Genomics of eukaryotic microorganisms

Isheng Jason Tsai

[2019 version]

Lecture objective

- Classification of eukaryotic microbes – a very general description
- Comparative genomics
  - Inferring orthology
  - Phylogenomics
- Case studies

Classification of eukaryotic microorganisms – a history

Robert Whittaker

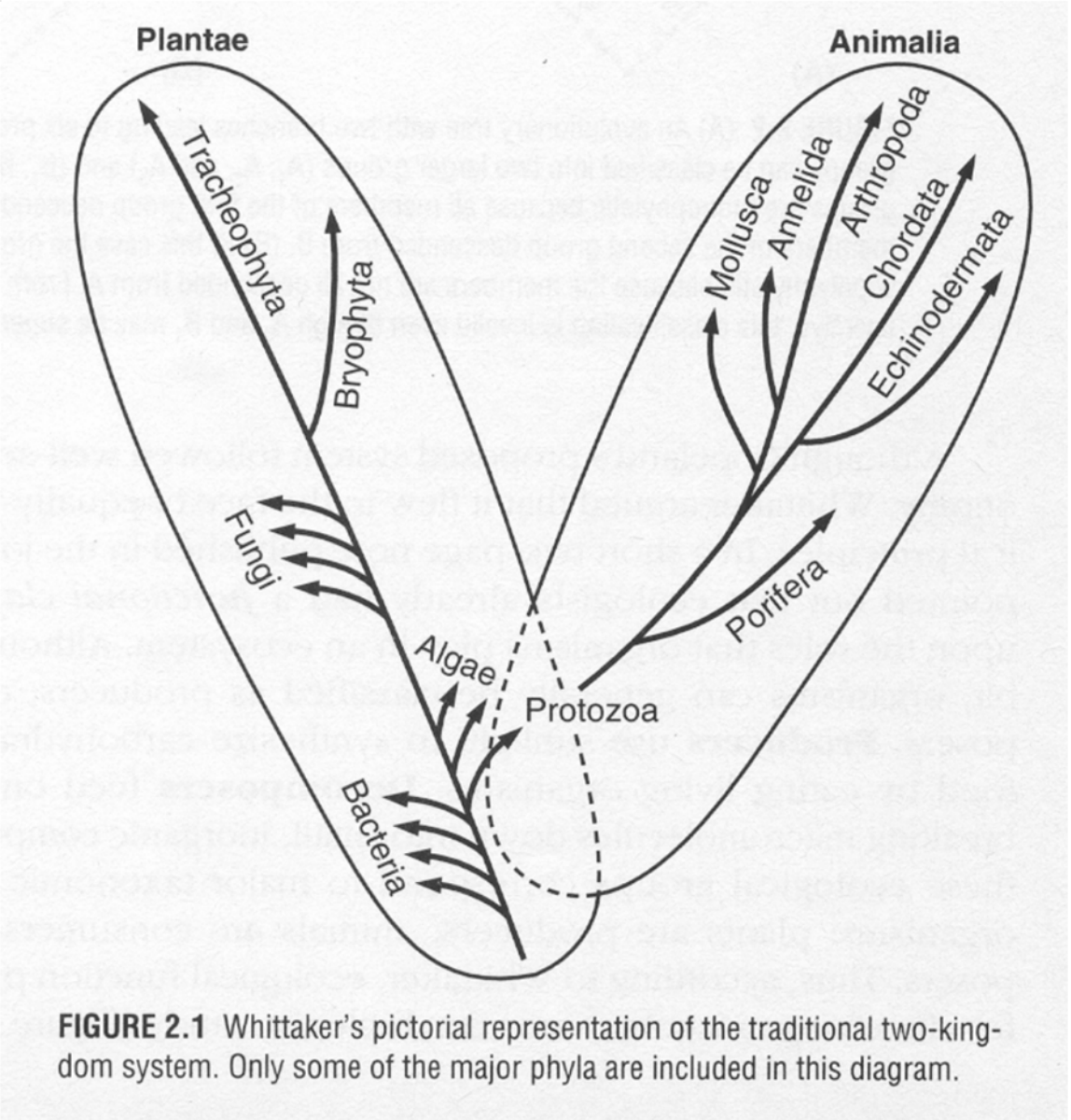


FIGURE 2.1 Whittaker's pictorial representation of the traditional two-kingdom system. Only some of the major phyla are included in this diagram.

Wiki http://shipseducation.net/db/whittaker.pdf

Classification of eukaryotic microorganisms – a history

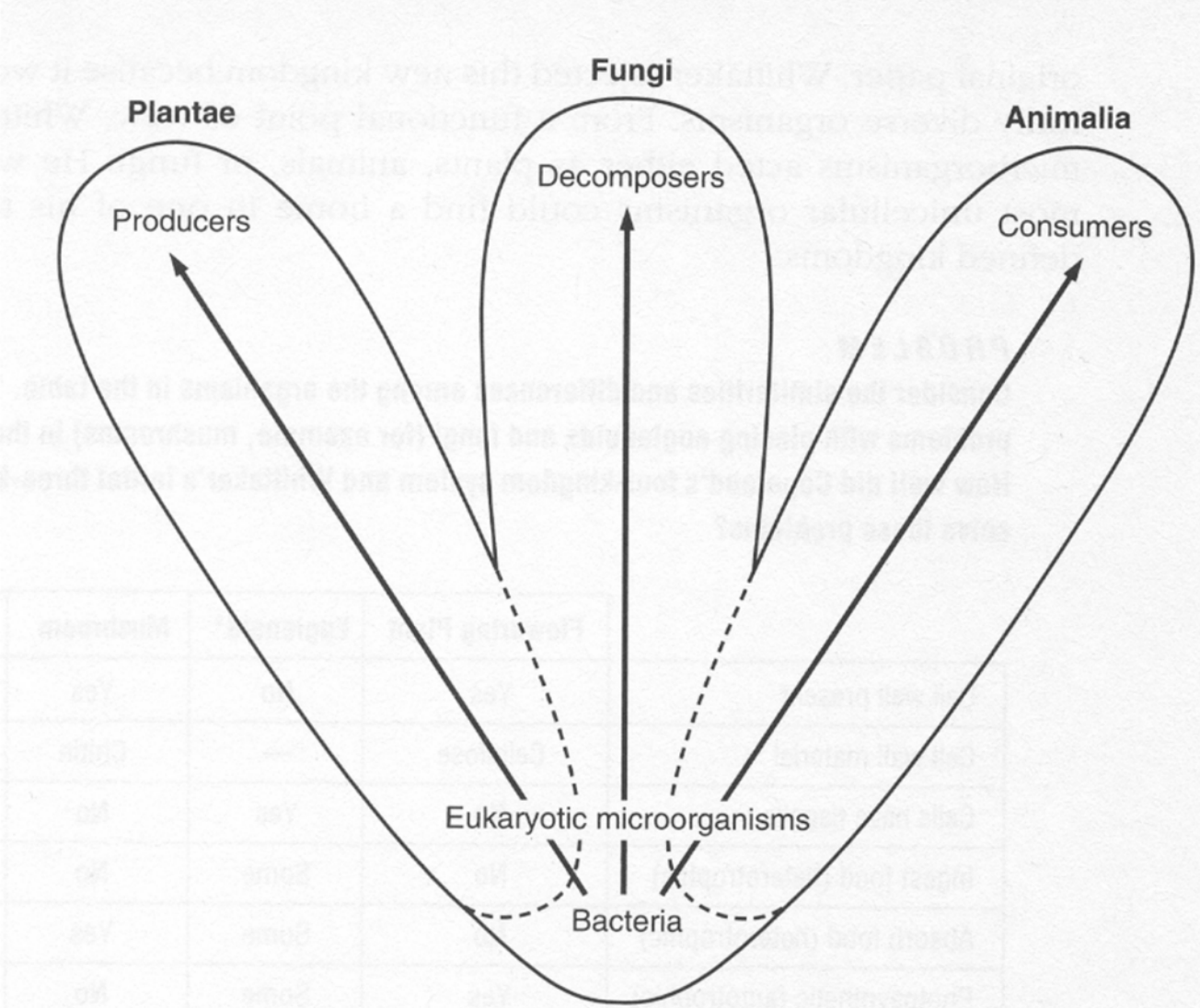


FIGURE 2.3 Whittaker's early three-kingdom system based upon ecological function.

- Ecological classification should reflect three major branches on the evolutionary tree
- Justification was appealing but had serious problems
- Example – Need to place most bacteria in kingdom Fungi
- How about algae and protozoans?

http://shipseducation.net/db/whittaker.pdf

Classification of eukaryotic microorganisms – four kingdoms (1957)

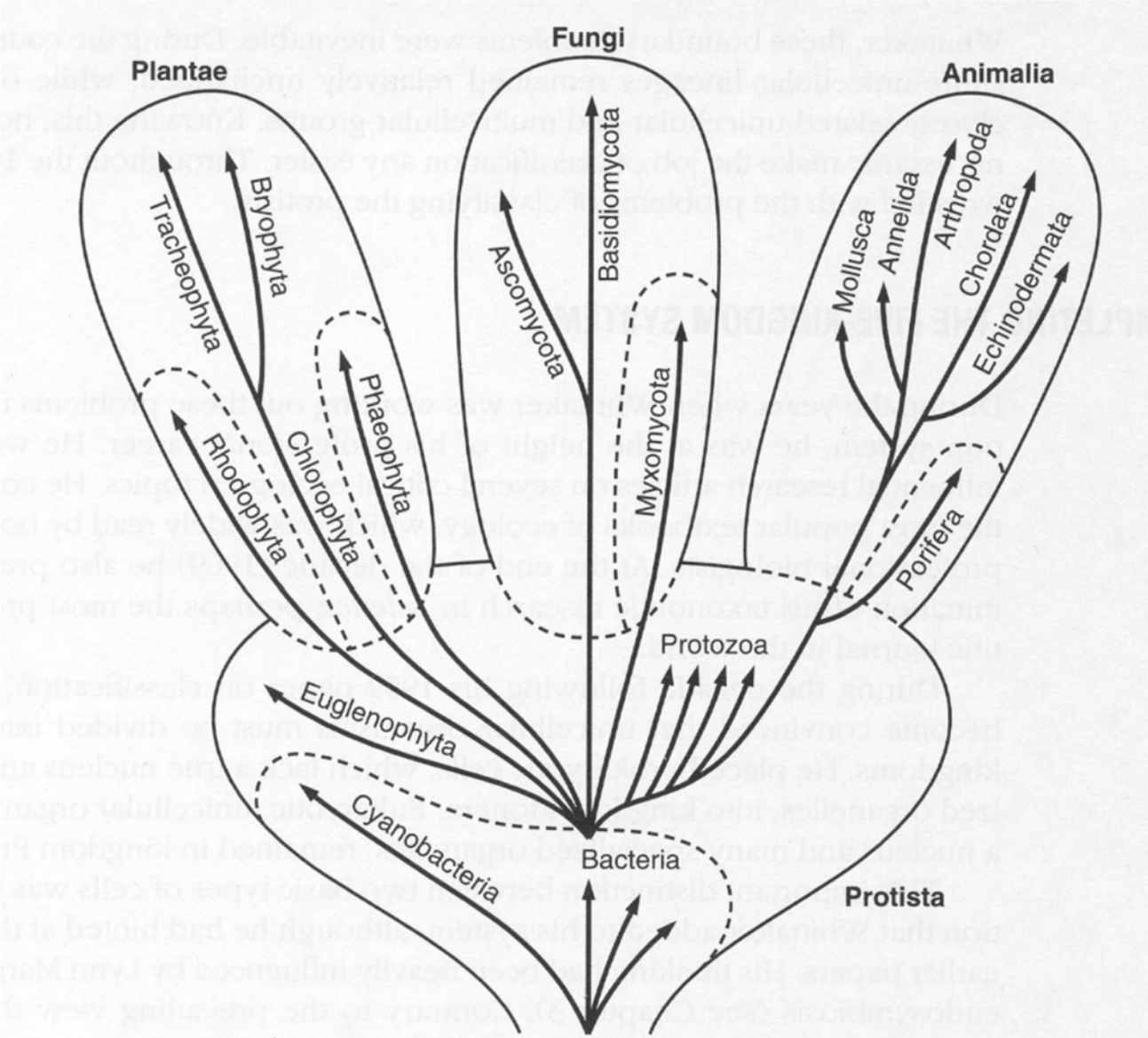
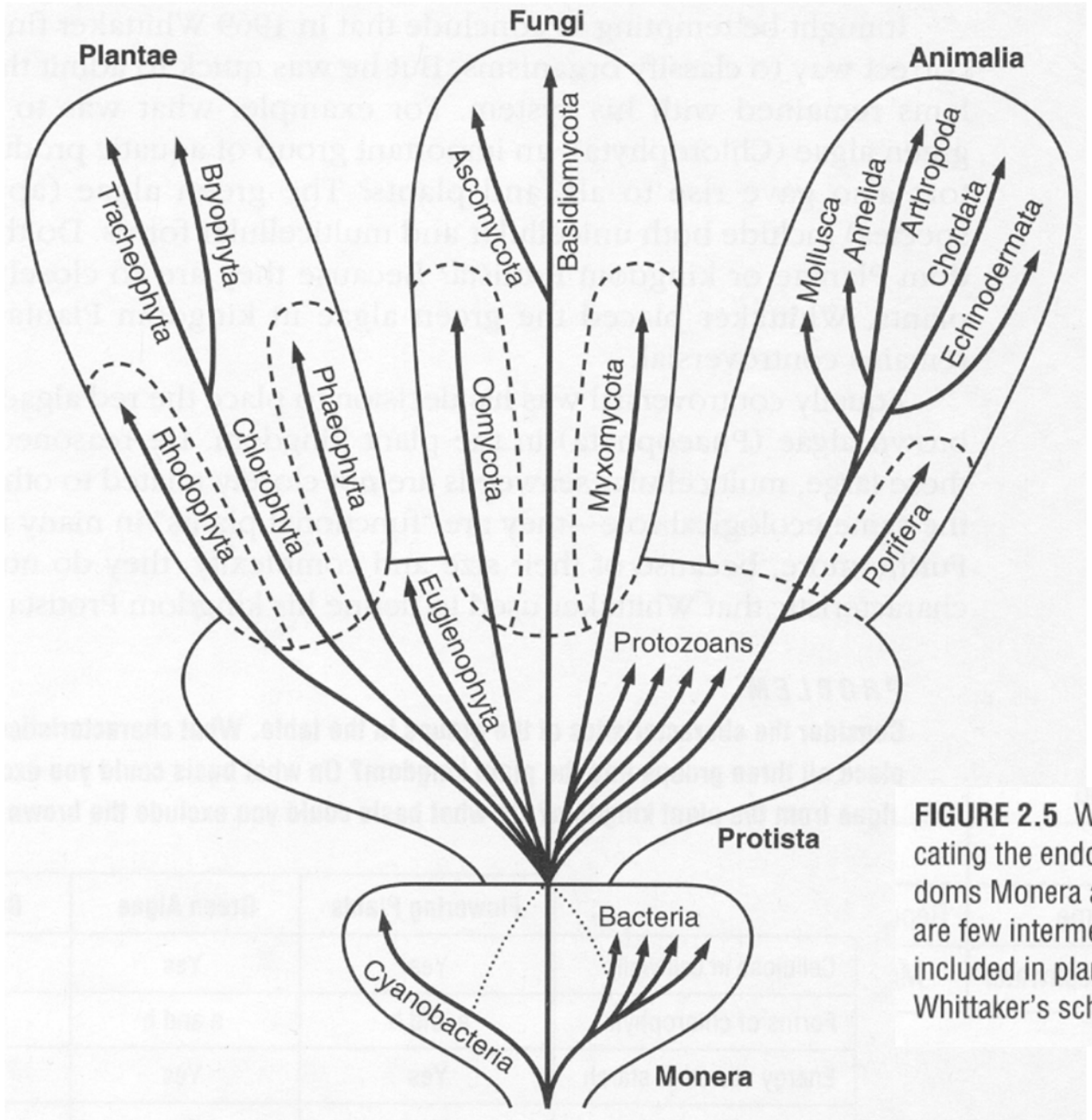


FIGURE 2.4 Whittaker's four-kingdom system. In this scheme, all unicellular organisms belong to kingdom Protista, which is divided into higher and lower subkingdoms.

http://shipseducation.net/db/whittaker.pdf

Classification of eukaryotic microorganisms – five kingdoms



- Widely adopted by all biologists in 1970s
- Still problematic in certain groups possessing both unicellular and multicellular organisms (for example, green algae in plantae or protista?

FIGURE 2.5 Whittaker's five-kingdom system. Notice the dotted lines in kingdom Monera indicating the endosymbiotic origin of eukaryotic cells. Also notice that the boundary between kingdoms Monera and Protista is very narrow, because according to the endosymbiotic theory there are few intermediaries between prokaryotic and eukaryotic cells. Ambiguous "problem groups" included in plants, fungi, and animals make each of these multicellular kingdoms polyphyletic in Whittaker's scheme.

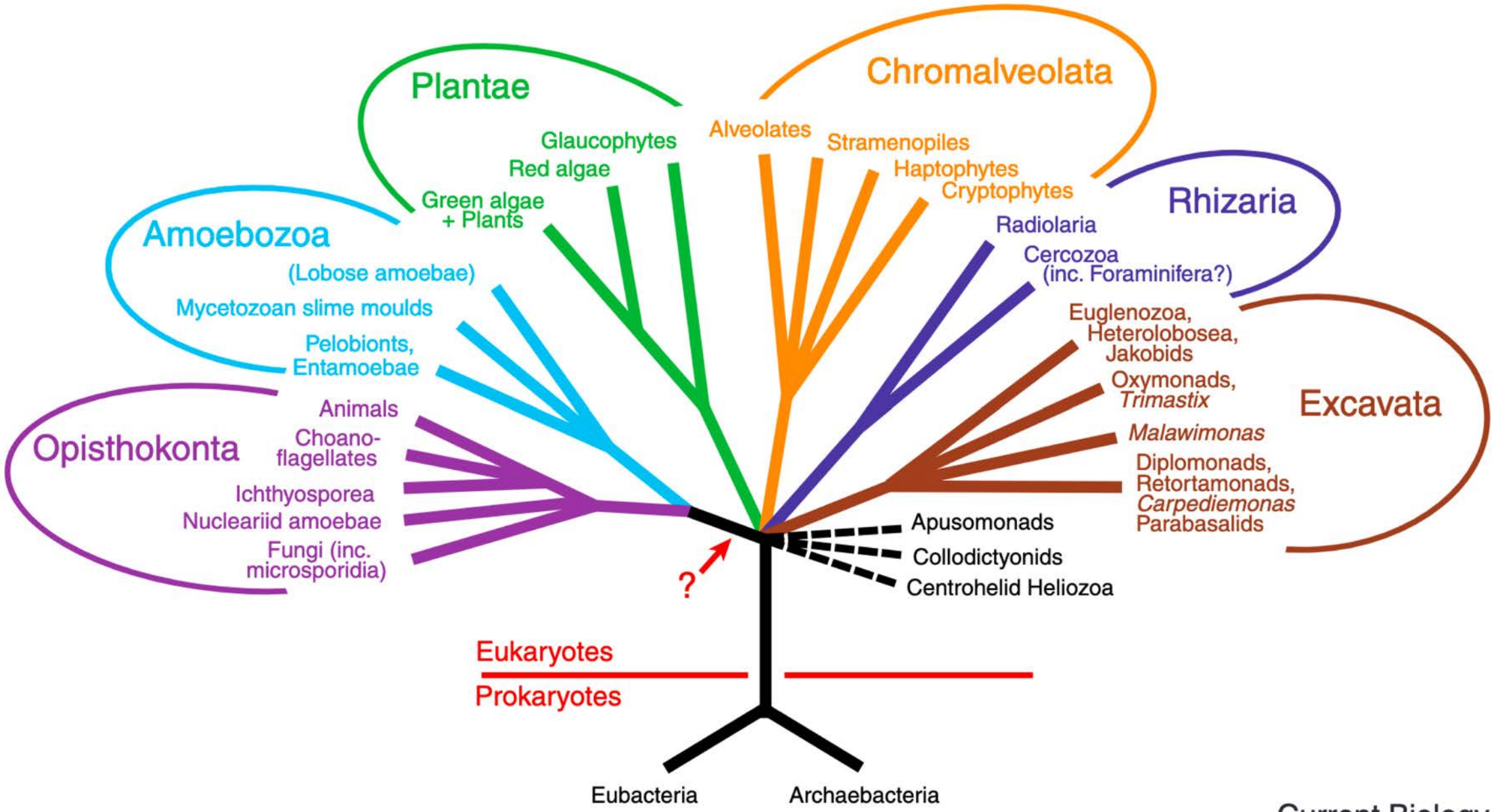
http://shipseducation.net/db/whittaker.pdf

Current standing of kingdoms

Linnaeus 1735[29]	Haeckel 1866[30]	Chatton 1925[31][32]	Copeland 1938[33][34]	Whittaker 1969[35]	Woese *et al.* 1977[36][37]	Woese *et al.* 1990[38]	Cavalier-Smith 1993[39][40][41]	Cavalier-Smith 1998[42][43][44]	Ruggiero *et al.* 2015[45]
2 kingdoms	3 kingdoms	2 empires	4 kingdoms	5 kingdoms	6 kingdoms	3 domains	8 kingdoms	6 kingdoms	7 kingdoms
(not treated)	Protista	Prokaryota	Monera	Monera	Eubacteria	Bacteria	Eubacteria	Bacteria	Bacteria
					Archaebacteria	Archaea	Archaebacteria		Archaea
		Eukaryota	Protista	Protista	Protista	Eucarya	Archezoa	Protozoa	Protozoa
							Protozoa		
							Chromista	Chromista	Chromista
Vegetabilia	Plantae		Plantae	Plantae	Plantae		Plantae	Plantae	Plantae
				Fungi	Fungi		Fungi	Fungi	Fungi
Animalia	Animalia		Animalia	Animalia	Animalia		Animalia	Animalia	Animalia

Wikipedia

The real ‘kingdoms’ of eukaryotes



Simpson and Roger (2004)**

The Revised Classification of Eukaryotes (2012)

		Super-groups	Examples
Eukaryota	Amorphea	Amoebozoa	Tubulinea Mycetozoa
		Opisthokonta	Fungi Choanomonada Metazoa
			Apusomonada
			Breviata
		Excavata	Metamonada Malawimonas Discoba
	Diaphoretickes		Cryptophyceae
			Centrohelida
			Telonemia
			Haptophyta
		Sar	Cercozoa Foraminifera "Radiolaria"
			Alveolata Stramenopiles
		Archaeplastida	Glaucophyta Rhodophyceae Chloroplastida
	Incertae sedis Eukaryota		Incertae sedis, and table 3

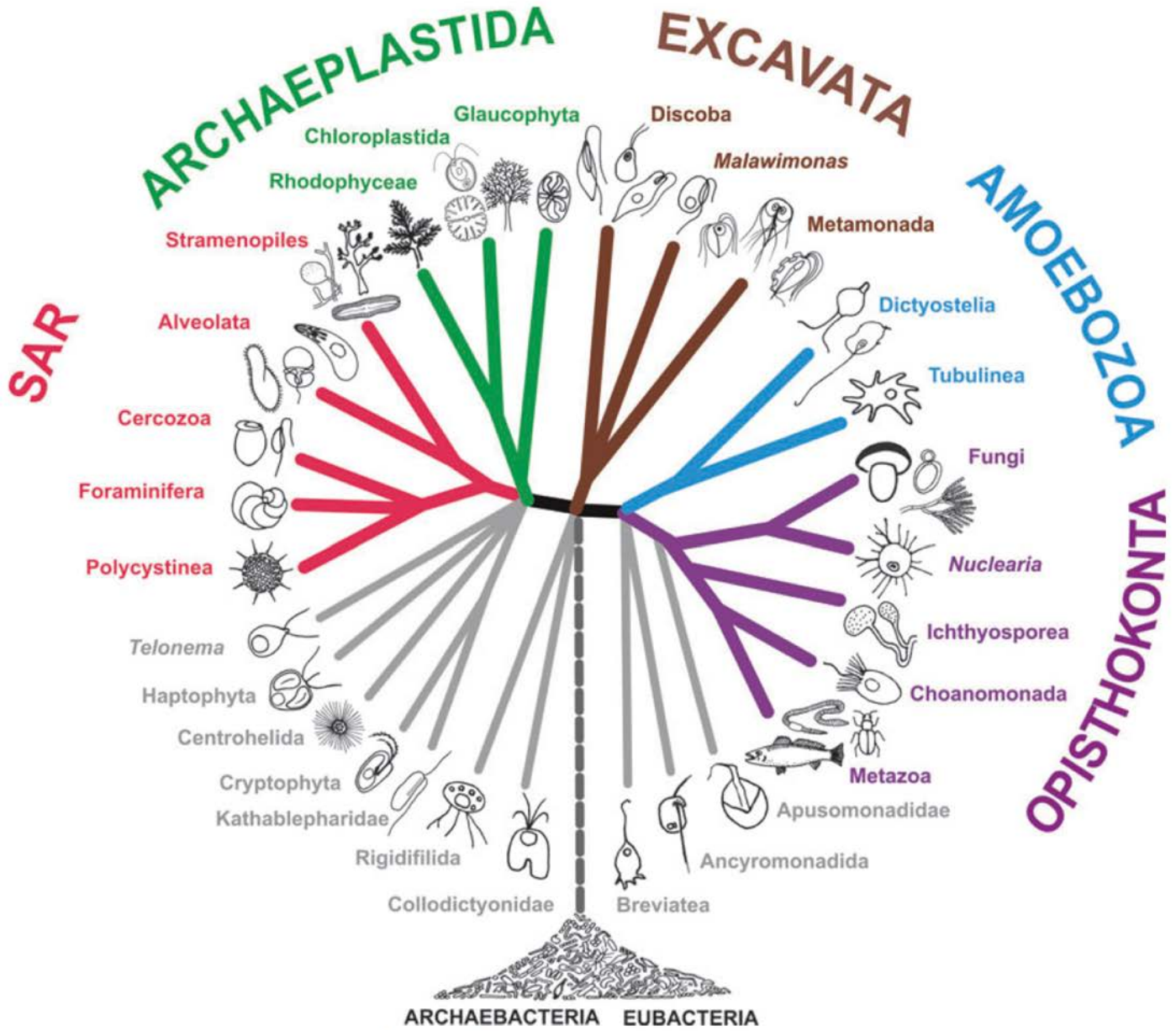


Fig. 1. A view of eukaryote phylogeny reflecting the classification presented herein.

Adl *et al* (2012)**

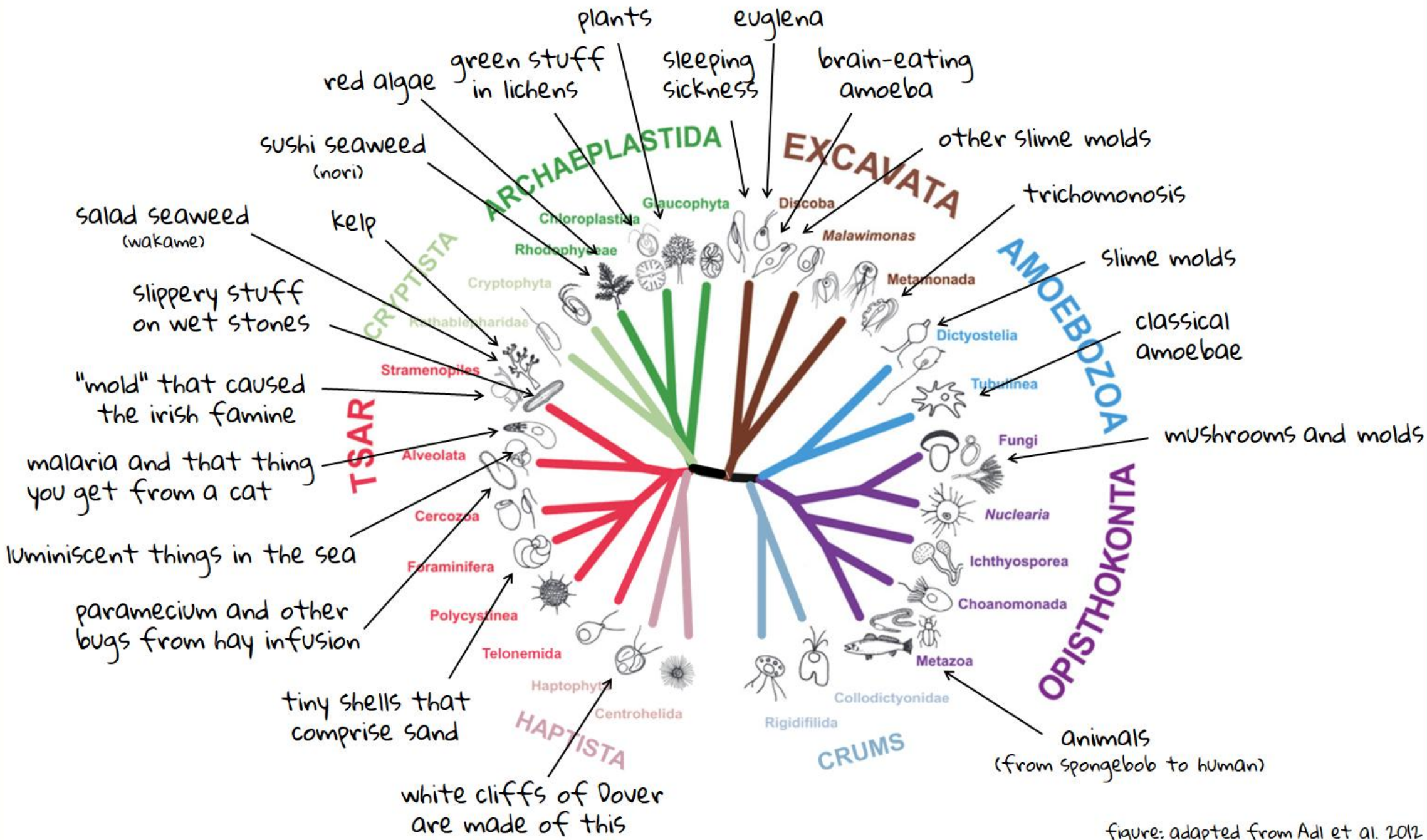
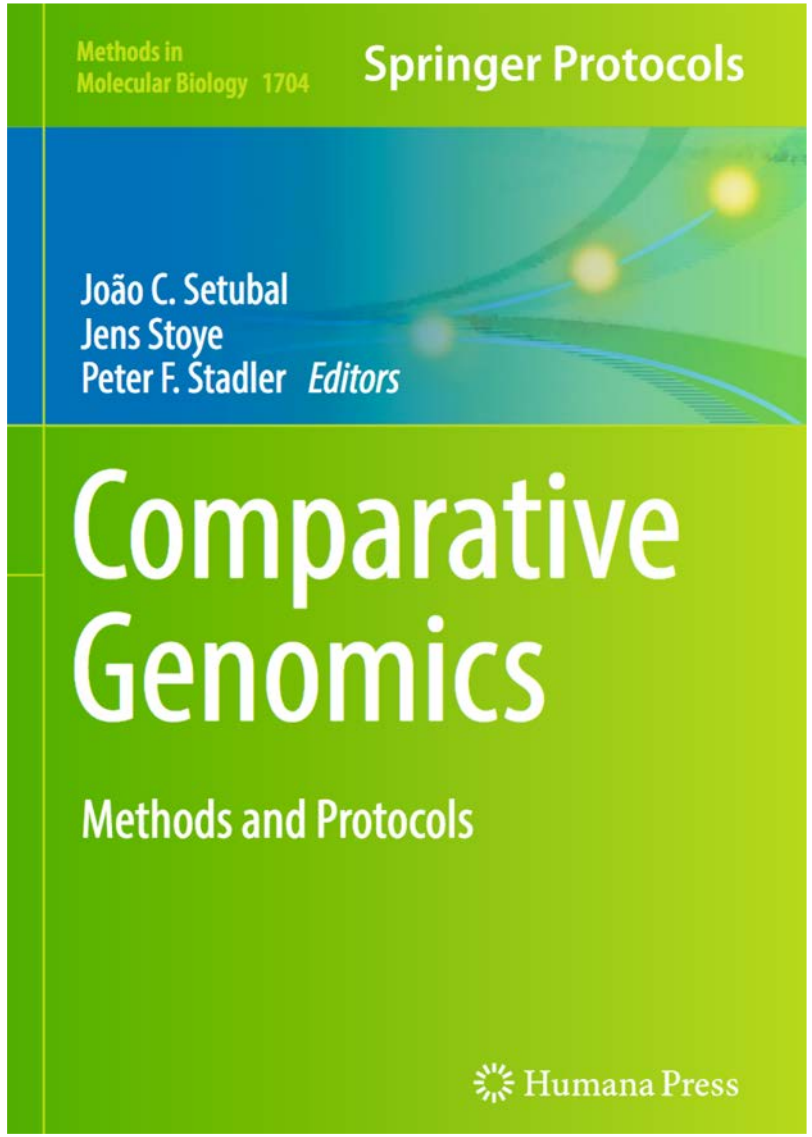


figure: adapted from Adl et al. 2012
text: @euglenaria

How to establish all these relationships?

And what can we reveal from these relationships?

Recommended book and paper



Orthology: definitions, inference, and impact on species phylogeny inference

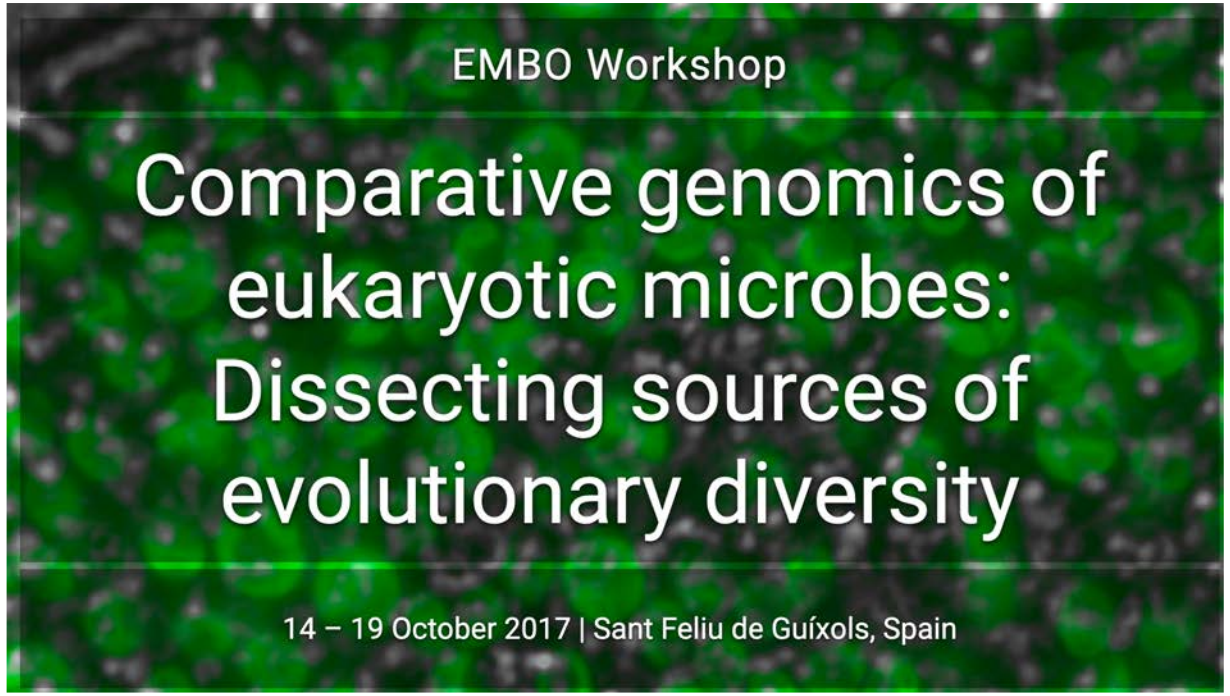
Rosa Fernández[1], Toni Gabaldón[1,2,3,*], Christophe Dessimoz[4,5,6,7,8,*]

Abstract: Orthology is a central concept in evolutionary and comparative genomics, used to relate corresponding genes in different species. In particular, orthologs are needed to infer species trees. In this chapter, we introduce the fundamental concepts of orthology relationships and orthologous groups, including some non-trivial (and thus commonly misunderstood) implications. Next, we review some of the main methods and resources used to identify orthologs. The final part of the chapter discusses the impact of orthology methods on species phylogeny inference, drawing lessons from several recent comparative studies.

https://link.springer.com/book/10.1007%2F978-1-4939-7463-4

https://arxiv.org/abs/1903.04530

Recommended references



Unicellular eukaryotes comprise the overwhelming majority of eukaryotic cellular and genomic diversity, pervading all branches of the eukaryotic tree of life. Recent sequencing efforts have significantly increased the number of unicellular eukaryotes for which genomic/cellular/proteomic data are available.

Metagenomics, transcriptomics, single-cell genomics, and other approaches are being applied to unravel the ecology, physiology, diversity and evolution of microbial eukaryotes and are shedding light on fundamental questions such as the origin of the eukaryotic cell, endosymbiosis, the origin of multicellularity and the evolution of major cellular systems in eukaryotes.

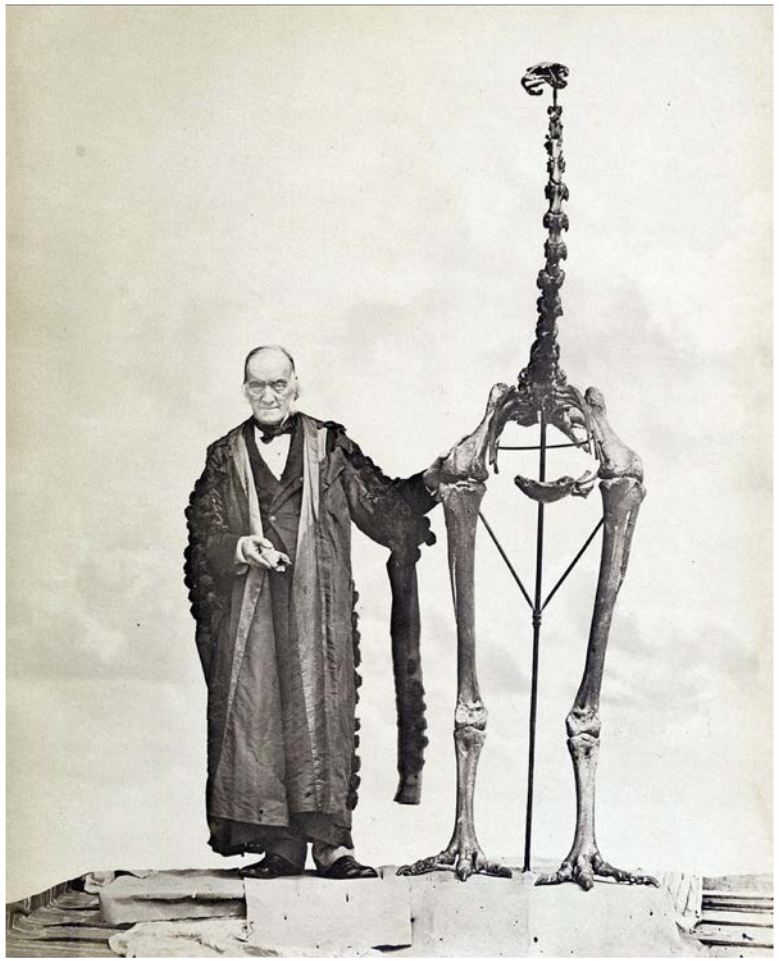
Although there are conferences devoted to genomics of prokaryotes and that of plants and animals, this EMBO Workshop will be the only forum bringing together this diverse community with a range of expertise and which concentrates on microbial eukaryotes.

There is a need, particularly in organisms without large communities, to have a forum where approaches, new methodologies and datasets can be shared to the advancement of this field.

http://meetings.embo.org/event/17-eukaryotes

Homology

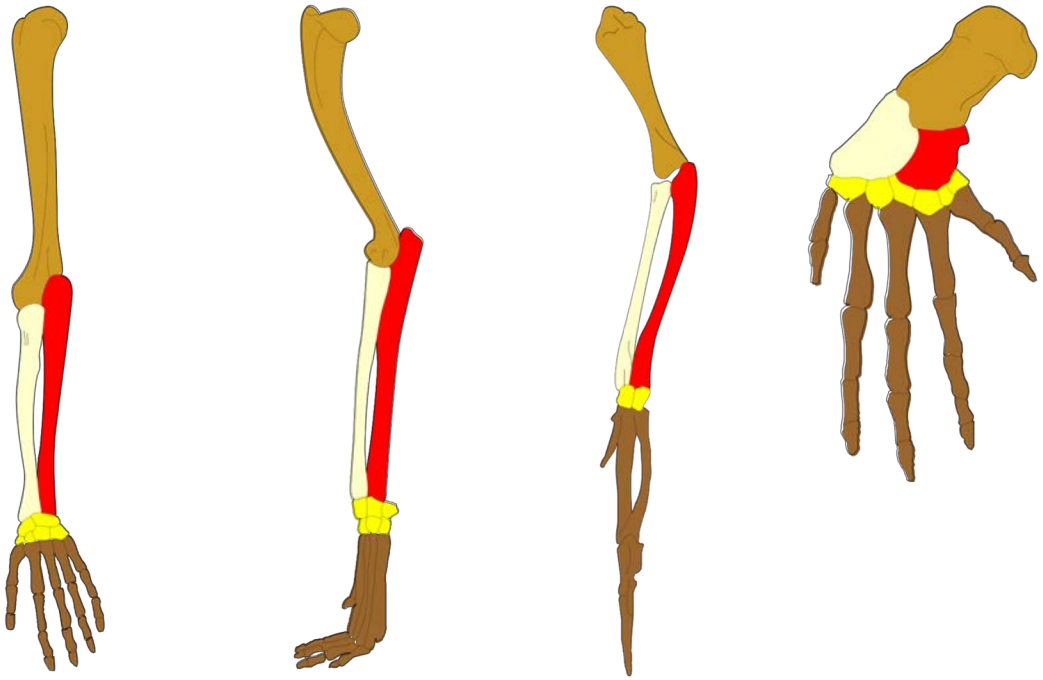
Termed before Darwin's time!



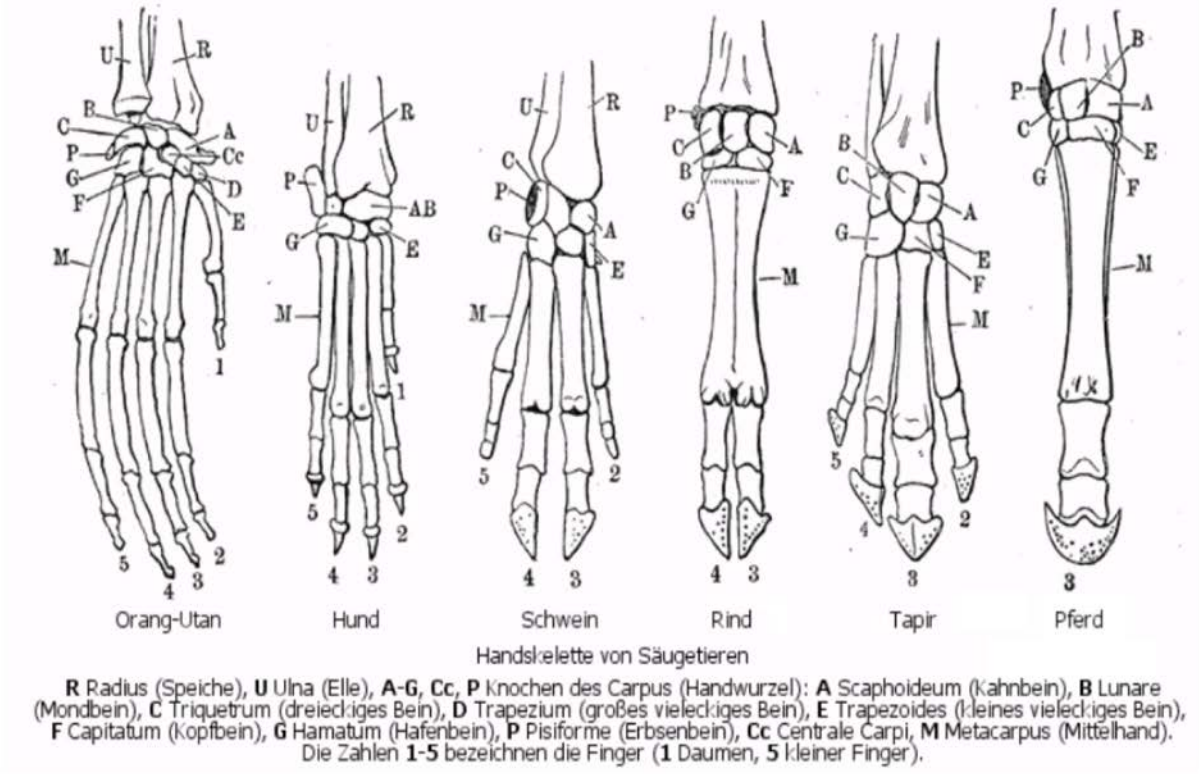
Sir Richard Owen KCB FRS (20 July 1804 – 18 December 1892) was an English biologist, comparative anatomist and paleontologist.

Wiki

Homology



Human Dog Bird Whale



Handskelette von Säugetieren

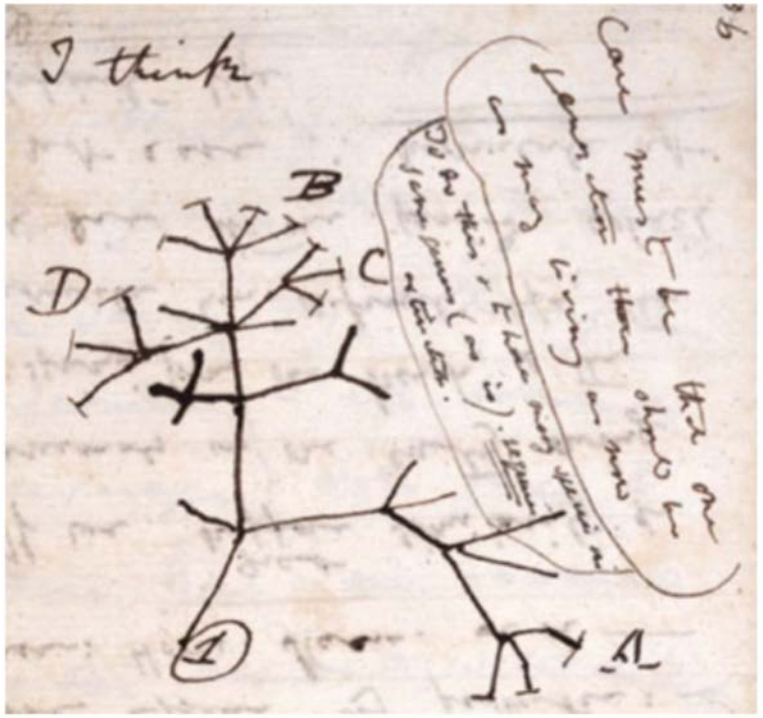
R Radius (Speiche), **U** Ulna (Elle), **A-G**, **Cc**, **P** Knochen des Carpus (Handwurzel): **A** Scaphoideum (Kahnbein), **B** Lunare (Mondbein), **C** Triquetrum (dreieckiges Bein), **D** Trapezium (großes vieleckiges Bein), **E** Trapezoides (kleines vieleckiges Bein), **F** Capitatum (Kopfbein), **G** Hamatum (Hafenbein), **P** Pisiforme (Erbsenbein), **Cc** Centrale Carpi, **M** Metacarpus (Mittelhand). Die Zahlen **1-5** bezeichnen die Finger (**1** Daumen, **5** kleiner Finger).

“the same organ in different animals under every variety of form and function” – Richard Owen

Owen 1843, p.379

https://en.wikipedia.org/wiki/Homology_(biology)

Darwin later reformulated homology as a result of **“descent with modification”**



CHAPTER VI.

DIFFICULTIES ON THEORY.

CHAPTER XIII.

MUTUAL AFFINITIES OF ORGANIC BEINGS: MORPHOLOGY: EMBRYOLOGY: RUDIMENTARY ORGANS.

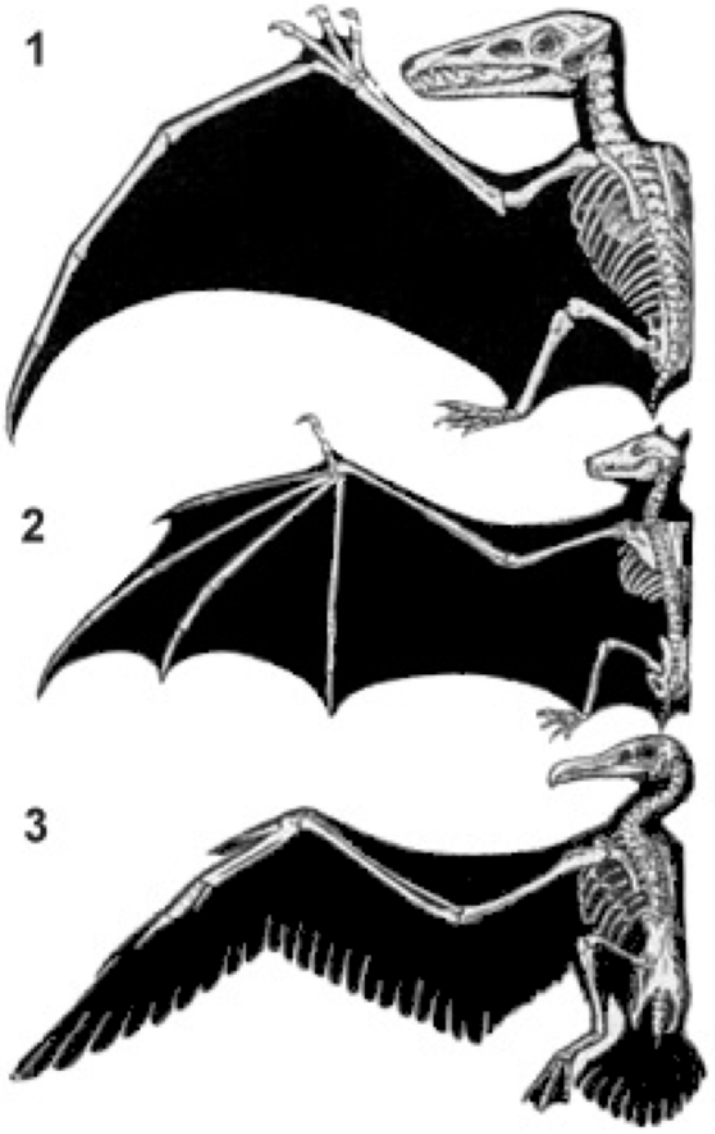
CLASSIFICATION, groups subordinate to groups—Natural system—Rules and difficulties in classification, explained on the theory of descent with modification—Classi-

http://darwin-online.org.uk/

Homology

The wings of pterosaus (1), bats(2) and birds (3) are **analogous** as wings, but **homologous** as forelimbs.

Homologs (any features: genes, trait, morphology) share **ancestry**

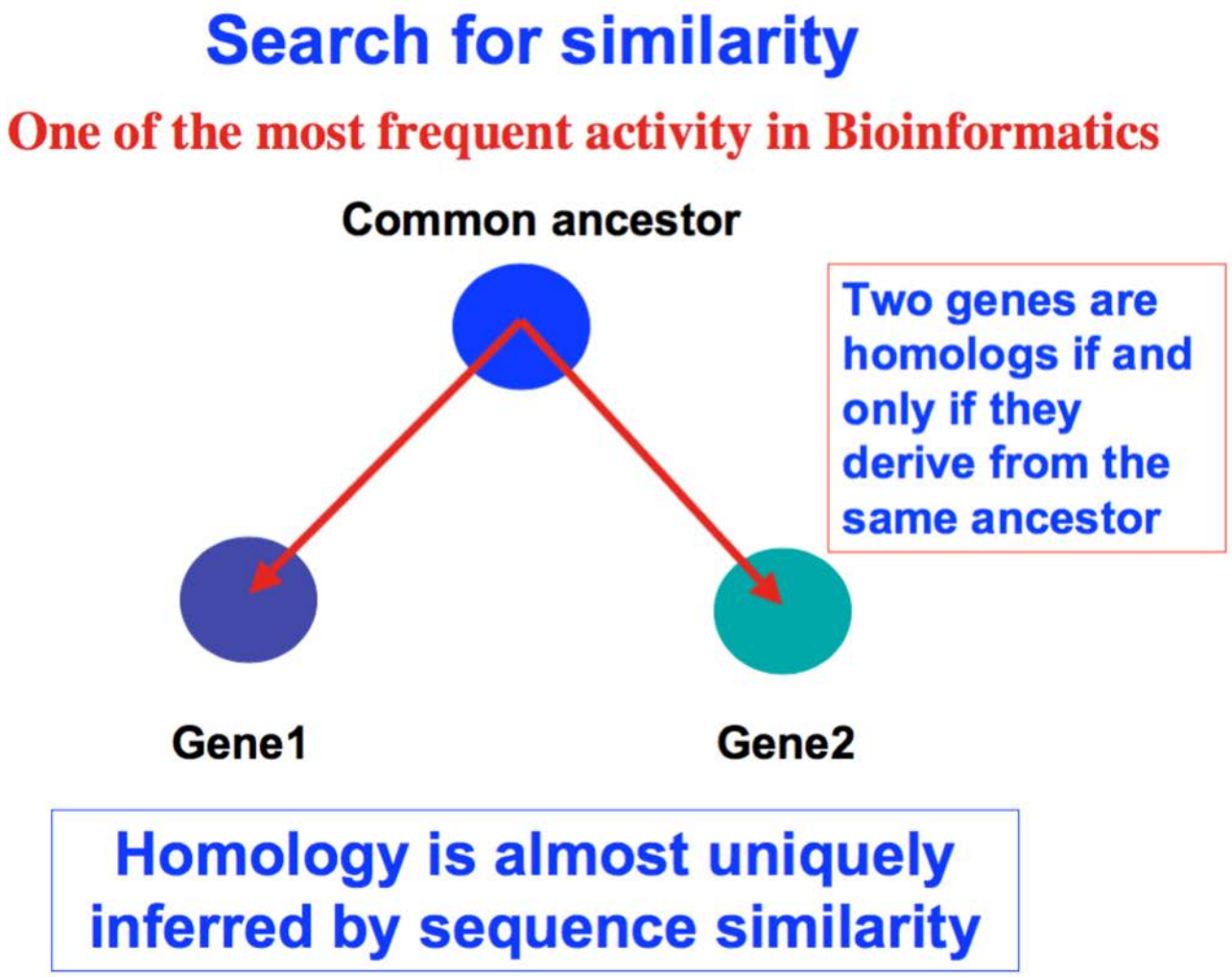


https://en.wikipedia.org/wiki/Homology_(biology)

Homology

Question: How do we establish homology at sequence level?

Search for similarity , collinearity, conservation of morphological characters



Slide of Fred Tekaia

Beware ; why?

~~Significant homology~~

55% married?
45% grandmom?

