Post-lecture notes: Class Session 3

Examplar data

- Yeast Genome Sequence (FASTA) – https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_00
- Yeast Genome Annotation (GFF) – https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_00
- For some details about the yeast genome see: https://www.ncbi.nlm.nih.gov/genome/15
  - see the table at the bottom that gives the correspondence between RefSeq IDs and chromosome #'s

Downloading files using `wget`

- `wget` is a command line tool that can be used to download a file directly to your VM (rather than having to download to your laptop and then re-upload to your VM). The most common usage of wget is of the form:

  ```
  wget URL_OF_FILE
  ```

- For example, to download the FASTA file above to the working directory on your VM you could execute the following command:

  ```
  wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_0
  ```

Decompressing compressed `.gz` files with `gunzip`

- Files that end with `.gz` are typically files that have been compressed with the `gzip` tool to make the data smaller and faster to transmit over the internet. `gzip` is similar to the `zip` utility which is found on most Windows and MacOS computers.
- Before we use such files we need to de-compress them using the `gunzip` command. The general form of this command is:

  ```
  gunzip COMPRESSED_FILE.gz
  ```

- For example, to decompress the FASTA file downloaded above we would execute this command:

  ```
  gunzip GCF_000146045.2_R64_genomic.fna.gz
  ```

  - The resulting decompressed file that is produced is `GCF_000146045.2_R64_genomic.fna`

Seting up symbolic lines with `ln -s`

- Symbolic links to files can be used to setup "shortcuts" or "aliases" to files with long names or that are stored in different parts of the filesystem
- The general form of creating a symbolic link is as follows:

  ```
  ln -s ORIGINAL_FILE SHORT_NAME
  ```

- For example, to create an alias with a more easily understood name than `GCF_000146045.2_R64_genomic.fna` we could do the following:

  ```
  ln -s GCF_000146045.2_R64_genomic.fna yeast.fna
  ```

- Once the symbolic link is created `yeast.fna` refers to `GCF_000146045.2_R64_genomic.fna` and we can substitute this symbolic link for the longer name in commands. For example, the following is more convenient, but in reality still reads from `GCF_000146045.2_R64_genomic.fna` "behind the scenes":

  ```
  less yeast.fna
  # equivalent to less GCF_000146045.2_R64_genomic.fna
  ```

- Note that deleting a symbolic link does NOT delete the original file

New commands introduced

For details about each of these commands: - Read my overview of the Unix Core Utilities - Then take a look at the `man` pages (e.g. `man echo`) to read about various options

- `less`
- `head`
- `tail`
- `echo`
- `cat` and `tac`
- `rev`
- `fold`
- Redirection operators:
  - `>` = redirect output to a file
    - `echo Hello World > hello.txt`
  - `>>` = append output to a file
    - `echo Goodbye World >> hello.txt`
  - `<` = redirect input to a command

Command we didn't have time to discuss but I'd like you to review

- `tr` – translate (substitute) or delete characters in input. Note that unlike most commands `tr` will not take a file as an argument, so typically you would use `cat` or input redirection to send the contents of a file through `tr`. Example
  - `tr 'e' '3' < hello.txt` – substitutes all occurences of "e" with "3"
  - `echo AATTAGACCAAC | tr "ATCG" "TAGC"` – computes the complemennt of a DNA nucleotide sequences

Examples of computations on genome annotation using grep and cut

- Filtering metadata lines out of a GFF file using `grep`

```
grep -E -v "^#" yeast.gff
```

- Getting specific columns (seqid = 1, feature type = 3) from the GFF file

```
grep -E -v "^#" yeast.gff | cut -f 1,3
```

- How many features are there on yeast chromosome II (NC_001134.8)?

```
grep -E -v "^#" yeast.gff | cut -f 1 | grep -c "NC_001134.8"
```

- How many exons are there on yeast chromosome II (NC_001134.8)?

```
grep -E -v "^#" yeast.gff | cut -f 1,3 | grep "NC_001134.8" |  grep -c "exon"
```