How good is my assembly?

QUAST (QUality ASsessment Tool)

The input data may be assemblies, reference genomes, gene and operon annotations, and raw reads files.

Sequences
The tool accepts assemblies and reference genomes in FASTA format. Files may be compressed with zip, gzip, or bzip2. A reference genome with multiple chromosomes can be provided as a single FASTA file with separate sequence for each chromosome inside.
Reads
QUAST accepts Illumina, PacBio, and Oxford Nanopore reads in FASTQ format (may be compressed) or in the aligned form in SAM/BAM formats.
Genes and operons
One can also specify files with gene and operon positions in the reference genome. QUAST will count fully and partially aligned regions, and output total values as well as cumulative plots.
The following file formats are supported:

GFF, versions 2 and 3;
BED: sequence name, start position, end position, gene/operon id, optional fields;
the format used by NCBI for genes ("Summary (text)");
four tab-separated columns: sequence name, gene/operon id, start position, end position.
Note that the sequence name has to fully match a name in the reference file.

Quast options

-o <output_dir>
Output directory. The default value is quast_results/results_<date_time>.
Note: QUAST reuses existing alignments if run repeatedly with the same output directory.
-r <path>
Reference genome file. Optional. 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.
--features (or -g) <path>
File with genomic feature positions in the reference genome. See details about the file format in section 2.2.
If you use GFF format and would like to count only a specific feature from it (e.g., only "CDS" or only "gene") you can specify this feature followed by a colon (":") as the filepath prefix (do not use spaces!). For example:
--features CDS:~/data/my_genome_annotation.gff otherwise, all features from the file will be considered.
If you do not have the annotated positions, you can make QUAST predict genes with --gene-finding.
--min-contig (or -m) <int>
Lower threshold for a contig length (in bp). Shorter contigs won't be taken into account (except for specific metrics, see section 3). The default value is 500.
--threads (or -t) <int>
Maximum number of threads. The default value is 25% of all available CPUs but not less than 1. If QUAST fails to determine the number of CPUs, maximum threads number is set to 4.

Quast advanced options

--eukaryote (or -e)

Genome is eukaryotic. Affects gene finding, conserved orthologs finding and contig alignment:

For prokaryotes (which is default), GeneMarkS is used. For eukaryotes, GeneMark-ES is used.

Barrnap will use eukaryotic database to predict ribosomal RNA genes.

BUSCO will use eukaryotic database to find conserved orthologs.

By default, QUAST assumes that a genome is circular and correctly processes its linear representation. This options indicates that the genome is not circular.

--fungus

Genome is fungal. Affects gene finding, conserved orthologs finding and contig alignment:

For gene finding, GeneMark-ES is used with --fungus option.

