

Course 1: High Throughput Sequencing: Genome assembly

Course Provider:

Dr. Tam Tran, USTH

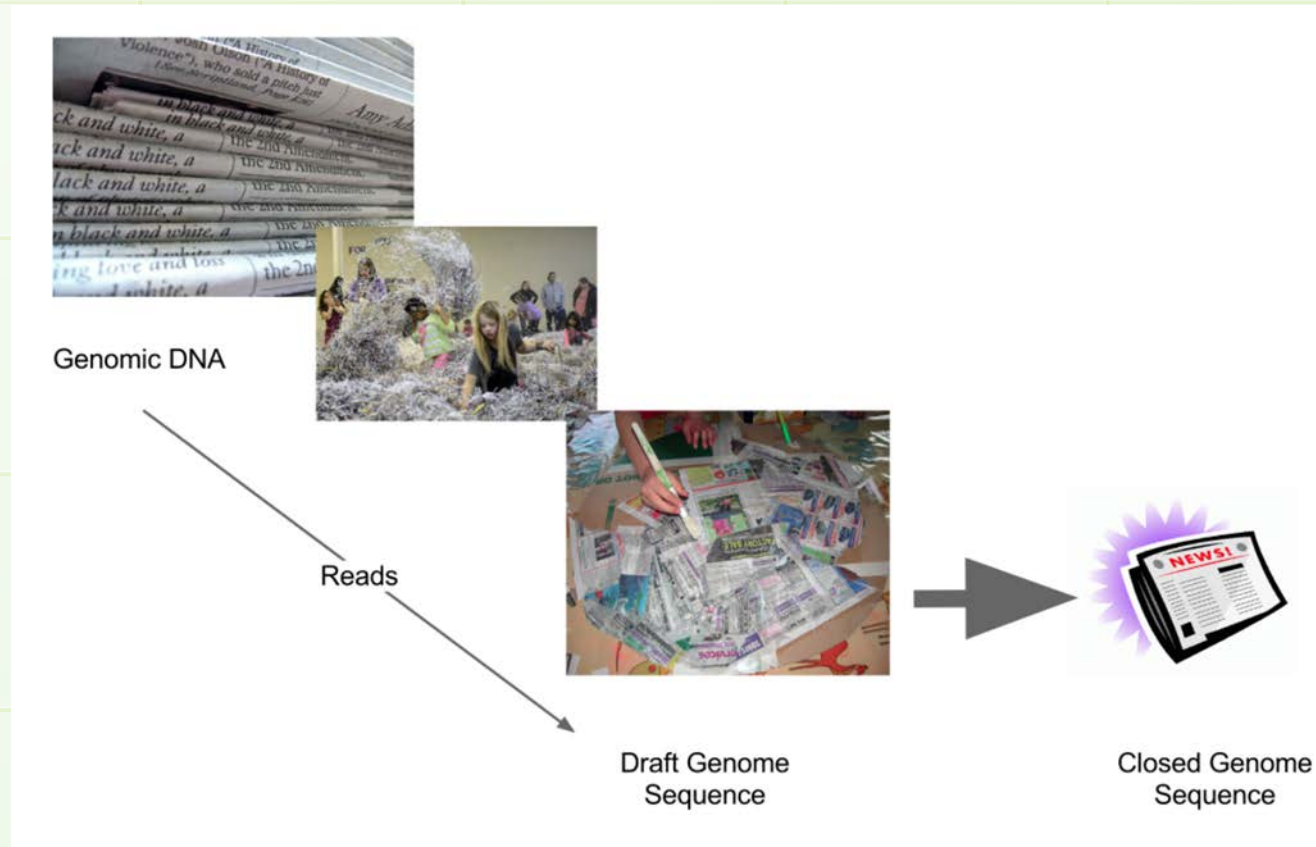
Mathis Arnal, IRD, Chiang Mai University

Bioinformatics for Infectious Disease Ecology Research (BInDER)

Outline

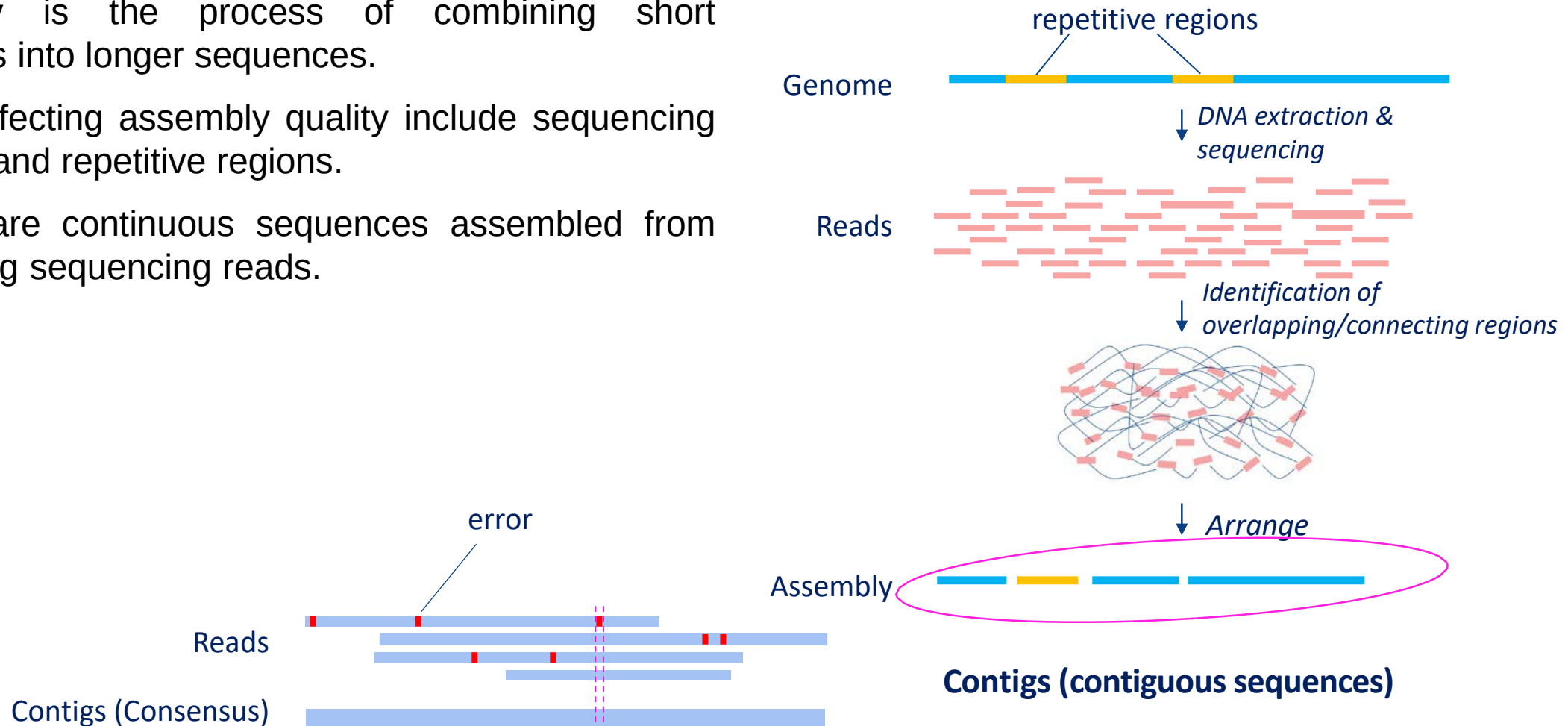
1. Introduction to Genome assembly
2. De novo genome assembly
3. Practical: Short-read assembly with SPAdes
4. Practical: Hybrid Assembly for short- and long-read data

