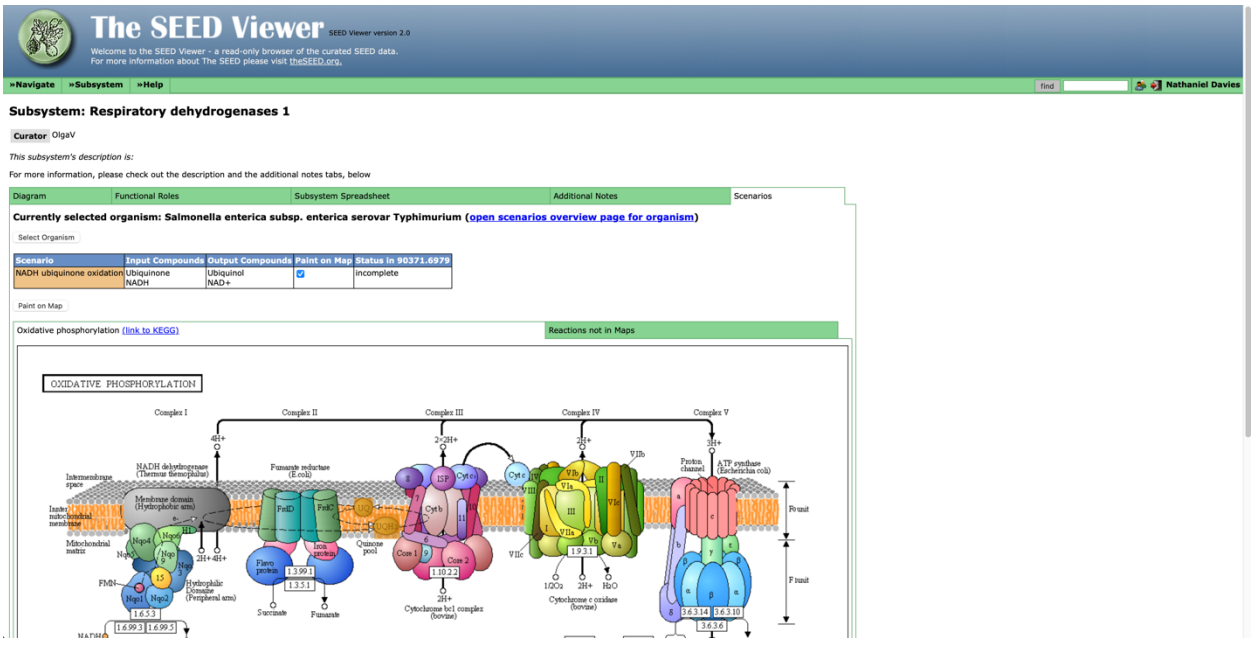
1.) Summarize the difference in outputs between ABySS and SPAdes

SPAdes and ABySS both assembled the genome of Salmonella successfully. However, the ABySS output was much more complicated to interpret than the SPAdes output. There were fifty-one items in my folder once the assembly had finished. Knowing which ones to use in further analysis is essential. In contrast, SPAdes only gave four files that were very easy to use. I think the interface of SPAdes is more user friendly than the interface of ABySS.

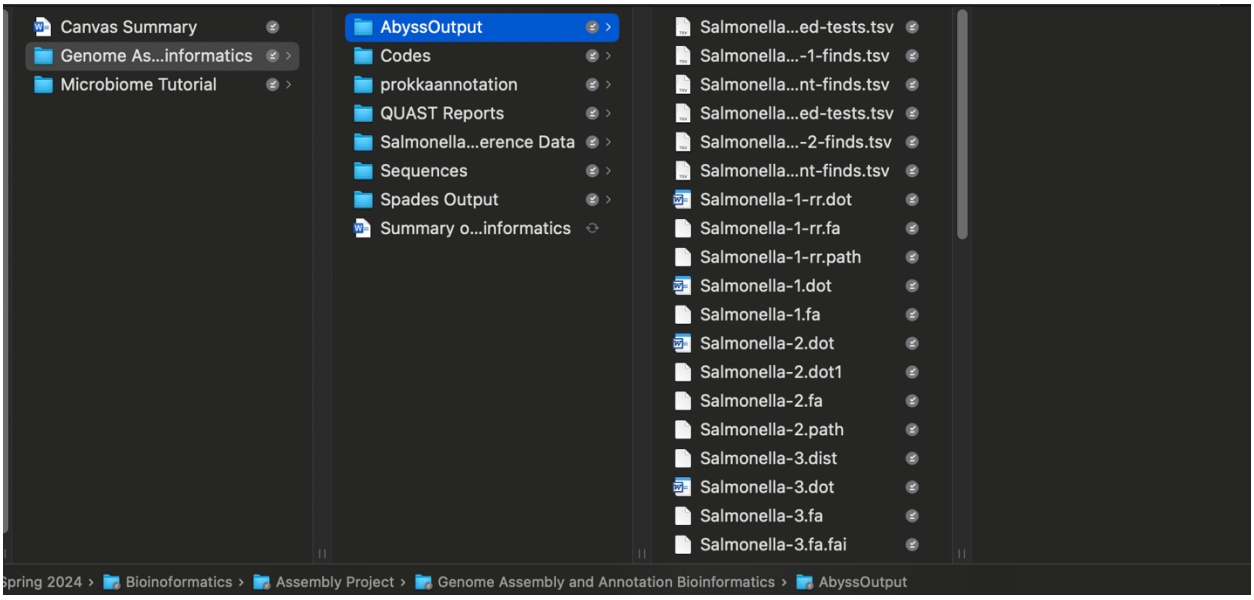
2.) Write a brief summary of the important output metrics from QUAST

Some of the important metrics from QUAST include the genome completeness, L_{50}, and N_{50} values. The genome completeness uses BUSCO (benchmarking universal single copy orthologs) which looks for genes that are common in all members of the taxon examined. By comparing the assembled genes to this set of common genes, BUSCO can determine the completeness of the assembled genome. The L_{50} is the number of contigs/scaffolds needed to have approximately 50% of the genome. A low L_{50} number corresponds to a higher quality genome because that means there are not a lot of little pieces that could be arranged in different ways. The N_{50} number means that 50% of contigs/scaffolds are this long. A higher N_{50} number corresponds to a higher quality genome, for similar reasons because it means there are a lot of big pieces, not a large quantity of small fragments. All fo these values from QUAST are important in determining the quality of an assembled genome.

Below is a picture of a pathway using the RAST Seed Viewer:



Below is a picture of my file organization:



Report

	Salmonella-8_fa	Salmonella-8_fa_broken
# contigs (>= 0 bp)	150	-
# contigs (>= 1000 bp)	44	50
Total length (>= 0 bp)	4871123	-
Total length (>= 1000 bp)	4851345	4849767
# contigs	48	58
Largest contig	537866	537866
Total length	4853853	4853353
Reference length	4951383	4951383
GC (%)	52.19	52.19
Reference GC (%)	52.24	52.24
N50	272520	223891
NG50	272520	220071
N90	59728	54580
NG90	54580	52399
auN	301881.7	245920.8
auNG	295935.4	241051.9
L50	6	7
LG50	6	8
L90	21	26
LG90	22	27
# misassemblies	10	10
# misassembled contigs	9	9
Misassembled contigs length	2096548	1436747
# local misassemblies	6	6
# scaffold gap ext. mis.	0	-
# scaffold gap loc. mis.	0	-
# unaligned mis. contigs	0	0
# unaligned contigs	0 + 5 part	0 + 5 part
Unaligned length	81853	81853
Genome fraction (%)	95.870	95.891
Duplication ratio	1.009	1.008
# N's per 100 kbp	10.34	0.04
# mismatches per 100 kbp	35.18	35.33
# indels per 100 kbp	3.63	3.46
# genomic features	13759 + 92 part	13748 + 106 part
Complete BUSCO (%)	98.65	98.65
Partial BUSCO (%)	0.00	0.00
# predicted rRNA genes	2 + 4 part	2 + 4 part
Largest alignment	537775	537775
Total aligned length	4769483	4769491
NA50	232147	205673
NGA50	232147	159815
NA90	46041	41640
NGA90	37078	33346
auNA	246636.6	211985.8
auNGA	241778.5	207788.8
LA50	7	8
LGA50	7	9
LA90	26	31
LGA90	28	34

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).

Misassemblies report

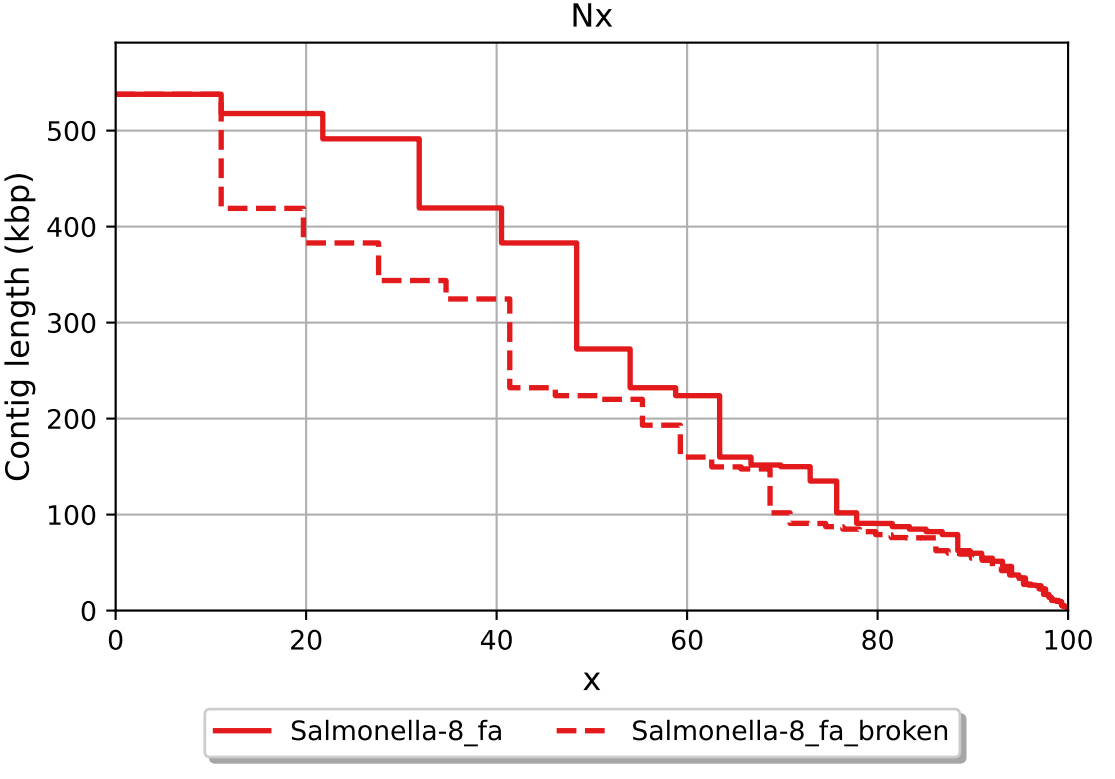
	Salmonella-8_fa	Salmonella-8_fa_broken
# misassemblies	10	10
# contig misassemblies	10	10
# c. relocations	10	10
# c. translocations	0	0
# c. inversions	0	0
# scaffold misassemblies	0	0
# s. relocations	0	0
# s. translocations	0	0
# s. inversions	0	0
# misassembled contigs	9	9
Misassembled contigs length	2096548	1436747
# local misassemblies	6	6
# scaffold gap ext. mis.	0	-
# scaffold gap loc. mis.	0	-
# unaligned mis. contigs	0	0
# mismatches	1678	1685
# indels	173	165
# indels (<= 5 bp)	131	131
# indels (> 5 bp)	42	34
Indels length	3542	3048

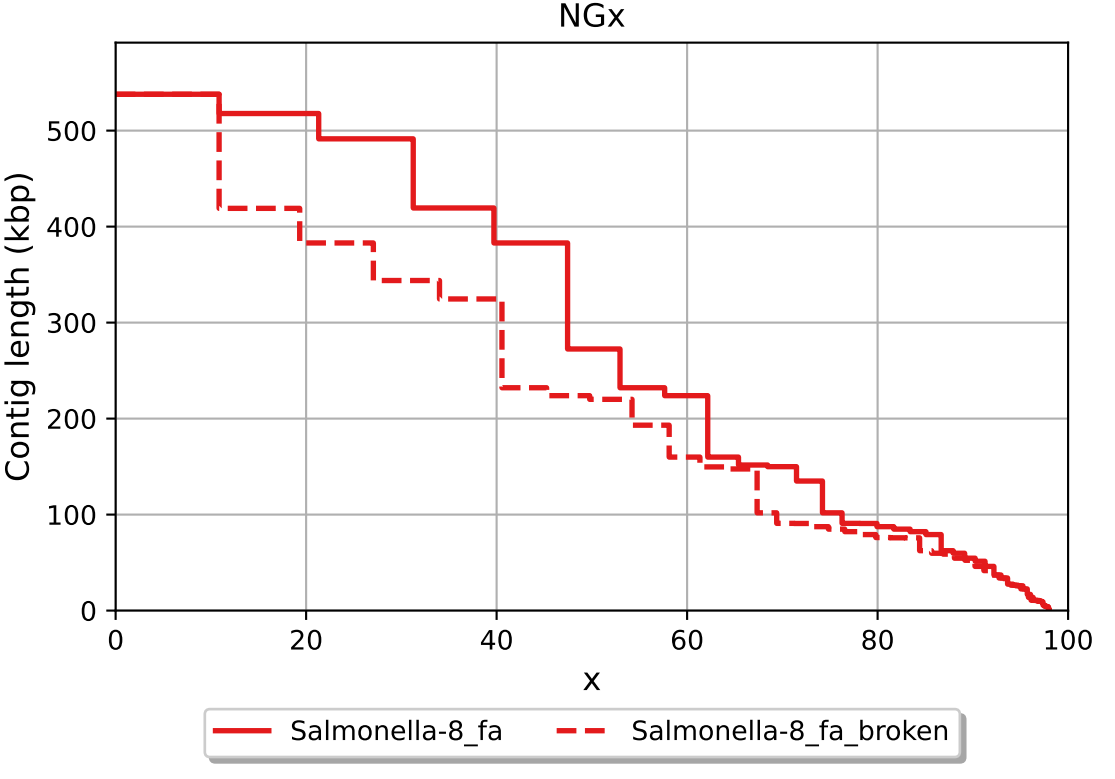
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

