REPORT

```
LegioCluster version: 24 September 2020
Date submitted: 2020-10-01
Submitted by: WH
Isolate name: Spy_sample_1_rerun
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: set_ref=M1_GAS Tutorial_part_3:_individual_submissions override_reference_selection
Folder name: WH201001_183415
```

"Spy_sample_1" had been run before, so a name change is required or the sample won't be processed.

Forces the isolate to be placed into the same cluster as the designated reference strain. Generally not a good idea, but might be of interest to epidemiologists.

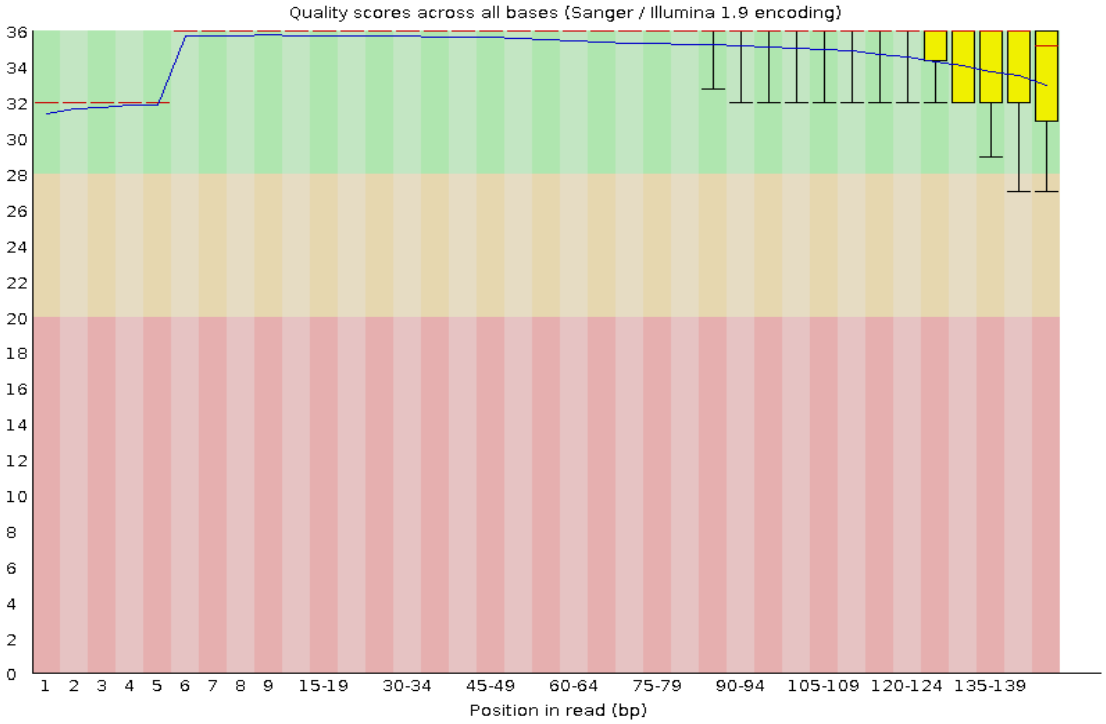
Read pre-processing (Trimmomatic):

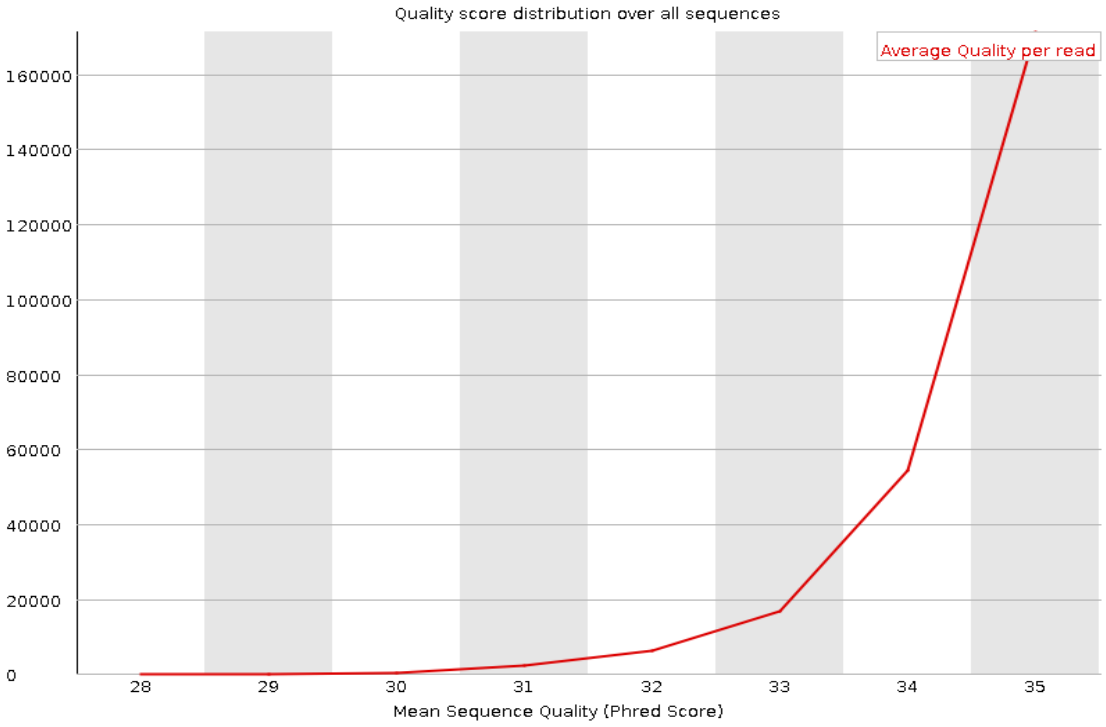
```
Adapters removed, low quality (< Q20) regions removed, short reads (<100) removed, ploy-G (>25) removed
Input read pairs: 299710
