Day 5: Looping, visualization & Assessment

2023/07/28

Samuel Hunter
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

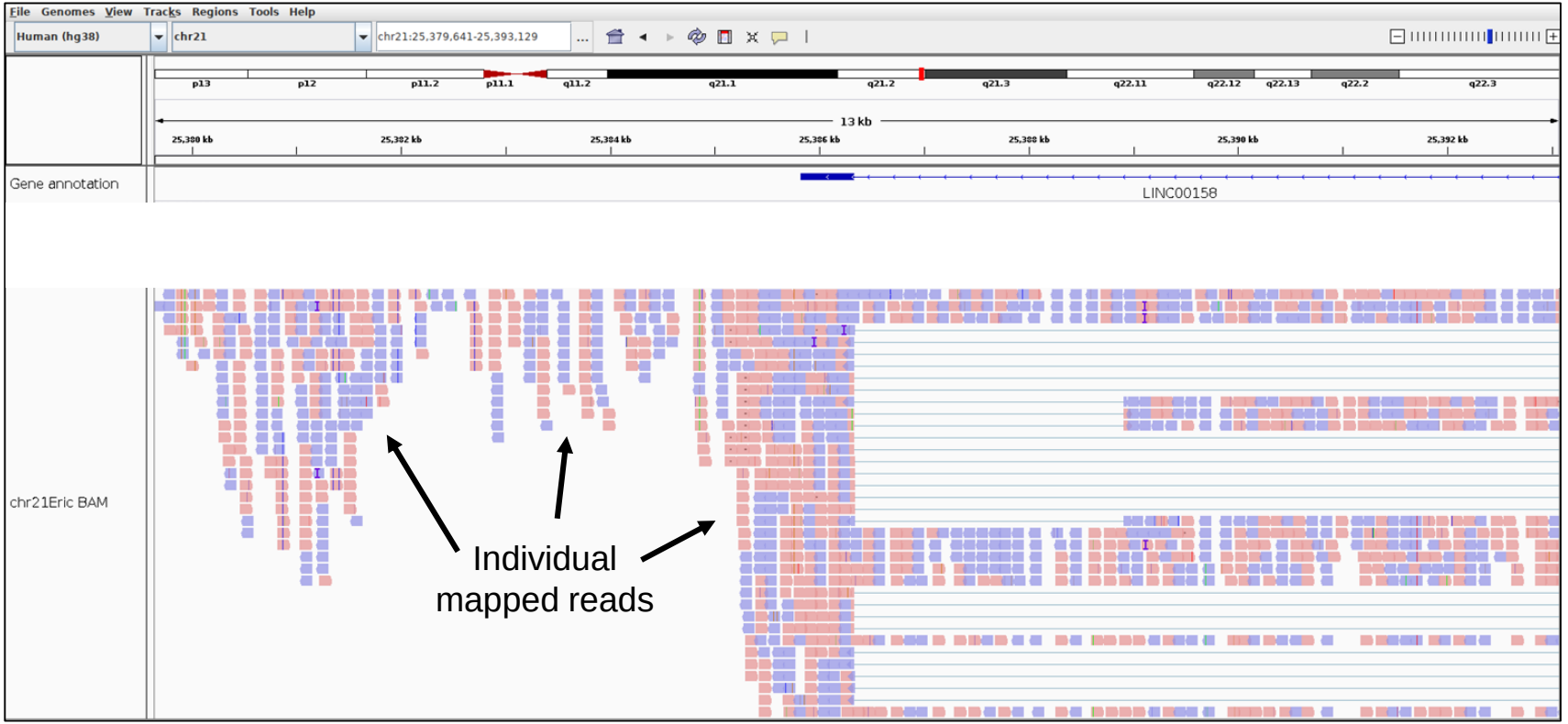
Goals for today:

Part 1 (1 hour): Create BedGraphs and TDFs for visualization (For loops if we have time)

