Gene and genome duplication

CUSO/SIB Course Next Generation Comparative Phylogenomics 2014

Marc Robinson-Rechavi

http://bioinfo.unil.ch/
marc.robinson-rechavi@unil.ch
@marc_rr

UNIL | Université de Lausanne
Département d'écologie
et évolution

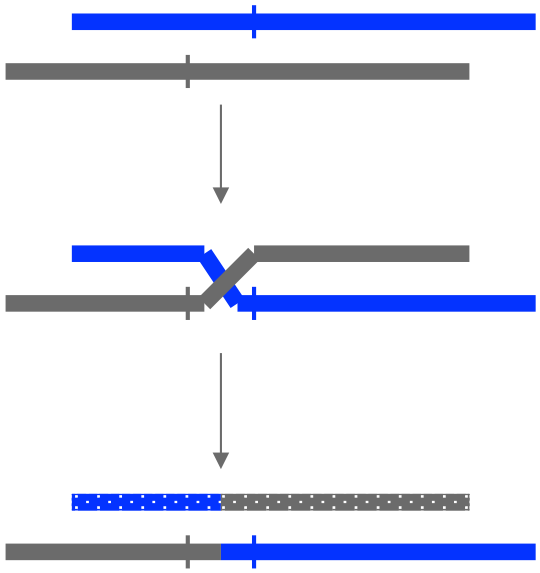
SIB
Swiss Institute of
Bioinformatics

Mechanisms of duplication:

- Polyploidy
- Aneuploidy
- Duplicative transposition
- Local tandem duplication

whole genome duplication

small-scale duplication, or gene duplication



deleated chromosome

chromosome with tandem duplication

Detecting duplications

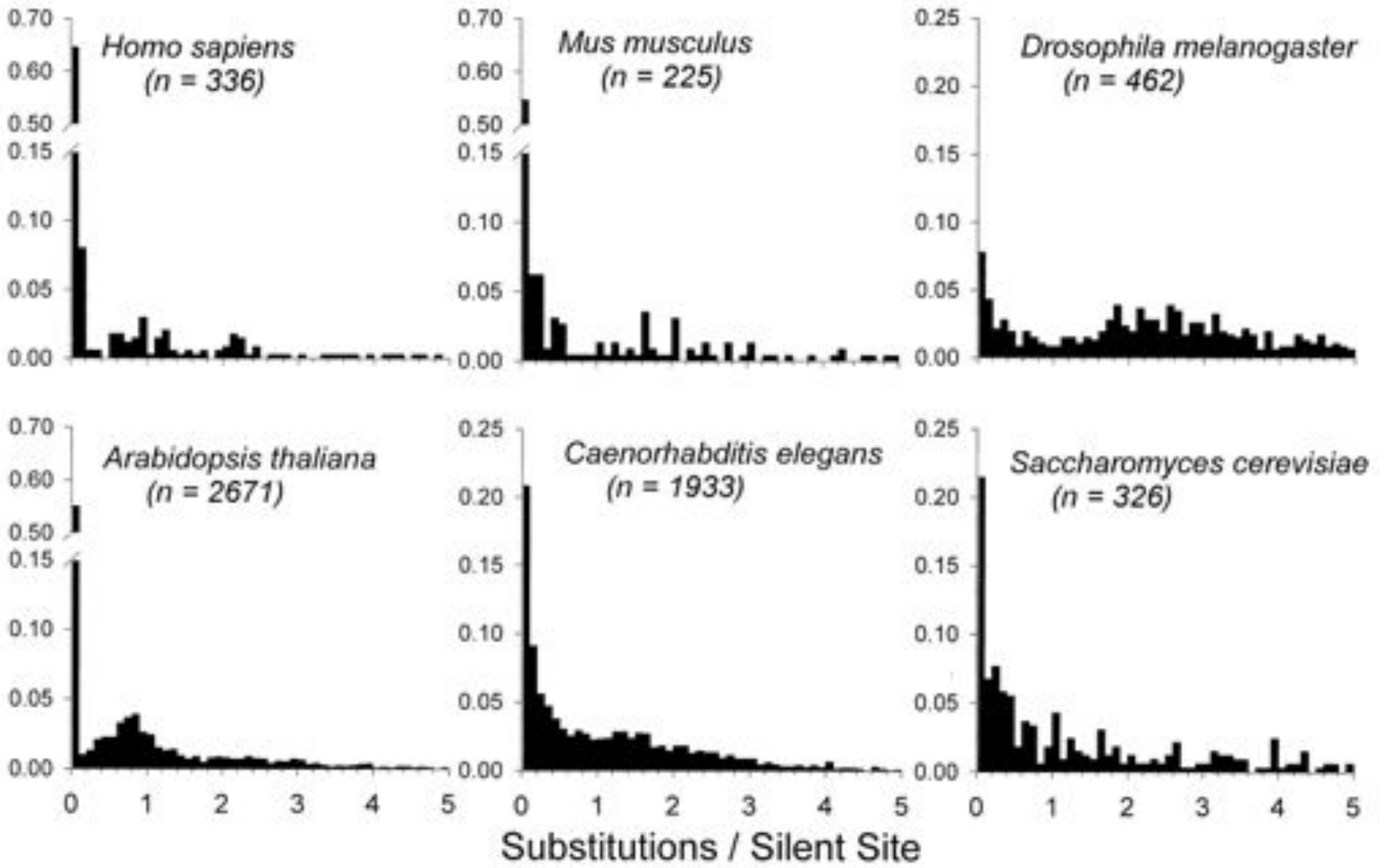
UNIL | Université de Lausanne
Département d'écologie et évolution

SIB
Swiss Institute of Bioinformatics

Discuss with your neighbor

- **What is more frequent: point mutations or duplications?**

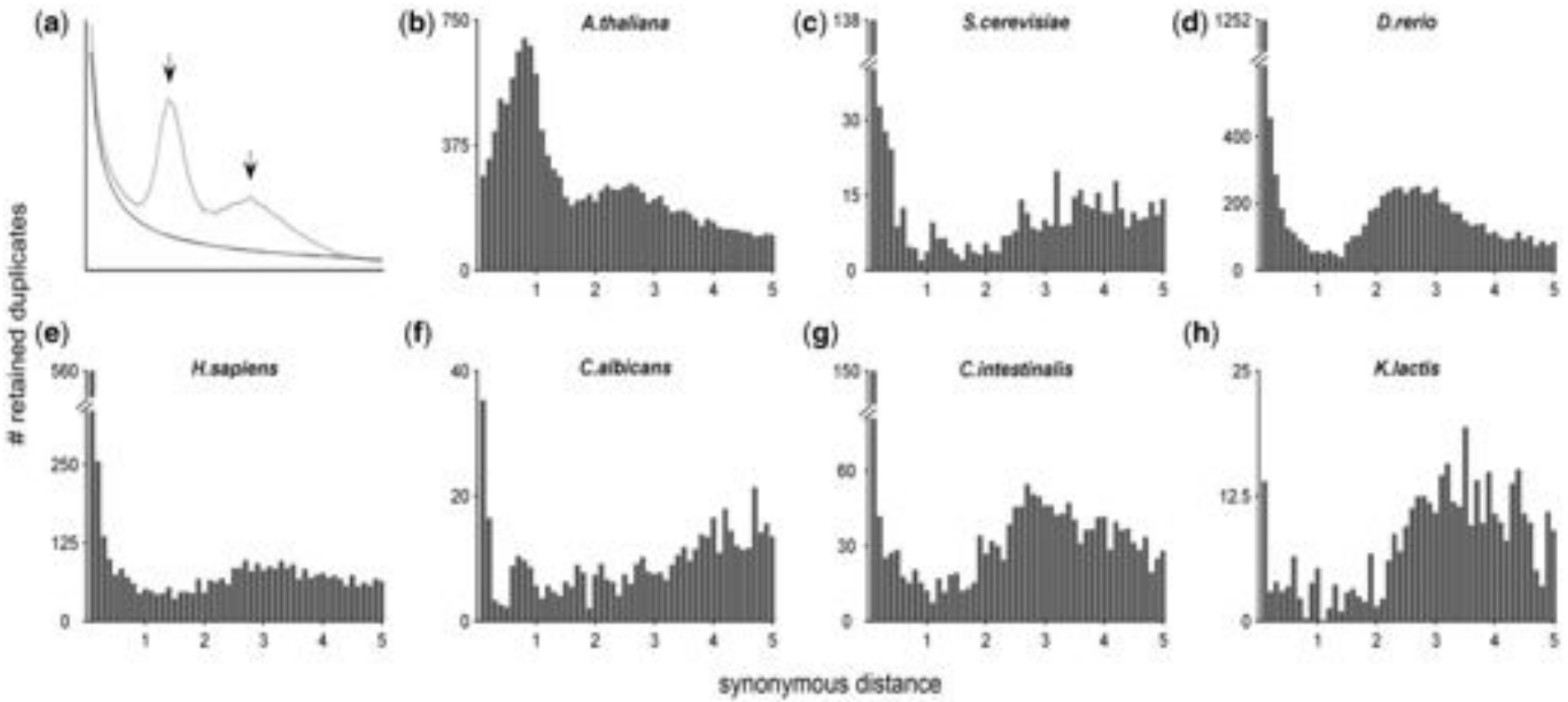
Duplication may be more frequent than single nucleotide subsitution in the human genome



Lynch and Conery (2000) Science 290: 1151-55

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

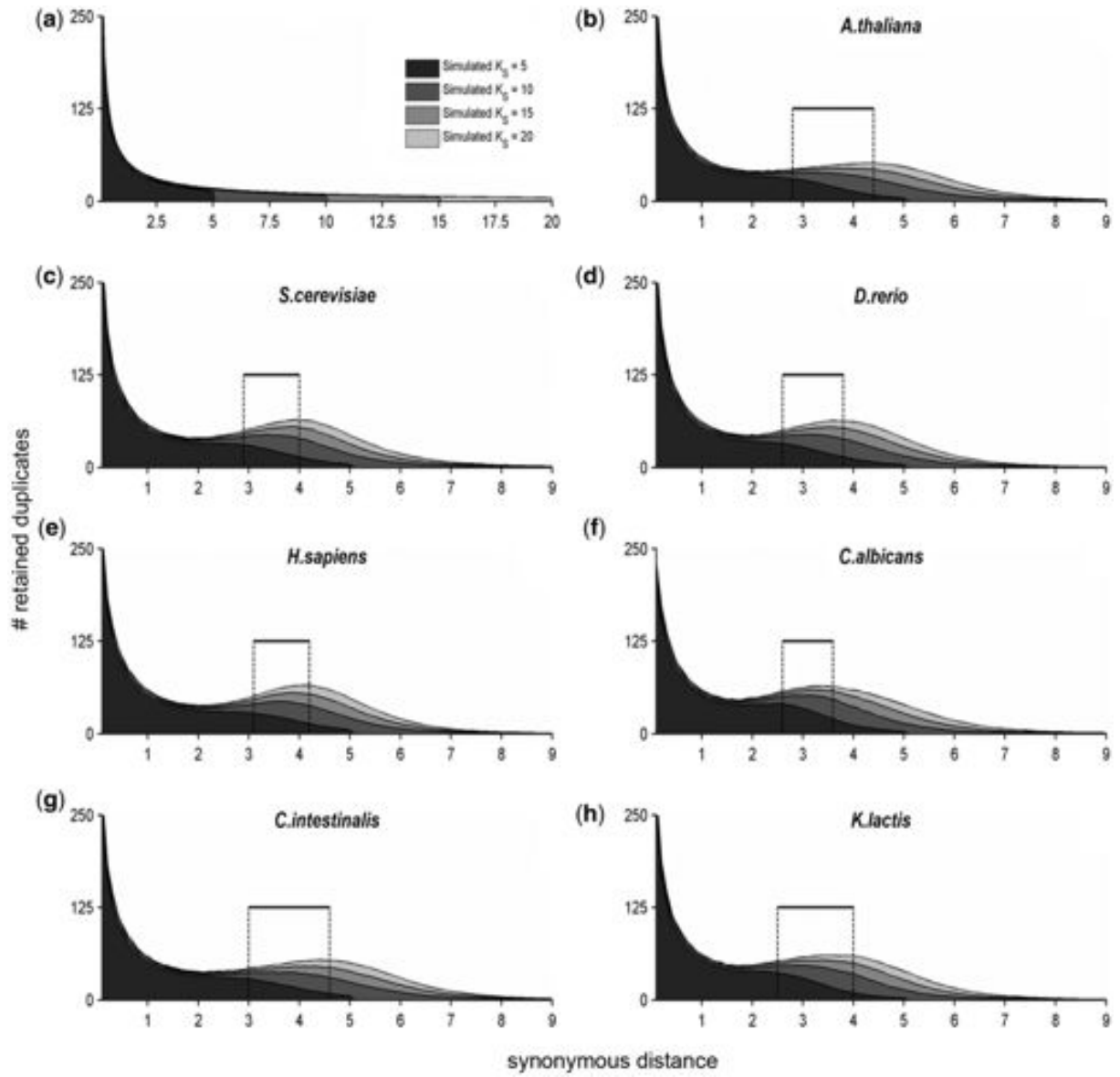
Can we trust Ks histograms?



Vanneste K et al. Mol Biol Evol 2013 30(1):177-90

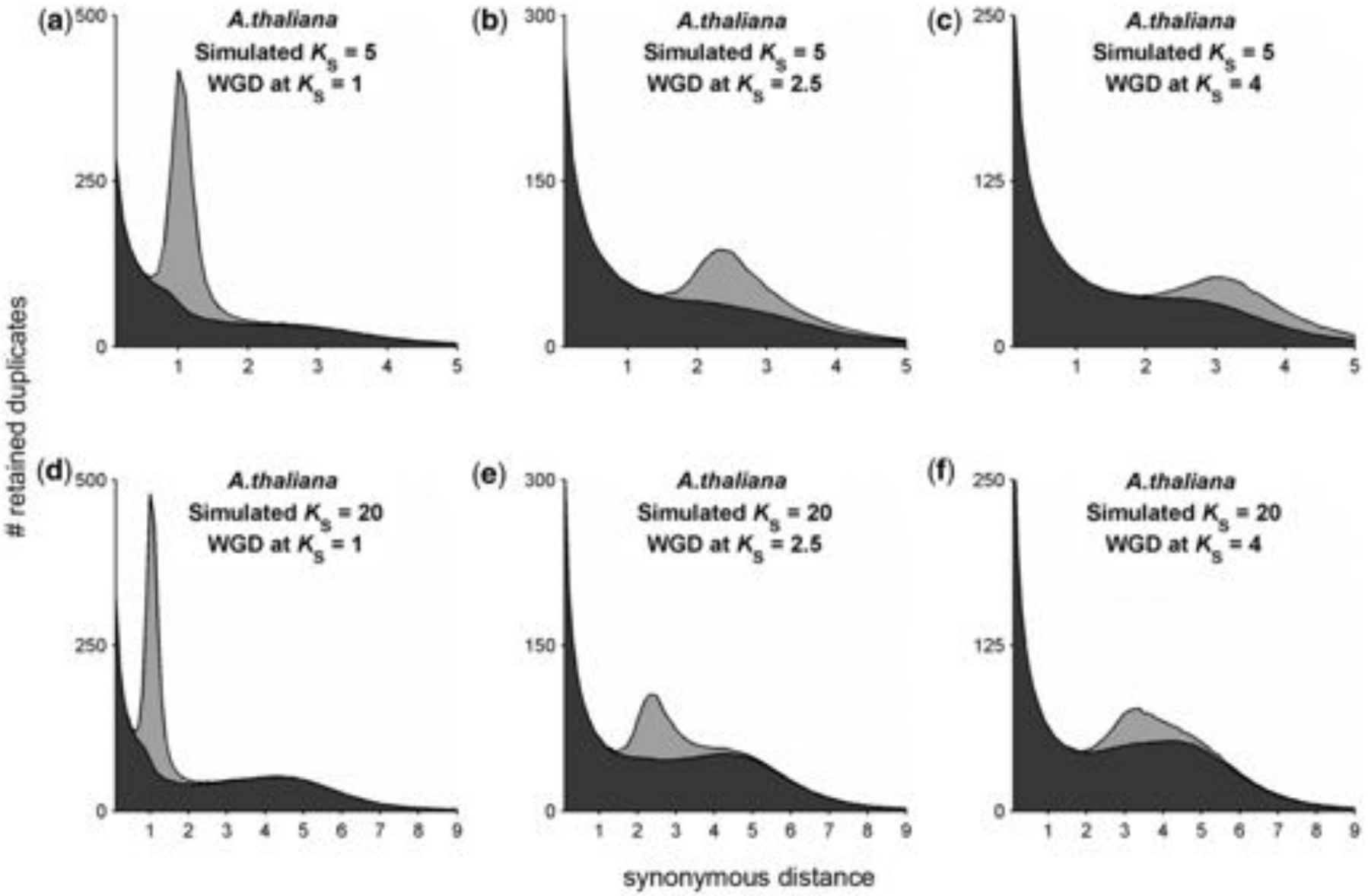
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

SSD age distributions are characterized by a saturation peak



Vanneste K et al. Mol Biol Evol 2013 30(1):177-90

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics



Vanneste K et al. Mol Biol Evol 2013 30(1):177-90

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Duplication great and small

Duplication can lead to lineage specific expansion of gene families

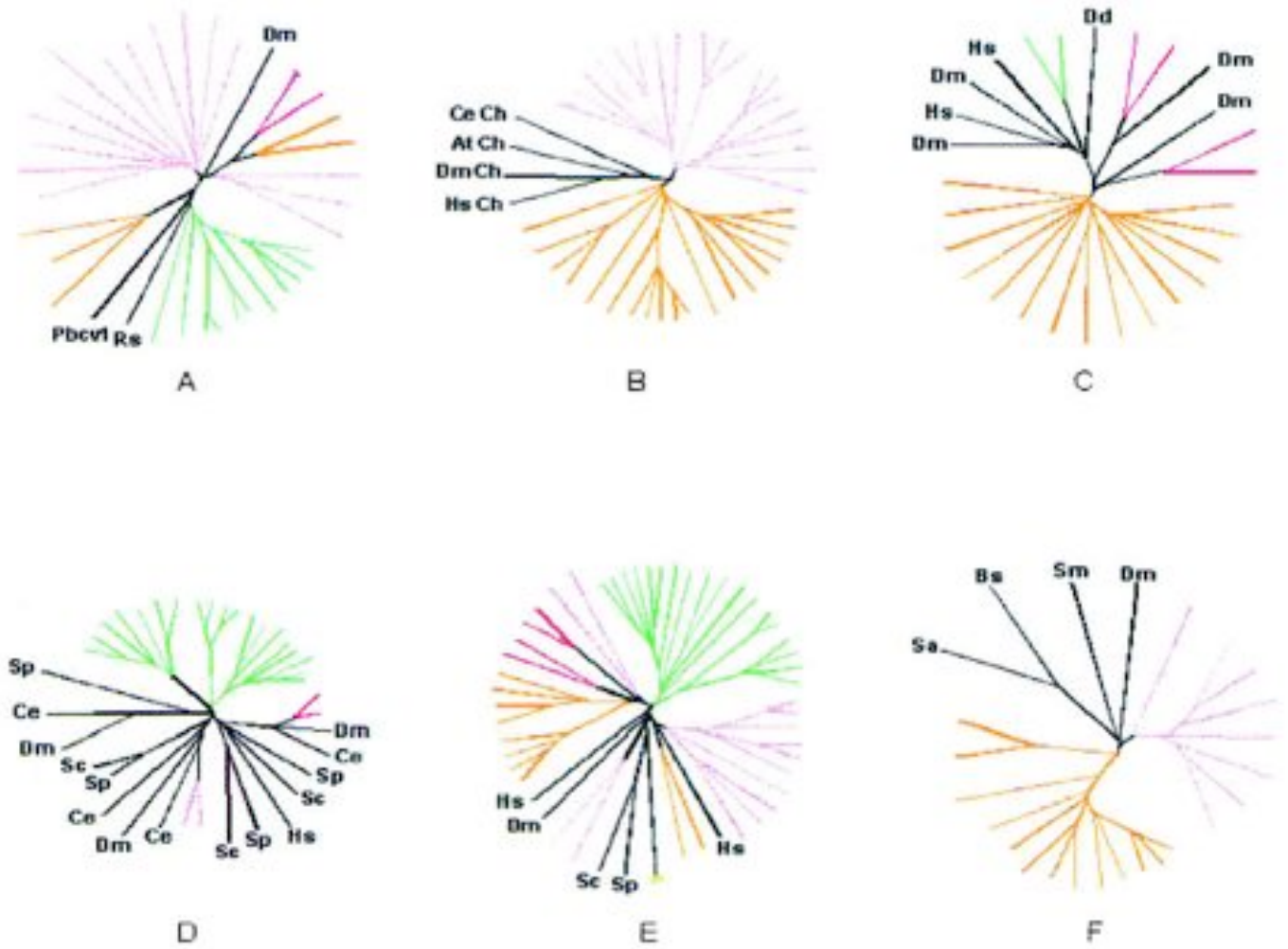


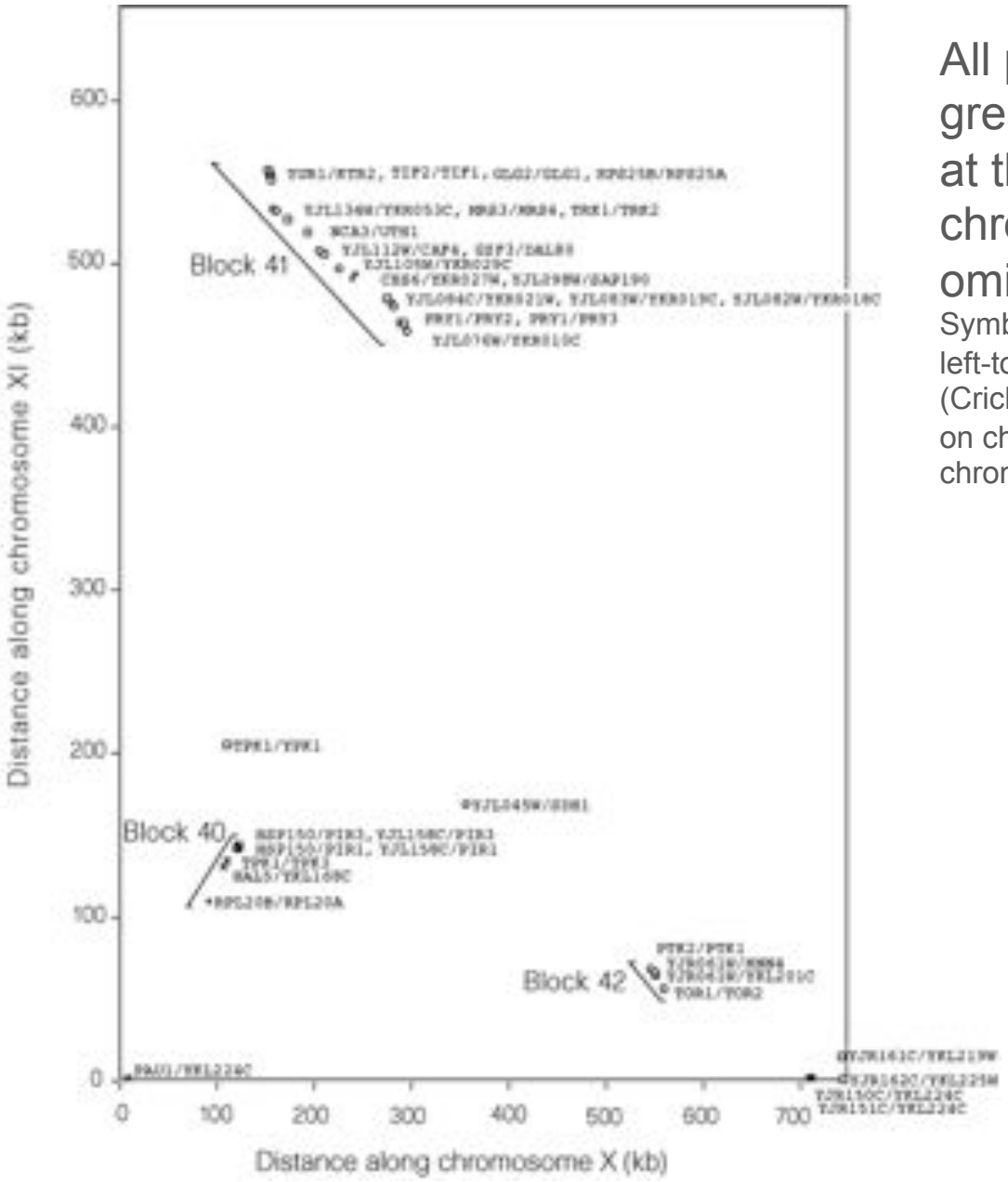
Figure 1 Phylogenetic analysis of selected eukaryotic lineage-specific expansions. Groups supported by a bootstrap value >70% are colored pink for *Drosophila melanogaster*, red for *Homo sapiens*, orange for *Caenorhabditis elegans*, green for *Arabidopsis thaliana*, and yellow for *Schizosaccharomyces pombe*. (*A*) Prolyl hydroxylases. (*B*) Small molecule kinases (Ch stands for choline kinase). (*C*) Patched-like protein. (*D*) MAP-Kinases. (*E*) P450 family hydroxylases. (*F*) MBOAT membrane acyltransferases. (At) *Arabidopsis thaliana*; (Bs) *Bacillus subtilis*; (Ce) *Caenorhabditis elegans*; (Dd) *Dictyostelium discoideum*; (Dm) (*Drosophila melanogaster*; (Hs) *Homo sapiens*; (Pbcv1) *Paramecium bursaria Chlorella* virus 1; (Rs) *Ralstonia solanacearum*; (Sa) *Staphylococcus aureus*; (Sc) *Saccharomyces cerevisiae*; (Sm) *Sinorhizobium meliloti*; (Sp) *Schizosaccharomyces pombe*. *Complete tree descriptions (full lists of GI numbers or gene names, and bootstrap values) are available in the Supplementary Material online at ftp://ncbi.nlm.nih.gov/pub/aravind/expansions, and http://www.genome.org.*.

Lespinet et al. (2002) Genome Res 12:1048-59

Whole genome duplications

UNIL | Université de Lausanne
Département d'écologie et évolution

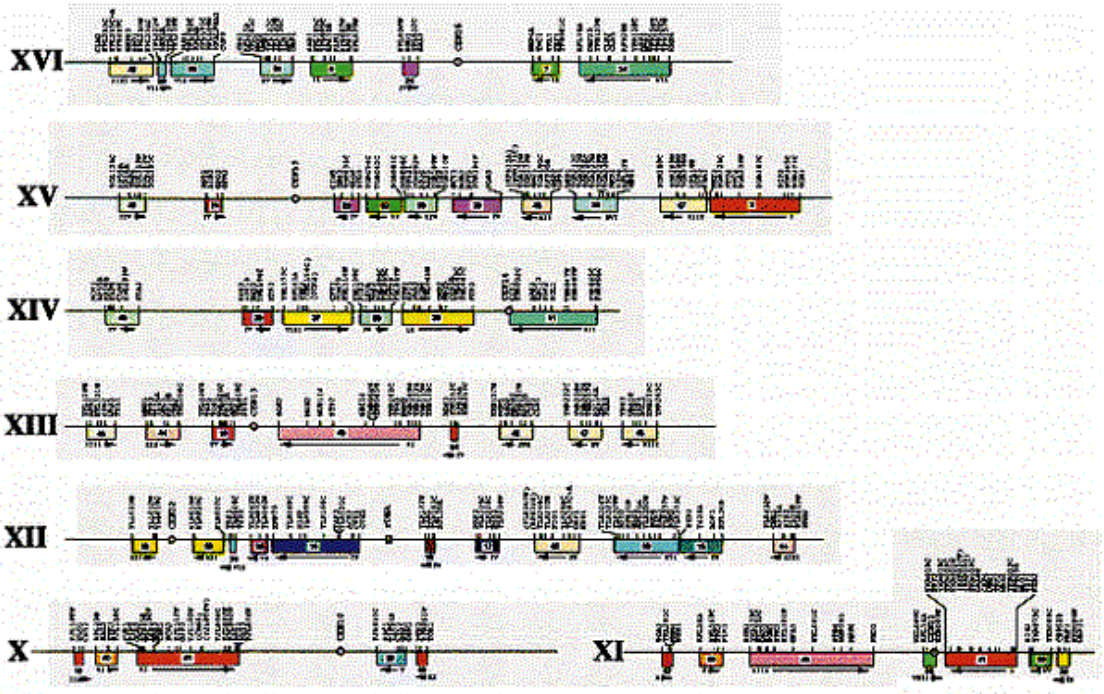
SIB Swiss Institute of Bioinformatics



All pairs of proteins with BLASTP scores greater than or equal to 200 are plotted at the position of their genes on the two chromosomes. Ty elements have been omitted.

Symbols indicate gene orientations: +, W (Watson strand; left-to-right transcription) on both chromosomes; times, C (Crick strand; right-to-left transcription) on both; squares, C on chromosome X but W on chromosome XI; circles, W on chromosome X but C on chromosome XI.

Wolfe and Shields (1997) Nature 387: 708-13



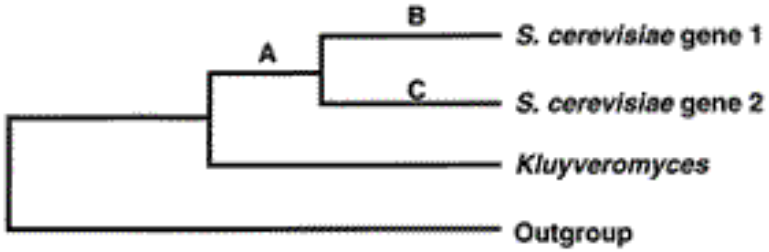
χ^2 relative to random distribution of transcriptional orientation and gene order $P < 0.0001$
less triplicates than expected by chance (Poisson) $P = 0.001$

≈ 85% loss of duplicates after whole genome duplication

55 duplicate regions were identified containing 376 pairs of homologous genes.
Amino-acid sequence identity between the pairs: 24% to 100%, with a mean of 63%.
Criteria used to define a duplicate region:

(1) BLASTP high scores of greater than or equal to 200 for each gene pair (these have an associated significance of $P = 10^{-18}$ or less)
(2) ≥ 3 pairs of homologues with intergenic distances ≤ 50 kb on each chromosome
(3) conservation of gene order and orientation (with allowance for small inversions within some blocks).

Duplicated regions on average 55 kb long, 6.9 duplicate gene pairs -> span 50% of the genome.



Block	Gene 1 / Gene 2	Bootstrap support (%)	Age	90% Confidence interval	Amino-acid sites	Branch length A	Branch length B	Branch length C	Outgroup
3	*NTH2 / NTH1*	100	0.70	(0.59 - 0.82)	714	0.052	0.131	0.115	*Candida albicans*
3	*GAL1 / GAL3*	99	0.55	(0.41 - 0.76)	468	0.100	0.115	0.130	*Homo sapiens*
37	*YCK1 / YCK2*	97	0.48		343	0.036	0.046	0.024	*Sch. pombe*
36	*HXK1 / HXK2*	94	0.77	(0.62 - 0.95)	484	0.037	0.147	0.107	*Sch. pombe*
45	*EXG1 / SPR1*	90	0.80	(0.67 - 0.97)	418	0.051	0.166	0.270	*Candida albicans*
10	*CRY1 / CRY2*	89	0.40		137	0.015	0.014	0.007	*Homo sapiens*
21	*SIR2 / HST1*	45	0.86	(0.66 - 1.00)	402	0.028	0.167	0.182	*Caenorhabditis elegans*
28	*CYC7 / CYC1*	34	0.82		106	0.017	0.110	0.057	*Sch. pombe*
30	*YGR043C / TAL1*	24	1.00		332	0.000	0.231	0.149	*Homo sapiens*
Genes with apparent gene conversion									
8	*RPS10A / RPS10B*	100	0.04		233	0.042	0.000	0.004	*Sch. pombe*
20	*SSB1 / SSB2*	100	0.06	(0.01 - 0.11)	613	0.049	0.002	0.004	*Candida albicans*
37	*RPL41B / RPL41A*	100	0.00		105	0.054	0.000	0.000	*Candida tropicalis*

Duplications which can be dated occured after the S. cerevisiae / Kluyveromyces speciation

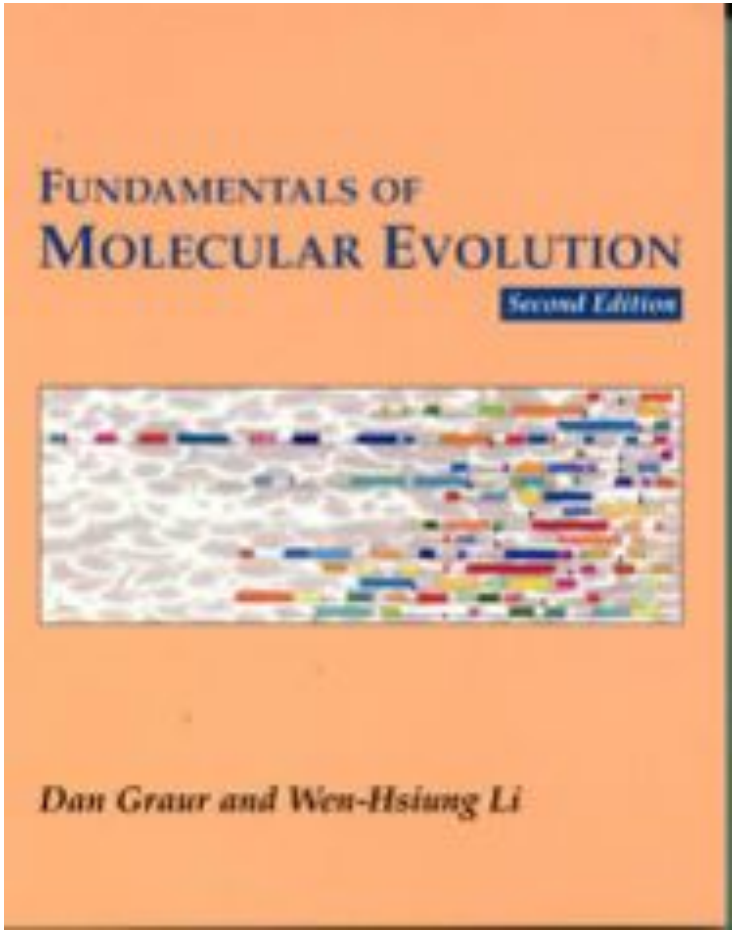
Wolfe and Shields (1997) Nature 387: 708-13

A few take home points:

• the ancient whole genome duplication was unsuspected before genome sequencing + bioinformatics

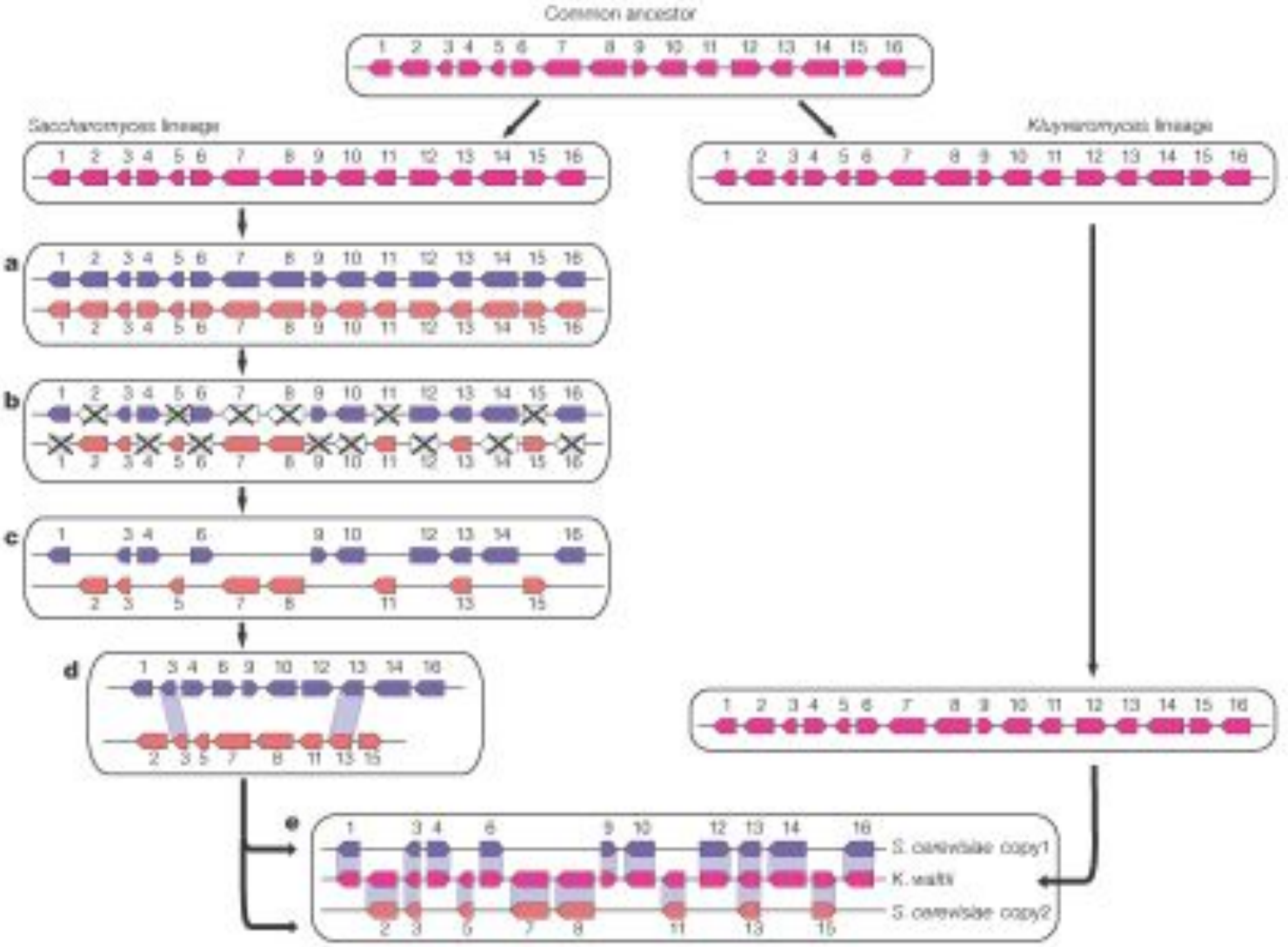
• synteny information provided key evidence

