Implementing RNA-Seq data in FlyBase chado

David Emmert

GMOD Community Meeting

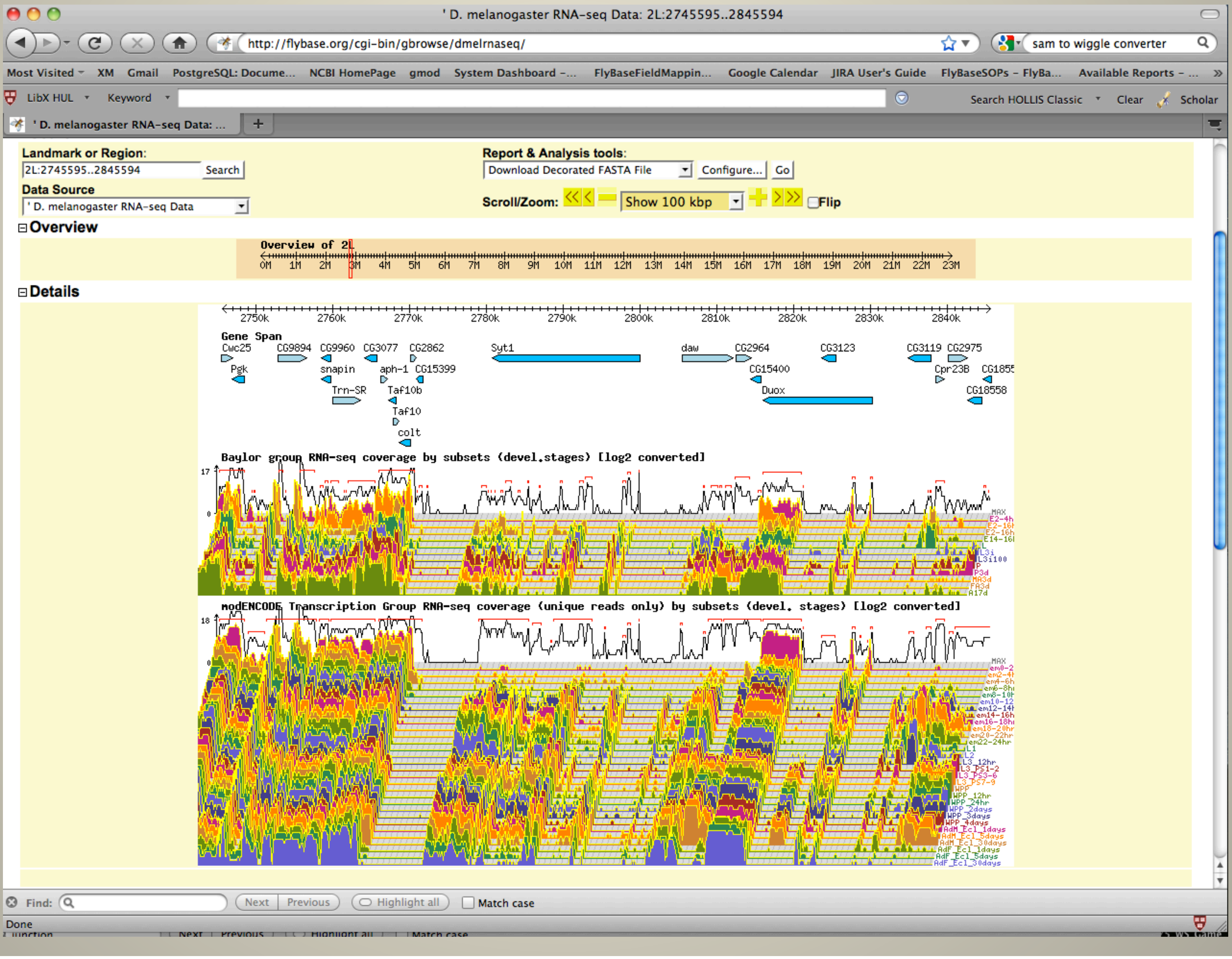
March 6&7, 2011

RNA-Seq Data

- Data sets:
  - – Daines et. al. 2010
    - Solexa/Illumina GAII 65, 75 & 100nt paired-end reads; 12 developmental stages
      - – 142.2 million uniquely mapped reads (unstranded)
      - – 54,594 unique junctions
  - – Graveley et. al. 2010 (modENCODE)
    - Illumina GAII 75 & 76nt single- and paired-end reads; 30 developmental stages
      - – 2.25 billion uniquely mapped reads (unstranded)
      - – 67,317 unique junctions

RNA-Seq Coverage Data

- Graphically represented in Gbrowse (but not integrated).



' D. melanogaster RNA-seq Data: 2L:2745595..2845594

http://flybase.org/cgi-bin/gbrowse/dmelrnaseq/

sam to wiggle converter

Landmark or Region: 2L:2745595..2845594

Data Source: ' D. melanogaster RNA-seq Data

Report & Analysis tools: Download Decorated FASTA File

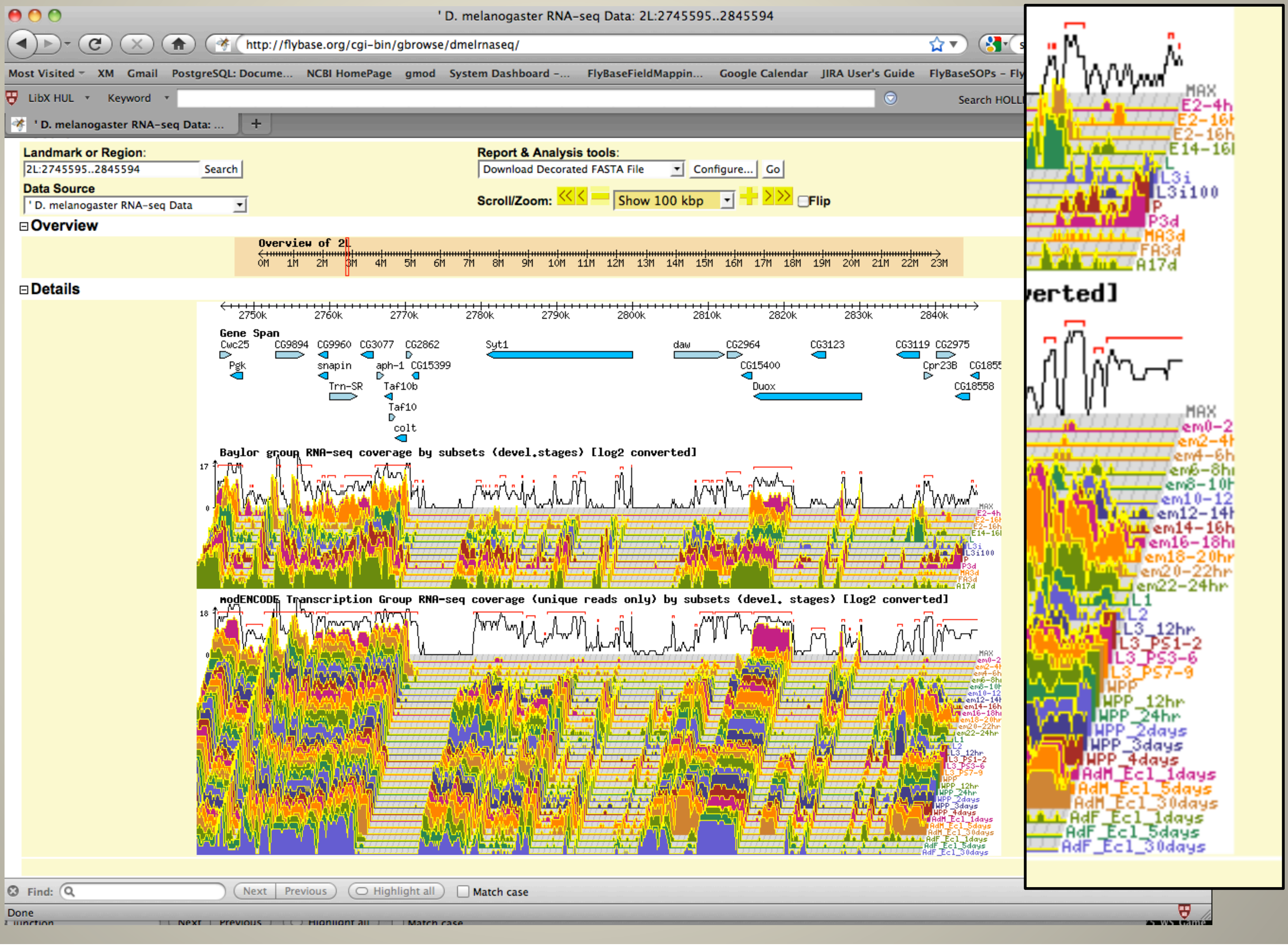
Scroll/Zoom: Show 100 kbp

Baylor group RNA-seq coverage by subsets (devel.stages) [log2 converted]

modENCODE Transcription Group RNA-seq coverage (unique reads only) by subsets (devel. stages) [log2 converted]

RNA-Seq Coverage Data

- Graphically represented in Gbrowse (but not integrated).



' D. melanogaster RNA-seq Data: 2L:2745595..2845594

http://flybase.org/cgi-bin/gbrowse/dmelrnaseq/

Landmark or Region: 2L:2745595..2845594

Data Source: ' D. melanogaster RNA-seq Data

Report & Analysis tools: Download Decorated FASTA File

Scroll/Zoom: Show 100 kbp

Baylor group RNA-seq coverage by subsets (devel.stages) [log2 converted]

modENCODE Transcription Group RNA-seq coverage (unique reads only) by subsets (devel. stages) [log2 converted]

Done

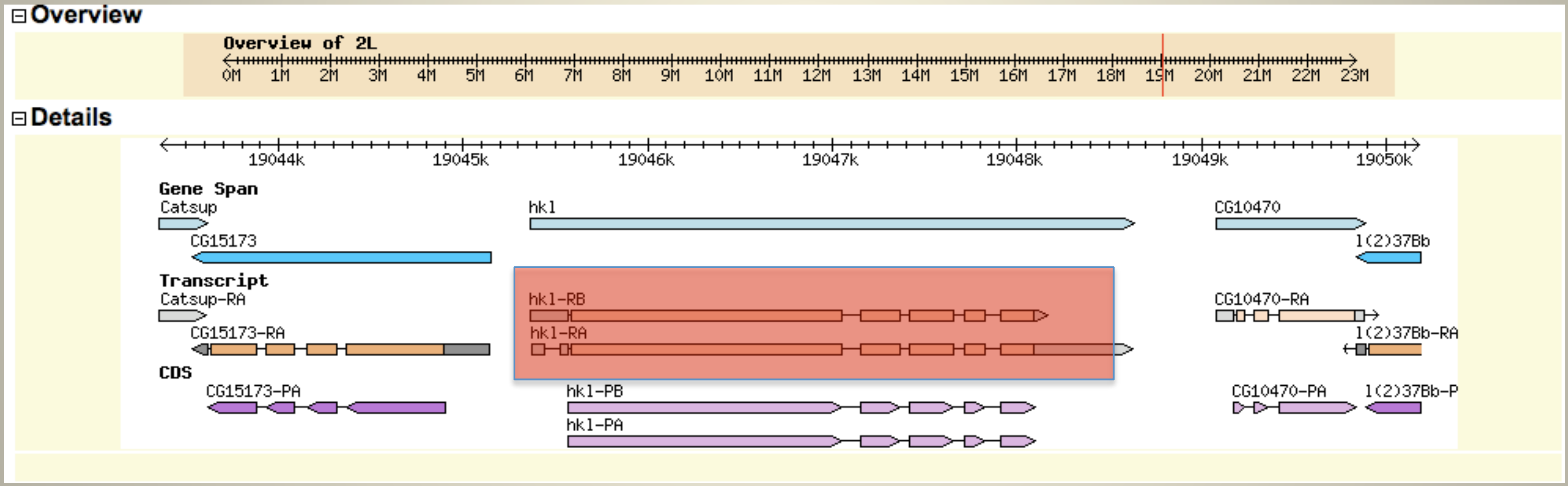
RNA-Seq Coverage Data

- Integration with genes:
  - Assign expression CV terms to genes based on coverage data.
  - Summarize HT expression data in gene reports.
  - Enable gene search by expression pattern.
  - Enable search for similarly expressed genes.