1. Introduction to Genome assembly

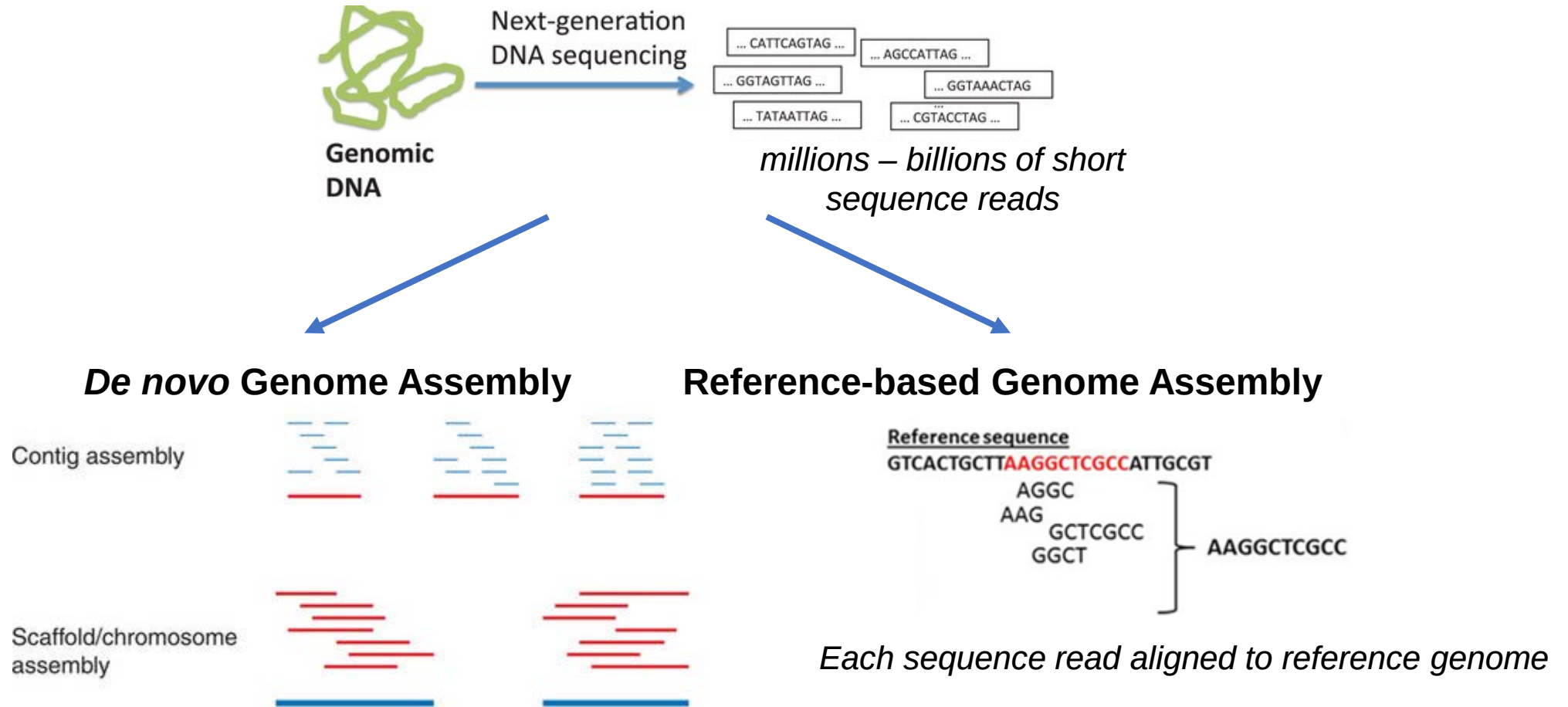


Assembly

- **Assembly** is the process of combining short sequences into longer sequences.
- Factors affecting assembly quality include sequencing coverage and repetitive regions.
- **Contigs** are continuous sequences assembled from overlapping sequencing reads.



Genome assembly approaches



Comparison of two genome assembly approaches

De novo assembly	Reference-based Genome Assembly
<i>Advantages:</i> ☑ Does not require a reference genome ☑ Can identify novel sequences and variants	<i>Advantages:</i> ☑ Requires lower sequencing coverage ☑ Fast and computationally efficient ☑ High accuracy when a suitable reference is available
<i>Disadvantages:</i> ✗ Requires high sequencing depth ✗ Computationally resource ✗ May produce gaps and assembly errors	<i>Disadvantages:</i> ✗ Depends on the reference genome ✗ May miss novel sequences

2. De novo genome assembly

De novo genome assembly approaches

- Overlap Layout Consensus
- De Bruijn Graphs

Workflow:

- ## Disadvantages:

Example: sequencing data containing 1 million reads

→ Number of read pairs to compare: 1 million $\times$ 1 million = 1 trillion comparisons.