• most genes were lost after the whole genome duplication



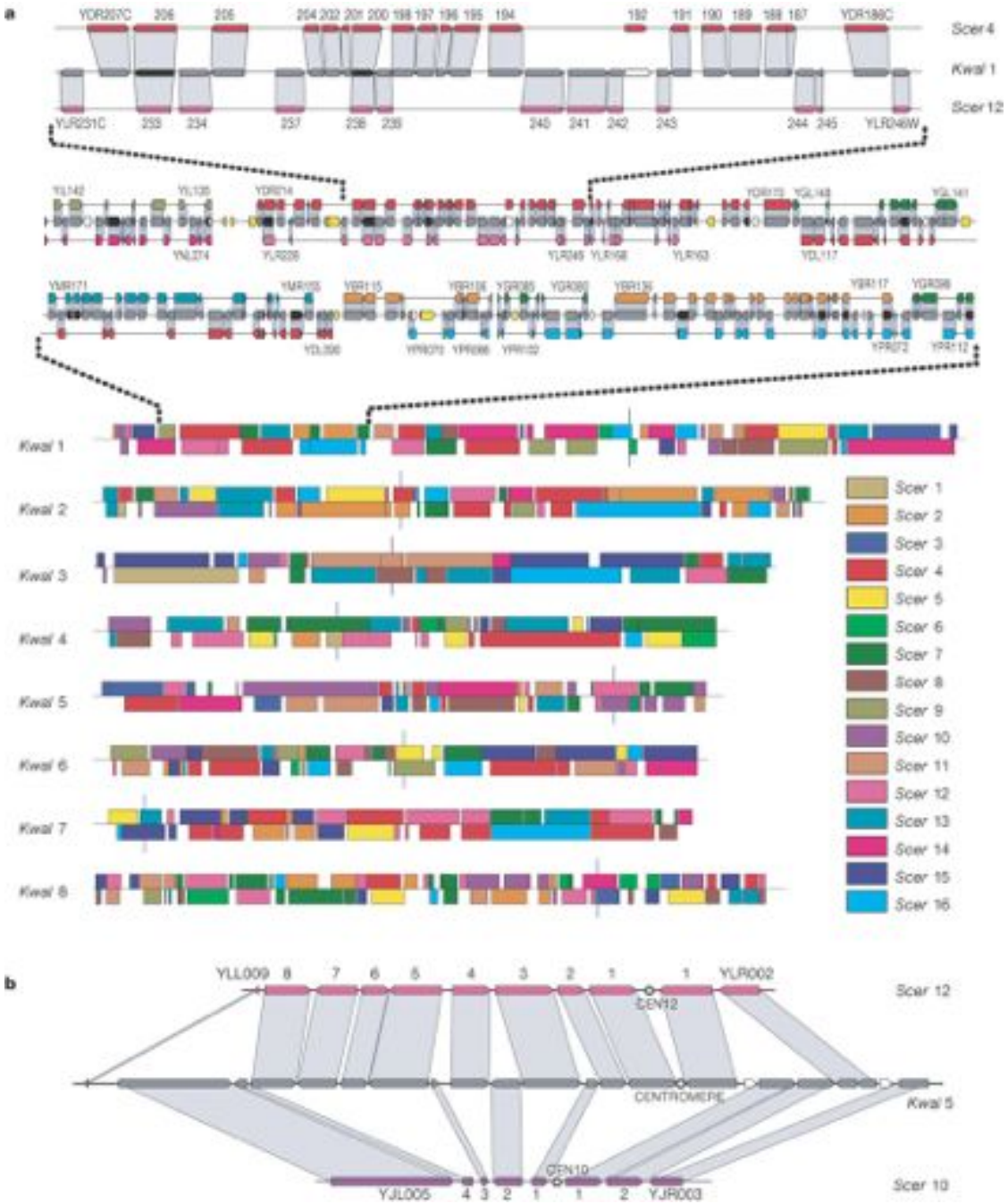
Of the importance of comparing genomes

Model of WGD followed by massive gene loss predicts gene interleaving in sister regions



Kellis et al (2004) Nature 428: 617-24

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics



Gene and region correspondence with *K. waltii* reveals Whole Genome Duplication

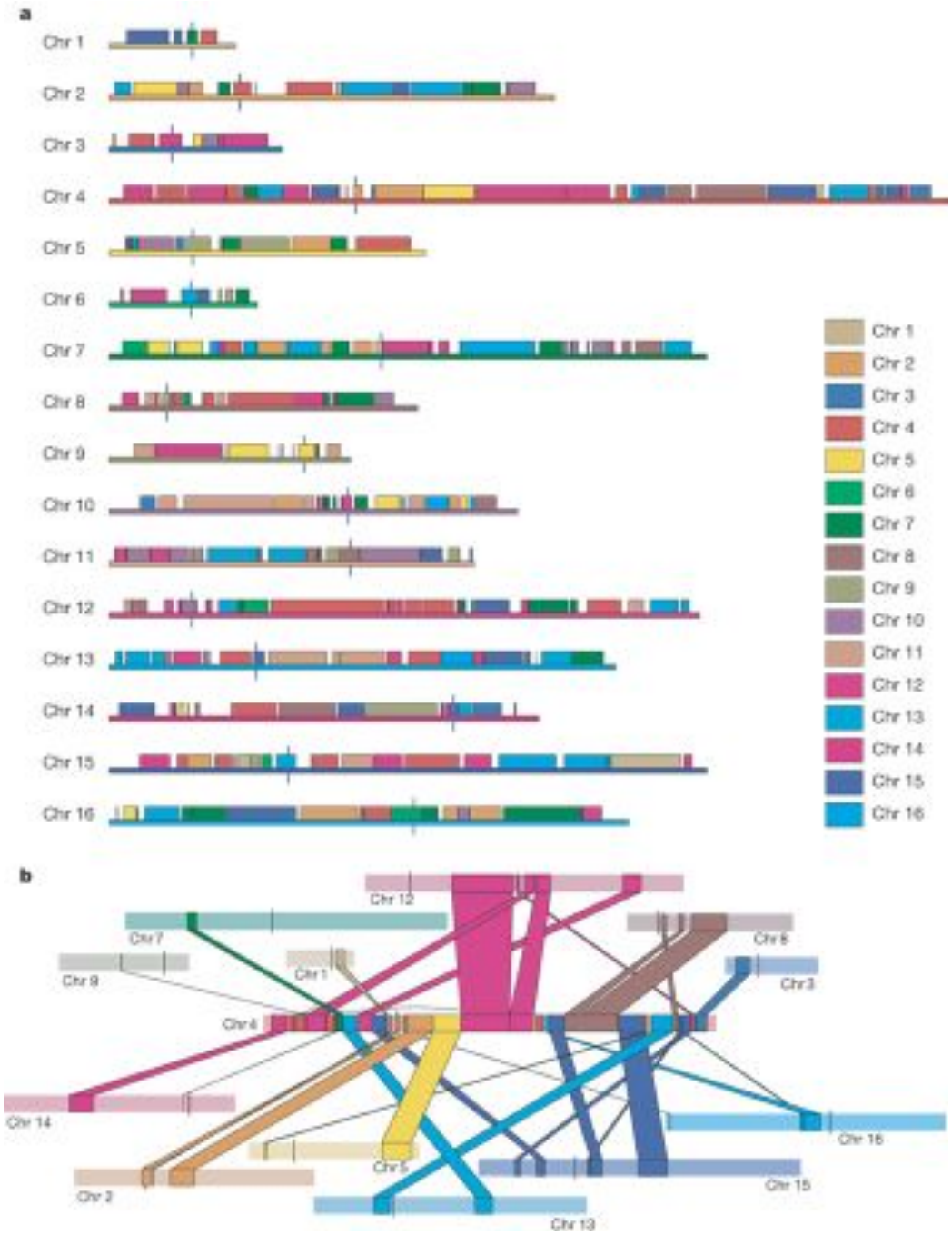
a- Each region of *K. waltii* (coloured by number of matches: white, 0; grey, 1; black, 2; yellow, > = 3) shows conserved gene order with two regions of S. cerevisiae (coloured by chromosome number). Spacing between *S. cerevisiae* genes is set to match *K. waltii* chromosomal positions. Vertical blue bars denote centromeres.

b- Doubly conserved synteny region showing duplicate mapping of centromeres (black circles). All sixteen *S. cerevisiae* centromeres show such duplicate mapping with *K. waltii* centromeres. This DCS region also illustrates that we can reliably recognize anciently duplicated segments even in the absence of any remaining two-copy genes. Evidence of WGD comes from gene interleaving and 2:1 mapping with orthologous *K. waltii* segments. The segments containing intervening genes are deleted, resulting in condensed sister regions. Kwal, *K. waltii*; Scer, *S. cerevisiae*.

Kellis et al (2004) Nature 428: 617-24

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Duplicated blocks in *S. cerevisiae*



Kellis et al (2004) Nature 428: 617-24

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

A few take home points:

• data alone is not sufficient to prove whole genome duplication, good bioinformatics is needed

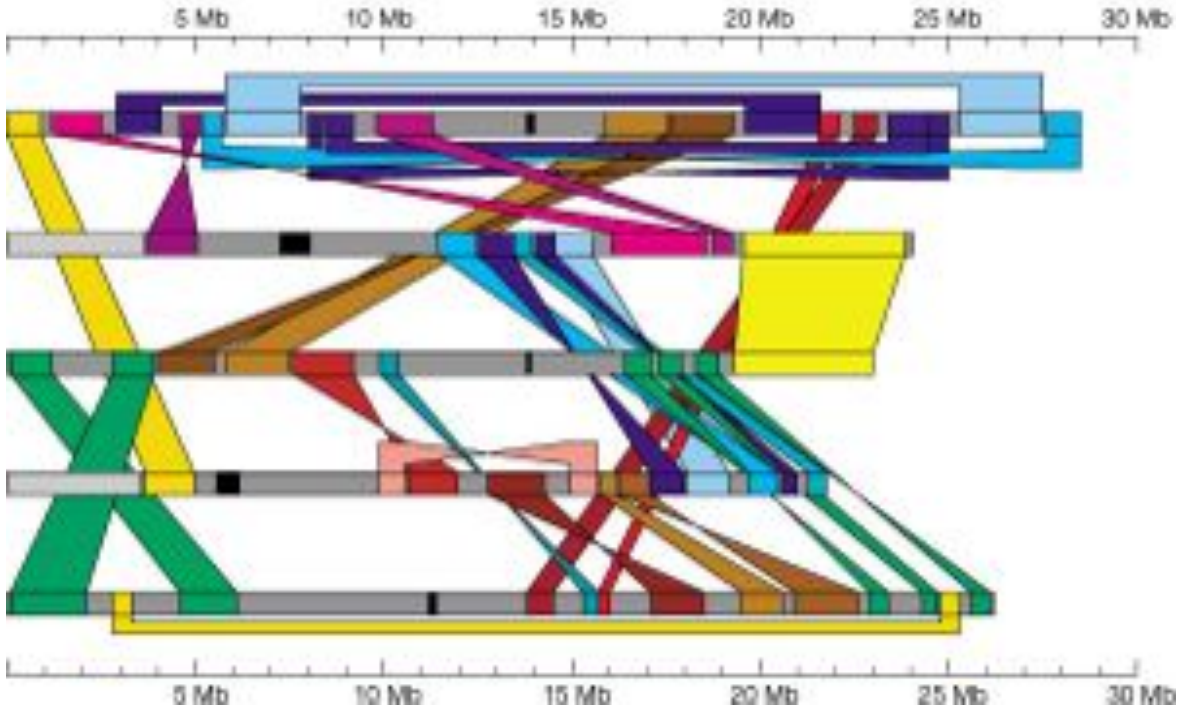
• no obvious link between genome duplication and organism complexity

• duplication can be proven even after secondary loss (e.g. there is only one gene left)

Also in plants

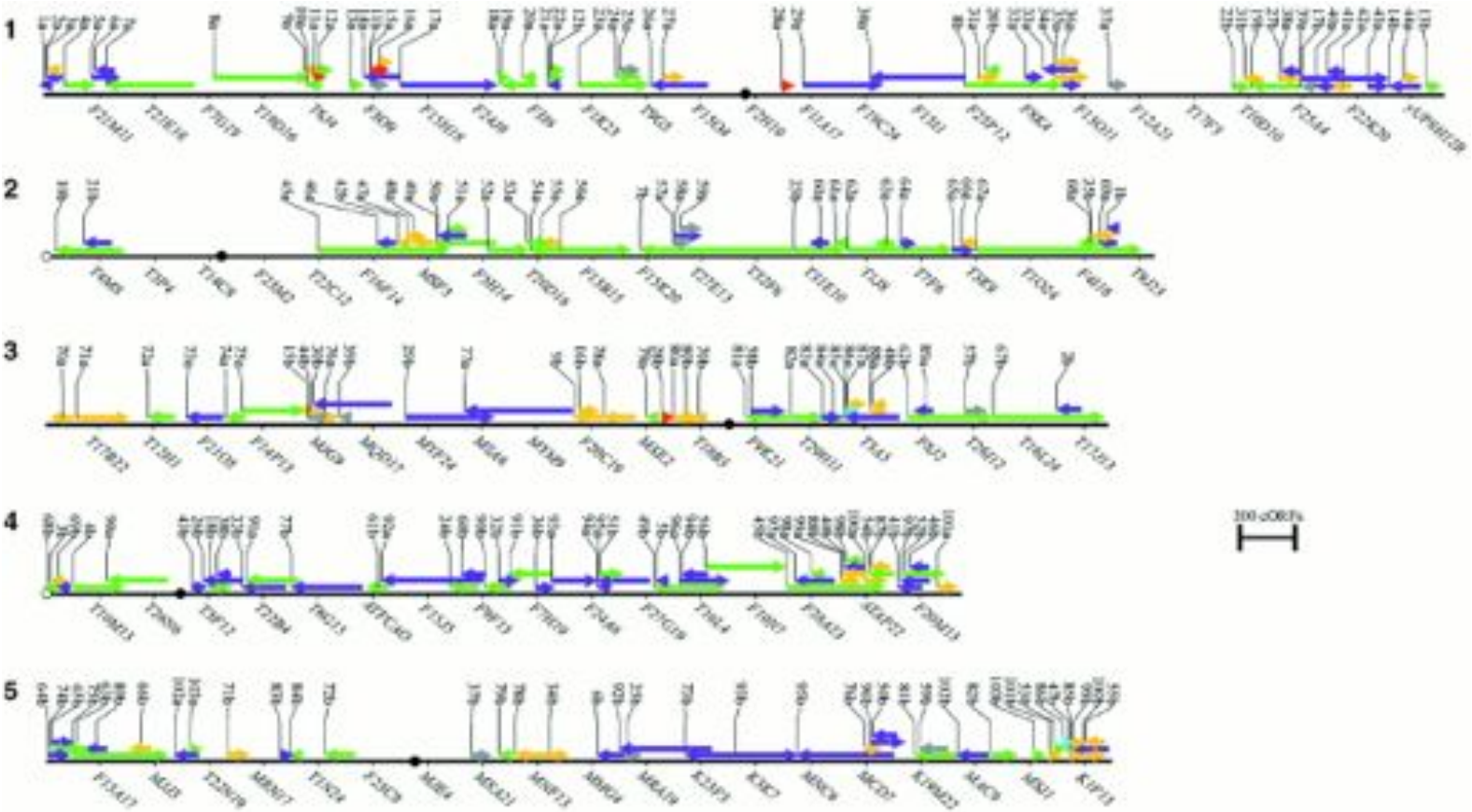
Arabidopsis genome:
DNA based: All five chromosomes aligned to each other in both orientations using MUMmer results filtered to identify all segments at least 1,000 bp in length with at least 50% identity.
-> 24 large duplicated segments of 100 kb or larger, comprising 65.6 Mb or 58% of the genome.

Protein based: TBLASTX to identify collinear clusters of genes residing in large duplicated chromosomal segments.
-> 67.9 Mb, 60% of the genome.
sequence conservation of the duplicated genes varies greatly:
6,303 (37%) of the 17,193 genes in the segments classified as highly conserved (E < 10 -30)
1,705 (10%) with less significant similarity up to E < 10-5.



Genomic map of duplicated blocks in Arabidopsis

The two copies of each putative duplicated block (e.g., 1a and 1b) are shown. Color denotes age class (red, A; blue, B; green, C; purple, D; orange, E; gray, F). Centromeres are shown with black circles, and ribosomal DNA with white circles. Direction of arrowhead indicates the predominant relative orientation of duplicated cORFs within each block (right, direct; left, inverted). Landmarks are given at 200 composite ORF intervals.



Vision et al. (2000) Science 290: 2114-17

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Table 1. Features of the five age classes of duplicated blocks. d_A is the minimum change in amino acids per between dispersed duplicated cORFs, averaged among all cORFs. Retained duplicates, ratio of presently duplicated to inferred ancestral cORFs. Block size, mean number of cORFs (including singletons) per copy.

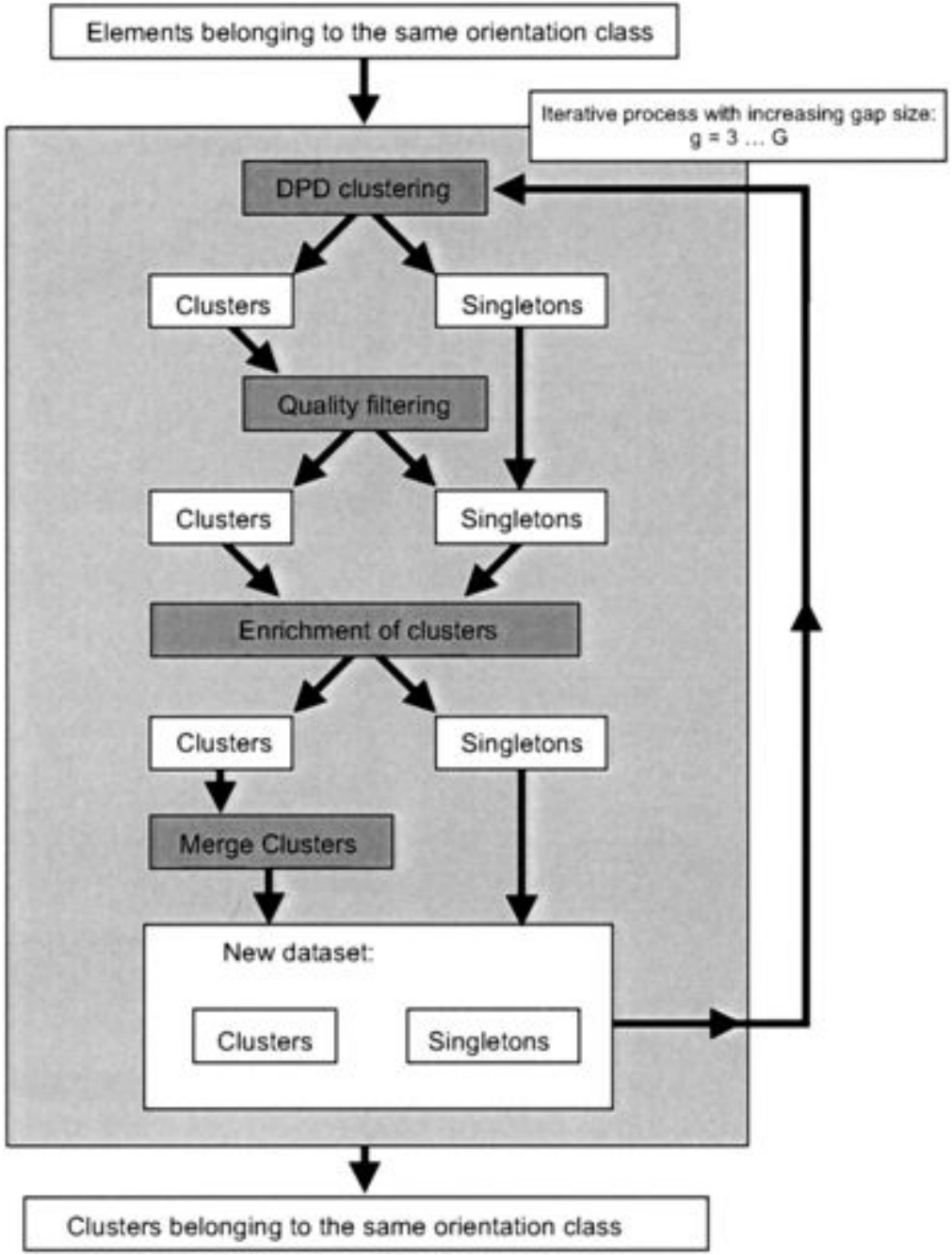
Age class	No. of blocks	d_A	Retained duplicates	Block size	Estimated age (Mya)
A*	2	0†	0.90	12.5	0
B	1	0.21†	0.38	14.5	50
C	35	0.45‡	0.15	149.5	100
D	36	0.57‡	0.13	128.0	140
E	23	0.71‡	0.11	60.0	170
F	6	0.84†	0.09	57.8	200

*Probable artifact of genome assembly. †Median among all duplicated genes. ‡Mean of best-fit normal distribution to block medians.

Overlaps between the blocks
duplicates of different "ages"
-> many block duplications in the past evolution of Arabidopsis?

Vision et al. (2000) Science 290: 2114-17

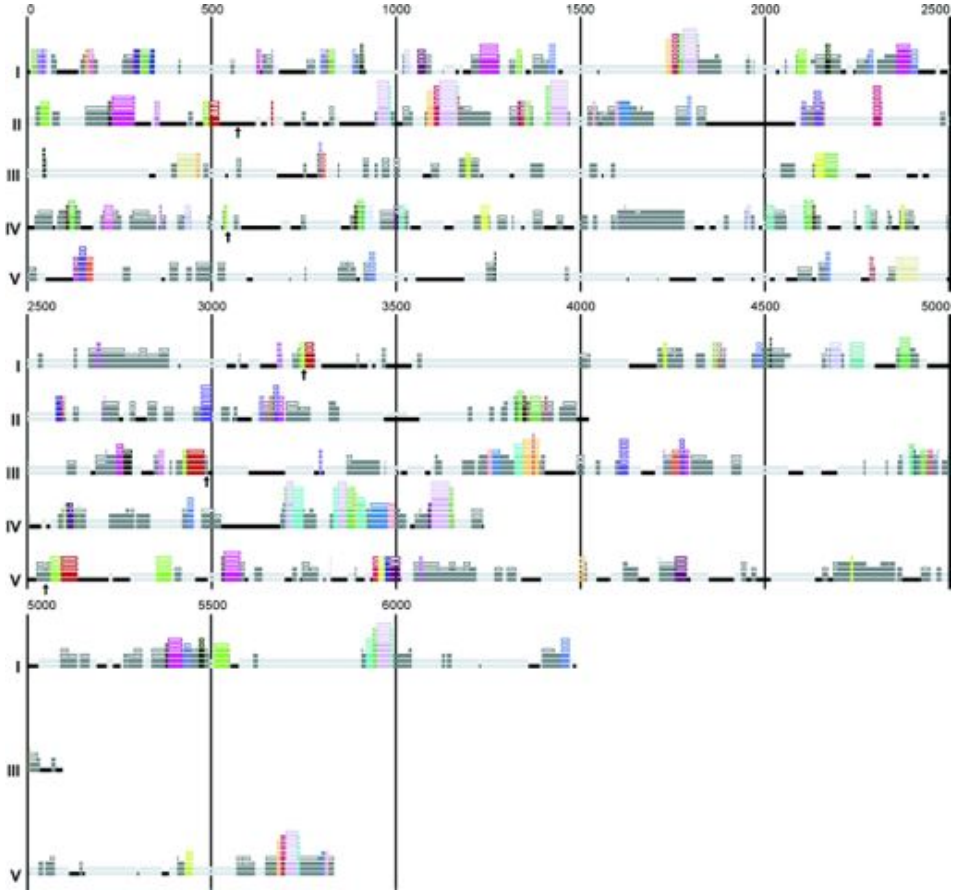
Sensitive methods are important



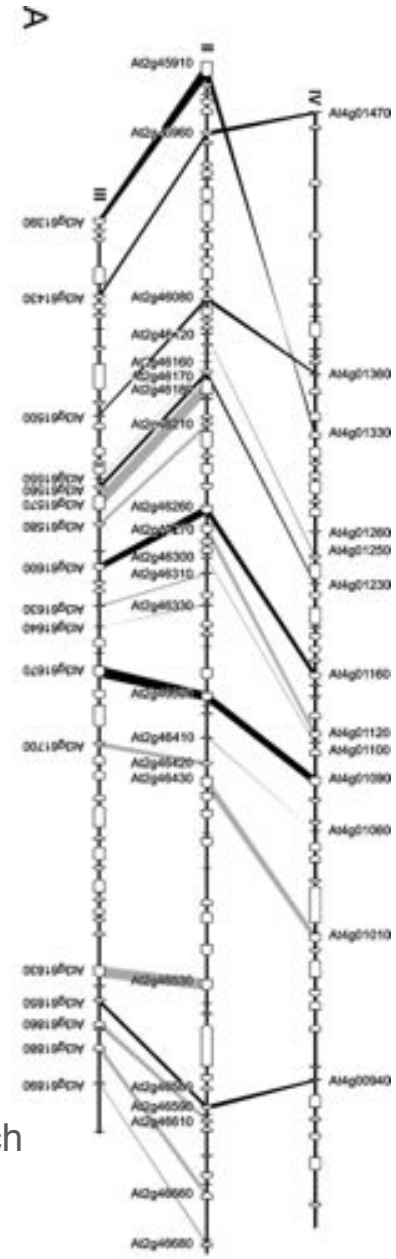
ADHoRe

Vandepoele et al. (2002) Genome Res. 12: 1792-1801

304 non hidden duplications with AHDoRe
total of 3571 anchor points
+ 1607 tandem repeats, involving 4193 individual genes.



Overview of the chromosomal location of all multiplicons detected in the Arabidopsis genome

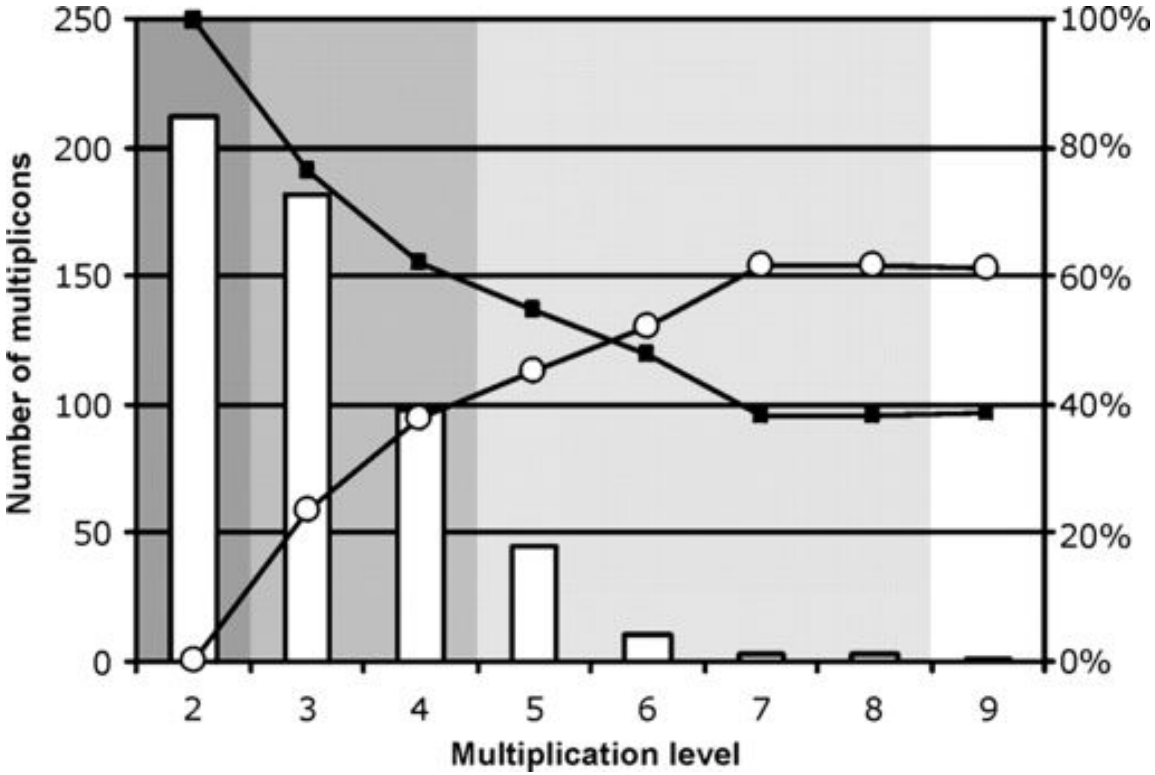


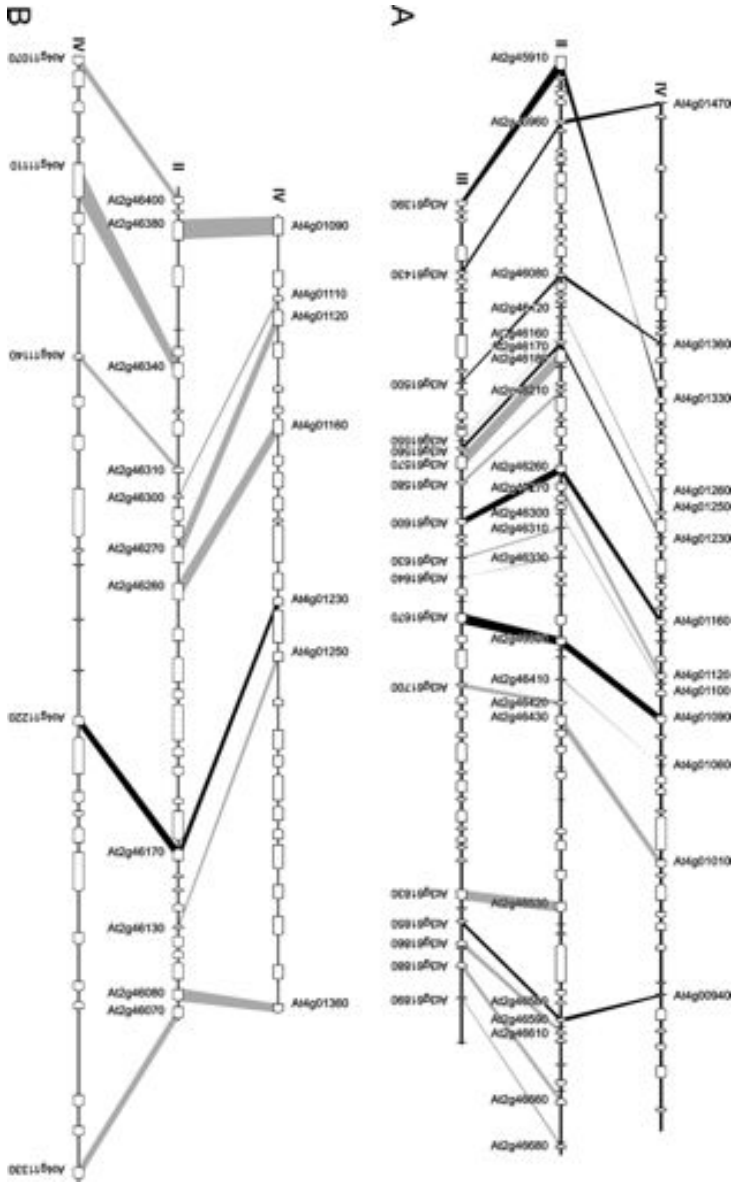
Example of a multiplicon in which nonhidden duplications can be observed between all three segments involved

Simillion et al. (2002) PNAS 99: 13627-32

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

53 hidden duplications detected:
10 genes (51 kb) to 218 genes (1.15 Mb)





Conclusion: 3 whole genome duplications in Arabidopsis evolution

Simillion et al. (2002) PNAS 99: 13627-32

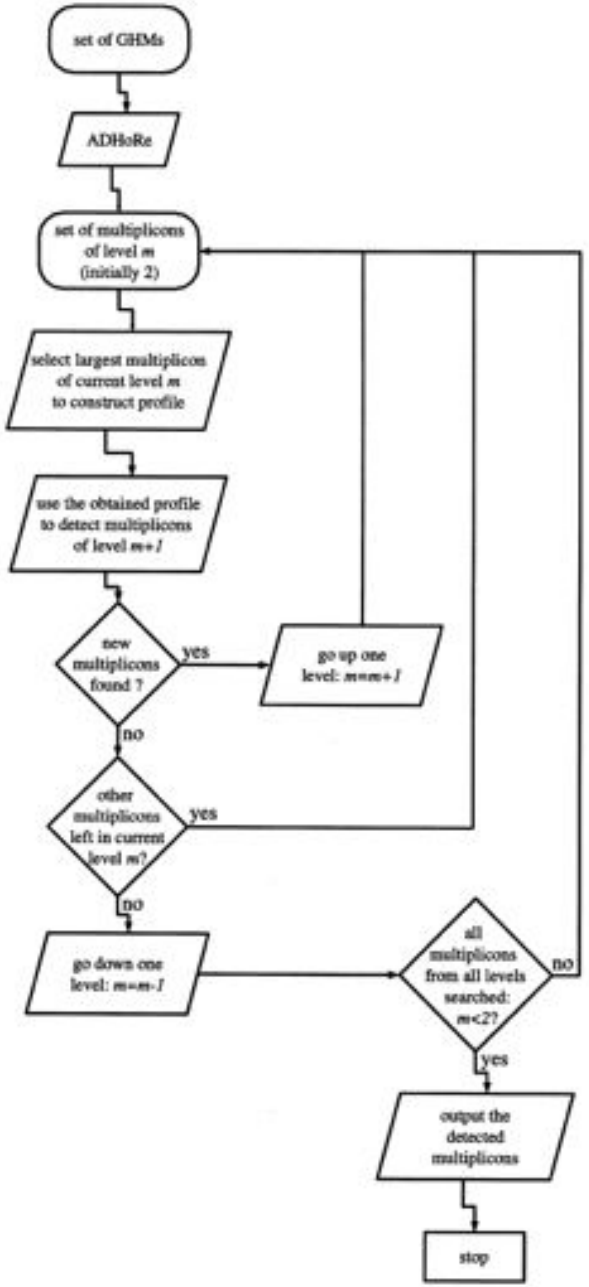
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

i-ADHoRe

Input: genes from one or several genomes
-> matrix of homology by BLAST

clusters defined as significant by probability of finding another anchor by chance, given the total density of homologous pairs

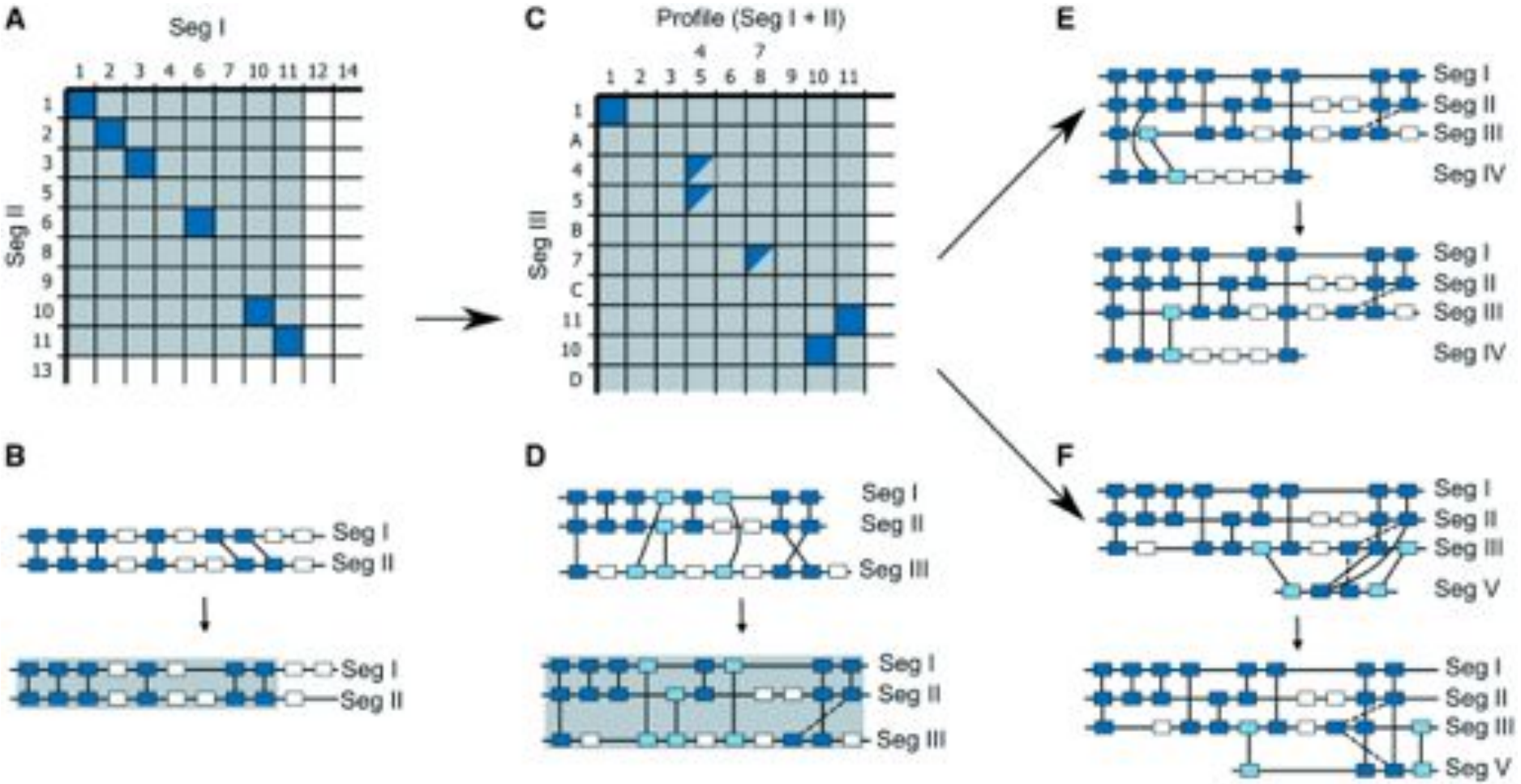
iterative procedure to find multiple duplications (multiplicons):
search for homology to previously defined blocks of high score



Simillion et al. (2004) Genome Res 14:1095-1106

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

i-ADHoRe



Proost S et al. Nucl. Acids Res. 2012;40:e11-e11

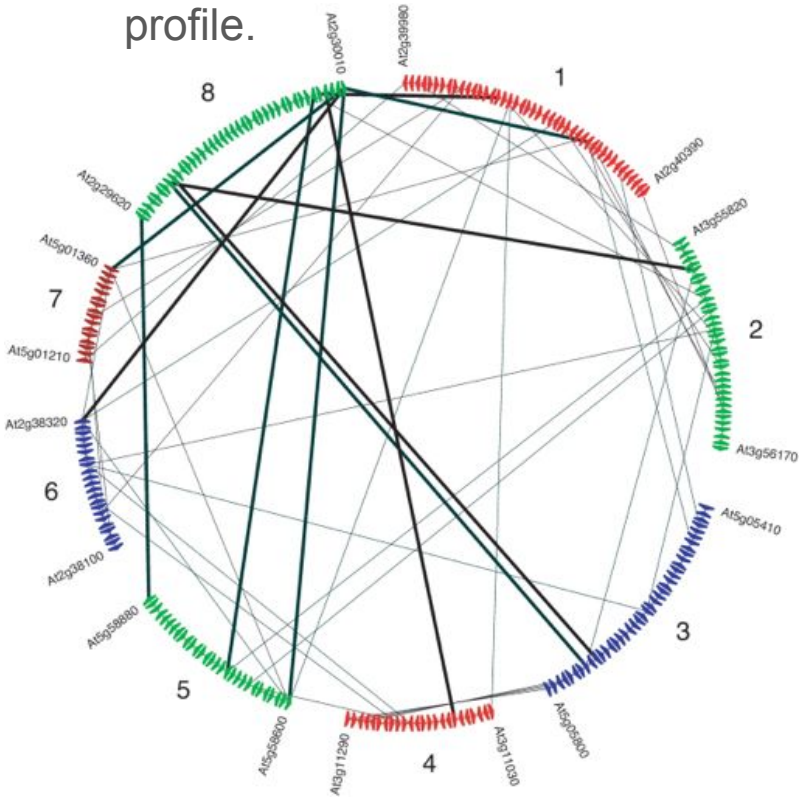
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

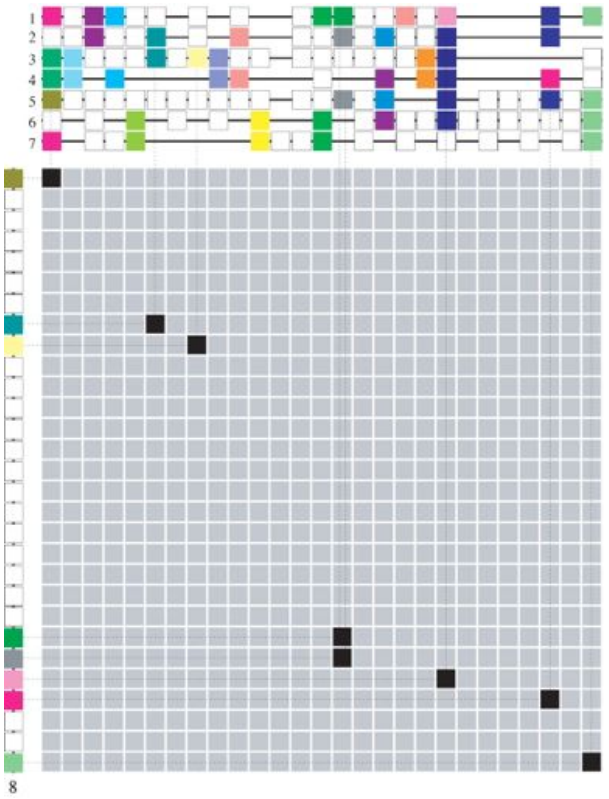
Example of an Arabidopsis multiplicon

Segments shown clockwise in the order in which they were added to the multiplicon:

Segments 1 and 2 were detected first as a pair of collinear segments that were consequently aligned to create a profile.

