SOLiD™ System Application Documentation: *solid_denovo_preprocessor.pl*

1.0	10/22/08	Vrunda Sheth	First draft.

.

Program

Program Name

solid_denovo_preprocessor.pl

Program Version

solid_denovo_preprocessor.pl v1

Description

Application

De novo assembly of SOLiD™ colorspace reads

Description

The solid_denovo_preprocessor.pl script is a part of the Velvet de novo assembly pipeline. It preprocesses SOLiD™ colorspace reads prior to submission to the colorspace-aware version of Velvet.

Release Notes

First release

Algorithm/Script Description

This script preprocesses SOLiD™ colorspace reads prior to submission to the colorspace aware version of Velvet .It trims the 1st base (primer) and the 1st color call of every read. For a fragment run it translates the colors 0123 to the 'pseudobases' ACGT. For a mate pair run it groups the mate pairs together discarding the reads without a mate, reverses the F3 reads to achieve the correct orientation for Velvet, and finally translates the colors to pseudobases.

Installation

Installation Instructions

Download the script and make it executable. You may need to change the first line of the script if your Perl is in a non-standard location (i.e., not /usr/bin/perl). You may need to install the Getopt::Long, IO::Handle, FileHandle, English, and File::Basename Perl modules from CPAN. Requires PERL version 5.8.5.

Usage

Usage Parameters - required

For a fragment run, at least one file should be specified:

--file <path to the input csfasta file>

Multiple fragment input files can be specified using additional –file arguments, e.g.,

--file <path to file1> --file <path to file2>

For a mate-pair run, both F3 and R3 files must be specified

--f3 <path to the F3 tag csfasta file>
--r3 <path to the R3 tag csfasta file>

Usage Parameters - optional

-r |--run fragment|mates

Specifies whether input files come from a mate-pair or a fragment run. Default is fragment.

-h|--help

Prints a help message.

-v|--version

Prints the program version.

Usage Parameters: is argument order important?

No.

Usage Example

Example1: fragment run

```
$ solid_denovo_preprocessor.pl --run fragment --file reads.csfasta.
```

Example2: mate-pair run

```
$ solid_denovo_preprocessor.pl --run mates --f3 reads_F3.csfasta
--r3 reads_R3.csfasta
```

List of programs/scripts included

This script belongs to the Velvet de novo assembly pipeline. Refer to Solid_denovo_HowTo for more information on how to do de novo assembly of SOLiD™ reads.

Path constraints

Output subdirectory will be created in the same directory in which the script was run.

System Input Files

One or more csfasta input files. If reads are from fragment runs, multiple input files may be specified. If reads are from a mate-pair run, input must consist of a single matched pair of F3 and R3 csfasta files.

Input File Versions supported

n/a

Additional Input Files

n/a

Input File Comments

All input files must be properly formatted csfasta. Blank lines at the start and end of the file are not permitted. Comments at the top of the file prefaced by '#' are allowed.

Example csfasta input file

```
# comments
>589_59_50_F3
T0000220302312101311121112
>589_59_92_F3
T1031320013001301110003332
>589_59_123_F3
T0300133101101030200320010
>589_59_200_F3
T2110011111003022003313101
>589_59_205_F3
T0110110010020131002320112
```

Output File(s)

Two output files are written to a subdirectory called 'output'.

1. colorspace_input.csfasta

   This file contains all of the input colorspace reads. For a fragment run with multiple files it concatenates all the files into a single file. For a mate pair run it orders the reads such that mate-pairs are adjacent to each other. This file is used as input to the post processor after running Velvet. For example,

   ```
   >589_59_200_F3
   T2110011111003022003313101
   >589_59_200_R3
   G0233012002200030122200323
   >589_59_92_F3
   T1031320013001301110003332
   >589_59_92_R3
   G2322213110111012003031031
   ```

2. doubleEncoded_input.de

This file is the double encoded file in which colors have been translated into 'pseudobases'. This is the input file for Velvet. For example,

```
>589_59_200_F3
CACTCTTAAGGATAACCCCCAACC
>589_59_200_R3
GTTACGAAGGAAATACGGGAATGT
>589_59_92_F3
GTTTAAACCCATCAATCAAGTCTA
>589_59_92_R3
TGGGCTCCACCCACGAATATCATC
```

Output File(s) Comments

The colorspace_input.csfasta is the csfasta file input to the Velvet postprocessor in the pipeline. The doubleEncoded_input.de is the input for Velvet.

Sample Input File(s)

- F3_reads.csfasta
- R3_read.csfasta

Sample Output File(s)

Sample output files will be in a subdirectory called output

- colorspace_input.csfasta
- doubleEncoded_input.de

Sample output files were generated using the following options:

```
$ solid_denovo_preprocessor.pl --run mates --f3 F3_reads.csfasta --r3 R3_read.csfasta
```

Supports AB kit or protocol or sample prep method

n/a

Other

Development language

Perl

Compiled for: (required for executables)

n/a

PBS is Required?

No

Comments

This script is a part of the Velvet de novo assembly pipeline for SOLiD reads.

Date (required)

10/06/08.