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

Carl Herrmann
IBDML & Univ. Méditerranée - Marseille

TAG'06 – LAPTH Annecy

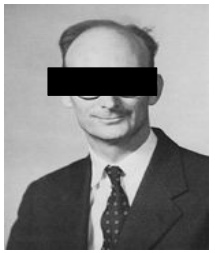
a fundamental question ...

Why am I smarter and more evolved than others ?

2 ways to ask the question

intra-species

"...than my colleagues ?"



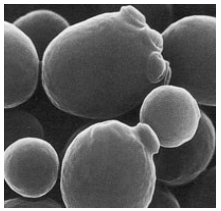
inter-species

"...than my guinea-pig ?"

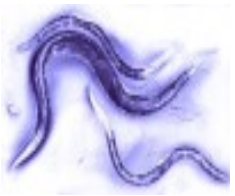
eukaryotic genome contest II

featuring :

yeast



worm



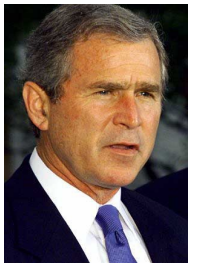
fly

arabidopsis

pufferfish

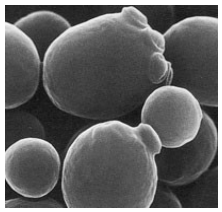
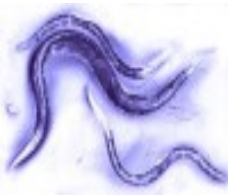
rice

human

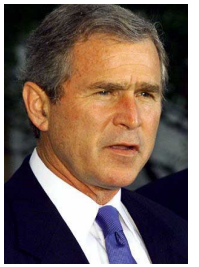


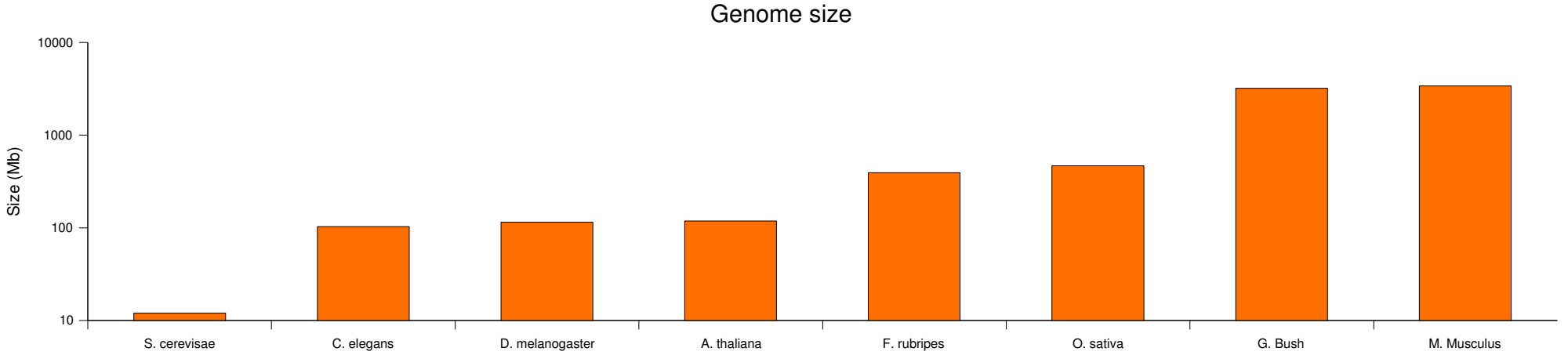
mouse

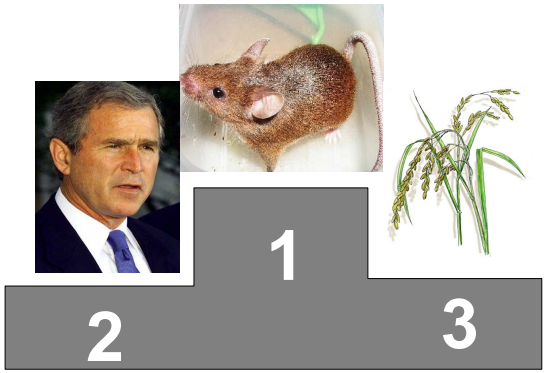
Genome size

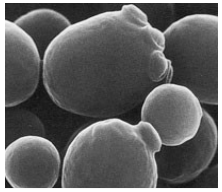
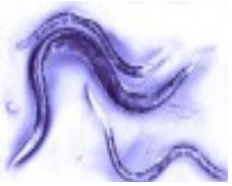


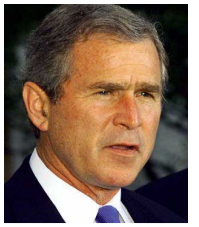
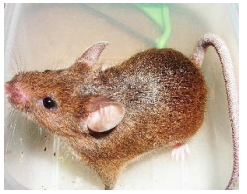


Genome size
Size (Mb)
10000
1000
100
10
S. cerevisae
C. elegans
D. melanogaster
A. thaliana
F. rubripes
O. sativa
G. Bush
M. Musculus

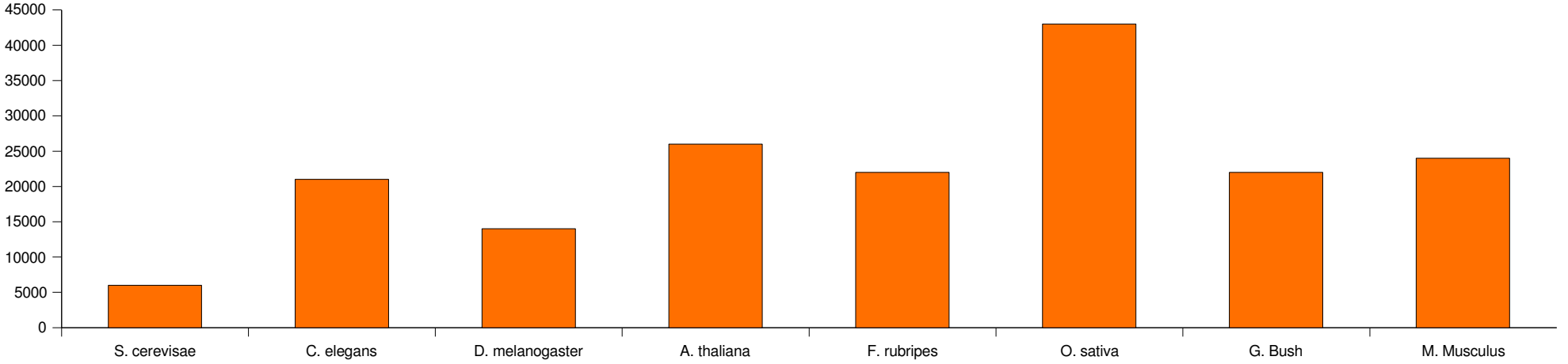

1
2
3

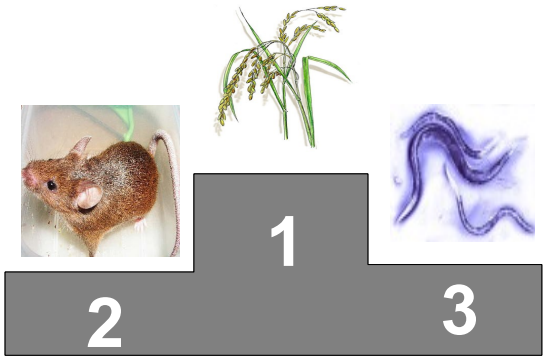
Number of genes

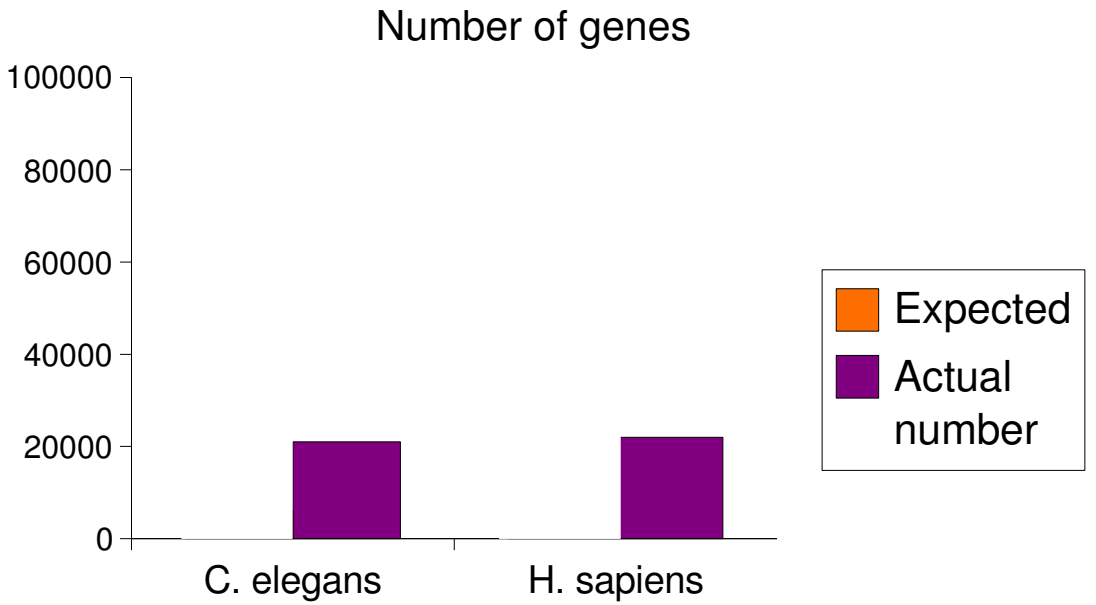


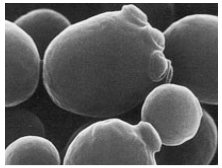
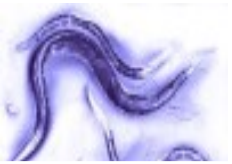




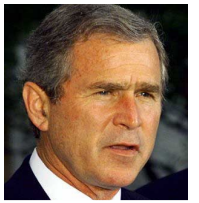
this is becoming a little embarassing for us ...

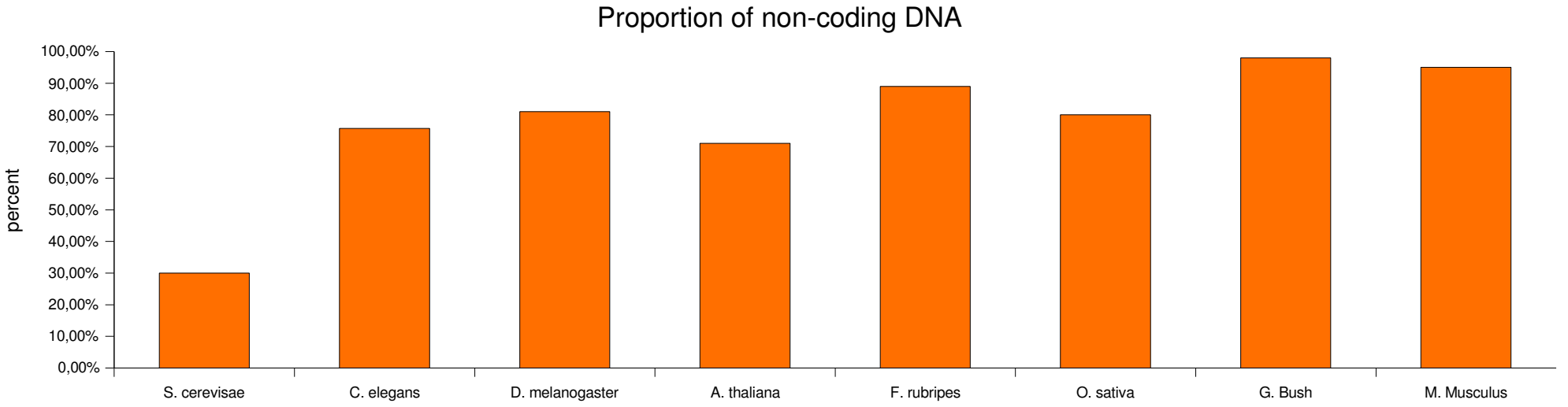


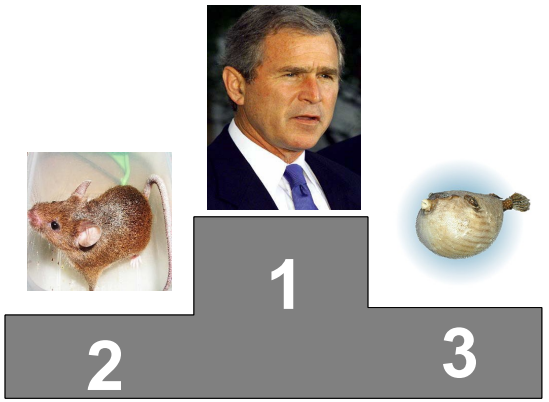
Non-coding DNA



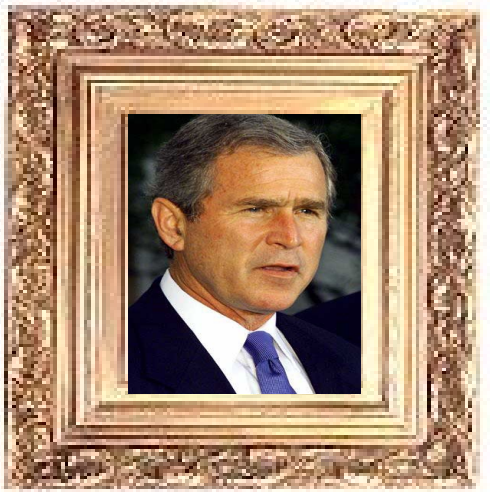





eukaryotic genome contest !!

