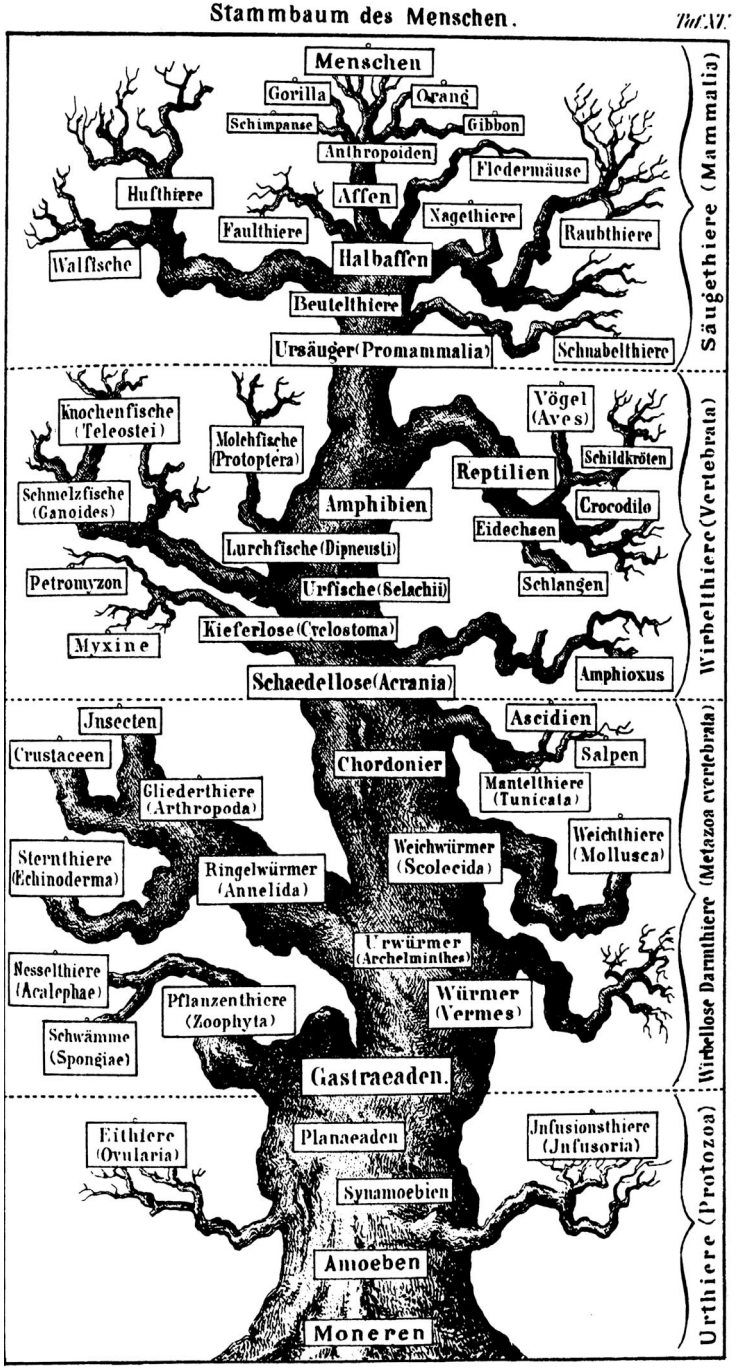
Evolution and Development

0. Evolution by changes in gene regulation (rather than creation of new genes)

1. (micro) evolution of early patterning in the fly embryo
   a. Locate regulatory sequence computationally.
   b. Molecular evolution (how seq. changes)
   c. How expression pattern of regulatory seq. changes.



Picturing Evolution (Haeckel 1874)

(development (egg->adult) of 'higher' organisms retraces adult forms of 'simpler' organisms.

Fig. 20. Haeckel's evolutionary tree. The first five stages (monera, amoeba, synamoeba, planaea, and gastrea) are reconstructed almost entirely from ontogeny of higher forms. (From Haeckel, 1874.)

Picturing Evolution (textbook ~1991)

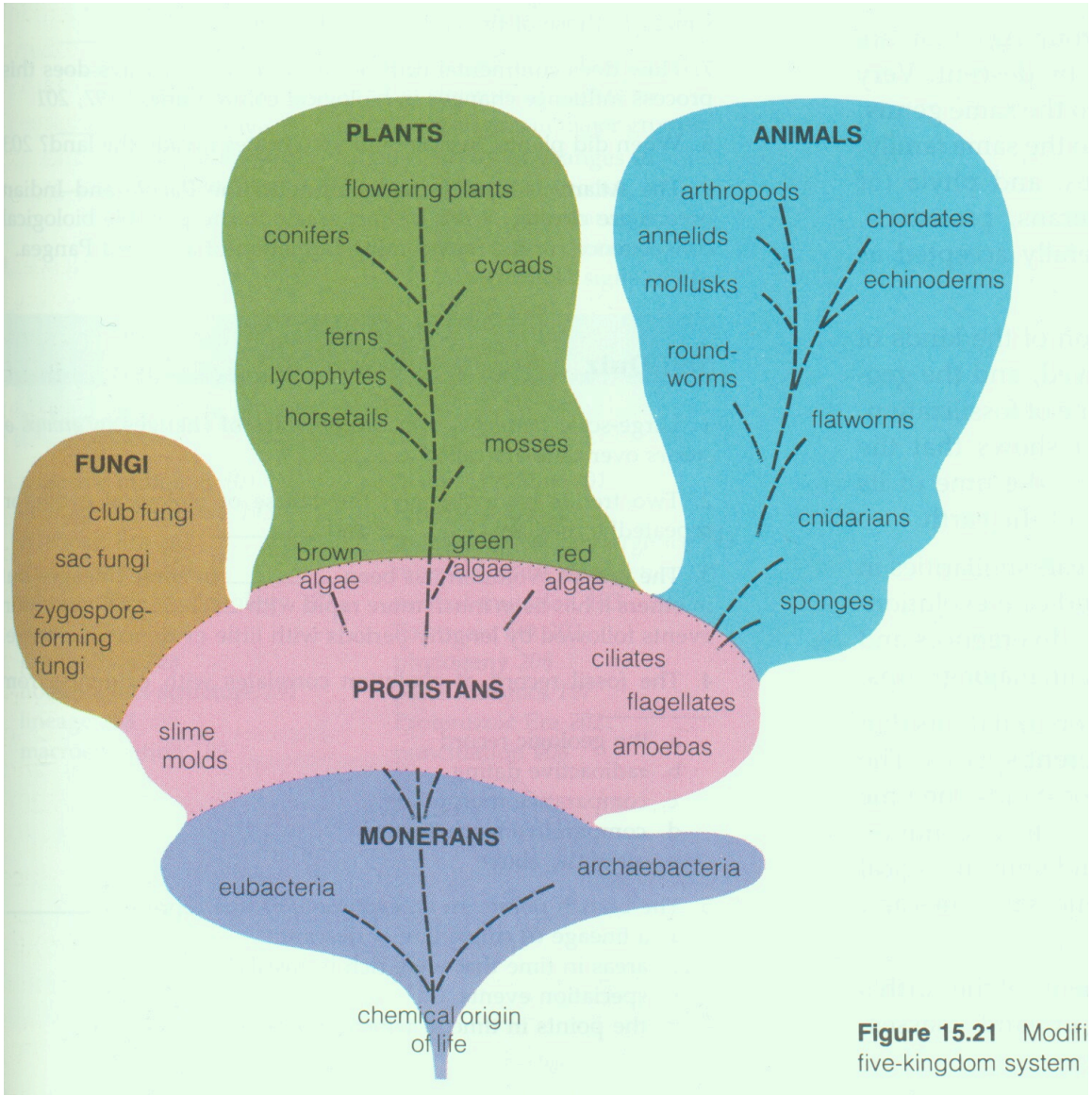
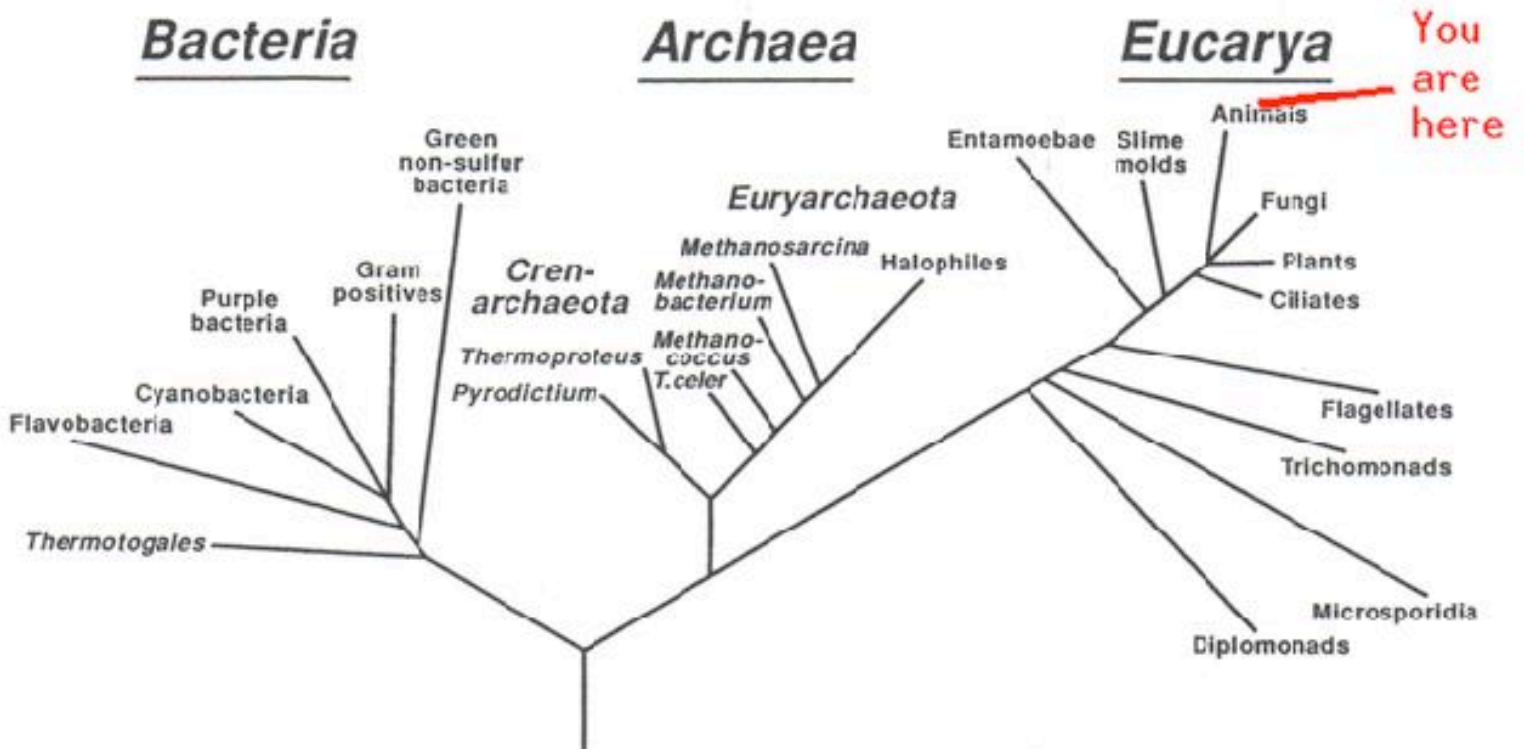


Figure 15.21 Modifi
five-kingdom system

The ‘5 kingdoms’

Picturing Evolution (Woese 1994)



From C. Woese, Microbiology Reviews 1994

Most multicellular life is restricted to the plant and animal branches in Eucarya

Genomic sequence supplies a metric with which to compare all life forms

Evolution via changes in gene regulation

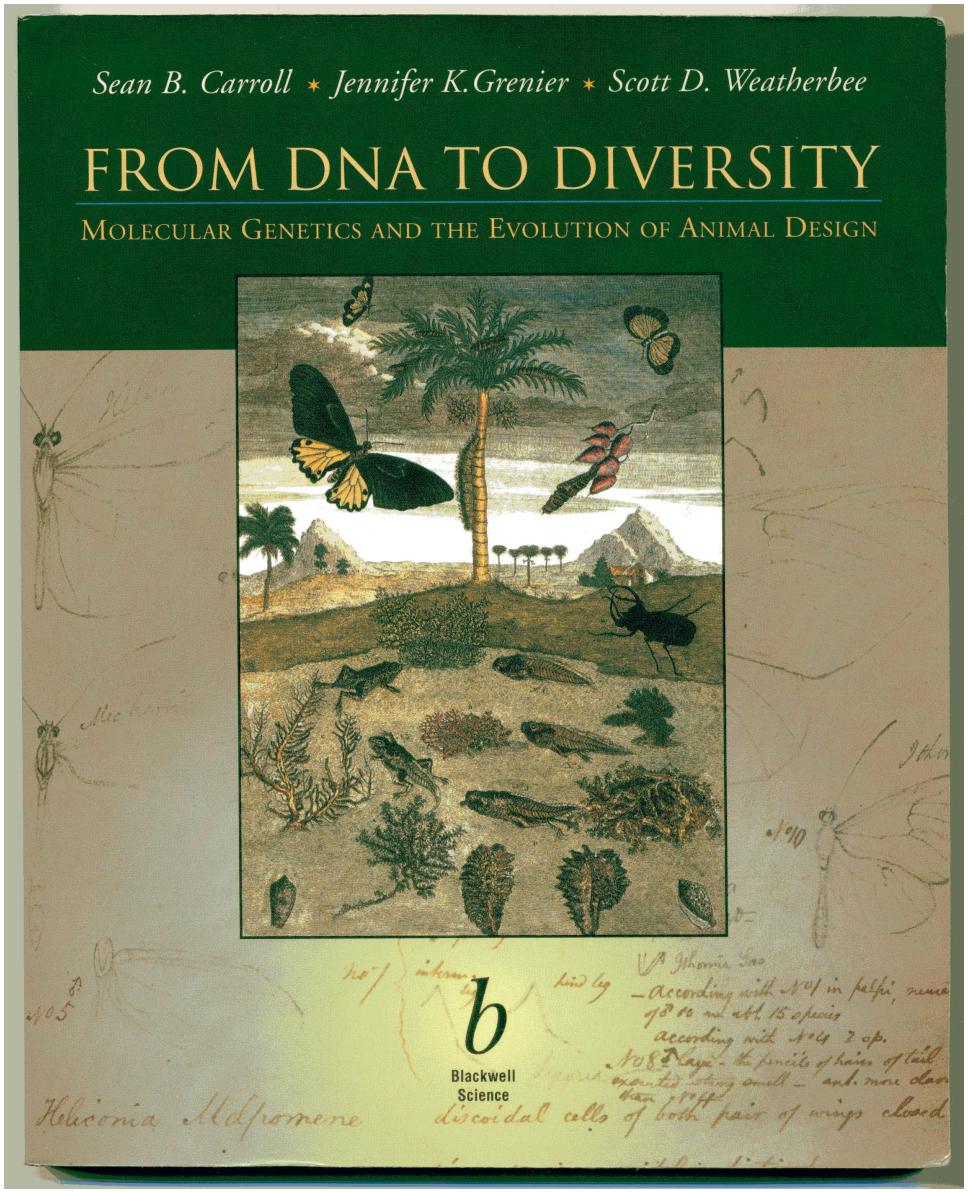
Homeotic mutations change body plan heterochrony change gene timing (early 20th c)

Gene numbers similar fly-human, gene types very similar, eg signaling pathways

All cells, same genome, development via regulatory cascade.

Organization of regulatory DNA modular: ~500bp pieces recapitulate a portion of native pattern in other genomic contexts. 'Quanta' of regulatory information

Genome as assembly manual rather than 'parts list'



Homeotic mutations

MATERIALS

FOR THE

STUDY OF VARIATION

TREATED WITH ESPECIAL REGARD TO

DISCONTINUITY

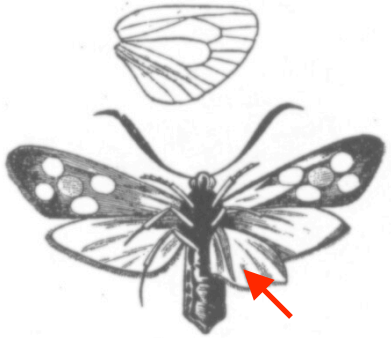
IN THE

ORIGIN OF SPECIES.

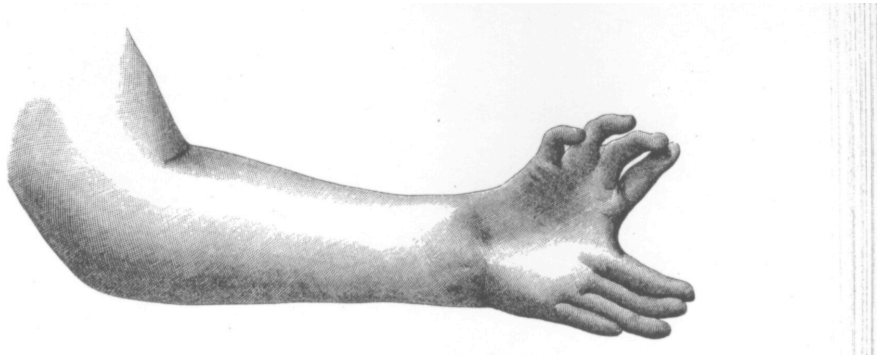
BY

WILLIAM BATESON, M.A.
FELLOW OF ST JOHN'S COLLEGE, CAMBRIDGE

London:
MACMILLAN AND CO.
AND NEW YORK.
1894



Ex. 78, leg->wing



Ex. 495 left hand, woman otherwise normal

Homeotic Mutations II

Chola dynasty 10-11th century, (vishnu, shiva)

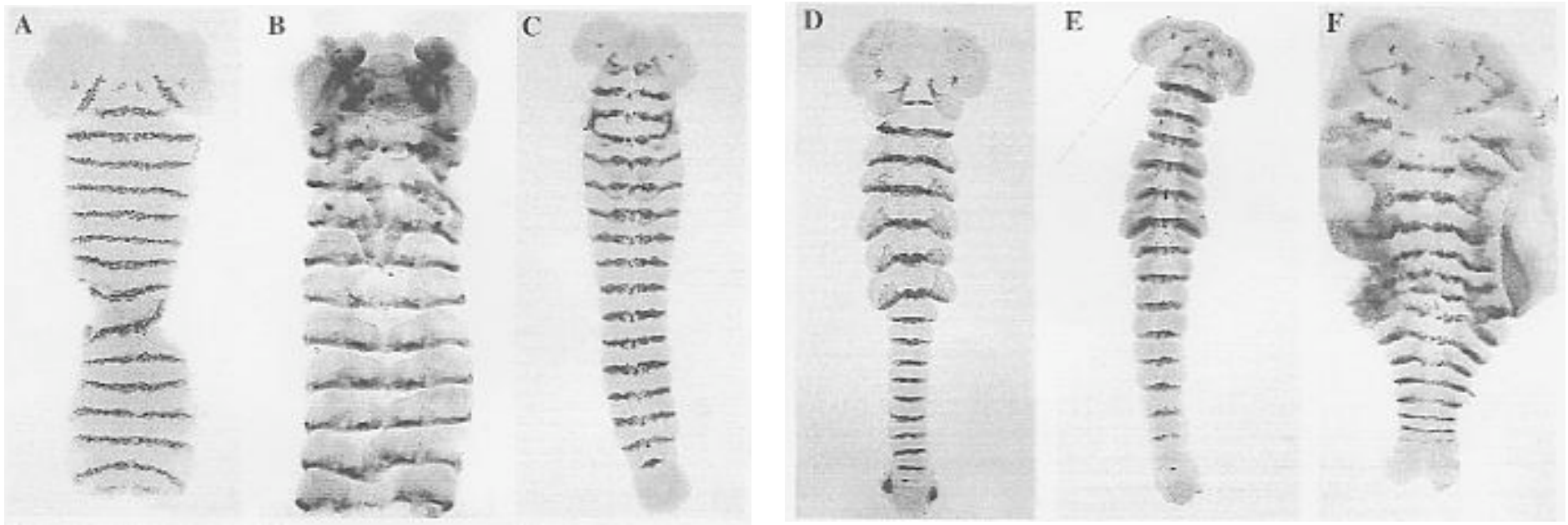
Ectopic Eyes induced by Pax6

Pax6 (regulatory protein) when expressed in antenna, leg, wing -> mini compound eyes, hence title of ‘master regulator’ (Walter Gehring Basel)

http://www.biozentrum.unibas.ch/gehring_pictures.html

Conservation of Engrailed expression

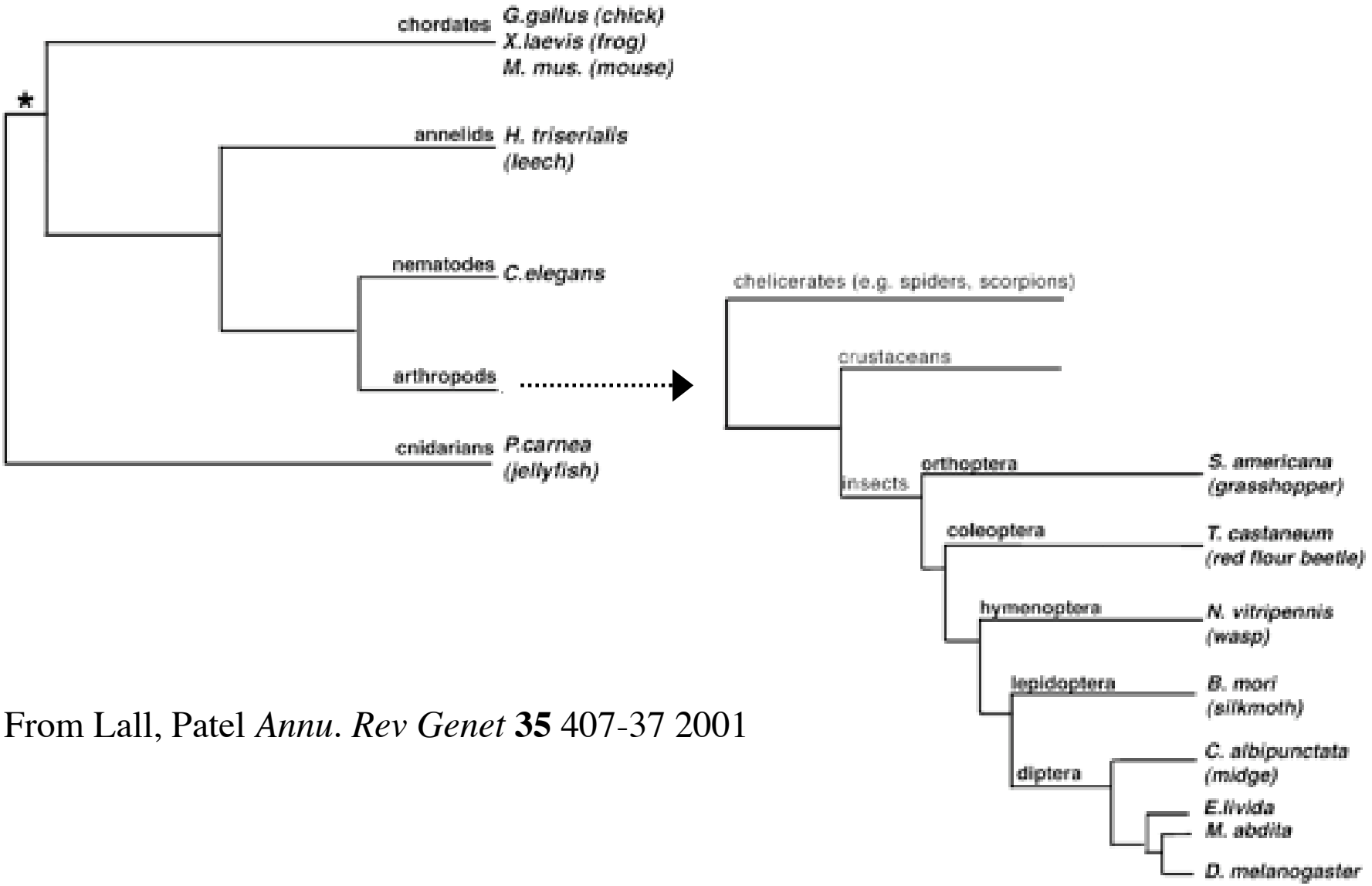
from N H Patel *Development Suppl* 201-207 1994



fly (dissected) moth beetle cricket grasshopper crayfish

Engrailed (and wg) marks segment boundaries in a variety of arthropods

Arthropod Phylogeny



From Lall, Patel *Annu. Rev Genet* **35** 407-37 2001

Segmented development of fly

The head-tail (A-P) differentiation of a fly embryo paradigm of development, proceeds via transcriptional cascade.

