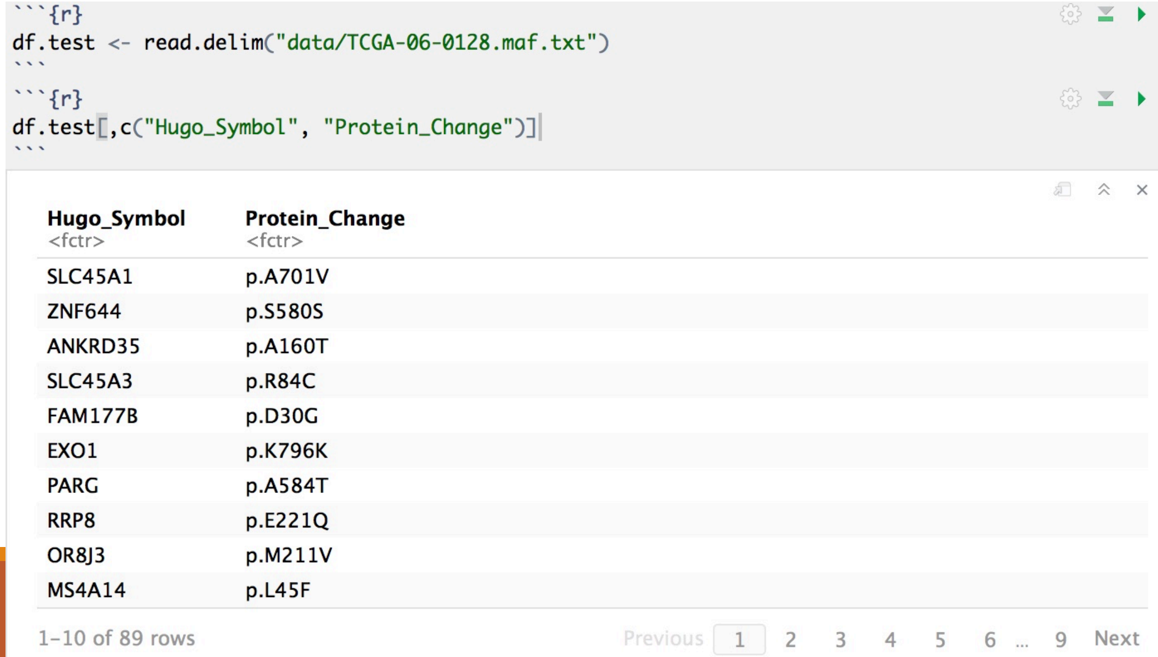
STAT115/BST282 Lab12

HW6 Questions overview

- Glioblastoma microarray data
- 2 subtypes identified by K-means clustering
- Differentially expressed genes
- Differentially methylated genes
- Survival analysis
- Cox model regression
- What next?

PartIII Q1 & Q2: Mutations

- Mutation files from samples of different GBM subtypes
- We are interested in the genes with mutations and the protein change in this mutation
- For each subtype, count the times a gene is mutated(Q1) & count the times a gene-protein change pair is mutated(Q2), the mutation that is specifically prevalent in one subtype might be the genetic factor that distinguishes the two subtypes



~~~
```{r}
df.test <- read.delim("data/TCGA-06-0128.maf.txt")
```

