

MSBI 32400 – LAB 3

LARRY HELSETH PHD &
JASON EDELSTEIN

January 23, 2019

Outline

2

- 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):

```
student@MSBI32400Lab1:~$ mkdir -p /data/lab3/data /data/lab3/doc /data/lab3/bin /data/lab3/results /data/lab3/src
student@MSBI32400Lab1:~$ tree /data/lab3
/data/lab3
├── bin
├── data
├── doc
├── results
└── src

5 directories, 0 files
student@MSBI32400Lab1:~$
```

- 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)

MSBI 32400 Lab 3 1/23/2019

Let's get the genes from file 9.5

4

- Head the document to see structure

```
student@MSBI32400Lab1:/data/lab3/data$ head WebDocument_9-05_101AutismPanel.bed.txt
chr1 71868625 72748417 NEGR1
chr1 187682629 188026000 NTNG1
chr1 86115106 86174116 ZNHIT6
chr10 89622870 89731687 PTEN
chr11 27676440 27743605 BDNF
chr11 71139239 71163914 DHCR7
chr11 71908602 71907345 FOLR1
chr11 126293254 126873355 KIRREL3
chr11 70313961 70963623 SHANK2
chr12 63539014 63544722 AVPR1A
student@MSBI32400Lab1:/data/lab3/data$
```

- 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

MSBI 32400 Lab 3 1/23/2019

5

```

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
KIFREL3
SHANK2
AVPR1A
CACNA1C
GRIN2B
PTPN11
SOX5
PCDH9
CHD8
FOXG1
GABRB3
SNRPN
UBE3A
ANKRD11
CREBBP
KCTD13
RFXO1
TSC2
PAFAH1B1
RAI1
SLC6A4
SMG6
C18orf1
KATNAL2
TCF4
PIK3P
ZNF507
CNTNAP5
DPP10
MBD5
NRXN1
SATB2
SCN1A
SCN2A
SPAST
ZEB2
--More--

```

MSBI 32400 Lab 3 1/23/2019

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**

assembly: Dec. 2013 (GRCh38/hg38) ▾
 Dec. 2013 (GRCh38/hg38)
 Feb. 2009 (GRCh37/hg19)
 Mar. 2006 (NCBI36/hg18)
 May 2004 (NCBI35/hg17)
 July 2003 (NCBI34/hg16)

output format: all fields from selected table ▾
 output file: all fields from selected table
 selected fields from primary and related tables
 file type returned: sequence
 GTF - gene transfer format
 CDS FASTA alignment from multiple alignment
 BED - browser extensible data
 custom track
 To reset all user hyperlinks to Genome Browser

output format: BED - browser extensible data ▾ Send output to
 output file: 101AutismGenelistExons.bed (leave blank to keep output in
 file type returned: ☒ plain text ☐ gzip compressed
 get output summary/statistics
 To reset all user cart settings (including custom tracks), [click here](#).

MSBI 32400 Lab 3 1/23/2019

Change output type to exon + 10bp

7

Output knownGene as BED

☐ Include custom track header:

name=

description=

visibility=

url=

Create one BED record per:

☐ Whole Gene

☐ Upstream by bases

☒ Exons plus bases at each end

☐ Introns plus bases at each end

☐ 5' UTR Exons

☐ Coding Exons

☐ 3' UTR Exons

☐ Downstream by bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream past the edge of the chromosome.