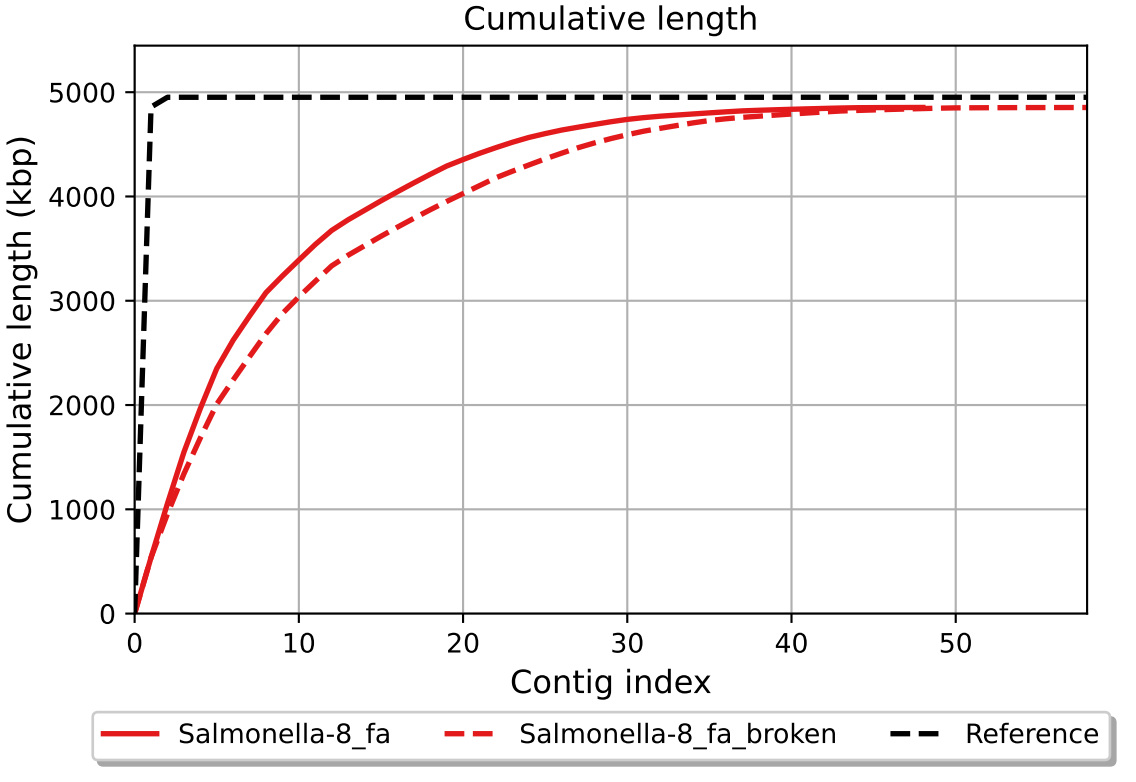
Unaligned report

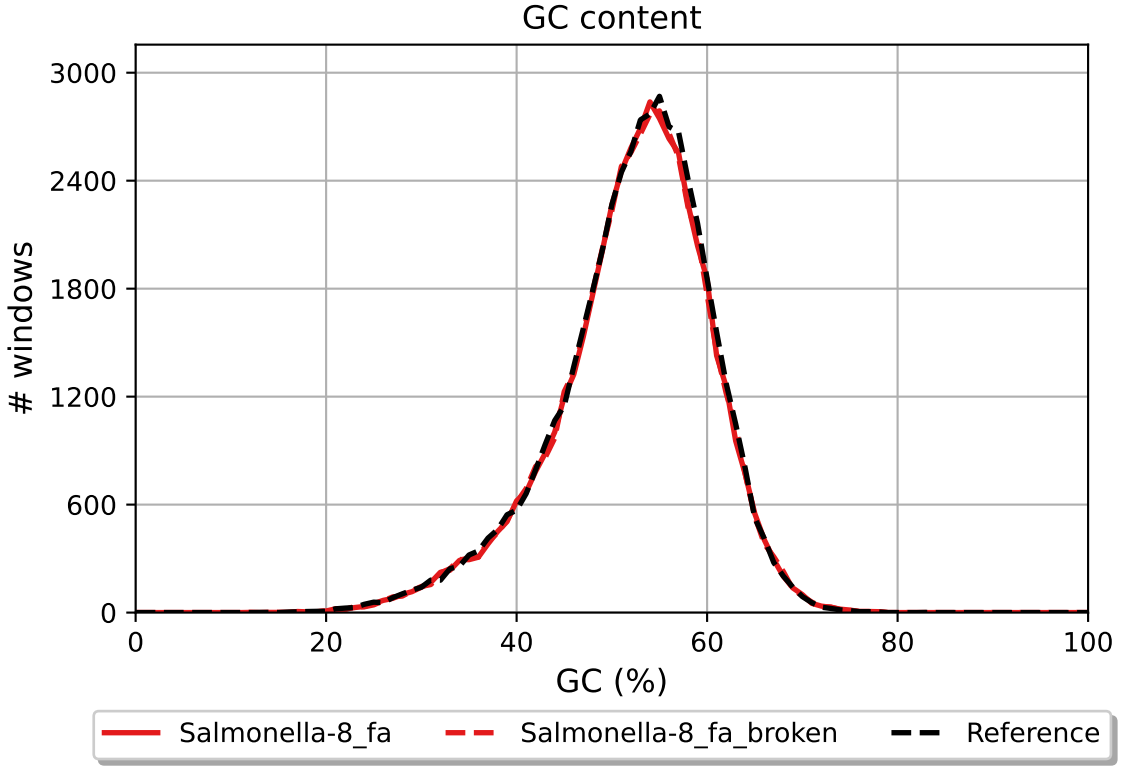
	Salmonella-8_fa	Salmonella-8_fa_broken
# fully unaligned contigs	0	0
Fully unaligned length	0	0
# partially unaligned contigs	5	5
Partially unaligned length	81853	81853
# N's	502	2

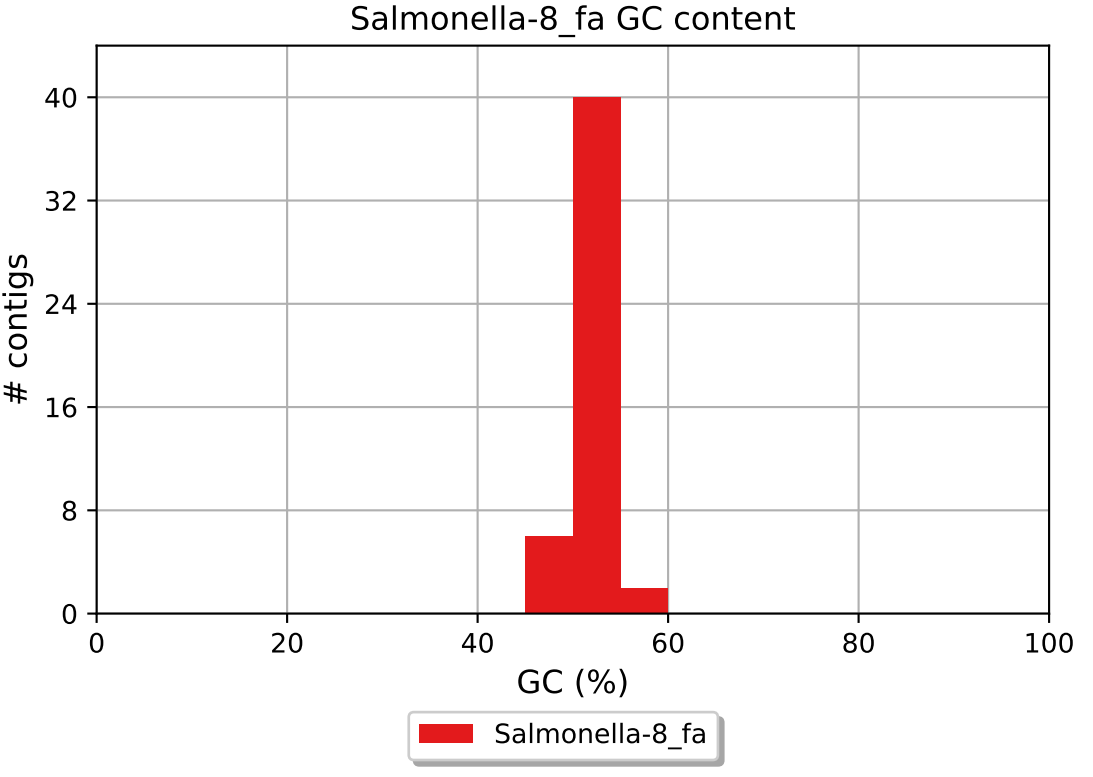
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

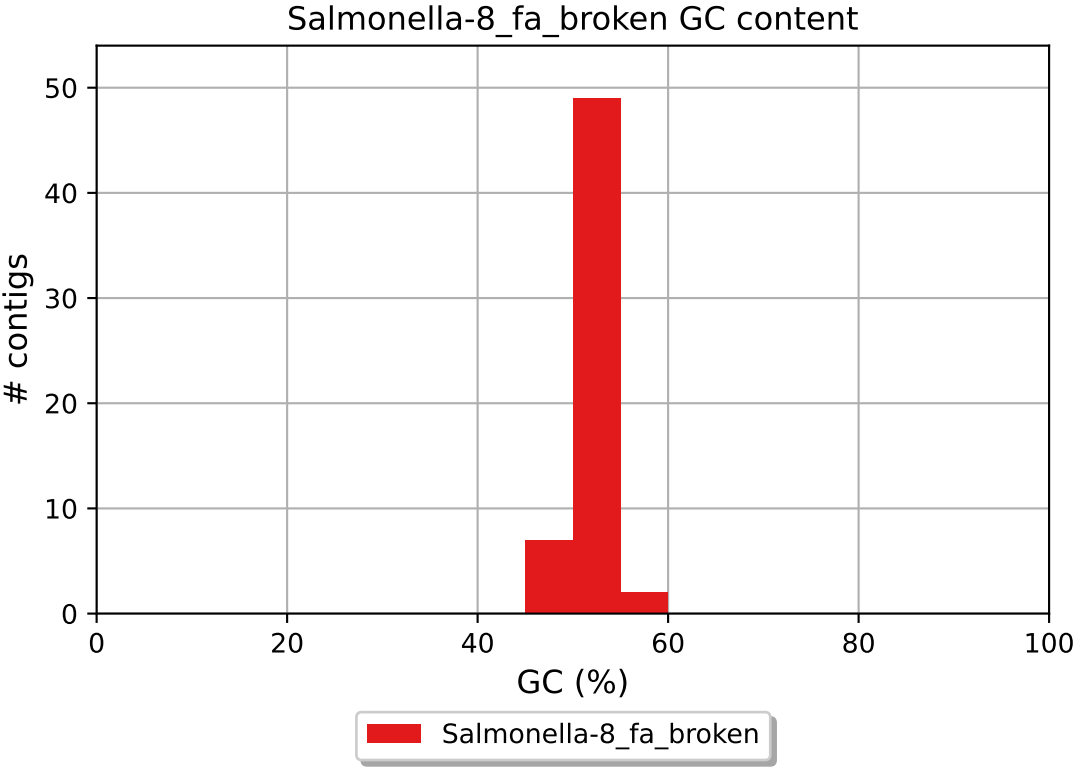

Nx
Contig length (kbp)
0
100
200
300
400
500
0
20
40
60
80
100
x
Salmonella-8_fa
Salmonella-8_fa_broken

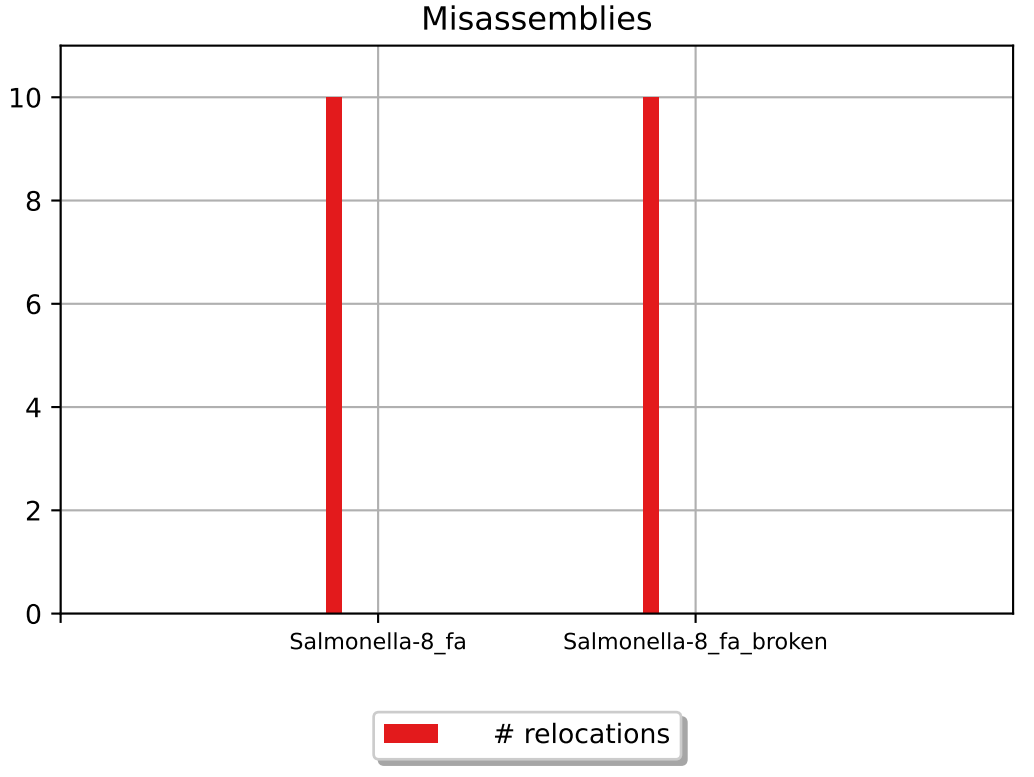

NGx
Contig length (kbp)
0
100
200
300
400
500
0
20
40
60
80
100
x
Salmonella-8_fa
Salmonella-8_fa_broken

