GENOME ASSEMBLY

velveth Velvet_output_Contig 31 -shortPaired -separate -fastq SR1.fastq SR2.fastq
velvetg Velvet_output_Contig -cov_cutoff 5

```
fastq
[0.000000] Reading FastQ file SR1.fastq;
[0.000041] Reading FastQ file SR2.fastq;
[5.997869] 2000000 sequences found in total in the paired sequence files
[5.997877] Done
[5.997931] Reading read set file Velvet_output_Contig/Sequences;
[6.282419] 2000000 sequences found
[7.662561] Done
[7.662572] 2000000 sequences in total.
[7.662640] Writing into roadmap file Velvet_output_Contig/Roadmaps...
[10.371306] Inputting sequences...
[10.371320] Inputting sequence 0 / 2000000
[21.875244] Inputting sequence 1000000 / 2000000
[34.844348]  === Sequences loaded in 24.473044 s
[34.844369] Done inputting sequences
[34.844372] Destroying splay table
[34.928982] Splay table destroyed
```

```
[0.000000] Reading roadmap file Velvet_output_Contig/Roadmaps
[4.574656] 2000000 roadmaps read
[4.576076] Creating insertion markers
[5.240727] Ordering insertion markers
[7.994133] Counting preNodes
[8.280513] 3391182 preNodes counted, creating them now
[11.340251] Sequence 1000000 / 2000000
[13.638452] Sequence 2000000 / 2000000
[13.638488] Adjusting marker info...
[13.957570] Connecting preNodes
[16.091905] Connecting 1000000 / 2000000
[18.683149] Connecting 2000000 / 2000000
```

quast.py contigs.fa -o quast_out

```
Version: 5.3.0

System information:
  OS: Linux-6.14.0-29-generic-x86_64-with-glibc2.39 (linux_64)
  Python version: 3.10.18
  CPUs number: 4

Started: 2025-10-16 21:00:31

Logging to /home/ibab/SEM3/RL/101625/assembly/quast_out/quast.log
NOTICE: Maximum number of threads is set to 1 (use --threads option to set it manually)

CWD: /home/ibab/SEM3/RL/101625/assembly
Main parameters:
  MODE: default, threads: 1, min contig length: 500, min alignment length: 65, min alignment IDY: 95.0, \
  ambiguity: one, min local misassembly length: 200, min extensive misassembly length: 1000

Contigs:
  Pre-processing...
  contigs.fa ==> contigs

2025-10-16 21:00:37
Running Basic statistics processor...
  Contig files:
    contigs
  Calculating N50 and L50...
    contigs, N50 = 12507, L50 = 150, auN = 14662.9, Total length = 6004592, GC % = 42.97, # N's per 100 kbp =  0.00
  Drawing Nx plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_out/basic_stats/Nx_plot.pdf
  Drawing cumulative plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_out/basic_stats/cumulative_plot.pdf
  Drawing GC content plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_out/basic_stats/GC_content_plot.pdf
  Drawing contigs GC content plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_out/basic_stats/contigs_GC_content_plot.pdf
  Drawing Coverage histogram (bin size: 1x)...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_out/basic_stats/coverage_histogram.pdf
```

```
/home/ibab/miniconda3/bin/quast.py contigs.fa -o quast_out

Version: 5.3.0

System information:
  OS: Linux-6.14.0-29-generic-x86_64-with-glibc2.39 (linux_64)
  Python version: 3.10.18
  CPUs number: 4

Started: 2025-10-16 21:00:31

Logging to /home/ibab/SEM3/RL/101625/assembly/quast_out/quast.log
NOTICE: Maximum number of threads is set to 1 (use --threads option to set it manually)

CWD: /home/ibab/SEM3/RL/101625/assembly
Main parameters:
  MODE: default, threads: 1, min contig length: 500, min alignment length: 65, min alignment IDY: 95.0, \
  ambiguity: one, min local misassembly length: 200, min extensive misassembly length: 1000

Contigs:
  Pre-processing...
  contigs.fa ==> contigs

2025-10-16 21:00:37
Running Basic statistics processor...
  Contig files:
    contigs
  Calculating N50 and L50...
    contigs, N50 = 12507, L50 = 150, auN = 14662.9, Total length = 6004592, GC % = 42.97, # N's per 100 kbp =  0.00
  Drawing Nx plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_out/basic_stats/Nx_plot.pdf
  Drawing cumulative plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_out/basic_stats/cumulative_plot.pdf
  Drawing GC content plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_out/basic_stats/GC_content_plot.pdf
  Drawing contigs GC content plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_out/basic_stats/contigs_GC_content_plot.pdf
  Drawing Coverage histogram (bin size: 1x)...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_out/basic_stats/coverage_histogram.pdf
```

