Phylogenetic inference in the context of whole genome duplication

Aoife McLysaght

Trinity College Dublin, The University of Dublin

Trinity College Dublin

Coláiste na Tríonóide, Baile Átha Cliath

The University of Dublin

Key papers and funding

ARTICLE

https://doi.org/10.1038/s41467-021-24573-z OPEN

Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution

Yoichiro Nakatani[1,4,5], Prashant Shingate[2,5], Vydianathan Ravi[2], Nisha E. Pillai[2], Aravind Prasad[2], Aoife McLysaght[1✉] & Byrappa Venkatesh[2,3✉]

Article

Independent rediploidization masks shared whole genome duplication in the sturgeon-paddlefish ancestor

Received: 11 July 2022

Accepted: 12 May 2023

Anthony K. Redmond[1], Deartháile Casey[1], Manu Kumar Gundappa[2], Daniel J. Macqueen[2] & Aoife McLysaght[1✉]

ARTICLE

https://doi.org/10.1038/s41467-021-22074-7 OPEN

Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding

Anthony K. Redmond[1] & Aoife McLysaght[1✉]

ARTICLE

Received 30 Jun 2016 | Accepted 20 Dec 2016 | Published 8 Feb 2017 DOI: 10.1038/ncomms14366 OPEN

Dosage sensitivity is a major determinant of human copy number variant pathogenicity

Alan M. Rice[1] & Aoife McLysaght[1]

letter

Extensive genomic duplication during early chordate evolution

Aoife McLysaght*, Karsten Hokamp* & Kenneth H. Wolfe

*These authors contributed equally to this work.

Ohnologs in the human genome are dosage balanced and frequently associated with disease

Takashi Makino[1] and Aoife McLysaght[2]

Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland

Check for updates

Letter

Recent de novo origin of human protein-coding genes

David G. Knowles and Aoife McLysaght[1]

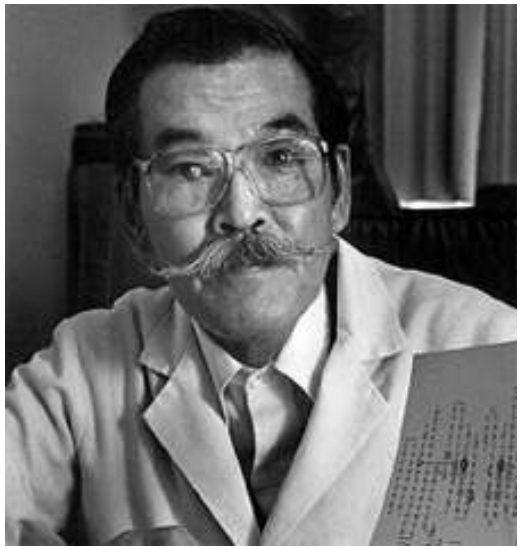
Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland

European Research Council

Established by the European Commission


Public Engagement

Whole Genome Duplication (WGD)/Polyploidy



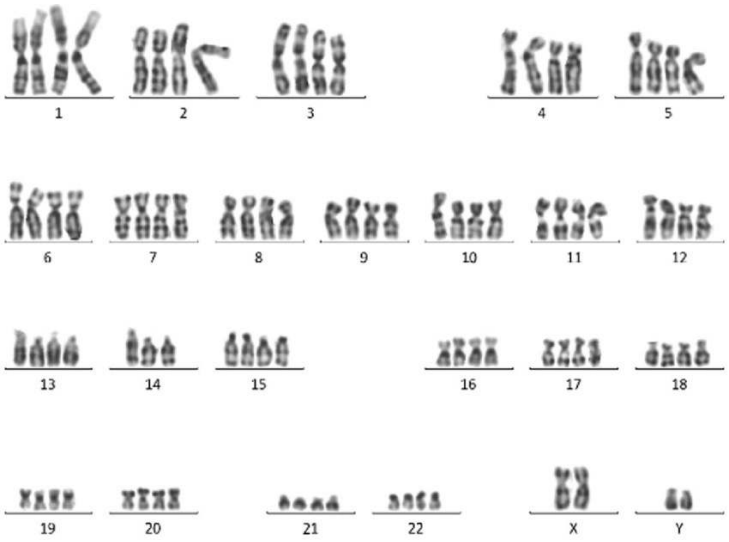
Susumo Ohno (1928-2000)

Evolution by Gene Duplication (1970)

- highly influential book

Ohno championed the idea that gene and genome duplication were powerful creative forces in evolution.

WGD now known to have occurred at the establishment of major lineages of plants and animals.



Tetraploid human cancer cell

(an example of the mutation – in cancer it is an evolutionary dead-end)

WGD on the tree of life

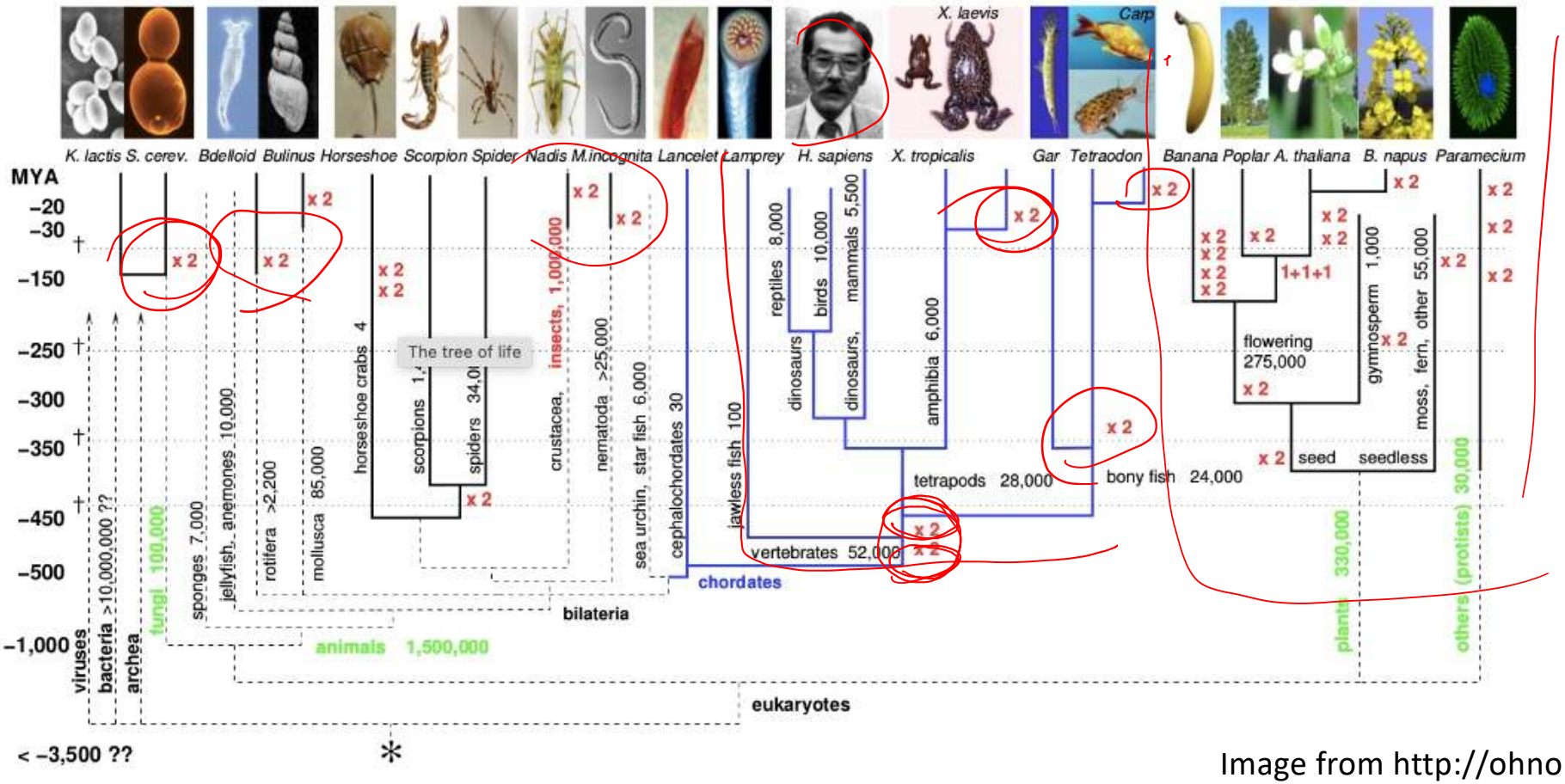


Image from http://ohnologs.curie.fr/

Gene or Genome duplication?

Whole Genome Duplication (WGD)

- All genes duplicated in a single event
- Including all regulatory sequences
- Preserves ratios of genes/products
- BUT: destabilises meiosis and other cellular processes
- Hard to find a mate.

- Potentially associated with rare major evolutionary events

Small-scale Duplication (SSD)

- One or several genes duplicated at a time
- A readily available route to new genes

- BUT: disturbs the ratios (stoichiometry) of some gene products in deleterious ways

- Common, less dramatic evolutionary change

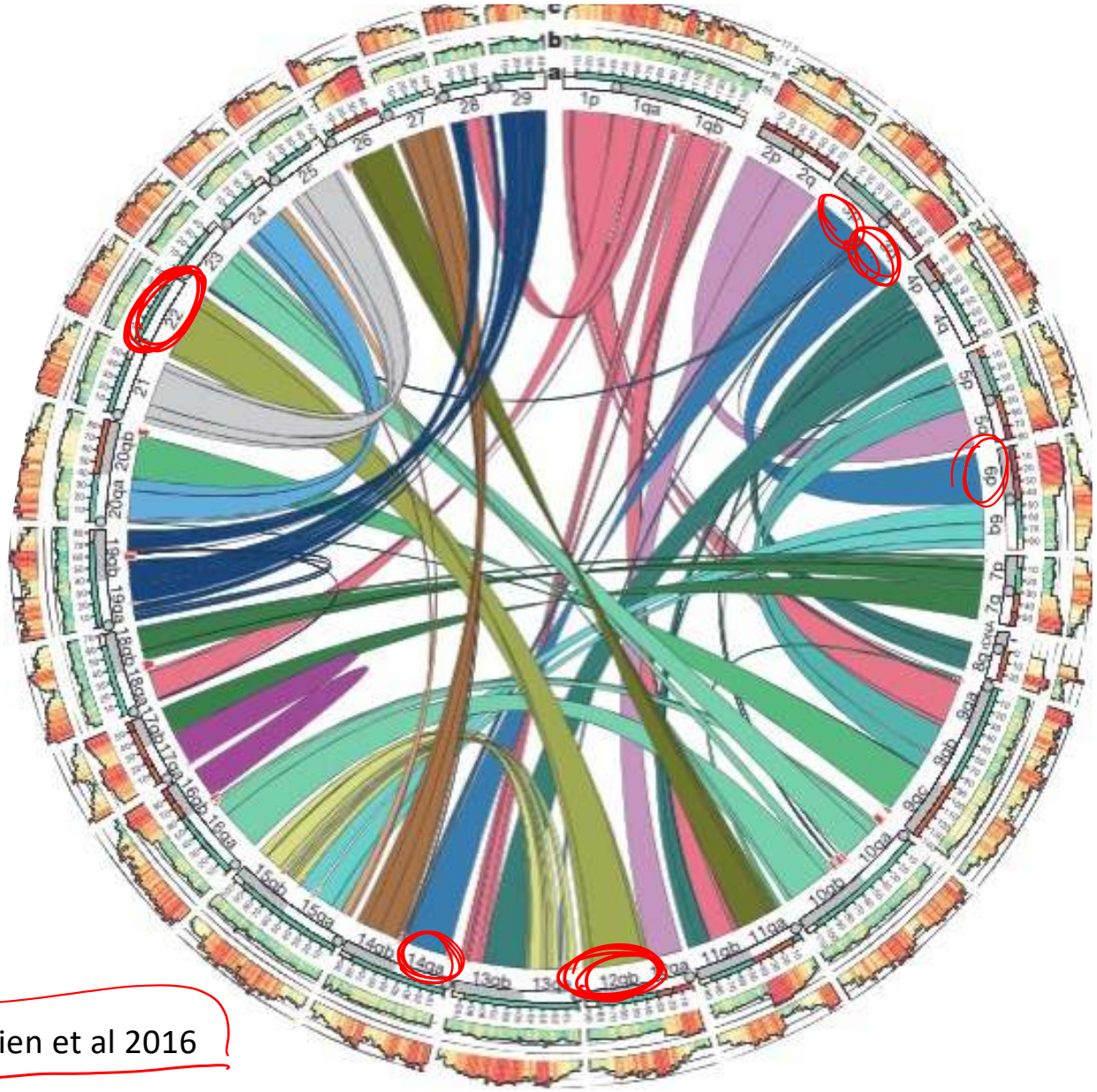
Major questions

- Where has WGD occurred in the tree of life?
  - How do we identify WGD?
- What evolutionary impact did it have?

- What implications does it have for understanding genomes?
- What is the legacy of WGD?

WGD has profound and lasting effects on genomes

- Massive genome rearrangements
- (evidence from salmonids and teleosts)

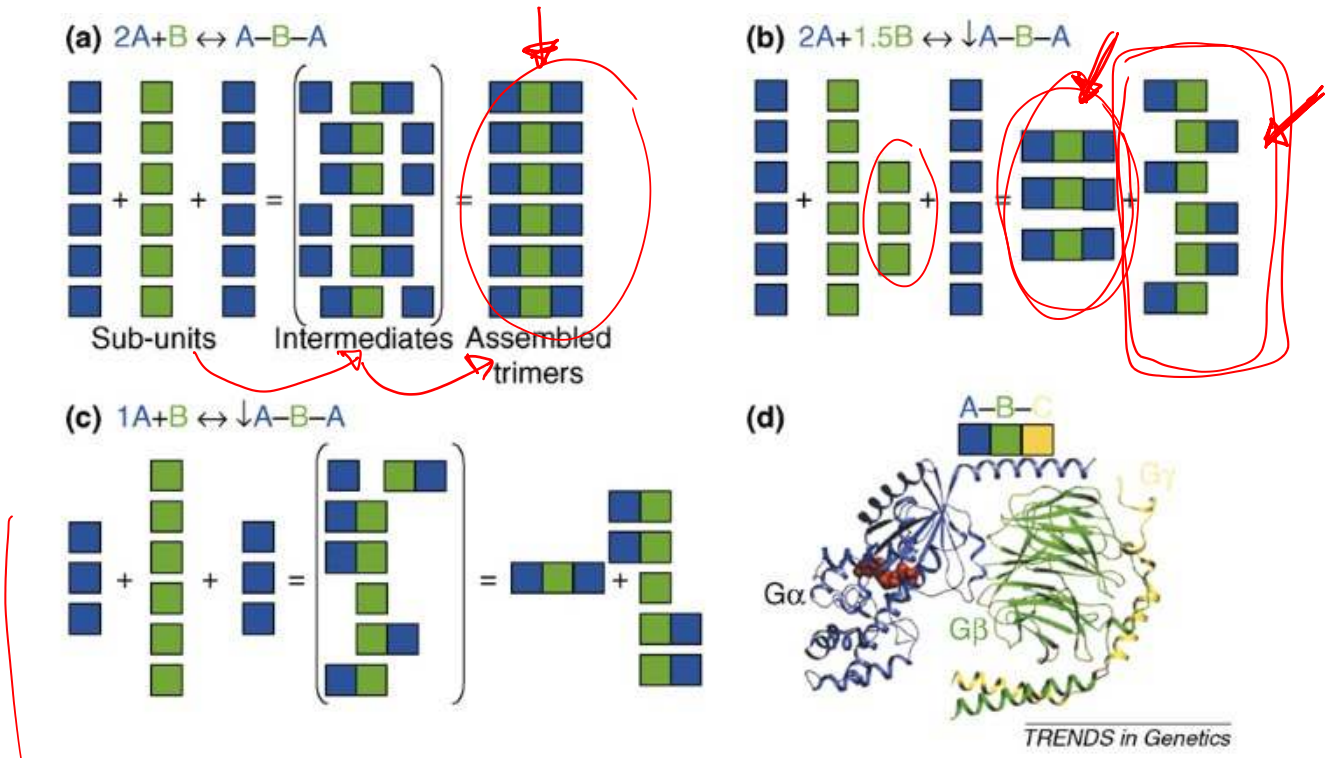


Lien et al 2016

WGD has profound and lasting effects on genomes

- Massive genome rearrangements
- Biased retention of dosage sensitive genes

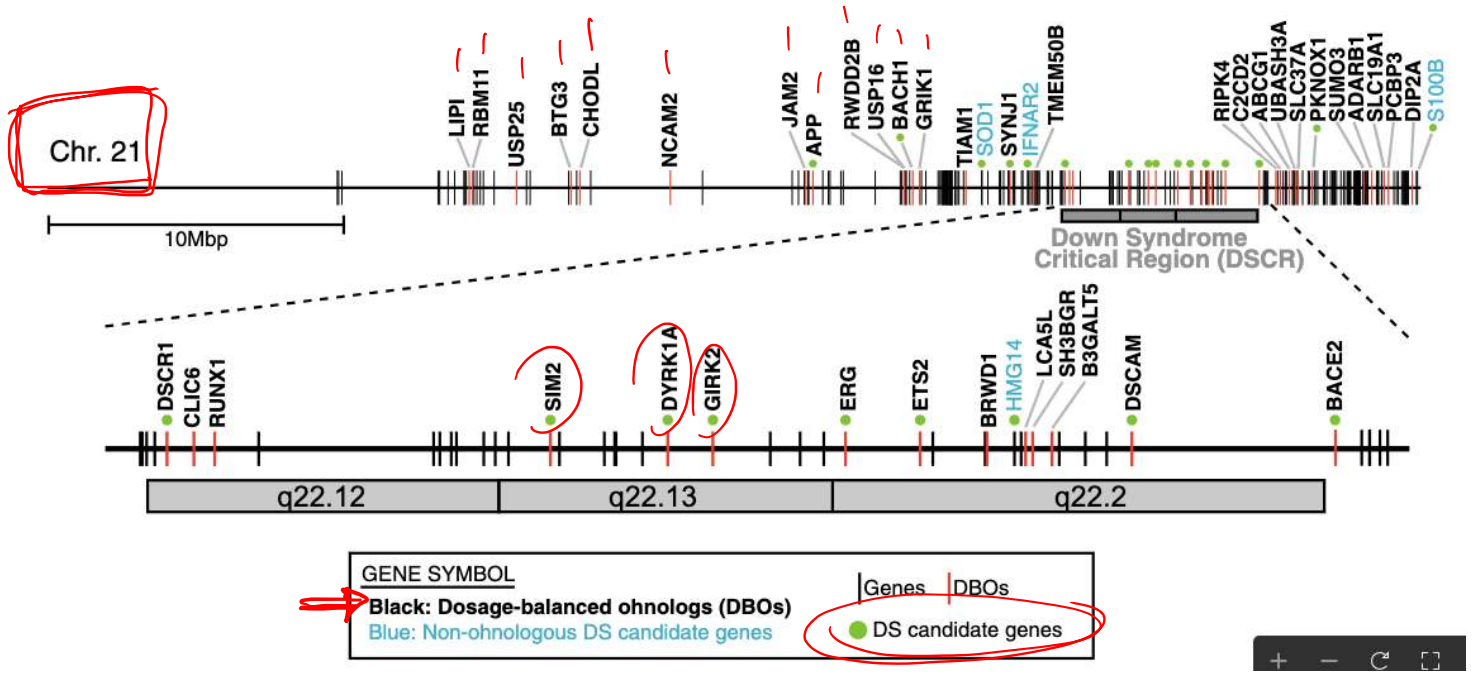
Dosage balanced



Veitia et al 2008

WGD has profound and lasting effects on genomes

- Massive genome rearrangements
- Biased retention of dosage sensitive genes



Makino et al 2010

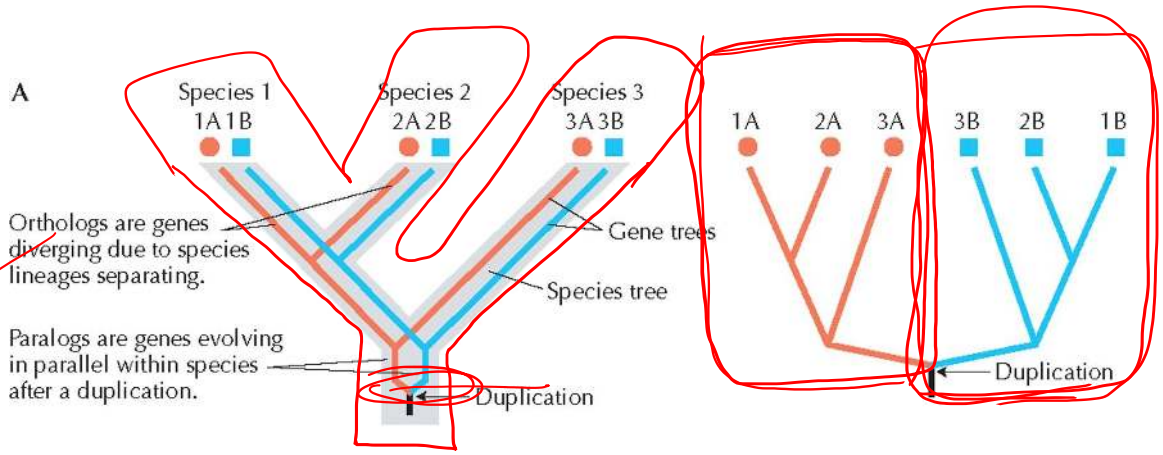
WGD has profound and lasting effect on genomes

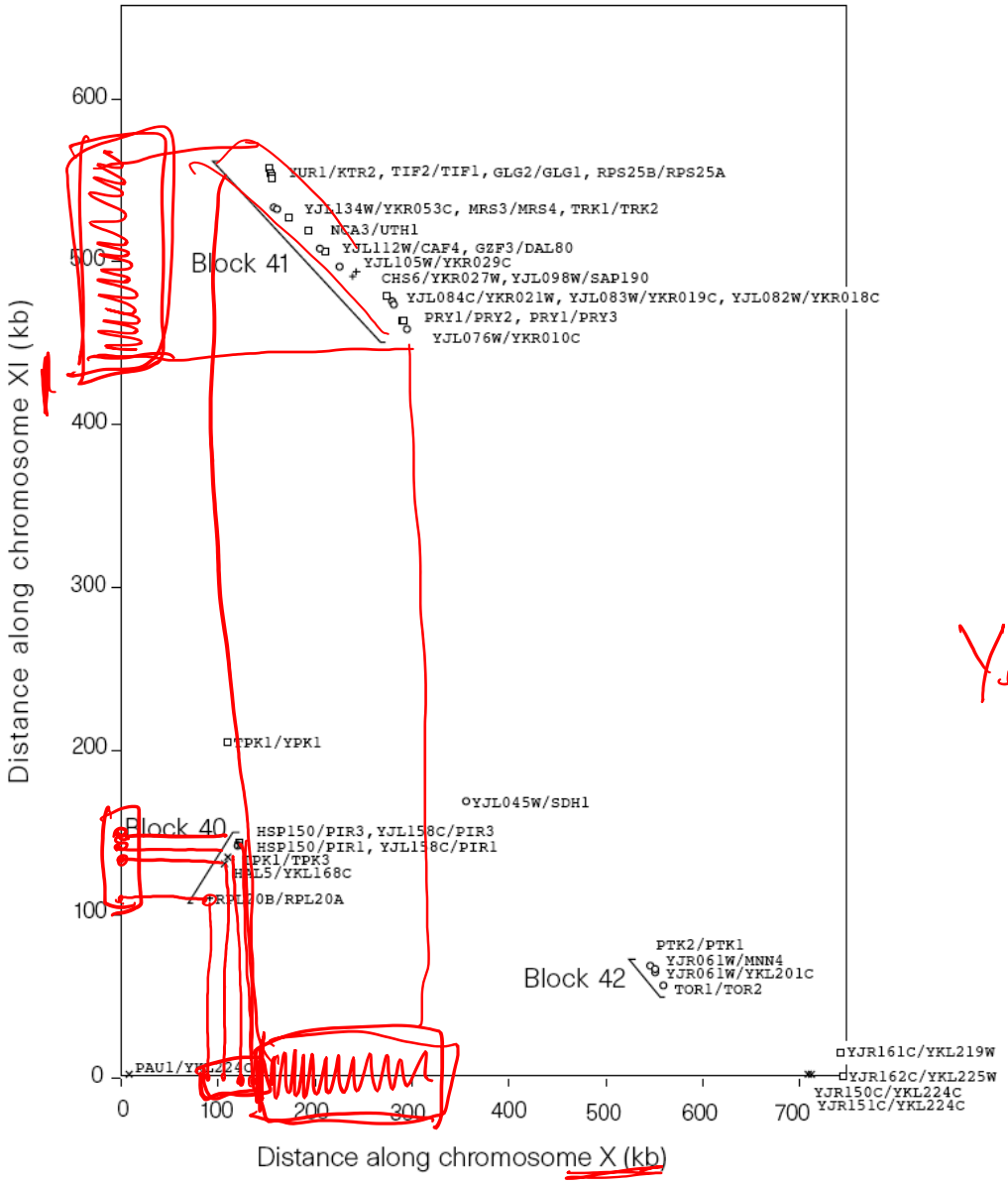
- Massive genome rearrangements
- Biased retention of dosage sensitive genes
  - implications for disease gene association

Rice & McLysaght 2017

How to detect WGD

- Testable predictions:
  - All genes in the genome (initially) duplicated
    - genome maps
    - gene order conserved blocks
    - Genome-wide 1:2 relationship with outgroups
  - At the same time
    - **Phylogenetics**
    - date estimates
    - tree topology (branching order)



