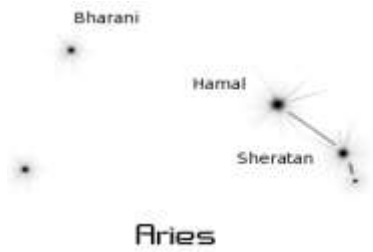
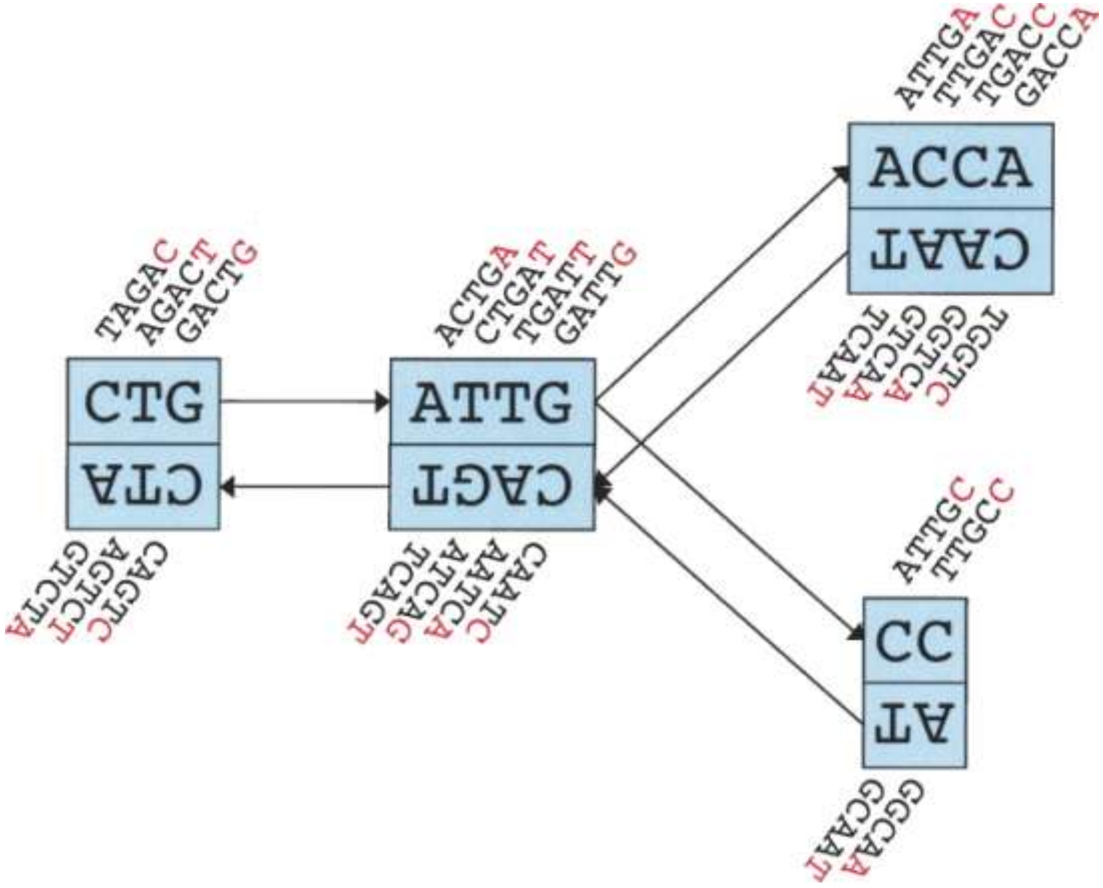
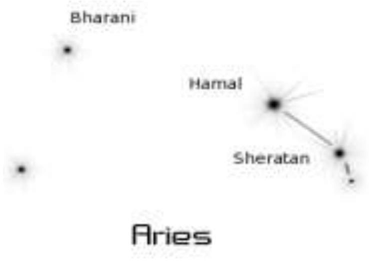


From Reads To Contigs

(in 30 min)





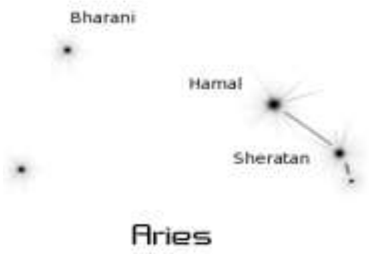
Before Starting…

- Which reads I’ve got? Length? Quality? Amount? Estimated coverage?

- Wwhich assembler? What parameters?

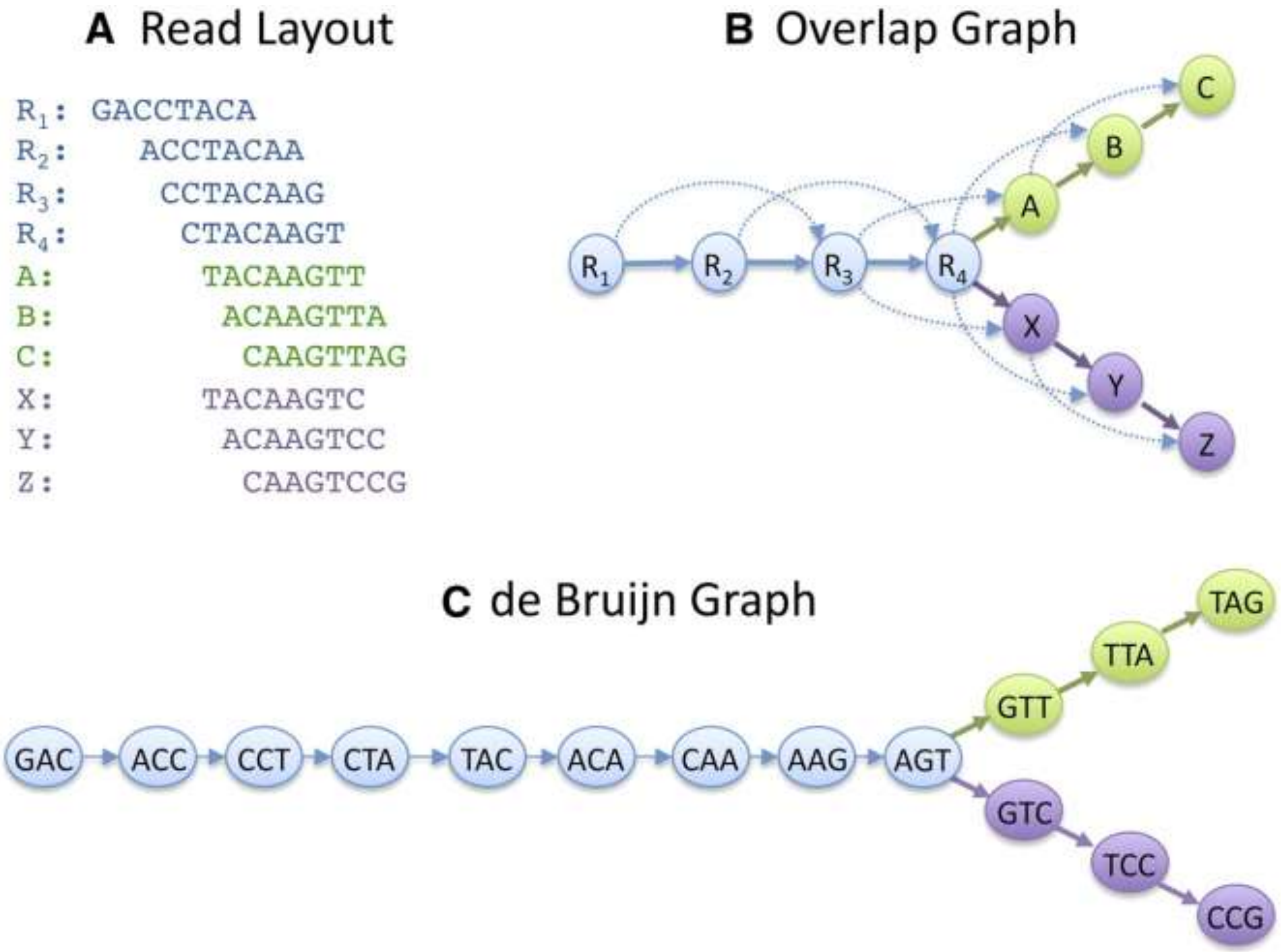
- Is there any potential reference? Is it close enough?