~~Weak homology~~

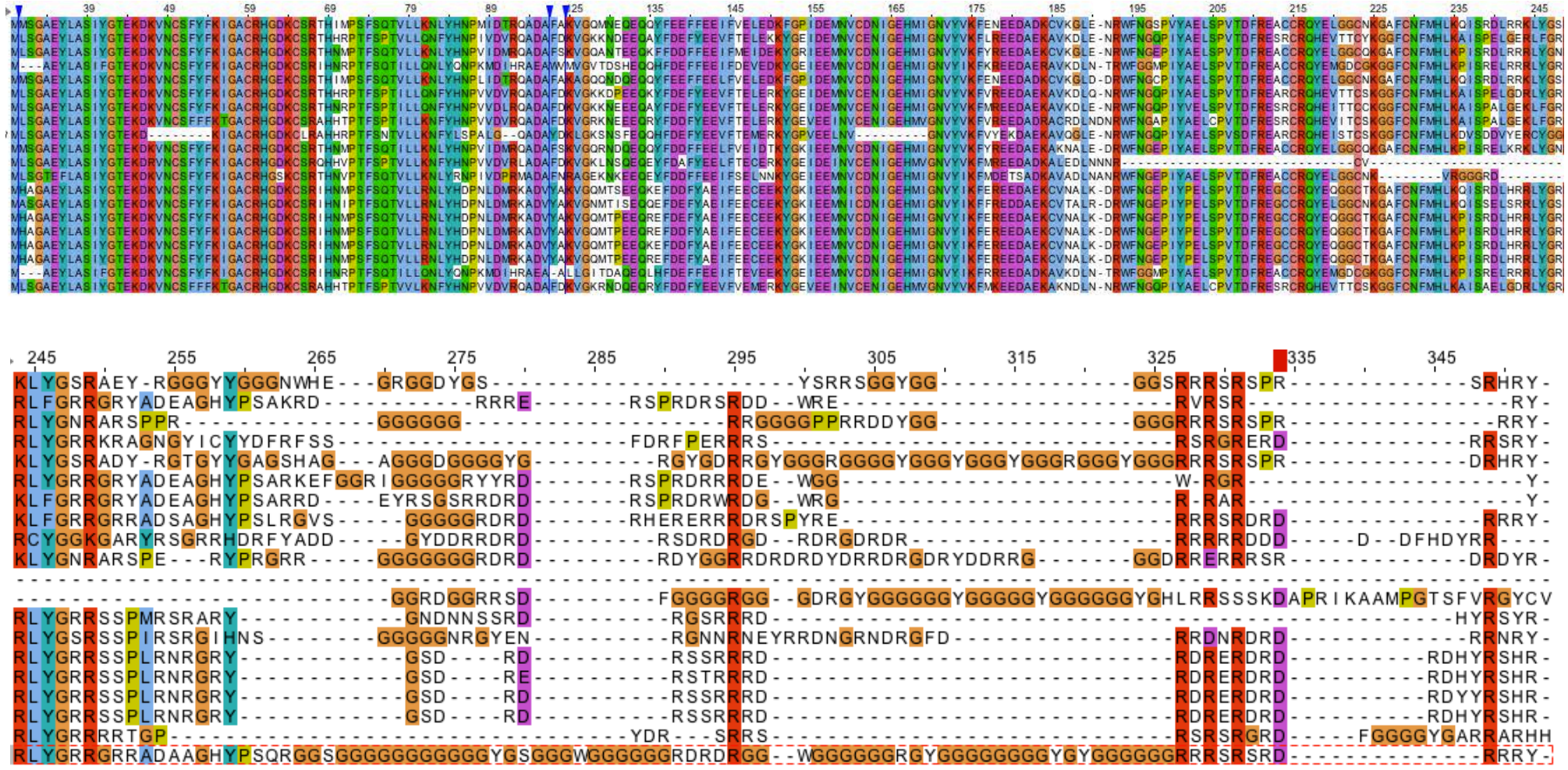
If you think about the meaning of homology,
then it really makes no sense

Significant similarity

Weak similarity

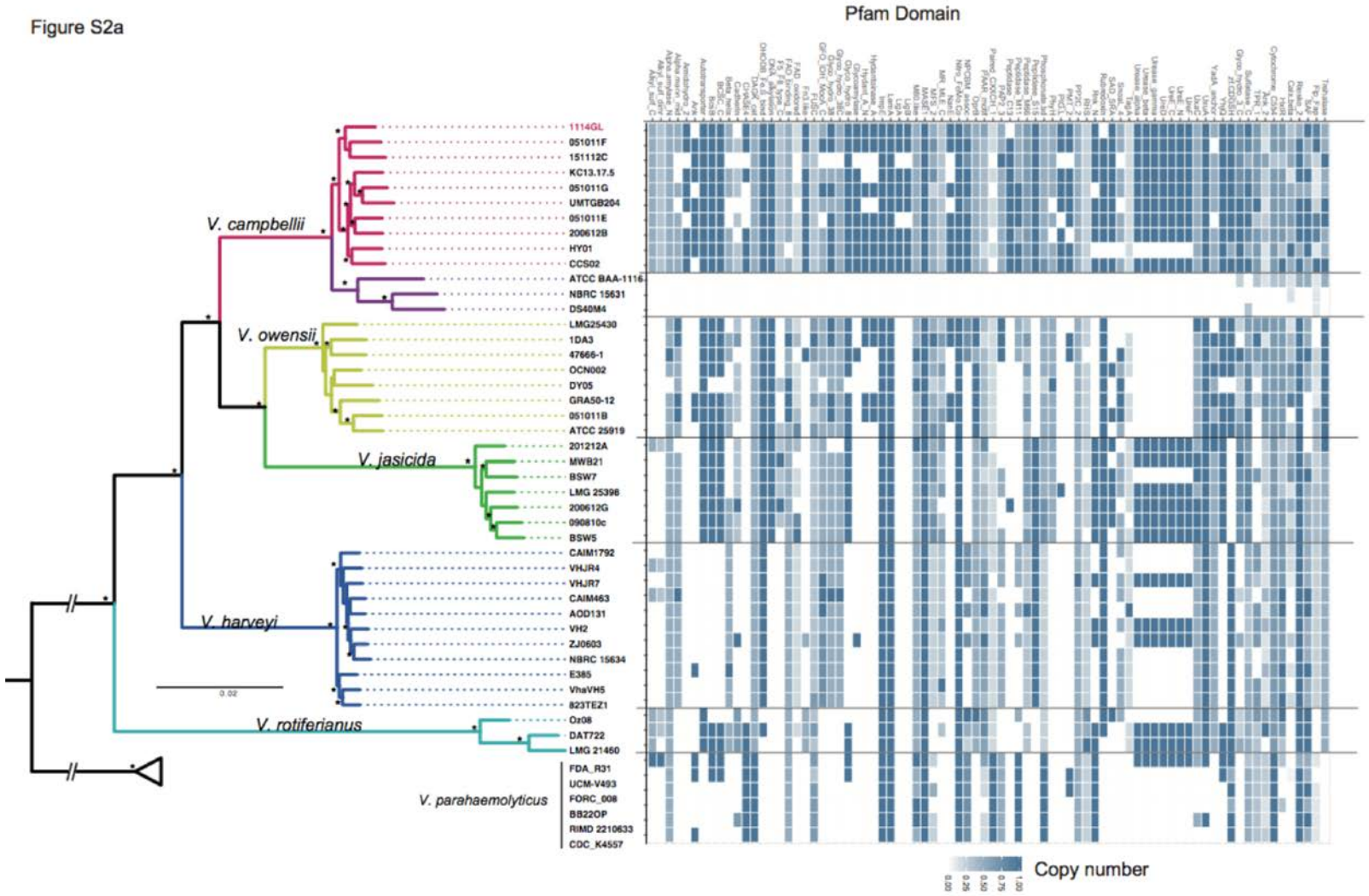
Extension of homology to sequences

Two sequences are homologous if they share the same a common ancestor



Extension of homology to genomes / species

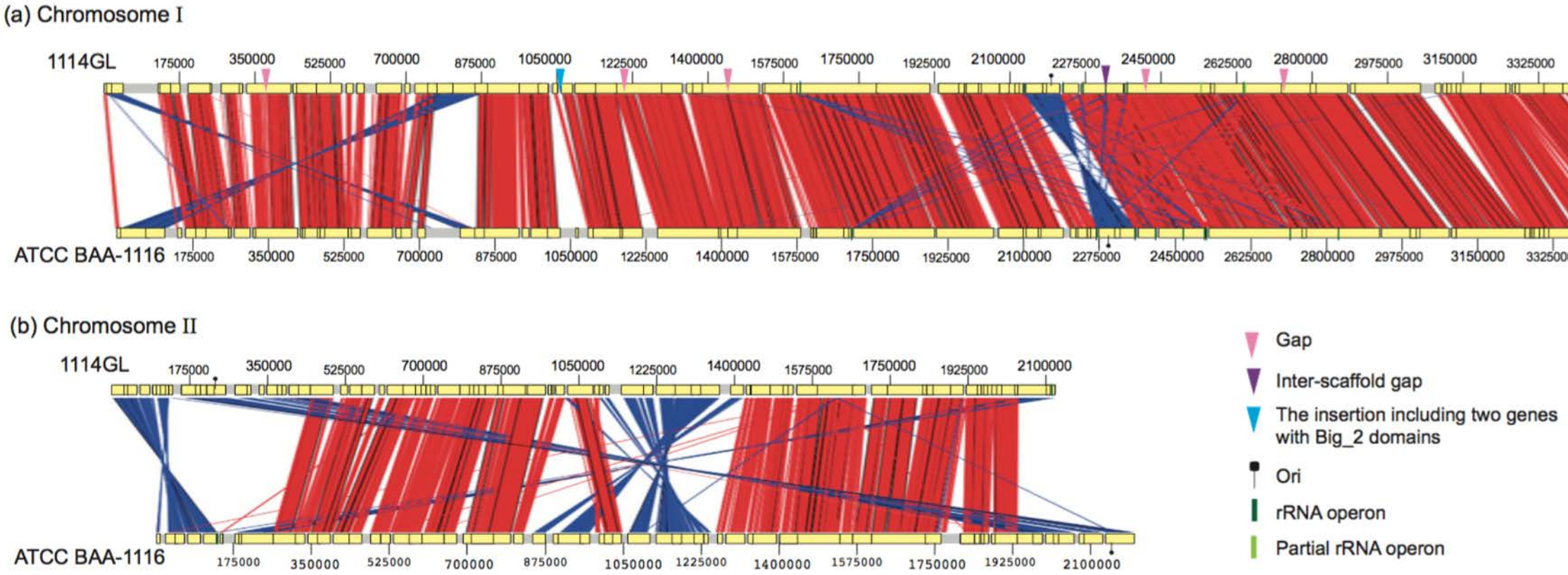
Similarity of individual sequences at different levels (sequence similarity ; domain combinations)



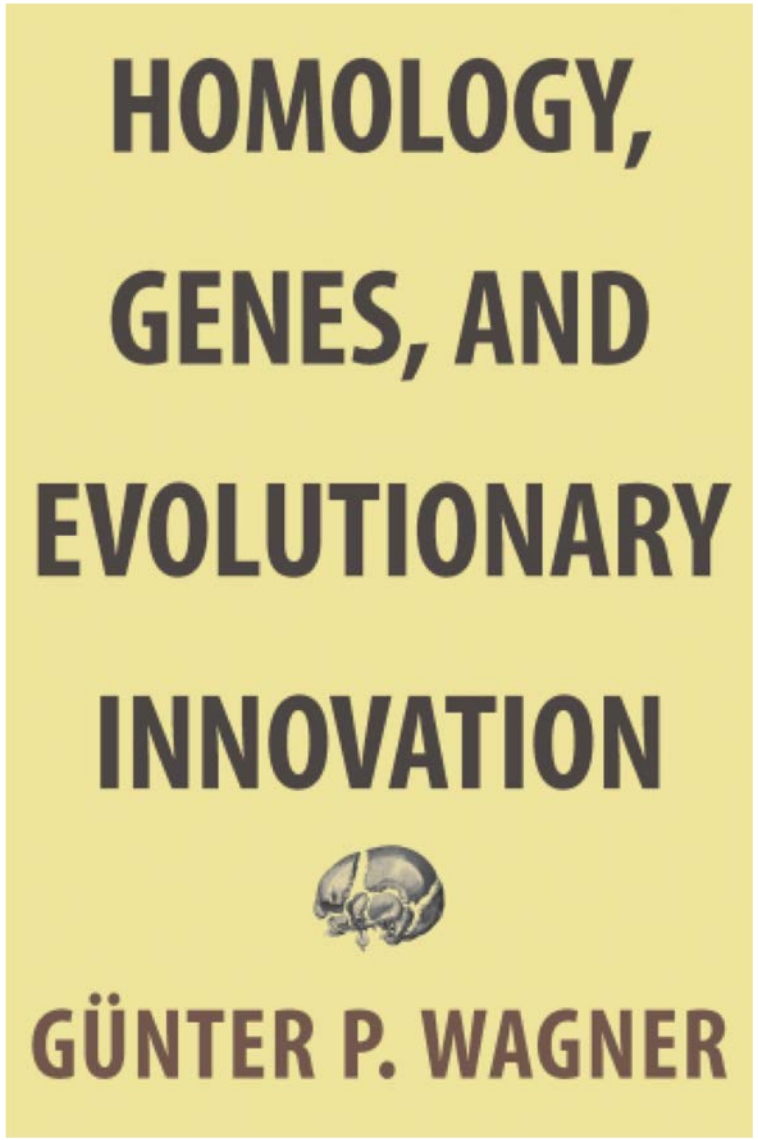
Hueimien Ke

Extension of homology to genomes / species

Similarity of individual features (ordering and rearrangement)



Hueimien Ke



Günter Wagner has thought long and hard about homology in relation to character identity, and in his new book he goes into great detail about why we should use **character identity as the basis for the homology of morphological characters.** For readers of *Systematic Biology*, the book is also a reminder that every **morphological character used in a phylogenetic analysis is a hypothesis of homology, and that great care is needed when deciding whether morphological characters in different organisms are likely to be homologs.**

…He also writes that "This book, although ostensibly about homology, is really a book on evolutionary developmental biology" (p. 3). Wagner argues that "the origin of novel characters and novel body plans is one of the most important but least researched questions in evolutionary biology" (p. 3)….

https://academic.oup.com/sysbio/article/64/2/365/1633524

Why comparative genomics? – A summary

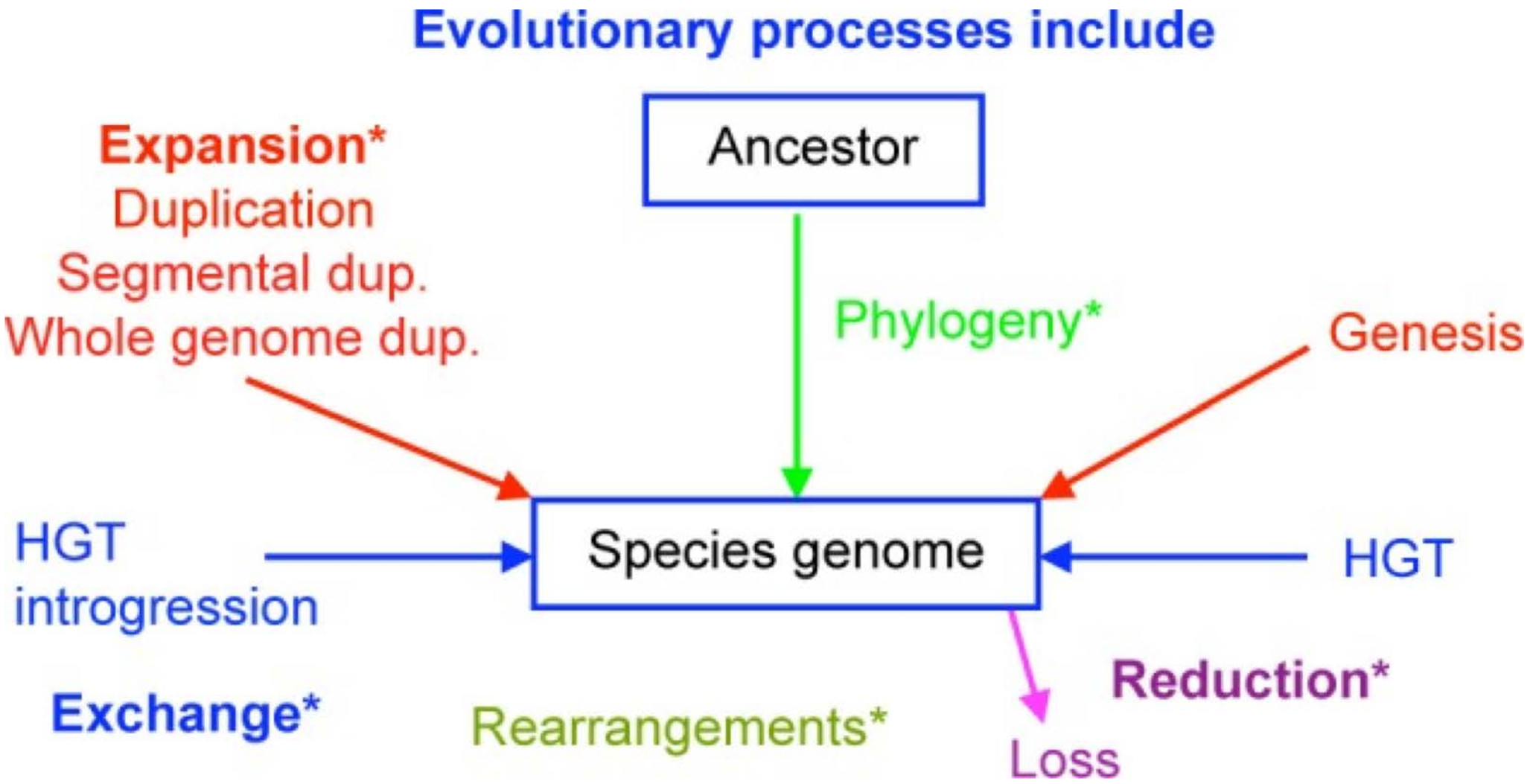
Compare multiple genomes now a norm

Similarity and differences between genomes

Use genomes to study evolution of these species:

- At various resolution (whole genome, chromosomes, regions, genes, base pairs)
- Identify the genomic basis of key phenotypes

Evolution process of a genome

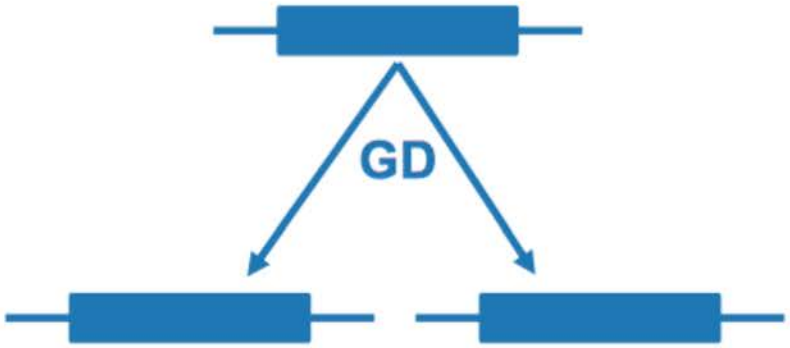


Tekaia (2016)

Sources of gene innovation
(Intuitive as genome gain genes of new functions)

Gene duplication (GD)

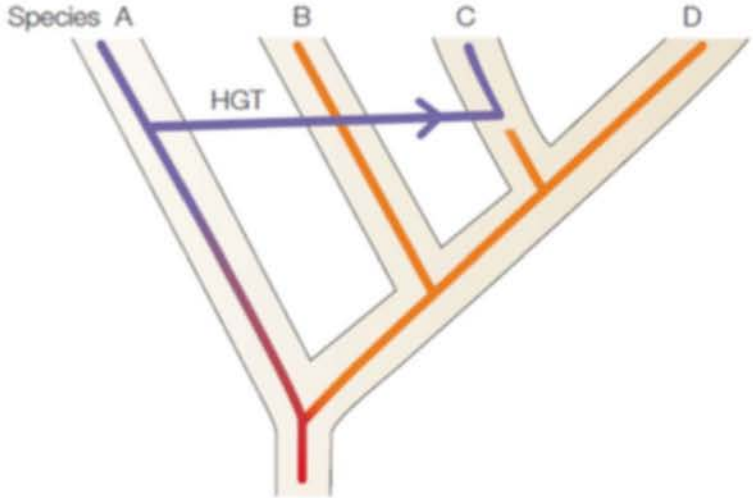
Any duplication of a region of DNA that contains a gene



- Plant organic material decay
- Starch catabolism
- Degradation of host tissues
- Toxin production

Horizontal gene transfer (HGT)

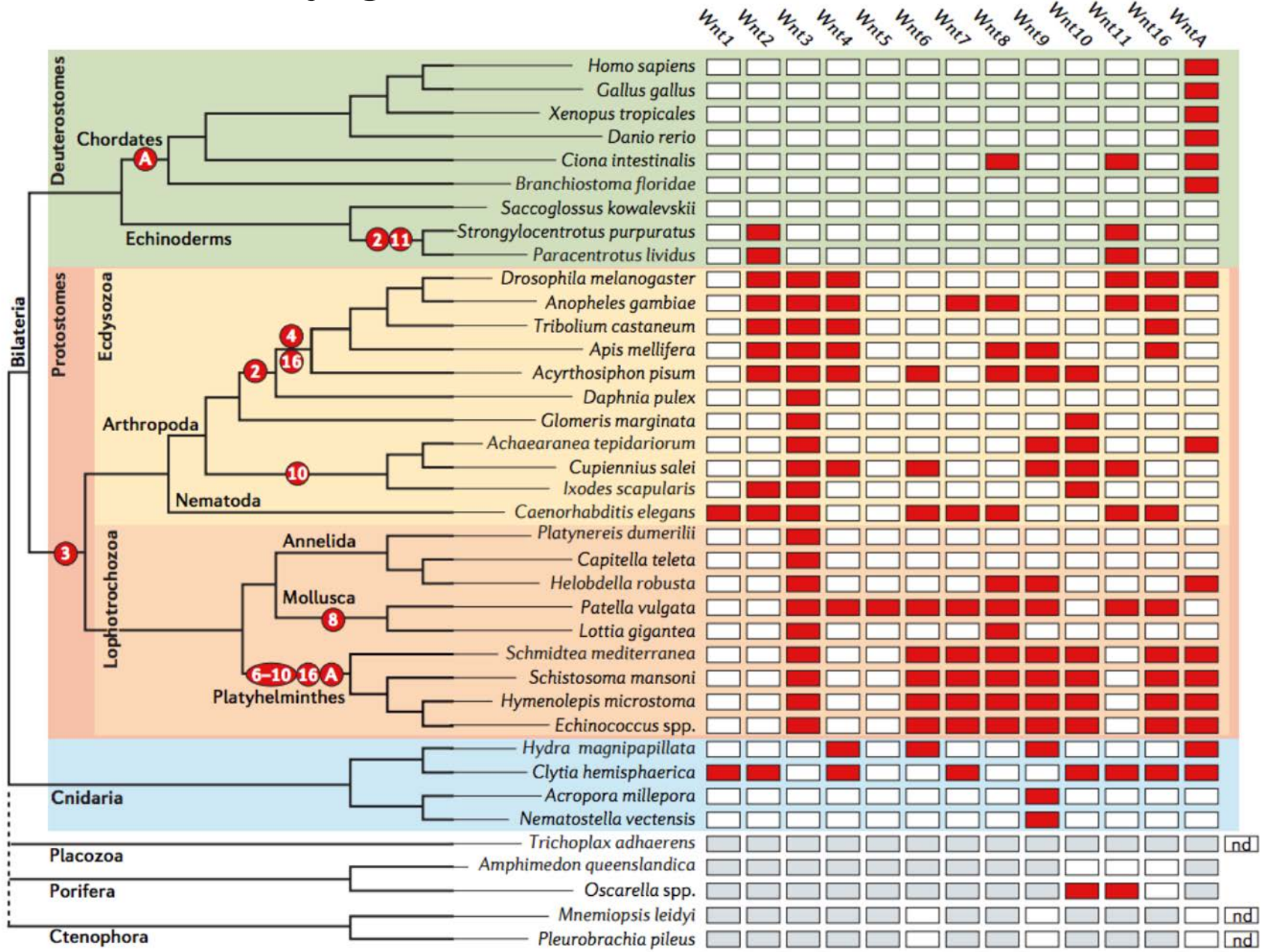
Exchange of genes between organisms other than through reproduction

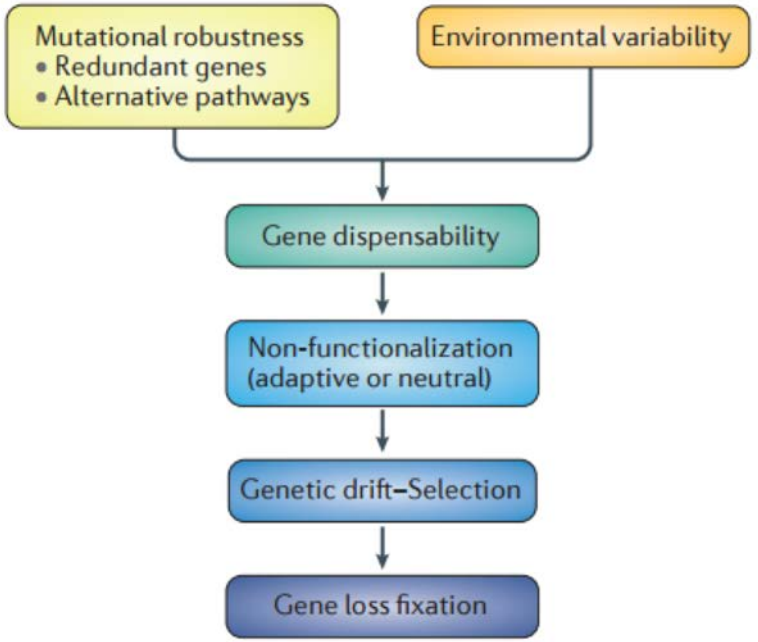


- Xenobiotic catabolism
- Toxin production
- Degradation of plant cell walls
- Wine fermentation

Slides of Antonis Rokas

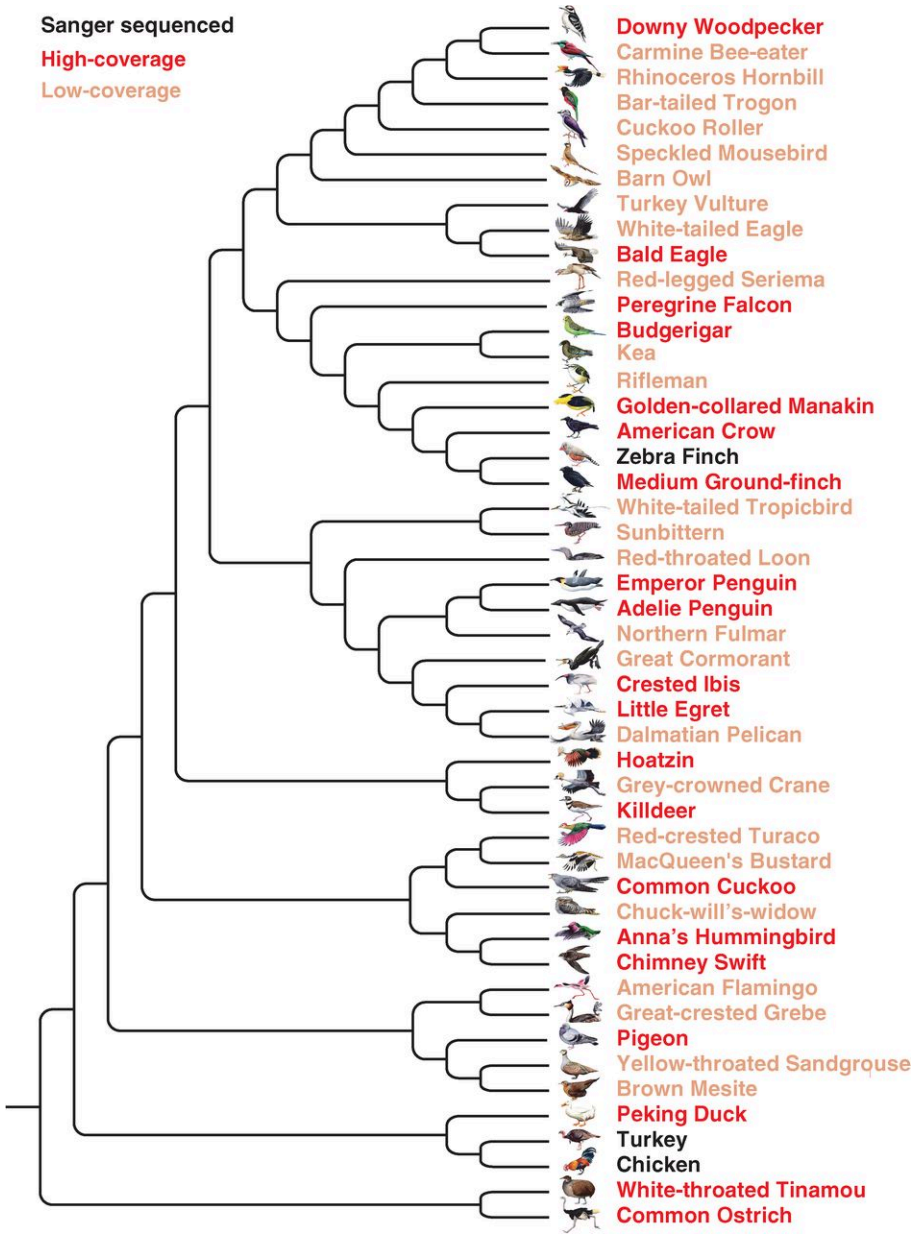
Evolution by gene loss

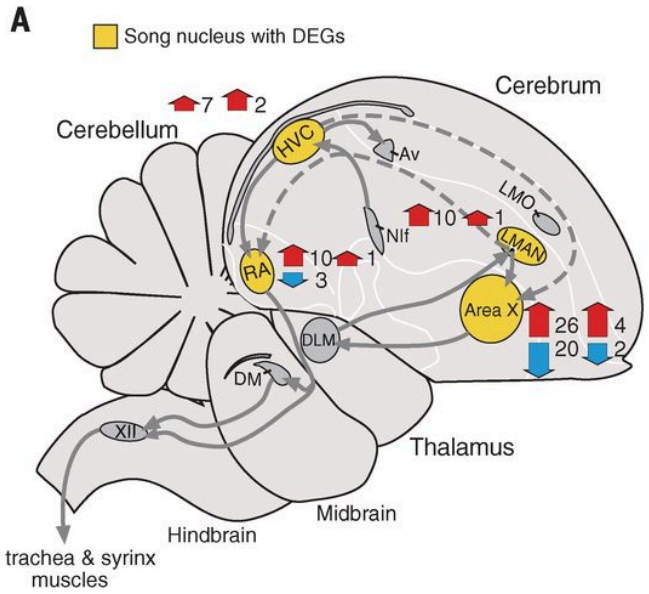


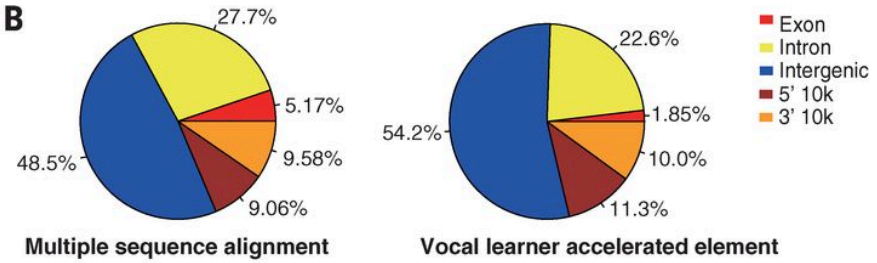


Albalat and Cañestro (2016)

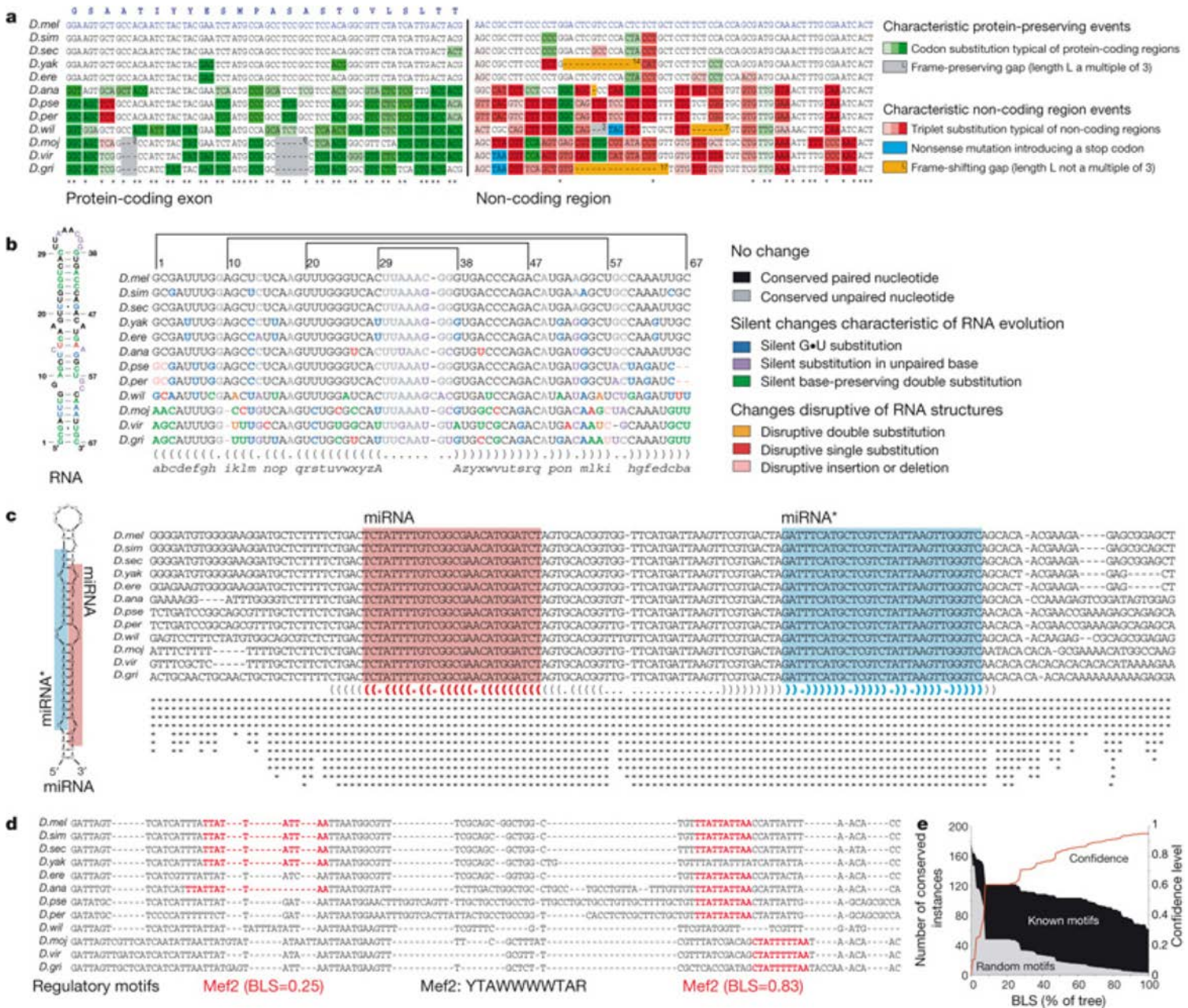
Reveal the evolutionary relationships among species







Link evolutionary processes with function

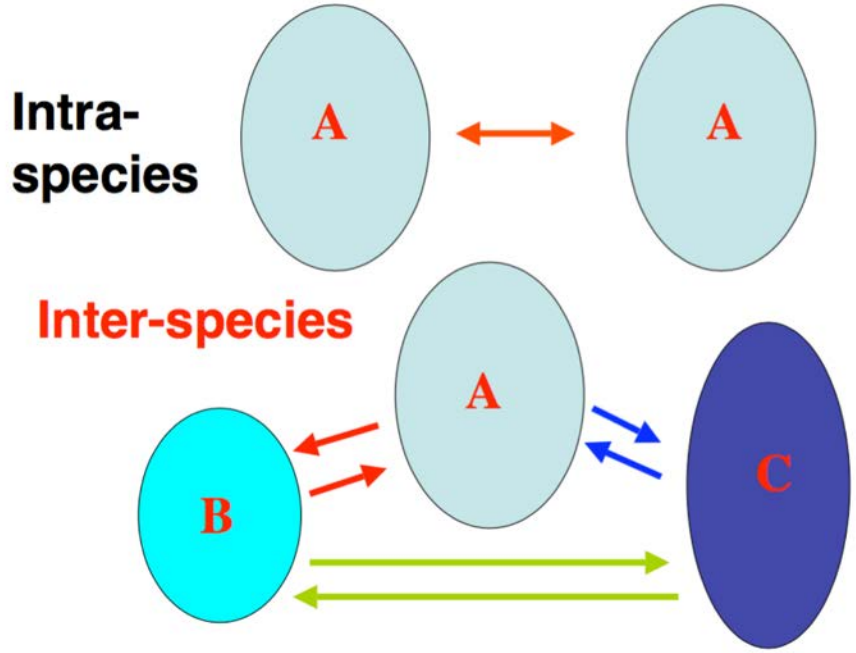


Clark et al (2007)

Comparing genomes

- Alignment of homologous regions
  - Inter-genomic: aligning genomic sequences from different species
  - Intra-genomic aligning genomic sequences from the same species

- Different levels of resolution
  - Comparative mapping (markers)
  - Synteny (~ gene content)
  - Colinearity (gene content + order conservation)
  - DNA-based alignments (base-to-base mapping)

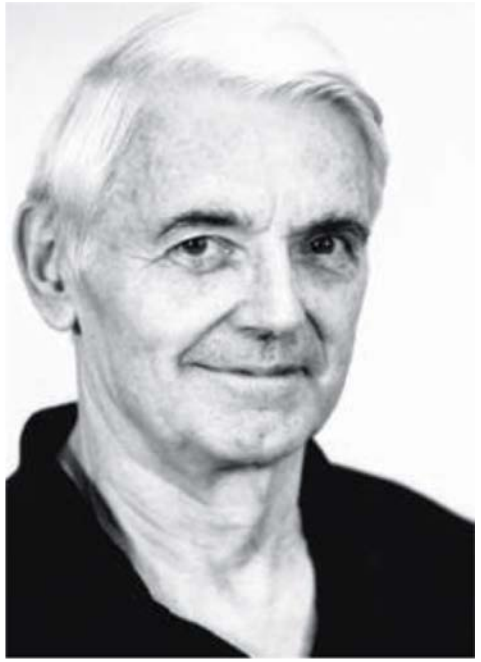


Orthology

Refining *how* homologous genes are related

DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS (1970)

WALTER M. FITCH



1929 - 2011

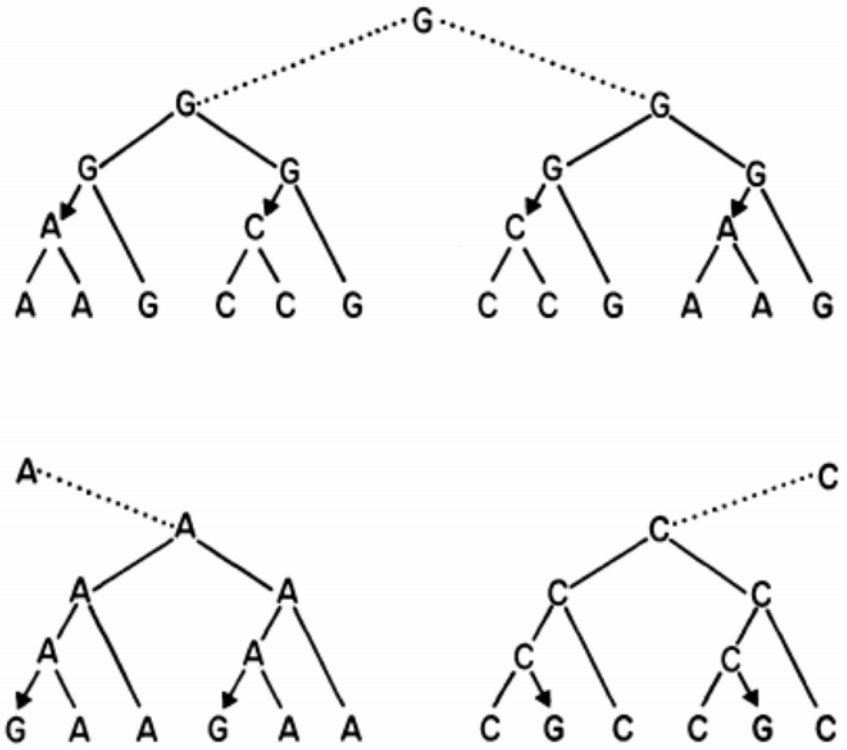


FIG. 1.—Distinguishing convergent from divergent types of nucleotide replacement patterns. Given are two groups of species (related within each group as shown by the solid lines) together with the nucleotide present at a specific position of the gene for each member species as shown at the branch tips. Given also the requirement that the ancestral nucleotide must permit the descendant nucleotides to be obtained in the minimum number of replacements, the ancestral nucleotide of the upper two groups must be set as G, with the required replacements indicated by the arrows. Were one to postulate a common ancestor for the two groups, no new mutations would need to be assumed; hence, this kind of pattern is called the divergent types. The lower two groups are identical except for rearranging the nucleotides at the branch tips, but now, in order to account for descendants in only four nucleotide replacements, the ancestral nucleotide of the lower two groups must be A and C. To postulate a common ancestor for these two groups would require, unlike the upper pair, an additional mutation. This situation shows different ancestral characters apparently converging toward the same descendant character, and hence is called the convergent type. One can calculate the frequency with which one might expect each type to be found in examining a large number of such nucleotide positions and compare that value to what is in fact found for a particular set of proteins. An abnormally large number of either type is evidence favoring that type of relation between the two groups examined.

From homology to orthology

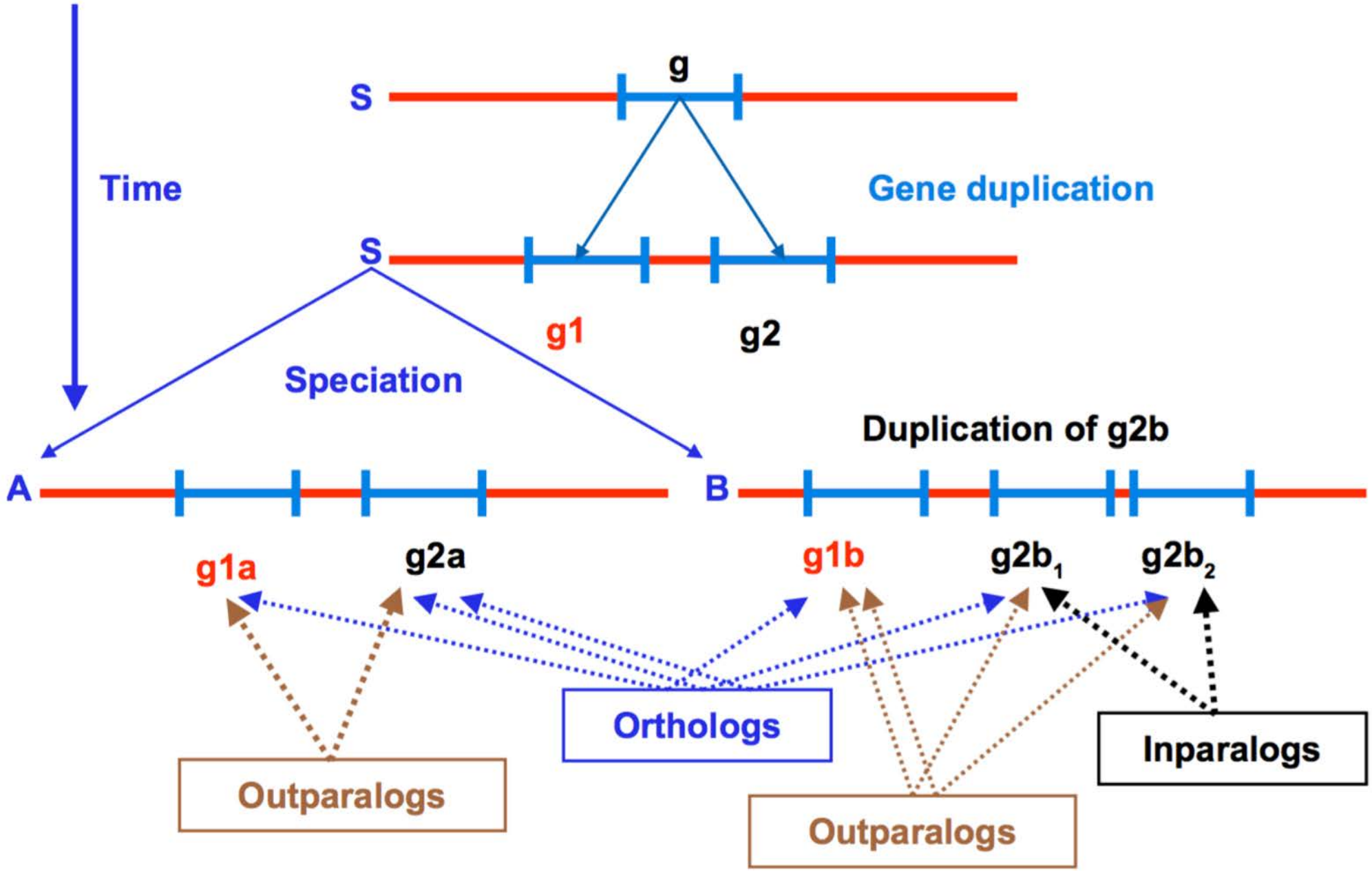
Homologues are sequences derived from a common ancestor...

- What are then orthologues? and paralogues?

Original definition of orthology and paralogy by Walter Fitch (1970, Systematic Zoology 19:99-113):

*"Where the homology is **the result of gene duplication** so that both copies have descended side by side during the history of an organism, (for example, alpha and beta hemoglobin) the genes should be called **paralogous** (para = in parallel).*

*Where the homology is **the result of speciation** so that the history of the gene reflects the history of the species (for example alpha hemoglobin in man and mouse) the genes should be called **orthologous** (ortho = exact)."*


Time
S
g
Gene duplication
S
g1
g2
Speciation
Duplication of g2b
A
B
g1a
g2a
g1b
g2b1
g2b2
Orthologs
Outparalogs
Outparalogs
Inparalogs

Tekaia (2016)

Why is orthology important?

Orthologs detection is of fundamental importance in:

- Reconstruction of the evolution of species and their genomes (Phylogenomics);

- Evolutionary studies of biological systems;

- Annotation of newly sequenced organisms;

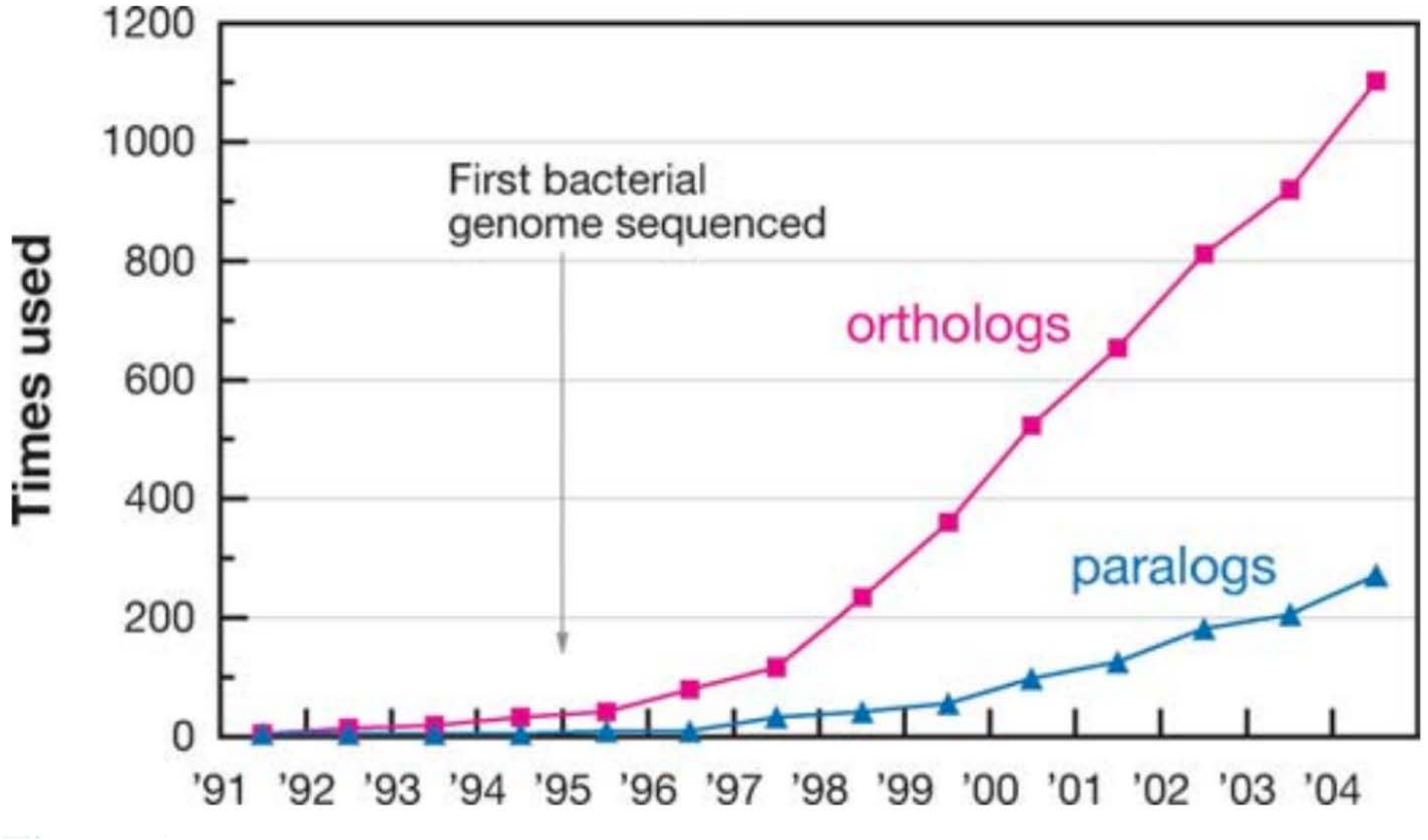
- Functional genomics (transfer of functional annotation predicted on "orthology-function conjecture");

- Gene organization in a given species.

Accurate determination of evolutionary relationships between orthologous gene families is of utmost importance for such goals.

Slide of Fred Tekaia

Usage of “ortholog” and ”paralog”



Koonin (2005)

Corollary

- Orthology definition is purely on evolutionary terms (not functional, not synteny…)
- There is no limit on the number of orthologs or paralogs that a given gene can have (when more than one ortholog exist, there is nothing such as "*the true ortholog*")
- Many-to-Many orthology relationships do exist (co-orthology)
- No limit on how ancient/recent is the ancestral relationship of orthologs and paralogs
- Orthology is non-transitive (as opposed to homology)

More precise definitions

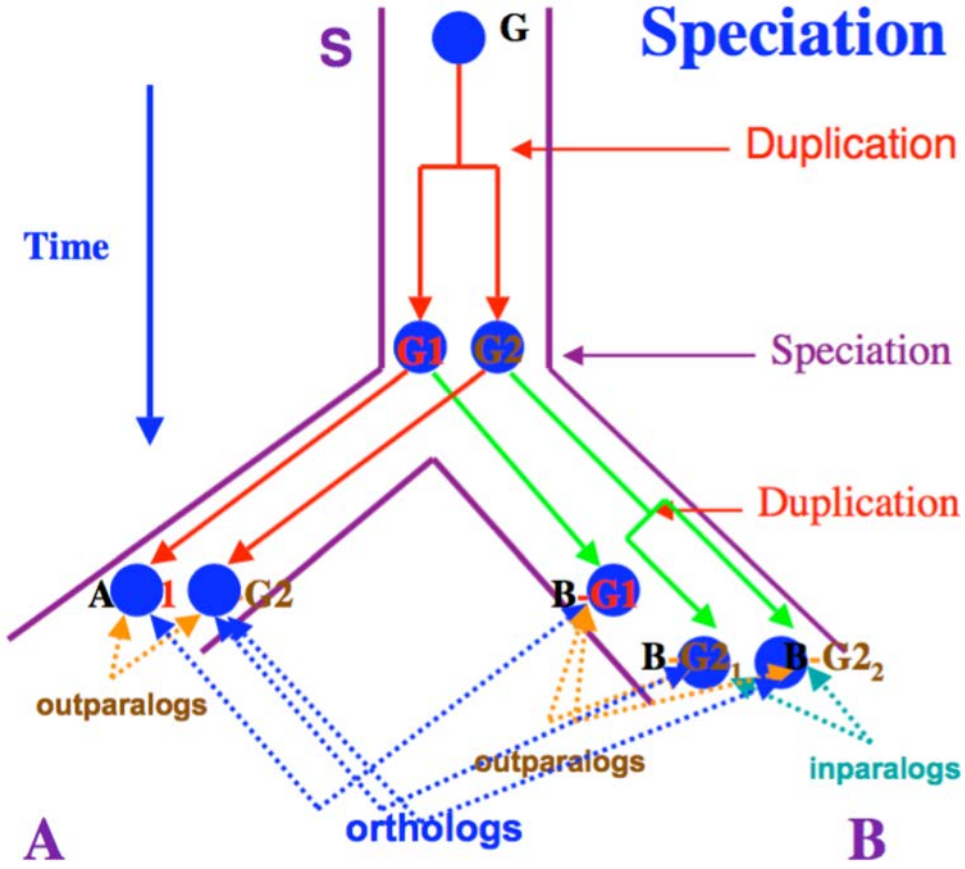


Table 1 Homology: terms and definitions

Homologs	**Genes sharing a common origin**
Orthologs	Genes originating from a single ancestral gene in the last common ancestor of the compared genomes.
Pseudoorthologs	Genes that actually are paralogs but appear to be orthologous due to differential, lineage-specific gene loss.
Xenologs	Homologous genes acquired via XGD by one or both of the compared species but appearing to be orthologous in pairwise genome comparisons.
Co-orthologs	Two or more genes in one lineage that are, collectively, orthologous to one or more genes in another lineage due to a lineage-specific duplication(s). Members of a co-orthologous gene set are inparalogs relative to the respective speciation event.
Paralogs	**Genes related by duplication**
Inparalogs (symparalogs)	Paralogous genes resulting from a lineage-specific duplication(s) subsequent to a given speciation event (defined only relative to a speciation event, no absolute meaning).
Outparalogs (alloparalogs)	Paralogous genes resulting from a duplication(s) preceding a given speciation event (defined only relative to a speciation event, no absolute meaning).
Pseudoparalogs	Homologous genes that come out as paralogs in a single-genome analysis but actually ended up in the given genome as a result of a combination of vertical inheritance and HGT.

Importance of assigning correct orthology

Important implications for phylogeny: only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions)

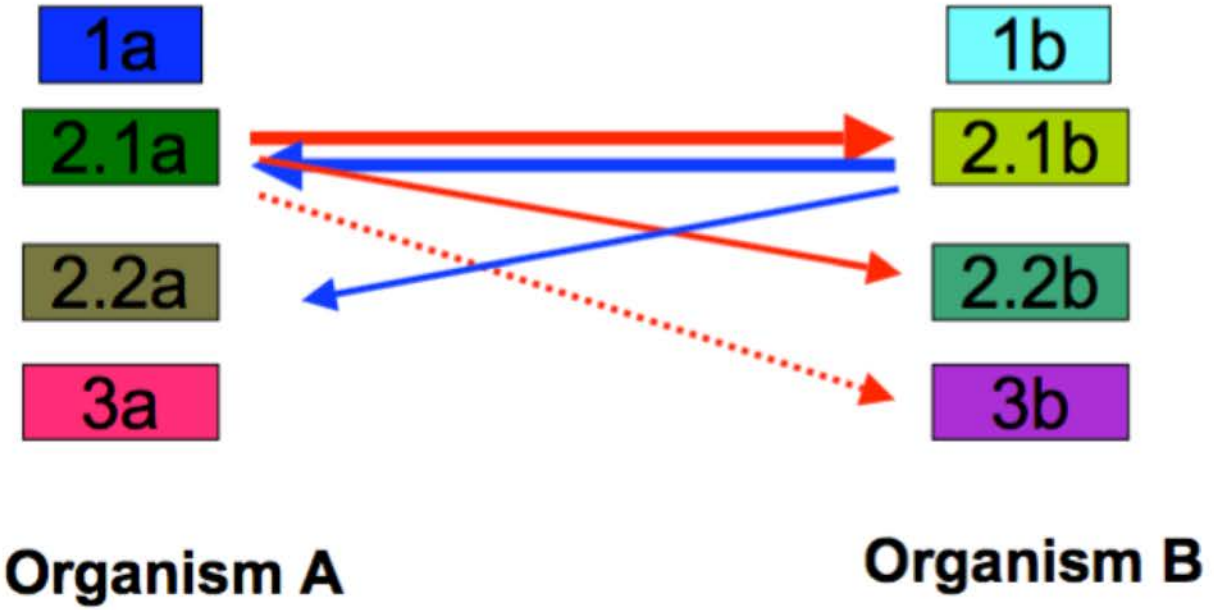
The most exact way of **comparing two (or more) genomes** in terms of their gene content. Necessary to uncover how genomes evolve.

Implications for **functional inference**: orthologs, as compared to paralogs, are more likely to share the same function

Ortholog inference methods

How to detect orthologous genes?

- The most intuitive way: **Best Reciprocal Hit (RBH)**

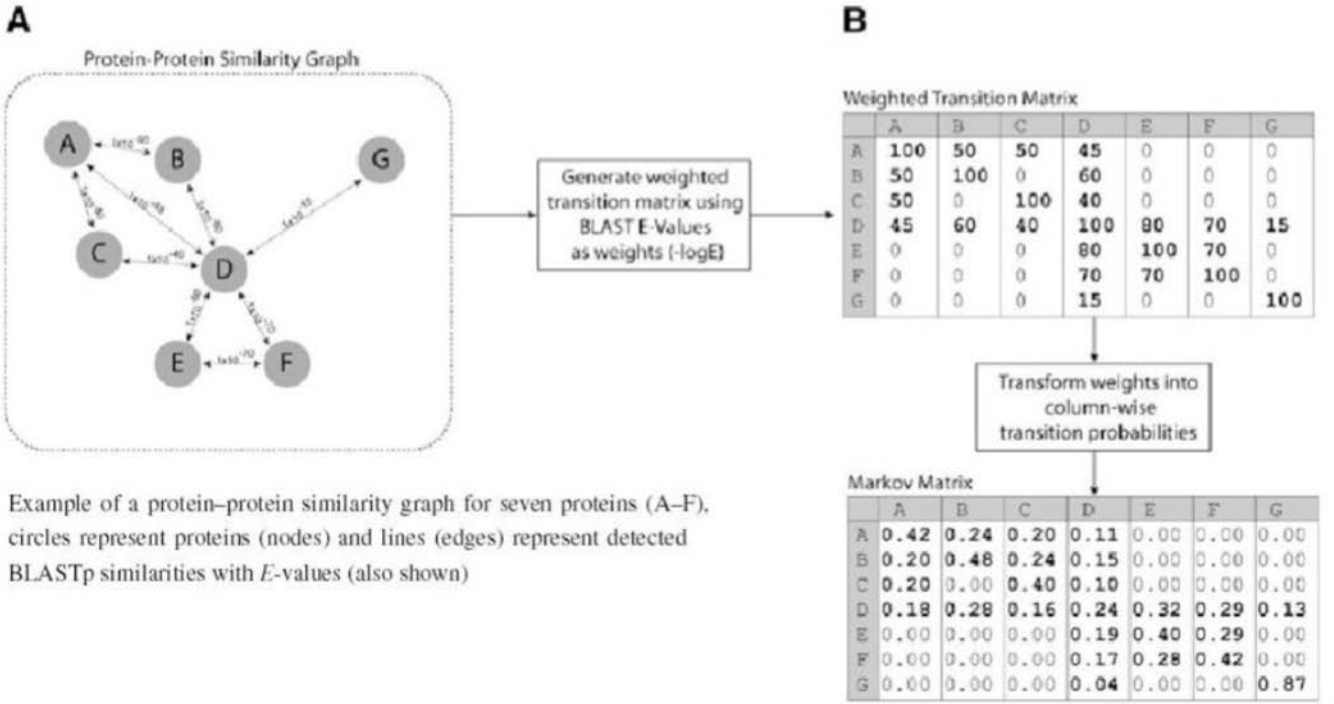


Slide of Fred Tekaia

Sequence by clustering

mcl: The Markov Cluster Algorithm

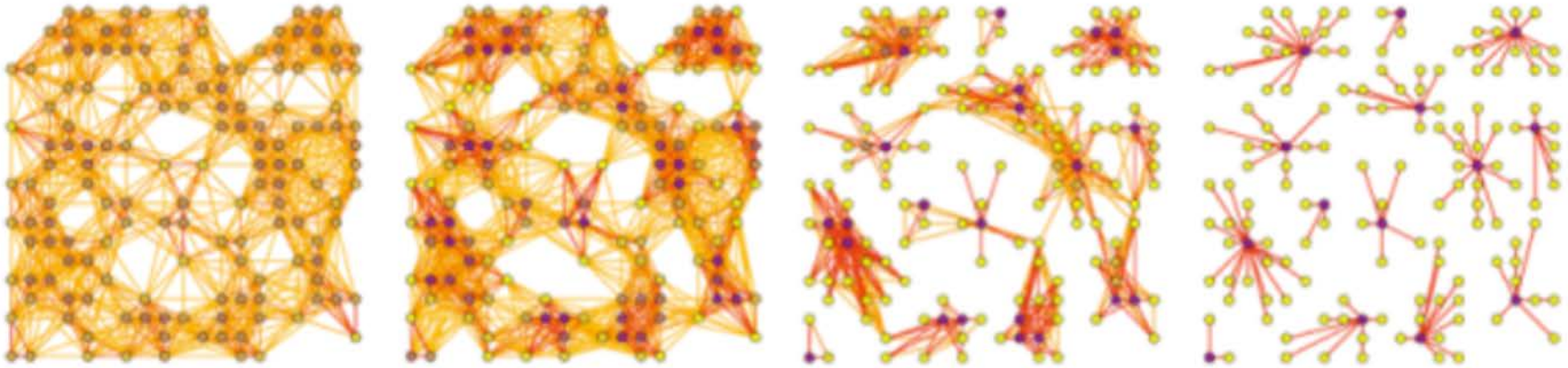
http://micans.org/mcl/ (Stijn Van Dongen)



Example of a protein–protein similarity graph for seven proteins (A–F), circles represent proteins (nodes) and lines (edges) represent detected BLASTp similarities with *E*-values (also shown)

Weighted transition matrix and associated column stochastic Markov matrix for the seven proteins shown in (A).

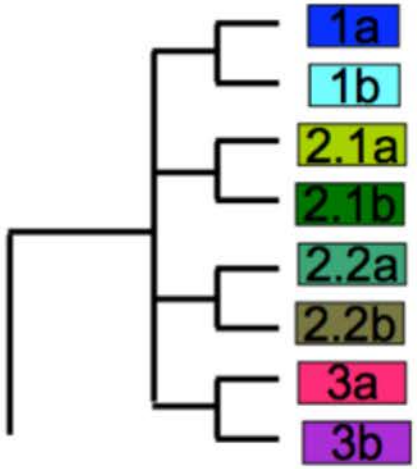
Produce clusters (gene families) using different inflation parameter



Slide of Fred Tekaia

How to detect orthologous genes?

- more rigorous: make a phylogenetic tree of the gene family



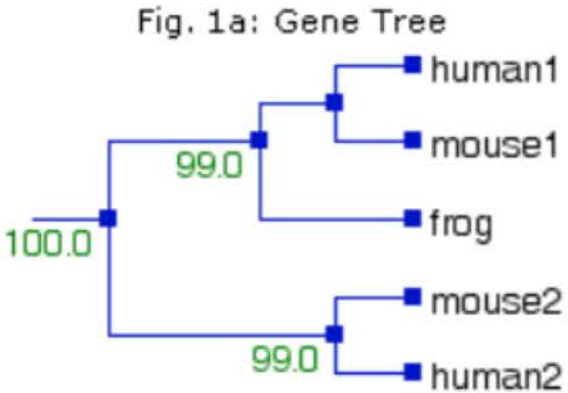
- more rigorous: look at synteny conservation

--> In fact inferring orthology is much more complicated particularly when considering more than 2 genomes!