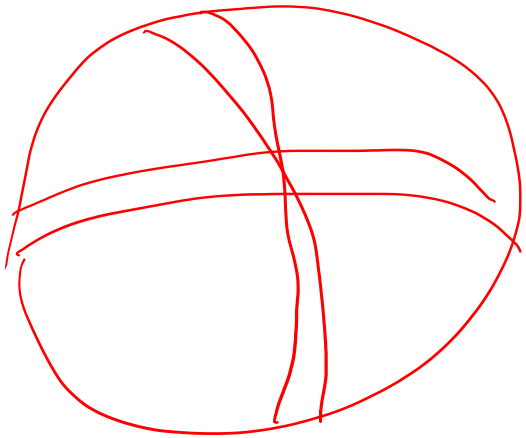


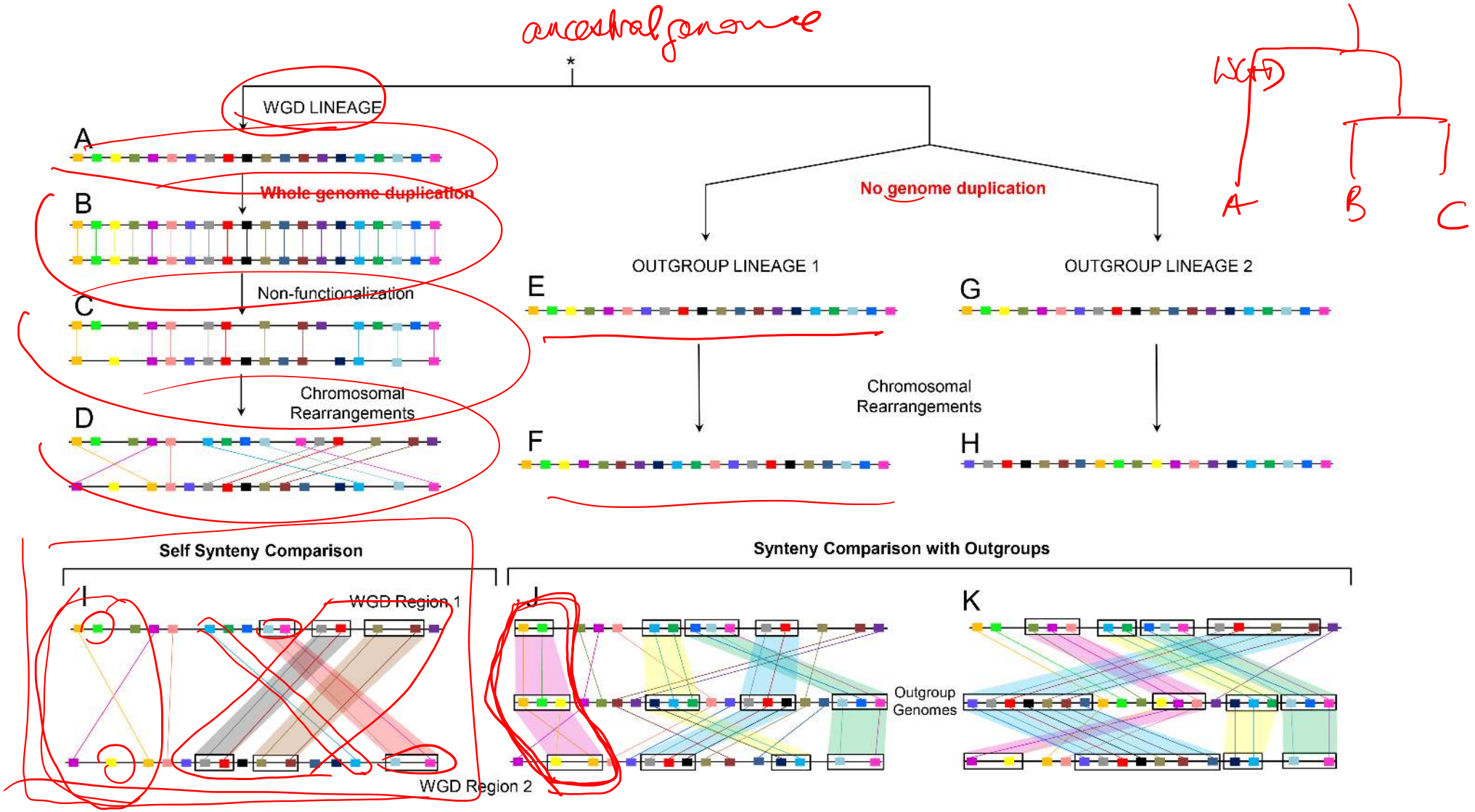
WGD in *Saccharomyces cerevisiae*:

Wolfe & Shields, Nature 387:708 (1997)
PMID 9192896

Genome map

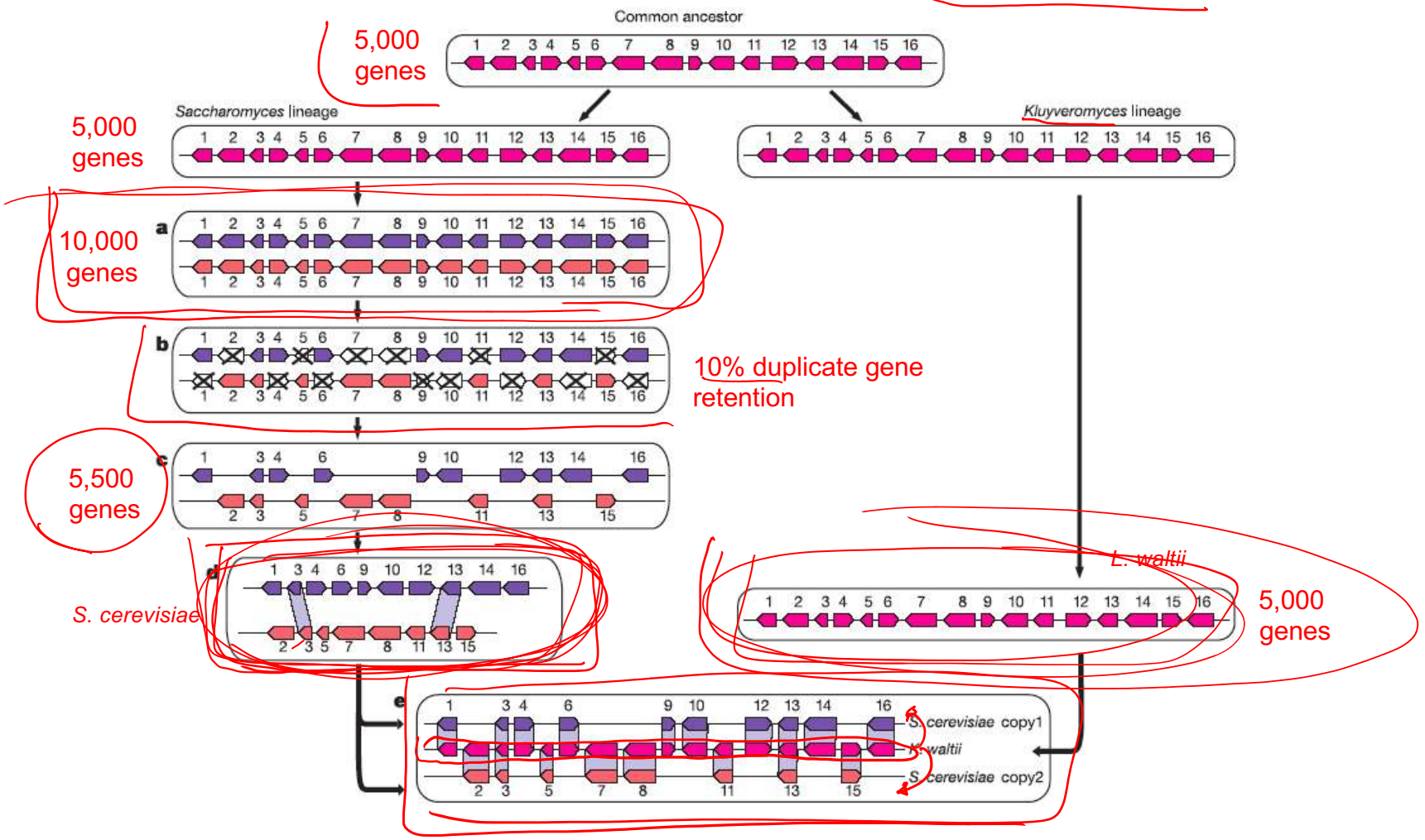
Yeast



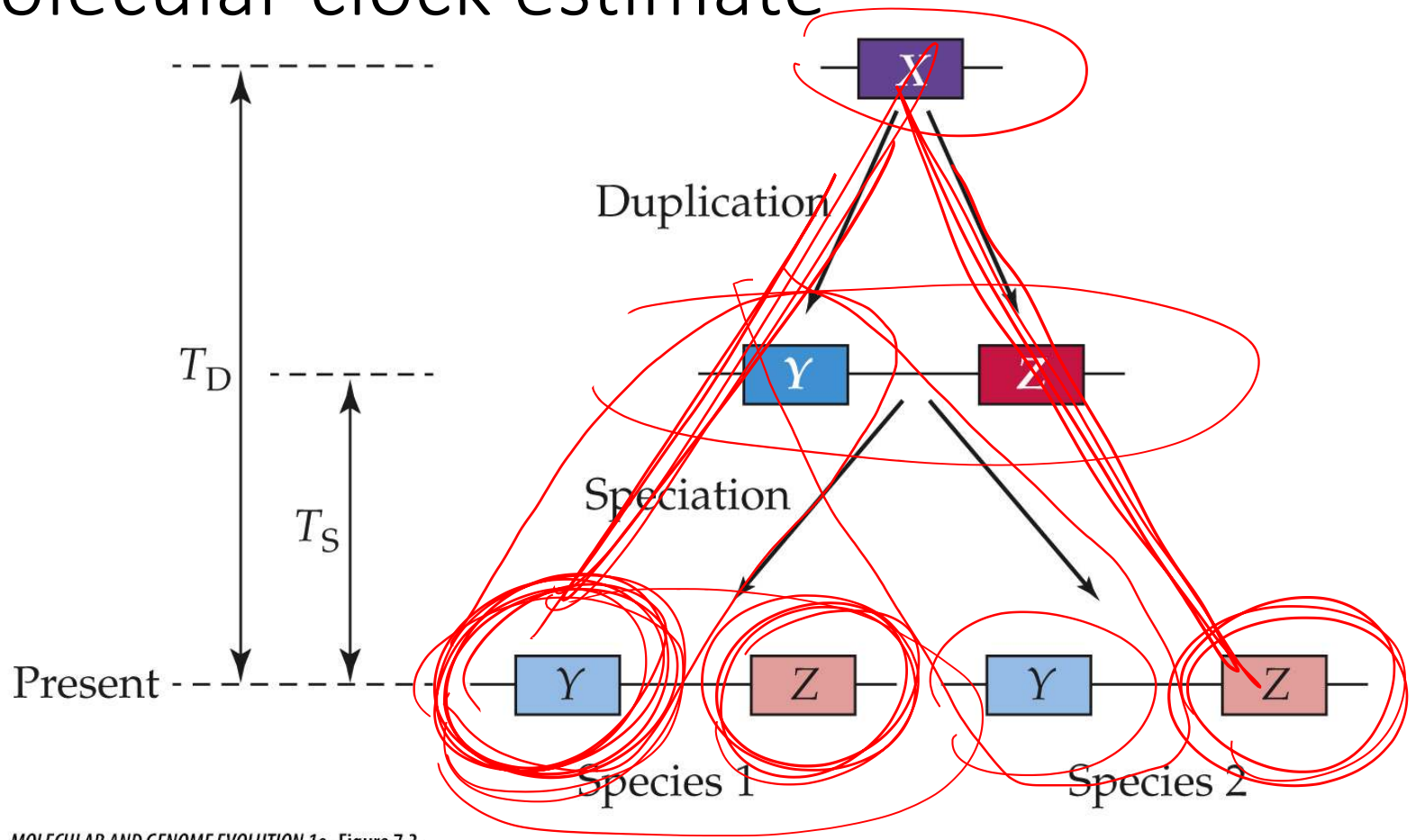


1.Singh, P. P., Arora, J. & Isambert, H. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLoS Computational Biology* **11**, e1004394 (2015).

** Kellis et al, Nature 428:617-624 (2004) PMID 15004568 – *Lachancea (Kluyveromyces) waltii* genome

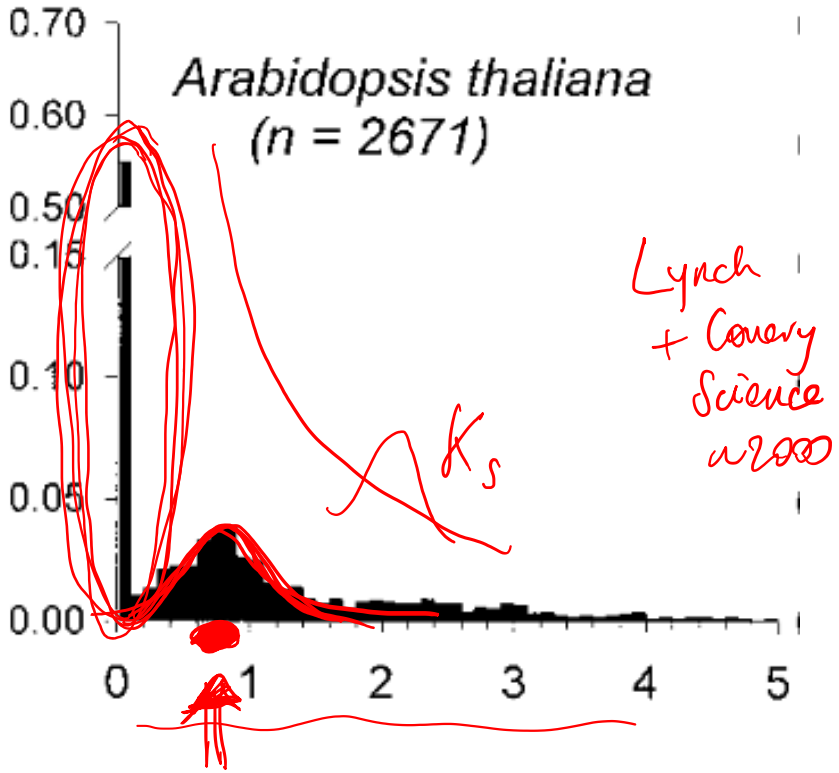


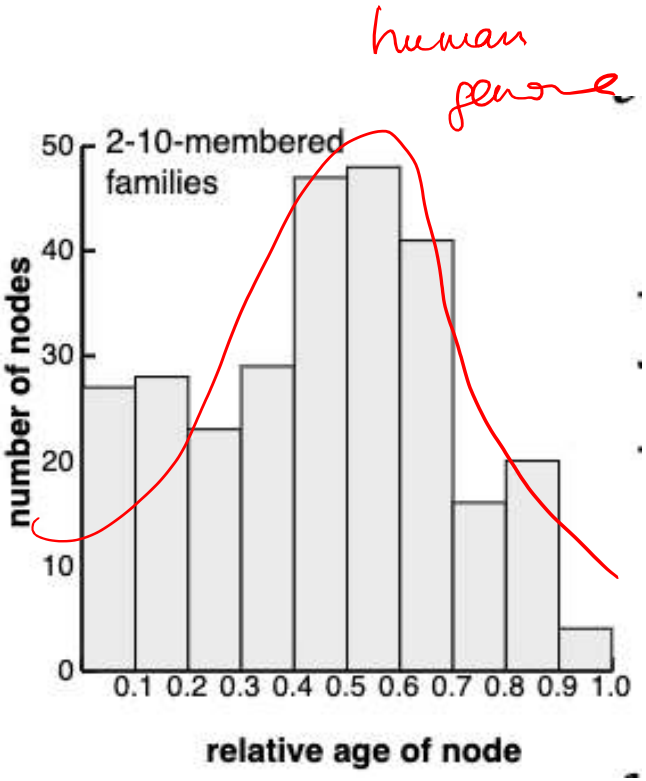
Evidence from inferred dates of duplication: molecular clock estimate



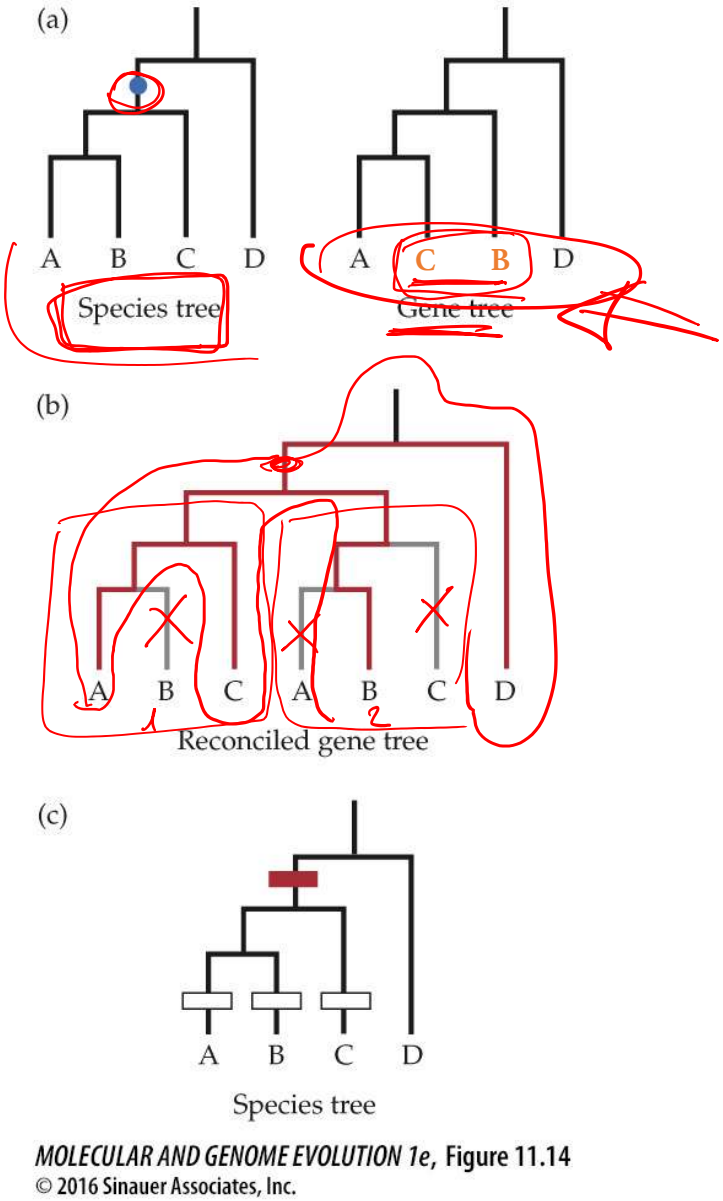
MOLECULAR AND GENOME EVOLUTION 1e, Figure 7.3

Evidence from inferred dates of duplication: molecular clock estimate



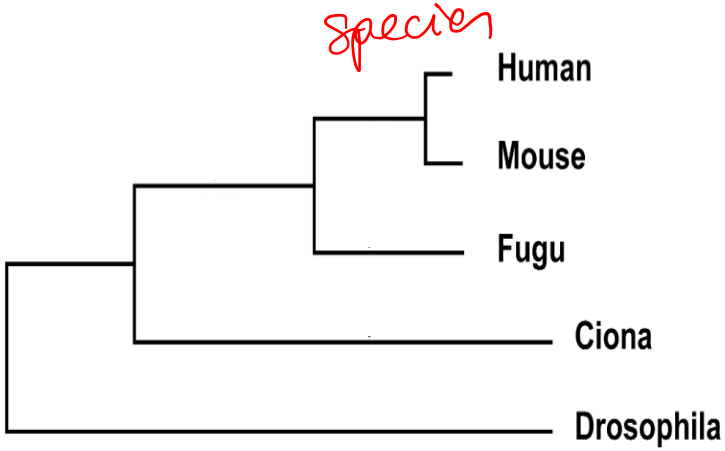


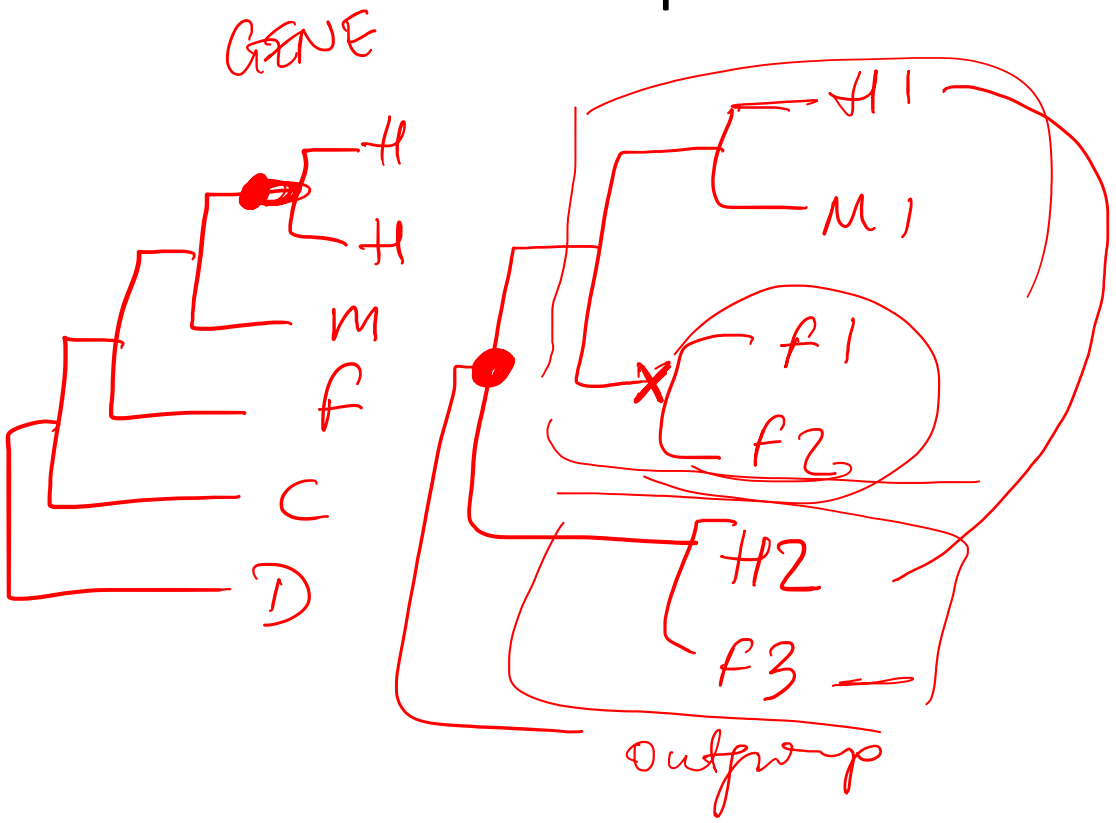
Evidence from inferred dates of duplication: branching order



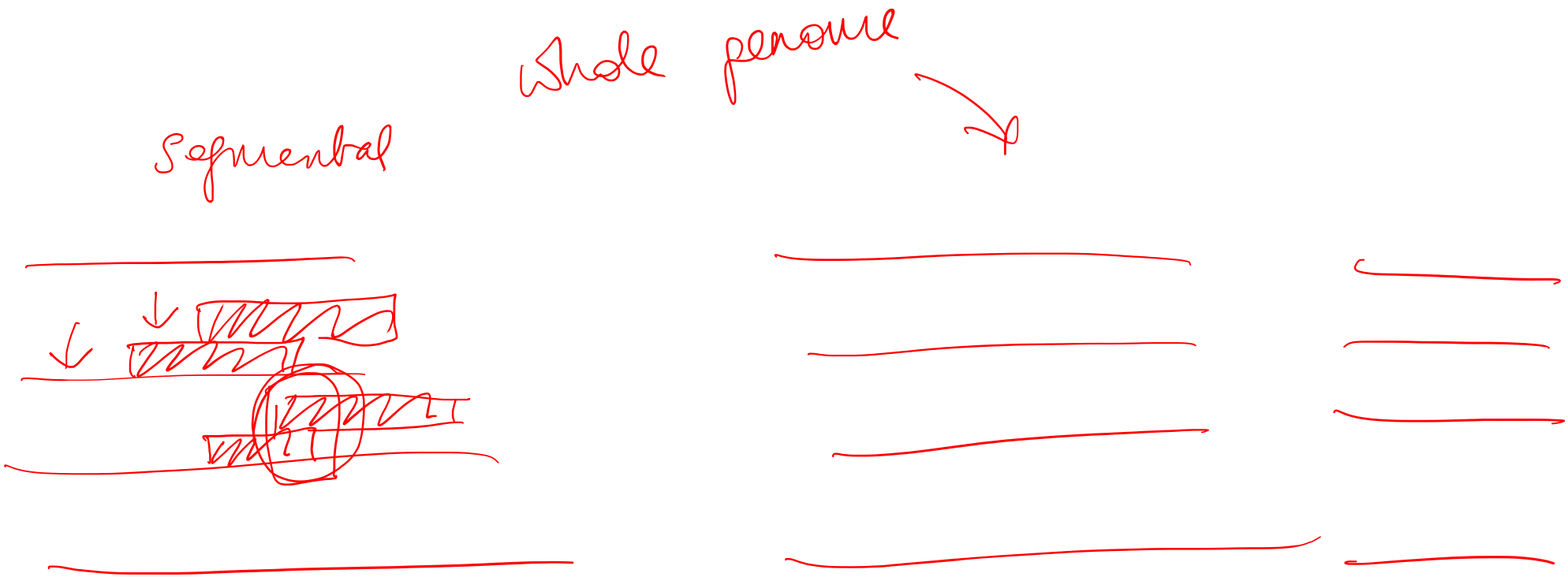
MOLECULAR AND GENOME EVOLUTION 1e, Figure 11.14