Hox + selector genes define cell fates, and specialization of serially homologous segments

Hox genes conserved fly-> human

Mutations -> extra appendages eg 4 winged fly (haltere -> wing), extra vertebrae (homeotic transformations)

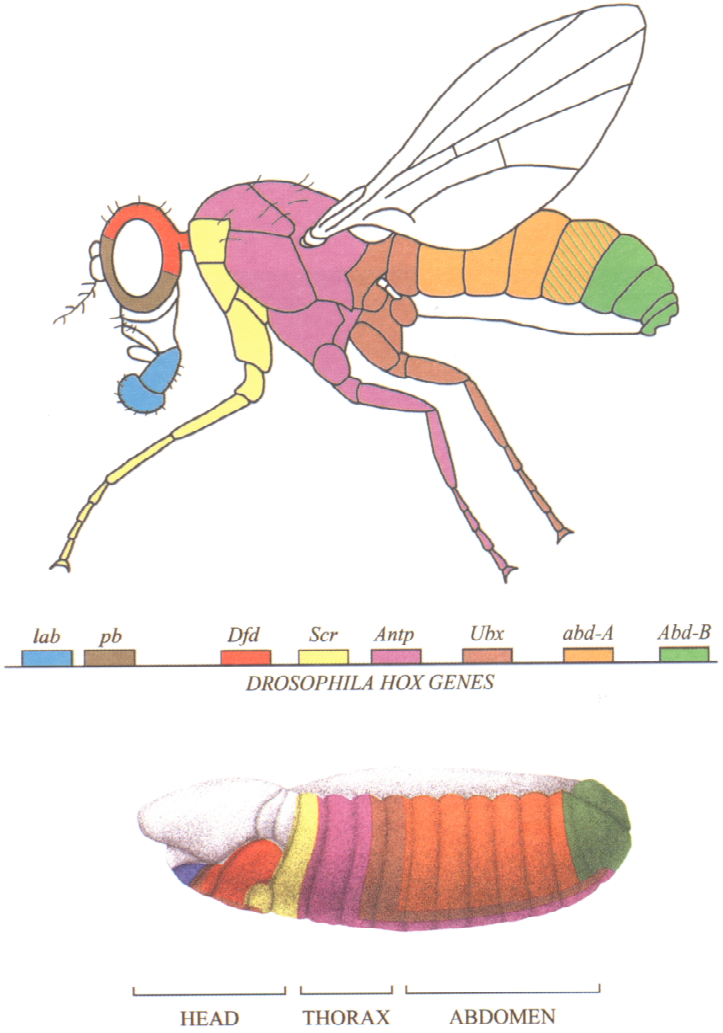


Figure 2.5
The *Hox* genes of *Drosophila*

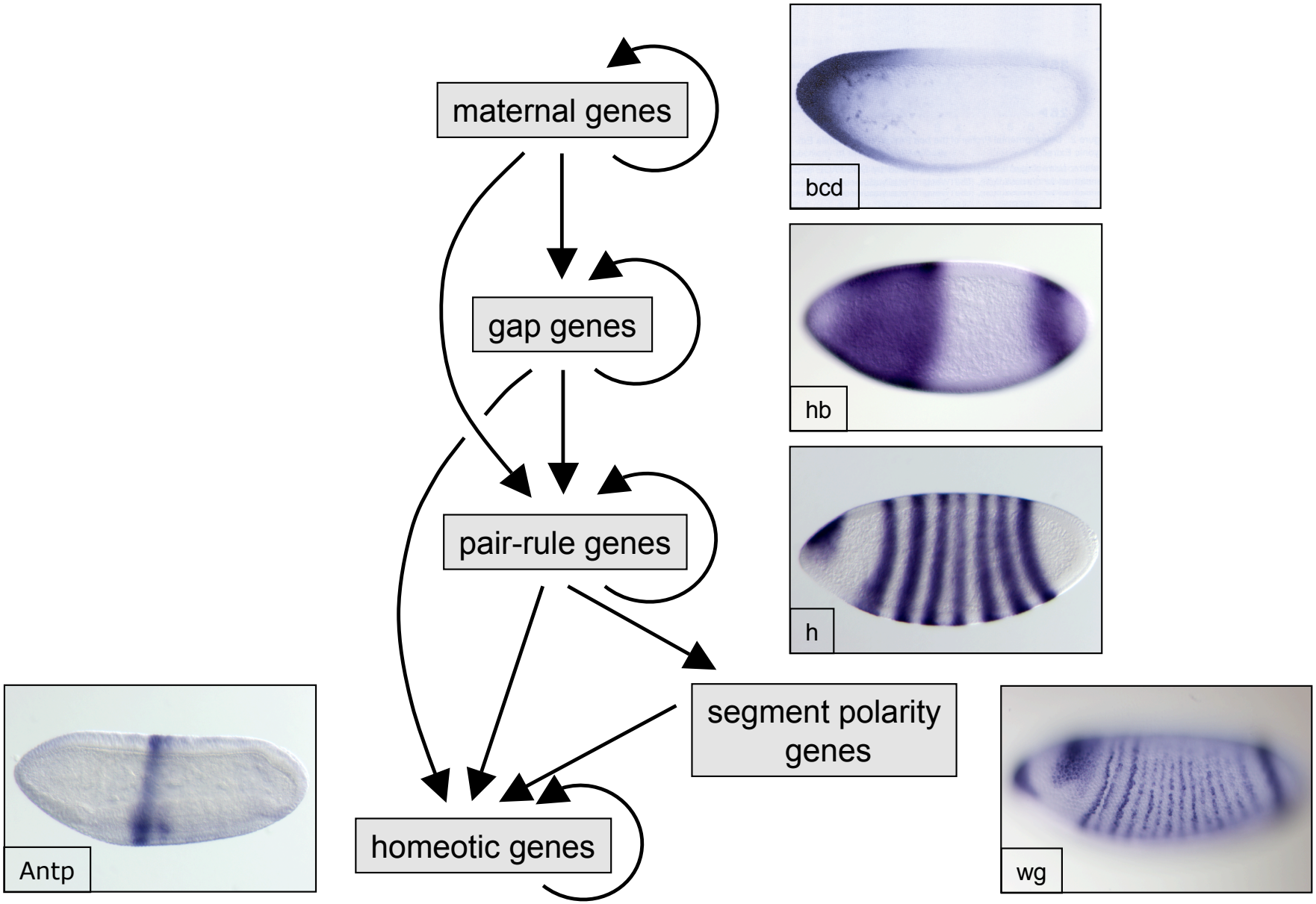
Eight *Hox* genes regulate the identity of regions within the adult (**top**) and embryo (**bottom**). The color coding represents the segments and structures that are affected by mutations in the various *Hox* genes.

Source: Modified from Carroll SB. Nature 1995;376:479–485.

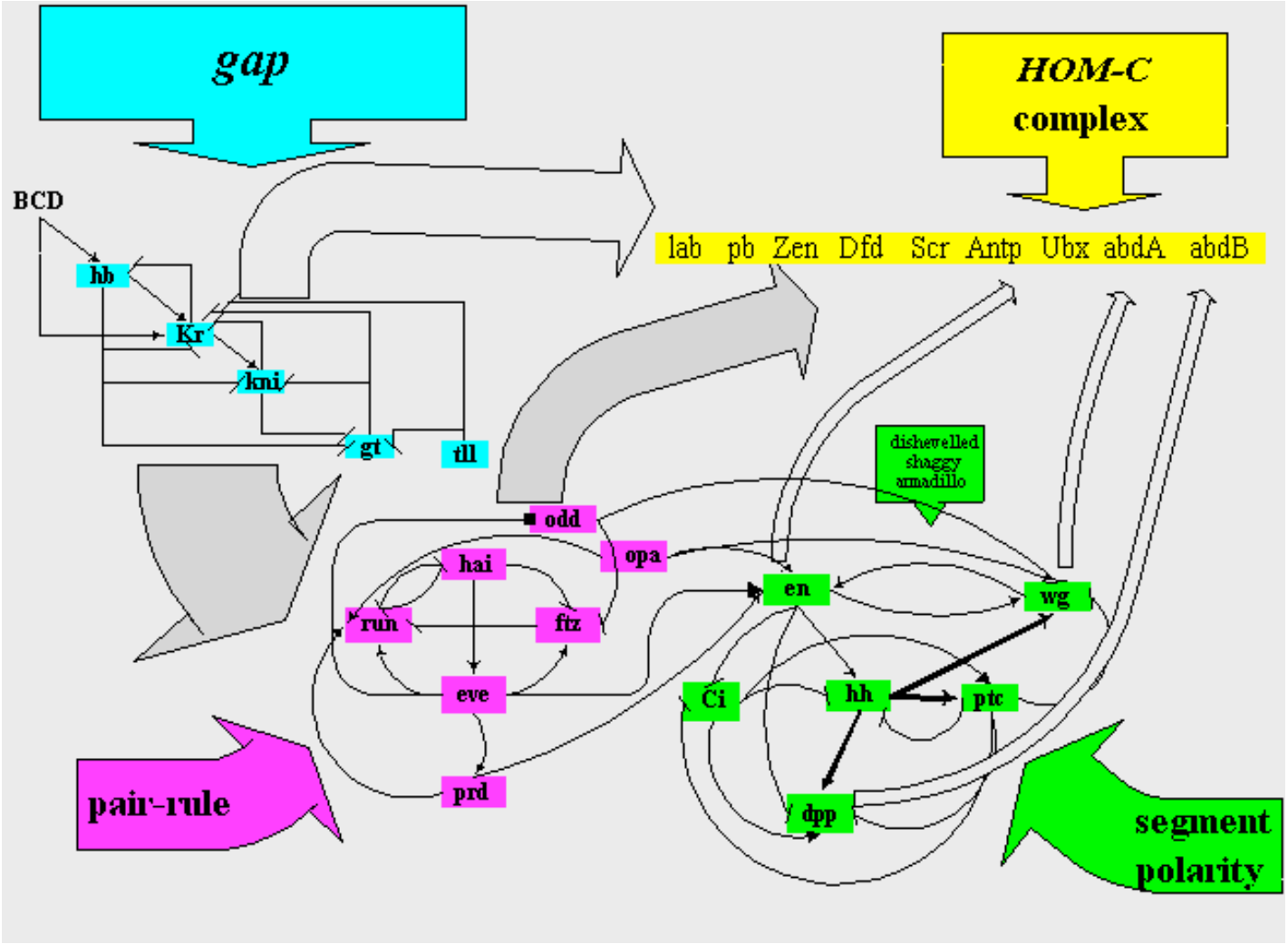
Techniques

- Biochemistry: visualize the location of mRNA and proteins in embryo; cut & paste DNA
- Genetic Screens: mutagenize and look for phenotype, determine location in genome, clone.
- Reverse Genetics: given genomic sequence, cutout, amplify and define function of any fragment.
- Transgenics: domesticate mobile elements, insert new DNA into genome eg attach regulatory DNA to reporter gene
- Crosses: mix mutations, eg mutant background for transgenic fly, or infer pathways

Establishment of segmented body pattern

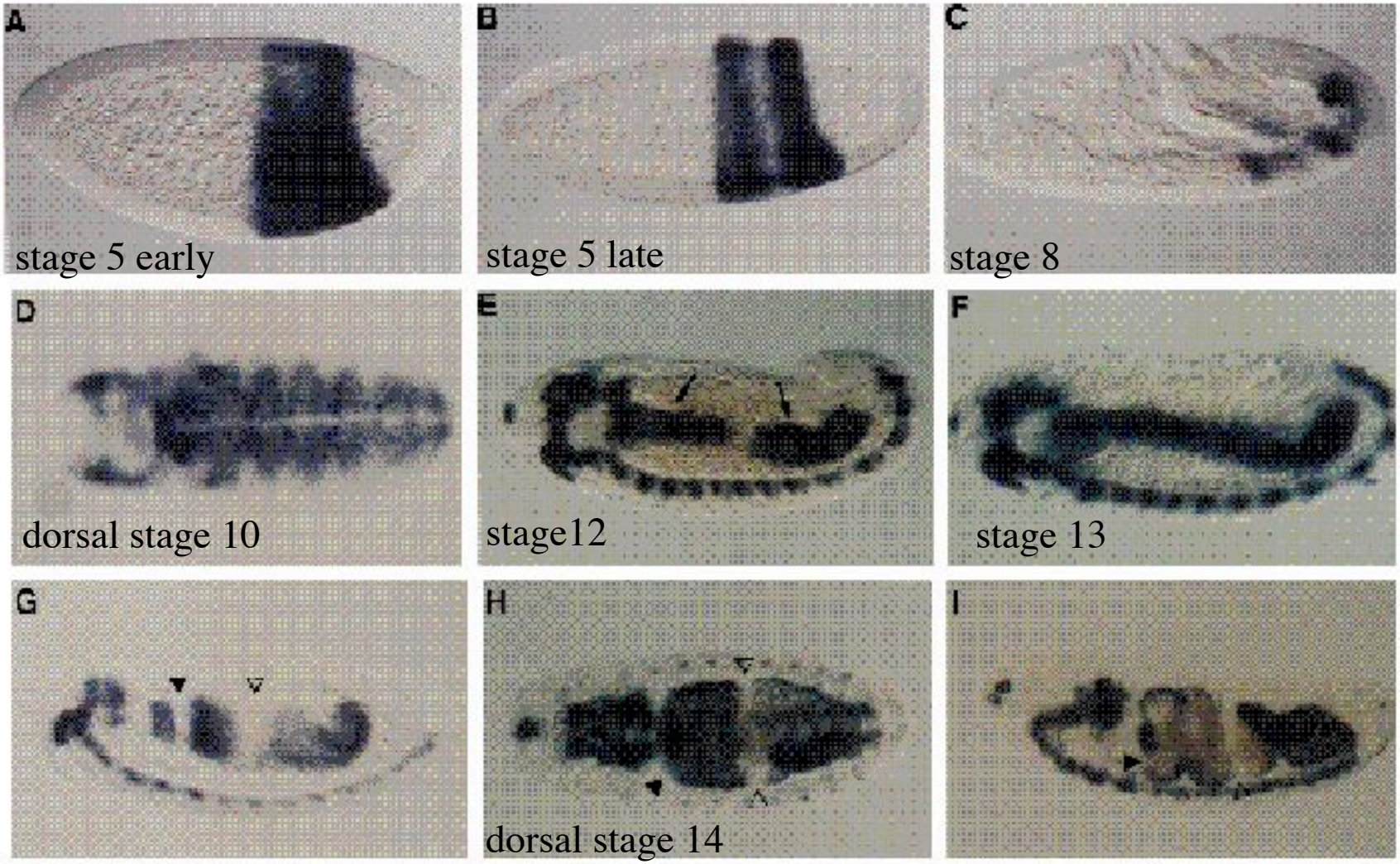


Segmentation gene hierarchy (from genetic screen)



But ~700 blastoderm patterned genes (BDGP insitus)

Expression of nub (24 hrs)



Affolter et al *Development* 117 1199-1210 (1993)

cis Regulatory modules in fly

Genome Statistics:

120 Mb assembled (Release 3)

~14k genes, ~20% genome codes for proteins, ~5% repeats 'junk??'

How is regulatory information encoded in the genome?

(eg analogue to exon/intron structure of genes)

Modules (small scale):

~0.2-1kb regions of genome which convey a position-time specific expression of gene

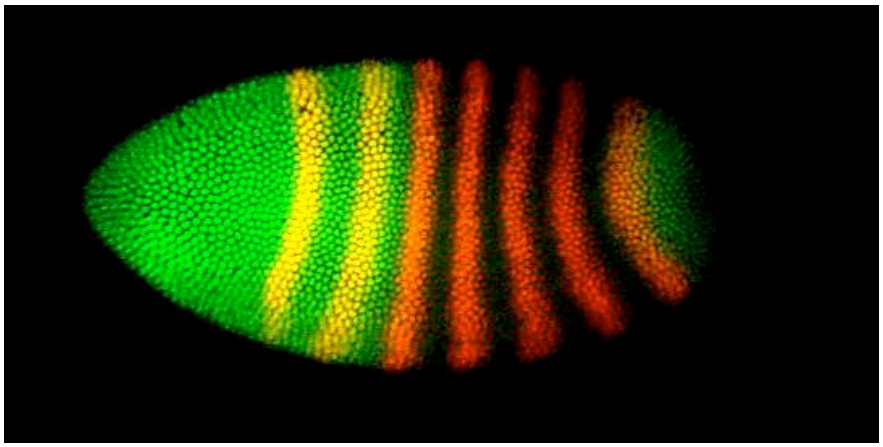
Retain properties in other genomic contexts.

Integrates multiple inputs.

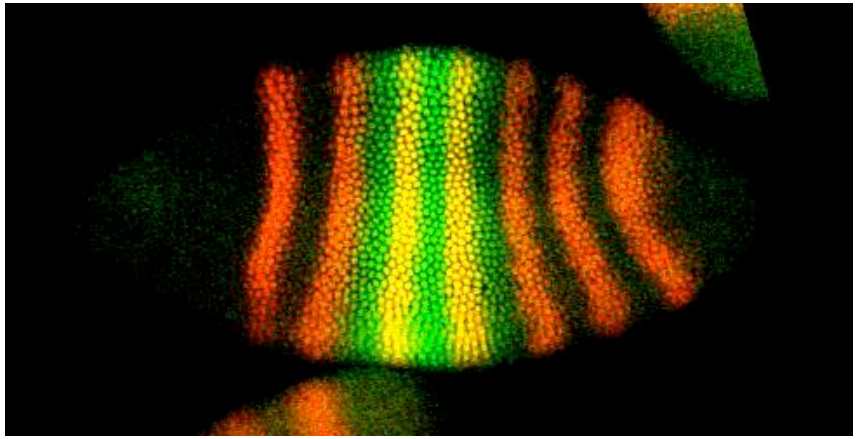
100's modules tested experimentally, expect several/gene

Chromosome domains (large scale)

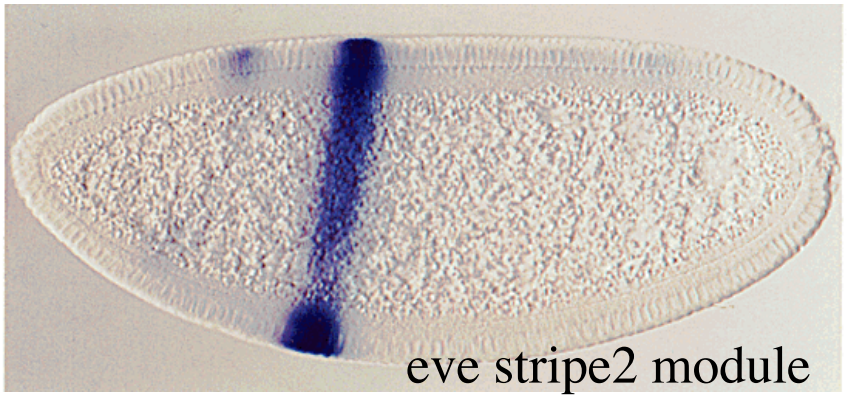
Inputs to pair rule gene *eve* stripe 2

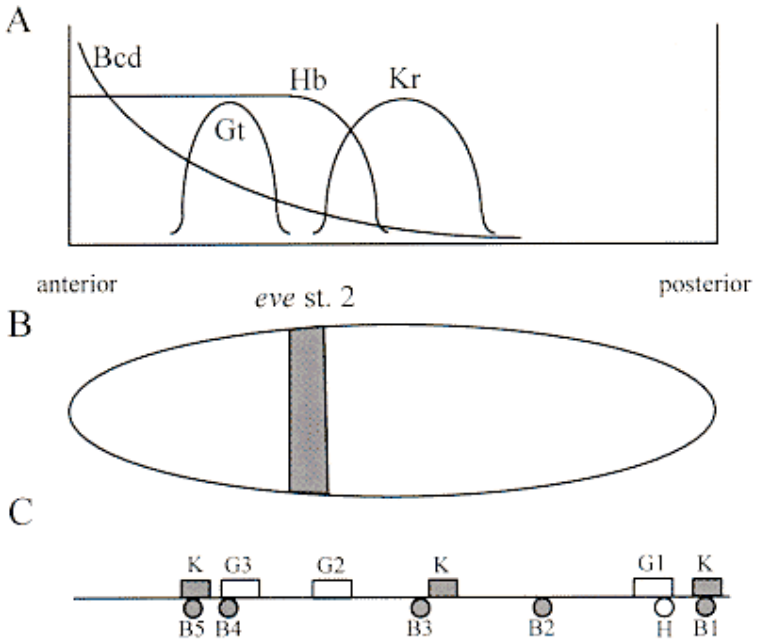


Hb + Eve



Kr + Eve from John Reinitz





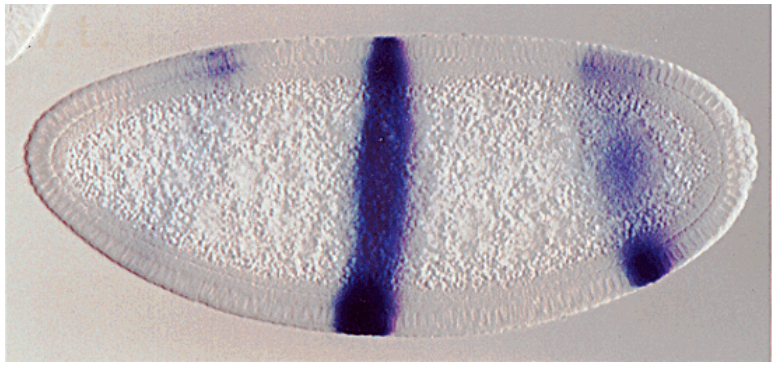
from Steve Small NYU

Finding regulatory modules computationally

Use known binding preferences of regulatory proteins in the pathway + genome sequence + location of genes:

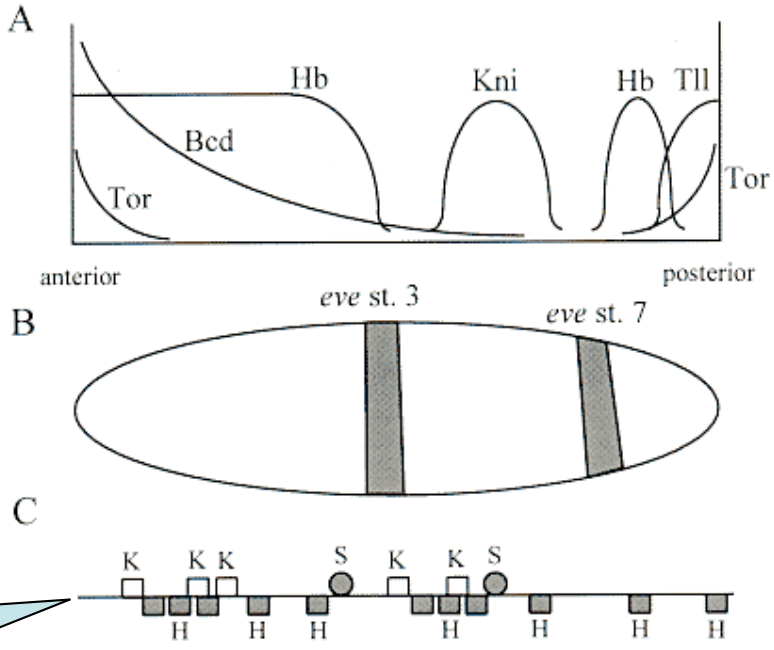
Compute binding free energy of proteins to moving ~500bp window of sequence.

