© David Deen

Day 5: TDF visualization & Assessment

2022/07/15
Daniel Ramírez
DnA Lab
University of Colorado Boulder, BioFrontiers

What you have learned so far on week 1 …

- Working on a unix-like command terminal.
- Connecting to computer cluster (AWS or BioFrontiers Fiji).
- Processing high throughput sequencing files (FASTQ) to obtain a reference genome aligned/mapped SAM/BAM files.

Today:

Part 1 (1 hour): Visualizing mapped files using TDF instead of BAM.

Part 2 (2 hours): Assessment of skills learned until today.

Day 5 - Part 1
(1 hour)

Visualizing mapped files using TDF instead of BAM

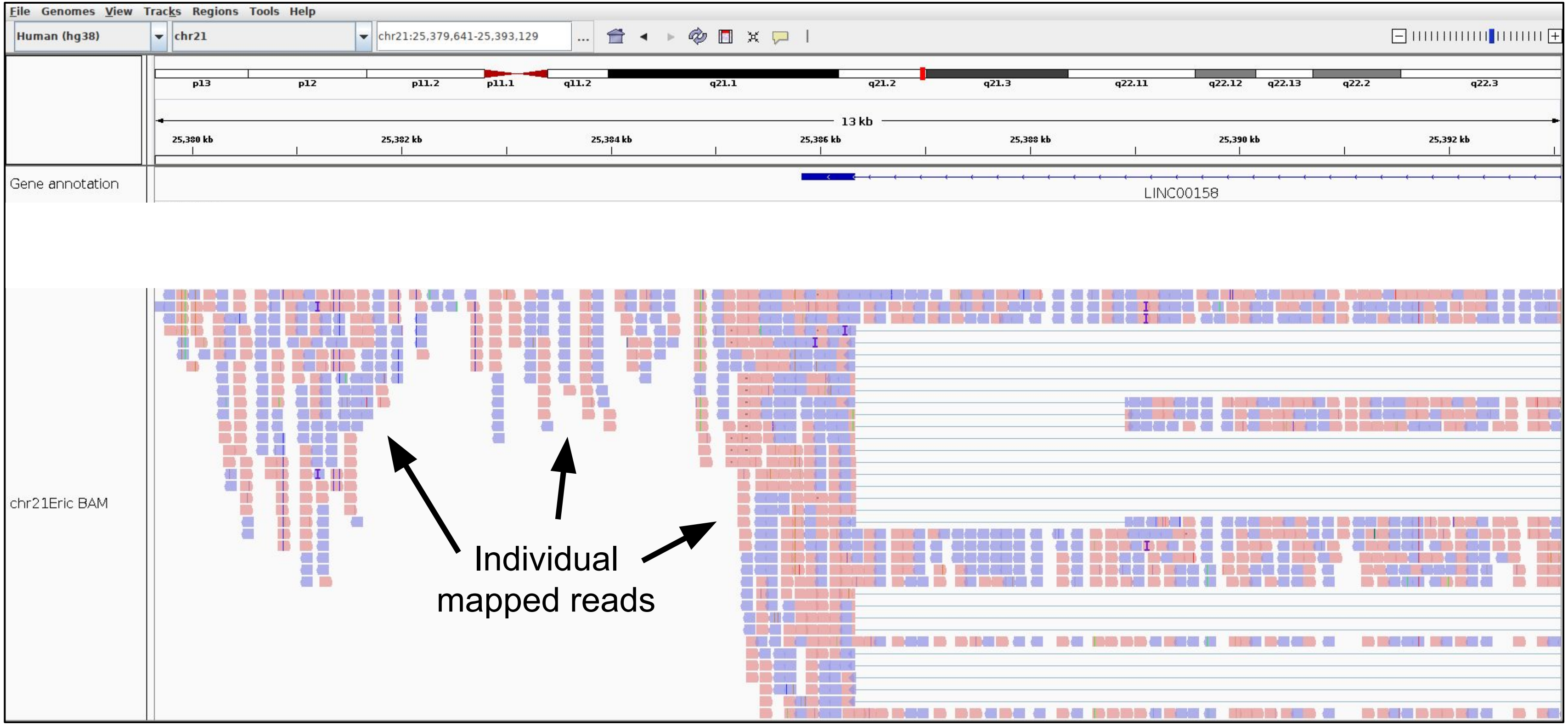
Comparing BAM and TDF files on IGV

