

Non-coding DNA (in Drosophila): no junk, but what else ??

Carl Herrmann

IBDML & Univ. Méditerranée - Marseille

TAG'06 – LAPTH Annecy

a fundamental question ...

*Why am I
smarter and more evolved
than others ?*

2 ways to ask the question

intra-species

"...than my colleagues ?"



inter-species

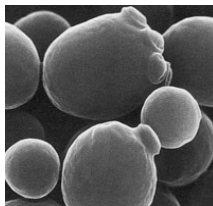
"...than my guinea-pig ?"



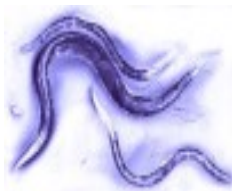
eukaryotic genome contest !!

featuring :

yeast



worm



fly



arabidopsis



pufferfish



rice



human



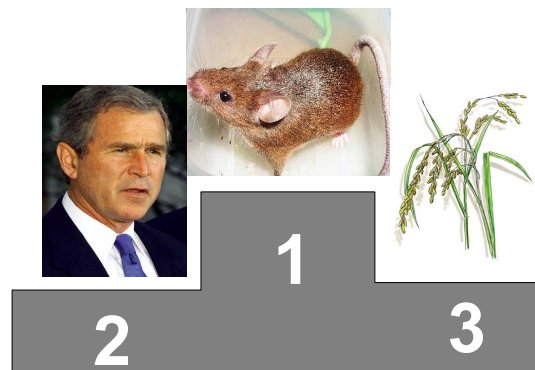
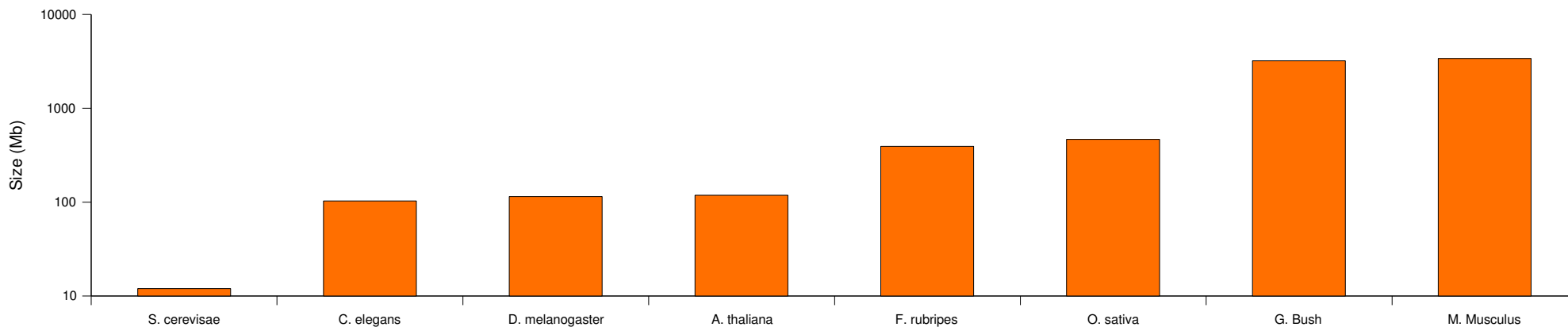
mouse



Genome size



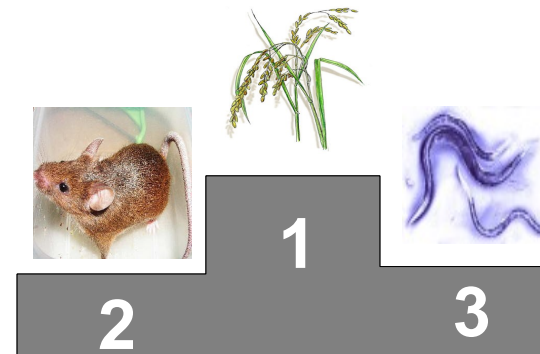
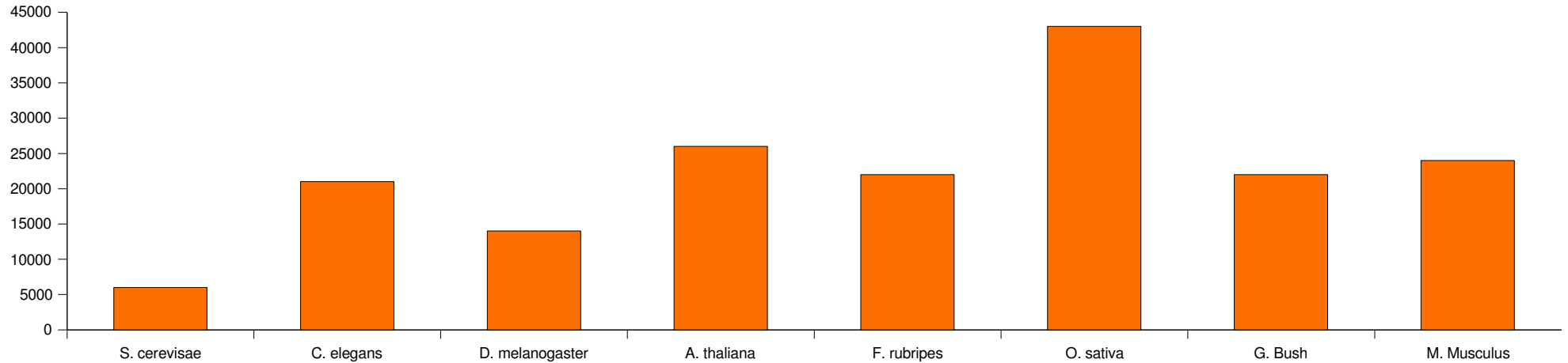
Genome size



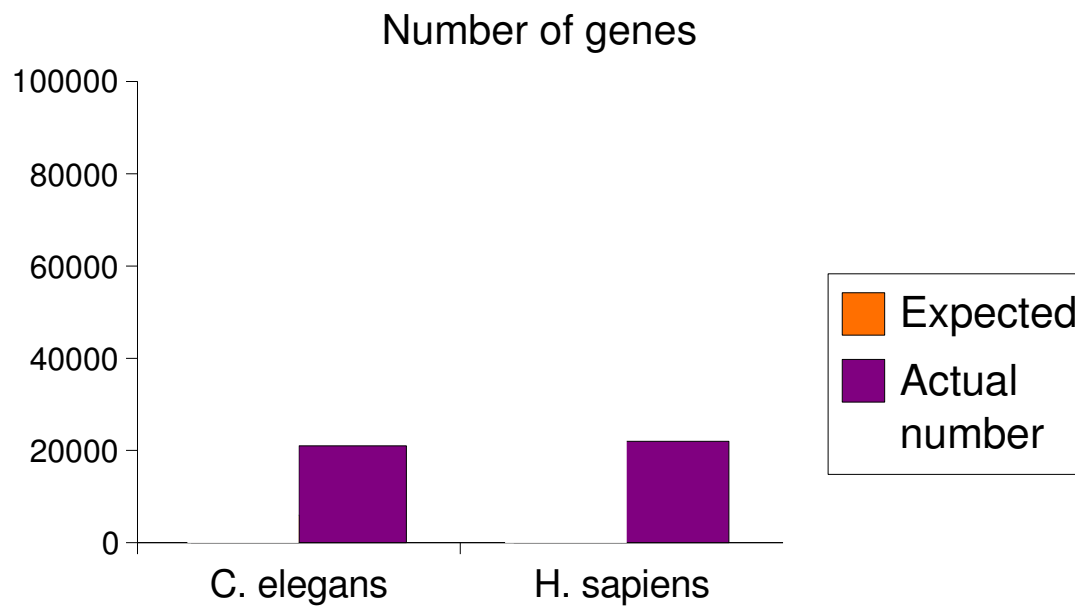
Number of genes



Number of genes



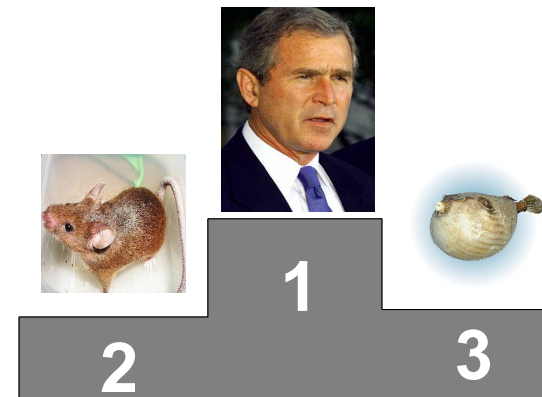
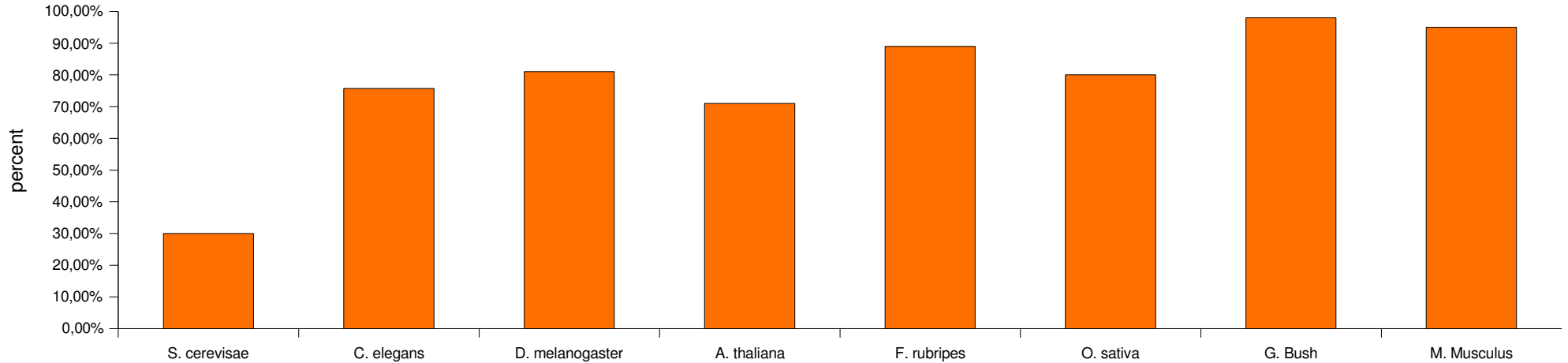
this is becoming a little
embarrassing for us ...



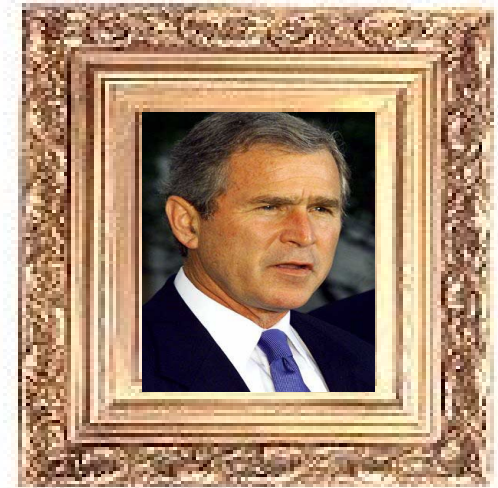
Non-coding DNA



Proportion of non-coding DNA

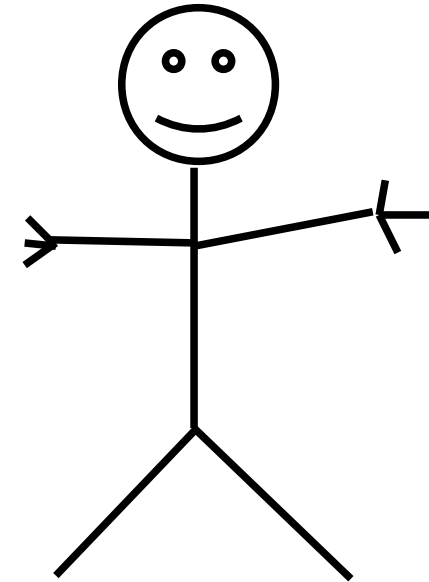
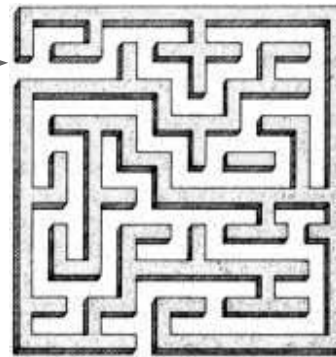
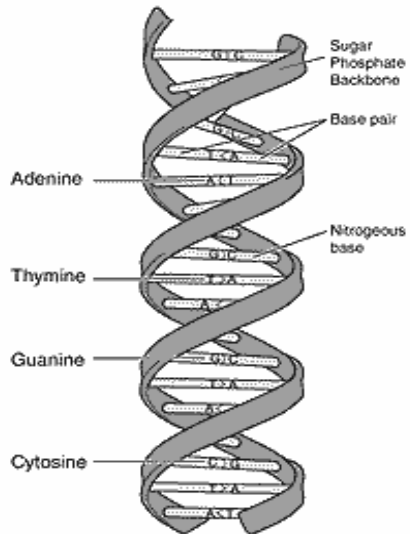


**eukaryotic genome
contest !!**



how glorious...

