VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

Exploring Gene Models in JBrowse

Learning objectives

- Examine gene models in JBrowse
- Assess gene models based on evidence from various datatypes, including
    - RNAseq data
    - ChIP-Chip and ChIP-seq data
    - Transcription start site data
- Determine if a gene model is accurate or if alternate models are possible

Introduction

Why do we need to examine gene models critically? In previous exercises, you spent some time learning about gene pages and examining genes in the context of the JBrowse genome browser. It is important to recognize that **gene models (structural annotation) are often open to interpretation**, especially with respect to

- transcript initiation and termination sites (5' and 3' untranslated regions, or UTRs).
- alternative processing events … if you sequence deep enough, virtually *all* genes (in organisms that process transcripts) display alternative splicing, even for single exon genes.
- the potential significance of non-coding RNAs.

Even actively curated genomes (*Plasmodium falciparum, Trypanosoma brucei, Saccharomyces cerevisiae*) do not fully reflect all available knowledge about stage-specific splicing, as new information is emerging all the time! In addition, many gene models were computationally derived using methods that may have not relied on experimental evidence supporting intron/exon boundaries (e.g. RNAseq data).

This exercise has two parts

- **Part A**: explore several lines of evidence to examine the gene model for a particular gene
- **Part B**: use your newfound knowledge to examine other genes in *T. gondii*

Part A: exploring lines of evidence for gene models

In this exercise, we will explore several lines of evidence (data types) to interpret gene models and assess their accuracy and completeness. We will use the genome browser track configuration options in greater detail, focusing on the interpretation of RNA-seq datasets, and using this information to examine the differentially spliced HXGPRT gene of *T. gondii*.

The screenshot (Figure) below shows a sample of data tracks that can be turned on and configured in JBrowse. There are a few tracks that are worth examining which help in determining the accuracy of annotated gene models and that help in defining possible alternative splice variants of a gene. The link below will display the JBrowse view from the screenshot, except for any special configurations which are not stored in the URL.

Step 1: Open this link: https://shorturl.at/l1H9v and observe the gene model and data tracks for the gene **TGME49_200320** (hypoxanthine-xanthine-guanine phosphoribosyl transferase HXGPRT)