Slide of Fred Tekaia

Tree reconciliation

Detection of speciation and duplication events using a species tree and gene family tree



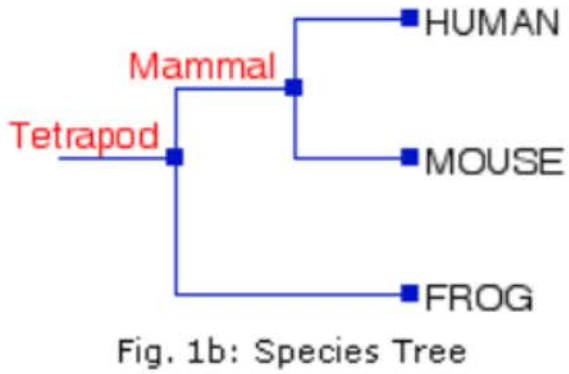
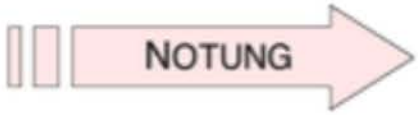


Fig. 1b: Species Tree



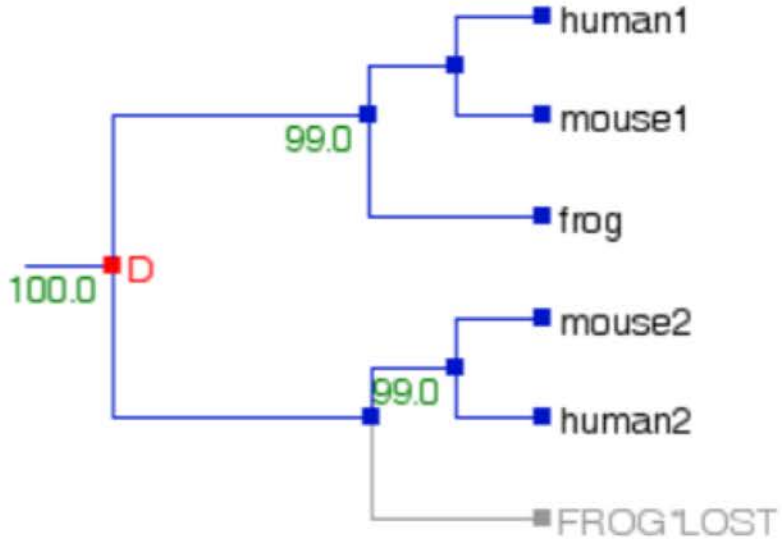
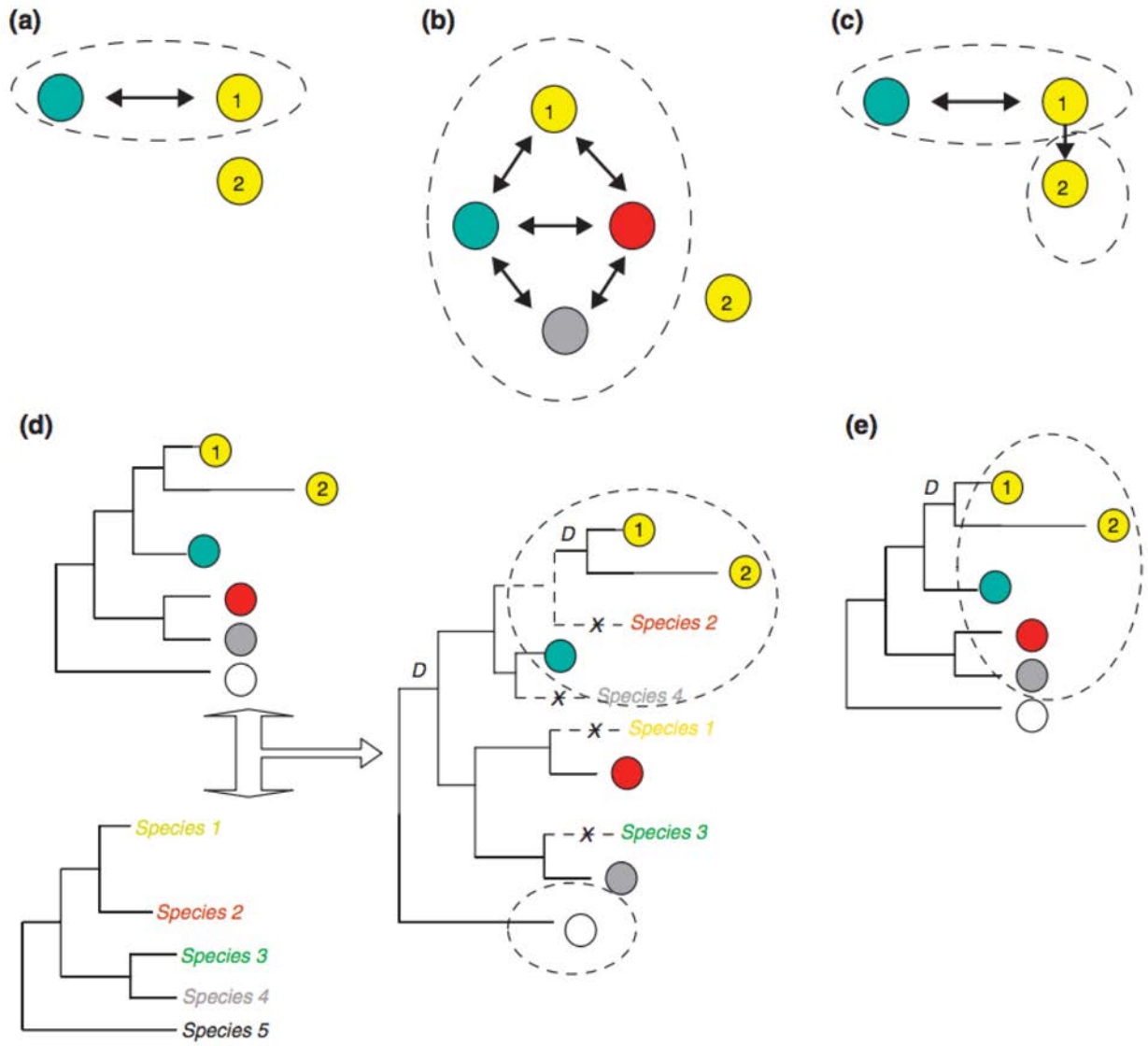


Fig. 1c: Reconciled Gene Tree

Stolzer et al (2012)

Orthology prediction methods



a) Best bidirectional hits
b) COG, MCL-clustering approach
c) InParanoid
d) Tree reconciliation
e) Species-overlap (PhylomeDB)

Gebaldon (2008)

Methods

Similarity

Rely on genome comparisons and clustering of highly similar genes to identify orthologous groups **(suitable for large genome datasets)**

Phylogeny

use candidate gene families determined by similarity and then rely on the reconciliation of the phylogeny of these genes with their corresponding species phylogeny to determine the subset of orthologs

(Good and more interpretable for small set of genomes)

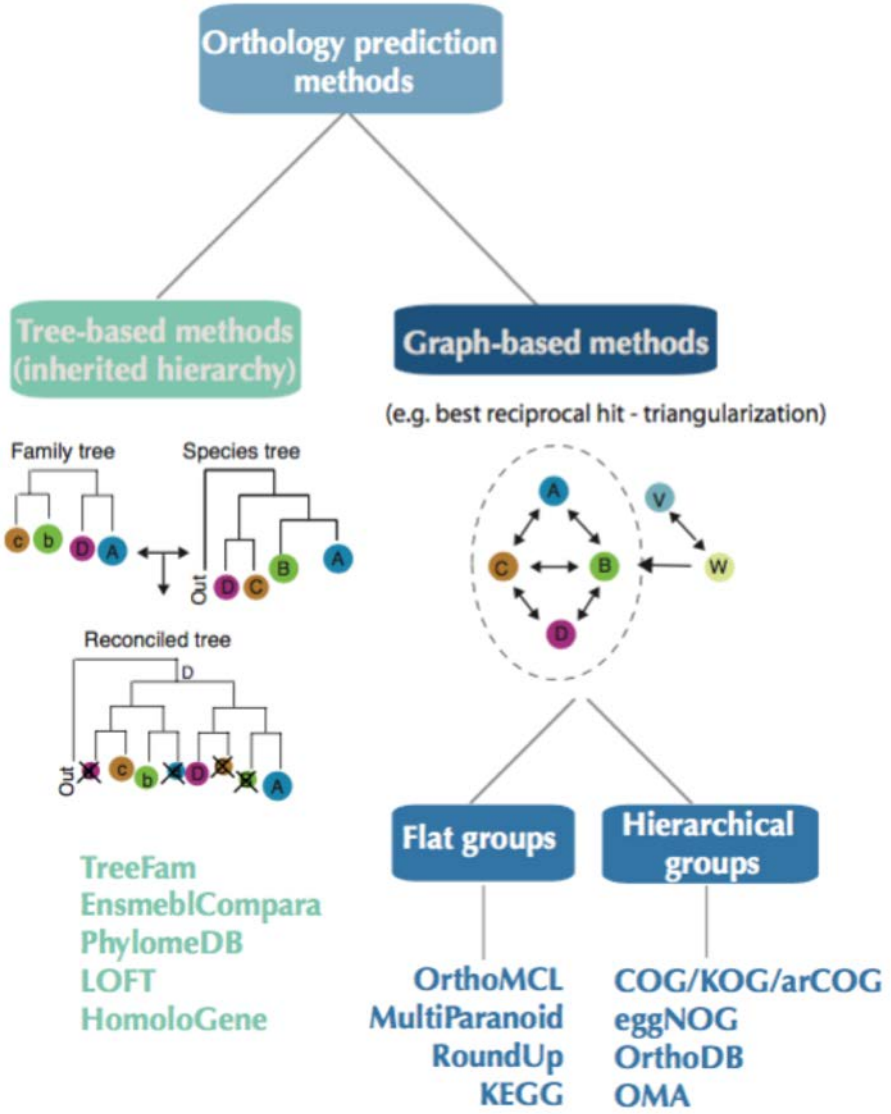
Others

Combination of (1) and (2)

Some uses synteny

Tools

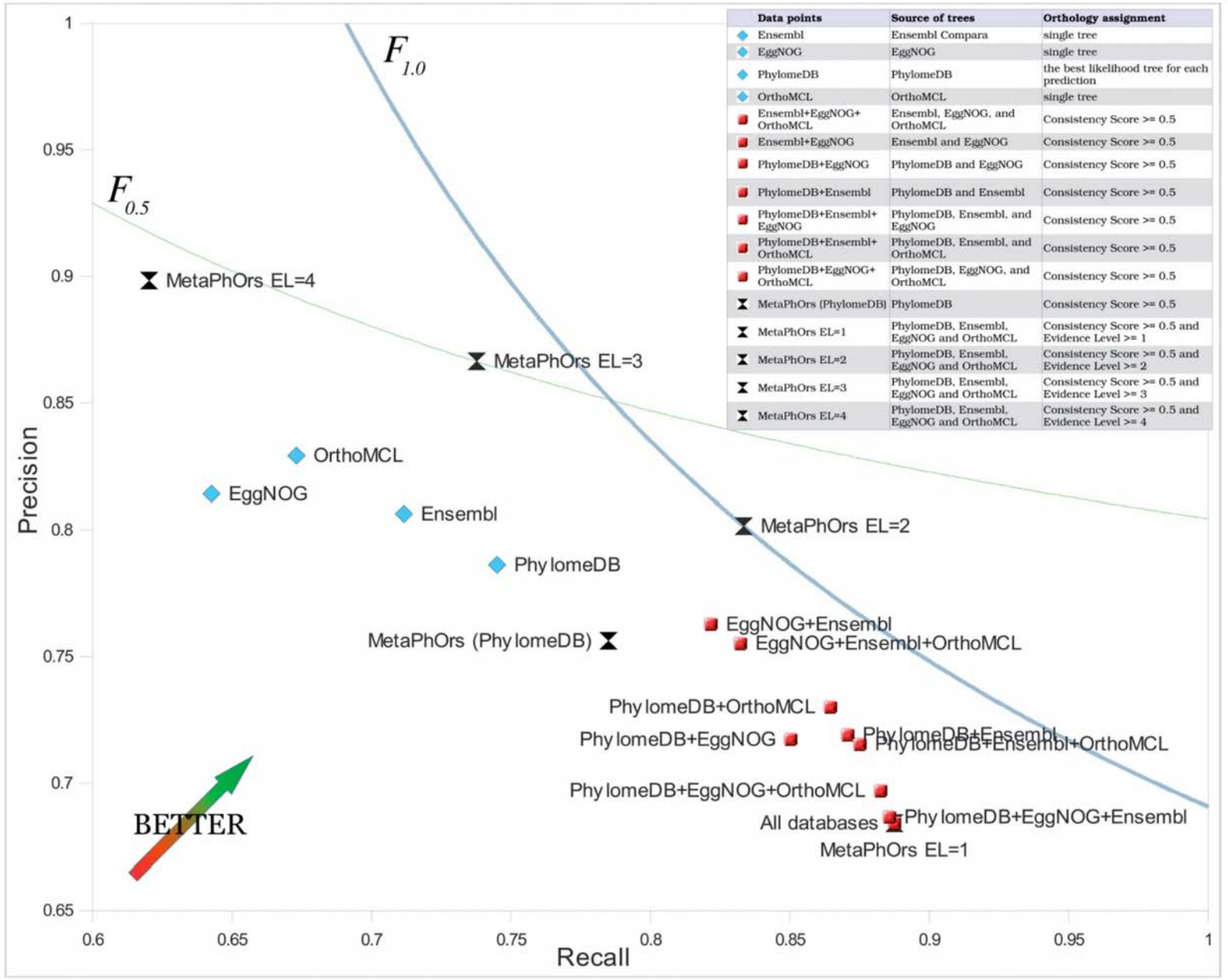
Method	Type	Comments	Reference
BUSCO	Graph	Based on precomputed "universal single-copy" genes (defined for a number of standard clades), and thus inherently limited to these. Originally developed to assess genome completeness.	(Waterhouse et al., 2017)
COG/KOG	Graph	One of the first methods, still widely used for prokaryotic data. Includes a manual curation step.	(Tatusov et al., 2003)
EggNOG	Hybrid	Originally developed as extension of COG/KOG. Recent versions also include tree-based refinements.	(Huerta-Cepas et al., 2016b)
ETE 3.0	Tree	General purpose tree analysis and visualisation package for Python, with species overlap function.	(Huerta-Cepas et al., 2016a)
Forester	Tree	General purpose tree analysis and visualisation software, including reconciliation function.	(Zmasek and Eddy, 2001)
GIGA	Tree	Gene/species tree reconciliation algorithm used in the PANTHER database. Also includes a heuristic for lateral gene transfer detection.	(Thomas, 2010)
GSR	Tree	Probabilistic gene/species tree reconciliation method	(Akerborg et al., 2009)
HaMSTR	Graph	The method uses a reference species to define one Hidden Markov Model per orthologous group, followed by reciprocal best hit within a family	(Ebersberger et al., 2009)
Hieranoid	Graph	Successor of Inparanoid to infer hierarchical orthologous groups from multiple species	(Kaduk et al., 2017)
Inparanoid	Graph	Infers orthologous groups independently for each pair of species.	(Sonnhammer and Östlund, 2015)
MetaPhOrs	Hybrid	Meta-method integrating predictions from multiple sources.	(Pryszcz et al., 2011)
Notung	Tree	Gene/species tree reconciliation software, with optional support for lateral gene transfer inference.	(Chen et al., 2000)
OMA	Graph	Infers both types of groups reviewed in this chapter: strict groups (suitable as markers for species tree inference) and hierarchical orthologous groups.	(Altenhoff et al., 2018a)
OrthoDB	Graph	Infers hierarchical orthologous groups. Used to infer the single-copy universal gene models of BUSCO.	(Zdobnov et al., 2017)
OrthoFinder	Graph	Infers hierarchical orthologous group with respect to the deepest speciation level only (the last common	(Emms and Kelly, 2015)



Tekaia (2016)
Fernández *et al* (2019)

https://questfororthologs.org/orthology_databases

Every tool kind of disagrees…



Data points	Source of trees	Orthology assignment
Ensembl	Ensembl Compara	single tree
EggNOG	EggNOG	single tree
PhylomeDB	PhylomeDB	the best likelihood tree for each prediction
OrthoMCL	OrthoMCL	single tree
Ensembl+EggNOG+ OrthoMCL	Ensembl, EggNOG, and OrthoMCL	Consistency Score >= 0.5
Ensembl+EggNOG	Ensembl and EggNOG	Consistency Score >= 0.5
PhylomeDB+EggNOG	PhylomeDB and EggNOG	Consistency Score >= 0.5
PhylomeDB+Ensembl	PhylomeDB and Ensembl	Consistency Score >= 0.5
PhylomeDB+Ensembl+ EggNOG	PhylomeDB, Ensembl, and EggNOG	Consistency Score >= 0.5
PhylomeDB+Ensembl+ OrthoMCL	PhylomeDB, Ensembl, and OrthoMCL	Consistency Score >= 0.5
PhylomeDB+EggNOG+ OrthoMCL	PhylomeDB, EggNOG, and OrthoMCL	Consistency Score >= 0.5
MetaPhOrs (PhylomeDB)	PhylomeDB	Consistency Score >= 0.5
MetaPhOrs EL=1	PhylomeDB, Ensembl, EggNOG and OrthoMCL	Consistency Score >= 0.5 and Evidence Level >= 1
MetaPhOrs EL=2	PhylomeDB, Ensembl, EggNOG and OrthoMCL	Consistency Score >= 0.5 and Evidence Level >= 2
MetaPhOrs EL=3	PhylomeDB, Ensembl, EggNOG and OrthoMCL	Consistency Score >= 0.5 and Evidence Level >= 3
MetaPhOrs EL=4	PhylomeDB, Ensembl, EggNOG and OrthoMCL	Consistency Score >= 0.5 and Evidence Level >= 4

Pryszcz et al (2011)

Caveats

Evolution of multi-domain proteins

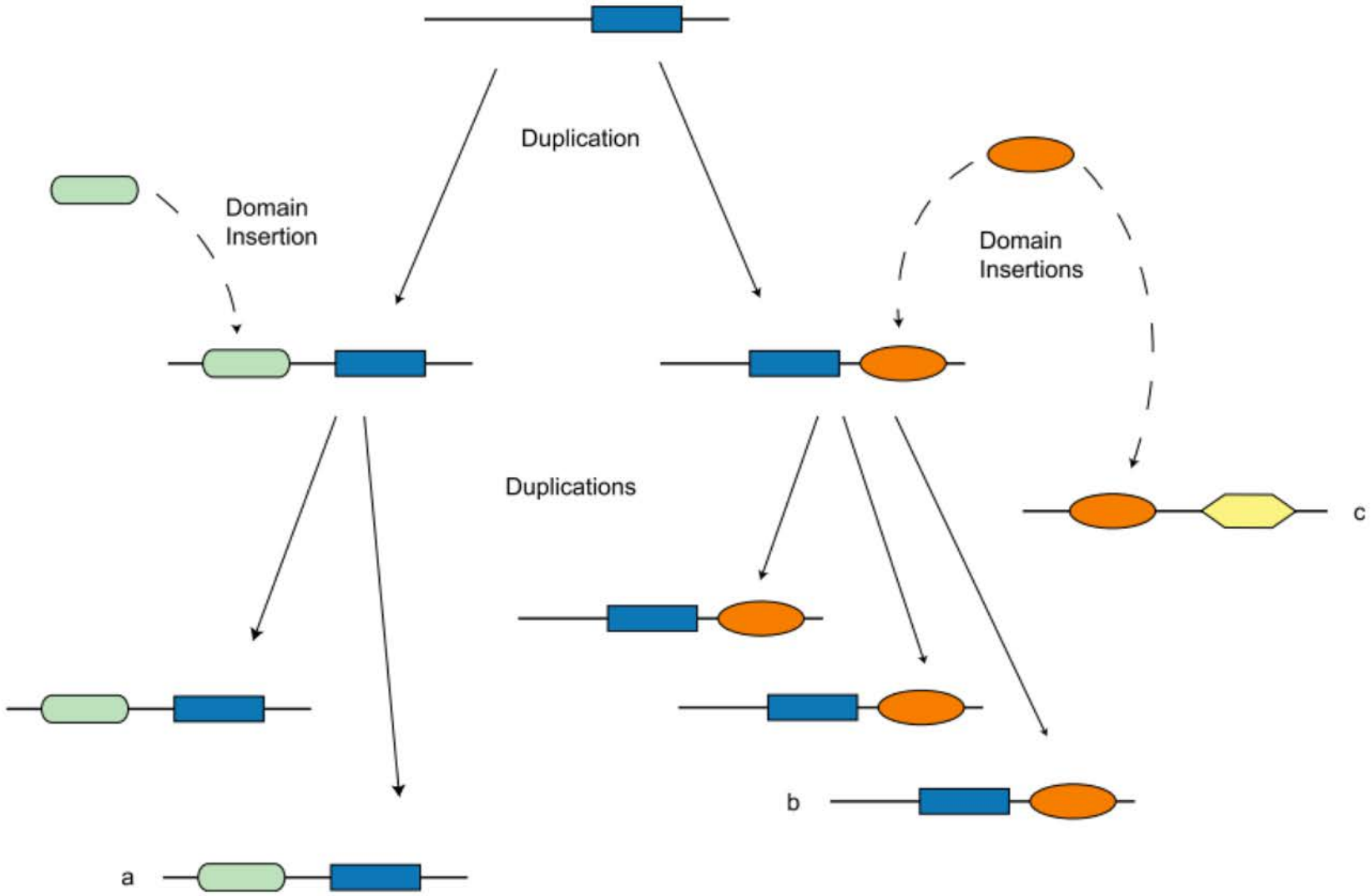
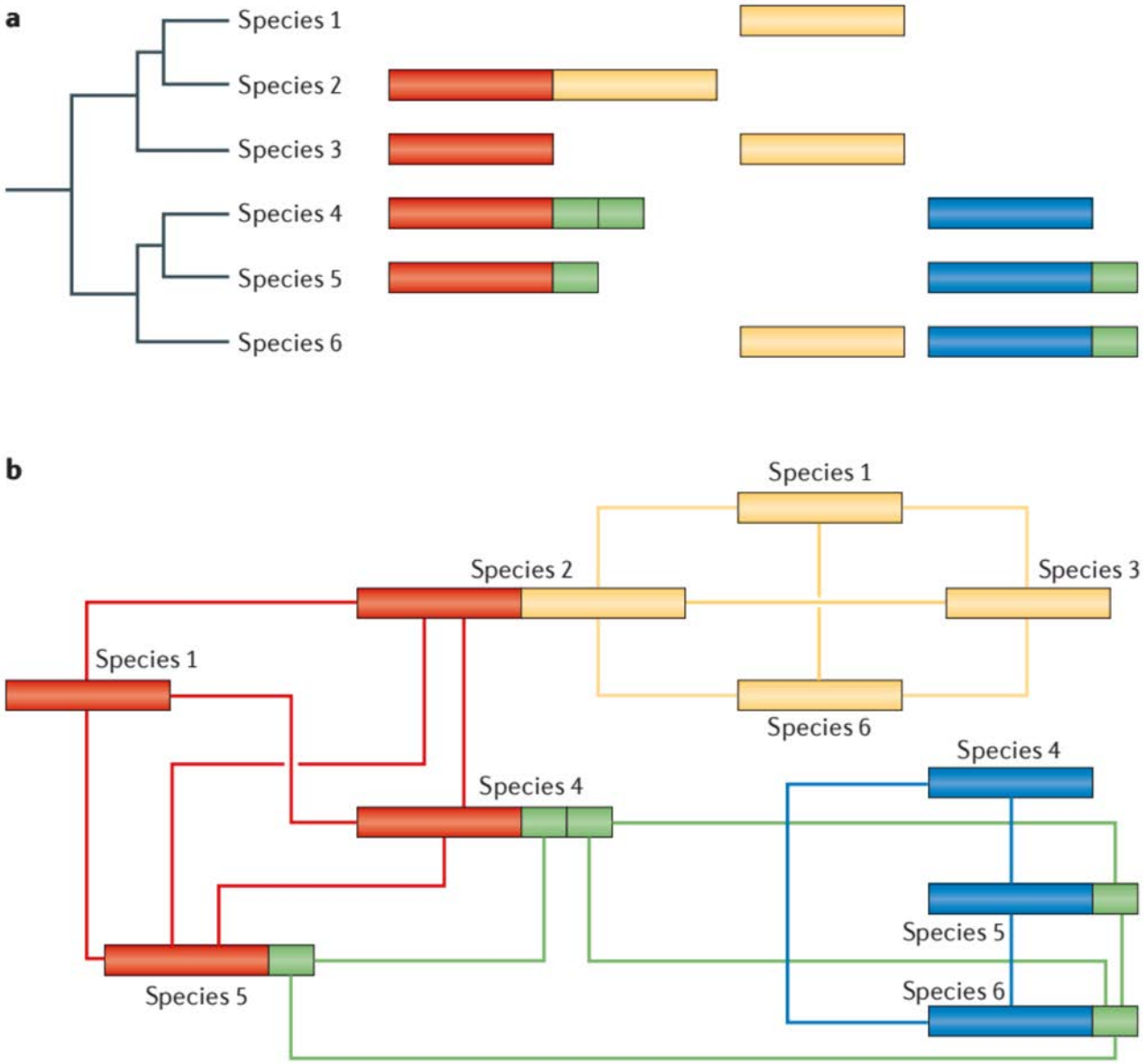


Figure 1. The evolution of a hypothetical multidomain family by gene duplication and domain insertion. Genes in the *a* and *b* subfamilies share a common ancestor but do not have identical domain composition. Gene *c* shares a homologous domain with genes in the *b* subfamily, but there is no gene that is ancestral to both *b* and *c*.
doi:10.1371/journal.pcbi.1000063.g001

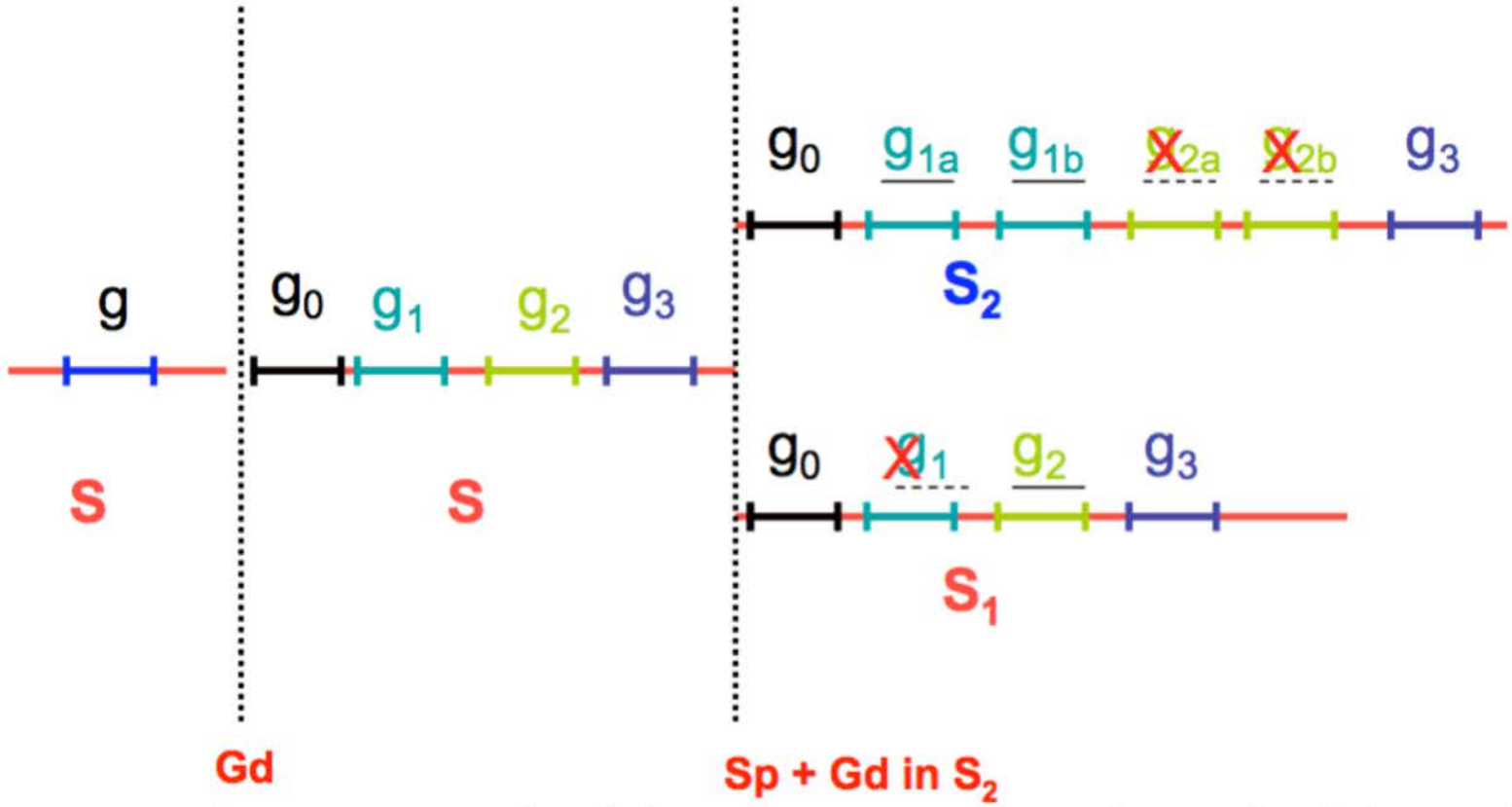
Song et al (2008)

Problem of clustering to assign gene families when comes to different domain combinations



Gabaldon and Koonin (2013)

Detection can go wrong: Example of an orthology misleading situation



We assume that gene g_1 (in S_1) and genes g_{2a} and g_{2b} (in S_2) are lost, similarity and phylogenetic methods for orthology detection will assign erroneously orthology to g_2, g_{1a} and g_{1b}. Indeed these are not orthologous, because g_2, g_{1a} and g_{1b} do not result from the same ancestral gene after the speciation event.

In this case solely the environment conservation, will help in detecting the gene duplication and loss event, and hypothesise their non-orthology.

Summary point

SUMMARY POINTS

1. Orthologs and paralogs are two types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication.
2. Distinguishing between orthologs and paralogs is crucial for successful functional annotation of genomes and for reconstruction of genome evolution.
3. A finer classification of orthologs and paralogs has been developed to reflect the interplay between duplication and speciation events, and effects of gene loss and horizontal gene transfer on the observed homologous relationship.
4. Methods for identification of sets of orthologous and paralogous genes involve phylogenetic analysis and various procedures for sequence similarity–based clustering.
5. Analysis of clusters of orthologous and paralogous genes is instrumental in genome annotation and in delineation of trends in genome evolution.
6. Rearrangements of gene structure confound orthologous and paralogous relationships.
7. The gene-centered concepts of orthology and paralogy can be generalized downward, to the level of strings of nucleotides and even single base pairs, and upward, to multigene arrays.

Koonin (2005)

Phylogenomics

Phylogenomics aims at inferring detailed information about the evolutionary histories of organisms by using whole genomes rather than just a single gene or a few genes. The term was coined by Jonathan Eisen in the context of prediction of gene function

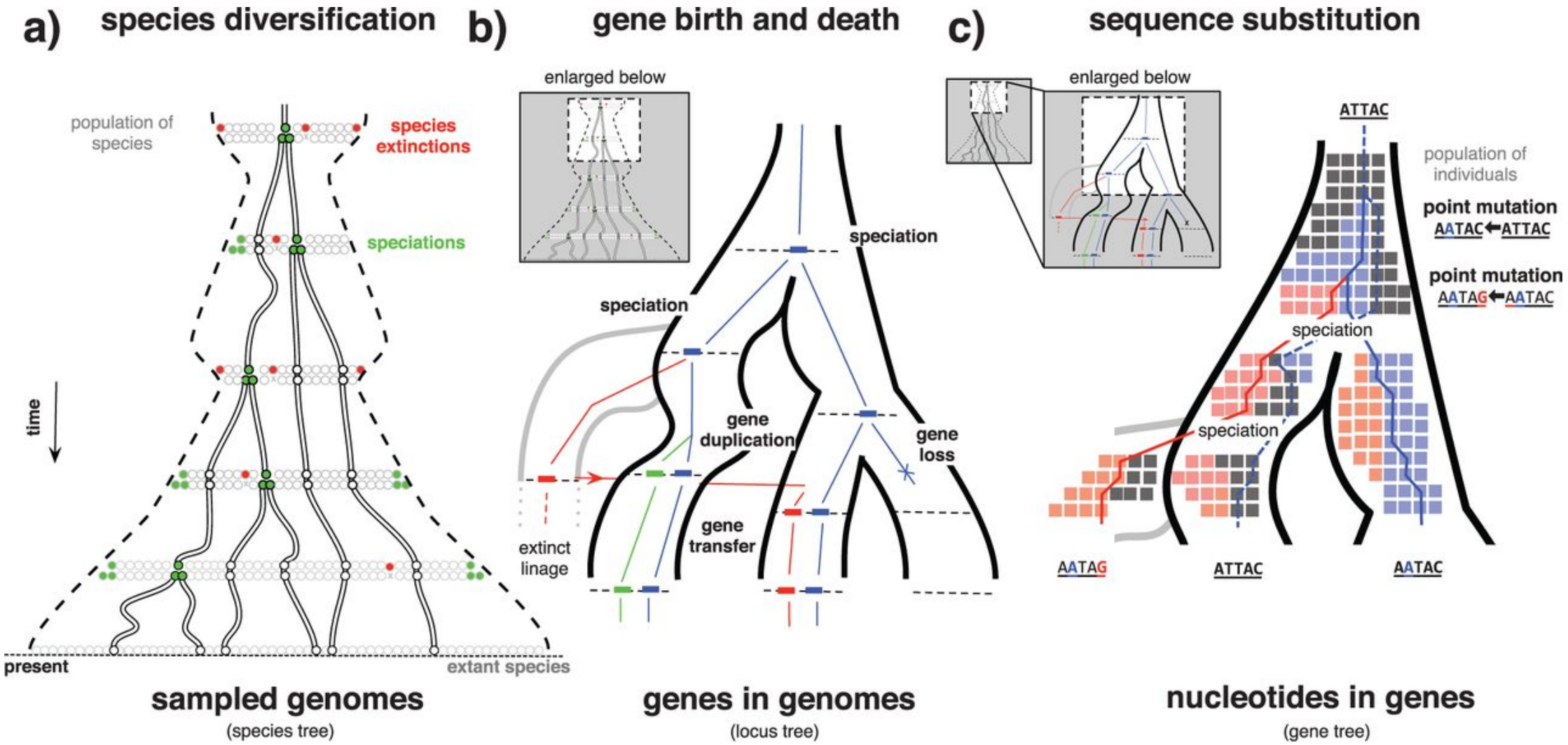
It would be difficult or impossible to understand the evolutionary history of an organism, even having available its whole genome sequence, in isolation. So it is always the case the phylogenomics is practiced for sets of genomes.

During the last 50 years, phylogeny has become more and more based on molecular data, increasingly **favoring homologous sequences over morphological characters**. This approach has been extremely fruitful, **producing constant improvement in the accuracy and resolution of phylogenetic reconstruction together with our understanding of evolutionary processes at the molecular level**.

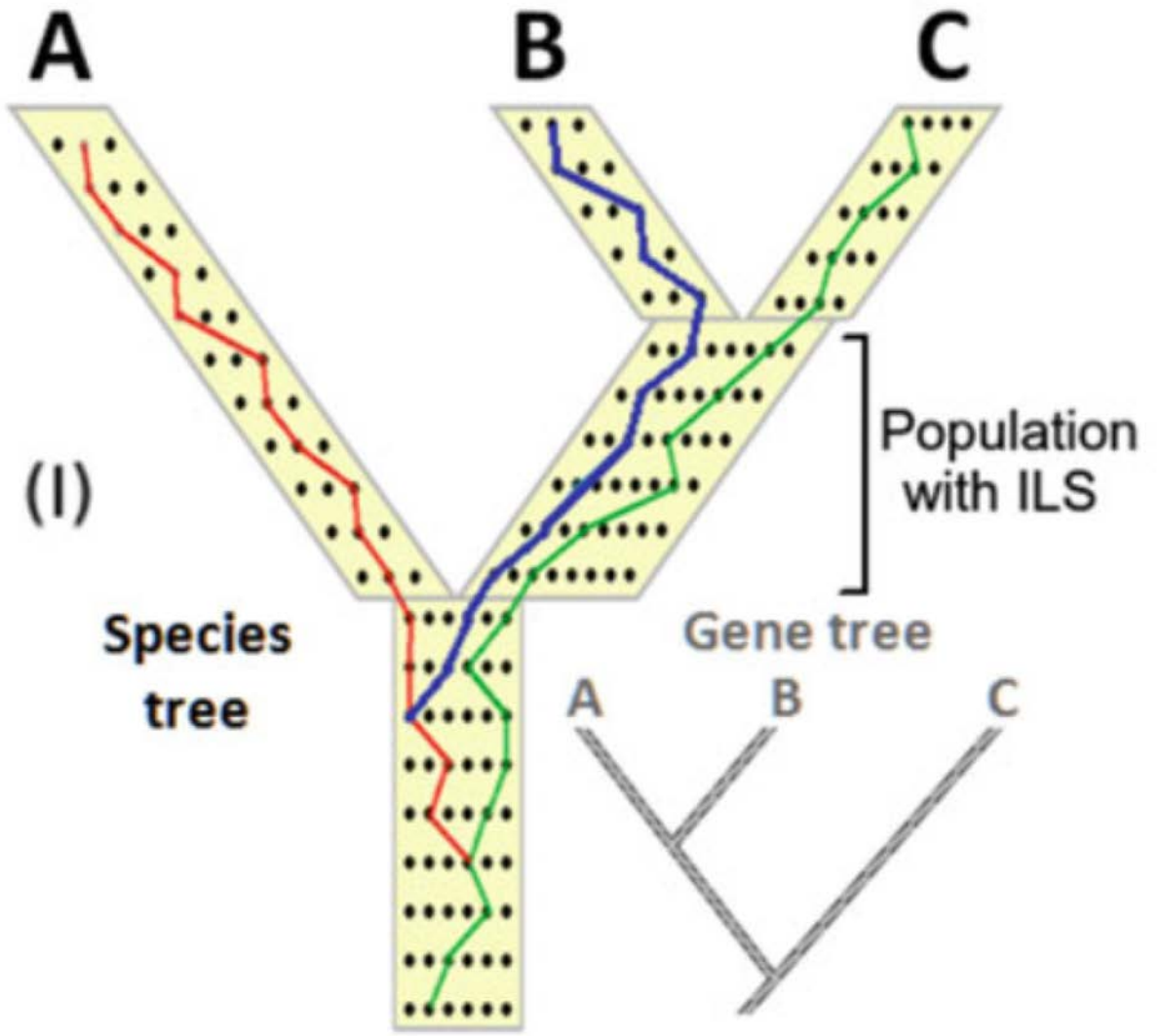
However, we have known all along that we are barking up the wrong trees: with increasing sophistication in the models of sequence evolution, **we have been reconstructing trees describing the history of fragments of genomic sequence, which we will liberally call "gene" in this review, but never the history of species. Gene trees are not species trees** (Maddison 1997).

https://academic.oup.com/sysbio/article/64/1/e42/1634124

Each level of the hierarchy contributes to generating phylogenetic signal that can lead to differences between reconstructed gene trees.



Processes that may induce gene trees that are different than the actual species tree

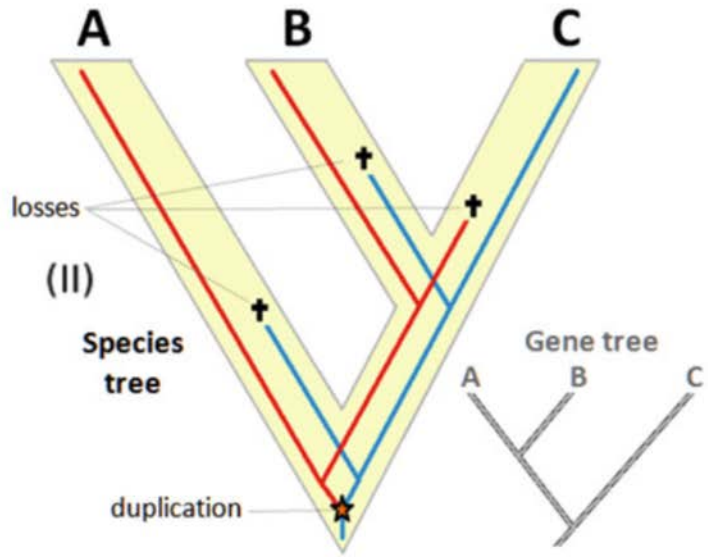


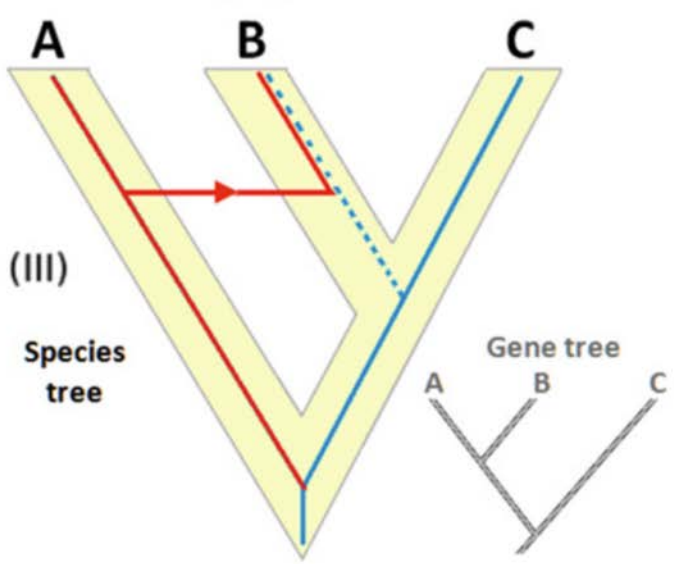
i) Incomplete lineage sorting

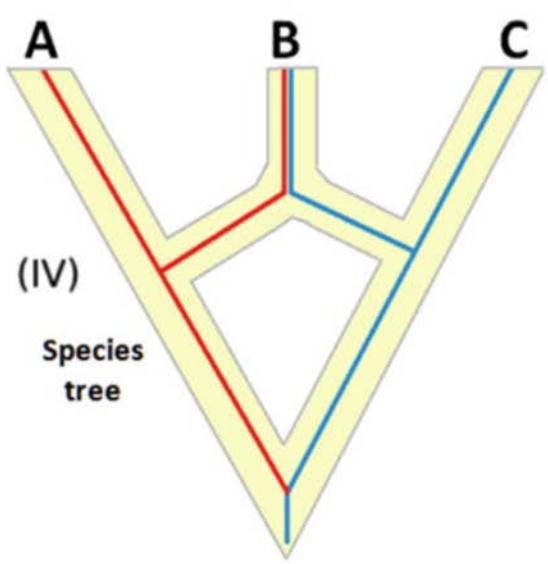
When a species splits in two, allelic lineages sort into the two descendant species, and this lineage sorting varies along the genome.

If speciation events are close in time, the lineage sorting process may be incomplete at the second speciation event and lead to gene genealogies that do not match the species phylogeny

Processes that may induce gene trees that are different than the actual species tree





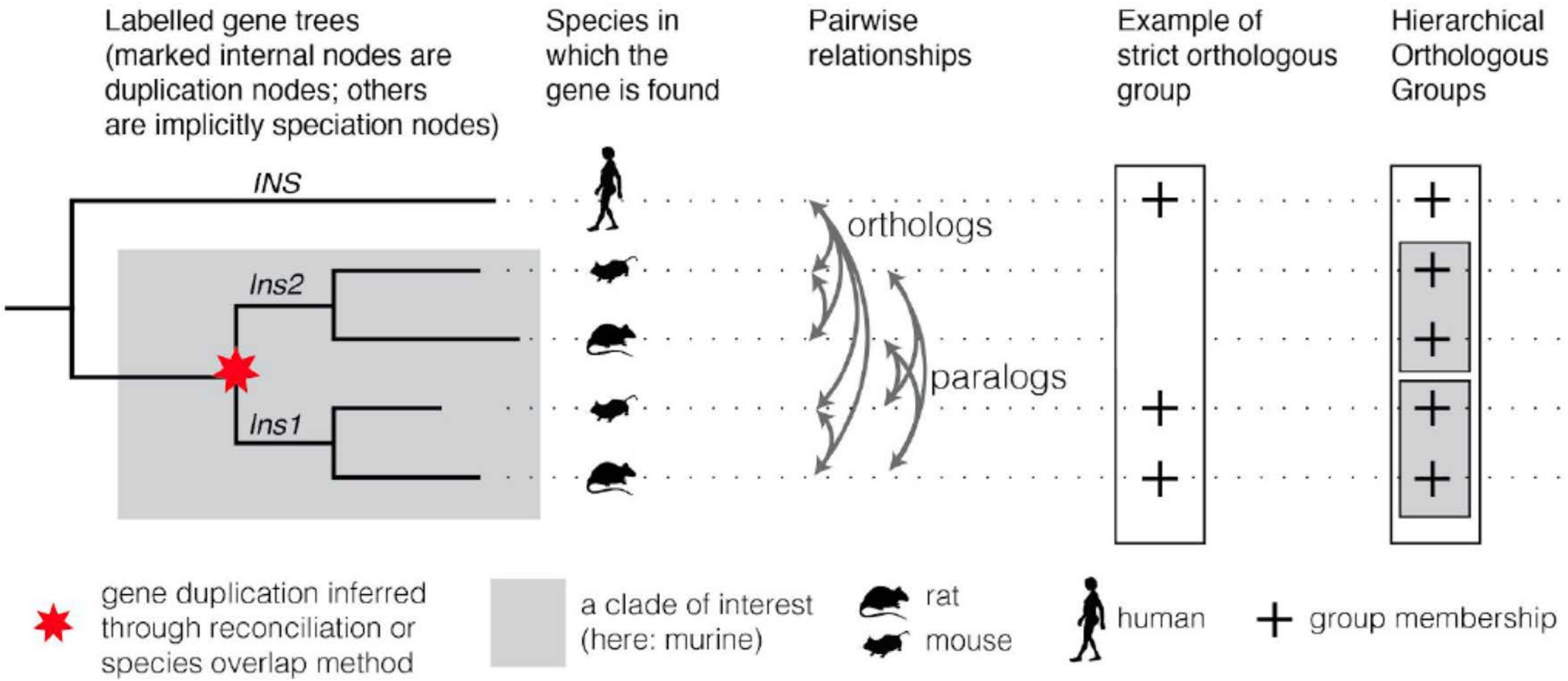


(II) **Duplication and Loss**
a locus may generate a duplicate somewhere in the genome, and then both may be inherited or just a single copy is maintained in each lineage.

(III) **Horizontal Gene Transfer**
(HGT): a donor DNA segment (from taxon A) is transmitted and incorporated into the host's genome (taxon B)

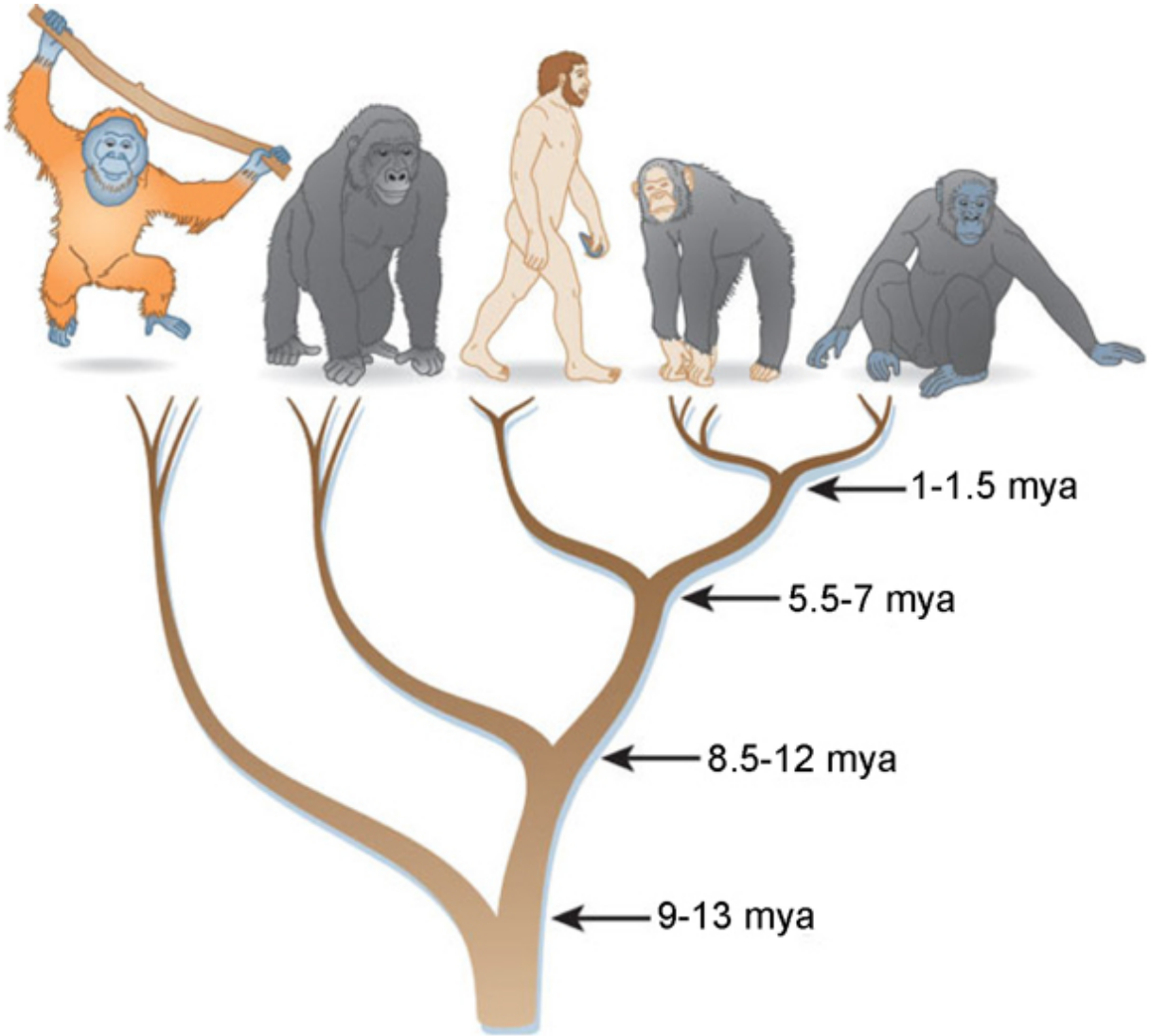
(IV) **Hybridization/Introgression**
in extreme cases of lateral transfer, or upon mixing of related species, different regions of the genome will bear two distinct evolutionary histories;

Problem of obtaining the ‘true’ orthologs for phylogenomics



Fernández *et al* (2019)

Why is Studying (Ape) Speciation Important? (Example)



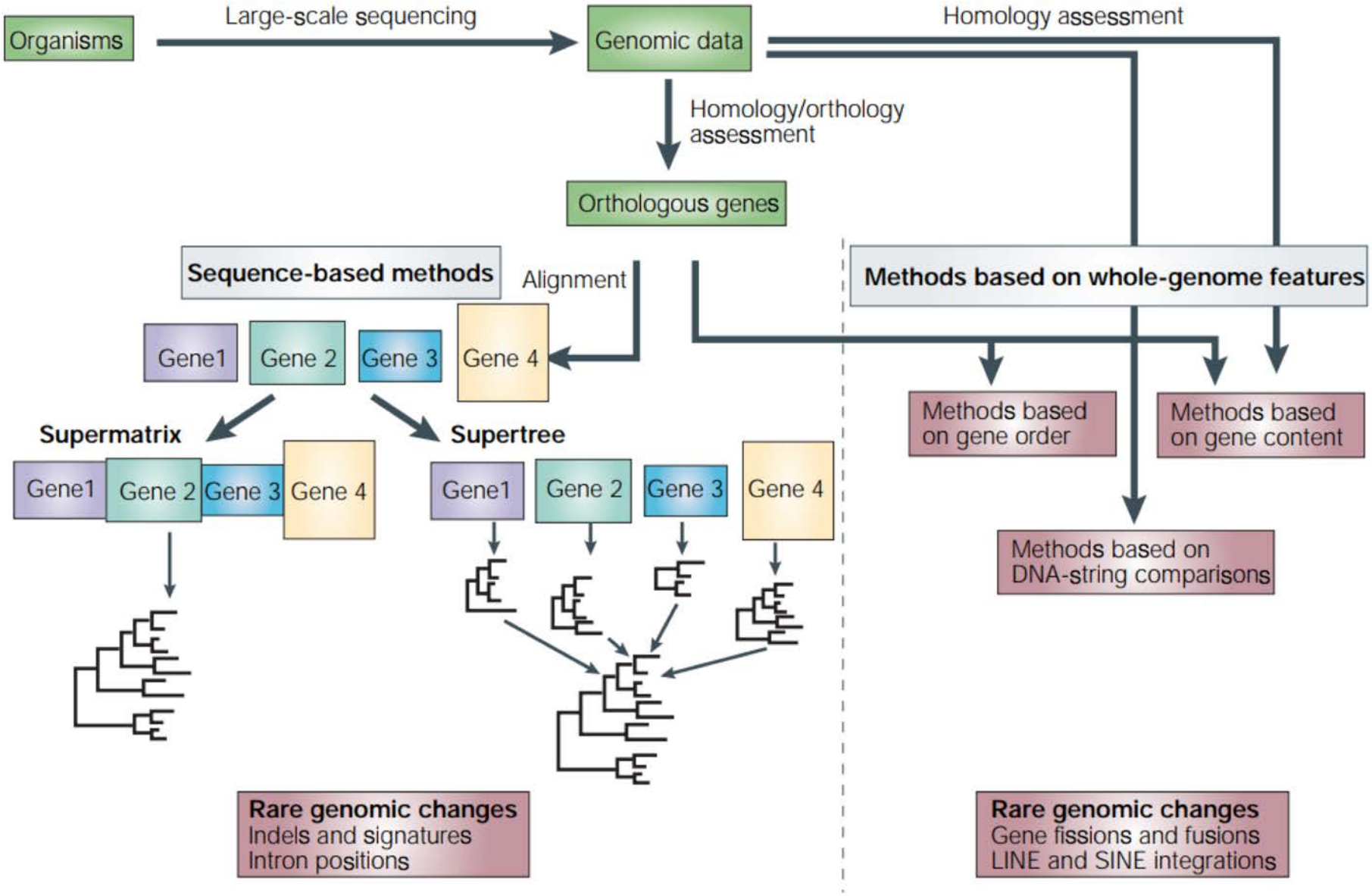
These studies also led to rich discussions about the suite of **factors that may have contributed to promoting speciation in the last common ancestor of humans and African apes**, as well as the **factors that might have contributed to creating the amazing diversity of Hominins that co-existed with each other during the Pliocene and Pleistocene** (Foley 2002).

For many years, there was considerable debate about which of the African apes is our closest relative…. The general consensus that emerged is that we share a more recent relationship with chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*) than we do with gorillas (*Gorilla gorilla*) (Ruvolo 1997, Chen & Li 2001).

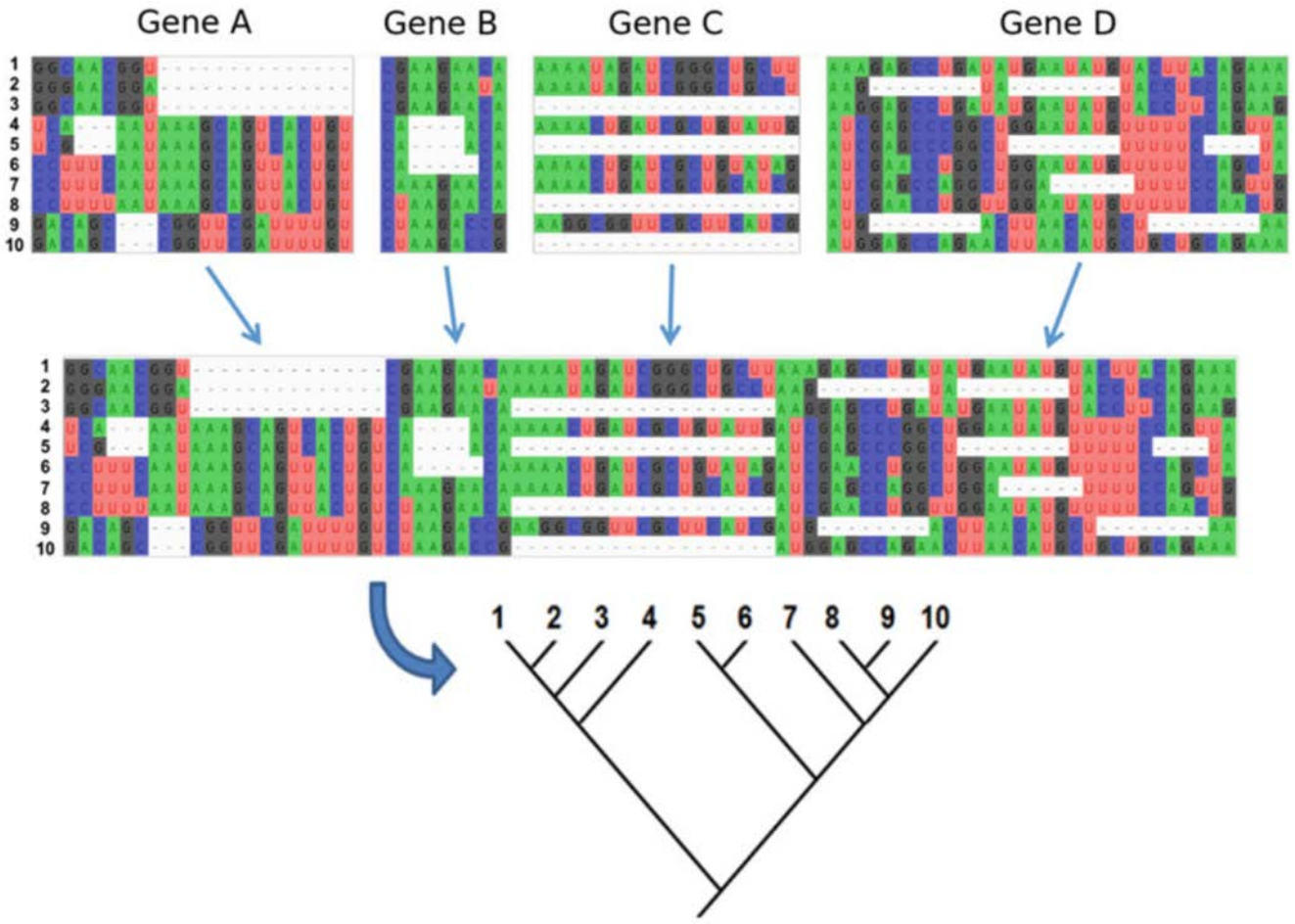
Current estimates indicate that up to 30% of the sequence of the human genome is more closely related to Gorilla than to Chimpanzee due to this process (Scally et al. 2012).

https://www.nature.com/scitable/knowledge/library/primate-speciation-a-case-study-of-african-96682434

Probably the most common (easy) way to construct alignment of concatenated gene shared across all species



Probably the most common (easy) way to construct alignment of concatenated gene shared across all species

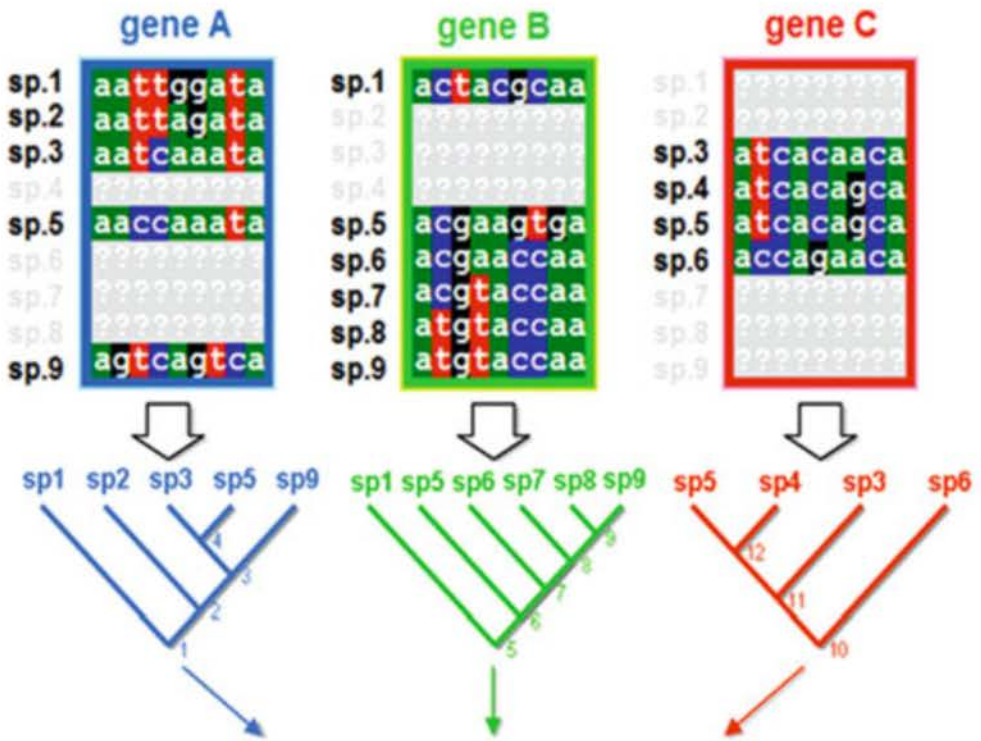


Important drawbacks:

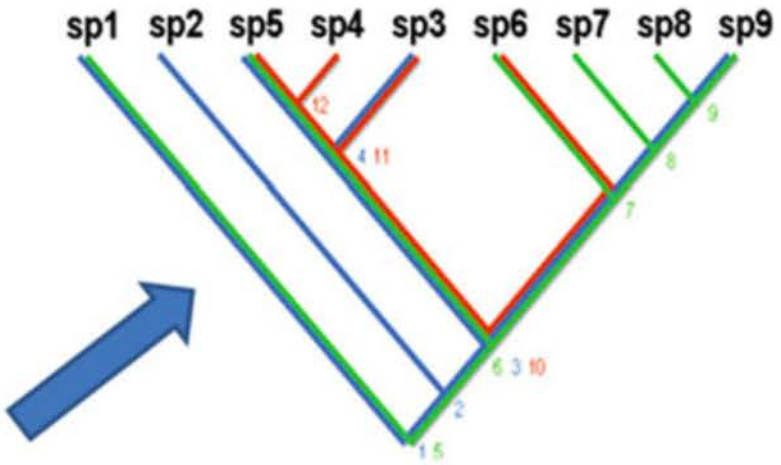
(1) it hinders variation among gene trees by assuming implicitly that all of them conform to a single species tree;
(2) if sampling was heterogeneous across species there may be too much missing data, which can affect topological reconstruction; Or limited number of genes shared among all species
(3) large data sampling effects inflate credibility in some clades;
(4) spurious hidden support can lead to support for non-existent clades; and
(5) in case of moderate to severe levels of ILS, supermatrix can become statistically inconsistent.