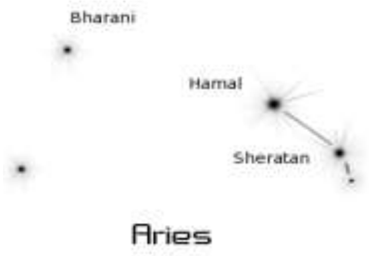
- Contigs Metrics
- Scaffolding
- Contigs Ordering
- Are my contigs good enough to be annotated?



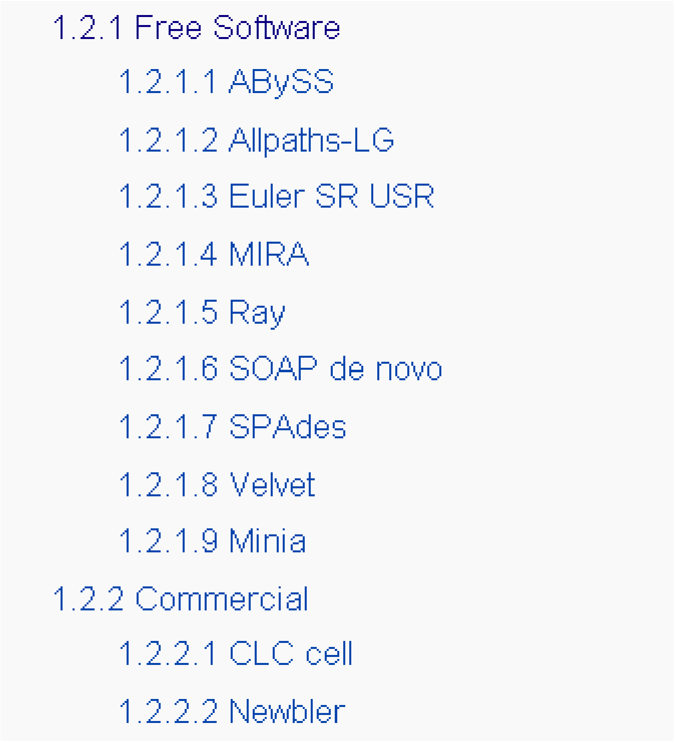
What Assembler?

Three main algorythms





Plethora of Software



- 1.2.1 Free Software
  - 1.2.1.1 ABySS
  - 1.2.1.2 Allpaths-LG
  - 1.2.1.3 Euler SR USR
  - 1.2.1.4 MIRA
  - 1.2.1.5 Ray
  - 1.2.1.6 SOAP de novo
  - 1.2.1.7 SPAdes
  - 1.2.1.8 Velvet
  - 1.2.1.9 Minia
- 1.2.2 Commercial
  - 1.2.2.1 CLC cell
  - 1.2.2.2 Newbler

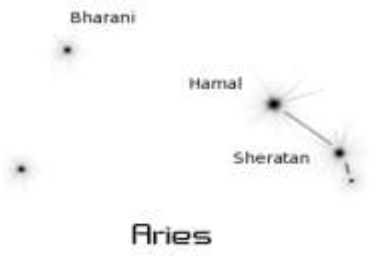
Different Algorythms

Different Requirements

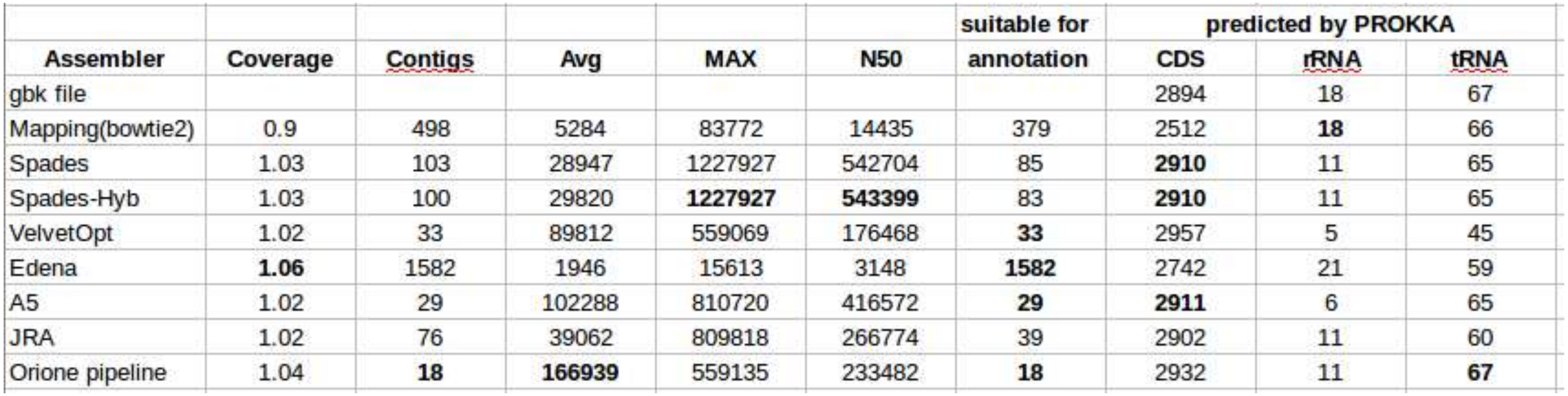
Different Performances

Different platforms, SE/PE

What can I choose?

