MSBI 32400 – LAB 3

LARRY HELSETH PHD &
JASON EDELSTEIN

January 23, 2019

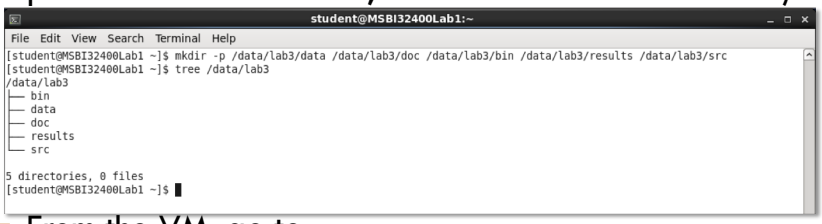
Outline

- First we'll download some files from Pevsner's website
- Use some text commands to trim files
- Generate BED files from UCSC
- Then install bedtools on the VM
- Use bedtools to sort and merge bed file
- Use samtools to work with BAM files

Pevsner's book companion site

3

- Create your project directories on the VM ('-p' creates parent if it doesn't exist; stack commands on one line):

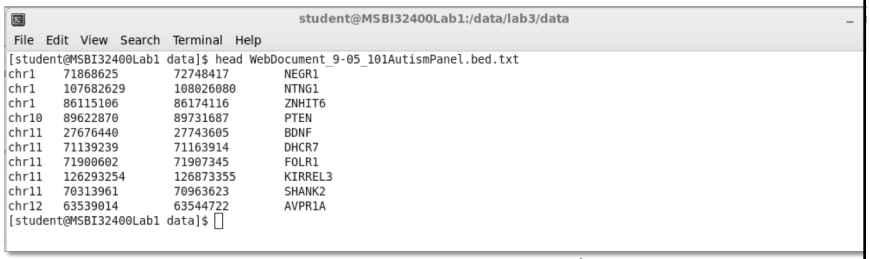


- From the VM, go to http://www.bioinfbook.org/php/C9E3k and download 9.5 and both 9.7 files into /data/lab3/data
  - May need to move ("mv") files from ~/Download to /data/lab3/data (hint- "mv <path to files> ." moves to current location)

Let's get the genes from file 9.5

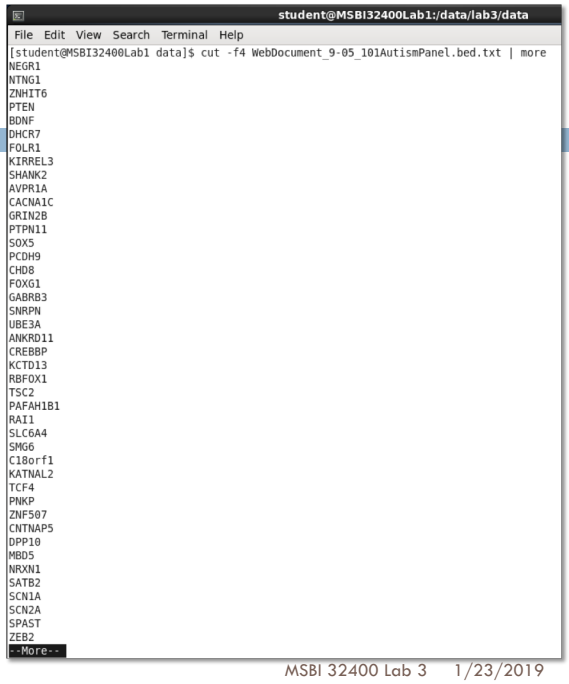
4

- Head the document to see structure



- Use the "cut" command to extract the 4th column (gene symbols) and send it to a text file:
  - cut -f4 WebDocument_9-05_101AutismPanel.bed.txt > genelist.txt