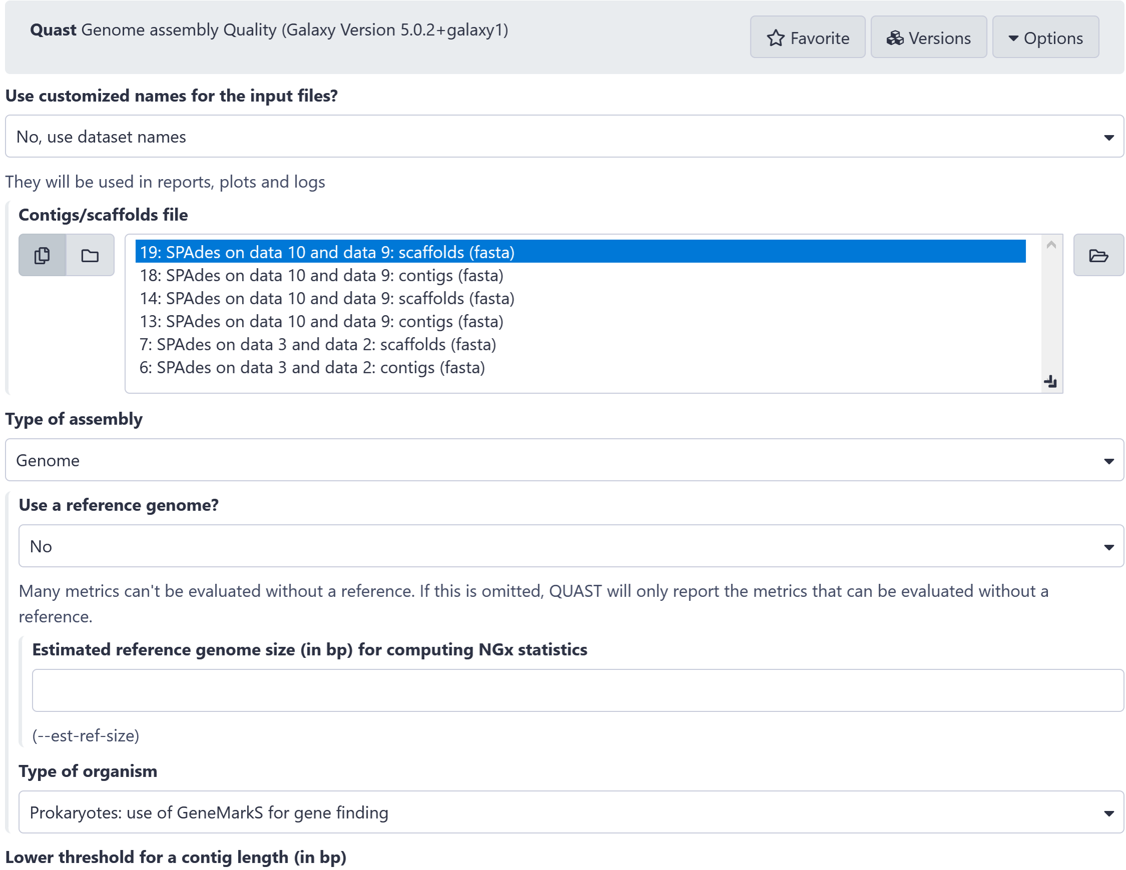
Barrnap will use eukaryotic database to predict ribosomal RNA genes.

>BUSCO will use fungal database to find conserved orthologs.

By default, QUAST assumes that a genome is circular and correctly processes its linear representation. This options indicates that the genome is not circular.

There are many, many more options. See http://quast.sourceforge.net/docs/manual.html

Quast in Galaxy



Quast Genome assembly Quality (Galaxy Version 5.0.2+galaxy1)

Use customized names for the input files?

No, use dataset names

They will be used in reports, plots and logs

Contigs/scaffolds file

19: SPAdes on data 10 and data 9: scaffolds (fasta)
18: SPAdes on data 10 and data 9: contigs (fasta)
14: SPAdes on data 10 and data 9: scaffolds (fasta)
13: SPAdes on data 10 and data 9: contigs (fasta)
7: SPAdes on data 3 and data 2: scaffolds (fasta)
6: SPAdes on data 3 and data 2: contigs (fasta)

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.

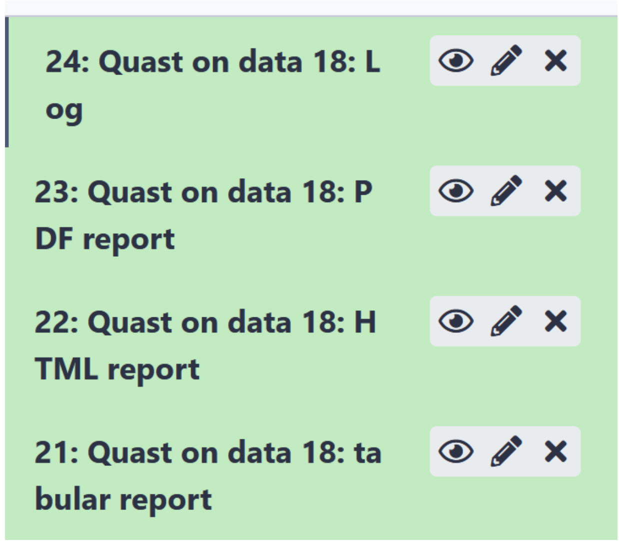
Estimated reference genome size (in bp) for computing NGx statistics

