6-8-2020

Summary

Today's aims were to set up the new macOS desktop computer to remotely access the rcapps5 research computing cluster. This included configuring the computer for Partners VPN compliance, and customizing the terminal for easy use. The permissions for VPN were already established and the computing cluster had been used in the past, so steps for those were not necessary this time.

Partners VPN Setup

Setup the mac to access Partners VPN. The key steps are:

1. installing the Partners peas self service app and the certificates that go along with it
2. Installing cisco anyconnect - through the self service app
3. Installing forescout secure connector - through the self service app

Accessing the VPN

1. Open terminal
2. Connect via ssh

```
$ ssh -XY esk17@rcapps5.dfci.harvard.edu
```

For the X features to work, I downloaded XQuartz for Mac.

Iterm Setup

I followed this tutorial:
https://www.freecodecamp.org/news/jazz-up-your-zsh-terminal-in-seven-steps-a-visual-guide-e81a8fd59a38/

Summary:

1. Use homebrew to install the newest version of zsh

```
$ brew install zsh
$ chsh -s /usr/local/bin/zsh
```

2. Install oh my zsh

```
$ sh -c "$(curl -fsSL
https://raw.github.com/robbyrussell/oh-my-zsh/master/tools/install.sh)"
```

3. Install syntax highlighting plugin

```
$ git clone https://github.com/zsh-users/zsh-syntax-highlighting.git
${ZSH_CUSTOM:-~/.oh-my-zsh/custom}/plugins/zsh-syntax-highlighting
```

4. Install the add zsh autosuggestion plugin

```
$ git clone https://github.com/zsh-users/zsh-autosuggestions
$ZSH_CUSTOM/plugins/zsh-autosuggestions
```

5. Activate plugins by adding "zsh-autosuggestions" and "zsh-syntax-highlighting" under the plugins in "~/.zshrc"

File Sharing Setup

I needed to download sshfs and FUSE from https://osxfuse.github.io/

I previously wrote the script for setting up scp but edited it for the current directories:

```
#!/bin/bash
#sshfs_setup.sh

echo "Unmount disk?"
select yn in "Yes" "No"; do
      case $yn in
      Yes ) echo "Disk to unmount:"; read pathToDir;diskutil unmount force
$pathToDir;hdiutil eject -force $pathToDir; break;;
      No ) break;;
      esac
done

echo "Mount disk?"
select yn in "Yes" "No"; do
      case $yn in
      Yes ) echo "Directory to mount:"; read toMount;sshfs -p 22
esk17@rcapps5.dfci.harvard.edu:/mnt/beegfs/home/esk17/data/
/Users/kim03/$toMount; break;;
      No ) exit;;
      esac
done
```

Lastly, I installed sublime text as a text editor.

6-9-2020

Summary

The goal for today was to subscribe to ReadCube for handling journal articles.

To-do

Tomorrow, I need to email Shengbao and Adrienne to schedule a meeting to talk about first steps. For the meeting I will bring up the bioconductor tutorial as a way to do robust analysis. I will also clean/reorganize the working environment in the server main directory. Ask about gene imputation. It was not recommended in the bioconductor tutorial but was part of the Regev lab's ulcerative colitis paper.

6-10-2020

Summary

Today, I subscribed to ReadCube for handling journal articles.

6-12-2020

Summary

I should start using tmux because the scp file transfer of the PD1-colitis fastq files broke overnight. In the meeting today, I got a clarification that gene imputation is probably not needed for 10X data, which is pretty good quality. We would rather not introduce biases into the data. Since our PD-1 data is in fastq format right now, I will install cellranger to obtain the counts matrix.

Transferring PD-1 patients' fastq files

```
$ cd /home/sv467/sv467_tmp/project/colon_PD1/
$ scp -r 0.raw.data/ esk17@rcapps5.dfci.harvard.edu:/home/esk17
```

```
# maybe try this in future
$ rsync -r -P -e ssh 0.raw.data/ esk17@rcapps5.dfci.harvard.edu:/home/esk17
the future
```

Installing Tmux for mac

```
$ brew install tmux
# example code for starting a new session: tmux new -s session1
```

Installing Cellranger

I followed the guide online:
https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest?

Another useful link:
https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/fastq-input

Example for unzipping tar.gz

```
$ tar -xzvf refdata-cellranger-GRCh38-and-mm10-3.1.0.tar.gz
$ tar -xzvf refdata-cellranger-GRCh38-3.0.0.tar.gz
```

Added this to my .bashrc file

```
$ export PATH=/home/esk17/cellranger/cellranger-3.1.0:$PATH
```

Post-installation steps for testing and registration

```
$ cellranger sitecheck > sitecheck.txt
$ cellranger upload your@email.edu sitecheck.txt
$ cellranger testrun --id=tiny
```

6-15-2020

Summary

I presented my thesis for an audience. I was notified that I need to establish a better hypothesis for my thesis. I expect differences in the inflammatory responses between the two types but the reasons why need to be justified by literature and theory. I also ran two cellranger processes using 12 cores respectively to run over night.

Log

Fastq files to convert to count matrices

```
DFCI-1294-CD3_S1_L001_I1_001.fastq.gz   DFCI-1618-CD3_S1_L001_R2_001.fastq.gz
DFCI-1294-CD3_S1_L001_R1_001.fastq.gz   p020620-GEX_S1_L001_I1_001.fastq.gz
DFCI-1294-CD3_S1_L001_R2_001.fastq.gz   p020620-GEX_S1_L001_R1_001.fastq.gz
DFCI-1294-CD45_S1_L001_I1_001.fastq.gz  p020620-GEX_S1_L001_R2_001.fastq.gz
DFCI-1294-CD45_S1_L001_R1_001.fastq.gz  p020620-TCR_S1_L001_I1_001.fastq.gz
DFCI-1294-CD45_S1_L001_R2_001.fastq.gz  p020620-TCR_S1_L001_R1_001.fastq.gz
DFCI-1294-TCR_S1_L001_I1_001.fastq.gz   p020620-TCR_S1_L001_R2_001.fastq.gz
DFCI-1294-TCR_S1_L001_R1_001.fastq.gz   p101519-CD3_S1_L001_I1_001.fastq.gz
DFCI-1294-TCR_S1_L001_R2_001.fastq.gz   p101519-CD3_S1_L001_R1_001.fastq.gz
DFCI-1545-CD3_S1_L001_I1_001.fastq.gz   p101519-CD3_S1_L001_R2_001.fastq.gz
DFCI-1545-CD3_S1_L001_R1_001.fastq.gz   p101519-TCR_S1_L001_I1_001.fastq.gz
DFCI-1545-CD3_S1_L001_R2_001.fastq.gz   p101519-TCR_S1_L001_R1_001.fastq.gz
DFCI-1618-CD3_S1_L001_I1_001.fastq.gz   p101519-TCR_S1_L001_R2_001.fastq.gz
DFCI-1618-CD3_S1_L001_R1_001.fastq.gz
```

CD3 Samples to be analyzed:
DFCI-1294-CD3
DFCI-1545-CD3
DFCI-1618-CD3
p101519-CD3

Started these processes at 5:20 PM CT

```
$ cellranger count --id=DFCI-1294-CD3 \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=DFCI-1294-CD3 \
                  --localcores=12 \
                  --localmem=64
#need to choose human only reference

$ cellranger count --id=DFCI-1545-CD3 \
  --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=DFCI-1545-CD3 \
```

```
                    --localcores=12 \
                    --localmem=256
```

In order to monitor CPU and memory usage, I open a new terminal displaying the htop monitor.

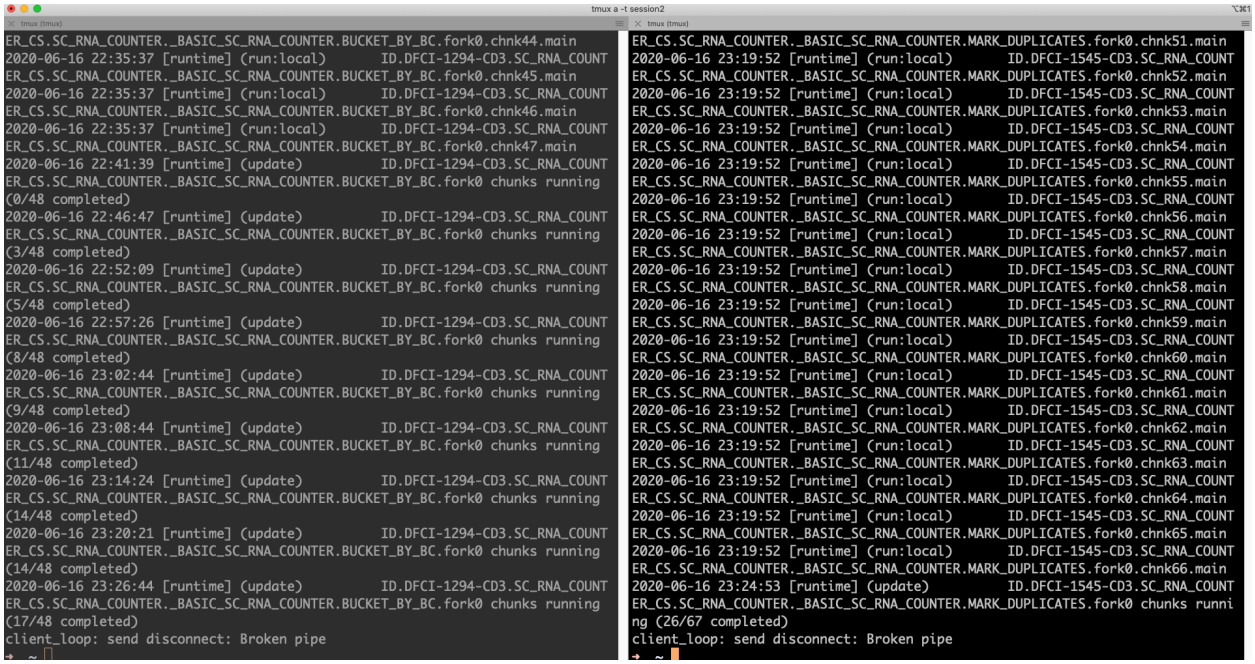
6-16-2020

Summary

Observed that the cellranger processes got disrupted by a broken ssh connection. I resumed them in the morning and they finished around 3PM in the afternoon. I ran two more samples for CD3, and I will see the results.

Log

I got a broken pipe result



Not sure if this means I should run it again. I checked and the ssh connection had been disconnected, even while using tmux.

I ran the commands again and the processes “resumed.” I suspect the connection broke when I disconnected the VPN.

I checked at night and found that the cellranger ran successfully. It took about a full day. I will try to see if there is a way to check the total time.

Output from a successful cellranger process:

```
Outputs:
- Run summary HTML:                         /mnt/beegfs/home/esk17/data/PD1/DFCI-1545-CD3/outs/web_summary.html
- Run summary CSV:                          /mnt/beegfs/home/esk17/data/PD1/DFCI-1545-CD3/outs/metrics_summary.csv
- BAM:                                      /mnt/beegfs/home/esk17/data/PD1/DFCI-1545-CD3/outs/possorted_genome_bam.bam
- BAM index:                                /mnt/beegfs/home/esk17/data/PD1/DFCI-1545-CD3/outs/possorted_genome_bam.bam.bai
- Filtered feature-barcode matrices MEX:    /mnt/beegfs/home/esk17/data/PD1/DFCI-1545-CD3/outs/filtered_feature_bc_matrix
- Filtered feature-barcode matrices HDF5:   /mnt/beegfs/home/esk17/data/PD1/DFCI-1545-CD3/outs/filtered_feature_bc_matrix.h5
- Unfiltered feature-barcode matrices MEX:  /mnt/beegfs/home/esk17/data/PD1/DFCI-1545-CD3/outs/raw_feature_bc_matrix
- Unfiltered feature-barcode matrices HDF5: /mnt/beegfs/home/esk17/data/PD1/DFCI-1545-CD3/outs/raw_feature_bc_matrix.h5
- Secondary analysis output CSV:            /mnt/beegfs/home/esk17/data/PD1/DFCI-1545-CD3/outs/analysis
- Per-molecule read information:            /mnt/beegfs/home/esk17/data/PD1/DFCI-1545-CD3/outs/molecule_info.h5
- CRISPR-specific analysis:                 null
- Loupe Cell Browser file:                  /mnt/beegfs/home/esk17/data/PD1/DFCI-1545-CD3/outs/cloupe.cloupe
- Feature Reference:                        null

Waiting 6 seconds for UI to do final refresh.
Pipestance completed successfully!

2020-06-17 15:09:53 Shutting down.
Saving pipestance info to "DFCI-1545-CD3/DFCI-1545-CD3.mri.tgz"
```

Samples starting tonight:

DFCI-1618-CD3

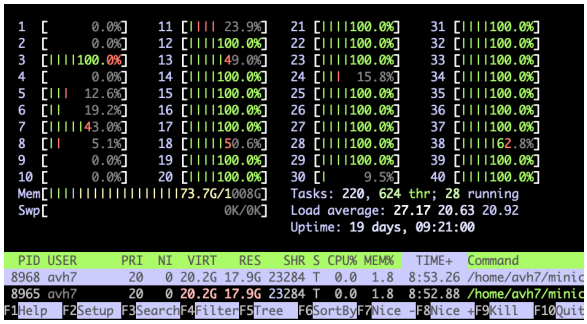
p101519-CD3

```
$ cellranger count --id=DFCI-1618-CD3 \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=DFCI-1618-CD3 \
                  --localcores=12 \
                  --localmem=64
#need to choose human only reference

$ cellranger count --id=p101519-CD3 \
  --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=p101519-CD3 \
                  --localcores=15 \
                  --localmem=256
```

My fingers are crossed that these run overnight without there being a broken pipe instance.

