

BIOL 343

Applied Bioinformatics I

Counting

Dr. Nic Wheeler

Counting converts mapped reads to expression values

Expression values will be used by statistical models to identify DEGs

- Recall the goal of our RNA-seq experiments...
 - Treatment vs Control
 - Mutant vs Wild type
 - *Identify differentially expressed genes (DEGs)*
- DEGs will be identified using statistical tests comparing **expression values** of transcripts/genes
- Expression values will be calculated based on the number of reads that **align/map** to a given genomic locus
- There are different ways to count reads

Counting converts mapped reads to expression values

Expression values will be used by statistical models to identify DEGs

Sample 1

mRNA from
Gene X



9 reads per million

RNA

Library

Reads

Expression values

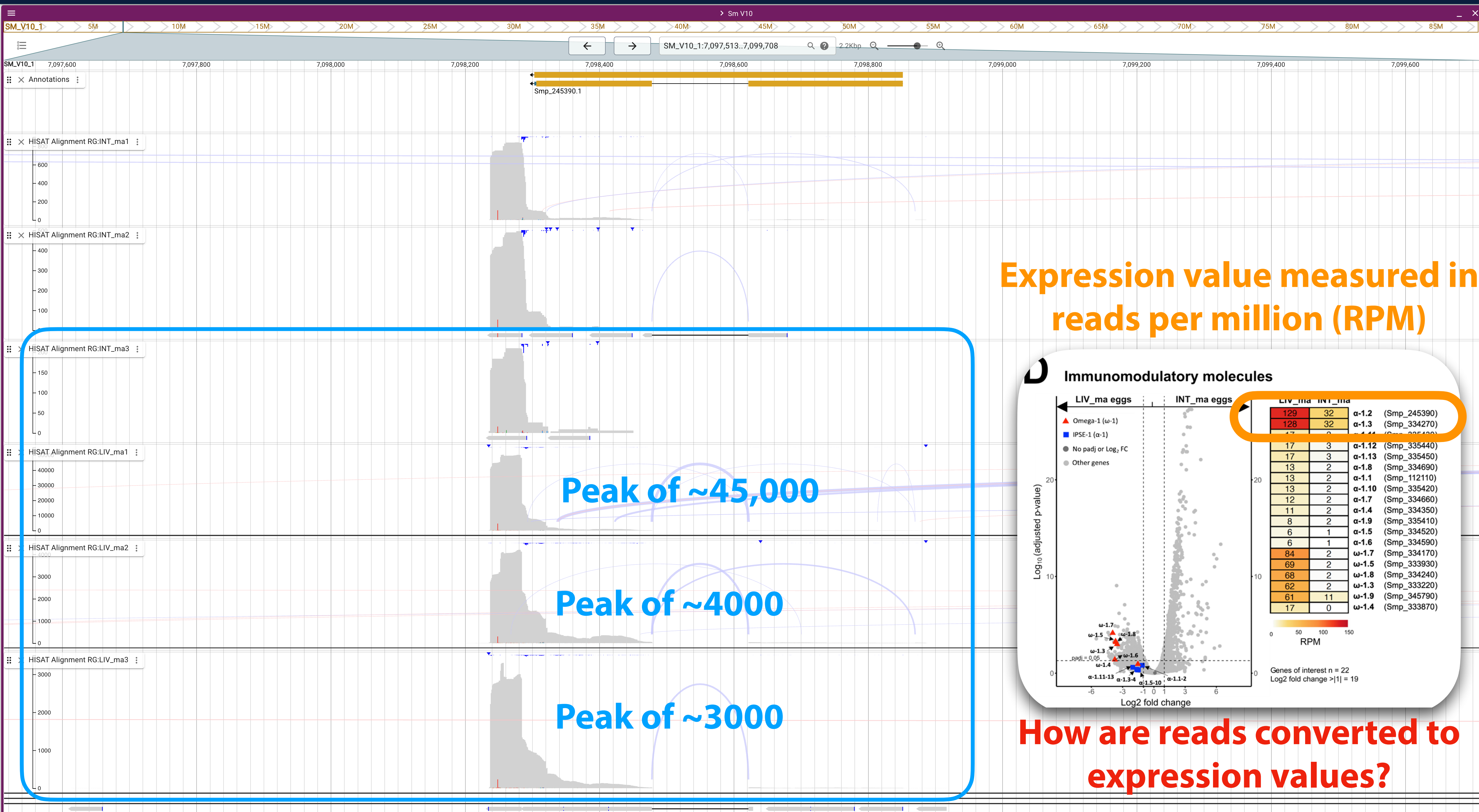
Differentially expressed gene?

