1.10 Exercises

1.10.1 Next Generation Sequencing Data

In this exercise we work with next generation sequencing (NGS) data. Unix is excellent at manipulating the huge FASTA files that are generated in NGS experiments.

FASTA files contain sequence data in text format. Each sequence segment is preceded by a single-line description. The first character of the description line is a "greater than" sign (>).[15]

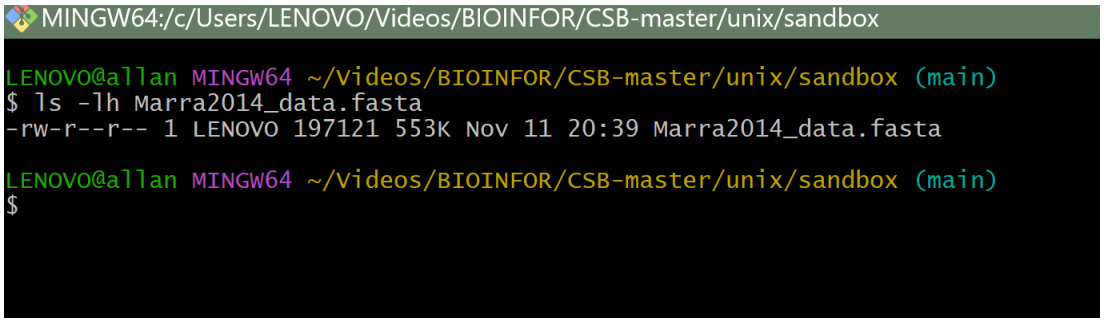
The NGS data set we will be working with was published by Marra and DeWoody (2014), who investigated the immunogenetic repertoire of rodents. You will find the sequence file `Marra2014_data.fasta` in the directory `CSB/unix/data`. The file contains sequence segments (contigs) of variable size. The description of each contig provides its length, the number of reads that contributed to the contig, its isogroup (representing the collection of alternative splice products of a possible gene), and the isotig status.

1. Change directory to `CSB/unix/sandbox`.
2. What is the size of the file `Marra2014_data.fasta`?[16]
3. Create a copy of `Marra2014_data.fasta` in the `sandbox` and name it `my_file.fasta`.
4. How many contigs are classified as `isogroup00036`?
5. Replace the original "two-spaces" delimiter with a comma.
6. How many unique isogroups are in the file?
7. Which contig has the highest number of reads (`numreads`)? How many reads does it have?

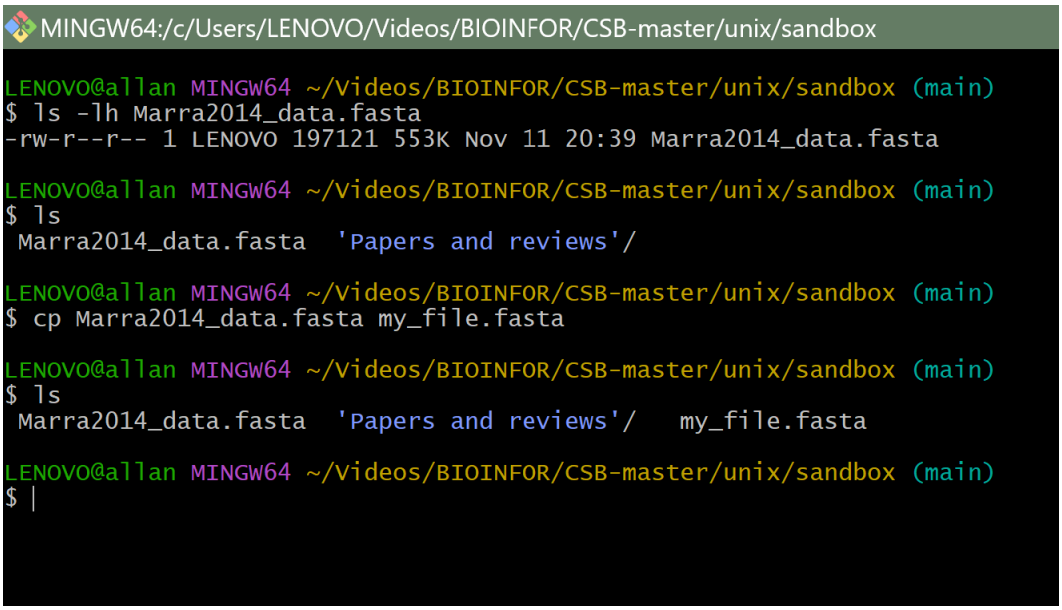
LIT 1

```
LENOVO@allan MINGW64 ~/Videos/BIOINFOR/CSB-master/unix/data (main)
$ ls
Buzzard2015_about.txt  Gesquiere2011_about.txt  Marra2014_about.txt   Pacifici2013_about.txt  Saavedra2013/           miRNA/
Buzzard2015_data.csv   Gesquiere2011_data.csv   Marra2014_data.fasta  Pacifici2013_data.csv   Saavedra2013_about.txt  sandbox

LENOVO@allan MINGW64 ~/Videos/BIOINFOR/CSB-master/unix/data (main)
$ mv Marra2014_data.fasta ../

LENOVO@allan MINGW64 ~/Videos/BIOINFOR/CSB-master/unix/data (main)
$ ls
Buzzard2015_about.txt  Gesquiere2011_about.txt  Marra2014_about.txt     Pacifici2013_data.csv  Saavedra2013_about.txt  sandbox
Buzzard2015_data.csv   Gesquiere2011_data.csv   Pacifici2013_about.txt  Saavedra2013/          miRNA/

LENOVO@allan MINGW64 ~/Videos/BIOINFOR/CSB-master/unix/data (main)
$ cd ../

LENOVO@allan MINGW64 ~/Videos/BIOINFOR/CSB-master/unix (main)
$ ls
Marra2014_data.fasta  data/  installation/  sandbox/  solutions/

LENOVO@allan MINGW64 ~/Videos/BIOINFOR/CSB-master/unix (main)
$ mv Marra2014_data.fasta sandbox/

LENOVO@allan MINGW64 ~/Videos/BIOINFOR/CSB-master/unix (main)
$ ls
data/  installation/  sandbox/  solutions/

LENOVO@allan MINGW64 ~/Videos/BIOINFOR/CSB-master/unix (main)
$ cd sandbox/

LENOVO@allan MINGW64 ~/Videos/BIOINFOR/CSB-master/unix/sandbox (main)
$ ls
 Marra2014_data.fasta  'Papers and reviews'/

LENOVO@allan MINGW64 ~/Videos/BIOINFOR/CSB-master/unix/sandbox (main)
$
```