(--est-ref-size)

Type of organism

Prokaryotes: use of GeneMarkS for gene finding

Lower threshold for a contig length (in bp)



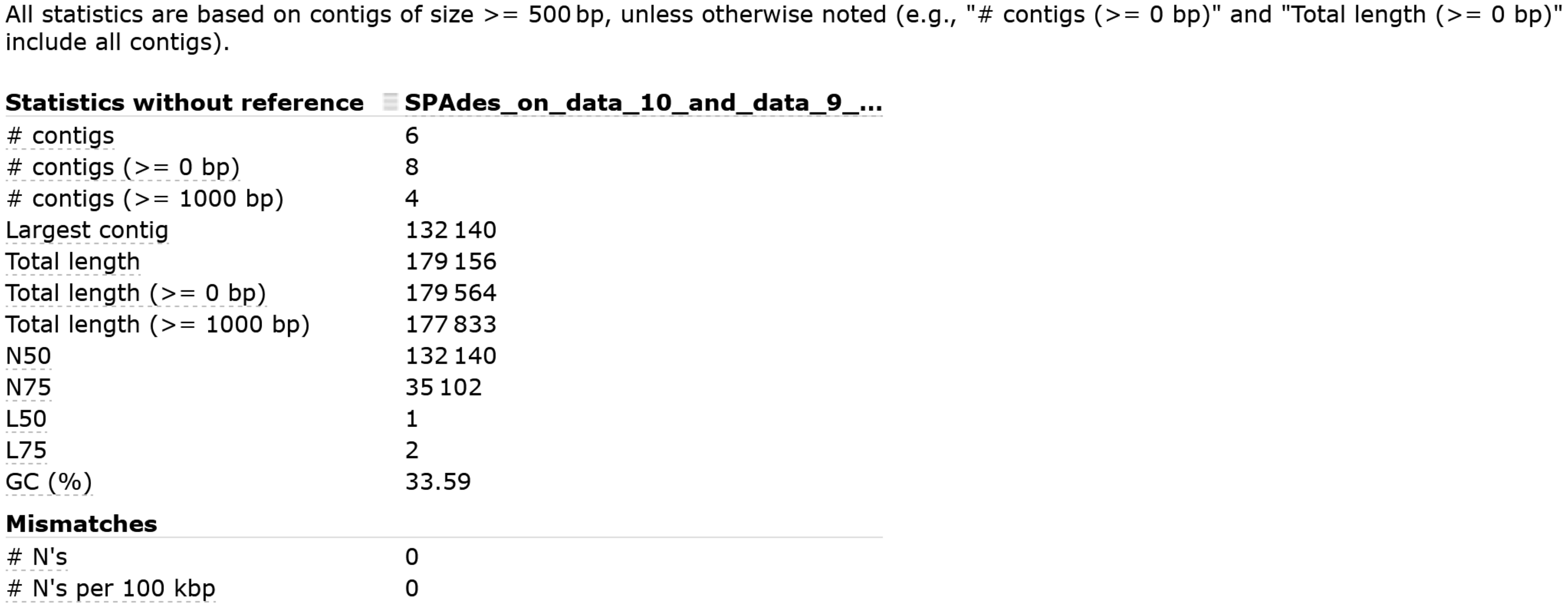
24: Quast on data 18: Log

23: Quast on data 18: PDF report

22: Quast on data 18: HTML report

21: Quast on data 18: tabular report

Quast output



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_data_10_and_data_9_...**
# contigs	6
# contigs (>= 0 bp)	8
# contigs (>= 1000 bp)	4
Largest contig	132 140
Total length	179 156
Total length (>= 0 bp)	179 564
Total length (>= 1000 bp)	177 833
N50	132 140
N75	35 102
L50	1
L75	2
GC (%)	33.59
Mismatches	
# N's	0
# N's per 100 kbp	0

N50 the size of the *smallest* contig, where it and all larger contigs make up at least 50% of the assembly size
L50 the number of contigs that make up N50
NG50 same as N50 but calculated relative to the reference genome
LG50 same as L50 but calculated relative to the reference genome

Output on HPC appears to be broken

Bandage

Bandage Info output on Galaxy

- **You need to select assembly graph as one of the outputs when running SPAdes**

Node count: The number of nodes in the graph. Only positive nodes are counted (i.e. each complementary pair counts as one).

