Unix exercise handout

**Warm up: unix commands**

1. Enter Unix environment (3 different ways depending on your computer).
2. Do the following using Unix command

- Check where you are

- See what files and directories exist here

- Go to your (My) Document folder

- Create a new directory for today exercise

- Enter your new directory

- See the content of your new directory (it should be empty :))

1. Download files from internet using wget command

* Go to this repository: <https://github.com/akoiwang/Unix_training/>
* Get URL addres of these files
  + real\_GFF.gff
  + real\_FASTA.fa
  + mini\_GFF.gff
  + mini\_FASTA.fa

1. Then do the following

* Move all 4 files to your newly-created directory from 2)
* How many lines are there in total?
* Try command cat / less / head / tail on mini\_FASTA.fa and mini\_GFF.gff
* Try the same commands on real\_FASTA.fa and real\_GFF.gff
* In real\_GFF.gff files, check how many lines contain information about “CDS”
* In mini\_FASTA and real\_FASTA, check how many sequences the file contain

1. More on less command

* Try looking into a file using less, and type forward slash (/) follow by a word/name you want to search for.
  + Hit “enter”, what happen?
  + Hit “n”, what happen?
  + Hit “N”, what happen?

1. On real\_GFF file, select only “CDS” information from chromosome “Schisto\_mansoni.Chr\_2.unplaced.SC\_0213” and save this information to a new file
2. Using file from 6) extract gene ID, keep only unique ID, then sort by ID and save this information to a new file

**Now, let’s work with bioinformatics files**

- Go here and download genomic sequence of *Paramecium tetraurelia* in FASTA format and genome annotation in GFF format. You may download using the GUI (graphical user interface, i.e. clicking), or continue using wget command

<http://protists.ensembl.org/Paramecium_tetraurelia/Info/Index?db=core>

or <https://www.ncbi.nlm.nih.gov/genome/275>

- Make sure the files are in your work directory

**Practice your Unix skills and learn something new too**

* The files from database are often downloaded as zipped file. How might you unzip it? (gunzip) Try to stay on command line as much as you can.
* Explore your downloaded files
* How many files do you have now in the directory?
* How many lines are in each/all of the files?
* List all files in a directory by chronological order (listed by date and time that the file has been modified)
* What do these commands mean?

o ls -lh

o ls -lt

o ls -R

o ls -l \*.fasta

o ls -l \*.\*[!z]

o ls -l {\*.fasta,\*.fa}

o ls -l \*f[a-z]

* How many sequences are there in the paramecium .fasta file?
* How many types of annotation are there in .gff file?
* How many genes are annotated in the genome?
* How can you get genomic coordinates of all CDS?
* How many genes have multiple exons
* How many genes are there on each of the + and - strands?
* How many genes (and transcripts) are there on each chromosome/contig?

**Try bacteria genomes**

- <https://bacteria.ensembl.org/Escherichia_coli_55989/Info/Index> how many header line is there? Is this the number you expect?

- What about this one?<http://bacteria.ensembl.org/Bacillus_cereus_e33l/Info/Index>

- And this?<https://bacteria.ensembl.org/Paeniclostridium_sordellii_8483/Info/Index>

- The header lines are quite long and contain a lot of information. Extract specific information from the header lines. For example, extract all the contig IDs

- Bacteria also have plasmids... how can you get plasmid sequences from the genome data?

- How many CDS are there in each chromosome/contigs?

- How many genes are there in each chromosome/contigs?

- What is contig? How does a contig differ from a chromosome?

- The bacteria files might be getting mixed up with planaria files.. let’s tidy up this workspace. Move genomics files into two directories. Delete any large intermediate files. Rename files if need be but retain necessary information.

**Further curiosity**

- On the Ensembl download page, there is a file called CHECKSUM...

o What is CHECKSUM? Why is it useful?

o How might you calculate CHECKSUM of your downloaded files?

o You might also see .dat file. How does a .dat file differ from a .fasta file?

o What are the difference between various .fasta or .gff files?

- How does Unix deal with file names or directory names that have “a space” in it? Also, files with # or - in its name?

**Species challenges**

- Of the 3 bacteria species, which one(s) contain gene insB-1 (Ensembl gene ID: EC55989\_1042, Ensembl transcript ID: CAU96904). And are there any homolog in the other two bacteria species? Try to extract information about this homolog (of the insB-1 gene if a homolog doesn’t exist)

- What genes are common between at least 2 bacteria species.

- Simplify the fasta headers by shortening the species name

- When you downloaded the genome, you might find some files with similar names, but same file size, are they different?

- We have fasta file (contain genome sequences) and gff (saying which location on the genome has what features e.g. exon, intron, protein coding regions)... how might we extract the information to get fasta of protein coding genes?