Both surviving: 252161 (84.13%)
Forward only surviving: 11725 (3.91%)
Reverse only surviving: 5794 (1.93%)
Dropped read pairs: 30030 (10.02%)
Mean (SD) lengths of trimmed 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
```

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
```

This information is printed for the user's information.
It has no impact on the execution of the pipeline.

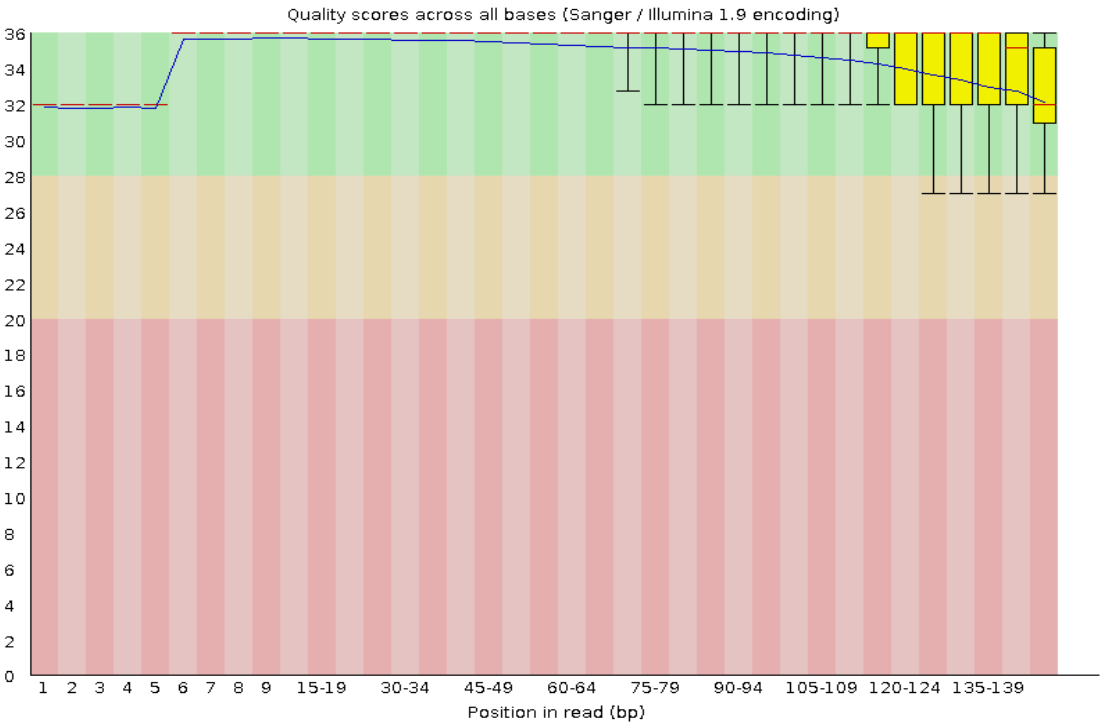


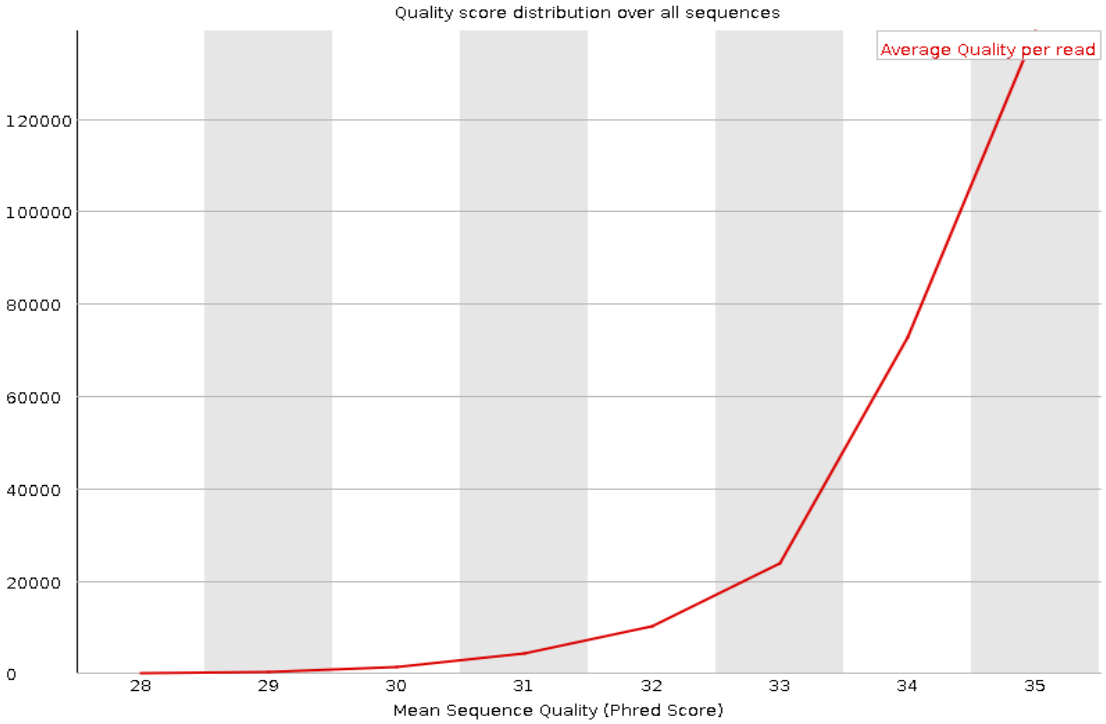


Read quality control (FastQC results):

```
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
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
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
```

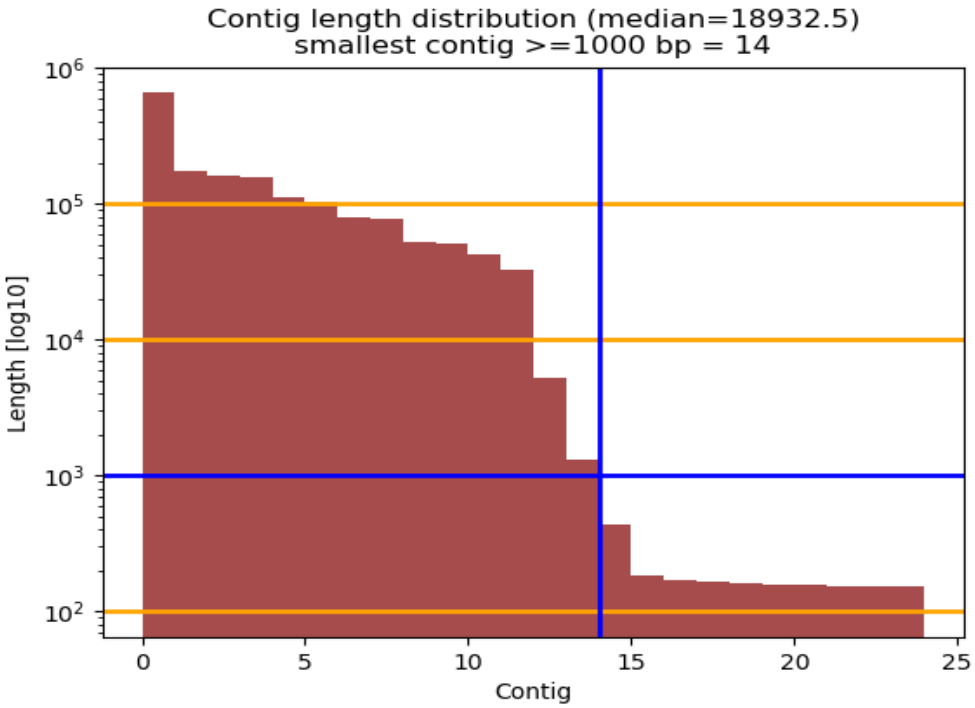


Figure: contigs vs length

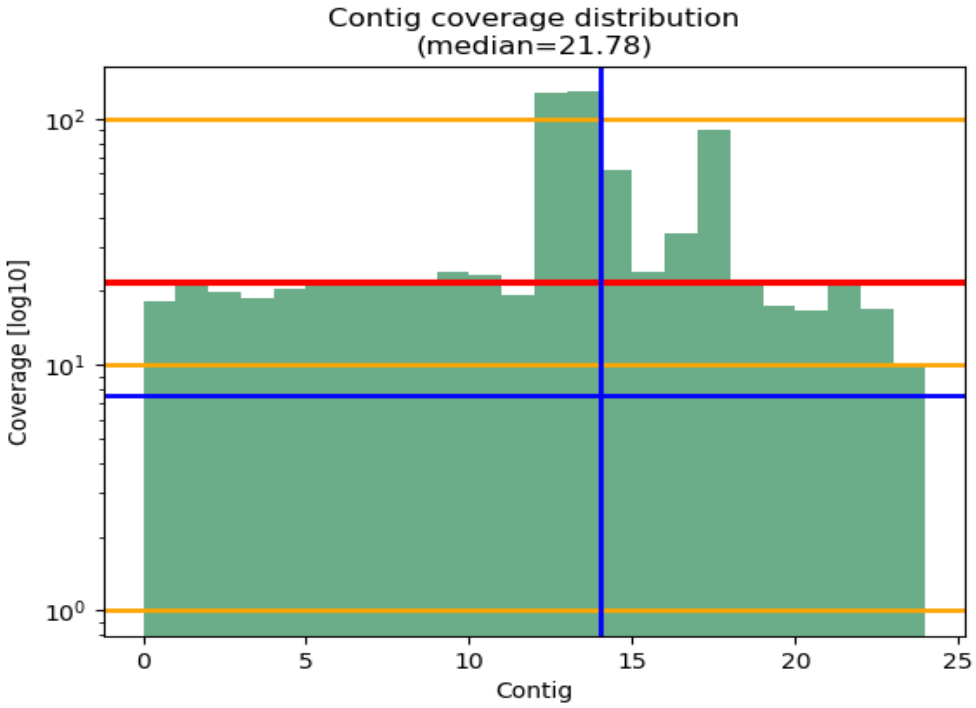


Figure: contigs vs coverage