Sample 2

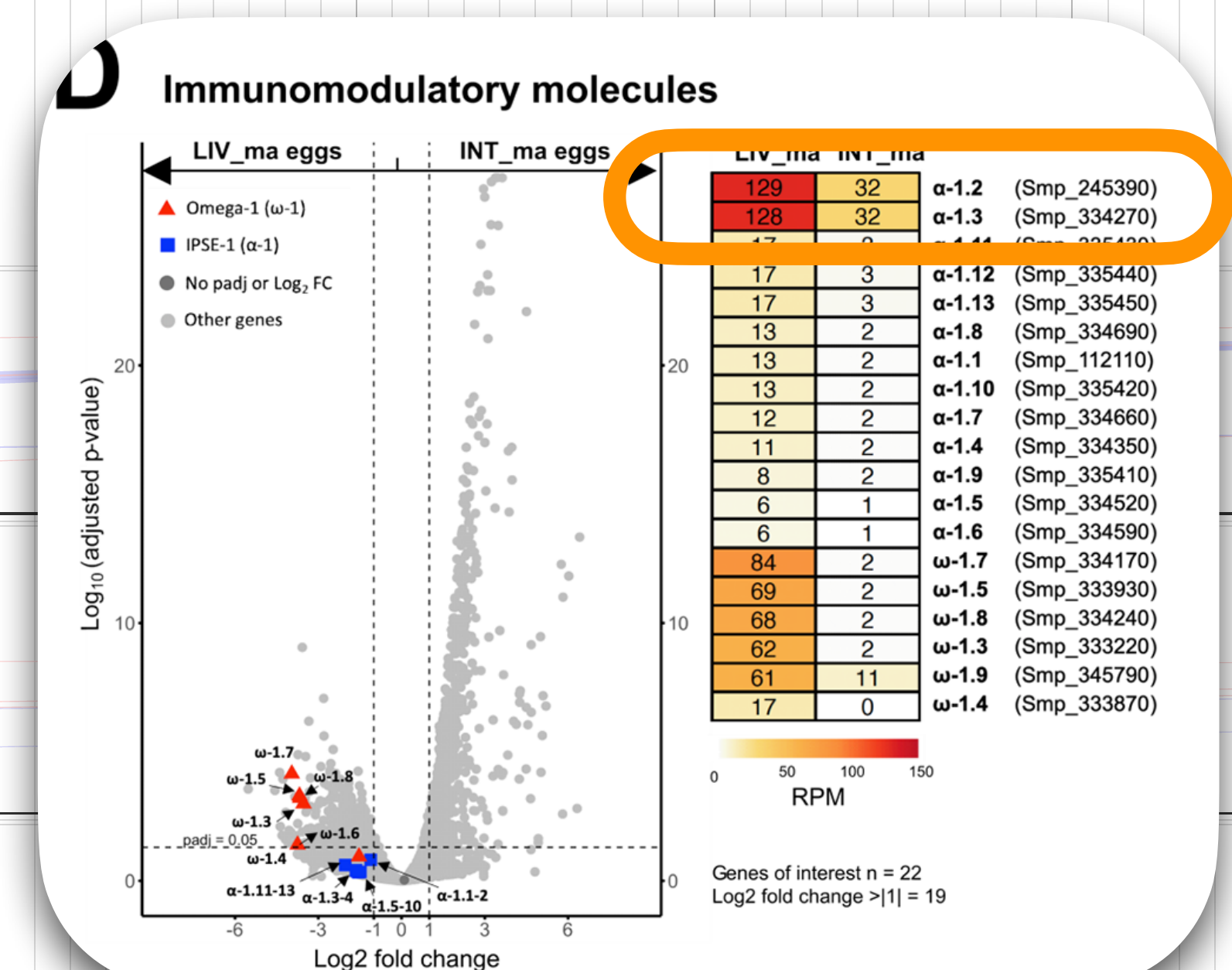
mRNA from
Gene X



18 reads per million



Expression value measured in reads per million (RPM)



How are reads converted to expression values?

How are reads converted to expression values?