The htop console:



6-17-2020

Summary

Observed that the cellranger processes got disrupted by a broken ssh connection. I resumed them in the morning and they finished around 3PM in the afternoon. I ran two more samples for CD3, and I will see the results.

This time, the ssh connection broke at 1:15 AM.

Resolution to the ssh connection problem

I was starting my tmux sessions from my local device and then connecting via ssh to the remote cluster. This resulted in the problem that anytime the local computer disconnected from the remote cluster, the tmux session would continue but the ssh connection within the tmux would discontinue. For example, as soon as the VPN expired overnight, the ssh connection broke. The solution was to start the tmux session within the remote cluster such that even if I disconnect, the tmux session and its processes will continue uninterrupted. Thus, the new way to run programs is to 1) connect to the remote cluster via ssh, 2) start a tmux session, 3) run programs in that session. These processes will be safe even if there is a surprise disconnection.

6-17-2020

Summary

Came across runtime errors. Sent help request to cellranger team. Also, I found it easier to ssh into the remote cluster once, start one tmux session, and use multiple windows within that session using tmux shortcuts.

Errors due to running two cellranger processes at the same time

```
2020-06-18 21:45:04 [runtime] Pipestance directory seems to have disappeared.
        stat /mnt/beegfs/home/esk17/data/PD1/DFCI-1618-CD3/_log: stale NFS file handle
2020-06-18 21:45:04 Shutting down.
rm: cannot remove '__DFCI-1618-CD3.mro': Stale file handle
/mnt/beegfs/home/esk17/cellranger/cellranger-3.1.0/cellranger-cs/3.1.0/bin/../tenkit/bin/common/_mrp: line 39: tarmri: command not found
```

After trying to resume (meaning inputting the same cellranger count command):

```
Martian Runtime - '3.1.0-v3.2.3'

RuntimeError: pipestance 'DFCI-1618-CD3' already exists and is locked by another Martian instance. If you are sure no other Martian instance is running, dele
te the _lock file in /mnt/beegfs/home/esk17/data/PD1/DFCI-1618-CD3 and start Martian again.

2020-06-20 00:30:36 Shutting down.
Saving pipestance info to "DFCI-1618-CD3/DFCI-1618-CD3.mri.tgz"
For assistance, upload this file to 10x Genomics by running:

cellranger upload <your_email> "DFCI-1618-CD3/DFCI-1618-CD3.mri.tgz"
```

Log

While waiting for a response, I decided to start this instead

```
$ cellranger count --id=DFCI-1294-CD45 \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=DFCI-1294-CD45 \
                  --localcores=12 \
                  --localmem=64
```

DFCI-1294-CD45
Started 6/20 at 12:39 AM ET

6-22-2020

Log

The run was successful! So I will run one more overnight.

```
$ cellranger count --id=DFCI-1294-TCR \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=DFCI-1294-TCR \
                  --localcores=12 \
                  --localmem=64
```

The run was successful! So I will run one more during the day

```
$ cellranger count --id=p020620-TCR \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=p020620-TCR \
                  --localcores=12 \
                  --localmem=64
```

6-23-2020

Summary

The run was successful. Today, I will focus on the pipeline and analysis code. My focus will be on following the Bioconductor tutorial: osca.bioconductor.org

Notes from Bioconductor Introduction:

- Workflows follow
    - Data import
    - Quality control and normalization
    - Feature selection
    - Dimensionality reduction
    - Analysis (clustering, differential expression)

Notes from Ch 2: Learning R and Bioconductor:

- Bioconductor is a repository with strict requirements of consistent data infrastructure, high quality documentation, and focus on genomic analysis
- For installing on mac and linux, use a package manager for R
- Bioconductor packages documentation guide:
    - Vignette - showcases basic functionality of the package
    - Reference manual - comprehensive listing of all the functions
    - E

I should also read this source: https://whattheyforgot.org/

Log

```
$ cellranger count --id=p101519-TCR \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=p101519-TCR \
                  --localcores=12 \
                  --localmem=64
```

6-24-2020

Summary

I continued to read the bioconductor online source and continued to produce counts matrices

Log

I moved the PD1 CD45 dataset for patient p101519, which had been left out.

```
$ ssh -XY esk17@rcapps4.dfci.harvard.edu
$ cd /home/sv467/sv467_tmp/project/colon_PD1/0.raw.data
$ scp p101519-CD45* esk17@rcapps5.dfci.harvard.edu:/home/esk17
```

```
DFCI-1294-CD3_S1_L001_I1_001.fastq.gz   p020620-GEX_S1_L001_I1_001.fastq.gz
DFCI-1294-CD3_S1_L001_R1_001.fastq.gz   p020620-GEX_S1_L001_R1_001.fastq.gz
DFCI-1294-CD3_S1_L001_R2_001.fastq.gz   p020620-GEX_S1_L001_R2_001.fastq.gz
DFCI-1294-CD45_S1_L001_I1_001.fastq.gz  p020620-TCR_S1_L001_I1_001.fastq.gz
DFCI-1294-CD45_S1_L001_R1_001.fastq.gz  p020620-TCR_S1_L001_R1_001.fastq.gz
DFCI-1294-CD45_S1_L001_R2_001.fastq.gz  p020620-TCR_S1_L001_R2_001.fastq.gz
DFCI-1294-TCR_S1_L001_I1_001.fastq.gz   p101519-CD3_S1_L001_I1_001.fastq.gz
DFCI-1294-TCR_S1_L001_R1_001.fastq.gz   p101519-CD3_S1_L001_R1_001.fastq.gz
DFCI-1294-TCR_S1_L001_R2_001.fastq.gz   p101519-CD3_S1_L001_R2_001.fastq.gz
DFCI-1545-CD3_S1_L001_I1_001.fastq.gz   p101519-CD45_S1_L001_I1_001.fastq.gz
DFCI-1545-CD3_S1_L001_R1_001.fastq.gz   p101519-CD45_S1_L001_R1_001.fastq.gz
DFCI-1545-CD3_S1_L001_R2_001.fastq.gz   p101519-CD45_S1_L001_R2_001.fastq.gz
DFCI-1618-CD3_S1_L001_I1_001.fastq.gz   p101519-TCR_S1_L001_I1_001.fastq.gz
DFCI-1618-CD3_S1_L001_R1_001.fastq.gz   p101519-TCR_S1_L001_R1_001.fastq.gz
DFCI-1618-CD3_S1_L001_R2_001.fastq.gz   p101519-TCR_S1_L001_R2_001.fastq.gz
```

I ran the following processes.

```
cellranger count --id=p020620-CD45 \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=p020620-GEX \
                  --localcores=12 \
                  --localmem=64

cellranger count --id=p101519-CD45 \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=p101519-CD45 \
                  --localcores=12 \
                  --localmem=64
```

Tomorrow my goal will be to manage my papers manager using Papers and work on sorting through my references. In addition, I will read about QC metrics and normalization

6-25-2020

Summary

Log

I got the following error, which I had observed previously:

```
2020-06-25 14:13:04 [runtime] Pipestance directory seems to have
disappeared.
            stat /mnt/beegfs/home/esk17/data/PD1/p020620-CD45/_log: stale
NFS file handle
2020-06-25 14:13:04 Shutting down.
rm: cannot remove '__p020620-CD45.mro': Stale file handle
(r_env) esk17@rcapps5:/mnt/beegfs/home/esk17/data/PD1$
ranger-cs/3.1.0/bin/../tenkit/bin/common/_mrp: line 39: tarmri: command not
found
```

I think the reason is that when two cellranger processes run at the same time and one process finishes early, the other cellranger process is locked/halted.

I also get the following error, so I need to exit the directory and re-enter the directory.

```
ls: cannot open directory '.': Stale file handle
```

I get the following error when I try to resume the cellranger.

I removed the lock file as per the instructions, and the cellranger process was able to continue.

The installations finished by 10PM CT. I organized my existing data into “0.raw.data” and “1.projects”. The projects directory contains the analyses and data such as filtered matrices (counts matrices). I also setup SublimeText on my local machine to interact with iTerm such that there is autocompletion and sending over to the terminal using Cmd+Enter.

I came across the problem that R 3.6.1 is not compatible with DoubletFinder.

```
install.packages("remotes")
remotes::install_github('chris-mcginnis-ucsf/DoubletFinder')
```

Setup Sublime Text as an R IDE

- Key features: interactivity between terminal and script editor; autocompletion
- Does not show attached data structures or plots

Steps in Sublime Text:

1. Open Command Palette (Cmd+Shift+P or Tools > Command Palette)
2. Enter “Install Package Control”
3. Use “Package Control: Install Package” in Command Palette to install “SendCode”
4. Use “SendCode: Choose Program” and select the appropriate terminal interface (in this case iTerm)
5. Use “Set Syntax: R” for syntax autocompletion

Notes from bioconductor

- They form their own distinct cluster(s), complicating interpretation of the results. This is most obviously driven by increased mitochondrial proportions or enrichment for nuclear RNAs after cell damage. In the worst case, low-quality libraries generated from different cell types can cluster together based on similarities in the damage-induced expression profiles, creating artificial intermediate states or trajectories between otherwise distinct subpopulations. Additionally, very small libraries can form their own clusters due to shifts in the mean upon transformation (Lun 2018).

I should be careful to read this paper in order to evaluate macrophages that have engulfed T cells

“The first few principal components will capture differences in quality rather than biology, reducing the effectiveness of dimensionality reduction.”

- They contain genes that appear to be strongly “upregulated” due to aggressive scaling to normalize for small library sizes. This is most problematic for contaminating transcripts (e.g., from the ambient solution) that are present in all libraries at low but constant levels. Increased scaling in low-quality libraries transforms small counts for these transcripts in large normalized expression values, resulting in apparent upregulation compared to other cells. This can be misleading as the affected genes are often biologically sensible but are actually expressed in another subpopulation.

Key QC metrics for UMI-based scRNA-seq:

- Low library size. The library size is defined as the total sum of counts across all relevant features for each cell.
- The number of expressed features in each cell.
- High mitochondrial reads.

The adaptive QC thresholds portion suggests log-transforming the data before applying the normal threshold:

For the 416B data, we identify cells with log-transformed library sizes that are more than 3 MADs below the median. A log-transformation is used to improve resolution at small values when `type="lower"`. Specifically, it guarantees that the threshold is not a negative value, which would be meaningless for a non-negative metric. Furthermore, it is not uncommon for the distribution of library sizes to exhibit a heavy right tail; the log-transformation avoids inflation of the MAD in a manner that might compromise outlier detection on the left tail. (More generally, it makes the distribution seem more normal to justify the 99% rationale mentioned above.)

6-27-2020

Summary

Today, I did most of the QC work for the p1294 CD3 sample. I checked model fitting and decided that the multiple hypothesis testing model was good and very close to the 3 MAD +- the sample median (which was recommended by Biocondutor for adaptive thresholding schemes). Tomorrow, I plan to move onto the HVF and scaling parts. On Monday, I plan to execute the integration.

Log

Following the Bioconductor tutorial, I wrote three functions (filter_by_mito, filter_by_libsize, filter_by_numfeats), which perform the filtering based on mitochondrial reads, library size, and number of features. For now, I will only use filter by numfeats and mito to filter quality cells. I am using the filtered counts matrix from Cellranger, so I do not need to use emptyDrops.

I analyzed the mitofiltering (as well as the other filtering criteria) and produced the following graphs to demonstrate how good the Norm(med, MAD) model is, and how good the procedure is. The figures show that the model fit is very good and that the procedure results in a threshold that is nearly equivalent to the 3 MAD above the sample median.

The following are templates for visualizing the distributions based on the output to the filtering. The first is for pre vs post comparison. The second is for model fitting.

```
### Figures to evaluate how good the mito filtering is performing
p <- ggplot(data=data.frame(x= c(rep("pre mito%",
length(q1.p1294$pre.mt.fraction)),rep("post mito%",
length(q1.p1294$post.mt.fraction))), mt.fraction =
c(q1.p1294$pre.mt.fraction,q1.p1294$post.mt.fraction))) +
geom_violin(aes(x= x, y=mt.fraction))
ggsave("figures/mitohist.pdf")

source("scripts/Stat111functions.R")
mito.data <- q1.p1294$pre.mt.fraction
p <- plot_kde(mito.data, kernel = "gaussian", bw=0.01, lab.x =
"mitochondrial gene fraction (per cell)", overlay = T, model="Norm",
    dmodel = function(x) {dnorm(x,median(mito.data),mad(mito.data))},
    support=seq(0,1,0.001))
p + geom_vline(xintercept= q1.p1294$mt.lim, color = "red") +
    geom_vline(xintercept= median(mito.data)+ 3*mad(mito.data), linetype =
"dotted")
ggsave("figures/mito-kde-Normal.pdf")

q1.p1294$total.removed/length(q1.p1294$pre.mt.fraction) #to find percentage
```

```
filtered
```

