presently consists of 12 species based on the strains' host preference and pathogenicity, including *B. melitensis* (mainly reported in small ruminants), *B. abortus* (cattle), *B. suis* (pigs), *B. canis* (dogs), *B. ovis* (sheep), and *B. neotomae* (woodrats) (7) along with six recently described novel species: *B. pinnipedialis* (cetaceans and seals), *B. ceti* (cetaceans and seals) (8), *B. microti* (common vole) (9), *B. inopinata* (breast implant) (10), *B. papionis* (baboons) (11), and *B. vulpis* (red foxes) (12).

Of all *Brucella* species, human brucellosis is mostly due to *B. melitensis*, *B. abortus*, and *B. suis*. Except for a few reports for *B. canis* and *B. ceti* (associated with marine mammals), the role of other *Brucella* species in human infection is questionable (13–15). Studies also showed greater variations in brucellosis prevalence in different animal species (16). In animals, cross-species transmission of *Brucella* spp. is also reported and can involve domestic and wild animals (17). Control of brucellosis in animal reservoirs is key to controlling infections in humans. However, the involvement of multiple bacterial species, zoonoses associated with the handling of infected animals and materials, and foodborne infections are indicative of the complex epidemiology of brucellosis. Across different continents, the prevalence of the disease is highly variable, showing an endemic nature in a few hosts in the African and Asian continents with an average prevalence range of 0 to 88.8% in sheep and goats and 0 to 68.8% in cattle (18).

Classical typing of *Brucella* species involves identification of different biovars. *Brucella* species can be classified based on serological differences, phenotypic features, and their susceptibility to phages and chemicals (7, 19). Genotypic approaches commonly employed for typing include single nucleotide polymorphism (SNP) analysis of the *rpoB* gene (20), multiple-locus variable-number tandem-repeat typing (MLVA) (21, 22), and hypervariable octameric oligonucleotide fingerprint (HOOF) variable number tandem repeats (23–25). MLVA typing is a widely used method characterizing different *Brucella* species (21, 26). Recently, sequence-based approaches based on multilocus sequence typing (MLST) (27, 28) that offer the possibility of web-based analysis and comparison are increasingly used for differentiating strains within the genus (29–32). Currently, two classical schemes based on nine and 21 MLST loci have been described (27, 28).

The development of typing tools that are also publicly available is of paramount importance for bacterial disease investigation, epidemiology, and outbreak surveillance. Whole-genome sequencing (WGS) has become an increasingly vital tool used by many laboratories to study the relatedness of bacterial strains in outbreaks. The most common methods include the gene-by-gene approach or SNP-based typing with or without a reference strain sequence. Typically, gene-by-gene allele calling comparison is based on the core genome (cgMLST), where only core genes are considered, or on the whole genome (wgMLST), which combines core and accessory genes for strain typing (33). For *Brucella*, a cgMLST scheme was initially developed based on 407 *Brucella* genomes (34). The scheme includes 164 targets and distinguishes *Brucella* down to the species and biovar level. However, the ultimate small number of target genes hampers the scheme's utility for epidemiological analyses. At the species level, cgMLST schemes have been described only for *B. melitensis*. The first scheme comprised 2,704 genes (35), and the other *B. melitensis*-specific scheme included 2,656 genes (36). The *B. melitensis* schemes have proven useful for high-resolution characterization in outbreak situations (30).

This study aimed to develop a universal, web-accessible cgMLST scheme that applies to all *Brucella* species. We included 1,325 *Brucella* genomes from public repositories and used them for cgMLST scheme development and evaluation (see Table S1a in the supplemental material). Here, we demonstrated the extensibility of the scheme for defining the global population structure of the genus *Brucella*. We compared the scheme resolution with the core-genome SNP approach and demonstrated the scheme's applicability for studying brucellosis outbreaks. In addition, we have made our scheme publicly available (https://pubmlst.org/bigsdb?db=pubmlst_brucella_seqdef&page=schemeInfo&scheme_id=3) and introduced a gene-by-gene nomenclature system to harmonize the genomic characterization of *Brucella* strains and support the tracking of transmission chains

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

in *Brucella* outbreaks. Finally, we benchmarked a data set of genomes with various read depths and examined the impact of read depth on the typing results of cgMLST.

MATERIALS AND METHODS

***Brucella* genome retrieval and processing.** In this study, we included 1,325 *Brucella* genomes from public repositories and used them for cgMLST scheme development and evaluation (see Table S1a in the supplemental material). For scheme development, we downloaded the *Brucella* genomes from the National Center for Biotechnology Information (NCBI) RefSeq database following the taxonomy number 234. Of the 783 genomes downloaded (March 2021), we excluded 57 genomes that were based only on sequencing technologies prone to high error rates, e.g., 454/Roche, ONT, and Ion Torrent. We also excluded 92 genomes with completeness scores less than 95% and contamination and heterogeneity scores more than 5% using CheckM v1.1.3 (37). The remaining genomes ($n = 634$) were used for taxonomical analysis by estimating the level of nucleotide similarity with FastANI v1.3 (38), and by performing a protein-based phylogeny using up to 400 universal bacterial proteins with PhyloPhlan v0.43 (39). We then used Mash v2.1 (40) for species confirmation. As a result, we included only 612 genomes for scheme development (Table S1a).

For scheme evaluation, we additionally used 600 random genomes from the Short Read Archive (SRA) database, for which the metadata on isolation sources were available and the sequencing was done with Illumina platforms (Table S1a). The fasterq-dump module of SRAtoolkit v2.9.2 was used to obtain files in FASTQ format (https://github.com/ncbi/sra-tools), fastp v0.20 was used to check the quality of the sequencing reads (41), and shovill v1.0.4 (SPAdes v3.12 [42]; flags –trim –minlen 500 –mincov 5) was used to perform genome assembly (43). The quality of genome assemblies was checked with Quast v4.3 (44), and we retained genomes with an N_{50} greater than 20,000 bases, and a number of contigs less than 1,000. Finally, Mash v2.1 was used to confirm the species of all genomes.

To evaluate the scheme for outbreak analysis, we reinvestigated published outbreak strains including 37 *B. melitensis* strains involved in outbreaks in central Italy (35), and a further 76 *B. abortus* strains implicated in several outbreaks in Northern Ireland (45) (Table S1a). The raw reads were obtained from the SRA database and processed using fastp v0.20, shovill v1.0.4, Quast v4.3, and Mash v2.1 as mentioned above.

Development of cgMLST for *Brucella*. We first performed high-resolution phylogeny of the genus. We identified consensus SNP positions with NASP pipeline v1.1.2 (46), for which we used as input genome assemblies aligned with NUCmer (47) to the reference genome of strain 16M. SNP positions falling into the duplicated regions were masked. Consensus SNPs were identified in each genome relative to the reference genome. We then used RAxML v8.2.9 (48) to perform a maximum-likelihood phylogenetic analysis using ascertainment bias correction and the gamma model of evolution.

Next, we used the genome of strain 16M (GenBank accession numbers NC_003317.1 and NC_003318.1; 25 October 2020) as a reference. The cgMLST Target Definer tool v1.5 of SeqSphere+ v7 was then applied in default mode to remove short, multicopy, and overlapping genes as well as genes with internal stop codons from the reference genome (49). As a result, 2,763 genes were kept. We then used these genes to genotype the quality-controlled 612 genomes from the RefSeq database. The percentage of the good targets was then estimated per each genotyped genome. A threshold of 98% was applied. Genes with an allele call rate below 98% were manually excluded from the final cgMLST scheme. The default settings of SeqSphere+ v7 for allele calling were used, including BLAST detection with BLASTN v.2.2.12 at sequence identity above 90% and sequence overlap above 99%.

Application of *Brucella* cgMLST. SeqSphere+ v7 (49) was used to detect cgMLST genes and allele calls. Gene detection was done using BLASTN (50) at a sequence identity of >90% and an overlap of >99%. Allele assignment involved only intact genes without frameshifts or sequence ambiguities. In-frame insertions or deletions of up to three codons relative to the reference genes were allowed. The cgMLST allelic profiles were then created using a combination of alleles identified for each gene. Finally, the allelic profiles were compared (uncalled genes were ignored in pairwise comparisons), and the resultant pairwise distances were used to generate a neighbor-joining (NJ) tree and a minimum spanning tree (MST) using SeqSphere+. The Simpson index of diversity and adjusted Wallace test of congruence were calculated using the Comparing Partitions tool (51). For the outbreak data set, we calculated core genome SNPs using the snippy pipeline v 4.6.0 (https://github.com/tseemann/snippy) with default settings.

Allele calling of cgMLST at various sequencing depths. We used a subset of 40 genomes available in the SRA database to examine the impact of sequencing depth on the allele calling rate of cgMLST and cluster definitions. Briefly, we assessed genome quality using fastp v0.20.0 in standard mode (41) and SeqKit v2.2.0 (52) to generate statistics of FASTQ files including Phred quality scores and total bases sequenced. ConFindr v0.7.4 (53) was then used to detect contamination. For ConFindr, the rMLST database (54) was also employed to identify contaminants based on variations in ribosomal proteins.

Next, we created a data set *in silico* with read depths between 10- and 100-fold. For this, we used rasusa v0.6.1 (55) to randomly downsample sequencing reads to the required coverage using the genome size of the *B. melitensis* strain as a reference (flag -g 3.2MB). BBMap (version dated 24 February 2022) (56) was then used to independently estimate the obtained sequencing depth and breadth of coverage.

For each downsampled data set, SKESA v2.4.0 (57), MEGAHIT v1.2.9 (58), and SPAdes v3.15.4 (42) were used for assembly. All assemblers were used with default settings except SPAdes, which was run with the flag (–careful) to reduce the number of mismatches and short indels. No refinement steps such as read trimming, polishing, or scaffolding were added before or after assembly. The 1,200 genomes generated by the three assemblers were allele called based on the BLASTN algorithm (50) using the SeqSphere+ program (49) and default settings (90% identity and 99% alignment to reference sequences required).