RNA-Seq Coverage Data

- Integration with genes:

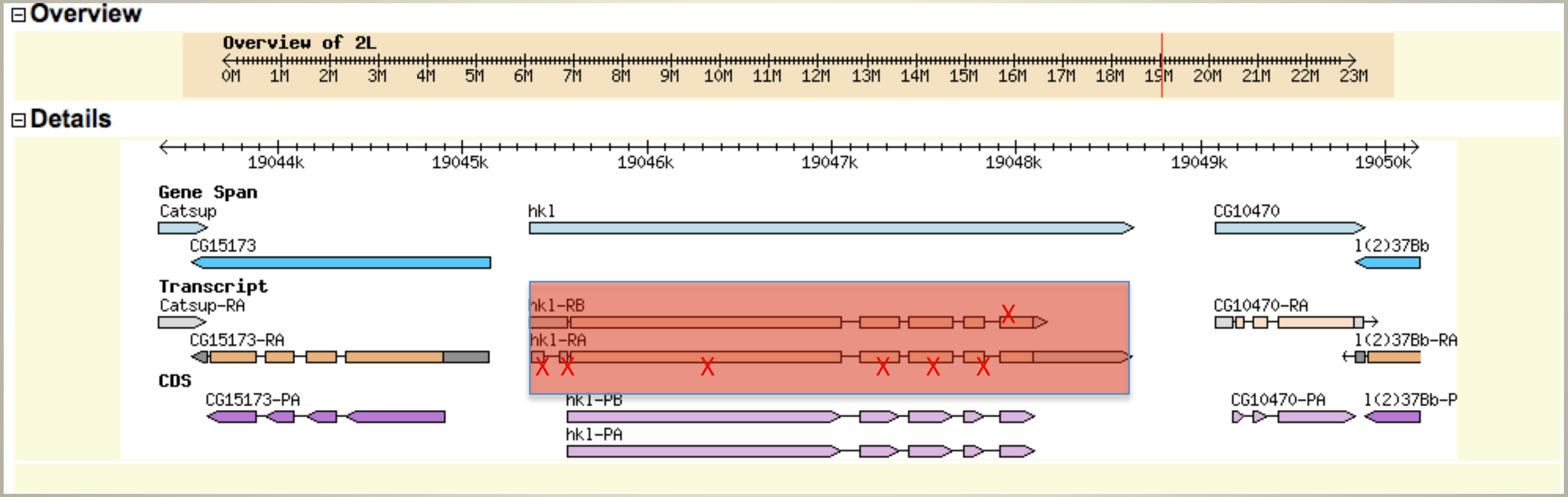
1) Determine unique transcribed region for each gene, e.g. *Dmel\hkl*:



RNA-Seq Coverage Data

- Integration with genes:

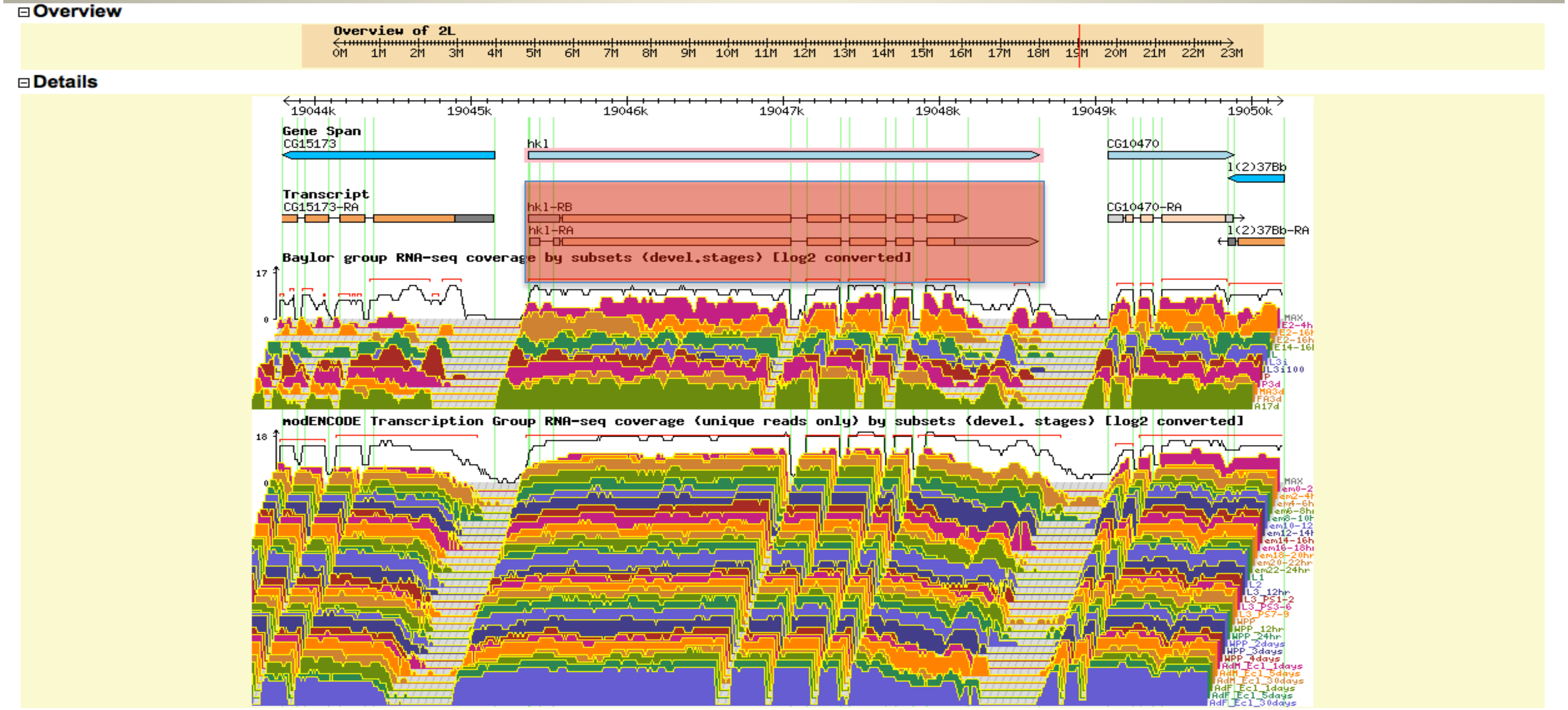
1) Determine unique transcribed region for each gene, e.g. *Dmel\hkl*:



RNA-Seq Coverage Data

- Integration with genes:

2) Correlate coverage for each stage over unique transcribed region:

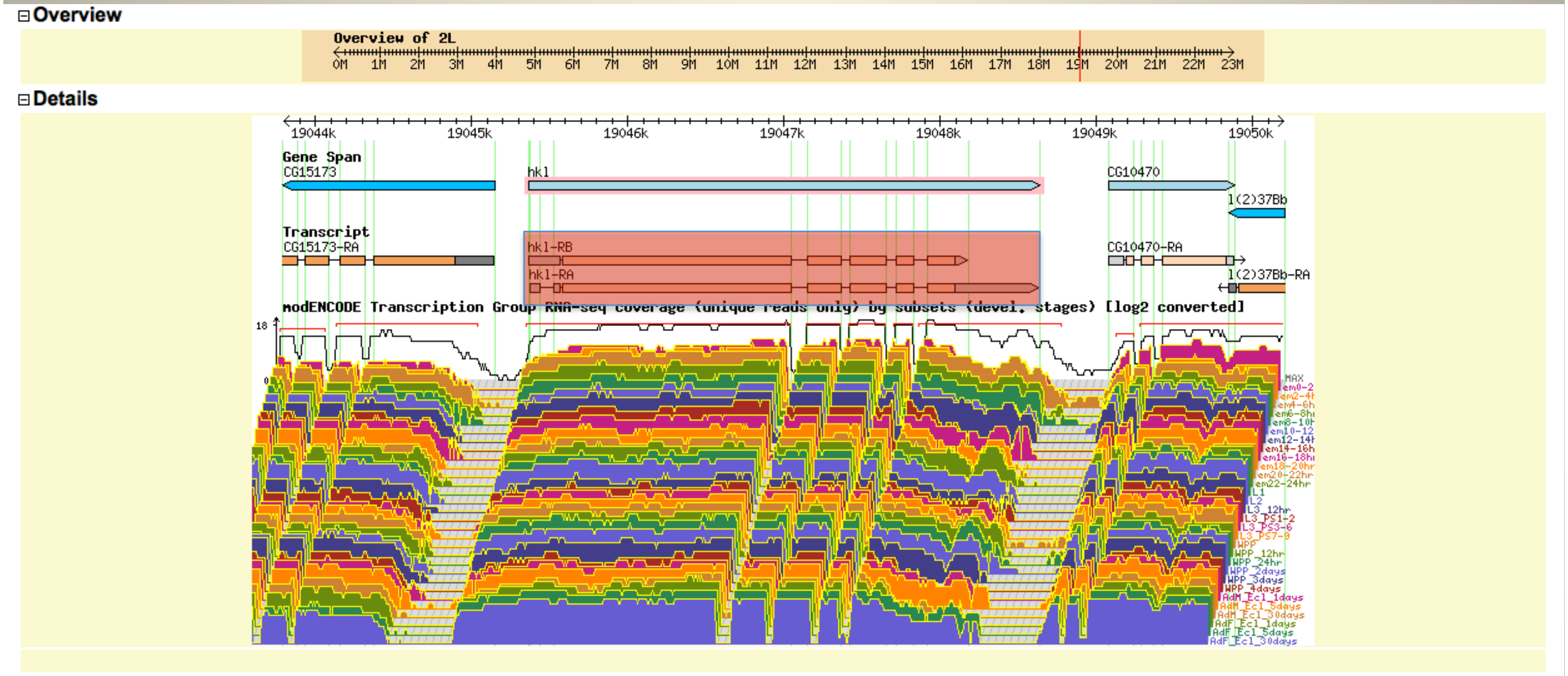


RNA-Seq Coverage Data

- Integration with genes:

3) Calculate $$RPKM = \frac{total\_exon\_reads}{exon\_length(KB) \times mapped\_reads(millions)}$$

*RPKM: **R**eads **p**er **k**ilobase of exon **m**odel per million mapped reads*

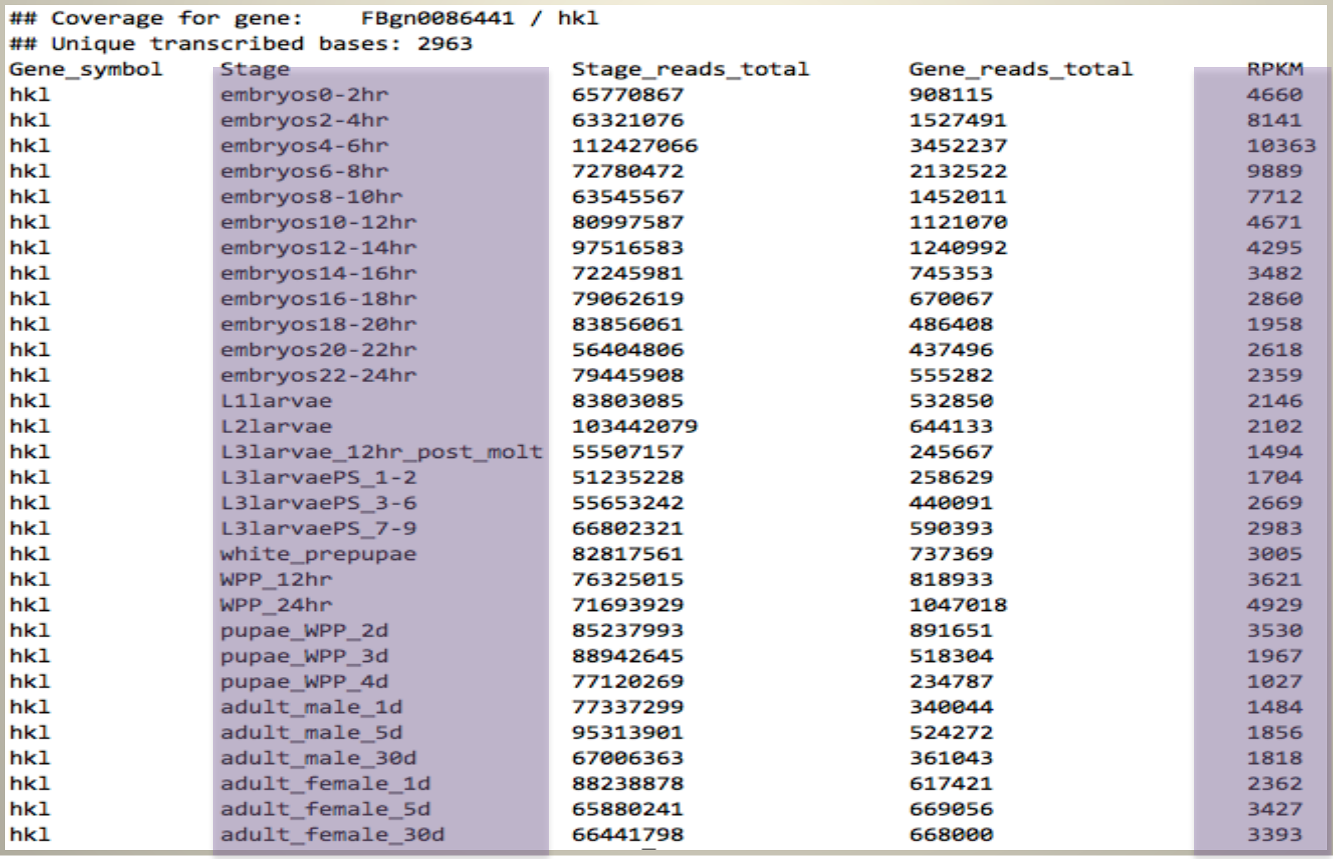


RNA-Seq Coverage Data

- Integration with genes:

