BrucellaID
Brucella species identification and MLST from reads

Quality assurance report

4 sample(s): 4 ok, 0 with a warning, 0 failed. BrucellaID 2.1.1, run on 2026-11-08T14:55:06-05:00.

This is an example report: a real run of the four public read sets of the BrucellaID tutorial, produced by scripts/make_example_report.py, which leaves out the computer it ran on. The user account, the host name, the operating system and the working directory read [removed], and the input and reference files are named without the folder they were in. Everything else is as BrucellaID wrote it. A report of your own names them all, which is what makes it worth signing.

1 Summary

Run started	2026-11-08T14:55:06-05:00
Run finished	2026-11-08T14:55:43-05:00
Operator	M-O Duceppe
Command	python3 scripts/make_example_report.py reads

SAMPLE	SPECIES	FROM	DECISION	ST	ALLELE PROFILE	CLOSEST REFERENCE	STATUS
Brucella abortus · ST 1 (1 sample(s))							
ERR3454415	Brucella abortus	MASH+ST	agree	1	2-1-1-2-1-3-1-1-1	GCA_000742195.1_ASM74219v1_genomic.fna.gz	ok
Brucella canis · ST 20 (1 sample(s))							
ERR2136546	Brucella canis	MASH+ST	arbitrated	20	1-6-4-1-5-3-5-4-6	GCA_000530495.1_Oliveri_genomic.fna.gz	ok
Brucella melitensis · ST 11 (1 sample(s))							
SRR19520329	Brucella melitensis	MASH+ST	agree	11	3-2-3-2-1-5-3-2-10	GCA_061325495.1_ASM6132549v1_genomic.fna.gz	ok
Brucella suis · ST 15 (1 sample(s))							
SRR8550484	Brucella suis	MASH+ST	agree	15	1-2-7-1-3-3-5-3-2	GCA_000742055.1_ASM74205v1_genomic.fna.gz	ok

REVIEWED BY

Name Date Signature

2 Notes, warnings and errors

SAMPLE	STATUS	WHAT IT SAYS
ERR2136546	ok	nearly as close: GCA_000741995.1_ASM74199v1_genomic.fna.gz (Brucella suis, 986/1000 hashes), 3 other Brucella canis genome(s)
ERR2136546	ok	MASH cannot separate Brucella canis and Brucella suis: ST 20 is Brucella canis in 70 of 70 isolate(s), which agrees with the closest reference
ERR3454415	ok	nearly as close: 9 other Brucella abortus genome(s)
SRR19520329	ok	nearly as close: 6 other Brucella melitensis genome(s)
SRR8550484	ok	nearly as close: 2 other Brucella suis genome(s)

3 Samples

ERR3454415 ok Brucella abortus ST 1

IDENTIFICATION: THE MASH DISTANCE

Closest reference	GCA_000742195.1_ASM74219v1_genomic.fna.gz
What it is	Brucella abortus str. 2308
NCBI accessions	GCA_000742195.1
MASH distance	0.000144
Estimated identity	99.986%
Shared min-hashes	994/1000 (at least 150 required)
K-mer copies required	14 (from the estimated 70x coverage)

THE SPECIES CALL

Species reported	Brucella abortus
Decided by	MASH+ST
Species within reach	Brucella abortus
Sequence type	1
What that ST is	ST 1 is Brucella abortus in 463 of 466 isolate(s) (the rest: Brucella melitensis:3)

agree The sequence type names the same species as the closest reference genome, which named it on its own: the two kinds of evidence agree.

Closest first. A shaded row is too close to the first to be separated by distance alone. The other 422 reference(s) are further away; brucellaid_mash.tsv has every distance.

REFERENCE GENOME	SPECIES	STRAIN	DISTANCE	IDENTITY %	SHARED HASHES
GCA_000742195.1_ASM74219v1_genomic.fna.gz	Brucella abortus	2308	0.000144	99.986	994/1000
GCA_050497595.1_ASM5049759v1_genomic.fna.gz	Brucella abortus	Ba44	0.000337	99.966	986/1000
GCA_000739315.1_ASM73931v1_genomic.fna.gz	Brucella abortus	544	0.000361	99.964	985/1000
GCA_000370045.1_Bruc_abor_63_168_V1_genomic.fna.gz	Brucella abortus	63/168	0.000386	99.961	984/1000
GCA_000472245.1_Denovo_assembly_genomic.fna.gz	Brucella abortus	S99	0.000533	99.947	978/1000
GCA_039649615.1_ASM3964961v1_genomic.fna.gz	Brucella abortus	X20_S4	0.000533	99.947	978/1000
GCA_050498085.1_ASM5049808v1_genomic.fna.gz	Brucella abortus	Ba732	0.000533	99.947	978/1000
GCA_022426865.1_ASM2242686v1_genomic.fna.gz	Brucella abortus	BANDDB_5033	0.000582	99.942	976/1000
GCA_030761375.1_ASM3076137v1_genomic.fna.gz	Brucella abortus	LBAB008	0.000607	99.939	975/1000
GCA_022426825.1_ASM2242682v1_genomic.fna.gz	Brucella abortus	BANDDB_3995	0.000631	99.937	974/1000

