# BIOL 343
# Applied Bioinformatics

## 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**

**Sample 2**

*mRNA from Gene X*

*mRNA from Gene X*
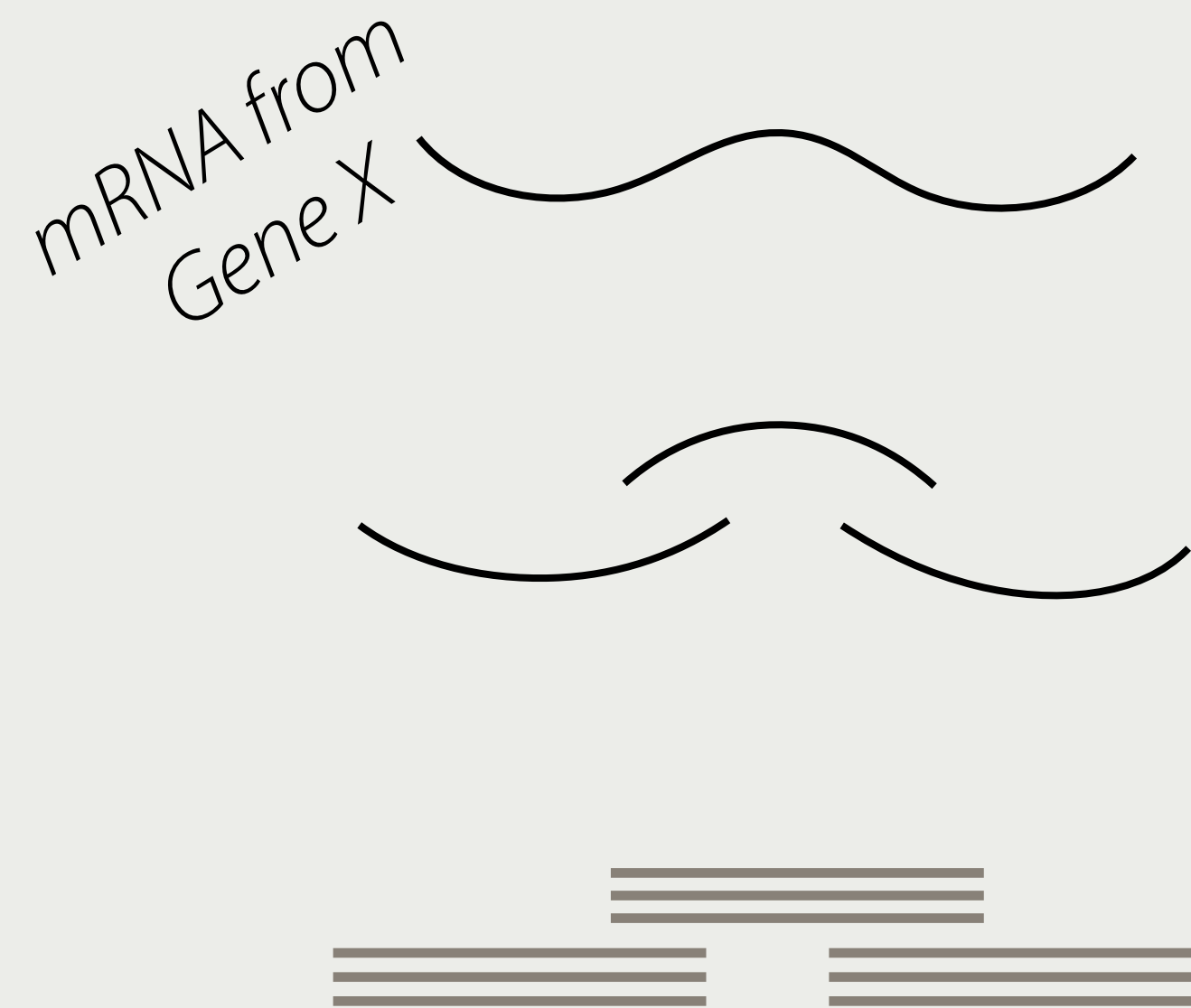
RNA

Library

Reads

Expression values

Differentially expressed gene?

