# Assigment – RNA-mapping

Today we will perform mRNA-analysis from start to finish, like in the paper. The main difference is that we will be running the entire analysis in R with a small subset of the data, but the concept is the same.

1. Download the zip files, unzip it and verify that it contains
   1. Fastq-files
   2. AlignRNA.R
   3. NC\_000913.gff3
   4. NC\_000913\_WG.fasta
2. In Rstudio, create a project in the folder you just unzipped and open the ‘alignRNA.R’ file
3. The ‘AlignRNA.R’ file contains all 5 steps needed to turn transcriptome-files into feature table as described in the lecture. There are, however, some missing values here and there (denoted by XXX), and you will have to fill them in as you go.
4. First, let’s setup the files (1).
   1. You will need to figure out what the path to the fastq-files is and replace the XXXs.
   2. Make sure you have all 12 samples for each read.
   3. Also make sure you understand why there are 2 files for each sample
5. Next, we make the index (2).
   1. You will need to work out from what file you make the index from.
   2. Why don’t we use a gapped index? Have a look in the help-file for ‘buildindex’.
6. Then we create the mapping file (3).
   1. What is the content of the resulting saf-object? Use the function head() to have a look at the first few lines.
   2. What exactly is the difference between the mapping file and the index we just made?
7. Finally, for the alignment (4)
   1. Add in the correct values for readfile1 and readfile2.
   2. This might take a while on some machines, but no more than a couple of minutes
8. Lastly, we will map our alignments to the genes, giving us a feature table (5)
   1. First, make sure you now have 12 bam-files
   2. The featureCounts-function returns a list of a bunch of things, but we are mostly interested in the mapping statistics (which we will keep in ‘stats’) and the features (which we keep in ‘features’)
   3. Have a look in the stats – which sample had the most assigned sequences? Which one had the fewest? Remember that we started with 10,000 read pairs for each sample.
   4. What are the main reasons for the Unassigned reads?
   5. Which gene has the most mapped reads in each sample? Is it the same and what do(es) they/it do?
   6. Are mapped reads directly interpretable as gene expression? Trick question, so what could be the bias here?