MSBI 32400 – LAB 8

LARRY HELSETH, PHD AND JASON EDELSTEIN

August 9, 2017

Outline

2

- DEMO-Working with Tumor Normal data
- DEMO-Annotating tumor/normal VCF
- Annotating gene panel data
- Using IGV to view cancer patient data, gene lists, networks, etc.
- DEMO-Vignette: How many genes do we have?

Demo

3

- Chr1 from Tumor and Normal samples for patient with pancreatic cancer. Sample = 60% tumor
- Aligned with BWA, then samtools mpileup using BED file and hg19 reference genome using Galaxy
- Genes on chr1 from KEGG Pancreatic Cancer pathway: CDC42, E2F2, JAK1, PIK3CD, PIK3R3, TGFB2

snpEff cancer annotation syntax

4

```
Annotations options:
        -cancer                         : Perform 'cancer' comparisons (Somatic vs Germline). Default: true
        -cancerSamples <file>           : Two column TXT file defining 'oringinal \t derived' samples.
        -formatEff                      : Use 'EFF' field compatible with older versions (instead of 'ANN').
        -geneId                         : Use gene ID instead of gene name (VCF output). Default: false
        -hgvs                           : Use HGVS annotations for amino acid sub-field. Default: true
        -hgvsOld                        : Use old HGVS notation. Default: false
        -hgvs1LetterAa                  : Use one letter Amino acid codes in HGVS notation. Default: false
        -hgvsTrId                       : Use transcript ID in HGVS notation. Default: false
        -lof                            : Add loss of function (LOF) and Nonsense mediated decay (NMD) tags.
        -noHgvs                         : Do not add HGVS annotations.
        -noLof                          : Do not add LOF and NMD annotations.
        -noShiftHgvs                    : Do not shift variants according to HGVS notation (most 3prime end).
        -oicr                           : Add OICR tag in VCF file. Default: false
        -sequenceOntology               : Use Sequence Ontology terms. Default: true
```

Running snpEff –cancer on 1033.chr1 T/N pair

5

- Edited the VCF so last line of header reads:
  #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 1033.chr1.Normal.bam 1033.chr1.Tumor.bam
- Prepared a cancer_samples.txt file with:
  1033.chr1.Normal 1033.chr1.Tumor
- time java -Xmx2G -jar /data/snpEff/snpEff.jar -v -cancerSamples samples_cancer.txt -cancer hg19 1033.chr1.vcf > 1033.chr1.cancer.ann.vcf

Results

Applications Places System Bioinfo Student Wed Feb 22, 10:21 AM

student@MSBI32400Lab1:/home/student/testing

File Edit View Search Terminal Help

```
[root@MSBI32400Lab1 testing]# time java -Xmx2G -jar /data/snpEff/snpEff.jar -v -cancerSamples samples_cancer.txt -cancer hg19 1033.chr1.vcf > 1033.chr1.cancer.ann.vcf
00:00:00        SnpEff version SnpEff 4.3i (build 2016-12-15 22:33), by Pablo Cingolani
00:00:00        Command: 'ann'
00:00:00        Reading configuration file 'snpEff.config'. Genome: 'hg19'
00:00:00        Reading config file: /home/student/testing/snpEff.config
00:00:00        Reading config file: /data/snpEff/snpEff.config
00:00:01        done
00:00:01        Reading database for genome version 'hg19' from file '/data/snpEff/./data/hg19/snpEffectPredictor.bin' (this might take a while)
00:00:14        done
00:00:14        Reading NextProt database from file '/data/snpEff/./data/hg19/nextProt.bin'
00:00:17        NextProt database: 542362 markers loaded.
00:00:17        Adding transcript info to NextProt markers.
00:00:17        NextProt database: 542362 markers added.
00:00:17        Loading Motifs and PWMs
00:00:17        Loading interactions from : /data/snpEff/./data/hg19/interactions.bin
00:00:29                Interactions: 1590613 added, 0 skipped.
00:00:29        Building interval forest
00:00:43        done.
00:00:43        Genome stats :
#-----------------------------------------------
# Genome name                 : 'Homo_sapiens (UCSC)'
# Genome version              : 'hg19'
# Genome ID                   : 'hg19[0]'
# Has protein coding info     : true
# Has Tr. Support Level info  : true
# Genes                       : 29583
# Protein coding genes        : 20797
#-----------------------------------------------
# Transcripts                 : 60834
# Avg. transcripts per gene   : 2.06
# TSL transcripts             : 0
#-----------------------------------------------
# Checked transcripts         :
#               AA sequences :      0 ( 0.00% )
#              DNA sequences :  52386 ( 86.11% )
#-----------------------------------------------
# Protein coding transcripts : 46522
#               Length errors :     93 ( 0.20% )
#  STOP codons in CDS errors :     78 ( 0.17% )
#       START codon errors :    117 ( 0.25% )
#        STOP codon warnings :     19 ( 0.04% )
#              UTR sequences :  45868 ( 75.40% )
#               Total Errors :    256 ( 0.55% )
#-----------------------------------------------
# Cds                         : 460256
# Exons                       : 570329
```

