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

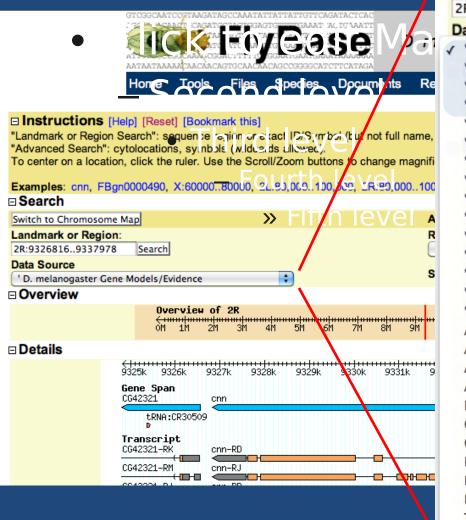
GBrowse link on top page

GTCGGCAATCOTTAAGATAGCCAAATATTATTATTGTTCA	SATACTCAC FB2012_02, released March 2nd, 2012	
	SEC A Data Data Data Di Di Di Di Schuig Gunes	& Cenumes
ANTANTANAN SURCEASE CARGE CARGA CARCAGE COGGGG CATE	CPTCATAGA	
	Cocurent Archives	Jump to Gene Go
	T3145	
D.melanogaster		FIELD M sequence
BLAST GBrowse	>> QueryEuider TermLink ImageBrowse	Batch Download
Fast-Track Your Paper	QuickSearch Simple Expression Phenotype GO References Data Type	De
FlyBase Forum	Species: Dinclude non-Dmel drosophilid species	QuickSearch help
Find a Fly Person	Search: ID/Symbol/Name All text	Search
News	°Enter text:	
modENCODE webinar MAR 1 17 Feb 12 In Memoriam: Arthur Chovnick 6 Jan 12 In Memoriam: James Crow 6 Jan 12	°QuickSearch autocomplete: Note: Wild cards (*) can be ad	ded to your search term
FlyBase 101 23 Dec 11 21 Dros genomes avail 29 Nov 11 26 new species in BLAST 18 Nov 11 2012 Release Schedule 27 Oct 11 Nobel Prize: Jules Hoffmann 6 Oct 11	Commentary See all commentaries Community curation effort a big success!	

GBrowse: Data Source default is Gene Models/Evidence

		MATATTATTATTGTTCAGATAC	TCAC	FB201	12_01, released January 201	th, 2012	
		0/Ees	ST/h=		r Com Mollele	Eviderco (A5.33	9
•	ATT COURSE CALLS OF GRACE	: Wing want a second	E I Y PC			Single and a	/
	AATAATAAAAACAACAGTG	CAACAACAGCOGGGGGCATCTTCA	TAGA				
	Home Tools F	iles Apedies Dr	ocuments F	Resources New	s Help Archives	Jump to Gen	e Go
Instructions	Help] [Reset] [Bookmark	this]					
"Landmark or Region	Search": sequence anio	poratact 12:5 mboble	ut not full name	e, for example, ct b	ut not cut), case-sensiti	ve, no wildcards.	
	cytolocations, symbols () on, click the ruler. Use the		o change magn	ification and positio	n		
	on, click the fuler. Use the		o change magn	inication and positio			
	gn0000490, X:60000800	00, QLleb, bd9.1100, b	, VR.90,0001	00,000, 3L:80,000	100,000, 3R:80,00010	0,000, 4:2000050000.	
□ Search							
Switch to Chromosome	e Map	>>> FITCh	level	Advanced Search:	Cytolocation	Start	
Landmark or Region				Report & Analysis			
2R:93268169337978	8 Search			Download Decorate	ed FASTA File	figure Go	
Data Source	no Modela (Evidence	•		Scroll/Zoom: K	Show 15.16 kt	n 🛊 🕂 🔀 🖂 Flip	
Overview	ne Models/Evidence	•					
Overview	Overview of 2R						
			н 7м ом ом	10M 11M 12M		17M 18M 19M 20M 2:	→ 1 M
		an an on on	in on si	1011 1111 1211	13/1 14/1 13/1 10/1	1711 1011 1911 2011 2.	
□ Details							
	9325k 9326k 9327k	9328k 9329k 93	30k 9331k	9332k 9333k 9	9334k 9335k 9336k	9337k 9338k 9339	++++++→ 9k
	Gene Span						
	CG42321 cnn					cbs	
	tRNA:CR30509 ₽			CG300	062 ⊒		
	Transcript						
	CG42321-RK cnn-R	H H	· · · · · · · · · · · · · · · · · · ·		062-RB ∎	cbs-RC	
	CG42321-RM cnn-R	J 				cbs-RB	
	0040004 DJ					-ka DA	

GBrowse: Data Sources



Landmark or Region: 2R:9326816..9337978 Search Data Source ' D. melanogaster Gene Models/Evidence D. melanogaster Expression/Regulation D. melanogaster MegaView ' D.melanogaster Gene Disruptions, Stocks and Reagents D. ananassae ' D. erecta D. grimshawi D. mojavensis D. persimilis D. pseudoobscura D. sechellia D. simulans D. virilis D. willistoni D. yakuba A. aegypti (mosquito) A. gambiae (mosquito) A. mellifera (honey bee) A. pisum (pea aphid) B. mori (silk worm) C. elegans (worm) C. pipiens (mosquito) I. scapularis (deer tick) N. vitripennis (wasp) P. humanus corporis (human body louse) T. castaneum (red flour beetle)

GTOGGCAA	ICOTTAAGATAC AAGIGCAGATG	GCCAAATA CCAACTEGA	PTATTATTGTTCA STG <u>TTTT</u> GAAATO	GATACTCAC AGTGAAATT			FB2012_0	01, release	ed January	20th, 2012		
	Thi ATO ATO	AGTI PUD ATCULA ACTTI ITS		se			Ge	ne Dr	nel\chi	nmo		
Home	Tools	Files	Species	Documents	Resou	rces	News	Help	Archive	S	Jump	to Gene Go
Profile Ma	inager		. 4	0					[Help	Open All	Close All
General	Information	1										
Symbol		C	mel\chinmo			Spec	ies		D. mela	anogaster		
Name			hronologicall	y inappropriate s		Anno	tation syn	nbol	CG316	66		
Feature	type	p	rotein_coding	_gene		FlyBa	ase ID		FBgn0	086758		
Gene M	odel Status	C	urrent			Stock	availabili	ty	14 pub	licly avail:	able	
Also Kn	own As	C	G17156									
Genomi	c Location											
Chromo	some (arm)) 2	L			Reco	mbination	map				
Cytoger	netic map	2	2A5-22B1			Sequ	ence locat	tion	2L:1,65	51,2601,	704,682 [+]	
View in modEN0 GBrow	/iew: (2) Models/E + GBrowse CODE	•		>>						6	corated FastA et genome re ne region Get FastA	gion
	mary Inform ent Updates											

GBrowse: Expression/Regulation

	Data Source				Scroll/Zoom: < <	Show 57.42 k	br 🗘 🕂 🔀 🕞		
		Expression/Regulation	•		Scroii/20011.	3110W 37.42 K		ΠP	
	F. Over /ie'.4	0 erviet u							vles
		4	tunnents, sund mus mus	the second se	un the set the set	15M 16M 17M 10	n uturunt an um di 1 azu 2v1 2a1 zzi		
	□ Details								J
	Details	(+ + + + + + + + + + + + + + + + + + +	.660k	16 '0k	1680k	1690k	1700k	· · · · · · · · · · · · · · · · · · ·	
		`er⊭Sjan	.0000	Tf. OK	1000K	1030K	17006		
	DEC	chi mu						cpb	gene
		Transcript chinmo-RH						cpb-RA	J
		chinmo-RH chinmo-RE				0		Cpb-RA IH⊟D→	exter
	•	ch nho RG							
		chirmo-RD							anno
				chinmo-RF		O			
				chirse-R I		O			trans
		Natural 16 (Soly	, visible below 33	сакь	blood{}289				trans natura
		Repeat region							Ilacuic
		cDNA 🔊							
		BT001697						AY069829	
		BT050405							cDN/
				AY060464					
		Tissue Expressio	on [Affy2] (FlyAt]	1636985_s_at			1629484_s_a	t <u>16</u> 305;	
					1		ante anno .		
									- FIVAt
				1636477_at			1628005_a	t	
							, d. Mill		FlyAt level
		Developmental st	tage subsets, unio	que reads (modENCODE	E Transcription Gro	up)			ovor
		LAM				7		\vvvvvvvv	expr
		"IV" N Preve	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	~	mmmn l			MAX em0-2	
					•	· //	1-7-5- P-7-1 VAV	TAN art-Sh	
			a phat is a state of the	to a state and the state	A AT			em10-12 em12-14 em14-16h	modl
								em18-20hr em20-22hr	
					•••			L1 L3_12hr	deve
				• • • • •	•		That a should be		
		****			1		S-WAWAYA	HPP 12hr HPP 24hr HPP 2days	timo
		N-W		<u>in :</u>			A STATISTICS	HPP Jaays HPP Jaays Add Ecl Idays	time
								Add Ecl 30days Adf Ecl 1days Adf Ecl 1days	
								Rdr_Ec1_30days	