Edge count: The number of edges in the graph. Only one edge in each complementary pair is counted.

Total length: The total number of base pairs in the graph.

Dead ends: The number of instances where an end of a node does not connect to any other nodes.

Percentage dead ends: The proportion of possible dead ends. The maximum number of dead ends is twice the number of nodes (occurs when there are no edges), so this value is the number of dead ends divided by twice the node count.

Connected components: The number of regions of the graph which are disconnected from each other.

Largest component: The total number of base pairs in the largest connected component.

N50: Nodes that are this length or greater will collectively add up to at least half of the total length.

Shortest node: The length of the shortest node in the graph.

Lower quartile node: The median node length for the shorter half of the nodes.

Median node: The median node length for the graph.

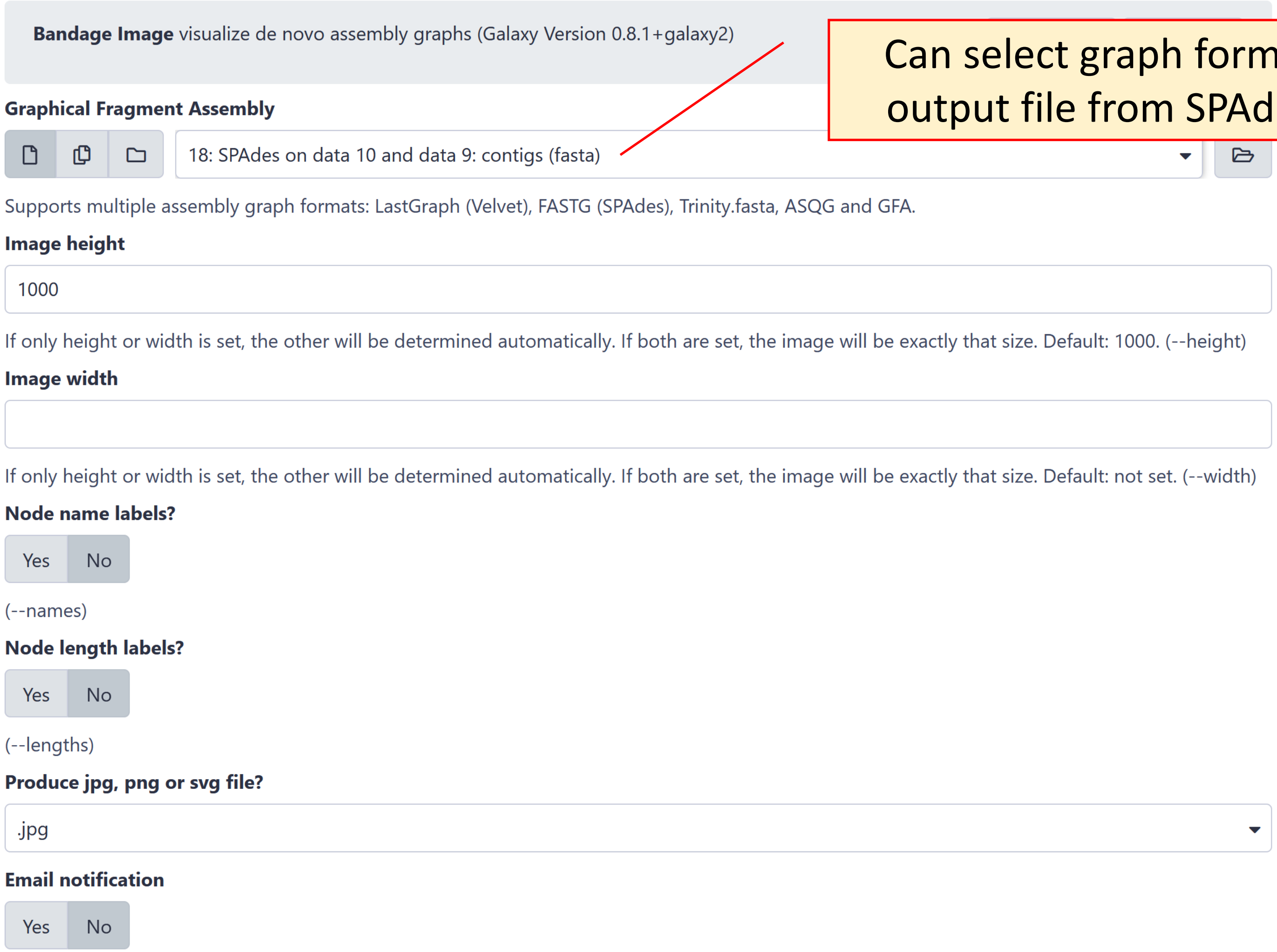
Upper quartile node: The median node length for the longer half of the nodes.