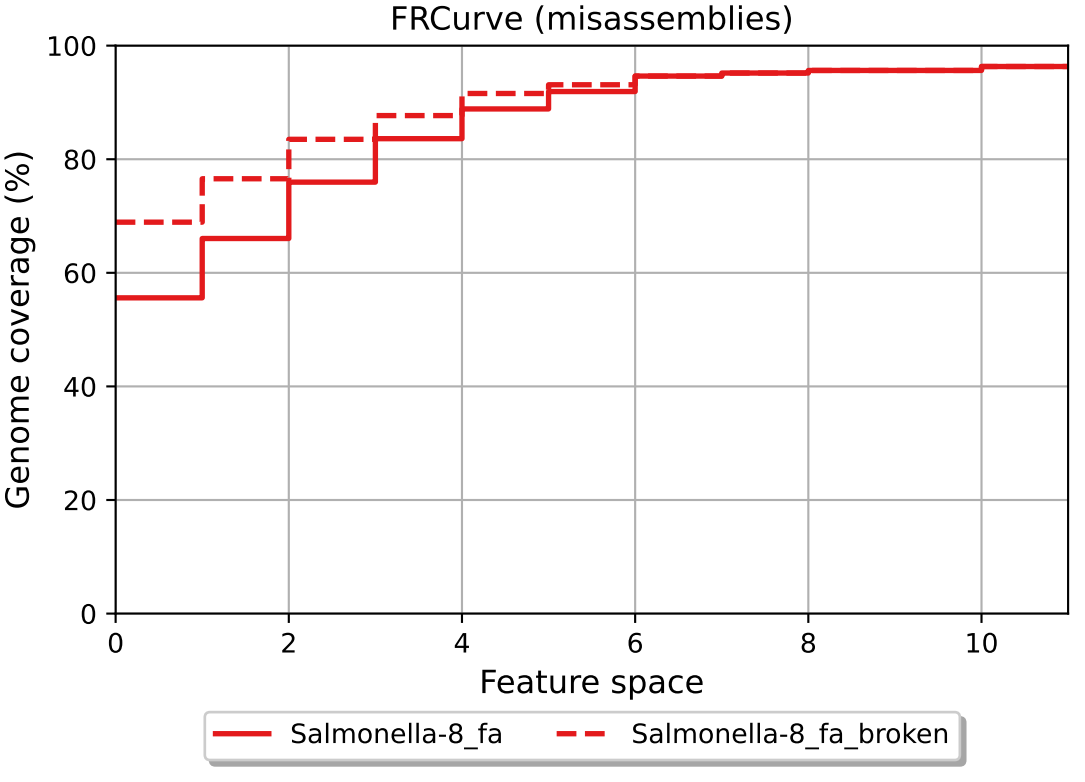

Cumulative length
Cumulative length (kbp)
5000
4000
3000
2000
1000
0
0
10
20
30
40
50
Contig index
Salmonella-8_fa
Salmonella-8_fa_broken
Reference

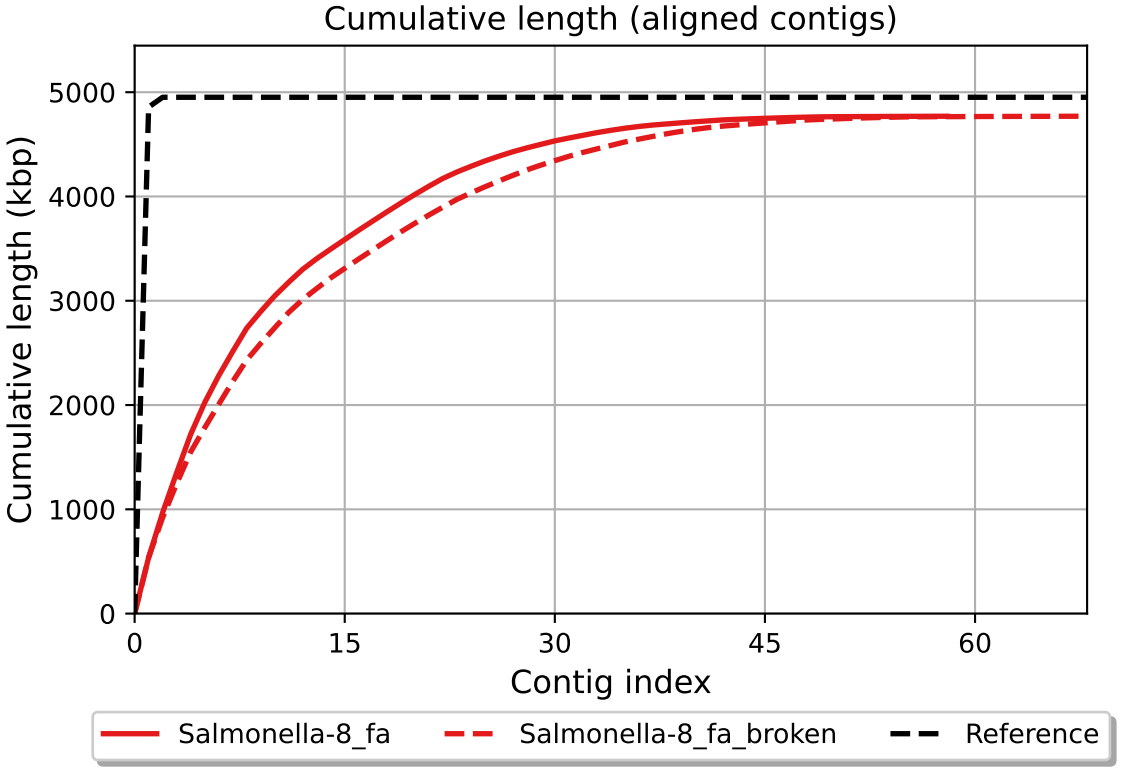

GC content
3000
2400
1800
1200
600
0
windows
0
20
40
60
80
100
GC (%)
Salmonella-8_fa
Salmonella-8_fa_broken
Reference

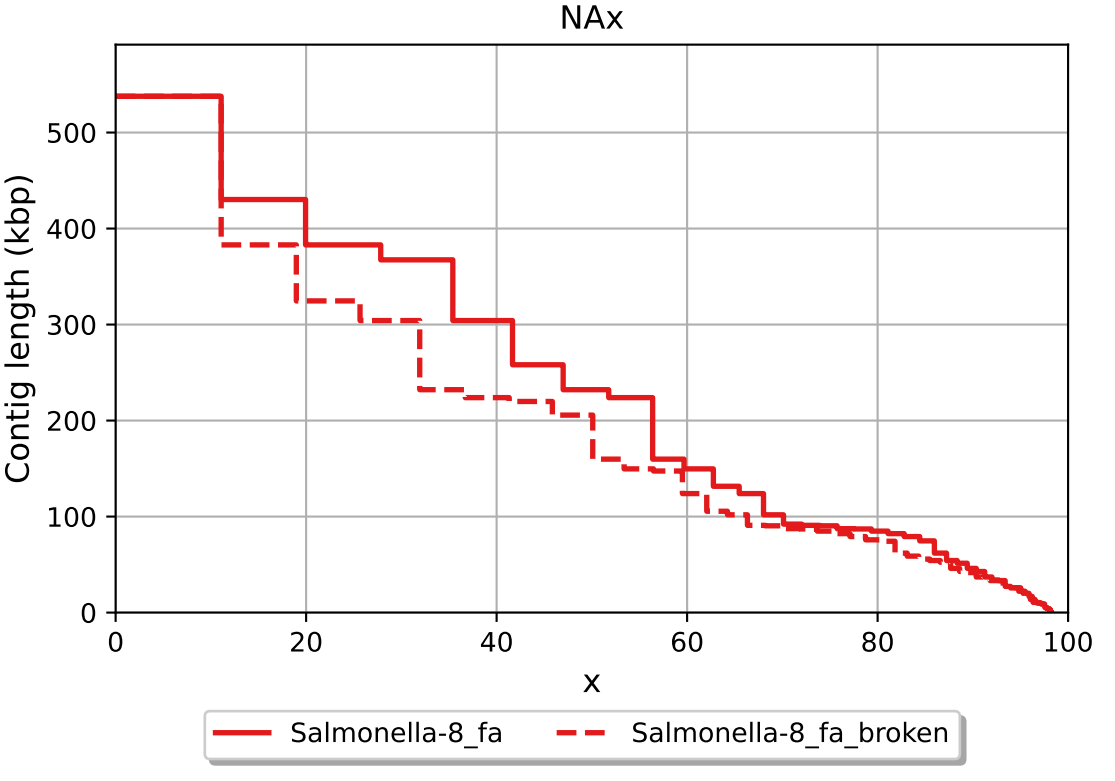

Salmonella-8_fa GC content
contigs
0
8
16
24
32
40
0
20
40
60
80
100
GC (%)
Salmonella-8_fa

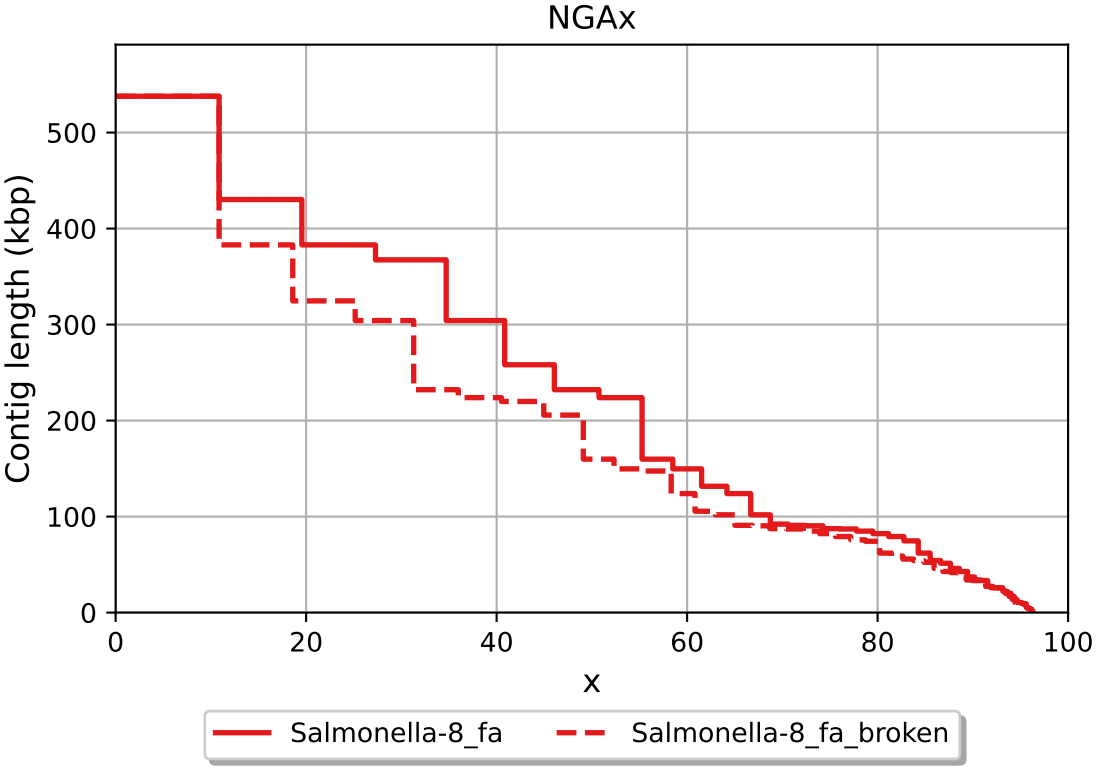

Salmonella-8_fa_broken GC content
contigs
0
10
20
30
40
50
0
20
40
60
80
100
GC (%)
Salmonella-8_fa_broken

