

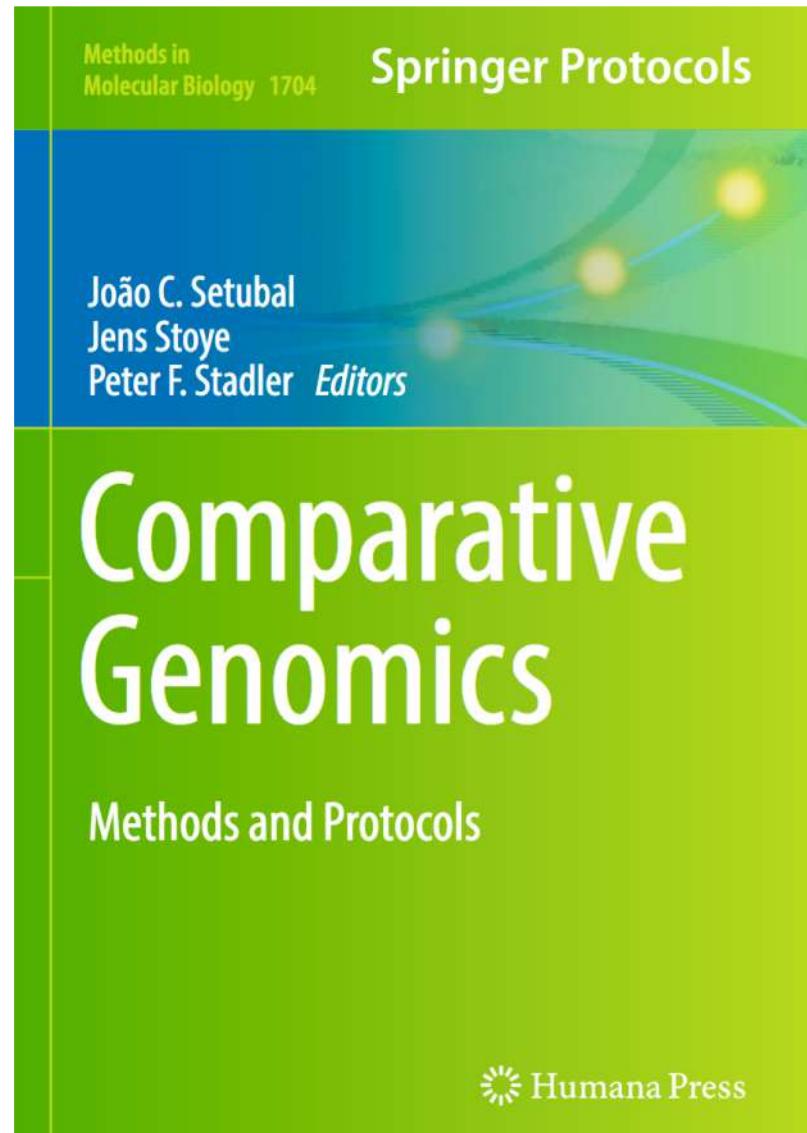
Comparative Genomics

Isheng Jason Tsai

30.10.2018



Recommended book



Comparative Genomics

Methods and Protocols

Edited by

João C. Setubal

Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, SP, Brazil

Jens Stoye

Faculty of Technology and Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany

Peter F. Stadler

*Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics,
University of Leipzig, Leipzig, Germany*

 Humana Press

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

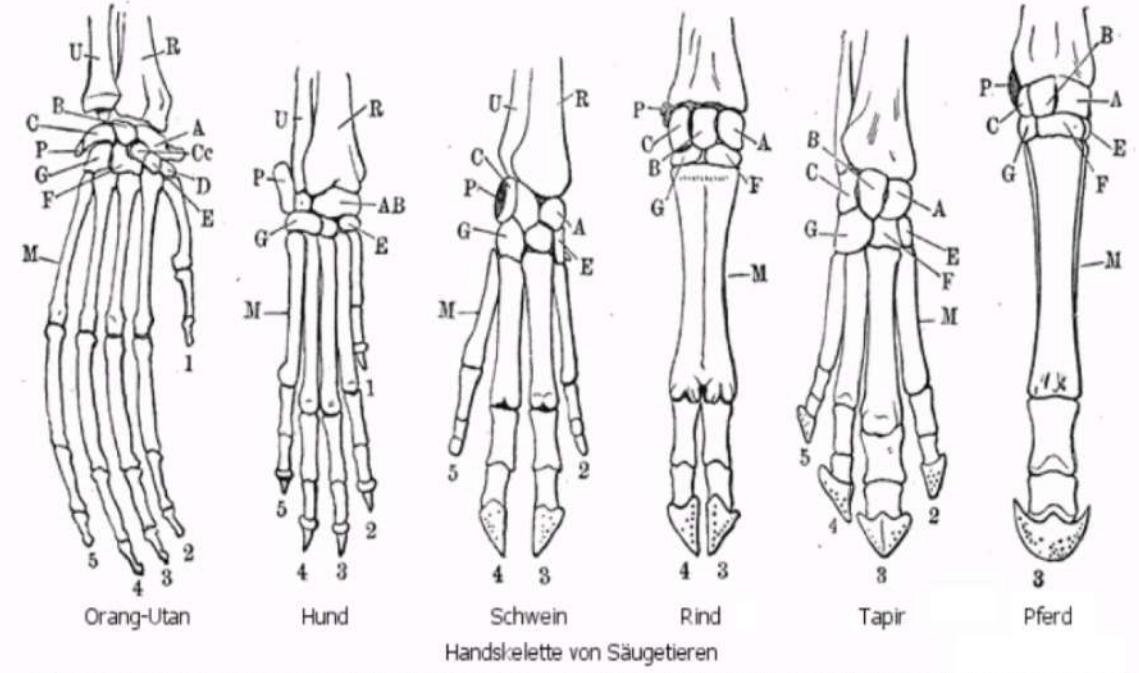
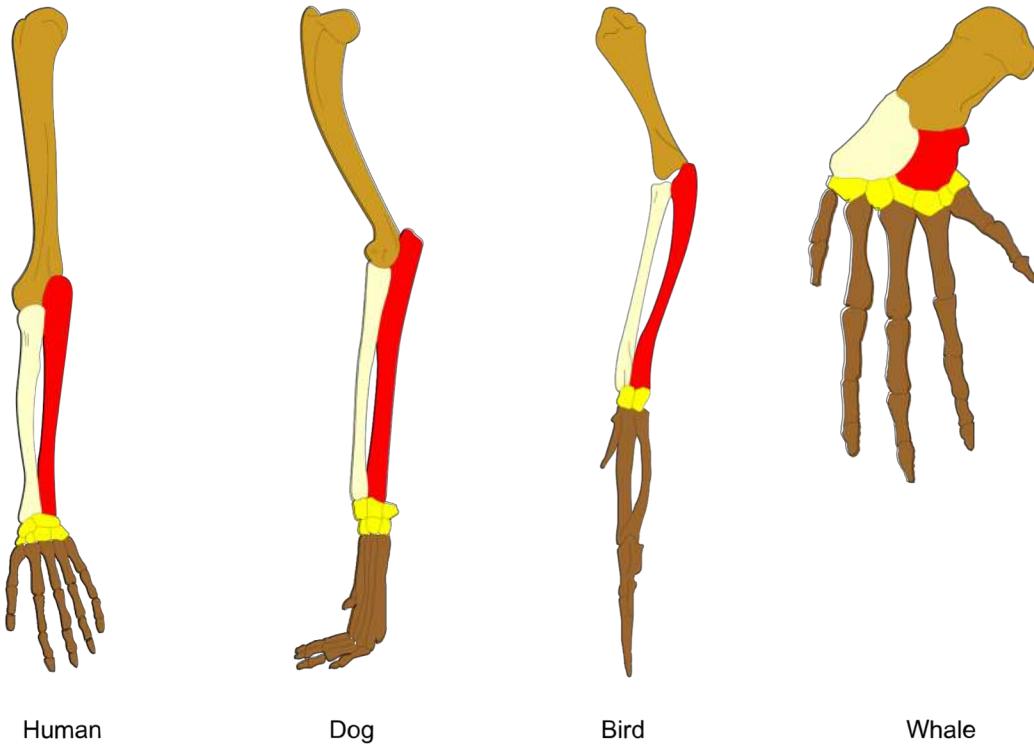
Homology

Termed before Darwin's time!



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

Homology



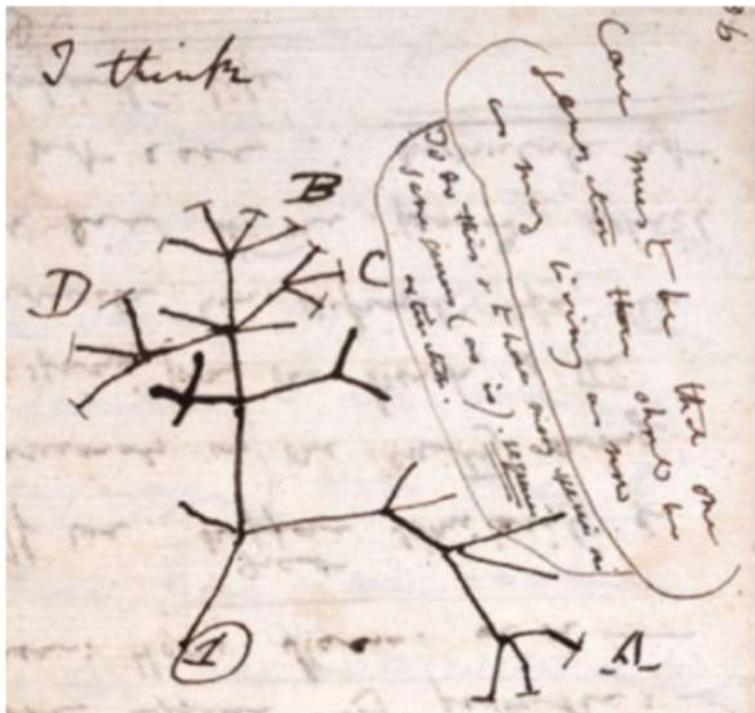
R Radius (Speiche), U Ulna (Elle), A-G, Cc, P Knochen des Carpus (Handwurzel): A Scaphoideum (Kahnbein), B Lunare (Mondbein), C Triquetrum (dreieckiges Bein), D Trapezium (großes vieleckiges Bein), E Trapezoides (kleines vieleckiges Bein), F Capitatum (Kopfbein), G Hamatum (Hafenbein), P Pisiforme (Erbsenbein), 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_\(biology\)](https://en.wikipedia.org/wiki/Homology_(biology))

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



CHAPTER VI.

DIFFICULTIES ON THEORY.

Difficulties on the theory of descent with modification—Transitions—Absence or rarity of transitional varieties—Transitions in habits of life—Diversified habits in the same species—Species with habits widely different from those of their allies—Organs of extreme perfection—Means of transition—Cases of difficulty—*Natura non facit saltum*—Organs of small importance—Organs not in all cases absolutely perfect—The law of Unity of Type and of the Conditions of Existence embraced by the theory of Natural Selection, 154

CHAPTER XIII.

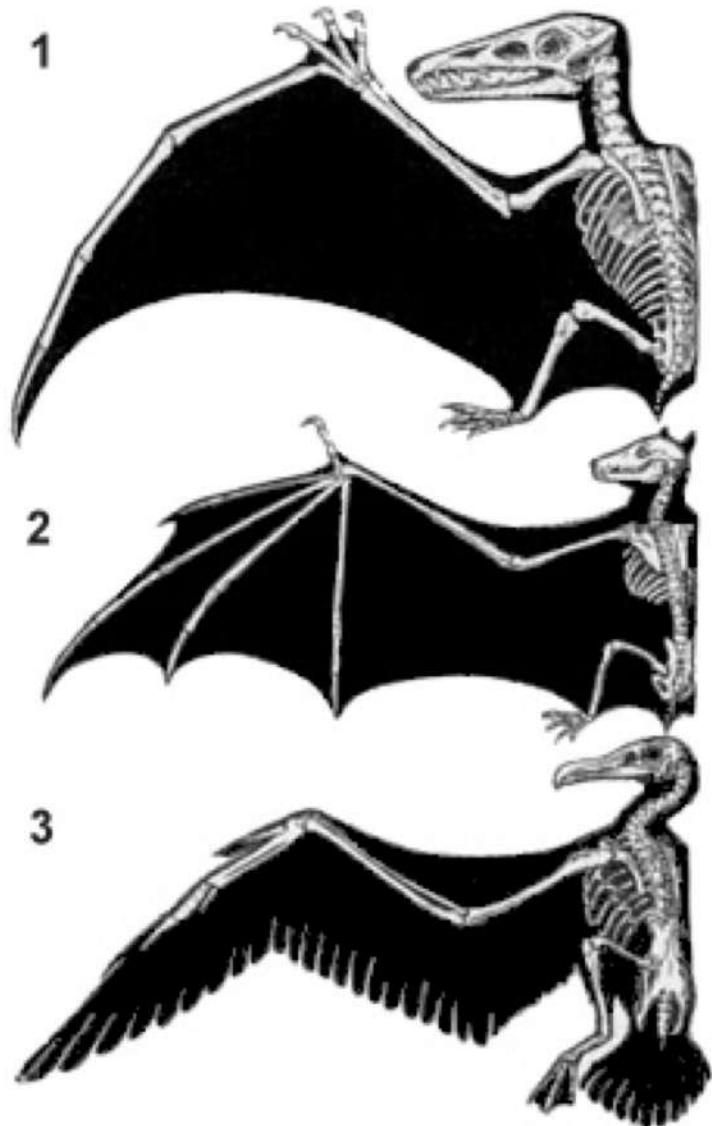
MUTUAL AFFINITIES OF ORGANIC BEINGS: MORPHOLOGY: EMBRYOLOGY: RUDIMENTARY ORGANS.

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

Homology

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

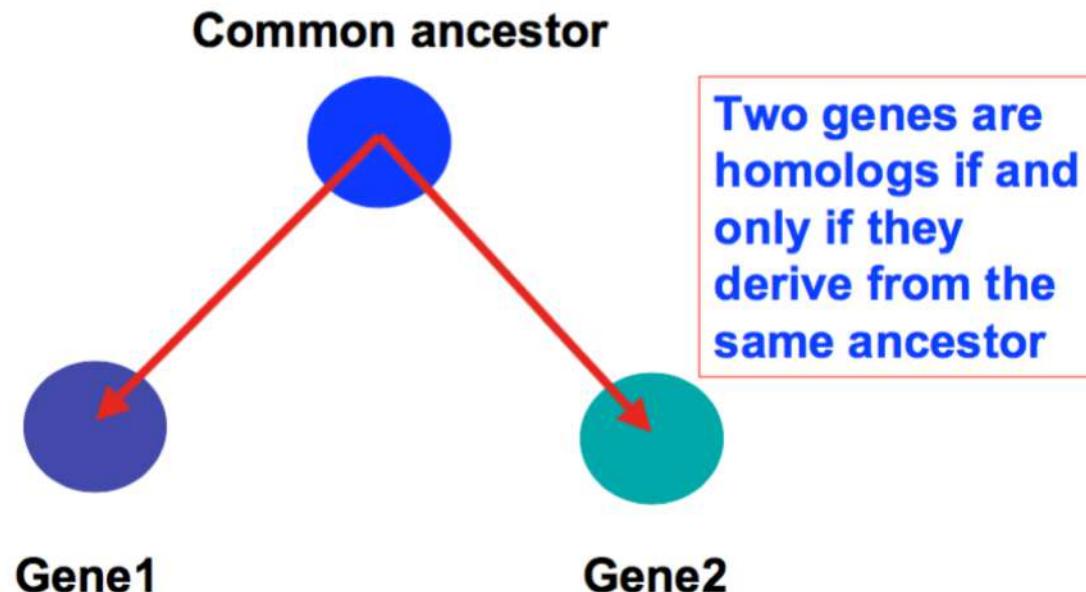
Homologs (any features: genes, trait, morphology) share **ancestry**



Search for similarity , collinearity, conservation of morphological characters

Search for similarity

One of the most frequent activity in Bioinformatics



Homology is almost uniquely inferred by sequence similarity

Beware ; why?

~~**Significant homology**~~

~~**Weak homology**~~

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

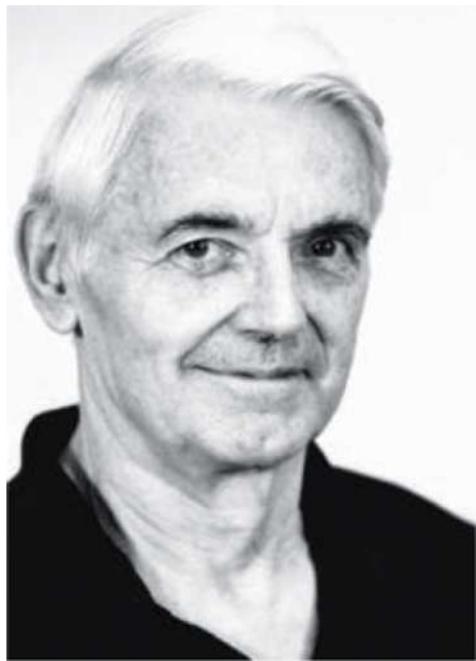
Significant similarity

55% married?
45% grandmom?

Weak similarity

DISTINGUISHING HOMOLOGOUS FROM
ANALOGOUS PROTEINS (1970)

WALTER M. FITCH



1929 - 2011

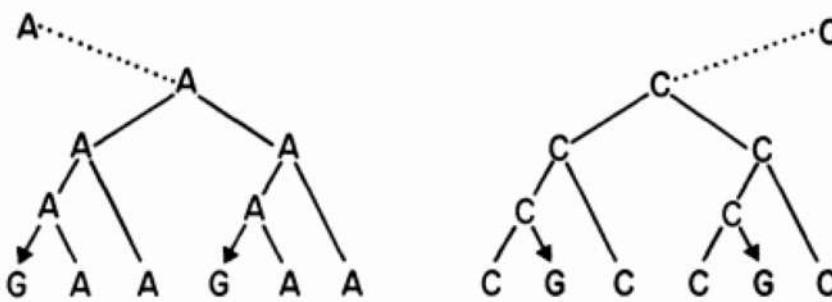
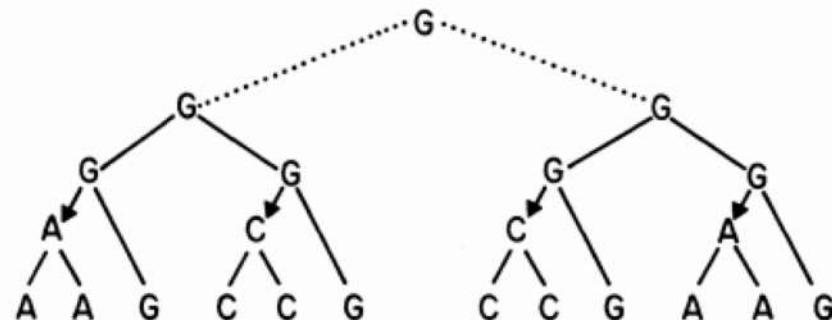
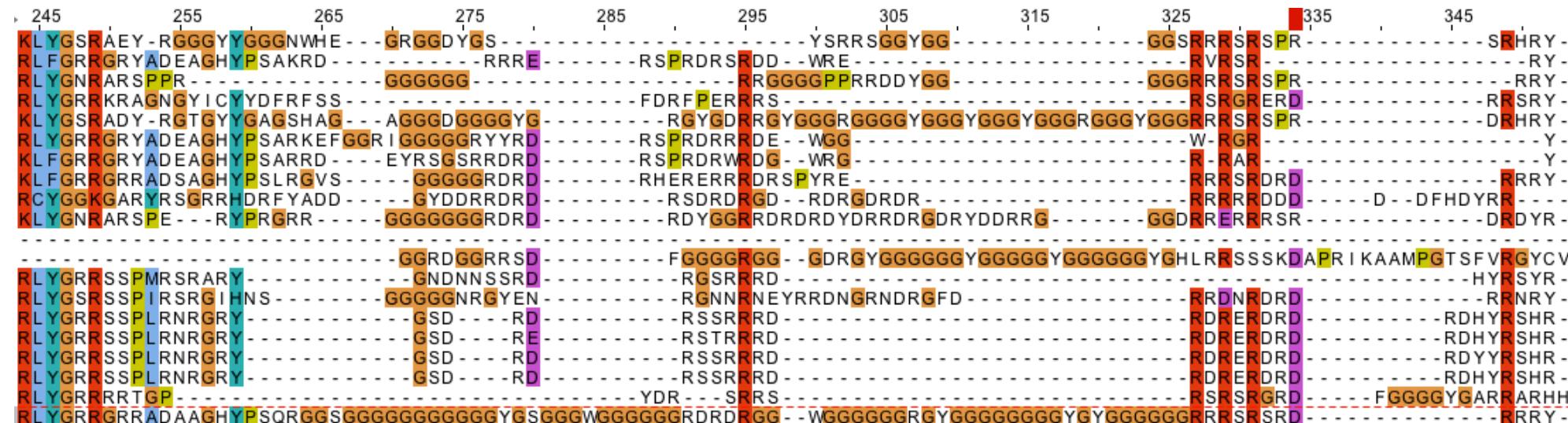
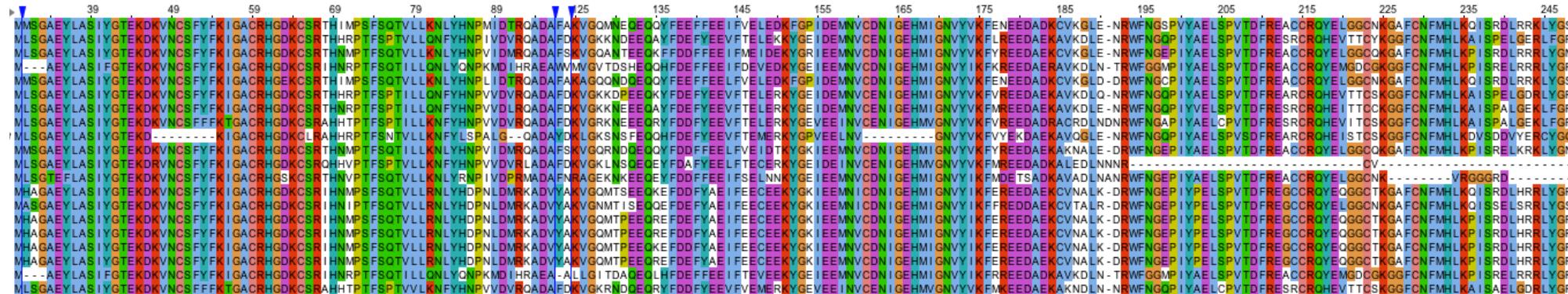


FIG. 1.—Distinguishing convergent from divergent types of nucleotide replacement patterns. Given are two groups of species (related within each group as shown by the solid lines) together with the nucleotide present at a specific position of the gene for each member species as shown at the branch tips. 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.

Extension of homology to sequences

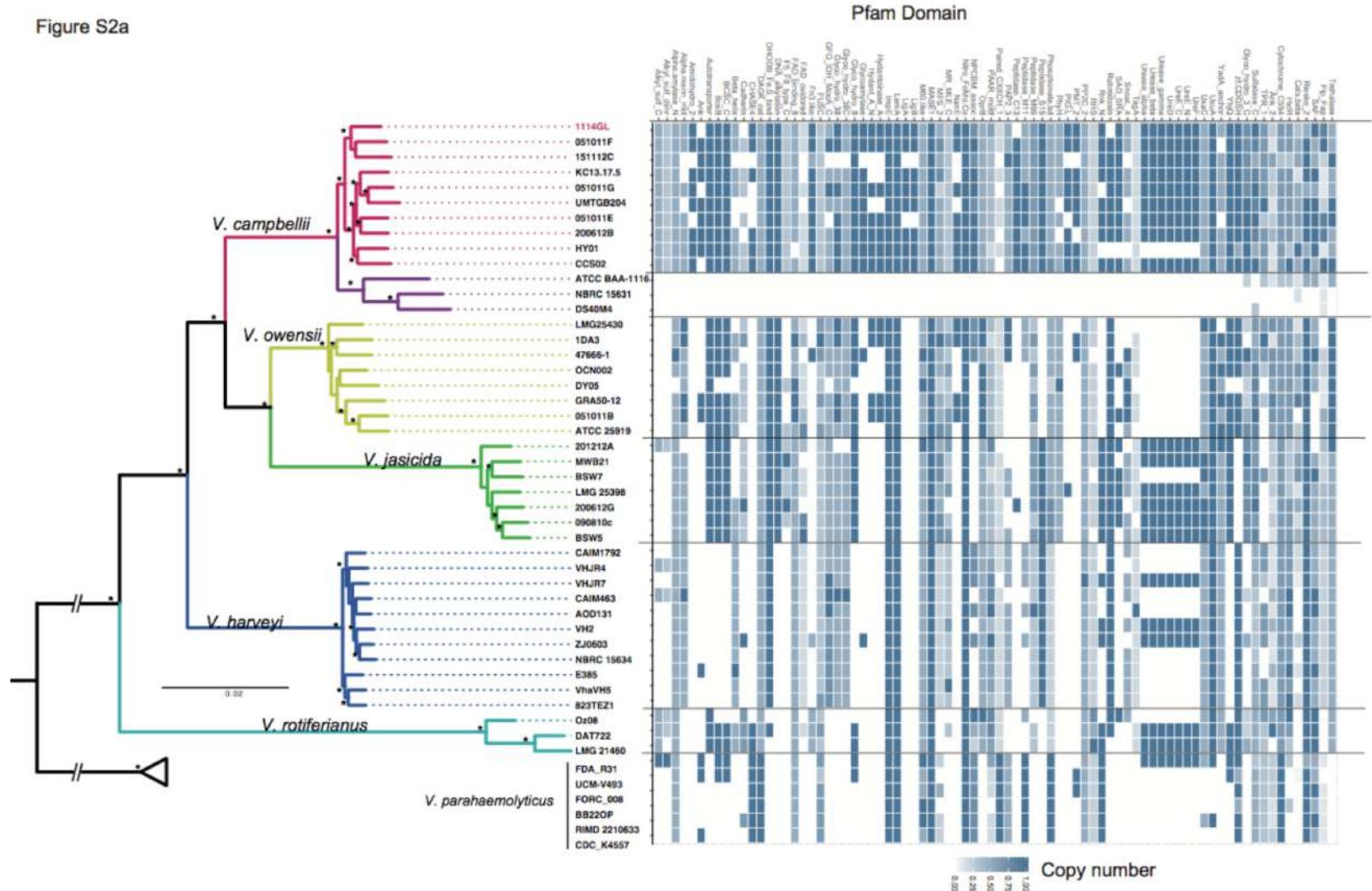
Two sequences are homologous if they share the same a common ancestor



Extension of homology to genomes / species

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

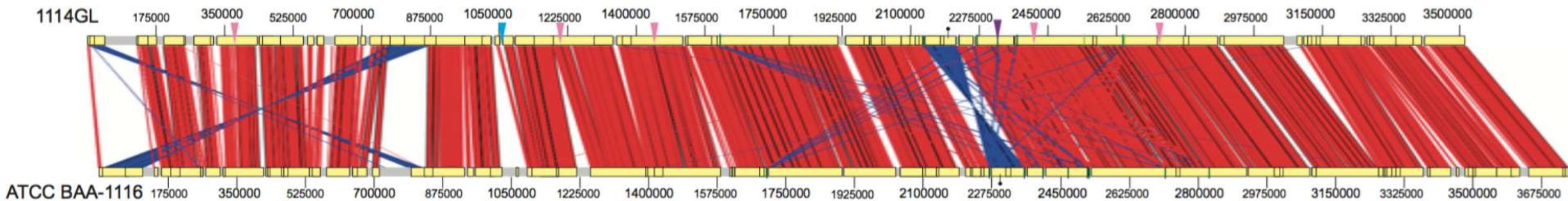
Figure S2a



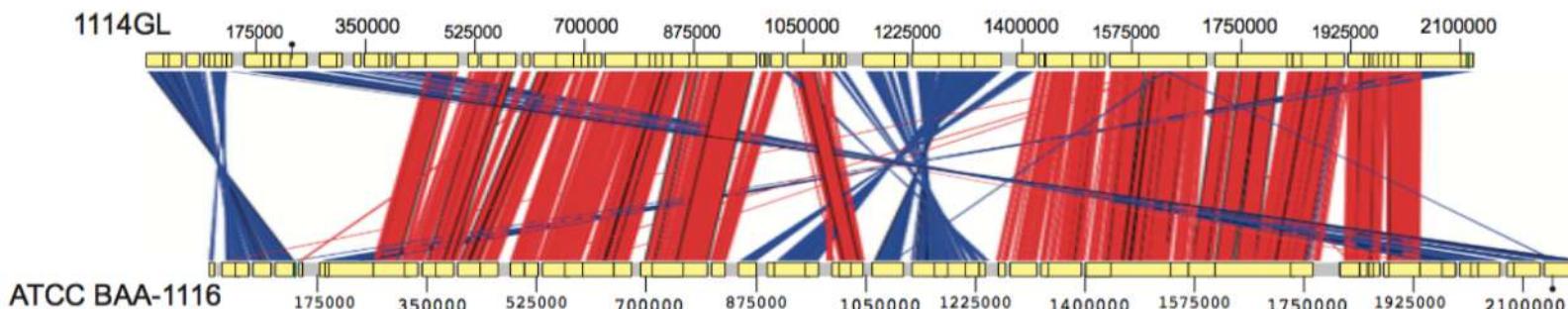
Extension of homology to genomes / species

Similarity of individual features (ordering and rearrangement)

(a) Chromosome I



(b) Chromosome II



- ▼ Gap
- ▼ Inter-scaffold gap
- ▲ The insertion including two genes with Big_2 domains
- Ori
- rRNA operon
- Partial rRNA operon

HOMOLOGY, GENES, AND EVOLUTIONARY INNOVATION



GÜNTER P. WAGNER

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)....

Why comparative genomics? – A summary

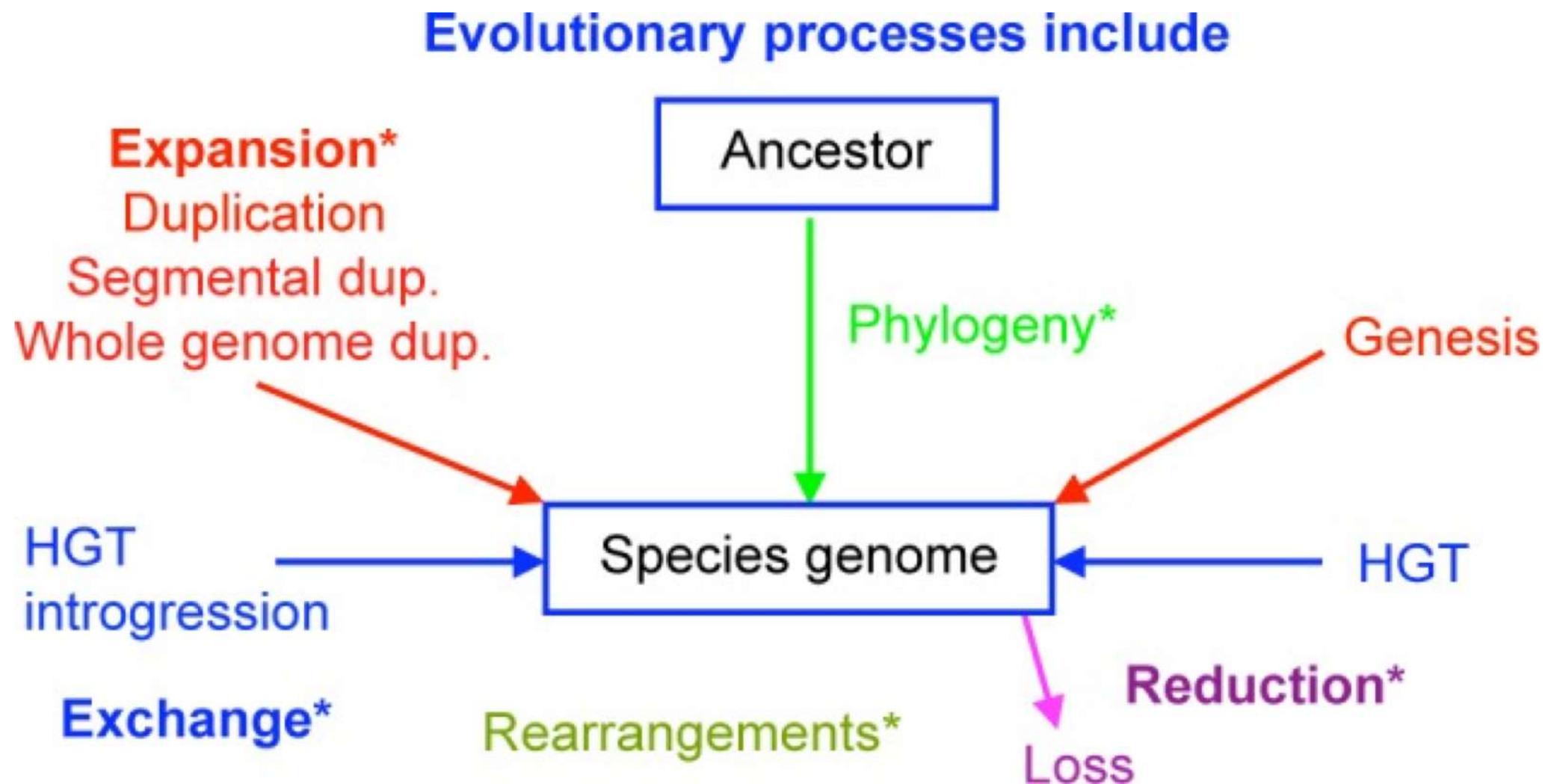
Compare multiple genomes now a norm

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

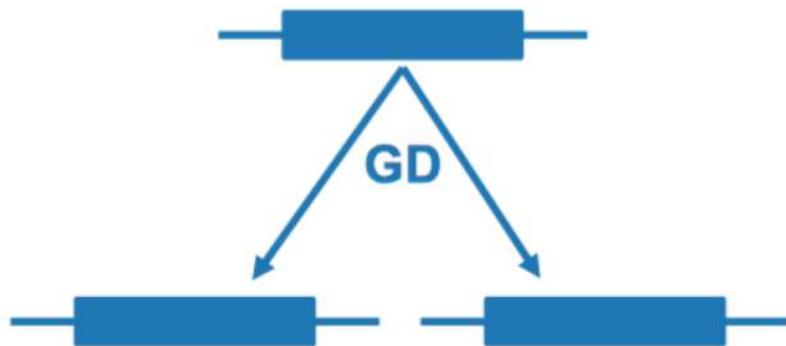


Sources of gene innovation

(Intuitive as genome gain genes of new functions)

Gene duplication (GD)

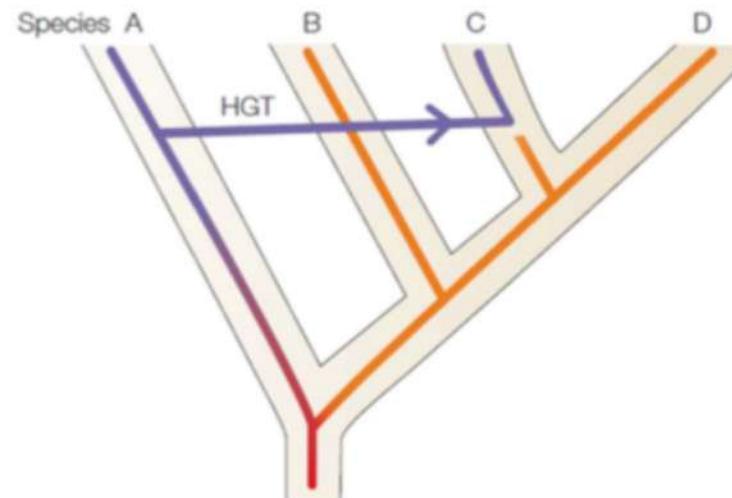
Any duplication of a region of DNA that contains a gene



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

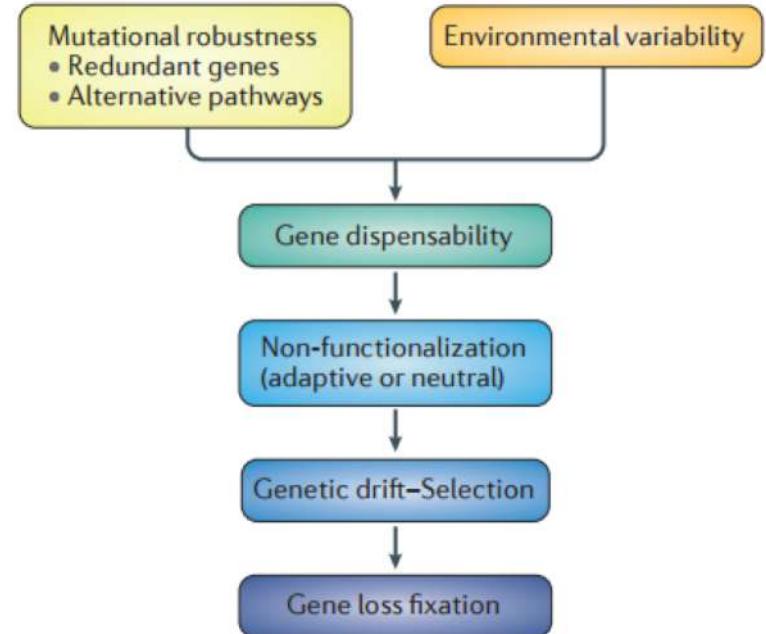
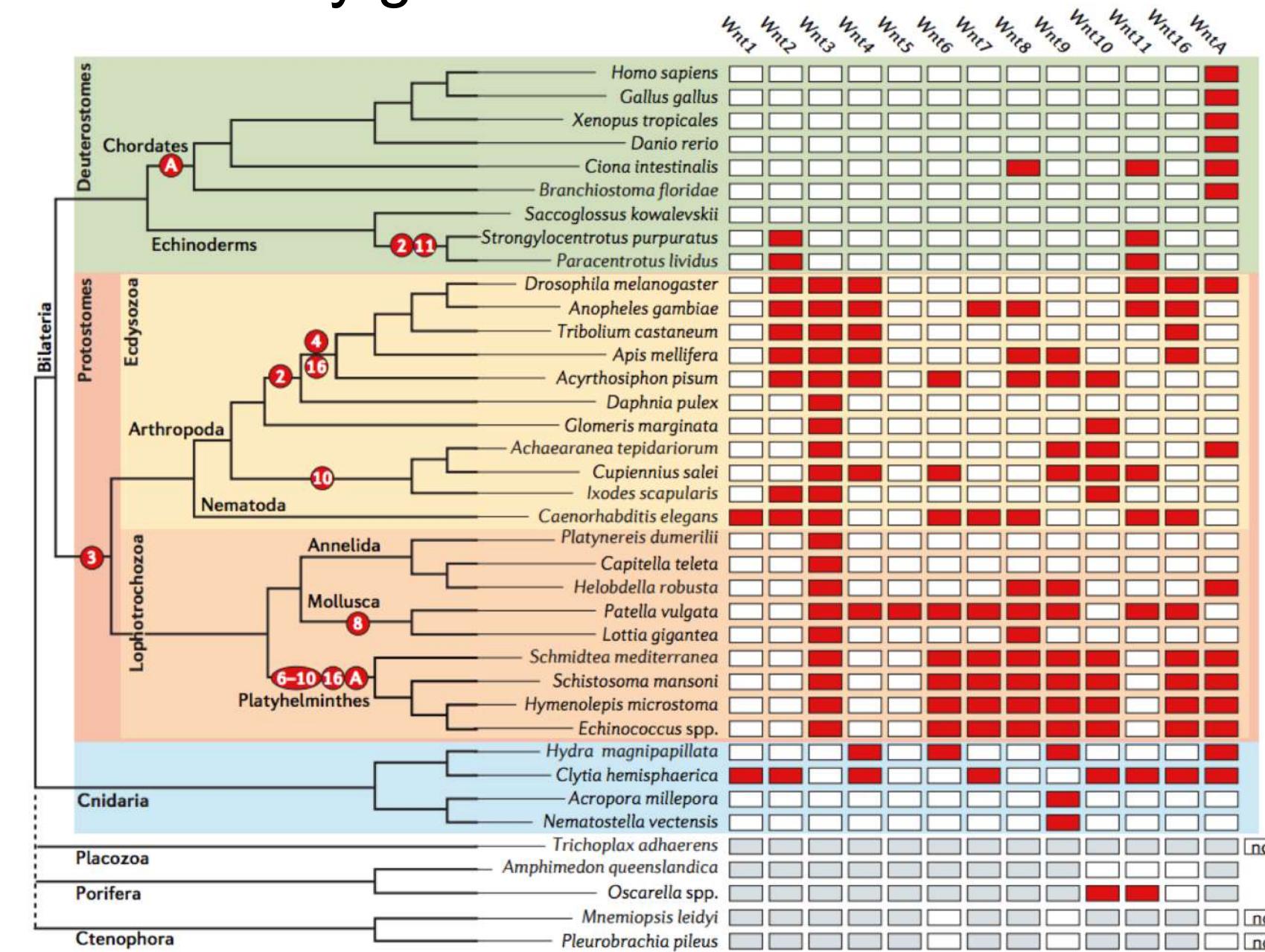
Horizontal gene transfer (HGT)

Exchange of genes between organisms other than through reproduction

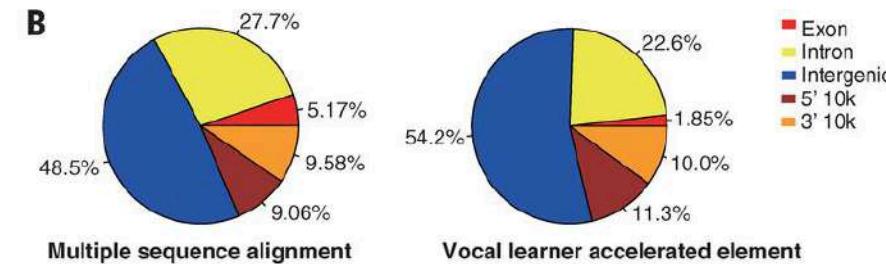
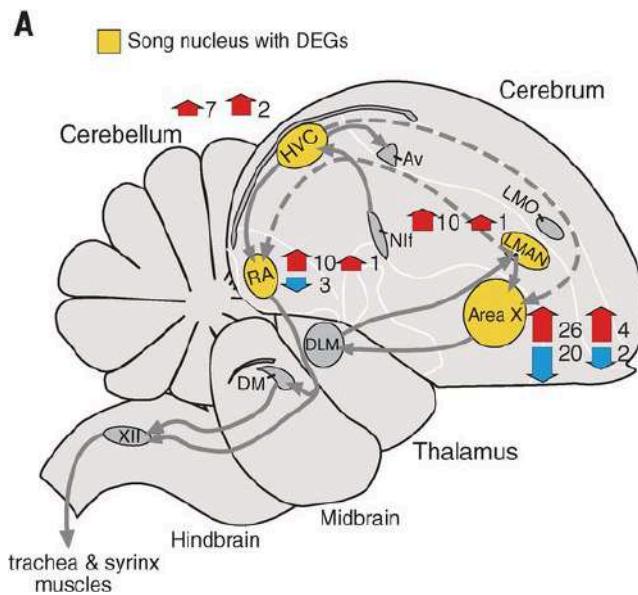
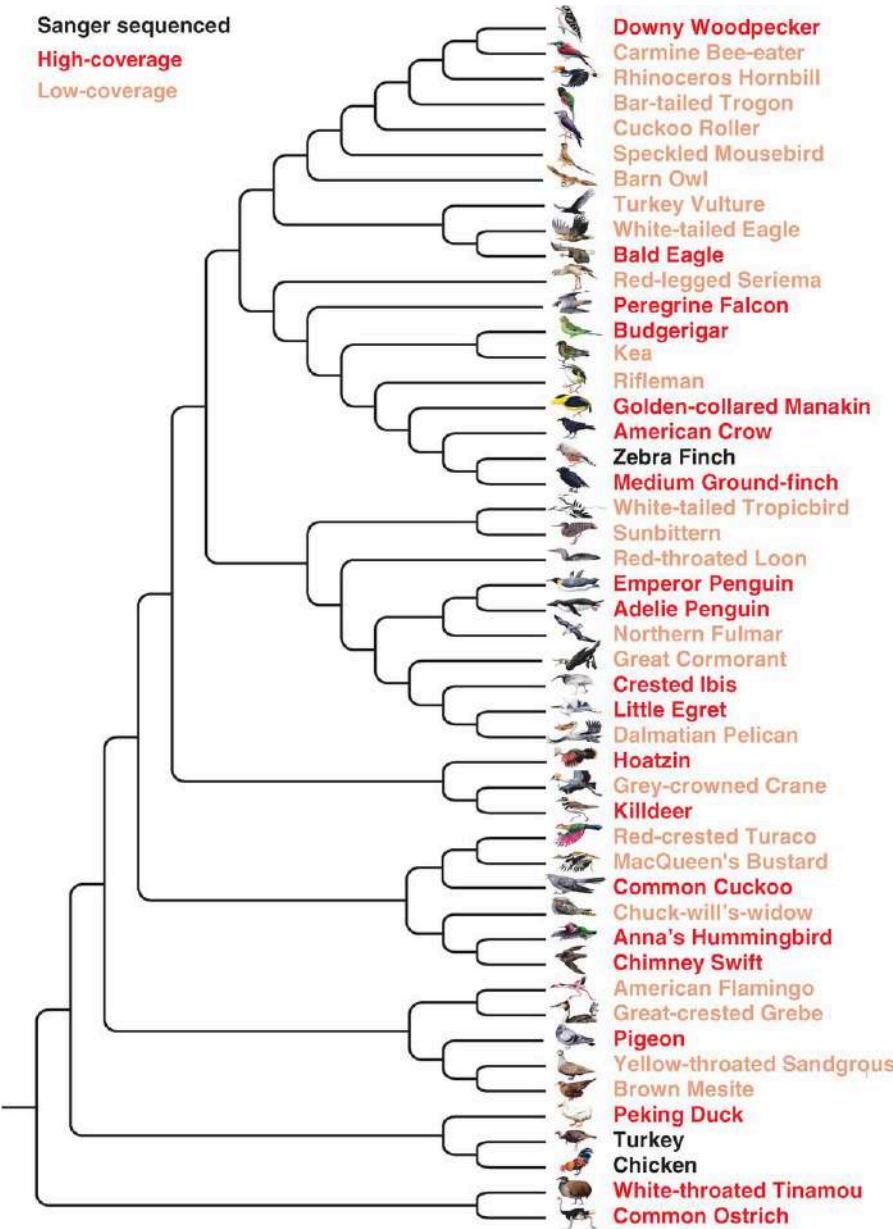


