Assemblies with different datasets

With this practical, we will all perform one (or more) different assemblies, testing several aspects, and compare the results.

Questions you can ask with these datasets

- what is the effect of coverage on assembly quality (both velvet and newbler)?
- what is the effect of read length on assembly quality (both velvet and newbler)?
- how well does IonTorrent data assemble with newbler, both shotgun only, and shotgun + mate pairs?
- how well does newbler work with Illumina data (paired end and mate pair)?
- can 454 and Illumina data be combined (using newbler) and what is the effect?

Datasets

I have provided a set of files to try different assemblies with:

- Please read *all of the points below* before you start your assembly
- To choose your input data combination, use the file called **'INF_BIO9120 fall 2012 assembly input data'**
- Please record your choice in this spreadsheet: **`bit.ly/INFBIO2`**
- Make sure no other person has chosen the same combination (but, see below)
- Please record the assembly results also in this spreadsheet: **`bit.ly/INFBIO2`**
- The following data types are available:
    - Illumina data as we used yesterday
    - 454 data as we used today
    - lower coverage versions of the 454 data (10X, 20X)
    - higher coverage versions of the 454 data
    - shorter versions of the MiSeq paired end data (50 nt and 100 nt)
    - IonTorrent data, both shotgun (sff files) and mate pair (fasta + qual files, prepared for newbler)
- When using velvet, you may want to estimate the best k-mer size first;

remember that it is no use to choose a kmer value longer than your read length
- The MiSeq datasets have *not been trimmed*, you need to do this yourself!
- It is no use trying 454 or Ion Torrent data with velvet
- It is OK to try to use illumina and Ion Torrent data with newbler, note however that for mate pair reads, special input files have been prepared (have a look inside the files!)
- newbler is your choice for hybrid assemblies (combining different sequencing technologies)
- for using newbler with Illumina and IonTorrent data, see the document '**Assemblies with newbler and Illumina or Ion Torrent data**'
- if you use more than 50X coverage with newbler, it is advisable to let the program now through the `-e` (expected coverage) flag. For example, if your *total* coverage is 80X, add '`-e 80`' to the command line
- If the dataset you wanted to try with newbler has been already been chosen by someone else, consider running it as well, but adding the `-scaffold` flag for comparison
- **Hint:** use 'nohup' before the command you write to start jobs which will continue to run when you close your terminal.
- Once you have chosen an assembly, **ask me to have a look at your plan!**

Once the assemblies are done, use the assemblathon_stats script and mauve contig mover to investigate your resulting sequences. Report your findings on this spreadsheet: **`bit.ly/INFBIO2`**

Precomputed assemblies

I have provided two assemblies with Celera here:

`/data/assembly/precomputed/celera`

`30X_18X_ovl` with 454 shotgun and mate pair data

`PBcR_18X` with error-corrected PacBio long reads and the 454 mate pair reads.

For practical reasons, please do not redo the Celera assemblies yourself at this time, these run for very long and may use up all the available cpus. You can consider trying this program in between other parts of this course, see the document **'INF-BIO9120 2012 Assembly using celera'**