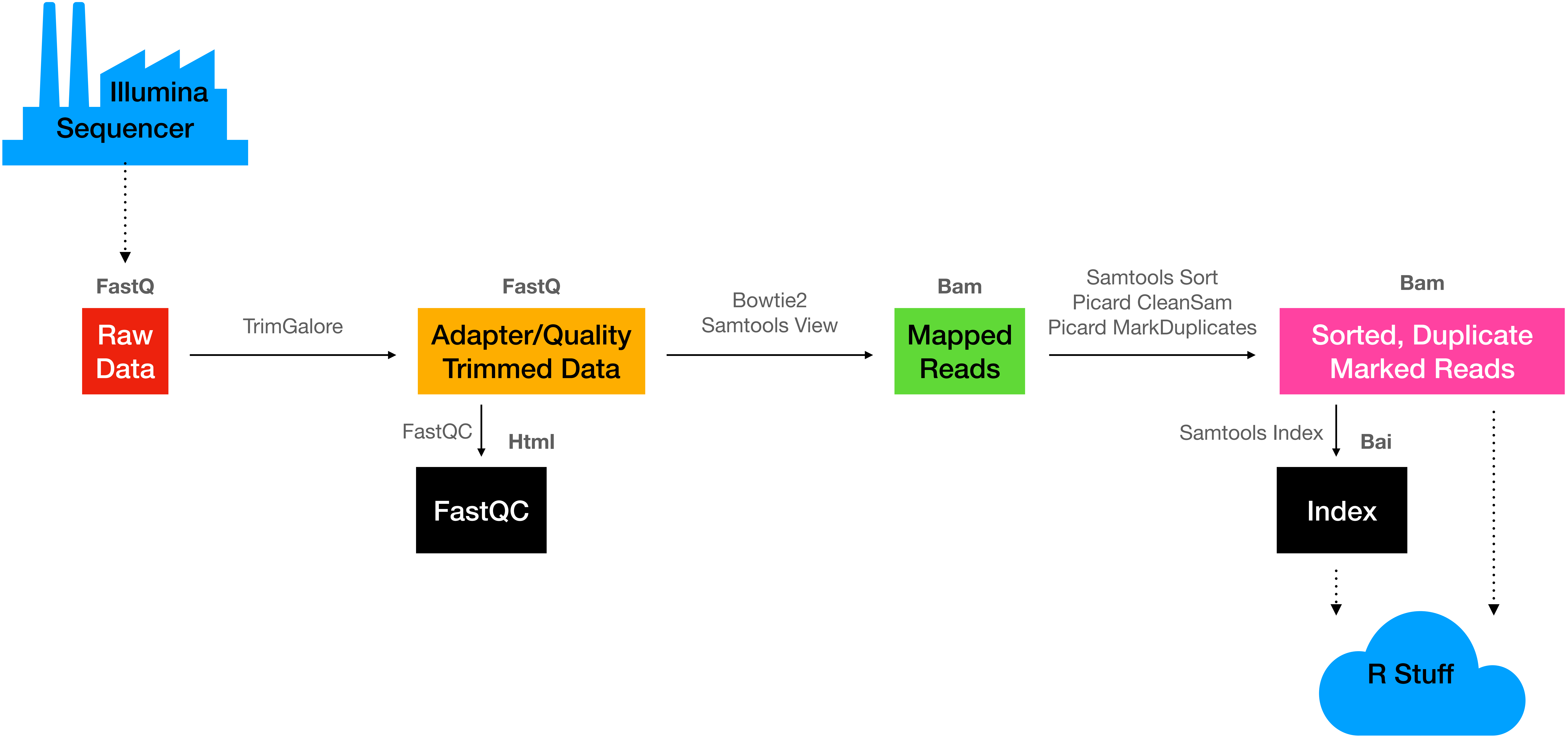
Intro to R for Biologists

IBiS Special Topics, Fall 2021

Class 11: Oct. 28, 2021 R: Genomic Ranges - Plotting Coverage

Erik Andersen and Shelby Blythe

General Workflow



Illumina Sequencer

FastQ

Raw Data

TrimGalore

FastQ

Adapter/Quality Trimmed Data

FastQC

Html

FastQC

Bowtie2
Samtools View

Bam

Mapped Reads

Samtools Sort
Picard CleanSam
Picard MarkDuplicates

Bam

Sorted, Duplicate Marked Reads

Samtools Index

Bai

Index

R Stuff

General Workflow

“R Stuff”

Measurement

Quantification

Taxonomy

.bam

But does my data look good?

