REPORT

LegioCluster version: 24 September 2020
Date submitted: 2020-10-01
Submitted by: WH
Isolate name: Spy_sample_2_rerun

New name or pipeline will quit.

Species: Spy
Forward reads: /projdata/WH_PL/Github/LegioCluster/reads/Spy/Spy_sample_2_R1_001.fastq.gz
Reverse reads: /projdata/WH_PL/Github/LegioCluster/reads/Spy/Spy_sample_2_R2_001.fastq.gz
Metadata: make_ref Tutorial_part_3:_individual_submissions make_new_cluster
Folder name: WH201001_184100

Forces the pipeline to add this isolate to the list of candidate references and to generate a new cluster.

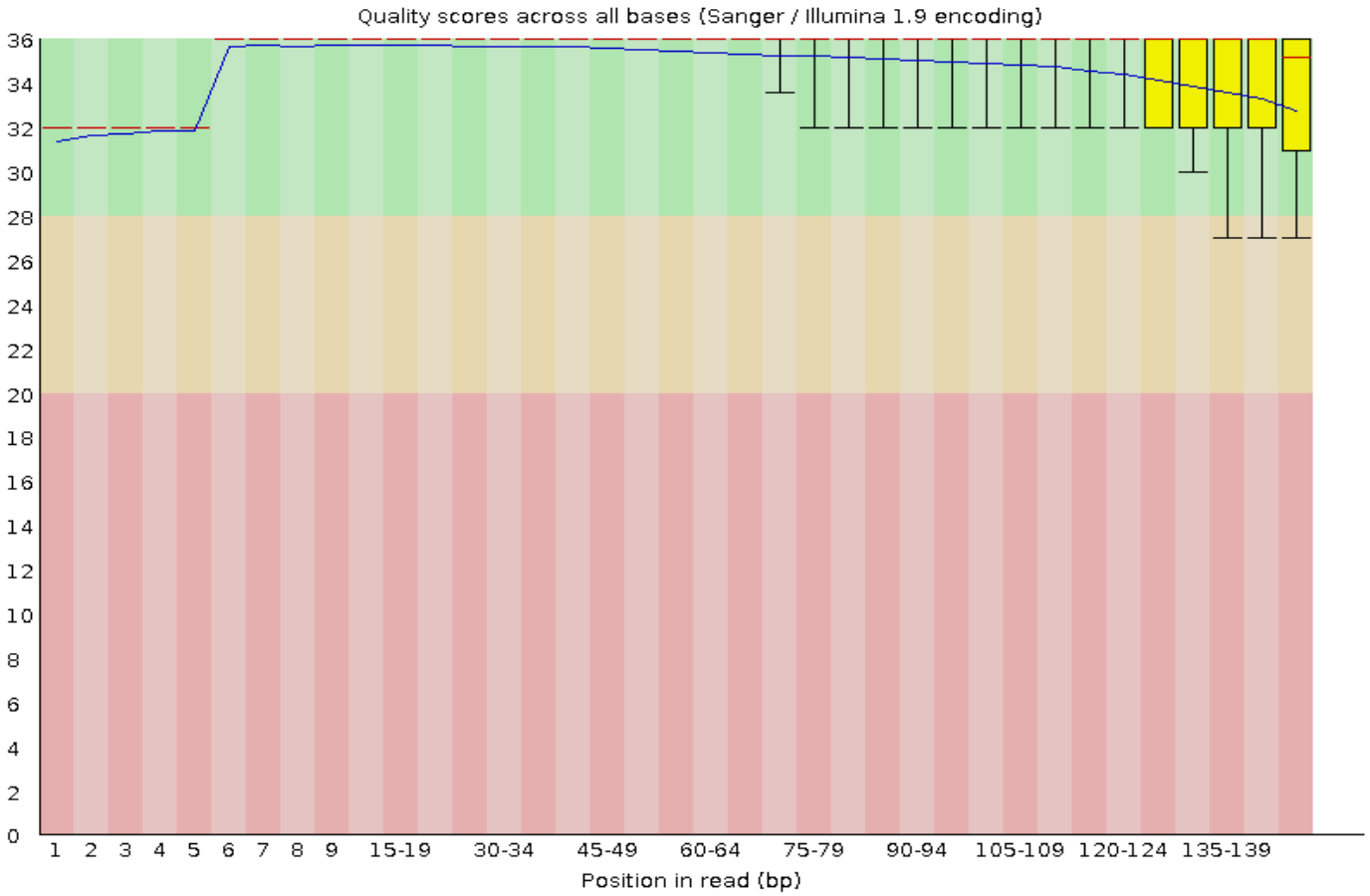
Read pre-processing (Trimmomatic):