Consensus

9

De Bruijn graphs are a key algorithmic technique in *de novo* genome assembly

Genome: ACGATGCTCAGACCCCCCCC

Short reads: ACGATGCTCAGA CTCAGACCC AGACCCC CCCCCC

k-mers:

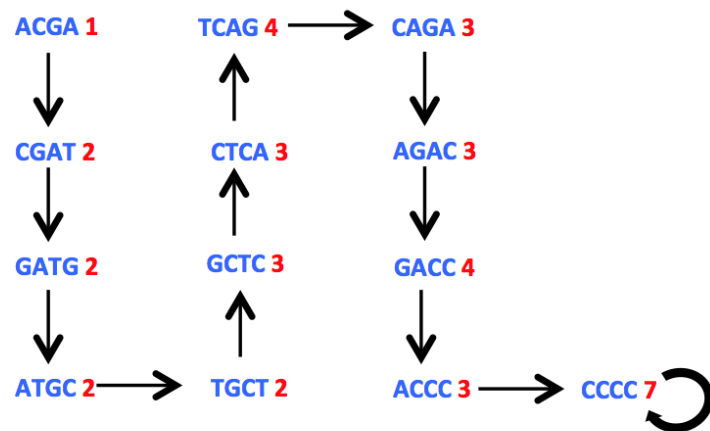
ACGAT
CGATG
GATGC
ATGCT
TGCTC
GCTCA
CTCAG
TCAGA

CTCAG
TCAGA
CAGAC
AGACC
GACCC

AGACC
GACCC
ACCCC

CCCCC
CCCCC
CCCCC

De Bruijn graph:



Assembled Contigs: ACGATGCTCAGACCCC

Disadvantages:

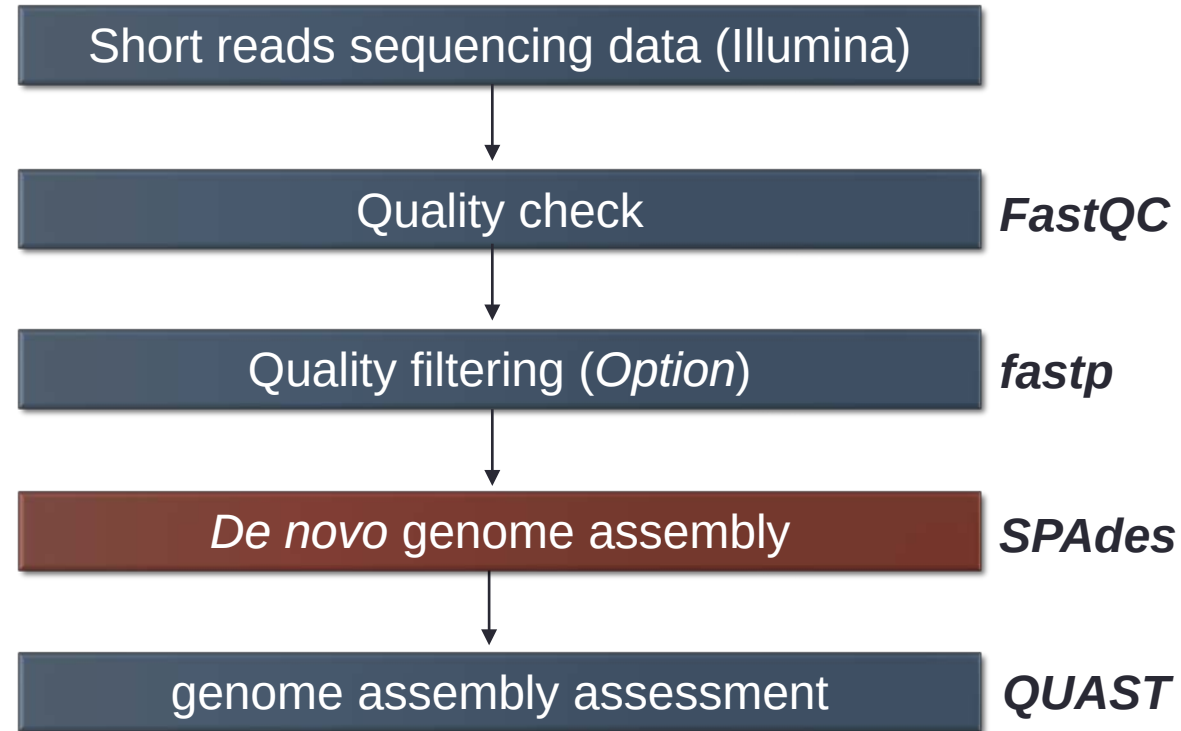
Difficult to resolve repetitive regions and sequencing errors in the genome

A *kmer* is a sub-string with length *k* extracted from a given DNA sequence

Tools used for genome assembly

Short Read	Long Read	Hybrid
SPAdes	Canu	Unicycler
ABYSS	Flye	MaSuRCA
SOAPdenovo2	NextDenovo	OPERA-MS
...	...	...

3. *Practical*: Short-read assembly with SPAdes



Short-read assembly with SPAdes

Tool: SPAdes

Input: Raw FASTQ files or paired-end data after quality filtering

Tool Parameters:

Operation mode: *Assembly and error correction*

Single-end or paired-end short-reads: *Paired-end: list of dataset pairs*

FASTA/FASTQ file(s): collection: *Trimmomatic on collection: paired*

Set coverage cutoff option: *auto*

Select optional output file(s): e.g. add additional output *Contigs stats*

Click **Run Tool**

SPAdes
genome assembler for genomes of regular and single-cell projects
(Galaxy Version 4.3.0+galaxy2)

Tool Parameters

Operation mode *

Assembly and error correction

To run read error correction, reads should be in FASTQ format.

