Interpreting RNA-seq data (*beta*)

You spent some time yesterday afternoon learning to use the GBrowse genome browser, familiarizing yourself with available datatypes, and track configurations. In last night's lecture, you also learned that gene models (structural annotation) are often open to interpretation, especially with respect to:

- transcript initiation and termination sites (5' and 3' UTRs)
- low abundance processing events (if you sequence deep enough, it is likely that *all* genes from organisms excise introns display alternative splicing)
- the potential significance of non-coding RNAs

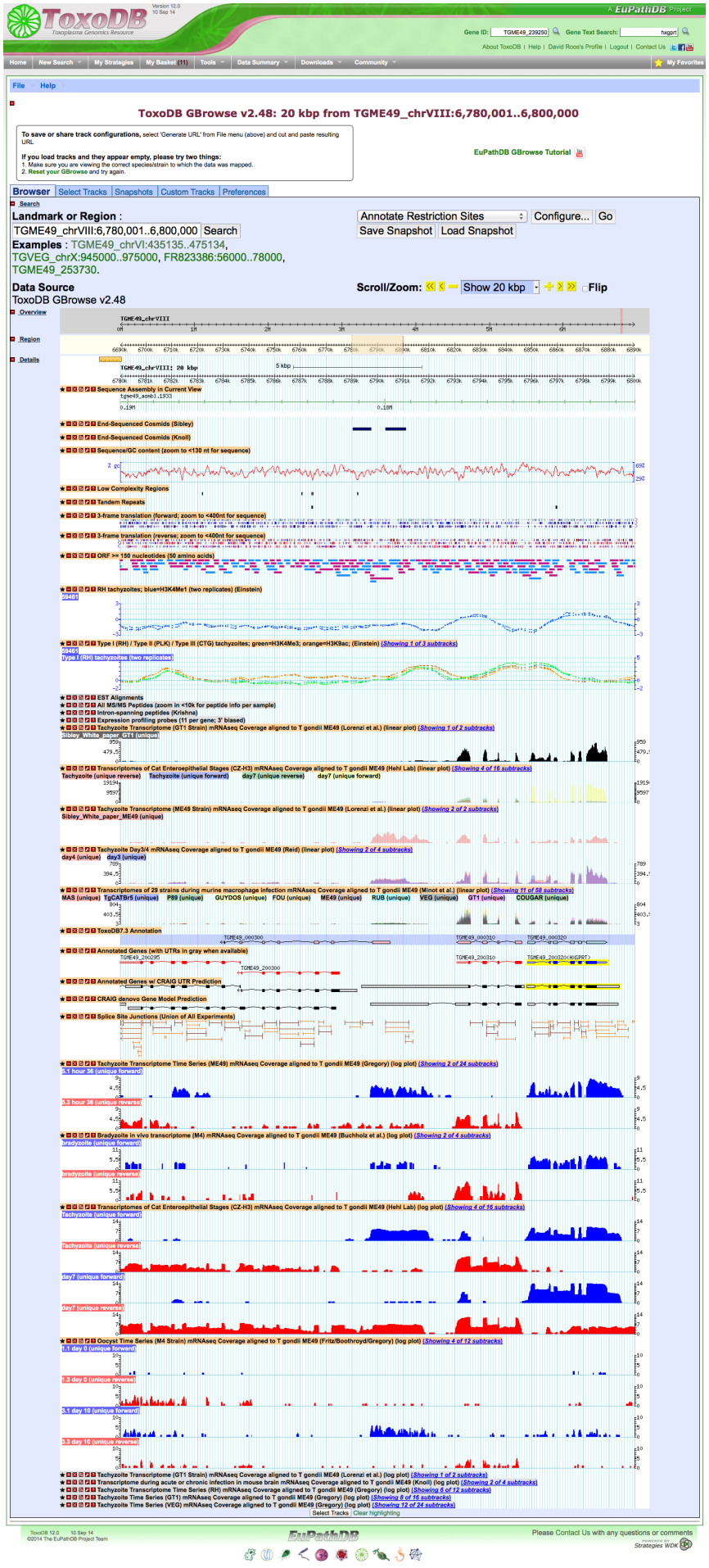
Today, we will explore track configuration options in much greater detail, focusing on the interpretation of RNA-seq datasets, using this information to examine the differentially-spliced HXGPRT gene of *T. gondii.*

You will apply your newfound skills to examine other instances of possible alternative splicing … and report your findings back to the group as a whole.