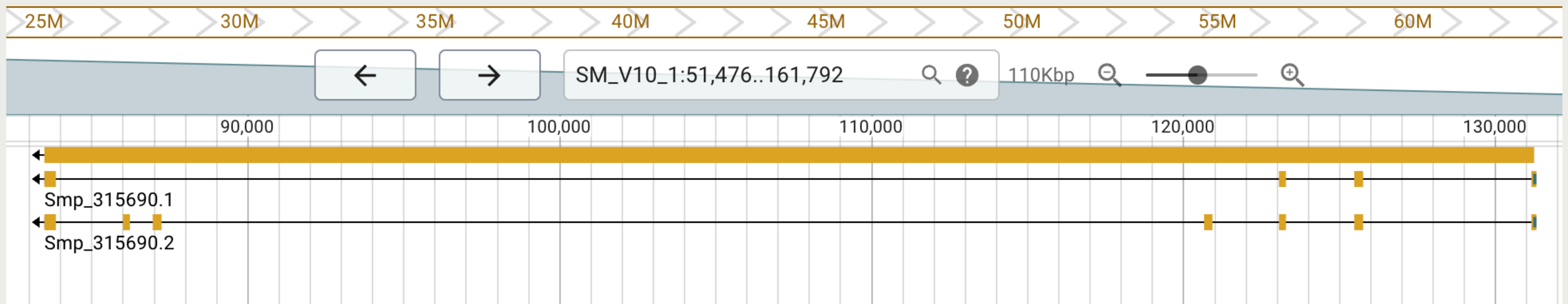
Many different situations must be considered

- What if a read only overlaps 1 nt of the gene?
- What if a read overlaps two genes?
- What if a gene has a lot of reads at one location and no reads at a different location?
- What if reads overlap a gene but the originated from the opposite strand?
- What if a read overlaps mostly intron?
- How do you link reads to different isoforms of the same gene?
- How do you normalize for genes that have variable lengths?

The featureCounts approach

Features and meta-features

- Consider SM_V10_1:80,002..135,160
 - Smp_315690 (gene)
 - Smp_315690.1 and Smp_315690.2 (transcripts)
 - Differ by the 3 exons (found in .2 but not .1)

