REPORT

```
LegioCluster version: 24 September 2020
Date submitted: 2020-10-01
Submitted by: WH
Isolate name: Spy_sample_1
Species: Spy
Forward reads: /projdata/WH_PL/Github/LegioCluster/reads/Spy/Spy_sample_1_R1_001.fastq.gz
Reverse reads: /projdata/WH_PL/Github/LegioCluster/reads/Spy/Spy_sample_1_R2_001.fastq.gz
Metadata: tutorial_part_2 multiple_submissions normal_operations
Folder name: WH201001_180341
```

Read pre-processing (Trimmomatic):

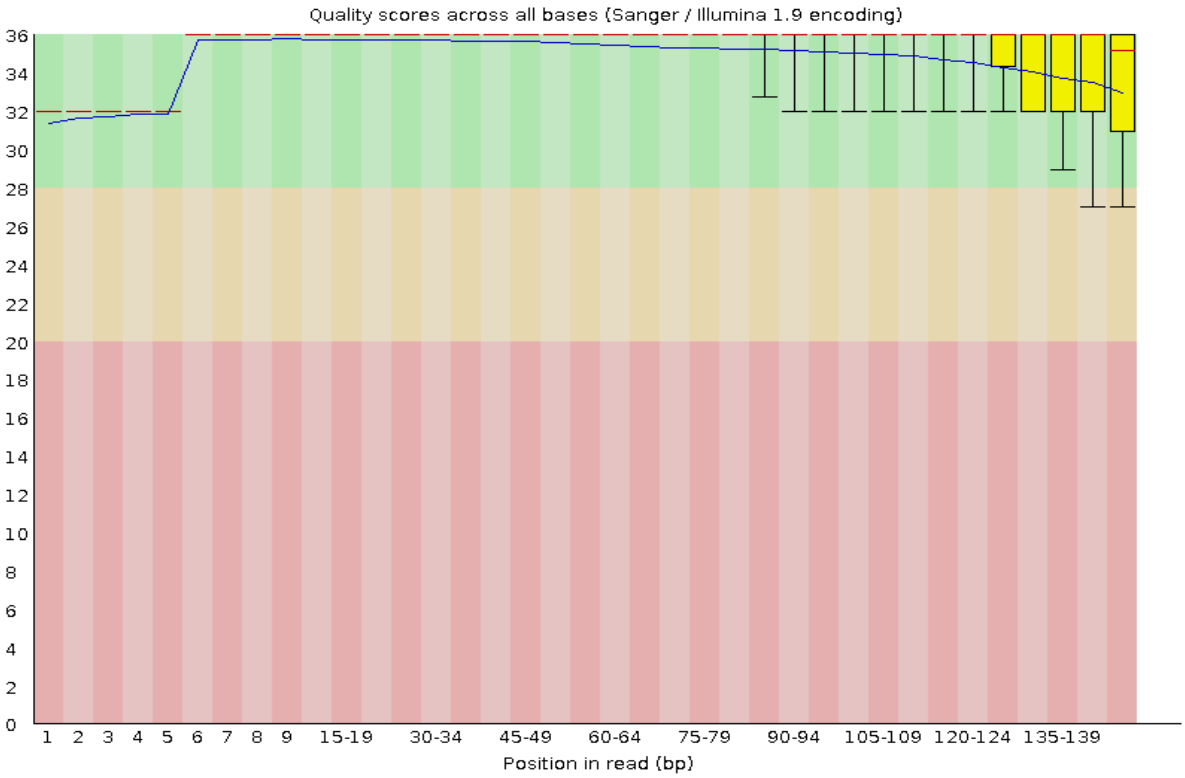
```
Adapters removed, low quality (< Q20) regions removed, short reads (<100) removed, ploy-G (>25) removed
Input read pairs: 299710
Both surviving: 252161 (84.13%)
Forward only surviving: 11725 (3.91%)
Reverse only surviving: 5794 (1.93%)
Dropped read pairs: 30030 (10.02%)
Mean (SD) lengths of trimmed F reads: 128.53 (48.037)
Mean (SD) lengths of trimmed R reads: 125.34 (51.115)
Mean (SD) no. of bases trimmed from 5' of F reads(*): 0.0 (0.013)
Mean (SD) no. of bases trimmed from 5' of R reads(*): 0.0 (0.093)
Mean (SD) no. of bases trimmed from 3' of F reads(*): 1.05 (5.202)
Mean (SD) no. of bases trimmed from 3' of R reads(*): 1.36 (5.856)
(*) if trimmed read length > 0
```