**9 reads per million**

**18 reads per million**
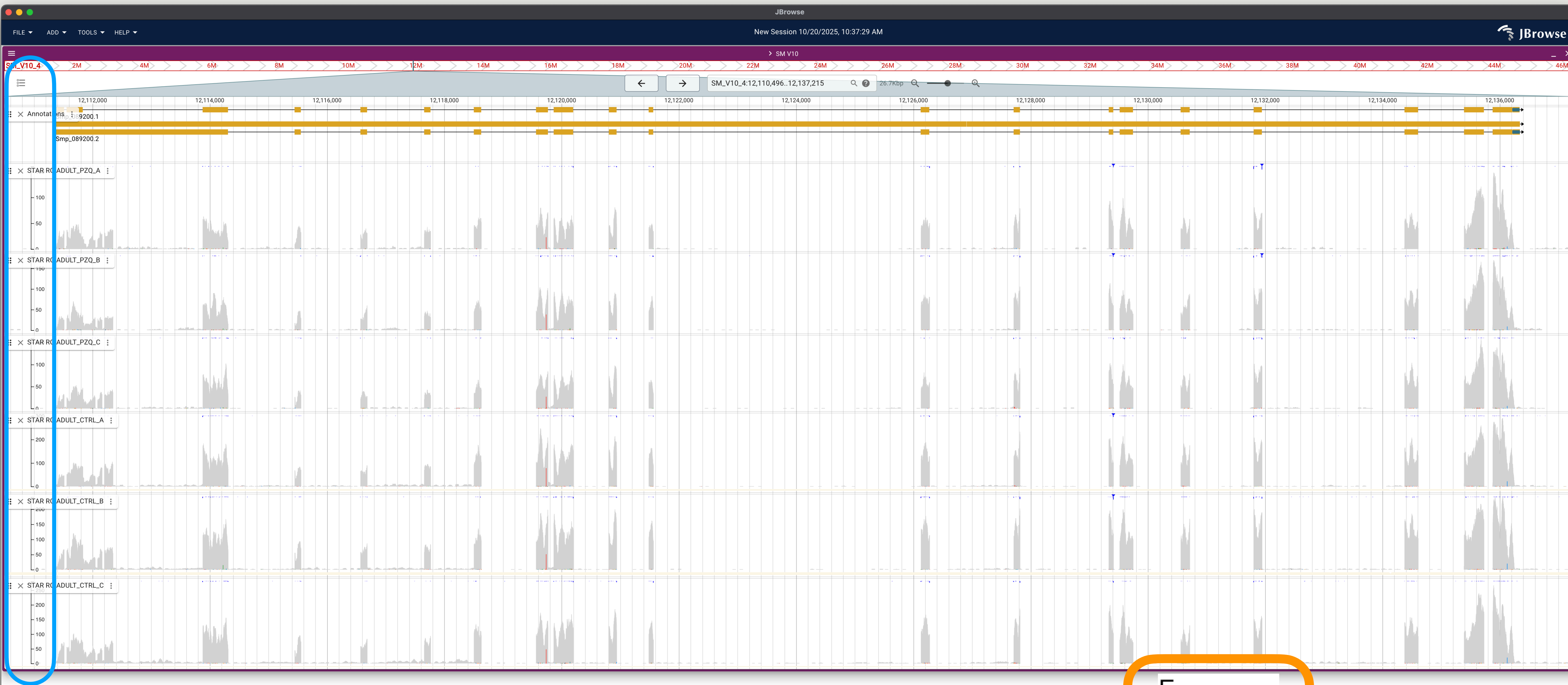
**Higher expression in CTRL than PZQ…do we see that?**

E

TPM

100
75
50
25
0

Smp_089200
(ABCB1-3)
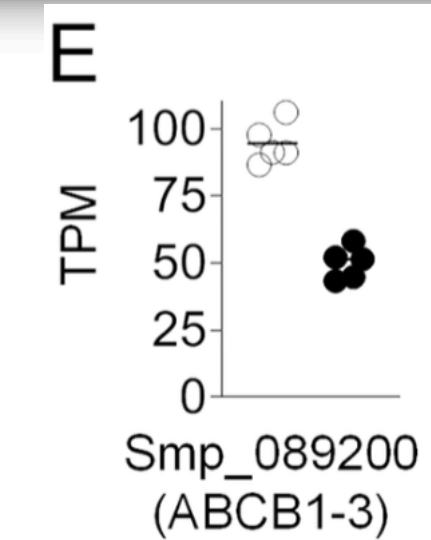
# How are reads converted to expression values?
## Many different situations must be considered

- 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 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)