Inputs to eve stripes 3+7



Both stripes see common Kni, Hb repression, and activation by Stat protein which is expressed uniformly.

Similar inputs to stripe 4,6 which share a module (S. Small)



examine binding sites computationally

Ahab (example)

Training data: eg 4 sites for factor 1, (AA, TA, AA, AA); --> matrix..

$$W_1 = \begin{pmatrix} 3/4 & 1 \\ 1/4 & 0 \end{pmatrix}$$ (rows A,T; columns position in motif)

Background: 50% A, T;

Model: Pick w_1 or background with independent probs' $p_1 + p_B = 1$, if select w_1 sample bases with freq, in matrix.

Prob(data = AAT | model) = all ways of partitioning data given w_1 ,backgnd -->

AA | T + A | AT + A | A | T $= p_1(.75*1)\, p_B(.5) + p_B(.5)\, p_1(.75*0) + p_B^3(.5)^3$ eg
($p(AA|w_1) = .75*1$; $p(AT|w_1) = .75*0$; $p(A|$ bckgnd$) = .5$ etc

$= (3/8)p_1p_B + (1/8)p_B^3$; maximize p_1, p_B subject to normalization -->

Prob(AAT | model) = 9/64 (eg $p_1 = 1/4$, $p_b = 3/4$)

Prob(AAT | background only) = 1/8 < 9/64

Score (-free energy) = ln(9/8) > 0

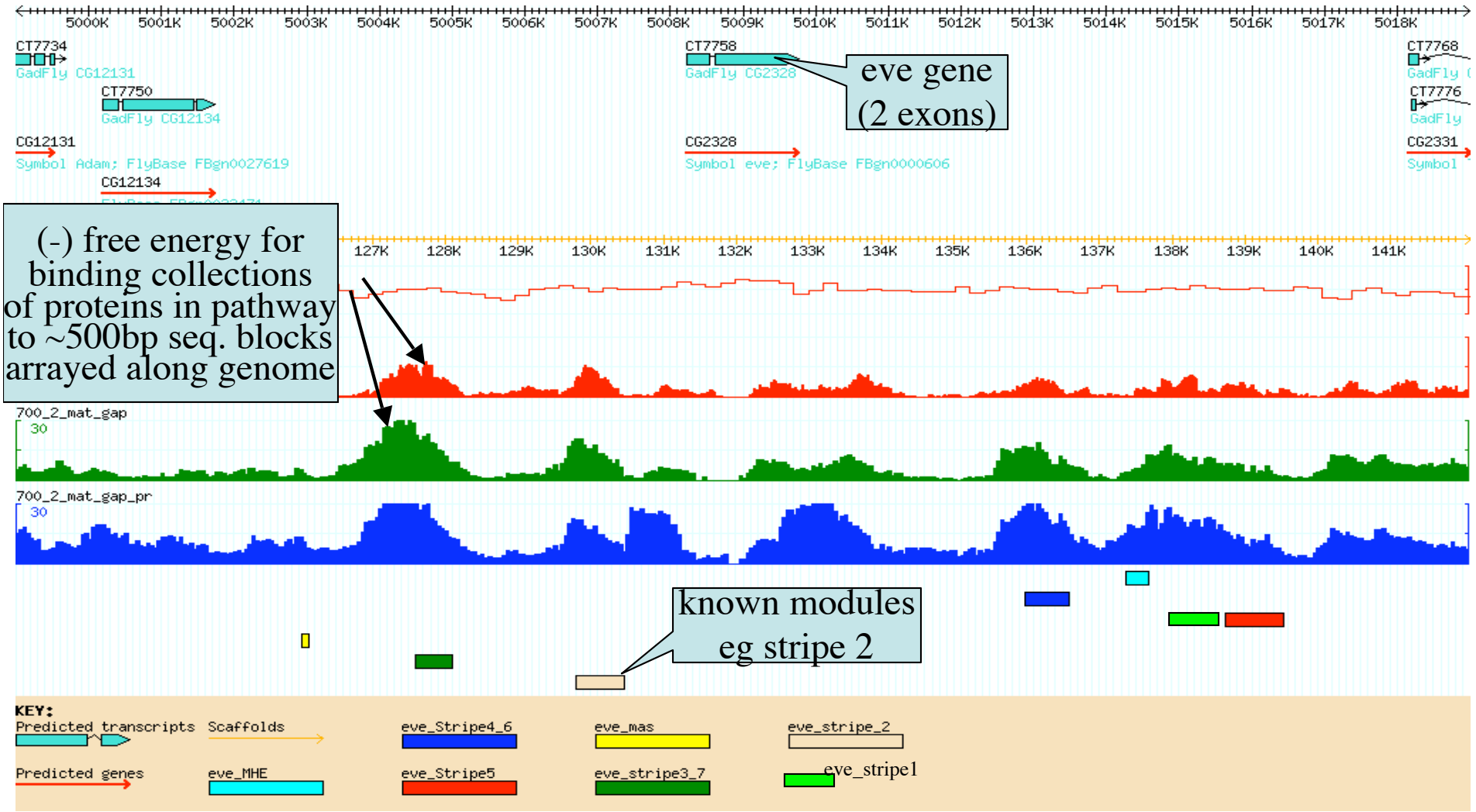
Finding modules in the genome

- Binding sites for all factors in pathway -> weight matrix
- Run through genome with eg 500 bp windows
- Ahab (Stubb 2 species) finds most probable way of generating window sequence from matrices or background.

  Assumes sequence generated by sampling from matrices or background with no correlations between motifs. Finds number and types of matrices which best fit sequence.

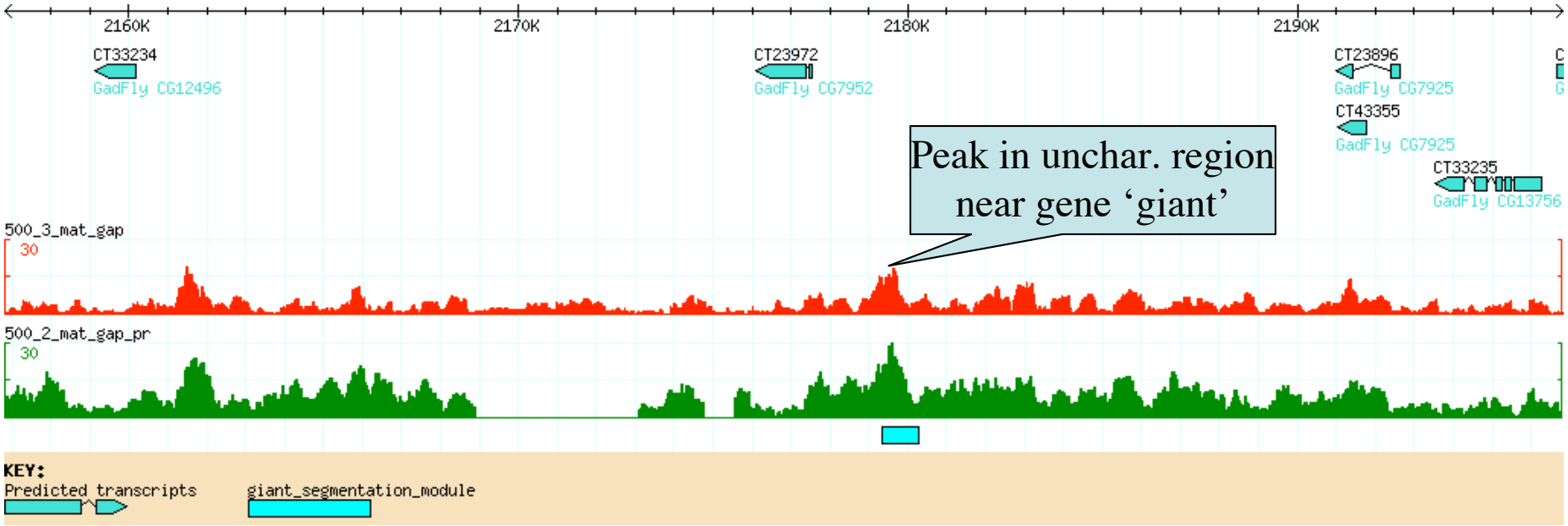
- Thus, no factor dependent thresholds, multiple weak sites scored along with strong sites -> one overall rank
- No free parameters.

Compute regulatory modules:
(using genome + binding energies of proteins in pathway)

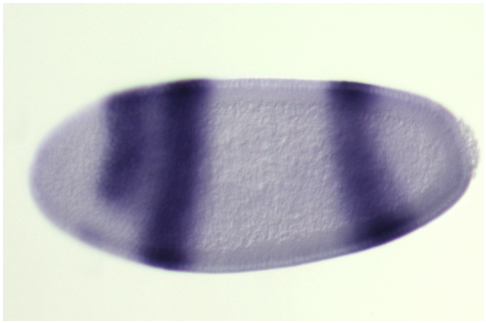


New module near giant

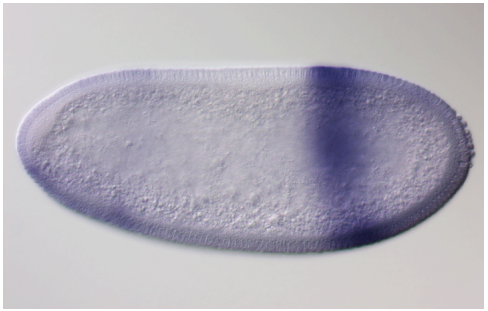
(+12 others: Schroeder et al PLoS ‘04)



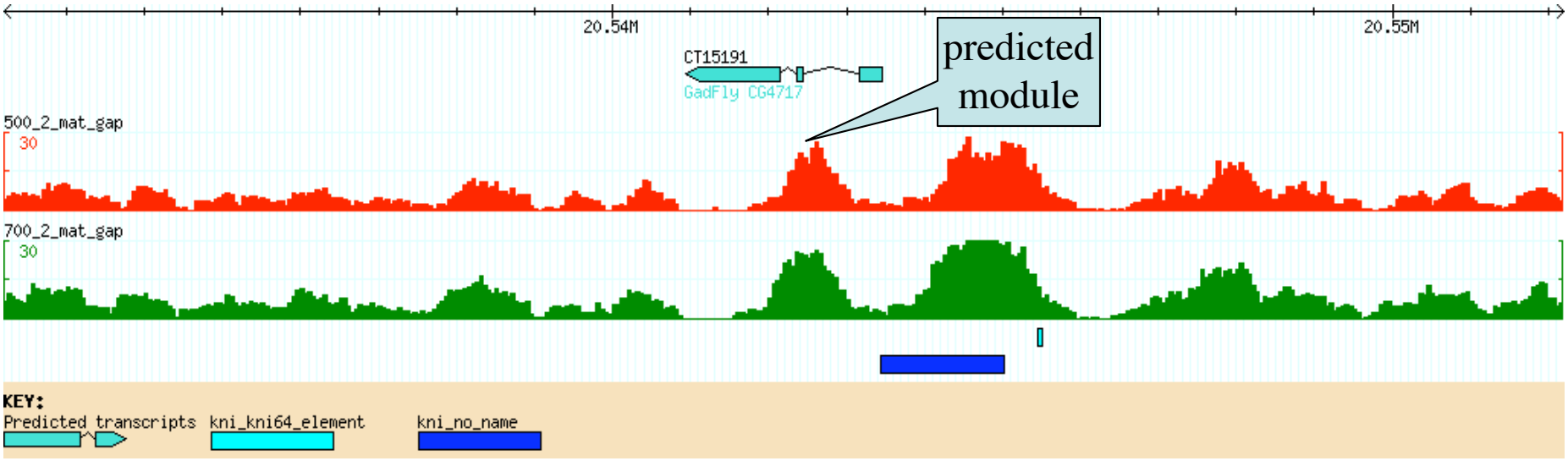
endogenous



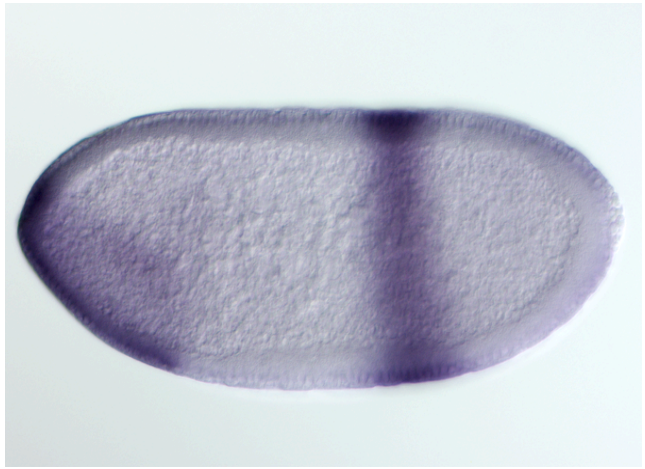
upstream module



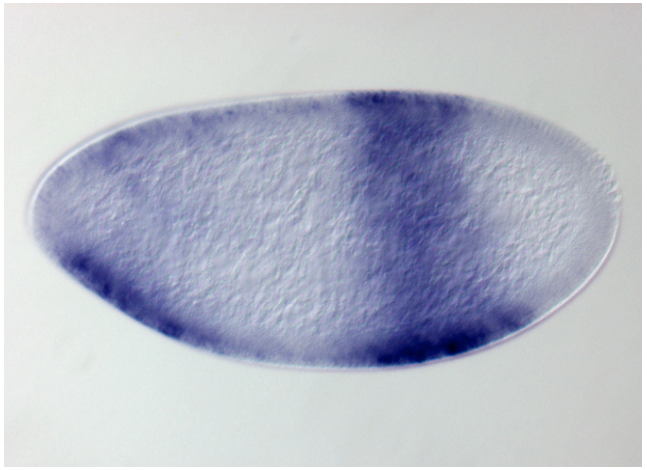
New module in knirps intron



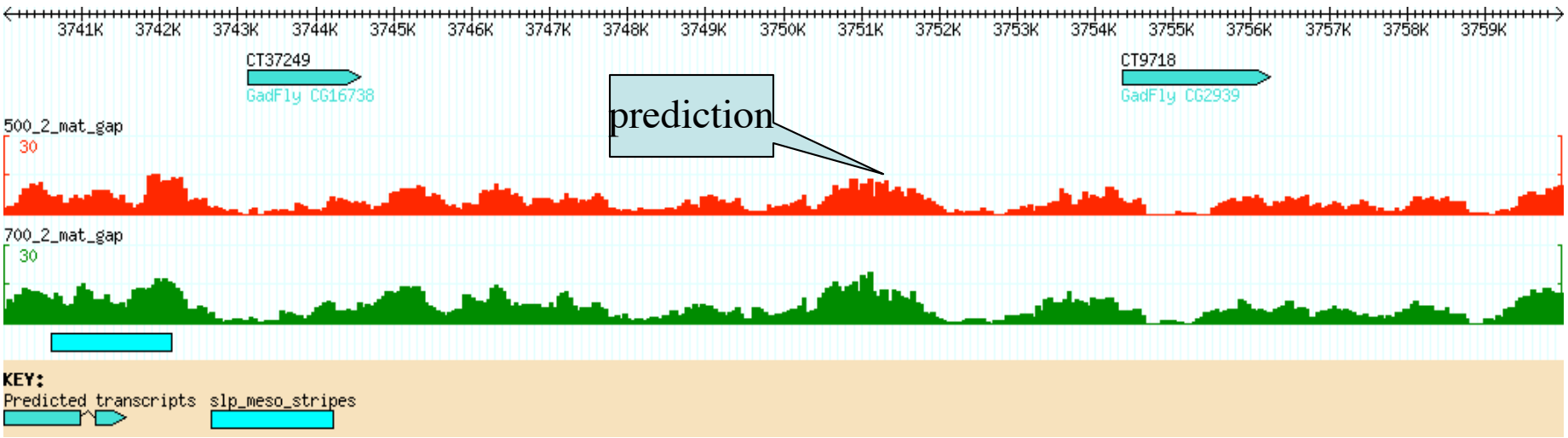
endogenous



module lacZ

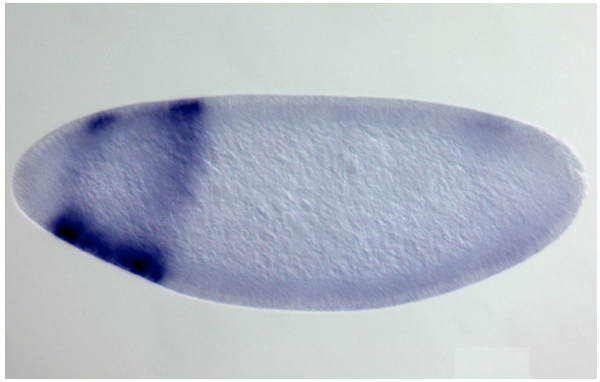


New module slp1,2

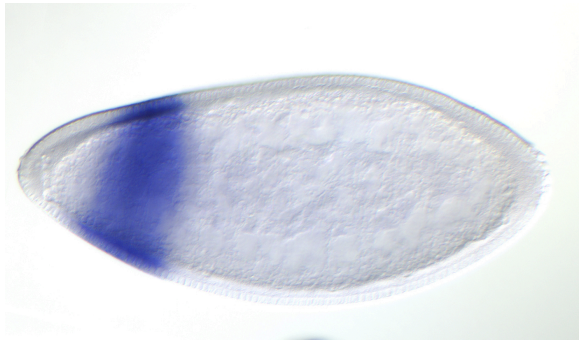


endogenous slp1

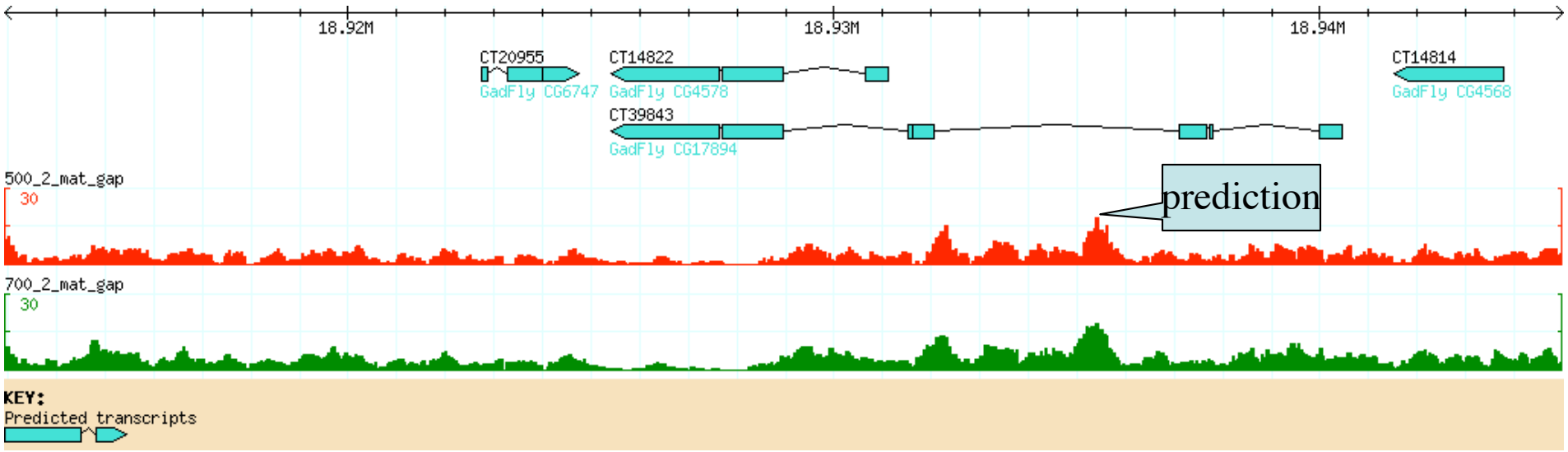
module -lacZ



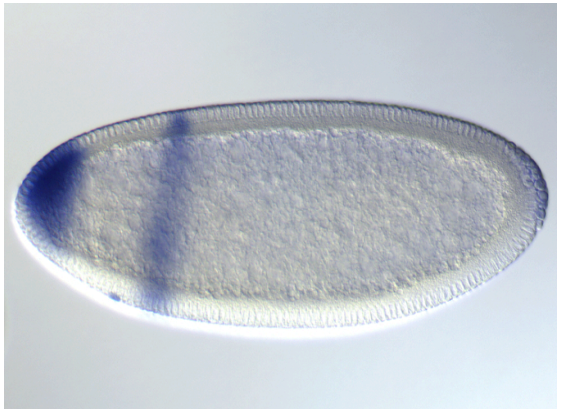
endogenous slp2 (BDGP)