- ❖ Xenobiotic catabolism
- ❖ Toxin production
- ❖ Degradation of plant cell walls
- ❖ Wine fermentation

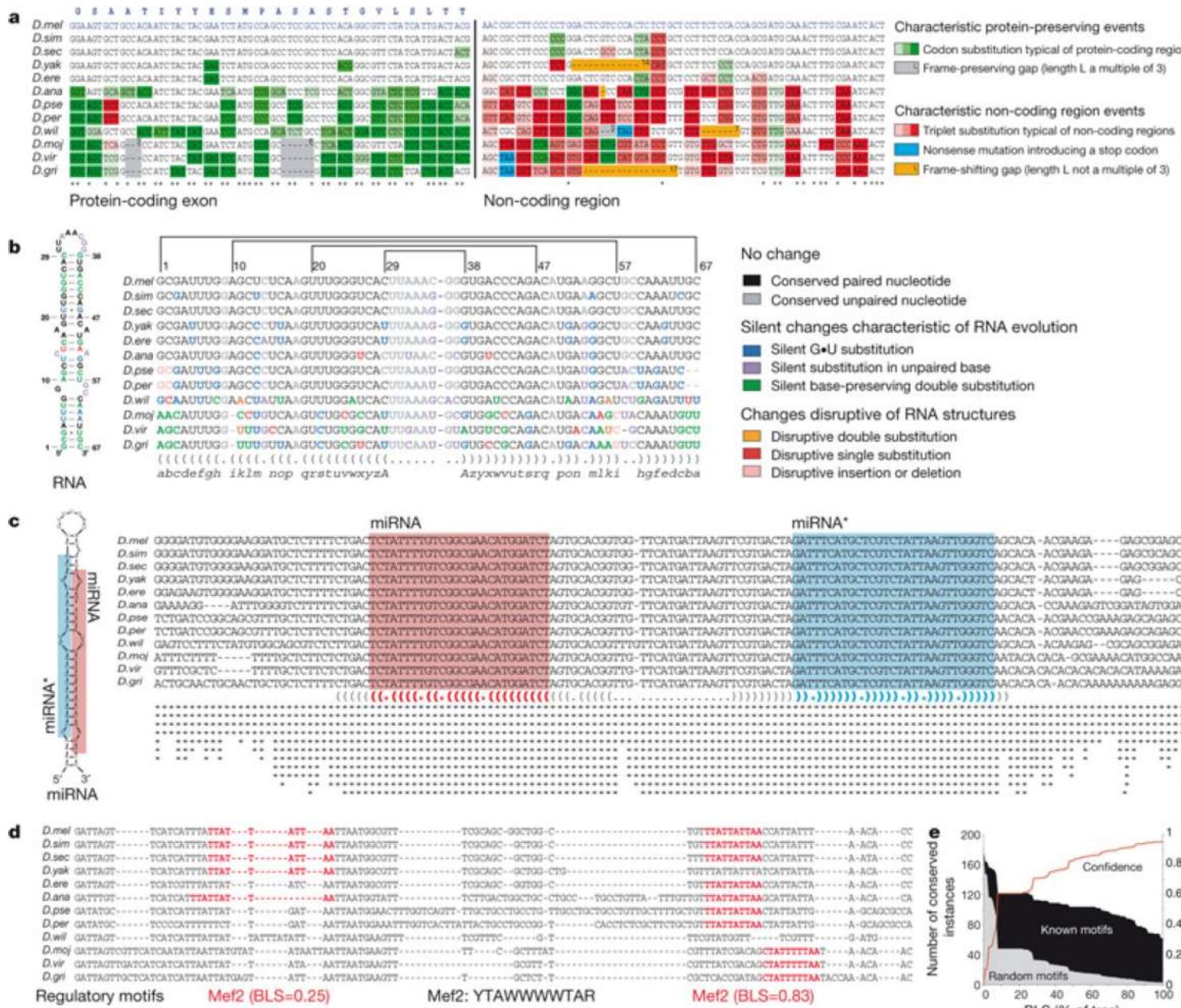
Evolution by gene loss



Reveal the evolutionary relationships among species



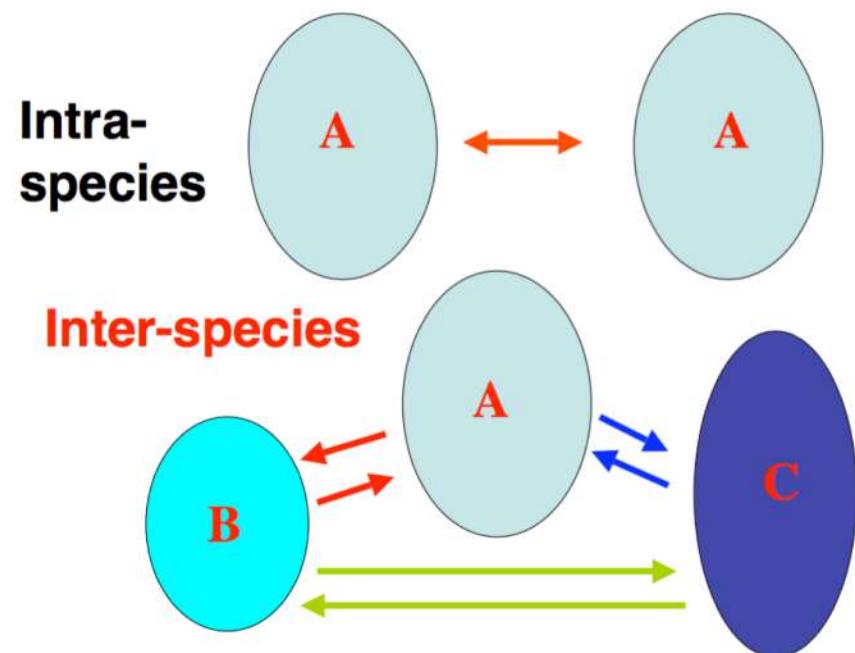
Link evolutionary processes with function



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

From homology to orthology

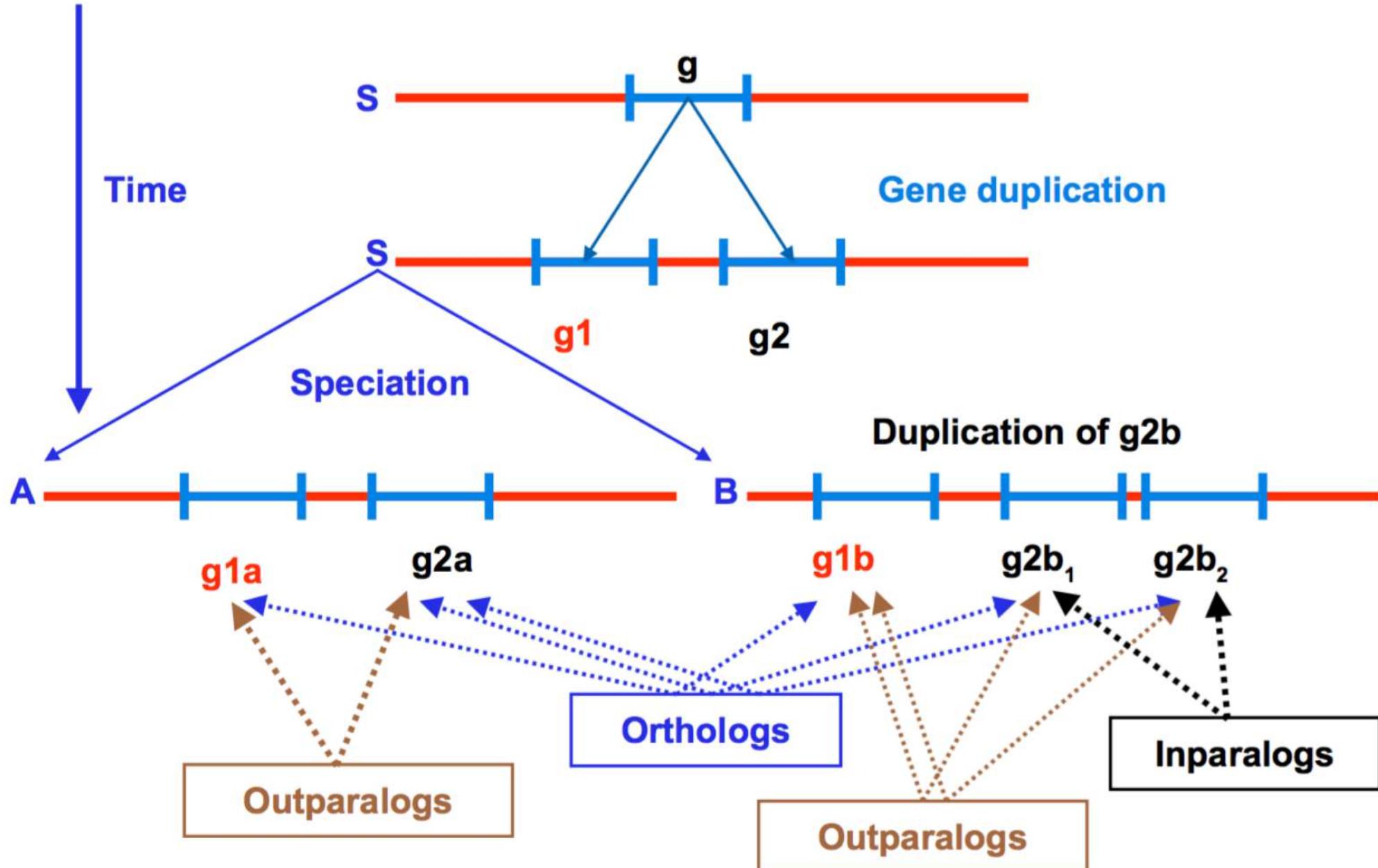
Homologues are sequences derived from a common ancestor...

- 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)."*



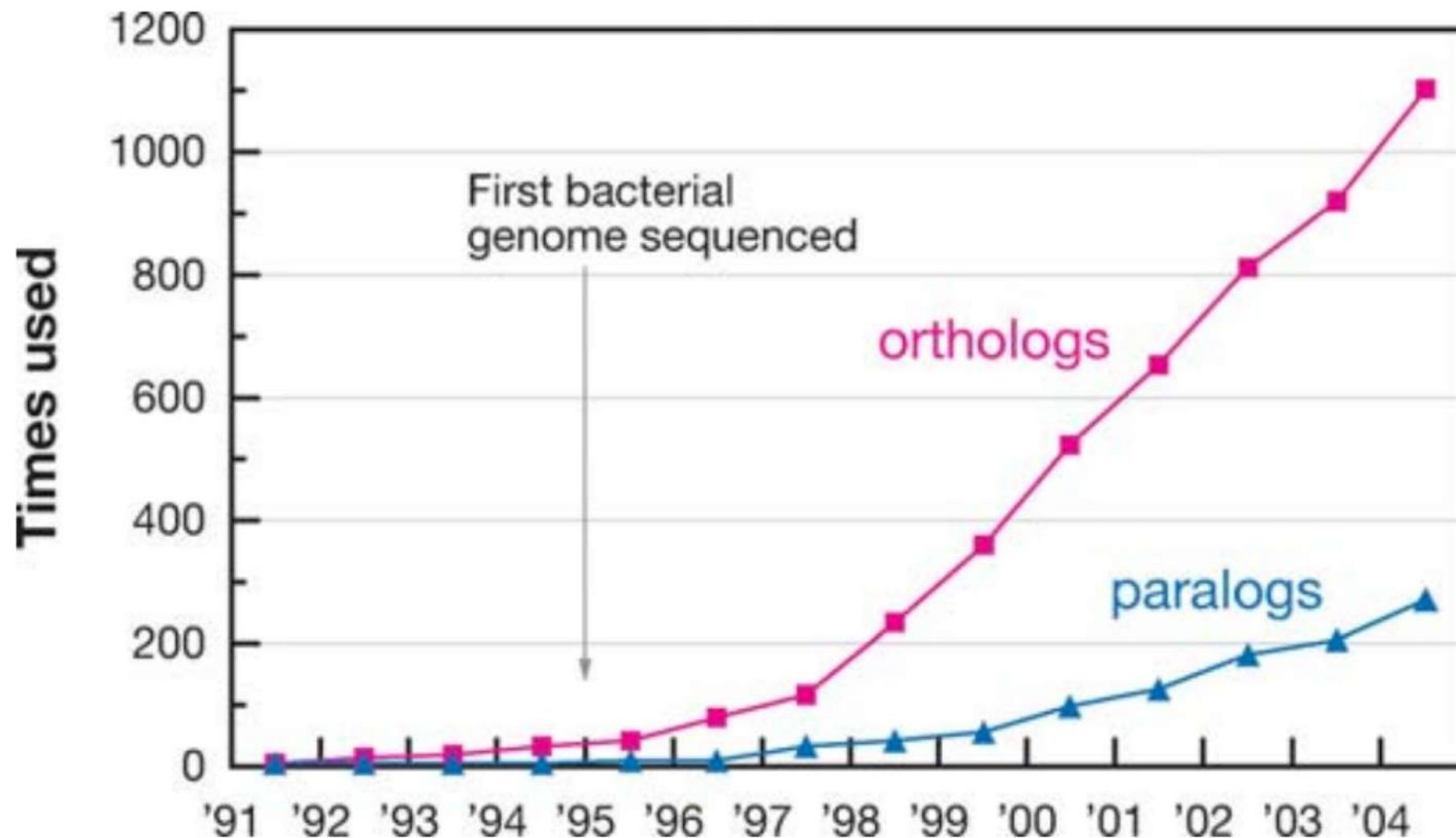
Why is orthology important?

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.

Usage of “ortholog” and “paralog”



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

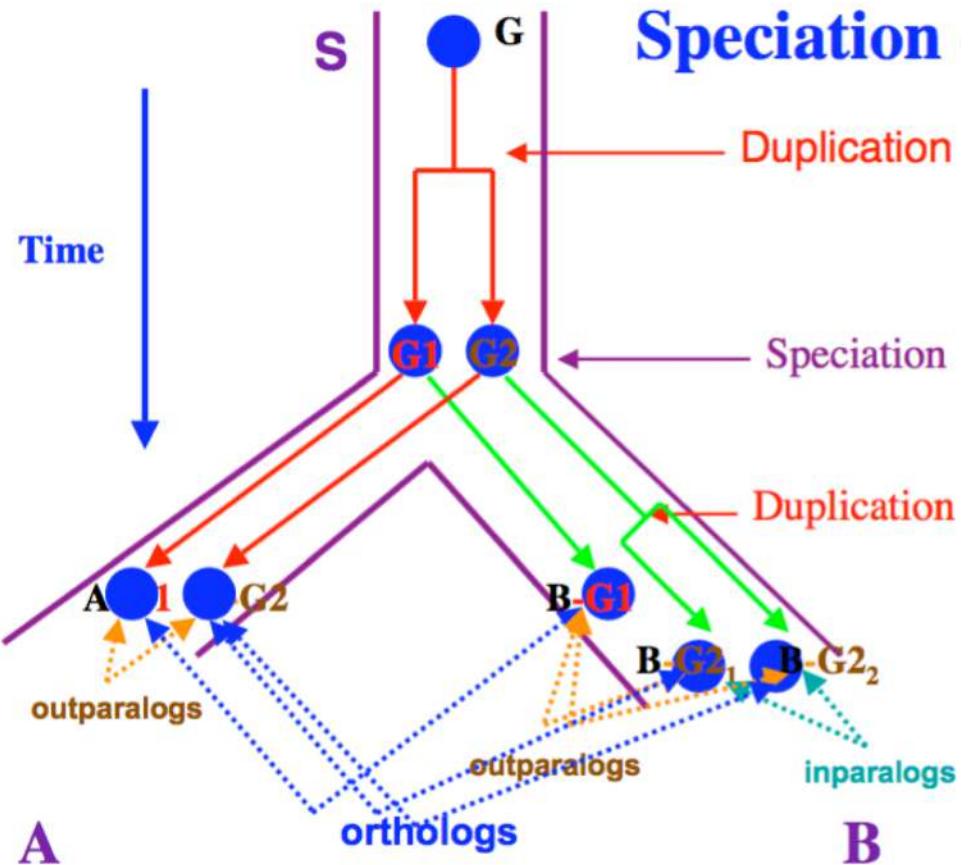


Table 1 Homology: terms and definitions

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)

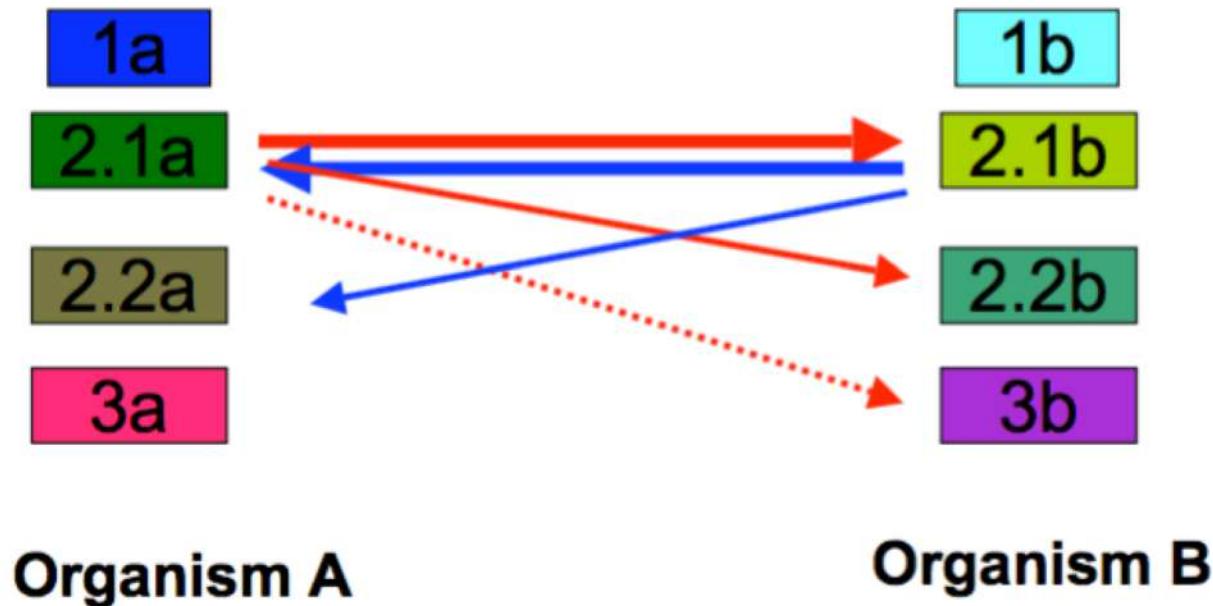
The most exact way of **comparing two (or more) genomes** in terms of their gene content. Necessary to uncover how genomes evolve.

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)**

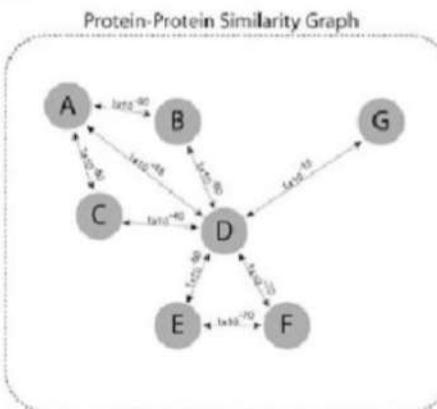


2.1a **2.1b** **Best Reciprocal Hits**

Sequence by clustering

mcl: The Markov Cluster Algorithm <http://micans.org/mcl/> (Stijn Van Dongen)

A



Generate weighted transition matrix using BLAST E-Values as weights (-logE)

B

Weighted Transition Matrix

	A	B	C	D	E	F	G
A	100	50	50	45	0	0	0
B	50	100	0	60	0	0	0
C	50	0	100	40	0	0	0
D	45	60	40	100	90	70	15
E	0	0	0	80	100	70	0
F	0	0	0	70	70	100	0
G	0	0	0	15	0	0	100

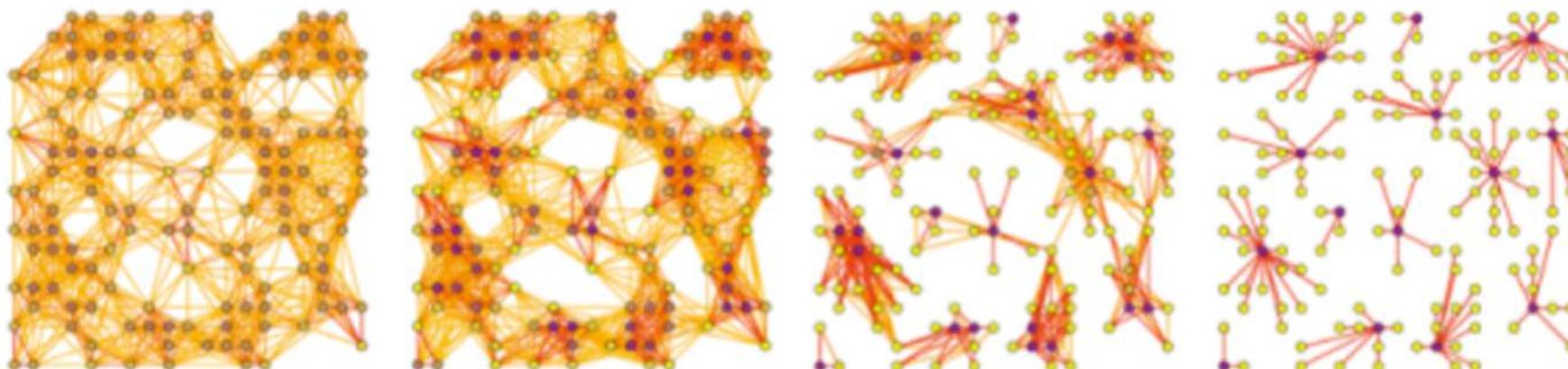
Transform weights into column-wise transition probabilities

Markov Matrix

	A	B	C	D	E	F	G
A	0.42	0.24	0.20	0.11	0.00	0.00	0.00
B	0.20	0.48	0.24	0.15	0.00	0.00	0.00
C	0.20	0.00	0.40	0.10	0.00	0.00	0.00
D	0.18	0.28	0.16	0.24	0.32	0.29	0.13
E	0.00	0.00	0.00	0.19	0.40	0.29	0.00
F	0.00	0.00	0.00	0.17	0.28	0.42	0.00
G	0.00	0.00	0.00	0.04	0.00	0.00	0.87

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)

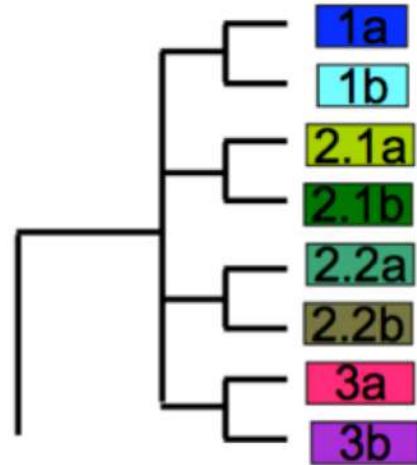
Produce clusters (gene families) using different inflation parameter



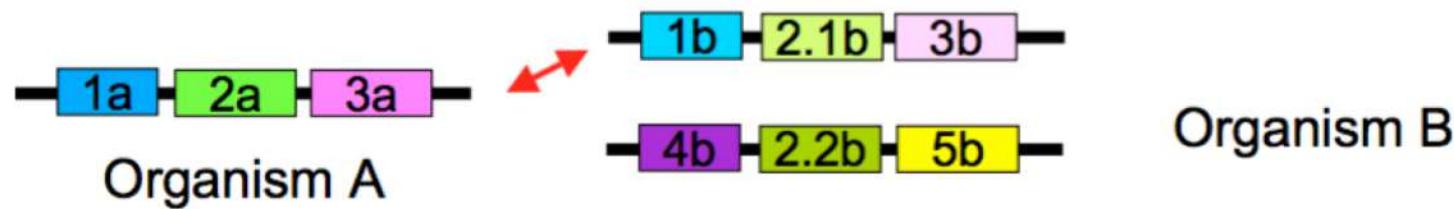
Weighted transition matrix and associated column stochastic Markov matrix for the seven proteins shown in (A).

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!

Tree reconciliation

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

Fig. 1a: Gene Tree

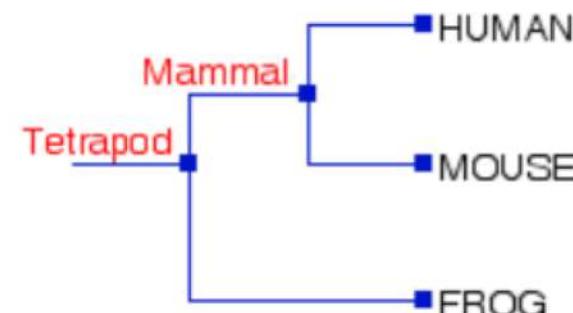
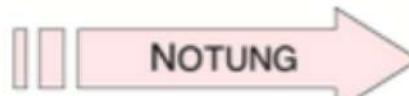
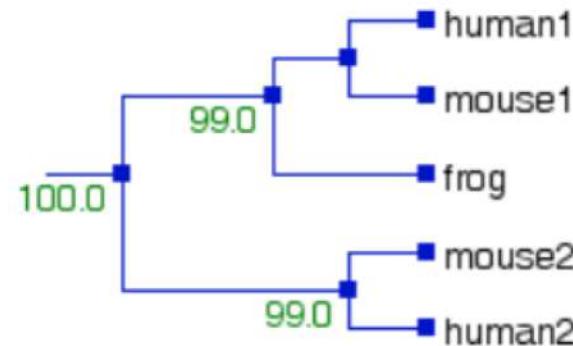


Fig. 1b: Species Tree

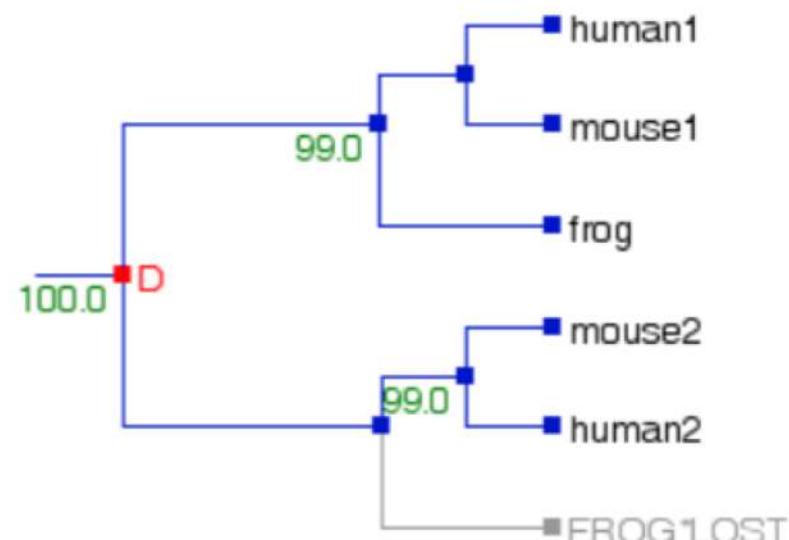
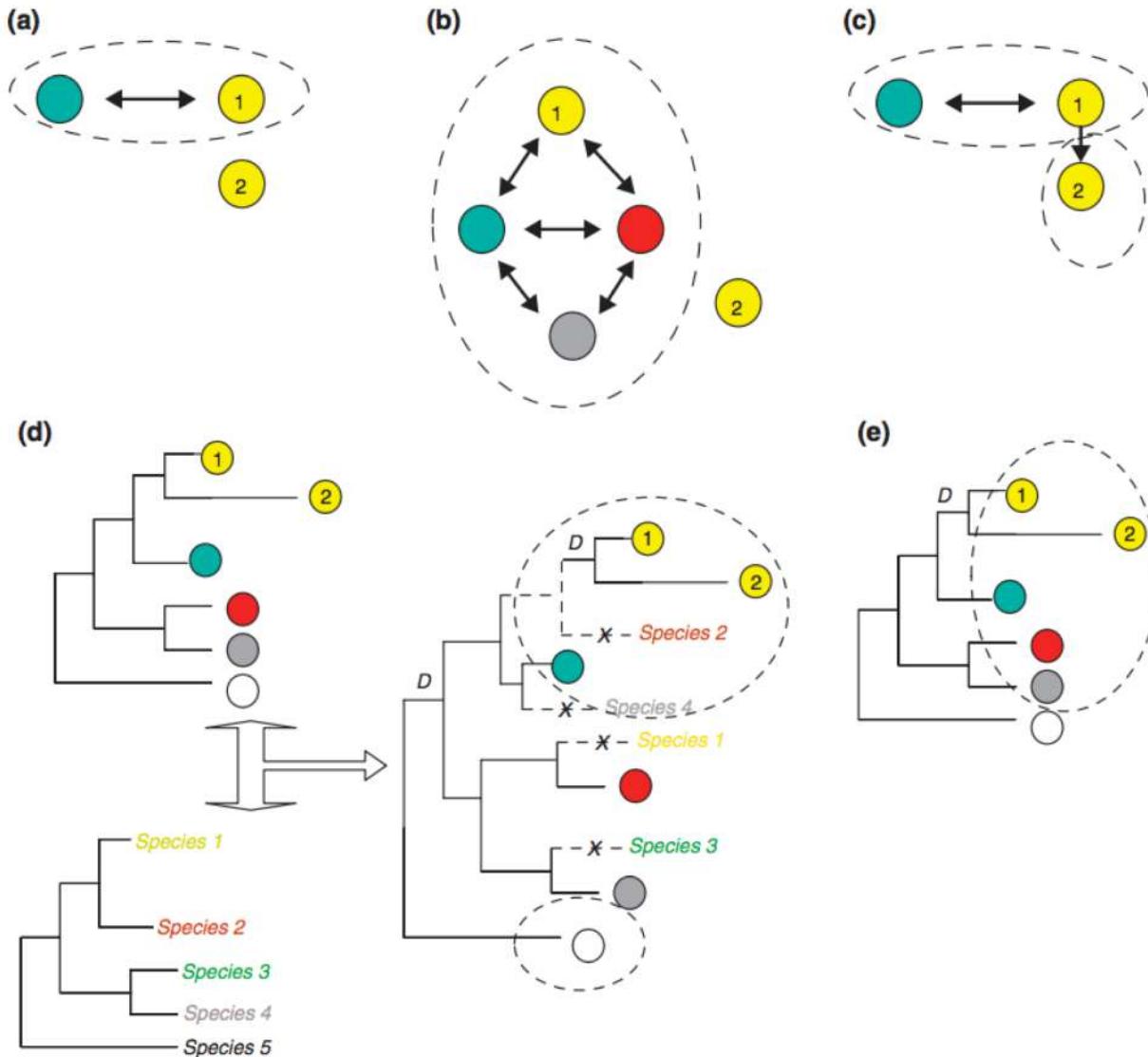


Fig. 1c: Reconciled Gene Tree

Orthology prediction methods



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

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

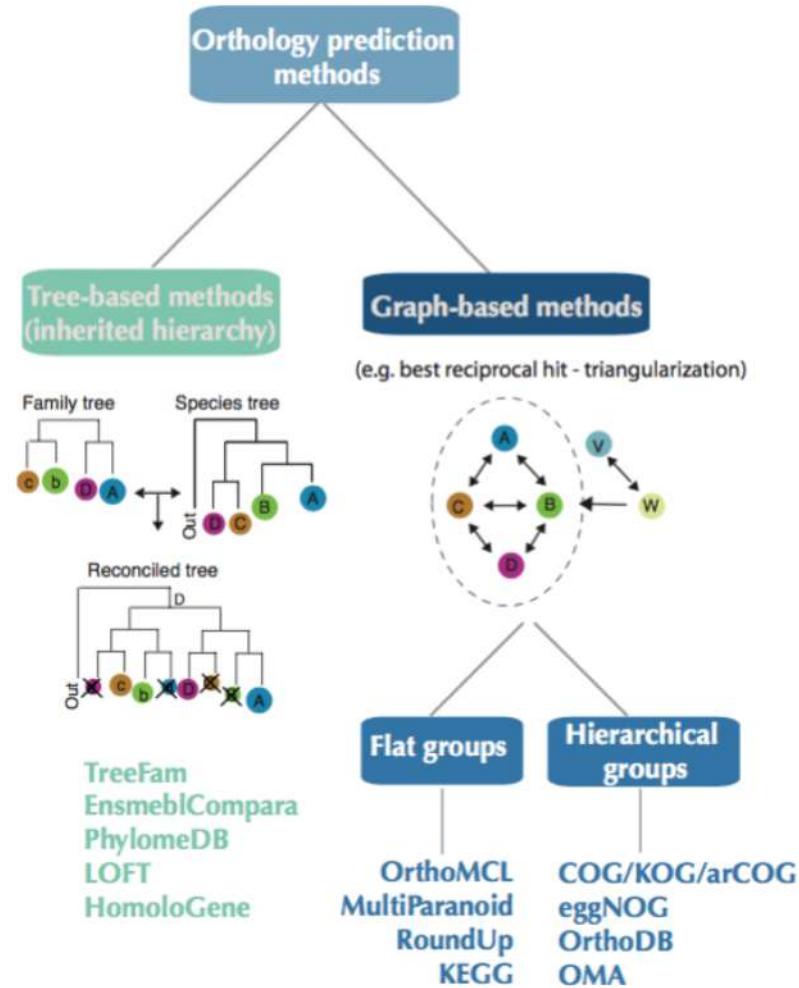
Combination of (1) and (2)

Some uses synteny

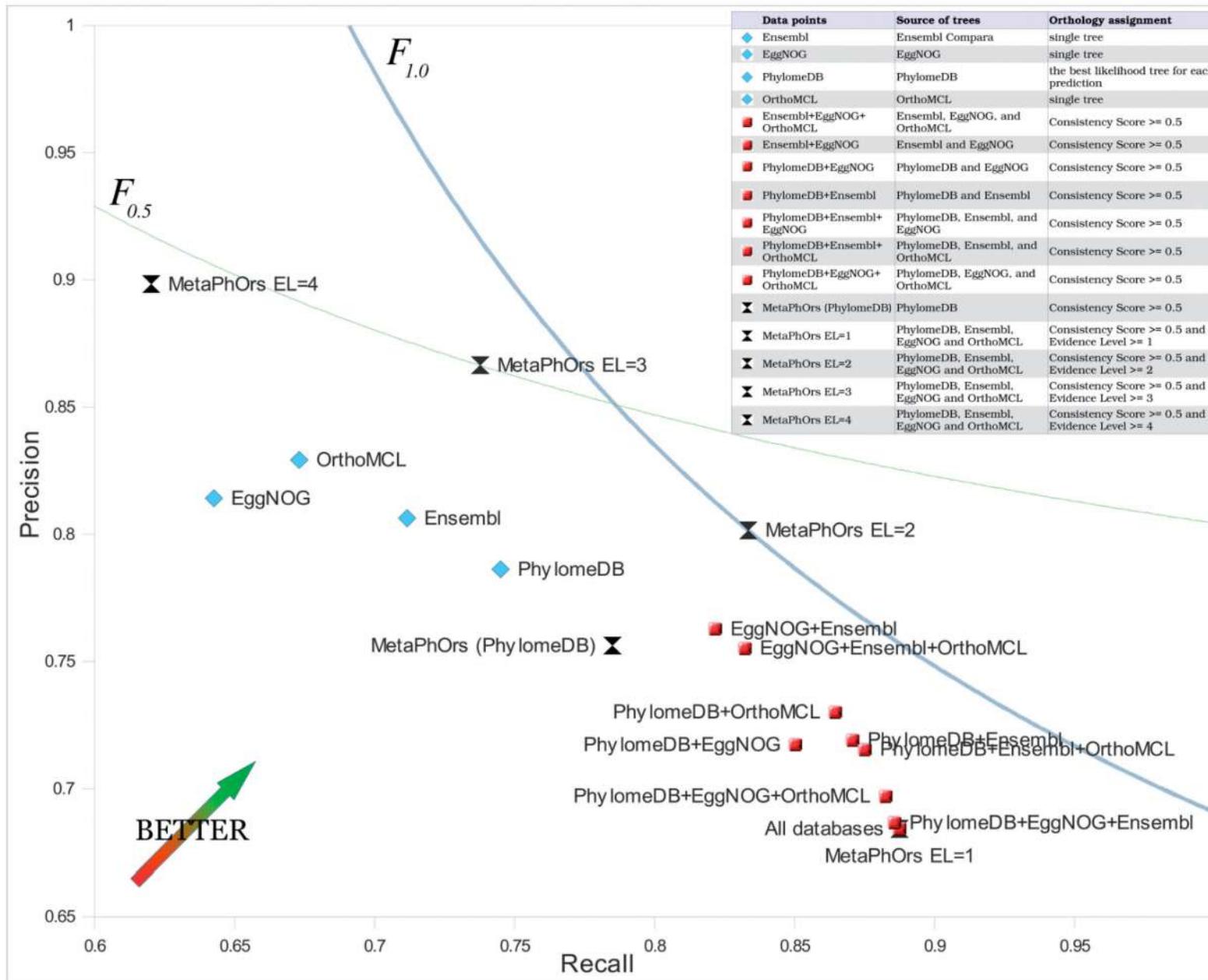
Tools

METHOD	ALGORITHM
COG ⁵⁴	Similarity—Single linkage clustering + Constraints
InParanoid/MultiParanoid ⁵⁵	Similarity (pair-wise species)/Extends to multiple species
OrthoMCL ⁵⁶	Similarity—MCL clustering algorithm
TribeMCL ⁵⁷	Similarity—MCL clustering algorithm
eggNOG ⁵⁸	Similarity—Detects false RBH due to gene fusion and protein domain shuffling
OrthoFocus ⁵⁹	Similarity—extended RBH to handle many-to-one and many-to-many relationships
OrthoInspector ⁶⁰	Similarity
SPO ⁶¹	Similarity (RBH)—Partition of orthologs includes Intra-species Partition and MCL clustering
OrthoFinder ⁶²	Similarity—Clustering
Roundup ⁶³	Reciprocal Smallest Distance
RSD ⁶⁴	Reciprocal Smallest Distance (evolutionary distance = estimated number of amino acid substitutions)
OMA ⁶⁵	Similarity—Global sequence alignment
ME ⁶⁶	Minimum Evolution Method
MSoAR ⁶⁷	Similarity—Genome rearrangement—duplication
Orthotrappster ⁶⁹	Phylogeny—bootstrap
RIO ⁷⁰	Similarity (HMMER)—bootstrap—Phylogeny
PhIGs ⁷¹	Similarity—Multiple sequence alignments—Phylogenetic trees
PhyOP ⁷²	Similarity (overlapping limits)—phylogeny based on d_s (synonymous substitution rates)
TreeFam ⁷³	Infer orthologs—paralog from the phylogenetic tree
LOFT ⁷⁴	Assigns hierarchical orthology numbers to genes based on a phylogenetic tree
EnsemblCompara GeneTrees ⁷⁵	Clustering—multiple alignment—tree generation based on TreeBeST method
SYNERGY ⁷⁶	Sequence similarity—species phylogeny—reconstruction of underlying gene evolutionary histories
PHOG ⁷⁷	Precomputed phylogenetic trees followed by identification of orthologs as sequences from different species that are each others reciprocal nearest neighbors
COCO-CL ⁷⁸	Similarity—Correlation between sequences—single linkage clustering

Note: This table shows some orthology inference methods with corresponding reference and a short description of their underlying algorithm.



Every tool kind of disagrees...