3) Calculate $RPKM = \dfrac{total\_exon\_reads}{exon\_length(KB) \times mapped\_reads(millions)}$

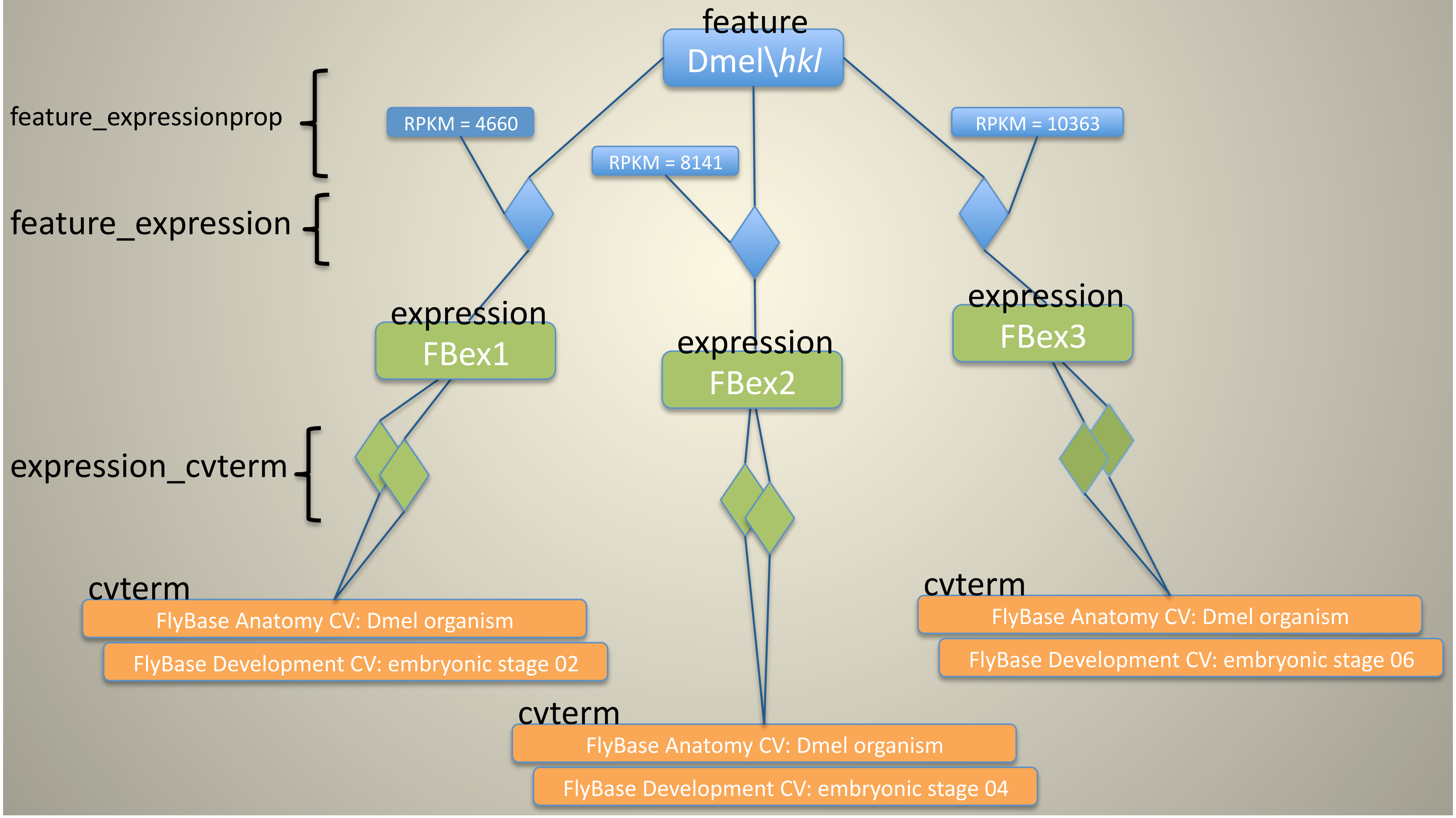
*RPKM: **R**eads **p**er **k**ilobase of exon **m**odel per million mapped reads*



```
## Coverage for gene:    FBgn0086441 / hkl
## Unique transcribed bases: 2963
Gene_symbol    Stage                     Stage_reads_total        Gene_reads_total         RPKM
hkl            embryos0-2hr              65770867                 908115                   4660
hkl            embryos2-4hr              63321076                 1527491                  8141
hkl            embryos4-6hr              112427066                3452237                  10363
hkl            embryos6-8hr              72780472                 2132522                  9889
hkl            embryos8-10hr             63545567                 1452011                  7712
hkl            embryos10-12hr            80997587                 1121070                  4671
hkl            embryos12-14hr            97516583                 1240992                  4295
hkl            embryos14-16hr            72245981                 745353                   3482
hkl            embryos16-18hr            79062619                 670067                   2860
hkl            embryos18-20hr            83856061                 486408                   1958
hkl            embryos20-22hr            56404806                 437496                   2618
hkl            embryos22-24hr            79445908                 555282                   2359
hkl            L1larvae                  83803085                 532850                   2146
hkl            L2larvae                  103442079                644133                   2102
hkl            L3larvae_12hr_post_molt   55507157                 245667                   1494
hkl            L3larvaePS_1-2            51235228                 258629                   1704
hkl            L3larvaePS_3-6            55653242                 440091                   2669
hkl            L3larvaePS_7-9            66802321                 590393                   2983
hkl            white_prepupae            82817561                 737369                   3005
hkl            WPP_12hr                  76325015                 818933                   3621
hkl            WPP_24hr                  71693929                 1047018                  4929
hkl            pupae_WPP_2d              85237993                 891651                   3530
hkl            pupae_WPP_3d              88942645                 518304                   1967
hkl            pupae_WPP_4d              77120269                 234787                   1027
hkl            adult_male_1d             77337299                 340044                   1484
hkl            adult_male_5d             95313901                 524272                   1856
hkl            adult_male_30d            67006363                 361043                   1818
hkl            adult_female_1d           88238878                 617421                   2362
hkl            adult_female_5d           65880241                 669056                   3427
hkl            adult_female_30d          66441798                 668000                   3393
```

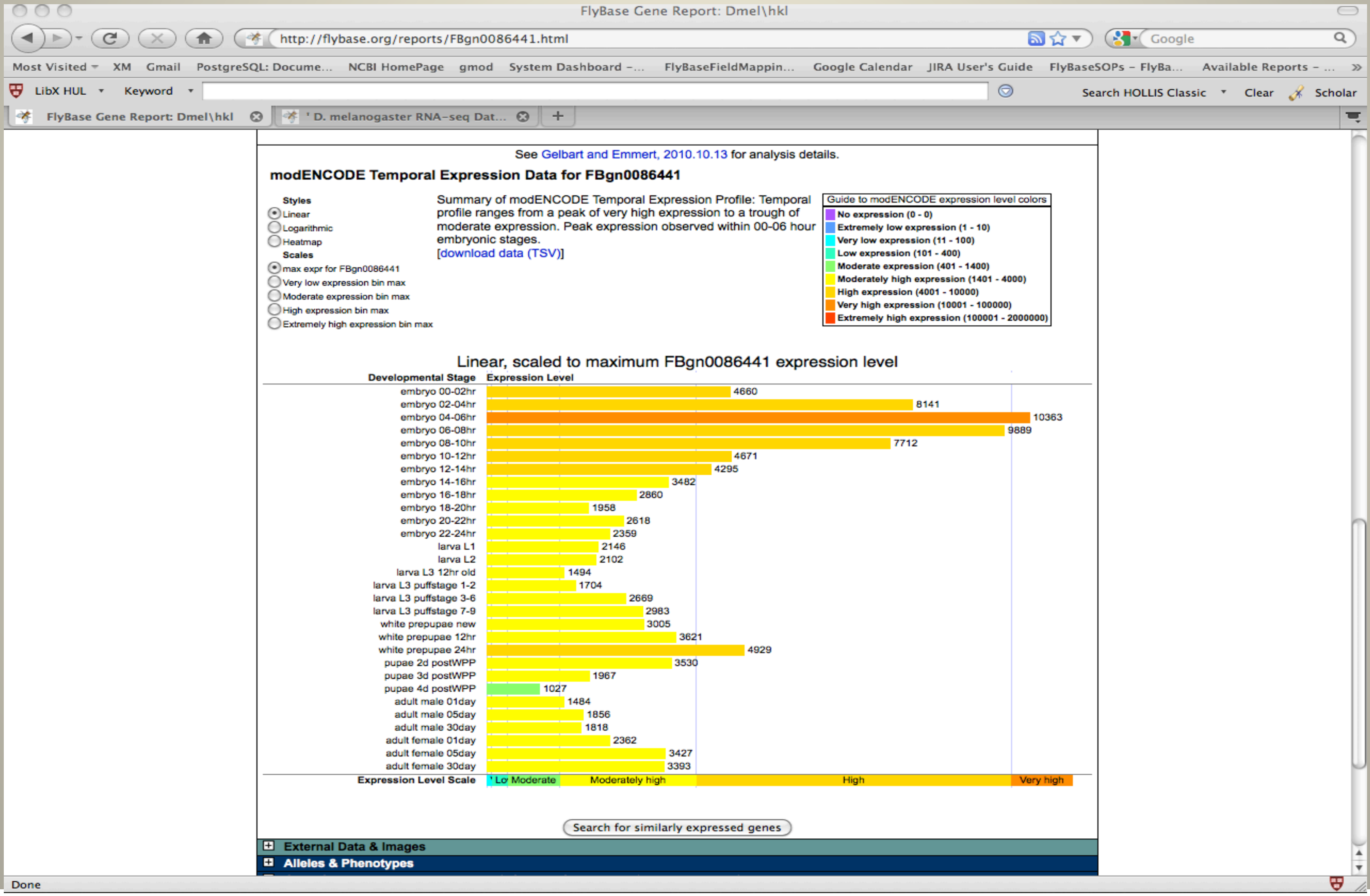
RNA-Seq Coverage Data

- Chado implementation:


feature
Dmel*hkl*
feature_expressionprop
RPKM = 4660
RPKM = 8141
RPKM = 10363
feature_expression
expression
FBex1
expression
FBex2
expression
FBex3
expression_cvterm
cvterm
FlyBase Anatomy CV: Dmel organism
FlyBase Development CV: embryonic stage 02
cvterm
FlyBase Anatomy CV: Dmel organism
FlyBase Development CV: embryonic stage 06
cvterm
FlyBase Anatomy CV: Dmel organism
FlyBase Development CV: embryonic stage 04

RNA-Seq Coverage Data

- Reporting & Searching:



See Gelbart and Emmert, 2010.10.13 for analysis details.

modENCODE Temporal Expression Data for FBgn0086441

Styles
● Linear
○ Logarithmic
○ Heatmap

Scales
● max expr for FBgn0086441
○ Very low expression bin max
○ Moderate expression bin max
○ High expression bin max
○ Extremely high expression bin max