[student@MSBI32400... student@MSBI32400L...

Results (cont)

```
Applications  Places  System                    Bioinfo Student                    Wed Feb 22, 10:21 AM
student@MSBI32400Lab1:/home/student/testing
File  Edit  View  Search  Terminal  Help
#               'Un_gl000239'   33824   Standard
#               '21_gl000210_random'    27682   Standard
#               'Un_gl000231'   27386   Standard
#               'Un_gl000229'   19913   Standard
#               'M'     16571   Vertebrate_Mitochondrial
#               'Un_gl000226'   15008   Standard
#               '18_gl000207_random'    4262    Standard
#-----------------------------------------------

00:01:16        Predicting variants
00:01:16        Reading cancer samples pedigree from file 'samples_cancer.txt'.
java.lang.RuntimeException: Cannot find pedigree Father/Original sample name '1033.chr1.Normal'
        at org.snpeff.vcf.PedigreeEnrty.sampleNumbers(PedigreeEnrty.java:74)
        at org.snpeff.snpEffect.commandLine.SnpEffCmdEff.readPedigree(SnpEffCmdEff.java:954)
        at org.snpeff.snpEffect.commandLine.SnpEffCmdEff.annotate(SnpEffCmdEff.java:171)
        at org.snpeff.snpEffect.commandLine.SnpEffCmdEff.annotateVcf(SnpEffCmdEff.java:465)
        at org.snpeff.snpEffect.commandLine.SnpEffCmdEff.annotate(SnpEffCmdEff.java:142)
        at org.snpeff.snpEffect.commandLine.SnpEffCmdEff.run(SnpEffCmdEff.java:1026)
        at org.snpeff.snpEffect.commandLine.SnpEffCmdEff.run(SnpEffCmdEff.java:981)
        at org.snpeff.SnpEff.run(SnpEff.java:1041)
        at org.snpeff.SnpEff.main(SnpEff.java:159)
Error: Error while processing VCF entry (line 28) :
        chr1    15029   .       G       A       69.0    .       DP=293;VDB=0.0386;AF1=0.25;AC1=1;DP4=54,137,24,43;MQ=18;FQ=70.2;PV4=0.28,1,1,0.49       GT:PL:GQ        0/0:0,116
,137:99 0/1:102,0,116:99
java.lang.RuntimeException: Cannot find pedigree Father/Original sample name '1033.chr1.Normal'
00:01:17        Loading sequences for chromosome '1' from file '/data/snpEff/./data/hg19/sequence.1.bin'
00:01:26        Building sequence tree for chromosome '1'
00:01:26        Done. Loaded 2071 sequences.

WARNINGS: Some warning were detected
Warning type    Number of warnings
WARNING_TRANSCRIPT_MULTIPLE_STOP_CODONS 31
WARNING_TRANSCRIPT_NO_START_CODON       34


00:01:35        Creating summary file: snpEff_summary.html
00:01:36        Creating genes file: snpEff_genes.txt
1 errors.
00:01:37        done.
00:01:37        Logging
00:01:38        Checking for updates...

real    1m39.627s
user    1m36.032s
sys     0m1.290s
[root@MSBI32400Lab1 testing]#
[student@MSBI32400...   student@MSBI32400L...
```