From genes to supertrees

Instead of forcing all gene trees to comply to a single tree, **supertree methods infer the best topology for each gene (using the same phylogenetic method for each), and then a topological consensus is obtained**. Such methods are able to make consensus trees even if the number of leaves among gene trees differs but overlaps to some extent, for example when a gene has not been sequenced for some taxa

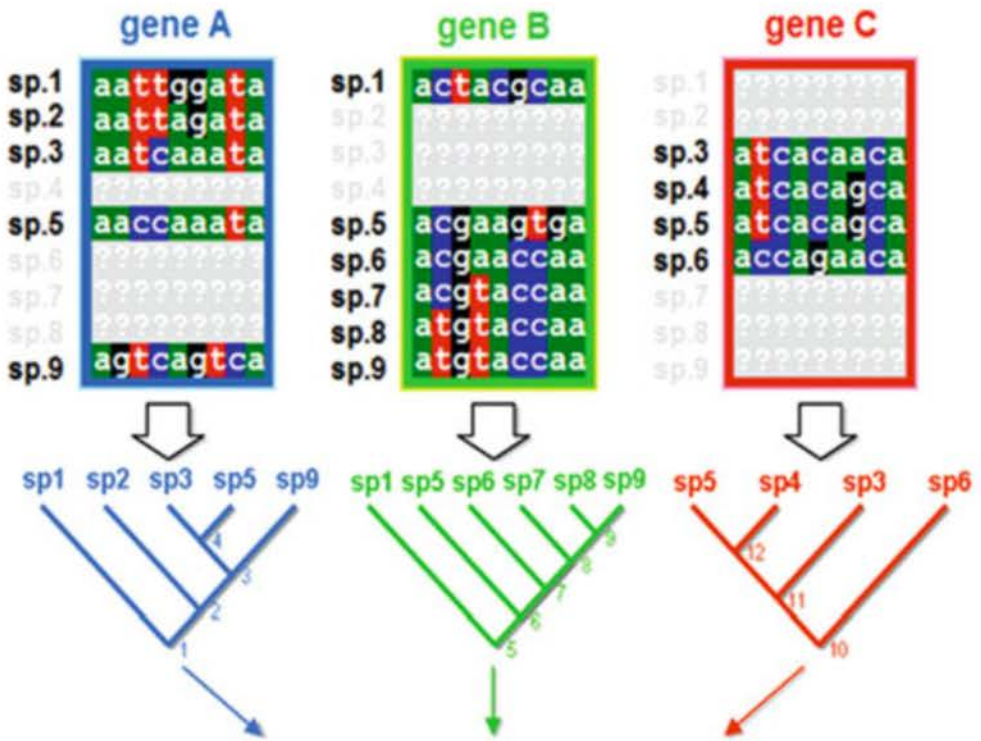


MRP	1	2	3	4	5	6	7	8	9	10	11	12
sp1	1	0	0	0	1	0	0	0	0	?	?	?
sp2	1	1	0	0	?	?	?	?	?	?	?	?
sp3	1	1	1	1	?	?	?	?	?	1	1	0
sp4	?	?	?	?	?	?	?	?	?	1	1	1
sp5	1	1	1	1	1	1	0	0	0	1	1	1
sp6	?	?	?	?	1	1	1	1	0	1	0	0
sp7	?	?	?	?	1	1	1	1	0	?	?	?
sp8	?	?	?	?	1	1	1	1	1	?	?	?
sp9	1	1	1	0	1	1	1	1	1	?	?	?

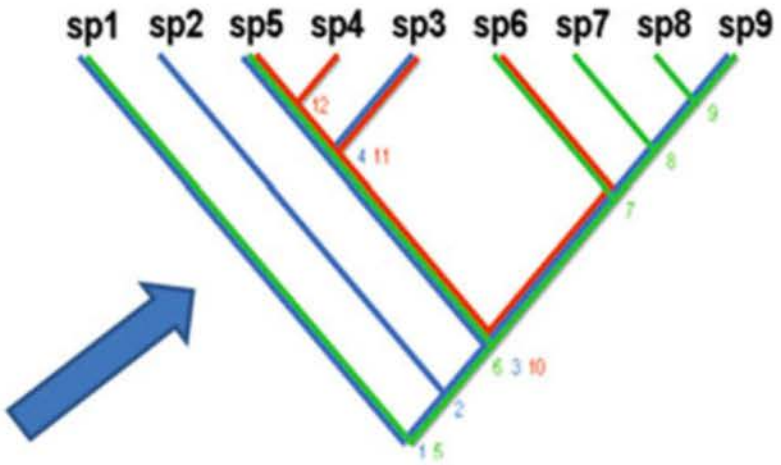


Current methods

A step beyond supertrees is the use of methods that take into consideration specific evolutionary processes that may be responsible for differences in gene topologies, and then estimate the species tree which would most likely have generated such gene trees, under different scenarios



MRP	1	2	3	4	5	6	7	8	9	10	11	12
sp1	1	0	0	0	1	0	0	0	0	?	?	?
sp2	1	1	0	0	?	?	?	?	?	?	?	?
sp3	1	1	1	1	?	?	?	?	?	1	1	0
sp4	?	?	?	?	?	?	?	?	?	1	1	1
sp5	1	1	1	1	1	1	0	0	0	1	1	1
sp6	?	?	?	?	1	1	1	1	0	1	0	0
sp7	?	?	?	?	1	1	1	1	0	?	?	?
sp8	?	?	?	?	1	1	1	1	1	?	?	?
sp9	1	1	1	0	1	1	1	1	1	?	?	?



nature ecology & evolution

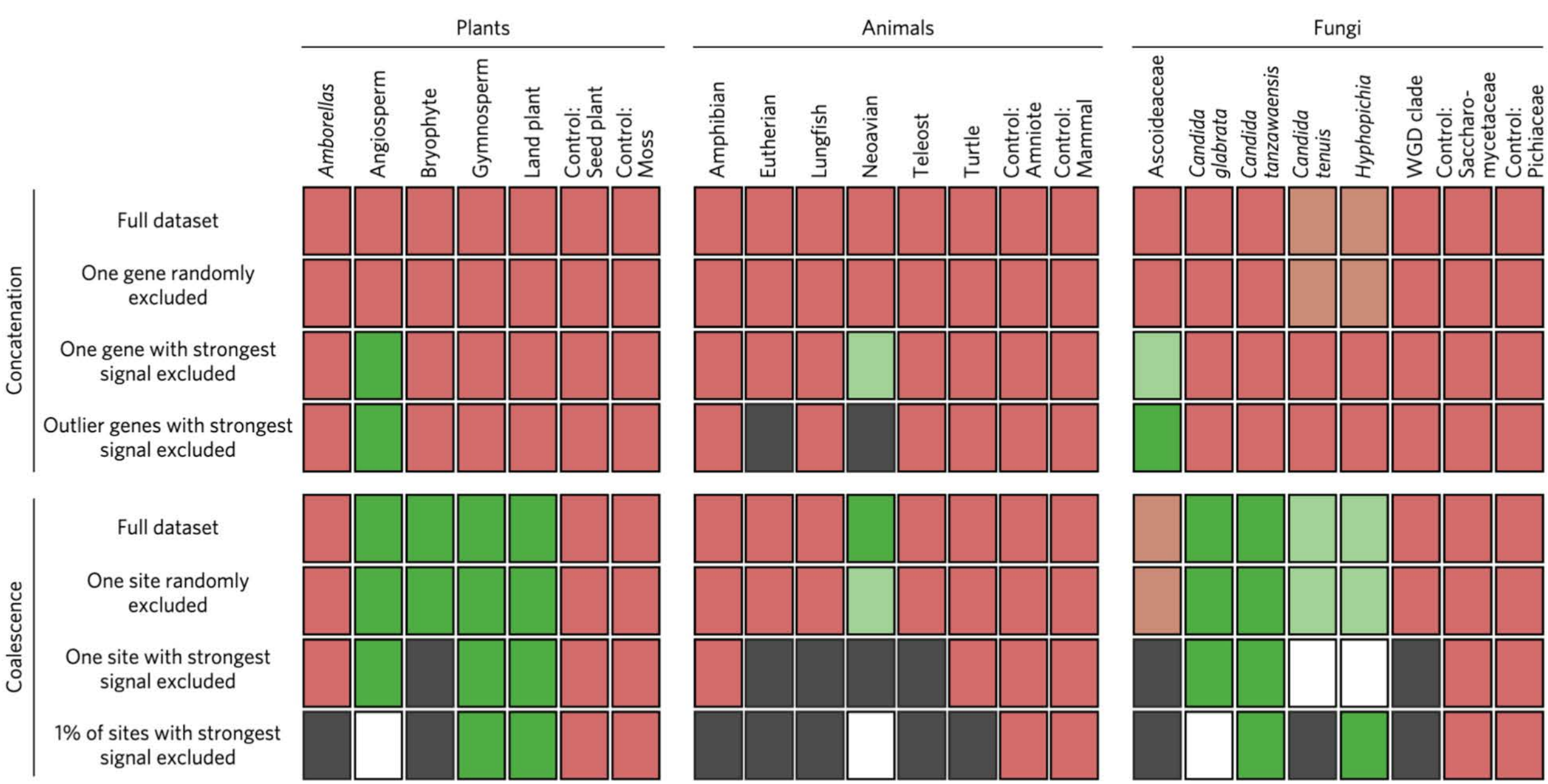
ARTICLES

PUBLISHED: 10 APRIL 2017 | VOLUME: 1 | ARTICLE NUMBER: 0126

Contentious relationships in phylogenomic studies can be driven by a handful of genes

Xing-Xing Shen[1], Chris Todd Hittinger[2] and Antonis Rokas[1]*

Phylogenomic studies have resolved countless branches of the tree of life, but remain strongly contradictory on certain, contentious relationships. Here, we use a maximum likelihood framework to quantify the distribution of phylogenetic signal among genes and sites for 17 contentious branches and 6 well-established control branches in plant, animal and fungal phylogenomic data matrices. We find that resolution in some of these 17 branches rests on a single gene or a few sites, and that removal of a single gene in concatenation analyses or a single site from every gene in coalescence-based analyses diminishes support and can alter the inferred topology. These results suggest that tiny subsets of very large data matrices drive the resolution of specific internodes, providing a dissection of the distribution of support and observed incongruence in phylogenomic analyses. We submit that quantifying the distribution of phylogenetic signal in phylogenomic data is essential for evaluating whether branches, especially contentious ones, are truly resolved. Finally, we offer one detailed example of such an evaluation for the controversy regarding the earliest-branching metazoan phylum, for which examination of the distributions of gene-wise and site-wise phylogenetic signal across eight data matrices consistently supports ctenophores as the sister group to all other metazoans.



Shen *et al* (2017)

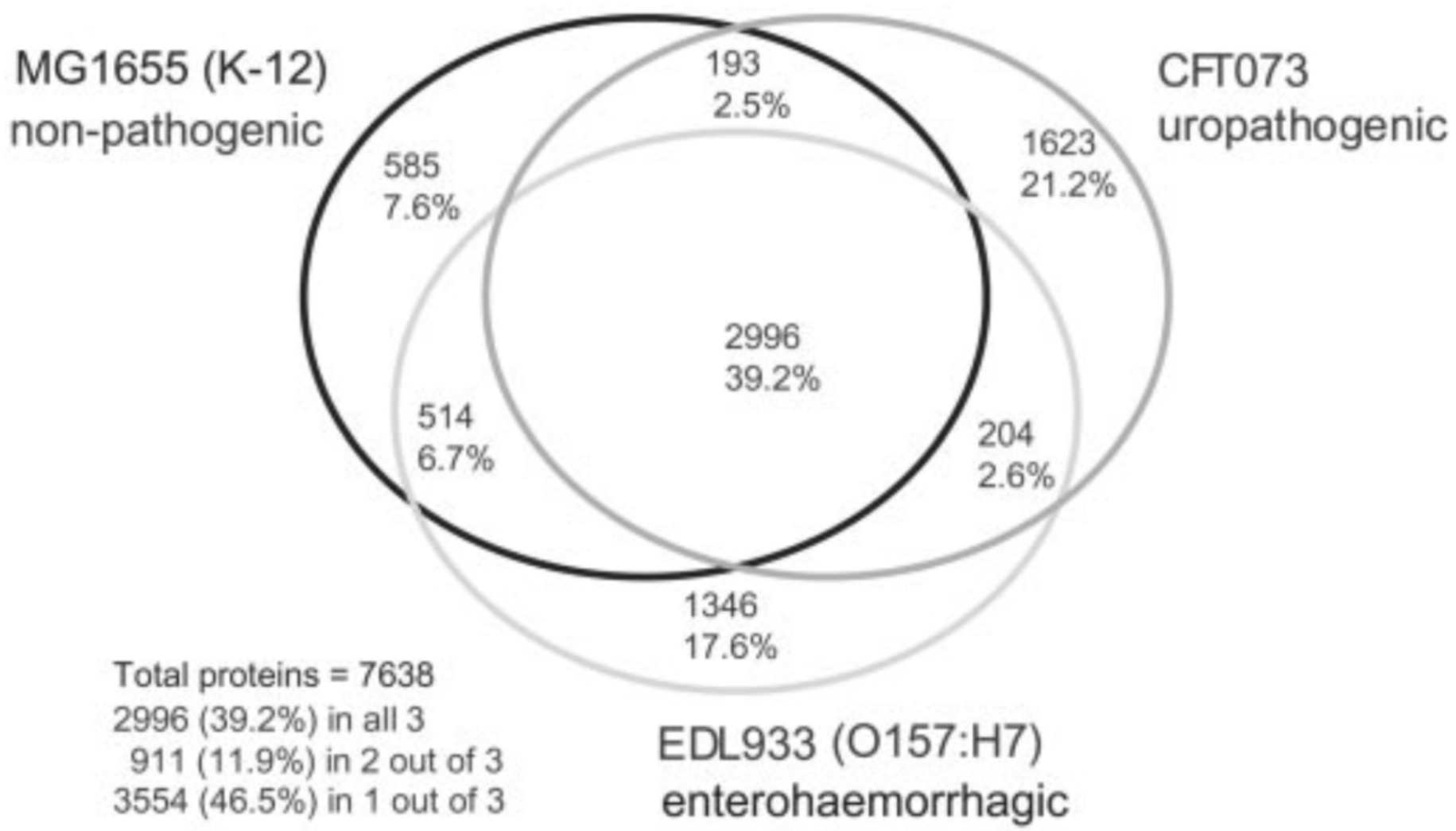
Visualisation of gene content / families

Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*

R. A. Welch*, V. Burland†‡, G. Plunkett III†, P. Redford*, P. Roesch*, D. Rasko§, E. L. Buckles¶, S.-R. Liou†‖, A. Boutin†**, J. Hackett†,††, D. Stroud†, G. F. Mayhew†, D. J. Rose†, S. Zhou†‡‡, D. C. Schwartz†‡‡, N. T. Perna§§, H. L. T. Mobley§, M. S. Donnenberg¶, and F. R. Blattner†

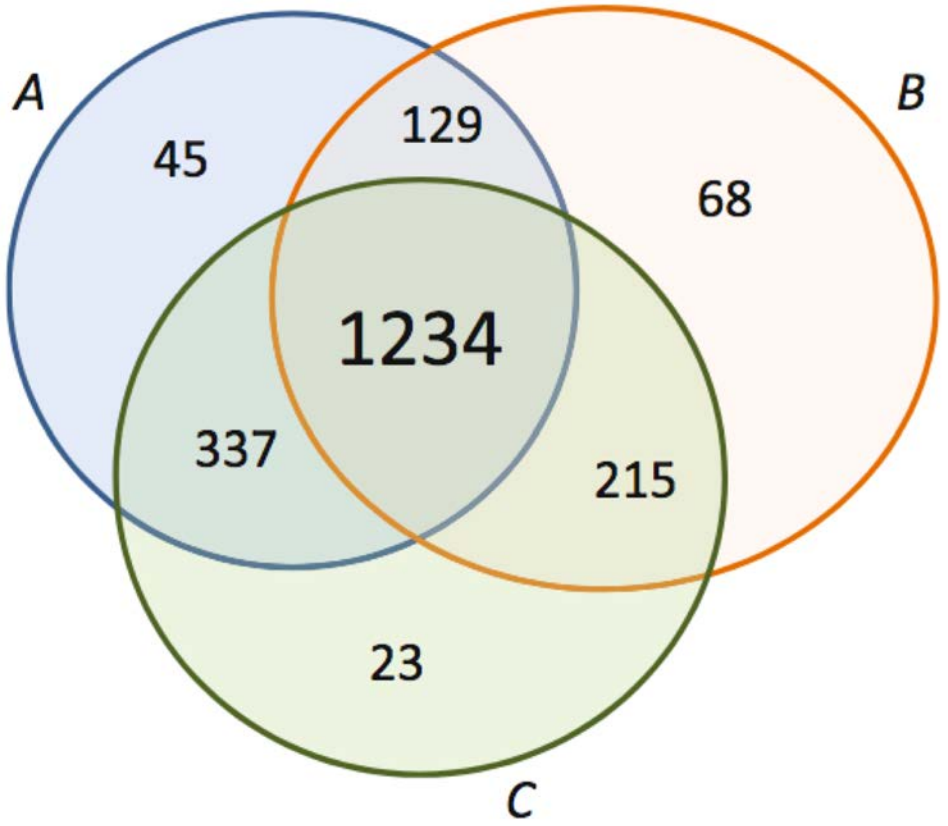
*Department of Medical Microl
Sciences, University of Wisconsi
Department of Medicine, Unive

Edited by John J. Mekalanos, H



Welch *et al* (2002)

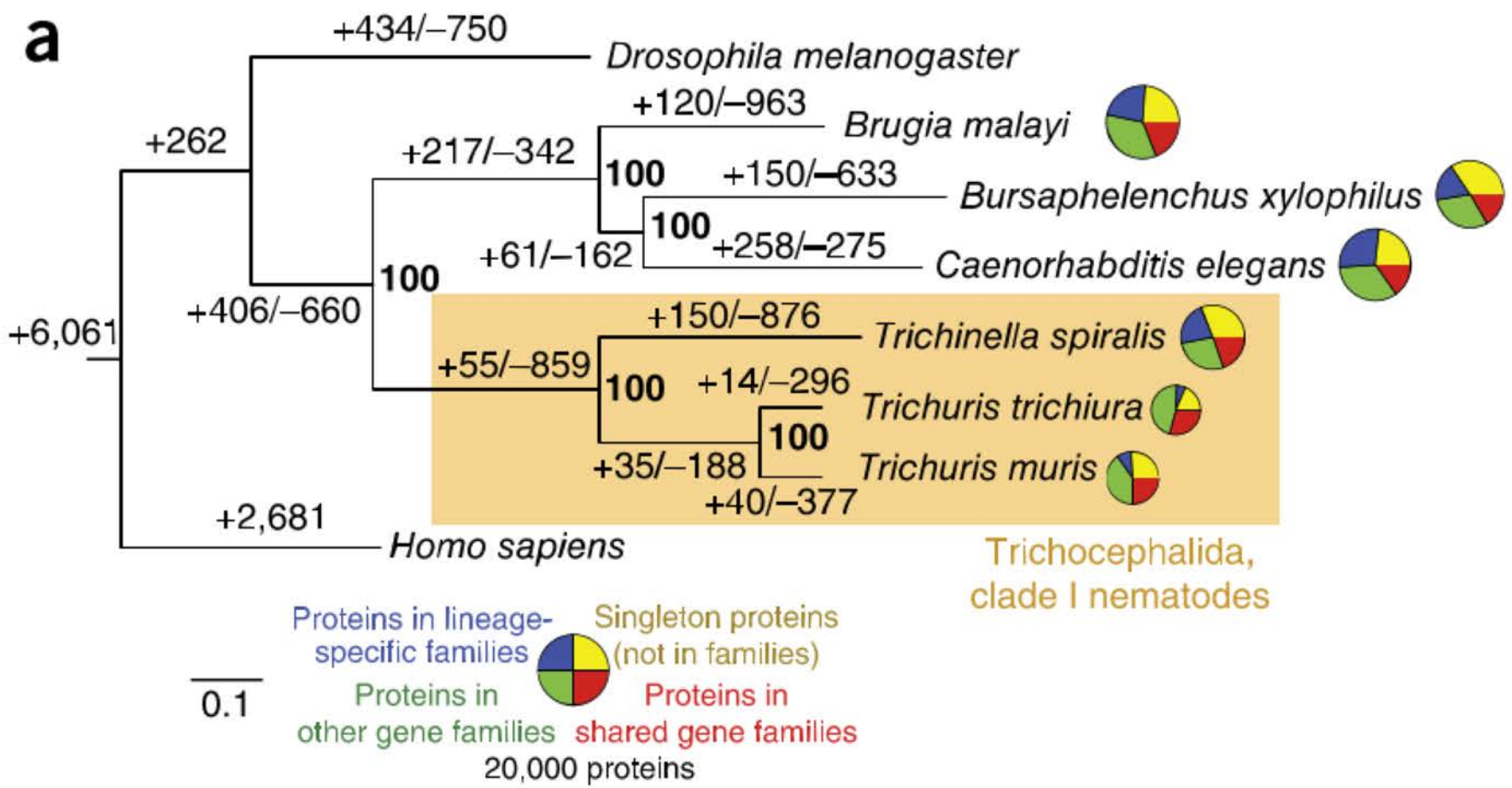
Illustration of a gene content Venn diagram for three hypothetical genomes A , B, and C

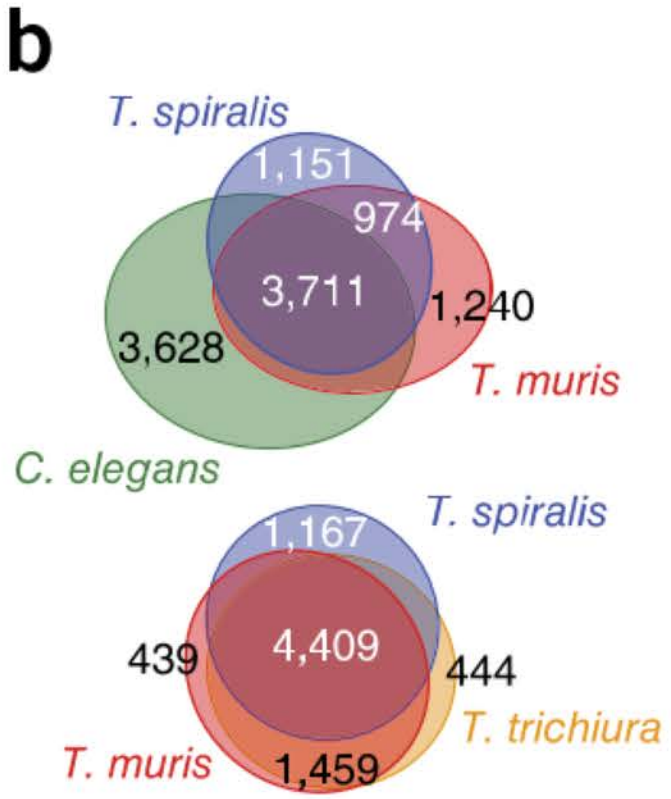


Gene	Genome						
	A	B	C	D	E	F	G
1	✓	✓			✓	✓	
2	✓		✓	✓	✓	✓	✓
3		✓		✓			
4		✓			✓		
5				✓			
6			✓		✓	✓	
7		✓		✓			✓

Schematic representation of a presence/absence gene matrix. Genomes are represented in columns, and gene families are represented in rows

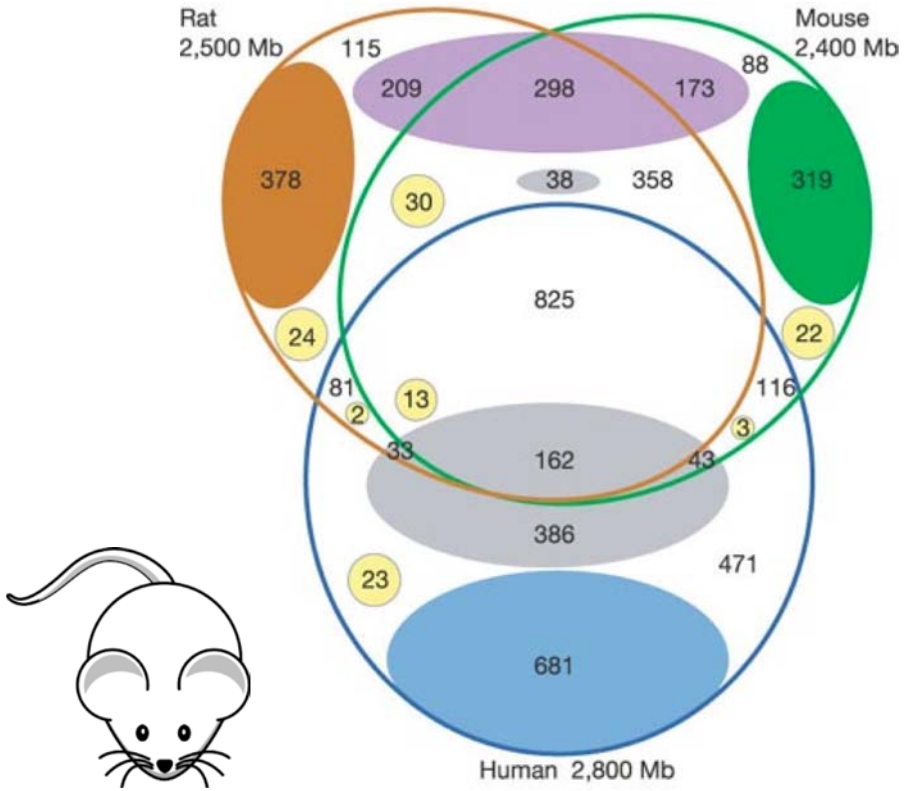
Phylogeny + Venn diagram to show expansion/loss

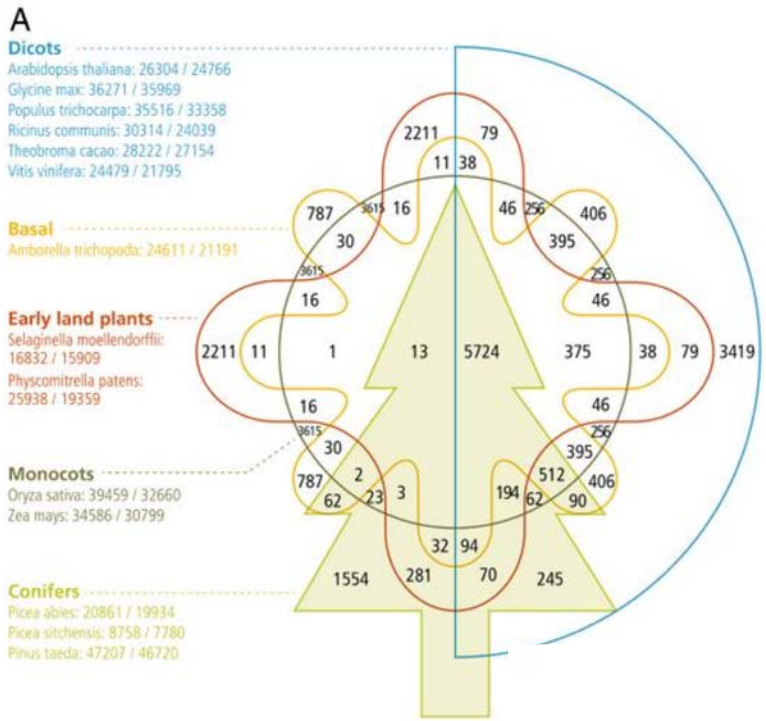


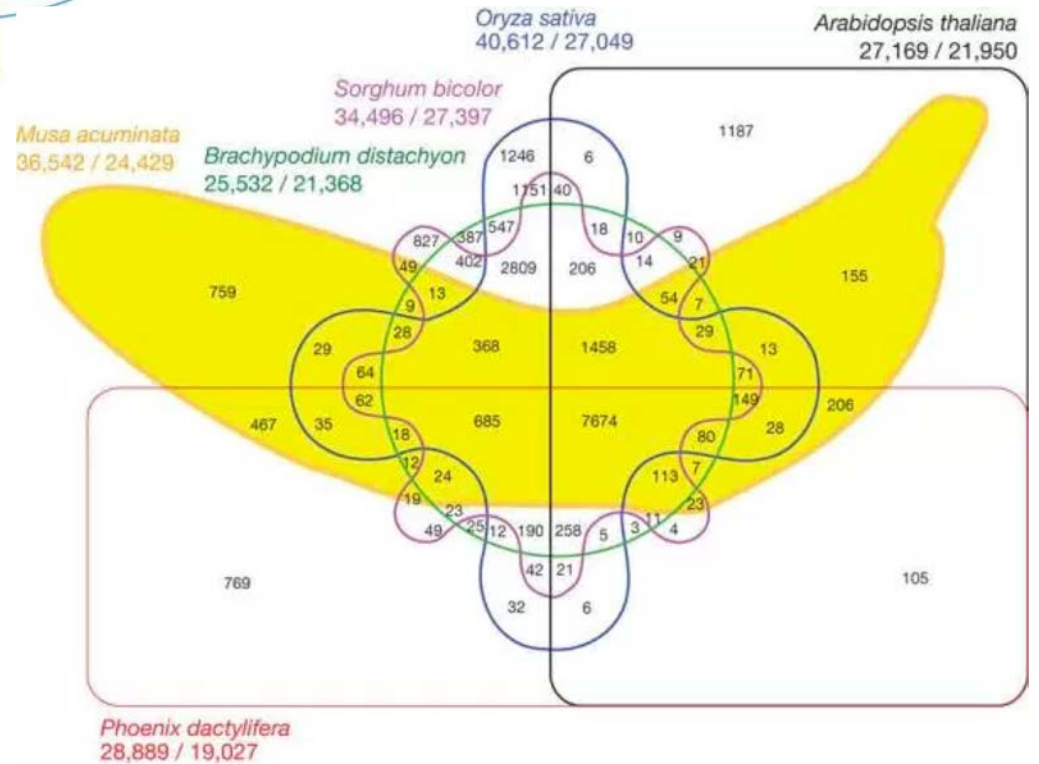


Foth (2014)

Trend of venn diagram…



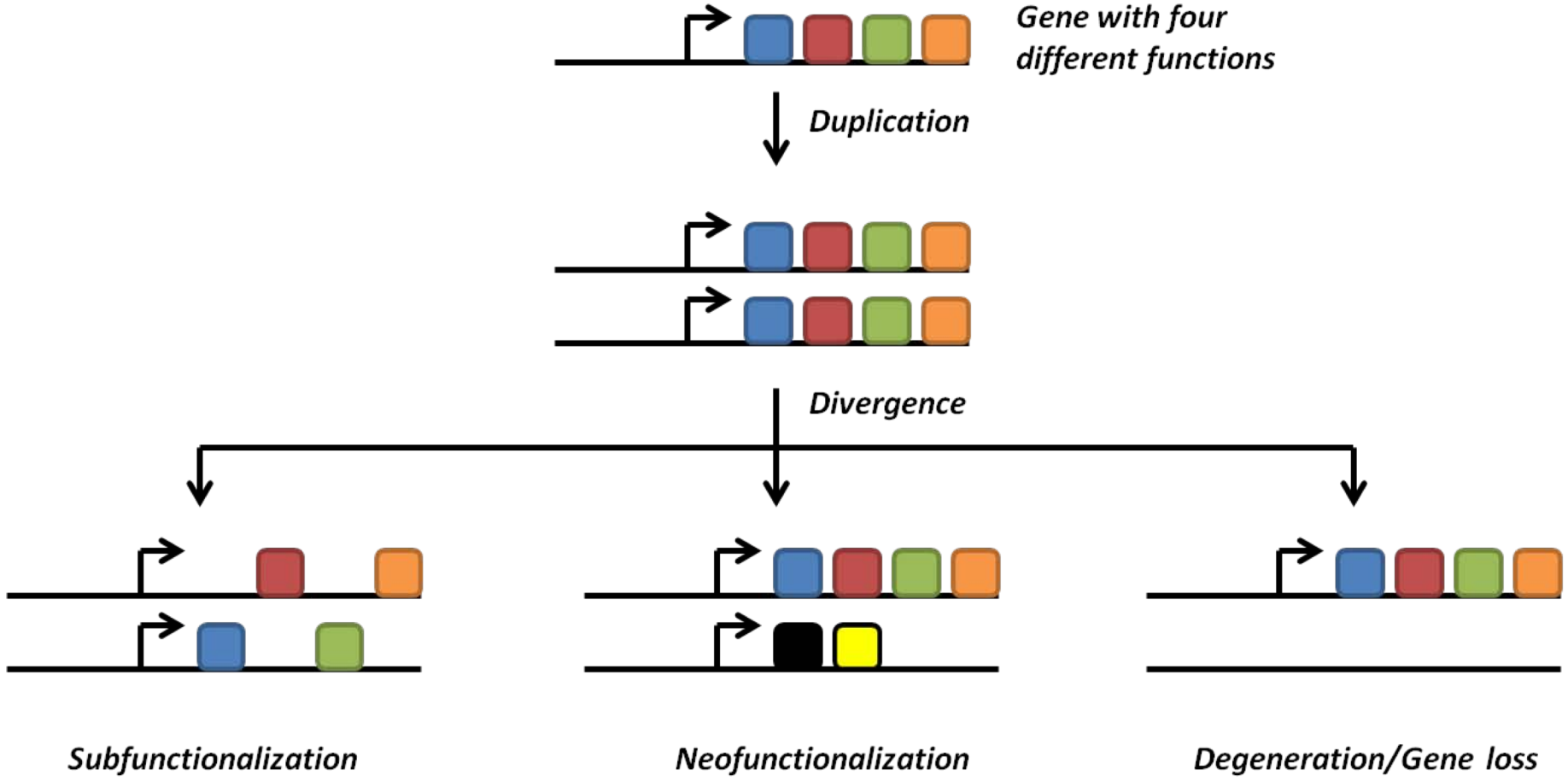




Gene and genome duplication

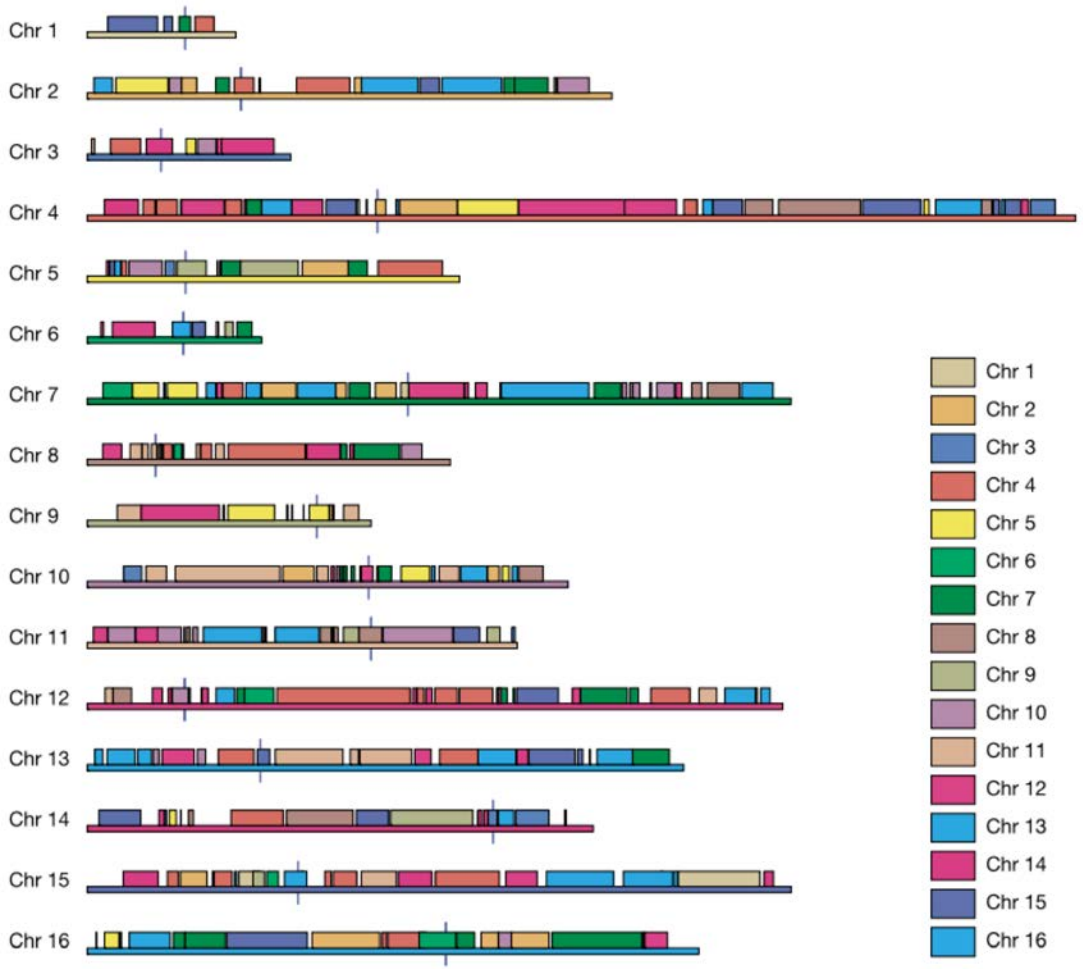
Why study gene duplication?

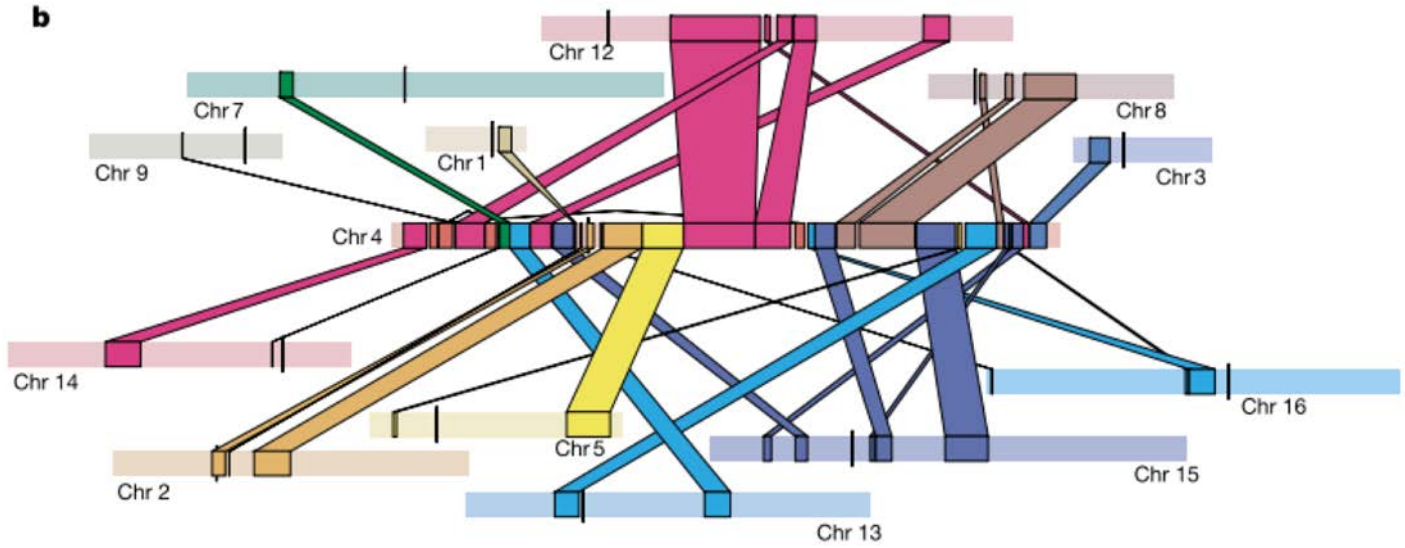
Gene duplications are traditionally considered as a major evolutionary source for protein new functions



Wiki

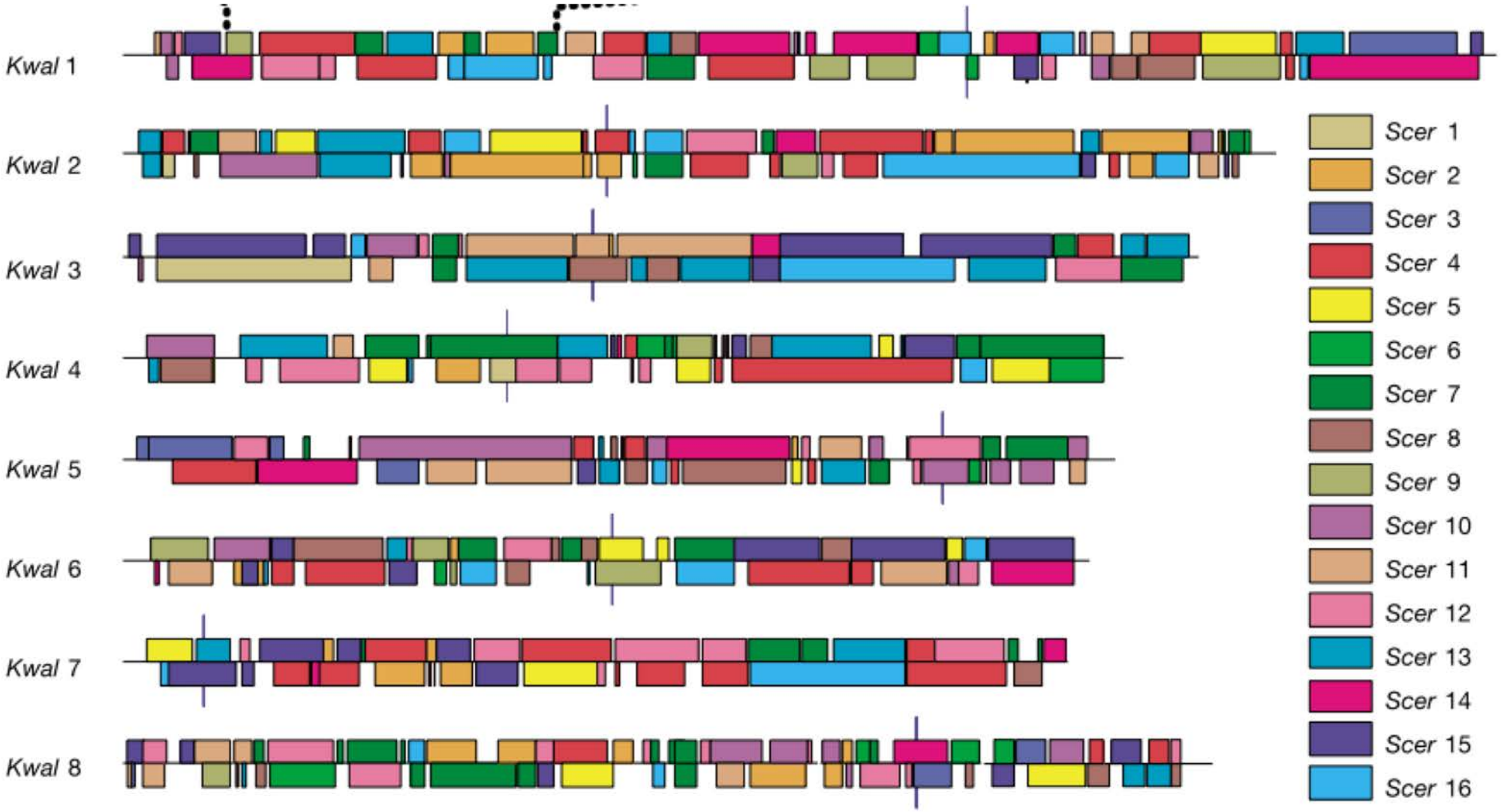
Within species





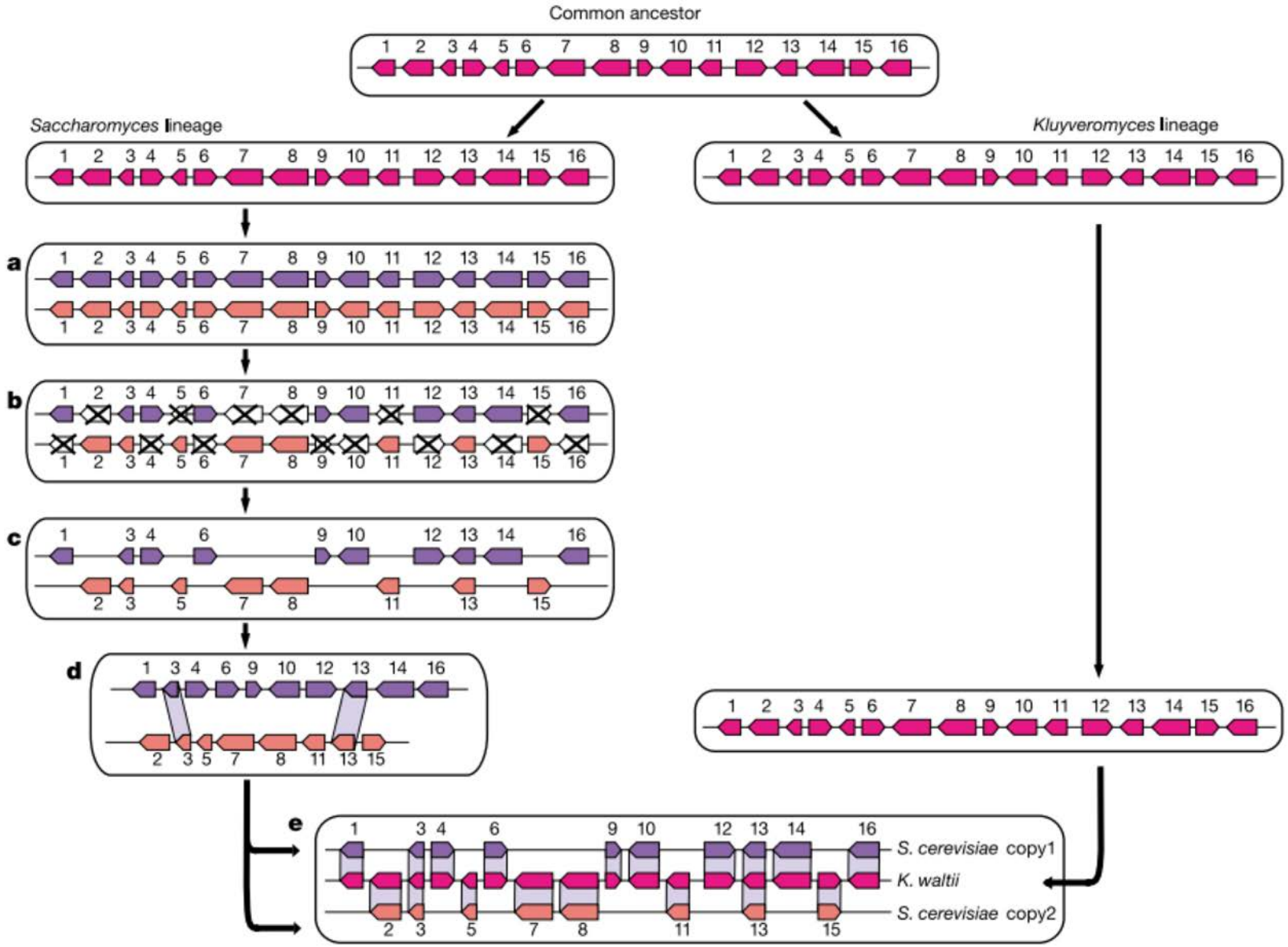
Kellis et al (2004)

Between species



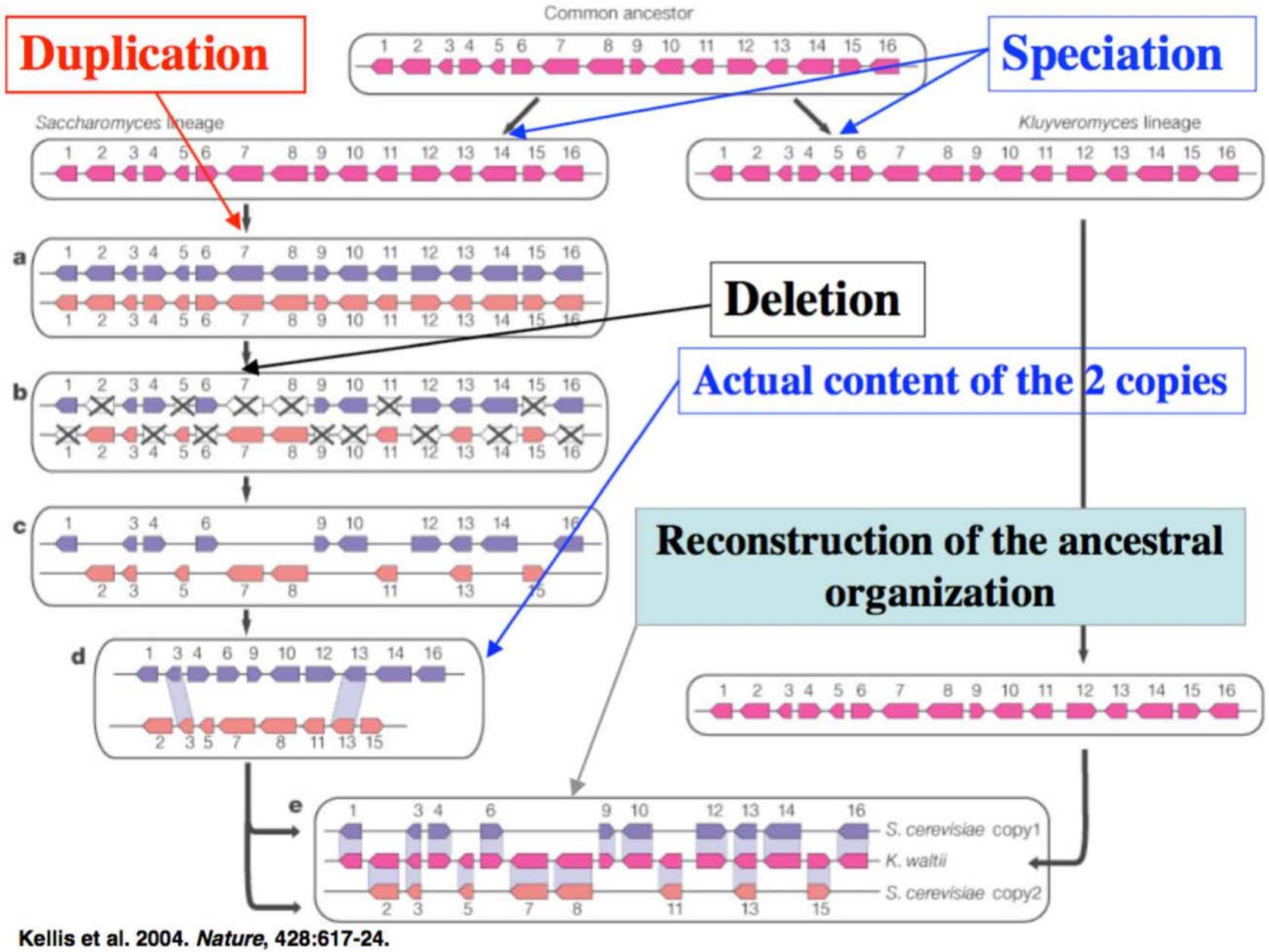
Kellis et al (2004)

Whole genome duplication model



Kellis et al (2004)

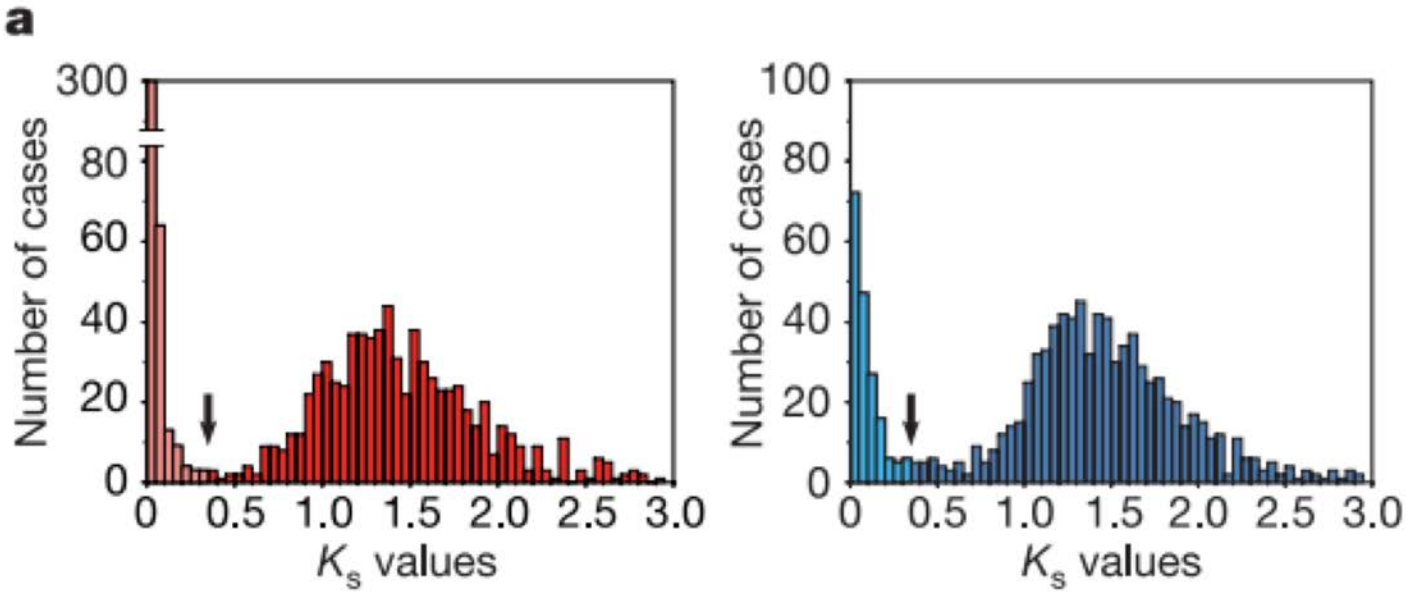
Determining ancestral conservation

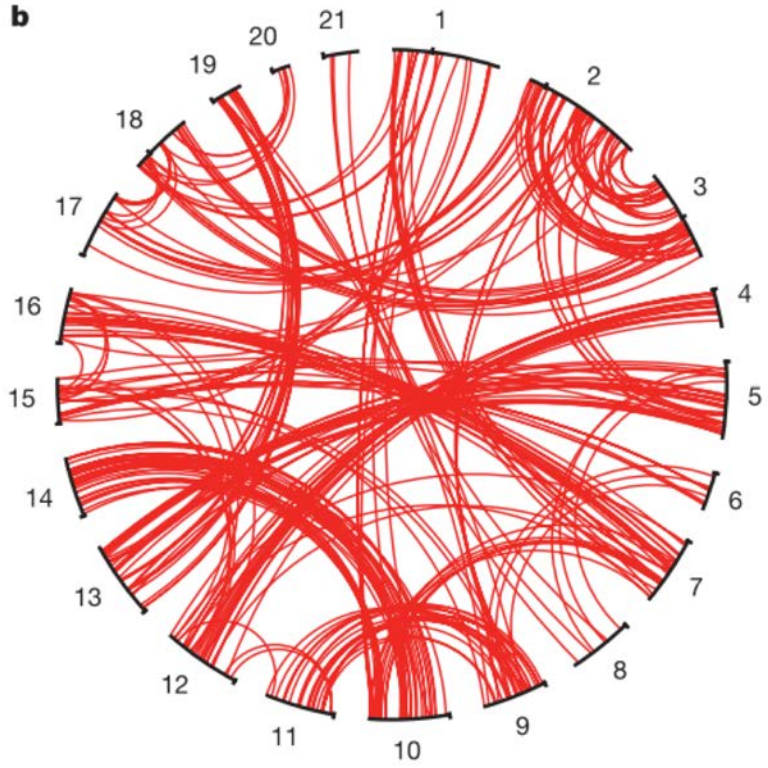


Kellis et al. 2004. *Nature*, 428:617-24.