Summary of modENCODE Temporal Expression Profile: Temporal profile ranges from a peak of very high expression to a trough of moderate expression. Peak expression observed within 00-06 hour embryonic stages.
[download data (TSV)]

Guide to modENCODE expression level colors
- **No expression (0 - 0)**
- **Extremely low expression (1 - 10)**
- **Very low expression (11 - 100)**
- **Low expression (101 - 400)**
- **Moderate expression (401 - 1400)**
- **Moderately high expression (1401 - 4000)**
- **High expression (4001 - 10000)**
- **Very high expression (10001 - 100000)**
- **Extremely high expression (100001 - 2000000)**

Linear, scaled to maximum FBgn0086441 expression level

Developmental Stage	Expression Level
embryo 00-02hr	4660
embryo 02-04hr	8141
embryo 04-06hr	10363
embryo 06-08hr	9889
embryo 08-10hr	7712
embryo 10-12hr	4671
embryo 12-14hr	4295
embryo 14-16hr	3482
embryo 16-18hr	2860
embryo 18-20hr	1958
embryo 20-22hr	2618
embryo 22-24hr	2359
larva L1	2146
larva L2	2102
larva L3 12hr old	1494
larva L3 puffstage 1-2	1704
larva L3 puffstage 3-6	2669
larva L3 puffstage 7-9	2983
white prepupae new	3005
white prepupae 12hr	3621
white prepupae 24hr	4929
pupae 2d postWPP	3530
pupae 3d postWPP	1967
pupae 4d postWPP	1027
adult male 01day	1484
adult male 05day	1856
adult male 30day	1818
adult female 01day	2362
adult female 05day	3427
adult female 30day	3393
Expression Level Scale	' Lo Moderate · Moderately high · High · Very high

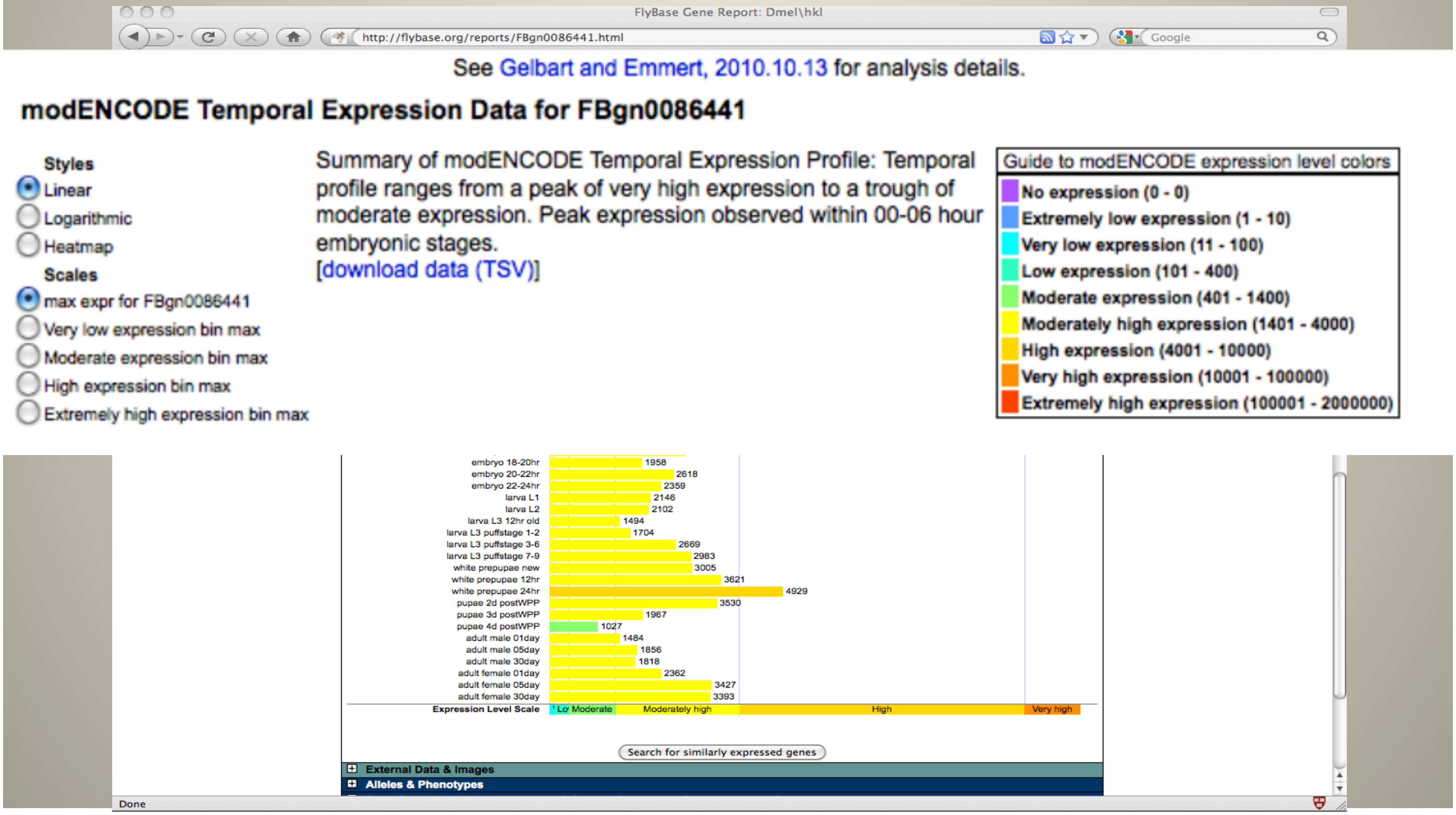
Search for similarly expressed genes

⊞ **External Data & Images**

⊞ **Alleles & Phenotypes**

RNA-Seq Coverage Data

- Reporting & Searching:


FlyBase Gene Report: Dmel\hkl
http://flybase.org/reports/FBgn0086441.html
Google

See Gelbart and Emmert, 2010.10.13 for analysis details.

modENCODE Temporal Expression Data for FBgn0086441

Styles

- ● Linear
- ○ Logarithmic
- ○ Heatmap

Scales

- ● max expr for FBgn0086441
- ○ Very low expression bin max
- ○ Moderate expression bin max
- ○ High expression bin max
- ○ Extremely high expression bin max

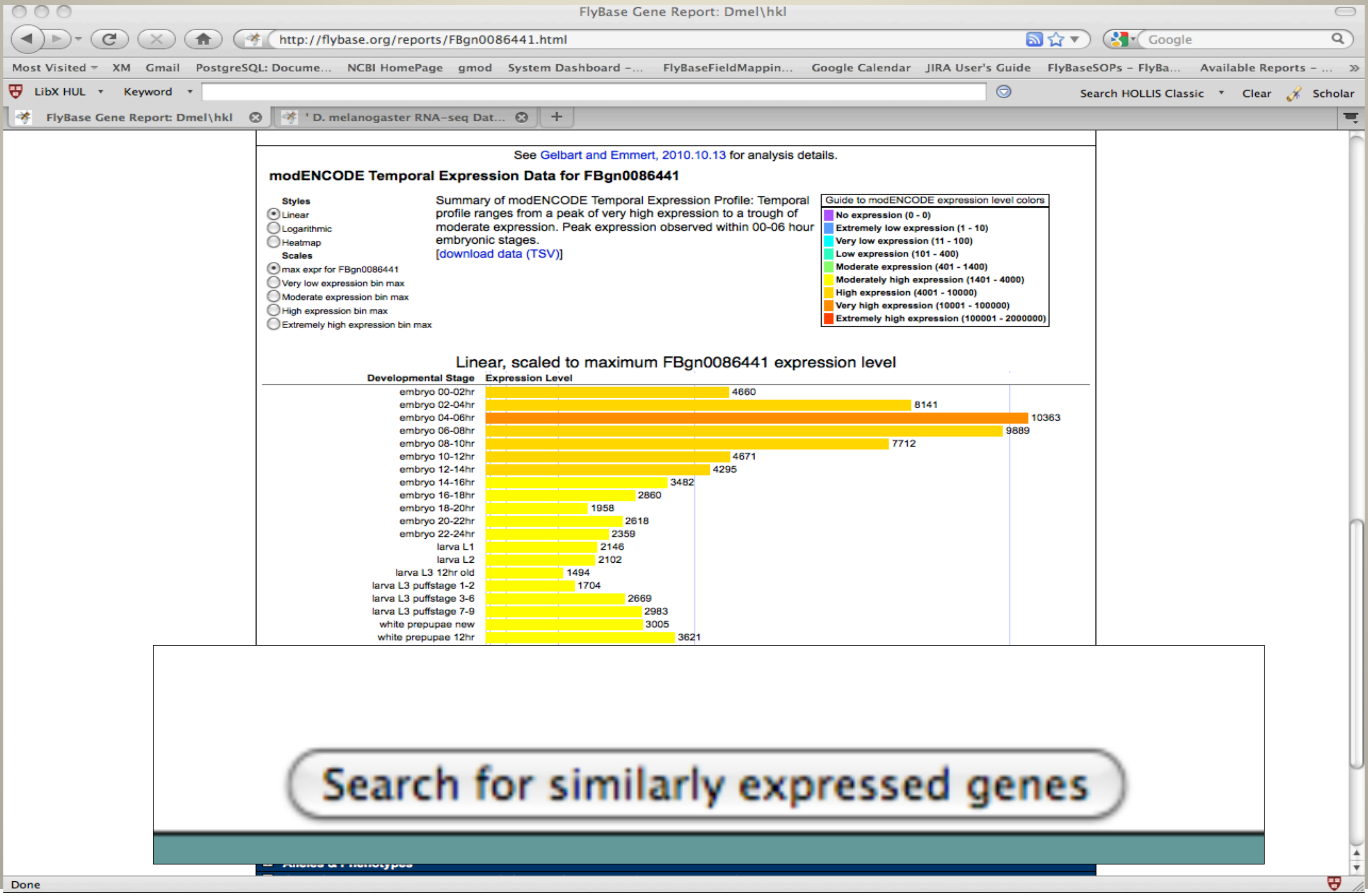
Summary of modENCODE Temporal Expression Profile: Temporal profile ranges from a peak of very high expression to a trough of moderate expression. Peak expression observed within 00-06 hour embryonic stages.
[download data (TSV)]

Guide to modENCODE expression level colors

- 🟪 **No expression (0 - 0)**
- 🟦 **Extremely low expression (1 - 10)**
- 🟦 **Very low expression (11 - 100)**
- 🟩 **Low expression (101 - 400)**
- 🟩 **Moderate expression (401 - 1400)**
- 🟨 **Moderately high expression (1401 - 4000)**
- 🟨 **High expression (4001 - 10000)**
- 🟧 **Very high expression (10001 - 100000)**
- 🟥 **Extremely high expression (100001 - 2000000)**

RNA-Seq Coverage Data

- Reporting & Searching:



See Gelbart and Emmert, 2010.10.13 for analysis details.

modENCODE Temporal Expression Data for FBgn0086441

Styles
● Linear
○ Logarithmic
○ Heatmap

Scales
● max expr for FBgn0086441
○ Very low expression bin max
○ Moderate expression bin max
○ High expression bin max
○ Extremely high expression bin max

Summary of modENCODE Temporal Expression Profile: Temporal profile ranges from a peak of very high expression to a trough of moderate expression. Peak expression observed within 00-06 hour embryonic stages.
[download data (TSV)]