QUAST

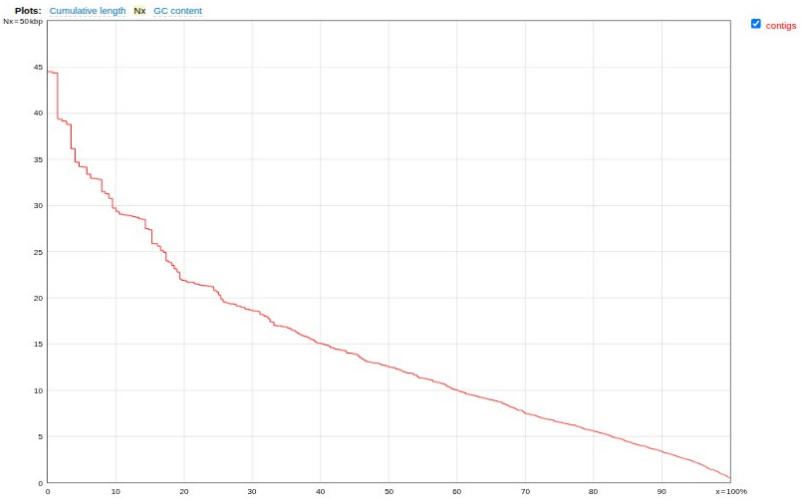
Quality Assessment Tool for Genome Assemblies

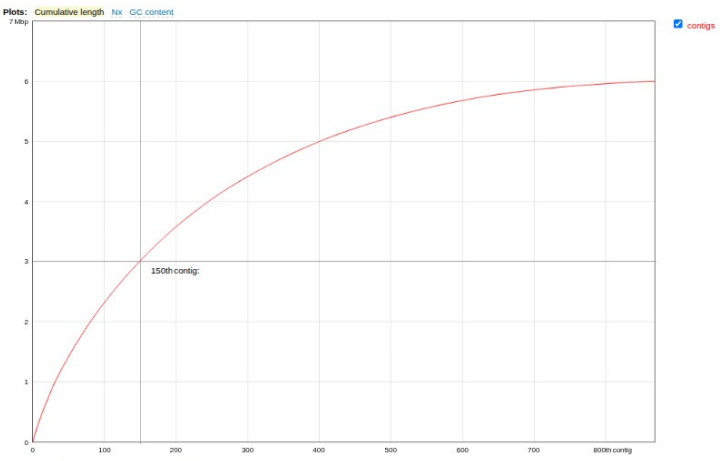
16 October 2025, Thursday, 21:00:40

View in Icarus contig browser

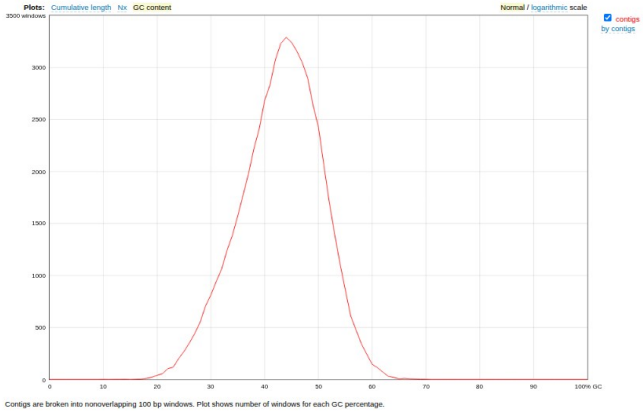
All statistics are based on contigs of size >= 500 bp, unless otherwise noted