Evidence from inferred dates of duplication: branching order

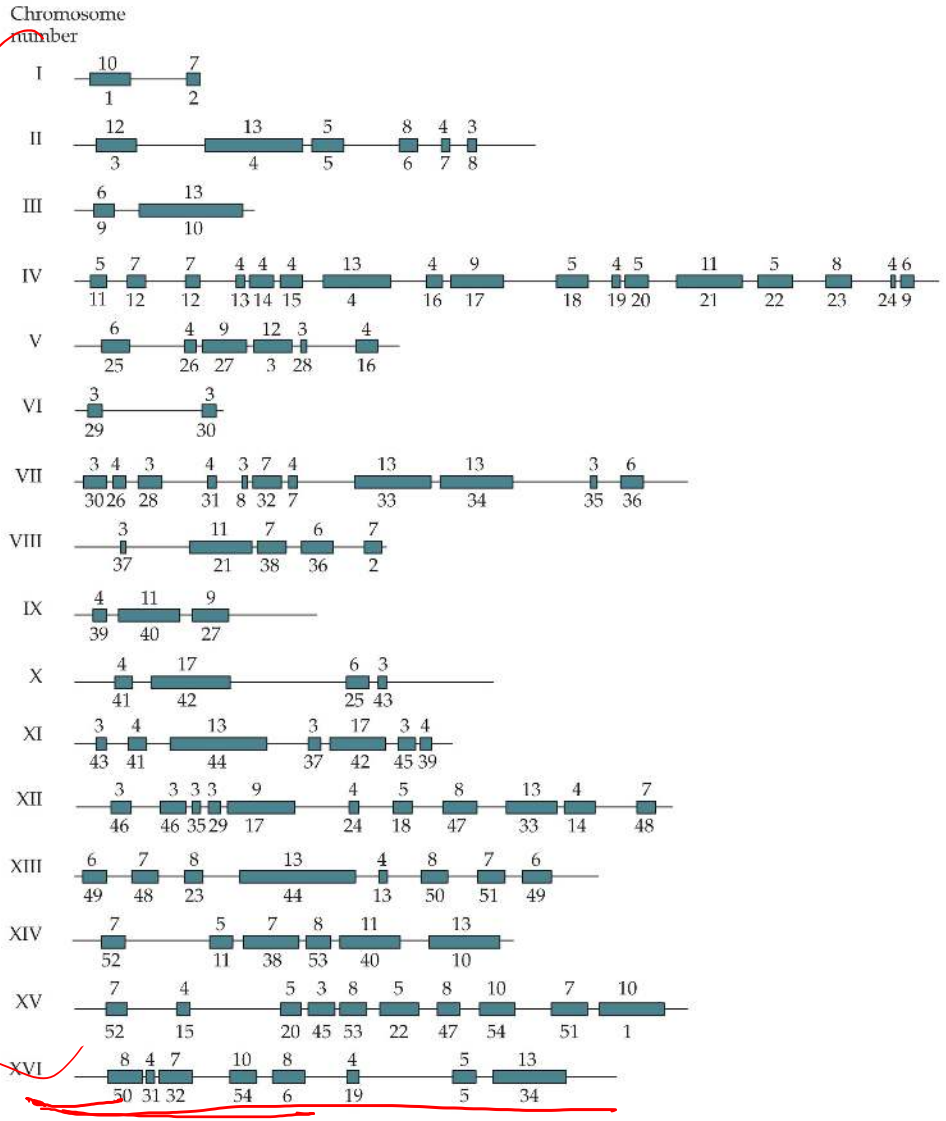




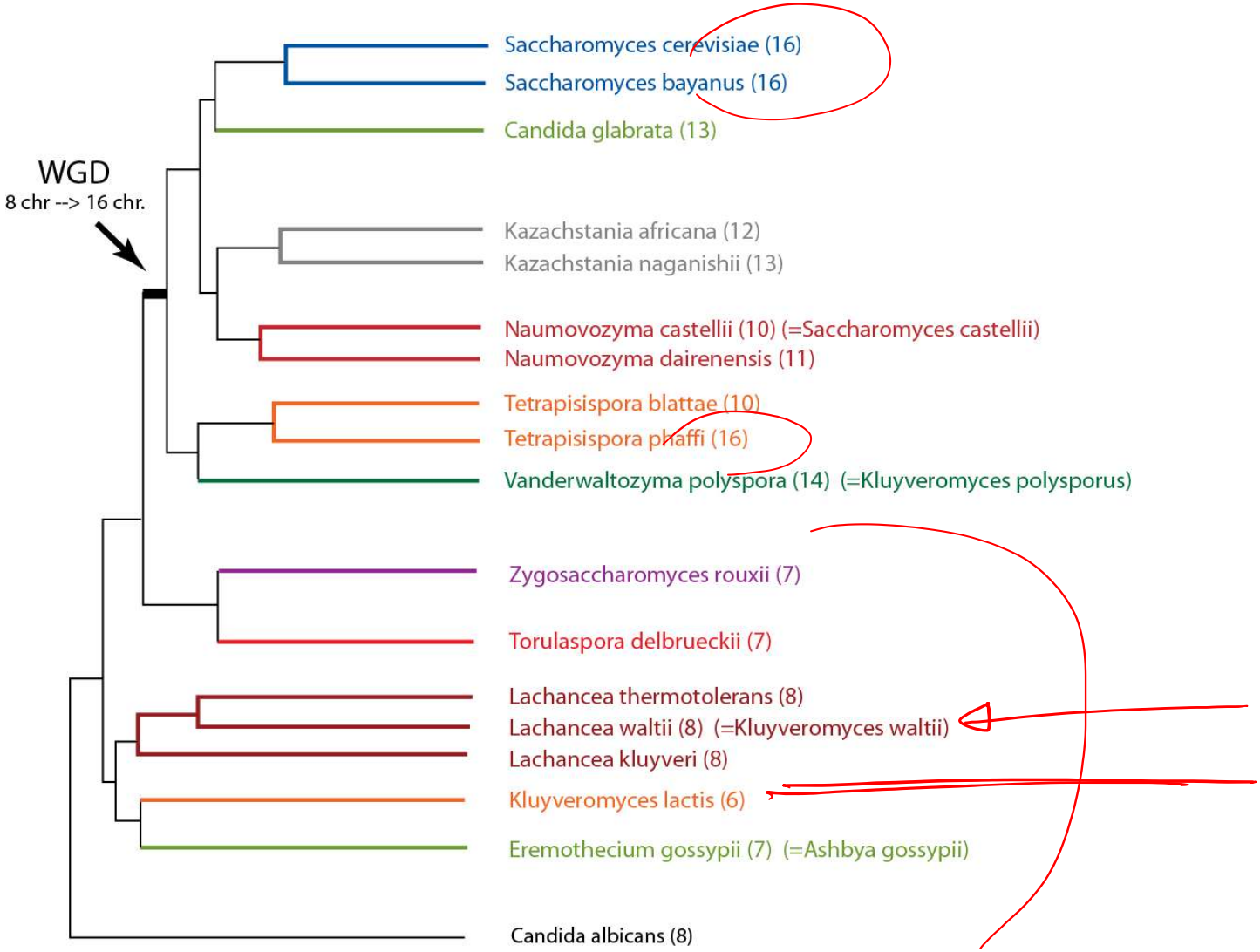
How can you distinguish WGD from segmental duplication?

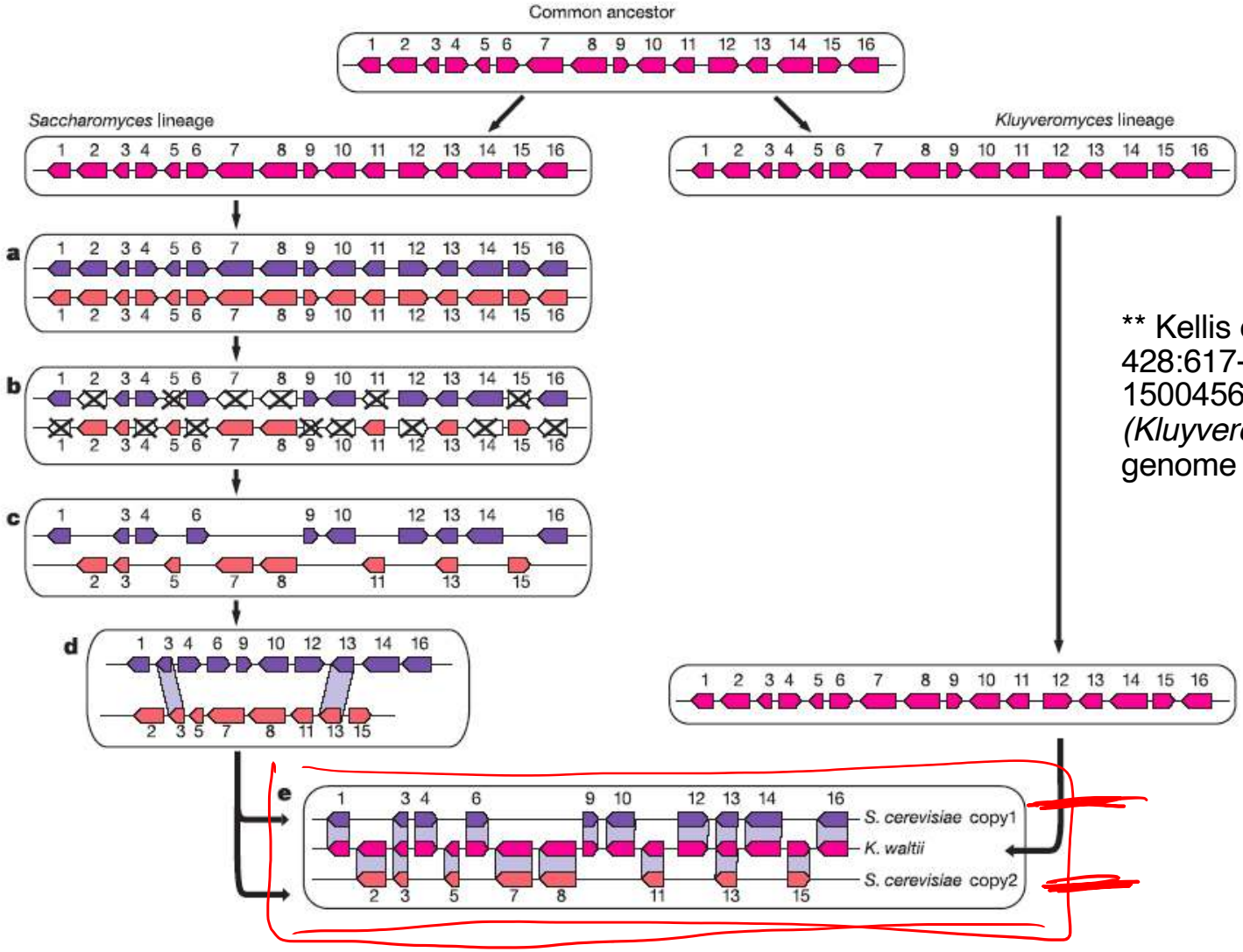


Non-overplapping blocks of doubly conserved synteny

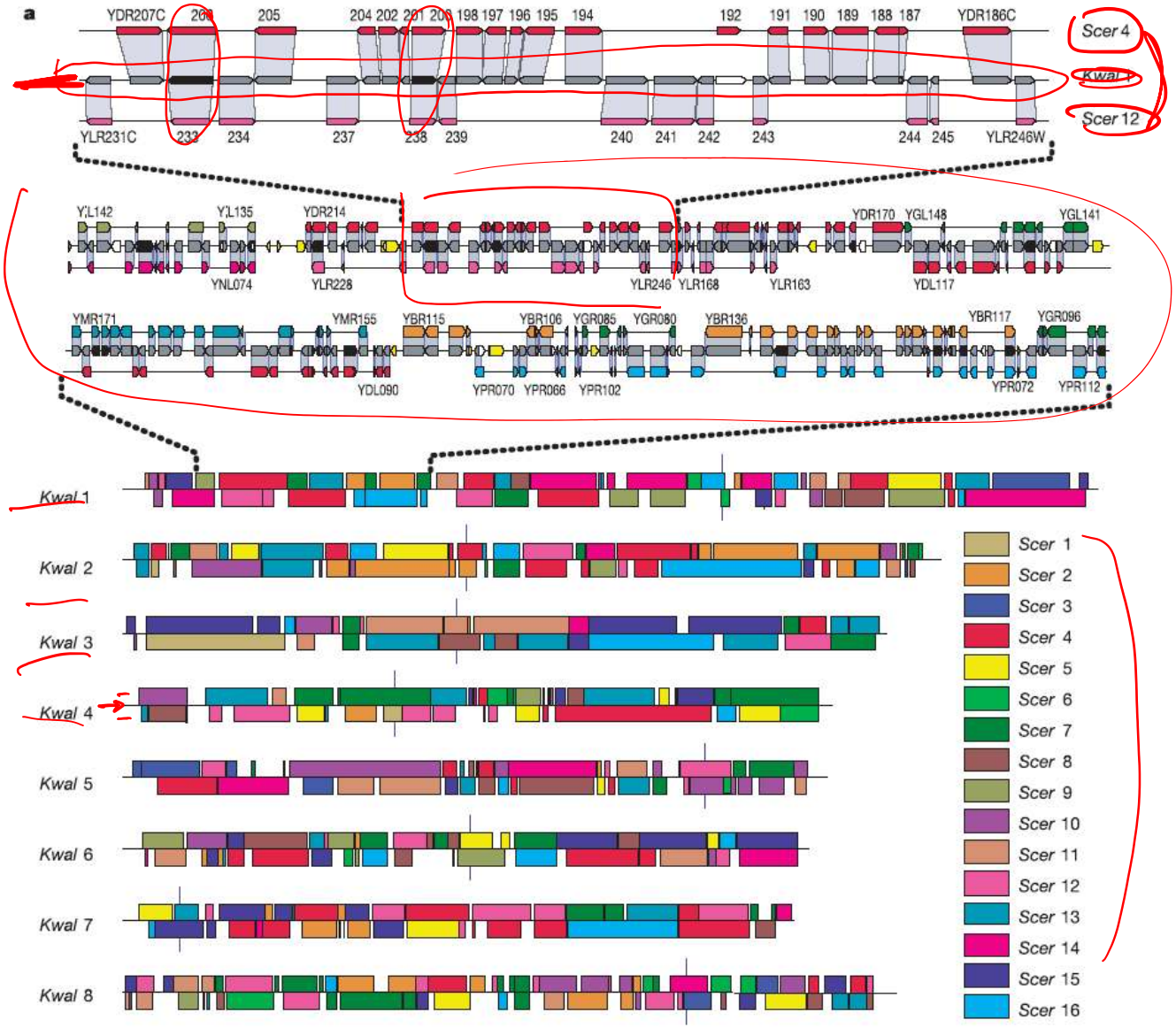


MOLECULAR AND GENOME EVOLUTION 1e, Figure 7.42


WGD
8 chr --> 16 chr.
Saccharomyces cerevisiae (16)
Saccharomyces bayanus (16)
Candida glabrata (13)
Kazachstania africana (12)
Kazachstania naganishii (13)
Naumovozyma castellii (10) (=Saccharomyces castellii)
Naumovozyma dairenensis (11)
Tetrapisispora blattae (10)
Tetrapisispora phaffi (16)
Vanderwaltozyma polyspora (14) (=Kluyveromyces polysporus)
Zygosaccharomyces rouxii (7)
Torulaspora delbrueckii (7)
Lachancea thermotolerans (8)
Lachancea waltii (8) (=Kluyveromyces waltii)
Lachancea kluyveri (8)
Kluyveromyces lactis (6)
Eremothecium gossypii (7) (=Ashbya gossypii)
Candida albicans (8)



** Kellis et al, Nature 428:617-624 (2004) PMID 15004568 – *Lachancea (Kluyveromyces) waltii* genome


a
YDR207C
205
204 202 201 198 197 196 195 194
192 191 190 189 188 187
YDR186C
Scer 4
Kwal 1
Scer 12
YLR231C
234
237
238 239
240 241 242 243
244 245
YLR246W
YDR214
YDR170
YGL148
YGL141
YNL074
YLR228
YLR246
YLR168
YLR163
YDL117
YMR171
YMR155
YBR115
YBR106
YGR085
YGR080
YBR136
YBR117
YGR096
YDL090
YPR070
YPR066
YPR102
YPR072
YPR112
Kwal 1
Kwal 2
Kwal 3
Kwal 4
Kwal 5
Kwal 6
Kwal 7
Kwal 8
Scer 1
Scer 2
Scer 3
Scer 4
Scer 5
Scer 6
Scer 7
Scer 8
Scer 9
Scer 10
Scer 11
Scer 12
Scer 13
Scer 14
Scer 15
Scer 16

WGD in *Saccharomyces cerevisiæ*

550 duplicated gene pairs out of 5600 genes in genome.

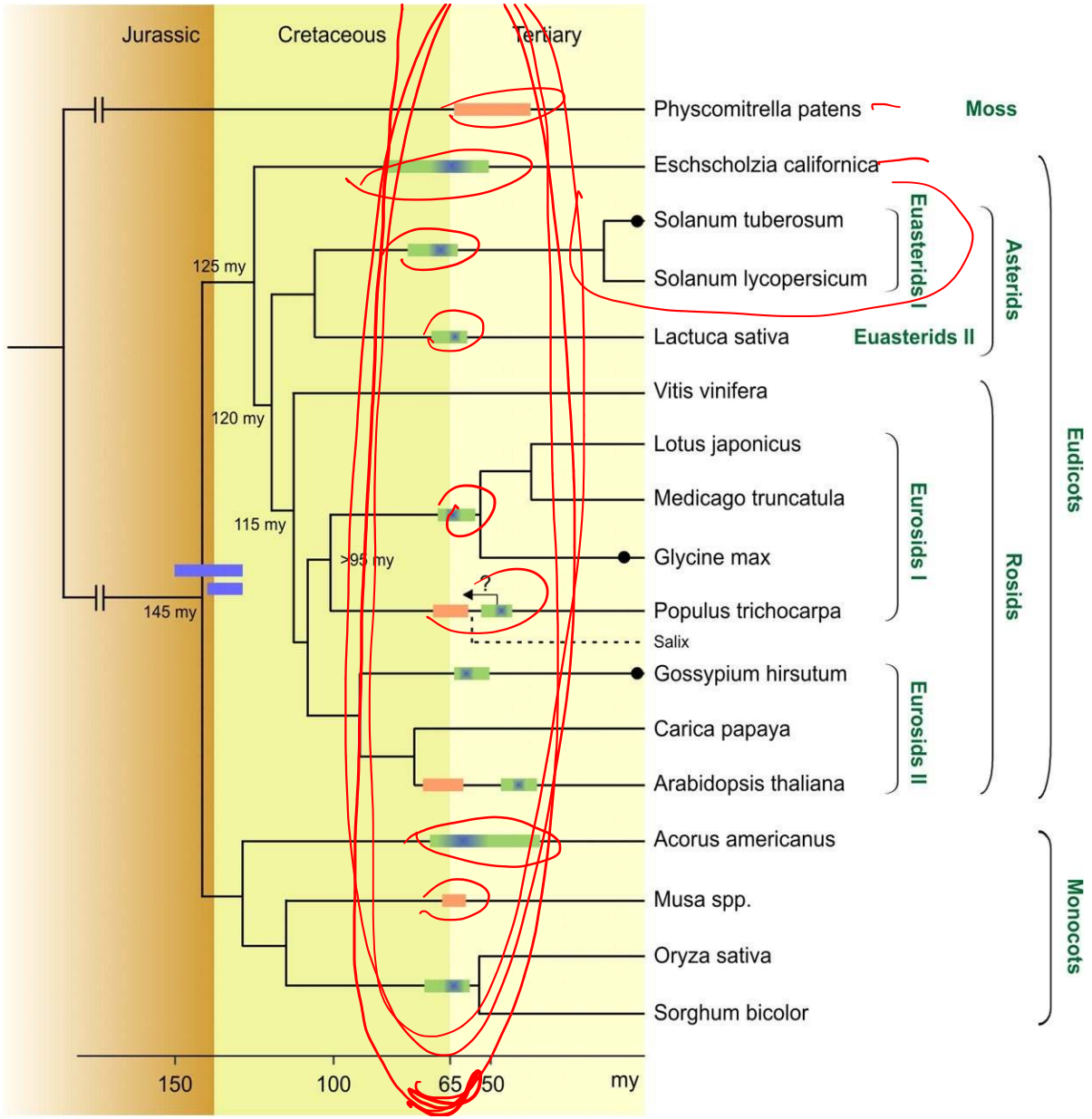
Average amino sequence identity is only 63%

Not diverged in function:

- About 80 pairs of ribosomal protein genes (e.g. *RPL2A*/*RPL2B*); each pair makes almost identical proteins.

Diverged in function:

- *ORC1* (origin recognition complex) / *SIR3* (silencing).
- *BUD8* / *BUD9*: landmarks for bud site choice for axial budding (diploids) vs. unipolar budding (haploids).
- *GAL1* (galactokinase) / *GAL3* (regulator).



WGD in plants

[Excerpt of fig legend] WGDs are indicated by green bars depicting the union of their 95% age confidence intervals calculated with various constraints (see Table S1). The dark green portions of the bars are centered on the best age estimates (see Table 1). Orange bars are WGD age estimates from literature.

Fawcett, J. A., Maere, S. & Peer, Y. V. de. *PNAS* **106**, 5737–5742 (2009).

I-ADHORE3.0

///

i-ADHoRe is a highly sensitive software tool to detect degenerated homology relations within and between different genomes.

This novel version of i-ADHoRe is designed to detect genomic homology in extremely large-scale data sets. Along with several under-the hood-improvements, resulting in a 30 fold reduction in runtime over previous versions, the implementation of multithreading and MPI now enables i-ADHoRe to take advantage of a parallel computing platform. As the scale of the data sets increased, the need for a new alignment algorithm able to cope with dozens of genomic segments became apparent. Therefore a new greedy graph

Open access, freely available online PLOS BIOLOGY

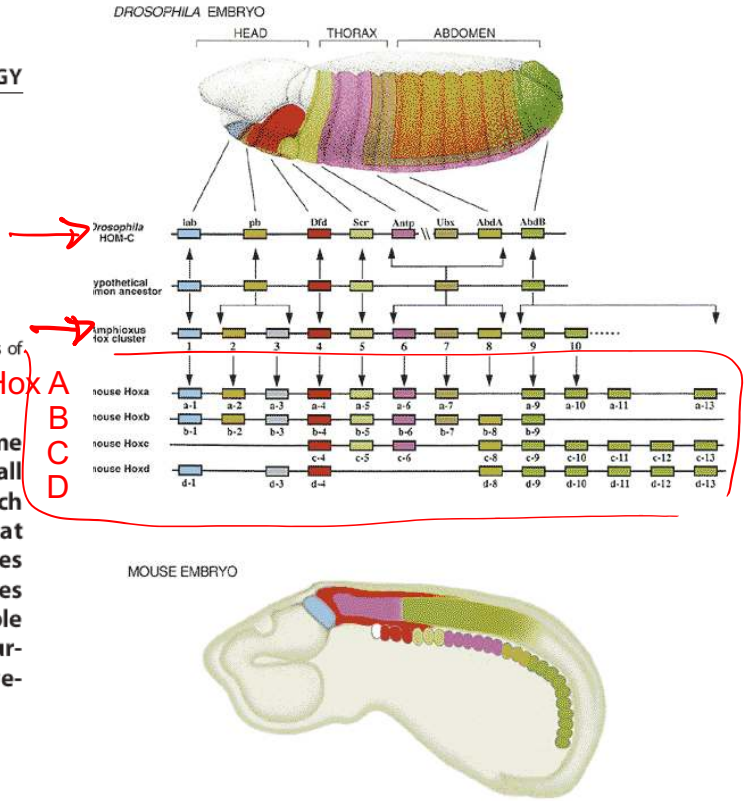
Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate

Paramvir Dehal[1], Jeffrey L. Boore[1,2]

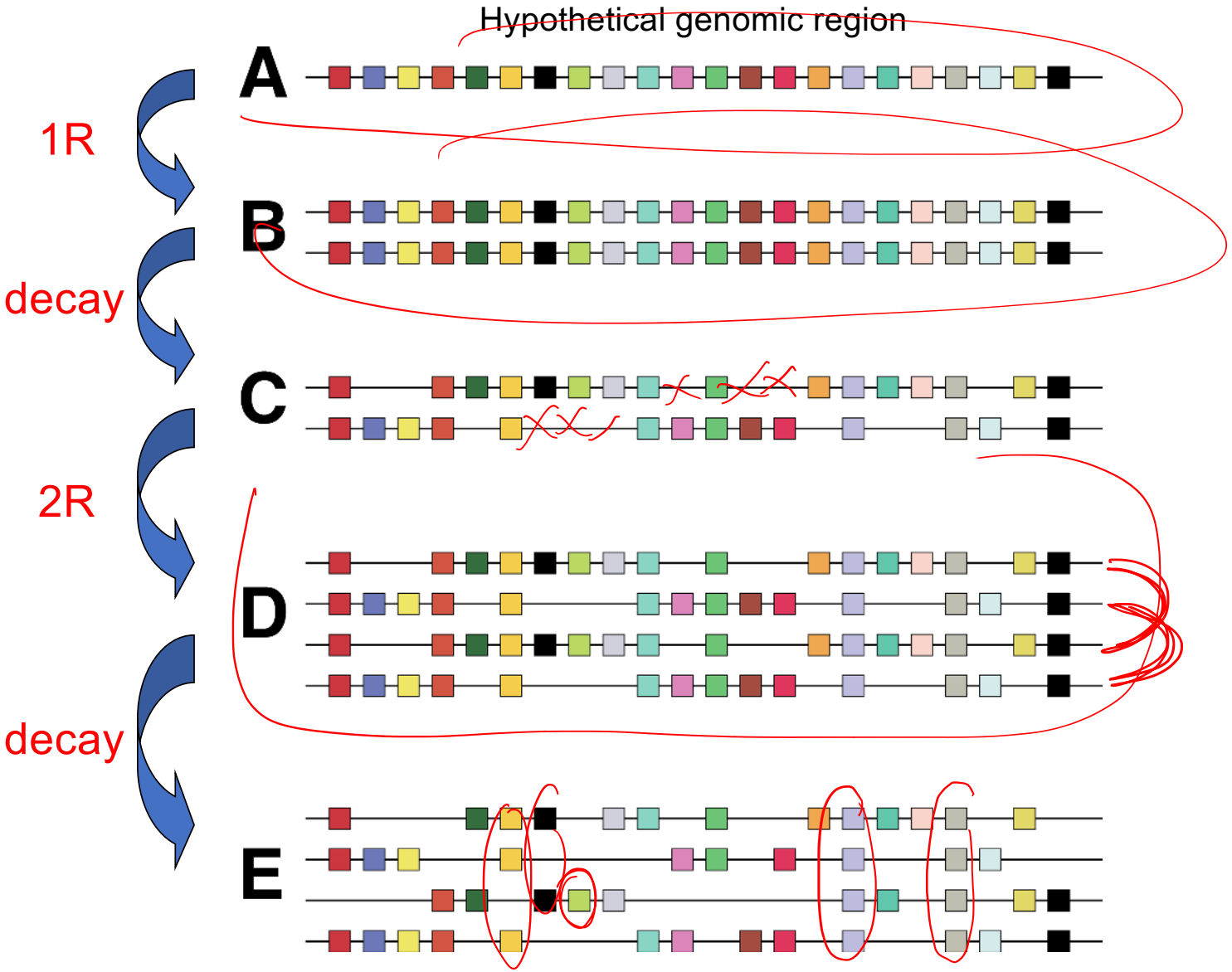
1 Evolutionary Genomics Department, Department of Energy Joint Genome Institute and Lawrence Berkeley National Laboratory, Walnut Creek, California, United States of America, **2** Department of Integrative Biology, University of California, Berkeley, California, United States of America

The hypothesis that the relatively large and complex vertebrate genome was created by two ancient, whole genome duplications has been hotly debated, but remains unresolved. We reconstructed the evolutionary relationships of all gene families from the complete gene sets of a tunicate, fish, mouse, and human, and then determined when each gene duplicated relative to the evolutionary tree of the organisms. We confirmed the results of earlier studies that there remains little signal of these events in numbers of duplicated genes, gene tree topology, or the number of genes per multigene family. However, when we plotted the genomic map positions of only the subset of paralogous genes that were duplicated prior to the fish–tetrapod split, their global physical organization provides unmistakable evidence of two distinct genome duplication events early in vertebrate evolution indicated by clear patterns of four-way paralogous regions covering a large part of the human genome. Our results highlight the potential for these large-scale genomic events to have driven the evolutionary success of the vertebrate lineage.

Citation: Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3(10): e314.