Guide to modENCODE expression level colors

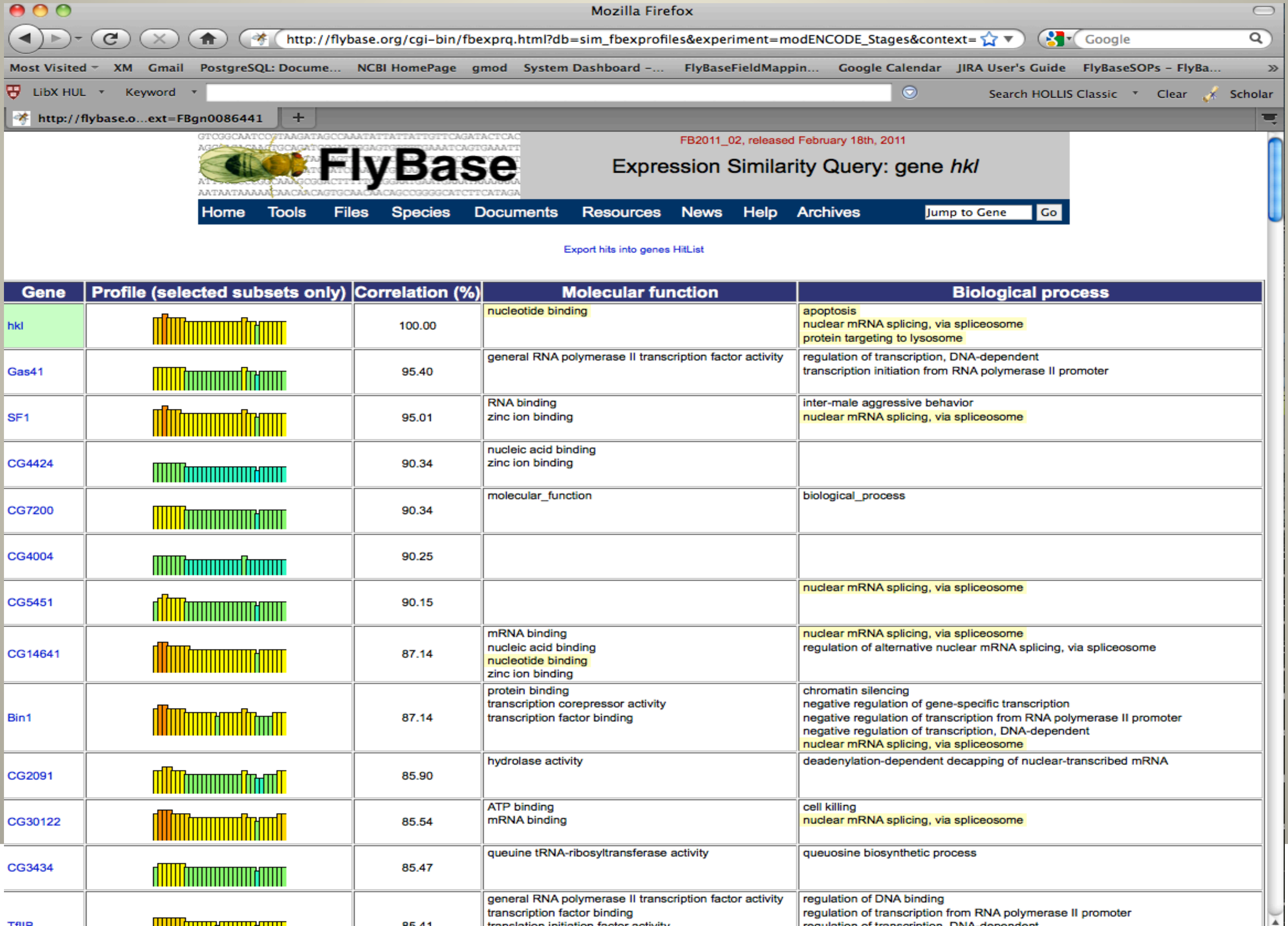
- 🟪 No expression (0 - 0)
- 🟦 Extremely low expression (1 - 10)
- 🟦 Very low expression (11 - 100)
- 🟩 Low expression (101 - 400)
- 🟩 Moderate expression (401 - 1400)
- 🟨 Moderately high expression (1401 - 4000)
- 🟨 High expression (4001 - 10000)
- 🟧 Very high expression (10001 - 100000)
- 🟥 Extremely high expression (100001 - 2000000)

Linear, scaled to maximum FBgn0086441 expression level

Developmental Stage	Expression Level
embryo 00-02hr	4660
embryo 02-04hr	8141
embryo 04-06hr	10363
embryo 06-08hr	9889
embryo 08-10hr	7712
embryo 10-12hr	4671
embryo 12-14hr	4295
embryo 14-16hr	3482
embryo 16-18hr	2860
embryo 18-20hr	1958
embryo 20-22hr	2618
embryo 22-24hr	2359
larva L1	2146
larva L2	2102
larva L3 12hr old	1494
larva L3 puffstage 1-2	1704
larva L3 puffstage 3-6	2669
larva L3 puffstage 7-9	2983
white prepupae new	3005
white prepupae 12hr	3621

RNA-Seq Coverage Data

- Reporting & Searching:



http://flybase.org/cgi-bin/fbexprq.html?db=sim_fbexprofiles&experiment=modENCODE_Stages&context=

FB2011_02, released February 18th, 2011

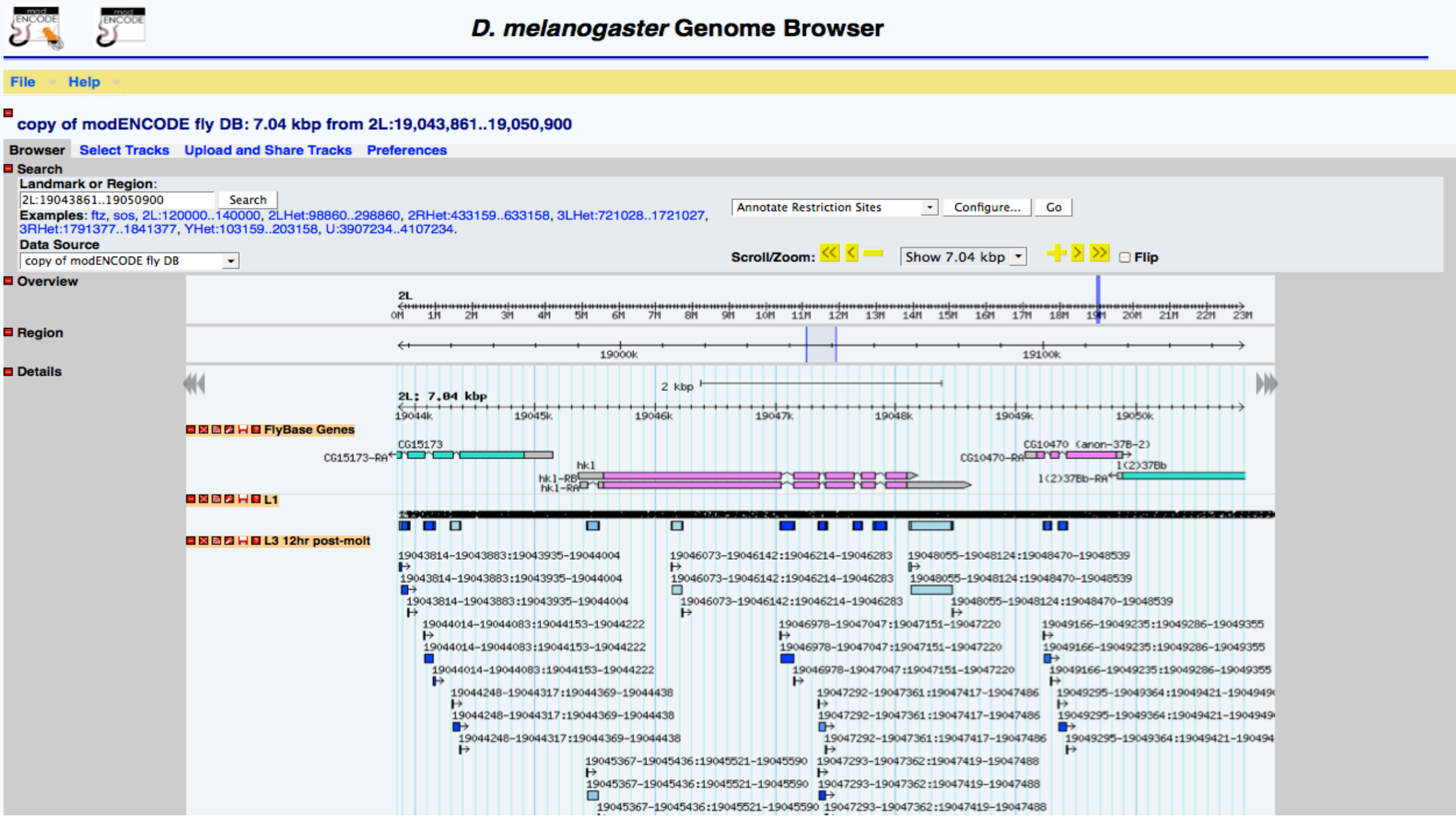
Expression Similarity Query: gene *hkl*

Export hits into genes HitList

Gene	Profile (selected subsets only)	Correlation (%)	Molecular function	Biological process
hkl		100.00	nucleotide binding	apoptosis nuclear mRNA splicing, via spliceosome protein targeting to lysosome
Gas41		95.40	general RNA polymerase II transcription factor activity	regulation of transcription, DNA-dependent transcription initiation from RNA polymerase II promoter
SF1		95.01	RNA binding zinc ion binding	inter-male aggressive behavior nuclear mRNA splicing, via spliceosome
CG4424		90.34	nucleic acid binding zinc ion binding	
CG7200		90.34	molecular_function	biological_process
CG4004		90.25		
CG5451		90.15		nuclear mRNA splicing, via spliceosome
CG14641		87.14	mRNA binding nucleic acid binding nucleotide binding zinc ion binding	nuclear mRNA splicing, via spliceosome regulation of alternative nuclear mRNA splicing, via spliceosome
Bin1		87.14	protein binding transcription corepressor activity transcription factor binding	chromatin silencing negative regulation of gene-specific transcription negative regulation of transcription from RNA polymerase II promoter negative regulation of transcription, DNA-dependent nuclear mRNA splicing, via spliceosome
CG2091		85.90	hydrolase activity	deadenylation-dependent decapping of nuclear-transcribed mRNA
CG30122		85.54	ATP binding mRNA binding	cell killing nuclear mRNA splicing, via spliceosome
CG3434		85.47	queuine tRNA-ribosyltransferase activity	queuosine biosynthetic process
TfIIB		85.41	general RNA polymerase II transcription factor activity transcription factor binding translation initiation factor activity	regulation of DNA binding regulation of transcription from RNA polymerase II promoter regulation of transcription, DNA-dependent

RNA-Seq Junction Data

- Graphically represented in Gbrowse (neither consolidated nor integrated).



D. melanogaster Genome Browser

File Help

copy of modENCODE fly DB: 7.04 kbp from 2L:19,043,861..19,050,900

Browser | Select Tracks | Upload and Share Tracks | Preferences

Search

Landmark or Region:

2L:19043861..19050900 Search

Examples: ftz, sos, 2L:120000..140000, 2LHet:98860..298860, 2RHet:433159..633158, 3LHet:721028..1721027, 3RHet:1791377..1841377, YHet:103159..203158, U:3907234..4107234.

