Hands-on with S. *cerevisiae* (~12 Mb, 16 chromosomes)

You will split in two teams (or more).

Your missions is to perform, compare and give information about different assemblies :

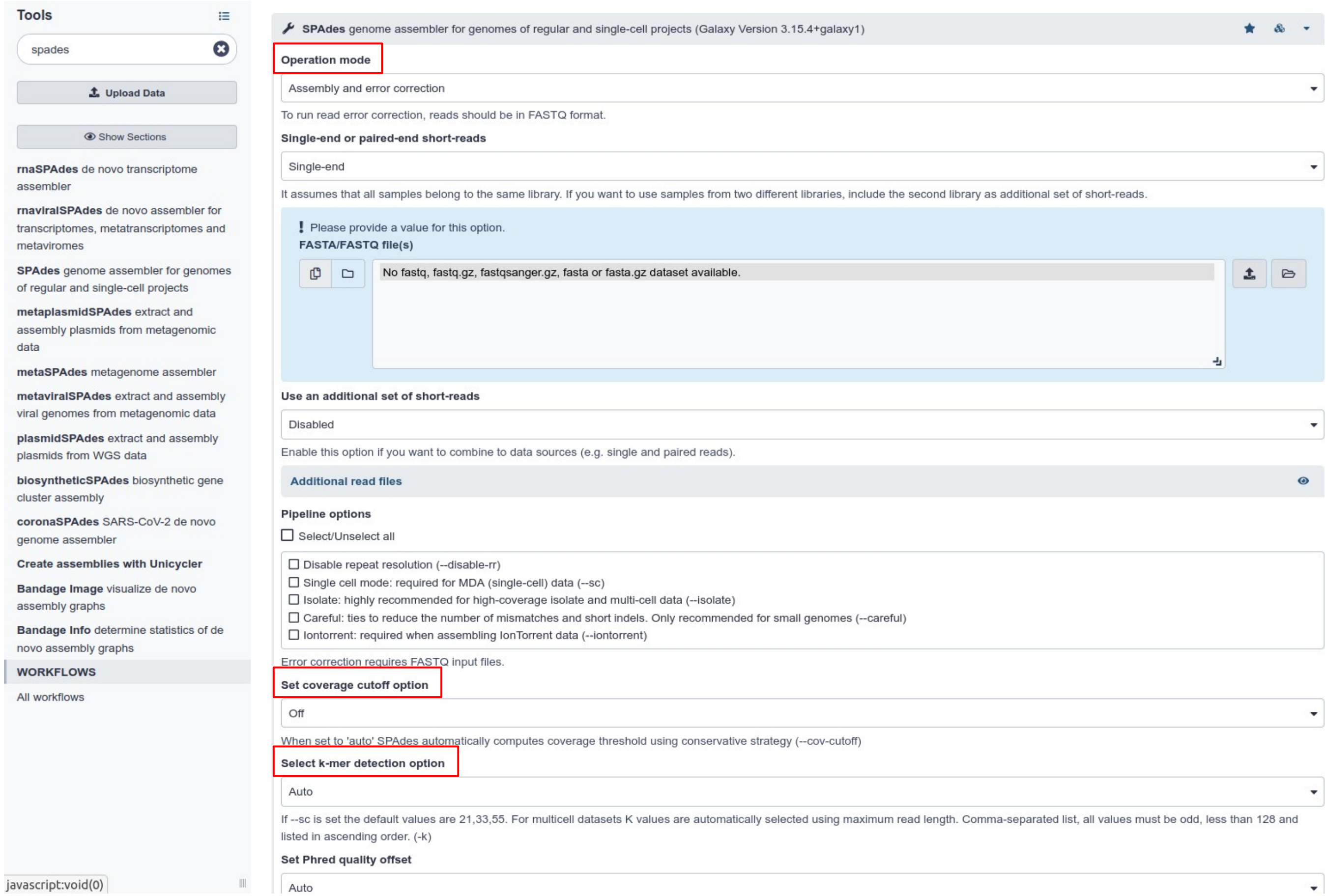
- using different data types (short reads (MiSeq PE), ONT reads (R9.4), HiFi reads),
- different coverages (from 5X to 60X),
- different software packages (SPAdes, flye, wtdbg2, Hifiasm).