gene extents annotated transcripts natural TE's

cDNAs

FlyAtlas tissuelevel expression data

modENCODE RNA-Seq developmental timecourse

GBrowse: Expression/Regulation

	Data Source			Scroll/Zoom: <	Show 57.42 k	br 🗘 🕂 😕 🕞 Flir		
	D. me anogaster E	x~ression/Regulation		301011/20011.	SIIUW 37.42 K		,	
	NOVEL NP X	0 erviet or zi						yles
			i ⊶n 5n 6n 7N 8M 9h	1011 11M 1211 1311 14	15M 16M 17M 10	n utunut an unu dan 1 1311 2011 211 2211	"23M	y ICJ
	□ Details							
((560k 1ť 0k	1680k	1690k	1700k	+++++>	
	XACI	`er;Sian chimo					cpb	aana
		Transcript						gene
		chinmo-RH			Oł		cpb-RA DH⊟D→	exter anno
		chinmo RE			O 			exter
	-	ch nmo RG I─ I── ── ── ── chirwo-RD			D ł			anno
		H	chinmo-RF		D			
			chinzo-R (O			trans
		Natural 18 (Snly visi	ille beluw CCOKb;		LI			natura
		- · ·		b1ood{}289				natura
		Repeat region cDNA						
		BT001697					AY069829	
		BT050405						cDN/
		Ti	AY060464					
		Tissue Expression [Af	1636985_s_at			1629484_s_at	16305;	
						- and a state -		
								FlyAt level
			1636477	_at		1628005_at		
								level
		Developmental stage s	subsets, unique reads (modENG	ODE Transcription Gr	up)			
	1	ំ សំភា			_ 7			expr
		1 Khomen	havenner a share and the	mmm	- And have	~~		CAPI
							em0-2 em2-4t em4-6h	
			and a high and			Law The Will	em8-10+ em10-12 em12-14+	modl
		an a		AR 0	Albury Dreek	- And	em14-16h em16-18hi em18-20hr	mou
							ep22-24hr L1 L2	deve
	2							ueve
		<u> </u>					HPP_12hr HPP_24hr	L :
	1		<u></u>			NUT AND A A	HPP Adays HPP Adays AdM_Ec1_Idays	time
	6						AdM_Ecl_30days AdM_Ecl_30days AdF_Ecl_1days AdF_Ecl_5days	
	-	······································	······································				tdr_Ec1_30days	

gene extents annotated transcripts natural TE's

cDNAs

FlyAtlas tissuelevel expression data

modENCODE RNA-Seq developmental timecourse

GB: Expression/regulation data options

• Cli	c k to	edit.	Ŵ	aster	tex	HPP HPP HPP HPP HPP HPP HPP HPP HPP HPP	yles yles
Clear highlighting							Update Image
[Help and key for FlyBase 6.	then se ther all						
□ Tracks							
Reference Genome Ar	notations (Iso-1) Select all	lear all					
Gene Span						Repea	t region
Transcript	Ni tu al 🗡 c	only is ole by love (00.1b)					
General Select all Clear	all						
3-frame translation (fo	rward)	DNA/GC (Content	0	3-frame translation (reve	erse)	
Aligned Evidence Selection	ct all Clear all						
I cDNA	mRNA				q exon junctions		
EST	other aligned sequences			Peptide	Atlas peptides		
Previous Gene Models	Select all Clear all	> Fifth lev					
Recently deleted trans	scripts						
Mapped Mutations Sel							
Transgenic Insertion S							
Noncoding Features							
Insulator class I	Silencer	TFBS - homeodomain		Chromatin Domains (5-sta	· · · · · · · · · · · · · · · · · · ·	-	ive Brain Enhancers (Pfeiffer et al)
Insulator class II	Regulatory Region	TFBS - helix-loop-helix dor		Chromatin Domains (9-sta		RNA	Editing Sites
Protein Binding Site	TFBS - HOT spot analysis	TFBS - BTB/POZ domain		Chromatin Domains (9-sta	te model, BG3 cells)		
Enhancer	TFBS - zinc finger domain	TFBS - other	6	Origin of Replication			
Microarray Features	Gelect all Clear all						
Affymetrix v1				Affymetrix v2			
	Clear all						
Timecourse [cDNA] (A		Timecourse [ampli]	cons] (Gau	ihar et al.)	Tissue Expre	ession [Af	fy2] (FlyAtlas)
Expression Levels: RM	·						
Developmental stage s			_	culture cells (modENCODE			Treatments/Conditions, by strand
	subsets, unique reads (modENC		Tissue	culture cells,by strand (modE	NCODE Transcription Gr	oup)	
	VA-Seq by Tissue Select all						
Digestive system		Imaginal disc and other c	arcass		Gonads and male acc	essory gla	inds
Fat body and salivary	·	CNS and adult head					
BLAST hit Select all Cl	ear all						
Blast HSP							
							Configure tracks Update Image

Select from Aligned Evidence: RNA-seq exon junctions

'Select all' from Expression Levels: RNA-Seq by Tissuelick to edit Master text styles

E General Select all Clear all						
□ 3-frame translation (forwar)		RC Content	🗌 3-fra	ame translation (revers	se)	
✓cDNA □mRNA		6	RNA-seq ex	on junctions		
EST otheraligned seque no		0	PeptideAtlas	s peptides		
Previous Gene Models Select all Clear all						
Mapped Mutations Select all Clear all						
Noncoding Features Select all Clear all						
Microarray Features Select all Clear all	CIELL L					
Expression Levels: MicroArrays Select all Cle						
Timecourse [cDNA] (Arbeitman et al.)	Timecourse [amplicons] (Gauhar et al.)		🗌 Tissue Expressi	on [Affy2] (FlyAtlas)	
Expression Levels: RNA-seq Select all Clear a	п					
Developmental stage subsets (Baylor)		Tissue culture cells (mo	dENCODE Tr	ranscription Group)	Treatments/Con strand	ditions, by
Developmental stage subsets, unique reads (mo Transcription Group)	dENCODE	Tissue culture cells,by s Transcription Group)	trand (modEf	NCODE		
Expression Levels: RNA-Seq by Tissue Select	t all Clear all					
☑ Digestive system	Imaginal disc and c	other carcass	🗹 Go	nads and male acces	sory glands	
Fat body and salivary glands	CNS and adult hea	d				
BLAST hit Select all Clear all						
Blast HSP						
					Configure tracks	Update Image

GBrowse: RNA-seq exon junctions

$\leftarrow + + + + +$		+ + + + + + + + + + + + + + + + + + +	-+ + + + + + + + + + + + + + + + + + +	1700k
Gene Span chinmo				
Transcript chinmo-RH	_			
chinmo-RE				
chinmo-RG	•			
chinmo-RD				
	chinmo-RF			
	chinmo-RA	· 		
Natural TE (only visible below 300K	★ blood():289		
		p100d{}209		
Repeat regio	n			
CDNA BT001697				
BT050405				
	AY060464			
RNA-seq exon	junctions	1		
	-			:
	N			::
	r -1			-
	:}			::#::
		:∔		
			:++::	