This profile was then used to detect segment 3, which, in turn, was aligned and added to the profile.

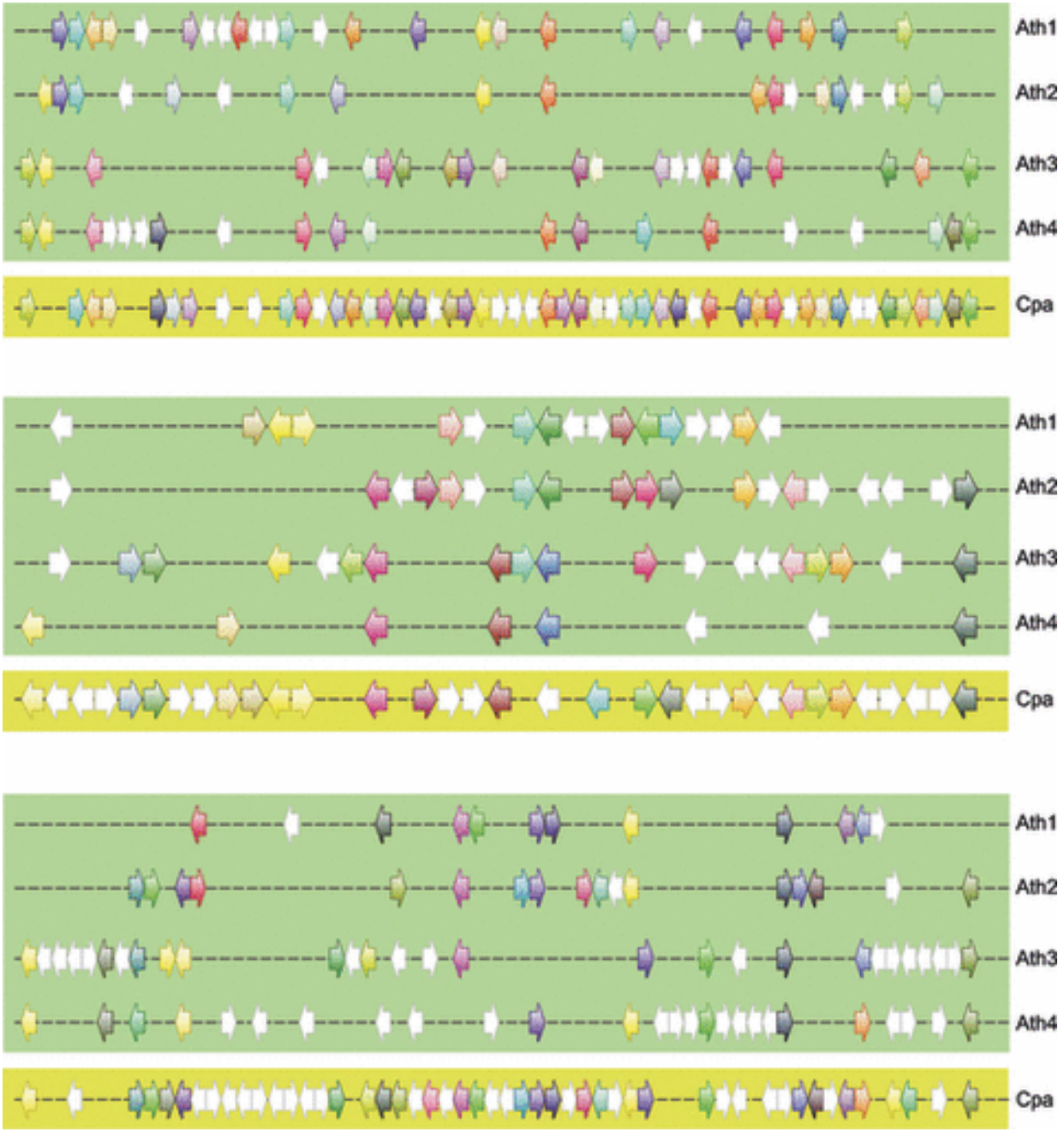




Based on the similarities with any of the other individual segments, the homology with segment 8 is far from statistically significant because there are too few anchor points and the number of intervening genes is too high. However, if we consider the seven other segments in the multiplicon (of which the mutual homology has already been established) together as a profile, we see that segment 8 shares in total eight genes with the profile.

Simillion et al. (2004) Genome Res 14:1095-1106

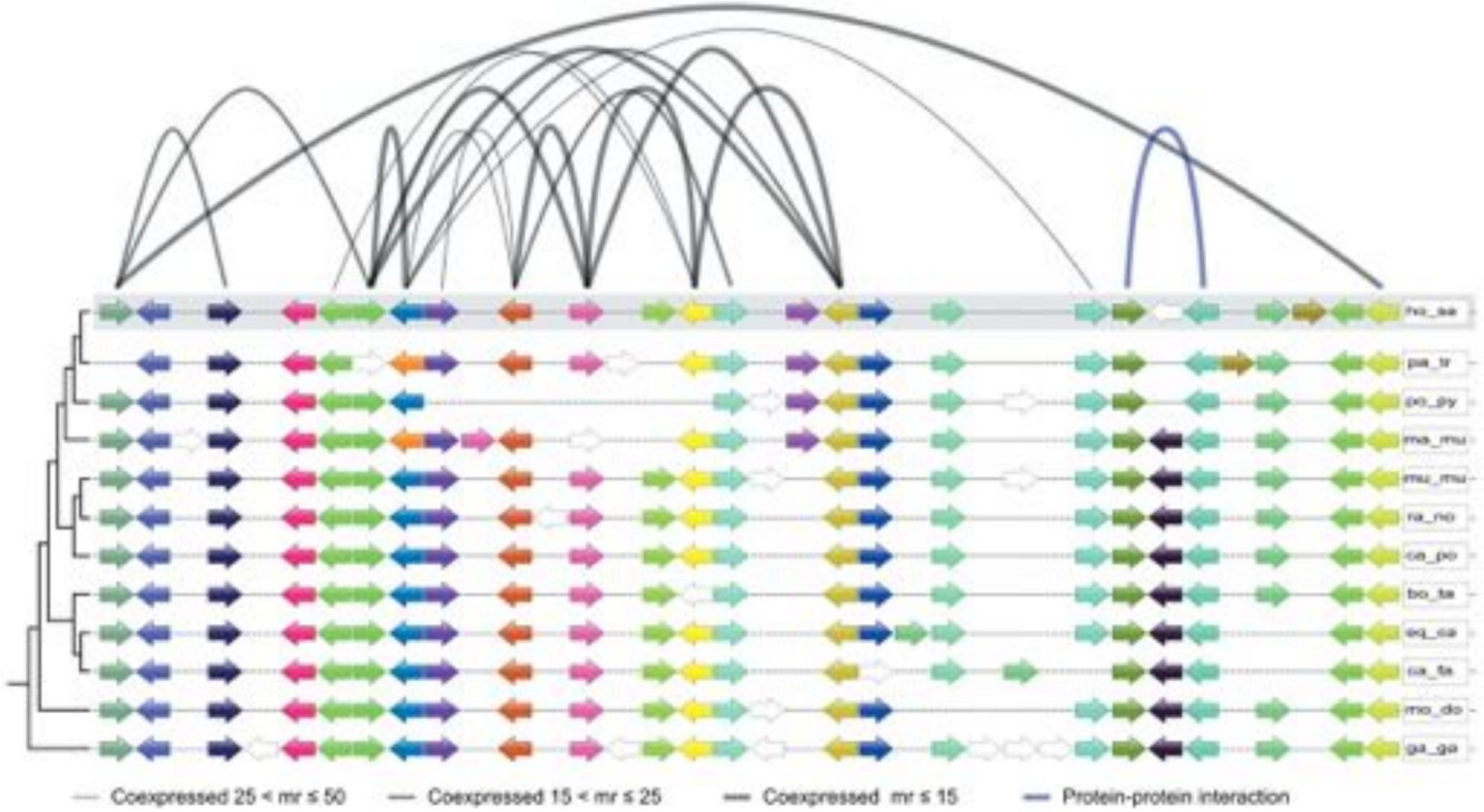
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics



Proost et al. 2011 The Plant Journal 66: 58-65

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Gene order alignment of collinear regions conserved over a large phylogenetic distance (human–chicken).



Proost S et al. Nucl. Acids Res. 2012;40:e11-e11

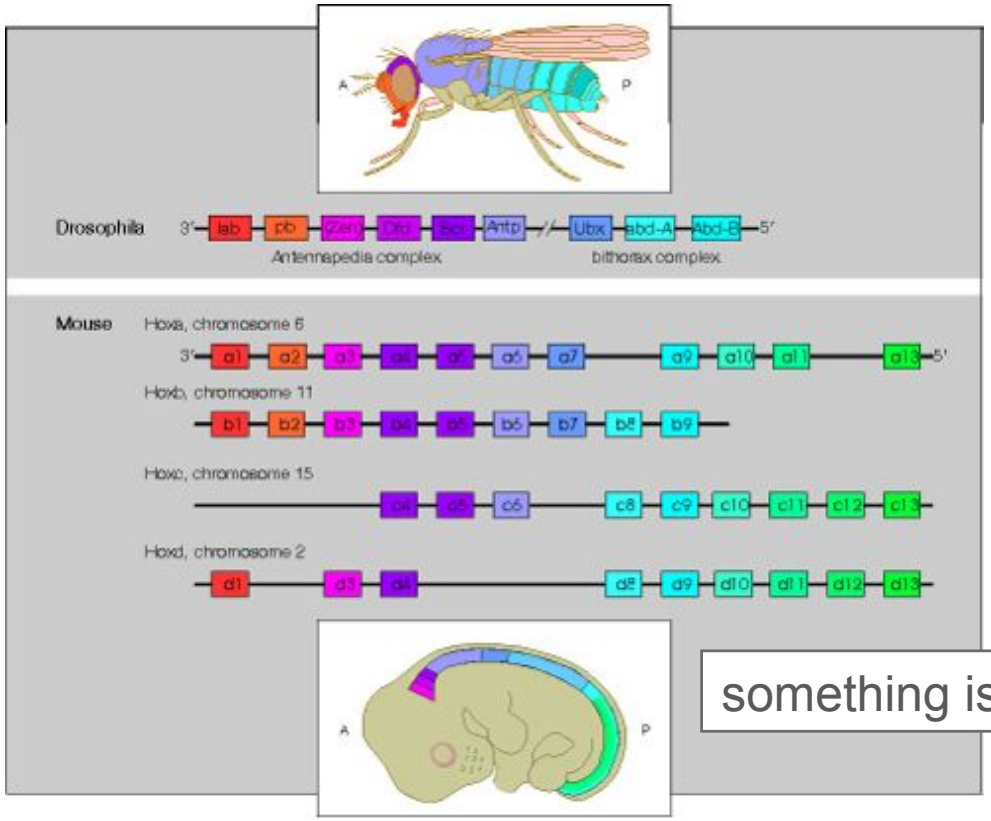
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Discuss with your neighbor

- What is more frequent: point mutations or duplications?
- **Where in animal evolution were there genome duplications?**

History of an idea:

Susumo Ohno (1970): observation of higher haploid DNA content in mammals than Ciona
-> hypothesis of two whole genome duplications which would account for "higher complexity"

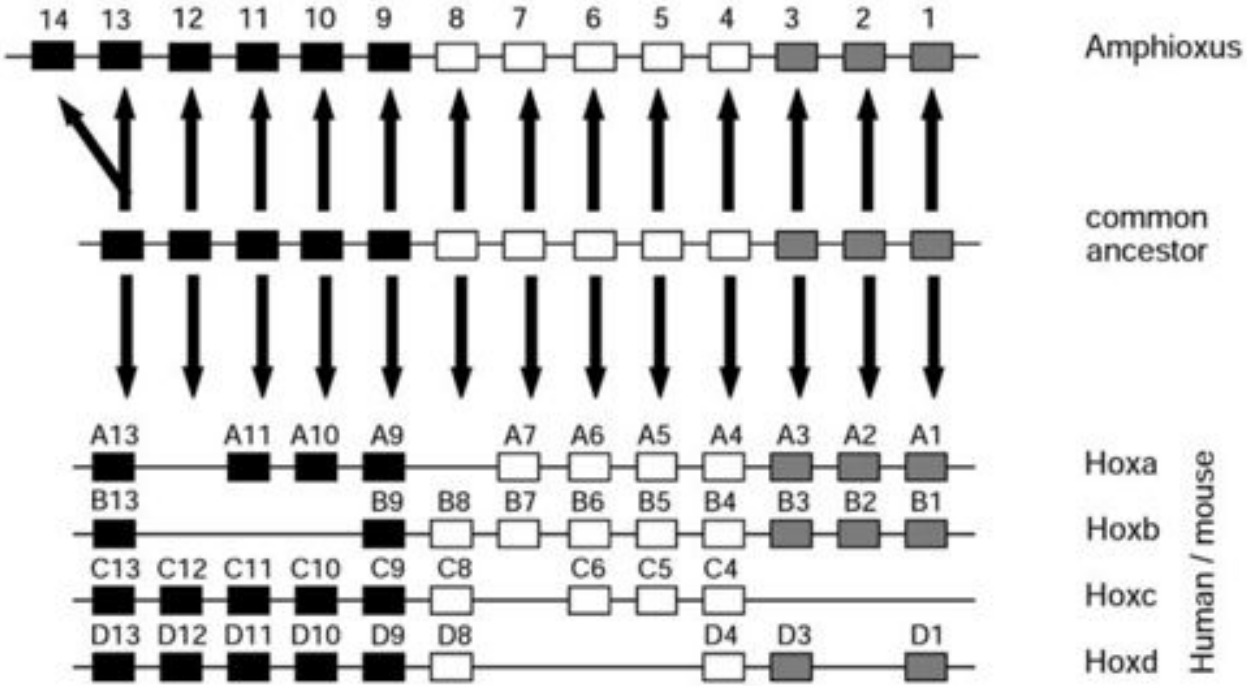


something is going on

http://www.mun.ca/biology/desmid/brian/BIOL3530/DB_Ch15/BIOL2900_EvoDevo.html

Two rounds of whole genome duplication in vertebrates

Garcia-Fernandez & Holland (1994): One Hox cluster in Amphioxus
-> duplications from 1 to 4 were vertebrate specific



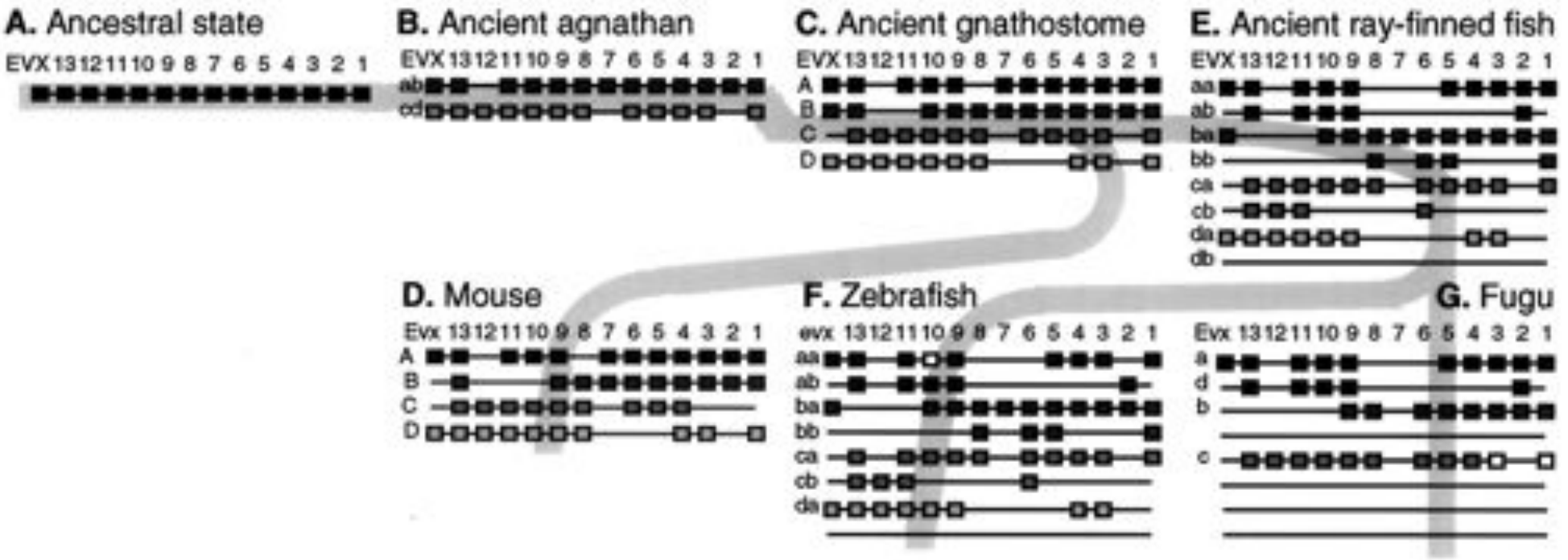
"2R hypothesis"

Holland (2003) J Struct Funct Genomics 3: 75–84

Spring (1997): many examples of gene families with 1 ortholog in Drosophila, 4 in human
but
anecdotal evidence
no phylogenetic or syntenic test

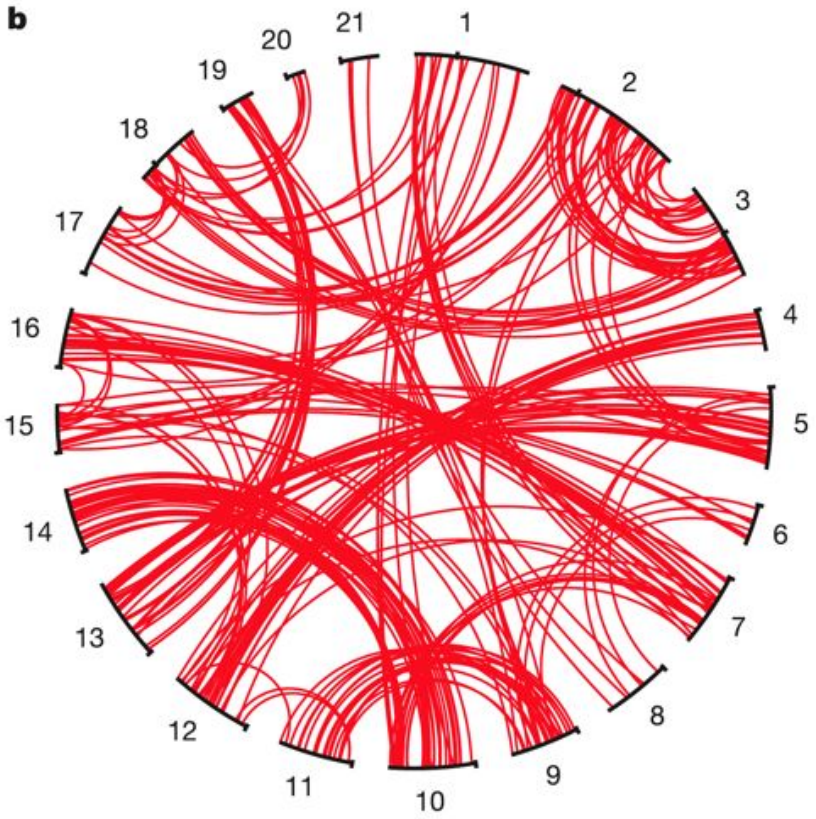
And more genes in fishes

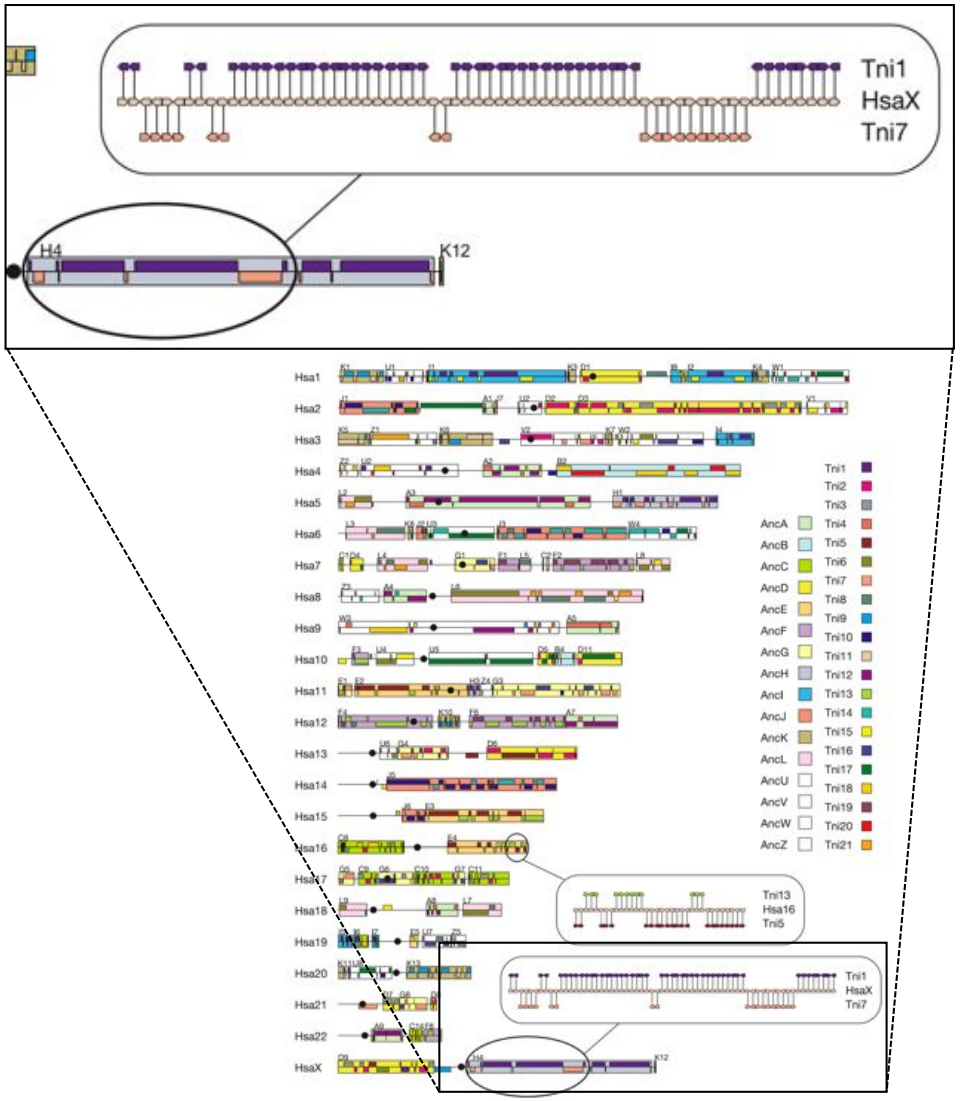
More Hox in fishes



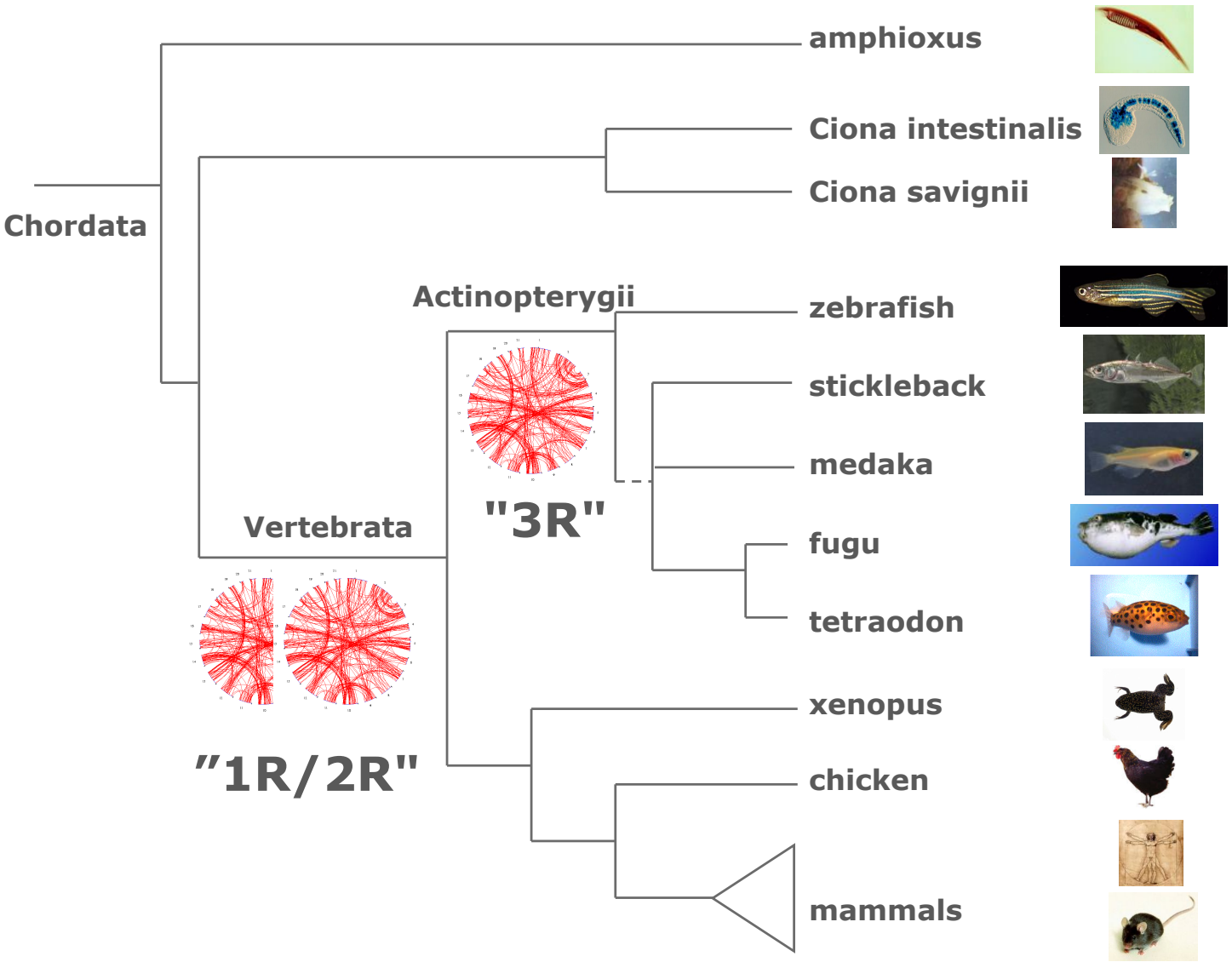
Amores et al. 1998 Science 282: 1711-1714

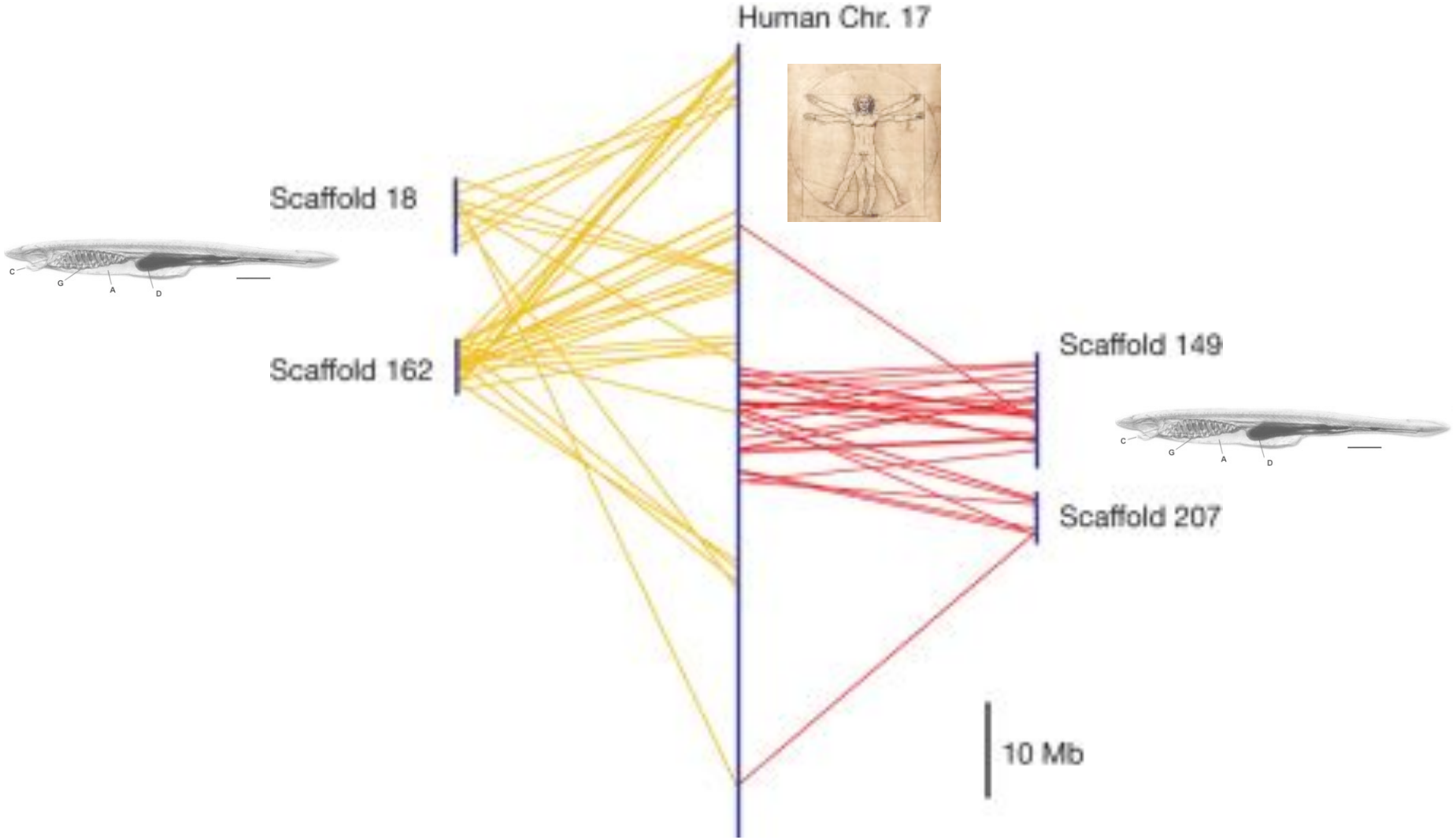
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics


b
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21


Tni1
HsaX
Tni7
H4
K12
Hsa1
Hsa2
Hsa3
Hsa4
Hsa5
Hsa6
Hsa7
Hsa8
Hsa9
Hsa10
Hsa11
Hsa12
Hsa13
Hsa14
Hsa15
Hsa16
Hsa17
Hsa18
Hsa19
Hsa20
Hsa21
Hsa22
HsaX
AncA
AncB
AncC
AncD
AncE
AncF
AncG
AncH
AncI
AncJ
AncK
AncL
AncU
AncV
AncW
AncZ
Tni1
Tni2
Tni3
Tni4
Tni5
Tni6
Tni7
Tni8
Tni9
Tni10
Tni11
Tni12
Tni13
Tni14
Tni15
Tni16
Tni17
Tni18
Tni19
Tni20
Tni21
Tni13
Hsa16
Tni5
Tni1
HsaX
Tni7

To understand vertebrates, study amphioxus

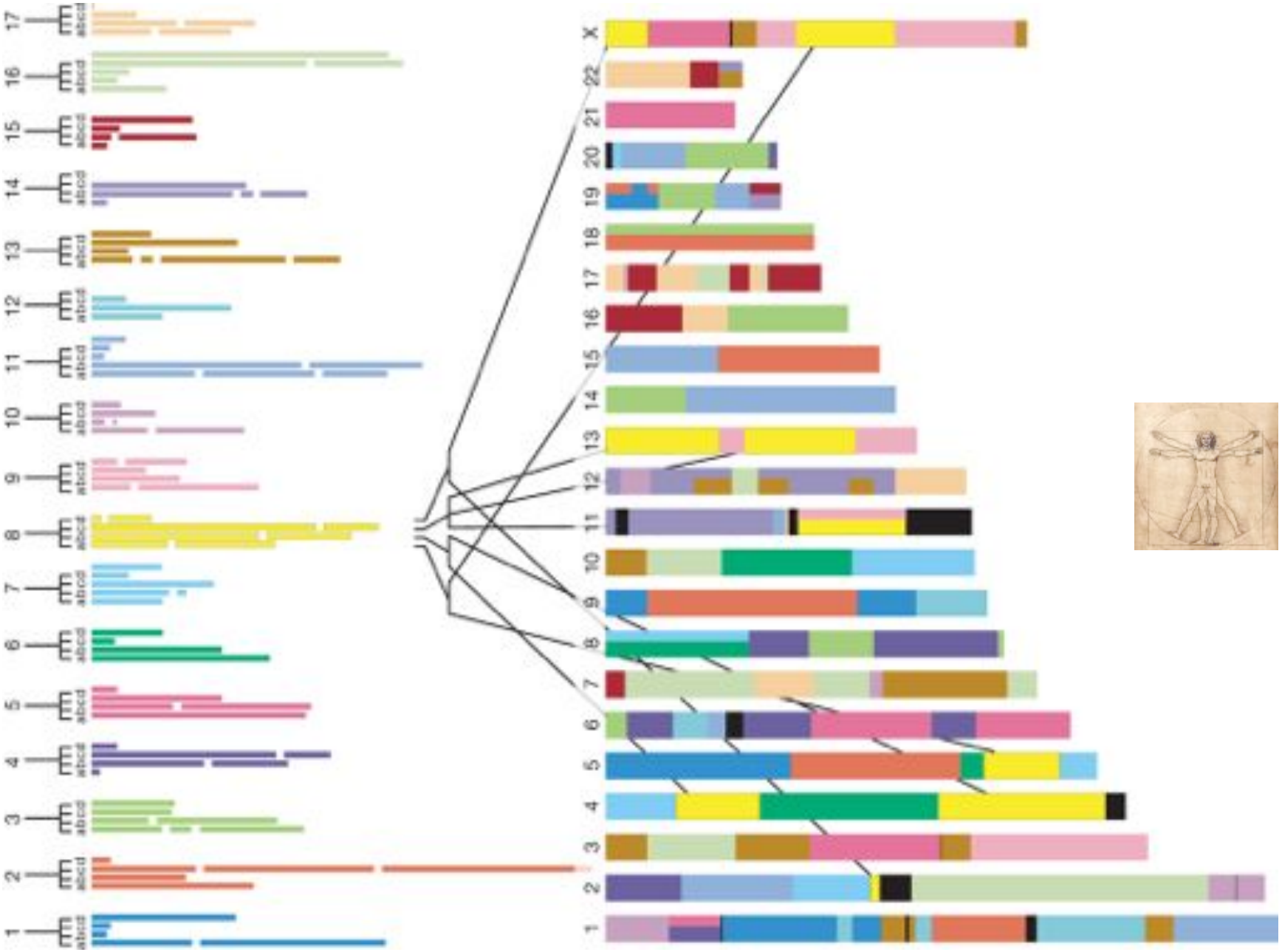

amphioxus
Ciona intestinalis
Ciona savignii
Chordata
Actinopterygii
zebrafish
stickleback
medaka
"3R"
Vertebrata
fugu
tetraodon
xenopus
chicken
"1R/2R"
mammals



Putnam et al. 2008 Nature 453, 1064-1071

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

ancestor



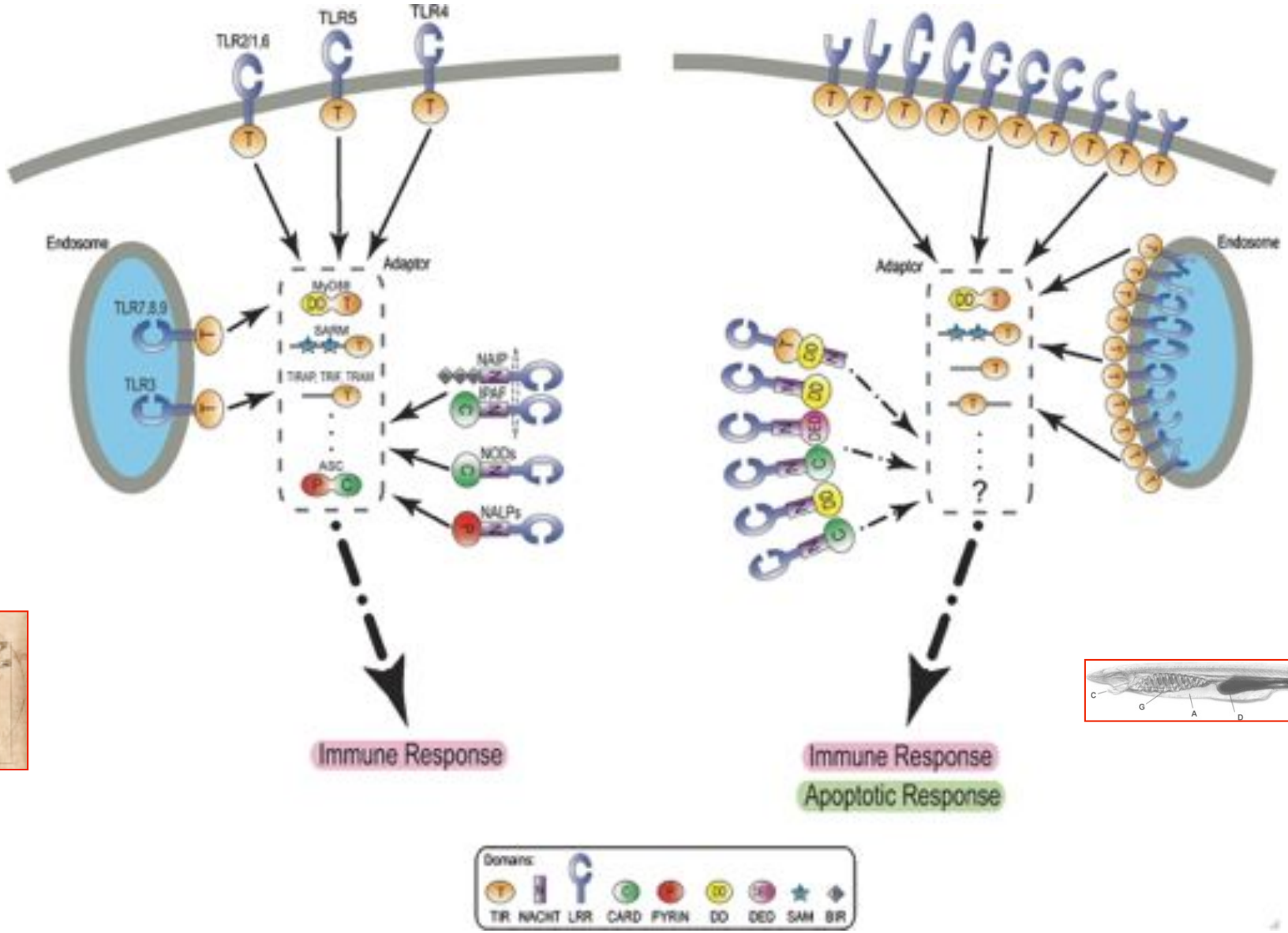
Putnam et al. 2008 Nature 453, 1064-1071

UNIL | Université de Lausanne
Département d'écologie et évolution

SIB Swiss Institute of Bioinformatics

It's not because it's amphioxus that it's always simple

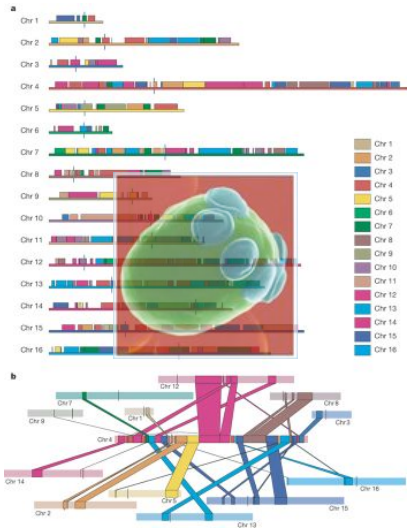
Expansion of innate immunity genes in amphioxus (right) compared with the human system (left).



