MSBI 32400 – LAB 8
LARRY HELSETH, PHD AND JASON EDELSTEIN

March 6, 2019

Outline

- 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

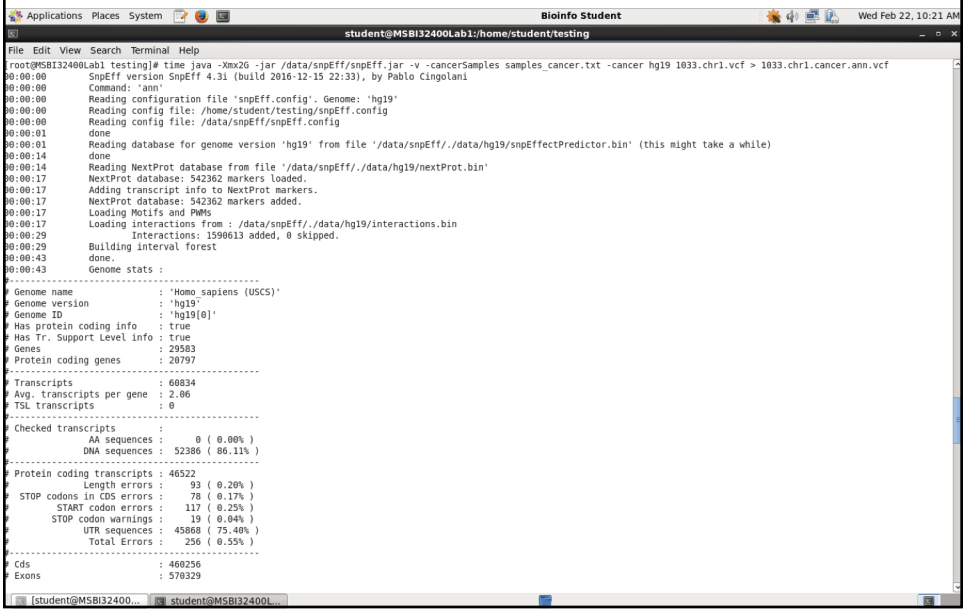
```
Annotations options:
        -cancer                         : Perform 'cancer' comparisons (Somatic vs Germline). Default: true
        -cancerSamples <file>           : Two column TXT file defining 'oringinal \t derived' samples.
        -formatEff                      : Use 'EFF' field compatible 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

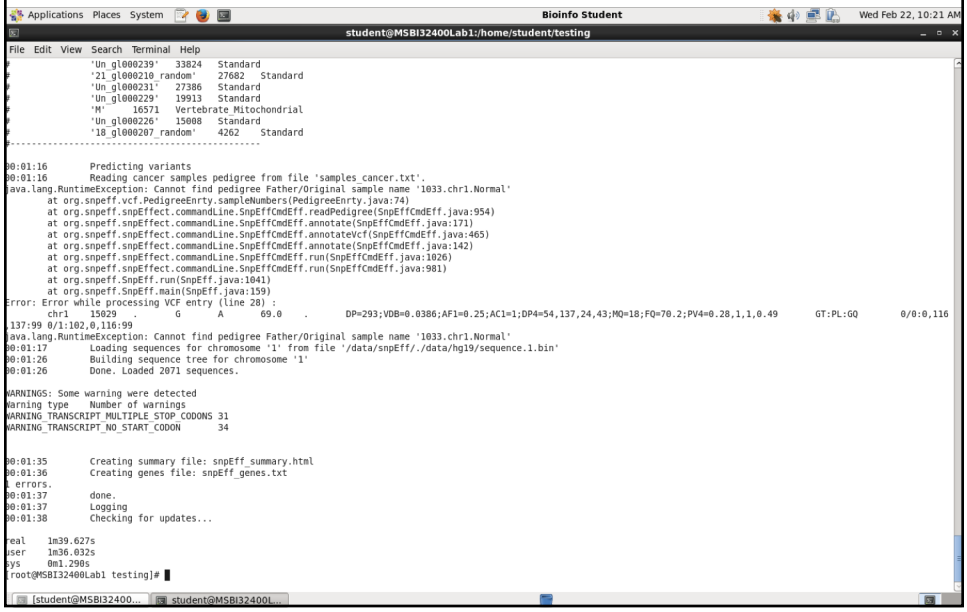
5

- 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



Results (cont)



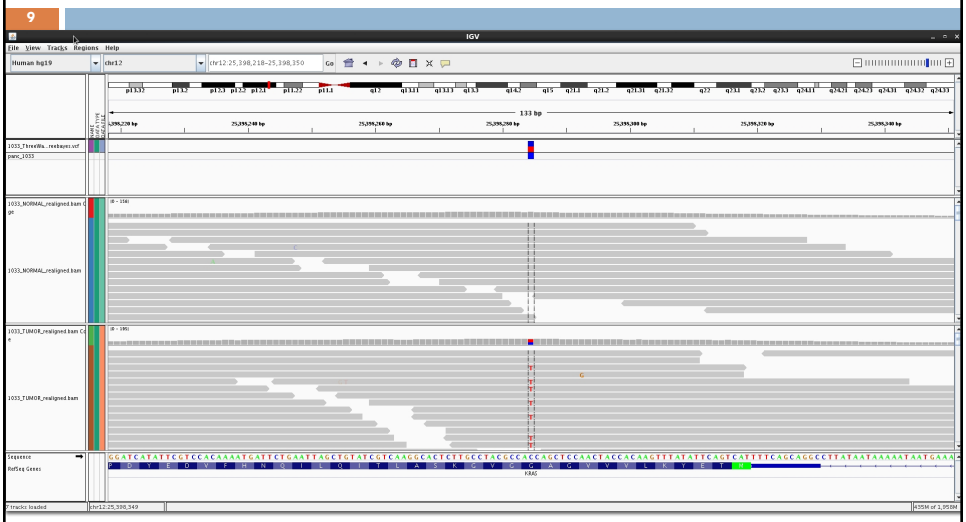
Searched for interesting variants…

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

