Updates on Peak Calling

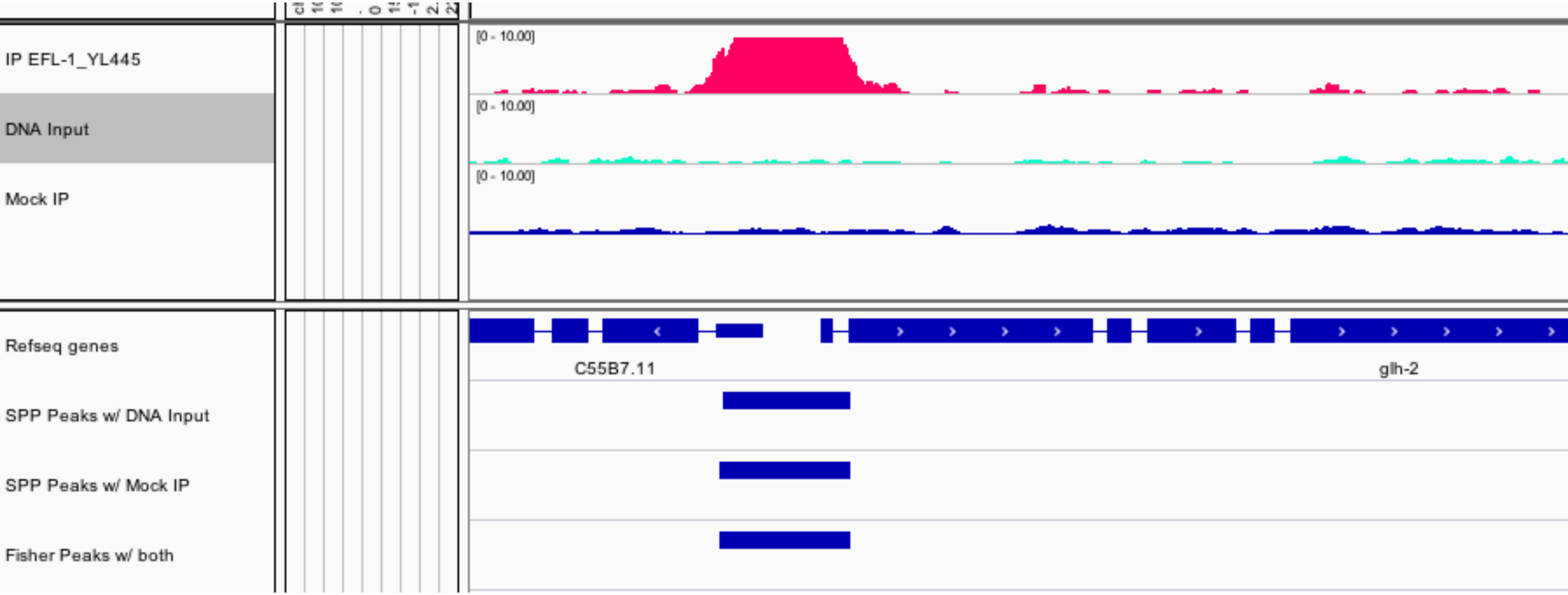
Jinrui Xu

02/23/2017

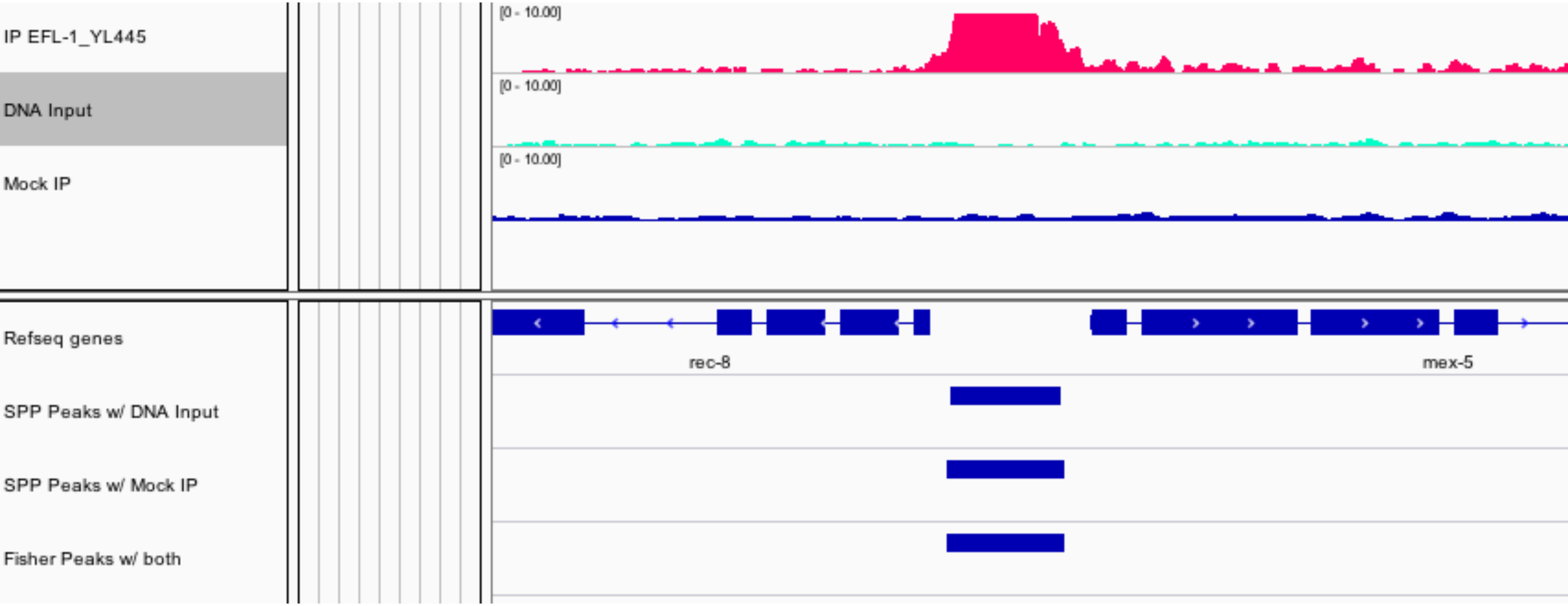
modERN Call

- Comparing peak calling results
  1 SPP using DNA input as control
  2 SPP using mock IP as control
  3 Fisher-exact test using both mock IP and DNA

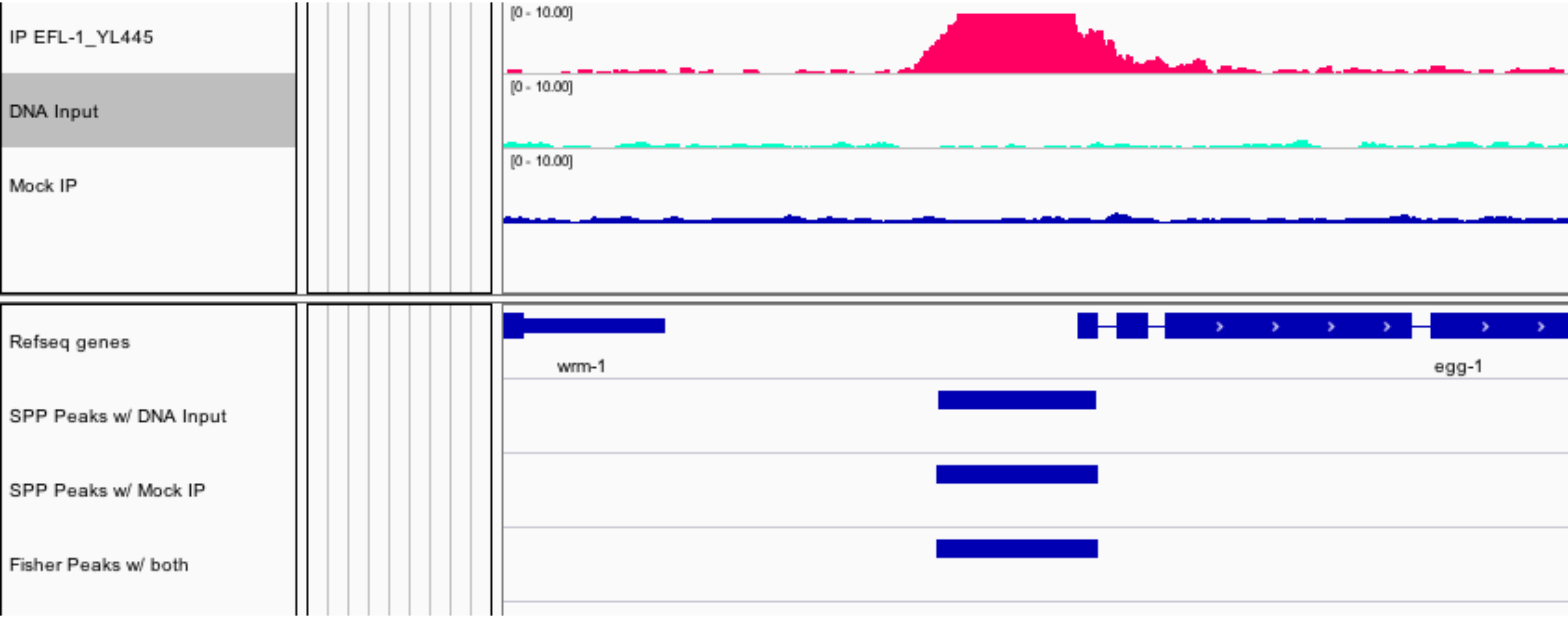
glh_2


IP EFL-1_YL445
[0 - 10.00]
DNA Input
[0 - 10.00]
Mock IP
[0 - 10.00]
Refseq genes
C55B7.11
glh-2
SPP Peaks w/ DNA Input
SPP Peaks w/ Mock IP
Fisher Peaks w/ both