Single-end or paired-end short-reads

Paired-end: list of dataset pairs

It assumes that all samples belong to the same library. If you want to use samples from two different libraries, include the second library as additional set of short-reads.

FASTA/FASTQ file(s): collection *

18: fastp on collection 1: Paired-end output

accepted formats ▼

Type of paired-reads *

Default (--pe)

In paired-end sequencing, the library preparation yields a set of fragments, and the machine sequences each fragment from both ends. In mate-pair sequencing, the library preparation yields two fragments that are distal to each other in the genome and in the opposite in orientation to that of a mate-paired fragment.

Set coverage cutoff option

Auto

When set to 'auto' SPAdes automatically computes coverage threshold using conservative strategy (--cov-cutoff)

Select k-mer detection option

Auto

kmer default: 21, 33, 55

If --sc is set the default values are 21,33,55. For multicell datasets K values are automatically selected using maximum read length. Comma-separated list, all values must be odd, less than 128 and listed in ascending order. (-k)

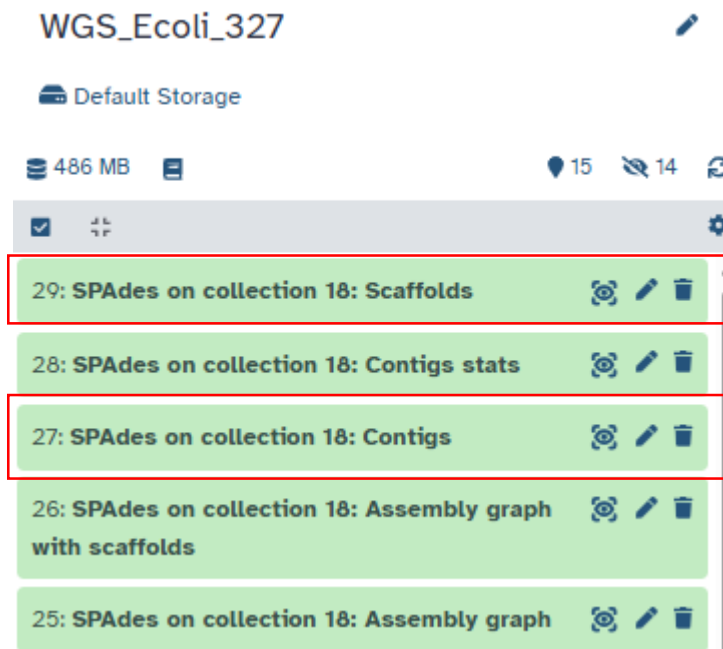
Select optional output file(s) *

Assembly graph x Assembly graph with scaffolds x Contigs x Scaffolds x Contigs stats x

switch to column select ▼

Only shown in history if selected here and generated by the specific run.

SPAdes output data



- The assembled genome data are stored in a FASTA file
 - ✓ *SPAdes scaffolds (fasta)*: scaffold sequences
 - ✓ *SPAdes contigs (fasta)*: contig sequences
- FASTA file extensions: *.fa*, *.fna*, hoặc *.fasta*
- Each contig/scaffold consists of:
 - ✓ A header (contig/scaffold name)
 - ✓ A sequence line.

```
>NODE 1 length 933 cov 35.767418
ACTATAAATTCAACTTTGCAACAGAACCATCTAATCTTCAACAACTGGCCCGTTTGTG
AACTACTCTTTAATAAAATAATTTTCCGGTCCCAATTCCACATTGCAATAATAGAAAAT
CCATCTTCATCGGCTTTTCGTGTCATCTGTATGAATCAAATCGCCTTCTTCTGTGTCA
TCAAGGTTTAATTTTTTATGTATTCTTTTAAACAAACCACCATAGGAGATTAACCTTTA
CGGTGTAAACCTTCCTCCAAATCAGACAAACGTTACAAATCTTTTCTTCATCATCGGTC
ATAAAATCCGTATCCTTTACAGGATATTTGTCAGTTTCGTCAATTGCCGATTGTATATCC
GATTTATATTTATTTTTCGGTCGAATCATTGAACTTTTACATTTGGATCATAGTCTAAT
TTCATTGCCTTTTTCCAAAATTGAATCCATTGTTTTGATTACGATGTTTCTGTATTC
TTAAATAAGTTGGTTCCACACATACCAATACATGCATGTGCTGATTATAAGAATTATCT
TTATTATTTATTGTCACTTCGGTTGCACGCATAAAACCAACAAGATTTTATTAATTTT
TTATATTGCATCATTCGGCGAAATCCTTGAGCCATATCTGACAACTCTTATTTAATTCT
TCGCCATCATAAACATTTTAACTGTTATTGTGAGAAACAACCAACGAAGTTGGCTTT
TGTTTAATAAATTCAGCAACAACCTTTTGTGACTGAATGCCATGTTTCATTGCTCTCCTC
CAGTTGCACATTGGACAAAGCCTGGATTTACAAAACCACTCGATACAACTTTCTTTTCG
CCTGTTTCACGATTTTGTGTTTATACTCTAATATTTTCAGCACAATCTTTTACTCTTCAGCC
TTTTTAAATTCAAGAAATATGCAGAAGTTCAAAGTAATCAACATTAGCGATTTTCTTTTCT
C
>NODE 2 length 70 cov 16.871429
AATGATATCGATAAAGACACAATGAGTATCGTACGACATGAGCCGATTGGCGTCGTAGGT
GCTGTTGTGCTTGGAACTTCCCAATGCTATTAGCTGC
>NODE 3 length 61 cov 15.213115
TTAGTACAATCAGGGCAATTGCTAATGGATATAAGAACATTAACACTGGGACTGAGTACA
TAATAATTTTAGTTAAGCCAACATTCGCG
```

Question:

How many contigs/scaffolds were assembled?

Quality Assessment Tool (QUAST)

Background:

QUAST can evaluate assemblies using reference genomes, as well as without reference genomes. QUAST produces detailed reports, tables and plots which show the different aspects of assemblies.