Caveats

Evolution of multi-domain proteins

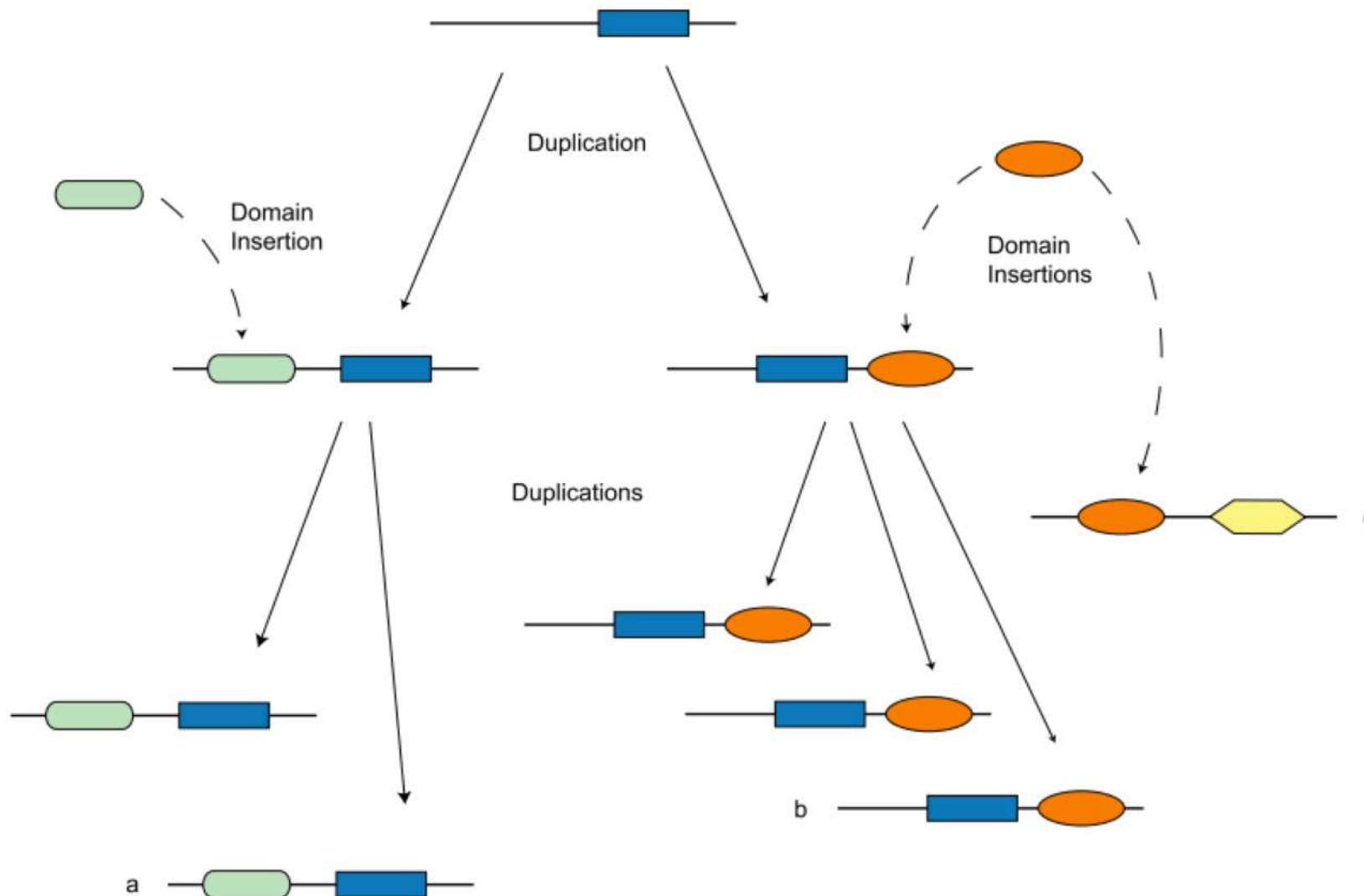
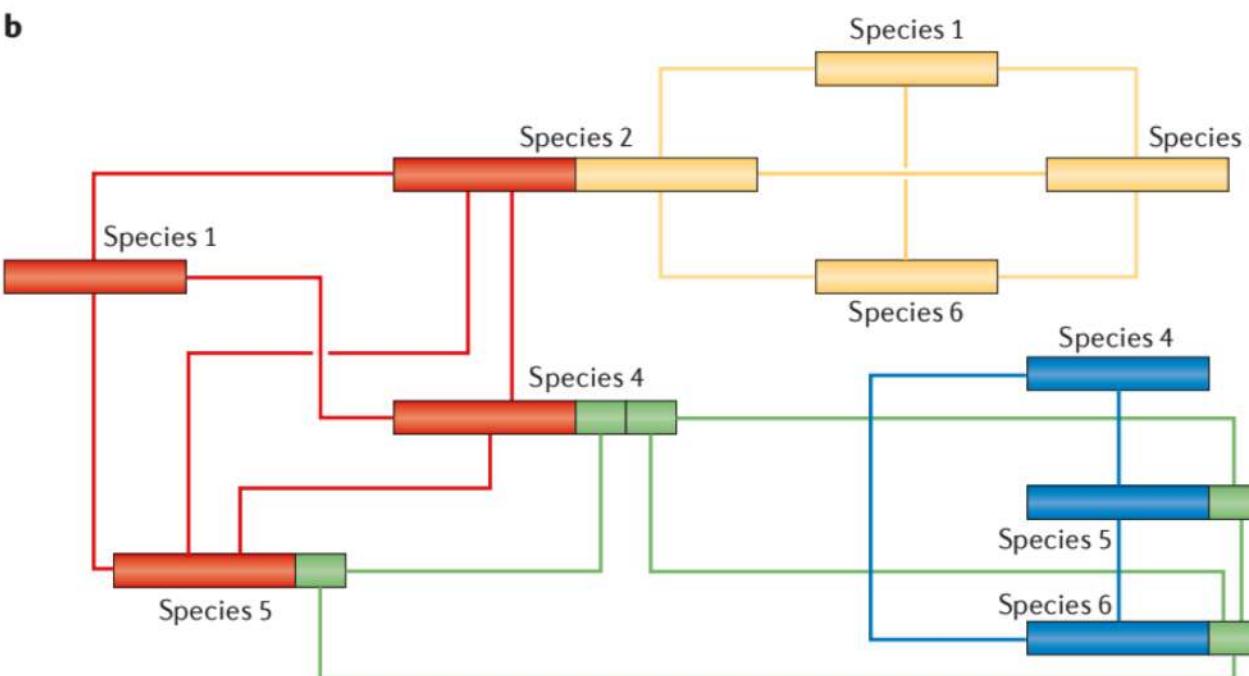
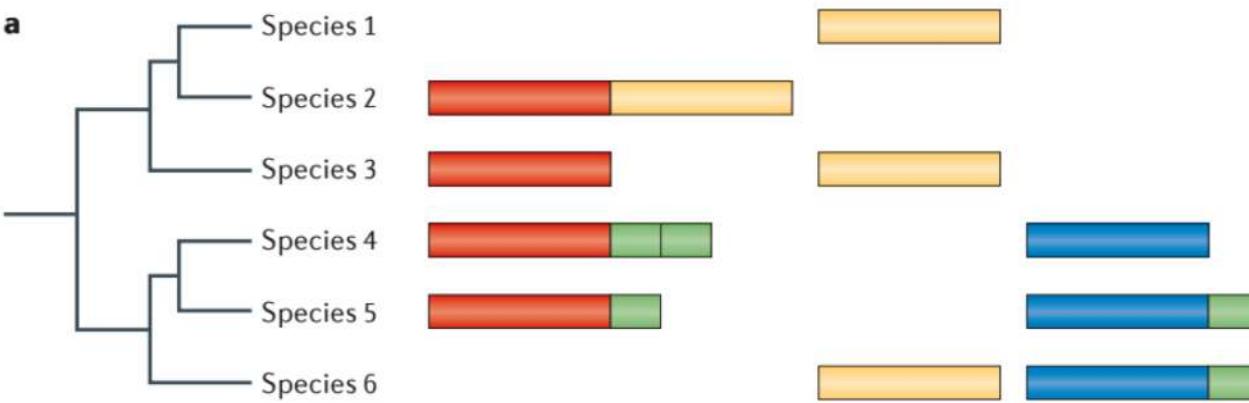


Figure 1. The evolution of a hypothetical multidomain family by gene duplication and domain insertion. Genes in the *a* and *b* subfamilies share a common ancestor but do not have identical domain composition. Gene *c* shares a homologous domain with genes in the *b* subfamily, but there is no gene that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

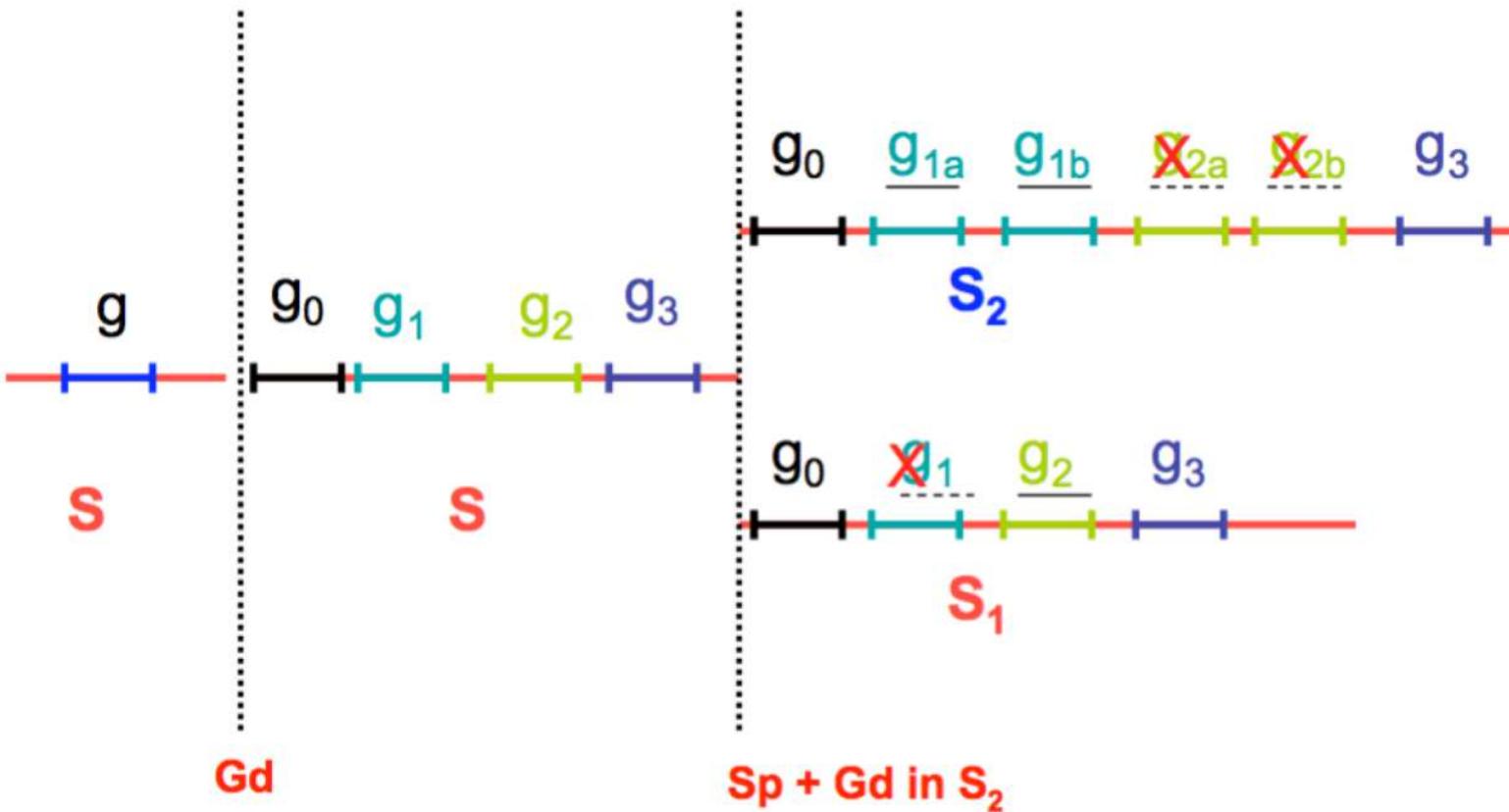
doi:10.1371/journal.pcbi.1000063.g001

Song et al (2008)

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



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.

Effect of HGT on orthology and paralogy (If orthology is simply inferred by gene content)

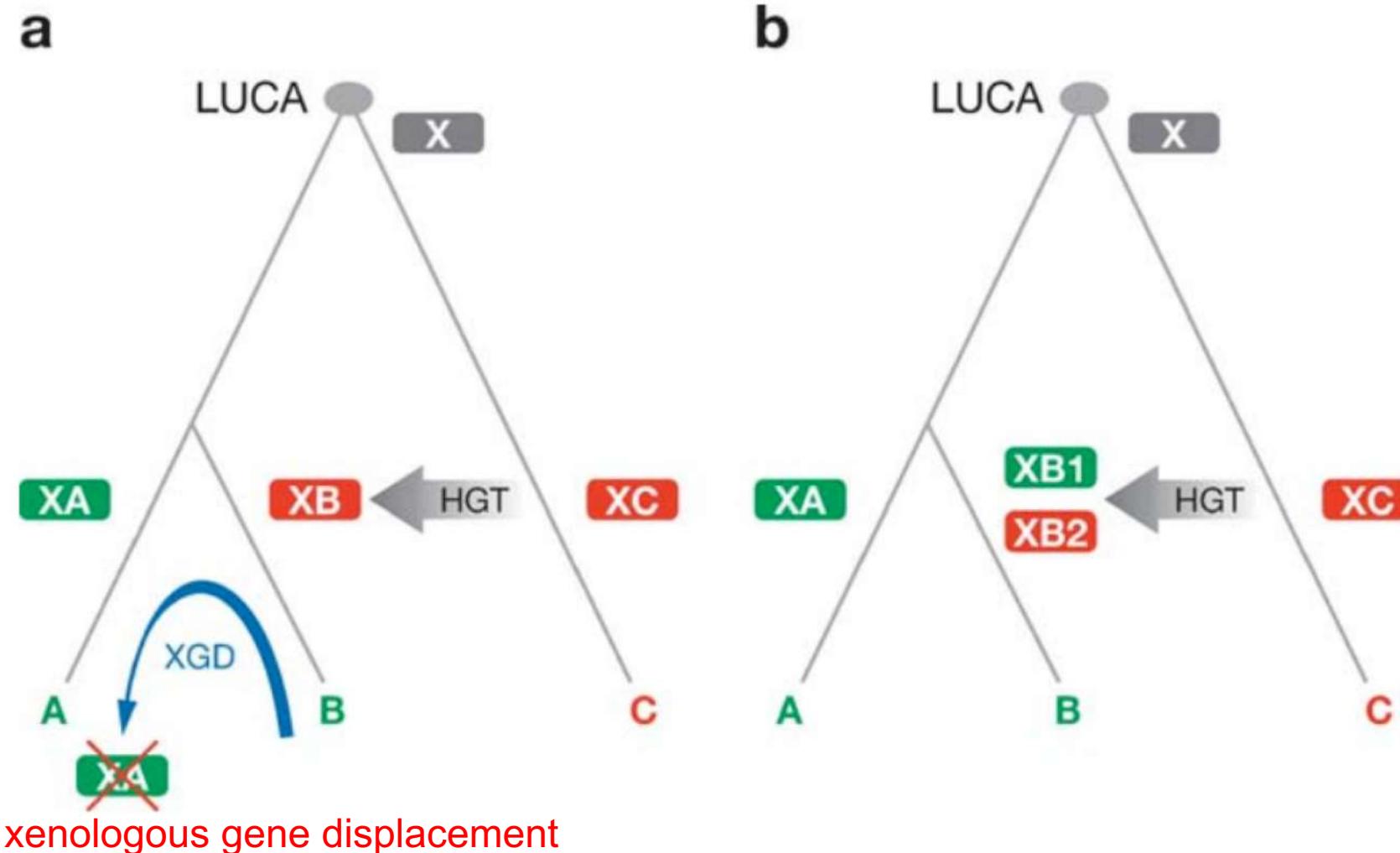


Figure 4

Effect of horizontal gene transfer on orthology and paralogy. (a) A hypothetical evolutionary scenario with HGT leading to xenology. (b) A hypothetical evolutionary scenario with HGT leading to pseudoparalogy. LUCA, Last Universal Common Ancestor (of all extant life forms).

Caveat: Do orthologs, as compared to paralogs, are more likely to share the same function?

How confident can we be that orthologs are similar, but paralogs differ?

Romain A. Studer and Marc Robinson-Rechavi

Department of Ecology and Evolution, Biophore, Lausanne University, CH-1015 Lausanne, Switzerland and Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

OPEN  ACCESS Freely available online

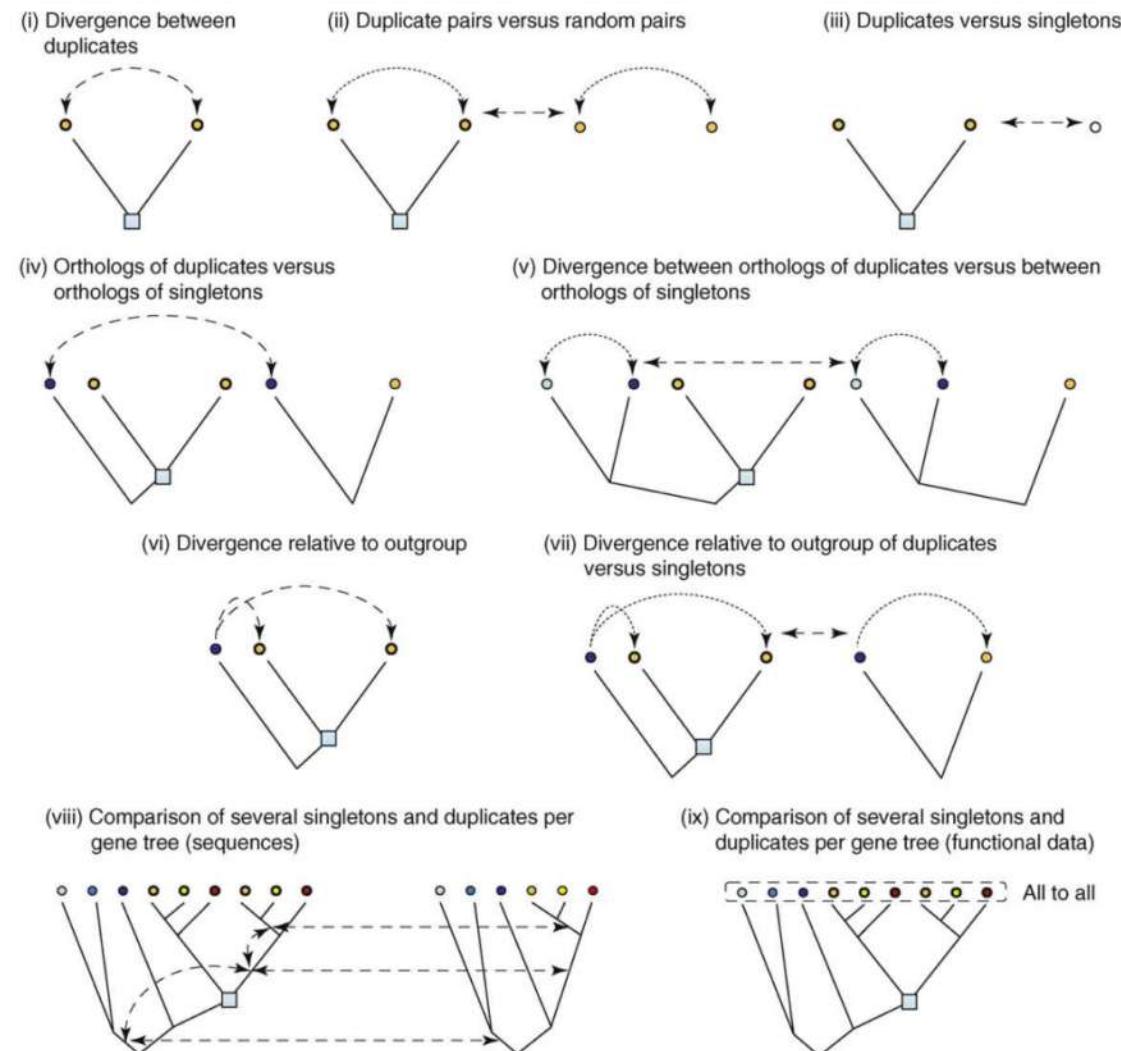
PLOS COMPUTATIONAL BIOLOGY

Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs

Adrian M. Altenhoff^{1,2}, Romain A. Studer^{2,3,4}, Marc Robinson-Rechavi^{2,3}, Christophe Dessimoz^{1,2,5*}

1 ETH Zurich, Department of Computer Science, Zürich, Switzerland, **2** Swiss Institute of Bioinformatics, Lausanne, Switzerland, **3** Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland, **4** Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London, United Kingdom, **5** EMBL-European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom

Some designs for the study of gene duplication.



TRENDS in Genetics

Table 1. The impact of study design on tests of evolution after duplication

Study design ^a	Data type ^b	Predictions under simple evolutionary models		Function change after duplication or speciation	Refs
		Preferential change after duplication	Subfunctionalization ^c Neofunctionalization		
(i) Divergence between duplicates	Functional	Differences between paralogs			[19,20,55]
(ii) Duplicate pairs versus random pairs	Functional	Paralogs more similar than random pairs, but not identical			[11,19,54]
(iii) Duplicates versus singletons	Functional	Measure of retention bias, confused by evolution after duplication			[11,19,25]
(iv) Orthologs of duplicates versus orthologs of singletons	Functional	Measure of retention bias			[12]
(v) Divergence between orthologs of duplicates versus between orthologs of singletons	Sequence	Measure of retention bias			[12,53]
(vi) Divergence relative to outgroup	Sequence	No prediction relative to symmetry, relaxed purifying selection	Asymmetry between paralogs, positive selection ^e		[11,17,58]
	Functional	Two paralogs different, complementary to full outgroup function	One paralog similar to outgroup, one different		[18,21]
(vii) Divergence relative to outgroup of duplicates versus singletons	Sequence	Higher divergence of duplicates ^d , confused by retention bias			[62]
	Functional	Two paralogs different, complementary to outgroup; singleton similar to outgroup	One paralog similar to outgroup, one different; singleton similar to outgroup	No specific prediction ^f	[18,24,25]
(viii) Comparison of several singletons and duplicates per gene tree (sequences)	Sequence	Higher relaxation of purifying selection on branches after duplication	More positive selection on branches after duplication	Positive selection in various branches of the tree ^g	[13,43,48,56]
(ix) <i>idem</i>	Functional	Conservation of pattern among singletons; sub-patterns in duplicates	Conservation in most homologs; new patterns ^h in some duplicates	Variation in pattern among homologs, with gain of new patterns ^h	

Testing duplication combining transcriptome dataset

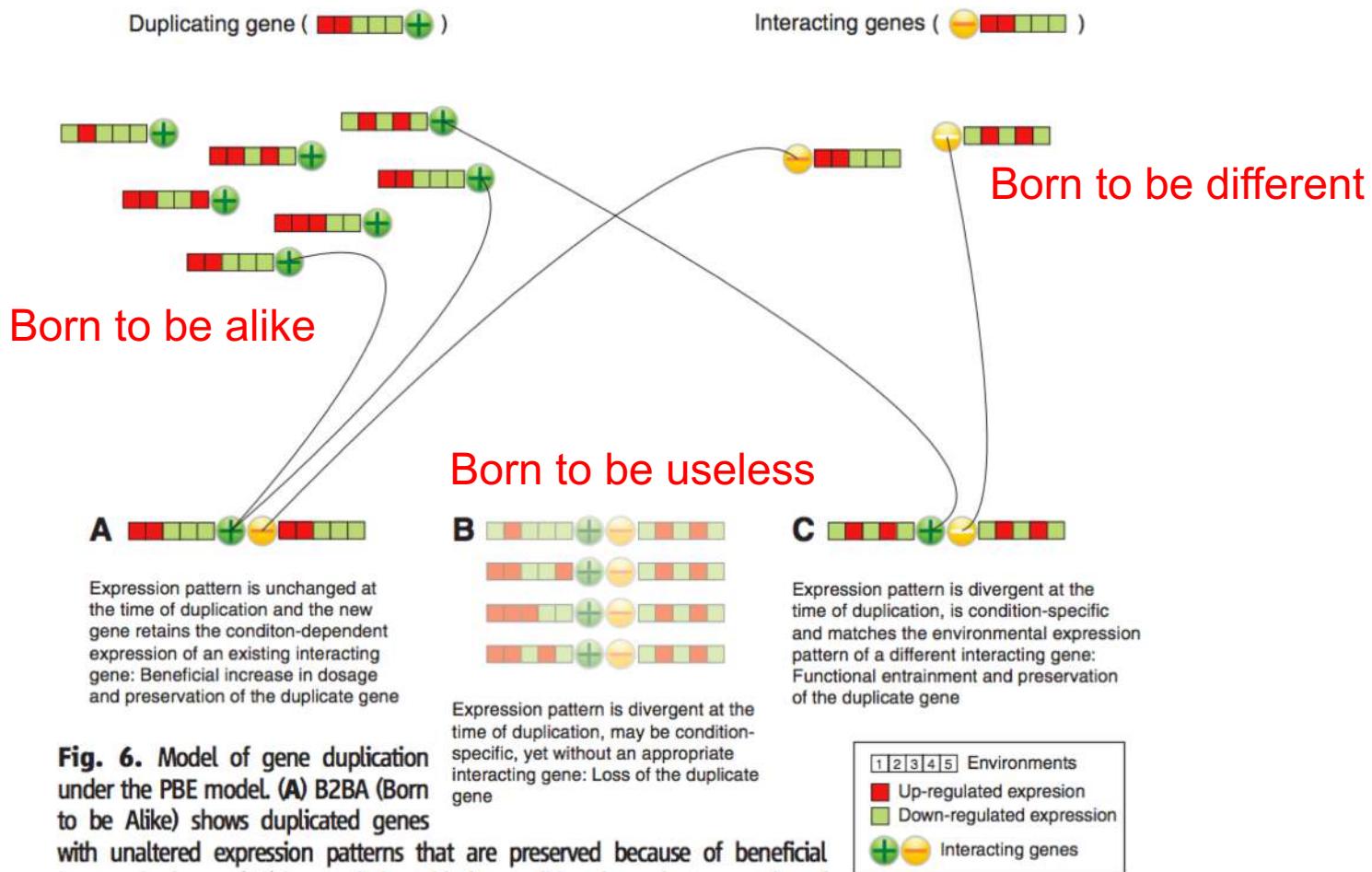


Fig. 6. Model of gene duplication under the PBE model. **(A)** B2BA (Born to be Alike) shows duplicated genes with unaltered expression patterns that are preserved because of beneficial increase in dosage (20) in association with the condition-dependent expression of an interacting gene. **(B)** B2BU (Born to be Useless) genes with initially divergent expression patterns and with inappropriate condition-dependent responses or interacting genes are most likely lost. **(C)** B2BD (Born to be Different). When the derived expression pattern of a paralog at the time of duplication is shared with a different interacting gene (white negative sign), and when the effect of their combined products is beneficial under a distinct environmental condition, the likelihood for preservation is increased. Color-coding represents condition-dependent expression patterns across multiple environments. Lines represent the process of functional entrainment.

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.

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

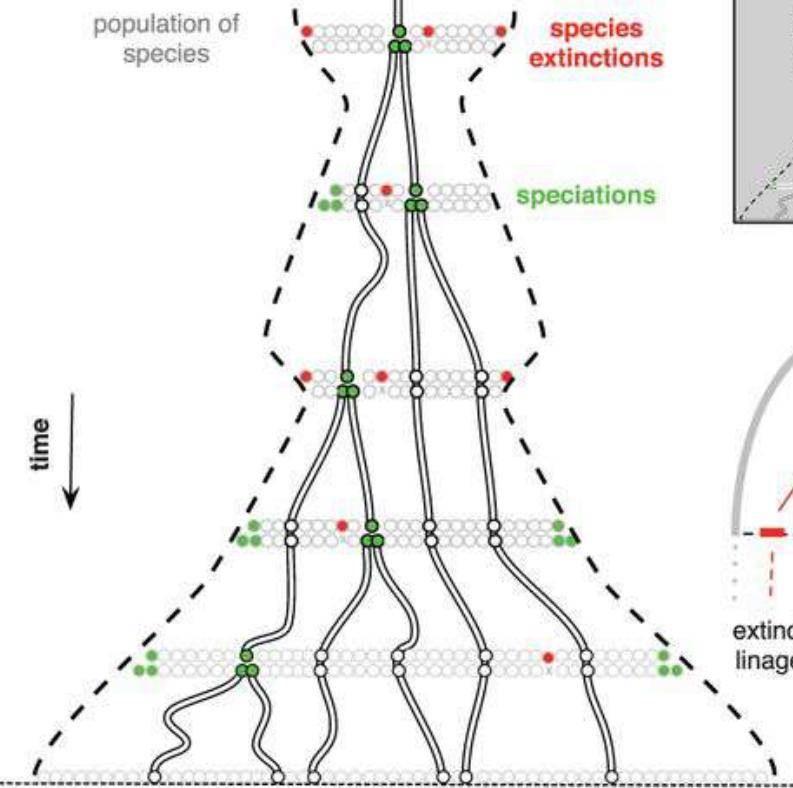
It would be difficult or impossible to understand the evolutionary history of an organism, even having available its whole genome sequence, in isolation. So it is always the case the phylogenomics is practiced for sets of genomes.

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).

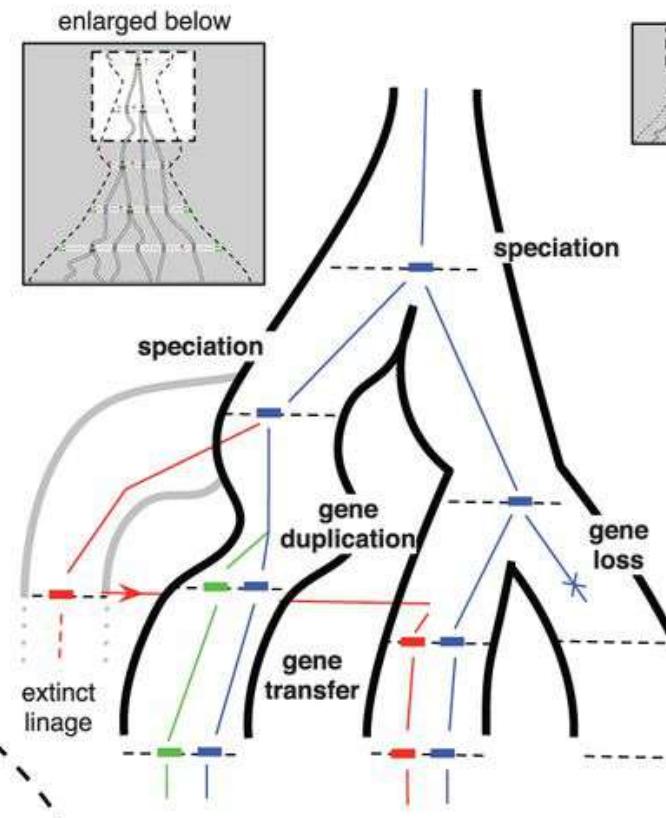
Each level of the hierarchy contributes to generating phylogenetic signal that can lead to differences between reconstructed gene trees.

a) species diversification



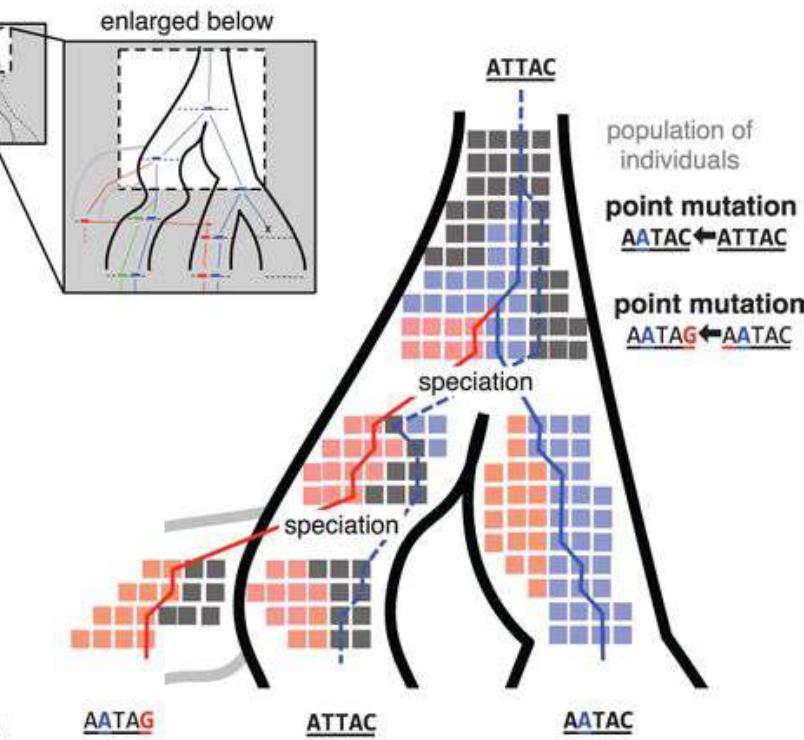
b)

gene birth and death



c)

sequence substitution



sampled genomes

(species tree)

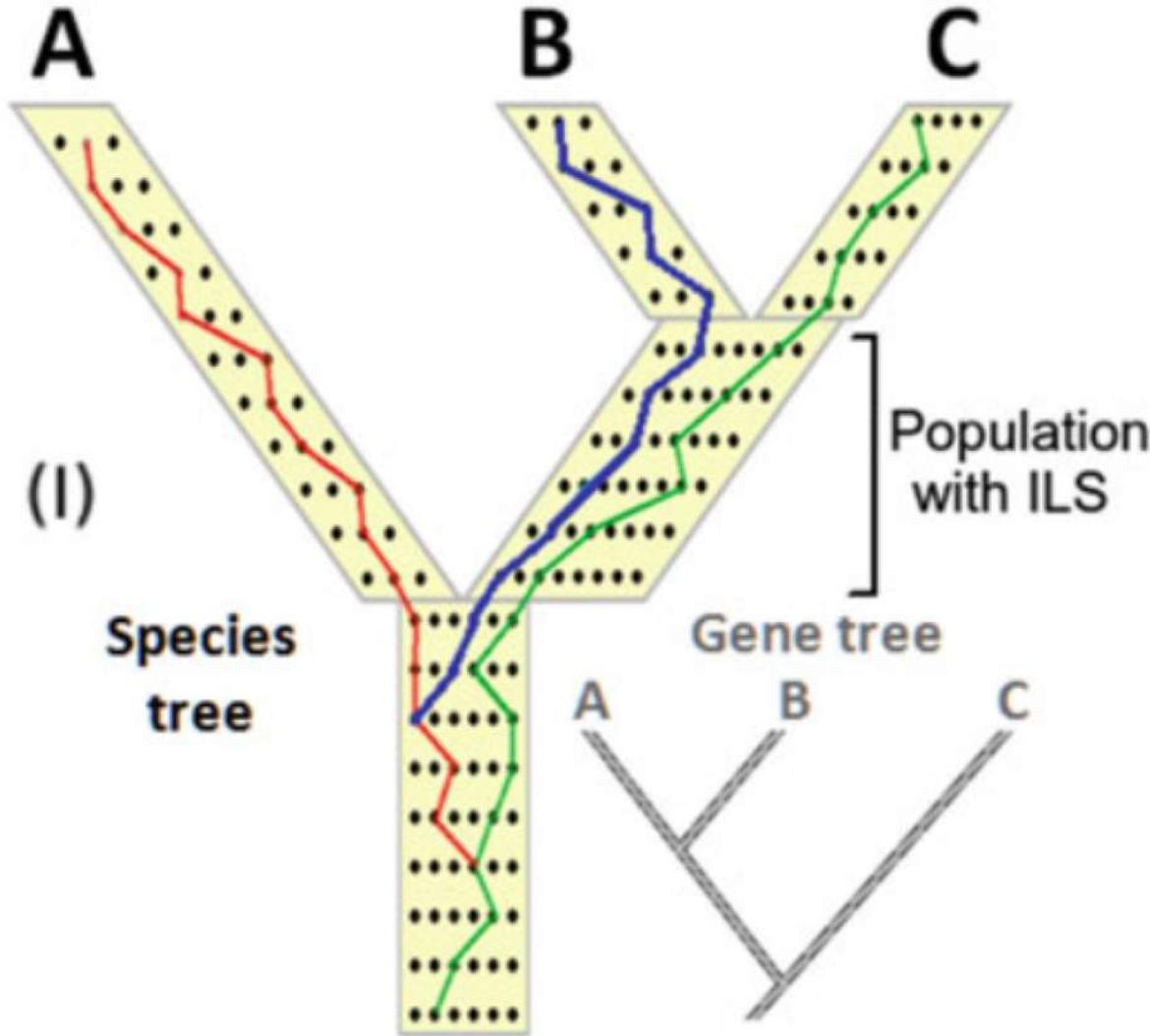
genes in genomes

(locus tree)

nucleotides in genes

(gene tree)

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

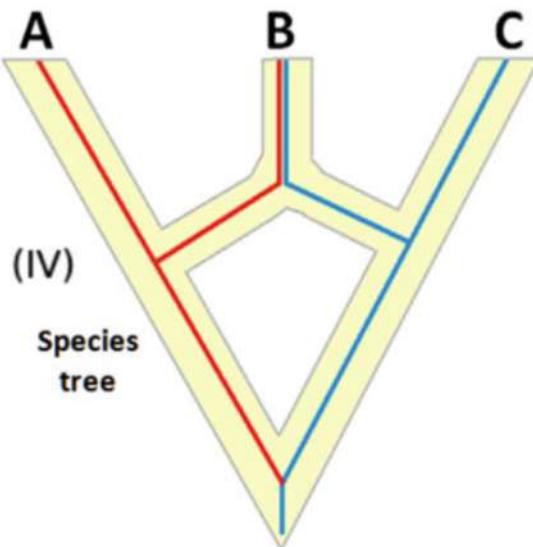
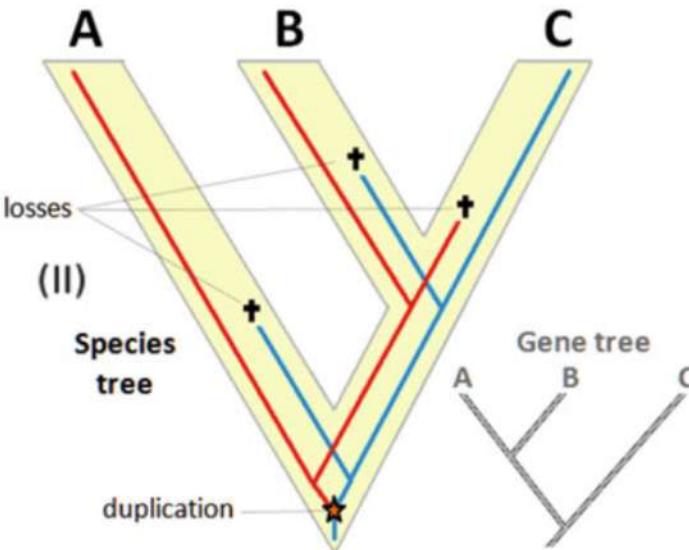
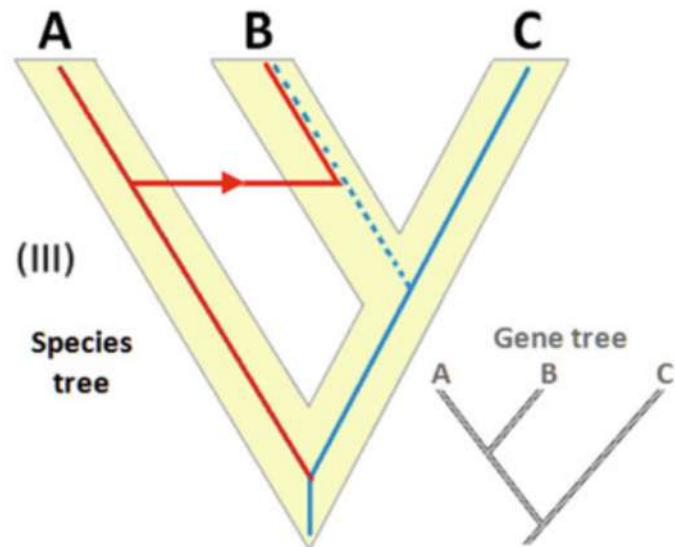


i) Incomplete lineage sorting

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.

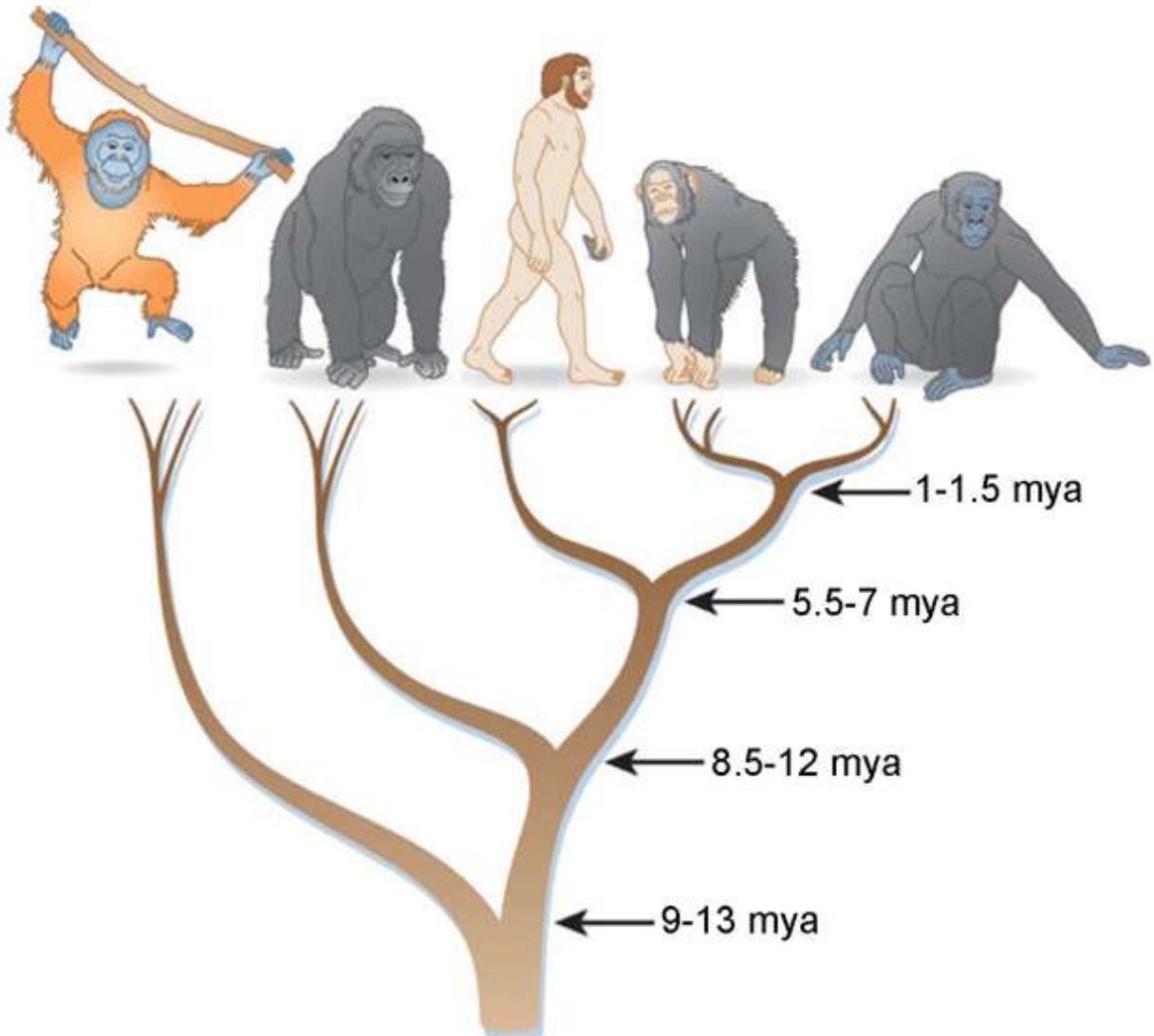
(III) Horizontal Gene Transfer

(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;

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

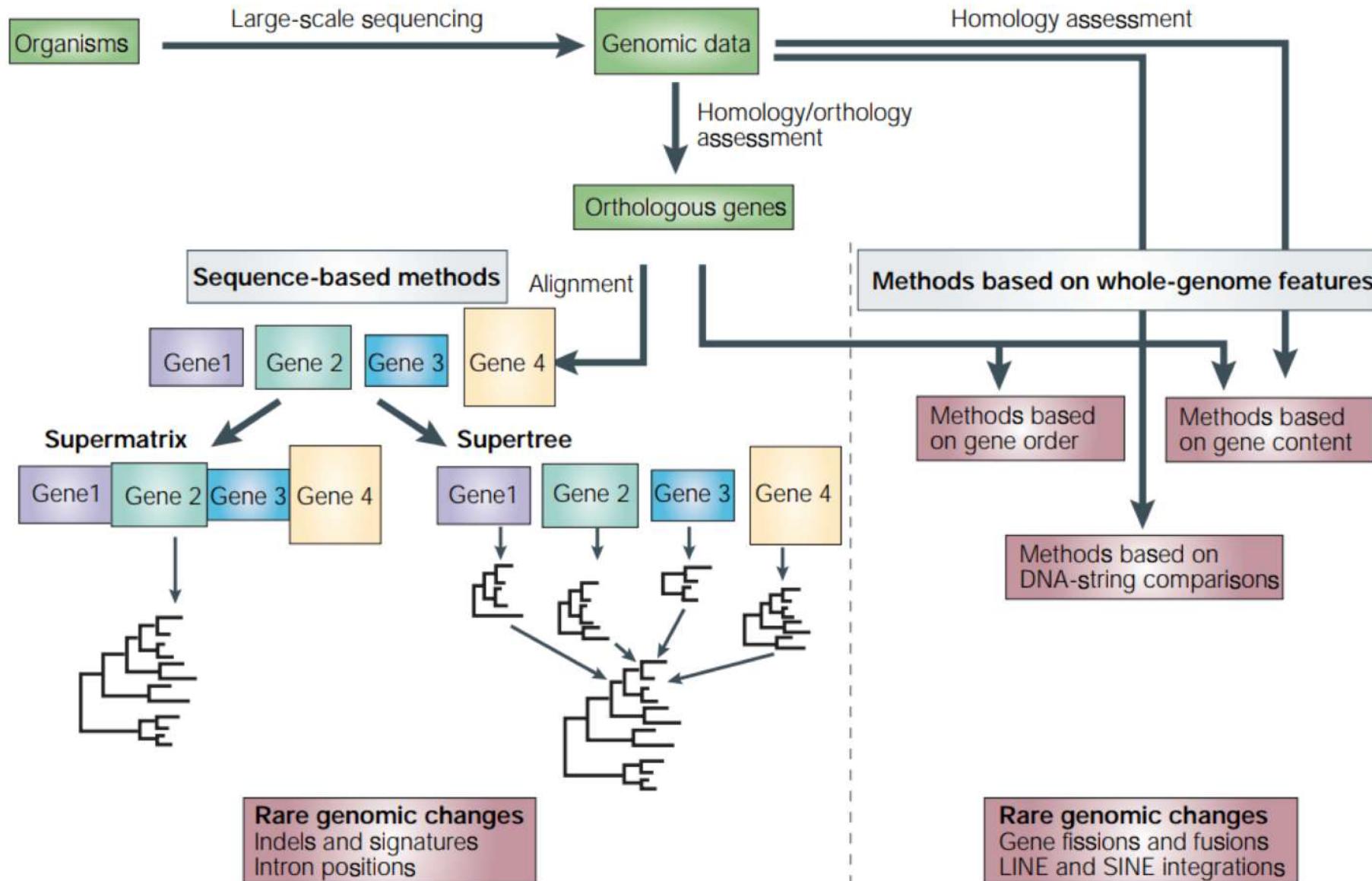


These studies also led to rich discussions about the suite of factors that may have contributed to promoting speciation in the last common ancestor of humans and African apes, as well as the factors that might have contributed to creating the amazing diversity of Hominins that co-existed with each other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

