NGW2010 workshop
Genome de novo assembly
Excercises, session 5, lecture 1

Tristan Lefébure

TNL7@cornell.edu

April 14, 2010

1 Data sets

First of all, copy the data sets and script templates to your home directory:

```
cd $HOME
mkdir session5
cd session5
mkdir lecture1
cp-r /home/gfs08/jp86/ngw2010/session5/lecture1/*
    $HOME/session5/lecture1/
```

The data sets are the same as the one described during the lecture (*E. coli* K12). It is usually a good idea to check that there was no problem during the transfer of the files. You can do this using the md5sum tool (type `md5sum <file name>`). The md5sum outputs should be exactly like the following (if not something went wrong during the transfer):

```
-sh-3.1$ md5sum *.gz *.sff
4382ad1a8ef22c070a8017b62549ba2a  SRR001665_1.fastq.gz
7fa01c962421fdfb57f783cd5f1b0f78  SRR001665_2.fastq.gz
4a40c4b2074c1d434a1eb72b3f9089c5  EcoliRL.sff
561cef1fd8fc4b46d3e6892d279615b0  ecoPEhalfSet8kb.sff
```

2 Newbler assembly

Run Newbler from the command line, use the file `newbler_cli.sh` as a template. Don't forget to change the project ID in the file header.

Describe the assembly you get:

- Coverage?
- Number of contigs?

- Number of scaffolds?

Blast one of the small scaffolds on the NR database using the NCBI web interface (`http://blast.ncbi.nlm.nih.gov/Blast.cgi`). What is this? Any clue why it was not assembled in the big scaffold?

3 Velvet assembly

3.1 Prepare the data set

Decompress the data set, and shuffle the two fastq files into one. Use as a template the file `prepare_fastq.sh`.

3.2 Manual run of Velvet

Run two Velvet assemblies, one with a kmer length of 23, the other a kmer length of 31. This involves running `velveth` and `velvetg` first, then assessing the kmer coverage (use a spreasheet or R to analyse the `stats.txt` file), and finally rerunning `velvetg` with the corresponding options. Use the file `run_velvet.sh` as a template. Which assembly is the best? If you had to close this *E. coli* genome would you use the Newbler or the Velvet assembly?