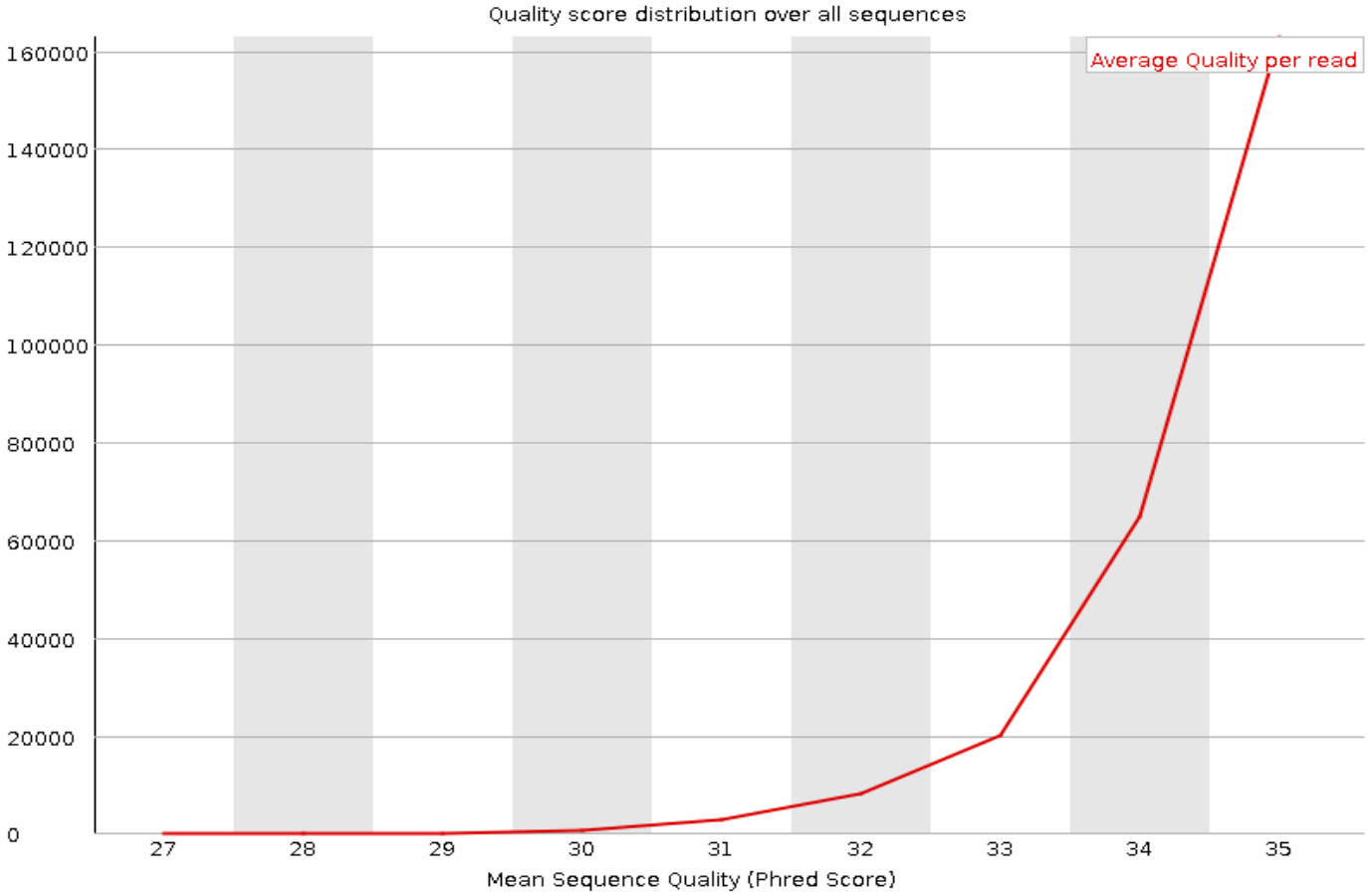
Adapters removed, low quality (< Q20) regions removed, short reads (<100) removed, ploy-G (>25) removed
Input read pairs: 299729
Both surviving: 260446 (86.89%)
Forward only surviving: 12102 (4.04%)
Reverse only surviving: 6819 (2.28%)
Dropped read pairs: 20362 (6.79%)
Mean (SD) lengths of trimmed F reads: 133.34 (42.694)
Mean (SD) lengths of trimmed R reads: 130.56 (46.079)
Mean (SD) no. of bases trimmed from 5' of F reads(*): 0.0 (0.012)
Mean (SD) no. of bases trimmed from 5' of R reads(*): 0.0 (0.102)
Mean (SD) no. of bases trimmed from 3' of F reads(*): 1.21 (5.607)
Mean (SD) no. of bases trimmed from 3' of R reads(*): 1.42 (6.001)
(*) if trimmed read length > 0

Read quality control (FastQC results):

Results for processed reads from:
/projdata/WH_PL/Github/LegioCluster/reads/Spy/Spy_sample_2_R1_001.fastq.gz
Filename paired_reads_1.fq
Total Sequences 260446
Sequences flagged as poor quality 0
Sequence length 100-149
%GC 38
PASS Basic Statistics
PASS Per base sequence quality

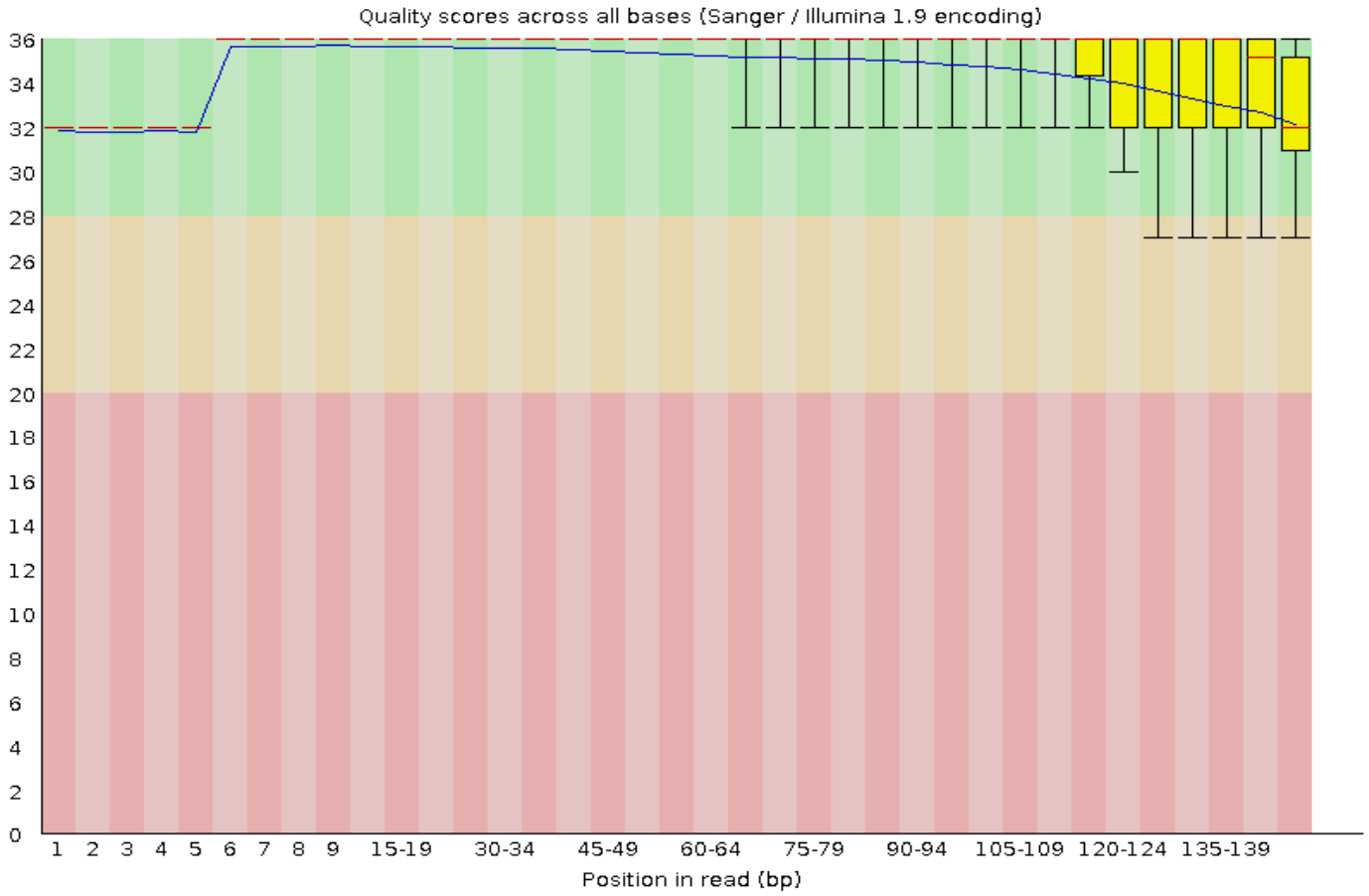
PASS Per tile sequence quality
PASS Per sequence quality scores
FAIL Per base sequence content
PASS Per sequence GC content
PASS Per base N content
WARN Sequence Length Distribution
PASS Sequence Duplication Levels
PASS Overrepresented sequences
PASS Adapter Content