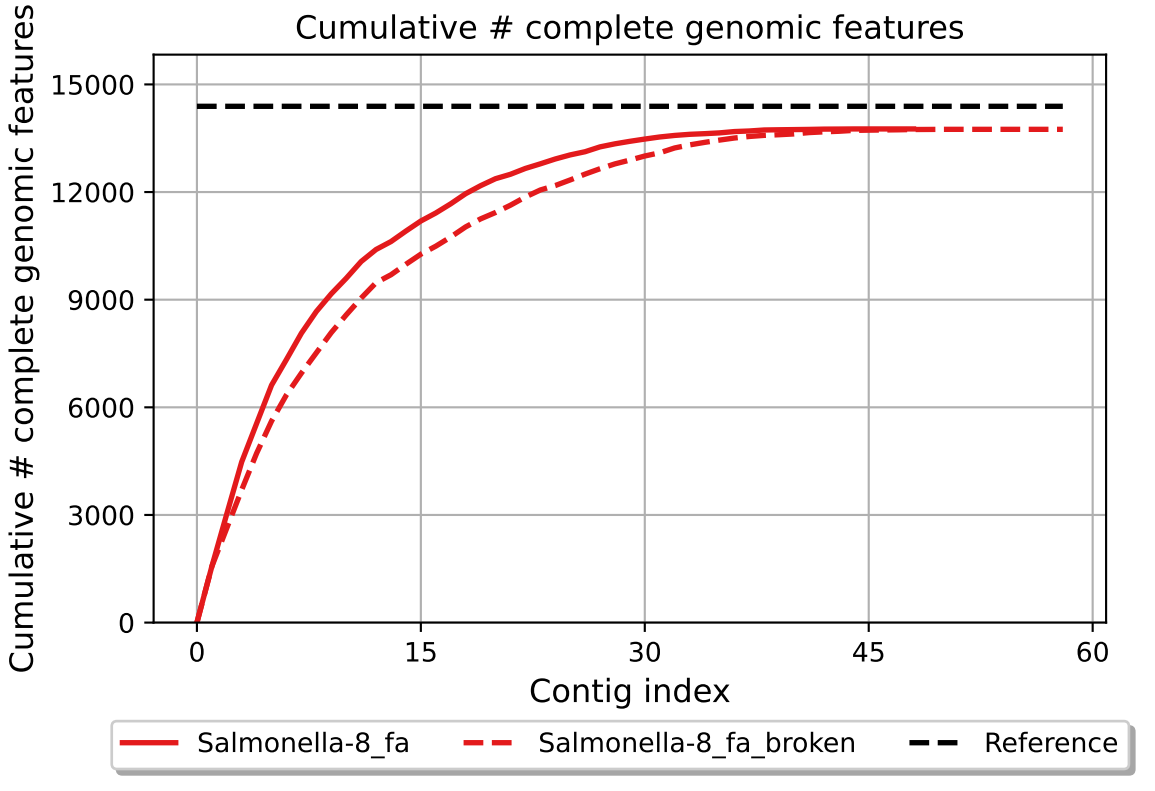

Misassemblies
10
8
6
4
2
0
Salmonella-8_fa
Salmonella-8_fa_broken
relocations

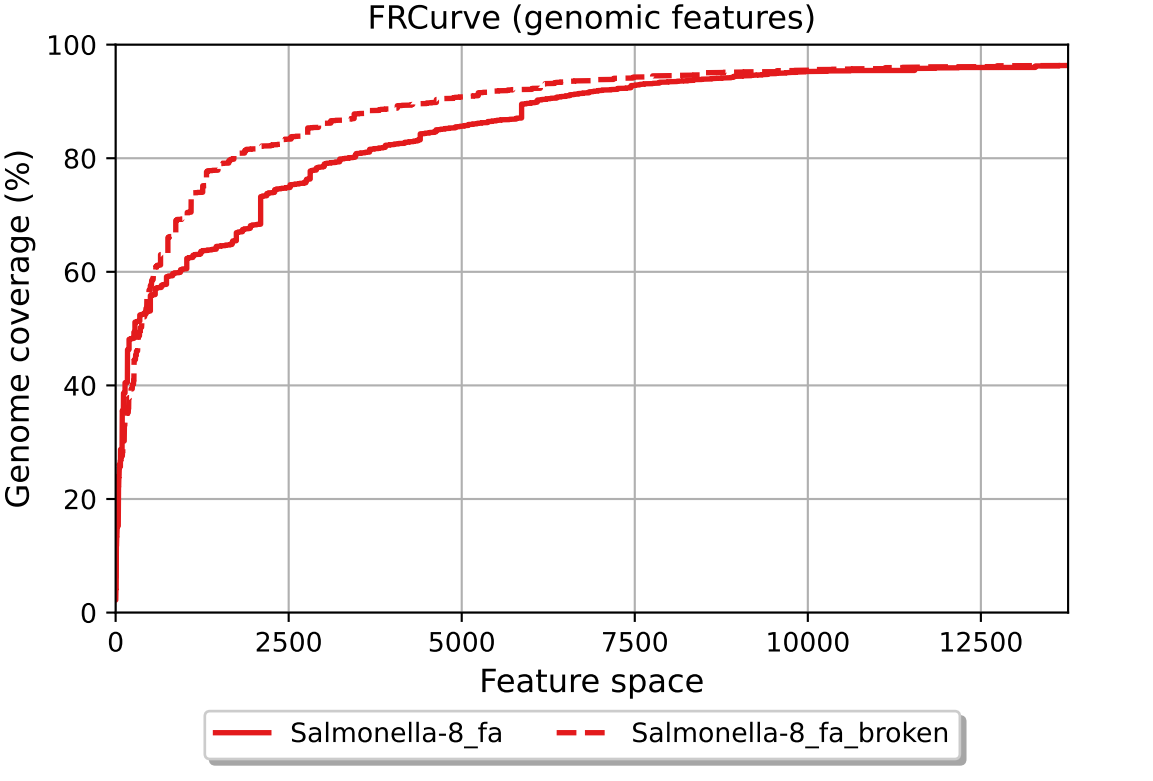

FRCurve (misassemblies)
Genome coverage (%)
100
80
60
40
20
0
0
2
4
6
8
10
Feature space
Salmonella-8_fa
Salmonella-8_fa_broken

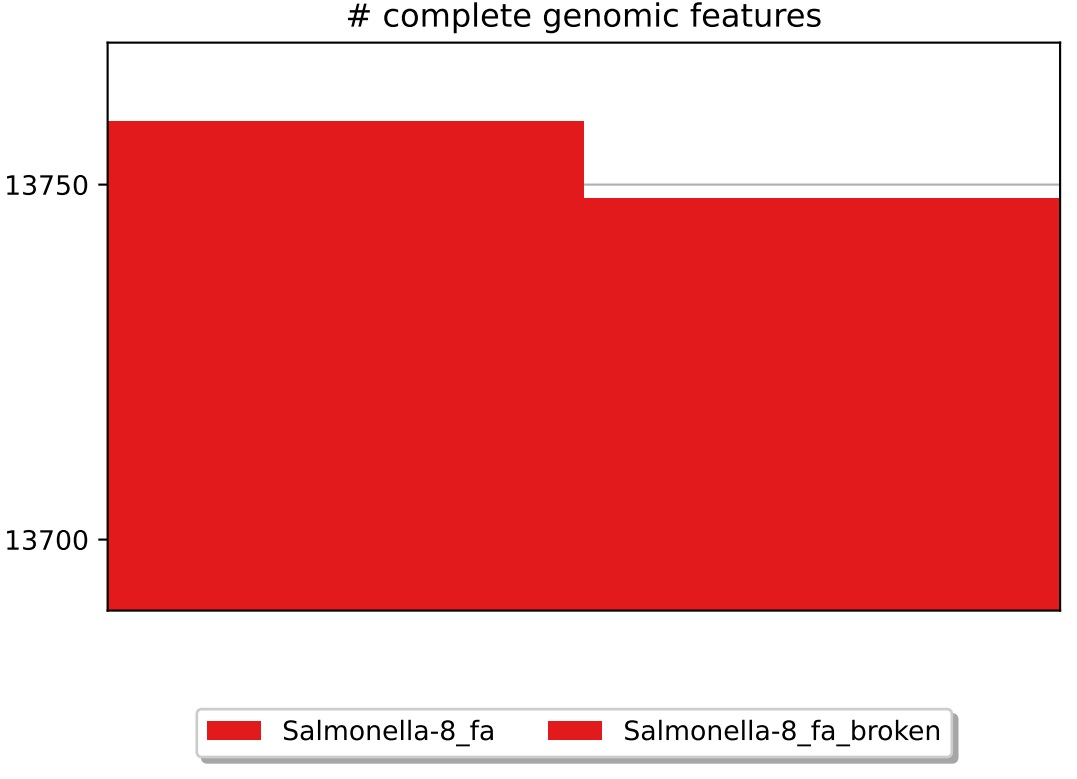

Cumulative length (aligned contigs)
Cumulative length (kbp)
5000
4000
3000
2000
1000
0
0
15
30
45
60
Contig index
Salmonella-8_fa
Salmonella-8_fa_broken
Reference

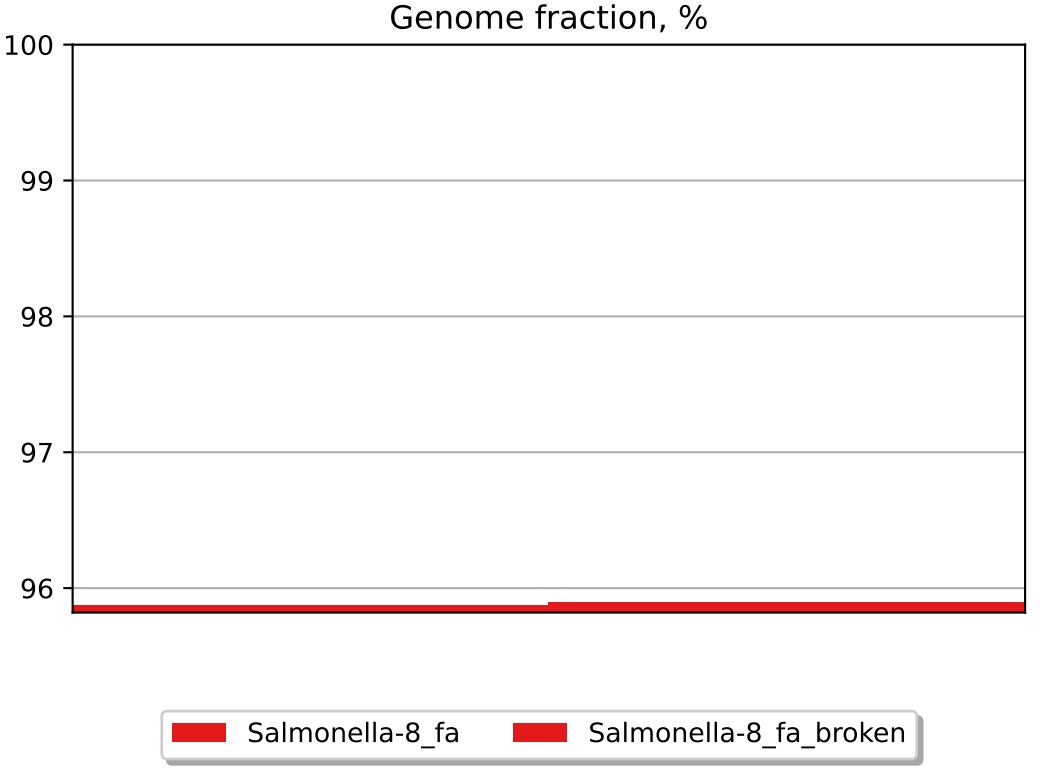

NAx
Contig length (kbp)
0
100
200
300
400
500
0
20
40
60
80
100
x
Salmonella-8_fa
Salmonella-8_fa_broken


NGAx
Contig length (kbp)
0
100
200
300
400
500
0
20
40
60
80
100
x
Salmonella-8_fa
Salmonella-8_fa_broken


Cumulative # complete genomic features
Cumulative # complete genomic features
15000
12000
9000
6000
3000
0
0
15
30
45
60
Contig index
Salmonella-8_fa
Salmonella-8_fa_broken
Reference


FRCurve (genomic features)
Genome coverage (%)
100
80
60
40
20
0
0
2500
5000
7500
10000
12500
Feature space
Salmonella-8_fa
Salmonella-8_fa_broken


complete genomic features
13750
13700
Salmonella-8_fa
Salmonella-8_fa_broken


Genome fraction, %
100
99
98
97
96
Salmonella-8_fa
Salmonella-8_fa_broken

Report

	SPAdes_on_data_2_and_data_1__Scaffolds	SPAdes_on_data_2_and_data_1__Scaffolds_broken
# contigs (>= 0 bp)	191	-
# contigs (>= 1000 bp)	63	71
Total length (>= 0 bp)	4813654	-
Total length (>= 1000 bp)	4788123	4786289
# contigs	72	82
Largest contig	549770	322638
Total length	4794213	4792805
Reference length	4951383	4951383
GC (%)	52.13	52.13
Reference GC (%)	52.24	52.24
N50	193475	168456
NG50	178662	149367
N90	53847	51504
NG90	44707	44707
auN	224465.2	166813.7
auNG	217340.1	161471.2
L50	8	10
LG50	9	11
L90	27	32
LG90	30	35
# misassemblies	7	7
# misassembled contigs	7	7
Misassembled contigs length	1416504	1035190
# local misassemblies	5	5
# scaffold gap ext. mis.	0	-
# scaffold gap loc. mis.	3	-
# unaligned mis. contigs	0	0
# unaligned contigs	0 + 5 part	0 + 5 part
Unaligned length	81933	81933
Genome fraction (%)	96.032	96.017
Duplication ratio	1.001	1.001
# N's per 100 kbp	22.94	0.00
# mismatches per 100 kbp	28.13	28.20
# indels per 100 kbp	3.65	3.46
# genomic features	13701 + 108 part	13685 + 112 part
Complete BUSCO (%)	98.65	98.65
Partial BUSCO (%)	0.00	0.00
# predicted rRNA genes	2 + 0 part	2 + 0 part
Largest alignment	376430	322563
Total aligned length	4710913	4709961
NA50	178662	129564
NGA50	152567	123652
NA90	35969	35969
NGA90	30214	26152
auNA	187266.9	155275.2
auNGA	181322.6	150302.2
LA50	9	11
LGA50	10	12
LA90	32	37
LGA90	36	42

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).