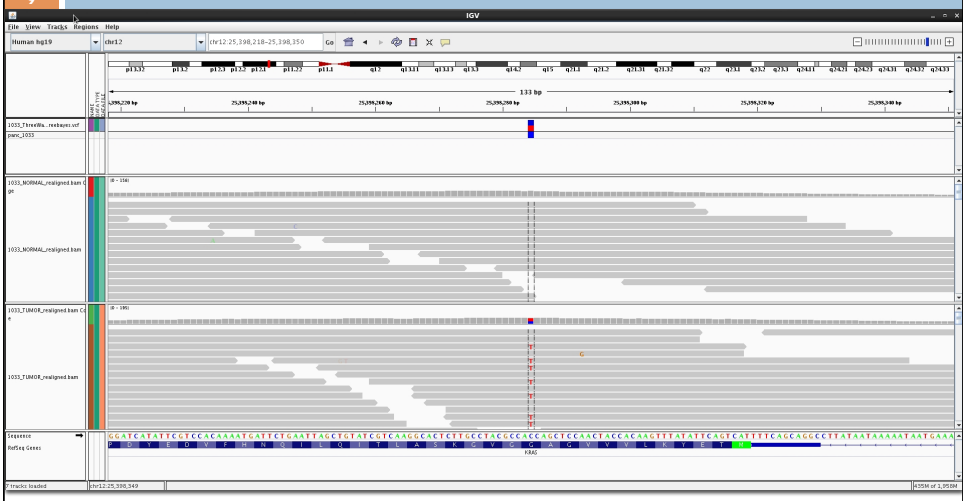
Searched for interesting variants…

8

- grep stop 1033.chr1.cancer.ann.vcf | grep ’0/1’ (& ‘1/1’) | grep '0/0’
  - NOTHING
- grep ‘<each of KEGG genes on chr1>’ 1033.chr1.cancer.ann.vcf | grep ‘0/1’ | grep ‘0/0’
  - NOTHING

Browsing the full VCF + 3 GB BAMs

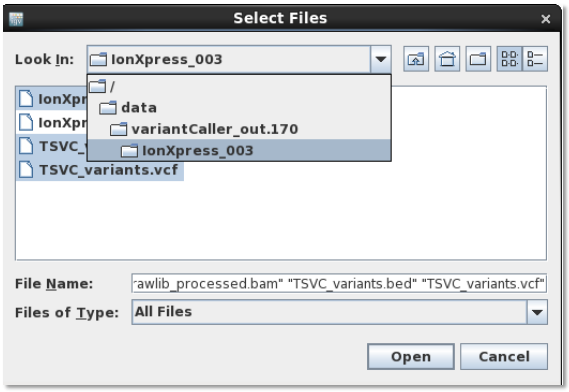
9



IGV view of Cancer Hotspot data

10

- Launch IGV in your VM, then open Cancer Hotspot data in /data/variantCaller_out.170/ IonXpress_003



Annotate to find interesting regions

11

- From your /data/lab8/results folder:
  java -Xmx2G -jar /data/snpEff/snpEff.jar eff -canon -noLog hg19 /data/variantCaller_out.170/IonXpress_003/TSVC_variants.vcf > TSVC_variants.snpEff.vcf
- java -Xmx2G -jar /data/snpEff/SnpSift.jar annotate -noLog /data/snpEff/data/hg19/clinvar/clinvar_20170701.vcf.gz TSVC_variants.snpEff.vcf > TSVC_variants.snpEff.clinvar.vcf

Do some quick filtering

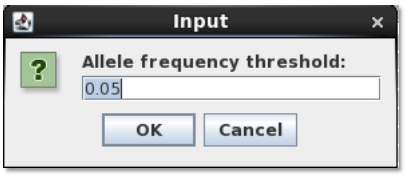
12

- grep -v "^#" TSVC_variants.snpEff.clinvar.vcf | grep -v '0/0' shows SNPs that aren't absent
- **grep -v "^#" TSVC_variants.snpEff.clinvar.vcf | grep -v '0/0' | grep stop** shows stop variants
- Open BAM + VCF + BED in IGV then go to region identified by grep as a stop.

