Interpreting RNA-seq data (Browser Exercise II)

Learning objectives:

- Examine gene models in JBrowse
- Assess gene models based on RNAseq and intron evidence data
- Assess gene models based on GC content
- Assess gene models based on protein mass spec data
- Explore ATAC-seq data
- Determine if a gene model is accurate or if alternate models are possible

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 (*Homo sapiens*, *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).

In this exercise, we will explore genome browser track configuration options in greater detail, focusing on the interpretation of RNA-seq and other datasets, and using this information 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/56y58vn5

- What evidence do each of the tracks provide?
- Is the GC content track useful in determining where coding sequence is compared to no-coding sequence?
- How many alternative splice variants does your gene have? Do you agree with them? Would you make additional gene model calls?
- Can the Mass spec peptide tracks help? Do they have limitations?
- Why is it useful to look at strand-specific RNA-Seq data?

*Figure 1: Screen shot from VectorBase JBrowse. **A**. Official gene models. **B**. Intron evidence (includes both those that match the official annotation and those that are un-annotated. **C**. GC content. **D**. ATAC-Seq **E**. Mass spectrometry peptides (collapsed view) **F**. Combined RNA-Seq data. **G**. Individual RNA-Seq coverage from strand-specific data.*

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 additional or alternative gene models can be made. We will then reconvene to hear a brief report from each group.

Group 1:	Group 6:
AGAP002559	AGAP012881
AGAP012081	AGAP009329
AGAP003374	AGAP004134
Group 2:	**Group 7:**
AGAP012396	AGAP013455
AGAP003471	AGAP003078
AGAP002315	AGAP002658
Group 3:	**Group 8:**
AGAP001238	AGAP004130
AGAP004237	AGAP012383
AGAP002327	AGAP001659
Group 4:	**Group 9:**
AGAP002272	AGAP028546
AGAP011054	AGAP005264
AGAP011980	AGAP012988
Group 5:	**Group 10:**
AGAP004321	AGAP012155
AGAP007086	AGAP007733
AGAP028034	AGAP007165