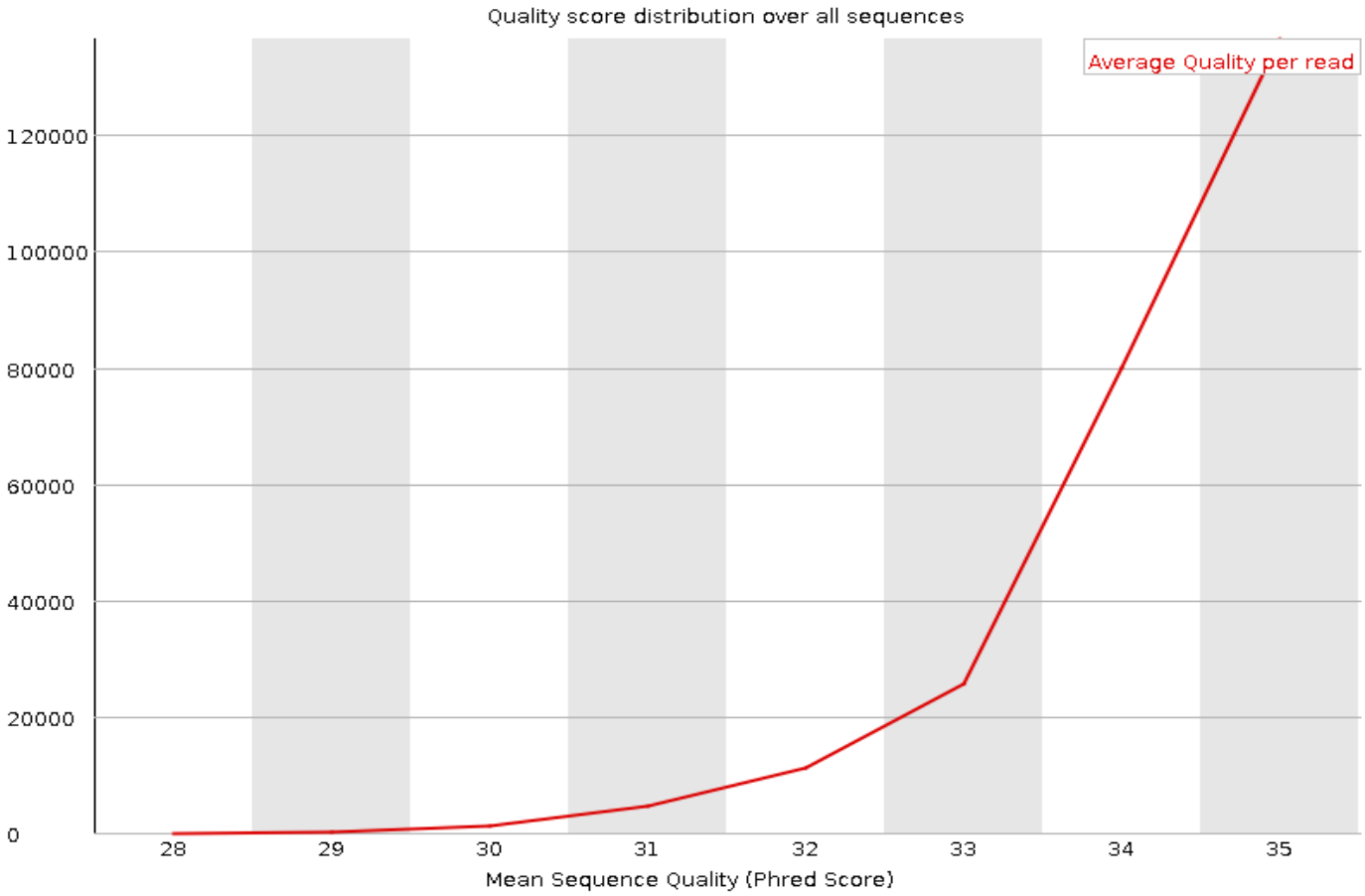




Read quality control (FastQC results):

```
Results for processed reads from:
/projdata/WH_PL/Github/LegioCluster/reads/Spy/Spy_sample_2_R2_001.fastq.gz
Filename paired_reads_2.fq
Total Sequences 260446
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
```


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 105-109 120-124 135-139
Position in read (bp)



Coverage: (260446 * 149 * 2) / 1831320 = 42.381

Percentage of bases with quality score >= Q30 (260150.0 * 100) / 260446.0 = 99.886

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: 0.000280715
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	656144	18.869692
2	176547	22.969326
3	162950	20.497345
4	156015	19.121696
5	102408	23.395345
6	89645	21.121740
7	79962	22.720911
8	77000	21.700311
9	52008	25.661609
10	51315	25.729439
11	42638	25.704918
12	32669	18.584990
13	22268	20.861340
14	5196	135.747998
15	1316	140.034705
16	808	43.979480
17	578	37.522954
18	268	19.753927
19	230	20.764706
20	183	17.132075
21	172	49.936842
22	168	96.527473
23	162	29.176471
24	159	19.414634
25	158	20.728395
26	155	19.128205
27	155	18.051282
28	155	15.230769
29	140	3.777778

