FlyBase

How High-Throughput Data are Informing Gene Models

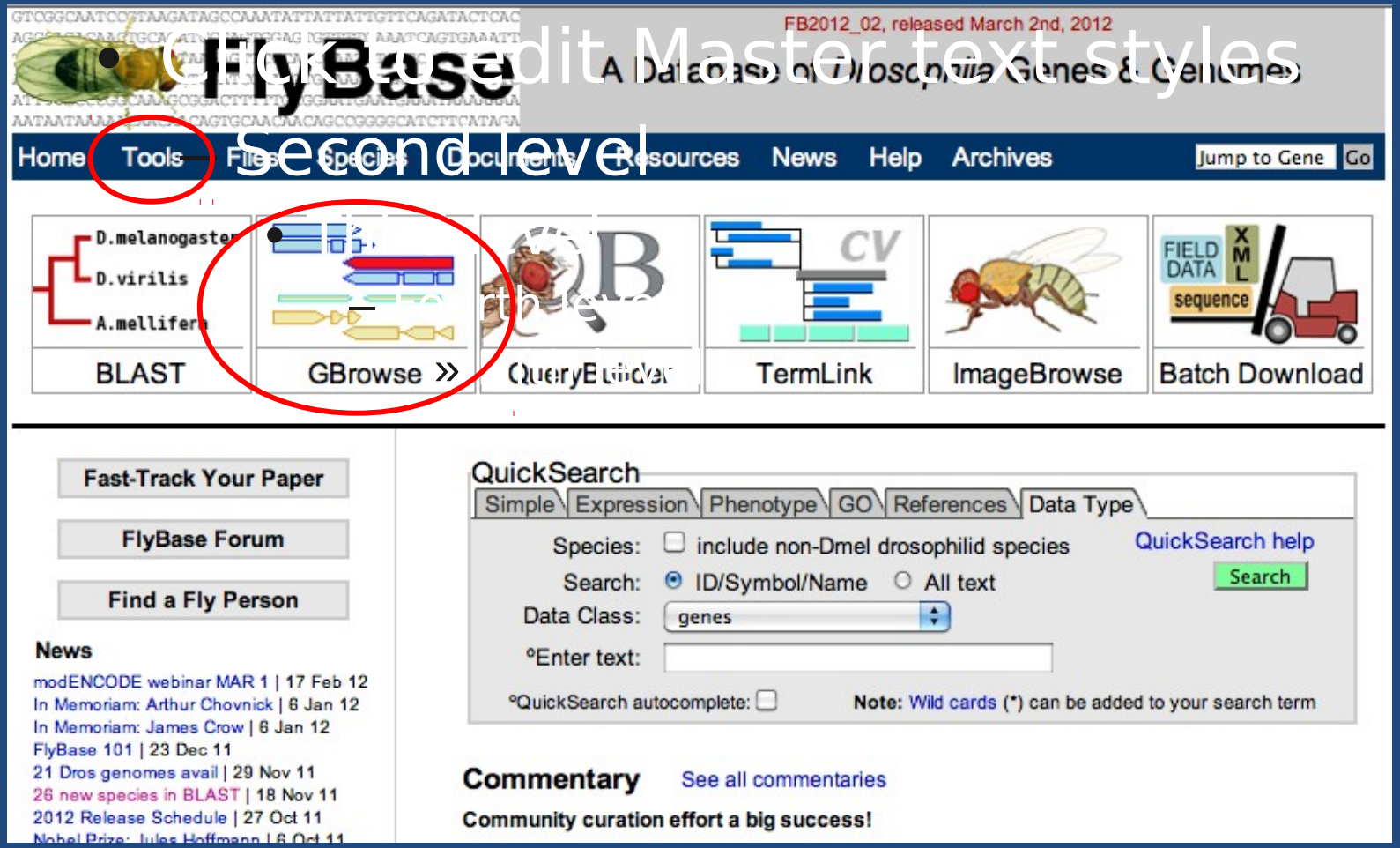
New high-throughput data

- RNA-Seq junction calls (introns)
- Stranded RNA-Seq coverage data from multiple tissues and cell lines
- Stop-codon readthrough predictions
- A-to-I RNA editing sites
- Transcription start site (TSS) mapping

Integration of High Throughput Data

- Presentations in GBrowse
- Sequence feature reports
- New and improved annotations
- Assessing the data
- Annotation on the genome of transcript features
- Use of gene model comments

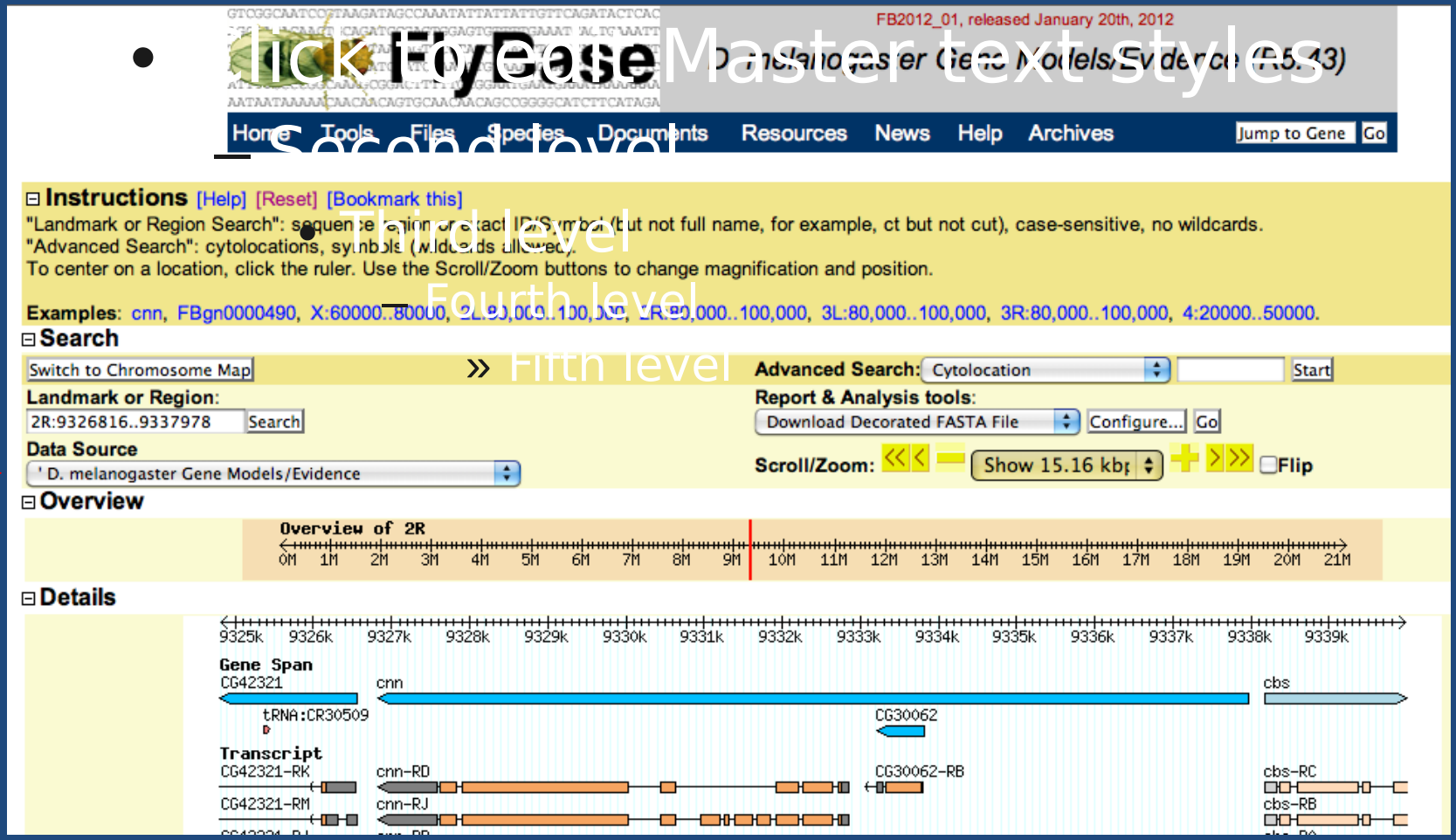
GBrowse link on top page



- Edit Master text styles
  - Second level

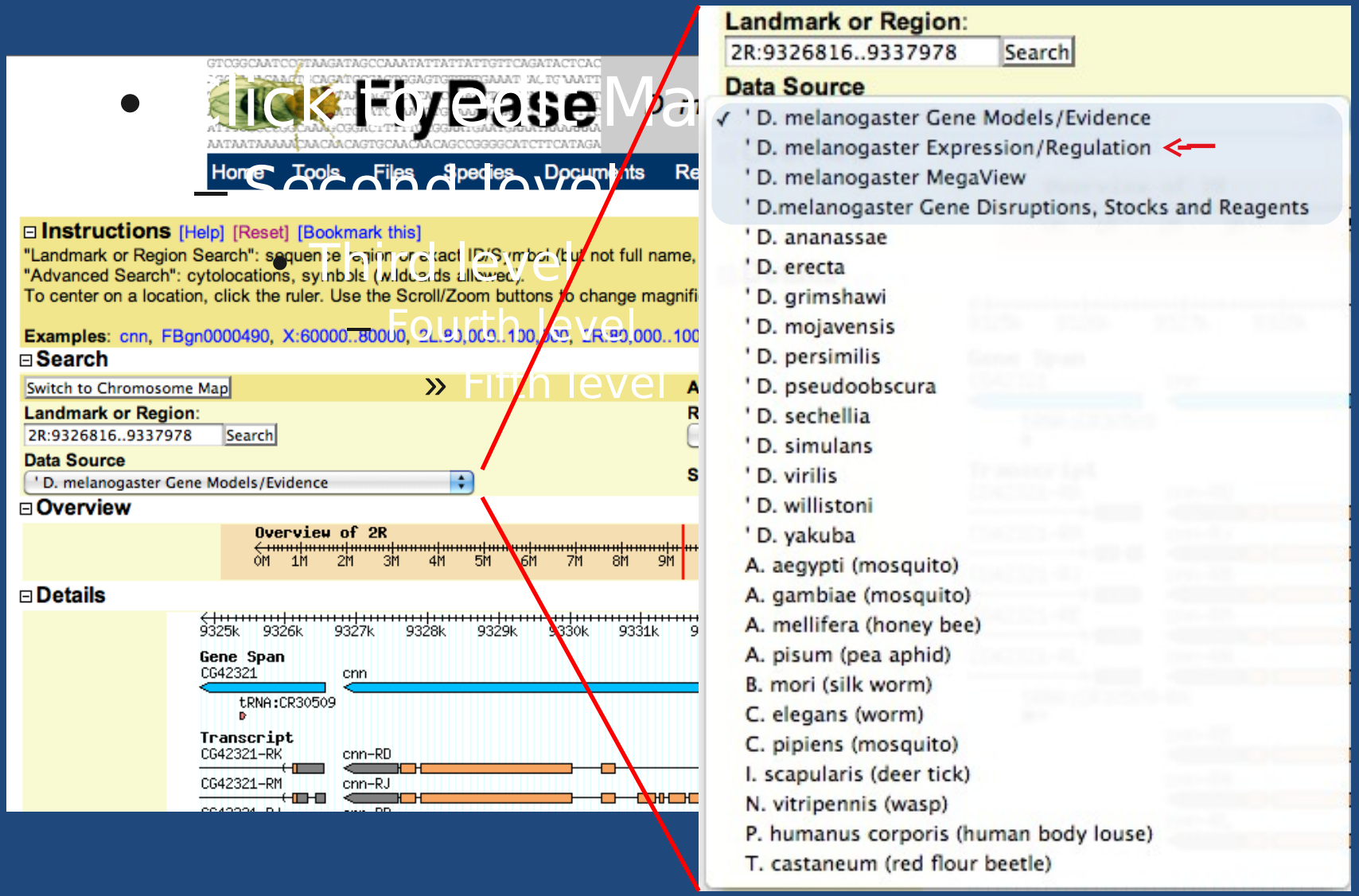
GBrowse: Data Source default is Gene Models/Evidence

- Click FlyBase Master text styles
  - Second level
    - Third level
      - Fourth level
        - » Fifth level



GBrowse: Data Sources

- Tick FlyBase Ma
  - Second level
    - Third level
      - Fourth level
        - Fifth level



FlyBase

FB2012_01, released January 20th, 2012

Gene Dmel\chinmo

Home | Tools | Files | Species | Documents | Resources | News | Help | Archives | Jump to Gene | Go

Profile Manager | Help | Open All | Close All

General Information			
Symbol	Dmel\chinmo	Species	*D. melanogaster*
Name	Chronologically inappropriate morphogenesis	Annotation symbol	CG31666
Feature type	protein_coding_gene	FlyBase ID	FBgn0086758
Gene Model Status	Current	Stock availability	14 publicly available
Also Known As	CG17156		
Genomic Location			
Chromosome (arm)	2L	Recombination map	
Cytogenetic map	22A5-22B1	Sequence location	2L:1,651,260..1,704,682 [+]