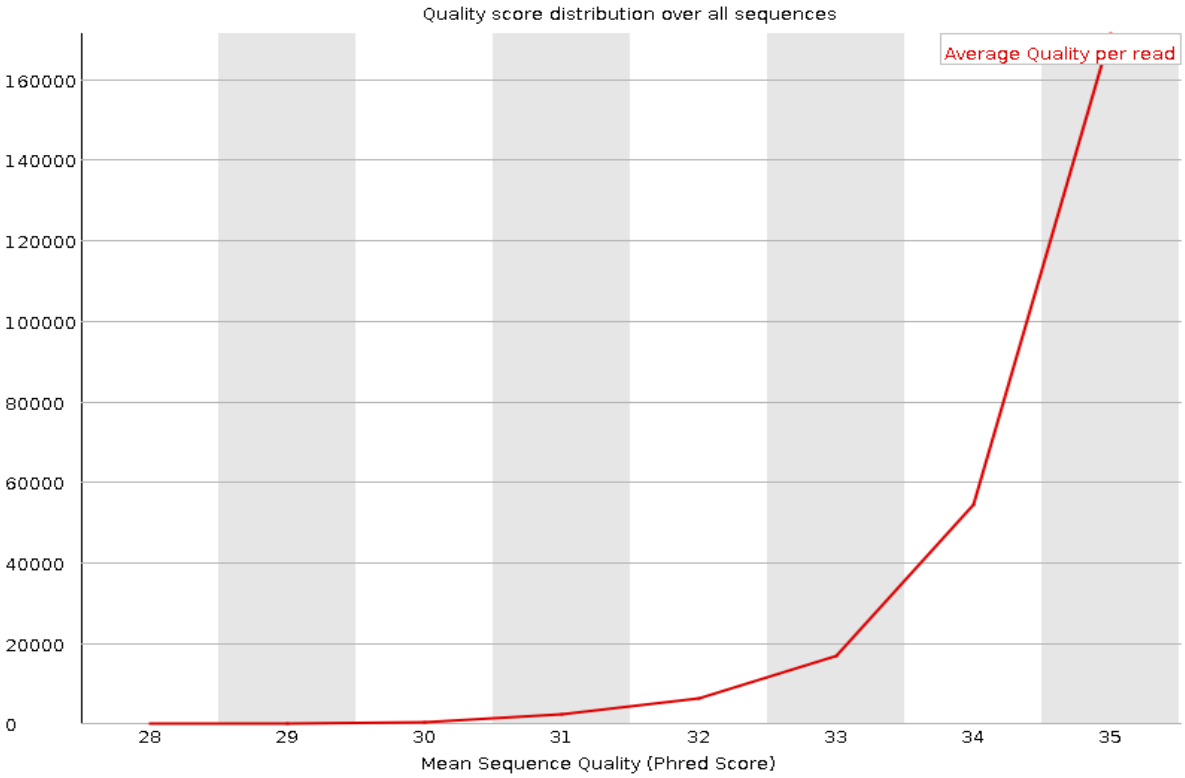
The pipeline will not continue if there are less than 150,000 read pairs.

Read quality control (FastQC results):

```
Results for processed reads from:
/projdata/WH_PL/Github/LegioCluster/reads/Spy/Spy_sample_1_R1_001.fastq.gz
Filename paired_reads_1.fq
Total Sequences 252161
Sequences flagged as poor quality 0
Sequence length 100-149
%GC 38
PASS Basic Statistics
PASS Per base sequence quality
PASS Per tile sequence quality
PASS Per sequence quality scores
FAIL Per base sequence content
PASS Per sequence GC content
PASS Per base N content
WARN Sequence Length Distribution
PASS Sequence Duplication Levels
PASS Overrepresented sequences
```