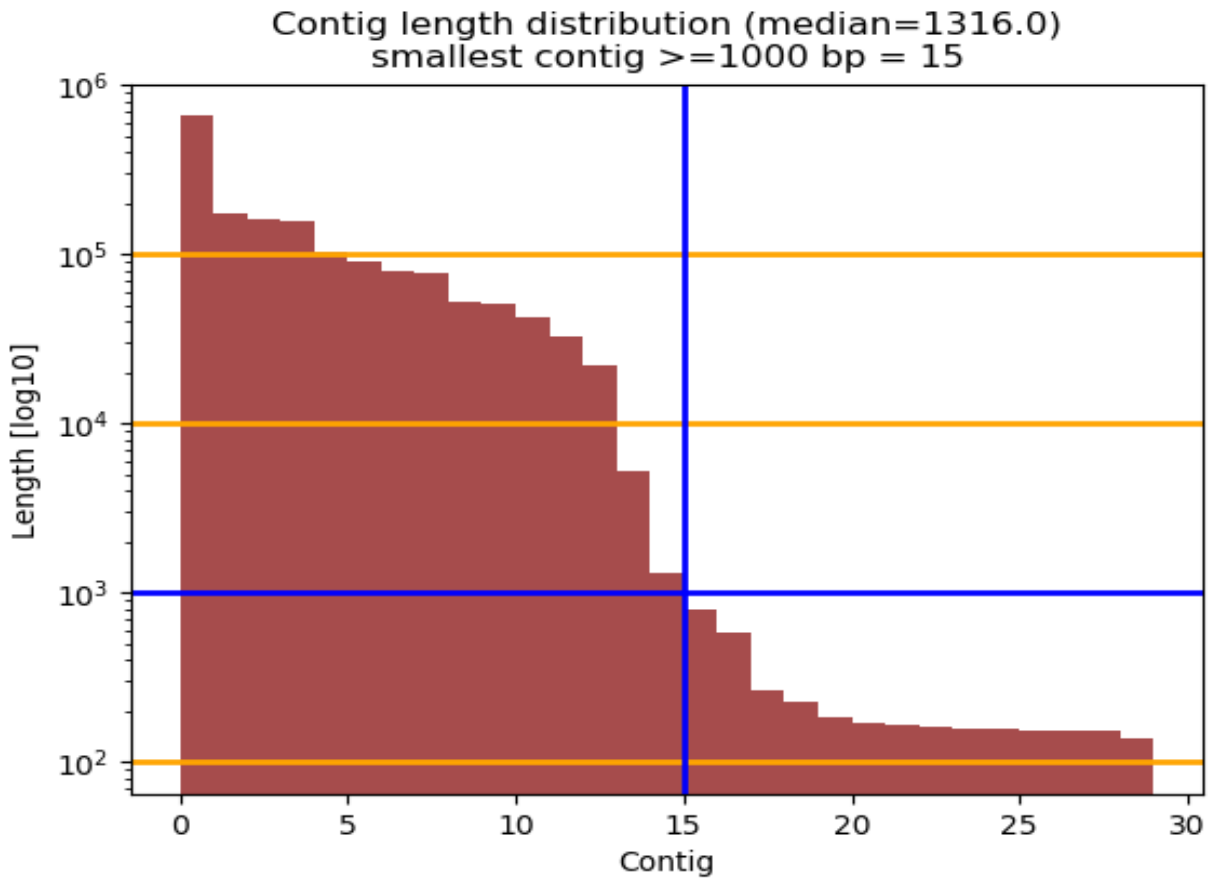


Figure: contigs vs length

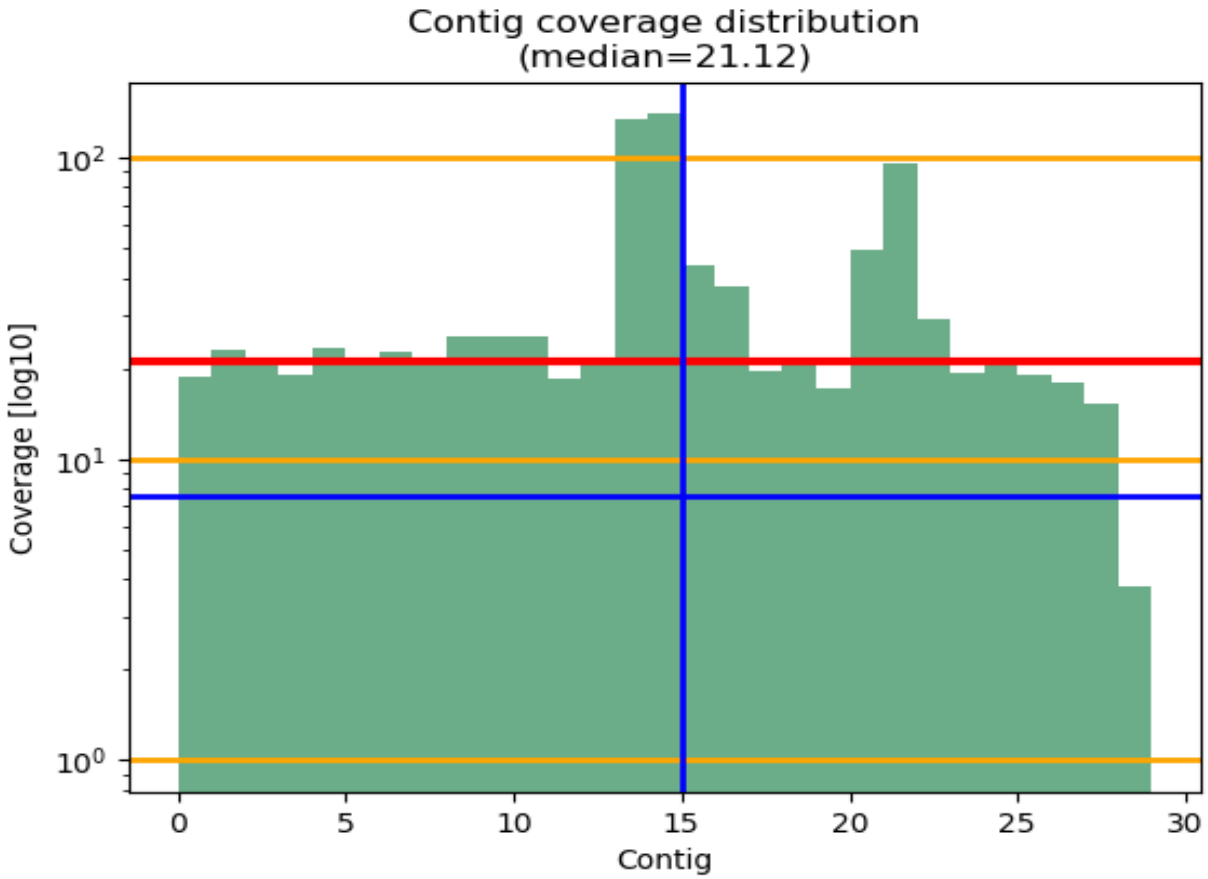


Figure: contigs vs coverage

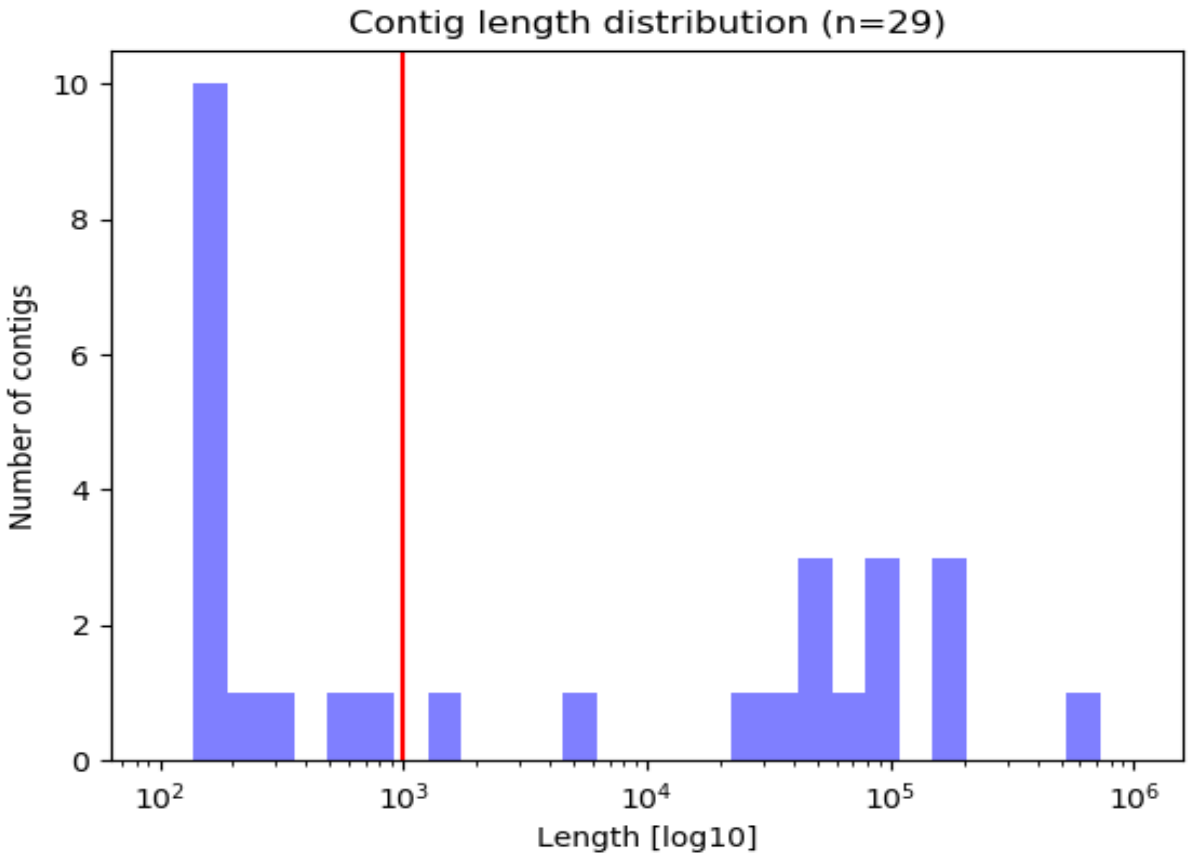


Figure: contig length distribution

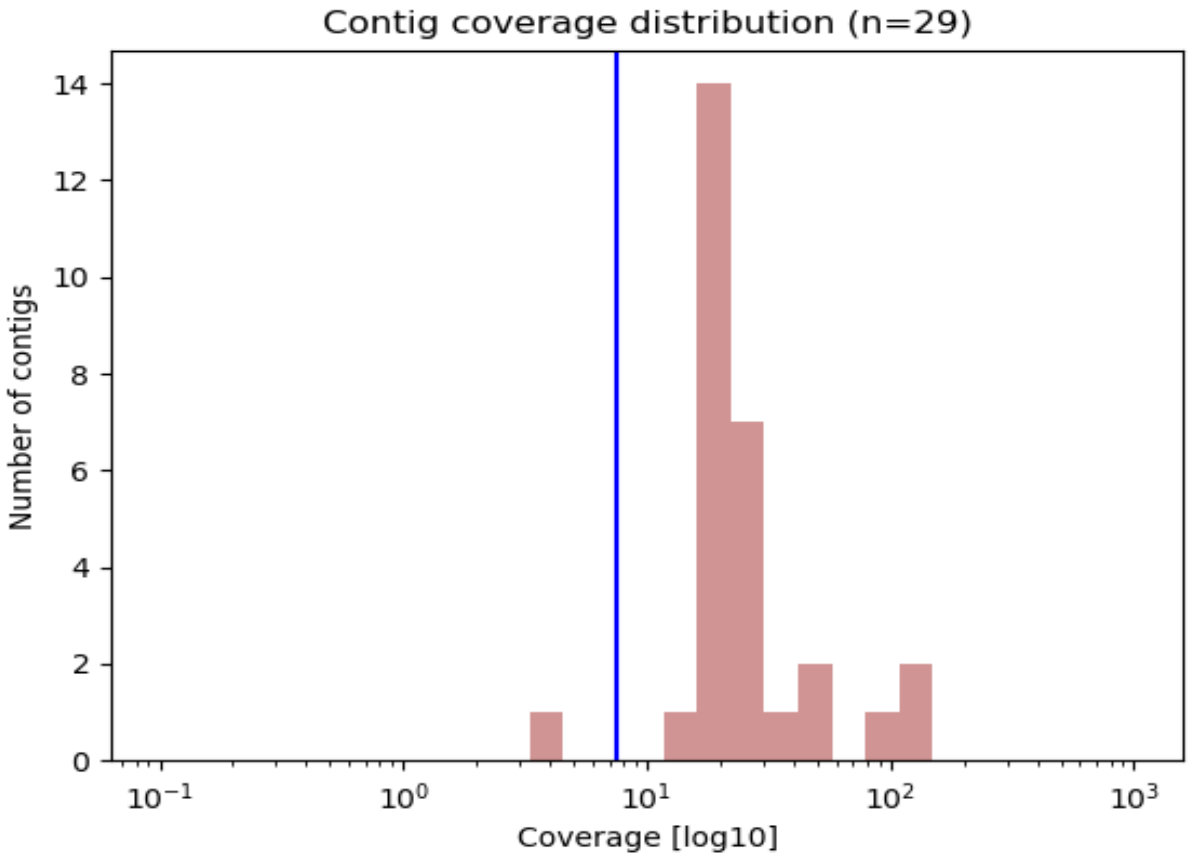


Figure: contig coverage distribution

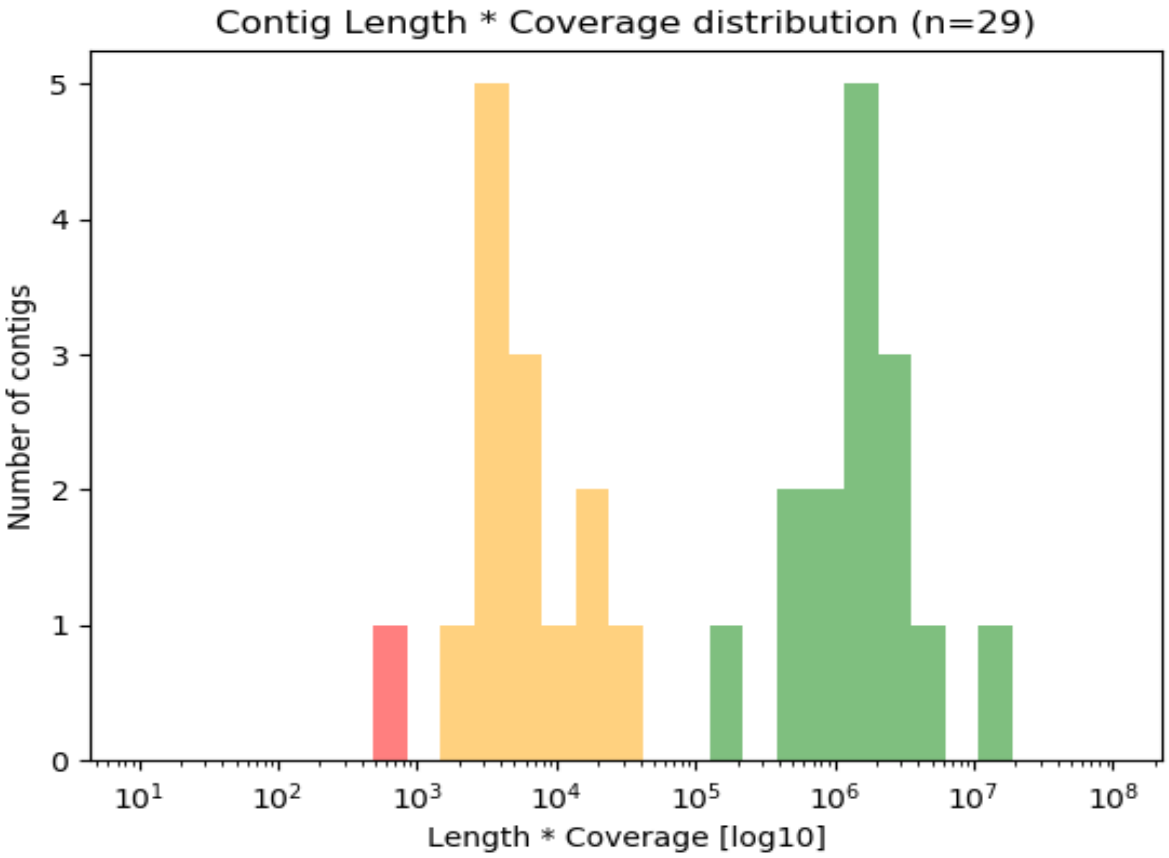


Figure: contig length * coverage distribution

Contig analysis:

(min length: 1000 bp, min coverage: 7.5x)
contigs that fail both thresholds: 3.45 %
contigs that are too short or have a low coverage: 44.83 %
contigs that meet both thresholds: 51.72 %
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: 4.76906e-05
P-value: 0.0
Matching hashes: 998/1000

Tutorial Part 2 showed that there is very little difference between Spy_sample_1 and Spy_sample_2, so this is the obvious choice for a reference.

```
Runner up
Strain name: M1_GAS
Mash distance: 0.0110882
P-value: 0.0
Matching hashes: 656/1000

Mash QC results: PASSED QC
```

Mapping the query against strain Spy_sample_1.fa (BWA MEM):

Even though Spy_sample_2_rerun is designated as a new candidate reference / cluster, it will first be mapped to the best matching reference for comparison.

Alignment QC (Samtools flagstat):

```
521123 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
231 + 0 supplementary
8704 + 0 duplicates
520968 + 0 mapped (99.97% : N/A)
520892 + 0 paired in sequencing
260446 + 0 read1
260446 + 0 read2