The figure on the following page (and also available for download) presents one example of an extensively configured GBrowse page used to examine alternative splicing in *Toxoplasma.* Many tracks have been turned on, some have been hidden, others reconfigured, diverse sub-tracks shown and/or overlaid, resolution of the display page has been increased, etc. Establishing this configuration took more than an hour … time well-spent if it is a configuration that you will want to reuse! As indicated at the top of all GBrowse pages, you can save a track configuration by selecting 'Generate URL' from the File menu at the top of the page. 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=1fa04ecd2a3ef5c0acd569703feda634;l=Scaffolds%1ECosmidsSibley%1ECosmidsLorenzi%1EGC%20Content%1ELowComplexity%1ETandemRepeat%1ETranslationF%1ETranslationR%1EORF%1EChIPEinsteinME1%1EChIPEinstein%1EEST%1EUnifiedMassSpecPeptides%1ERiteshPeptide%1EAffymetrixExpressionNuclearCoding%1ETgonME49_Sibley_White_paper_GT1_rnaSeq_RSRCCoverageUnlogged%1ETgonME49_DBP_Hehl-Grigg_rnaSeq_RSRCCoverageUnlogged%1ETgonME49_Sibley_White_paper_ME49_rnaSeq_RSRCCoverageUnlogged%1ETgonME49_Reid_tachy_rnaSeq_RSRCCoverageUnlogged%1ETgonME49_Saeij_Jeroen_strains_rnaSeq_RSRCCoverageUnlogged%1EOldVersionGenes%1EGene%1Eutr_only_union%1Edenovo_union%1ERUMIntronUnified%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_Sibley_White_paper_GT1_rnaSeq_RSRCCoverage%1ETgonME49_Knoll_Laura_Pittman_rnaSeq_RSRCCoverage%1ETgonME49_Gregory_RH_mRNA_rnaSeq_RSRCCoverage%1ETgonME49_Gregory_GT1_mRNA_rnaSeq_RSRCCoverage%1ETgonME49_Gregory_VEG_mRNA_rnaSeq_RSRCCoverage;h_feat=Tgme49_200320%40yellow

Fortunately, you can also use TinyURL or other plug-ins to create a more manageable bookmark. The gigantic above can also be accessed at **http://tinyurl.com/lgomads** . You may also wish to install in your browser a screen capture plugin, such as Awesome Screenshot, which was used to grab the figure below as a single image.



Please navigate to the above (short) URL for the following exercises, questions and comments, which are (mostly) organized from top to bottom with respect to the figure (but not necessarily in order of importance!) They address issues you may wish to think about … but note that it is not necessary to explore them all in detail!

First, examine the entire page returned … it may differ in some respects from the image shown, as not all parameters are stored in the URL. Using the information provided below, however, you should be able to reconfigure tracks as desired, if you wish to do so.

The HXGPRT gene is highlighted in yellow, in a track displaying current curated annotation near the middle of this page. Recall from yesterday's exercises that individual tracks can be dragged and dropped to change their position on the page. In general, tracks you turn on will NOT appear where you want them, but configurations are stored as cookies on your computer, and will persist from one session to the next. (This information is lost with each new release, however, and at present there is no way to save your preferred track displays. If this is important to you, please say so by clicking the Contact Us button at the top of the page!)

You can also use tools shown as buttons associated with the orange track names to get more information, show, hide, remove, share, or reconfigure tracks. Hidden tracks are indicated in gray; the tool icon allows you to show, hide or reorder substracks, change track height, labeling, display colors, etc.