Genomic Maps

Select View: Gene Models/E

View in GBrowse

modENCODE GBrowse

Decorated FastA

Get genome region

Gene region

Get FastA

Summary Information

Recent Updates

GBrowse: Expression/Regulation

- Click to edit Master text styles
  - Sec
    - Th

gene extents

annotated transcripts

natural TE's

cDNAs

FlyAtlas tissue-level expression data

modENCODE RNA-Seq developmental timecourse

GBrowse: Expression/Regulation

- Click to edit Master text styles
  - Sec
    - Th

gene extents

annotated transcripts

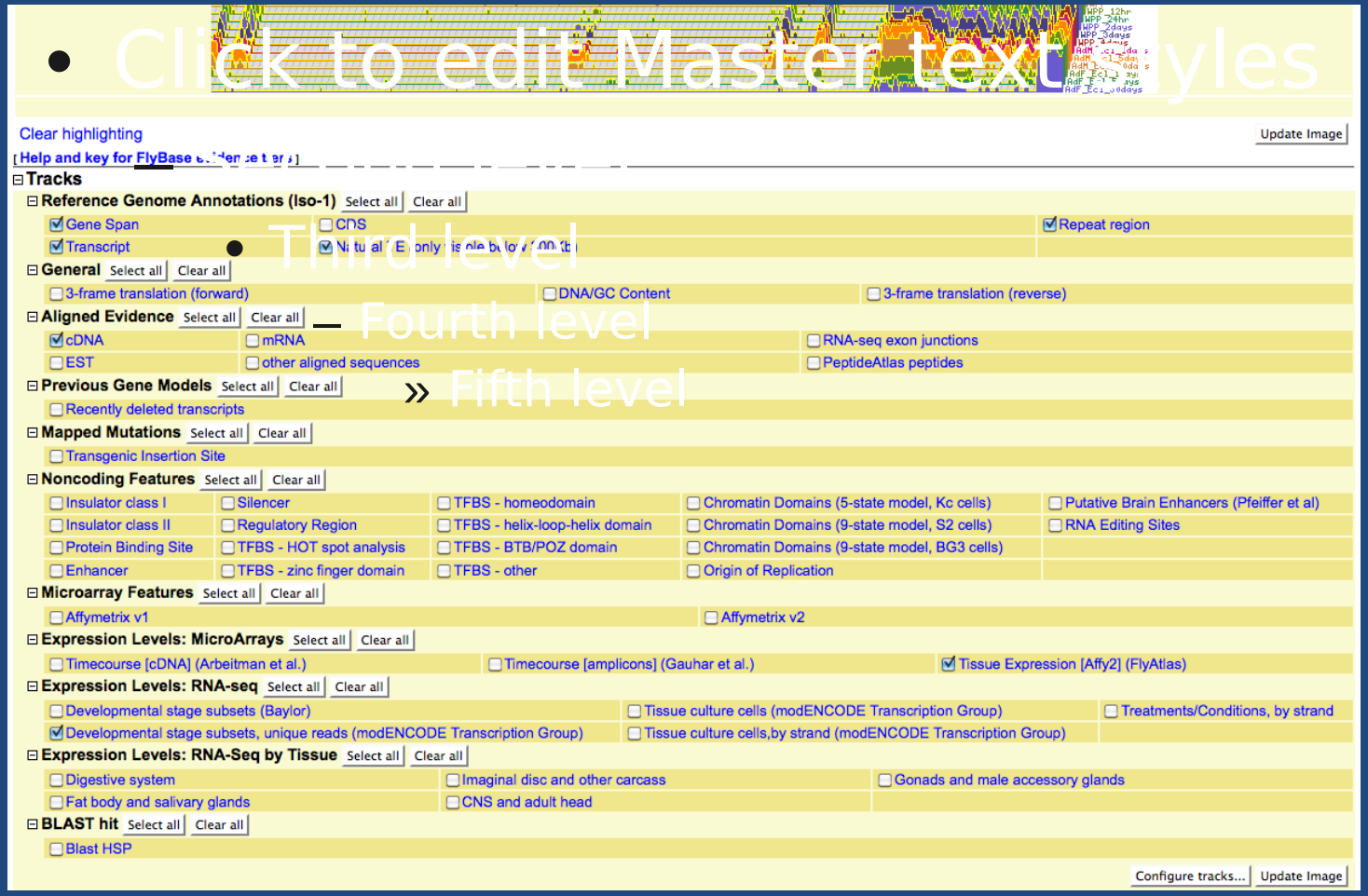
natural TE's

cDNAs

FlyAtlas tissue-level expression data

modENCODE RNA-Seq developmental timecourse

GB: Expression/regulation data options



Select from Aligned Evidence: RNA-seq exon junctions

'Select all' from Expression Levels: RNA-Seq by Tissue

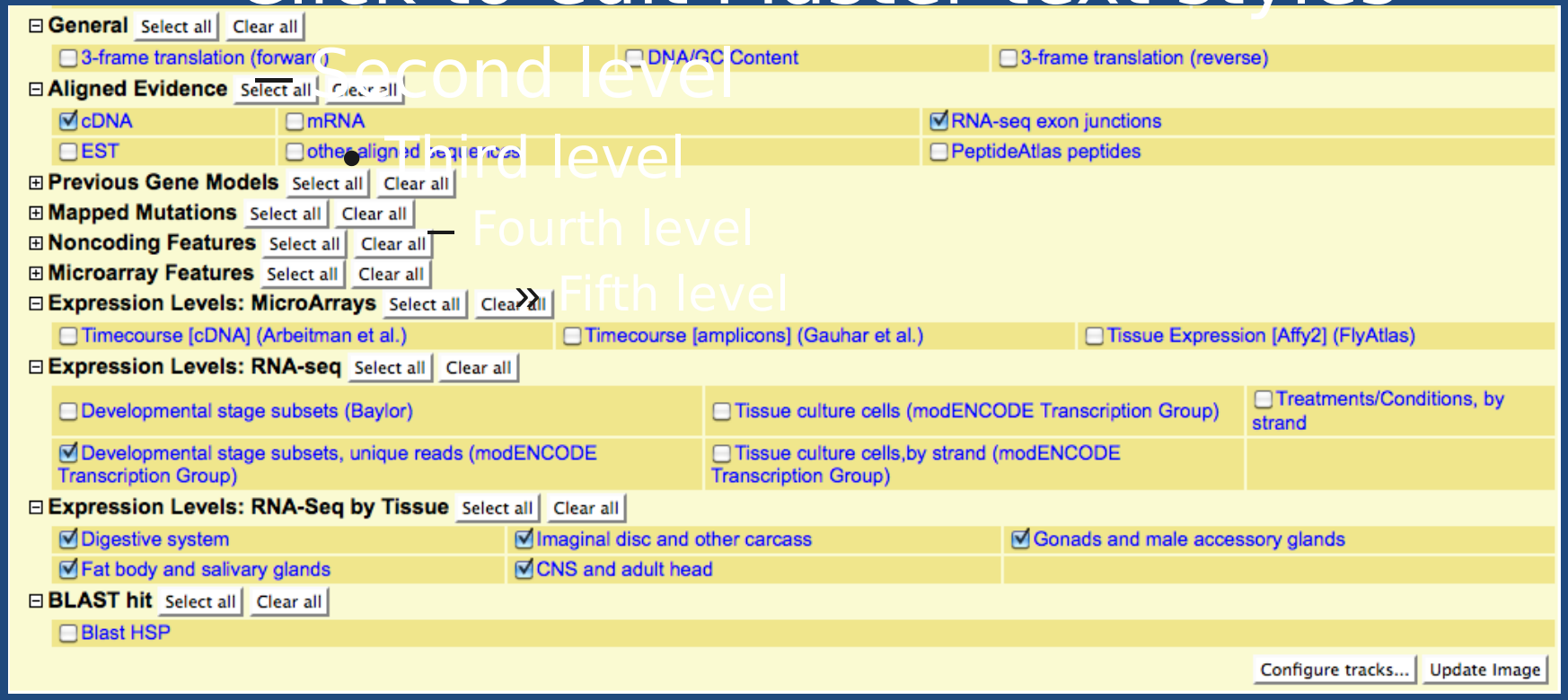
Click to edit Master text styles

Second level

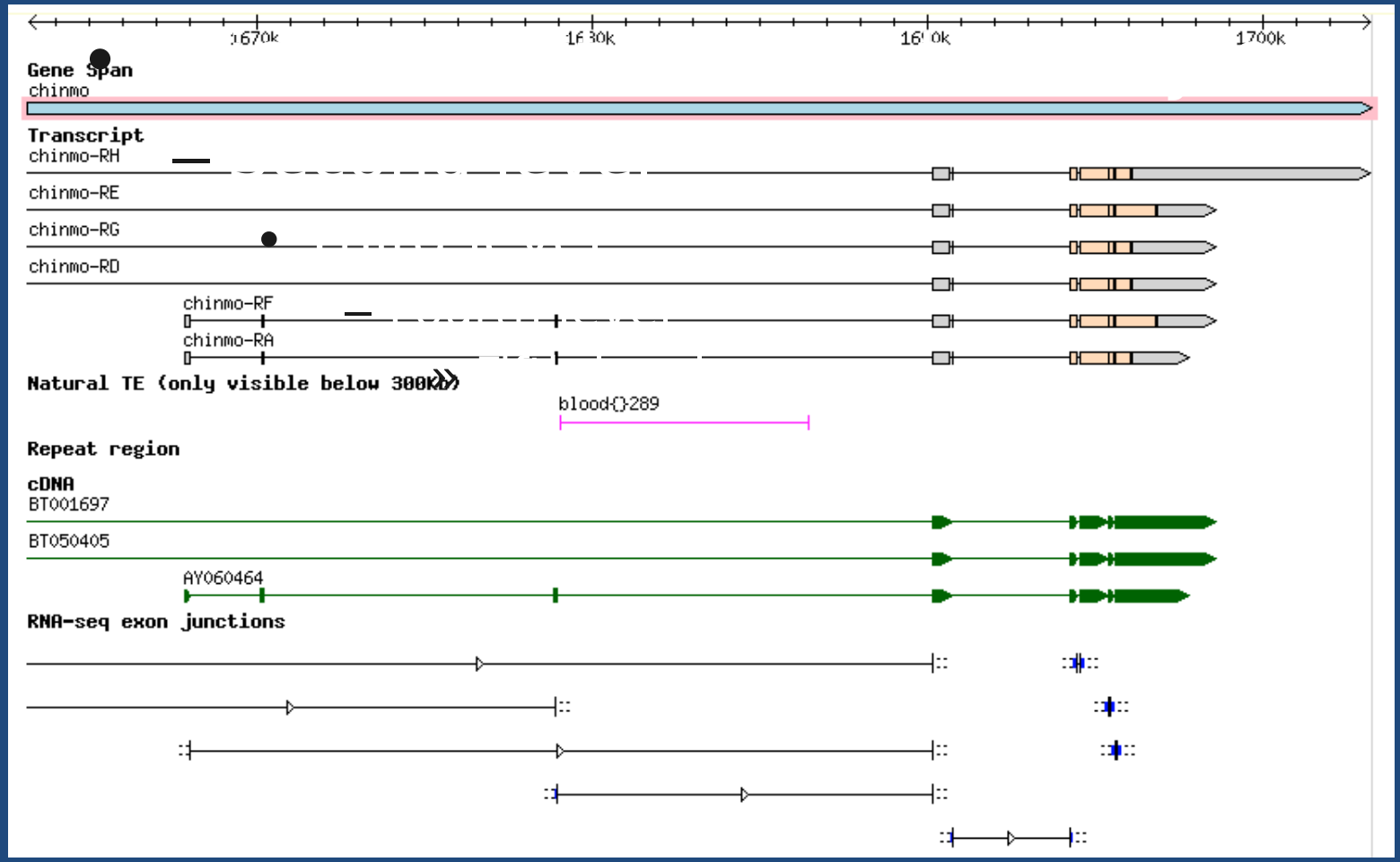
Third level

Fourth level

Fifth level

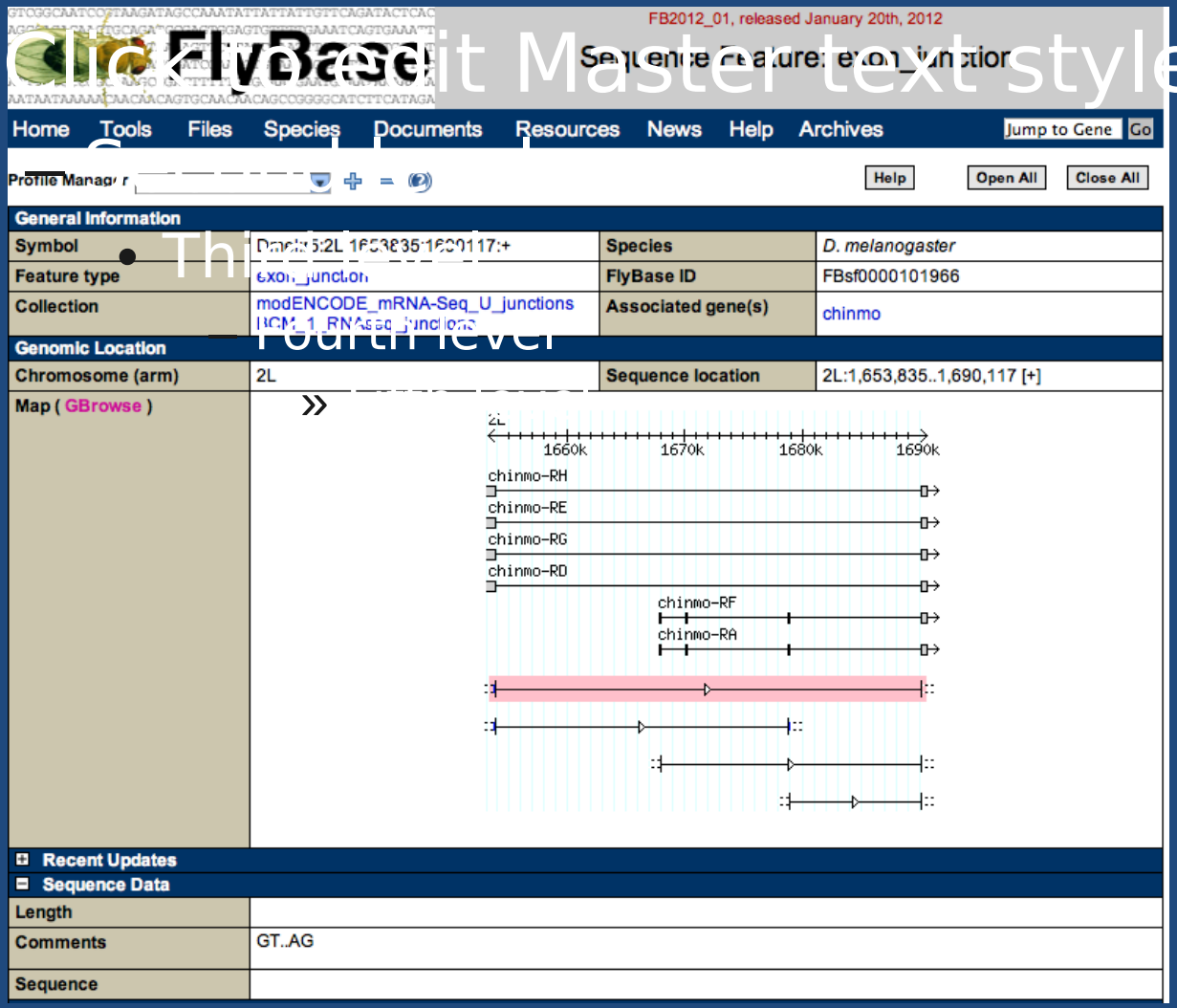


GBrowse: RNA-seq exon junctions



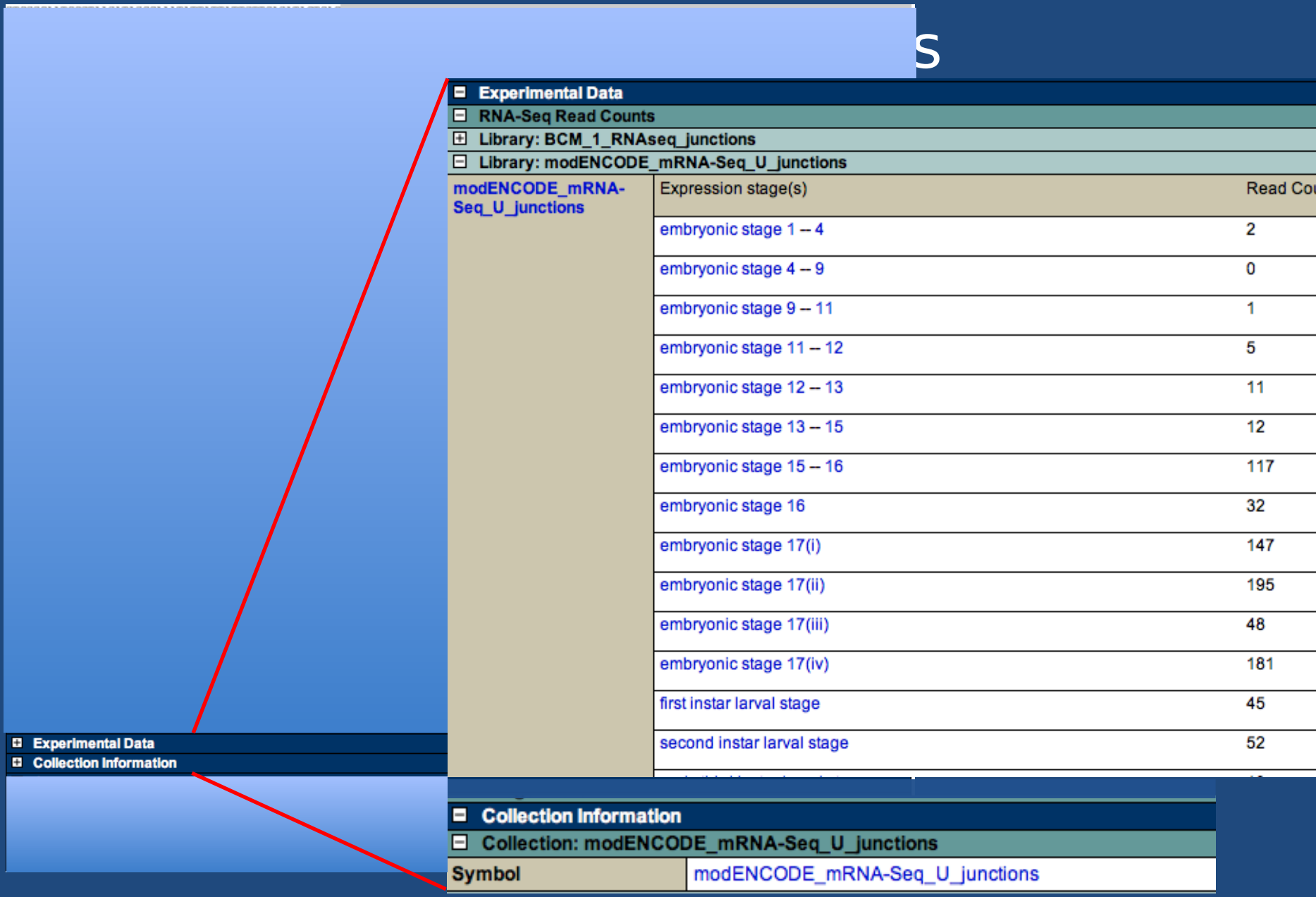
Sequence feature reports for RNA-Seq exon junctions

- Click FlyBase
- Edit Master text styles
  - Second level
    - Third level
      - Fourth level



s

Experimental Data
Collection Information



Experimental Data
RNA-Seq Read Counts
Library: BCM_1_RNAseq_junctions
Library: modENCODE_mRNA-Seq_U_junctions

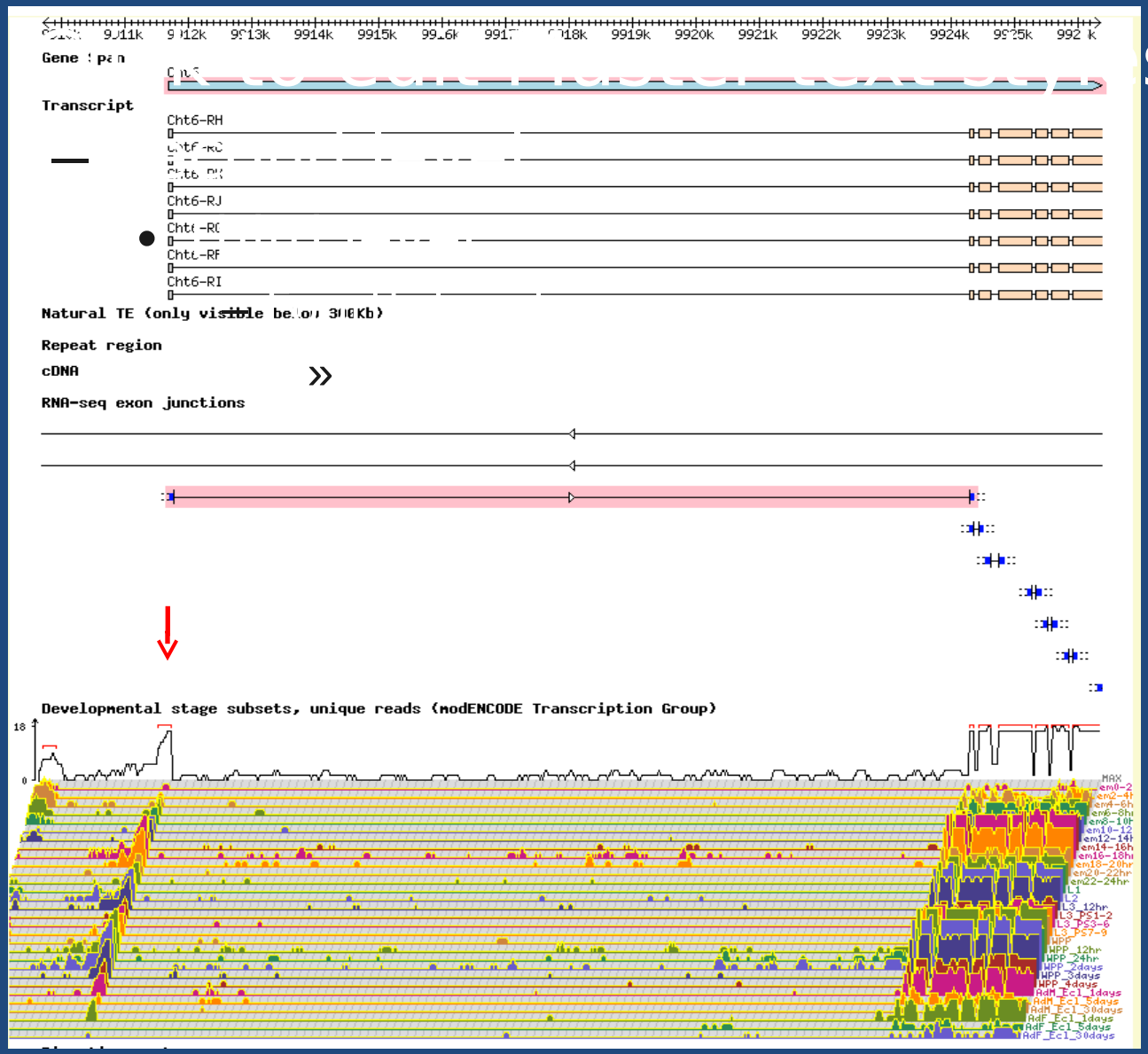
modENCODE_mRNA-Seq_U_junctions	Expression stage(s)	Read Cou
	embryonic stage 1 -- 4	2
	embryonic stage 4 -- 9	0
	embryonic stage 9 -- 11	1
	embryonic stage 11 -- 12	5
	embryonic stage 12 -- 13	11
	embryonic stage 13 -- 15	12
	embryonic stage 15 -- 16	117
	embryonic stage 16	32
	embryonic stage 17(i)	147
	embryonic stage 17(ii)	195
	embryonic stage 17(iii)	48
	embryonic stage 17(iv)	181
	first instar larval stage	45
	second instar larval stage	52