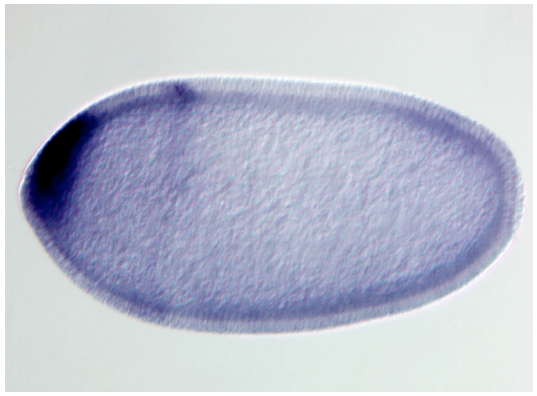
New cnc module in intron



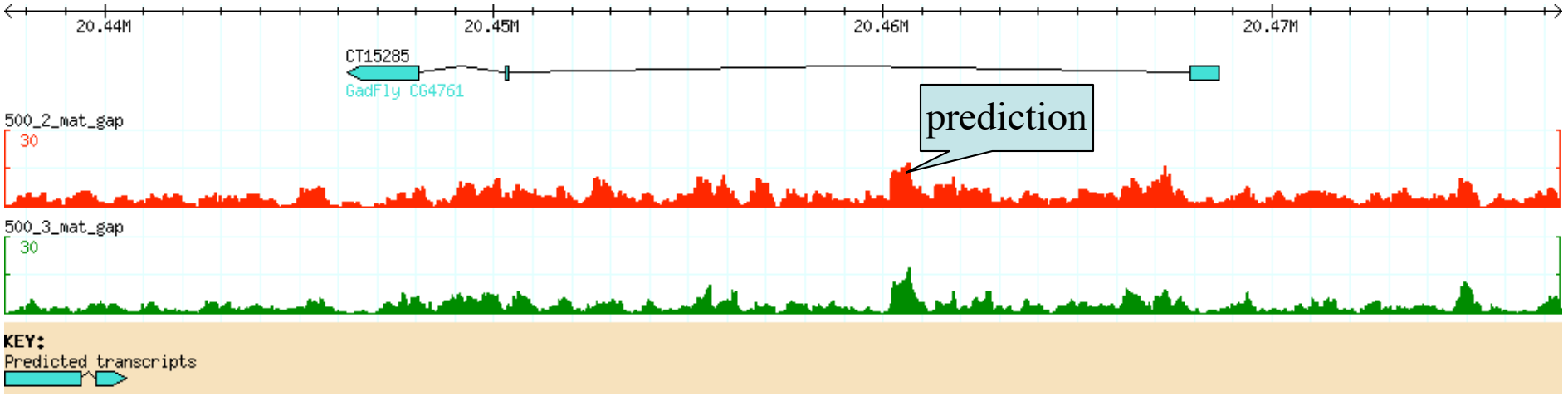
endogenous



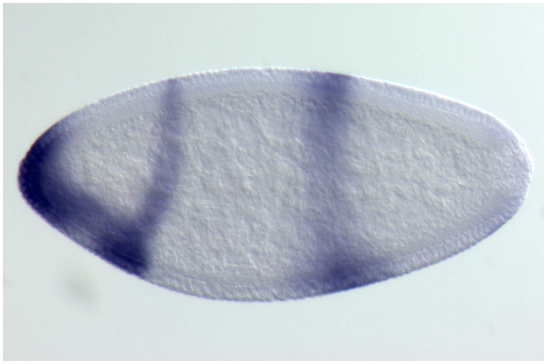
module -lacZ



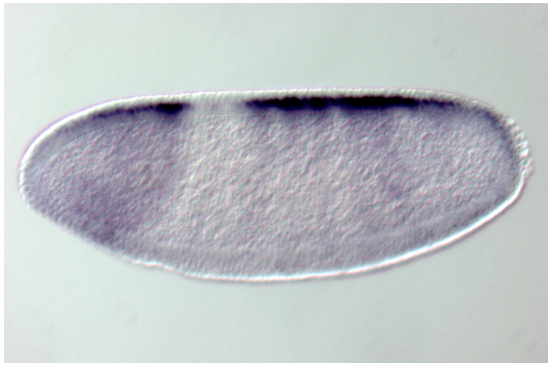
New module knrl intron



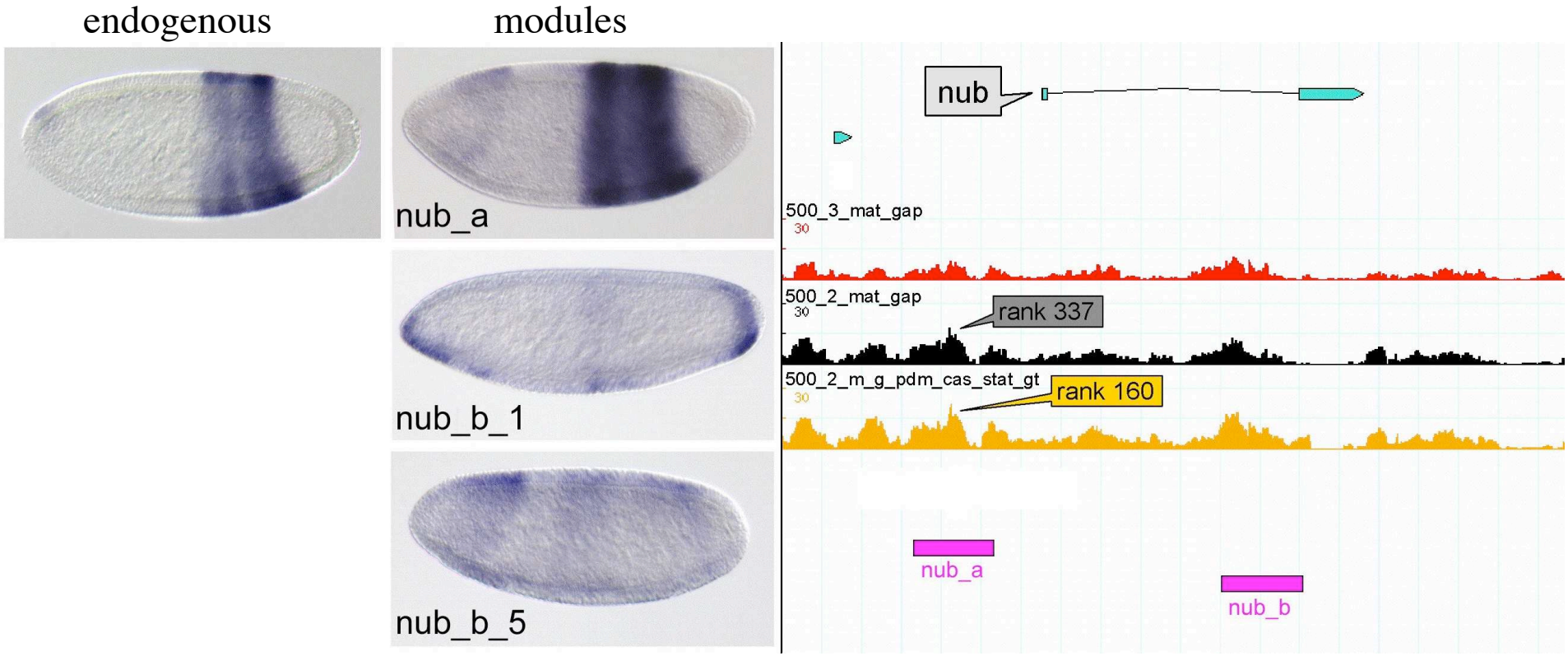
endogenous



module -lacZ



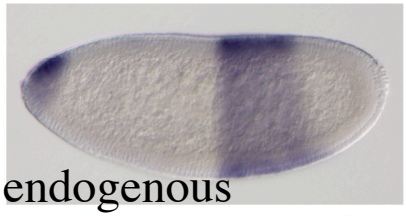
nub (pdm1)

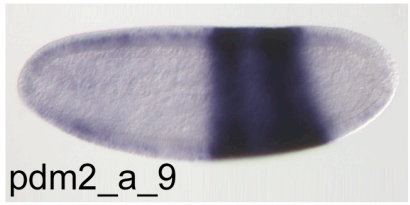


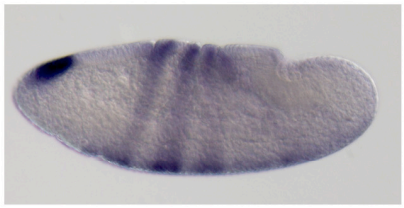
The module *nub_a* gave strong reproducible blastoderm expr, nub_b weak, variable in blastoderm, but strong in neural.

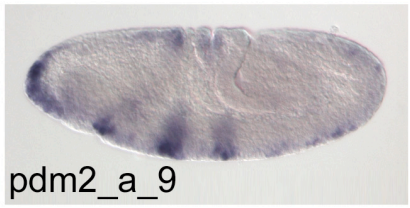
pdm2

(b) POU domain protein 2 (pdm2)

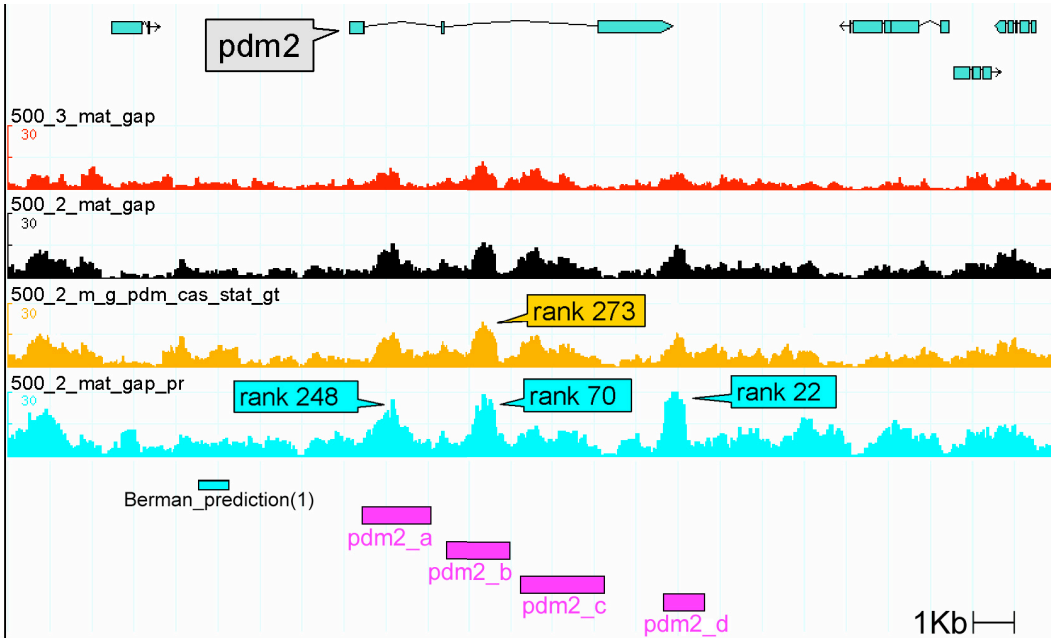




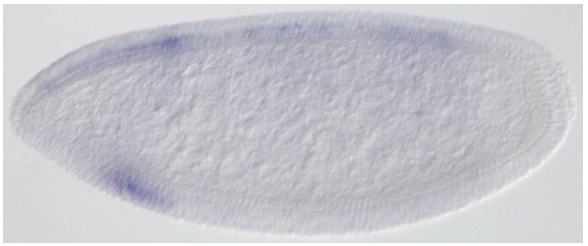




pdm2_a strong invariant expr.

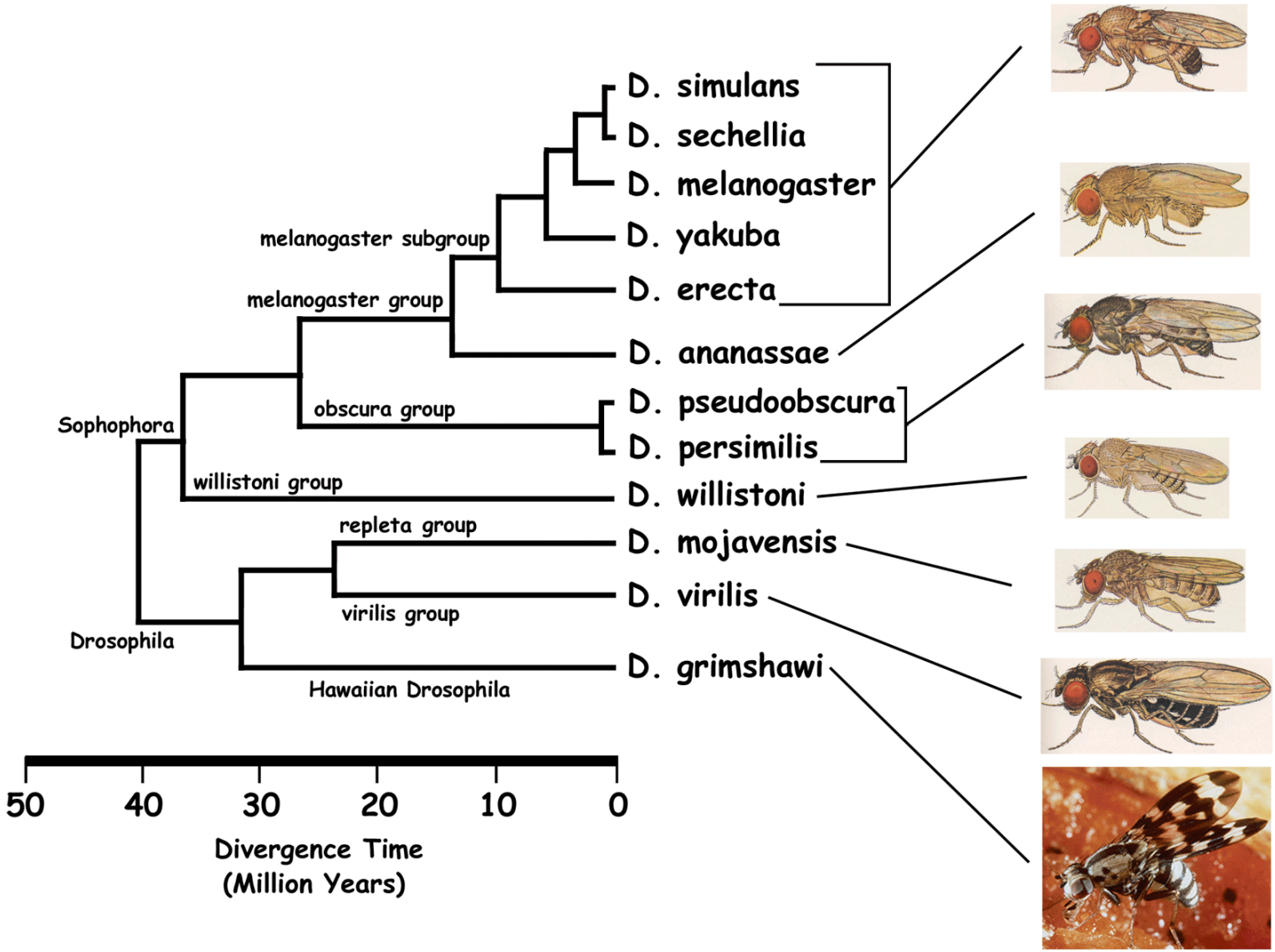


pdm2_b module: variable weak expression from various lines (auto-repress)



Fly Genome Collection

(Scale: D.mel-D.ere ~ human-mouse via synon. mutations)



Why Compare Species

1. Homologous sequence is easy to map, and infer changes
   (eg D.mel vs D.psu, noncoding, order preserving map on ~10kb scale with 30-40% nucleotide conservation.)

2. What sequ. is conserved -> functional modules, binding sites
   (ie functional regions evolve more slowly than neutral rate eg useful screen for functional genes in yeast)

3. What's different -> evolution of gene regulation
   (how does regulatory sequence change eg point mutations, inserts.. how do binding sites move, and does that change expression of cis regulatory module. Can expression of individual modules change, with the gene invariant.)

Known binding sites vs interspecies conservation**

•Optimize parameters in alignment-synteny code for prediction of binding sites.
•Assess significance via randomizing sites.
•Conclusion: weak correlation between known binding sites and conserved interspecies sequence.

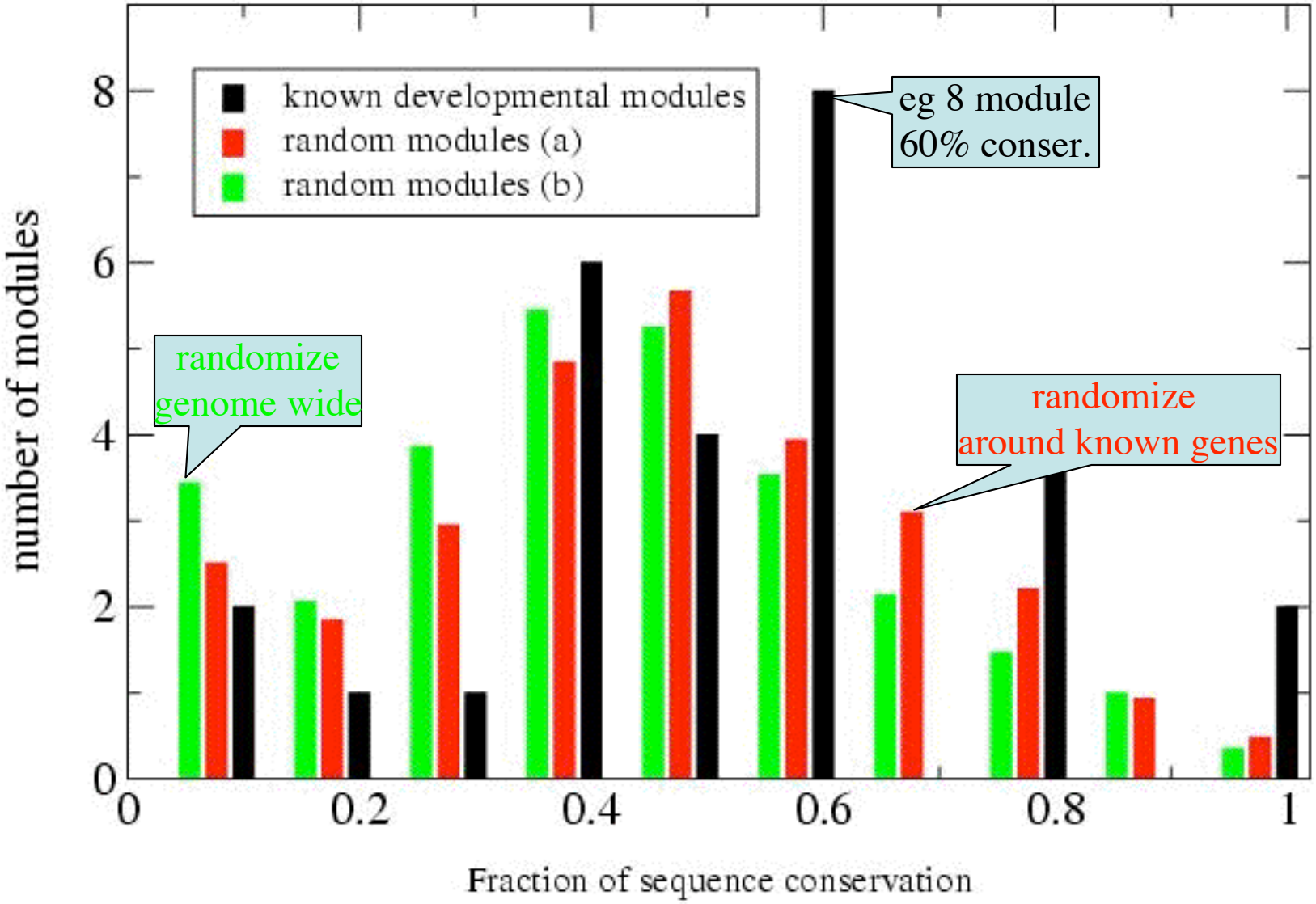
	genome sequ.		known sites(bp)		randomize sites	
Species:	total	cons.	total	cons.	cons.	z-score
D.pseudo	223k	98k	3300	1520	1130	4.3
D.virilis	33k	10k	2200	700	550	2.

(two alignment programs tried; 'hit' defined by inclusion of site or bp overlap)

(*Emberly et al BMC *Bioinformatics* **4**, 51 2003)

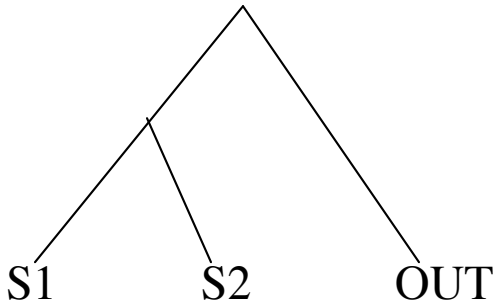
Sequence conservation in 30 known modules

(known position and length of each module used)



Inferring sequence change from alignments

S1 AAAA ACTTACTTCGTT
S2 AATAACGTACTT CGTT
Out AAAATCGTAGTA CGTT



Substitution (S1-S2) those between S1,2 and OUT not scored.
Deletion from S1 (not insert to S2),
Insertion to S1 via tandem duplication of ACTT.
Tandem duplications can mutate ACTT *T* ACTT AGTT (3 copies ACTT)

A C T
ACGGTTA T *ACT T* . . . *unpairing of replicating strand, and recopying of TGAA*
TGCCAATTGAAGCAA *(strand being copied)*

How sequence changes

Data:

Compare D.mel with D.yakuba (10 My) with D.psu as outgroup (25 My)

1. More base pairs change due to insertion/deletion events than point mutations (but fewer events)

2. Excess of insertions over deletions (most events local copying)

Given rapid loss of nonfunctional sequence (half life 10^7 y) (2) -> most sequence functional (fly genome ~ estimates of functional size of human genome ie 5% of total, based on human-mouse comps)

Statistics of changes in noncoding sequence

Data:

Compare D.mel with D.yakuba (10 My) with D.psu as outgroup

use 100kb embryo patterning modules (Statistics as fn of gap penalty)

Indels/substitutions (D.mel vs D.yak)

0.13 (events), 1.2 (coverage)

Insert/Deletion

1 (events), 2 (coverage) D.mel (resp 2, 8 D.yak)

(both species incr. vs common ancestor D.yak more than D.mel)

Tandem Repeats: ('mini-satellites', 2-3 copies of 5-10bp)

25% (coverage) D.mel (45% D.yak) vs 3.7% random.

Given recovery rate implied by point mutations, all inserts could be tandem repeats.

Most noncoding sequence is functional

- Nonfunctional sequence (ie pseudogenes) lost from D.mel with half life of 14My, thus shrinks 4x since common ancestor D.mel, D.psu. (Petrov Hartl 1997)

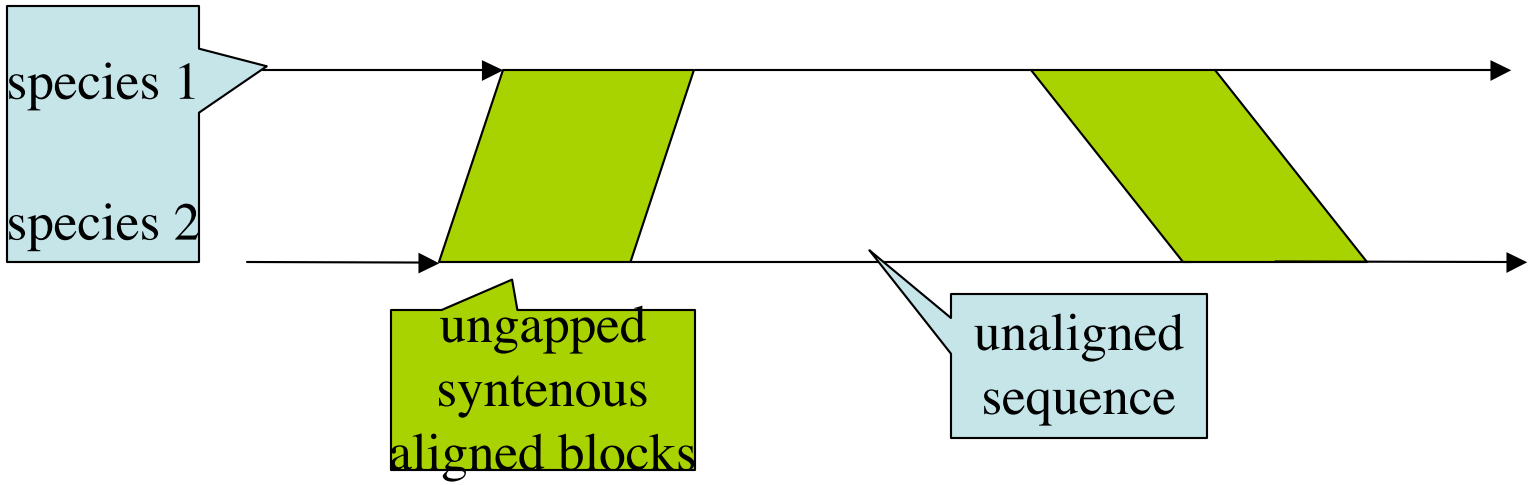
- Length of homologous ~1kb pieces noncoding DNA retains size between D.mel , D.psu.

- Most of D.mel and D.yak (D.psu outgroup) noncoding sequence increasing vs common ancestor (rules out uniform shrinkage, which would preserve pairwise lengths)

Locating modules and binding sites

Stubb* fits number and type of binding sites from a set of weight matrices for factors in pathway. No empirical thresholds/cutoffs. For two species, evolution model used for aligned blocks, other sequence assumed independent.

*(S. Sinha et al *Bioinformatics* **19** i292-301 2003, http://stubb.rockefeller.edu stubb on line)



Stubb Algorithm

Model for data = ∑all ways of creating sequence by sampling

1. wtmx's (with optional positional correlations)
2. 'background' (n-th order markov model)

Maximize *P(data | model)* over p_{wtmx}, $\sum p_{wtmx}$, $p_{bckgnd} = 1$.

In aligned blocks, create sequence in ancestor, then biased mutation model (same wtmx binds) ancestor(a)->species(b_1, b_2..)