Genotype / Phenotype



"why would a perfect God create flawed DNA which is primarily composed of useless non-coding regions ?"



noncodingDNA.com - home - Firefox

File Edit View Go Bookmarks Tools Help

http://www.noncodingdna.com/

noncodingDNA.com exploring the noncoding portion of the genome

Genomics

home noncoding

Updates

ncDNA/tgDNA
Figure 1, in *The hidden* presents a 'cartoon' presented in the Table. Click here for more information.

Human Gene Number Decreased
The number of human genes was once predicted to be over 100,000, then decreased to 35,000 and is now reduced to 20,000 - 25,000. Read the latest **ncDNA news** here.

More nc/tgDNA values added
Six more values added. Plus, a revision to the general ncDNA/tgDNA trend.

Nature Reviews Genetics
RNA regulation: a new genetics? (April 2004)

Visit NRG's website here
Visit PubMed Here

New to genetics and noncoding DNA? Click here for a jargon-free explanation.

Done

A recent paper - the reason for the design of this site - suggests that the amount of noncoding DNA (or junk DNA) per genome is a more accurate indicator of biological complexity than either gene number or genome size. It is therefore highly likely that these sequences are functional, and the idea of "junk" DNA may need to be trashed.

billions and billions of stars

*anybody out
there ???*

billions and billions of nucleotides

*anything in
there ???*

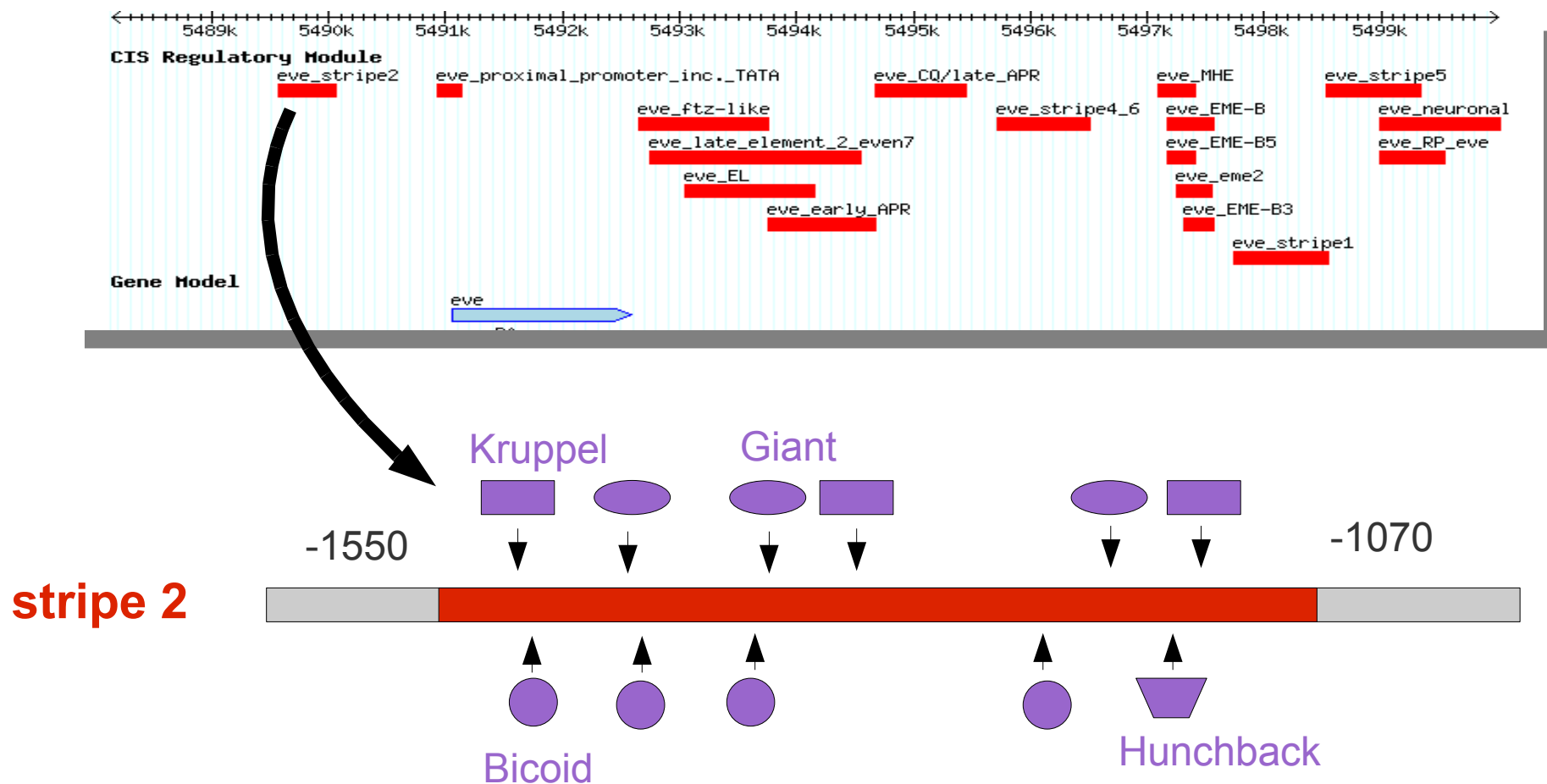
"junk" DNA ??

- ~ 40-50% of the human genome consists of **repetitive elements**
 - single repeats
 - transposons (SINEs, LINEs)
 - (micro)satellites (eg. $(CA)_n$)
- large presence of non-functional **pseudo-genes**

functional role ???

... but many functional elements

- enhancers & transcription factor binding sites



... but many functional elements

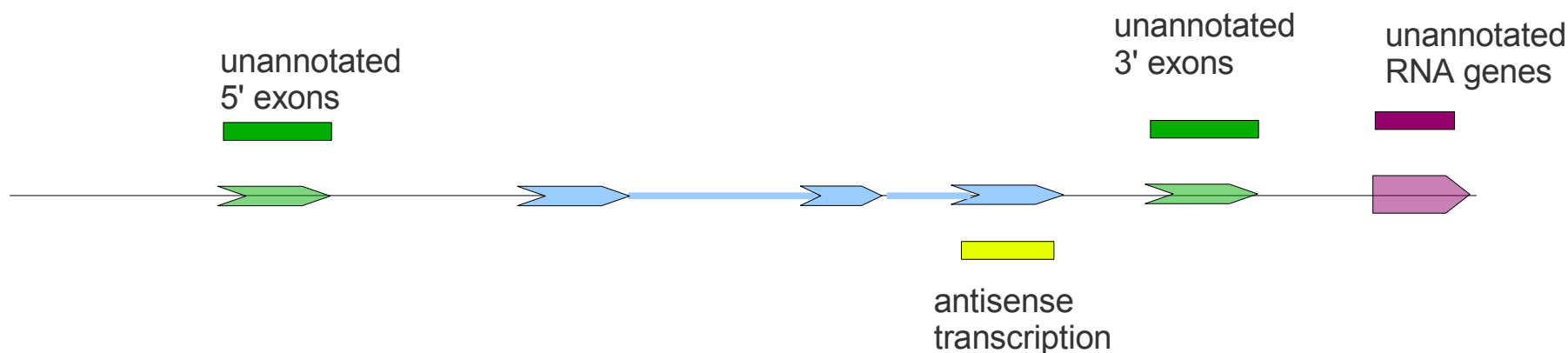
- chromatin insulators
- PcG & trxG protein binding sites
- small RNA genes
 - rRNA
 - tRNA
 - snRNA, snoRNA
 - miRNA



... and a huge level of transcription !!

- Fly: 85% of genome transcribed [Manak et al, 2006]
- Mouse: 62% of the genome transcribed [FANTOM consortium, 2005]
- Yeast: >85% of genome transcribed [David et al., 2006]

TUF = transcript of unknown function



[cf. Willingham & Gingeras, Cell 125 (2006) 1215-1220]

yes !

yes !!

who's asking ?

sure !

sure !

yeap !

oui !

anything in there ???

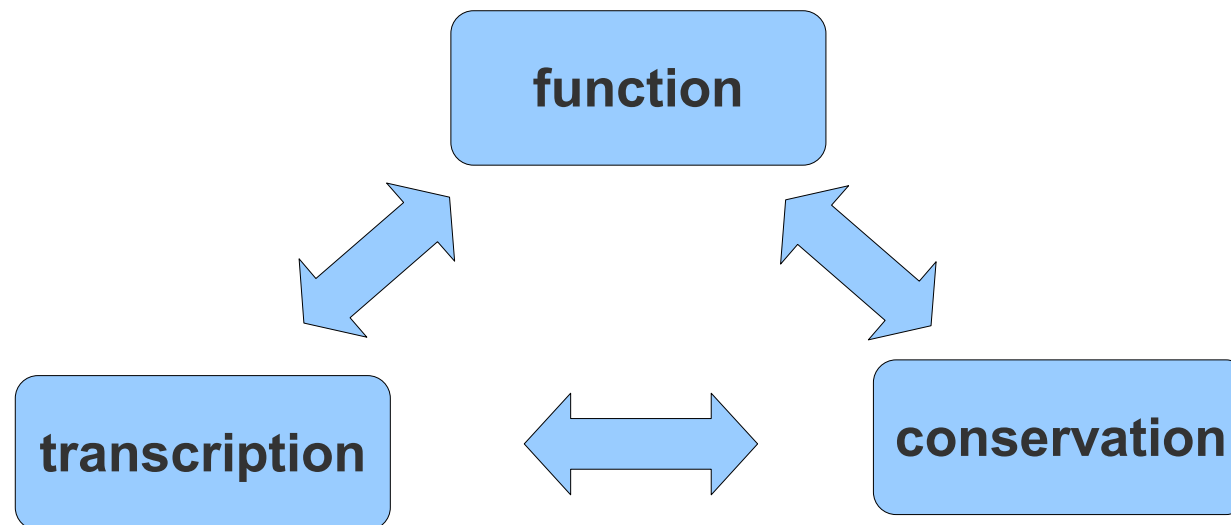
hello !

I'm here !

yes !

so there's something in here.....

... but how do we identify it ?

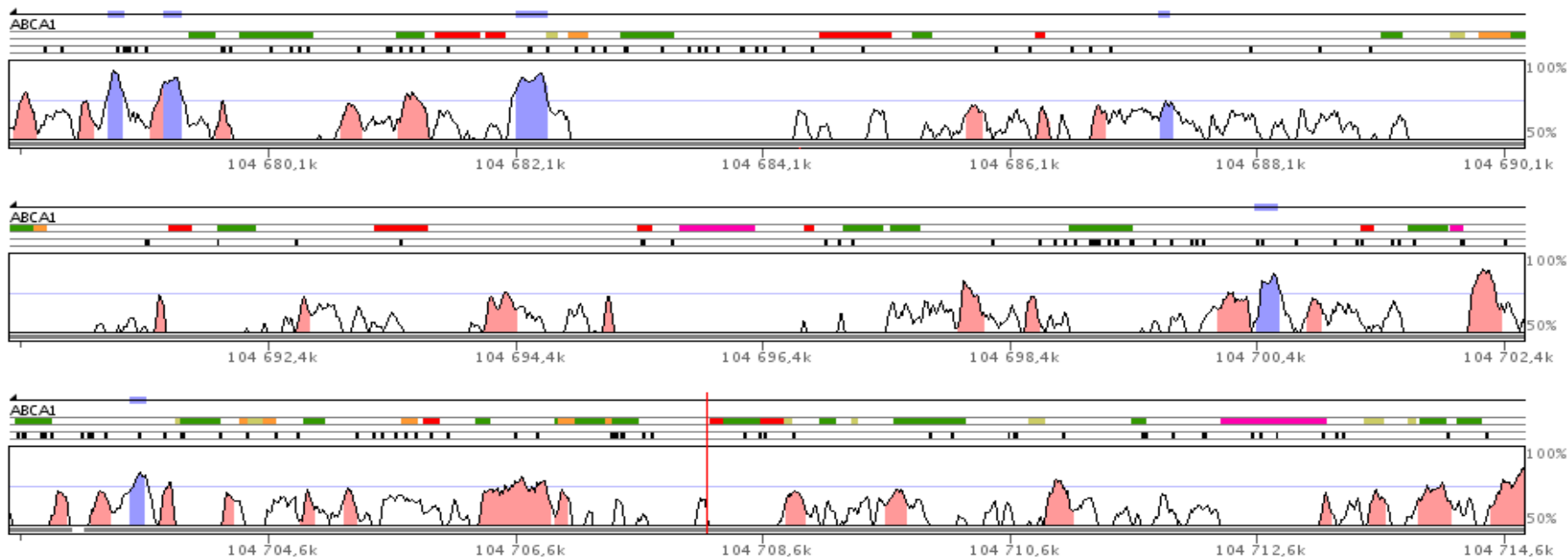


Comparative genomics :
what is under selective pressure ?

what type of conservation ?