Sequence feature reports for RNA-Seq exon junctions

TOGGCANTCOTTANGATAGCCAAN GC CA TAA TIGCAGA TOGAATAG	ATTATIATIGTICAGATACICAC	FB2012_01, released	d January 20th, 2012	
	VB22561+		re: eາv/n ກ່າວເ	ctv/k
NTATIONAL CARGO GA TITTI I ATAATAANAA CAACAGIGCAA	ACAGCOGGGGCATCTTCATAGA	IUSCU		JUYIY
Home <u>Tools</u> Files	Species Documents R	esources News Help	Archives	Jump to Gene Go
Protile Manager			Help Ope	n All Close All
	_ + = Ø		neip ope	
General Information	Denet: 5:21 1653835:1600117:+	Species	D. malanagastar	
Symbol Feature type		Species FlyBase ID	D. melanogaster FBsf0000101966	
Collection	exon_junction modENCODE_mRNA-Seg_U_junction			
	BGM_1_RNAseq_junctions		chinmo	
Genomic Location	Tourchieve			
Chromosome (arm) Map (GBrowse)	2L	Sequence location	2L:1,653,8351,690,11	7 [+]
	chinm chinm chinm chinm chinm chinm	o-RH o-RE o-RG	80k 1690k 	
 Recent Updates Sequence Data 				
Length				
Comments	GTAG			
Sequence				
ocquence				

	5	
Experimenta	al Data	
🗆 RNA-Seq Re	ead Counts	
E Library: BC	M_1_RNAseq_junctions	
🗆 Library: mo	dENCODE_mRNA-Seq_U_junctions	
modENCODE_r Seq_U_junction		Read Col
	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
Experimental Data Collection Information	second instar larval stage	52
Collection	Information	
	n: modENCODE_mRNA-Seq_U_junctions	

Symbol

modENCODE_mRNA-Seq_U_junctions

New 5' exon (UTR) added to Cht6

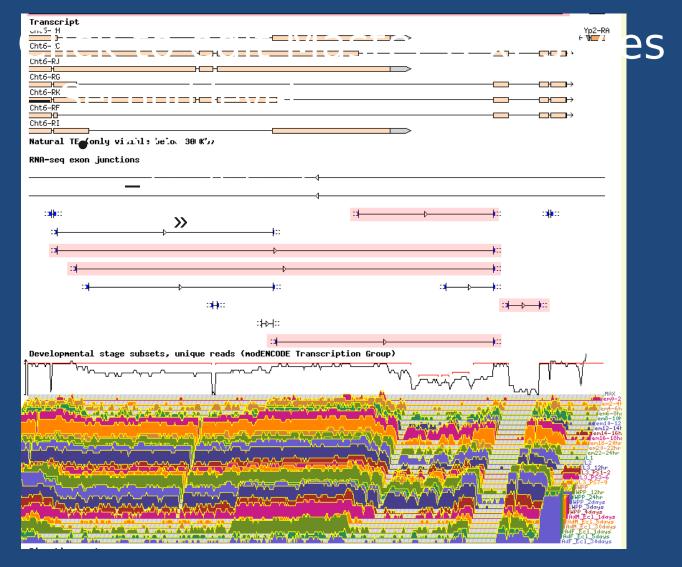
e : pan		
nscript	Cht6-RH	
	uhtfinka	
-	2:46 C	
	Cht6-RJ	
	Cht/ -R(
•	0	
	0 Cht6-RI	
unal TE (a	0	
eat region -		
A	»	
-seq exon	junctions	
	7	
		10 012
		:=+=
		20 4 022
		: ::
1	V	:::
elopmental	stage subsets, unique reads (modENCODE Transcription Group)	
•	7	<u>ורת היוראי</u> ר
~~ [~]		
<i>, , , , , , , , , , , , , , , , , , , </i>		
1		
- and -		
	······································	

New 5' exon (UTR) added to Cht6

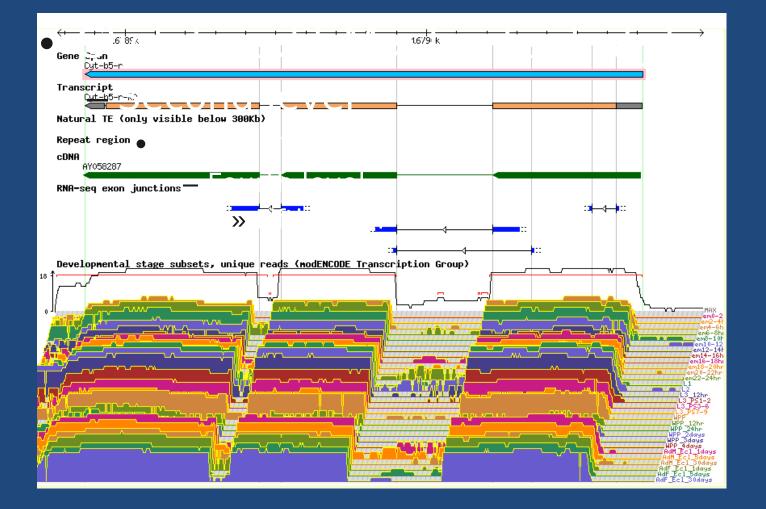
2a i	9917 ──18k 9919k 9920k 9921k 9922k 9923k 9924k 9 ⁻² 5k 997 5k C
ript Cht6-RH	
инсо-кп 0	
Chite - Chi	
Cht6-RJ	
Cht6 -RG	
•	
0 Cht6-RI	
TE (only visible below 340Kb)	
-	
region	
»	
exon junctions	
:•	↓::
	50 4 00
	··· · · · · · · · · · · · · · · · · ·
- I	
U U U U U U U U U U U U U U U U U U U	10 H 00
•	
	a
mental stage subsets, unique reads (modENC	:ODE Transcription Group)
4	
	end-fh end-fh
	English (1997)
	em14-16h em16-18h
	A Constraint of the second sec
	And

Gene merge: Cht6 + downstream genes

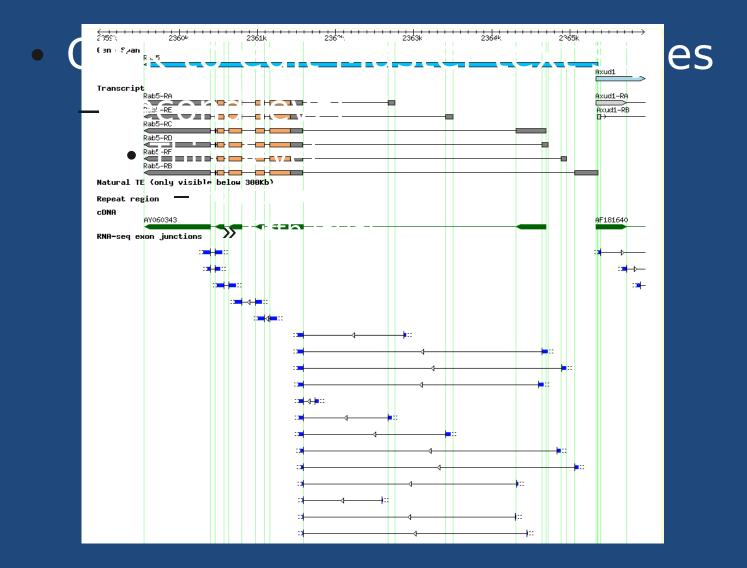




Junctions: low frequency junctions may not be annotated



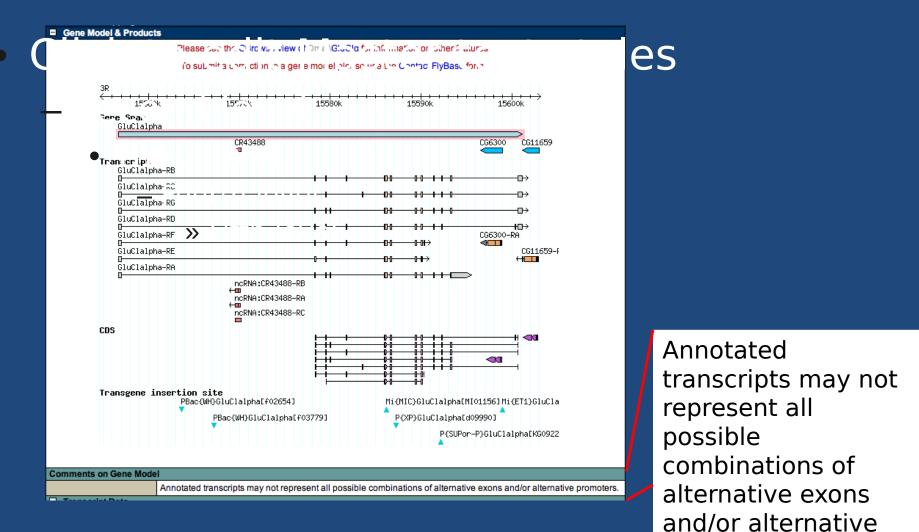
Junctions: variations in 5' UTR may not be annotated



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)



promoters.

