

Genome Assembly Improvement and Quality Assessment



On the other hand, the genome of *Arabidopsis lyrata* is available in a **draft form**; partially solved with contigs or scaffolds not necessarily in the right order, and some may be missing or misoriented. The **goal** is to **improve** the assembly of this draft genome using the reference genome of *A. thaliana*. To do this, both genome files (**.fna.gz**) need to be **downloaded** from NCBI and then **uncompressed**. This step ensures that the files are accessible and properly formatted for use with genome scaffolding and quality assessment tools.

The **RagTag** performs **reference-guided scaffolding**, which means it uses the known chromosome layout of a closely related organism (*A. thaliana*) to guide the reconstruction of the *A. lyrata* genome. This works on the assumption that since these species are evolutionarily

close, their genomes are largely **syntenic** — meaning they share similar gene order and structure.

Internally, RagTag works by **aligning** the contigs of the draft genome to the reference genome using a **fast sequence alignment algorithm (like minimap2)**. It then analyzes where and how these contigs map to the reference chromosomes. Based on this mapping, RagTag determines the **best order and orientation** of the contigs and attempts to **link** them together **into longer scaffolds**, inserting Ns where gaps exist. The result is a new genome assembly that maintains the biological structure of the reference, but contains the actual sequence data from the draft genome. This step is crucial for making the draft genome biologically interpretable and useful for downstream analyses like gene prediction, synteny mapping, or evolutionary comparison.

Step-3: Quality Assessment of the Original Draft

This step gives us a baseline measurement of the quality of the draft assembly before any improvement is attempted. It's essential for scientifically validating the impact of the scaffolding process. That's why we run QUAST on the unscaffolded draft genome of *A. lyrata*.

- **Number of Contigs / Scaffolds:**

A **contig** is a continuous stretch of DNA sequence with no gaps — assembled from overlapping reads. A **scaffold** is a higher-order structure, made by joining contigs together using information like paired-end read distance or alignment to a reference; it may contain gaps (usually filled with N's). The **fewer the contigs/scaffolds**, the more continuous the genome. A high number suggests fragmentation — possibly due to low sequencing coverage, repetitive regions, or complex genome structure.

- **Total Length of Assembly:**

This is the **cumulative length of all contigs or scaffolds** in the assembly. For a well-assembled genome, this should be **close to the expected genome size** (e.g., ~135 Mb for *A. thaliana*). If the total length is **much smaller**, it may mean incomplete assembly (missing regions). If it's **much larger**, it could indicate contamination, misassemblies, or duplicated regions (overassembly).

- **N50:**

It is the length of the shortest contig/scaffold such that 50% of the total assembly is in contigs/scaffolds of that size or longer. A **higher N50** means longer contiguous sequences and a better assembly. A **low N50** indicates high fragmentation.

- **L50:**

This is the number of contigs/scaffolds that make up 50% of the genome, corresponding to the N50 value. A **lower L50 is better** — fewer large pieces are covering more of the genome.

- **Largest Contig / Scaffold:**

A long largest contig means some part of the genome has been assembled very well — often a gene-rich or non-repetitive region. If scaffolding has been performed (e.g., using RagTag), this number usually increases because small contigs are linked into bigger ones.

- **GC Content:**

GC content is the **percentage of G and C bases** in the assembled genome. It is mostly used as a **sanity check** to make sure the genome resembles what's expected biologically.

- **Number of Ns (Ambiguous Bases):**

It gives two main insights – number of **gaps** exist and how much of the genome is **unknown** or unsequenced. A higher number of Ns after scaffolding is normal, but too many may mean low-quality reference alignment or misassembly.

Step-4: Quality Assessment of the Scaffolded Genome (QUAST)

The purpose here is to compare the metrics from before and after scaffolding. Ideally, we will observe that the N50 has increased, indicating that contigs have been successfully merged into longer scaffolds. We should also see a reduction in the total number of contigs, and possibly a longer total assembly size (as some gaps are now filled). This second QUAST run allows us to assess how effective RagTag was in improving the assembly.

Step-5: Comparative QUAST Analysis

This comparative mode of QUAST allows for a direct evaluation of how each assembly performs across all metrics. Comparative analysis is important because it removes subjectivity — we no longer rely on memory or flipping between reports. It also helps highlight trade-offs: for example, a scaffolded genome may have longer contigs but more Ns (gaps).

