More datasets

This session will list and introduce a range of **whole genome shotgun** experiments. They all are *Escherichia coli* isolates (different strains, unfortunately), but sequenced with different platforms:

1. Roche 454 (and 454 Junior)
2. Thermo IonTorrent (Proton)
3. Illumina (MiSeq)
4. Oxford Nanopore (MinION)
5. Pacific Biosciences (PacBio)

Note that these datasets have been downloaded from a public repository called **Short Reads Archive** hosted by the NCBI. It's a useful source of published and publicly available NGS datasets, that can be very useful to test pipelines or add “controls” to your analyses.

Where are these datasets

The general path to explore these datasets is `/bsb/denovo/datasets/`.

What's inside this directory?

Let's simply list the content of the directory with the datasets:

```
ls -l /bsb/denovo/datasets/
```

The output is something like:

```
total 28
drwxrwsr-x 2 ubuntu ubuntu 4096 Nov 21 17:45 454
drwxrwsr-x 2 ubuntu ubuntu 4096 Nov 21 17:48 454JR
drwxrwsr-x 2 ubuntu ubuntu 4096 Nov 21 17:45 illumina
drwxrwsr-x 2 ubuntu ubuntu 4096 Nov 21 17:49 ionproton
drwxrwsr-x 2 ubuntu ubuntu 4096 Nov 21 17:45 mixed
drwxrwsr-x 2 ubuntu ubuntu 4096 Nov 21 17:51 nanopore
drwxrwsr-x 2 ubuntu ubuntu 4096 Nov 21 17:51 pacbio
```

Basically a set of directories reminding us which platform generated the dataset. If we wanted to list all files with `.fastq` as extension contained in the directory *and its subdirectories* we could simply use the **find** command:

```
find /bsb/denovo/datasets/ -name "*.fastq"
```

How many sequences in each dataset?

Last week we introduced the SeqKit package to analyse and manipulate FASTA and FASTQ files. The

`seqkit stats` program quickly counts the reads, giving also the total amount of bases and maximum, average and minimum read length.

Let's try analysing the output of the good old 454 run:

```
seqkit stats /bsb/denovo/datasets/454/SRP001673.fastq
```

We totally have about 95Mbp, that for an *E. coli* genome means we produced a 20X coverage shotgun. Not that bad!

How do reads look like?

Different dataset vary. A simple way to have a look is using the `less` command. Remember that when using `less` you can interact with keystrokes (arrows, page up/down, and finally q to exit!). Example:

```
less -S /bsb/denovo/datasets/454/SRP001673.fastq
```

We can use the `-S` parameter to avoid word wrap, and keep the sequences in one line (use left/right arrows to scroll).

Where does this data comes from?

Most datasets have been downloaded from the NCBI Short Reads Archive, and they keep their Accession ID as filename. This means that you can search the SRA for the Accession, that in most cases is a single sequencing from a project and in the 454 case is the accession of an entire project.