Interpreting RNA-seq data (Browser Exercise II)

In previous exercises, you spent some time learning about gene pages and examining genes in the context of the GBrowse 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

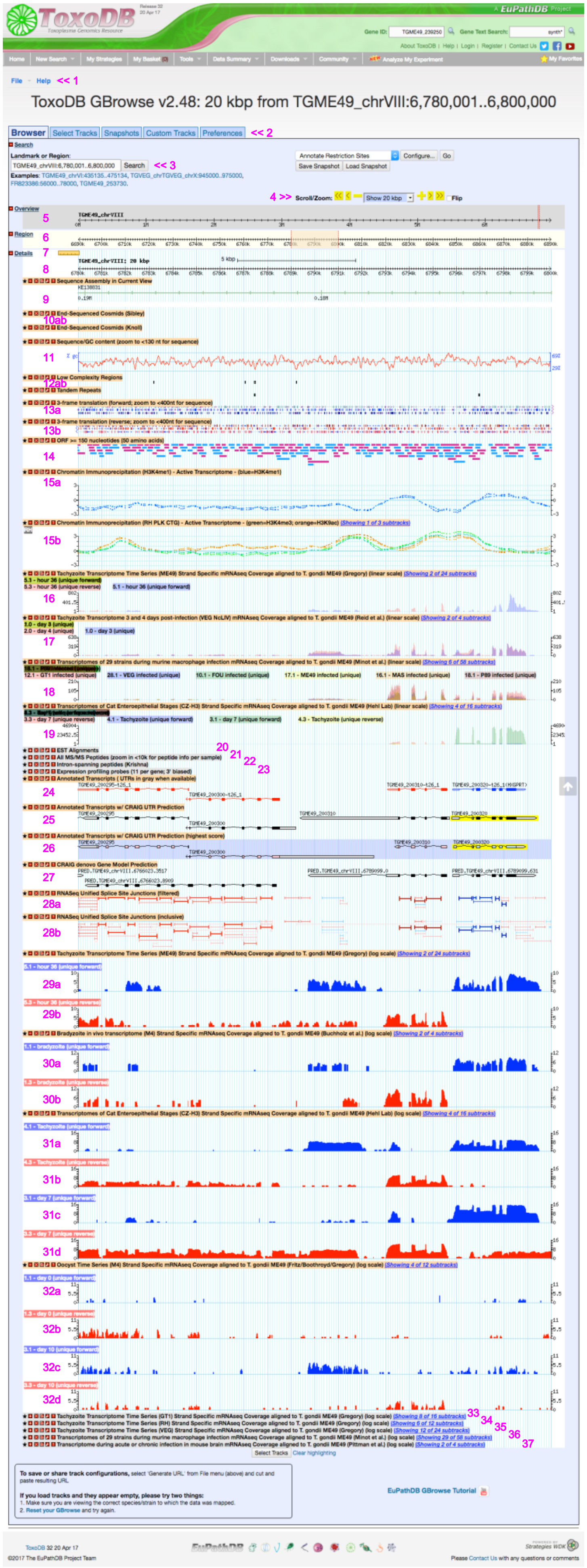
Even heavily curated genomes (*Plasmodium falciparum, Trypanosoma brucei, Saccharomyces cerevisiae*) may not fully reflect all available knowledge about stage-specific splicing, as new information is emerging all the time!

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 large figure shown on the following page presents one example of a GBrowse view that has been extensively reconfigured to explore alternative splicing in *Toxoplasma.* The resolution of the display page has been increased, many additional tracks have been turned on, and some of these have been overlaid, or reconfigured in other ways. Setting this up can take some time … but may be worth the effort, if you wish to use this particular configuration repeatedly. Much (but not all) track configuration data is stored in the URL, which can be generated by selecting 'Generate URL' from the File menu at the top of the page **(1)**. This particular URL is gigantic!