Note the tabs at top: **Browser** presents the graphical view, **Select Tracks** displays available datasets … open this tab and examine the datasets currently available in ToxoDB.org. *You may also wish to examine datasets available in other component databases, e.g. TriTrypDB.org. What other datasets are currently available that should be included in these databases? What other datasets would be useful for your research?* **Snapshot** allows you to save this view (if you are logged in) … but note that it may not faithfully reflect your display. For example, hidden tracks are opened. As noted above, you may find Awesome Screenshot and other browser plugins to be more effective. **Custom Tracks** allows you to upload your own data … as we will do in the RNAseq mapping exercise later today and tomorrow. **Preferences** allows you to configure your display, including the 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. The HXGPRT gene TgME49_200320 is highlighted in yellow.

While you may often navigate to the Genome Browser from an individual Gene Page (e.g. TgME49_200320), you can also enter specific chromosomes, contigs, or regions of interest. Try changing the region displayed. In the later exercises, you will probably want to navigate to specific regions of interest. Recall from yesterday's exercises that you can also zoom in and out, and scroll left or right using the menus and buttons at right.

The **Overview** panel displays the entire chromosome (or contig) -- ~7 Mb in this case – highlighted to show the **Region** of interest. Some datasets at the bottom of the Select Tracks list can be displayed in these panels. Try turning on these tracks for the Overview, to examine completeness of the chromosomal assembly, centromere location, gene density, etc (SNP density may not function properly at present … if this is important to you, use the Contact Us link!) *It is particularly important to be aware of assembly gaps … if your gene/region of interest contains gaps, your interpretation will likely be flawed!*

Let's briefly jump to the middle of the page, to examine current **Annotated Genes**, showing the HXGPRT gene (TgME49_200320, highlighted in yellow), and several adjacent genes. As you know from yesterday, EuPathDB databases use the convention that genes indicated in red are transcribed from right to left (on the bottom strand), and those in blue from left to right. The track above displays previous annotation from **ToxoDB release 7.3**, which is identical for TgME49_200320 (formerly TgME49_000320), but differs for TgME49_000300 (compare with TgME49_200300 & TgME49_200295). Two additional tracks display alternative gene predictions generated by the CRAIG gene finder: one adding UTR predictions to the annotated genes, and the other making *de novo* predictions. In your evaluation of different genes, you may wish to consider the performance of these differing algorithms.

Returning to the top of the **Details** tracks, note the ruler at left, which you should try clicking and dragging as a marker to facilitate the analysis of feature coordinates and alignment across tracks. In addition to using the tools noted above for navigation, you can also click and drag to define a region to zoom in, or click on individual features in various tracks for more data.

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

Why might tracks relating to **GC Content, Low Complexity Regions** and **Tandem Repeats** be useful? How might the presence of low complexity regions affect the uniformity of RNA-seq mapping results?

Why might tracks displaying **3-frame translations** (forward & reverse) and **ORFs** (open reading frames) >150nt be of interest? 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 are shown illustrating chromatin mark data from chromatin immunoprecipitation experiments: **H3K4Me1** and **H3K4Me3 + H3K9ac** data. Why do you think that these datasets are grouped in this way … either based on your prior knowledge of function, or based on what you see by comparing the observed patterns with the **Annotated Genes** track below? To get a better feel for these datasets, try zooming out to 200 kb, try clicking on the tool icon in the orange bars to add additional subtracks. You may also wish to explore additional tracks (open the **Select Tracks** tab). How do you think this picture might change with ChIP-seq data (not currently available for *T. gondii*)?

The next four tracks in the above figure are hidden: **EST Alignments, MS/MS Peptides, Intron-spanning peptides,** and **Expression profiling probes.** What are these datasets, and why might (or might not) they be useful for evaluating gene model preditions shown in the Annotation track(s)? Feel free to open, reconfigure, etc, if interested.

The following five tracks display RNA-seq data on a linear vertical scale. The first (**White paper GT1 unique**). Shows significant expression of TgME49_200310 & TgME49_200310 (labeled on the Annotation track, below). Why are these peaks so heterogeneous in height? Why are they not shown in red for TgME49_200310 and blue for TgME49_200310? Can you see evidence for the published fact that HXGPRT is alternatively spliced … some transcripts lack exon 3, and some read through intron 1 (in the 5' UTR)?