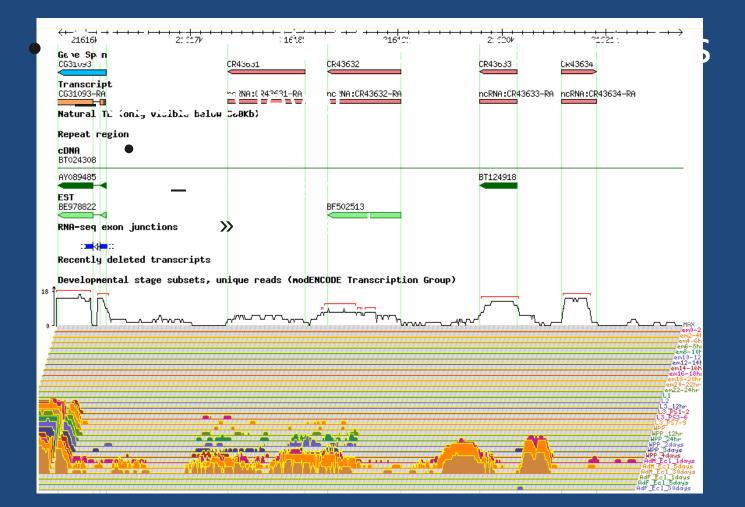
GBrowse: expression data

<	16 30k	16 [°] 0k	1700k
Gene Span chinmo			
Transcript chinmo-RH			
chinmo-RE			
chinmo-RG			
chinmo-RD			
chinmo-RF			
chinmo-RA D			
Natural TE (only visible below 300kb)	I		
-	blood{}289		
Repeat region			
CDNA			
BT001697			
BT050405			
AY060464			
RNA-seq exon junctions	•		
↓ ↓	::		# 12
:			:#::
		::	
		:;	