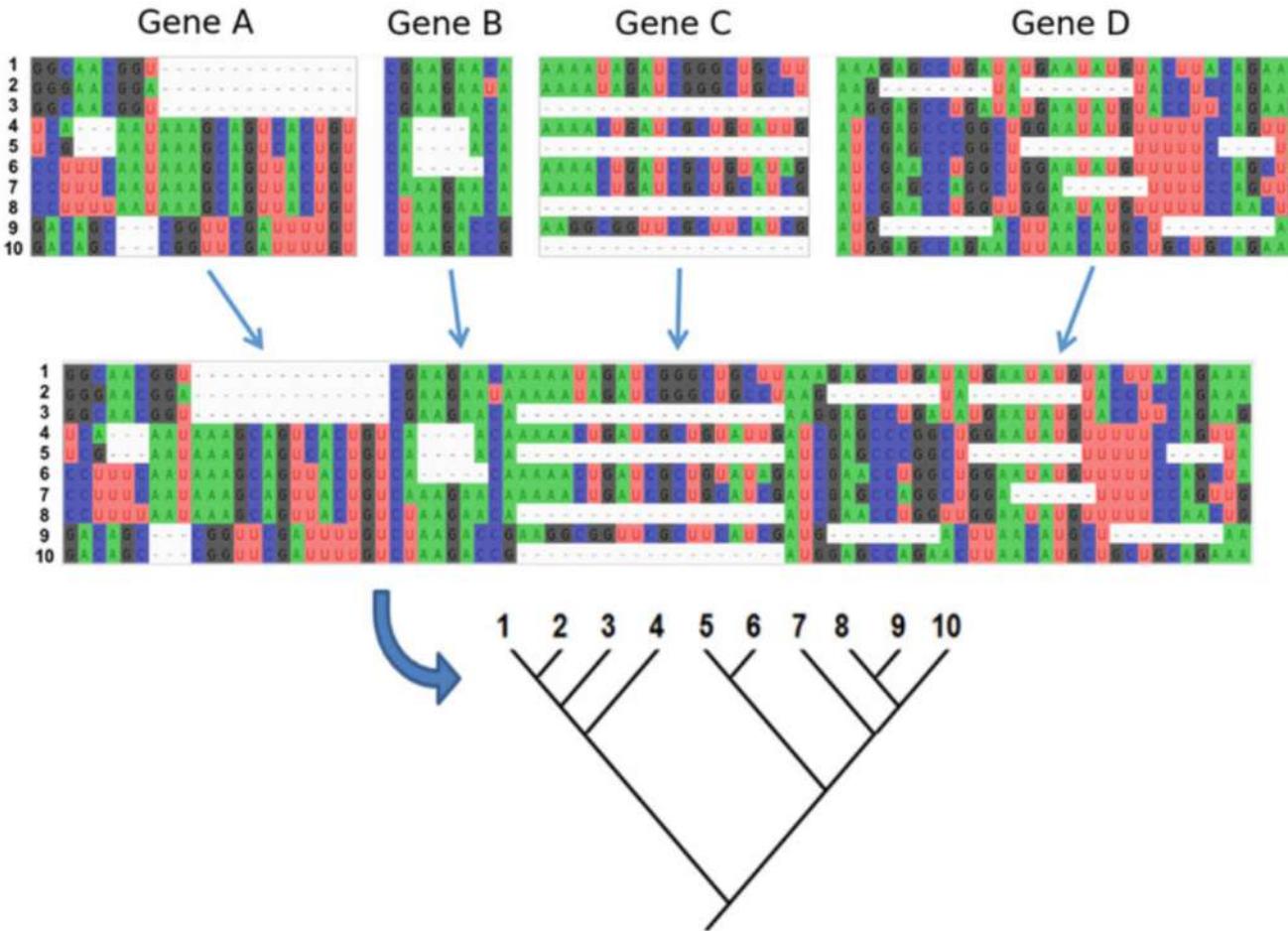
For many years, there was considerable debate about which of the African apes is our closest relative.... The general consensus that emerged is that we share a more recent relationship with chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*) than we do with gorillas (*Gorilla gorilla*) (Ruvolo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

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



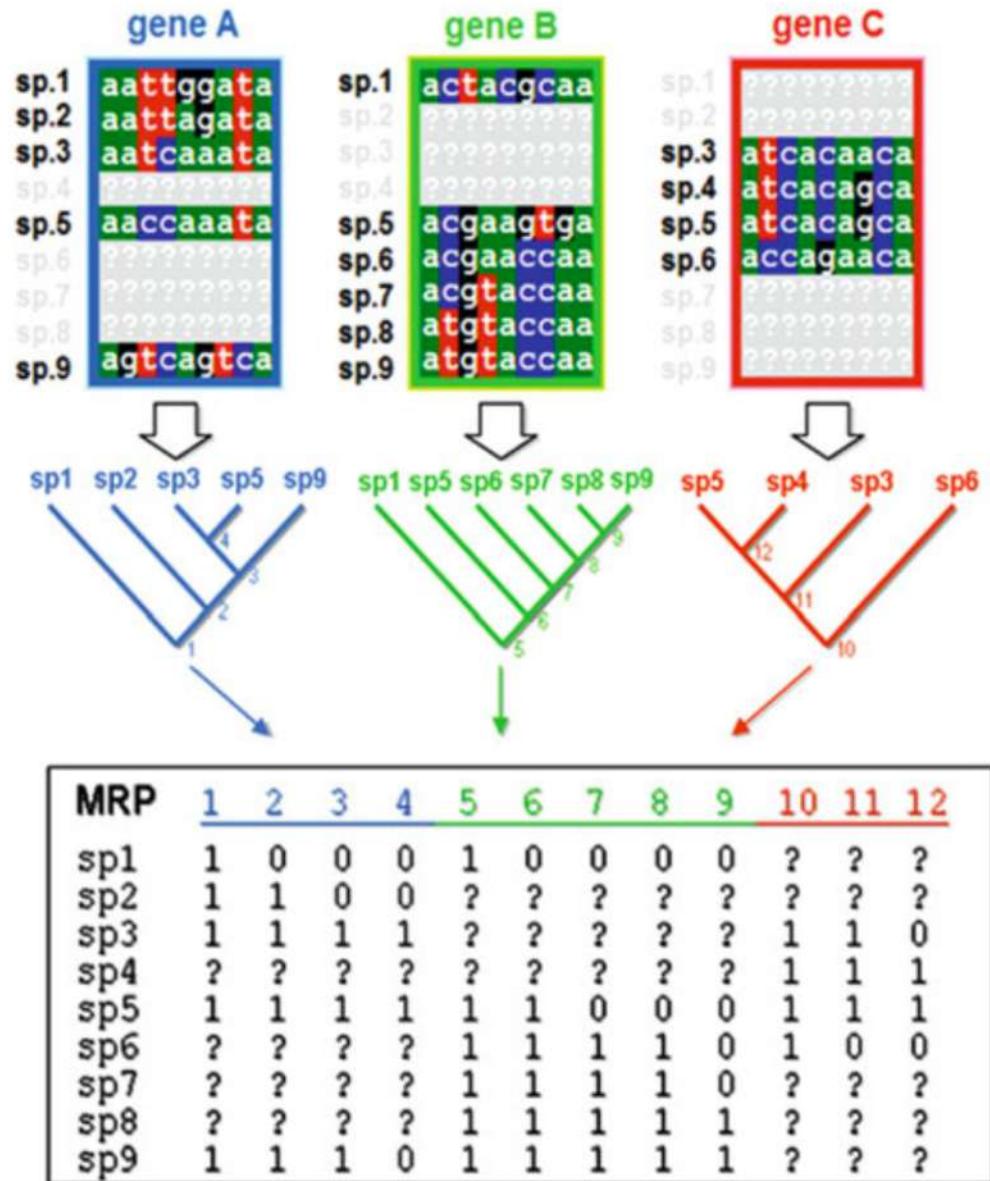
Probably the most common (easy) way to construct alignment of concatenated gene shared across all species



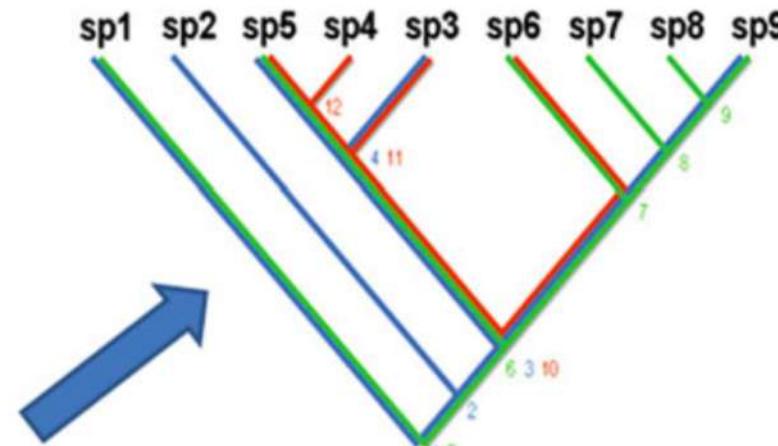
Important drawbacks:

- (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.

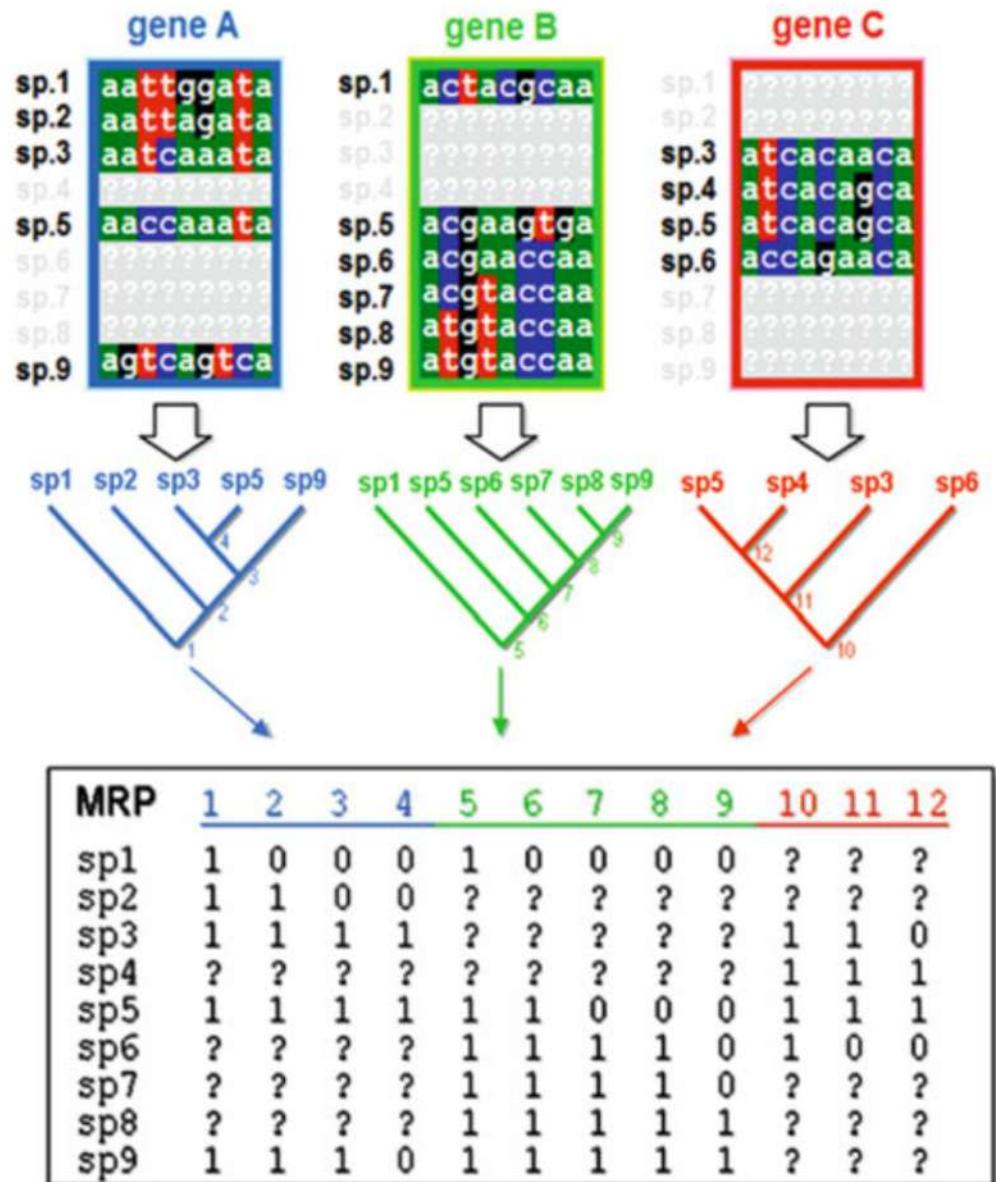
From genes to supertrees



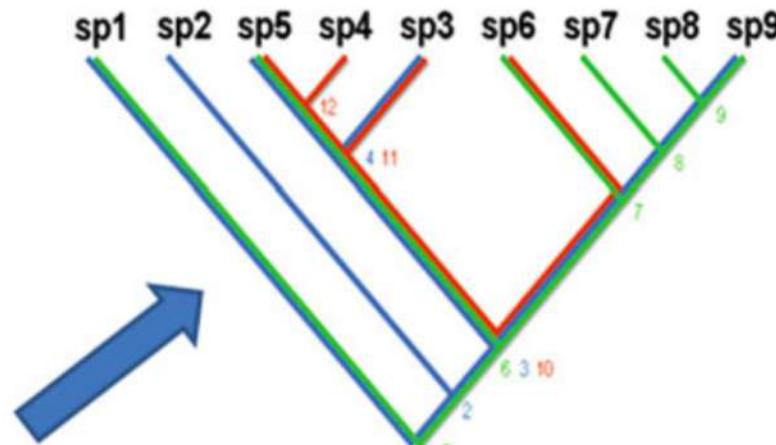
Instead of forcing all gene trees to comply to a single tree, **supertree methods infer the best topology for each gene (using the same phylogenetic method for each)**, and then a topological consensus is obtained. Such methods are able to make consensus trees even if the number of leaves among gene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



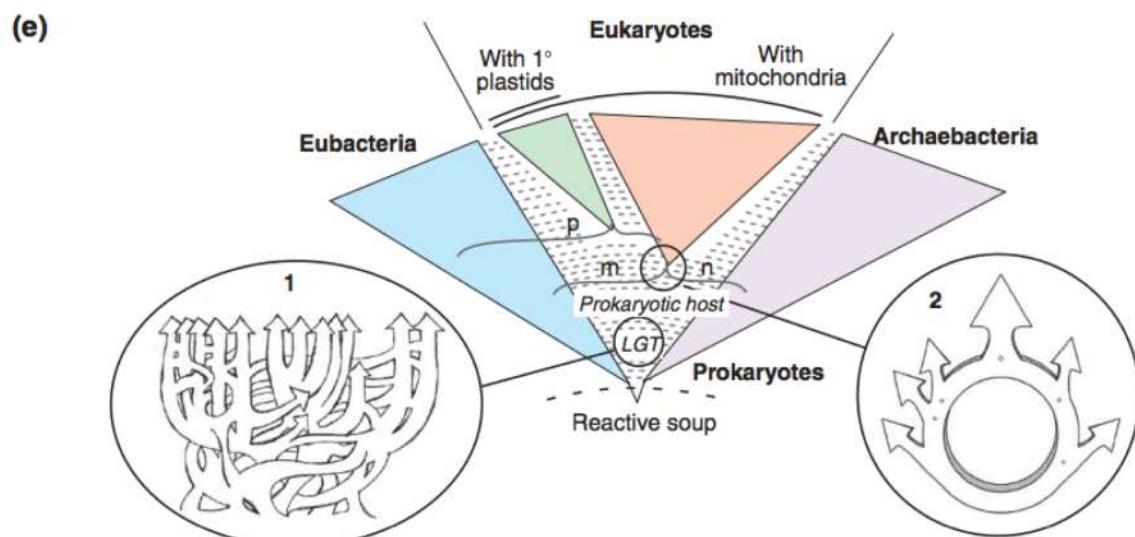
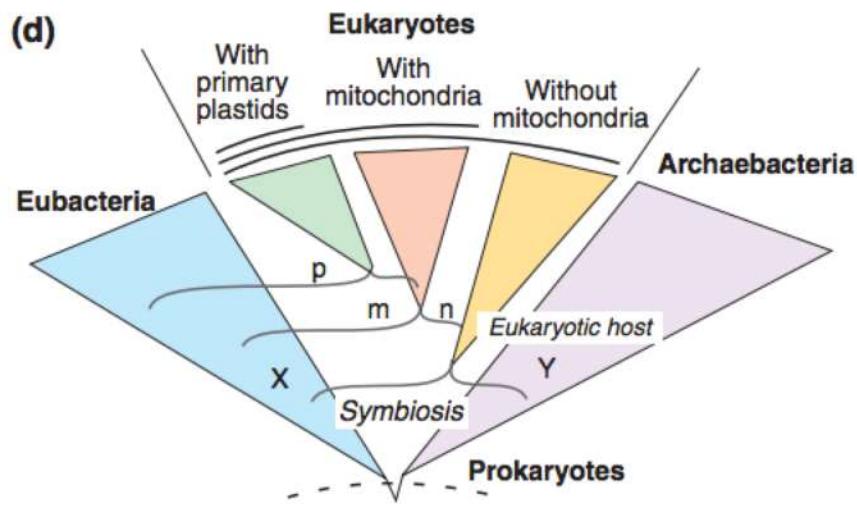
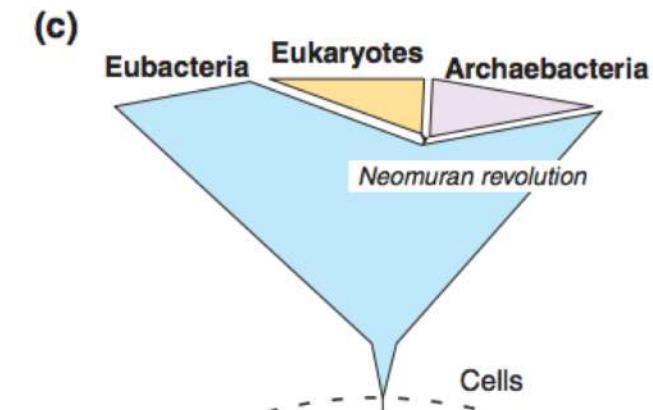
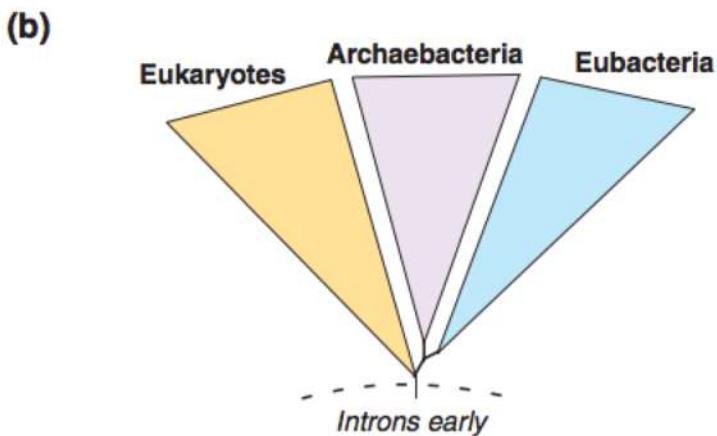
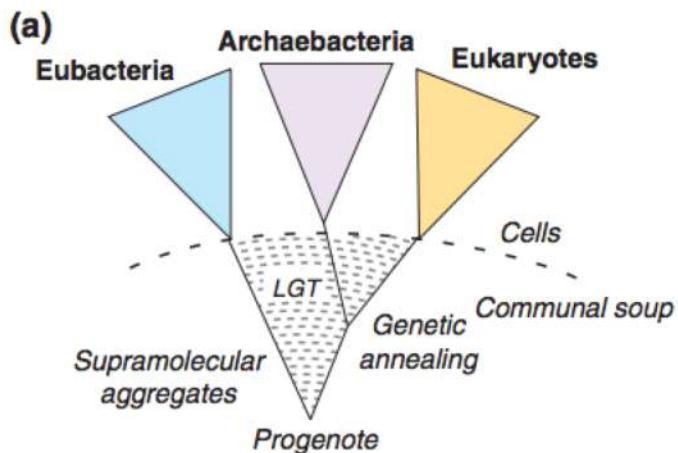
Current methods



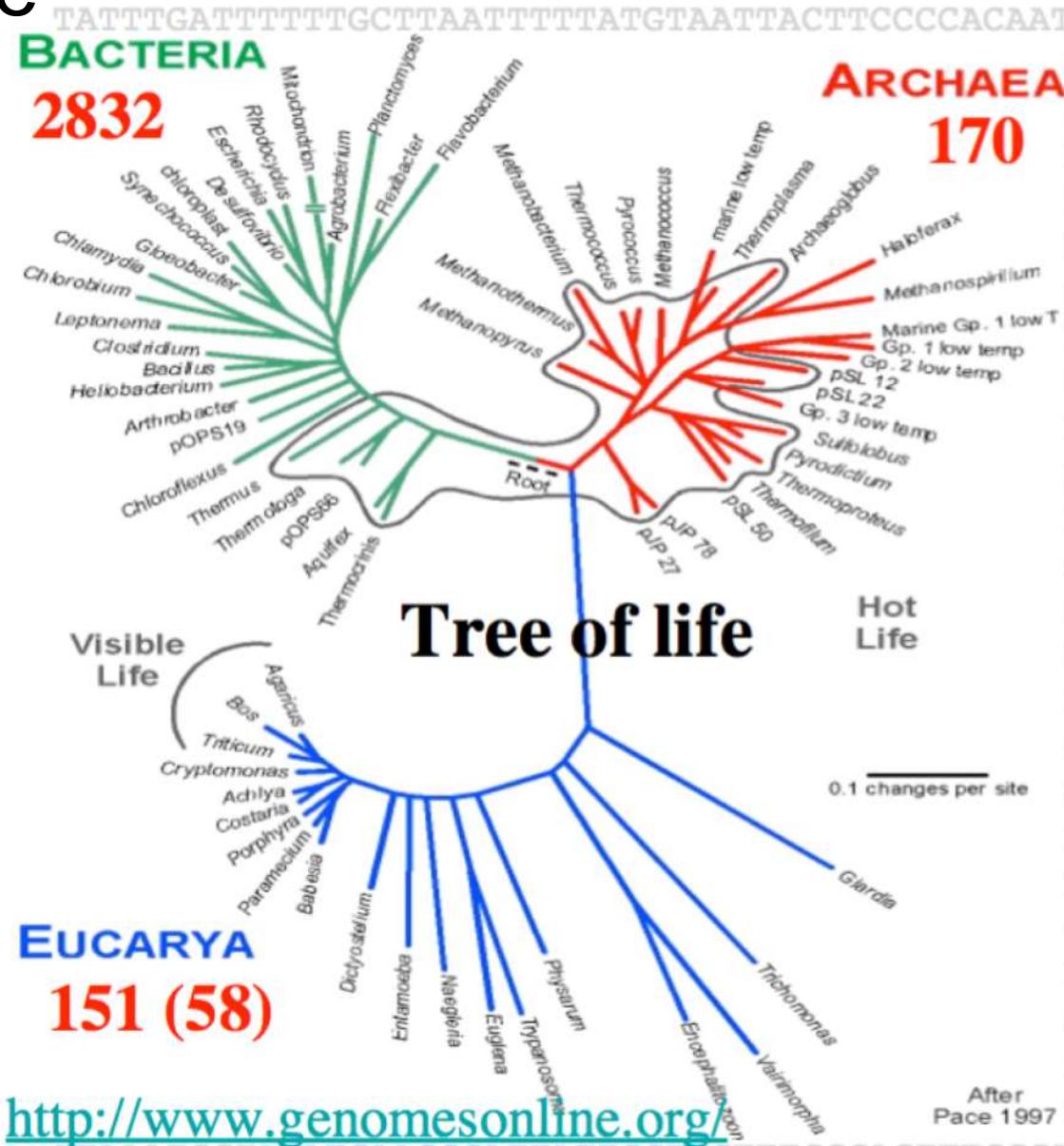
A step beyond supertrees is the use of methods that take into consideration specific evolutionary processes that may be responsible for differences in gene topologies, and then estimate the species tree which would most likely have generated such gene trees, under different scenarios



Five models models of tree of life



Tree of life



<http://www.genomesonline.org/>

ATAAAATGGTAGATGAAGCGATTATTCGCTTGTCTTCCCACCTTTTCCCCA

AATAATAATTAATAAAAAGGGGTAAAAAT

Viruses: 3333

Complete finished genomes: 3060

(04/09/14)

- 2832 Bacteria
 - 170 Archaea
 - 58 eukaryotes

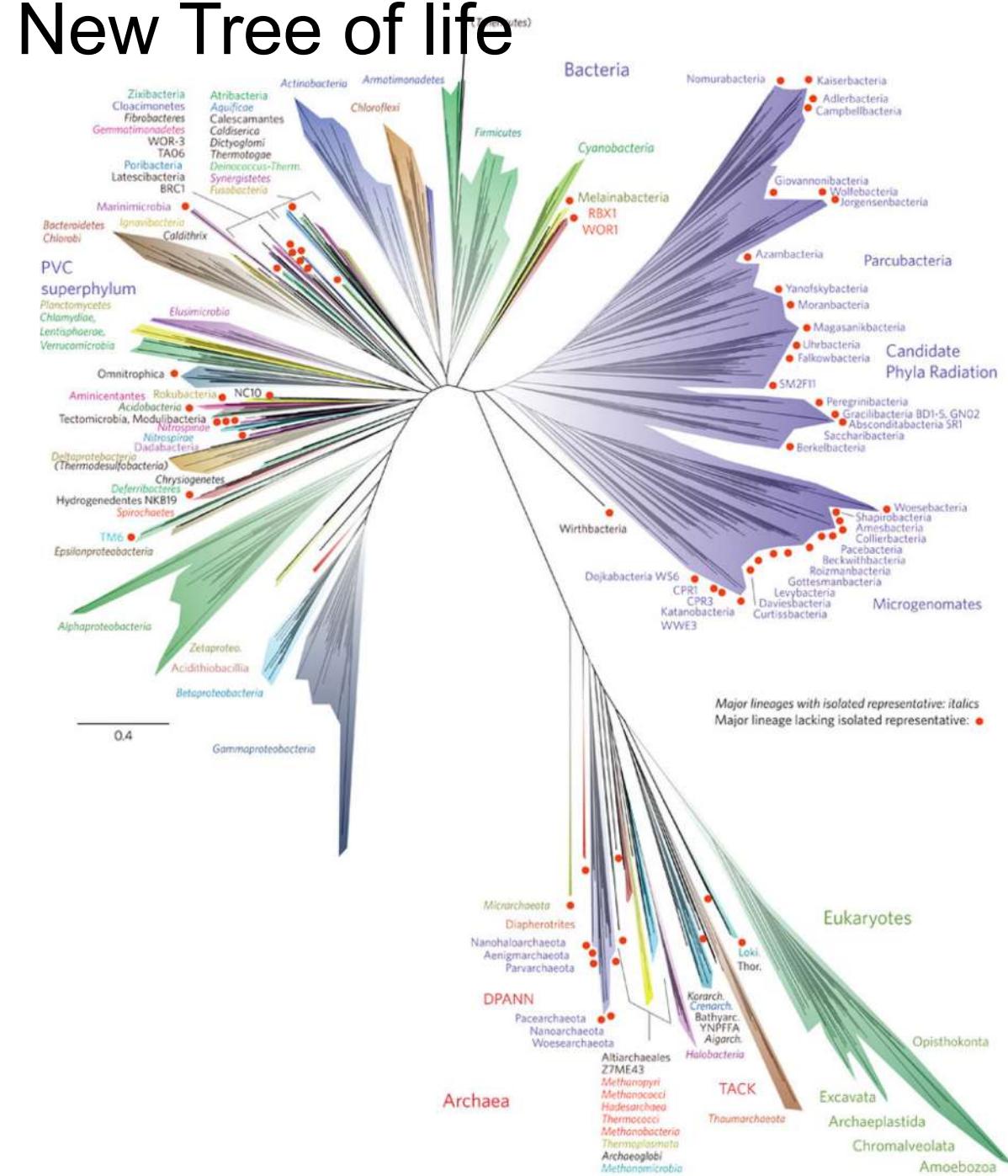
Incomplete genomes projects: 38262

- 32068 Bacteria
 - 664 Archaea
 - 5530 Eukaryotes

Transcriptomes: 947

- 76 Bacteria
 - 11 Archaea
 - 860 Eukaryota

New Tree of life



The third trunk that Woese and his colleagues identified included little-known microbes that live in extreme places like hot springs and oxygen-free wetlands. Woese and his colleagues called this third trunk Archaea.

Dr. Banfield said she expected new branches to be discovered for eukaryotes, especially for tiny species such as microscopic fungi. “That’s where I think the next big advance might be found,” Dr. Banfield said.

Dr. Hug disagreed that scientists were done with bacteria. “I’m less convinced we’re hitting a plateau,” she said. “There are a lot of environments still to survey.”

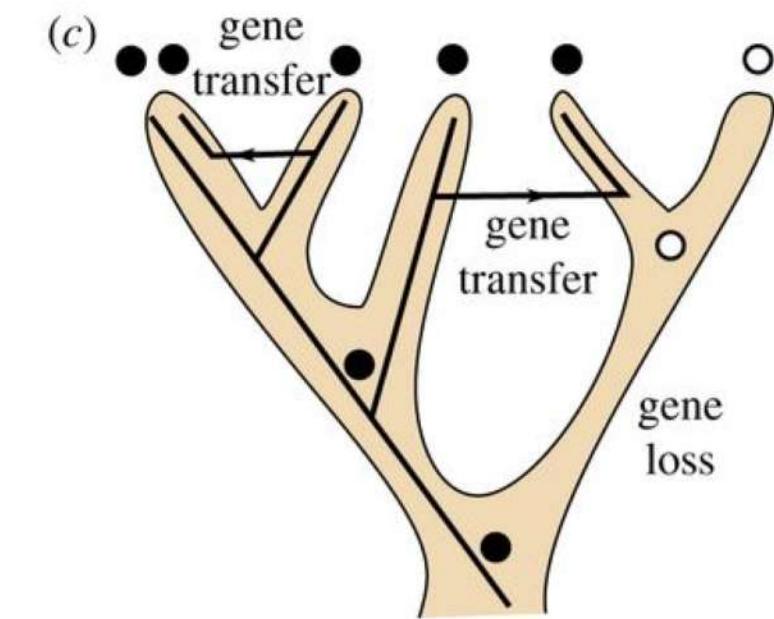
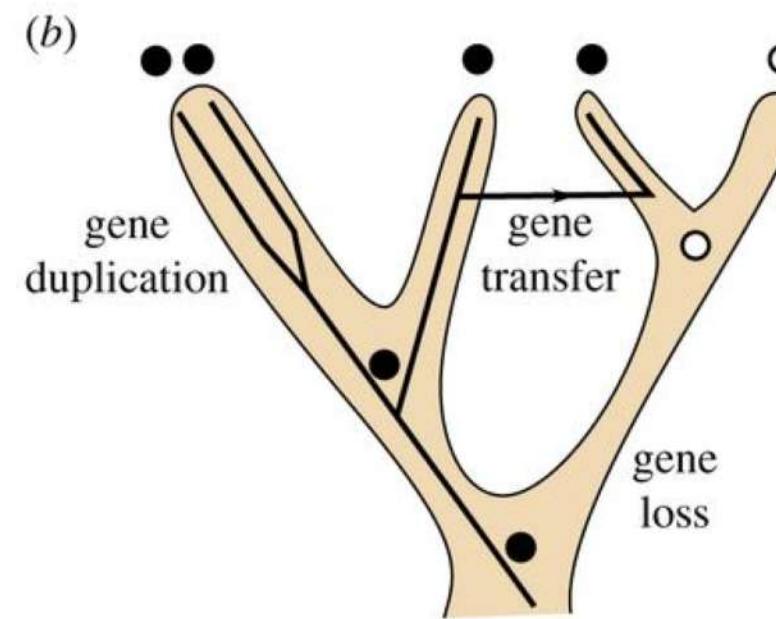
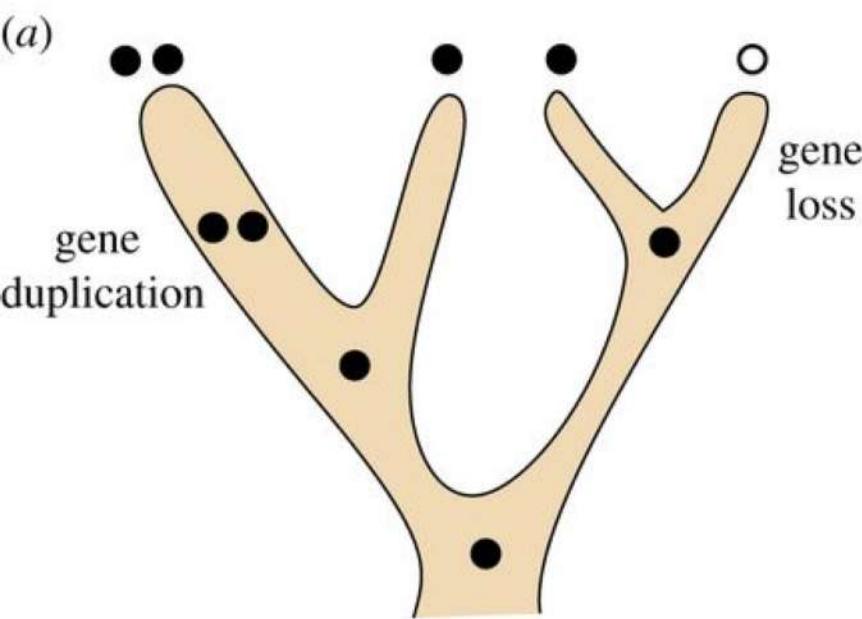
Hug et al (2016)

http://www.nytimes.com/2016/04/12/science/scientists-unveil-new-tree-of-life.html?_r=0

Horizontal gene transfer (HGT)

Inferring HGT require

- 1) species phylogeny ; 2) gene phylogeny
- 3) extensive taxon sampling



● presence ○ absence

Complicated history of genes: dig into finer details

Gene fusion



Gene fission



Domains shuffling



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 Microbiology and Immunology, University of Wisconsin-Madison, WI, USA
†Department of Medical Microbiology and Immunology, University of Wisconsin-Madison, WI, USA
‡Department of Biochemistry, University of Wisconsin-Madison, WI, USA

Edited by John J. Mekalanos, Harvard University, MA, USA

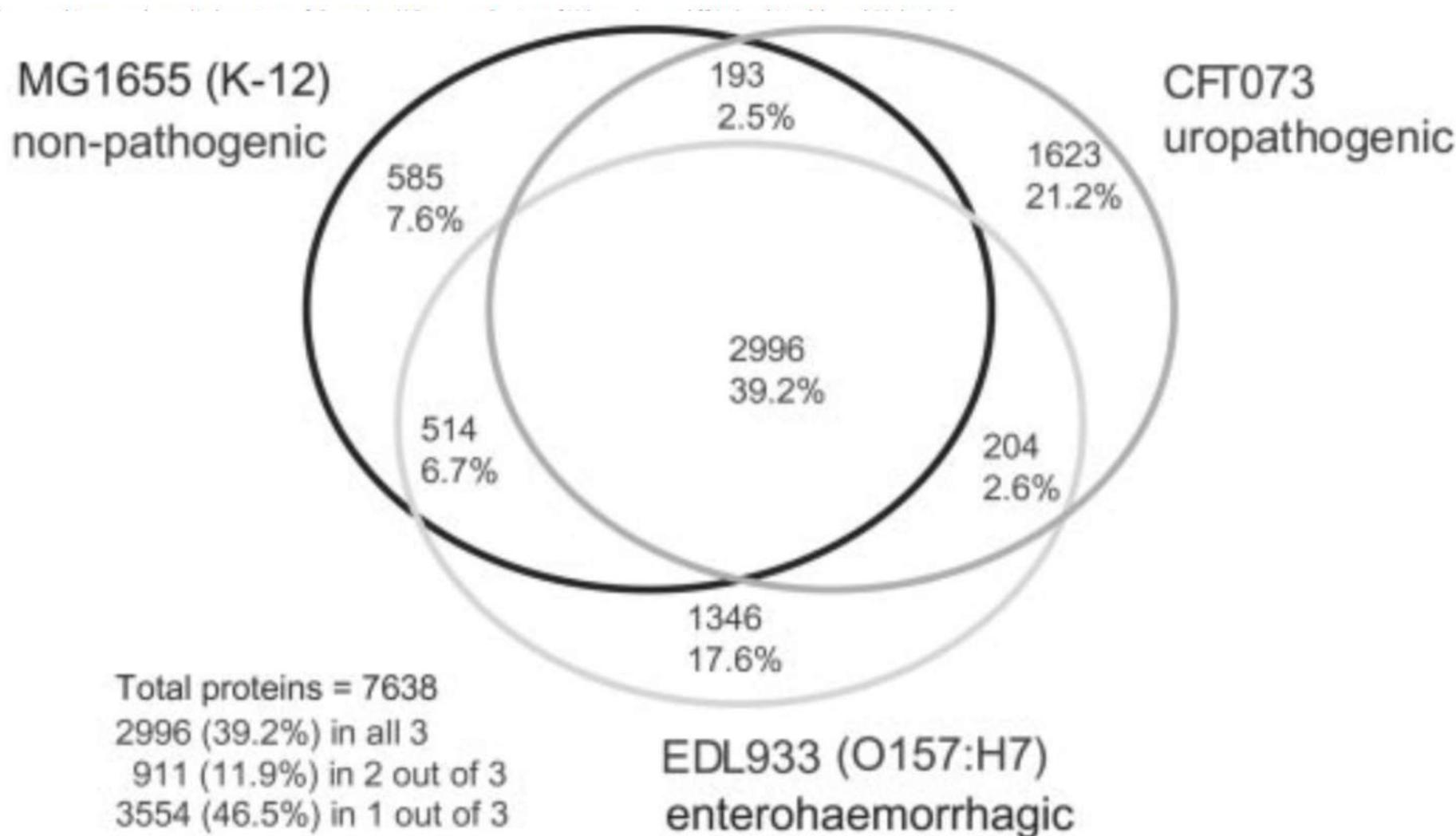


Illustration of a gene content Venn diagram for three hypothetical genomes A, B, and C

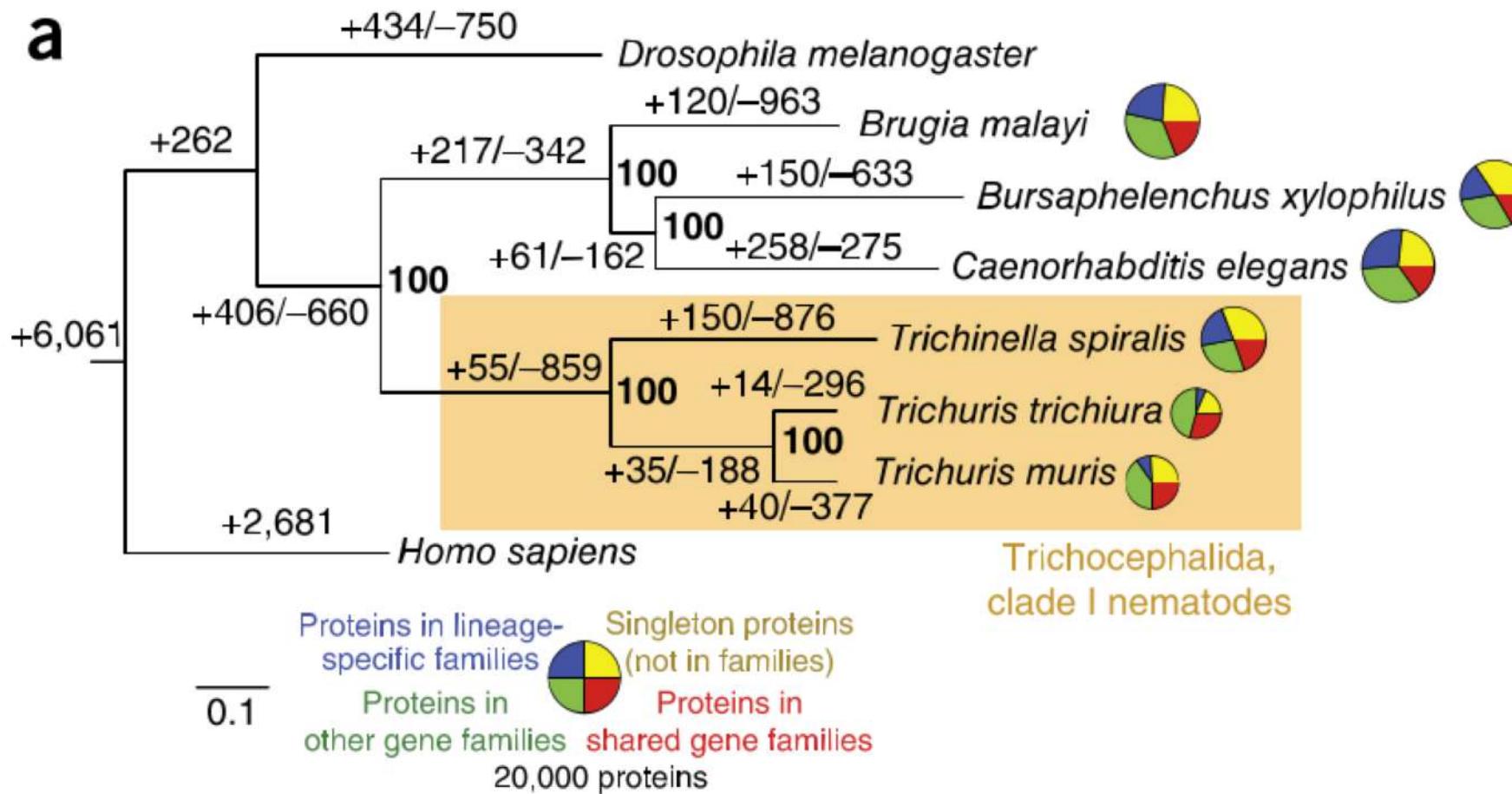
Gene	Genome						
	A	B	C	D	E	F	G
1	✓	✓			✓	✓	
2	✓		✓	✓	✓	✓	✓
3		✓		✓			
4		✓			✓		
5				✓			
6			✓		✓	✓	
7		✓		✓			✓



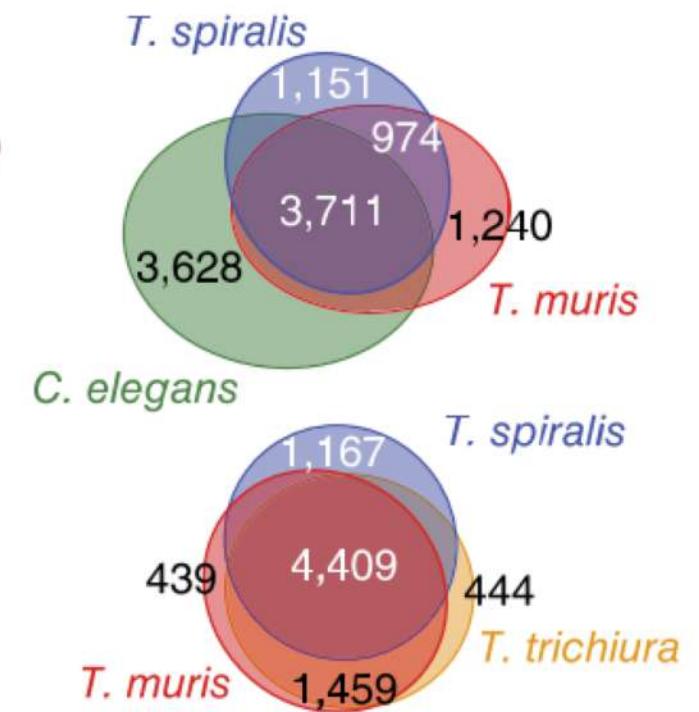
Schematic representation of a presence/absence gene matrix. Genomes are represented in columns, and gene families are represented in rows