Import Reads to GRanges

Coverage/ Wiggle Plot

Regions of Interest

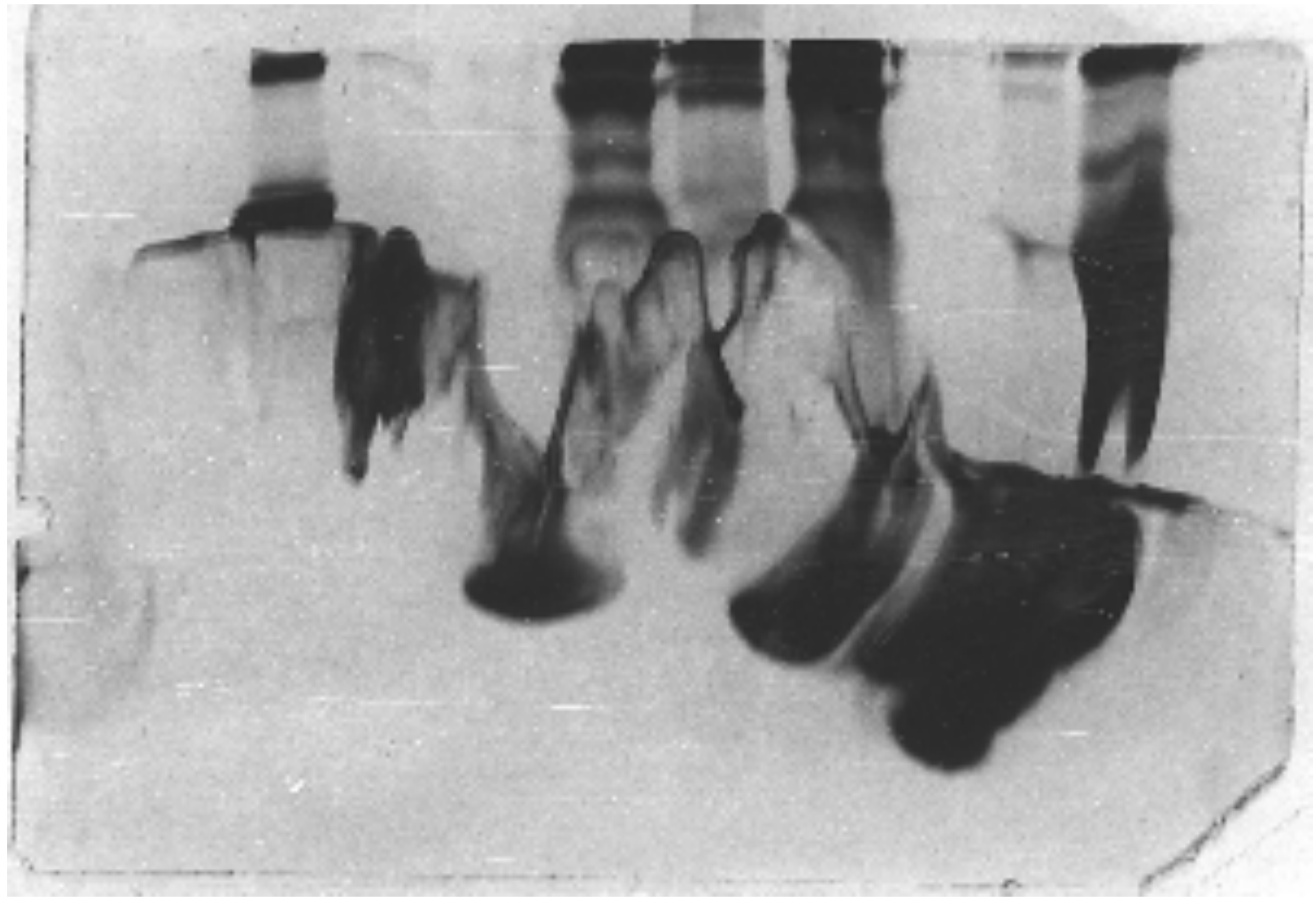
Counts per Region of Interest

Differential Enrichment Analysis

Clustering, Grouping, Categorizing…

For most things, we know how to tell if they “worked”…

We usually rely on the power of observation and evaluating a set of prior expectations.



SDS-PAGE “Hall of Shame”: https://www.ruf.rice.edu/~bioslabs/studies/sds-page/sdsgoofs.html

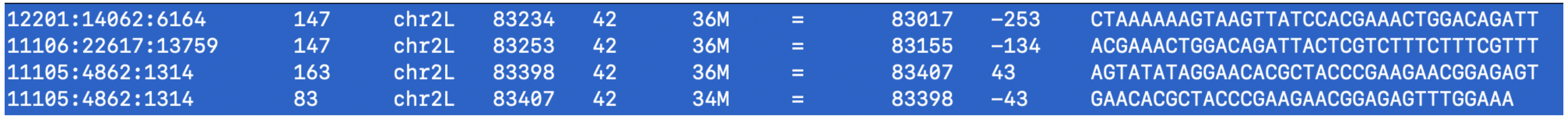
How do you know if your genomics “worked”?

