Assembly using celera

The Celera assembler uses the overlap-layout-consensus paradigm, and is the only program from the Sanger era of sequencing that continuously keeps adapting to new sequencing technologies. It is also one of the very few programs that can use the new, long PacBio reads. Celera is more complicated than the other programs used for this course, with many parameters to choose from. It also gives a huge number of output files and folders, besides the actual scaffolds and contigs. It has a fairly long run time, which is why it is not used for the main part of this course. Feel free to try it out, but make sure you have at least **22 GB of free disk space** if you want to run the 454 shotgun + mate pair assembly. You can check the available space by typing `df -h` and look for the `/home` entry.

Here you will be given two simple examples:

- with 454 shotgun and mate pair data
- with error-corrected PacBio long reads and the 454 mate pair reads.

For more information about the program, see **wgs-assembler.sourceforge.net**

Where is what

Celera requires input files in its own 'frg' (fragment) format. Also, PacBio reads need to be error-corrected with high-quality short reads, which was done for you, as this takes a lot of computational time and resources.

All files are located in the folder `/data/assembly/frg`

`HUFU8GI02_30X.fr.frg`	454 shotgun reads
`8kb_FJ4Q8OU01_18X.frg`	454 mate pair reads
`PBcR_10kb_CLR_with_PacBio_2kb_CCS.frg`	PacBio reads*

* error-corrected using PacBio CCS (short circular consensus) reads and the PacBioToCA pipeline of the Celera package

General setup

First, setup your environment, i.e. add the path to the celera program to your $PATH:

```
export PATH=/site/infbio9120/bin/celera/wgs-7.0/Linux-amd64/bin:$PATH
```

Assemblies are generally started by running this command:

```
runCA -d directory -p prefix -s specfile input-files
```

-d: name of the outputfolder
-p: name prepended to all output files
-s: specification files with parameters

Celera assembly with 454 shotgun and mate pair reads

Make a spec file with the following information:

```
overlapper    = ovl
unitigger     = bog
utgErrorRate  = 0.03
```

Run as follows:

```
runCA -d asm_name -p asm_name -s asm_name.spec \
     /data/assembly/frg/HUFU8GI02_30X.fr.frg \
     /data/assembly/frg/8kb_FJ4Q8OU01_18X.frg
```

Celera assembly with long, error-corrected PacBio reads and 454 mate pair reads

Make a spec file with the following information:

```
cnsErrorRate = 0.10
ovlErrorRate = 0.10

overlapper = ovl
unitigger = bogart
utgBubblePopping = 1

merSize = 14
merCompression = 1

doToggle=0
toggleNumInstances = 0
toggleUnitigLength = 2000
```

```
doOverlapBasedTrimming = 1
doExtendClearRanges = 2
```

Run as follows:

```
runCA -d asm_name -p asm_name -s asm_name.spec \
  /data/assembly/frg/PBcR_10kb_CLR_with_PacBio_2kb_CCS.fr
g \
  /data/assembly/frg/8kb_FJ4Q8OU01_18X.frg
```