```{r}
df.test[,c("Hugo_Symbol", "Protein_Change")]
```
~~~

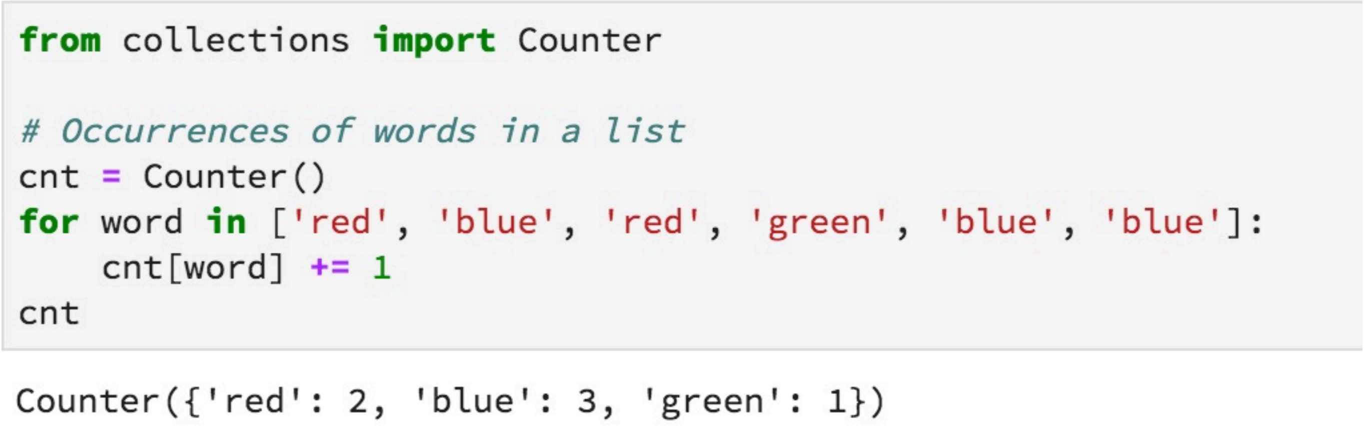
```
Hugo_Symbol    Protein_Change
<fctr>         <fctr>
SLC45A1        p.A701V
ZNF644         p.S580S
ANKRD35        p.A160T
SLC45A3        p.R84C
FAM177B        p.D30G
EXO1           p.K796K
PARG           p.A584T
RRP8           p.E221Q
OR8J3          p.M211V
MS4A14         p.L45F
1-10 of 89 rows                Previous  1  2  3  4  5  6 ...  9  Next
```

Gene Names	Count
…	…
…	…

Gene-Protein change pair	Count
…	…
…	…

PartIII Q1 & Q2: Tips

- Check out the Counter class from the collections module in Python



```
from collections import Counter

# Occurrences of words in a list
cnt = Counter()
for word in ['red', 'blue', 'red', 'green', 'blue', 'blue']:
    cnt[word] += 1
cnt
```

```
Counter({'red': 2, 'blue': 3, 'green': 1})
```

- You can do math things like add up two Counter objects
- Check out the documentation here: https://docs.python.org/2/library/collections.html