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

Data availability. Genome sequence data examined in this study are publicly available under the accession numbers given in Table S1a.

RESULTS

Definition of a cgMLST scheme for *Brucella*-wide genotyping. To begin with the scheme setup, we first quality-filtered 634 *Brucella* genomes from the Reference Sequences (RefSeq; accessed March 2021) database and used them to define the overall genetic diversity of the genus (see Materials and Methods and see Table S1a in the supplemental material). The midpoint-rooted maximum-likelihood (ML) phylogeny, supported by the average nucleotide identity (ANI) analysis, divided the RefSeq genomes into two distinct groups: a small divergent group (group I) of 22 genomes from the species *Brucella intermedia* (formerly *Ochrobactrum intermedium* [59]) and *Brucella pituitosa* and a large group (group II) with 612 genomes (Table S1a and b). The genomes of the second group include the following *Brucella* species: *B. melitensis*, *B. abortus*, *B. suis*, *B. canis*, *B. ovis*, *B. neotomae*, *B. ceti*, *B. pinnipedialis*, *B. microti*, and *B. inopinata* (Table S1a). These species fall collectively into a single genomic species as described previously (60), with ANI values greater than 97%. Of the 612 genomes, 601 (98.2%) had ANI concordance more than 99% compared to the *B. melitensis* reference 16M genome (Fig. 1). ANI results between groups I and II were 80 to 83% (Fig. 1). Based on this divergence level, we excluded the 22 genomes of group I and focused our genome typing scheme on group II *Brucella* (Fig. 1).

We used 612 *Brucella* genomes (group II) to perform a high-resolution phylogeny based on the genomes' nucleotide content (Fig. S1). We determined 159,572 SNP positions (4.84%) in the genomes with the NASP pipeline (46). The phylogenetic analysis, consistent with previous reports (2, 26, 28, 61), showed characteristics of the global *Brucella* phylogeny, including (i) simultaneous evolutionary radiations of several descendant lineages, each lineage including a single or pair of species; (ii) clear separation of the genomes of each species into a distinct clade; (iii) ancestral relationships between some *Brucella* species, such as *B. canis*, was shown to have evolved later from the species *B. suis*, as was *B. pinnipedialis* in relation to *B. ceti* (2, 62); and (iv) a genetically more distant group of genomes (n = 11 genomes) positioned at the base of the phylogenetic tree, including one genome of the species *B. inopinata* and another 10 genomes of unassigned species. We note that 52 genomes in the data set were not originally reported for a specific *Brucella* species and that potential mislabeling of the species was identified (Table S1a). These results indicate that the 612 genomes cover the known global population structure of *Brucella* and are therefore suitable to set up a cgMLST scheme. However, two *Brucella* species were not considered: *B. papionis*, for which no assembly is available, and *B. vulpis*, for which the only two available genomes were not incorporated in the RefSeq database because of anomalous assembly (March 2021). The compiled genomic data set represented diverse geographic origins (54 countries of the six continents), isolation from diverse ecological sources including humans (n = 184) and animals (n = 291), and diverse isolation times (from 1931 to 2020; Table S1a).

Next, we used the 612 genomes to create a "soft defined" cgMLST scheme as described previously (63). Briefly, SeqSphere v7 removed 340 genes (192 incomplete, 16 repetitive, and 132 overlapping genes) from the reference genome (strain 16M; accession numbers, NC_003317.1 and NC_003318.1; 25 October 2020; total genes, 3,103). The retained 2,763 genes were then used to genotype the 612 genomes. The genes that were present and allele typed in at least 98% of the 612 genomes were placed in a core genome scheme, whereas the remaining genes were moved to an accessory genome scheme. As a result, 1,764 genes formed the core genome MLST scheme. The genes range in size from 60 to 4,134 bp, with a total length of 1,556,385 bp, covering 47.2% of the reference genome size and 56.8% of the reference genes with coding DNA sequences (Table S1c). The accessory genome MLST scheme included 1,131 genes with a total length of 1,093,278 bp.

Evaluation of cgMLST for high-resolution genotyping of *Brucella* genomes. To evaluate the performance of the novel scheme, we used an additional 600 genomes randomly obtained from the Short Read Archive (SRA) database, representing strains collected between 1940 and 2020: 184 were from humans, 450 from animals, and 42

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

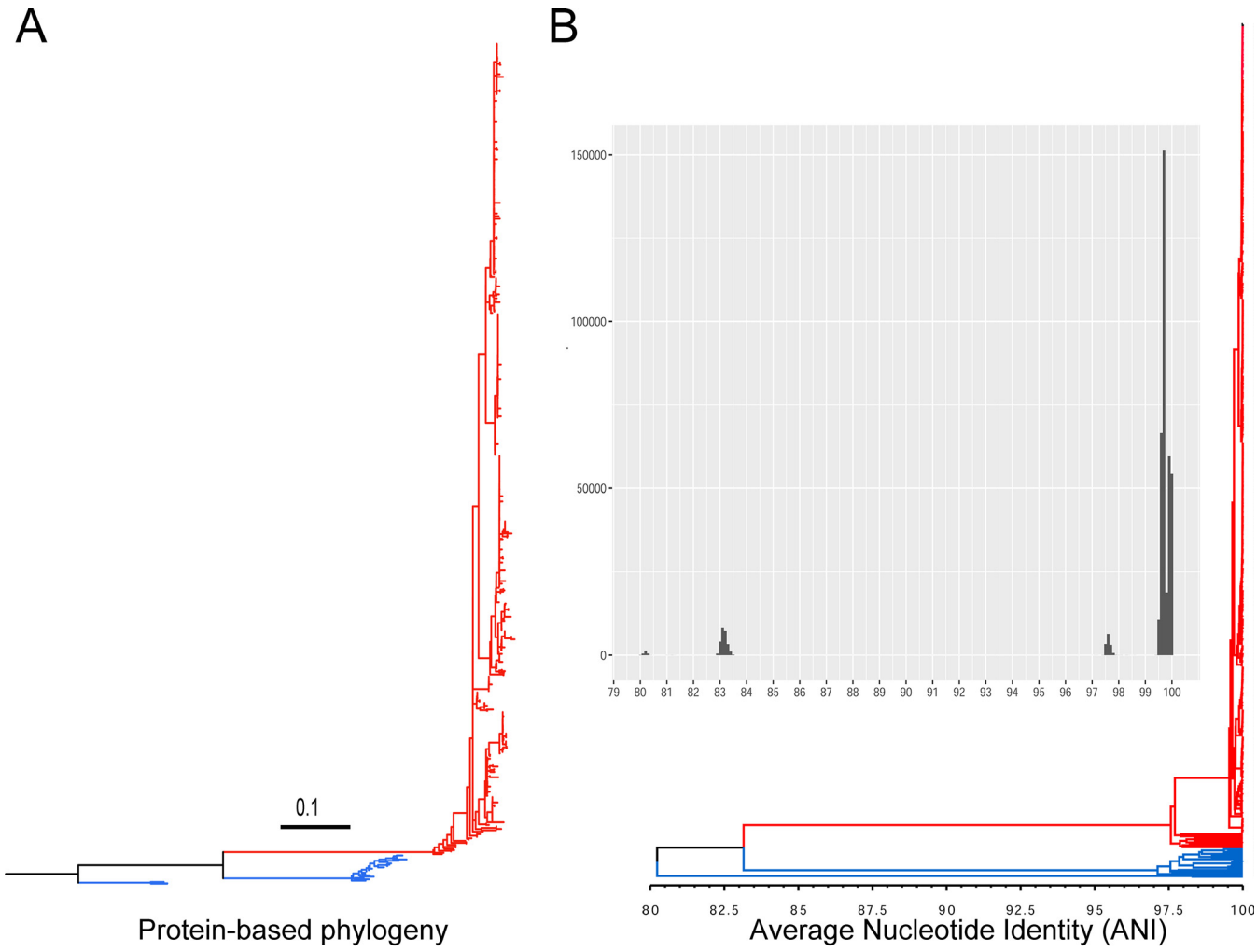

FIG 1 Taxonomic classification of 634 *Brucella* genomes downloaded from the NCBI Reference Sequence (RefSeq) database (March 2021). (A) A midpoint-rooted maximum-likelihood (ML) phylogeny of *Brucella* genomes calculated with PhyloPhlAn. The phylogeny is based on the concatenated alignment of 24,110 amino acid positions of up to 400 universally conserved bacterial proteins. (B) The results of pairwise average nucleotide identity (ANI) between all *Brucella* genomes calculated with FastANI and plotted with bactaxR. The distribution of the ANI values is represented in the histogram, and the relatedness between all genomes is illustrated by a dendrogram created using the average linkage hierarchical clustering method where the tree height corresponds to ANI similarity. The red branch denotes the genomes of all *Brucella* species while the blue branches denote divergent genomes from the species *B. intermedia* (*O. intermedium*) and *Brucella pituitosa*, as explained in the text.

from unknown hosts. The strains originated from 49 countries on the six continents. The 1,764 cgMLST targets were highly conserved in the 600 genomes, with an average allele call rate of 99.5% (median, 99.9%; standard deviation, 0.7%). The average number of variants found for each cgMLST gene was 12 ± 6.4 variants (range, 1 to 47), with a total number of SNP positions ranging from 0 to 177 per gene and a total of 49,705 SNP positions in all genes (Table S1c). The length of cgMLST genes correlated directly with the number of variants detected for each gene (Table S1c).