(QUAST: quality assessment tool for genome assemblies, *Bioinformatics*, Volume 29, Issue 8, 15 April 2013, Pages 1072–1075, <https://doi.org/10.1093/bioinformatics/btt086>)

Perform QUAST on Galaxy

Tool: Quast Genome assembly Quality

Tool Parameters:

- *Assembly mode:* Individual assembly
- *Contigs/Scaffolds:* SPAdes on data.: scaffolds/ or contigs (fasta)
- *Type of assembly:* Genome
- *Use a reference genome:* “No”. Select “Yes” to compare the assembly with the reference genome.

Quast Genome assembly Quality
(Galaxy Version 5.3.0+galaxy1)

☆ ⚙️ ▼ 📄 ▶ Run Tool

Assembly mode?

Individual assembly (1 contig file per sample) ▼

Useful to know if contigs have been generated all samples together (co-assembly) or on each sample individually (individual assembly)

Use customized names for the input files?

No, use dataset names ▼

They will be used in reports, plots and logs

Contigs/scaffolds file *

📄 📄 📁 ... 27: SPAdes on collection 18: Contigs ▼

accepted formats ▼

Reads options

Disabled ▼

Currently, the supported read types are Illumina unpaired, paired-end and mate-pair reads, PacBio SMRT, and Oxford Nanopore long reads.

Type of assembly

Genome ▼

Use a reference genome?

No ▼

Many metrics can't be evaluated without a reference. If this is omitted, QUAST will only report the metrics that can be evaluated without a reference.

QUAST results report: HTML format

QUAST

Quality Assessment Tool for Genome Assemblies

16 October 2026, Friday, 16:57:25

[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).

Statistics without reference SPAdes_on_collection_18__Cont...

# contigs	153
# contigs (≥ 0 bp)	273
# contigs (≥ 1000 bp)	117
# contigs (≥ 5000 bp)	75
# contigs (≥ 10000 bp)	64
# contigs (≥ 25000 bp)	46
# contigs (≥ 50000 bp)	32
Largest contig	268 673
Total length	5 338 373
Total length (≥ 0 bp)	5 366 236
Total length (≥ 1000 bp)	5 314 661
Total length (≥ 5000 bp)	5 209 743
Total length (≥ 10000 bp)	5 130 079
Total length (≥ 25000 bp)	4 858 775
Total length (≥ 50000 bp)	4 407 690
N50	134 517
N90	26 695
auN	136 167
L50	15
L90	44
GC (%)	50.66

Per base quality

# N's per 100 kbp	0
# N's	0

Sample ID: E.coli 327 (ERR13363635)

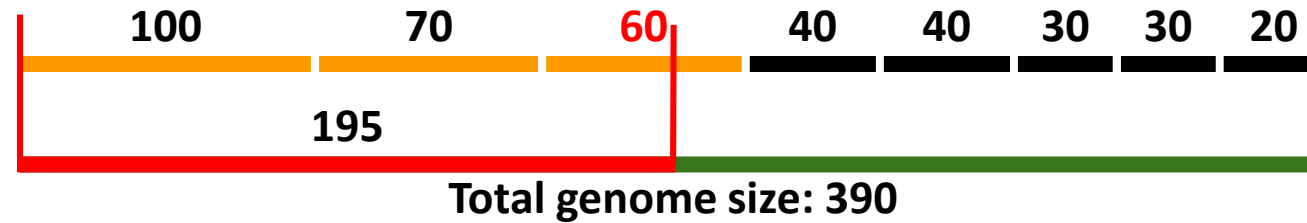
Basic statistics

Contig sizes

- Number of contigs
- Number of large contigs (i.e. > 1000 bp)
- Largest contig length
- Total assembly length
- Misassemblies

N50

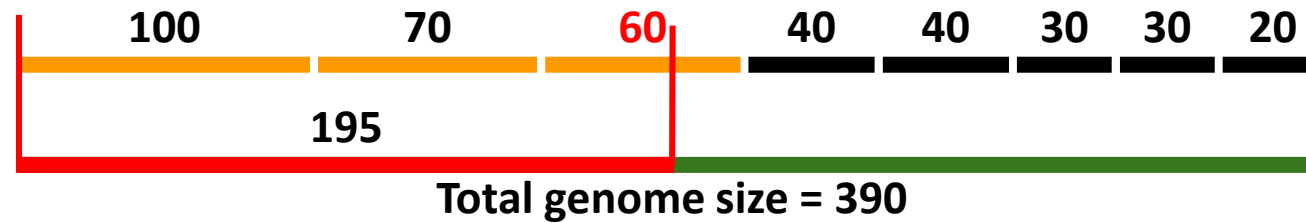
The length of the contig crossing 50% mark of the assembly



$$\text{N50} = 60$$

L50

The number of contigs crossing 50% mark of the assembly



$$\text{L50} = 3$$

Question:

Based on the QUAST analysis, evaluate the assembly quality of E. coli 327 (ERR13363635) using the following criteria:

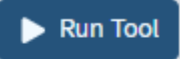
1. Number of contigs and scaffolds
2. Largest contig
3. Assembly quality metrics: N50, L50, and misassemblies
4. Compare the total genome length and GC content (%) of the assembly from E. coli 327 with the E. coli reference genome (Escherichia coli str. K-12 substr. MG1655: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000005845.2/).

Filter contigs/scaffolds by length (optional)

Tool: Filter sequences by length

Tool Parameters

- *Fasta file*: Assembled genome sequences
- *Minimal length*: Enter the minimum contig length to retain.