Observed region on screen = **13 kb**

annotation

TDF

BAM



Individual mapped reads

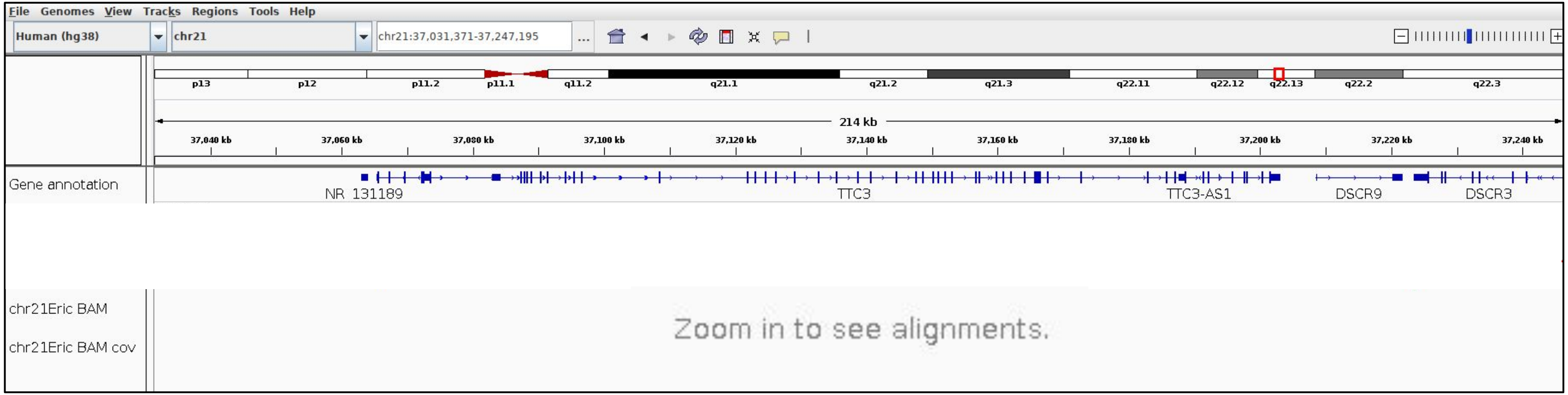
Comparing BAM and TDF files on IGV

Observed region on screen = **214 kb**

annotation

TDF

BAM



Human (hg38) | chr21 | chr21:37,031,371-37,247,195

Zoom in to see alignments.

IGV does not display individually mapped reads on such a big region at once!

But IGV does okay displaying TDF coverage across any zoom region.

Obtaining “per million” scaling factor to normalize

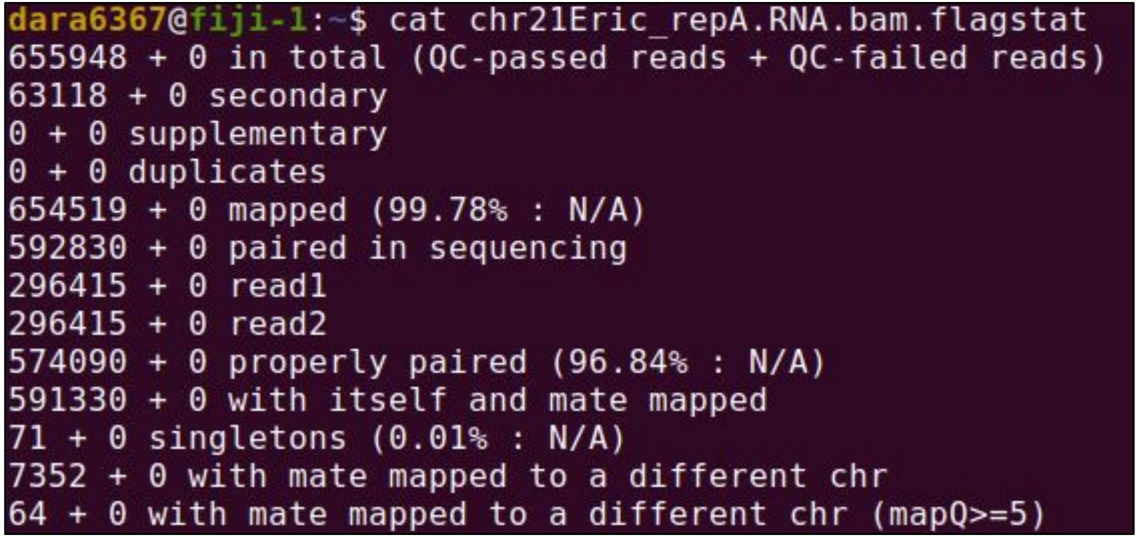
BAM file had a **63,118** paired-reads mapped flagged as secondary alignments.

BAM file had a total of **654,519** paired-reads mapped.

