MSBI 32400 – LAB 3

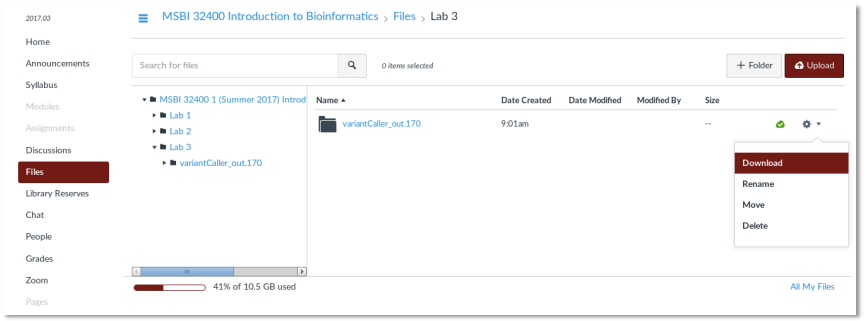
LARRY HELSETH PHD &
JASON EDELSTEIN

July 5, 2017

Replace some missing files first

2

- Sign into Canvas from your VM and download the variantCaller_out.170 folder from Lab 3 files:



Extract the folder into /data

3

- cd /data then unzip ~/Downloads/Lab_3_export.zip

Outline

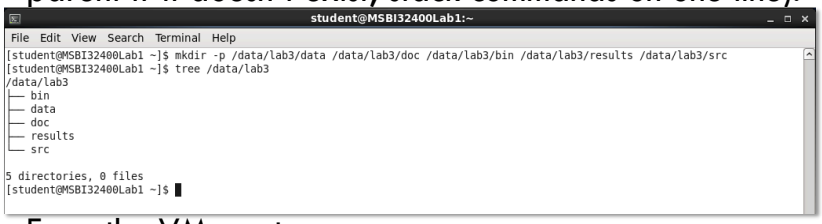
4

- 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

5

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

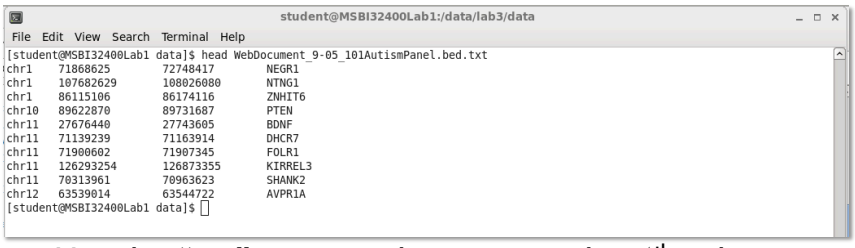


- From the VM, go to http://www.bioinfbook.org/php/C9E3k and download 9.5 and both 9.7 files into /data/lab3/data
  - May need to "mv" files from ~/Download to /data/lab3/data (hint- "mv <path to files> ." moves to current location)

Let's get the genes from file 9.5

6

- 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

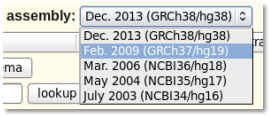
7

```
student@MSBI32400Lab1:/data/lab3/data
File  Edit  View  Search  Terminal  Help
[student@MSBI32400Lab1 data]$ cut -f4 WebDocument_9-05_101AutismPanel.bed.txt | more
NEGR1
NTNG1
ZNHIT6
PTEN
BDNF
DHCR7
FOLR1
KIRREL3
SHANK2
AVPR1A
CACNA1C
GRIN2B
PTPN11
SOX5
PCDH9
CHD8
FOXG1
GABRB3
SNRPN
UBE3A
ANKRD11
CREBBP
KCTD13
RBFOX1
TSC2
PAFAH1B1
RAI1
SLC6A4
SMG6
C18orf1
KATNAL2
TCF4
PNKP
ZNF507
CNTNAP5
DPP10
MBD5
NRXN1
SATB2
SCN1A
SCN2A
SPAST
ZEB2
--More--
```