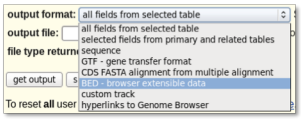
5

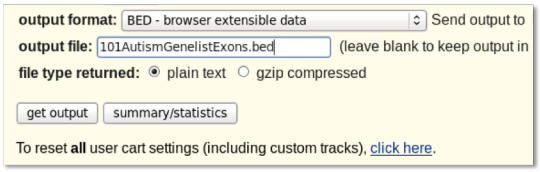


Now, go to UCSC and upload genes

6

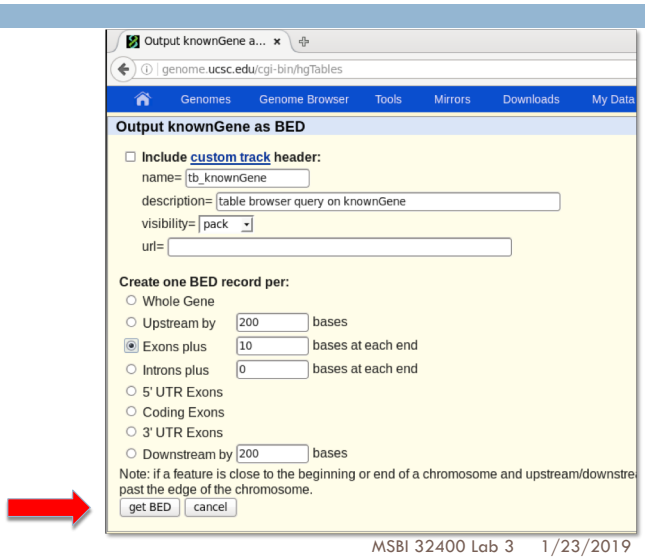
- http://genome.ucsc.edu/cgi-bin/hgTables
- Change Assembly to hg19
- "Upload list" in Identifiers and browse to your new **genelist.txt** & Submit
- Change output type to BED
- Give a filename and click **get output**





Change output type to exon + 10bp

7



Move the new BED file to lab3/data

8

```
student@MSBI32400Lab1:/data/lab3/data
File  Edit  View  Search  Terminal  Help
[student@MSBI32400Lab1 data]$ ls -ltr
total 336832
-rw-rw-r--. 1 student student      3126 Jan 15 14:07 WebDocument_9-05_101AutismPanel.bed.txt
-rw-rw-r--. 1 student student   3131856 Jan 15 14:08 WebDocument_9-7_mysample1.bai
-rw-rw-r--. 1 student student 341767805 Jan 15 14:26 WebDocument_9-7_mysample1.bam
-rw-rw-r--. 1 student student       715 Jan 15 14:48 genelist.txt
[student@MSBI32400Lab1 data]$ ls -ltr ~/Downloads/
total 1452
-rw-rw-r--. 1 student student 711132 Dec 22 21:03 setuptools-32.2.0.zip
-rw-r--r--. 1 student student   3102 Dec 22 21:29 README_install_notes.md
-rw-rw-r--. 1 student student   4116 Dec 22 21:29 README_install_notes.html
-rw-rw-r--. 1 student student 186570 Dec 27 15:01 crossroads___bioinformatics_by_tuxedopengu.jpg
-rw-rw-r--. 1 student student 572374 Jan 15 14:52 101AutismGenelistExons.bed
[student@MSBI32400Lab1 data]$ mv ~/Downloads/101AutismGenelistExons.bed .
[student@MSBI32400Lab1 data]$
```

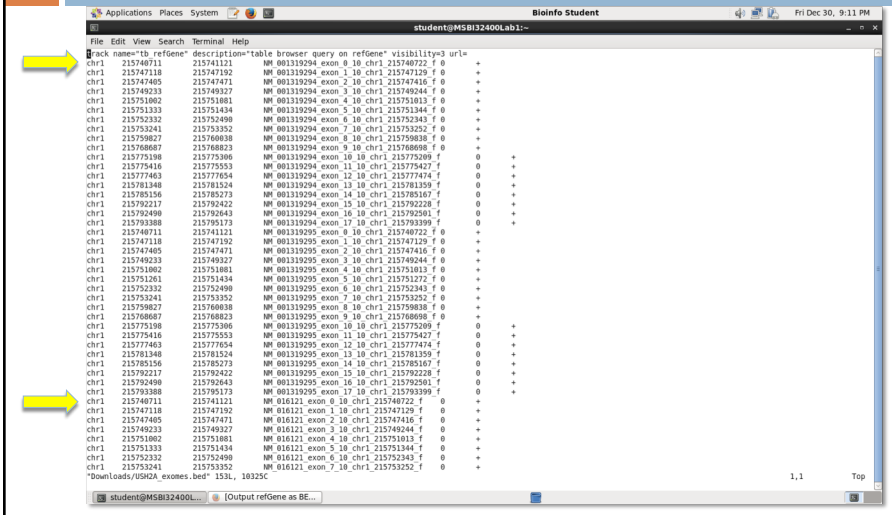
Open IGV then add BAM + BEDs

9

- File/Open in IGV then browse to /data/lab3/data
- Go to a gene like MEF2C and compare BED coverages
  - Pevsner's file is full gene, ours is just exons + 10 bp
  - Zoom in to compare the coverage
  - If you hover over one of the exons in BED file you'll see multiple transcripts
    - Need bedtools to clean that up!