genome size award

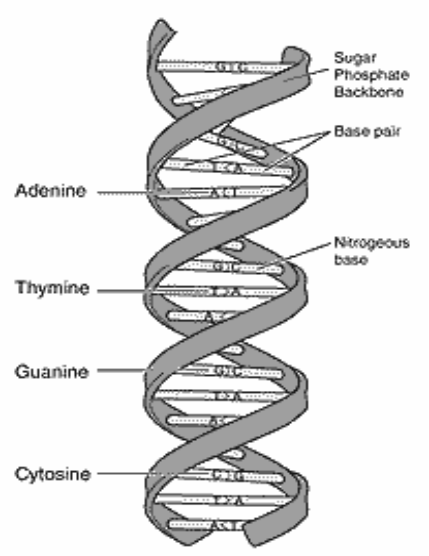
number of gene award

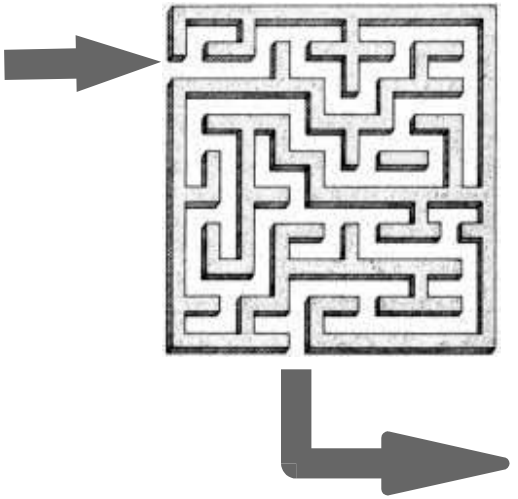


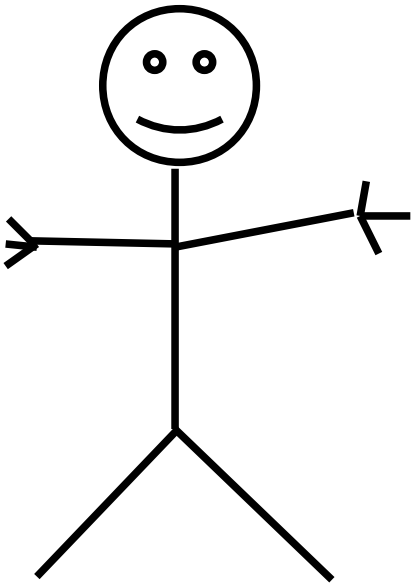
junk DNA award

how glorious...

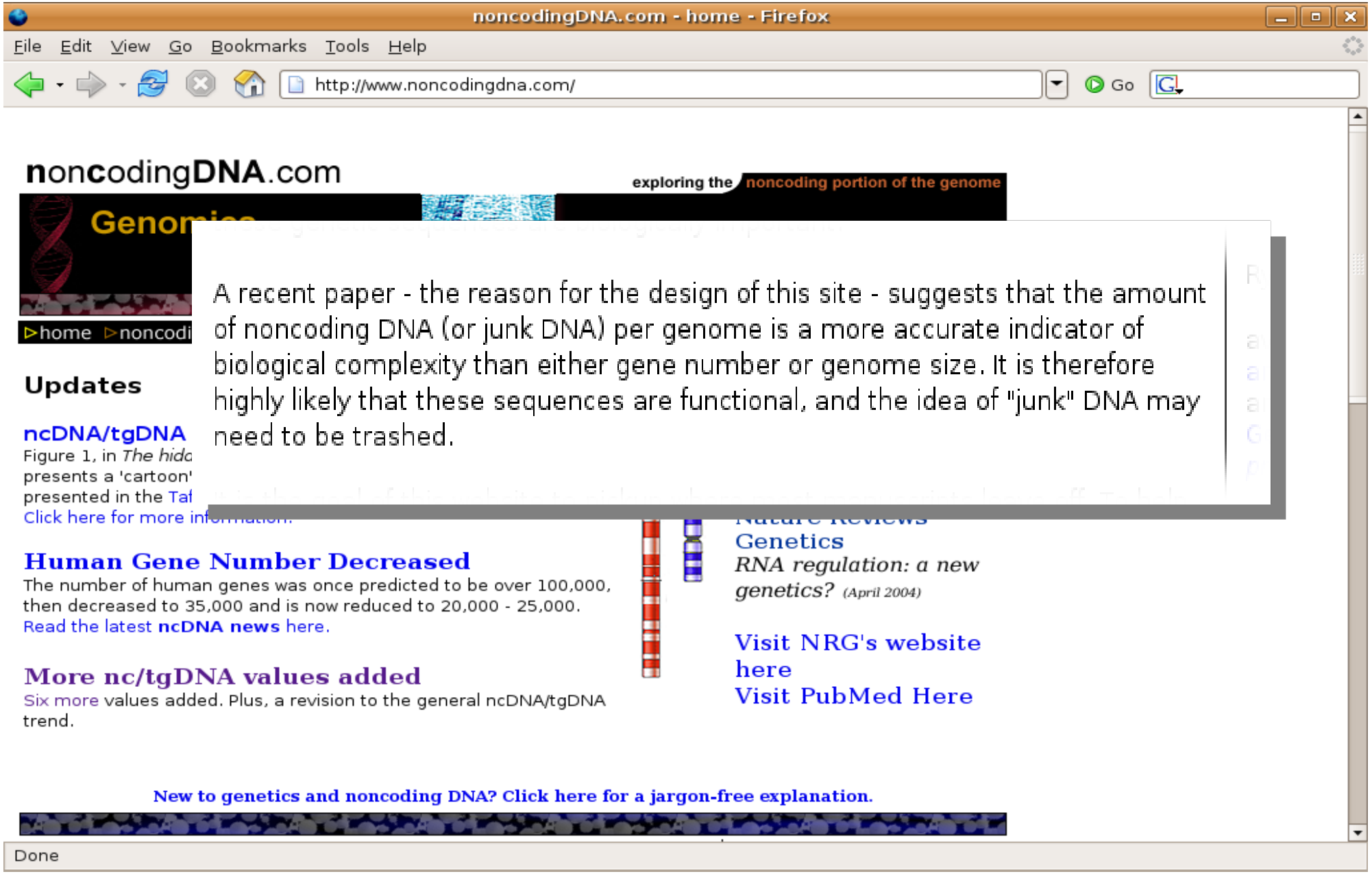
Genotype / Phenotype





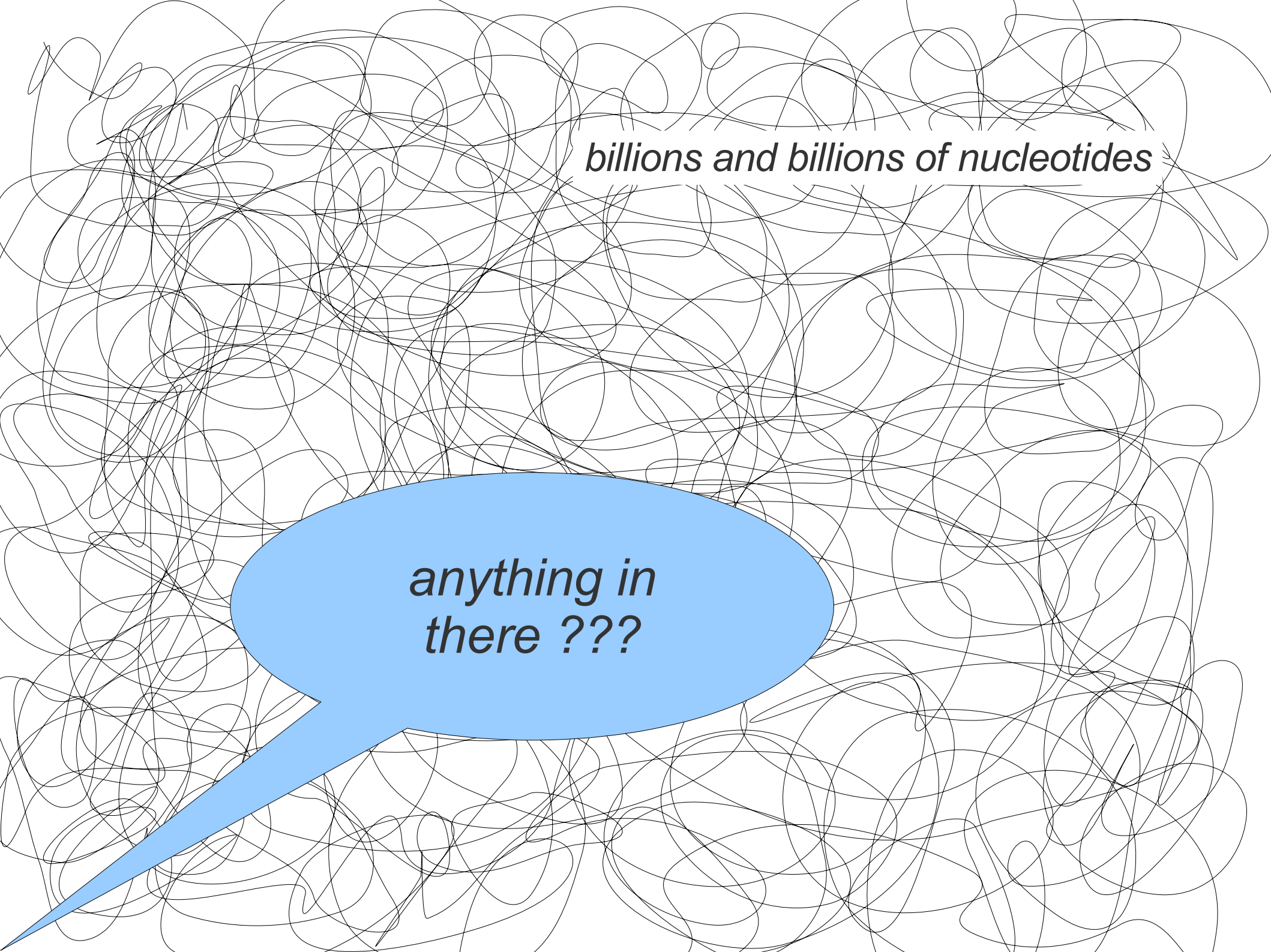


"why would a perfect God create flawed DNA which is primarily composed of useless non-coding regions ?"



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 ???

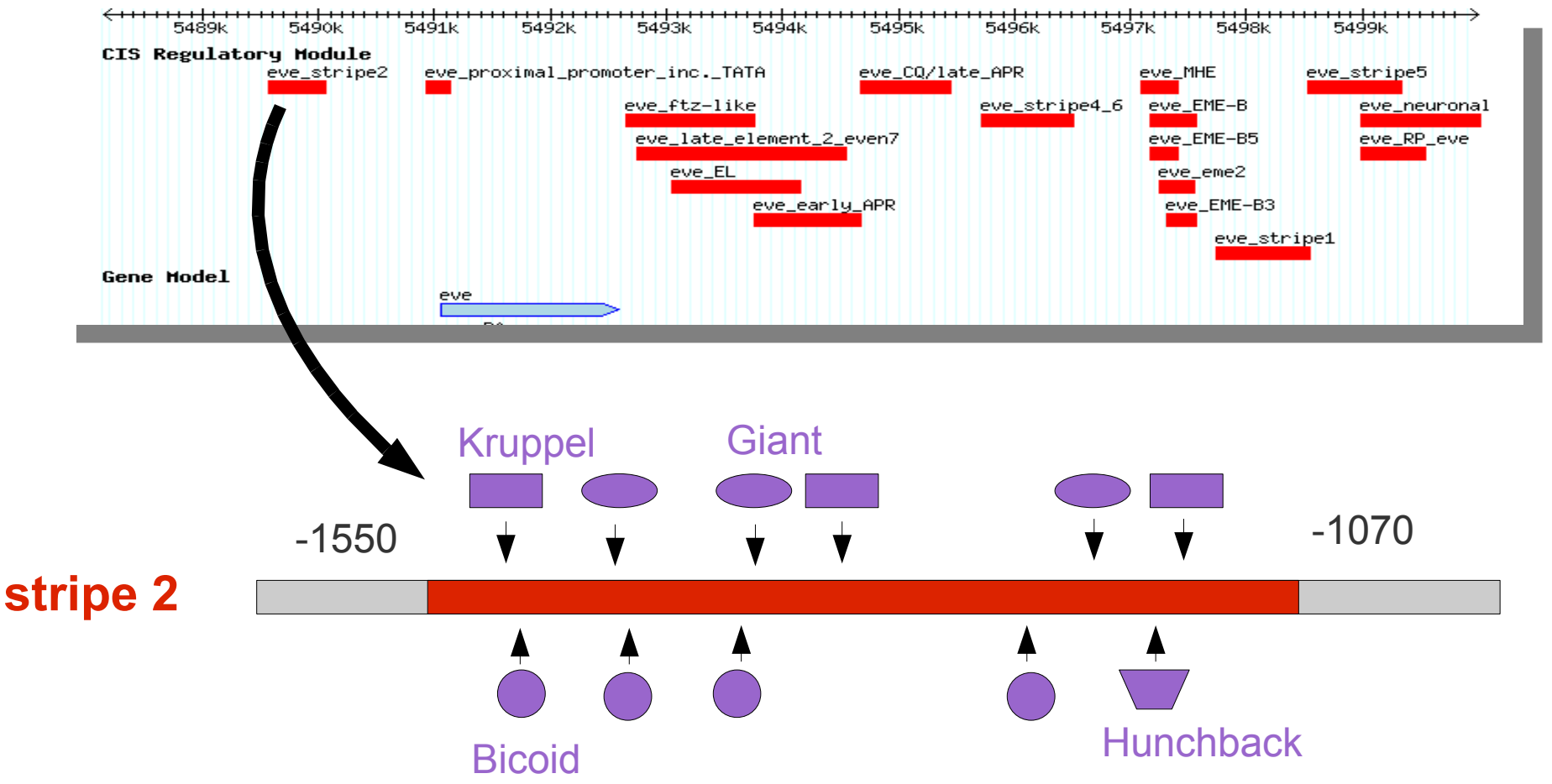
"junk" DNA ??