Totally project-dependent. *It is up to you to decide.*

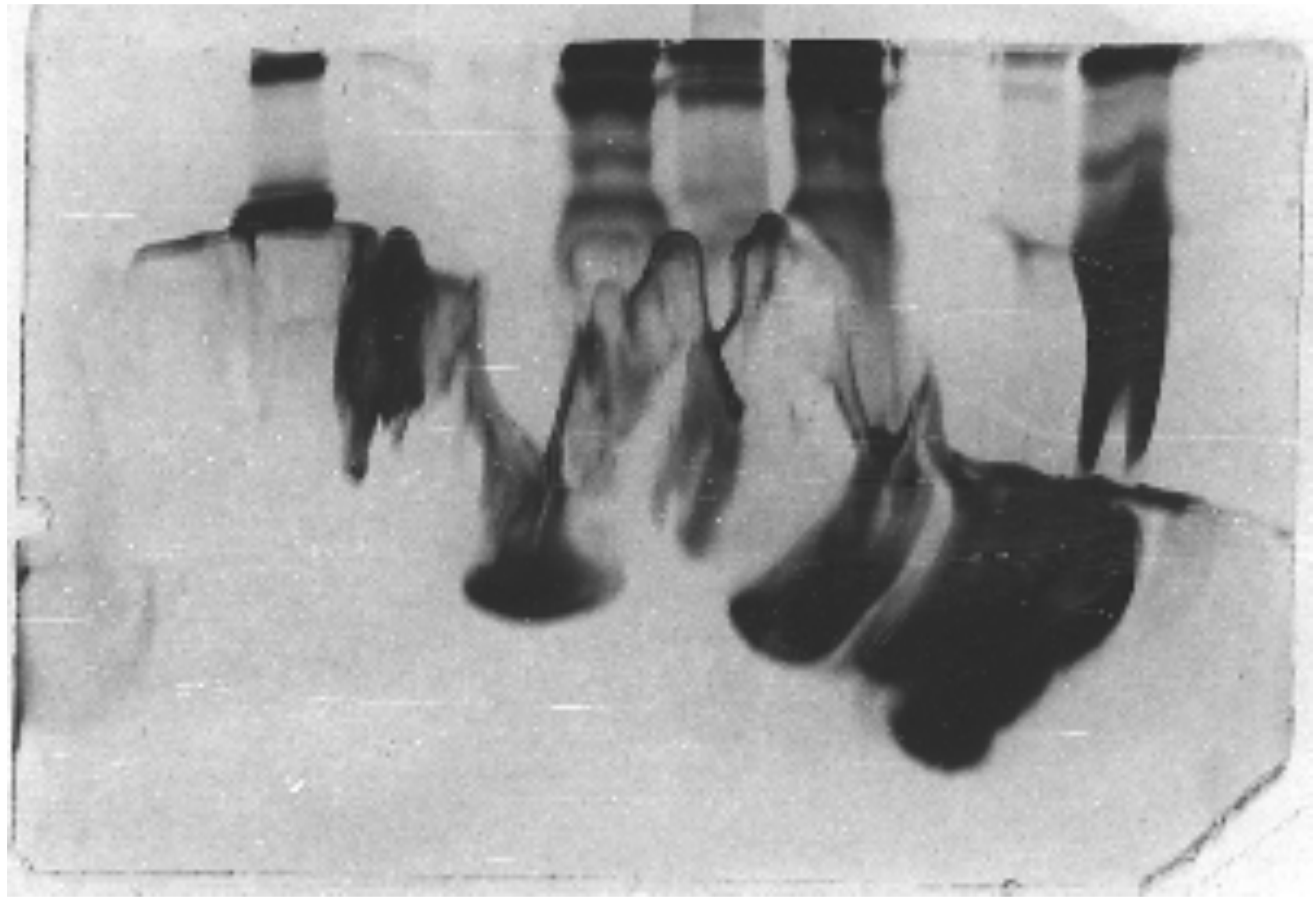
- For a ChIP-seq experiment, you may want to know that you have enrichments in your ChIP (that are not evident in negative controls)…
- For RNA-seq, you may want to know whether the signal is limited (mostly) to genic regions, and whether coverage is uniform or biased…
- **It is important to have a set of prior expectations, and to objectively evaluate them early on in the process.**
  - **Ideally, you know some places to check in the genome that should and should not ‘have stuff going on’.**

How to observe?