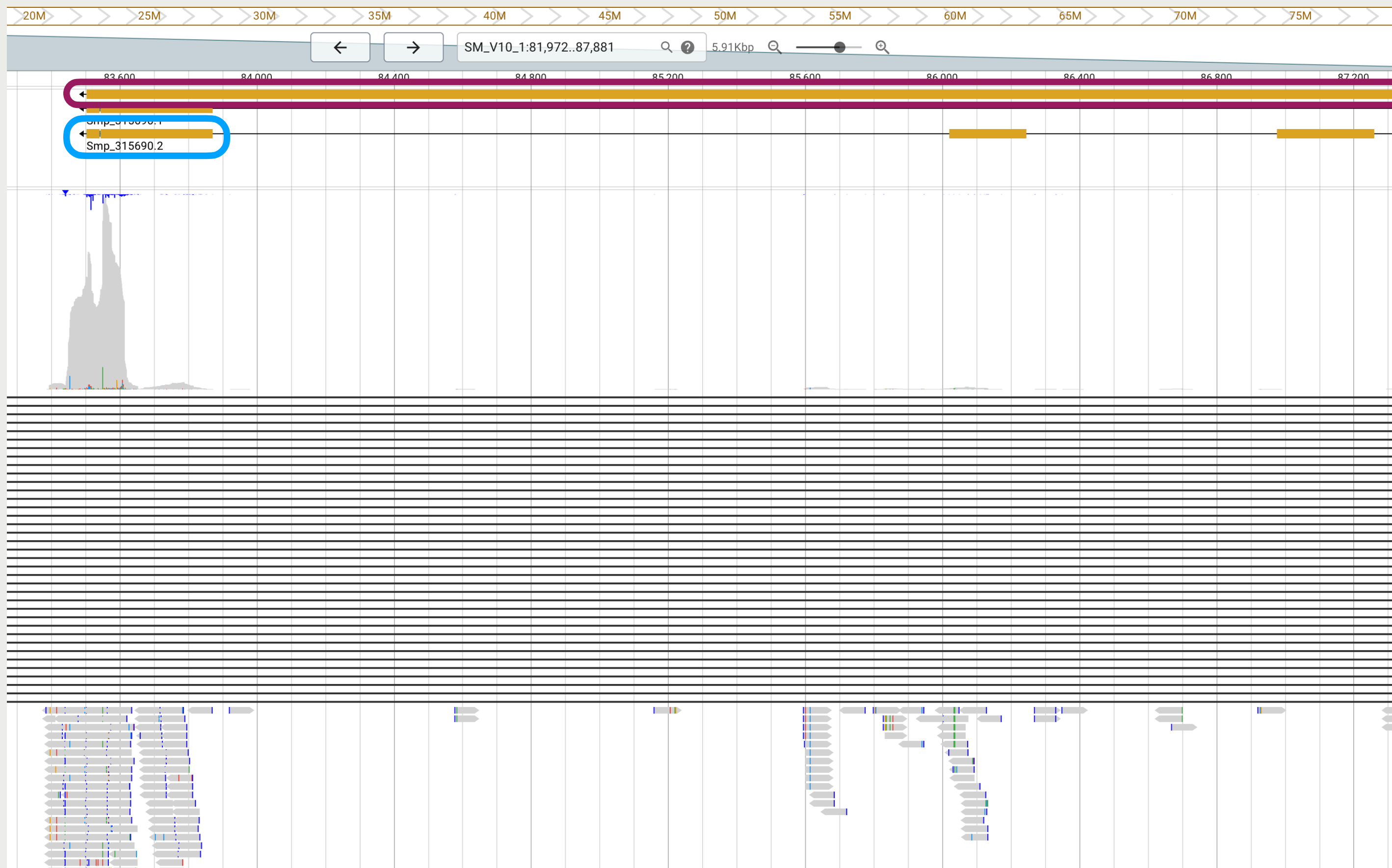


[illegible]

The featureCounts approach

Features and meta-features

- Most reads align to the 3' end:



Feature = exon (by default)

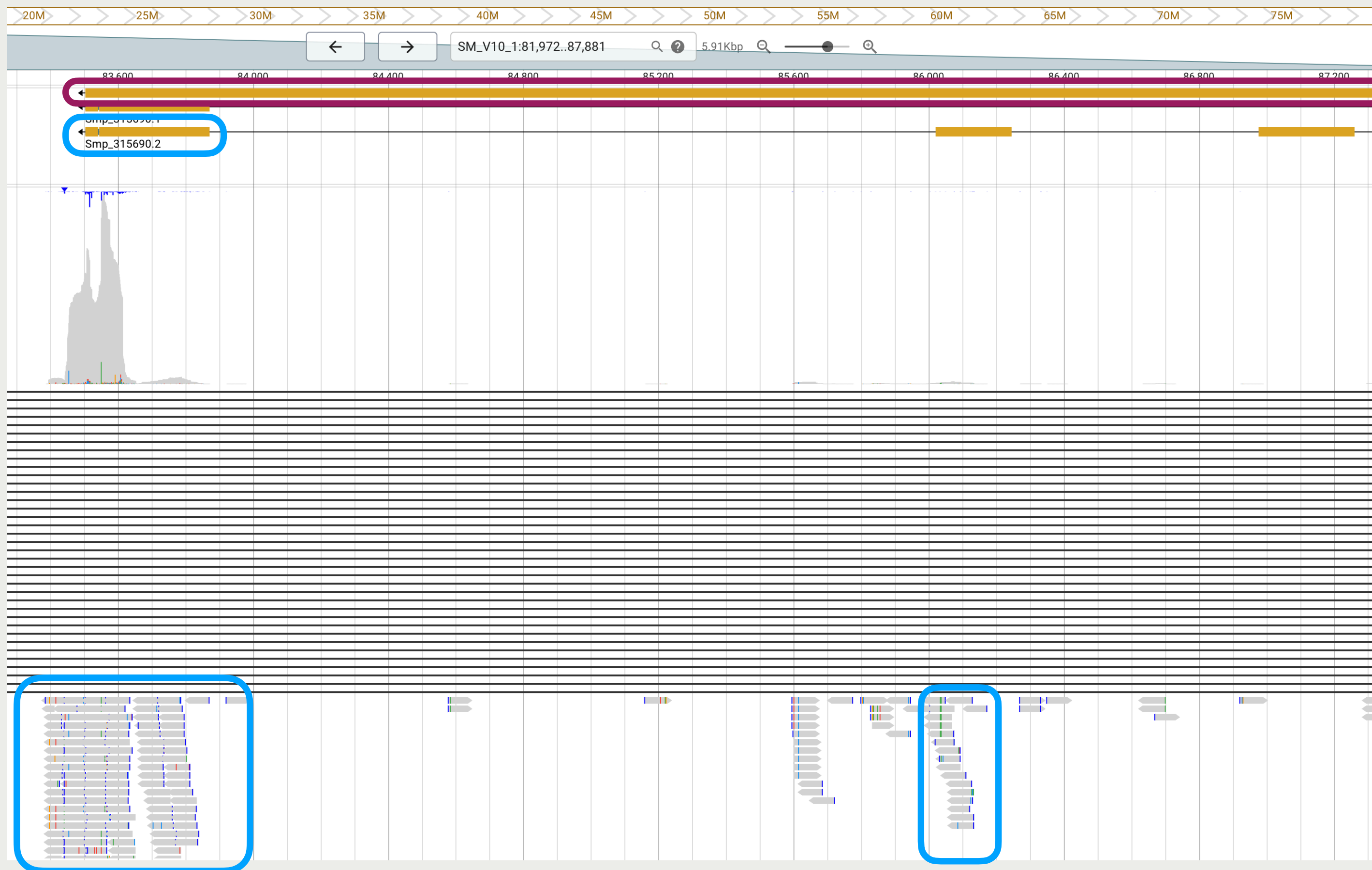
Meta-feature = gene_id (by default)