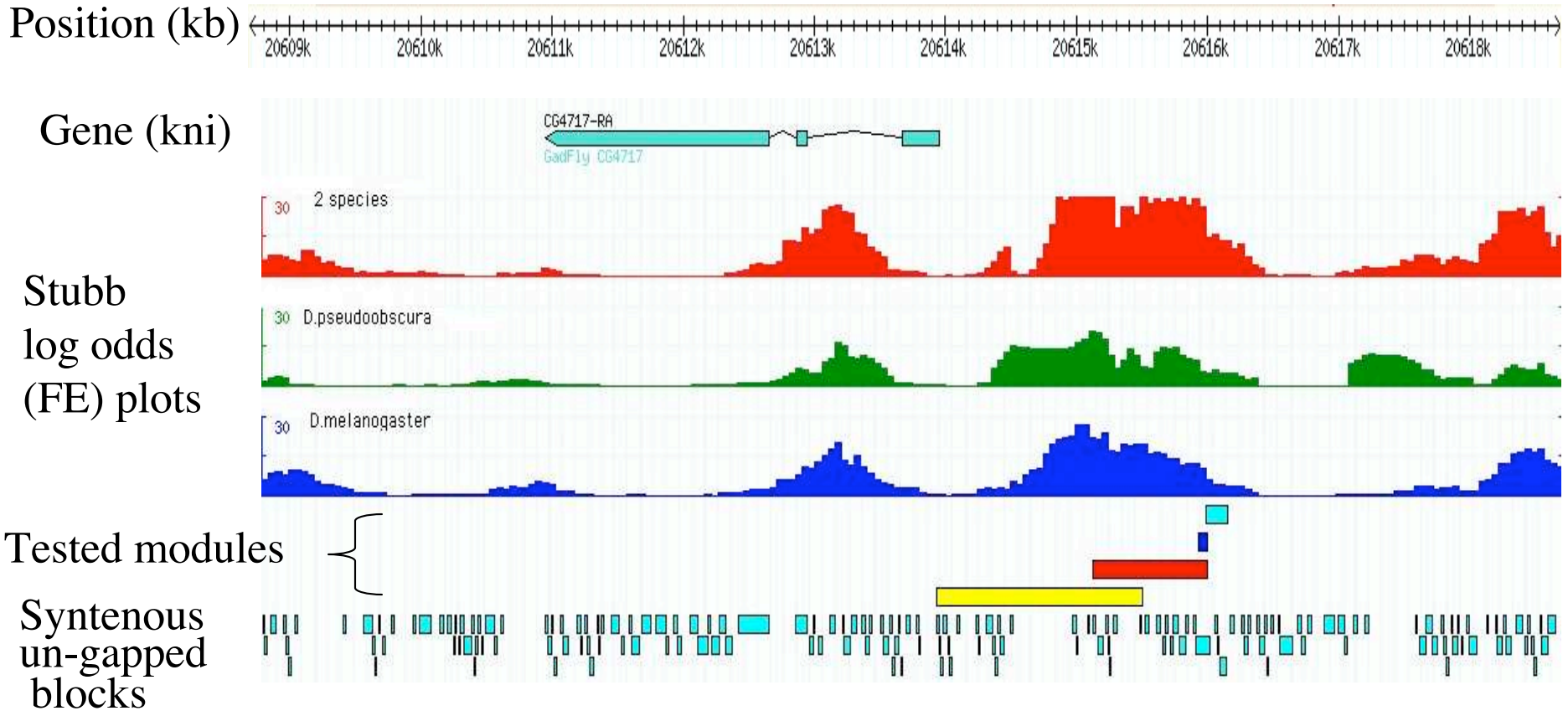
$$P(b_1, b_2) = \sum_{ancestor\ \ a} P(a)P(b_1 \mid a)P(b_2 \mid a)$$

$$P_t(b \mid a) = \ _{a,b}e^{\ t} + p_b(1 \quad e^{\ t}) \qquad P_{(t+s)}(c \mid a) = \sum_{b} P_t(c \mid b)P_s(b \mid a)$$

where conditional probability $P_t(b \mid a)$ has correct limits as t->0, inf, and composition in time,

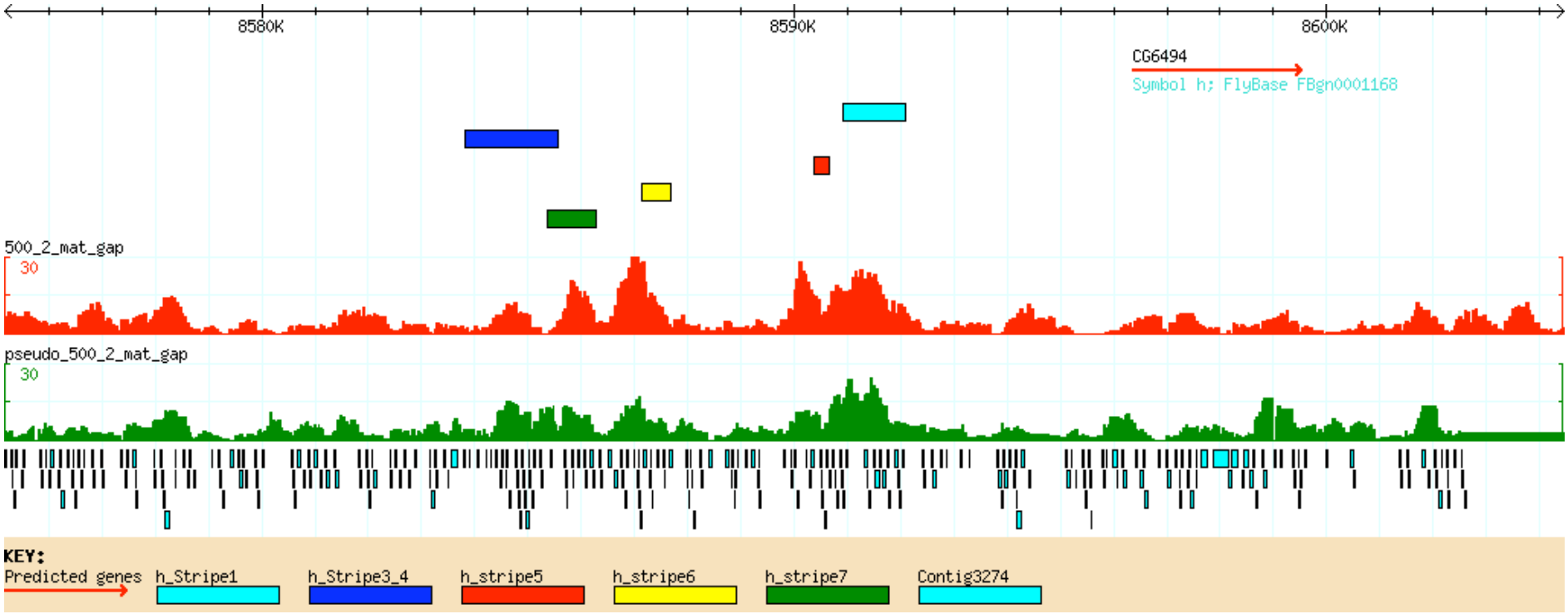
p_b -> probability of base *b* (over each column of wtmx)

Interspecies comparisons: integrated display



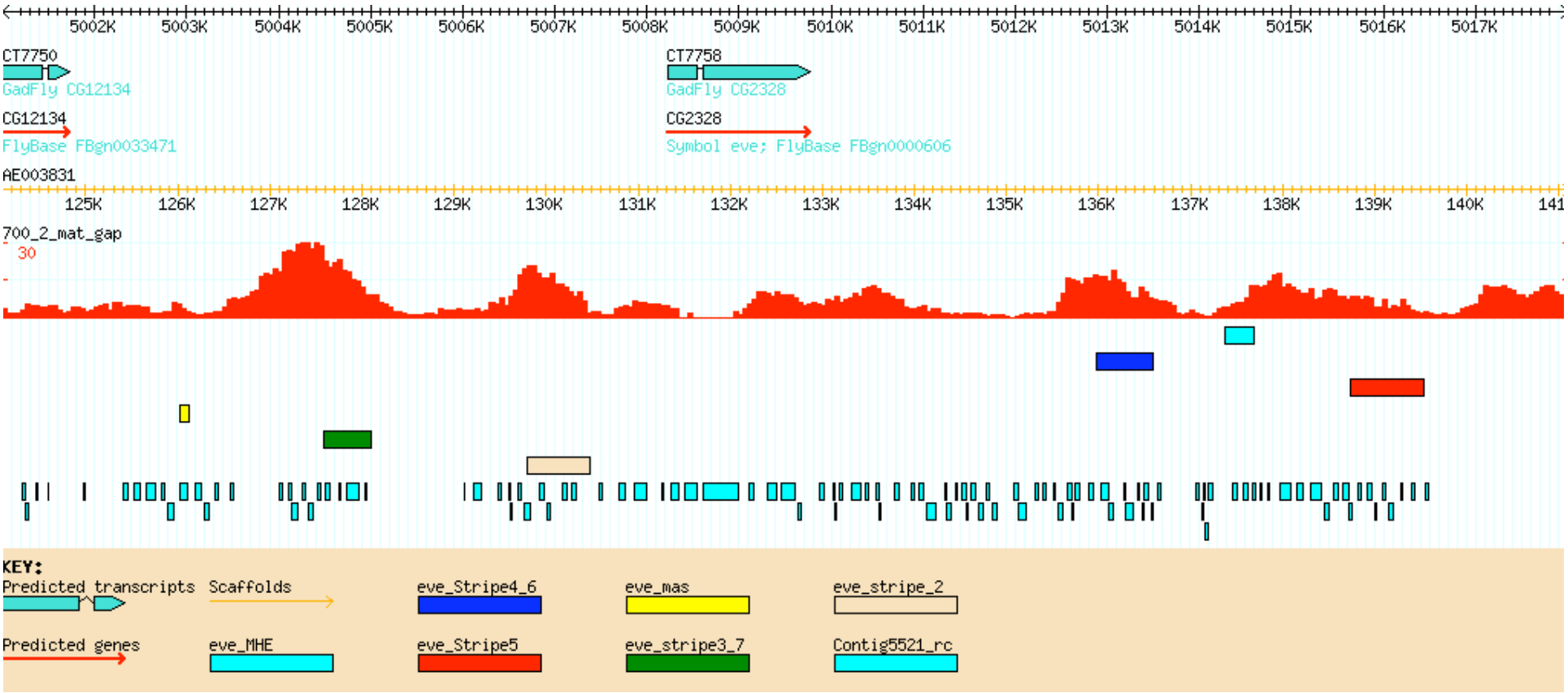
Interspecies comparisons *h* (D.mel vs D.psu)

hairy upstream region. Ahab plots D.mel, D.psu



Interspecies comparisons *eve* (D.pseudo)

region surrounding eve



Regulatory evolution: Comparing cis modules

Do computational screen ~100 modules; find examples of following:

synergistic $FE_{2species} = FE_{mel} + FE_{psu}$

(one set of parameters fits data of both species)

antagonize: $FE_{2species} \leq \max(FE_{mel}, FE_{psu})$

(contents of homologous module has changed)

For a given gene, examples of duplicate (paralogous) modules

(expression of these modules more likely to change)

Interesting sequence changes:

Large (~module) inserts with plausible binding sites,

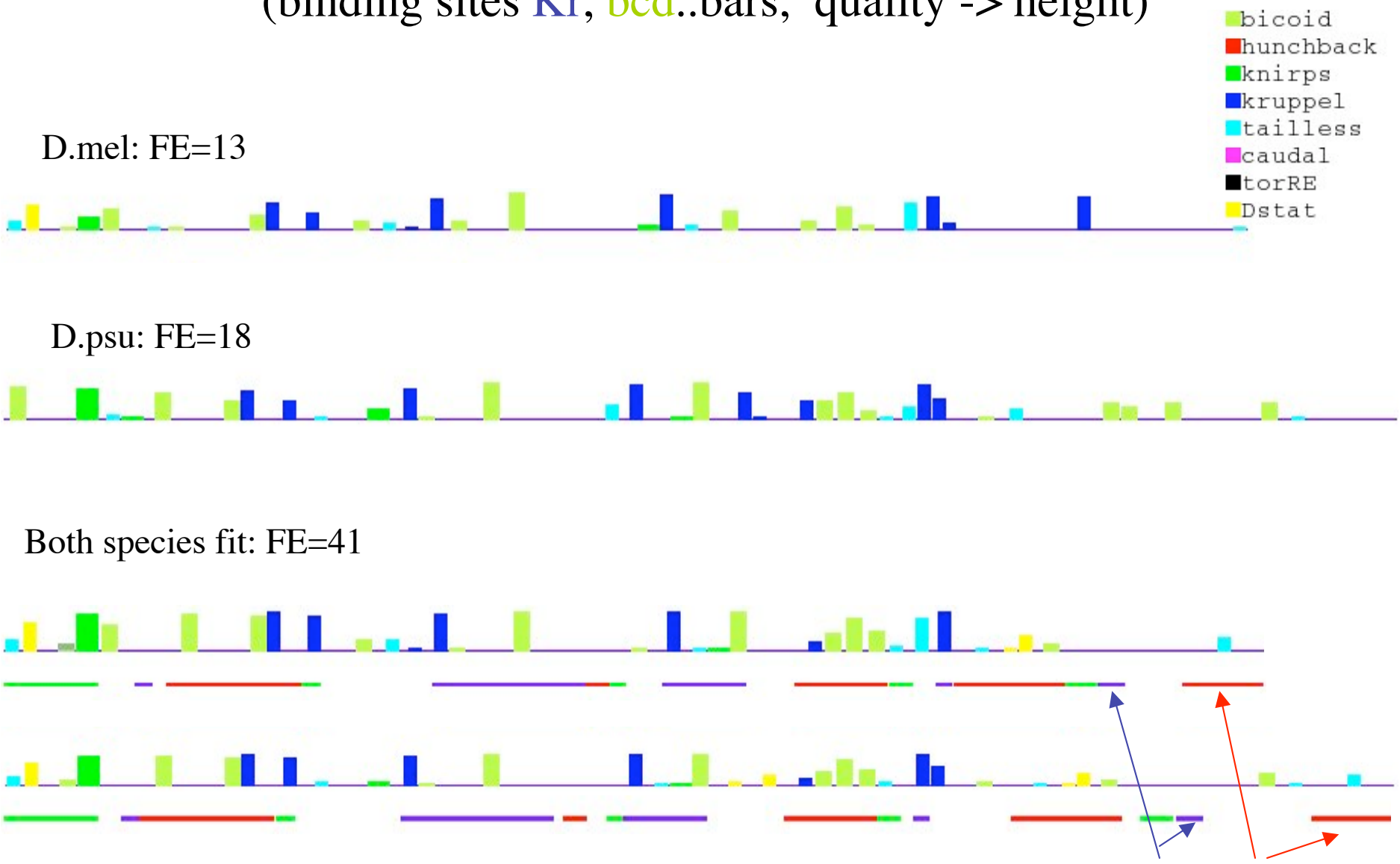
Tandem repeats carry interesting binding sites.

Test a subset in each category.

(Subject intertwined with 'module grammar' for functional module)

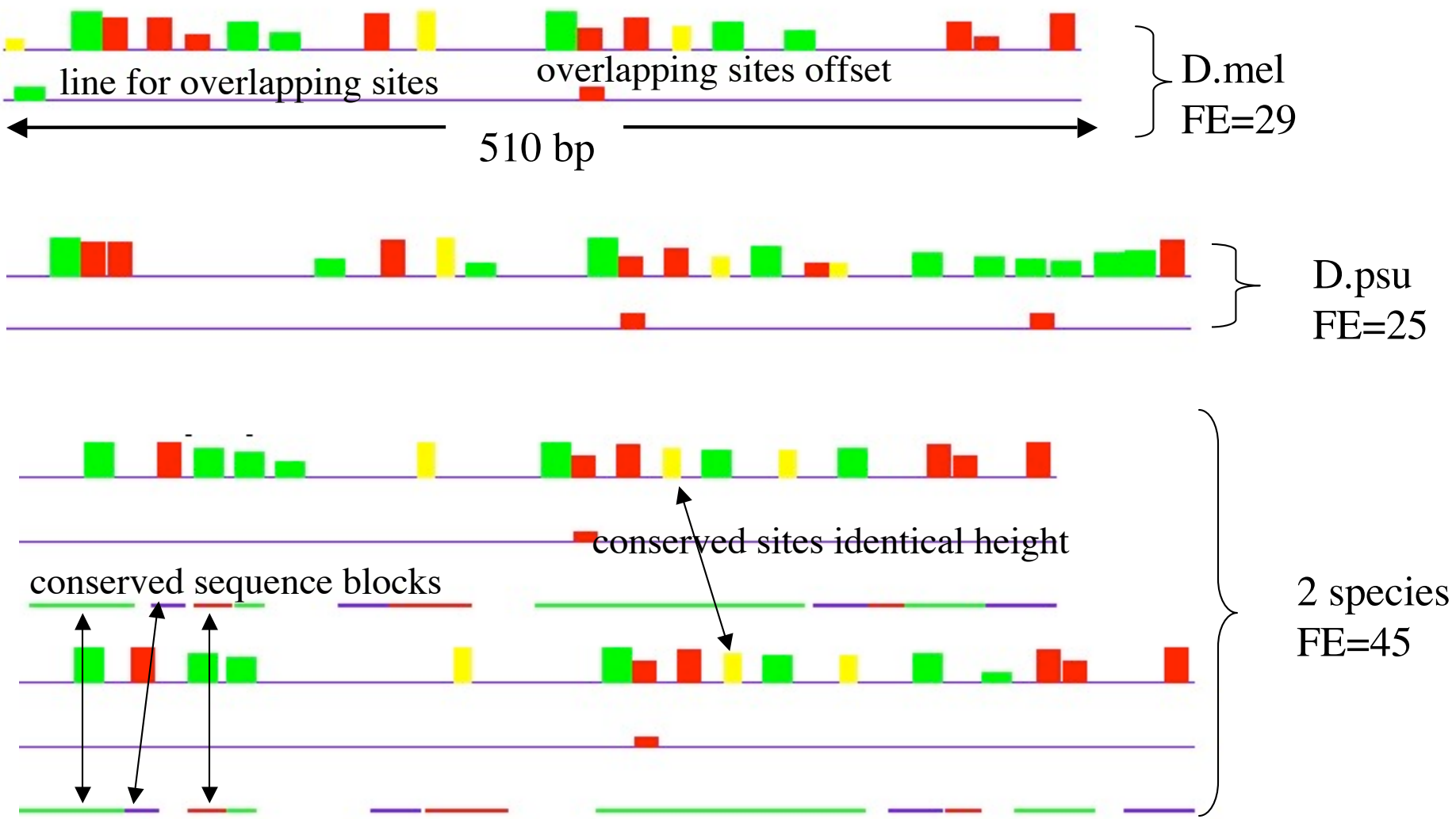
Inputs to eve stripe 1: Synergistic

(binding sites Kr, bcd..bars, quality -> height)

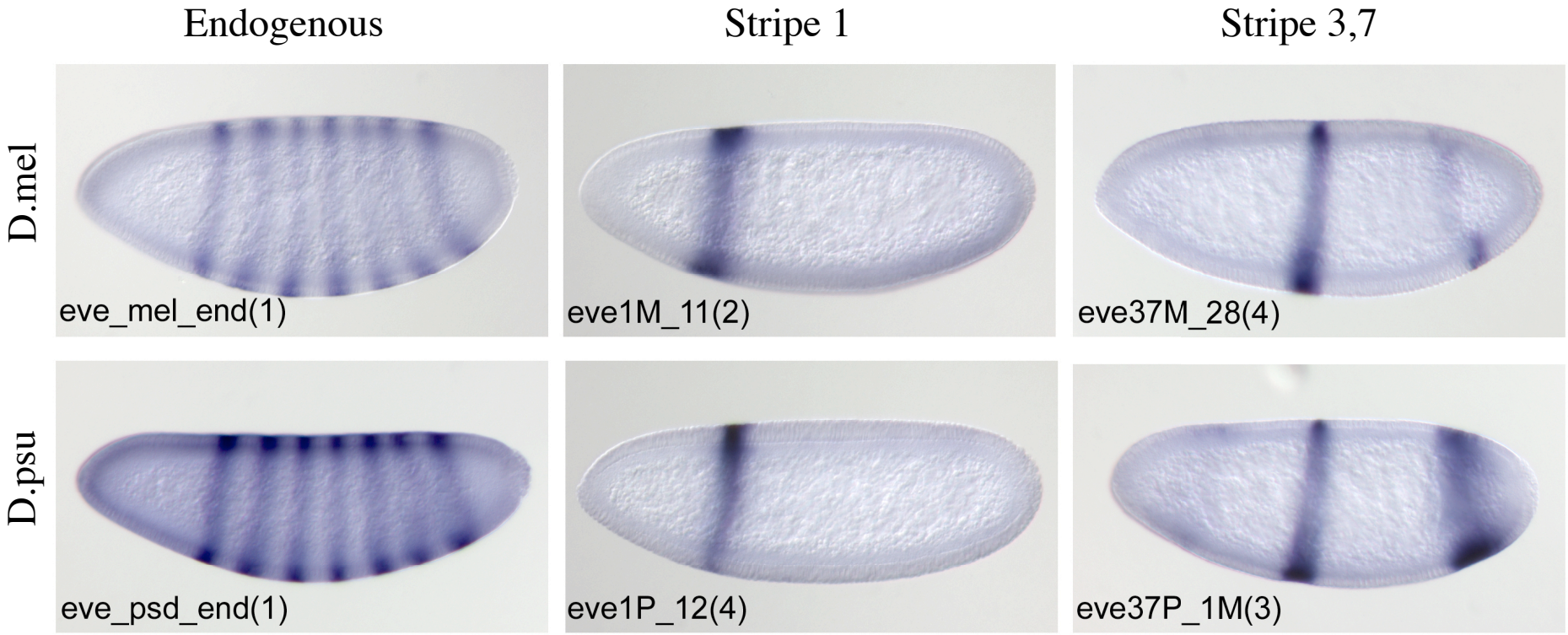


Homologous syntenous ungapped blocks co-colored

Inputs: hb, kni, stat to eve37 module:
(less synergy, more conserved seq.)