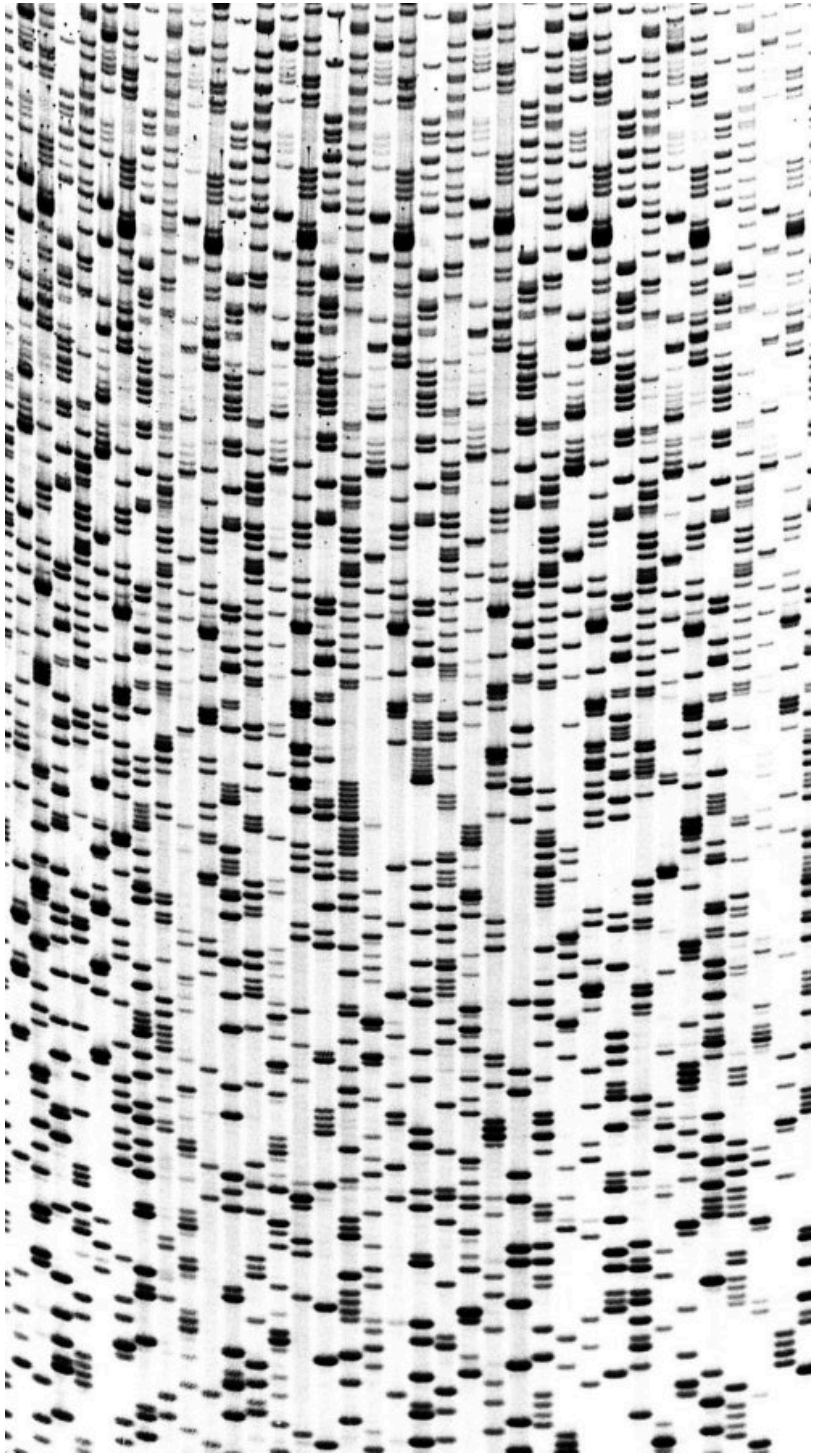
We have millions of lines that look like this…



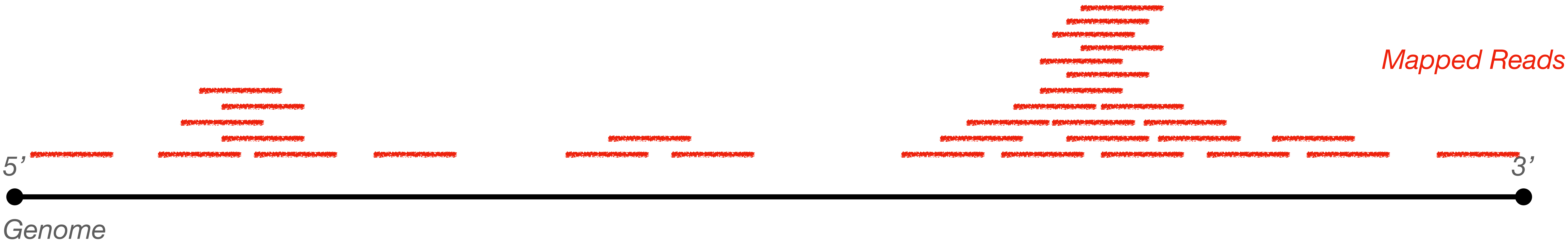
```
12201:14062:6164      147    chr2L   83234   42     36M      =       83017   -253    CTAAAAAAGTAAGTTATCCACGAAACTGGACAGATT
11106:22617:13759     147    chr2L   83253   42     36M      =       83155   -134    ACGAAACTGGACAGATTACTCGTCTTTCTTTCGTTT
11105:4862:1314       163    chr2L   83398   42     36M      =       83407   43      AGTATATAGGAACACGCTACCCGAAGAACGGAGAGT
11105:4862:1314       83     chr2L   83407   42     34M      =       83398   -43     GAACACGCTACCCGAAGAACGGAGAGTTTGGAAA
```



How can you tell if your data “looks like” the gel on the right or the gel on the left?