Slide of Fred Tekaia

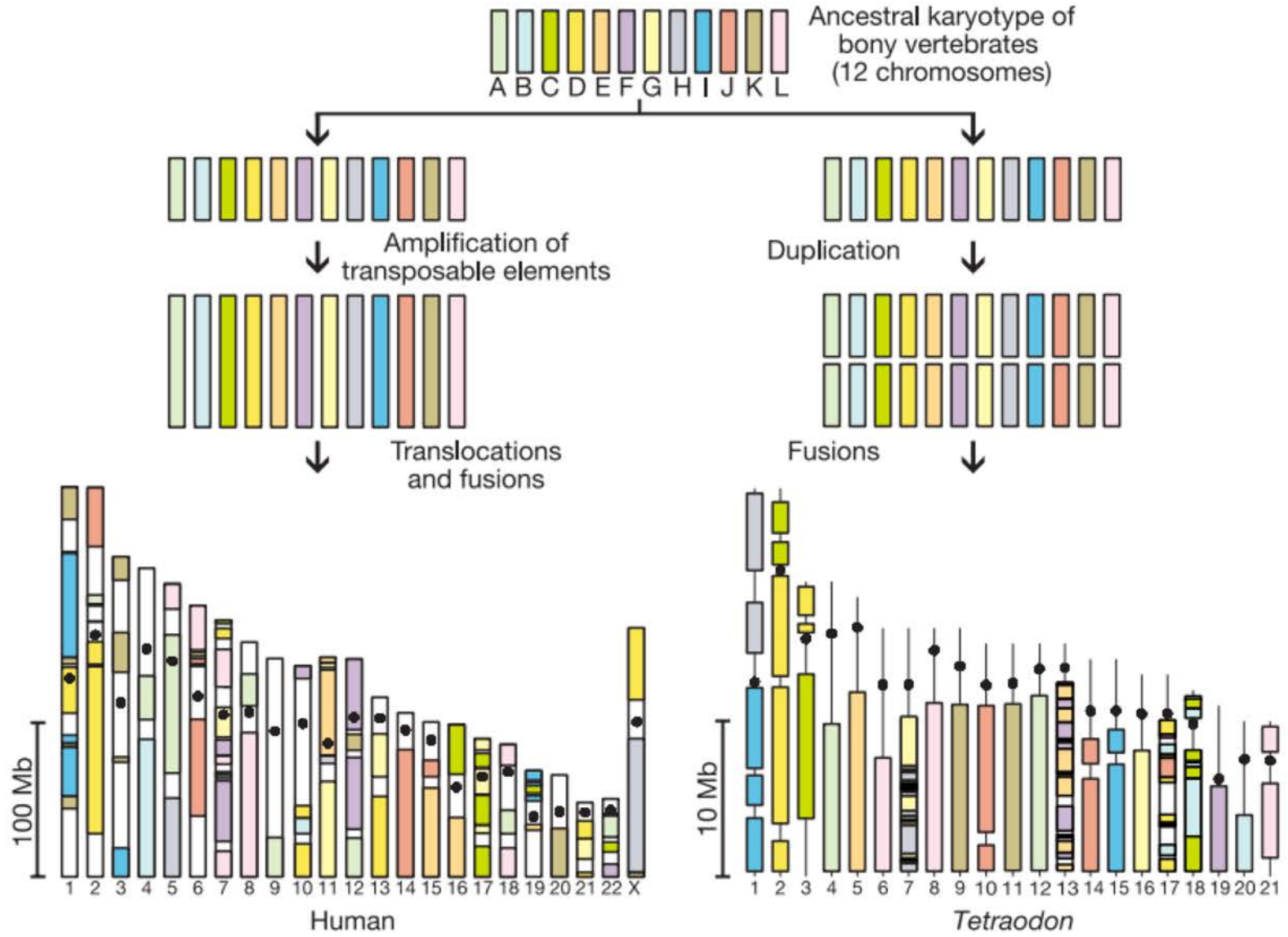
Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype





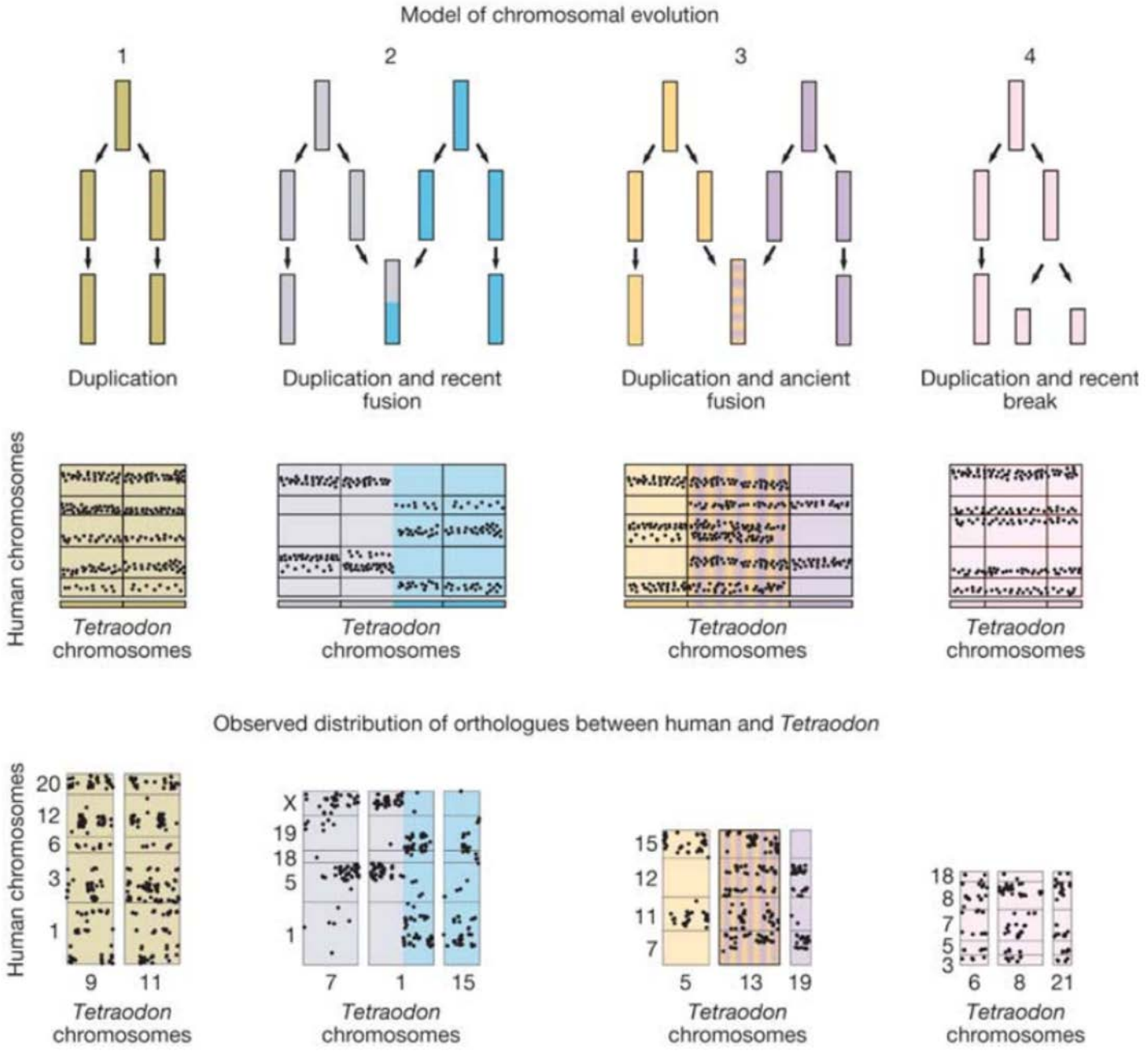
Jaillon et al (2004)

Reconstructing ancient genome rearrangement



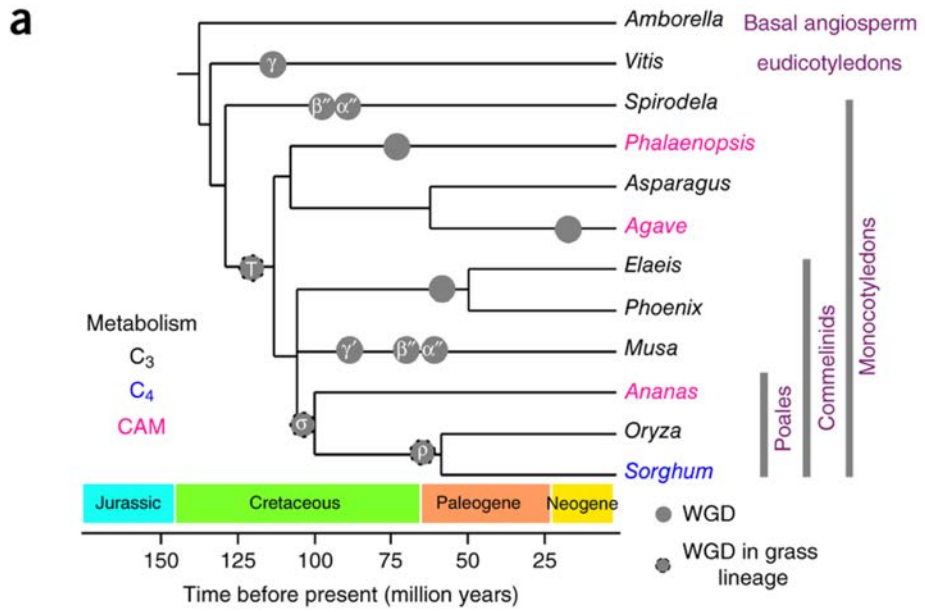
Jaillon et al (2004)

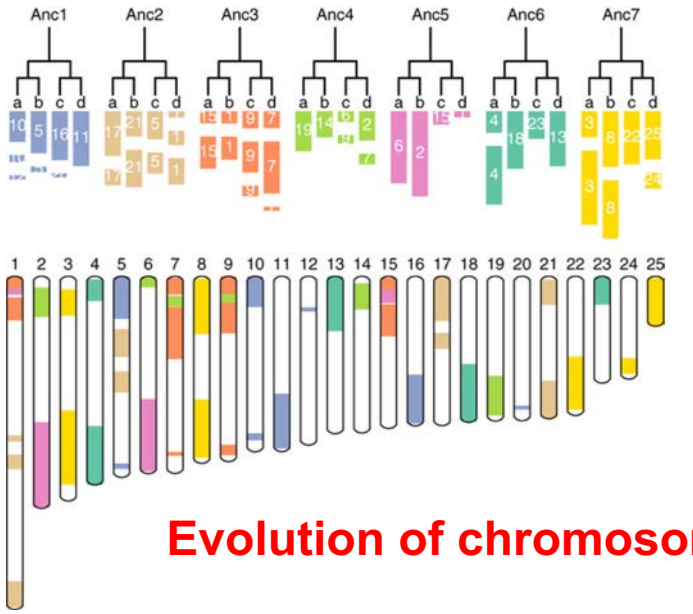
Reconstructing ancient genome rearrangement



Jaillon et al (2004)

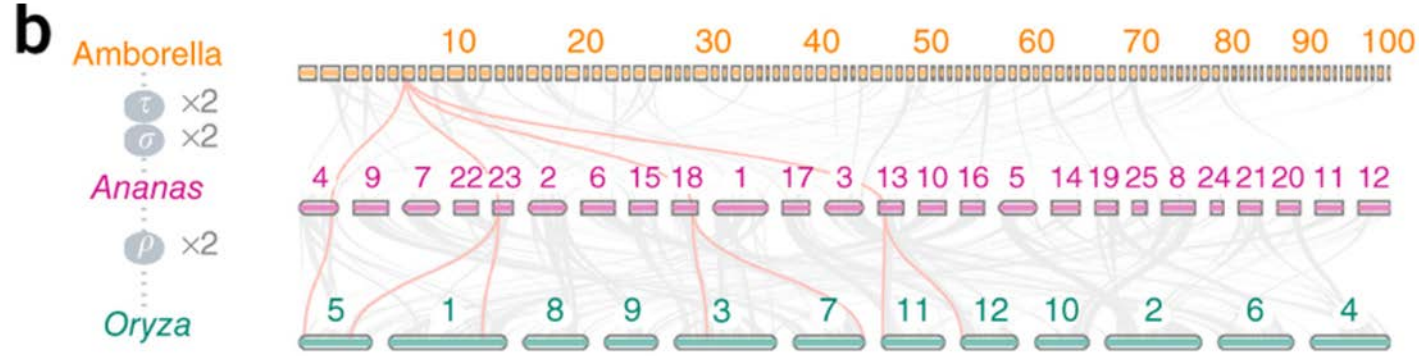
Pineapple genome

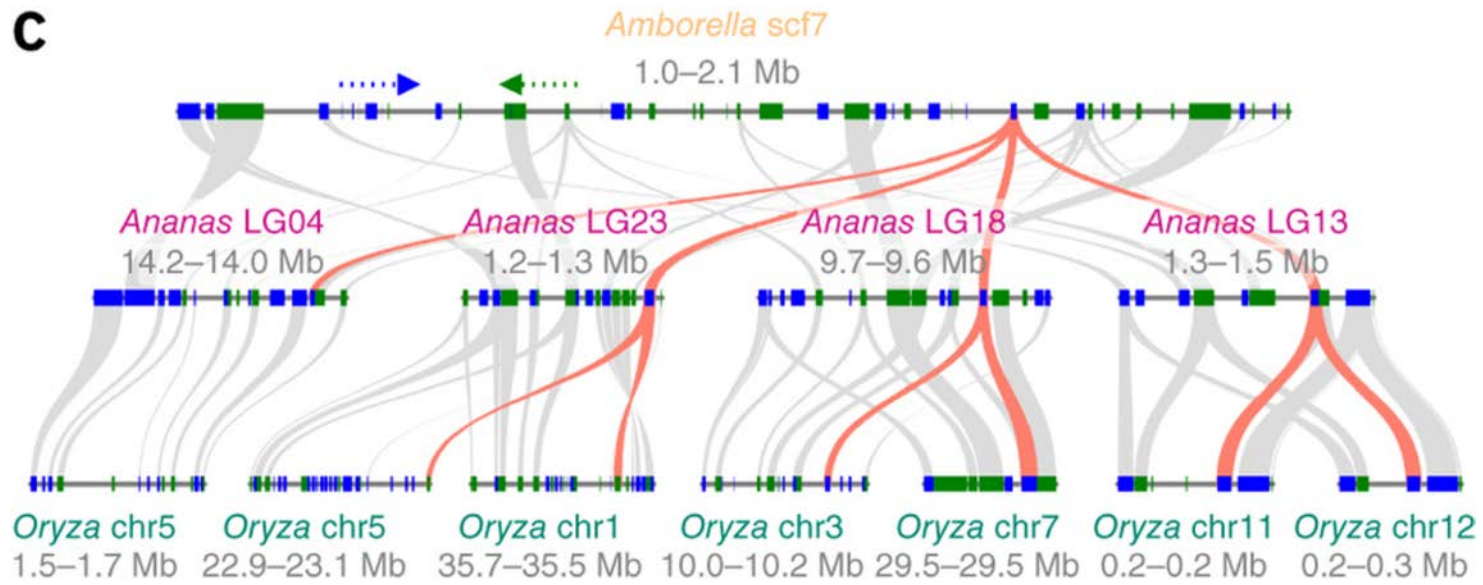




Evolution of chromosomes

Genome alignment





Colinearity of genes

Ming et al (2015)

Case study: lost of gene families

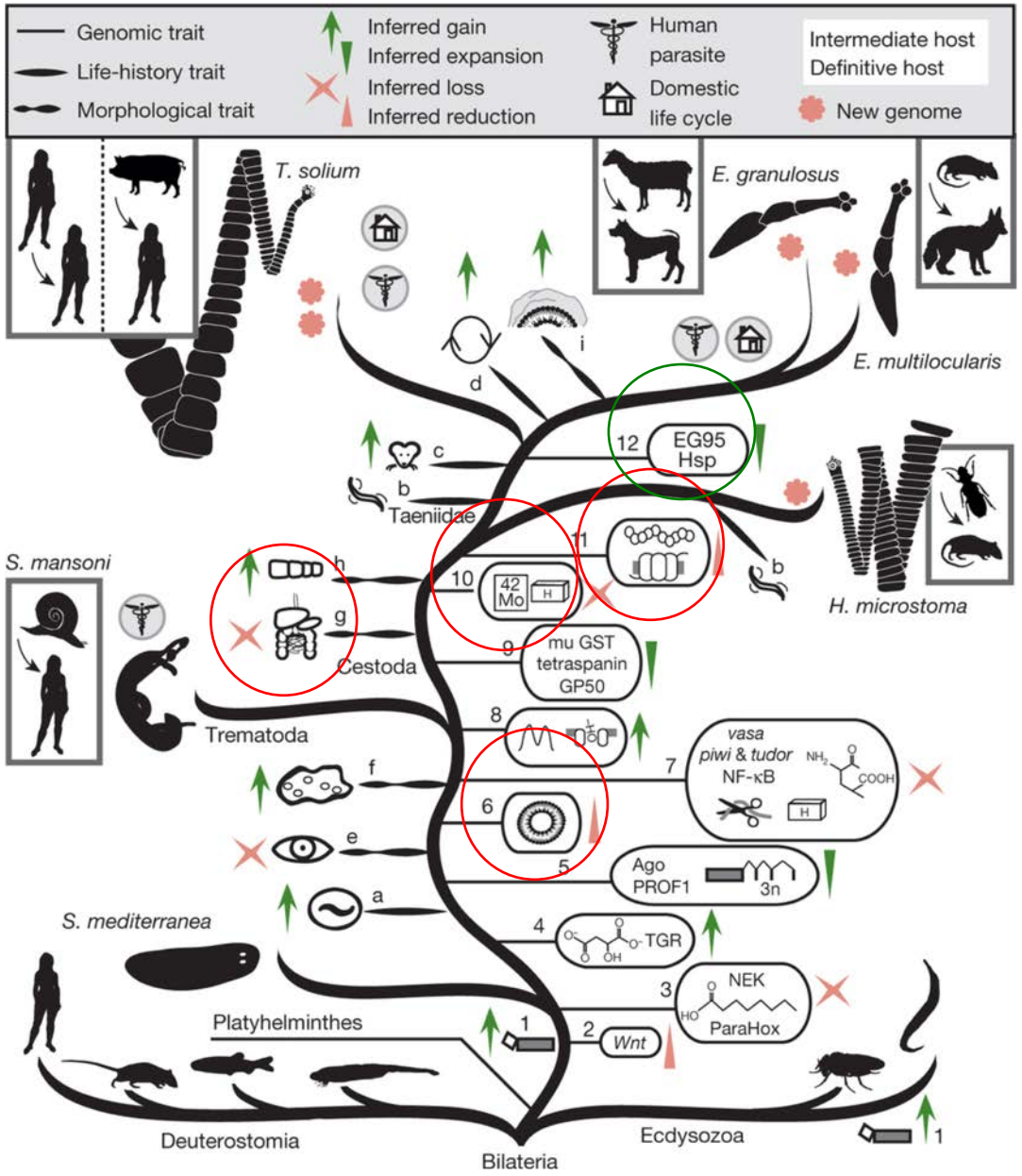
Comparative genomics of tapeworms

tapeworms

Blood fluke

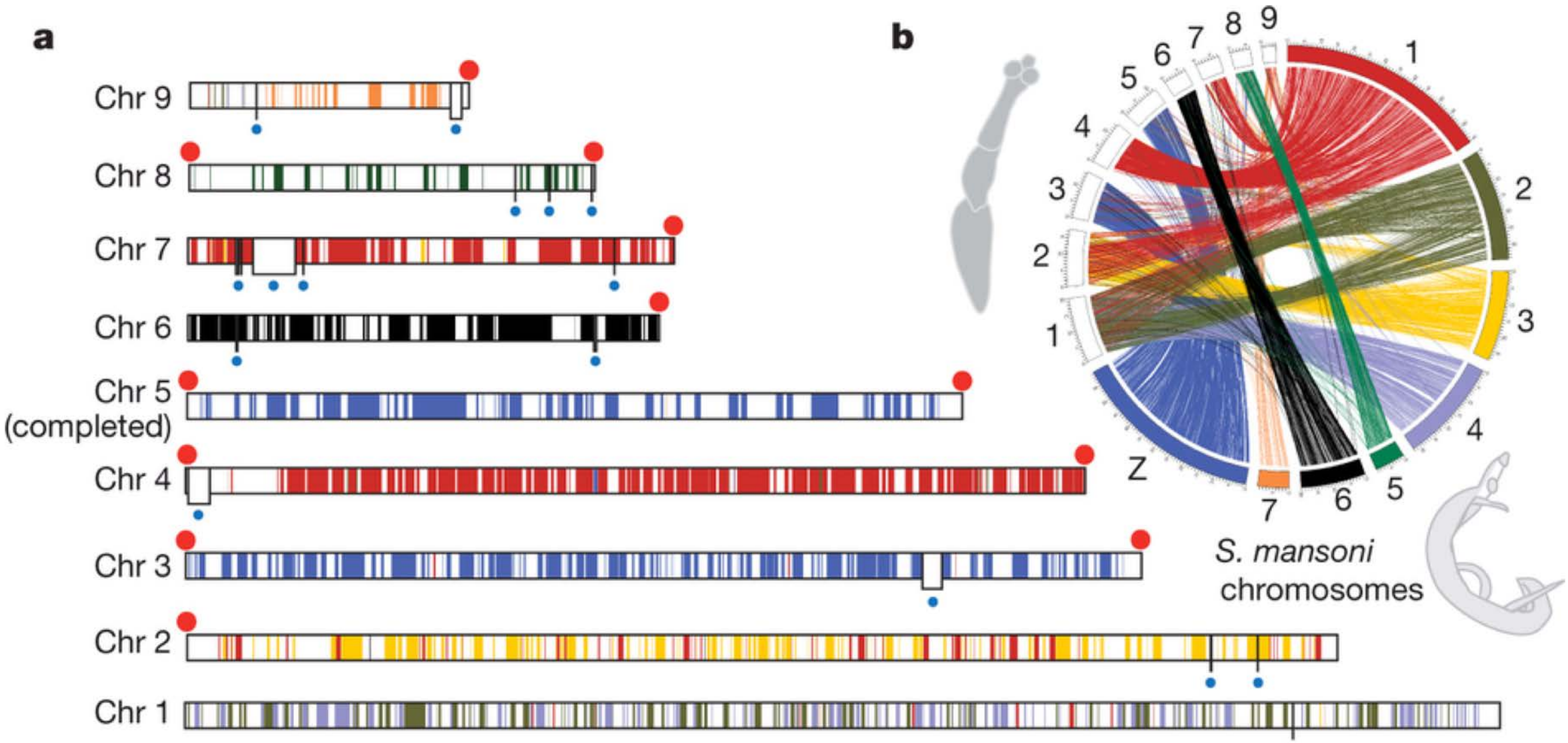
Free-living

Model

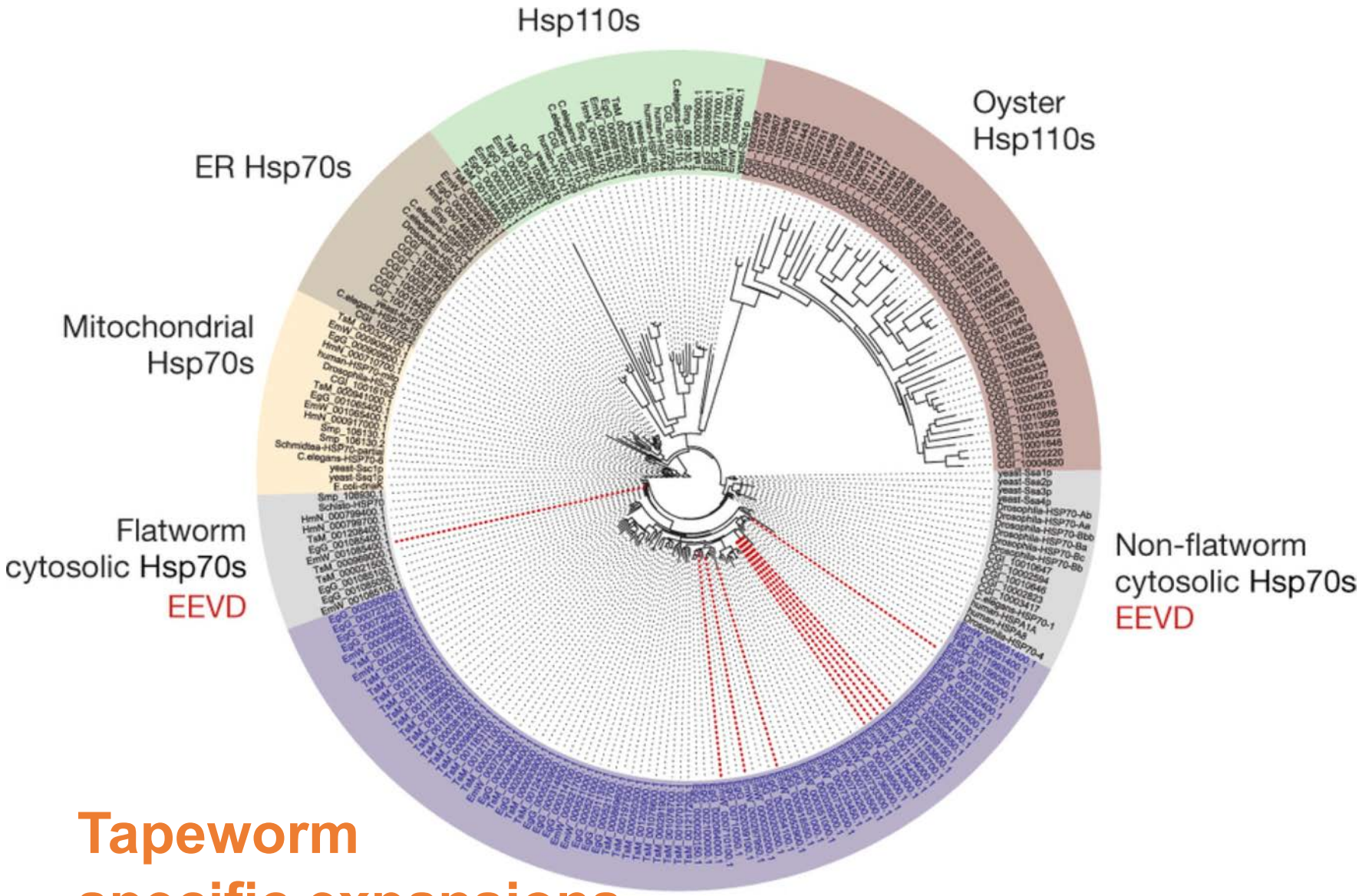


- A total of four tapeworm genomes were sequenced
- We compare with free-living and other parasite genomes
- ‘A route’ to complete parasitism

Genome of *E. multilocularis*

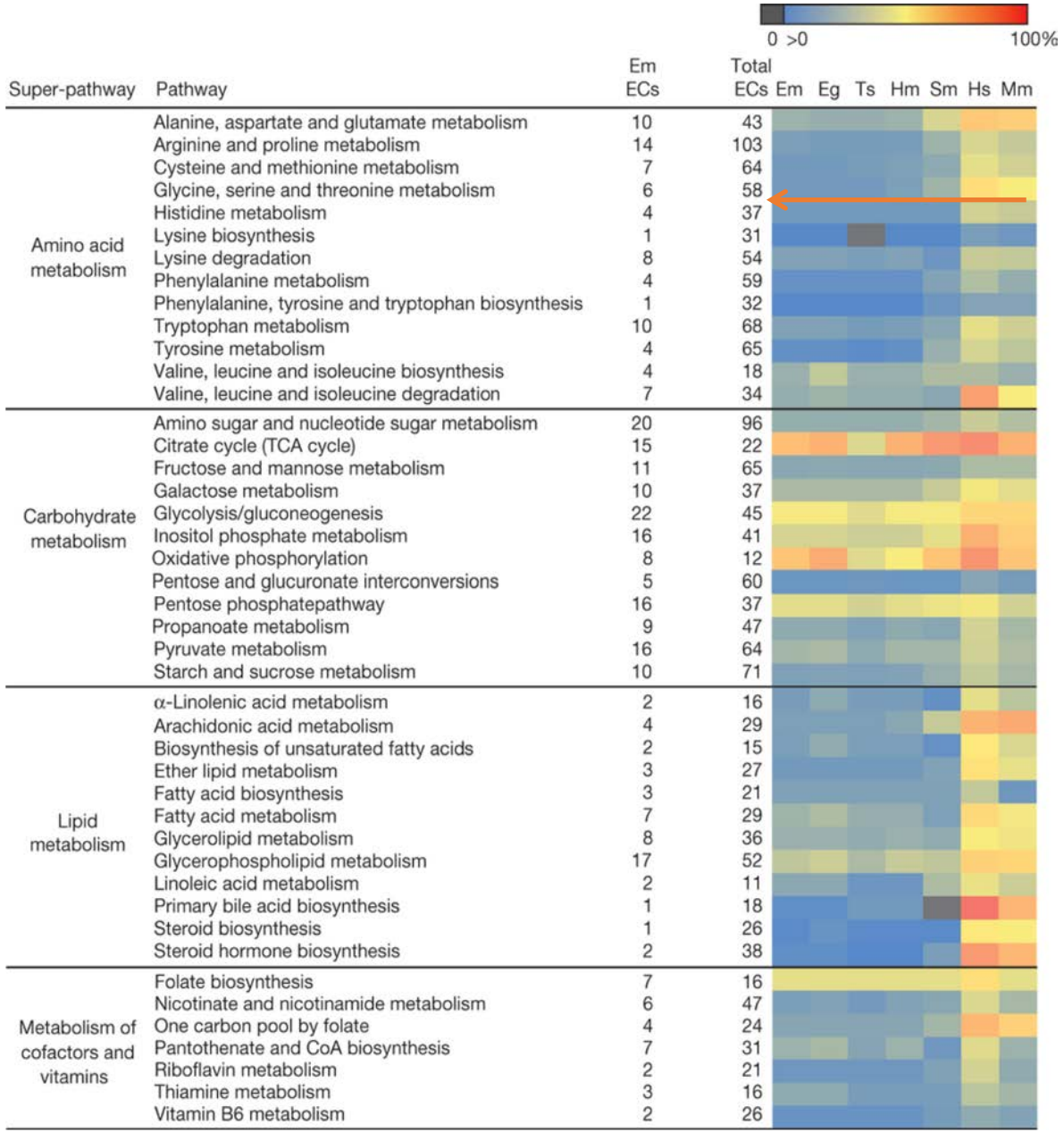


Heat shock protein expansion in tapeworms



Tapeworm specific expansions

Reduced metabolism in tapeworms



Super-pathway	Pathway	Em ECs	Total ECs
Amino acid metabolism	Alanine, aspartate and glutamate metabolism	10	43
	Arginine and proline metabolism	14	103
	Cysteine and methionine metabolism	7	64
	Glycine, serine and threonine metabolism	6	58
	Histidine metabolism	4	37
	Lysine biosynthesis	1	31
	Lysine degradation	8	54
	Phenylalanine metabolism	4	59
	Phenylalanine, tyrosine and tryptophan biosynthesis	1	32
	Tryptophan metabolism	10	68
	Tyrosine metabolism	4	65
	Valine, leucine and isoleucine biosynthesis	4	18
	Valine, leucine and isoleucine degradation	7	34
Carbohydrate metabolism	Amino sugar and nucleotide sugar metabolism	20	96
	Citrate cycle (TCA cycle)	15	22
	Fructose and mannose metabolism	11	65
	Galactose metabolism	10	37
	Glycolysis/gluconeogenesis	22	45
	Inositol phosphate metabolism	16	41
	Oxidative phosphorylation	8	12
	Pentose and glucuronate interconversions	5	60
	Pentose phosphatepathway	16	37
	Propanoate metabolism	9	47
	Pyruvate metabolism	16	64
	Starch and sucrose metabolism	10	71
Lipid metabolism	α-Linolenic acid metabolism	2	16
	Arachidonic acid metabolism	4	29
	Biosynthesis of unsaturated fatty acids	2	15
	Ether lipid metabolism	3	27
	Fatty acid biosynthesis	3	21
	Fatty acid metabolism	7	29
	Glycerolipid metabolism	8	36
	Glycerophospholipid metabolism	17	52
	Linoleic acid metabolism	2	11
	Primary bile acid biosynthesis	1	18
	Steroid biosynthesis	1	26
	Steroid hormone biosynthesis	2	38
Metabolism of cofactors and vitamins	Folate biosynthesis	7	16
	Nicotinate and nicotinamide metabolism	6	47
	One carbon pool by folate	4	24
	Pantothenate and CoA biosynthesis	7	31
	Riboflavin metabolism	2	21
	Thiamine metabolism	3	16
	Vitamin B6 metabolism	2	26

Reduced metabolism

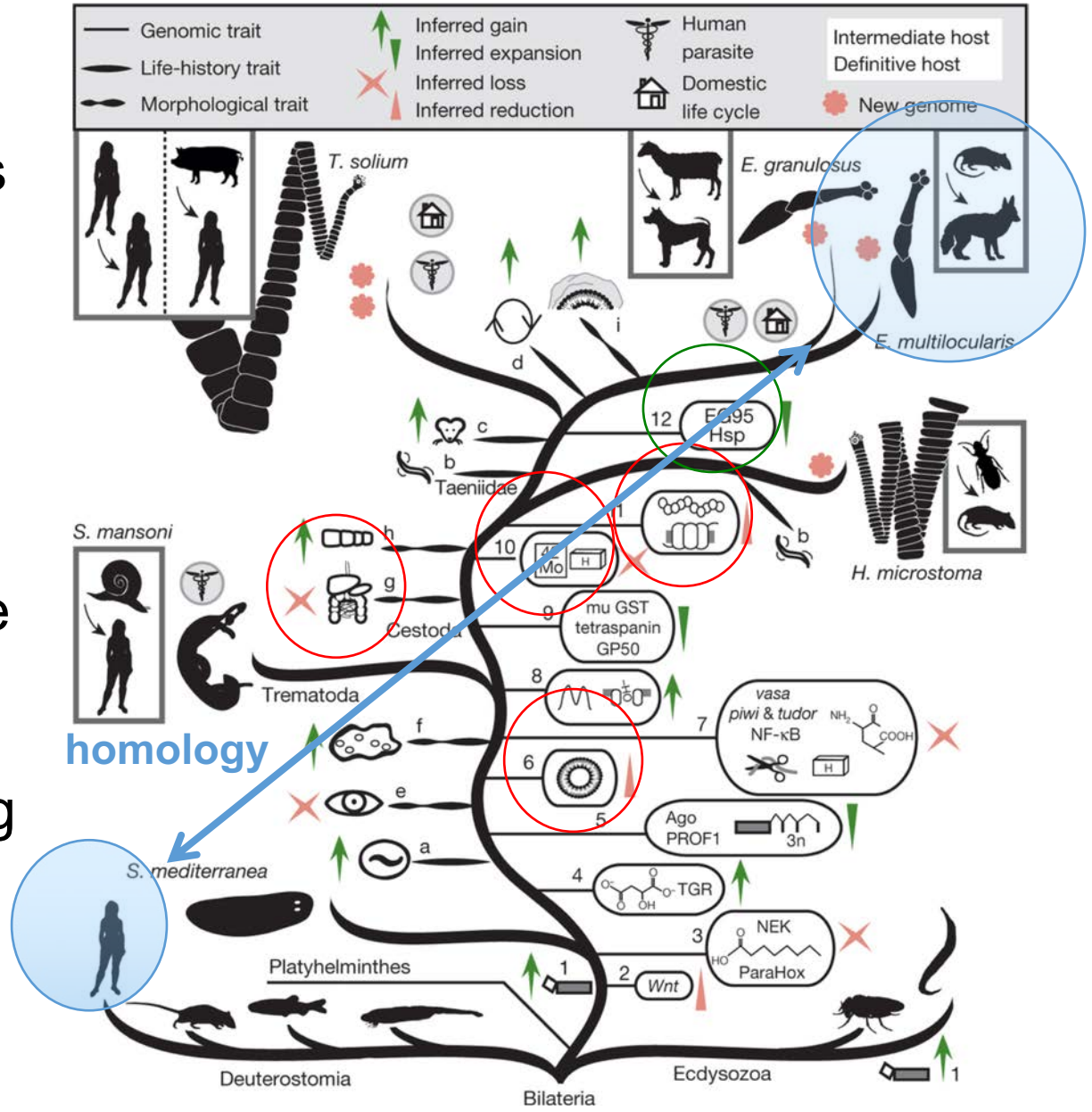
Tapeworm’s road to parasitism

tapeworms

Blood fluke

Free-living

Model

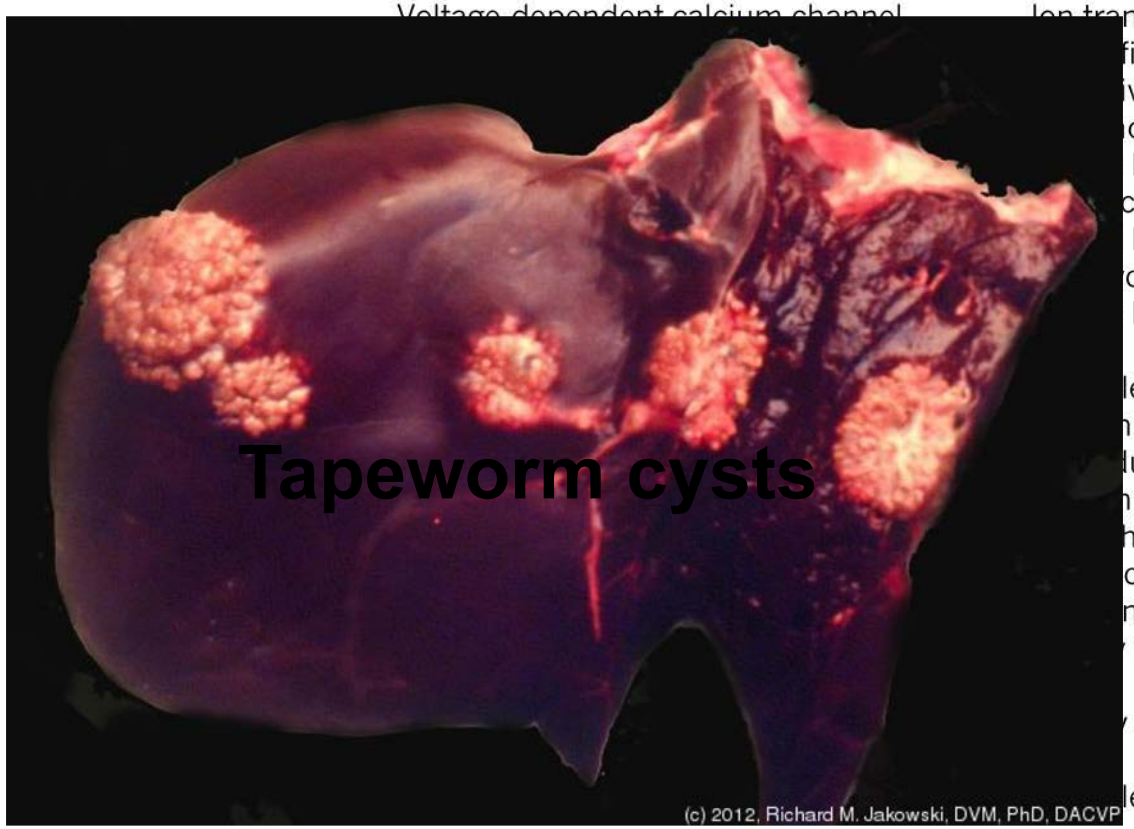


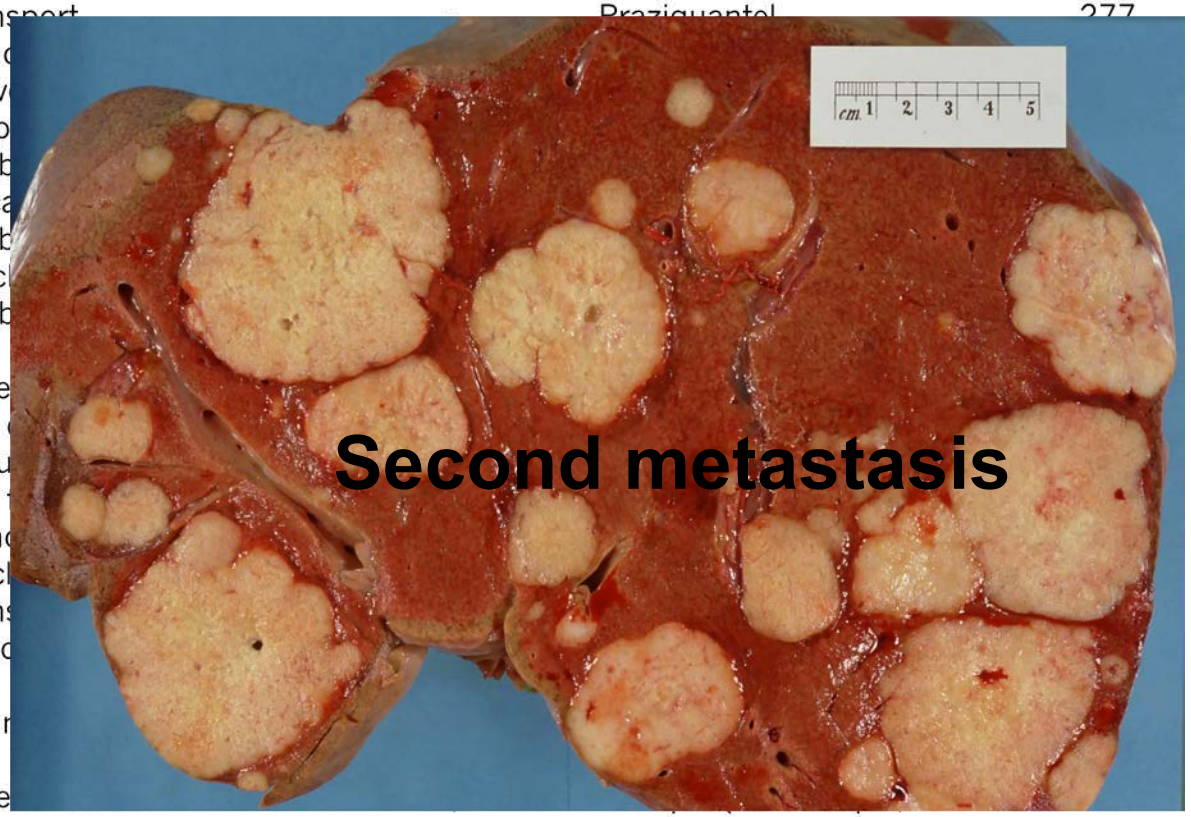
Predict candidate drugs

Promising drug targets in tapeworms

Table 1 | Top 20 promising targets in *E. multilocularis*

Target category	Target	Action	Expression	Drug	Rank
Current targets	Tubulin β-chain	Cytoskeleton	M,A	Albendazole	406
	Voltage-dependent calcium channel	Ion transport		Praziquantel	277
				Lorazepam)	
	Elongation factor 2	Translation	M,A	Experimental compounds	54
	Cathepsin B	Protease	M	Experimental compounds	55
	Dual-specificity mitogen activated protein	Signalling, activation of p38	M	Experimental compounds	56
	Purine nucleoside phosphorylase	Purine metabolism	M,A	Didanosine	63





http://en.wikipedia.org/wiki/Metastasis
http://ocw.tufts.edu/data/

Comparing genomes beyond gene (copy) numbers

Extension of homology to genomes

Gene family gains and losses in previous lecture

Comparing genomes at **different resolution**

- Synteny (gene content on the same chromosome)
- Colinearity (gene content + order conservation)
- DNA-based alignments (base-to-base mapping)

Extension of homology to genomes: synteny

Synteny Conservation and Chromosome Rearrangements During Mammalian Evolution

Jason Ehrlich,*,[1] David Sankoff† and Joseph H. Nadeau*,[2]

**Jackson Laboratory, Bar Harbor, Maine 04609 and †Centre de recherches mathématiques, Université de Montréal, Montréal, Québec, H3C 3J7 Canada*

Manuscript received December 13, 1996
Accepted for publication June 4, 1997

MAPS of LINKAGE and SYNTENY HOMOLOGIES between MOUSE and MAN

JOSEPH H. NADEAU

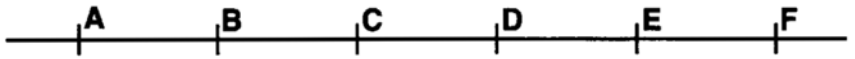
1989

Synteny refers to the occurrence of two or more genes on the same chromosome, whereas *conserved synteny* refers to two or more homologous genes that are syntenic in two or more species, regardless of gene order on each chromosome, *i.e.*, synteny but not necessarily gene order is conserved (Figure 2; see also NADEAU 1989). *Conserved linkage* pertains to the conservation of both synteny and order of homologous genes between species (Figure 2; see also NADEAU 1989). A *disrupted synteny* refers to circumstances where a pair of genes are located on the same chromosome in one species but their homologues are located on different chromosomes in another species, *i.e.*, the genes are syntenic in only one of the two species. Syntenic genes can be identified by examining published genetic maps and conserved segments can be identified by comparing

Synteny

conservation of gene content

A. Genetic map in reference species



Each unit is gene

Conserved synteny and linkage

Gene arrangement:

Definition: **Same gene order and similar genetic distances.**

Count:
One *conserved linkage* involving genes,
one conserved synteny. involving genes A,B,C,E,F.

Possible cause:
No inter-chromosomal rearrangement.
No intra-chromosomal rearrangement.

Conserved synteny, conserved linkage, disrupted linkage

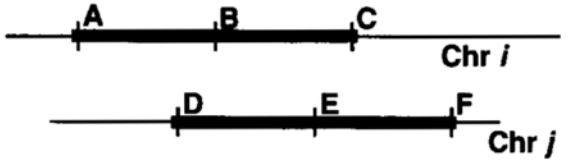
Gene arrangement:

Count:
One *conserved linkage* involving genes B,C,D;
One *conserved linkage* involving genes E,F.
One *disrupted linkage* involving genes B,C,D vs E,F vs A.
One *conserved synteny* involving genes A,B,C,D,E,F.

Possible causes:
An intra-chromosomal rearrangement,
such as a paracentric inversion.

Conserved synteny, disrupted synteny, conserved linkage, disrupted linkage

Gene arrangement:



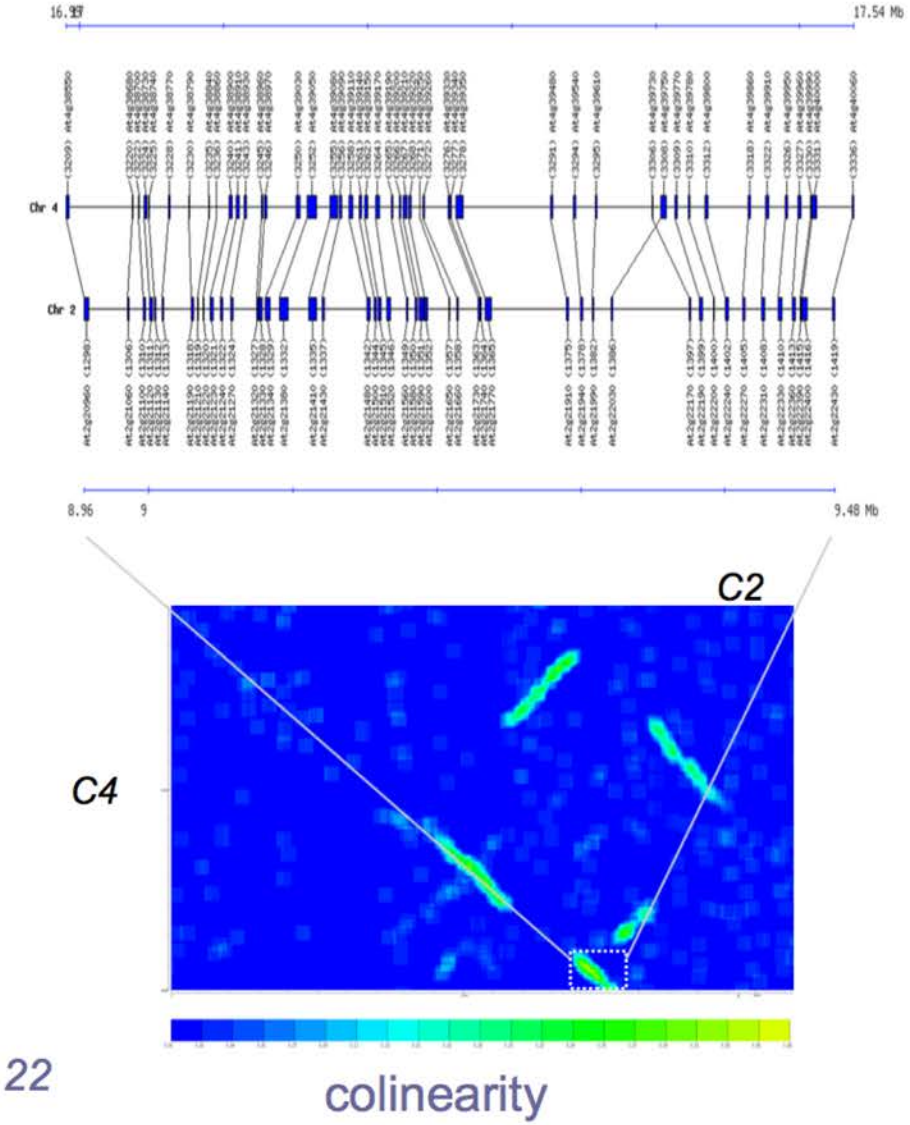
Count:
One *conserved linkage* involving genes A,B,C;
One *conserved linkage* involving genes D,E,F.
One *disrupted linkage* involving genes A,B,C vs D,E,F .
One *conserved synteny* involving genes A,B,C.
One *conserved synteny* involving genes D,E,F.
One *disrupted synteny* involving genes A,B,C vs D,E,F.

Possible causes:
An inter-chromosomal rearrangement,
such as a reciprocal translocation.

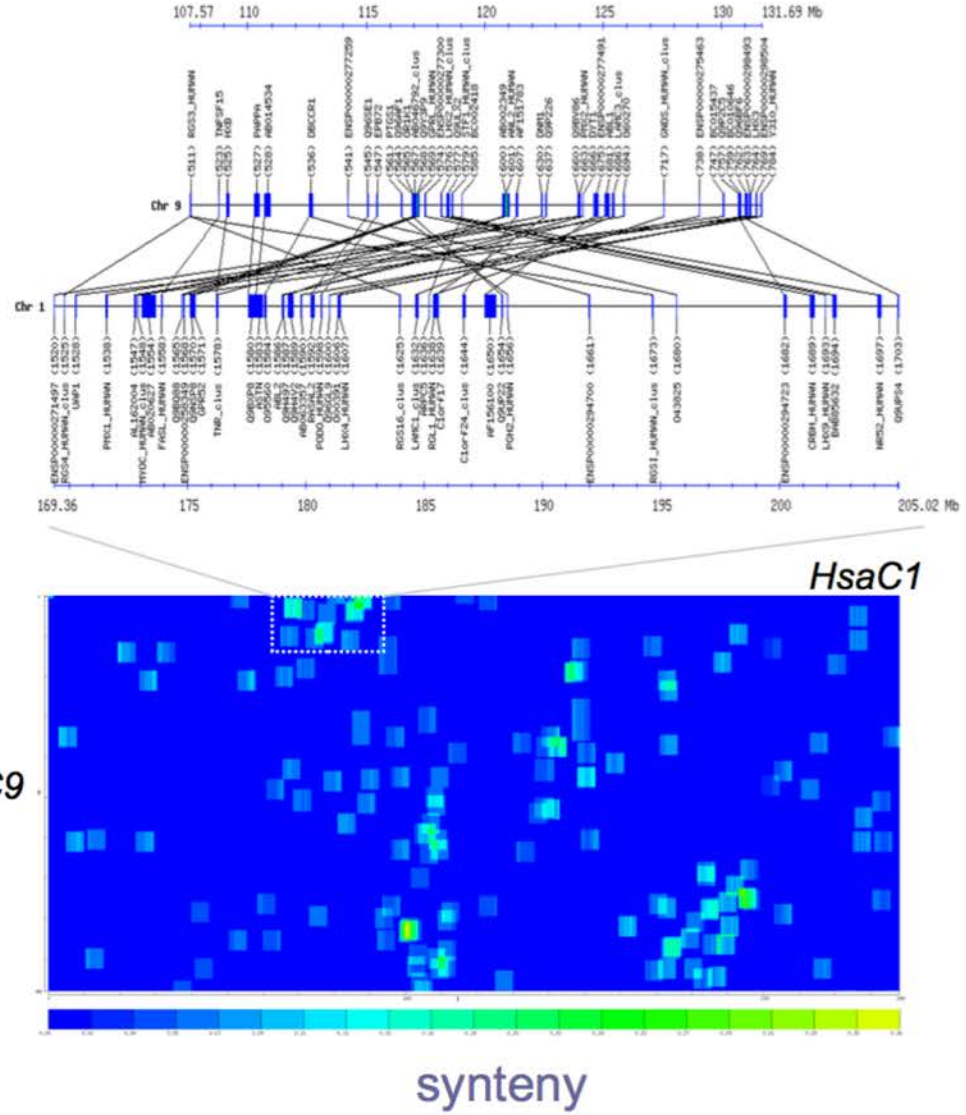
Ehrlich (1997)

Synteny and colinearity

recent duplication

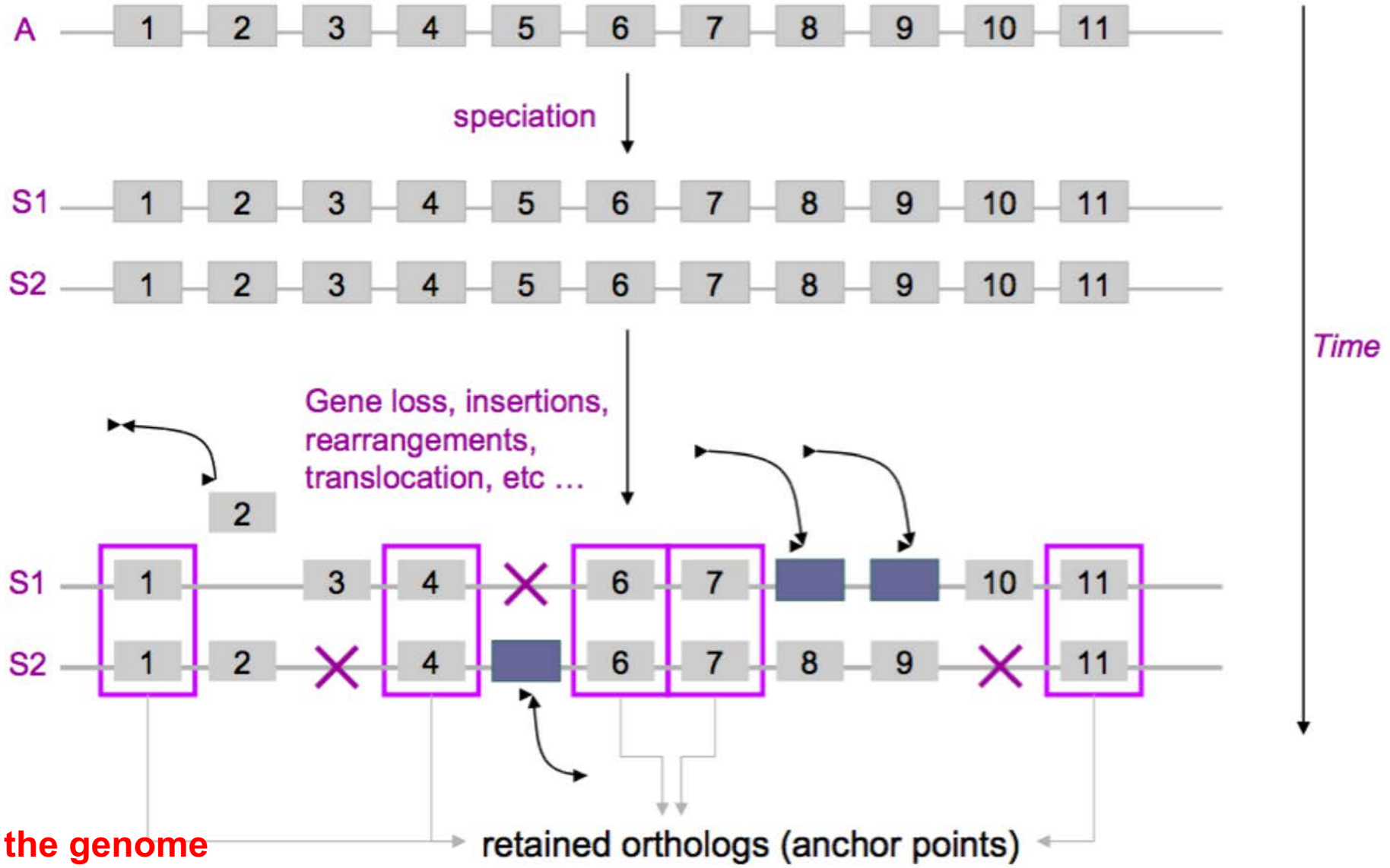


ancient duplication



Slide by Klaas Vandepoele

Inferring gene collinearity



Correctly identify orthologs

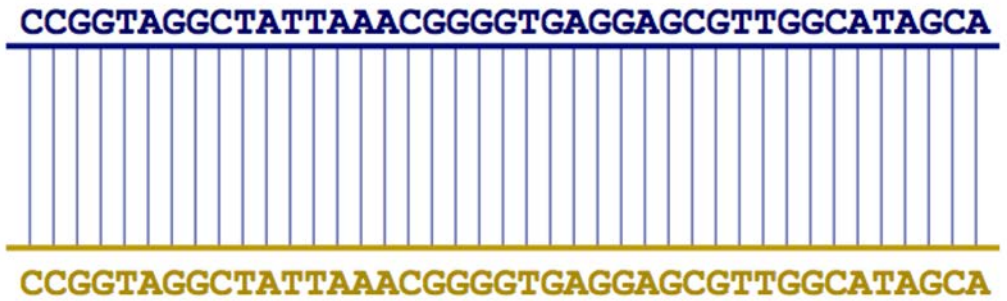
+

Determine their position on the genome

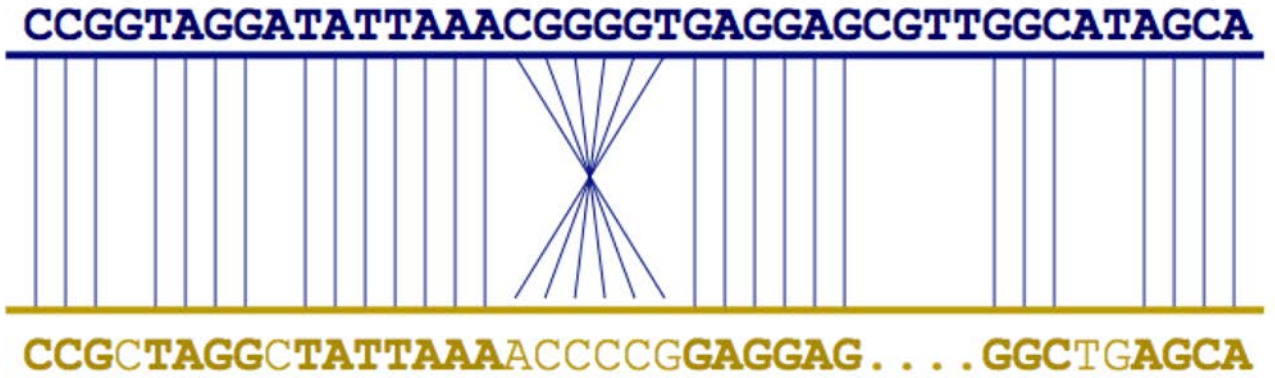
Slide by Klaas Vandepoele

Whole genome alignment

For two genomes, A and B, find a mapping from each position in A to its corresponding position in B

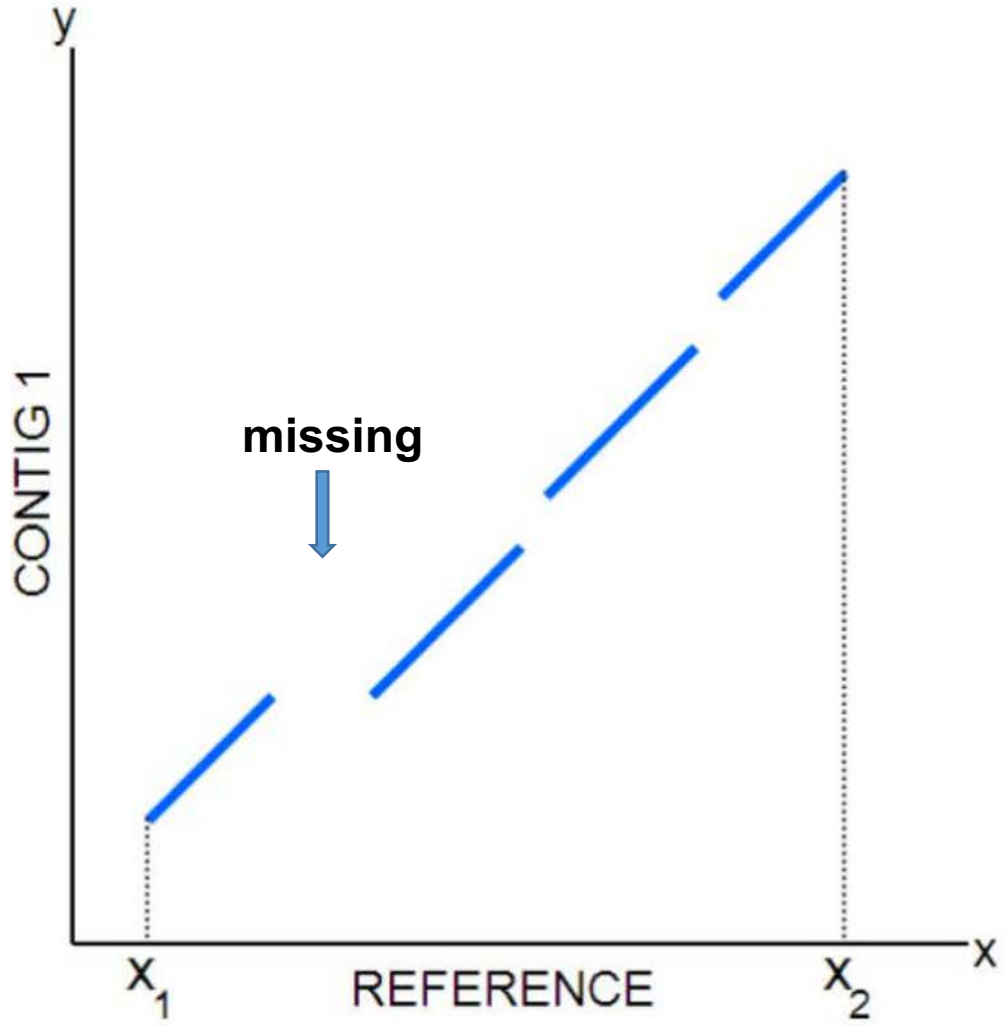


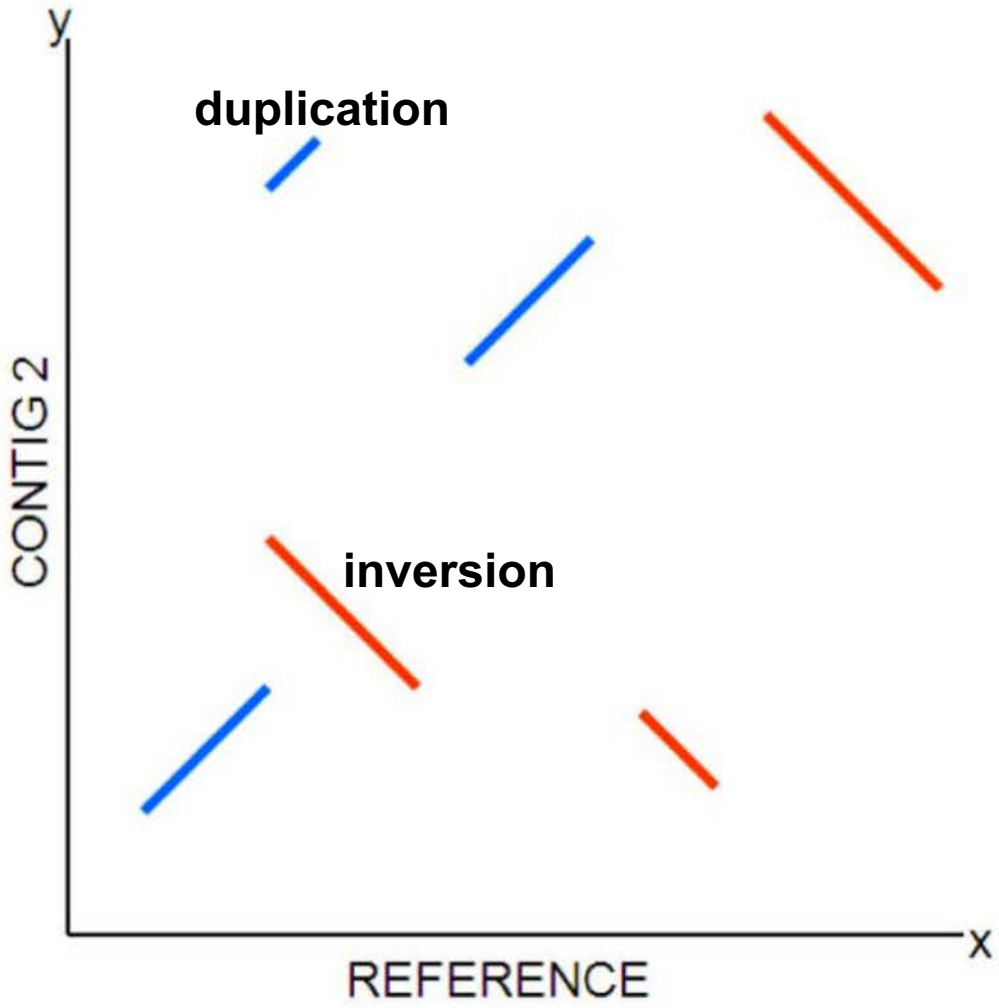
In reality, Genome A may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to B (sometimes all of the above)