Longest node: The length of the longest node in the graph.

Bandage Info example output on Galaxy

```
Node count:                       8
Edge count:                       0
Smallest edge overlap (bp):       0
Largest edge overlap (bp):        0
Total length (bp):                179564
Total length no overlaps (bp):    179564
Dead ends:                        16
Percentage dead ends:             100%
Connected components:             8
Largest component (bp):           132140
Total length orphaned nodes (bp): 132140
N50 (bp):                         132140
Shortest node (bp):               111
Lower quartile node (bp):         474
Median node (bp):                 2728
Upper quartile node (bp):         13219
Longest node (bp):                132140
Median depth:                     10.6238
Estimated sequence length (bp):   187798
```

Bandage Image in Galaxy



Bandage Image visualize de novo assembly graphs (Galaxy Version 0.8.1+galaxy2)

Graphical Fragment Assembly

18: SPAdes on data 10 and data 9: contigs (fasta)

Supports multiple assembly graph formats: LastGraph (Velvet), FASTG (SPAdes), Trinity.fasta, ASQG and GFA.

Image height

1000

If only height or width is set, the other will be determined automatically. If both are set, the image will be exactly that size. Default: 1000. (--height)

Image width

If only height or width is set, the other will be determined automatically. If both are set, the image will be exactly that size. Default: not set. (--width)

Node name labels?

Yes [✓ No]

(--names)

Node length labels?

Yes [✓ No]

(--lengths)

Produce jpg, png or svg file?