Mex-5


IP EFL-1_YL445
DNA Input
Mock IP
[0 - 10.00]
[0 - 10.00]
[0 - 10.00]
Refseq genes
rec-8
mex-5
SPP Peaks w/ DNA Input
SPP Peaks w/ Mock IP
Fisher Peaks w/ both

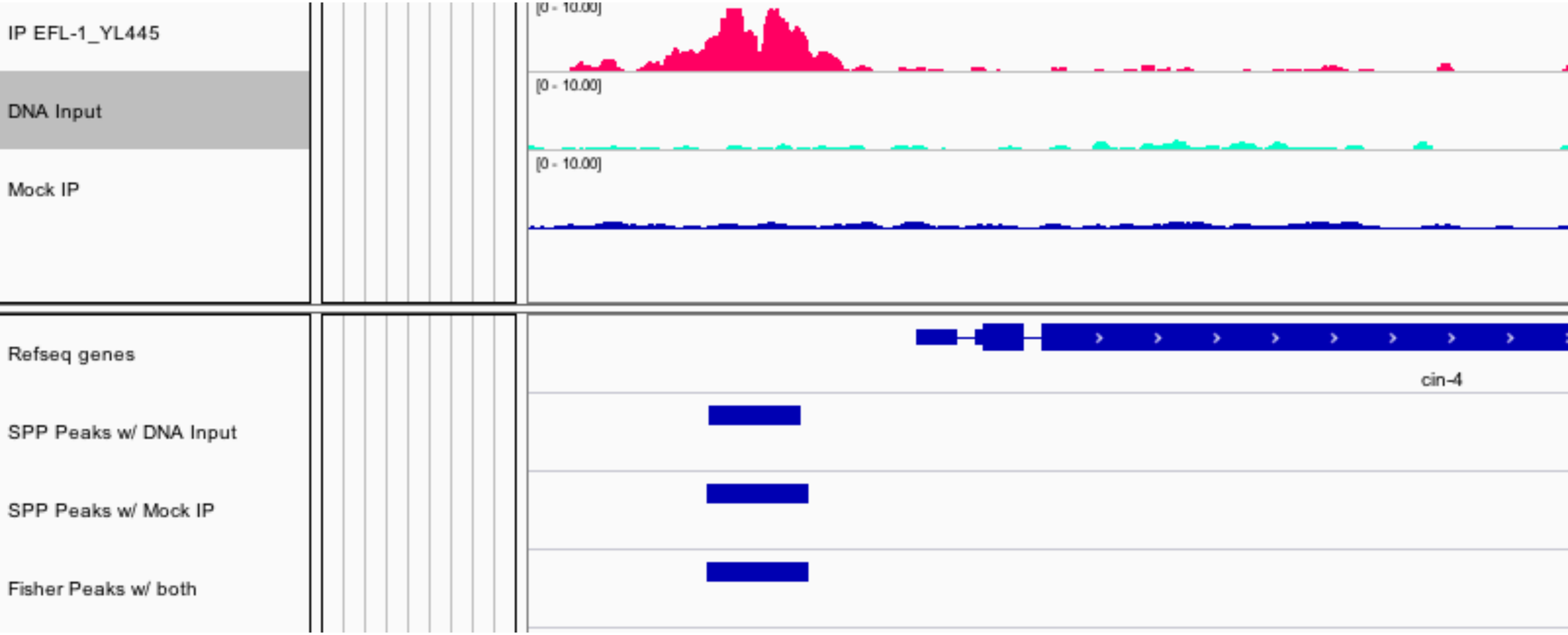
Egg-1


IP EFL-1_YL445
[0 - 10.00]
DNA Input
[0 - 10.00]
Mock IP
[0 - 10.00]
Refseq genes
wrm-1
egg-1
SPP Peaks w/ DNA Input
SPP Peaks w/ Mock IP
Fisher Peaks w/ both

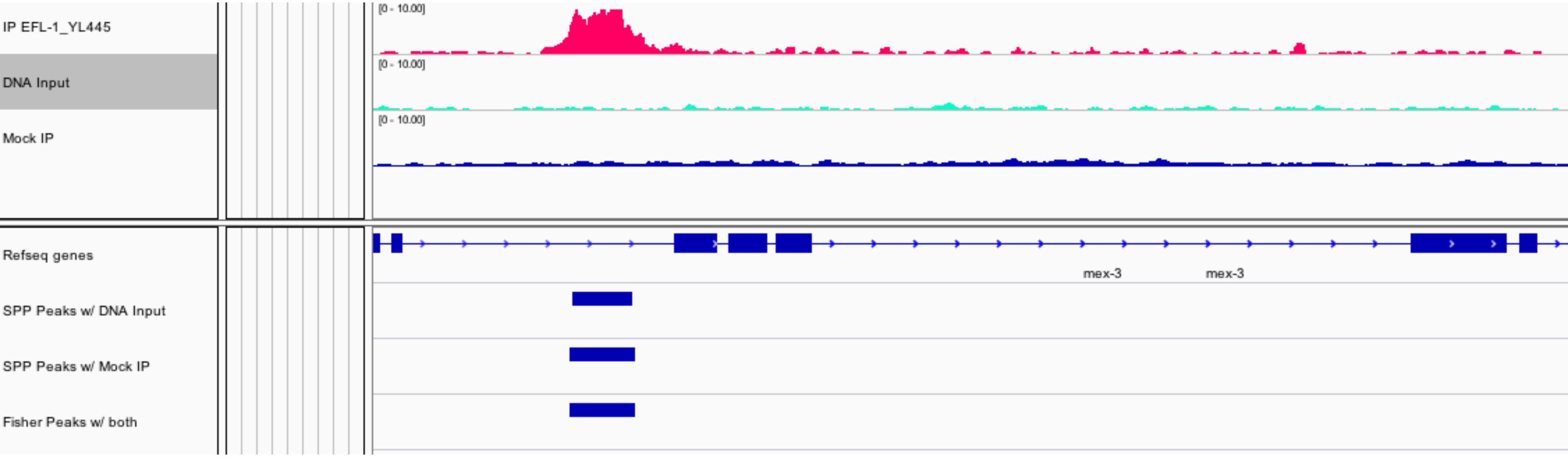
glp-1


IP EFL-1_YL445
[0 - 10.00]
DNA Input
[0 - 10.00]
Mock IP
[0 - 10.00]
Refseq genes
F02A9.4
glp-1
SPP Peaks w/ DNA Input
SPP Peaks w/ Mock IP
Fisher Peaks w/ both

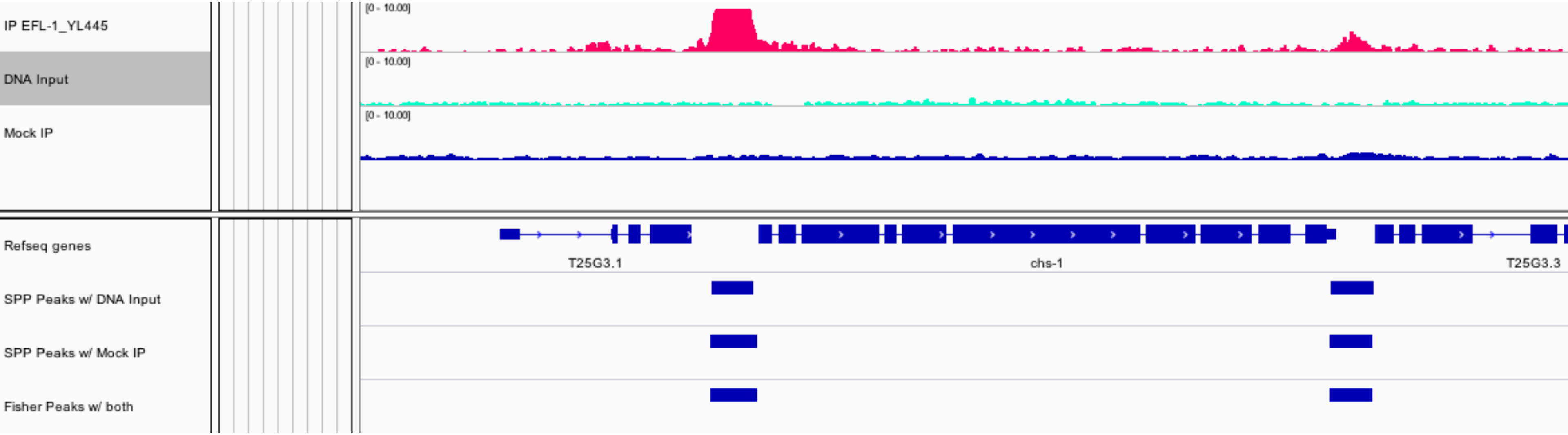
Cin-4


IP EFL-1_YL445
DNA Input
Mock IP
[0 - 10.00]
[0 - 10.00]
[0 - 10.00]
Refseq genes
cin-4
SPP Peaks w/ DNA Input
SPP Peaks w/ Mock IP
Fisher Peaks w/ both