http://toxodb.org/cgi-bin/gbrowse/toxodb/?start=6780001;stop=6800000;ref=TGME49_chrVIII;width=1024;version=100;flip=0;grid=1;id=3dcd3db760a437dbc26f58aaaa59d3d0;l=Scaffolds%1ECosmidsSibley%1ECosmidsLorenzi%1EGC%20Content%1ELowComplexity%1ETandemRepeat%1ETranslationF%1ETranslationR%1EORF%1EtgonME49_chipChipExper_Einstein_ME1_RSRC_chipChipSmoothed%1EtgonME49_chipChipExper_Einstein_RSRC_chipChipSmoothed%1EtgonME49_Gregory_ME49_mRNA_rnaSeq_RSRCCoverageUnlogged%1EtgonME49_Reid_tachy_rnaSeq_RSRCCoverageUnlogged%1EtgonME49_Saeij_Jeroen_strains_rnaSeq_RSRCCoverageUnlogged%1EtgonME49_DBP_Hehl-Grigg_rnaSeq_RSRCCoverageUnlogged%1EEST%1EUnifiedMassSpecPeptides%1ERiteshPeptide%1EAffymetrixExpressionNuclearCoding%1EGene%1Eutr_only_union%1Eutr_only.scores%1Edenovo_union%1EGSNAPUnifiedIntronJunctionRefined%1EGSNAPUnifiedIntronJunctionInclusive%1EtgonME49_Gregory_ME49_mRNA_rnaSeq_RSRCCoverage%1EtgonME49_Buchholz_Boothroyd_M4_in_vivo_bradyzoite_rnaSeq_RSRCCoverage%1EtgonME49_DBP_Hehl-Grigg_rnaSeq_RSRCCoverage%1EtgonME49_Boothroyd_oocyst_rnaSeq_RSRCCoverage%1EtgonME49_Gregory_GT1_mRNA_rnaSeq_RSRCCoverage%1EtgonME49_Gregory_RH_mRNA_rnaSeq_RSRCCoverage%1EtgonME49_Gregory_VEG_mRNA_rnaSeq_RSRCCoverage%1EtgonME49_Saeij_Jeroen_strains_rnaSeq_RSRCCoverage%1EtgonME49_Knoll_Laura_Pittman_rnaSeq_RSRCCoverage;h_feat=tgme49_200320%40yellow

Fortunately, TinyURL.com (or other plug-ins) allows us to create a more manageable bookmark; the above URL can be accessed at **http://tinyurl.com/GeneStructureGrowse** … *please navigate to this URL to begin the following exercises.* Note that your screen may differ from the image shown, as not all parameters are stored in the URL. Using the information provided in this exercise, however, you should be able to reconfigure tracks exactly as shown, should you wish to do so. You may also find it helpful to install in your browser a screen capture plugin,


