Manual Annotation of Features in DNA sequences

Overview

- What is manual annotation?
- Why do manual annotation?
- Tools for manual annotation
  - GBrowse
  - Apollo
  - Artemis
- Critical evaluation of evidence

What is manual annotation?

- Automated gene structure annotations are often imperfect and can benefit from manual refinement.

- A manual annotator uses different tools to view and curate the gene structure.

Why do manual annotation?

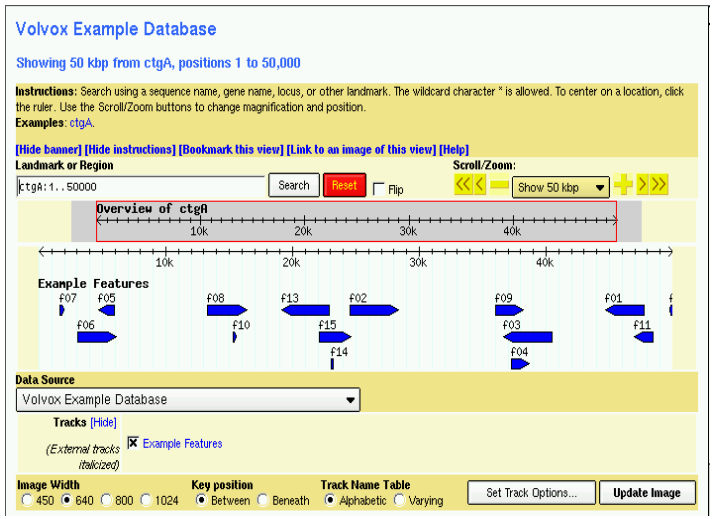
- Finding the genes is a primary goal of genome sequencing, and the gene set is the foundation for further studies. Accuracy is essential!
- Scientists can make informed decisions to correct faulty automated annotations.
- Model organisms have to be annotated to GOLD standards.

Manual annotation tools

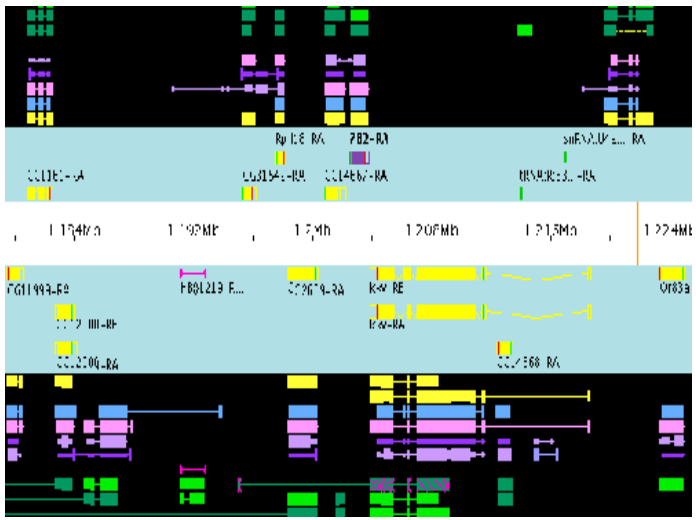
- Annotation tools allow you to visualize the location of features with respect to each other.
- They can also display evidence alignments and other information relevant to the process of annotation.
- Annotation editors allow the user to make changes in the annotation of a sequence and save these changes to a database.
- This talk will mainly focus on some open source annotation viewers and editors, however many commercial solutions are also available.

Open source annotation tools/viewers

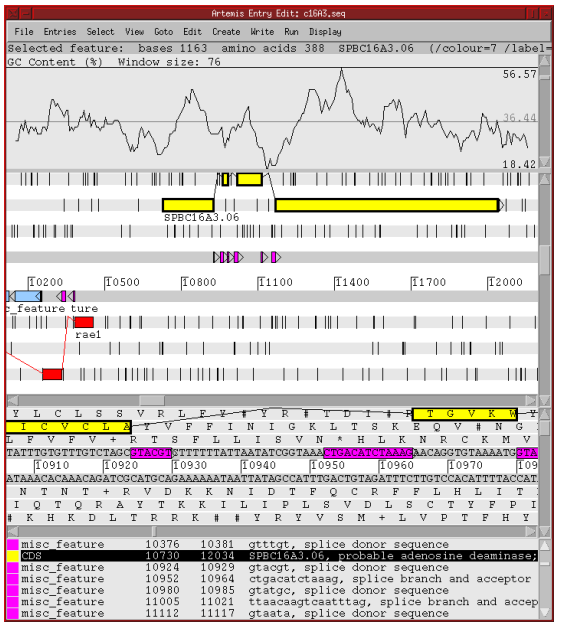
GBROWSE



APOLLO



ARTEMIS

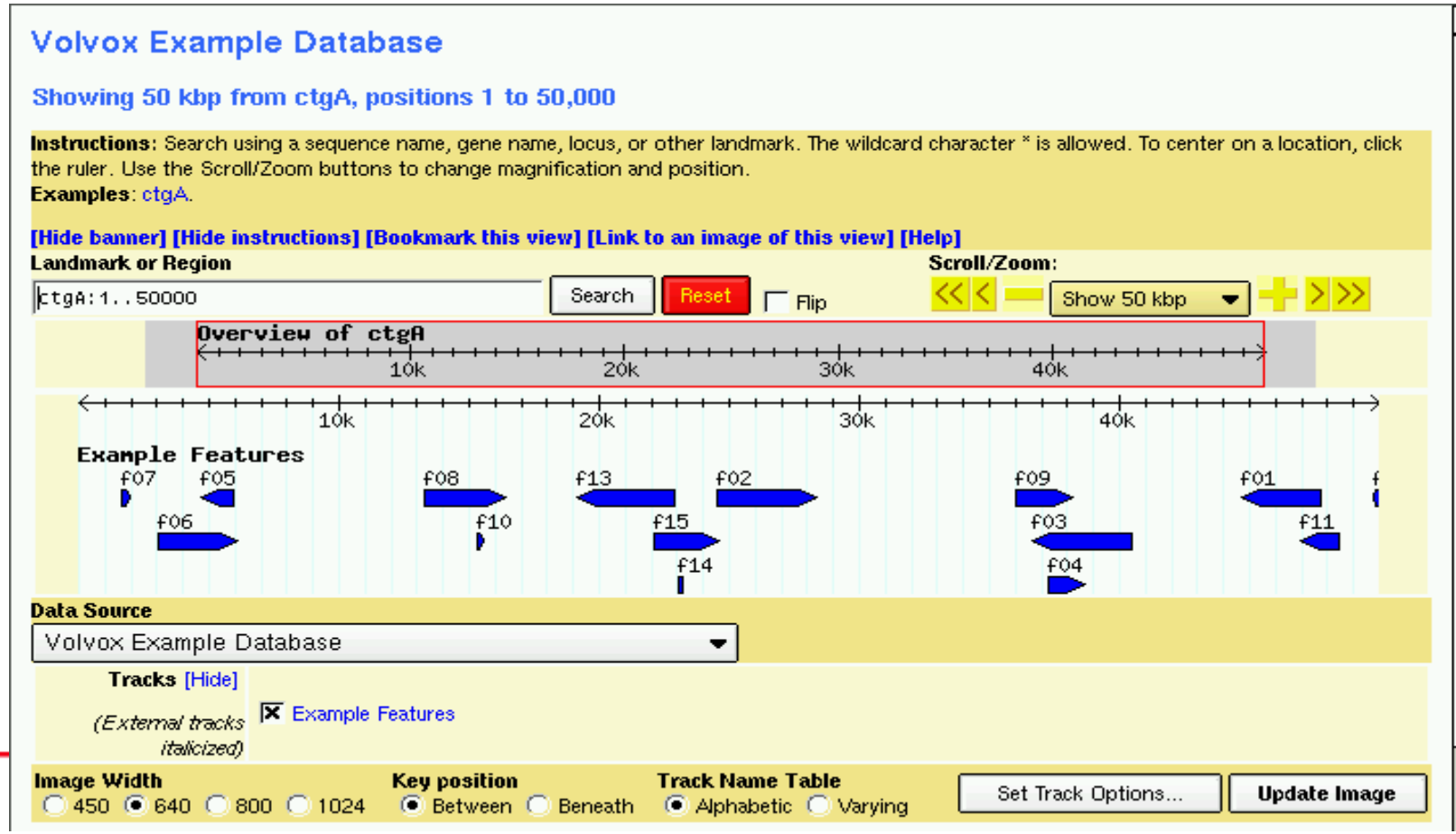


GBrowse

http://www.gmod.org/gbrowse

The Generic Genome Browser (also referred to as 'gbrowser') is a combination of database and interactive web page for manipulating and displaying annotations on genomes. It is similar to the genome browser used at UCSD and Ensemble

Hands on this afternoon



J. Craig Venter INSTITUTE

Important Features of GBrowse

It is mainly used to view and critically evaluate the various evidence

- Web based annotation Viewer
- Simultaneous bird's eye and detailed views of the genome.
- Scroll, zoom and center functions
- Attach arbitrary URLs to any annotation.
- Order and appearance of tracks are customizable by administrator and end-user.
- Search by annotation ID, name, or comment.
- Supports third party annotation using GFF format
- Settings persist across sessions.
- Allows DNA and GFF dumps.
- DNA sequence can be loaded into the database and displayed in a track along with the annotation.
- Annotation provided by the user can also be displayed in a track. The annotation data can be in a file on a remote server.

GBrowse needs:

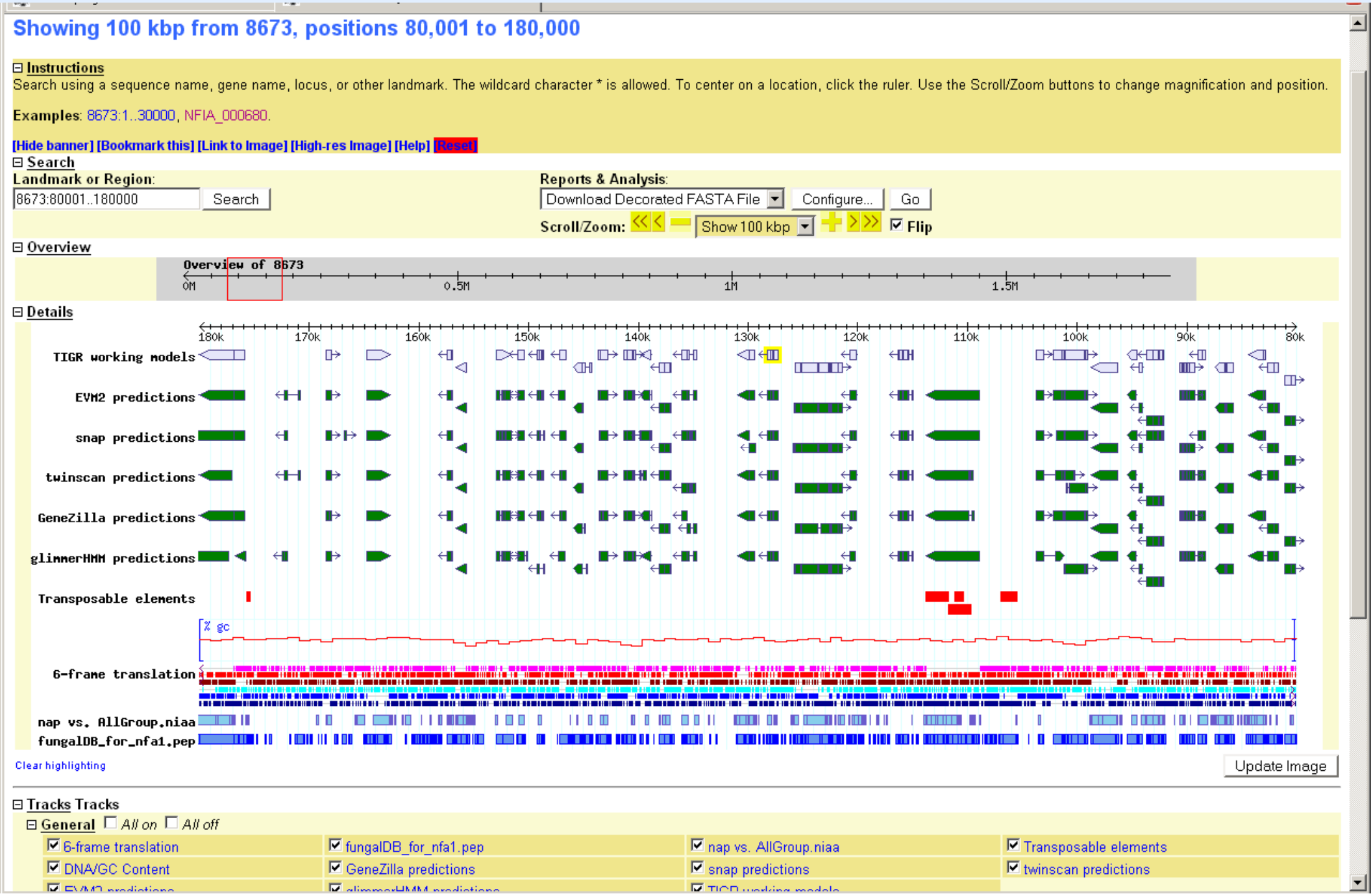
- Data files

  DNA and features files to load into the local database. e.g. gff files

- Conf files

  GBrowse configuration files for you to take and modify.

