De novo assembly

The process of assembling a genome or any sequence without the use of an external reference.

SETUP ENVIRONMENT

a) Set up your environment

```
cd /vault/home/curso/userXX/data
```

b) have a look at the files and do the quality control with **fastqc**
(same as mapping course, you can skip this step if already done).

```
fastqc
```
Then from the menu, load the file `s_1_1_sequence.txt`

You can also ask for non-interactive behavior and view output (fastqc_report.html) with a browser
```
fastqc s_1_2_sequence.txt &
```

What do think of the quality of the data?
Do you need to trim the reads? if yes at which positions?

De novo ASSEMBLY

c) Assembly steps: select one of the program (either ABySS or Velvet or SOAPdenovo) and try several k values to run the assembly.
As it will take 1-2 hours to finish, please open a new terminal window and jump to section REMAPPING below.

ABYSS (60-90min) (manual):

```
mkdir abyss
cd abyss
for k in 33 31 29 27 25 23; do abyss-pe k=$k n=5 name=aby_$k lib='pe35'
pe35='../s_1_1_sequence.txt.filtered ../s_1_2_sequence.txt.filtered'; done
```

Velvet (40-60min) (manual.pdf):

#merge paired sequences in a single file (mandatory for velvet)
```
shuffleSequences_fastq.pl s_1_1_sequence.txt.filtered s_1_2_sequence.txt.filtered merged.fq
mkdir velvet
cd velvet
```
#run velvet
```
for k in 33 31 29 27 25 23; do velveth vel_$k $k -fastq -shortPaired ../merged.fq && velvetg
vel_$k -ins_length 600 -ins_length_sd 100 -unused_reads yes -min_contig_lgth 100 -cov_cutoff
auto -exp_cov auto; done
```

SOAPdenovo (40-60min) (manual):

```
mkdir soap
cd soap
cp /vault/course2011/soap.config .
for k in 33 31 29 27 25 23; do SOAPdenovo-63mer all -s soap.config -K $k -R -L 100 -o soap_$k -p
```

```
8 && GapCloser -b soap.config -a soap_$k.scafSeq -p 31 -t 8 -o soap_$k.closed; done
```

Have a look at the files generated by the assembler.
What are their sizes? Where are the assembled sequences?
What is the difference between a scaffold and a contig?

d) compare assemblies by their metrics

```
fac2.pl <any_contigs_file>
```

```
e.g.,
for abyss:
fac2.pl abyss/aby_*-contigs.fa
for velvet
fac2.pl velvet/vel_*/contigs.fa
for soapdenovo
fac2.pl soap/soap_*.closed
```

Which K value gives the best assembly? why?
Compare with your neighbors

e) compare the largest contigs to the reference (CP001844.gbk) with **MAUVE**:

sort the contigs by size and keep only those > 1000 bp
```
sort_contigs.pl -b -m 1000 -p -z <your_best_contigs> mybestsorted.fa
```

Start Mauve

```
Mauve
```

and import the reference and your contigs (File->Align with progressiveMauve, then Add Sequence...), then start the alignment by clicking "Align..." button and enter a name.
Once finished (takes a few minutes), select View->Color Scheme->Backbone color and unselect View->Style->LCB outlines and View->Style->LCB connecting lines.

or

use **nucmer** to align contigs onto a reference:
#calculate the alignments
```
nucmer CP001844.fa mybestsorted.fa -p refVSmybest
```
show matching contigs positions
```
show-coords -lro -L 2000 refVSmybest.delta
```
output the pseudo genome and the list of unused contigs
```
show-tiling -a -c -g -1 -l 1000 -v 80.0 -p mybest.pseudo -u mybest.unused refVSmybest.delta
```

Are these good contigs? why?
Do you see extensions in the SCCmec region (37kbp-90kbp)?

REMAPPING

f) go to your Bowtie assembly (you did it yesterday, if not please refer to the practical page "Assembly by mapping").

```
cd /vault/home/curso/userXX/bowtie
```

create a directory for assembling the unmapped reads

```
mkdir asm_unmapped
cd asm_unmapped
```

g) assemble the unmapped reads with velvet to find missing sequences

```
for k in 33 31 29 27 25 23; do abyss-pe k=$k n=5 name=unref_$k lib="unse35" unse35="../ref_un";
```

```
done
```

h) check the largest contigs by running BLASTn vs NR at NCBI

Note: To extract the largest contig use:

```
fac2.pl unref_XX-contigs.fa
```

keep the `max` size and type:

```
sort_contigs.pl -m <max_size> unref_XX-contigs.fa largestcontig.fa
```

What do you find? a plasmid? a phage? a genomic region?