Changing IGV

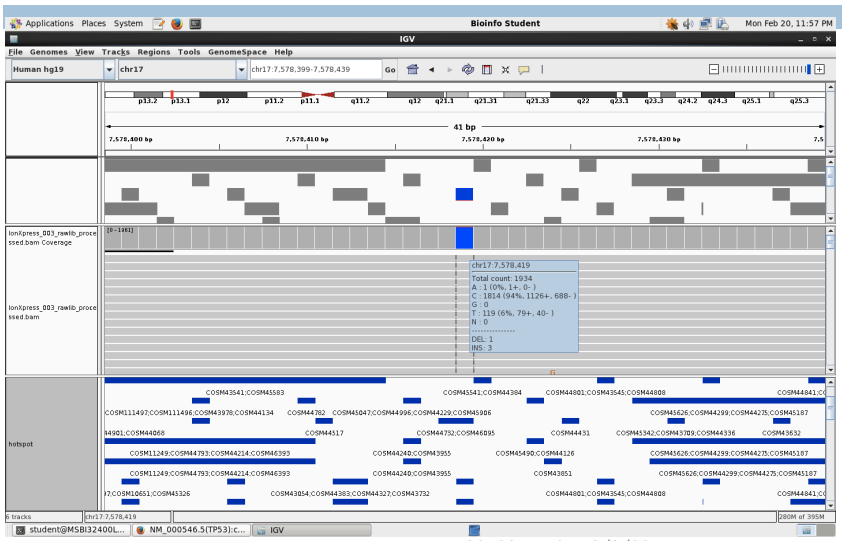
13

- CHANGE allele frequency threshold from 0.2 to 0.05 to view low frequency variants in TP53
- Either Command-Click (Mac) on depth of coverage or View/Preferences then change on the Alignments tab



View Hotspot files in stop region

14



Select COSMIC in hotspot bed

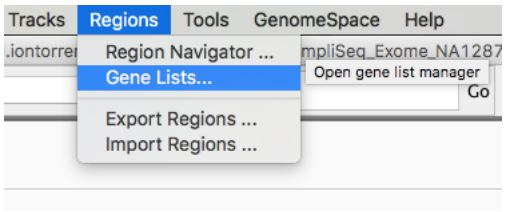
15

- Copy info for COSMIC ID that best matches SNP then look up on COSMIC database (GRCh37 archive)