PartIII Q1 & Q2: Clarification

- Q1: Total counts combining two subtypes

Gene Names	Count
…	…
…	…

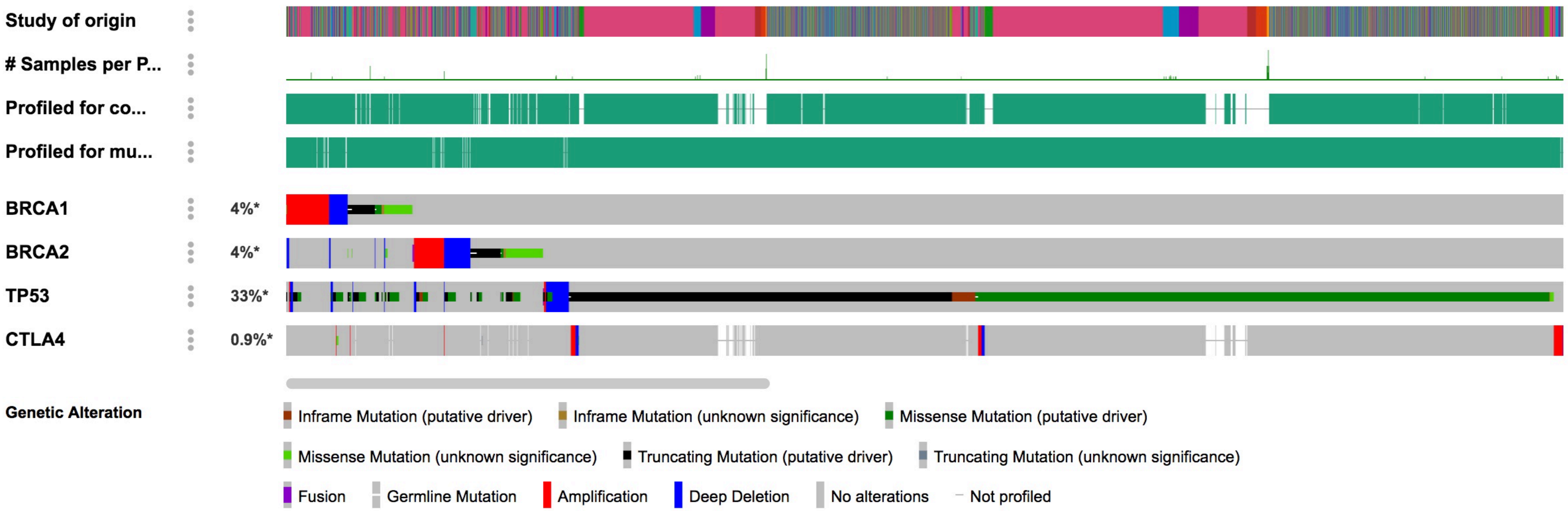
- Q2: For each subtype

Gene-Protein change pair	Count
…	…
…	…

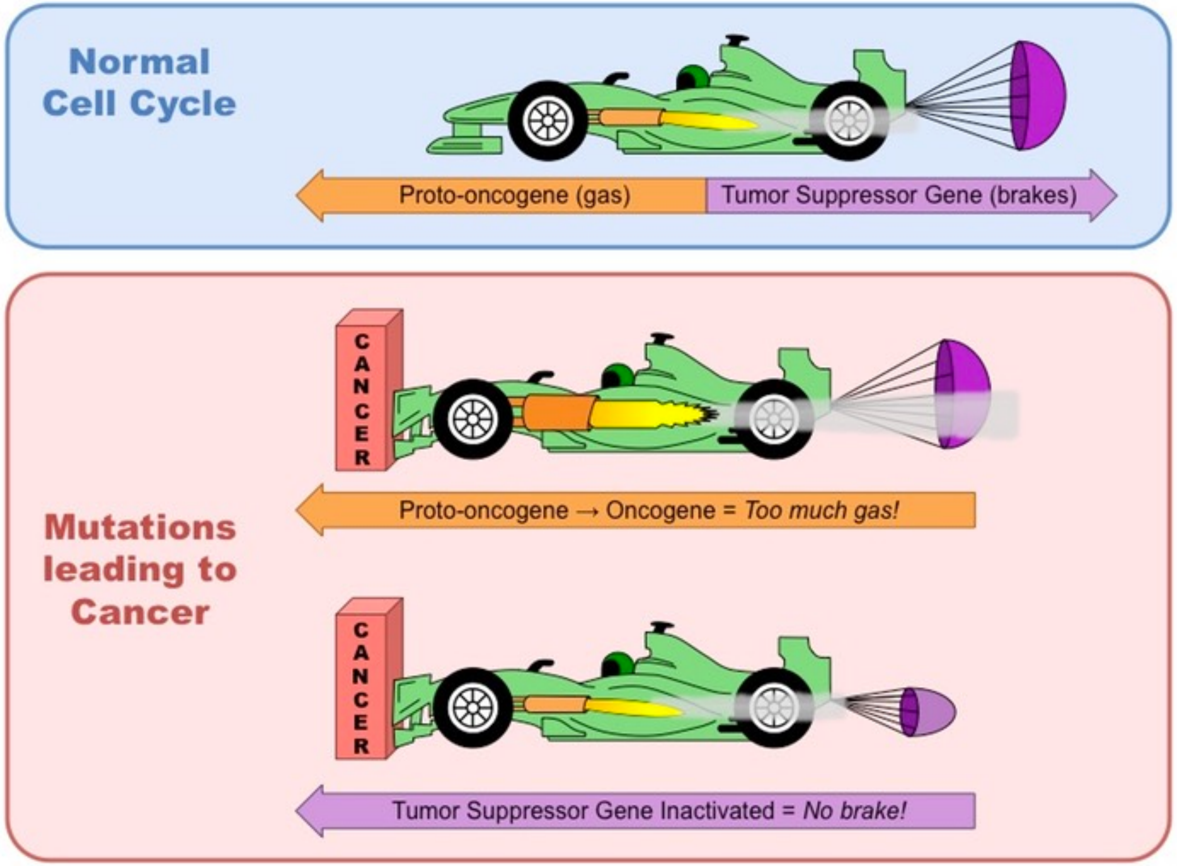
Gene-Protein change pair	Count
…	…
…	…

PartIII Q3 & Q4 & Q5

- http://www.cbioportal.org/
- Choose the studies you are interested in, submit the genes that you want to look into
- Here is a sample view of invasive breast carcinoma and 4 genes
- Gain of functions or loss of functions?


Study of origin
# Samples per P...
Profiled for co...
Profiled for mu...
BRCA1
4%*
BRCA2
4%*
TP53