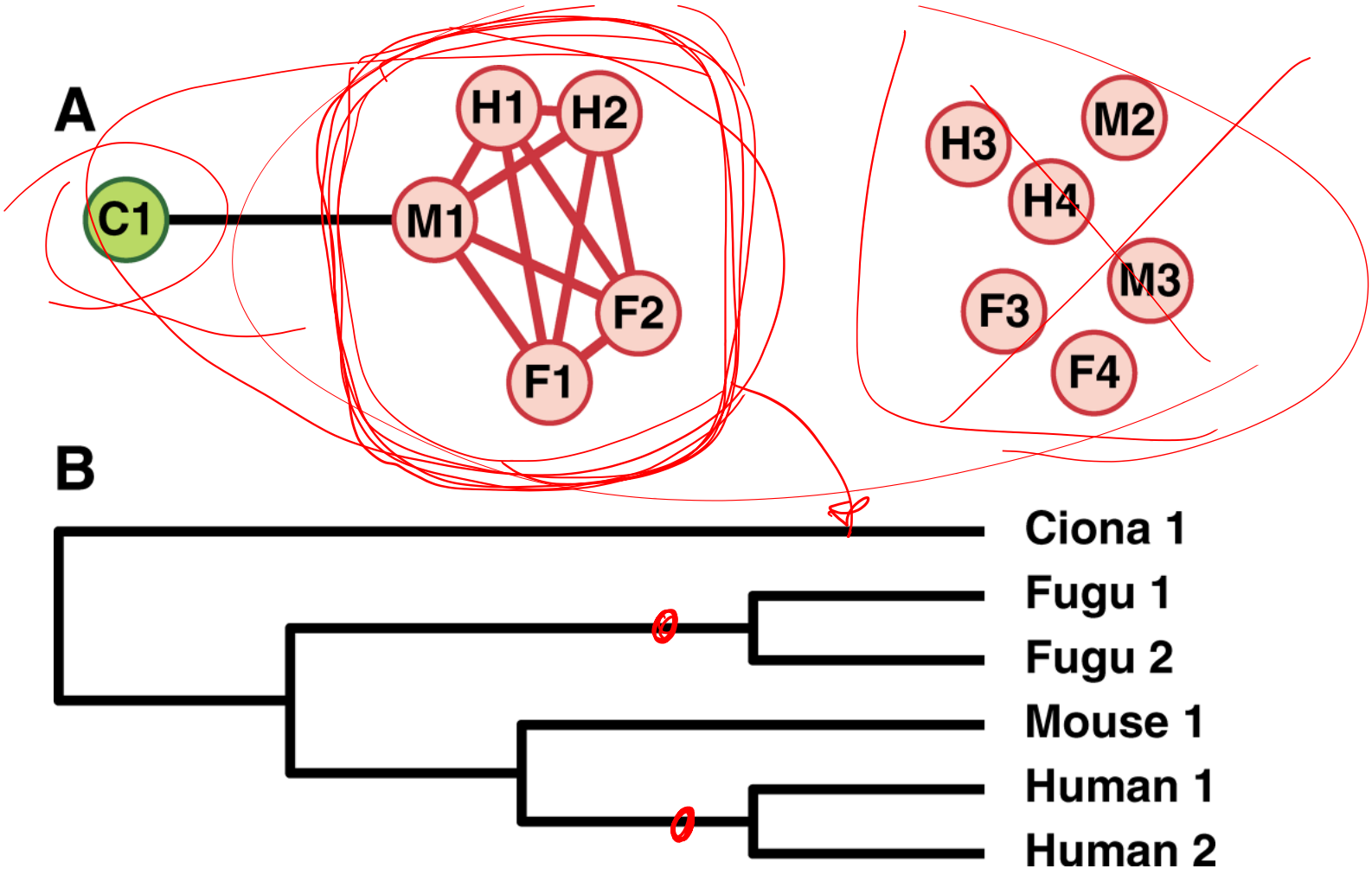


Dehal P, Boore JL. PLoS Biol. 3, e314 (2005). PMID 16128622



Dehal & Boore (2005)

Dehal & Boore (2005)



An example of a gene family with one duplication in *Fugu* and another in human. *Ciona intestinalis* (sea squirt) is the outgroup.

By perezoso - Self-photographed, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1771320

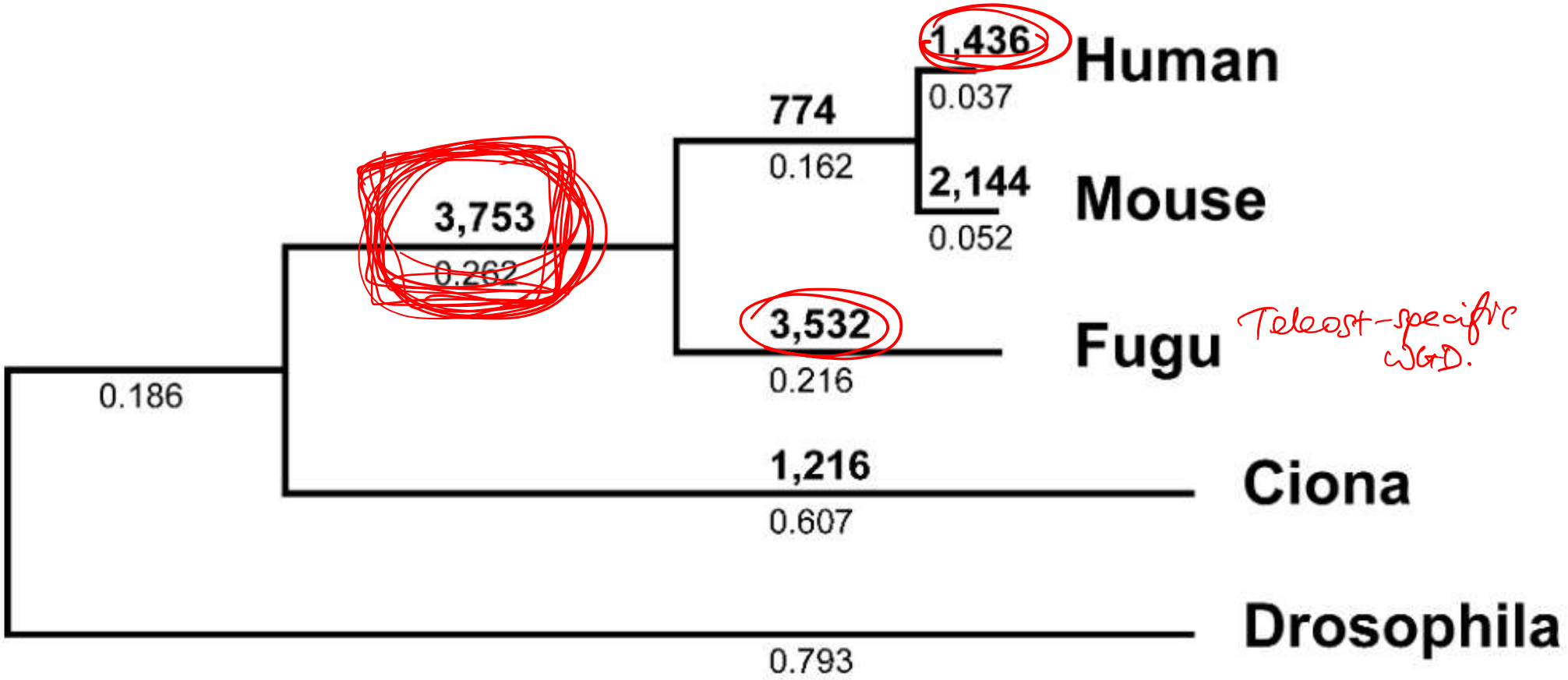
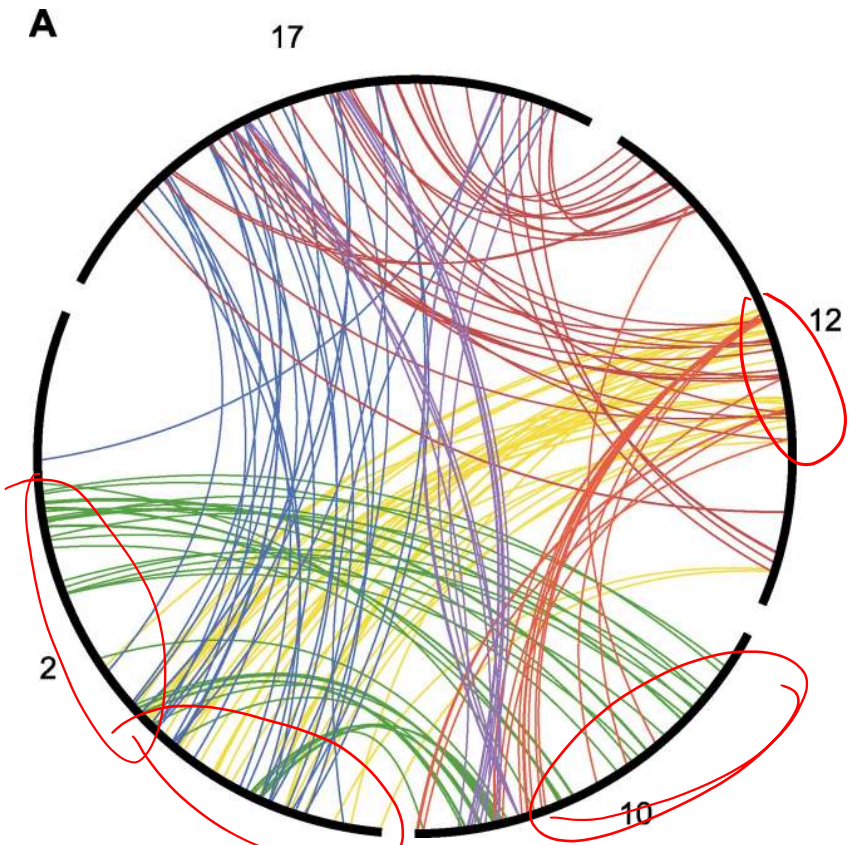
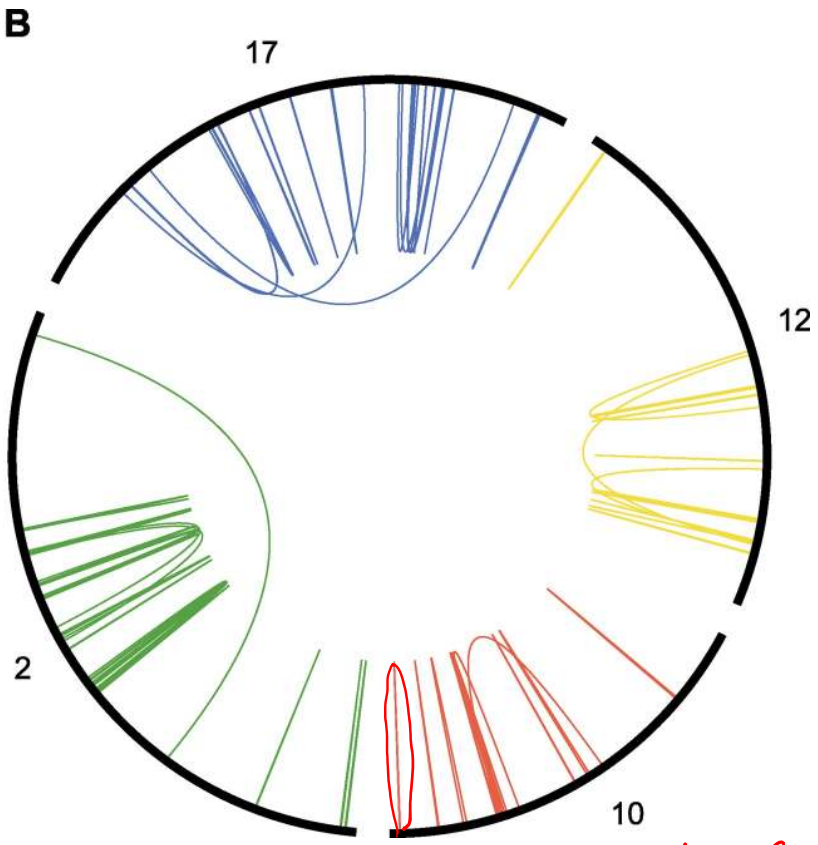


Figure 4. Phylogenetic Analysis of the Four Chordates with *Drosophila* as an Outgroup



Old duplications

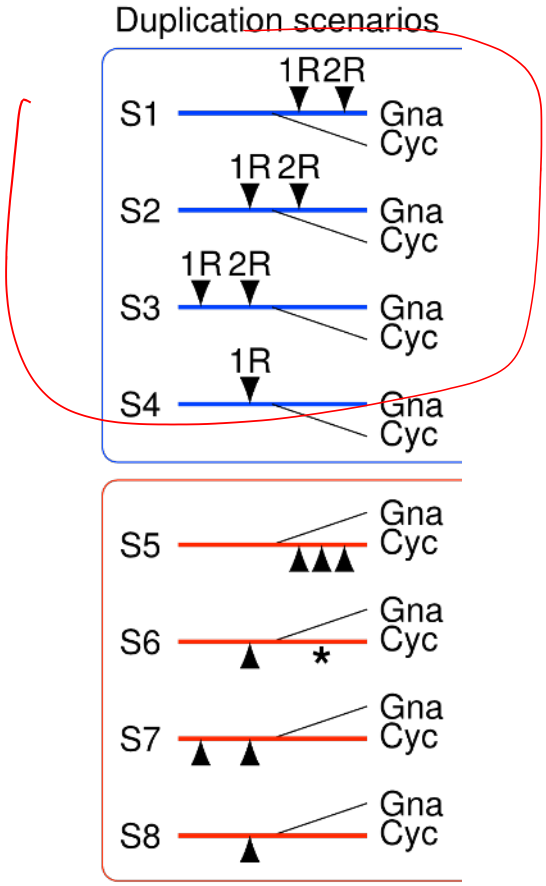


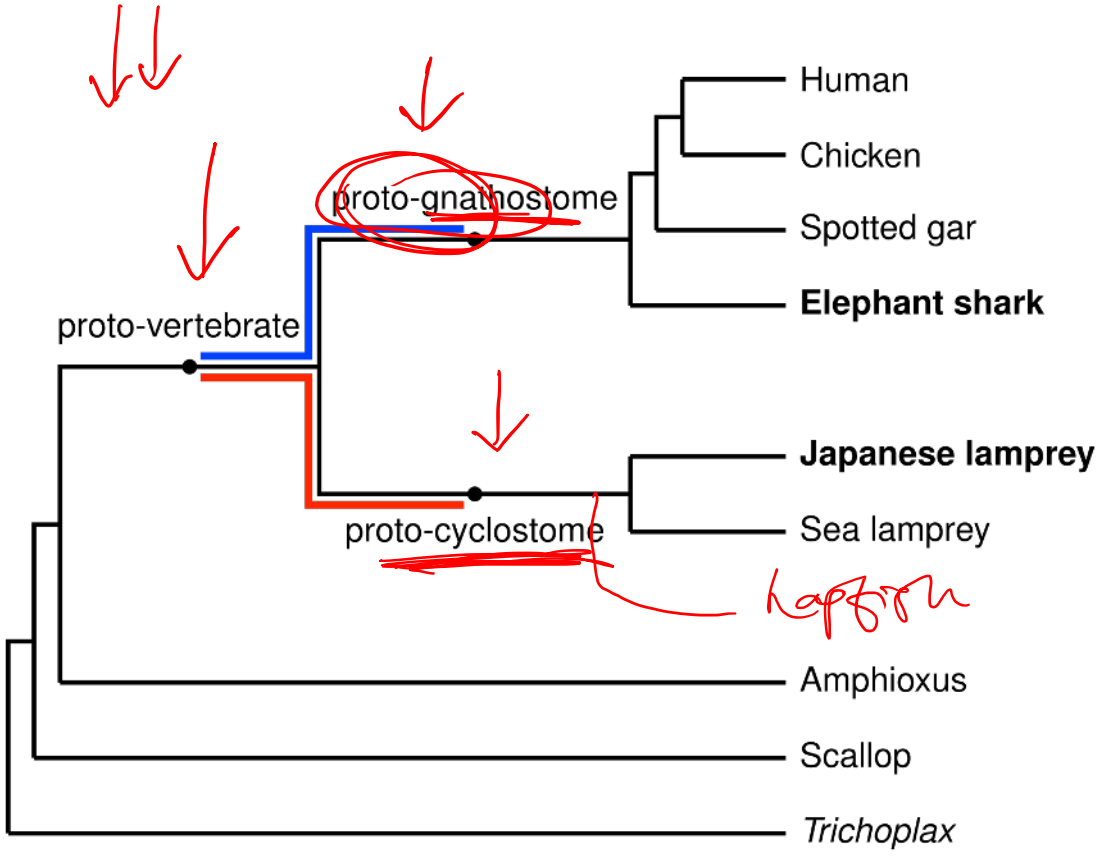
Recent duplications

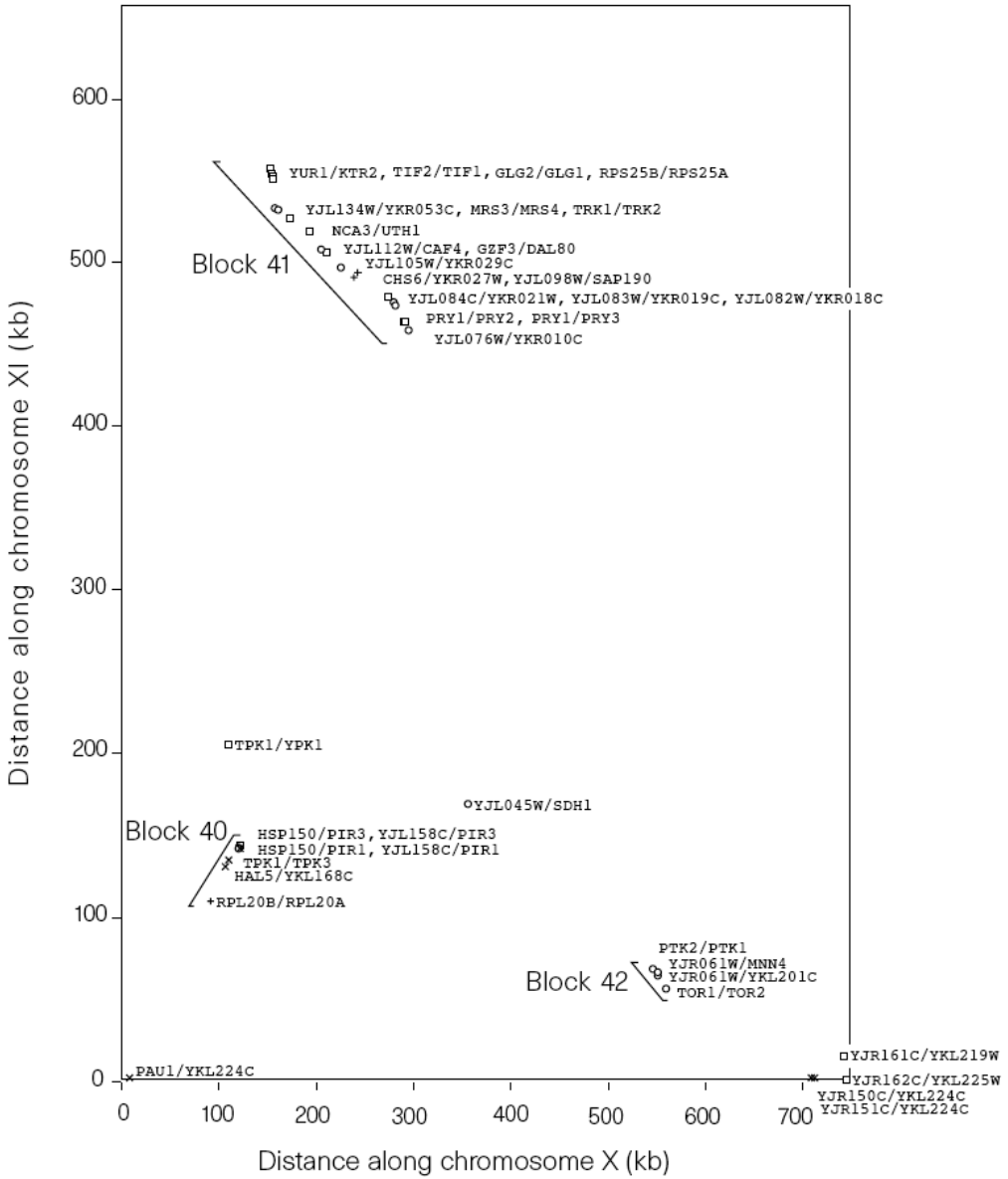
Tandem dup

Dehal & Boore (2005)

Open questions regarding number and timing of vertebrate WGDs



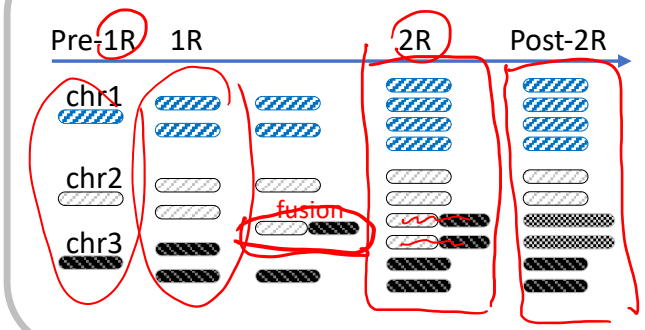




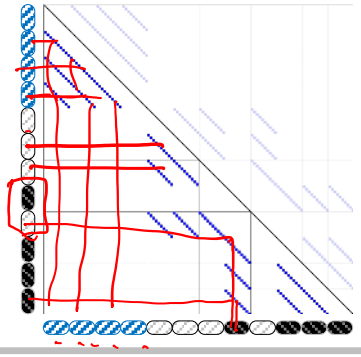
Dotplots are useful for discovery of synteny blocks within or between genomes

But gets messy with time….

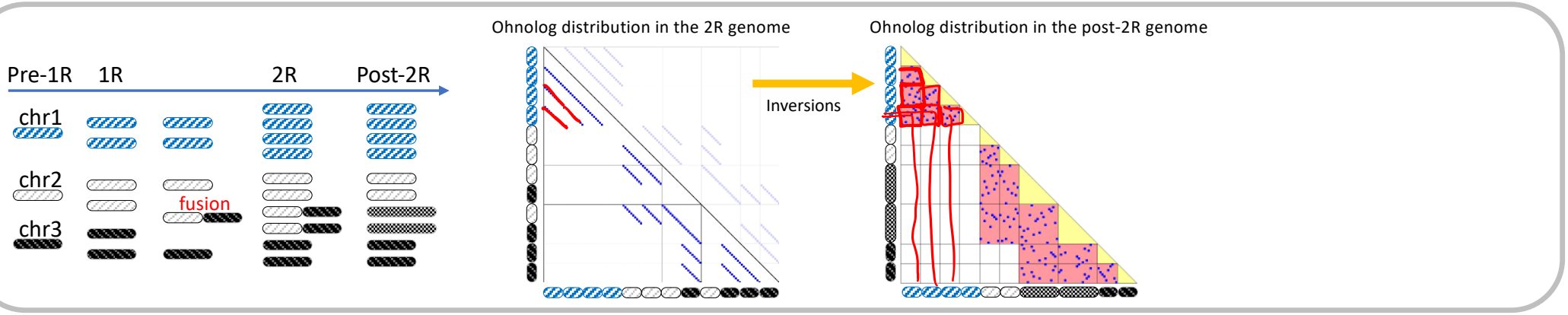
Model of vertebrate genome evolution and reconstruction of the post-2R chromosomes



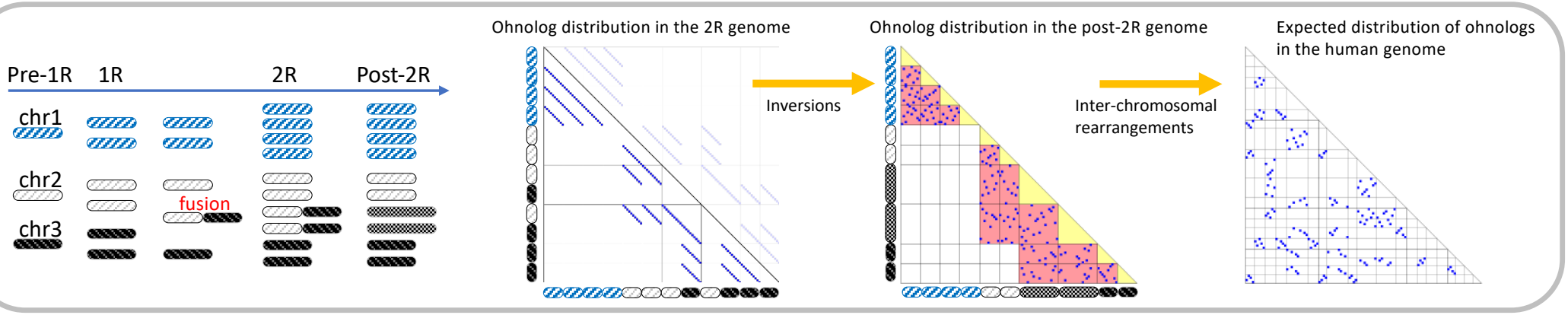
Ohnolog distribution in the 2R genome



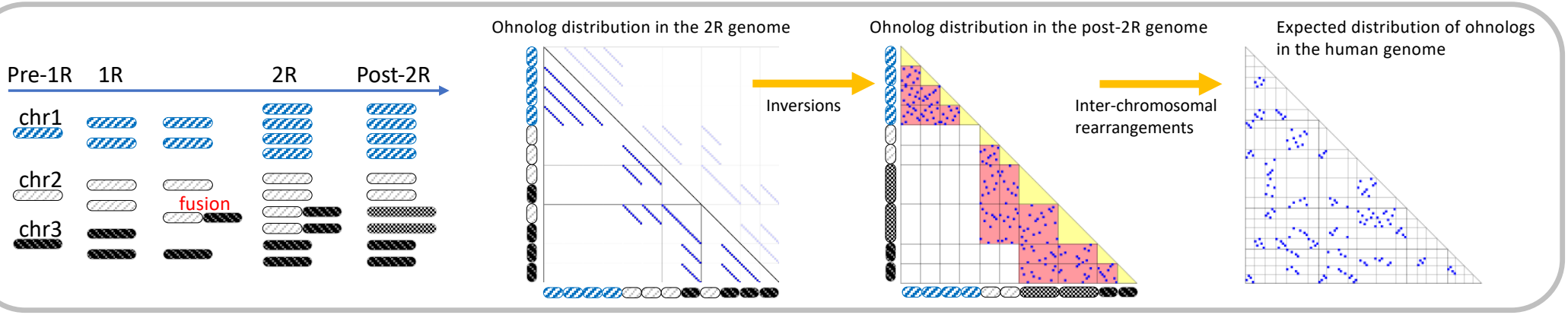
Model of vertebrate genome evolution and reconstruction of the post-2R chromosomes

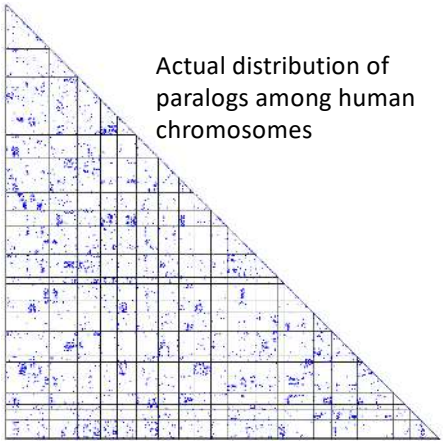


Model of vertebrate genome evolution and reconstruction of the post-2R chromosomes