ToxoDB
Toxoplasma Genomics Resource
Release 32
20 Apr 17
EuPathDB Project
Gene ID:
Gene Text Search:
Home
New Search
My Strategies
My Basket (0)
Tools
Data Summary
Downloads
Community
Analyze My Experiment
My Favorites
File
Help
<< 1
Browser
Select Tracks
Snapshots
Custom Tracks
Preferences
<< 2
Search
Landmark or Region:
TGME49_chrVIII:6,780,001..6,800,000
Search
<< 3
Annotate Restriction Sites
Configure...
Go
Save Snapshot
Load Snapshot
4 >>
Scroll/Zoom:
Show 20 kbp
Flip
Overview
5
TGME49_chrVIII
Region
6
Details
7
8
TGME49_chrVIII: 20 kbp
5 kbp
9
Sequence Assembly in Current View
10ab
End-Sequenced Cosmids (Sibley)
End-Sequenced Cosmids (Knoll)
11
Sequence/GC content (zoom to <130 nt for sequence)
12ab
Low Complexity Regions
Tandem Repeats
13a
3-frame translation (forward; zoom to <400nt for sequence)
13b
3-frame translation (reverse; zoom to <400nt for sequence)
14
ORF >= 150 nucleotides (50 amino acids)
15a
Chromatin Immunoprecipitation (H3K4me1) - Active Transcriptome - (blue=H3K4me1)
15b
Chromatin Immunoprecipitation (RH PLK CTG) - Active Transcriptome - (green=H3K4me3; orange=H3K9ac) (Showing 1 of 3 subtracks)
16
Tachyzoite Transcriptome Time Series (ME49) Strand Specific mRNAseq Coverage aligned to T. gondii ME49 (Gregory) (linear scale) (Showing 2 of 24 subtracks)
17
Tachyzoite Transcriptome 3 and 4 days post-infection (VEG NcLIV) mRNAseq Coverage aligned to T. gondii ME49 (Reid et al.) (linear scale) (Showing 2 of 4 subtracks)
18
Transcriptomes of 29 strains during murine macrophage infection mRNAseq Coverage aligned to T. gondii ME49 (Minot et al.) (linear scale) (Showing 6 of 58 subtracks)
19
Transcriptomes of Cat Enteroepithelial Stages (CZ-H3) Strand Specific mRNAseq Coverage aligned to T. gondii ME49 (Hehl Lab) (linear scale) (Showing 4 of 16 subtracks)
20
21
22
23
EST Alignments
All MS/MS Peptides (zoom in <10k for peptide info per sample)
Intron-spanning peptides (Krishna)
Expression profiling probes (11 per gene; 3' biased)
Annotated Transcripts (UTRs in gray when available)
24
TGME49_200320-t26_1 (HXGPRT)
25
Annotated Transcripts w/ CRAIG UTR Prediction
26
Annotated Transcripts w/ CRAIG UTR Prediction (highest score)
27
CRAIG denovo Gene Model Prediction
28a
RNASeq Unified Splice Site Junctions (filtered)
28b
RNASeq Unified Splice Site Junctions (inclusive)
Tachyzoite Transcriptome Time Series (ME49) Strand Specific mRNAseq Coverage aligned to T. gondii ME49 (Gregory) (log scale) (Showing 2 of 24 subtracks)
29a
5.1 - hour 36 (unique forward)
29b
5.3 - hour 36 (unique reverse)
Bradyzoite in vivo transcriptome (M4) Strand Specific mRNAseq Coverage aligned to T. gondii ME49 (Buchholz et al.) (log scale) (Showing 2 of 4 subtracks)
30a
1.1 - bradyzoite (unique forward)
30b
1.3 - bradyzoite (unique reverse)
Transcriptomes of Cat Enteroepithelial Stages (CZ-H3) Strand Specific mRNAseq Coverage aligned to T. gondii ME49 (Hehl Lab) (log scale) (Showing 4 of 16 subtracks)
31a
4.1 - Tachyzoite (unique forward)
31b
4.3 - Tachyzoite (unique reverse)
31c
3.1 - day 7 (unique forward)
31d
3.3 - day 7 (unique reverse)
Oocyst Time Series (M4) Strand Specific mRNAseq Coverage aligned to T. gondii ME49 (Fritz/Boothroyd/Gregory) (log scale) (Showing 4 of 12 subtracks)
32a
1.1 - day 0 (unique forward)
32b
1.3 - day 0 (unique reverse)
32c
3.1 - day 10 (unique forward)
32d
3.3 - day 10 (unique reverse)
Tachyzoite Transcriptome Time Series (GT1) Strand Specific mRNAseq Coverage aligned to T. gondii ME49 (Gregory) (log scale) (Showing 8 of 16 subtracks)
Tachyzoite Transcriptome Time Series (RH) Strand Specific mRNAseq Coverage aligned to T. gondii ME49 (Gregory) (log scale) (Showing 6 of 12 subtracks)
Tachyzoite Transcriptome Time Series (VEG) Strand Specific mRNAseq Coverage aligned to T. gondii ME49 (Gregory) (log scale) (Showing 12 of 24 subtracks)
Transcriptomes of 29 strains during murine macrophage infection mRNAseq Coverage aligned to T. gondii ME49 (Minot et al.) (log scale) (Showing 29 of 58 subtracks)
Transcriptome during acute or chronic infection in mouse brain mRNAseq Coverage aligned to T. gondii ME49 (Pittman et al.) (log scale) (Showing 2 of 4 subtracks)
33
34
35
36
37
Select Tracks
Clear highlighting
EuPathDB GBrowse Tutorial
ToxoDB 32 20 Apr 17
©2017 The EuPathDB Project Team
EuPathDB
Strategies WDK
Please Contact Us with any questions or comments