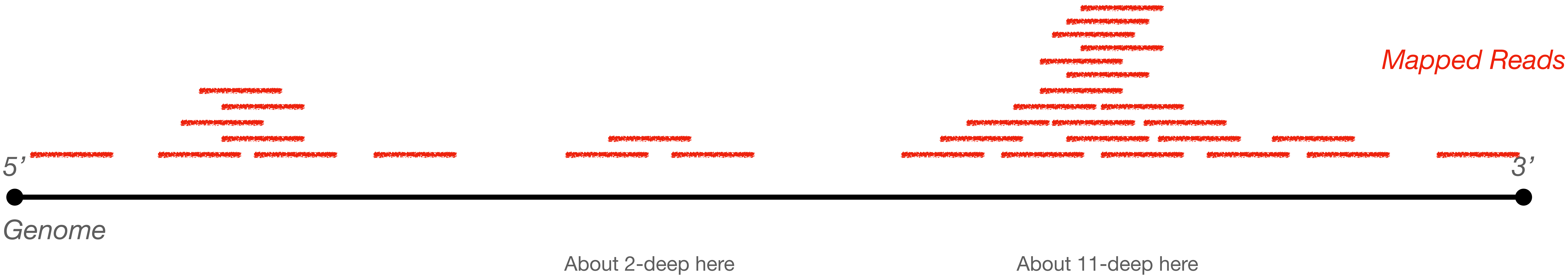


One approach is to plot the coverage depth



- This is something we can have an intuition about…

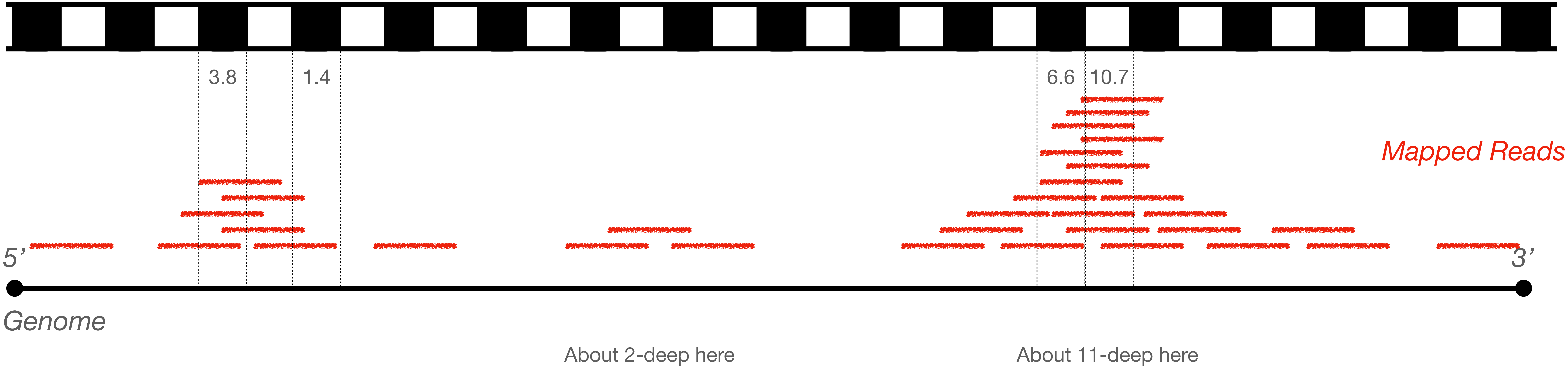
One approach is to plot the coverage depth



- This is something we can have an intuition about…

One approach is to plot the coverage depth

Potential Problem: Genomes are big.



- Solution: rather than calculate coverage on a per-bp basis, first tile the genome into N-bp bins, and then calculate average coverage per bin.

Approach:

- Make a GRanges object with your N-bp genome tiles using the `BSgenome` object as input.

```
GenomicRanges::tileGenome()
```