You will compare the results using quast and dgenies (https://dgenies.toulouse.inra.fr/).

You will fill the table : https://lite.framacalc.org/f952lz03mb-9vey

You will present you findings (impact of data types, coverage, software package,...) : 5 minutes per team

SPAdes parameters



rnaSPAdes de novo transcriptome assembler

rnaviralSPAdes de novo assembler for transcriptomes, metatranscriptomes and metaviromes

SPAdes genome assembler for genomes of regular and single-cell projects

metaplasmidSPAdes extract and assembly plasmids from metagenomic data

metaSPAdes metagenome assembler

metaviralSPAdes extract and assembly viral genomes from metagenomic data

plasmidSPAdes extract and assembly plasmids from WGS data

biosyntheticSPAdes biosynthetic gene cluster assembly

coronaSPAdes SARS-CoV-2 de novo genome assembler

Create assemblies with Unicycler

Bandage Image visualize de novo assembly graphs

Bandage Info determine statistics of de novo assembly graphs

SPAdes genome assembler for genomes of regular and single-cell projects (Galaxy Version 3.15.4+galaxy1)

Operation mode

Assembly and error correction

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

Single-end or paired-end short-reads

Single-end

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.

! Please provide a value for this option.

FASTA/FASTQ file(s)

No fastq, fastq.gz, fastqsanger.gz, fasta or fasta.gz dataset available.

Use an additional set of short-reads

Disabled

Enable this option if you want to combine to data sources (e.g. single and paired reads).

Additional read files

Pipeline options

☐ Select/Unselect all

- ☐ Disable repeat resolution (--disable-rr)
- ☐ Single cell mode: required for MDA (single-cell) data (--sc)
- ☐ Isolate: highly recommended for high-coverage isolate and multi-cell data (--isolate)
- ☐ Careful: ties to reduce the number of mismatches and short indels. Only recommended for small genomes (--careful)
- ☐ Iontorrent: required when assembling IonTorrent data (--iontorrent)

Error correction requires FASTQ input files.