Arabidopsis Genomes

Reference genome (*A. thaliana*): GCF_000001735.1 (replaced by - GCF_000001735.4)

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/735/GCF_000001735.4_TAIR10.1/GCF_000001735.4_TAIR10.1_genomic.fna.gz

Assembly statistics

	RefSeq
Genome size	119.7 Mb
Total ungapped length	119.5 Mb
Number of chromosomes	7
Number of scaffolds	7
Scaffold N50	23.5 Mb
Scaffold L50	3
Number of contigs	102
Contig N50	11.2 Mb
Contig L50	5
GC percent	36
Assembly level	Chromosome
View sequences	view RefSeq sequences

Draft Genome (*Arabidopsis lyrata*): GCA_963993105.1

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/963/993/105/GCA_963993105.1_pu6_sup/GCA_963993105.1_pu6_sup_genomic.fna.gz

Assembly statistics

	GenBank
Genome size	548.5 Mb
Total ungapped length	530 Mb
Number of scaffolds	63,864
Scaffold N50	58.6 kb
Scaffold L50	88
Number of contigs	116,658
Contig N50	7.4 kb
Contig L50	18,838
GC percent	36.5
Genome coverage	171x
Assembly level	Scaffold

Avian Genomes

Reference genome (*Aquila chrysaetos chrysaetos* / Golden Eagle): GCF_900496995.1

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/900/496/995/GCF_900496995.1_bAquChr1.2/GCF_900496995.1_bAquChr1.2_genomic.fna.gz

Assembly statistics

	RefSeq
Genome size	1.2 Gb
Total ungapped length	1.2 Gb
Number of chromosomes	27
Number of scaffolds	142
Scaffold N50	46.9 Mb
Scaffold L50	9
Number of contigs	333
Contig N50	21.9 Mb
Contig L50	19
GC percent	42
Genome coverage	60x
Assembly level	Chromosome
View sequences	view RefSeq sequences

Draft genome (*Falco cherrug* / Saker Falcon): GCF_000337975.1

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/337/975/GCF_000337975.1_F_cherrug_v1.0/GCF_000337975.1_F_cherrug_v1.0_genomic.fna.gz

Assembly statistics

	RefSeq
Genome size	1.2 Gb
Total ungapped length	1.2 Gb
Number of scaffolds	5,863
Scaffold N50	4.2 Mb
Scaffold L50	84
Number of contigs	75,898
Contig N50	31.3 kb
Contig L50	10,949
GC percent	41.5
Genome coverage	147x
Assembly level	Scaffold
View sequences	

Workflow Pipeline

```
#!/bin/bash

set -e # Exit immediately on error

### GENOME IMPROVEMENT PIPELINE (Run from Genome_improvement directory)

qst=~/.NGS/Packages/quast/quast.py

# Step-1: Download genomes
read -p "Enter URL for reference genome (.fna.gz): " ref_url
read -p "Enter URL for draft genome (.fna.gz): " draft_url
wget -q --show-progress "$ref_url"
wget -q --show-progress "$draft_url"

# Step-2: Extract filenames from URLs
ref_gz=$(basename "$ref_url")
draft_gz=$(basename "$draft_url")
ref_fa="${ref_gz%.gz}" # Remove .gz extension to get uncompressed reference filename
draft_fa="${draft_gz%.gz}"

# Step-3: Uncompress genomes
gunzip -c "$ref_gz" > "$ref_fa"
gunzip -c "$draft_gz" > "$draft_fa"

# Step-4: Correct the draft genome using reference
echo "Running RagTag correction..."
ragtag.py correct -u "$ref_fa" "$draft_fa" -o ragtag_correction

# Step-5: Check if correction succeeded
corrected_fa="ragtag_correction/ragtag.correct.fasta" # Set path to corrected output
if [ ! -s "$corrected_fa" ]; then # If corrected file doesn't exist or is empty
    echo "ERROR: Correction failed or produced empty output. Exiting."
    exit 1 # Exit script with error
fi

# Step-6: Scaffold corrected genome using reference
echo "Running RagTag scaffolding..."
ragtag.py scaffold -u "$ref_fa" "$corrected_fa" -o ragtag_scaffold

# Step-7: QUAST on original draft genome
echo "Initial Quality Assessment..."
python3 "$qst" "$draft_fa" -o quast_pre_ragtag

# Step-8: QUAST on scaffolded genome
echo "Improved Assembly Assessment..."
python3 "$qst" "ragtag_scaffold/ragtag.scaffold.fasta" -o quast_post_ragtag

# Step-9: QUAST on comparative mode
echo "Comparative Assembly Assessment..."
python3 "$qst" "$draft_fa" "ragtag_scaffold/ragtag.scaffold.fasta" -o quast_comparative

echo "Genome improvement successfully done."
```

```
(ragtag_env) ibab@LAPTOP-BVSTVK8Q:~/NGS/Genome_improvement/ragtag_correction$ ls -lh
total 553M
-rw-r--r-- 1 ibab ibab 4.3M Aug 3 17:06 ragtag.correct.agp
-rw-r--r-- 1 ibab ibab 20M Aug 3 16:54 ragtag.correct.asm.paf
-rw-r--r-- 1 ibab ibab 824 Aug 3 16:54 ragtag.correct.asm.paf.log
-rw-r--r-- 1 ibab ibab 0 Aug 3 16:54 ragtag.correct.err
-rw-r--r-- 1 ibab ibab 526M Aug 3 17:06 ragtag.correct.fasta
-rw-r--r-- 1 ibab ibab 3.4M Aug 3 17:07 ragtag.correct.fasta.fai
(ragtag_env) ibab@LAPTOP-BVSTVK8Q:~/NGS/Genome_improvement/ragtag_correction$ cd .
(ragtag_env) ibab@LAPTOP-BVSTVK8Q:~/NGS/Genome_improvement$ cd ragtag_scaffold/
(ragtag_env) ibab@LAPTOP-BVSTVK8Q:~/NGS/Genome_improvement/ragtag_scaffold$ ls -lh
total 555M
-rw-r--r-- 1 ibab ibab 5.4M Aug 3 17:32 ragtag.scaffold.agp
-rw-r--r-- 1 ibab ibab 24M Aug 3 17:07 ragtag.scaffold.asm.paf
-rw-r--r-- 1 ibab ibab 827 Aug 3 17:07 ragtag.scaffold.asm.paf.log
-rw-r--r-- 1 ibab ibab 67K Aug 3 17:32 ragtag.scaffold.confidence.txt
-rw-r--r-- 1 ibab ibab 0 Aug 3 17:06 ragtag.scaffold.err
-rw-r--r-- 1 ibab ibab 526M Aug 3 17:33 ragtag.scaffold.fasta
-rw-r--r-- 1 ibab ibab 122 Aug 3 17:33 ragtag.scaffold.stats
```