Phylogeny + Venn diagram to show expansion/loss

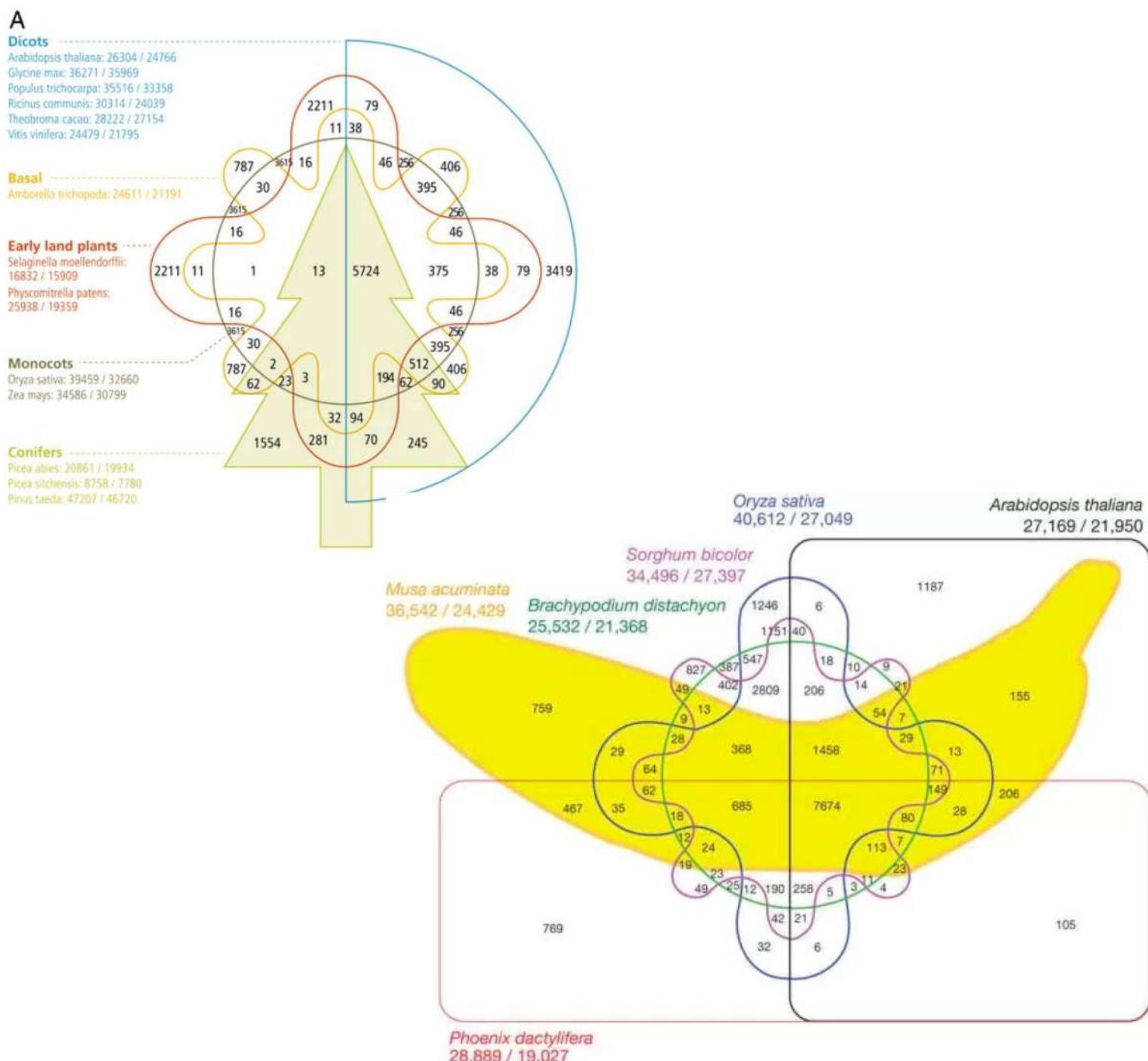
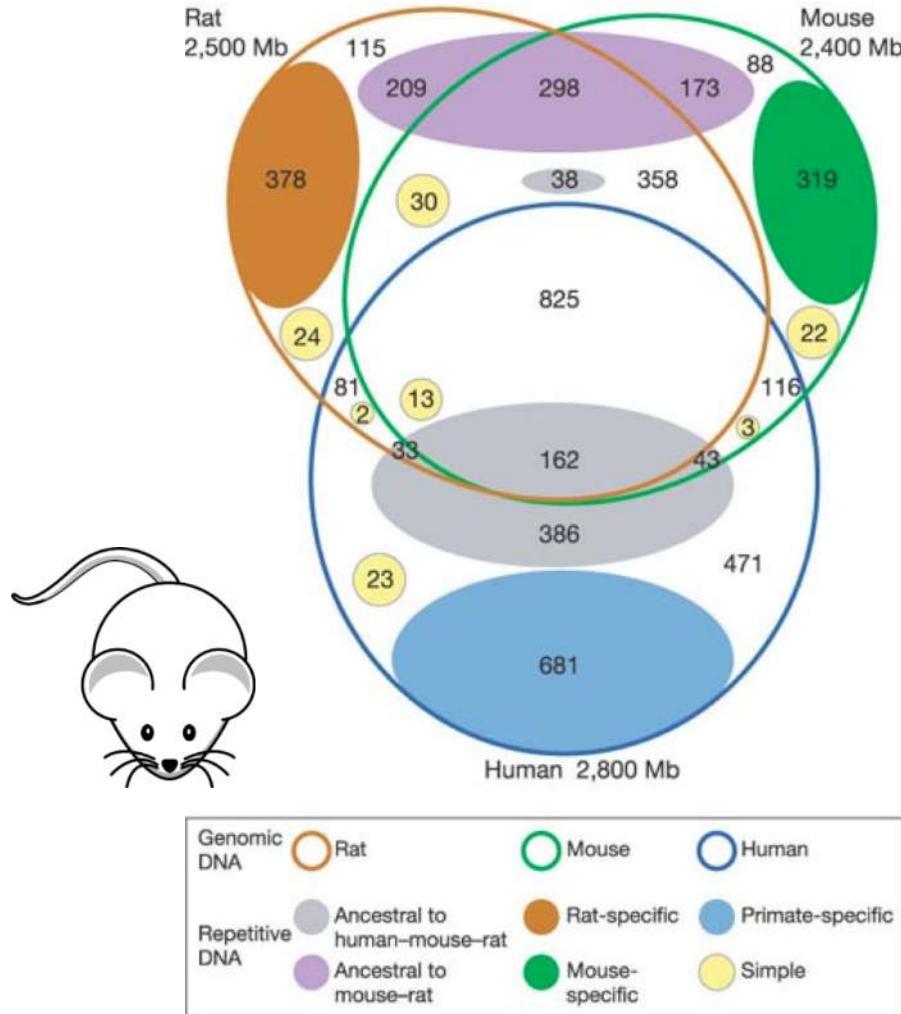
a



b



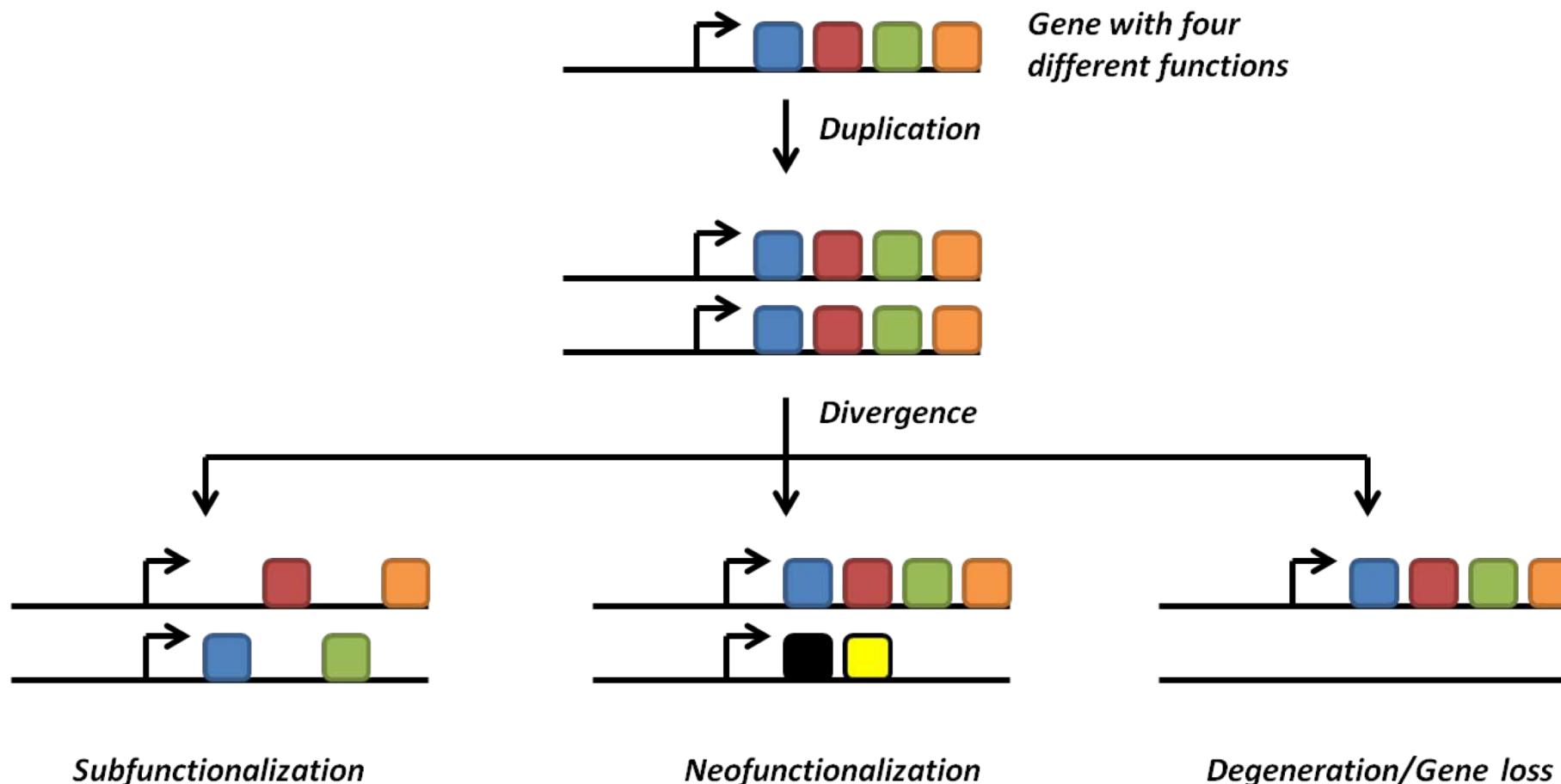
Trend of venn diagram...



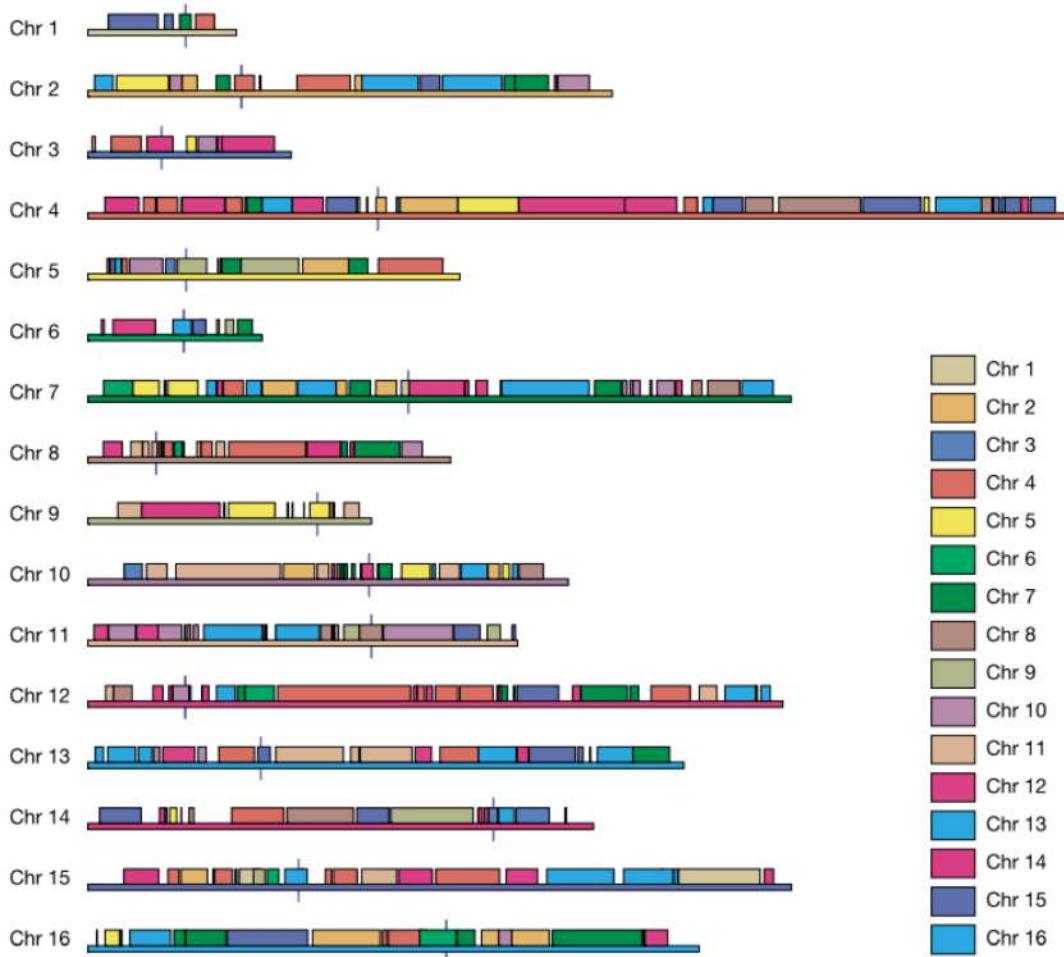
Gene and genome duplication

Why study gene duplication?

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

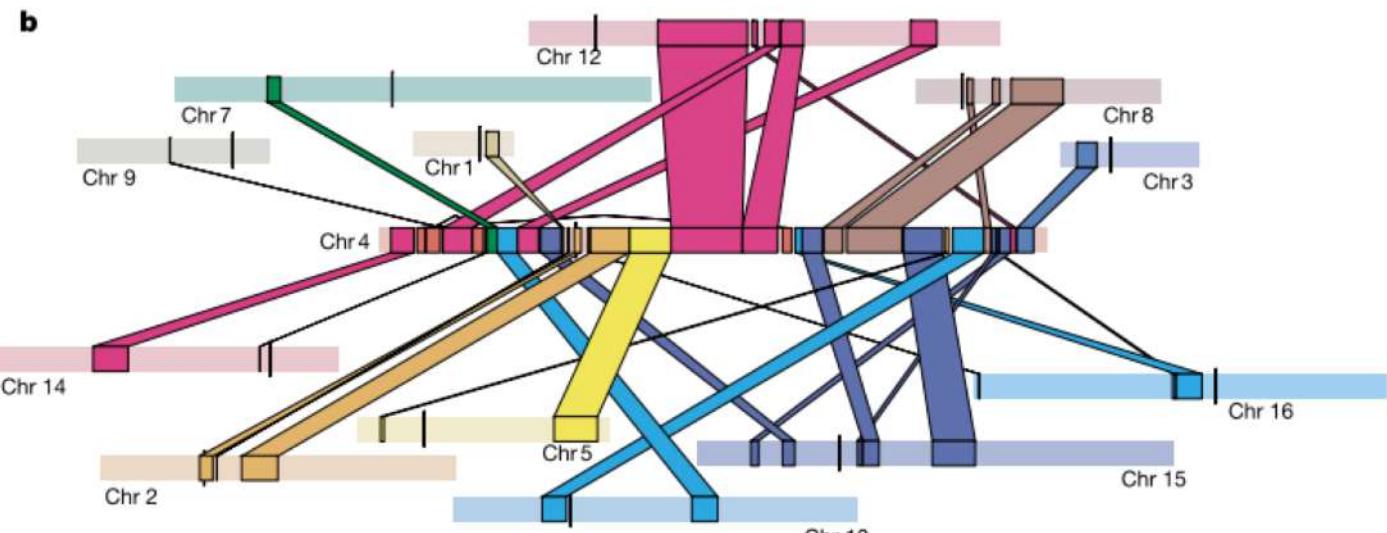


Within species

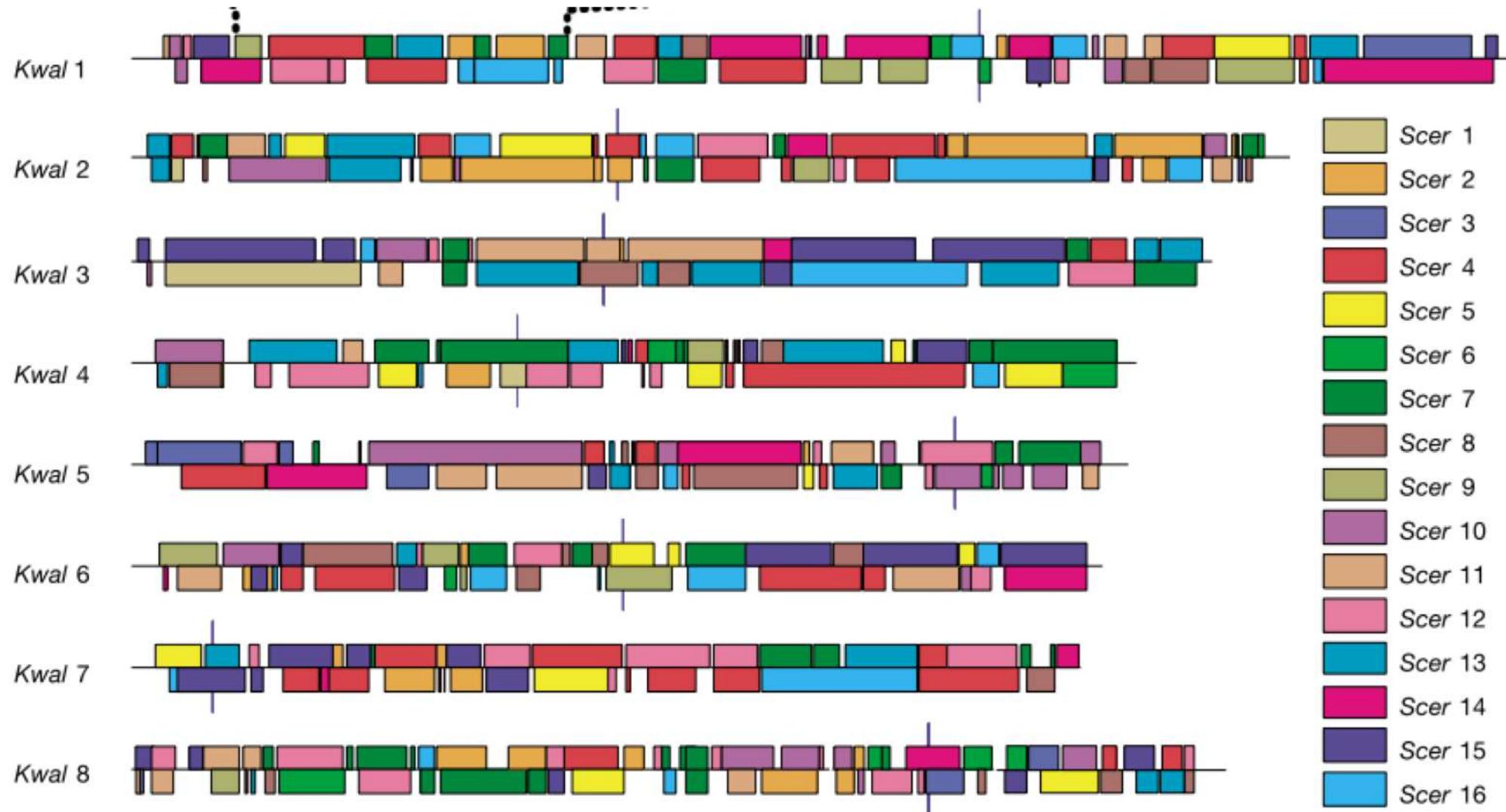


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16

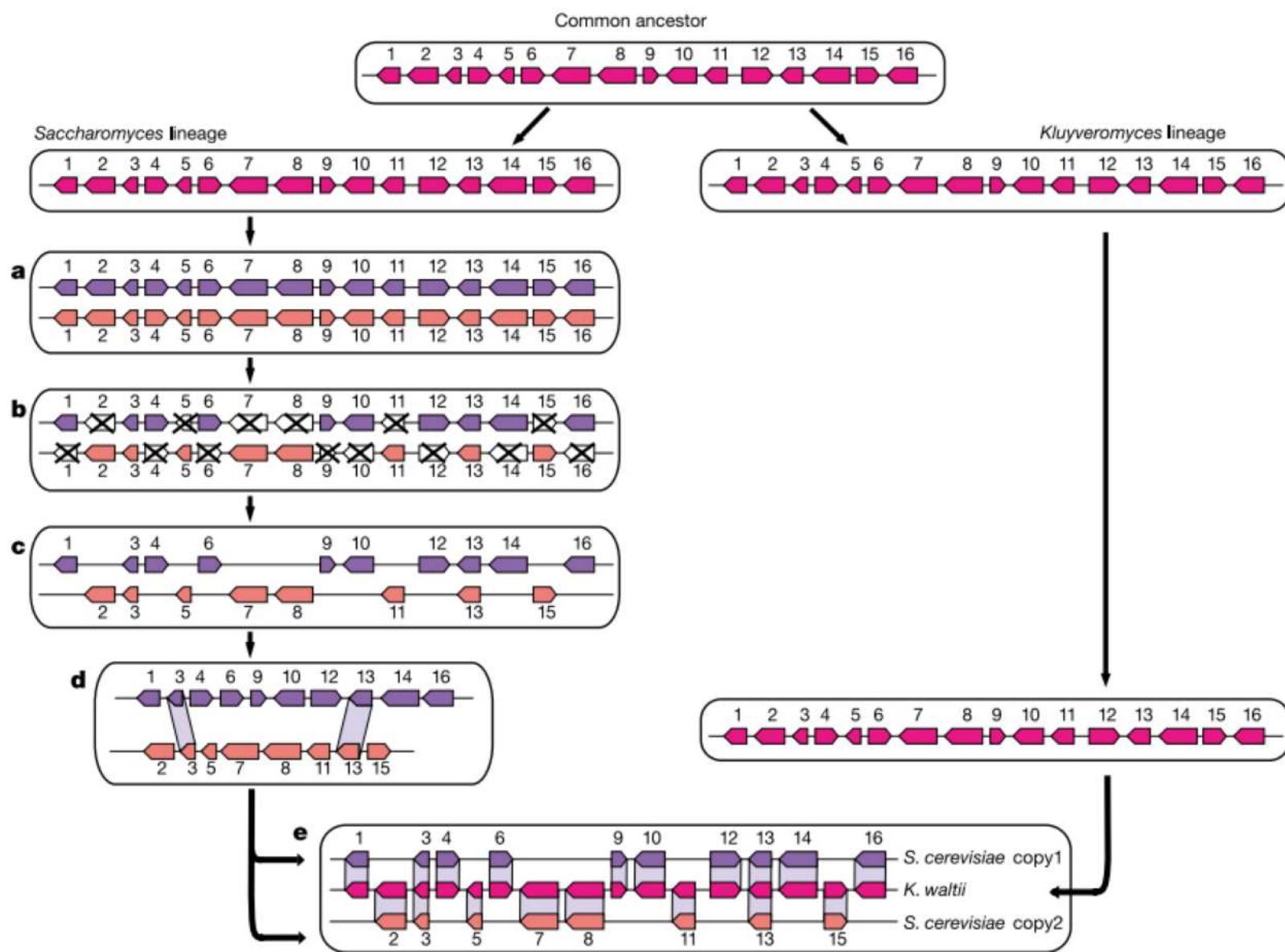
b



Between species

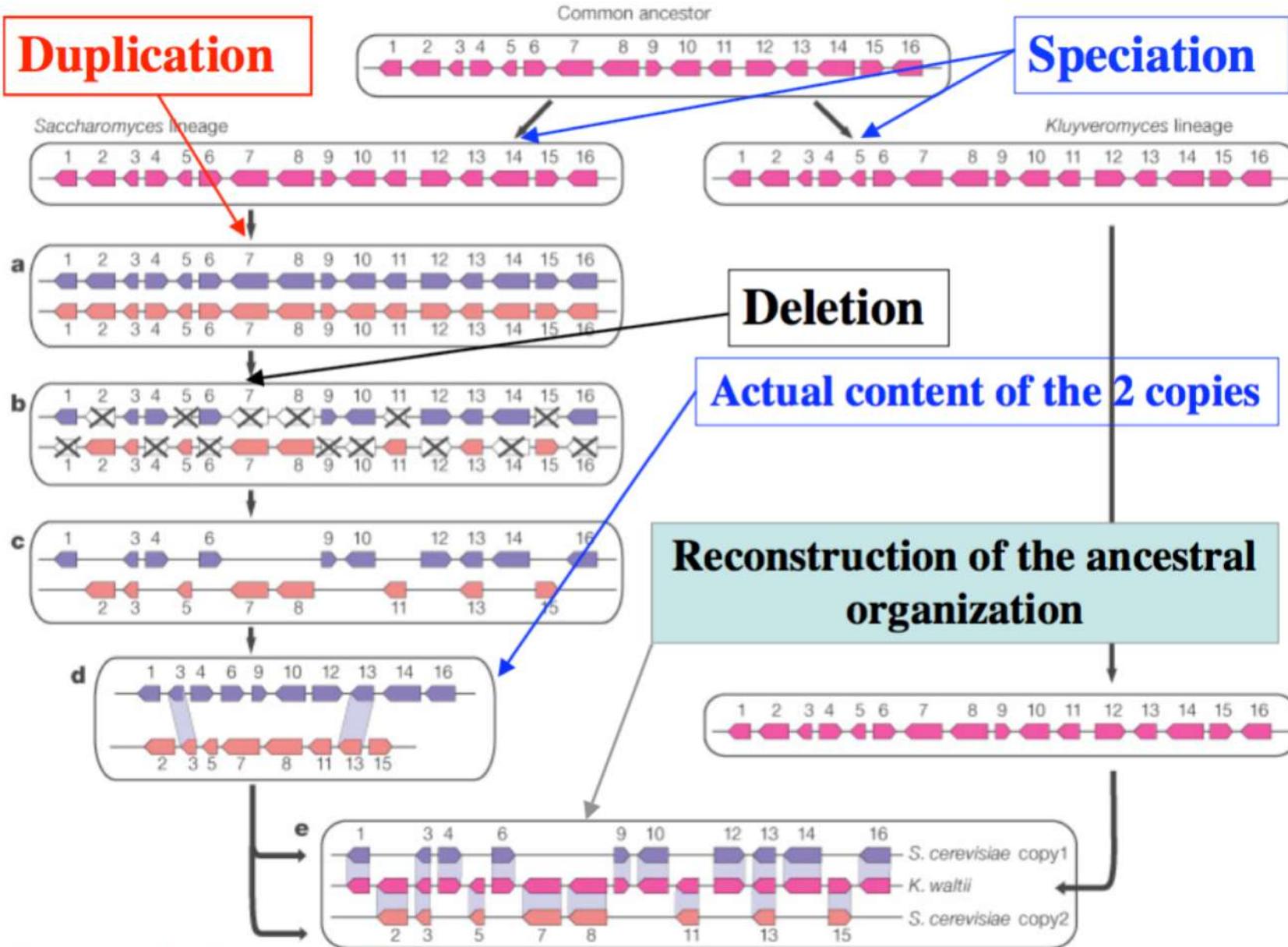


Whole genome duplication model



Kellis et al (2004)

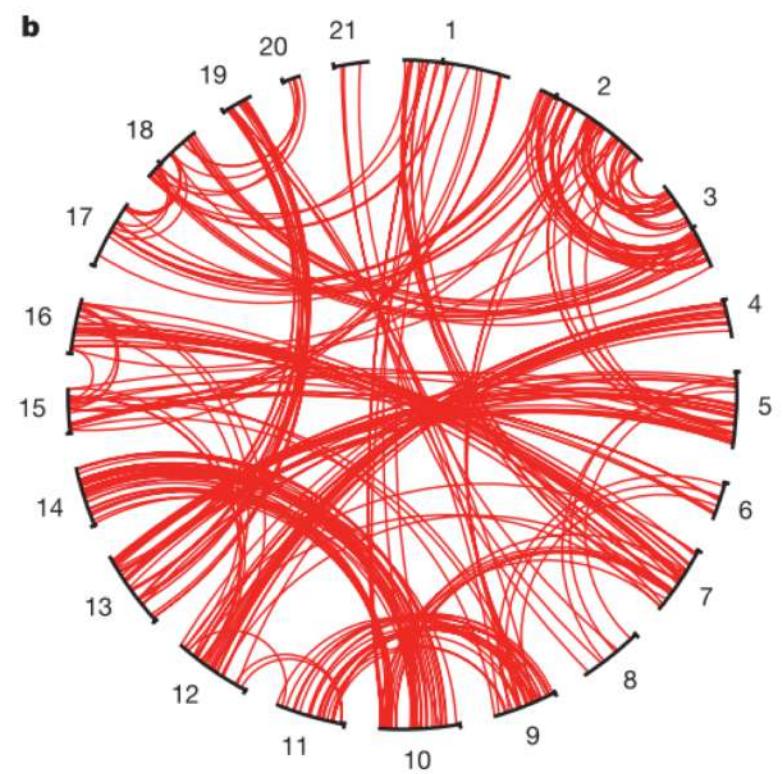
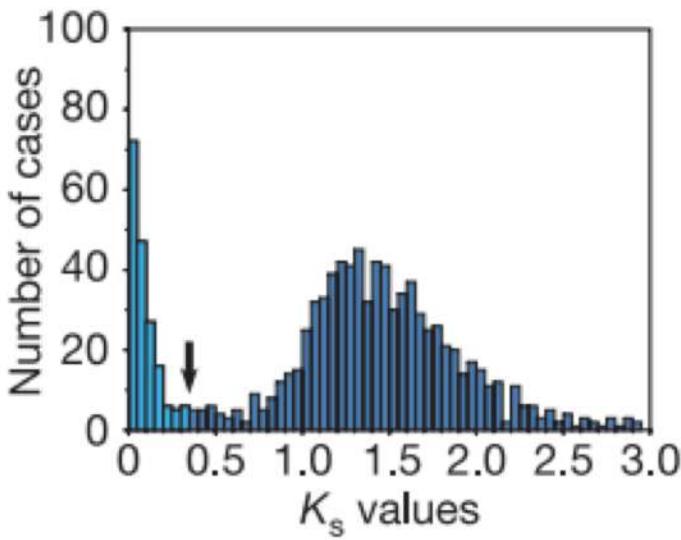
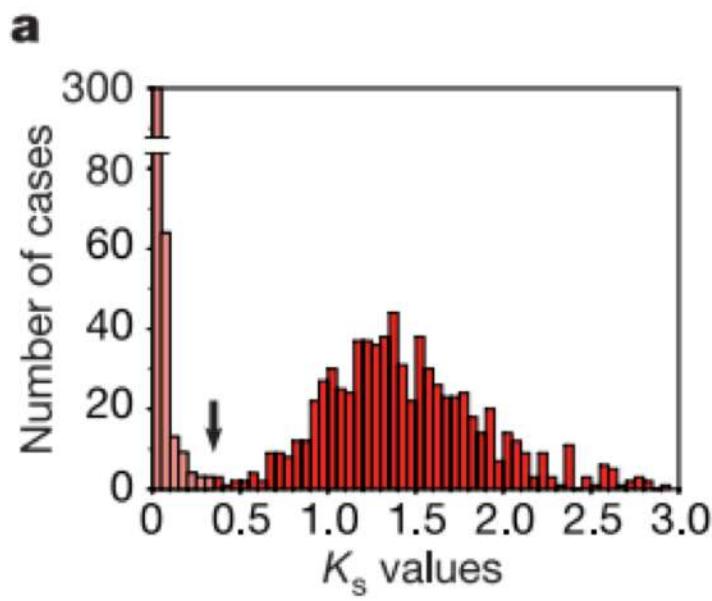
Determining ancestral conservation



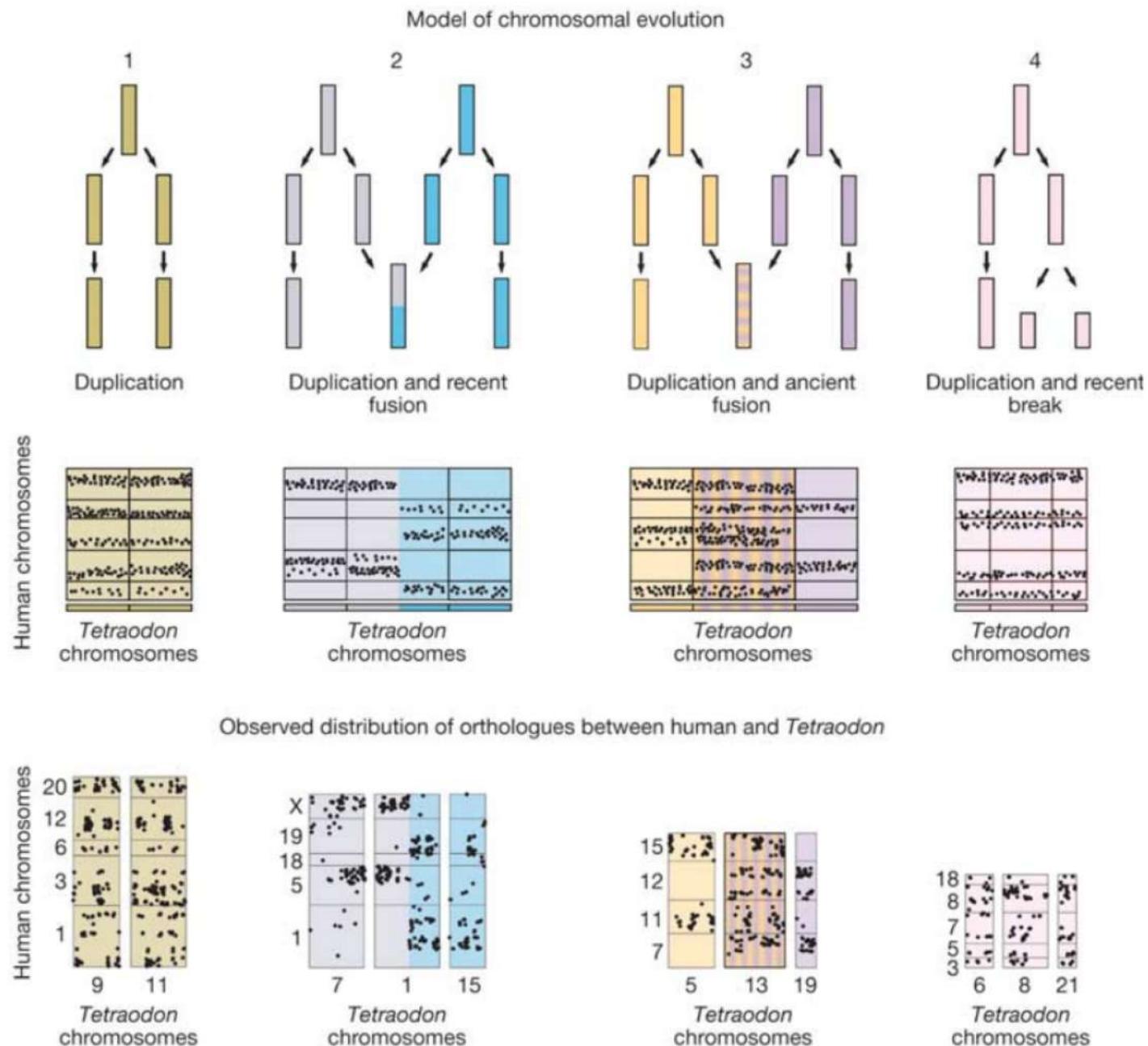
Kellis et al. 2004. *Nature*, 428:617-24.

Slide of Fred Tekaia

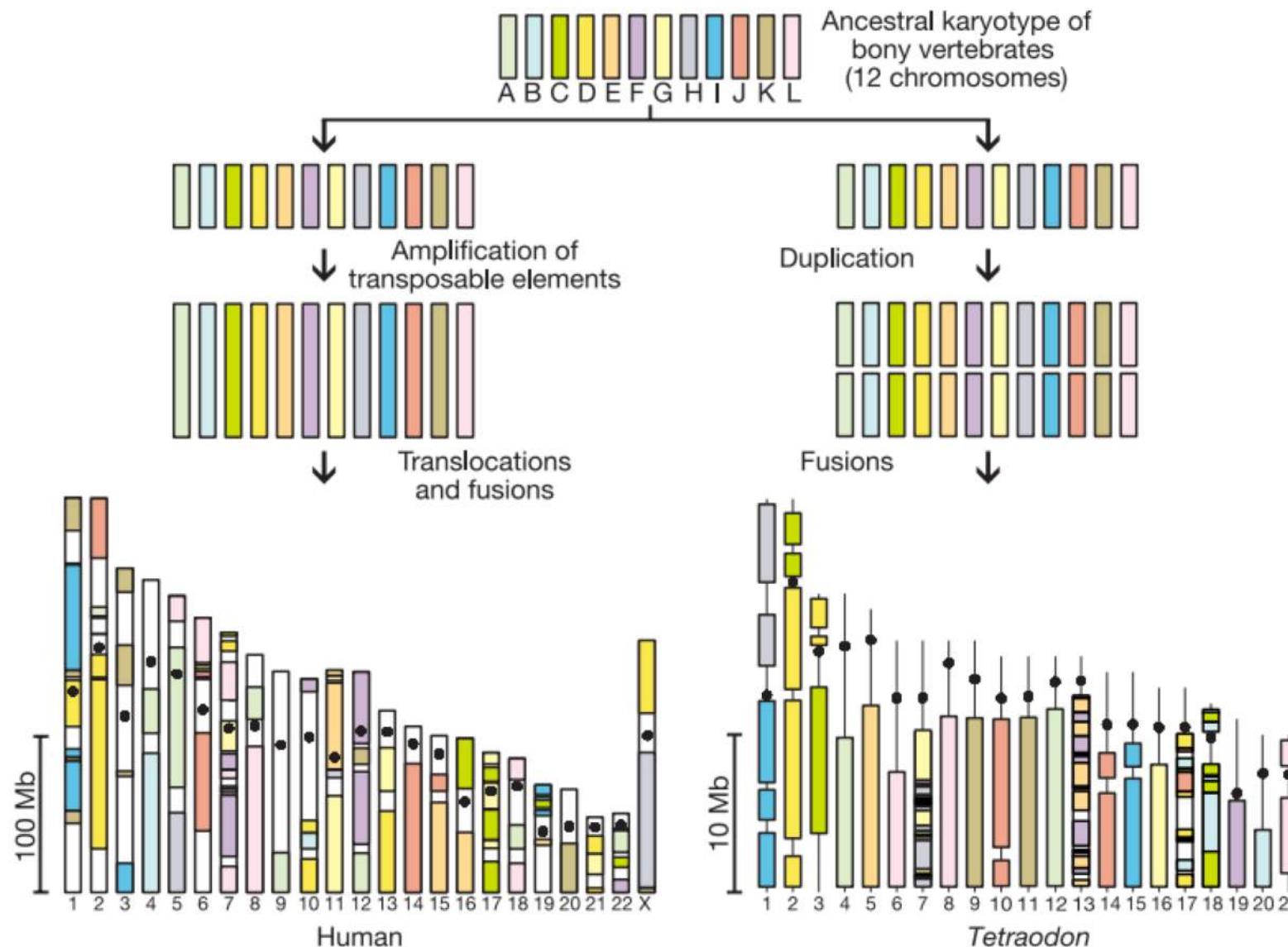
Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype



Reconstructing ancient genome rearrangement

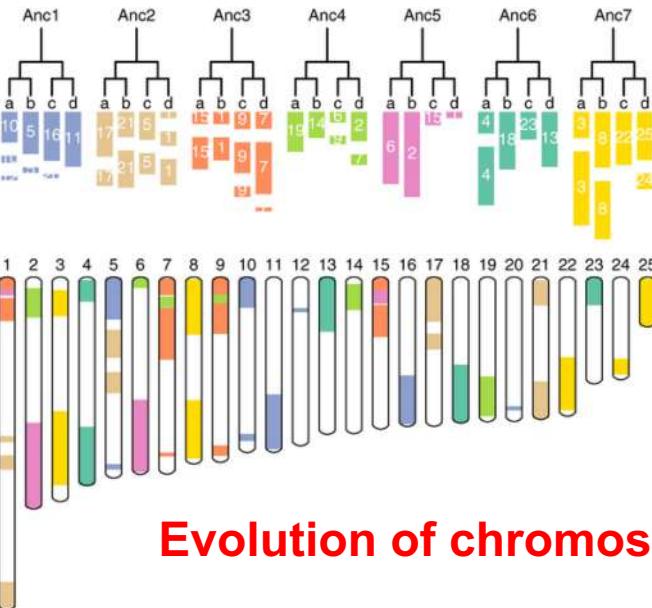
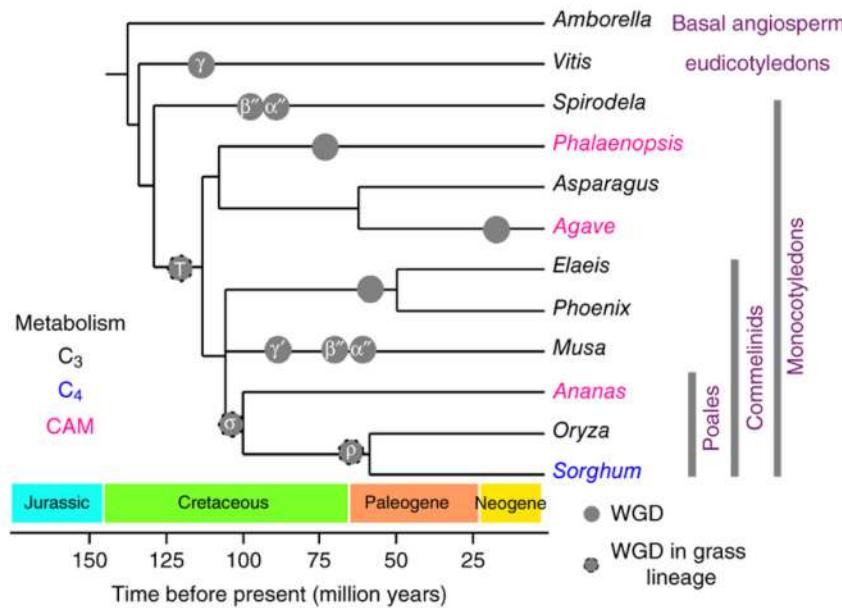