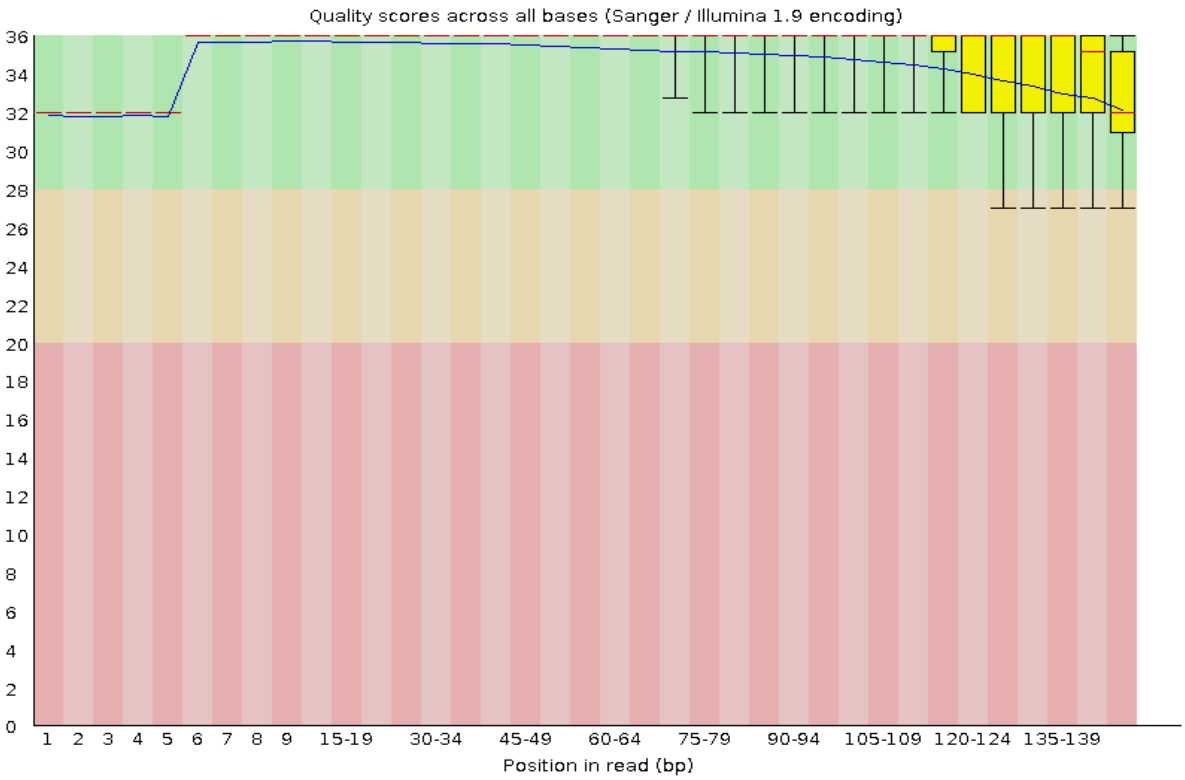
PASS Adapter Content

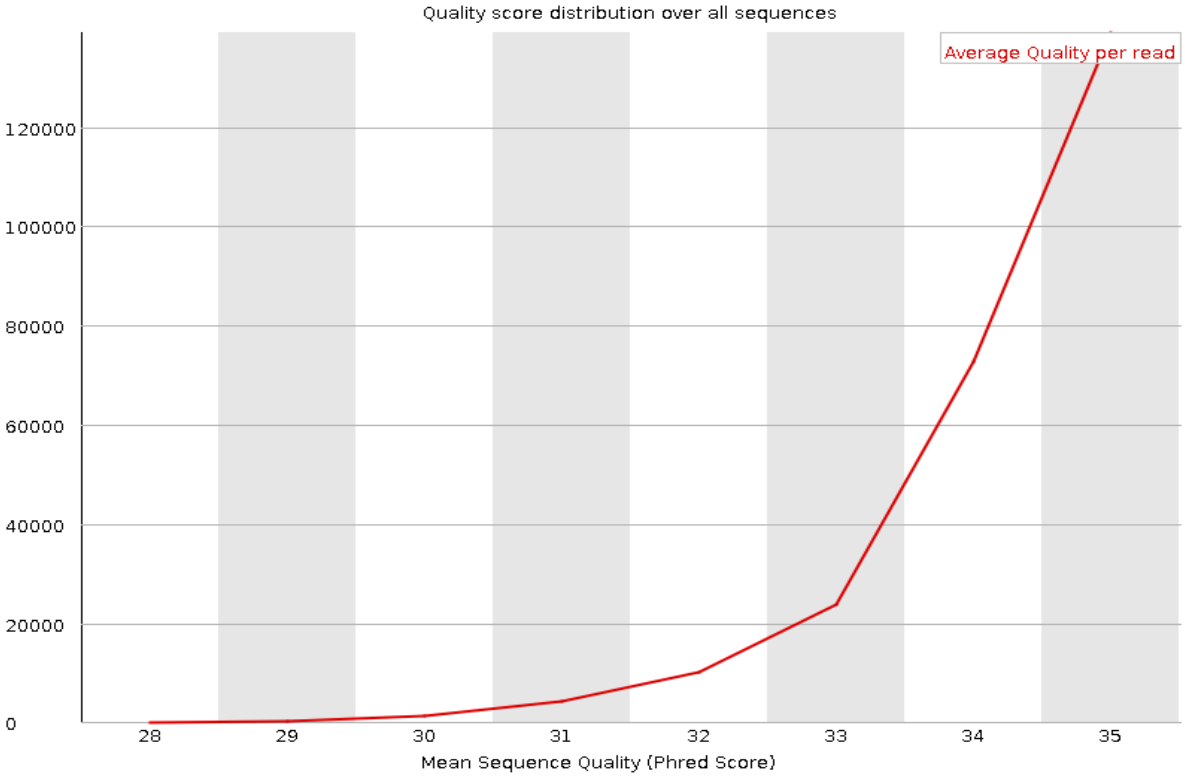




Read quality control (FastQC results):

```
Results for processed reads from:
/projdata/WH_PL/Github/LegioCluster/reads/Spy/Spy_sample_1_R2_001.fastq.gz
Filename paired_reads_2.fq
Total Sequences 252161
Sequences flagged as poor quality 0
Sequence length 100-149
%GC 38
PASS Basic Statistics
PASS Per base sequence quality
PASS Per tile sequence quality
PASS Per sequence quality scores
FAIL Per base sequence content
PASS Per sequence GC content
PASS Per base N content
WARN Sequence Length Distribution
PASS Sequence Duplication Levels
PASS Overrepresented sequences
PASS Adapter Content
```





Coverage: (252161 * 149 * 2) / 1831320 = 41.033

Percentage of bases with quality score >= Q30 (251862.0 * 100) / 252161.0 = 99.881

Contamination check (Mash):

Reference with the shortest distance
Strain name: Spy
Mash distance: 0.0113587
P-value: 0.0
Matching hashes: 286/400

The pipeline will not continue if the wrong species returns the highest score or if the Mash distance is above threshold.

These reads seem to have come from: Streptococcus pyogenes or a related species.

Runner up
Strain name: Cdi
Mash distance: 0.262949
P-value: 0.000289154
Matching hashes: 3/400

There should be only a few matching hashes for the runner-up. Anything above 20/400 might indicate a contamination.

These reads seem to have come from: Clostridioides difficile or a related species.

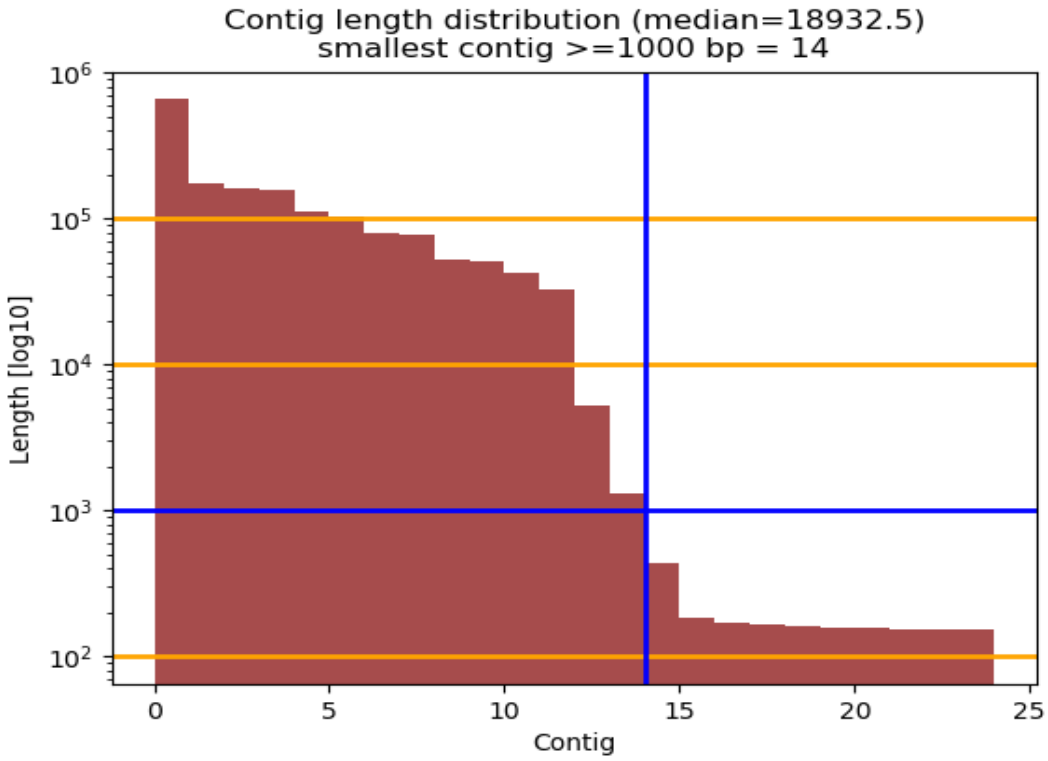
Mash QC results: PASSED QC

De novo assembly (SPAdes):

