David Halvorsen
UCSC Data Wrangler Interview Handouts
My Blog: https://DaveHalvorsen.github.io
My Code: https://github.com/DaveHalvorsen

- **Education**
    - 2018 UCSC Database and Data Analytics Certificate
    - 2011 SUNY Biochem BS w/ Physics minor
- **Employment**
    - Ski Instructor 2003-2007: Freestyle
    - TA/Tutor 2009-2011: OCHEM
    - Research Associate 2012-2017: 384-well ImmunoFISH, Flow Cytometry, 96-well +/- Phi29 Telomeric qPCR, Mammalian Cell Culture, Dot Blots, ELISA, Western Blots, TRAP, 96-well DNA Purification, PicoGreen, Transfection, 384-well Plate Reader (Absorbance, Fluorescence, FRET, Luminescence), DNA Circle Quantification, 20 GB of Electronic Notes
- **Fundraising**
    - $72,000 Lifespan.io Cancer Crowdfunding Campaign 2016
    - $25,000 LEF Grant 2014
    - $200K Tour de Cure collab grant Jeremy Henson 2015
- **Presentations**
    - Halvorsen, D., Hunt, T., Aggrawal, M., Silva, H. (2014). Development of a high-content, automated platform for rapid analysis of alternative lengthening of telomeres (ALT)-associated promyelocytic leukemia nuclear bodies (APBs) in human cancer cells. Telomeres, Telomerase & Disease Poster Session, Brussels, Belgium.
    - Halvorsen, D., Hunt, T., Moody, K., Silva, H. (2013). ALTered cancer cells: Uncovering the genetic basis of ALT (Alternative Lengthening of Telomeres). SENS 6 Lecture, Cambridge, England.
- **Scientific Writing**
    - Silva H; Halvorsen D; Henson JD, 2015, 'Control ALT, delete cancer', Scientist, vol. 29
    - Halvorsen, D., Silva, H. 2015. High Throughput Telomeric Circle Assay. U.S. Patent Application US2015031831, filed May 2015. Patent Pending
- **Hobbies**
    - Sketch Comedy, Improv Comedy, Skiing, Cosplay, Web Development
- **Biological Data Wrangling Examples (Python, R, MySQL, Shell & Terminal):**
    - **I get the following error when I try to use the “Name files as submitted and put into subdirectories” shell script for downloading the kriegsteinRadialGliaStudy1 dataset: “curl: (60) server certificate verification failed”. Here’s how I fixed the script in Python: #Python #ShellScript**

```
#!/usr/bin/env python
file = open("kriegsteinGliaName_files_as_submitted_and_put_into_subdirectories.sh")
lines = file.readlines()
new_file = open("Working_Glia_Submitted_Many_Directories.sh", "a")
for line in lines:
    entry = line.rstrip() + " --insecure \n"
    new_file.write(entry)
file.close()
new_file.close()
```

- ○ **Kallisto is a program for quantifying transcripts from RNA-Seq data. The github.io manual for Kallisto says that abundance.tsv est_counts is "estimated counts" and tpm is "Transcripts Per Million". Presumably, you would be interested in the samples with the highest transcript abundance. Here's an R script I wrote to do that: #R**

SEE https://github.com/DaveHalvorsen/Wrangling_CIRM_Data/tree/master/kriegsteinGlia_Projects

- ○ **The 217 GB quakeBrainGeo1 data set is going to take FOREVER to download :(… I'll make a MySQL database for its file info with Python and MySQL: #Python #MySQL**

SEE https://github.com/DaveHalvorsen/Wrangling_CIRM_Data/tree/master/quakeBrain_Projects

```
mysql root@localhost 13:09 [quakeBrain] > SELECT * FROM Quake_Shell_Table WHERE
file_name = "SRR1974678_1.fastq.gz";
+-----------+-----------+------------+-----------------------+
| accession | file_type | meta_name  | file_name             |
+-----------+-----------+------------+-----------------------+
| sc000AUC  | reads     | SRR1974678 | SRR1974678_1.fastq.gz |
+-----------+-----------+------------+-----------------------+
```

- ○ **I received this error at least twice (currently @127/217 GB) while downloading the quakeBrainGeo1 dataset: "curl: (56) GnuTLS recv error (-9)". There's no record of what files failed, BUT the shell script file order should match the terminal output :D #Python**

SEE https://github.com/DaveHalvorsen/Wrangling_CIRM_Data/tree/master/quakeBrain_Projects

```
##################### this is the output of the python code ##########################
failure at download number 271
"SRR1974678_1.fastq.gz"
"51  644M   51  335M    0     0  4945k      0  0:02:13  0:01:09  0:01:04 4836k"
'https://cirm.ucsc.edu/cgi-bin/cdwGetFile?acc=sc000AUC'
failure at download number 563
SRR1974824_1.fastq.gz
25  122M   25 30.8M    0     0   379k      0  0:05:31  0:01:23  0:04:08  500k
'https://cirm.ucsc.edu/cgi-bin/cdwGetFile?acc=sc000BFM'
```

- ○ **The failed downloads mentioned above WERE downloaded, BUT they are smaller in size than the files on your webserver. HOWEVER, the files still open … this could lead to downstream data analysis errors. Here's some terminal sleuthing:**

```
################## the files that failed are present ########################################
$ find . -name SRR1974678_1.fastq.gz
./raw/reads-ByExp-sra-SRX-SRX995-SRX995996-SRR1974678-/SRR1974678_1.fastq.gz
$ ls -l ./raw/reads-ByExp-sra-SRX-SRX995-SRX995996-SRR1974678-/SRR1974678_1.fast
q.gz
-rwxrwxrwx 1 david david 351404032 Dec 16 03:33 ./raw/reads-ByExp-sra-SRX-SRX995-
SRX995996-SRR1974678-/SRR1974678_1.fastq.gz

################# HOWEVER, my files are smaller than your server files ###################
My SRR1974678_1.fastq.gz" is 351.4 MB, BUT CIRM website says it should be 676105674 (676 MB)
and Chrome download of file from CIRM is 676.1 MB. Therefore: MY DOWNLOAD FAILED
My SRR1974824_1.fastq.gz is 32.4 MB, BUT CIRM website says it should be 128736201 (128 MB)
and Chrome download of file from CIRM is 128 MB. Therefore: MY DOWNLOAD FAILED
```