GBrowse : An example



Configure Tracks

Settings for TIGR Neosartorya fischeri browser

Undo Changes | Revert to Defaults | Refresh | Cancel Changes and Return... | Accept Changes and Return...

Track Options

The *Show* checkbox turns the track on and off. The *Compact* option forces the track to be condensed so that annotations will overlap. The *Expand* and *Hyperexpand* options turn on collision control using slower and faster layout algorithms. The *Expand & label* and *Hyperexpand & label* options force annotations to be labeled. If *Auto* is selected, the collision control and label options will be set automatically if space permits. To change the track order use the *Change Track Order* popup menu to assign an annotation to a track. To limit the number of annotations of this type shown, change the value of the *Limit* menu.

Track		Show	Format	Limit	Change Track Order
Track 1	TIGR working models	☑	Compact	No limit	
Track 2	EVM2 predictions	☑	Auto	No limit	
Track 3	snap predictions	☑	Expand	No limit	
Track 4	twinscan predictions	☑	Expand	No limit	
Track 5	GeneZilla predictions	☑	Auto	No limit	
Track 6	glimmerHMM predictions	☑	Auto	No limit	
Track 7	Transposable elements	☑	Auto	No limit	
Track 8	DNA/GC Content	☑	Auto	No limit	
Track 9	6-frame translation	☑	Auto	No limit	
Track 10	nap vs. AllGroup.niaa	☑	Expand & Label	No limit	
Track 11	fungalDB_for_nfa1.pep	☑	Auto	No limit	

Undo Changes | Revert to Defaults | Refresh | Cancel Changes and Return... | Accept Changes and Return...

External tracks italicized

*Overview track

**Region track

Detailed information about a gene

TIGR *Aedes aegypti* genome browser

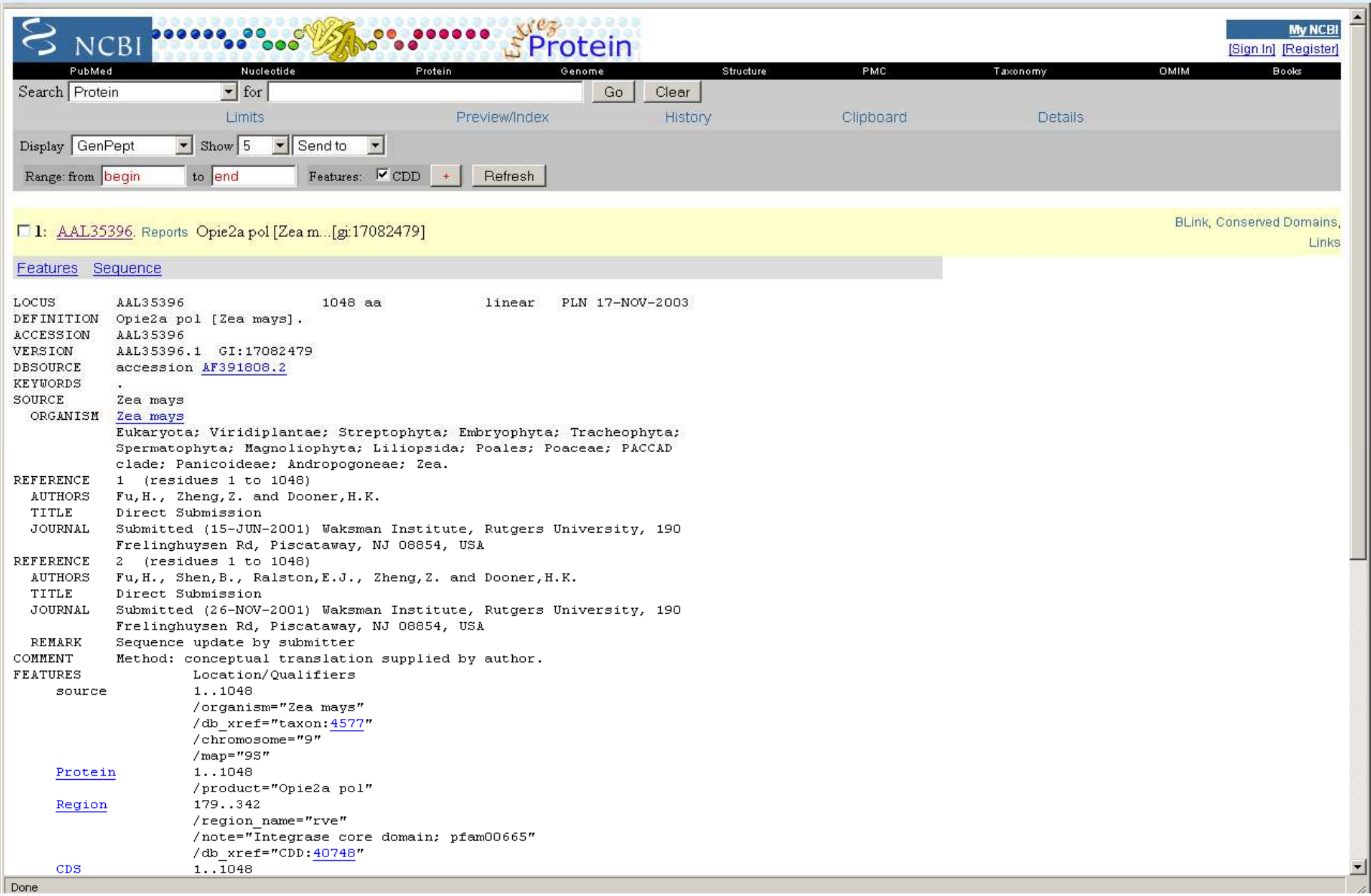
Name:	**24834.m01550**	
Class:	mRNA	
Type:	processed_transcript	
Source:	augustus	
Position:	supercont1.1:5258362..5274955 (+ strand)	
Length:	16594	
Parts:	**Type:**	CDS
	Source:	augustus
	Position:	supercont1.1:5258362..5258562 (+ strand)
	Length:	201
	Type:	CDS
	Source:	augustus
	Position:	supercont1.1:5258638..5258764 (+ strand)
	Length:	127
	Type:	CDS
	Source:	augustus
	Position:	supercont1.1:5258819..5259310 (+ strand)
	Length:	492
	Type:	CDS
	Source:	augustus
	Position:	supercont1.1:5259526..5259578 (+ strand)
	Length:	53
	Type:	CDS
	Source:	augustus
	Position:	supercont1.1:5259643..5259924 (+ strand)
	Length:	282
	Type:	CDS
	Source:	augustus
	Position:	supercont1.1:5274029..5274955 (+ strand)
	Length:	927

>24834.m01550 class=mRNA position=supercont1.1:5258362..5274955 (+ strand)

ATGGAGATTGGAGAAATGGCGATGTCCAATTTCGCATCGTGGTACCACGCGATGAATCCTGGCCATCTGACCACAGATGA
GCTGGAGCACGAGATGAAAATCCGTGGCATTCCGTTAGATGTTTCTCGTAGCGCTGCGGAACGAACGTTGCGTAGCCACC
TGAAAGAGGAAAAAGGCAAATCGGACATTAGAATGAAATTTGTGAGTGAATCGGTAGTTGCTGAGTTAGAGGTTTGTGAC
GATAAACTGAATGGGATAAAAAATCACCTATAAAAGCGAGTGTCTAAGAAAGCGCCAGATCAGTTGTACAAAACGAAATT
GGTTCATATCTTTTTCCGAATGCAGCGATTGCAAGCTCACACTGAGGAAGAAGACGAATTGAATAGCCTCGCGTTGATTG
CTGGTGAATGTGTAAAGCTTCTCAACGTGTATTATTCACCCACATCCCACCTTTTAGAAGTGCGAGAGGCCGAATTGGCG
ATTCTAAACAAGAGTTTGAATCAAATAAGGCAAGAGTTAGTACCGAATGATGGTGGACAAGGGGAGAATAATATGAATGA
AAGATCAAATGGGAGATCGACAGGTGAAAACGTGACTGAGCCGAATGGTAGGAATGTAGAAAATGATAGGATGAGTACCG
GAGGTGAGCTATCTGAGAAGACGACGGATAGTGAGAAAGAGGGATTAGCTTCTGAGATCCGTCAGCTAACGGCGGTGGTG
AGCCAACTACTGCAACGCATCACGGTACTTGAAGCTAATCAAAACGTTGCTCAACCGTTGAATCTTCATACAGGTACAGC
CAACAGTACGCAGATTGAGAGGAATATGGTGGTAGAGAAAGAGGTAGAAGAAGTTCAGGAAAAGTCAGTAGATTTCTTAG
GCTGGTTGAAACAGAGAAATGAGTCGTTAGATACGCTGAAAAATTCGGATGAAGACAAGCACGTCGGTAGTAAGCTGAAT
GAGAGGACGGAAAGGAGAGTTGAAAATAGCAGGCCATTGAGTATTGGAAATAAATTTCCGATTCATAAATGGTCTGTGAG
ATATGACGGAATGGATAATGGCAGGCGGTTGAATGAATTCTTGAAAGAGGTGGAATTCAACGCCCGGTCAGAAGGATTTT

Done

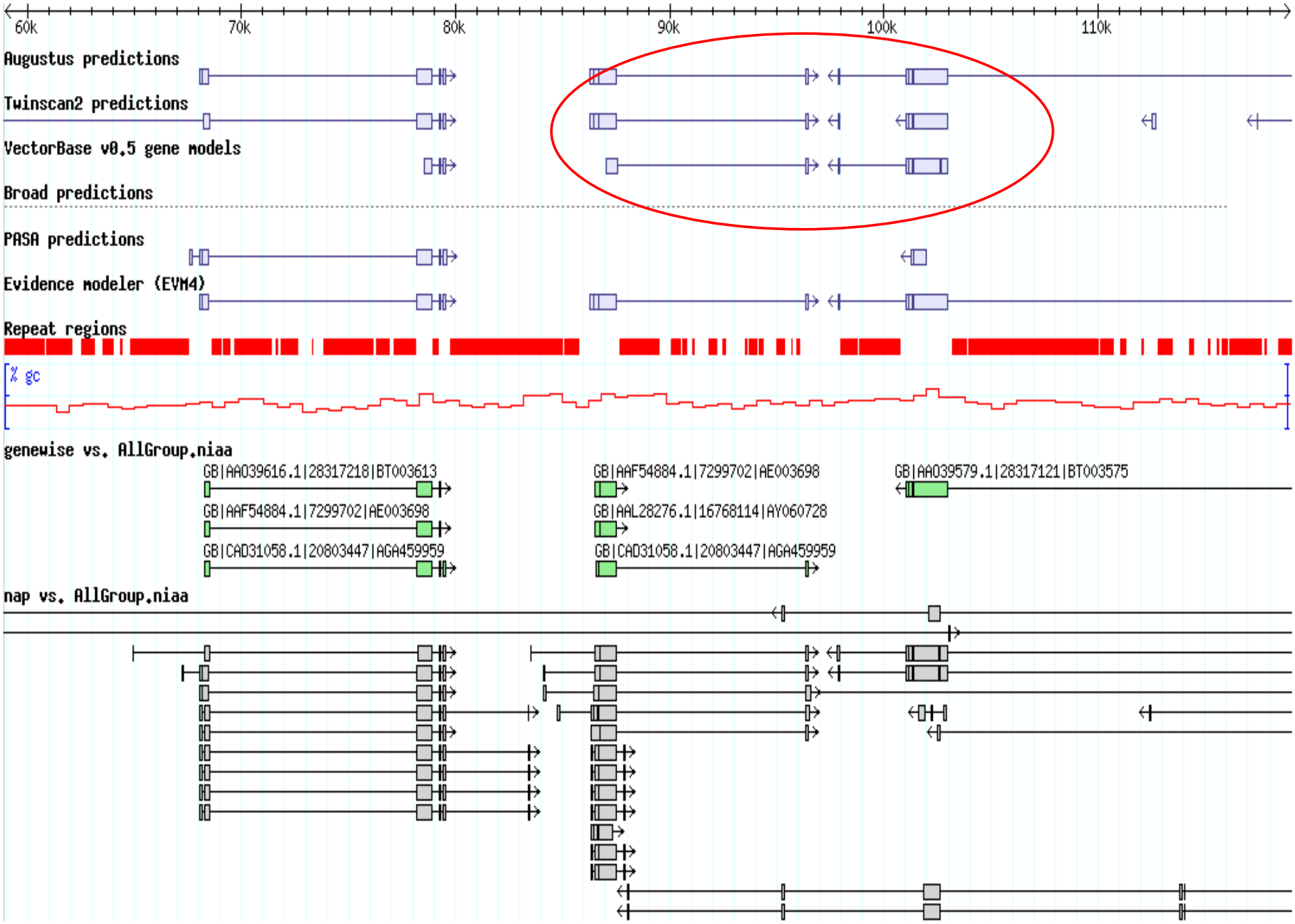
Clickable links



Load and compare annotations

Details

Visually compare Annotations By Different groups



Repeat region information

TIGR *Aedes aegypti* genome browser

Name:	**repeat_15051**
Class:	Match
Type:	match
Source:	repeat_regions
Position:	supercont1.1:5209777..5210002 (+ strand)
Length:	226
Note:	trf