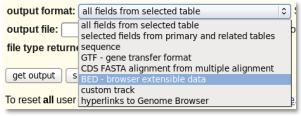

Now, go to UCSC and upload genes

8

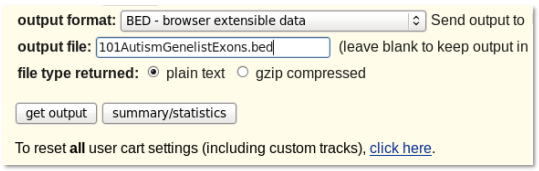
- http://genome.ucsc.edu/cgi-bin/hgTables
- Change Assembly to hg19



- "Upload list" in Identifiers and browse to your new **genelist.txt** & Submit
- Change output type to BED

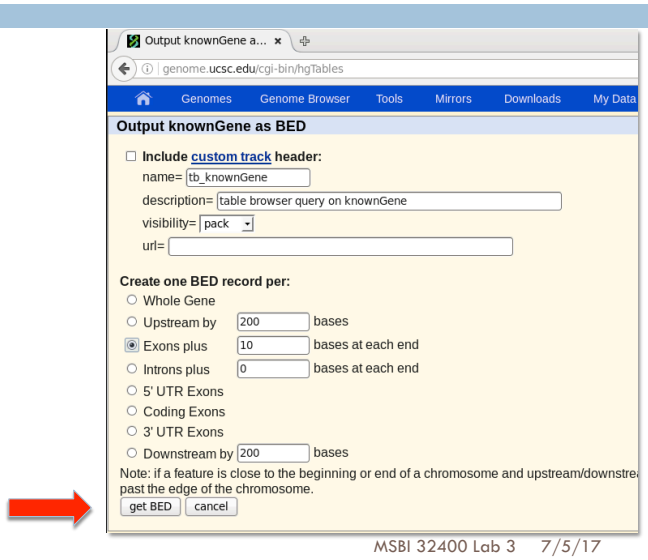


- Give a filename and click **get output**



Change output type to exon + 10bp

9



Move the new BED file to lab3/data

10

```
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]$
```

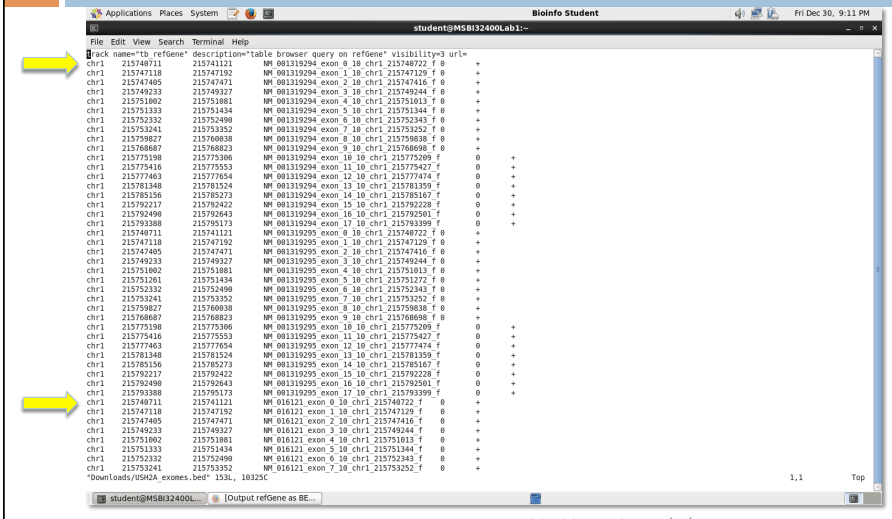
Open IGV then add BAM + BEDs

11

- File/Open in IGV then browse to /data/lab3/data
- Go to a gene like MEFC2 and compare BED coverages MEF2C***
  - 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

12



Installing bedtools (as non-root)

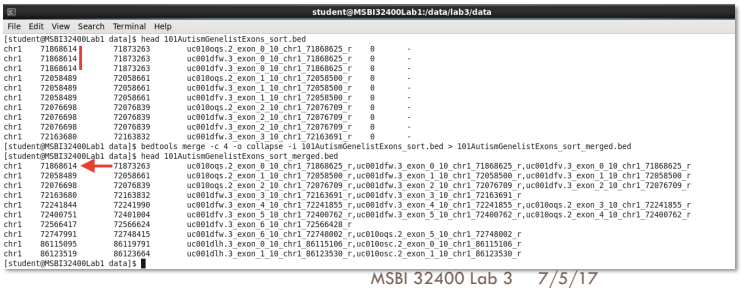
13

- Modified from http://bedtools.readthedocs.io/en/latest/content/installation.html (see also Pevsner pg. 413)
- Let's use the **wget** command from ~/Downloads folder:
  - wget https://github.com/arq5x/bedtools2/releases/download/v2.26.0/bedtools-2.26.0.tar.gz
- 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