- ~ 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



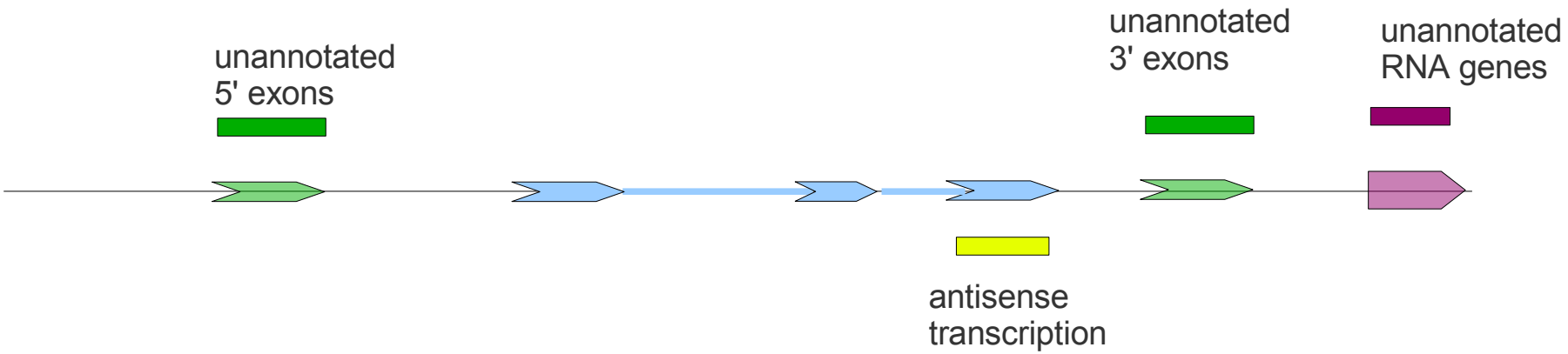
... but many functional elements

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

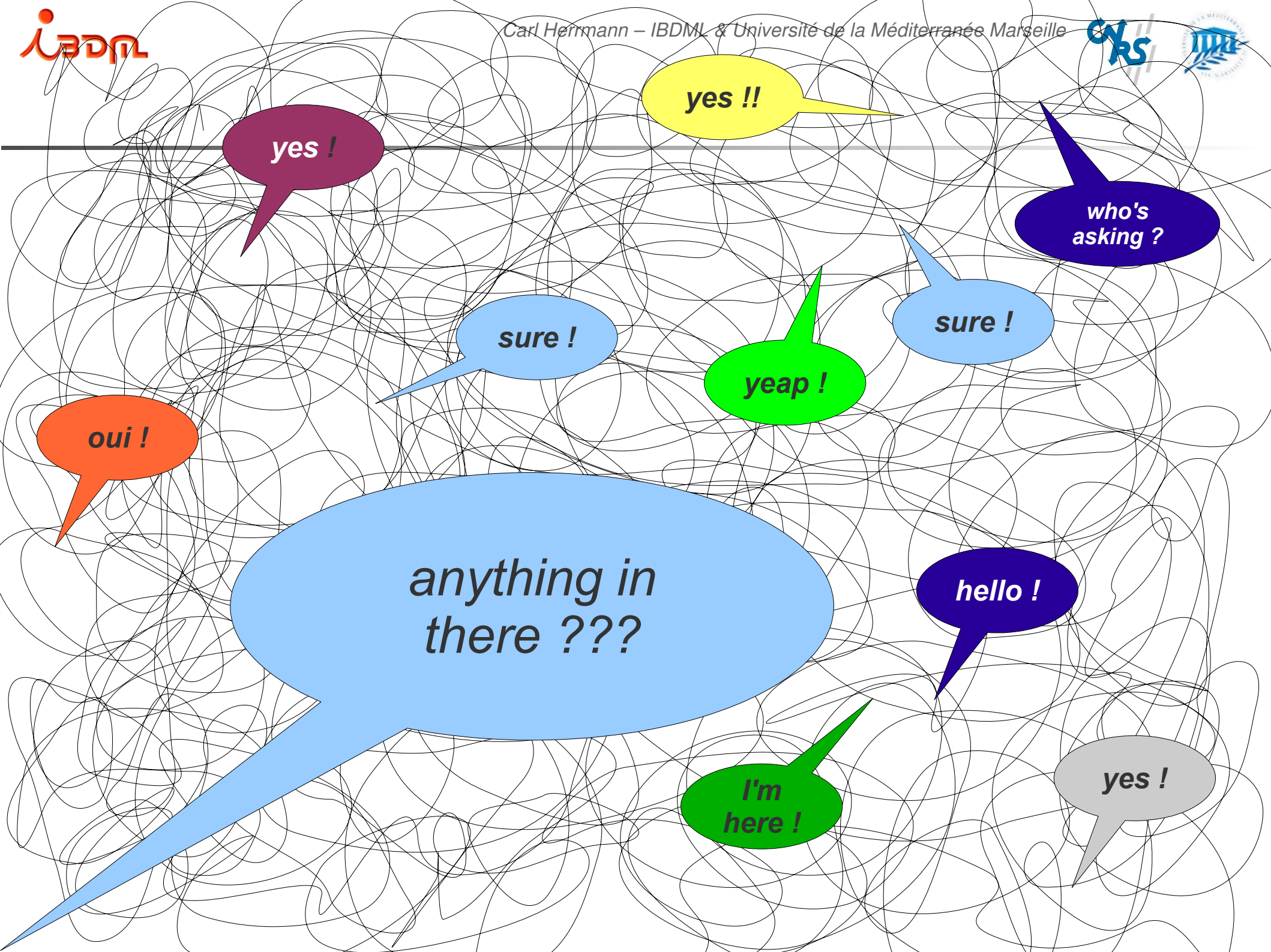
... and a huge level of transcription !!

- 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

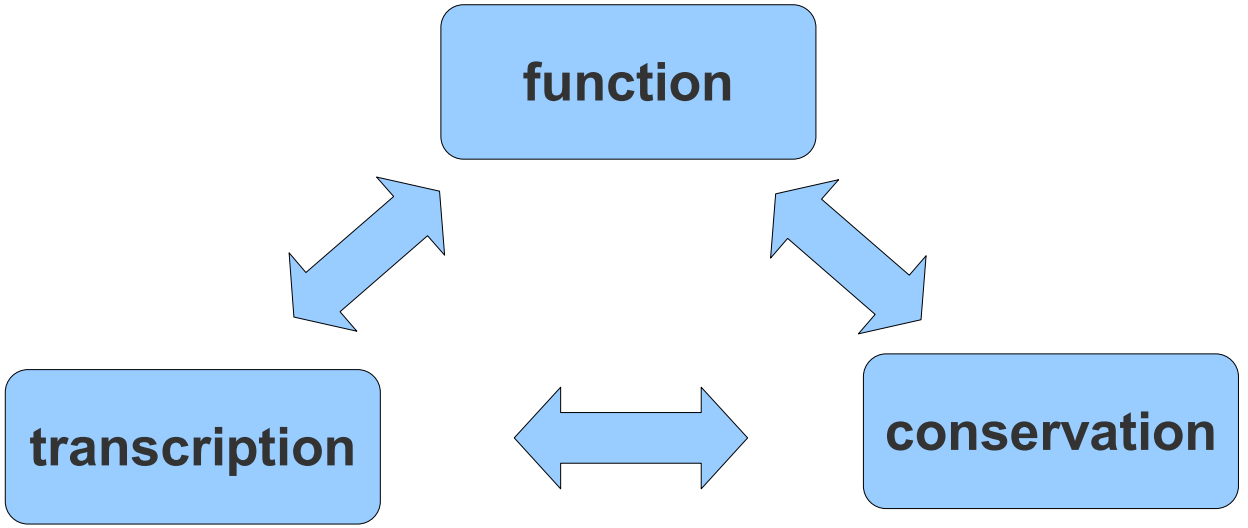


[cf. Willingham & Gingeras, Cell 125 (2006) 1215-1220]


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 ?



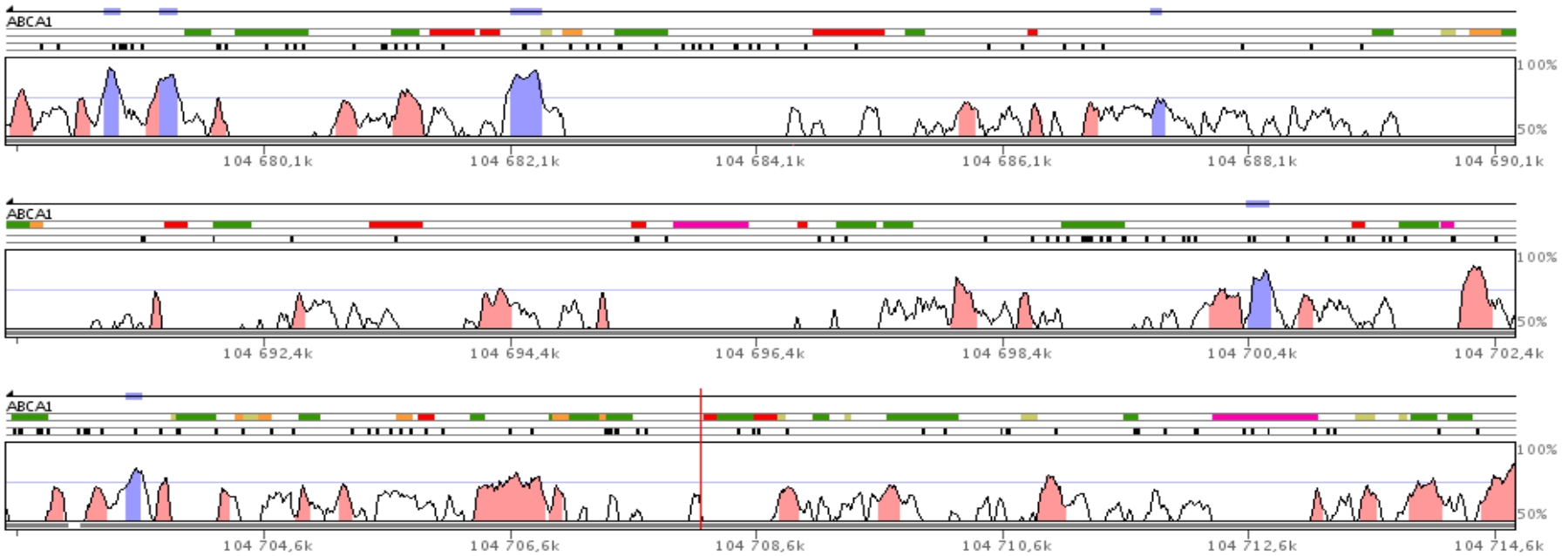
Comparative genomics :
what is under selective pressure ?

what type of conservation ?

- nucleotide conservation: conserve **primary** sequence

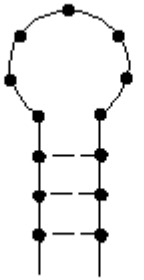
A C C G T T A C A T G G T A A
A G C G A T A C A T C T T A C

66% identity



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 ?

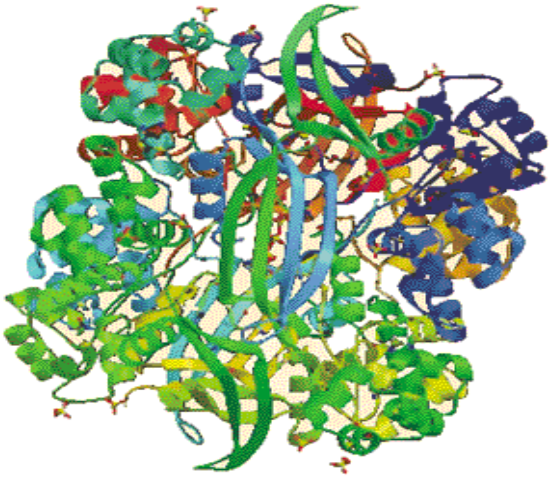
- codon conservation: conserve **protein sequence/structure**

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

93% nucl. id, 80% AA id



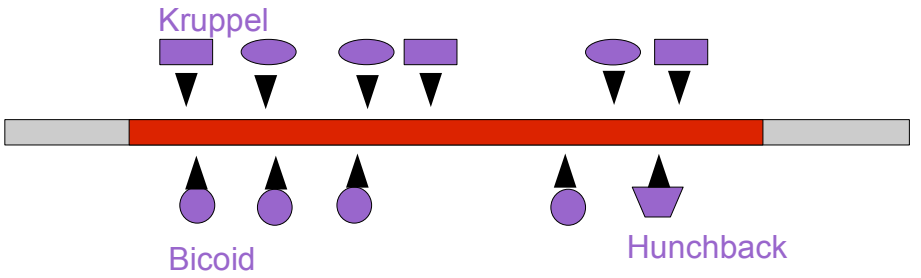
what type of conservation ?

different **constraints**

different definitions of conservation

what type of constraints for non-coding DNA ?

