**Unix test**

Your name:

write your answer and/or what you did

1. Go to Taito and make directory MCOMPOP to your wrk-directory
2. get fastq-files for 10 samples from <http://www.nature.com/ncomms/2015/150930/ncomms9452/full/ncomms9452.html> and unzip them. The links can be found from “**Accession codes”** at the end of the paper. You need total of 20 fastq-files, 10 forward and 10 matching reverse.
3. make a tab delimited file in a text editor/excel with

sample\_name year R1 R2

ERXXX 2012 16S-AMP-1089\_S1\_L001\_R2\_001.fastq 16S-AMP-1089\_S1\_L001\_R2\_001.fastq

ERXYY 2012 16S-AMP-1089\_S1\_L001\_R2\_001.fastq 16S-AMP-1089\_S1\_L001\_R2\_001.fastq

1. calculate how many sequence reads there are in each multifastq-files
2. make copy of the one fastq-files and from the copy take 1000 first and 1000 last sequences to a new file called “subset.fastq”. Count that you indeed have 2000 sequences.
3. Go to Taito and make directory FASTA to your wrk-directory and get the multifasta file from the following location ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF\_000018885.1\_ASM1888v1/[GCF\_000018885.1\_ASM1888v1\_protein.faa.gz](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000018885.1_ASM1888v1/GCF_000018885.1_ASM1888v1_protein.faa.gz)
4. Extract archive and count how many FASTA headers the file contains
5. Substitute in the FASTA headers the underscore (“\_”) with the hyphen (“-“)
6. Delete the directory FASTA in your wrk directory