Filter IGV view with gene list

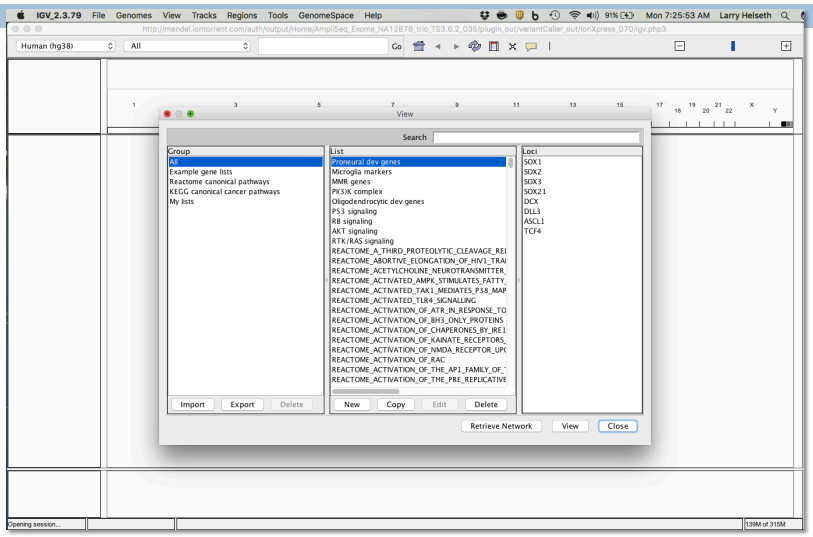
16

- Open Regions/Gene Lists



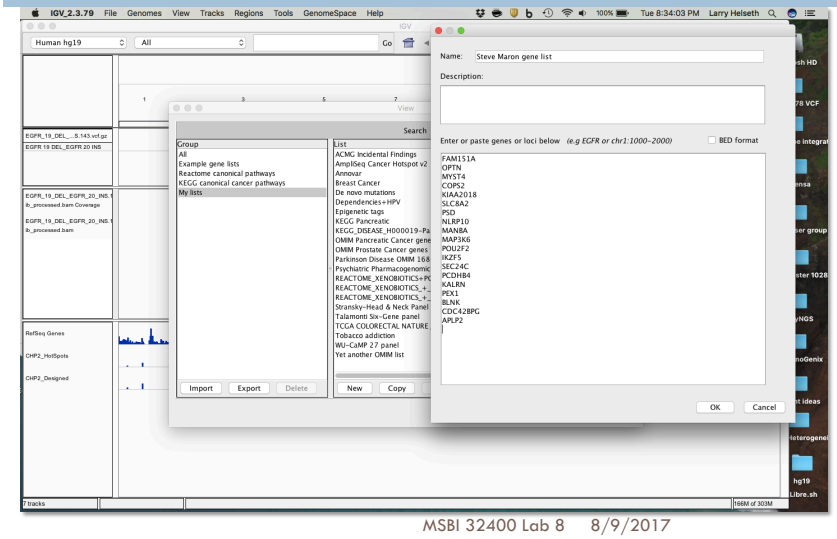
Gene list view

17



Can make your own gene list

18



Download EGFR hotspot sample

19

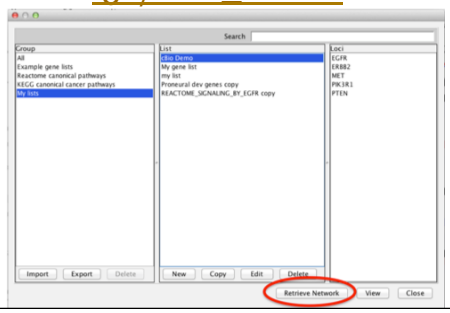
- From Canvas Files/Lab8 download the EGFR Hotspot folder
- Open BAM, VCF.gz, hotspots bed and designed bed file in IGV
- Go to gene EGFR
- Zoom & inspect Exon 19
- Expand hotspot bed track
- Identify COSMIC ID that matches observed change
  - Include that in write-up, along with coordinates & full description from COSMIC web site. What tumor type is this most commonly seen in?
- Use KEGG gene list for above cancer type and examine other genes for SNPs in coding regions.
  - Report at least two from different genes

MSBI 32400 Lab 8 8/9/2017

Visualizing cBio Network (~~BROKEN~~)

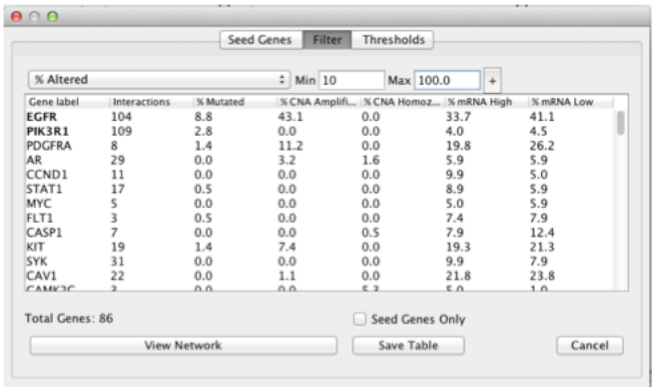
20

- Allows us to look at selected genes, their network "neighbors" and drugs which act on them
- Launch from IGV
- https://software.broadinstitute.org/software/igv/cbio_viewer