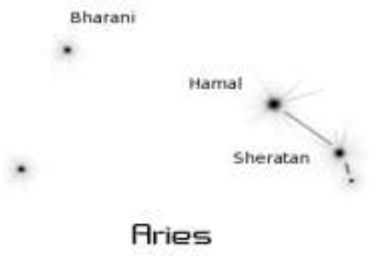


The only answer is trying…



						suitable for	predicted by PROKKA		
Assembler	**Coverage**	**Contigs**	**Avg**	**MAX**	**N50**	**annotation**	**CDS**	**rRNA**	**tRNA**
gbk file							2894	18	67
Mapping(bowtie2)	0.9	498	5284	83772	14435	379	2512	**18**	66
Spades	1.03	103	28947	1227927	542704	85	**2910**	11	65
Spades-Hyb	1.03	100	29820	**1227927**	**543399**	83	**2910**	11	65
VelvetOpt	1.02	33	89812	559069	176468	**33**	2957	5	45
Edena	**1.06**	1582	1946	15613	3148	**1582**	2742	21	59
A5	1.02	29	102288	810720	416572	**29**	**2911**	6	65
JRA	1.02	76	39062	809818	266774	39	2902	11	60
Orione pipeline	1.04	**18**	**166939**	559135	233482	**18**	2932	11	**67**

Data from:
Listeria monocytogenes, Illumina next500, cov > 100x, 135+135PE

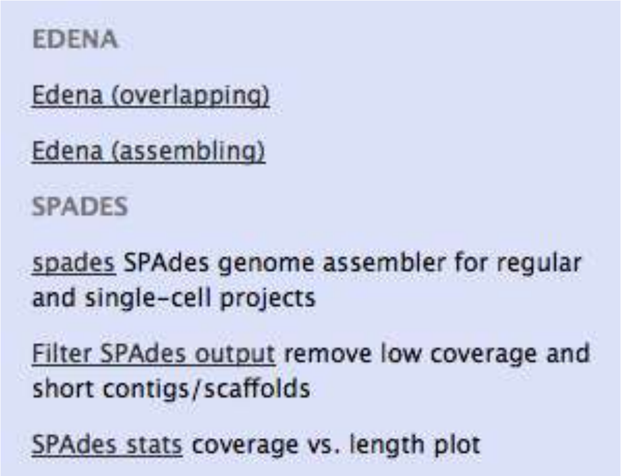


Our Suggested Strategy

Platform **Ion torrent**: Spades (best performances)

Illumina: **Edena** (best speed and resources management)

Coming Soon:
Integrated pipelines A5_miseq, JRA, ..



EDENA

Edena (overlapping)

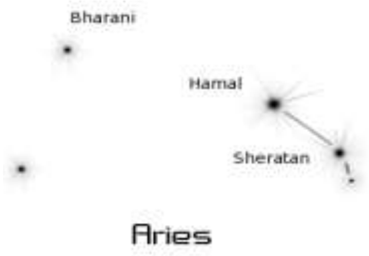
Edena (assembling)

SPADES

spades SPAdes genome assembler for regular and single-cell projects

Filter SPAdes output remove low coverage and short contigs/scaffolds

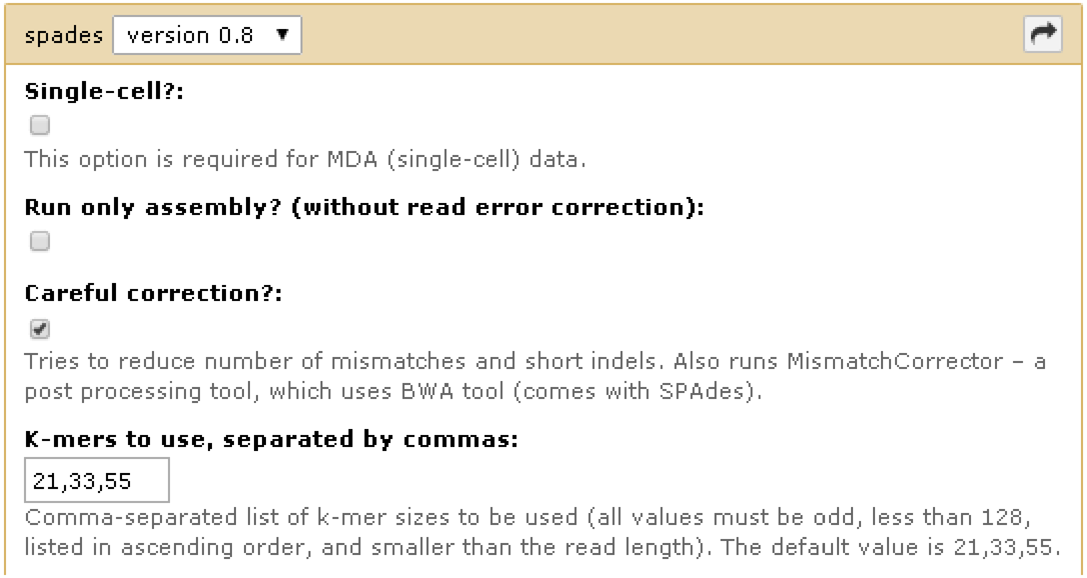
SPAdes stats coverage vs. length plot



Spades 1

Algorythm section

Kmer section



spades version 0.8 ▼

Single-cell?:

□

This option is required for MDA (single-cell) data.