33%*
CTLA4
0.9%*
Genetic Alteration
Inframe Mutation (putative driver)
Inframe Mutation (unknown significance)
Missense Mutation (putative driver)
Missense Mutation (unknown significance)
Truncating Mutation (putative driver)
Truncating Mutation (unknown significance)
Fusion
Germline Mutation
Amplification
Deep Deletion
No alterations
Not profiled

Oncogene & Tumor Suppressor Gene

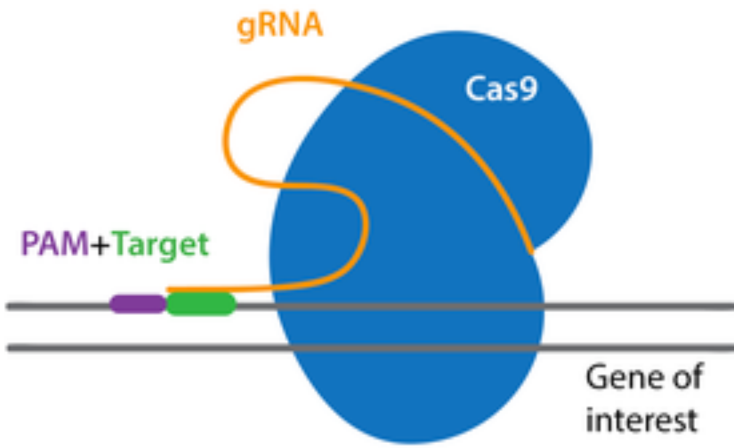

Normal Cell Cycle
Proto-oncogene (gas)
Tumor Suppressor Gene (brakes)
Mutations leading to Cancer
CANCER
Proto-oncogene → Oncogene = *Too much gas!*
CANCER
Tumor Suppressor Gene Inactivated = *No brake!*