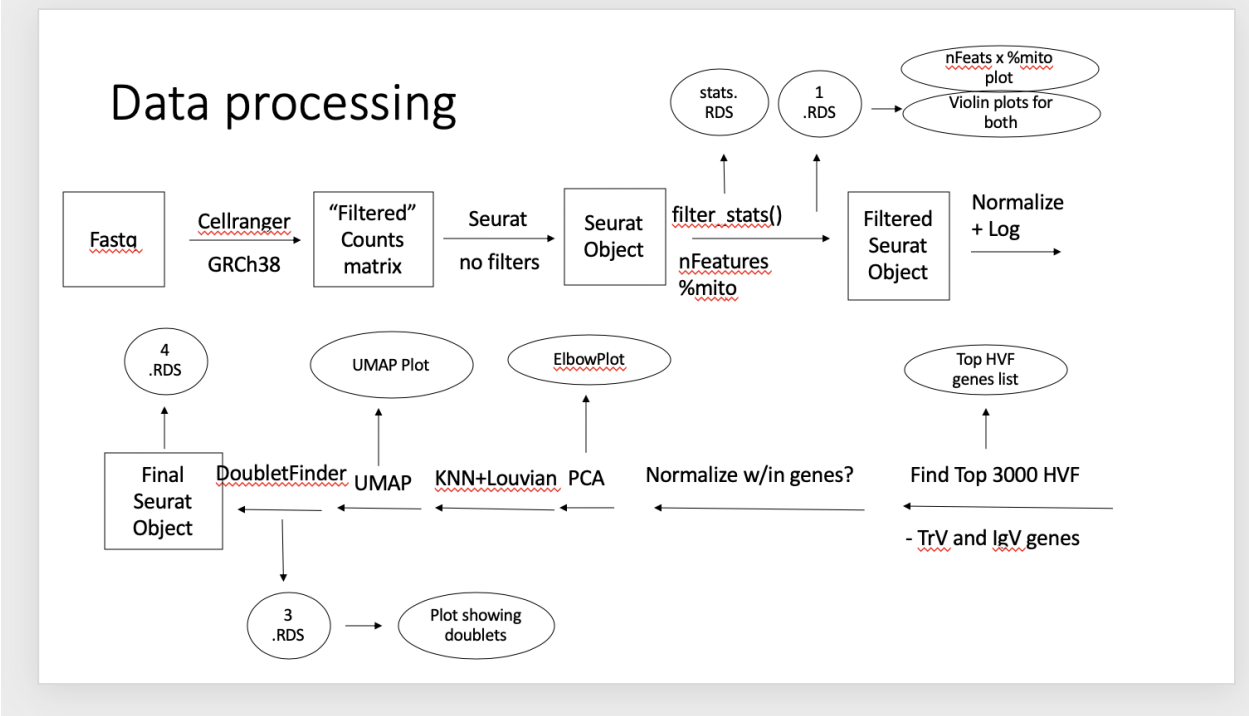
I also checked the rationale behind the multiplicative scaling factor behind the log-transformation, which is 10,000. This is because after the calling, most genes need a 10,000 multiplier to become a value greater than 1.

I then applied doublet removal on the p1294 data. I observed which cells were marked by Doublet Finder.

Information about the T cell Seurat object from Shengbao: 17874 genes and 67926 cells.
I also looked at the distributions of the cells and nFeature_RNA values.

Finally, based on observing the different violin plots, I decided that the best metrics for low quality cell removal are the same as the ones suggested on the Seurat vignette page, namely a *union* of the number of unique features and percent mitochondrial reads. I wrote a function "filter_stats", which outputs relevant stats about the filtering including the cells that need to be removed. I used this to highlight the cells that need to be removed. I think this can be applied to all the analyses moving forward.

I think the following early pipeline will be the best: (ovals are saved plots or objects)



6-29-2020

Summary

Today, I edited the find.pK function to remove the plotting outputs, which were resulting in errors. I called the edited function “find.pK.noPlot”. I wrote the wrapper function “remove_doublets” to execute DoubletFinder using the “find.pK.noPlot” and label the cells as high, low, or no likelihood of being a doublet. I included these new functions in a R script that can be sourced, called “new_clustering_functions.R”.

Tomorrow, my plan is to **implement the data processing pipeline** laid out yesterday. I might write a wrapper function for saving the data at the right points for QC figures. This means I will need to refine the pipeline for subsetting the T cells from the patient with only CD45 data. Afterwards, I plan to integrate the dataset with Shengbao’s T cell dataset, which includes the ipi colitis, no colitis cases, and others.

Log

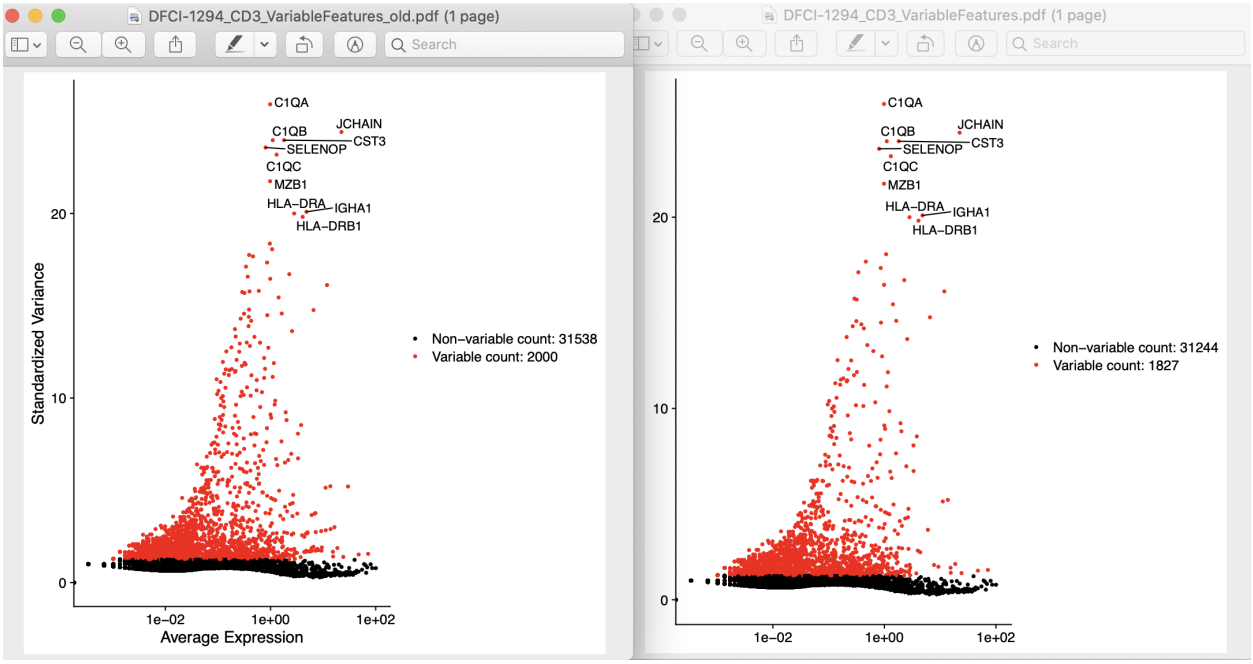
View codelog3 and codelog4 for the code that I was working on. The gist of what I did was noted in the summary. The only thing I left out was that I experimented with SCTransform and the variable features list that it output. The list included some TCR_V genes and probably some IG_V genes, so I will filter those out if I end up using SCTransform.

6-30-2020

Summary

I implemented the workflow. I learned that the grepl("^TR.V") recognized some genes that I did not want to remove, including TRPV genes, which encode a calcium channel, and a TROVE2 gene. Also not sure about the TRDV gene and will confirm with Adrienne.

I edited the Variablefeatures plot to exclude the genes that were removed from the hvf list (new function is in the new_clustering_functions.R file named "VariableFeaturePlot.Tcell". Comparison of resultant plot below:



Plan for tomorrow

- I need to think about whether to get rid of the variable genes during the variance calculation and standardization.
    - I do not think so. This is because the variance calculation and standardization is attempting to exclude technical noise. This is separate from the biological significance of the variance. In reality the variance should fall in a large range of insignificant to large because of biological clonotype differentiation. We should remove these from the variable genes list but not before the standardization procedure.
- I also need to redo the process if needed. It seems like the top 2000 variable genes were not affected by the correction in TR*V gene removal.
    - 2000 gene selection seems good from the variable features plots because most of the highly variable genes were captured and 2000 was not below the technical noise line

- I will repeat the procedure for all the samples.
- subsetting the T cells from the patient with only CD45 data.
- subsetting the T cells from the patient with only CD45 data.

7-1-2020

Summary

Executed the implemented workflow on the remaining CD3 samples, found that DFCI1545 has a small number of cells, and discovered that DFCI1618 and p101519 CD3 samples were not fully processed by Cellranger. Re-ran one overnight. (Also, installed tree command from source.)

Log

DFCI1545
33538 243
A very small number of cells!

DFCI1618 CD3 cellranger was not finished.

Installed tree command (unix):

```
$ wget http://mama.indstate.edu/users/ice/tree/src/tree-1.8.0.tgz
$ tar zxvf tree-1.8.0.tgz
$ cd tree-1.8.0
$ make
# edit the Makefile's prefix to be local
$ make install
#add the bin path to the ~/.bashrc file
```

DFCI-1618-CD3 and p101519-CD3 Cellranger directories did not have outs. I.e. They were not finished. I removed the locks and resumed them...

```
$ cellranger count --id=p101519-CD3 \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=p101519-CD3 \
                  --localcores=12 \
                  --localmem=64
```

It looks like they were corrupted. Because the process above was unable to resume. I deleted the existing directory of progress and restarted the process.

7-2-2020

Summary

Fixed my conda environment, updated R to 4.0.2, installed “scran” package.

Log

Problem 1: Cellranger had halted.

```
2020-07-02 08:25:09 [runtime] Pipestance directory seems to have disappeared.
        stat /mnt/beegfs/home/esk17/1.projects/PD1/data/p101519-CD3/_log: stale NFS file handle
2020-07-02 08:25:09 Shutting down.
rm: cannot remove '__p101519-CD3.mro': Stale file handle
/mnt/beegfs/home/esk17/cellranger/cellranger-3.1.0/cellranger-cs/3.1.0/bin/../tenkit/bin/common/_mrp: line 39: tarmri: command not found
```

Solution: Not sure what the cause is, but the server was down overnight, which may have halted the progress. I resumed it and it was able to finish.

Problem 2: “scran” package function “bootstrap_cluster” unavailable for R-3.6.1.

Solution: Install new R environment using conda and with the latest version of R (managed by the conda package manager)

I had some issues with updating the current R environment because I had done some local-system-level modifications that had really messed with the updating, so I installed a fresh copy of conda (miniconda) and used that instead of the old conda system.

1. Installed miniconda3 (did not remove anaconda3, just in case; luckily it did not interfere)
2. Created a new miniconda3 environment with R-4.0.2:

```
$ conda config --add channels conda-forge
$ conda create -n r_env2
$ conda activate r_env2
$ conda search r-base
$ conda install r-base=4.0.2
$ conda install -c anaconda libgcc-ng
```

Past anaconda settings that I commented out and removed.

```
## >>> conda initialize >>>
## !! Contents within this block are managed by 'conda init' !!
#__conda_setup="$('/home/esk17/anaconda3/bin/conda' 'shell.bash' 'hook' 2>
/dev/null)"
#if [ $? -eq 0 ]; then
#     eval "$__conda_setup"
#else
#     if [ -f "/home/esk17/anaconda3/etc/profile.d/conda.sh" ]; then
#     . "/home/esk17/anaconda3/etc/profile.d/conda.sh"
#     else
#     export PATH="/home/esk17/anaconda3/bin:$PATH"
```

```
#     fi
#fi
#unset __conda_setup
# <<< conda initialize <<<

# alias rconda="conda activate r_env"
```

Problem 3: Issues installing some packages in conda's R installation (gcc and g++ compiler issues)

Makeconf and ldpaths are fine: they are pointing to the conda's updated GNU compilers.
Installing bioconductor using the r-base R:
Obtained the following error:

```
g++ -I"/home/esk17/miniconda3/envs/r_env2/lib/R/include" -DNDEBUG -I../inst/include/  -DNDEBUG -D_FORTIFY_SOURCE=2 -O2 -isystem /home/esk17/miniconda3/envs/r
_env2/include -I/home/esk17/miniconda3/envs/r_env2/include -Wl,-rpath-link,/home/esk17/miniconda3/envs/r_env2/lib   -fpic  -fvisibility-inlines-hidden  -fmes
sage-length=0 -march=nocona -mtune=haswell -ftree-vectorize -fPIC -fstack-protector-strong -fno-plt -O2 -ffunction-sections -pipe -isystem /home/esk17/minico
nda3/envs/r_env2/include -fdebug-prefix-map=/home/conda/feedstock_root/build_artifacts/r-base_1593070357708/work=/usr/local/src/conda/r-base-4.0.2 -fdebug-pr
efix-map=/home/esk17/miniconda3/envs/r_env2=/usr/local/src/conda-prefix  -c api.cpp -o api.o
g++: error: unrecognized command line option '-fno-plt'
make: *** [/home/esk17/miniconda3/envs/r_env2/lib/R/etc/Makeconf:175: api.o] Error 1
ERROR: compilation failed for package 'Rcpp'
* removing '/mnt/beegfs/home/esk17/miniconda3/envs/r_env2/lib/R/library/Rcpp'
```

Later in the execution:

```
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (locfit)
ERROR: dependency 'Rcpp' is not available for package 'RcppAnnoy'
* removing '/mnt/beegfs/home/esk17/miniconda3/envs/r_env2/lib/R/library/RcppAnnoy'
ERROR: dependency 'Rcpp' is not available for package 'RcppHNSW'
* removing '/mnt/beegfs/home/esk17/miniconda3/envs/r_env2/lib/R/library/RcppHNSW'
ERROR: dependency 'Rcpp' is not available for package 'sitmo'
* removing '/mnt/beegfs/home/esk17/miniconda3/envs/r_env2/lib/R/library/sitmo'
* installing *source* package 'S4Vectors' ...
```

Later:

```
ERROR: dependencies 'rhdf5', 'Rhdf5lib' are not available for package
'HDF5Array'
* removing
'/mnt/beegfs/home/esk17/miniconda3/envs/r_env2/lib/R/library/HDF5Array'
ERROR: dependencies 'Rcpp', 'RcppAnnoy', 'RcppHNSW' are not available for
package 'BiocNeighbors'
```