The cgMLST resulted in high-resolution genotyping. The 600 genomes were clustered into 476 cgMLST sequence types (cgST), ignoring missing values for pairwise comparisons. This resulted in a Simpson index of discrimination of 0.998 (95% confidence interval [CI] = 0.997 to 0.999). Analysis of the whole-genome SNP provided a slightly higher resolution than that of the cgMLST. We detected 92,651 SNPs across the 600 genomes. The number of unique SNP profiles was 531 (Simpson's index = 0.999, 95% CI = 0.999 to 1.000). Considering only SNP sites in the 1,764 cgMLST targets led to identifying an almost similar number of SNP profiles as with core genome MLST (n = 482).

Finally, the cgMLST had a higher resolution than the classical MLST schemes of the genus *Brucella*. In total, the 600 genomes were classified into 40 ST using the 9-gene

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

MLST scheme, whereas 596 of the 600 genomes (four genomes of undetermined MLST21 ST) were classified into 65 ST using the 21-gene MLST scheme, resulting in a Simpson index of diversity of 0.868 (95% CI = 0.853 to 0.882) and 0.898 (95% CI = 0.883 to 0.914), respectively (Table S1a).

Reanalysis of *Brucella* outbreaks with the novel scheme. We used the developed scheme to characterize published sets of epidemiologically linked cases of brucellosis involving *B. melitensis* and *B. abortus*. For *B. melitensis*, we included the genome of 37 strains of known epidemiological links described in a previous study (35). The strains were collected from a single outbreak in 21 farms from four provinces in central Italy (35). The study compared the resolution of core genome SNPs and classical MLVA typing with a *B. melitensis*-specific cgMLST scheme. The results showed that two distinct clusters were involved in the outbreak and that the inter- and intracluster distance resolution achieved with WGS methods was higher than that with classical MLVA.

We retrieved the raw sequence data from this previous study and reprocessed them (see Materials and Methods). Because the assembly workflow in our study differed from that of the original study, we first wanted to examine the effects of assembly on the cgMLST results. Therefore, we first reapplied the previously described *B. melitensis*-specific cgMLST (https://www.cgmlst.org/ncs/) to the genomes (35). While the results were highly reproducible, there were some minor differences in the number of different alleles between the genomes in each cluster (Fig. S2). The allelic distances within the two clusters were higher than previous results (missing values ignored for pairwise comparisons; Fig. S2). Cluster 1 had a mean difference of 3.8 (range 0 to 15) while cluster 2 had mean allelic differences of 5 alleles (range 2 to 7) (Fig. S2). One hundred sixty-one gene variants represent the intercluster genetic distance.

Next, we applied the *Brucella*-wide cgMLST to the 37 genomes. The results show that at least 99.8% of the scheme targets were detected in each genome, with all targets detected in 29 genomes (100% allele call rate) (Table S1a). The typing results were congruent with the original study, defining the two genetic clusters involved in the outbreak using a clustering threshold of three different genes between any two neighboring isolates; one isolate had 5 genes different from the closest neighbor (Fig. 2). The 37 genomes were indexed into 20 cgST (missing values were ignored for pairwise comparisons). Eighteen of the 37 genomes formed a single central cgST of cluster 1. The mean genetic variation of cluster 1 was 1.9 $\pm$ 1.9 alleles (median, 1; range, 0 to 8) and of cluster 2 was 2.3 $\pm$ 0.8 alleles (median, 2.5; range, 1 to 3) (Table S1d). The intercluster allelic variation was 100 distinct alleles. These results agree with the typing output of the *B. melitensis*-specific scheme, indicating that the two schemes produce comparable resolutions suitable for identifying the closely related strains.

For *B. abortus*, we reinvestigated the genomes of 76 strains involved in brucellosis outbreaks in cattle in nine different locations related to Divisional Veterinary Offices (DVOs) in Northern Ireland from 1991 to 2011 (45). The strains had limited genetic distances and were previously characterized with classical MLVA and WGS-based SNP analysis (45, 64). Two phylogenetic lineages were characterized, a small lineage of 4 strains and a large lineage of 72 strains. The cgMLST classified the genomes into 23 cgST where the allele call rate was 99.8% to 100% in all genomes (only 1 to 3 targets were missing in 7 genomes). In congruence with the SNP analysis, the MST of the cgMLST profiles indicated a limited heterogeneity for all genomes (Fig. 3). In addition, the two genetic lineages were identified with 67 to 72 differing alleles between the two lineages. Lineage 1 included two cgST that were distant by four alleles. Lineage 2 included 72 genomes with mean allele differences of 1.6 $\pm$ 1.5 (median, 1; range, 0 to 8 alleles) and a maximum of three different genes between any two neighboring isolates (Fig. 3 and Table S1e). These cgMLST results confirm the results of the previous whole-genome SNP analysis, indicating comparable performance (45).

Deciphering the population structure of *Brucella* with cgMLST. We deciphered the population structure of the genus *Brucella* using cgMLST (Fig. 4). Therefore, we used 1,322 (out of 1,325) genomes with an allele calling ratio over 90%. The characteristic population structure of *Brucella* as determined by cgMLST was consistent with the core

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

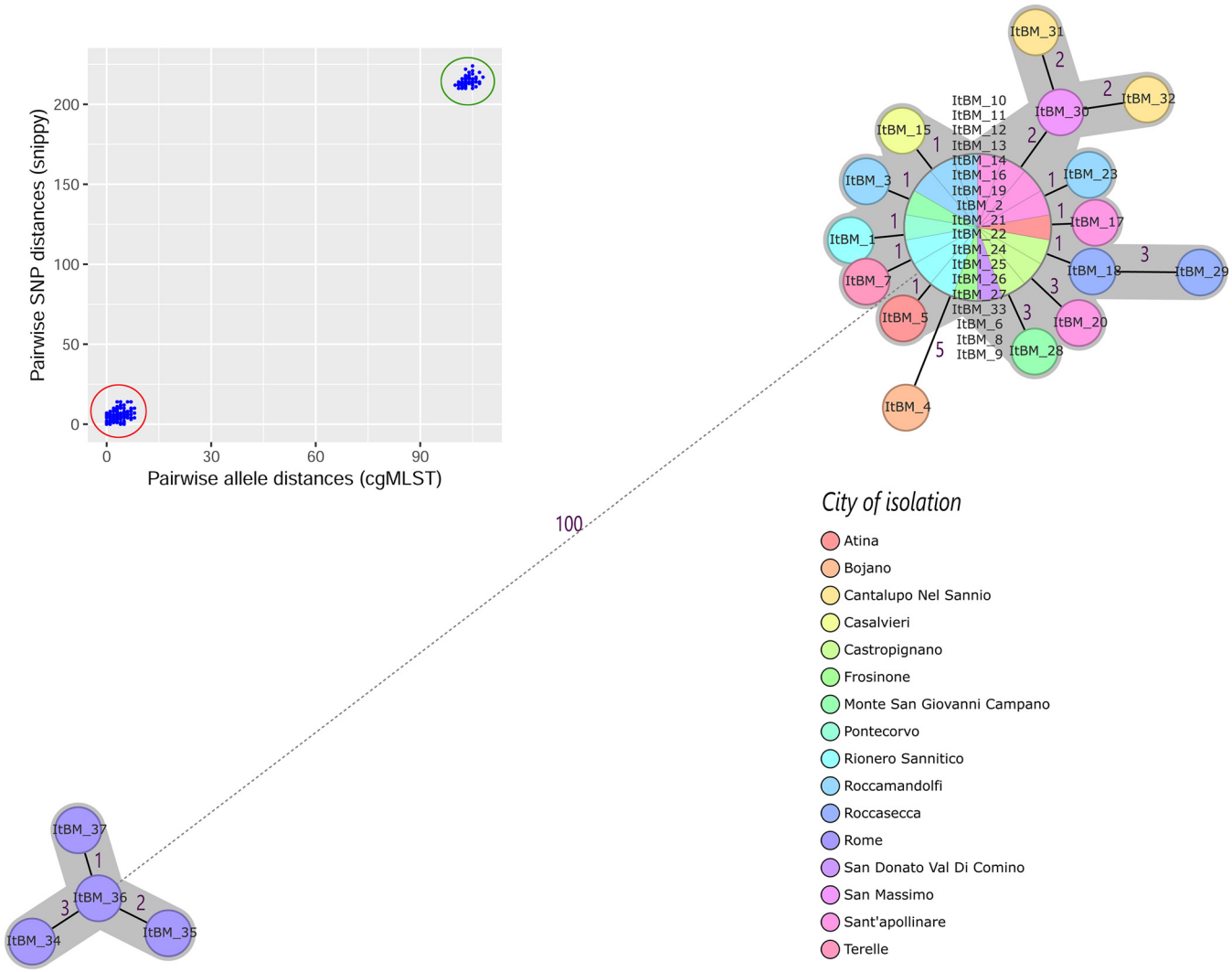

FIG 2 Minimum spanning tree (MST) calculated for 37 *Brucella melitensis* genomes with known epidemiological linkage using 1,764 *Brucella*-wide core genome MLST targets. The MST was generated with Ridom SeqSphere, ignoring missing values in pairwise comparisons. Each circle represents a unique cgMLST profile and is labeled according to the strain's city of isolation. The circle size is proportional to the number of genomes per each cgMLST genotype. The number of different alleles between cgMLST profiles is indicated on the connecting lines. Solid and dashed lines represent allele differences below and above 10, respectively. Clusters are highlighted with gray-shaded areas based on a threshold of three allele mismatches between any two neighbors. The inset box shows the comparison of cgMLST allele distance and core genome SNPs for each genome pair. Genomic distances (cgMLST alleles and SNPs) within and between outbreak strains are highlighted by red and blue circles, respectively. For a phylogeny based on the core genome SNPs, we refer the reader to the original publication (35).