PartIII Q6

- Clinical Trials for conditions
- https://www.clinicaltrials.gov

Part IV: CRISPR Screens

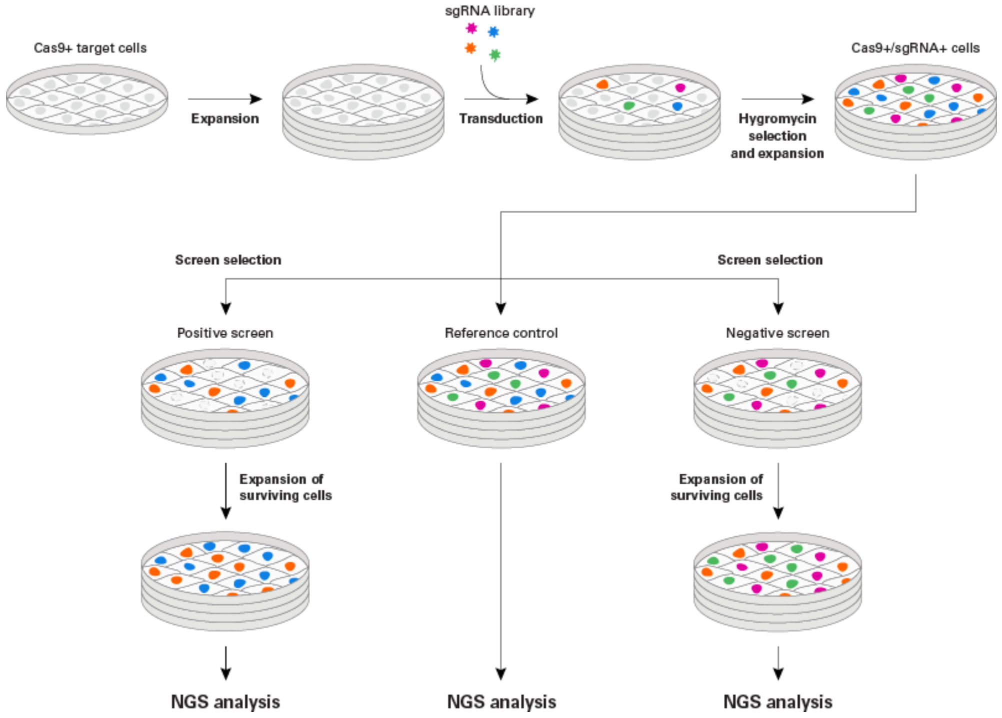
- CRISPR systems
  - Adaptive immune system in bacteria modified for genome engineering
  - Two components:
    - sgRNA
    - Cas protein
  - Can be used to perform gene knock-out



Genome Engineering
Transcriptional Regulation
Other Applications

Part IV: CRISPR Screens

- sgRNA library to transduce cells
- Genome-wide knock-outs(KOs)
- Apply positive or negative selection
  - Positive: look for genes that KOs make the cells survive
    - Cells expressing sgRNAs for these genes will be enriched
    - For finding drug resistance mechanism
  - Negative: look for cells that are lost due to Kos
    - Cells expressing sgRNAs for these genes will be lost
    - Cells expressing other sgRNAs overrepresented
    - For finding survival-essential genes

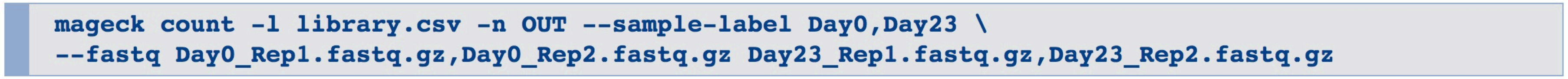


Part IV: CRISPR Screens

- Analyzing CRISPR Screen Data with MAGeCK
- **Installation** on Odyssey:
  - Copy the folder /n/stat115/2020/HW6/mageck-0.5.8 to your home directory
  - cp -r /n/stat115/2020/HW6/mageck-0.5.8 ~
  - cd ~/mageck-0.5.8
  - module load Anaconda/5.0.1-fasrc01
  - python setup.py install –user
  - test that the command works with mageck –help
- Data stored at /n/stat115/2020/HW6/crispr_data