Reconstructing ancient genome rearrangement

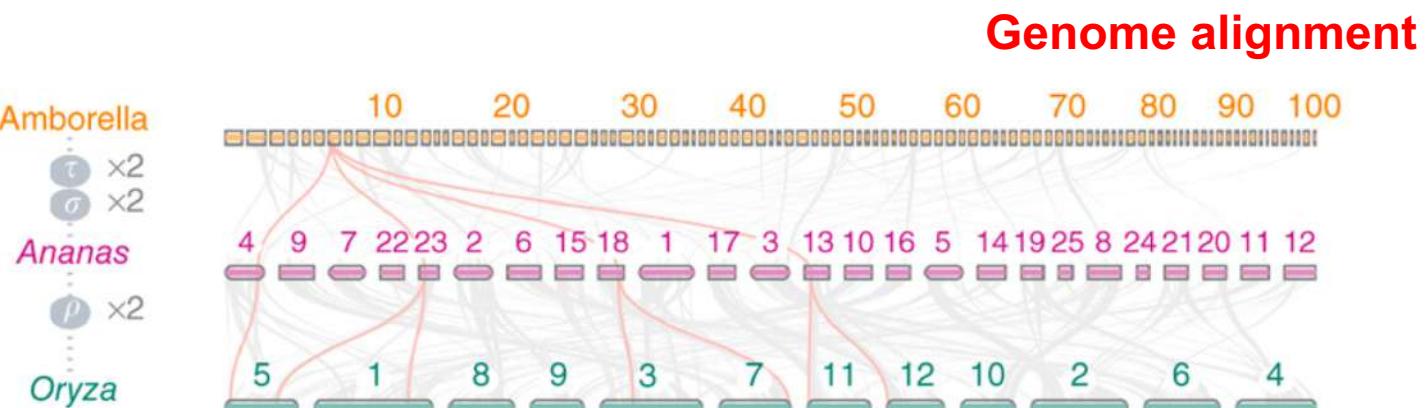


Pineapple genome

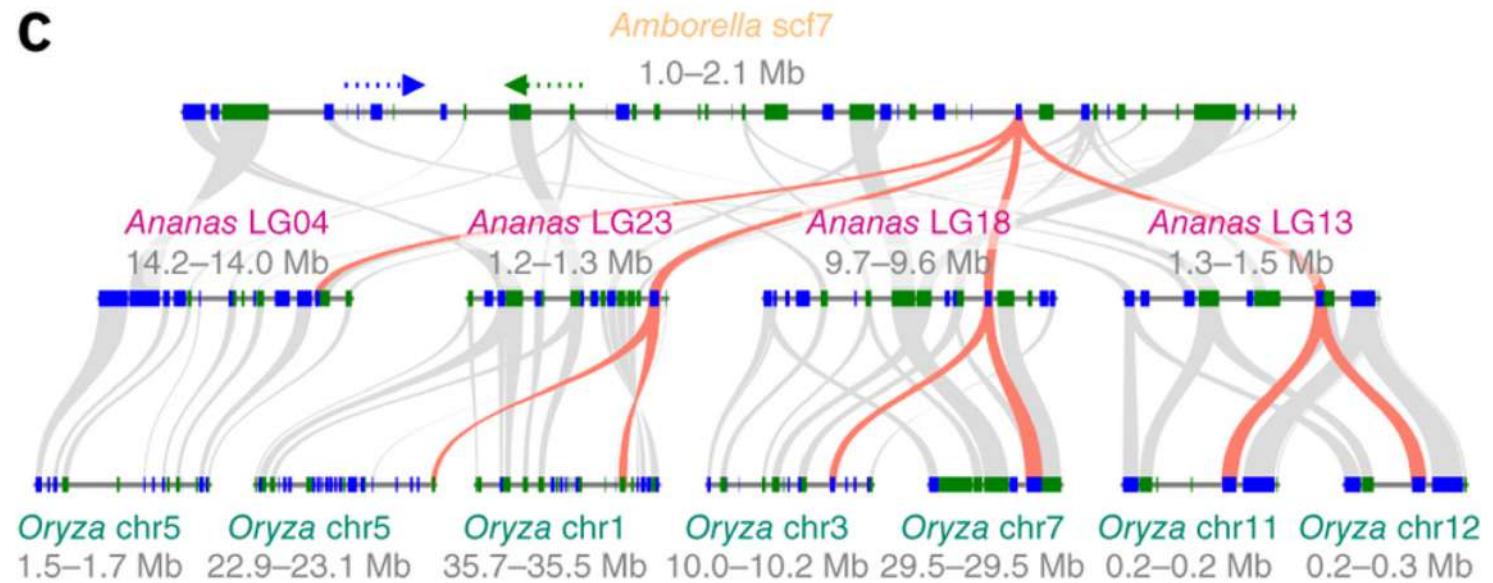
a



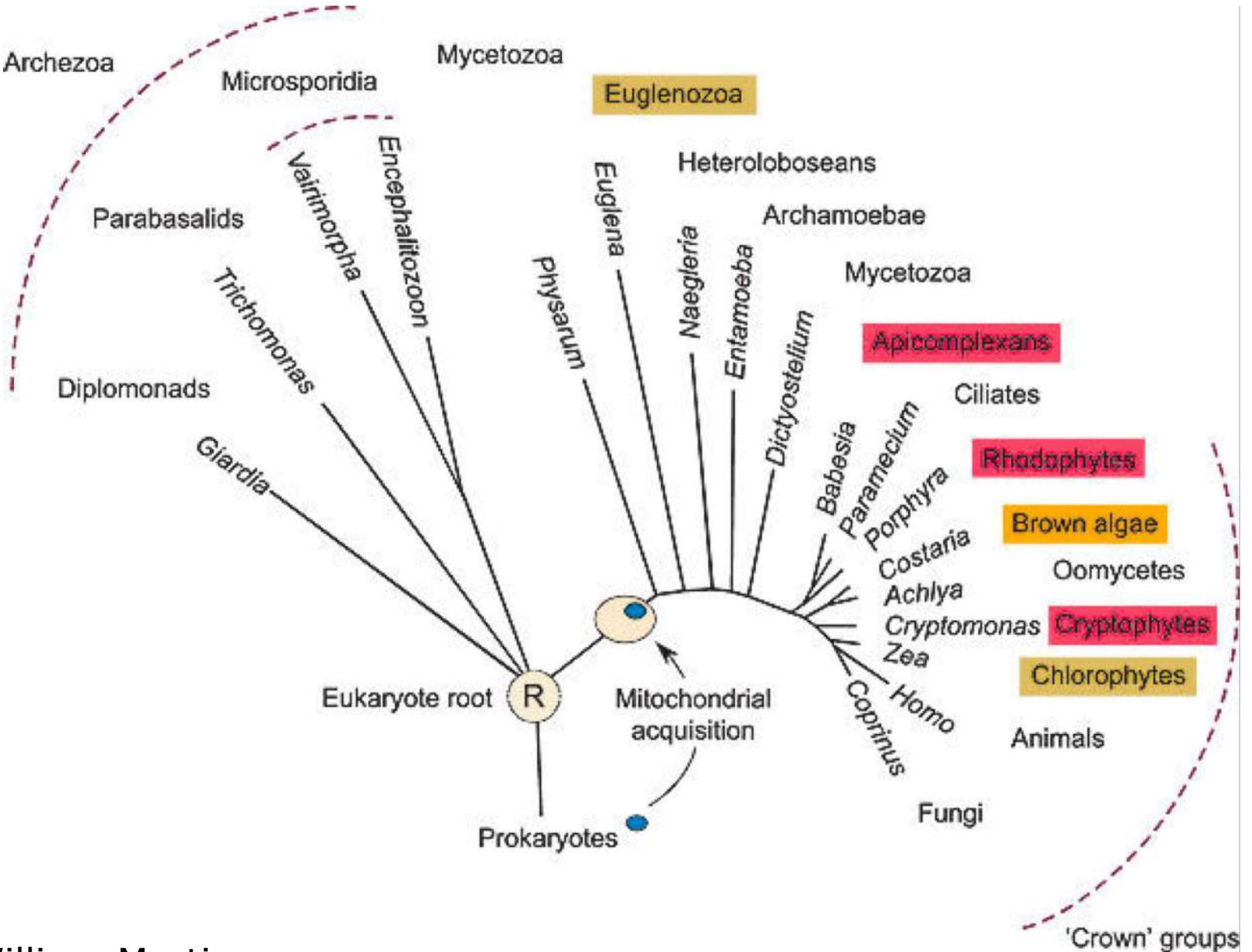
b



c

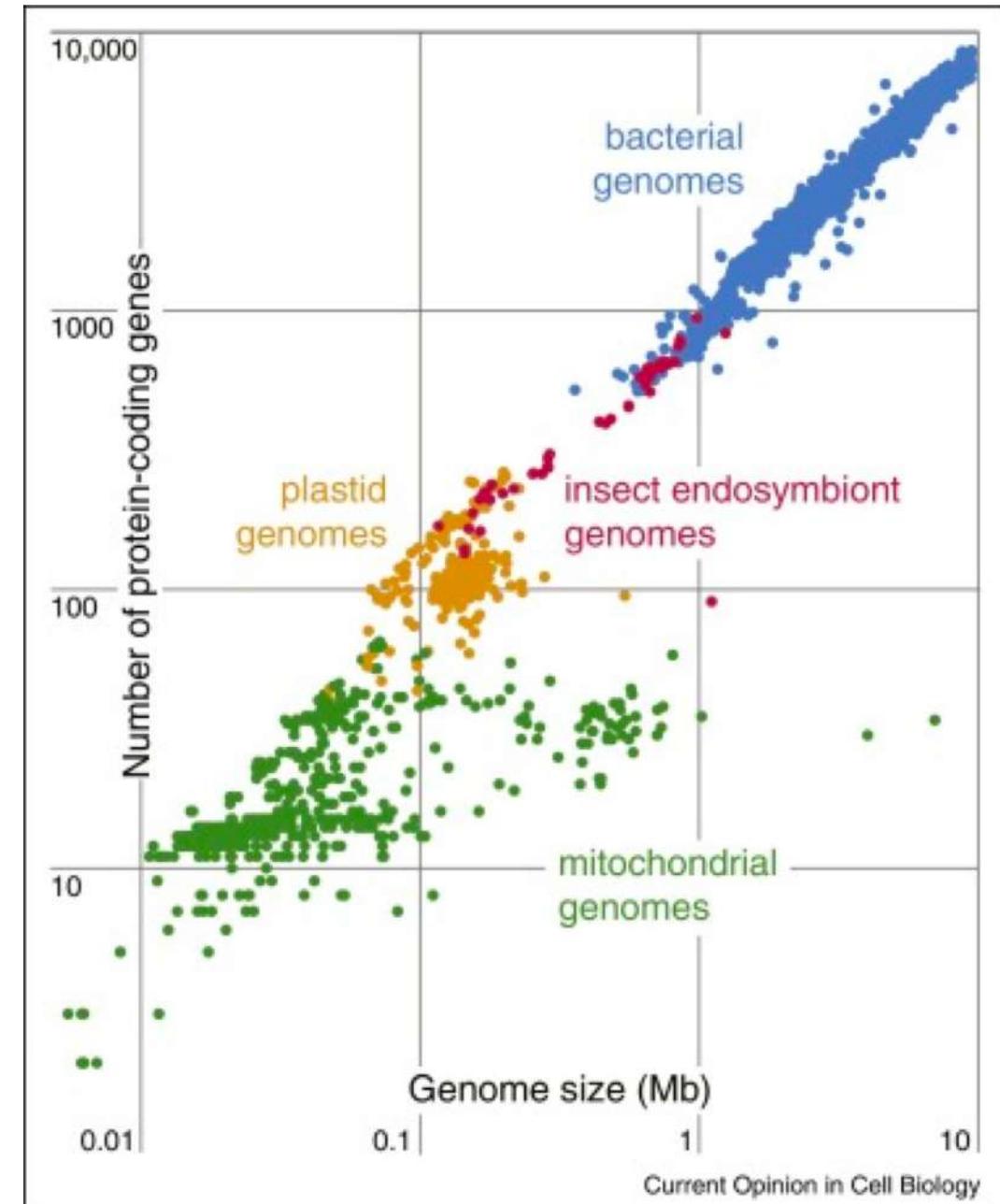


Symbiosis



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.

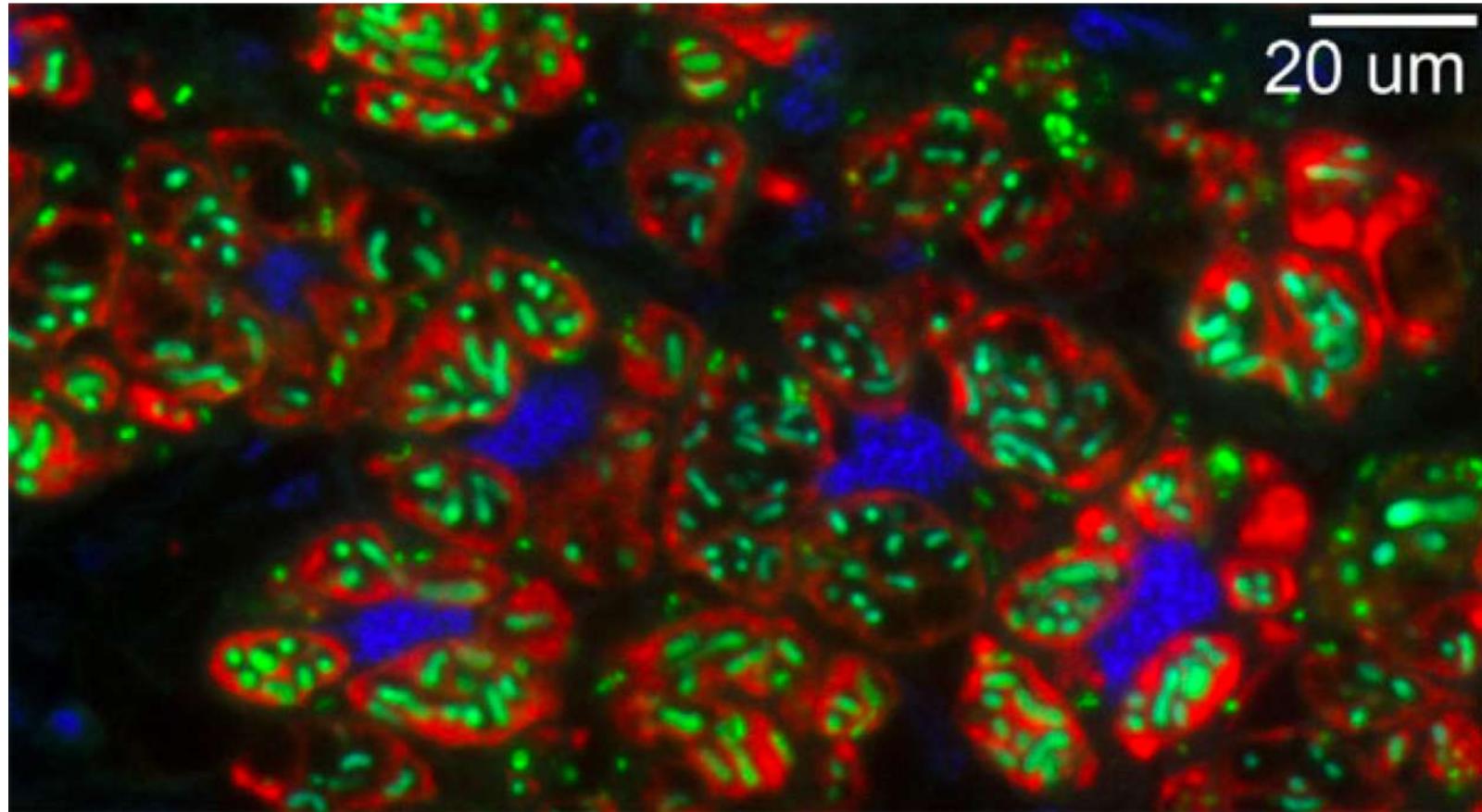
“Insect endosymbionts are missing (genomic) links between bacteria and organelles. It is now widely appreciated that all animals form symbioses with bacteria. Insects are especially interesting in this regard because they form many intracellular symbioses — that is, they allow bacteria to live inside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



Case study: Mealybugs

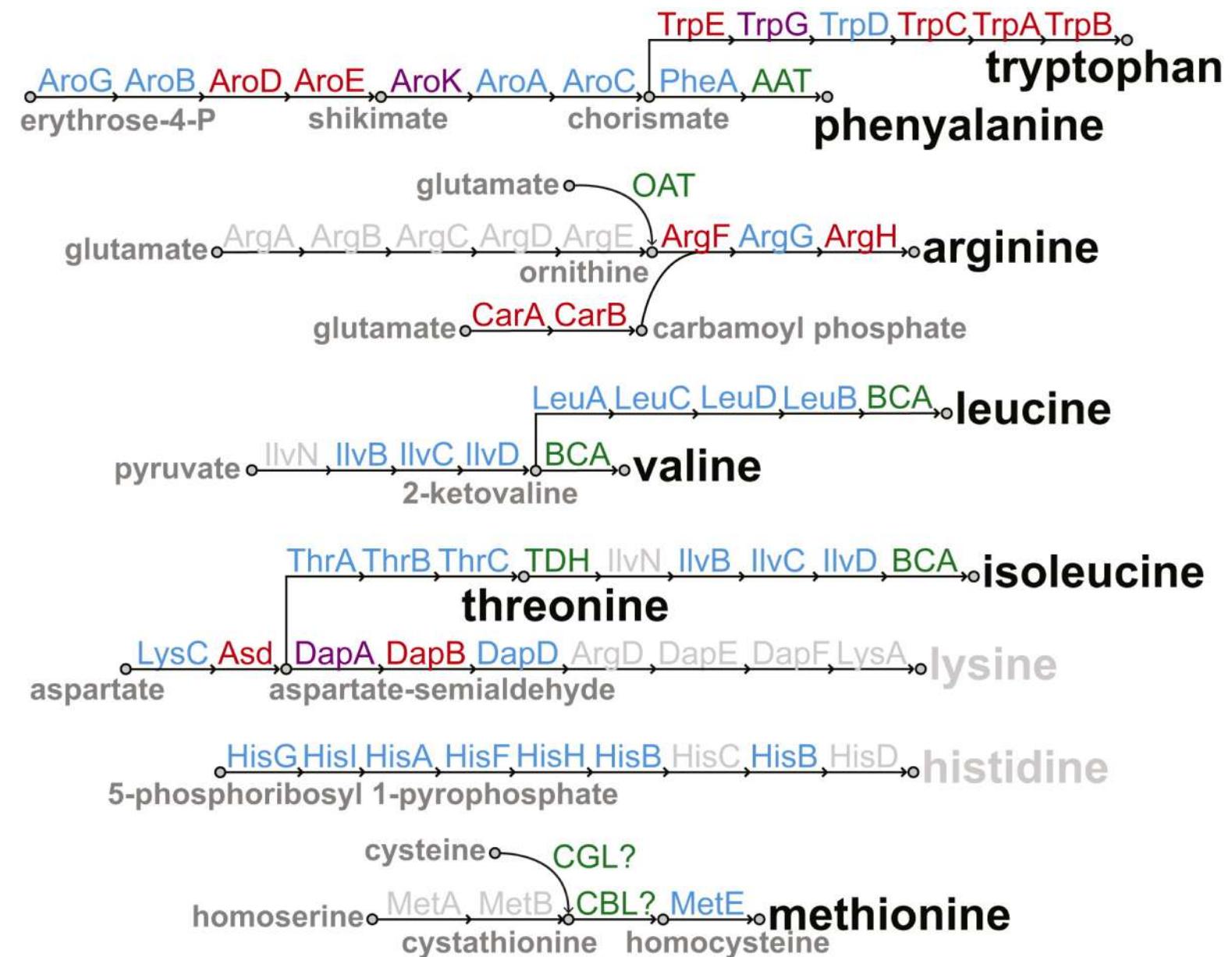


Triple Symbiotic Relationship between Mealybugs, *Tremblaya princeps*, and *Moranella endobia*



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

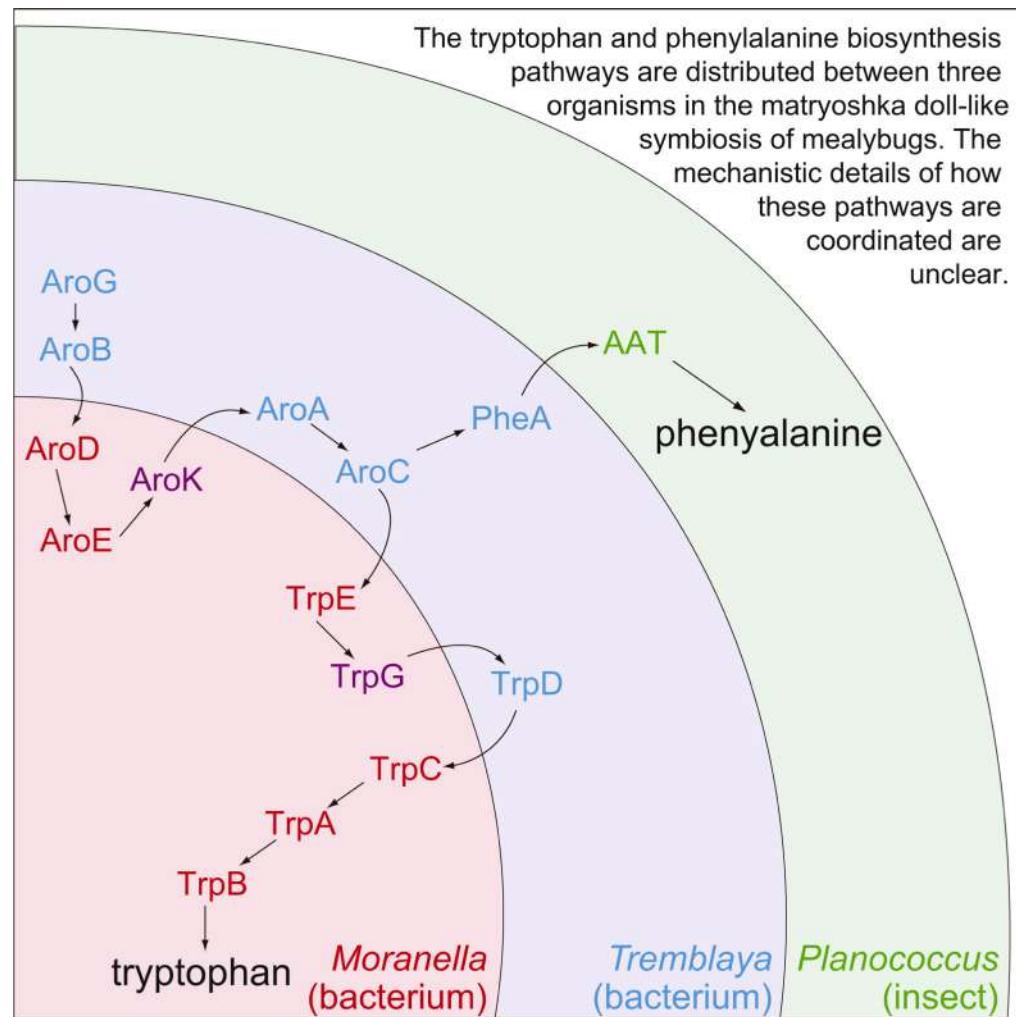
Predicted Essential Amino Acid Metabolic Contributions of the Mealybug-Tremblaya-Moranella Symbiosis



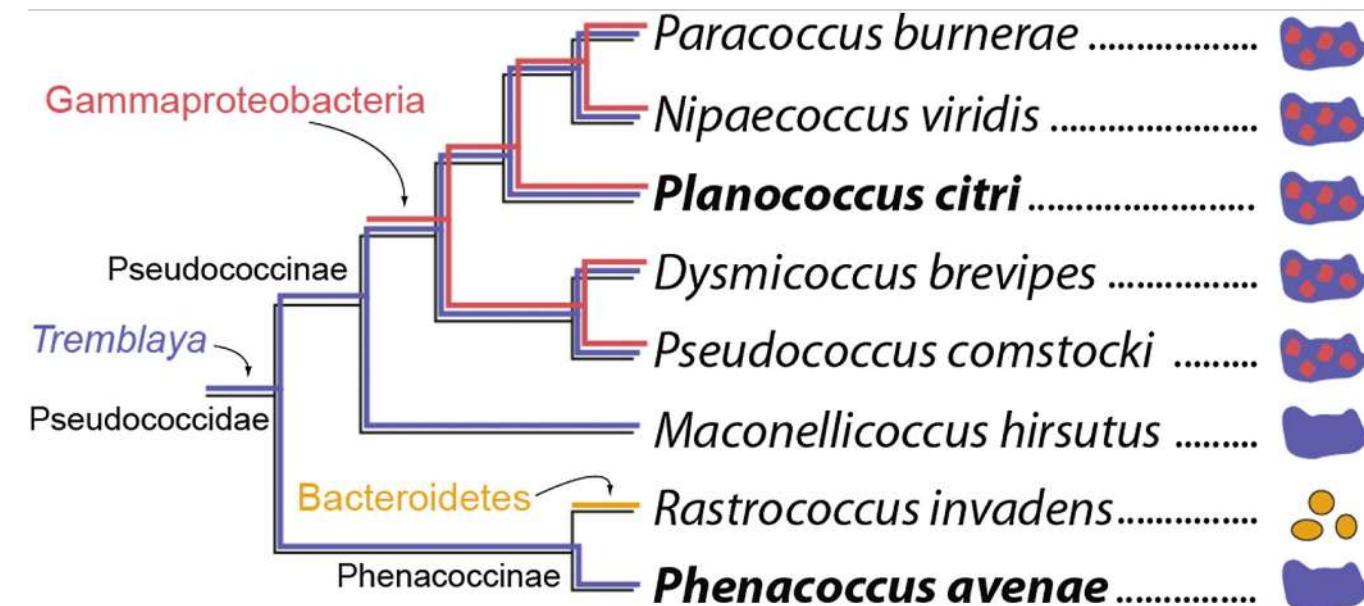
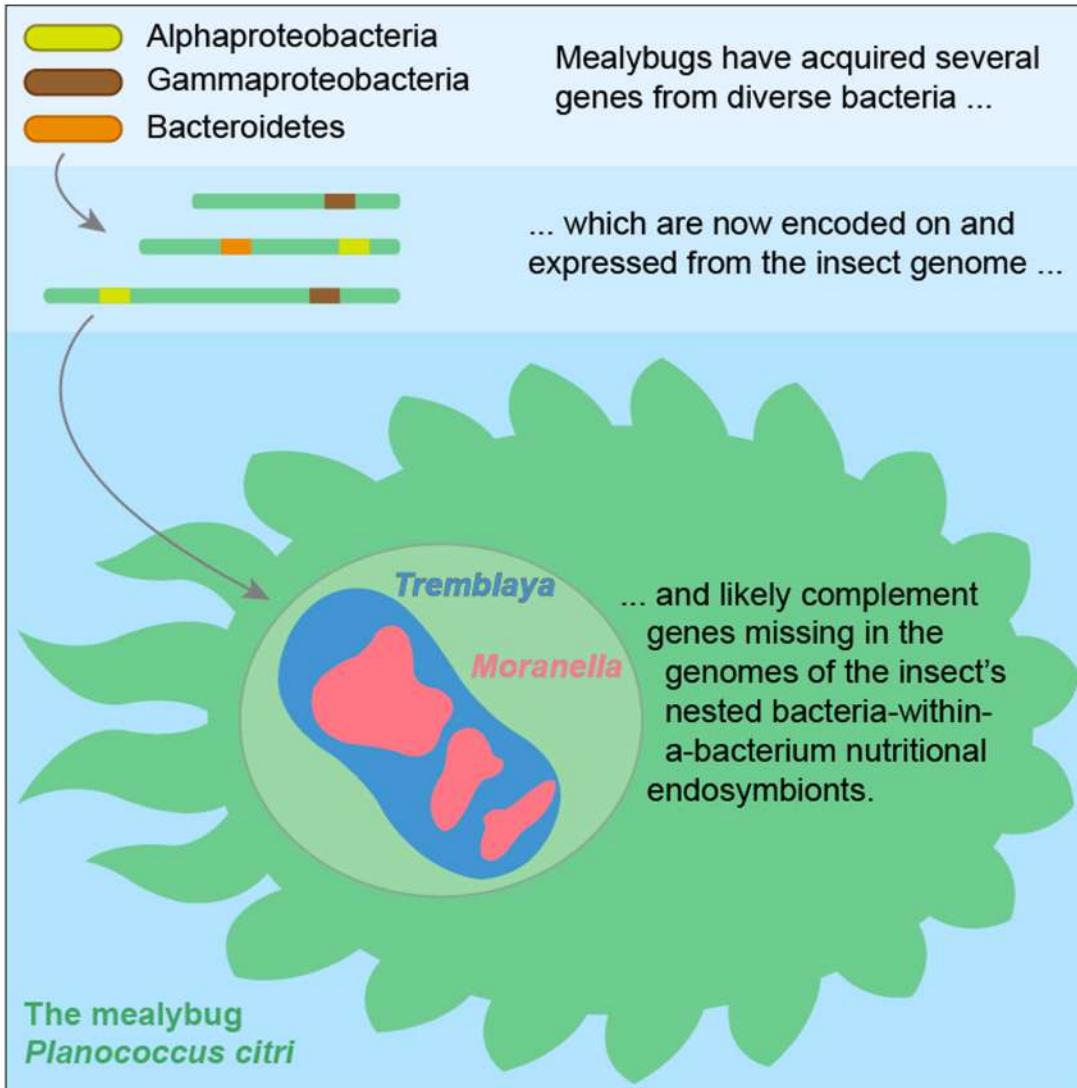
Gene homologs found in the Tremblaya genome are blue; the Moranella genome, red; both the Tremblaya and Moranella genomes, purple; neither the Tremblaya nor the Moranella genome, gray; activities not found in either bacterial genome but predicted to be encoded in the mealybug genome, green.

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

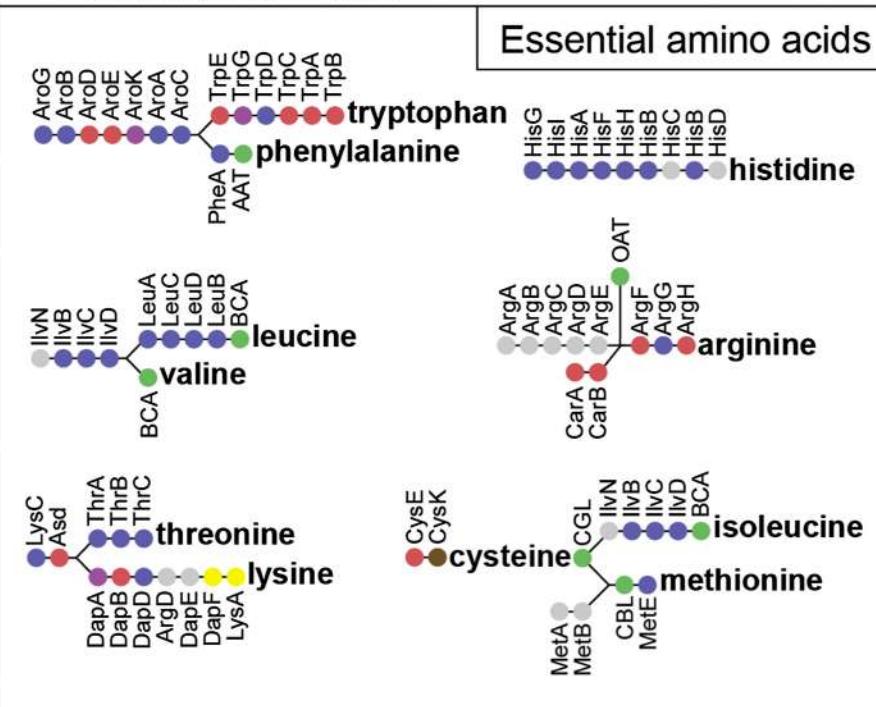
- HGT from diverse bacteria to the insect host genome support the three-way symbiosis
- Endosymbiont genomes can massively degrade without transfer of genes to the host



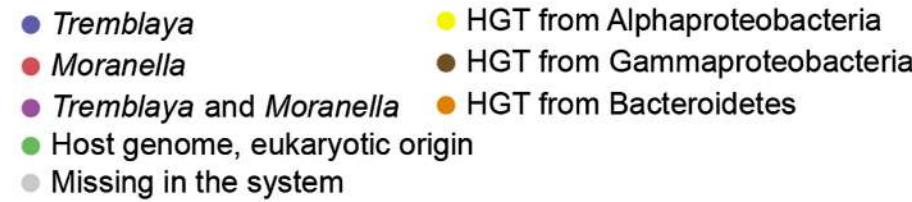
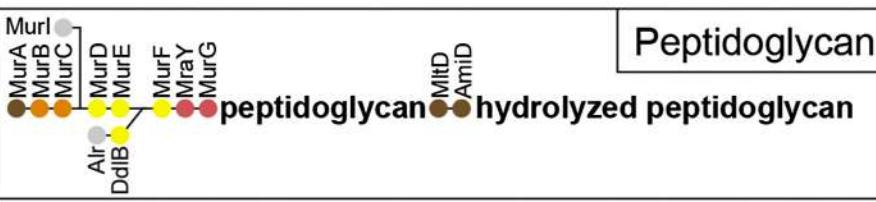
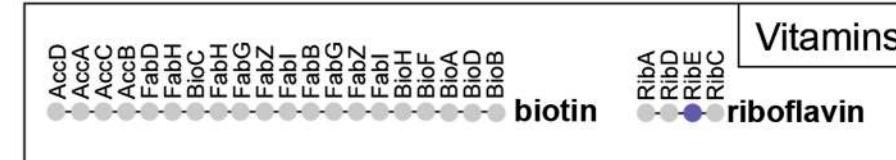
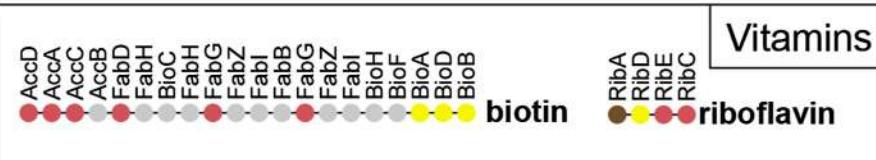
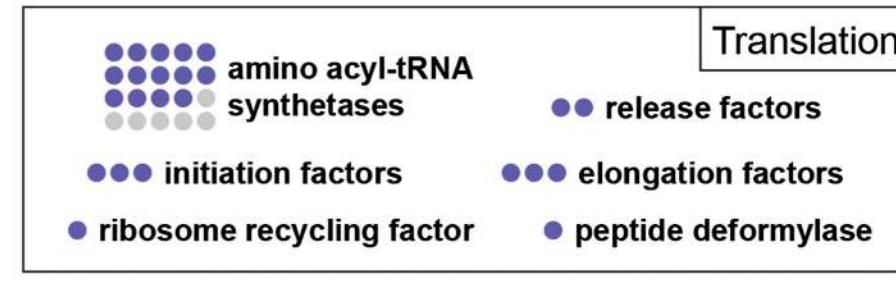
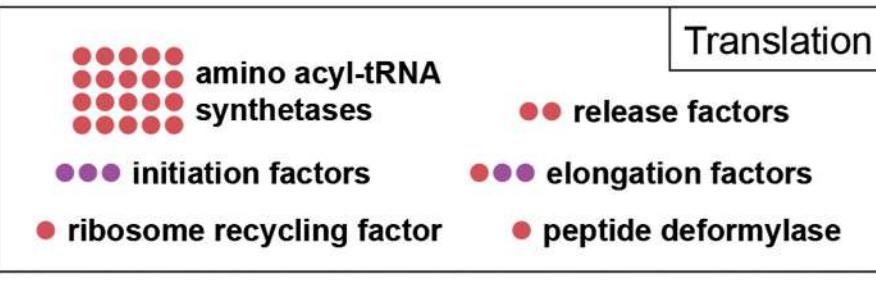
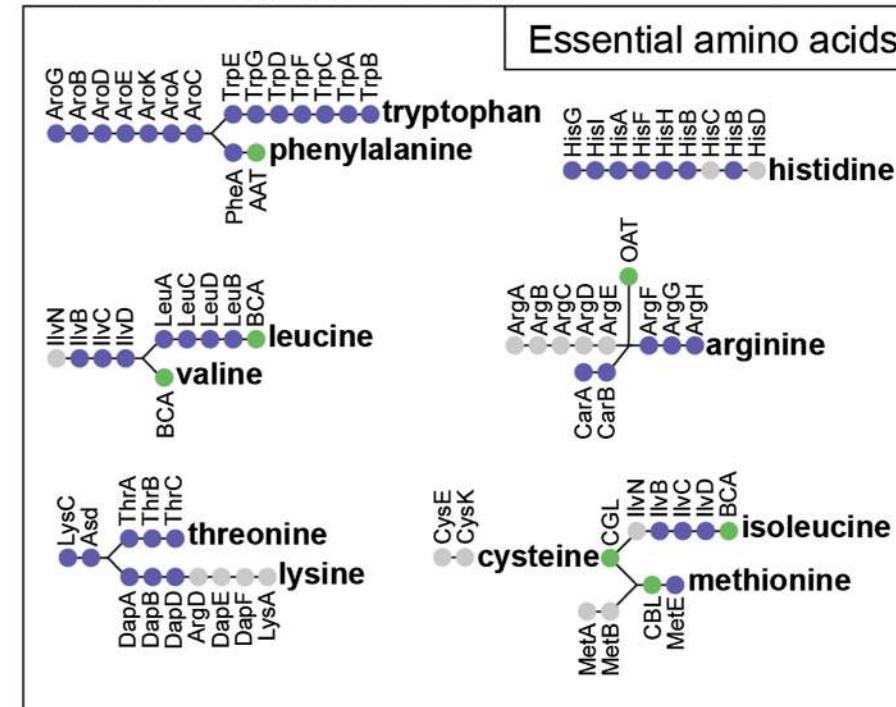
Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis



A *Planococcus citri*



B *Phenacoccus avenae*



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,*}

¹Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA

²Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA

³Canadian 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>

Some cicadas contain two bacterial symbionts, *Sulcia* and *Hodgkinia*.



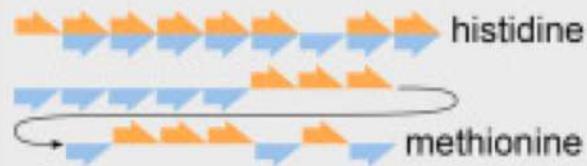
Other cicadas contain three symbionts: *Sulcia* and two versions of *Hodgkinia*.



The two new *Hodgkinia* genotypes arose from an unusual speciation event.



The single *Hodgkinia* genome encodes genes needed for the production of histidine and methionine.



The new *Hodgkinia* genotypes partition these pathways, requiring both species for the production of histidine and methionine.

Comparing genomes (at gene level)

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:

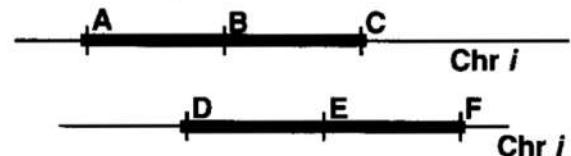
One *conserved linkage* involving genes B,C,D;
One *conserved linkage* involving genes E,F.
One *disrupted linkage* involving genes B,C,D vs E,F vs A.
One *conserved synteny* involving genes A,B,C,D,E,F.

Possible causes:

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

Conserved synteny, disrupted synteny, conserved linkage, disrupted linkage

Gene arrangement:



Count:

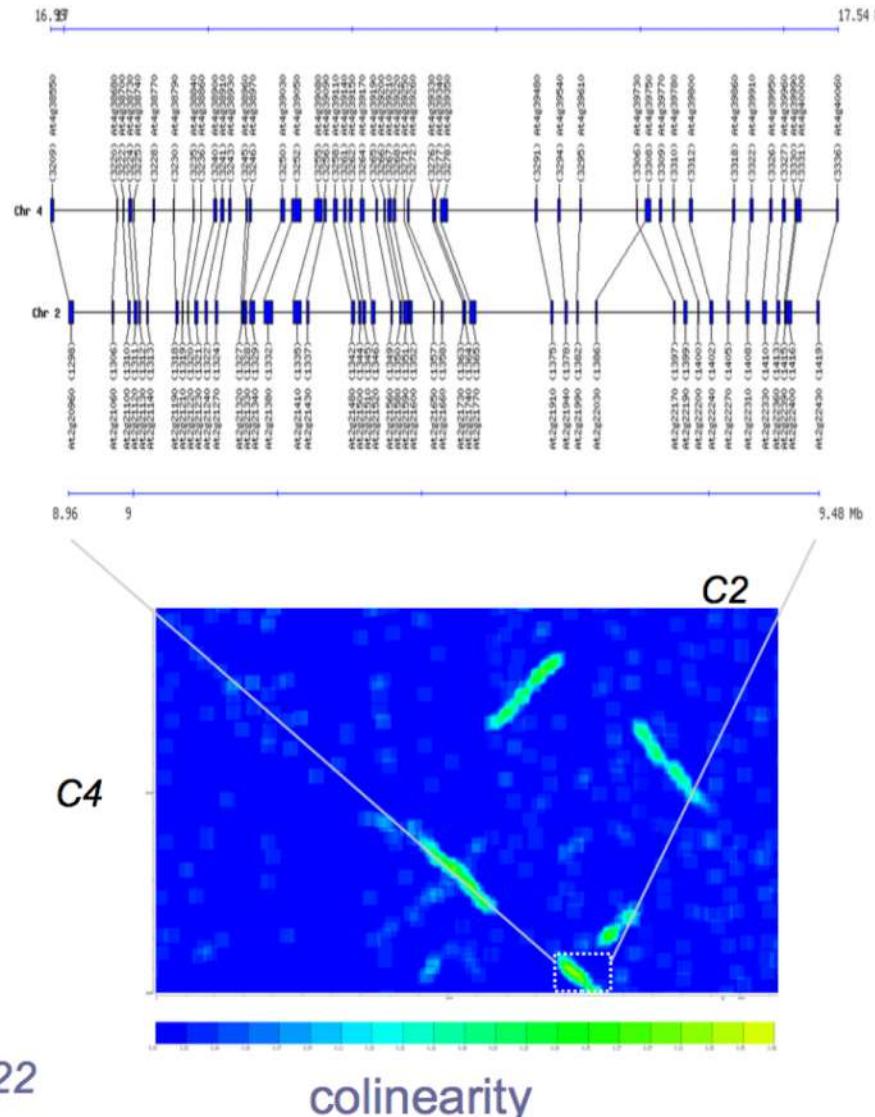
One *conserved linkage* involving genes A,B,C;
One *conserved linkage* involving genes D,E,F.
One *disrupted linkage* involving genes A,B,C vs D,E,F .
One *conserved synteny* involving genes A,B,C.
One *conserved synteny* involving genes D,E,F.
One *disrupted synteny* involving genes A,B,C vs D,E,F.

Possible causes:

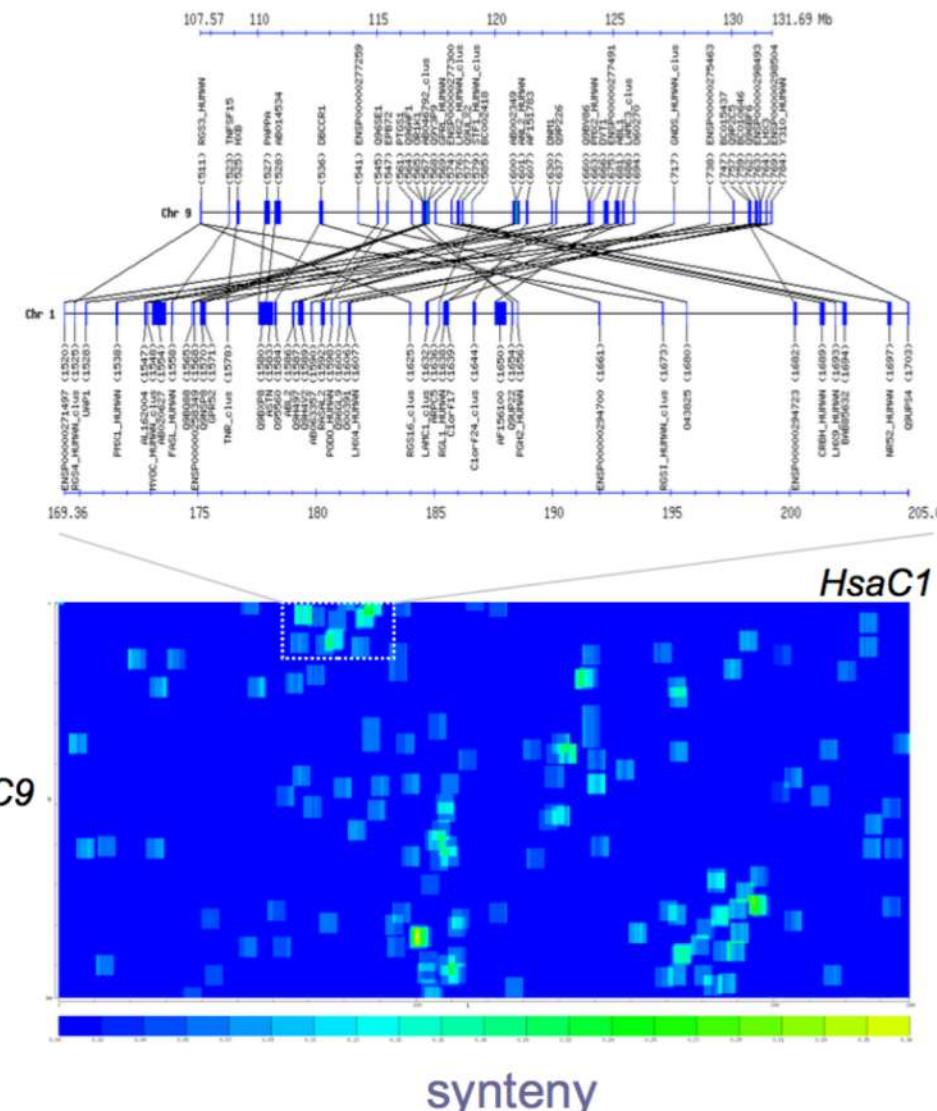
An inter-chromosomal rearrangement,
such as a reciprocal translocation.