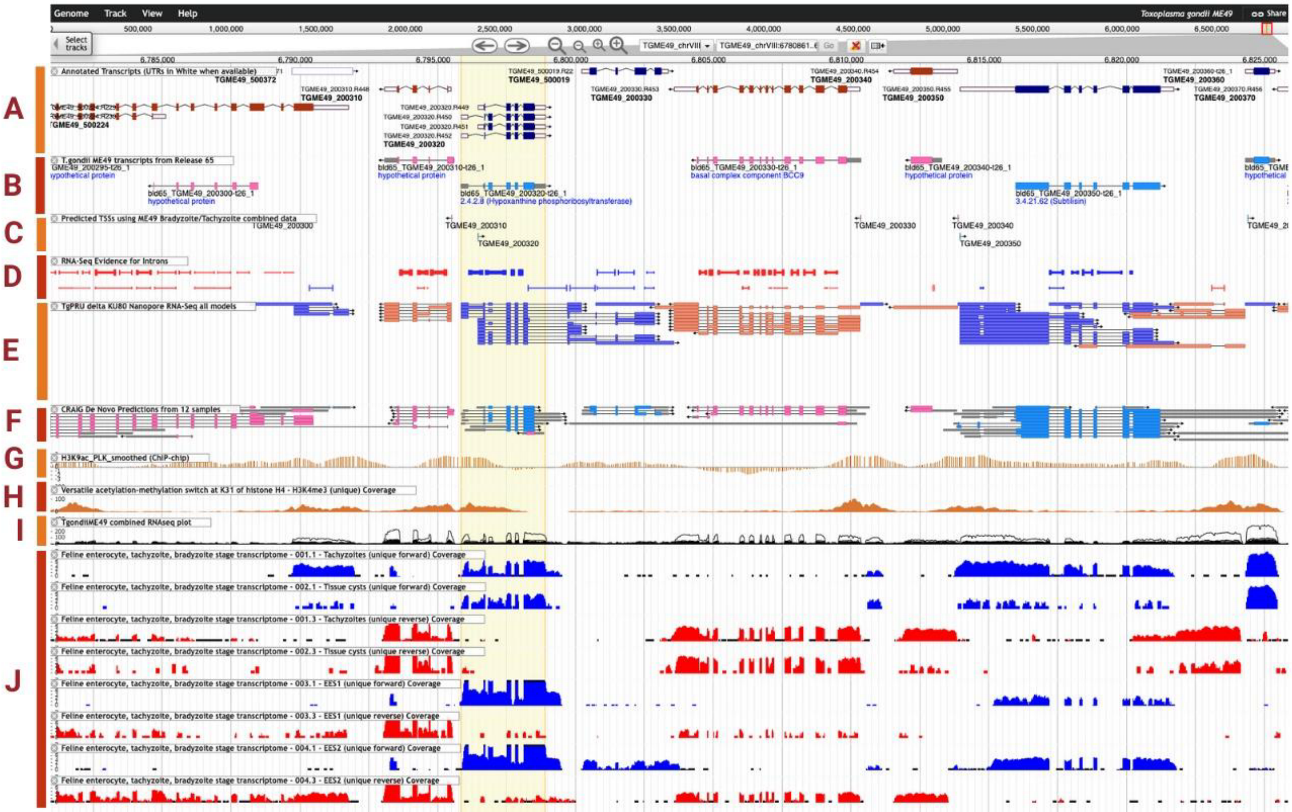

Genome
Track
View
Help
Toxoplasma gondii ME49
Share
Select tracks
TGME49_chrVIII
A
Annotated Transcripts (UTRs in White when available)
TGME49_200310
TGME49_500019
TGME49_200330
TGME49_200340
TGME49_200350
TGME49_200360
TGME49_200370
TGME49_500224
TGME49_200320
B
T.gondii ME49 transcripts from Release 65
bld65_TGME49_200320-t26_1
2.4.2.8 (Hypoxanthine phosphoribosyltransferase)
basal complex component BCC9
hypothetical protein
3.4.21.62 (Subtilisin)
C
Predicted TSSs using ME49 Bradyzoite/Tachyzoite combined data
D
RNA-Seq Evidence for Introns
E
TgPRU delta KU80 Nanopore RNA-Seq all models
F
CRAIG De Novo Predictions from 12 samples
G
H3K9ac_PLK_smoothed (ChIP-chip)
H
Versatile acetylation-methylation switch at K31 of histone H4 - H3K4me3 (unique) Coverage
I
TgondiiME49 combined RNAseq plot
J
Feline enterocyte, tachyzoite, bradyzoite stage transcriptome - 001.1 - Tachyzoites (unique forward) Coverage
Feline enterocyte, tachyzoite, bradyzoite stage transcriptome - 002.1 - Tissue cysts (unique forward) Coverage
Feline enterocyte, tachyzoite, bradyzoite stage transcriptome - 001.3 - Tachyzoites (unique reverse) Coverage
Feline enterocyte, tachyzoite, bradyzoite stage transcriptome - 002.3 - Tissue cysts (unique reverse) Coverage
Feline enterocyte, tachyzoite, bradyzoite stage transcriptome - 003.1 - EES1 (unique forward) Coverage
Feline enterocyte, tachyzoite, bradyzoite stage transcriptome - 003.3 - EES1 (unique reverse) Coverage
Feline enterocyte, tachyzoite, bradyzoite stage transcriptome - 004.1 - EES2 (unique forward) Coverage
Feline enterocyte, tachyzoite, bradyzoite stage transcriptome - 004.3 - EES2 (unique reverse) Coverage

Figure Screenshot from JBrowse showing different tracks:
(**A**) Annotated transcripts (official gene models),
(**B**) Release 65 transcripts (earlier gene models),
(**C**) Predicted TSSs (transcription start sites) using ME49 Bradyzoite/Tachyzoite combined data,
(**D**) RNA-Seq evidence for introns (splice junction evidence),
(**E**) Nanopore RNA-Seq (long-read transcriptomic data),
(**F**) CRAIG de novo predictions from 12 samples (alternative gene models using RNA-seq evidence from 12 experiments),
(**G**) ChIP-Chip H3K9ac,
(**H**) ChIP-Seq H3K4me3,
(**I**) Combined RNA-Seq,
(**J**) RNA-seq coverage plots in feline transcriptome (strand specific).