genome SNP analysis (Fig. S3), demonstrating that each species is clearly divided into a distinct clade and that the individual lineages of each species are well resolved. In *B. melitensis* (n = 498 genomes), four main lineages were identified: (i) the West Mediterranean lineage, which was detected in the Middle East and Europe; (ii) the African lineage in Africa; (iii) the American lineage in North and South America and Europe; and (iv) the East Mediterranean lineage in Asia, Europe, and the Middle East. Here, the Mediterranean lineages were more abundant in the *B. melitensis* data set (90.7%, n = 452 genomes). In *B. abortus* (n = 472 genomes), the three main clades, A, B, and C, were identified. Clades A and B were less common and restricted to African countries (9.7%, n = 44 genomes). These two clades represent early-branching groups, in contrast to clade C, which represents a relatively recent branch of *B. abortus*. Clade C (subdivided into subclades C1 and C2) has a wide geographic distribution (90.6%, n = 428 genomes). In *B. suis*, two clades were identified: clade 1, including biovar 2 genomes, was mainly attributed to Europe, while clade 2, comprising biovars 1, 3, and 4 as well as *B. canis* genomes, was mostly

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

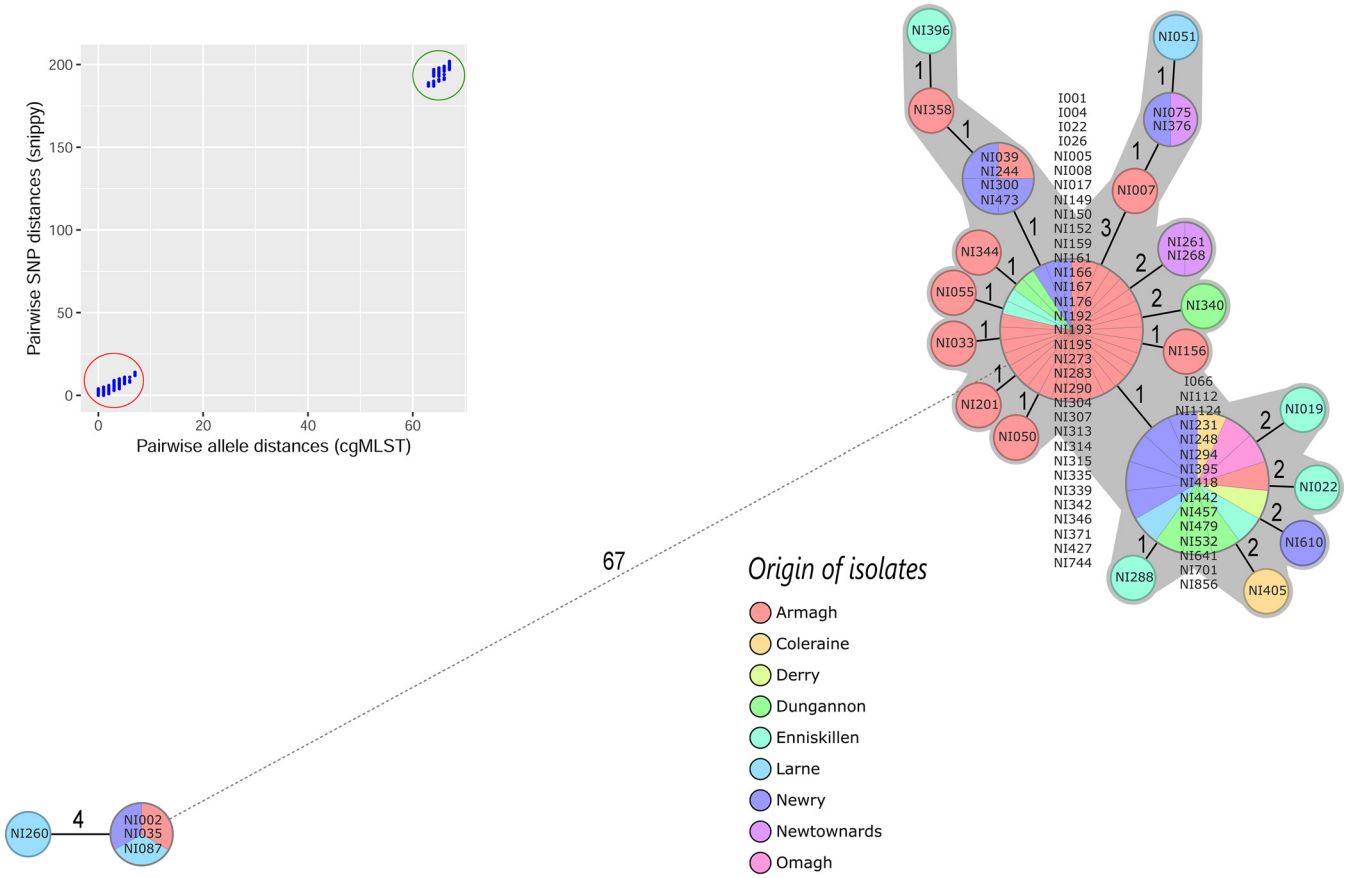

FIG 3 Minimum spanning tree (MST) calculated for 76 *Brucella abortus* genomes using 1,764 *Brucella*-wide core genome MLST targets. The MST was generated with Ridom SeqSphere, ignoring missing values in pairwise comparisons. Each circle represents a unique cgMLST profile and is labeled according to the origin of strains. The circle size is proportional to the number of genomes per each cgMLST genotype. The number of different alleles between cgMLST profiles is indicated on the connecting lines. Solid and dashed lines represent allele differences below and above 10, respectively. Clusters are highlighted with gray-shaded areas based on a threshold of three allele mismatches between any two neighbors. The inset box shows the comparison of cgMLST allele distance and core genome SNPs for each genome pair. Genomic distances (cgMLST alleles and SNPs) within and between outbreak strains are highlighted by red and blue circles, respectively. For a phylogeny based on the core genome SNPs, we refer the reader to the original publication (45).

attributed to North and South America, Asia, and Tonga. In addition, two *B. suis* genomes corresponding to MLST21 ST19 (*B. suis* biovar 5) formed a separate branch from the other *B. suis* genomes. The close phylogenetic relationship of *B. suis* and *B. canis* was also demonstrated with cgMLST, as was the close relationship between *B. ceti* and *B. pinnipedialis*, which were isolated mainly from marine mammals. *B. ovis* comprised five genomes from MLST21 ST13 that form a distinct lineage. Overall, these results show that the new scheme provides sufficient resolution to delineate the global population structure of *Brucella* strains.

Among the 1,322 genomes examined with the scheme, we observed nearly identical cgMLST profiles detected in different countries/laboratories (Table S1f). This included 58 genomes classified into 12 cgMLST cluster groups with fewer than three different alleles. Clusters included reference strains of the species and vaccine strains processed remotely in different laboratories. Unfortunately, the detailed epidemiological information on all genomes was sparse. However, some were reported in previous studies (31, 65), tracing strains back to patients with travel or migration history. This was shown, for example, in patients from Norway and Sweden who had a travel history to countries of endemicity such as African or Middle Eastern countries, but patients with no travel history also were reported, possibly indicating hidden transmission routes, e.g., via food vehicle or illegal animal trade. Similar cgMLST profiles were also observed in different host species, confirming transmission of *Brucella* between hosts.

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

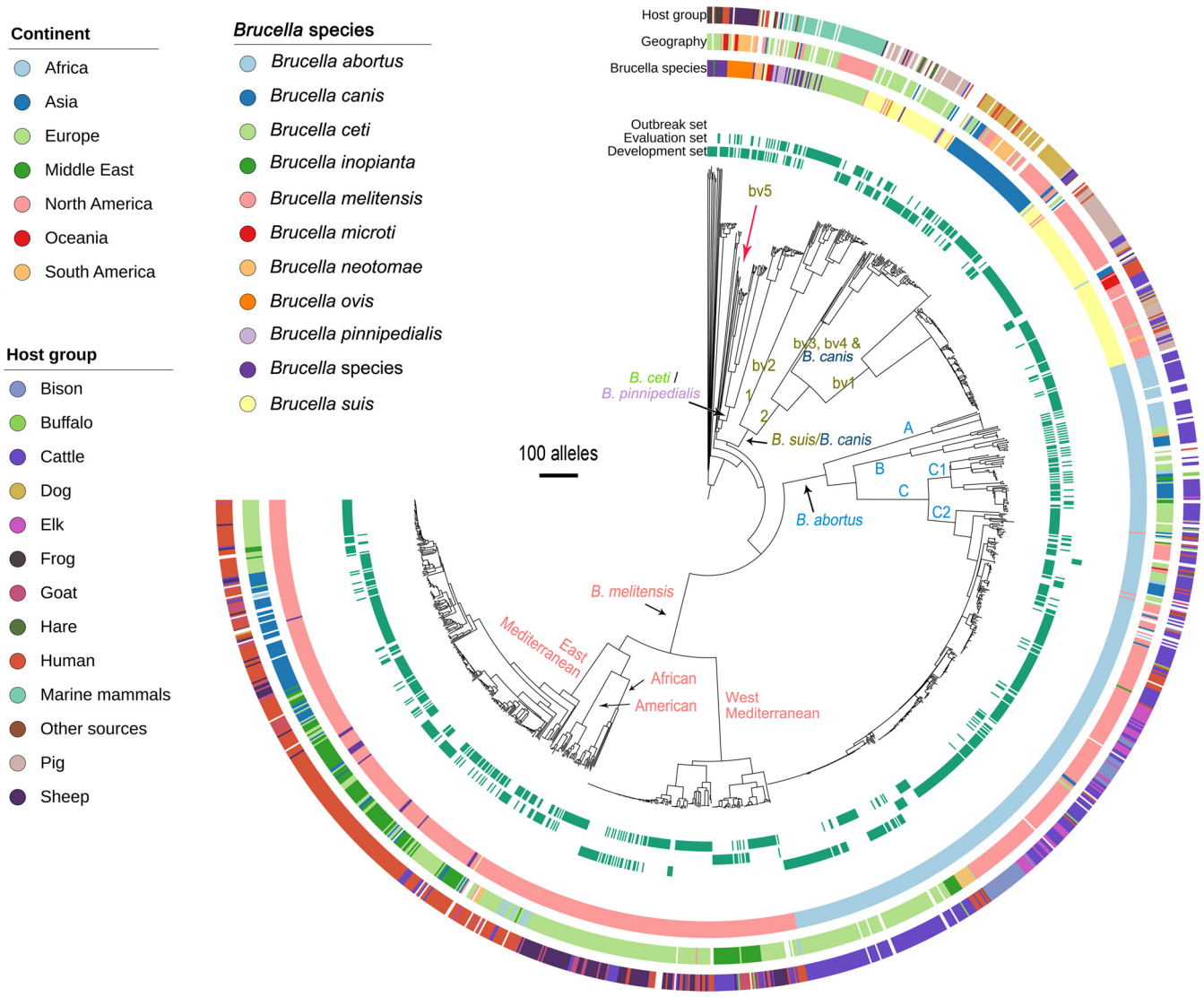

FIG 4 Neighbor-joining tree constructed for the 1,322 *Brucella* genomes based on the cgMLST allelic profiles, deciphering the characteristic population structure of pathogenic *Brucella*. Tree visualization was performed using iTOL.