Set coverage cutoff option

Off

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

Select k-mer detection option

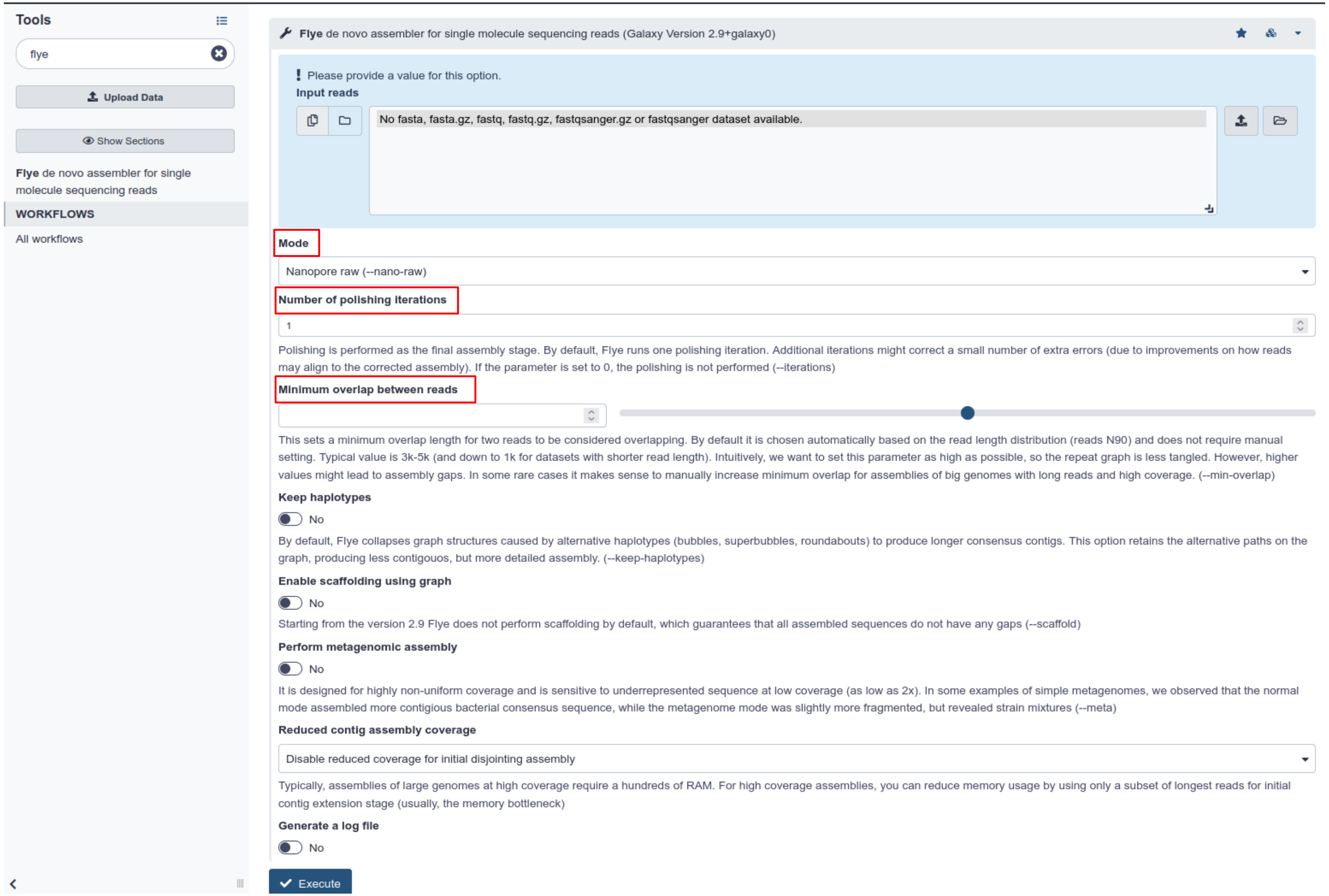
Auto

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)

Set Phred quality offset

Auto

Flye parameters



Flye de novo assembler for single molecule sequencing reads (Galaxy Version 2.9+galaxy0)

! Please provide a value for this option.

Input reads

No fasta, fasta.gz, fastq, fastq.gz, fastqsanger.gz or fastqsanger dataset available.

Mode

Nanopore raw (--nano-raw)

Number of polishing iterations

1

Polishing is performed as the final assembly stage. By default, Flye runs one polishing iteration. Additional iterations might correct a small number of extra errors (due to improvements on how reads may align to the corrected assembly). If the parameter is set to 0, the polishing is not performed (--iterations)

Minimum overlap between reads

This sets a minimum overlap length for two reads to be considered overlapping. By default it is chosen automatically based on the read length distribution (reads N90) and does not require manual setting. Typical value is 3k-5k (and down to 1k for datasets with shorter read length). Intuitively, we want to set this parameter as high as possible, so the repeat graph is less tangled. However, higher values might lead to assembly gaps. In some rare cases it makes sense to manually increase minimum overlap for assemblies of big genomes with long reads and high coverage. (--min-overlap)

Keep haplotypes

No

By default, Flye collapses graph structures caused by alternative haplotypes (bubbles, superbubbles, roundabouts) to produce longer consensus contigs. This option retains the alternative paths on the graph, producing less contigouos, but more detailed assembly. (--keep-haplotypes)

Enable scaffolding using graph

No

Starting from the version 2.9 Flye does not perform scaffolding by default, which guarantees that all assembled sequences do not have any gaps (--scaffold)

Perform metagenomic assembly

No

It is designed for highly non-uniform coverage and is sensitive to underrepresented sequence at low coverage (as low as 2x). In some examples of simple metagenomes, we observed that the normal mode assembled more contigious bacterial consensus sequence, while the metagenome mode was slightly more fragmented, but revealed strain mixtures (--meta)