14

- 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

15

- 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

16

- 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

17

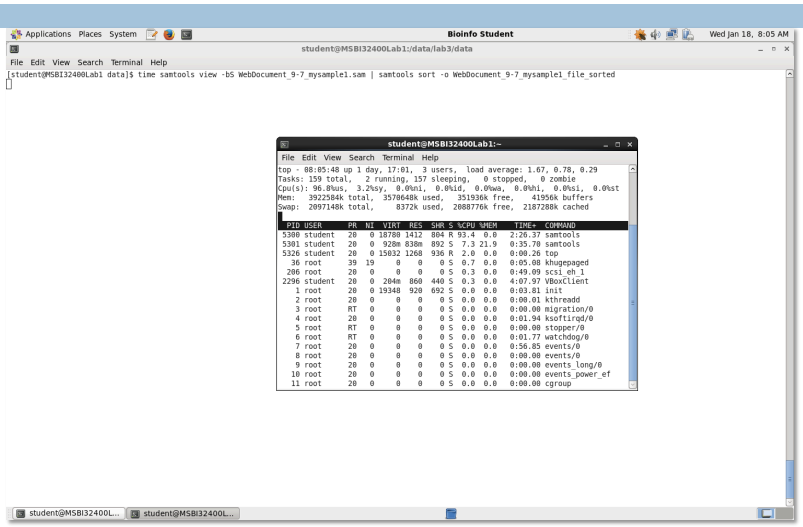
- Use samtools to check the sample name:
  - samtools view -H WebDocument_9-7_mysample1.bam | grep '@RG'
- Use samtools to extract the FASTQ and SAM from BAM file (3620774 reads)
  - samtools fastq WebDocument_9-7_mysample1.bam > WebDocument_9-7_mysample1.fastq
- Use samtools to view SAM file
  - samtools 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