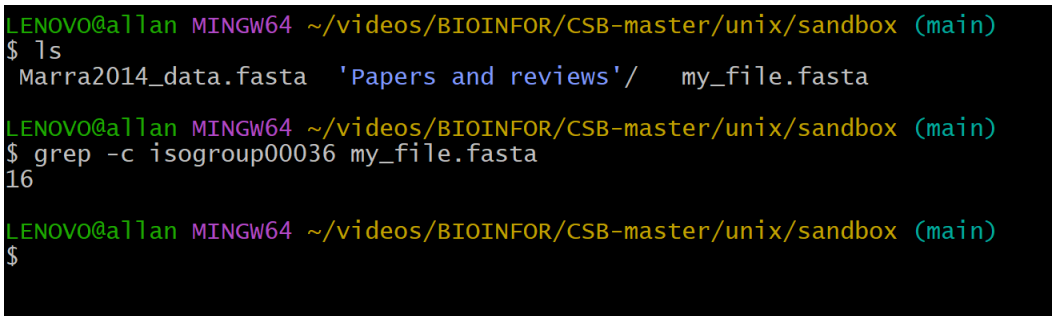
LIT 2



LIT 3



LIT 4



LIT 5

```
LENOVO@allan MINGW64 ~/videos/BIOINFOR/CSB-master/unix/sandbox (main)
$ cat my_file.fasta | tr ' ' ',' | head -n 10
>contig00001,,length=527,,numreads=2,,gene=isogroup00001,,status=it_thresh
ATCCTAGCTACTCTGGAGACTGAGGATTGAAGTTCAAAGTCAGCTCAAGCAAGAGATTTG
TTTACAATTAACCCACAAAAGGCTGTTACTGAAGGTGTGGCTTAAGTGTCAGAGCAACAG
CTATGAGTGGAGGAATTTTCTATTACAATATAATTTCATCTCTGGTAAATTGACCAATTA
ACTGGAACTTTTTCCAACTGAAATAAATGGTAAACTTTTTATCCACCATTCTGCCATCTG
ACTCACAAAGACCCATGGGAATGGGTGATGAAATCCAACATGCTTCTTTGTAGCAAAAAT
AAATAAAATCCCCAGAAGGGTGAGGTAAATGGAAAACTCCAAACTCGCCCCTCAGGTGGG
TGTAATTTACCCAAGTCTGAGAGGAGGCAGAGTTTTTCCCAATGGACTTTGGTTAAGTGA
GATATGCTGGTCTGTAGAAGGAGGGAGTTCTAGGAAAACAGACACTTAAGTAGGGCCGAA
CTAAAAATTGTATCAGTCAGATCTTCATGTGAAGTCCTGTGTGCCCA

LENOVO@allan MINGW64 ~/videos/BIOINFOR/CSB-master/unix/sandbox (main)
$
```

LIT 6

```
LENOVO@allan MINGW64 ~/videos/CSB-master/CSB-master/unix/sandbox (main)
$ grep " gene" my_file.fasta | cut -d '=' -f 4-4 | uniq -c
    147 isogroup00001  status
     47 isogroup00002  status
     30 isogroup00003  status
     28 isogroup00004  status
     26 isogroup00005  status
     27 isogroup00006  status
     27 isogroup00007  status
     26 isogroup00008  status
     15 isogroup00009  status
     21 isogroup00010  status
     21 isogroup00011  status
     17 isogroup00012  status
     22 isogroup00013  status
     21 isogroup00014  status
     15 isogroup00015  status
     23 isogroup00016  status
     18 isogroup00017  status
     20 isogroup00018  status
     16 isogroup00019  status
     12 isogroup00020  status
     21 isogroup00021  status
     15 isogroup00022  status
     22 isogroup00023  status
     18 isogroup00024  status
     19 isogroup00025  status
     18 isogroup00026  status
     20 isogroup00027  status
     16 isogroup00028  status
     16 isogroup00029  status
     17 isogroup00030  status
     13 isogroup00031  status
     14 isogroup00032  status
     19 isogroup00033  status
     16 isogroup00034  status
     15 isogroup00035  status
     16 isogroup00036  status
     17 isogroup00037  status
     16 isogroup00038  status
     19 isogroup00039  status
     17 isogroup00040  status
     16 isogroup00041  status
     13 isogroup00042  status
      3 isogroup00043  status

LENOVO@allan MINGW64 ~/videos/CSB-master/CSB-master/unix/sandbox (main)
$
```

LIT 7

```
LENOVO@allan MINGW64 ~/videos/CSB-master/CSB-master/unix/sandbox (main)
$ grep "contig" my_file.fasta | cut -d '=' -f 1-1 | sort -n | tail -n 1
>contig01385  length

LENOVO@allan MINGW64 ~/videos/CSB-master/CSB-master/unix/sandbox (main)
$
```