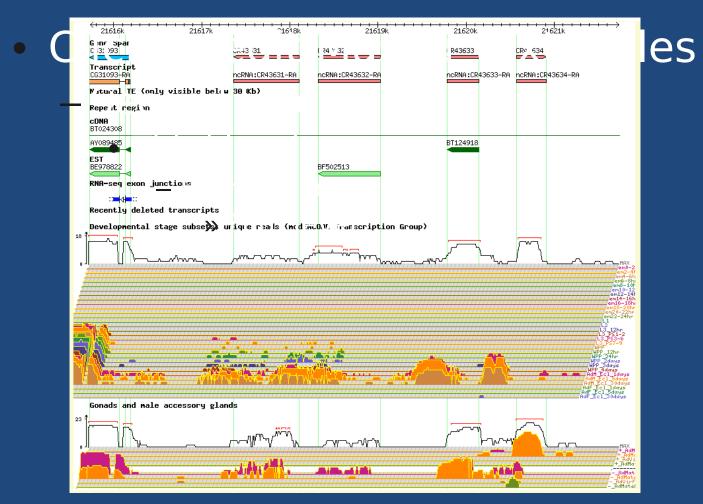


strand-specific RNA-Seq data from isolated L/A tissues

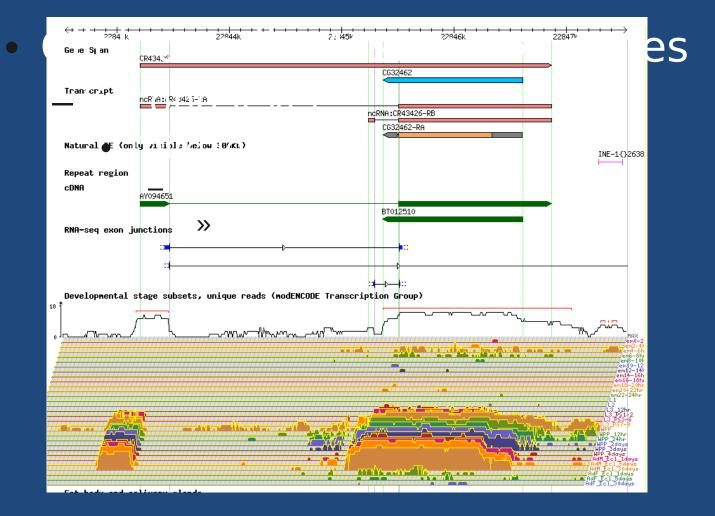
New genes: ncRNA genes



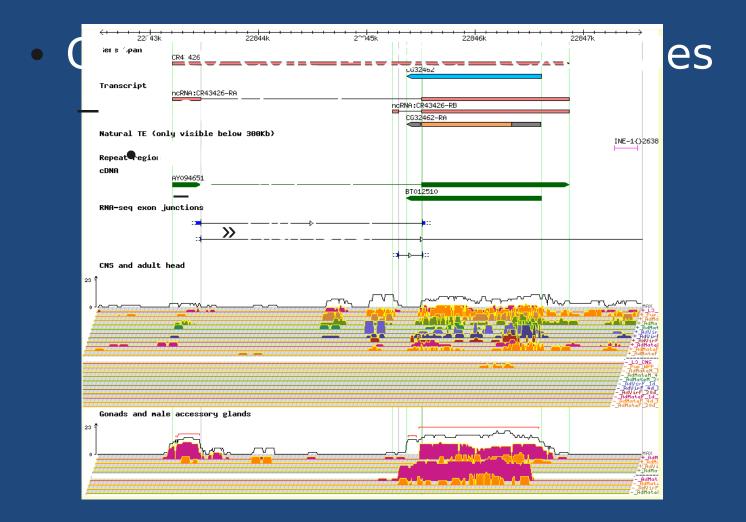
Stranded RNA-Seq data: ncRNA genes



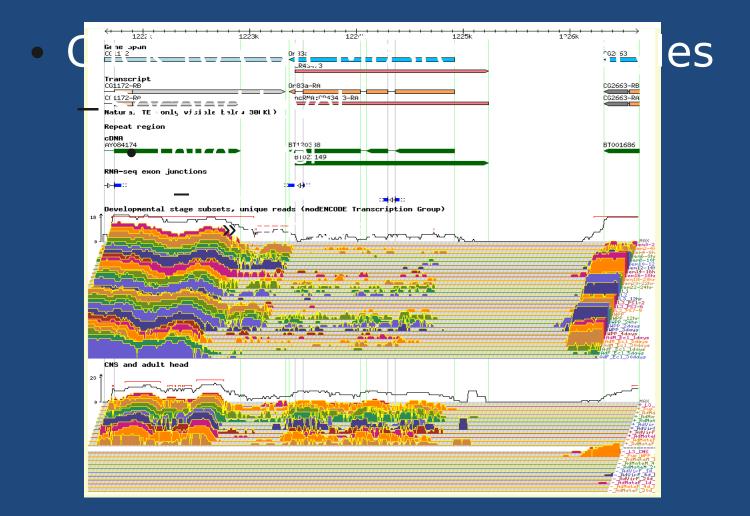
New genes: anti-sense RNAs



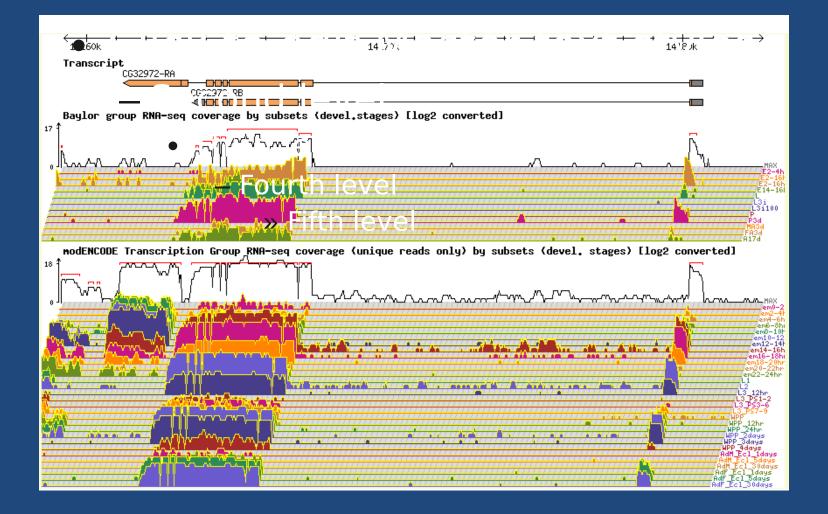
Stranded RNA-Seq data: anti-sense RNAs



Stranded RNA-Seq data: anti-sense RNAs



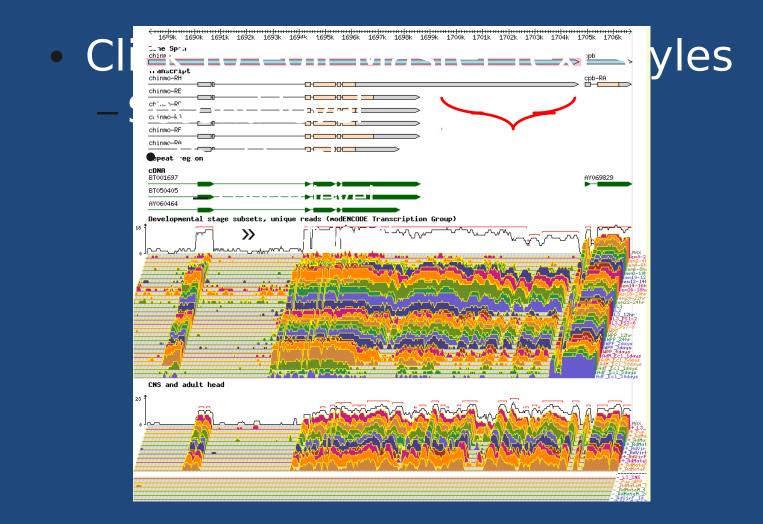
Gene split: CG32972



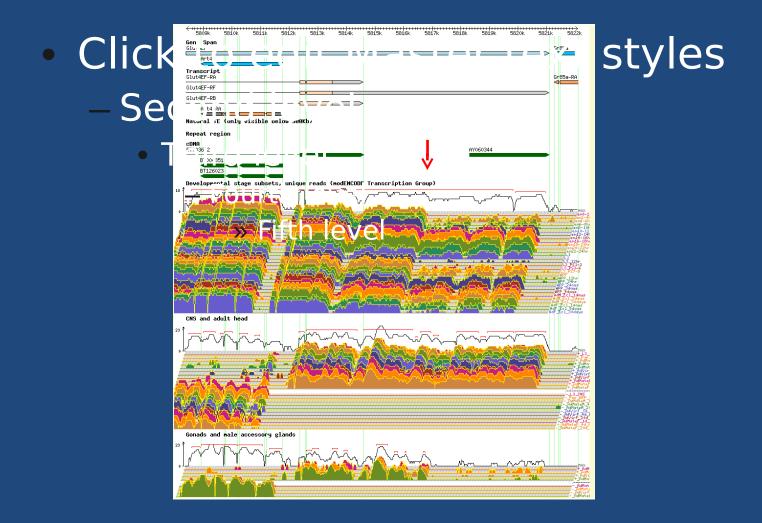
Result of split: CG43332 + CG43333



Improved annotations: extended 3' UTRs



Intermediate 3' UTR extensions not annotated



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

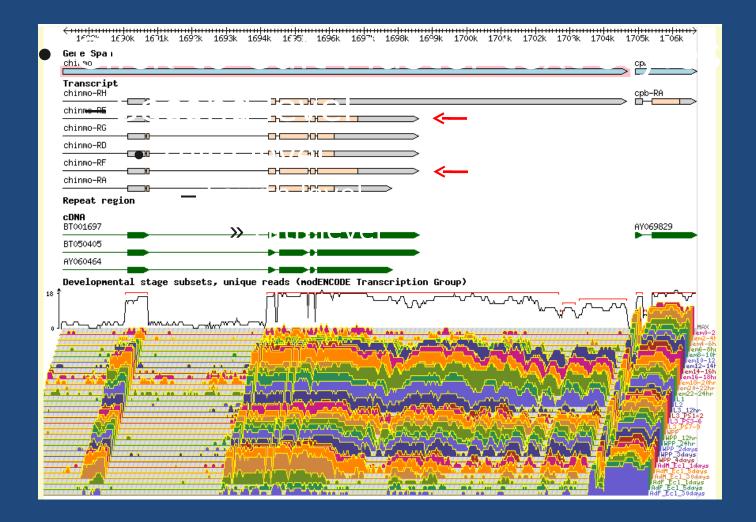
Impact of RNA-Seq data Major source of reannotation flags:

-Unannotated junctions (complete categorized and prioritized list)

-Targeted MIP cDNAs

-RNA-Seq coverage data without corresponding annotation

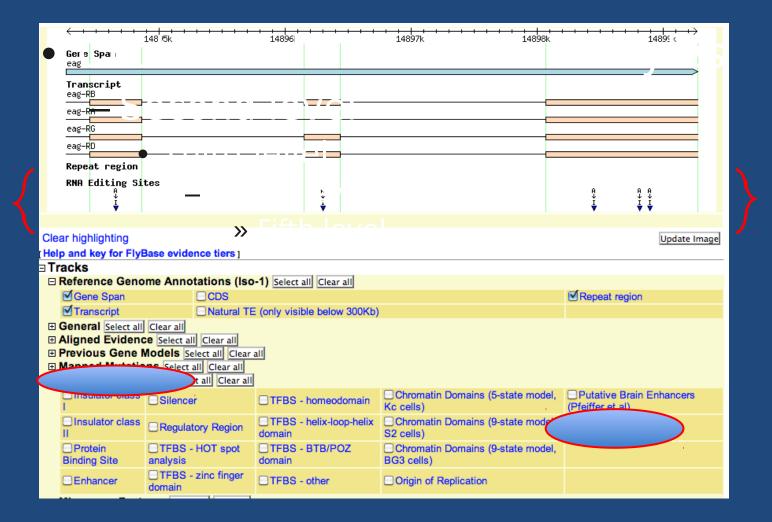
New data: stop-codon readthrough



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



GTUGGUNTUUTINGATAGUNAN AGGE GENAATGUNANTGUNAN	TATTATTATTGTTCAGATACTCAC GAGTGTTTTTGATATTCAGTGAAATT	FB2012_01, rele	ased January 20th, 2012
ATT T	yBase s	equence Feature: m	odified_RNA_base_feature
Home Tools Files	s Species Documents	Resources News Help	Archives Jump to Gene
Profile Manager	• + - 0		Help Open All Close Al
General Information			
Symbol	A-I_edit_000938	Species	D. melanogaster
Fea ure tyje	mildified_ RNA _helse _felsture		FB #01:00.734202
Collection	mE_A-to-I_RNA_Editing_Sites	Associated gene(s)	
Genomic Location	У.	Servence legation	VitA 808 800 44 808 800 til
Chromosom (arr)		Sequence location	X:14,898,80914,898,809 [+]
Length			
Comments	Gite of Aria Ted ting in CDS: At nits	o Asp substitution.	
Sequence			
Associated information	Eourth lev	/e	
Gene(s) (targeted or local)	eag		
Allele(s)	>>>		
Transcripts(s)	eag-RA eag-RB eag-RD eag-RG		
Experimental Data	•		
RNA Editing Sites	A Editing Citor		
Library: mE_A-to-I_RN	Expression stage(s)	Read Count	Editing Frequency
I_RNA_Editing_Sites		0	N/A
	embryonic stage 1 4		
	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:200000220000, 2L:80,000100,000, 2R:80,000100,000, 3L:80,000100,000, 3R:80,000100,000, 4:10000012
--