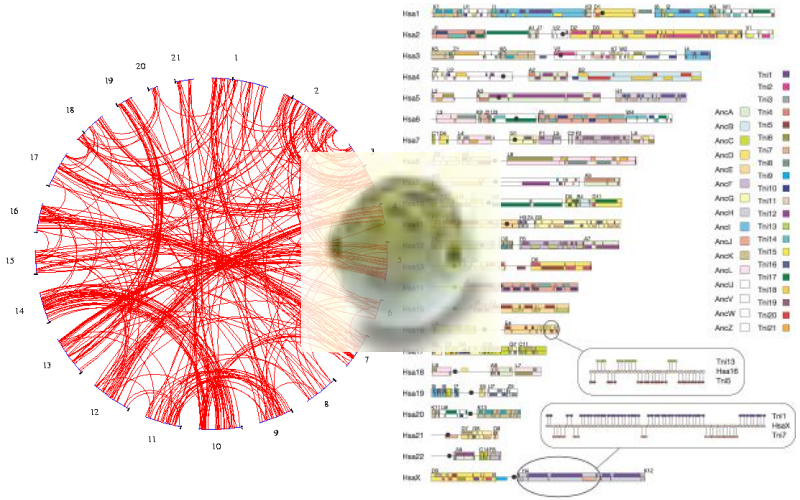
Holland L Z et al. Genome Res. 2008;18:1100-1111

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

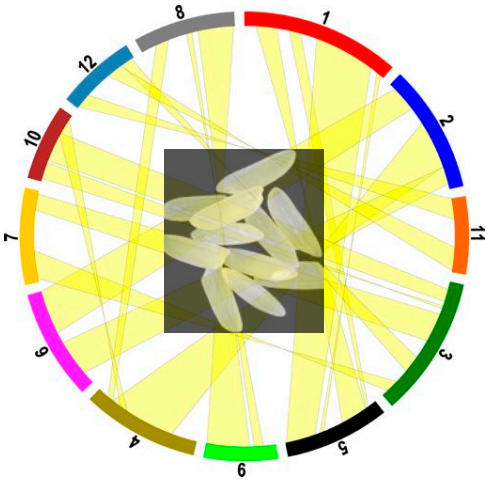
Bref, whole genome duplications abund



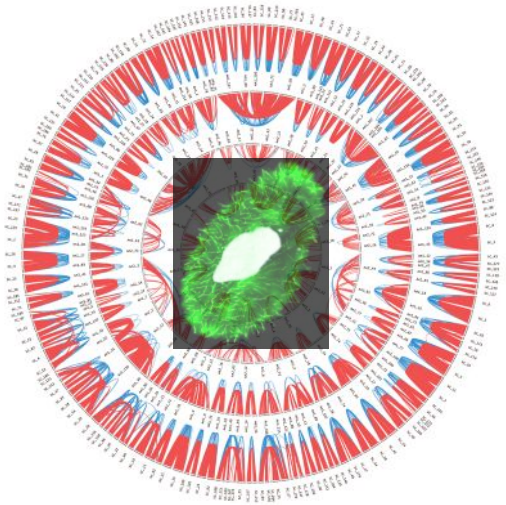
Kellis et al 2004 Nature 428:617-624



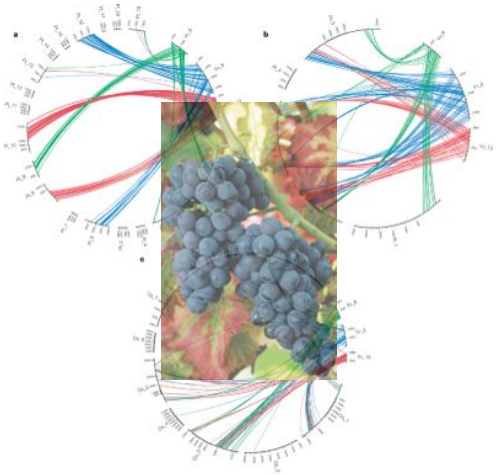
Jaillon et al. 2004 Nature 431:946-957



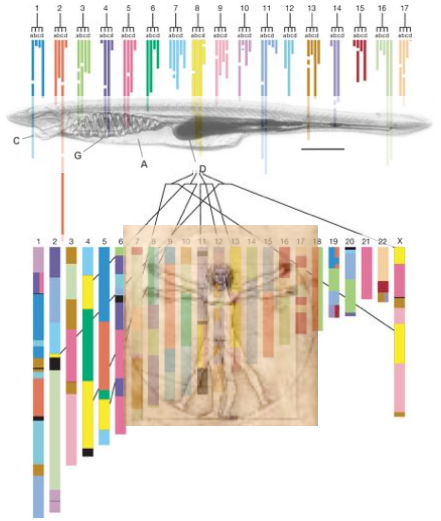
Yu et al 2005 PLoS Biol 3:e38



Jaillon et al 2006 Nature 444, 171-178

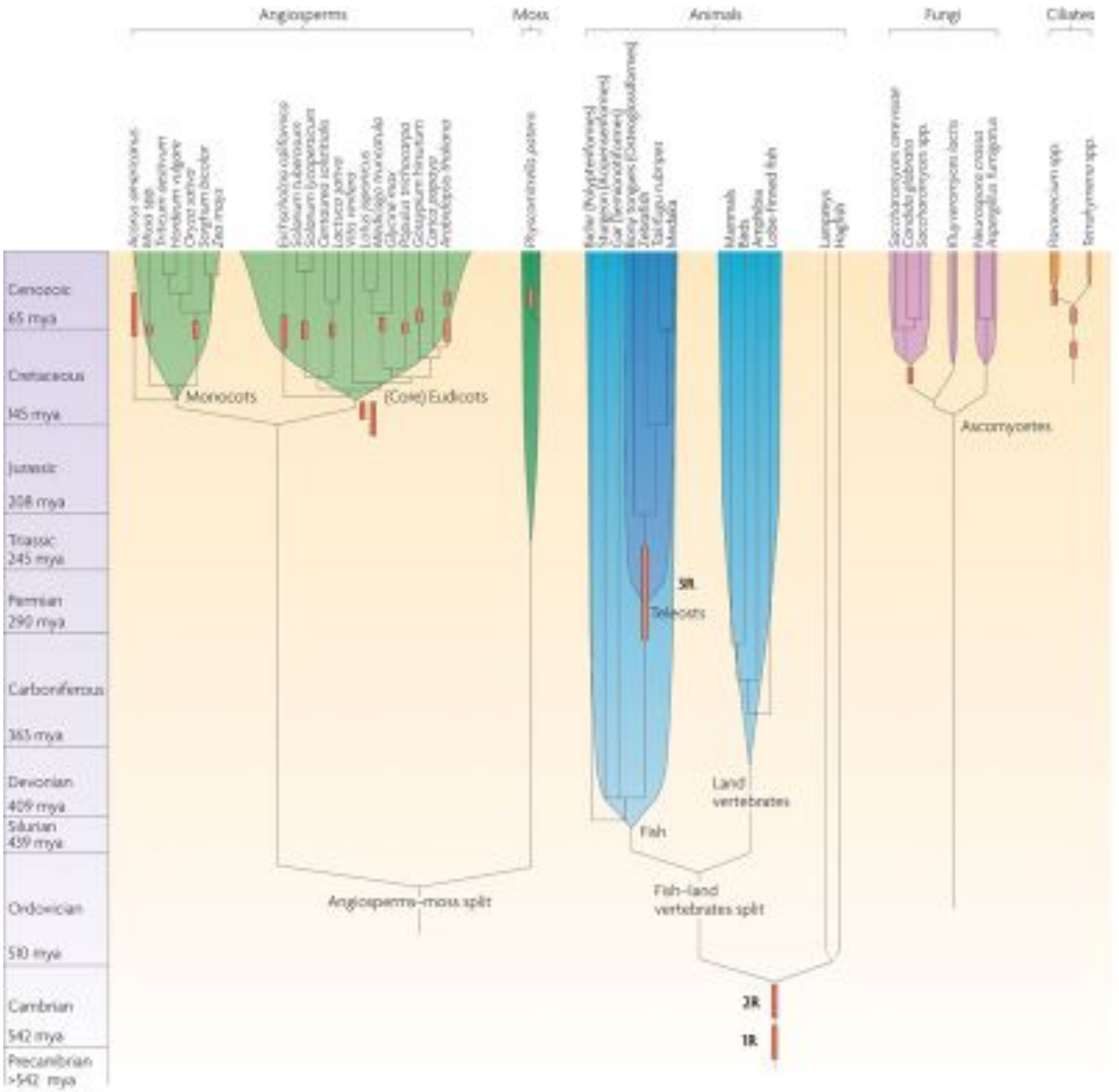


Jaillon et al. 2007 Nature 449, 463-467



Putnam et al 2008 Nature 453:1064-1071

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics



Nature Reviews | Genetics

Van de Peer et al. (2009) Nature Rev Genet 10:725-732

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Let's work!

pick one of FRE1-8; examine region in:
http://ygob.ucd.ie/

pick a human gene from list, examine region in:
http://www.genomicus.biologie.ens.fr/

Consequences of duplication

Discuss with your neighbor

- What is more frequent: point mutations or duplications?
- Where in animal evolution were there genome duplications?
- **Two reasons why we should distinguish orthologs and paralogs**

Ohno (1970): after duplication, either one copy gains a new function, or one copy is lost

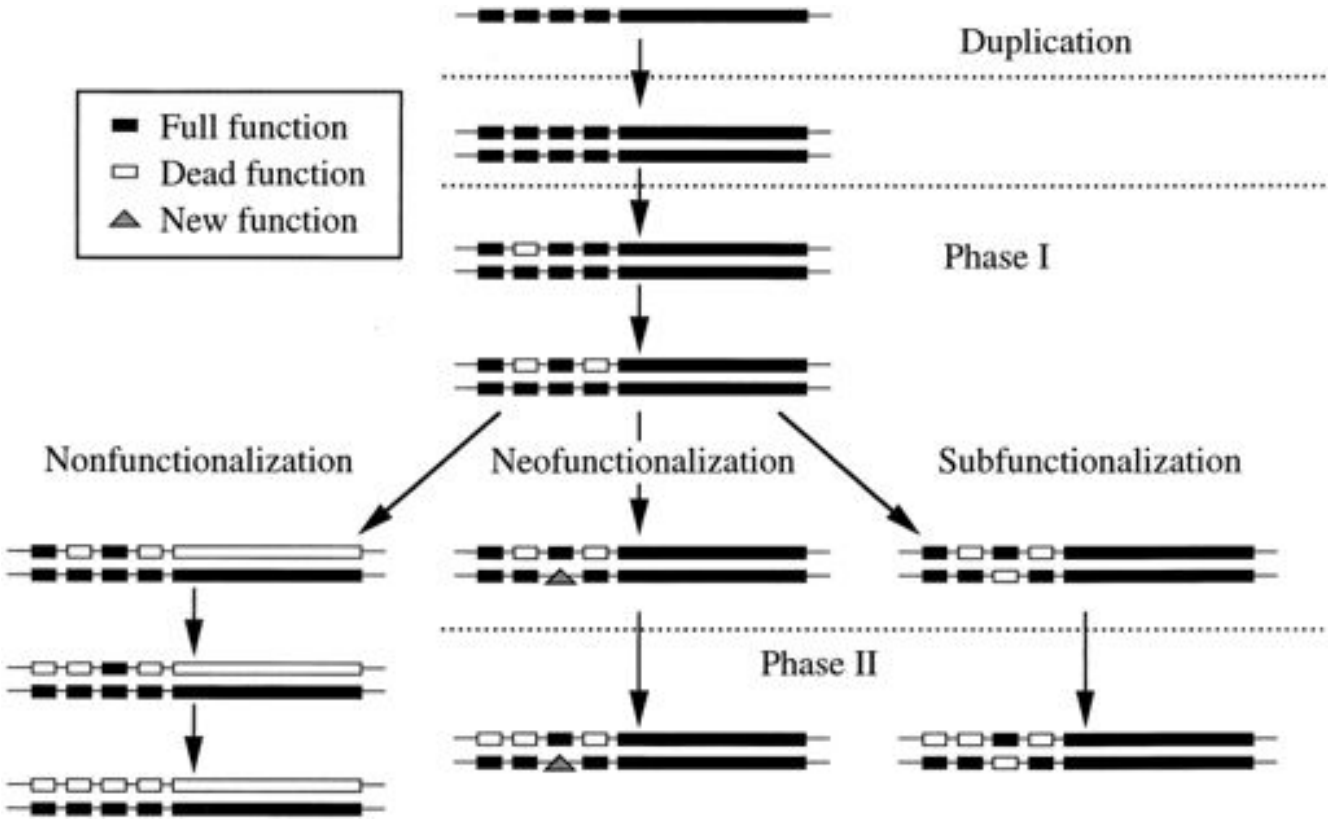
nonfunctionalization neofunctionalization subfunctionalization

Lynch lab (1999-2001): probability of new functional mutation before loss too low relative to frequency of duplicate genes

Schubert et al. (2006) TREE 21: 269-77

DDC model = Duplication Degeneration Complementation

Original model based on regulatory regions (i.e. gene expression patterns):

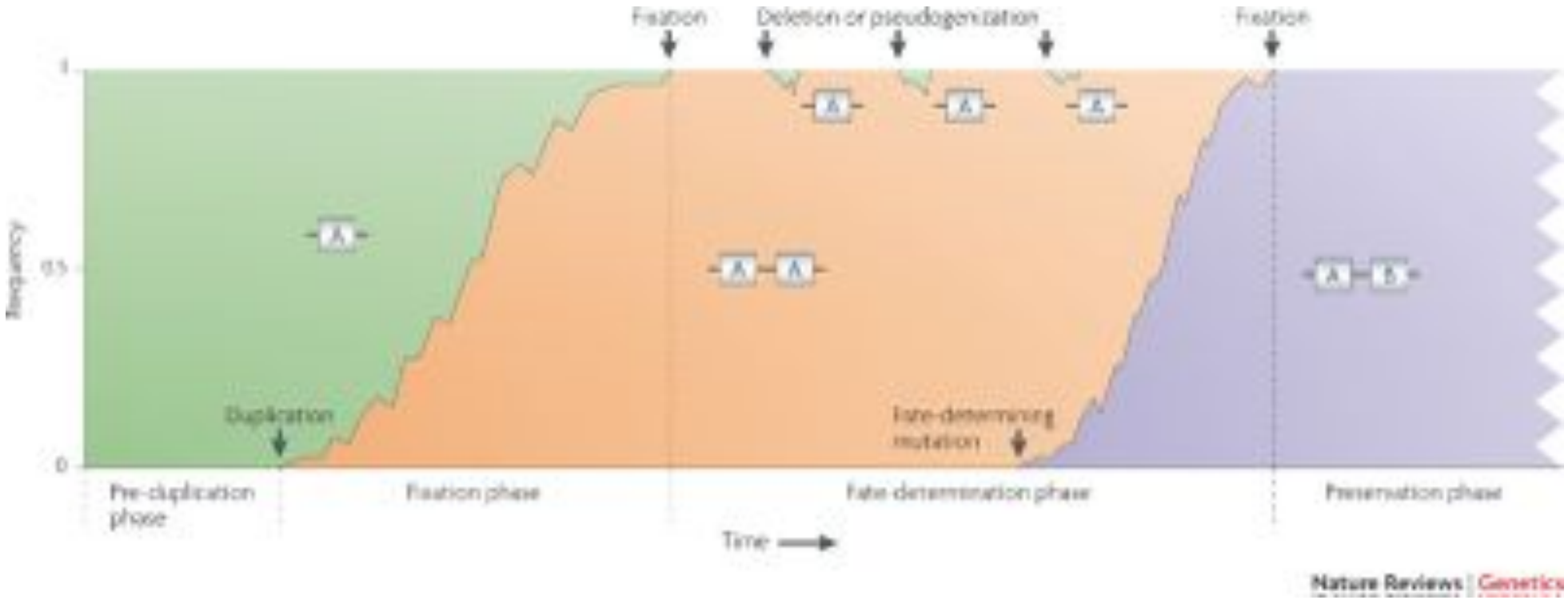


Force et al. (1999) Genetics 151: 1531-45

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Conant & Wolfe 2008 Nature Rev Genet 9:938-950

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics



Innan and Kondrashov 2010 Nature Reviews Genetics 11, 97-108

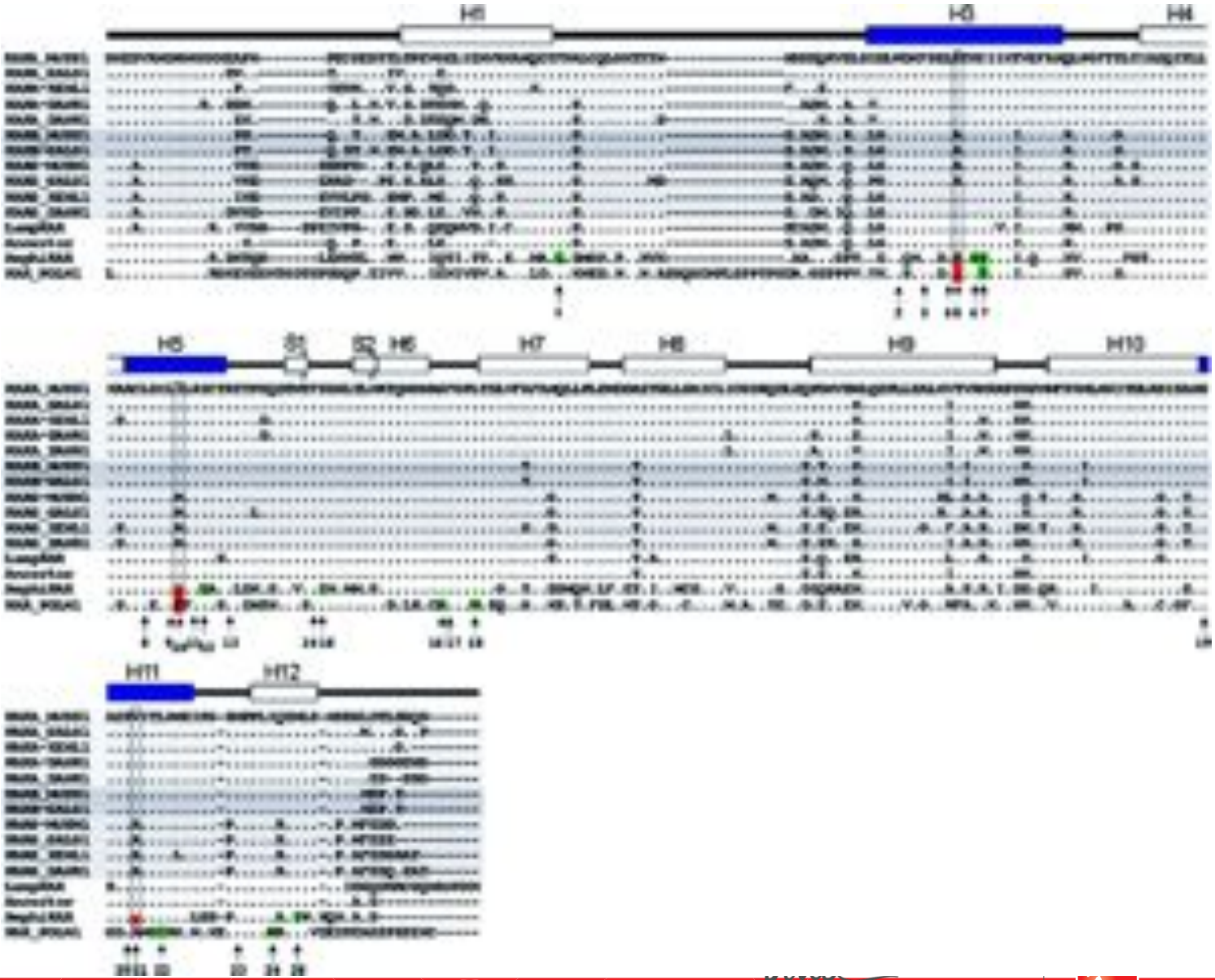
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Forms of functional divergence

Discuss with your neighbor

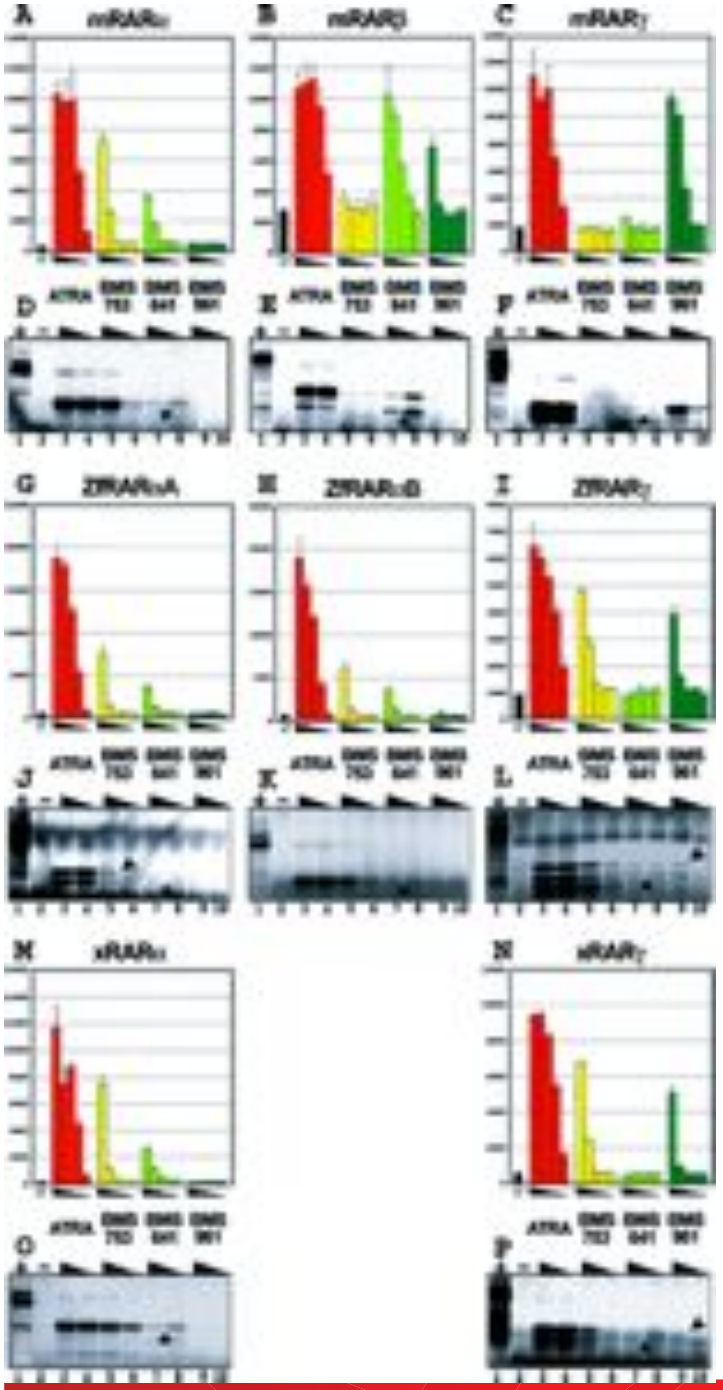
- What is more frequent: point mutations or duplications?
- Where in animal evolution were there genome duplications?
- Two reasons why we should distinguish orthologs and paralogs
- **What is gene function?**

Sequence divergence: Biochemical function



Escriva et al. 2006 PLoS Genet 2: e102

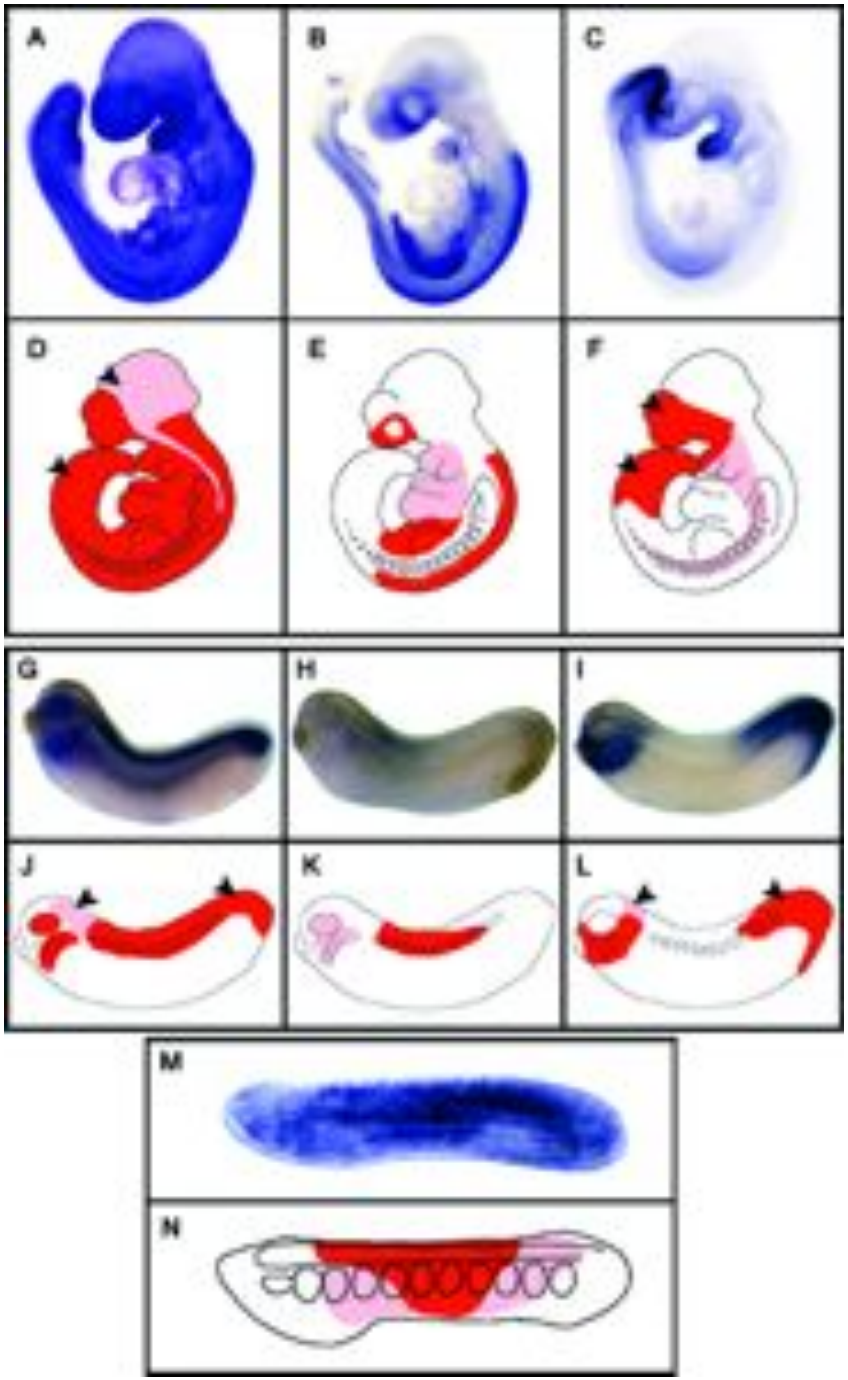
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics



Transcriptional Activity and Binding Selectivity of Vertebrate RARs

Escriva et al. 2006 PLoS Genet 2: e102

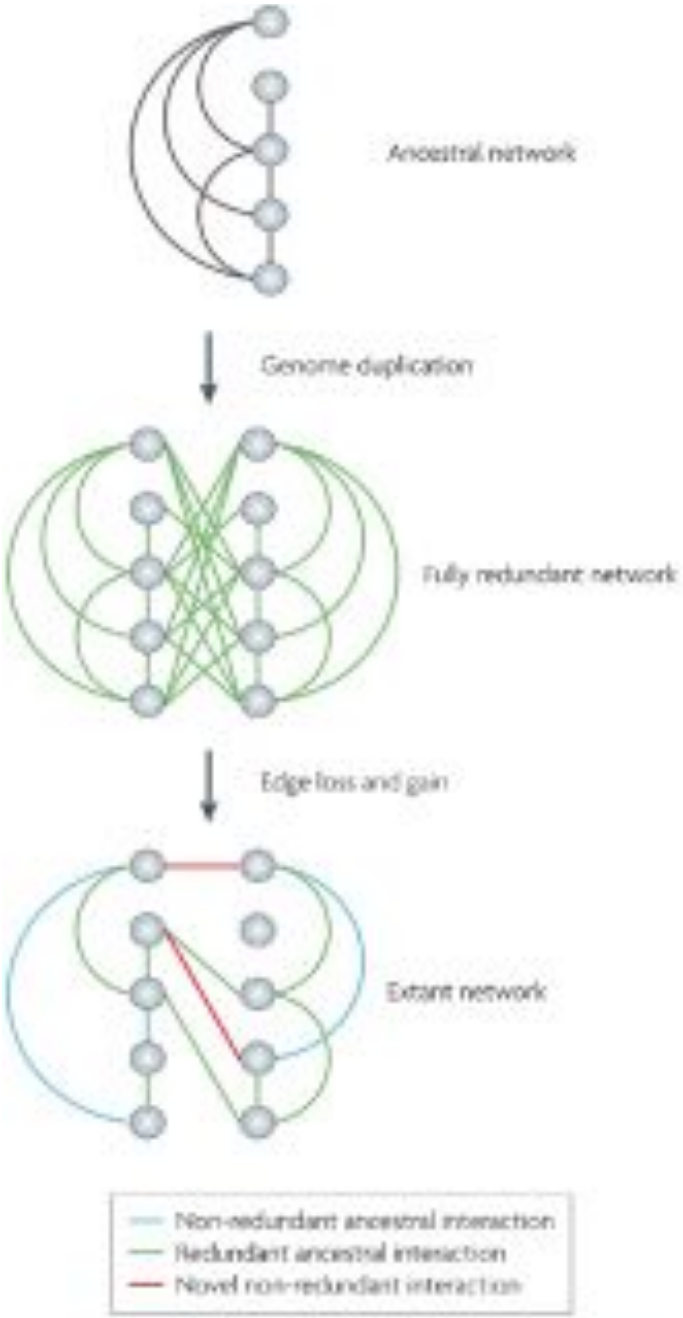
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics



Staining of embryos indicates expression of mRARα (A), mRARβ (B), and mRARγ (C) in mouse embryos at E9; of xRARα (G), xRARβ (H), and xRARγ (I) in stage 30 Xenopus embryos, and of AmphiRAR (M) in 20 h old amphioxus larvae. Schematic representations are shown of the expression territories of mRARs (D–F), xRARs (J–L), and AmphiRAR (N) in mouse, Xenopus, and amphioxus embryos, respectively. Regions with high levels of expression are red and those with lower levels of expression are pink. Arrowheads indicate regions in mouse and Xenopus embryos where the RAR expression cannot be correlated with AmphiRAR expression and can be described as "new expression territories."

Escriva et al. 2006 PLoS Genet 2: e102

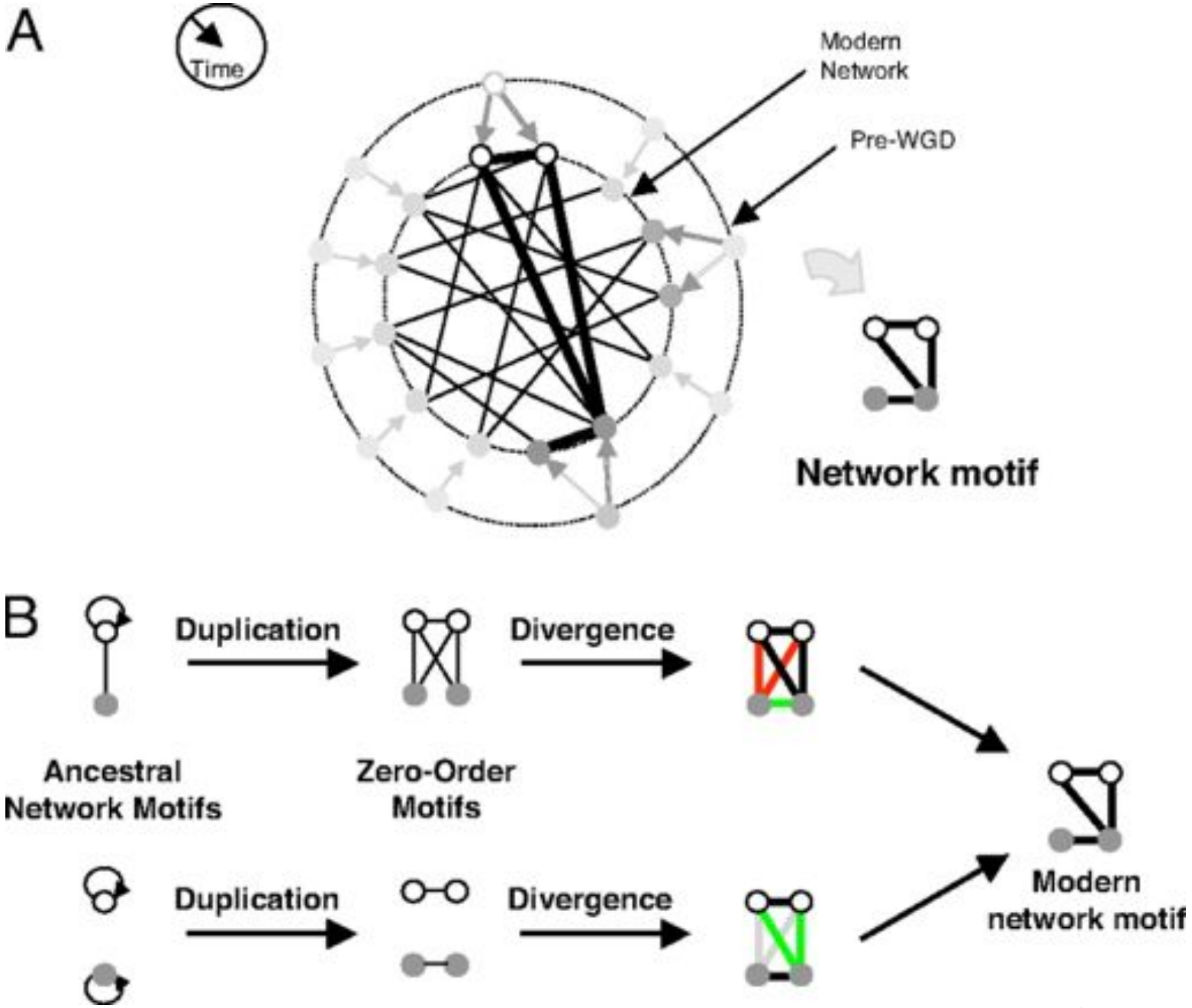
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics



Nature Reviews | Genetics

Conant and Wolfe 2008 Nature Reviews Genetics 9, 938-950

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics



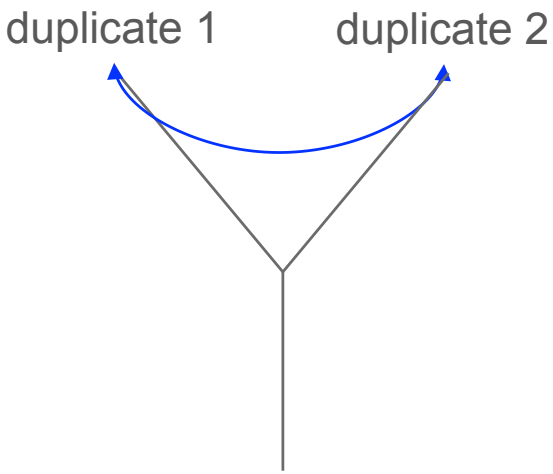
Presser A et al. PNAS 2008;105:950-954

UNIL | Université de Lausanne
Département d'écologie
et évolution

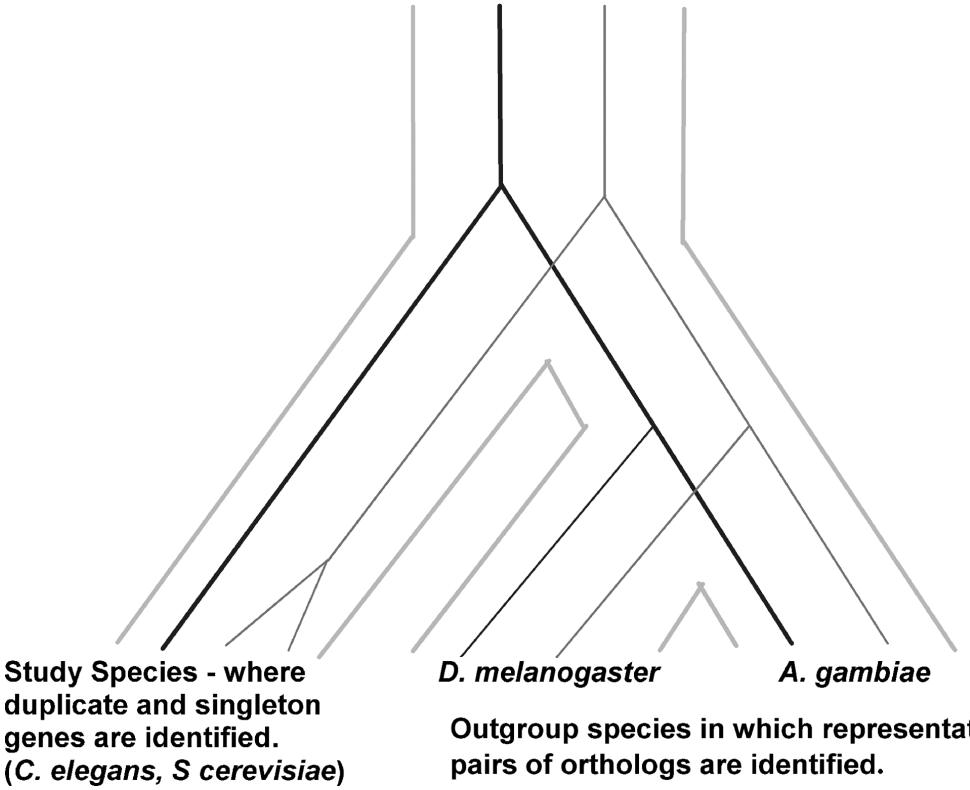
SIB
Swiss Institute of
Bioinformatics

Not all genes are kept in duplicate

Are genes kept in duplicate biased relative to selective pressure and evolutionary rates?