Synteny and colinearity

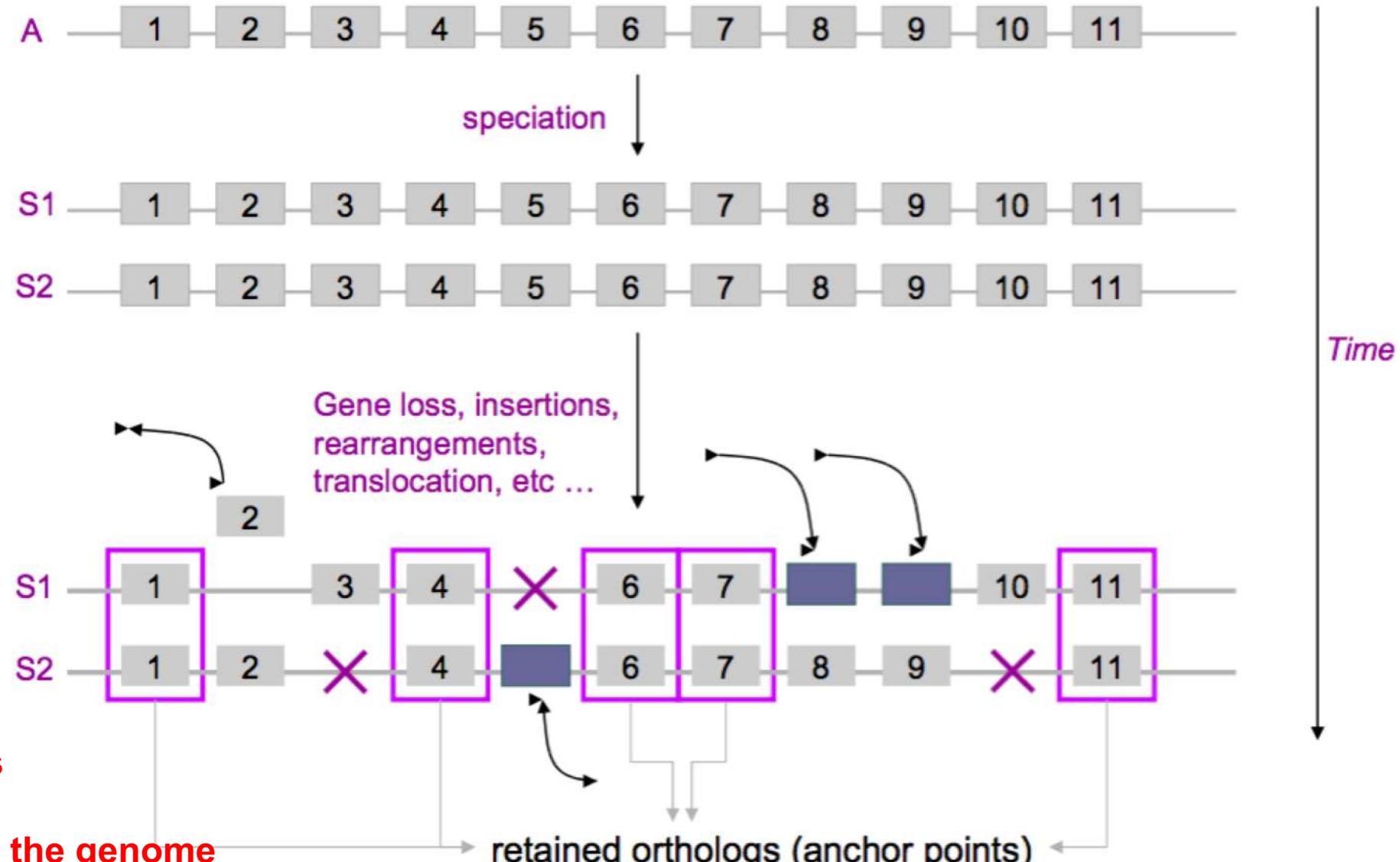
recent duplication



ancient duplication



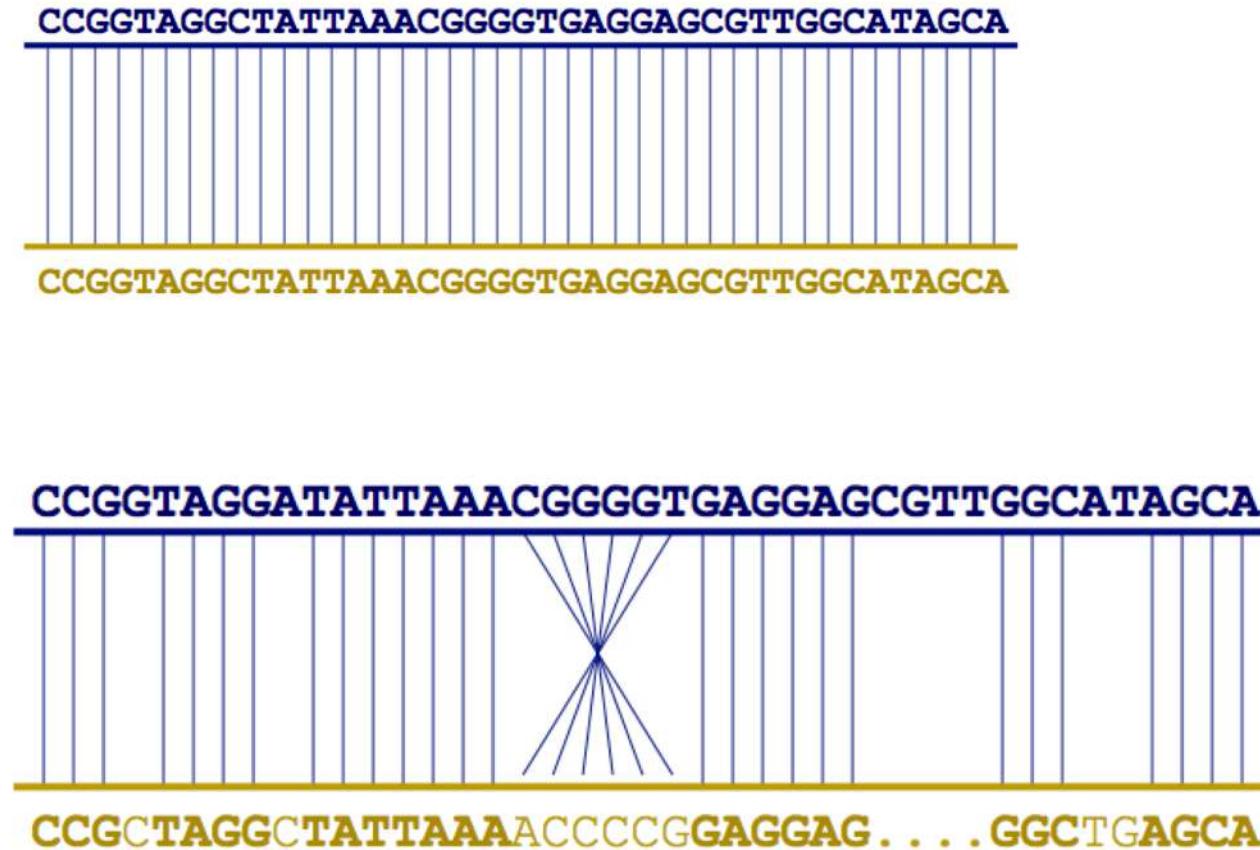
Inferring gene collinearity



Whole genome alignment

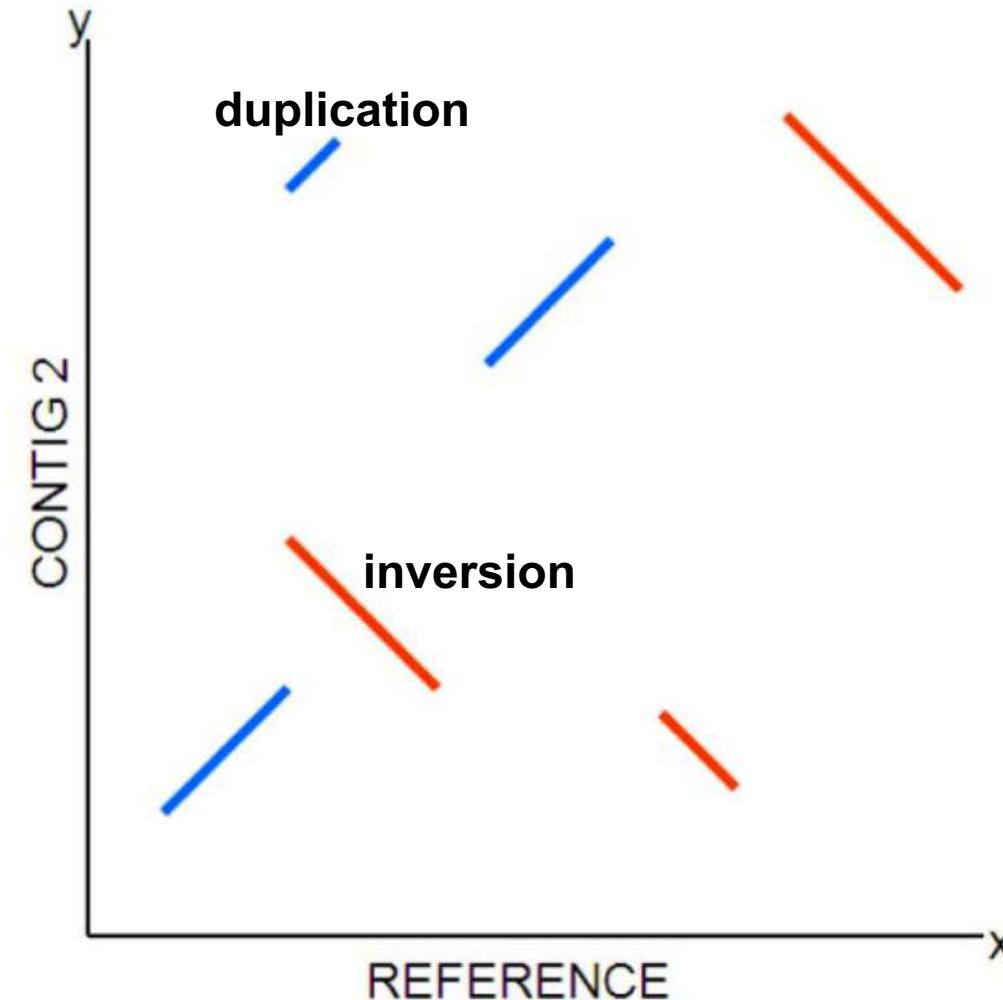
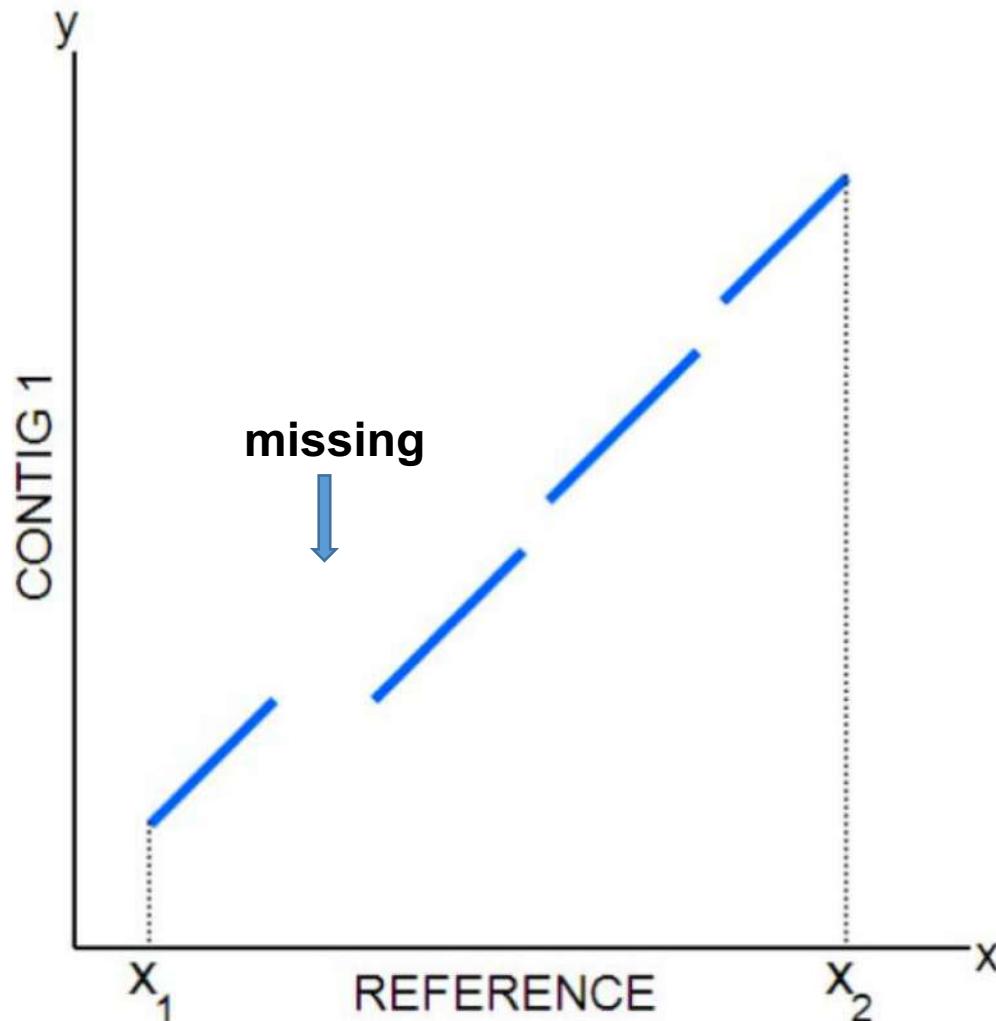
For two genomes, A and B,
find a mapping from each
position in A to its
corresponding position in B

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



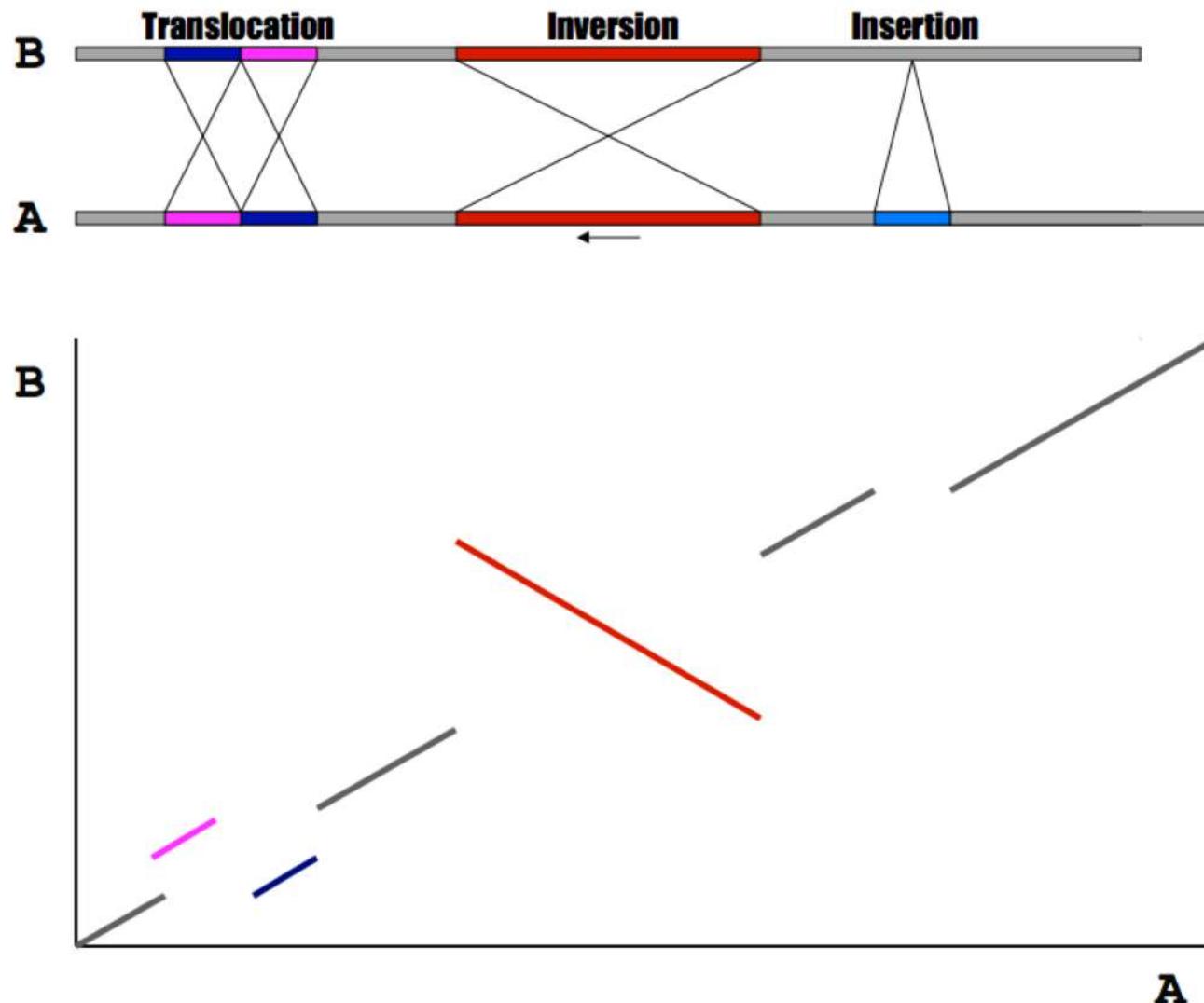
Aligning genome at nucleotide / amino acid level

Visualise through dotplot



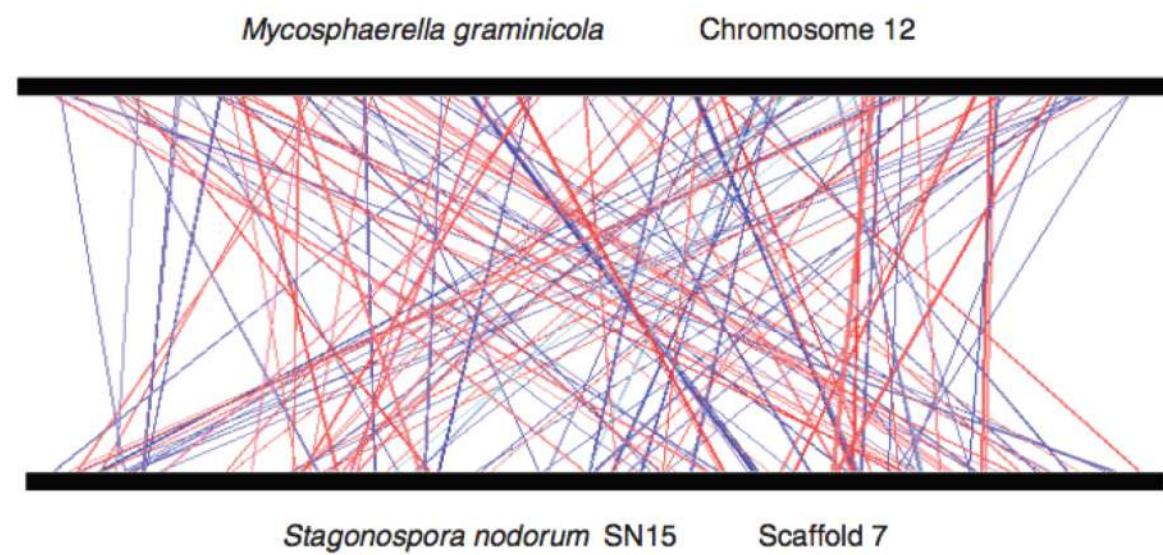
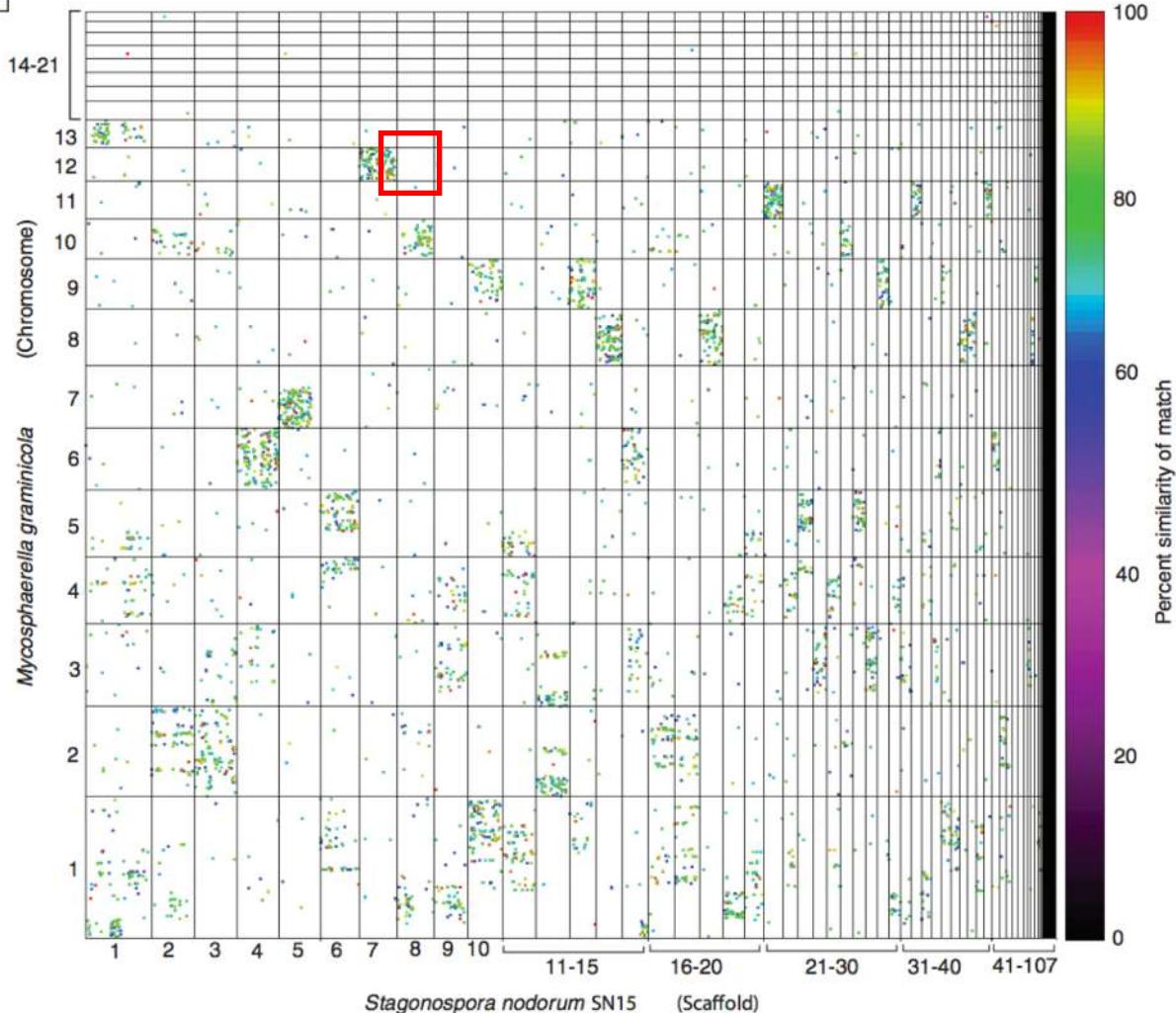
Aligning genome at nucleotide / amino acid level

Visualise through dotplot

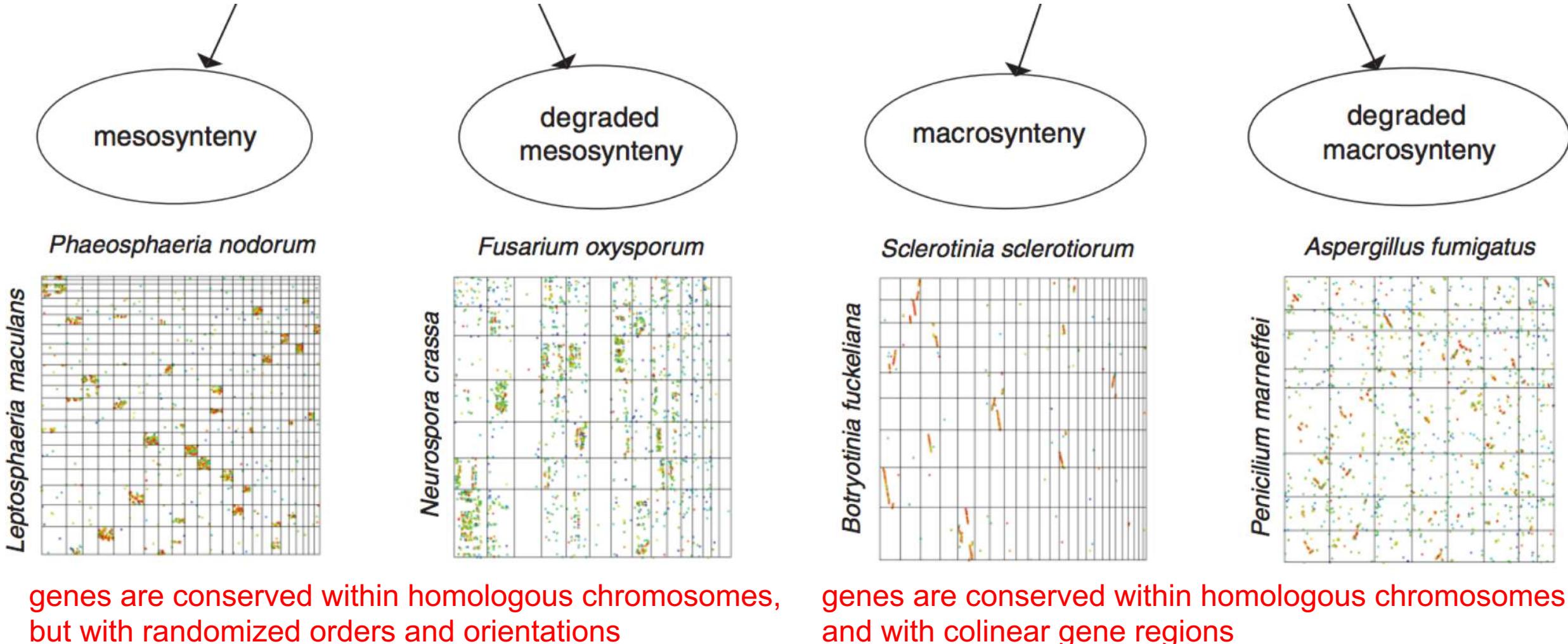


Relationship between genome synteny, syntenic orthologs and duplications

(a)

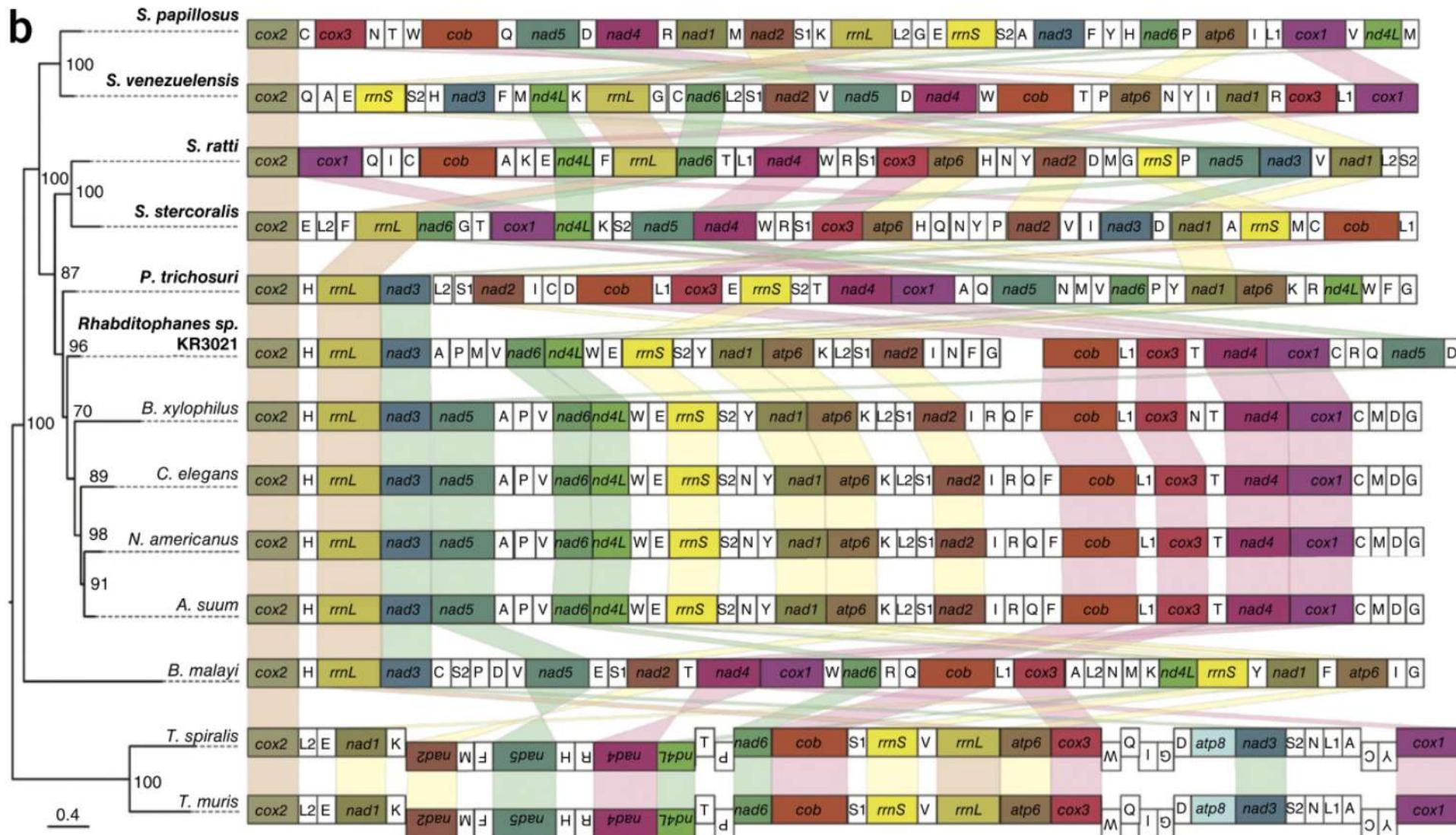


Different kinds of genome synteny



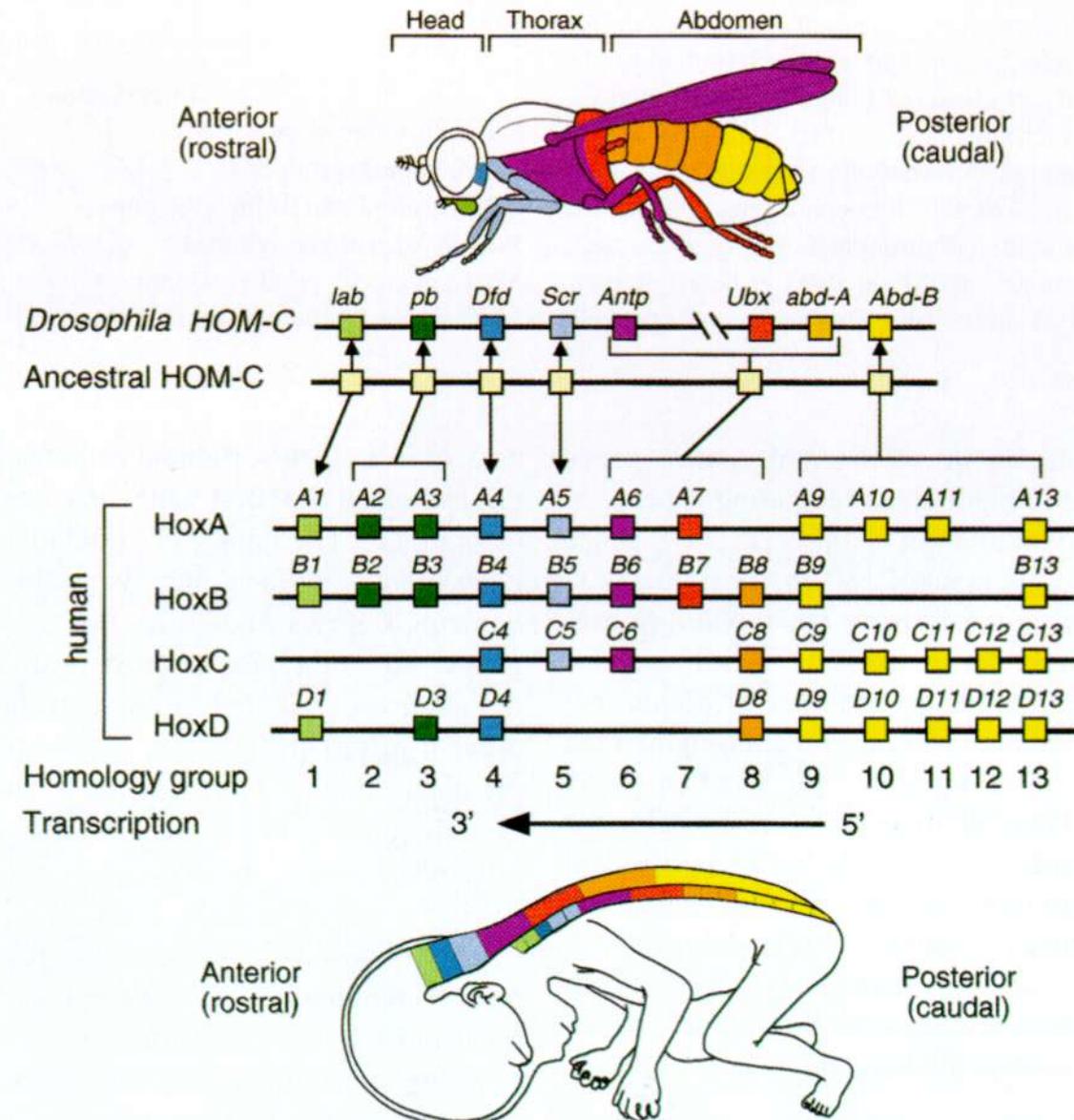
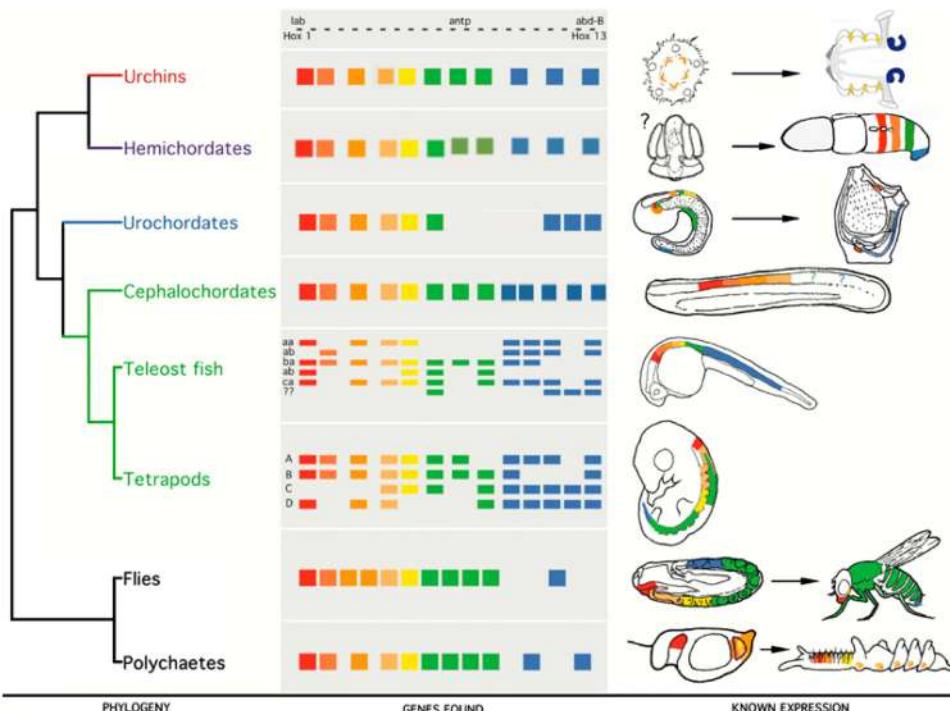
Why are we interested in synteny and collinearity?

Establish relationship between species



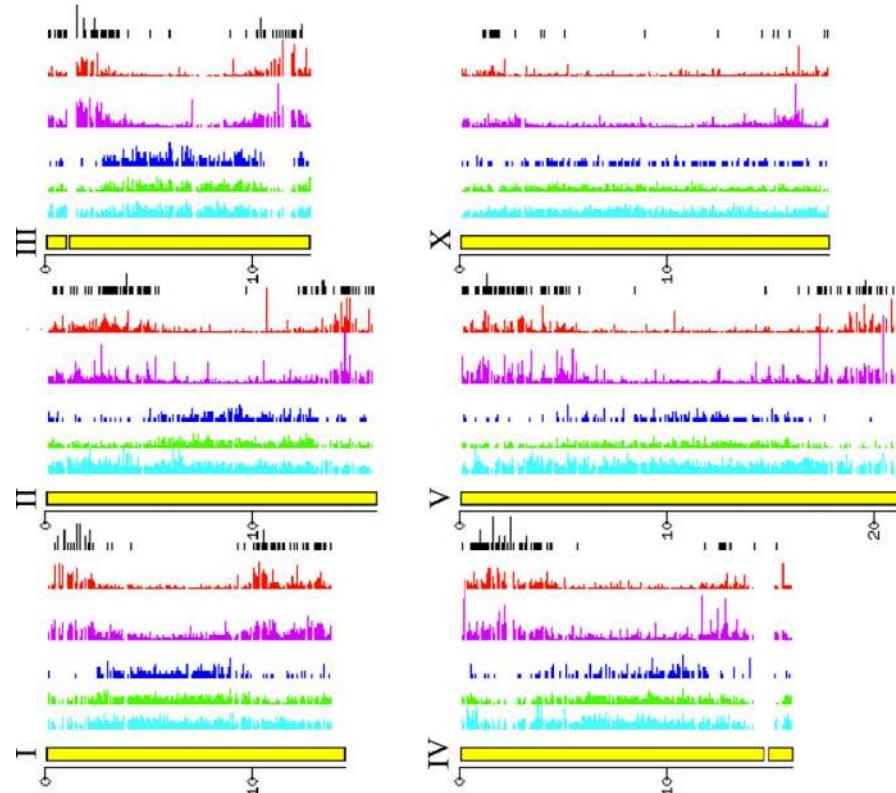
Why are we interested in synteny and collinearity?

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

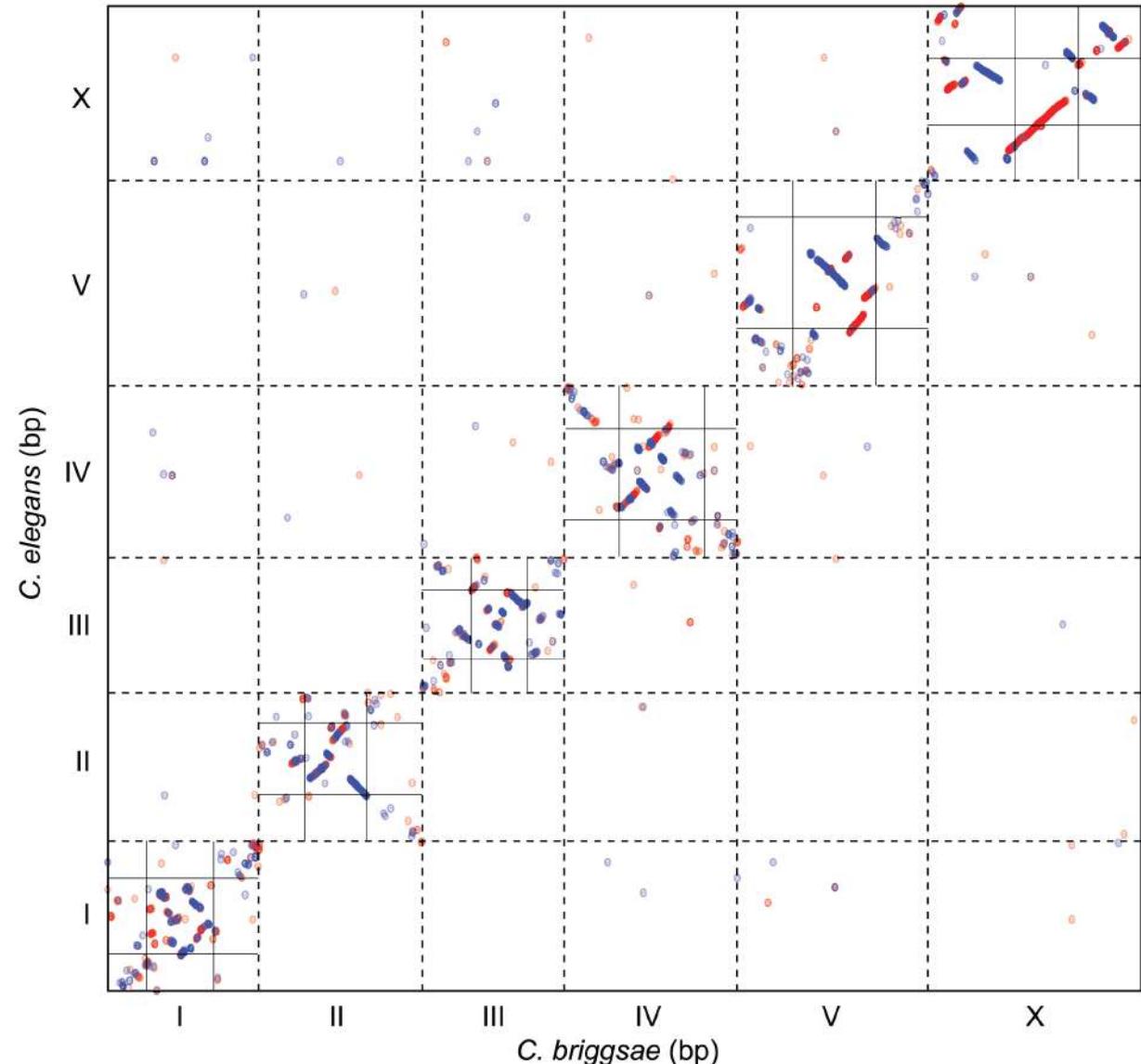


Why are we interested in synteny and collinearity?

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



Yeast similarities: Inverted repeats: Tandem repeats: TTAGGC repeats:
Sequence: Predicted genes: EST matches:



Stein et al., PLOS Biology 2003

The *C. elegans* Sequencing Consortium Science 1998

MGC in fungi

Fungi contain a remarkable diversity of both primary and secondary metabolic pathways involved in ecologically specialized or accessory functions. Genes in these pathways are frequently physically linked on fungal chromosomes, forming metabolic gene clusters (MGCs)

a Standard configuration of fungal MGCs

Primary metabolism

Nitrate assimilation MGC
in *Ogataea polymorpha*



Secondary metabolism

Gliotoxin MGC in
Aspergillus fumigatus



MGC in fungi

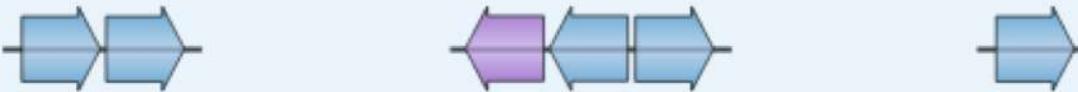
b Variation in gene content and structure among fungal MGCs

Primary metabolism

Galactose MGC and unclustered pathway genes in *Saccharomyces cerevisiae*



Biotin MGCs and unclustered pathway gene in *Saccharomyces cerevisiae*

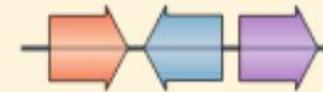


Secondary metabolism

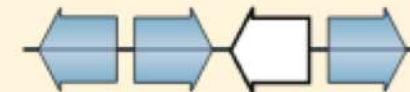
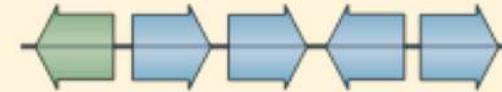
Emericellamide MGC in *Aspergillus nidulans*



Kojic acid MGC in *Aspergillus oryzae*



Dothistromin MGCs and unclustered gene in *Dothistroma septosporum*



The intertwined fumagillin (F) and pseurotin (P) super MGC in *Aspergillus fumigatus*

MGC in fungi

Galactose MGC

Saccharomyces cerevisiae



Torulaspora delbrueckii



Aspergillus nidulans



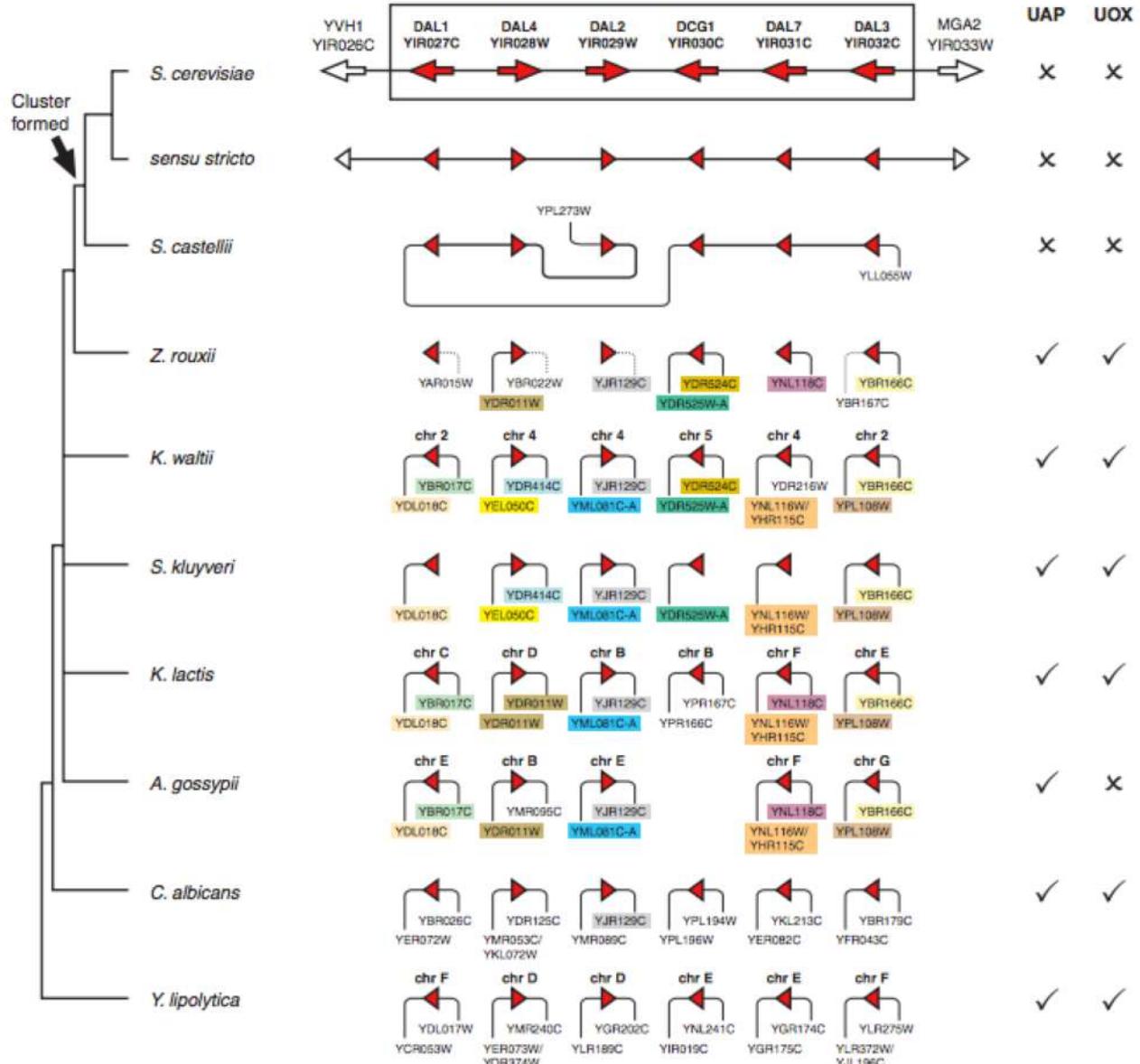
■ Enzyme ■ Transporter ■ Regulator ■ Non-MGC gene

Why are we interested in synteny and collinearity?