MLST: THE SEQUENCE TYPE

Sequence type	1
Allele profile	2-1-1-2-1-3-1-1-1
Every locus called	yes
Reads in the input	1,175,082
Reads matching an allele	2,948

INPUT FILES

FILE	SIZE	MODIFIED
ERR3454415_1.fastq.gz	87.4 MB	2026-11-07 15:47:04 EST
ERR3454415_2.fastq.gz	83.5 MB	2026-11-07 15:48:02 EST

```
ERR3454415_1.fastq.gz
MD5 8eaddc2c8044657712ec586ba5482c70
ERR3454415_2.fastq.gz
MD5 b2706df339382a63f38bba34222b4053
```

The margin each call had over the next allele of its locus; a ratio near 1.00 is the one to check.

LOCUS	ALLELE	READS	RUNNER-UP (ITS READS)	AHEAD	RATIO	NOTE
gap	2	432	gap_6 (300)	132	1.44	
aroA	1	322	aroA_30 (210)	112	1.53	
glk	1	222	glk_21 (158)	64	1.41	
dnaK	2	340	dnaK_26 (222)	118	1.53	
gyrB	1	334	gyrB_22 (224)	110	1.49	
trpE	3	320	trpE_4 (242)	78	1.32	
cobQ	1	308	cobQ_12 (196)	112	1.57	
int_hyp	1	368	int_hyp_4 (256)	112	1.44	
omp25	1	292	omp25_12 (186)	106	1.57	

NOTES

- nearly as close: 9 other Brucella abortus genome(s)

ERR2136546 ok Brucella canis ST 20

IDENTIFICATION: THE MASH DISTANCE

Closest reference	GCA_000530495.1_Oliveri_genomic.fna.gz
What it is	Brucella canis str. Oliveri
NCBI accessions	GCA_000530495.1
MASH distance	0.000119
Estimated identity	99.988%
Shared min-hashes	995/1000 (at least 150 required)
K-mer copies required	9 (from the estimated 47x coverage)

THE SPECIES CALL

Species reported	Brucella canis
Decided by	MASH+ST
Species within reach	Brucella canis, Brucella suis
Sequence type	20
What that ST is	ST 20 is Brucella canis in 70 of 70 isolate(s)

arbitrated A reference of a second species was near enough to the closest one that the distance did not really choose. The sequence type chose between the species it left within reach.

Closest first. A shaded row is too close to the first to be separated by distance alone. The other 427 reference(s) are further away; brucellaid_mash.tsv has every distance.

REFERENCE GENOME	SPECIES	STRAIN	DISTANCE	IDENTITY %	SHARED HASHES
GCA_000530495.1_Oliveri_genomic.fna.gz	Brucella canis	Oliveri	0.000119	99.988	995/1000
GCA_000741995.1_ASM74199v1_genomic.fna.gz	Brucella suis	BSQ	0.000337	99.966	986/1000
GCA_056582615.1_ASM5658261v1_genomic.fna.gz	Brucella canis	M-24	0.000361	99.964	985/1000
GCA_000367265.1_Bruc_cani_CNGB_1172_V1_genomic.fna.gz	Brucella canis	CNGB 1172	0.000483	99.952	980/1000
GCA_050496145.1_ASM5049614v1_genomic.fna.gz	Brucella canis	Bc1066	0.000483	99.952	980/1000

MLST: THE SEQUENCE TYPE

Sequence type	20
Allele profile	1-6-4-1-5-3-5-4-6
Every locus called	yes
Reads in the input	625,576
Reads matching an allele	1,824

INPUT FILES

FILE	SIZE	MODIFIED
ERR2136546_1.fastq.gz	57.0 MB	2026-11-07 15:50:28 EST
ERR2136546_2.fastq.gz	63.9 MB	2026-11-07 15:51:19 EST

```
ERR2136546_1.fastq.gz
MD5 7573e407362c1458a464a4d92fec2db4
ERR2136546_2.fastq.gz
MD5 a83a9dceaee29371219b98f0a2f1dc39
```

The margin each call had over the next allele of its locus; a ratio near 1.00 is the one to check.