contig	length (bp)	coverage
1	656942	18.186882
2	176609	21.963939
3	162950	19.848551
4	156517	18.693863
5	112615	20.321038
6	102408	22.289062
7	79962	21.811792
8	76995	21.152799
9	52008	22.033390
10	50951	23.933188
11	42638	23.335800
12	32669	19.175933
13	5196	127.540926
14	1316	129.384988
15	441	62.197802
16	183	24.000000
17	172	34.600000
18	168	91.252747
19	162	21.752941
20	159	17.524390
21	158	16.691358
22	155	21.051282
23	155	16.910256
24	155	9.846154

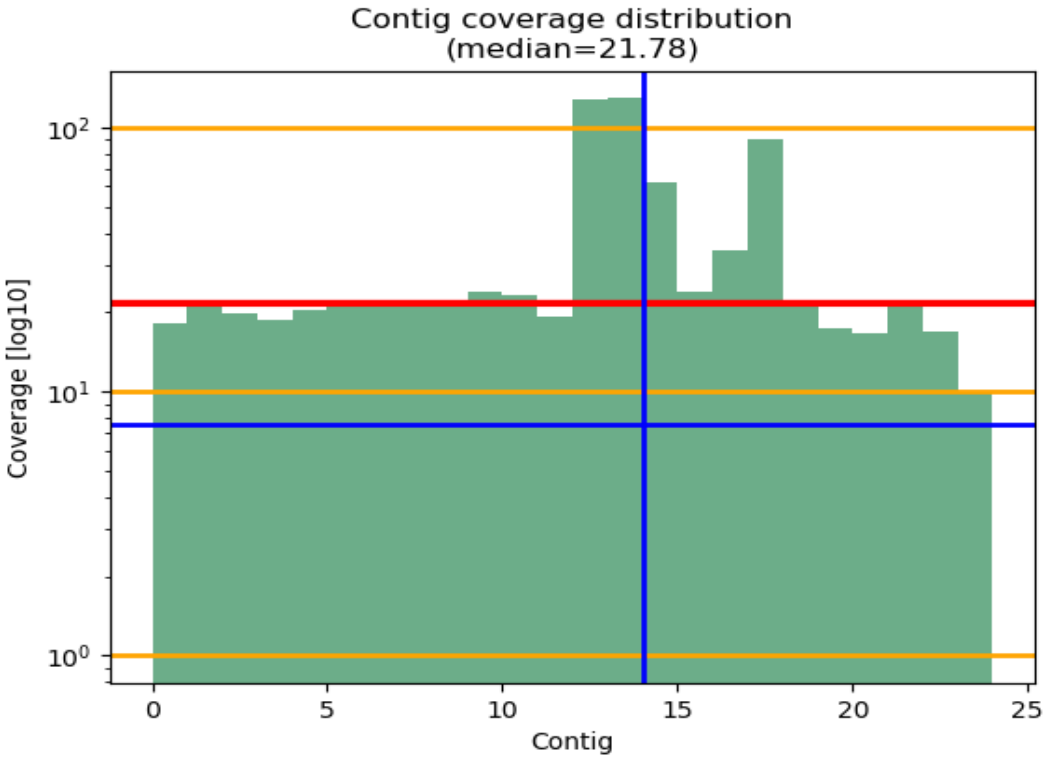
Contigs with an unusually high coverage indicate repeat regions or plasmids, which might be of interest.

Depending on the species, too many reads indicate poor data quality and/or contamination issues. The pipeline will not continue if there are too many contigs (350 for Legionella, 500 for other species).



Looking for contigs with a length below 1000 bp that indicate poor data quality.

Figure: contigs vs length



Looking for contigs with
a coverage below 7.5
that indicate poor data quality.