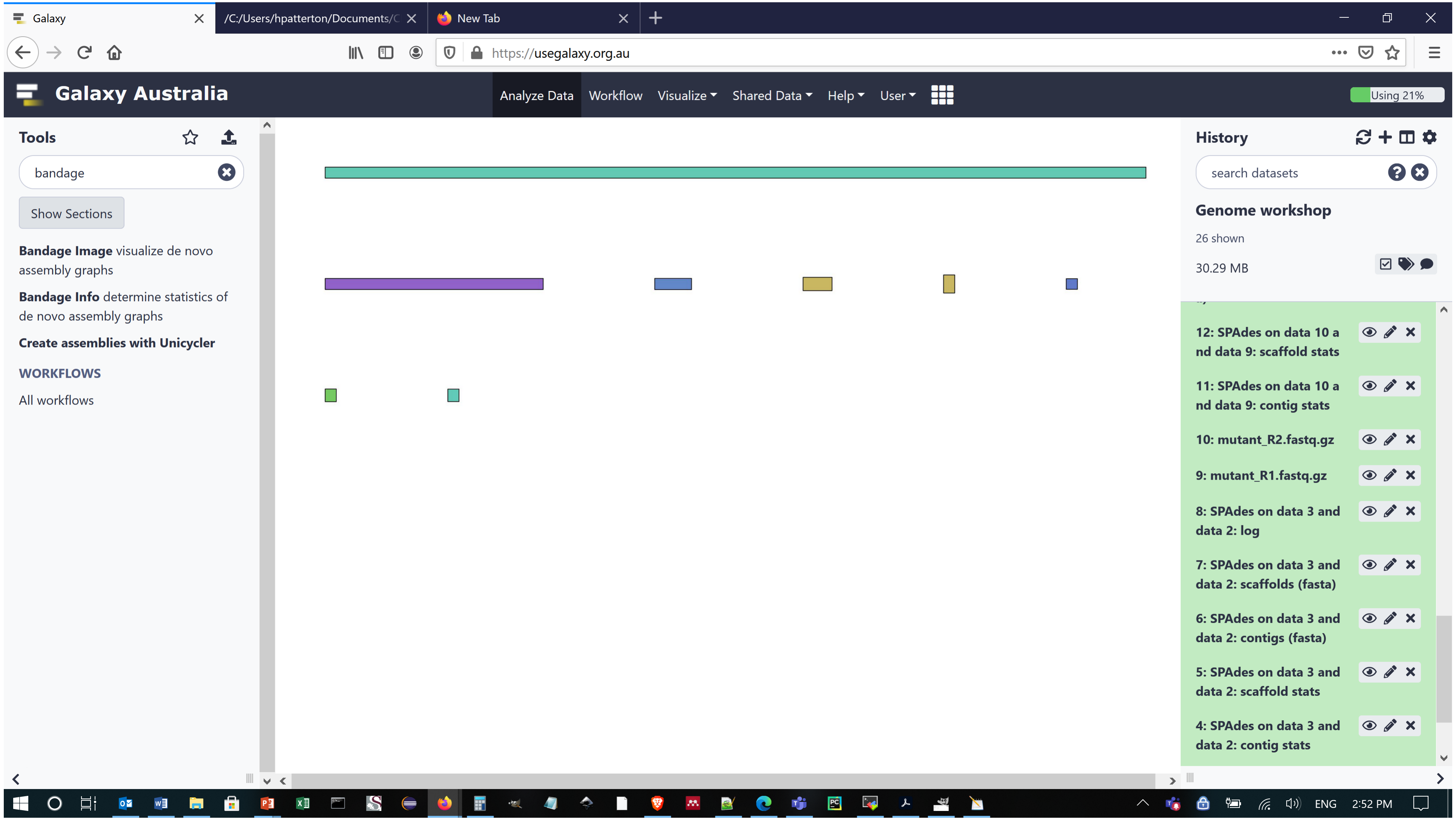
.jpg

Email notification

Yes [✓ No]