Arabidopsis
(File Size)







```

ibab@LAPTOP-BVSTVK8Q:~/NGS/Genome_improvement$ cd Avian_genome/
ibab@LAPTOP-BVSTVK8Q:~/NGS/Genome_improvement/Avian_genome$ ls -lh
total 3.0G
-rw-r--r-- 1 ibab ibab 1.2G Aug 3 19:55 GCF_000337975.1_F_cherrug_v1.0_genomic.fna
-rw-r--r-- 1 ibab ibab 209K Aug 3 22:33 GCF_000337975.1_F_cherrug_v1.0_genomic.fna.fai
-rw-r--r-- 1 ibab ibab 351M Jun 24 2019 GCF_000337975.1_F_cherrug_v1.0_genomic.fna.gz
-rw-r--r-- 1 ibab ibab 1.2G Aug 3 19:54 GCF_900496995.1_bAquaChr1.2_genomic.fna
-rw-r--r-- 1 ibab ibab 365M Aug 7 2019 GCF_900496995.1_bAquaChr1.2_genomic.fna.gz
drwxr-xr-x 4 ibab ibab 4.0K Aug 4 03:34 quast_comparative
drwxr-xr-x 4 ibab ibab 4.0K Aug 4 03:35 quast_post_ragtag
drwxr-xr-x 4 ibab ibab 4.0K Aug 4 03:35 quast_pre_ragtag
drwxr-xr-x 2 ibab ibab 4.0K Aug 4 03:35 ragtag_correction
drwxr-xr-x 2 ibab ibab 4.0K Aug 4 03:36 ragtag_scaffold
ibab@LAPTOP-BVSTVK8Q:~/NGS/Genome_improvement/Avian_genome$ cd ragtag_scaffold/
ibab@LAPTOP-BVSTVK8Q:~/NGS/Genome_improvement/Avian_genome/ragtag_scaffold$ ls -lh
total 1.1G
-rw-r--r-- 1 ibab ibab 528K Aug 3 23:33 ragtag.scaffold.agp
-rw-r--r-- 1 ibab ibab 1.1M Aug 3 23:32 ragtag.scaffold.asm.paf
-rw-r--r-- 1 ibab ibab 906 Aug 3 23:32 ragtag.scaffold.asm.paf.log
-rw-r--r-- 1 ibab ibab 44K Aug 3 23:33 ragtag.scaffold.confidence.txt
-rw-r--r-- 1 ibab ibab 0 Aug 3 22:33 ragtag.scaffold.err
-rw-r--r-- 1 ibab ibab 1.1G Aug 3 23:33 ragtag.scaffold.fasta
-rw-r--r-- 1 ibab ibab 120 Aug 3 23:33 ragtag.scaffold.stats
ibab@LAPTOP-BVSTVK8Q:~/NGS/Genome_improvement/Avian_genome/ragtag_scaffold$ cd ..
ibab@LAPTOP-BVSTVK8Q:~/NGS/Genome_improvement/Avian_genome$ cd quast_comparative/
ibab@LAPTOP-BVSTVK8Q:~/NGS/Genome_improvement/Avian_genome/quast_comparative$ ls -lh
total 528K
drwxr-xr-x 2 ibab ibab 4.0K Aug 4 03:21 basic_stats
-rw-r--r-- 1 ibab ibab 53K Aug 4 03:27 icarus.html
drwxr-xr-x 2 ibab ibab 4.0K Aug 4 03:34 icarus_viewers
-rw-r--r-- 1 ibab ibab 2.5K Aug 4 03:27 quast.log
-rw-r--r-- 1 ibab ibab 433K Aug 4 03:27 report.html
-rw-r--r-- 1 ibab ibab 1.6K Aug 4 03:26 report.tex
-rw-r--r-- 1 ibab ibab 768 Aug 4 03:26 report.tsv
-rw-r--r-- 1 ibab ibab 2.1K Aug 4 03:26 report.txt
-rw-r--r-- 1 ibab ibab 1.4K Aug 4 03:26 transposed_report.tex
-rw-r--r-- 1 ibab ibab 768 Aug 4 03:26 transposed_report.tsv
-rw-r--r-- 1 ibab ibab 1.5K Aug 4 03:26 transposed_report.txt