Figure: contigs vs coverage

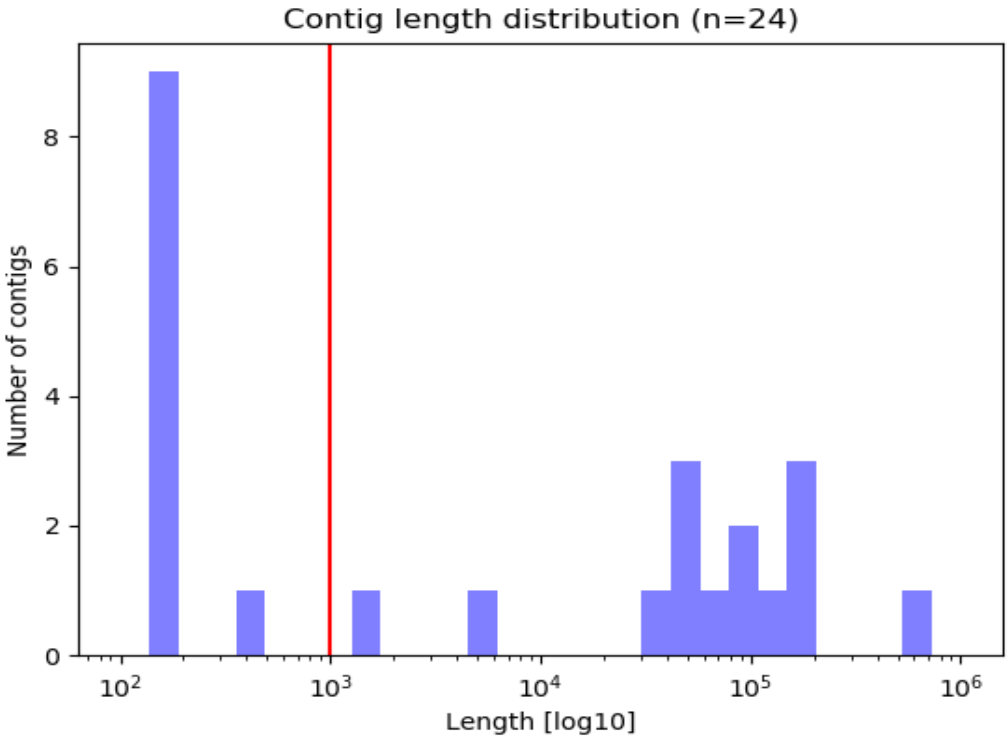


Figure: contig length distribution

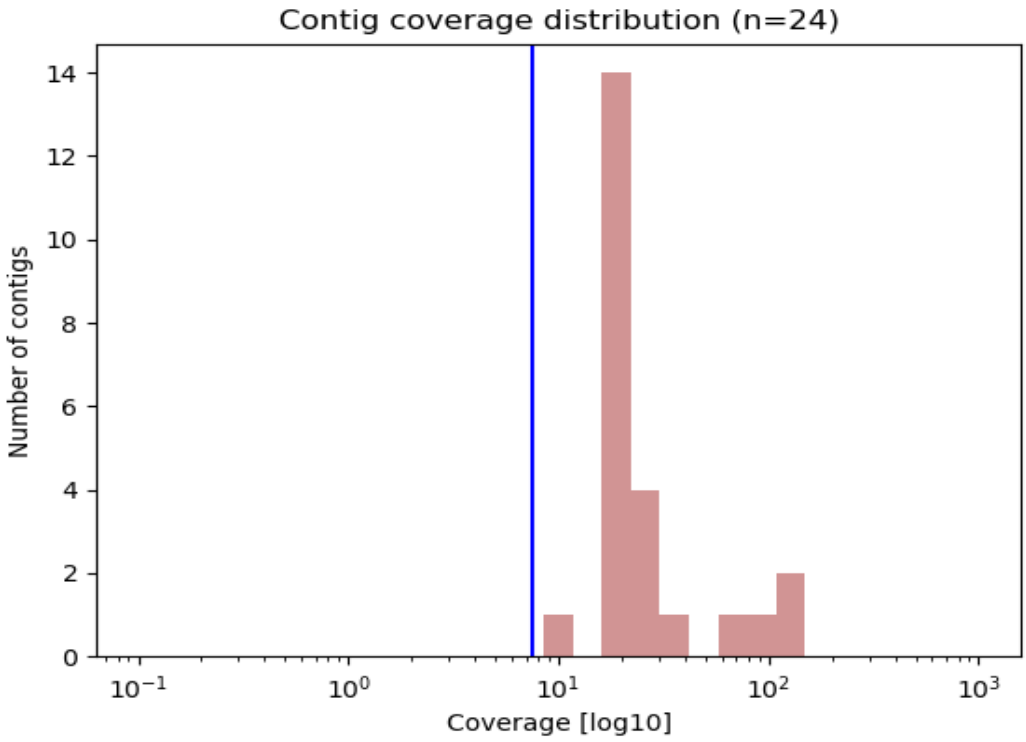
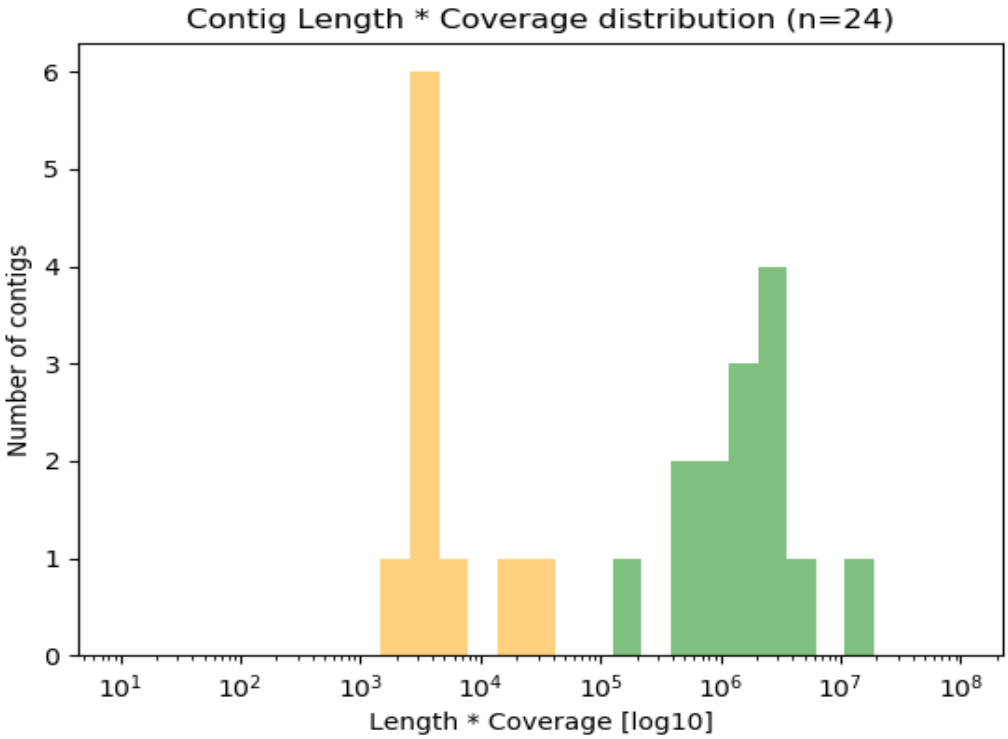


Figure: contig coverage distribution



Green:
contigs >= 1 kb and with >= 7.5 fold coverage = high quality contigs

Yellow:
contig length < 1 kb or coverage < 7.5 fold

Red:
contig length < 1 kb and coverage < 7.5 fold = potential low grade contamination

Black:
contigs >= 1 kb and with >= 250 fold coverage = potential plasmids.

