Interpreting RNA-seq data (Browser Exercise II)

Learning objectives:

- Examine gene models in JBrowse
- Assess gene models based on RNAseq data
- Assess gene models based on ChIP-chip and ChIP-Seq data
- Determine if a gene model is accurate or if alternate models are possible
- Explore transcription start site data

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, however, 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 heavily 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 replied on experimental evidence supporting intron/exon boundaries (e.g. RNAseq data).

In this exercise, we will explore 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. You will then apply your newfound skills to examine other genes that may be alternatively spliced … and report your findings back to the group as a whole.

The screen shot below (Fig. 1) 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 figure 1, except for any special configurations with are not stored in the URL. For example, tracks 1c and 1d are collapsed in figure one but will appear expanded in the JBrowse view after clicking on the link:

https://tinyurl.com/379c5r8v

- What evidence do each of the tracks provide?
- Are the ChiP-ChIP and Chip-seq tracks similar in what they show?
- How many alternative splice variants of HXGPRT would you be willing to annotate based on the evidence?
- Are there other data tracks that might be useful to examine?

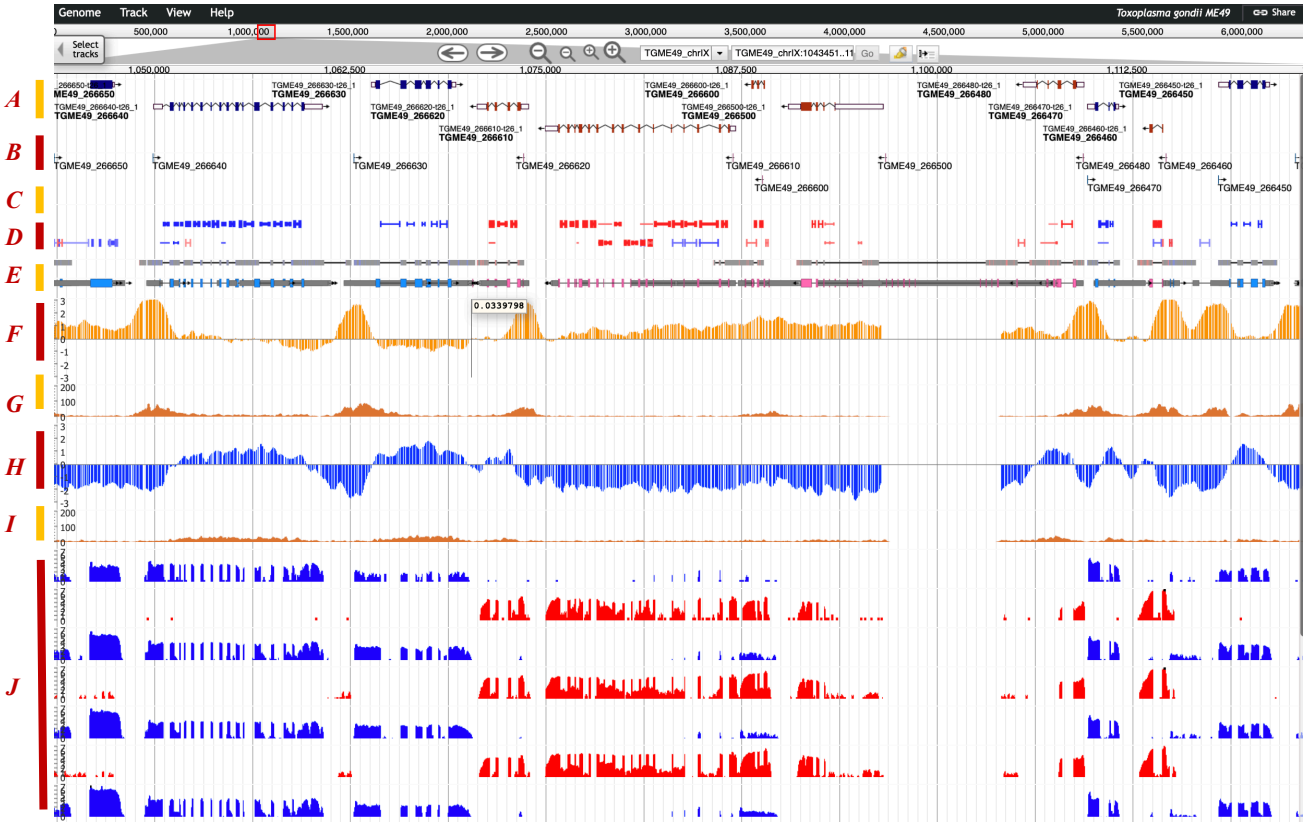

*Figure 1: Screen shot from ToxoDB JBrowse. **A**. Official gene models. **B**. Predicted transcription start sites. **C**. Splice junction evidence based on available RNAseq data. **D**. Nanopore long-read transcriptomic data (collapse view). **E**. Alternative gene models using RNAseq evidence from 12 experiments (collapsed view). **F**. Chip-ChIP H3K9ac. **G**. Chip-Seq H3K4me3. **H**. Chip-ChIP H3K4me1. I. Chip-Seq H3K4me1. **J**. RNAseq coverage from Toxoplasma gondii strain CZ clone H3 in feline enteroepithelial stage (strand specific).*

Working in groups, please examine the genes in your list, to evaluate their official gene models based on RNA-seq data and any other available evidence. See if you can discover which exon(s) were represented … and determine whether

these genes are actually alternatively spliced (constitutively or stage-specifically). We will then reconvene to hear a brief report from each group.

Group 1:	**Group 4:**	**Group 7:**
TGME49_246490	TGME49_211420	TGME49_281440
TGME49_256650	TGME49_214440	TGME49_279390
TGME49_283540	TGME49_250115	TGME49_202770
Group 2:	**Group 5:**	**Group 8:**
TGME49_226410	TGME49_261720	TGME49_217490
TGME49_225730	TGME49_268610	TGME49_292150
TGME49_213610	TGME49_270520	TGME49_276170
Group 3:	**Group 6:**	**Group 9:**
TGME49_213660	TGME49_280380	TGME49_266610
TGME49_297160	TGME49_293720	TGME49_299010
TGME49_211250	TGME49_248445	TGME49_230180