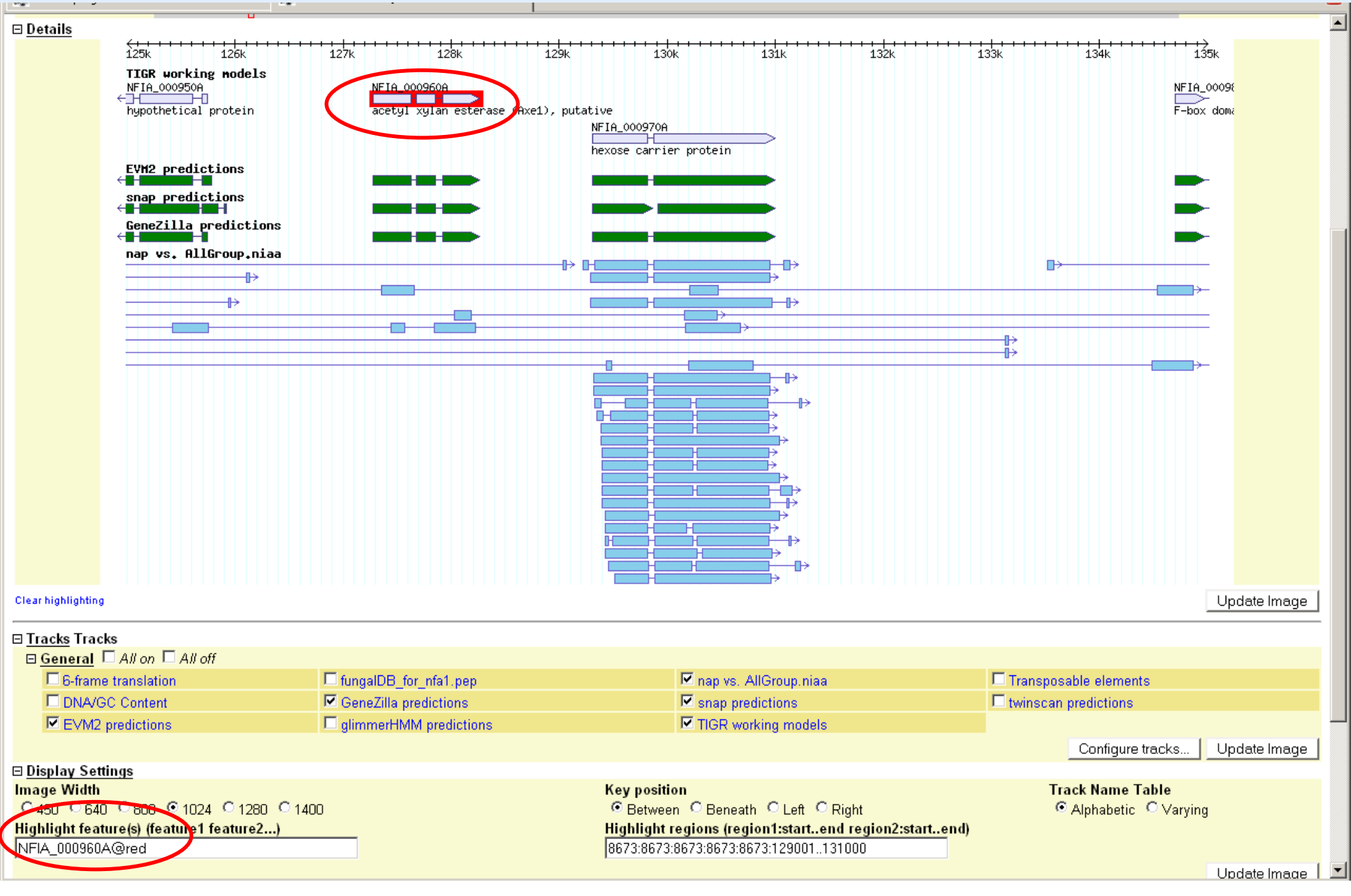
>repeat_15051 class=Match position=supercont1.1:5209777..5210002 (+ strand)
TATCGATTTGAGAATCGATTGAGAAATATCGATTTGAGAATCAAGGGAGTAATTTTAGTTTGAGGATCGATGGAATGGTA
TCGATTTGAGAATCTATTGAGAAATATTGATTTGAGGATTAATTGAGTAATATCGATTTGAGAATCAATTGAATAGTATC
GATTTGCGAATCGATGGAATGGTATCGATTTGAGAATCGATGGAGTAATATCGATTTGAGGATCGA

Name:	**repeat_15051**
Class:	Match
Type:	match
Source:	transposonPSI
Position:	supercont1.29:1540354..1540551 (+ strand)
Length:	198
Note:	LINE

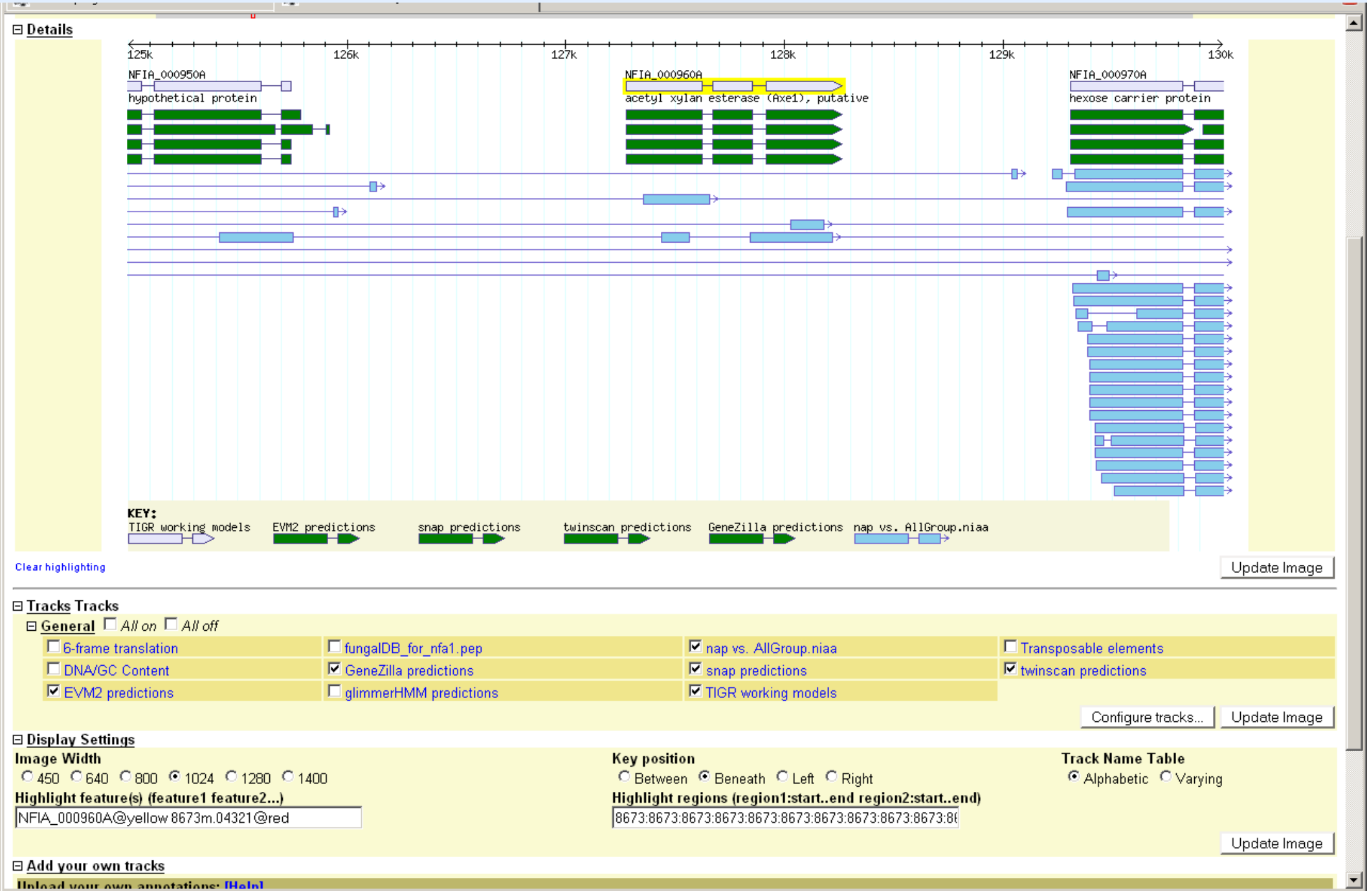
>repeat_15051 class=Match position=supercont1.29:1540354..1540551 (+ strand)
GACATAGGCCAACATGTCAAAACATCTAAAAACATGAAGGCCCCAGGTTTCGACAGCATCCTGAACATGGAACTCAAGCA
ATTAAGTCTCCAGTTCTATCTACATAGCTACTTCCGACTTAGCTACTTCCCCTCGTCGTGGAAGTCAGCAAAAGTCATCC
CCATCACAAAACCTGGGAAGGATCCACCAAAATCTGTT

Highlight Regions

Highlight Features



Display options



Links to Manatee

Rama and Linda will talk
about Manatee tomorrow
And Mathangi will demo manatee

Neosartorya fischeri **NFIA_000970**

▸ Download sequence ▸ Show genomic region on 1099437636265

Gene Identification

Gene Product Name: **hexose carrier protein**
Locus Name: NFIA_000970
Comment:

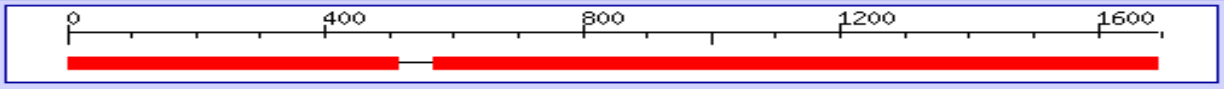
Gene Ontology Classification |

▸ **None found**

Attributes

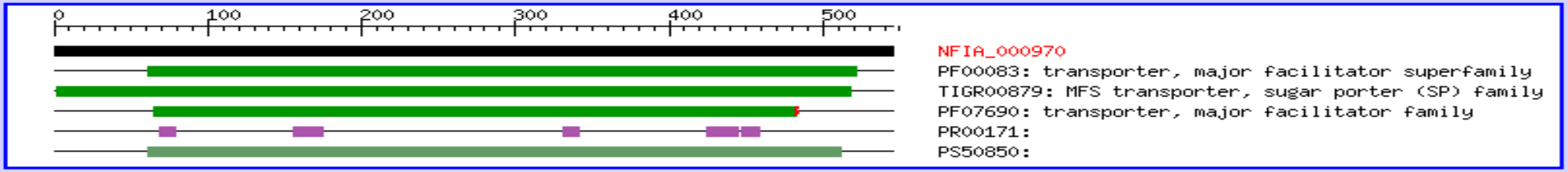
Chromosome:	Unknown
Coordinates (5' - 3'):	129307 - 130996 on assembly 1099437636265
Nucleotide length:	1635
Predicted protein length:	545
Predicted molecular weight:	0.00
Predicted pI:	0.00

Gene Structure



Intron/Exon structure in nucleotide coordinates relative to the start of the gene. Click on the picture to display evidence supporting this gene structure.

Protein Features

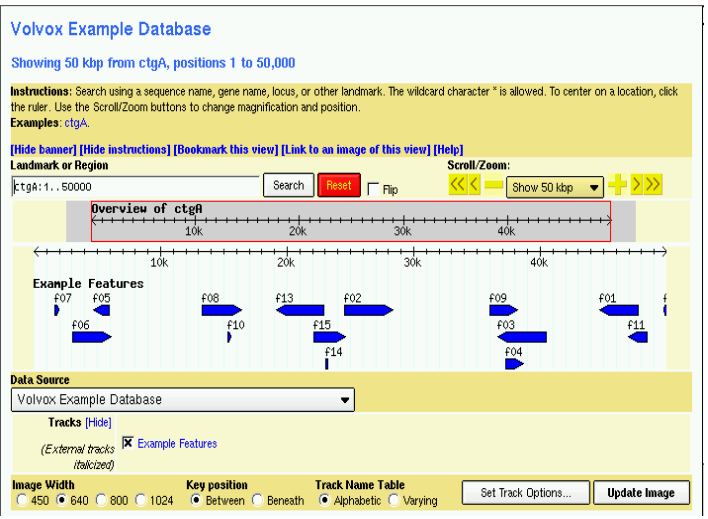


HMM/PFAM Hits

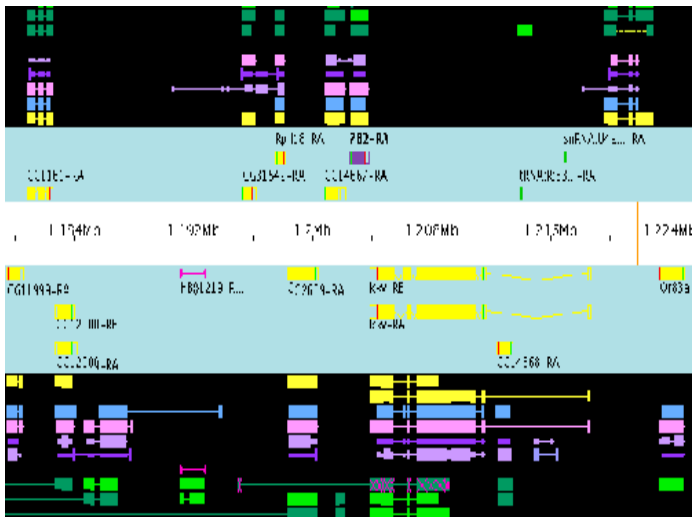
Accession	Name	Total Score	Trusted cutoff	Noise cutoff	Total expect	Curated
PF00083	transporter, major facilitator superfamily	386.0	-85.00	-85.10	6.7e-113	
TIGR00879	MFS transporter, sugar porter (SP) family	347.1	150.00	100.00	3.5e-101	