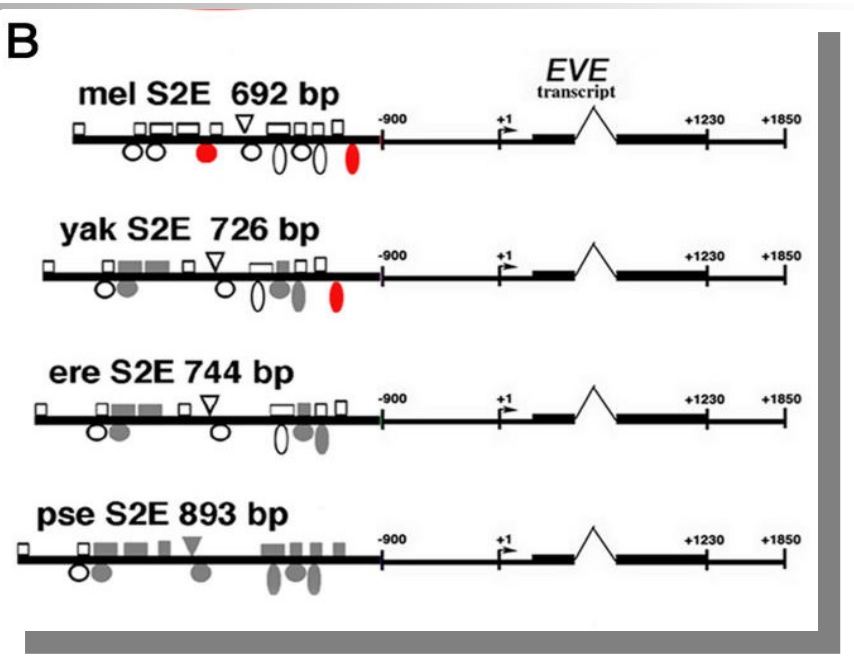
- example: **enhancers**
  - *conserve full sequence?*
  - *conserve sequence of BS alone ?*
  - *conserve number of binding sites ?*
  - *conserve order/spacing between binding sites ?*

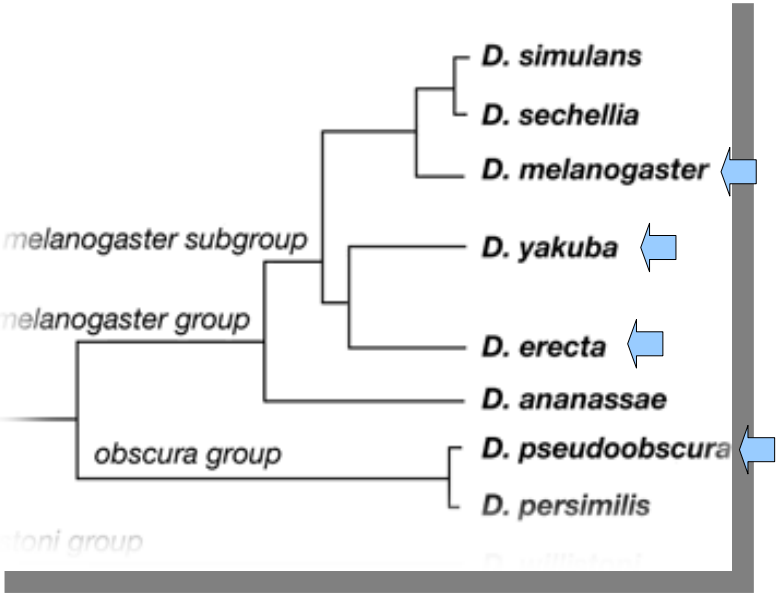


Enhancer evolution

- 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

BioMed Central

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"

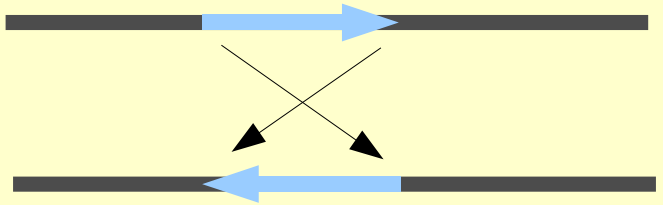
Article

Comparat
pseudoobscu
and *cis*-ele

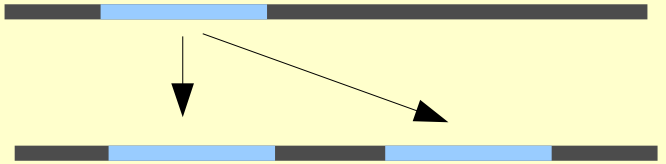
identical bases in 14, whereas the nearby sites, at ~66% identical, would be expected to have 9.24 identical bases in 14. That difference, though statistically significant, amounts to <1 bp of excess conservation per site. Such a slight difference in conservation would appear to offer scant hope of identifying 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

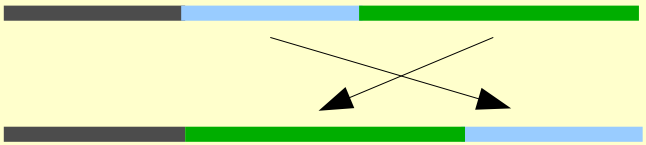
inversions



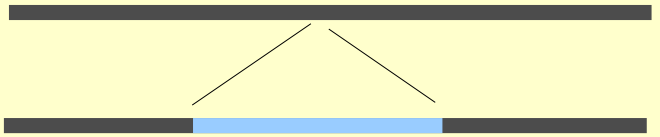
duplications



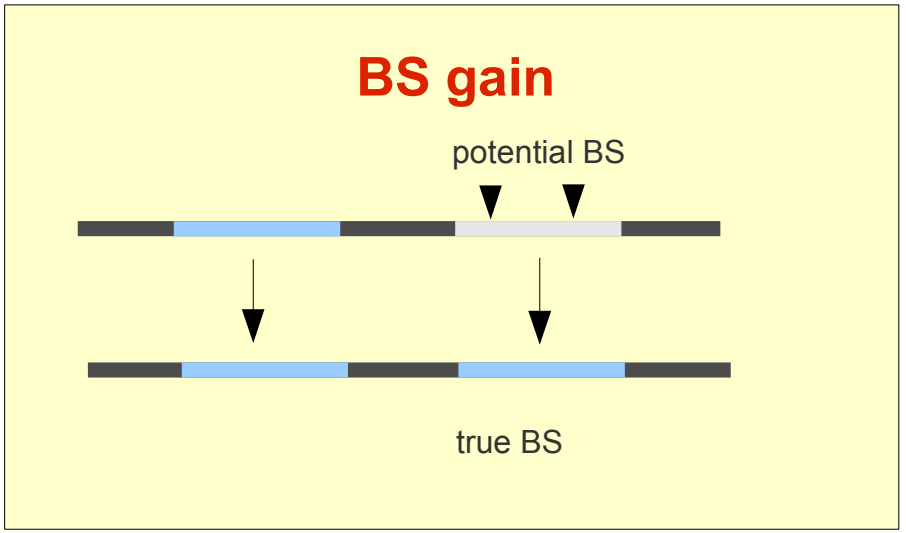
translocations

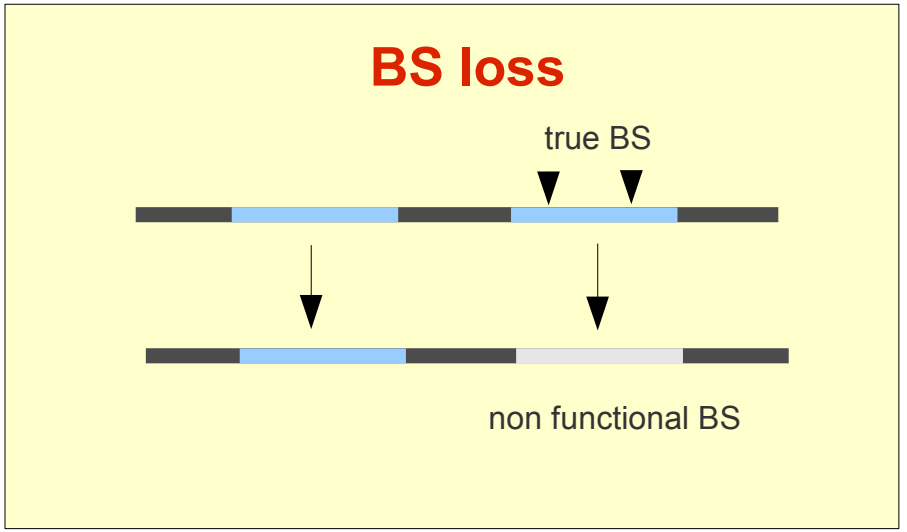


insertions



Binding site turnover

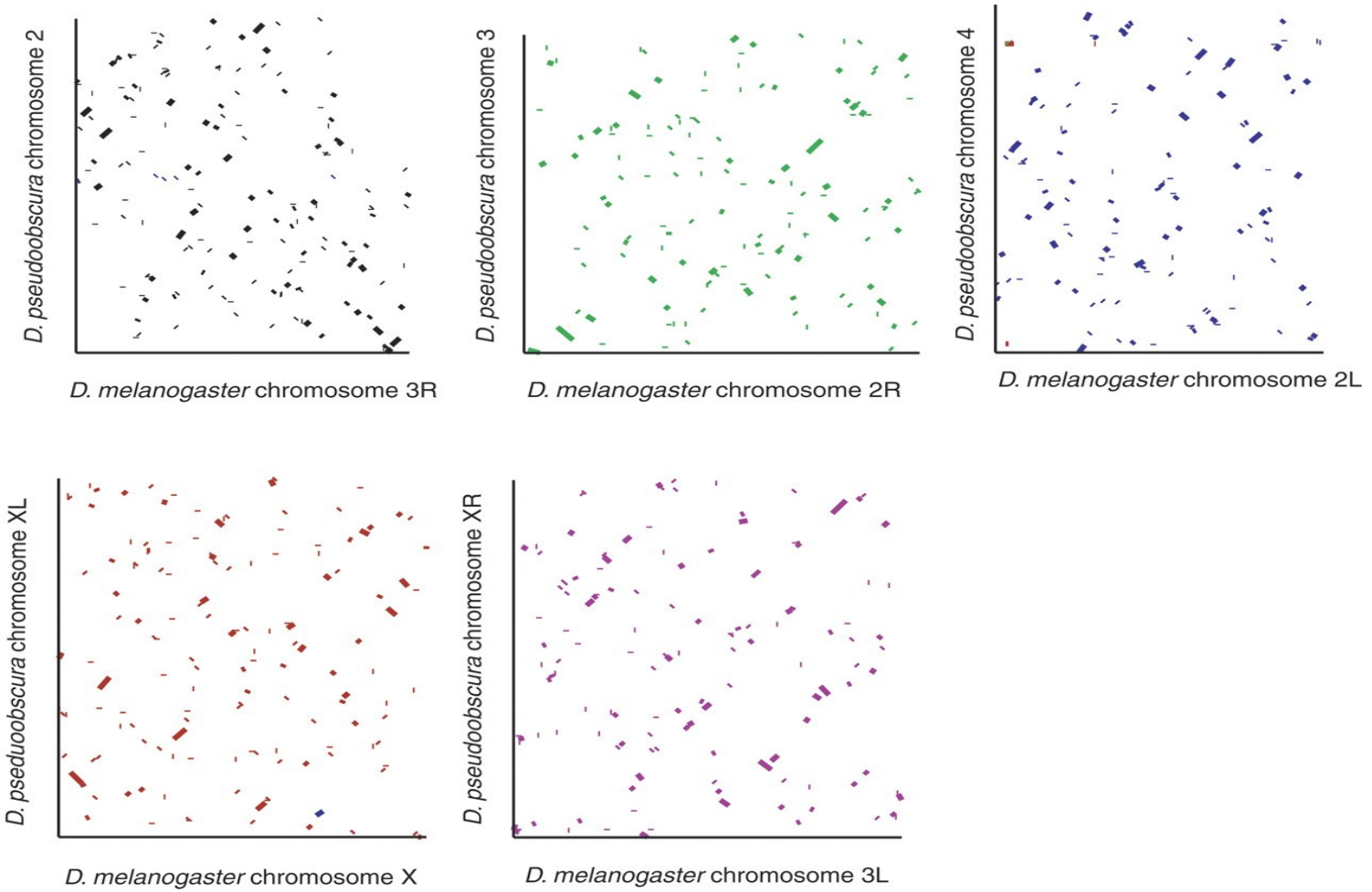




- study of *Zeste* BS turnover in drosophila [Moses et al, 2006]

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

D. pseudoobscura vs. D. melanogaster



Richards et al, Gen. Res. (2005)

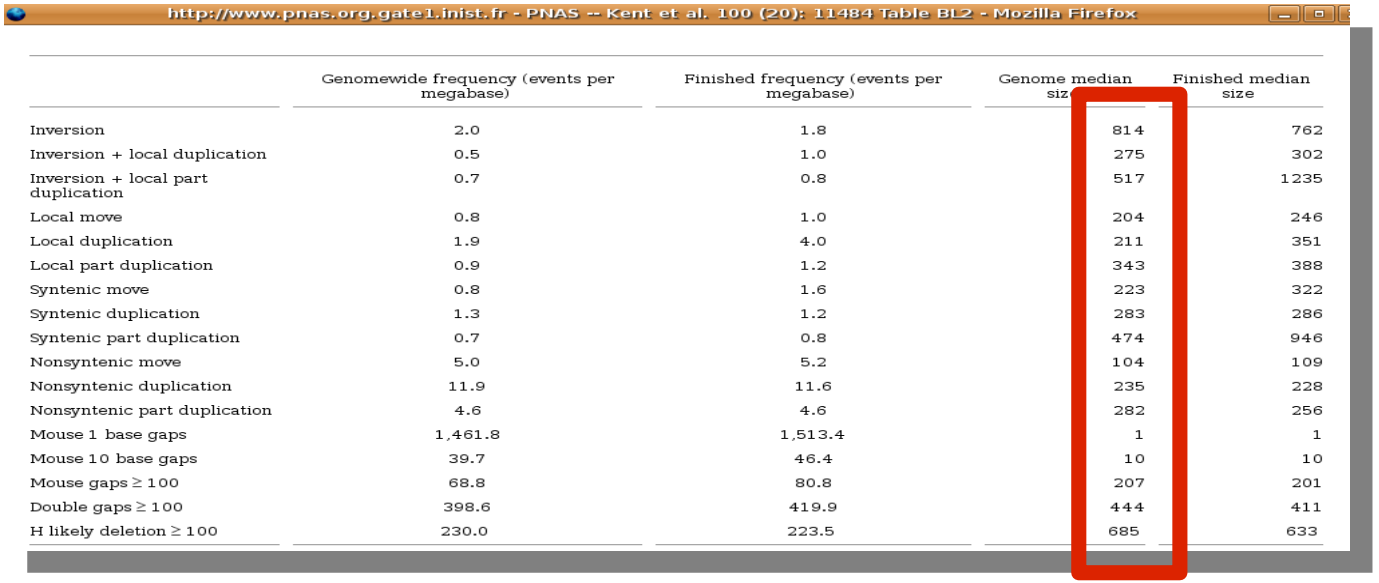
"evolution's cauldron" (Kent et al.)

something has happened here ...

different scales involved here !