Total primary paired-reads mapped are
654,519 - 63,118 = **591,401**

“Per million” factor is then
591,401 / 1,000,000 = **0.591401**

chr21Eric_repA.RNA.bam.flagstat

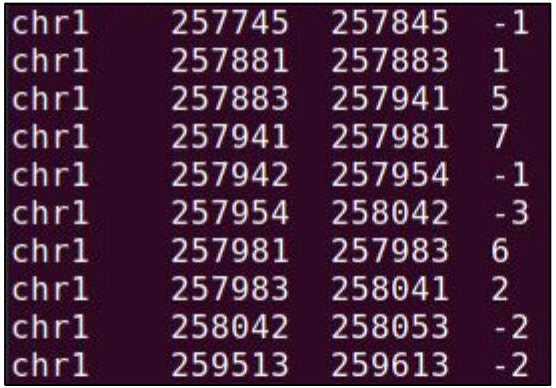


```
dara6367@fiji-1:~$ cat chr21Eric_repA.RNA.bam.flagstat
655948 + 0 in total (QC-passed reads + QC-failed reads)
63118 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
654519 + 0 mapped (99.78% : N/A)
592830 + 0 paired in sequencing
296415 + 0 read1
296415 + 0 read2
574090 + 0 properly paired (96.84% : N/A)
591330 + 0 with itself and mate mapped
71 + 0 singletons (0.01% : N/A)
7352 + 0 with mate mapped to a different chr
64 + 0 with mate mapped to a different chr (mapQ>=5)
```

Normalize read coverage by sequencing depth

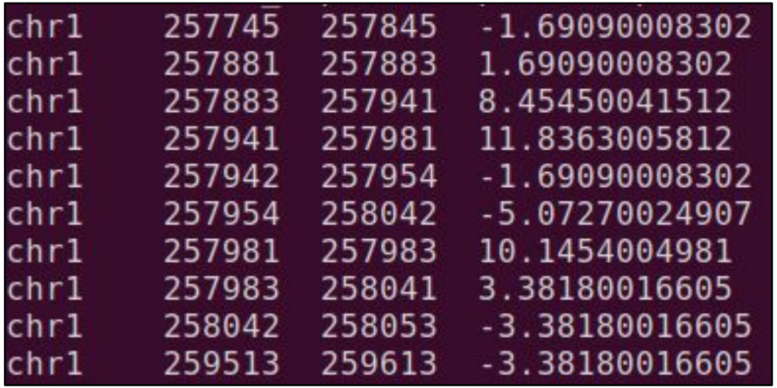
4th column in BedGraph file contains the read density information
for the interval denoted in columns 1st, 2nd, and 3rd

chr21Eric_repA.RNA.BedGraph



```
chr1     257745  257845  -1
chr1     257881  257883  1
chr1     257883  257941  5
chr1     257941  257981  7
chr1     257942  257954  -1
chr1     257954  258042  -3
chr1     257981  257983  6
chr1     257983  258041  2
chr1     258042  258053  -2
chr1     259513  259613  -2
```

chr21Eric_repA.RNA.**mp**.BedGraph



```
chr1     257745  257845  -1.69090008302
chr1     257881  257883  1.69090008302
chr1     257883  257941  8.45450041512
chr1     257941  257981  11.8363005812
chr1     257942  257954  -1.69090008302
chr1     257954  258042  -5.07270024907
chr1     257981  257983  10.1454004981
chr1     257983  258041  3.38180016605
chr1     258042  258053  -3.38180016605
chr1     259513  259613  -3.38180016605
```

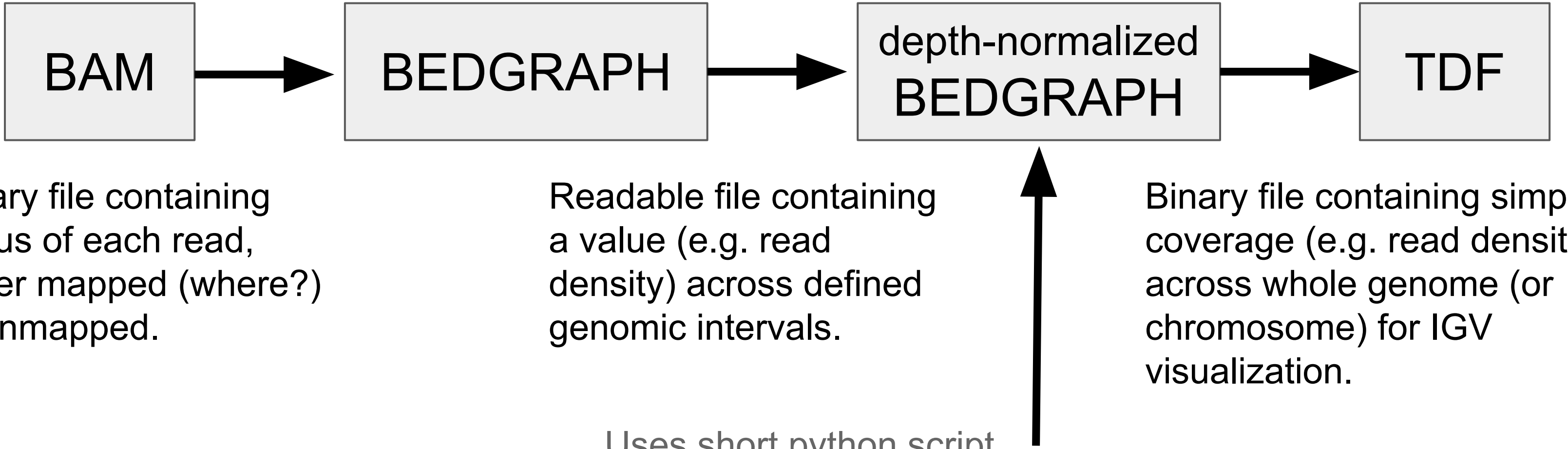
÷ by “per million” factor (0.591401)