Statistics without reference	contigs
# contigs	868
# contigs (>= 0 bp)	1645
# contigs (>= 1000 bp)	750
# contigs (>= 5000 bp)	392
# contigs (>= 10000 bp)	202
# contigs (>= 25000 bp)	32
# contigs (>= 50000 bp)	0
Largest contig	44 501
Total length	6 004 592
Total length (>= 0 bp)	6 121 977
Total length (>= 1000 bp)	5 918 868
Total length (>= 5000 bp)	4 960 526
Total length (>= 10000 bp)	3 601 097
Total length (>= 25000 bp)	1 018 216
Total length (>= 50000 bp)	0
N50	12 507
N90	3364
auN	14 663
L50	150
L90	500
GC (%)	42.97
Per base quality	
# N's per 100 kbp	0
# N's	0





Contigs are ordered from largest (contig #1) to smallest.



Contigs are broken into nonoverlapping 100 bp windows. Plot shows number of windows for each GC percentage.

augustus --species=E_coli_K12 contigs.fa > Augustus_out/contig_1.gff

gffread -g Velvet_output_Contig/contigs.fa -x Augustus_out/output_gene.fa Augustus_out/contig_1.gff

gffread -g Velvet_output_Contig/contigs.fa -y Augustus_out/output_protein.fa Augustus_out/contig_1.gff

```
NODE_1285_length_8726_cov_34.961380     AUGUSTUS        transcript      3689    5242    1       -       .       g2936.t1
Warning: invalid GTF record, transcript_id not found:
NODE_1285_length_8726_cov_34.961380     AUGUSTUS        transcript      5325    5891    0.95    -       .       g2937.t1
Warning: invalid GTF record, transcript_id not found:
NODE_1285_length_8726_cov_34.961380     AUGUSTUS        transcript      6074    7543    0.99    -       .       g2938.t1
Warning: invalid GTF record, transcript_id not found:
NODE_1285_length_8726_cov_34.961380     AUGUSTUS        transcript      7568    8497    0.97    -       .       g2939.t1
Warning: invalid GTF record, transcript_id not found:
NODE_1286_length_1350_cov_31.194075     AUGUSTUS        transcript      52      444     0.97    +       .       g2940.t1
Warning: invalid GTF record, transcript_id not found:
NODE_1294_length_2005_cov_33.484787     AUGUSTUS        transcript      942     1850    1       -       .       g2941.t1
Warning: invalid GTF record, transcript_id not found:
NODE_1301_length_12419_cov_30.485144    AUGUSTUS        transcript      121     2016    0.52    +       .       g2942.t1
Warning: invalid GTF record, transcript_id not found:
NODE_1301_length_12419_cov_30.485144    AUGUSTUS        transcript      2054    4327    0.96    +       .       g2943.t1
Warning: invalid GTF record, transcript_id not found:
```

busco -i Velvet_output_Contig/contigs.fa -m genome -l bacteria_odb10 -o busco_bacteria_results --cpu 4

```
2025-10-16 21:15:27 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:15:27 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:15:27 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:15:27 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:15:28 INFO:       Results:        C:95.1%[S:91.9%,D:3.2%],F:4.8%,M:0.1%,n:124

2025-10-16 21:15:28 INFO:

        ---------------------------------------------------
        |Results from dataset bacteria_odb10              |
        ---------------------------------------------------
        |C:95.1%[S:91.9%,D:3.2%],F:4.8%,M:0.1%,n:124      |
        |118    Complete BUSCOs (C)                       |
        |114    Complete and single-copy BUSCOs (S)       |
        |4      Complete and duplicated BUSCOs (D)        |
        |6      Fragmented BUSCOs (F)                     |
        |0      Missing BUSCOs (M)                        |
        |124    Total BUSCO groups searched               |
        ---------------------------------------------------
2025-10-16 21:15:28 INFO:       BUSCO analysis done. Total running time: 250 seconds
2025-10-16 21:15:28 INFO:       Results written in /home/ibab/SEM3/RL/101625/assembly/busco_bacteria_results
2025-10-16 21:15:28 INFO:       For assistance with interpreting the results, please consult the userguide: https://busco.ezlab.org/busco_userguide.html

2025-10-16 21:15:28 INFO:       Visit this page https://gitlab.com/ezlab/busco#how-to-cite-busco to see how to cite BUSCO
```

```
        ---------------------------------------------------
        |Results from dataset bacteria_odb10
        ---------------------------------------------------
        |C:95.1%[S:91.9%,D:3.2%],F:4.8%,M:0.1%,n:124
        |118    Complete BUSCOs (C)
        |114    Complete and single-copy BUSCOs (S)
        |4      Complete and duplicated BUSCOs (D)
        |6      Fragmented BUSCOs (F)
        |0      Missing BUSCOs (M)
        |124    Total BUSCO groups searched
        ---------------------------------------------------
```