Collection Information
Collection: modENCODE_mRNA-Seq_U_junctions

Symbol	modENCODE_mRNA-Seq_U_junctions

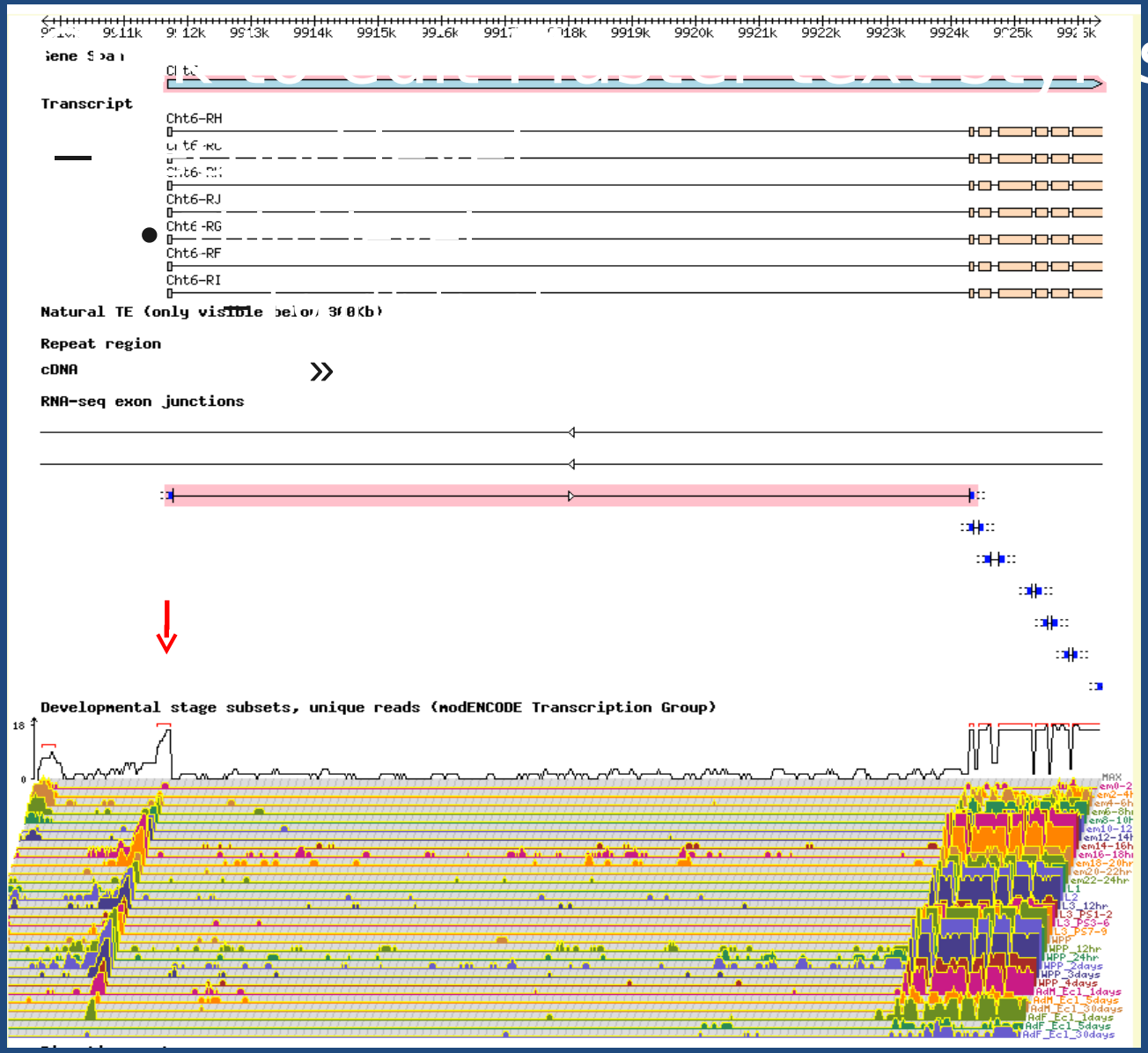
New 5’ exon (UTR) added to Cht6

- [illegible]s



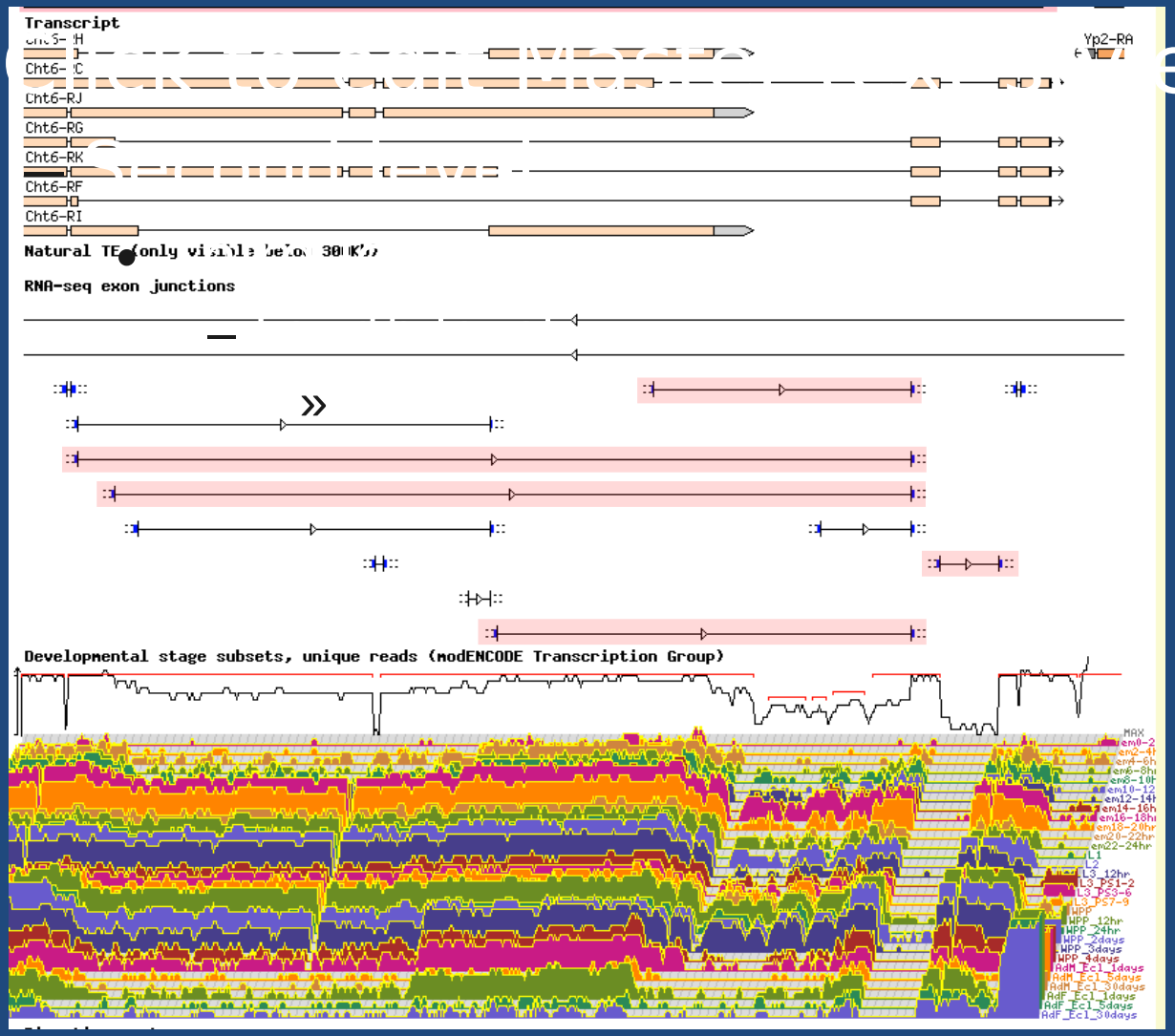
New 5’ exon (UTR) added to Cht6

- S

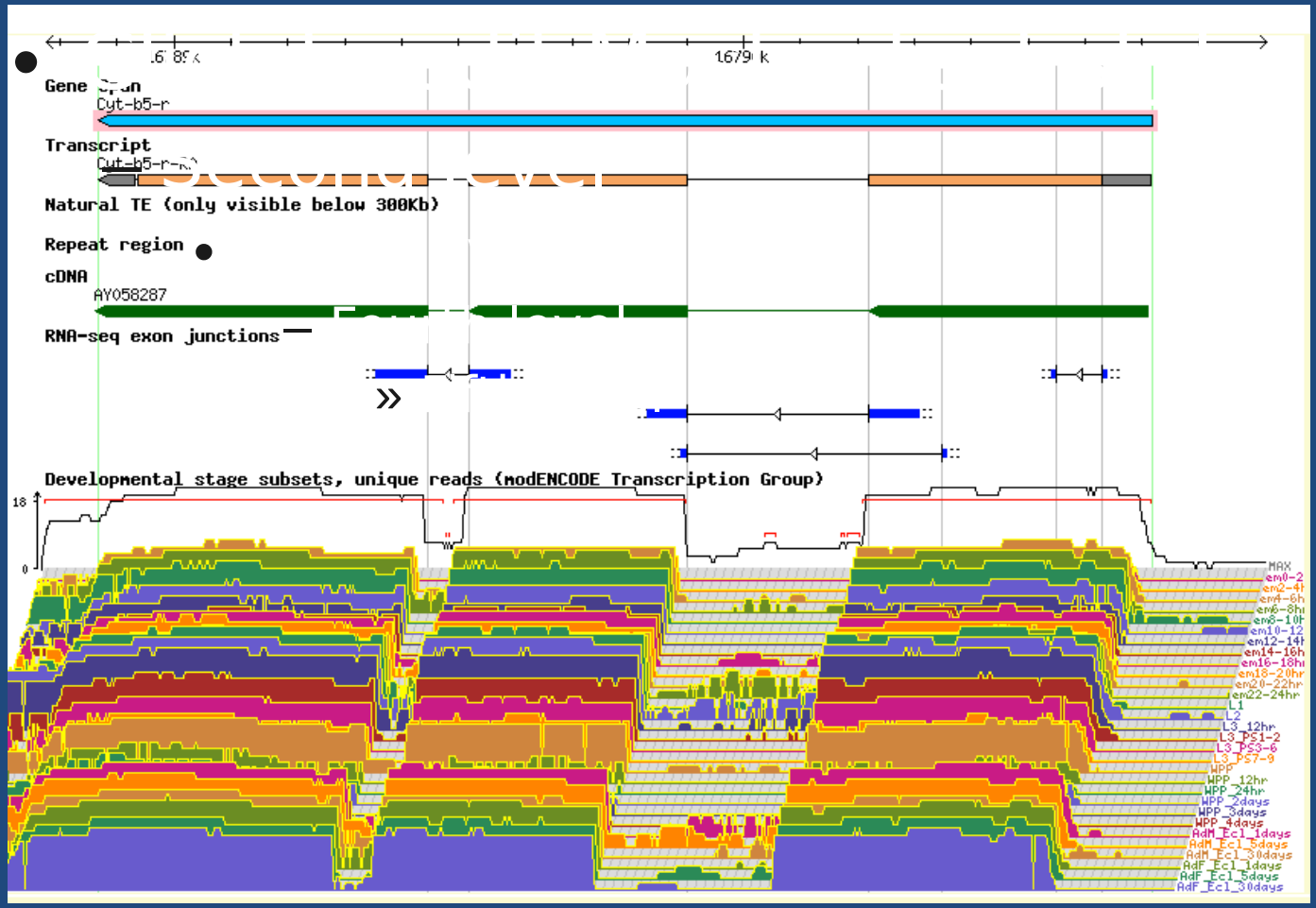


Gene merge: Cht6 + downstream genes

- [illegible]es

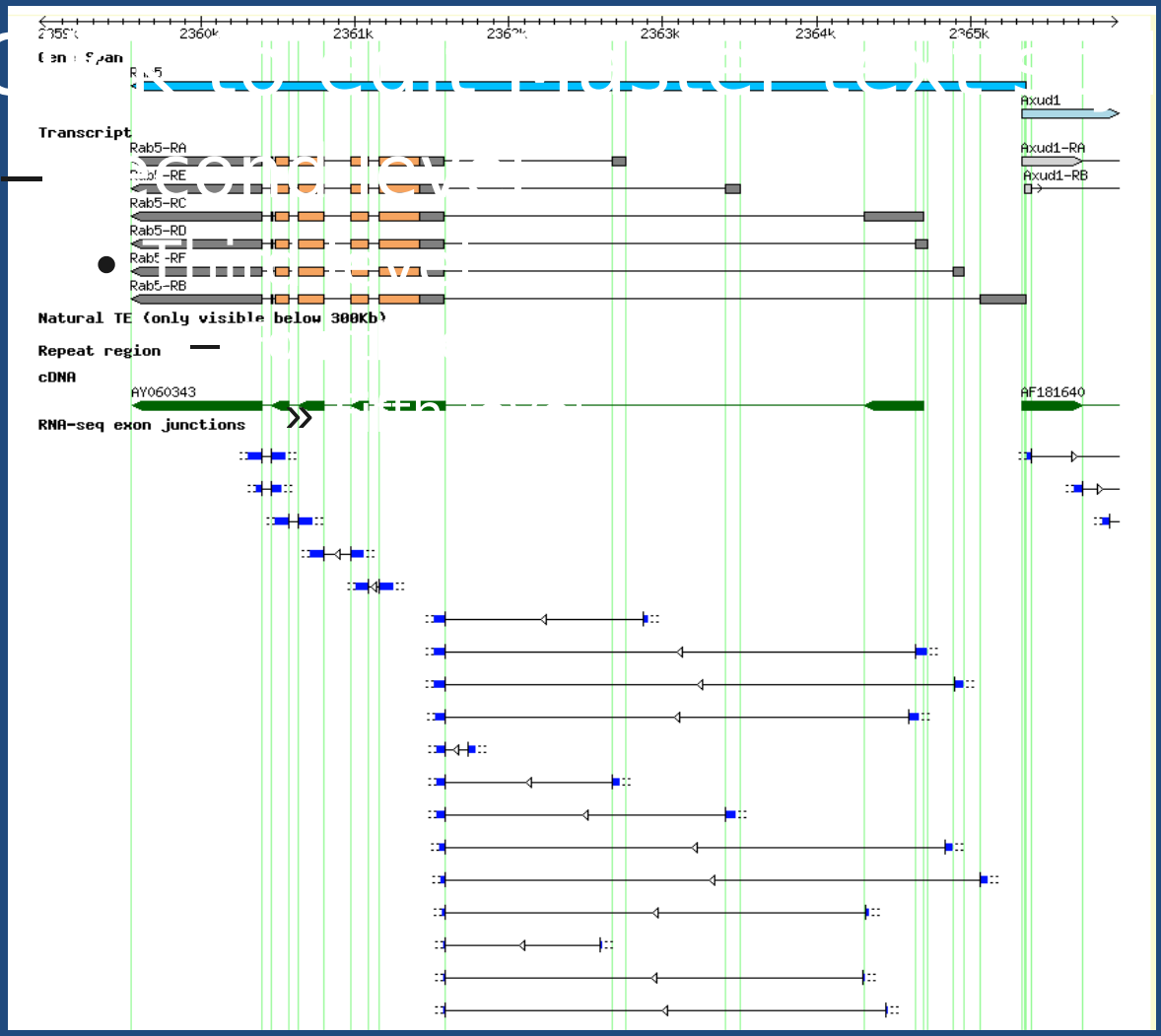


Junctions: low frequency junctions may not be annotated



Junctions: variations in 5’ UTR may not be annotated

- C[illegible]es

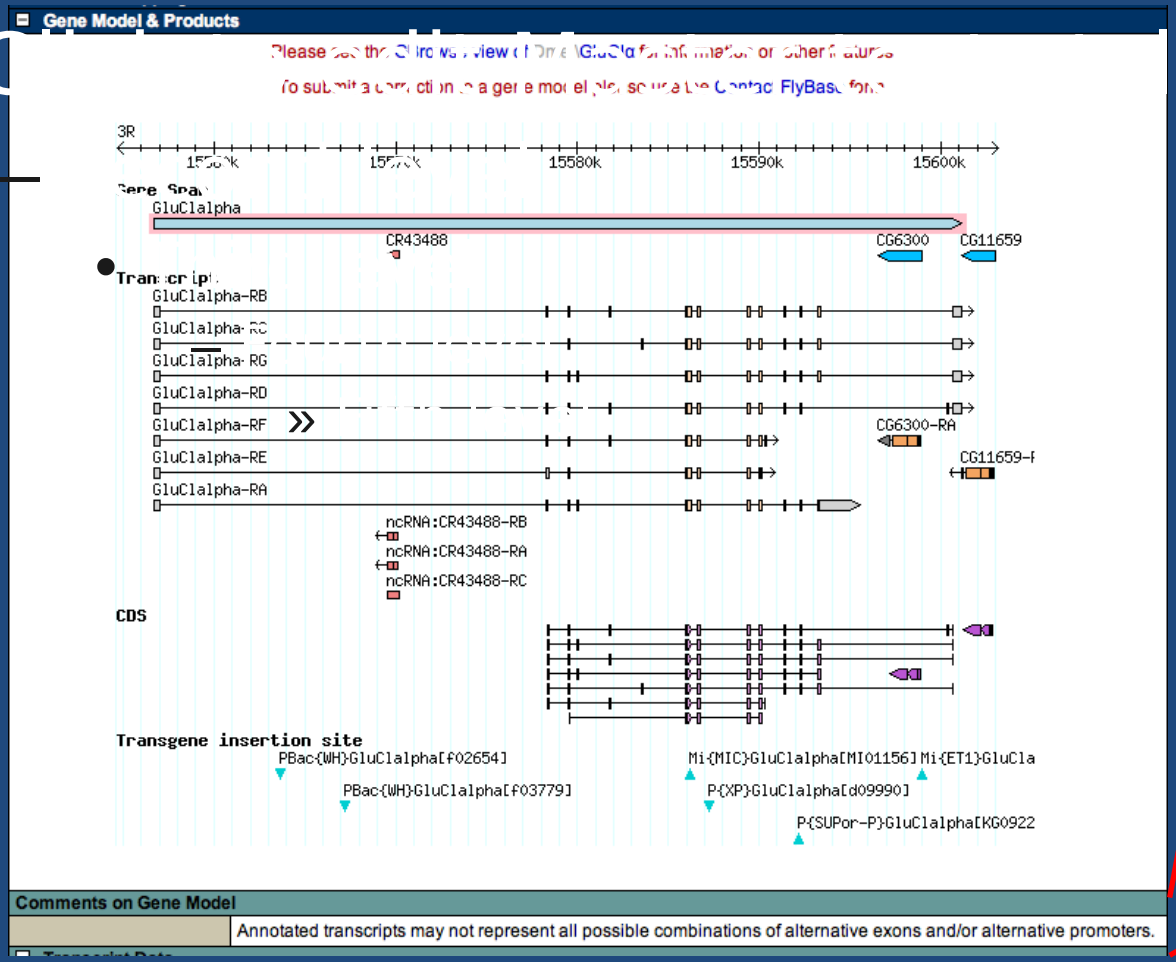


Gene model comments

- Low-frequency RNA-Seq exon junction(s) not annotated.
- Annotated transcripts do not represent all supported alternative splices within 5' UTR.

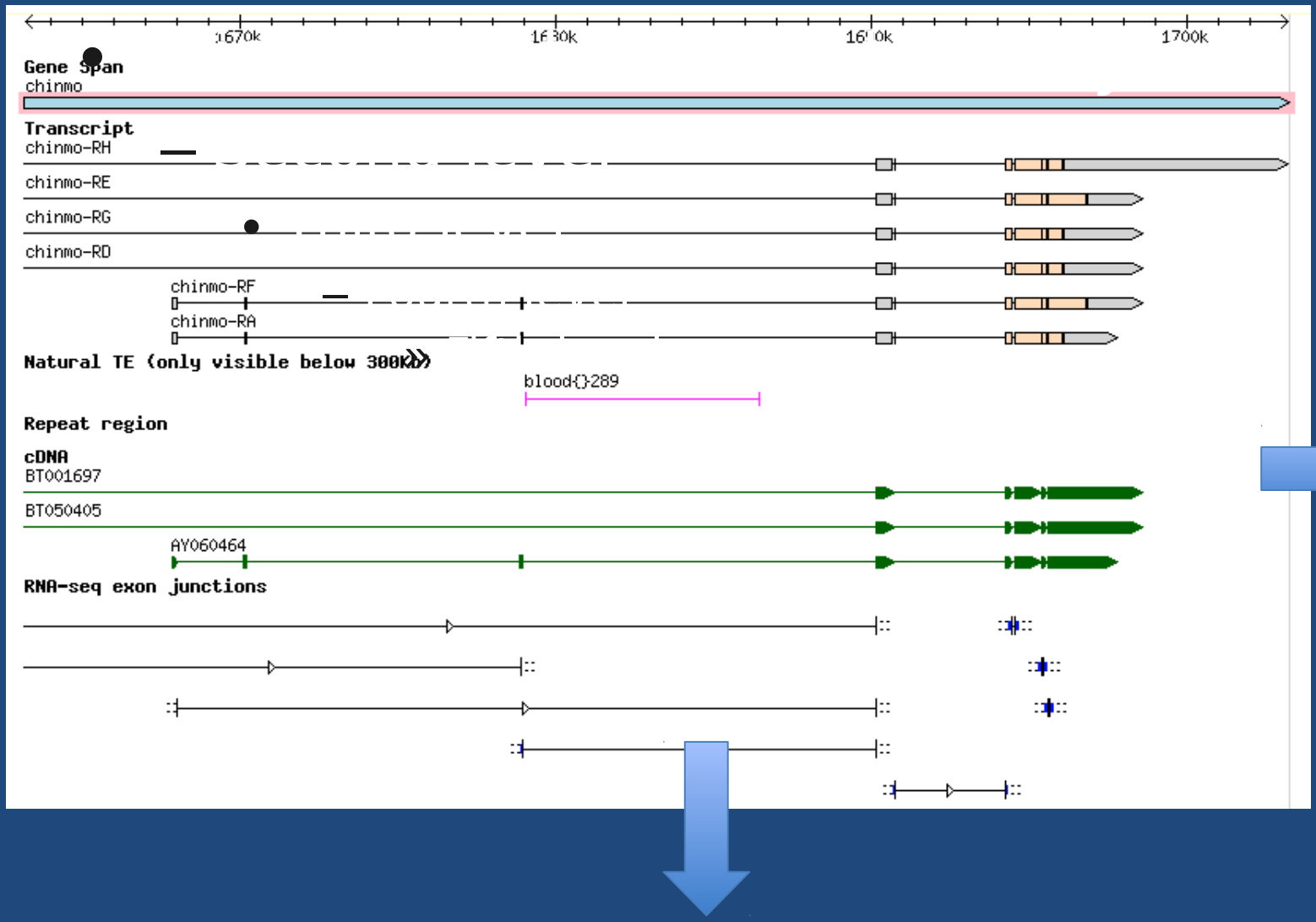
Comments on Gene Model (permutation problem)

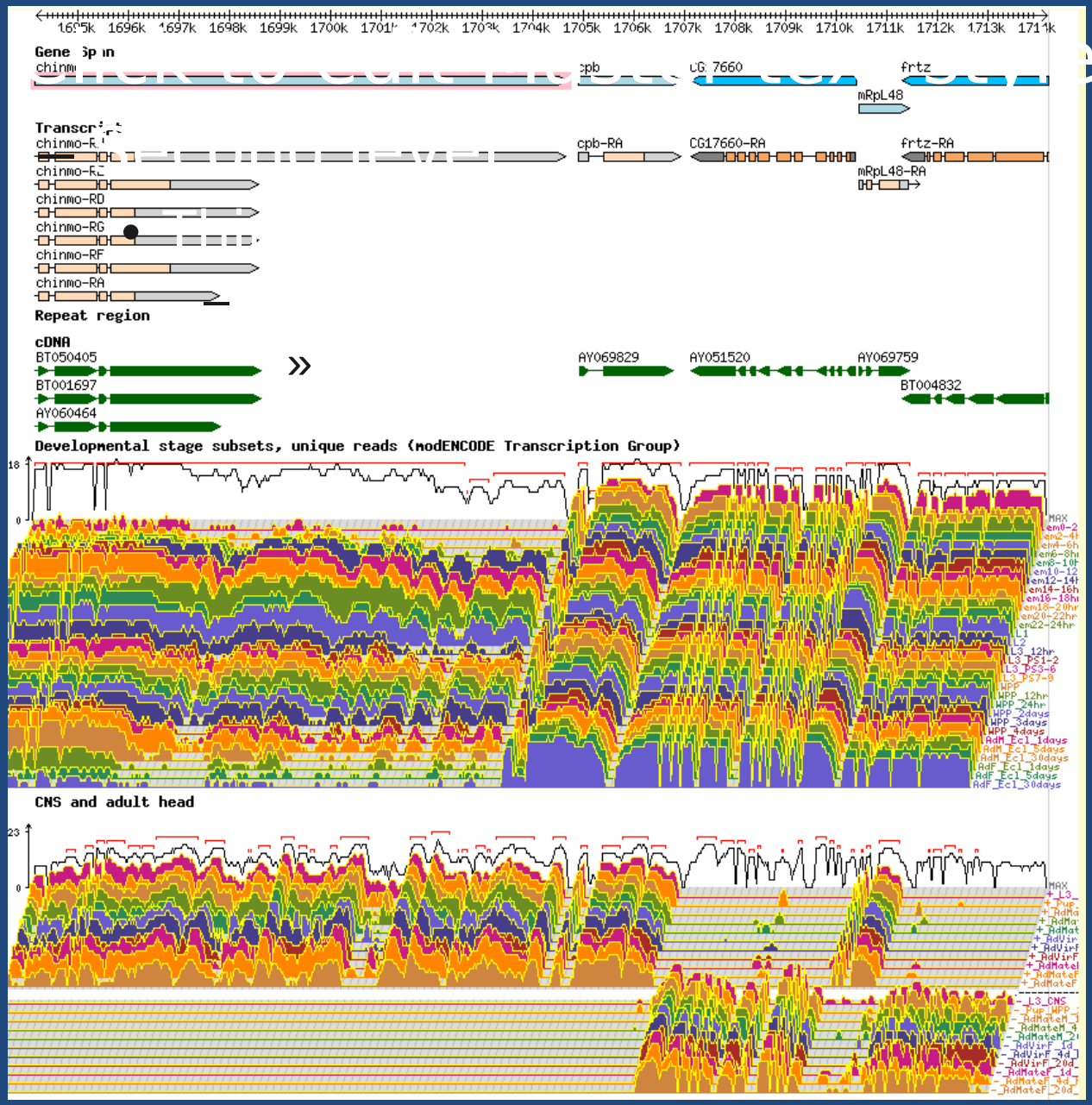
- C[illegible]es



Annotated transcripts may not represent all possible combinations of alternative exons and/or alternative promoters.