Open source annotation tools/viewers

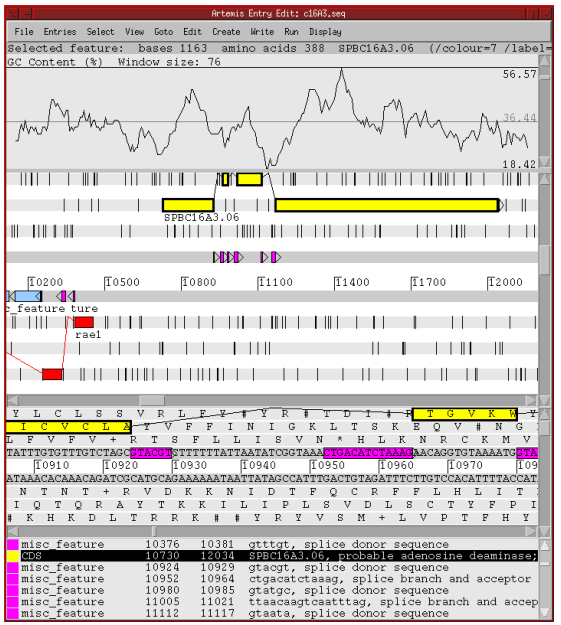
GBROWSE



APOLLO

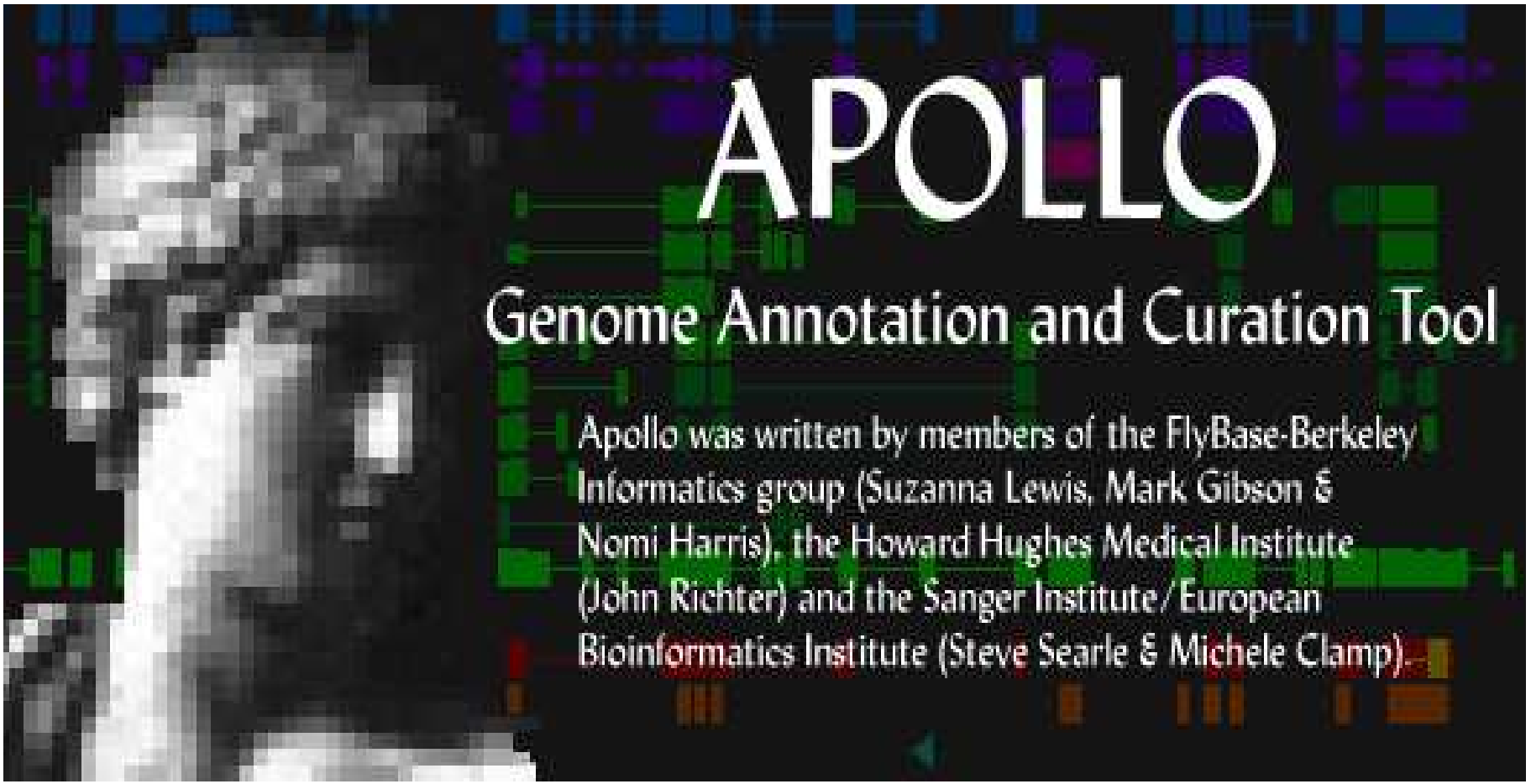


ARTEMIS



APOLLO

http://www.fruitfly.org/annot/apollo/



About Apollo

- Apollo is a genome annotation **viewer** and **editor**
- It was developed as a collaboration between the Berkeley Drosophila Genome Project (part of the FlyBase consortium) and The Sanger Institute in Cambridge, UK.
- Apollo allows researchers to explore genomic annotations at many levels of detail, and to perform expert annotation curation, all in a graphical environment.
- The Generic Model Organism Database (GMOD) project, which aims to provide a complete ready-to-use toolkit for analyzing whole genomes, has adopted Apollo as its **annotation workbench**.
- Apollo is a **Java** application that can be downloaded and run on Windows, Mac OS X, or any Unix-type system (including Linux).

Important features of Apollo

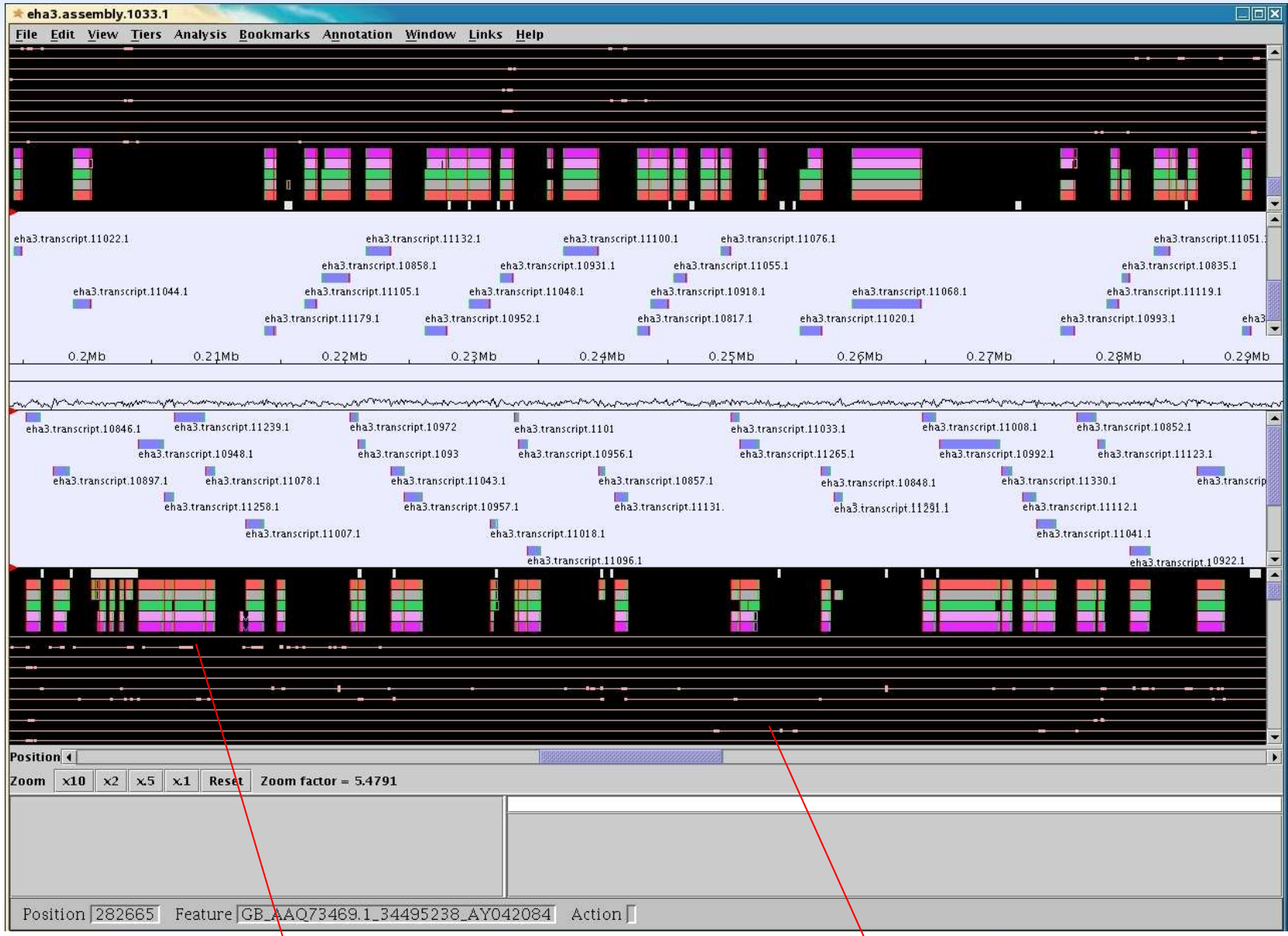
- Apollo provides 'semantic zooming'
- You can drag and drop features from evidence panel.
- Simple interface for editing gene models (merging, splitting, deleting, adding exons/transcripts, set 5', 3' or calculate longest open reading frame)
- By default it can display both strands at the same time.
- Sequence-level adjustments are possible with Apollo's exon editor.
- Apollo allows you to tag results with a comment
- Apollo automatically creates alternatively spliced transcripts for a gene whenever the Open Reading Frames (ORFs) of transcripts overlap.

Reading Data

Apollo can transparently access data across the network from remote machines, as well as reading files that reside locally.

- Apollo can read and load data directly from a chado database
- Apollo can also read Chado XML or GAME XML files
- It can read Ensembl GFF files
- It can read Ensembl schema databases
- It can read GenBank or EMBL format