```

Avian
(File Size)

Results & Interpretation (Comparative Analysis)








 ragtag.correct.agp	03-08-2025 17:06	AGP File	4,344 KB
 ragtag.correct.asm.paf	03-08-2025 16:54	PAF File	20,272 KB
 ragtag.correct.asm.paf	03-08-2025 16:54	Text Document	1 KB
 ragtag.correct.err	03-08-2025 16:54	ERR File	0 KB
 ragtag.correct	03-08-2025 17:06	FASTA File	5,37,630 KB
 ragtag.correct.fasta.fai	03-08-2025 17:07	FAI File	3,465 KB

ragtag.correct.fasta: This file contains the corrected draft genome after aligning it to the reference genome. Used to compare with the original draft to check whether misassemblies or errors were fixed.

ragtag.correct.agp: Describes how the corrected contigs map to the reference.

ragtag.correct.asm.paf: Pairwise alignment format (used internally by RagTag).

ragtag.correct.fasta.fai: Index for fast access to sequences.

Name	Date modified	Type	Size
 ragtag.scaffold.agp	03-08-2025 17:32	AGP File	5,489 KB
 ragtag.scaffold.asm.paf	03-08-2025 17:07	PAF File	24,533 KB
 ragtag.scaffold.asm.paf	03-08-2025 17:07	Text Document	1 KB
 ragtag.scaffold.confidence	03-08-2025 17:32	Text Document	67 KB
 ragtag.scaffold.err	03-08-2025 17:06	ERR File	0 KB
 ragtag.scaffold	03-08-2025 17:33	FASTA File	5,38,177 KB
 ragtag.scaffold.stats	03-08-2025 17:33	STATS File	1 KB

ragtag.scaffold.fasta: This is the **final improved genome** after correction **and** scaffolding. It integrates the contigs and arranges them based on reference genome structure. Used for **final downstream analyses**.

ragtag.scaffold.agp: How contigs were ordered and oriented.

ragtag.scaffold.confidence: Confidence scores for scaffold joins.

ragtag.scaffold.stats: Assembly statistics (e.g., number of scaffolds, N50).

.err files are empty (0 KB), indicating no critical errors during run.

