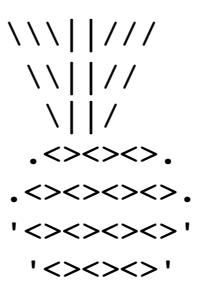

Ananas Software Manual

Version 0.0.0

1. Requirements

Ananas runs on Linux 64 and requires gcc. (tested versions include gcc 4.6.3 and gcc 4.8.2).

2. Installation

Check out the source code:

```
> svn co svn+ssh://<username>@snigel.imbim.uu.se/storage/svn/ryggrad
```

Compile:

```
> cd ryggrad
> make -j 20
```

or, to enable OpenMPI support:

```
> make -j 20 OPEN_MP=yes
```

Note: the overlap finder currently uses parallelization via OpenMPI. This will change.

3. Test

To run a small test data set:

```
> ./test_ananas
```

At the end, the script calls diff to compare outputs. If there are no differences, the software works correctly.

4. Assemble RNA-Seq data

Run

```
./Ananas
```

with the following parameters (for more detailed explanation, see below):

```
-i<string> : input fasta file
-o<string> : output directory (def=ananas_out)
-m<double> : minimum overlap identity (def=0.99)
-dir<string> : direction of pairs: fr fowards each other, ff same direction, na
```

```
unpaired
-b<int> : bandwidth of alignments (maximum indel size) (def=0)
-strand<string> : strand specificity (0=no 1=yes) (def=0)
-ml<int> : minimum overlap (for alignments) (def=25)
-s<int> : step size (for alignments) (def=30)
-n<int> : number of CPU cores (def=1)
-no<int> : number of processes for overlap finding (def=1)
-group<bool> : group identical reads (recommended for large data sets) (def=0)
```

Input and sequencing technology: Ananas requires the input data in a single fasta file. The current version infers read pairing from the read names (“/1” and “/2”) as provided by Illumina data. The parameter '-dir' specifies the relative orientation of read pairs, i.e. 'fr' for reads pointing towards each other, 'ff' for reads pointing in the same direction, and 'na' for unpaired reads. This parameter is mandatory.

Error patterns are assumed to be mismatches by default. To process data with homopolymer indels (454, IonProton), set the alignment bandwidth to '-b 2'. The option '-strand' will preserve the strandedness of the data. Note that the results are currently output as reverse complement. Note: to build contig consensus for data with indels, run ./KmerKonsensus at the end.

Parallelization: overlap finding currently uses OpenMPI (to be changed), whereas layout uses native multithreading. Use the '-n' parameter to specify the number of threads. Note: the final processes are parallelized as processes and might require significant amounts of memory (currently, each process loads the entire set of reads).

Outputs: the final output is in contigs_altsplic.fasta (all contig sequences) and contigs_altsplic.layout (read placements). Contigs are listed (not necessarily in order) following the convention 'Contig_<process id>_<locus id>_<transcript id>' (similar to Trinity), where the process id indicates which process or thread processed this locus, the locus id, and a set of transcripts or isoforms. Note that only the combination of all three identifiers is guaranteed to be unique.

5. Known issues

!!! Currently, Ananas will not run out-of-the-box on large data sets. !!!

We are working on a proper implementation, for now, you will have to manually pre-process the reads in two steps. Note: pairing information is retained ONLY for Illumina data. Run:

```
> ./SimplicityFilter
```

Available arguments:

```
-i<string> : input fasta file
-o<string> : output fasta file
-n<int> : maximum run (def=20)
```

On the output, run:

```
> ./BuildReadGroups
```

Available arguments:

```
-i<string> : input fasta file
-o<string> : output fasta file
-n<int> : maximum redundancy (def=5)
-m<int> : maximum mismatch (def=1)
```