ToxoDB GBrowse v2.48: 20 kbp from TGME49_chrVIII:6,780,001..6,800,000

To save or share track configurations, select 'Generate URL' from File menu (above) and cut and paste resulting URL

If you load tracks and they appear empty, please try two things:
1. Make sure you are viewing the correct species/strain to which the data was mapped.
2. Reset your GBrowse and try again.

such as Awesome Screenshot, which was used to grab the figure shown below as a single image.

This genome segment encompasses the *Toxoplasma* HXGPRT gene, TgME49_200320, highlighted in yellow in the annotation tracks, near the middle of the page (**labeled 24-26 in *magenta***). As you learned in the Genome Browser exercise, tracks can be dragged and dropped elsewhere on the page to reorder … *try this!* In general, new tracks you turn on won't appear where you want them, but configurations are stored as cookies on your computer, and will persist from one session to the next (although this information may be lost in new releases).

Track names are highlighted in orange (hidden tracks are shown in gray). Small icons to the left of each track name allow you to flag, delete, share, reconfigure, or get more information. The tool icon (wrench) controls track height, labeling, color, *etc.* Clicking the link at right (***Showing … subtracks***) allows you to show/hide, reorder or overlay sub-tracks … *try this!*

As you already know, the **Browser** tab at the top of the page (**2**) displays the graphical view shown here; **Select Tracks** shows available datasets … *take a few minutes to explore the datasets currently available in ToxoDB.org*. You may also wish to explore datasets available in your home database (FungiDB.org, PlasmoDB.org, TriTrypDB.org, *etc*). You might also want to consider what other datasets would be useful for your research, and whether there are additional datasets in the public domain that should be integrated into ToxoDB or other databases (let us know about these by clicking on the 'Contact Us' link).

Snapshot allows you to save this view, if you are logged in, but note that this may not faithfully reflect your display (for example, hidden tracks are opened; Awesome Screenshot or other browser plugins may be more effective). **Custom Tracks** allows you to upload your own data, as we will do in the RNA-seq mapping exercise. **Preferences** allows you to configure your display, including image width and highlighted regions. The image shown was set to width=1024 (which may be larger than your display will support unless you have a high resolution or large-screen monitor).

Users commonly reach the Genome Browser via individual Gene Pages (*e.g.* TgME49_200320), but note that you can also enter specific chromosomes, contigs, or regions of interest. *Try changing the region displayed,* for example by focusing on HXGPRT (change 6,780 to 6,790 or 6,795; **3**). In the later exercises, you will probably want to navigate to specific regions of interest. For example, you may wish to examine the junctional region between TgME49_200295 and TgME49_200300. Recall from the Genome Browser exercise that you can also zoom in and out, or scroll left or right using the menus and buttons at right (**4**). You can also click in individual features (*e.g.* TgME49_200310), or click on any of the rulers (**5,6,8**) and drag to zoom in on a particular region.

Let's return to the genome annotation (**24-27**) for further analysis. EuPathDB databases use the convention that genes transcribed from left to right (*i.e.* on the top strand) are colored blue (*e.g.* TgME49_200320) and those transcribed from right to left are colored red (*e.g.* TgME49_ 200310). Track **24** presents the current annotation. Three additional tracks display alternative gene predictions generated by the CRAIG gene finder: one (**25**) adds the longest UTR predictions to the existing annotation, another one adds UTR predictions with high confidence (score) (**26**) to the existing annotation and the other presents *de novo* predictions (**27**). Craig gene models (**25,27**) suggest a longer 3' UTR than that shown in the official annotation, for both TgME49_200320 *… which 3' UTR do you think is correct? What evidence would you require to decide?*

Above the annotation, four sets of visible tracks display mRNA-seq data, using a **linear** vertical scale (**16-19**), *i.e.* transcripts represented by twice as many reads are twice as high. *Compare the observed transcript abundance with introns and exons in the TgME49_ 200320 (HXGPRT) gene model … is this what you expected?* Note the ruler at left (**7**), which you can activate by clicking and dragging to identify precise coordinate and facilitate the analysis of vertical feature alignment across tracks, allowing you to compare how RNA-seq read abundance maps to the HXGPRT annotation.