More:

```
ERROR: dependency 'isoband' is not available for package 'ggplot2'
* removing
'/mnt/beegfs/home/esk17/miniconda3/envs/r_env2/lib/R/library/ggplot2'
ERROR: dependency 'ggplot2' is not available for package 'ggbeeswarm'
* removing
'/mnt/beegfs/home/esk17/miniconda3/envs/r_env2/lib/R/library/ggbeeswarm'
ERROR: dependency 'ggplot2' is not available for package 'viridis'
* removing
'/mnt/beegfs/home/esk17/miniconda3/envs/r_env2/lib/R/library/viridis'
```

```
ERROR: dependencies 'ggplot2', 'ggbeeswarm', 'DelayedMatrixStats',
'viridis', 'Rcpp', 'BiocNeighbors', 'BiocSingular' are not available for
package 'scater'
* removing
'/mnt/beegfs/home/esk17/miniconda3/envs/r_env2/lib/R/library/scater'
ERROR: dependencies 'Rcpp', 'scater', 'edgeR', 'BiocNeighbors', 'igraph',
'DelayedMatrixStats', 'BiocSingular', 'dqrng' are not available for package
'scran'
* removing
'/mnt/beegfs/home/esk17/miniconda3/envs/r_env2/lib/R/library/scran'
```

...

```
configure: error: couldn't find zlib library.
      Please specify a location using --with-zlib=/path/to/zlib
ERROR: configuration failed for package 'Rhdf5lib'
```

...

```
ERROR: dependency 'Rhdf5lib' is not available for package 'rhdf5'
```

I did the following to install some of the packages. For example “Rcpp” is included in conda’s r-essentials package.

```
$ conda install r-essentials r-igraph
```

Still some packages cannot be installed:

```
1: In install.packages(...) :
  installation of package 'Rhdf5lib' had non-zero exit status
2: In install.packages(...) :
  installation of package 'sitmo' had non-zero exit status
3: In install.packages(...) :
  installation of package 'igraph' had non-zero exit status
4: In install.packages(...) :
  installation of package 'rhdf5' had non-zero exit status
5: In install.packages(...) :
  installation of package 'dqrng' had non-zero exit status
6: In install.packages(...) :
  installation of package 'HDF5Array' had non-zero exit status
7: In install.packages(...) :
  installation of package 'DelayedMatrixStats' had non-zero exit status
8: In install.packages(...) :
  installation of package 'scater' had non-zero exit status
9: In install.packages(...) :
```

```
  installation of package 'scran' had non-zero exit status
```

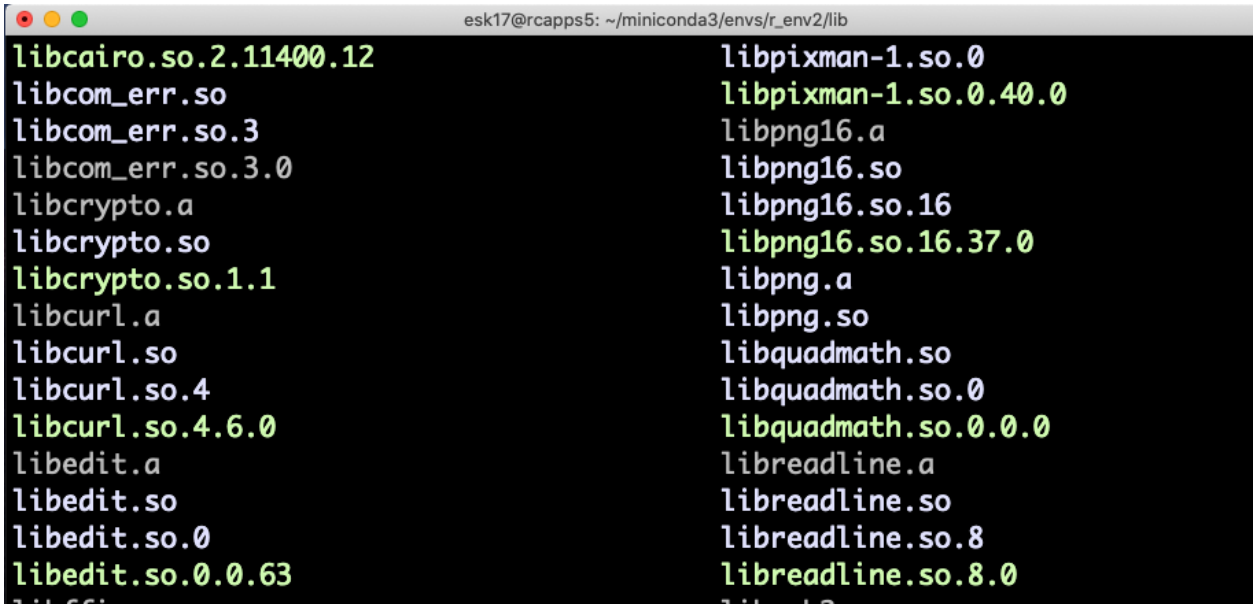
Solution:

```
ln -s
/home/esk17/miniconda3/envs/r_env2/bin/x86_64-conda_cos6-linux-gnu-gcc
/home/esk17/miniconda3/envs/r_env2/bin/gcc
ln -s
/home/esk17/miniconda3/envs/r_env2/bin/x86_64-conda_cos6-linux-gnu-g++
/home/esk17/miniconda3/envs/r_env2/bin/g++
```

Note 1: for some reason, even though the Makeconf file for the conda R installation notates environmental variables that point to the conda gcc (7.3.0) executables, the install.packages() functions sometimes defaults to using the local gcc and g++ (perhaps by using the executables found through “which gcc” or “which g++”. This solution circumvents the problem by creating a symbolic gcc and g++ executable in the environment bin directory such that “which gcc” and “which g++” will point to the conda gcc and g++ versions (7.3.0). The library paths seem to be fine - nothing needs to be done there (although I did previously set LD_LIBRARY_PATH=”/home/esk17/miniconda3/envs/r_env2/lib:/usr/local/lib/” , which could be having an effect).
Note 2: remove gcc and g++ symbolic links when trying to update conda. For example, only update conda outside of r_env2, which is the environment that stores the symbolic links.

Problem 3: Issues with installing R-4.0.2 from source
I continued the attempt to install R-4.0.2 from source even though it had been successfully installed via the conda package manager in the r_env2 environment. The error I encountered was that the readline libraries were not being found (even though they were present in the environment, i.e. “conda install readline” did not update or install anything).

```
checking how to hardcode library paths into programs... immediate
checking for cos in -lm... yes
checking for sin in -lm... yes
checking for dlopen in -ldl... yes
checking readline/history.h usability... no
checking readline/history.h presence... no
checking for readline/history.h... no
checking readline/readline.h usability... no
checking readline/readline.h presence... no
checking for readline/readline.h... no
checking for rl_callback_read_char in -lreadline... no
checking for main in -lncurses... no
checking for main in -ltermcap... no
checking for main in -ltermlib... no
checking for rl_callback_read_char in -lreadline... no
configure: error: --with-readline=yes (default) and headers/libs are not a
```



Solution:
Fixed the readline problem by setting the following before executing “configure”:

```
$ export CONDA_BUILD=1
$ export
LD_LIBRARY_PATH="/home/esk17/miniconda3/envs/r_env2/lib:/usr/local/lib/"
$ conda activate r_env2
```

Note: https://takehomessage.com/2020/01/07/virtual-environment-r-development/
Not sure how to use “conda build” ...

```
$ conda install conda-build
$ conda build . # <- froze at this command
```

--- could be relevant in the future, as of now, the above solution works amazingly (even without the symbolic link gcc and g++ executables).

Ran the following overnight (last one!):

```
cellranger count --id=DFCI-1618-CD3 \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data \
                  --sample=DFCI-1618-CD3 \
                  --localcores=12 \
                  --localmem=64
```

7-3-2020

Summary

Finished the QC for the samples. Now will integrate into the Tcell dataset. Integration script was written in “codelog5.R” (sidenote: moved metadata key for Shengbao’s T cell collection from codelog4 to codelog5). I also subsetted the T cells from the p020620-CD45 population.

	Orig. count	Removed (Mito&lnf)	Mito upper thres	Feats lower thres	Doublet hi	Doublet lo	Singlet	Final count
DFCI-1294_CD3	3198	267	10.40	983.00	235	58	2638	2696
DFCI-1545_CD3	243	25	9.78	1004.00	8	14	196	210
DFCI-1618_CD3	4103	415	10.67	890.00	310	59	3319	3378
p101519_CD3	4316	349	8.04	897.00	313	84	3570	3654
p020620_CD45	9519	435	11.88	291.00	797	111	8176	8287
p020620_CD45v2	9519	435	11.88	291.00	796	112	8176	8288

Regarding the different clustering methods between v1 (npc 20 and VariableFeatures list excluding T cell variable genes) and v2 (npc 15 and variable features list with no filtering), the number of final cells is essentially the same, so I will go with v1. Subsetting the T cells from this dataset, I got 3488 total T cells.

7-4-2020

Summary

I did some data wrangling to match meta.data features of Shengbao's T cell dataset, which means adding cells. Merging everything we get: 33538 features and 81352 cells. I trimmed the features to only include features that have more than 10 cells expressing it. This resulted in 18188x81352.

Found this amazing resource: https://broadinstitute.github.io/2019_scWorkshop/

Check: do I need to add 1 to the nFeatures in order to do the log properly???

Check: how to compare runtimes in R?

- library(microbenchmark)
  http://adv-r.had.co.nz/Performance.html#faster-r

7-5-2020

Summary

https://www.r-bloggers.com/rstudio-server-part-3-using-an-ssh-tunnel-for-high-performance/

```
$ ssh -f -N -L 1234:localhost:8787 esk17@rcapps5.dfci.harvard.edu
```

Kill background ssh tunnel:

```
# First, find the PID of the background ssh tunnel
$ ps aux | grep ssh
# Then, kill the PID process
$ kill <insert PID>
```

^ issue is that I can only use R-3.6.2 with the Rstudio-server installation on this server.

Manually installed uuid: http://www.ossp.org/pkg/lib/uuid/

```
$ mkdir build
$ ./configure --prefix=/home/esk17/uuid-1.6.2/build
$ make
$ make check
$ make install
# add the path to the build/bin in ~/.bashrc
```

7-6-2020

Summary

Ran SCTransform online using “multicore” setting for future. The result is a cluster that resembles the TOX paper in Science, but the clusters are much less clearly partitioned. In addition, I had to modify the VariableFeaturePlot() once more to display the highly variable genes based on the VariableFeatures() list of modified highly variable genes. For some reason, the list of HVG was not being modified in the HVGInfo() function, so I implemented the removal of those genes in a new function called VariableFeaturePlot.Tcells.SCT().

7-13-2020

Summary

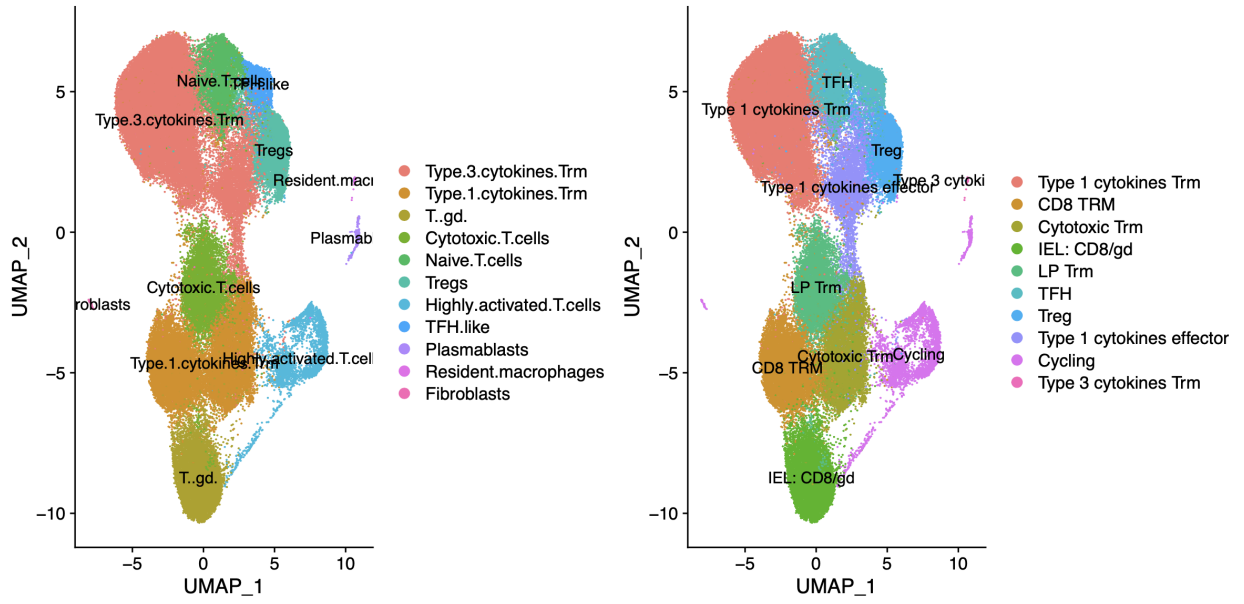
Recently, I have been working on my thesis proposal. I am modifying it and rewriting it as I receive feedback.

