WebApollo: A Web-based Sequence Annotation Editor for Distributed Community Annotation

Ed Lee[1], Gregg Helt[1], Nomi Harris[1], Robert Buels[3], Christopher Childers[2], Justin Reese[2], Mónica Muñoz-Torres[2], Christine Elsik[2], Ian Holmes[3], and Suzanna Lewis[1]

1) Berkeley Bioinformatics Open-source Projects, Lawrence Berkeley National Laboratory, Berkeley, CA; 2) Georgetown University, Washington DC; 3) University of California at Berkeley, Berkeley, CA

Summary

As technical advances make sequencing faster and cheaper, genomic annotation efforts must adapt to keep pace. The upward trend in the number of genome sequencing projects means there will be a larger reliance on contributions from domain specialists. Thus the curation environment is shifting from a traditional centralized model, in which all curators for a given genome project share the same physical location, to a geographically dispersed community annotation model—which requires new tools to support community annotation efforts. WebApollo was designed to provide an easy to use, web-based environment that allows multiple distributed users to edit and share sequence annotations.

Curators and investigators from the bee genome research community are currently beta-testing WebApollo. Their curation efforts, findings and interactions will dramatically upgrade the quality of the annotation data for the genomes of honeybee (Apis mellifera), and two bumble bees (Bombus impatiens and Bombus terrestris), which will lead to a better understanding of the biology of these social insects.

WebApollo is comprised of three components: a web-based client, a server-side annotation editing engine, and a server-side service that provides the client with data from different source databases. All three software components are open source and freely available.

The web-based client is designed as an extension to JBrowse, a JavaScript-based genome browser that provides a fast, highly interactive interface for the visualization of genomic data. This JBrowse extension provides the gestures needed for editing annotations, such as dragging and dropping features to create new annotations of genes, transcripts and other genomic elements, dragging to change exon boundaries of existing annotations, and using context-specific menus to modify features. It has support for deep sequencing visualization (e.g., BAM data). The extension also connects to the annotation-editing service and the data-providing services.

The server-side annotation-editing engine is written in Java. It handles all the necessary logic for editing and deals with the complexities of modifications in a biological context, where a single change can have multiple cascading effects (e.g., when splitting or merging transcripts). Edits are stored persistently in the server, allowing users to quickly recover their data in the event of unexpected browser or server crashes. The server provides synchronized updates over multiple browser instances, so that every edit is immediately visible to all users who are viewing or editing the same region. It offers multiple levels of user accessibility, allowing project owners to decide with whom to share their work, and whether to allow read-only or both read and write access.

The server-side service that provides data to the client is built on top of Trellis, a Distributed Annotation System (DAS) server framework. It sends JBrowse-supported JavaScript Object Notation (JSON) data, rather than the more verbose DAS XML. We developed Trellis plugins to access data from the UCSC MySQL genome database, Ensembl DAS services, and GMOD Chado databases.