Apollo – Main window



Evidence

Forward Strand

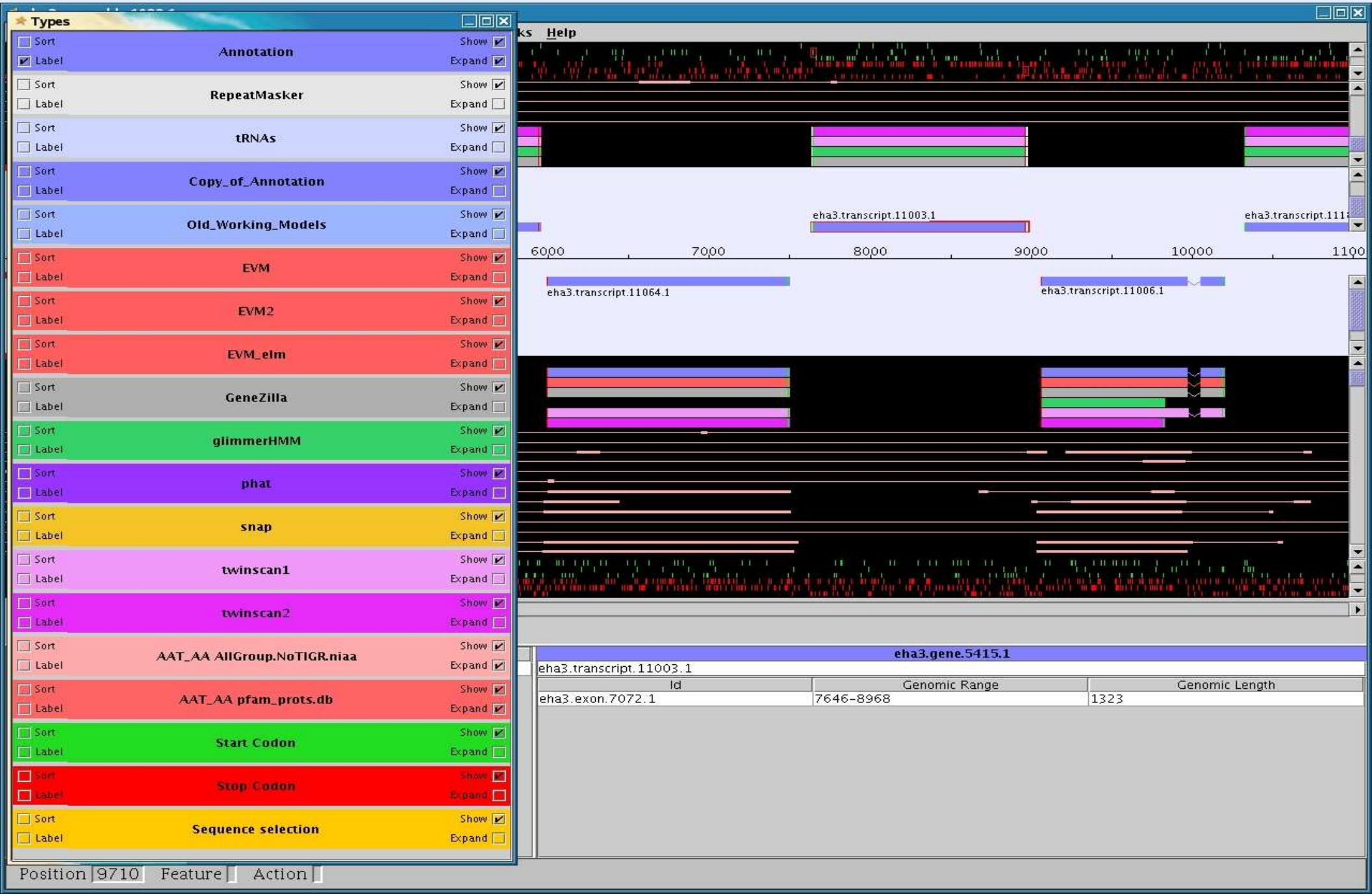
Reverse Strand

Evidence

Gene Finders

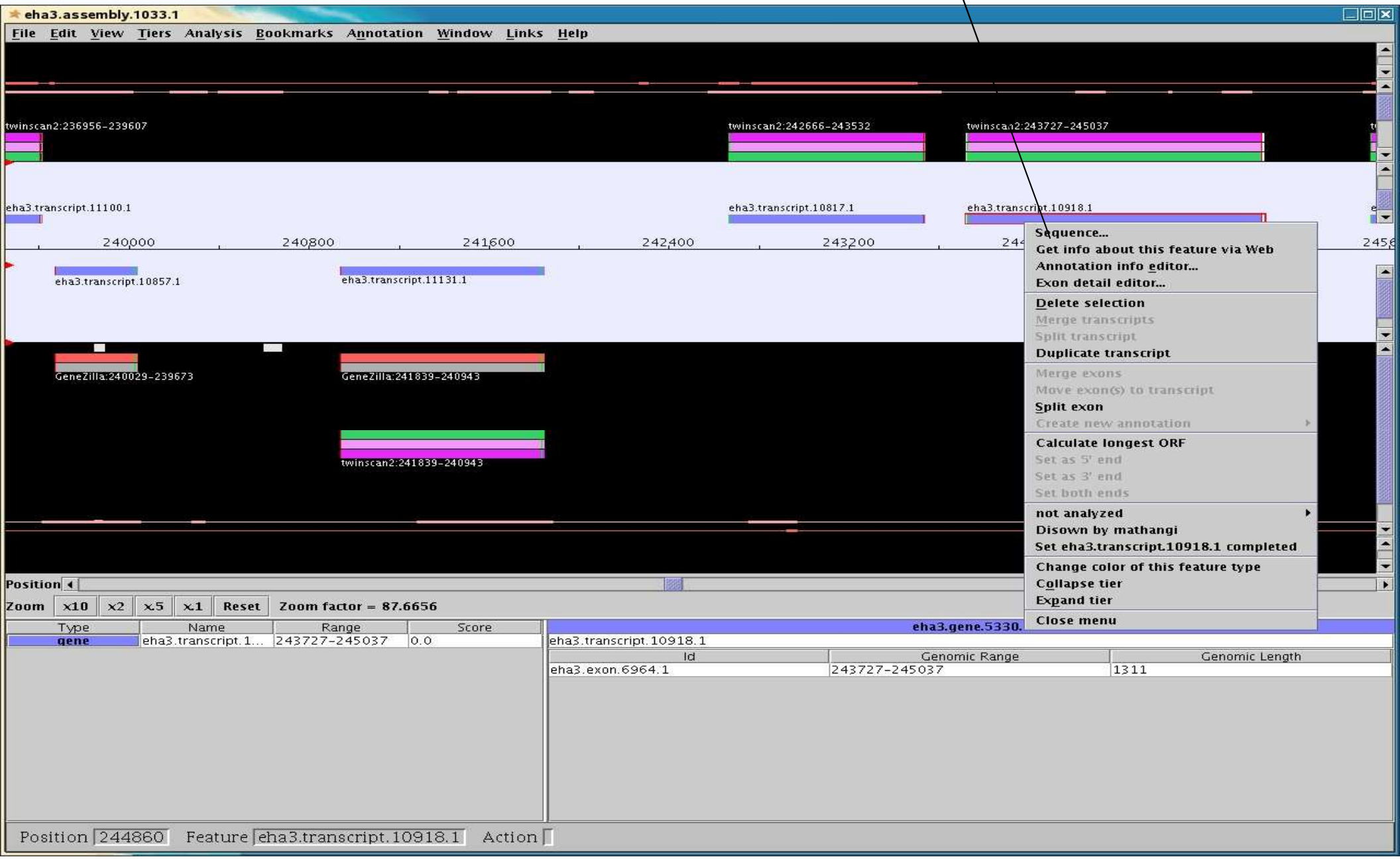
Database Searches

Types Panel

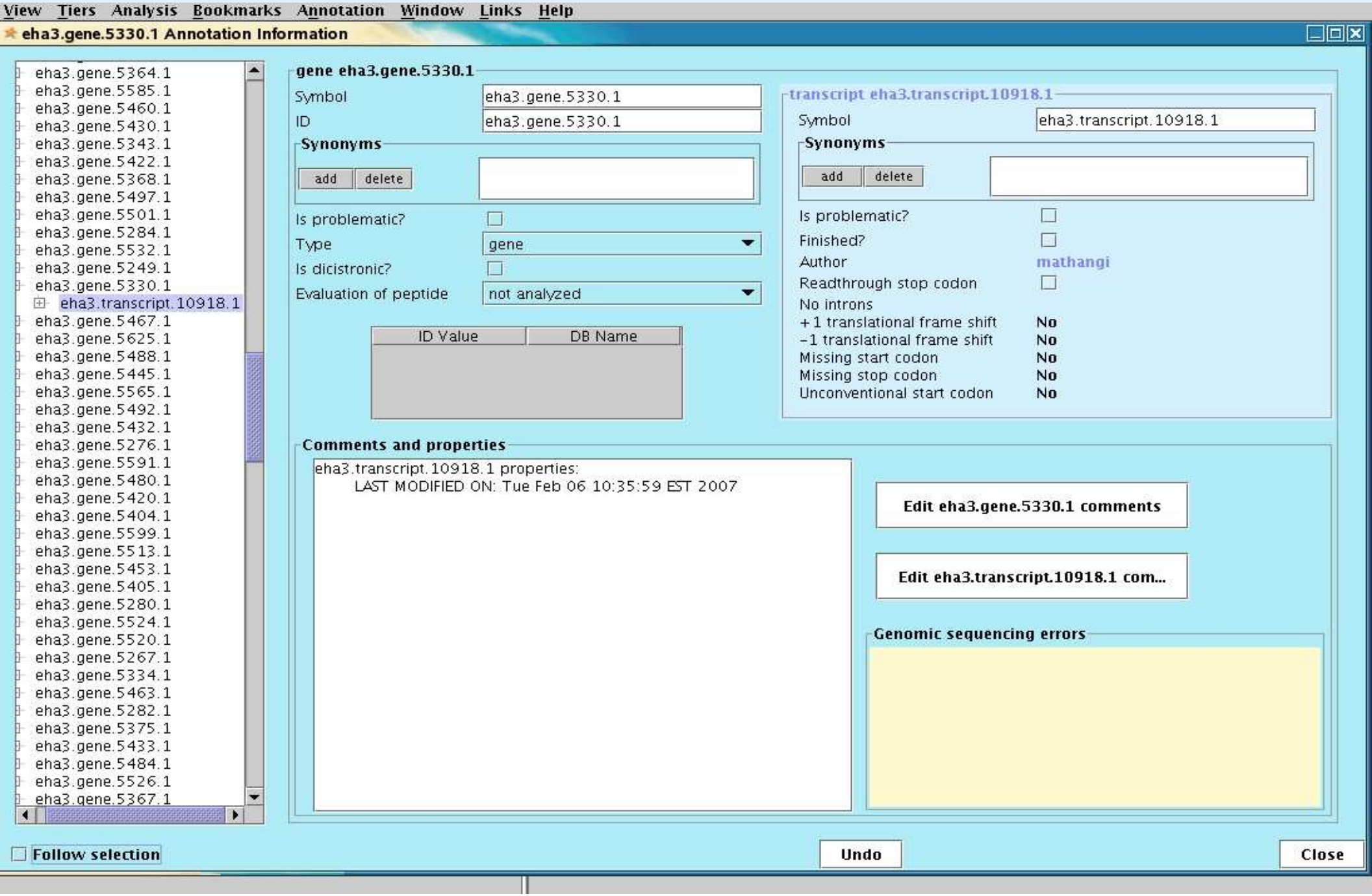

Types
Sort
Label
Annotation
Show
Expand
RepeatMasker
tRNAs
Copy_of_Annotation
Old_Working_Models
EVM
EVM2
EVM_elm
GeneZilla
glimmerHMM
phat
snap
twinscan1
twinscan2
AAT_AA AllGroup.NoTIGR.niaa
AAT_AA pfam_prots.db
Start Codon
Stop Codon
Sequence selection
Help
eha3.transcript.11003.1
6000
7000
8000
9000
10000
eha3.transcript.11064.1
eha3.transcript.11006.1
eha3.gene.5415.1
eha3.transcript.11003.1
Id
Genomic Range
Genomic Length
eha3.exon.7072.1
7646-8968
1323
Position
9710
Feature
Action