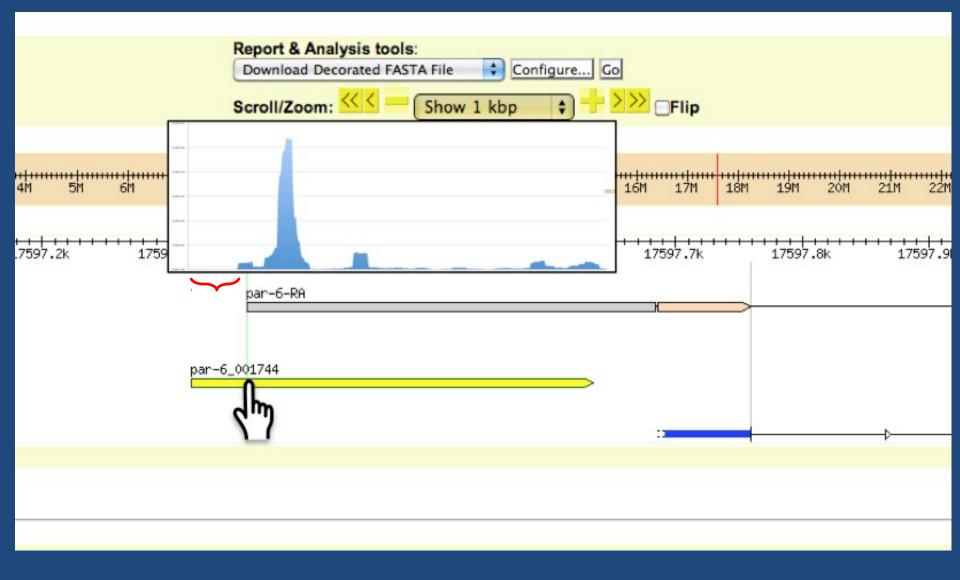
⊟ Search			
Landmark or Region:	Report & Analysis tools		
X:1759697617597975 Search	Download Decorated FAS	TA File 🗘 Configure Go	
Data Source	Scroll/Zoom: < < 💳	Show 1 kbp 🔹 🕂 😕 🕞 Flip	
('D. melanogaster RNA-seq Data	Scion/200m.		
Overview			
Overview of the second	x manufilia di anti di a m 3m 4m 5m 6m 7m 6m 9m 10m 11m	12M 13M 14M 15M 16M 17M 18M 19M 20M	21M 22M
□ Details			
Clear highlighting [Help and key for Stock Center Data evidence ti] Tracks Clear all Clear all	par-6-RA par-6_001744	17597.6k 17597.7k 17597.8k	· · · · · · · · · · · · · · · · · · ·
Chromatin Domains (5-state model, Kc cel	Ils) Chromatin Domains (9-state model,	S2 cells) Chromatin Domains (9-sta	te model, BG3 cells)
Gene Model features Select all Clear all			
Gene Span		✓ Transcript	
Genome reagents and data Select all Clea	ar 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 conv	/erted]	ssion 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 Set	lect all Clear all		
digestive system (by strand) [log2 converted	ed]	and) [log2 converted]	glands (by strand) [log2 converted]
fat body and salivary glands (by strand) [lo	og2 converted] CNS and adult head (by strand) [log2 converted]	pnverted]	
External Annotation Tracks Select all Cle	ear all		

Configure tracks... Update Image

Transcription start sites range represented as extent on genome

Click to edit Moster to to the styles Nerview of X ON 1H 2H 3H 4H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	++>
Overview of X or 1h Or 1h <th>++-> </th>	++->
Construction Construction<	++>
Control Control <t< th=""><th>+++> </th></t<>	+++>
(+++++++++++++++++++++++++++++++++++++	>
Transcript CG8188-RA	>
CG8188-RB	
TSS CG8188_001743 → pa ~-€_1001714	
RNA-seq exon junctions	
ck Center Data evidence tiers j	
Select all Clear all	
nains (5-state model, Kc cells) Chromatin Domains (9-state model, S2 cells) Chromatin Domains (9-state model, BG3	cells)
tures Select all Clear all	conc)
✓Transcript	
Its and data Select all Clear all	
□aligned cDNA □aligned mRNA	
es Select all Clear all	
pson Transgene insertion site Estimated Cytological band	

TSS data: frequency distributions



TSS annotation at 90% point

	par-6
as ■ cumulative frequency	
■ 90% mark ■ frequency	
	Gray = cumulative frequency Red = 90% point Black = frequency at each
	nucleotide
D2 -	

TSS annotation at 90% point peaked promoter class

1	TwdIL	
03	 cumulative frequency 90% mark frequency 	
0.6	Gray = cumulative frequency Red = 90% point Black = frequency at each	
	nucleotide	
02		

TSS annotation at 90% point broad promoter class

1			
Pot Area			CG10435
0.8 -	cumulative frequency		
	90% mark		
26 -	■ frequency		
		cumulative freque	ncy
		90% point	
	Black =	frequency at each	1
	nucleo		
02 -			

TSS annotation at 90% point broad promoter class

18 -		tyn (SP71)
	Cumulative frequency	
	90% mark	
	■ frequency	
0.6		
	Gray = cumulative frequency Red = 90% point Black = frequency at each	
0.4 -	nucleotide	
0.2 -		

- 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

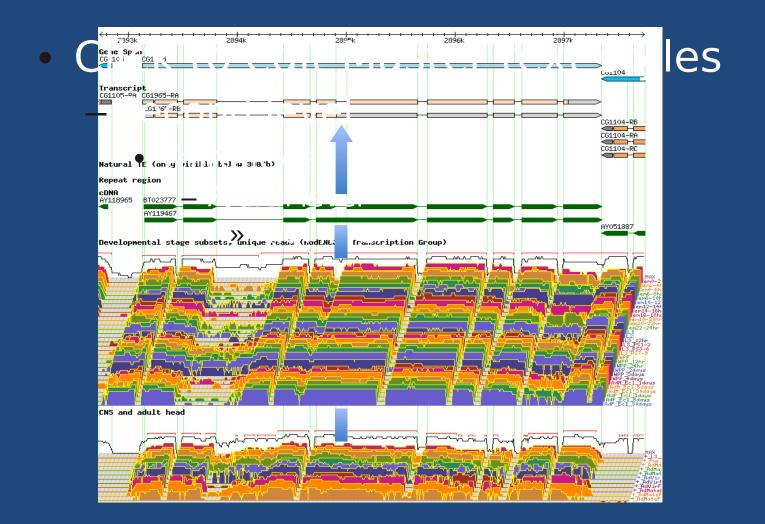
A-to-I RNA editing

Graveley, et al. (2011) Nature 471:473-479

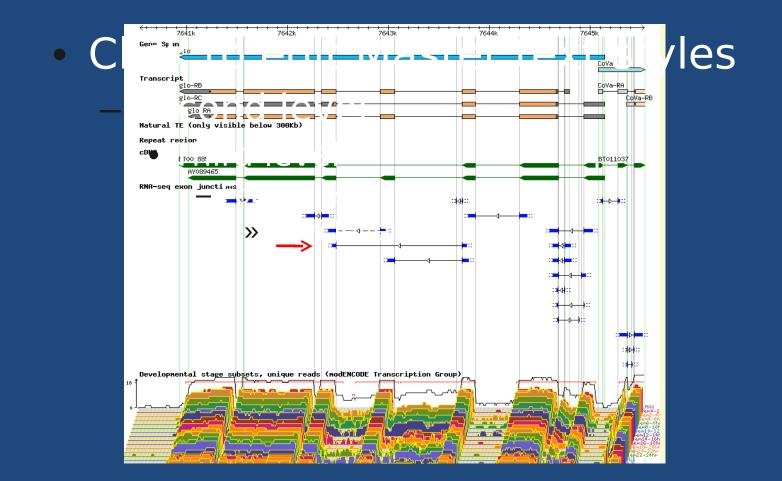
Transcription start site data

Hoskins, et al. (2011) Genome Res. 21:182-192

Transcripts that result in truncated polypeptides: retained introns



Transcripts that result in truncated polypeptides: alternative splices



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 noncoding 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