LOCUS	ALLELE	READS	RUNNER-UP (ITS READS)	AHEAD	RATIO	NOTE
gap	1	232	gap_26 (170)	62	1.36	
aroA	6	194	aroA_31 (80)	114	2.42	
glk	4	148	glk_2 (94)	54	1.57	
dnaK	1	244	dnaK_2 (162)	82	1.51	
gyrB	5	224	gyrB_4 (110)	114	2.04	
trpE	3	178	trpE_5 (114)	64	1.56	
cobQ	5	172	cobQ_30 (108)	64	1.59	
int_hyp	4	234	int_hyp_1 (180)	54	1.30	
omp25	6	172	omp25_2 (98)	74	1.76	

NOTES

- nearly as close: GCA_000741995.1_ASM74199v1_genomic.fna.gz (Brucella suis, 986/1000 hashes), 3 other Brucella canis genome(s)
- MASH cannot separate Brucella canis and Brucella suis: ST 20 is Brucella canis in 70 of 70 isolate(s), which agrees with the closest reference

SRR19520329 ok Brucella melitensis ST 11

IDENTIFICATION: THE MASH DISTANCE

Closest reference	GCA_061325495.1_ASM6132549v1_genomic.fna.gz
What it is	Brucella melitensis str. 48_S36
NCBI accessions	GCA_061325495.1
MASH distance	0.000119
Estimated identity	99.988%
Shared min-hashes	995/1000 (at least 150 required)
K-mer copies required	11 (from the estimated 53x coverage)

THE SPECIES CALL

Species reported	Brucella melitensis
Decided by	MASH+ST
Species within reach	Brucella melitensis
Sequence type	11
What that ST is	ST 11 is Brucella melitensis in 546 of 546 isolate(s)

agree The sequence type names the same species as the closest reference genome, which named it on its own: the two kinds of evidence agree.

Closest first. A shaded row is too close to the first to be separated by distance alone. The other 425 reference(s) are further away; brucellaid_mash.tsv has every distance.

REFERENCE GENOME	SPECIES	STRAIN	DISTANCE	IDENTITY %	SHARED HASHES
GCA_061325495.1_ASM6132549v1_genomic.fna.gz	Brucella melitensis	48_S36	0.000119	99.988	995/1000
GCA_002191375.1_ASM219137v1_genomic.fna.gz	Brucella melitensis	BwIM_ITA_55	0.000264	99.974	989/1000
GCA_006517305.1_ASM651730v1_genomic.fna.gz	Brucella melitensis	2011.TE.11844.1.1	0.000337	99.966	986/1000
GCA_047040135.1_M2206_genomic.fna.gz	Brucella melitensis	M2206	0.000361	99.964	985/1000
GCA_006506815.1_ASM650681v1_genomic.fna.gz	Brucella melitensis	2011.TE.6837.1.1	0.000410	99.959	983/1000
GCA_006506735.1_ASM650673v1_genomic.fna.gz	Brucella melitensis	2013.TE.15003.1.1	0.000483	99.952	980/1000
GCA_017177365.1_ASM1717736v1_genomic.fna.gz	Brucella melitensis	759	0.000557	99.944	977/1000

MLST: THE SEQUENCE TYPE

Sequence type	11
Allele profile	3-2-3-2-1-5-3-2-10
Every locus called	yes
Reads in the input	1,136,924
Reads matching an allele	2,824

INPUT FILES

FILE	SIZE	MODIFIED
SRR19520329_1.fastq.gz	35.9 MB	2026-11-07 15:45:34 EST
SRR19520329_2.fastq.gz	40.6 MB	2026-11-07 15:46:03 EST

```
SRR19520329_1.fastq.gz
MD5 12edf200445bafd03d033fac3d98118c
SRR19520329_2.fastq.gz
MD5 a81d3abf43d7a2849c9062061cd8603a
```

The margin each call had over the next allele of its locus; a ratio near 1.00 is the one to check.