How do you explain the heterogeneity in abundance, across the transcript, and even within a single exon? Can you see evidence for alternative splicing in TgME49_ 200320? Note that the HXGPRT gene is known to be alternatively spliced, sometimes reading through the first intron (located within the 5' UTR), and sometimes skipping the third exon, removing an acylation domain responsible for membrane association of one HXGPRT protein isoform.

The track (**Tachyzoite Transcriptome, strain ME49; 16**) presents strand-specific data (blue = forward; red = reverse), overlaid. *Click on the tool icon (wrench) to learn how to undo (or redo) the semitransparent overlay. How do these data compare with the track below? How do you* *explain any differences?* Note that transcript reads for this track extend further to the right, explaining in the longer UTR prediction in **25**. *Is this appropriate?*

The next track (**17**) displays nonstrand-specific mRNA-seq data for parasites cultivated *in vitro* for **3 vs 4 days** (note that in this instance red and blue reflect different time points, not different strands). *How do you interpret the observed differences, if any?*

Track **18** (**Transcriptomes of 29 strains**) presents a transparent overlay of tachyzoite gene expression in 10 strains. Not much is visible here … *but try removing the transparent overlay mode to examine more closely.* If you are particularly interested in strain-specific differences, you may wish to look at additional strains as well, zoom in or out, or move to different regions of this (or other) chromosome(s). Note that all of these sequences have been mapped to the ME49 reference genome … *does this concern you (why or why not)? What additional information would be helpful for analyzing cross-strain mapping?*

The last track of linear mRNA-seq data presented in this figure overlays four datasets: **forward** and **reverse** strand data from **tachyzoites** and **d7 gametocytes** (**19**). In the overlay, the yellow track is most abundant. *What does this tell you about stage-specificity of HXGPRT expression?*

Now skip down to red and blue tracks displayed below the Annotation tracks (**29-32**). These are additional RNA-seq datasets, from strand-specific sequencing experiments, but presented on a **log-2** vertical scale (**16-19**), *i.e.* transcripts represented by twice as many reads are just one unit higher. The first set (**29a&b**) corresponds precisely to the data shown in track (**16**). Similarly, track (**31a-d**) corresponds precisely to track (**19**), but on a log scale, rather than linear scale (and without the semi-transparent overlay. *How does log vs linear representation of the data affect your interpretation of gene model accuracy, splicing, transcript abundance (and the uniformity of coverage)?* mRNA-seq data is most commonly displayed on a linear scale … *which representation do you prefer?*

These tracks display data from different life cycle stages: **tachyzoites** (ME49 strain, cultivated *in vitro*; **29**), **bradyzoites** (M4 strain, isolated from mouse brains; **30**), **day 7 gametocytes** (Cz-H3 strain, isolated from feline intestinal epithelium, along with tachyzoite controls from the same strain; **31**), and **unsporulated (day 0)** and **sporulated (day 10)** oocysts (M4 strain sporozoites; **32**). *Do you see evidence of stage-specific expression? Are you concerned that these sequencing data derive from different strains* (some of whose genomes have not been sequenced), but all are mapped to the ME49 reference? Five additional datasets are hidden at the very bottom (**33-37**), and other tracks may be turned on or off, if you are interested in exploring further.