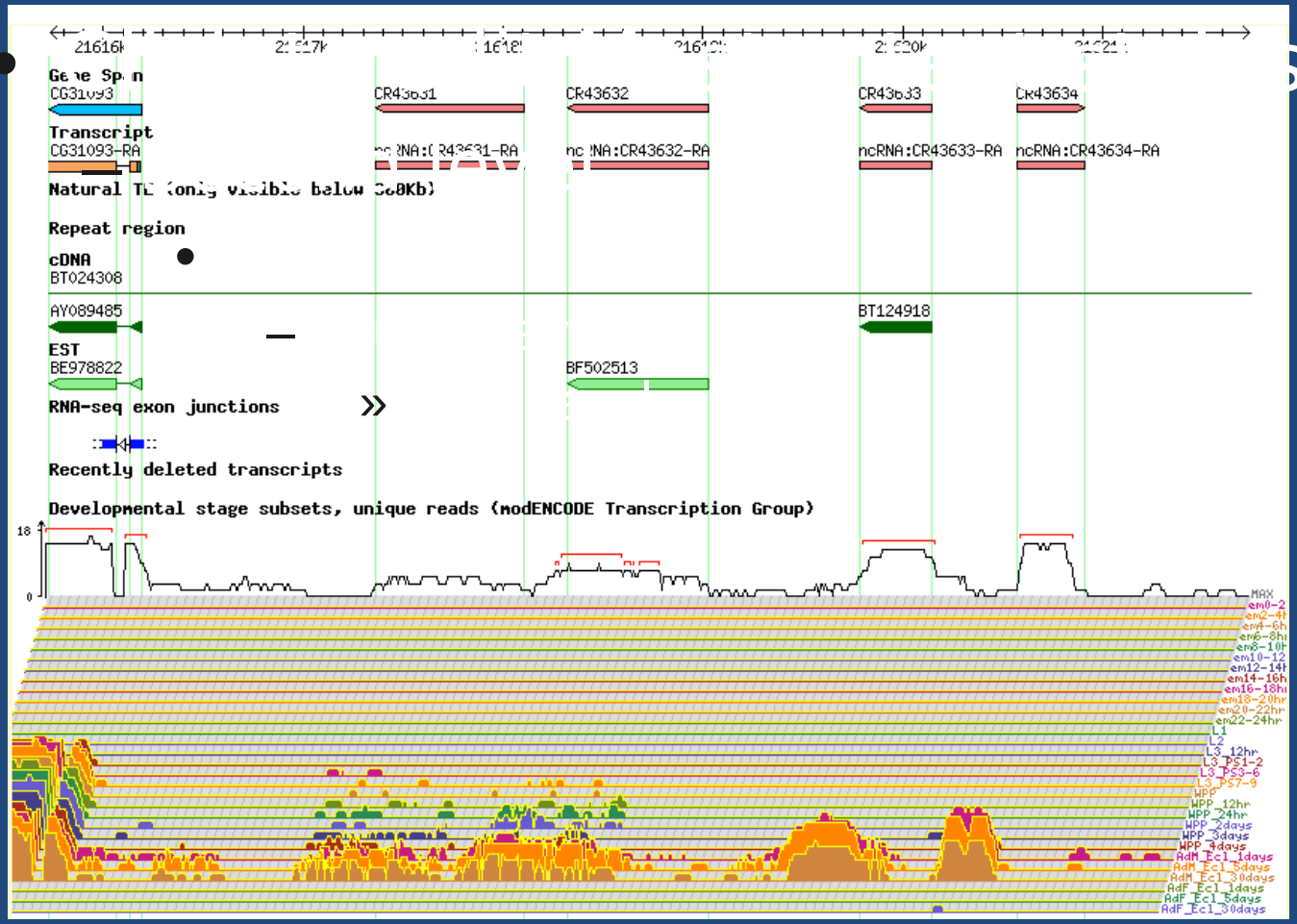
GBrowse: expression data





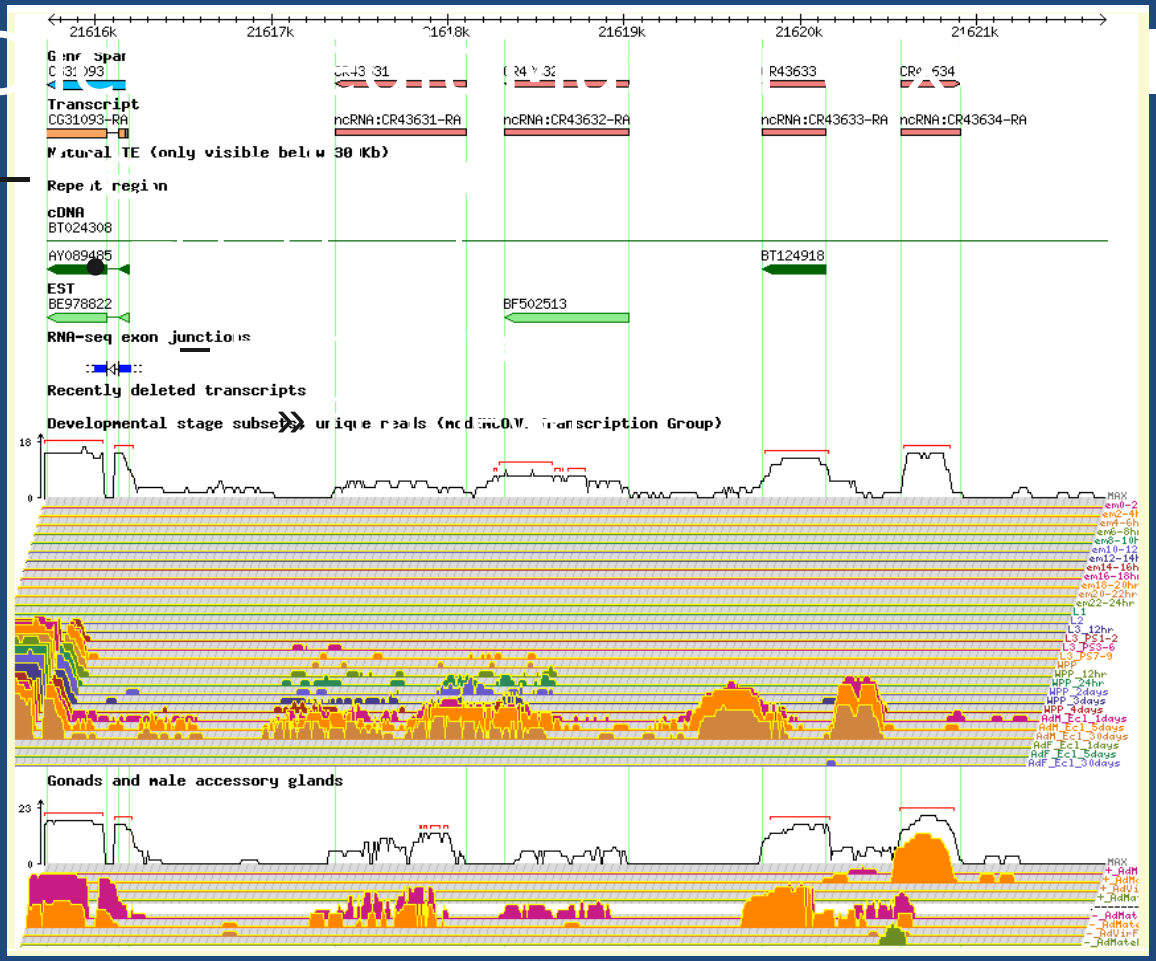
strand-specific RNA-Seq data from isolated L/A tissues

New genes: ncRNA genes

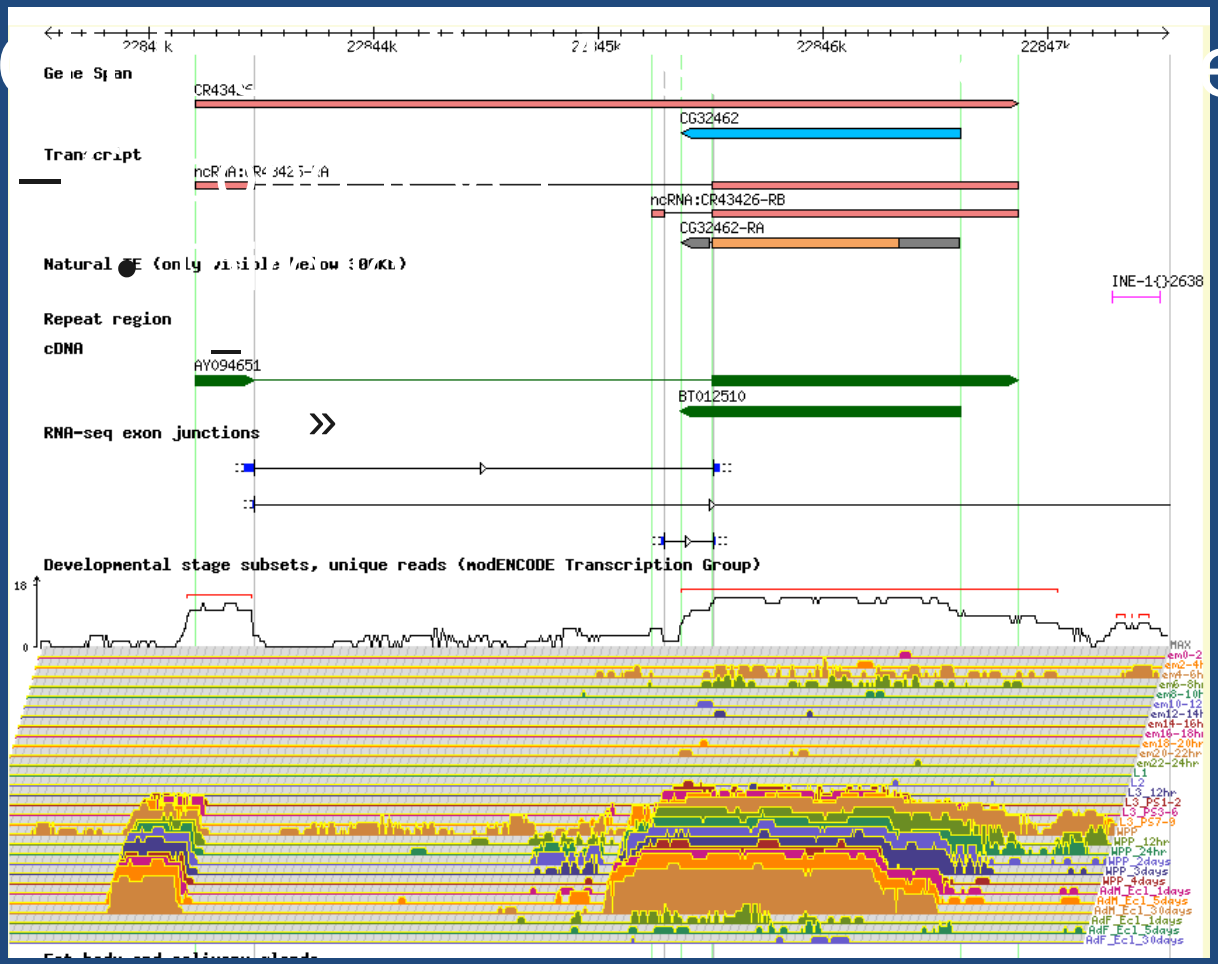


Stranded RNA-Seq data: ncRNA genes

- C les

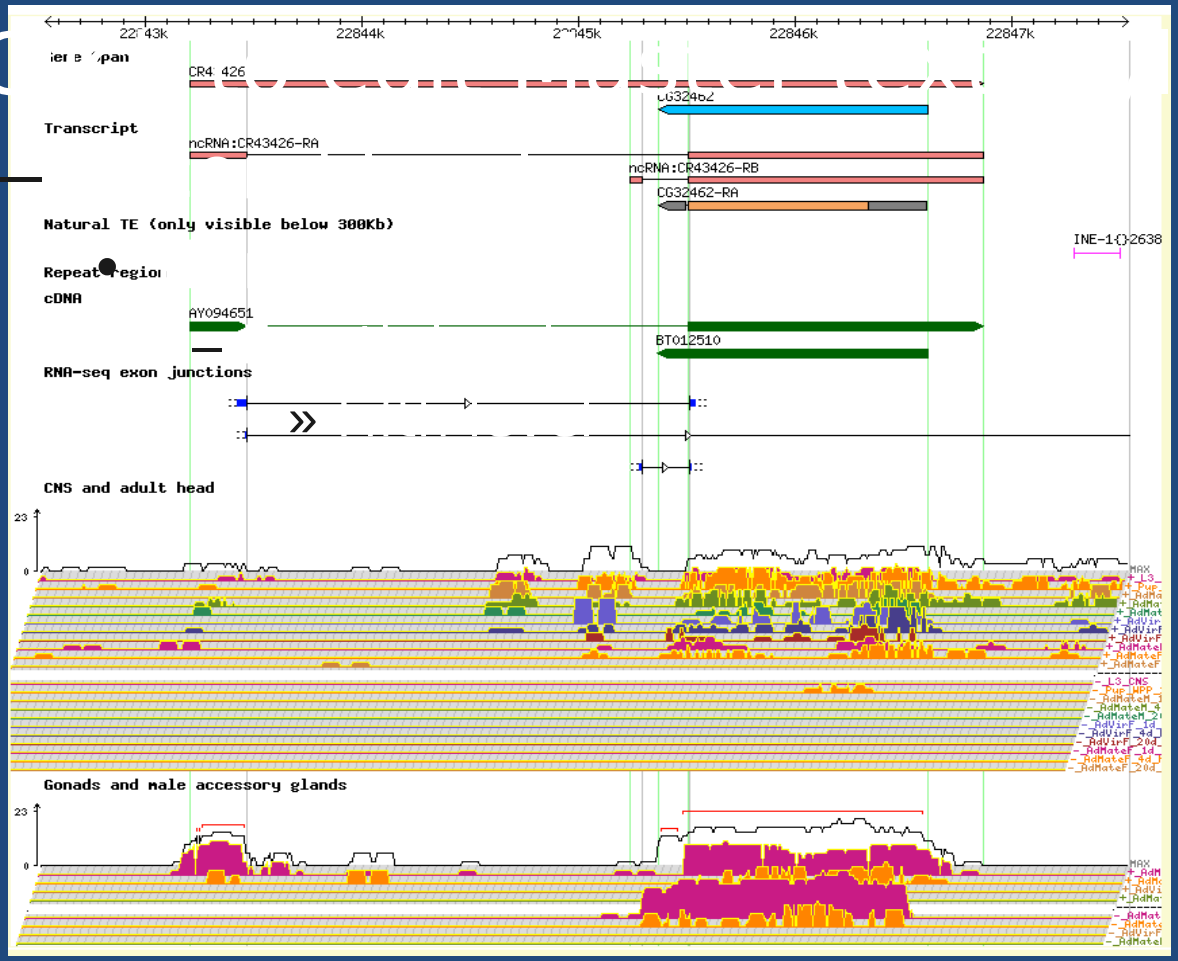


New genes: anti-sense RNAs

- 

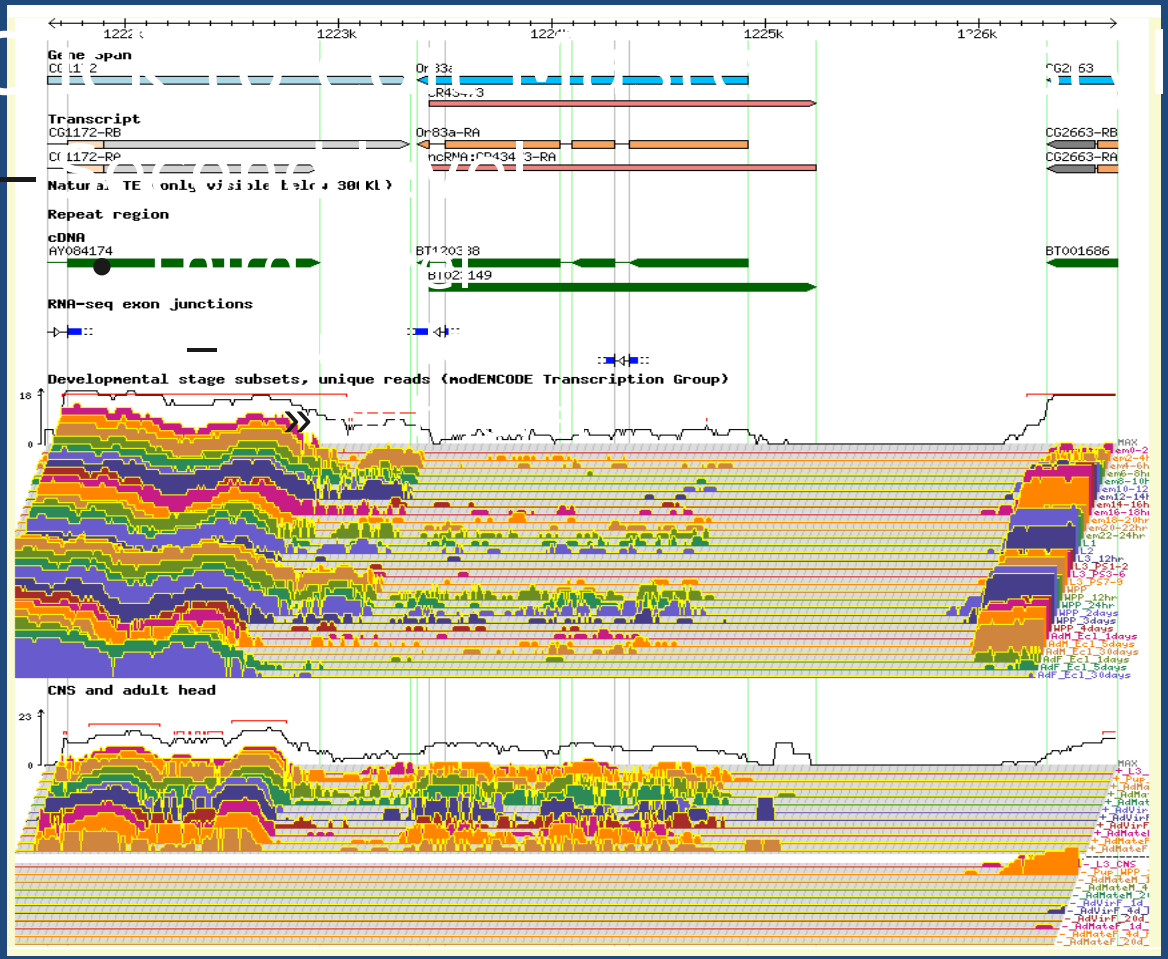
Stranded RNA-Seq data: anti-sense RNAs

- C[illegible]es

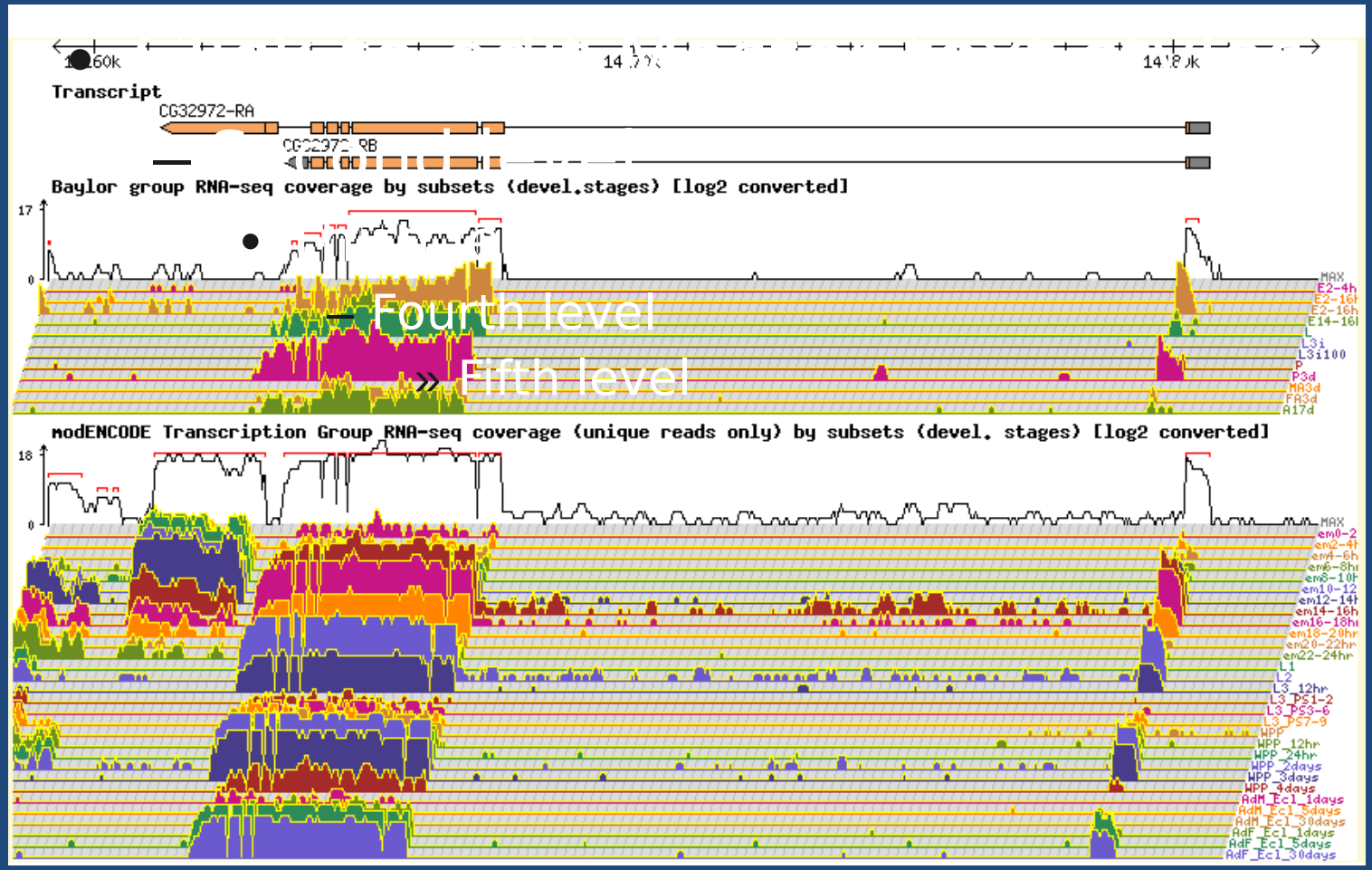


Stranded RNA-Seq data: anti-sense RNAs

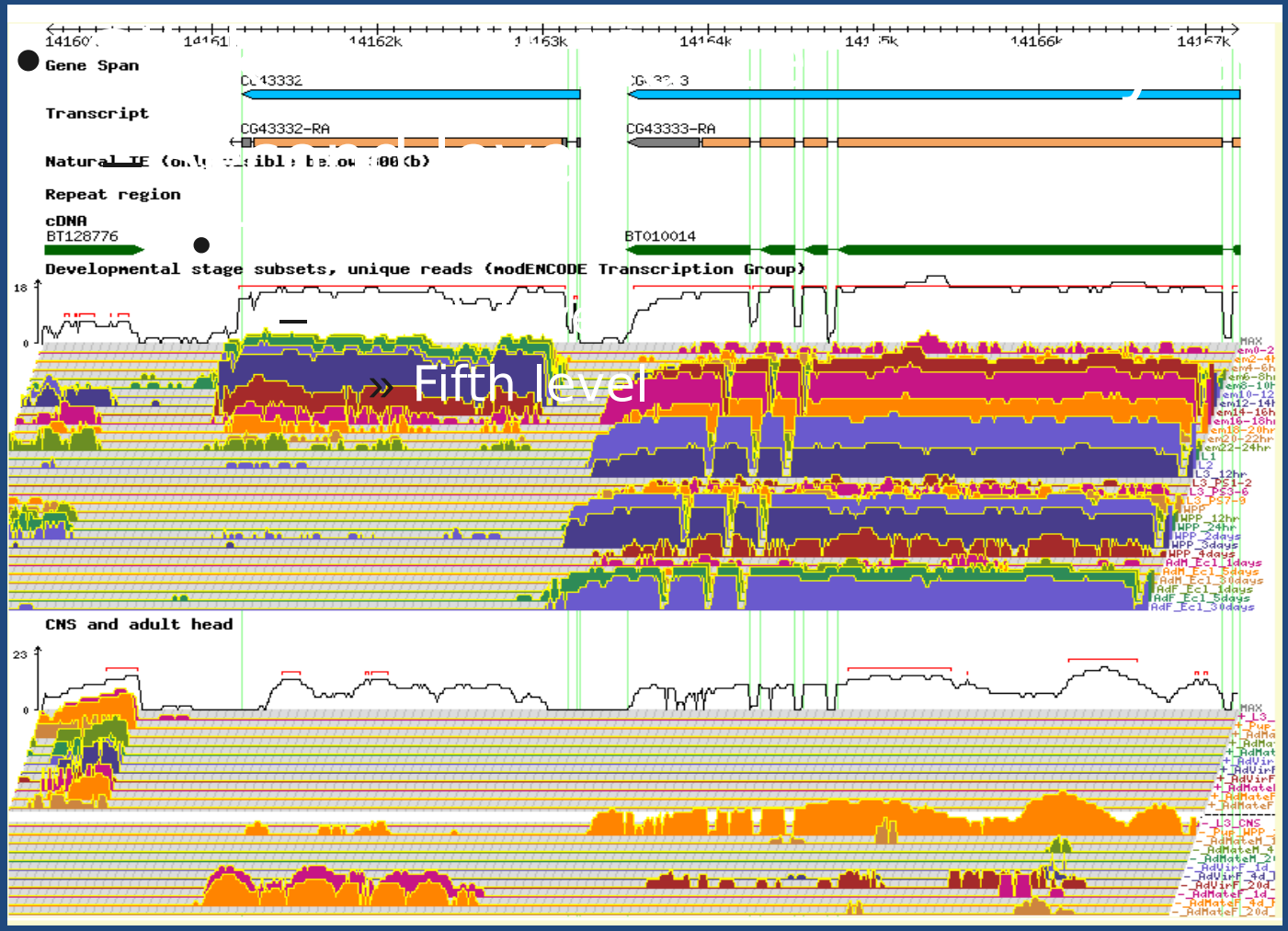
- C[illegible]es
  - [illegible]