Taken together, these results indicate that the typing efficacy enables correlation with the source country of infections and thus that trace-back analysis is possible based on related strains reported after a time span in different regions, laboratories, or hosts.

Impact of sequencing read depth on allele calling rate and cluster analysis of cgMLST. We aimed to evaluate the impact of sequencing depth on allele calling rate and cgST cluster definition. To this end, we collected 40 genomes from the SRA database that were sequenced with paired-end libraries with read lengths of 100, 150, 250, and 300 bp. This was based on our observation that 92.5% (n = 24,18) of *Brucella* genomes in the SRA database (accessed 31 May 2022, WGS data = 2,965 genomes) were based on Illumina platforms and were predominantly produced using paired-end libraries (94%, n = 2,278). As such, we based our evaluation procedure on Illumina data. For all included genomes, no evidence of contamination was found using the ConFindr tool (53) and the rMLST database (54). In addition, at least 80% of raw reads had Phred quality scores above 30. The estimated sequencing depth for all genomes was higher than 100-fold.

Based on these 40 genomes, we benchmarked a data set of downsampled reads. We randomly downsampled the genomes to obtain target read depth in the range of 10- to 100-fold (Table S1g). The outcome of this random downsampling procedure was

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

independently evaluated with the BBMap mapping tool, and we found high agreement between the target depth (i.e., the depth specified by downsampling) and the depth reported by BBMap (56) ($R^2 > 99.9\%$; Pearson test, P value $< 2.2e{-}16$), with an absolute mean difference of 0.74 ± 1.05 (Table S1g). In all genomes, the breadth of coverage was higher than 99.2% over the two *Brucella* chromosomes (Table S1g). Using this benchmarked data set, we performed *de novo* assembly using three assembly software programs, SPAdes (42) (correction mode), SKESA (57) (default settings), and MEGAHIT (58) (default settings). The final data set comprised 1,200 genomes (40 genomes $\times$ 10 various read depths $\times$ 3 assembly software programs). The results show that the contiguity statistics of all assemblers did not show significant improvement at read sequence depth of more than 40-fold (Table S1h). Above this depth, the highest N_{50} was reported with SPAdes (mean, 224,714 bp) compared to MEGAHIT (mean, 158,700 bp) and SKESA (mean, 155,671 bp) (Table S1h). At low read depth (10- and 20-fold), SPAdes and MEGAHIT generally showed higher contiguity values (Table S1h). At read depth above 30, the three assemblers produced genomes with overall contig size covering >98% of the reference genome. Despite the good contiguity metrics of all assemblers, three genomes (ERR471311, ERR471313 [*B. ceti*], and SRR11449060 [*B. melitensis*]) produced by SPAdes and MEGAHIT were found to deviate from the reference genome size by more than 20% (i.e., the total genome size was 4.1 to 7 Mbp compared to 3.2 Mbp of the reference genome), and GC content deviated from the reference genome by more than 10% (i.e., GC content was 44 to 50% compared to 57.2% of the reference genome) (Table S1h).

For the 1,200 genomes generated by the three assemblers, we determined the number of cgMLST loci found and allele called. We clustered the genomes with an allele calling rate above 90% into cluster types at fewer than three differing alleles. Cluster analysis was consistent for SPAdes and SKESA genomes at a coverage depth of 40 or more, i.e., the downsampled genomes from the same sample were grouped into the respective cluster type (CT) (Fig. 5). However, for MEGAHIT, a higher sequencing depth of 90-fold was required to avoid inconsistencies in cluster definition of three (out of 40) genomes (Fig. 5). Above these depth thresholds, cluster definition (3 differing alleles) was highly reproducible with complete agreement between the three assemblers (adjusted Rand coefficients and adjusted Wallace from two comparison directions = 1) (Fig. 5). However, reproducible assignment of the cgST profiles (zero differing alleles) across different assemblers showed incomplete agreement even at higher depth values. SKESA showed the best overall reproducibility (reproducible cgST for 38 out of the 40 genomes at read depth of >40-fold), followed by SPAdes (25 out of 40 genomes; read depth, >40) and MEGAHIT (25 out of 40 genomes; read depth, >40) (Table S1h). Additionally, the cluster mismatch was minimal in SKESA assemblies compared to SPAdes and MEGAHIT (Fig. 5). On the other hand, SKESA genomes had the highest number of unidentified alleles at low read depths (10- and 20-fold) due to the high fragmentation of the genome (Fig. 5).

Scheme incorporation in PubMLST *Brucella* database. The cgMLST scheme was made accessible via the PubMLST *Brucella* species database (https://pubmlst.org/organisms/brucella-spp) (66). An allele nomenclature database for the proposed scheme was also integrated. The scheme is also accessible via the PubMLST RESTful application programming interface (https://rest.pubmlst.org) (67). Classification of the cgMLST allelic profiles is performed using the single-linkage method with cutoff thresholds defined hierarchically at 200, 100, 50, 25, 10, 5, and 3 different alleles in order to reveal different levels of relationships between the strains.

DISCUSSION

Brucellosis is a highly contagious bacterial zoonosis that affects numerous animal species and can cause significant economic losses in livestock, as well as severe illness and death in humans (16, 68). The classical (core) species of the genus *Brucella* are genetically monomorphic with limited recombination and horizontal gene transfers,

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

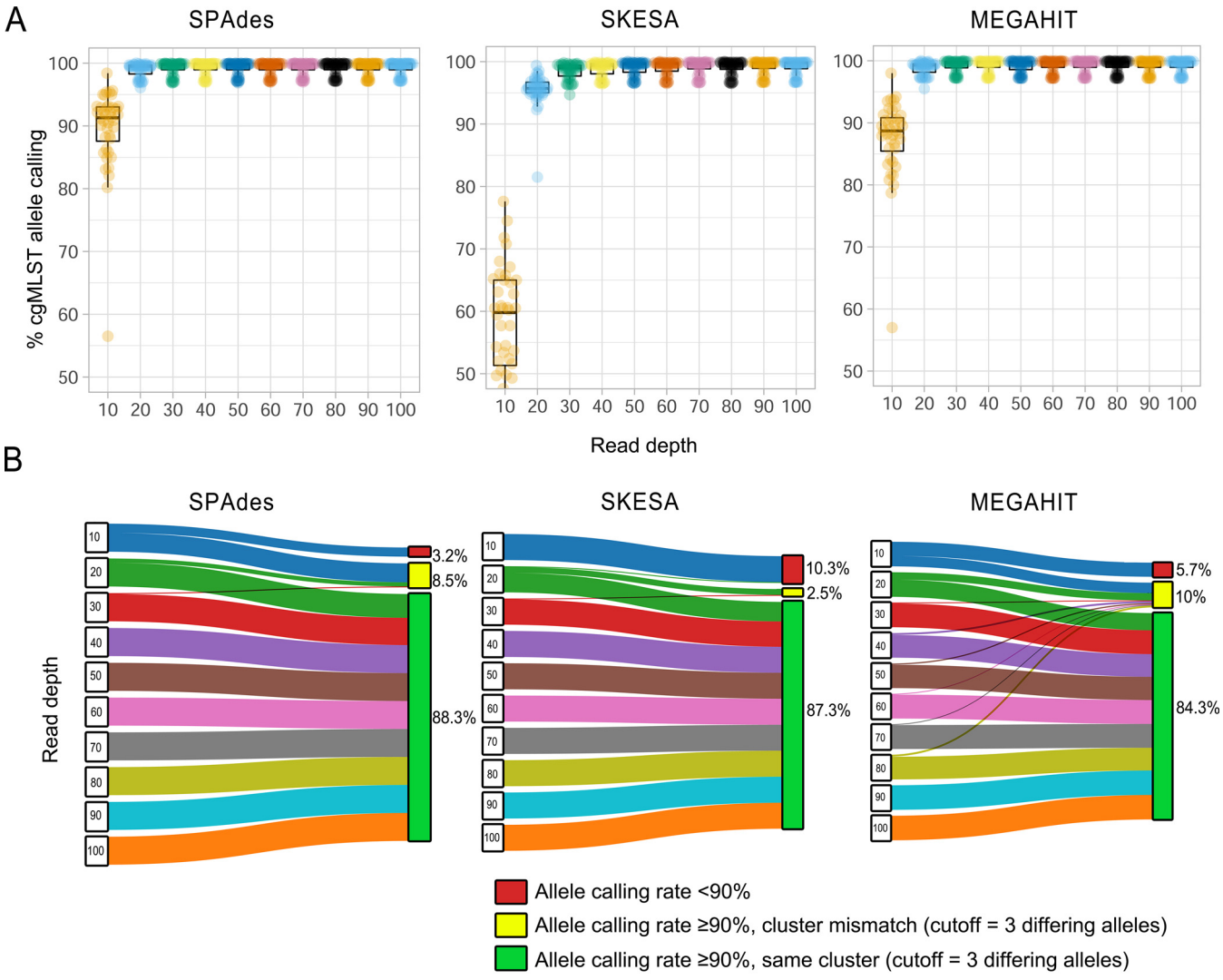

FIG 5 Effect of various sequencing depths and assembly software on allele calling rate and cluster analysis with *Brucella* cgMLST. (A) Box plots of the mean percentage of allele calling rate according to assembler for data generated with different coverage depths (n = 40 genomes per group). (B) Alluvial plots showing the frequency of the effect of read depth on clustering results of cgMLST data.