Slide of Mike Schatz

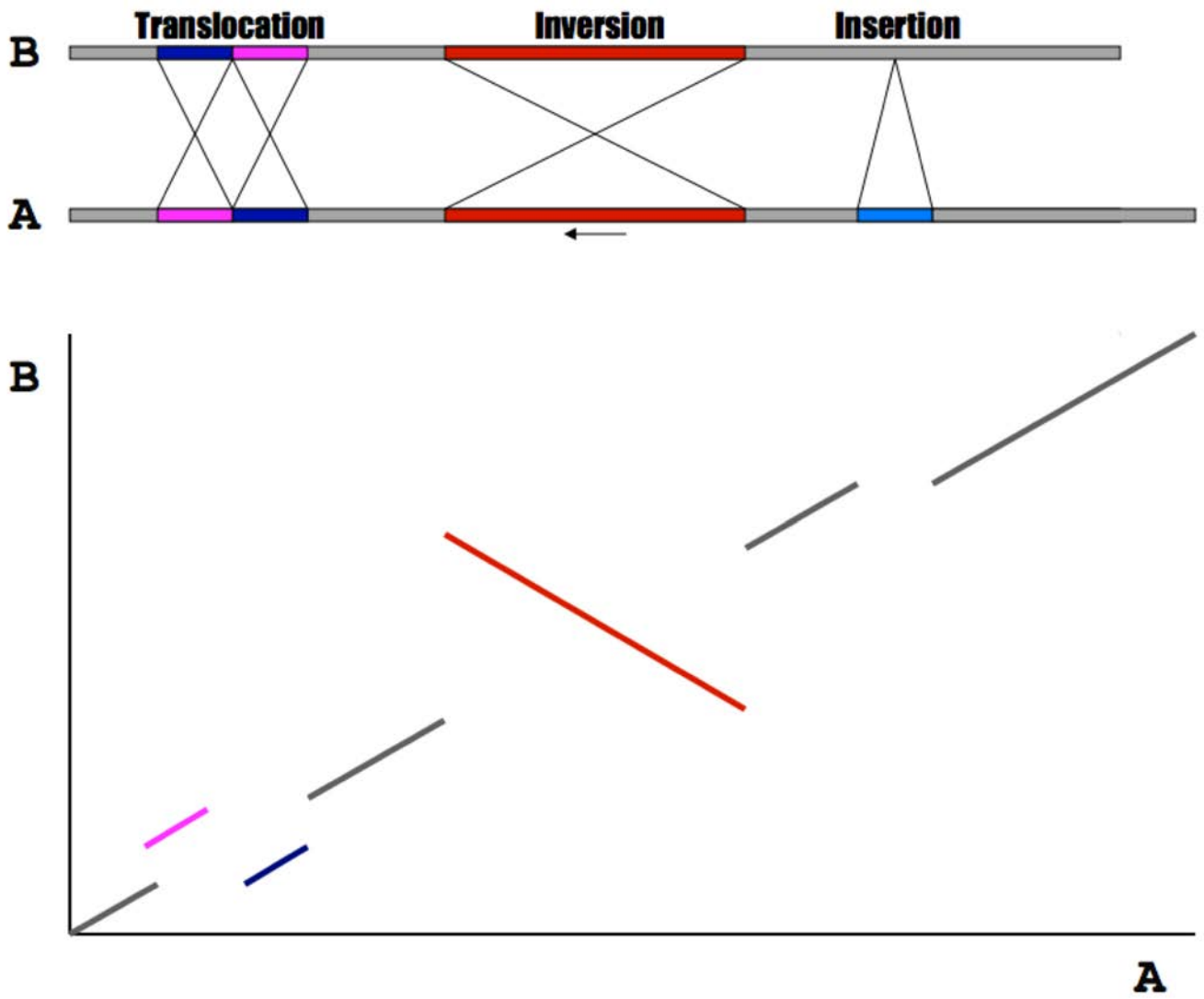
Aligning genome at nucleotide / amino acid level
Visualise through **dotplot**





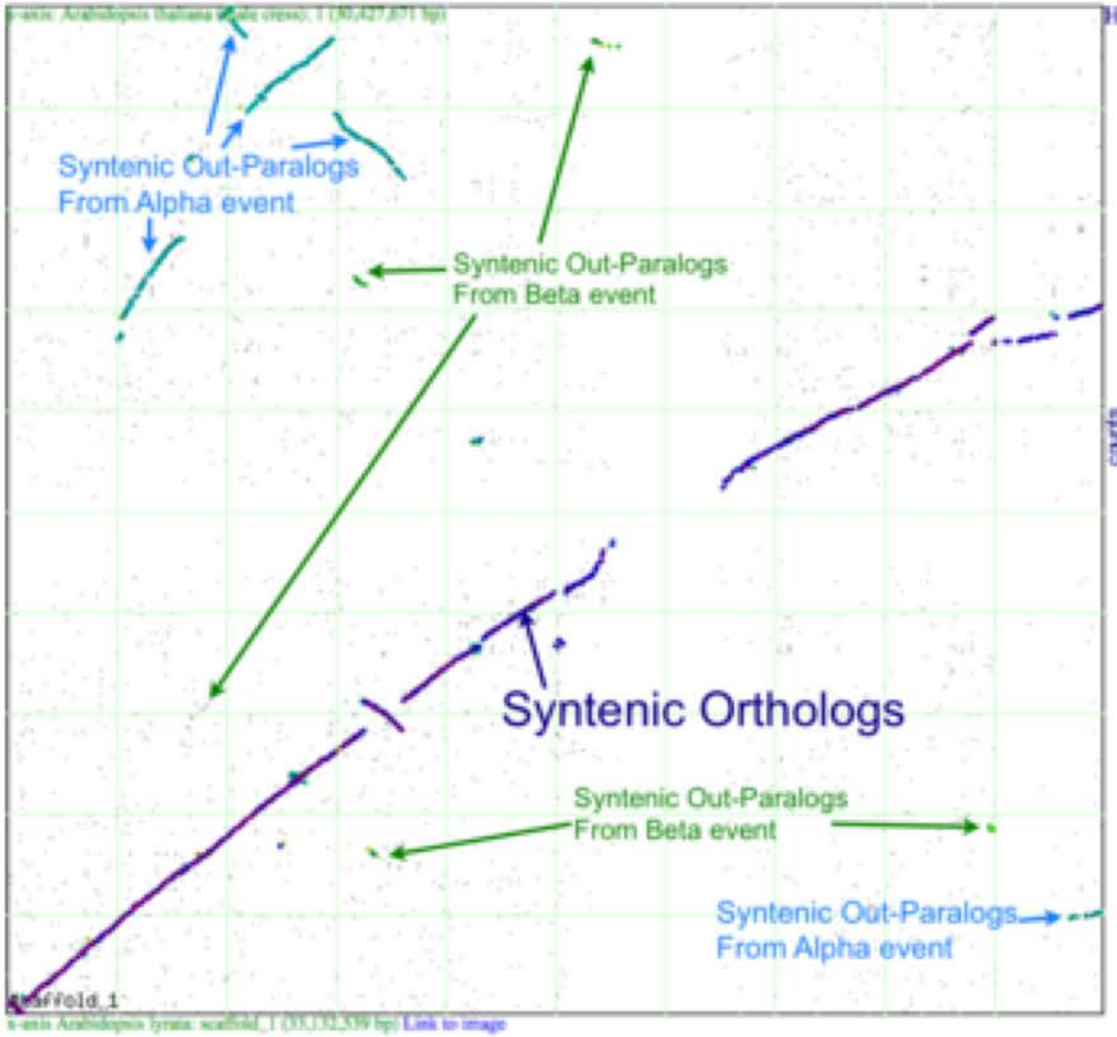
González and Liao (2008)

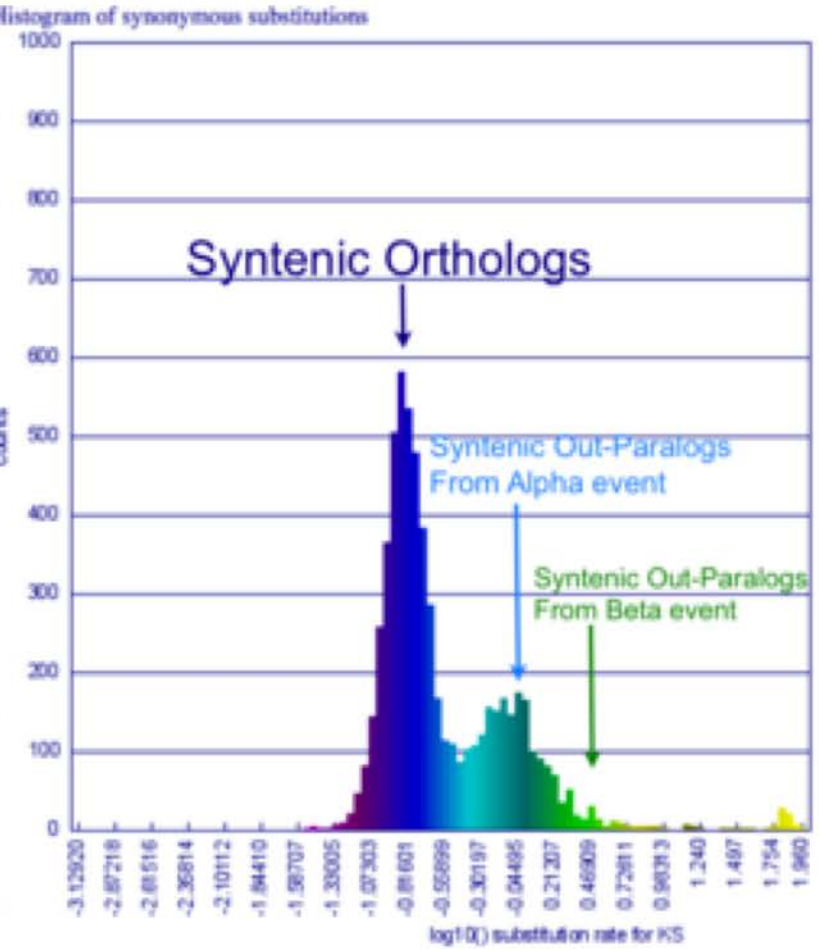
Aligning genome at nucleotide / amino acid level
Visualise through **dotplot**

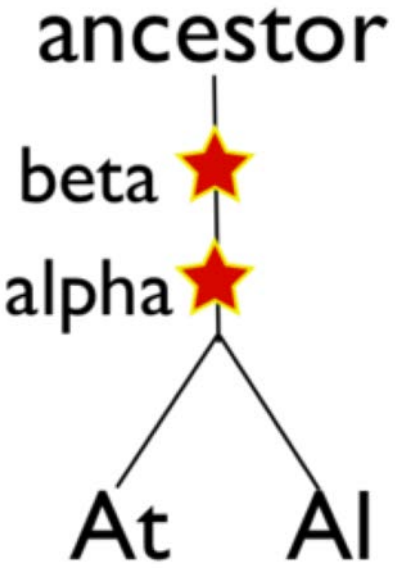


Slide of Mike Schatz

Relationship between genome synteny, syntenic orthologs and duplications

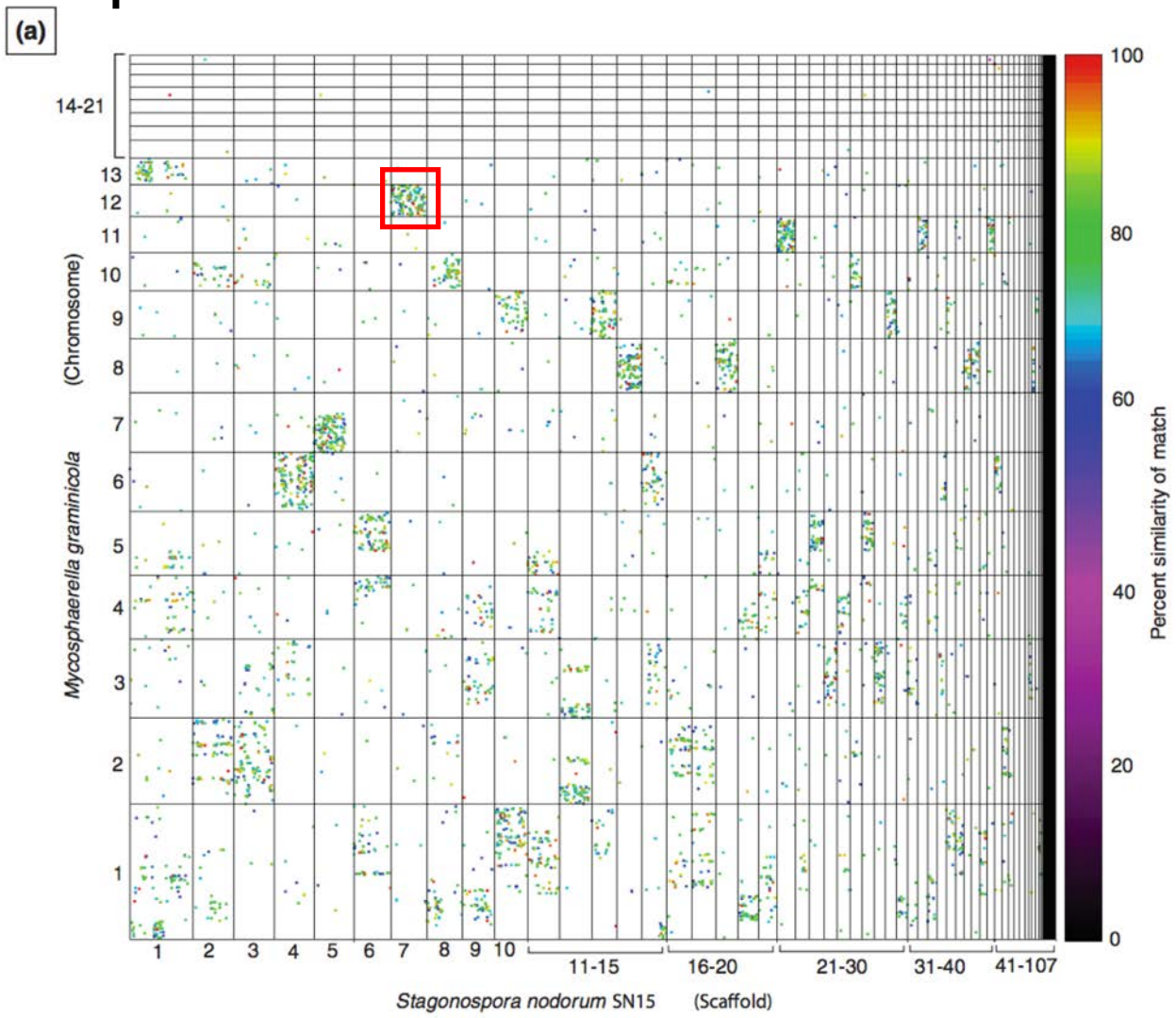


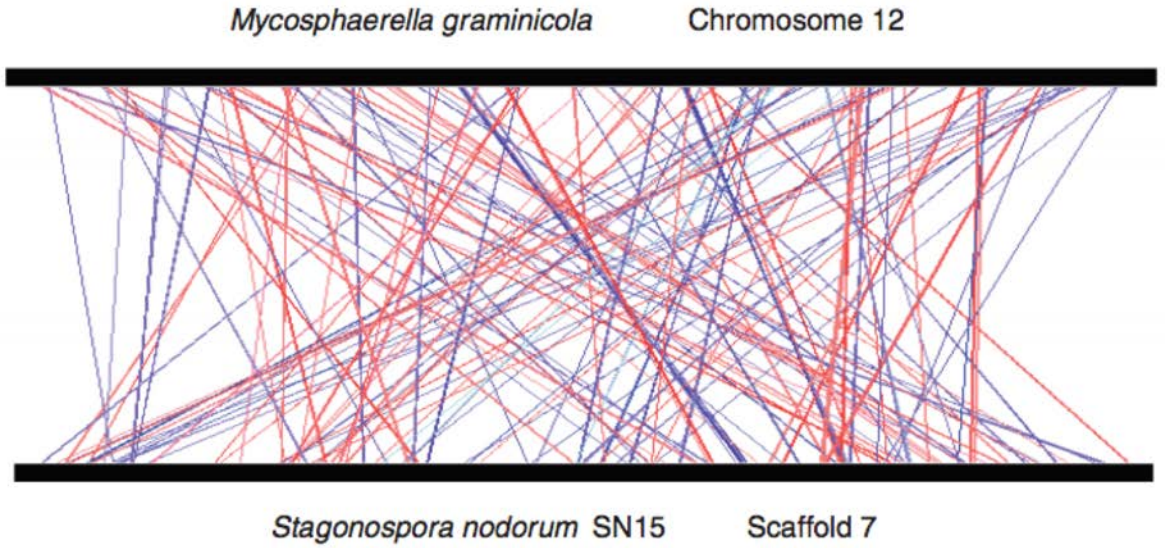




https://genomevolution.org/wiki/index.php/Syntenic_comparison_of_Arabidopsis_thaliana_and_Arabidopsis_lyrata

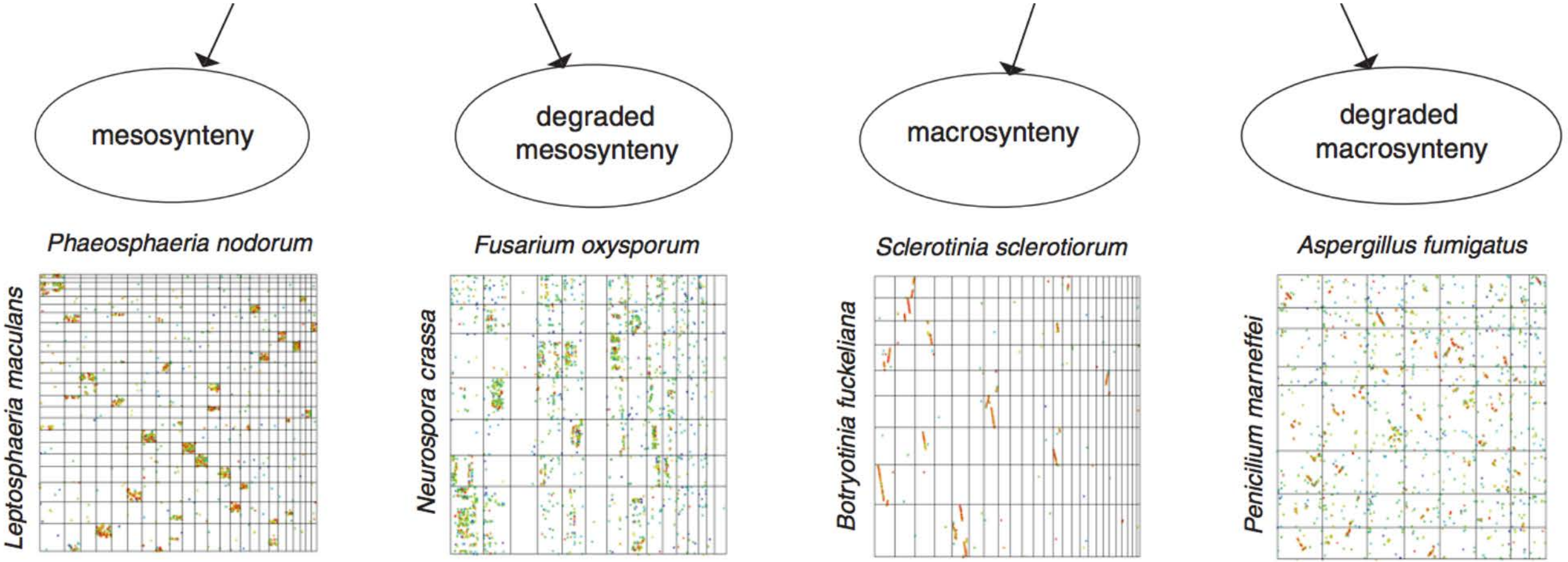
Relationship between genome synteny, syntenic orthologs and duplications





Hane et al (2011)

Different kinds of genome synteny



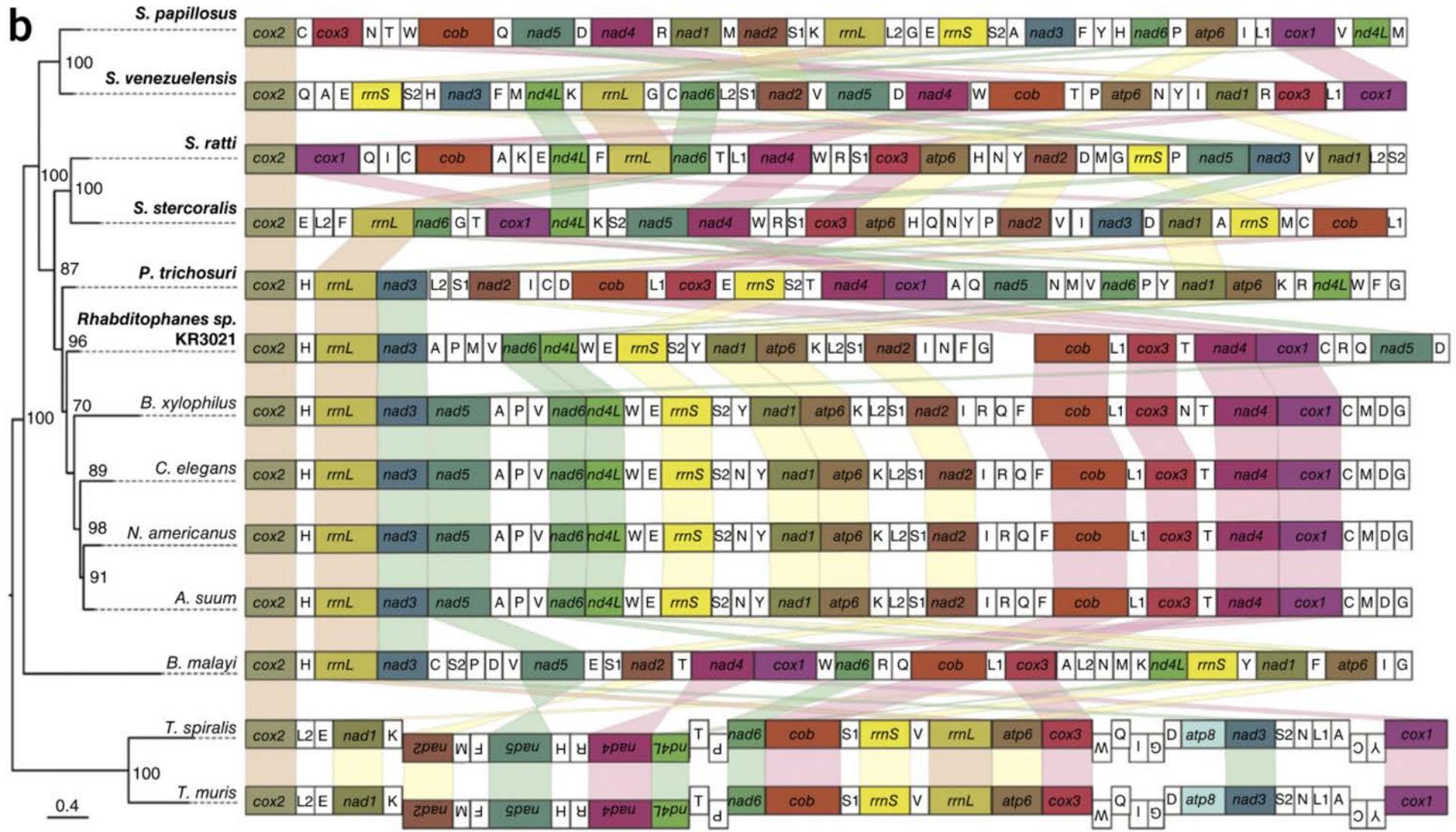
genes are conserved within homologous chromosomes, but with randomized orders and orientations

genes are conserved within homologous chromosomes, and with colinear gene regions

Hane et al (2011)

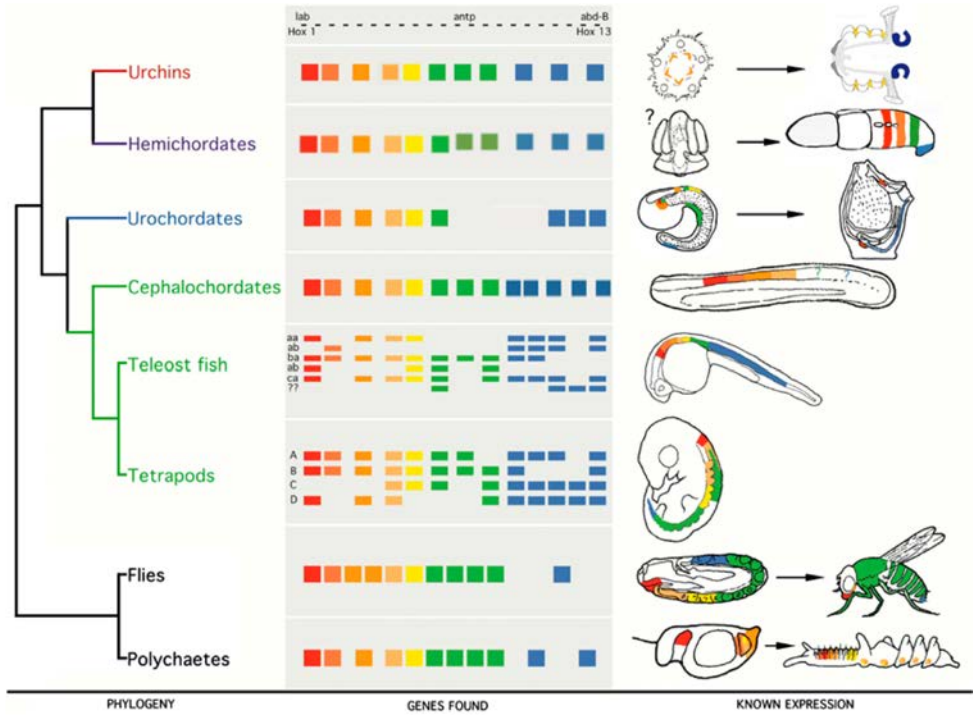
Why are we interested in synteny and collinearity?

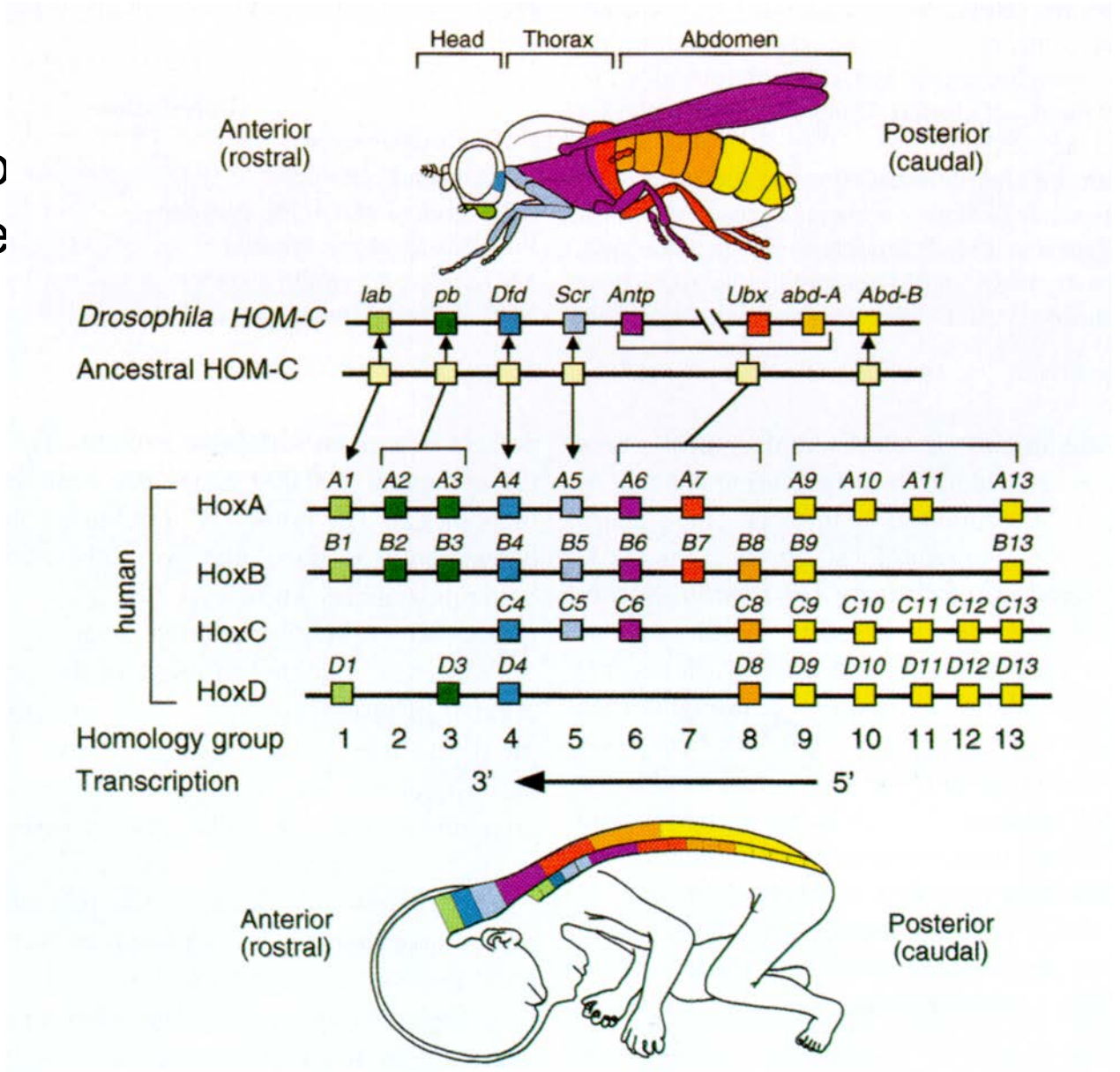
Establish relationship between species



Why are we interested in synteny and collinearity?

Evolutionary conserved features (orthologs, synteny, collinearity) are goo indicators of functionally important genome regions

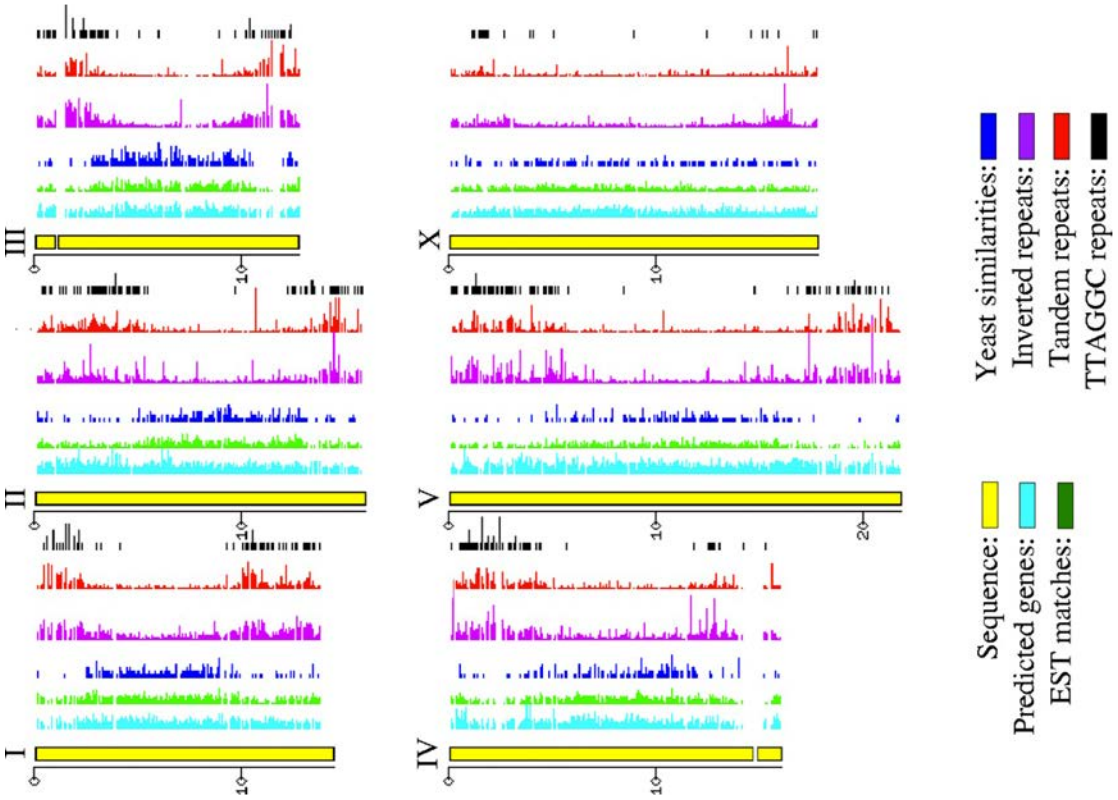


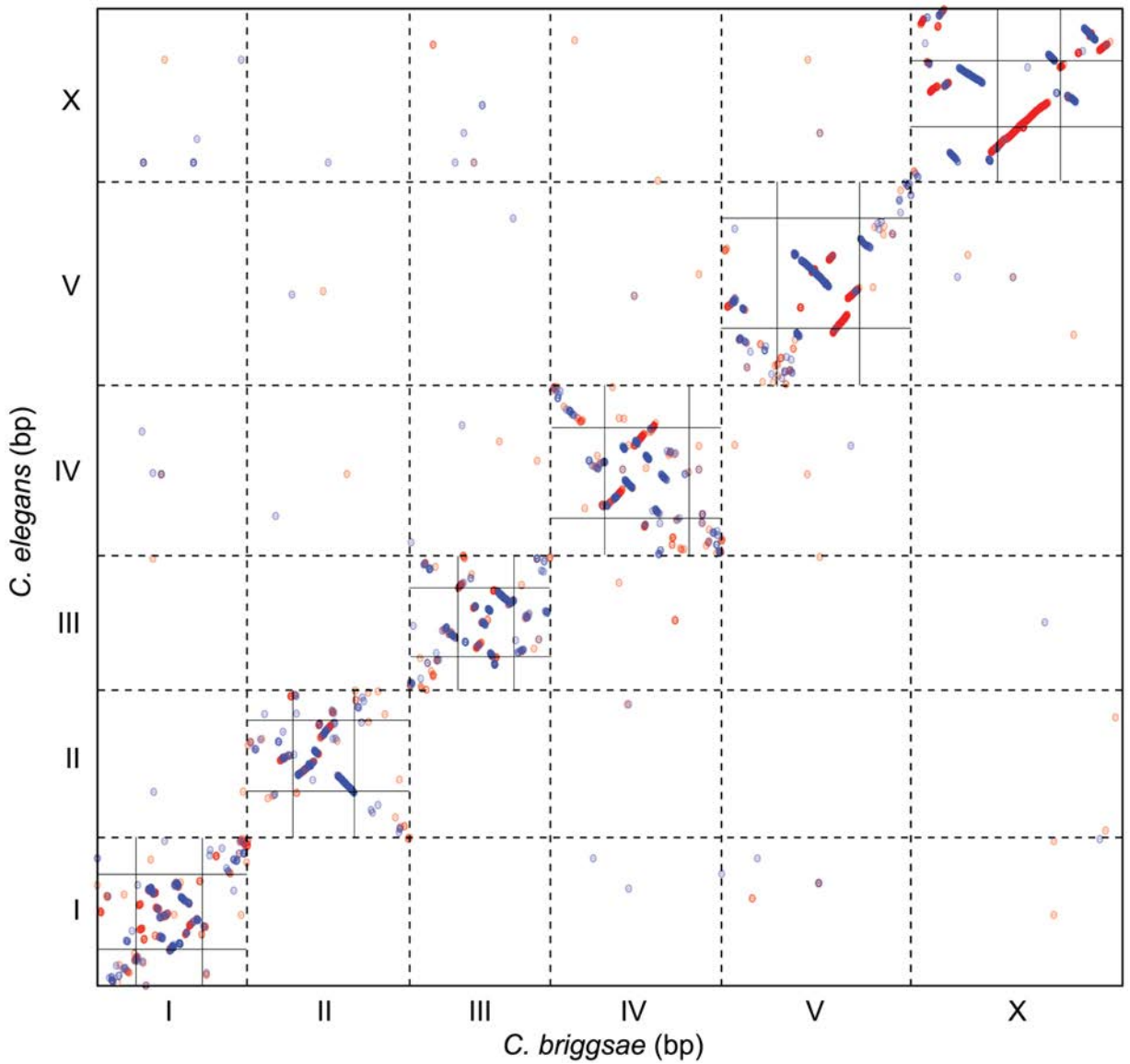


Swalla (2006) ; Mark et al (1997)

Why are we interested in synteny and collinearity?

Evolutionary conserved features (orthologs, synteny, collinearity) relate to genome biology



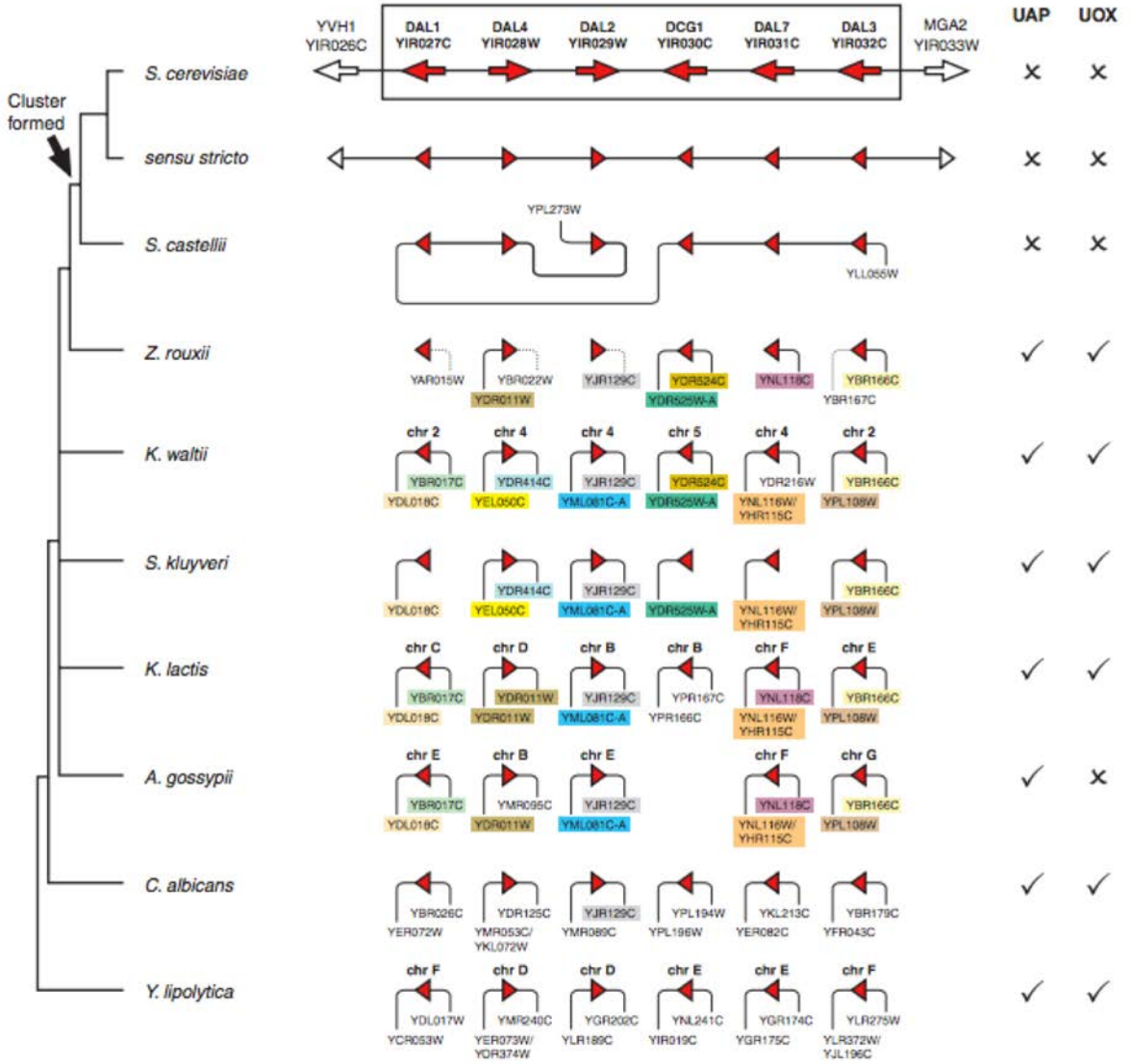


Stein *et al.*, PLOS Biology 2003

The *C. elegans* Sequencing Consortium Science 1998

Why are we interested in synteny and collinearity?

We can **reconstruct evolutionary histories of gene & gene families** and eventually lead to functioning of species

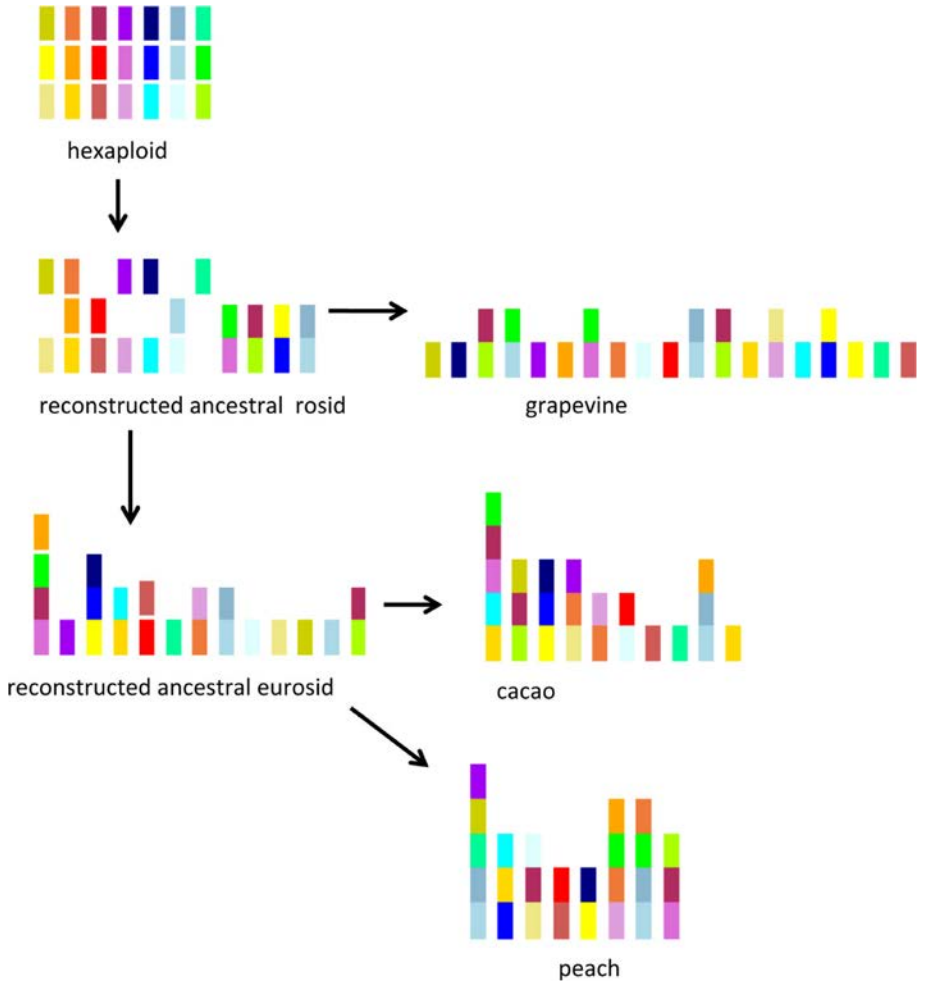


Birth of a metabolic gene cluster in yeast by adaptive gene relocation

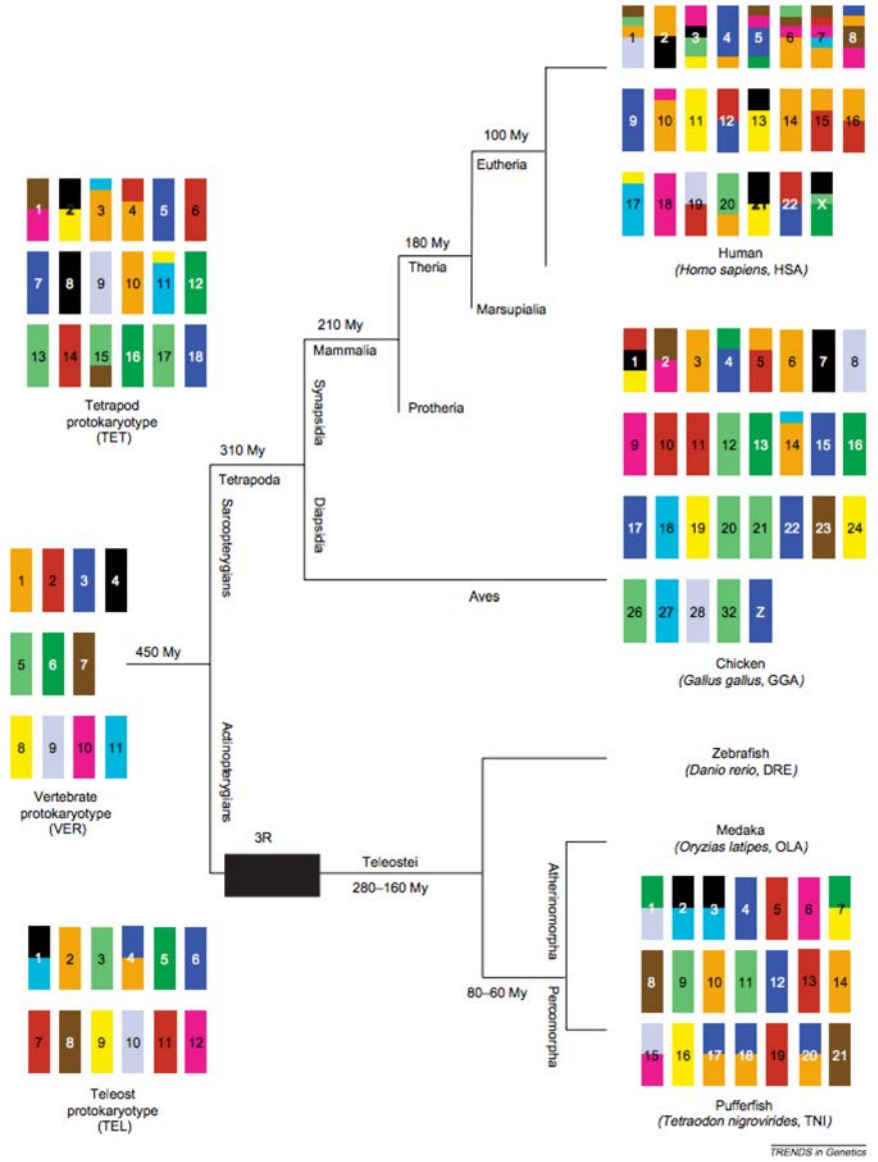
Simon Wong & Kenneth H Wolfe

Why are we interested in synteny and collinearity?

We can **reconstruct ancient karyotypes** that eventually lead to better understanding of evolution of species



Zheng et al (2013)



Kohn et al (2006)

Case study:

nature plants

ARTICLES

https://doi.org/10.1038/s41477-018-0337-0

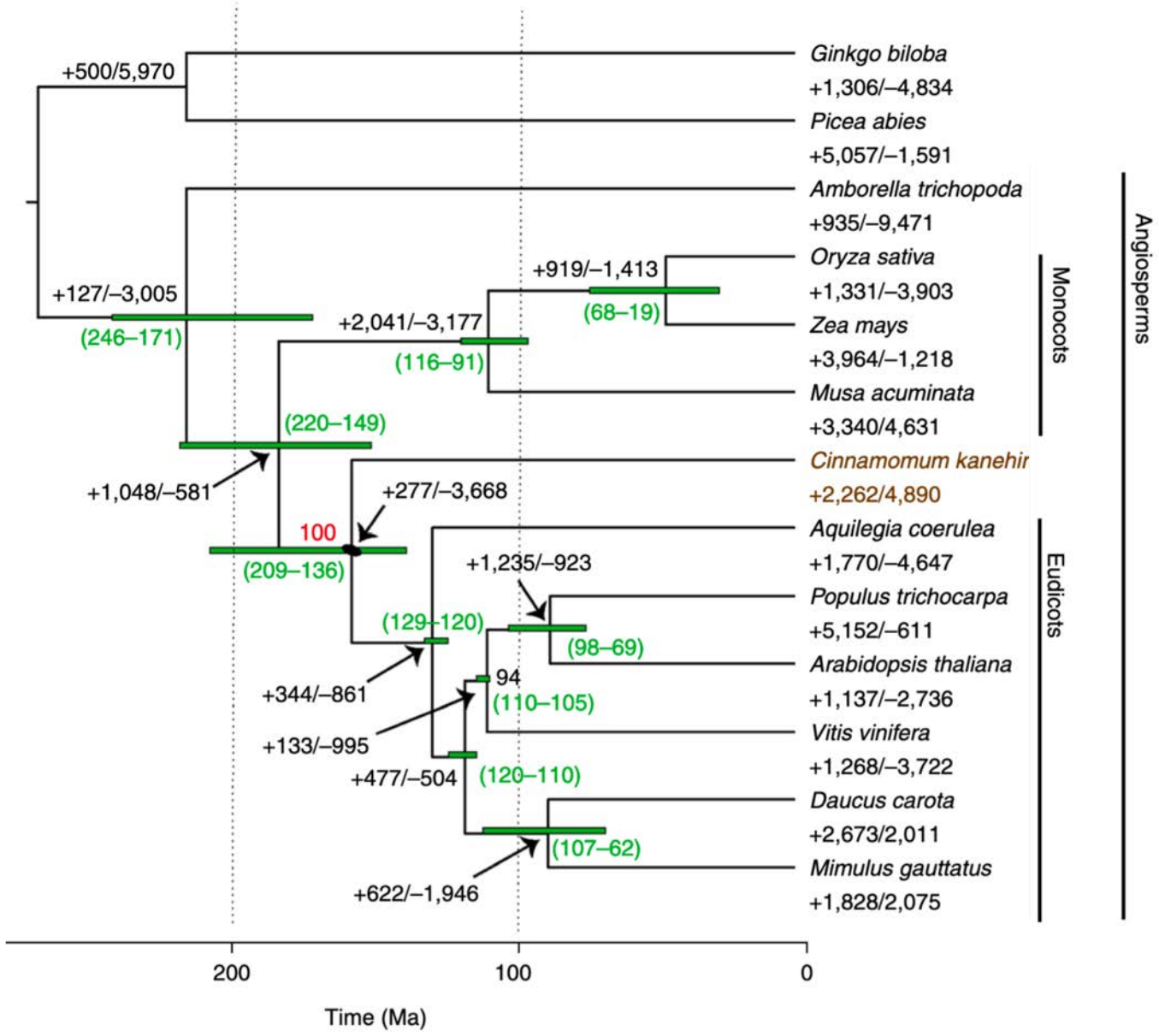
OPEN

Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution

Shu-Miaw Chaw[1,6*], Yu-Ching Liu[1], Yu-Wei Wu[2], Han-Yu Wang[1], Chan-Yi Ivy Lin[1], Chung-Shien Wu[1], Huei-Mien Ke[1], Lo-Yu Chang[1,3], Chih-Yao Hsu[1], Hui-Ting Yang[1], Edi Sudianto[1], Min-Hung Hsu[1,4], Kun-Pin Wu[4], Ling-Ni Wang[1], James H. Leebens-Mack[5] and Isheng J. Tsai[1,6*]

Phylogenomic placement

Phylogenomic placement of *C. kanehirae* sister to eudicots. To resolve the long-standing debate over the phylogenetic placement of magnoliids relative to other major flowering plant lineages, we constructed a phylogenetic tree based on 211 strictly single-copy orthologue sets (that is, one and only one homologue in all species) identified through OrthoFinder[21] gene family circumscription of all gene models from the SCT and 12 other seed plant genomes (see Methods). A single species tree was recovered through maximum likelihood analysis[27] of a concatenated supermatrix of the single-copy gene alignments and coalescent-based analysis using the 211 gene trees[28] (Fig. 2 and Supplementary Fig. 11). SCT, representing the magnoliid lineage, was placed as sister to the eudicot clade (Fig. 2). This topology remained robust when we included a transcriptome data set of an additional 22 species of magnoliids order from the 1,000 plants initiative[29] (1KP), although lower bootstrap support was obtained (Supplementary Fig. 12). Using MCMCtree[30] with fossil calibrations, we calculated a 95% confidence interval for the time of divergence between magnoliids and eudicots to be 136.0–209.4 Ma (Fig. 2), which overlaps with two other recent estimates (114.8–164.1 Ma[31] and 118.9–149.9 Ma[32]).



Chaw et al (2019)

Signature of whole genome duplication

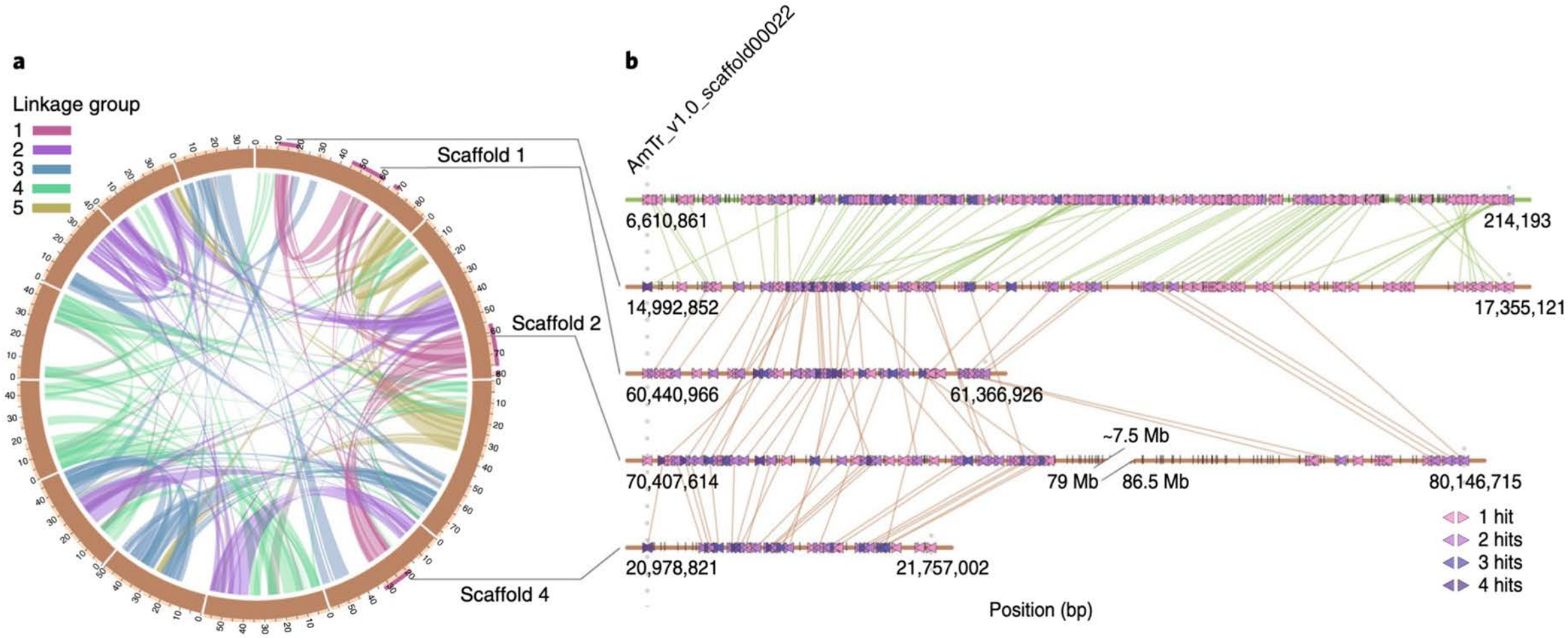
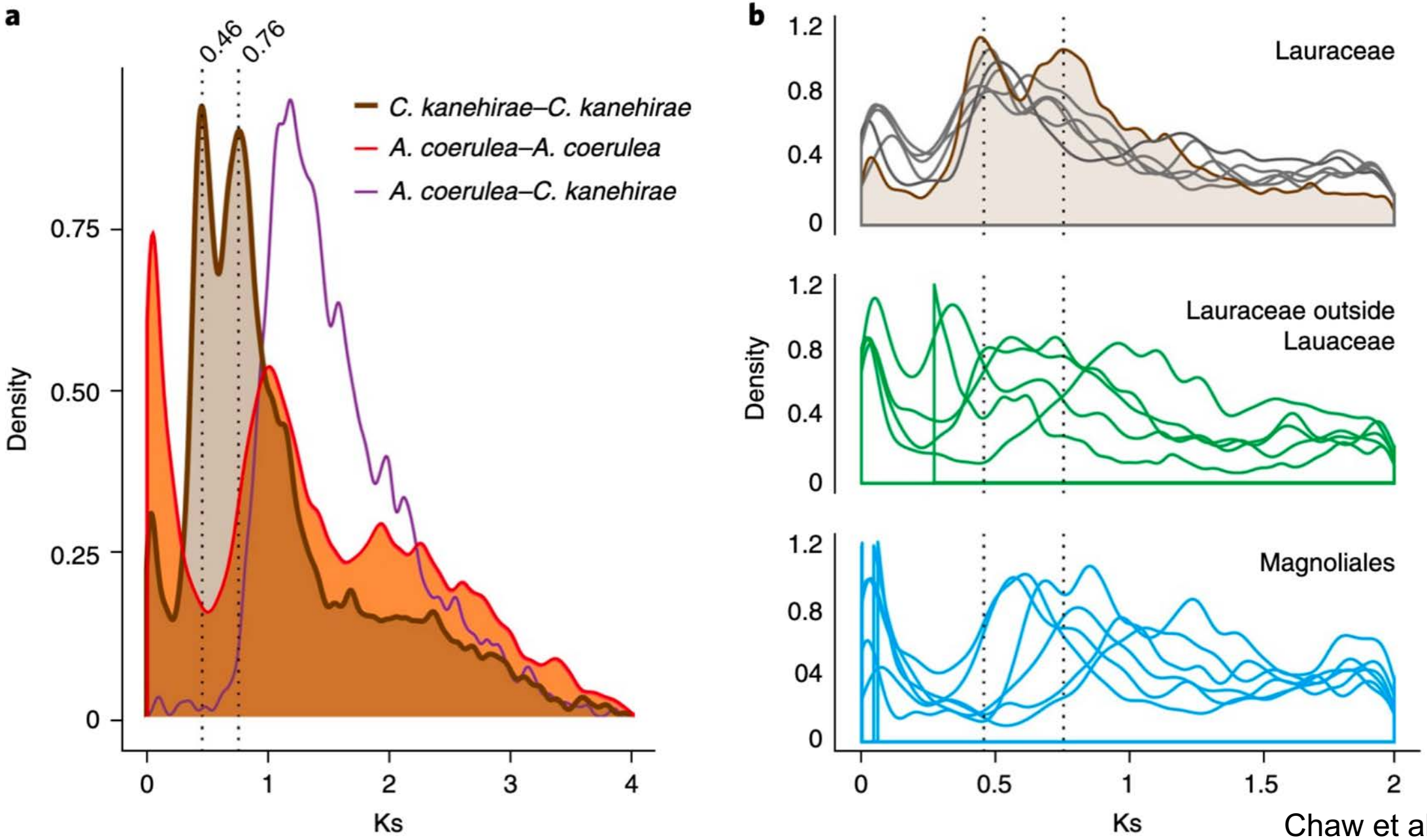


Fig. 3 | Evolutionary analysis of the SCT genome. a, Schematic representation of the intragenomic relationship among the 637 synteny blocks in the SCT genome. Synteny blocks (denoted by peach blocks) were assigned unambiguously into five linkage clusters representing ancient karyotypes and are colour coded. Purple blocks denote the synteny block assigned in the first linkage group (see also Supplementary Fig. 13). **b**, Schematic representation of the first linkage group within the SCT genome and their corresponding relationship in *A. trichopoda*.