Gene split: CG32972

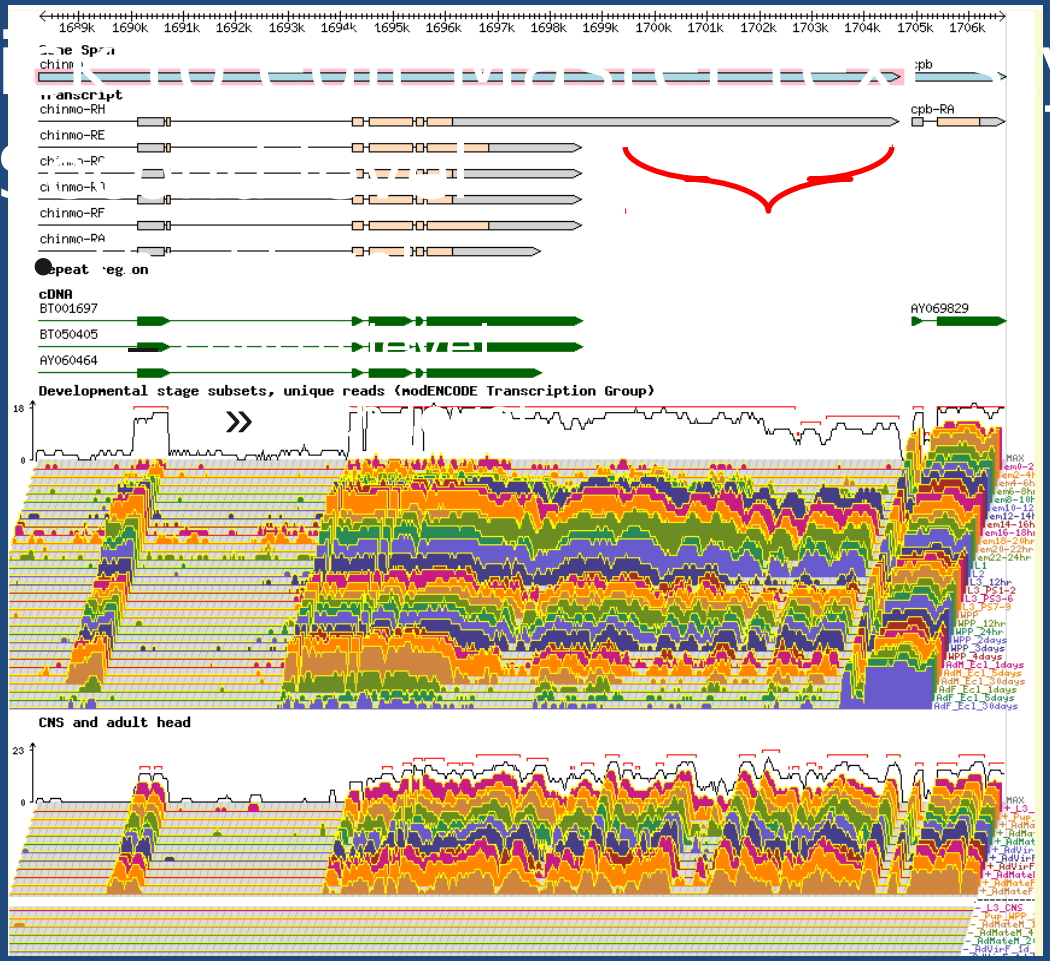


Result of split: CG43332 + CG43333



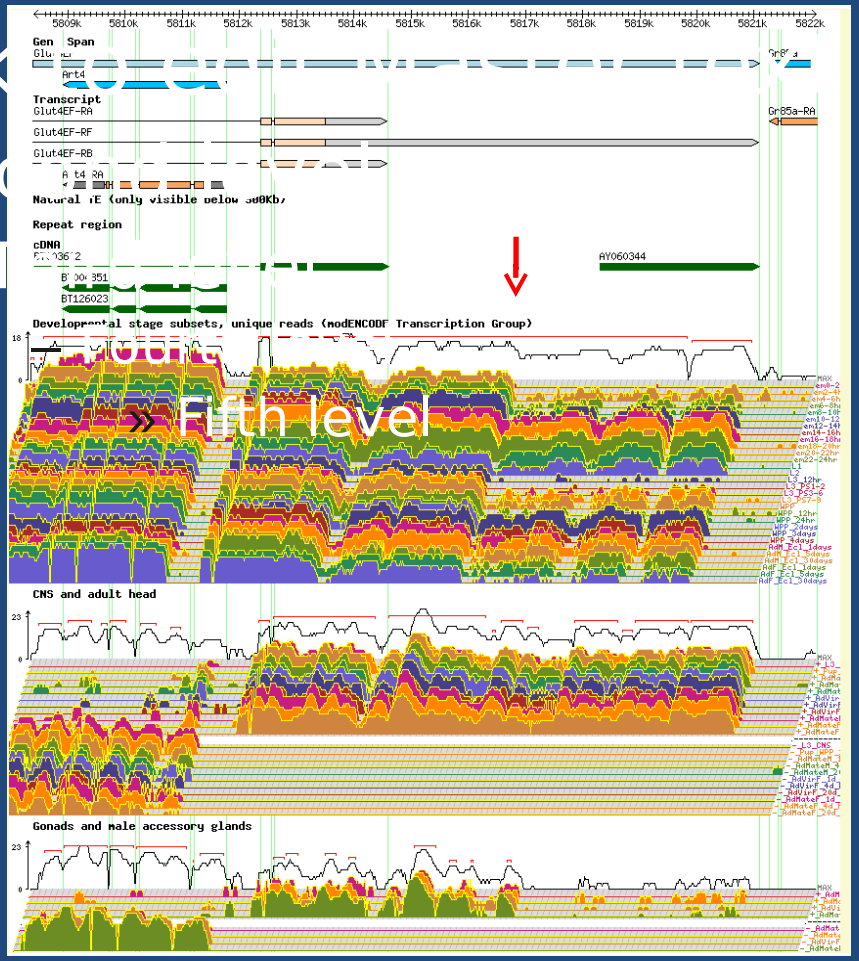
Improved annotations: extended 3’ UTRs

- Cli[illegible]yles
  - [illegible]



Intermediate 3’ UTR extensions not annotated

- Click [illegible] styles
  - Sec[illegible]
    - T[illegible]
      - Fifth level



Gene model comments

- RNA-Seq data support multiple isoforms with extended 3' UTRs of differing length, some of which appear to be stage and/or tissue specific.

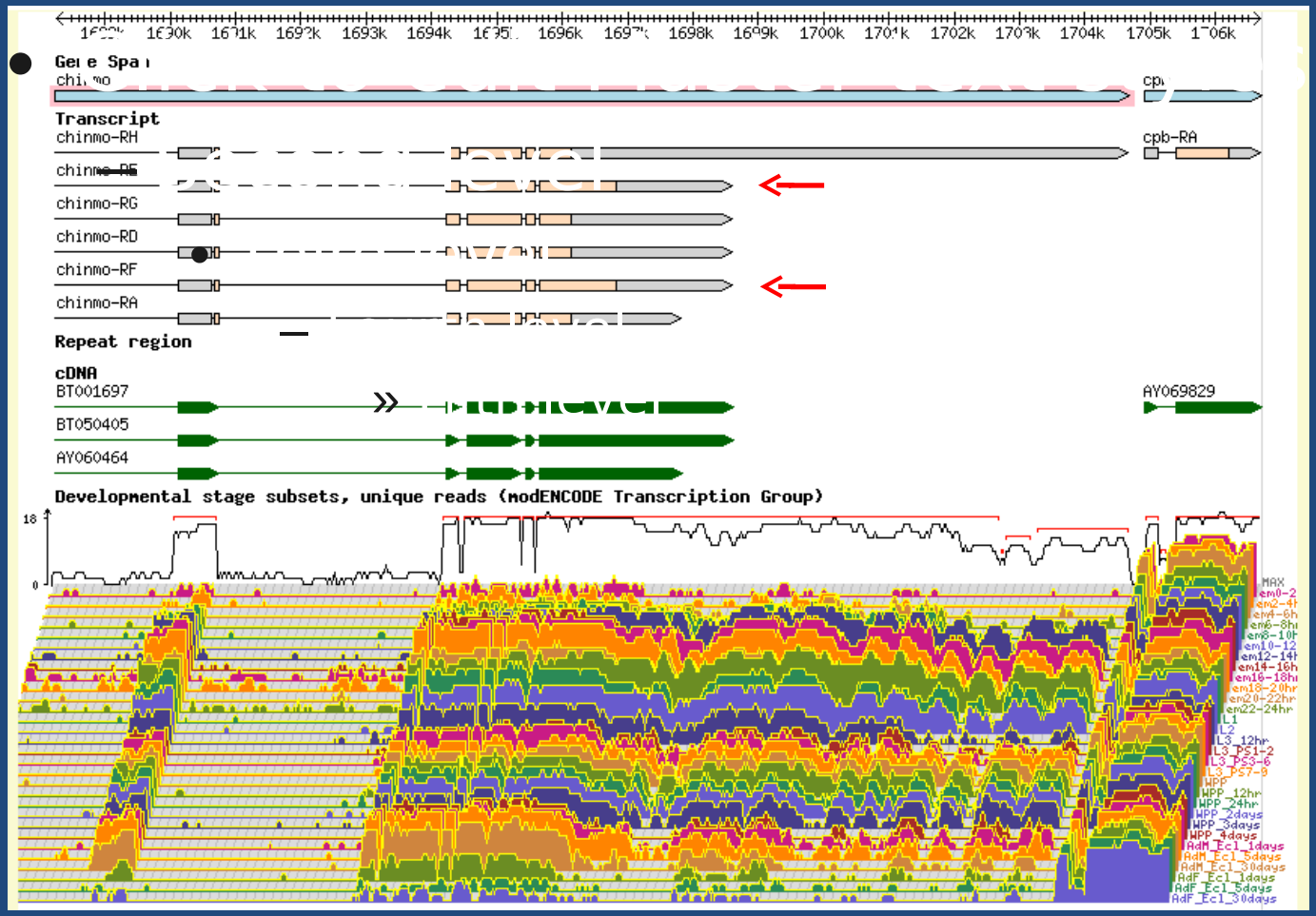
Impact of RNA-Seq data

- New and improved annotations:
  - New genes, coding and non-coding
  - New 5’ exons/promoters
  - New internal exons
  - Extended 3’ UTRs
  - Merges & splits

Impact of RNA-Seq data

- Major source of reannotation flags:
  - Unannotated junctions (complete categorized and prioritized list)
  - Targeted MIP cDNAs
  - RNA-Seq coverage data without corresponding annotation

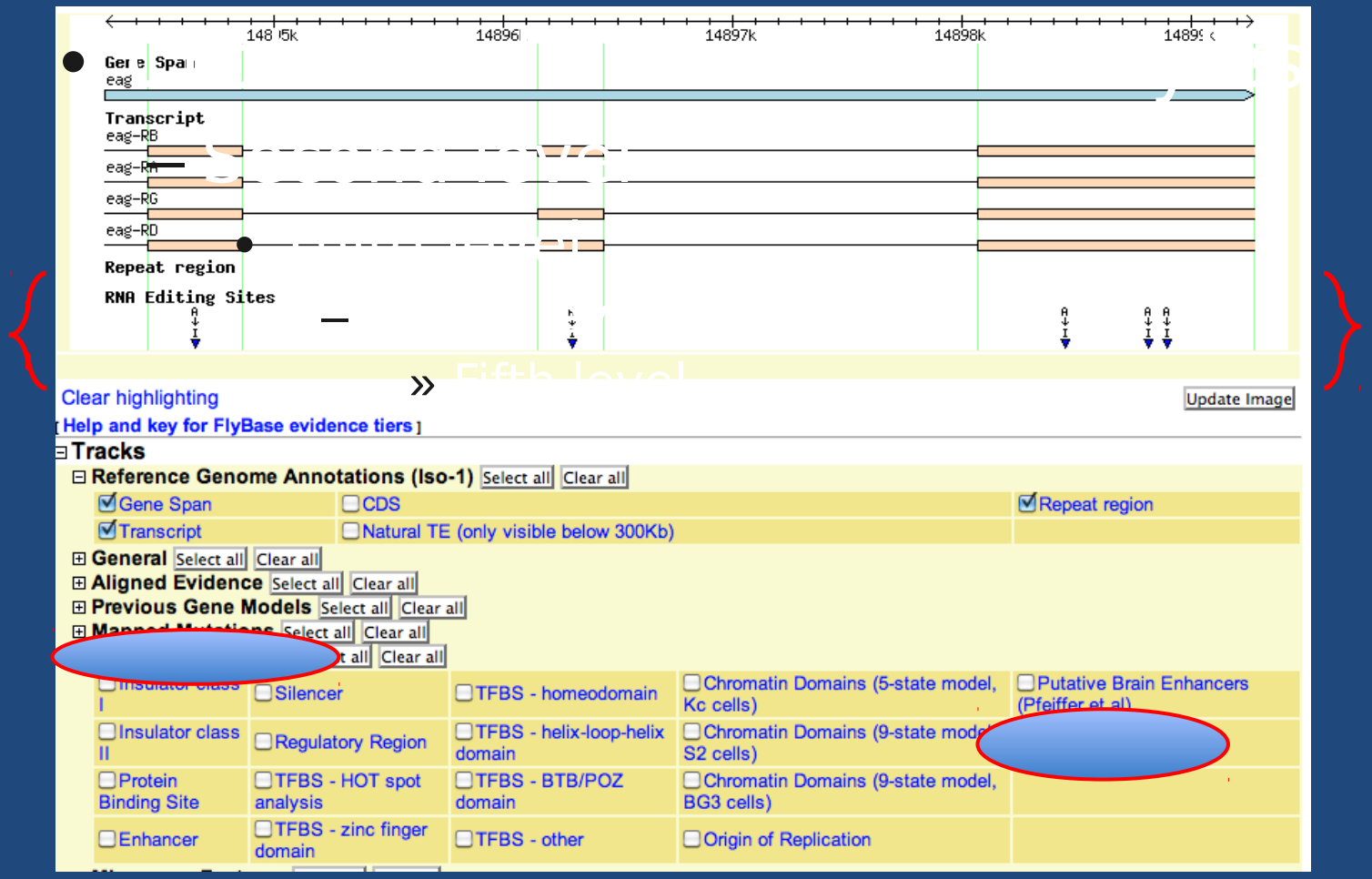
New data: stop-codon readthrough

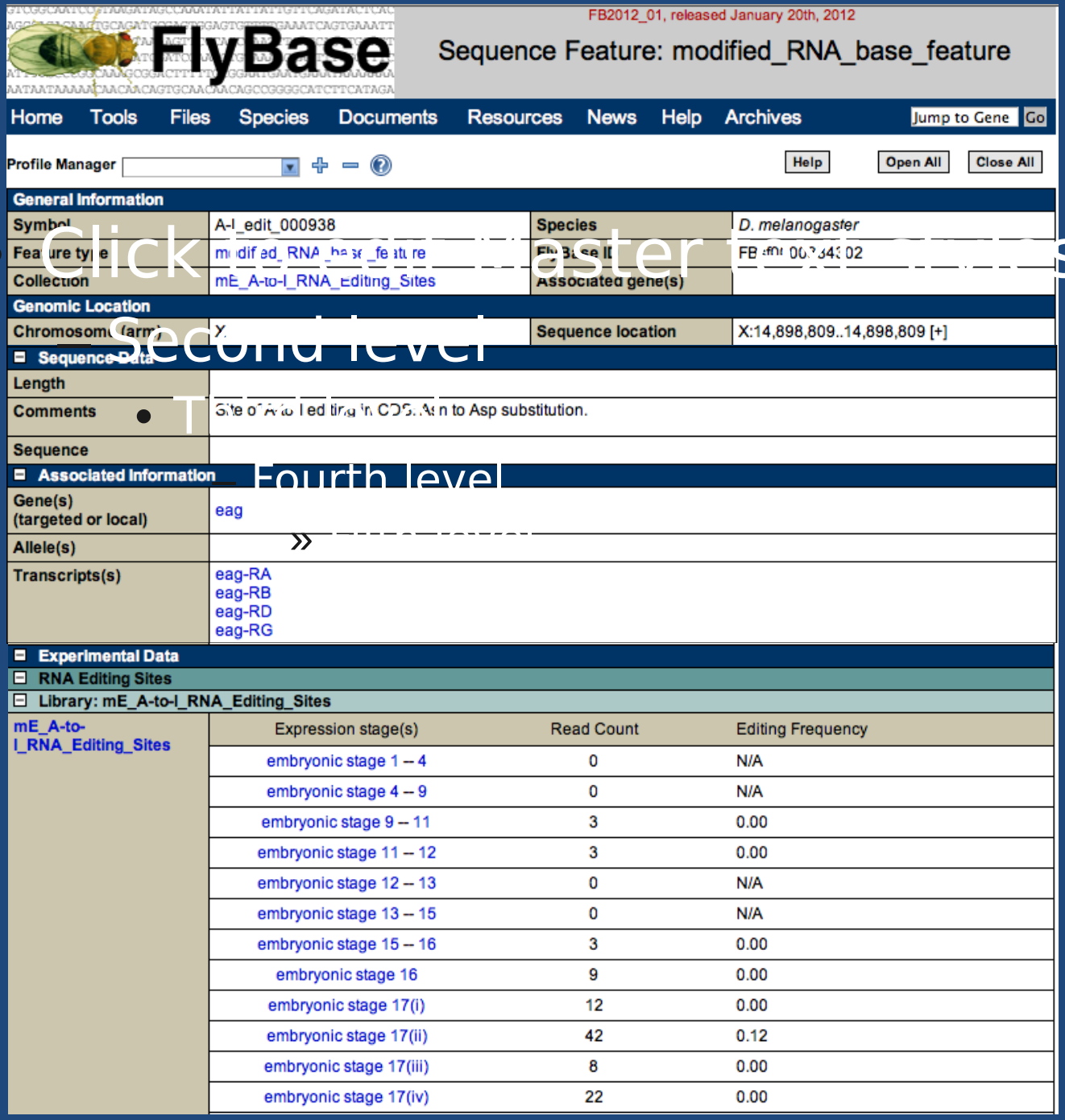


Gene model comments

- gene_with_stop_codon_read_through ; SO:0000697
- Stop-codon suppression (UGA) postulated; FBrf0216884.

Transcript features annotated on the genome: RNA editing sites





FlyBase

FB2012_01, released January 20th, 2012

Sequence Feature: modified_RNA_base_feature

Home Tools Files Species Documents Resources News Help Archives Jump to Gene Go

Profile Manager Help Open All Close All

- Click on Fourth Master Features
- Second level
- Third level
- Fourth level

General Information

Symbol	A-I_edit_000938	Species	*D. melanogaster*
Feature type	modified_RNA_base_feature	FlyBase ID	FBsf0[illegible]00[illegible]34[illegible]02
Collection	mE_A-to-I_RNA_Editing_Sites	Associated gene(s)	

Genomic Location

Chromosome (arm)	X	Sequence location	X:14,898,809..14,898,809 [+]

Sequence Data

Length	
Comments	Site of A-to-I editing in CDS: Asn to Asp substitution.
Sequence	

Associated Information

Gene(s) (targeted or local)	eag
Allele(s)	»
Transcripts(s)	eag-RA eag-RB eag-RD eag-RG

Experimental Data

RNA Editing Sites

Library: mE_A-to-I_RNA_Editing_Sites

mE_A-to-I_RNA_Editing_Sites

Expression stage(s)	Read Count	Editing Frequency
embryonic stage 1 -- 4	0	N/A
embryonic stage 4 -- 9	0	N/A
embryonic stage 9 -- 11	3	0.00
embryonic stage 11 -- 12	3	0.00
embryonic stage 12 -- 13	0	N/A
embryonic stage 13 -- 15	0	N/A
embryonic stage 15 -- 16	3	0.00
embryonic stage 16	9	0.00
embryonic stage 17(i)	12	0.00
embryonic stage 17(ii)	42	0.12
embryonic stage 17(iii)	8	0.00
embryonic stage 17(iv)	22	0.00

modENCODE TSS: FlyBase GBrowse

Instructions [Help] [Reset] [Bookmark this]

Search using a sequence name or range, gene symbol (but not full name, for example, ct but not cut), gene identifier, insertion symbol, insertion identifier, or other landmark. Searches are case-sensitive. Wildcards are not supported at this time. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: cnn, FBgn0000490, X:200000..220000, 2L:80,000..100,000, 2R:80,000..100,000, 3L:80,000..100,000, 3R:80,000..100,000, 4:100000..120000.