Figure: contig length * coverage distribution

Contig analysis:

```
(min length: 1000 bp, min coverage: 7.5x)
contigs that fail both thresholds: 0.0 %
contigs that are too short or have a low coverage: 41.67 %
contigs that meet both thresholds: 58.33 %
contigs with a high coverage (> 250x): 0.0 %
```

Finding a reference strain (Mash):

```
Reference with the shortest distance
Strain name: M1_GAS
Mash distance: 0.0110444
P-value: 0.0
Matching hashes: 657/1000

Reference with the shortest distance
Strain name: ATCC_19615
Mash distance: 0.0123068
P-value: 0.0
Matching hashes: 629/1000
```

Two candidate reference strains yielded similar Mash results. In that case, the reads will be mapped to both genomes and the one with the higher percentage of mapped reads chosen for the SNP/indel calling.

```
Runner up
Strain name: ATCC_19615
Mash distance: 0.0123068
P-value: 0.0
Matching hashes: 629/1000
```

Since there were only two candidate reference genomes, the one with the lower score appears a second time, as runner-up.

```
Mash QC results: PASSED QC
```

Mapping the query against strain M1_GAS.fa (BWA MEM):

Alignment QC (Samtools flagstat):

```
505610 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
1288 + 0 supplementary
7321 + 0 duplicates
492633 + 0 mapped (97.43% : N/A)
504322 + 0 paired in sequencing
252161 + 0 read1
252161 + 0 read2
486694 + 0 properly paired (96.50% : N/A)
489258 + 0 with itself and mate mapped
2087 + 0 singletons (0.41% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

```
Percentage of mapped reads: 97.43
```

Strain M1_GAS has the higher percentage of mapped reads and will be used for SNP/indel calling by Freebayes.

97.45% indicates that 2.55% of reads were unmapped. These unmapped reads could be just "noise" or correspond to unique DNA in the query genome, such as phages or plasmids.

Alignment QC (Samtools idxstats):

```
ref_fa_file len mapped unmapped
Streptococcus.pyogenes.M1.GAS.complete.sequence_NC.002737.2_length_1852433_cov_1.000 1852433 492633 2087
* 0 0 10890
```

Genomic fragments:

```
Smallest fragment: 0
Mean length: 3136.84
S.D.: 57508.52
median: 375.0
Largest fragment: 1852433
```

Mapping the query against strain ATCC_19615.fa (BWA MEM):

Alignment QC (Samtools flagstat):

```
505517 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
1195 + 0 supplementary
7297 + 0 duplicates
480637 + 0 mapped (95.08% : N/A)
504322 + 0 paired in sequencing
252161 + 0 read1
252161 + 0 read2
473580 + 0 properly paired (93.90% : N/A)
476714 + 0 with itself and mate mapped
2728 + 0 singletons (0.54% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

```
Percentage of mapped reads: 95.08
```

Here, the second reference candidate is less ideal than the first one.

If none of the reference candidates' mapping results exceed 90%, then the query genome will found a new cluster and the genome added to the list of candidate references, so that subsequent isolate can be mapped to it.

Alignment QC (Samtools idxstats):

```
ref_fa_file len mapped unmapped
Streptococcus.pyogenes.strain.ATCC.19615.chromosome.complete.genome_NZ.CP008926.1_length_1844804_cov_1.000
1844804 480637 2728
* 0 0 22152
```

Genomic fragments:

```
Smallest fragment: 0
Mean length: 3123.77
S.D.: 50752.13
median: 363.0
Largest fragment: 1844597
```

Assembly quality check (Quast results) for SPAdes_contigs.fa:

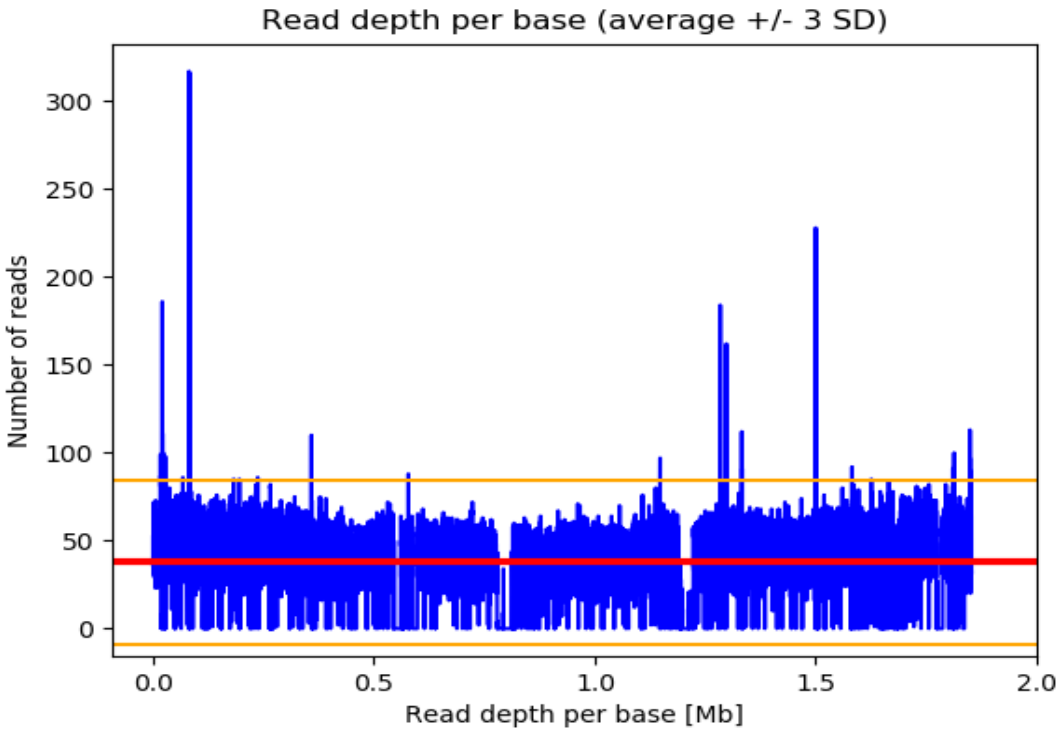
```
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0
bp)" and "Total length (>= 0 bp)" include all contigs).

Assembly SPAdes_contigs
# contigs (>= 0 bp) 24
# contigs (>= 1000 bp) 14
# contigs (>= 5000 bp) 13
# contigs (>= 10000 bp) 12
# contigs (>= 25000 bp) 12
# contigs (>= 50000 bp) 10
Total length (>= 0 bp) 1711684
Total length (>= 1000 bp) 1709776
Total length (>= 5000 bp) 1708460
Total length (>= 10000 bp) 1703264
Total length (>= 25000 bp) 1703264
Total length (>= 50000 bp) 1627957
# contigs 14
Largest contig 656942
Total length 1709776
Reference length 1852433
Reference GC (%) 38.51
N50 162950
NG50 162950
N75 102408
NG75 79962
L50 3
LG50 3
L75 6
LG75 7
# misassemblies 21
# misassembled contigs 9
Misassembled contigs length 1430650
# local misassemblies 25
# scaffold gap ext. mis. 0
```

```
# scaffold gap loc. mis. 0
# unaligned mis. contigs 0
# unaligned contigs 0 + 7 part
Unaligned length 49625
Genome fraction (%) 89.408
Duplication ratio 1.002
# N's per 100 kbp 0.00
# mismatches per 100 kbp 953.86
# indels per 100 kbp 34.05
Largest alignment 197931
Total aligned length 1658336
NA50 76526
NGA50 74071
NA75 48894
NGA75 36346
LA50 7
LGA50 8
LA75 14
LGA75 16
```