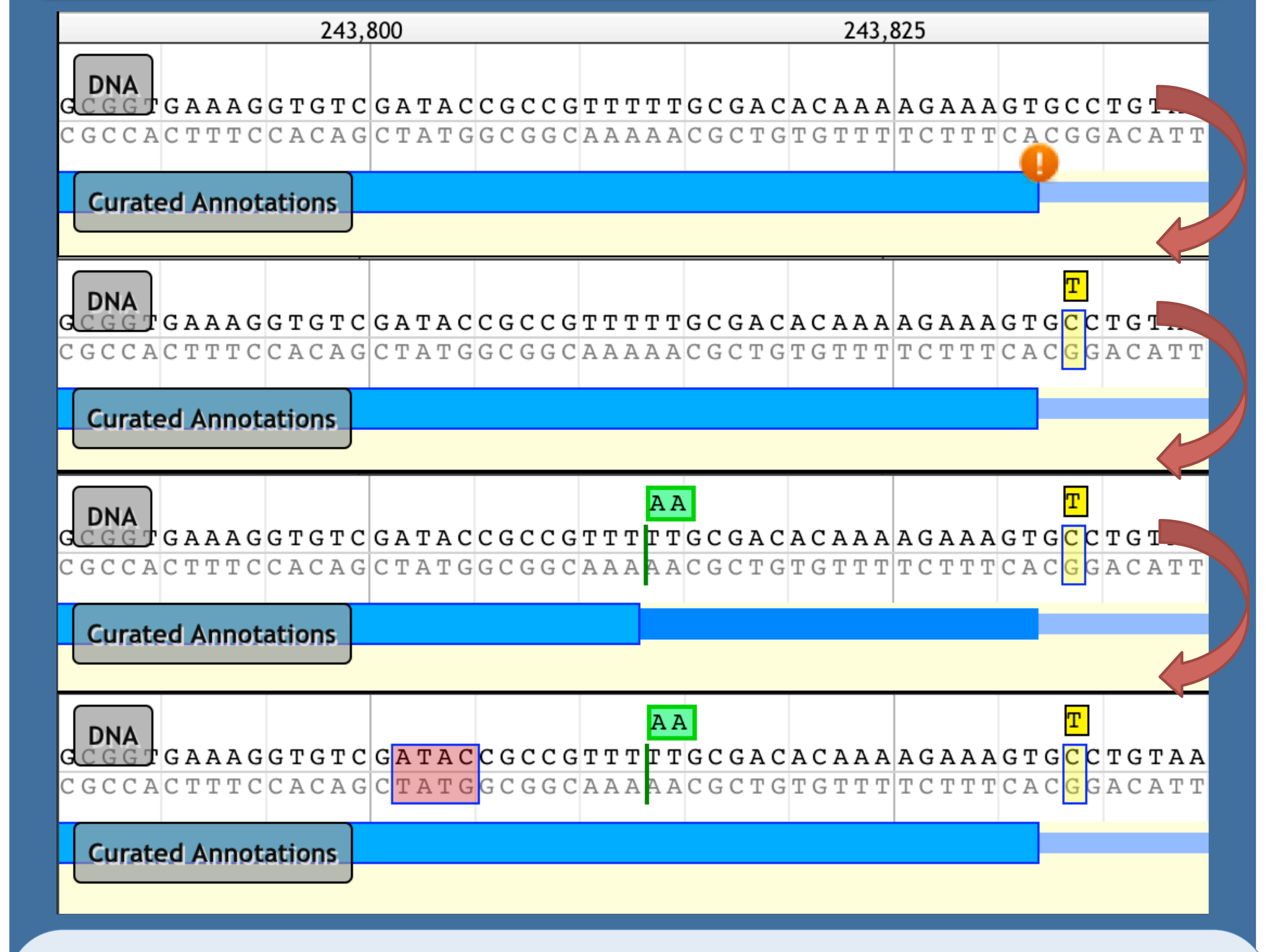
WebApollo will be publicly released in the 4rth quarter of 2012

Public demo at http://icebox.lbl.gov:8080/ApolloWebDemo

More info at http://gmod.org/wiki/WebApollo

Curating Genomic Sequence Alterations

Genome assemblies with lower sequencing coverage can often have small errors in the assembled genomic sequence. These are often first spotted by curators, when the errors cause problems with gene structures. To enable curation in the presence of genome sequencing errors, WebApollo allows curators to annotate genomic sequence insertions, deletions, and substitutions. These annotations do not alter the underlying assembly on the server. However they are taken into account when determining the sequence of an annotation.