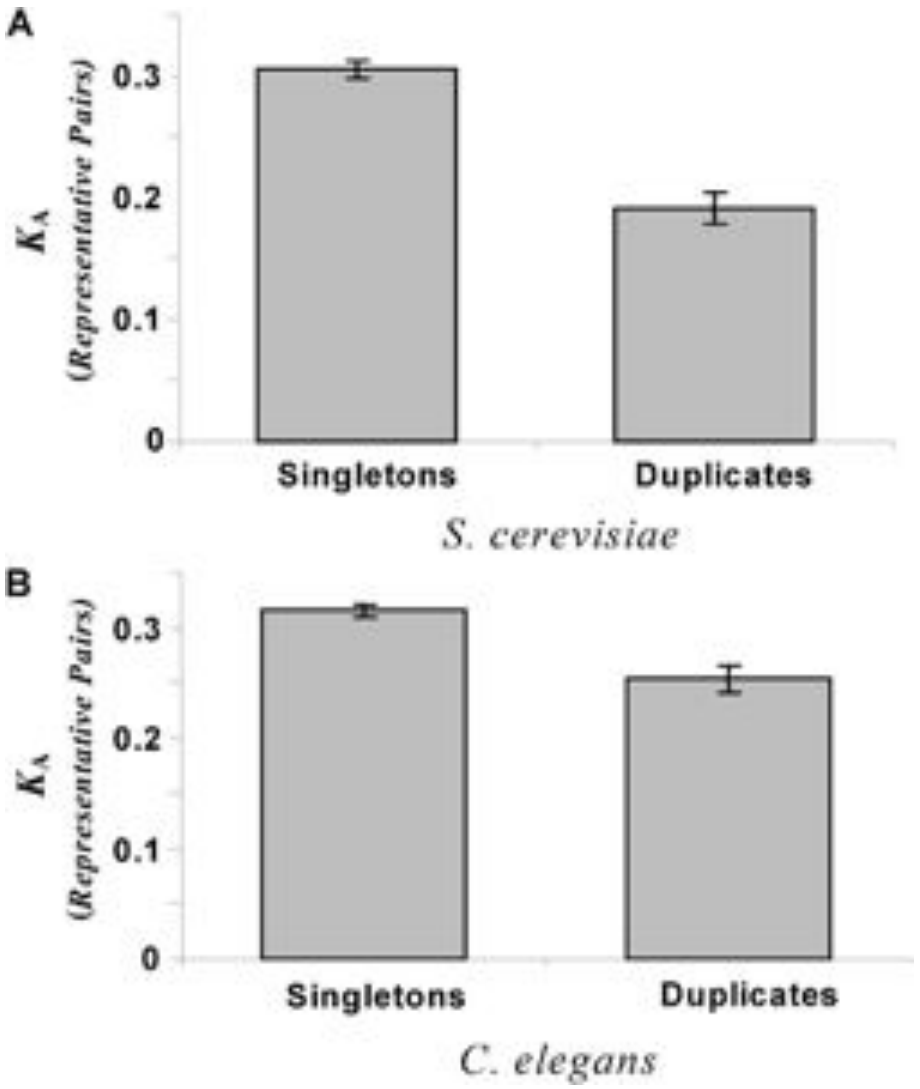


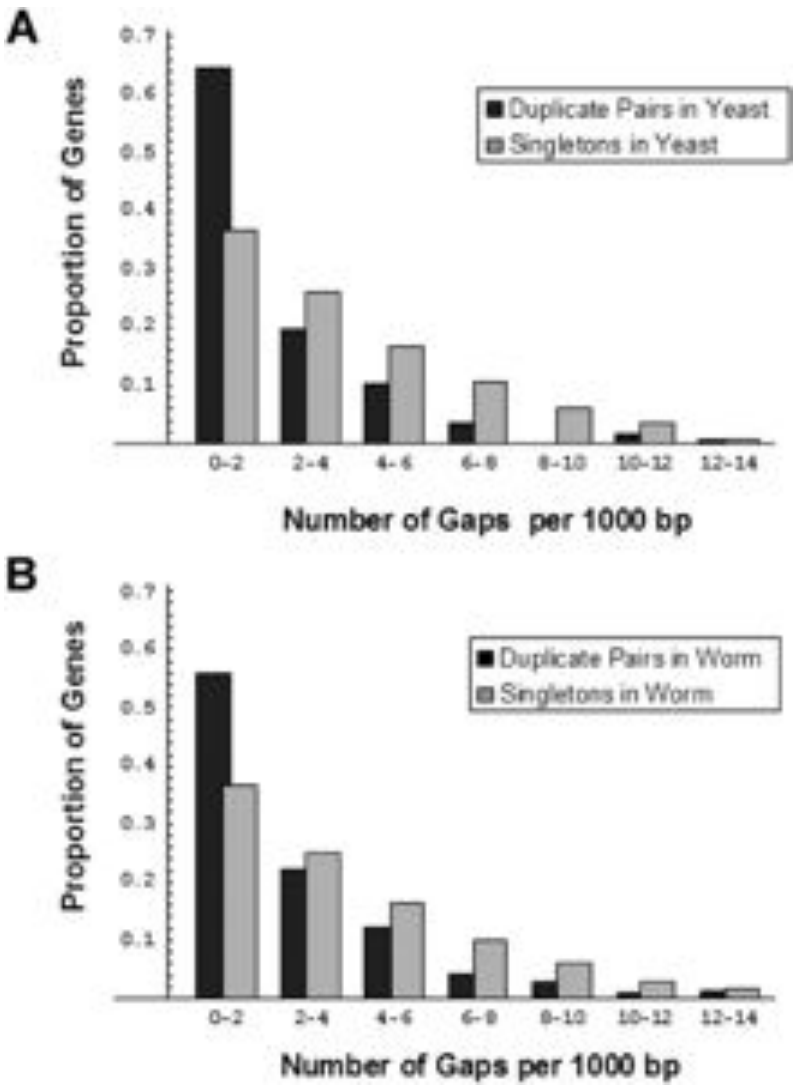
Ka = original rate of gene
+ possible change due to duplication



Ka Drosophila / Anopheles = independent of changes after duplication in yeast or nematode

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics





UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Comparison	*S. cerevisiae*		*C. elegans*	
	Spearman Correlation (r_{ab})	Partial Correlation Coefficient (r_{abc})	Spearman Correlation (r_{ab})	Partial Correlation Coefficient (r_{abc})
Class[a] versus CAI	0.40***	0.37***	0.09*	0.12*
Class[a] versus K_S	−0.18***	−0.07[ns]	0.08[ns]	0.11*
K_S versus CAI	−0.31***	−0.26***	−0.27***	−0.28***

Significance was tested for the direct and partial correlation coefficients using the statistics $t_r = r\sqrt{\frac{n-2}{1-r^2}}$ and $t_r = r\sqrt{\frac{n-2-m}{1-r^2}}$, respectively, where n is the sample size, m is the number of variables held constant, and r is the rank correlation coefficient (Sokal and Rohlf 1995).

[a]For this parameter, representative pairs were given a value of either 0 (for a singleton) or 1 (for a duplicate).

NS, nonsignificant; *, $p = 0.05$; **, $p = 0.01$; ***, $p = 0.001$.

DOI: 10.1371/journal.pbio.0020055.t002

In yeast, biases differ between WGD and smaller scale duplications (SSD):

Table 1. Evolutionary rate estimates for the three sets of study genes obtained by measuring the divergence between outgroup orthologs

Gene set	Nematode outgroup lineages[b]		Fly outgroup lineages[c]	
	K_A (all genes)	K_A (without ribsomal genes)	K_A (all genes)	K_A (without ribsomal genes)
Singletons	0.095	0.097	0.075	0.077
WGD duplicates	0.061 **[d]	0.089 *	0.045 **	0.061 *
SSD duplicates	0.079 **	0.087 *	0.063 **	0.070 *

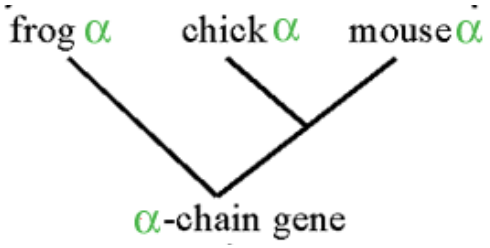
[a] The gene studied included singltons, WGD duplicates and SSD duplicates. Both types of duplicates appear to arise from conserved genes.
[b]Pairs of outgroup orthologs were identified in *C. elegans* and *C. briggsae* where possible (see methods in supplementary material online) and mean divergence between these pairs for each group, both with and without ribosomal proteins is shown.
[c]Similar mean divergence estimates for representative pairs from *D. melanogaster* and *D. pseudoobscura* are shown.
[d]Significance levels are * $P<0.05$, ** $P<0.01$ for Mann-Whitney U-test, compared with the set of singleton genes.

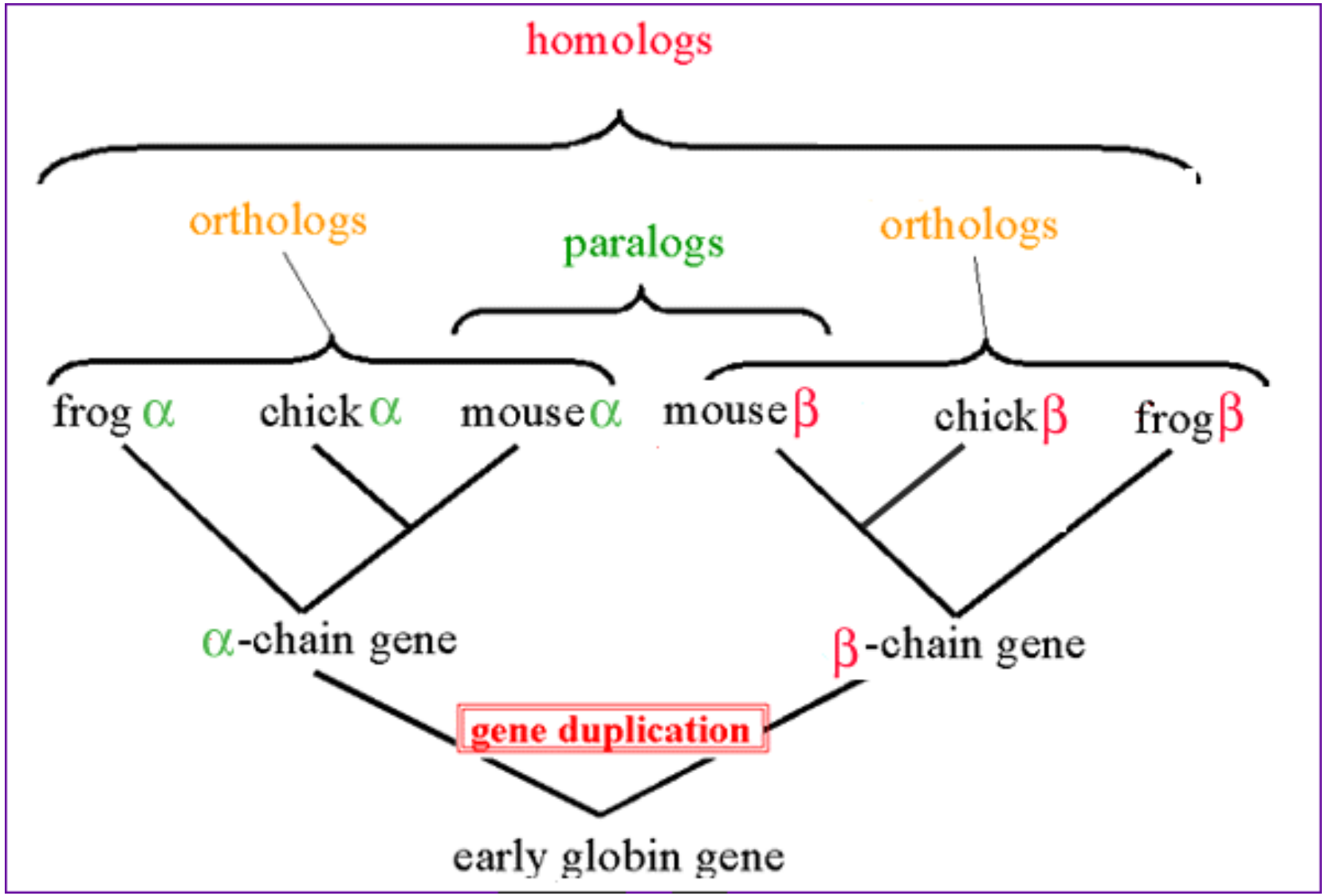
Table 2. The percentage of genes in each paralogy class belonging to each high-level GO functional category[a]

GO function	Including ribosomal genes			Without ribosomal genes		
	Singletons	WGD pairs	SSD pairs	Singletons	WGD pairs	SSD pairs
Catalytic	48.54%	45.21%	48.74%	51.64%	63.06% *[b]	54.19% ns
Binding	24.73%	21.00%	26.38%	26.19%	21.66% ns	28.77% ns
Transcription regulator	10.48%	9.13%	4.77%	11.24%	12.74% ns	5.31% *
Structural molecule	10.39%	30.59%	11.81%	9.54%	4.46% *	10.89% ns
Transporter	8.90%	3.20%	9.81%	3.92%	3.18% ns	1.96% ns
Enzyme regulator	3.56%	2.74%	6.78%	3.82%	3.82% ns	7.54% *
Signal transducer	0.69%	1.83%	2.01%	1.80%	1.91% ns	1.40% ns

[a]Only the seven largest functional classes are shown, genes of unknown function have been removed. The total percentages can exceed 100% because a gene can be annotated as belonging to more than one class.
[b]Significance levels are ** $P<0.001$, * $P<0.05$ and not significant (ns) compared with the set of singleton genes.

Orthologs vs. paralogs





http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html

Homologs are most commonly defined as orthologs, paralogs, or xenologs.

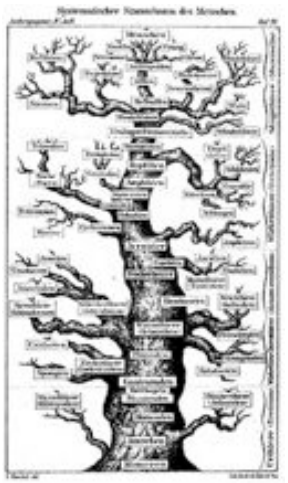
Orthologs are homologs produced by speciation—they represent genes derived from a common ancestor that diverged because of divergence of the organism. Orthologs tend to have similar function.

Paralogs are homologs produced by gene duplication and represent genes derived from a common ancestral gene that duplicated within an organism and then diverged. Paralogs tend to have different functions.

Xenologs are homologs resulting from the horizontal transfer of a gene between two organisms. The function of xenologs can be variable, depending on how significant the change in context was for the horizontally moving gene. In general, though, the function tends to be similar.

http://www.ncbi.nlm.nih.gov/About/primer/phylo.html

"Ortholog conjecture"

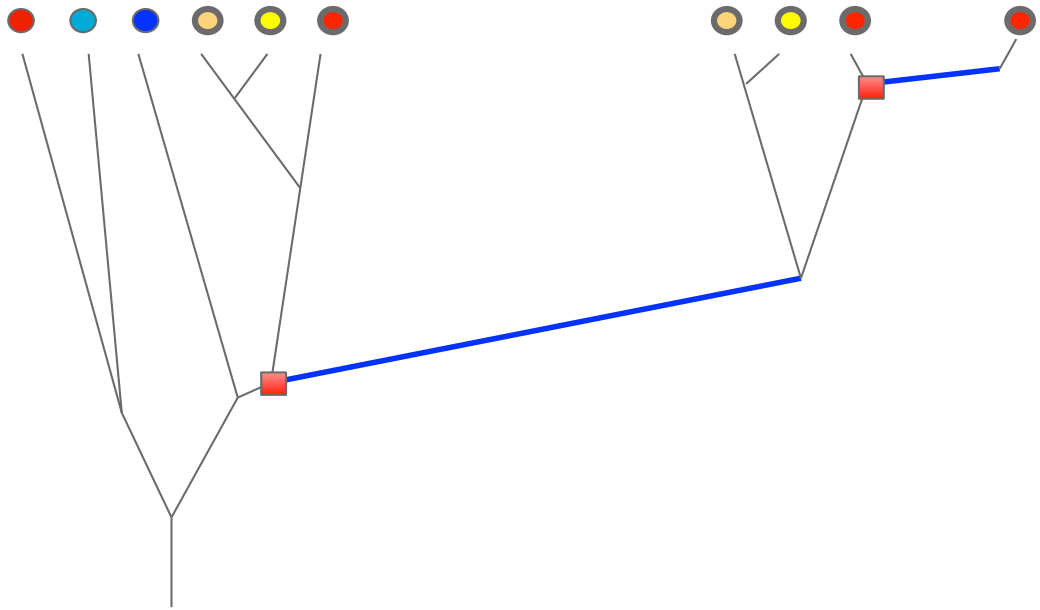


UNIL | Université de Lausanne
Département d'écologie et évolution

SIB
Swiss Institute of Bioinformatics

This should not be an assumption
It should be an hypothesis to test

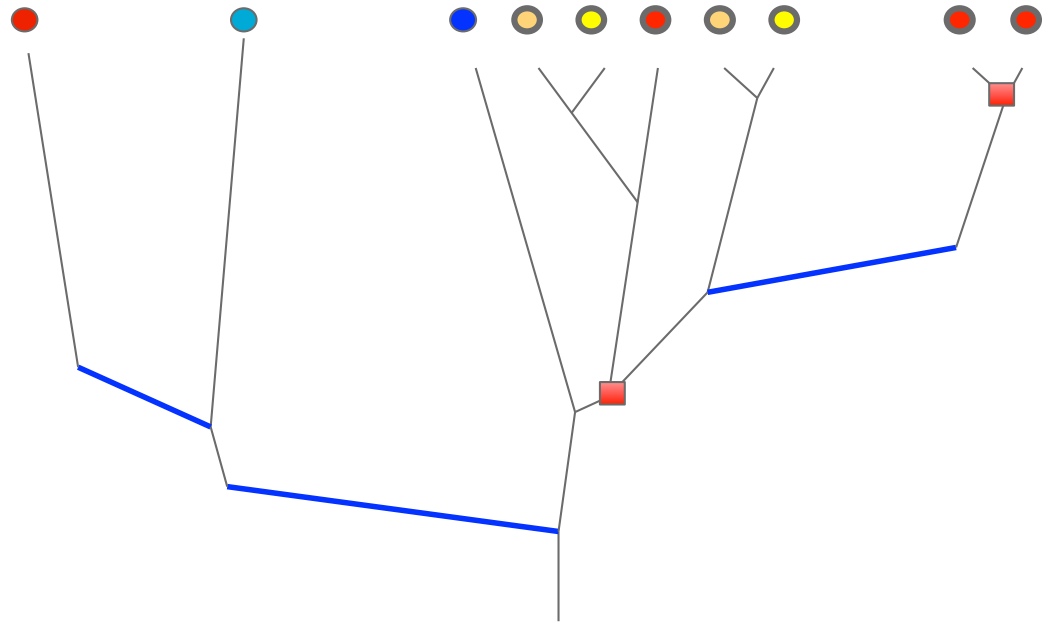
ortholog conjecture:
functional change after duplication



Studer and Robinson-Rechavi 2009 Trends in Genetics

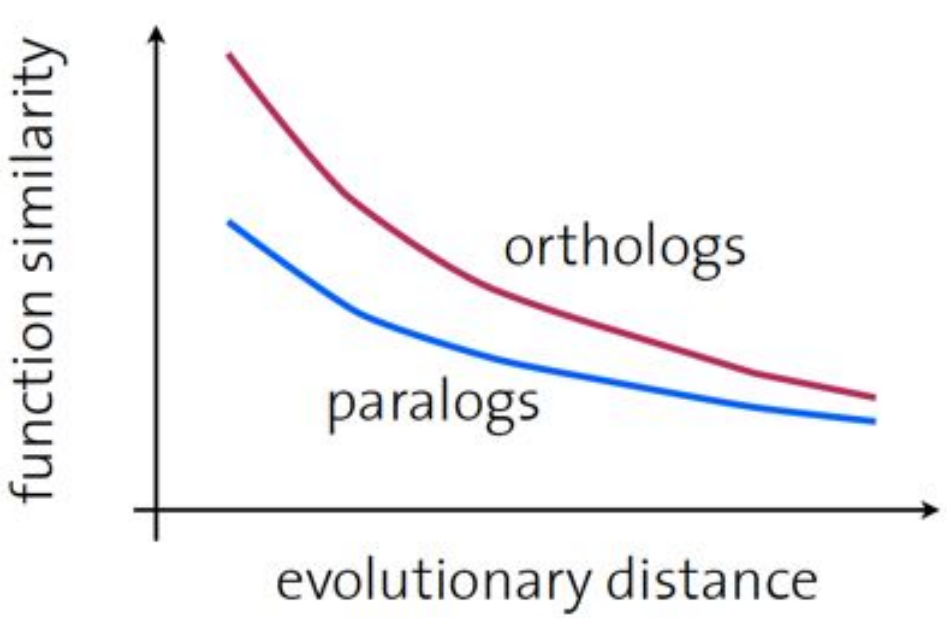
Alternative hypothesis:

Uniform distribution of positive selection

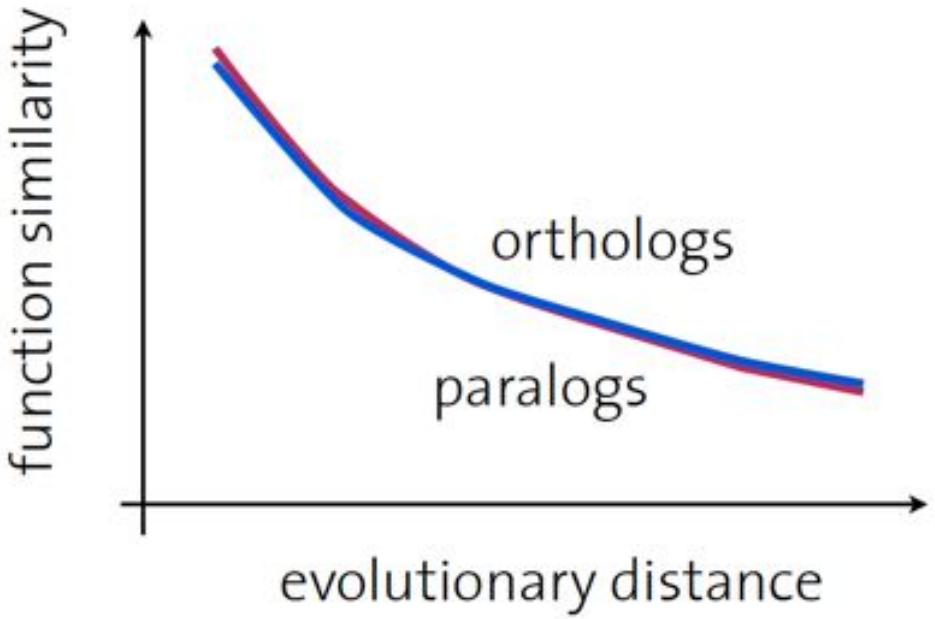


Studer and Robinson-Rechavi 2009 Trends in Genetics

Another way of putting it



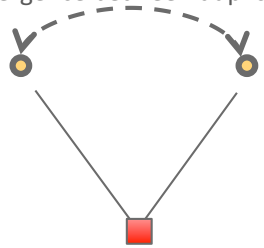
ortholog conjecture



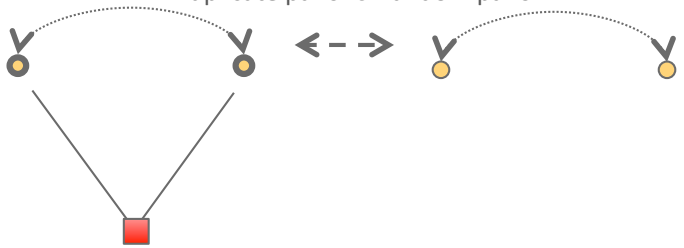
uniform model

Study design

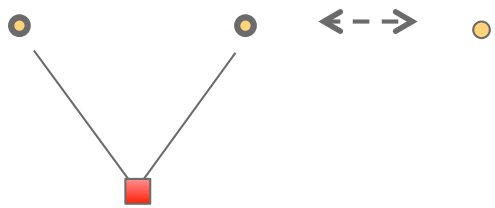
I. Divergence between duplicates



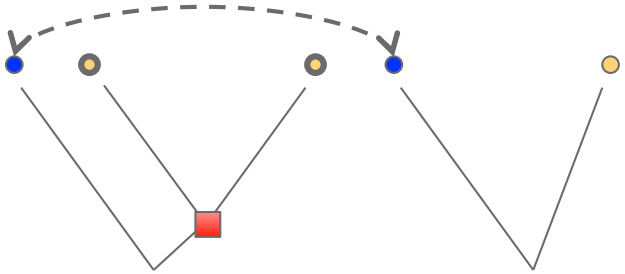
II. Duplicate pairs *vs.* random pairs



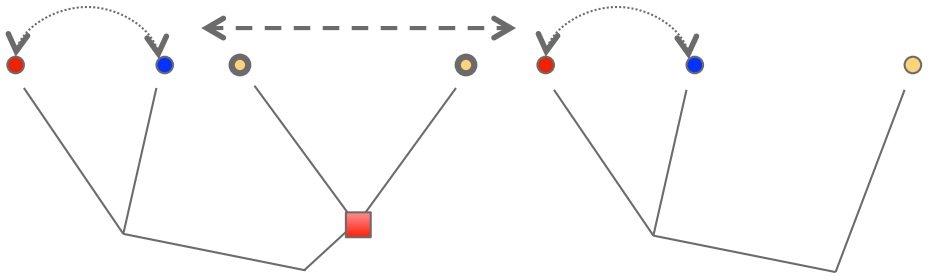
III. Duplicates *vs.* singletons



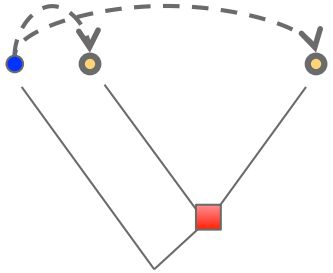
IV. Orthologs of duplicates *vs.* orthologs of singletons



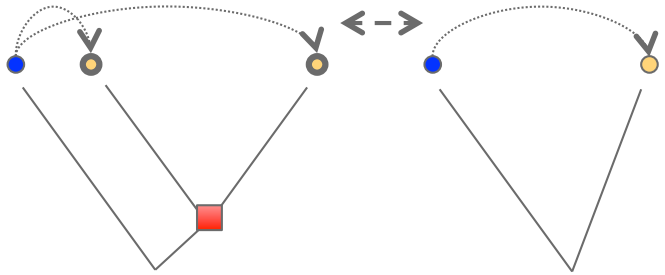
V. Divergence between orthologs of duplicates *vs.* between orthologs of singletons



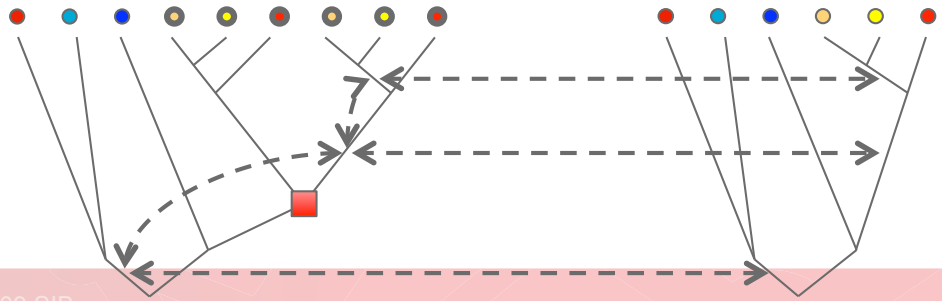
VI. Divergence relative to outgroup



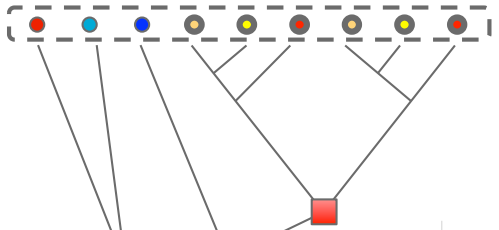
VII. Divergence relative to outgroup of duplicates *vs.* singletons



VIII. Comparison of several singletons and duplicates per gene tree (sequences)



IX. Comparison of several singletons and duplicates per gene tree (functional data)

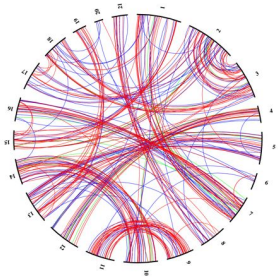


Three cartoon models of evolution

- DDC subfunctionalization (loss of function) after duplication only
- Neofunctionalization after duplication only
- Neofunctionalization with time, irrespective of duplication

➢ Different predictions in different study designs?

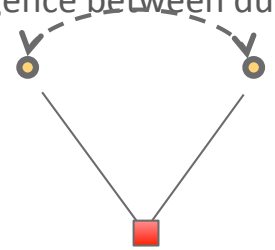
Comparisons inside one genome



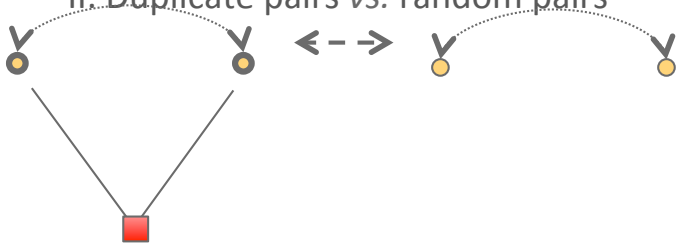
UNIL | Université de Lausanne
Département d'écologie et évolution

SIB Swiss Institute of Bioinformatics

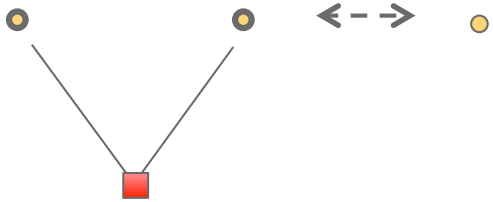
I. Divergence between duplicates



II. Duplicate pairs *vs.* random pairs



III. Duplicates *vs.* singletons



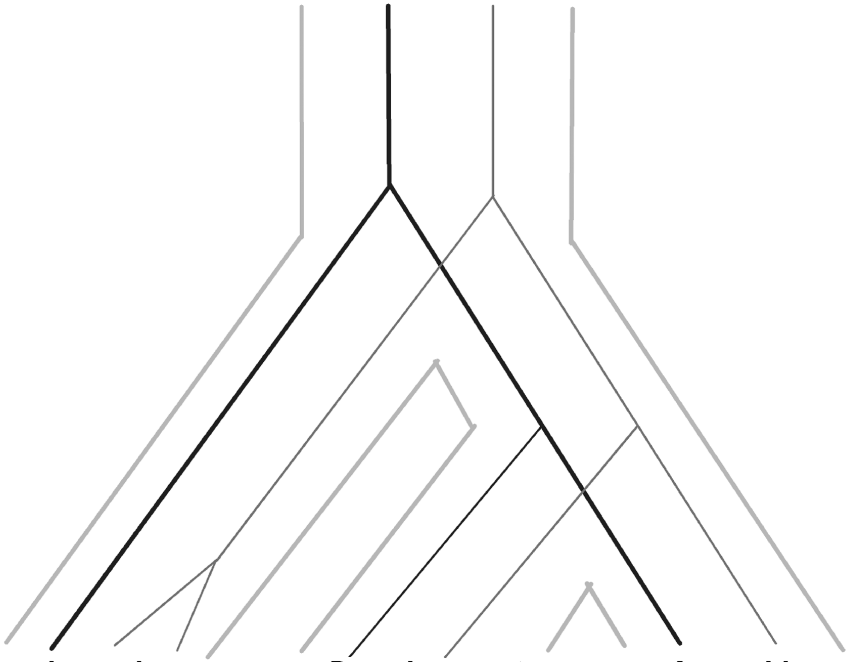
What can we see under different models of evolution?

I- *All models*: Differences between paralogs

II- *All models*: Paralogs more similar than random pairs, but not identical

III- *All models*: Measure of retention bias, confused by evolution after duplication

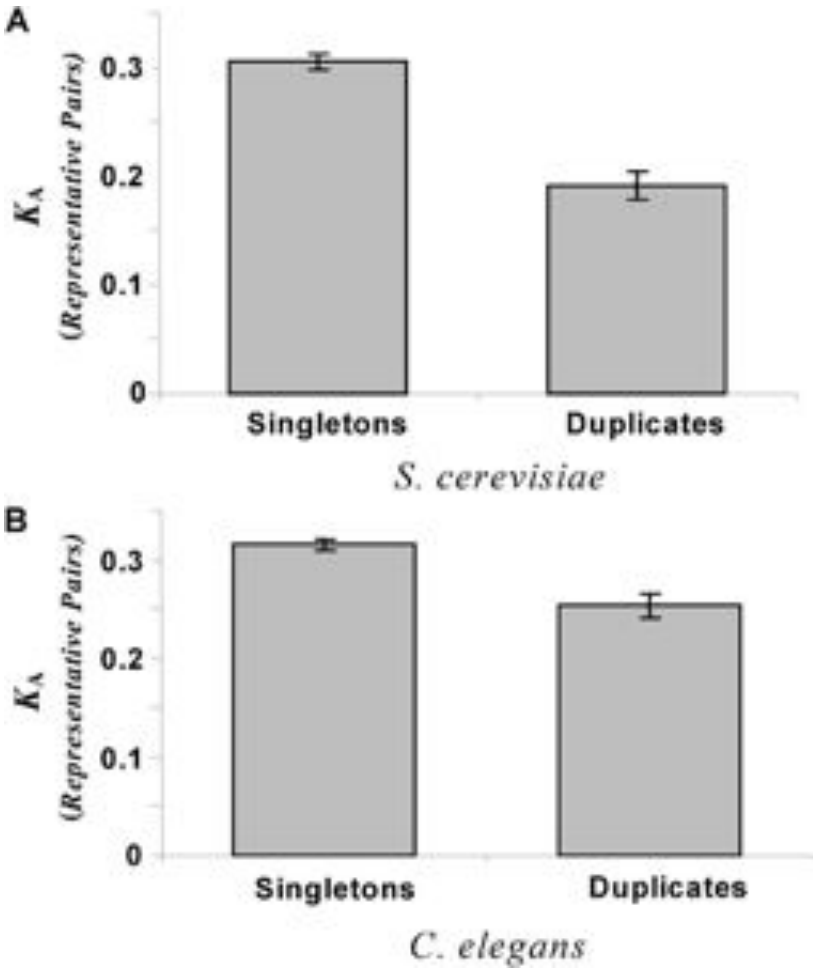
Biased gene retention



Study Species - where duplicate and singleton genes are identified. (*C. elegans, S cerevisiae*)

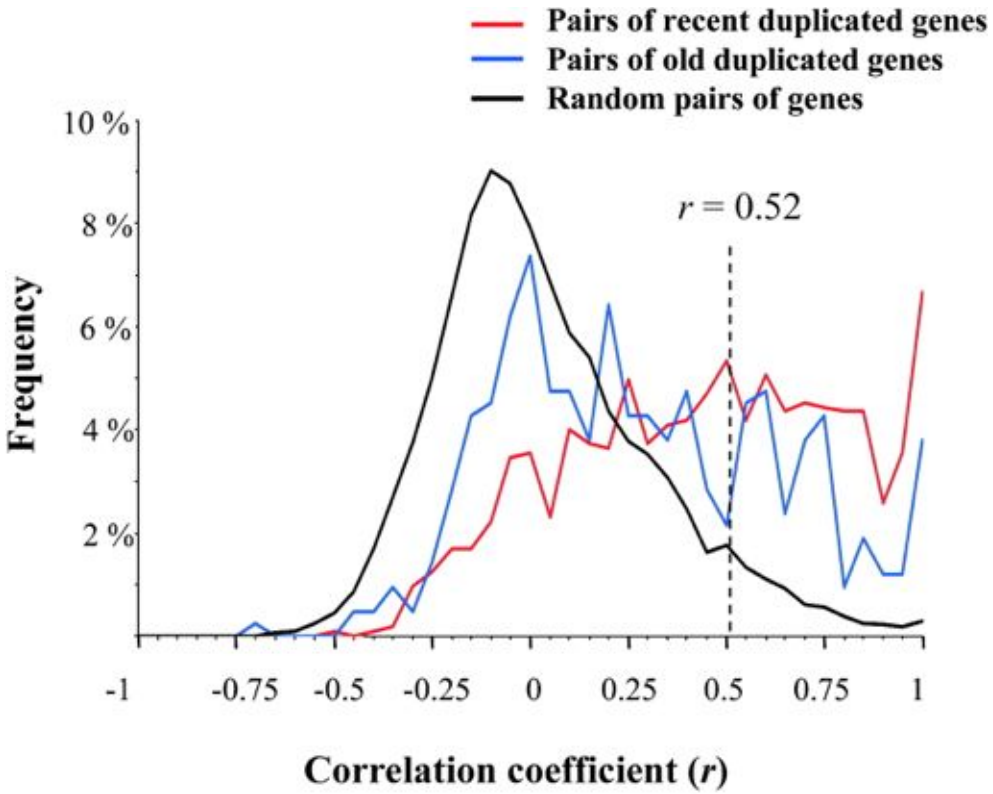
D. melanogaster ***A. gambiae***

Outgroup species in which representative pairs of orthologs are identified.