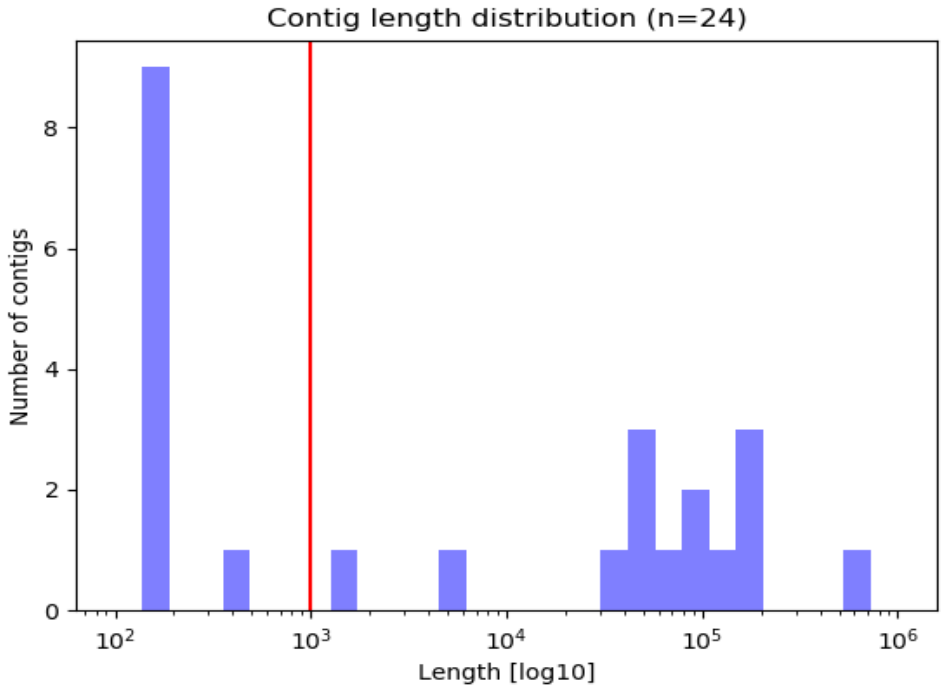


Figure: contig length distribution

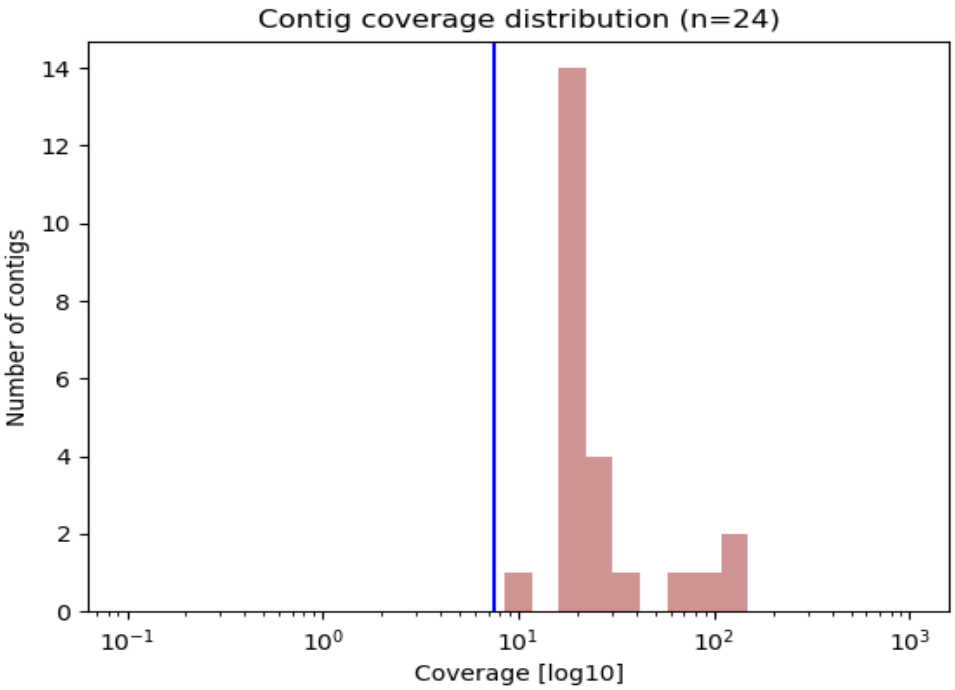


Figure: contig coverage distribution

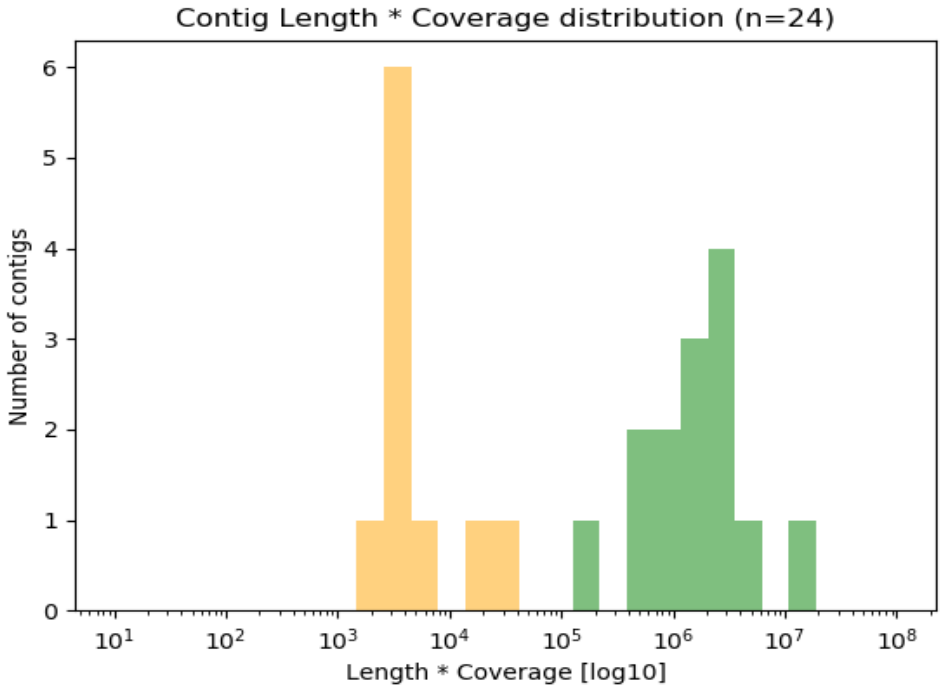


Figure: contig length * coverage distribution

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: Spy_sample_1
Mash distance: 0.0
P-value: 0.0
Matching hashes: 1000/1000

Runner up
Strain name: M1_GAS
Mash distance: 0.0110444
P-value: 0.0
Matching hashes: 657/1000

Mash QC results: PASSED QC
```

Spy_sample_1 was added during Tutorial part 2 as a new candidate reference and is now the best possible reference.

```
Over-writing Mash-selected reference with User pre-selected reference(s): ['M1_GAS.fa']
```

Spy_sample_1 is the better reference, but that selection was over-written by the user.