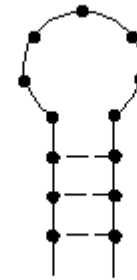
- nucleotide conservation: conserve **primary** sequence

A **C** C G **T** T A C A T **G** **G** T A **A** 66% identity
 A **G** C G **A** T A C A T **C** **T** T A **C**



what type of conservation ?

- RNA conservation: conserve **secondary** structure



Hairpin loop

← A C C G U U A C A U A C **G** G U →
 A C C G U U A C A U A C **U** G U

93% identity but **no conservation of the secondary structure!**

← **A C C G U** U A C A U **A C G G U** →
G U U A C U A C A U **G U A A C**

30% identity but **conservation of the secondary structure!**

what type of conservation ?

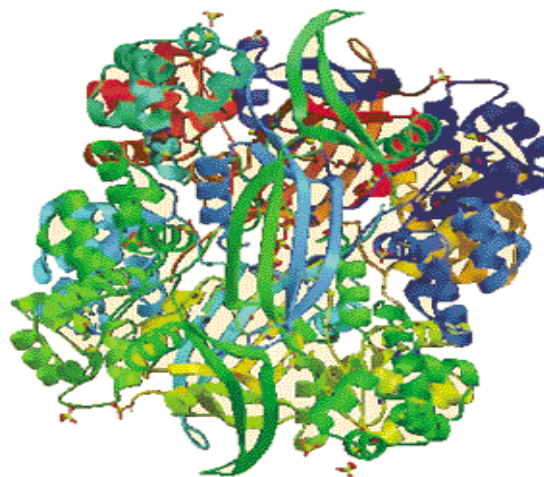
- codon conservation: conserve **protein sequence/structure**

T		V		T		W		Y					
A	C	C	G	T	A	C	A	T	G	G	T	A	T
A	C	G	G	T	A	C	C	T	G	G	T	A	C
T		V		T		W		Y					

73% nucl. id, 100% AA id

T		V		T		W		Y					
A	C	C	G	T	T	A	C	A	T	G	G	T	A
A	C	C	G	T	T	A	T	A	T	G	G	T	A
T		V		I		W		Y					

93% nucl. id, 80% AA id



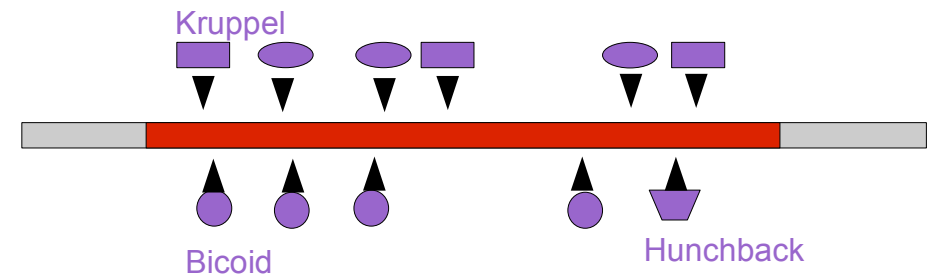
what type of conservation ?

different **constraints**
 ⇕
 different definitions of conservation

what type of constraints for non-coding DNA ?

example: **enhancers**

- *conserve full sequence?*
- *conserve sequence of BS alone ?*
- *conserve number of binding sites ?*
- *conserve order/spacing between binding sites ?*

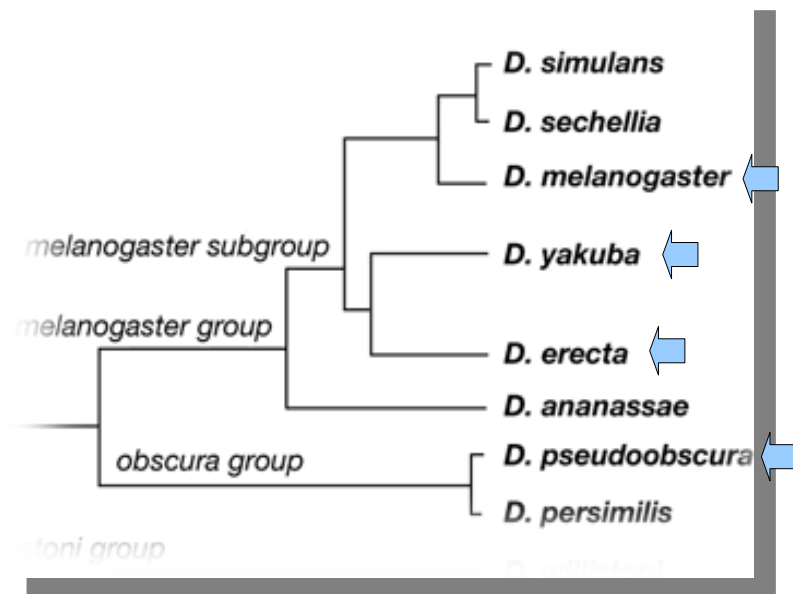
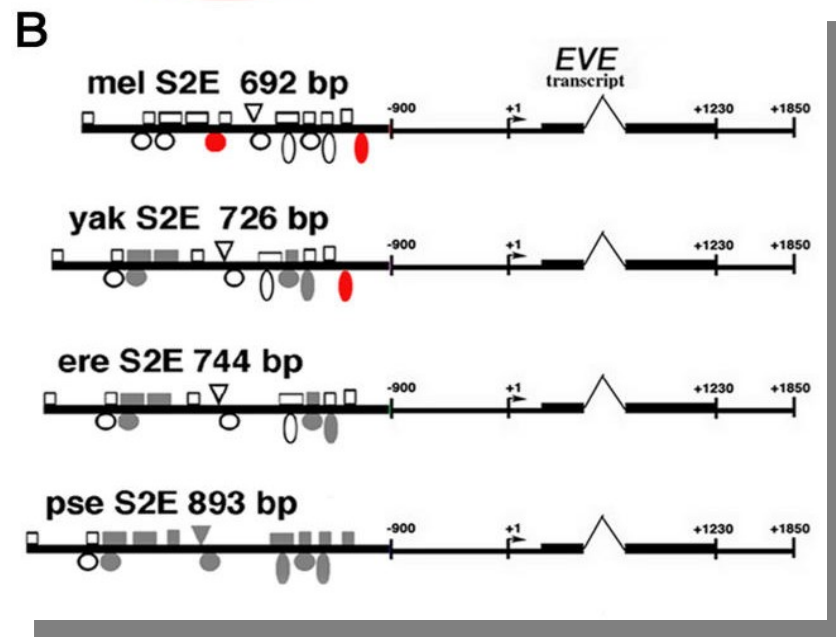


Enhancer evolution

➤ study of *eve* Stripe 2 enhancer (S2E) in *Drosophila*

- $S2E_{mel} \neq S2E_{yak/ere}$
- $S2E_{mel} \approx S2E_{pse}$

**functional conservation
not correlated to
sequence conservation !!**



Binding site conservation

BMC Bioinformatics



networks in multicellular organisms, the body patterning of the fly embryo.

Results: We find that 50%–70% of known binding sites reside in conserved sequence blocks, but these percentages are not greatly enriched over what is expected by chance. Finally, a computational genome-wide search in both species for regulatory modules based on clusters of binding sites suggests that genes central to the regulatory network are consistently recovered.

Conclusions: Our results indicate that binding sites remain clustered for these "core modules"

Downloaded from www.genome.org on November 7, 2006

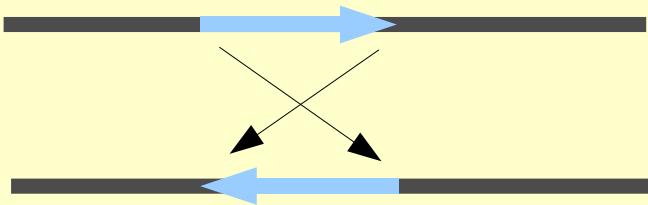
Article

Comparative conservation of cis-elements and cis-elements

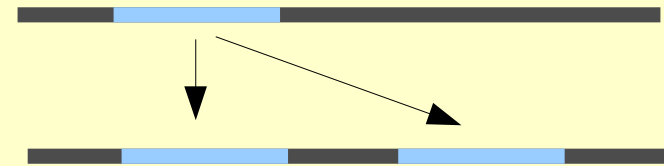
In aligned CREs of 72%¹⁰ amounts to ~10 identical bases in 14, whereas the nearby sites, at ~66% identical, would be expected to have 9.24 identical bases in 14. That difference, though statistically significant, amounts to <1 bp of excess conservation per site. Such a slight difference in conservation would appear to offer scant hope of identifying CREs through pairwise sequence conservation alone in these species. Additional information, such as knowledge of gene expression patterns and known motifs, as in Grad et al. (2004), or genome sequences of additional related species, as in Kellis et al. (2003), will be needed.

Mechanisms of evolution

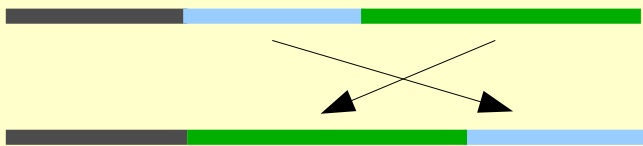
inversions



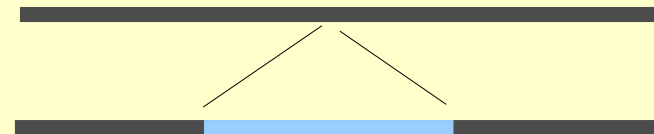
duplications



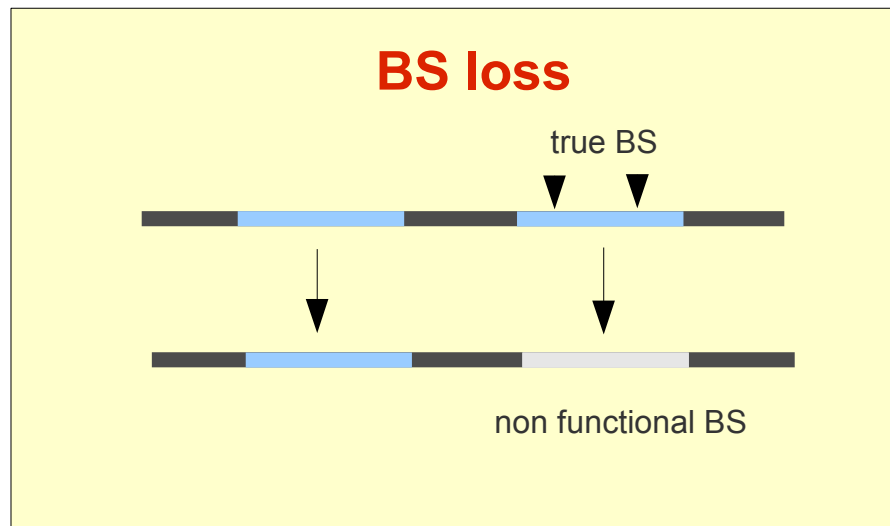
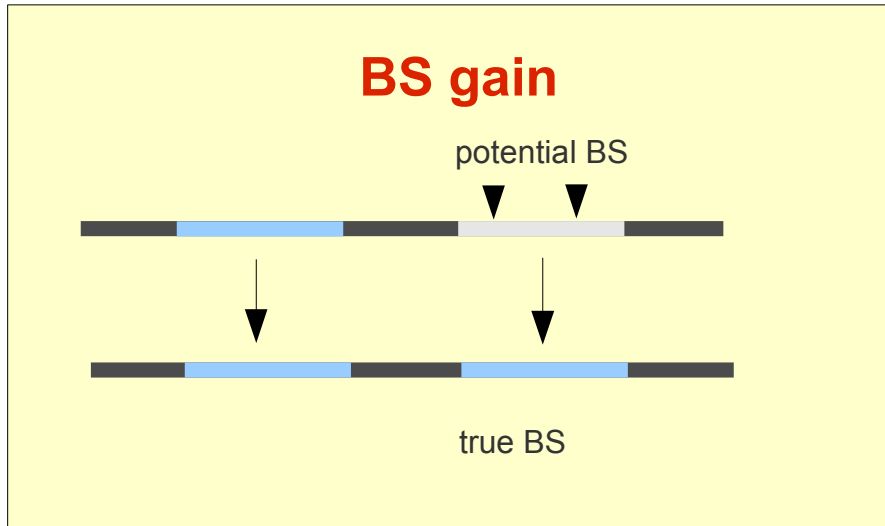
translocations



insertions



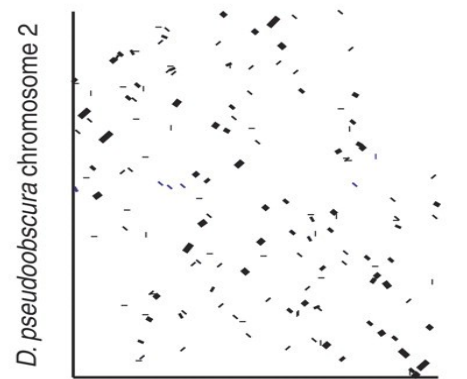
Binding site turnover



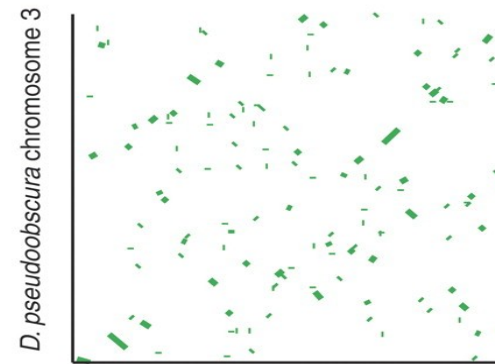
- study of *Zeste* BS turnover in drosophila [Moses et al, 2006]