Search

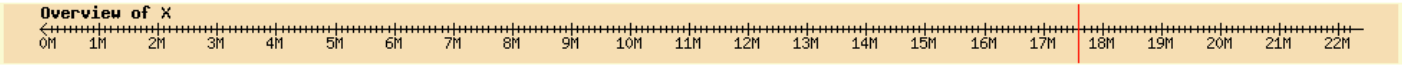
Landmark or Region: X:17596976..17597975 Search

Data Source ' D. melanogaster RNA-seq Data

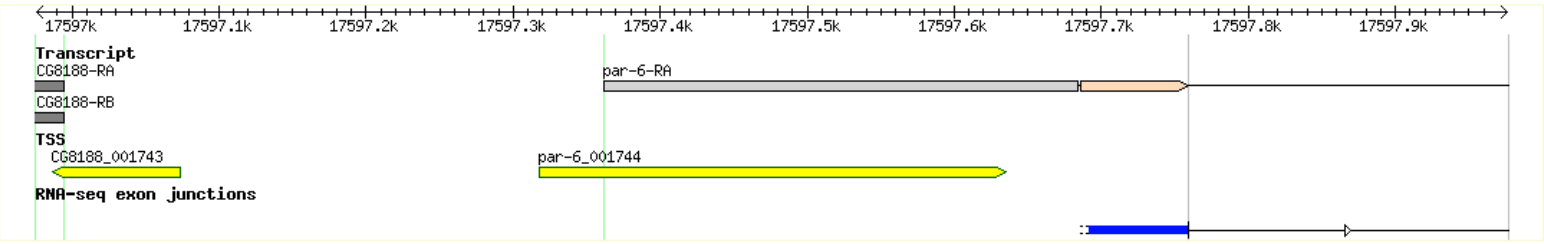
Report & Analysis tools: Download Decorated FASTA File Configure... Go

Scroll/Zoom: Show 1 kbp Flip

Overview



Details



Clear highlighting Update Image

[Help and key for Stock Center Data evidence tiers]

Tracks

Chip-seq data Select all Clear all

- Chromatin Domains (5-state model, Kc cells)
- Chromatin Domains (9-state model, S2 cells)
- Chromatin Domains (9-state model, BG3 cells)

Gene Model features Select all Clear all

- Gene Span
- Transcript

Genome reagents and data Select all Clear all

- aligned EST
- aligned cDNA
- aligned mRNA

Genomic features Select all Clear all

- Natural transposon
- Transgene insertion site
- Estimated Cytological band

Mapped features Select all Clear all

- Insulator class I
- Insulator class II

Other RNA-seq data Select all Clear all

- cell lines expression (by strand) [log2 converted]
- expression under treatments (by strand) [log2 converted]

RNA-seq data Select all Clear all

- RNA-seq exon junctions
- modENCODE Transcription Group RNA-seq coverage (unique reads only) by subsets (devel. stages) [log2 converted]
- modENCODE Transcription Group RNA-seq coverage (embryonic stages, by strand) [log2 converted]
- K. White group RNA-seq coverage by subsets (devel.stages) [log2 converted]
- Baylor group RNA-seq coverage by subsets (devel.stages) [log2 converted]
- modENCODE Transcription Group RNA-seq coverage (all reads) by subsets (devel. stages) [log2 converted]
- modENCODE Transcription Group RNA-seq coverage by subsets (tissue culture cells) [log2 converted]

RNA-seq data, expression by tissue Select all Clear all

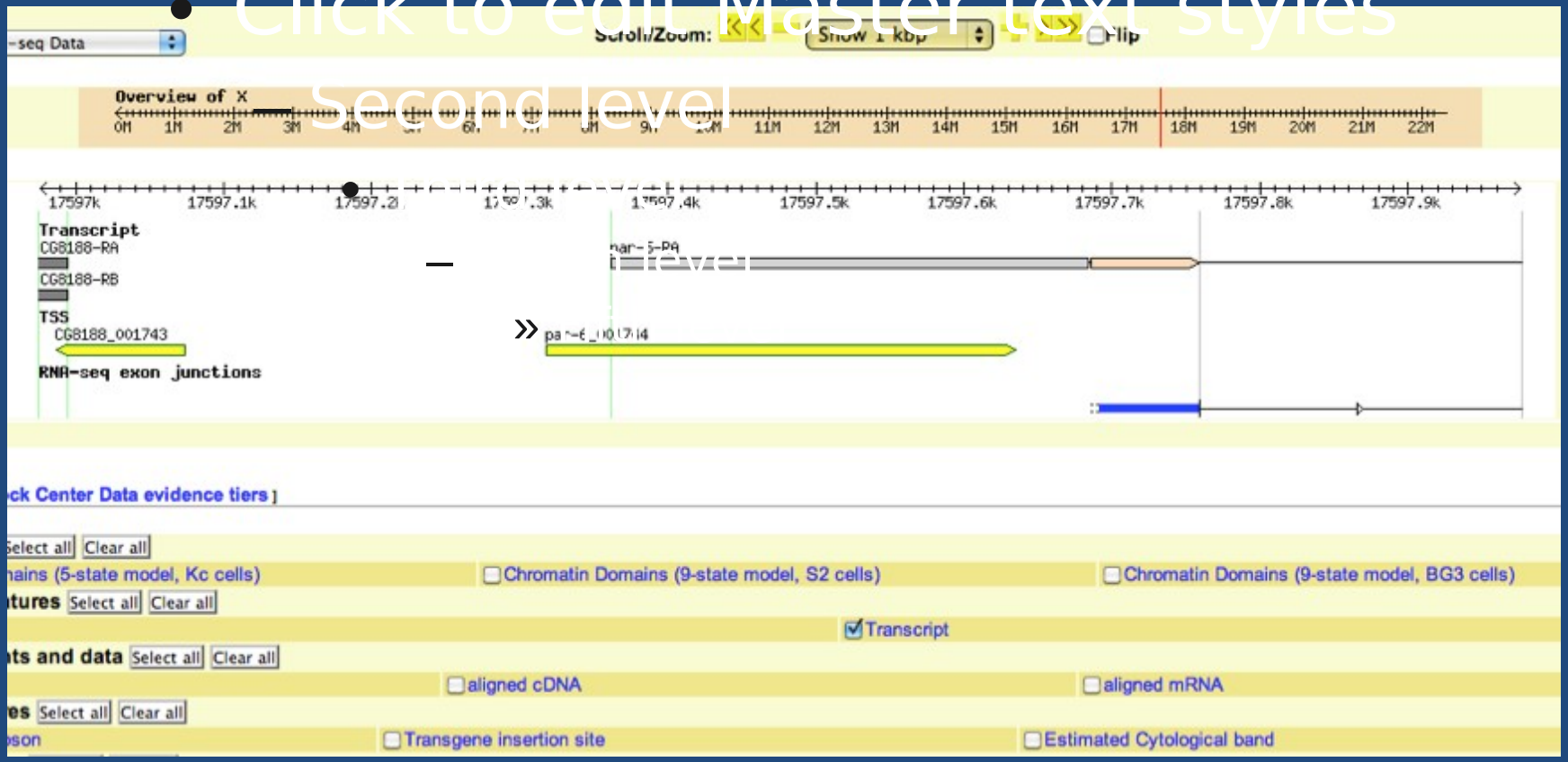
- digestive system (by strand) [log2 converted]
- imaginal disc and other carcass (by strand) [log2 converted]
- gonads and male accessory glands (by strand) [log2 converted]
- fat body and salivary glands (by strand) [log2 converted]
- CNS and adult head (by strand) [log2 converted]

External Annotation Tracks Select all Clear all

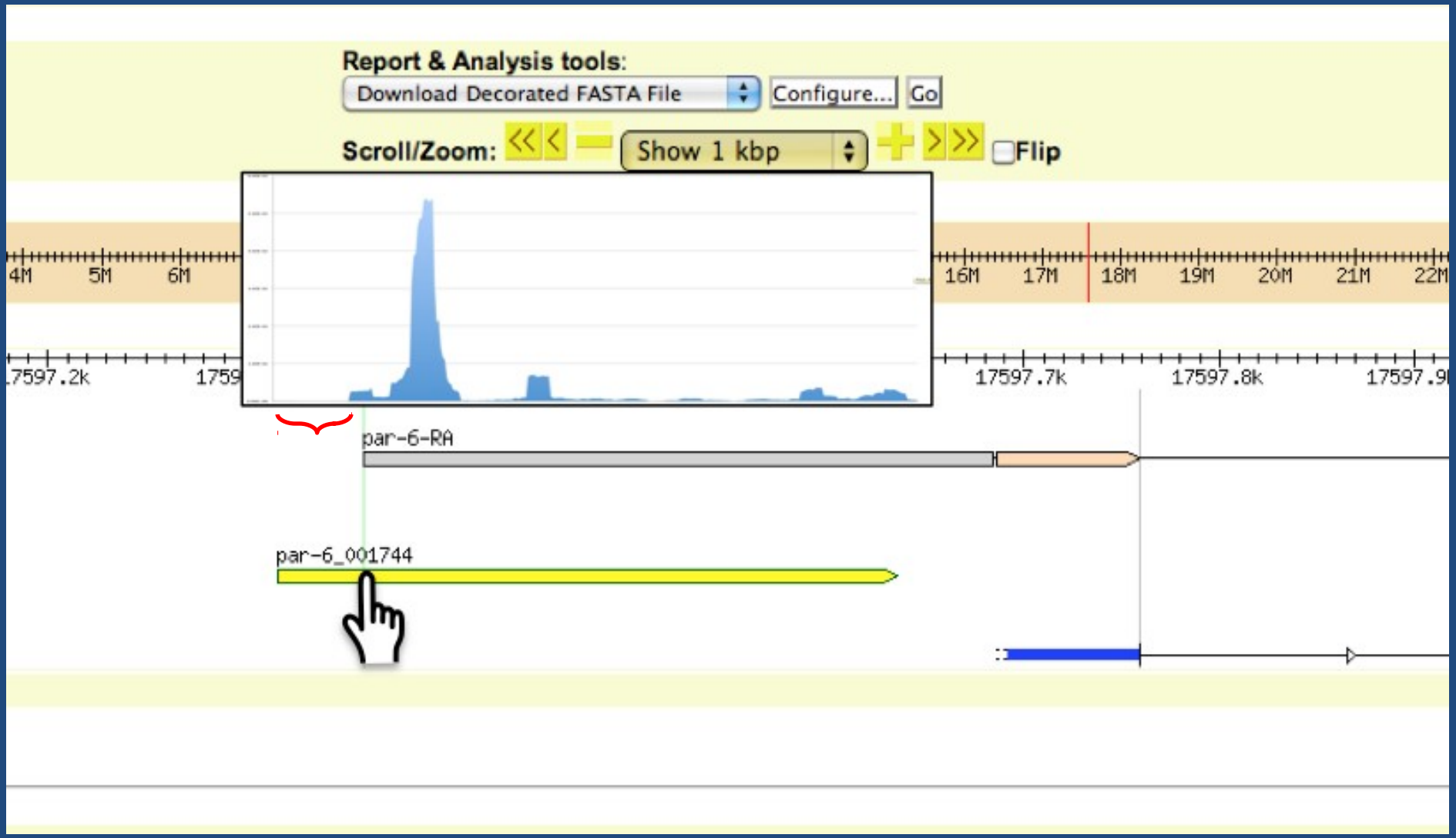
Configure tracks... Update Image

Transcription start sites
range represented as extent on genome

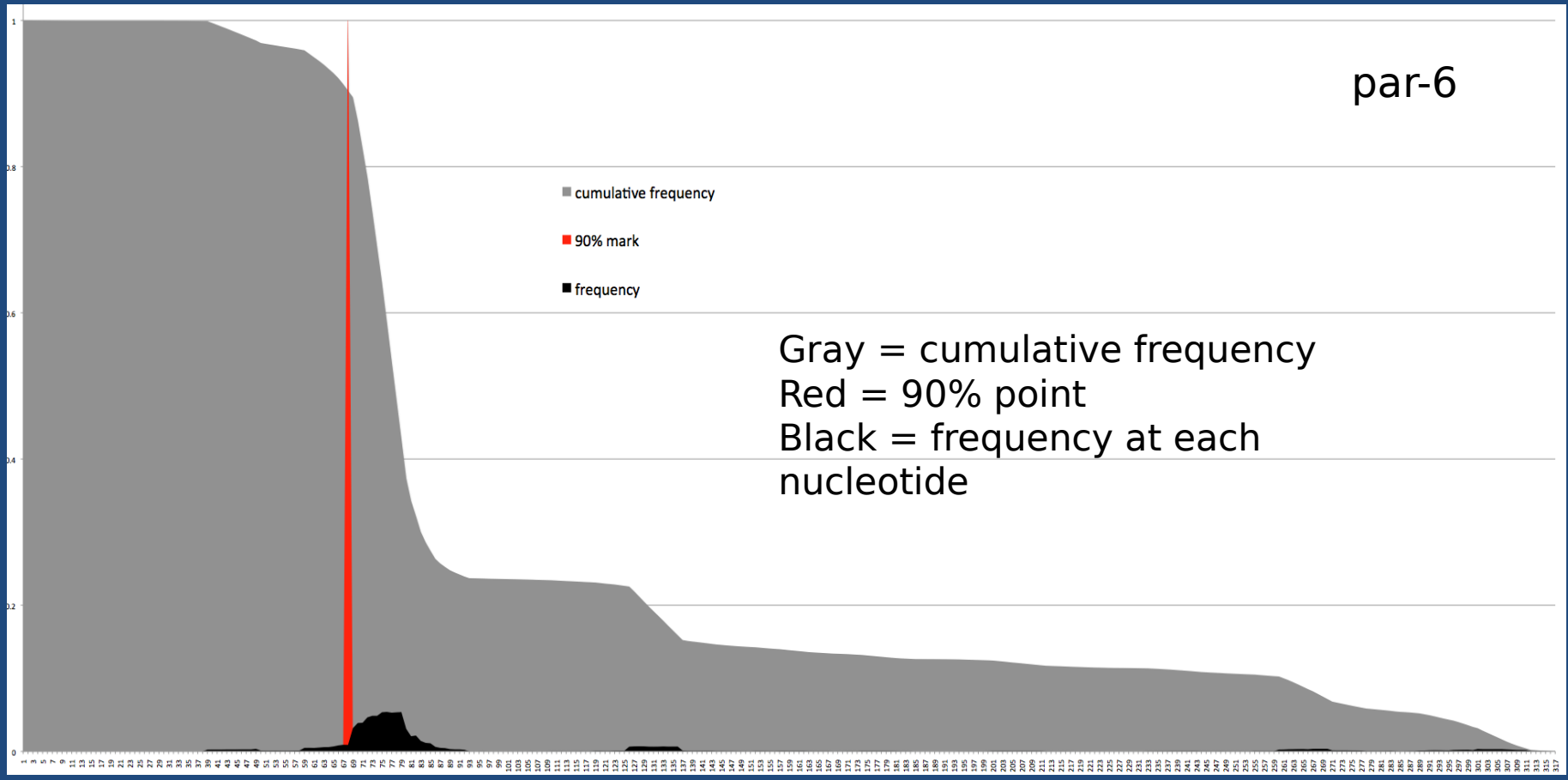
- Click to edit Master text styles
  - Second level



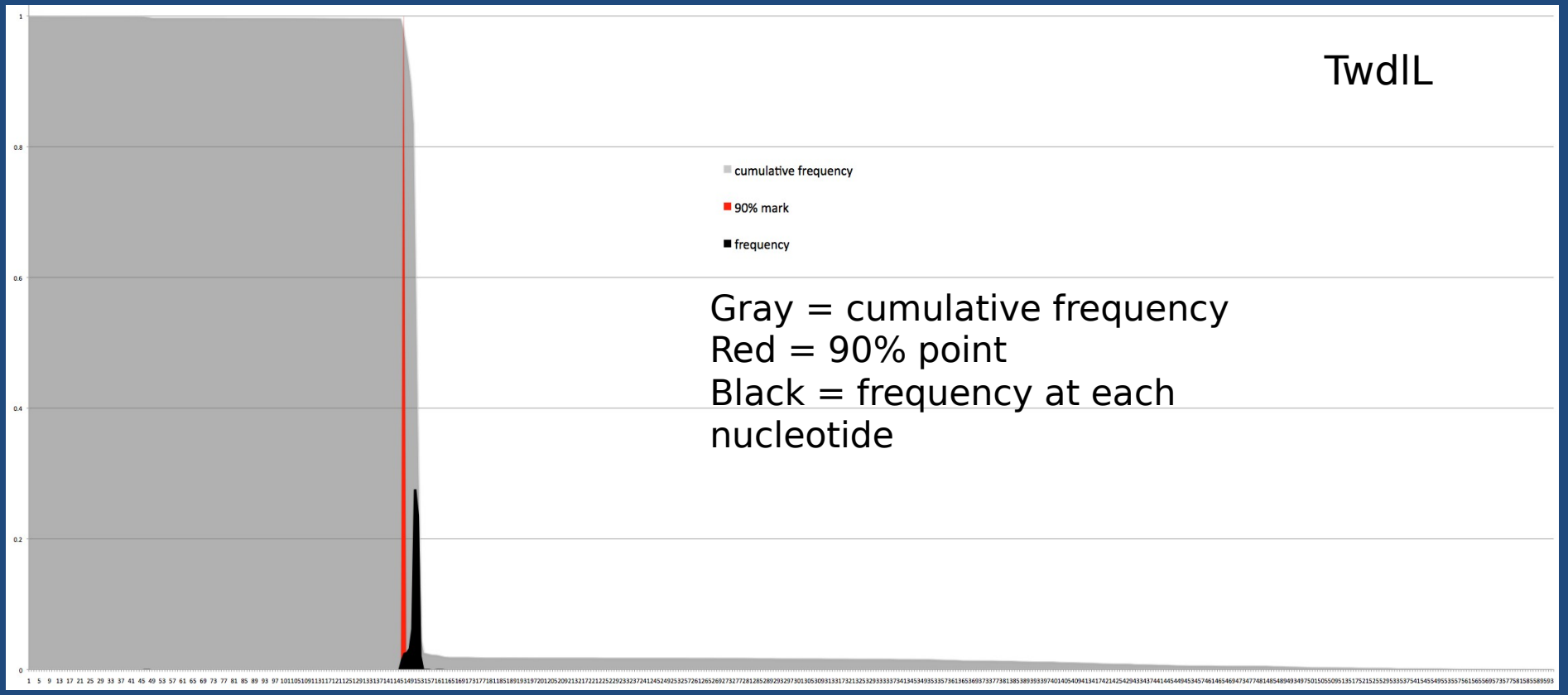
TSS data: frequency distributions



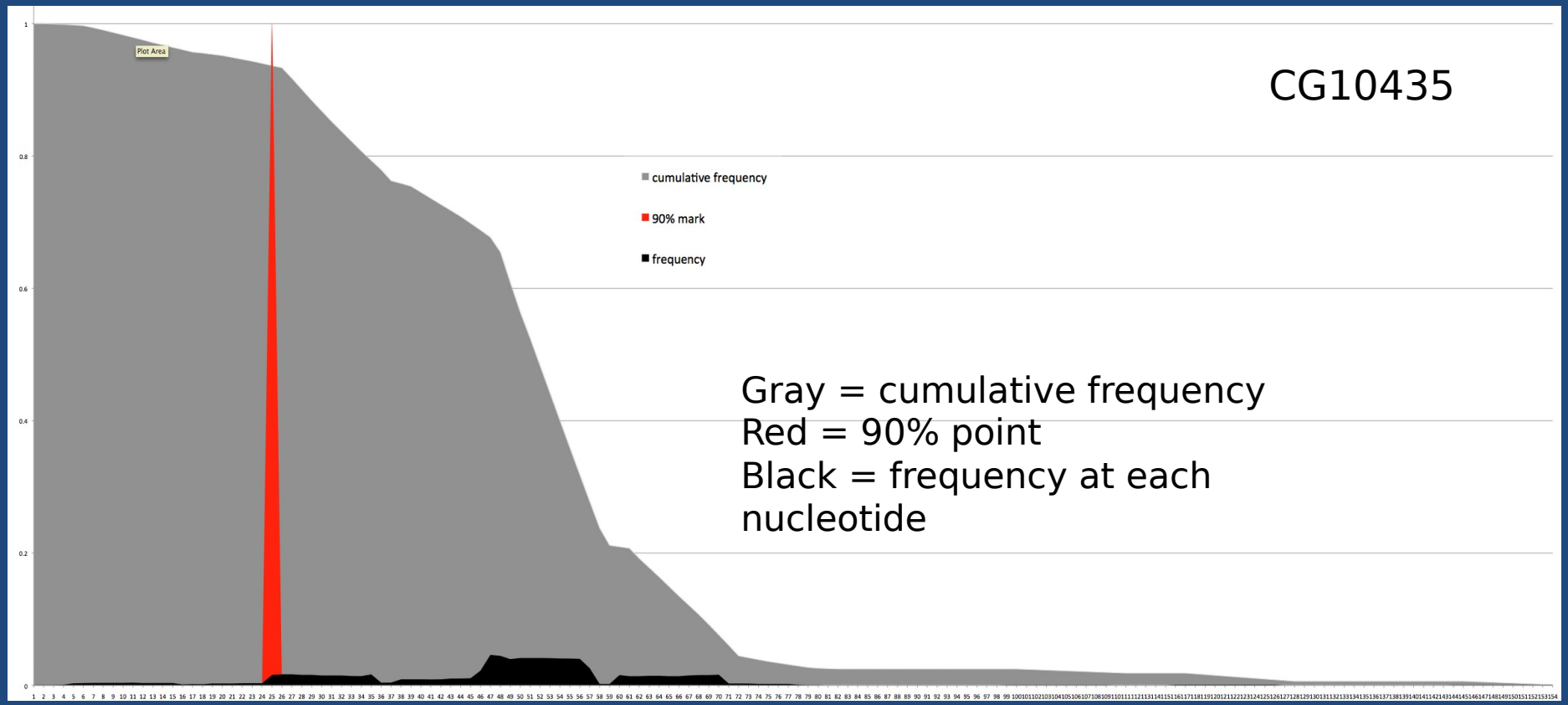
TSS annotation at 90% point



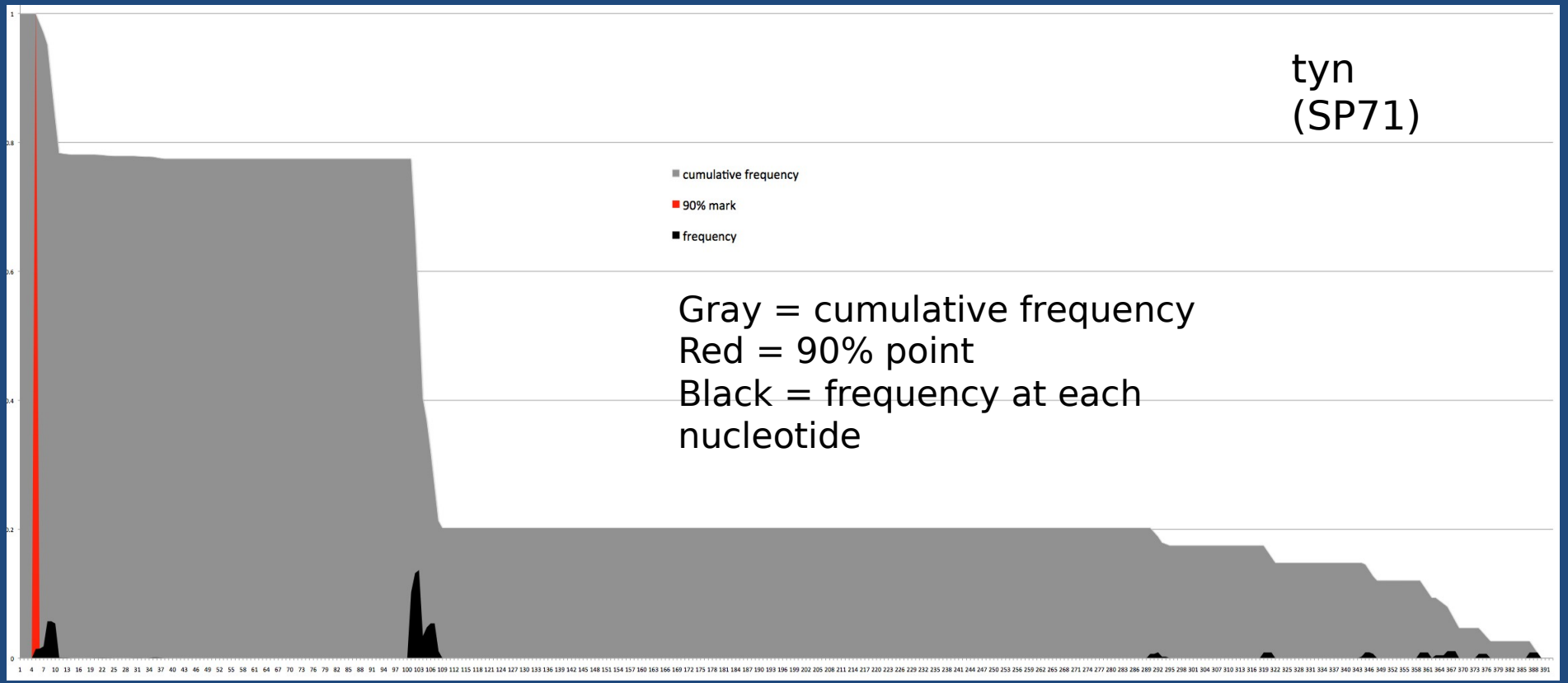
TSS annotation at 90% point peaked promoter class