LOCUS	ALLELE	READS	RUNNER-UP (ITS READS)	AHEAD	RATIO	NOTE
gap	3	350	gap_34 (250)	100	1.40	
aroA	2	324	aroA_1 (256)	68	1.27	
glk	3	282	glk_26 (184)	98	1.53	
dnaK	2	280	dnaK_6 (204)	76	1.37	
gyrB	1	374	gyrB_2 (266)	108	1.41	
trpE	5	348	trpE_31 (242)	106	1.44	
cobQ	3	260	cobQ_2 (194)	66	1.34	
int_hyp	2	328	int_hyp_20 (222)	106	1.48	
omp25	10	270	omp25_8 (182)	88	1.48	

NOTES

- nearly as close: 6 other Brucella melitensis genome(s)

SRR8550484 ok Brucella suis ST 15

IDENTIFICATION: THE MASH DISTANCE

Closest reference	GCA_000742055.1_ASM74205v1_genomic.fna.gz
What it is	Brucella suis str. Thompsen
NCBI accessions	GCA_000742055.1
MASH distance	0.000264
Estimated identity	99.974%
Shared min-hashes	989/1000 (at least 150 required)
K-mer copies required	14 (from the estimated 69x coverage)

THE SPECIES CALL

Species reported	Brucella suis
Decided by	MASH+ST
Species within reach	Brucella suis
Sequence type	15
What that ST is	ST 15 is Brucella suis in 99 of 103 isolate(s) (the rest: Brucella melitensis:2, Brucella ovis:1, Brucella sp.:1)

agree The sequence type names the same species as the closest reference genome, which named it on its own: the two kinds of evidence agree.

Closest first. A shaded row is too close to the first to be separated by distance alone. The other 427 reference(s) are further away; brucellaid_mash.tsv has every distance.

REFERENCE GENOME	SPECIES	STRAIN	DISTANCE	IDENTITY %	SHARED HASHES
GCA_000742055.1_ASM74205v1_genomic.fna.gz	Brucella suis	Thompsen	0.000264	99.974	989/1000
GCA_000365565.1_Bruc_suis_F8_06-2_V1_genomic.fna.gz	Brucella suis	F8/06-2	0.000459	99.954	981/1000
GCA_014884445.1_ASM1488444v1_genomic.fna.gz	Brucella suis	CVI_105	0.000582	99.942	976/1000
GCA_000292105.2_ASM29210v2_genomic.fna.gz	Brucella suis	BCB032	0.001839	99.816	927/1000
GCA_055657665.1_ASM5565766v1_genomic.fna.gz	Brucella suis	25Ref0017	0.001839	99.816	927/1000

MLST: THE SEQUENCE TYPE

Sequence type	15
Allele profile	1-2-7-1-3-3-5-3-2
Every locus called	yes
Reads in the input	1,693,672
Reads matching an allele	3,204

INPUT FILES

FILE	SIZE	MODIFIED
SRR8550484_1.fastq.gz	71.6 MB	2026-11-07 15:48:53 EST
SRR8550484_2.fastq.gz	72.4 MB	2026-11-07 15:49:45 EST

```
SRR8550484_1.fastq.gz
MD5 68833febebe273fadd9dc6ee0e95ed16
SRR8550484_2.fastq.gz
MD5 9b8438792c71722cd7fb4fb0d18e414c
```

The margin each call had over the next allele of its locus; a ratio near 1.00 is the one to check.

LOCUS	ALLELE	READS	RUNNER-UP (ITS READS)	AHEAD	RATIO	NOTE
gap	1	508	gap_14 (412)	96	1.23	
aroA	2	284	aroA_16 (232)	52	1.22	
glk	7	282	glk_4 (202)	80	1.40	
dnaK	1	302	dnaK_27 (224)	78	1.35	
gyrB	3	384	gyrB_1 (290)	94	1.32	
trpE	3	346	trpE_5 (296)	50	1.17	
cobQ	5	294	cobQ_4 (220)	74	1.34	
int_hyp	3	444	int_hyp_1 (272)	172	1.63	
omp25	2	330	omp25_21 (250)	80	1.32	

NOTES

- nearly as close: 2 other Brucella suis genome(s)

4 Definitions and methods

The references listed for each sample

A sample block lists the five closest reference genomes, and any other near enough to the closest that the distance does not separate them, which is the evidence the species call rests on. brucellaid_mash.tsv holds the distance to every reference of the sketch.

The species, from a MASH distance

The reads of a sample are reduced to the 1000 smallest hashes of their 21-mers, ignoring the k-mers seen fewer times than the copy threshold derived from the coverage, and that sketch is compared with the sketch of each reference genome. The distance is an estimate of the fraction of positions that differ, so 0.000024 is about 99.998% identity. The closest reference names the species, provided the reads share at least the required number of its 1000 min-hashes; below that the sample is reported as ND, because the closest genome of the set is not an answer. The reference data table above says which set this run measured against.

