Canu

Canu is single molecule sequence assembler.

An example invocation could look like

```
module load Canu/2.1
canu \
  -p asm \
  -d /path/to/output \
  genomeSize=1.6g \
  -pacbio-hifi /path/to/input \
  -gridOptions="-A 'project' -l walltime=72:00:00 -r y" \
  -preExec="module load Canu/2.1" \
  -gridEngineResourceOption="-l select=1:ncpus=THREADS:mem=MEMORY:arch=skylake" \
  -shell="/bin/bash"
```

If canu is run in grid mode it should be started on a login node. If canu is started as a job it will detect the jobid and only use on a single node.

The following arguments are necessary for canu to work in grid mode for us.

```
-gridOptions="-A 'projectname' -l walltime=72:00:00 -r y"
```

Projectname and walltime must be set for every job. The walltime is per subjob, e.g. if canu runs 100 subjobs in a jobarray each subjob would have 72 hours to work with. If you run multiple similar assemblies this argument should be adjusted ones a proper value is known - too high and the jobs will hang in the queue for a significant amount of time, too low and the jobs won't finish. The `-r y` flag allows jobs to restart which is necessary for array jobs.

The `-gridEngineResourceOptions` argument is a template into which proper THREADS and MEMORY values are set by canu. For some steps like overlapInCore and mhap canu can estimate proper values on its own.
If similar jobs are run frequently it might be worthwhile to specify more accurate cpu, memory, and walltime requirements for each pass https://canu.readthedocs.io/en/latest/parameter-reference.html#grid-options .

The `-preExec="module load Canu/2.1"` command is run at the beginning of each job in the canu pipeline and ensures all dependencies are loaded. Running the module command in the generated canu scripts requires `-shell="/bin/bash"`.