 **Filter sequences by length** (Galaxy Version 1.2)    

Tool Parameters

Fasta file *

    52: SPAdes on collection 8: Scaffolds 

accepted formats ▼

Minimal length *

500|  *e.g. Select contigs longer than 500 bp* 

Maximum length *

0

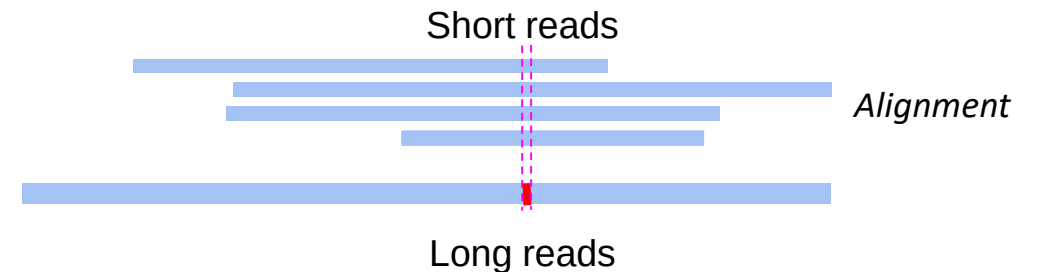
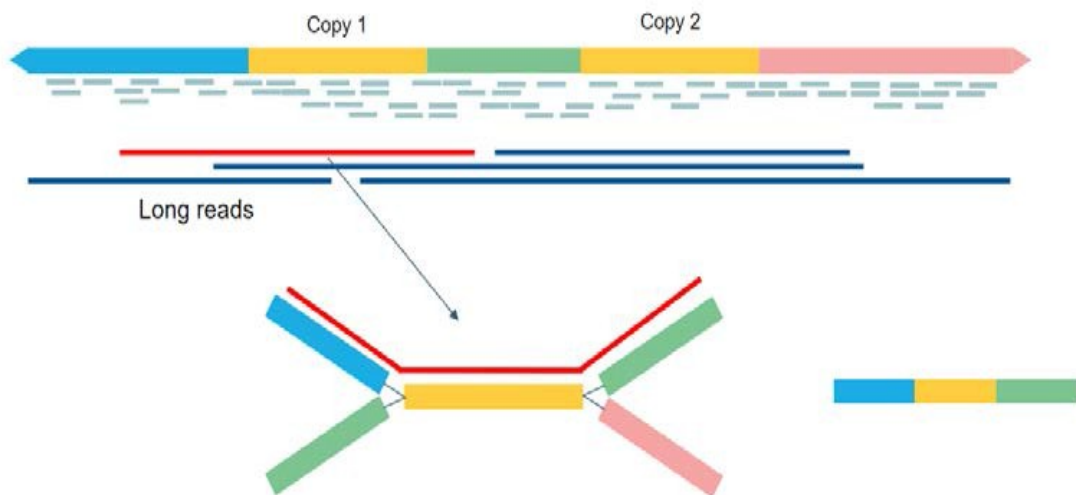
Setting to '0' will return all sequences longer than the 'Minimal length'

4. *Practical:* Hybrid Assembly for short- and long-read data

Hybrid Assembly:

First, assemble the genome using either short-read or long-read data, then polish the assembly using the remaining read data.

- **Short-read-first assembly:** Assemble the genome using short reads first, then use long reads to provide structural information about the genome.
- **Long-read-first assembly:** Assemble the genome using long reads first, then use short reads to correct errors in the contig.



Hybrid assembly with Unicycler

Tool: *Unicycler*

Tool Parameters:

- *Paired or Single end data:* Paired Collection
- *Select a pair collection:* Raw FASTQ files or paired-end data after quality filtering (short reads)
- *Select long reads:* ONT sequencing data

Create assemblies with Unicycler
pipeline for bacterial genomes (Galaxy Version 0.5.1+galaxy0) ☆ 🖨 ⌵ 📁 ▶ Run Tool

Tool Parameters

Paired or Single end data?

Paired Collection ←

Select between paired and single end data

Select a paired collection *

18: fastp on collection 1: Paired-end output ←

accepted formats ▼

Select long reads. If there are no long reads, leave this empty - optional

📄 📄 📁 ⋮ 2: Single-end data (fastq-dump) ↘

accepted formats ▼

! The supplied input will be mapped over this tool.
(--long)

Select the Illumina data format.

Click to navigate to the folder containing the short- and long-read data

Unicycler output data

WGS_Ecoli_327

Default Storage

582 MB

19 31

33: Create assemblies with Unicycler on collection 2 and 18: Final Assembly
a list with 1 **fasta** dataset

32: Create assemblies with Unicycler on collection 2 and 18: Final Assembly Graph
a list with 1 **gfa1** dataset

31: Create assemblies with Unicycler on collection 2 and 18: SPAdes graphs
a list with 1 **gfa1** list

- **Final Assembly:** Genome sequence assembled using Unicycler
- **Final Assembly Graph:** Assembly graph constructed from short- and long-read data
- **SPAdes graphs:** Assembly graph constructed from short-read data using SPAdes

Unicycler Input Data Options:

Unicycler can run with Illumina data alone.

Unicycler can run with Nanopore long-read data alone.

Unicycler can run with a combination of Illumina and Nanopore data.

Edit Dataset Attributes for Unicycler output (optional)

1. Select the assembly file to edit (File: Final Assembly)

33: Create assemblies with Unicycler on collection 2 and 18: Final Assembly

a list with 1 **fasta** dataset

2. Click "View"

1: ERR13363635

298 sequences

format **fasta**, database ?

Starting Unicycler (2026-10-16 17:11:53)
Welcome to Unicycler, an assembly pipeline for bacterial genomes. Since you

>1 length=1373458 depth=1.08x

GATGAACTGGCCCTGACGCTGGAAGCTGACAGCGTTAACGTACAGGCACTGAAC/
GTTGACGTCGATGGTGTGAGCTTGCCGAACTGATTAACTGGTCTGCGATAATGC
TTGTTGATGAATCTGACCGGACCTCAACAGACAGCATACCCCCCTCTGCTGCCCTC
TAGTACCGCGCATATCACGGAAACAGACAACGCCTGGCTGACTCGCTGTCGCATC

3. In Edit Dataset Attributes → Datatypes, set Target datatype to fasta.gz → Create Dataset

35: Create assemblies with Unicycler on dataset 5:

Final Assembly

ok

298 sequences

format **fasta** database ? size **5.6 MB**

Preview Visualize Details Edit

Edit Dataset Attributes

Attributes Datatypes Permissions

Assign Datatype

New Type

fasta

This will change the datatype of the existing dataset but not modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

Save Auto-detect

Convert to Datatype

Target datatype

fasta.gz (using 'Convert uncompressed file to compressed')

This will create a new dataset with the contents of this dataset converted to a new format.