Run only assembly? (without read error correction):

□

Careful correction?:

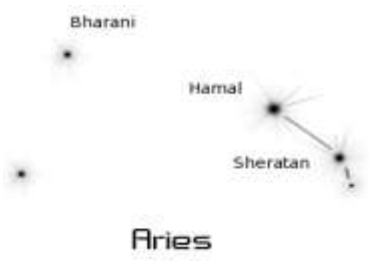
☑

Tries to reduce number of mismatches and short indels. Also runs MismatchCorrector – a post processing tool, which uses BWA tool (comes with SPAdes).

K-mers to use, separated by commas:

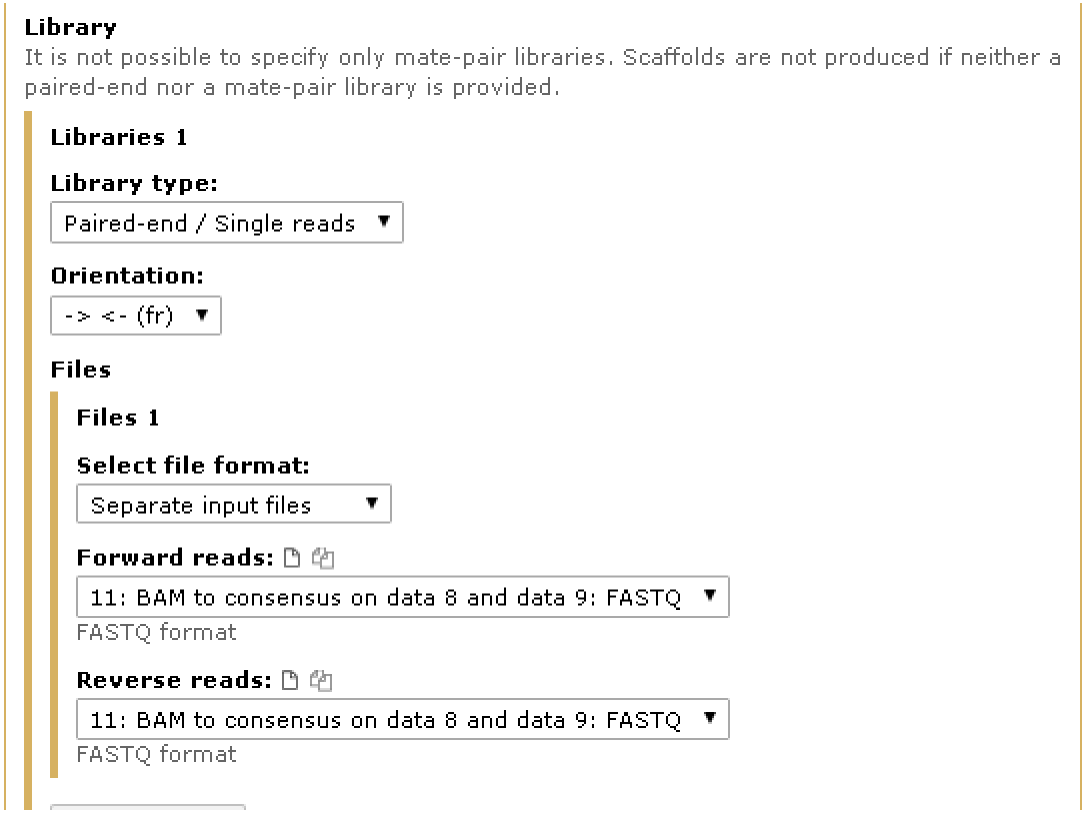
21,33,55

Comma-separated list of k-mer sizes to be used (all values must be odd, less than 128, listed in ascending order, and smaller than the read length). The default value is 21,33,55.



Spades 2

Library section



Library

It is not possible to specify only mate-pair libraries. Scaffolds are not produced if neither a paired-end nor a mate-pair library is provided.

Libraries 1

Library type: Paired-end / Single reads ▼

Orientation: -> <- (fr) ▼

Files

Files 1

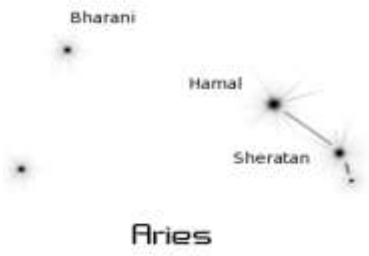
Select file format: Separate input files ▼

Forward reads: 11: BAM to consensus on data 8 and data 9: FASTQ ▼

FASTQ format

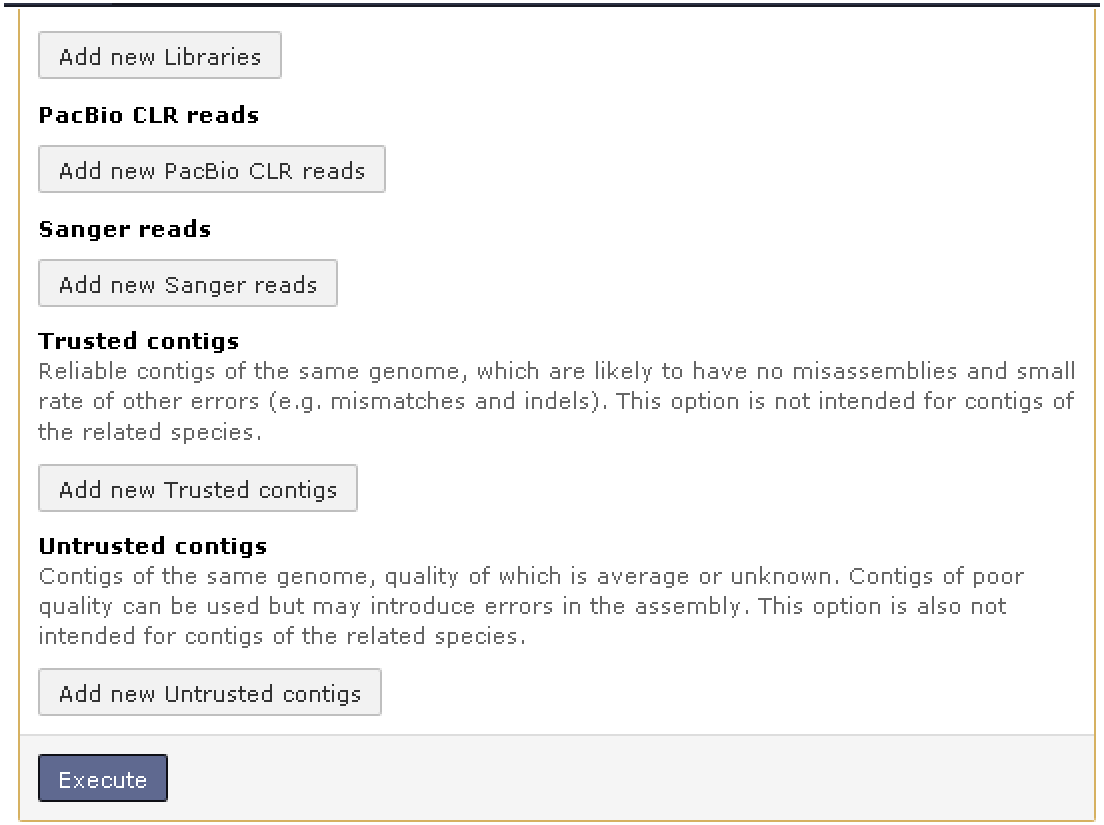
Reverse reads: 11: BAM to consensus on data 8 and data 9: FASTQ ▼