Divide each row entry by the “per million” scaling factor to normalize
BedGraph to account for total number of mapped reads (or library depth)

Normalize read coverage by sequencing depth

What does this script do?

d5-bam-to-tdf-pairedend.sbatch



BAM → BEDGRAPH → depth-normalized BEDGRAPH → TDF

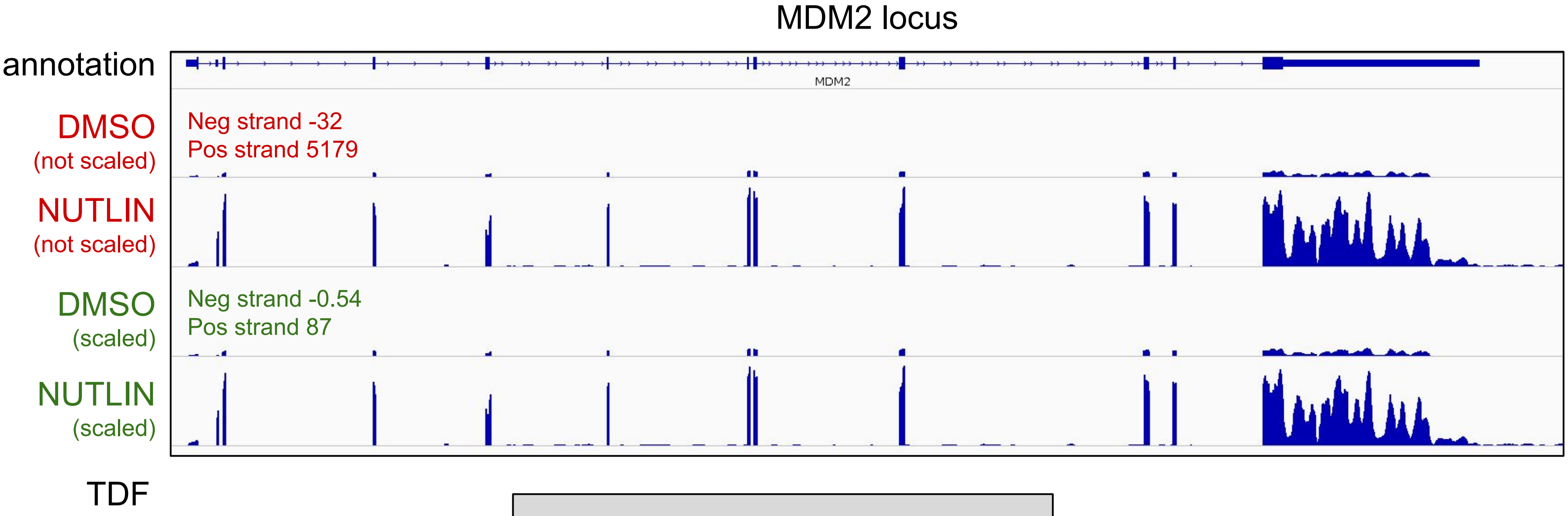
Binary file containing status of each read, either mapped (where?) or unmapped.

Readable file containing a value (e.g. read density) across defined genomic intervals.

Binary file containing simply coverage (e.g. read density) across whole genome (or chromosome) for IGV visualization.

Uses short python script readcount_corrected_geneomeBedgraphs.py

Comparing TDFs: not scaled vs scaled datasets



“per million mapped” scaling factor

DMSO dataset = 50.7

Nutlin dataset = 59.2

Day 5 - Part 2
(2 hours)

Assessment of skills learned from Day 2 to Day 5

What have you learned this week?

Let’s work on the next 2 hours (10 am - 12 pm) on processing new high-throughput sequencing datasets.

Check out available FASTQ datasets on Day5 files.
(Extra-very-real-points: Do all files with a loop!)

START FASTQ ➡ SAM ➡ BAM ➡
➡ BEDGRAPH ➡ scaled BEDGRAPH ➡ TDF FINISH