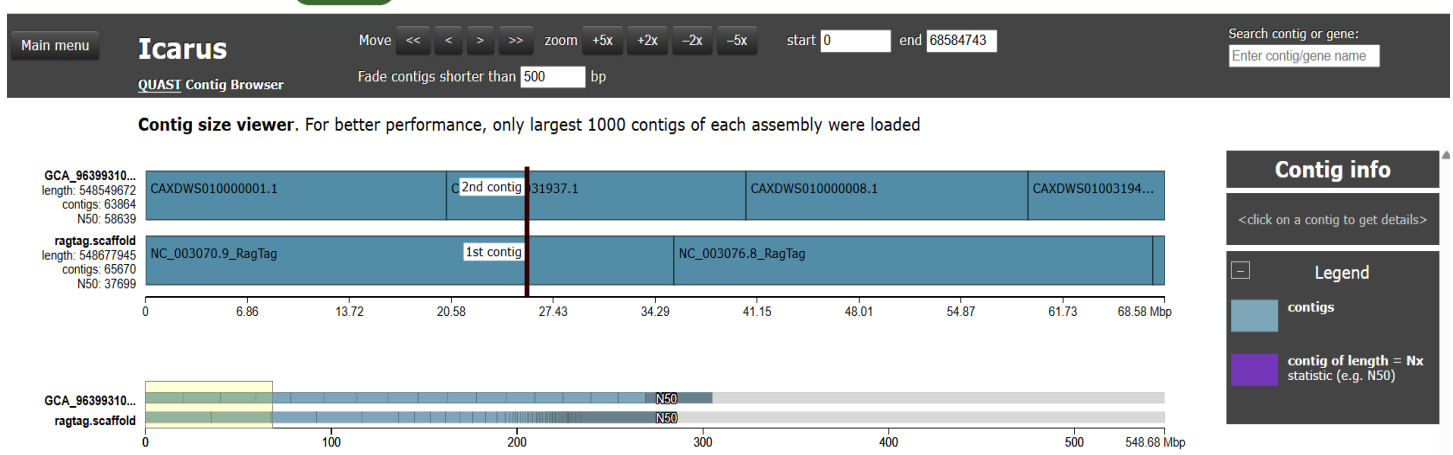
For Arabidopsis –

QUAST		
Quality Assessment Tool for Genome Assemblies		
03 August 2025, Sunday, 18:10:28		
View in Icarus contig browser		
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).		
<div> <div>Worst</div> <div>Median</div> <div>Best</div> </div> <input checked="" type="checkbox"/> Show heatmap		
Statistics without reference	GCA_963993105.1_pu6_sup_genom...	ragtag.scaffold
# contigs	63 864	65 670
# contigs (≥ 0 bp)	63 864	65 733
# contigs (≥ 1000 bp)	63 864	65 576
# contigs (≥ 5000 bp)	15 428	16 950
# contigs (≥ 10000 bp)	4965	5607
# contigs (≥ 25000 bp)	889	1217
# contigs (≥ 50000 bp)	156	444
Largest contig	20 232 408	35 538 529
Total length	548 549 672	548 677 945
Total length (≥ 0 bp)	548 549 672	548 698 972
Total length (≥ 1000 bp)	548 549 672	548 605 094
Total length (≥ 5000 bp)	431 703 796	431 288 426
Total length (≥ 10000 bp)	361 291 463	354 640 264
Total length (≥ 25000 bp)	302 502 590	291 783 798
Total length (≥ 50000 bp)	278 014 435	265 860 060
N50	58 639	37 699
N90	2862	2850
auN	8 358 891	7 925 034
L50	87	641
L90	31 544	33 237
GC (%)	36.35	36.35
Per base quality		
# N's per 100 kbp	3375.1	3401.52
# N's	18 514 090	18 663 380

Interpretation:

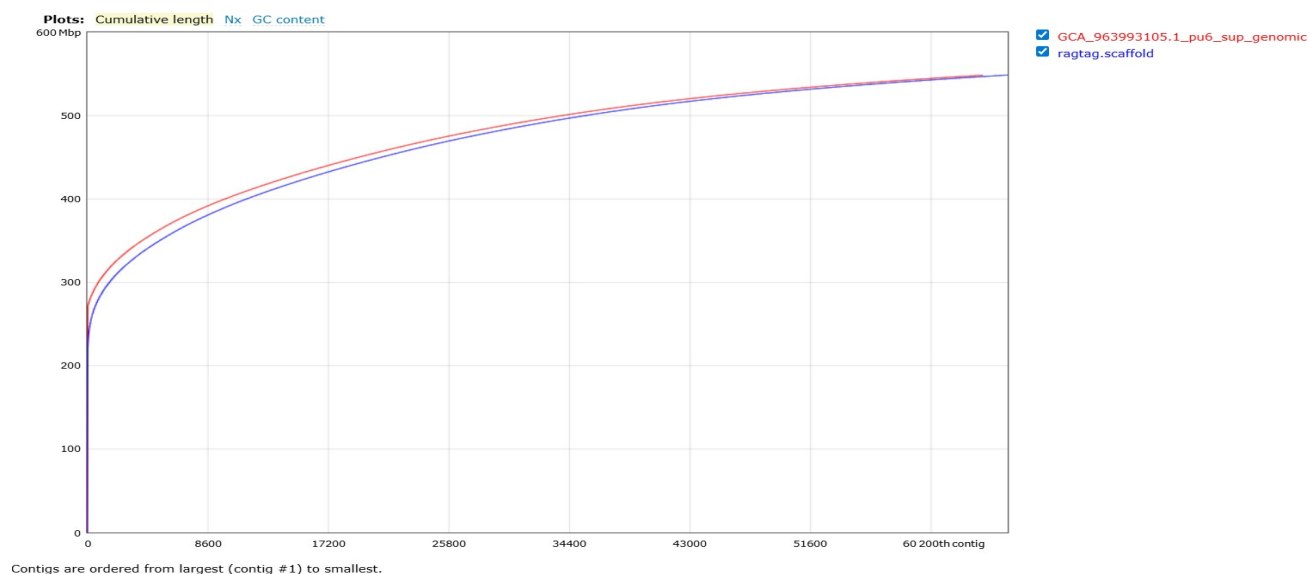
- ❖ The pre-RagTag assembly contained **63,864 contigs**, which slightly **increased to 65,670** in the post-RagTag assembly. This might initially suggest fragmentation; however, it is a typical result of RagTag introducing conservative breaks at suspected misassemblies and inserting gaps to scaffold the genome based on a reference.
- ❖ In the pre-RagTag assembly, the **largest contig** was 20.23 Mb, whereas after RagTag scaffolding, this increased dramatically to **35.54 Mb - an approx. 75% improvement**. This change indicates that RagTag effectively utilized the reference genome to link and order contigs, producing longer scaffolds that likely correspond to entire chromosomal regions.