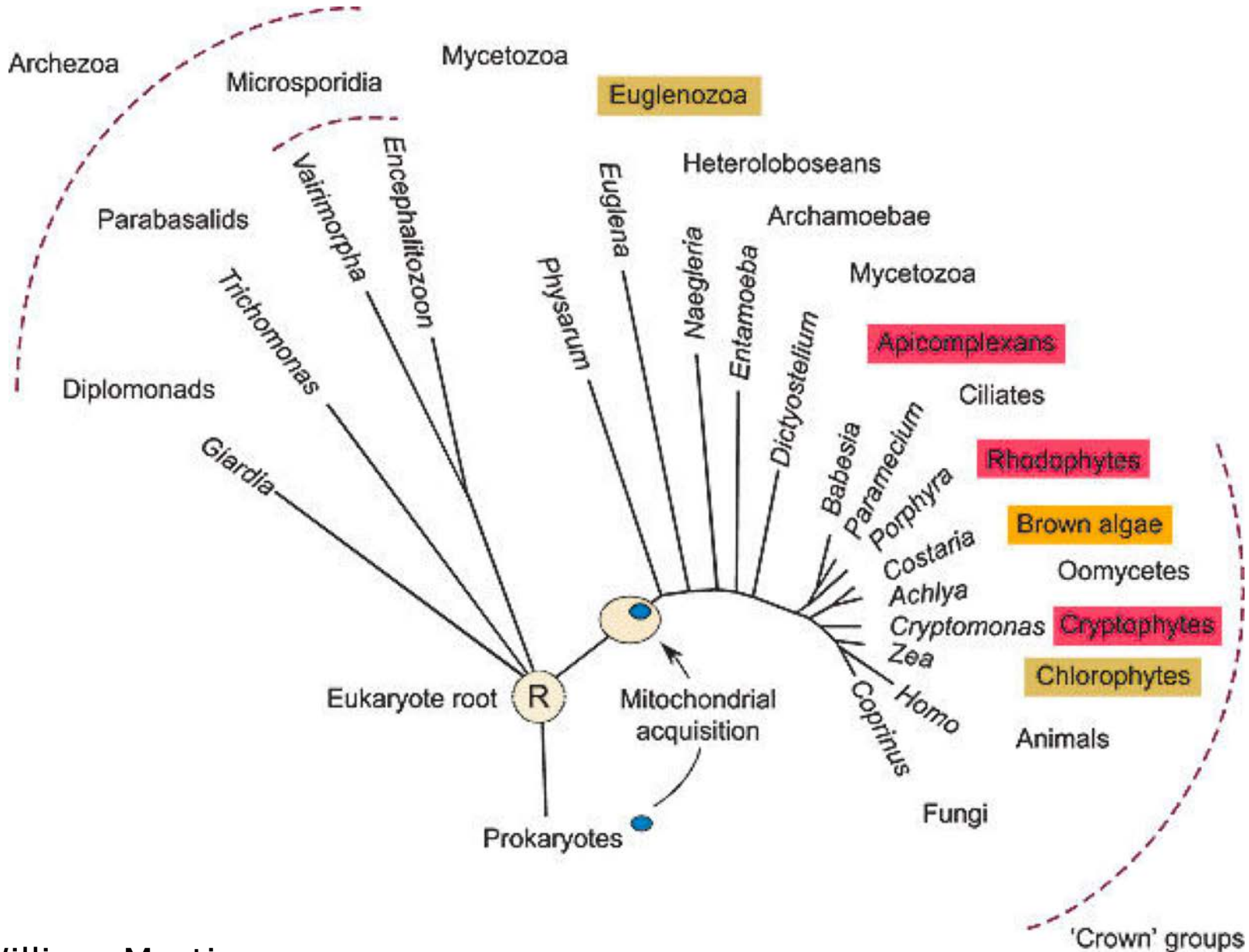
Chaw et al (2019)

Signature of whole genome duplication



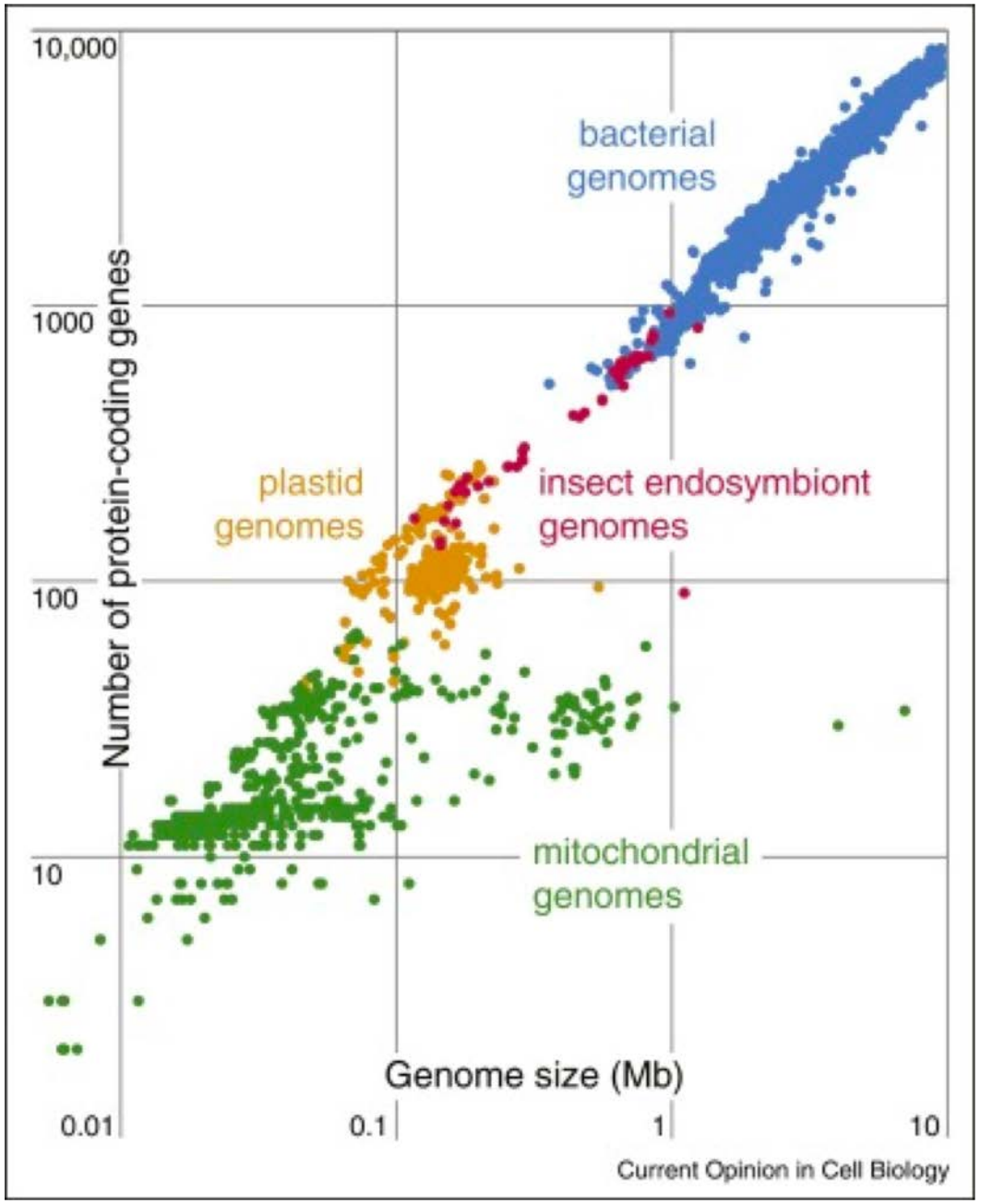
Chaw et al (2019)

How do we study origin of organelles?



Martin Embley & William Martin
Nature **440**, 623-630(30 March 2006)

Genomes from bacteria, insect endosymbionts, chloroplasts, and mitochondria form an unbroken continuum of size and coding density. The plot is truncated at 10 Mb and 10,000 genes.

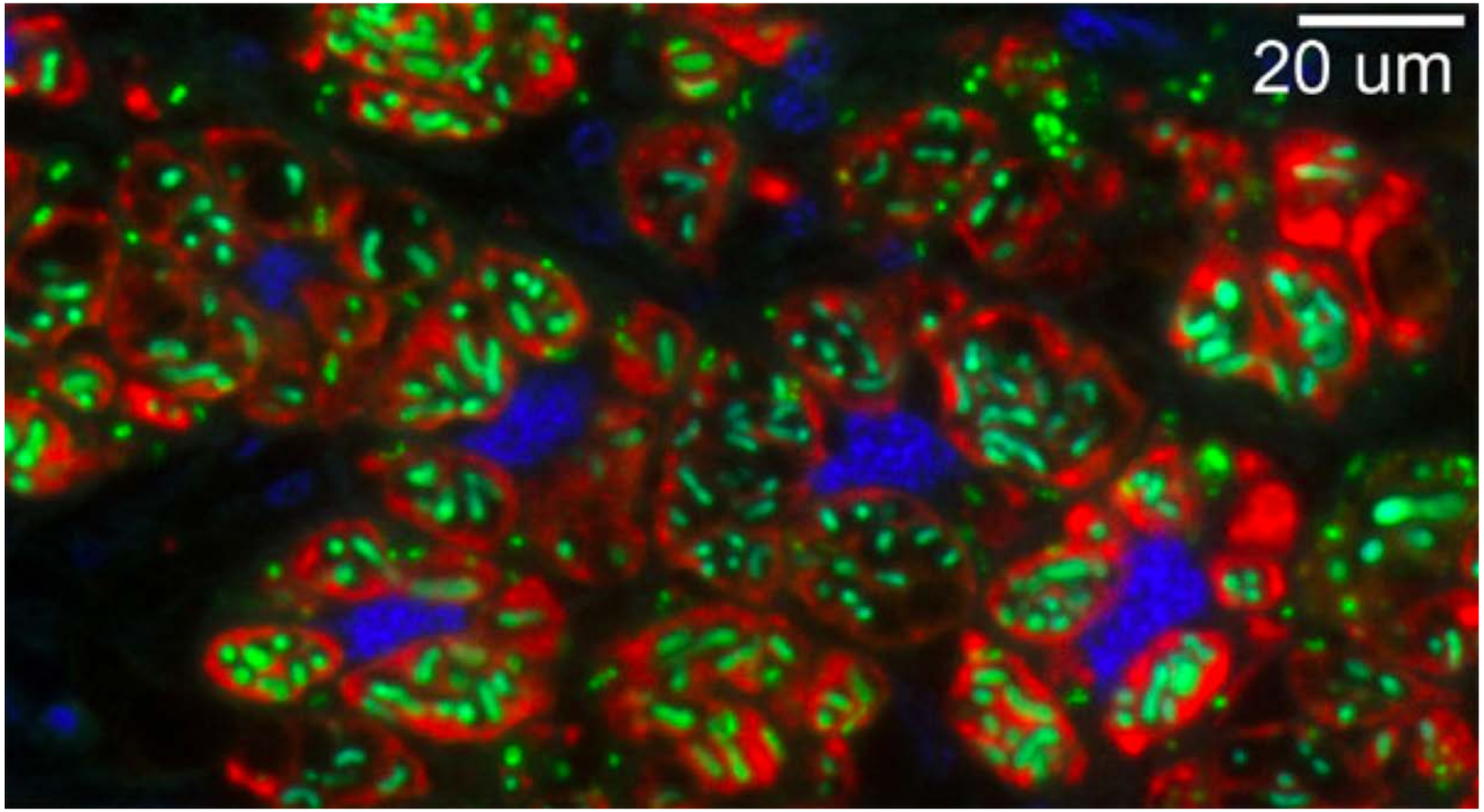


“Insect endosymbionts are missing (genomic) links between bacteria and organelles. It is now widely appreciated that all animals form symbioses with bacteria. Insects are especially interesting in this regard because they form many intracellular symbioses — that is, they allow bacteria to live inside their cells — that are not pathogenic from the host perspective”

McCutcheon (2016) Current Opinion in Cell Biology

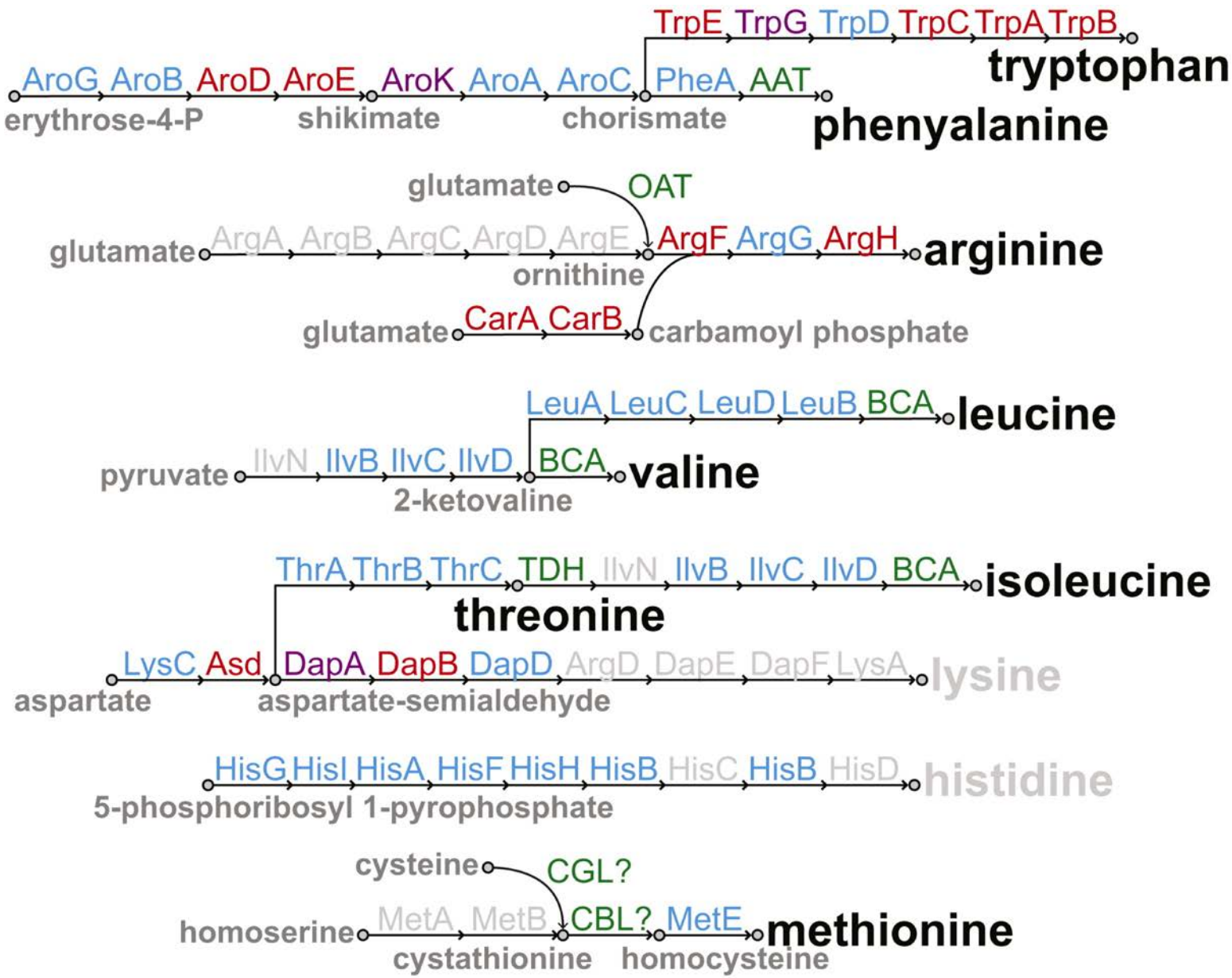
Case study: Mealybugs

Triple Symbiotic Relationship between Mealybugs, *Tremblaya princeps*, and *Moranella endobia*



Mealybug cells, showing Tremblaya (red), Moranella (green) and mealybug nuclei (blue).
Credit: Ryuichi Koga, National Institute of Advanced Industrial Science and Technology, Japan

Predicted Essential Amino Acid Metabolic Contributions of the Mealybug-Tremblaya-Moranella Symbiosis



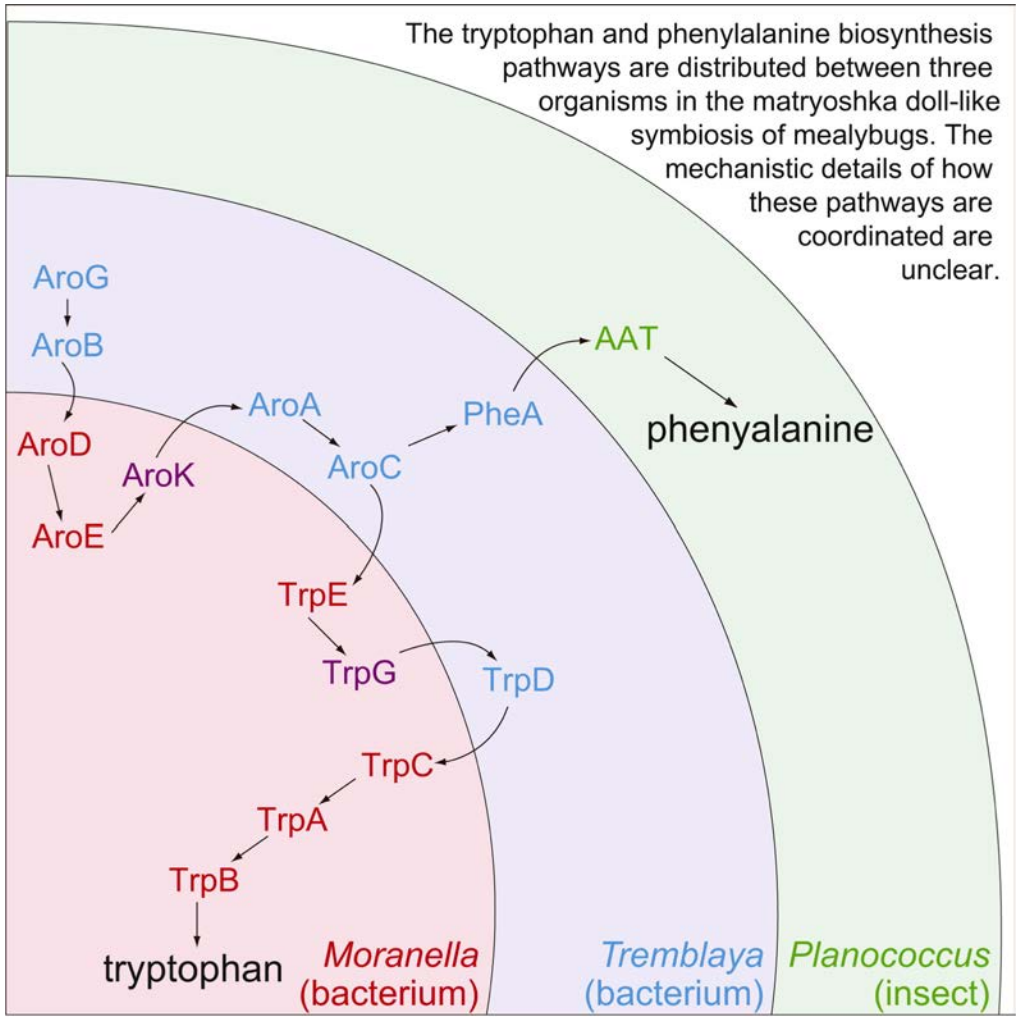
Gene homologs found in the Tremblaya genome are blue; the Moranella genome, red; both the Tremblaya and Moranella genomes, purple; neither the Tremblaya nor the Moranella genome, gray; activities not found in either bacterial genome but predicted to be encoded in the mealybug genome, green.

McCutcheon and Dohlen (2011) Current Biology

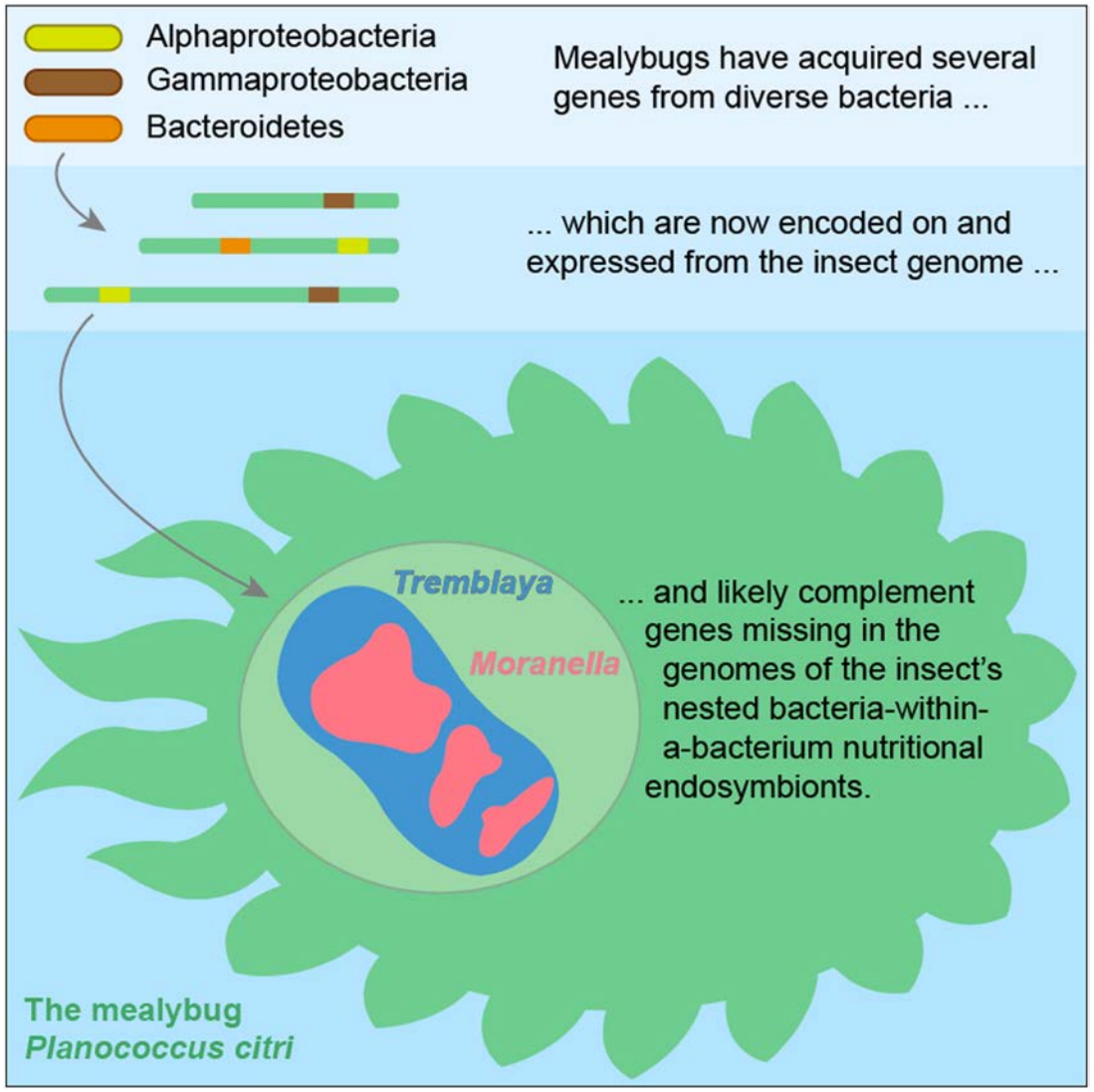
Genome degeneracy of a bacterial endosymbiont is driven by its own endosymbiont

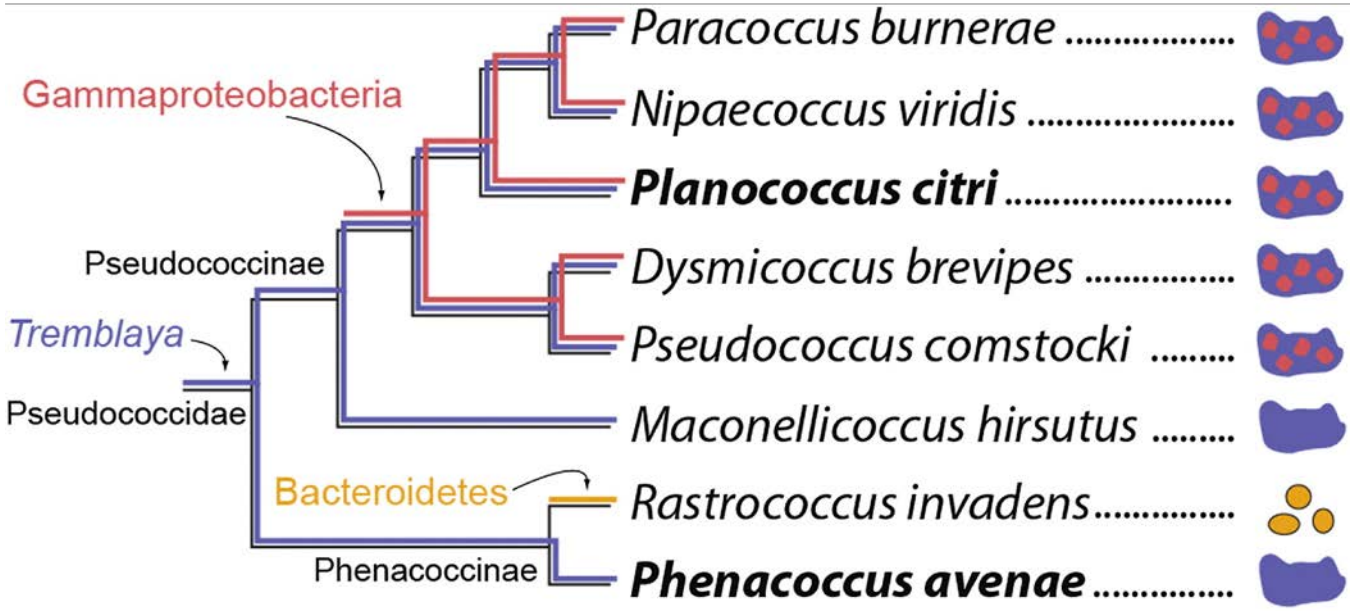
•HGT from diverse bacteria to the insect host genome support the three-way symbiosis

•Endosymbiont genomes can massively degrade without transfer of genes to the host

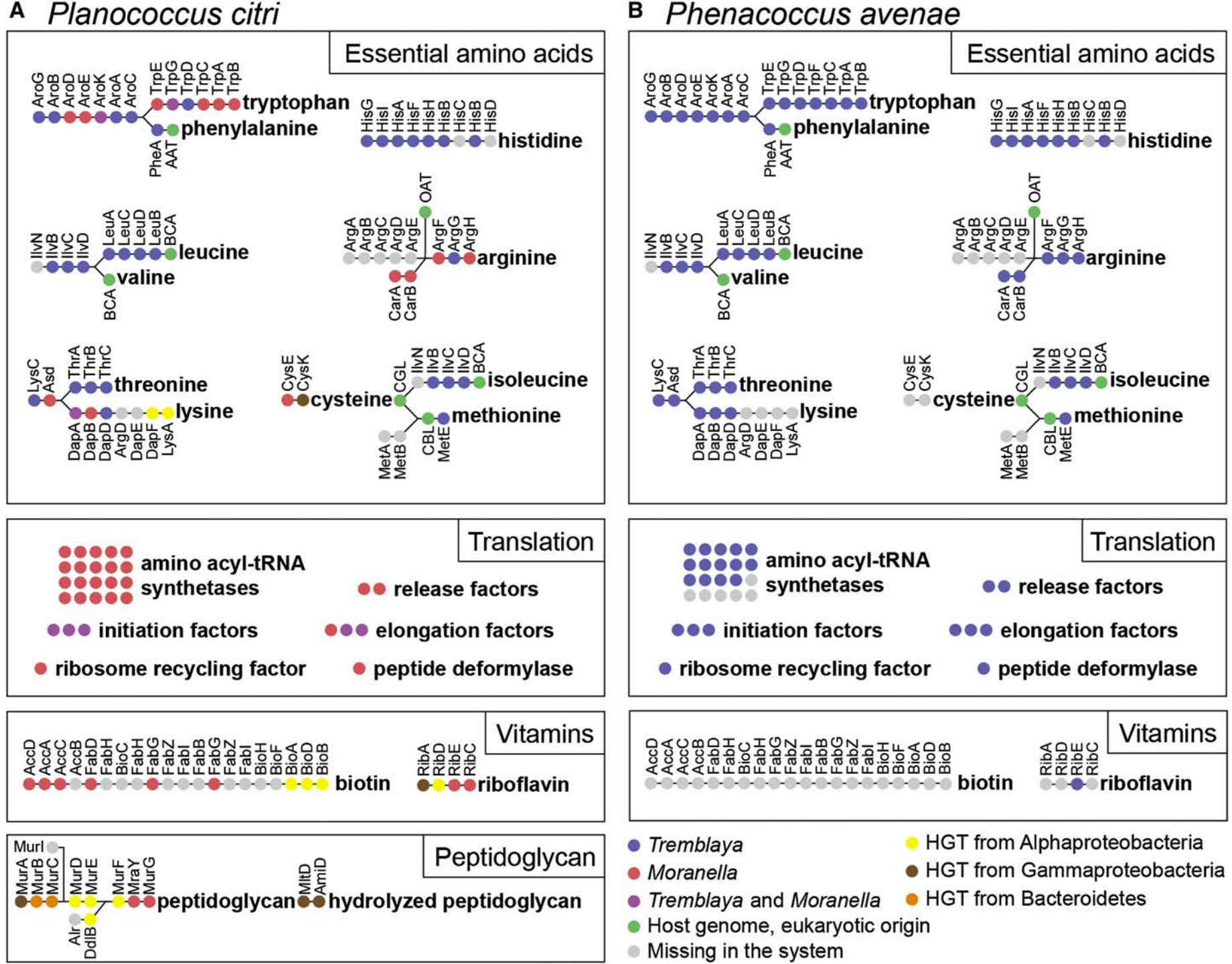


Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis





FilipHusnik et al., (2013) Cell


A Planococcus citri
Essential amino acids
tryptophan
phenylalanine
histidine
leucine
valine
arginine
threonine
lysine
isoleucine
cysteine
methionine
Translation
amino acyl-tRNA synthetases
release factors
initiation factors
elongation factors
ribosome recycling factor
peptide deformylase
Vitamins
biotin
riboflavin
Peptidoglycan
peptidoglycan
hydrolyzed peptidoglycan
B Phenacoccus avenae
Essential amino acids
tryptophan
phenylalanine
histidine
leucine
valine
arginine
threonine
lysine
isoleucine
cysteine
methionine
Translation
amino acyl-tRNA synthetases
release factors
initiation factors
elongation factors
ribosome recycling factor
peptide deformylase
Vitamins
biotin
riboflavin
Tremblaya
Moranella
Tremblaya and Moranella
Host genome, eukaryotic origin
Missing in the system
HGT from Alphaproteobacteria
HGT from Gammaproteobacteria
HGT from Bacteroidetes

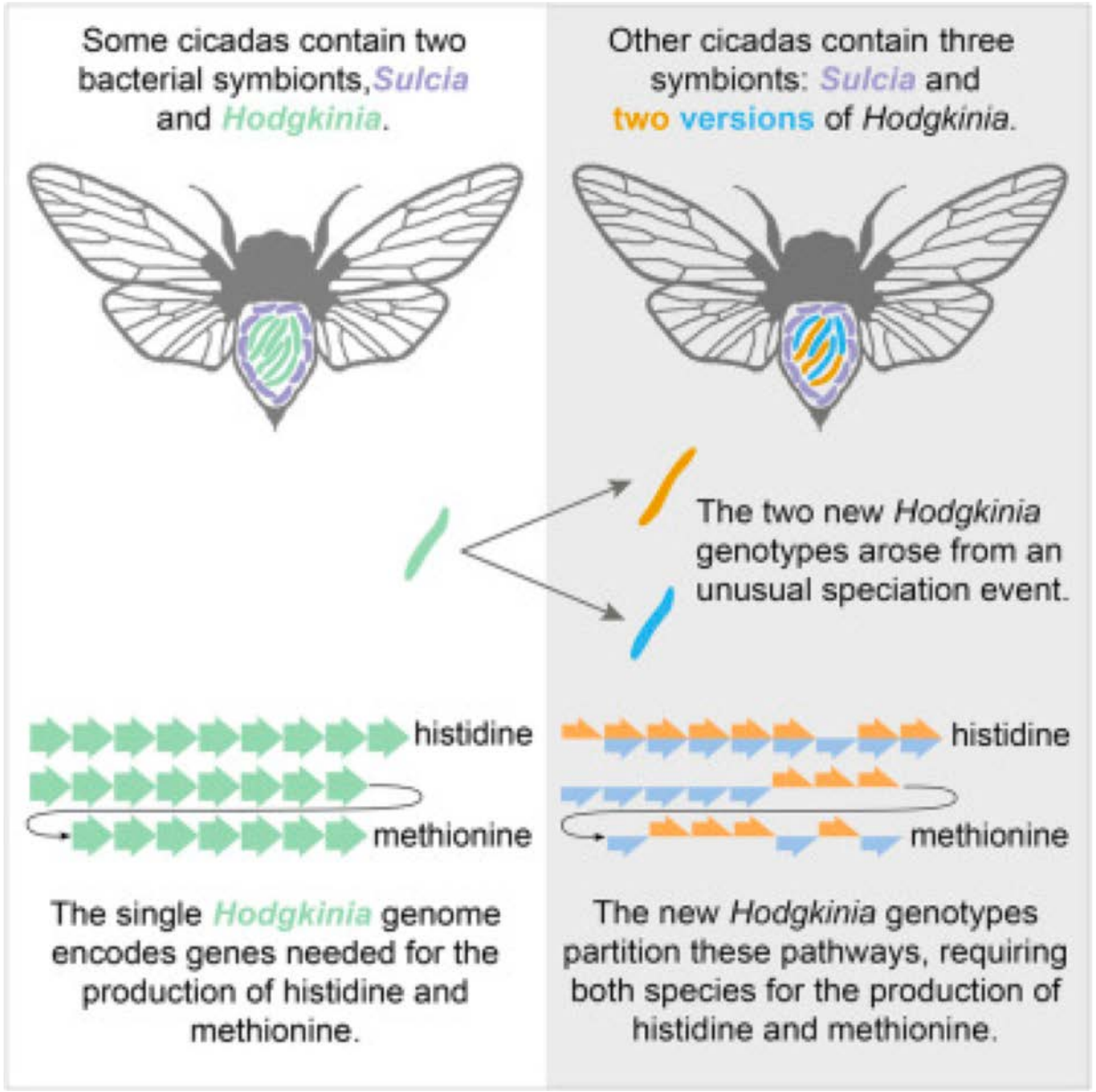
Even more fascinating case

Cell

Sympatric Speciation in a Bacterial Endosymbiont Results in Two Genomes with the Functionality of One

James T. Van Leuven,[1] Russell C. Meister,[2] Chris Simon,[2] and John P. McCutcheon[1,3,*]
[1]Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA
[2]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA
[3]Canadian Institute for Advanced Research, CIFAR Program in Integrated Microbial Biodiversity, Toronto, ON M5G 1Z8, Canada
*Correspondence: john.mccutcheon@umontana.edu
http://dx.doi.org/10.1016/j.cell.2014.07.047

https://www.youtube.com/watch?v=XRI2JxTzJ-0&list=UUlSV2Tk7x-wBBXP6-VCNbNw

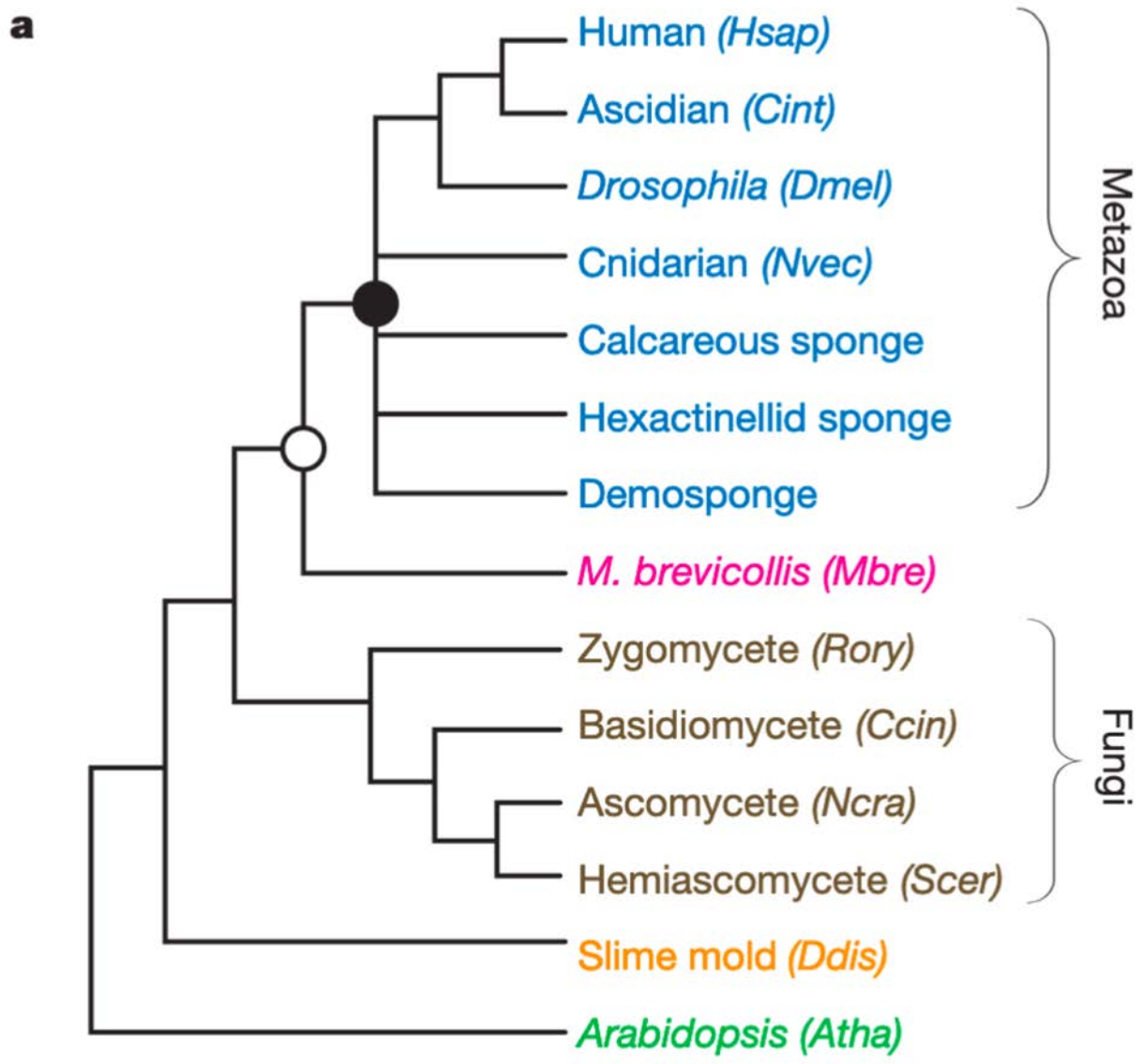


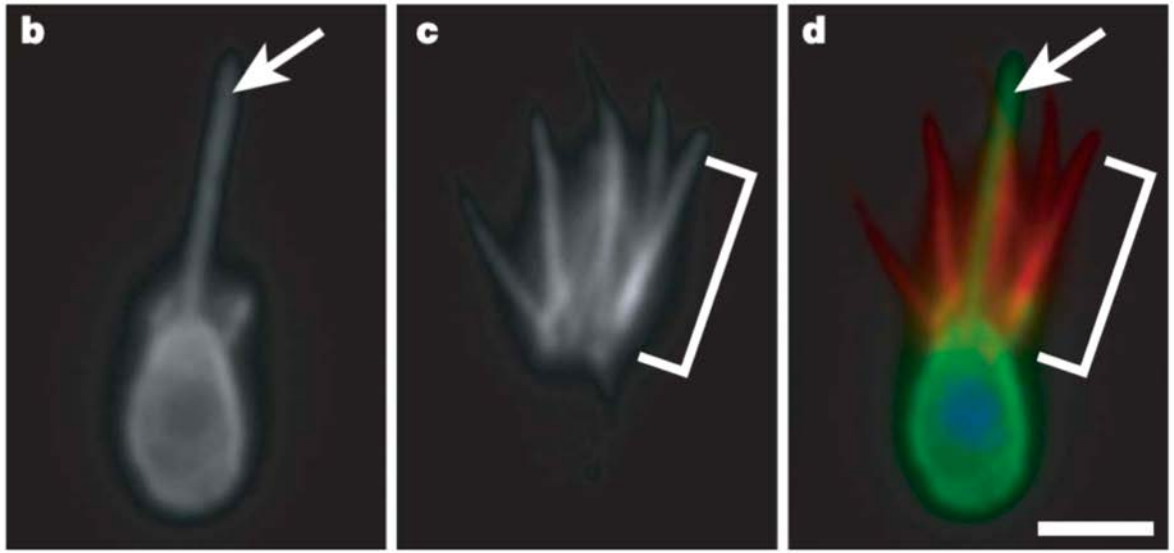
How do we study origin of animals (metazoans)?

Question: what constitute the first animal?

- Unicellular -> Multicellular
  - What's needed for multicellularity?
    - Interactions between cells
    - Formation of aggregates? How?

Choanoflagellate





Ddis) and *Arabidopsis* (*Arabidopsis thaliana*, *Atha*). **b–d**, Choanoflagellate cells bear a single apical flagellum (arrow, **b**) and an apical collar of actin-filled microvilli (bracket, **c**). **d**, An overlay of β-tubulin (green), polymerized actin (red) and DNA localization (blue) reveals the position of the flagellum within the collar of microvilli. Scale bar, 2 μm.

King et al., (2008) Nature

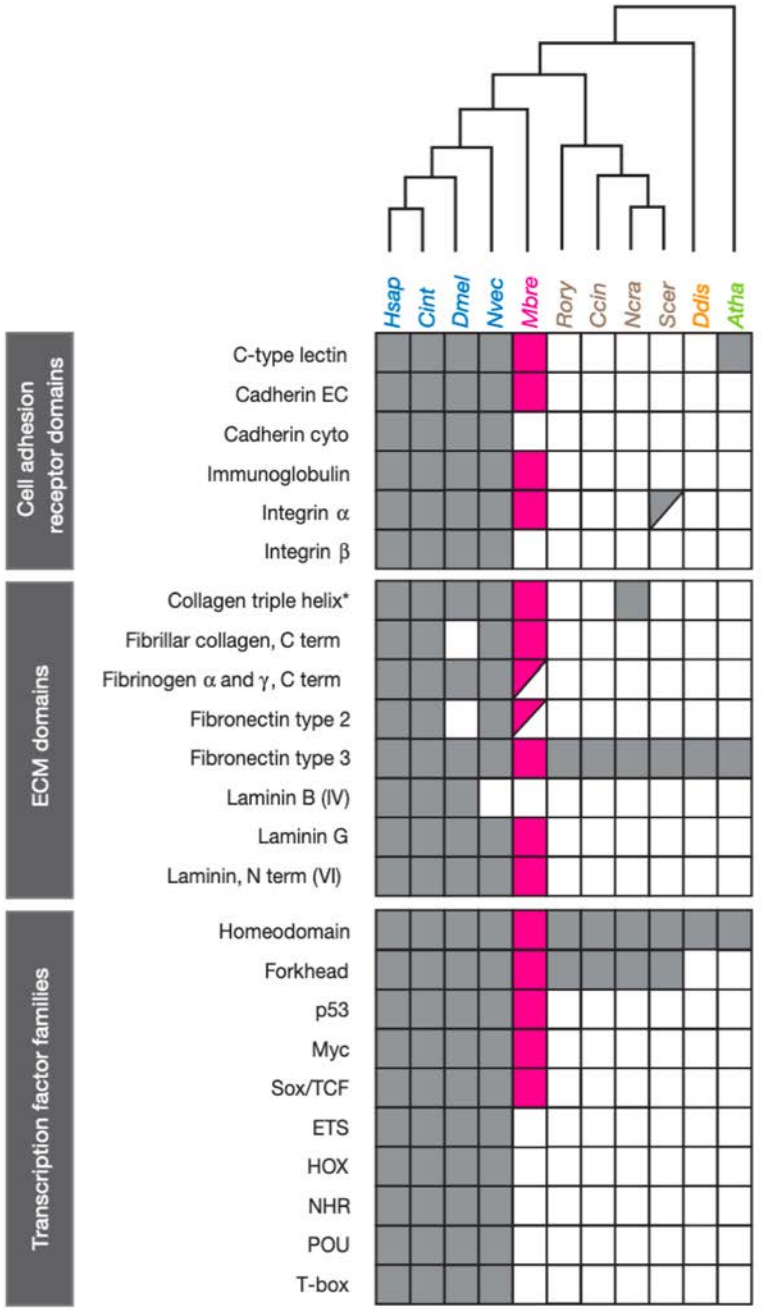
Choanoflagellate – example findings

An abundance of cell adhesion domains

A critical step in the transition to multicellularity was the evolution of mechanisms for stable cell adhesion. *M. brevicollis* encodes a diverse array of cell adhesion and extracellular matrix (ECM) protein domains previously thought to be restricted to metazoans (Fig. 3).

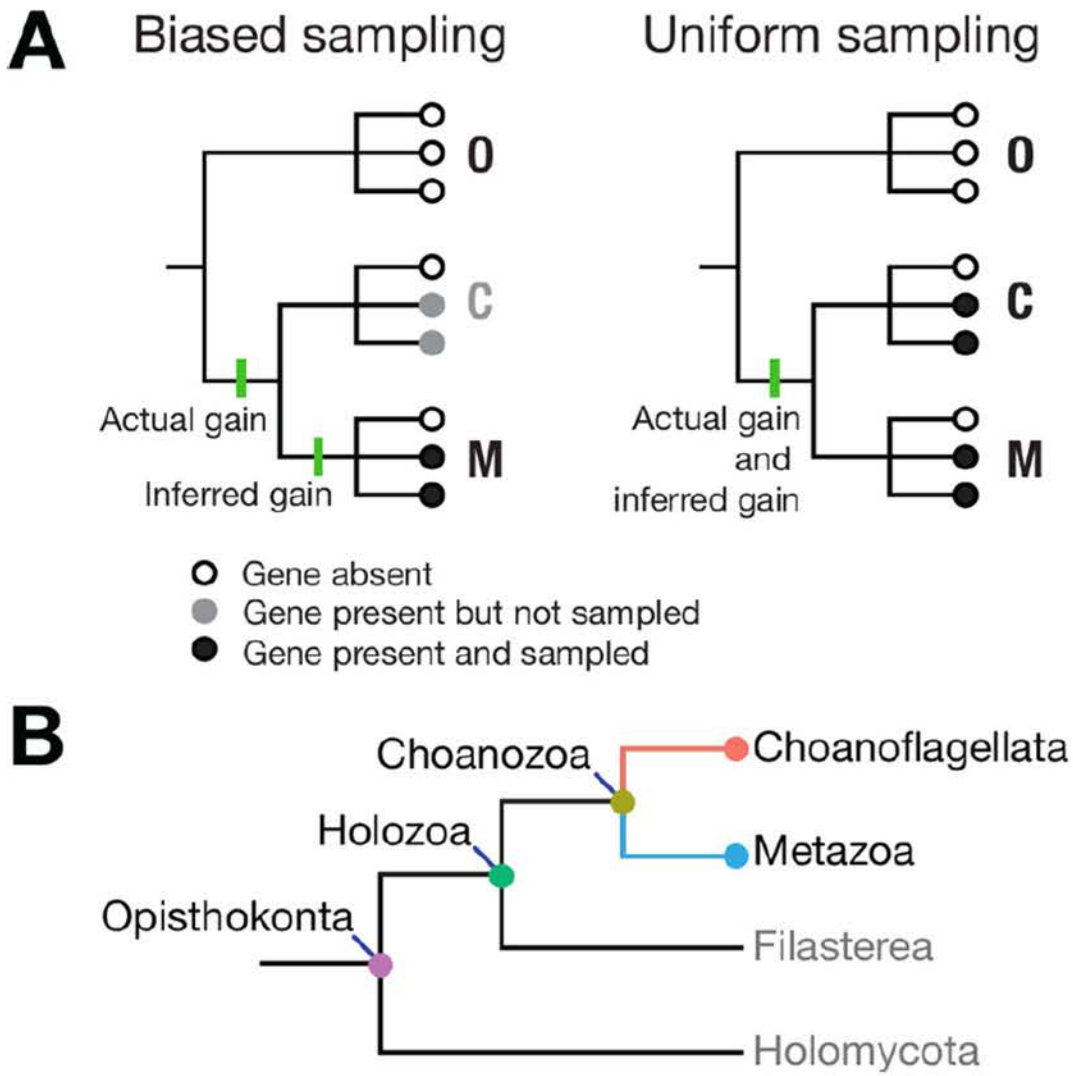
The finding in *M. brevicollis* of cell adhesion domains that were previously known only in metazoans has two important implications. First, the common ancestor of metazoans and choanoflagellates possessed several of the critical structural components used for multicellularity in modern metazoans. Second, given the absence of evidence for stable cell adhesion in *M. brevicollis*, this also suggests that homologues of metazoan cell adhesion domains may act to mediate interactions between *M. brevicollis* and its extracellular environment.

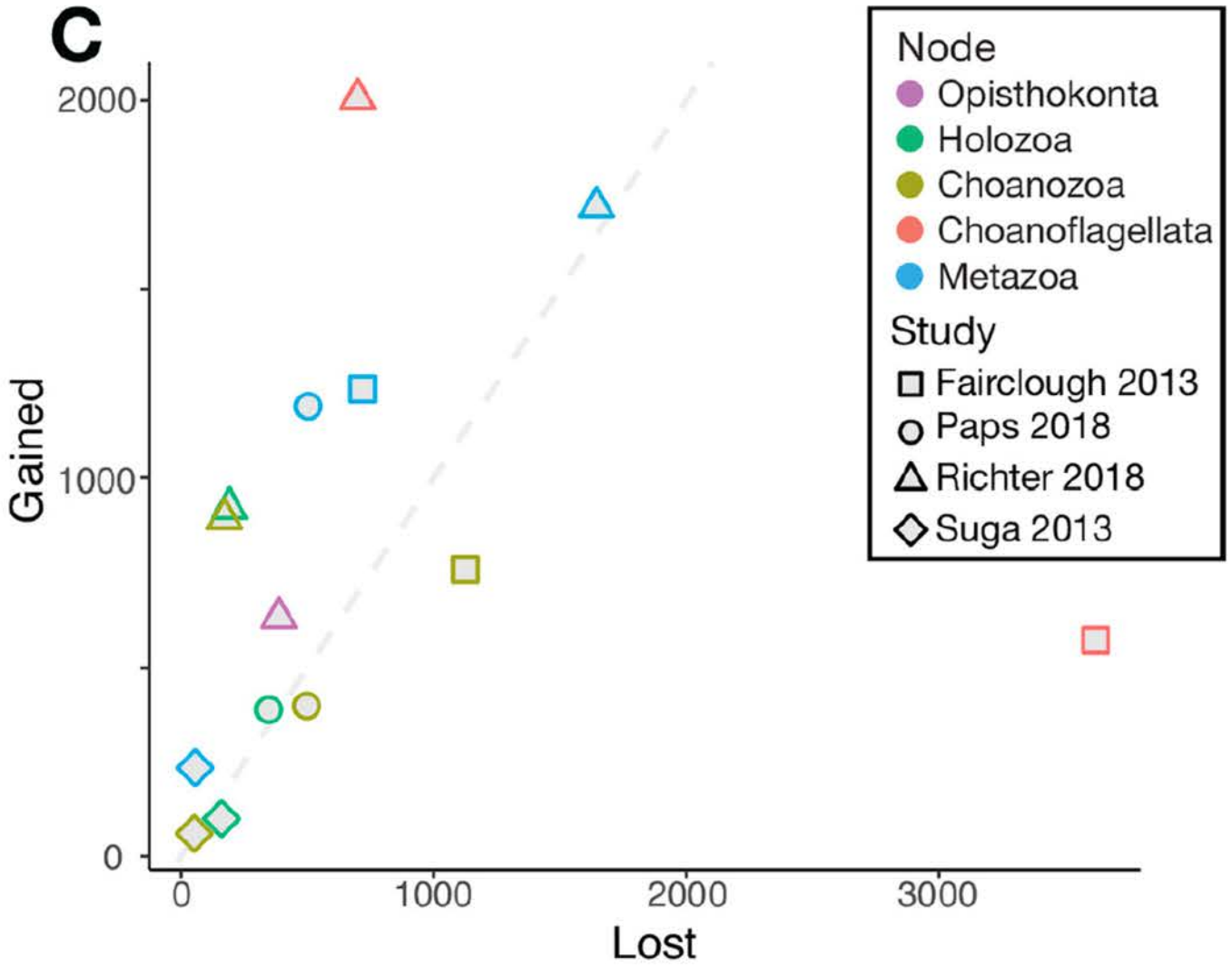
The discovery of putatively secreted ECM proteins in a free-living choanoflagellate suggests that elements of the metazoan ECM evolved in contact with the external environment before being sequestered within an epithelium. Although some choanoflagellates secrete extracellular structures or adhere to form colonial assemblages[19,33,34], *M. brevicollis* is not known to do so. Instead, these ECM protein homologues in *M. brevicollis* may mediate an analogous process such as substrate attachment.