Step 2: Consider the following questions

- What evidence do each of the tracks provide?
    - A- Annotated transcripts (official gene models)
    - B- Release 65 transcripts (earlier gene models)
    - C- Predicted TSSs (transcription start sites) using ME49 Bradyzoite/Tachyzoite combined data
    - D- RNA-Seq evidence for introns (splice junction evidence)
    - E- Nanopore RNA-Seq (long-read transcriptomic data)[1]
    - F- CRAIG de novo predictions from 12 samples (alternative gene models using RNA-seq evidence from 12 experiments)
    - G- ChIP-Chip H3K9ac
    - H- ChIP-Seq H3K4me3
    - I- Combined RNA-Seq
    - J- RNA-seq coverage plots in feline transcriptome (strand specific)

- Are the ChIP-ChIP and Chip-seq tracks similar in what they show?

- Do you agree with the current annotated alternative splice forms of HXGPRT? Would you include any other splice forms?

- Are there other data tracks that might be useful to examine?

[1] You may notice that while the annotated transcripts (track A) correspond closely to the bulk RNAseq data (J), the long read RNAseq data (E) appears to contradict these data, suggesting that annotated genes TgME49_200320 & TgME49_500019 form a single transcription unit: a 1.4 kb unannotated intron disrupts the last exon of TgME49_200320, and the mRNA continues through TgME49_500019. This is because the TgPRU delta KU80 strain used for Nanopore sequencing was an HXGPRT knock-out mutant generated by deleting a 1.4 kb genomic restriction fragment; the putative "intron" at chrVIII:6798268-6799708 corresponds to this DNA deletion.

Part B: evaluating gene models

Working in groups, examine the *Toxoplasma gondii* genes in your assigned list (see table below). Using RNA-seq data and any other available evidence in JBrowse, evaluate the accuracy of each gene's official gene model.

- Determine which exons are supported by the data
- Whether the gene shows evidence of alternative splicing—either constitutive or stage-specific
- Whether the current gene model appears accurate or should be revised

We will reconvene at the end of the exercise to hear a brief report from each group.

Group 1:	**Group 4:**	**Group 7:**
TGME49_278510	TGME49_201270	TGME49_281440
TGME49_256650	TGME49_236630	TGME49_208718
TGME49_283540	TGME49_250115	TGME49_222930
Group 2:	**Group 5:**	**Group 8:**
TGME49_265390	TGME49_261720	TGME49_217490
TGME49_225730	TGME49_268610	TGME49_292150
TGME49_288000	TGME49_266310	TGME49_276170
Group 3:	**Group 6:**	**Group 9:**
TGME49_213660	TGME49_280380	TGME49_297850
TGME49_297160	TGME49_293720	TGME49_299010
TGME49_256025	TGME49_248445	TGME49_240470

Tips for evaluating gene models

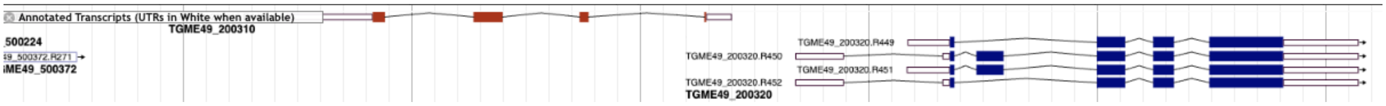
Take home message for the "gene model day"- Biology is complicated. We want to know what the gene looks like, but the gene model is often only an approximation. For many genes, there are alternate transcripts that start upstream and downstream; whether they should be annotated is a judgement call.

We use ToxoDB for this exercise because it has vast amounts of data in JBrowse. Note that this diversity of tracks is not available for all genomes.

Things to remember

- Check the genome at the top
- Red genes are read right to left while blue ones are read left to right, relative to orientation of chromosome
- Use the "about the track" feature to learn more
- Like with any data, information presented in many JBrowse tracks was generated in a lab- consider factors such as the strain, experimental conditions, developmental stage

1. Annotated transcripts (official gene models).



a. **What it is**: A representation of the gene's structure. These models often come from an automated pipeline that predicts the gene models and represent the current best guess for how the gene is organized and transcribed.

b. **How to read it**:

   i. Thick colored boxes representing **exons** (coding sequence). Blue boxes mean that the gene is on the forward/plus (+) strand, while red boxes represent genes on the reverse/minus (–) strand

   ii. Thin white boxes represent untranslated regions (**UTR**)