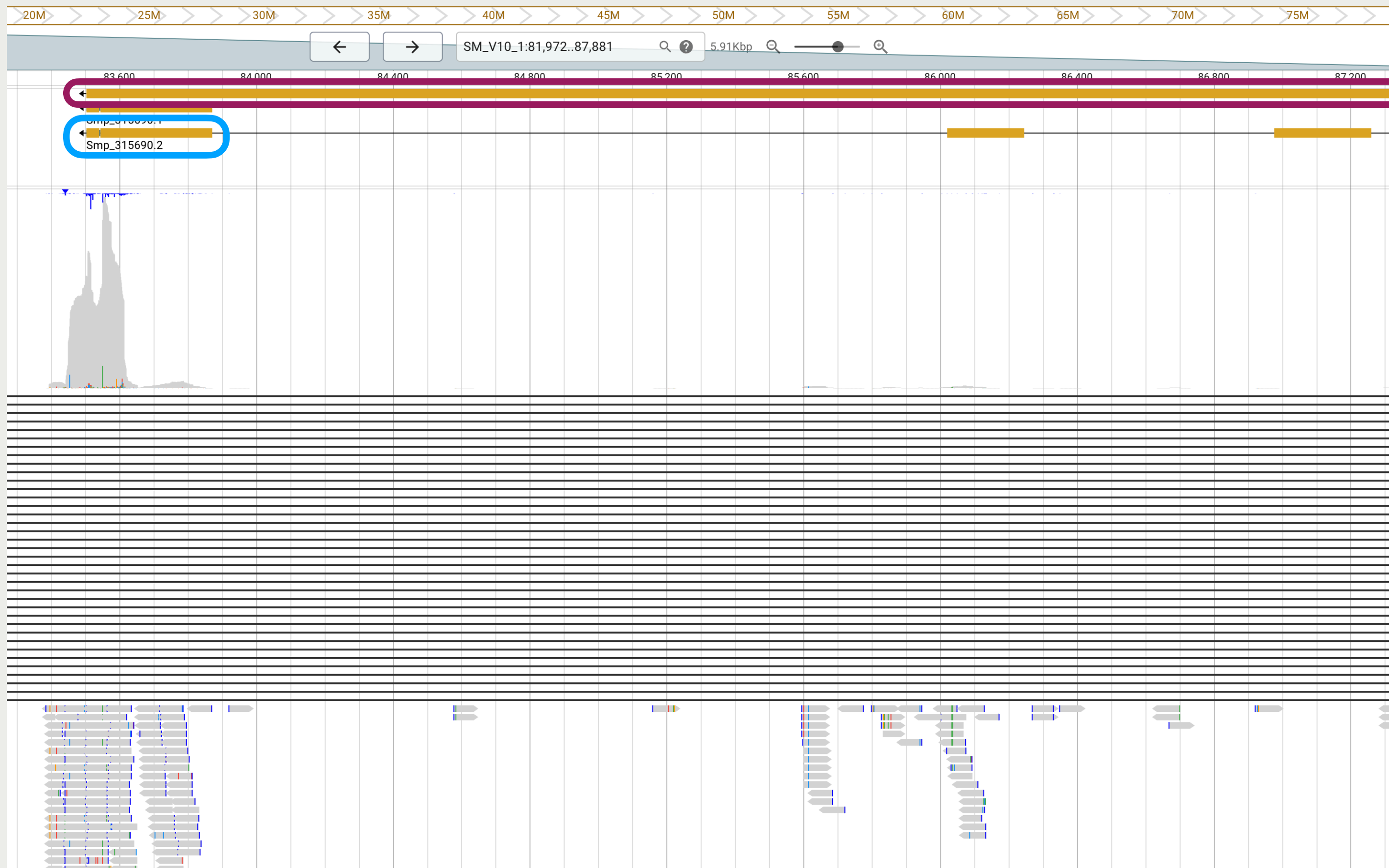| SM_V1O_1 | WormBase | gene | 83503 | 131262 | . | - | . | gene_id "Smp_315690"; gene_version "1"; gene_source "WormBase"; gene_biotype "protein_coding"; |
|---|---|---|---|---|---|---|---|---|
| SM_V1O_1 | WormBase | transcript | 83503 | 131262 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; gene_source "WormBase"; gene_biotype "protein_coding"; transcript_source "Wor |
| SM_V1O_1 | WormBase | exon | 131175 | 131262 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; exon_number "1"; gene_source "WormBase"; gene_biotype "protein_coding"; trans |
| SM_V1O_1 | WormBase | CDS | 131175 | 131232 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; exon_number "1"; gene_source "WormBase"; gene_biotype "protein_coding"; trans |
| SM_V1O_1 | WormBase | start_codon | 131230 | 131232 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; exon_number "1"; gene_source "WormBase"; gene_biotype "protein_coding"; trans |
| SM_V1O_1 | WormBase | exon | 125495 | 125778 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; exon_number "2"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | CDS | 125495 | 125778 | . | - | 2 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; exon_number "2"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | exon | 123081 | 123305 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; exon_number "3"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | CDS | 123081 | 123305 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; exon_number "3"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | exon | 83503 | 83870 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; exon_number "4"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | CDS | 83544 | 83870 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; exon_number "4"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | stop_codon | 83541 | 83543 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; exon_number "4"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | five_prime_utr | 131233 | 131262 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; gene_source "WormBase"; gene_biotype "protein_coding"; transcript_source "Wo |
| SM_V1O_1 | WormBase | three_prime_utr | 83503 | 83540 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.1"; gene_source "WormBase"; gene_biotype "protein_coding"; transcript_source "Wo |
| SM_V1O_1 | WormBase | transcript | 83503 | 131262 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; gene_source "WormBase"; gene_biotype "protein_coding"; transcript_source "Wo |
| SM_V1O_1 | WormBase | exon | 131175 | 131262 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "1"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | CDS | 131175 | 131232 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "1"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | start_codon | 131230 | 131232 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "1"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | exon | 125495 | 125778 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "2"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | CDS | 125495 | 125778 | . | - | 2 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "2"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | exon | 123081 | 123305 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "3"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | CDS | 123081 | 123305 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "3"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | exon | 120682 | 120949 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "4"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | CDS | 120682 | 120949 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "4"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | exon | 86976 | 87259 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "5"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | CDS | 86976 | 87259 | . | - | 2 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "5"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | exon | 86020 | 86244 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "6"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | CDS | 86020 | 86244 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "6"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | exon | 83503 | 83870 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "7"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | CDS | 83544 | 83870 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "7"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | stop_codon | 83541 | 83543 | . | - | 0 | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; exon_number "7"; gene_source "WormBase"; gene_biotype "protein_coding"; tran |
| SM_V1O_1 | WormBase | five_prime_utr | 131233 | 131262 | . | - | . | gene_id "Smp_315690"; gene_version "1"; transcript_id "Smp_315690.2"; gene_source "WormBase"; gene_biotype "protein_coding"; transcript_source "We |