8-3-2020

Summary

Tasks are labeling and prelim analysis of clusters

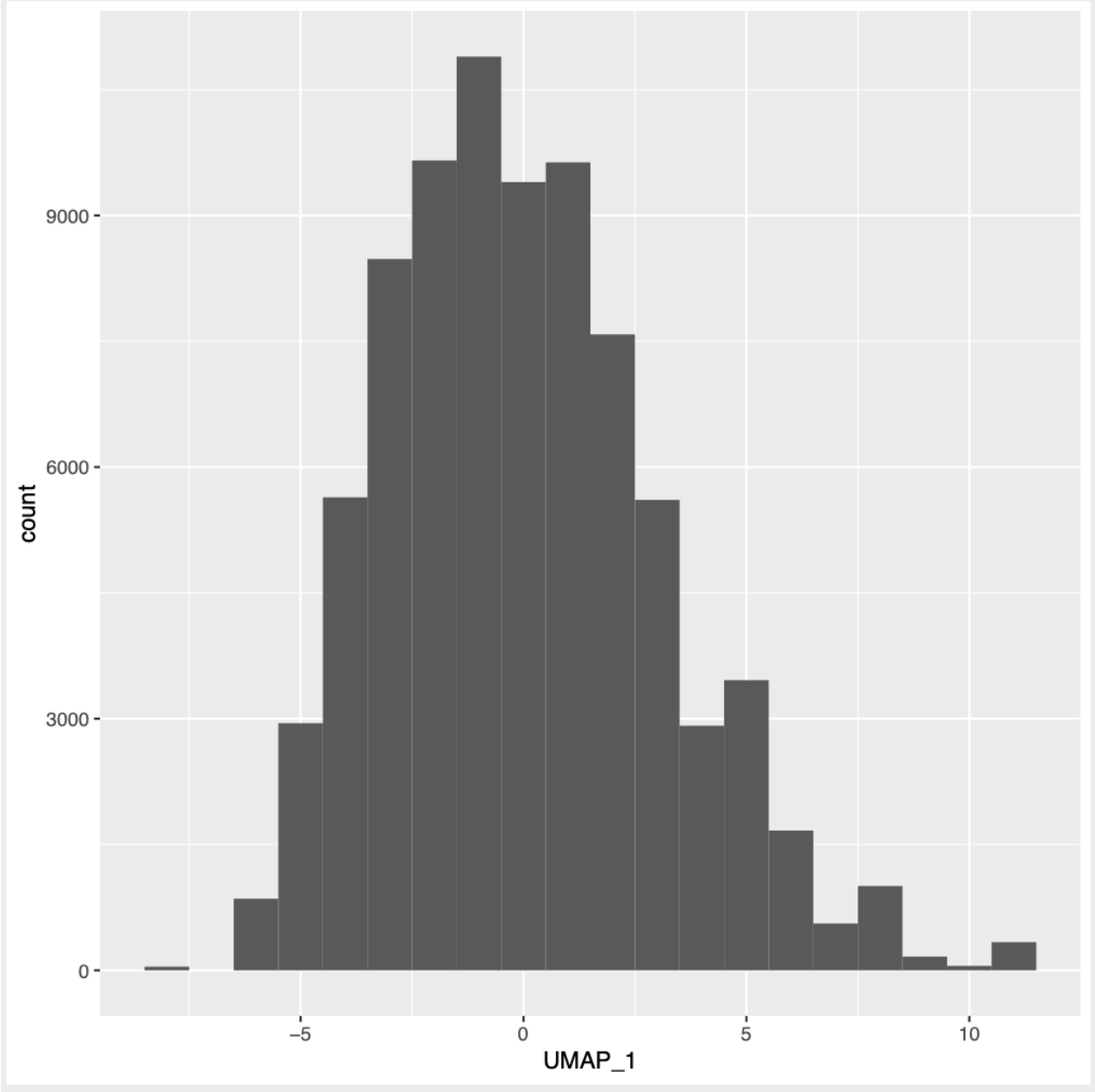
- Installed SingleR
- Modified previously written wrapper functions:
    - process_references , which does the data organization and processing of the Martin 2019 dataset and Luoma dataset with appropriate cluster labels
    - label_singleR , which is a wrapper function for adding SingleR labels with respect to the martin and luoma cluster definitions onto a seurat object.
- Note: the SingleR function takes more than an hour to execute. (I do not have an accurate time estimate.)



8-4-2020

Summary

Tasks are labeling and prelim analysis of clusters



Histogram to determine which cells to remove. The outlier cells (non-T cells) were removed by their UMAP coordinates.

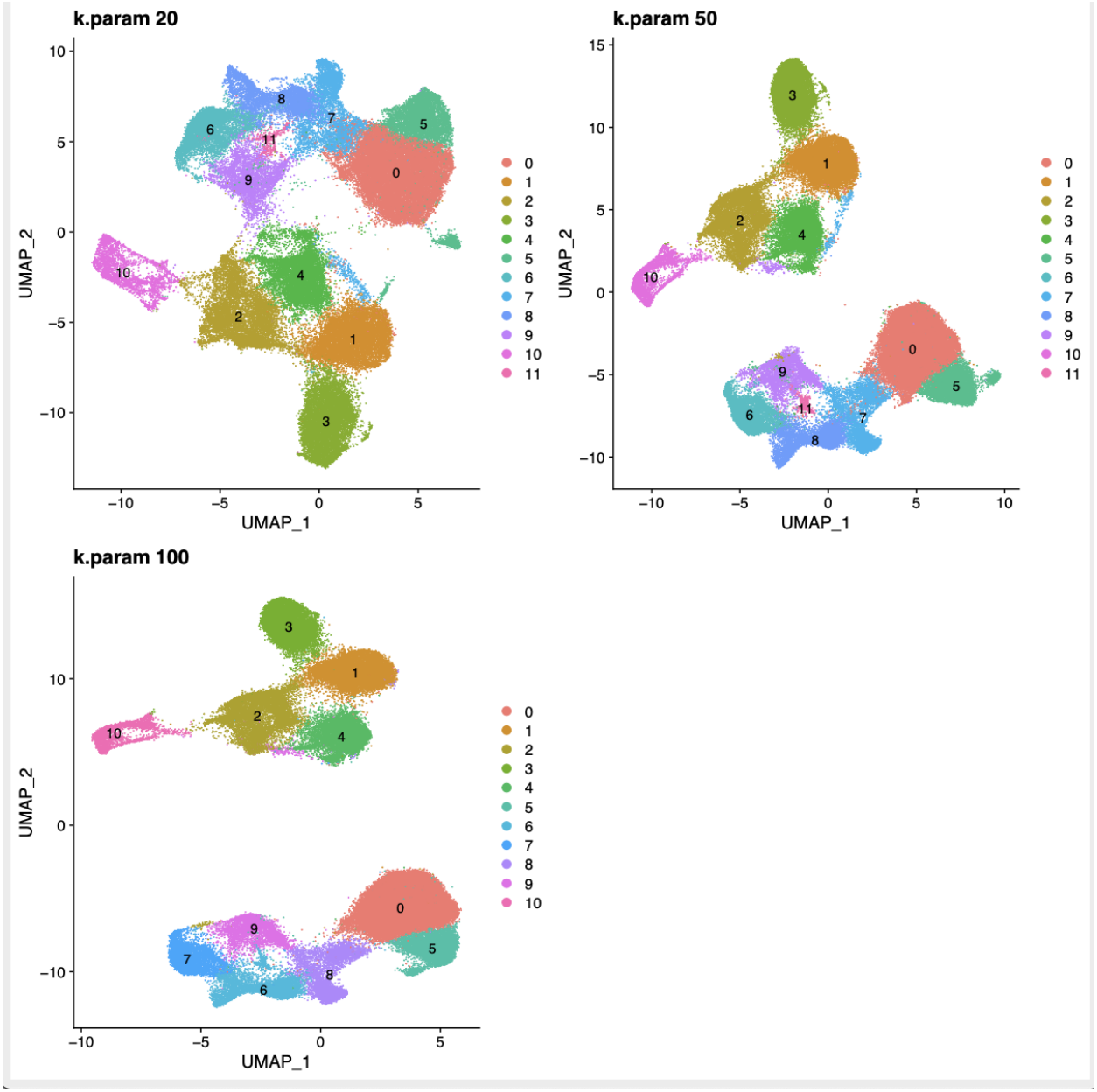
8-5-2020

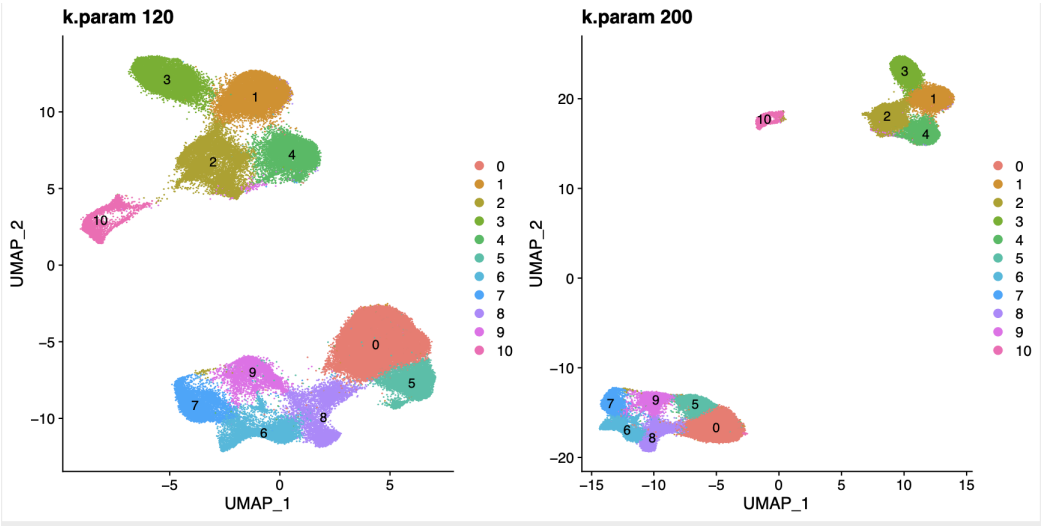
Summary

Tasks are labeling and prelim analysis of clusters

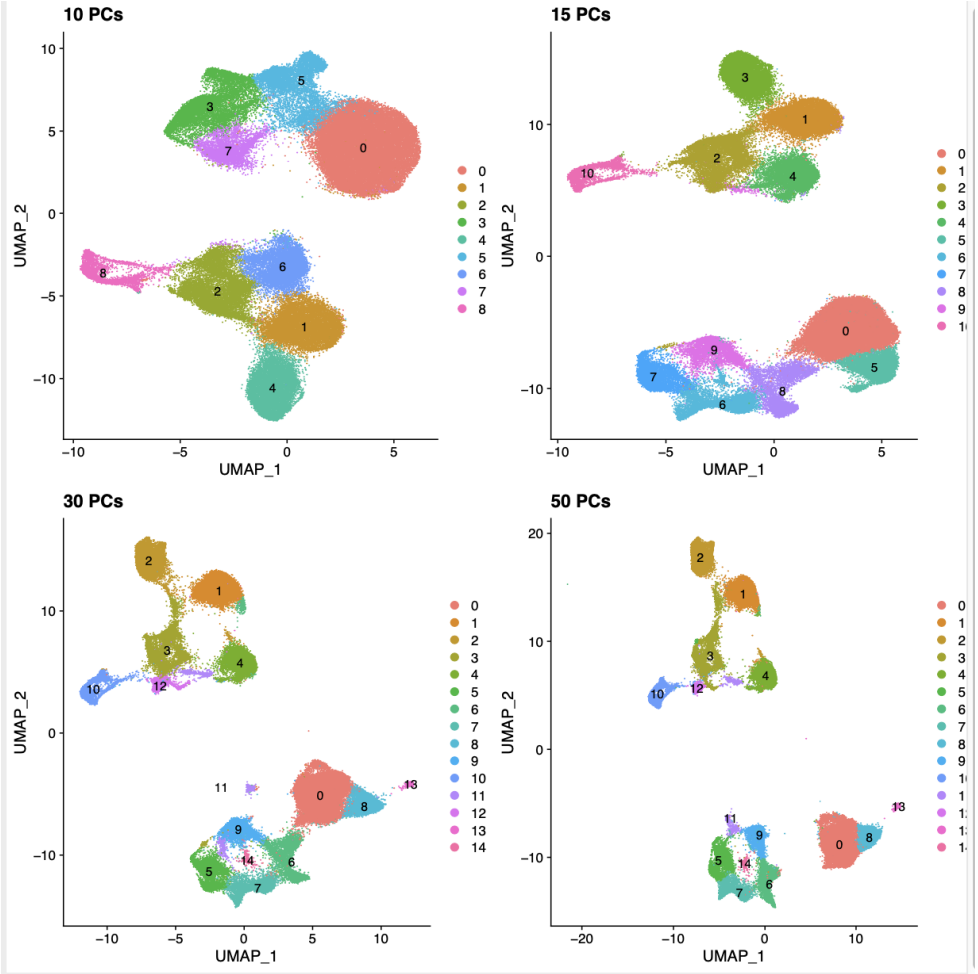
- Wrote the wrapper function: cluster_umap() which applied SNN+louvain graph-based clustering onto a scaled seurat object with PCA. Variables are number of PCA's to use, k.param for the nearest neighbors algorithm (?), resolution for cluster resolution, and min.dist for spacing of the clusters (smaller means tighter looking clusters).

Experiment 1: varying the k.param. The bottom right corner was supposed to be k.param=500, but the function ran out of memory. I repeated the experiment by increasing the memory limit and choosing 200 versus 500. The larger the k.param value, the longer the clustering takes.