>5% of the BS were gained in D. over the last 10M years

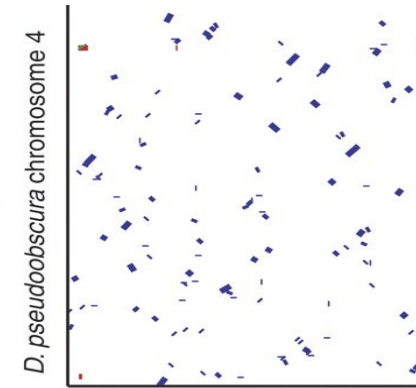
D. pseudoobscura vs. *D. melanogaster*



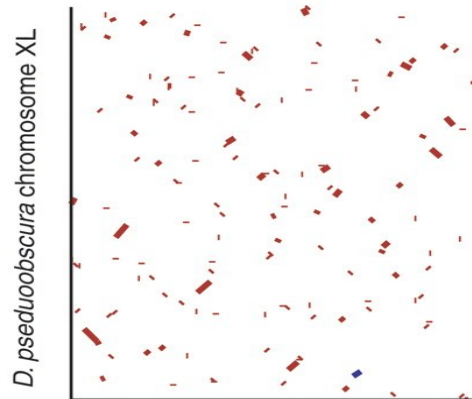
D. melanogaster chromosome 3R



D. melanogaster chromosome 2R



D. melanogaster chromosome 2L

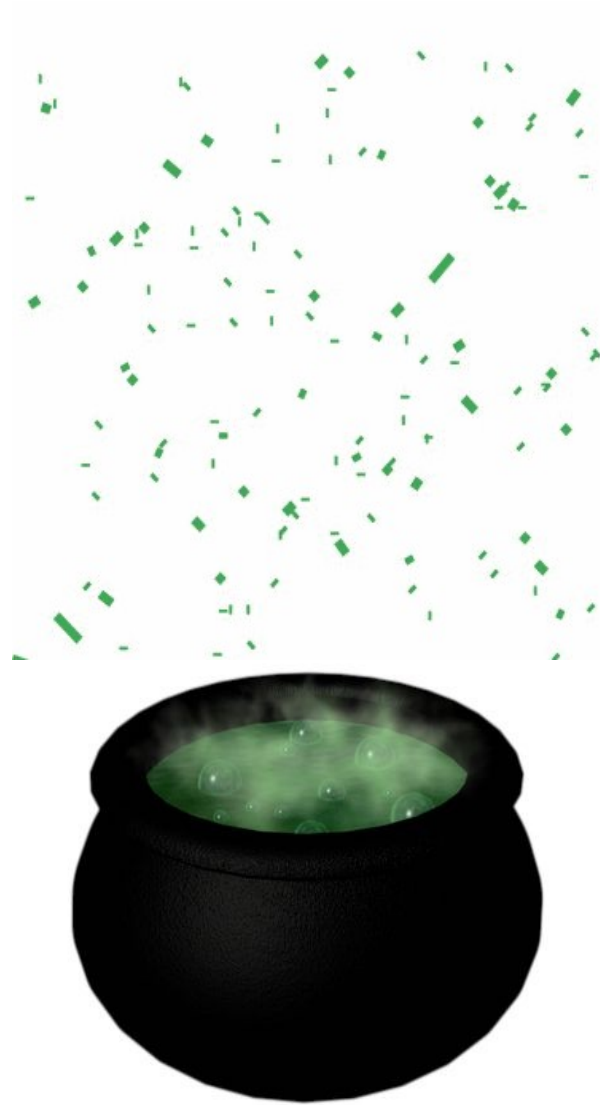


D. melanogaster chromosome X



D. melanogaster chromosome 3L

"evolution's cauldron" (Kent et al.)



*something has
happened here ...*

different scales involved here !

- average size of syntenous block *D.mel./D.pseudo*: ~83 kb
- small scale rearrangements \leq a few kb

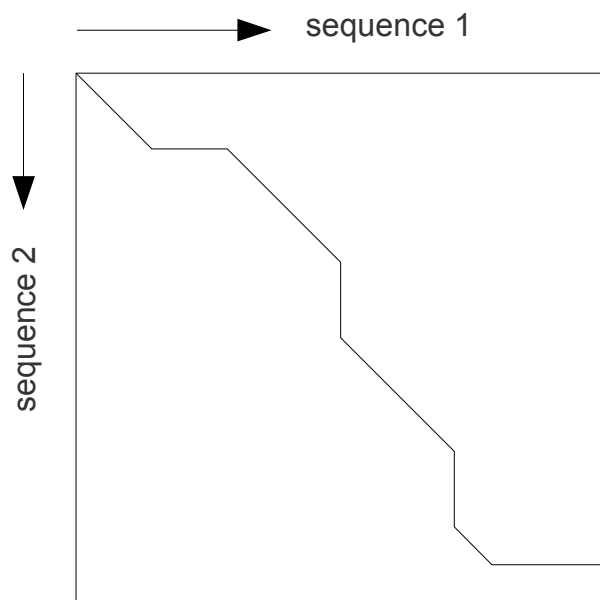
Human-mouse analysis [Kent et al, PNAS 2003]

http://www.pnas.org.gate1.inist.fr - PNAS -- Kent et al. 100 (20): 11484 Table BL2 - Mozilla Firefox

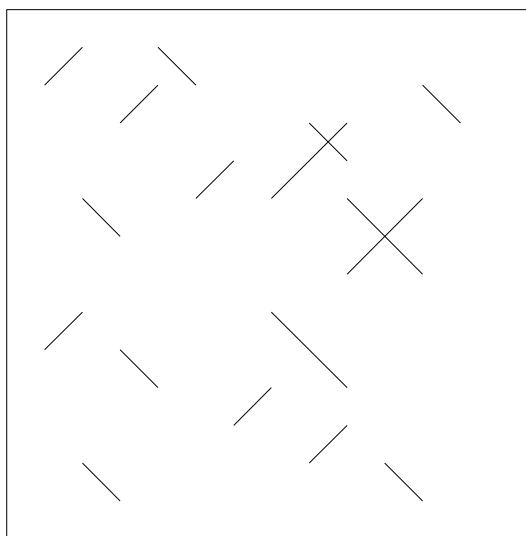
	Genomewide frequency (events per megabase)	Finished frequency (events per megabase)	Genome median size	Finished median size
Inversion	2.0	1.8	814	762
Inversion + local duplication	0.5	1.0	275	302
Inversion + local part duplication	0.7	0.8	517	1235
Local move	0.8	1.0	204	246
Local duplication	1.9	4.0	211	351
Local part duplication	0.9	1.2	343	388
Syntenic move	0.8	1.6	223	322
Syntenic duplication	1.3	1.2	283	286
Syntenic part duplication	0.7	0.8	474	946
Nonsyntenic move	5.0	5.2	104	109
Nonsyntenic duplication	11.9	11.6	235	228
Nonsyntenic part duplication	4.6	4.6	282	256
Mouse 1 base gaps	1,461.8	1,513.4	1	1
Mouse 10 base gaps	39.7	46.4	10	10
Mouse gaps \geq 100	68.8	80.8	207	201
Double gaps \geq 100	398.6	419.9	444	411
H likely deletion \geq 100	230.0	223.5	685	633

how can we study these events ?
how can we relate them to coding regions ?

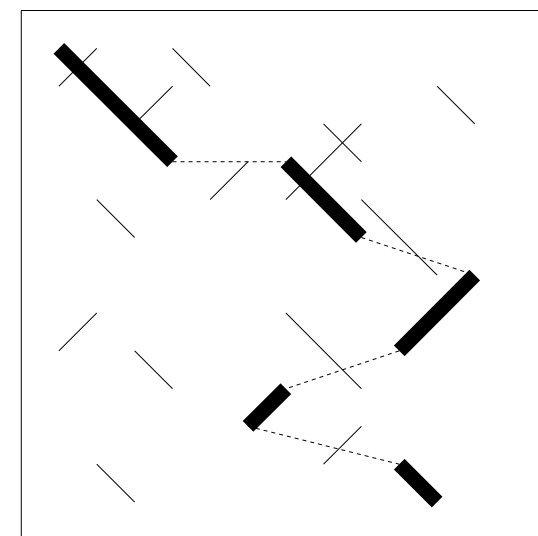
comparing sequences



global



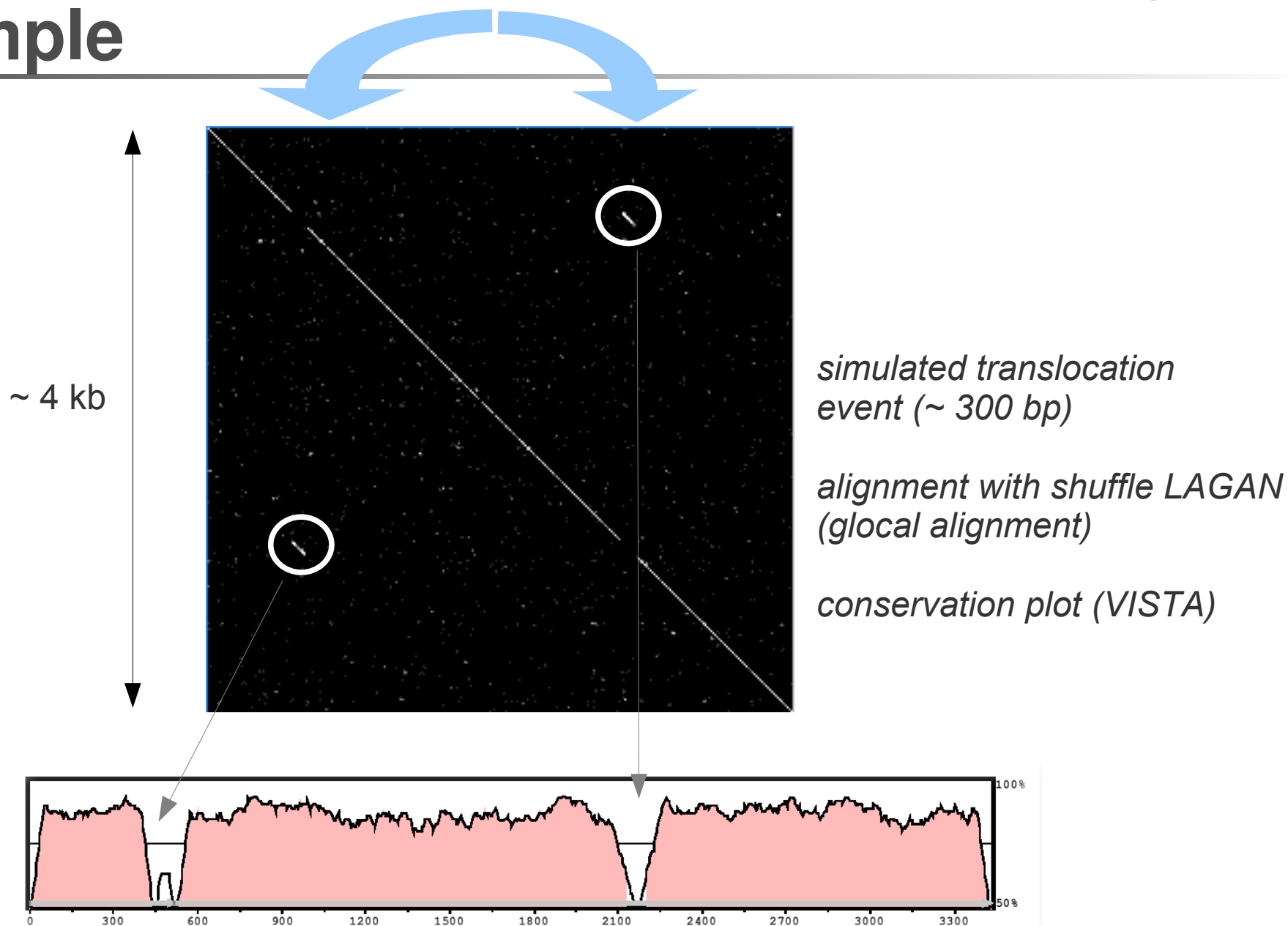
local



"glocal"
(shuffle LAGAN)

do they see these genomic events ?