517884 + 0 properly paired (99.42% : N/A)
520608 + 0 with itself and mate mapped
129 + 0 singletons (0.02% : N/A)
2600 + 0 with mate mapped to a different chr
2553 + 0 with mate mapped to a different chr (mapQ>=5)
```

```
Percentage of mapped reads: 99.97
```

NOTE:

The threshold has been changed to 100% to ensure that the query isolate is added to the list of candidate references.

```
Percentage of mapped reads below threshold.
Adding the isolate to the list of candidate reference genomes.
```

Alignment QC (Samtools idxstats):

```
ref_fa_file len mapped unmapped
NODE_1_length_656942_cov_18.186882 656942 178239 0
```

```
NODE_2_length_176609_cov_21.963939 176609 58354 16
NODE_3_length_162950_cov_19.848551 162950 47923 0
NODE_4_length_156517_cov_18.693863 156517 43102 0
NODE_5_length_112615_cov_20.321038 112615 34011 3
NODE_6_length_102408_cov_22.289062 102408 34332 0
NODE_7_length_79962_cov_21.811792 79962 26069 0
NODE_8_length_76995_cov_21.152799 76995 23971 0
NODE_9_length_52008_cov_22.033390 52008 19060 0
NODE_10_length_50951_cov_23.933188 50951 18827 45
NODE_11_length_42638_cov_23.335800 42638 15690 0
NODE_12_length_32669_cov_19.175933 32669 8790 2
NODE_13_length_5196_cov_127.540926 5196 10099 63
NODE_14_length_1316_cov_129.384988 1316 2501 0
* 0 0 26
```

Genomic fragments:

```
Smallest fragment: 0
Mean length: 395.06
S.D.: 371.32
median: 399.0
Largest fragment: 170465
```

