

# Using SPAdes

# SPAdes pipeline

**SPAdes comes in several separate modules**

## **BayesHammer**

read error correction tool for Illumina reads, which works well on both single-cell and standard data sets.

## **IonHammer**

read error correction tool for IonTorrent data, which also works on both types of data.

## **SPAdes**

iterative short-read genome assembly module; values of K are selected automatically based on the read length and data set type.

## **MismatchCorrector**

a tool which improves mismatch and short indel rates in resulting contigs and scaffolds; this module uses the BWA tool

MismatchCorrector is turned off by default, but we recommend to turn it on

We recommend to run SPAdes with BayesHammer/IonHammer to obtain high-quality assemblies.

However, if you use your own read correction tool, it is possible to turn error correction module off. It is also possible to use only the read error correction stage, if you wish to use another assembler. S

# Get a new, blank history

Galaxy

https://galaxy-new.sun.ac.za/galaxy/datasets/edit

Analyze Data Workflow Visualize Shared Data Admin Help User Using 281.2 GB

**Tools**

- Get Data
- Send Data
- Collection Operations
- Expression Tools
- Lift-Over
- Text Manipulation
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- Fetch Alignments/Sequences
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Phenotype Association
- NGS: Mapping
- NGS: ChIP-seq
- NGS: QC
- NGS: RNA-seq

**Edit dataset attributes**

Attributes Convert Datatypes Permissions

Edit attributes Auto-detect Save

**Name**

SPAdes on data 3 and data 2: log

**Info**

Command line: /scratch/galaxy/database/dependencies/\_conda/envs/\_spades@3.12.0/bin/spades.py -o /scratch/galaxy/database/jobs\_directory/001/1611/working --disable-gzip-output --careful -t 1 -m 250 -k 21,33,55 --pe1-fr --pe1-1 fastq/scratch/galaxy/datab

**Annotation**

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

**Database/Build**

----- Additional Species Are Below -----

**History**

search datasets

**Genome workshop**

26 deleted

666 MB

This history is empty. You can load your own data or get data from an external source

Give your new history a descriptive name

# Gettings data to Galaxy

The screenshot shows the Galaxy web interface with a 'File Upload' dialog box open. The dialog box has a title bar 'File Upload' and a search bar containing 'pacbio'. Below the search bar is a file list with columns for Name, Date modified, Type, and Size. The file list shows a folder 'Workshops' and a file 'pacbio.fa' with a size of 19,522 KB. Below the file list is a 'File name' field and a file type dropdown set to 'All Files (\*.\*)'. At the bottom of the dialog box are 'Open' and 'Cancel' buttons. A red box highlights the text 'Select Get Data > Upload File > Choose local File'.

Download from web or upload from disk

Regular Composite Collection Rule-based

File Upload

Search pacbio

Workshops

Name	Date modified	Type	Size
pacbio.fa	10/21/2020 12:18	FQ File	19,522 KB
		FQ File	336 KB
		FQ File	336 KB

pacbio

File name:

All Files (\*.\*)

Open Cancel

Type (set all): Auto-detect Genome (set all): ----- Additional S...

Choose local file Choose FTP file Paste/Fetch data Pause Reset Start Close

Database/Build

----- Additional Species Are Below -----

Tools

search tools

descriptive stats for BAM dataset

CalMD recalculate MD/NM tags

BedCov calculate read depth for a set of genomic intervals

SAM-to-BAM convert SAM to BAM

BAM-to-SAM convert BAM to SAM

Filter BAM datasets on a variety of attributes

Upload File from your computer

UCSC Main table browser

UCSC Archaea table browser

EBI SRA ENA SRA

modENCODE fly server

InterMine server

Flymine server

modENCODE modMine server

MouseMine server

Ratmine server

YeastMine server

modENCODE worm server

WormBase server

ZebrafishMine server

History

SPAdes workshop

Using 281.2 GB

# Select the file type

Download from web or upload from disk

Regular Composite Collection Rule-based

You added 3 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
 pacbio.fq	19.1 MB	fastqsan... 	----- Additional S... 		0% 
 R1.fq	335.8 KB	fastqsan... 	----- Additional S... 		0% 
 R2.fq	335.8 KB	fastqsan... 	----- Additional S... 		0% 

Type (set all):   Genome (set all):  

 Choose local file  Choose FTP file  Paste/Fetch data  Pause  Reset  Start  Close

Select *fastqsanger* as type for all

# SPAdes settings

The screenshot shows the Galaxy Australia interface with the SPAdes tool configuration page. The 'Tools' panel on the left lists various tools, with 'SPAdes' selected. A red box highlights the 'Select SPAdes' button. The main panel shows the tool's settings, including options for single-cell data, assembly type, correction, k-mer values, coverage cutoff, and library type. A yellow box in the foreground contains a table summarizing these settings.