- ❖ The **total assembly size remained highly consistent** between the two versions, with the pre-RagTag assembly at 548,549,672 bp and the post-RagTag assembly at 548,698,972 bp.
- ❖ The **GC content remained constant at 36.35%** in both assemblies, further confirming that no compositional bias was introduced during scaffolding.
- ❖ Interestingly, the **N50**—a commonly used metric indicating assembly contiguity—**decreased from 58,639 bp pre-RagTag to 37,699 bp post-RagTag**. At first glance, this reduction might appear to indicate a drop in assembly quality. However, RagTag intentionally breaks potentially chimeric or misassembled regions and inserts gaps (Ns) to more accurately reflect genome structure. Supporting this, the **auN metric**, which takes both contig length and number into account, **decreased only modestly from 8.36 Mb to 7.93 Mb**.
- ❖ The **no. of Ns per 100 kbp slightly increased from 3375.1 to 3401.52**, and the **total number of Ns rose from 18.51 million to 18.66 million**. These small increases are typical consequences of scaffolding, where gaps are inserted to bridge contigs based on the reference layout.



Interpretation:

In the **pre-RagTag assembly** (GCA_963993105.1), **contigs are shorter and more fragmented**, with multiple large contigs spread across the same genomic region. This is evident from the presence of several independent contigs (e.g., **CAXDWS010000001.1**, **CAXDWS010000008.1**) displayed side-by-side, indicating a lack of connectivity. In contrast, the **post-RagTag assembly** (ragtag.scaffold) shows a significantly **longer and contiguous scaffold** (e.g., **NC_003070.9_RagTag**) covering the same region, suggesting successful stitching of shorter contigs into a more complete pseudochromosome. While **N50 is slightly reduced** due to introduced gaps, the larger and fewer scaffolds in the RagTag output indicate a clearer and more accurate representation of chromosomal structure.



Interpretation:

The cumulative length plot compares the original genome assembly (**GCA_96393105.1, red**) with the RagTag-scaffolded version (**ragtag.scaffold, blue**). The **x-axis** shows **contigs** ordered from largest to smallest, and the **y-axis** represents the **cumulative genome length**. The red curve rises more steeply at the beginning, indicating that the original assembly had longer top-ranked contigs contributing more quickly to the total length. In contrast, the blue curve rises more gradually, showing that the RagTag scaffolding produced a more balanced distribution of contig lengths, though with fewer extremely long contigs early on. Despite these differences, **both assemblies reach a similar total genome size (~550 Mbp)**. This suggests that **RagTag preserved overall genome completeness** while redistributing contig lengths for structural refinement.