ragtag.py correct GCA_014131755.1_ASM1413175v1_genomic.fna contigs.fa

```
(base) ibab@IBAB-MSc14-Comp011:~/SEM3/RL/101625/assembly$ ragtag.py correct GCA_014131755.1_ASM1413175v1_genomic.fna contigs.fa
Thu Oct 16 21:17:27 2025 --- VERSION: RagTag v2.1.0
Thu Oct 16 21:17:27 2025 --- CMD: ragtag.py correct GCA_014131755.1_ASM1413175v1_genomic.fna contigs.fa
Thu Oct 16 21:17:27 2025 --- WARNING: Without '-u' invoked, some component/object AGP pairs might share the same ID. Some external pro
grams/databases don't like this. To ensure valid AGP format, use '-u'.
Thu Oct 16 21:17:27 2025 --- INFO: Mapping the query genome to the reference genome
Thu Oct 16 21:17:27 2025 --- INFO: Running: minimap2 -x asm5 -t 1 /home/ibab/SEM3/RL/101625/assembly/GCA_014131755.1_ASM1413175v1_geno
mic.fna /home/ibab/SEM3/RL/101625/assembly/contigs.fa > /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.asm.paf 2> /ho
me/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.asm.paf.log
Thu Oct 16 21:17:27 2025 --- INFO: Finished running : minimap2 -x asm5 -t 1 /home/ibab/SEM3/RL/101625/assembly/GCA_014131755.1_ASM1413
175v1_genomic.fna /home/ibab/SEM3/RL/101625/assembly/contigs.fa > /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.asm.
paf 2> /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.asm.paf.log
Thu Oct 16 21:17:27 2025 --- INFO: Reading whole genome alignments
Thu Oct 16 21:17:27 2025 --- INFO: Filtering and merging alignments
Thu Oct 16 21:17:28 2025 --- INFO: Writing: /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.agp
Thu Oct 16 21:17:28 2025 --- INFO: Writing broken contigs
Thu Oct 16 21:17:28 2025 --- INFO: Running: ragtag_break_query.py /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.agp
/home/ibab/SEM3/RL/101625/assembly/contigs.fa > /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.fasta 2> /home/ibab/SE
M3/RL/101625/assembly/ragtag_output/ragtag.correct.err
Thu Oct 16 21:17:28 2025 --- INFO: Finished running : ragtag_break_query.py /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.co
rrect.agp /home/ibab/SEM3/RL/101625/assembly/contigs.fa > /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.fasta 2> /ho
me/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.err
Thu Oct 16 21:17:28 2025 --- INFO: Goodbye
```

