Introduction to Scientific Computing: A Crash Course

Presented by Travis J Lawrence and Dana L Carper
Quantitative and Systems Biology
University of California, Merced

Worksheet 3.1.1

Exploring genome annotations

1. Change to the directory containing the genome annotation file. What command would you use to list only the genome annotation files in the directory? You will need to use a wildcard.
2. Using `less` explore one of the annotation files.
3. Some lines start with `#`. Do these appear to be annotation records?
4. Using `grep` how would you select lines that begin with `#`?
5. How would you select lines that don't start with `#`? Use the `manpage` for grep to find an option to do this.
6. Using the output from the command in `question 5` figure out the number of annotations each genome contains. Remember that you can use the output from one command as the input to another command using the `|` character.
7. Using `cut` how many fields does each annotation line contain?
8. What information does each field contain? Some fields might not be clear, however, you can lookup the `gff` format online.
9. Interested in the number of features annotated for each chromosome develop a series of commands that will provide this information. Use `grep`, and `wc` to develop this pipeline.
10. Using `head` or `less` What is the first annotation? Should we include this `feature` in our counts for each chromosome? Modify the pipeline from `question 9` to exclude this annotation type from your counts.
11. The solution to `question 9 and 10` requires that you run the pipeline for each chromosome. Modify the pipeline from `question 10` using `cut`, `sort`, `uniq` and an option for `uniq` to reproduce the results from `question 10` but in one command.
12. Based on how `uniq` functions did you need to use the `sort` command in the pipeline for `question 11`?
13. Sort the results from `question 11` in descending order by the number of features.
14. Interested in the type of features annotated in the genome develop a pipeline that reports each `feature type` once.
15. Develop a pipeline to count the number of times each `feature type` occurs.
16. Sort the results from `question 15` in descending order by the number of times each feature occurs.
17. Interested in the distribution of features on each chromosome develop a pipeline that give this results. You will need to use `grep`, `cut`, `sort`, and `uniq`.
18. Wanting to use the results from `question 17` in a later analysis redirect the output to a file with

a descriptive name.

19. Looking at the results from `question 17` does the ratio of `genes` to `mRNA` have a biological explanation?

Preparing RNA-seq annotation file

Several RNA-seq pipelines require annotation files in a format that differs from `gff`. Even when a pipeline accepts `gff` formatted annotations there are multiple reasons you might want to modify the available annotation. The questions below will lead you through producing a `gene` level annotation in another popular format called `SAF`. `SAF` format starts with a header line with the names of the five required columns separated by tabs. The column names are `GeneID`, `Chr`, `Start`, `End`, `Strand`, and additional columns with supplemental annotation information may be added.

20. Using a text editor open a new file and write the header line with column names then save and exit your text editor.
21. How would you select only the `gene` features from the `gff` annotation file?
22. The `GeneID` column in a `SAF` file is a unique identifier for each feature. The gene accession tends to be a good choice for this field because it makes down stream analyses easier. Which field in the `gff` annotation contains this information?
23. What delimiter is used to separate the meta data fields in the answer to `question 22`?
24. Develop a command pipeline that extracts the gene accession number for each gene in the `gff` annotation file. You will need to us `grep`, and `cut`. Redirect the output of this command to a new file.
25. Next you need to extract the chromosome, start, end, and strand information from the `gff` file. Using `cut` you should be able to get this information in one command. Redirect the output of this command to a new file.
26. You now have all the information you need for your `SAF` annotation file, but in three separate files. First you should combine the `GeneID` file and the file that contains the chromosome, start, end, and strand information. You can do this using the `paste` command. We did not cover this command in the lecture. Read the `manpage` and experiment with the command to see if you can get the result you want. Once you are getting the results you want redirect the output to a new file.
27. Finally, you need to combine the header file you made in `question 20` and file you produced in `question 26`. You can do this by using the `cat` command and the append redirect `>>`. Before attempting to do this make backup copies of your original files from `questions 20 and 26`. Again we did not lecture on the `cat` command, but quick experimentation should reveal the function of this command. The append redirect `>>` causes the redirected output to be appended to the bottom of the file instead of overwriting it.