Single-cell?	No	
Run only assembly? (without read error correction)		No
Careful correction?	Yes	
Automatically choose k-mer values	No	
K-mers to use, separated by commas	21,33,55	
Coverage Cutoff	Off	
Libraries are IonTorrent reads?		No

# Select the forward and reverse Illumina read files

The screenshot displays the Galaxy Australia web interface. The main configuration area is titled "Libraries" and "Files". Under "Files", the "Select file format" is set to "Separate input files". The "Forward reads" field is set to "2: R1.fq" and the "Reverse reads" field is set to "3: R2.fq". The "History" panel on the right shows a list of datasets: "3: R2.fq", "2: R1.fq", and "1: pacbio.fq". The "Tools" panel on the left lists various tools, including SPAdes, Shovill, metaSPAdes, and Unicycler.

Select

# Data compression/decompression with gzip

Usage: gzip [OPTION]... [FILE]...

Compress or uncompress FILEs (by default, compress FILEs in-place).

Mandatory arguments to long options are mandatory for short options too.

- c, --stdout write on standard output, keep original files unchanged
- d, --decompress decompress
- f, --force force overwrite of output file and compress links
- h, --help give this help
- l, --list list compressed file contents
- L, --license display software license
- n, --no-name do not save or restore the original name and time stamp
- N, --name save or restore the original name and time stamp
- q, --quiet suppress all warnings
- r, --recursive operate recursively on directories
- S, --suffix=SUF use suffix SUF on compressed files
- t, --test test compressed file integrity
- v, --verbose verbose mode
- V, --version display version number
- 1, --fast compress faster
- 9, --best compress better
- rsyncable Make rsync-friendly archive

# SPAdes output

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   

3: R2.fq   

2: R1.fq   

1: pacbio.fq   

- There should now be 5 new files in your history
- You can select the eye icon to view the file
- Note the small contigs with very high coverage
- Scaffolds = contigs: np scaffold assembly

name	length	coverage
#name	length	coverage
NODE_1	132140	12.619781
NODE_2	35102	12.666048
NODE_3	5925	13.405451
NODE_4	4666	25.835610
NODE_5	790	65.717007
NODE_6	533	10.721757
NODE_7	297	21.004132
NODE_8	111	13.571429

# SPAdes on the HPC

```
#!/bin/bash
```

```
#PBD -N workshop_spades
```

```
#PBS -l walltime=1:00:00
```

```
#PBS -l ncpus=1
```

```
#PBS -l mem=1GB
```

```
#PBS -e myprog.err
```

```
#PBS -o myprog.out
```

```
#PBS -M hpatterton@sun.ac.za
```

```
#PBS -m abe
```

```
cd $PBS_O_WORKDIR
```

```
module load app/SPAdes/3.14.0
```

```
spades.py -o $PBS_O_WORKDIR --careful -1 $PBS_O_WORKDIR/reads/mutant_R1.fastq.gz
```

```
-2 $PBS_O_WORKDIR/reads/mutant_R2.fastq.gz
```

```
echo Done!
```