Reduced contig assembly coverage

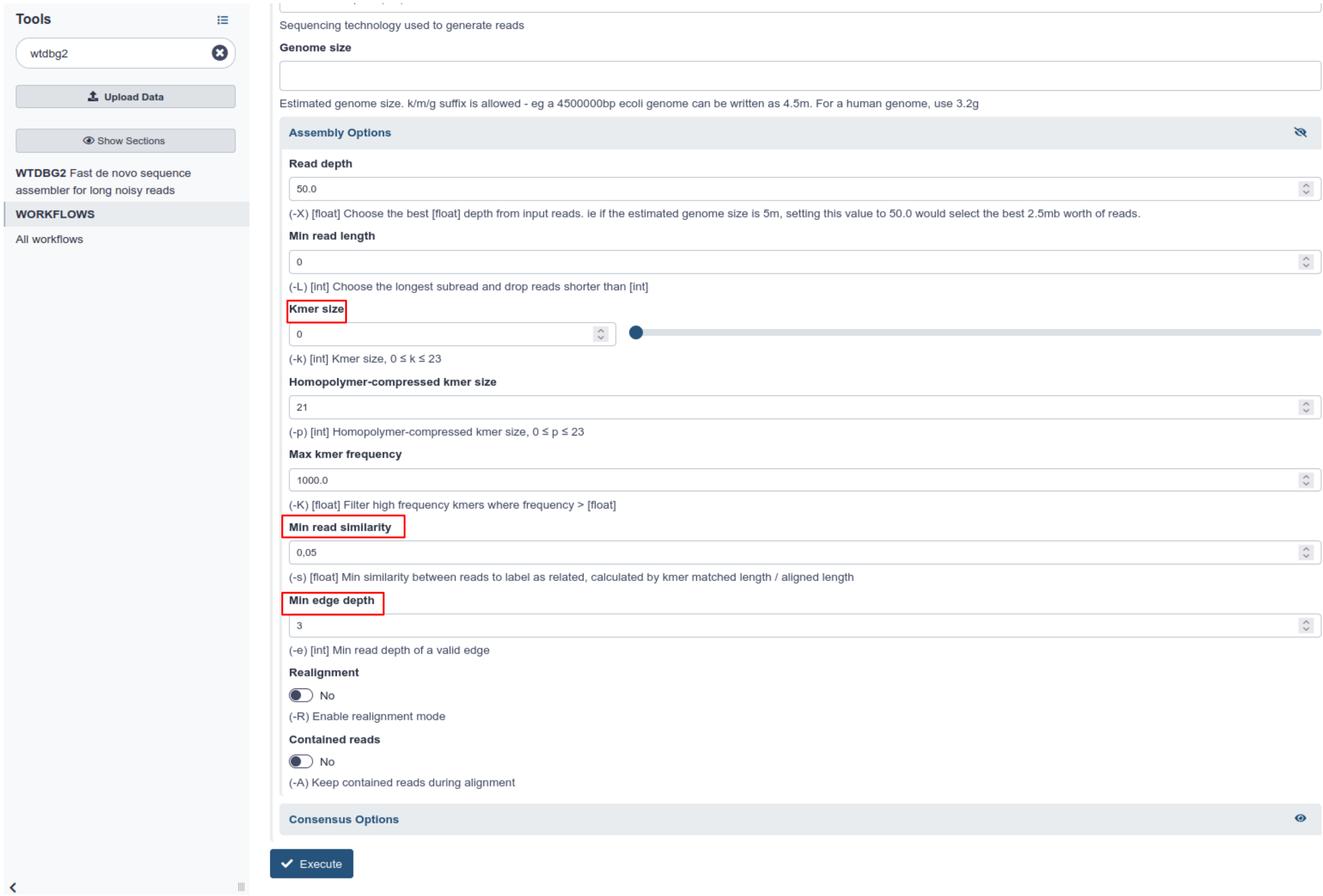
Disable reduced coverage for initial disjointing assembly

Typically, assemblies of large genomes at high coverage require a hundreds of RAM. For high coverage assemblies, you can reduce memory usage by using only a subset of longest reads for initial contig extension stage (usually, the memory bottleneck)

Generate a log file

No

wtdbg2 parameters



WTDBG2 Fast de novo sequence assembler for long noisy reads

Sequencing technology used to generate reads

Genome size

Estimated genome size. k/m/g suffix is allowed - eg a 4500000bp ecoli genome can be written as 4.5m. For a human genome, use 3.2g

Assembly Options

Read depth

50.0

(-X) [float] Choose the best [float] depth from input reads. ie if the estimated genome size is 5m, setting this value to 50.0 would select the best 2.5mb worth of reads.