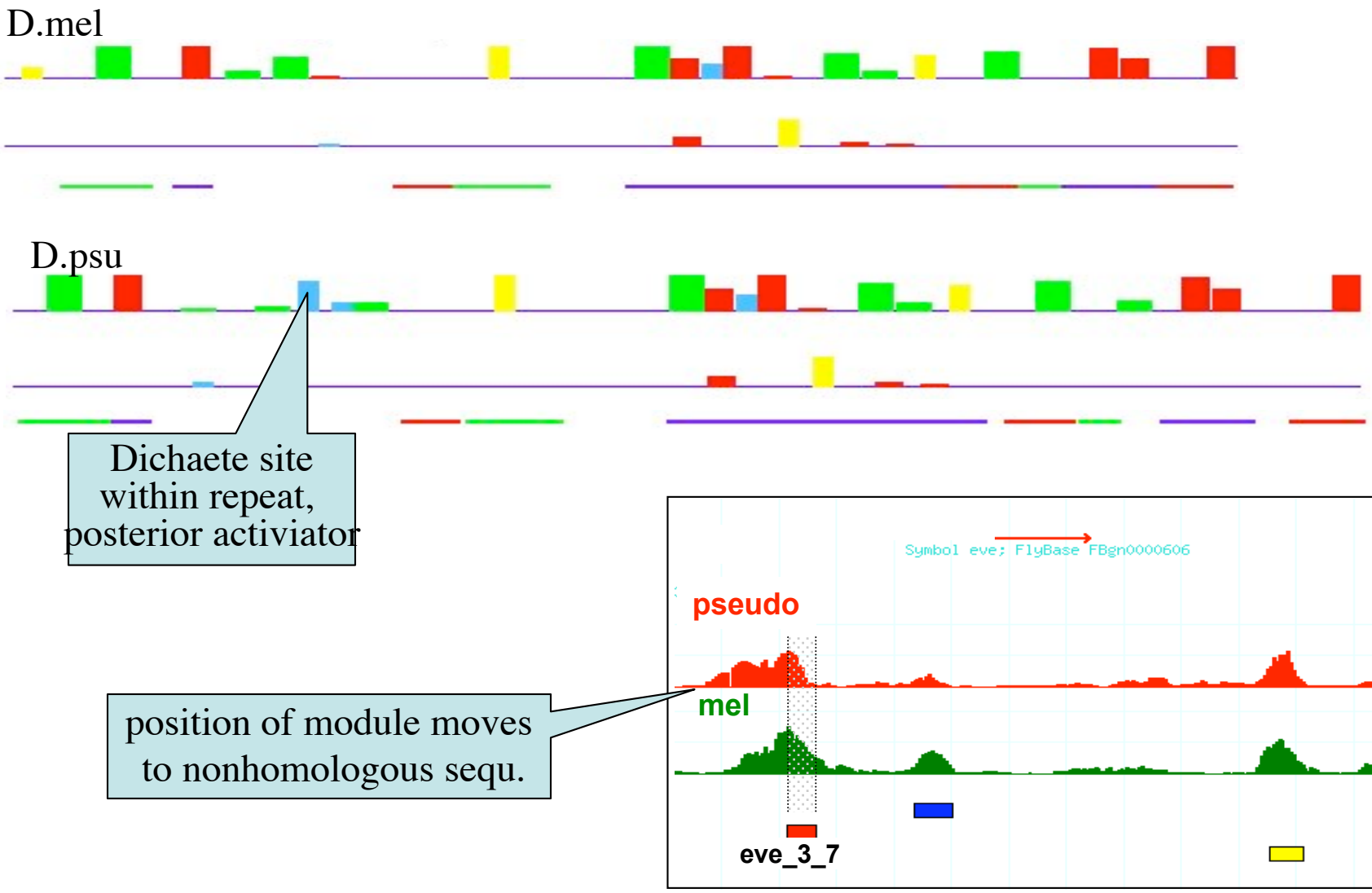


Evolution of eve modules
(all constructs in D.mel)



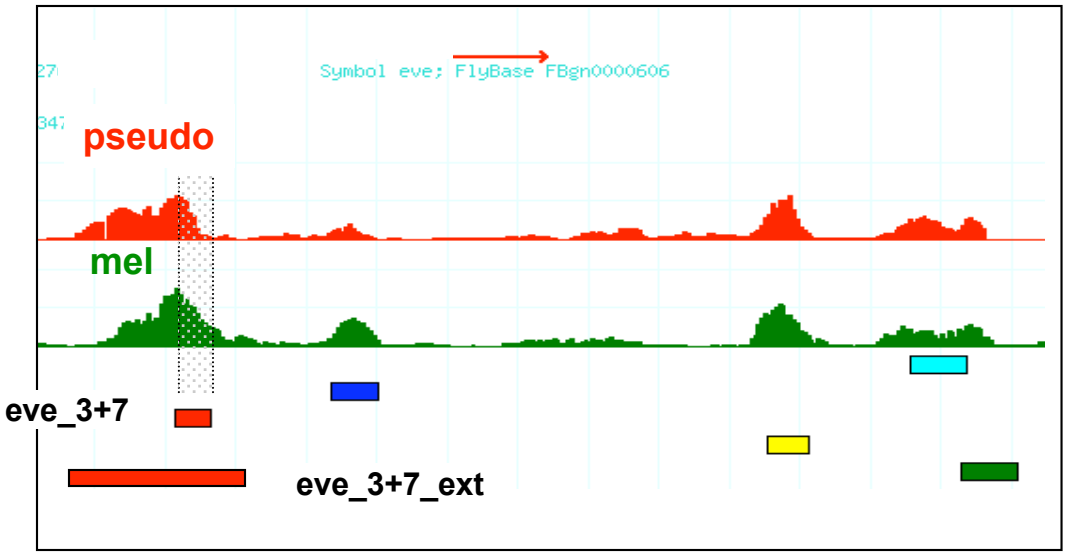
Stripe 1 identical; stripe 7 weaker than 3 in D.mel (known) but stronger, broadened in D.psu. Homologous pieces of two genomes different expression patterns. Eve gene same, hence functional module in D.psu has moved. (inputs same both species)

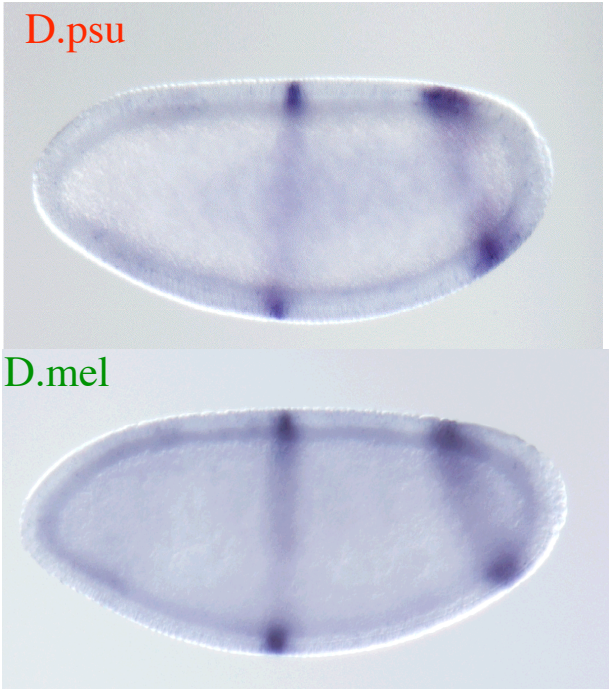
eve37: additional factors or ‘treadmilling’ of module



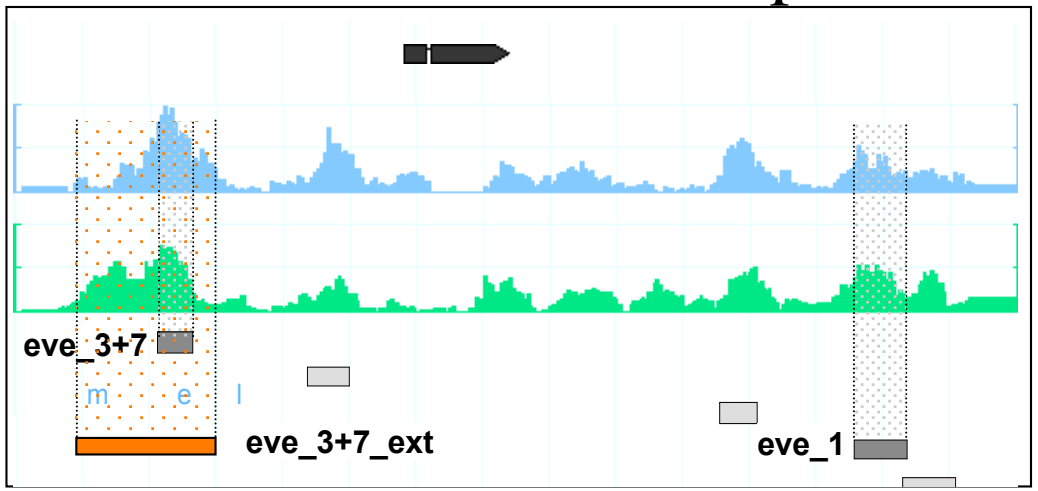
Re-delineation of eve3+7

Redefine the module in both species so that it included entire peak in D.psu.

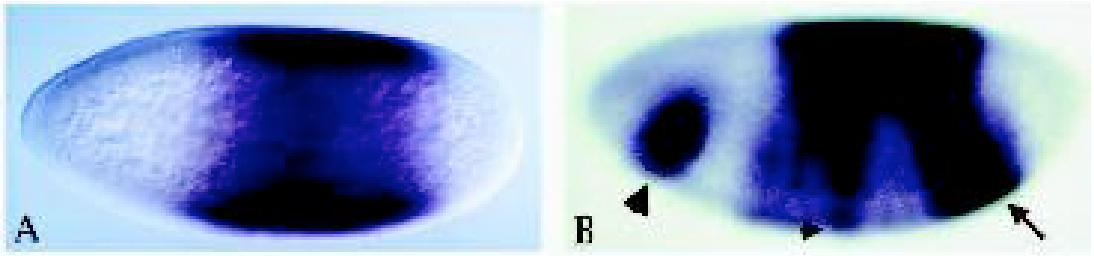




Misc D. pictures

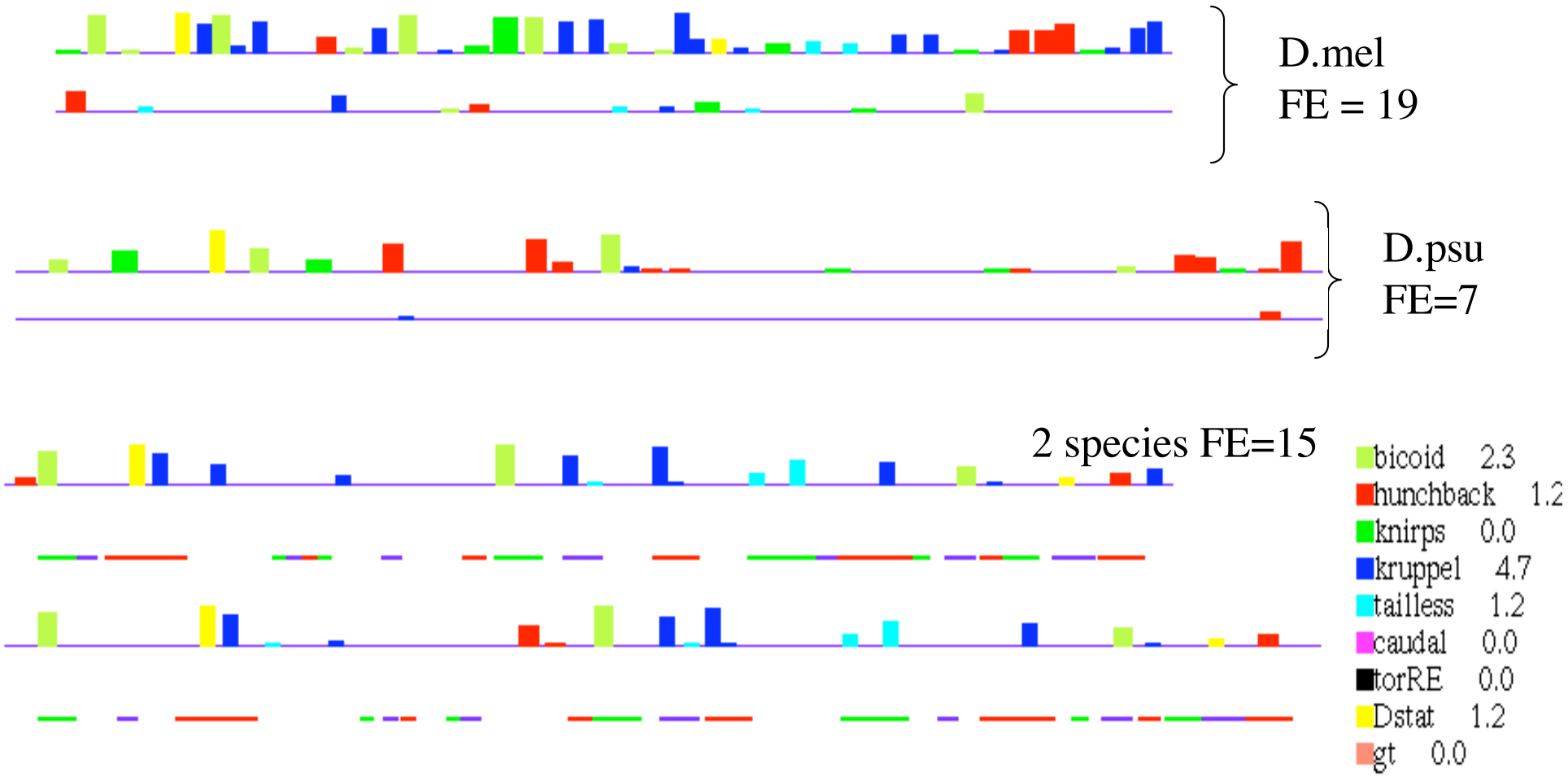


cycle 13 cycle 14



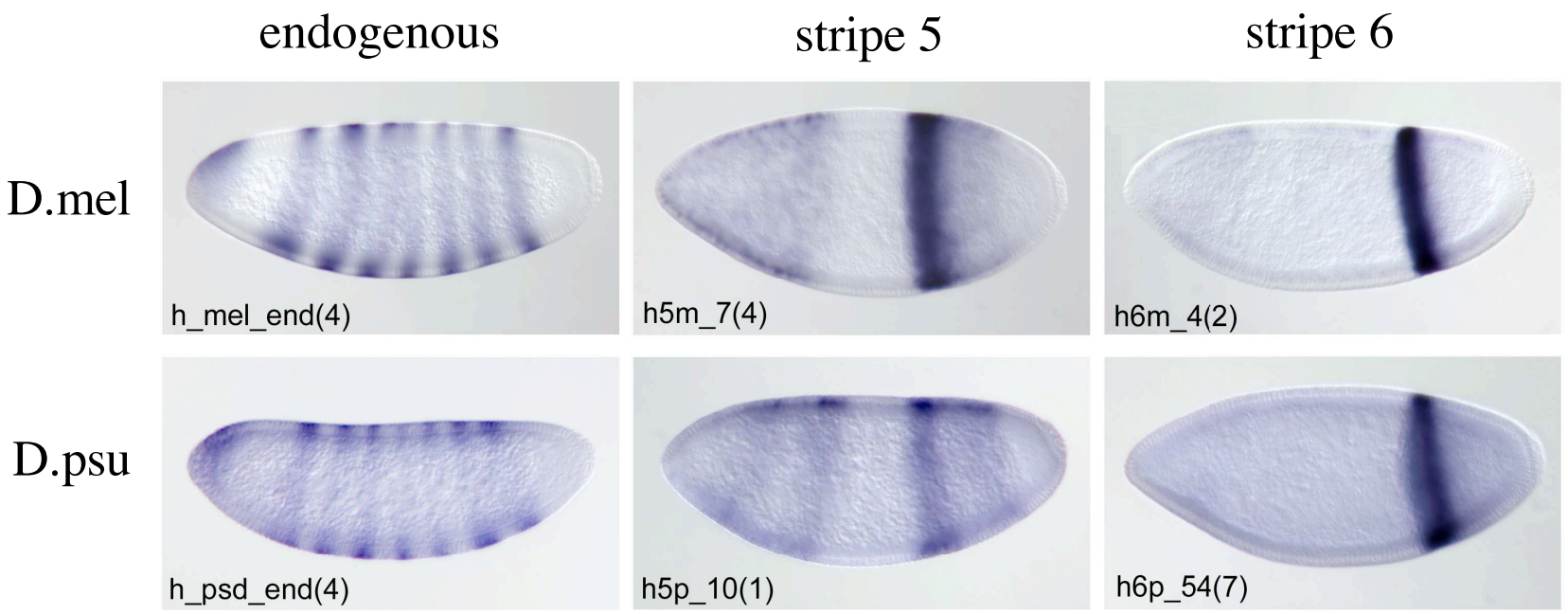
Nambu & Nambu *Devel*, **122** 3467 1996

Inputs to h stripe 5 : antagonistic

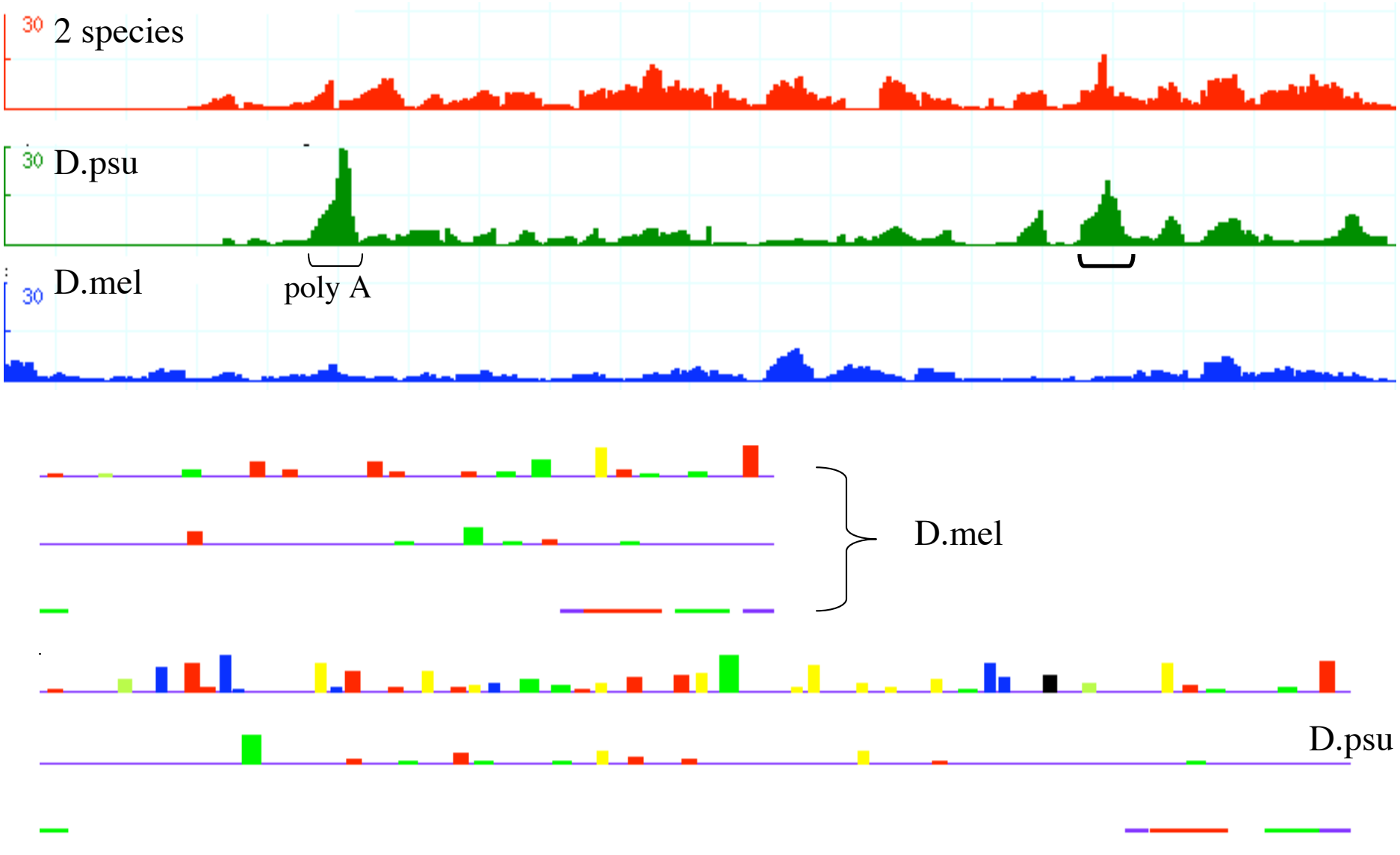


Sequences similar: genetic exper Kr, gt define anterior, posterior boundaries, kni weak activator (indirect), cad activates h5-7 Stat, bcd inputs not suspected

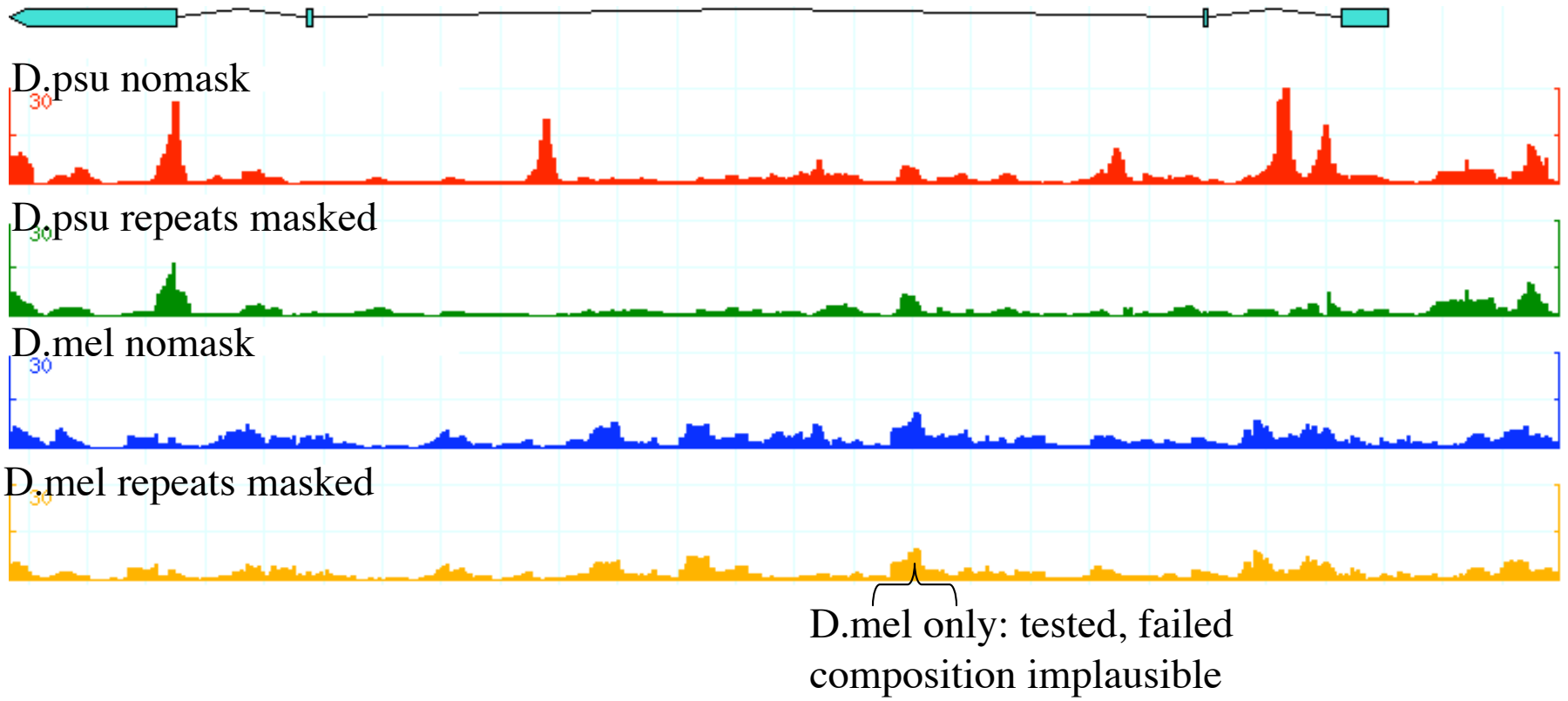
Evolution of h modules



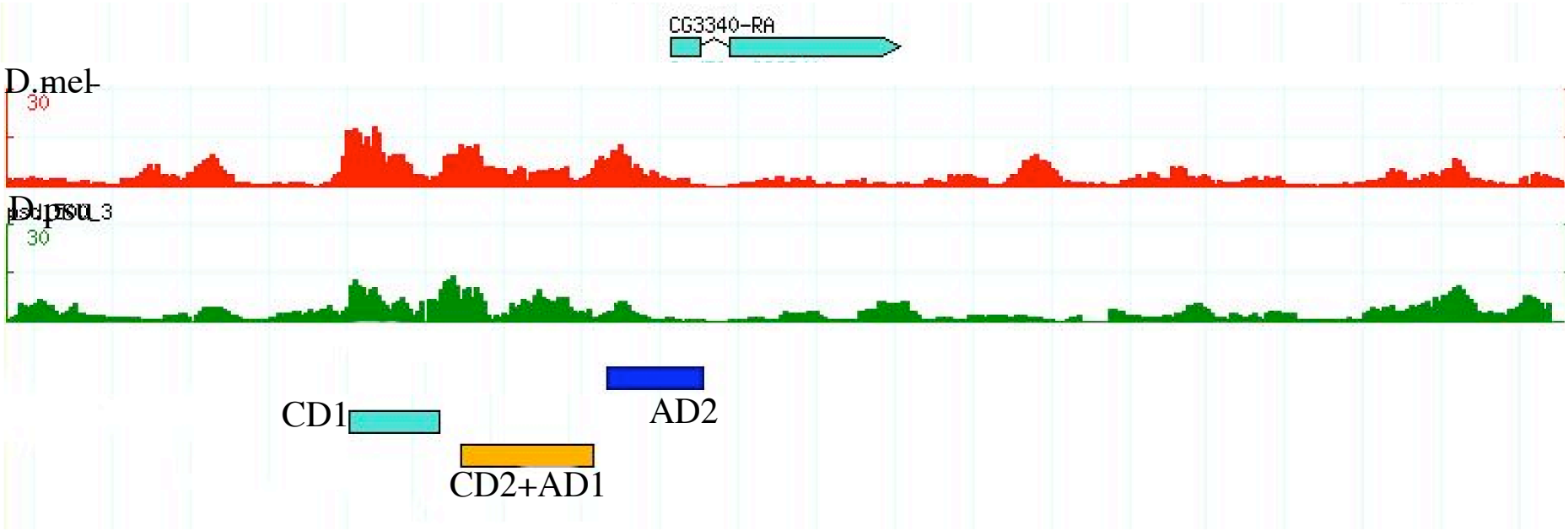
Insert to D.psu upstream of Antp



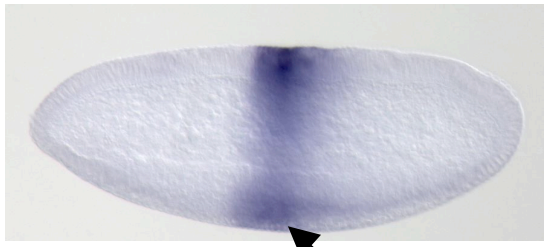
Modules composed of tandem repeats (knrl)



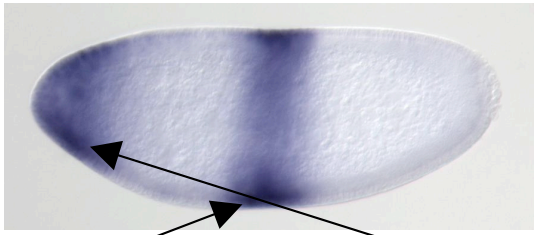
Kr: duplicate modules (D.mel)
(for CD, central domain; AD, anterior domain)



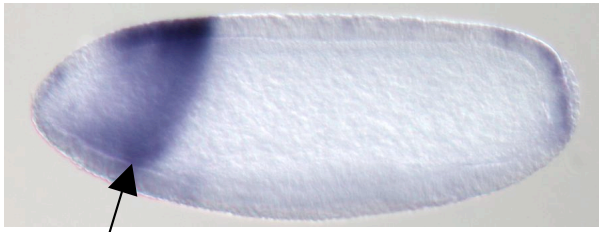
D.mel Kr_CD1



D.mel Kr_CD2_AD1



D.mel Kr_AD2



Both modules cause central domain

Both modules cause anterior domain

Changed contents of Kr CD1

D.mel -> D.psu looses most hb (9->2) sites, and all gt (2->0) sites, which repress in anterior. Activator, bcd unchanged (11->9 sites).