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) 29
# contigs (>= 1000 bp) 15
# contigs (>= 5000 bp) 14
# contigs (>= 10000 bp) 13
# contigs (>= 25000 bp) 12
# contigs (>= 50000 bp) 10
Total length (>= 0 bp) 1711572
Total length (>= 1000 bp) 1708081
Total length (>= 5000 bp) 1706765
Total length (>= 10000 bp) 1701569
Total length (>= 25000 bp) 1679301
Total length (>= 50000 bp) 1603994
# contigs 17
```

```
Largest contig 656144
Total length 1709467
Reference length 1709776
Reference GC (%) 38.38
N50 162950
NG50 162950
N75 89645
NG75 89645
L50 3
LG50 3
L75 6
LG75 6
# misassemblies 0
# misassembled contigs 0
Misassembled contigs length 0
# local misassemblies 0
# scaffold gap ext. mis. 0
# scaffold gap loc. mis. 0
# unaligned mis. contigs 0
# unaligned contigs 0 + 0 part
Unaligned length 0
Genome fraction (%) 99.959
Duplication ratio 1.000
# N's per 100 kbp 0.00
# mismatches per 100 kbp 0.06
# indels per 100 kbp 0.29
Largest alignment 656144
Total aligned length 1709098
NA50 162950
NGA50 162950
NA75 89645
NGA75 89645
LA50 3
LGA50 3
LA75 6
LGA75 6
```

Alignment QC (Samtools depth):

```
Total number of bases: 1709776
Number (percent) of bases with read depth < 1: 1 (0.0%)
Number (percent) of bases with read depth >= 1: 1709775 (100.0%)
Average read depth (S.D.): 43.93 (16.757)
Average read depth (S.D., count) for bases with read depth >= 1: 43.93 (16.76, 1709775)
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, 1)
```