Data Source

copy of modENCODE fly DB

Annotate Restriction Sites | Configure... | Go

Scroll/Zoom: Show 7.04 kbp ☐ Flip

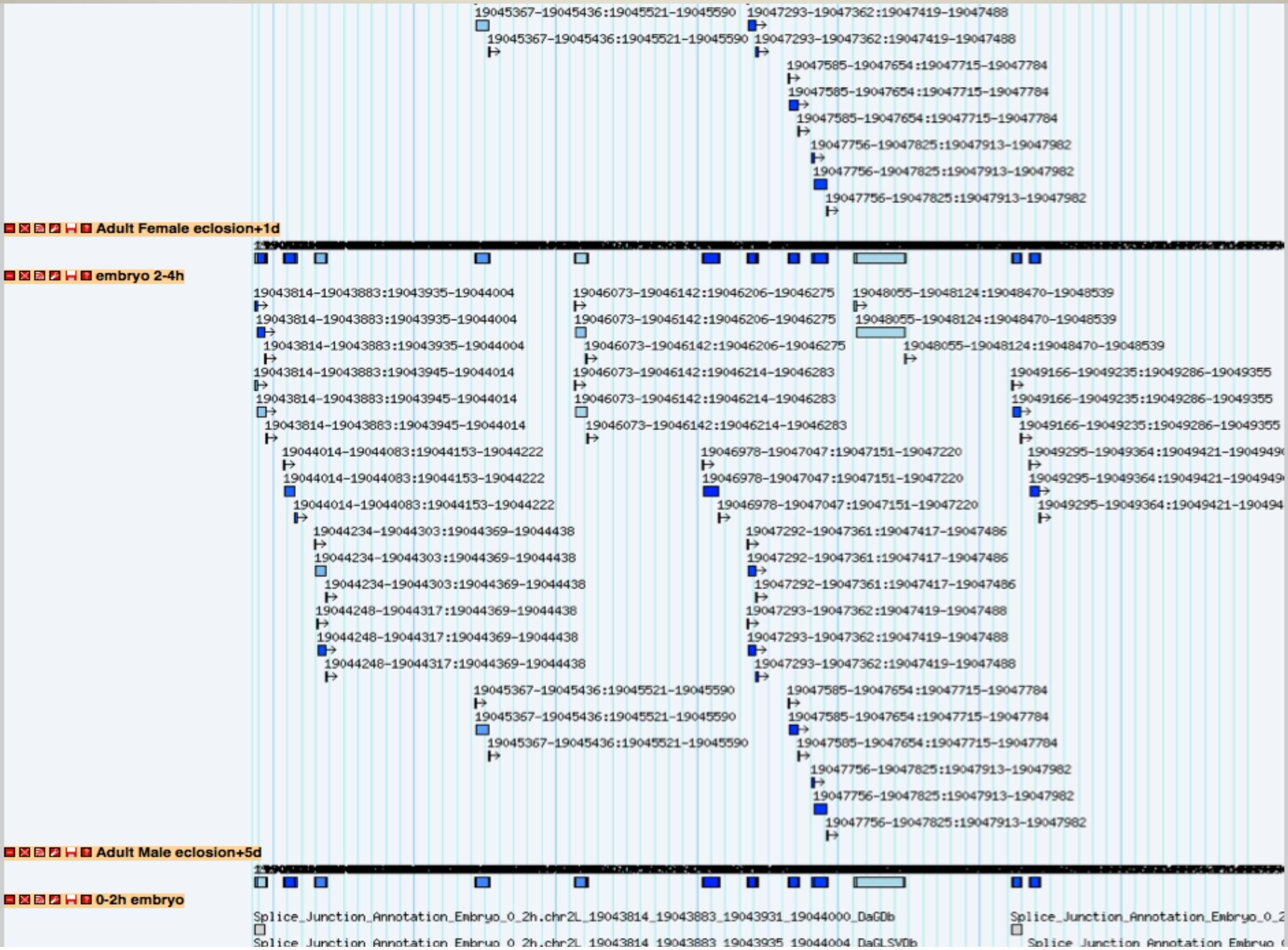
Overview

Region

Details

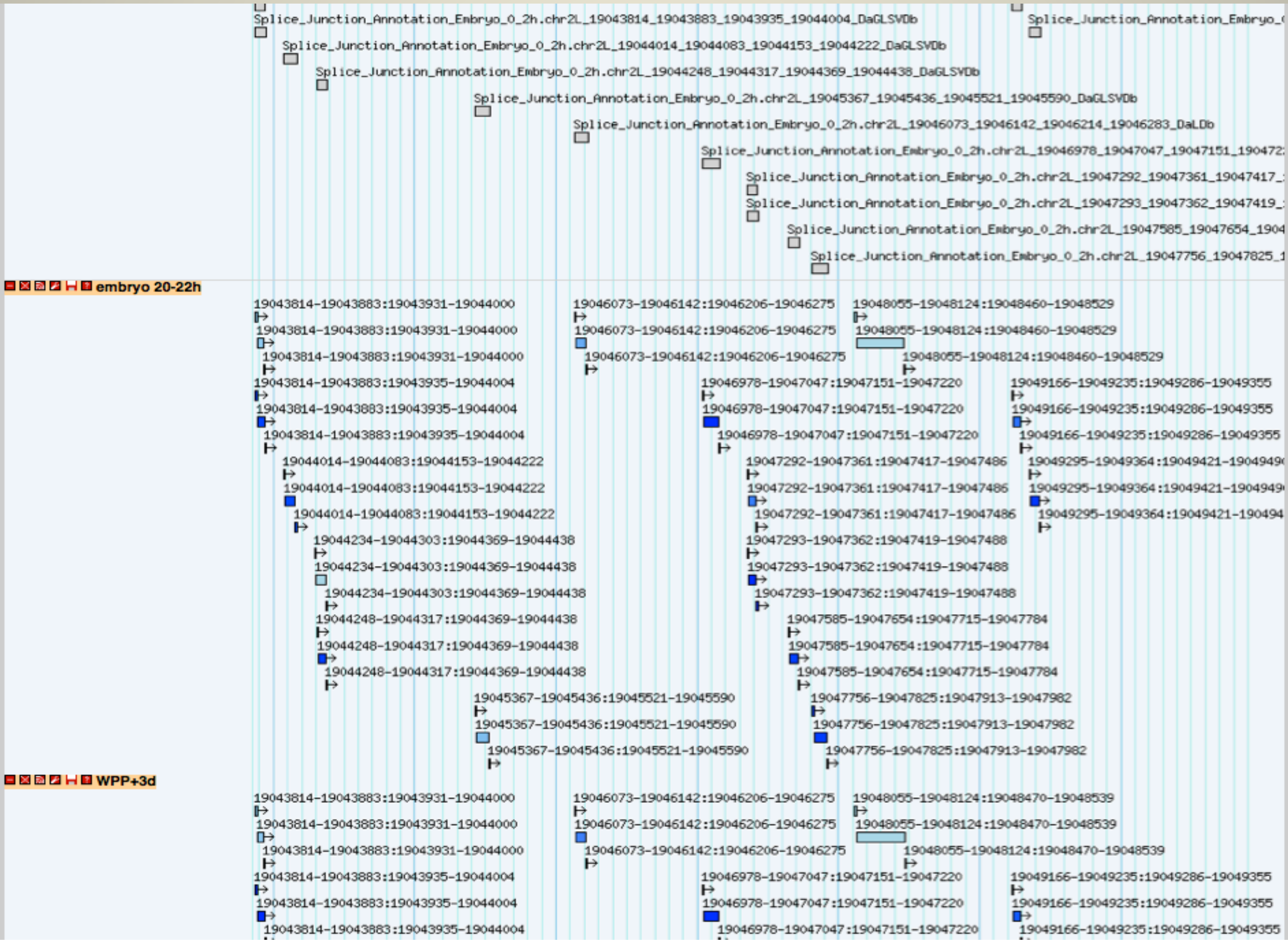
RNA-Seq Junction Data

- Graphically represented in Gbrowse (neither consolidated nor integrated).



RNA-Seq Junction Data

- Graphically represented in Gbrowse (neither consolidated nor integrated).



```
Splice_Junction_Annotation_Embryo_0_2h.chr2L_19043814_19043883_19043935_19044004_DaGLSVDb
Splice_Junction_Annotation_Embryo_0_2h.chr2L_19044014_19044083_19044153_19044222_DaGLSVDb
Splice_Junction_Annotation_Embryo_0_2h.chr2L_19044248_19044317_19044369_19044438_DaGLSVDb
Splice_Junction_Annotation_Embryo_0_2h.chr2L_19045367_19045436_19045521_19045590_DaGLSVDb
Splice_Junction_Annotation_Embryo_0_2h.chr2L_19046073_19046142_19046214_19046283_DaLDb
Splice_Junction_Annotation_Embryo_0_2h.chr2L_19046978_19047047_19047151_190472
Splice_Junction_Annotation_Embryo_0_2h.chr2L_19047292_19047361_19047417_
Splice_Junction_Annotation_Embryo_0_2h.chr2L_19047293_19047362_19047419_
Splice_Junction_Annotation_Embryo_0_2h.chr2L_19047585_19047654_1904
Splice_Junction_Annotation_Embryo_0_2h.chr2L_19047756_19047825_1
Splice_Junction_Annotation_Embryo_
```

embryo 20-22h

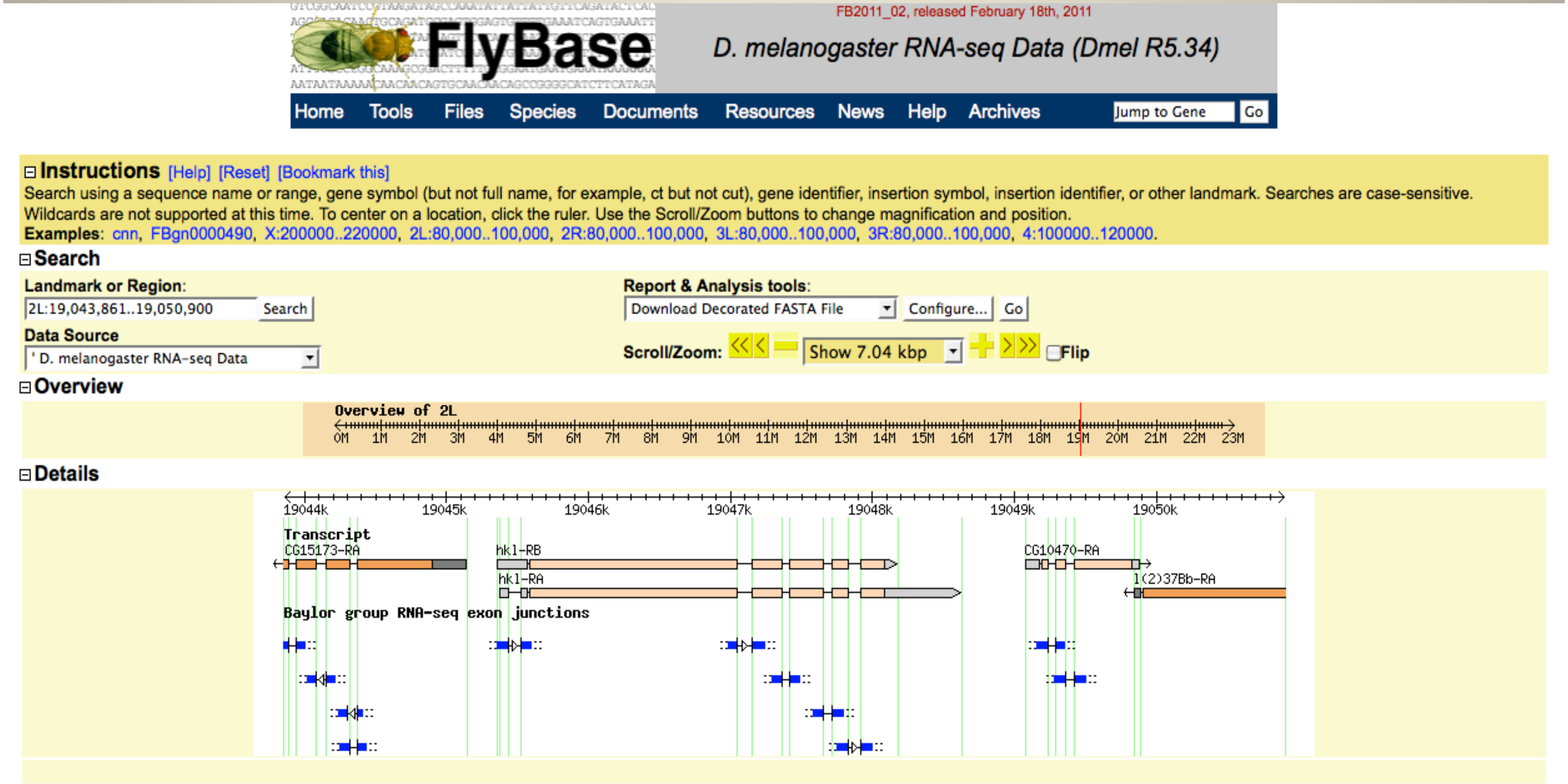
```
19043814-19043883:19043931-19044000
19043814-19043883:19043931-19044000
19043814-19043883:19043931-19044000
19043814-19043883:19043935-19044004
19043814-19043883:19043935-19044004
19043814-19043883:19043935-19044004
19044014-19044083:19044153-19044222
19044014-19044083:19044153-19044222
19044014-19044083:19044153-19044222
19044234-19044303:19044369-19044438
19044234-19044303:19044369-19044438
19044234-19044303:19044369-19044438
19044248-19044317:19044369-19044438
19044248-19044317:19044369-19044438
19044248-19044317:19044369-19044438
19045367-19045436:19045521-19045590
19045367-19045436:19045521-19045590
19045367-19045436:19045521-19045590
19046073-19046142:19046206-19046275
19046073-19046142:19046206-19046275
19046073-19046142:19046206-19046275
19046978-19047047:19047151-19047220
19046978-19047047:19047151-19047220
19046978-19047047:19047151-19047220
19047292-19047361:19047417-19047486
19047292-19047361:19047417-19047486
19047292-19047361:19047417-19047486
19047293-19047362:19047419-19047488
19047293-19047362:19047419-19047488
19047293-19047362:19047419-19047488
19047585-19047654:19047715-19047784
19047585-19047654:19047715-19047784
19047585-19047654:19047715-19047784
19047756-19047825:19047913-19047982
19047756-19047825:19047913-19047982
19047756-19047825:19047913-19047982
19048055-19048124:19048460-19048529
19048055-19048124:19048460-19048529
19048055-19048124:19048460-19048529
19049166-19049235:19049286-19049355
19049166-19049235:19049286-19049355
19049166-19049235:19049286-19049355
19049295-19049364:19049421-1904949
19049295-19049364:19049421-1904949
19049295-19049364:19049421-190494
```

WPP+3d

```
19043814-19043883:19043931-19044000
19043814-19043883:19043931-19044000
19043814-19043883:19043931-19044000
19043814-19043883:19043935-19044004
19043814-19043883:19043935-19044004
19043814-19043883:19043935-19044004
19046073-19046142:19046206-19046275
19046073-19046142:19046206-19046275
19046073-19046142:19046206-19046275
19046978-19047047:19047151-19047220
19046978-19047047:19047151-19047220
19046978-19047047:19047151-19047220
19048055-19048124:19048470-19048539
19048055-19048124:19048470-19048539
19048055-19048124:19048470-19048539
19049166-19049235:19049286-19049355
19049166-19049235:19049286-19049355
19049166-19049235:19049286-19049355
```

RNA-Seq Junction Data

- Graphically represented in Gbrowse (consolidated but not integrated).