- featureCounts first counts “hits” to features (i.e., exons)
- “Hit” is any read that overlaps by 1 bp (by default)
- It then counts “hits” to meta-features (i.e., genes)
 - Tallies all hits to an exon that is apart of that gene

The featureCounts approach

Features and meta-features

- Most reads align to the 3' end:



Feature = exon (by default)

Meta-feature = gene_id (by default)

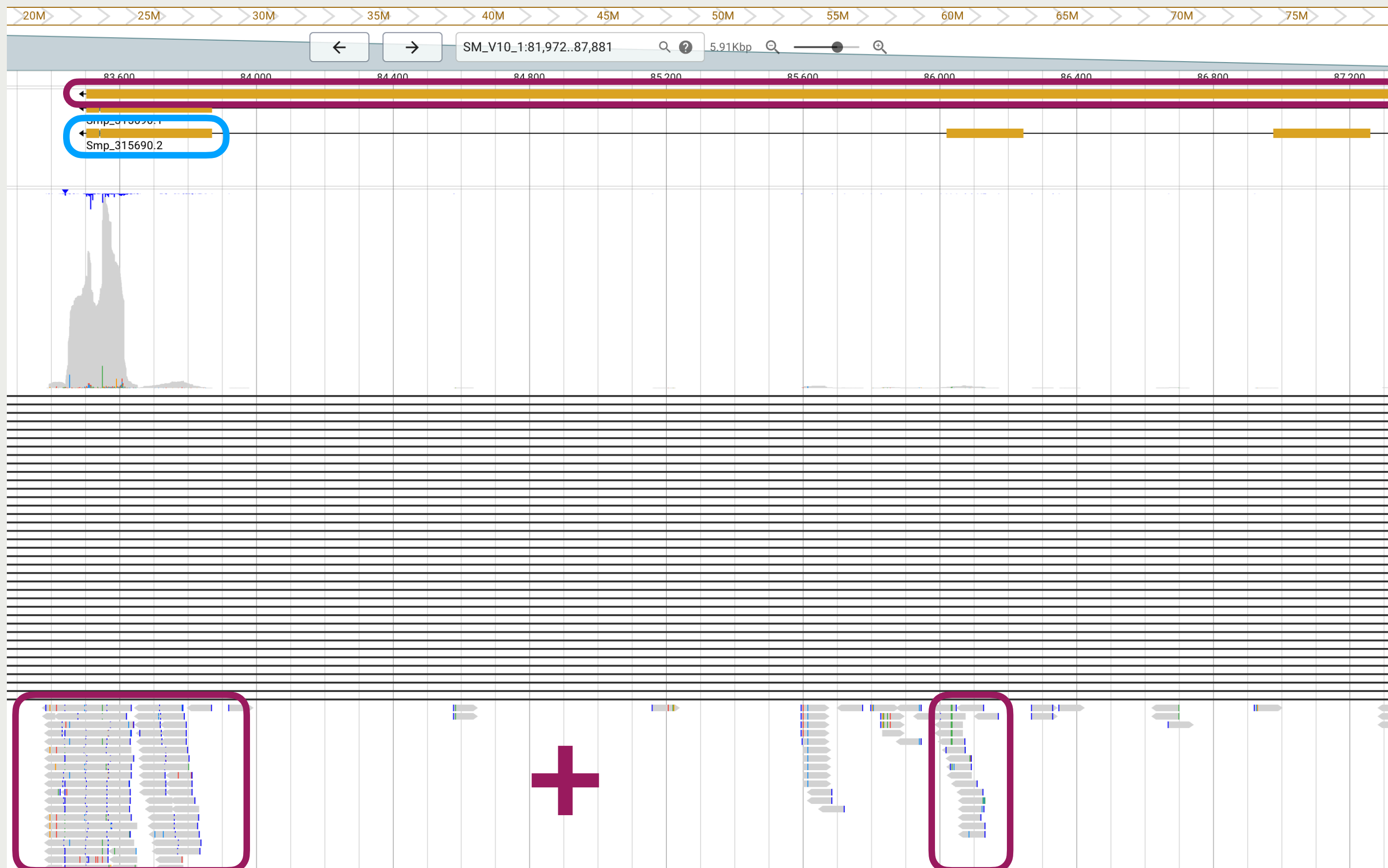
- featureCounts first counts "hits" to features (i.e., exons)
- "Hit" is any read that overlaps by 1 bp (by default)
- It then counts "hits" to meta-features (i.e., genes)
 - Tallies all hits to an exon that is apart of that gene