Finding more about annotation and features

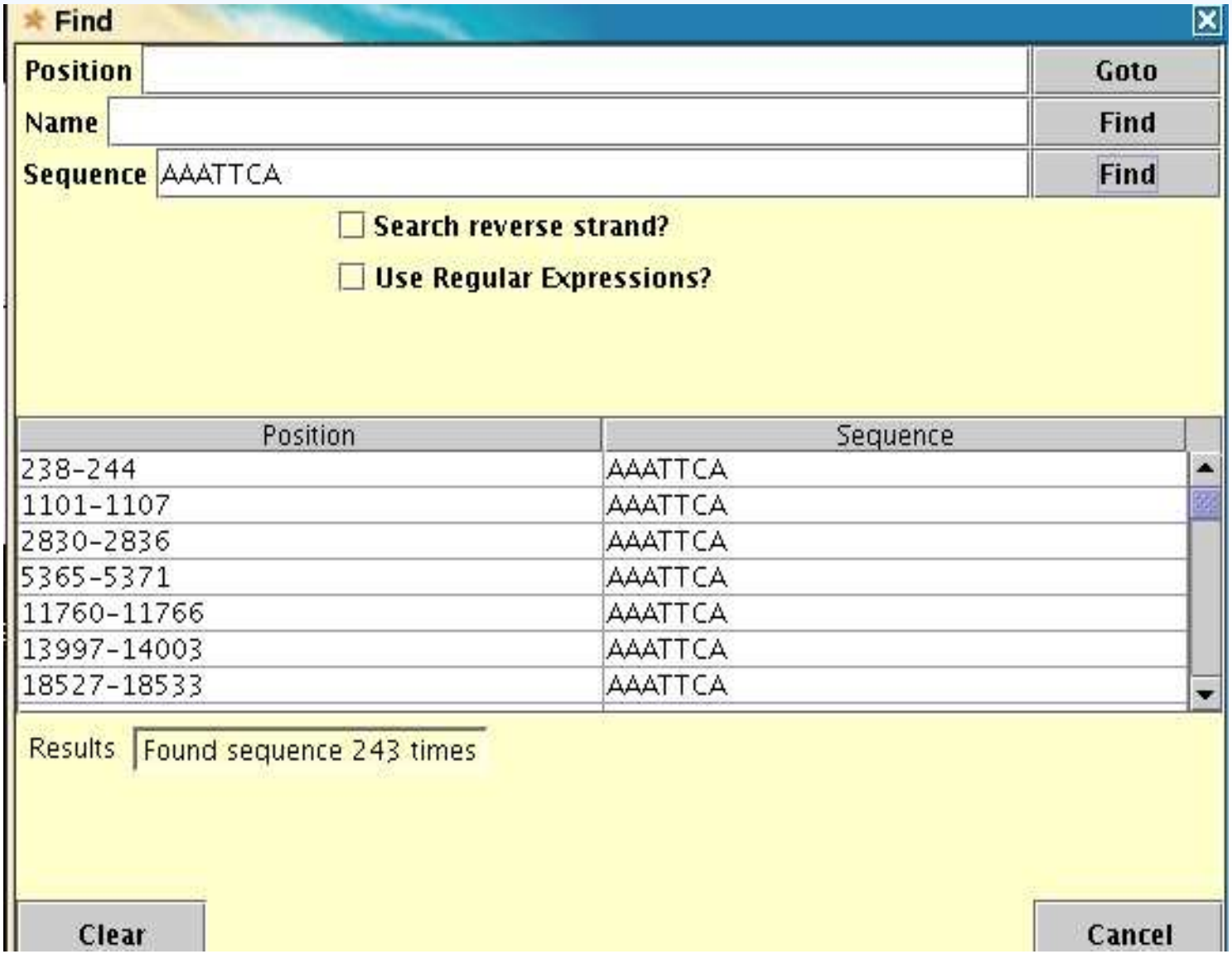
Right click on a gene



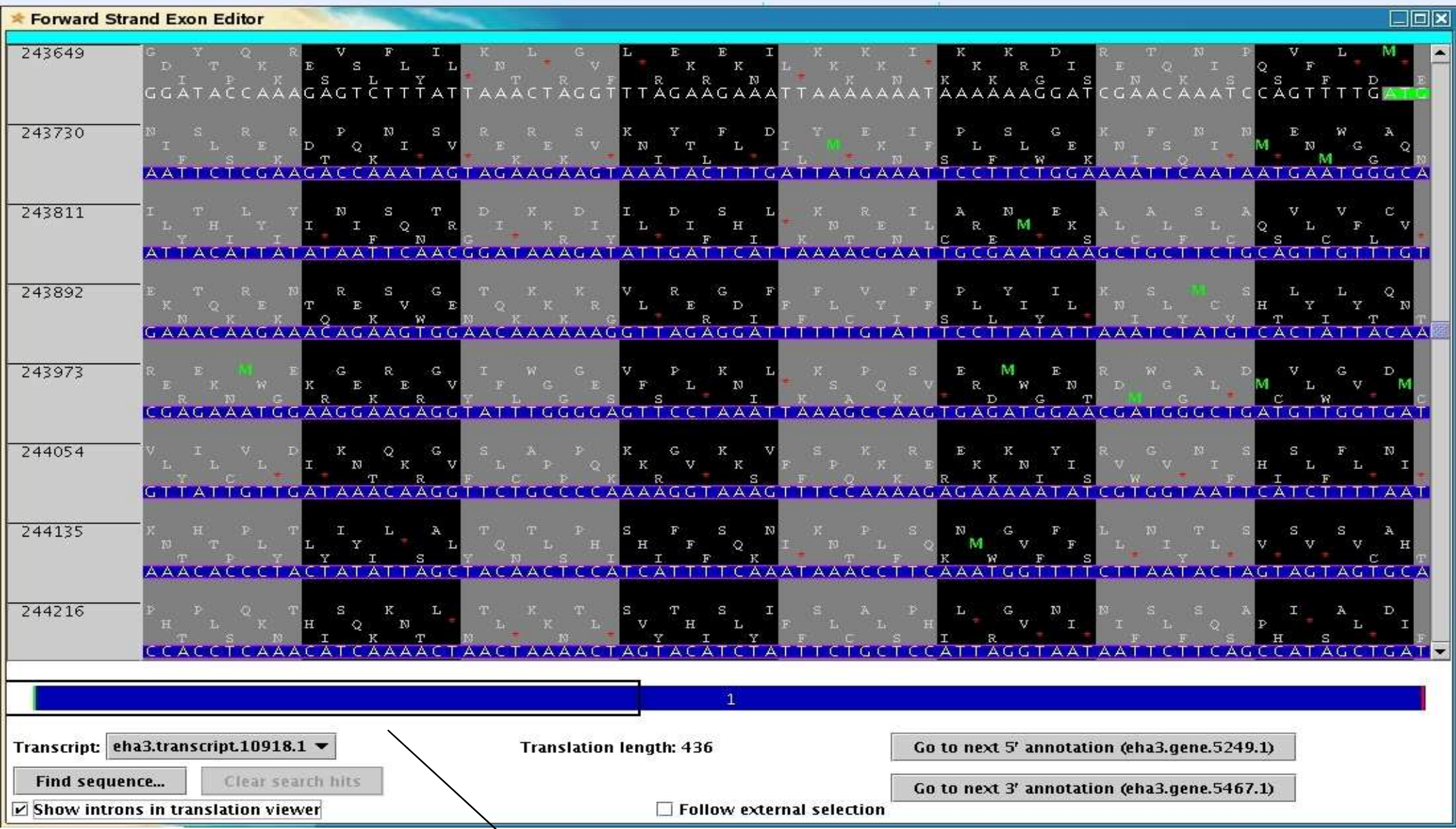
Annotation info editor



FIND Function



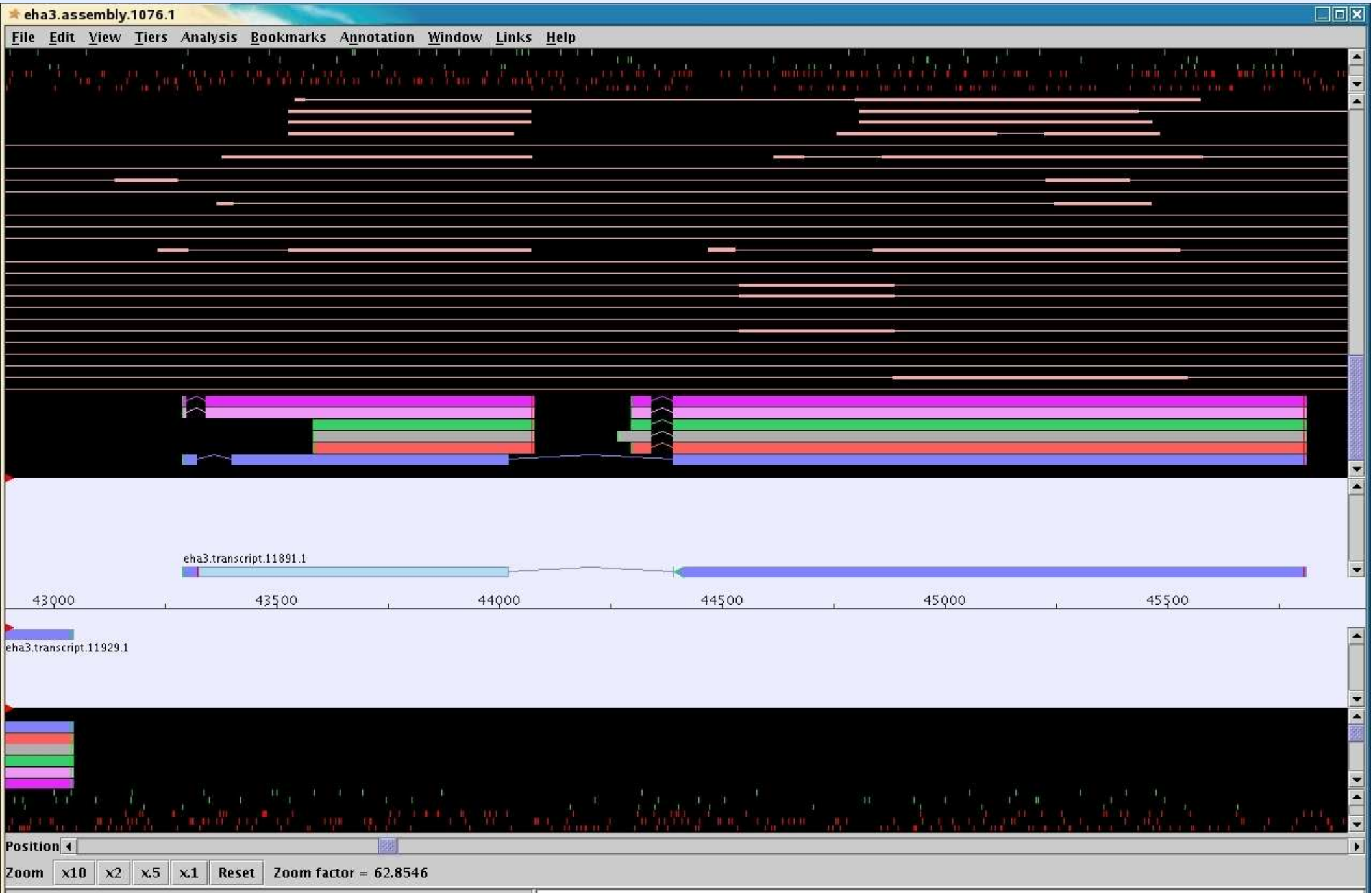
Exon detail editor



J. Craig Venter INSTITUTE

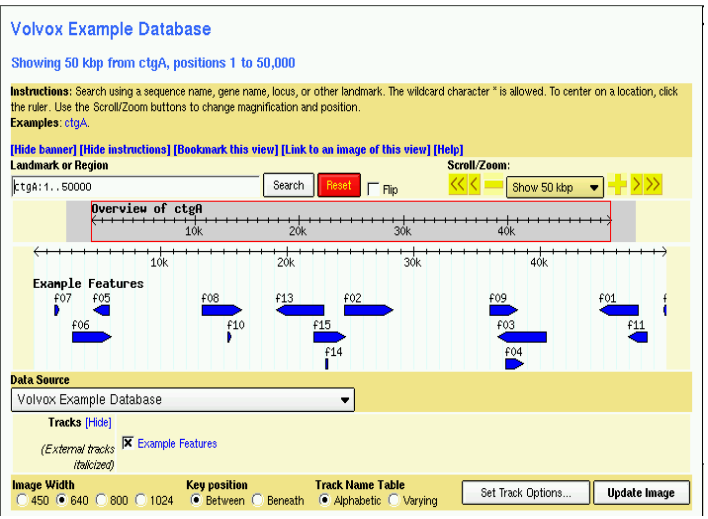
Sliding window

An example

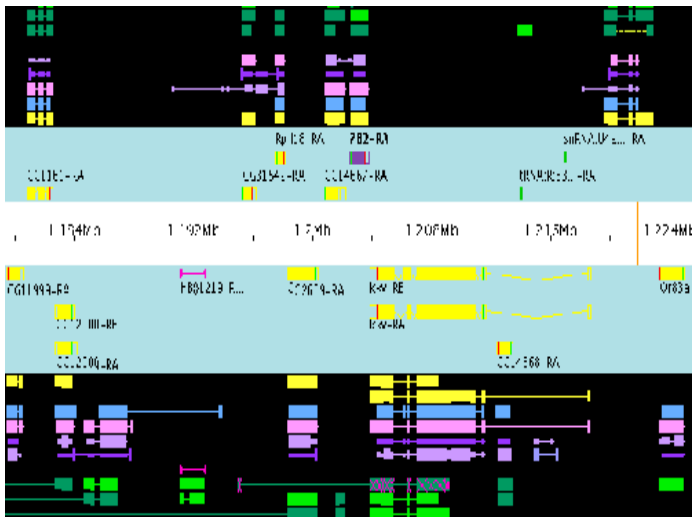

eha3.assembly.1076.1
File Edit View Tiers Analysis Bookmarks Annotation Window Links Help
eha3.transcript.11891.1
43000
43500
44000
44500
45000
45500
eha3.transcript.11929.1
Position
Zoom x10 x2 x.5 x.1 Reset Zoom factor = 62.8546

Open source annotation tools/viewers

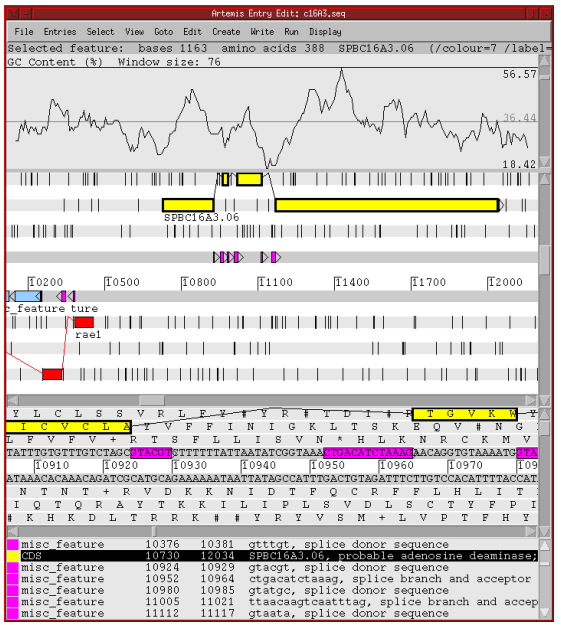
GBROWSE



APOLLO



ARTEMIS



Artemis

http://www.sanger.ac.uk/Software/Artemis/

- Developed by Sanger Institute, UK
- Reads in Genbank, EMBL and GFF3 files and sequence in FASTA format.
- Displays annotation derived from a feature table.
- Has a selection of tools such as a G/C graph.
- Allows the direct display of blast results.
- Artemis is Java based (cross platform), easy to run and install.
- Allows editing the annotation and export it in GenBank, EMBL, and GFF3 format