FASTQ format



Spades 3

Additional Input section



PacBio CLR reads

Sanger reads

Trusted contigs

Reliable contigs of the same genome, which are likely to have no misassemblies and small rate of other errors (e.g. mismatches and indels). This option is not intended for contigs of the related species.

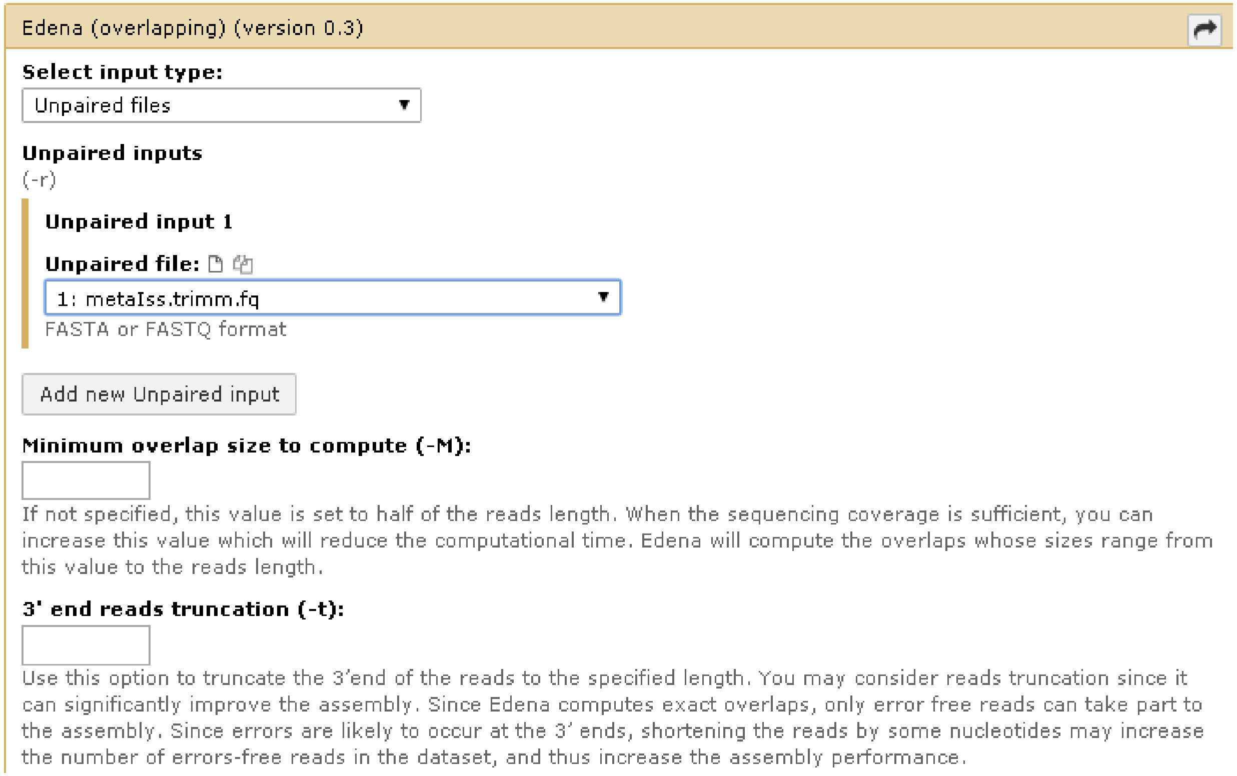
Untrusted contigs

Contigs of the same genome, quality of which is average or unknown. Contigs of poor quality can be used but may introduce errors in the assembly. This option is also not intended for contigs of the related species.

Edena Overlapping

Input section

Overlapping parameters



Edena (overlapping) (version 0.3)

Select input type:

Unpaired files ▾

Unpaired inputs

(-r)

Unpaired input 1

Unpaired file:

1: metaIss.trimm.fq ▾

FASTA or FASTQ format

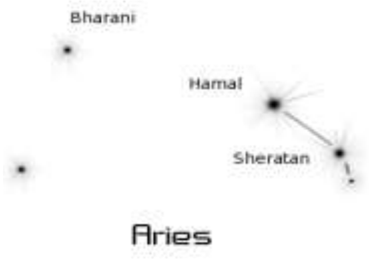
Minimum overlap size to compute (-M):

If not specified, this value is set to half of the reads length. When the sequencing coverage is sufficient, you can increase this value which will reduce the computational time. Edena will compute the overlaps whose sizes range from this value to the reads length.

3' end reads truncation (-t):