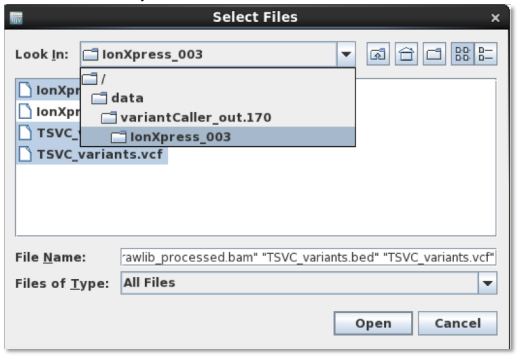
9



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_20190211.vcf.gz** TSVC_variants.snpEff.vcf > TSVC_variants.snpEff.clinvar.vcf

Do some quick filtering

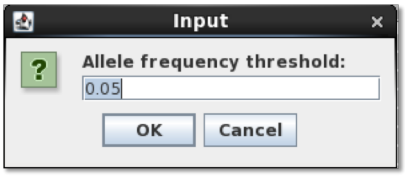
12

- 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

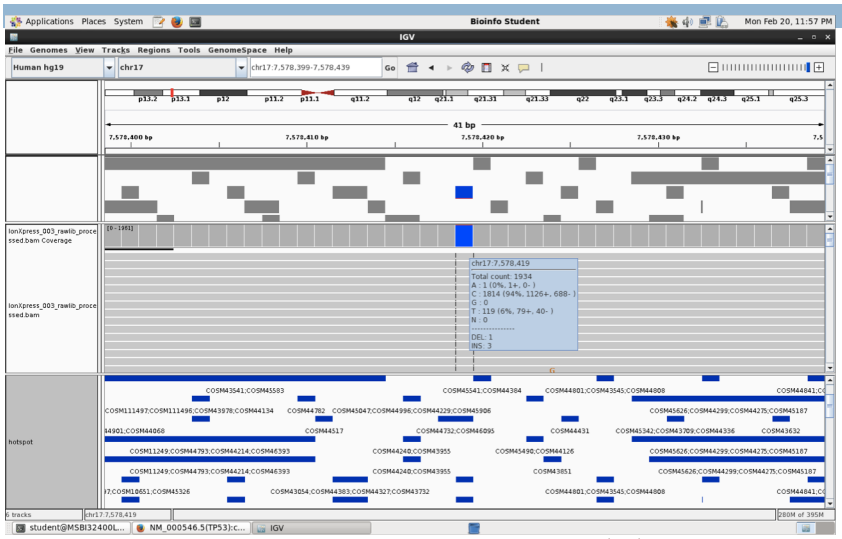
13

- 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

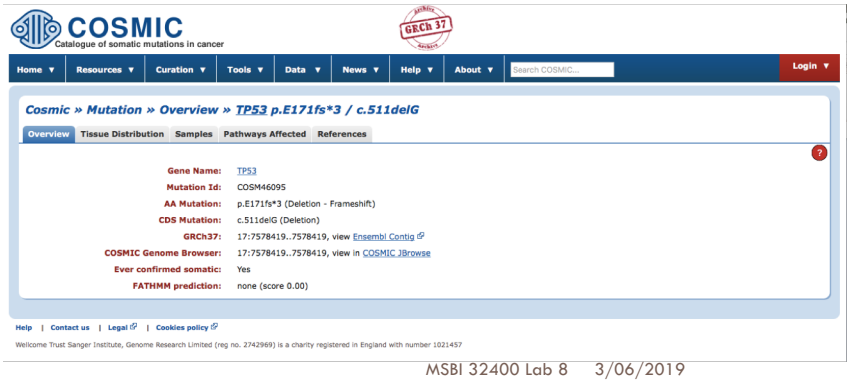
14



Select COSMIC in hotspot bed

15

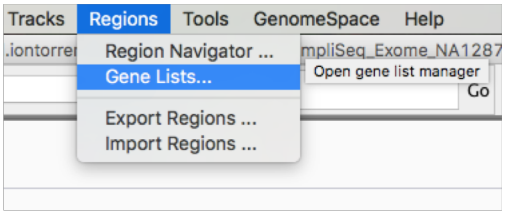
- Copy info for COSMIC ID that best matches SNP then look up on COSMIC database (GRCh37 archive)



Filter IGV view with gene list

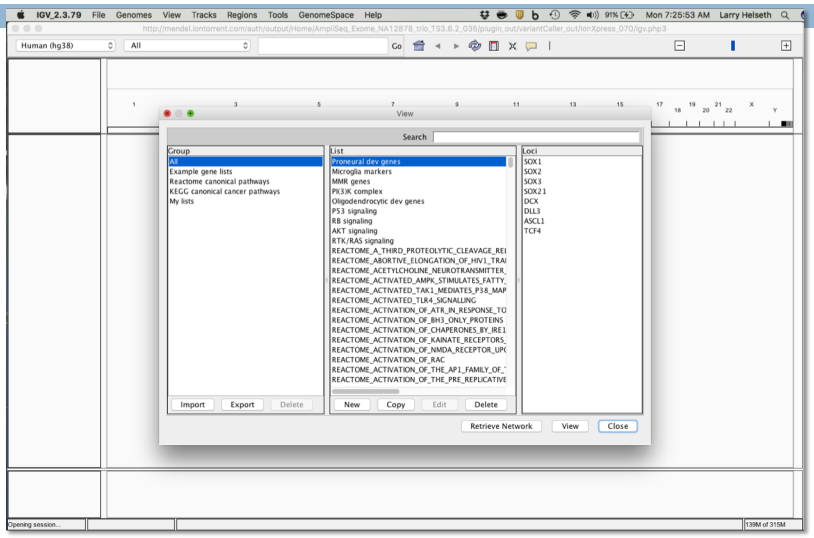
16

- Open Regions/Gene Lists



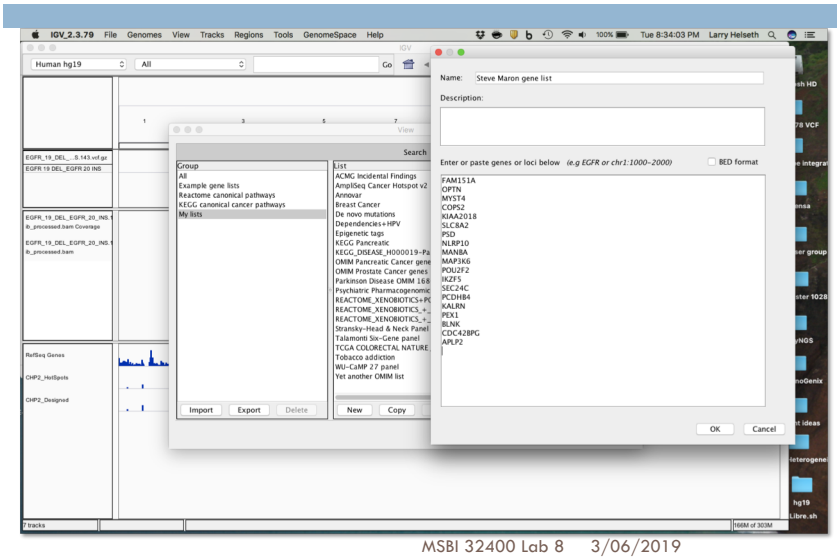
Gene list view

17



Can make your own gene list

18



EGFR hotspot sample

19

- Change privileges on folder from /data:
  - chmod 775 EGFR_19_DEL_EGFR_20_INS.143
- 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

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

How Many Proteins in our Proteome?

21

- 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

22

```
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

23

- Redirect previous command to a file
- Cut the output to first column (the number of genes, not the chromosome), then use Linux **paste** to string the numbers together, separated by "+" sign ("paste -sd+")
- Pipe to **bc** to calculate the sum (Linux "basic calculator")

Homework

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