```
Number of gaps >= 100 bases: 0
List of gaps >= 100 bases: []
Total number of bases in gaps >= 100 bases: 0
```

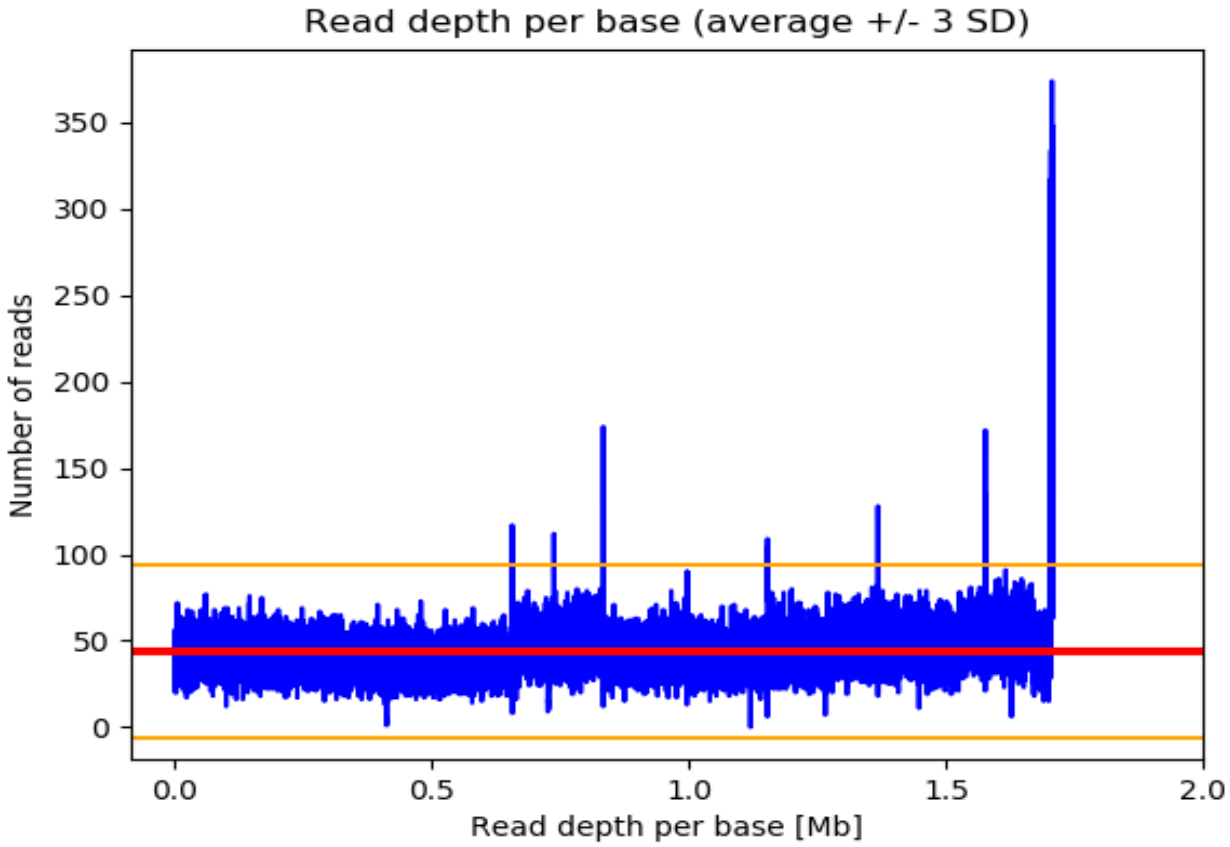


Figure: Read depth per base_1 (plot)

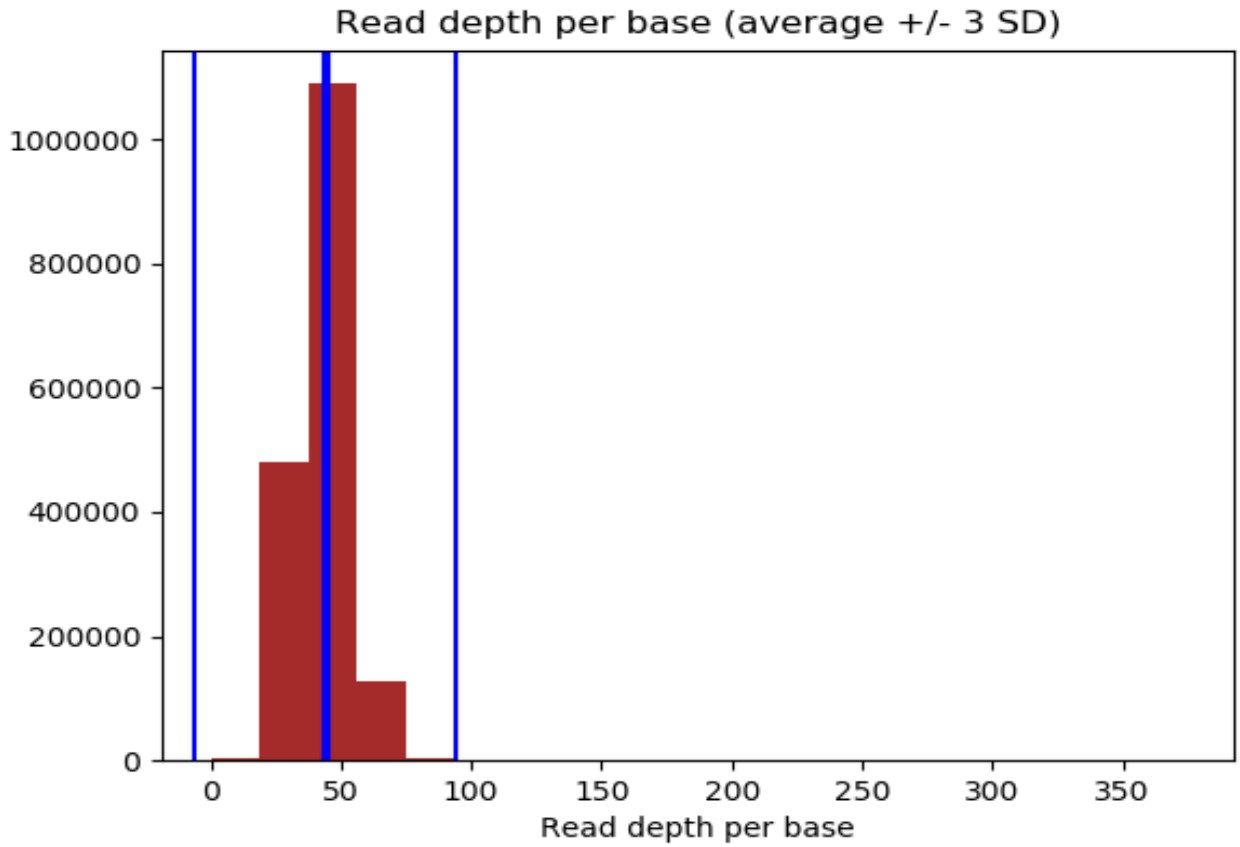


Figure: Read depth per base_1 (histogram)

Mapping quality check (Qualimap results):