Misassemblies report

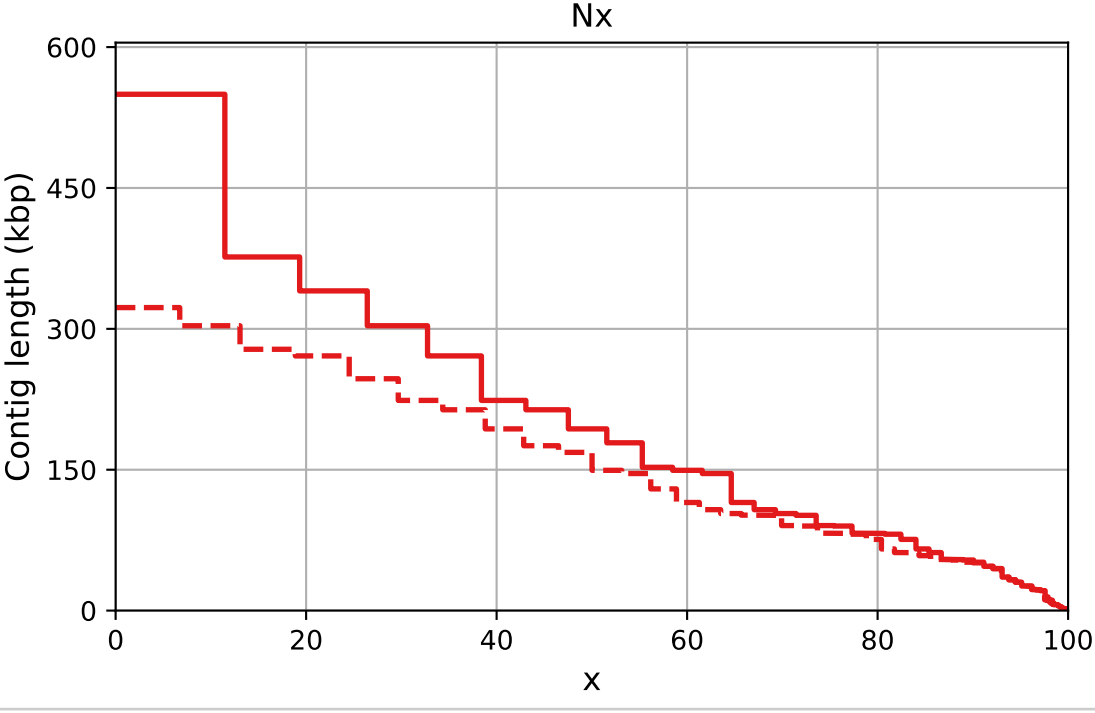
	SPAdes_on_data_2_and_data_1__Scaffolds	SPAdes_on_data_2_and_data_1__Scaffolds_broken
# misassemblies	7	7
# contig misassemblies	7	7
# c. relocations	7	7
# c. translocations	0	0
# c. inversions	0	0
# scaffold misassemblies	0	0
# s. relocations	0	0
# s. translocations	0	0
# s. inversions	0	0
# misassembled contigs	7	7
Misassembled contigs length	1416504	1035190
# local misassemblies	5	5
# scaffold gap ext. mis.	0	-
# scaffold gap loc. mis.	3	-
# unaligned mis. contigs	0	0
# mismatches	1325	1328
# indels	172	163
# indels (<= 5 bp)	137	136
# indels (> 5 bp)	35	27
Indels length	3722	2336

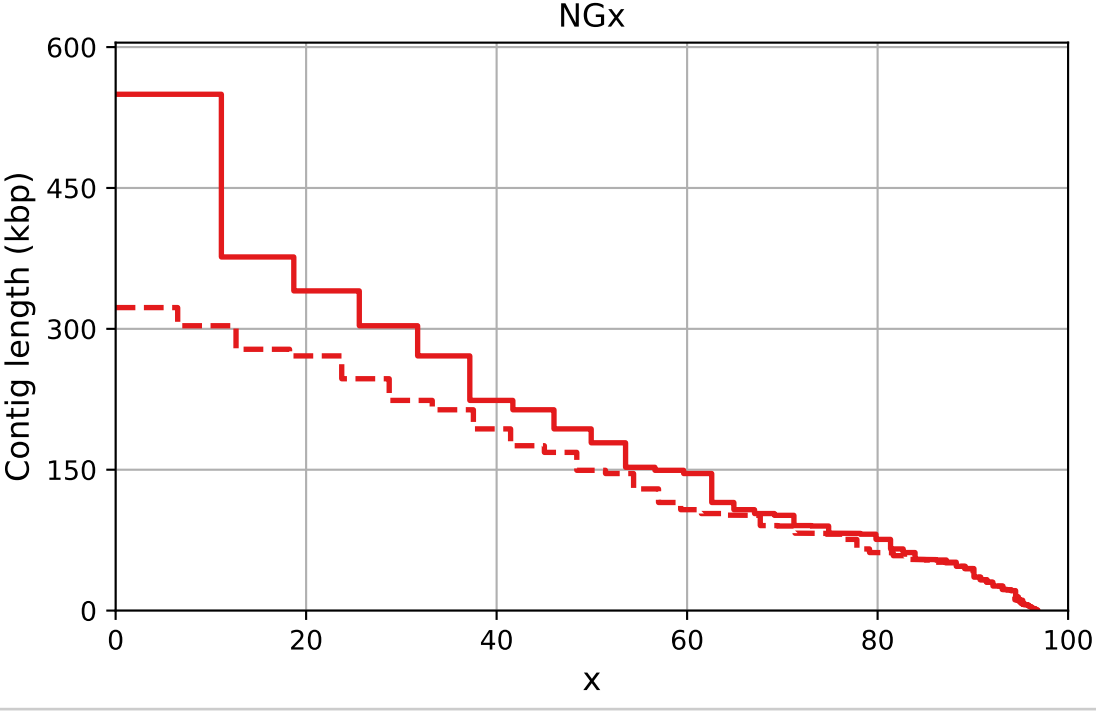
All statistics are based on contigs of size >= 500 bp, unless otherwise noted
(e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

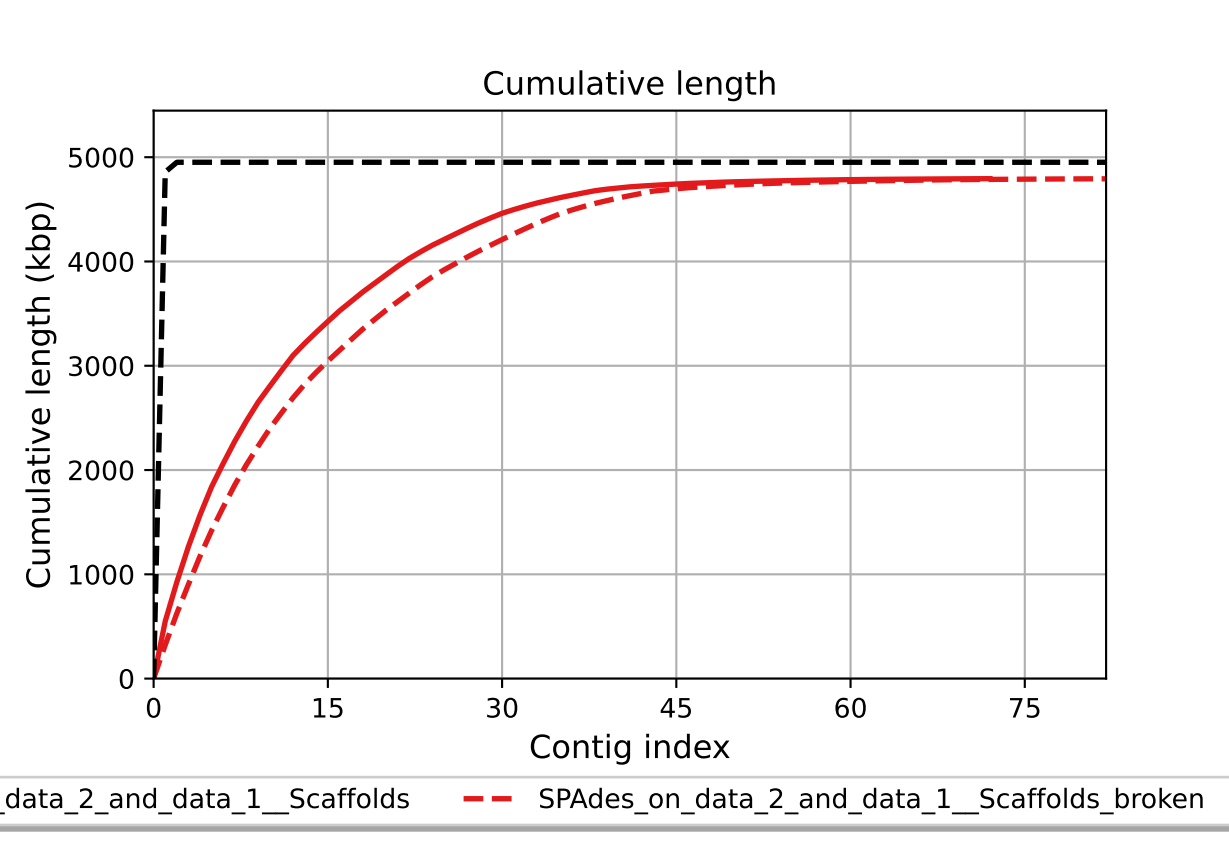
Unaligned report

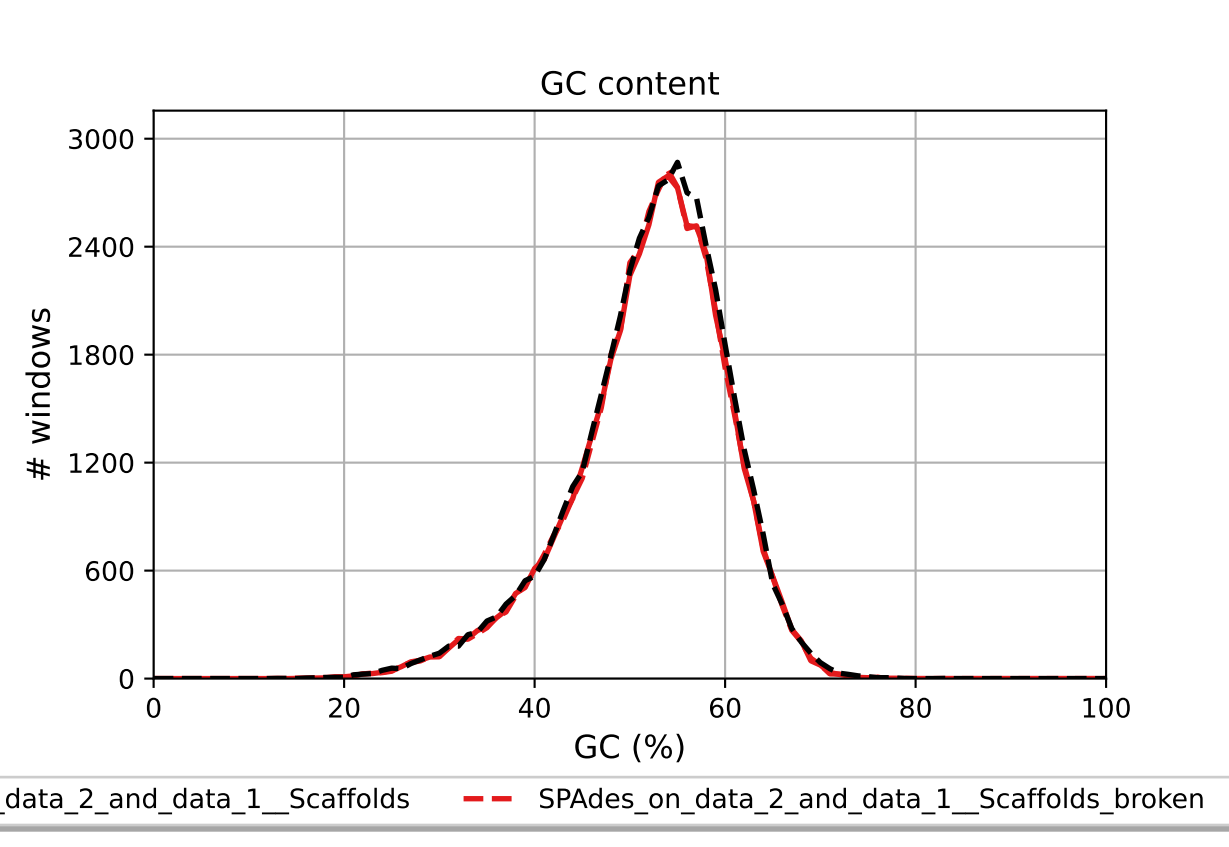
	SPAdes_on_data_2_and_data_1__Scaffolds	SPAdes_on_data_2_and_data_1__Scaffolds_broken
# fully unaligned contigs	0	0
Fully unaligned length	0	0
# partially unaligned contigs	5	5
Partially unaligned length	81933	81933
# N's	1100	0

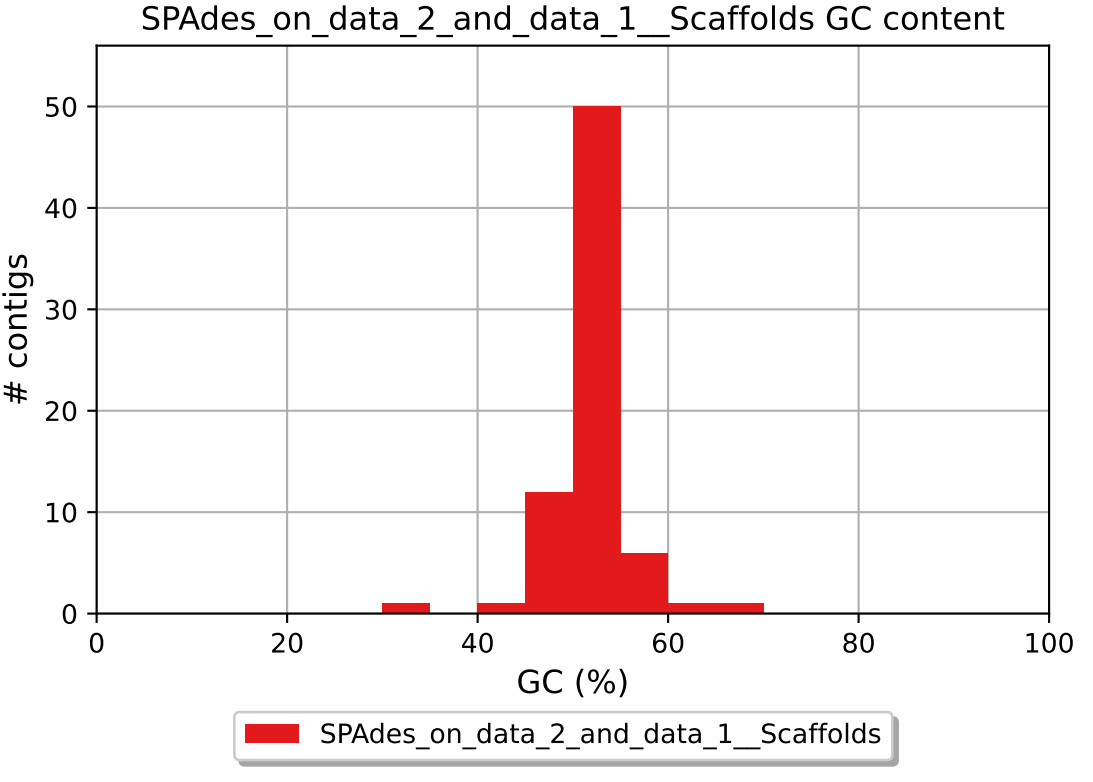
All statistics are based on contigs of size >= 500 bp, unless otherwise noted
(e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