Min read length

0

(-L) [int] Choose the longest subread and drop reads shorter than [int]

Kmer size

0

(-k) [int] Kmer size, 0 ≤ k ≤ 23

Homopolymer-compressed kmer size

21

(-p) [int] Homopolymer-compressed kmer size, 0 ≤ p ≤ 23

Max kmer frequency

1000.0

(-K) [float] Filter high frequency kmers where frequency > [float]

Min read similarity

0,05

(-s) [float] Min similarity between reads to label as related, calculated by kmer matched length / aligned length

Min edge depth

3

(-e) [int] Min read depth of a valid edge

Realignment

No

(-R) Enable realignment mode

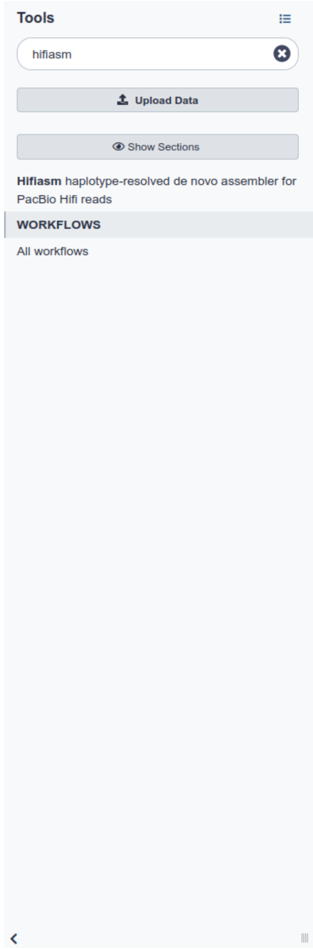
Contained reads

No

(-A) Keep contained reads during alignment

Consensus Options

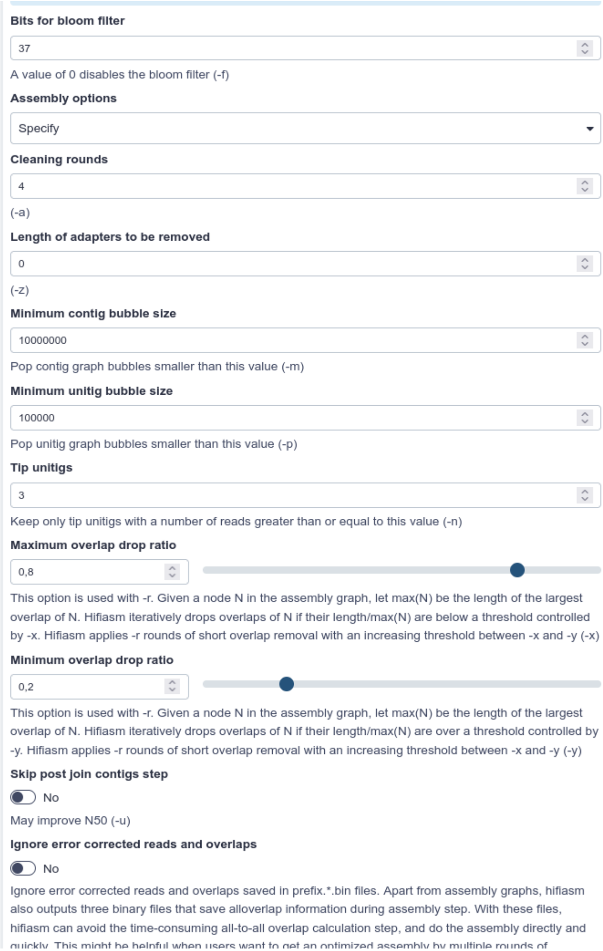
Hifiasm parameters



Hifiasm haplotype-resolved de novo assembler for PacBio Hifi reads

WORKFLOWS

All workflows



Bits for bloom filter

37

A value of 0 disables the bloom filter (-f)