MSBI 32400 Lab 3 1/23/2019

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 341767885 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]$

```

MSBI 32400 Lab 3 1/23/2019

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!

MSBI 32400 Lab 3 1/23/2019

BED file has overlapping transcripts

10

```

student@MSBI32400Lab1:~$ cat bed
chr1 215740711 215741121 NM_001319294 exon 8 chr1 215740722 f 0 +
chr1 215747118 215747192 NM_001319294 exon 1 chr1 215747129 f 0 +
chr1 215747405 215747471 NM_001319294 exon 2 chr1 215747416 f 0 +
chr1 215748233 215749327 NM_001319294 exon 3 chr1 215748244 f 0 +
chr1 215751802 215751881 NM_001319294 exon 4 chr1 215751813 f 0 +
chr1 215751333 215751434 NM_001319294 exon 5 chr1 215751344 f 0 +
chr1 215752332 215752490 NM_001319294 exon 6 chr1 215752343 f 0 +
chr1 215753241 215753352 NM_001319294 exon 7 chr1 215753252 f 0 +
chr1 215759827 215760838 NM_001319294 exon 8 chr1 215759838 f 0 +
chr1 215768687 215768823 NM_001319294 exon 9 chr1 215768698 f 0 +
chr1 215773198 215773386 NM_001319294 exon 10 chr1 215773209 f 0 +
chr1 215775416 215775553 NM_001319294 exon 11 chr1 215775427 f 0 +
chr1 215777463 215777604 NM_001319294 exon 12 chr1 215777474 f 0 +
chr1 215781348 215781524 NM_001319294 exon 13 chr1 215781359 f 0 +
chr1 215781556 215781524 NM_001319294 exon 14 chr1 215781567 f 0 +
chr1 215792217 215792422 NM_001319294 exon 15 chr1 215792228 f 0 +
chr1 215792498 215792643 NM_001319294 exon 16 chr1 215792501 f 0 +
chr1 215793388 215793513 NM_001319294 exon 17 chr1 215793399 f 0 +
chr1 215740711 215741121 NM_001319295 exon 8 chr1 215740722 f 0 +
chr1 215747118 215747192 NM_001319295 exon 1 chr1 215747129 f 0 +
chr1 215747405 215747471 NM_001319295 exon 2 chr1 215747416 f 0 +
chr1 215748233 215749327 NM_001319295 exon 3 chr1 215748244 f 0 +
chr1 215751802 215751881 NM_001319295 exon 4 chr1 215751813 f 0 +
chr1 215751333 215751434 NM_001319295 exon 5 chr1 215751344 f 0 +
chr1 215752332 215752490 NM_001319295 exon 6 chr1 215752343 f 0 +
chr1 215753241 215753352 NM_001319295 exon 7 chr1 215753252 f 0 +
chr1 215759827 215760838 NM_001319295 exon 8 chr1 215759838 f 0 +
chr1 215768687 215768823 NM_001319295 exon 9 chr1 215768698 f 0 +
chr1 215773198 215773386 NM_001319295 exon 10 chr1 215773209 f 0 +
chr1 215775416 215775553 NM_001319295 exon 11 chr1 215775427 f 0 +
chr1 215777463 215777604 NM_001319295 exon 12 chr1 215777474 f 0 +
chr1 215781348 215781524 NM_001319295 exon 13 chr1 215781359 f 0 +
chr1 215781556 215781524 NM_001319295 exon 14 chr1 215781567 f 0 +
chr1 215792217 215792422 NM_001319295 exon 15 chr1 215792228 f 0 +
chr1 215792498 215792643 NM_001319295 exon 16 chr1 215792501 f 0 +
chr1 215793388 215793513 NM_001319295 exon 17 chr1 215793399 f 0 +
chr1 215740711 215741121 NM_016121 exon 8 chr1 215740722 f 0 +
chr1 215747118 215747192 NM_016121 exon 1 chr1 215747129 f 0 +
chr1 215747405 215747471 NM_016121 exon 2 chr1 215747416 f 0 +
chr1 215748233 215749327 NM_016121 exon 3 chr1 215748244 f 0 +
chr1 215751802 215751881 NM_016121 exon 4 chr1 215751813 f 0 +
chr1 215751333 215751434 NM_016121 exon 5 chr1 215751344 f 0 +
chr1 215752332 215752490 NM_016121 exon 6 chr1 215752343 f 0 +
chr1 215753241 215753352 NM_016121 exon 7 chr1 215753252 f 0 +

```

MSBI 32400 Lab 3 1/23/2019

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)

MSBI 32400 Lab 3 1/23/2019

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**