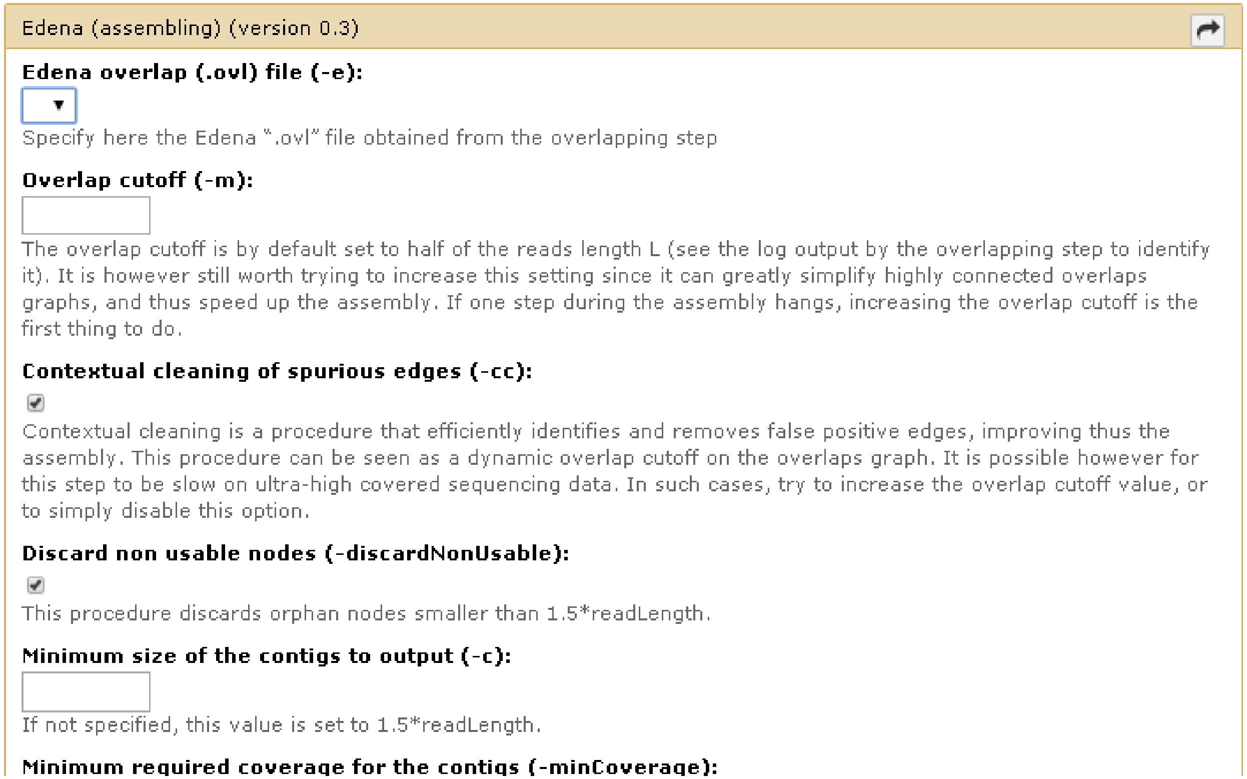
Use this option to truncate the 3'end of the reads to the specified length. You may consider reads truncation since it can significantly improve the assembly. Since Edena computes exact overlaps, only error free reads can take part to the assembly. Since errors are likely to occur at the 3' ends, shortening the reads by some nucleotides may increase the number of errors-free reads in the dataset, and thus increase the assembly performance.

Not suitable for
Ion torrent



Edena Assembling 1

Extension parameters



Edena (assembling) (version 0.3)

Edena overlap (.ovl) file (-e):

▾

Specify here the Edena ".ovl" file obtained from the overlapping step

Overlap cutoff (-m):

The overlap cutoff is by default set to half of the reads length L (see the log output by the overlapping step to identify it). It is however still worth trying to increase this setting since it can greatly simplify highly connected overlaps graphs, and thus speed up the assembly. If one step during the assembly hangs, increasing the overlap cutoff is the first thing to do.

Contextual cleaning of spurious edges (-cc):

☑

Contextual cleaning is a procedure that efficiently identifies and removes false positive edges, improving thus the assembly. This procedure can be seen as a dynamic overlap cutoff on the overlaps graph. It is possible however for this step to be slow on ultra-high covered sequencing data. In such cases, try to increase the overlap cutoff value, or to simply disable this option.

Discard non usable nodes (-discardNonUsable):

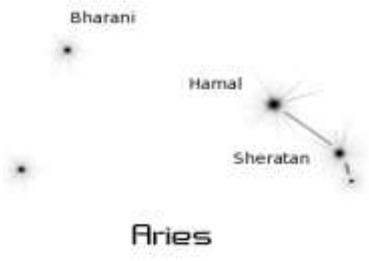
☑

This procedure discards orphan nodes smaller than 1.5*readLength.

Minimum size of the contigs to output (-c):

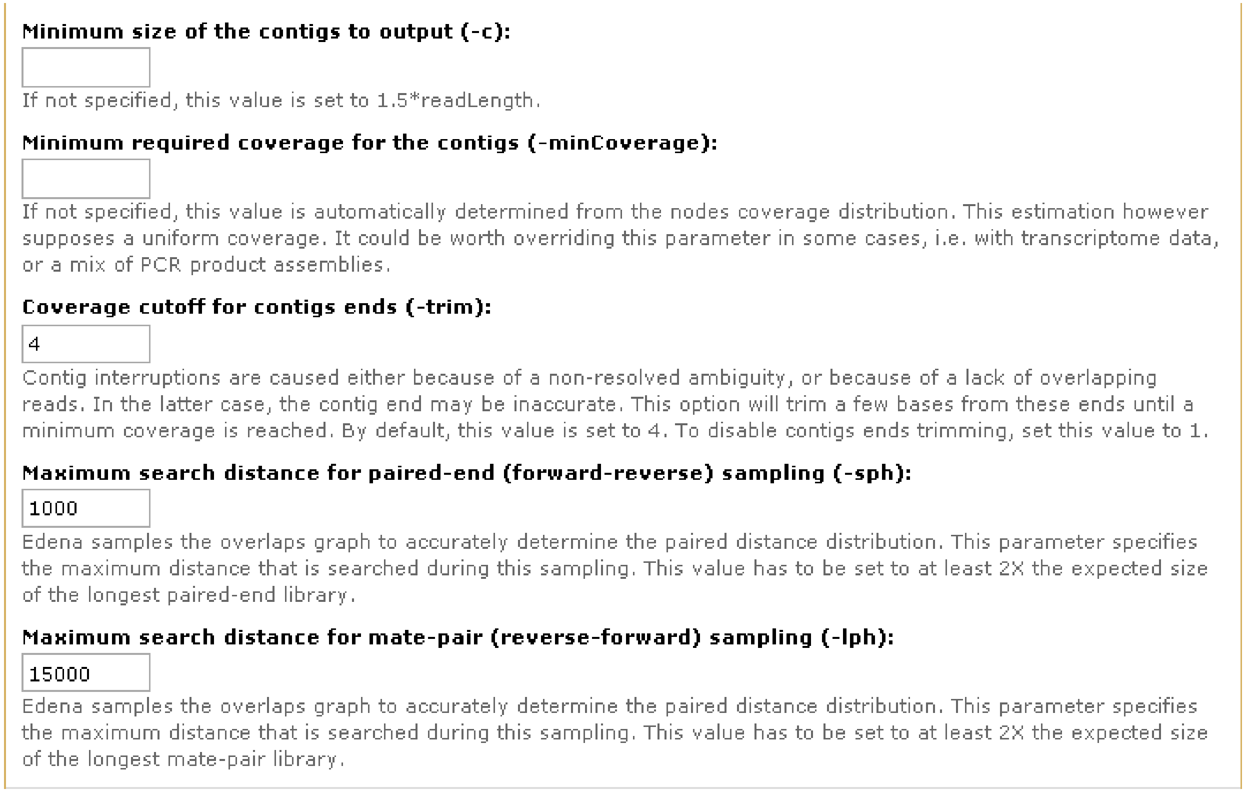
If not specified, this value is set to 1.5*readLength.

