INF-BIO9120 2012 Assembly: sources

NOTE all course material has been released through github at https://github.com/lexnederbragt/INF-BIO9120_fall2012_de_novo_assembly

Sequence data

Much of the sequence data used in this practical is available to download.

Illumina paired end
The MiSeq run was downloaded from http://www.illumina.com/systems/miseq/scientific_data.ilmn (E. coli strain MG1655 Read1.fastq and Read2.fastq).
To generate subsets of reads, seqtk (https://github.com/lh3/seqtk) was used as follows. Additionally, the sequence headers were updated with '/1' and '/2' to indicate the pairing (see the fastq entry on wikipedia)

150 nt, 50X coverage (2 x 773279 reads)

```
seqtk sample -s 11 \
MiSeq_Ecoli_MG1655_110721_PF_R1.fastq 773279 \
|awk '{if(/^@:114/){print $0"/1"}else{print $0}}' \
>MiSeq_Ecoli_MG1655_50x_R1.fastq

seqtk sample -s 11 \
MiSeq_Ecoli_MG1655_110721_PF_R2.fastq 773279 \
|awk '{if(/^@:114/){print $0"/2"}else{print $0}}' \
>MiSeq_Ecoli_MG1655_50x_R2.fastq
```

100 nt, 50X coverage (2 x 1159919 reads)

```
seqtk sample -s 11 \
MiSeq_Ecoli_MG1655_110721_R1.fastq 1159919 \
| cut -c 1-100 \
| awk '{if(/^@:114/){print $0"/1"}else{print $0}}' \
>MiSeq_50x_R1_100nt.fastq

seqtk sample -s 11 \
MiSeq_Ecoli_MG1655_110721_R2.fastq 1159919 \
```

```
| cut -c 1-100 \
| awk '{if(/^@:114/){print $0"/2"}else{print $0}}' \
>MiSeq_50x_R2_100nt.fastq
```

150 nt, 50X coverage (2 x 2319838 reads)

```
seqtk sample -s 11 \
MiSeq_Ecoli_MG1655_110721_R1.fastq 2319838 \
| cut -c 1-50 \
| awk '{if(/^@:114/){print $0"/1"}else{print $0}}' \
>MiSeq_50x_R1_50nt.fastq

seqtk sample -s 11 \
MiSeq_Ecoli_MG1655_110721_R2.fastq 2319838 \
| cut -c 1-50 \
| awk '{if(/^@:114/){print $0"/2"}else{print $0}}' \
>MiSeq_50x_R2_50nt.fastq
```

Illumina mate pairs

Data from Ribeiro et al 2012 (the Allpaths_LG paper describing the use of PacBio reads, http://genome.cshlp.org/content/early/2012/09/27/gr.141515.112) was downladed for E coli:
ftp://ftp.broadinstitute.org/pub/papers/assembly/Ribeiro2012/data/ecoli_data.tar.gz

To fix non-identical sequence names between fasta and quala

```
sed 's/sequence_//' jump_reads.Solexa-42866.A.quala \
     >jump_reads.Solexa-42866.A.qual
```

Convert to fastq using fastaqual2fastq.py

The original reads are in ← → orientation. However, Velvet expects → ←, so I reverse complemented both files

```
seqtk seq -r jump_reads.Solexa-42866.A.fastq \
     >jump_reads.Solexa-42866.A_RC.fastq

seqtk seq -r jump_reads.Solexa-42866.B.fastq \
     >jump_reads.Solexa-42866.B_RC.fastq
```

Prepared files for newbler:

```
cat jump_reads.Solexa-42866.A.fasta |awk '{if (!/>/){print $0}
else{print $1" template="substr($1,2)" dir=f library=42866"}}'
|revcomp_fasta.pl>jump_reads.Solexa-42866.A_RC_for_newbler.fas
ta

cat jump_reads.Solexa-42866.A.quala \
|sed 's/sequence_//' |awk '{if (!/>/){print $0}
else{print $1" template="substr($1,2)" dir=f library=42866"}}'
|revcomp_qual.pl>jump_reads.Solexa-42866.A_RC_for_newbler.qual

cat jump_reads.Solexa-42866.B.fasta |awk '{if (!/>/){print $0}
else{print $1" template="substr($1,2)" dir=r library=42866"}}'
|revcomp_fasta.pl>jump_reads.Solexa-42866.B_RC_for_newbler.fas
ta

cat jump_reads.Solexa-42866.B.quala \
|sed 's/sequence_//' |awk '{if (!/>/){print $0}
else{print $1" template="substr($1,2)" dir=r library=42866"}}'
|revcomp_qual.pl>jump_reads.Solexa-42866.B_RC_for_newbler.qual
```

454 Shotgun reads

The library and beads for this run were preapred by 454 Life Sciences in the US, and sequenced at the Norwegian Sequencing Centre as part of a test run after an instrument upgrade. The reads (two lanes of a plate divided into four lanes) are available upon request.