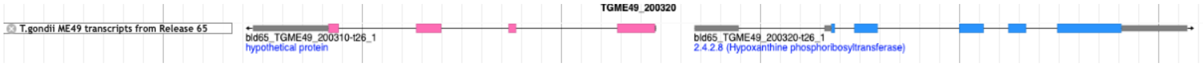
   iii. Connecting lines represent **introns**

   iv. Start/stop codons (where gene model begins and ends)

   v. Transcript isoforms (if any) shown on successive rows

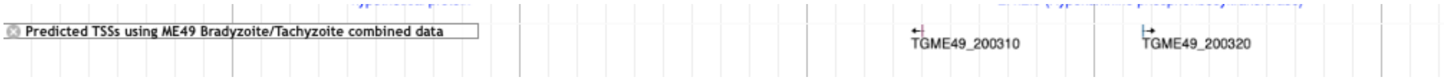
c. **How to interpret**: Annotated transcripts represent the **hypothesized structure**, but data (RNA-seq, ChIP-seq, TSS predictions, etc.- described below) shows whether that model is correct. *Note*: It is helpful to pin the "Annotated transcripts" track to the top of the page while evaluating gene models in the light of other evidence.

2. **Release 65 transcripts (earlier gene models)**


TGME49_200320
T.gondii ME49 transcripts from Release 65
bld65_TGME49_200310-t26_1
hypothetical protein
bld65_TGME49_200320-t26_1
2.4.2.8 (Hypoxanthine phosphoribosyltransferase)

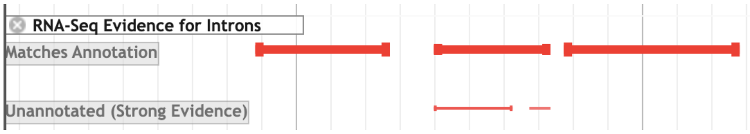
    a. **What it is**: Where available, these tracks are previous annotation releases, older predictions of exon–intron structure before the current official gene model was adopted.
    b. **How to read it**: Same structure as annotated transcripts (described above), but JBrowse uses color intensity to visually separate older models from the current ones. Lighter blue and lighter pink shades tell you that "**this is no longer the official annotation**."
    c. **How to interpret**:
        i. Look for differences between these and the current annotated transcript. Older models may differ in exon count, exon length, intron boundaries, UTR annotations, and isoform predictions. These differences help you understand what changed and why.
        ii. Compare older models to RNA-Seq data. You may be able to see why the gene model was corrected. For a summary of updates, see this spreadsheet.

3. **Predicted TSSs (transcription start sites)**


Predicted TSSs using ME49 Bradyzoite/Tachyzoite combined data
TGME49_200310
TGME49_200320

    a. **What it is**: Computationally predicted locations where transcription is likely to begin for a gene. These predictions usually come from algorithms that analyze features associated with promoters, such as promoter sequence motifs, CpG-rich regions, transcription factor binding patterns, chromatin features (if available), comparative genomics.
    b. **How to read it**: Track shows a single vertical bar for each predicted transcription start site. Each bar represents a single predicted TSS position (one base) with an assigned confidence or strength score.
    c. **How to interpret in relation to the gene model**: Look at "about the track" to learn more about what the score means. Then click on the marker to check the score. You can compare the bar's position to the first exon of the gene model to assess whether the annotated 5' end is correct, too long, or too short.

4. **RNA-Seq evidence for introns (splice junction evidence)**

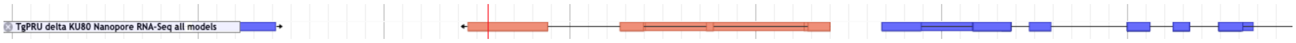


    a. **What it is**: Splice junction evidence comes from RNA-seq reads that span exon–exon boundaries. These "junction reads" do not align continuously; instead, they align in two pieces with a gap that corresponds to an intron. There are two sub-tracks for RNA-Seq evidence for introns. One is "**matches annotation**" (thousands of reads) and the other is "**unannotated" (strong evidence)**"- there are introns which have good evidence but they don't match the annotations.

b. **How to read the graph**: Each junction is visualized as a horizontal line = the intron, two small vertical bars = the *donor* (5') and *acceptor* (3') splice site, the thickness of the horizontal line = number of reads supporting that splice junction (appears upon hovering). A thick bar exactly matching an annotated intron confirms the current annotation. A thick bar connecting exons not joined in the current model could be evidence for alternative splicing or a missing intron. Clicking on the bar shows a statistics table, including the %MAI (percent of the Most Abundant Intron), which tells you how strong the support is relative to other introns in the gene. High %MAI suggests it is the dominant splice form.

c. **How to interpret in relation to the gene model**: It could be noise or it could be another exon which suggests an alternative gene model. By comparing these bars to the annotated gene model, you can tell whether introns are correct, missing, or alternatively used. Stage-specific or low-abundance junctions may reveal rare but important splice variants.