Finally, return to the **Splice Site Junctions** tracks (**28a&b**), located immediately below the Annotation tracks. These are probably the most useful tracks for evaluating gene models, including intron annotation, as it presents RNA-seq reads that span a gap (presumably due to intron excision) … from *all* available RNA-seq experiments (**28a)**. Color intensity indicates the total number of intron-spanning reads, and mousing over the spans indicates the distribution by experiment (**Select Tracks** also allows you to display separate tracks for individual experiment). *Do these data support the published annotation of alternative splicing of HXGPRT, as described above? Is there any evidence of stage or strain-specific alternative splicing?*

Note that there are a lot of putative introns: 2.7 million in the genome, <4% of which are annotated! Indeed, if you sequence deep enough, *all* genes (even genes that likely include only a single exon), and many unannotated regions, show evidence of splicing, at least at low abundance. *What do you make of the additional candidate introns associated with TgME49_200320? Do you believe them all? Do you believe any of them? Should the official GenBank annotation for TgHXGPRT be changed?*

Several other genes are also visible upstream of TgME49_200300 (HXGPRT) in the above figure. Returning to this larger view (TGME49_chrVIII:6780001..6800000), *what do you make of gene TgME49_200310? Does it contain introns? Is the annotated gene model correct? Is there evidence of alternative splicing? Why are there such extreme disparities in annotation of the 3' UTR? Why might the Craig gene finder have inserted an intron into the UTR, when there is no evidence for any corresponding intron-spanning reads?*

What about TgME49_200300 & TgME49_200295? *Is this one gene or two … or perhaps not a gene at all:note that expression is vanishingly low (<1% the level of HXGPRT)? And what do you make of the region around 6790K,* that is not currently annotated as a gene. This region appears to be transcribed from one strand in tachyzoites (**29a,30a,31a**) and sporozoites (**32c**), but the other strand in gametocytes (**31d**)!

In analyzing these results, you may also wish to display other lines of evidence, including the sequence of Expressed Sequence Tags (ESTs) (**20**), peptides observed by tandem mass spectrometry (**21**), including peptides that span annotated introns (**22**). In addition, you might wish to consider prior studies conducted by array hybridization … in which case it will be important to know what probes were used for those studies (**23**).

You might also need to understand more about the context of your gene(s) of interest, including the genomic sequence, chromatin marks, *etc.* The **Overview** panel at top (**5**) displays the entire chromosome or contig (~7 Mb in this case), highlighted to show the **Region** of interest (**6**), which is, in turn, highlighted to show under **Detail** (**8**) the region specified in (**3**). Some datasets at the bottom of the list under the **Select Tracks** tab can be displayed in these panels (assembly, centromere, gene density, *etc*). *Try turning on these tracks for the Overview,* to examine completeness of the chromosomal assembly, centromere location, gene density, *etc*. Note the importance of understanding genome assembly quality … if your gene/region of interest contains gaps, your interpretation will likely be flawed.

Two tracks display **Cosmid End Sequences** (**10**). *What are cosmids? Why might these be of interest* (in general, and in this particular application for evaluating genome assembly and gene models)? *Try zooming out from 20 kb to 200 kb to see how the picture changes … why do horizontal bars appear for some cosmid ends but not others.*

Other tracks may also be useful, such as those displaying **GC Content** (**11**)**, Low Complexity Regions** (**12a**) and **Tandem Repeats** (**12b**), **3-frame translations** (forward & reverse) (**13a&b**) and **ORFs** (open reading frames) >150nt (**14**)? *How might the presence of low complexity regions affect the uniformity of RNA-seq mapping results? How does the presence (or absence) of open reading frames affect your assessment of gene models?* Note that many tracks change their displays at different levels of resolution. Try zooming in from 20kb to 200nt to see how 3-frame translations are displayed (what would you expect to see)?

Two tracks in the figure illustrate chromatin mark data from immunoprecipitation experiments: **H3K4Me1** (**15a**) and **H3K4Me3 + H3K9ac** (**15b**). *What does the relationship of these peaks to each other suggest*, based on comparing the observed patterns with the **Annotated Genes** track below (or based on your prior knowledge of chromatin mark function)? To get a better feel for these datasets, try zooming out to 200 kb, and try adding additional subtracks. *How do you think this picture might change with ChIP-seq data* (not currently available for *T. gondii*)?