Create Dataset

Question:

Compare the quality of short-read and hybrid short- and long-read genome assemblies of *E. coli* 327 against the reference genome *Escherichia coli* str. K-12 substr. MG1655.

1. Which assembly is closest to the reference genome?
2. Which assembly has the longest contig?
3. Compare the total assembly length with that of the reference genome.
4. Compare N50 and L50 values.
5. Compare the number of misassemblies.
6. Compare the GC content (%) of each assembly with that of the reference genome.

Reference genomes retrieval

<https://www.ncbi.nlm.nih.gov/datasets/genome/>

The screenshot shows the NCBI Genomes dataset page. At the top, there is a search bar with the text "Search NCBI ...". Below the search bar, there are navigation links: "NCBI Datasets", "Taxonomy", "Genome" (which is highlighted), "Gene", "Command-line tools", and "Documentation".

The main section is titled "Genome". Below the title, there is a search instruction: "Search by taxonomic name or ID, Assembly name, BioProject, BioSample, WGS or Nucleotide accession". Below this is a search input field with the placeholder text "Enter a search term". A grey callout box with an arrow points to this field, containing the text: "Enter the species name for which genome data will be retrieved".

Below the search field, there are example search terms: "Try examples: Homo sapiens GCF_000001405.40 PRJNA489243 SAM".

The section "Genomic data available from NCBI Datasets" follows, with a sub-instruction: "Click below to learn more about the genomic data available from NCBI Datasets." Below this are three categories: "Eukaryota", "Archaea", and "Bacteria", each with a corresponding icon.

To the right of these categories is a large box titled "All Genomes" containing three circular statistics:

Category	Value
Total	3.31M
Reference	44.17K
Annotated	2.66M

Example: Reference genome data for *Escherichia coli* str. K-12 substr. MG1655

(DOI: 10.1038/s41598-020-64074-5)

- Download link for the ***Escherichia coli* str. K12** strain:
https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000005845.2/

Genome assembly ASM584v2 reference

[Download](#) [datasets](#) [API](#) [FTP](#) 

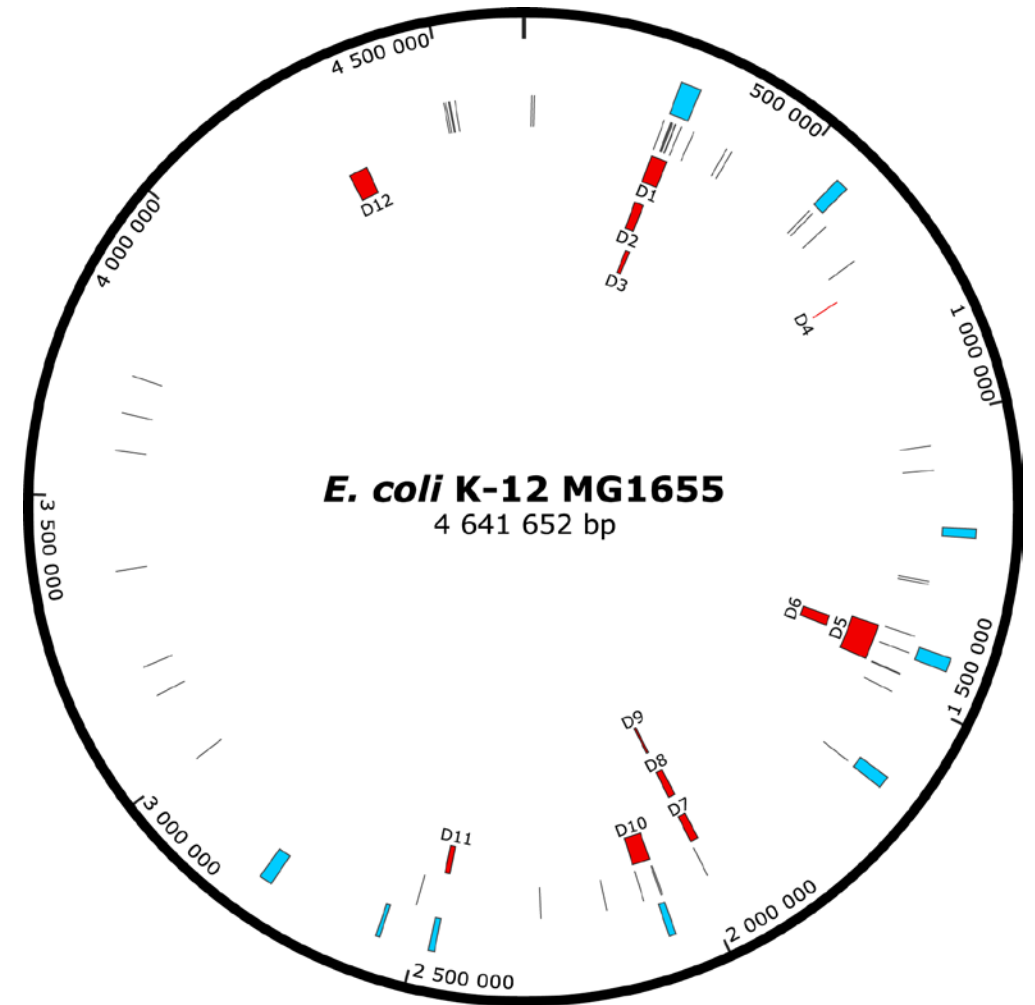
NCBI RefSeq assembly	GCF_000005845.2
Submitted GenBank assembly	GCA_000005845.2
Taxon	Escherichia coli str. K-12 substr. MG1655
Strain	K-12 substr. MG1655
Submitter	Univ. Wisconsin
Date	Sep 26, 2013

=> Click FTP to open the data download directory.