BED file has overlapping transcripts

10



Installing bedtools (as non-root)

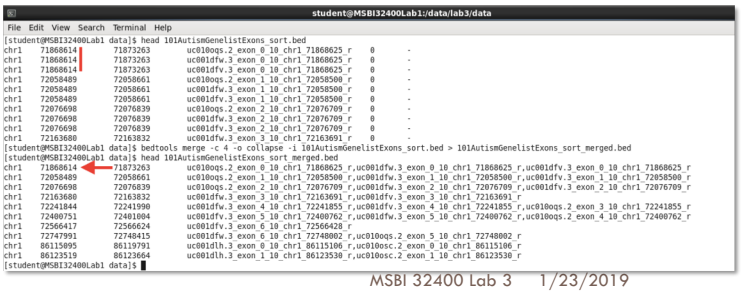
11

- Modified from http://bedtools.readthedocs.io/en/latest/content/installation.html
- Let's use the **wget** command from ~/Downloads folder:
  - wget https://github.com/arq5x/bedtools2/releases/download/v2.25.0/bedtools-2.25.0.tar.gz
  - Newer version (v2.27.1) won't compile on our VM
- Next extract using **tar** command
  - tar -zxvf bedtools-2.26.0.tar.gz
- cd to new bedtools2 directory
- Type "**make**" and wait while code is compiled, then copy the entire bin subdirectory to your home directory
  - cp -rp bin/ ~
- Add your new bin directory to your PATH
  - export PATH=~/bin:$PATH ("echo $PATH" to verify)

Let's clean up the BED file

12

- bedtools sort -i 101AutismGenelistExons.bed > 101AutismGenelistExons_sort.bed
- bedtools merge -c 4 -o collapse -i 101AutismGenelistExons_sort.bed > 101AutismGenelistExons_sort_merged.bed



Load merged BED in IGV

13

- Point to an exon and compare information with that from original UCSC downloaded BED file
- Advantage of merging is single track name
- Disadvantage is can't expand track to see all transcripts
- Your approach should vary based upon needs

SAMtools and other toys

14

- Copy /data/bds-files/chapter-11-alignment/NA12891_CEU_sample.vcf.gz to lab3/data folder
- Look at the file structure using **zcat** (pipe to more)
- Extract the text VCF using **gunzip** NA12891_CEU_sample.vcf.gz
- Count lines without comments using grep
  - grep -v "^#" | wc -l
  - Can also use: zcat NA12891_CEU_sample.vcf.gz | grep -v '^#' | wc -l if you didn't gunzip file
- Look for the samtoolsCommand & reference file in VCF header (it tells how VCF was generated)

SAMtools

15

- Use samtools to check the sample name:
  - samtools view -H WebDocument_9-7_mysample1.bam | grep '@RG'
- Use samtools to extract the FASTQ and SAM from BAM file (3620774 reads)
  - samtools fastq WebDocument_9-7_mysample1.bam > WebDocument_9-7_mysample1.fastq
- Use samtools to view SAM file
  - samtools view –h WebDocument_9-7_mysample1.bam > WebDocument_9-7_mysample1.sam
- Compare file sizes and send to Jason then delete .fastq to conserve space

Monitor what's going on

16

- Open a second terminal and use **top**
- Use **time** before your command to time the process
- My BAM → SAM took ~40 seconds.
- My SAM → sorted BAM took ~9 ½ minutes
- Can use second terminal to monitor intermediate file sizes
  - Click File/Open Terminal to open second terminal in same directory (can also Open Tab but harder to switch between using Alt-Tab)
  - Hint-Can open Terminal from File Explorer

Create a new BAM file from SAM

17

- samtools view -bS WebDocument_9-7_mysample1.sam | samtools sort -o WebDocument_9-7_mysample1_file_sorted.bam
- samtools index WebDocument_9-7_mysample1_file_sorted.bam
- Check that header information is intact:
  - samtools view -H WebDocument_9-7_mysample1_file_sorted.bam
- Compare the file sizes of the regenerated files with the original BAMs (send to Jason). Delete the SAM and new BAM files to conserve space on the VM.

Homework

18

- Upload to Canvas or e-mail Jason (jasone@uchicago.edu) the screenshots and file information requested above before next class with "**Lab #3**" in the subject line