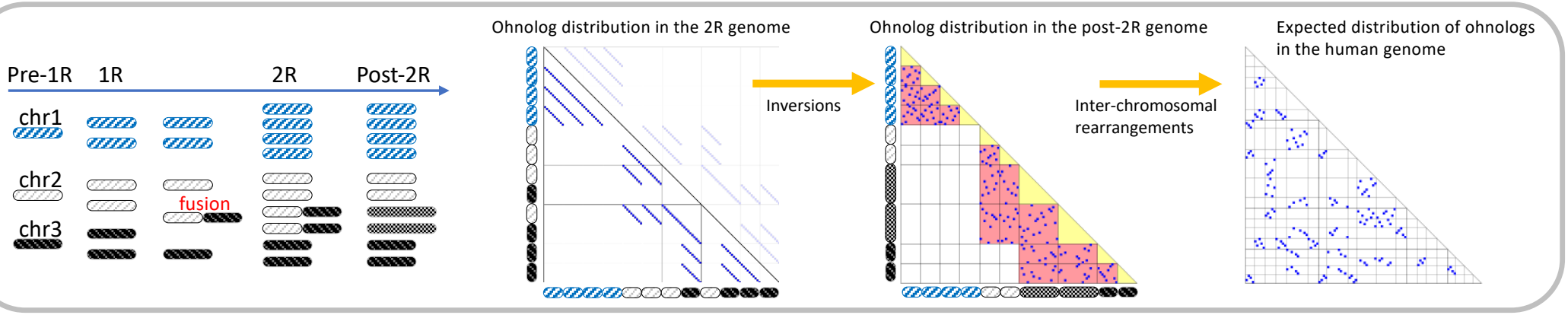


Model of vertebrate genome evolution and reconstruction of the post-2R chromosomes

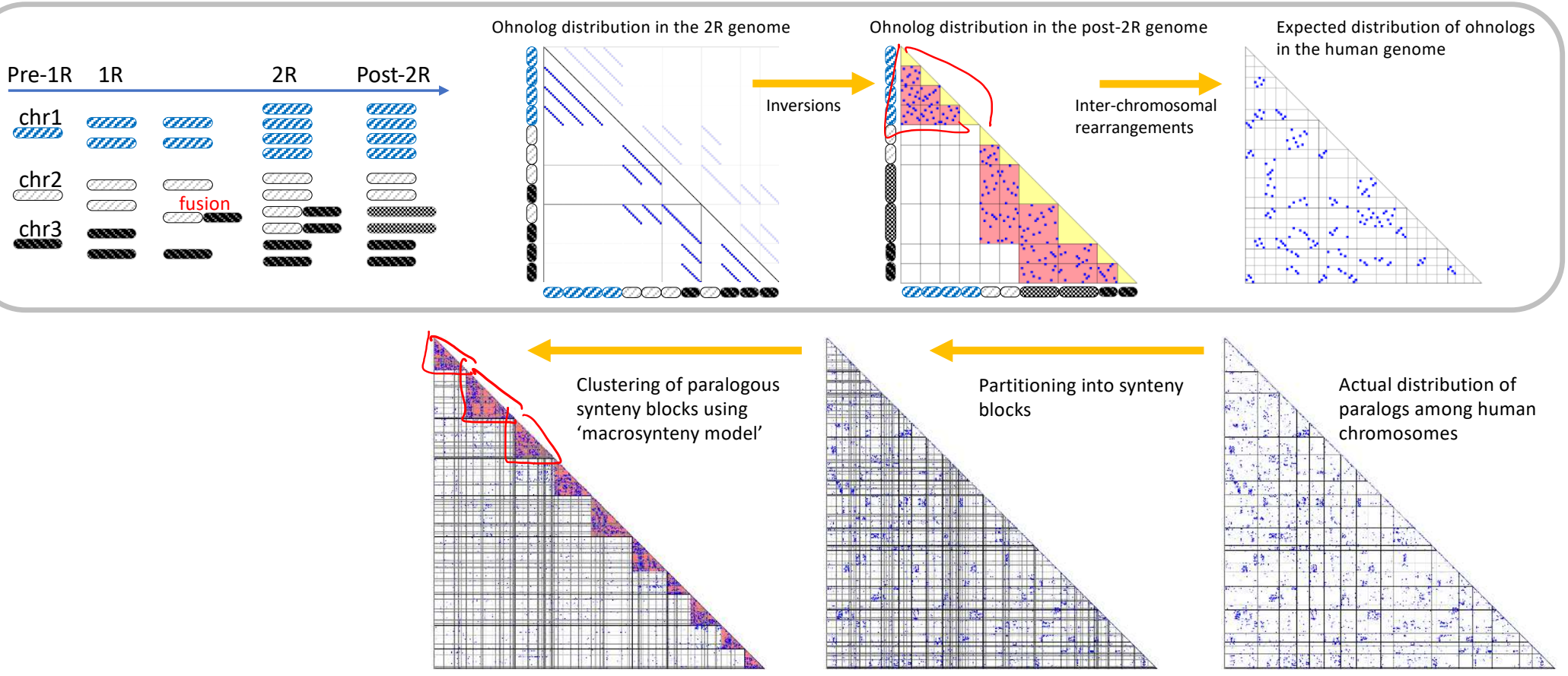




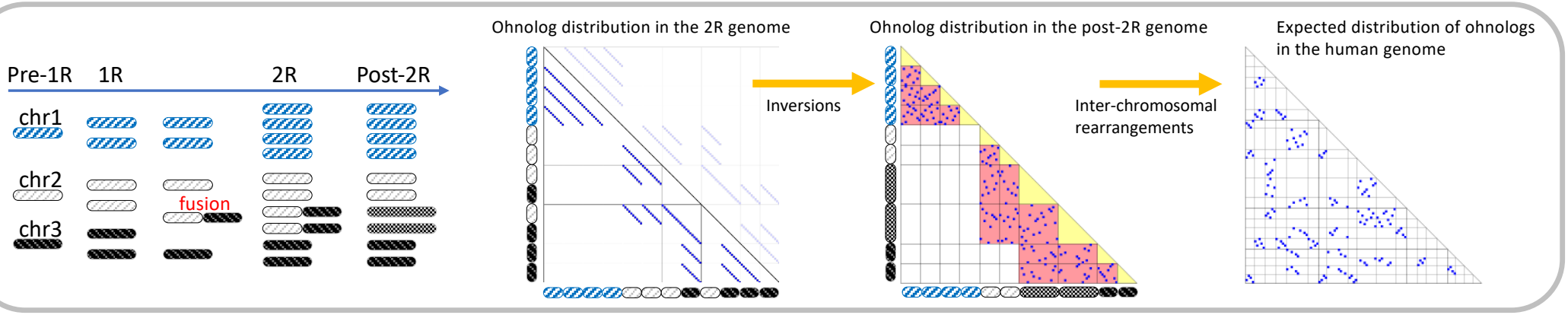
Model of vertebrate genome evolution and reconstruction of the post-2R chromosomes

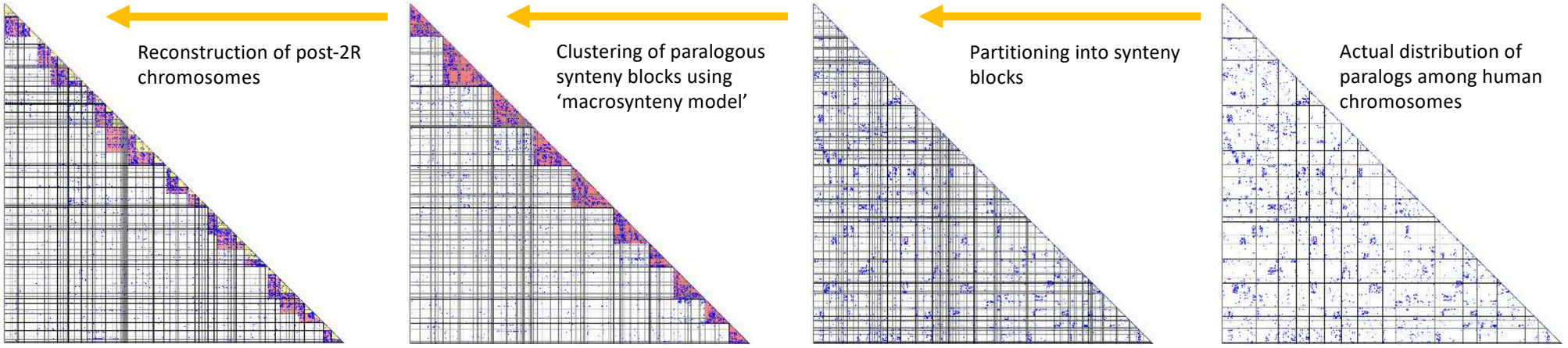


Model of vertebrate genome evolution and reconstruction of the post-2R chromosomes

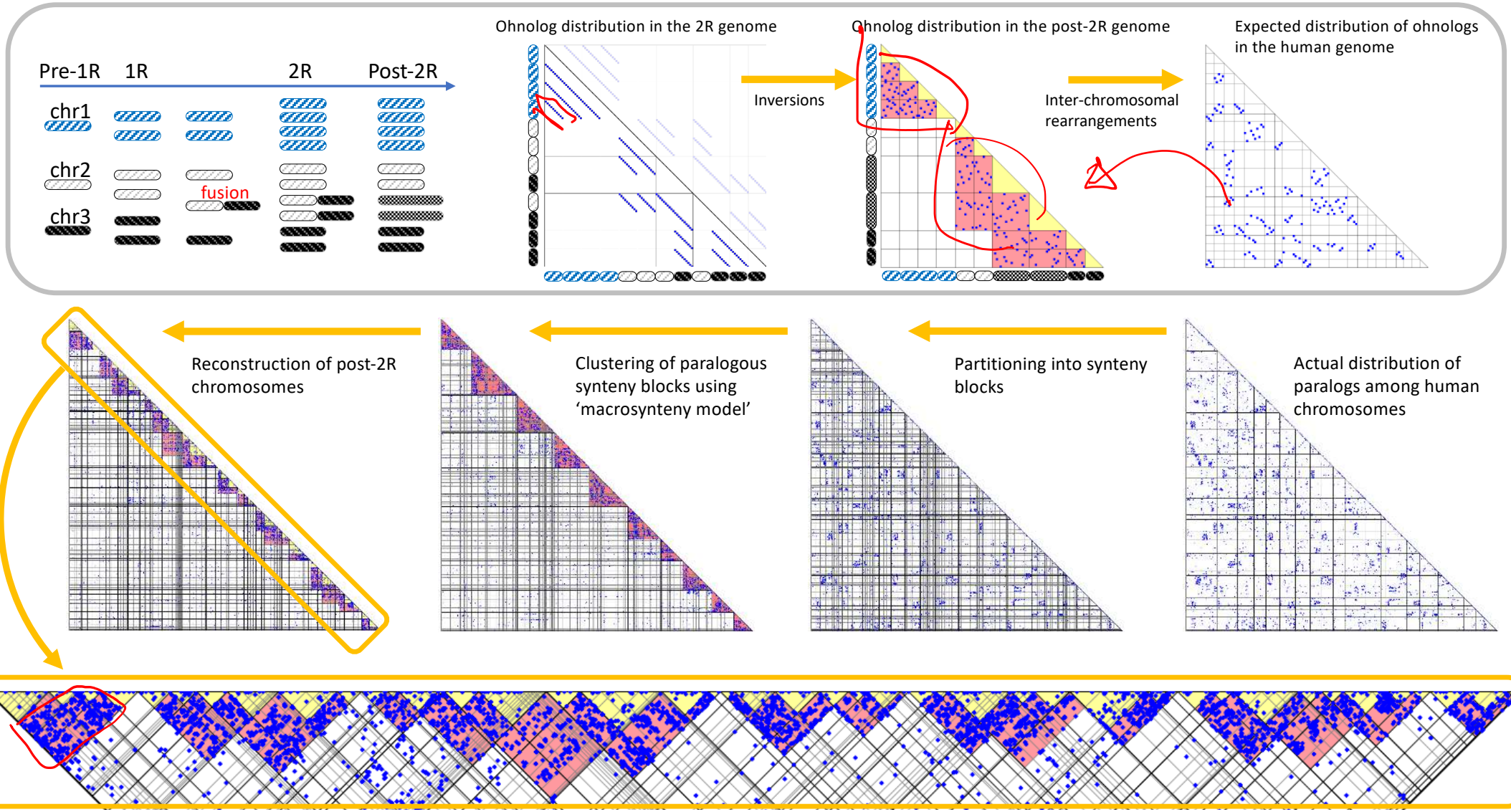


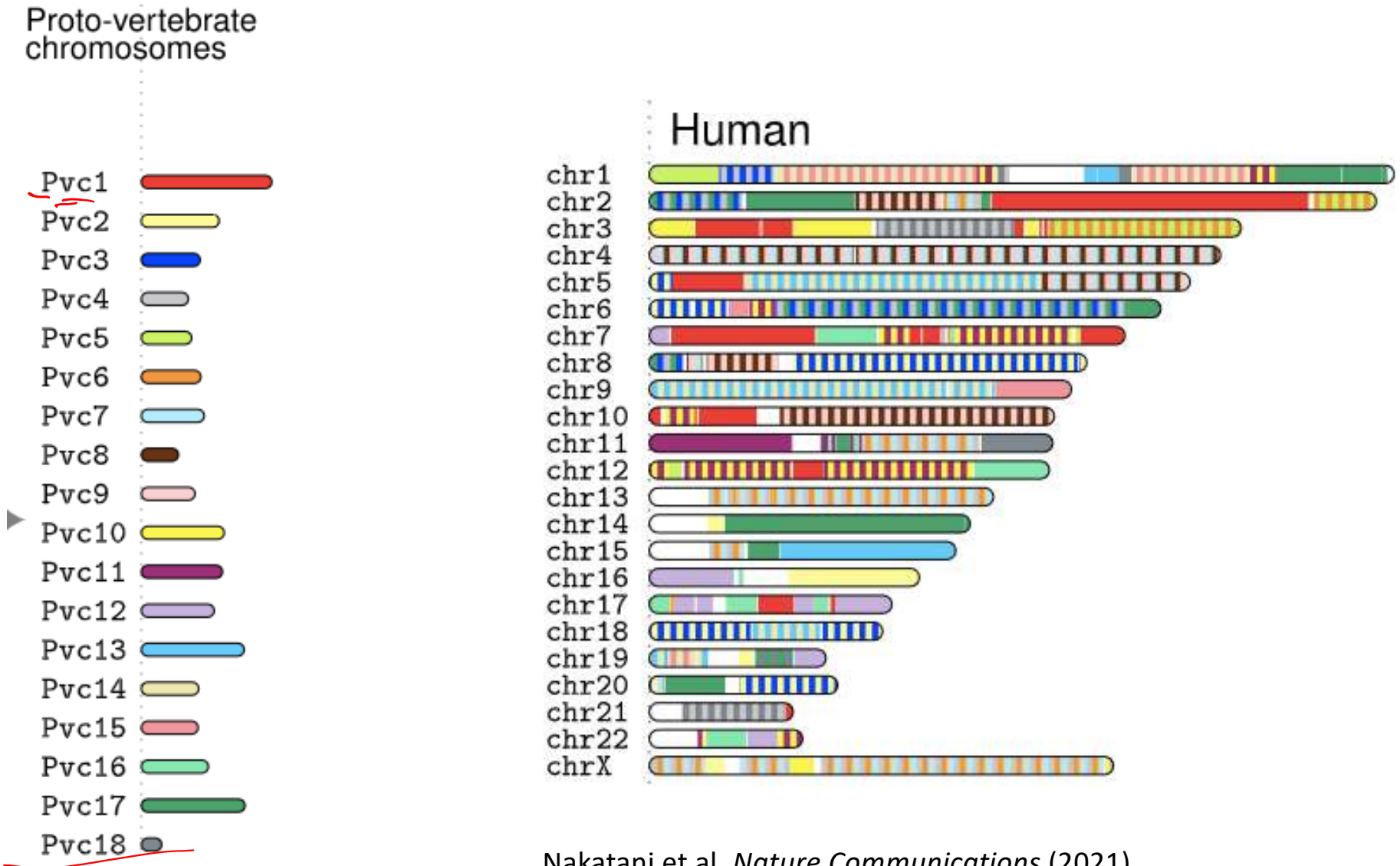
Model of vertebrate genome evolution and reconstruction of the post-2R chromosomes





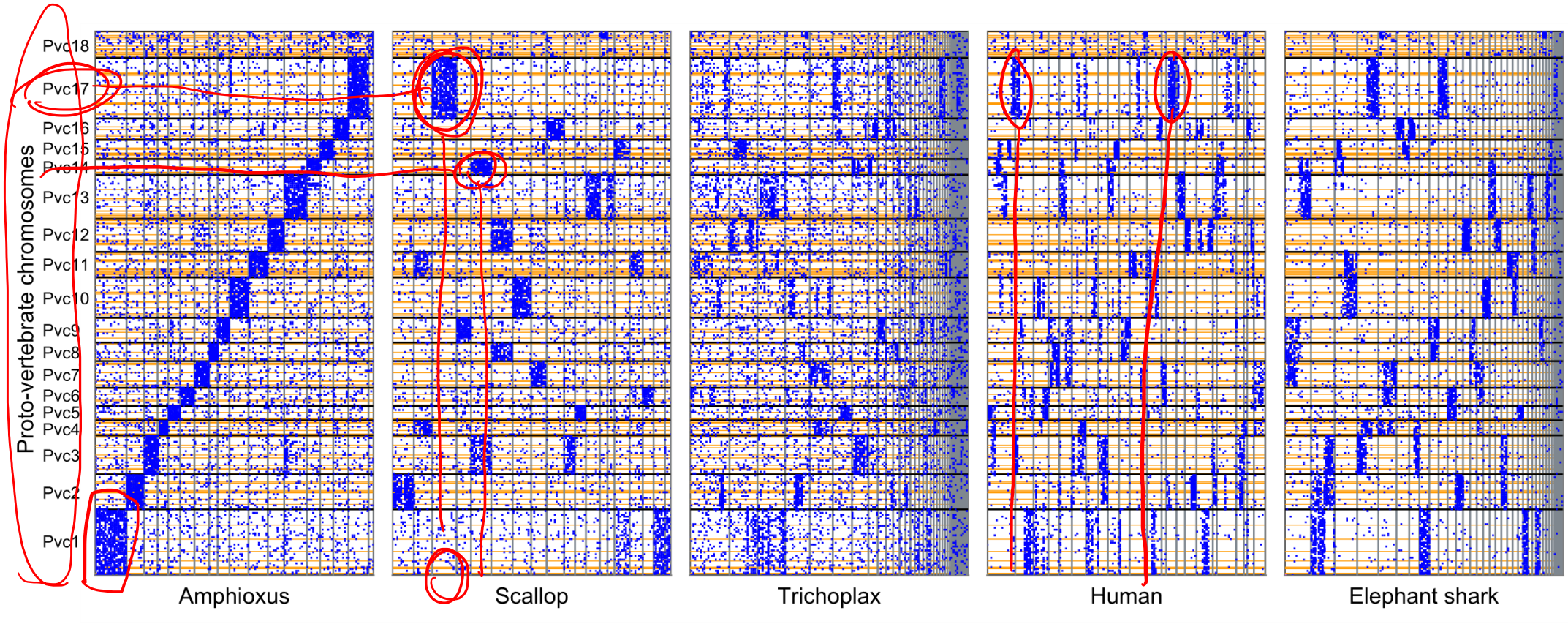
Model of vertebrate genome evolution and reconstruction of the post-2R chromosomes





Nakatani et al, *Nature Communications* (2021)

Inferred ancestral chromosomes have distinct ortholog distributions across other animal genomes

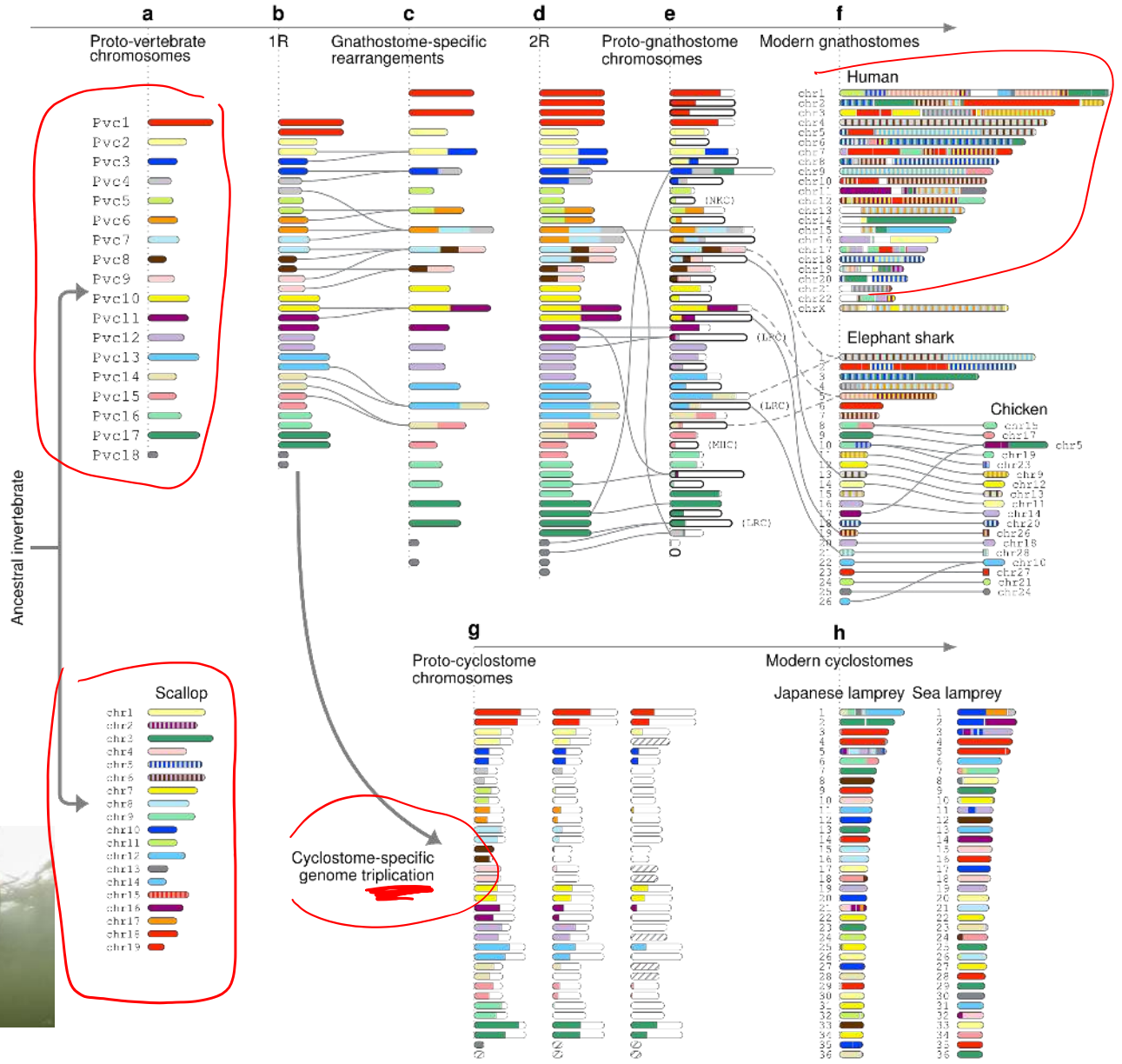


Reconstruction:

1. Ancestral vertebrate genome of 18 chromosomes

2. First WGD, but not second is shared by gnathostomes and cyclostomes (evidence from chromosome fusions events)

3. Cyclostomes had an independent hexaploidy



Gene loss post-WGD

- Paralogs and orthologs can be more challenging to distinguish
- Sometimes a locus can be single-copy in all genomes, but be paralogs not orthologs

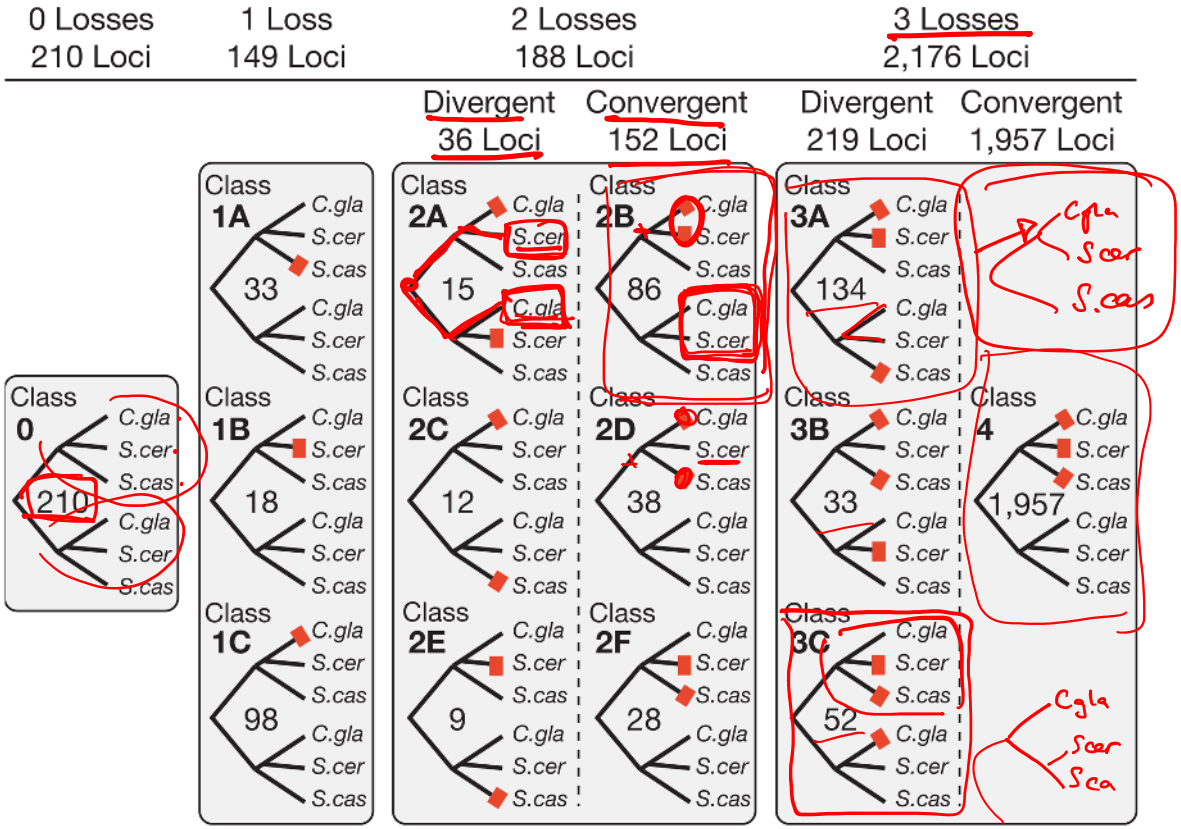
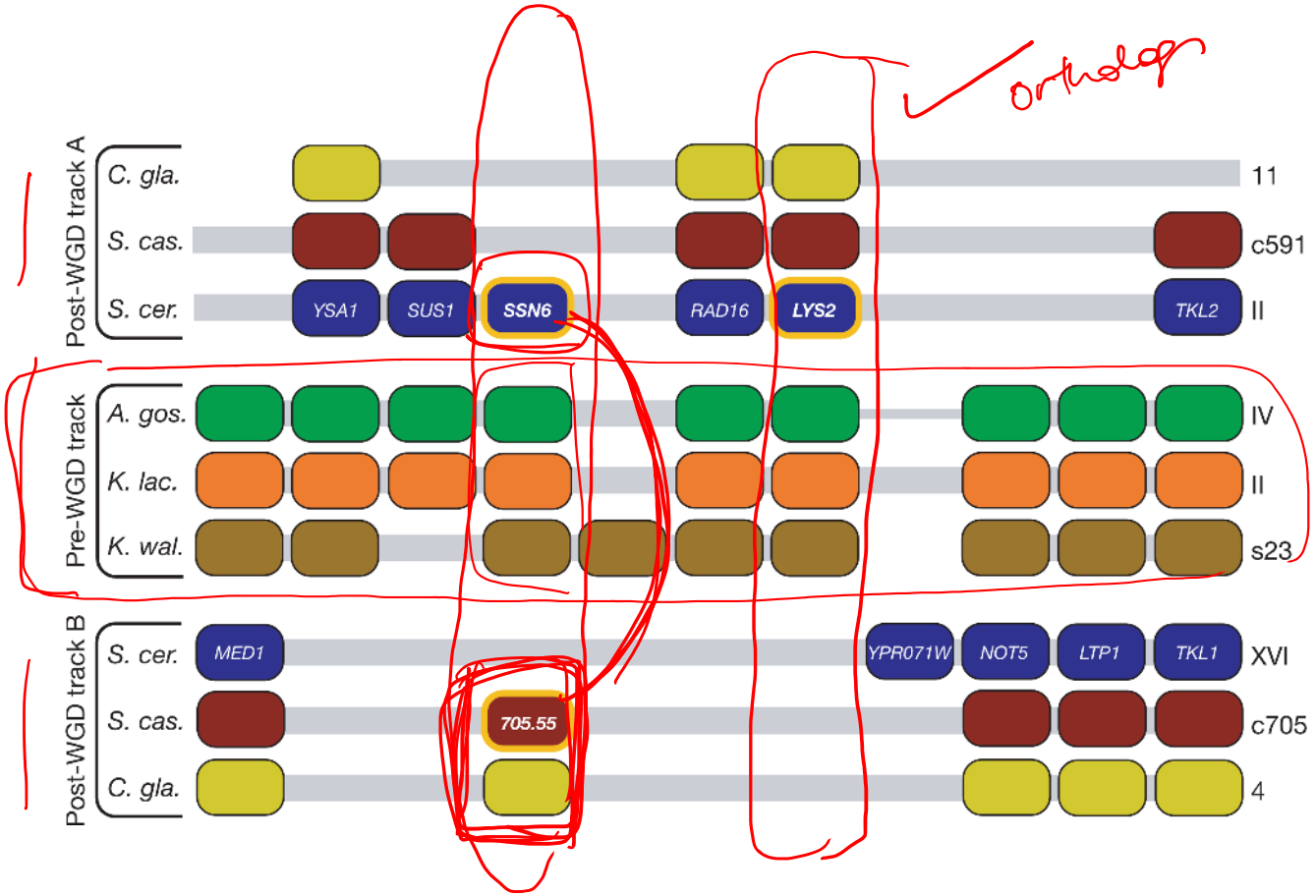


Figure 2 | Classes of gene loss pattern among 2,723 ancestral loci in *S. cerevisiae*, *S. castellii* and *C. glabrata*, and their frequencies. Red marks denote gene absence and are used to group ancestral loci into 14 gene-loss

Scannell *et al* Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**, 341–345 (2006).