- Calculate coverage over your .bam file

```
GenomicRanges::coverage()
```

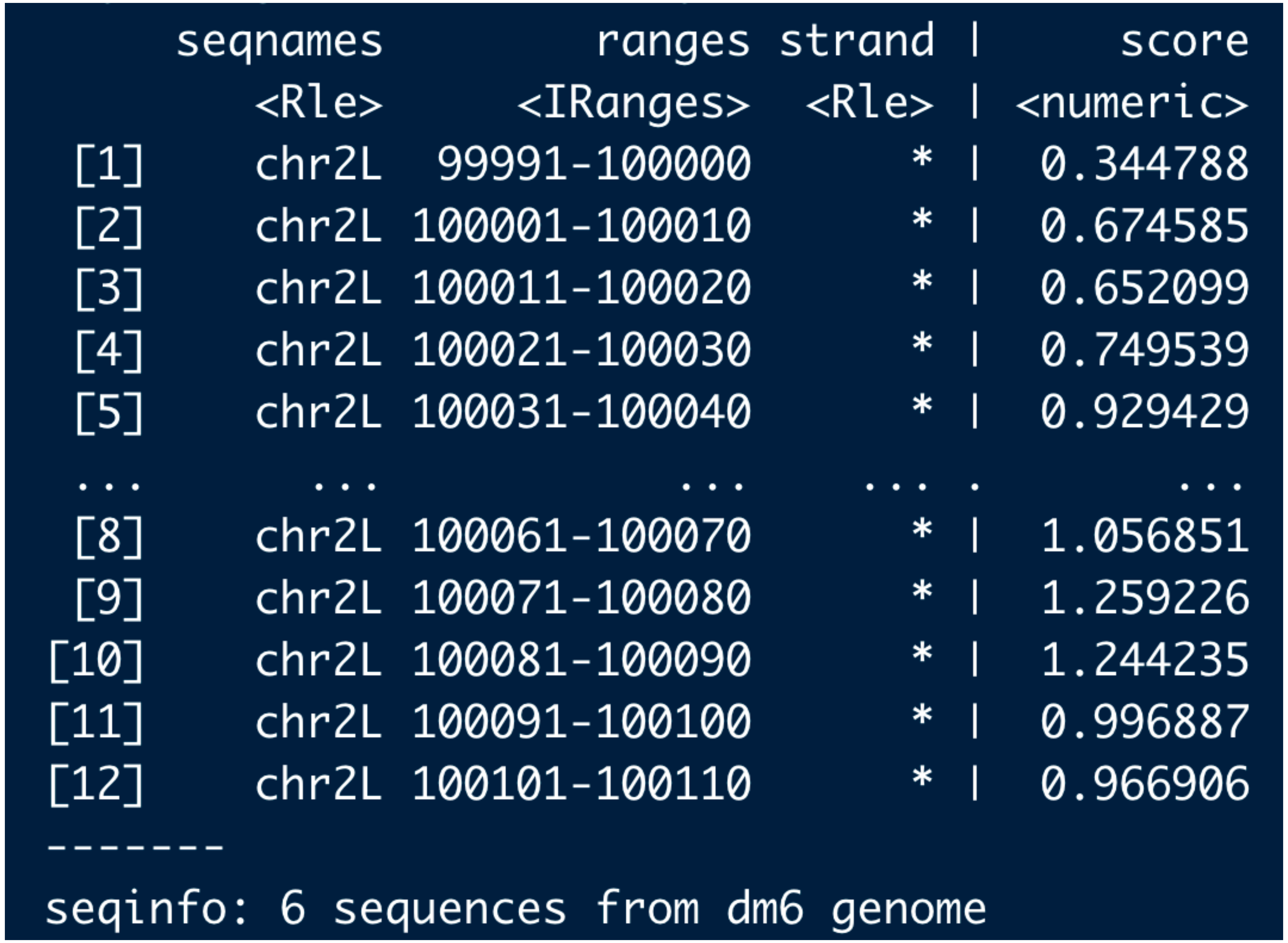
- Calculate the binned average coverage

```
GenomicRanges::binnedAverage()
```

Note, you need to be aware of whether you have gapped or non-gapped data.

This yields something like this:

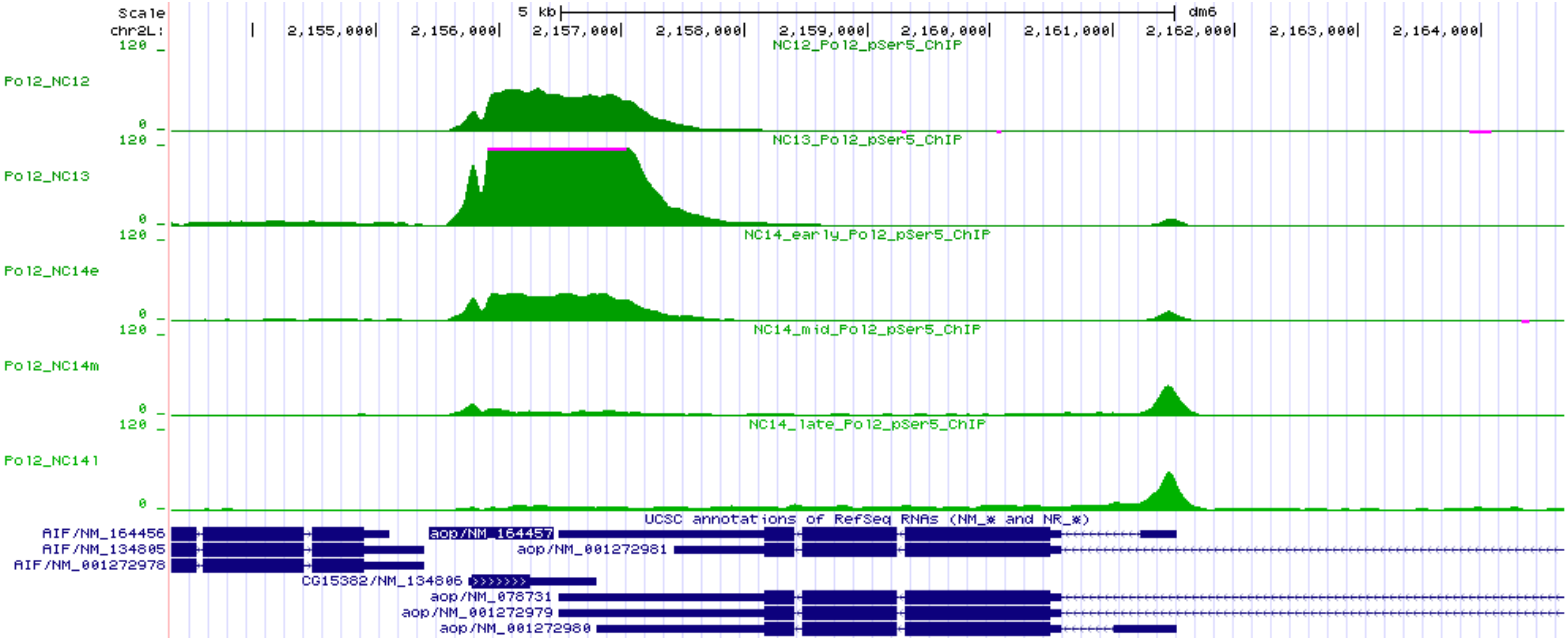
GRanges object with 10-bp bins (could be bigger) + a *score*.



```
      seqnames          ranges strand |     score
         <Rle>       <IRanges>  <Rle> | <numeric>
  [1]    chr2L   99991-100000      * |  0.344788
  [2]    chr2L  100001-100010      * |  0.674585
  [3]    chr2L  100011-100020      * |  0.652099
  [4]    chr2L  100021-100030      * |  0.749539
  [5]    chr2L  100031-100040      * |  0.929429
  ...      ...            ...    ... .       ...
  [8]    chr2L  100061-100070      * |  1.056851
  [9]    chr2L  100071-100080      * |  1.259226
 [10]    chr2L  100081-100090      * |  1.244235
 [11]    chr2L  100091-100100      * |  0.996887
 [12]    chr2L  100101-100110      * |  0.966906
 -------
 seqinfo: 6 sequences from dm6 genome
```

What we want to do is to be able to throw it up on the genome browser

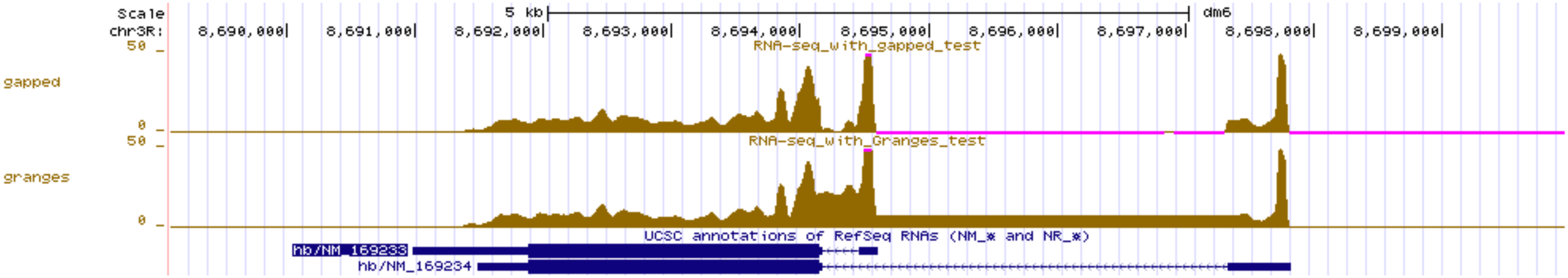
We need to export, and perhaps add a ‘trackline’ for UCSC



Do you mind the gaps?

- When you import a .bam to Genomic Ranges, you lose all information about gaps. This can be OK for certain approaches, like ChIP and ATAC, but can lead to undesirable outcomes with RNA-seq.

- All of the approaches we will cover in today's activity work equally well with `GenomicRanges` and `GappedAlignment` objects.



Same coverage data, but with GAlignments as input (top) and GRanges as input (bottom). There are some cases it could be useful to know where the 'junction straddling' reads are (like if you didn't know where the introns were, or for resolving complex alternative splicing). Usually, however, the bottom approach is not ideal.