• ade3

• svr

• CG344 22

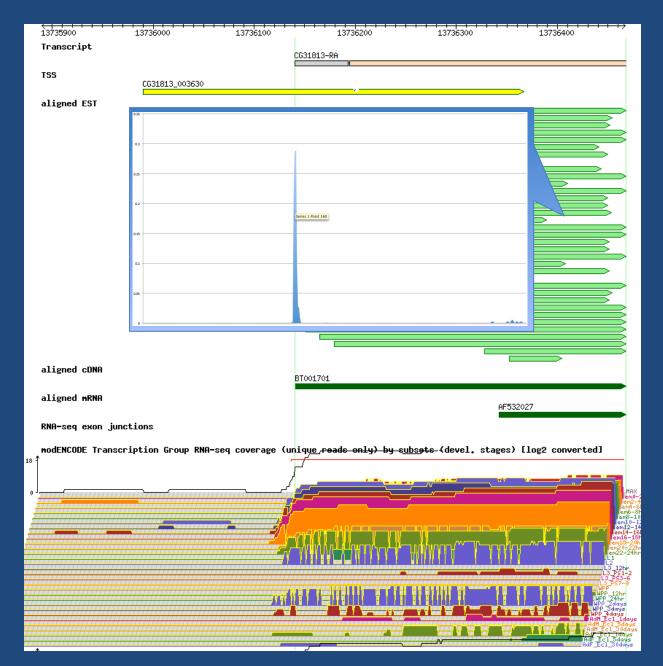
psq

scf
AGO
2
osk

Gene annotation using modENCODE TSS

13735900	13736000	13736100	13736200	13736300	13736400
Transcript			CG31813-RA		
TSS	CG31813_003630				
aligned EST					>
arighed LJI					
				>	
				<u> </u>	
aligned cDNA					
attRuen cowe			BT001701		
aligned mRNA					
				AF5	32027
RNA-seq exon ju	Inctions				
modENCODE Trans	cription Group RNA	-seg coverage (unique ,reads only) by s	ubs ets -{devel. stages) [log2 converted]
18		- - - -	·		
			/		
0 1					HAX em0-
			Ţ	_	en4-6 en6-8 en6-10
					iem10-12 iem12-14 iem14-16
					em18-18h em18-20h en20-22h
					L2 L2 L2 L2 L2
				19999997 	L3 PS1-2 L3 PS3-6 L3 PS7-9
					HPP 12hr HPP 24hr
					HPP 2days HPP 3days HPP 4days
					Adh Ecl Sdays Adh Ecl Sdays Adh Ecl Sdays
*					Adr Ec1 5days Adr Ec1 30days

Gene annotation using modENCODE TSS

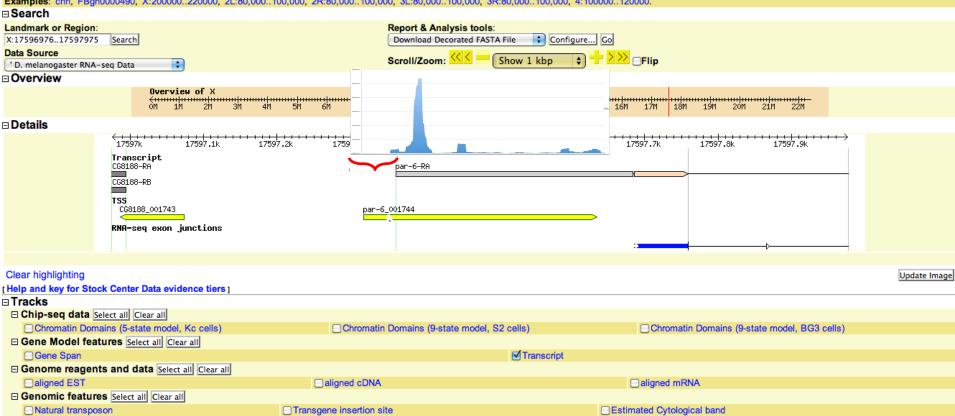


modENCODE TSS: FlyBase GBrowse

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.



Insulator class II

imaginal disc and other carcass (by strand) [log2 converted]

CNS and adult head (by strand) [log2 converted]

expression under treatments (by strand) [log2 converted]

modENCODE Transcription Group RNA-seg coverage

modENCODE Transcription Group RNA-seq coverage by

(embryonic stages, by strand) [log2 converted]

subsets (tissue culture cells) [log2 converted]

Mapped features Select all Clear all
Insulator class I
Other RNA-seq data Select all Clear all

E RNA-seg data Select all Clear all modENCODE Transcription Group RNA-seq coverage (unique RNA-seq exon junctions reads only) by subsets (devel. stages) [log2 converted] Baylor group RNA-seq coverage by modENCODE Transcription Group RNA-seg coverage (all reads) subsets (devel.stages) [log2 converted] by subsets (devel. stages) [log2 converted]

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

digestive system (by strand) [log2 converted] [] fat body and salivary glands (by strand) [log2 converted]

External Annotation Tracks Select all Clear all

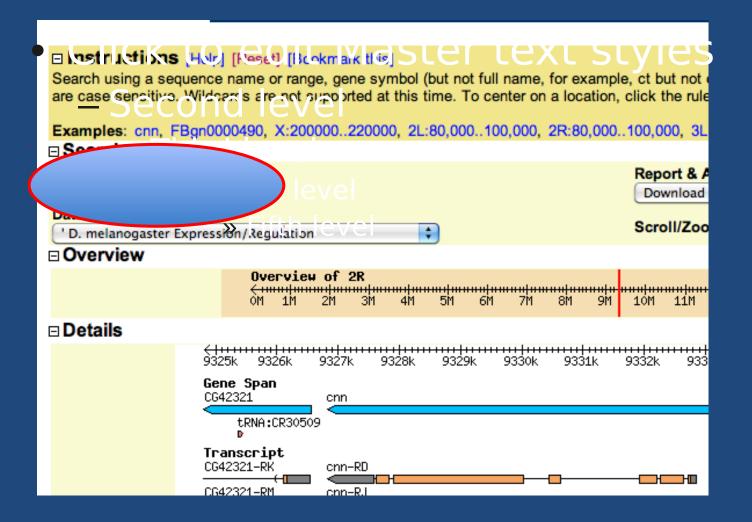
cell lines expression (by strand) [log2 converted]

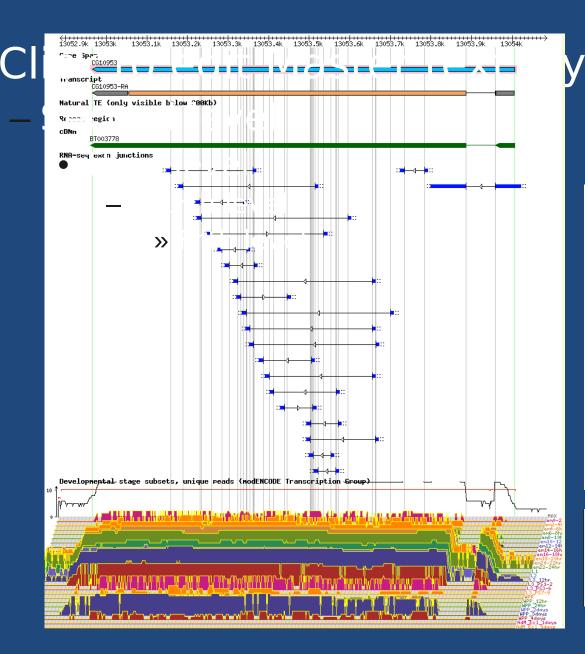
K. White group RNA-seq coverage by

subsets (devel.stages) [log2 converted]

gonads and male accessory glands (by strand) [log2 converted]

GBrowse: navigation options





yles Spurious junctions in region containing repeats

AGAAGCTACTCGATCGCTCTTGGAGTAACAGATTGCAGCTGTCATCATGA AATTCCTGGTTGTTGCCCTTTGCCGTTCTGGCCTGCGCCTATGGTGATGTC TCCCATTTGGGTTACGACTACTCTCCTCCGGCCCCAGCTCCGGTTTACCA GCCCGCCCCGGCTCCGGTCTACCAGCCCGCCCCCGCTCCGGTCTACCAGC CTGCTCCAGCACCAGTTGTGATCCCTGCTCCTGCTCCAGCTCCAGTTGTG ATCCCTGCTCCTGCTCCCGCTCCAGTTGTGATTCCGGCTCCAGCTCCCGT GAAGACGTACGTCCCACCCGCACCCATCAGCATCCCAGCCCCAGTTTACC CCTGCTCCCATCAGCATCCCA TGAGATCCCTGCCCC AGCTCCAGTGAACACCTACATTCCCCCCGCACCAGCACCAGCTCCGGTTT ACCAGCCAGCCCCTGCTCCCATTCCCGTGAGCATCCCGGCTCCAGCTCCC GTTTACCAGCCTGCTCCCGCTCCCGTTGTGATCCCAGCCCCTGCTCCAGC ACCAGTTGTGATCCCTGCTCCTGCTCCCGCTCCAGTTGTGATCCCAGCCC CAGCTCCCGTTAAGTCCTATGTGCCACCTGCACCAATCAGCATCCCAGCC CCGGCTCCCGTTTACCAGCCAGCTCCCATCAGCATCCCGGCCCCTGCCCC TTCTGGAGGAGATCGAGCCCGCTTCCAATGATGGATACCGCTACAAGACC GTGCGTCGTCGCGTCTACCGTCACCGTTTCTAAACTTACCAAAGCCCAAC GAAAAACTGTTTCCTAACGAATCCCTACGGGGACTCGCAATAAAAAATTT CCGCCTTCAATTTGTTAAAC

