Body-site specific assemblies of unassembled reads
HMP

Author: Mihai Pop

Version: 0.1
Effective Date: 10/28/2011

1 Abstract

This SOP describes the process used to identify HMP reads that were not assembled as part of the HMP Whole Metagenome Assembly pipeline (see hmpdacc.org/HMASM/) then assembling these using the same pipeline. For the HMP Whole Metagenome Assembly step, please refer to the relevant SOP at hmpdacc.org/HMASM/.

Note: This SOP is primarily relevant for historical purposes and to assist those who want to reproduce the results generated by the HMP. The SOP will need to be revised to take into account changes in technology and the availability of new software tools better suited for metagenomic assembly.

2 Introduction

The body-specific assembly pipeline identifies the reads that were not assembled by the SOAPdenovo-based whole metagenome assembly. Since SOAPdenovo does not report the placement of reads in the assembly, the placement is inferred through alignment of the reads against the assembled contigs. Since it is possible that pairing relationships between paired-end reads are broken by the alignment process, the pairing is re-computed at the end of the alignment process.

3 Requirements

3.1 Data requirements

- FASTA-formatted contigs constructed by SOAPdenovo. Pipeline assumes they are in the format produced by the whole metagenome assembly pipeline (a single directory that contains a file named <NAME>.contigs.fa containing all the contigs)

- FASTQ-formatted reads that were provided as input to SOAPdenovo. Pipeline assumes the file names follow the format: <NAME>.denovo_duplicates_marked.trimmed.[1|2|un].fastq where the [1|2|un] indicates whether the reads are forward (1), reverse (2), or unpaired (un).

3.2 Software requirements

- Perl (any recent version) with at least package Getopt::Long installed - www.perl.org

- Bowtie (any recent version) – bowtie-bio.sf.net

- get_singles.pl perl script – provided in the MetAMOS package https://github.com/treangen/metAMOS

3.3 Compute requirements

Linux machine with at least 4GB of RAM.

4 Procedure

Script invocation

```
./get_singles.pl -reads SRS051116 -assembly SRS051116_LANL/ > get_singles.log
```

Assumes read files are prefixed with SRS051116 and the assembly is located in the directory SRS051116_LANL

Log file can be redirected to /dev/null for most applications – it is only useful for debugging.

The outputs are written in the assembly directory and contain the following files:

```
SRS051116.denovo_duplicates_marked.trimmed.unplaced.1.fastq

SRS051116.denovo_duplicates_marked.trimmed.unplaced.2.fastq

SRS051116.denovo_duplicates_marked.trimmed.unplaced.singleton.fastq
```

These contain the unplaced reads that are mated (.1 & .2) and the unpaired unplaced reads (.singleton).

These files can be directly used as input to the whole metagenome assembly pipeline.

Running whole metagenome assembly pipeline

The exact whole metagenome assembly pipeline is described in a different SOP, available at **hmpdacc.org/HMASM/**. The main caveat when running a body-site specific assembly is that each library may have a different library size. These need to be directly encoded as separate sections of the `config.txt` configuration file.

5 Implementation

Code is written in Perl.

6 Discussion

This SOP is specific to the HMP data products, as used within the HMP, and is also based on possibly obsolete software.

The alignment process only looks at the first 25bp of each read. Thus, it is possible that more reads are considered as 'assembled' than actually used in the SOAPdenovo assembly. This restriction is necessary both for reducing the runtime of the program, and for ensuring that alignments can be found given that SOAPdenovo creates contigs much smaller than the read length (smallest contig is k-mer size + 1bp).

7 Revision History

Version	Author/Reviewer	Date	Change Made
0.01	Mihai Pop	10/28/2011	Establish SOP