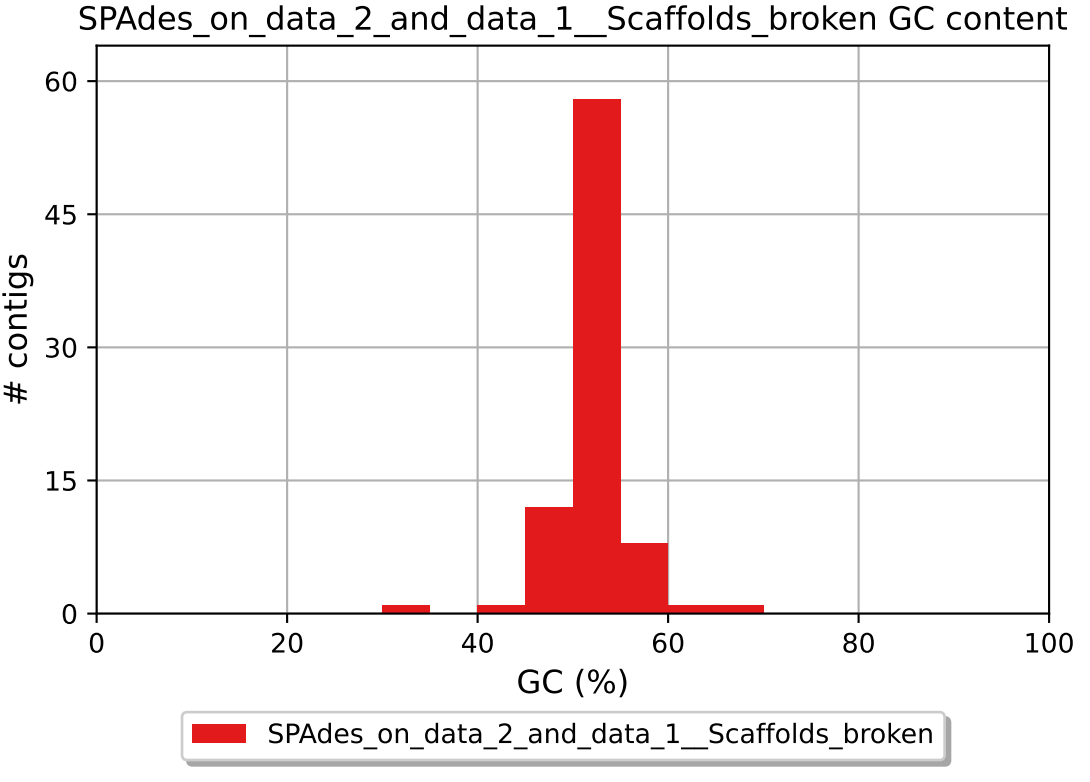

Nx
600
450
300
150
0
Contig length (kbp)
0
20
40
60
80
100
x
PAdes_on_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds

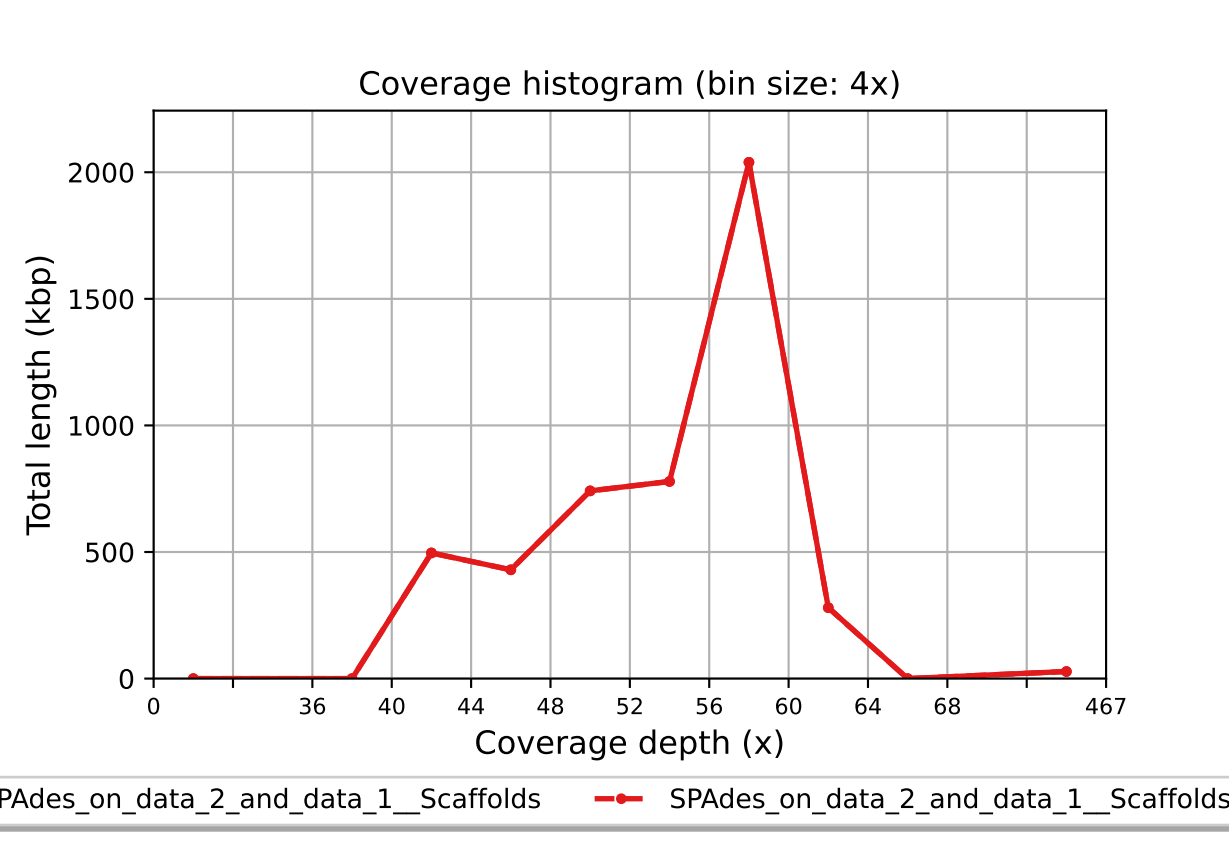

NGx
Contig length (kbp)
600
450
300
150
0
0
20
40
60
80
100
x
PAdes_on_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds

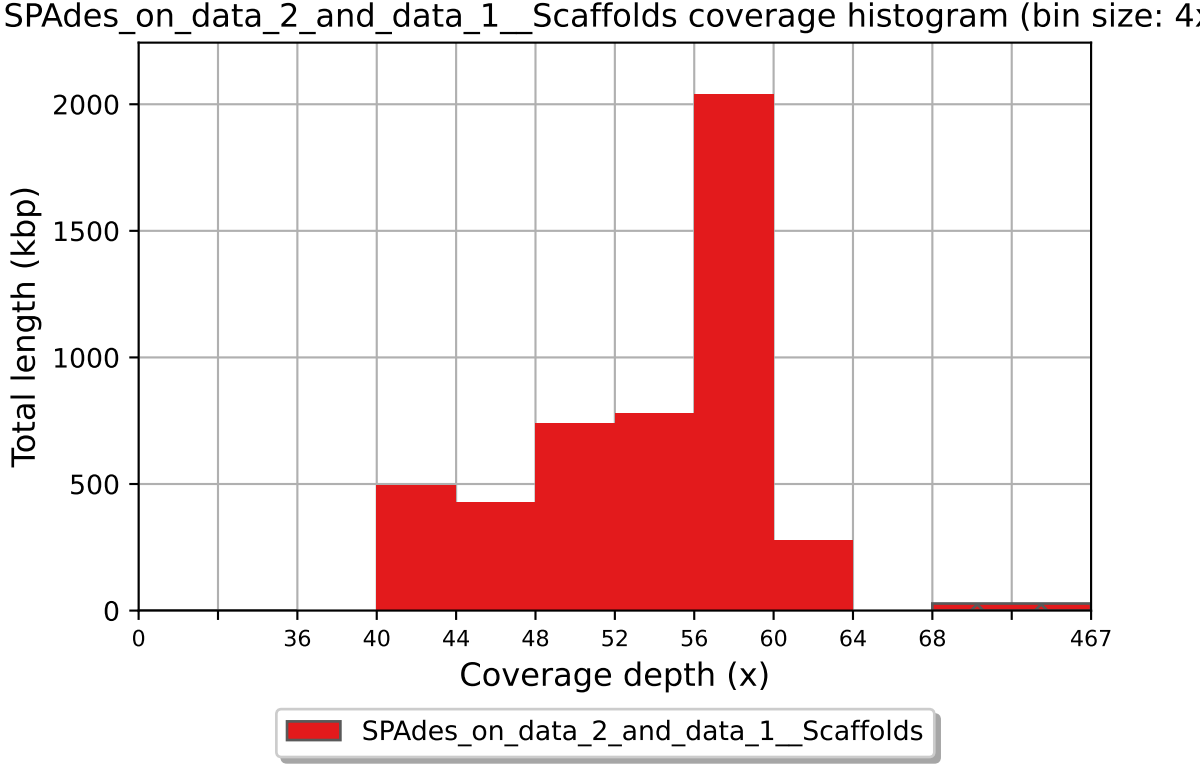

Cumulative length
Cumulative length (kbp)
5000
4000
3000
2000
1000
0
0
15
30
45
60
75
Contig index
_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds_broken

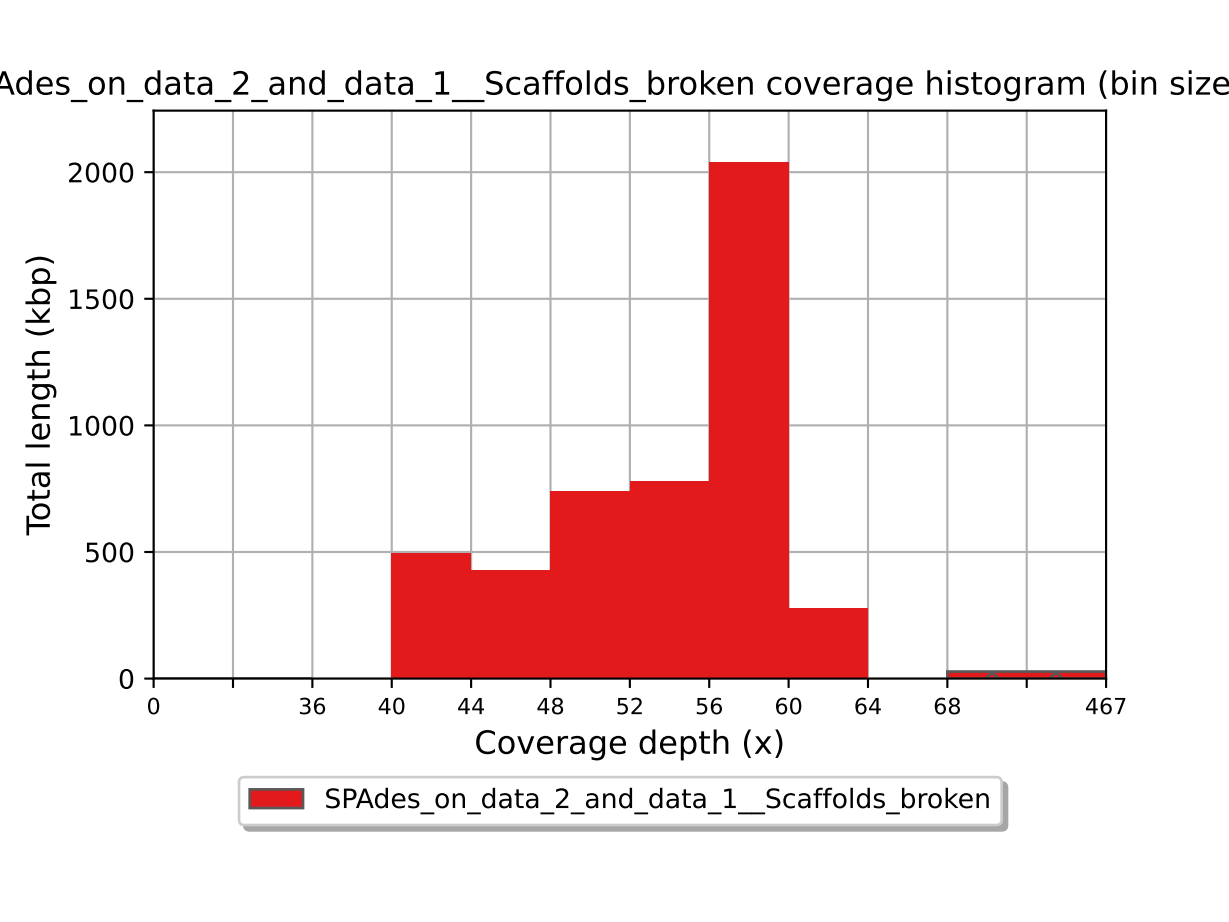

GC content
3000
2400
1800
1200
600
0
windows
0
20
40
60
80
100
GC (%)
_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds_broken