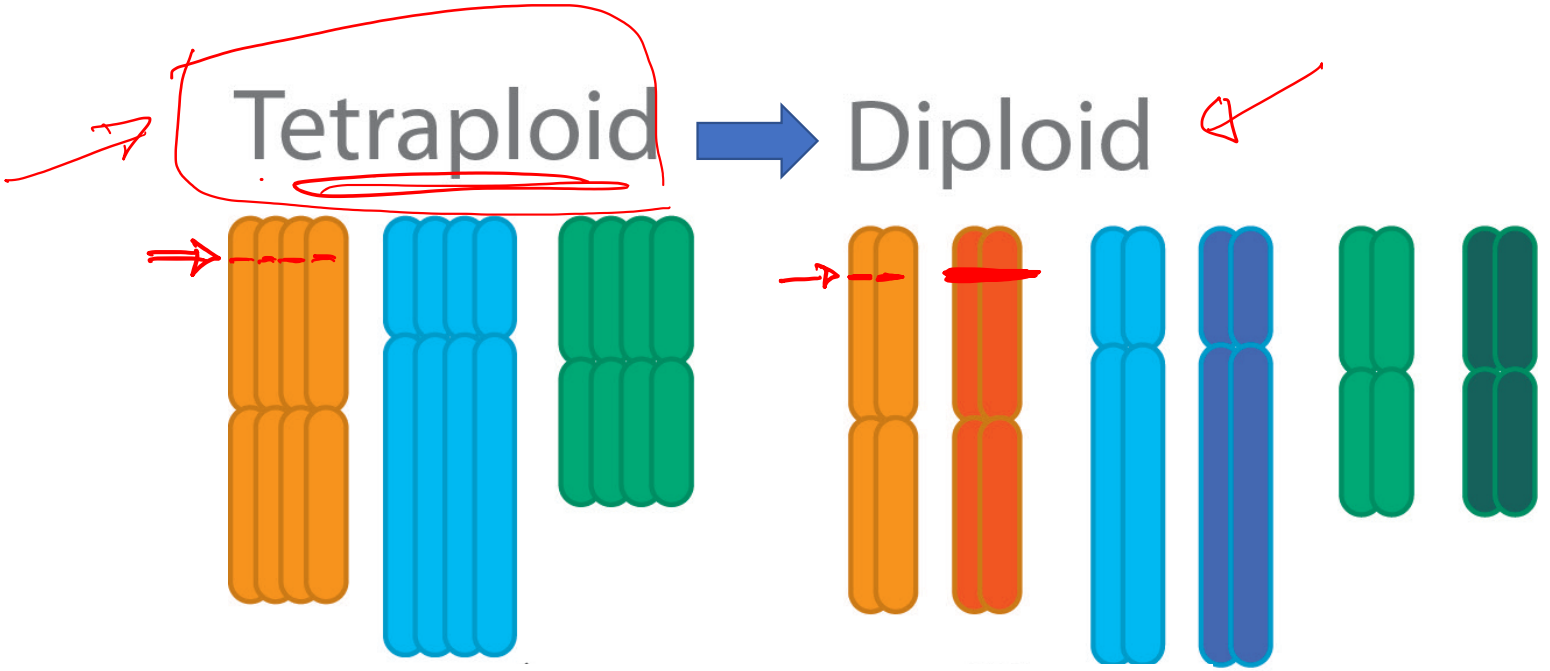
In some cases orthologs and paralogs can most efficiently be distinguished through synteny relationships



YGOB yeast genome browser

paleotetraploid

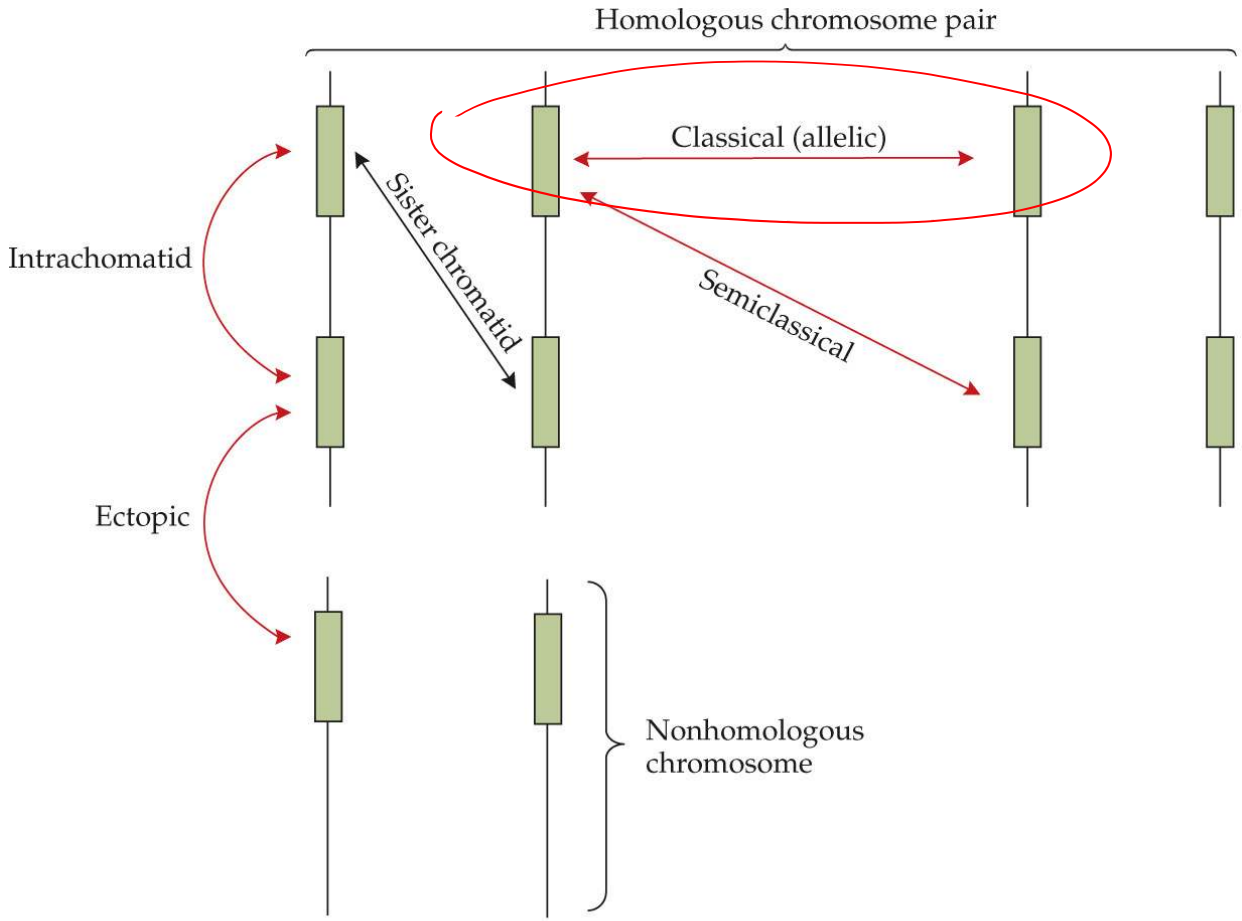
Return to diploid inheritance - rediploidisation



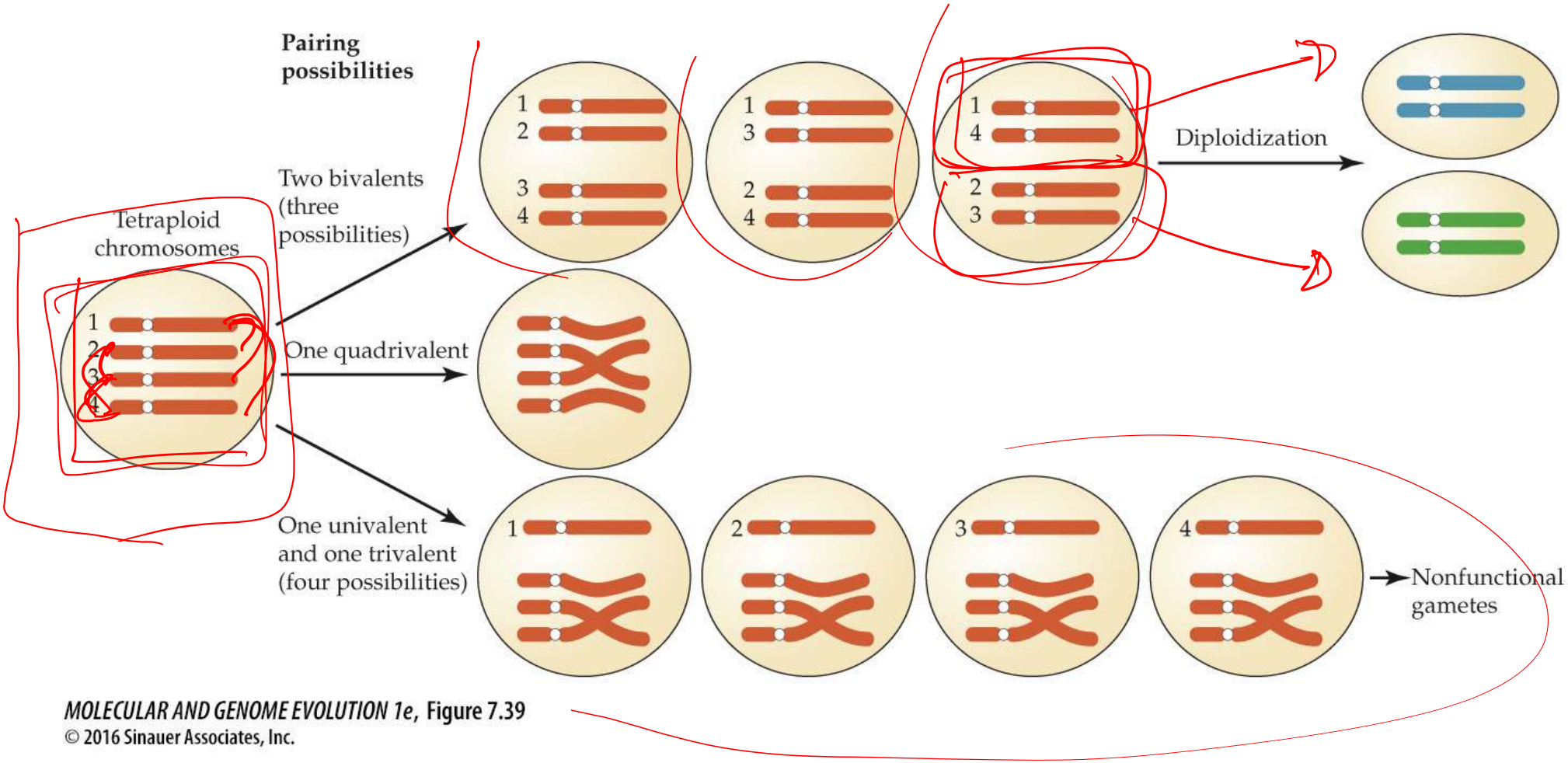
4 alleles at a locus
(one gene)

2 independent loci
(2 genes – paralogs)

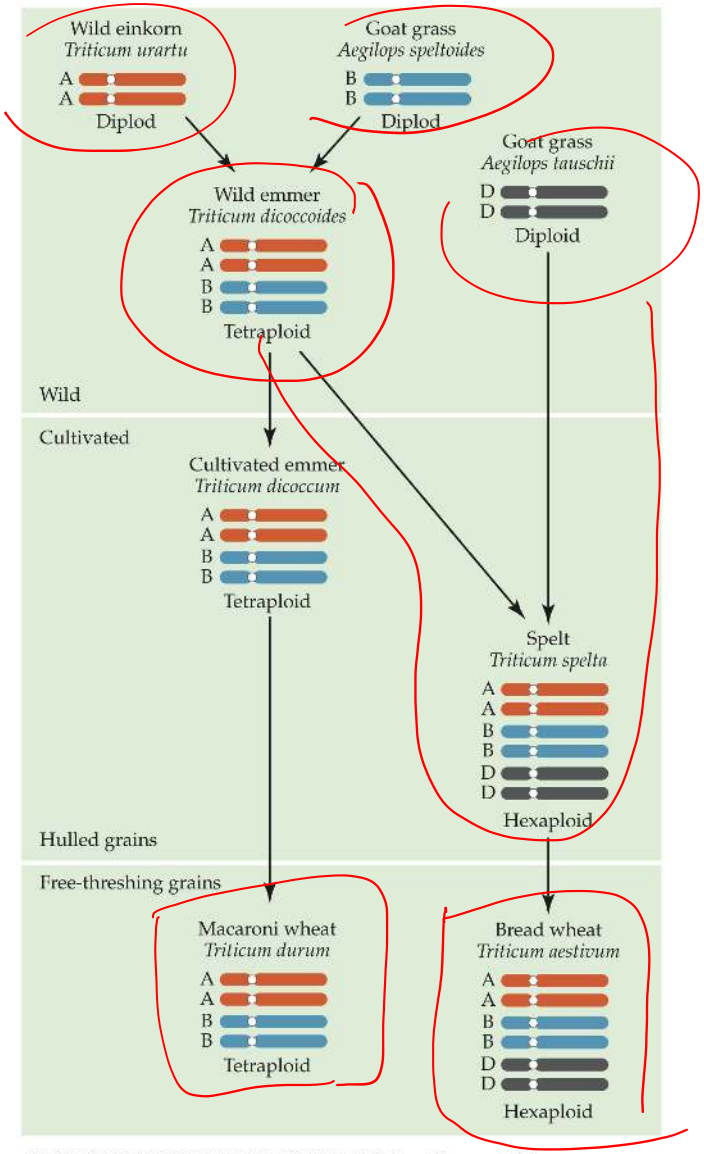
Gene conversion



MOLECULAR AND GENOME EVOLUTION 1e, Figure 7.23



MOLECULAR AND GENOME EVOLUTION 1e, Figure 7.39

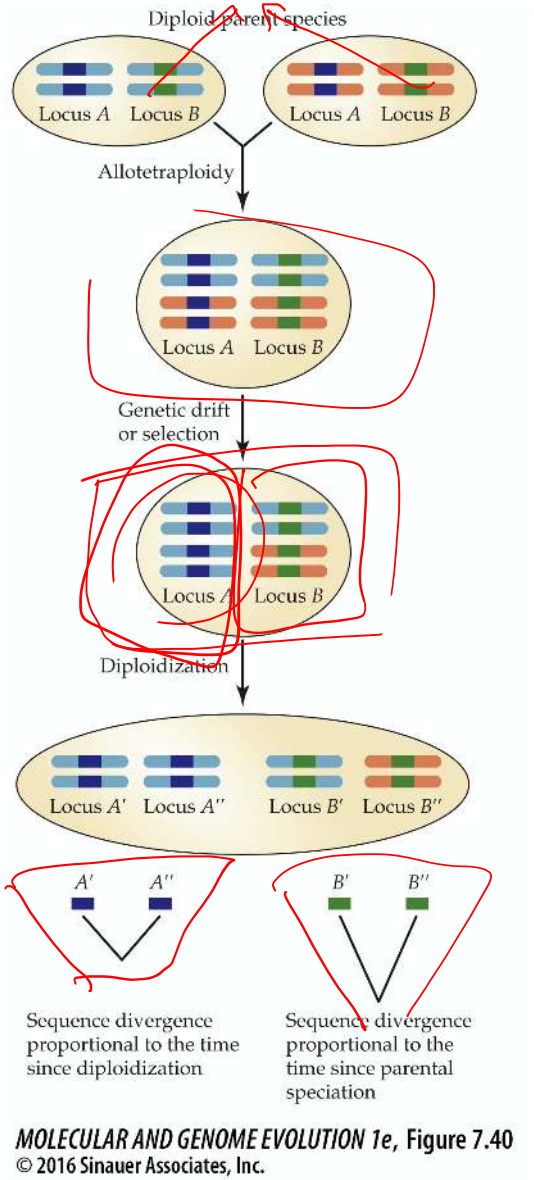


MOLECULAR AND GENOME EVOLUTION 1e, Figure 7.38

Allopolyploidy vs autopolyploidy

hybridisation → wheat

self duplication → perfect copy



MOLECULAR AND GENOME EVOLUTION 1e, Figure 7.40

Phylogenetic consequences of tetrasomic locus resolution

ARTICLE

OPEN

doi:10.1038/nature17164

The Atlantic salmon genome provides insights into rediploidization

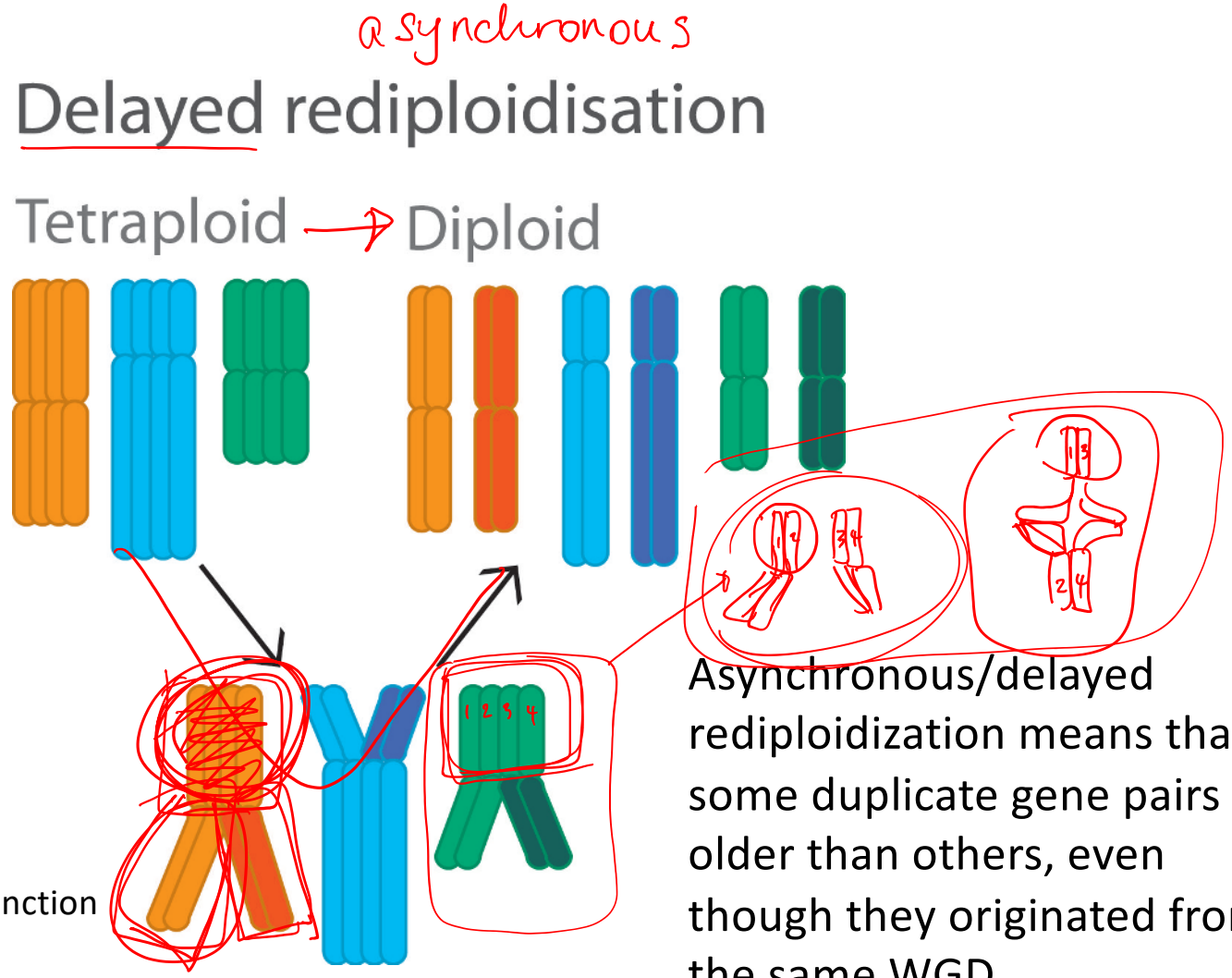
Sigbjørn Lien[1], Ben F. Koop[2], Simen R. Sandve[1], Jason R. Miller[3], Matthew P. Kent[1], Torfinn Nome[1], Torgeir R. Hvidsten[4,5], Jong S. Leong[2], David R. Minkley[2], Aleksey Zimin[6], Fabian Grammes[1], Harald Grove[1], Arne Gjuvsland[1], Brian Walenz[3], Russell A. Hermansen[7,8,9], Kris von Schalburg[2], Eric B. Rondeau[3], Alex Di Genova[10,11], Jeevan K. A. Samy[1], Jon Olav Vik[1], Magnus D. Vigeland[12], Lis Caler[3], Unni Grimholt[13], Sissel Jentoft[14], Dag Inge Våge[1], Pieter de Jong[15], Thomas Moen[16], Matthew Baranski[17], Yniv Palti[18], Douglas R. Smith[19,20], James A. Yorke[6], Alexander J. Nederbragt[14], Ave Tooming-Klunderud[14], Kjetill S. Jakobsen[14], Xuanting Jiang[21], Dingding Fan[21], Yan Hu[21], David A. Liberles[8,9], Rodrigo Vidal[22], Patricia Iturra[23], Steven J. M. Jones[24,25], Inge Jonassen[26], Alejandro Maass[10,11], Stig W. Omholt[27] & William S. Davidson[25]

The whole-genome duplication 80 million years ago of the common ancestor of salmonids (salmonid-specific fourth vertebrate whole-genome duplication, Ss4R) provides unique opportunities to learn about the evolutionary fate of a duplicated vertebrate genome in 70 extant lineages. Here we present a high-quality genome assembly for Atlantic salmon (*Salmo salar*), and show that large genomic reorganizations, coinciding with bursts of transposon-mediated repeat expansions, were crucial for the post-Ss4R rediploidization process. Comparisons of duplicate gene expression patterns across a wide range of tissues with orthologous genes from a pre-Ss4R outgroup unexpectedly demonstrate far more instances of neofunctionalization than subfunctionalization. Surprisingly, we find that genes that were retained as duplicates after the teleost-specific whole-genome duplication 320 million years ago were not more likely to be retained after the Ss4R, and that the duplicate retention was not influenced to a great extent by the nature of the predicted protein interactions of the gene products. Finally, we demonstrate that the Atlantic salmon assembly can serve as a reference sequence for the study of other salmonids for a range of purposes.