Alignment QC (Samtools depth):

```
Total number of bases: 1852433
Number (percent) of bases with read depth < 1: 159017 (8.58%)
Number (percent) of bases with read depth >= 1: 1693416 (91.42%)
Average read depth (S.D.): 37.88 (15.563)
Average read depth (S.D., count) for bases with read depth >= 1: 41.44 (10.84, 1693416)
Average read depth (S.D., count) for bases with read depth > 0 and < 1: 0 (0, 0)
Average read depth (S.D., count) for bases with read depth == 0: 0.0 (0.0, 159017)
Number of gaps >= 100 bases: 88
List of gaps >= 100 bases: [106, 107, 113, 114, 119, 120, 120, 124, 126, 130, 136, 138, 138, 141, 142,
153, 154, 155, 157, 163, 167, 171, 174, 175, 177, 178, 179, 188, 201, 205, 213, 214, 214, 215, 240, 242,
252, 265, 267, 280, 282, 297, 298, 324, 324, 370, 380, 386, 389, 398, 403, 406, 435, 468, 552, 553, 581,
613, 714, 835, 903, 936, 972, 1053, 1208, 1275, 1299, 1313, 1355, 1498, 1569, 1695, 2080, 2101, 2237,
2525, 3096, 3118, 4185, 5779, 6666, 7783, 10395, 12247, 12458, 14569, 14934, 19720]
Total number of bases in gaps >= 100 bases: 154850
```

The percentage of unmapped reads (see above) is an indication of DNA unique to the query.
The total number of bases in gaps indicates DNA that is unique to the reference genome.
An alignment and visualization (e.g. with Mauve), helps to identify unique stretches of DNA.



Unusually high peaks indicate repeat regions.

Gaps suggest DNA unique to the reference.

Figure: Read depth per base_1 (plot)

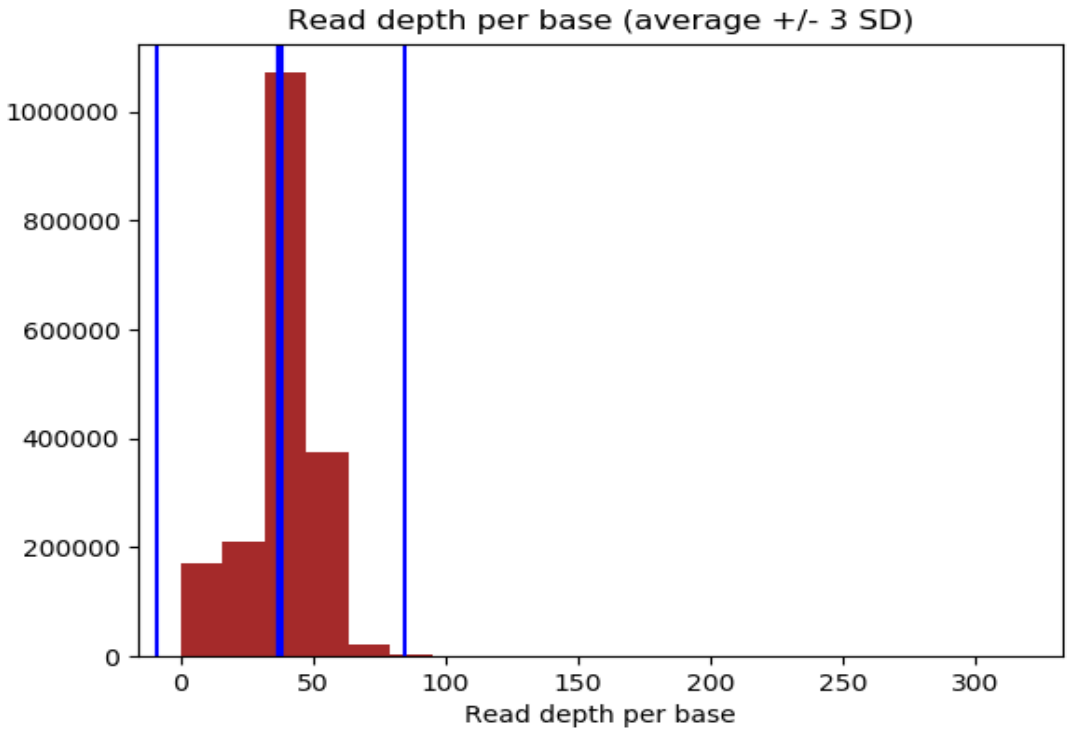


Figure: Read depth per base_1 (histogram)