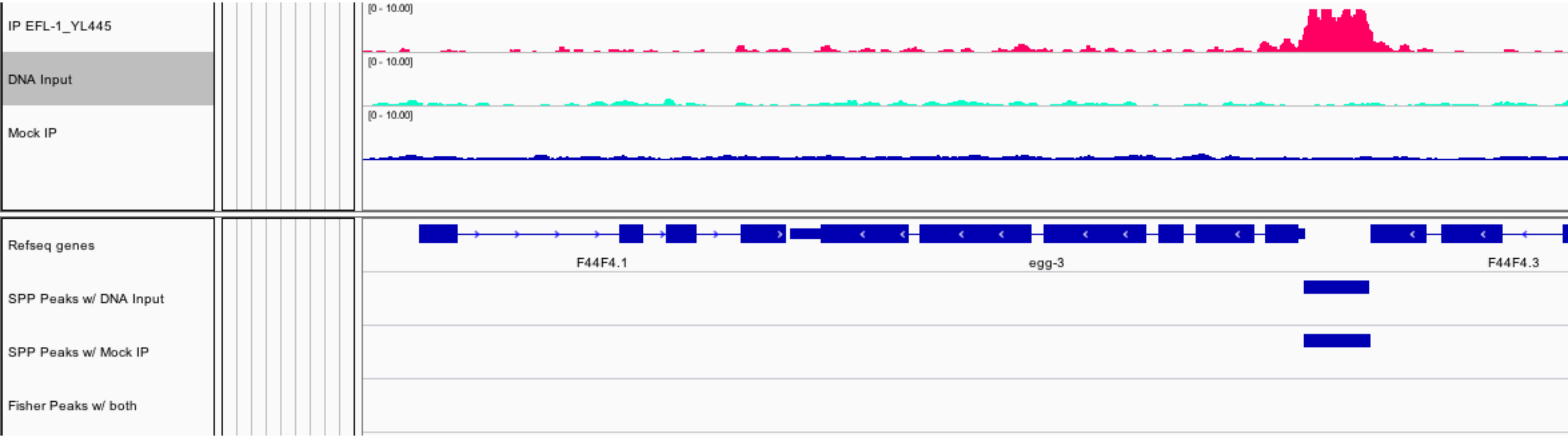
Mex-3


IP EFL-1_YL445
[0 - 10.00]
DNA Input
[0 - 10.00]
Mock IP
[0 - 10.00]
Refseq genes
mex-3
mex-3
SPP Peaks w/ DNA Input
SPP Peaks w/ Mock IP
Fisher Peaks w/ both

Chs-1


IP EFL-1_YL445
DNA Input
Mock IP
[0 - 10.00]
[0 - 10.00]
[0 - 10.00]
Refseq genes
SPP Peaks w/ DNA Input
SPP Peaks w/ Mock IP
Fisher Peaks w/ both
T25G3.1
chs-1
T25G3.3

Egg-3


IP EFL-1_YL445
DNA Input
Mock IP
[0 - 10.00]
[0 - 10.00]
[0 - 10.00]
Refseq genes
SPP Peaks w/ DNA Input
SPP Peaks w/ Mock IP
Fisher Peaks w/ both
F44F4.1
egg-3
F44F4.3

- EFL-1_YL455
  - DNA input peaks: 2192 vs Mock IP peaks: 1978
- SNPC-4_OP179: 1430 vs 1328
- SNPC_4_YL457: 1110 vs 908
- SNPC-4_YL524: 2327 vs 2129
- DPL-1_YL425: 3901 vs 3887

- Mock IP controls
  - N2_WA_yAd_IP_Rep0.tagAlign
    - N2_WA_yAd_IP_GoatV-aGFP_Rep1.tagAlign
    - N2_WA_yAd_IP_GoatV-aGFP_Rep2.tagAlign

Cele_controls

GFP	N2	yAd	IP	GoatV-aGFP	1	2016-1967	/raid/modencode/ce/processing/20161005
GFP	N2	yAd	IP	GoatV-aGFP	2	2016-1968	/raid/modencode/ce/processing/20161005

- Mock IP w/ GFP-Ab generates small number of significant peaks, in fly
  - Median: 184 (mock IP) vs 3550 (DNA input)
  - 25% of TFs have sig. peaks < 40

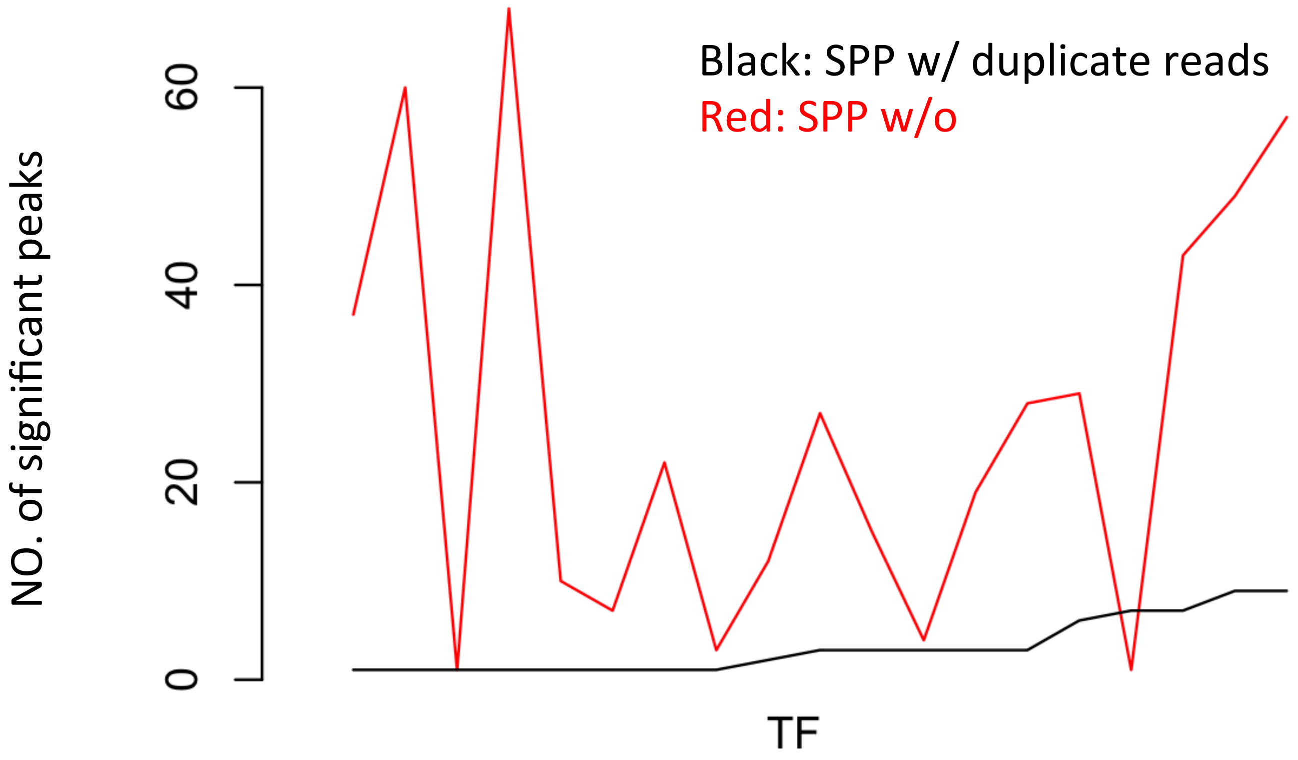
- The 42 TFs with lowest NO. of peaks in the 168 TFs:
  - Slowly expressed (not significant)
  - Less number of replicate experiments ($p < 0.03$)
    - In the 42 TFs, 14 have 3 replicates, 28 TFs have 2 replicates
    - In the rest TFs, the ratio is 65: 61