TSS annotation at 90% point broad promoter class



TSS annotation at 90% point broad promoter class



- Annotation on the genome of transcript features
  - RNA editing sites
  - Transcription start sites (as regions)
- Annotated transcripts will reflect TSS 90% point

New high-throughput data

- RNA-Seq junction calls
- Stranded RNA-Seq coverage data from multiple tissues (cell lines, treatments)
- Stop-codon readthrough predictions
- A-to-I RNA editing sites
- Transcription start site (TSS) mapping

Integration of High Throughput Data

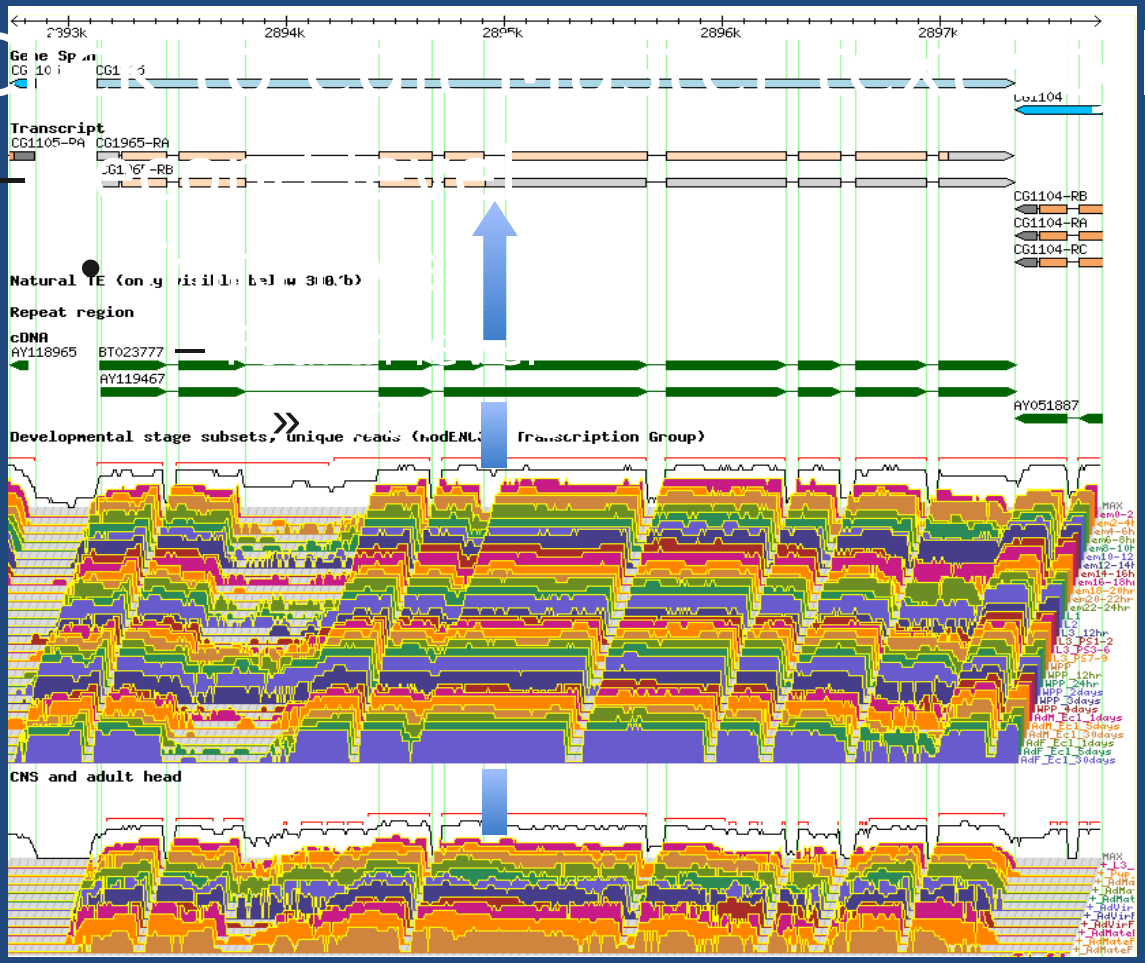
- Presentations in GBrowse
- Sequence feature reports
- New and improved annotations
- Assessing the data
- Annotation on the genome of transcript features
- Use of gene model comments

modENCODE papers

- RNA-Seq (coverage, junctions, 3’ UTRs)
  - Graveley, et al. (2011) Nature 471:473-479
  - Cherbas, et al. (2011) Genome Res. 21:301-314
  - Smibert, et al. (2012) Cell Reports 23 Feb. 2012
- Stop-codon readthrough
  - Jungreis, et al. (2011) Genome Res. 21:2096-2113
- A-to-I RNA editing
  - Graveley, et al. (2011) Nature 471:473-479
- Transcription start site data
  - Hoskins, et al. (2011) Genome Res. 21:182-192

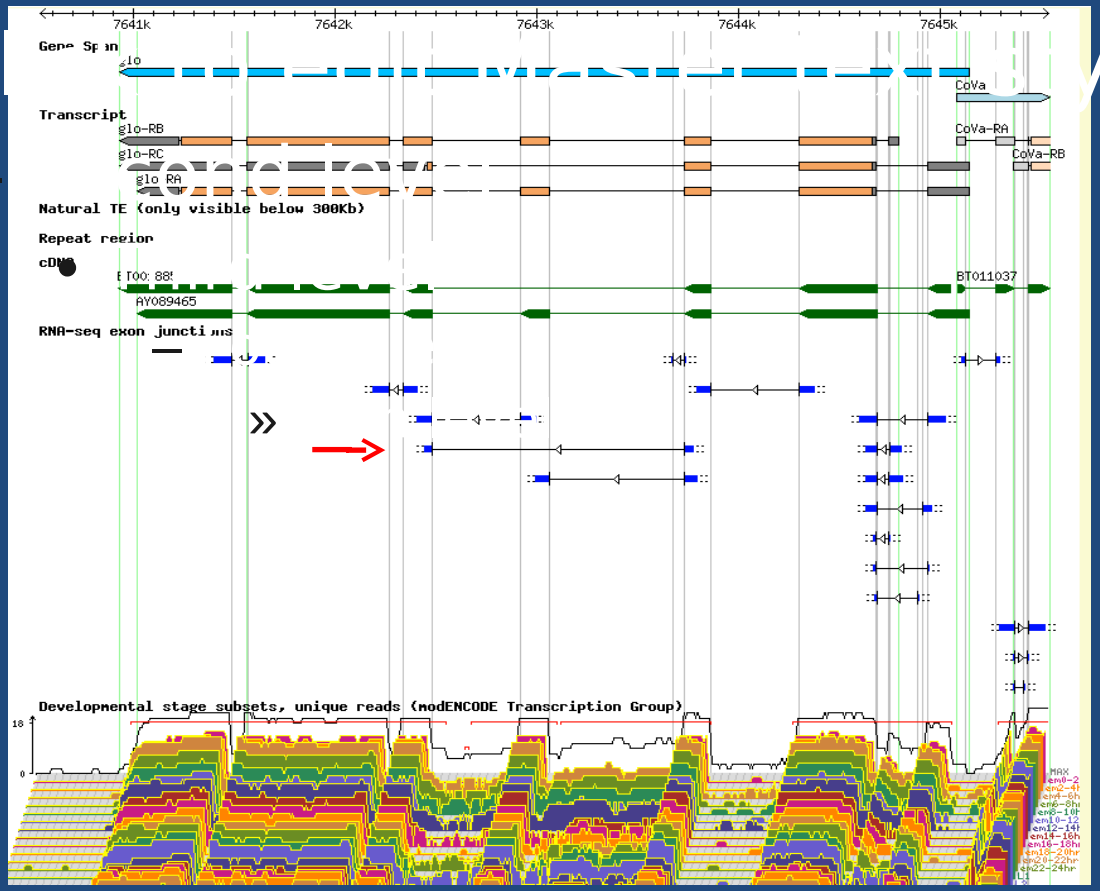
Transcripts that result in truncated polypeptides: retained introns

- C[illegible]les



Transcripts that result in truncated polypeptides: alternative splices

- C[illegible]yles
  - [illegible]



Annotated transcripts with truncated polypeptides (520 transcripts, 413 genes)

- Retained intron ⎓ premature stop
- Retained intron ⎓ downstream start
- Alternative splice ⎓ premature stop
- Alternative splice ⎓ downstream start

Comments (associated with transcript)

- Evidence supports alternative splice leading to premature stop codon and/or downstream start; may or may not produce functional polypeptide.
- Based on cDNA(s) with retained intron; results in premature stop codon and/or downstream start; may or may not produce functional polypeptide.

A few cases addressed in the literature

- Truncated polypeptides appear to be functional – can annotate properly
- The variants thought to be non-coding transcripts – cannot annotate properly
  - Change: create non-coding transcripts within a coding gene model
  - How do we handle the rest??

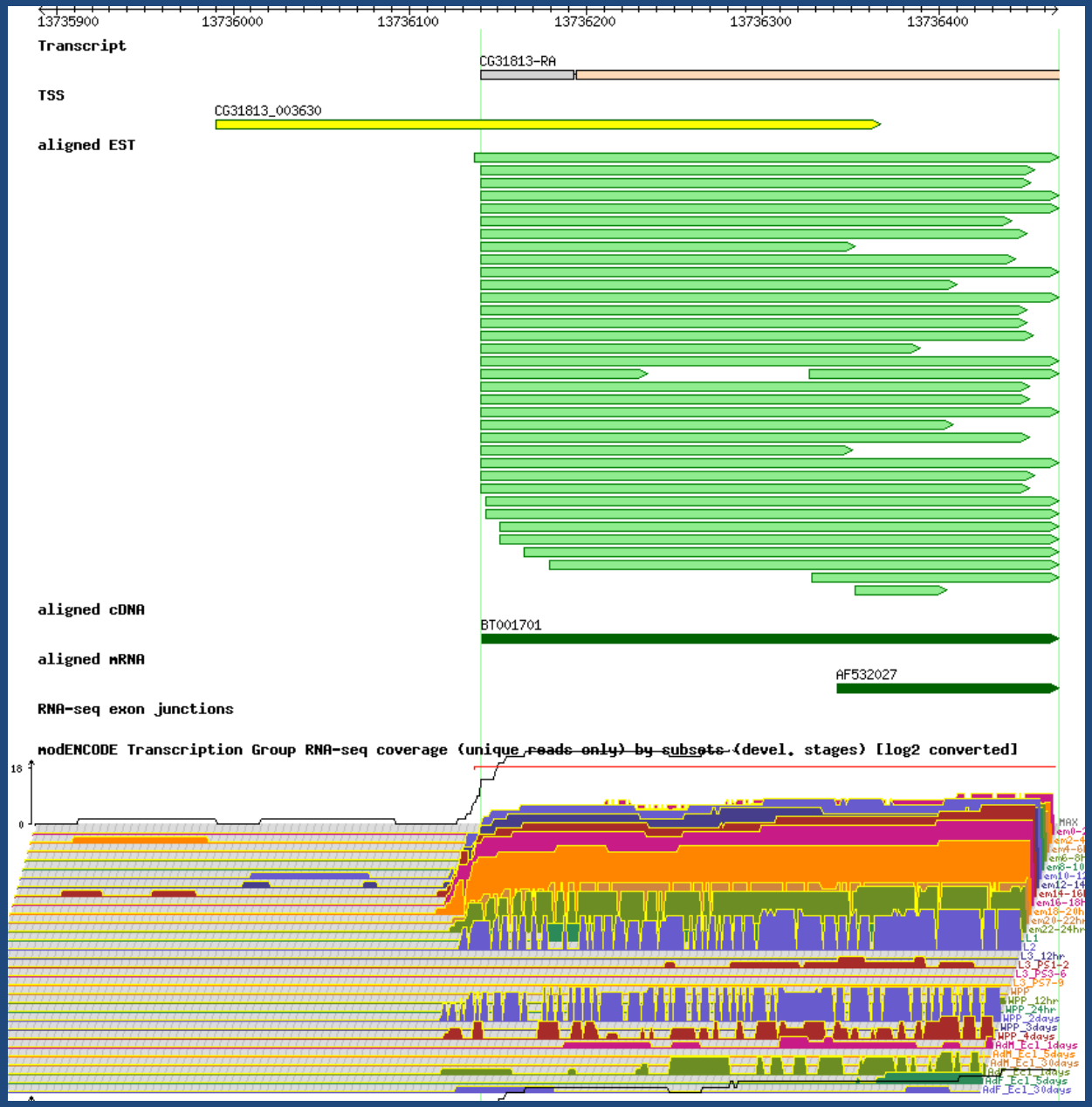
Some genes involved in RNA-splicing with annotated truncated isoforms

- tra
- su(w[a])
- Sxl
- B52
- Moca-cyp
- Rm62
- sqd
- CG1965
- CG1622
- CG1646
- Ars2
- Pep
- SC35
- scaf6
- U2af50
- Zn72D
- CG11266
- CG5316
- CG6686
- CG7757

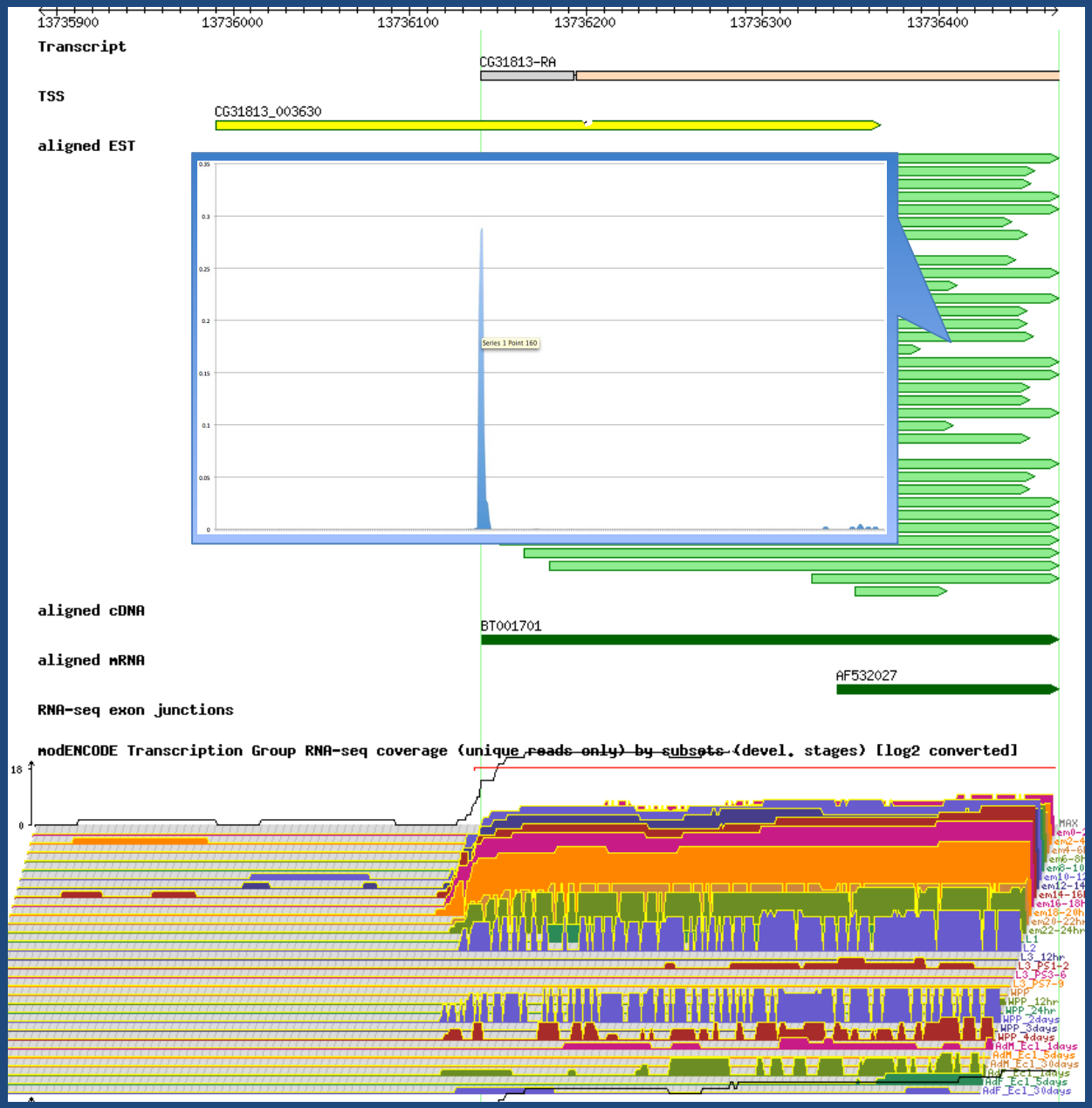
Short polypeptides supported in literature

- wah
- ade3
- svr
- CG34422
- psq
- scf
- AGO2
- osk

Gene annotation using modENCODE TSS



Gene annotation using modENCODE TSS



modENCODE TSS: FlyBase GBrowse

Instructions [Help] [Reset] [Bookmark this]

Search using a sequence name or range, gene symbol (but not full name, for example, ct but not cut), gene identifier, insertion symbol, insertion identifier, or other landmark. Searches are case-sensitive. Wildcards are not supported at this time. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: cnn, FBgn0000490, X:200000..220000, 2L:80,000..100,000, 2R:80,000..100,000, 3L:80,000..100,000, 3R:80,000..100,000, 4:100000..120000.

Search

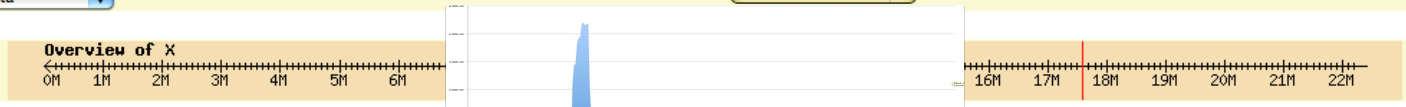
Landmark or Region: X:17596976..17597975 Search

Data Source: ' D. melanogaster RNA-seq Data

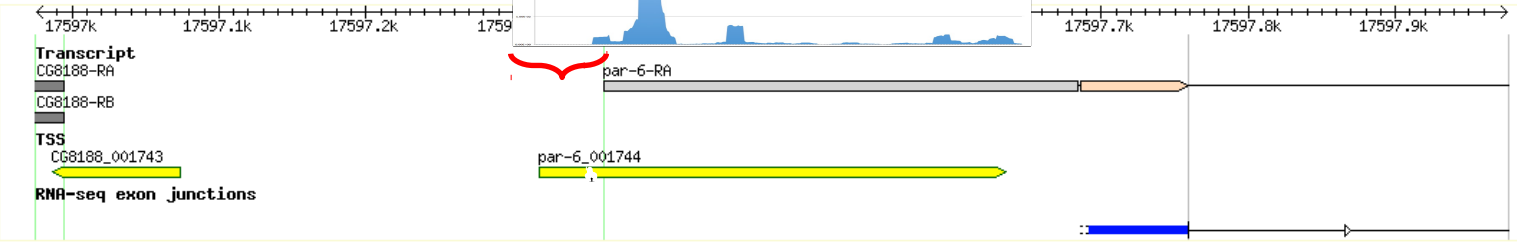
Report & Analysis tools: Download Decorated FASTA File Configure... Go

Scroll/Zoom: Show 1 kbp Flip

Overview



Details



Clear highlighting Update Image

[Help and key for Stock Center Data evidence tiers]

Tracks

Chip-seq data Select all Clear all

- Chromatin Domains (5-state model, Kc cells)
- Chromatin Domains (9-state model, S2 cells)
- Chromatin Domains (9-state model, BG3 cells)

Gene Model features Select all Clear all

- Gene Span
- Transcript

Genome reagents and data Select all Clear all

- aligned EST
- aligned cDNA
- aligned mRNA

Genomic features Select all Clear all

- Natural transposon
- Transgene insertion site
- Estimated Cytological band

Mapped features Select all Clear all

- Insulator class I
- Insulator class II

Other RNA-seq data Select all Clear all

- cell lines expression (by strand) [log2 converted]
- expression under treatments (by strand) [log2 converted]

RNA-seq data Select all Clear all

- RNA-seq exon junctions
- modENCODE Transcription Group RNA-seq coverage (unique reads only) by subsets (devel. stages) [log2 converted]
- modENCODE Transcription Group RNA-seq coverage (embryonic stages, by strand) [log2 converted]
- K. White group RNA-seq coverage by subsets (devel.stages) [log2 converted]
- Baylor group RNA-seq coverage by subsets (devel.stages) [log2 converted]
- modENCODE Transcription Group RNA-seq coverage (all reads) by subsets (devel. stages) [log2 converted]
- modENCODE Transcription Group RNA-seq coverage by subsets (tissue culture cells) [log2 converted]