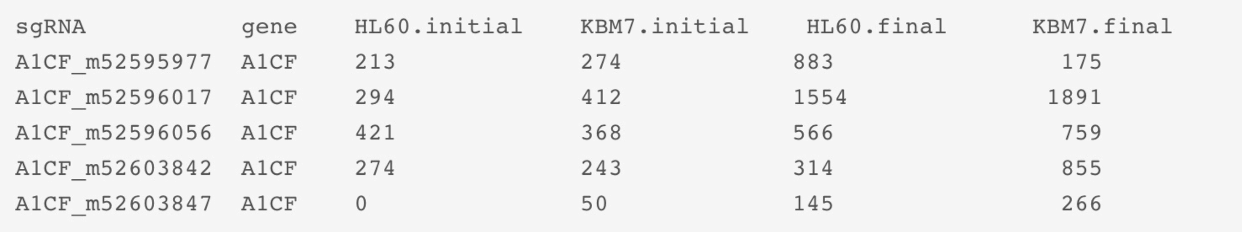
Part IV: How to run

- First, have to convert the fastq files into counts for each gene



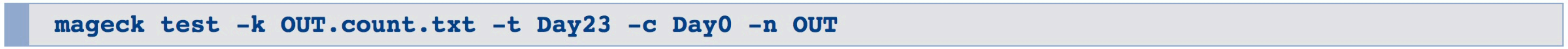
```
mageck count -l library.csv -n OUT --sample-label Day0,Day23 \
--fastq Day0_Rep1.fastq.gz,Day0_Rep2.fastq.gz Day23_Rep1.fastq.gz,Day23_Rep2.fastq.gz
```

- -l: The provided sgRNA information, including the sgRNA id, the sequence, and the gene it is targeting
- Replicates separated by comma, while samples from different conditions separated by space



```
sgRNA           gene    HL60.initial    KBM7.initial    HL60.final      KBM7.final
A1CF_m52595977  A1CF    213             274             883               175
A1CF_m52596017  A1CF    294             412             1554             1891
A1CF_m52596056  A1CF    421             368             566               759
A1CF_m52603842  A1CF    274             243             314               855
A1CF_m52603847  A1CF    0               50              145               266
```

- Then, test if the counts are significant or not

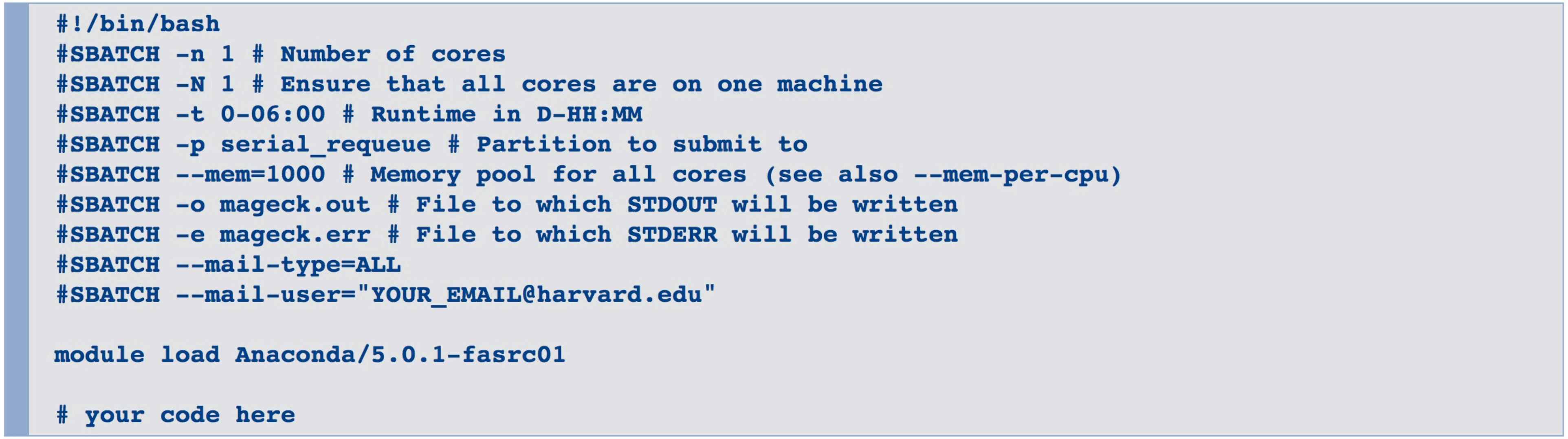


```
mageck test -k OUT.count.txt -t Day23 -c Day0 -n OUT
```

- Make sure that the labels match when running `mageck test`
- Output files: https://sourceforge.net/p/mageck/wiki/output/