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

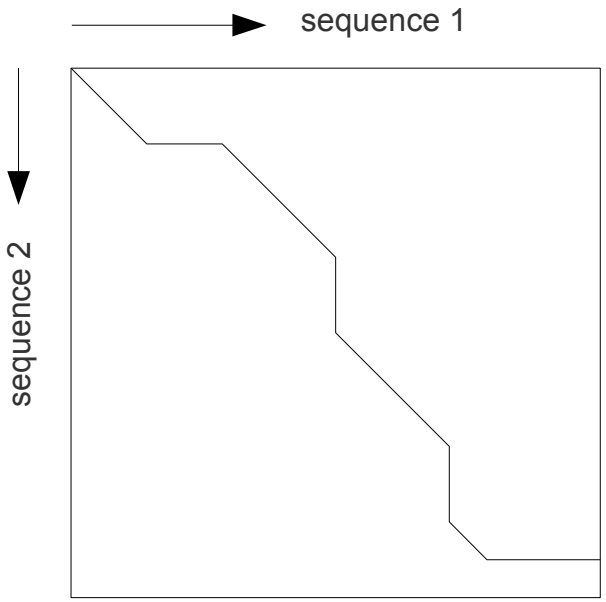
Human-mouse analysis [Kent et al, PNAS 2003]



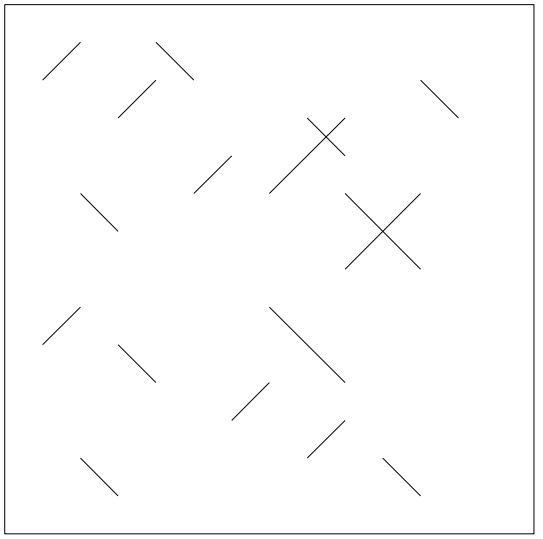
	Genomewide frequency (events per megabase)	Finished frequency (events per megabase)	Genome median siz	Finished median size
Inversion	2.0	1.8	814	762
Inversion + local duplication	0.5	1.0	275	302
Inversion + local part duplication	0.7	0.8	517	1235
Local move	0.8	1.0	204	246
Local duplication	1.9	4.0	211	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 ≥ 100	68.8	80.8	207	201
Double gaps ≥ 100	398.6	419.9	444	411
H likely deletion ≥ 100	230.0	223.5	685	633

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

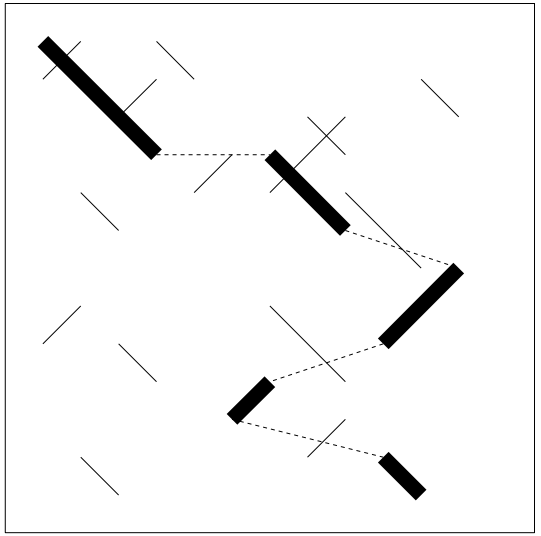
comparing sequences



global



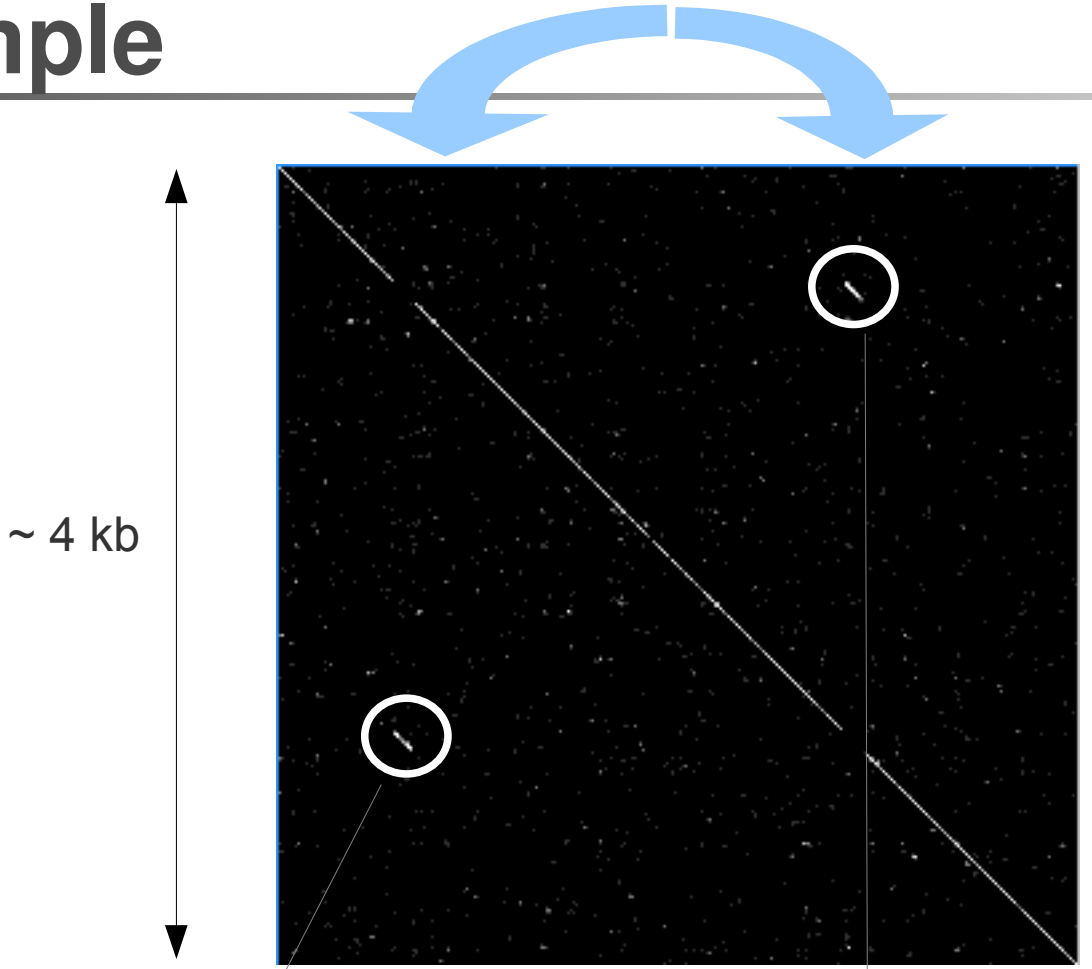
local



"glocal"
(shuffle LAGAN)

do they see these genomic events ?

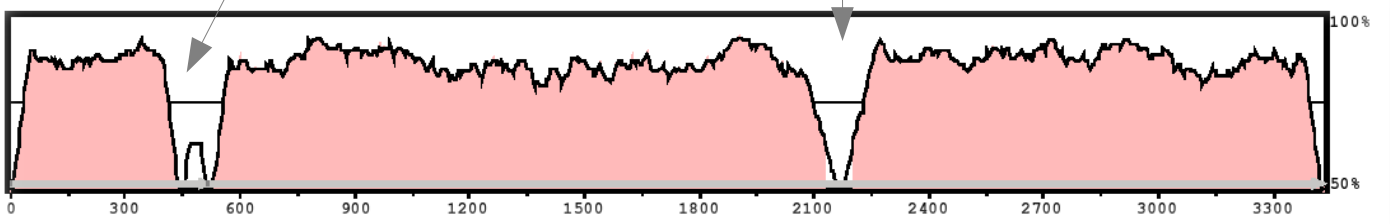
example



simulated translocation event (~ 300 bp)

alignment with shuffle LAGAN (glocal alignment)

conservation plot (VISTA)



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

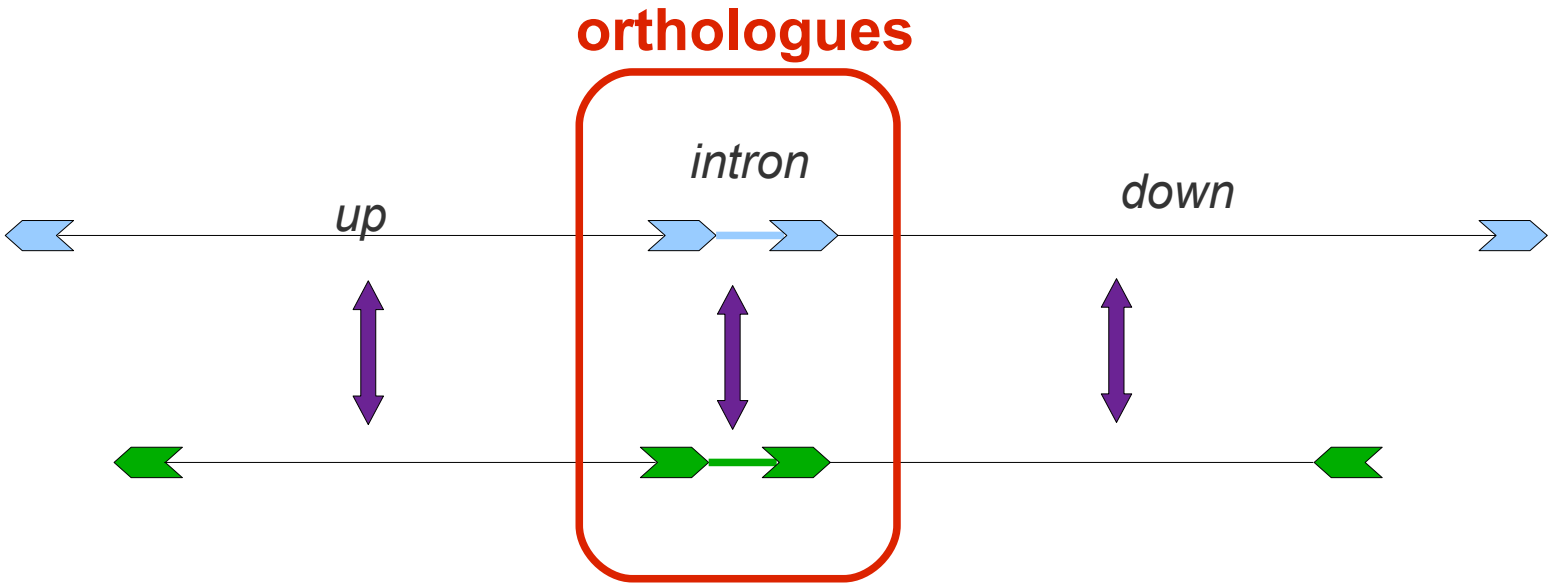
1) identify conserved non-coding blocks in Drosophila
2) make inventory of small scale rearrangement events in non-coding DNA
3) study evolution of known functional elements
(e.g. enhancers)
"phylogeny" of non-coding DNA ?

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)

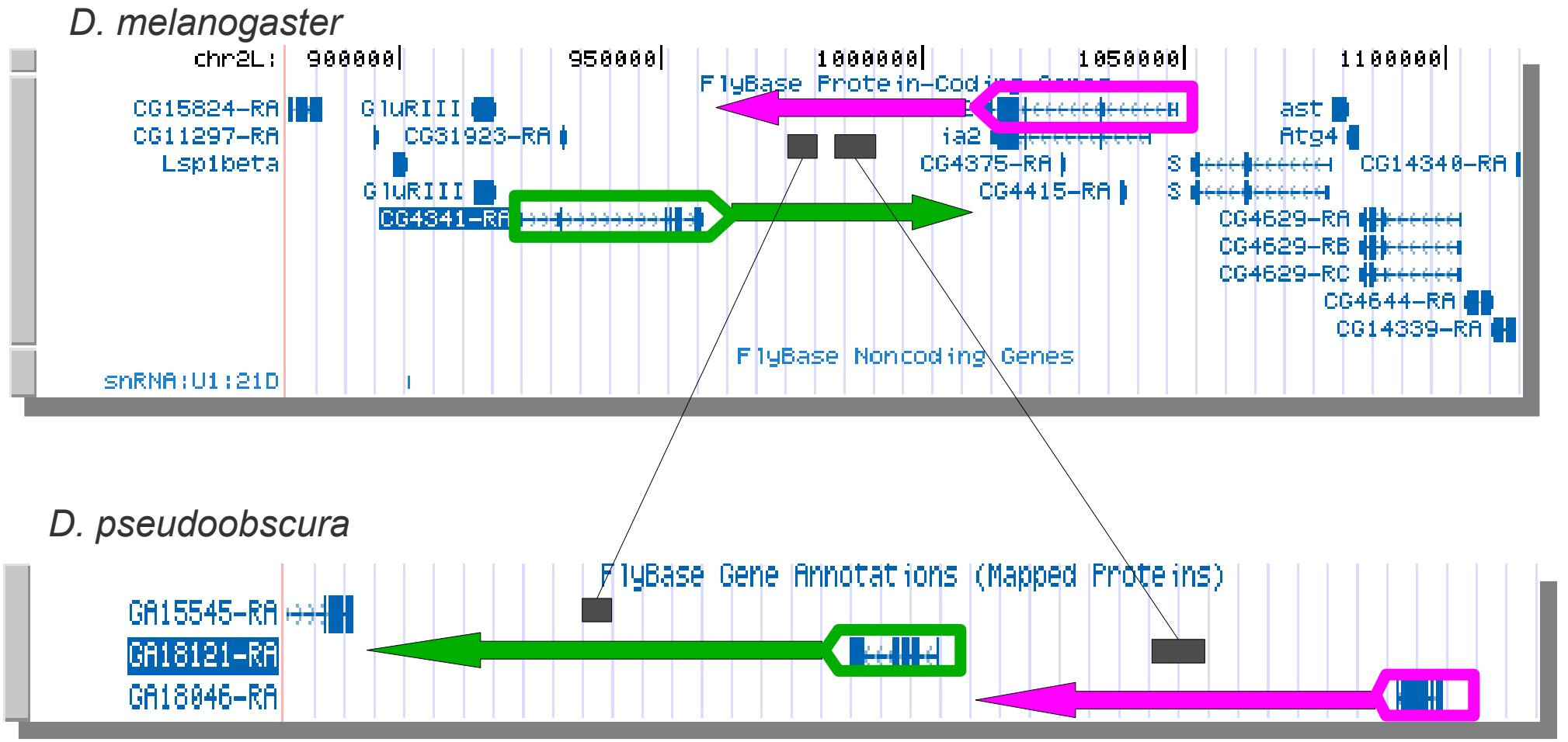
align orthologous regions



- 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)