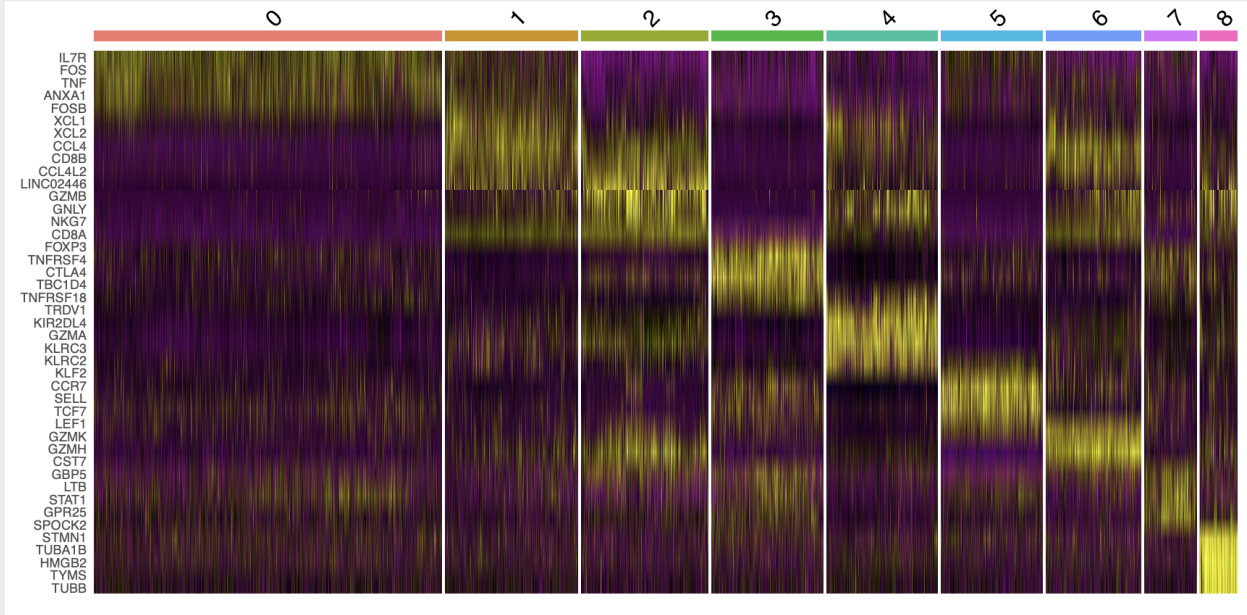




Experiment 2: varying the number of PCs. It seems 15 looks ok. 10 also looks ok. Since there are no good metrics (other than bootstrapping, which takes a long time to run) to evaluate whether a clustering is good, my eye tells me that 15 and 10 look better than 30 and 50.



Below is the heatmap of the differentially expressed genes per cluster in the final “onlyTcells” sample.



Can try GSA instead of GSEA: see Efron and Tibshirani

8-7-2020

Summary

Ran SingleR - did not finish executing. Took too long because it was assigning labels to each cell. Decided to change to clusters.

8-9-2020

Summary

Identified a bug in the Monaco reference singleR execution, so fixed that and left to run overnight.

8-12-2020

Summary

The singleR process was still running, and it was taking too long and using 0% CPU (probably because memory limit?), so I decided to remove the Monaco reference and try label_singleR once more. This is all recorded in codelog6. Eventually, I restarted the R session and tried again. I decided not to use the function wrapper that I had written for executing singleR (label_singleR), and it worked. The expected time for ~100K cells is 15 minutes for cluster mode and roughly 1 hour for single cell mode.

Setup Git with Remote Server

Credits to Jonathan Chu and stackoverflow: https://stackoverflow.com/questions/7152607/git-force-push-current-working-directory/7152800#7152800

1. On your remote server, create a directory for your files and initialize it as a git repository without the --bare option

```
$ mkdir [directory_name]
$ cd [directory_name]
$ git init
```

2. Configure the remote repository with receive.denyCurrentBranch ignore

```
$ git config --local receive.denyCurrentBranch updateInstead
```

3. In your personal computer (local git repo), add the remote git repo

```
$ git init # in the directory of interest
$ git remote add [some_name] ssh://user@host/path/to/remote/repo
# my case:
# git remote add rcapps5
ssh://esk17@rcapps5.dfci.harvard.edu/home/esk17/1.projects/colitis/scripts
```

4. Minor setting configurations (optional)

```
$ git config --global core.editor nano
```

```
$ git config --global --edit #edit the name and email
# then setup github if you have:
https://kbroman.org/github_tutorial/pages/first_time.html
```

Explanation: 1. The --bare option initializes without a working directory, i.e. it'll have the git info to keep track of history / version control, but not the actual directory of files you can use. --bare is usually used when the git repo is just there for version control
2. This configures the remote so that the files automatically update upon push. Should be no issue if you don't make any modifications to the files via the server side

Basic Git Workflow

```
$ git add [insert filename]
$ git commit -m [insert comment in quotes]
$ git push [name of remote] master
```

Key source: https://rogerdudler.github.io/git-guide/

You need an initial commit to start pushing
https://stackoverflow.com/questions/4181861/message-src-refspec-master-does-not-match-any-when-pushing-commits-in-git

To remove a remote, do:

```
$ git remote rm [name of remote]
```

Changing commit when there is a wrong name:
https://stackoverflow.com/questions/41349644/commit-with-wrong-username-to-github

Tutorial on GitSavvy package for Sublime Text:
https://mijingo.com/lessons/using-git-in-sublime-text/

Setup Rsync with Remote Server
I made a bash script (sync_figures.sh) with the following:

```
#!/bin/bash
#sync_figures.sh

rsync -avzh
esk17@rcapps5.dfci.harvard.edu:/home/esk17/1.projects/PD1/figures  ~/
```

Rsync Cheatsheet: https://devhints.io/rsync

8-13-2020

Summary

First, I chose to use the clustering settings of # of principle components as 15 and k.param for nearest neighbors as 50 because this resembled the previous analysis done by Shengbao the best. Also, this version is slightly more spaced out. The reason for using 50 vs 200 is that with 200, some of the subtle differences will not be captured by the clustering whereas lower k.param values will reveal more of the heterogeneity (also computationally faster). I named all the files done with this clustering “onlyTcells1550” or “onlyTcells_npc15_k50”.

I was able to successfully apply singleR (codelog6_labeling_only.R). I found that there is one portion of the cluster at the bottom right that should be split into two clusters. The island of cells should be MAIT cells but it is clumped in with another group of cells that mask it.

FeaturePlots are not yet informative. However, other inference measures and basic descriptive analysis revealed that the two PD-1 colitis positive samples are very different (codelog7). We might need a larger sample size to get a better understanding of the full picture.

8-14-2020

Summary

I implemented a Chi-squared goodness of fit test (codelog7.R). It showed that both no-colitis vs control and colitis vs control have significantly different distributions. However, this is to be expected that they are slightly different. The goodness of fit test seems to be too sensitive, especially since the sample size is high, so the confidence that the distributions are different (albeit slightly) increase.

Instead, I utilized the differential abundance (DA) analysis suggested by the “Orchestrating scRNA-seq analysis using Bioconductor” textbook. The DA analysis uses the “edgeR” packaged and a QM fitness model that I have yet to fully understand. However, its performance seems to align with the previously done rank-based hypothesis testing.

8-19-2020

Summary

I reorganized the projects directory. I combined the “martin” and “PD1” directories. I merged the data, saved objects and figures. I renamed this directory “colitis” By doing so, I had to update the git version control system and the sync_figures.sh script.

In addition, I also began the re-analysis of the Martin Crohns disease dataset. I took the post-qc, merged sample (s2.rds) and I re-clustered and used SingleR for labeling.

I ran the filter_stats() function to get some information but did not actually apply the filtering because I had previously already removed high percent.mt reads and redoing the percent.mt

(%mitochondrial genes) threshold calculation based on the empirical model would falsely remove more cells. However, I did remove some cells with abnormally large number of features. I did this based on 3MADs above the MEDIAN of the natural log of the number of features empirical distribution. I cut off cells with # of features greater than 2177. As an aside, I had wanted to do a more flexible gating because looking at the LNFxMITO plot, the LNF (log number of features) cutoff should vary with the %mito reads. However, I could not find a good way to do this, and the number of cells removed was very small (869 cells) compared to the total number of cells.

I did this to identify the CD3+ clusters. However, something with the normalization results in several ambiguous clusters. I decided to redo the clustering and labeling without regressing across the patient ids. This makes much more sense because I am not sure how the regression would actually work with categorical variables. I left this running overnight.

8-20-2020

Summary

Codelog8: The clustering was much more successful without regressing out by “patient.id” and only regressing out by “percent.mt.” I chose 15 principle components and k.param to be 100. The umap looks good. I used both cell annotations (Martin and Monaco as reference) and featureplots of T cell genes to identify the T cell cluster. I excluded the ILCs because they had low CD3 count.

Codelog9: I used the martin cluster and attempted a merge (no batch effect correction) with the CPI T cells. Surely enough, I observed batch effects. After the merge, I did no quality control filtrations.

I downloaded the UC colitis dataset from the link below and moved it to the data directory:
https://singlecell.broadinstitute.org/single_cell/study/SCP259/intra-and-inter-cellular-rewiring-of-the-human-colon-during-ulcerative-colitis#study-download

On Friday, I will try the anchor-based integration procedure in Seurat v3, following the guide here: https://satijalab.org/seurat/v3.2/integration.html

8-22-2020

Summary

Codelog9_anchors: I integrated the Crohns disease (Martin et al. 2019) and the CPI colitis CD3 T cells using the reference-based method with log-normalization as shown in the Satija lab tutorial. (https://satijalab.org/seurat/v3.2/integration.html)

The PCA figure is interesting: cluster 11 seems to be associated with a large spread.
The UMAP looks much more promising. I ran the cluster labeling annotations overnight.

A nice workshop for clustering:
https://nbisweden.github.io/workshop-scRNAseq/labs/compiled/seurat/seurat_04_clustering.html

8-23-2020

Summary

Codelog9_anchors: The batch effects were removed using the anchor-based integration method. The cluster labeling also looks promisingly consistent. Also proportionality bar graphs show that the

Codelog10: QC pipeline on the cells:

```
> sum(s$cells.to.remove==T)
[1] 10049
> sum(s$cells.to.remove==F)
[1] 200565
```

^ removed 5% of cells - all based on high mitochondrial reads.

8-24-2020

Summary

Codelog10: I integrated the UC dataset (downloaded previously) with the CD and CPI using anchors. The results do not look as great as the CD integration to CPI. The reason is probably because the UC dataset was not filtered properly and had some noticeable batcheffects that were not corrected (at least, this was mentioned in the paper). The batches were per patient and this information is available, so I will incorporate that.

tBET metric for comparison

```
CD_martin  CD3_Tcell UC_smillie
    38604      80459      75445
```

Checking storage usage and free space in my home directory

```
(base) esk17@rcapps5:~$ du -sh -- *
314G    0.raw.data
660G    1.projects
88K     2.code
13G     anaconda3
91G     cellranger
4.1G    miniconda3
255M    OpenBLAS
1.0K    R
1.1G    R-3.6.1
102K    rstudio-server-conda
2.2M    scrublet
758K    tree-1.8.0
2.9M    uuid-1.6.2
388K    uuid-1.6.2.tar.gz
```

```
(base) esk17@rcapps5:~$ df -h .
Filesystem                      Size  Used Avail Use% Mounted on
172.24.224.189:/mnt/storage/home  786T  501T  286T  64% /mnt/beegfs/home
```

Check QC-stats

By now you should know how to plot different features onto your data. Take the QC metrics that were calculated in the first exercise, that should be stored in your data object, and plot it onto your UMAP and as violin plots per cluster using the clustering method of your choice. For example, plot number of UMIS, detected genes, percent mitochondrial reads. Then, check carefully if there is any bias in how your data is separated due to quality metrics. Could it be explained biologically, or could you have technical bias there?

8-25-2020

Summary

Codelog10_redo_wo_mast_cells: I removed the mast cells, which I had accidentally included from the Smillie dataset. However, the clustering was still not much improved. Specifically, the IL1 and IL17 Trms cells were fused as a cluster and several low quality clusters were observed.

Codelog11: I implemented DE_heatmap(), which calculates the average gene expression per cluster and visualizes the avg. expression of the top 5 differentially expressed genes.

8-26-2020

Summary

Today, I redid the clustering and T cell selection from the Smillie dataset.

8-27-2020

Summary

I integrated the T cells with the

The issue is that there are some genes with cells that have NA for gene expression. This causes the entire variance to have NA.