which implies the importance of whole-genome sequencing-based approaches for accurate *Brucella* strain typing. WGS enables high-resolution genotyping of bacterial isolates at the deepest taxonomic level as well as characterization of bacterial pathogens (69). WGS has been successfully used in outbreaks to illustrate transmission routes and pathogen spread and to assess the epidemiological linkage between isolates (69). Nevertheless, careful harmonization of sequence data generation and interpretation is essential to improve reproducibility and transferability of results across different sectors and disciplines. The cgMLST is a WGS genotyping system based on a predefined set of core genes that, on the one hand, provides superior resolution compared to classical methods. On the other hand, cgMLST has made it possible to establish a harmonized approach for bacterial genotyping in different clinical laboratories (33). In *Brucella*, genomic approaches based on cgMLST analysis have been successfully established only for *B. melitensis* (35, 36), which is the most commonly isolated *Brucella* species from humans. However, other *Brucella* species such as *B. abortus* and *B. suis* are frequently responsible for infections in humans and are associated with substantial economic losses in livestock production (70). In this study, we established a unified cgMLST scheme to characterize all pathogenic *Brucella* species in humans and animals. This scheme has been systematically developed to capture the diversity of all classical *Brucella* species,

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

B. melitensis, *B. abortus*, *B. suis*, *B. canis*, *B. ovis*, *B. neotomae*, *B. ceti*, and *B. pinnipedialis*, as well as the nonclassical species *B. microti* and *B. inopinata* and the recently described *Brucella* strains from amphibians (4). This was initiated by performing a primary phylogenetic analysis of all available RefSeq *Brucella* genomes. The phylogenetic analysis was concordant with prior studies (1, 71), indicating that the different *Brucella* species form a distinct phylogenetic lineage and meet the criteria for classification into a single genomic species (Fig. 1). Similar observations on close relatedness between all classical *Brucella* species averaging core genome nucleotide identities of >99% and identical 16S rRNA and *recA* gene sequences were reported in earlier studies (72, 73).

The quality of sequenced genomes is a major concern for WGS-based methods, notably cgMLST. In order to minimize the effect of accidental incorporation of erroneous genomes on the scheme definition, we applied stringent quality filters on all genomes. We additionally followed the definition of soft-core genomes to select core targets. Therefore, we selected the genes that were present in more than 98% of the 612 *Brucella* genomes in the RefSeq database. A similar approach was used, for example, in *Burkholderia pseudomallei* (63) and *Campylobacter jejuni*/*Campylobacter coli* schemes (74), where a 95% threshold was used to select core genes. This approach has the advantage of allowing screening of a large number of initial genomes, including draft incomplete genome sequences. As a result, the final *Brucella*-wide scheme comprised 1,764 targets, which was consistent with previous estimations of the core genome (75) but included significantly more loci than the previously described 164-gene-based *Brucella* scheme (34).

The reproducibility of allele calling was found to depend on the quality of the assembled genomes, for which the sequencing depth was considered relevant for obtaining high-quality genomes (76, 77). For paired-end Illumina data, our results show that the sequencing depth affects the accuracy and rate of allele calling of cgMLST and can also lead to cluster mismatches that affect next-generation sequencing (NGS) diagnosis. The required depth threshold may also vary depending on the assembler used. For *Brucella* cgMLST, we recommend a depth of coverage of 40 or more for raw paired-end Illumina data assembled by SKESA or SPAdes (along with prior quality assessment of raw data for strain contamination and an average Q30 of no less than 80%). It is worth noting that the required quality criteria and depth of sequencing may be different for other sequencing platforms, e.g., Pacific Biosciences and Oxford Nanopore Technologies (78). Additionally, we have shown that the reproducibility of cluster types across different assemblers is rather more possible than that of cgST profiles. Therefore, definition of cluster nomenclature when interpreting cgMLST data using predefined or hierarchical cluster thresholds should allow comparability of samples produced using distinct bioinformatics approaches while also revealing epidemiological links between isolates.

Genome-based analyses enable unambiguous taxonomic resolution for species identification and strain typing (69). Considering the long incubation period of brucellosis in humans and animals, the outbreak and source are often not identified until later confirmation of the diagnosis. Again, tracing with high-resolution genotypic methods is practically more valuable, especially in areas of endemicity where frequent exposure and repeated outbreaks, animal movements, and livestock trade are also common. In areas of endemicity where mixed agricultural practices are often adopted, infections with multiple *Brucella* species and genotypes may also be possible. Hence, the availability of a single scheme that covers multiple species of *Brucella* and has a resolution comparable to that of a single-species scheme is valuable to reduce the need to establish multiple separate cgMLST schemes for each species typed and also to improve accurate species diagnosis without the requirement of extensive phenotypic characterization.

The applicability of genomic analysis proved valuable for outbreak investigations in strains with documented and with missing epidemiological data (79), e.g., to identify the source of infection (80). In this study, we have also shown that the novel cgMLST had high discriminatory power and hence can resolve the closely related strains in

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

outbreaks, making the new scheme suitable for epidemiological surveillance of pathogenic *Brucella*. The cgMLST was clearly superior to the classical MLST in terms of resolution. The results of cgMLST were comparable to those of core genome-based SNP typing in outbreaks. This was observed based on 37 *B. melitensis* and 76 *B. abortus* outbreak strains (45), leading to similar conclusions as in the original studies.

Whole-genome-based bacterial genotyping methods involving the inference of single point mutations (SNP typing) or the detection of gene variants in the form of numbered allele profiles (cgMLST allele typing) are considered compatible methods with different advantages. The SNP typing methods allow robust phylogenetic reconstruction of SNP alignment data as they are based on base-by-base comparison and consider evolutionary models with masked recombination sites and mobile elements. The allele typing methods, on the other hand, are more likely to produce concordant and comparable results between laboratories since the typing data are easily portable via online nomenclature databases, and the underlying workflow is less computationally intensive and does not require complex bioinformatics skills. There is also the possibility to maximize the resolution of allele typing methods by including the core and accessory genes (whole-genome MLST). However, the incorporation of accessory genes faces several hurdles related to the versatile presence of accessory genes and accuracy in distinguishing core from accessory genes and in resolving orthologous and paralogous genes (81).

In conclusion, we have developed a new cgMLST scheme for *Brucella* strains. The new scheme can be a valuable tool for the study of brucellosis outbreaks and should represent a valuable genomic resource, especially for countries with endemic brucellosis. In turn, better tracing of strains and their genotyping from animal and human hosts will contribute to a better surveillance system that will support the implementation of a better control program.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 2.1 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.5 MB.

ACKNOWLEDGMENTS

D. Harmsen is a cofounder of Ridom GmbH (Münster, Germany). The other authors declare that there are no conflicts of interest.

REFERENCES

1. Suárez-Esquivel M, Chaves-Olarte E, Moreno E, Guzmán-Verri C. 2020. *Brucella* genomics: macro and micro evolution. Int J Mol Sci 21:7749. https://doi.org/10.3390/ijms21207749.
2. Whatmore AM, Foster JT. 2021. Emerging diversity and ongoing expansion of the genus *Brucella*. Infect Genet Evol 92:104865. https://doi.org/10.1016/j.meegid.2021.104865.
3. de Figueiredo P, Ficht TA, Rice-Ficht A, Rossetti CA, Adams LG. 2015. Pathogenesis and immunobiology of brucellosis: review of *Brucella*-host interactions. Am J Pathol 185:1505–1517. https://doi.org/10.1016/j.ajpath.2015.03.003.
4. Al Dahouk S, Köhler S, Occhialini A, Jiménez de Bagüés MP, Hammerl JA, Eisenberg T, Vergnaud G, Cloeckaert A, Zygmunt MS, Whatmore AM, Melzer F, Drees KP, Foster JT, Wattam AR, Scholz HC. 2017. *Brucella* spp. of amphibians comprise genomically diverse motile strains competent for replication in macrophages and survival in mammalian hosts. Sci Rep 7:44420. https://doi.org/10.1038/srep44420.
5. Pappas G, Panagopoulou P, Christou L, Akritidis N. 2006. *Brucella* as a biological weapon. Cell Mol Life Sci 63:2229–2236. https://doi.org/10.1007/s00018-006-6311-4.
6. Valderas MW, Roop RM. 2006. *Brucella* and bioterrorism, p 139–153. *In* Anderson B, Friedman H, Bendinelli M (ed), Microorganisms and bioterrorism. Springer US, Boston, MA.
7. Corbel MJ, Brinley Morgan WJ. 1975. Proposal for minimal standards for descriptions of new species and biotypes of the genus *Brucella*. Int J Syst Evol Microbiol 25:83–89. https://doi.org/10.1099/00207713-25-1-83.
8. Foster G, Osterman BS, Godfroid J, Jacques I, Cloeckaert A. 2007. *Brucella ceti* sp. nov. and *Brucella pinnipedialis* sp. nov. for *Brucella* strains with cetaceans and seals as their preferred hosts. Int J Syst Evol Microbiol 57:2688–2693. https://doi.org/10.1099/ijs.0.65269-0.
9. Scholz HC, Hubalek Z, Sedlácek I, Vergnaud G, Tomaso H, Al Dahouk S, Melzer F, Kämpfer P, Neubauer H, Cloeckaert A, Maquart M, Zygmunt MS, Whatmore AM, Falsen E, Bahn P, Göllner C, Pfeffer M, Huber B, Busse HJ, Nöckler K. 2008. *Brucella microti* sp. nov., isolated from the common vole Microtus arvalis. Int J Syst Evol Microbiol 58:375–382. https://doi.org/10.1099/ijs.0.65356-0.
10. Scholz HC, Nöckler K, Göllner C, Bahn P, Vergnaud G, Tomaso H, Al Dahouk S, Kämpfer P, Cloeckaert A, Maquart M, Zygmunt MS, Whatmore AM, Pfeffer M, Huber B, Busse HJ, De BK. 2010. *Brucella inopinata* sp. nov., isolated from a breast implant infection. Int J Syst Evol Microbiol 60:801–808. https://doi.org/10.1099/ijs.0.011148-0.
11. Whatmore AM, Davison N, Cloeckaert A, Al Dahouk S, Zygmunt MS, Brew SD, Perrett LL, Koylass MS, Vergnaud G, Quance C, Scholz HC, Dick EJ, Hubbard G, Schlabritz-Loutsevitch NE. 2014. *Brucella papionis* sp. nov., isolated from baboons (*Papio* spp.). Int J Syst Evol Microbiol 64:4120–4128. https://doi.org/10.1099/ijs.0.065482-0.