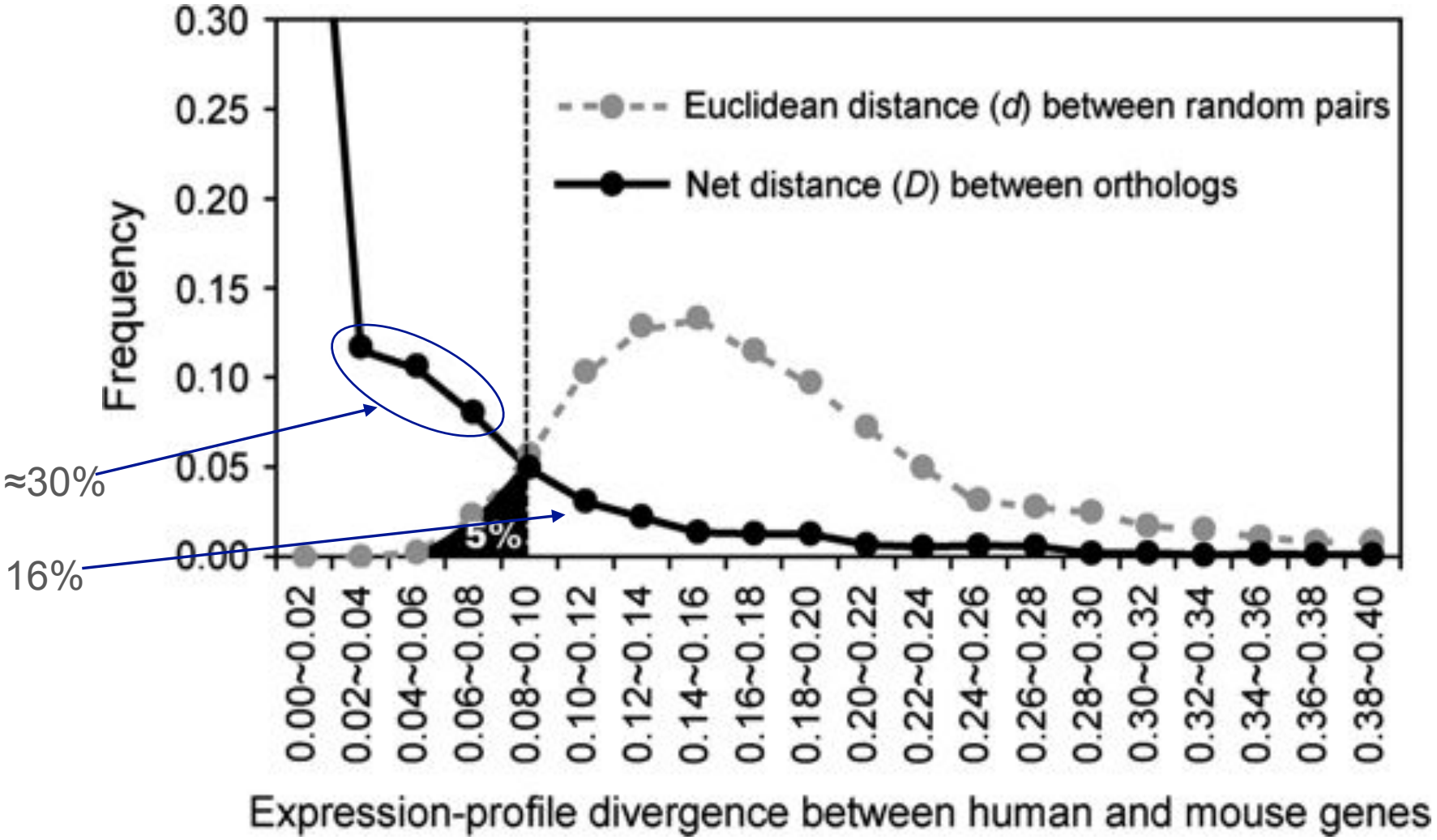
Arabidopsis WGD genes:

Affymetrix data from 62 different conditions and tissues



expression divergence:
r < 0.52:
57% of pairs of recent duplicates
73% of pairs of ancient duplicates
r < 0:
15% of pairs of recent duplicates
29% of pairs of ancient duplicates

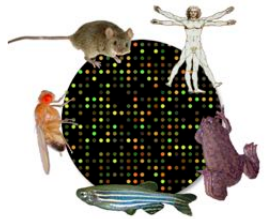
Blanc & Wolfe (2004) Plant Cell 16:1679-91



Liao & Zhang 2006 Mol Biol Evol 23, 530-40

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

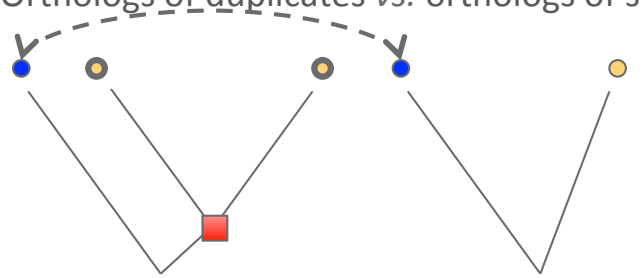
Two genomes: functional data



UNIL | Université de Lausanne
Département d'écologie et évolution

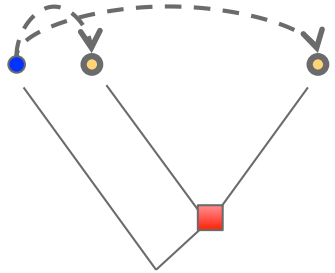
SIB Swiss Institute of Bioinformatics

IV. Orthologs of duplicates *vs.* orthologs of singletons



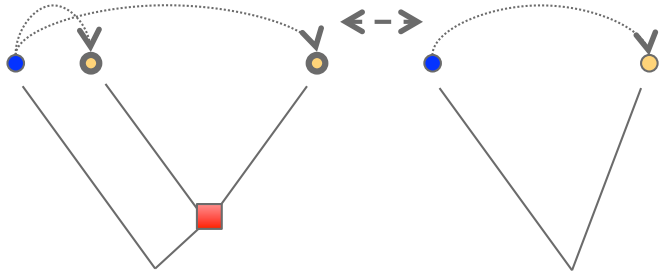
IV- *All models*: Measure of retention bias

VI. Divergence relative to outgroup

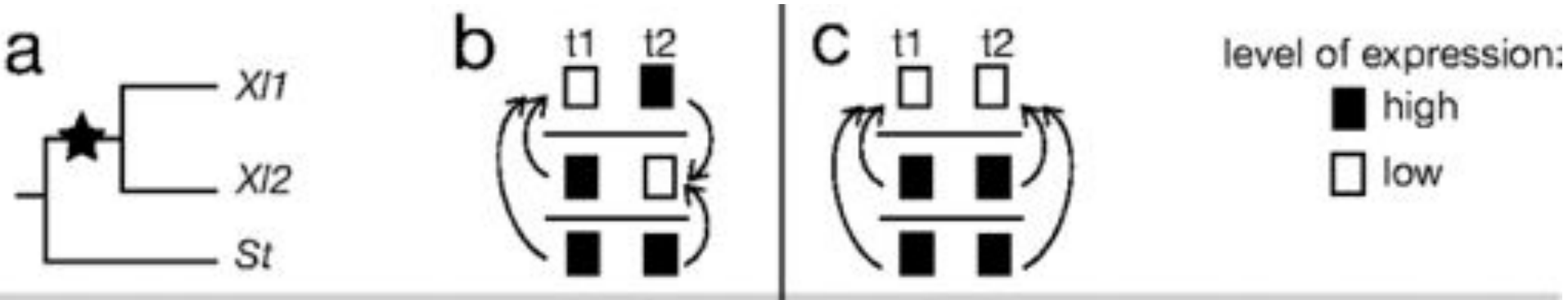


VI- *DDC subfunctionalization after duplication*: Two paralogs different, complementary to full outgroup function
Neofunctionalisation after duplication or *with time*: One paralog similar to outgroup, one different

VII. Divergence relative to outgroup of duplicates *vs.* singletons

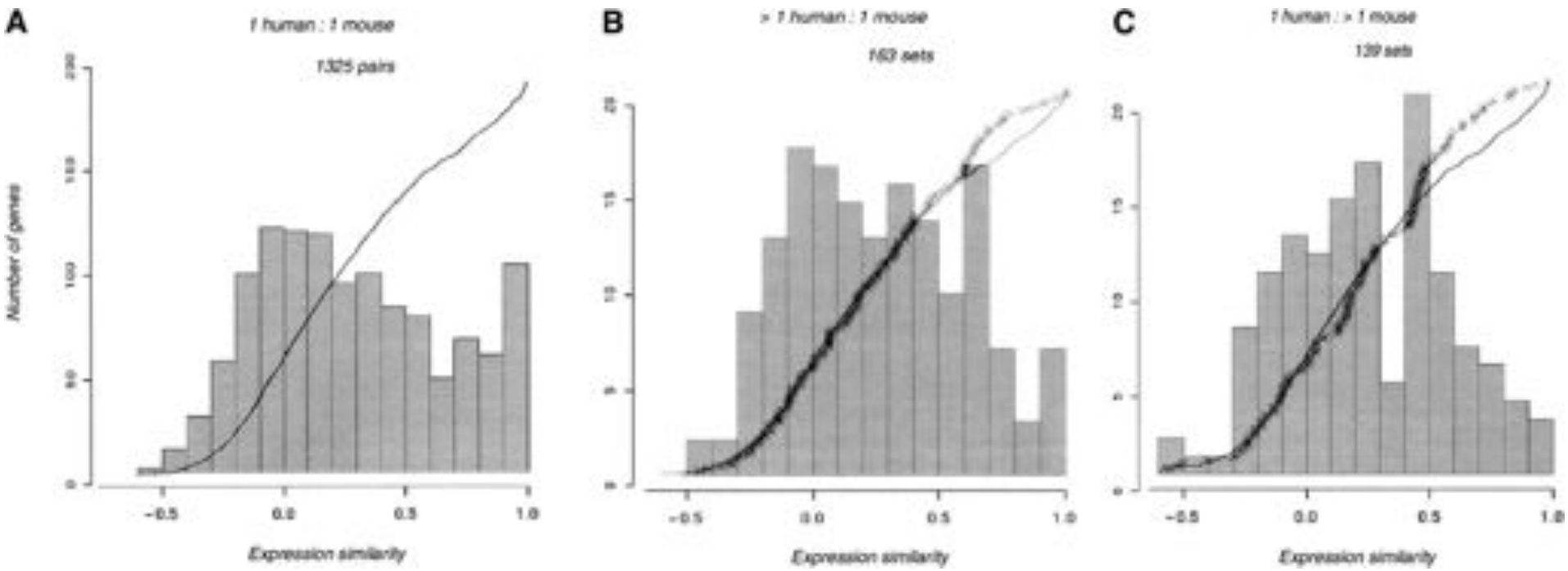


VII- *DDC subfunctionalization after duplication*: Two paralogs different, complementary to outgroup; singleton similar to outgroup
Neofunctionalisation after duplication: One paralog similar to outgroup, one different; singleton similar to outgroup
Neofunctionalisation with time: No specific prediction



		subfunctionalization	asymmetric	Fisher test p-value
d	divergence in at least 2 tissues	19 (61)	49 (109)	0.01 (0.01)
e	divergence in exactly 2 tissues	15 (44)	41 (93)	0.02 (0.003)

Semon & Wolfe 2008 PNAS 105, 8333-38

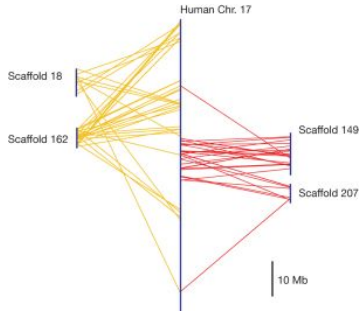


Genes highly similar in expression are underrepresented in ortholog sets with recent human- or mouse-specific gene duplications. Histograms of expression correlation coefficients (R) across 16 tissues are shown for (A) 1325 one-to-one orthologs between human and mouse; (B) 163 ortholog sets with one mouse sequence and more than one human co-ortholog; (C) 135 ortholog sets with one human sequence and more than one mouse co-ortholog

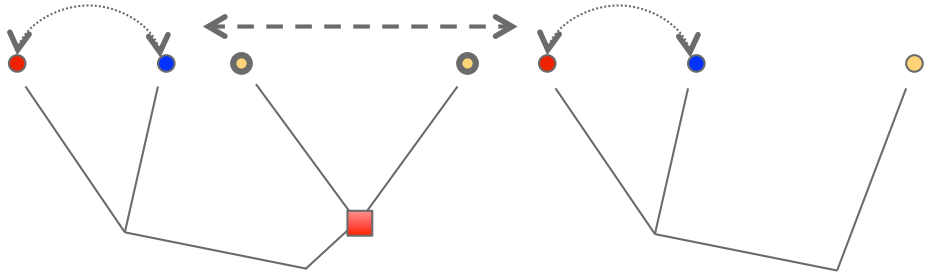
Huminiecki & Wolfe 2004 Genome Res 14, 1870-79

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Two genomes: sequences

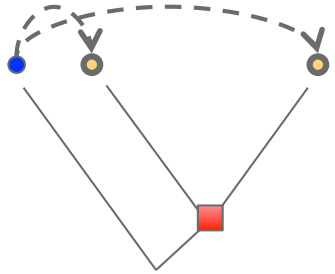


V. Divergence between orthologs of duplicates *vs.* between orthologs of singletons



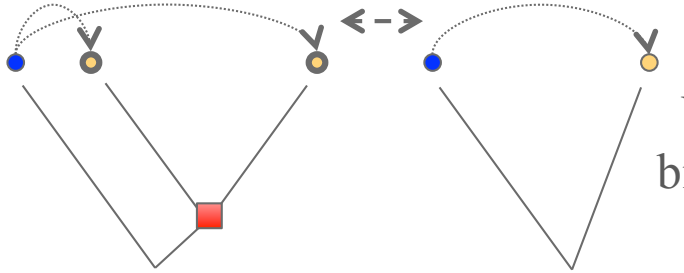
V- *All models*: Measure of retention bias

VI. Divergence relative to outgroup



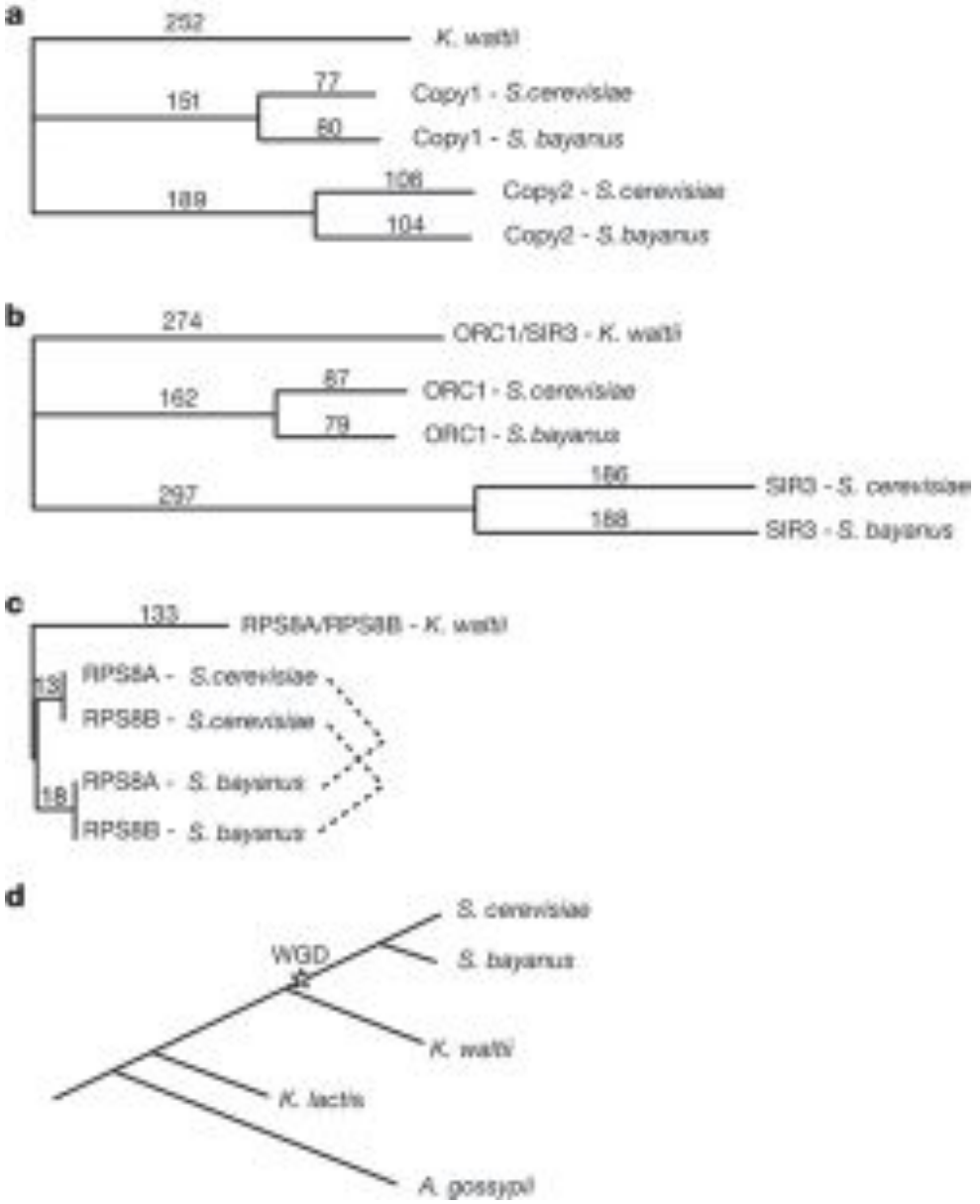
VI- *DDC subfunctionalization after duplication*: No prediction relative to symmetry, relaxed purifying selection *Neofunctionalisation after duplication* or *with time*: Asymmetry between paralogs, positive selection (low power)

VII. Divergence relative to outgroup of duplicates *vs.* singletons



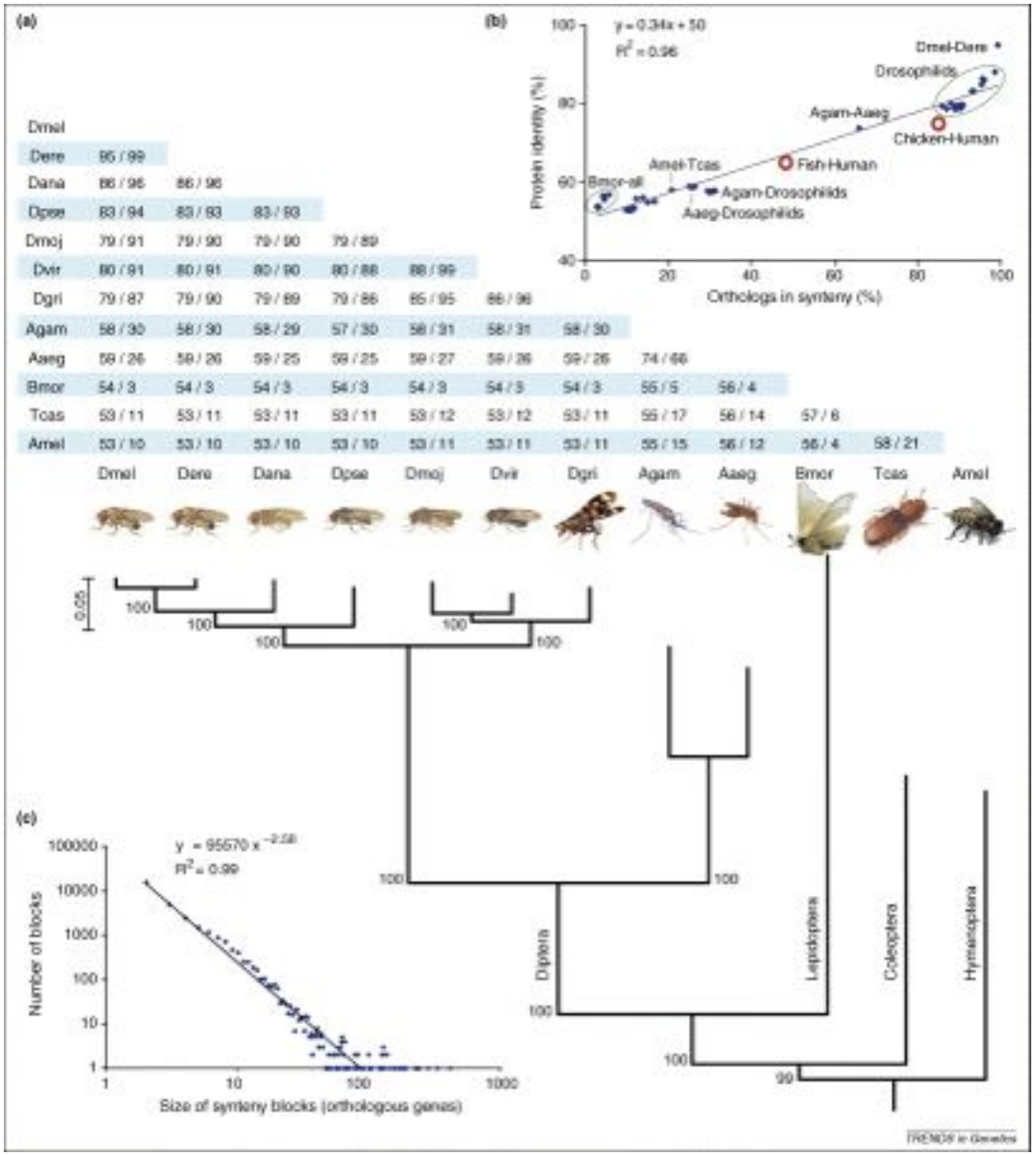
VII- Higher divergence of duplicates, confused by retention bias

After WGD in yeast:
76 / 457 gene pairs (17%) show accelerated protein evolution
defined as instances in which the amino acid substitution rate along one or both of the S. cerevisiae branches was at least 50% faster than the rate along the K. waltii branch.

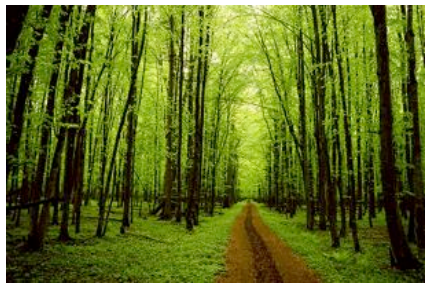


Kellis et al (2004) Nature 428: 617-24

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics


(a)
(b)
(c)
Dmel
Dere
Dana
Dpse
Dmoj
Dvir
Dgri
Agam
Aaeg
Bmor
Tcas
Amel
Protein identity (%)
Orthologs in synteny (%)
Drosophilids
Agam-Aaeg
Chicken-Human
Fish-Human
Amel-Tcas
Agam-Drosophilids
Aaeg-Drosophilids
Number of blocks
Size of synteny blocks (orthologous genes)
Diptera
Lepidoptera
Coleoptera
Hymenoptera

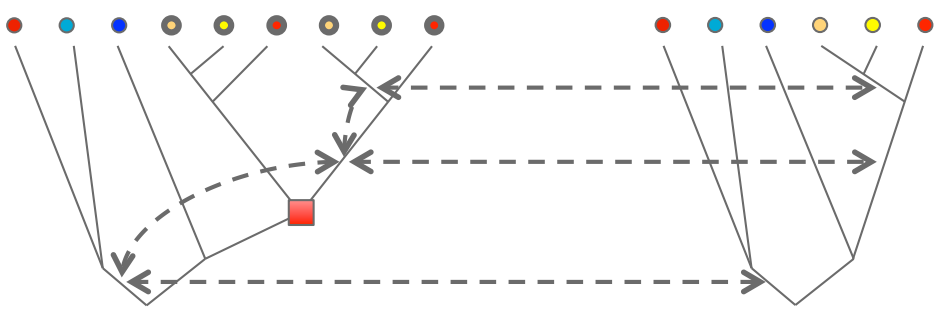
Real phylogenomics



UNIL | Université de Lausanne
Département d'écologie et évolution

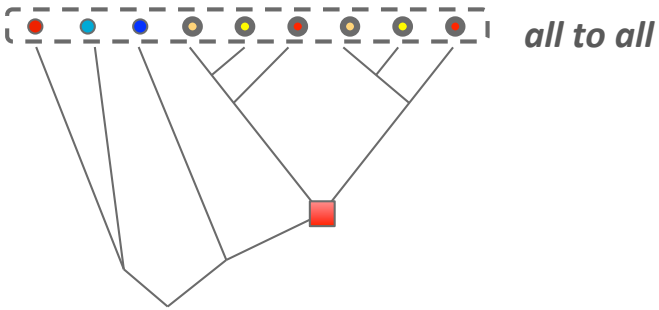
SIB
Swiss Institute of Bioinformatics

VIII. Comparison of several singletons and duplicates per gene tree (sequences)



VII- *DDC subfunctionalization after duplication*: Higher relaxation of purifying selection on branches after duplication
Neofunctionalisation after duplication: More positive selection on branches after duplication
Neofunctionalisation with time: Positive selection in various branches of the tree

IX. Comparison of several singletons and duplicates per gene tree (functional data)

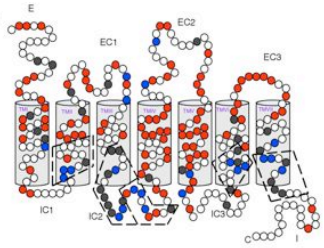


IX- *DDC subfunctionalization after duplication*: Conservation of pattern among singletons; sub-patterns in duplicates
Neofunctionalisation after duplication: Conservation in most homologs; new patterns in some duplicates
Neofunctionalisation with time: Variation in pattern among homologs, with gain of new patterns

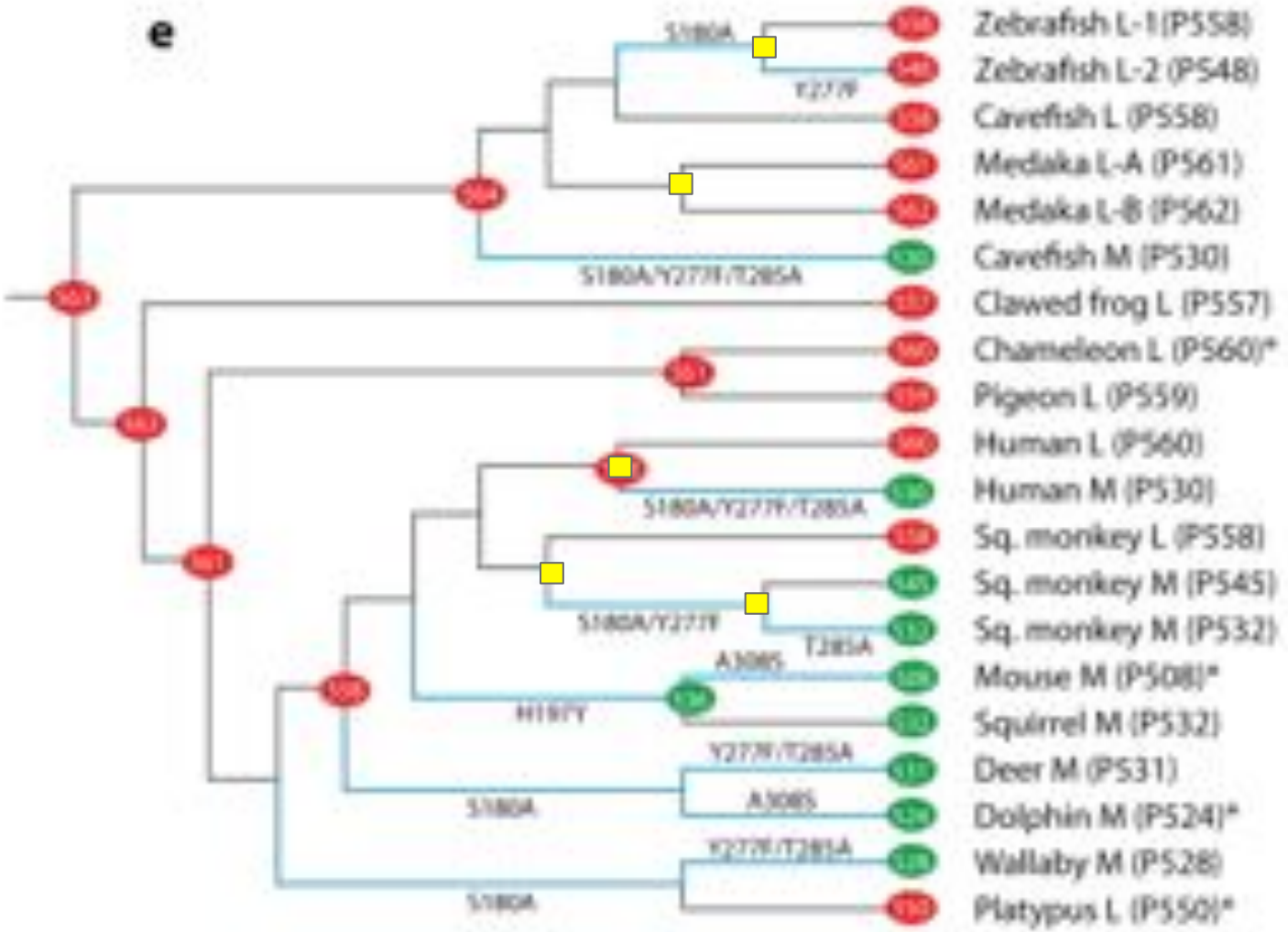
Discuss with your neighbor

- What is more frequent: point mutations or duplications?
- Where in animal evolution were there genome duplications?
- Two reasons why we should distinguish orthologs and paralogs
- What is gene function?
- **Find an example of orthologs with different functions**

Real data can be confusing



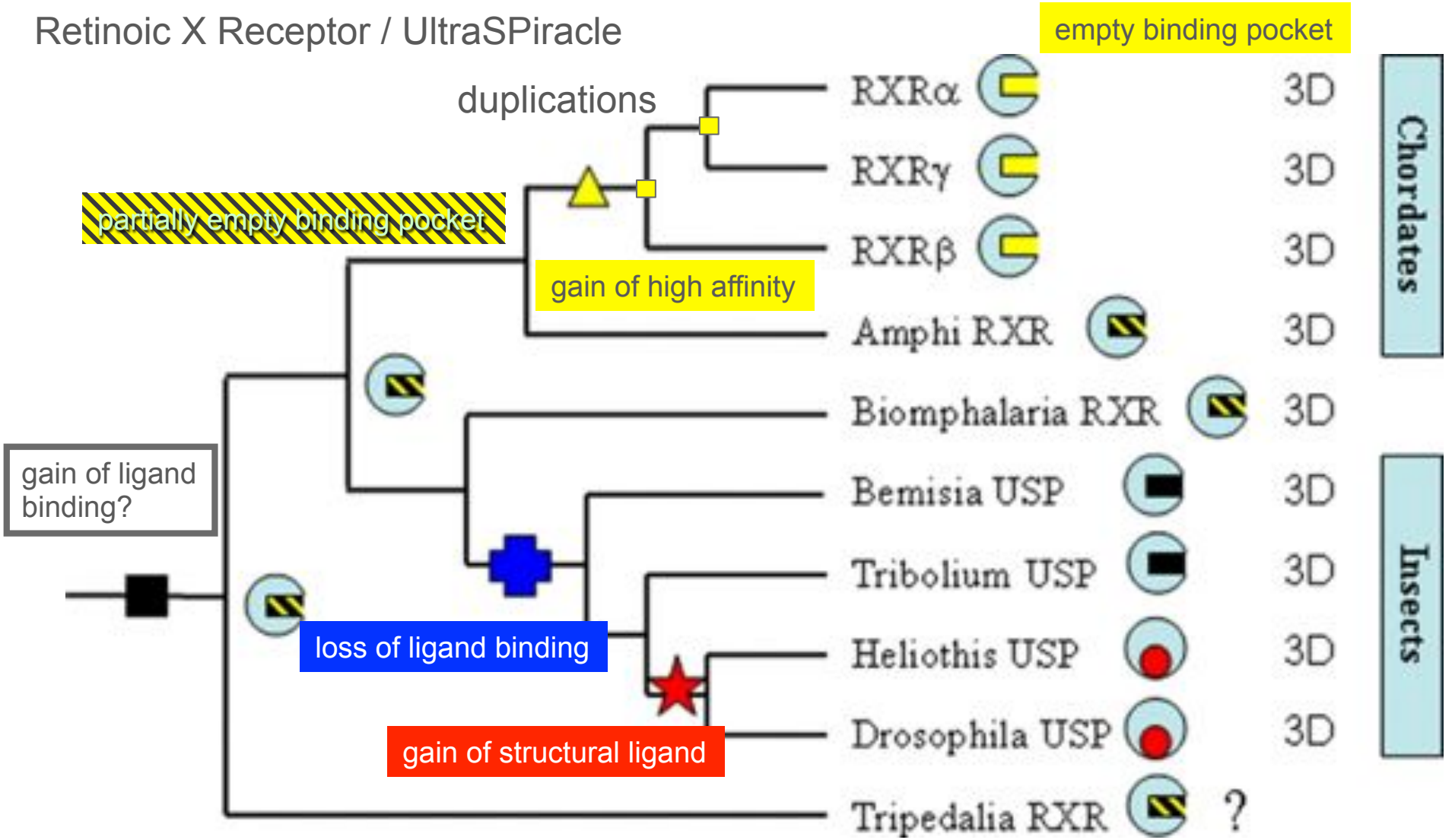
M/LWS pigments



shifts in wavelength

Yokoyama (2008) Ann Rev Genomics Human Genet 9: 259-282

Retinoic X Receptor / UltraSPiracle



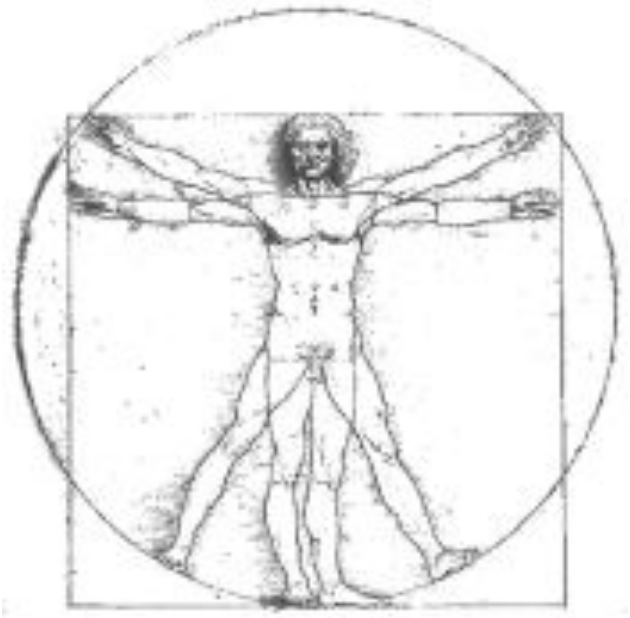
Tocchini-Valentini et al. 2009 J. Biol. Chem. 284: 1938-1948

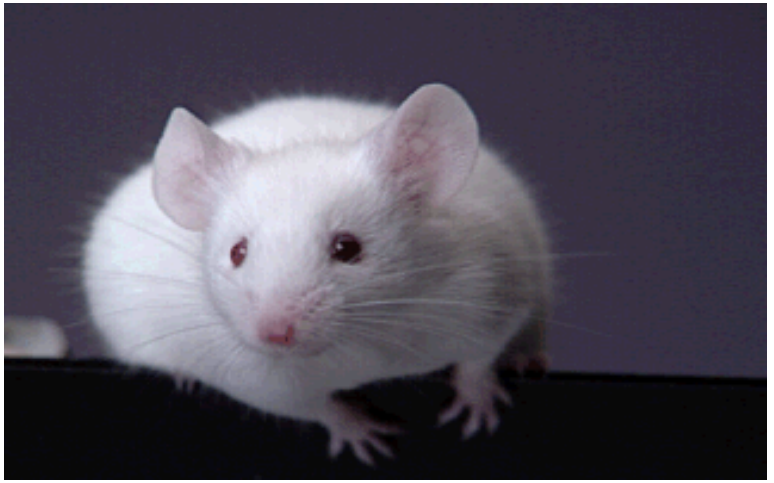
Humans are not mice

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Review of human-mouse ortholog comparisons:

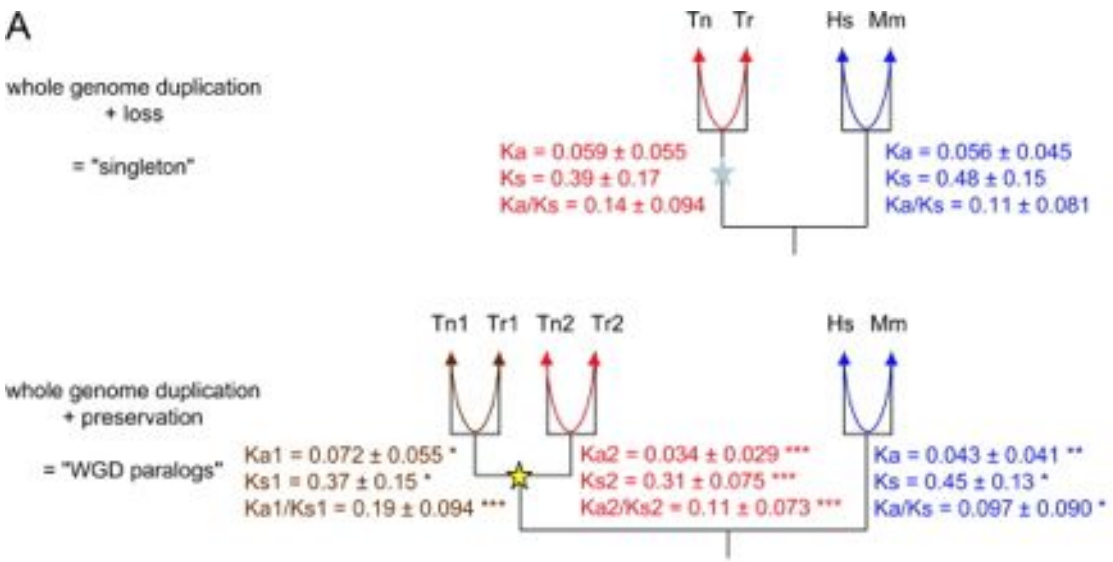
- ≥16% divergence of expression
- ≥13% divergence of splice forms
- 9% with secondary duplications
- ≥20% mutant lethal in human but not in mouse

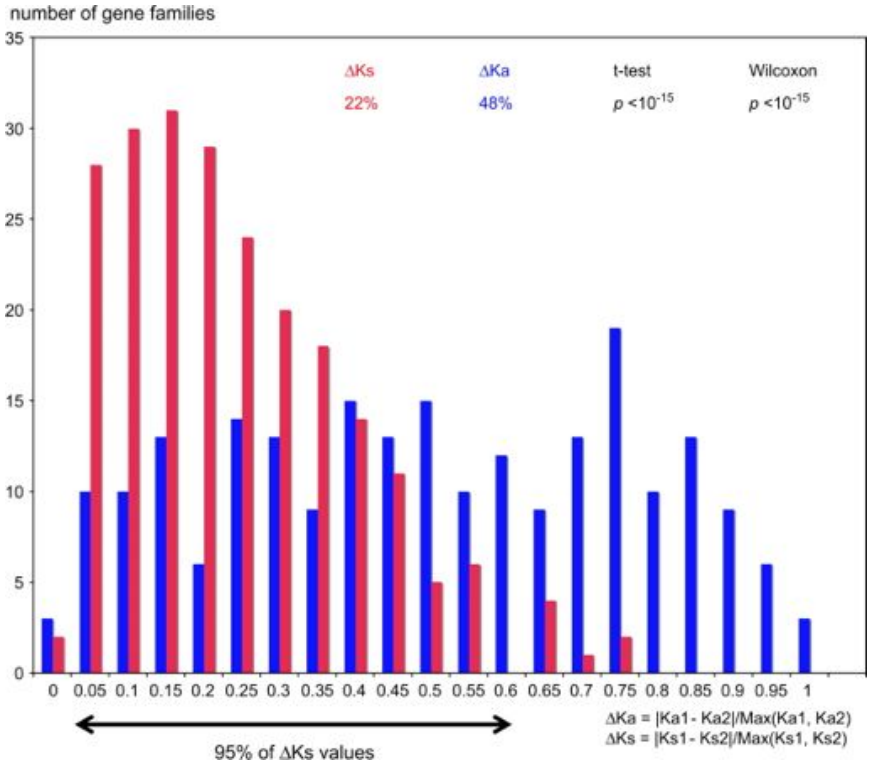




Gharib and Robinson-Rechavi 2011 Briefings Bioinf 12, 436-441

Duplicates evolve asymmetrically





Paralogs from fish WGD:

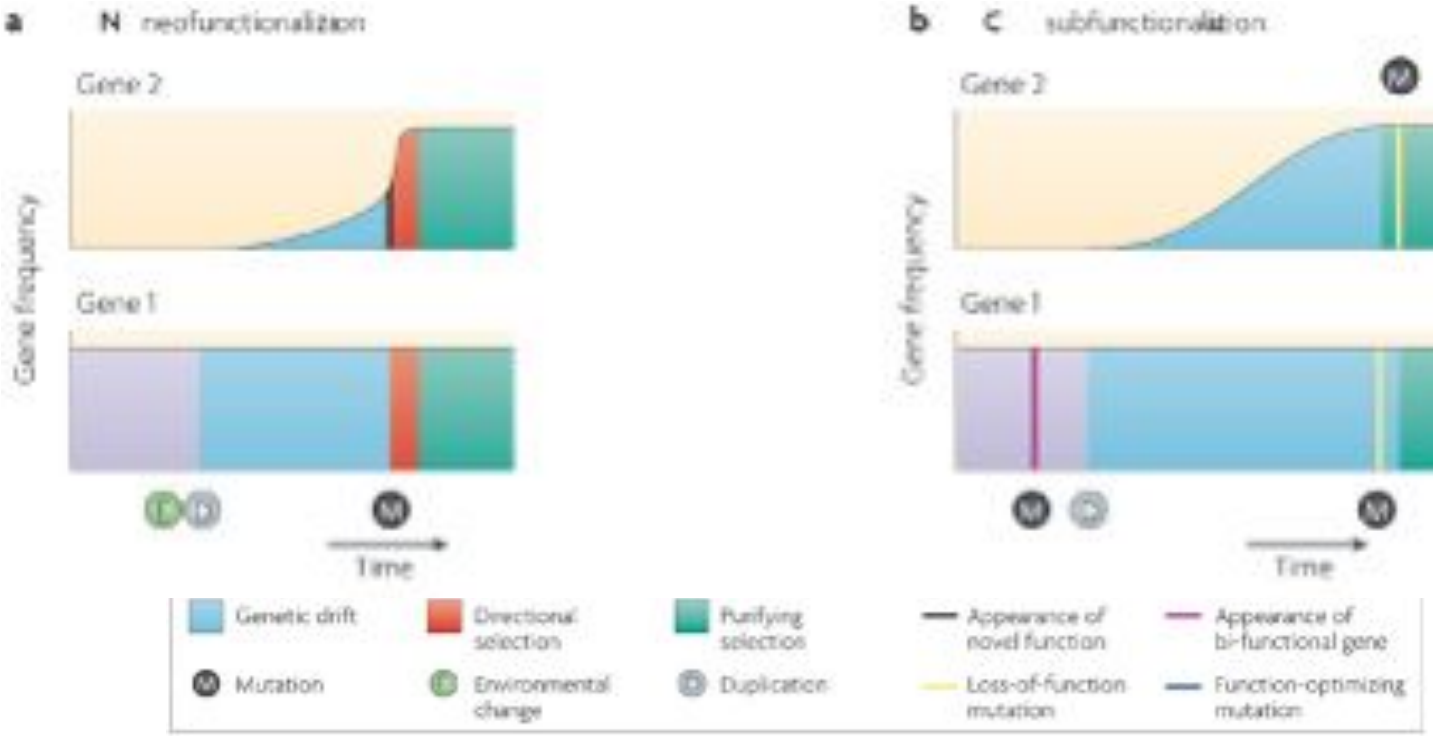
81% of paralogous pairs: ΔKa > ΔKs
36% of paralogous pairs: ΔKa > 97.5% percentile ΔKs

Ka/Ks of slowest paralog ≈ Ka/Ks of outgroup
Ka/Ks of fastest paralog >> Ka/Ks of outgroup
-> acceleration of one paralog

Brunet et al. (2006) Mol Biol Evol 23: 1808-16

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

What is asymmetry indicative of?