For Avian –

QUAST

Quality Assessment Tool for Genome Assemblies

04 August 2025, Monday, 03:27:18

[View in Icarus contig browser](#)

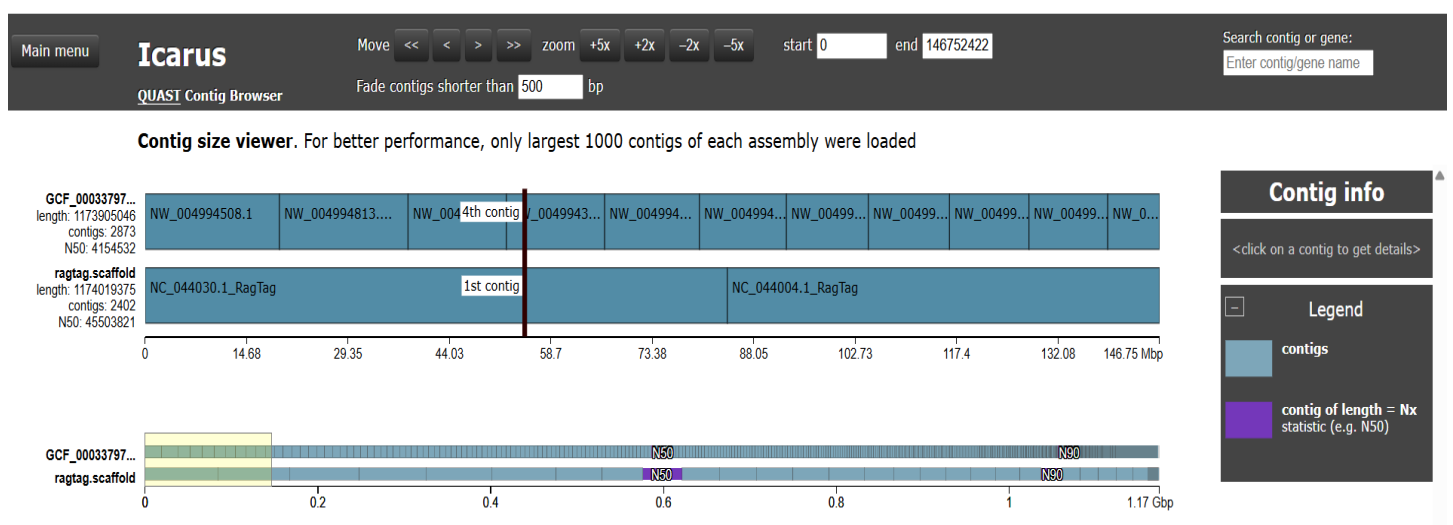
All statistics are based on contigs of size ≥ 500 bp, unless otherwise noted (e.g., "# contigs (≥ 0 bp)" and "Total length (≥ 0 bp)" include all contigs).

Worst Median Best ☒ Show heatmap

Statistics without reference	GCF_000337975.1_F_cherrug_v1....	ragtag.scaffold
# contigs	2873	2402
# contigs (≥ 0 bp)	5863	5358
# contigs (≥ 1000 bp)	1563	1131
# contigs (≥ 5000 bp)	819	410
# contigs (≥ 10000 bp)	731	271
# contigs (≥ 25000 bp)	653	141
# contigs (≥ 50000 bp)	585	82
Largest contig	19 410 955	84 206 865
Total length	1 173 905 046	1 174 019 375
Total length (≥ 0 bp)	1 174 811 715	1 174 913 915
Total length (≥ 1000 bp)	1 172 978 694	1 173 121 150
Total length (≥ 5000 bp)	1 171 627 445	1 171 764 487
Total length (≥ 10000 bp)	1 171 049 556	1 170 819 807
Total length (≥ 25000 bp)	1 169 806 192	1 168 772 104
Total length (≥ 50000 bp)	1 167 247 145	1 166 757 195
N50	4 154 532	45 503 821
N90	988 184	23 996 820
auN	5 298 155	53 152 201
L50	84	9
L90	295	22
GC (%)	41.7	41.7
Per base quality		
# N's per 100 kbp	2028.95	2037.46
# N's	23 817 946	23 920 146