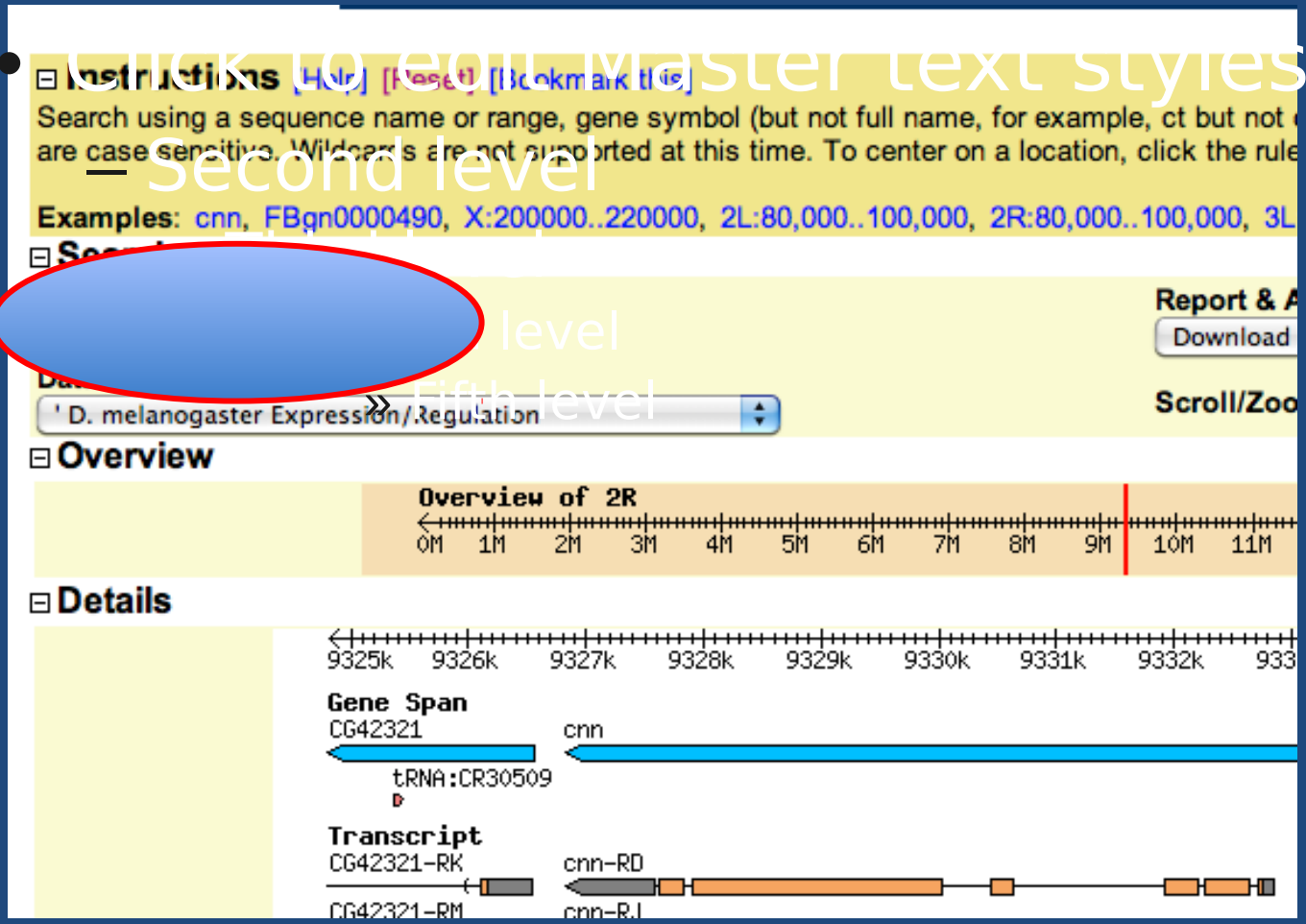
RNA-seq data, expression by tissue Select all Clear all

- digestive system (by strand) [log2 converted]
- imaginal disc and other carcass (by strand) [log2 converted]
- gonads and male accessory glands (by strand) [log2 converted]
- fat body and salivary glands (by strand) [log2 converted]
- CNS and adult head (by strand) [log2 converted]

External Annotation Tracks Select all Clear all

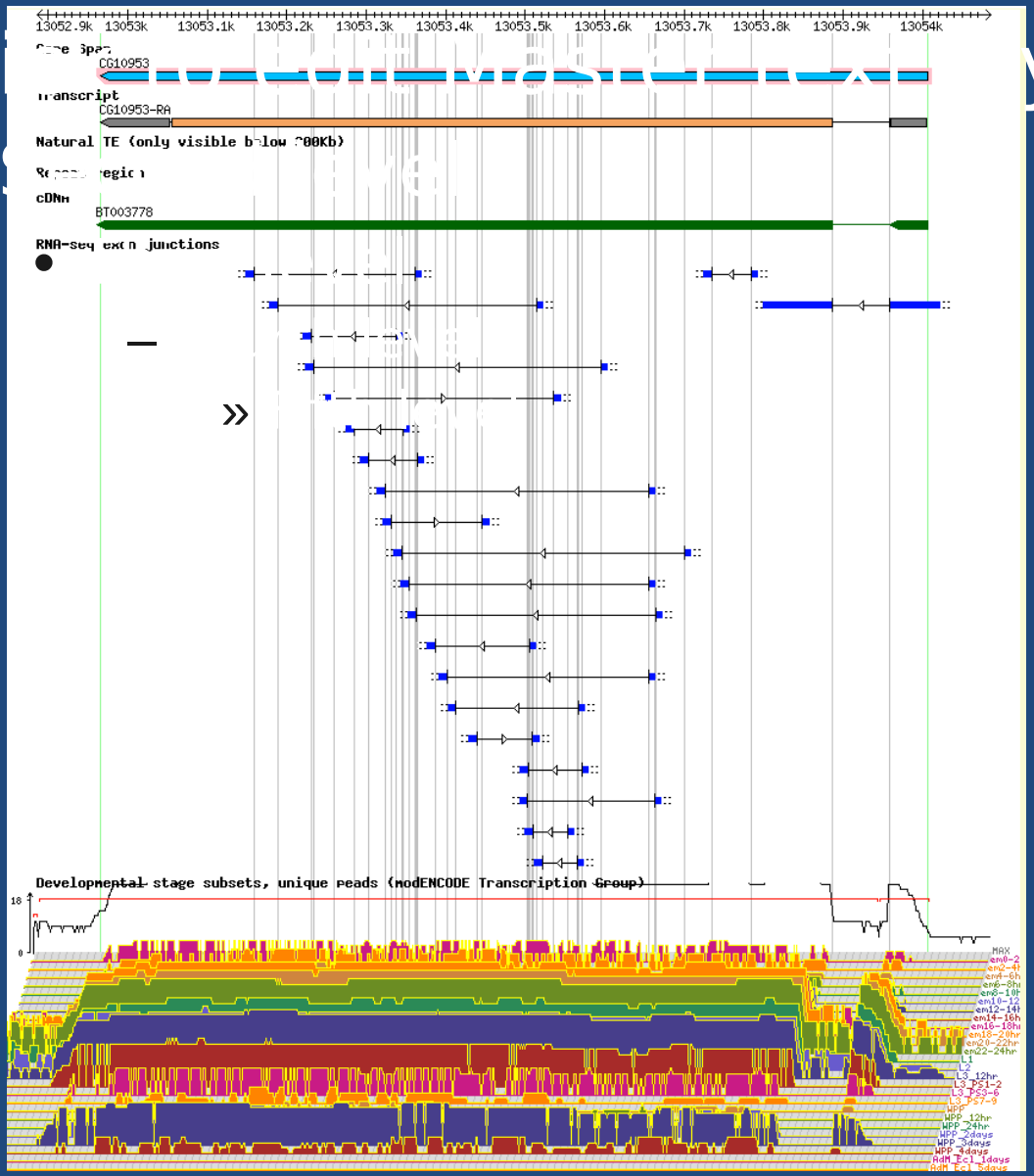
Configure tracks... Update Image

GBrowse: navigation options



- Click to edit Master text styles
  - Second level
    - Third level
      - Fifth level

Cl… yles

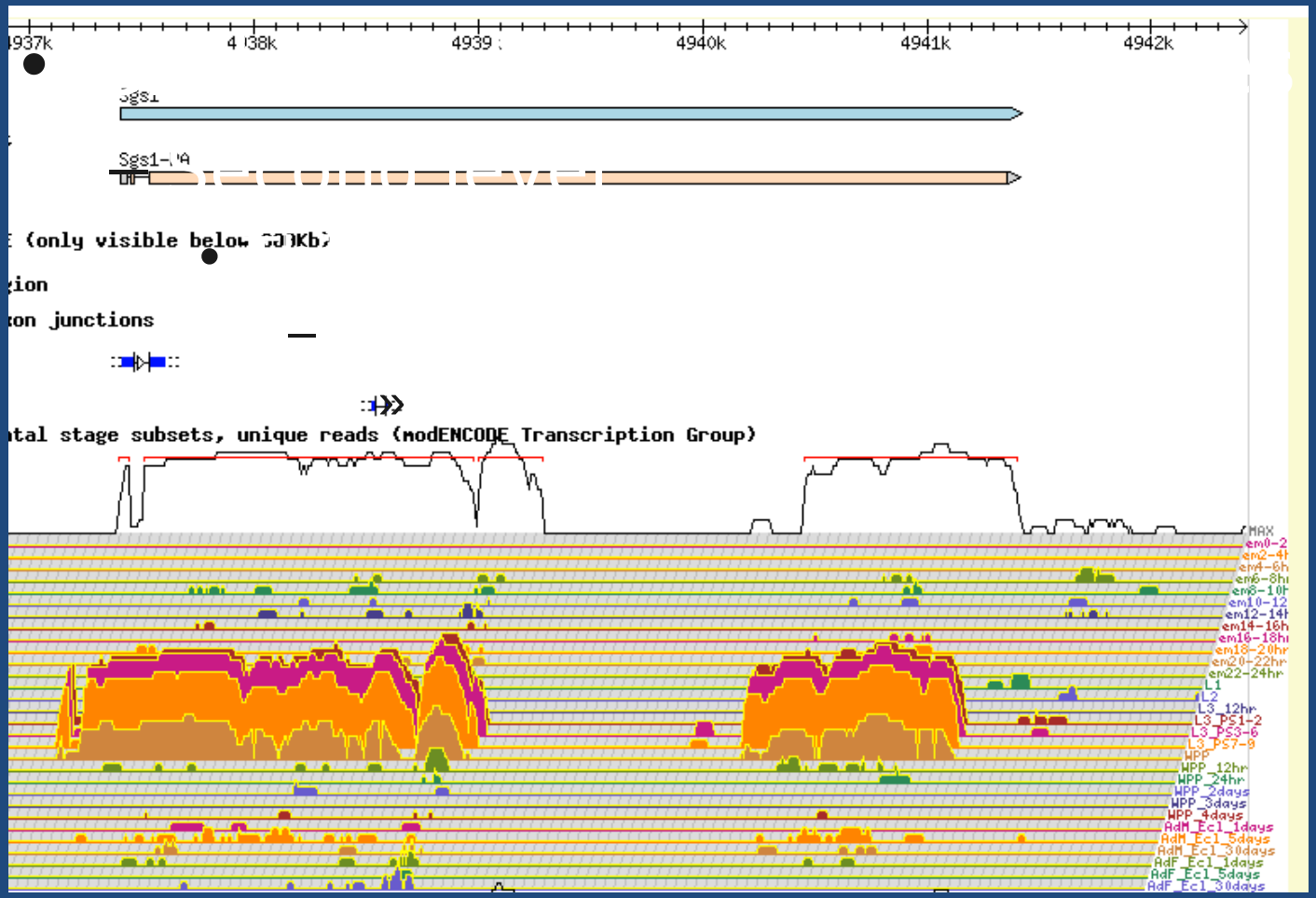


Spurious junctions in region containing repeats

AGAAGCTACTCGATCGCTCTTGGAGTAACAGATTGCAGCTGTCATCATGA
AATTCCTGGTTGTTGCCTTTGCCGTTCTGGCCTGCGCCTATGGTGATGTC
TCCCATTTGGGTTACGACTACTCTCCTCCGGCCCCAGCTCCGGTTTACCA
GCCCGCCCCGGCTCCGGTCTACCAGCCCGCCCCCGCTCCGGTCTACCAGC
CTGCTCCAGCACCAGTTGTGATCCCTGCTCCTGCTCCAGCTCCAGTTGTG
ATCCCTGCTCCTGCTCCCGCTCCAGTTGTGATTCCGGCTCCAGCTCCCGT
GAAGACGTACGTCCCACCCGCACCCATCAGCATCCCAGCCCCAGTTTACC
AGCCAGCTCCTGCTCCCATTAGGATCCCAGCCCCAGTTTACCAGCCAGCT
CCTGCTCCCATCAGCATCCCA[redacted]TGAGATCCCTGCCCC
AGCTCCAGTGAACACCTACATTCCCCCCGCACCAGCACCAGCTCCGGTTT
ACCAGCCAGCCCCTGCTCCCATTCCCGTGAGCATCCCGGCTCCAGCTCCC
GTTTACCAGCCTGCTCCCGCTCCCGTTGTGATCCCAGCCCCTGCTCCAGC
ACCAGTTGTGATCCCTGCTCCTGCTCCCGCTCCAGTTGTGATCCCAGCCC
CAGCTCCCGTTAAGTCCTATGTGCCACCTGCACCAATCAGCATCCCAGCC
CCGGCTCCCGTTTACCAGCCAGCTCCCATCAGCATCCCGGCCCCTGCCCC
AGTTTACCAACCAGCTCCCGCTCCCGTCTACCAGCCAACCAACACTCAGG
TTCTGGAGGAGATCGAGCCCGCTTCCAATGATGGATACCGCTACAAGACC
GTGCGTCGTCGCGTCTACCGTCACCGTTTCTAAACTTACCAAAGCCCAAC
GAAAAACTGTTTCCTAACGAATCCCTACGGGGACTCGCAATAAAAAATTT
CCGCCTTCAATTTGTTAAAC

MKFLVVAFAVLACAYGDVSHLGYDYSPPAPAPVYQPAPAPVYQPAPAPVY
QPAPAPVVIPAPAPAPVVIPAPAPAPVVIPAPAPVKTYVPPAPISIPAPV
YQPAPAPIRIPAPVYQPAPAPISIPAPAPIEIPAPAPVNTYIPPAPAPAP
VYQPAPAPIPVSIPAPAPVYQPAPAPVVIPAPAPAPVVIPAPAPAPVVIP
APAPVKSYVPPAPISIPAPAPVYQPAPISIPAPAPVYQPAPAPVYQPTNT
QVLEEIEPASNDGYRYKTVRRRVYRHRF

Repeat and low complexity regions create alignment problems



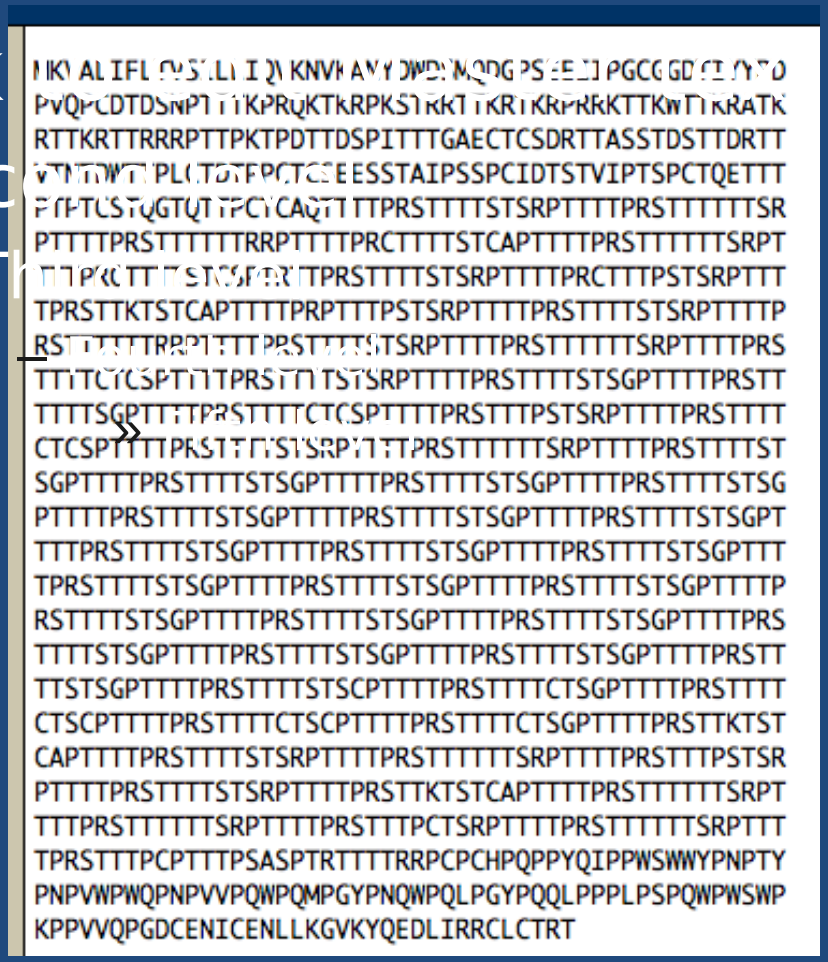
Sgs1 repeats foil RNA-Seq alignments

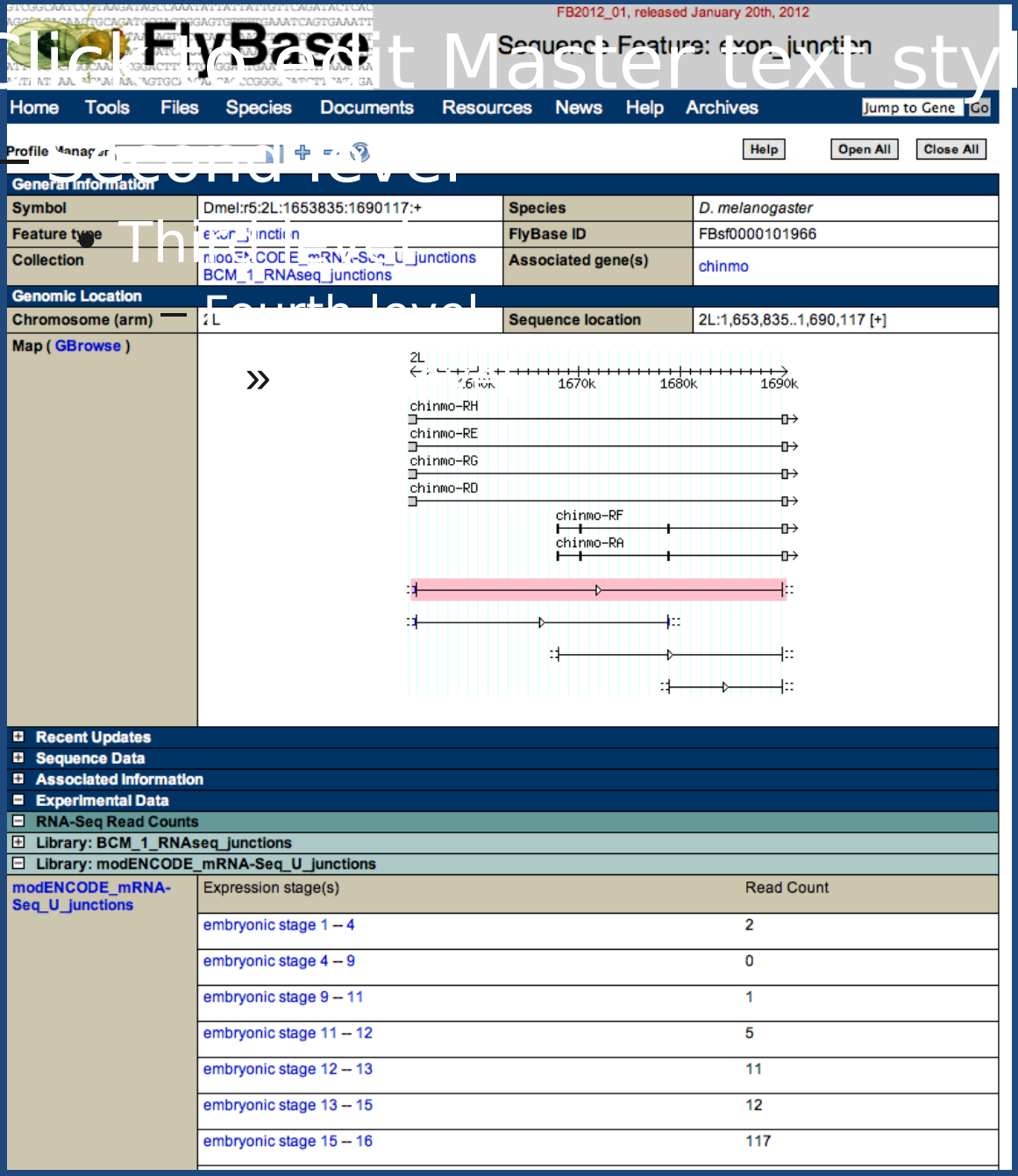
lick

Seq

- Th

styles



ATCCACAACAACAACTTCCACAAGTGGCCCAACGACAACTACTCCTAGAT
CCACAACAACAACTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCC
ACAACAACAACTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCCAC
AACAACAACTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAA
CAACAACTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACA
ACAACTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAAC
AACTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAA
CTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAACT
TCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAACTTC
CACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAACTTCCA
CAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAACTTCCACA
AGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAACTTCCACAAG
TGGCCCAACGACAACTACTCCTAGATCCACAACAACAACTTCCACAAGTG
GCCCAACGACAACTACTCCTAGATCCACAACAACAACTTCCACAAGTGGC
CCAACGACAACTACTCCTAGATCCACAACAACAACTTCCACAAGTTGCCC
AACGACAACTACTCCTAGATCCACAACAACAACTTGCACAAGTGGCCCAA
CGACAACTACTCCTAGATCCACAACAACAACTTGCACAAGTTGCCCAACG
ACAACTACTCCTAGATCCACAACAACAACTTGCACAAGTTGCCCAACGAC
AACTACTCCTAGATCCACAACAACAACTTGCACAAGTGGCCCAACGACAA
CTACTCCTAGATCCACAACAAAAACCTCCACTTGTGCCCCAACGACAACT
ACTCCTAGATCCACAACAACAACTTCCACAAGTCGCCCAACGACAACTAC
TCCTAGATCCACAACAACAACTACCACAAGTCGCCCAACGACGACTACTC
CTAGATCCACAACAACACCTTCCACAAGTCGCCCAACCACAACTACTCCT
AGATCCACAACAACTACTTCCACAAGTCGCCCAACGACAACTACTCCTAG
ATCCACAACAAAAACTTCCACTTGTGCCCCAACGACAACTACTCCTAGAT

- Click to edit Master text styles
  - Second level
    - Third level
      - Fourth level