To make lower-coverage datasets, the sfffile command (part of newbler was used). An example (-pickb: randomly select a number of bases, 139m stands for 139 Mbp, 30X coverage):

```
sfffile -o HUFU8GI02_30X.sff -pickb 139m HUFU8GI02.sff
```

To prepare shorter, titanium length, reads:

```
sfffile -o HUFU8GI02+4_Titanium_length.sff -xlr HUFU8GI02.sff
HUFU8GI04.sff
sfffile -o HUFU8GI02+4_30X_Titanium_length.sff -pick 139m
HUFU8GI02+4_Titanium_length.sff
```

454 mate pair reads

These reads were a gift from 454 Life Sciences, and I am not sure if I can 'give them away'. If you are interested in working with these reads, contact me!

Subsampling (pickr: randomly select a number of reads):

```
sfffile -pickr 259000 -o FJ4Q8OU01_18X.sff FJ4Q8OU01.sff
```

Ion Torrent shotgun and mate pair reads

To get access to these, one has to obtain a (free) account at the Ion Community: http://ioncommunity.lifetechnologies.com/welcome

The reads belonging to this application note: 'De Novo Sequencing on the Ion Torrent PGM' http://ioncommunity.lifetechnologies.com/docs/DOC-2656 were downloaded, see http://ioncommunity.lifetechnologies.com/docs/DOC-2265:

`C11-127.sff.zip`	Raw shotgun reads in SFF format
`FRA-257.sff.zip`	Raw 3.5kb Long Mate-pair reads in SFF format
`C28-140.sff.zip`	Raw 8.9 Long Mate-pair reads in SFF format

Downsampling the shotgun reads was done as for the 454 reads (using `sfffile`)

The description on how to prepare the mate pair files for newbler is given on my blog:
http://flxlexblog.wordpress.com/2012/03/02/ion-torrent-mate-pairs-and-a-single-scaffold-for-e-coli-k12-substr-mg1655/

PacBio long reads

Following the instructions in http://www.smrtcommunity.com/Share/Datasets/E-coli-K12-De-Novo, raw long reads and CCS short reads were downloaded and error-corrected using PacBioToCA from the Celera assembler.

Software

- fastqc: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/
- cutadapt: http://code.google.com/p/cutadapt/
- newbler: you can ask a copy through a form at the 454 website (http://my454.com/contact-us/software-request.asp)
- velvet: http://www.ebi.ac.uk/~zerbino/velvet/
- celera: wgs-assembler.sourceforge.net
- Tablet assembly viewer http://bioinf.scri.ac.uk/tablet/
- MAUVE multiple genome alignment http://asap.ahabs.wisc.edu/software/mauve/
- mauve assembly metrics: http://code.google.com/p/ngopt/ NOTE: here is a trick to make working with mauve assembly metrics easier:

Make a script called mauveAssemblyMetrics.sh with this command (one single line):

```
java -cp
/site/infbio9120/bin/mauve_snapshot_2011-07-18/Mauve.jar
org.gel.mauve.assembly.ScoreAssembly $*
```

Make sure the script is in your $PATH
Also, make sure progressiveMauve, mauveAssemblyMetrics.pl and mauveAssemblyMetrics.R are in your $PATH

Now you can run, for each assembly:

```
mauveAssemblyMetrics.sh \
-reference /path/to/ref.gbk \
-assembly ../path/to/contigs.fa \
-outputDir asm_name \
-reorder asm_name
```

Scripts

- perlscripts and other scripts not mentioned below are available upon request
- interleave_pairs.py see https://github.com/lexnederbragt/denovo-assembly-tutorial
- pair_up_reads.py see https://github.com/lexnederbragt/denovo-assembly-tutorial

Contact me

lex.nederbragt@bio.uio.no
@lexnederbragt on twitter
flavors.me/flxlex
flxlexblog.wordpress.com and contig.wordpress.com