Also attempted the following:
https://github.com/satijalab/seurat/issues/1698

8-28-2020

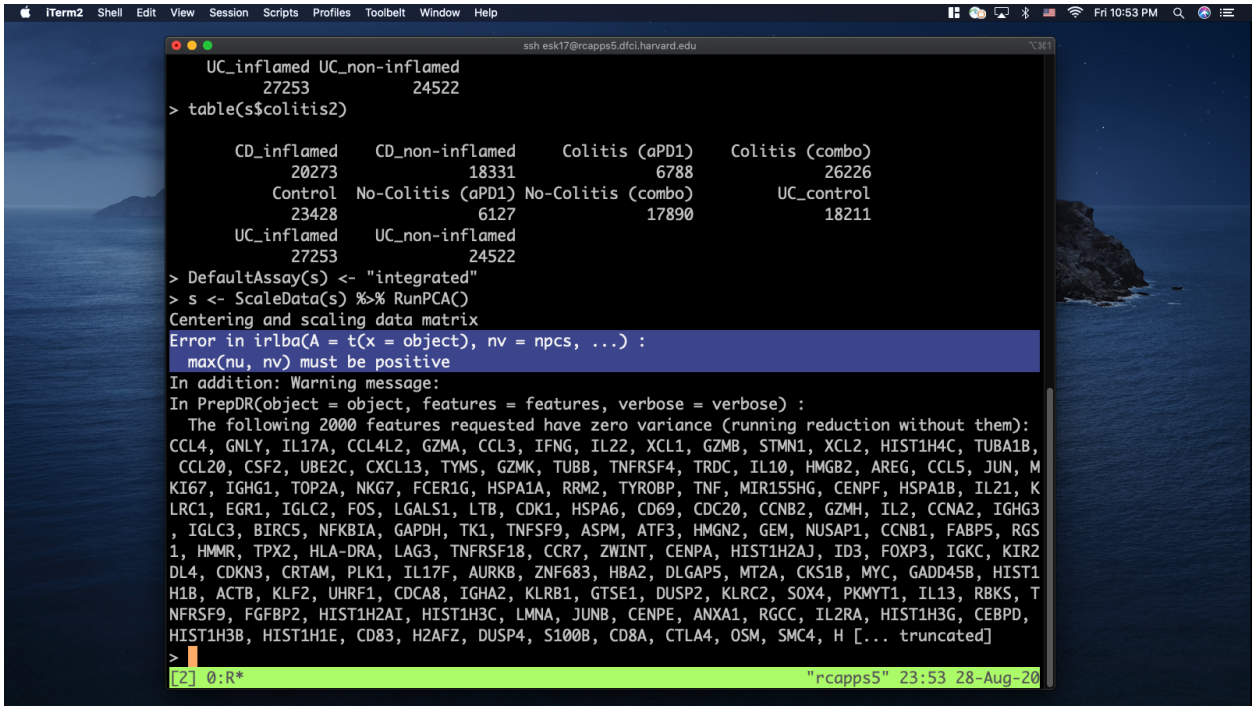
Summary

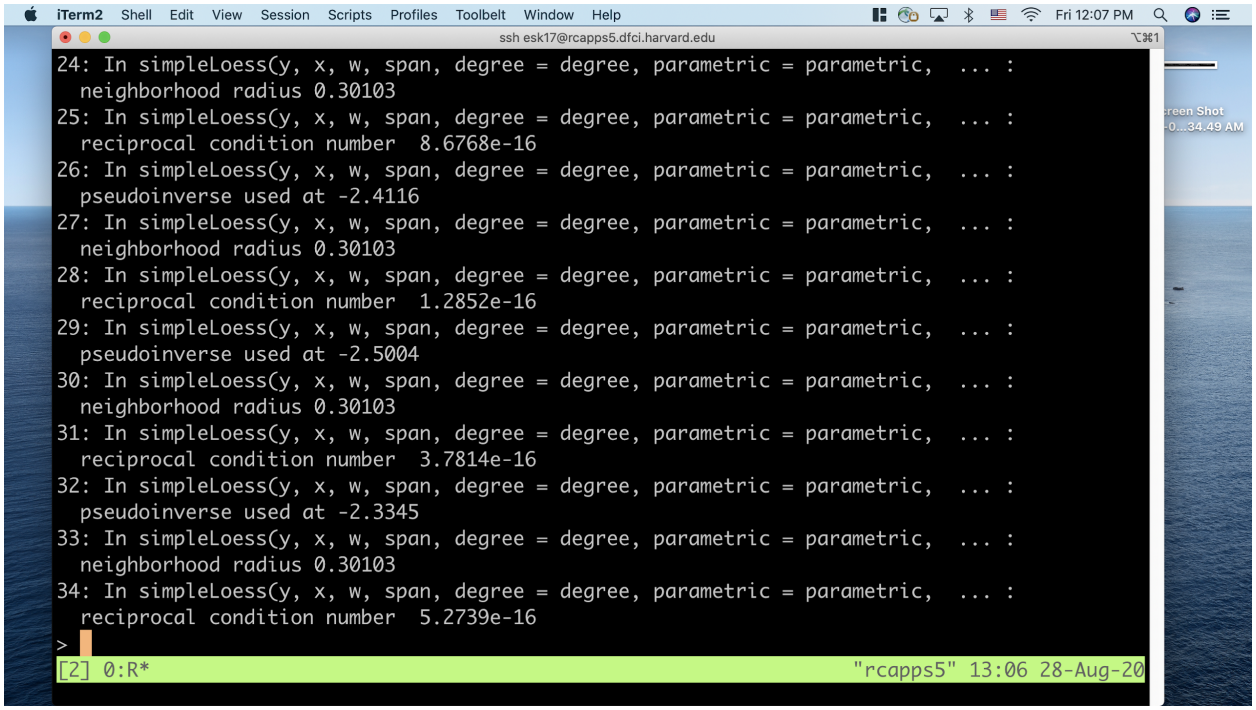
Type these commands in this exact order:
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge

$ conda install -c bioconda bioconductor-sva
conda update -n base conda
https://alexanderlabwhoi.github.io/post/anaconda-r-sarah/

Spent the day troubleshooting the CD+CPI+UC integration.

```
Error in nn2(data = c(-2.24030283596717, -2.24030283596717, -2.71147518059928,  :
  Cannot find more nearest neighbours than there are points
```





8-30-2020

Summary

codelog11_CDCPI_anchor_redo: Reintegrated CD+CPI after renormalizing, scaling, and clustering the martin CD3 T cells, which in the previous version was left at the CD45+ normalization and clustering stage. I called this new integration version cdcpi2.

9-17-2020

Summary

https://resources.aertslab.org/cistarget/help.html

wget https://resources.aertslab.org/cistarget/zsync_curl
chmod a+x zsync_curl
ZSYNC_CURL="${PWD}/zsync_curl"
echo "${ZSYNC_CURL}"

Error:
libcurl: Problem with the SSL CA cert (path? access rights?)
could not read control file from URL
https://resources.aertslab.org/cistarget/databases/homo_sapiens/hg19/refseq_r45/mc9nr/gene_based/hg19-500bp-upstream-7species.mc9nr.feather.zsync

Just downloaded using wget:
feather_database="${feather_database_url##*/}"
wget "${feather_database_url}"

9-18-2020

conda install -c bioconda nextflow

9-21-2020

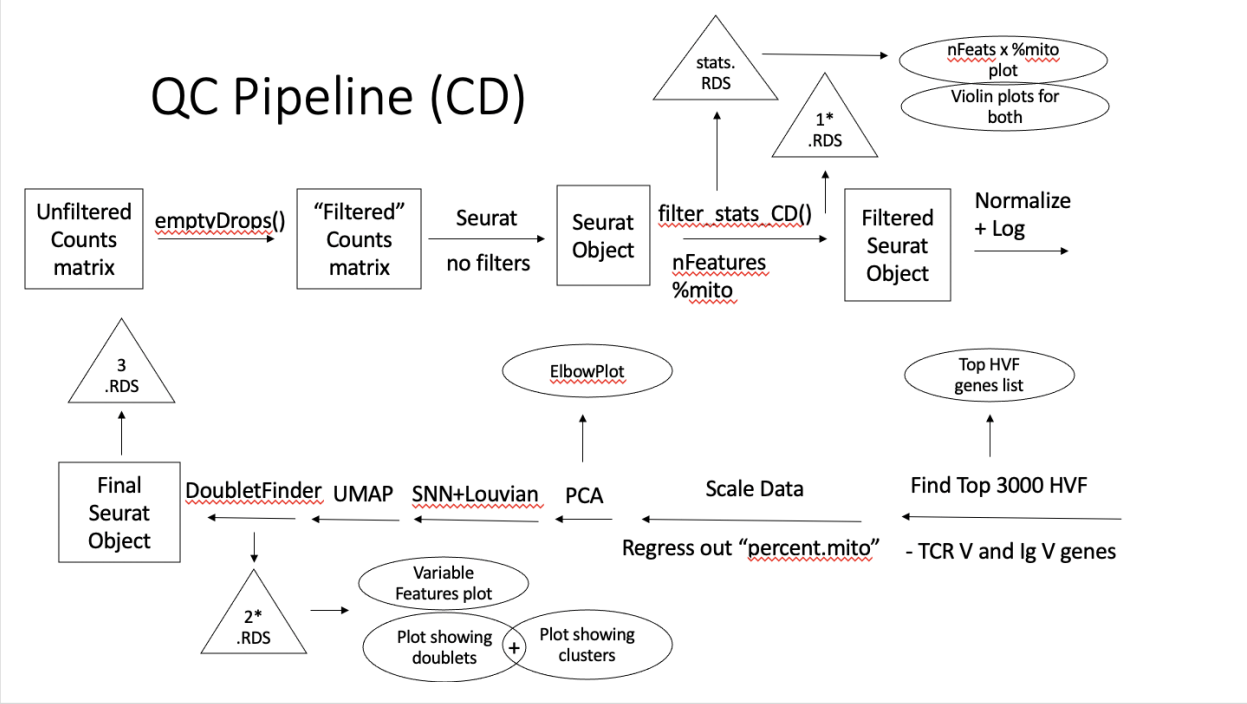
I redid the martin with the workflow used from

10-6-2020

Discrepancy in code
#33694 features across 737280
33694 features across 737268 (merged)

4334 cells that is a cell
403 cells to remove
3931
I had a bug in the code and my 1star.rds files include fewer cells than should be.

Fixed code to remove 1star



	Orig. count	is.cell fraction	is.cell counts	Removed (Mito & Lnf)	Mito upper thres	Feats lower thres	Doublet hi	Doublet lo	Singlet	Final count
GSM3972009_69	737280	5.88e-03	4334	403	5.80	38	335	58	3538	3596
GSM3972010_68	737280	2.00e-02	14710	1598	8.01	0	1101	210	11801	12011
GSM3972011_122	737280	2.52e-03	1861	17	73.10	0	158	26	1660	1686
GSM3972012_123	737280	4.94e-03	3645	0	Inf	0	296	68	3281	3349
GSM3972013_128	737280	5.34e-03	3937	562	11.47	0	295	43	3037	3080
GSM3972014_129	737280	4.59e-03	3382	936	11.94	105	187	58	2201	2259
GSM3972015_135	737280	7.36e-03	5429	1551	18.63	64	316	72	3490	3562
GSM3972016_138	737280	7.03e-03	5180	728	7.90	130	353	92	4007	4099
GSM3972017_158	737280	8.38e-03	6175	1328	24.46	0	414	71	4362	4433
GSM3972018_159	737280	9.04e-03	6667	784	42.38	0	500	88	5295	5383
GSM3972019_180	737280	5.52e-03	4070	903	16.41	0	271	46	2850	2896
GSM3972020_181	737280	1.60e-02	11811	868	8.03	121	884	210	9849	10059
GSM3972021_186	737280	1.06e-02	7835	1097	9.57	0	570	104	6064	6168
GSM3972022_187	737280	5.06e-03	3731	161	6.72	0	288	69	3213	3282
GSM3972023_189	737280	9.15e-03	6749	997	16.91	0	483	92	5177	5269
GSM3972024_190	737280	8.54e-03	6294	1298	13.37	159	439	61	4496	4557
GSM3972025_192	737280	1.23e-02	9090	1182	17.21	135	649	142	7117	7259
GSM3972026_193	737280	1.52e-02	11217	1562	12.23	118	803	163	8689	8852
GSM3972027_195	737280	7.94e-03	5856	0	Inf	0	521	65	5270	5335
GSM3972028_196	737280	5.63e-03	4152	716	28.41	0	303	41	3092	3133
GSM3972029_208	737280	6.11e-03	4503	549	17.34	0	333	62	3559	3621
GSM3972030_209	737280	8.47e-03	6248	638	20.05	0	489	72	5049	5121

Noticeably the lower threshold for feature count is 0 for many of the samples. Also the proportion of non-empty cells is very low (around 0.001) but varies greatly with some samples having two times the number of cells.
Doing some manual investigation, I found that Mito is not cut off for samples 123 and 195 and is high for 122. This is not a good sign. 208 is also bimodal

The density curves for the LNF for 129 and 138 are bimodal and have a much smaller peak before.

181 looks good
192 not too bad

195 seems to have many dead cells, which is visible from the LNF vs MITO

Conclusion is that I will need manual cleaning of each of the samples.
New strategy: min(20, fraction) ----> remove dead cells and from existing cells determine LNF lower bound with max(log(200),lnf.lim).

10-7-2020

Removed bad clusters individually

	Orig. count	is.cell fraction	is.cell counts	Removed (Mito & Lnf)	Mito upper thres	Feats lower thres	Doublet hi	Doublet lo	Singlet	Final count
GSM3972009_69	737280	5.88e-03	4334	727	5.80	199	305	56	3246	3302
GSM3972010_68	737280	2.00e-02	14710	3621	8.01	199	897	212	9980	10192
GSM3972011_122	737280	2.52e-03	1861	879	20.00	199	81	17	884	901
GSM3972012_123	737280	4.94e-03	3645	2518	20.00	199	91	22	1014	1036
GSM3972013_128	737280	5.34e-03	3937	911	11.47	199	265	38	2723	2761
GSM3972014_129	737280	4.59e-03	3382	1280	11.94	293	148	62	1892	1954
GSM3972015_135	737280	7.36e-03	5429	2041	18.63	199	276	63	3049	3112
GSM3972016_138	737280	7.03e-03	5180	879	7.90	199	348	82	3871	3953
GSM3972017_158	737280	8.38e-03	6175	1987	20.00	199	359	60	3769	3829
GSM3972018_159	737280	9.04e-03	6667	2748	20.00	199	331	61	3527	3588
GSM3972019_180	737280	5.52e-03	4070	1097	16.41	199	262	35	2676	2711
GSM3972020_181	737280	1.60e-02	11811	1116	8.03	199	862	208	9625	9833
GSM3972021_186	737280	1.06e-02	7835	1800	9.57	199	526	78	5431	5509
GSM3972022_187	737280	5.06e-03	3731	795	6.72	199	246	48	2642	2690
GSM3972023_189	737280	9.15e-03	6749	1459	16.91	199	428	101	4761	4862
GSM3972024_190	737280	8.54e-03	6294	1486	13.37	236	419	62	4327	4389
GSM3972025_192	737280	1.23e-02	9090	1331	17.21	199	637	139	6983	7122
GSM3972026_193	737280	1.52e-02	11217	1813	12.23	199	777	163	8464	8627
GSM3972027_195	737280	7.94e-03	5856	3319	20.00	199	219	35	2283	2318
GSM3972028_196	737280	5.63e-03	4152	1229	20.00	199	253	39	2631	2670
GSM3972029_208	737280	6.11e-03	4503	860	17.34	199	302	62	3279	3341
GSM3972030_209	737280	8.47e-03	6248	1301	20.00	199	428	67	4452	4519

Final cell counts after removing clusters with high proportions of doublets:

```
 [1] 2910 9319  901  973 2648 1774 2917 3709 3776 3319 2489 8664 5244 2470 4123
[16] 4139 6840 8282 2073 2539 3097 4102
```

10-12-20

Today, I subsetted the tcells from the new martin integrated data (all of the relevant figures are in “saved_objects/CD_marin_qc_100720” and “figures/CD_martin_100720”). I subsetted based on CD3 expression. Then the resulting cells were clustered. The best clustering seemed to happen with 15 PCs and kparam set to 100; I called this clustering “martin_naive_CD3_15100” where naive indicates that I did not apply integration and only regressed based on percent mito and nFeature RNA. The resulting clusters were integrated into the CPI data with no reference dataset. I labeled this integration version as cdcpi3 (or cdcpi3_anchor). Note that unlike the previous integrated versions, there is no selected reference. Surprisingly, even without selecting the reference, the cpi dataset had much better mappings than the previous.