# The featureCounts approach
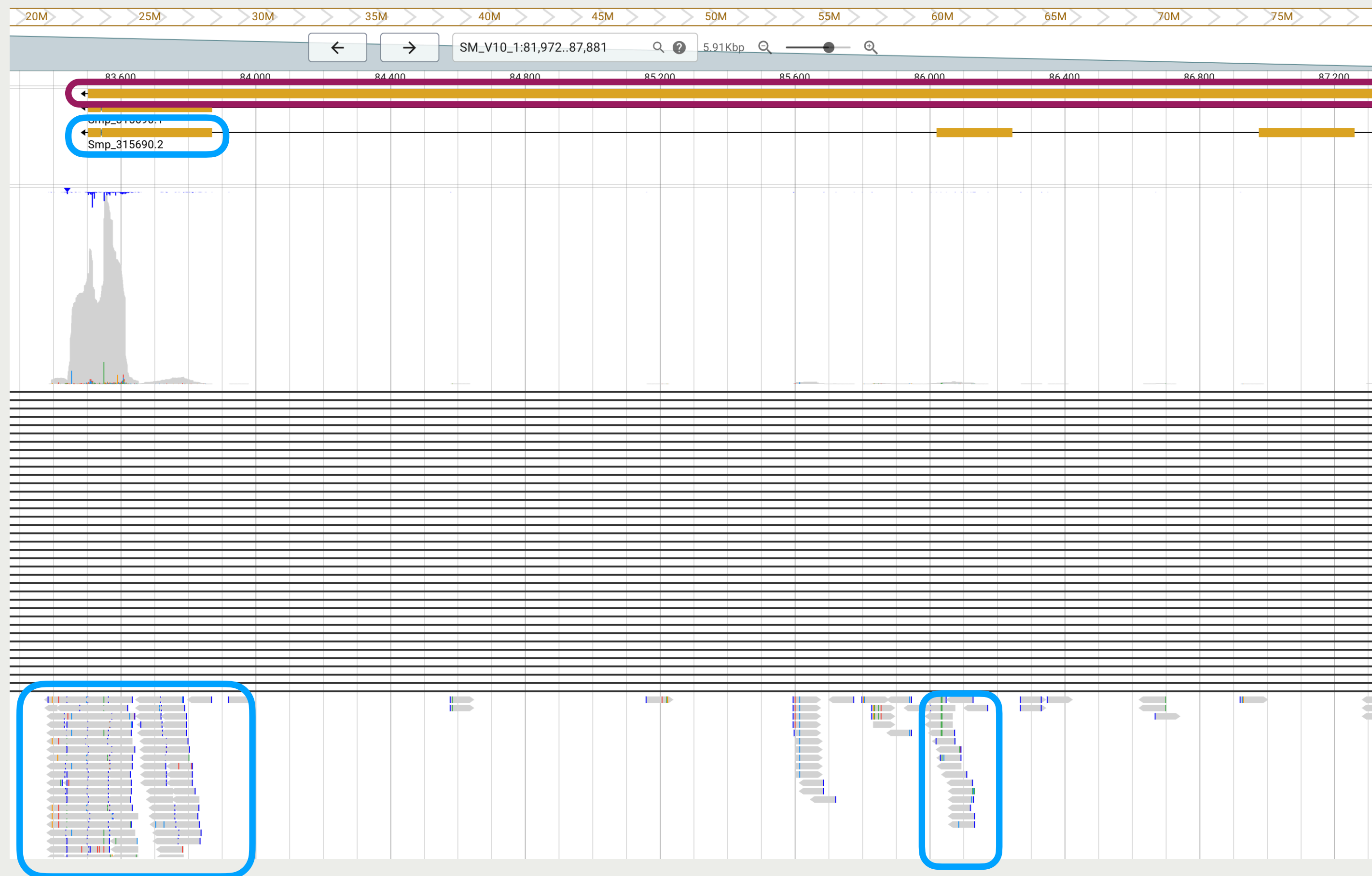