Minimum required coverage for the contigs (-minCoverage):



Edena Assembling 2

Contigs filters



Minimum size of the contigs to output (-c):

If not specified, this value is set to 1.5*readLength.

Minimum required coverage for the contigs (-minCoverage):

If not specified, this value is automatically determined from the nodes coverage distribution. This estimation however supposes a uniform coverage. It could be worth overriding this parameter in some cases, i.e. with transcriptome data, or a mix of PCR product assemblies.

Coverage cutoff for contigs ends (-trim):

4

Contig interruptions are caused either because of a non-resolved ambiguity, or because of a lack of overlapping reads. In the latter case, the contig end may be inaccurate. This option will trim a few bases from these ends until a minimum coverage is reached. By default, this value is set to 4. To disable contigs ends trimming, set this value to 1.

Maximum search distance for paired-end (forward-reverse) sampling (-sph):

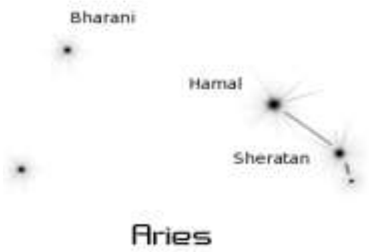
1000

Edena samples the overlaps graph to accurately determine the paired distance distribution. This parameter specifies the maximum distance that is searched during this sampling. This value has to be set to at least 2X the expected size of the longest paired-end library.

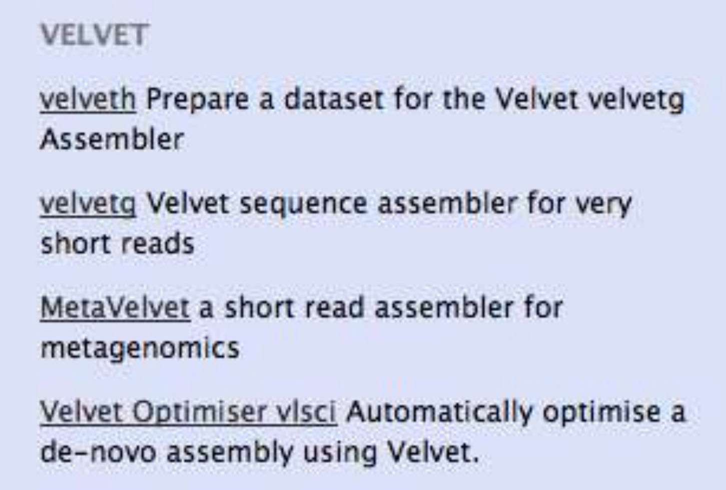
Maximum search distance for mate-pair (reverse-forward) sampling (-lph):

15000

Edena samples the overlaps graph to accurately determine the paired distance distribution. This parameter specifies the maximum distance that is searched during this sampling. This value has to be set to at least 2X the expected size of the longest mate-pair library.



Other Assemblers



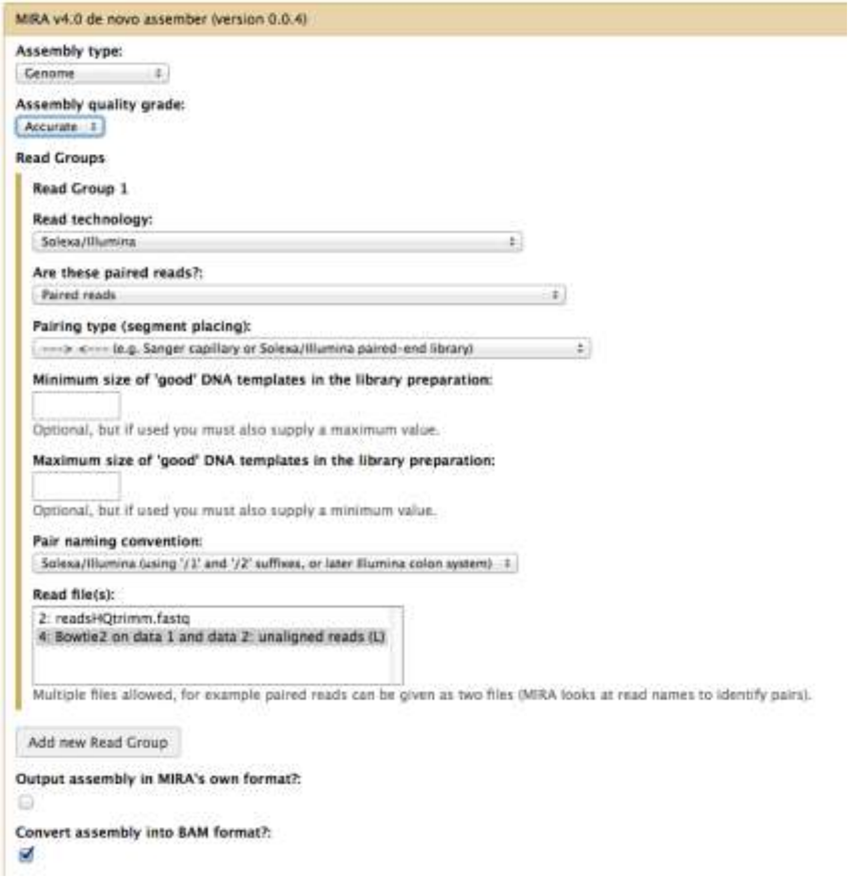
VELVET

velveth Prepare a dataset for the Velvet velvetg Assembler

velvetg Velvet sequence assembler for very short reads

MetaVelvet a short read assembler for metagenomics

Velvet Optimiser vlsci Automatically optimise a de-novo assembly using Velvet.