# SPAdes output

Name	Size (KB)	Last modified	Owner	Group	Access
..					
K21		2020-10-24 21:18	hpatterton	ldapuser	drwxr-xr-x
K33		2020-10-24 21:18	hpatterton	ldapuser	drwxr-xr-x
K55		2020-10-24 21:19	hpatterton	ldapuser	drwxr-xr-x
K77		2020-10-24 21:19	hpatterton	ldapuser	drwxr-xr-x
misc		2020-10-24 21:19	hpatterton	ldapuser	drwxr-xr-x
mismatch_corrector		2020-10-24 21:18	hpatterton	ldapuser	drwxr-xr-x
pipeline_state		2020-10-24 21:19	hpatterton	ldapuser	drwxr-xr-x
tmp		2020-10-24 21:19	hpatterton	ldapuser	drwxr-xr-x
assembly_graph.fastg	359	2020-10-24 21:19	hpatterton	ldapuser	-rw-r--r--
assembly_graph_with_scaffolds.gfa	176	2020-10-24 21:19	hpatterton	ldapuser	-rw-r--r--
before_rr.fasta	179	2020-10-24 21:19	hpatterton	ldapuser	-rw-r--r--
contigs.fasta	178	2020-10-24 21:19	hpatterton	ldapuser	-rw-r--r--
contigs.paths	1	2020-10-24 21:19	hpatterton	ldapuser	-rw-r--r--
dataset.info	1	2020-10-24 21:18	hpatterton	ldapuser	-rw-r--r--
input_dataset.yaml	1	2020-10-24 21:18	hpatterton	ldapuser	-rw-r--r--
params.txt	1	2020-10-24 21:18	hpatterton	ldapuser	-rw-r--r--
run_spades.sh	2	2020-10-24 21:18	hpatterton	ldapuser	-rw-r--r--
run_spades.yaml	4	2020-10-24 21:18	hpatterton	ldapuser	-rw-r--r--
scaffolds.fasta	178	2020-10-24 21:19	hpatterton	ldapuser	-rw-r--r--
scaffolds.paths	1	2020-10-24 21:19	hpatterton	ldapuser	-rw-r--r--
spades.log	133	2020-10-24 21:19	hpatterton	ldapuser	-rw-r--r--

- Always make sure that myprog.err is empty, and that contents of myprof.out has no information that could indicate problems
- The contigs.fasta and scaffolds.fasta in the top directory are the final output files
- Assembly graph and assembly graph with scaffold files are also produced
- You can view these files in *bandage* (see later)

# Output of contigs

```
>NODE_7_length_577_cov_6.114000
CATGCTAGCAAGTTAAGCGAACACTGACATGATAAATTAGTGGTTAGCTATATTTTTTTTA
CTTTGCAACAGAACCGAAAATAATCTCTTCAATTTATTTTTATATGAATCCTGTGACTCA
ATGATTGTAATATCTAAAGATTTTCAGTTCATCATAGACAATGTTCTTTTCAACATTTTTT
ATAGCAAATTGATTAAATAAATTCTCTAATTTCTCCCGTTTGATTTCACTACCATAGATT
ATATTATCATTGATATAGTCAATGAATAATGACAAATTATCACTCATAACAGTCCCAACC
CCTTTATTTTTGATAGACTAATTATCTTCATCATTGTAAAACAAATTACACCCTTTAAATT
TAACTCAACTTAAATATCGACAAATTAAAAACAATAAAATTACTTGAATATTATTCATA
ATATATTAACAACCTTTATTATACTGCTCTTTATATATAAAATCATTAATAATTAAACAAG
CCTTAAAATATTTAACTTTTTTGTGATTATTACACATTATCTTATCTGCTCTTTATCACC
ATAAAAATAGAAAAACAAGATTCCTAAAGAATATAG
>NODE_8_length_319_cov_11.933884
TCGTATTCTTCGACTGATAATTGCTCTCTAGATTCTAGCATATTTAAGTGTTCCTTTTA
TCTAATGCTTTGTCATATCCTTTAACGATTGAACCACTAAAGATTTCTCCTACTGCTCCT
GAACCATAACTAAATAGACATACTTTCTCTTCTGGTTGGAATGTGTGGTTCTGTAATAAC
GAAATTAACTTAAAGTATAATGATCCTGTATAAATGTTACCAACATCTCTATTCCATAAT
ACGGTTCCTGTTGCAAAGTTGAATTTATAGTATAATTTTAAACAAAAGGAGTCTTCTGTAT
GAACTATTTTCAGATATAAA
>NODE_9_length_155_cov_7.538462
T TACTCTTTCAGCCTTTTTTAAATTCAAGAATATGCAGAAGTTCAAAGTAATCAACATTAG
CGATTTTCTTTTCTCTCCATGGTCTCACTTTTCCACTTTTTGTCTTGTCCACTAAAACCC
TTGATTTTTTCATCTGAATAAATGCTACTATTAGTA
```