Can select graph format output file from SPAdes

Bandage Image output in Galaxy



Bandage Image visualize de novo assembly graphs

Bandage Info determine statistics of de novo assembly graphs

Create assemblies with Unicycler

WORKFLOWS

All workflows

Genome workshop

26 shown

30.29 MB

12: SPAdes on data 10 and data 9: scaffold stats

11: SPAdes on data 10 and data 9: contig stats

10: mutant_R2.fastq.gz

9: mutant_R1.fastq.gz

8: SPAdes on data 3 and data 2: log

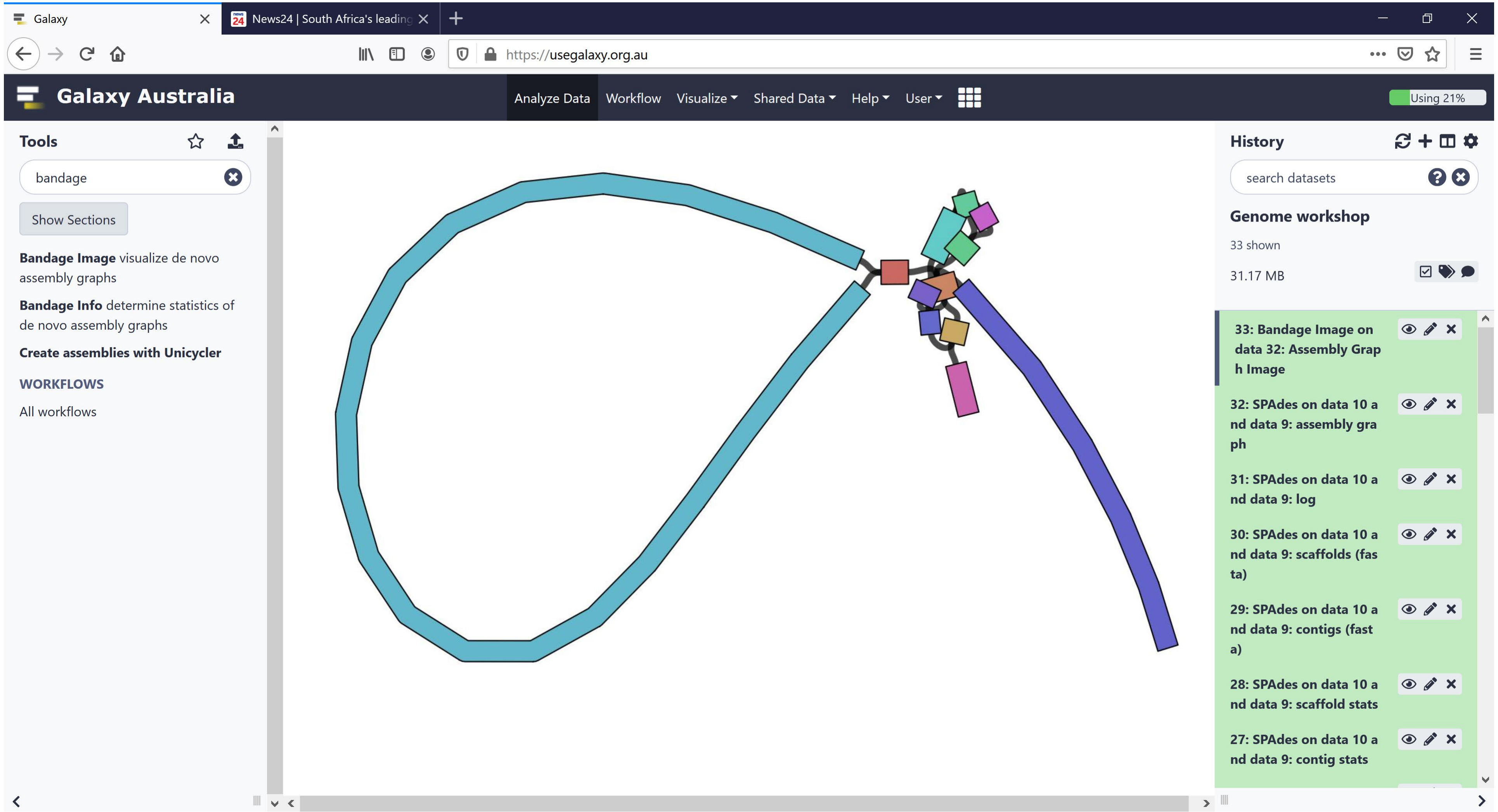
7: SPAdes on data 3 and data 2: scaffolds (fasta)

6: SPAdes on data 3 and data 2: contigs (fasta)

5: SPAdes on data 3 and data 2: scaffold stats

4: SPAdes on data 3 and data 2: contig stats

Bandage Image output



Bandage Image visualize de novo assembly graphs

Bandage Info determine statistics of de novo assembly graphs

Create assemblies with Unicycler

Genome workshop

33 shown

31.17 MB

33: Bandage Image on data 32: Assembly Graph Image

32: SPAdes on data 10 and data 9: assembly graph

31: SPAdes on data 10 and data 9: log

30: SPAdes on data 10 and data 9: scaffolds (fasta)

29: SPAdes on data 10 and data 9: contigs (fasta)

28: SPAdes on data 10 and data 9: scaffold stats

27: SPAdes on data 10 and data 9: contig stats