5. **Nanopore RNA-Seq (long-read transcriptomic data)**

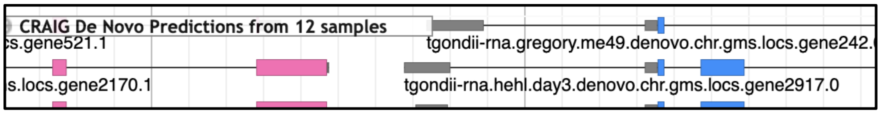


a. **What it is**: Nanopore RNA-Seq is a long-read sequencing technology that sequences full-length RNA molecules (or cDNA derived from them). Unlike short-read RNA-Seq (Illumina), Nanopore reads can span entire transcripts. This allows detection of alternative splicing (different isoforms of a gene), fusion transcripts, polyadenylation sites, novel transcripts. Output is aligned to a reference genome to show where transcripts map.

b. **How to read the graph**: It is represented as gene models displayed as boxes and lines. Boxes = exons, lines = introns. Arrows indicate transcription direction.

c. **How to interpret in relation to the gene model**: Compare to annotated gene model track, if exons and introns match, then transcript corresponds to known isoform. If there are extra or missing exons, there is potential novel alternative splicing. Novel intron junctions could suggest new splice sites. Strand misalignment may indicate antisense transcription or misannotation.

6. **CRAIG de novo predictions from 12 samples (alternative gene models using RNA-seq evidence from 12 experiments)**

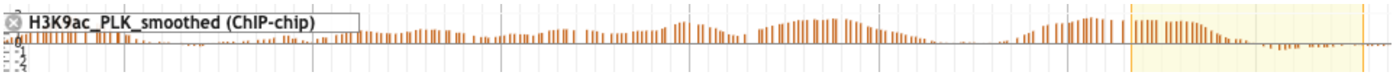

CRAIG De Novo Predictions from 12 samples
s.gene521.1
tgondii-rna.gregory.me49.denovo.chr.gms.locs.gene242.
s.locs.gene2170.1
tgondii-rna.hehl.day3.denovo.chr.gms.locs.gene2917.0

a. **What it is**: **CRAIG** (Constraint-based Reconstruction and Analysis of Initial Gene models) is a computational tool that predicts gene structures *directly from genomic sequence*, without relying on existing annotations.

b. **How to read the graph**: This track looks like the usual gene annotation tracks (linked exon blocks).

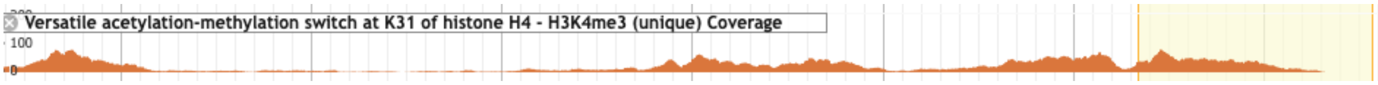
c. **How to interpret in relation to the gene model**: This track is useful for checking the accuracy of existing gene models. CRAIG can propose alternative exon boundaries. It can predict exons missing from the official annotation. CRAIG may find ORFs not present in the reference annotation. Gives an independent “second opinion” compared to RNA-seq. Provides a starting point for manual curation when experimental data are sparse. CRAIG predictions are especially helpful for organisms like *T. gondii* where some gene models remain incomplete. If CRAIG’s model matches RNA-seq evidence, it may reveal a correct or improved gene model. If CRAIG predicts exons not supported by RNA-seq, those predictions are likely false positives. CRAIG is most powerful when used together with evidence-based tracks.