MIRA v4.0 de novo assember (version 0.0.4)

Assembly type:
Genome

Assembly quality grade:
Accurate

Read Groups

Read Group 1

Read technology:
Solexa/Illumina

Are these paired reads?:
Paired reads

Pairing type (segment placing):
---> <--- (e.g. Sanger capillary or Solexa/Illumina paired-end library)

Minimum size of 'good' DNA templates in the library preparation:

Optional, but if used you must also supply a maximum value.

Maximum size of 'good' DNA templates in the library preparation:

Optional, but if used you must also supply a minimum value.

Pair naming convention:
Solexa/Illumina (using '/1' and '/2' suffixes, or later Illumina colon system)

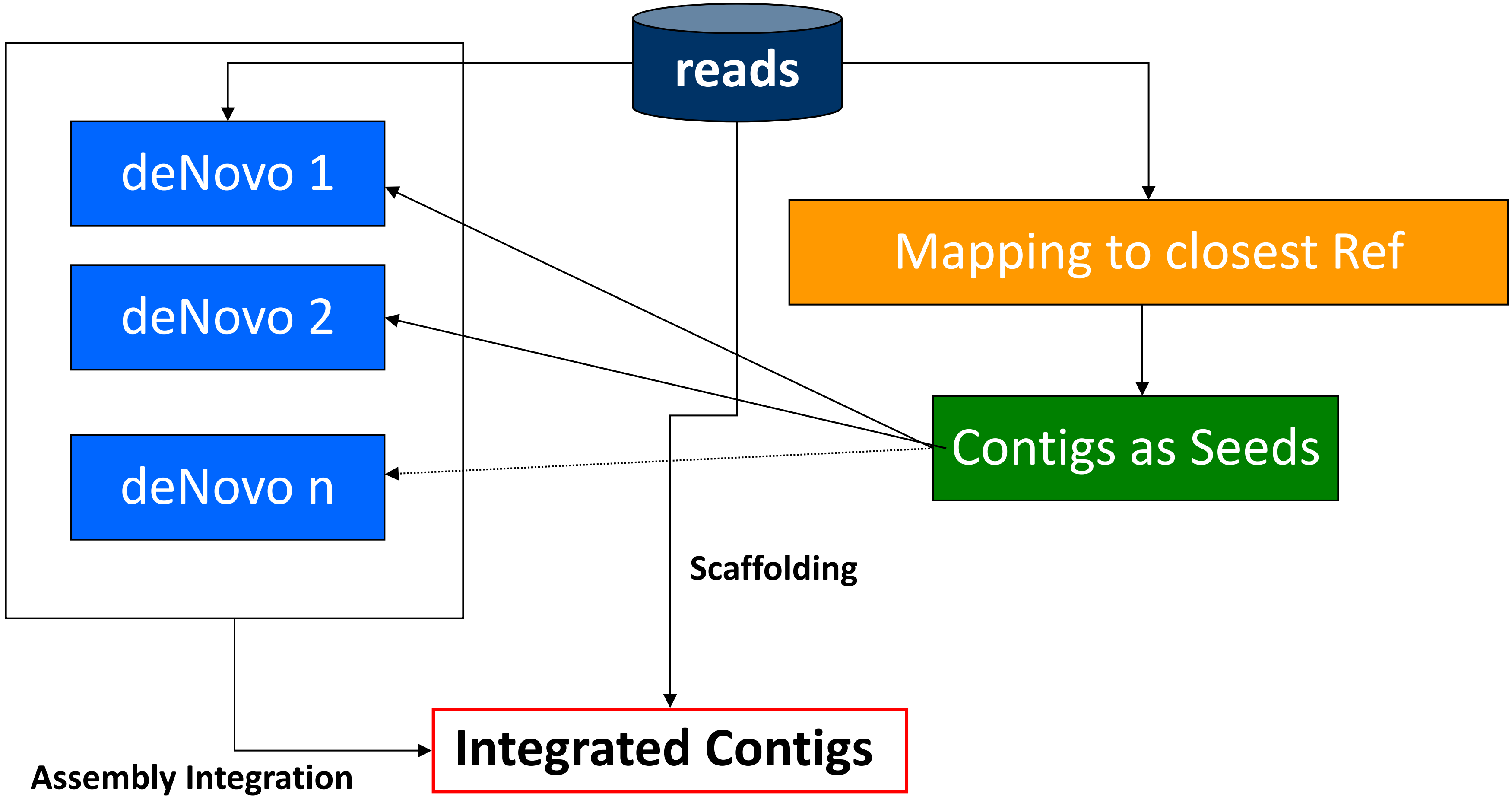
Read file(s):
2: readsHQtrimm.fastq
4: Bowtie2 on data 1 and data 2: unaligned reads (L)

Multiple files allowed, for example paired reads can be given as two files (MIRA looks at read names to identify pairs).

Output assembly in MIRA's own format?:
□

Convert assembly into BAM format?:
☑

Hybrid Strategy



reads

deNovo 1

deNovo 2

deNovo n

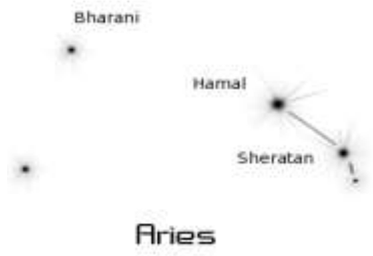
Mapping to closest Ref

Contigs as Seeds

Scaffolding

Integrated Contigs

Assembly Integration
CISA, Blast, Mauve



After Assembling?

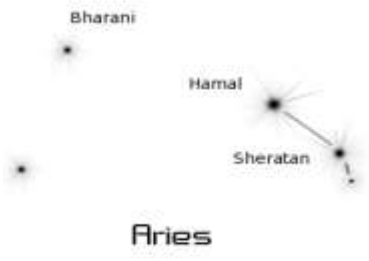
Contigs Evaluation (metrics or specialized sftw es quast)

Scaffolding

Contigs Ordering

Contigs integration

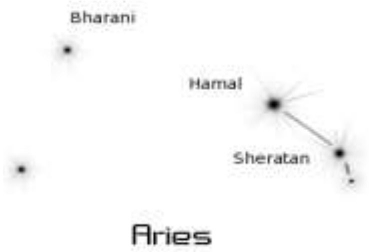
Are my contigs good enough? Annotate them!



Suggestion…

Try to assembly a dataset with the three softwares…

Please consider:
cpu time
metrics
unmapping reads over contigs



Time out

are you ok?

any question?