Mapping quality check (Qualimap results):

```
number of bases = 1,852,433 bp
number of contigs = 1
number of reads = 505,610
number of mapped reads = 492,633 (97.43%)
number of mapped bases = 71,233,605 bp
mean mapping quality = 56.3461
```

SNPs and INDEL events between Spy_sample_1 and reference M1_GAS (FreeBayes):

ME (IE, BID, SNP)

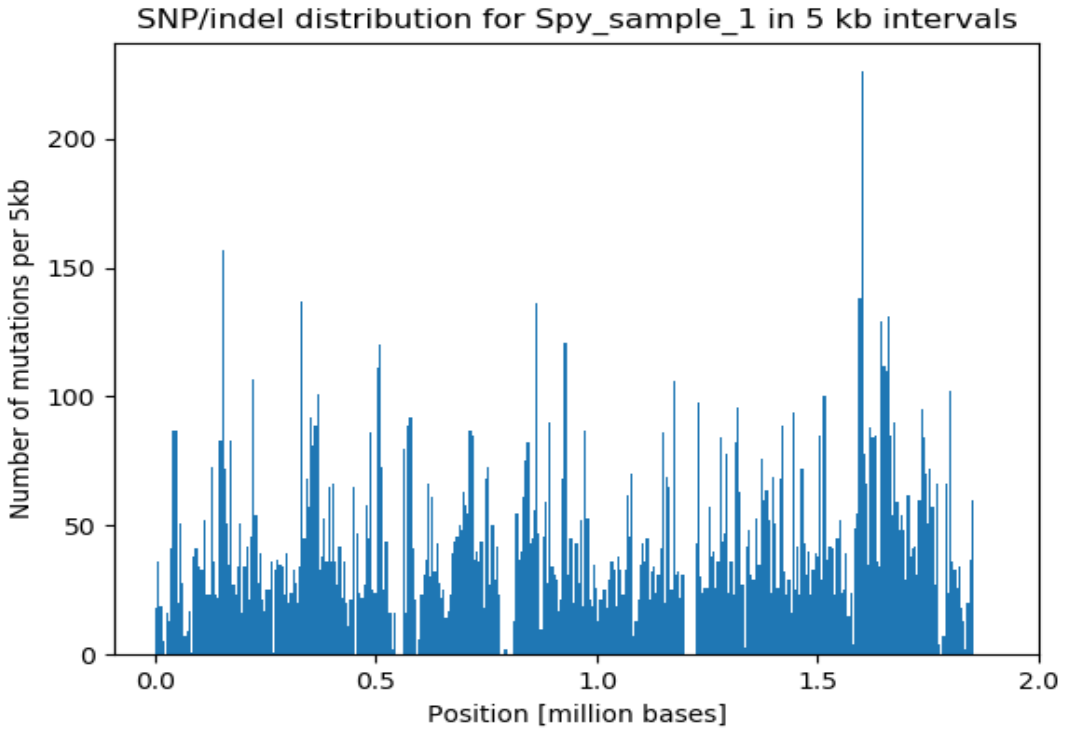
Found 15052 (334, 1071, 14718) SNPs and INDEL events compared to a reference genome of 1852433 bp.
(Note that the indel event count might be slightly lower in the SNP-matrix.)

NOTE:

ME = mutation events (IE + SNP)
IE = insertions and deletion events
BID = bases in insertions or deletions
SNP = single nucleotide polymorphism

Number of SNPs and INDEL events above threshold. Adding the isolate to the list of reference genomes.

Too many mutation events suggest that the reference and query are very distinct. In this case, the query genome will found a new cluster and the genome added to the list of candidate references, so that subsequent isolate can be mapped to it.



Hot spots could indicate regions of horizontal gene transfer.

Figure: SNP/INDEL distribution

Note:

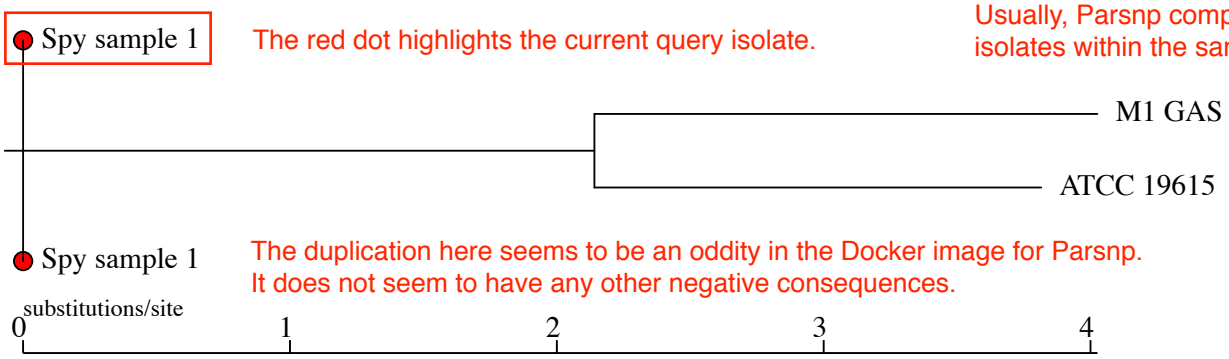
The isolate passed the QC check for new references.

Each potential new reference candidate will have to pass stringent quality control checks to ensure that the genome data are robust. If an isolate fails the QC check, it will not be added to any existing cluster or form a new one. However, the user can run the isolate again and force the creation of a new cluster based on this isolate, if re-sequencing the sample is not an option.

```
Please run a Blast search on the isolate.fa file to make sure the sample is not contaminated.
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch
```

Phylogentic analysis of the core genome (Parsnp):

Here, we compare the new reference to all other references. Usually, Parsnp compares all isolates within the same cluster.



The red dot highlights the current query isolate.

The duplication here seems to be an oddity in the Docker image for Parsnp. It does not seem to have any other negative consequences.