

## Bash: A Command Language Interpreter

GFF files are genomic data files that contain both a table of features and FASTA formatted genome sequences. The description of the format is described at <https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md>. In this assignment, you will be reverse engineering a bash script (on page 2) that downloads and analyzes [https://downloads.yeastgenome.org/curation/chromosomal\\_feature/saccharomyces\\_cerevisiae.gff](https://downloads.yeastgenome.org/curation/chromosomal_feature/saccharomyces_cerevisiae.gff).

The script is complicated because it will give you an opportunity to learn several scripting techniques and utilities. Using nano on Agave, copy this script into a file called 01-script.bash. Note that you will learn better if you go through the effort to type the script instead of copying-and-pasting.

1. As you are typing, add a comment describing what each line does. You will likely need to consult man pages and reverse engineer the script (delete/change lines to see how it effects the output) to determine what it is doing.
2. Using comment lines at the beginning of the script, write a detailed explanation of what the script does and how it works over all. Take what you know about each individual line and pull them into an over all understanding of the script.

## Additional Resources

1. *Bioinformatics Data Skills*: chapters 3, 4, 7, & 11 (on Canvas)
2. *Data Science at the Command Line*: chapters 1, 2, & 4 (on Canvas)
3. <https://stackoverflow.com/a/3871336>
4. <https://www.gnu.org/savannah-checkouts/gnu/bash/manual/faq.html>

## Submitting

Feel free to discuss your answers with each other and ask for help. Google is also very helpful to understand what commands do.

When you are finished upload your annotated script to Canvas. Remember to include the information/references of any resources you used to complete the assignment (apart from manual pages and help messages). This includes fellow students.

```
1 # Usage: bash 01-script.bash
2
3 url=https://downloads.yeastgenome.org/curation/\
4 chromosomal_feature/saccharomyces_cerevisiae.gff
5
6 filename=$(basename "$url")
7
8 echo "Chr,Genes"
9 curl -s "$url" |
10 sed -e '###FASTA/,$d' |
11 grep -v '^#' |
12 awk '$3 == "gene"' |
13 cut -f1 |
14 sort |
15 uniq -c |
16 sort -k1,1nr |
17 awk '{print $2 "," $1}'
```

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs