The Edena v3 assembler, development version 110814 Reference manual

License

Edena is free only for non-profit academic uses. Using this program implies the acceptance by the user of the terms of the license "LICENSE.TXT" included in the package.

Development version

You downloaded a development version of Edena. This means that some features are incomplete and that the program has not been thoroughly tested. If you experience weird results or behaviors, please report the case to david.hernandez@genomic.ch.

Overview

A quick description of the program options is obtained by typing: `edena -h`

Edena is an overlaps graph based short reads assembler. The maximum supported read length is 128 nucleotides. However, reads that are too long may impair the assembly performance. In such a case, you should consider the "3' end reads truncation" procedure described below.

This program requires the reads to be all the same length, as Illumina GA reads are. This is due to historical reasons and because it greatly simplifies several computational steps. 454 or Sanger reads are therefore not suited to Edena. If you provide multiple files with different read lengths, Edena will trim the 3' end of the reads so that the reads are all the same length as the shortest reads in the file.

The program was developed in a framework of whole genome bacterial assemblies. It is therefore more suited for this kind of task though we also successfully used it for other types of projects.

An assembly with Edena is a two step process: *overlapping* and *assembling*:

Overlapping mode: The reads files are provided to the program which computes the transitively reduced overlaps graph. This structure is then stored together with the sequence reads in a binary file suffixed with ".ovl".

Assembling mode: The ".ovl" file is provided to the program, as well as some assembly parameters. A set of contigs in FASTA format is outputted. The purpose of having a two step process is that the .ovl file is computed only once and can then be used to produce assemblies with different parameters.

Usage: Overlapping mode

Edena can accept both unpaired and paired files, fastq and fasta format. Note that for technical reasons, all reads are required to be of the same length. You can however provide the program with different files containing different reads length. In such case, Edena will trim the 3' ends of the longer reads so that they fit the shorter length. It is however required that reads within each individual file are the same length (as Illumina GA reads are). (See also the 3'end reads truncation option).

-r <files…> **unpaired files**

Unpaired files are provided using the flag -r

```
>edena -r file1 file2 …
```

-paired <files…> **paired-end files**

Paired files are specified by pairs, multiple pairs can be specified. Edena samples the inserts length distribution independently for each pair of files.

```
>edena -paired file1_pair1 file1_pair2 file2_pair1 file2_pair2 …
```

-peHorizon <value> **maximum search distance for paired-end reads connection**

Edena samples the overlaps graph to accurately determine the paired distance distribution. This parameter specifies the maximum distance that is searched during this sampling. By default, this value is set to 500 nucleotides. If the insert size of the sequenced sample (or of one of the sequenced sample) is larger than 400 approximately, you may increase this value to 1.5 x the expected library size.

-t <value> **3' end reads truncation**

Use this option to truncate the 3'end of the reads such that the resulting length is <value>. 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.

-M <value> **minimum overlap size to compute**

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.

-p <name> **prefix name for output files**

Usage: Assembling mode

-e <.ovl file> **Edena ".ovl" file**

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

-m <value> **overlap cutoff**

The optimal setting of this parameter may require some trials. However, this setting is now less critical than it was for the version 2 of Edena, where it had to be specified in a way that allowed for elimination of as many false positive overlaps as possible, while preserving true positive overlaps. The current version now includes a contextual cleaning enabled by default (see next option –cc) which does much of the job.

If the overlap cutoff is not specified, the minimum overlaps size will be the one that was used during the computation of the overlaps. 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…

-cc <yes/no> **contextual cleaning**

This option is enabled by default. 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 very slow, particularly on ultra high covered sequencing data. In such cases, try to increase the overlap cutoff value (-m), or to simply disable this option by using "-cc no".

- discardNonUsable <yes/no>

Enabled by default, this procedure discards nodes smaller than 1.5*readLength and that are not connected to any other nodes.

-c <value> **minimum size of the contigs to output**

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

-minCoverage <value> **minimum required coverage for the contigs**

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.

-trim <value> **coverage cutoff for contigs ends**

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 such 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.

Suggestions, comment, bugs

david.hernandez@genomic.ch.