Unix/Linux Tutorial for Beginners

Session I – Unix & File Formats

Mandatory exercises

1. Which of the following statements are true?
   a) The term Unix is used for a family of computer operating systems that derive from the original AT&T Unix.
   b) Unix/Linux can run several programs simultaneously, although only one user can be connected at a time.
   c) Nowadays the command line interface is still in use as there are tasks that simply can't be done with a graphical user interface.
   d) Unix/Linux can run several programs simultaneously, and manage several users at the same time.

2. A graphical user interface
   a) can't be used simultaneously with a command line interface.
   b) provides windows, icons and menus, which the user can use to send commands to the computer.

3. In Unix/Linux CLI is used as acronym for
   a) Children's Learning Institute
   b) Command Line Interface
   c) Critical Limb Ischemia
   d) Command Line Interpreter

4. In a fasta file
   a) the end of a sequence is indicated by a '!'
   b) the sign '*' indicates conserved residues
   c) description lines are marked with a '>', while the sequences are written in IUPAC single-letter code with no space between letters
   d) both, sequences and quality scores, are contained.

5. The characteristic difference between a FASTA and a FASTQ file is:
   a) there is no difference
   b) a FASTQ file extends the FASTA format by including 2 addtional lines, namely one with the base quality scores and one containing the sequence description
   c) the third line of a FASTQ entry is empty
   d) the FASTQ description line starts with '@', while the FASTA description line starts with '>'

6. Which of the following statements are correct:
   a) a valid SAM/BAM file does not need a header section
   b) a SAM file stores the alignment results between raw reads and a reference genome
   c) a sam-file and a bam-file differ in their content
   d) the alignment section contain 12 tab-delimited fields.

7. The mapping quality score in a SAM/BAM file reflects
   a) the probability that the read is aligned correctly
   b) how similar two reads are
   c) the sequencing quality of a read
   d) the probability that the read is aligned incorrectly.

8. Which of the following statements is false:
   a) the header of a VCF file contain the alignment between reads and reference genome
   b) VCF stands for "variant call format"
   c) The allele depth is displayed by the shortcut 'FA'
   d) VCF is a text file format, which describes SNP, indel, and structural variation calls.

Optional exercises

1. Consider the following FASTA-format file:

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
>NM_026810.2 Mus musculus mutL homolog 1 (Mlh1)
AGGACTGCAGTCGGCCGAAGCTGAAGGAAGAACTTGAGCGTGAGGAGCTCGAGTGATTGGCTGACTGGGA
ACTCGGGCGCCAATATGGCGTTTGTAGCAGGAGTTATTCGGCGTCTGGACGAGACGGTAGTGAACCGCAT
AGCGGCGGGGGAAGTCATTCAGCGGCCGGCCAATGCTATCAAAGAGATGATAGAAAACTGTTTAGATGCA
AAATCTACAAATATTCAAGTGGTTGTTAAGGAAGGTGGCCTGAAGCTAATTCAGATCCAAGACAATGGCA
CTGGAATCAGGAAGGAAGATCTGGATATTGTGTGTGAGAGGTTCACTACGAGTAAACTGCAGACTTTTGA
GGATTTAGCCAGTATTTCTACCTATGGCTTTCGTGGTGAGGCATTGGCAAGCATAAGCCATGTGGCCCAT
```

What does the segment above display?

a) two DNA sequences, named 'GENE X PROTEIN (OVALBUMIN-RELATED)' and '*Mus musculus* mutL homolog 1 (Mlh1)'

b) nine DNA sequences, of 70 nucleotides each. The first three are named '*P01013*' and the last six are named '*NM_026810.2*'

c) two RNA sequences, named '*P01013*' and '*NM_026810*', with lengths of 210 and 420 base pairs, respectively

d) one protein and one DNA sequence, named '*P01013*' and '*NM_026810.2*'

2. Which of the following sequences in FASTA-format are not valid?

a) gene_1:

```
>gene_1
AAGCCCCTCCTAAACCCTGTGCAGGCAACCAGGGCACCCTGATCACGGTGGAAGACCTT
```

b) gene_2:

```
> gene_2
CACTGCCCAATGCCACAACCGTGGACAACATTCGCTCCATCTTTGGAAATGCGGTTAGT
```

c) gene_3:

```
>gene_3
 GTGAAGAAGTGCATTTTCCTACTCTTCATCAACCACCGTCTGGTAGAATCAGCTGCCT
CACTGCCCAATGCCACAACCGTGGACAACATTCGCTCCATCTTTGGAAATGCGGTTAGT
```

d) gene_4:

```
>gene_4
AATCACAAGGAGGAAAGCTTTAAAAAATCCAAGTGAAGAGTACGGAAAAATTTTGGAAG
TATTCAATACACAATTCAGGCATTAGTTTCTCAGTTAAAAAACAAGGTGAGACAGTATC
```

3. Each sequence in a FASTQ file is represented by a four-line. Consider the following two segments of a fastq file. Which one is correct?

a) segment 1:

```
@HWI-ST1097:104:D13TNACXX:4:1101:1715:2142 1:N:0:CGATGT
GCGTTGGTGGCATAGTGGTGAGCATAGCTGCCTTCCAAGCAGTTATGGGAG
+
=<@BDDD=A;+2C9F<CB?;CGGA<<ACEE*1?C:D>DE=FC*0BAG?DB6
```

b) segment 2:

```
@HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 2:N:0
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
hhhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y
+HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 2:N:0
```

4. Consider the following alignment:

```
REF: AGCTAGCATCGTGTCGCCCGTCTAGCATACGCATGATCGACTGTCAGCTAGTCAGA
READ:          gggGTGTAACC-GACTAGgggg
```

a) 11M1I10M

b) 3S8M1D6M4S

c) 9M32N8M

d) 3H8M1D6M4H

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.