ragtag.py scaffold -u GCA_014131755.1_ASM1413175v1_genomic.fna ragtag_output/ragtag.correct.fasta

```
(base) ibab@IBAB-MSc14-Comp011:~/SEM3/RL/101625/assembly$ ragtag.py scaffold -u GCA_014131755.1_ASM1413175v1_genomic.fna ragtag_output
/ragtag.correct.fasta
Thu Oct 16 21:19:24 2025 --- VERSION: RagTag v2.1.0
Thu Oct 16 21:19:24 2025 --- CMD: ragtag.py scaffold -u GCA_014131755.1_ASM1413175v1_genomic.fna ragtag_output/ragtag.correct.fasta
Thu Oct 16 21:19:24 2025 --- INFO: Mapping the query genome to the reference genome
Thu Oct 16 21:19:24 2025 --- INFO: Running: minimap2 -x asm5 -t 1 /home/ibab/SEM3/RL/101625/assembly/GCA_014131755.1_ASM1413175v1_geno
mic.fna /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.fasta > /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragta
g.scaffold.asm.paf 2> /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.asm.paf.log
Thu Oct 16 21:19:25 2025 --- INFO: Finished running : minimap2 -x asm5 -t 1 /home/ibab/SEM3/RL/101625/assembly/GCA_014131755.1_ASM1413
175v1_genomic.fna /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.fasta > /home/ibab/SEM3/RL/101625/assembly/ragtag_ou
tput/ragtag.scaffold.asm.paf 2> /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.asm.paf.log
Thu Oct 16 21:19:25 2025 --- INFO: Reading whole genome alignments
Thu Oct 16 21:19:25 2025 --- INFO: Filtering and merging alignments
Thu Oct 16 21:19:25 2025 --- INFO: Ordering and orienting query sequences
Thu Oct 16 21:19:25 2025 --- INFO: Writing scaffolds
Thu Oct 16 21:19:25 2025 --- INFO: Writing: /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.agp
Thu Oct 16 21:19:25 2025 --- INFO: Running: ragtag_agp2fa.py /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.agp /hom
e/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.fasta > /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.f
asta 2> /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.err
Thu Oct 16 21:19:25 2025 --- INFO: Finished running : ragtag_agp2fa.py /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffol
d.agp /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.correct.fasta > /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.
scaffold.fasta 2> /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.err
Thu Oct 16 21:19:25 2025 --- INFO: Running: ragtag_stats.py /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.agp /home
/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.confidence.txt > /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.s
caffold.stats 2> /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.err
Thu Oct 16 21:19:25 2025 --- INFO: Finished running : ragtag_stats.py /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold
.agp /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.confidence.txt > /home/ibab/SEM3/RL/101625/assembly/ragtag_outpu
t/ragtag.scaffold.stats 2> /home/ibab/SEM3/RL/101625/assembly/ragtag_output/ragtag.scaffold.err
Thu Oct 16 21:19:25 2025 --- INFO: Goodbye
```

quast.py ragtag_output/ragtag.scaffold.fasta -o quast_ragtag_out

```
Started: 2025-10-16 21:20:36

Logging to /home/ibab/SEM3/RL/101625/assembly/quast_ragtag_out/quast.log
NOTICE: Maximum number of threads is set to 1 (use --threads option to set it manually)

CWD: /home/ibab/SEM3/RL/101625/assembly
Main parameters:
  MODE: default, threads: 1, min contig length: 500, min alignment length: 65, min alignment IDY: 95.0, \
  ambiguity: one, min local misassembly length: 200, min extensive misassembly length: 1000

Contigs:
  Pre-processing...
  ragtag_output/ragtag.scaffold.fasta ==> ragtag.scaffold

2025-10-16 21:20:37
Running Basic statistics processor...
  Contig files:
    ragtag.scaffold
  Calculating N50 and L50...
    ragtag.scaffold, N50 = 5810422, L50 = 1, auN = 5461183.0, Total length = 6182734, GC % = 42.95, # N's per 100 kbp =  1848.70
  Drawing Nx plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_ragtag_out/basic_stats/Nx_plot.pdf
  Drawing cumulative plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_ragtag_out/basic_stats/cumulative_plot.pdf
  Drawing GC content plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_ragtag_out/basic_stats/GC_content_plot.pdf
  Drawing ragtag.scaffold GC content plot...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_ragtag_out/basic_stats/ragtag.scaffold_GC_content_plot.pdf
  Drawing Coverage histogram (bin size: 3x)...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_ragtag_out/basic_stats/coverage_histogram.pdf
  Drawing ragtag.scaffold coverage histogram (bin size: 3x)...
    saved to /home/ibab/SEM3/RL/101625/assembly/quast_ragtag_out/basic_stats/ragtag-scaffold_coverage_histogram.pdf
```

QUAST

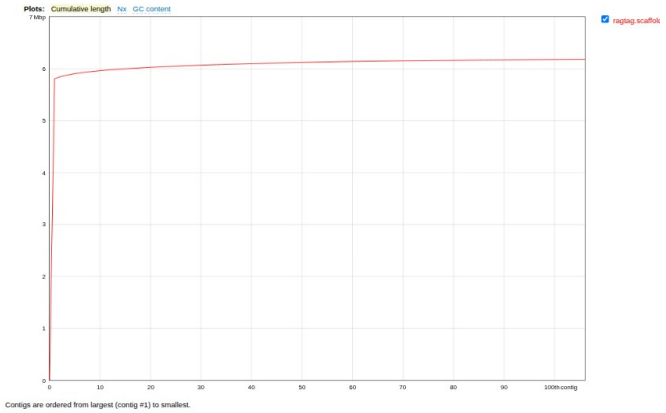
Quality Assessment Tool for Genome Assemblies