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

12. Scholz HC, Revilla-Fernández S, Dahouk SA, Hammerl JA, Zygmunt MS, Cloeckaert A, Koylass M, Whatmore AM, Blom J, Vergnaud G, Witte A, Aistleitner K, Hofer E. 2016. *Brucella vulpis* sp. nov., isolated from mandibular lymph nodes of red foxes (*Vulpes vulpes*). Int J Syst Evol Microbiol 66:2090–2098. https://doi.org/10.1099/ijsem.0.000998.
13. De Massis F, Zilli K, Di Donato G, Nuvoloni R, Pelini S, Sacchini L, D'Alterio N, Di Giannatale E. 2019. Distribution of *Brucella* field strains isolated from livestock, wildlife populations, and humans in Italy from 2007 to 2015. PLoS One 14:e0213689. https://doi.org/10.1371/journal.pone.0213689.
14. McDonald WL, Jamaludin R, Mackereth G, Hansen M, Humphrey S, Short P, Taylor T, Swingler J, Dawson CE, Whatmore AM, Stubberfield E, Perrett LL, Simmons G. 2006. Characterization of a *Brucella* sp. strain as a marine-mammal type despite isolation from a patient with spinal osteomyelitis in New Zealand. J Clin Microbiol 44:4363–4370. https://doi.org/10.1128/JCM.00680-06.
15. Nomura A, Imaoka K, Imanishi H, Shimizu H, Nagura F, Maeda K, Tomino T, Fujita Y, Kimura M, Stein G. 2010. Human *Brucella canis* infections diagnosed by blood culture. Emerg Infect Dis 16:1183–1185. https://doi.org/10.3201/eid1607.090209.
16. Gwida M, Al Dahouk S, Melzer F, Rösler U, Neubauer H, Tomaso H. 2010. Brucellosis - regionally emerging zoonotic disease? Croat Med J 51:289–295. https://doi.org/10.3325/cmj.2010.51.289.
17. Olsen SC, Palmer MV. 2014. Advancement of knowledge of *Brucella* over the past 50 years. Vet Pathol 51:1076–1089. https://doi.org/10.1177/0300985814540545.
18. Franc KA, Krecek RC, Häsler BN, Arenas-Gamboa AM. 2018. Brucellosis remains a neglected disease in the developing world: a call for interdisciplinary action. BMC Public Health 18:125. https://doi.org/10.1186/s12889-017-5016-y.
19. Morgan WJ, Corbel MJ. 1976. Recommendations for the description of species and biotypes of the genus *Brucella*. Dev Biol Stand 31:27–37.
20. Marianelli C, Ciuchini F, Tarantino M, Pasquali P, Adone R. 2006. Molecular characterization of the *rpoB* gene in *Brucella* species: new potential molecular markers for genotyping. Microbes Infect 8:860–865. https://doi.org/10.1016/j.micinf.2005.10.008.
21. Al Dahouk S, Flèche PL, Nöckler K, Jacques I, Grayon M, Scholz HC, Tomaso H, Vergnaud G, Neubauer H. 2007. Evaluation of *Brucella* MLVA typing for human brucellosis. J Microbiol Methods 69:137–145. https://doi.org/10.1016/j.mimet.2006.12.015.
22. Le Flèche P, Jacques I, Grayon M, Al Dahouk S, Bouchon P, Denoeud F, Nöckler K, Neubauer H, Guilloteau LA, Vergnaud G. 2006. Evaluation and selection of tandem repeat loci for a *Brucella* MLVA typing assay. BMC Microbiol 6:9. https://doi.org/10.1186/1471-2180-6-9.
23. Valdezate S, Cervera I, Hernandez P, Navarro A, Saéz Nieto JA. 2007. Characterisation of human outbreaks of brucellosis and sporadic cases by the use of hyper-variable octameric oligonucleotide fingerprint (HOOF) variable number tandem repeats. Clin Microbiol Infect 13:887–892. https://doi.org/10.1111/j.1469-0691.2007.01768.x.
24. Valdezate S, Navarro A, Villalón P, Carrasco G, Saéz-Nieto JA. 2010. Epidemiological and phylogenetic analysis of Spanish human *Brucella melitensis* strains by multiple-locus variable-number tandem-repeat typing, hypervariable octameric oligonucleotide fingerprinting, and *rpoB* typing. J Clin Microbiol 48:2734–2740. https://doi.org/10.1128/JCM.00533-10.
25. Maquart M, Le Flèche P, Foster G, Tryland M, Ramisse F, Djønne B, Al Dahouk S, Jacques I, Neubauer H, Walravens K, Godfroid J, Cloeckaert A, Vergnaud G. 2009. MLVA-16 typing of 295 marine mammal *Brucella* isolates from different animal and geographic origins identifies 7 major groups within *Brucella ceti* and *Brucella pinnipedialis*. BMC Microbiol 9:145. https://doi.org/10.1186/1471-2180-9-145.
26. Vergnaud G, Hauck Y, Christiany D, Daoud B, Pourcel C, Jacques I, Cloeckaert A, Zygmunt MS. 2018. Genotypic expansion within the population structure of classical *Brucella* species revealed by MLVA16 typing of 1404 *Brucella* isolates from different animal and geographic origins, 1974–2006. Front Microbiol 9:1545. https://doi.org/10.3389/fmicb.2018.01545.
27. Whatmore AM, Perrett LL, MacMillan AP. 2007. Characterisation of the genetic diversity of *Brucella* by multilocus sequencing. BMC Microbiol 7:34. https://doi.org/10.1186/1471-2180-7-34.
28. Whatmore AM, Koylass MS, Muchowski J, Edwards-Smallbone J, Gopaul KK, Perrett LL. 2016. Extended multilocus sequence analysis to describe the global population structure of the genus *Brucella*: phylogeography and relationship to biovars. Front Microbiol 7:2049. https://doi.org/10.3389/fmicb.2016.02049.
29. Liu Z-G, Wang M, Zhao H-Y, Piao D-R, Jiang H, Li Z-J. 2019. Investigation of the molecular characteristics of *Brucella* isolates from Guangxi province, China. BMC Microbiol 19:292. https://doi.org/10.1186/s12866-019-1665-6.
30. Janowicz A, De Massis F, Zilli K, Ancora M, Tittarelli M, Sacchini F, Di Giannatale E, Sahl JW, Foster JT, Garofolo G. 2020. Evolutionary history and current distribution of the West Mediterranean lineage of Brucella melitensis in Italy. Microb Genom 6:mgen000446. https://doi.org/10.1099/mgen.0.000446.
31. Sacchini L, Wahab T, Di Giannatale E, Zilli K, Abass A, Garofolo G, Janowicz A. 2019. Whole genome sequencing for tracing geographical origin of imported cases of human brucellosis in Sweden. Microorganisms 7:398. https://doi.org/10.3390/microorganisms7100398.
32. Shevtsova E, Vergnaud G, Shevtsov A, Shustov A, Berdimuratova K, Mukanov K, Syzdykov M, Kuznetsov A, Lukhnova L, Izbanova U, Filipenko M, Ramankulov Y. 2019. Genetic diversity of *Brucella melitensis* in Kazakhstan in relation to world-wide diversity. Front Microbiol 10:1897. https://doi.org/10.3389/fmicb.2019.01897.
33. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11:728–736. https://doi.org/10.1038/nrmicro3093.
34. Sankarasubramanian J, Vishnu US, Gunasekaran P, Rajendhran J. 2019. Development and evaluation of a core genome multilocus sequence typing (cgMLST) scheme for *Brucella* spp. Infect Genet Evol 67:38–43. https://doi.org/10.1016/j.meegid.2018.10.021.
35. Janowicz A, De Massis F, Ancora M, Cammà C, Patavino C, Battisti A, Prior K, Harmsen D, Scholz H, Zilli K, Sacchini L, Di Giannatale E, Garofolo G. 2018. Core genome Multilocus sequence typing and single nucleotide polymorphism analysis in the epidemiology of *Brucella melitensis* infections. J Clin Microbiol 56:e00517-18. https://doi.org/10.1128/JCM.00517-18.
36. Pelerito A, Nunes A, Núncio MS, Gomes JP. 2020. Genome–scale approach to study the genetic relatedness among *Brucella melitensis* strains. PLoS One 15:e0229863. https://doi.org/10.1371/journal.pone.0229863.
37. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.
38. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9.
39. Segata N, Börnigen D, Morgan XC, Huttenhower C. 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun 4:2304. https://doi.org/10.1038/ncomms3304.
40. 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 Biol 17:132. https://doi.org/10.1186/s13059-016-0997-x.
41. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.
42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
43. Seemann T. 2021. Shovill: assemble bacterial isolate genomes from Illumina paired-end reads. https://github.com/tseemann/shovill. Accessed 22 June 2021.
44. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
45. Allen AR, Milne G, Drees K, Presho E, Graham J, McAdam P, Jones K, Wright L, Skuce R, Whatmore AM, Graham J, Foster JT. 2020. Genomic epizootiology of a *Brucella abortus* outbreak in Northern Ireland (1997–2012). Infect Genet Evol 81:104235. https://doi.org/10.1016/j.meegid.2020.104235.
46. Sahl JW, Lemmer D, Travis J, Schupp JM, Gillece JD, Aziz M, Driebe EM, Drees KP, Hicks ND, Williamson CHD, Hepp CM, Smith DE, Roe C, Engelthaler DM, Wagner DM, Keim P. 2016. NASP: an accurate, rapid method for the identification of SNPs in WGS datasets that supports flexible input and output formats. Microb Genom 2:e000074. https://doi.org/10.1099/mgen.0.000074.