It appears that all clusters have CD3, which is a good sign. It seems that cluster 8 has B cell contaminants however.

10-17-20

Agenda:

- Finding if Ig genes are in the DE list in the CD Cell paper
- Composition by dataset to find a suitable control population and use it as a metric for control
- Colitis visualization
- Mono vs combo summaries

Git clone https://github.com/effiken/martin_et_al_cell_2019.git
conda install -c conda-forge r-devtools

10/31/2020
Warning: cdcpi2_CD_inf_vs_unif DA analysis figures are corrupt (I overrode them with martin_naive_CD315100_CD_inf_vs_uninf information).

11-8-20

Tasks:

- Finished the metrics comparing harmony and seurat. The results suggest that both are similar but the anchor-based has slightly higher resolution.

11-15-20

Tasks:

- Fixed cluster comparison figure. Found the error in the row matching

11-30-20

Transfered the IBD sample and processed them:

```
$ cellranger count --id=p1089neg-GEX-Pool7 \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data/IBD \
                  --sample=p1089neg-GEX-Pool7 \
                  --localcores=12 \
                  --localmem=64
```

Edited

```
$ cellranger count --id=p1089pos-GEX-Pool6 \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                    --fastqs=/home/esk17/0.raw.data/IBD \
                    --sample=p1089pos-GEX-Pool6 \
                    --localcores=12 \
                    --localmem=64
```

12-7-20

Tasks:

- Received storage on a new server: esk17@rcgpus.dfci.harvard.edu

```
(base) esk17@rcGPUs:~$ du -sh -- *
431G    0.raw.data
908G    1.projects
89K     2.code
13G     anaconda3
30M     CellBender
91G     cellranger
14G     miniconda3
16K     nextflow
255M    OpenBLAS
1.0K    R
1.1G    R-3.6.1
102K    rstudio-server-conda
2.2M    scrublet
758K    tree-1.8.0
2.9M    uuid-1.6.2
388K    uuid-1.6.2.tar.gz
```

- Getting error running cellbender:

```
/home/esk17/miniconda3/envs/cellbender/lib/python3.7/site-packages/torch/cu
da/__init__.py:52: UserWarning: CUDA initialization: The NVIDIA driver on
your system is too old (found version 10010). Please update your GPU driver
by downloading and installing a new version from the URL:
http://www.nvidia.com/Download/index.aspx Alternatively, go to:
https://pytorch.org to install a PyTorch version that has been compiled
with your version of the CUDA driver. (Triggered internally at
/pytorch/c10/cuda/CUDAFunctions.cpp:100.)
  return torch._C._cuda_getDeviceCount() > 0
Traceback (most recent call last):
  File "/home/esk17/miniconda3/envs/cellbender/bin/cellbender", line 33, in
<module>
      sys.exit(load_entry_point('cellbender', 'console_scripts',
```

```
'cellbender')())
  File "/mnt/beegfs/home/esk17/CellBender/cellbender/base_cli.py", line 98,
in main
      args = cli_dict[args.tool].validate_args(args)
  File
"/mnt/beegfs/home/esk17/CellBender/cellbender/remove_background/cli.py",
line 69, in validate_args
      assert torch.cuda.is_available(), "Trying to use CUDA, " \
AssertionError: Trying to use CUDA, but CUDA is not available.
```

GPU information:

```
Tue Dec  8 00:58:10 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 418.67       Driver Version: 418.67       CUDA Version: 10.1     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           Off  | 00000000:04:00.0 Off |                    0 |
| N/A   38C    P0    59W / 149W |  11032MiB / 11441MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla K80           Off  | 00000000:05:00.0 Off |                    0 |
| N/A   28C    P0    74W / 149W |      0MiB / 11441MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla K80           Off  | 00000000:83:00.0 Off |                    0 |
| N/A   34C    P0    59W / 149W |      0MiB / 11441MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla K80           Off  | 00000000:84:00.0 Off |                    0 |
| N/A   28C    P0    72W / 149W |      0MiB / 11441MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   4  Tesla K80           Off  | 00000000:89:00.0 Off |                    0 |
| N/A   28C    P0    57W / 149W |      0MiB / 11441MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   5  Tesla K80           Off  | 00000000:8A:00.0 Off |                    0 |
| N/A   37C    P0    73W / 149W |      0MiB / 11441MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   6  Tesla K80           Off  | 00000000:8D:00.0 Off |                    0 |
| N/A   30C    P0    66W / 149W |      0MiB / 11441MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   7  Tesla K80           Off  | 00000000:8E:00.0 Off |                    0 |
| N/A   43C    P0    86W / 149W |      0MiB / 11441MiB |     64%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0     26444      C   ...da3/envs/iscore_deepscore/bin/python3.7 11019MiB |
+-----------------------------------------------------------------------------+
```

FIxed error: Fresh installation

New runtime error:

```
RuntimeError: CUDA out of memory. Tried to allocate 24.00 MiB (GPU 0; 11.17
GiB total capacity; 84.30 MiB already allocated; 1.56 MiB free; 98.00 MiB
reserved in total by PyTorch)
```

Fixed by choosing a different GPU (the default GPU 0 was being used by another person).

```
CUDA_VISIBLE_DEVICES=3 cellbender remove-background \
                  --input
/home/esk17/1.projects/colitis/data/CD_martin/GSM3972009_69/ \
                  --output
/home/esk17/1.projects/colitis/data/CD_martin_cellbender/GSM3972009_69cb.h5
\
                  --cuda \
                  --expected-cells 5000 \
                  --total-droplets-included 7000 \
                  --fpr 0.01 \
                  --epochs 150
```

```
$ cellranger count --id=p1089pos-TCR-Pool6 \
   --transcriptome=/home/esk17/cellranger/refdata-cellranger-GRCh38-3.0.0 \
                  --fastqs=/home/esk17/0.raw.data/IBD \
                  --sample=p1089pos-TCR-Pool6 \
                  --localcores=12 \
                  --localmem=64
```

1-25-21

Tasks:

- Created custom bash commands

1. First create the invisible file to hold the custom commands

```
$ touch ~/.custom_bash_commands.sh
```

2. Added the following in the script file using a text editor (nano):

```
#!/bin/bash

# sync figures from remote to local
function sync_figures() {
      rsync -avzh
esk17@rcapps5.dfci.harvard.edu:/home/esk17/1.projects/colitis/figures ~/
}

# transfer files from local to remote
function transfer_files() {
      rsync -apvzh ~/transfer/
esk17@rcapps5.dfci.harvard.edu:/home/esk17/1.projects/colitis/transfer
}

function lazygit_ncf() {
      git add new_clustering_functions.R
      git commit -m "$1"
      git push
      git push rcapps5 master
}
```

3. Added the following to the end of the .zshrc file:

```
source ~/.custom_bash_commands.sh #custom commands for colitis project
1/25/2021
```

Tutorial link:
https://shanelonergan.github.io/streamline-your-workflow-with-custom-bash-commands/

1-26-21

Tasks:

- Fix qc_CD, save_figures_CD, filter_stats functions. New features were added such as custom QC thresholding and improved plots.
- Clustered both the p1089neg and p1089pos (codelog17)
    - Used a manual threshold for min Log number of features for p1089pos sample

1-26-21

Tasks:

- Update to Seurat 4

Backup conda environment:
$ conda activate r_env2
$ conda env export > environment_r_env2.yml
$ conda list --explicit > spec_file_r_env2.txt

$ conda create --name seurat4 --clone r_env2
https://stackoverflow.com/questions/40700039/how-can-you-clone-a-conda-environment-into-the-root-environment

Conda activate seurat4

- Mapped p1089pos to CD+CPI anchor dataset as reference
    - Anchors took quite a long time to calculate ~1.5 hours (started at 11pm and ended at 12:30am)
    - Need to map again because the mapping was done with all CD45+ cells and not just T cells.

1-30-21

Tasks:

- Clustered UC Chang pbmc scRNA-seq data; merged without batch correction. Not an issue for subsetting T cells; however, some batch effects seem evident.
- Subset the T cells

1-31-21

Tasks:

- Integrated with CCA and anchors for UC Chang T cells
- Subset the T cells from MGH CD sample and mapped it onto the CD+CPI anchor dataset as reference

2-4-21

Tasks:

- Analyzed MGH CD sample with cluster proportion comparison. Wrote a function for it. Also performed DE per cluster. Found only B cell contaminating genes. This is strange because it was also observed for the Martin et al. CD dataset. Potentially this suggests that there are a significantly larger number of B cells in CD samples.
- Subset the T cells from MGH CD sample and mapped it onto the CD+CPI anchor dataset as reference

2-8-21

Tasks:

- Clustered and integrated the UC rectum samples (codelog18.1)
- DE

2-10-21

Tasks:

- DE gave some B cell markers
- Obtained DA and DE analysis for the UC rectum samples versus the CPI colitis.

3-8-21

Tasks:

- Fresh install of anaconda3
- Cellphonedb fails to work on the server.

```
Last login: Mon Mar  8 20:18:44 2021 from 10.251.21.107
esk17@rcGPUs:~$ export PATH="/home/sv467/anaconda3/env/cpdb/bin:$PATH"
esk17@rcGPUs:~$ cd 1.projects/colitis/cpdb/test/
esk17@rcGPUs:~/1.projects/colitis/cpdb/test$ cellphonedb method analysis test_meta.txt test_counts.t
xt
/home/esk17/.local/lib/python3.8/site-packages/sklearn/utils/deprecation.py:144: FutureWarning: The
sklearn.cluster.k_means_ module is  deprecated in version 0.22 and will be removed in version 0.24.
The corresponding classes / functions should instead be imported from sklearn.cluster. Anything that
 cannot be imported from sklearn.cluster is now part of the private API.
  warnings.warn(message, FutureWarning)
[ ][APP][08/03/21-20:25:17][WARNING] Latest local available version is `v2.0.0`, using it
[ ][APP][08/03/21-20:25:17][WARNING] User selected downloaded database `v2.0.0` is available, using
it
[ ][CORE][08/03/21-20:25:17][INFO] Initializing SqlAlchemy CellPhoneDB Core
[ ][CORE][08/03/21-20:25:17][INFO] Using custom database at /mnt/beegfs/home/esk17/.cpdb/releases/v2
.0.0/cellphone.db
[ ][APP][08/03/21-20:25:17][INFO] Launching Method cpdb_analysis_local_method_launcher
[ ][APP][08/03/21-20:25:17][INFO] Launching Method _set_paths
[ ][APP][08/03/21-20:25:17][INFO] Launching Method _load_meta_counts
[ ][CORE][08/03/21-20:25:17][INFO] Launching Method cpdb_method_analysis_launcher
[ ][CORE][08/03/21-20:25:17][INFO] Launching Method _counts_validations
```

- Reinstalled anaconda and also re-created the r-environment to work with Seurat and CellChat
    - conda install r-FNN r-igraph r-ggrepel r-RANN r-reticulate r-Rtsne r-RcppEigen r-catools r-RSpectra r-dqrng r-leiden r-gplots r-spatstat r-uwot r-Rocr r-future r-hdf5r r-gridExtra r-devtools r-systemfonts
    - conda install -c bioconda bioconductor-biocgenerics bioconductor-biobase bioconductor-complexheatmap

```
ln -s /home/esk17/anaconda3/envs/r_4.0.3/bin/x86_64-conda_cos6-linux-gnu-gcc
/home/esk17/anaconda3/envs/r_4.0.3/bin/gcc
ln -s /home/esk17/anaconda3/envs/r_4.0.3/bin/x86_64-conda_cos6-linux-gnu-g++
/home/esk17/anaconda3/envs/r_4.0.3/bin/g++
```

```
ln -s /home/esk17/anaconda3/envs/r_env2/bin/x86_64-conda_cos6-linux-gnu-gcc
/home/esk17/anaconda3/envs/r_env2/bin/gcc
ln -s /home/esk17/anaconda3/envs/r_env2/bin/x86_64-conda_cos6-linux-gnu-g++
/home/esk17/anaconda3/envs/r_env2/bin/g++
```

```
Error:
TypeError: simplicial_set_embedding() missing 3 required positional
arguments: 'densmap', 'densmap_kwds', and 'output_dens'
```