* [_genomic.gbff.gz](#): GenBank data

* [_genomic.fna.gz](#): the genome sequence file

* [_genomic.gff.gz](#): genomic feature file in GFF3 format



Upload reference genome files to Galaxy

- Upload the genomic feature file (GFF) and the genome sequence file (FASTA)
- How to upload data: Choose **Upload** → **Choose local file** → **Start** → **Close**

The screenshot displays the Galaxy Australia web interface. On the left sidebar, the 'Upload' button is highlighted with an orange circle and the number '1'. The main panel shows the 'Tools' section with a search bar and various tool categories. A modal window titled 'Upload from Disk or Web to WGS_Ecoli_327' is open, showing two files in the queue: 'GCF_000005845.2' (1.4 MB) and 'GCF_000005845.2' (387.6 KB). Both files are set to 'Auto-detect' and 'unspecified (?)'. At the bottom of the modal, the 'Choose local file' button is highlighted with an orange circle and the number '2', and the 'Start' button is highlighted with an orange circle and the number '3'. Other buttons visible include 'New upload Beta', 'Choose from repository', 'Paste/Fetch data', 'Pause', 'Reset', and 'Close'.

Multiple genome assembly quality assessment with QUAST

Tools: Quast Genome assembly Quality

Tool Parameters:

- *Assembly mode*: Co-assembly
- *Contigs/Scaffolds*: Assembled genome sequences
- *Type of assembly*: Genome
- *Use a reference genome*: “Yes”
 - ✓ *Reference genome*: Genome sequence in FASTA format (.fna)
 - ✓ *Genomic feature positions in the reference genome*: gff file

Select the previously
uploaded reference
genome files

Type of assembly

Genome

Use a reference genome?

Yes

Many metrics can't be evaluated without a reference. If this is omitted, QUAST will only report the metrics that can be evaluated without a reference.

Reference genome *

51: GCF_000005845.2_ASM584v2_genomic.fna.gz (as fasta)

switch to column select

accepted formats

(-f)

Genomic feature positions in the reference genome - optional

52: GCF_000005845.2_ASM584v2_genomic.gff.gz (as gff3)

accepted formats

Gene coordinates for the reference genome (--features)

Quast Genome assembly Quality
(Galaxy Version 5.3.0+galaxy1)

Run Tool

Tool Parameters

Assembly mode?

Co-assembly (1 contig file for several samples)

Useful to know if contigs have been generated all samples together (co-assembly) or on each sample individually (individual assembly)

Use customized names for the input files?

No, use dataset names

They will be used in reports, plots and logs

Contigs/scaffolds file *

54: Create assemblies with Unicycler on dataset 5: Final Assembly compressed (as fasta)

27: SPAdes on collection 18: Contigs

switch to column select

accepted formats

Select the genome assembly files

QUAST results with the reference genome

QUAST

Quality Assessment Tool for Genome Assemblies

17 October 2026, Saturday, 00:55:42

[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).

Aligned to "dataset_efae4a02-2db7-4b1c-b0fc-426664a9ffb4" | 4 641 652 bp | 1 fragment | 50.79 % G+C
9523 genomic features

 ☒ Show heatmap
Worst Median Best

Alignment-based statistics

	Create_assemblies_with_Unicyc...	SPAdes_on_collection_18_Cont...
Genome fraction (%)	84.676	84.541
Duplication ratio	1.008	1.001
# genomic features	7733 + 392 part	7571 + 432 part
Largest alignment	147 900	113 100
Total aligned length	3 912 979	3 878 903
NGA50	32 006	28 579
LGA50	39	48

Misassemblies

# misassemblies	158	151
Misassembled contigs length	4 921 318	4 313 782

Per base quality

# mismatches per 100 kbp	2533.21	2543.06
# indels per 100 kbp	51.19	47.18
# N's per 100 kbp	0	0

Statistics without reference

# contigs	74	153
Largest contig	1 373 458	268 673
Total length	5 477 436	5 338 373
Total length (≥ 1000 bp)	5 459 681	5 314 661
Total length (≥ 10000 bp)	5 398 737	5 130 079
Total length (≥ 50000 bp)	5 150 005	4 407 690

*Click here to view
the full report*

→ [Extended report](#)

The analysis history of E. coli 327 (ERR13363635) has been shared on the Galaxy Australia platform and can be accessed at the following link:

<https://usegalaxy.org.au/u/tam-tran/h/wgs-ecoli-327>

The screenshot displays the Galaxy Australia interface for a history named 'WGS_Ecoli_327'. The interface includes a search bar, a list of datasets, and a list of analysis steps. A red arrow points to the 'Copy History' option in the 'History Options' menu.

Histories / WGS_Ecoli_327 current

search datasets

WGS_Ecoli_327

Default Storage

488 MB

27: SPAdes on collection 18: Contigs

26: SPAdes on collection 18: Assembly graph with scaffolds

25: SPAdes on collection 18: Assembly graph

20: fastp on collection 1: JSON report

a list with 1 json dataset

19: fastp on collection 1: HTML report

a list with 1 html dataset

18: fastp on collection 1: Paired-end output

a list with 1 fastqsanger.gz pair

14: NanoPlot on collection 2: NanoStats post filtering

a list with 1 tabular dataset

13: NanoPlot on collection 2: NanoStats

a list with 1 tabular dataset

12: NanoPlot on collection 2: HTML report

a list with 1 html dataset

7: FastQC on collection 1: RawData

a list with 1 txt pair

Switch to this history

Manage History

History Options

Resume Paused Jobs

Copy History

Copy History to a New History

Delete History

Export Tool References

Export History to File

Archive History

Extract Workflow

Show Invocations

Show History Graph

Show History Notebooks

Share & Manage Access

Summary

1. Two approaches to genome assembly

- De novo genome assembly
- Reference-based genome assembly

2. Genome assembly tools

- Many bioinformatics tools are available for assembling genomes from sequencing data. Choose the appropriate tool and workflow based on the sequencing data and genome characteristics.
- Higher sequencing coverage generally leads to a more complete assembly.
- Combines short-read data with long-read data to leverage their complementary strengths.