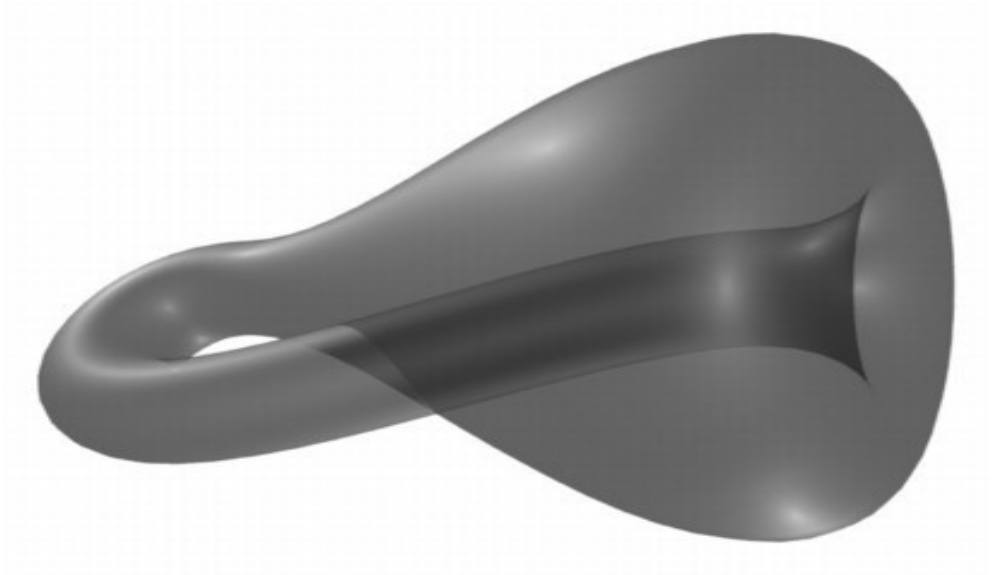
cf. AAA Drosophila comparative genomics site [http://rana.lbl.gov/drosophila/]

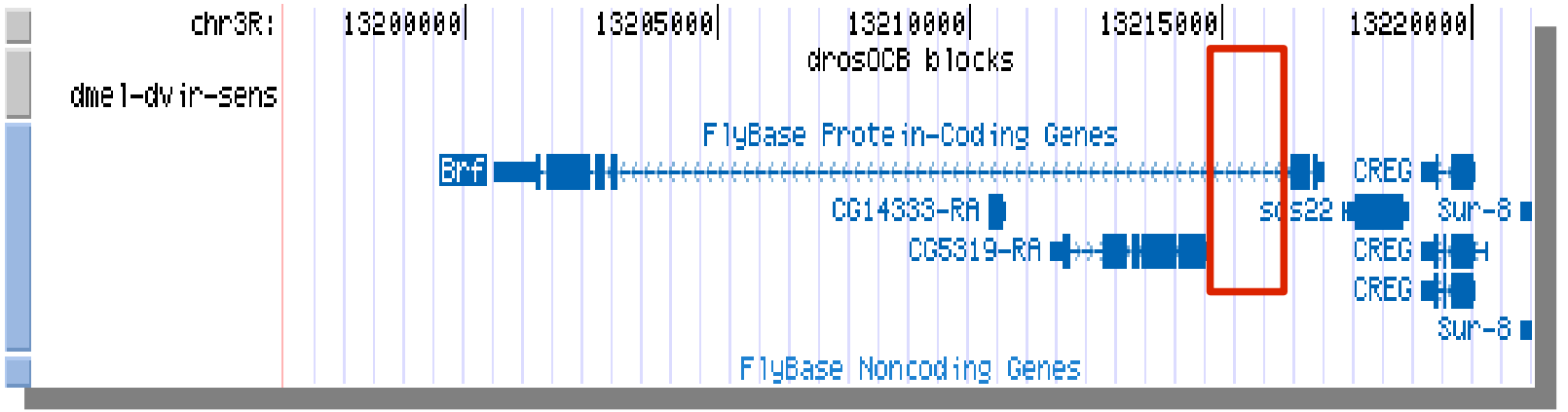
"genome tiling alignment"



blocks are attributed to one or several genes

where is in, where is out ?





flowchart

mask genomes

extract up/down/intron regions

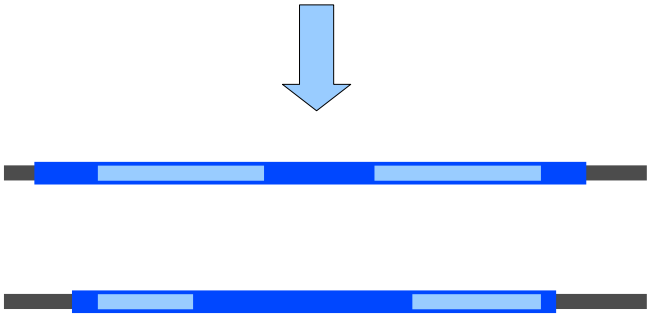
align orthologous regions of orthologous genes

merge blocks which overlapp in *D.mel*. and *D.xxx*

CHAOS

"double seed"

L = 7 (-1) : sensitive version
L = 10(-1) : default version



[Chaos: Brudno et al., BMC Bioinfo]

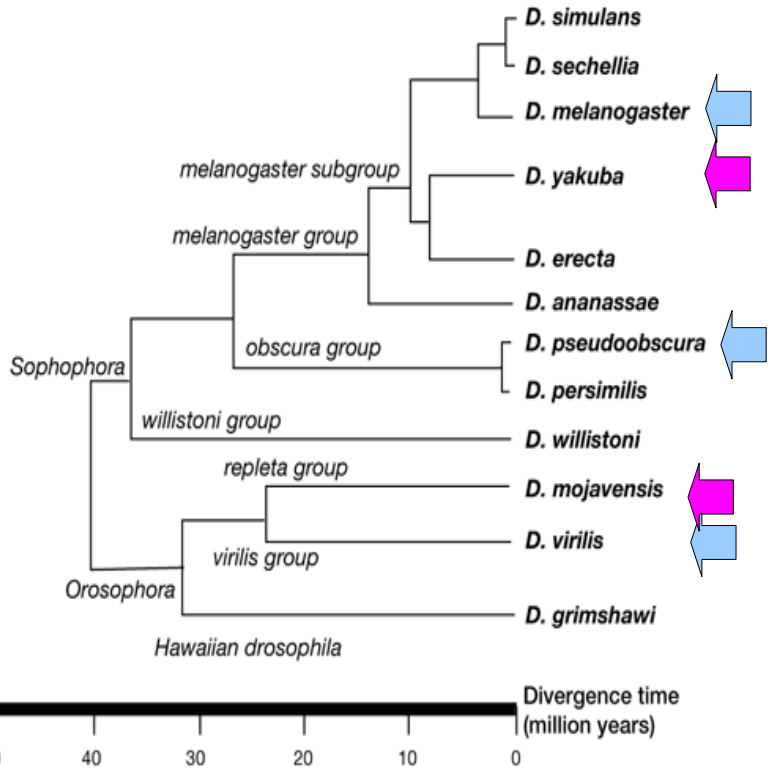
Some statistics

D.melanogaster vs. D.pseudoobscura

	Intergenic		Intronic	
Region aligned	**55.5 Mb**		**38.5 Mb**	
	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	
Region aligned	**48.1 Mb**		**34.7 Mb**	
	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

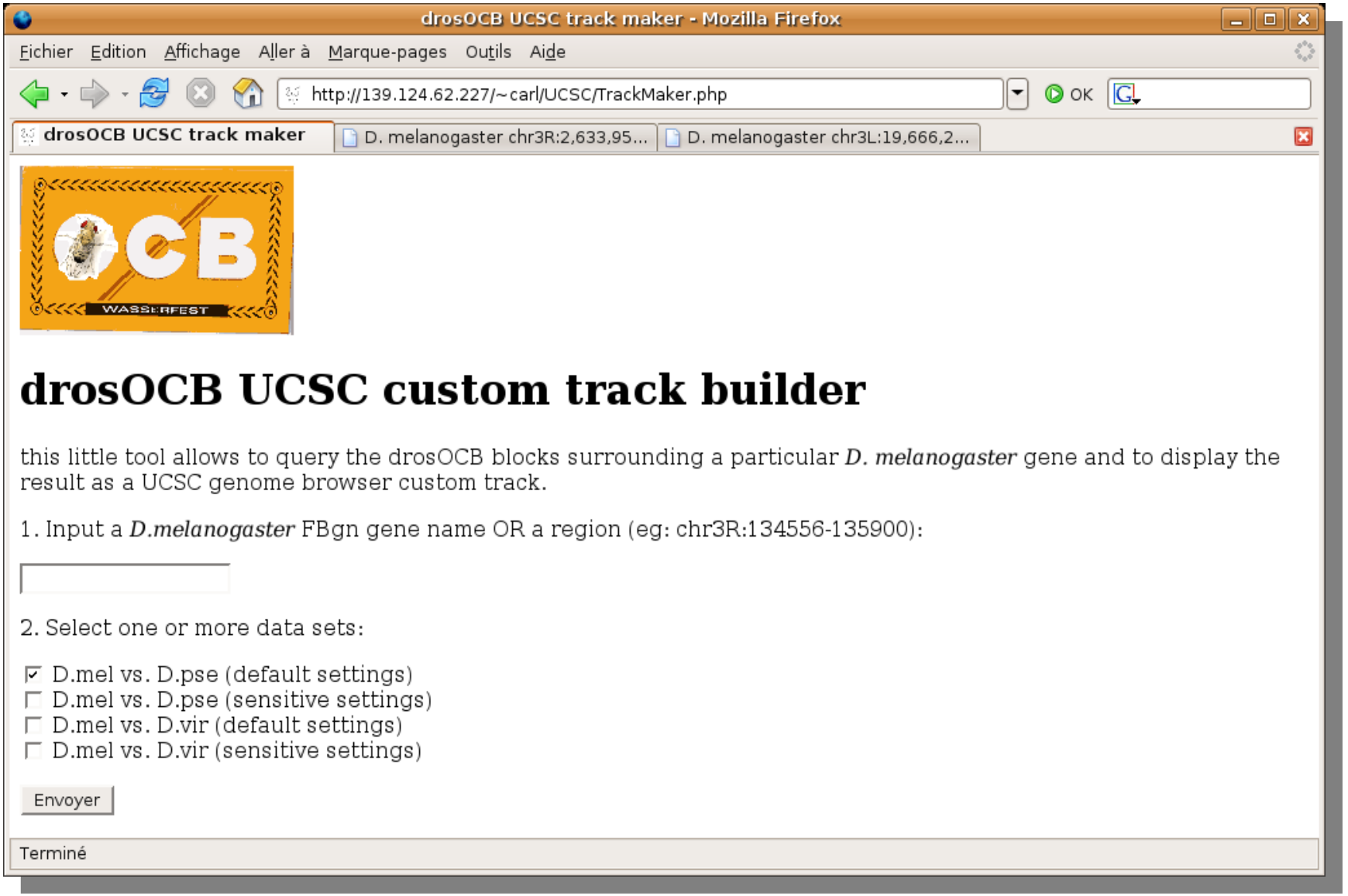
local
(BLAST, CHAOS,...)

lobal©
(drosOCB)