MKFLVVAFAVLACAYGDVSHLGYDYSPPAPAPVYQPAPAPVYQPAPAPVY QPAPAPVVIPAPAPAPVVIPAPAPAPVVIPAPAPVKTYVPPAPISIPAPV YQPAPAPIRIPAPVYQPAPAPISIPAPAPIEIPAPAPVNTYIPPAPAPAP VYQPAPAPIPVSIPAPAPVYQPAPAPVVIPAPAPAPVVIPAPAPAPVVIP APAPVKSYVPPAPISIPAPAPVYQPAPISIPAPAPVYQPAPAPVYQPTNT QVLEEIEPASNDGYRYKTVRRRVYRHRF

Repeat and low complexity regions create alignment problems

1937k	+ + + + + + + + + + + + + + + + + + + 	4939 :	4940k	4941k	4942k
	ວ໌g≋⊥ 			>	
-	Sgs1-, 'A 007-00-1-0-1-0-1-0-1-0-1			⊳	
: (only vi	sible below GOOKb)				
ion	•				
on juncti	ons				
:	→ } =::				
utal stage	subsets, unique reads	wodENCODE Transc	ription Group)		
101 30050					
	(V	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,			
					en2-41 en4-6h en5-8h
					eri8-10 em10-12 em12-14 em14-16h
					em16-18hi em18-20hr em20-22hr
		W. W	~~~~~~~~~~~~~~~~~~~~~~~~~~		em22-24hr
					L3 P51-2 L3 P53-6 L3 P57-9
					HPP HPP 12hr HPP 24hr
					HPP 12hr HPP 24hr HPP 2days HPP 3days HPP 3days HPP 3days HPT 4days RdH 2c1 1days RdH 2c1 1days
					Hill Held Ec1 30days
		11 To			AdF Ecl Idays AdF Ecl Idays AdF Ecl Idays AdF Ecl Idays

Sgs1 repeats foil RNA-Seq alignments

NKY ALIFE CYSTLLLED KNVK AMY DWD: MODG PSGEDD PGCG GD (TE (Y)70 PVOPCDTDSNP1 I I KPRUKTKRPKS1 KK1 I KR I KREKKKTTKWT I KRATK RTTKRTTRRPTTPKTPDTTDSPITTTGAECTCSDRTTASSTDSTTDRTT Sec WITH PLUTOTEPOTOSEESSTAIPSSPCIDTSTVIPTSPCTQETTT FIFTUS TUGTUT TPC1 CAUTT TTPRSTTTTSTSRPTTTTPRSTTTTTSR PTTTTPRSTTTTTRRPTTTTPRCTTTTSTCAPTTTTPRSTTTTTSRPT TITPROTT TISTOSPON TPRSTTTTSTSRPTTTPRCTTPSTSRPTTT TPRSTTKTSTCAPTTTPRPTTTPSTSRPTTTTPRSTTTTSTSRPTTTTP RS____TTTPRO1TTPRSTTTSSRPTTTPRSTTTTSRPTTTPRS TTITCTCSPTITTPRSTTTSTSRPTTTTPRSTTTTSTSGPTTTTPRSTT TTTTSGPTTTTPRSTTTCTCSPTTTTPRSTTTPSTSRPTTTTPRSTTTT CTCSPTTTPKSTTFISTSRPTTTPRSTTTTTSRPTTTPRSTTTTST SGPTTTTPRSTTTTSTSGPTTTTPRSTTTTSTSGPTTTTPRSTTTTSTSG PTTTTPRSTTTTSTSGPTTTTPRSTTTTSTSGPTTTTPRSTTTTSTSGPT TTTPRSTTTTSTSGPTTTTPRSTTTTSTSGPTTTTPRSTTTTSTSGPTTT TPRSTTTTSTSGPTTTTPRSTTTTSTSGPTTTTPRSTTTTSTSGPTTTTP RSTTTTSTSGPTTTTPRSTTTTSTSGPTTTTPRSTTTTSTSGPTTTTPRS TTTTSTSGPTTTTPRSTTTTSTSGPTTTTPRSTTTTSTSGPTTTTPRSTT TTSTSGPTTTTPRSTTTTSTSCPTTTTPRSTTTCTSGPTTTTPRSTTTT CTSCPTTTTPRSTTTTCTSCPTTTTPRSTTTTCTSGPTTTTPRSTTKTST CAPTTTTPRSTTTTSTSRPTTTTPRSTTTTTSRPTTTTPRSTTTPSTSR PTTTPRSTTTSTSRPTTTPRSTTKTSTCAPTTTPRSTTTTTSRPT TTTPRSTTTTTSRPTTTPRSTTTPCTSRPTTTTPRSTTTTTSRPTTT TPRSTTTPCPTTTPSASPTRTTTTRRPCPCHP0PPY0IPPWSWWYPNPTY PNPVWPWQPNPVVPQWPQMPGYPNQWPQLPGYPQQLPPPLPSPQWPWSWP KPPVV0PGDCENICENLLKGVKY0EDLIRRCLCTRT

ick

styles

ATCCACAACAACAACTTCCACAAGTGGCCCCAACGACAACTACTCCTAGAT CCACAACAACAACTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCC ACAACAACAACTTCCACAAGTGGCCCCAACGACAACTACTCCTAGATCCAC AACAACAACTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAA CAACAACTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACA ACAACTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAAC AACTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAA CTTCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAACT TCCACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAACTTC CACAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAACTTCCA CAAGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAACTTCCACA AGTGGCCCAACGACAACTACTCCTAGATCCACAACAACAACTTCCACAAG TGGCCCAACGACAACTACTCCTAGATCCACAACAACAACTTCCACAAGTG GCCCAACGACAACTACTCCTAGATCCACAACAACAACTTCCACAAGTGGC CCAACGACAACTACTCCTAGATCCACAACAACAACTTCCACAAGTTGCCC AACGACAACTACTCCTAGATCCACAACAACAACTTGCACAAGTGGCCCCAA CGACAACTACTCCTAGATCCACAACAACAACTTGCACAAGTTGCCCCAACG ACAACTACTCCTAGATCCACAACAACAACTTGCACAAGTTGCCCCAACGAC AACTACTCCTAGATCCACAACAACAACTTGCACAAGTGGCCCAACGACAA CTACTCCTAGATCCACAACAAAAACCTCCACTTGTGCCCCCAACGACAACT ACTCCTAGATCCACAACAACAACTTCCACAAGTCGCCCAACGACAACTAC TCCTAGATCCACAACAACAACTACCACAAGTCGCCCAACGACGACTACTC CTAGATCCACAACAACACCTTCCACAAGTCGCCCAACCACAACTACTCCT AGATCCACAACAACTACTTCCACAAGTCGCCCAACGACAACTACTCCTAG ATCCACAACAAAAACTTCCACTTGTGCCCCCAACGACAACTACTCCTAGAT

	NTAL TAY JOGGGG TATISTA GA						V I
Home Tools File	es Species Docume	ents Resour	ces News	Help Ar	chives	Jump to Gene	Go
Profile Managar					Help	Open All Close	All
General Information							
Symbol	Dmel:r5:2L:1653835:1690	117:+	Species	L	D. melanogaster		
Feature type	exon_function		FlyBase ID	F	Bsf0000101966	ò	
Collection	nuooShCOE E_mRN/eSeg BCM_1_RNAseq_junction	_U_junctions s	Associated gen	e(s)	chinmo		
Genomic Location	Courth I						
Chromosome (arm) — Map (GBrowse)	2L		Sequence locat	on 2	2L:1,653,8351,6	690,117 [+]	
	»	:1600 chinmo-RH chinmo-RE chinmo-RG chinmo-RD	chinmo-RF				
Recent Updates							
Sequence Data							
Associated Informat	lon						
 Experimental Data RNA-Seq Read Court 	nte						
Library: BCM_1_RN/							
Library: modENCOD	E_mRNA-Seq_U_junctions						
	Expression stage(s)				Read Cou	int	
	embryonic stage 1 – 4				2		
	embryonic stage 1 4 embryonic stage 4 9				2		
modENCODE_mRNA- Seq_U_junctions							-
	embryonic stage 4 – 9				0		
	embryonic stage 4 – 9 embryonic stage 9 – 11				0		
	embryonic stage 4 – 9 embryonic stage 9 – 11 embryonic stage 11 – 12				0 1 5		