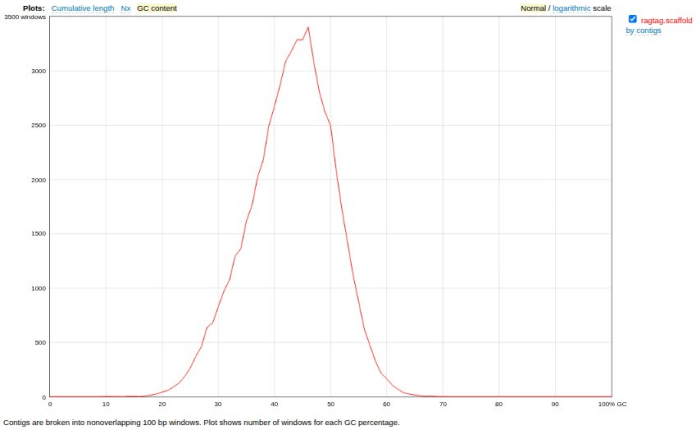
16 October 2025, Thursday, 21:20:39

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless other

Statistics without reference	ragtag.scaffold
# contigs	106
# contigs (>= 0 bp)	502
# contigs (>= 1000 bp)	74
# contigs (>= 5000 bp)	20
# contigs (>= 10000 bp)	8
# contigs (>= 25000 bp)	3
# contigs (>= 50000 bp)	1
Largest contig	5 810 422
Total length	6 182 734
Total length (>= 0 bp)	6 236 277
Total length (>= 1000 bp)	6 159 924
Total length (>= 5000 bp)	6 029 873
Total length (>= 10000 bp)	5 946 354
Total length (>= 25000 bp)	5 870 520
Total length (>= 50000 bp)	5 810 422
N50	5 810 422
N90	5 810 422
auN	5 461 183
L50	1
L90	1
GC (%)	42.95
Per base quality	
# N's per 100 kbp	1848.7
# N's	114 300






```
busco -i ragtag_output/ragtag.correct.fasta -o busco_ragtag_correct -m genome --cpu 4
```

```
2025-10-16 21:25:19 INFO:       [bbmap] 1 of 1 task(s) completed
2025-10-16 21:25:19 INFO:       ***** Run Prodigal on input to predict and extract genes *****
2025-10-16 21:25:19 INFO:       Genetic code 11 selected as optimal
2025-10-16 21:25:19 INFO:       ***** Run HMMER on gene sequences *****
2025-10-16 21:25:19 INFO:       Running 124 job(s) on hmmsearch, starting at 10/16/2025 21:25:19
/usr/bin/busco:13: SyntaxWarning: invalid escape sequence '\w'
  "cannot import name '(?P<module_name>[\w]+)", err.msg
/usr/bin/busco:13: SyntaxWarning: invalid escape sequence '\w'
  "cannot import name '(?P<module_name>[\w]+)", err.msg
/usr/bin/busco:13: SyntaxWarning: invalid escape sequence '\w'
  "cannot import name '(?P<module_name>[\w]+)", err.msg
/usr/bin/busco:13: SyntaxWarning: invalid escape sequence '\w'
  "cannot import name '(?P<module_name>[\w]+)", err.msg
2025-10-16 21:25:20 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:25:20 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:25:20 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:25:20 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:25:20 INFO:       [hmmsearch]     13 of 124 task(s) completed
2025-10-16 21:25:20 INFO:       [hmmsearch]     25 of 124 task(s) completed
2025-10-16 21:25:20 INFO:       [hmmsearch]     38 of 124 task(s) completed
2025-10-16 21:25:20 INFO:       [hmmsearch]     50 of 124 task(s) completed
2025-10-16 21:25:21 INFO:       [hmmsearch]     63 of 124 task(s) completed
2025-10-16 21:25:21 INFO:       [hmmsearch]     75 of 124 task(s) completed
2025-10-16 21:25:21 INFO:       [hmmsearch]     87 of 124 task(s) completed
2025-10-16 21:25:21 INFO:       [hmmsearch]     112 of 124 task(s) completed
2025-10-16 21:25:22 INFO:       [hmmsearch]     124 of 124 task(s) completed
/usr/bin/busco:13: SyntaxWarning: invalid escape sequence '\w'
  "cannot import name '(?P<module_name>[\w]+)", err.msg
/usr/bin/busco:13: SyntaxWarning: invalid escape sequence '\w'
  "cannot import name '(?P<module_name>[\w]+)", err.msg
/usr/bin/busco:13: SyntaxWarning: invalid escape sequence '\w'
  "cannot import name '(?P<module_name>[\w]+)", err.msg
/usr/bin/busco:13: SyntaxWarning: invalid escape sequence '\w'
  "cannot import name '(?P<module_name>[\w]+)", err.msg
2025-10-16 21:25:22 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:25:22 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:25:22 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:25:22 INFO:       NumExpr defaulting to 4 threads.
2025-10-16 21:25:22 INFO:       Results:        C:95.1%[S:91.9%,D:3.2%],F:4.8%,M:0.1%,n:124
```

