BIOL495\_lab2\_excersices4 Thomas Sanchez

**2.14 Exercises IV**

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 (>).

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 lab1\_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. Open a shell, change working directory to lab1\_unix/sandbox.
2. Create a copy of Marra2014\_data.fasta in the sandbox and name it my\_file.fasta.
3. How many lines are there in my\_file.fasta?
4. How many sequences are there in my\_file.fasta (hint: each new try of sequence starts with “>contig”, when using grep remember to add “\” before “>”, as “>” means directing to a file without the backslash )
5. save all the lines start with “>contig” to a new file.
6. How many contigs are classified as isogroup00030?
7. Replace “=” and double spaces with a comma. (hint: remove consecutive spaces, use “-s” option in tr)
8. How many unique isogroups are in the file?
9. Which contig has the highest number of reads (numreads)? How many reads does it have?

$ wc -l n[23][0-9].txt

(provides linecount for n20.txt to n39.txt