Hits to the feature

The featureCounts approach

Features and meta-features

- Most reads align to the 3' end:



Feature = exon (by default)

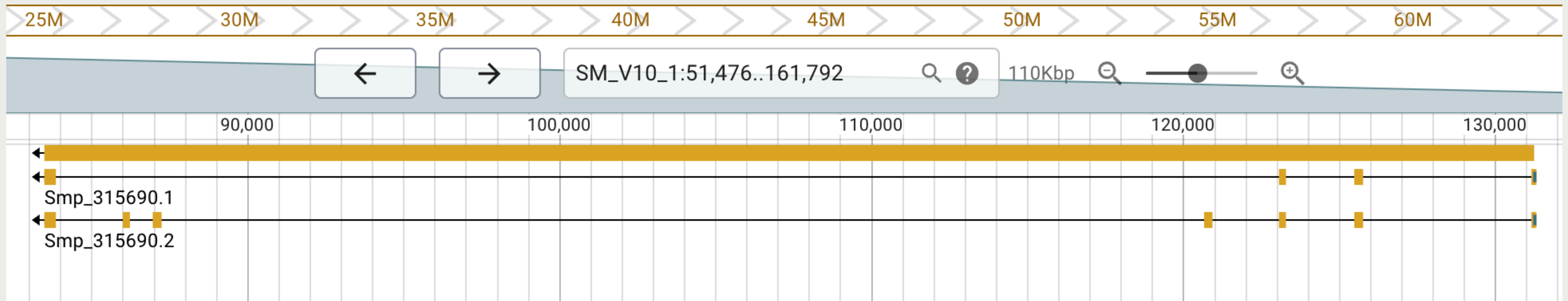
Meta-feature = gene_id (by default)

- featureCounts first counts "hits" to features (i.e., exons)
- "Hit" is any read that overlaps by 1 bp (by default)
- It then counts "hits" to meta-features (i.e., genes)
 - Tallies all hits to an exon that is apart of that gene

The featureCounts approach

Features and meta-features

- featureCounts has excellent performance (written in C)
- Works well when transcript-level differential expression is not of interest
 - What about when you want to differentiate isoforms (different transcripts from the same gene) like Smp_315690.1 and Smp_315690.2?
- Only use hits that map unambiguously...



The featureCounts approach

Features and meta-features

- Output: integer counts at meta-feature level
- Matrix of integer values - the value in the i -th row and the j -th column represents the reads assigned to gene i in sample j
- Cannot do differential expression with only these counts
 - Unnormalized
 - Longer genes will necessarily have more reads
 - Counts will vary by library size (i.e., number of reads that aligned)
 - Differential expression tools will do the normalization for us