1. Movement

1. Explore your file system!
2. Whats in your home folder?
3. How many users are there on your system?
4. Try an absolute path to get to home!

2. File manipulation I

1. Create a new folder at home
2. Create an empty file in it
3. List the file
4. Rename the file
5. Move the file one level up
6. Remove the file. Test if it's really gone.
7. Remove the folder. Test if it's really gone.
8. Go to root
9. Try to create a directory in root

3. File manipulation II

1. Create a directory called python_course.
2. Move the course files from your Downloads folder to the course folder
3. Unpack them with 'tar xzvf [files]'
4. What files are in there?

4. Investigate files I

1. How big are your course files?
2. And how long?
3. Create a new file
4. How big and long is it?

5. Investigate files II

1. Open a new file in nano
2. Write something into it
3. Save it as test1.txt
4. Do it again for a 2nd file called test2.txt
5. Cat both into a new file called test3.txt
6. Open it with less
7. Remove all 3 files in a single command (use a wildcard!).

6. Real data

In all following tasks we will try to analyse the file 'quasar.tsv' in your course files. It's a list of mutations and their effects in several colorectal cancer samples.

Each line corresponds to one mutation. The columns stand for:

sample-chrom-pos-ref-alt-altreads-refreads-quality-effect-effectimpact-class-gene-type-dbsnp

1. Understand the contents
2. Try head and tail
3. How many variants are there in total?.

7. Extract and filter

1. Is there a variant in the gene NPM1?
2. How many variants in TP53?
3. Any insertions in TP53? Don't look manually, find the proper pipe!
4. How many mutations from "C" to "T"?

8. Extract & count unique entries

1. Which sample has the most variants?
2. Which gene has the most variants of high impact?
3. How many different effect categories are there?.
4. Which type of snp is the most common?

9. Manipulate and filter

Assume we want to unite the 2 weak effect categories so that LOW and MODIFIER are both renamed to 'WEAK'.

1. Use sed to make that change and save the results in a new file.

Let's now try to investigate the following hypothesis using only commandline tools:

"If many of our C>T snps are false-positives (due to deamination), we would expect to find more of them in low-quality variants. For comparison we'll use T>C variants which are not affected by deamination."

1. Use the 'length' function of awk to select only snps.
2. Set a cutoff for low qual. snps (below 20) and high qual (above 80). How many are there each?
3. Count C>T and T>C in both sets (2 commands!) and compare them (manually).
4. What's the conclusion?