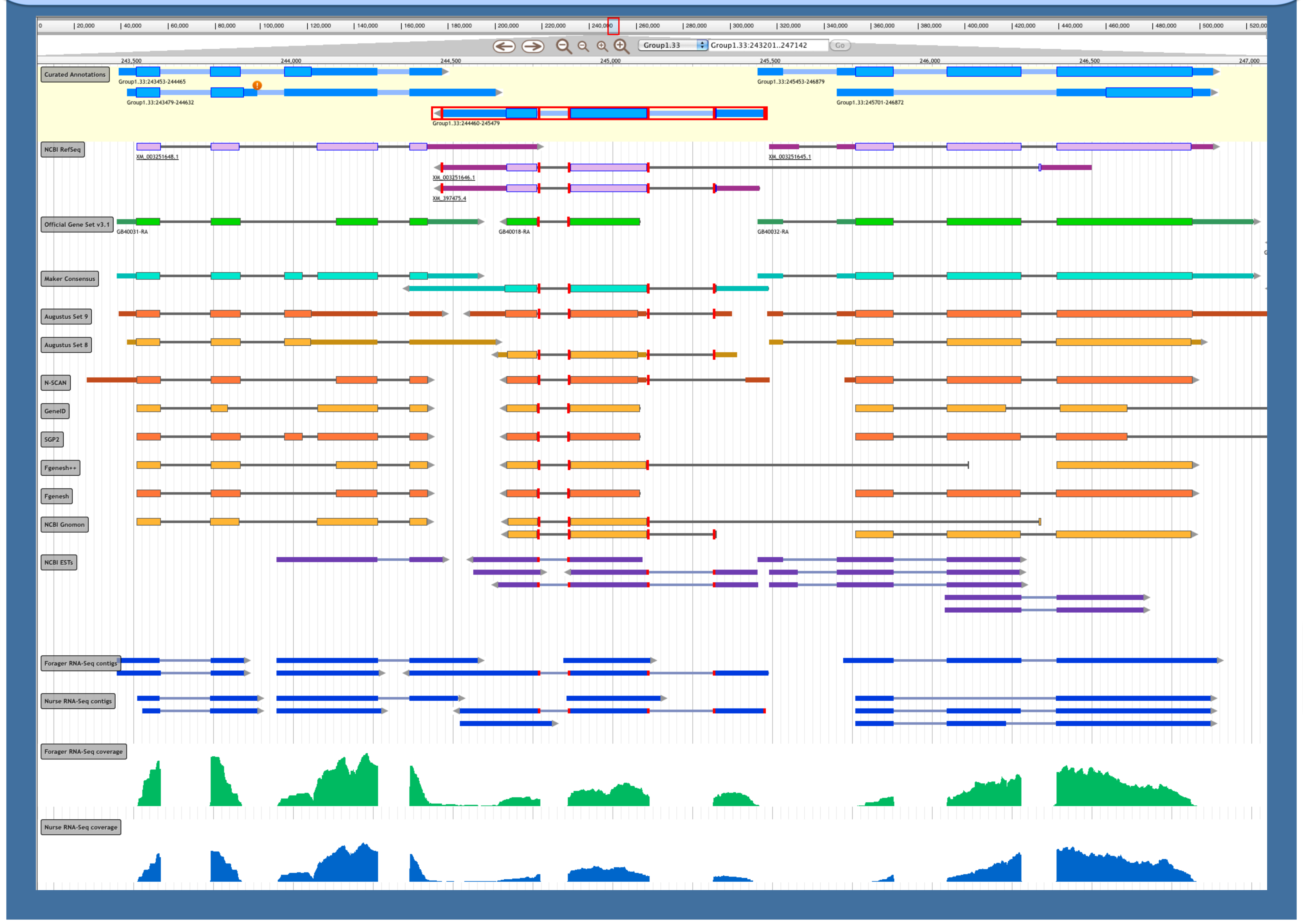


The above image shows the results of a series of sequence alteration editing operations in WebApollo. The top panel shows no sequence alterations, but the transcript annotation (in blue) is flagged with an orange exclamation icon which indicates that the curated intron-exon junction does not follow the canonical splice site pattern of having a "GT" immediately 3' of the junction. In the second panel a curator has examined this issue and determined that a base was mis-called in the assembly, and has therefore added a substitution annotation (shown in yellow), substituting a "T" for a "C". This change immediately triggers removal of the non-canonical warning icon, since with the substitution the splice junction now has the canonical "GT". In the third panel a curator has created a sequence insertion annotation (shown in green) upstream of the splice, and for the transcript annotation this leads to a stop codon which truncates the CDS. In the last panel a sequence deletion annotation has been created (shown in red), which causes a frame shift for the annotation transcript, resulting in reversal of the CDS truncation.