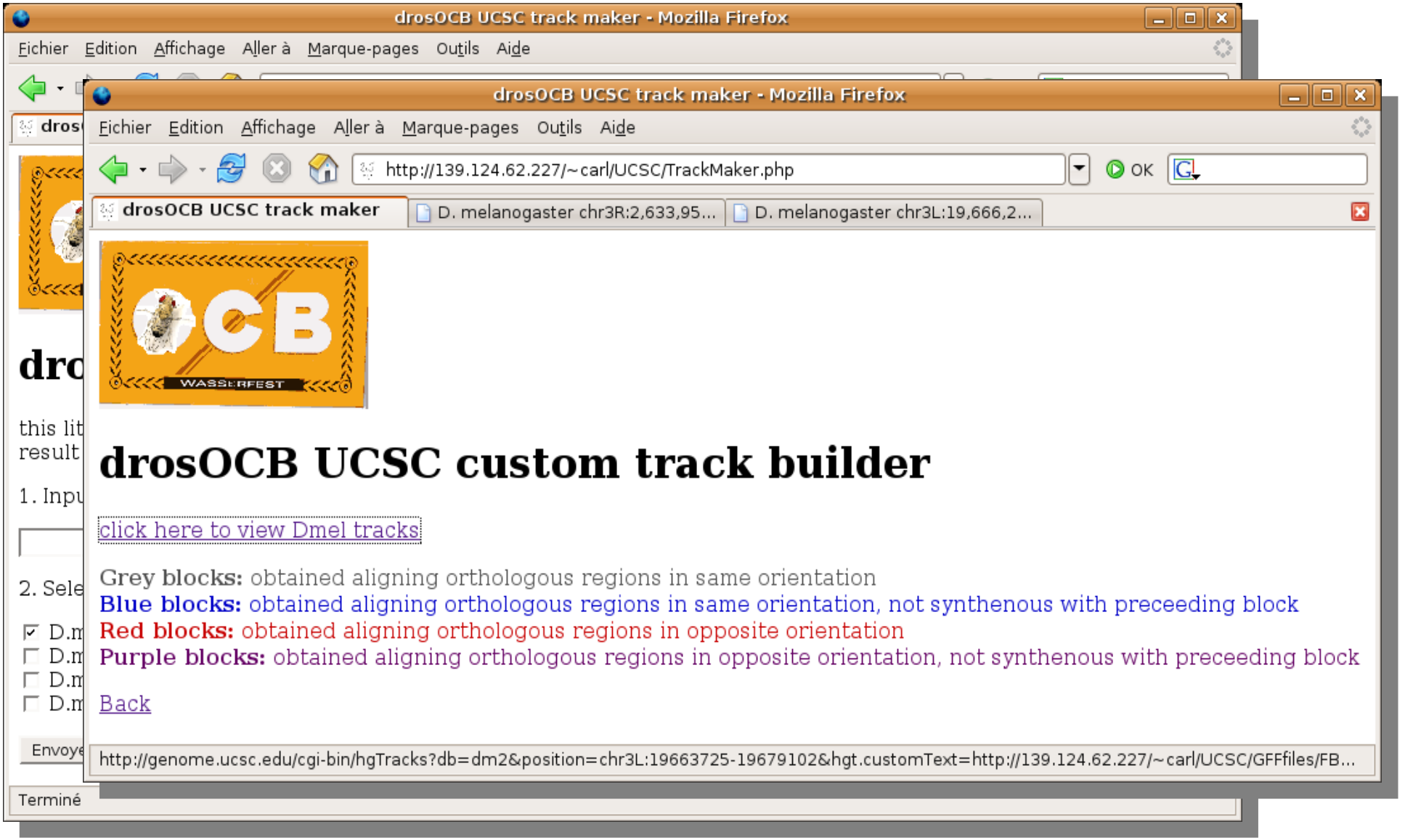
glocal
(shuffleLAGAN...)

global

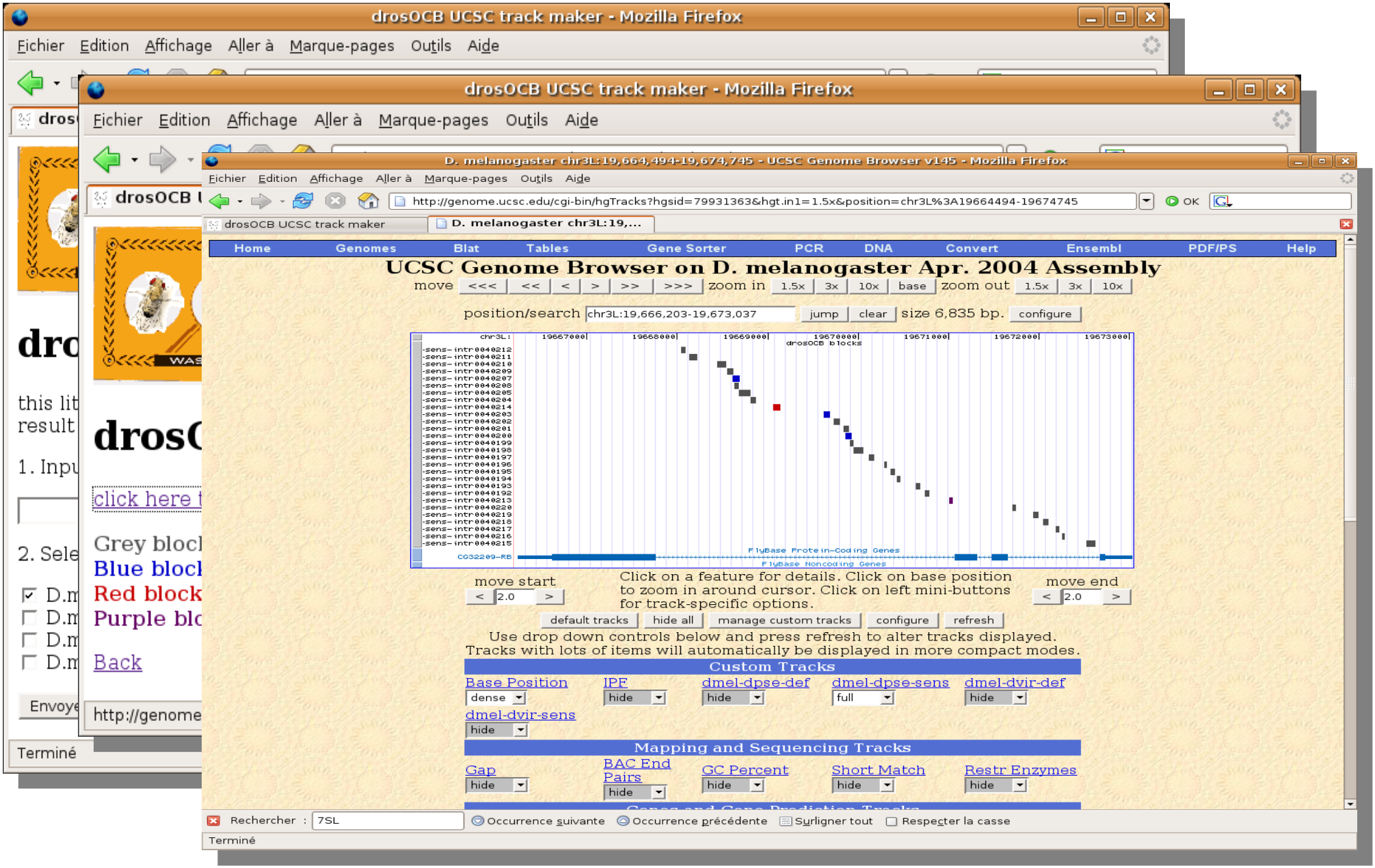
"database" "interface"



"database" "interface"

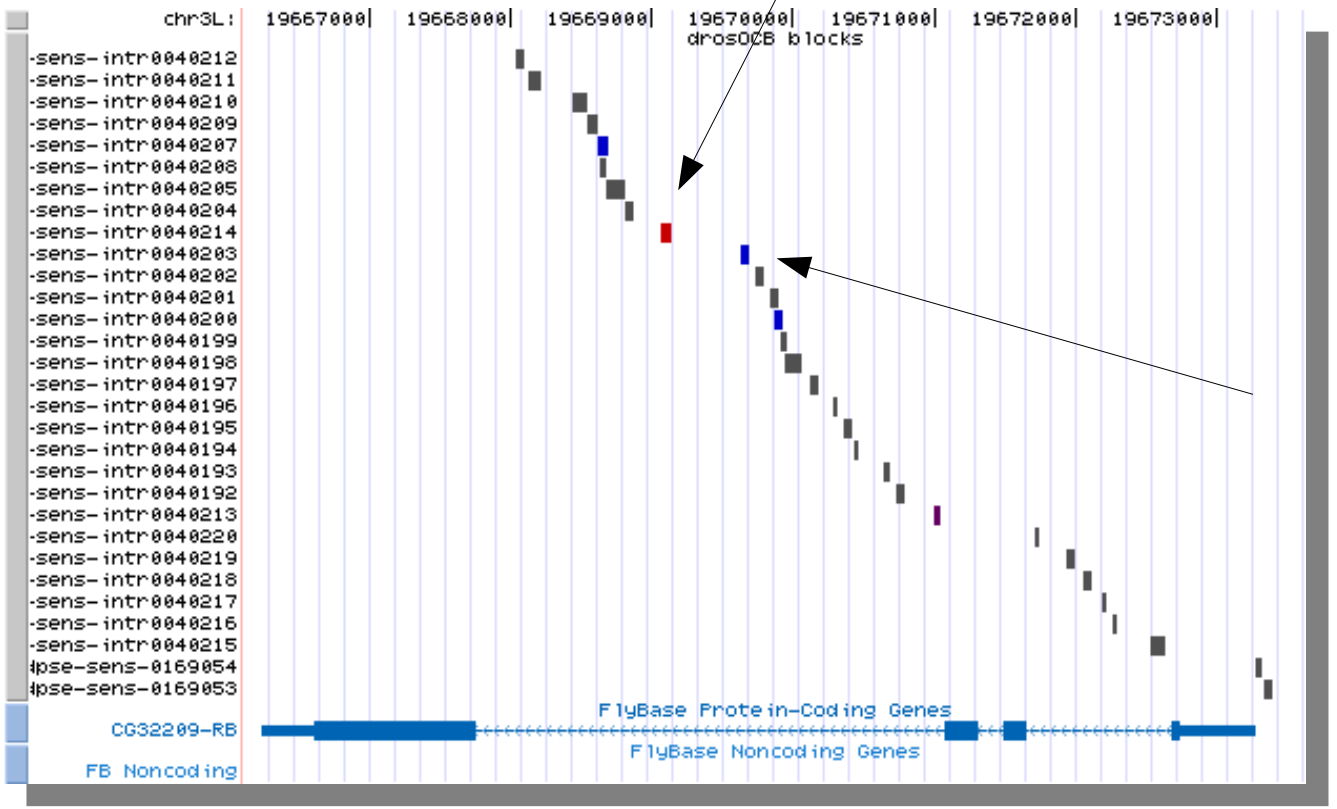


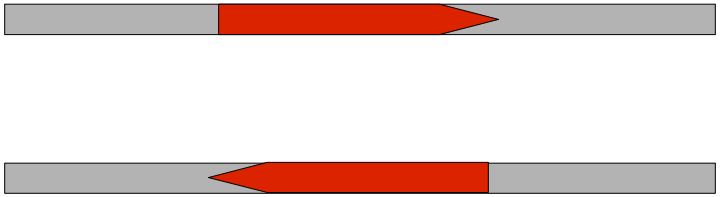
"database" "interface"



UCSC genome browser custom track

RED : inverted block *(inversion ?)*





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

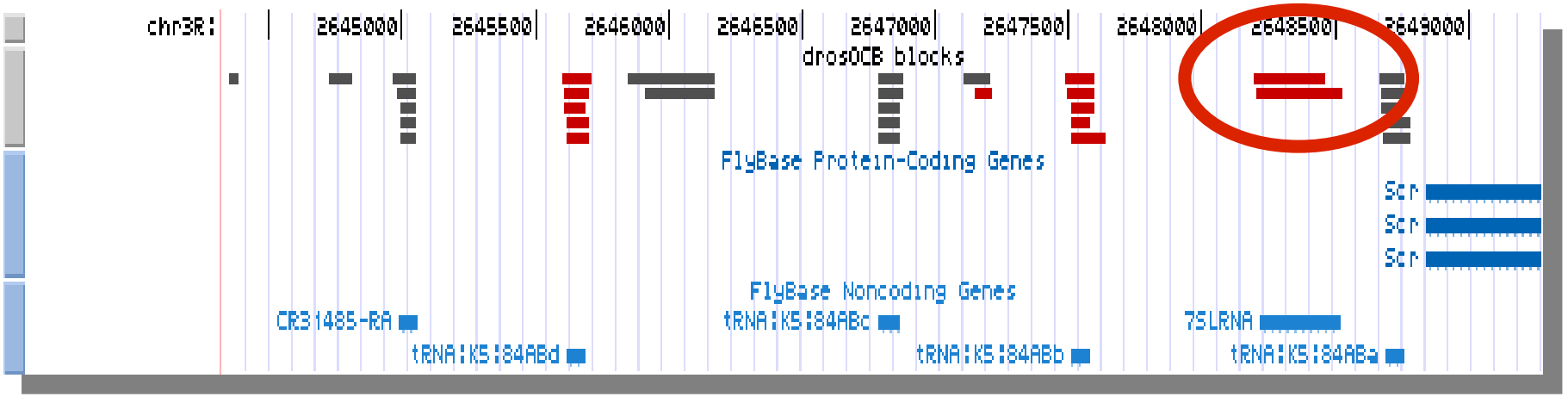
Many putative inversion events

- 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 !