Conant & Wolfe 2008 Nature Rev Genet 9:938-950

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Real neofunctionalization has postive selection

Species 1	A	L	P	H	Y
	GCC	CTT	CCT	CAT	TAT
Species 2	A	R	P	H	Y
	GCC	CGT	CCT	CAT	TAC

Measure of mutation, genetic drift and time:

- $dS = \frac{\text{number of } \mathbf{synonymous} \text{ substitutions}}{\text{synonymous sites}}$

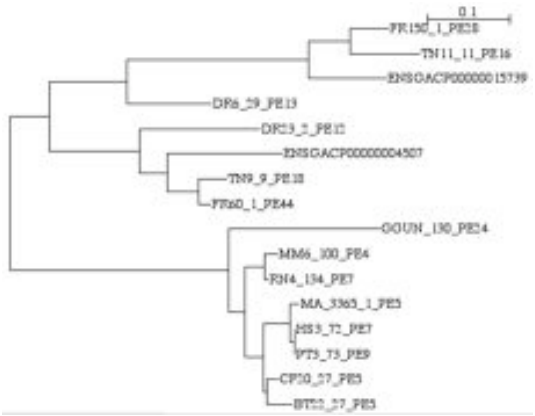
Measure of mutation, genetic drift and time and selection:

- $dN = \frac{\text{number of } \mathbf{non\text{-}synonymous} \text{ substitutions}}{\text{non-synonymous sites}}$

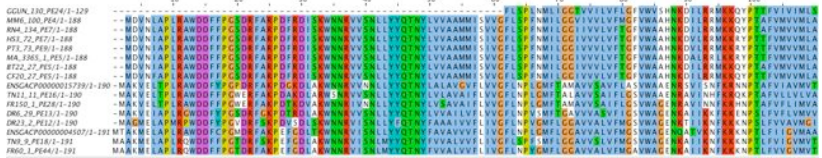
Measure of selective pressure: ratio dN/dS (ω):

- dN/dS < 1 -> purifying selection
- dN/dS = 1 -> neutral evolution
- dN/dS > 1 -> positive selection

Codon models
(PAML Package, Yang 2007)

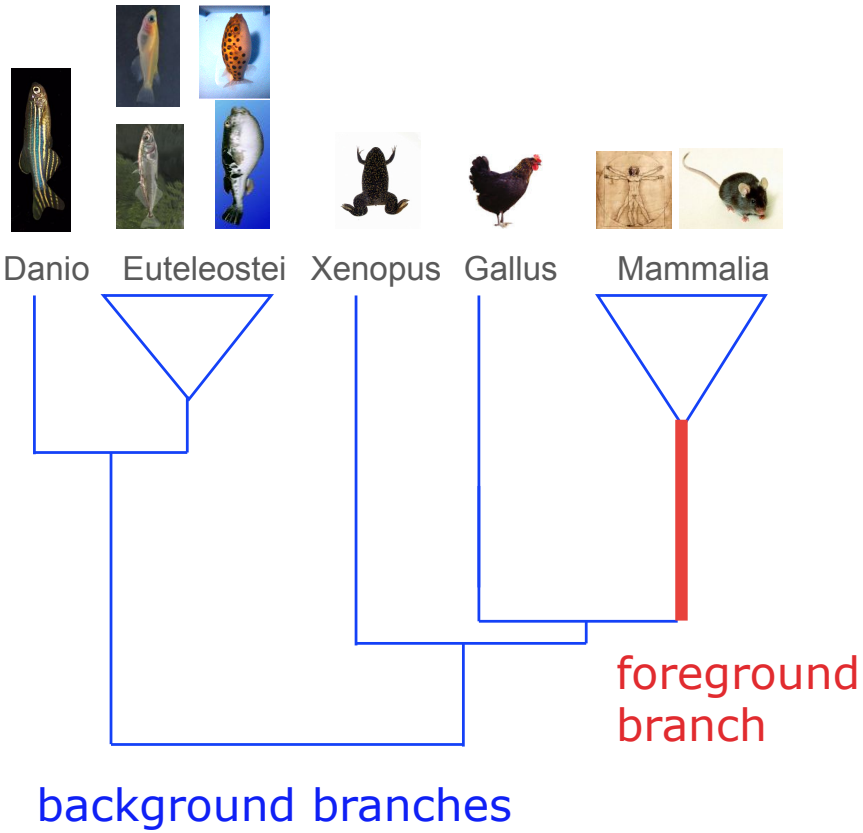


- One ratio model
  - Same dN/dS for all branches and sites.
- Branch models
  - Estimate different dN/dS among branches.
- Site models
  - Estimate different dN/dS among sites.



- Branch-site models
  - Estimate different dN/dS among branches and sites.

Branch-site test



ω = dN/dS

$\omega_0 < 1$
$\omega_1 = 1$
proportions of sites: $p_0 + p_1 = 100\%$

$\omega_0 < 1$
$\omega_1 = 1$
ω_2
proportions of sites: $p'_0 + p'_1 + p'_2 = 100\%$

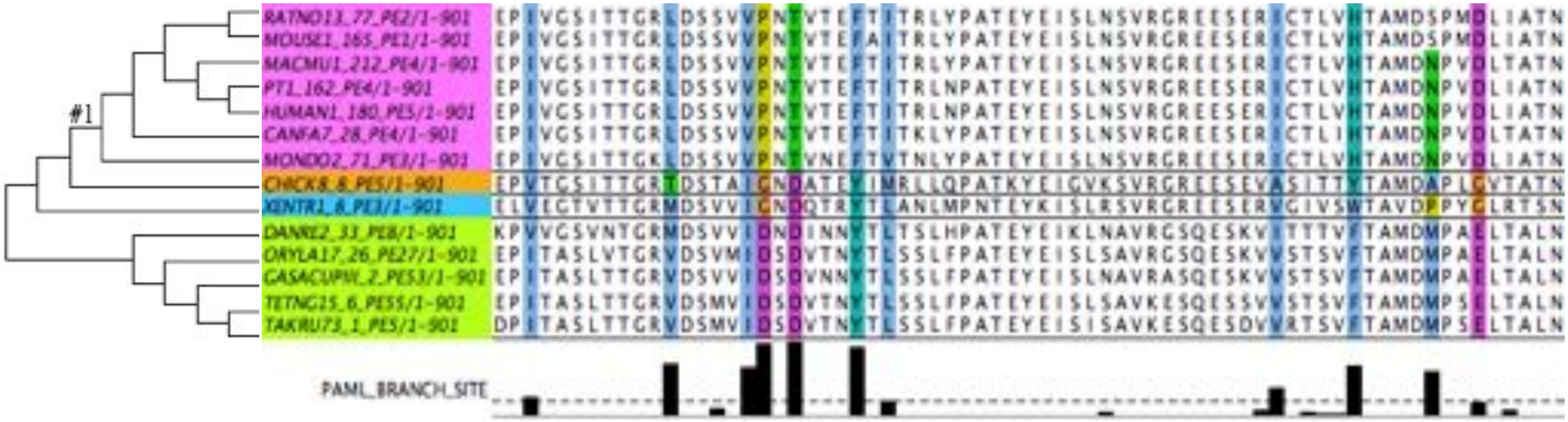
$\omega_2 = 1$

$\omega_2 \geq 1$

Likelihood Ratio Test:
Is the difference between the foreground and background branches due to positive selection?

Zhang et al (2005) Mol Biol Evol 22: 2472-2479

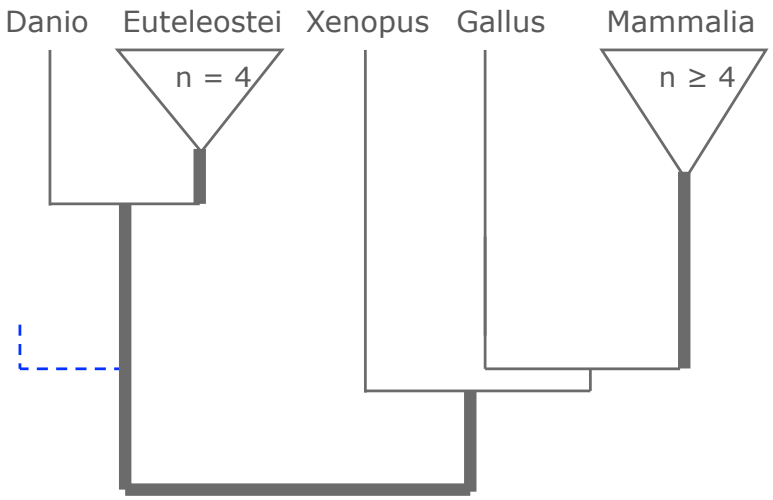
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics



Bayesian Empirical Bayes posterior probability of positive selection for each site on this branch

Positive selection in vertebrate genes

Positive selection in vertebrate evolution

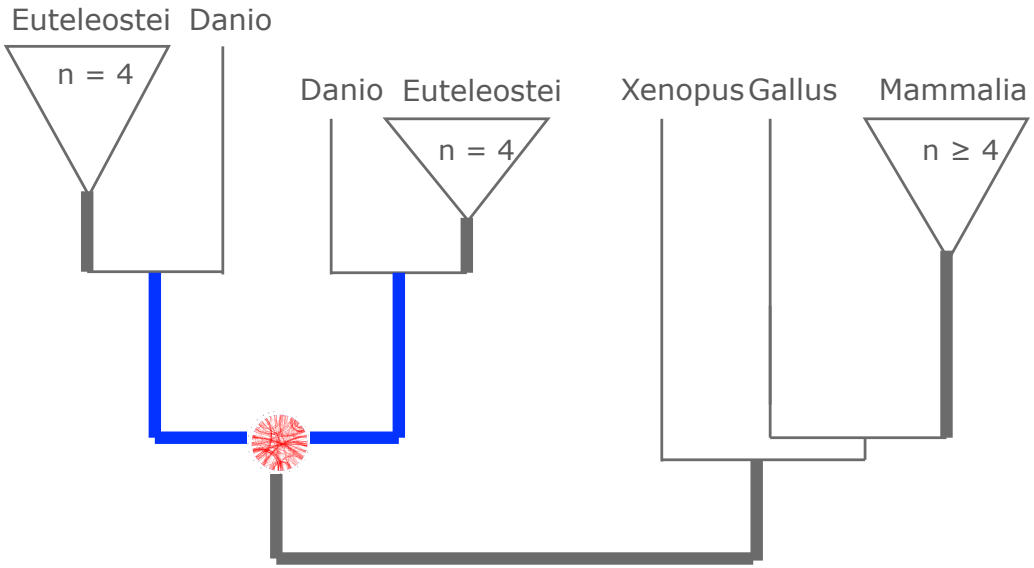


2673 branches tested
(correction for multiple testing: q = 10%)

Positive selection significant for 45% of branches

Only 0.9% to 4.7% of sites

≥ 1 branch significant for 77% of genes



Studer et al. 2008 Genome Res 18:1393-1402

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Positive selection and duplication

Positive selection in vertebrate evolution

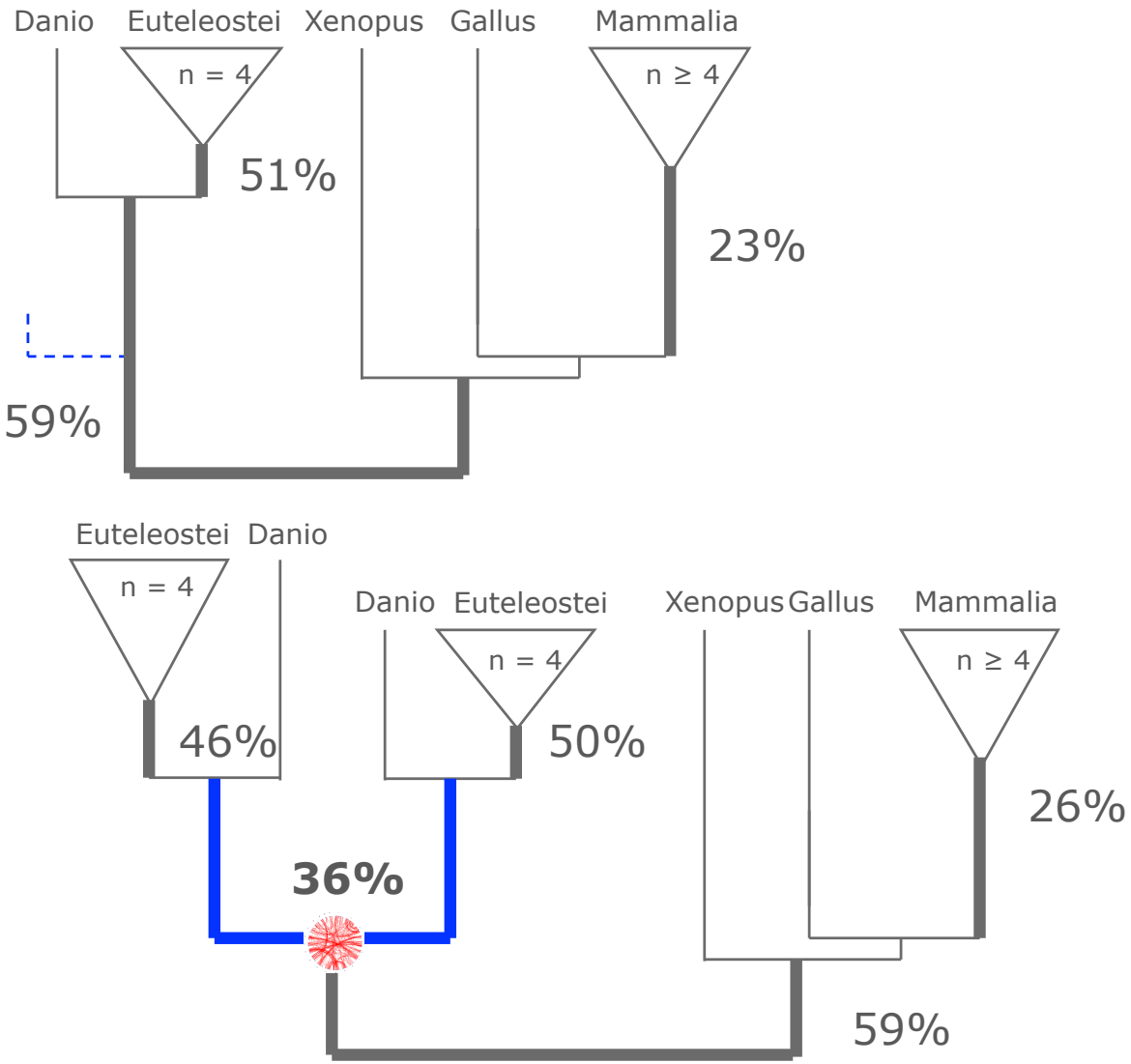


Table 5. Influence of duplication on substitution parameters

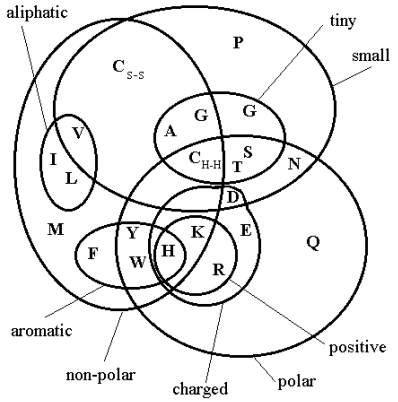
Branch behavior[a]	Parameter	FSGD vs. other branches[b]		Euteleosts branch: FSGD topology vs. singleton		Bony vertebrates branch: 2R detected vs. not detected	
		P-value[c]	Difference[d]	*P*-value[c]	Difference[d]	*P*-value[c]	Difference[d]
LRT significant	ΔLnL	0.14	8.9–10.7	0.57	10–11	0.94	11–11
	Branch length[e]	0.017	0.085–0.11	0.014	0.093–0.12	0.035	0.13–0.14
	Mean ω_0	0.93	0.075–0.075	0.015	0.061–0.075	0.0098	0.072–0.079
	Percent of sites ω_0[f]	0.92	83%–84%	0.0029	89%–85%	0.67	83%–82%
	Percent of sites ω_1[f]	0.27	8.6%–10%	**3.4×10^{-4}**	**6.6%–11%**	0.035	9.3%–11%
	Percent of sites ω_2[f]	0.12	8.1%–5.6%	0.56	4.1%–4.1%	0.075	8.1%–7.2%
LRT nonsignificant	ΔLnL	0.41	0.69–0.65	0.30	0.79–0.65	0.33	0.91–1.0
	Branch length[e]	0.43	0.082–0.069	0.46	0.10–0.077	0.018	0.095–0.12
	Mean ω_0	0.13	0.054–0.062	0.65	0.061–0.061	**3.8×10^{-6}**	**0.048–0.064**
	Percent of sites ω_0[f]	**0.00026**	**80%–86%**	0.28	84%–87%	0.011	88%–83%

Boldface indicates that the difference is significant.
[a]Classification according to whether the LRT for positive selection is significant on each branch ($q = 0.10$).
[b]Only branches from tree topologies for which the FSGD branch exists were used.
[c]Nonpaired Wilcoxon test. In bold if the difference is significant after Bonferroni correction ($\alpha = 0.05/30 = 0.0017$).
[d]Duplication or post-duplication mean − nonduplication mean.
[e]In amino acid substitutions/site.
[f]Values for the foreground branch of each test.
(LRT) Likelihood ratio test.

- No impact on positive selection
- Relaxation of purifying selection
- Preferential retention of genes under strong purifying selection (Davis & Petrov 2004; Brunet et al 2006)

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Amino acid shifts

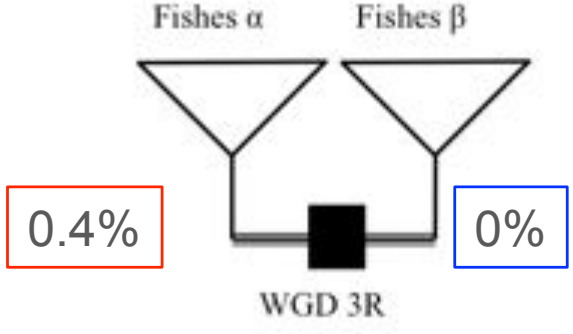


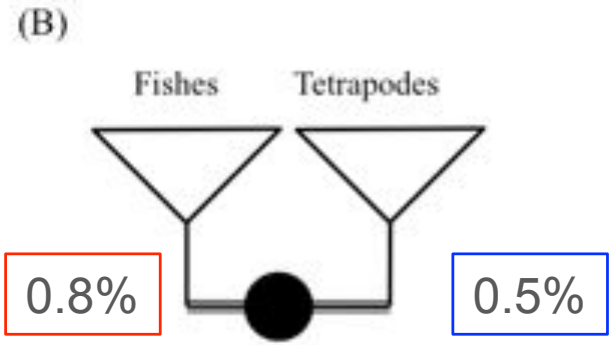
Duplication and amino acid patterns

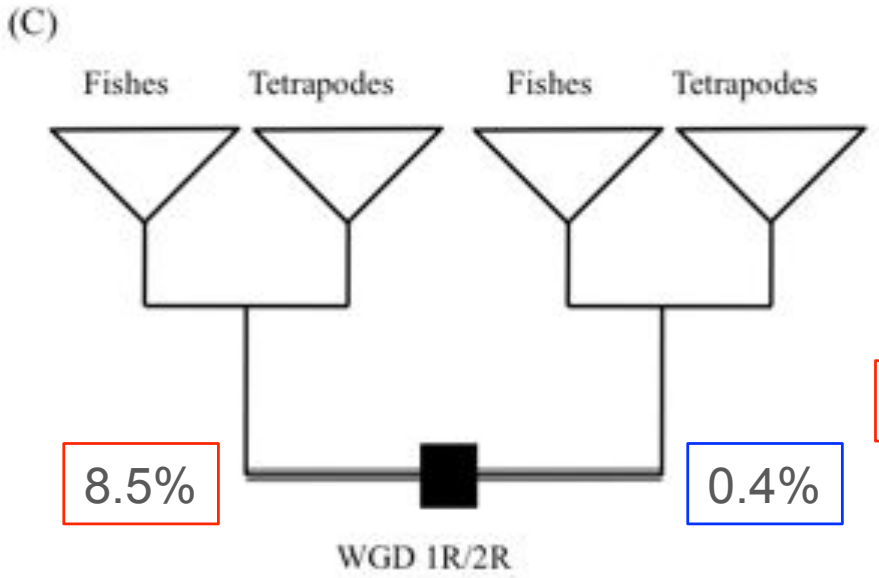
covarions

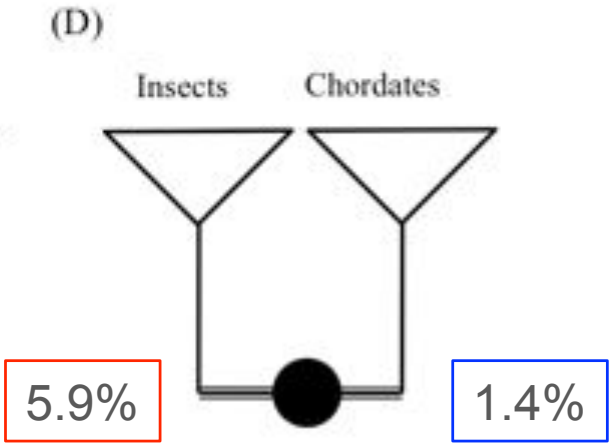
constant
but different

```
KYYLDVLVSDASKPSKWRQ
KFYLDVLVSDISKPSKWRQ
KFYLDVLVSDASKPSKWRQ
KYYLDVMVSDLSKPSKWRQ
NYYLDVMVSDSSKPSKWRH
KYYLDVLVSDLSKPSKWRH
EFYLDVMISDASKPSKWRH
HYYLDVMVSDISKPSKWRH
```



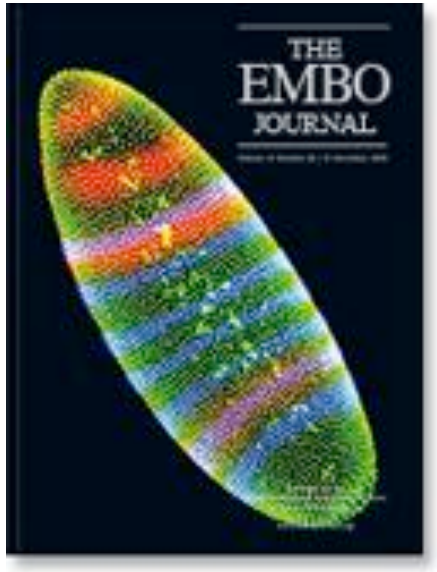






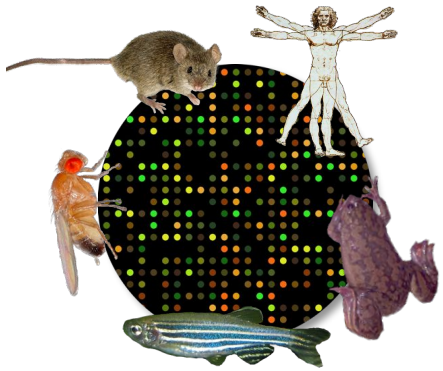
Studer & Robinson-Rechavi 2010 Mol Biol Evol 27: 2618-27

Gene expression



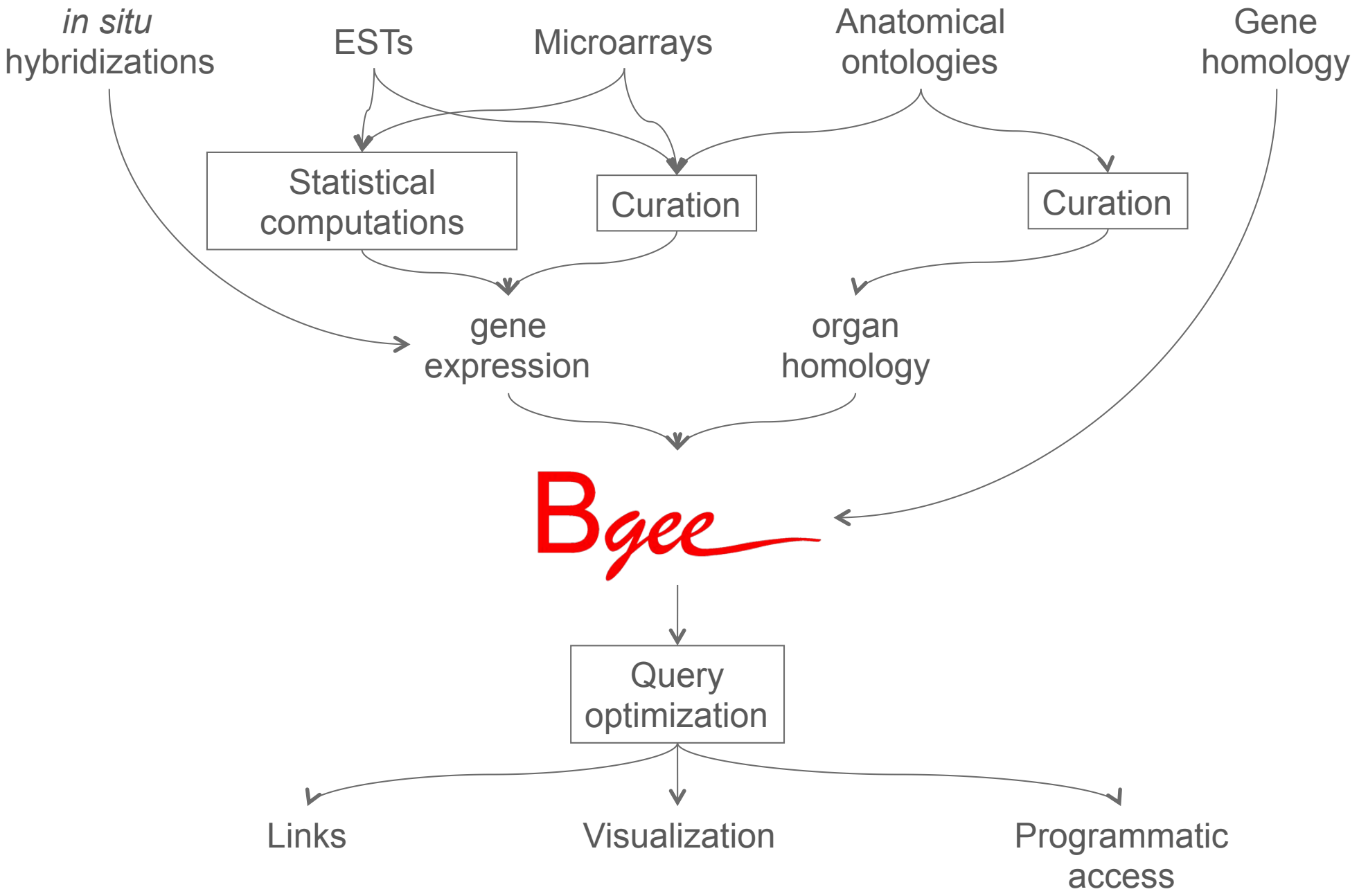
Database:

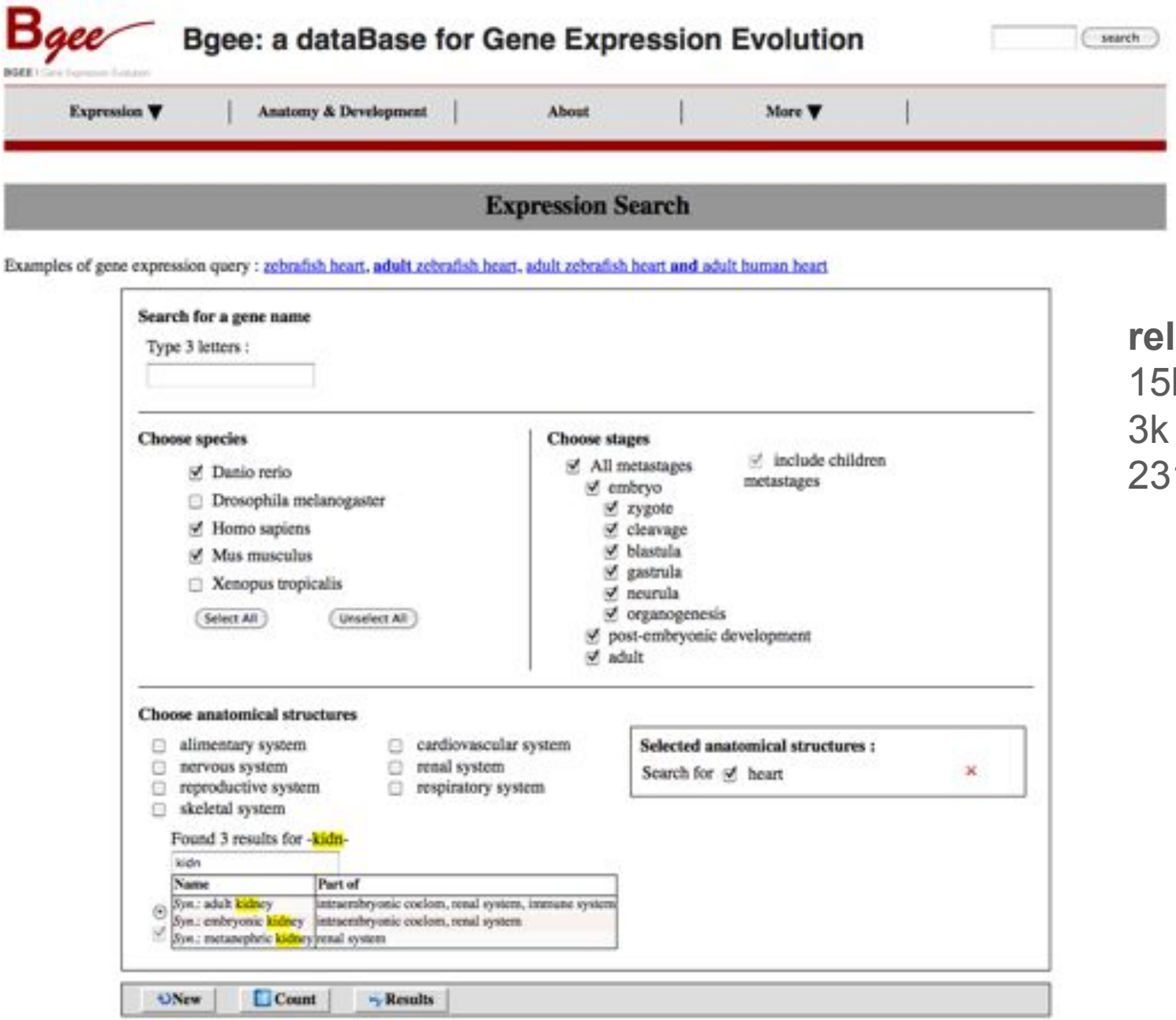
Bgee



Bastian F., Parmentier G., Roux J., Moretti S., Laudet V., Robinson-Rechavi M., 2008.
Lecture Notes in Computer Science 5109: 124-131

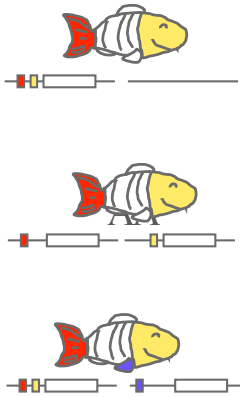
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics


in situ hybridizations
ESTs
Microarrays
Anatomical ontologies
Gene homology
Statistical computations
Curation
Curation
gene expression
organ homology
Bgee
Query optimization
Links
Visualization
Programmatic access



release 10:
15k Affymetrix chips
3k EST libraries
231k *in situ* experiments

Expression after duplication

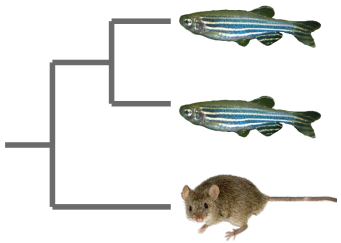
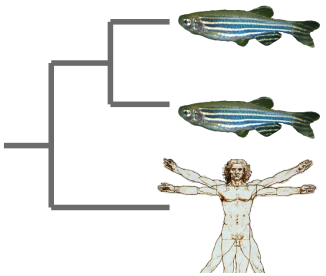
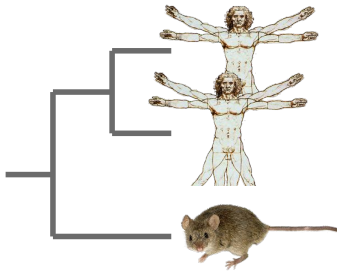
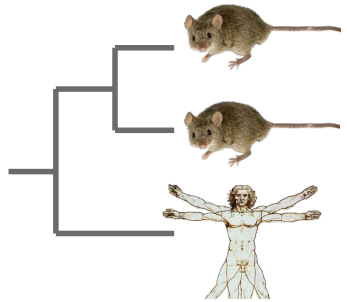


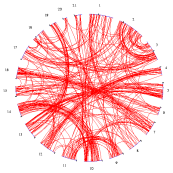
Bastian F., Roux J., Robinson-Rechavi M., *unpublished*

UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

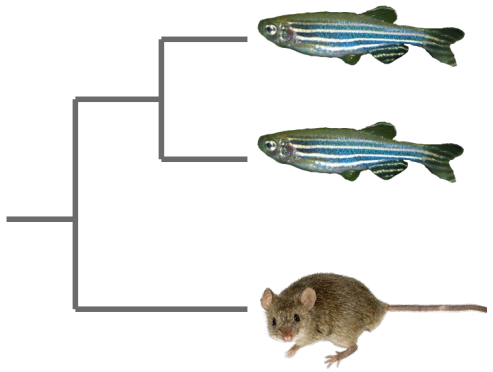
Evolution of expression after duplication

- 4 types of "triplets":

Whole-genome duplicates		Small scale duplicates	
			
Zebrafish duplicates Mouse singleton	Zebrafish duplicates Human singleton	Human duplicates Mouse singleton	Mouse duplicates Human singleton