MSBI 32400 Lab 8 8/9/2017

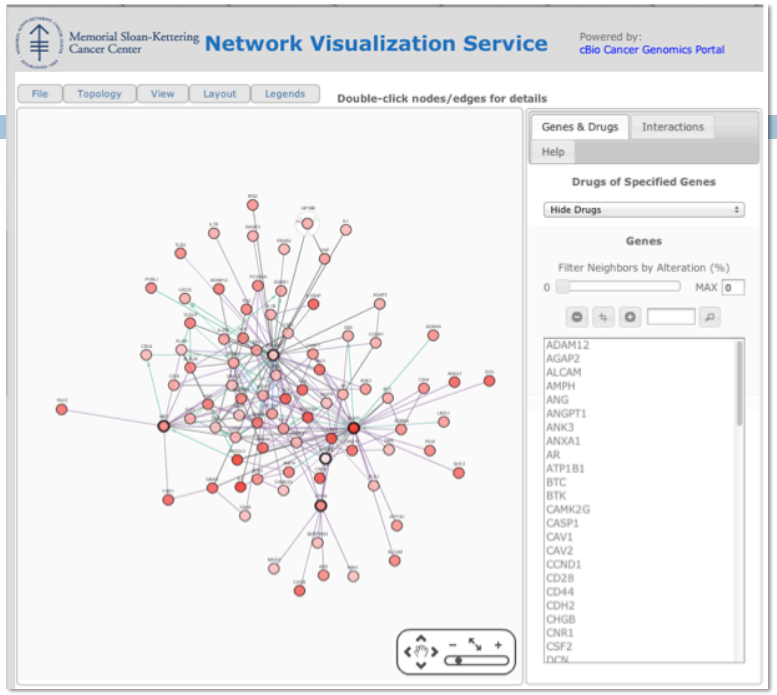
21



Gene label	Interactions	% Mutated	% CNA Amplifi...	% CNA Homoz...	% mRNA High	% mRNA Low
EGFR	104	8.8	43.1	0.0	33.7	41.1
PIK3R1	109	2.8	0.0	0.0	4.0	4.5
PDGFRA	8	1.4	11.2	0.0	19.8	26.2
AR	29	0.0	3.2	1.6	5.9	5.9
CCND1	11	0.0	0.0	0.0	9.9	5.0
STAT1	17	0.5	0.0	0.0	8.9	5.9
MYC	5	0.0	0.0	0.0	5.0	5.9
FLT1	3	0.5	0.0	0.0	7.4	7.9
CASP1	7	0.0	0.0	0.5	7.9	12.4
KIT	19	1.4	7.4	0.0	19.3	21.3
SYK	31	0.0	0.0	0.0	9.9	7.9
CAV1	22	0.0	1.1	0.0	21.8	23.8

MSBI 32400 Lab 8 8/9/2017

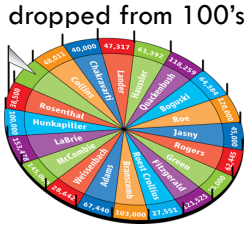
22



How Many Proteins in our Proteome?

23

- Estimates of the number of human genes have dropped from 100's of thousands to < 21,000
- The number of proteins is based on the proteome
- Numbers of proteins also vary based on alternative splicing
- Several labs report >10,000 protein IDs



Pennisi, E. "Working the (gene count) numbers: finally, a firm answer?" Science. 2007 May 25;316(5828):1113.

- ENCODE* found 20,687 protein-coding genes with 6.3 alternatively spliced transcripts per locus, whose coding exons encompass 1.22% of the genome (20,687 X 6.3 = 130,328.1 proteins)
- UniProtKB+ lists 131,333 human proteins, of which 68,079 represent the "complete proteome" (as of 11/28/12).

*30 papers published 6 Sept 2012 at http://nature.com/encode
+http://www.uniprot.org/

Pevsner example - Chapter 20 #3

24

```
for chr in {1..22} X Y MT
do
  esearch -db gene -query "Homo sapiens [ORGN]
AND $chr [CHR]" |
  efilter -query "alive [PROP] AND genetype
protein coding [PROP]" |
  efetch -format docsum |
  xtract -pattern DocumentSummary -NAME Name \
  -block GenomicInfoType -match "ChrLoc:$chr" \
  -tab "\n" -element ChrLoc,"&NAME" |
  grep '.' | sort | uniq | cut -f 1 |
  sort-uniq-count-rank
done
```

Script outputs a list of counts by chr

25

- Use Linux **paste** to string the numbers together, separated by "+" sign
- Use Linux **bc** to calculate the sum

student@MSBI32400Lab1:~/testing

File Edit View Search Terminal Help

```
[student@MSBI32400Lab1 testing]$ cat genes_by_chr.txt
2069    1
1267    2
1070    3
754     4
872     5
1037    6
935     7
690     8
800     9
738     10
1294    11
1026    12
335     13
611     14
605     15
863     16
1185    17
277     18
1404    19
545     20
248     21
446     22
850     X
71      Y
13      MT
[student@MSBI32400Lab1 testing]$ cut -f1 genes_by_chr.txt | paste -sd+ | bc
20005
[student@MSBI32400Lab1 testing]$ ls -ltr
total 8
-rwxrwxr-x. 1 student student 391 Feb 19 16:29 pevsner_ch20_3_problem.sh
-rw-rw-r--. 1 student student 170 Feb 19 16:33 genes_by_chr.txt
[student@MSBI32400Lab1 testing]$
```

Homework

27

- E-mail Jason (jasone@uchicago.edu) the README with the file information requested above before next class with "**Lab #8**" in the subject line