example



small scale events are hard to see in global (glocal) alignments

our scope

provide a **genome wide**, **high resolution** map of conserved non-coding blocks in *Drosophila*

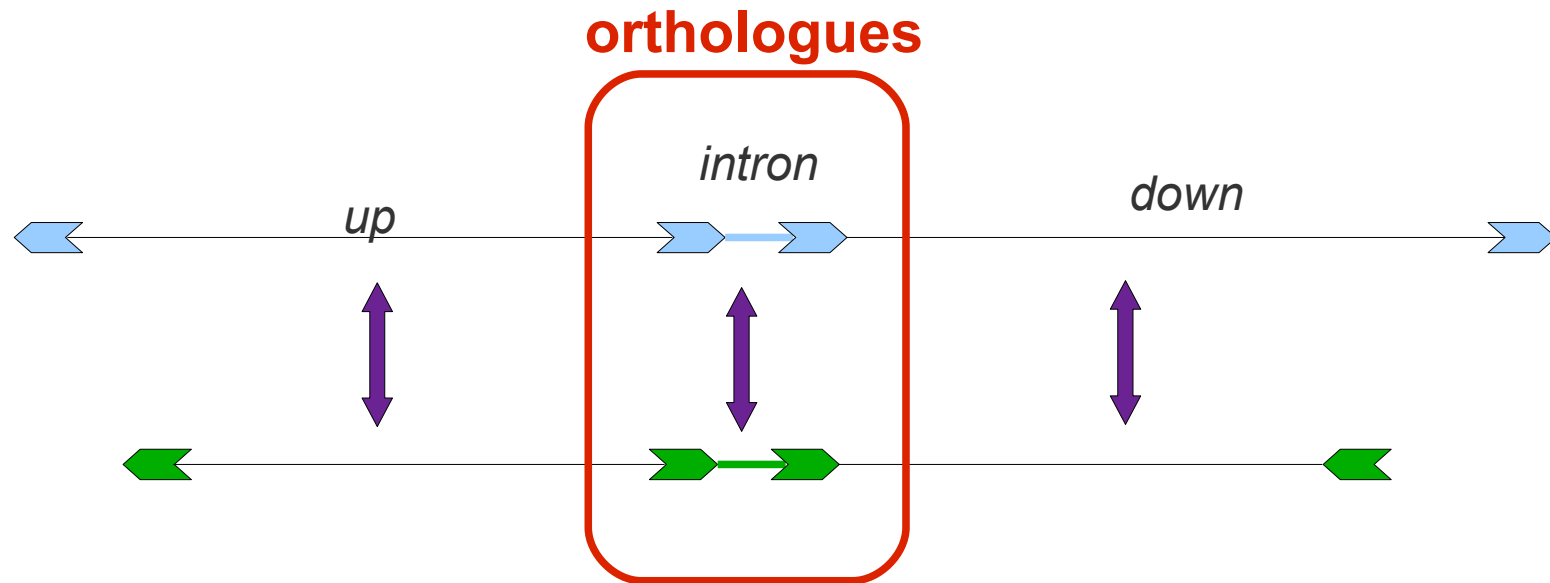
- 1) identify conserved non-coding blocks in *Drosophila*
- 2) make inventory of small scale rearrangement events in non-coding DNA
- 3) study evolution of known functional elements (e.g. enhancers)
"phylogeny" of non-coding DNA ?

drosOCB drosOphila Conserved Blocks

- a catalogue of **non-coding conserved elements** in drosophila
- *D.mel./D.xxx* **pairwise** alignments
- **gene-centric, local** alignments (CHAOS)

in collaboration with Loredana Martignetti (Univ. of Turin)

align orthologous regions

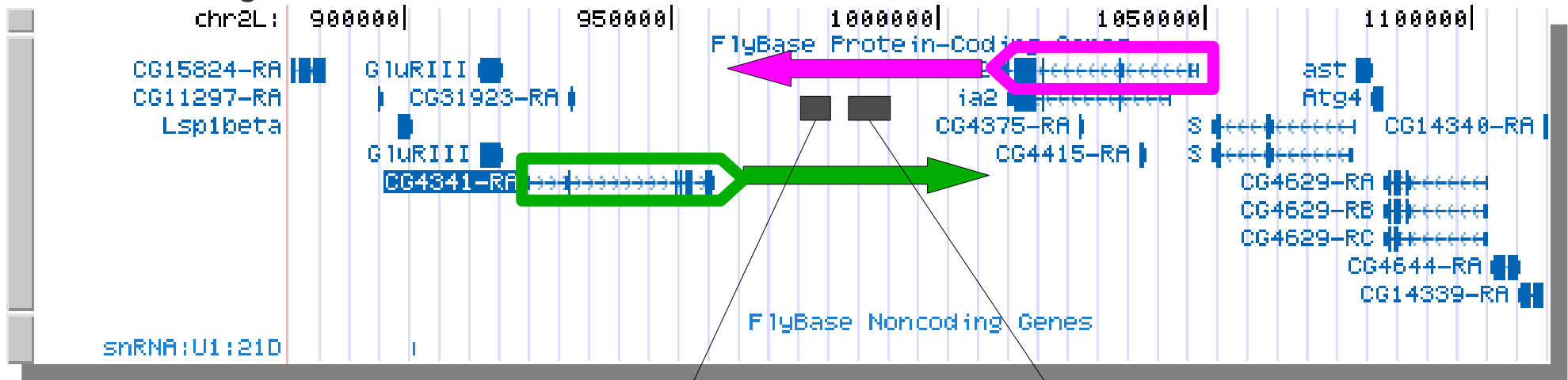


- depends on orthology relationships
D.mel/D.pse: 12160 pairs; D.mel/D.vir: ~ 10000 pairs
- depend **crucially** on annotations ! (use recent consensus annotations, July 2006)

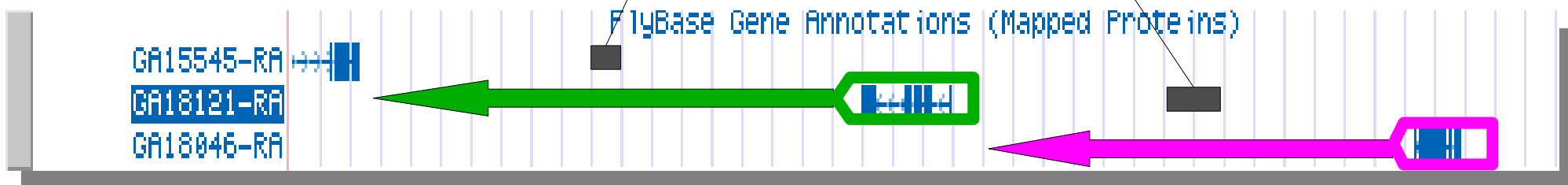
cf. AAA Drosophila comparative genomics site [<http://rana.lbl.gov/drosophila/>]

"genome tiling alignment"

D. melanogaster

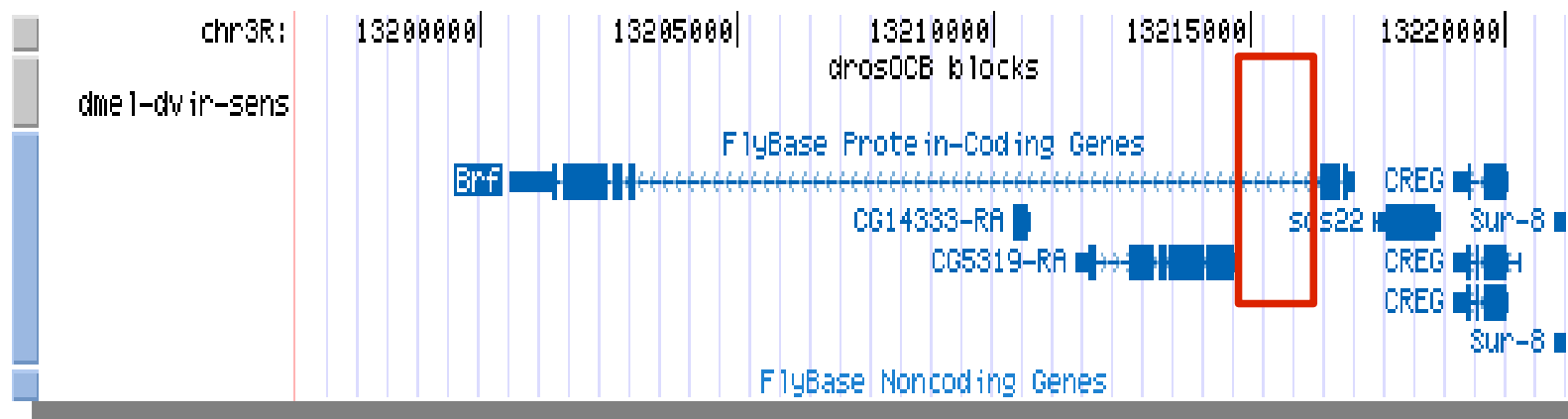
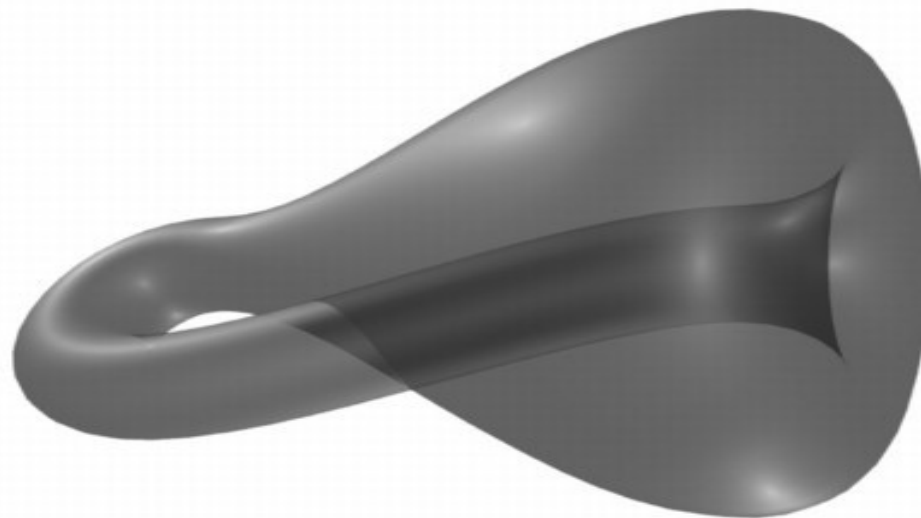