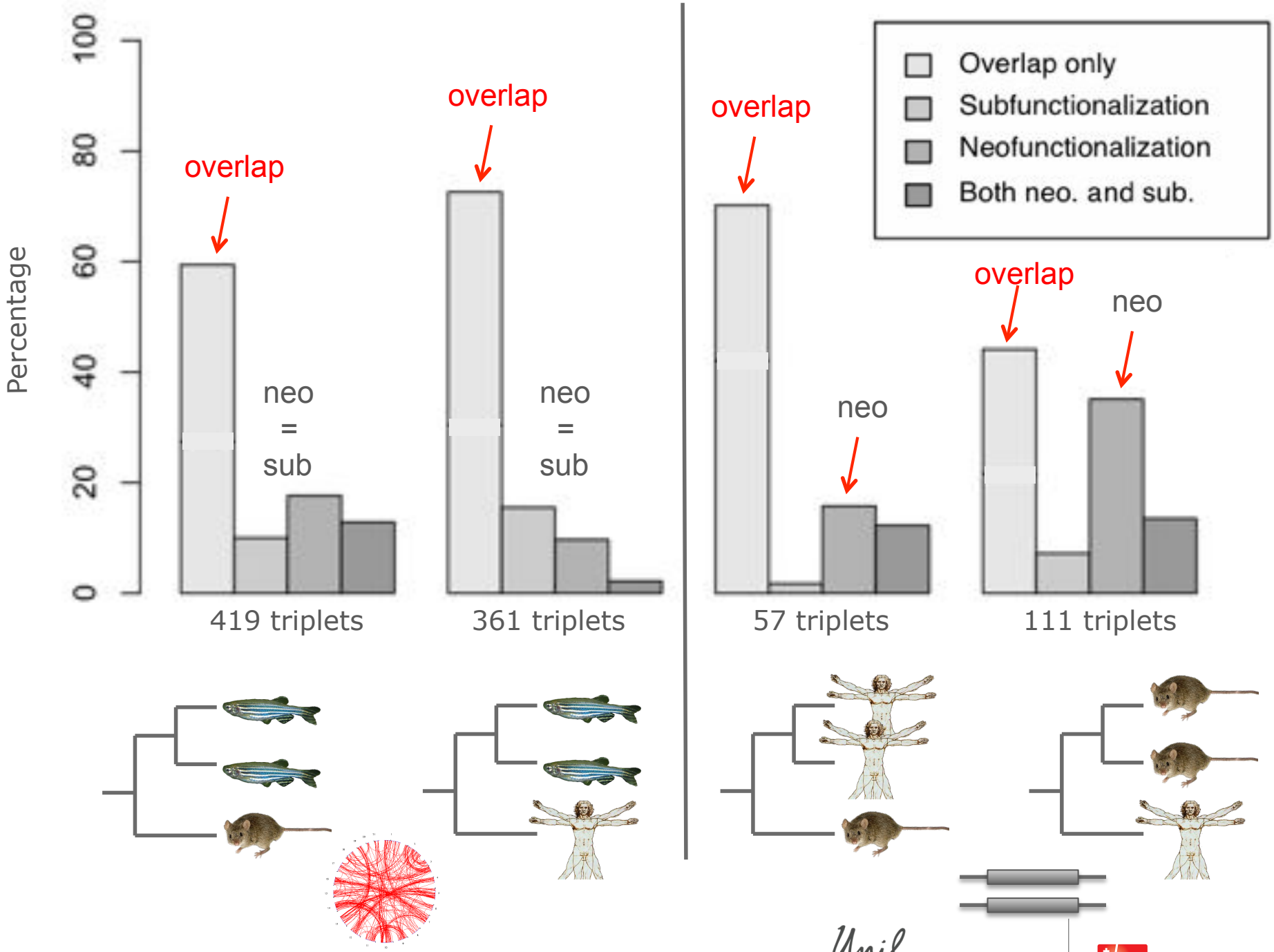


An example: Khdrbs1 related genes



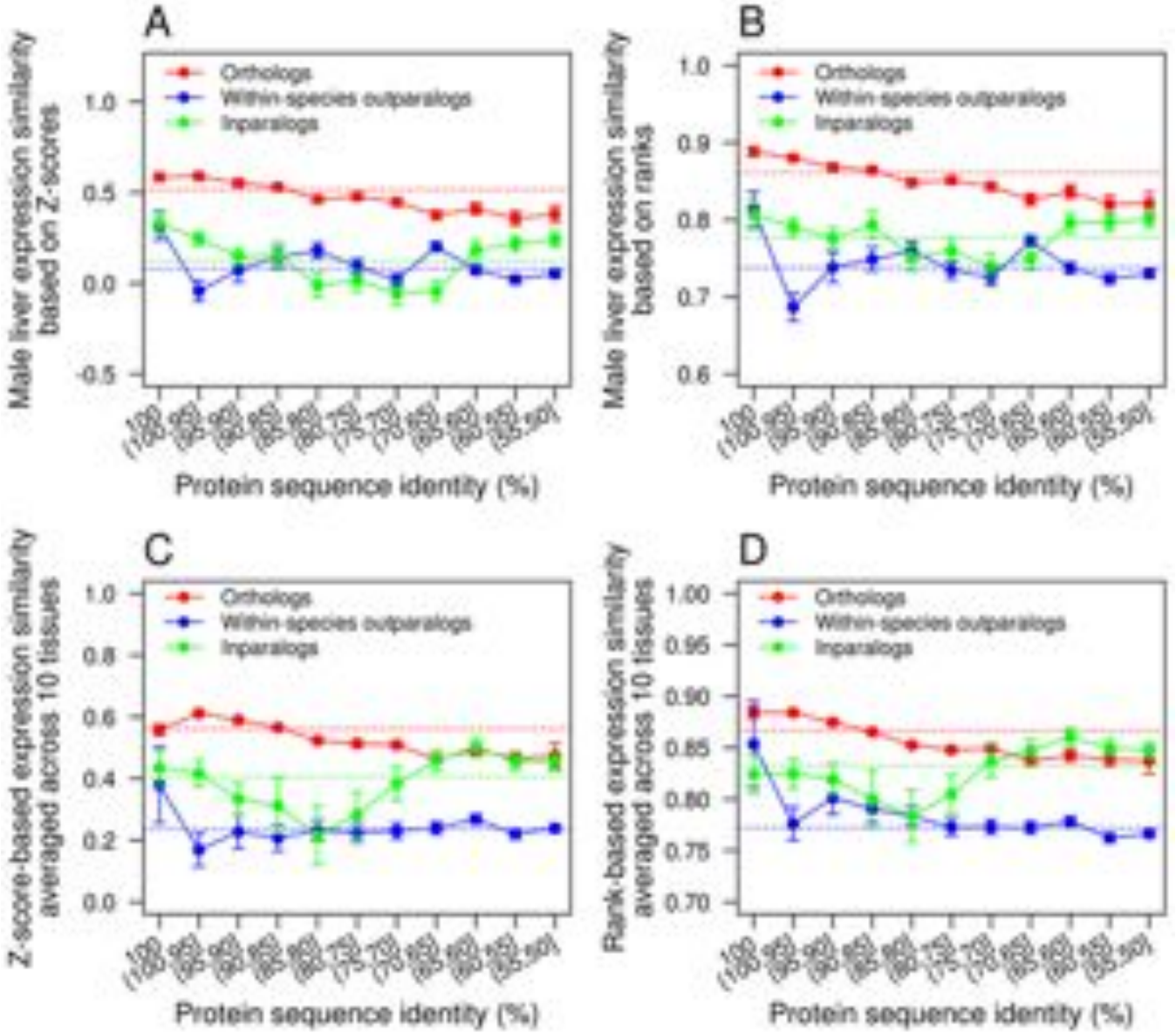
Neofunctionalization

HOG name	Category	Expression singleton	Expression dupl. 1	Expression dupl. 2
cerebellum	Overlap	Expressed	Expressed	Expressed
archinephric duct	Overlap	Expressed	Expressed	Expressed
mesenchyme	Overlap	Expressed	Expressed	Expressed
somite	Overlap	Expressed	Expressed	Expressed
retina	Overlap	Expressed	Expressed	Expressed
ovary	Overlap	Expressed	Expressed	Expressed
testis	Overlap	Expressed	Expressed	Expressed
liver	Overlap	Expressed	Expressed	Expressed
heart	Overlap	Expressed	Expressed	Expressed
telencephalon	Overlap	Expressed	Expressed	Expressed
inner ear	Overlap	Expressed	Expressed	Expressed
olfactory organ	Overlap	Expressed	Expressed	Expressed
viscerocranium	Overlap	Expressed	Expressed	Expressed
diencephalon	Overlap	Expressed	Expressed	Expressed
presomitic mesoderm	Overlap	Expressed	Expressed	Expressed
skin	Overlap	Expressed	Expressed	Expressed
tectum	Overlap	Expressed	Expressed	Expressed
forelimb - pectoral fin	Overlap	Expressed	Expressed	Expressed
mesonephros	Overlap	Not expressed	Expressed	Expressed
pharynx	Overlap	Not expressed	Expressed	Expressed
neural tube	Neofunctionalization	Not expressed	Expressed	Not expressed



UNIL | Université de Lausanne
Département d'écologie
et évolution

SIB
Swiss Institute of
Bioinformatics



Chen X, Zhang J (2012) PLoS Comput Biol 8(11): e1002784

Expression diverges (a bit) between paralogs

Let's work!

examine human gene you studied in Genomicus in:

http://selectome.unil.ch/

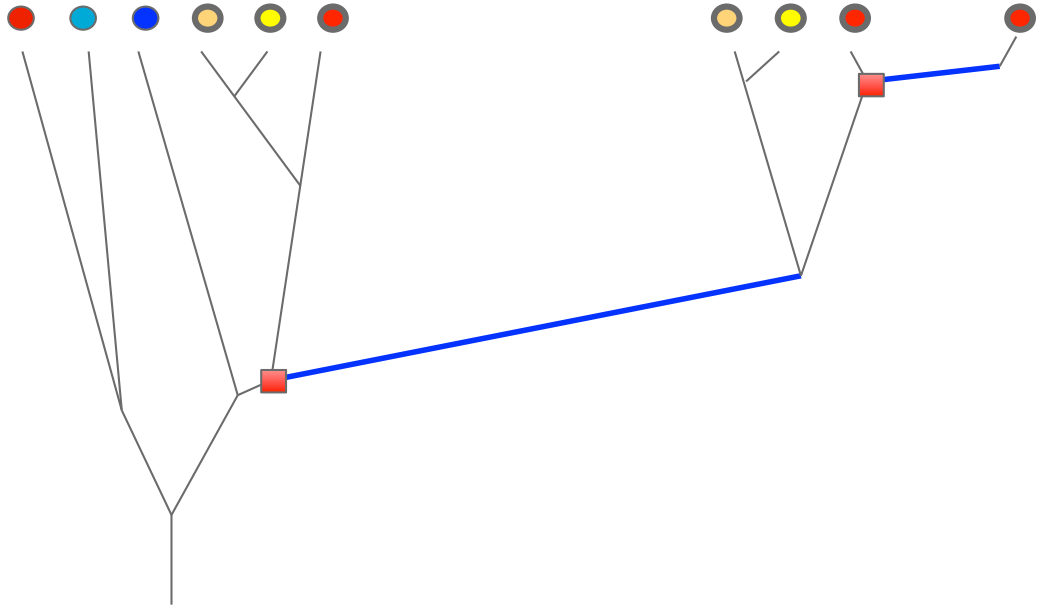
http://bgee.unil.ch/

- exclude Xenopus
- chose Data quality "high"
- show gene details
- display all homologous organ groups

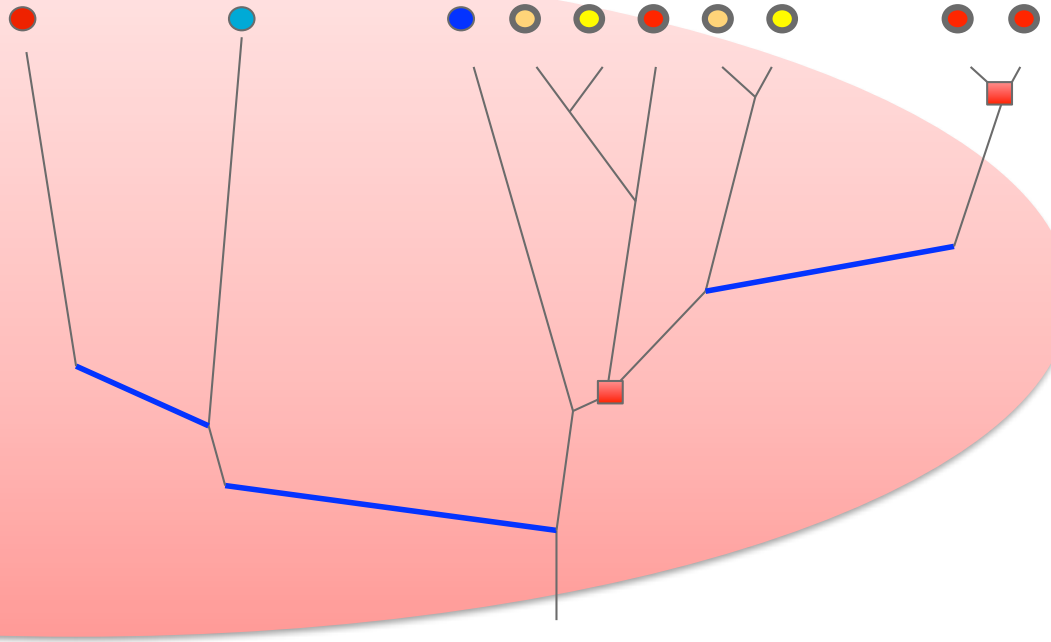
- **overall picture of evolution of each of these genes**

Support for alternative model

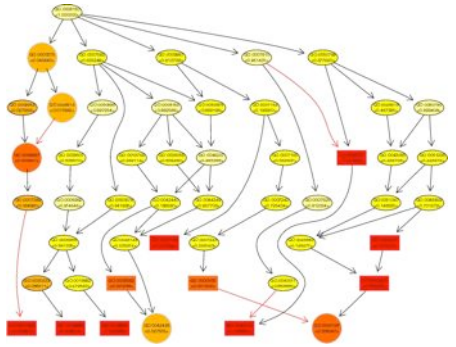
standard model:
functional change after duplication



alternative model:
functional change with time

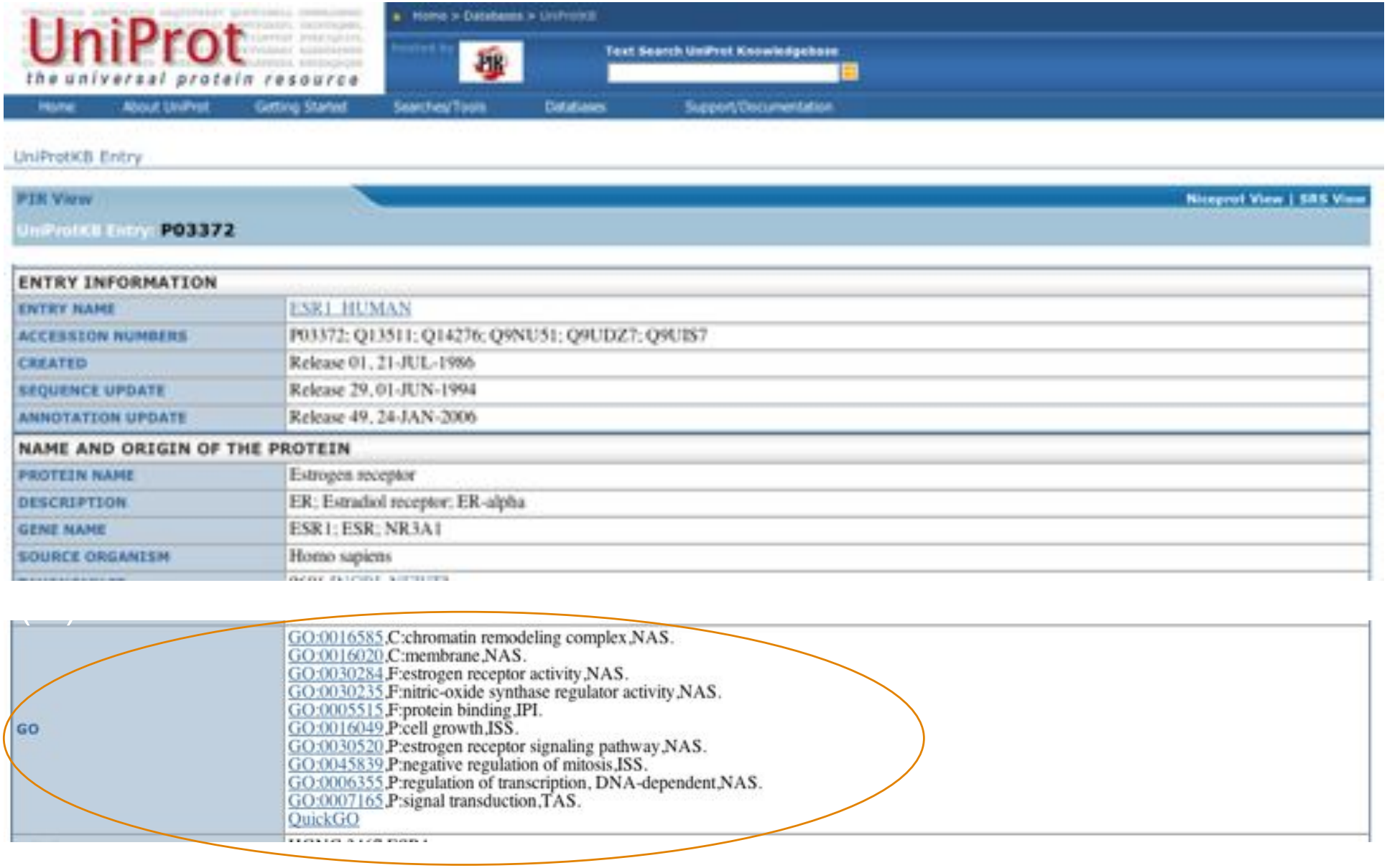


Phylogenomics of function



UNIL | Université de Lausanne
Département d'écologie et évolution

SIB
Swiss Institute of Bioinformatics


UniProt
the universal protein resource
Home > Databases > UniProtKB
Text Search UniProt Knowledgebase
Home
About UniProt
Getting Started
Searches/Tools
Databases
Support/Documentation
UniProtKB Entry
PIR View
Niceprot View | SRS View
UniProtKB Entry: P03372
ENTRY INFORMATION
ENTRY NAME
ESR1_HUMAN
ACCESSION NUMBERS
P03372; Q13511; Q14276; Q9NU51; Q9UDZ7; Q9UES7
CREATED
Release 01, 21-JUL-1986
SEQUENCE UPDATE
Release 29, 01-JUN-1994
ANNOTATION UPDATE
Release 49, 24-JAN-2006
NAME AND ORIGIN OF THE PROTEIN
PROTEIN NAME
Estrogen receptor
DESCRIPTION
ER; Estradiol receptor; ER-alpha
GENE NAME
ESR1; ESR; NR3A1
SOURCE ORGANISM
Homo sapiens
GO
GO:0016585,C:chromatin remodeling complex,NAS.
GO:0016020,C:membrane,NAS.
GO:0030284,F:estrogen receptor activity,NAS.
GO:0030235,F:nitric-oxide synthase regulator activity,NAS.
GO:0005515,F:protein binding,IPI.
GO:0016049,P:cell growth,ISS.
GO:0030520,P:estrogen receptor signaling pathway,NAS.
GO:0045839,P:negative regulation of mitosis,ISS.
GO:0006355,P:regulation of transcription, DNA-dependent,NAS.
GO:0007165,P:signal transduction,TAS.
QuickGO

Gene Ontology

- 17 genomes with highest GO coverage
- experimental evidence codes: EXP, IDA, IEP, IGI, IMP, IPI
- from GOA and Ensembl
- orthologs and paralogs inferred by OMA
- confirmation with Ensembl

Per species pair:
- one-to-one orthologs
- 1-to-many and many-to-many orthologs
- different-species paralogs.

Per species:
- same-species paralogs

-> 26 330 pairs of genes with experimental GO annotations for both

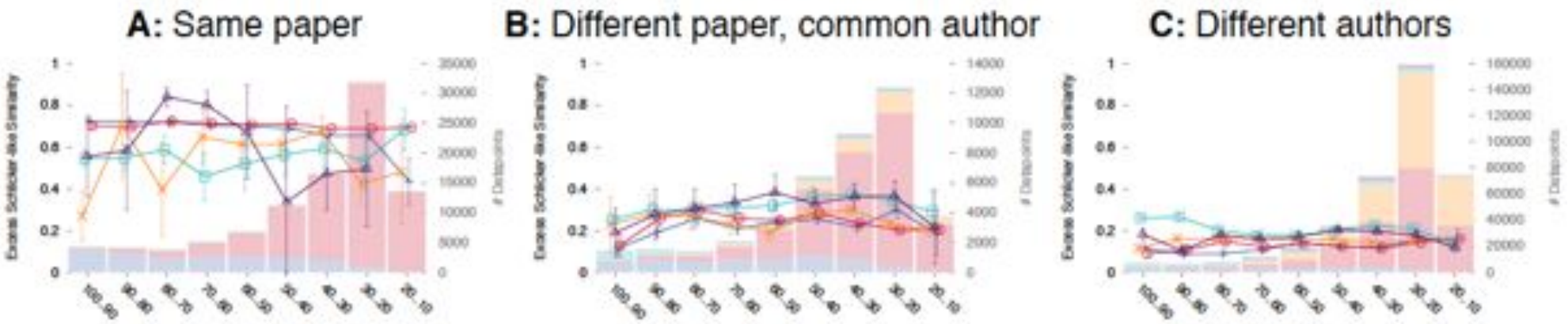
UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

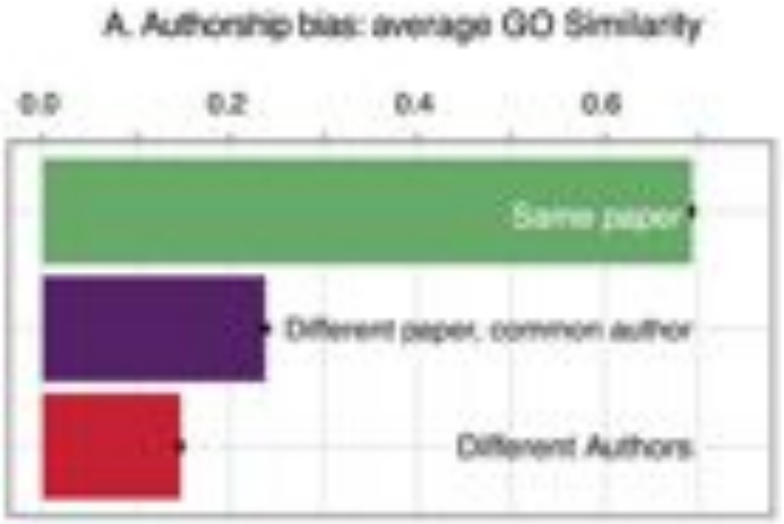
Beware of the bias

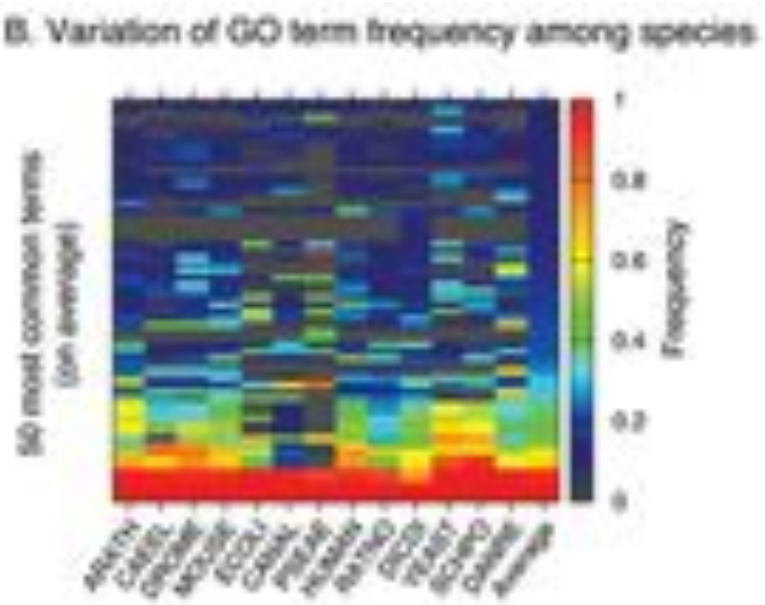
Dataset	Same-Species Paralogs	Different-Species Paralogs	1:1 Orthologs	Other Orthologs
Same Publications	1573	18	154	44
Different Publications, Same Authors	613	382	874	434
No Common Author	2492	20719	13149	11296
All Experimental Annotations	3312	20766	13309	11371

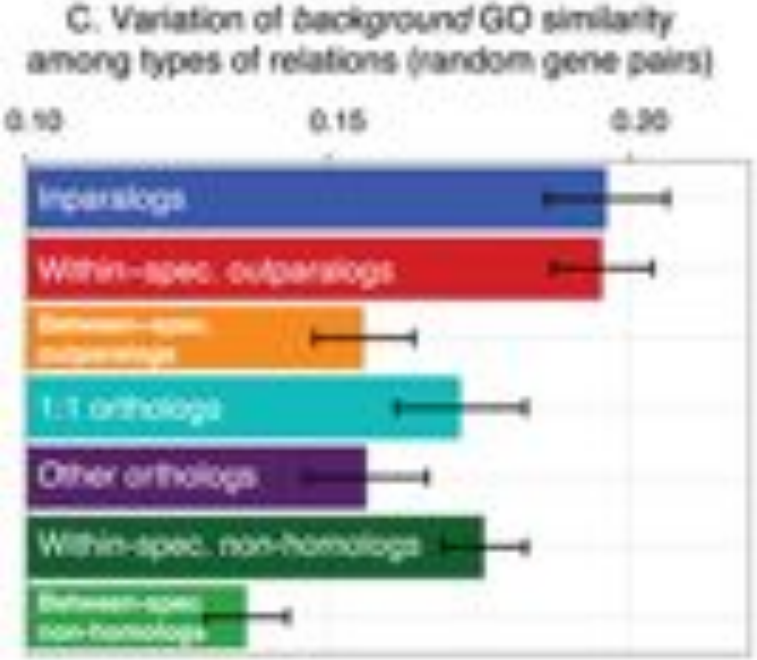
47%

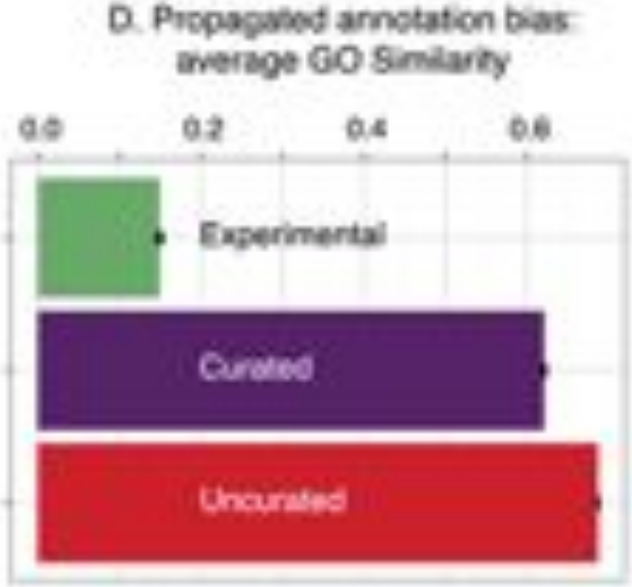
1%

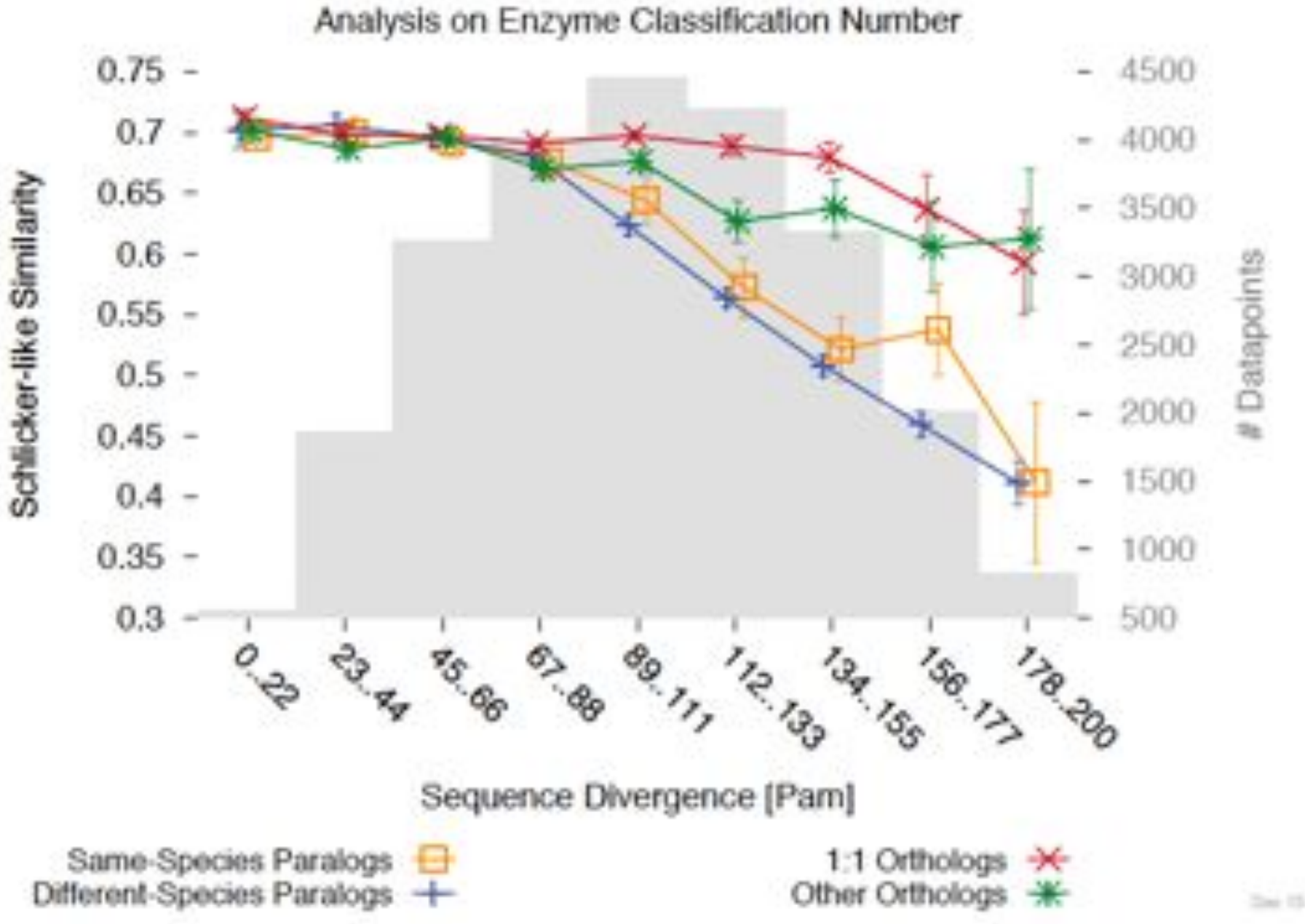



A. Authorship bias: average GO Similarity
0.0
0.2
0.4
0.6
Same paper
Different paper, common author
Different Authors


B. Variation of GO term frequency among species
50 most common terms (on average)
Frequency
1
0.8
0.6
0.4
0.2
0

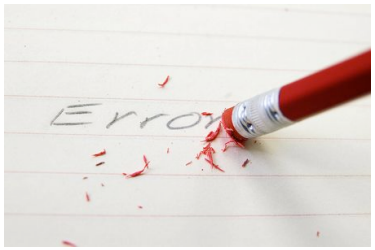

C. Variation of background GO similarity among types of relations (random gene pairs)
0.10
0.15
0.20
Inparalogs
Within-spec. outparalogs
1:1 orthologs
Other orthologs
Within-spec. non-homologs

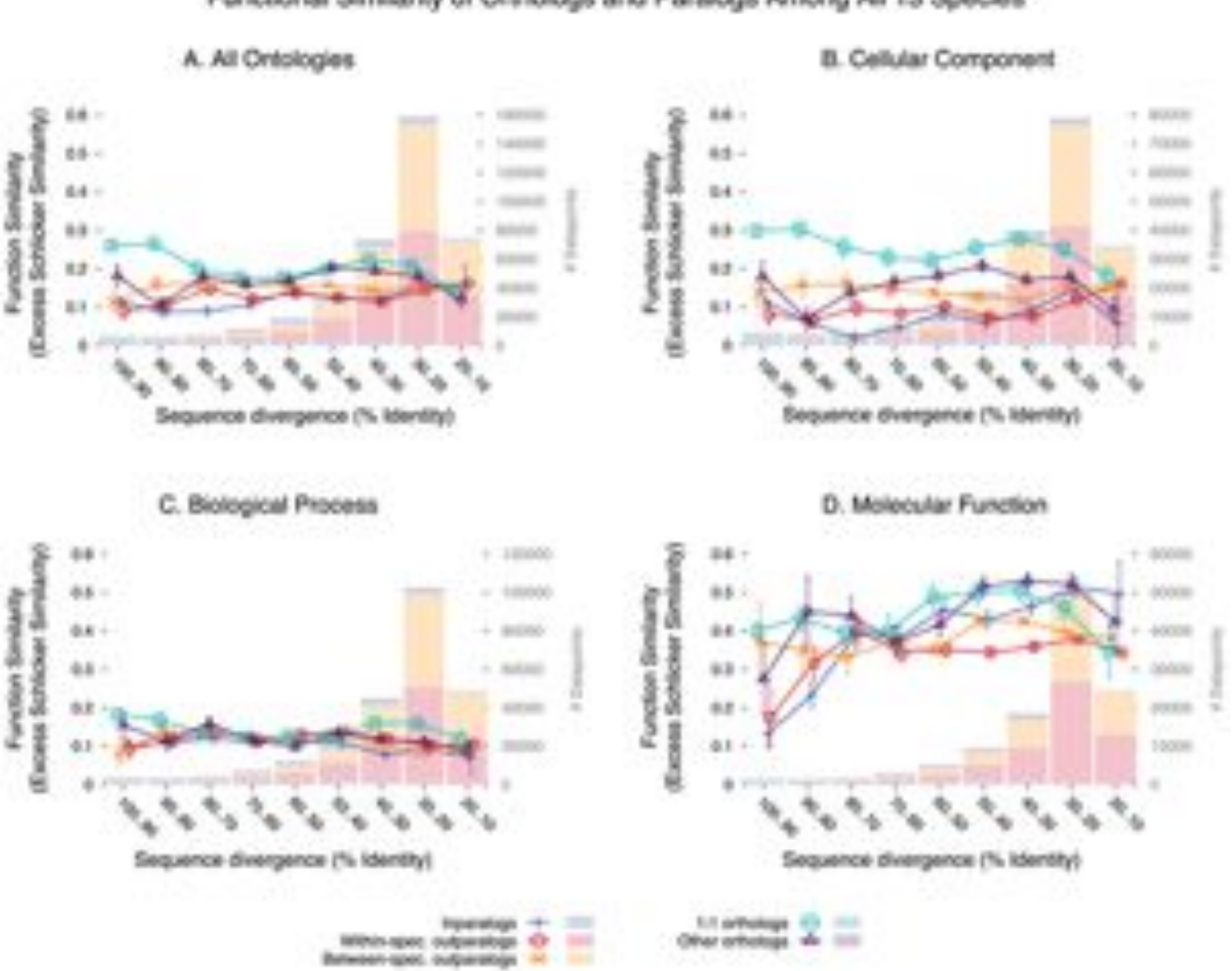

D. Propagated annotation bias: average GO Similarity
0.0
0.2
0.4
0.6
Experimental
Curated
Uncurated



Altenhoff et al, u*npublished*

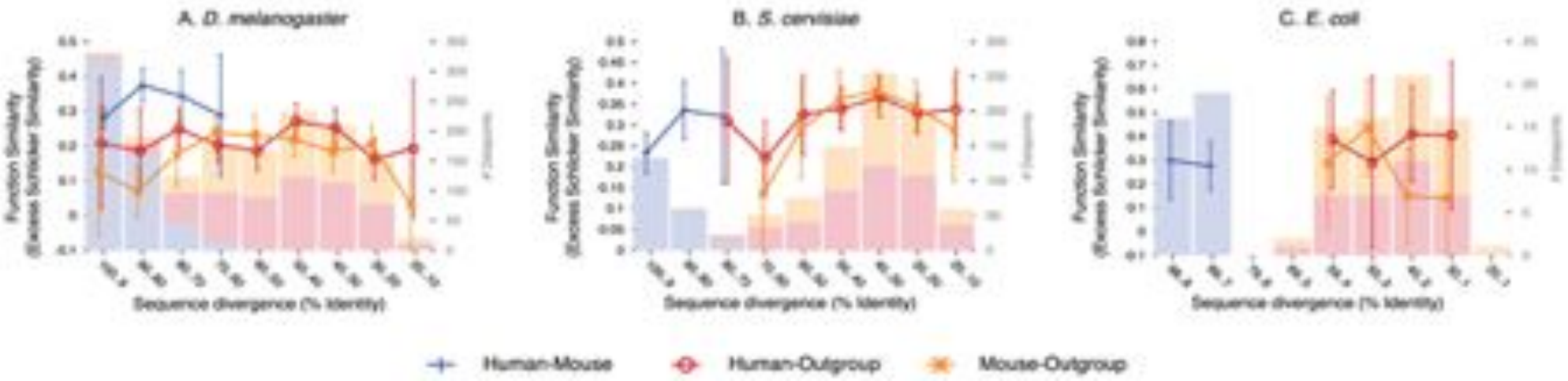
Comparisons after correcting for biases





UNIL | Université de Lausanne
Département d'écologie et évolution
SIB Swiss Institute of Bioinformatics

Orthologs are useful


Functional Similarity of corresponding 1:1 Orthologs in Human, Mouse and Outgroup
A. D. melanogaster
B. S. cerevisiae
C. E. coli
Function Similarity (Excess Schlicker Similarity)
Sequence divergence (% identity)
Human-Mouse
Human-Outgroup
Mouse-Outgroup

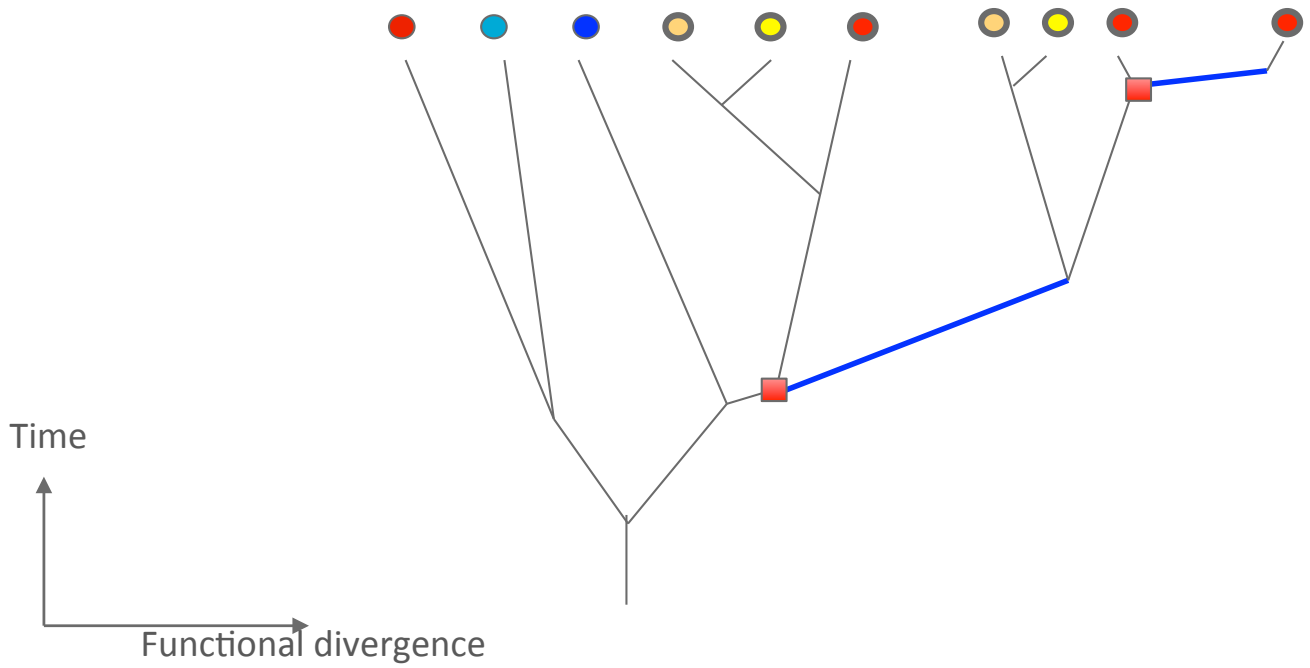
Conclusion orthologs/ paralogs

UNIL | Université de Lausanne
Département d'écologie et évolution

SIB Swiss Institute of Bioinformatics

The standard model holds, but is weaker than expected.

Duplication does affect gene evolution
Other factors are significant in gene function divergence



General conclusion

- Duplication happens:
  - Duplication is frequent
  - Genome duplication has shaped many eukaryotic genomes

- Function:
  - Duplicate genes can diverge and add new functions
  - Changes in function can concern
    - expression
    - protein sequence
    - interactions
    - and more
  - Duplicate genes can be conserved
  - Orthologs can diverge

Duplication is a major force in the evolution of genomes

UNIL | Université de Lausanne
Département d'écologie
et évolution

SIB
Swiss Institute of
Bioinformatics