Before you start:

The computer has a special environment variable $PATH to specify a set of directories where executable programs are located. On the command line type

```
echo $PATH
```

to see which directories are currently in your PATH.

Some software tools for this exercise (e.g. velveth) are in /usr/bin which is already on your PATH. Other tools are stored in directories that are not on your path yet:

```
/local/prog/augustus/bin
/local/prog/braker
/local/prog/wgs-8.1/Linux-amd64/bin
```

To add these directories to the environment variable $PATH, you need to edit a file in your home directory, named `.profile` (file name starts with a dot)

In this file, add the following lines:

```
PATH=/local/prog/wgs-8.1/Linux-amd64/bin:$PATH
export PATH
```

You need one of these PATH=xxx/xxx:$PATH lines for each software tool that you want to add.

Save the file, and close it.

After changing the file, type on the command line

```
source ~/.profile
```

Verify the changes to your path by typing

```
echo $PATH
```

Executable programs like “runCA” are now available. To check this, type

```
which runCA
```

If you install other tools (in progs_nobackup in your home): same procedure.

The data for this exercise is stored in a tar archive:
http://www.bioinformatics.nl/courses/BIF-30806/docs/yeast.tgz

Use wget to download it to a directory on the server.

Type `tar –xvzf yeast.tgz` to unpack the tar archive.

Assembly, annotation, and mapping

Chromosome III of a specific yeast strain, which is often used in industrial biotechnology, CEN.PK 113-7D, was sequenced using an Illumina HiSeq. Paired-end sequencing of 300bp fragments resulted in read pairs of 101 bases each. The pairs are stored on corresponding lines in two files, “cenpk-chr3_1.fastq” and “cenpk-chr3_2.fastq”. For this strain also some RNA was sequenced (CENPK_RNA_1.fastq and CENPK_RNA_2.fastq)

This particular strain, CEN.PK 113-7D, was used in an experiment in which cultures were evolved in the lab to find alternative transporters of lactic acid. To this end, the only known gene coding for a lactate transporter, Jen1, was knocked out and for 100 generations the cells best surviving on a medium containing lactate as a carbon source were selected. The resulting yeast strain *IMW004* was able to grow well on lactate. This strain was also sequenced, resulting in the files imw004-chr3_1.fastq and imw004-chr3_1.fastq.

Yeast is of course a well-studied model organism, so for *Saccharomyces cerevisiae* strain S288C, the reference genome of chromosome III (chr3.fasta) and annotation (chr3.gff) are available.

This assignment is **an invitation to explore and analyze** this yeast data set. You may design and implement a pipeline in Python to perform genome assembly and/or annotation and/or read mapping. Alternatively you can read some review papers.

1. **Assembly**
   Assemble the genomic reads of CEN.PK 113-7D. Calculate the N50 size and index for this assembly. Try to improve the assembly, in terms of N50, by pre-processing the reads or changing parameters of the assembly tools. For example, what is the effect of the k-mer size?

Assembly tools
velvet
wgs-assembler
soapdenovo
abyss

2. **Annotation**
   Annotate one of your own assemblies or the reference genome of chr3. What is the output format that augustus produces? Try to visualize this data using the IGV. Can you find some genes with introns? How similar is your annotation to the reference annotation?

Annotation tools
augustus
braker1
tRNAscan

3. **Mapping**
   Map the genomic reads of CEN.PK on your assembly or on the reference genome. Study the resulting SAM/BAM file. Do you understand the format? Try to visualize the data using IGV. Play around with samtools to generate some mapping statistics. What is the effect of the tool parameters? Try to map the RNA reads to the assembly, using tophat. Visualize these data also. What's the difference with the mapped genomic reads? How good is the correspondence between the annotation and the mapped RNA reads?

Mapping tools
bowtie2
bwa
tophat

4. **Reading (see Week 3&4 Literature on BlackBoard)**
   - Review NGS platforms
   - Review assembly (GAGE and Assemblathon 2)
   - Review annotation
   - Review mapping