- IP has ave. 12% duplicate reads

- Mock IP has ave. 16% duplicate reads

- DNA input has ave. ~8% duplicate reads

- Removing duplicate reads improves discovery



Black: SPP w/ duplicate reads

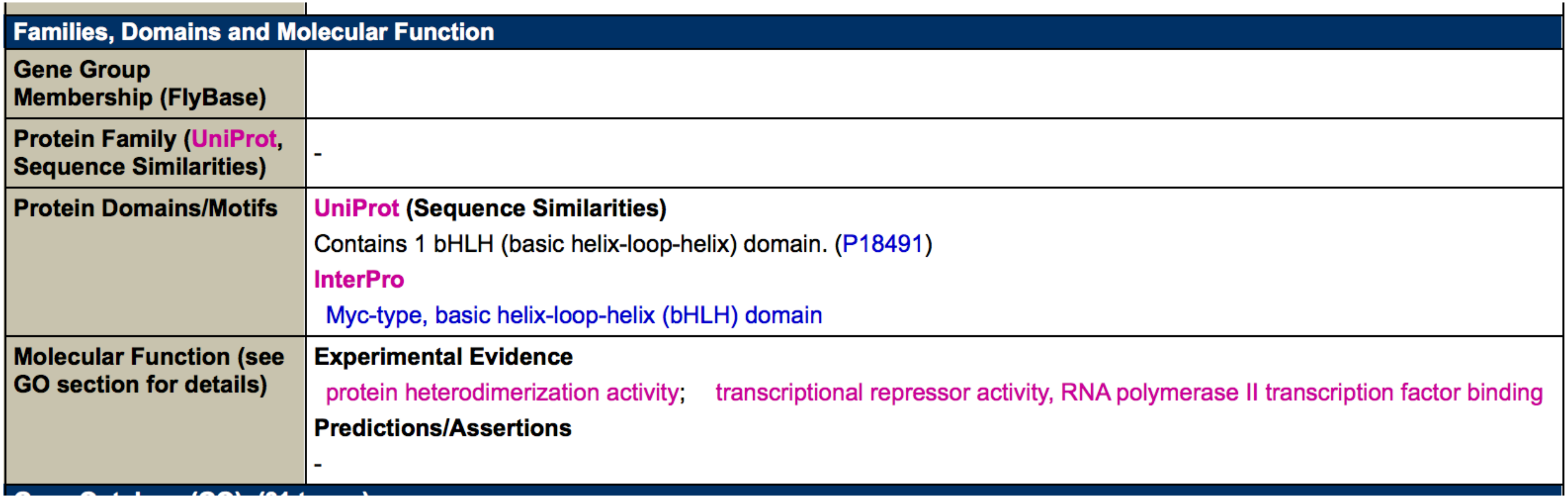
Red: SPP w/o

The 19 TFs with lowest NO. of sig. peaks

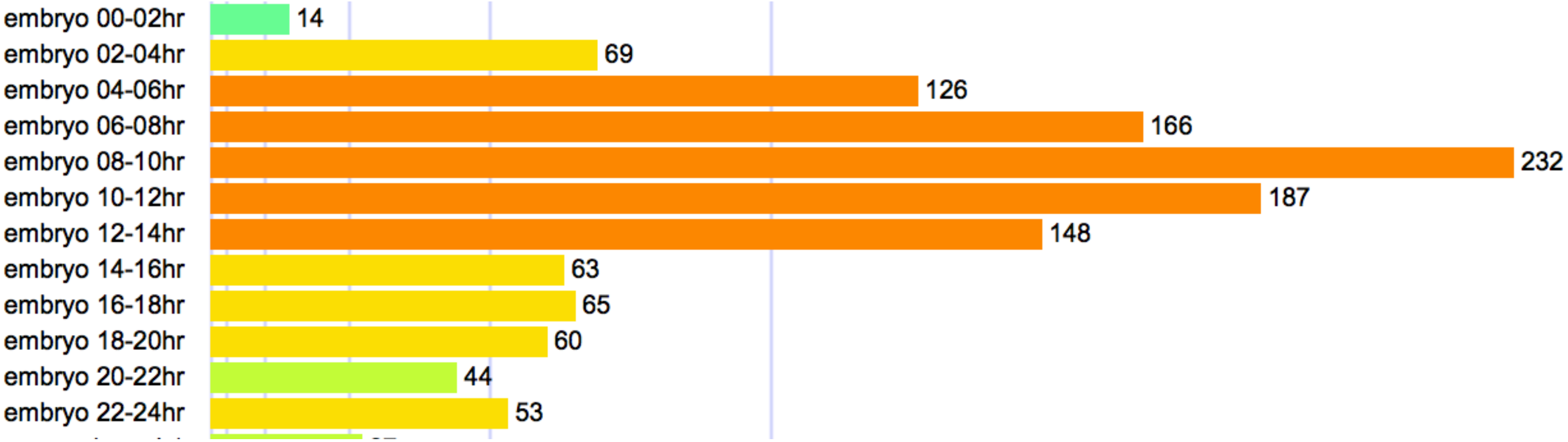
Screening the 19 TFs

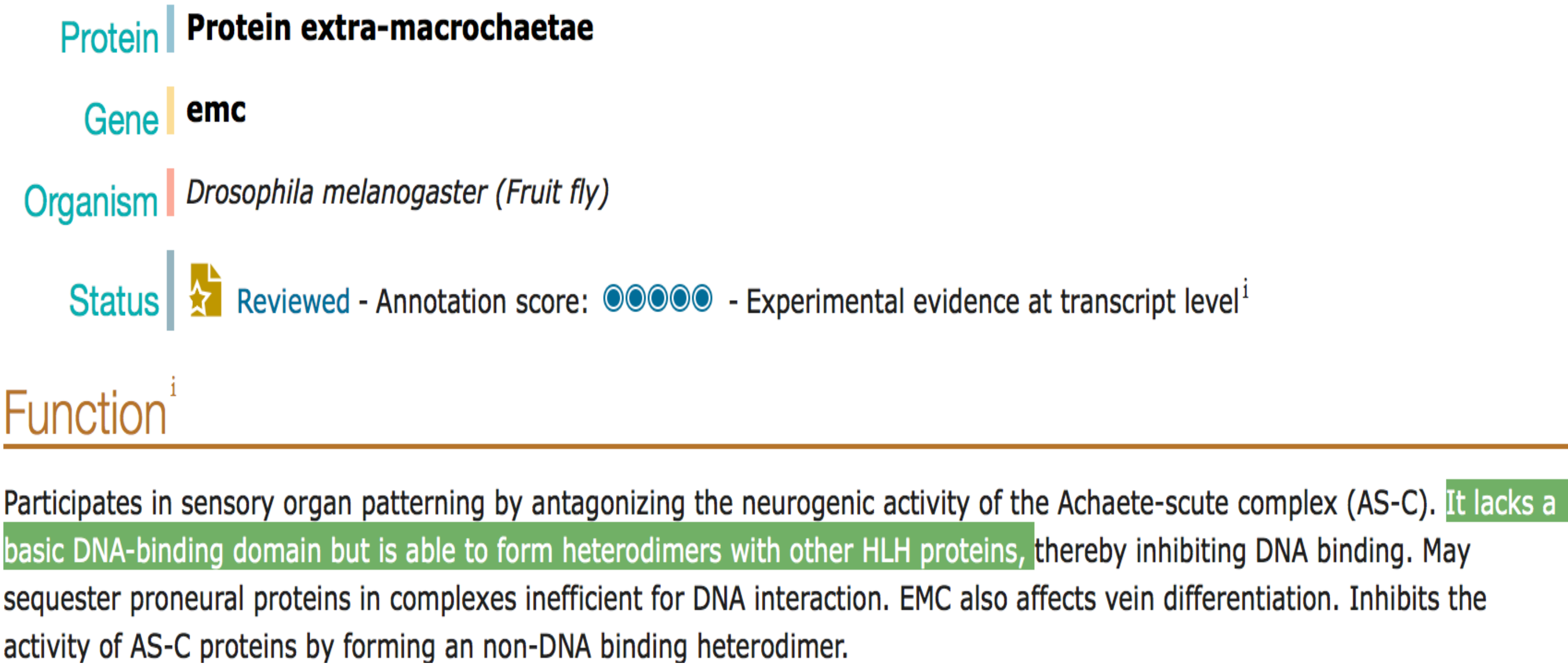
- TFs with annotations suggesting no bindings peaks

- emc: co-activator, 10 sig. peaks



Families, Domains and Molecular Function	
Gene Group Membership (FlyBase)	
Protein Family (UniProt, Sequence Similarities)	-
Protein Domains/Motifs	**UniProt (Sequence Similarities)** Contains 1 bHLH (basic helix-loop-helix) domain. (P18491) **InterPro** Myc-type, basic helix-loop-helix (bHLH) domain
Molecular Function (see GO section for details)	**Experimental Evidence** protein heterodimerization activity; transcriptional repressor activity, RNA polymerase II transcription factor binding **Predictions/Assertions** -