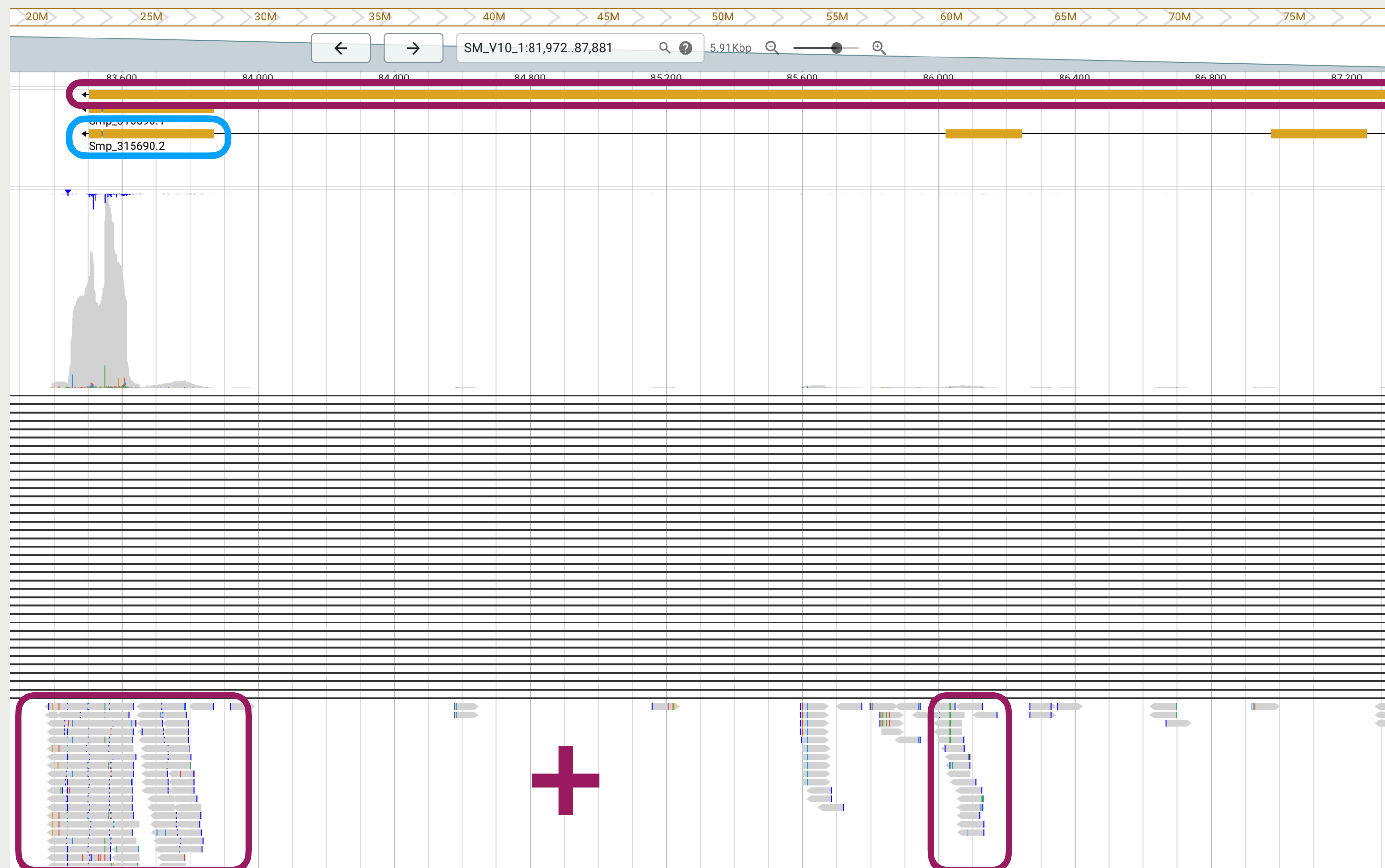## Features and meta-features