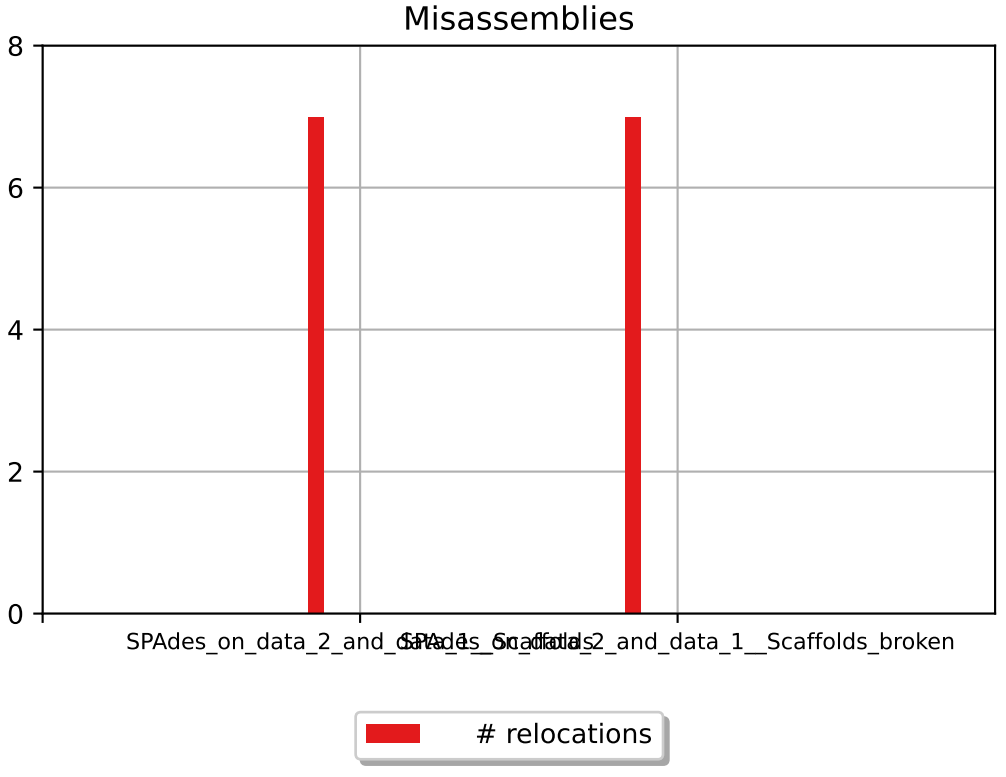

SPAdes_on_data_2_and_data_1__Scaffolds GC content
contigs
GC (%)
0
10
20
30
40
50
0
20
40
60
80
100
SPAdes_on_data_2_and_data_1__Scaffolds

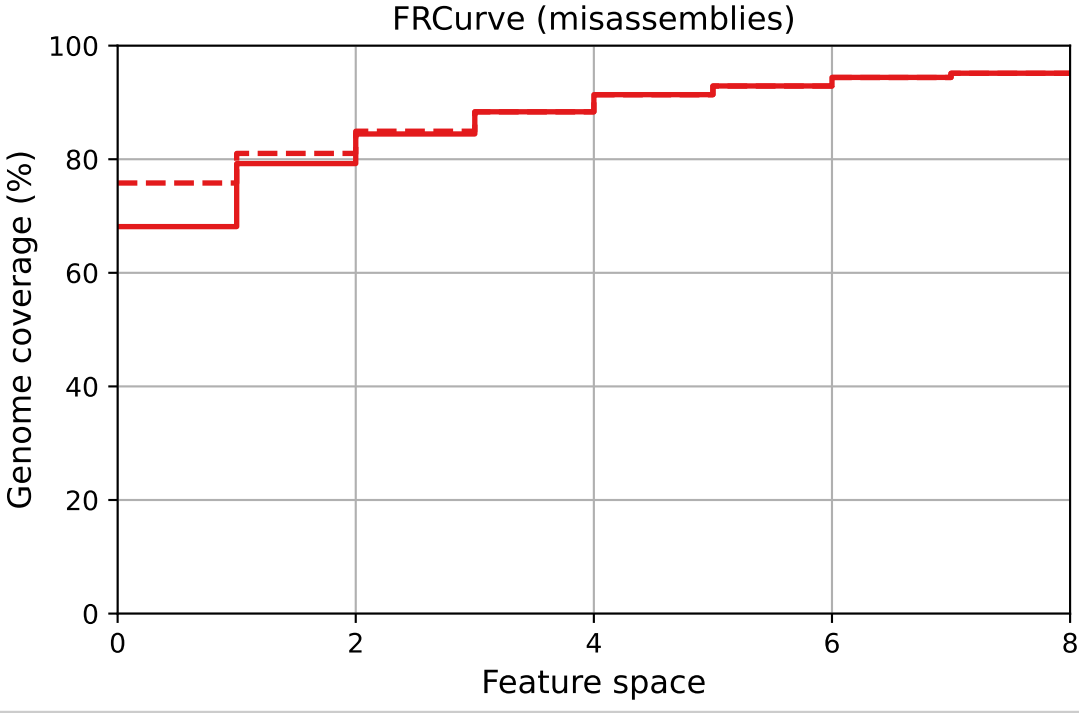

SPAdes_on_data_2_and_data_1__Scaffolds_broken GC content
contigs
GC (%)
0
15
30
45
60
0
20
40
60
80
100
SPAdes_on_data_2_and_data_1__Scaffolds_broken

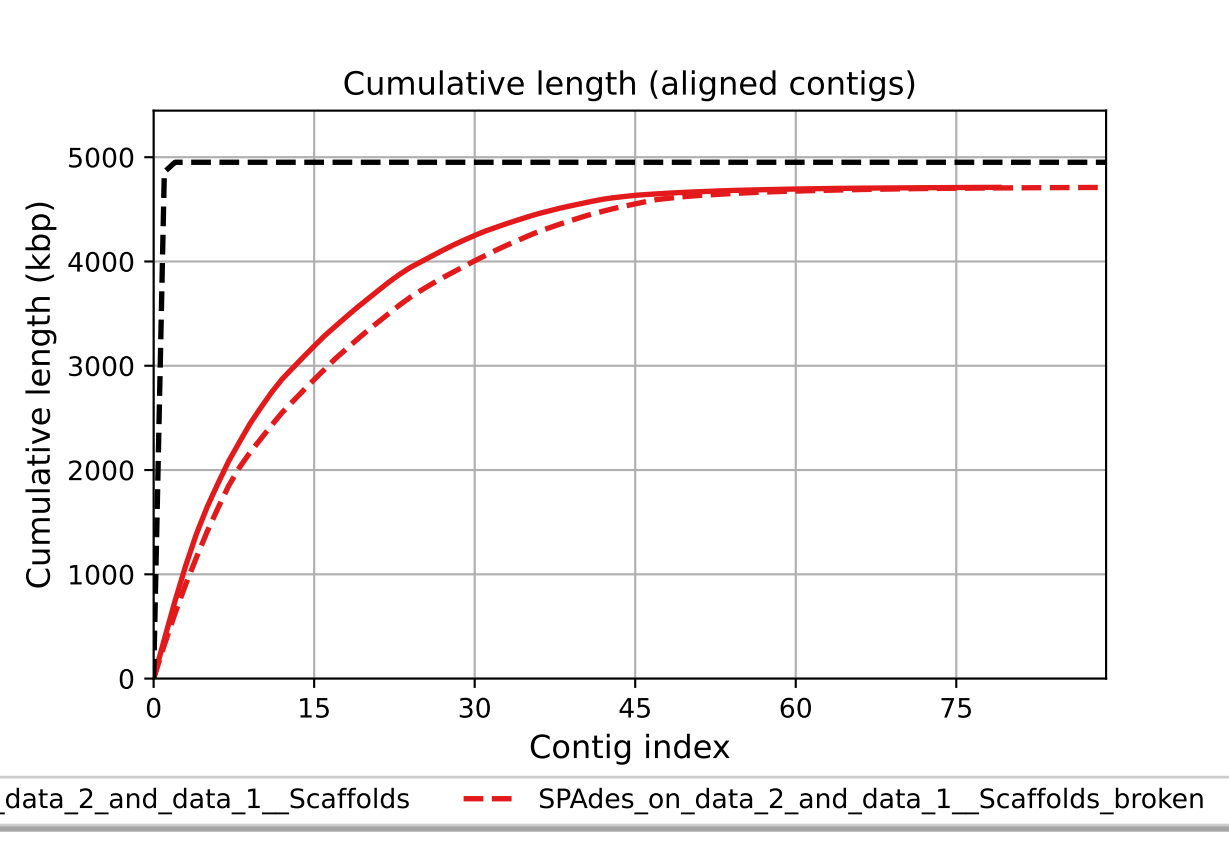

Coverage histogram (bin size: 4x)
Total length (kbp)
2000
1500
1000
500
0
0
36
40
44
48
52
56
60
64
68
467
Coverage depth (x)
SPAdes_on_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds

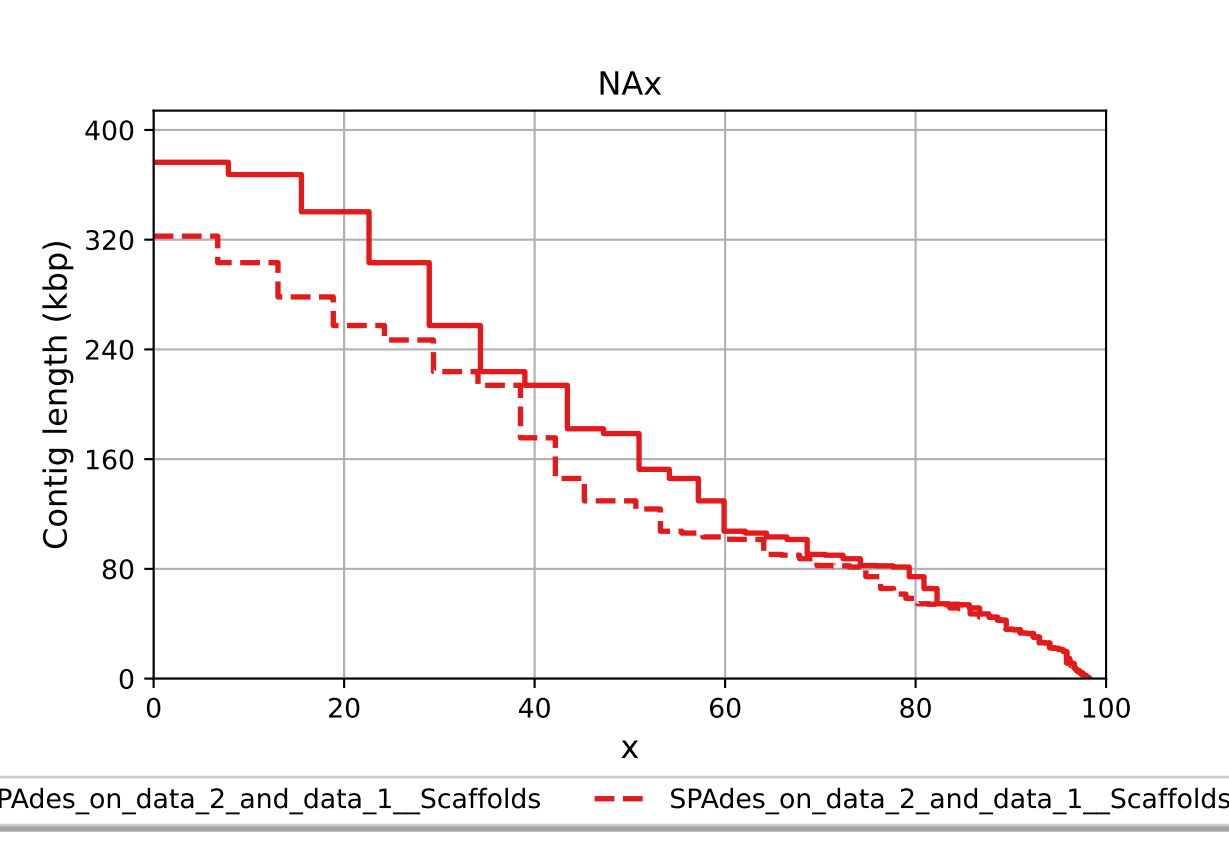

SPAdes_on_data_2_and_data_1__Scaffolds coverage histogram (bin size: 4x
Total length (kbp)
2000
1500
1000
500
0
0
36
40
44
48
52
56
60
64
68
467
Coverage depth (x)
SPAdes_on_data_2_and_data_1__Scaffolds

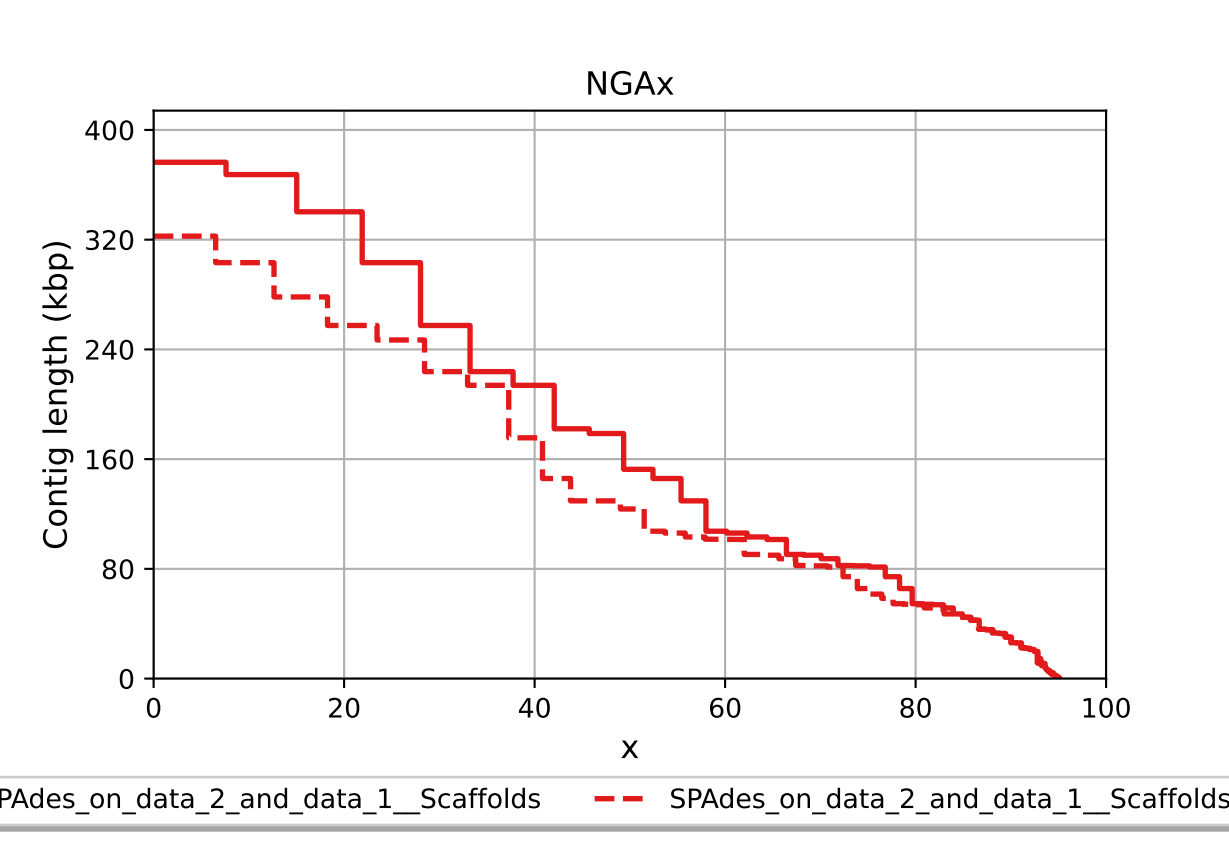

Ades_on_data_2_and_data_1__Scaffolds_broken coverage histogram (bin size
Total length (kbp)
2000
1500
1000
500
0
0
36
40
44
48
52
56
60
64
68
467
Coverage depth (x)
SPAdes_on_data_2_and_data_1__Scaffolds_broken