Protein | **Protein extra-macrochaetae**

Gene | **emc**

Organism | *Drosophila melanogaster (Fruit fly)*

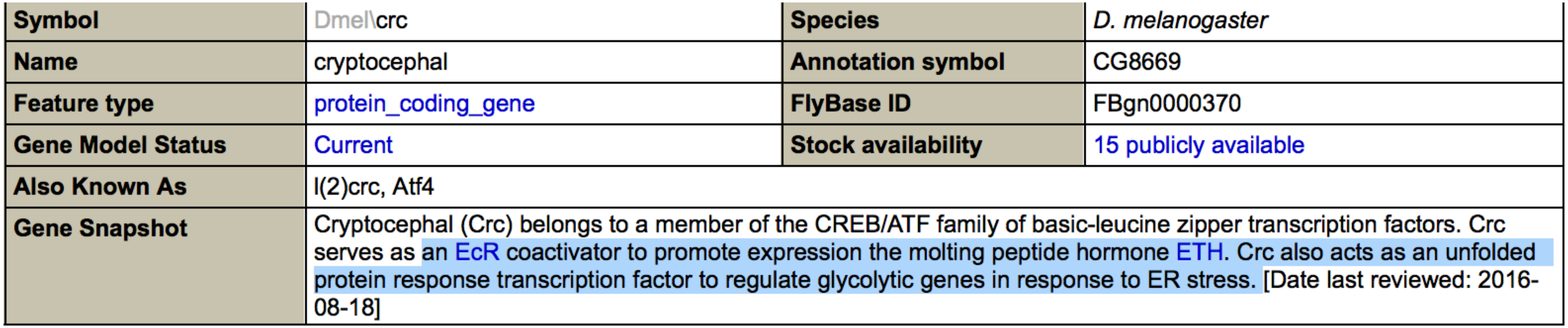
Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at transcript level

Function

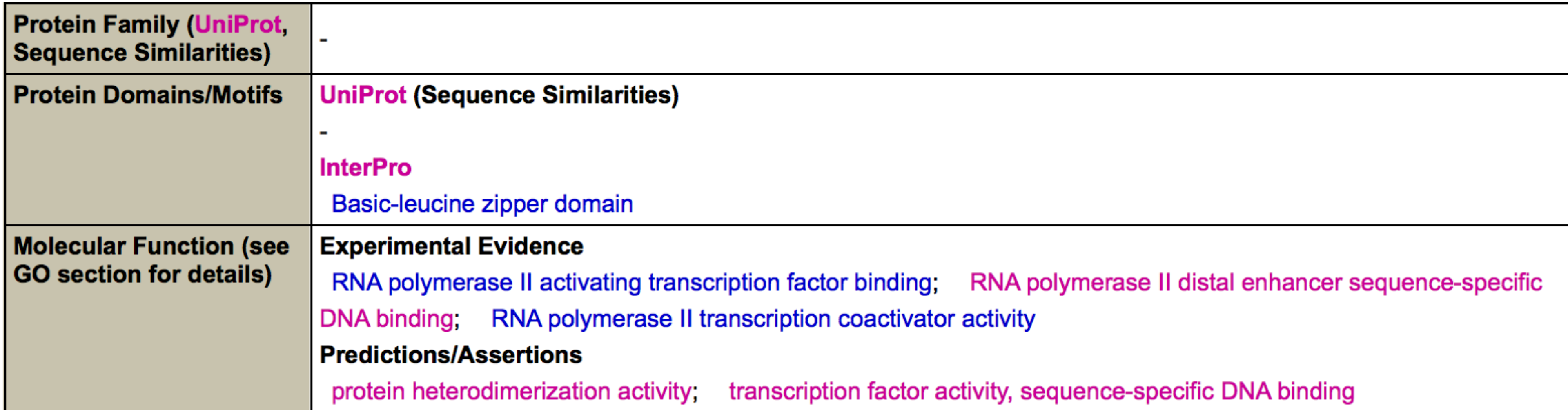
Participates in sensory organ patterning by antagonizing the neurogenic activity of the Achaete-scute complex (AS-C). It lacks a basic DNA-binding domain but is able to form heterodimers with other HLH proteins, thereby inhibiting DNA binding. May sequester proneural proteins in complexes inefficient for DNA interaction. EMC also affects vein differentiation. Inhibits the activity of AS-C proteins by forming an non-DNA binding heterodimer.

http://www.uniprot.org/uniprot/P18491

- crc: annotated as a co-activator, 15 sig. peaks

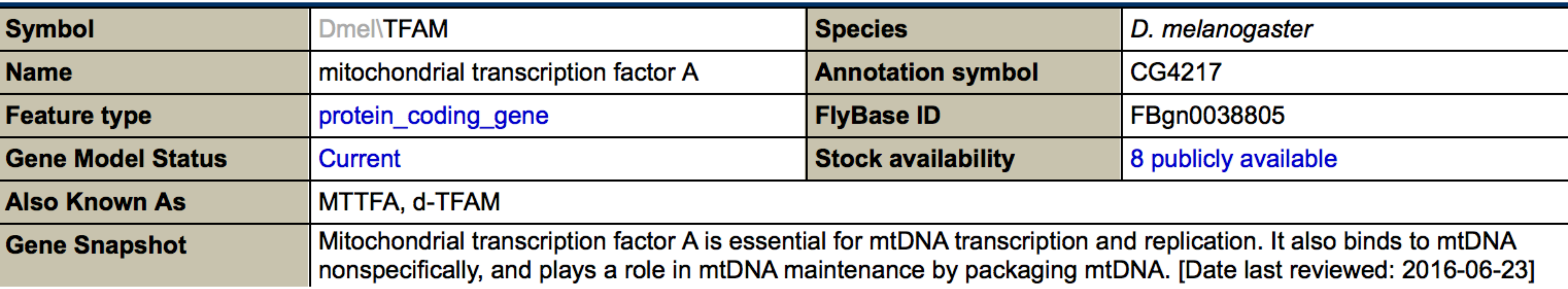


Symbol	Dmel\crc	**Species**	*D. melanogaster*
Name	cryptocephal	**Annotation symbol**	CG8669
Feature type	protein_coding_gene	**FlyBase ID**	FBgn0000370
Gene Model Status	Current	**Stock availability**	15 publicly available
Also Known As	l(2)crc, Atf4		
Gene Snapshot	Cryptocephal (Crc) belongs to a member of the CREB/ATF family of basic-leucine zipper transcription factors. Crc serves as an EcR coactivator to promote expression the molting peptide hormone ETH. Crc also acts as an unfolded protein response transcription factor to regulate glycolytic genes in response to ER stress. [Date last reviewed: 2016-08-18]		

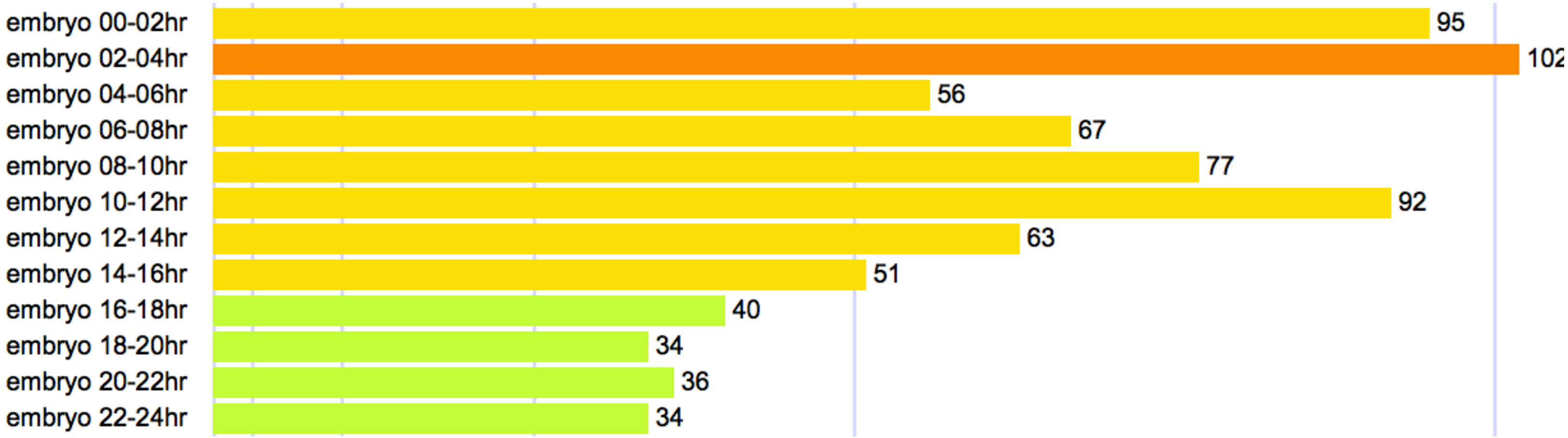


Protein Family (UniProt, Sequence Similarities)	-
Protein Domains/Motifs	**UniProt (Sequence Similarities)** - **InterPro** Basic-leucine zipper domain
Molecular Function (see GO section for details)	**Experimental Evidence** RNA polymerase II activating transcription factor binding; RNA polymerase II distal enhancer sequence-specific DNA binding; RNA polymerase II transcription coactivator activity **Predictions/Assertions** protein heterodimerization activity; transcription factor activity, sequence-specific DNA binding