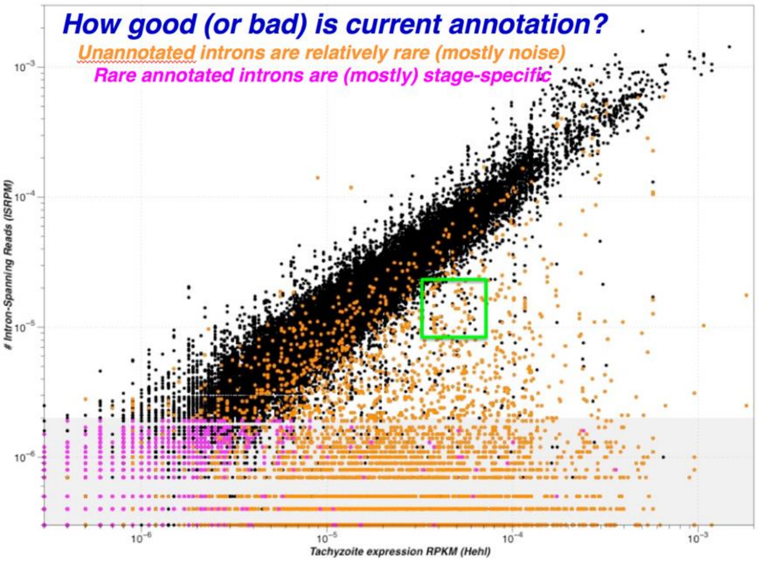
On cautionary note: remember that all of these experiments reflect studies on steady-state transcript abundance. *What datasets would you need to generate to assess transcription rates, rather than steady-state levels?* Are any such datasets available for *Toxoplasma?* What about fungal species, or *Plasmodium* (check out the available datasets under **Select Tracks** in the genome browser within PlasmoDB)?

Group exercise. From the above exercises, we know that hundreds of thousands of introns identified by RNA-seq experiments are not represented in the reference *T. gondii* annotation … but most of these are observed at lower levels (often *far* lower) than expected based on transcripts mapping to the annotated coding sequence.

The graph at below shows each putative intron in the entire genome, plotted based on the number of intron-spanning reads detected per million RNA-seq reads (on the vertical axis), and the number of reads mapping to predicted gene coding sequence (on the horizontal axis, normalized to account for differences in gene size). Annotated introns are shown as black dots; unannotated introns are colored orange.

The majority of *annotated* introns are represented in RNA-seq data at the frequency expected based on reads mapping to the annotated gene as a whole. Moreover, many annotated introns not expressed in tachyzoites are expressed in other life cycle stages (pink dots).

Unannotated introns fall in the lower part of this graph: while most are reproducibly observed in multiple experiments, they are less frequent (usually far less frequent) than reads corresponding to annotated introns. These are probably the molecular equivalent of typographical errors, although of course the possibility that some may be functionally significant under appropriate conditions cannot be excluded.



Functional alternative splicing (*e.g.* the excision of introns 2 & 3 in

HXGPRT, *vs* the exon skip polymorphism resulting in a single larger intron) would be expected to fall just slightly below the diagonal black line. The following list includes genes represented by the green box in the above figure, i.e. candidate instances of alternative splicing (this list also includes some genes that display possible alternative splicing in tachzyzoites, but not in gametocytes, or vice-versa.

Working in groups of four, please select at least two genes from this list to evaluate, 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.

TgME49_200320 (HXGPRT)
TGME49_246490
TGME49_256650
TGME49_283540
TGME49_226410
TGME49_225730
TGME49_213610
TGME49_213660
TGME49_297160
TGME49_211250
TGME49_211420
TGME49_214440
TGME49_250115
TGME49_261720
TGME49_268610
TGME49_270520
TGME49_280380
TGME49_293720
TGME49_248445
TGME49_230180
TGME49_281440
TGME49_279390
TGME49_202770
TGME49_217490
TGME49_292150
TGME49_276170
TGME49_266640
TGME49_266920
TGME49_299010