D. pseudoobscura

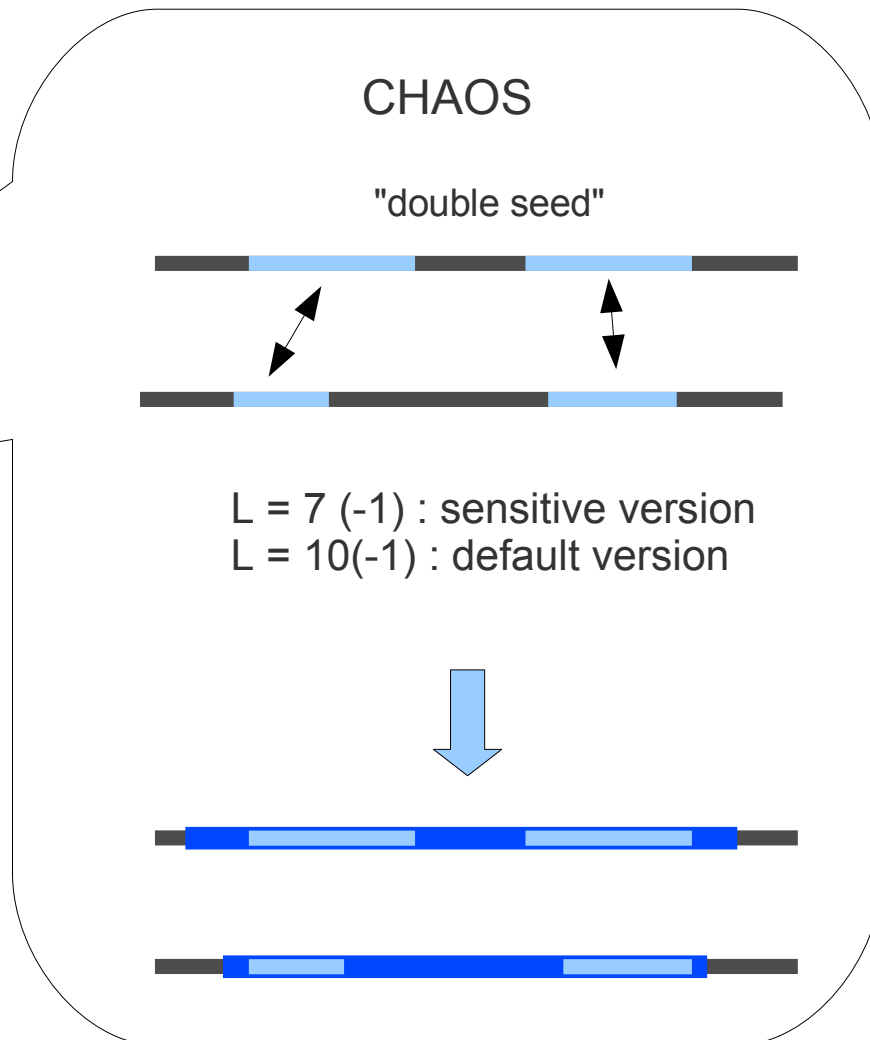
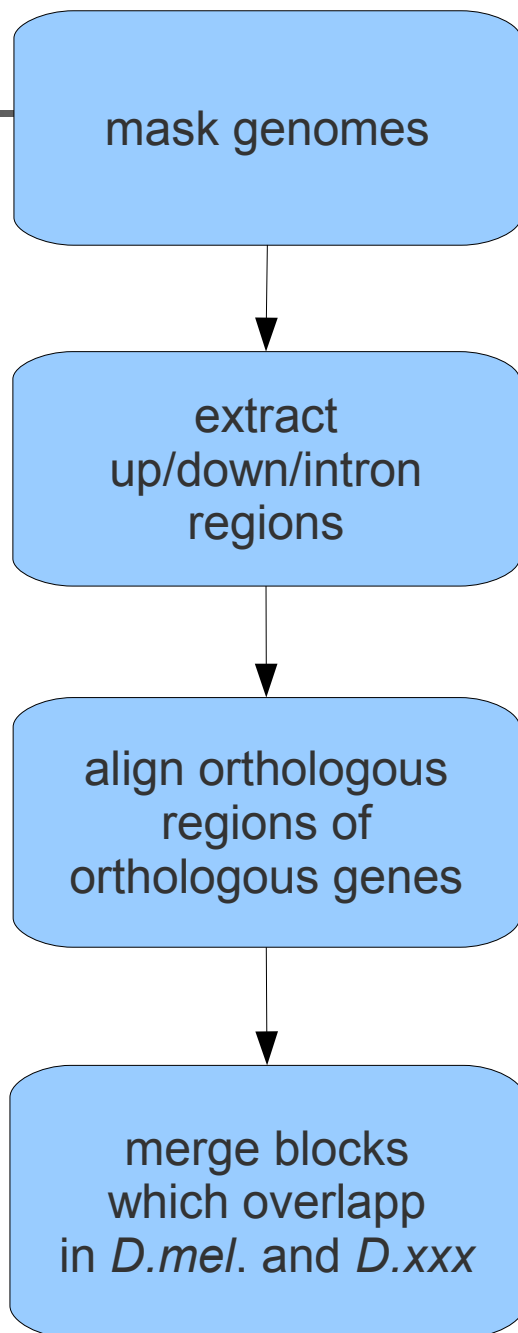


blocks are attributed to one or several genes

where is in, where is out ?



flowchart



[Chaos: Brudno et al., BMC Bioinfo]

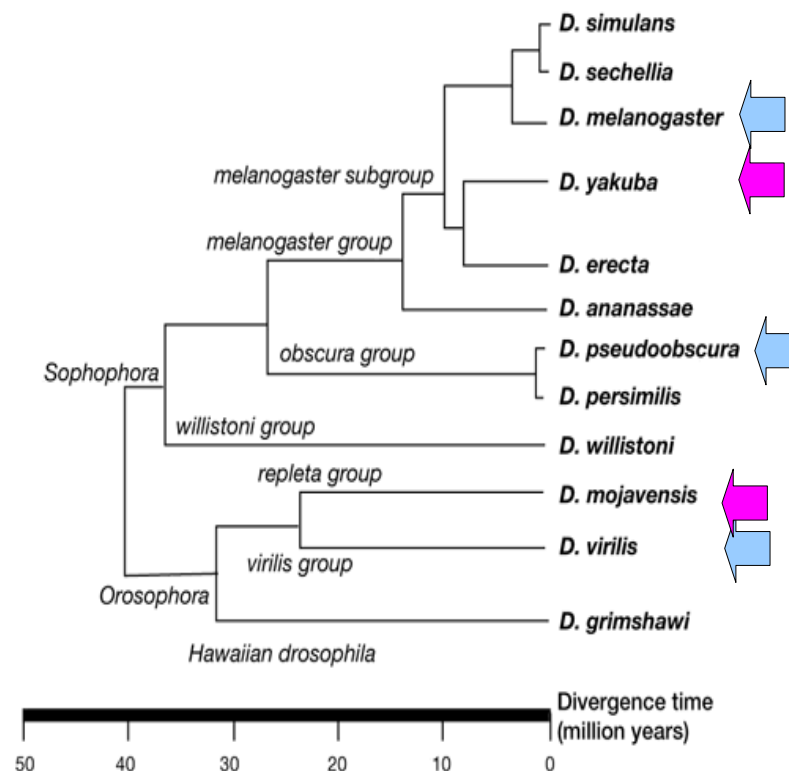
Some statistics

D.melanogaster vs. *D.pseudoobscura*

	Intergenic		Intronic	
	55.5 Mb		38.5 Mb	
Region aligned	default	sensitive	default	sensitive
# CB	156283	256023	35825	62864
total size of CB	9.1 Mb	16.1 Mb	2.2 Mb	4.3 Mb
% of aligned reg.	16.4%	29.0%	5.6%	11.1%
mean size of CB	59.5 bp	72.0 bp	61 bp	73.5 bp

D.melanogaster vs. *D.virilis*

	Intergenic		Intronic	
	48.1 Mb		34.7 Mb	
Region aligned	default	sensitive	default	sensitive
# CB	50022	103487	23429	63791
total size of CB	2.6 Mb	6.1 Mb	1.2 Mb	3.6 Mb
% of aligned reg.	5.0%	12.7%	3.4%	10.4%
mean size of CB	53.2 bp	65.7 bp	54.6 bp	67.4 bp



alignments vs. scale



s c a l e

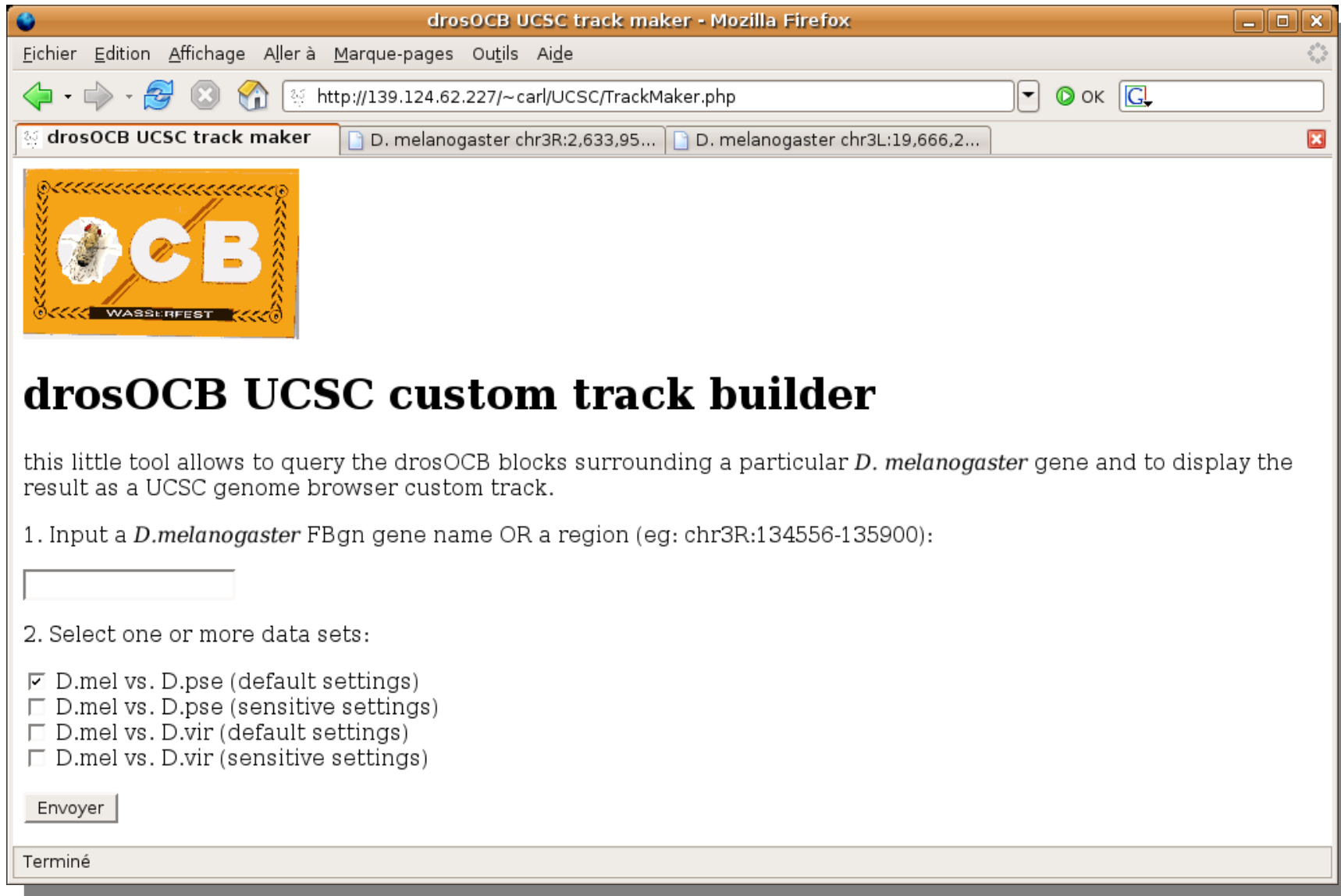
local
(BLAST, CHAOS,...)

lobal©
(drosOCB)

glocal
(shuffleLAGAN...)

global

"database" "interface"




drosOCB UCSC track maker - Mozilla Firefox

Eichier Edition Affichage Aller à Marque-pages Outils Aide

http://139.124.62.227/~carl/UCSC/TrackMaker.php

drosOCB UCSC track maker



drosOCB UCSC custom track builder

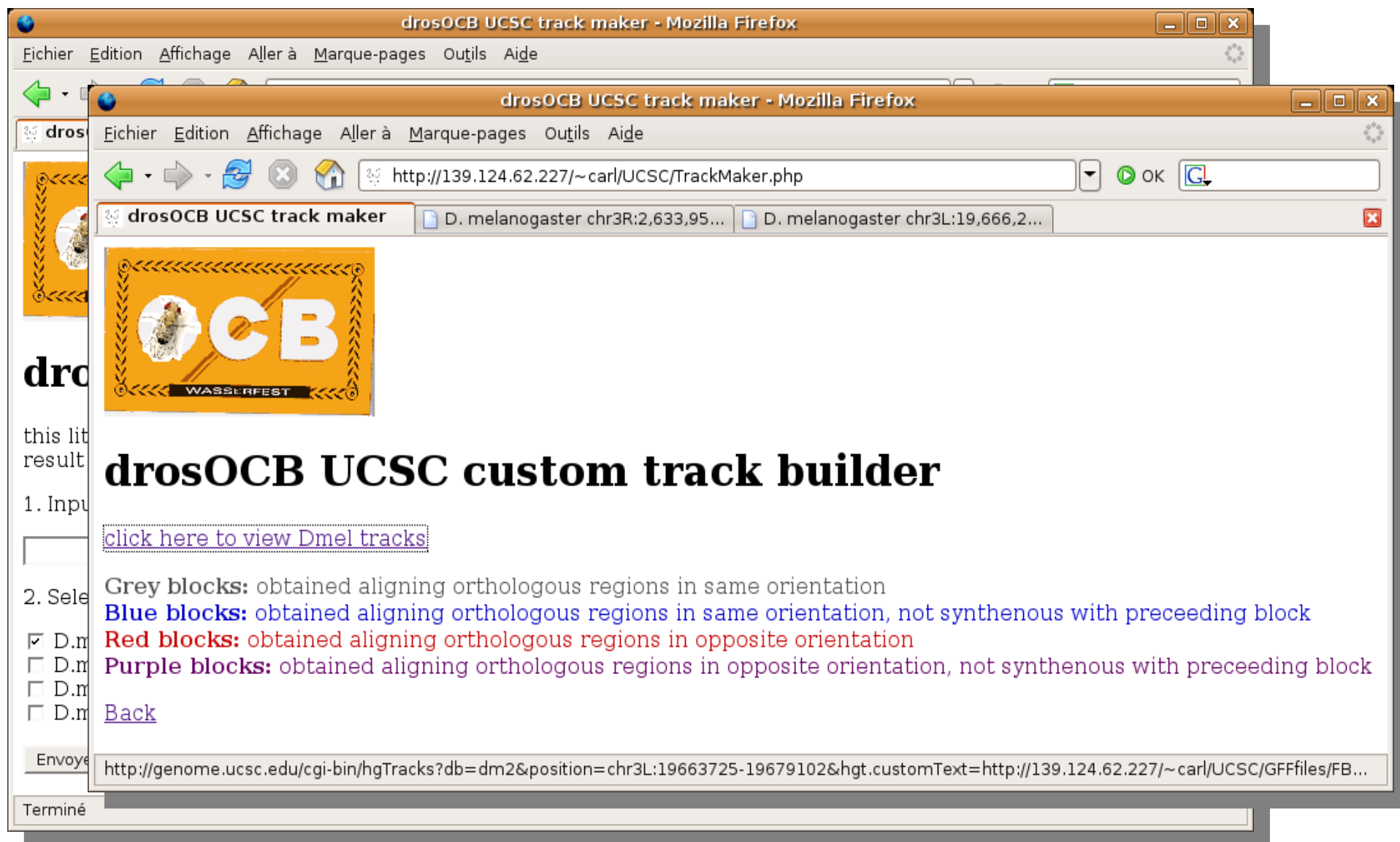
this little tool allows to query the drosOCB blocks surrounding a particular *D. melanogaster* gene and to display the result as a UCSC genome browser custom track.