C:95.1%[S:91.9%,D:3.2%],F:4.8%,M:0.1%,n:124

Assembly Type	Complete (C)	Single-Copy (S)	Duplicated (D)	Fragmented (F)	Missing (M)	Total BUSCOs (n)
Contigs	118 (95.1%)	114 (91.9%)	4 (3.2%)	6 (4.8%)	0 (0.1%)	124
Scaffolded	118 (95.1%)	114 (91.9%)	4 (3.2%)	6 (4.8%)	0 (0.1%)	124

 # 1. Genome Assembly with Velvet

 # Initialize Velvet assembly directory with k-mer size 31 using paired-end reads
velveth Velvet_output_Contig 31 -shortPaired -separate -fastq SR1.fastq SR2.fastq

 # Build the assembly graph; discard contigs with coverage <5
velvetg Velvet_output_Contig -cov_cutoff 5

 # 2. Assembly Quality Assessment

 # Evaluate assembly quality (N50, total length, contig count, etc.)
quast.py contigs.fa -o quast_out

 # 3. Gene Prediction with Augustus

 # Predict genes on assembled contigs using E. coli K12 model; output in GFF format
augustus --species=E_coli_K12 contigs.fa > Augustus_out/contig_1.gff

 # Extract predicted gene sequences in nucleotide FASTA format from GFF
gffread -g Velvet_output_Contig/contigs.fa -x Augustus_out/output_gene.fa Augustus_out/contig_1.gff

 # Extract predicted protein sequences from GFF

```
gffread -g Velvet_output_Contig/contigs.fa -y Augustus_out/output_protein.fa Augustus_out/contig_1.gff


# 4. Genome Completeness Assessment with BUSCO


# Evaluate completeness of assembled contigs using BUSCO with bacterial dataset
busco -i Velvet_output_Contig/contigs.fa -m genome -l bacteria_odb10 -o busco_bacteria_results --cpu 4

# Example BUSCO output interpretation:
# C:95.1%[S:91.9%,D:3.2%],F:4.8%,M:0.1%,n:124
# C = Complete BUSCOs, S = Single-copy, D = Duplicated, F = Fragmented, M = Missing, n = total BUSCO groups


# 5. Reference-Guided Scaffolding with RagTag


# Correct contigs using a reference genome
ragtag.py correct GCA_014131755.1_ASM1413175v1_genomic.fna contigs.fa

# Scaffold corrected contigs against the reference genome
ragtag.py scaffold -u GCA_014131755.1_ASM1413175v1_genomic.fna ragtag_output/ragtag.correct.fasta

# Assess quality of scaffolded assembly
quast.py ragtag_output/ragtag.scaffold.fasta -o quast_ragtag_out


# 6. BUSCO on Corrected Assembly


# Evaluate completeness of corrected (ragtag) assembly
busco -i ragtag_output/ragtag.correct.fasta -o busco_ragtag_correct -m genome --cpu 4
```