The next track overlays four datasets: **forward** and **reverse** strand data from **tachyzoites** and **d7 gametocytes**. Click on the tool icon to learn how to undo (or redo) the transparent overlay. In the overlay (as shown in the figure), the HXGPRT gene shows up in yellow. Why?

The third track in this set displays parasites cultivated *in vitro* for **3 vs 4 days** (blue and red, respectively, overlaid). How do you interpret the observed differences, if any?

The next track (**Sibley white paper**) shows one track of ME49 tachyzoite transcriptomic data … but the orange bar says that 2 subtracks are shown. Why are they not visible? *Hint: zoom out to 50 kb and look again!* Why might this display be useful (in theory, if not in this particular instance?

The last track of linear RNA-seq data (**Transcriptomes of 29 strains …**) shows a transparent overlay of tachyzoite gene expression in 11 strains. Not much is visible here … but try clicking on the tool icon to reconfigure the scale, and try adding additional subtracks (probably best *not* viewed in transparent overlay mode). Do you observe any strain-specific differences in the isolates shown? Try zooming out and moving to other regions of this (or other) chromosome(s) … how common is strain-specific expression? Does it concern you that all of these sequences have been mapped to the ME49 reference genome? Why or why not? What additional data would be helpful to alleviate possible concerns on this score?

The next track (**Sibley white paper**) shows one track of ME49 tachyzoite transcriptomic data … but the orange bar says that 2 subtracks are shown. Why are they not visible? *Hint: zoom out to 50 kb and look again!* Why might this display be useful (in theory, if not in this particular instance?

Now skip down to the four sets of red and blue tracks below the Annotation tracks. These are additional RNA-seq datasets, from strand-specific sequencing experiments, and displayed on a log scale. Five additional datasets are included at the very bottom; all are hidden in the figure, but you may wish to open them to explore further.

What additional information can you glean from the first (**Tachyzoite**) datasets? How does the representation of TgME49_200320 change (you may wish to turn on log and linear representations of the exact same datasets, and move tracks next to each other for comparison)? Why are these graphs less 'spiky' than the linear representations discussed above? Does your interpretation of the data change, especially with respect to putative exons, alternative splicing to remove or include exon 3? Should the existing annotation in GenBank be changed for TgME49_200320? What about TgME49_200310? TgME49_200300?

The following three sets of tracks display data from different life cycle stages: **bradyzoites** (M4 strain, isolated from mouse brains), **day 7 gametocytes** (Cz-H3 strain, isolated from feline intestinal epithelium, along with tachyzoite controls from the same strain), and **unsporulated (day 0)** and **sporulated (day 10)** oocysts (sporozoites, M4 strain). Why do these various datasets show differing degrees of smoothness? Do you see any evidence of stage-specific expression? Are you concerned that these sequencing data from different strains (whose genomes have not been sequenced) are all mapped to the ME49 reference strain? Considering TgME49_200320, what conclusions do you draw about alternative splicing of HXGPRT? What about strain-specific transcription of this gene? What about stage-specific expression?

What conclusions do you draw about TgME49_200310? What about TgME49_200300? What about other regions shown? Any evidence of non-coding RNAs? What about other regions in the genome? Feel free to explore your favorite gene(s)!

Finally, return to the **Splice Site Junctions (Union of All Experiments)** track, located immediately below the Annotation tracks. This is perhaps the single most useful track for evaluating structural gene models, including intron annotation, as it sums *all* intron-spanning reads (RNA-seq reads that jump a gap, presumably due to intron excision) … from *all* available RNA-seq experiments. There certainly are a lot of possible introns, corresponding to all annotated genes and even many unannotated regions (supporting the contention that if you sequence deep enough, all genes show evidence of alternative splicing. Color intensity indicates the total number of intron-spanning reads, and mousing over the spans indicates the distribution by experiment. You may also wish to open additional tracks for individual experiments (under the **Select Tracks** tab).