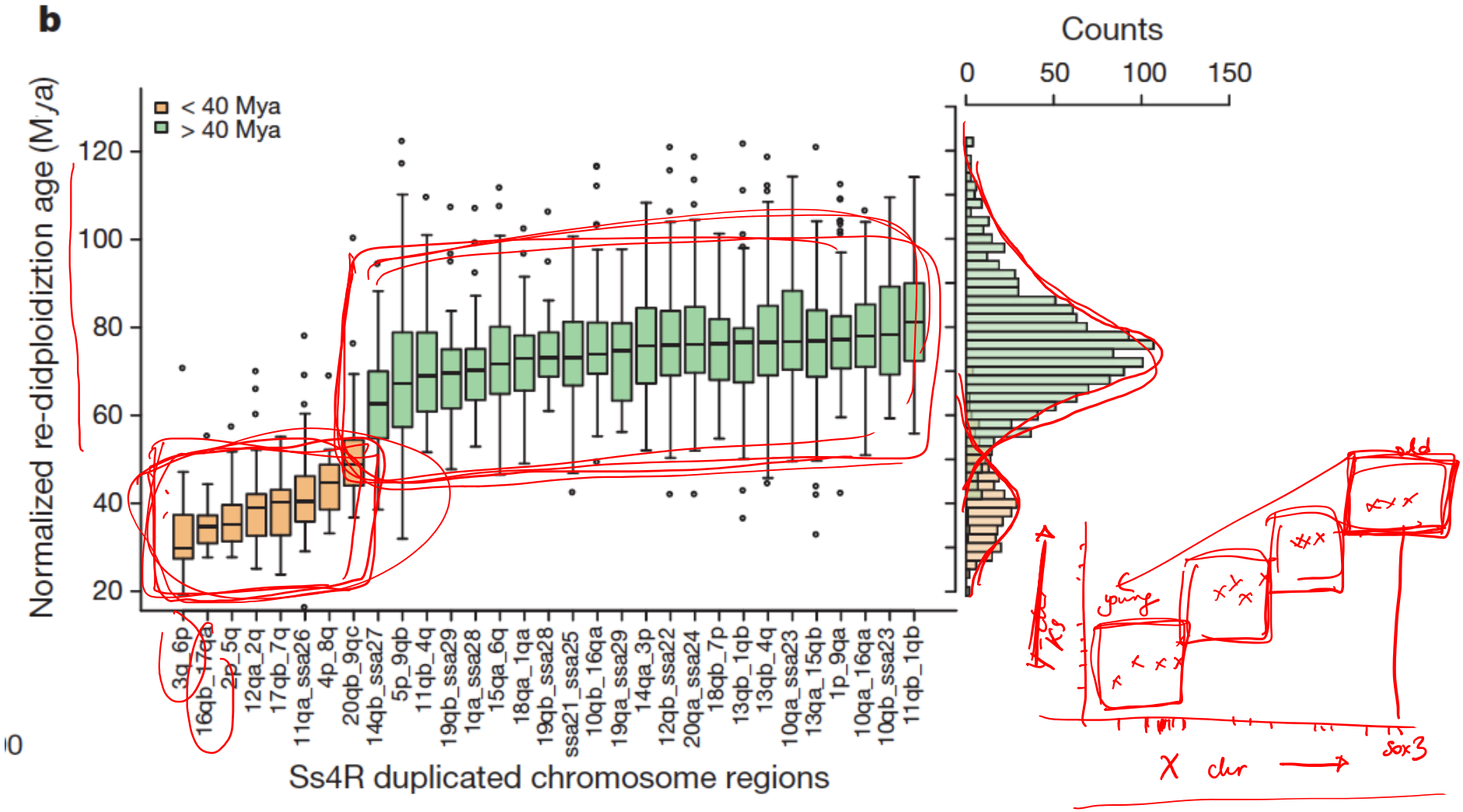
2016 - Nature 533 (7602): 200–205. doi:10.1038/nature17164.

Alleles at a locus
- allelic variation is possible
- sequence divergence is limited by gene conversion

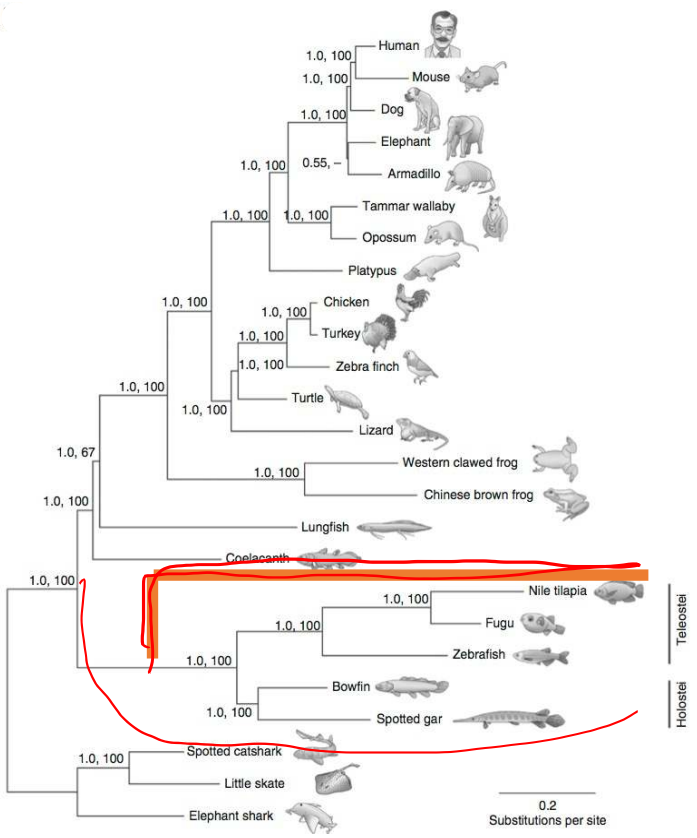
Duplicated loci (duplicate genes)
- free to diverge in sequence and function

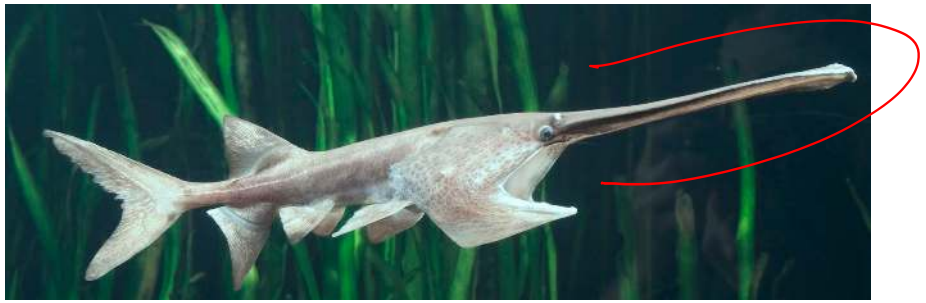


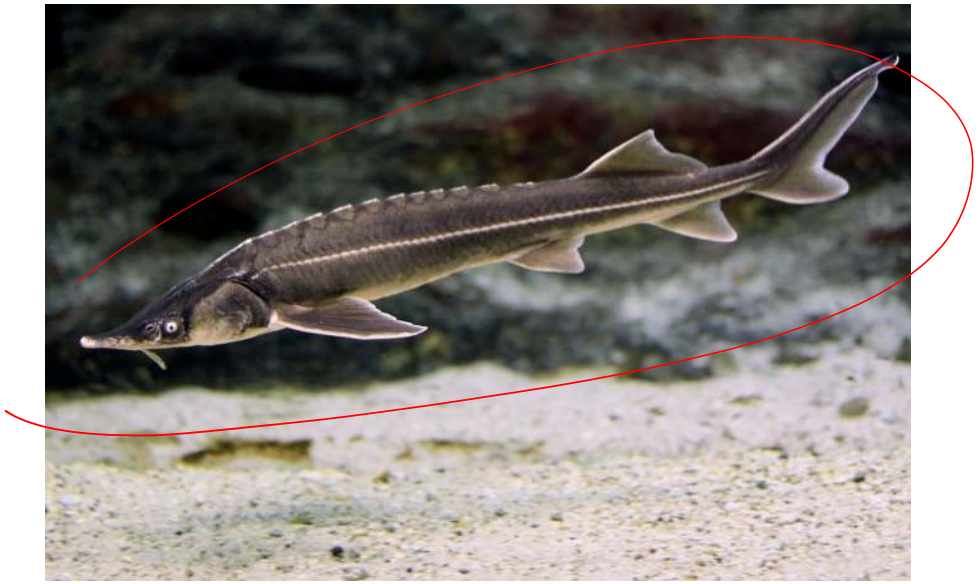
Asynchronous/delayed rediploidization means that some duplicate gene pairs are older than others, even though they originated from the same WGD.


b
Counts
0
50
100
150
< 40 Mya
> 40 Mya
Normalized re-didploidiztion age (Mya)
120
100
80
60
40
20
3q_6p
16qb_17qa
2p_5q
12qa_2q
17qb_7q
11qa_ssa26
4p_8q
20qb_9qc
14qb_ssa27
5p_9qb
11qb_4q
19qb_ssa29
1qa_ssa28
15qa_6q
18qa_1qa
19qb_ssa28
ssa21_ssa25
10qb_16qa
19qa_ssa29
14qa_3p
12qb_ssa22
20qa_ssa24
18qb_7p
13qb_1qb
13qb_4q
10qa_ssa23
13qa_15qb
1p_9qa
10qa_16qa
10qb_ssa23
11qb_1qb
Ss4R duplicated chromosome regions
young
old
X chr
Sox3

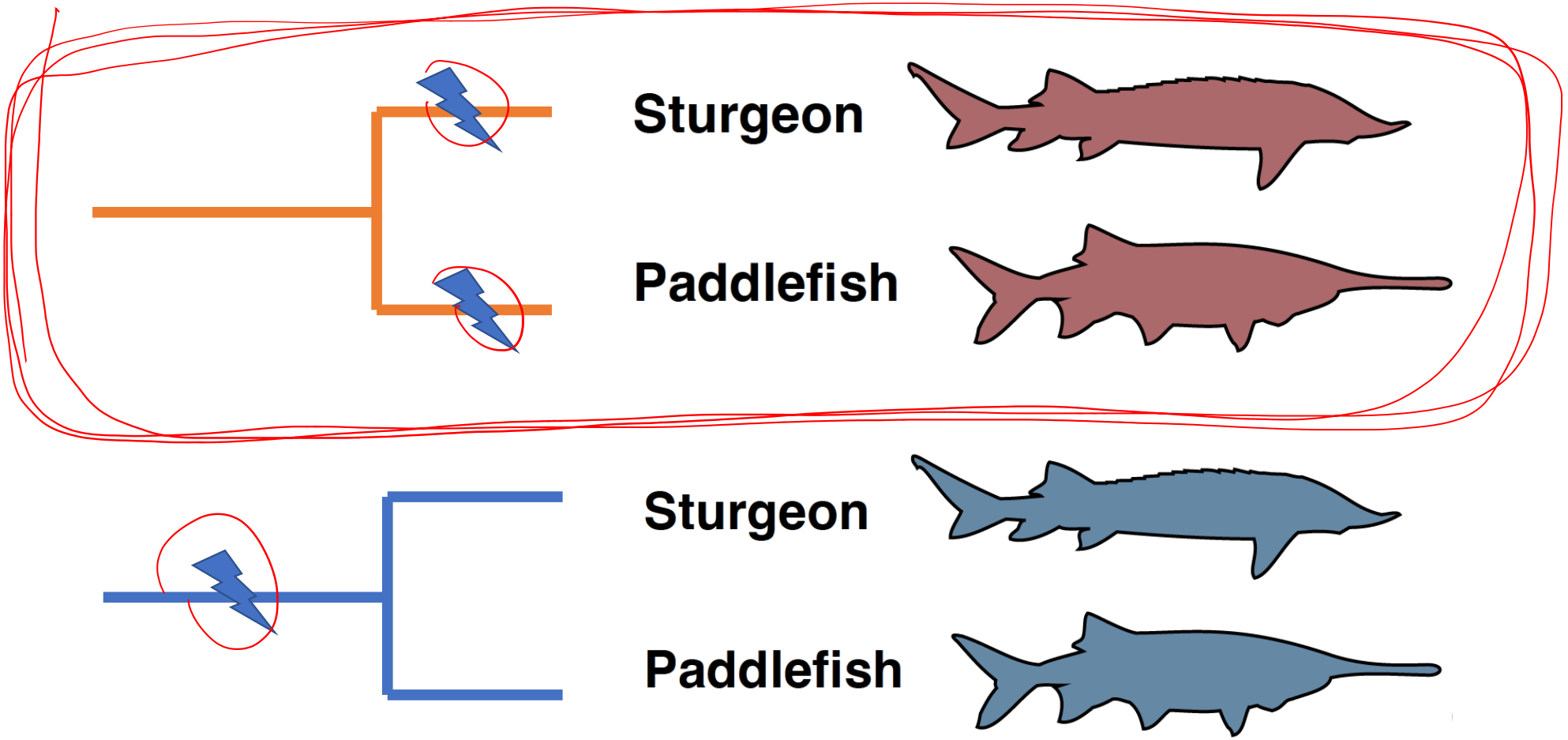
Genome duplications in sturgeon and paddlefish



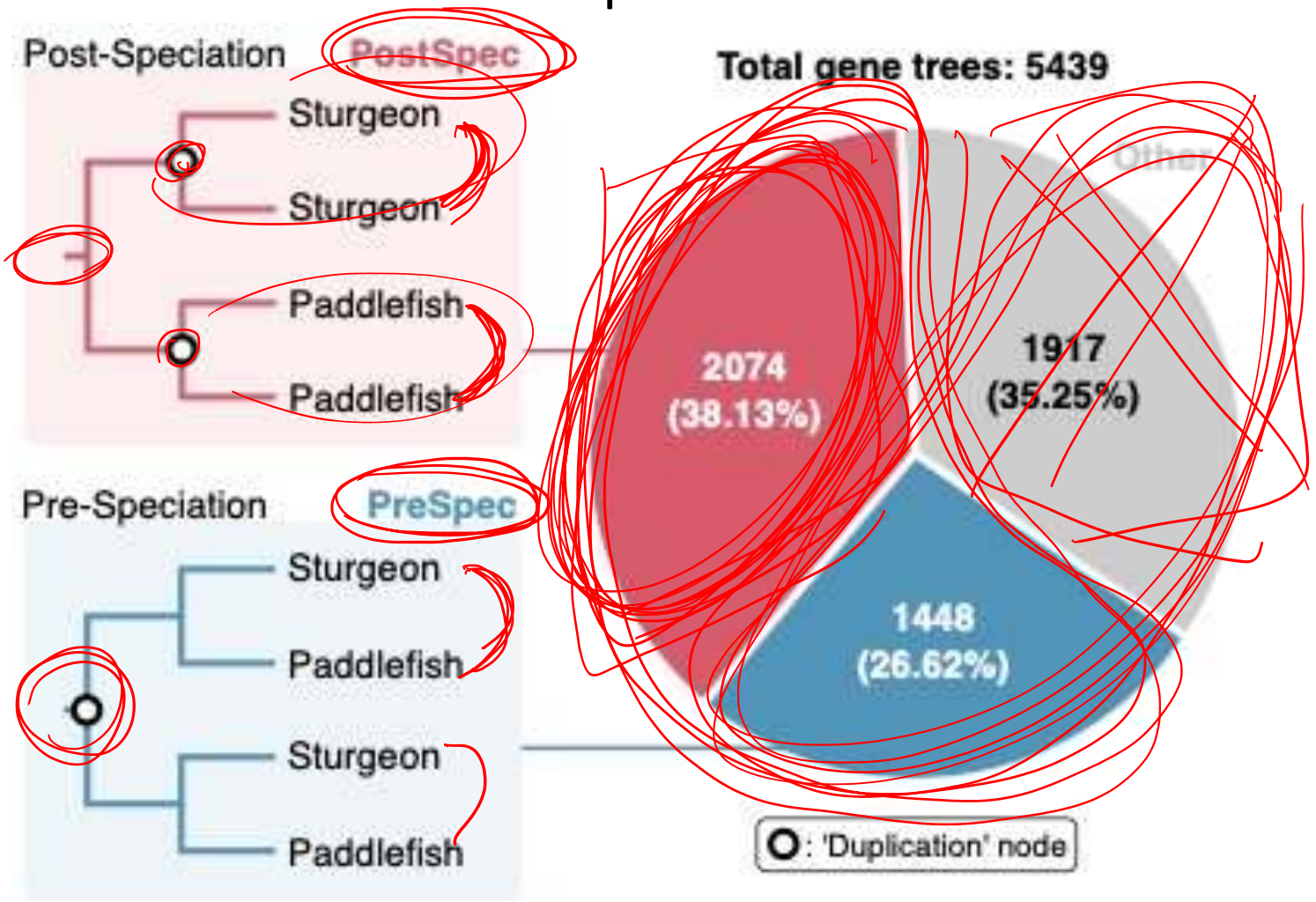


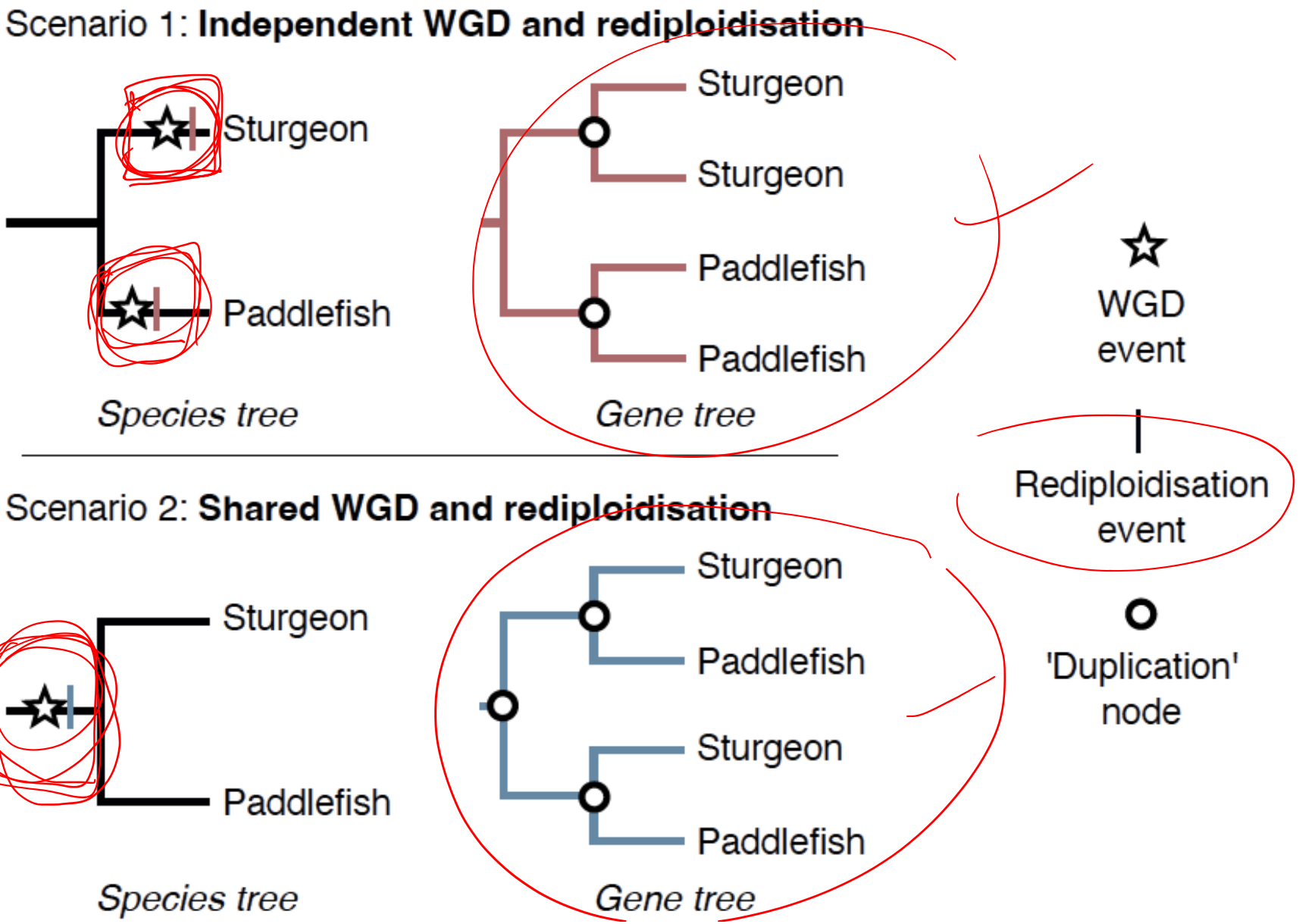


One whole genome duplication or two?

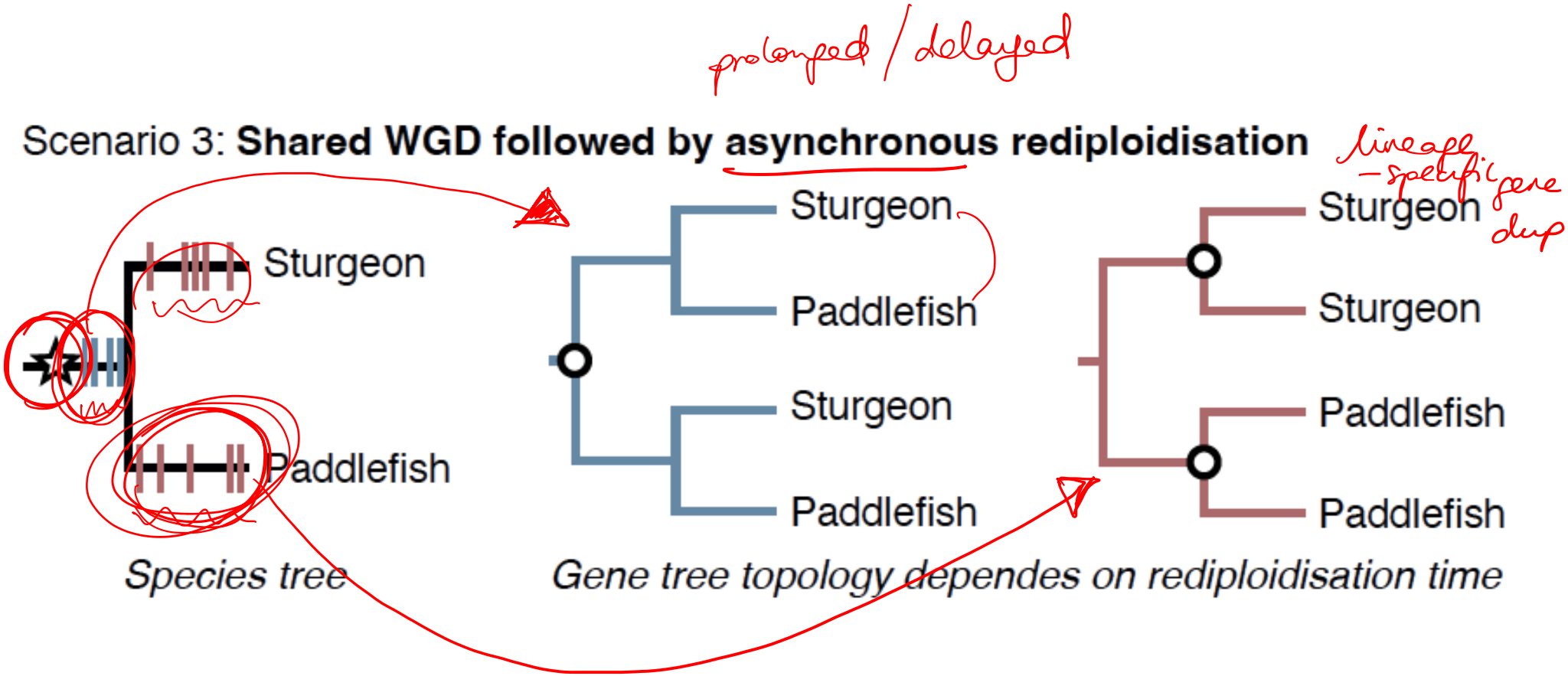


A plurality of trees show independent duplication

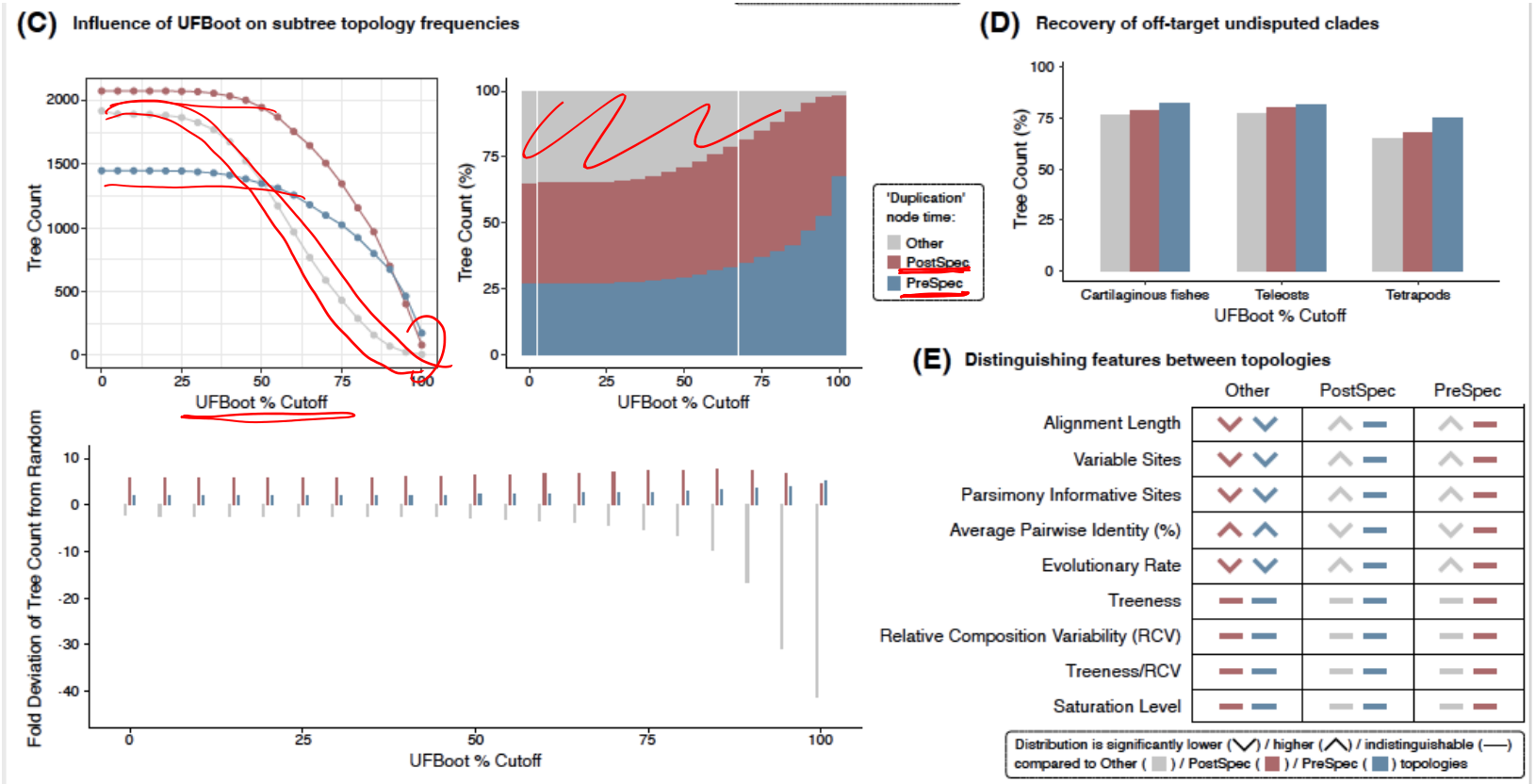



Scenario 1: Independent WGD and rediploidisation
Sturgeon
Paddlefish
Species tree
Sturgeon
Sturgeon
Paddlefish
Paddlefish
Gene tree
WGD event
Rediploidisation event
Scenario 2: Shared WGD and rediploidisation
Sturgeon
Paddlefish
Species tree
Sturgeon
Paddlefish
Sturgeon
Paddlefish
Gene tree
'Duplication' node