D.mel, FE=21

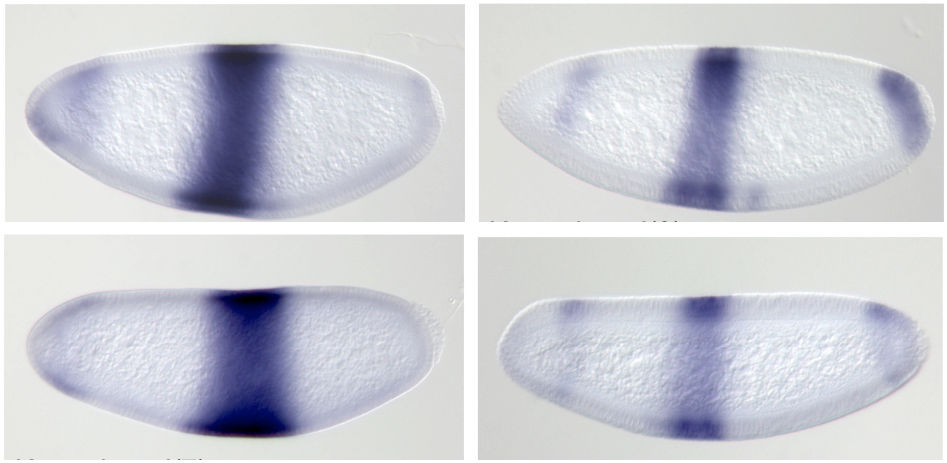
D.psu, FE=18

Evolution of Kr

Kr endogenous (time -->)

D.mel

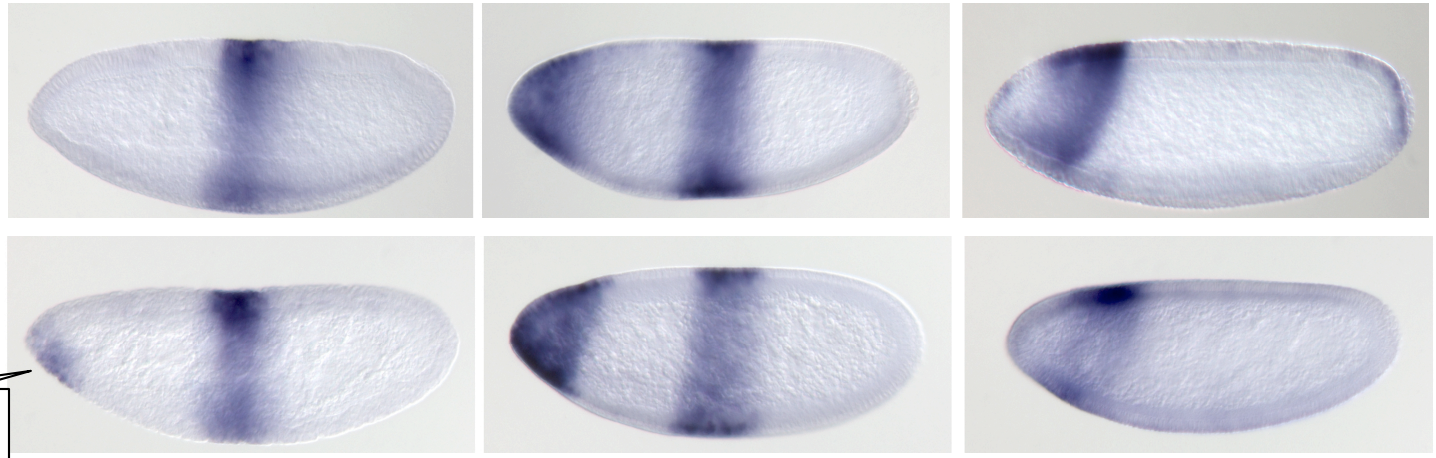
D.psu



CD1 CD2+AD1 AD2

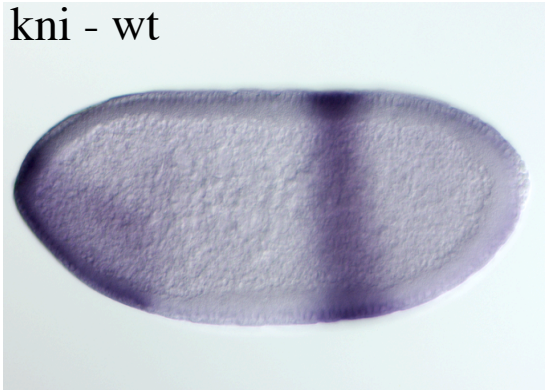
D.mel

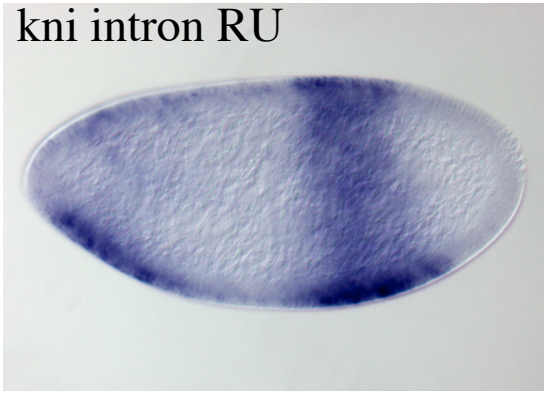
D.psu

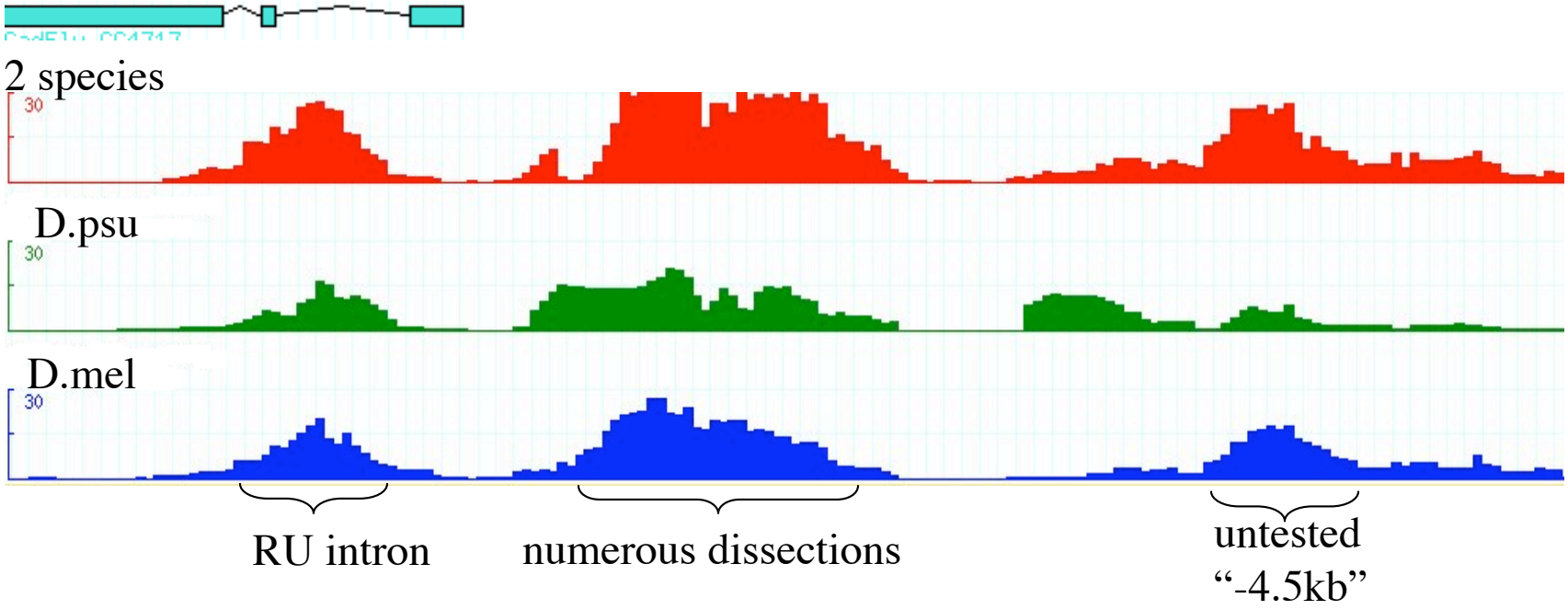


anterior
D.psu only

Duplicate modules for kni

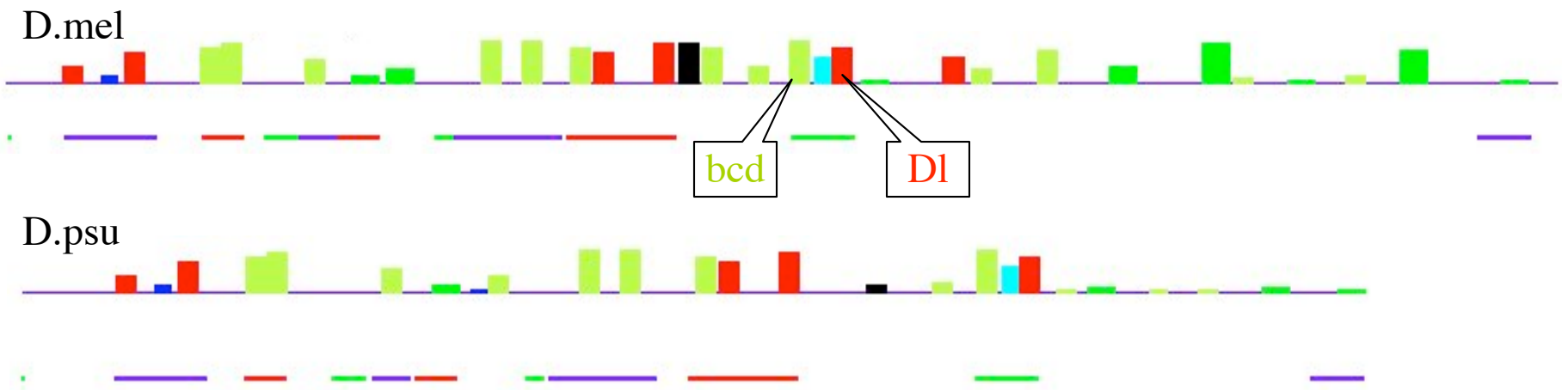






Contents of kni -4.5kb module

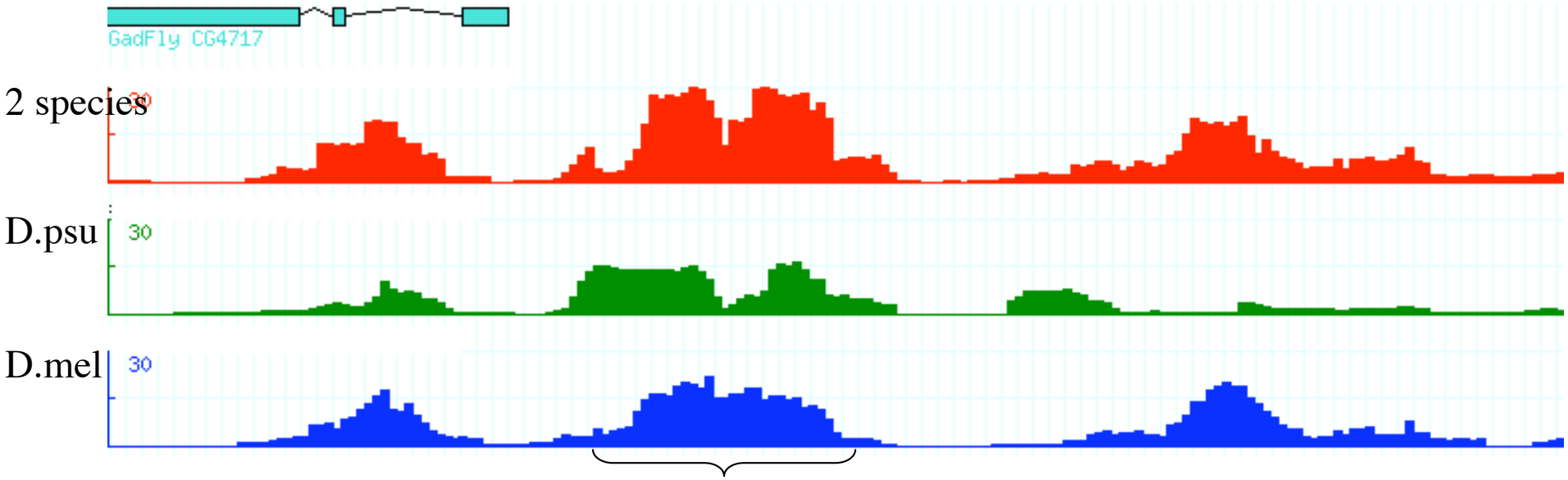
(Synergistic: anterior, ventral expression, Dl, bcd)



bicoid	10.2
hunchback	0.6
knirps	3.7
kruppel	0.4
tailless	0.6
caudal	0.0
torRE	1.1
Dstat	0.0
gt	0.2
pdm1_2	0.0
Dl_10	0.3
Dl_11	3.8

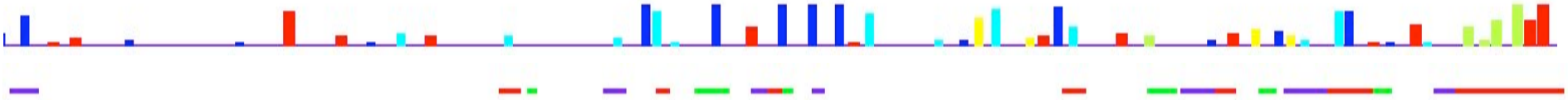
D.mel module expresses ventral anterior
(RU experiments)

Delineation of kni_composite upstream module



Contents of kni_composite

D.mel FE=27

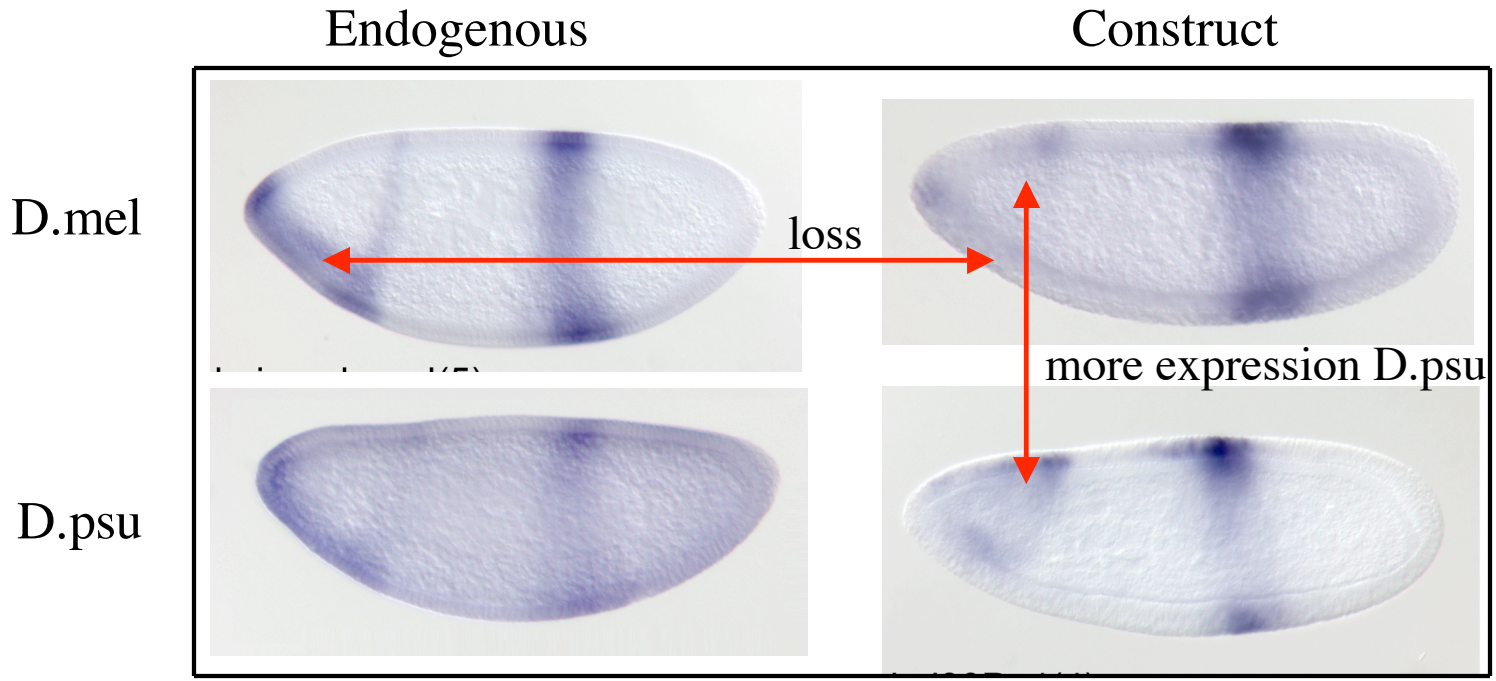


D.psu FE=19

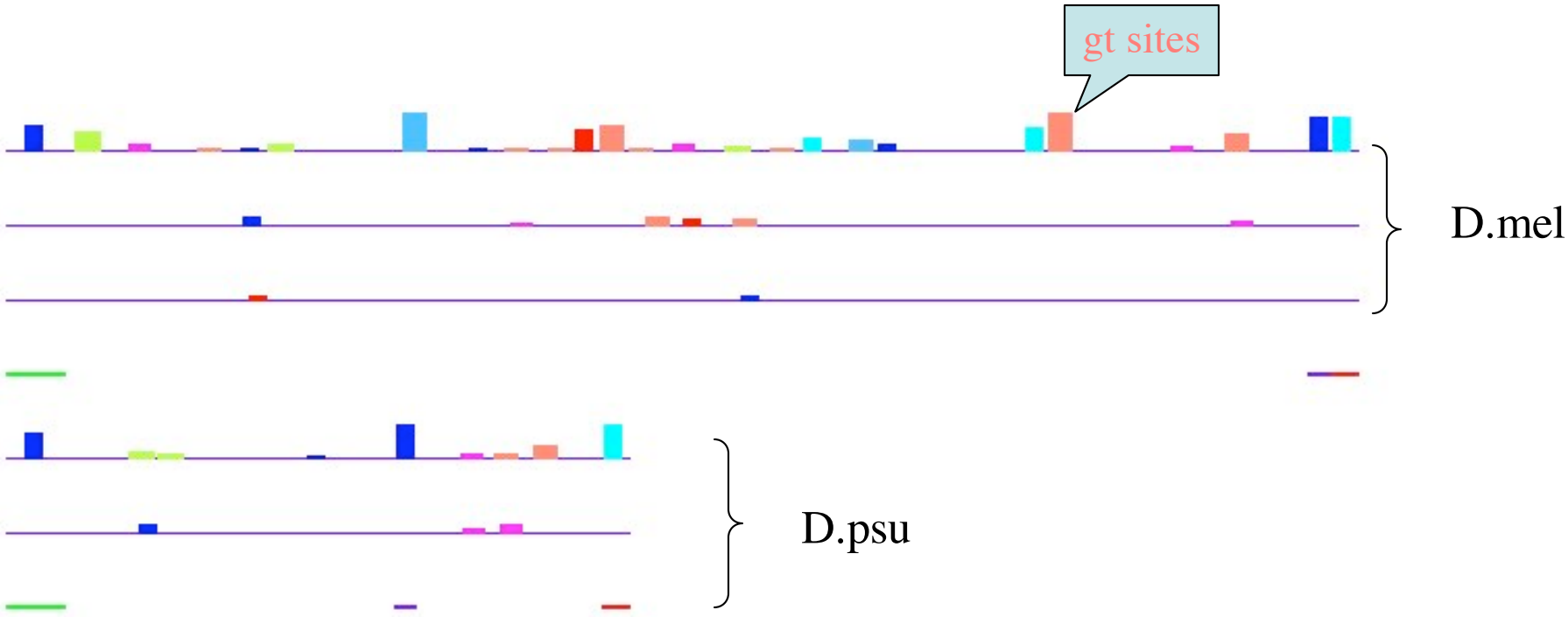
Evolution of kni_composite

D.mel: 2.2kb construct with predicted bcd, hb, tll, Kr, gt sites:
(lit: hb, tll, gt known repressors, bcd,cad activate)

D.psu: half of hb, and all gt sites (via ~100bp indel)

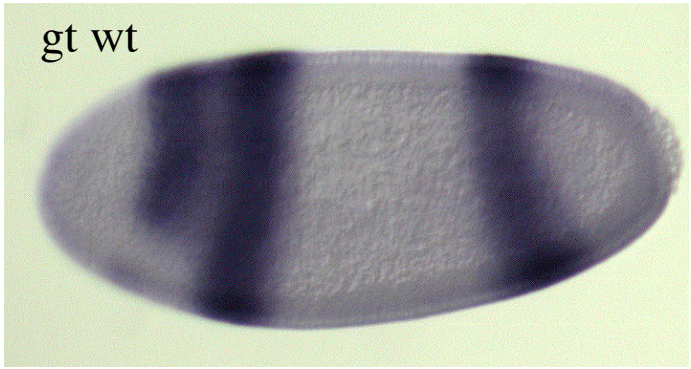


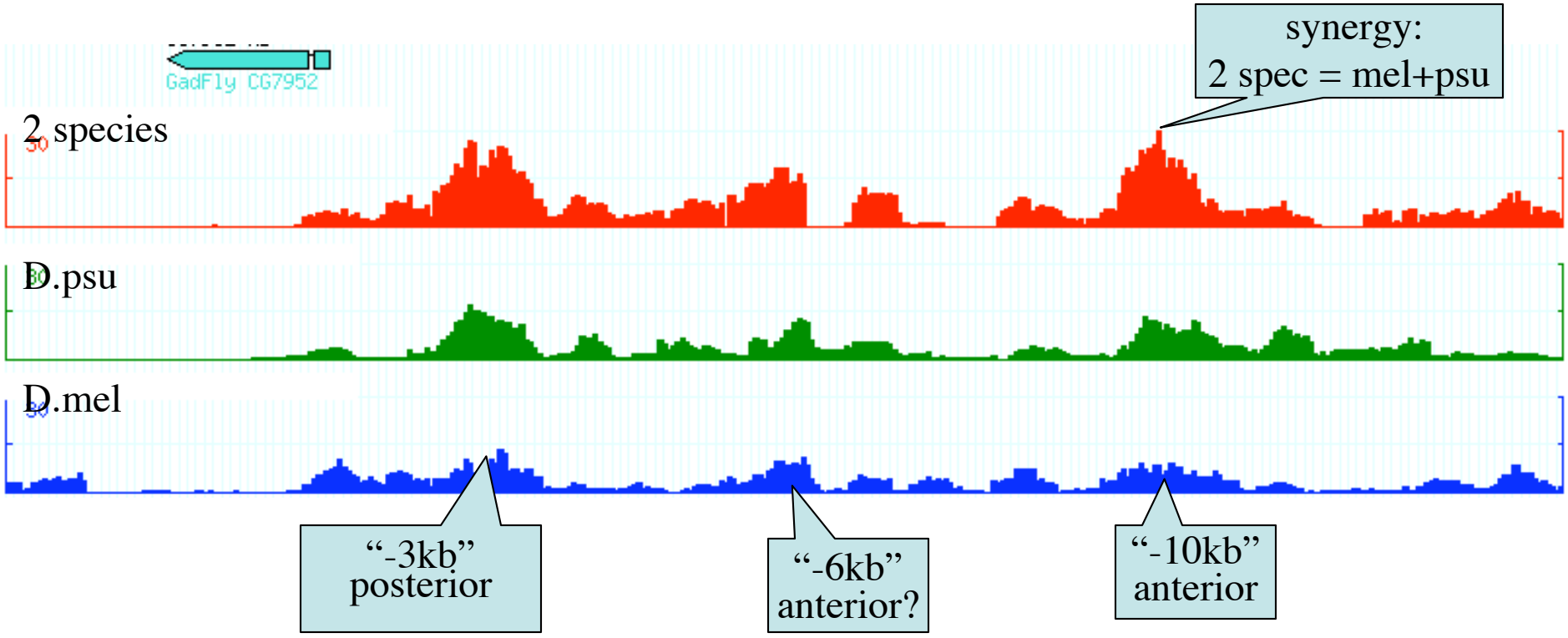
Insert into kni composite module (2 species plot)



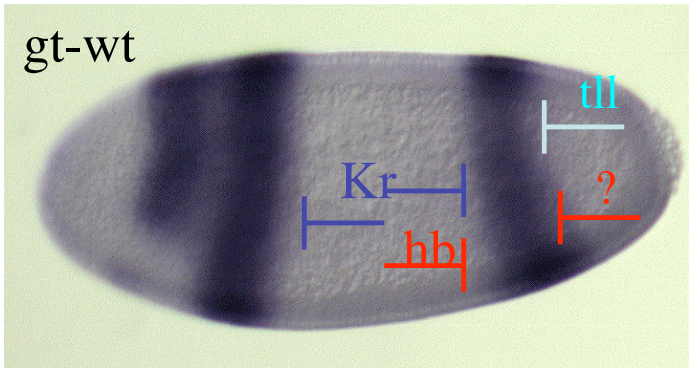
D.mel module previously tested, gt known genetically to repress kni, but homologous sites lost in D.psu. (additional synteny up/down stream.)