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

Comparing BAM and TDF files on IGV

Observed region on screen = **13 kb**



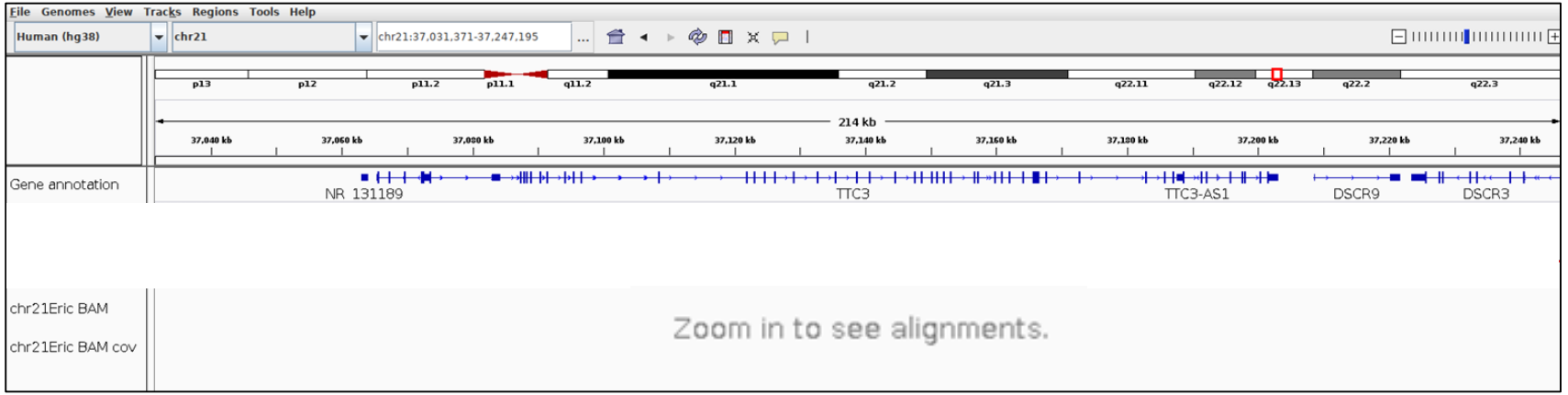
Comparing BAM and TDF files on IGV

Observed region on screen = **214 kb**

annotation

TDF

BAM

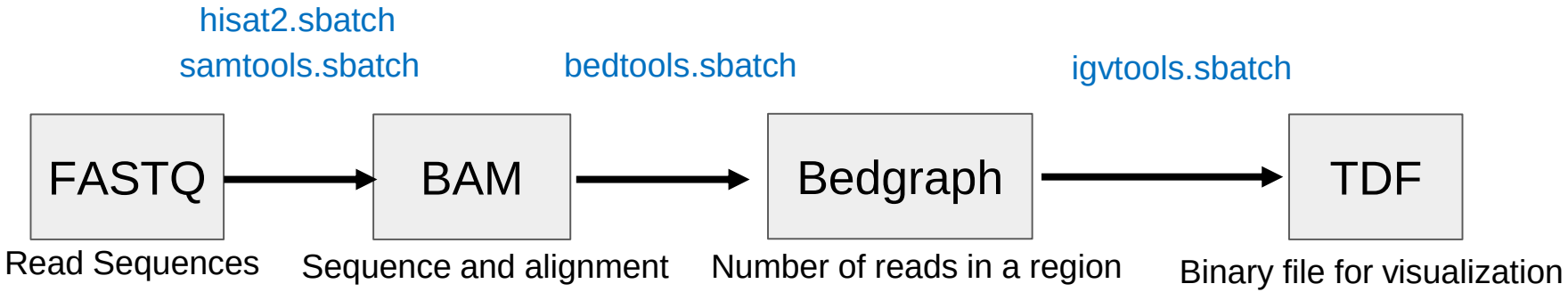


IGV does not display individually mapped reads on such a big region at once!

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

Go to the Day5 worksheet on Github, and get started on today’s exercise

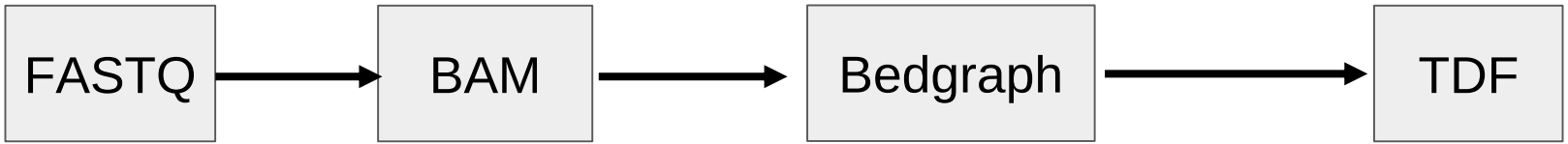
Each dataset is processed the same way, so that we end up with usable files for analysis



But for each FASTQ file, I'll need to edit all of my variable names and submit the jobs individually!

For Loops in a nutshell:

For every dataset in (my/directory/of/fastqs)
Do



Done!

We feed each FASTQ file into the scripts without needing to submit the jobs individually or manually edit the variable names

Day 5: Assessment

Check out the other FASTQ datasets in Day5.

Imagine these files are fresh off the sequencer… can you write your own scripts to turn them from raw FASTQs to TDFs? What QC checks should you perform?

(Extra-very-real-points: Do all files with a loop!)

FASTQ ➞ SAM ➞ BAM ➞
➞ BEDGRAPH ➞ TDF