- TFAM: mitochondrial TF, 28 sig. paeks



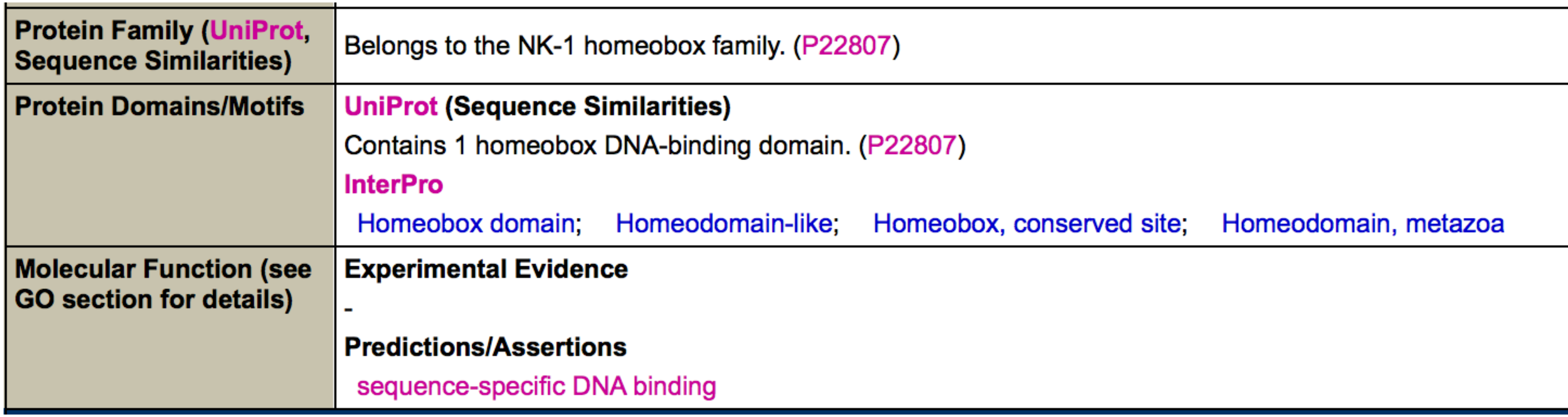
Symbol	Dmel\TFAM	**Species**	*D. melanogaster*
Name	mitochondrial transcription factor A	**Annotation symbol**	CG4217
Feature type	protein_coding_gene	**FlyBase ID**	FBgn0038805
Gene Model Status	Current	**Stock availability**	8 publicly available
Also Known As	MTTFA, d-TFAM		
Gene Snapshot	Mitochondrial transcription factor A is essential for mtDNA transcription and replication. It also binds to mtDNA nonspecifically, and plays a role in mtDNA maintenance by packaging mtDNA. [Date last reviewed: 2016-06-23]		



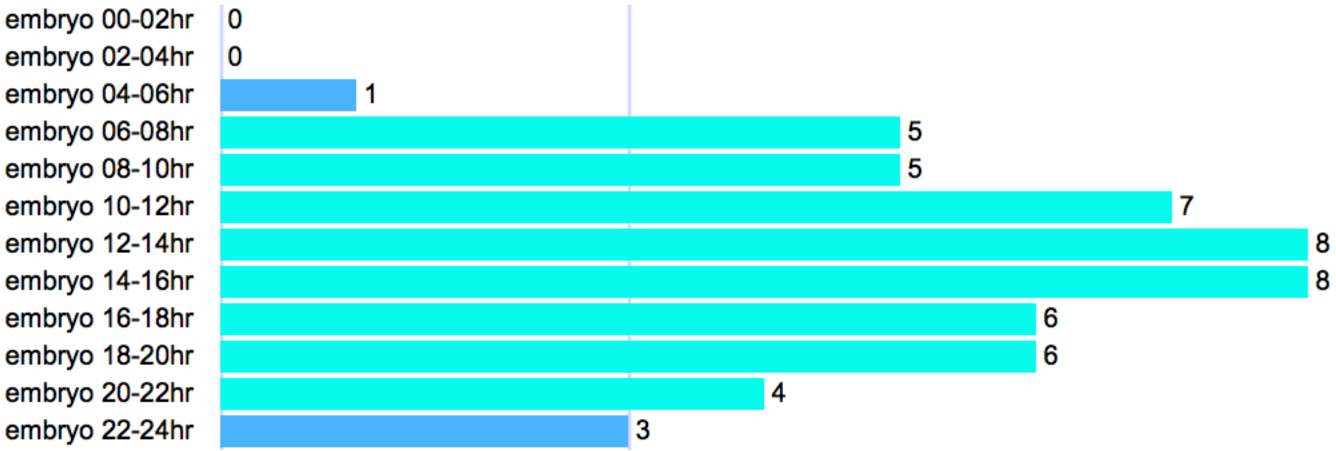
- Lowly expressed (, maybe highly expressed in a small number of cells)

- Temporal expressed TFs

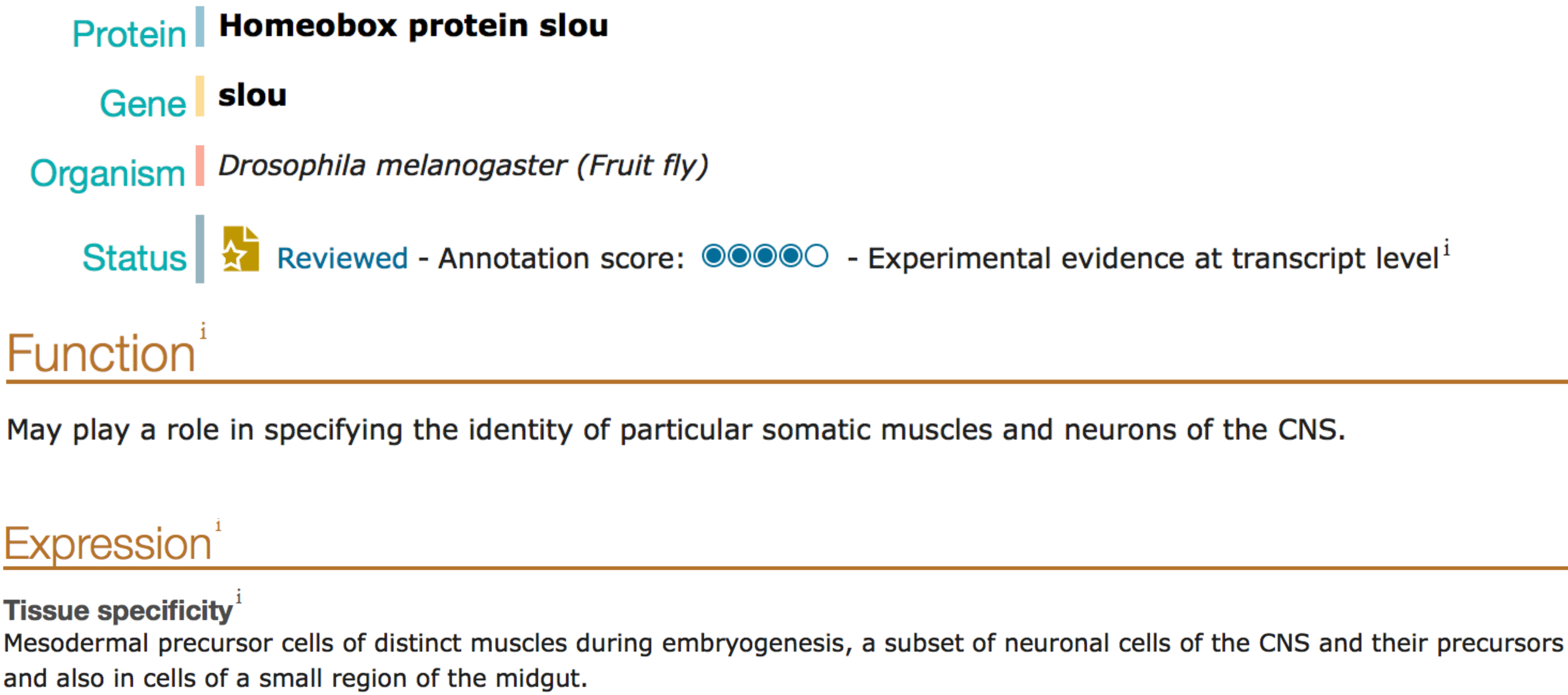
- Slou: 37 sig. peaks



Protein Family (UniProt, Sequence Similarities)	Belongs to the NK-1 homeobox family. (P22807)
Protein Domains/Motifs	**UniProt (Sequence Similarities)** Contains 1 homeobox DNA-binding domain. (P22807) **InterPro** Homeobox domain; Homeodomain-like; Homeobox, conserved site; Homeodomain, metazoa
Molecular Function (see GO section for details)	**Experimental Evidence** - **Predictions/Assertions** sequence-specific DNA binding



Very low expression: 0-3
Low: 4-10
Moderate: 11-25
Moderately high: 25-50



Protein **Homeobox protein slou**

Gene **slou**

Organism *Drosophila melanogaster (Fruit fly)*

Status Reviewed - Annotation score: ●●●●○ - Experimental evidence at transcript level

Function

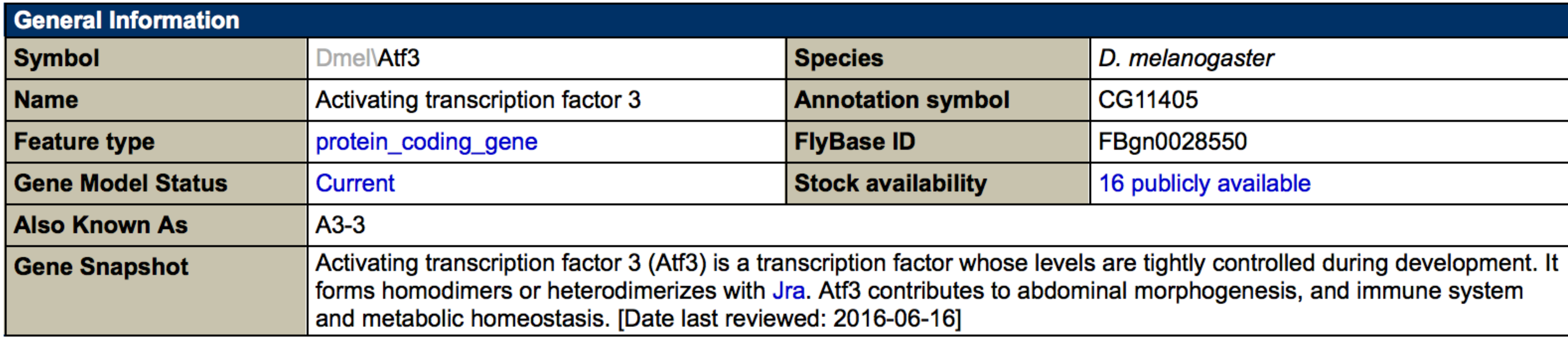
May play a role in specifying the identity of particular somatic muscles and neurons of the CNS.