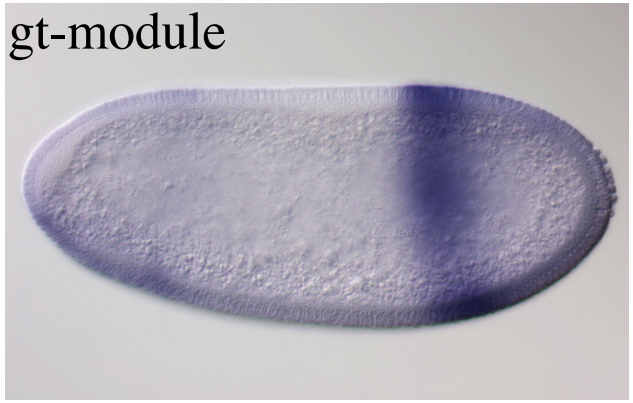
gt modules: synergy&duplication



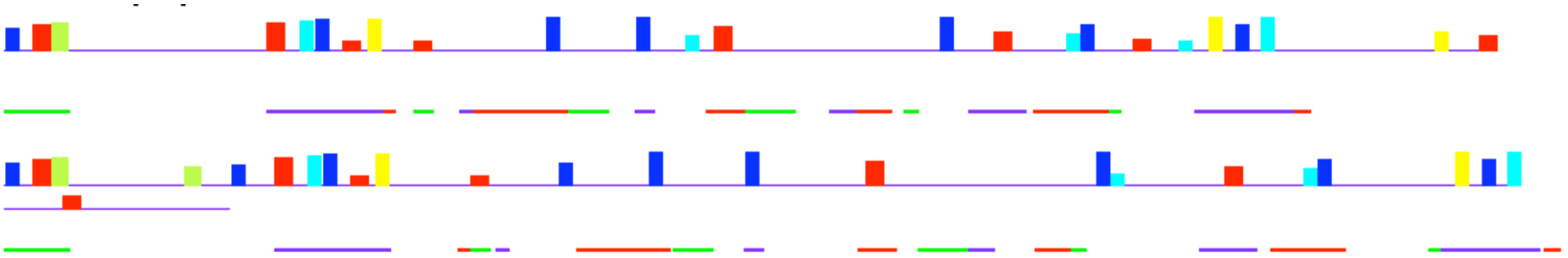


Inputs to gt posterior module



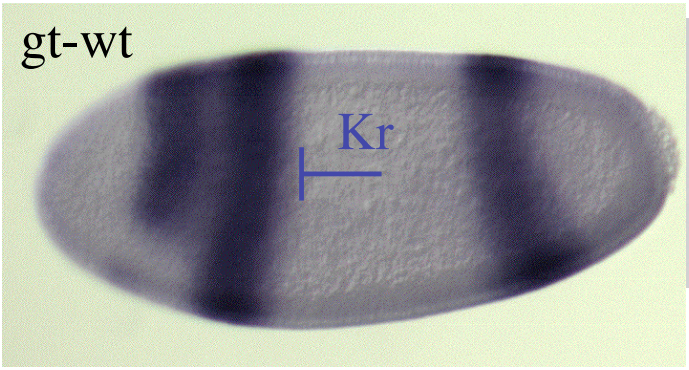


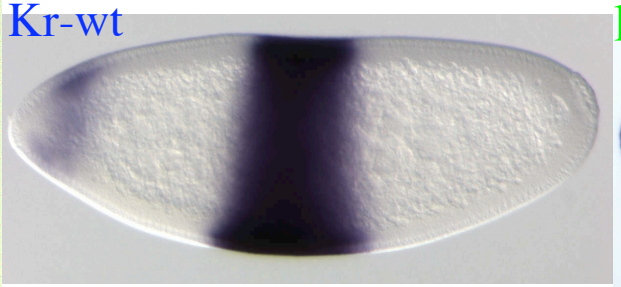
Genetic experi Kr, hb control anterior boundary, tll controls posterior boundary,
Unclear direct vs indirect and role of stat (ubiquitous activator) not known.

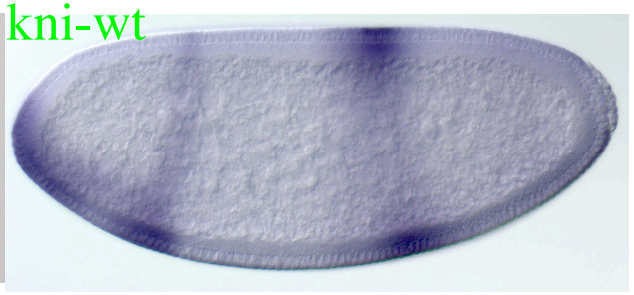


Consensus inputs to gt anterior modules

Kr, kni repress, bcd hb activate





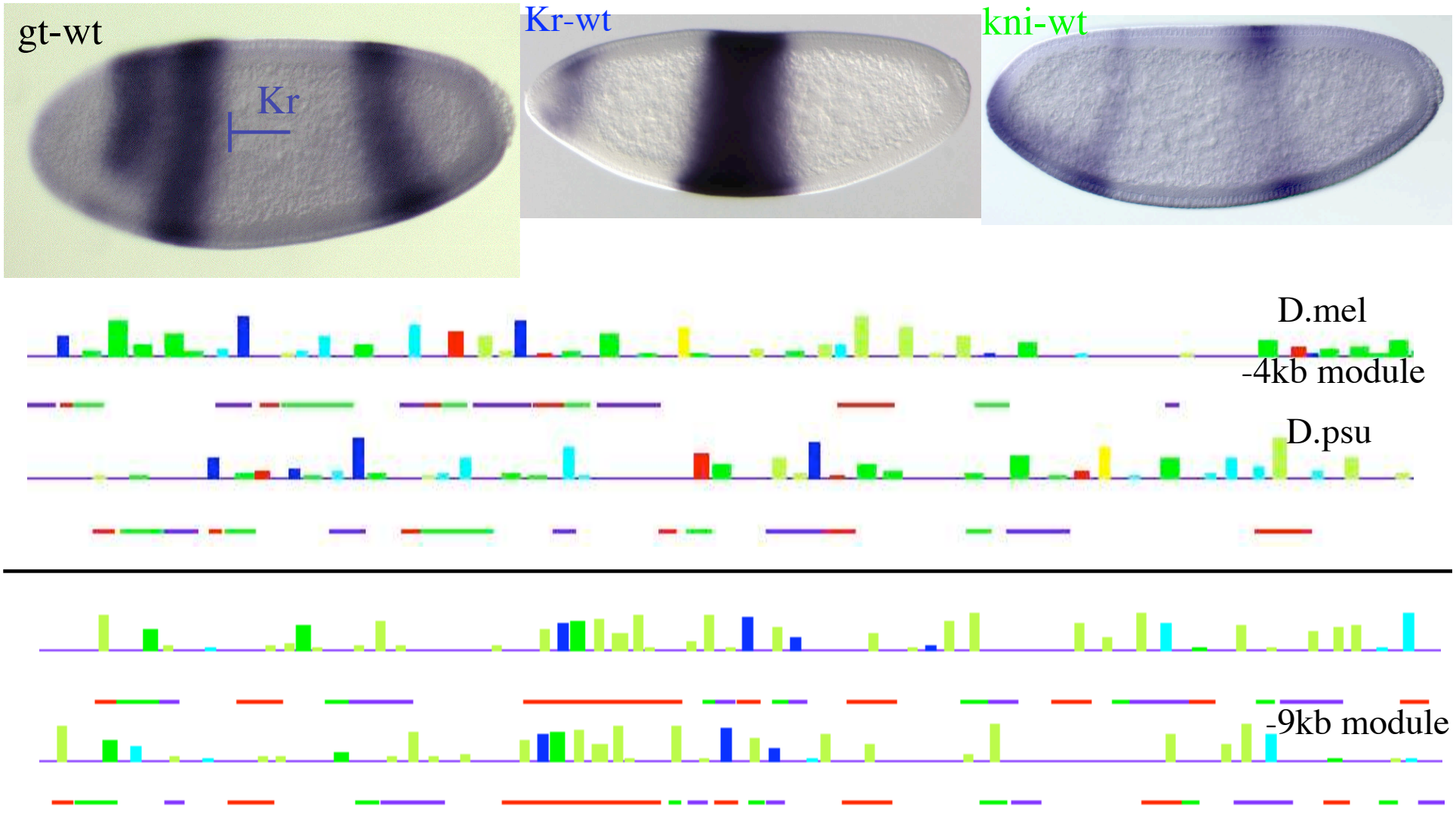


Patterns of Kr, kni repressors, complimentary to gt-anterior.

Both anterior modules have similar Kr, kni; more bcd in -9kb module

Inputs to gt anterior modules (detail)

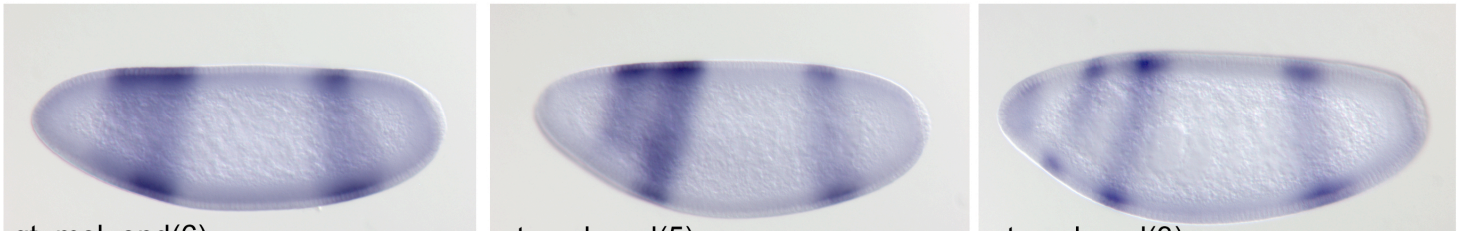
Kr, kni repress, bcd hb activate



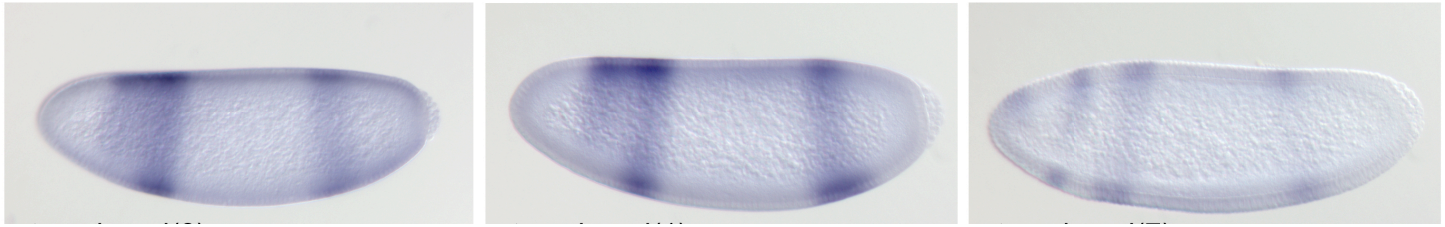
Evolution of gt anterior expression

Endogenous (time -->)

D.mel



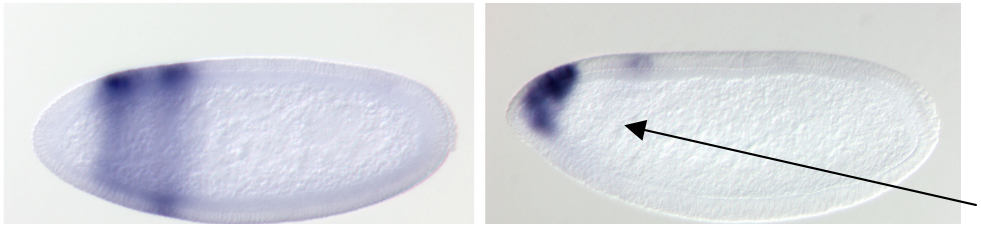
D.psu



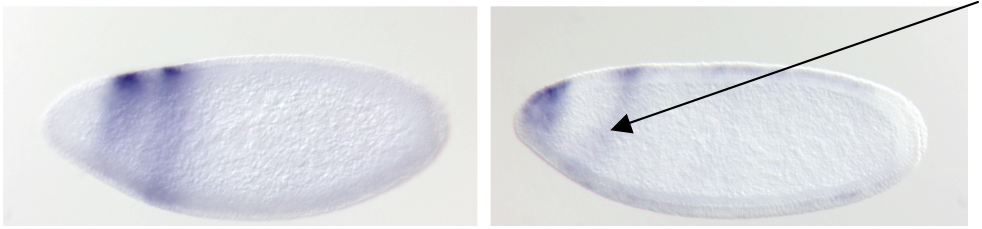
-10kb module

-6kb module

D.mel

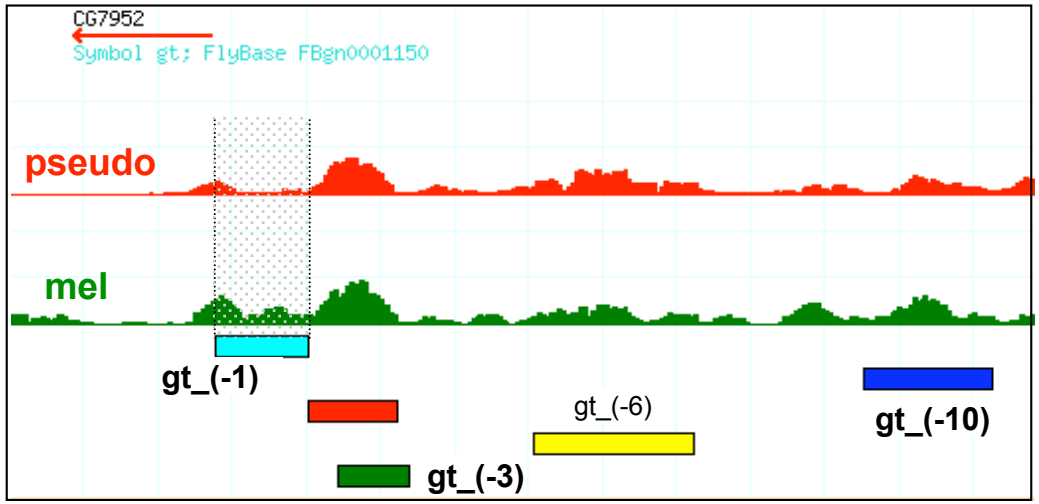


D.psu

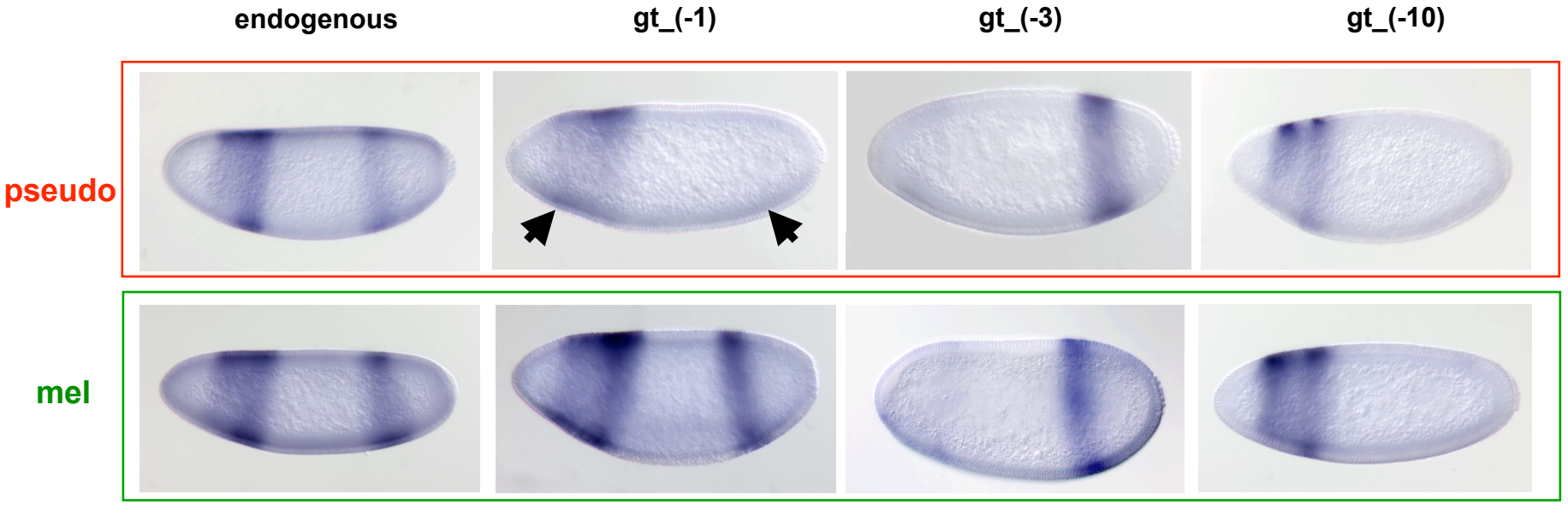


More anterior stripe in D.psu

gt:module with duplicate functions lost



gt -1kb upstream module both number and strength of sites lost in D.pseudo.



Conclusions (evolution)

Computational screen of all known and predicted modules for ~30 AP patterned genes -> experiments 22 constructs (all in D.mel) + 11 pairs of insitu's for homologous genes.

Evolution of regulation: analogy with gene evolution

- *duplicate modules* drive very similar patterns, --> relative contribution of homologous modules may change? (ex kni, gt, Kr, tll). Sequence duplication or convergent evolution??
- pseudomodules, on the way in/out, weak sites, lack activator, self repress.
- large indels (esp tandem repeats) can carry plausible binding sites but influence not modular. Events not recent, compensatory changes elsewhere??
- more functional conservation on larger scales eg binding sites change a lot, modules less, and genes less (eg hb in D.vir, Lukowitz 1994).
- ~700 blastoderm patterned genes ,BDGP, essential or parasitic

Ability to calculate expression change of homologous modules limited, but evolution provides useful data for module logic

Conclusions (module finding)

How is regulatory signal encoded in the genome?

Counting binding sites from set of related factors works for blastoderm and signaling modules. Two species fit in parallel marginal improvement AP, possibly better for wing

Spacing matters for short range repressors (eg Kr) however positional correl don't improve predictions for AP but significant for hox/exd hth, (dl+cofactors?).

Architecture of sets of modules surrounding gene defined by insulators and promoter accessibility.

Accuracy of calculations:

For known blastoderm patterned genes, if computed inputs sensible > 75%

For genome scans, no other information predicted 150 genes, ~50% blastoderm patterned, 'failures' of calculations informative.

AP axis atypical: Intercell Signaling (eye, wing, leg, glial,.

Conclusions

How is regulatory signal encoded in the genome?

Using a collection of factors that define a pathway, counting number and quality of sites picks out regulatory modules.

Other constraints (eg positional), what's the phase space, 'grammar rules'

Alternative approach, make random 'sentences' and ask cell if meaningful.

What is the logical function (AND, OR…) computed by the module

Evolution of regulation: analogy with gene evolution

more functional conservation on larger scales eg binding sites change a lot more than gene expression patterns.

pseudomodules, on the way in/out, weak sites, lack activator, self repress.

duplicate modules drive very similar patterns, does relative contribution vary between species? (ex Kr, tll, gt, kni)

treadmilling: drift of functional module to contiguous regions of genome.

Mosquitos?

Conclusions

How is regulatory signal encoded in the genome?

Using a collection of factors that define a pathway, counting number and quality of sites picks out regulatory modules (13/16 predictions work exp)

Does this work for other developmental and cell signaling systems?

Grammar rules for functional modules??

Evolution of regulation: analogy with gene evolution

More functional conservation on larger scales eg binding sites change a lot modules less, genes less still.

duplicate modules drive very similar patterns, but relative contribution can vary between species? (ex Kr, tll, gt, kni).

Treadmilling: modules change by drifting to contiguous regions of genome.

Questions:

Fitness correlates (if any) of the changes we see

Drosophilids split x 10^7 years, can one infer mosquito patterning from fly, 10^8 y, embryonic patterning similar.

Collaborators

N. Rajewsky (NYU bio), M. Vergassola (Inst Pasteur)
(single species)

E. Emberly (SFU phys), S. Sinha (U. Ill cs)
(2 species and evolution)

U. Gaul: J.Fak, M. Pearce, M. Schroeder

Refs…

References:

History of Evolution:

SJ Gould, "Ontogony & Phylogeny"

" "Wonderful Life"

W. Provine, "Sewall Wright & Evol. Bio."

" , "Origins of Theor. Pop. Genetics"

Evo-Devo

A.S. Wilkins, "Evol. of Developmental Pathways

E.H. Davidson, "Regulatory Genome"

" "Genomic Regulatory Systems"

S.B .Carroll, "DNA and Diversity"

Fly Pop. Biology

J.R. Powell, "Evol. Bio. Drosophila Model"

Speculations

Fly vs single celled life:

Decoding the regulatory sequence may be easier in fly, eg less compressed code more redundancy easier to decipher.

Classifying vs clustering

Organisms decode regulatory sequence with proteins, computationally we can deduce binding sites given module(s). Its inherently easier to classsify sites against template, than discover sites by statistical overrepresentation. Multiple genomes improve binding site prediction since supply multiple copies of same message. How many copies of a fly module, or bacterial promoter are needed to detect its architecture??

Regulatory components reused eg gap genes in neural patterning

Kr, tll, neural regulation uses same DNA as AP modules, caveat on autonomy of 'module'

Gene variability between species/strains:

many experiments, yeast, fly. Scale Berkeley insitu project, ~700 blastoderm patterned genes, <100 in screens, imagine the remainder parasitic variation. (a number of these detected in genome wide scans)