FB2011_02, released February 18th, 2011

D. melanogaster RNA-seq Data (Dmel R5.34)

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: 2L:19,043,861..19,050,900

Data Source ' D. melanogaster RNA-seq Data

Report & Analysis tools: Download Decorated FASTA File

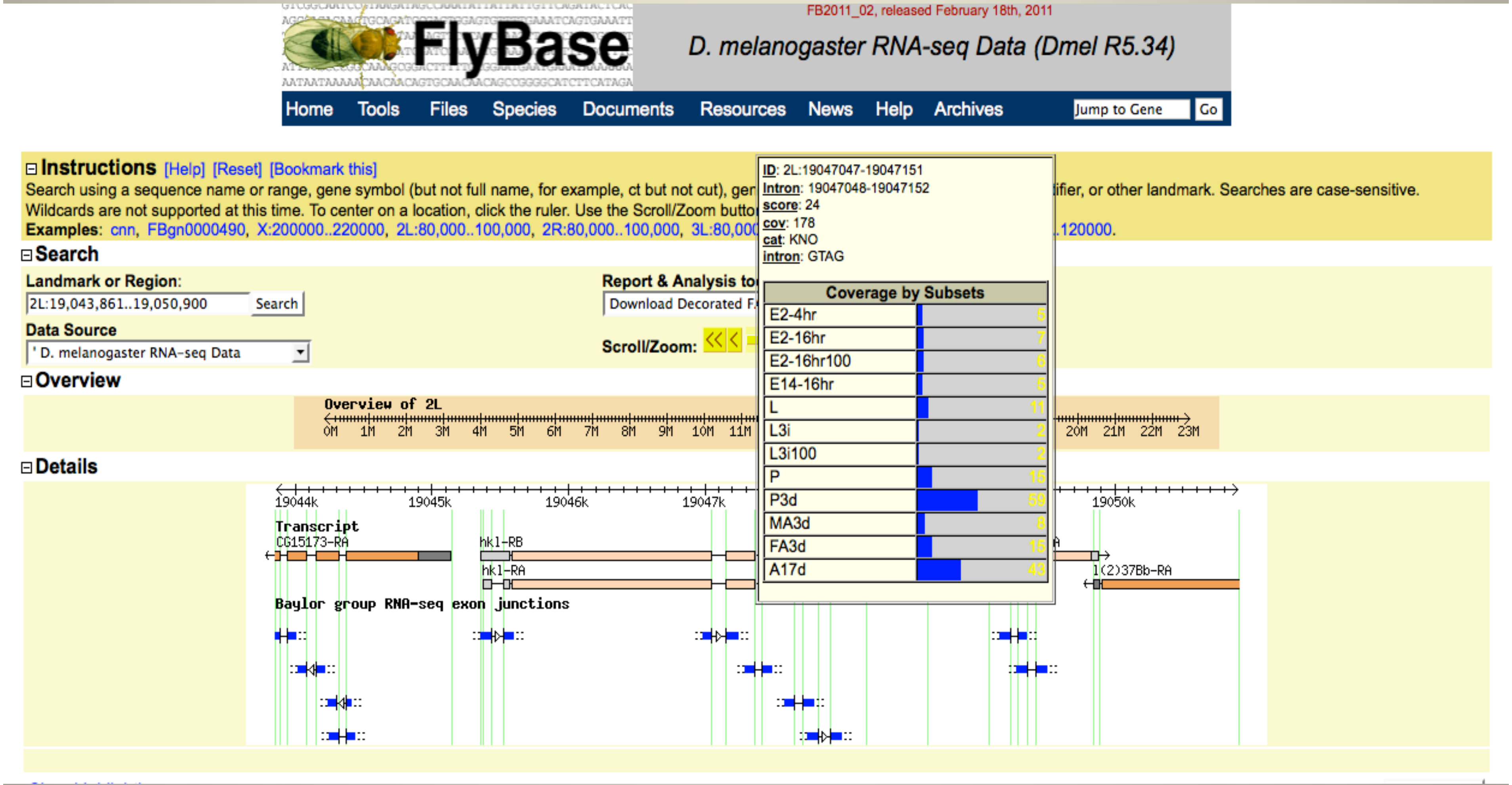
Scroll/Zoom: Show 7.04 kbp □Flip

Overview

Details

RNA-Seq Junction Data

- Graphically represented in Gbrowse (consolidated but not integrated).



FB2011_02, released February 18th, 2011

D. melanogaster RNA-seq Data (Dmel R5.34)

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

⊟ **Instructions** [Help] [Reset] [Bookmark this]

Search using a sequence name or range, gene symbol (but not full name, for example, ct but not cut), gen ifier, 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 butto

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

⊟ **Search**

Landmark or Region: 2L:19,043,861..19,050,900 Search

Data Source: ' D. melanogaster RNA-seq Data

Report & Analysis to Download Decorated F

Scroll/Zoom:

⊟ **Overview**

⊟ **Details**

ID: 2L:19047047-19047151
Intron: 19047048-19047152
score: 24
cov: 178
cat: KNO
intron: GTAG

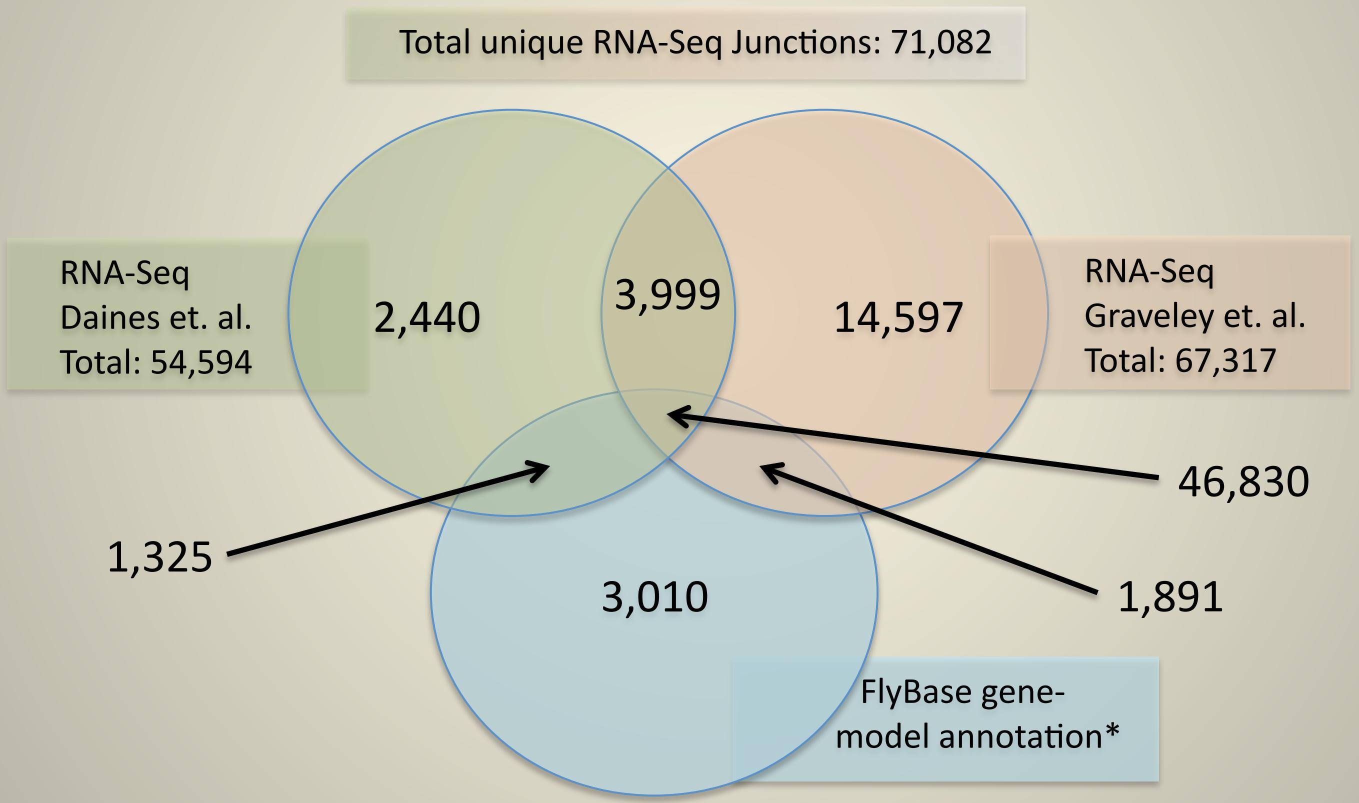
Coverage by Subsets	
E2-4hr	5
E2-16hr	7
E2-16hr100	6
E14-16hr	5
L	11
L3i	2
L3i100	2
P	15
P3d	59
MA3d	8
FA3d	15
A17d	43

RNA-Seq Junction Data

- Consolidation
  - – Create persistent exon_junction records in chado.
  - – Consolidate junctions by location.
  - – Reduces number of records required in DB.
    - E.g., modENCODE set reduces from 1.7M to 67K unique
- Integration
  - – Correlate predicted junctions with junctions from gene-model annotations.
  - – Correlate junctions with transcripts, genes, cDNAs.