King et al., (2008) Nature

Choanoflagellate – biased sampling can lead to different results

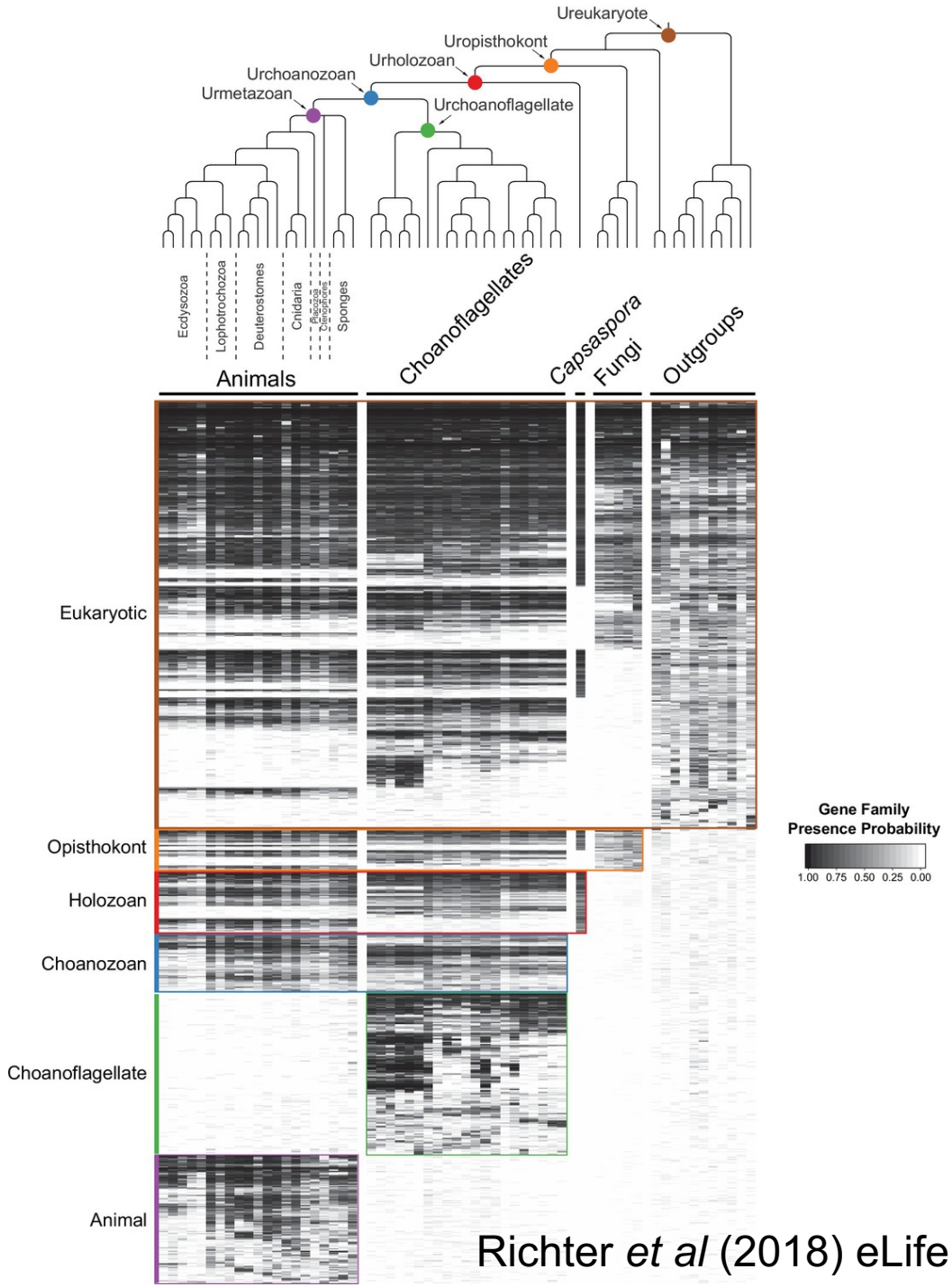




Lewis and Dunn (2018) eLife

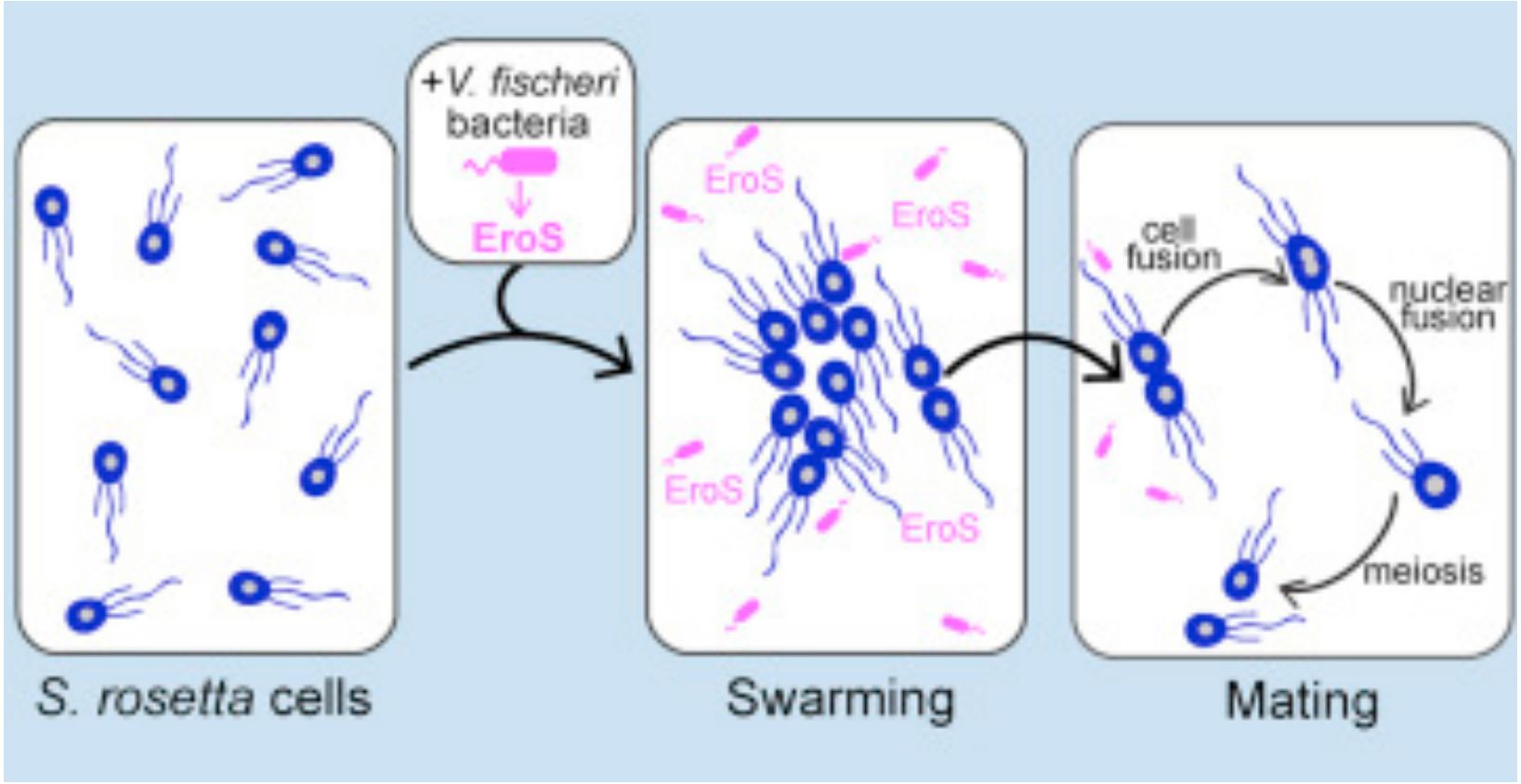
Deeper sampling reveal Differential retention and loss of ancestral gene families in extant animals and choanoflagellates

"The patchwork ancestry of the Urmetazoan genome is illustrated by the fact that many gene families responsible for animal development, immunity and multicellular organization evolved through shuffling of protein domains that first originated in the choanozoan stem lineage together with ancient or animal-specific domains"



Richter *et al* (2018) eLife

Choanoflagellate – environmental cues to aggregate



The bacterium *Vibrio fischeri* induces mating in the choanoflagellate *S. rosetta*

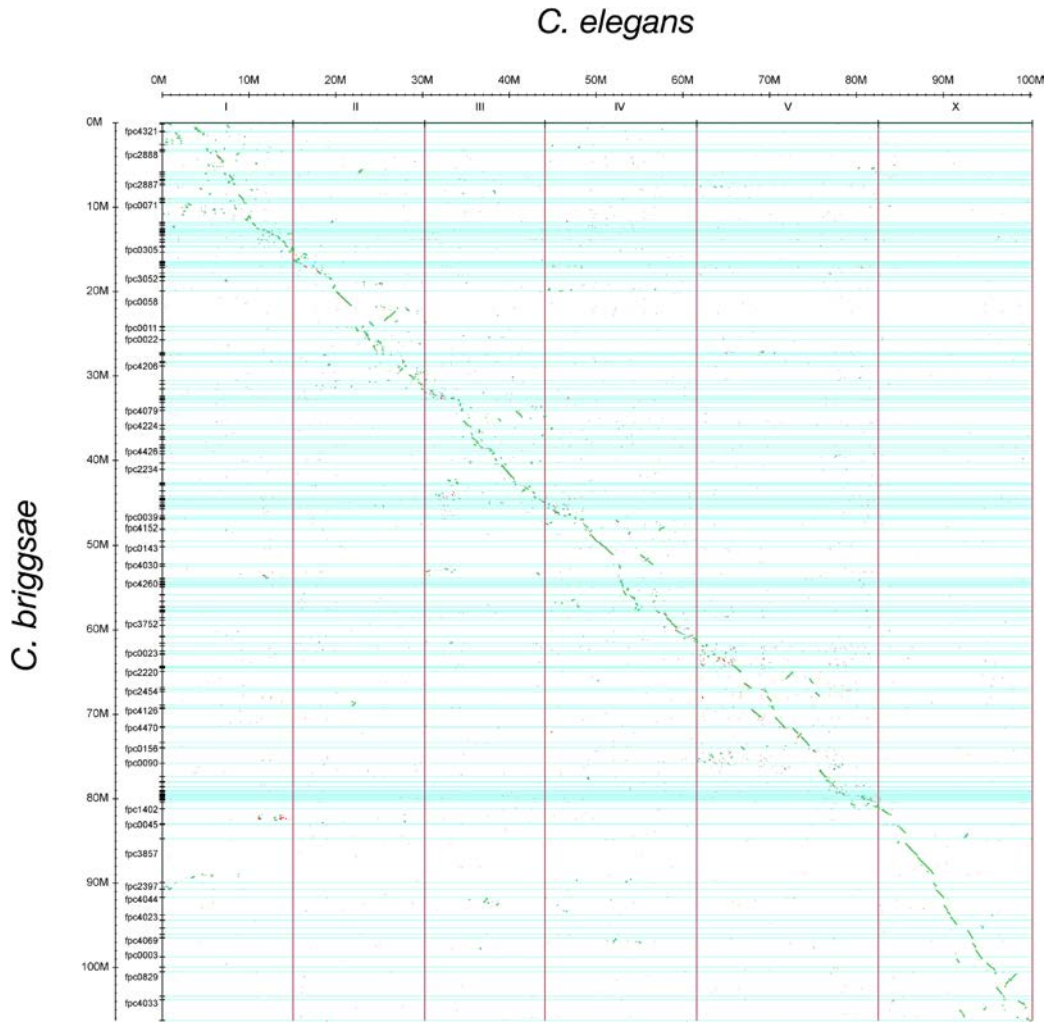
- The "aphrodisiac" produced by *V. fischeri* is a chondroitinase that we name EroS
- The enzymatic activity of EroS is required for this function
- Chondroitin sulfate, the EroS substrate, evolved before the origin of animals

https://www.sciencedirect.com/science/article/pii/S0092867417309303?via%3Dihub

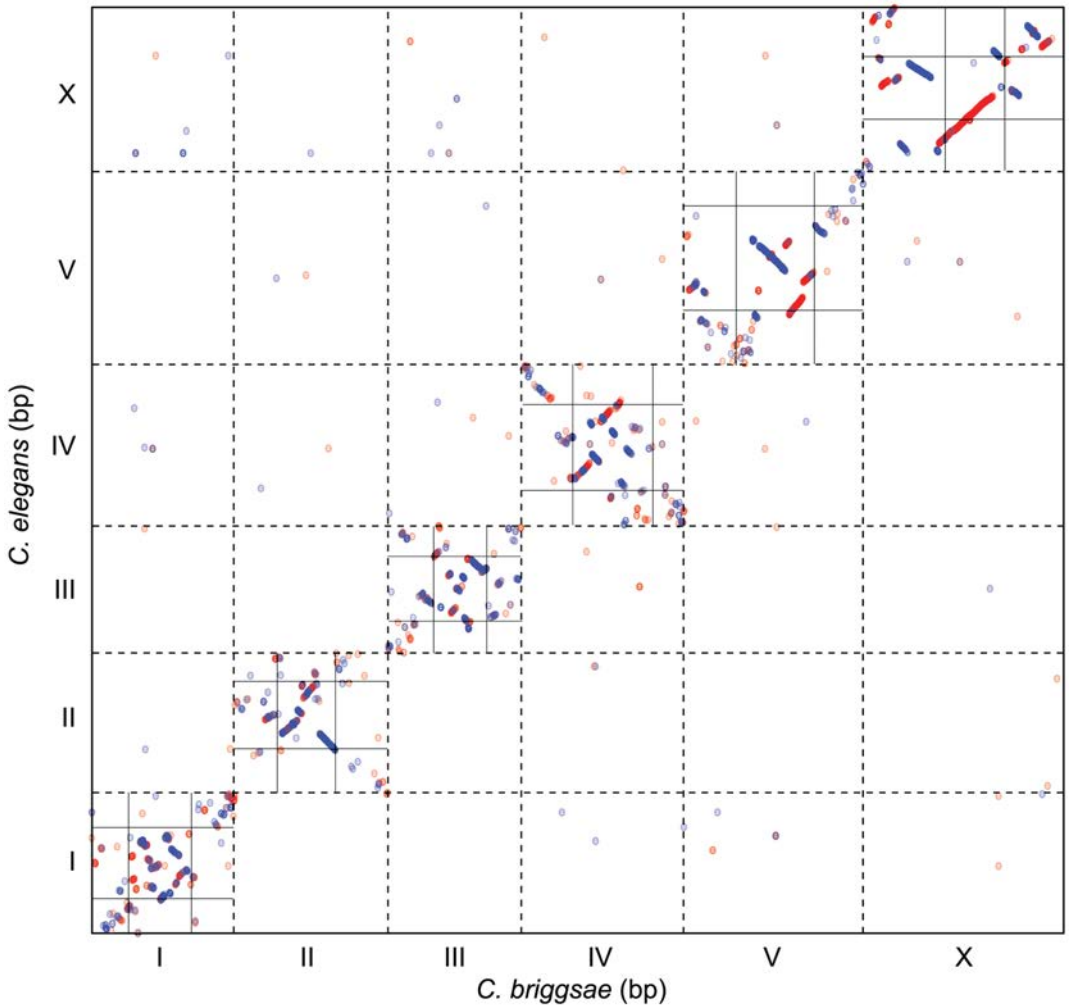
Woznica *et al* (2017) Cell

Some caveats

Assembly quality likely to influence synteny observation



Stein *et al.,* PLOS Genetics (2003)



Ross *et al.,* PLOS Genetics (2011)

Syteny based scaffolding: use with caution

Tang *et al. Genome Biology* (2015) 16:3
DOI 10.1186/s13059-014-0573-1

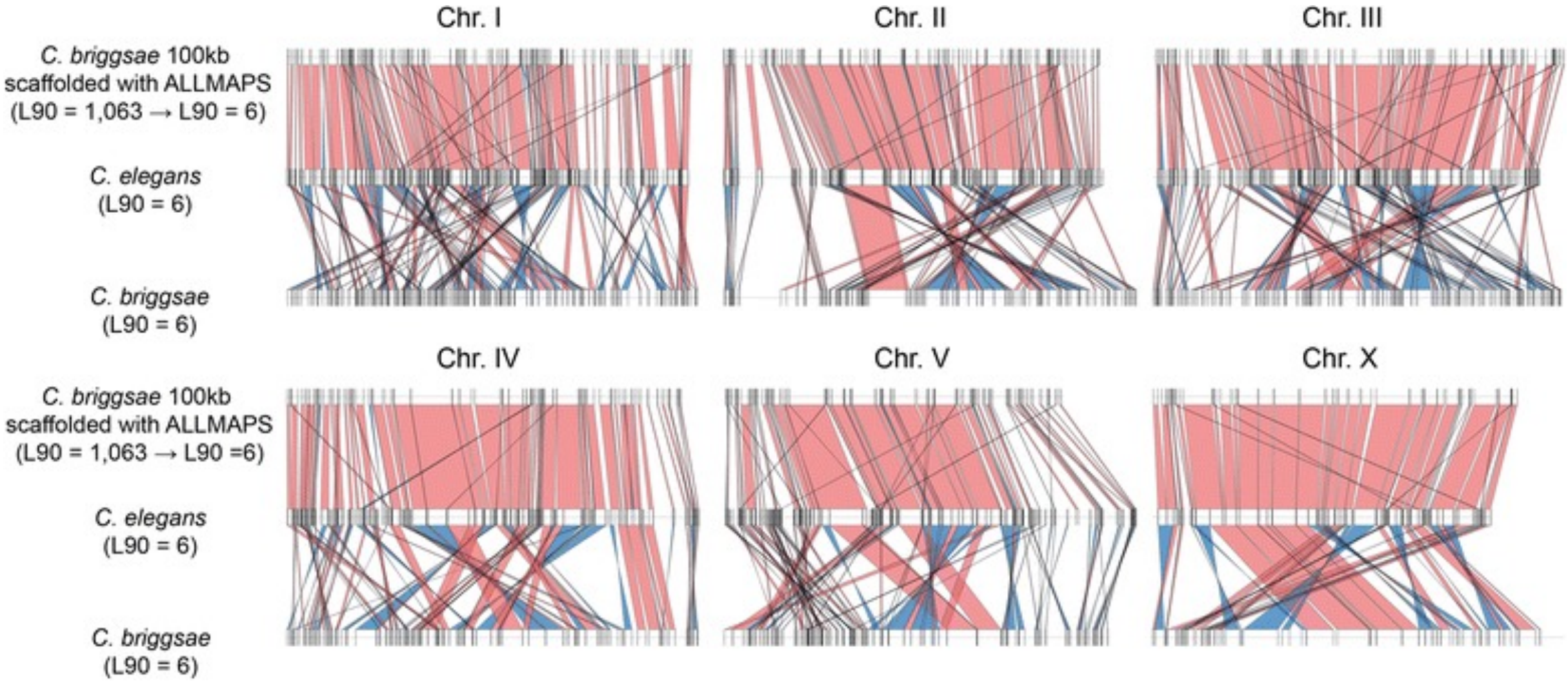
METHOD Open Access

ALLMAPS: robust scaffold ordering based on multiple maps

Haibao Tang[1,2,3*], Xingtan Zhang[4], Chenyong Miao[1], Jisen Zhang[1], Ray Ming[1], James C Schnable[3,5], Patrick S Schnable[3,6], Eric Lyons[2] and Jianguo Lu[7]

for example, in 'orphan' species where there is little research investment in the past, **we can still create consensus chromosomal assemblies based on comparative maps against multiple, closely-related genomes as a collection of 'references'** … Correct?

Syteny based scaffolding: use with caution



Homology beyond level of genes

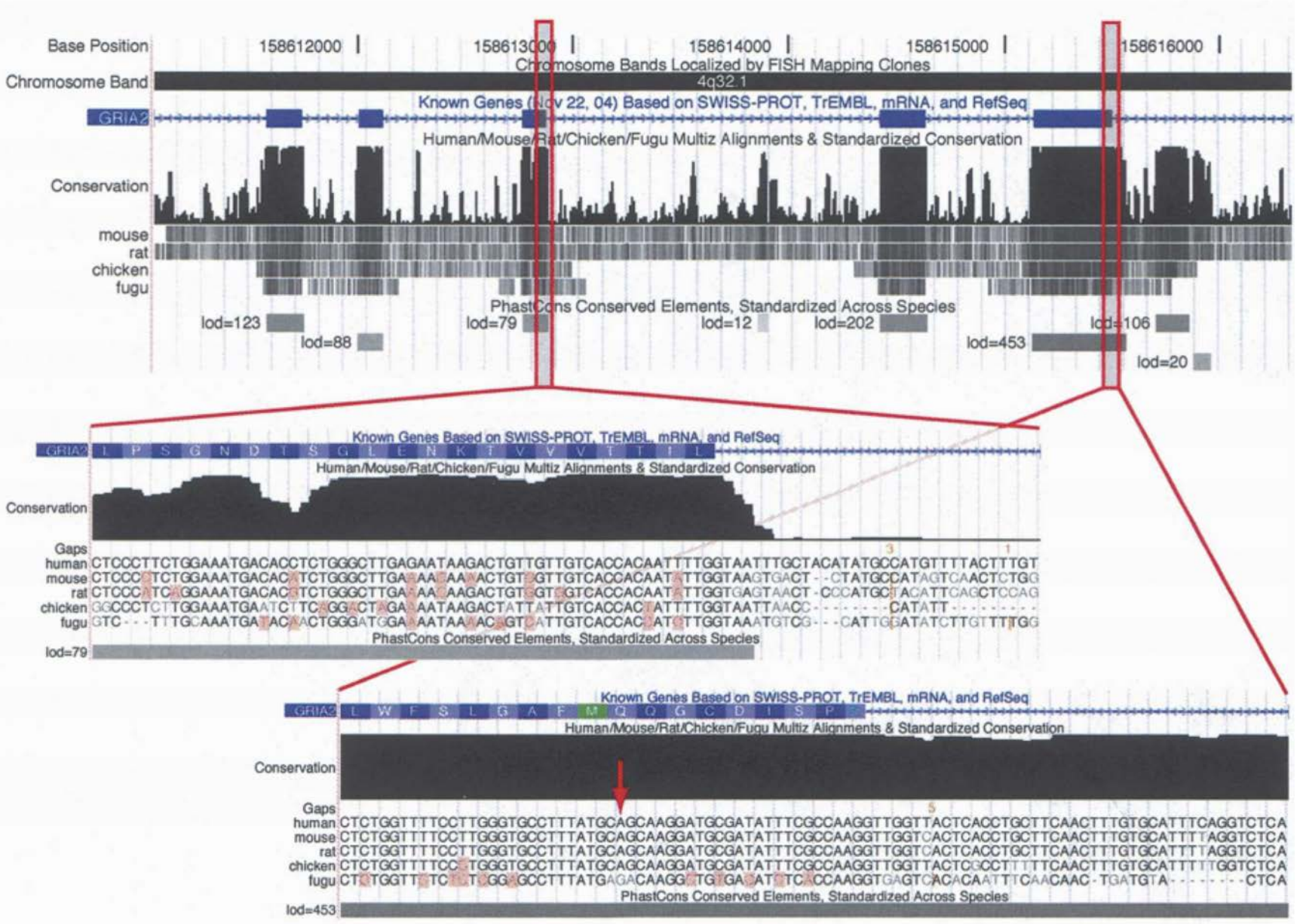
Article

Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes

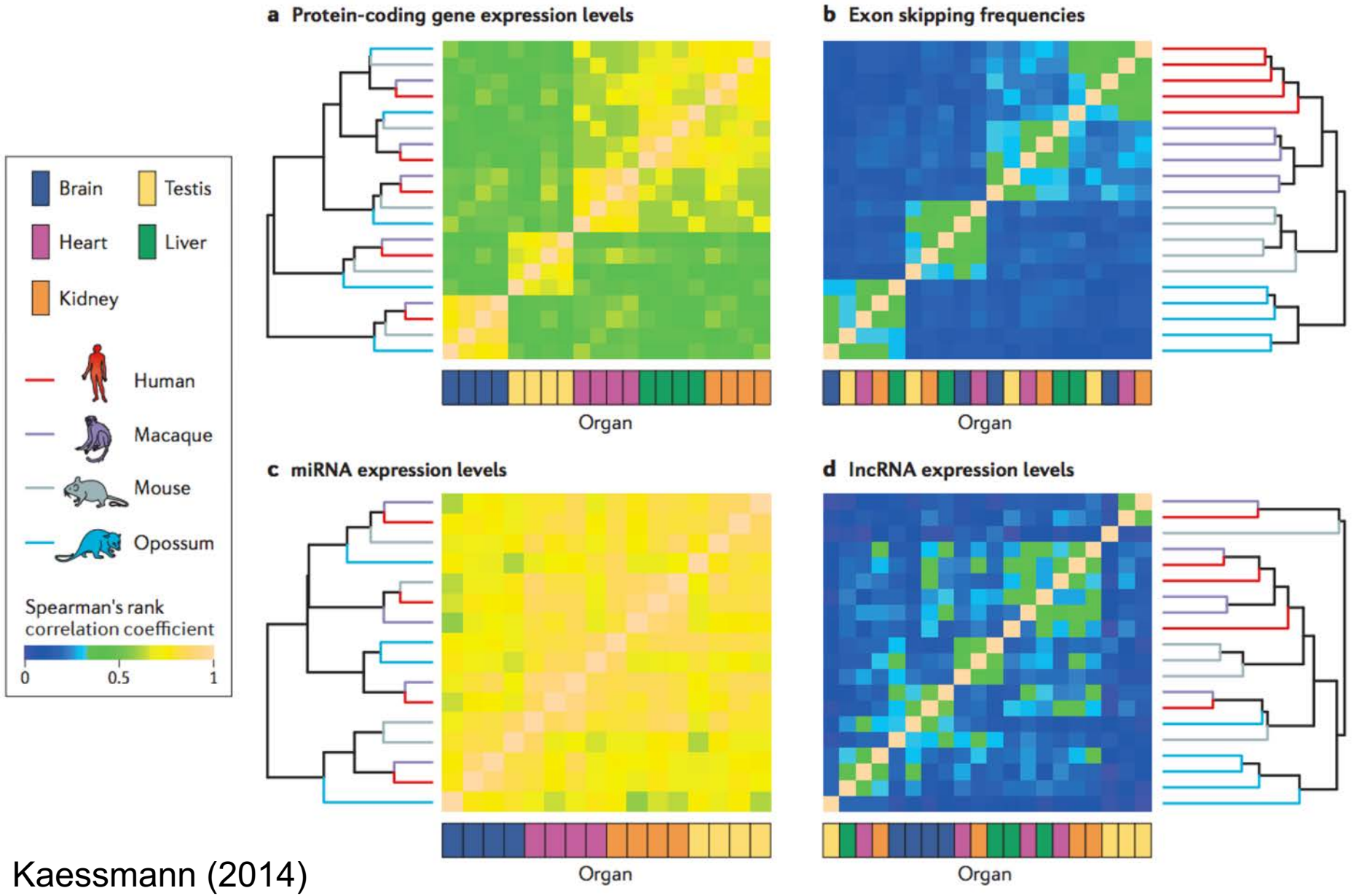
Adam Siepel,[1,6] Gill Bejerano,[1] Jakob S. Pedersen,[1] Angie
Kate Rosenbloom,[1] Hiram Clawson,[1] John Spieth,[4] LaDe
Stephen Richards,[5] George M. Weinstock,[5] Richard K. W
W. James Kent,[1] Webb Miller,[3] and David Haussler[1,2]

*[1]Center for Biomolecular Science and Engineering, [2]Howard Hughes Medical Institut
Cruz, California 95064, USA; [3]Center for Comparative Genomics and Bioinformatics
Park, Pennsylvania 16802, USA; [4]Genome Sequencing Center, Washington Universi
63108, USA; [5]Human Genome Sequencing Center, Department of Molecular and H
Houston, Texas 77030, USA*

PhastCons

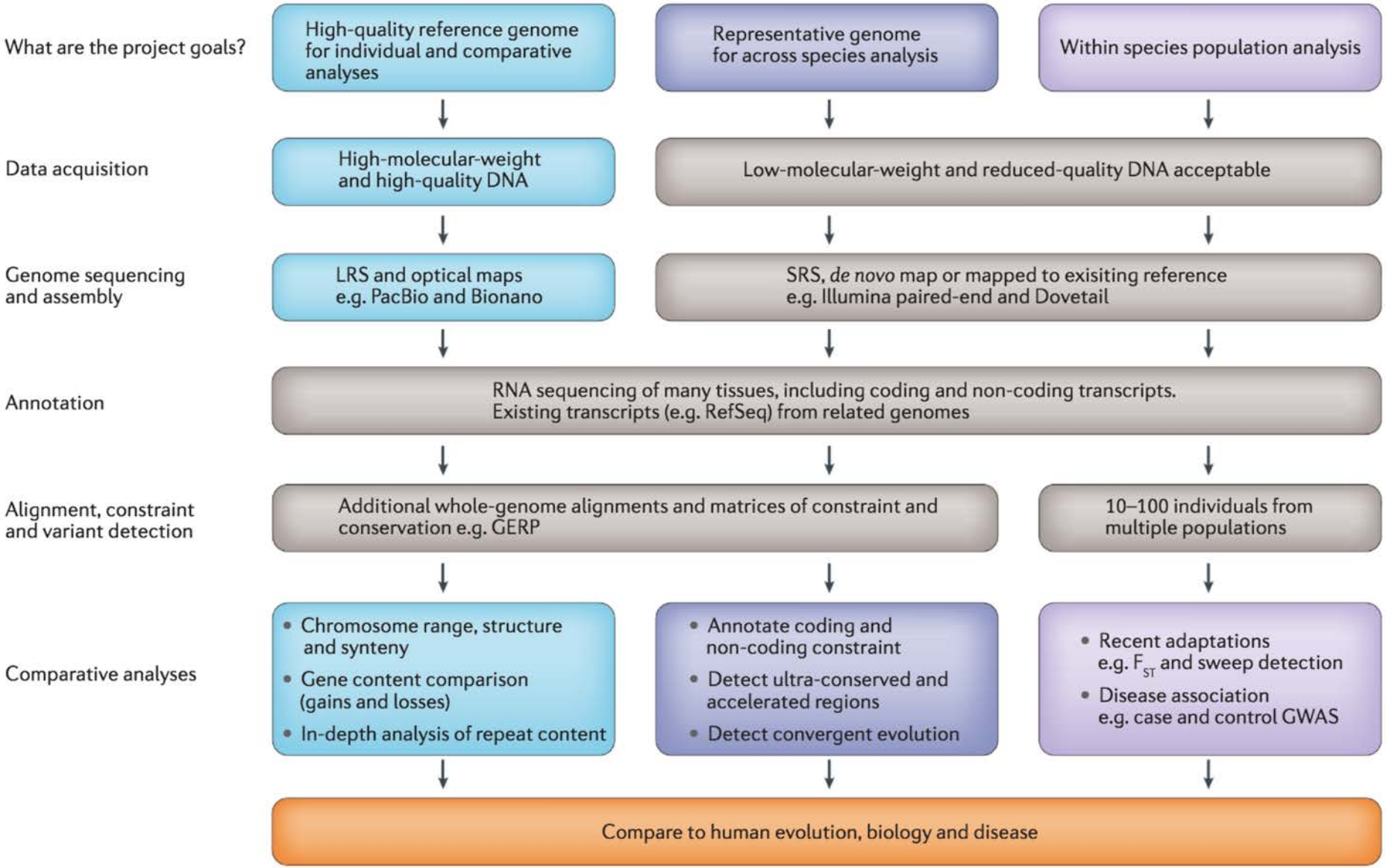


Global patterns of evolution for different aspects of the transcriptome



Necsulea and Kaessmann (2014)

Designing a sequencing project: 2017 version



Meadows and Toh (2017) Nature Review Genetics

Evolution at chromosome level

- Autosomes vs. Sex chromosomes
- Euchromatin / Heterochromatin
- Chromosome arm versus center
- Large and small chromosomes
- Ploidy and polysomy

Looking back in 2003

Group	Species	Common	Size (Mb)	Chromosome (1N)	Gene no.	Repeat %
Mammal	*Homo sapiens*	Human	2900	23	30,000	46
Mammal	*Mus musculus*	House mouse	2500	20	30,000	38
Fish	*Takifugu rubripes*	Tiger pufferfish	400	22 (?)	30,000	<10
Urochordate	*Ciona intestinales*	Sea squirt	155	14	16,000	~10
Insect	*Anopheles gambiae*	Malaria mosquito	280	3	14,000	16
Insect	*Drosophila melanogaster*	Fruit fly	137	4	13,600	2
Nematode	*Caenorhabditis elegans*	Nematode worm	97	6	19,100	<1
Apicomplexa	*Plasmodium falciparum*	Human malaria parasite	23	14	5,300	<1
Apicomplexa	*Plasmodium yoelli*	Rodent malaria parasite	25	14	5,300	<1
Dictyosteliida	*Dictyostelium discoideum**	Social amoeba	34	6	2,800	<1
Protozoan	*Leishmania major**	Intracellular parasite	34	36	9,800	<1
Fungi	*Saccharomyces cerevisiae*	Brewer's yeast	12	16	5,700	2.4
Fungi	*Schizosaccharomyces pombe*	Fission yeast	13.8	3	4,900	0.35
Microsporidium	*Encephalitozoon cuniculi*	Intracellular parasite	2.5	11	2,000	<0.1
Angisoperm	*Arabidopsis thaliana*	Mustard weed	125	5	25,500	14
Angiosperm	*Oryza sativa*	Rice	400	12	32000–50000	?

Chromosomal Rearrangements and Repeats: Cause or Consequence?

Centromeric and Telomeric Regions—Sites of Rapid Genomic Change

Duplications: Engines of Gene and Genome Evolution?

Synteny: Fragile Versus Random Breakage Model?

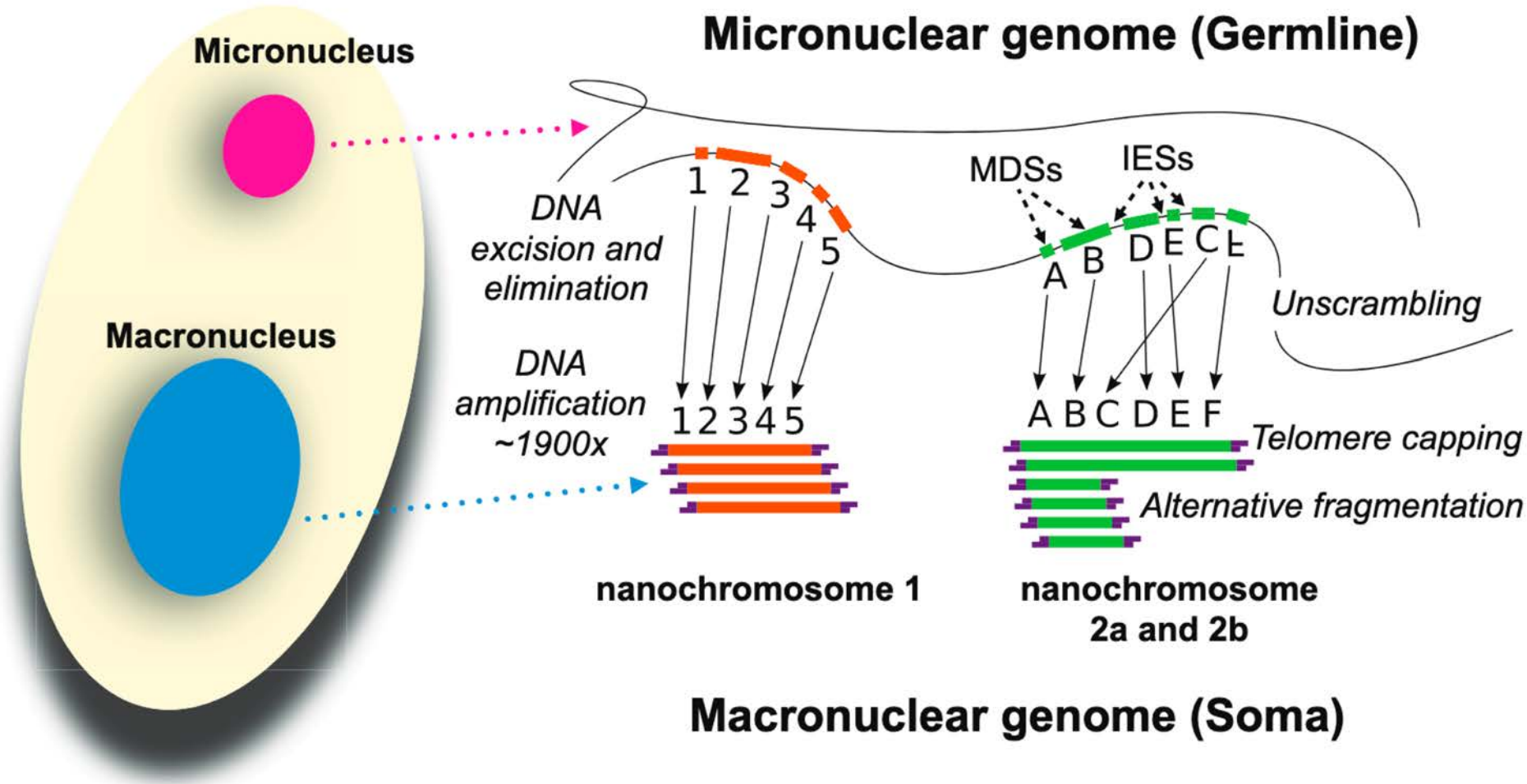
http://science.sciencemag.org/content/sci/301/5634/793.full.pdf

OPEN ACCESS Freely available online

PLOS | BIOLOGY

The *Oxytricha trifallax* Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes

Estienne C. Swart[1],
Jaspreet S. Khurana
Robert S. Fulton[2,3],
John C. Matese[7], La
Thomas A. Jones[6],
Elaine R. Mardis[2,3],



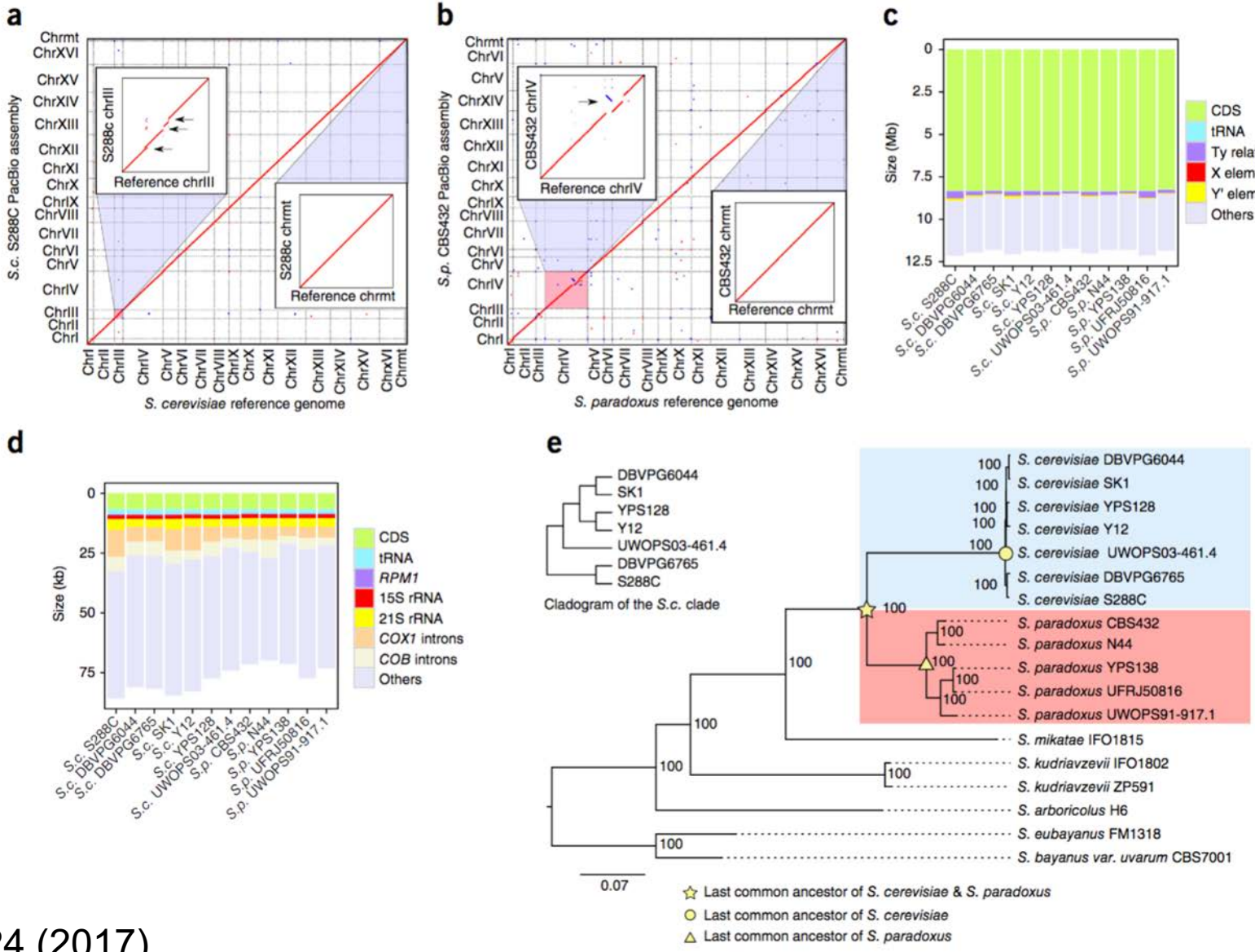
	Genome size	Genes	Chromo-somes	Ploidy	Alternative fragmen-tation
Oxytricha	**~50 Mb**	**~18,400**	**~15,600**	**Variable ~1,900[a]**	**Yes**
Stylonychia	~50 Mb[b]	~12,000[c]	~10-15,000[b]	Variable ~15,000[b]	Yes
Euplotes	~50 Mb?	?	nano ?	Variable ~2,000[d]	No?
Nyctotherus	~50 Mb[e]	?	nano ?	Variable[e] ?	?
Tetrahymena	105 Mb[f]	24,700[g]	225[f]	45[f]	limited[h,i]
Ichthyophthirius	49 Mb[j]	8,100[j]	71[j]	~12,000[j]	?
Paramecium	72 Mb[k]	40,000[k]	~200[k]	~800[l]	limited[l]
Perkinsus	87 Mb	23,700	?	1	NA
Plasmodium	23 Mb[m]	5,300[m]	14[m]	1	NA

Swart *et al* (2013) PLOS Biology

Contrasting evolutionary genome dynamics between domesticated and wild yeasts

Jia-Xing Yue[1], Jing Li[1], Louise Aigrain[2], Johan Hallin[1], Karl Persson[3], Karen Oliver[2], Anders Bergström[2], Paul Coupland[2,5], Jonas Warringer[3], Marco Cosentino Lagomarsino[4], Gilles Fischer[4], Richard Durbin[2] & Gianni Liti[1]

- long-read sequencing to generate **end-to-end genome assemblies** for **12 strains** representing major subpopulations of the partially domesticated yeast *Saccharomyces cerevisiae* and its wild relative *S. paradoxus*.



Nature Genetics **volume 49**, pages 913–924 (2017)

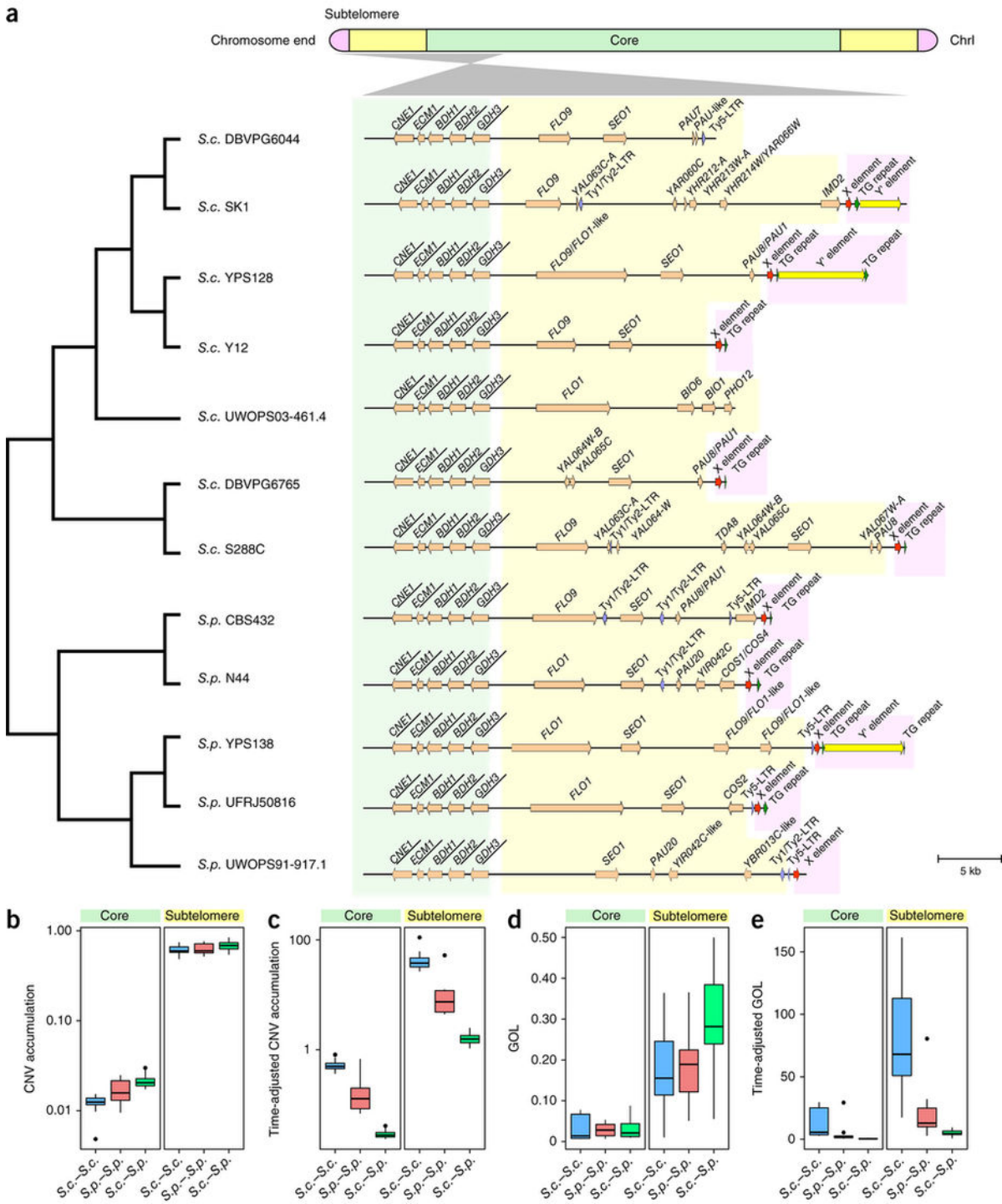
Contrasting evolutionary genome dynamics between domesticated and wild yeasts

Jia-Xing Yue[1], Jing Li[1], Louise Aigrain[2], Johan Hallin[1], Karl Persson[3], Karen Oliver[2], Anders Bergström[2], Paul Coupland[2,5], Jonas Warringer[3], Marco Cosentino Lagomarsino[4], Gilles Fischer[4], Richard Durbin[2] & Gianni Liti[1]

- enable precise definition of chromosomal boundaries between cores and subtelomeres
- *S. paradoxus* shows faster accumulation of balanced rearrangements (inversions, reciprocal translocations and transpositions), *S. cerevisiae* accumulates unbalanced rearrangements (novel insertions, deletions and duplications) more rapidly.

- Such striking contrasts between wild and domesticated yeasts are likely to reflect the influence of human activities on structural genome evolution.

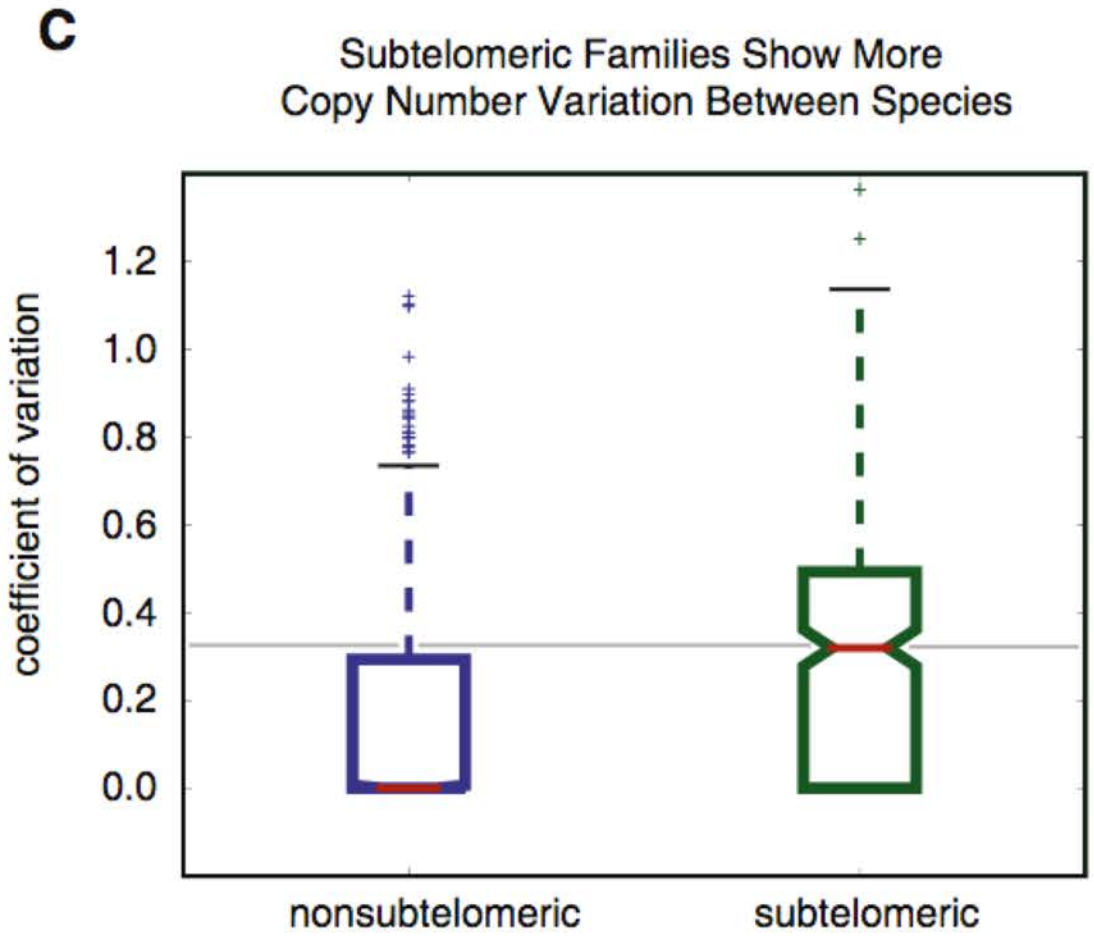
Nature Genetics **volume 49**, pages 913–924 (2017)

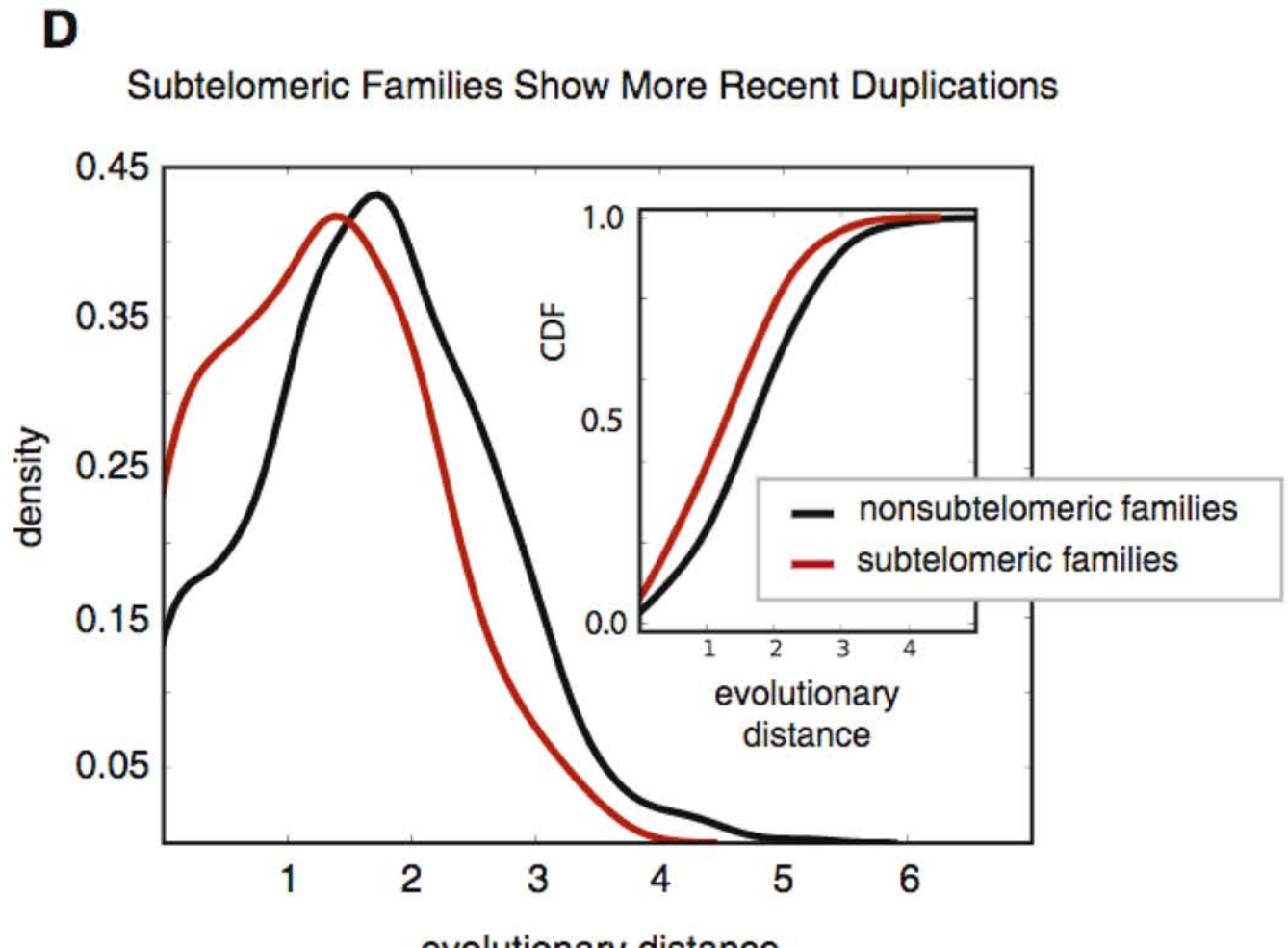


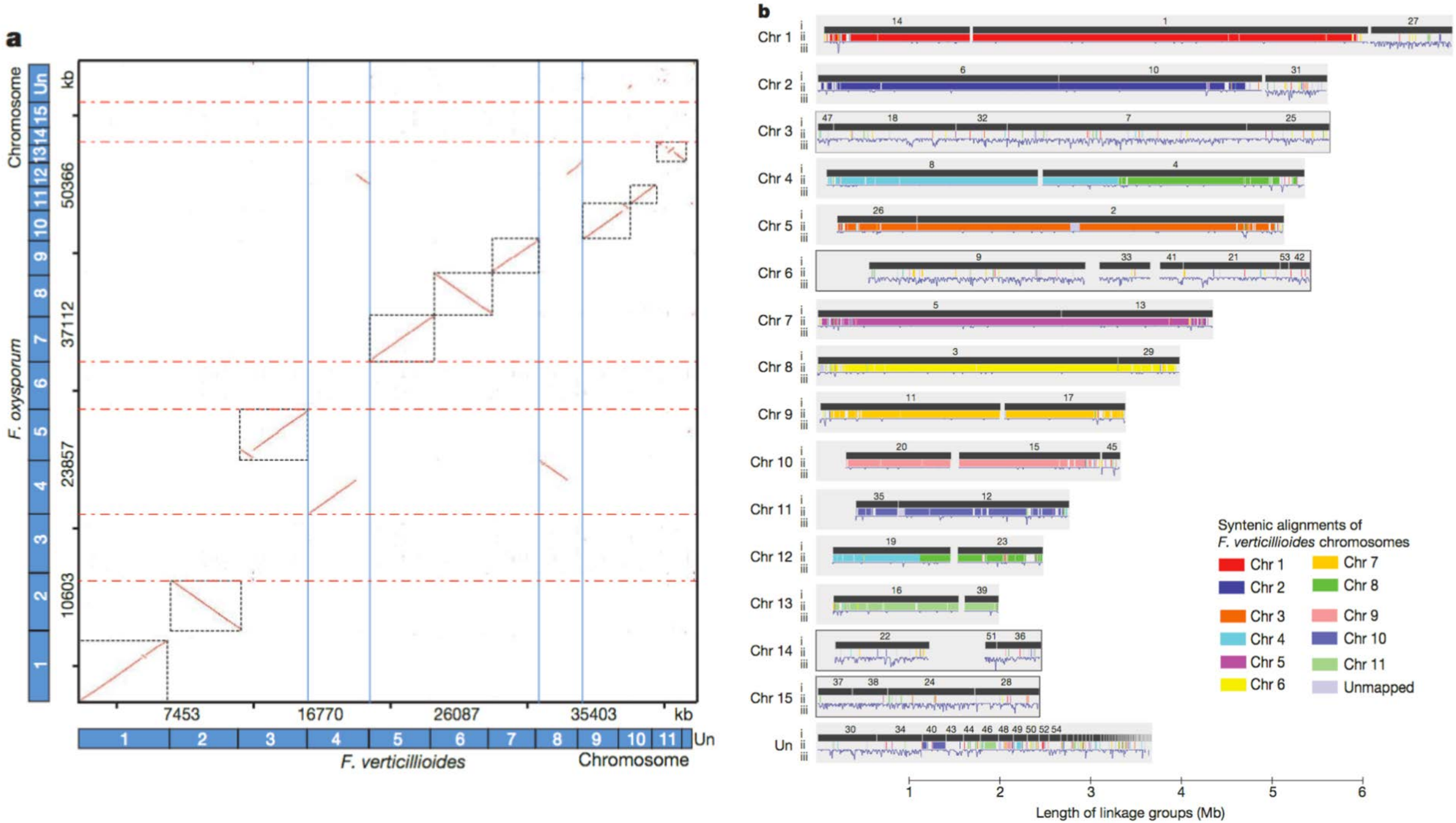
Current Biology 20, 895–903, May 25, 2010 ©2010 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2010.04.027

Rapid Expansion and Functional Divergence of Subtelomeric Gene Families in Yeasts

…our computational and experimental analyses show that the **extraordinary instability of eukaryotic subtelomeres supports rapid adaptation to novel niches by promoting gene recombination and duplication followed by functional divergence of the alleles**






a
Chromosome
F. oxysporum
kb
50366
37112
23857
10603
7453
16770
26087
35403
kb
F. verticillioides
Chromosome
b
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Un
Syntenic alignments of
F. verticillioides chromosomes
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Unmapped
Length of linkage groups (Mb)

LETTER

doi:10.1038/nature12235

Reversal of an ancient sex chromosome to an autosome in *Drosophila*

Beatriz Vicoso[1] & Doris Bachtrog[1]

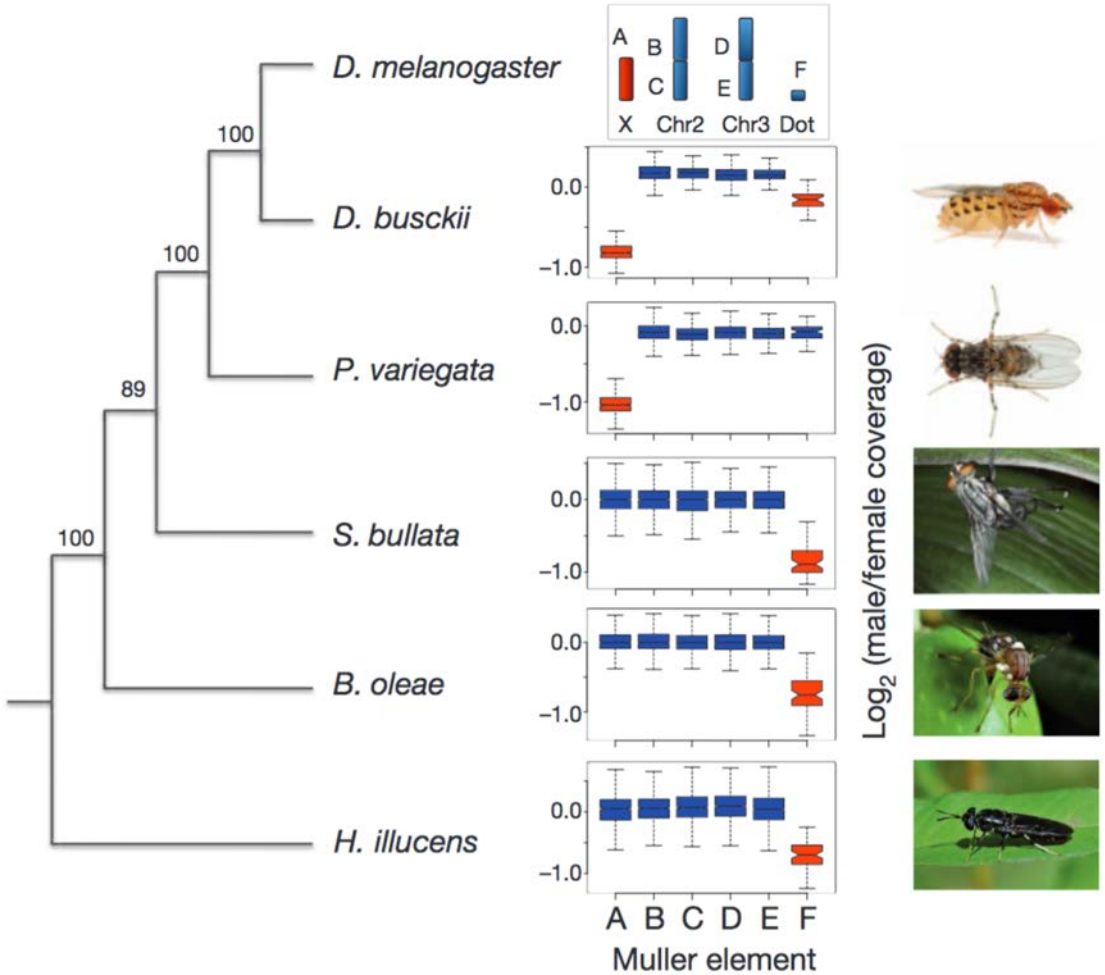


Figure 1 | Sex chromosomes in higher Diptera revealed by genome analysis. Evolutionary relationship inferred from 185 conserved protein-coding genes (93,134 amino acids) using PhyML (with bootstrap values indicated at the nodes), and male-to-female coverage ratio across chromosome elements (Muller elements A–F) in the Diptera species studied. X chromosomes (red) have only half the read coverage in males versus females. Boxes extend from the first to the third quartile and whiskers to the most extreme data point within 1.5 times the interquantile range.

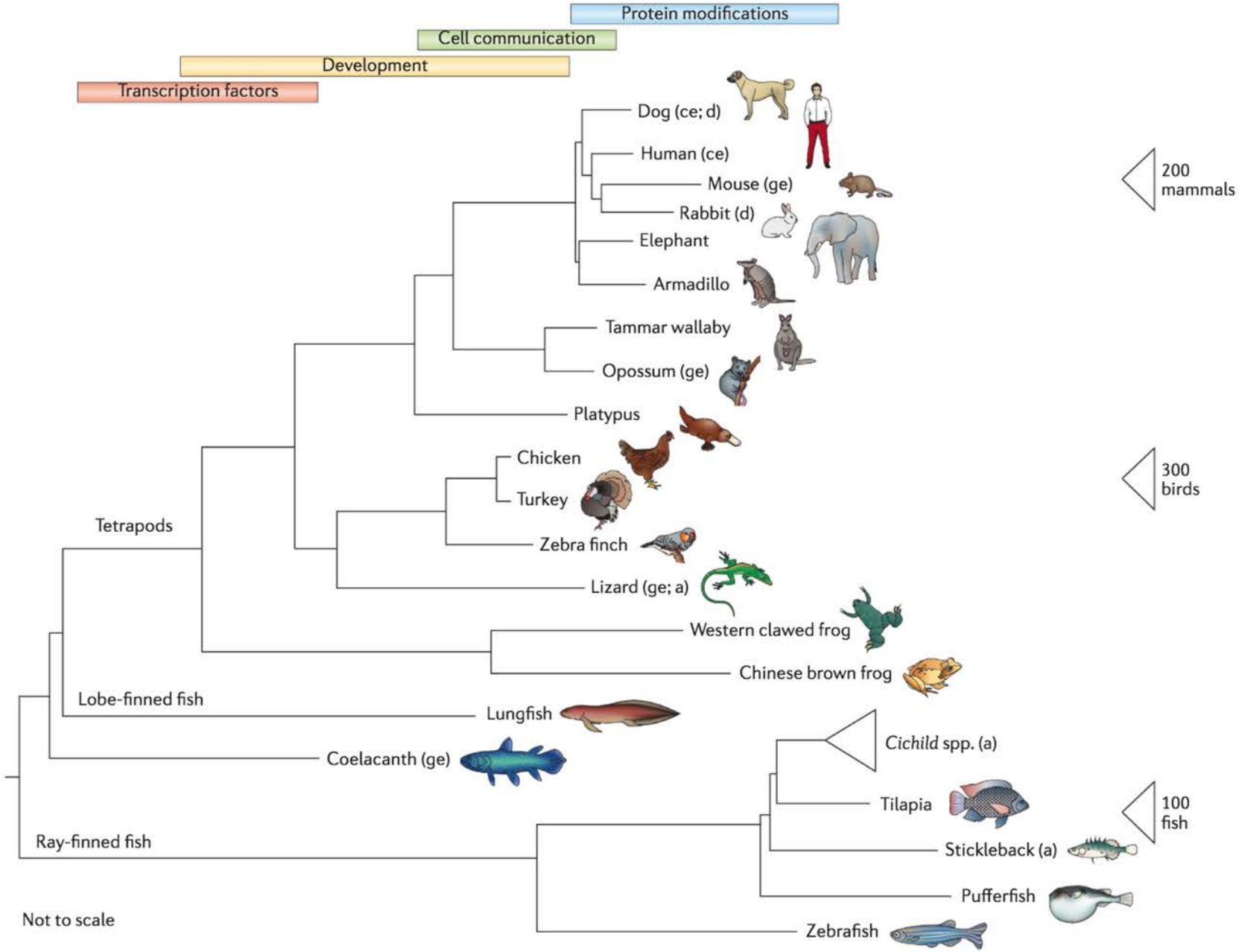
Good review – recent update

Dissecting evolution and disease using comparative vertebrate genomics

Jennifer R. S. Meadows[1] and Kerstin Lindblad-Toh[1,2]

Abstract | With the generation of more than 100 sequenced vertebrate genomes in less than 25 years, the key question arises of how these resources can be used to inform new or ongoing projects. In the past, this diverse collection of sequences from human as well as model and non-model organisms has been used to annotate the human genome and to increase the understanding of human disease. In the future, comparative vertebrate genomics in conjunction with additional genomic resources will yield insights into the processes of genome function, evolution, speciation, selection and adaptation, as well as the quantification of species diversity. In this Review, we discuss how the genomics of non-human organisms can provide insights into vertebrate biology and how this can contribute to the understanding of human physiology and health.

Meadows and Toh (2017) Nature Review Genetics



Meadows and Toh (2017) Nature Review Genetics

Why comparative genomics? – a summary

How genome evolved;
How genome functions

- At various resolution (whole genome, chromosomes, regions, genes, base pairs)
- Conservation, Duplication, Species specific genes
- Inferring Orthologs and paralogs
- Gene families (clusters) of paralogs, of orthologs
- Conserved or specialized domains in clusters of paralogs, orthologs
- Gene transfer, introgression between species
- Relate genotypes to phenotypes
- Identify the genomic basis of key phenotypes

Fred Tekaia (slideshare)

Genomics of Eukaryotic microorganisms – a summary

How genome evolved;
How genome functions

- We are only at the beginning phase of
- Untapped diversity and mechanisms waiting to be discovered
- Arguably much more fascinating (?) than animals

Fred Tekaia (slideshare)

https://www.nature.com/subjects/comparative-genomics