Expression

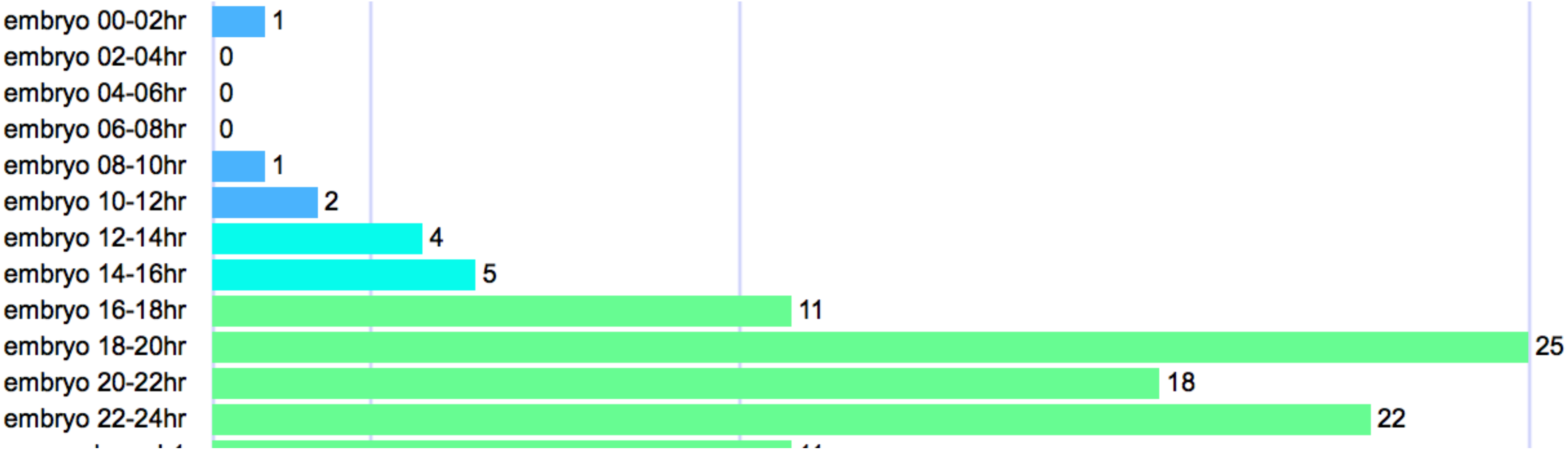
Tissue specificity

Mesodermal precursor cells of distinct muscles during embryogenesis, a subset of neuronal cells of the CNS and their precursors and also in cells of a small region of the midgut.

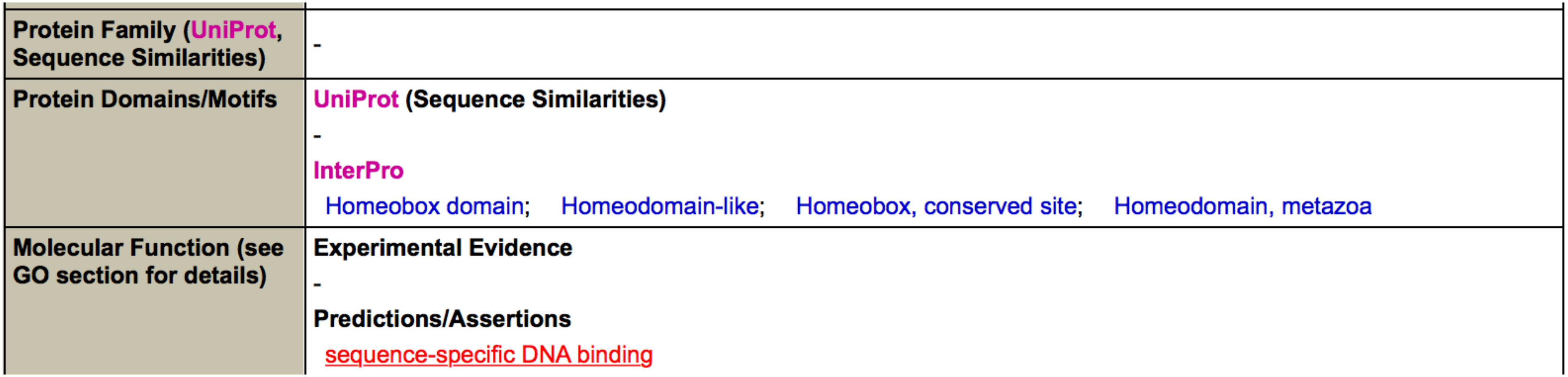
- Atf3, 49 sig. peaks, ChIP-Seq labeled as E0-24



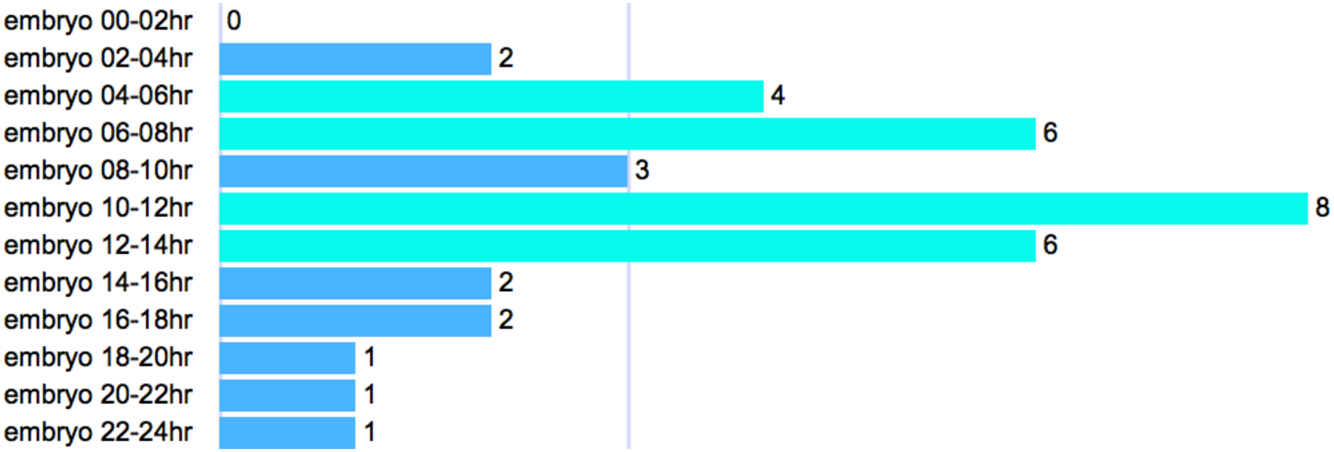
General Information			
Symbol	Dmel\Atf3	**Species**	*D. melanogaster*
Name	Activating transcription factor 3	**Annotation symbol**	CG11405
Feature type	protein_coding_gene	**FlyBase ID**	FBgn0028550
Gene Model Status	Current	**Stock availability**	16 publicly available
Also Known As	A3-3		
Gene Snapshot	Activating transcription factor 3 (Atf3) is a transcription factor whose levels are tightly controlled during development. It forms homodimers or heterodimerizes with Jra. Atf3 contributes to abdominal morphogenesis, and immune system and metabolic homeostasis. [Date last reviewed: 2016-06-16]		