```

student@MSBI32400Lab1:~/data/lab3/data$ head 101AutismGenelistExons_sort.bed
chr1 71868614 71873263 uc010oqs.2 exon 0 10 chr1 71868625 r 0 -
chr1 71868614 71873263 uc001dfv.3 exon 0 10 chr1 71868625 r 0 -
chr1 71868614 71873263 uc001dfv.3 exon 0 10 chr1 71868625 r 0 -
chr1 72058489 72058661 uc010oqs.2 exon 1 10 chr1 72058500 r 0 -
chr1 72058489 72058661 uc001dfv.3 exon 1 10 chr1 72058500 r 0 -
chr1 72058489 72058661 uc001dfv.3 exon 1 10 chr1 72058500 r 0 -
chr1 72076098 72076839 uc010oqs.2 exon 2 10 chr1 72076709 r 0 -
chr1 72076098 72076839 uc001dfv.3 exon 2 10 chr1 72076709 r 0 -
chr1 72076098 72076839 uc001dfv.3 exon 2 10 chr1 72076709 r 0 -
chr1 72163686 72163832 uc001dfv.3 exon 3 10 chr1 72163691 r 0 -
student@MSBI32400Lab1:~/data/lab3/data$ bedtools merge -c 4 -o collapse -i 101AutismGenelistExons_sort.bed > 101AutismGenelistExons_sort_merged.bed
student@MSBI32400Lab1:~/data/lab3/data$ head 101AutismGenelistExons_sort_merged.bed
chr1 71868614 71873263 uc010oqs.2 exon 0 10 chr1 71868625 r,uc001dfv.3 exon 0 10 chr1 71868625 r
chr1 72058489 72058661 uc010oqs.2 exon 1 10 chr1 72058500 r,uc001dfv.3 exon 1 10 chr1 72058500 r
chr1 72076098 72076839 uc010oqs.2 exon 2 10 chr1 72076709 r,uc001dfv.3 exon 2 10 chr1 72076709 r
chr1 72163686 72163832 uc001dfv.3 exon 3 10 chr1 72163691 r,uc001dfv.3 exon 3 10 chr1 72163691 r
chr1 72241844 72241990 uc001dfv.3 exon 4 10 chr1 72241855 r,uc001dfv.3 exon 4 10 chr1 72241855 r,uc010oqs.2 exon 3 10 chr1 72241855 r
chr1 72480751 72481094 uc001dfv.3 exon 5 10 chr1 72480762 r,uc001dfv.3 exon 5 10 chr1 72480762 r,uc010oqs.2 exon 4 10 chr1 72480762 r
chr1 72566417 72566624 uc001dfv.3 exon 6 10 chr1 72566428 r
chr1 72747991 72748415 uc001dfv.3 exon 6 10 chr1 72748002 r,uc010oqs.2 exon 5 10 chr1 72748002 r
chr1 86115995 86115991 uc001dlh.3 exon 0 10 chr1 86115106 r,uc010osc.2 exon 0 10 chr1 86115106 r
chr1 86125519 86123664 uc001dlh.3 exon 1 10 chr1 86125538 r,uc010osc.2 exon 1 10 chr1 86125538 r
student@MSBI32400Lab1:~/data/lab3/data$

```

MSBI 32400 Lab 3 1/23/2019

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

MSBI 32400 Lab 3 1/23/2019

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)

MSBI 32400 Lab 3 1/23/2019

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

MSBI 32400 Lab 3 1/23/2019

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 1/2 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

MSBI 32400 Lab 3 1/23/2019

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.

MSBI 32400 Lab 3 1/23/2019

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

MSBI 32400 Lab 3 1/23/2019