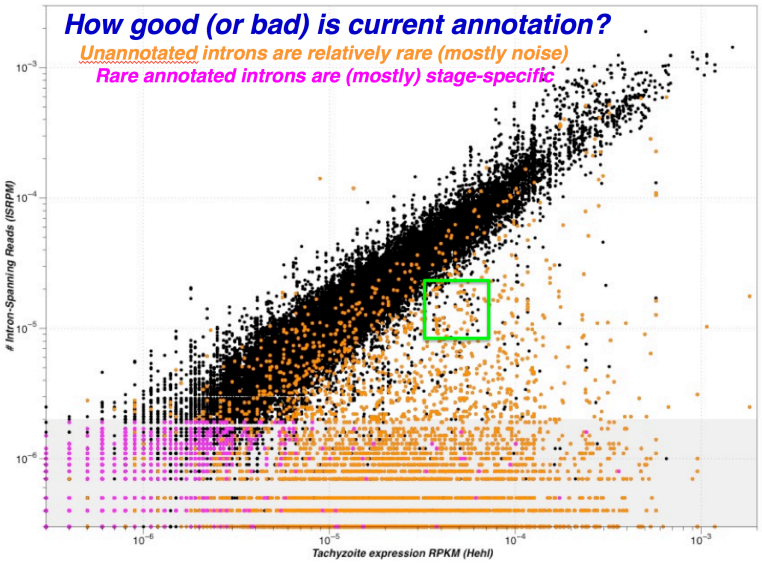
Do these data support the published annotation of alternative splicing of HXGPRT? What do you make of additional candidate introns associated with TgME49_200320? Do you believe them all? Do you believe any? Is there any evidence of stage or strain-specific alternative splicing? Should the existing annotation in GenBank be changed for TgME49_200320? What about TgME49_200310? TgME49_200300?

How do you interpret the annotation of TgME49_200310? Is the existing annotation correct? Why do you think the CRAIG UTR track shows such a long 3' UTR? Why do you think that the CRAIG *de novo* track shows a 1 kb intron in the 3' UTR?

Note 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 for *Plasmodium* (check out the available datasets under **Select Tracks** in the genome browser within PlasmoDB).

Group exercise. In last night's lecture – and from the above exercises – we saw that while many thousands of introns identified by RNA-seq experiments are not represented in the reference *T. gondii* annotation, most of these are observed at lower levels (often *far* lower) than expected based on transcripts mapping to the annotated coding sequence.

On the graph at right, each putative intron in the entire genome is represented by a single dot, 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 shown in orange.



It is clear that the vast majority of annotated introns are represented in RNA-seq at the same frequency as reads that map entirely within a single exon. Most introns that are annotated but not expressed in tachyzoites, turn out to be expressed in gametocytes (pink dots).

Unannotated introns fall in the lower part of this graph: while most are reproducibly observed in multiple experiments, they are far less frequent than reads corresponding to annotated introns. These are probably the molecular biological equivalent of typographical errors … although it cannot be ruled out that some may be functionally significant, under appropriate considtions.

Functional alternative splicing (such as intron 2 & 3 in HXGPRT, and the larger intron 2-3, which elimitates exon 3) 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_219485	TgME49_260270
TgME49_201260	TgME49_222060	TgME49_262630
TgME49_201820	TgME49_224520	TgME49_263070
TgME49_202770	TgME49_225120	TgME49_263100
TgME49_205430	TgME49_229010	TgME49_266080
TgME49_208740	TgME49_230180	TgME49_266640
TgME49_209950	TgME49_234410	TgME49_266920
TgME49_210700	TgME49_235920	TgME49_270520
TgME49_211330	TgME49_236620	TgME49_271610
TgME49_211630	TgME49_239560	TgME49_278510
TgME49_212200	TgME49_239700	TgME49_278830
TgME49_213030	TgME49_242415	TgME49_279390
TgME49_213325	TgME49_242570	TgME49_286932
TgME49_213460	TgME49_245660	TgME49_289910
TgME49_214920	TgME49_247195	TgME49_294400
TgME49_217510	TgME49_252210	TgME49_305290
TgME49_218490	TgME49_253170	TgME49_309420
TgME49_218830	TgME49_253690	TgME49_309980
TgME49_218910	TgME49_254470	TgME49_313480
TgME49_219230	TgME49_258880	TgME49_315860