1. Input a *D.melanogaster* FBgn gene name OR a region (eg: chr3R:134556-135900):
2. Select one or more data sets:
 - D.mel vs. D.pse (default settings)
 - D.mel vs. D.pse (sensitive settings)
 - D.mel vs. D.vir (default settings)
 - D.mel vs. D.vir (sensitive settings)

Envoyer

Terminé

"database" "interface"



dro

this lit
result

1. Inpu

2. Sele

D.m
 D.m
 D.m
 D.m

Envoye

Terminé

drosOCB UCSC custom track builder

[click here to view Dmel tracks](#)

Grey blocks: obtained aligning orthologous regions in same orientation
Blue blocks: obtained aligning orthologous regions in same orientation, not synthenous with preceeding block
Red blocks: obtained aligning orthologous regions in opposite orientation
Purple blocks: obtained aligning orthologous regions in opposite orientation, not synthenous with preceeding block

[Back](#)

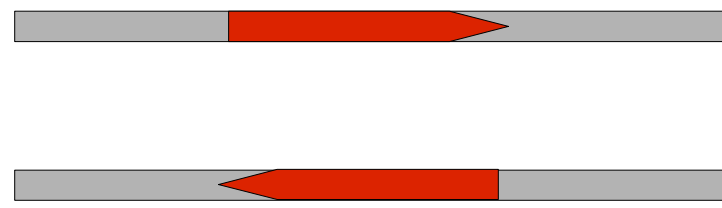
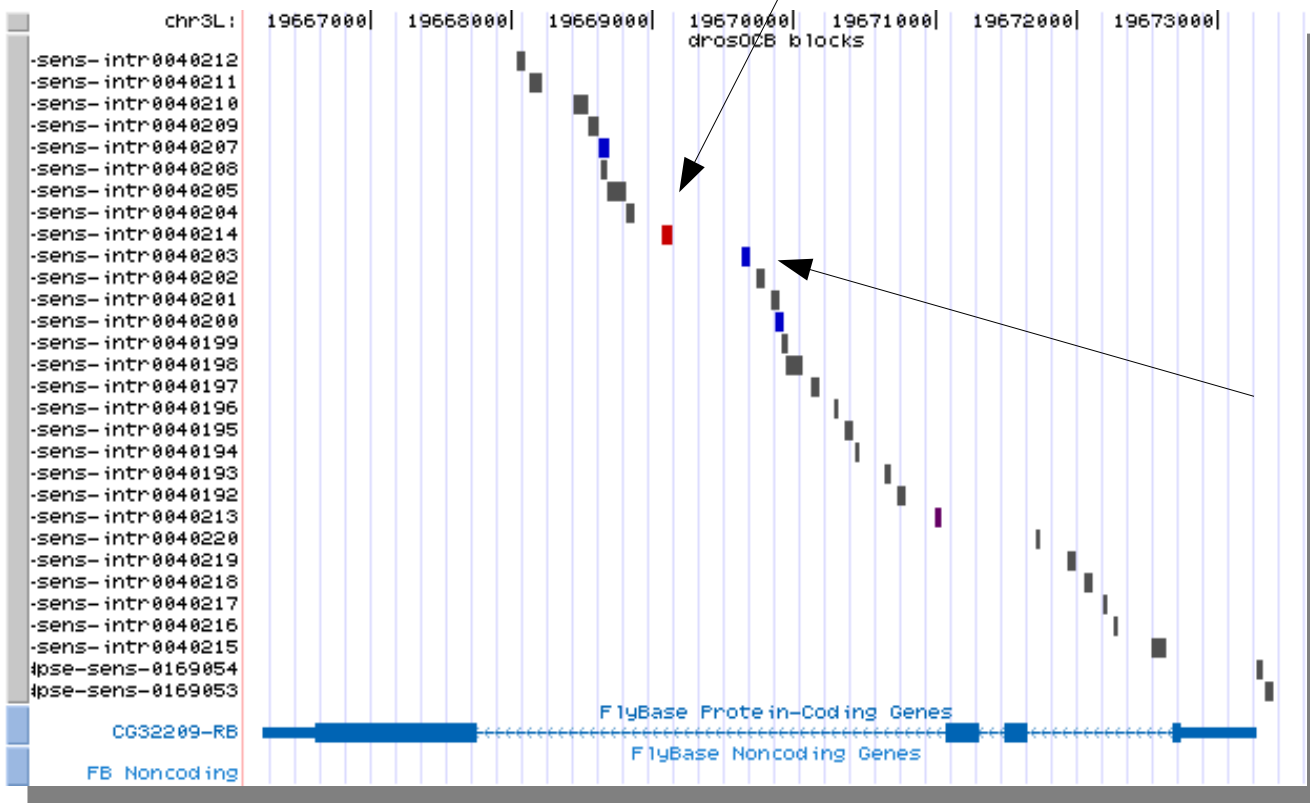
http://genome.ucsc.edu/cgi-bin/hgTracks?db=dm2&position=chr3L:19663725-19679102&hgt.customText=http://139.124.62.227/~carl/UCSC/GFFfiles/FB...

"database" "interface"

The screenshot displays a multi-layered browser window. The top-most window is titled "drosOCB UCSC track maker - Mozilla Firefox" and shows a search bar with the URL "http://genome.ucsc.edu/cgi-bin/hgTracks?hgid=79931363&hgt.in1=1.5x&position=chr3L%3A19664494-19674745". Below this, another window shows the UCSC Genome Browser interface. The main heading is "UCSC Genome Browser on D. melanogaster Apr. 2004 Assembly". The search bar contains "chr3L:19,666,203-19,673,037" and indicates a "size 6,835 bp.". The visualization shows a genomic track with various features, including "FlyBase Prote in-Coding Genes" and "FlyBase Noncoding Genes". The interface includes navigation controls like "move", "zoom in", and "zoom out", and a "Custom Tracks" section with dropdown menus for "Base Position", "IPF", "dmel-dpse-def", "dmel-dpse-sens", "dmel-dvir-def", "dmel-dvir-sens", "Gap", "BAC End Pairs", "GC Percent", "Short Match", and "Restr Enzymes". The bottom of the browser shows a search bar with "7SL" and a "Terminé" status.

UCSC genome browser custom track

RED : inverted block (*inversion ?*)



BLUE: non-synthenous block (*translocation?*)

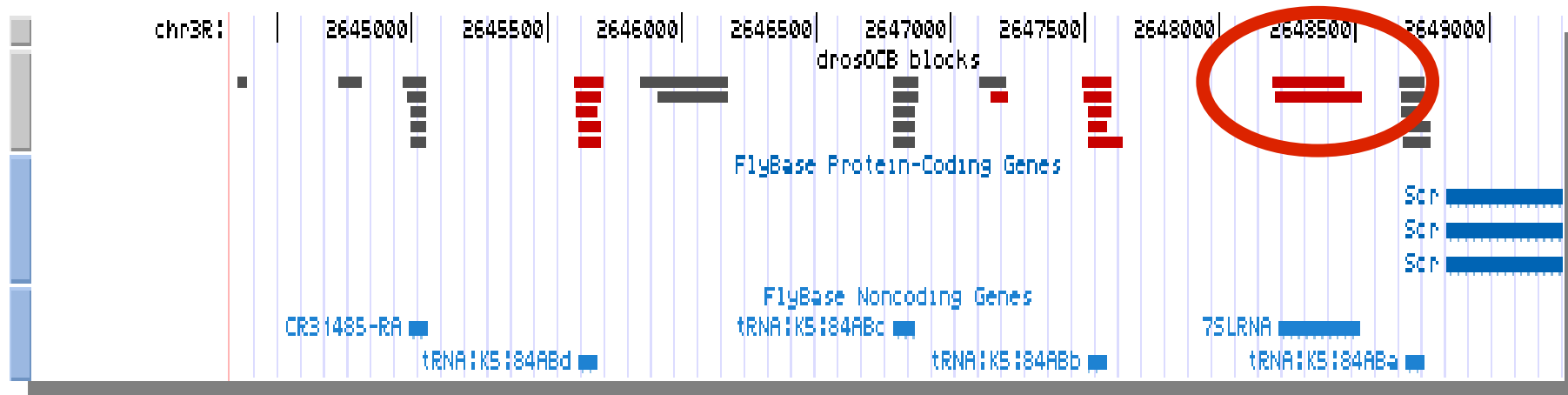
Many putative inversion events

- many conserved blocks obtained by aligning **opposite strands**
 - intergenic regions: **15%** of the blocks
 - introns: **12.5 %** of the blocks
- **not evenly distributed** over the aligned regions
 - present in **38%** of the aligned regions

Duplication of non-coding RNA genes

96% conservation
over 250 bp !

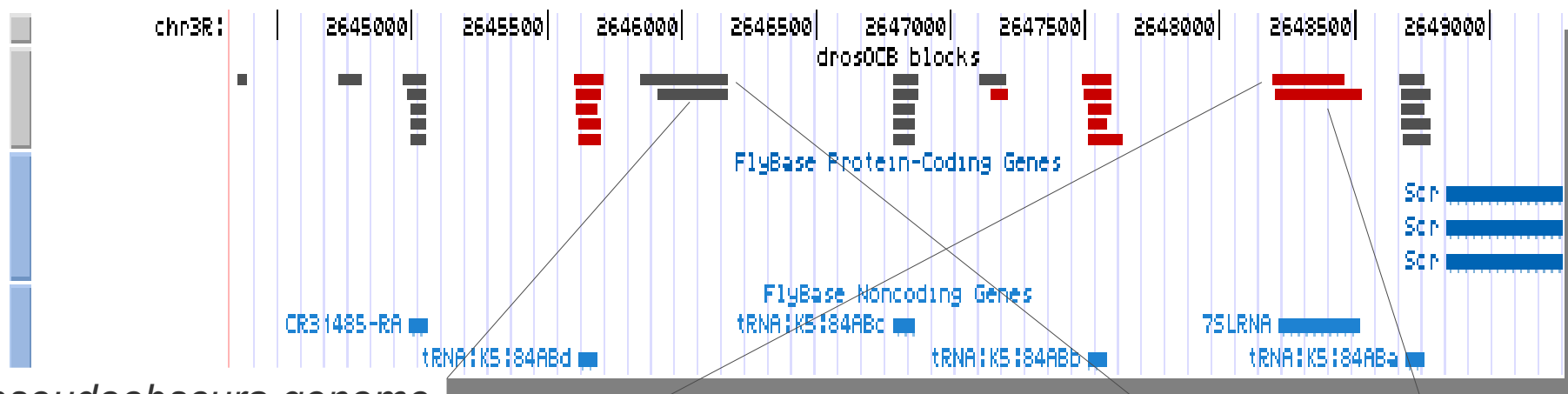
D. melanogaster genome



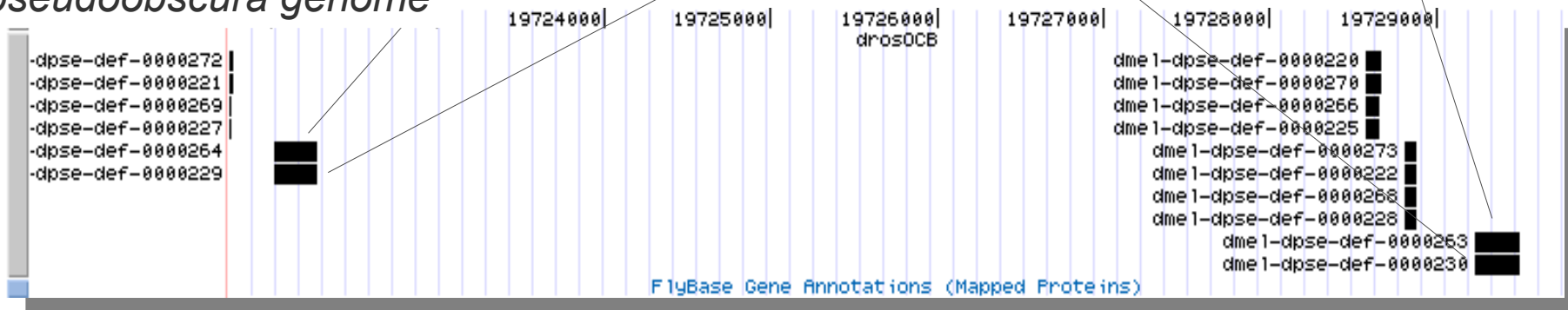
overlapping, but not merged:
map to **distinct** locations in the *D.pse* genome