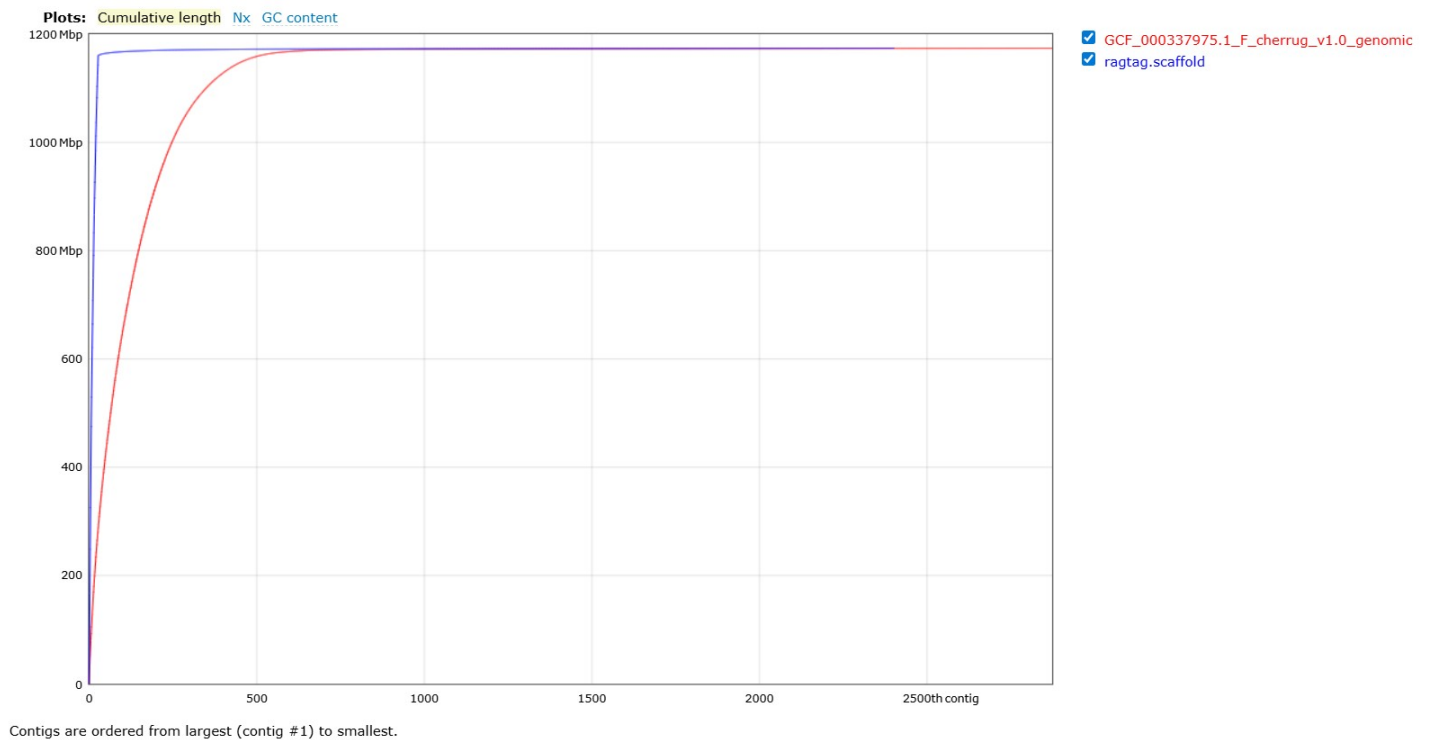
Interpretation:

The QCAST results clearly indicate that RagTag scaffolding significantly enhanced the avian genome assembly's structural quality. They have **nearly identical total lengths** (~1.174 Gbp) and **GC content** (41.7%), confirming no sequence loss or change in base composition. The number of **contigs decreased from 2,873 to 2,402**, reflecting **reduced fragmentation**. Most notably, the **N50 value increased sharply to 45.5 Mb**, while the **L50 dropped from 84 to just 9**, showing that **longer scaffolds now make up a larger portion of the genome**. The **largest contig length** also rose from 19.4 Mb to **84.2 Mb**, highlighting improved scaffold continuity. Although there was a **slight increase in ambiguous bases** (N's per 100 kbp rose from 2028.95 to 2037.46), this is expected during reference-based scaffolding.



Interpretation:

In the top panel, blue blocks represent the largest contigs arranged from left to right by size. The reference assembly (top row) contains many shorter, fragmented contigs, while the RagTag assembly (bottom row) has longer, more continuous scaffolds. Notably, the first RagTag contig spans approximately 58.7 Mbp, indicating successful scaffolding of smaller contigs into a single larger unit. The middle line shows the genomic scale, allowing direct size comparison. In the bottom Nx panel, the cumulative length distribution further highlights assembly quality. The RagTag assembly reaches N50 and N90 values faster, meaning fewer contigs contribute more significantly to genome length—evidence of improved contiguity. The sidebar indicates color coding for contigs (blue) and N-statistics (purple). Overall, this visualization clearly demonstrates that the RagTag assembly is less fragmented and structurally more complete than the original.



Interpretation:

The cumulative length plot clearly shows that the **ragtag.scaffold assembly** (blue line) **achieves the full genome length with far fewer and larger contigs** than the original GCF_000337975.1 assembly (red line). The blue curve **rises steeply and plateaus early**, meaning just a few large scaffolds account for most of the genome. In contrast, the **red curve** builds up more gradually and continues rising across many more contigs, **indicating fragmentation**. Despite both assemblies having similar total genome sizes (~1.17 Gbp), the distribution of sequence across contigs is much more efficient in the RagTag assembly. This confirms that RagTag has substantially improved the continuity of the genome, reducing fragmentation and generating a higher-quality assembly.