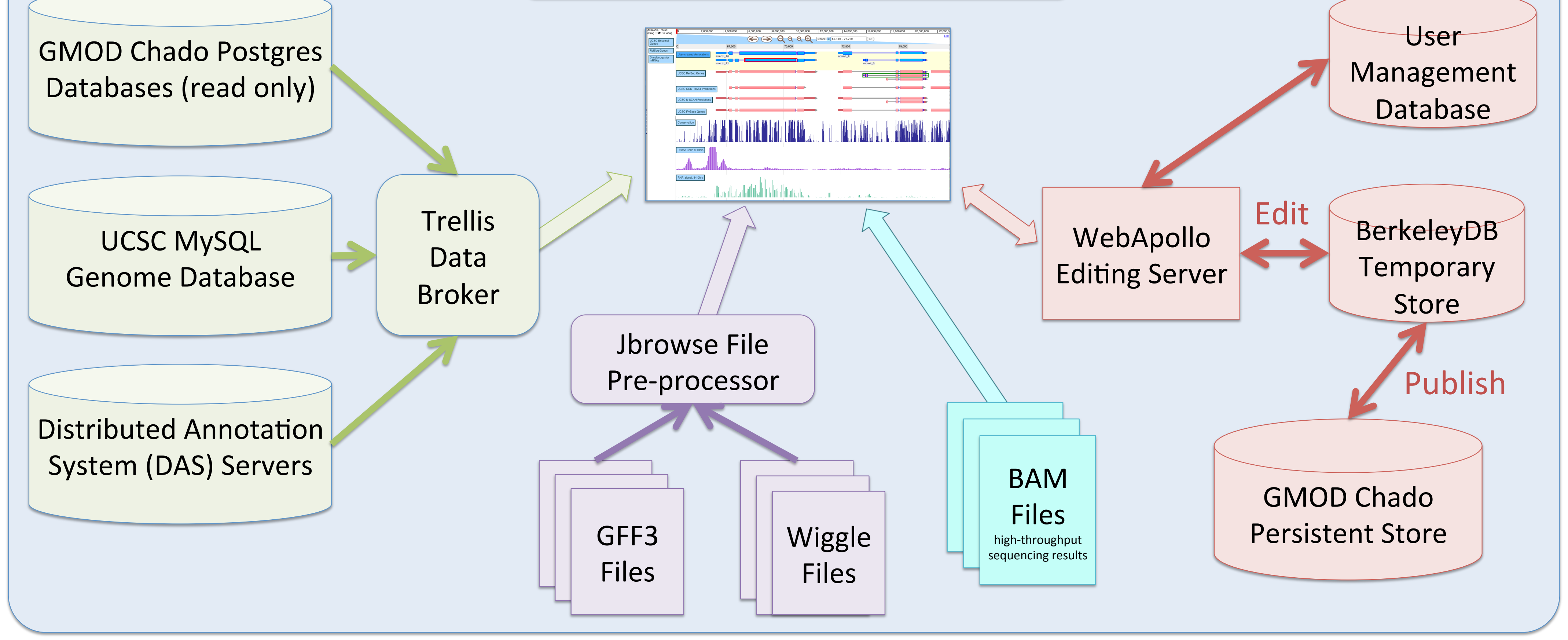
Manual Curation of Gene Structures: a crucial component of genome analysis

Shown below is WebApollo displaying tracks of genomic annotations along a small region of a scaffold from Honey Bee (Apis Mellifera) genome assembly 4.5. In the top track are in-progress gene models being created and edited in WebApollo. Results from various gene prediction programs are shown in yellow & orange. Results from MAKER, an analysis pipeline that builds consensus results from a number of the other computational analyses, is displayed in teal. The Official Gene Set, created using GLEAN, is shown in green. Transcripts from the NCBI RefSeq database are shown in pink. Aligned ESTs are shown in purple Results from high throughput sequencing (RNA-Seq) are displayed as assembled contigs (using exonerate) in dark blue, and as coverage graphs for two different experiments at the bottom.

This region illustrates the problem with relying solely on computational analyses for determining gene structures. Of the gene prediction results, no program is in complete agreement with any other in calling intron-exon boundaries across the region. MAKER, the Official Gene Set, and RefSeq also disagree. Results like these are common, and curators and tools to enable curation are needed to manually resolve these differences in order to create a more accurate gene set for the sequenced organism. WebApollo allows curators to build gene models via an intuitive drag-and-drop user interface. Curators can create an initial annotation based on any computational result, then add or delete exons, extend exons, and merge or split transcripts.



Data Sources


GMOD Chado Postgres Databases (read only)
UCSC MySQL Genome Database
Distributed Annotation System (DAS) Servers
Trellis Data Broker
Jbrowse File Pre-processor
GFF3 Files
Wiggle Files
BAM Files
high-throughput sequencing results
WebApollo Editing Server
User Management Database
BerkeleyDB Temporary Store
Edit
Publish
GMOD Chado Persistent Store