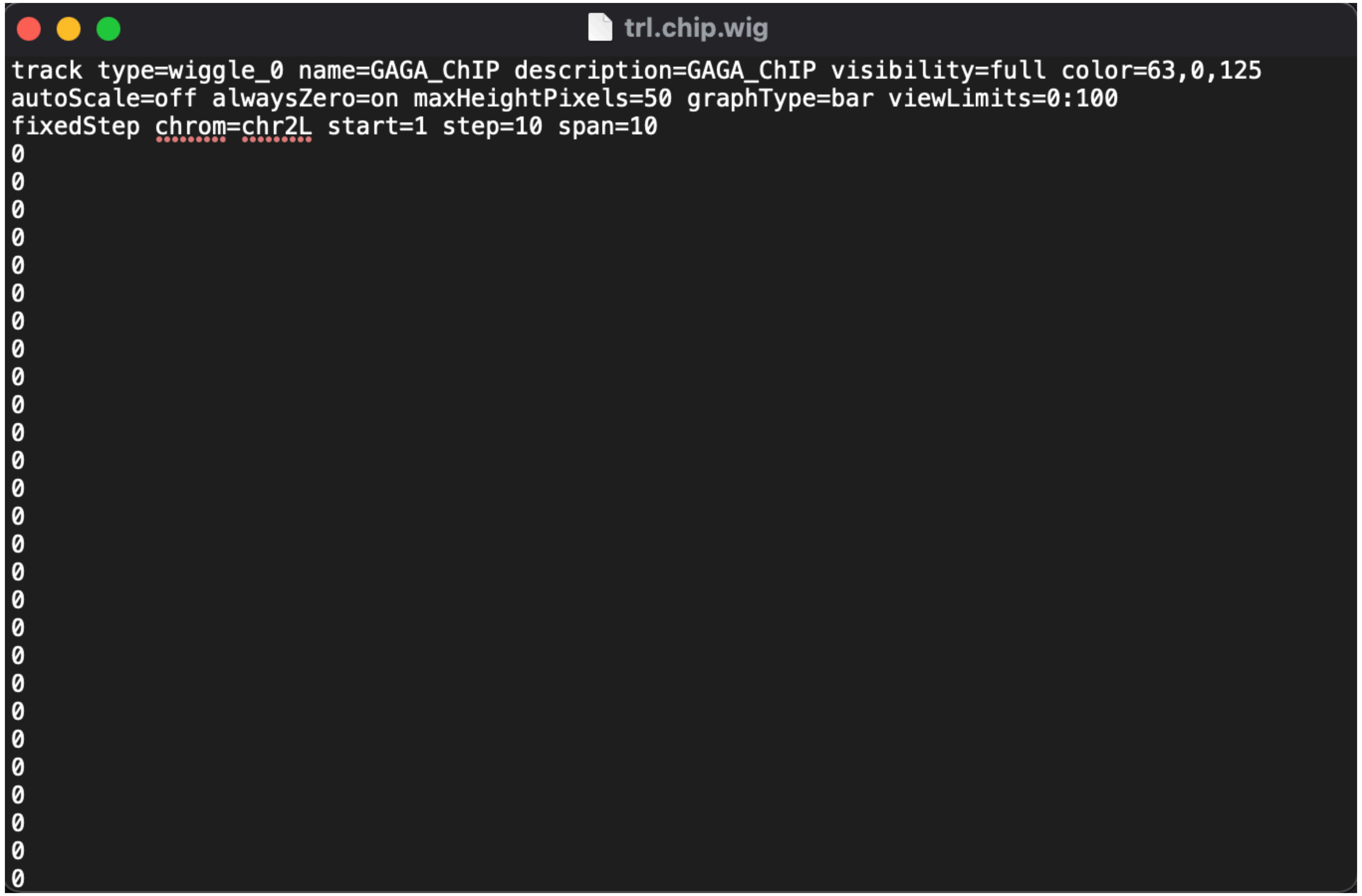
To export, use `rtracklayer::export`

- You need to know what format you want, and this will depend on *where* you want to display your data.
- **Formats:**
  - **wiggle_0** (.wig): uncompressed, fixed width ranges
    - **UCSC**: good but can be slow, needs trackline.
    - **IGV**: good but slow/makes angry
  - **bigWig** (.bigWig or .bw): compressed, fixed width ranges
    - **UCSC**: better but needs URL.
    - **IGV**: ideal.

I prefer UCSC but IGV is good in a pinch.

We often use .wig files for initial data exploration

- Making a trackline for .wig files and appending it to the file using R is a pain in the rear.
- The trackline will tell UCSC how to plot your data, what to name it, and additional options.
- Specifications here:
- https://genome.ucsc.edu/goldenPath/help/wiggle.html



```
track type=wiggle_0 name=GAGA_ChIP description=GAGA_ChIP visibility=full color=63,0,125
autoScale=off alwaysZero=on maxHeightPixels=50 graphType=bar viewLimits=0:100
fixedStep chrom=chr2L start=1 step=10 span=10
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
```

Activity:

- Worked example of calculating binned averages from .bam file to GRanges
- A discussion of export for .wig and .bigWig formats.
  - (Including minor details not covered in the lecture)
- A nifty home-made function that exports a .wig file and automatically adds a custom UCSC trackline.
- **Bonus:** making publication quality plots of coverage using the `Gviz` package.