```
number of bases = 1,709,776 bp
number of contigs = 14
number of reads = 521,123
number of mapped reads = 520,968 (99.97%)
number of mapped bases = 76,383,635 bp
mean mapping quality = 59.9317
```

SNPs and INDEL events between Spy_sample_2_rerun and reference Spy_sample_1 (FreeBayes):

```
Found 4 (3, 7, 1) SNPs and INDEL events compared to a reference genome of 1709776 bp.
(Note that the indel event count might be slightly lower in the SNP-matrix.)
```

NOTE: The threshold has been changed to zero to ensure that the query isolate is added to the list of candidate references.

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

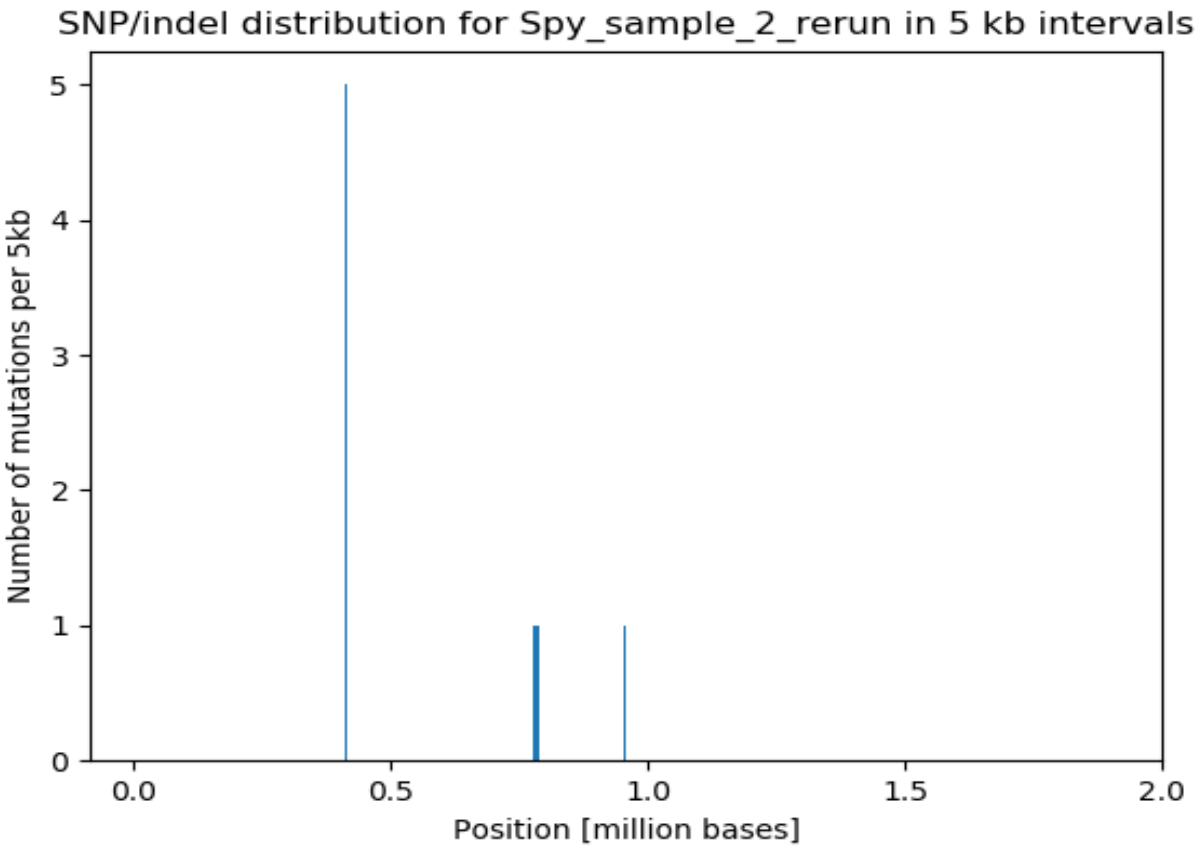
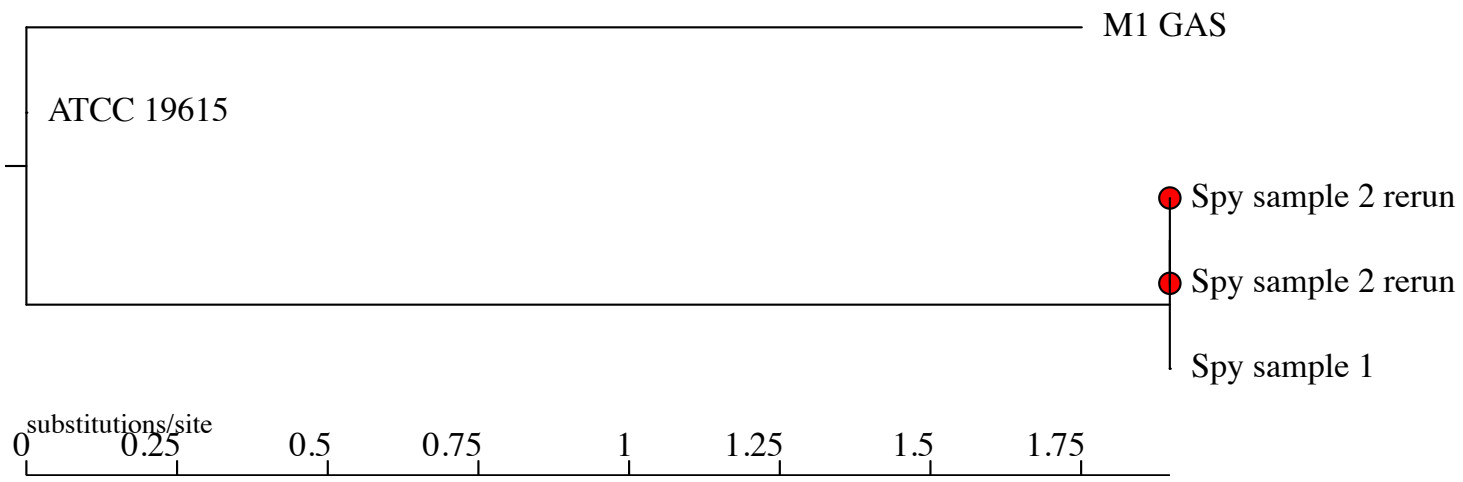


Figure: SNP/INDEL distribution

Note:

The isolate passed the QC check for new references.
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):


M1 GAS
ATCC 19615
Spy sample 2 rerun
Spy sample 2 rerun
Spy sample 1
substitutions/site
0
0.25
0.5
0.75
1
1.25
1.5
1.75