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
```

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
```

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
```

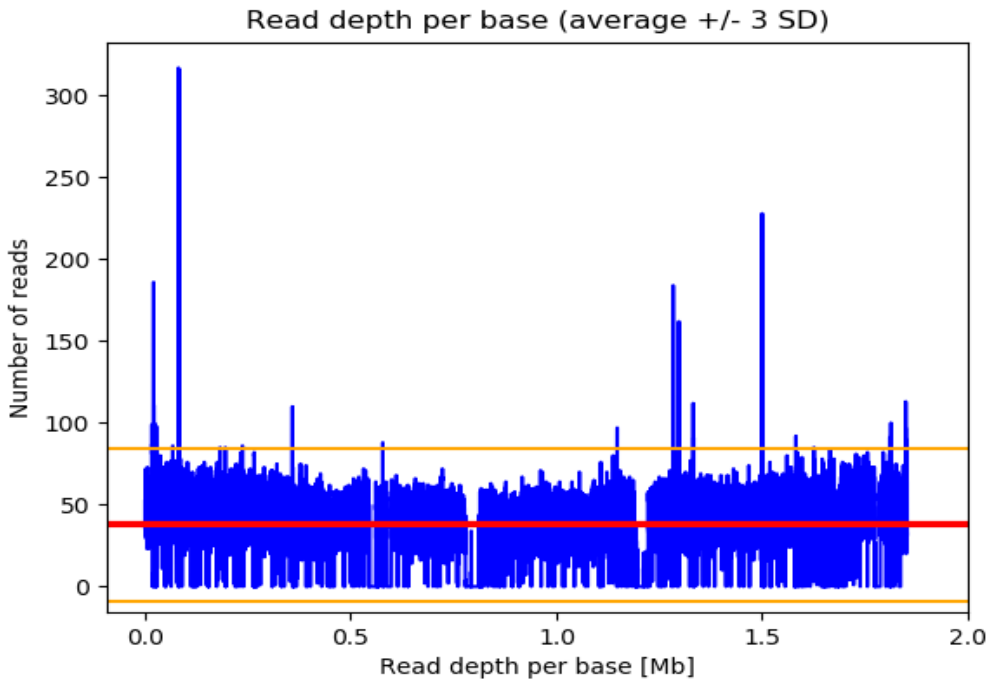


Figure: Read depth per base_1 (plot)

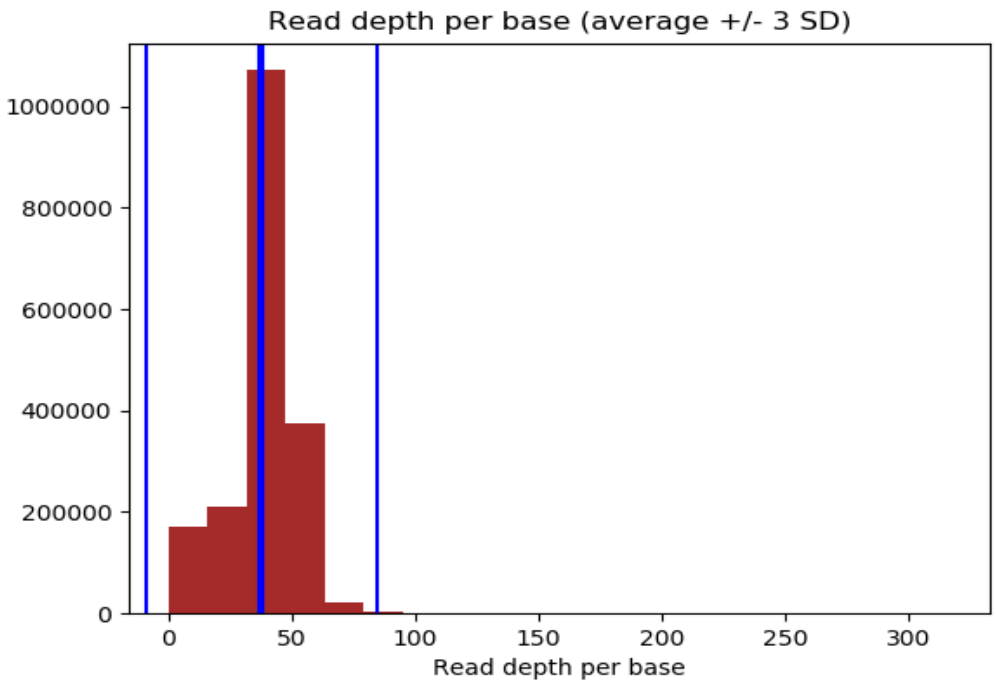


Figure: Read depth per base_1 (histogram)

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_rerun and reference M1_GAS (FreeBayes):

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.)