Part IV: How to run

- Sample Slurm script



```
#!/bin/bash
#SBATCH -n 1 # Number of cores
#SBATCH -N 1 # Ensure that all cores are on one machine
#SBATCH -t 0-06:00 # Runtime in D-HH:MM
#SBATCH -p serial_requeue # Partition to submit to
#SBATCH --mem=1000 # Memory pool for all cores (see also --mem-per-cpu)
#SBATCH -o mageck.out # File to which STDOUT will be written
#SBATCH -e mageck.err # File to which STDERR will be written
#SBATCH --mail-type=ALL
#SBATCH --mail-user="YOUR_EMAIL@harvard.edu"

module load Anaconda/5.0.1-fasrc01

# your code here
```

Part IV Q1: QC

- Look at the countsummary.txt file generated after mageck count. Look at documentation here for guide to QC metrics:
- https://sourceforge.net/p/mageck/wiki/output/
- We want:
  - Percentage of reads mapped to be above 0.6
  - Zero counts less than 0.1
  - Gini index less than 0.1
- Ribosomal genes
  - Ribosomal genes are survival essential, thus KOs will definitely result in death, often put as negative control
  - Check the genesummary.txt for ribosomal genes
  - Genes ranked by how negatively selected they are (most to least)
  - Ribosomal genes start with “RP”, so you can get all the rows that have ribosomal genes using grepl("^RP", genesummary$id) on the id column of the genesummary.txt file

Part IV Q1: Replicate Consistency

- Count each replicate separately
- Code to count separately:

```
mageck count -l library.csv -n OUT_SEPARATE --sample-label Day0_Rep1,Day0_Rep2,Day23_Rep1,Day23_Rep2 \
--fastq Day0_Rep1.fastq.gz Day0_Rep2.fastq.gz Day23_Rep1.fastq.gz Day23_Rep2.fastq.gz
```

- Note that now we don’t put a comma between the replicates because we want them to be considered separately
- The resultant count.txt file will contain one column for each of the 4 samples
- Plot the counts for Rep1 against Rep2 and look at the correlation

sgRNA	Gene	D0_Rep1	D0_Rep2	D23_Rep1	D23_Rep2
…	…	…	…	…	…

Part IV Q2: Positive and Negative Selection Genes

- Again, see genesummary.txt
- Can use FDR < 0.05 to identify the genes
- Can use DAVID for pathway enrichment

Part IV Q3: Drug target

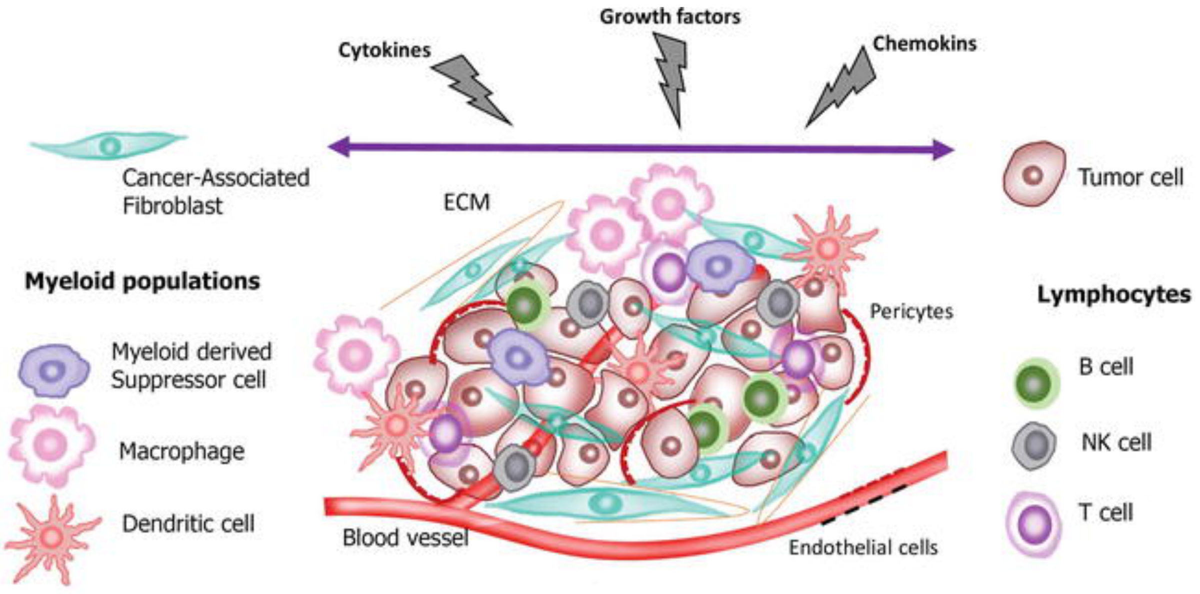
- Negatively selected genes→ Vital for GBM cancer cells survival
- Potential drug target
- But are they also negatively selected in other normal cell types? (i.e. vital for other cells as well)
- If so, not ideal for targeting them → Toxicity
- Visualize Expression vs Dependency in many cell types
  - A more negative dependency: more depleted in a CRISPR screen
- If you can’t find a gene, try to look for its alias