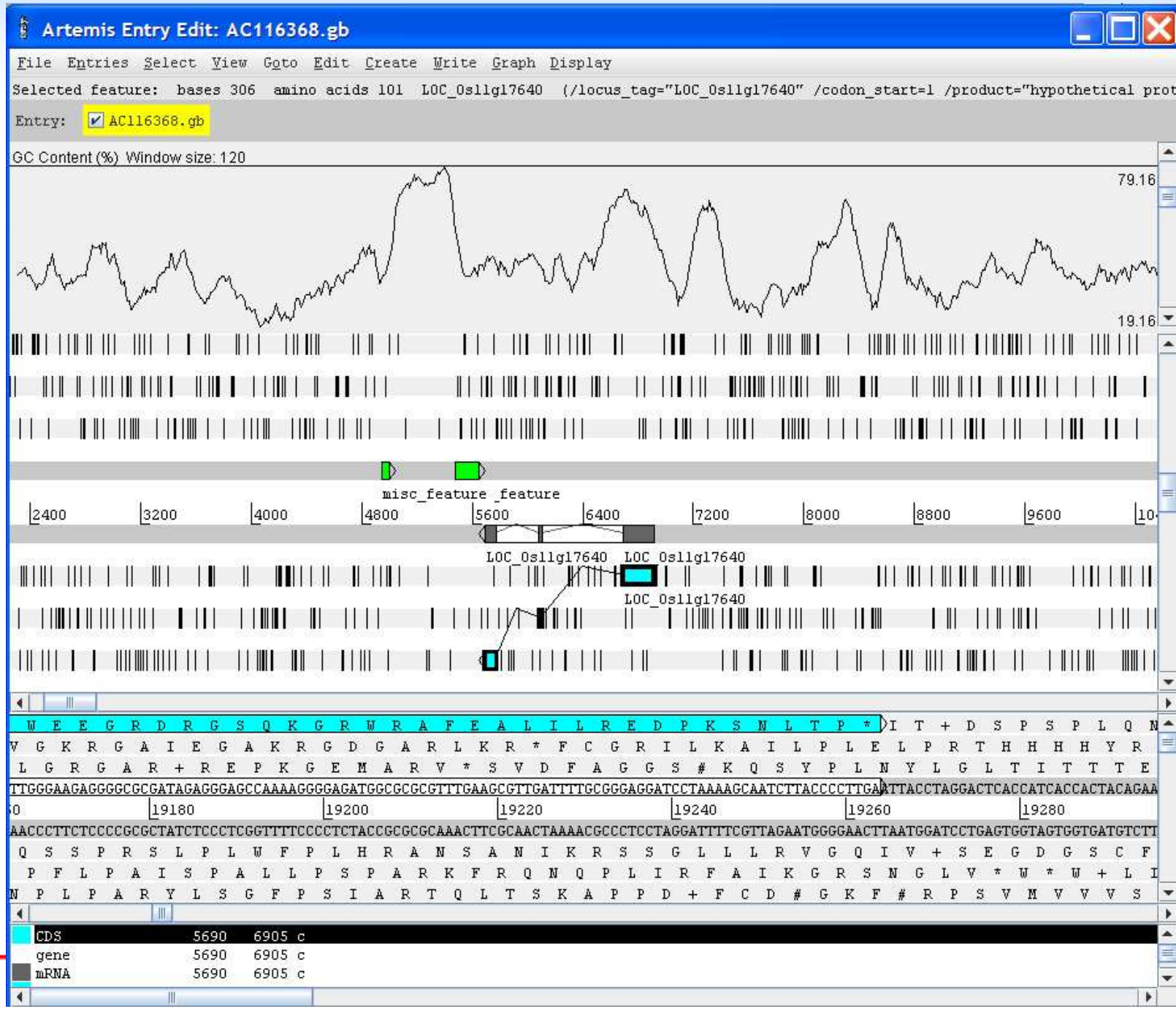
Sequence Viewing

GC Plot

Sequence And Feature Overview

DNA View And 6 Frame Translation

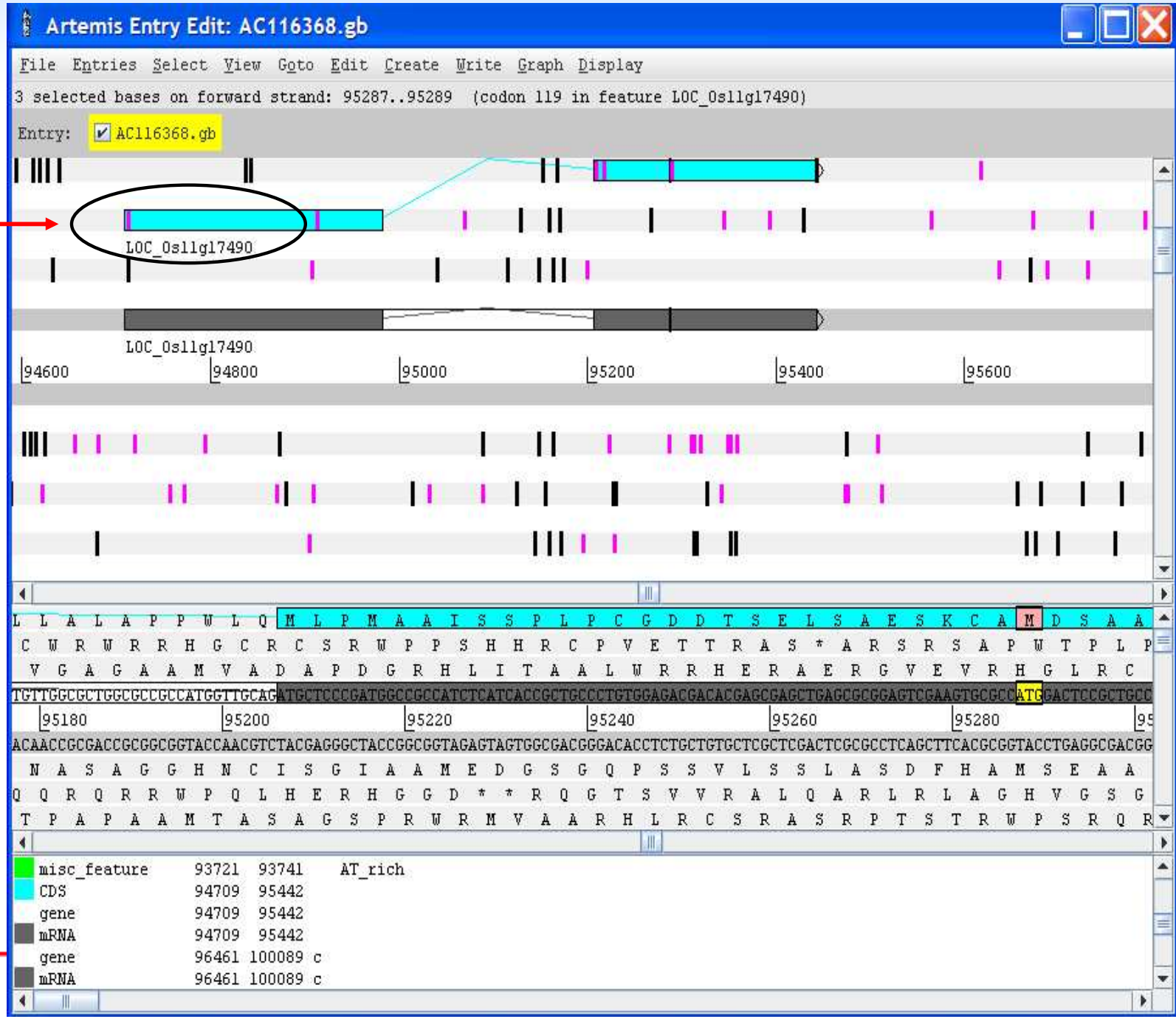
Feature Table



J. Craig Venter INSTITUTE

Annotating with Artemis

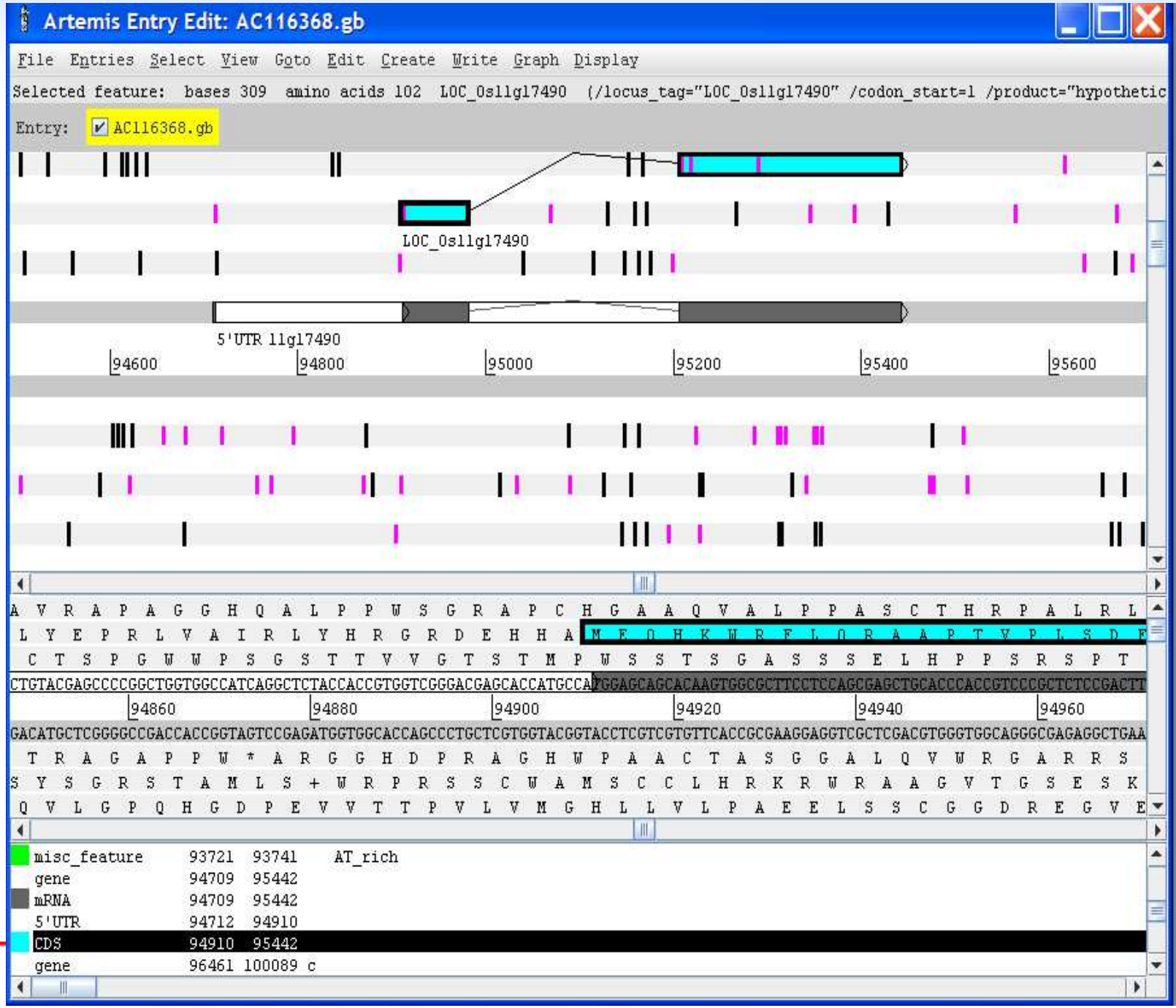
Evidence shows that this region should be annotated as a 5'-UTR (hypothetically)



J. Craig Venter INSTITUTE

Annotating with Artemis

5' UTR inserted in Gene model



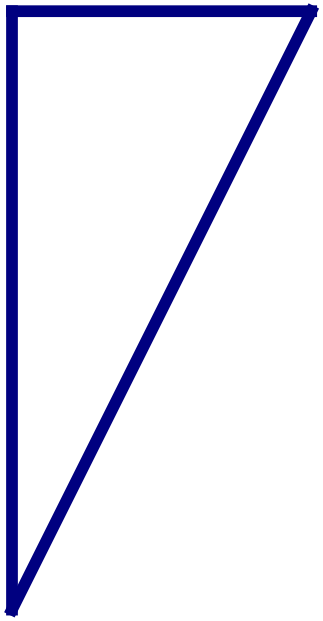
J. Craig Venter INSTITUTE

Comparing the viewers

	Artemis	Apollo	Genome Browser
Data Source	Flat file	Flat file or database	Flat file or database
Sharing	Not designed for sharing	Users can share databases	Convenient via the web, possible to share databases
Annotating	Designed for annotating sequence	Designed for annotating sequence	Users may add “private” annotation

Critical evaluation of evidence

Weighing the Evidence



full-length cDNA

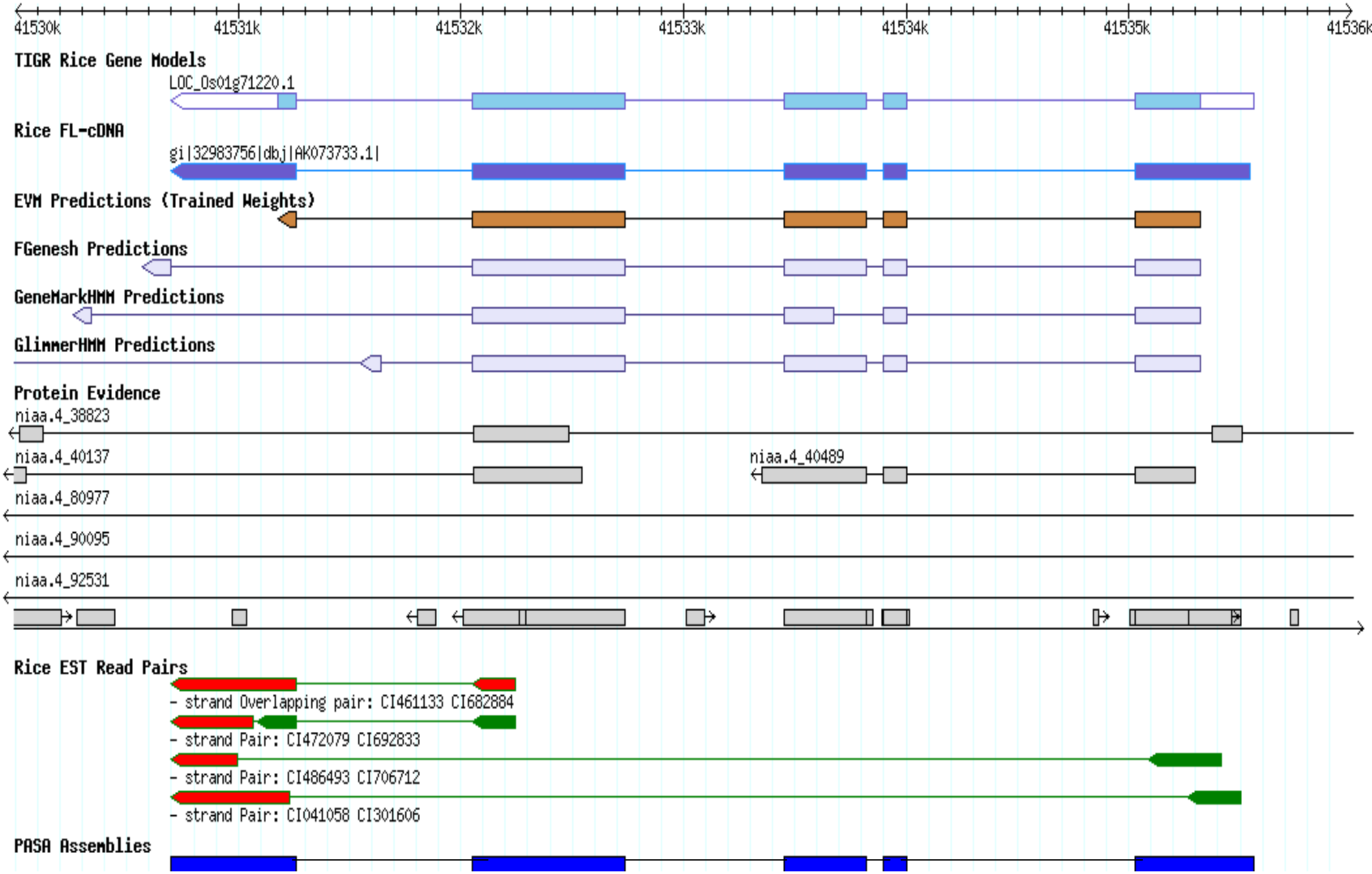
EST

protein

gene prediction

Full-length cDNAs are the most reliable pieces of evidence, followed by EST, protein, and finally gene prediction evidence. Evidence of the highest reliability is used for structural annotation.