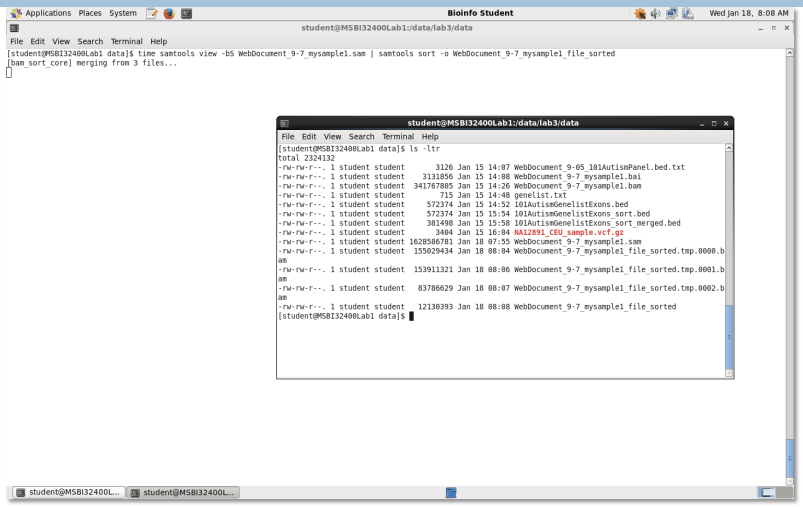
18

- 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

19


Applications Places System
Bioinfo Student
Wed Jan 18, 8:05 AM
student@MSBI32400Lab1:/data/lab3/data
File Edit View Search Terminal Help
[student@MSBI32400Lab1 data]$ time samtools view -bS WebDocument_9-7_mysample1.sam | samtools sort -o WebDocument_9-7_mysample1_file_sorted
student@MSBI32400Lab1:~
top - 08:05:48 up 1 day, 17:01, 3 users, load average: 1.67, 0.78, 0.29
Tasks: 159 total, 2 running, 157 sleeping, 0 stopped, 0 zombie
Cpu(s): 96.8%us, 3.2%sy, 0.0%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st
Mem: 3922584k total, 3570648k used, 351936k free, 41956k buffers
Swap: 2097148k total, 8372k used, 2088776k free, 2187288k cached
PID USER PR NI VIRT RES SHR S %CPU %MEM TIME+ COMMAND
5300 student 20 0 18780 1412 804 R 93.4 0.0 2:26.37 samtools
5301 student 20 0 928m 838m 892 S 7.3 21.9 0:35.70 samtools
5326 student 20 0 15032 1268 936 R 2.0 0.0 0:00.26 top
36 root 39 19 0 0 0 S 0.7 0.0 0:05.08 khugepaged
206 root 20 0 0 0 0 S 0.3 0.0 0:49.09 scsi_eh_1
2296 student 20 0 204m 860 440 S 0.3 0.0 4:07.97 VBoxClient
1 root 20 0 19348 920 692 S 0.0 0.0 0:03.81 init
2 root 20 0 0 0 0 S 0.0 0.0 0:00.01 kthreadd
3 root RT 0 0 0 0 S 0.0 0.0 0:00.00 migration/0
4 root 20 0 0 0 0 S 0.0 0.0 0:01.94 ksoftirqd/0
5 root RT 0 0 0 0 S 0.0 0.0 0:00.00 stopper/0
6 root RT 0 0 0 0 S 0.0 0.0 0:01.77 watchdog/0
7 root 20 0 0 0 0 S 0.0 0.0 0:56.85 events/0
8 root 20 0 0 0 0 S 0.0 0.0 0:00.00 events/0
9 root 20 0 0 0 0 S 0.0 0.0 0:00.00 events_long/0
10 root 20 0 0 0 0 S 0.0 0.0 0:00.00 events_power_ef
11 root 20 0 0 0 0 S 0.0 0.0 0:00.00 cgroup
student@MSBI32400L...
student@MSBI32400L...

20


Applications Places System
Bioinfo Student
Wed Jan 18, 8:08 AM
student@MSBI32400Lab1:/data/lab3/data
File Edit View Search Terminal Help
[student@MSBI32400Lab1 data]$ time samtools view -bS WebDocument_9-7_mysample1.sam | samtools sort -o WebDocument_9-7_mysample1_file_sorted
[bam_sort_core] merging from 3 files...
student@MSBI32400Lab1:/data/lab3/data
[student@MSBI32400Lab1 data]$ ls -ltr
total 2324132
-rw-rw-r--. 1 student student 3126 Jan 15 14:07 WebDocument_9-05_101AutismPanel.bed.txt
-rw-rw-r--. 1 student student 31318856 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
-rw-rw-r--. 1 student student 572374 Jan 15 14:52 101AutismGenelistExons.bed
-rw-rw-r--. 1 student student 572374 Jan 15 15:54 101AutismGenelistExons_sort.bed
-rw-rw-r--. 1 student student 381498 Jan 15 15:58 101AutismGenelistExons_sort_merged.bed
-rw-rw-r--. 1 student student 3404 Jan 15 16:04 NA12891_CEU_sample.vcf.gz
-rw-rw-r--. 1 student student 1628586781 Jan 18 07:55 WebDocument_9-7_mysample1.sam
-rw-rw-r--. 1 student student 155029434 Jan 18 08:04 WebDocument_9-7_mysample1_file_sorted.tmp.0000.bam
-rw-rw-r--. 1 student student 153911321 Jan 18 08:06 WebDocument_9-7_mysample1_file_sorted.tmp.0001.bam
-rw-rw-r--. 1 student student 83786629 Jan 18 08:07 WebDocument_9-7_mysample1_file_sorted.tmp.0002.bam
-rw-rw-r--. 1 student student 12130393 Jan 18 08:08 WebDocument_9-7_mysample1_file_sorted
[student@MSBI32400Lab1 data]$
student@MSBI32400L...
student@MSBI32400L...

Create a new BAM file from SAM

21

- 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

22

- E-mail Jason (jasone@uchicago.edu) the screenshots and file information requested above before next classwith "**Lab #3**" in the subject line