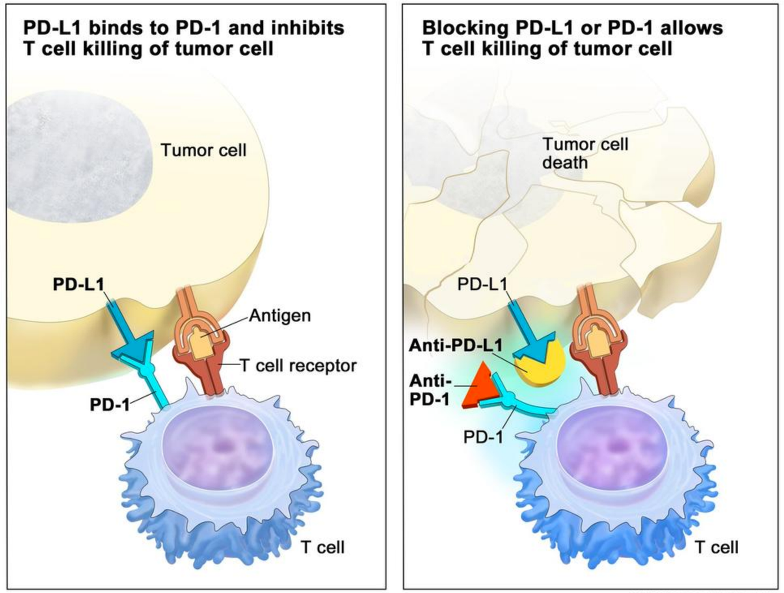
Part IV Q4: Drug Target

- Remove genes that are in the PanEssential.txt file from the negatively selected genes
- Can sort by FDR to find the top ones
- OASIS genomics website
  - Make sure to select GBM

PartV: Cancer immunology and immunotherapy

- cancer microenvironment
- Immune checkpoint and checkpoint inhibitors
  - Well-known checkpoints: CTLA4, PD1-PDL1, etc.




PD-L1 binds to PD-1 and inhibits T cell killing of tumor cell
Tumor cell
PD-L1
Antigen
T cell receptor
PD-1
T cell
Blocking PD-L1 or PD-1 allows T cell killing of tumor cell
Tumor cell death
PD-L1
Anti-PD-L1
Anti-PD-1
PD-1
T cell

© 2015 Terese Winslow LLC
U.S. Govt. has certain rights

Part V: Cancer immunology and immunotherapy

- TIMER
- Q1: Are there immune infiltrates into the microenvironment?
- Q2: Are immune checkpoints present?
- Q3: Does the presence of immune infiltrates improve survival?
- **Several corrections** (website updated):
- Q1: **Gene_DE** tab to look at differential expression
- Q2: Gene tab, you may include **only T cell CD8+** as the infiltrate
- Q2: PD1 = PDCD1, PD1L = CD274
- Q3: Look at the “**Outcome”** tab to find survival outcome, again, include only **T cell CD8+** would be good

Thank you!

Acknowledgement

- Andy Shi
- Dr. Shirley Liu