Under normal circumstances, this many mutation events for S. pyogenes would cause the pipeline to add the isolate to the list of reference candidates and create a new cluster. However, that function has been suppressed after forcing strain M1_GAS to be the designated reference.

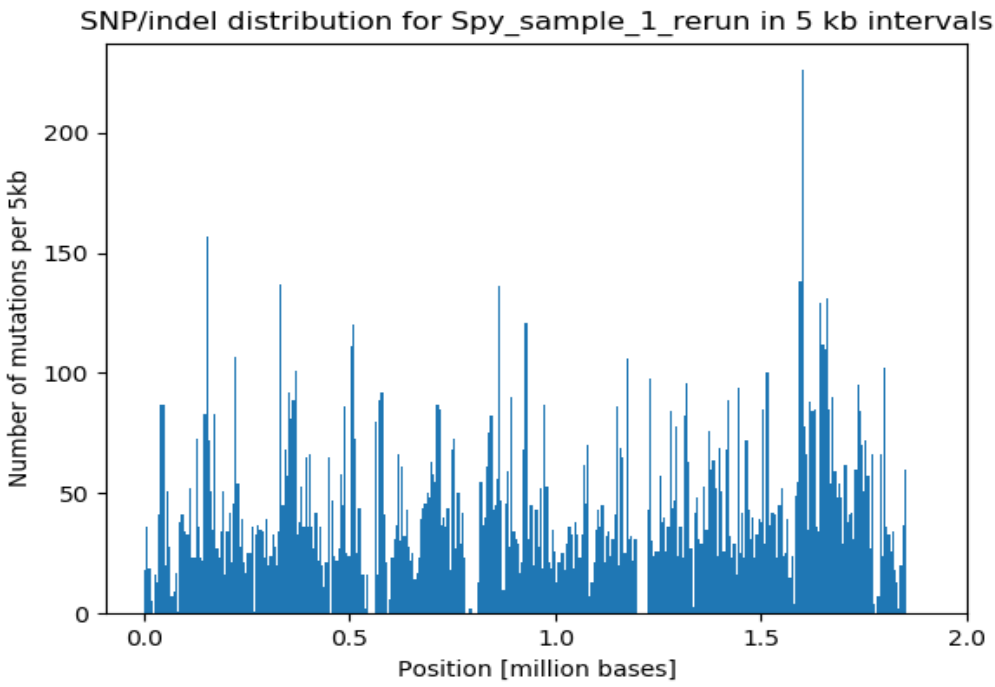


Figure: SNP/INDEL distribution

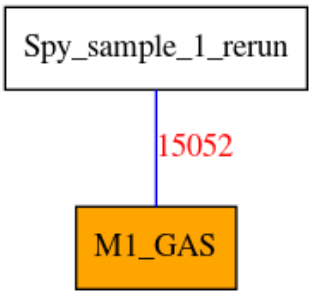


Figure: Minimum Spanning tree (ME)

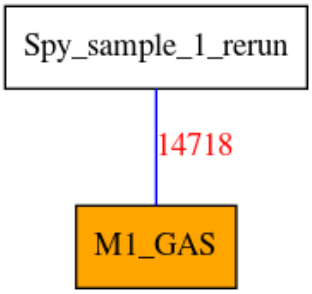


Figure: Minimum Spanning tree (SNP)

Phylogentic analysis of the core genome (Parsnp):

No phylogenetic tree is made if: a) there are less than three isolates in a cluster b) an isolate did not pass the QC check for new references

No tree since Parsnp needs more than two isolates.