Gene tree topology depends on rediploidisation time

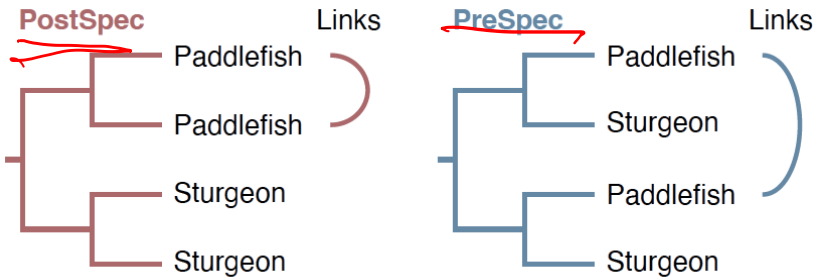


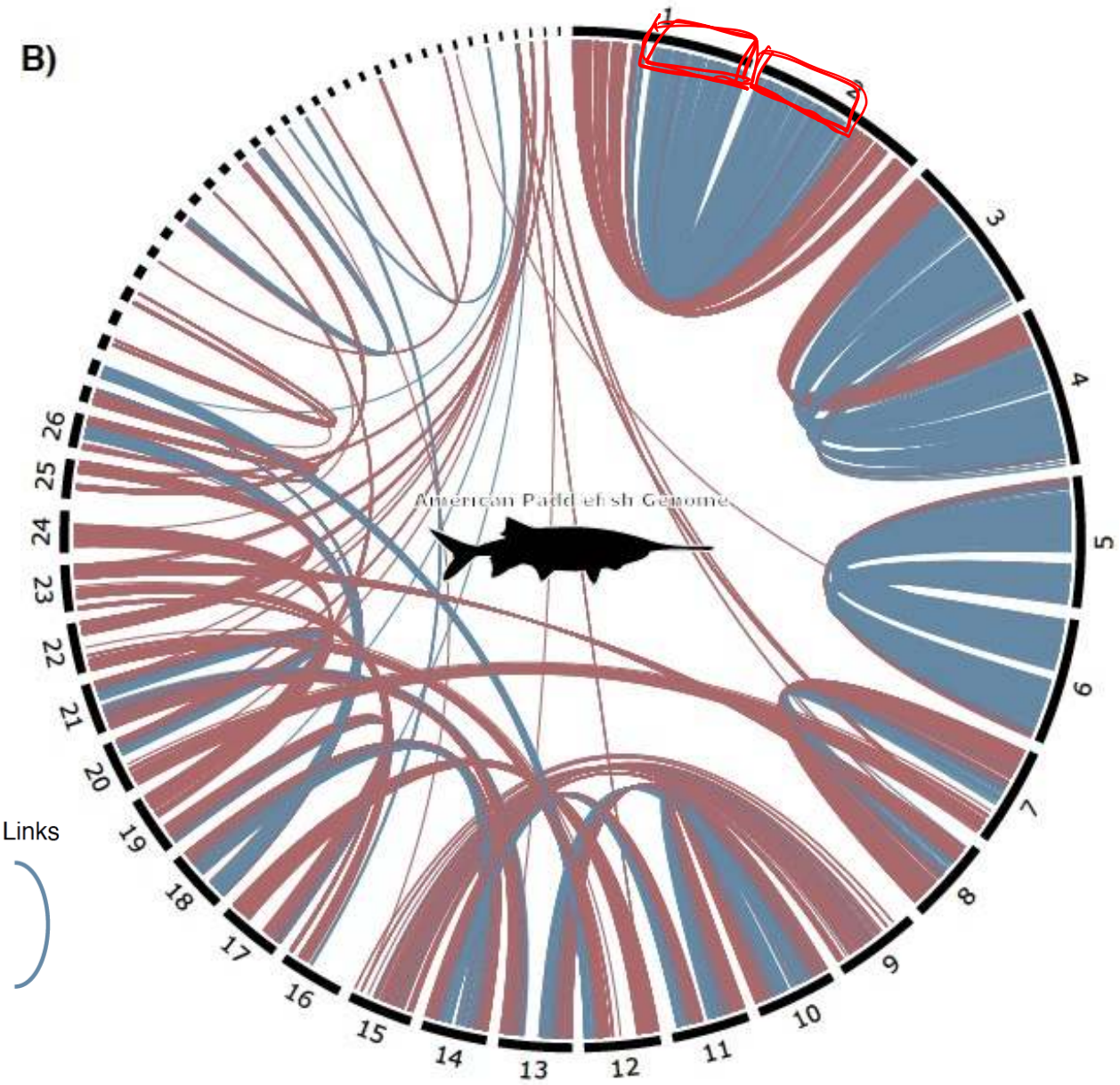
The trees are good : both postSpec and preSpec topologies are robust – not due to error



Topologies are clustered in the genome

Looks like ‘blocks’ of rediploidisation

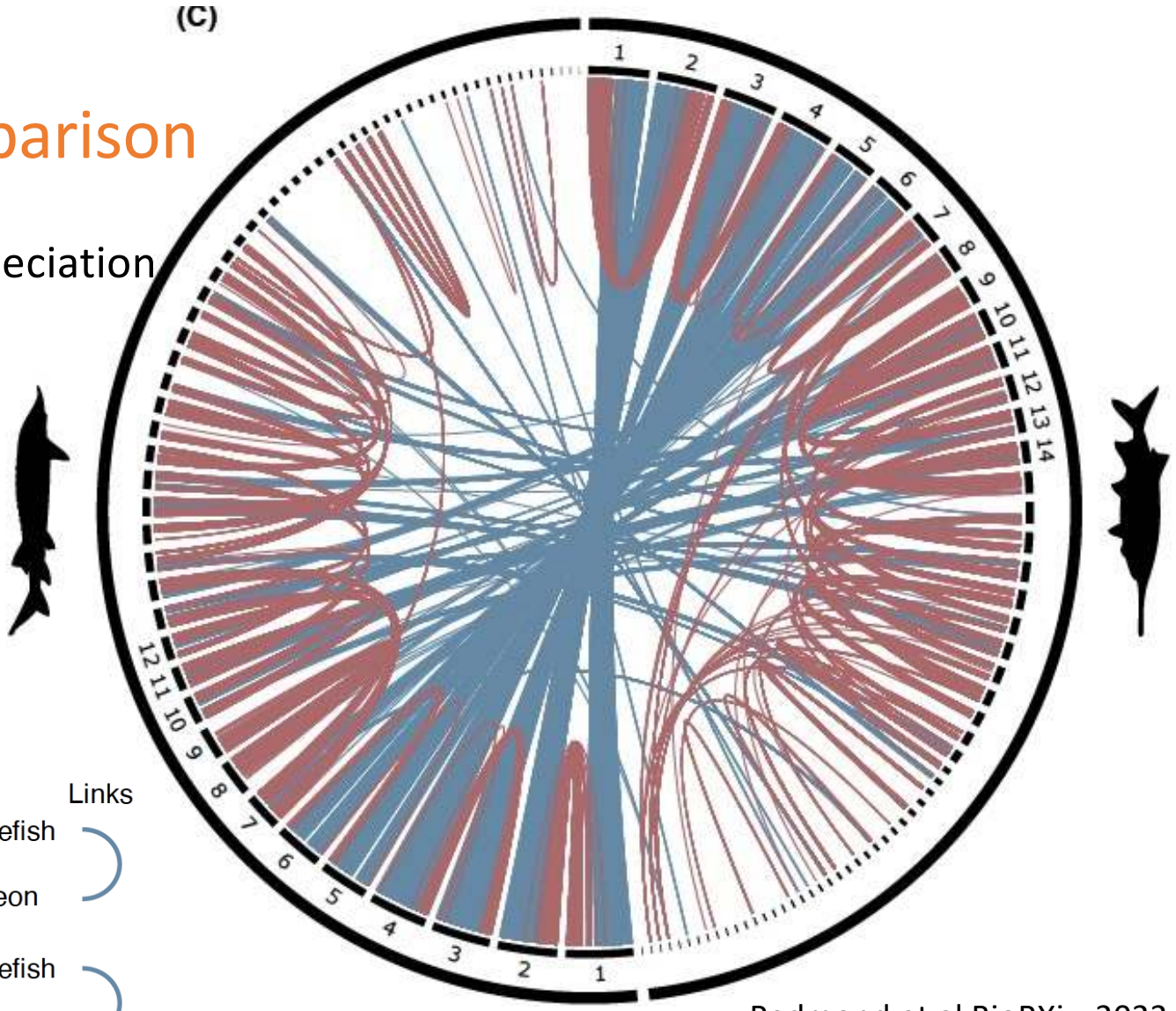


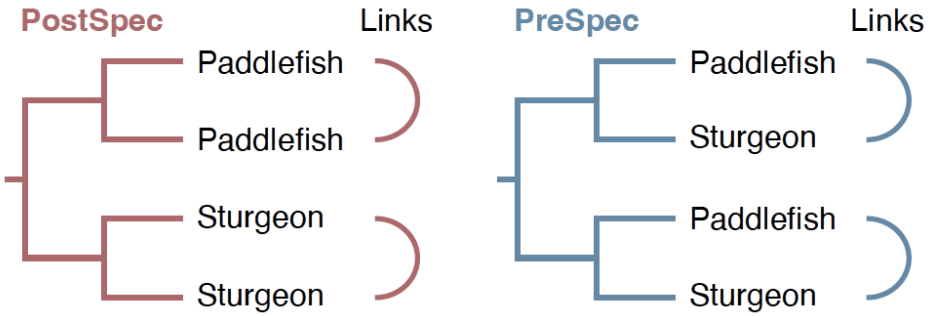


Between genome comparison

Blue blocks rediploidised before speciation

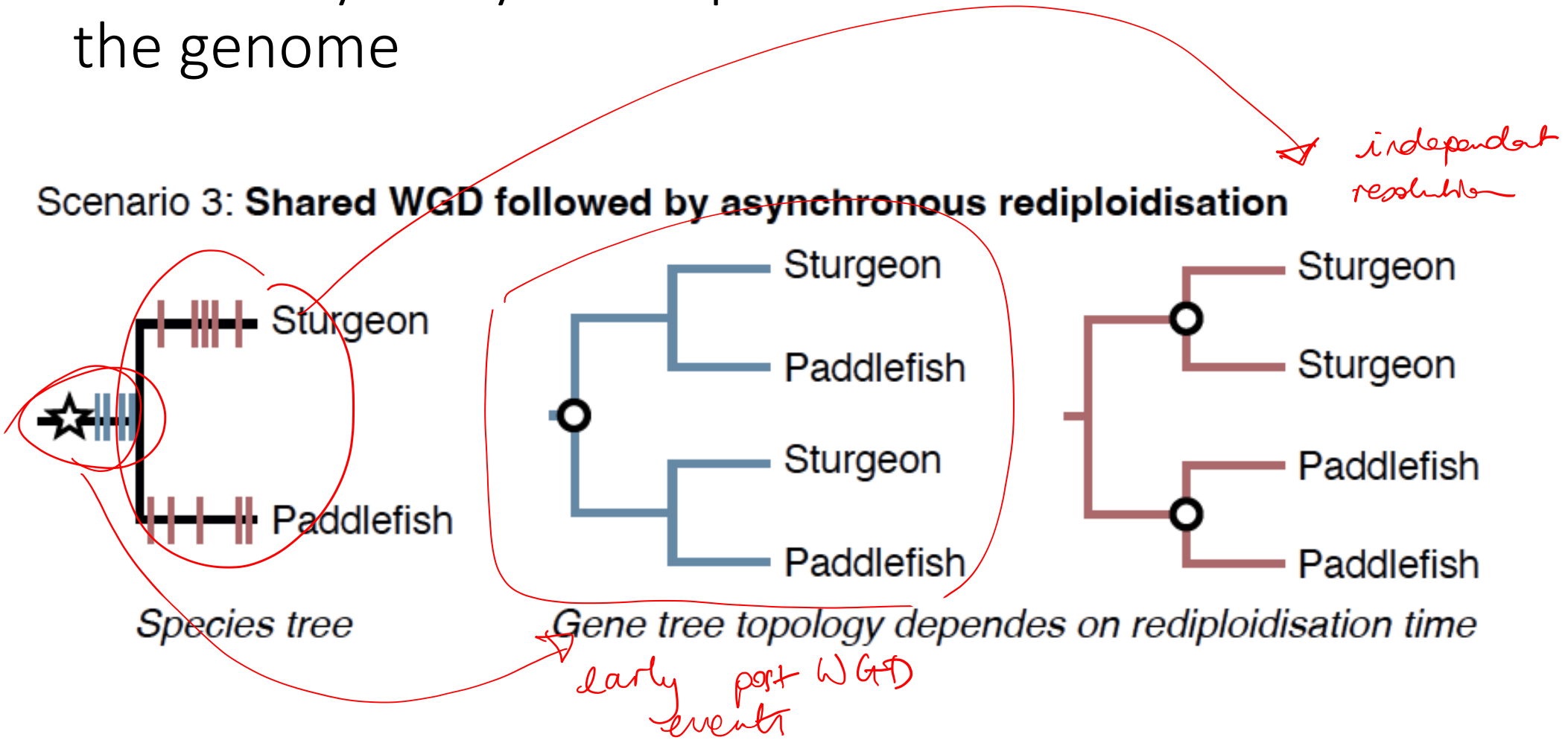
Red blocks after speciation

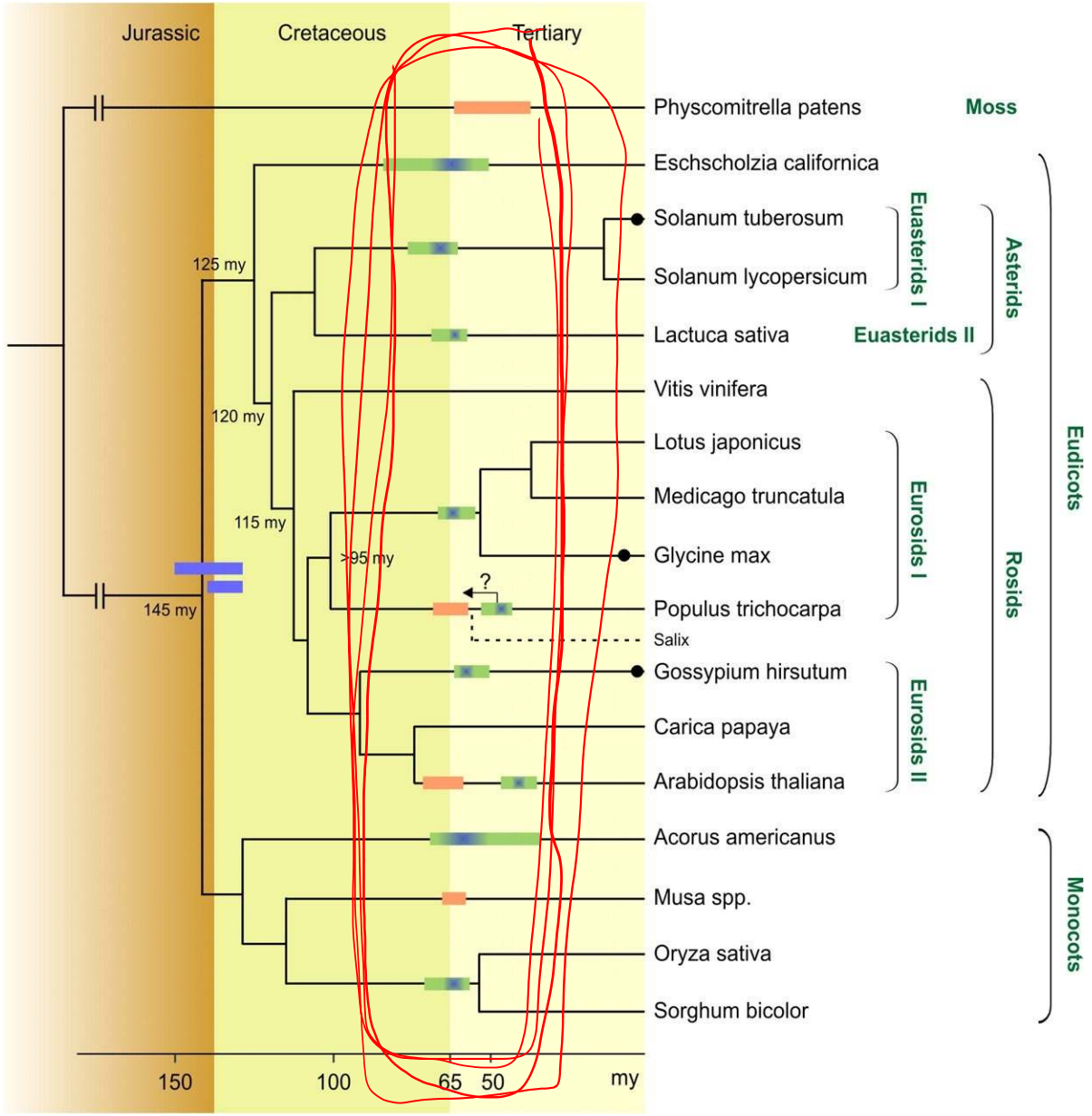




Redmond et al BioRXiv, 2022

Shared WGD in sturgeon-paddlefish ancestor is masked by delayed rediploidization of over half the genome

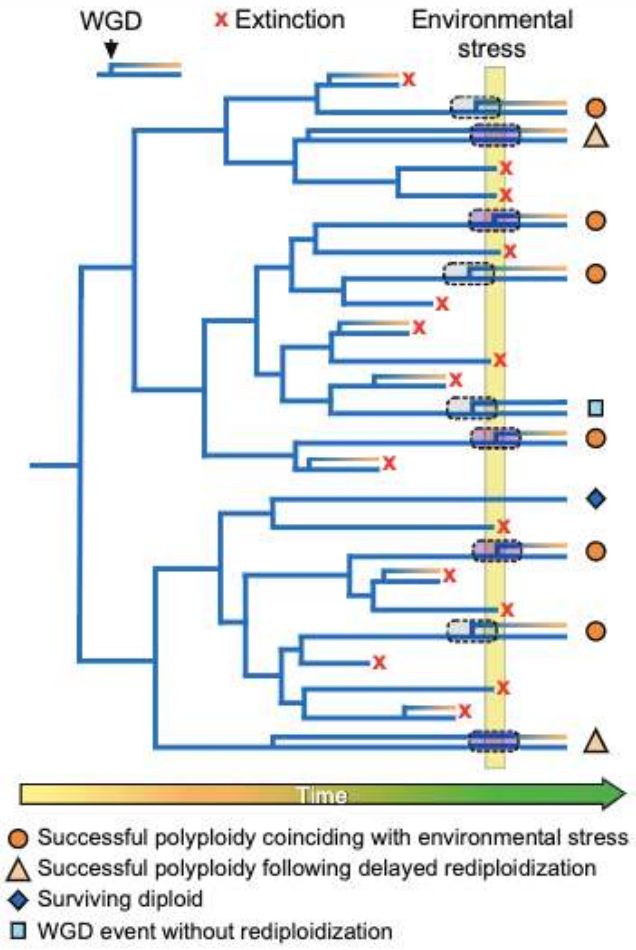




Does WGD help organisms survive periods of stress or environmental change?

Fawcett, J. A., Maere, S. & Peer, Y. V. de. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *PNAS* **106**, 5737–5742 (2009).

Polyploidy and stress



- Polyploidy is destabilising – why is it sometimes successful?
  - increased genetic variation (allopolyploids)
  - buffering effect of paralogs - increased mutational robustness
  - Polyploidy-mediated gene flow & hybridization
  - Immediate changes in anatomical structures and some physiological processes
  - Gene expression changes, epigenetic remodelling
- Many of these are suboptimal for the ancestral environment, but in a changing environment may present a faster route to adaptation.

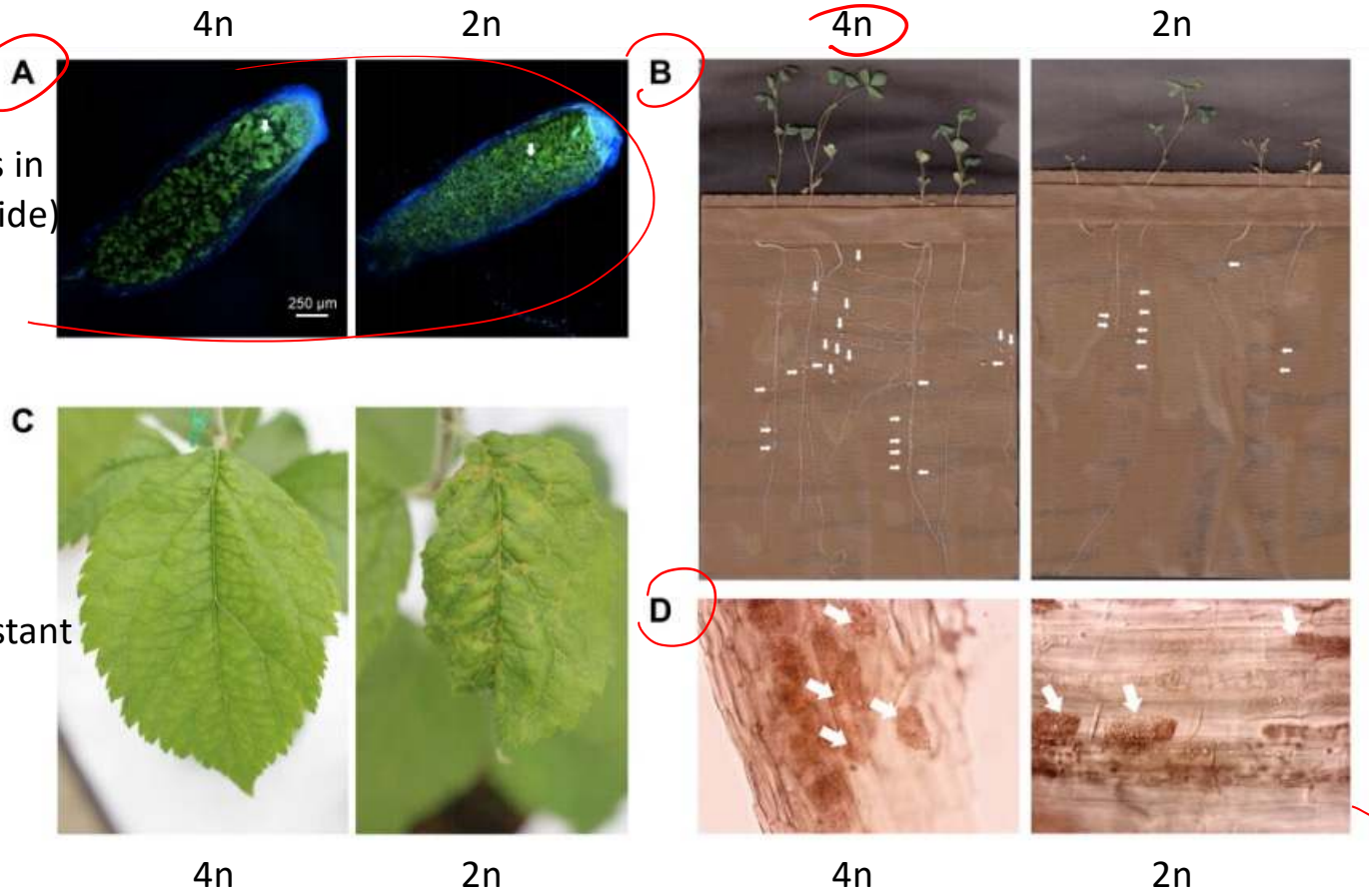
Peer, Y. V. de, Ashman, T.-L., Soltis, P. S. & Soltis, D. E. Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell* **33**, koaa015 (2020).

Neotetraploid (4n) plants have enhanced mutualistic interactions (A,B,D) and better resistance to pathogenic interactions (C).

Larger root nodules in 4n (A close-up, B wide) alfalfa (Medicago)

4n apple more resistant to fungal infection

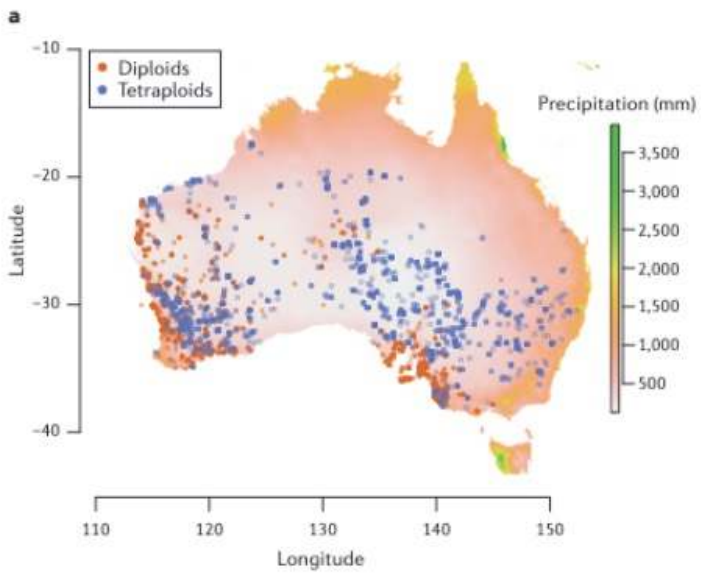
Higher colonization by mutualistic fungi in 4n Heuchera roots

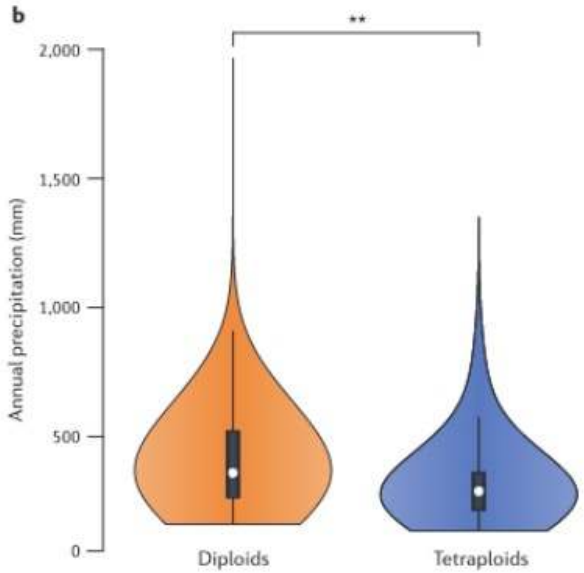


Peer, Y. V. de, Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nature Reviews Genetics* **18**, 411–424 (2017).

Tetraploid frogs have a wider distribution and tolerate drought better than diploids

Polyploidy-mediated gene flow & hybridization





Peer, Y. V. de, Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nature Reviews Genetics* **18**, 411–424 (2017).
Novikova, P. Yu. *et al.* Polyploidy breaks speciation barriers in Australian burrowing frogs Neobatrachus. *Plos Genet* **16**, e1008769 (2020).

Take home messages

- WGD is a powerful evolutionary event that causes great genomic upheaval and adds lots of genes to the genome
- Common in plants, but comparatively rare in animals
- Evidence for WGD comes from the distribution of paralogs in the genome – clusters/blocks of duplicated genes; many blocks; blocks originated in same/similar time period and are (mostly) non-overlapping
- 2 rounds of WGD happened around the origin of vertebrates
- The legacy of these WGD events includes many paralogs (ohnologs) in vertebrate genomes. These ohnologs are disproportionately associated with disease.
- WGD is often associated with species diversification and innovation
- Mounting evidence that links WGD to survival of stress, including climate change
- Generally speaking, it takes a combination of phylogenetics and synteny analysis to study WGD …. phylogenomics