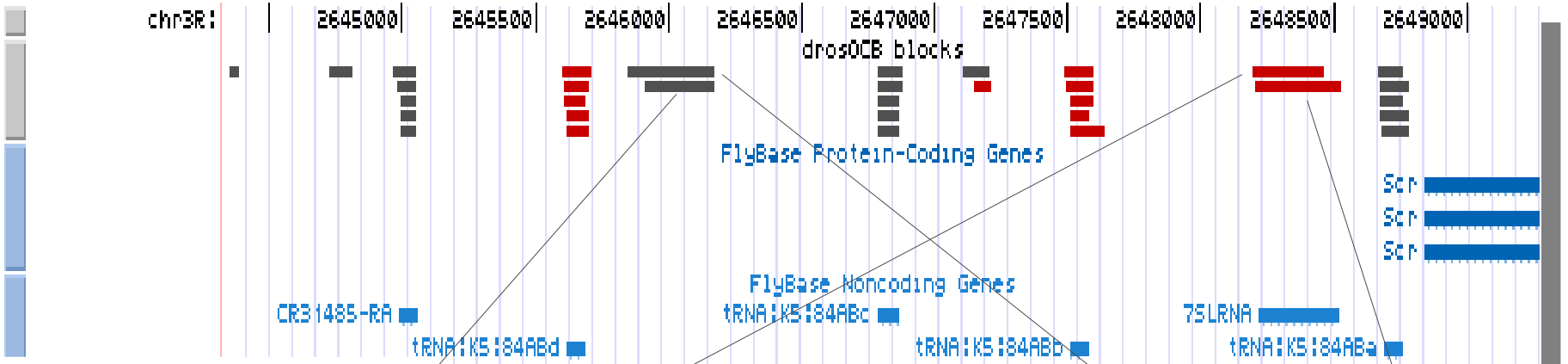
D. melanogaster genome



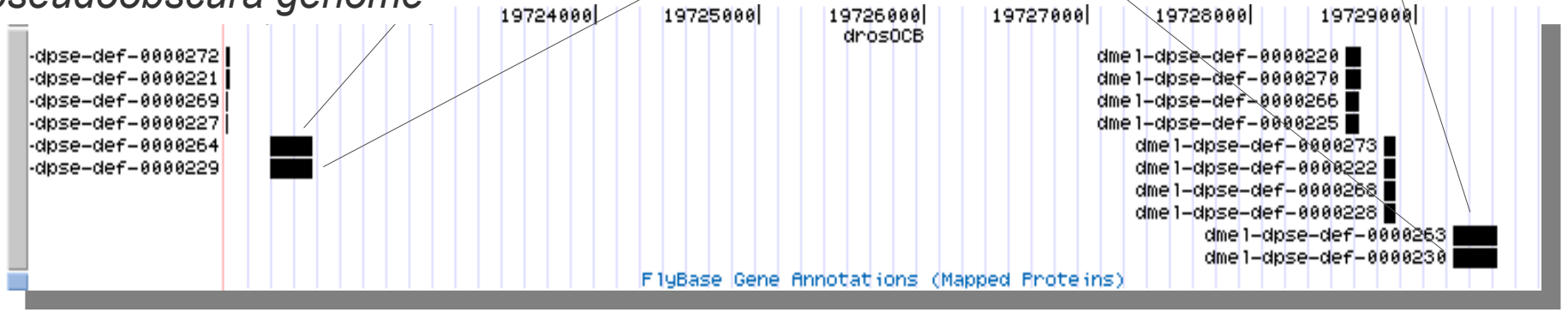
overlapping, but not merged:
map to **distinct** locations in the *D.pse* genome

Duplication of non-coding RNA genes

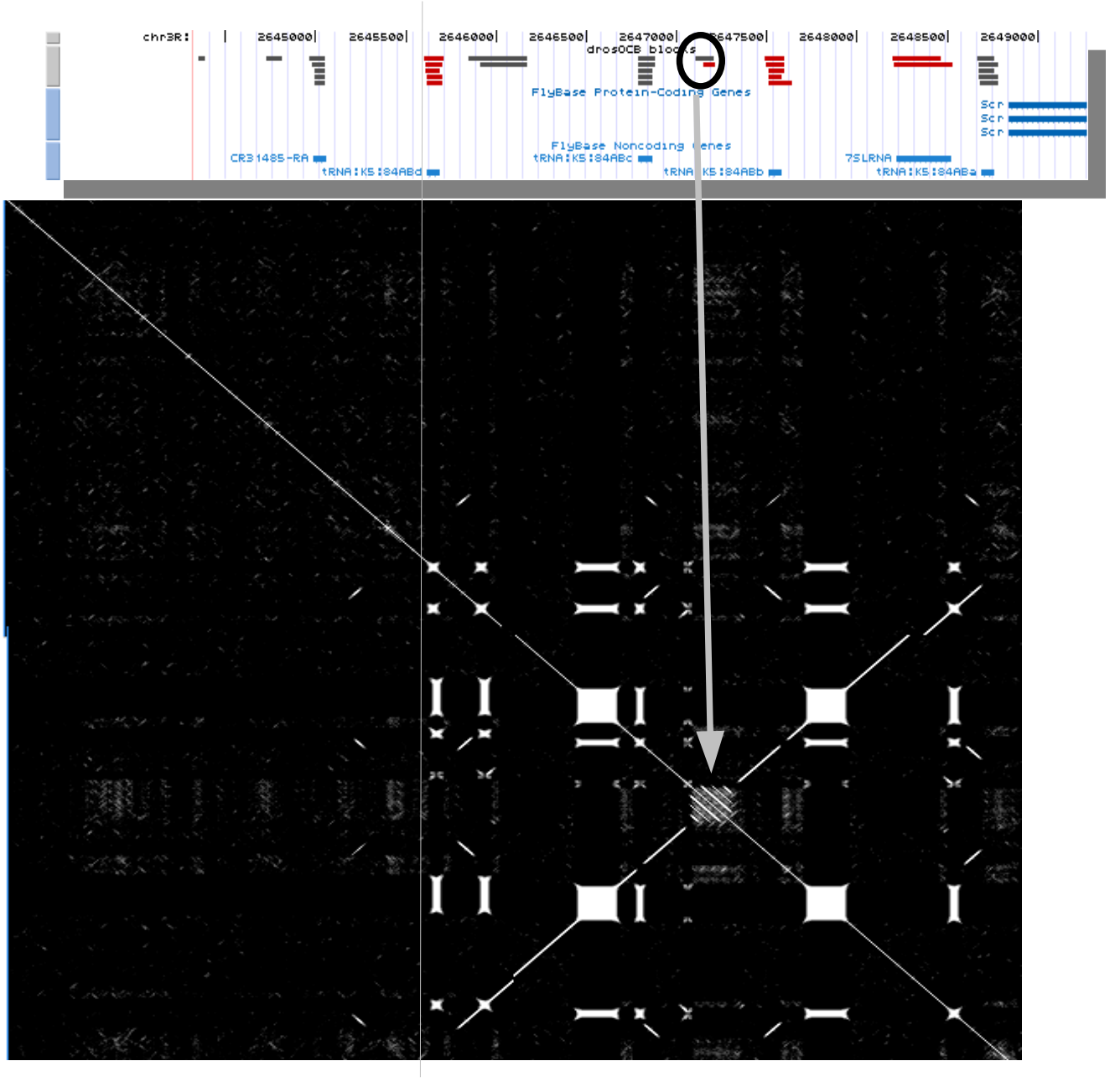
D. melanogaster genome



D. pseudoobscura genome

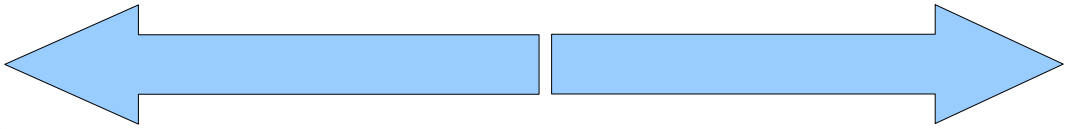


dotplot

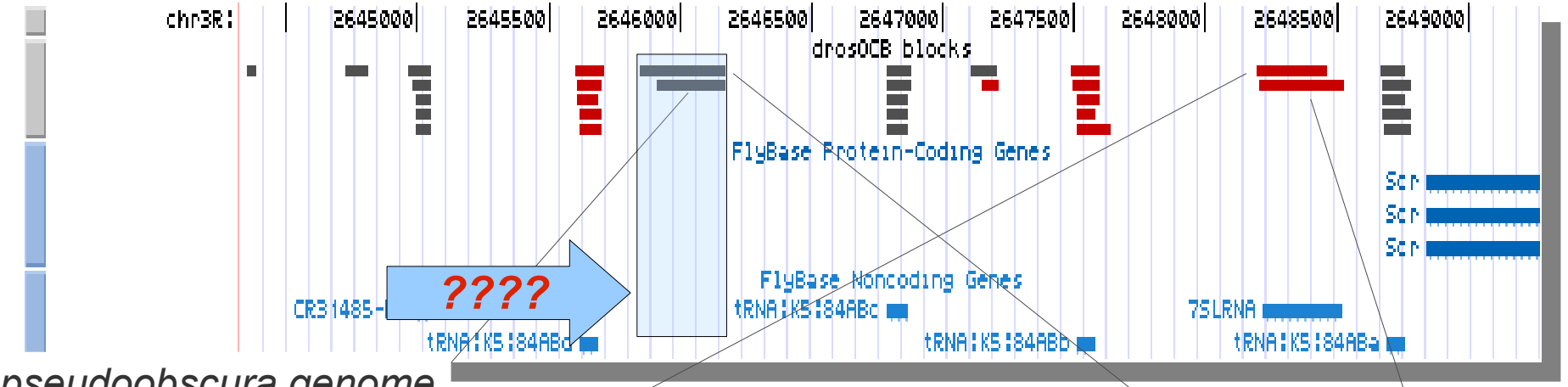

chr3R:
2645000
2645500
2646000
2646500
2647000
2647500
2648000
2648500
2649000
FlyBase Protein-Coding Genes
FlyBase Noncoding Genes
CR31485-RA
tRNA:K5:84ABc
7SLRNA
tRNA:K5:84ABa

Duplication of non-coding RNA genes

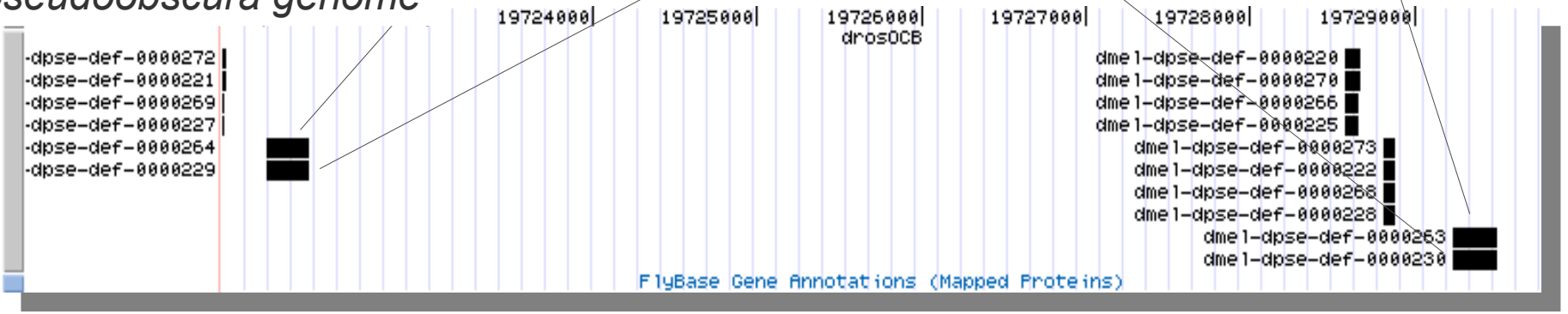
duplicated inverted region



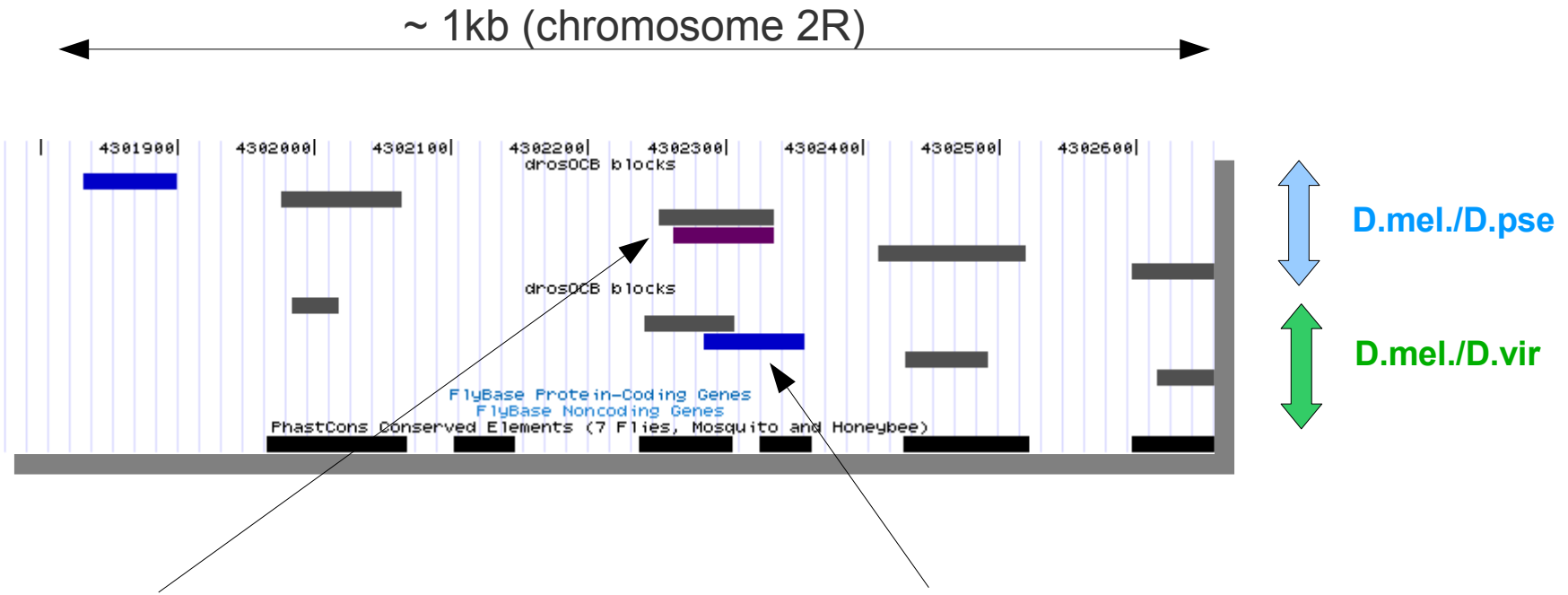
D. melanogaster genome



D. pseudoobscura genome



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

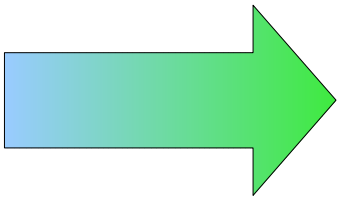
- starting point for phylogenetic footprinting
  application to discovery of cardiac enhancers in Drosophila
- systematic study of small scale genomic rearrangement events
  coding vs. non-coding fate
- "phylogeny" of non-coding DNA:
  lineage specific elements/events ?

General conclusions

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

"Actually, He didn't".

protein centric view



ncDNA/RNA centric view

the upcoming (r)evolution ?