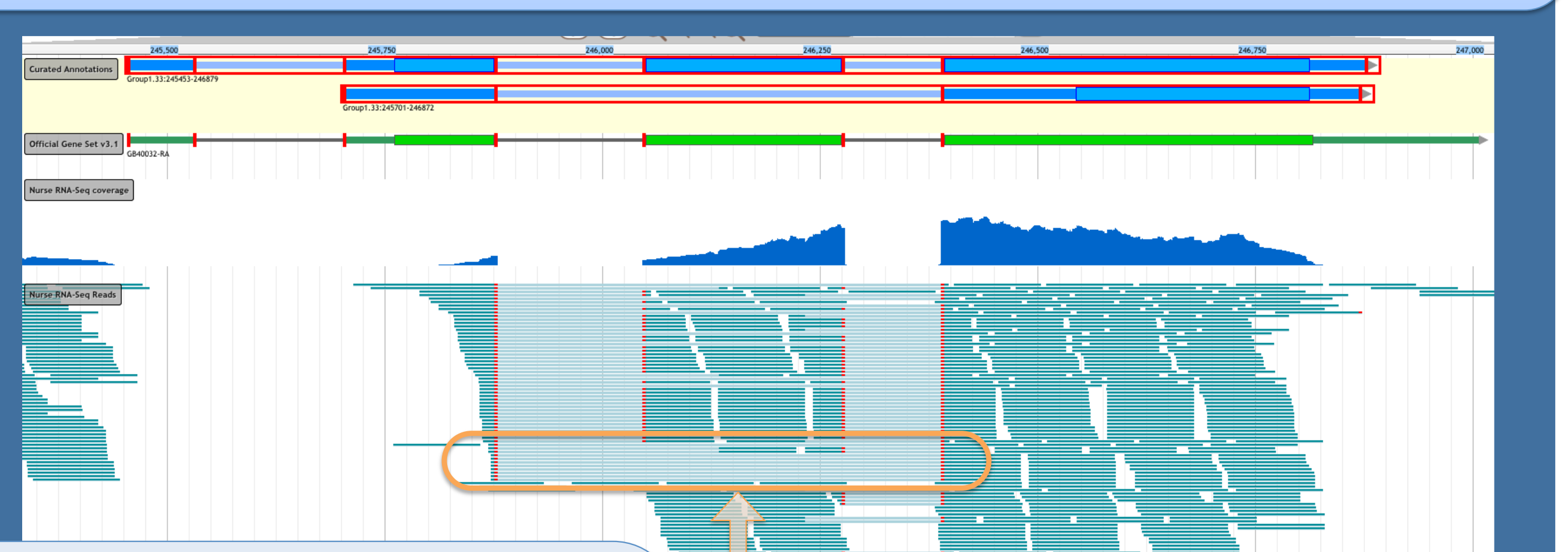
Other User Interface Highlights

- Configurable Undo/Redo History: transcript based, each transcript has a separate history for Undo/Redo
- Real-time updating: edits in one client are instantly pushed to all other clients
- Edge Matching: when annotation is selected, any other annotations with edges that match coordinates of selected annotation are highlighted.
- Tracking of non-canonical splice sites in curated annotations
- Ability to add comments, either chosen from a pre-defined set of comments or as freeform text.
- Can set start of translation for a transcript or let server determine automatically
- Convenient management of user login, authentication, and edit permissions
- Two-stage curation process: edit, then publish

Near-Future Plans

- Test with additional genomes before public release.
- Better documentation of both user interface and server components
- Add ability to search for a region by sequence residues
- Loading data directly from GFF3 files, both remotely and from user's local machine.
- Enable curation of additional annotation types (e.g. tRNAs)
- Enable adding DB-xrefs (e.g. for GO functional annotation)
- User track configuration to set annotation colors, height, etc.
- Hierarchical organization of tracks
- Public release in fourth quarter of 2012

Taking Advantage of Deep RNA Sequencing



Incorporation of high-throughput RNA sequencing data is rapidly becoming a requirement for thorough genome structure curation. WebApollo can display high-throughput RNA sequencing data in different ways, as shown here. The genomic coverage plot in dark blue was generated by server-side pre-processing of RNA-seq aligned reads, and shows for each base position in the genome how many aligned reads cover that position. WebApollo can also display individual aligned reads as shown in teal. Due to the large number of reads that can be generated by a single RNA-seq experiment, the aligned reads are typically stored in a binary file format (BAM) for efficient storage and retrieval. WebApollo is capable of using BAM file indexing to efficiently access only the slices of the BAM file needed for the current view. This allows WebApollo to directly load aligned reads from BAM files on any standard web server. The example shown highlights the importance of utilizing deep RNA sequencing for curation. A previously unknown alternatively spliced form of an annotated gene is indicated by a number of RNA seq reads that skip one particular exon. This example also highlights the usefulness of selection edge-matching (in red).

This work was supported by the National Institutes of Health grant numbers 5R01GM080203 from the National Institute of General Medical Sciences; and by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231