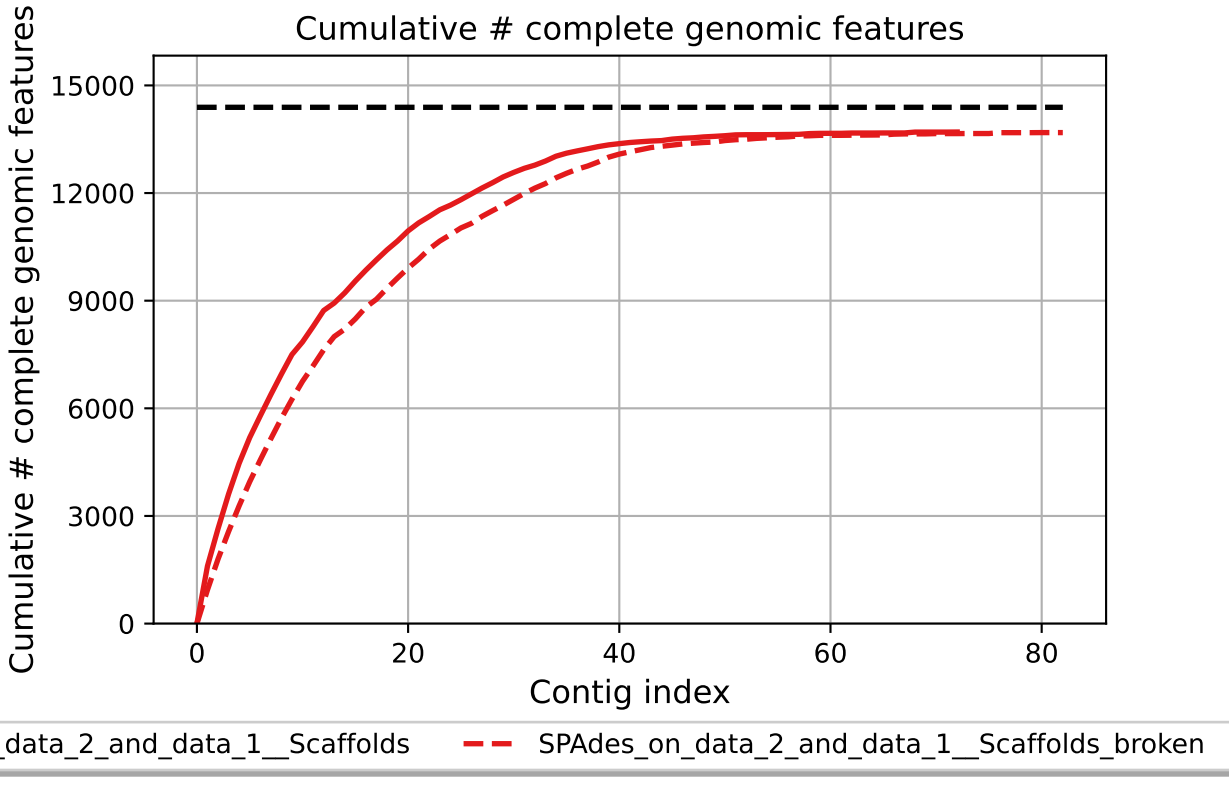

Misassemblies
8
6
4
2
0
SPAdes_on_data_2_and_data_1__Scaffolds_broken
relocations

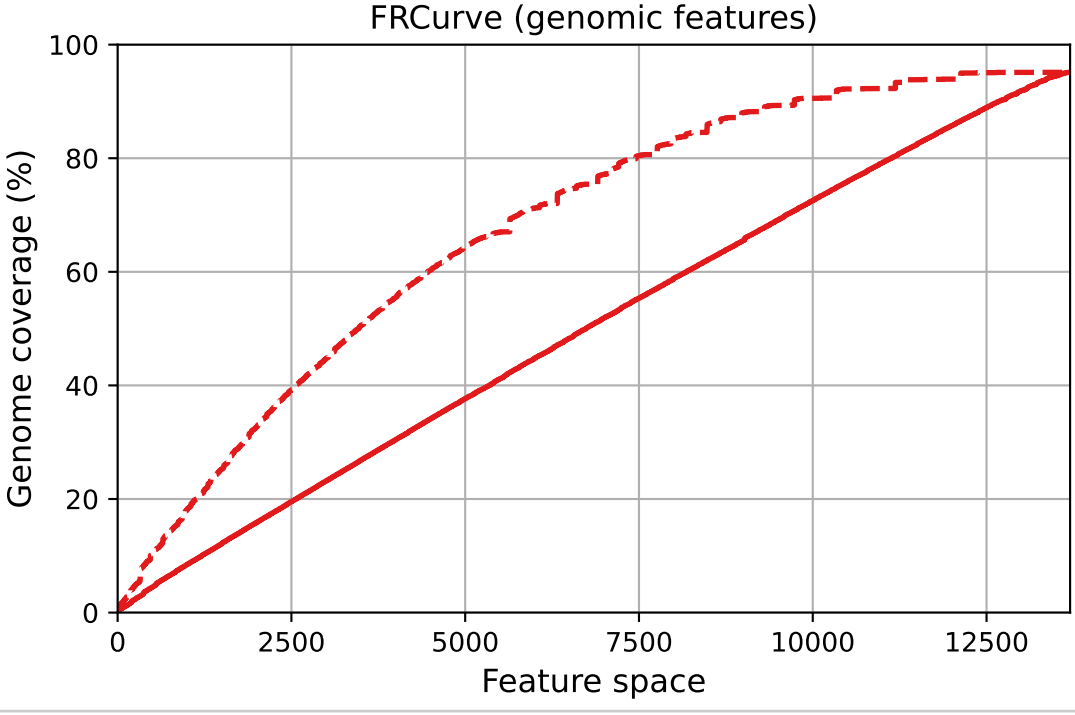

FRCurve (misassemblies)
Genome coverage (%)
100
80
60
40
20
0
0
2
4
6
8
Feature space
PAdes_on_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds

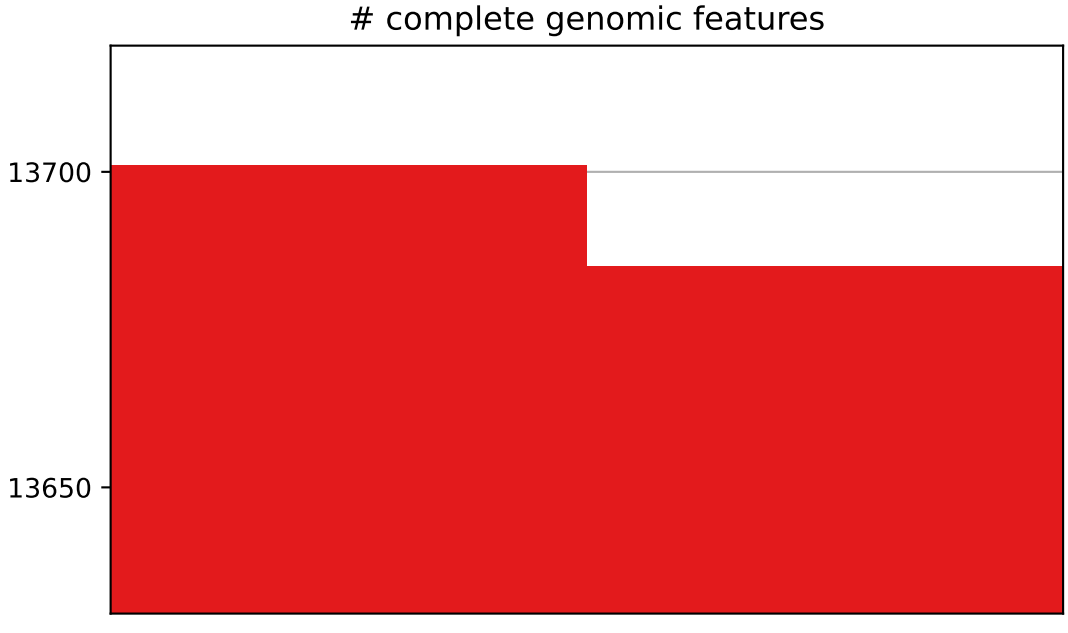

Cumulative length (aligned contigs)
Cumulative length (kbp)
5000
4000
3000
2000
1000
0
0
15
30
45
60
75
Contig index
_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds_broken


NAx
Contig length (kbp)
400
320
240
160
80
0
0
20
40
60
80
100
x
PAdes_on_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds

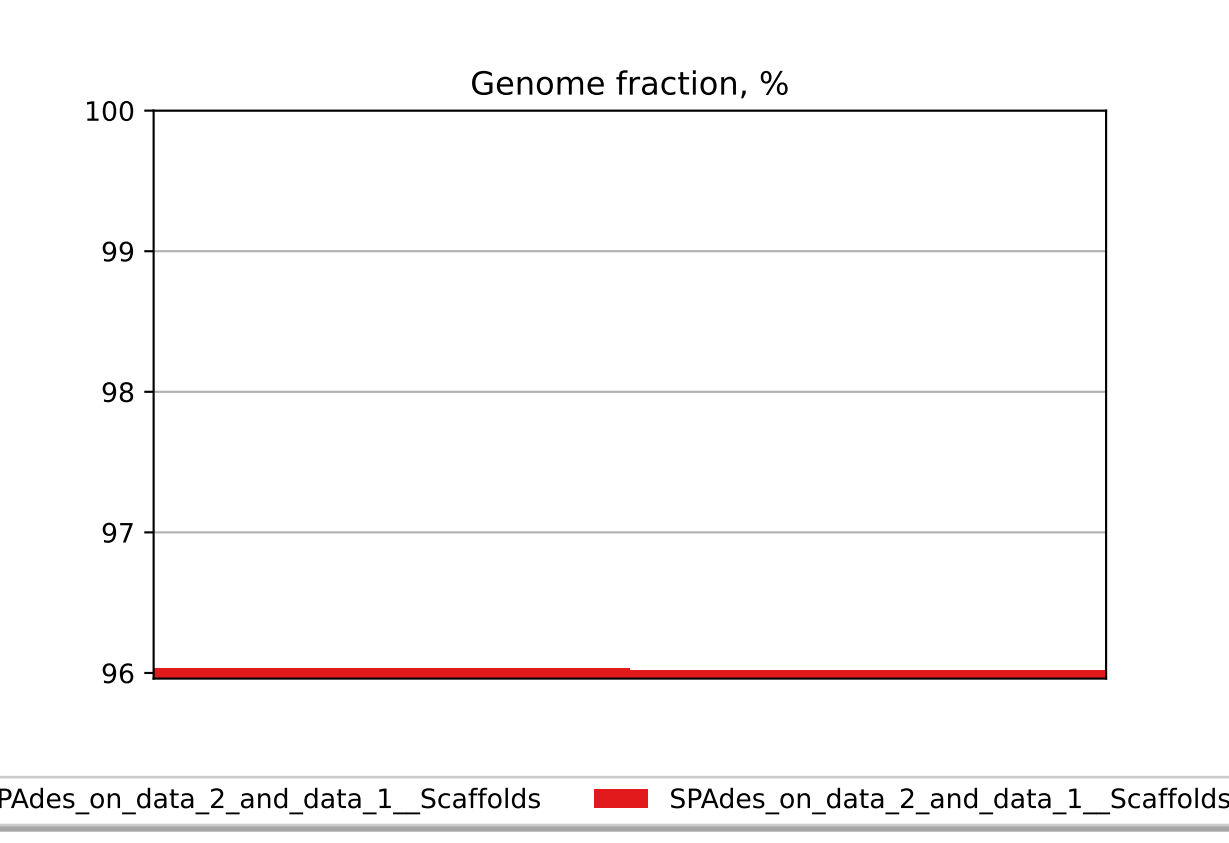

NGAx
Contig length (kbp)
400
320
240
160
80
0
0
20
40
60
80
100
x
PAdes_on_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds


Cumulative # complete genomic features
Cumulative # complete genomic features
15000
12000
9000
6000
3000
0
0
20
40
60
80
Contig index
_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds_broken


FRCurve (genomic features)
Genome coverage (%)
Feature space
0
20
40
60
80
100
0
2500
5000
7500
10000
12500
PAdes_on_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds



SPAdes_on_data_2_and_data_1__Scaffolds SPAdes_on_data_2_and_data_1__Scaffolds


Genome fraction, %
100
99
98
97
96
SPAdes_on_data_2_and_data_1__Scaffolds
SPAdes_on_data_2_and_data_1__Scaffolds