Why there is no biovar in this report

BrucellaID does not report a biovar, because it cannot measure one. Three things stand in the way. Hardly any reference carries a biovar at all: 7 of the 432 genomes of the default set do, because the NCBI records of the rest carry none, so 93% of the biovar-labelled B. abortus isolates of the PubMLST isolate database are of a biovar no reference has. The distance cannot separate the biovars that are there: the two B. suis biovar 2 genomes are 0.00163 apart, further than the biovar 3 genome is from one of them (0.00150). And a biovar is a phenotype that genome clustering does not track -- the best any rule from sequence type to biovar could manage over the isolate database is 66% for B. abortus and 63% for B. melitensis, because one ST spans several biovars. This report therefore shows no biovar at all, not even the ones recorded for the ST: of the 118 sequence types in the table, 42 have a single biovar on record, and showing one next to a sample invites reading it as a measurement of that sample. The counts are in the packaged mlst_st_species.tsv for anyone who wants them. For a biovar, use classical biotyping, or brucetyper, which reads the Bruce-ladder and Suis-ladder assays in silico.

The sequence type, from the alleles

The same reads are given to Seal with the alleles of the nine loci as references, counting each read for every allele it shares the highest number of 25-mers with and for no other. The allele the sample carries is therefore counted for every read of its locus, and a neighbouring allele only for the reads that fall where the two agree, so the allele with the most reads wins. No mismatch is allowed in a k-mer, because the alleles of a locus differ by as little as one base. Two alleles with the same count are reported as a tie rather than separated by a coin toss, and a locus below the minimum read count is not called.

What the ST values mean

A number is a profile that is in the table. "new" means all nine alleles were called but the combination is not in the table, which is either a genuinely new ST or a novel allele called as its nearest neighbour. "ND" means at least one locus had no call; the note says which.

How the species is decided when two are equally close

Some Brucella species are closer to each other than the variation within either of them: in the sketches that ship with BrucellaID, the closest pair of different species is always a B. canis and a B. suis: 0.00024 apart in the 16 genomes of vSNP, 0.000264 in the 432 of the default set. Where a reference of a second species is within the close-match margin of the closest one, the distance has not really chosen, and the sequence type chooses among those species, because the two share no ST. An ST is only used when enough of its isolates agree on a species, and an ST that contradicts a species the distance picked cleanly is reported as a conflict and not obeyed.

What this report does not establish

That the sample is what the label on the tube says, that the library is not a mixture of two isolates, or that the reference data is current. It establishes what the software was given, what it measured, and what it concluded from those measurements.

5 Run information

Operator	M-O Duceppe
User account	[removed]
Computer	[removed]
Operating system	[removed]
Working directory	[removed]
Started	2026-11-08T14:55:06-05:00
Finished	2026-11-08T14:55:43-05:00
Command	python3 scripts/make_example_report.py reads

Software

PROGRAM	VERSION
brucellaid	2.1.1
Python	3.12.16
MASH	2.3
BBTools	40.05
Java	openjdk version "25.0.4.1" 2026-08-18

Parameters

PARAMETER	VALUE
Threads	16
Parallel samples	2
MLST	True
Input checksums	MD5

Reference data

DATA	WHAT IT HOLDS	SOURCE	MD5
MASH sketch	432 reference genome(s), k = 21, 1000 hashes	the package's	800b94a5f74a266e7c08309e78c891a2
Reference table	432 genome(s) with their species, biovar, strain and accessions	the package's	621f43f317759e9fb04eece7f9ec4acb
MLST alleles	309 allele(s) of the 9 loci	the package's	95fcf5e43f4cbf7ebe840c8e54bfbf46
MLST profiles	122 sequence type(s)	the package's	dd4bc06867bd58ce2a8f362755aa79a7
ST to species	118 sequence type(s) with a species	the package's	3a6751b8a1ab1273f3bc0e19e08efb7f

```
.../brucellaid/data/references_extended.msh
.../brucellaid/data/references_extended.tsv
.../brucellaid/data/mlst_alleles.fasta
.../brucellaid/data/mlst_profiles.tsv
.../brucellaid/data/mlst_st_species.tsv
```

References

1. Whatmore AM, Perrett LL, MacMillan AP. Characterisation of the genetic diversity of Brucella by multilocus sequencing. BMC Microbiology 2007;7:34. doi:10.1186/1471-2180-7-34
2. Ondov BD, Treangen TJ, Melsted P, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology 2016;17:132. doi:10.1186/s13059-016-0997-x
3. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Research 2018;3:124. doi:10.12688/wellcomeopenres.14826.1
4. Bushnell B. BBTools. https://sourceforge.net/projects/bbmap/