**Feature = exon (by default)**

**Meta-feature = gene_id (by default)**

- featureCounts first counts reads aligning to features (i.e., exons)
- Read is counted if it overlaps by 1 bp (by default)
- Then counts reads aligning to meta-features (i.e., genes)
  - Tallies all reads to an exon that is apart of that gene

# The featureCounts approach
## Features and meta-features



**Hits to the feature**

**Feature = exon (by default)**

**Meta-feature = gene_id (by default)**

- featureCounts first counts reads aligning to features (i.e., exons)
- Read is counted if it overlaps by 1 bp (by default)
- Then counts reads aligning to meta-features (i.e., genes)
  - Tallies all reads to an exon that is apart of that gene

# The featureCounts approach
## Features and meta-features



**Hits to the meta-feature**

**Feature = exon (by default)**

**Meta-feature = gene_id (by default)**

- featureCounts first counts reads aligning to features (i.e., exons)
- Read is counted if it overlaps by 1 bp (by default)
- Then counts reads aligning to meta-features (i.e., genes)
  - Tallies all reads 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…
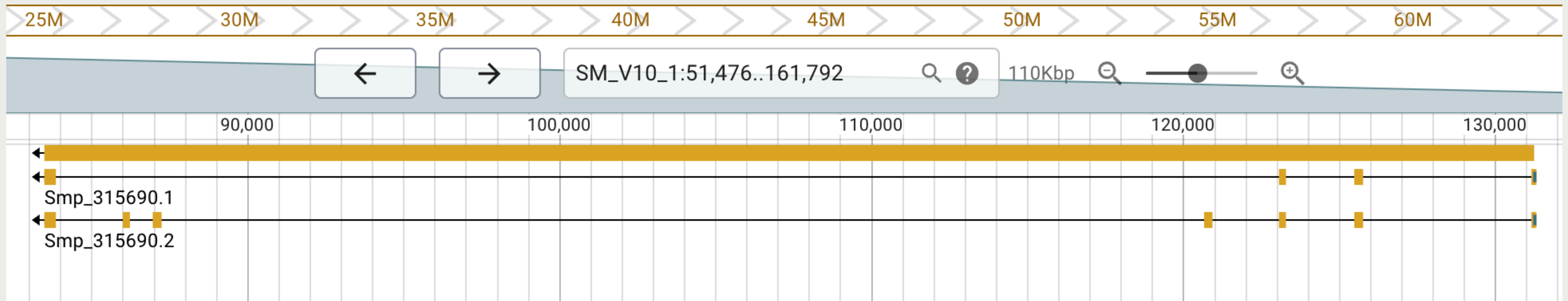
# The featureCounts approach
## Features and meta-features

- 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

# The featureCounts approach
## Features and meta-features

| Geneid | Chr | Start | End | Strand | Length | ADULT_PZQ_E | ADULT_PZQ_D | ADULT_CTRL_E | ADULT_CTRL_D |
|---|---|---|---|---|---|---|---|---|---|
| **Smp_329140** | SM_V10_1 | 68427 | 68783 | - | 357 | 2 | 1 | 0 | 0 |
| **Smp_318880** | SM_V10_1 | 261046 | 277059 | + | 16014 | 19 | 15 | 5 | 3 |