Assembly options

Specify

Cleaning rounds

4

(-a)

Length of adapters to be removed

0

(-z)

Minimum contig bubble size

10000000

Pop contig graph bubbles smaller than this value (-m)

Minimum unitig bubble size

100000

Pop unitig graph bubbles smaller than this value (-p)

Tip unitigs

3

Keep only tip unitigs with a number of reads greater than or equal to this value (-n)

Maximum overlap drop ratio

0,8

This option is used with -r. Given a node N in the assembly graph, let max(N) be the length of the largest overlap of N. Hifiasm iteratively drops overlaps of N if their length/max(N) are below a threshold controlled by -x. Hifiasm applies -r rounds of short overlap removal with an increasing threshold between -x and -y (-x)

Minimum overlap drop ratio

0,2

This option is used with -r. Given a node N in the assembly graph, let max(N) be the length of the largest overlap of N. Hifiasm iteratively drops overlaps of N if their length/max(N) are over a threshold controlled by -y. Hifiasm applies -r rounds of short overlap removal with an increasing threshold between -x and -y (-y)

Skip post join contigs step

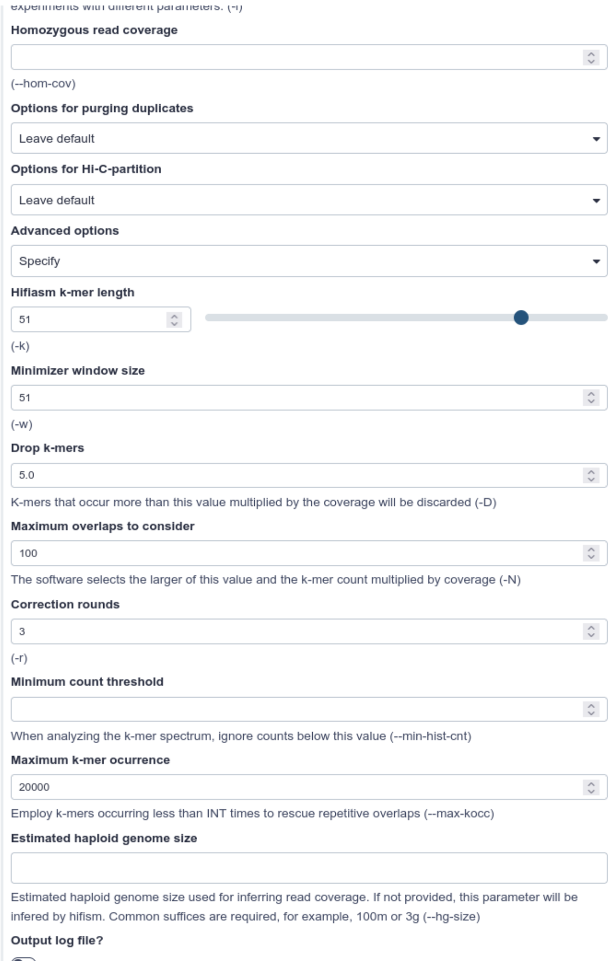
No

May improve N50 (-u)

Ignore error corrected reads and overlaps

No

Ignore error corrected reads and overlaps saved in prefix.*.bin files. Apart from assembly graphs, hifiasm also outputs three binary files that save alloverlap information during assembly step. With these files, hifiasm can avoid the time-consuming all-to-all overlap calculation step, and do the assembly directly and quickly. This might be helpful when users want to get an optimized assembly by multiple rounds of



experiments with different parameters. (-i)

Homozygous read coverage

(--hom-cov)

Options for purging duplicates

Leave default

Options for Hi-C-partition

Leave default

Advanced options

Specify

Hifiasm k-mer length

51

(-k)