7. **ChIP-Chip data**



   a. **What it is**: Chromatin Immunoprecipitation combined with microarray (“chip”) technology. You use an antibody to pull down DNA associated with a protein or histone modification (for instance H3K9ac: Histone H3 acetylated at lysine 9, usually associated with active promoters and enhancers, indicating transcriptionally active regions.) Then the bound DNA is hybridized to a microarray to detect enrichment across the genome. Output is signal intensity at array probes across the genome (continuous smoothed signal). Limited by microarray detection and limited by probe spacing on array.
   b. **How to read the graph**: Peaks = high enrichment (more H3K9ac). Higher peaks= stronger evidence of active chromatin.Troughs = low/no enrichment.
   c. **How to interpret in relation to the gene model**: H3K9ac is often enriched just upstream of the transcription start site (TSS). Peaks overlapping a promoter suggest active transcription initiation. Moderate enrichment across gene bodies can indicate actively transcribed genes. Lack of enrichment may indicate silent genes. Compare H3K9ac track with RNA-Seq (expression) tracks. Genes with strong H3K9ac at the promoter often have high transcription.

8. **ChIP-seq data**

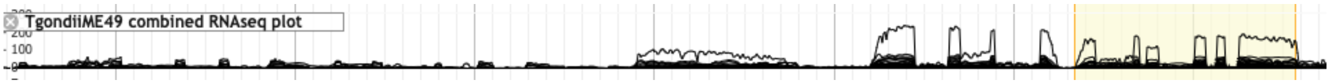

Versatile acetylation-methylation switch at K31 of histone H4 - H3K4me3 (unique) Coverage
100
0

   a. **What it is**: Chromatin Immunoprecipitation + Sequencing uses an antibody to pull down DNA associated with a specific histone modification or protein. The bound DNA is sequenced and aligned to the genome. For instance, H3K4me3, trimethylation of histone H3 at lysine 4 typically marks active promoters, often at transcription start sites. Only reads that map uniquely to one location in the genome are considered, reducing ambiguity. Coverage track shows the density of ChIP-Seq reads along the genome. High resolution, often single-base or nucleosome resolution (~10–50 bp). Sharp peaks can resolve individual binding sites. Can discover novel sites genome-wide, not limited by array design.
   b. **How to read the graph**: filled area where X axis is genomic coordinates and Y axis is number of ChIP-seq reads mapped to each region- enrichment. High peaks = strong enrichment of the histone modification. Broad peaks = extended regions of

modification (sometimes gene bodies or enhancers). Narrow peaks = usually promoters or transcription factor binding sites.

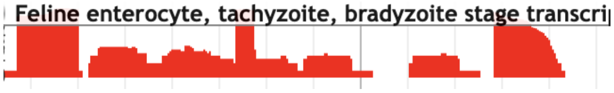
c. **How to interpret in relation to the gene model**: Promoters / TSS: H3K4me3 is concentrated at active promoters. Peaks overlapping the annotated TSS suggest transcriptionally active genes. Gene body: Sometimes H3K4me3 spreads slightly into the gene body of highly expressed genes. Lack of peaks suggests low or inactive transcription.

9. **Combined RNA-seq plot**



   a. **What it is**: This track usually represents gene expression from multiple RNA-Seq experiments combined into one view. Overlay of multiple conditions. Purpose: to visualize overall transcript expression patterns across the genome.
   b. **How to read the graph**: Y axis shows expression levels. Peaks show regions of high transcript abundance while troughs/zeros show low or no expression. Split peaks or multiple overlapping tracks suggest alternative isoforms or condition-specific gene expression.
   c. **How to interpret in relation to the gene model**: Exon correspondence: Coverage peaks should align with annotated exons. Gaps indicate introns, skipped exons, or low expression. Transcript isoforms: Long-read data or junction-spanning short reads can reveal alternative splicing. Compare aligned reads or transcript blocks to gene models to identify known vs novel isoforms.

10. **RNA-seq coverage tracks**



   a. **What it is**: A coverage plot shows the number of RNA-seq reads that align to each position along the genome. It provides a quantitative view of transcription across a gene or region.
   b. **How to read the graph**: At each genomic position, the height of the coverage signal represents: “How many sequencing reads overlap this base (or small window)?” Peaks appear where expression is high. Valleys or flat areas appear where expression is low or absent. This graph is often automatically normalized.
   c. **How to interpret in relation to the gene model**:
      i. Identifying exons: coverage is usually high over exons and low in introns
      ii. Evaluating alternative splicing: Missing coverage over an annotated exon may indicate that it is not used
      iii. Differences between stage-specific datasets highlight stage-specific isoforms
      iv. Checking gene boundaries: coverage may start upstream or extend downstream