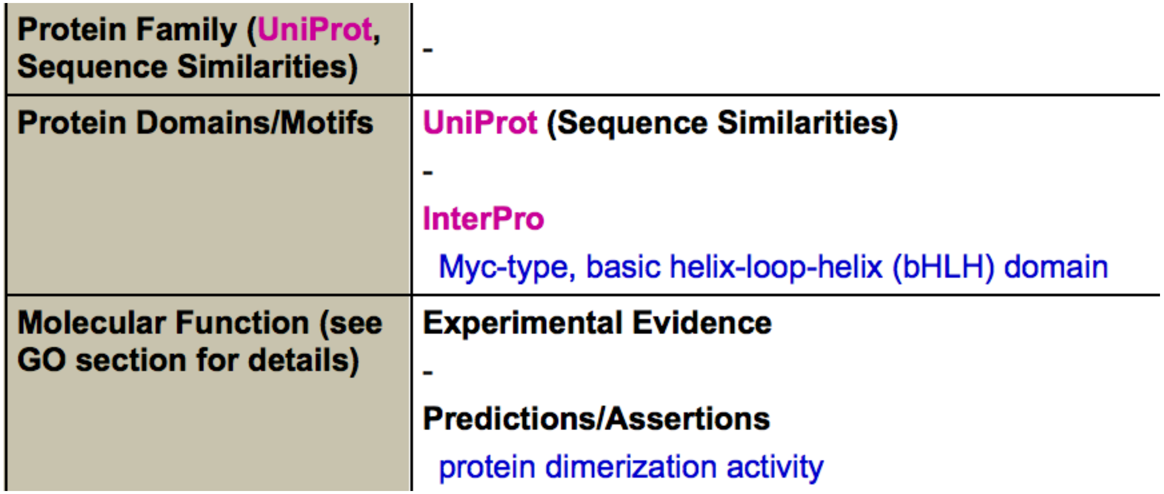
- btn, 1 sig. peaks, , Chip-Seq during E0-14



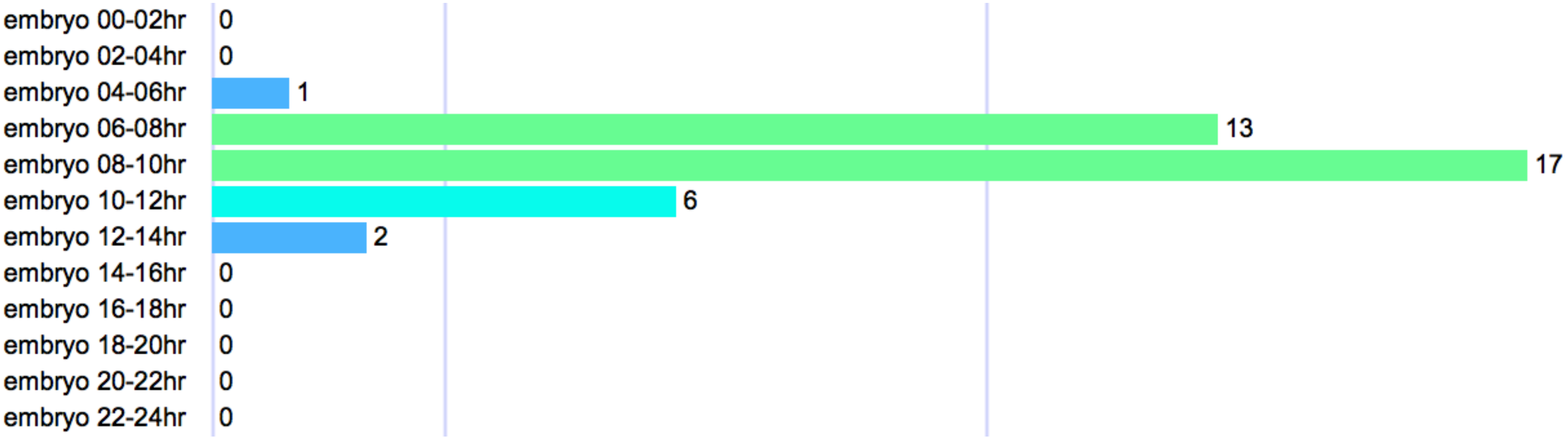
Protein Family (UniProt, Sequence Similarities)	-
Protein Domains/Motifs	**UniProt (Sequence Similarities)** - **InterPro** Homeobox domain; Homeodomain-like; Homeobox, conserved site; Homeodomain, metazoa
Molecular Function (see GO section for details)	**Experimental Evidence** - **Predictions/Assertions** sequence-specific DNA binding



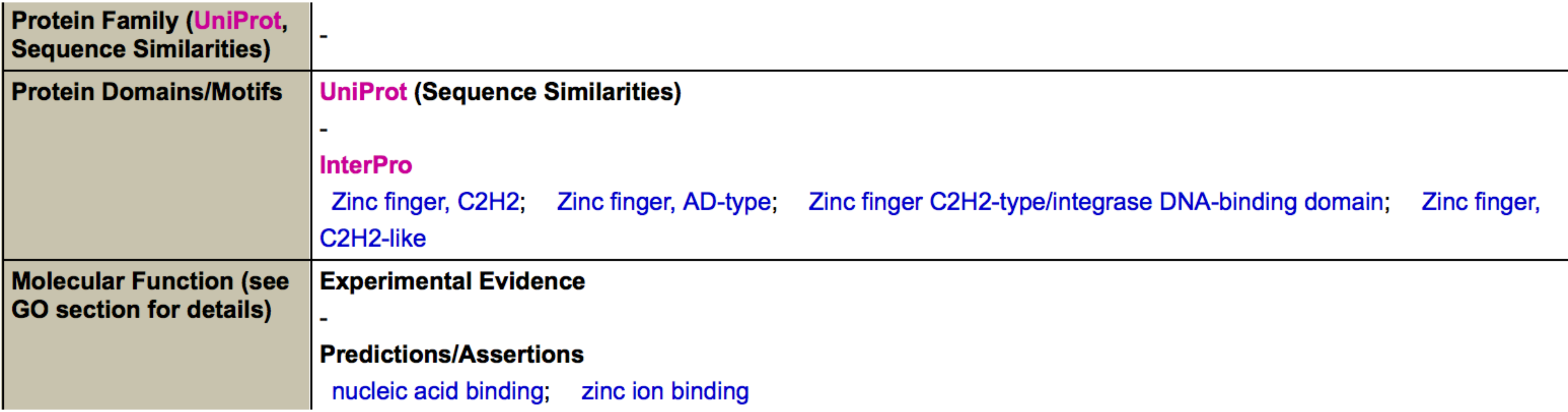
- cato, 60 sig. peaks, , Chip-Seq during E0-24



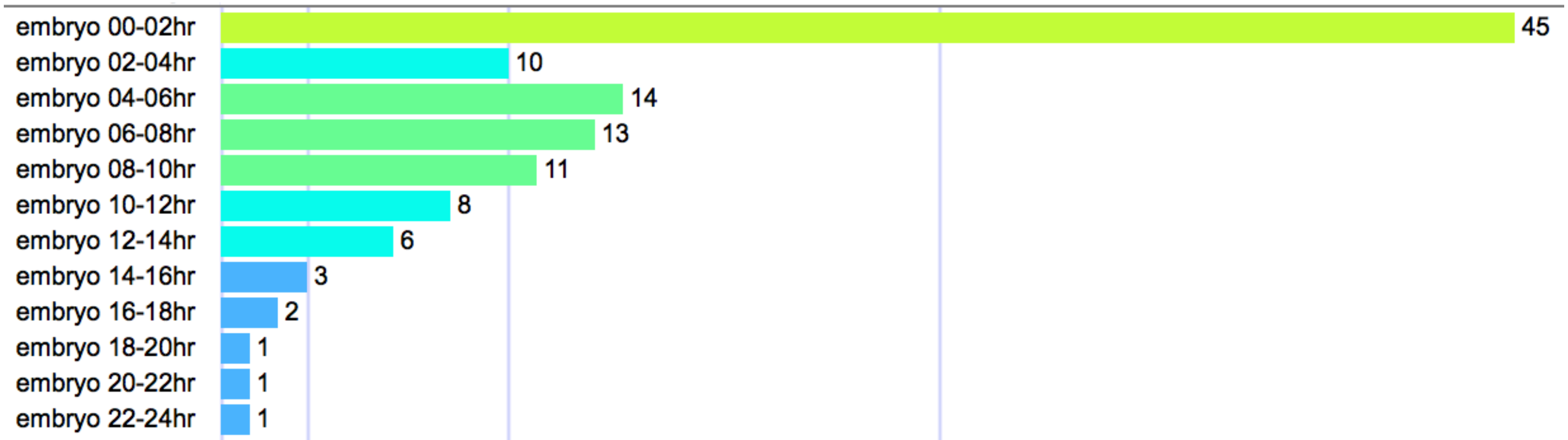
Protein Family (UniProt, Sequence Similarities)	-
Protein Domains/Motifs	**UniProt (Sequence Similarities)** - **InterPro** Myc-type, basic helix-loop-helix (bHLH) domain
Molecular Function (see GO section for details)	**Experimental Evidence** - **Predictions/Assertions** protein dimerization activity



- CG12942, 43 sig. peaks, Chip-Seq during E0-24



Protein Family (UniProt, Sequence Similarities)	-
Protein Domains/Motifs	**UniProt (Sequence Similarities)** - **InterPro** Zinc finger, C2H2; Zinc finger, AD-type; Zinc finger C2H2-type/integrase DNA-binding domain; Zinc finger, C2H2-like
Molecular Function (see GO section for details)	**Experimental Evidence** - **Predictions/Assertions** nucleic acid binding; zinc ion binding



- Several TFs are lack of experimental evidence suggesting binding peaks