Minimizer window size

51

(-w)

Drop k-mers

5.0

K-mers that occur more than this value multiplied by the coverage will be discarded (-D)

Maximum overlaps to consider

100

The software selects the larger of this value and the k-mer count multiplied by coverage (-N)

Correction rounds

3

(-r)

Minimum count threshold

When analyzing the k-mer spectrum, ignore counts below this value (--min-hist-cnt)

Maximum k-mer ocurrence

20000

Employ k-mers occurring less than INT times to rescue repetitive overlaps (--max-kocc)

Estimated haploid genome size

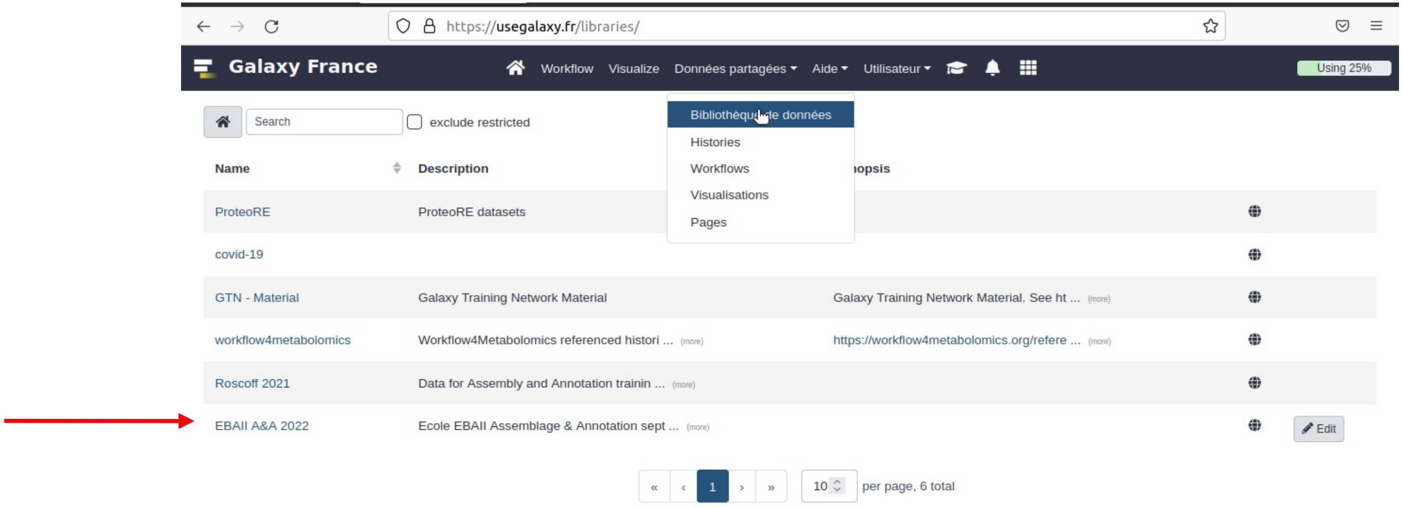
Estimated haploid genome size used for inferring read coverage. If not provided, this parameter will be infered by hifism. Common suffices are required, for example, 100m or 3g (--hg-size)

Output log file?

And

Processing will be performed using :https://usegalaxy.fr/

The data files are located at : shared libraries/EBAII A&A 2022/Assembly



Name	Description	iopsis	
ProteoRE	ProteoRE datasets		🌐
covid-19			🌐
GTN - Material	Galaxy Training Network Material	Galaxy Training Network Material. See ht ... (more)	🌐
workflow4metabolomics	Workflow4Metabolomics referenced histori ... (more)	https://workflow4metabolomics.org/refere ... (more)	🌐
Roscoff 2021	Data for Assembly and Annotation trainin ... (more)		🌐
EBAII A&A 2022	Ecole EBAII Assemblage & Annotation sept ... (more)		🌐 Edit

You can ask question using : https://semestriel.framapad.org/p/abgi5b9vdm-9vey?lang=fr

Let’s go!