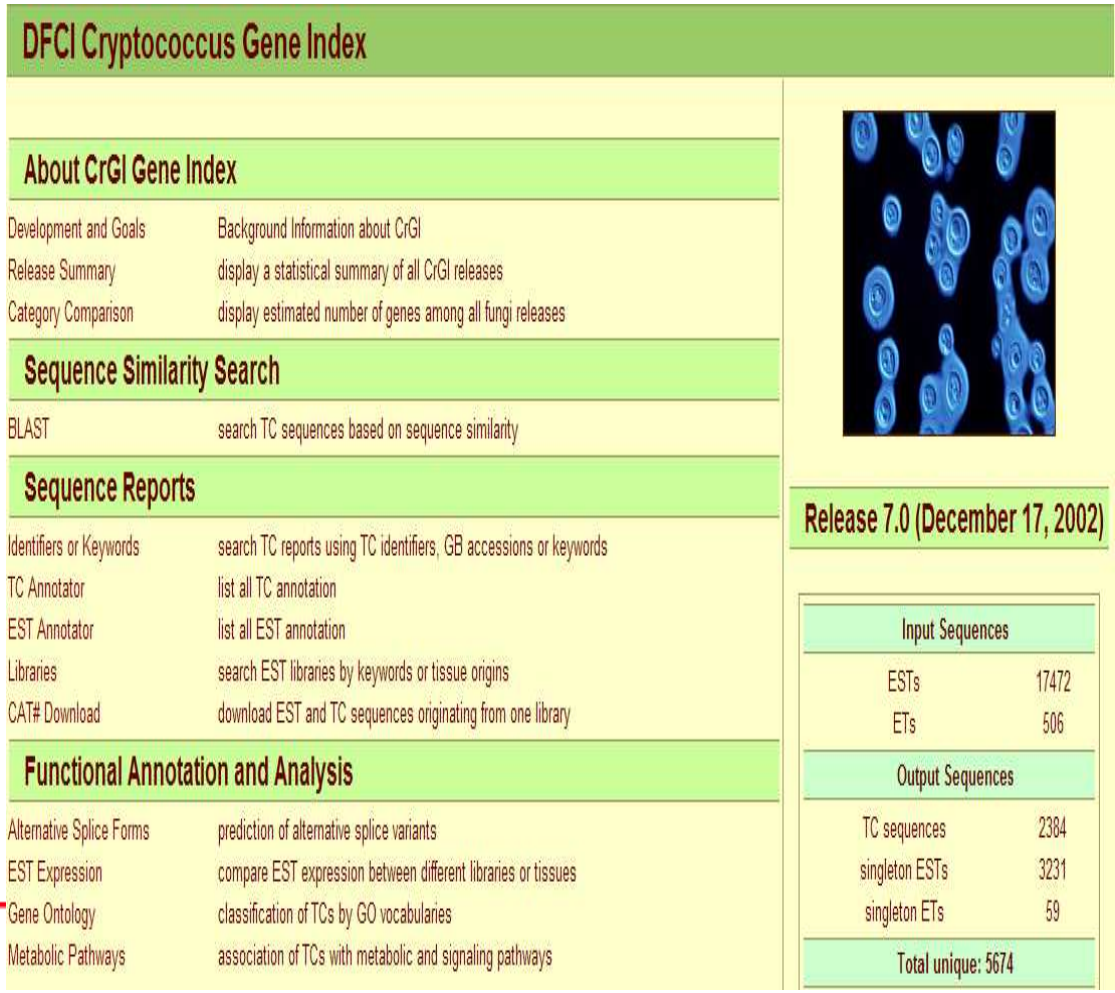
full-length cDNA Evidence



EST Evidence

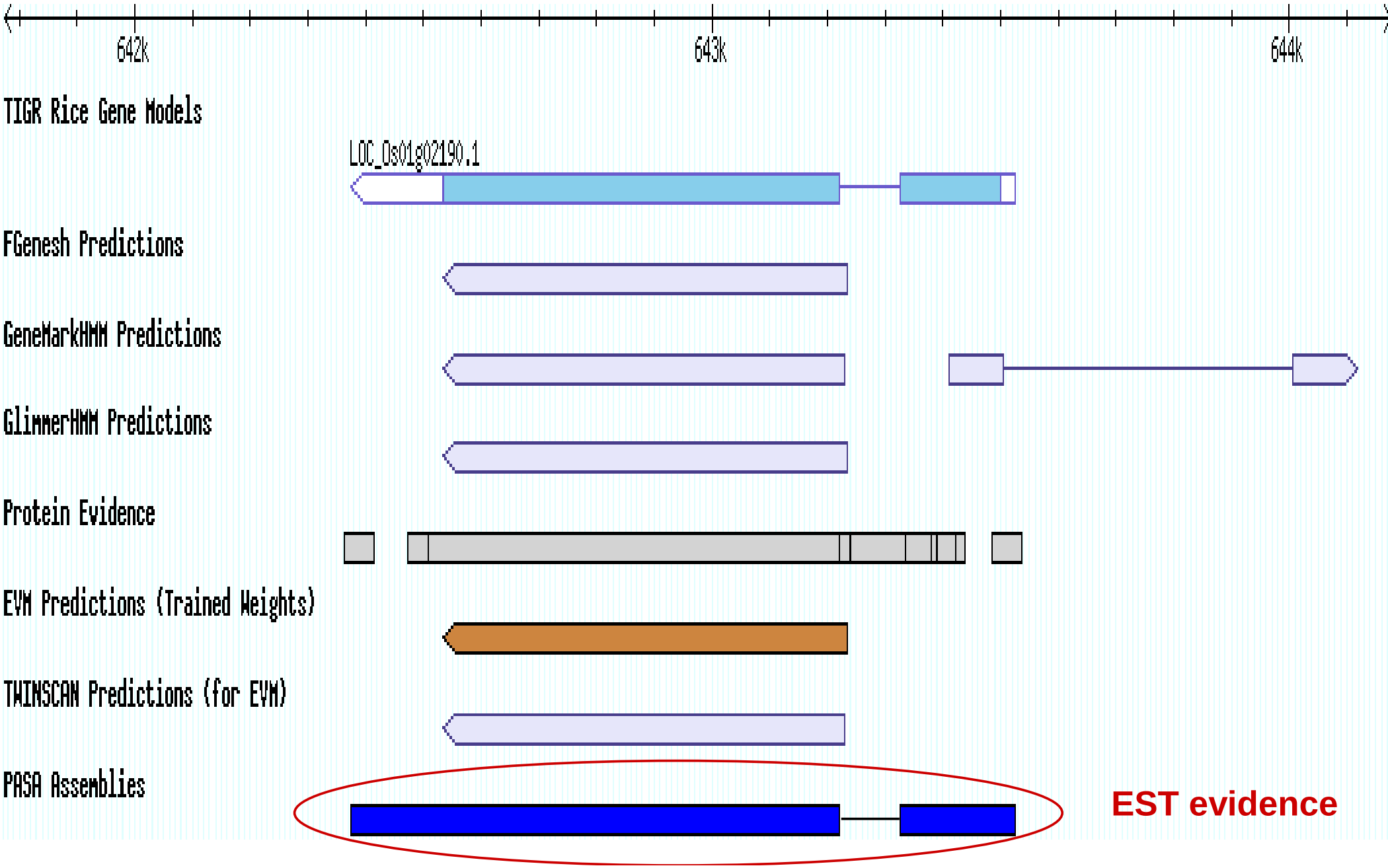
When encountering EST evidence, one can examine information about the EST or the ESTs used to make a Tentative Consensus (TC) sequence at Gene Indices WebPages:

http://compbio.dfci.harvard.edu/tgi/

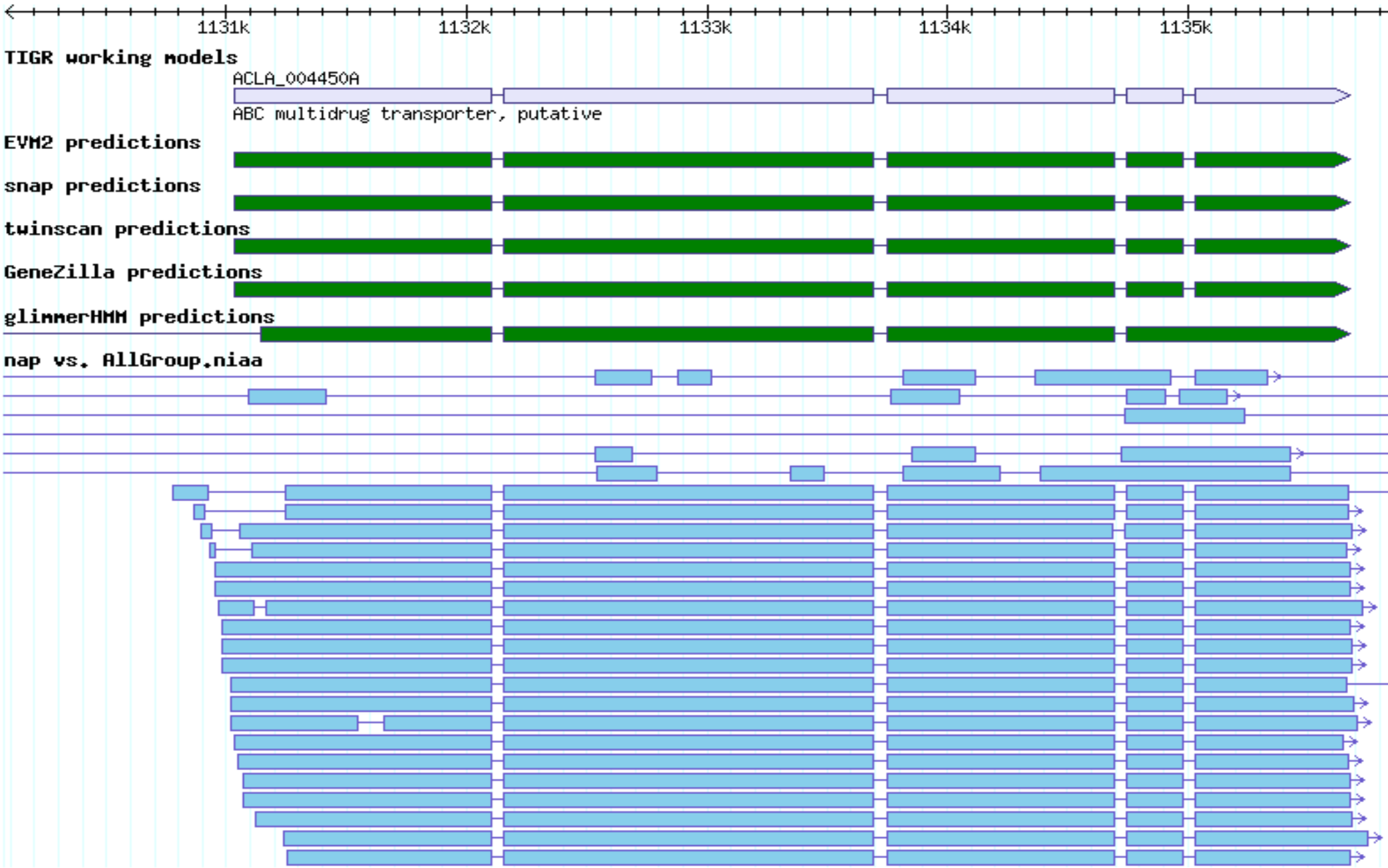


J. Craig Venter
INSTITUTE

EST Evidence



Protein Evidence



Questions?