Unix/Linux Tutorial for Beginners
Session V – Pairwise exercises

1. Create the folder *'pairEx'* in the directory *~/myLinuxProject/results/courseDay_02/* and in the directory *~/myLinuxProject/myData/raw/*. Use only one command to create both folders at the same time.

   Answer: the directory 'pairEx' was created successfully if you can change to these directories and the pwd command returns:

   ```
   /home/user/myLinuxProject/results/courseDay_02/pairEx
   /home/user/myLinuxProject/myData/raw/pairEx
   ```

   Please replace 'user' with your user name.

   Used command(s):

2. Copy the files *anno.gff, myCDS.fa, nameList.txt,* and *animalSounds.tab* from the directory *~/data/pairEx* to the folder *~/myLinuxProject/myData/raw/pairEx*.

   Answer: the data was successfully copied to your folder, if the listing of your directory content ls returns:

   ```
   $ ls ~/myLinuxProject/myData/raw/pairEx
   animalSounds.tab  anno.gff  myCDS.fa  nameList.txt
   ```

   Used command(s):

3. The columns in the file *anno.gff* are separated by '::' instead of tabs.

   ```
   C10002475::GeneWise::mRNA::3::104::56.14::-::.::ID=Pad_R000001;Source=ENSTGUT00000006161;
   C10002475::GeneWise::CDS::3::104::.::-::0::Parent=Pad_R000001;
   C10002475::GeneWise::mRNA::16::291::65.91::+::.::ID=Pad_R000002;Source=ENSGALT00000035625;
   C10002475::GeneWise::CDS::16::136::.::+::0::Parent=Pad_R000002;
   C10002475::GeneWise::CDS::239::291::.::+::2::Parent=Pad_R000002;
   ...
   ```

   Replace all occurances of the string '::' with a tab and write the file *anno_corrected.gff* in the results folder, which was generated in step 1. Use the command man <command> to get help and see which options are available for the command of interest.

   Answer: The new created file should be located in the folder *~/myLinuxProject/results/courseDay_02/pa* and should look as follow:

```
$ ls ~/myLinuxProject/results/courseDay_02/pairEx/
anno_corrected.gff

$ less anno_corrected.gff
C10002475  GeneWise  mRNA  3    104  56.14  -  .  ID=Pad_R000001;Source=ENSTGUT00000006161;
C10002475  GeneWise  CDS   3    104  .      -  0  Parent=Pad_R000001;
C10002475  GeneWise  mRNA  16   291  65.91  +  .  ID=Pad_R000002;Source=ENSGALT00000035625;
C10002475  GeneWise  CDS   16   136  .      +  0  Parent=Pad_R000002;
C10002475  GeneWise  CDS   239  291  .      +  2  Parent=Pad_R000002;
```

Used command(s):

4. Extract all duplicated and unique identifier from the file *nameList.txt* and save the results in the files *duplicated_names.txt* and *uniq_names.txt*. How many unique and how many duplicated identifier have you found?

   Answer: There are 3 duplicated names (Anna, Lina, and Tyler) and 30 unique ones. The result files should be located in the folder *~/myLinuxProject/results/courseDay_02/pair*

   Used command(s):

5. The tab-separated file *animalSounds.tab* list animals and the sounds they make. Extract the second column and determine which sound is the most common among the listed animals and how often it occurs.

   Answer: The most common sound in the animal list is *'grunt'* and it occurs 3 times.

   Used command(s):

6. Count how many sequences are stored in the file *myCDS.fa*. In a next step reverse complement the sequences (not the sequence identifier) and convert at the same time the lowercase letters to uppercase letters. The complements of the nucleotides A, C, G, T are T, G, C, A. Write the output in your result folder in a file named *myCDS_revCompl.fa*. Does the file fulfill the standard FASTA file format?

   Answer: There are 10 sequences stored in the fasta file *myCDS.fa*. The reverse complemented and to uppercase letters converted file looks as follow:

```
$ less ~/myLinuxProject/results/courseDay_02/pairEx/myCDS_revCompl.fa
AACAAGTACCTTGTCAGGCAGGTCCTCTAACCACAGATGTAGGCCTGCCTGTTGAGAGGCACCAAAAGCTGCAAGAGCATGCACCAA
GTTATTACATGCCTTGGGGGCAAAATAAAACGTGCAAGCACTGAAACTTAGACGAGCGAAGTCCCAAATCTCTCTCAATAAAATTCC
TTCTGGTGCCCTGTCATAAACTGAACTTTGGAGCGCTTTTATTAGTGTCTGGCAATTAGACTCGACCTCCACATTTTTCAT
GATTGATACTGCAAGACTACTGATAAATCTGCCATAACTAAGCTTCAATTCCTGCCAATTTCCAACATCAACCAATCCTTCCTCATG
TCCATTGGAAGTGCGTGCCACAAGGGATTCGAACTTGCGGTCATTGCTGGTGAGAATACAGTAGGCCCTCAGCAGATCACCCCGTCC
TAAGTTTGGTATCTCTTGAAGTGCCGCAAAGATTTCCTTCGAAGATGCATTCGCTGGTTCCTCTCGAGGTTCATTTGGCTTCACAGA
TATCTGAAATGAATTATTGATGTTCCCAGACATGCAAAAGAGTGCATCATCTGTTCTAAACCTCTTTCTCCTCCTGGCCTGTAGAGT
GGGTTGCGTTGATGTCGGCGTTCCAACCTGGTTTTGACTGCTAGGTTCAGATGCTGAAGGTGAATTTTGCATAAGCTCCTGGACATC
TTCAGCACCCTTCAATGCCATTTCTTCGGCATTCTCCAT
....
```

Used command(s):

Exercises are in part derived by material from ©Software Carpentry (`http://software-carpentry.org`, license: CC BY 4.0) that was adapted from me for this course. Another part is from a BILS course given by Martin Dahlö and used here by his kind agreement. Remaining exercises by M. Martis.