RNA-Seq Junction Data

- Junctions from all sources:



Total unique RNA-Seq Junctions: 71,082

RNA-Seq
Daines et. al.
Total: 54,594

RNA-Seq
Graveley et. al.
Total: 67,317

FlyBase gene-model annotation*

2,440

3,999

14,597

46,830

1,325

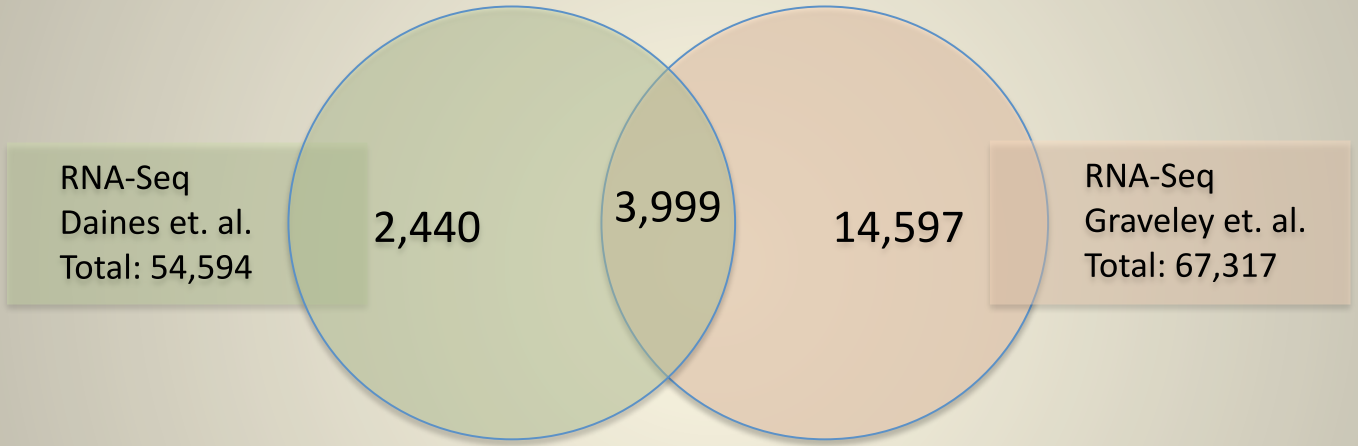
3,010

1,891

* FlyBase release FB2011_02; Dmel r5.34

RNA-Seq Junction Data

- Correlation of non-matching predicted junctions to annotated gene-model junctions*



RNA-Seq
Daines et. al.
Total: 54,594

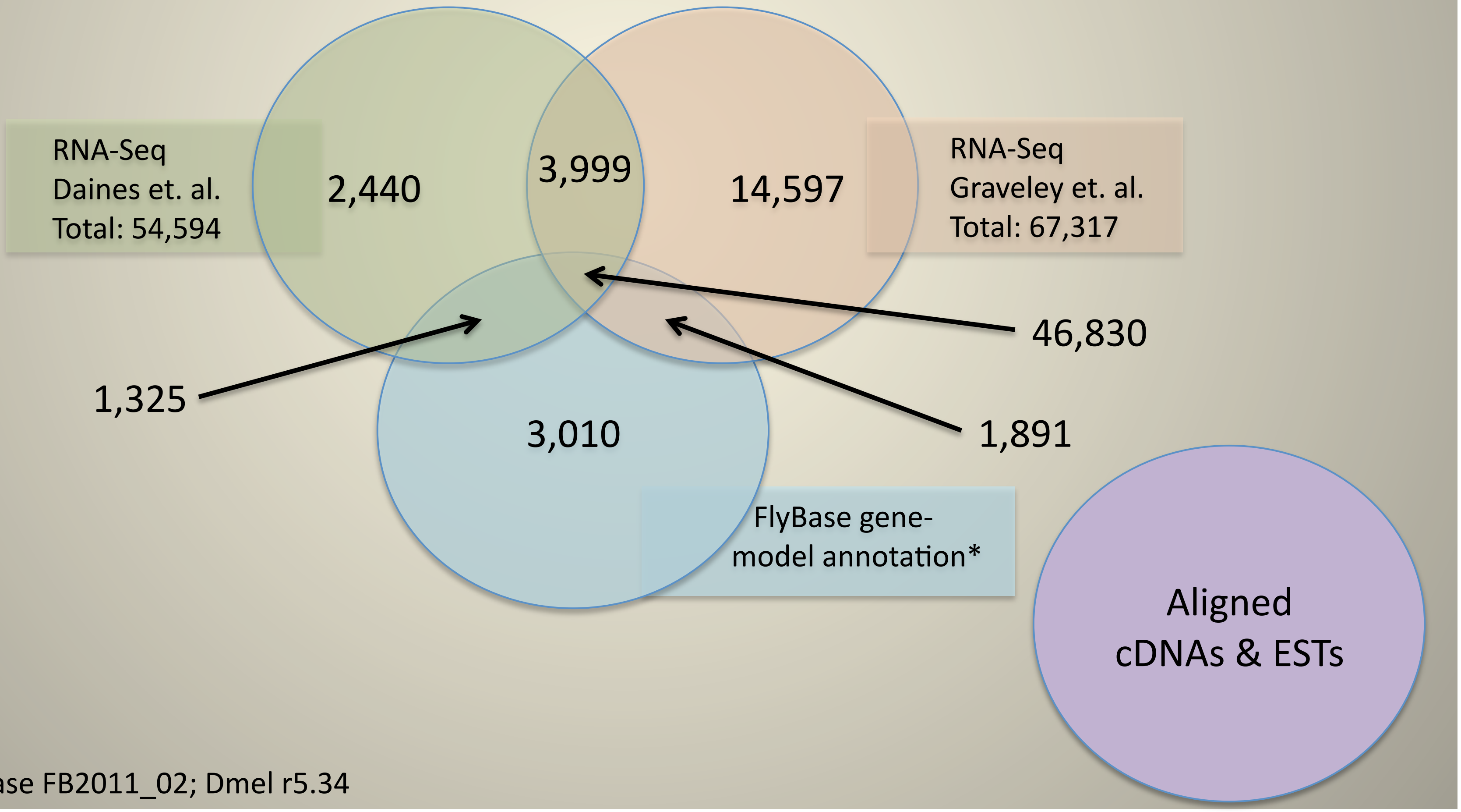
RNA-Seq
Graveley et. al.
Total: 67,317

Type of Correlation	**Count**
Predicted Alternative Junction	1794
Predicted Exon-Skip Alt. Junction	606
Predicted Alt. Junction Differing in Splice Donor Site	7488
Predicted Alt. Junction Differing in Splice Acceptor Site	5257
Predicted Novel Junction	5891

* FlyBase release FB2011_02; Dmel r5.34

RNA-Seq Junction Data

- Eventually incorporate junctions from aligned cDNAs & ESTs…



* FlyBase release FB2011_02; Dmel r5.34

RNA-Seq Junction Data

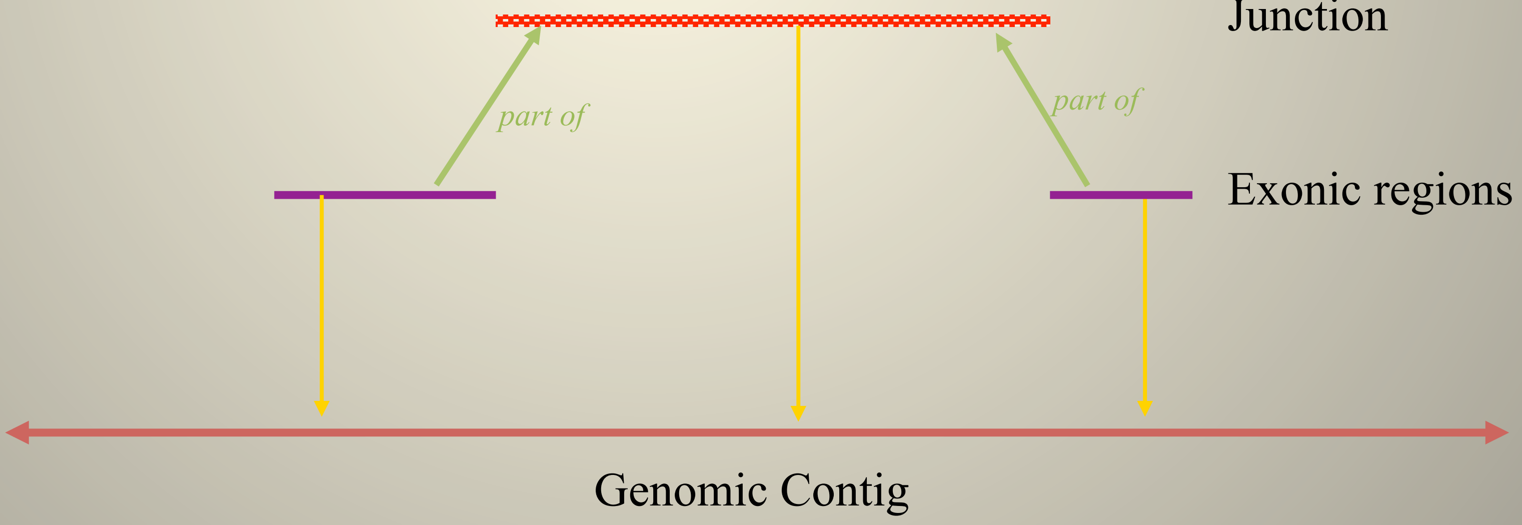
- Chado implementation (localization):

— feature.type = match

▬ feature.type = exon_junction

→ featureloc

→ feature_relationship



Junction

part of

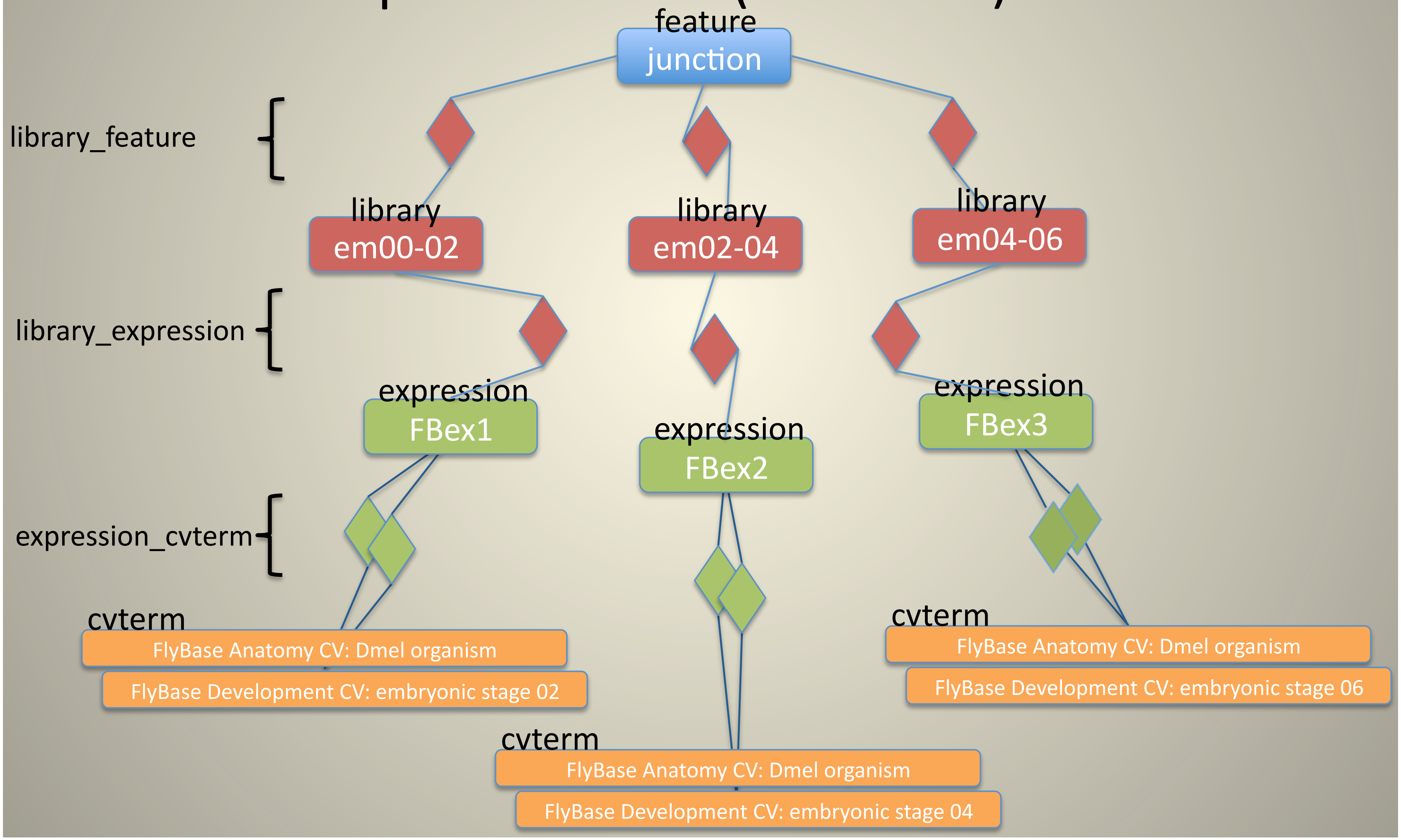
part of

Exonic regions

Genomic Contig

RNA-Seq Junction Data

- Chado implementation (metadata):



feature
junction
library_feature
library
em00-02
library
em02-04
library
em04-06
library_expression
expression
FBex1
expression
FBex2
expression
FBex3
expression_cvterm
cvterm
FlyBase Anatomy CV: Dmel organism
FlyBase Development CV: embryonic stage 02
cvterm
FlyBase Anatomy CV: Dmel organism
FlyBase Development CV: embryonic stage 06
cvterm
FlyBase Anatomy CV: Dmel organism
FlyBase Development CV: embryonic stage 04

Acknowledgements

Bill Gelbart
Jim Thurmond
Victor Strelets

Andy Schroeder
Pinglei Zhou
Josh Goodman