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.

47. Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics Chapter 10:Unit 10.3. https://doi.org/10.1002/0471250953.bi1003s00.
48. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
49. Junemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A, Goesmann A, von Haeseler A, Stoye J, Harmsen D. 2013. Updating benchtop sequencing performance comparison. Nat Biotechnol 31:294–296. https://doi.org/10.1038/nbt.2522.
50. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389.
51. Carrico JA, Silva-Costa C, Melo-Cristino J, Pinto FR, de Lencastre H, Almeida JS, Ramirez M. 2006. Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant *Streptococcus pyogenes*. J Clin Microbiol 44:2524–2532. https://doi.org/10.1128/JCM.02536-05.
52. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11:e0163962. https://doi.org/10.1371/journal.pone.0163962.
53. Low AJ, Koziol AG, Manninger PA, Blais B, Carrillo CD. 2019. ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. PeerJ 7:e6995. https://doi.org/10.7717/peerj.6995.
54. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MCJ. 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology (Reading) 158:1005–1015. https://doi.org/10.1099/mic.0.055459-0.
55. Hall MB. 2022. Rasusa: randomly subsample sequencing reads to a specified coverage. J Open Source Softw 7:3941. https://doi.org/10.21105/joss.03941.
56. Bushnell B. 2014. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab (LBNL), Berkeley, CA.
57. Souvorov A, Agarwala R, Lipman DJ. 2018. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol 19:153. https://doi.org/10.1186/s13059-018-1540-z.
58. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676. https://doi.org/10.1093/bioinformatics/btv033.
59. Hördt A, López MG, Meier-Kolthoff JP, Schleuning M, Weinhold L-M, Tindall BJ, Gronow S, Kyrpides NC, Woyke T, Göker M. 2020. Analysis of 1,000+ type-strain genomes substantially improves taxonomic classification of *Alphaproteobacteria*. Front Microbiol 11:468. https://doi.org/10.3389/fmicb.2020.00468.
60. Moreno E. 2020. The one hundred year journey of the genus *Brucella* (Meyer and Shaw 1920). FEMS Microbiol Rev 45:fuaa045. https://doi.org/10.1093/femsre/fuaa045.
61. Wattam AR, Foster JT, Mane SP, Beckstrom-Sternberg SM, Beckstrom-Sternberg JM, Dickerman AW, Keim P, Pearson T, Shukla M, Ward DV, Williams KP, Sobral BW, Tsolis RM, Whatmore AM, O'Callaghan D. 2014. Comparative phylogenomics and evolution of the *Brucellae* reveal a path to virulence. J Bacteriol 196:920–930. https://doi.org/10.1128/JB.01091-13.
62. Guzmán-Verri C, González-Barrientos R, Hernández-Mora G, Morales J-A, Baquero-Calvo E, Chaves-Olarte E, Moreno E. 2012. *Brucella ceti* and brucellosis in cetaceans. Front Cell Infect Microbiol 2:3. https://doi.org/10.3389/fcimb.2012.00003.
63. Lichtenegger S, Trinh TT, Assig K, Prior K, Harmsen D, Pesl J, Zauner A, Lipp M, Que TA, Mutsam B, Kleinhappl B, Steinmetz I, Wagner GE. 2021. Development and validation of a *Burkholderia pseudomallei* core genome multilocus sequence typing scheme to facilitate molecular surveillance. J Clin Microbiol 59:e0009321. https://doi.org/10.1128/JCM.00093-21.
64. Allen A, Breadon E, Byrne A, Mallon T, Skuce R, Groussaud P, Dainty A, Graham J, Jones K, Pollock L, Whatmore A. 2015. Molecular epidemiology of *Brucella abortus* in Northern Ireland-1991 to 2012. PLoS One 10:e0136721. https://doi.org/10.1371/journal.pone.0136721.
65. Johansen TB, Scheffer L, Jensen VK, Bohlin J, Feruglio SL. 2018. Whole-genome sequencing and antimicrobial resistance in *Brucella melitensis* from a Norwegian perspective. Sci Rep 8:8538. https://doi.org/10.1038/s41598-018-26906-3.
66. Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res 3:124. https://doi.org/10.12688/wellcomeopenres.14826.1.
67. Jolley KA, Bray JE, Maiden MCJ. 2017. A RESTful application programming interface for the PubMLST molecular typing and genome databases. Database (Oxford) 2017:bax060. https://doi.org/10.1093/database/bax060.
68. Moreno E. 2014. Retrospective and prospective perspectives on zoonotic brucellosis. Front Microbiol 5:213. https://doi.org/10.3389/fmicb.2014.00213.
69. Gilchrist CA, Turner SD, Riley MF, Petri WA, Jr, Hewlett EL. 2015. Whole-genome sequencing in outbreak analysis. Clin Microbiol Rev 28:541–563. https://doi.org/10.1128/CMR.00075-13.
70. Scholz HC, Vergnaud G. 2013. Molecular characterisation of *Brucella* species. Rev Sci Tech 32:149–162. https://doi.org/10.20506/rst.32.1.2189.
71. Leclercq SO, Cloeckaert A, Zygmunt MS. 2020. Taxonomic organization of the family *Brucellaceae* based on a phylogenomic approach. Front Microbiol 10:3083. https://doi.org/10.3389/fmicb.2019.03083.
72. Scholz HC, Mühldorfer K, Shilton C, Benedict S, Whatmore AM, Blom J, Eisenberg T. 2016. The change of a medically important genus: worldwide occurrence of genetically diverse novel *Brucella* species in exotic frogs. PLoS One 11:e0168872. https://doi.org/10.1371/journal.pone.0168872.
73. Gee JE, De BK, Levett PN, Whitney AM, Novak RT, Popovic T. 2004. Use of 16S rRNA gene sequencing for rapid confirmatory identification of *Brucella* isolates. J Clin Microbiol 42:3649–3654. https://doi.org/10.1128/JCM.42.8.3649-3654.2004.
74. Cody AJ, Bray JE, Jolley KA, McCarthy ND, Maiden MCJ. 2017. Core genome multilocus sequence typing scheme for stable, comparative analyses of *Campylobacter jejuni* and *C. coli* human disease isolates. J Clin Microbiol 55:2086–2097. https://doi.org/10.1128/JCM.00080-17.
75. Yang X, Li Y, Zang J, Li Y, Bie P, Lu Y, Wu Q. 2016. Analysis of pan-genome to identify the core genes and essential genes of *Brucella* spp. Mol Genet Genomics 291:905–912. https://doi.org/10.1007/s00438-015-1154-z.
76. Liu Y-Y, Chen B-H, Chen C-C, Chiou C-S. 2021. Assessment of metrics in next-generation sequencing experiments for use in core-genome multilocus sequence type. PeerJ 9:e11842. https://doi.org/10.7717/peerj.11842.
77. Palma F, Mangone I, Janowicz A, Moura A, Chiaverini A, Torresi M, Garofolo G, Criscuolo A, Brisse S, Di Pasquale A, Cammà C, Radomski N. 2022. In vitro and in silico parameters for precise cgMLST typing of Listeria monocytogenes. BMC Genomics 23:235. https://doi.org/10.1186/s12864-022-08437-4.
78. Fu S, Wang A, Au KF. 2019. A comparative evaluation of hybrid error correction methods for error-prone long reads. Genome Biol 20:26. https://doi.org/10.1186/s13059-018-1605-z.
79. Schaeffer J, Revilla-Fernández S, Hofer E, Posch R, Stoeger A, Leth C, Schmoll F, Djordjevic V, Lakicevic B, Matovic K, Hufnagl P, Indra A, Allerberger F, Ruppitsch W. 2021. Tracking the origin of Austrian human brucellosis cases using whole genome sequencing. Front Med 8:635547. https://doi.org/10.3389/fmed.2021.635547.
80. Bardenstein S, Gibbs RE, Yagel Y, Motro Y, Moran-Gilad J. 2021. Brucellosis outbreak traced to commercially sold camel milk through whole-genome sequencing, Israel. Emerg Infect Dis 27:1728–1731. https://doi.org/10.3201/eid2706.204902.
81. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. 2019. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res 29:304–316. https://doi.org/10.1101/gr.241455.118.

Downloaded from https://journals.asm.org/journal/jcm on 18 August 2022 by 193.22.115.2.