Duplication of non-coding RNA genes

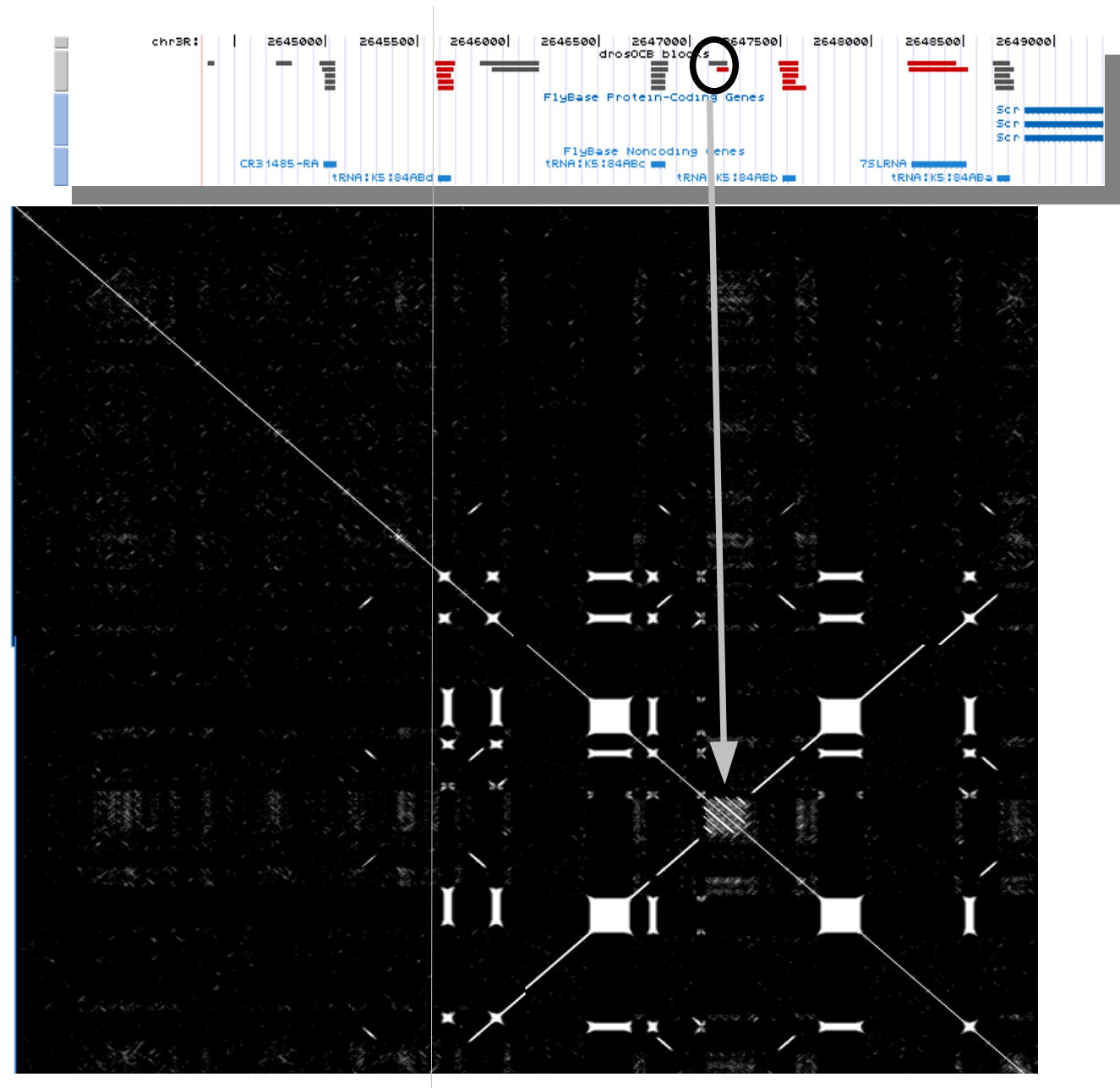
D. melanogaster genome



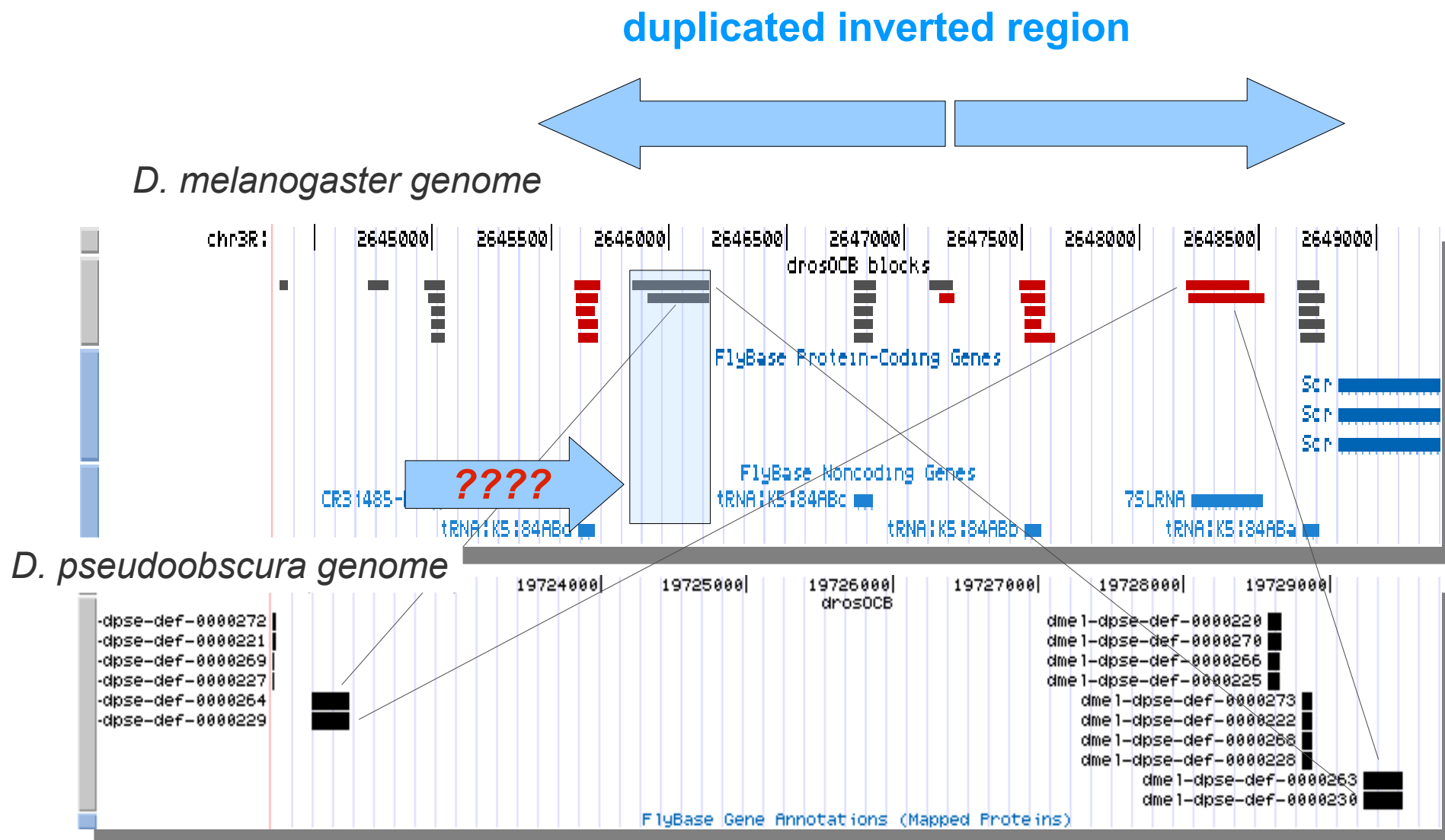
D. pseudoobscura genome



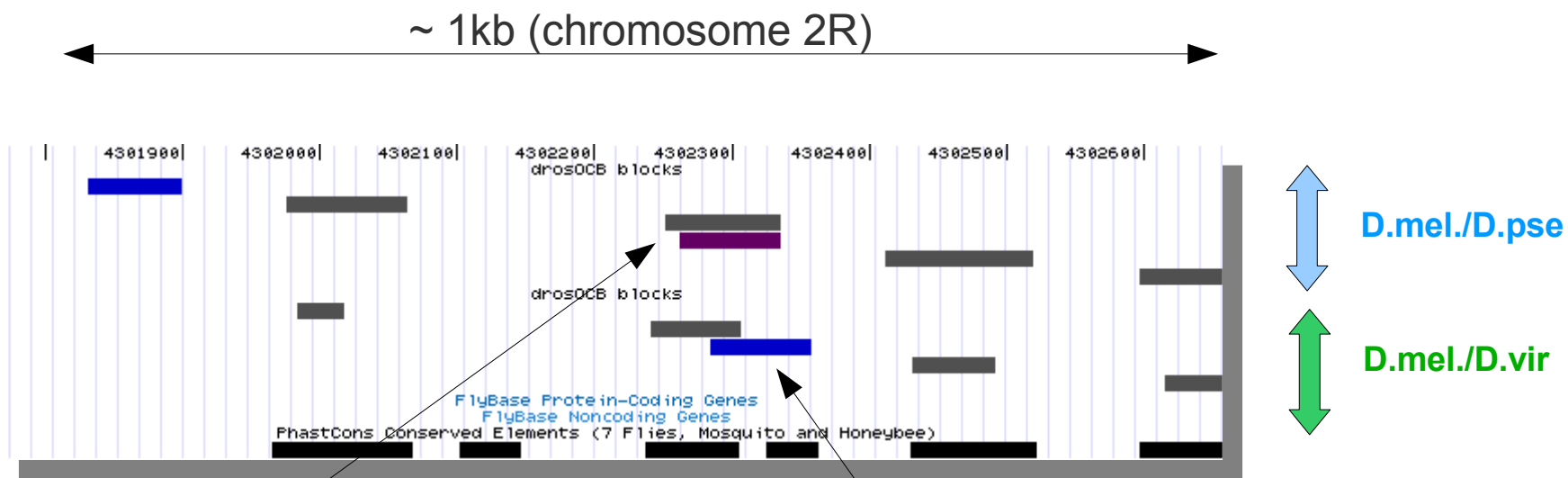
dotplot



Duplication of non-coding RNA genes



lineage specific events ?



blocks on *D.pse.*
are 18 kb away from
each other

blocks on *D.vir.*
are 31 kb away from
each other

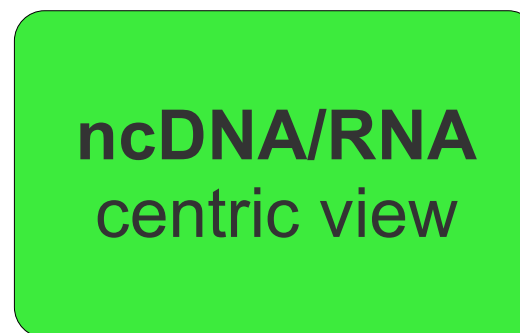
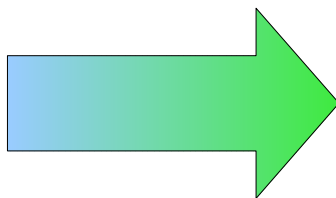
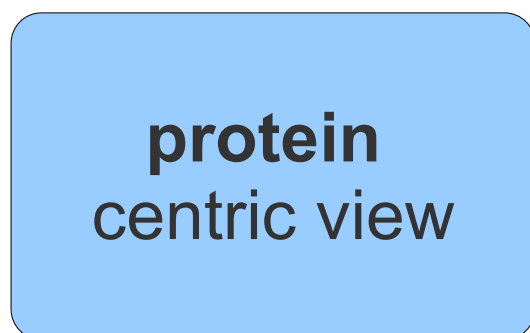
Possible usage of drosOCB

- starting point for phylogenetic footprinting
application to discovery of cardiac enhancers in Drosophila
- systematic study of small scale genomic rearrangement events
coding vs. non-coding fate
- "phylogeny" of non-coding DNA:
lineage specific elements/events ?

General conclusions

"Why would a perfect God create flawed DNA which is primarily composed of useless, non-coding regions?"

"Actually, He didn't".



the upcoming (r)evolution ?