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

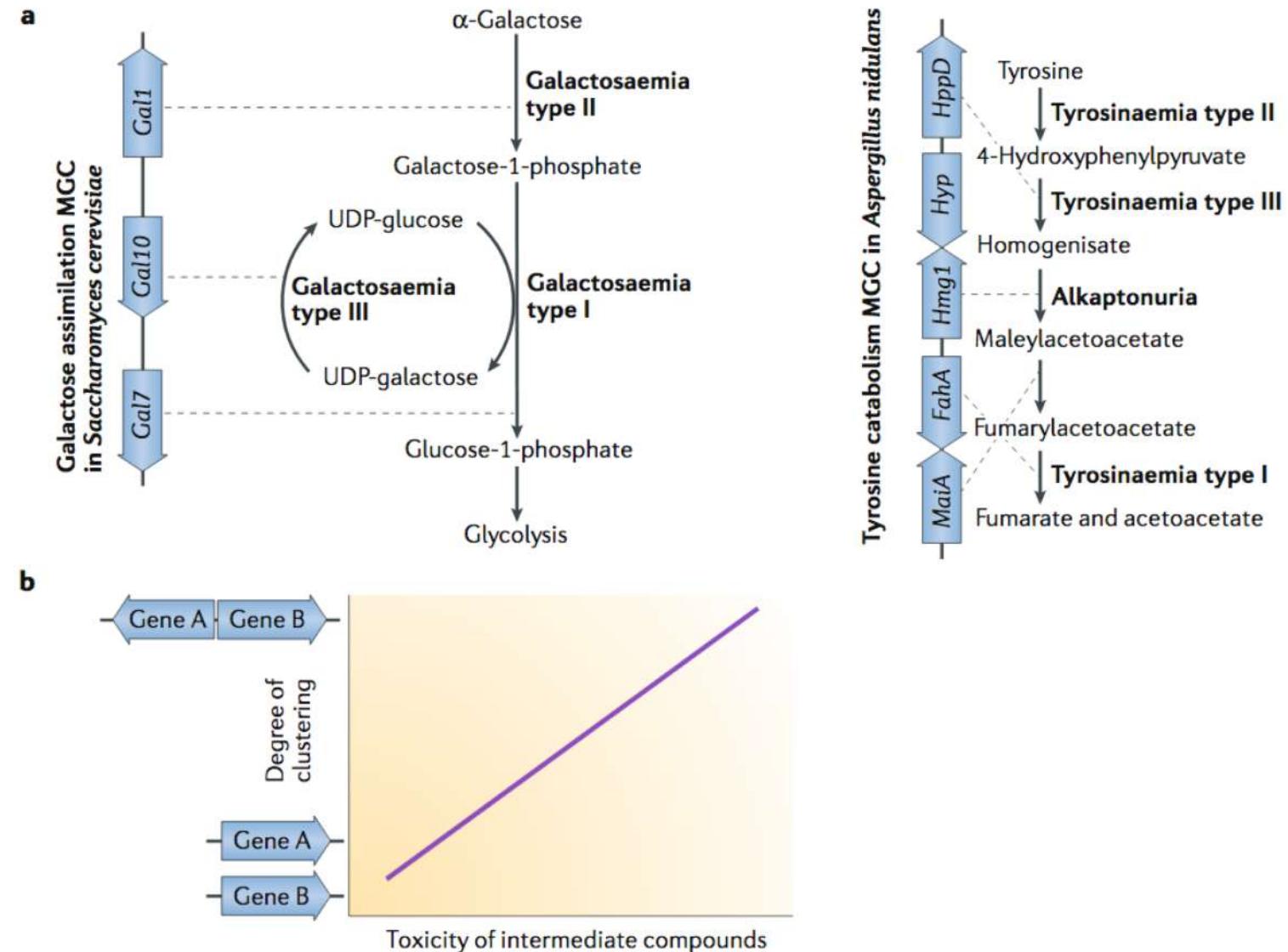
Birth of a metabolic gene cluster in yeast by adaptive gene relocation

Simon Wong & Kenneth H Wolfe



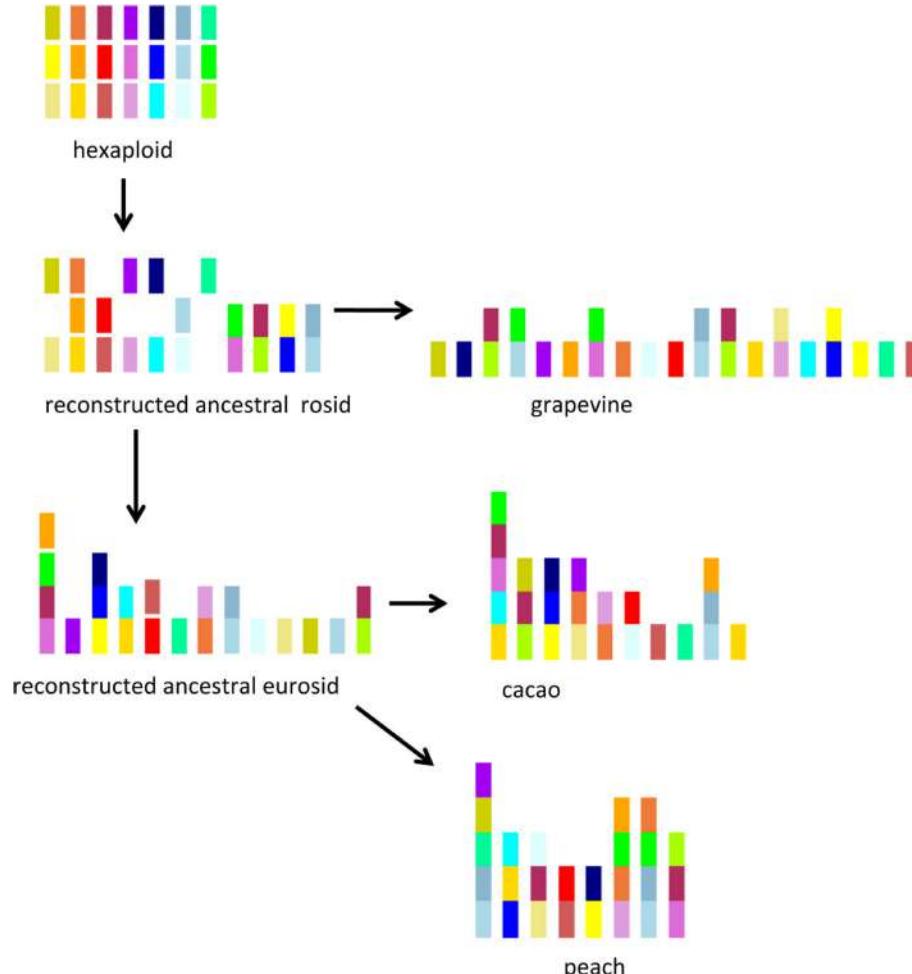
The toxicity avoidance hypothesis

A | The intermediate compounds of several fungal MGCs, such as those involved in galactose assimilation and tyrosine catabolism, are toxic. **b |** According to the toxicity avoidance hypothesis, MGCs are expected to be more common in pathways that generate very toxic metabolic intermediates.

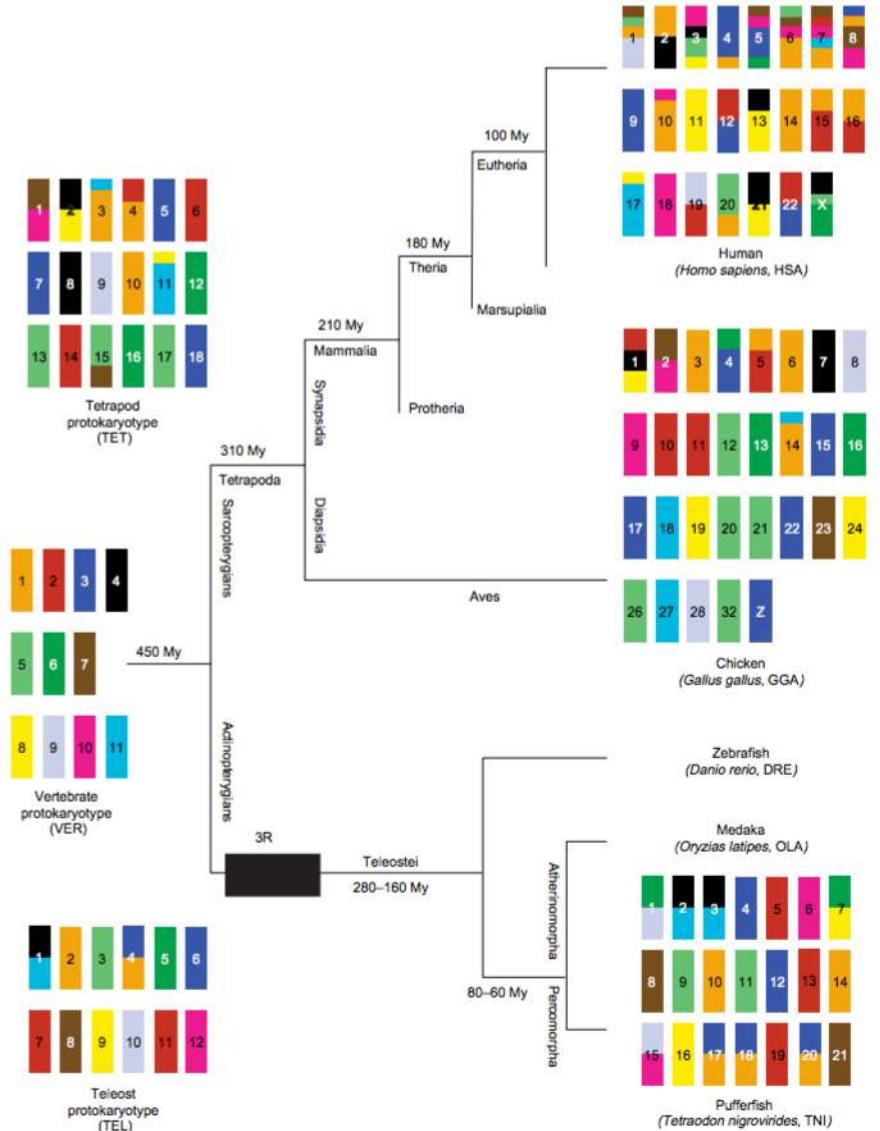


Why are we interested in synteny and collinearity?

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



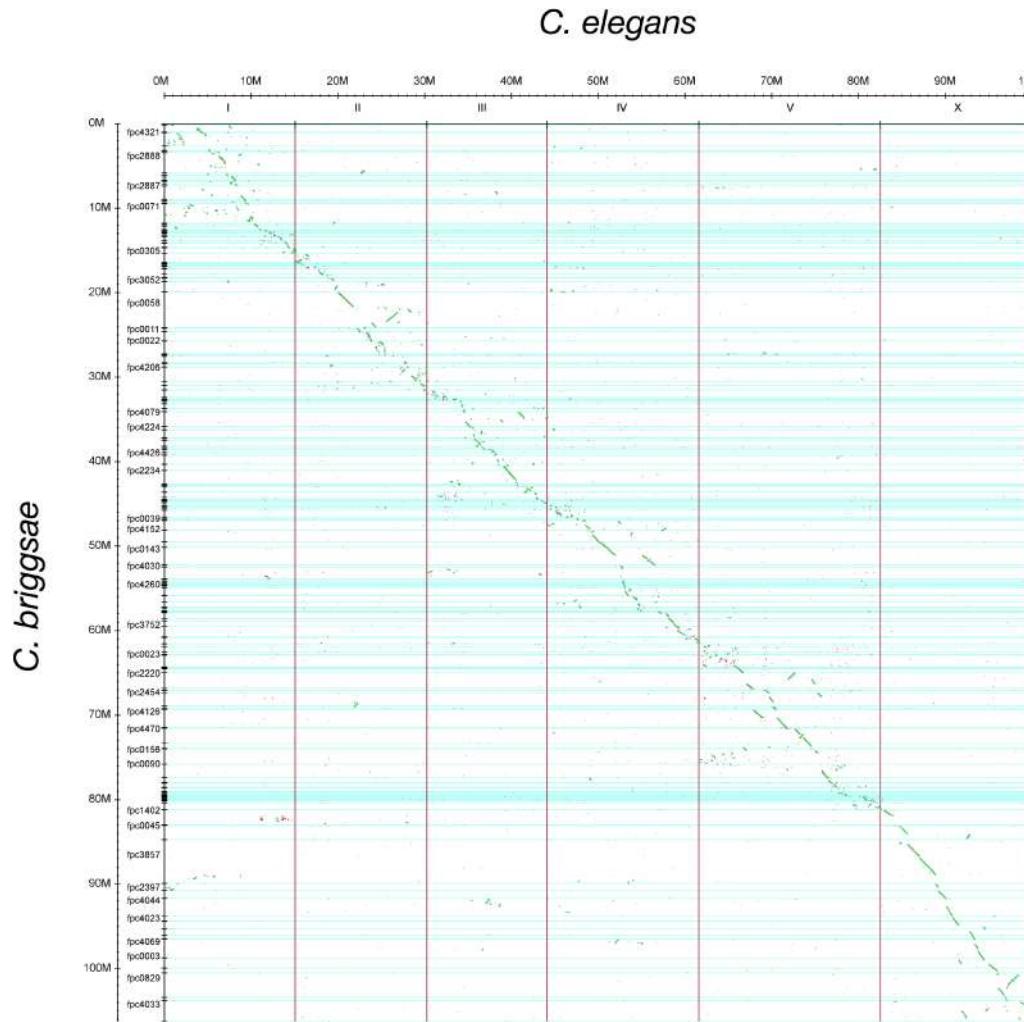
Zheng et al (2013)



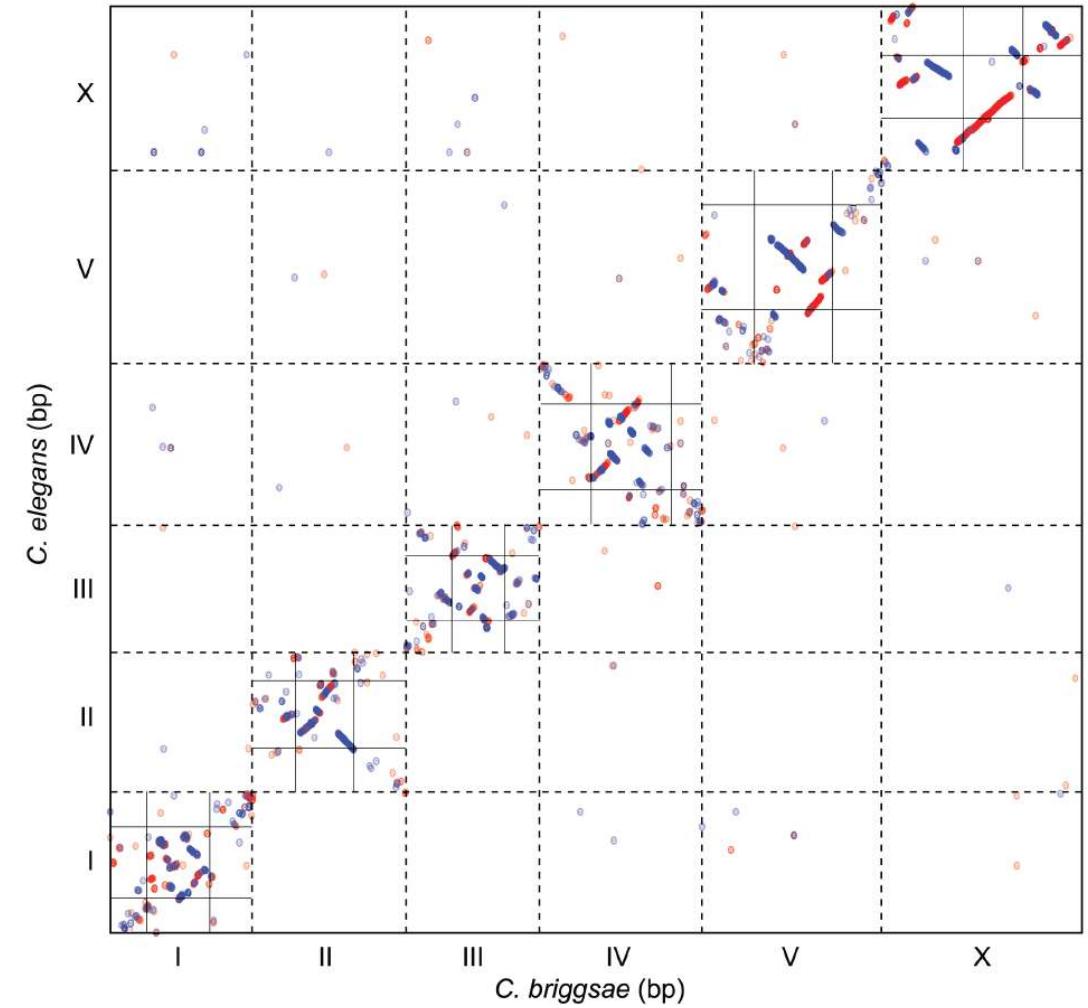
Kohn et al (2006)

Some caveats

Assembly quality likely to influence synteny observation



Stein et al., PLOS Genetics (2003)



Ross et al., PLOS Genetics (2011)

Synteny based scaffolding: use with caution

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



METHOD

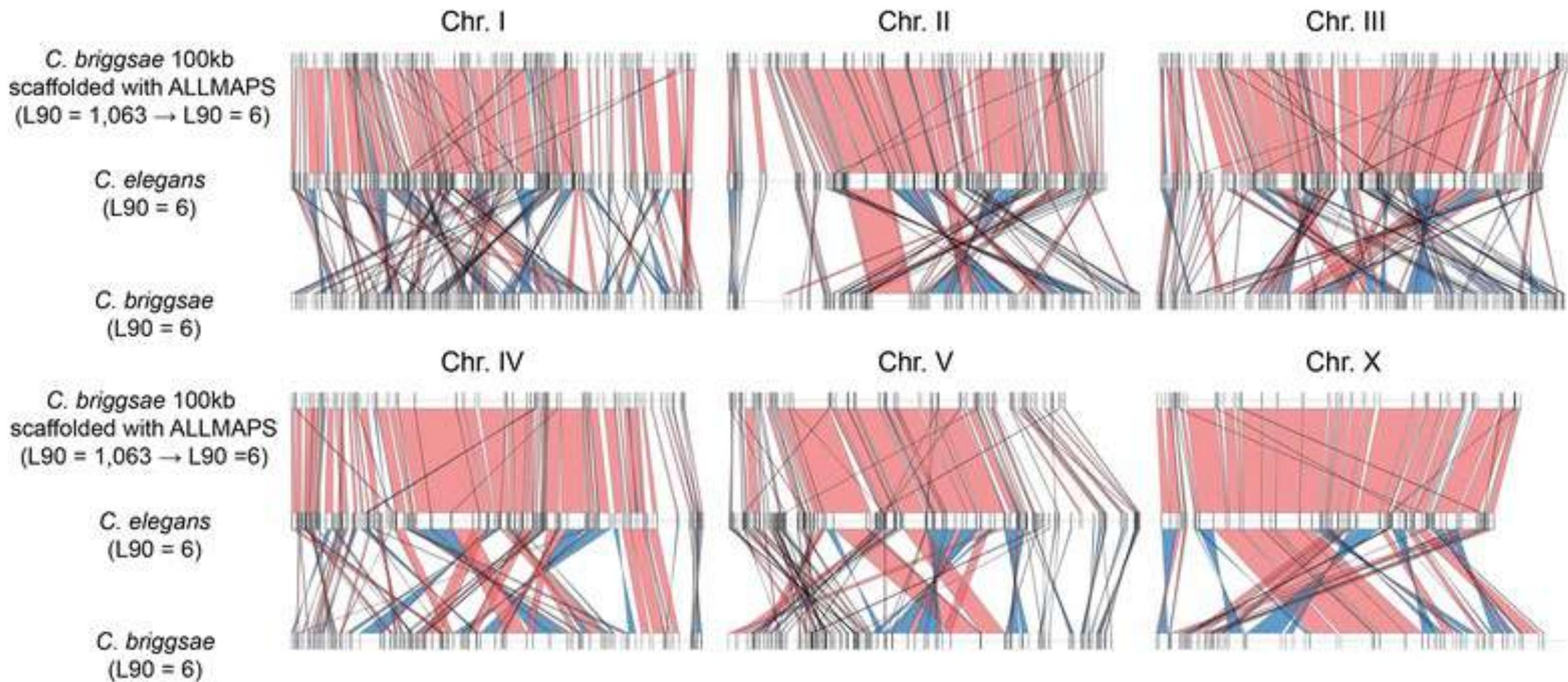
Open Access

ALLMAPS: robust scaffold ordering based on multiple maps

Haibao Tang^{1,2,3*}, Xingtian Zhang⁴, Chenyong Miao¹, Jisen Zhang¹, Ray Ming¹, James C Schnable^{3,5},
Patrick S Schnable^{3,6}, Eric Lyons² and Jianguo Lu⁷

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?**

Synteny based scaffolding: use with caution



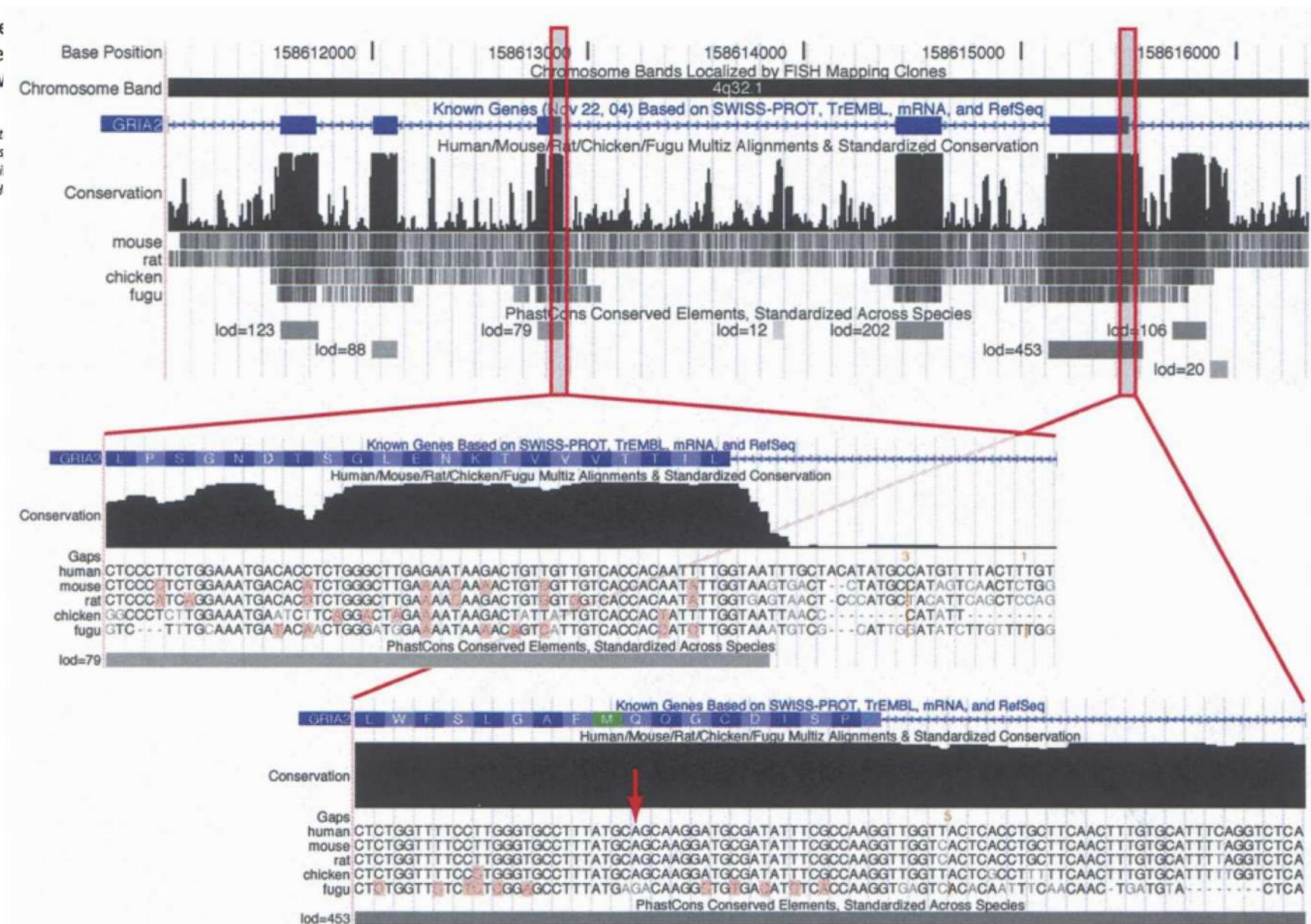
Comparing genomes beyond gene level

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

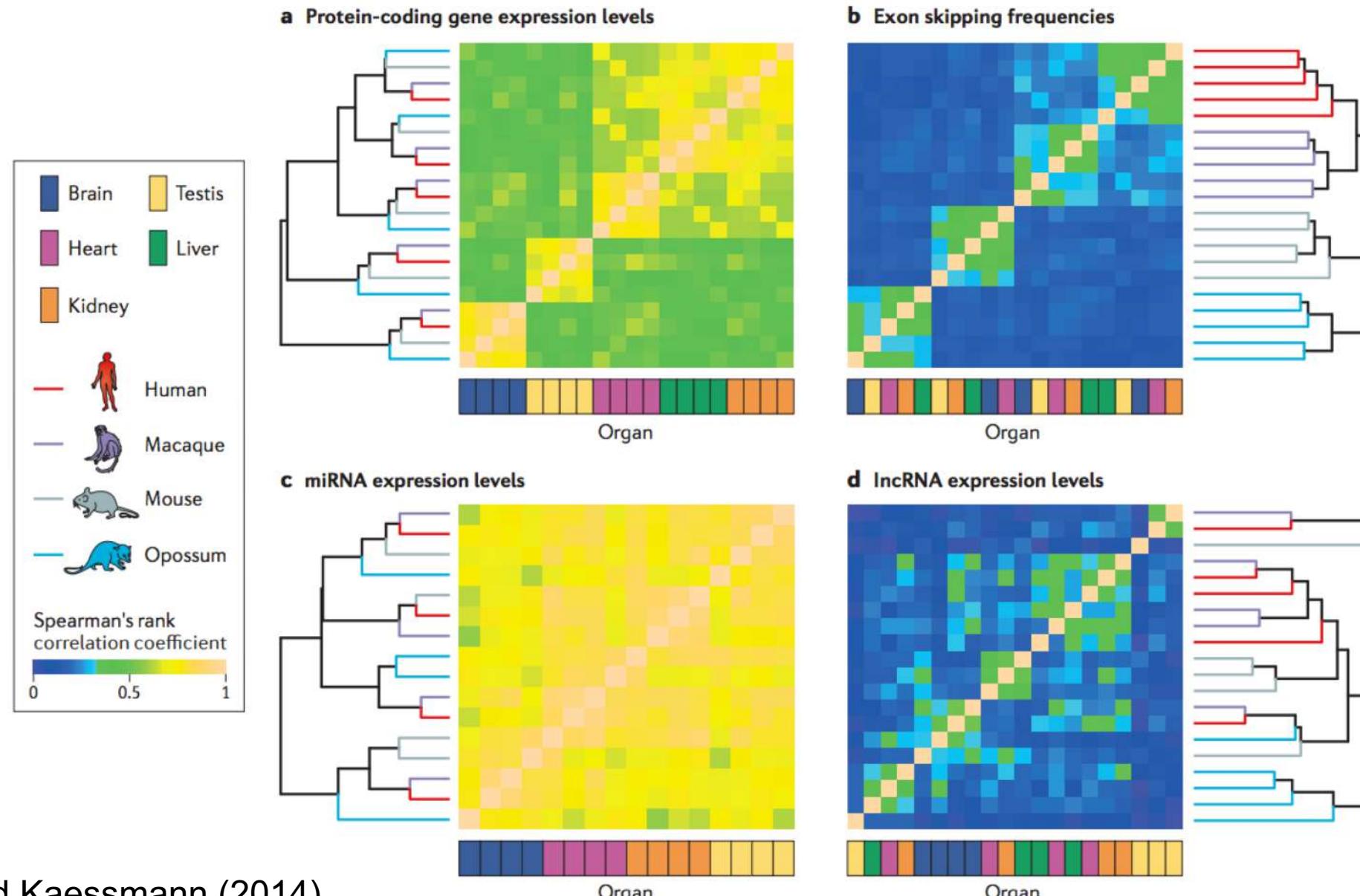
Adam Siepel,^{1,6} Gill Bejerano,¹ Jakob S. Pedersen,¹ Angie Raskin,¹ Michael A. Giardine,¹ Kate Rosenbloom,¹ Hiram Clawson,¹ John Spieth,⁴ LaDean M. Miller,³ Stephen Richards,⁵ George M. Weinstock,⁵ Richard K. Wilson,⁵ Webb Miller,³ W. James Kent,¹ Webb Miller,³ and David Haussler^{1,2}

¹Center for Biomolecular Science and Engineering, ²Howard Hughes Medical Institute, University of California at Santa Cruz, California 95064, USA; ³Center for Comparative Genomics and Bioinformatics, Penn State University, University Park, Pennsylvania 16802, USA; ⁴Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63108, USA; ⁵Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA

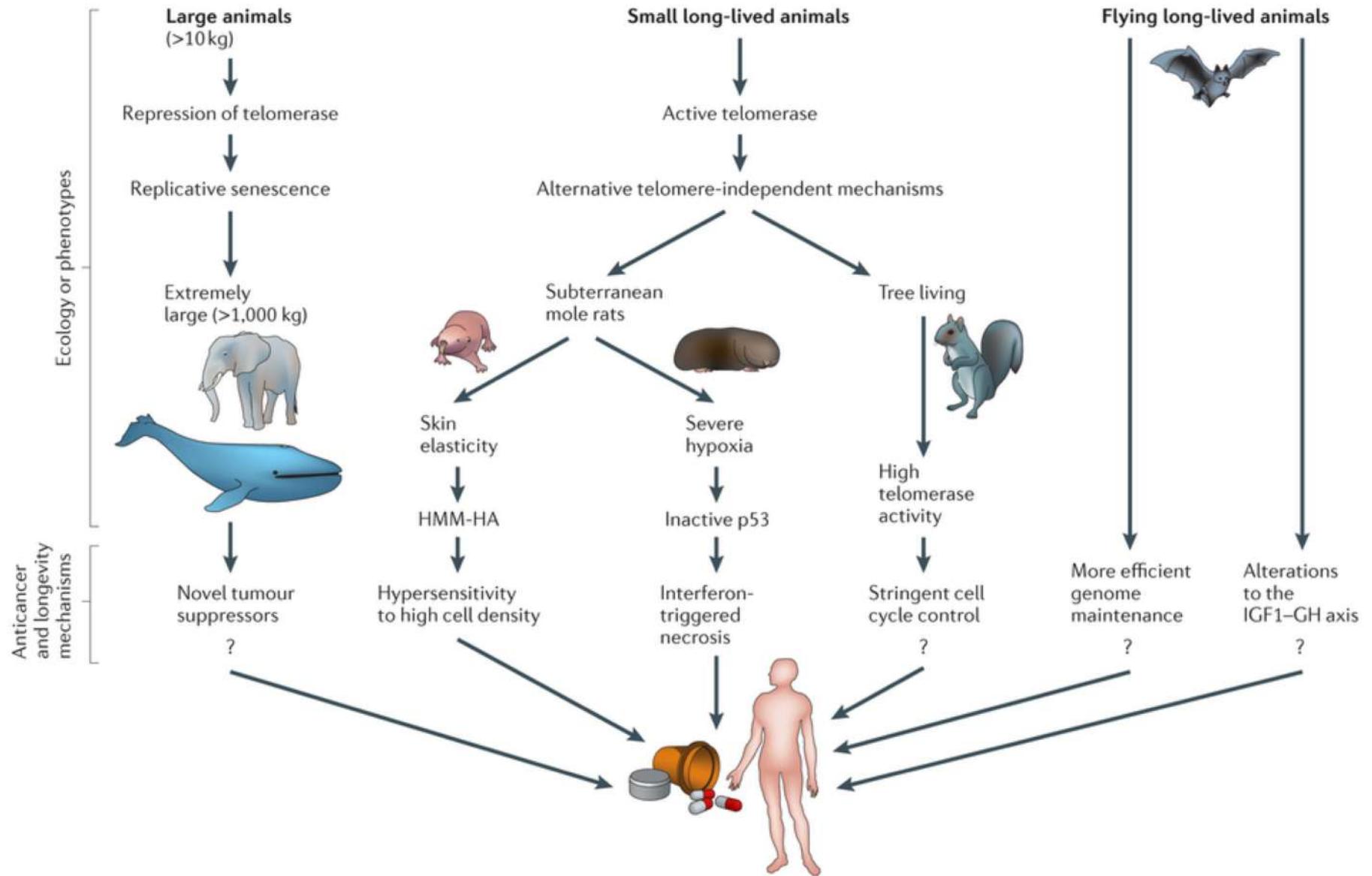
PhastCons



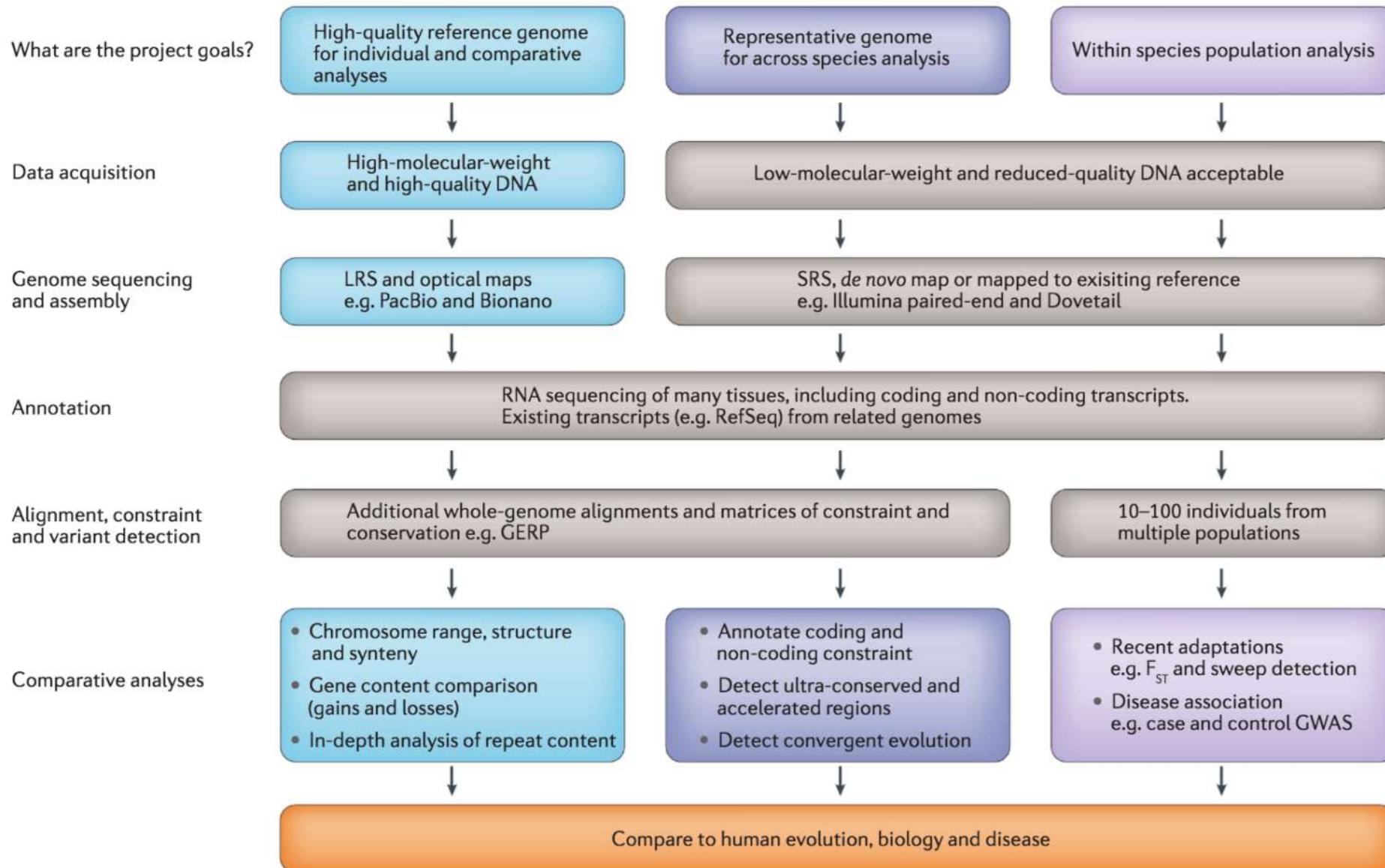
Global patterns of evolution for different aspects of the transcriptome



Comparative genomics of longevity ageing (with focus)

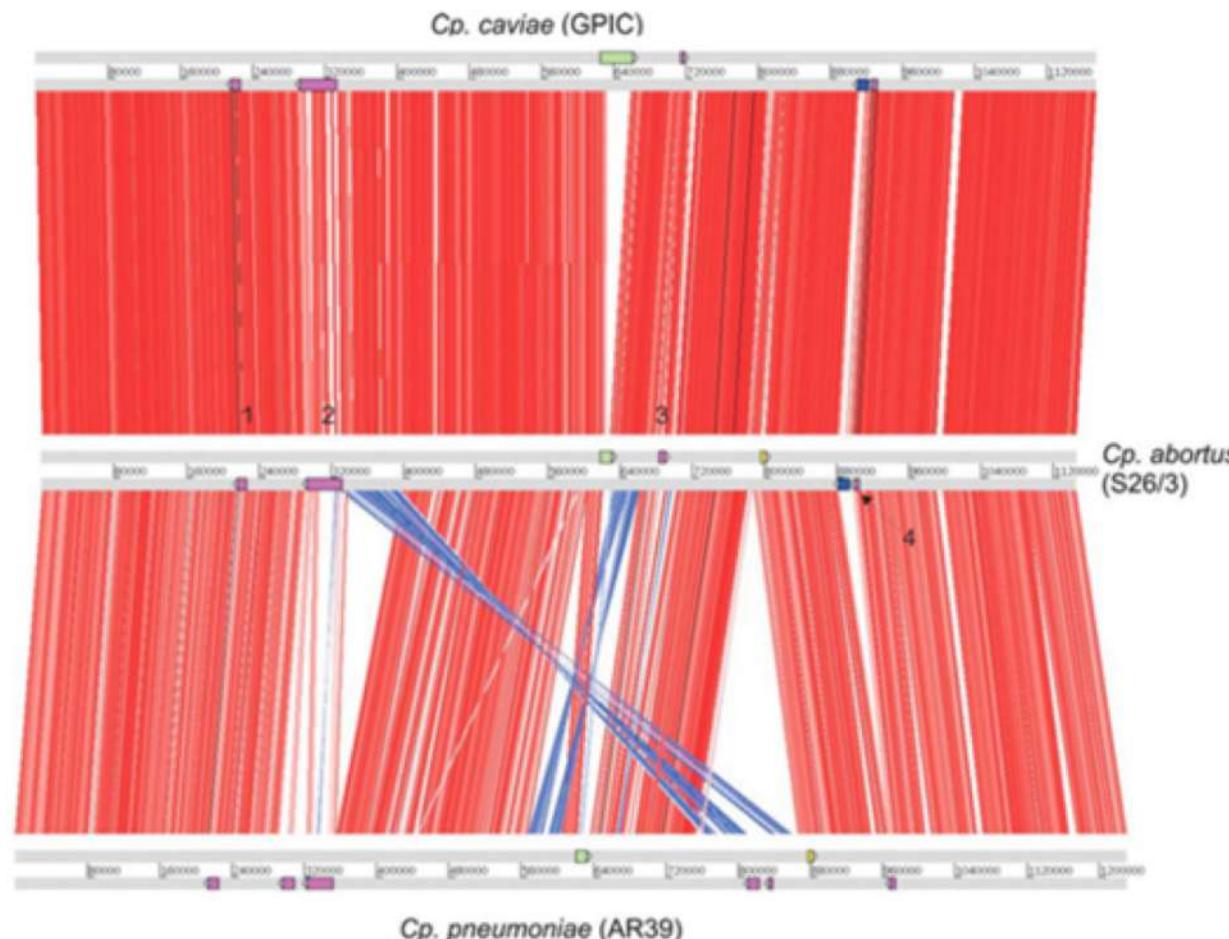


Designing a sequencing project: 2017 version



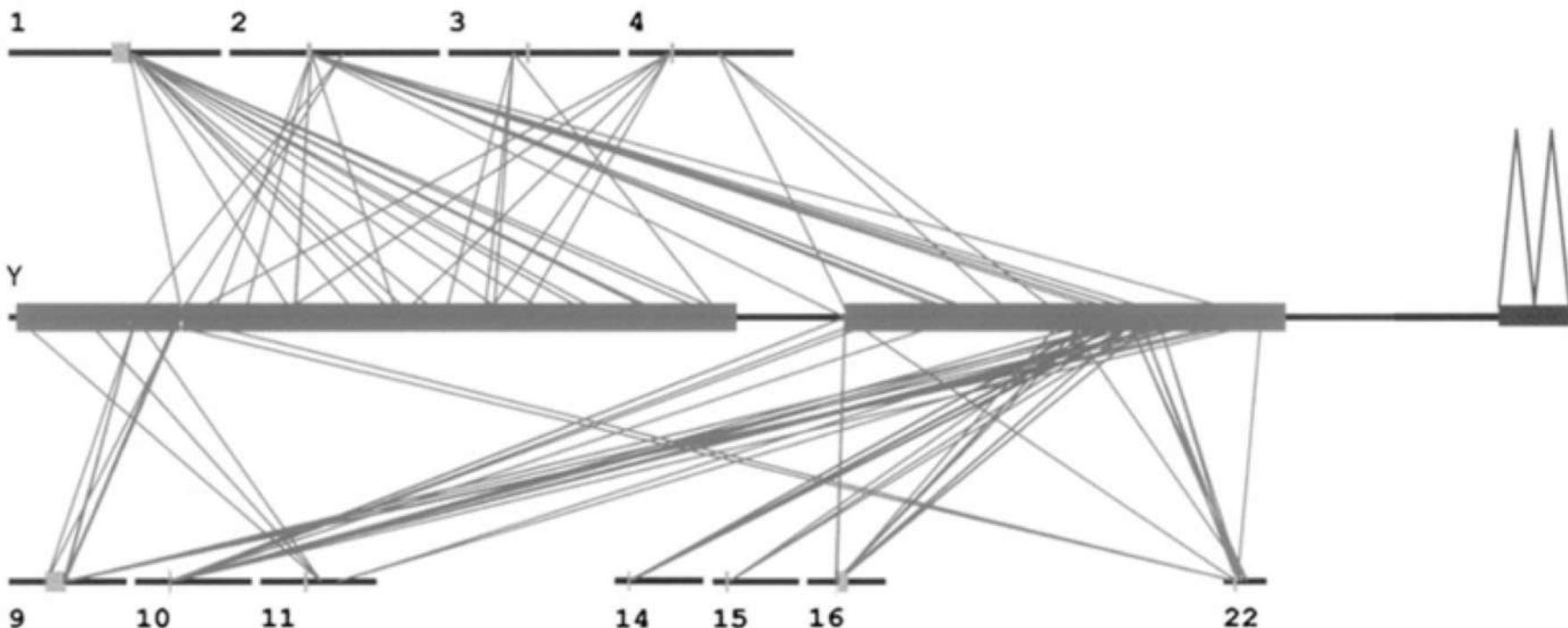
Genome visualisation

- this is the most common way to represent relationships within genomic positions
 - works when the number of cross-overs is limited



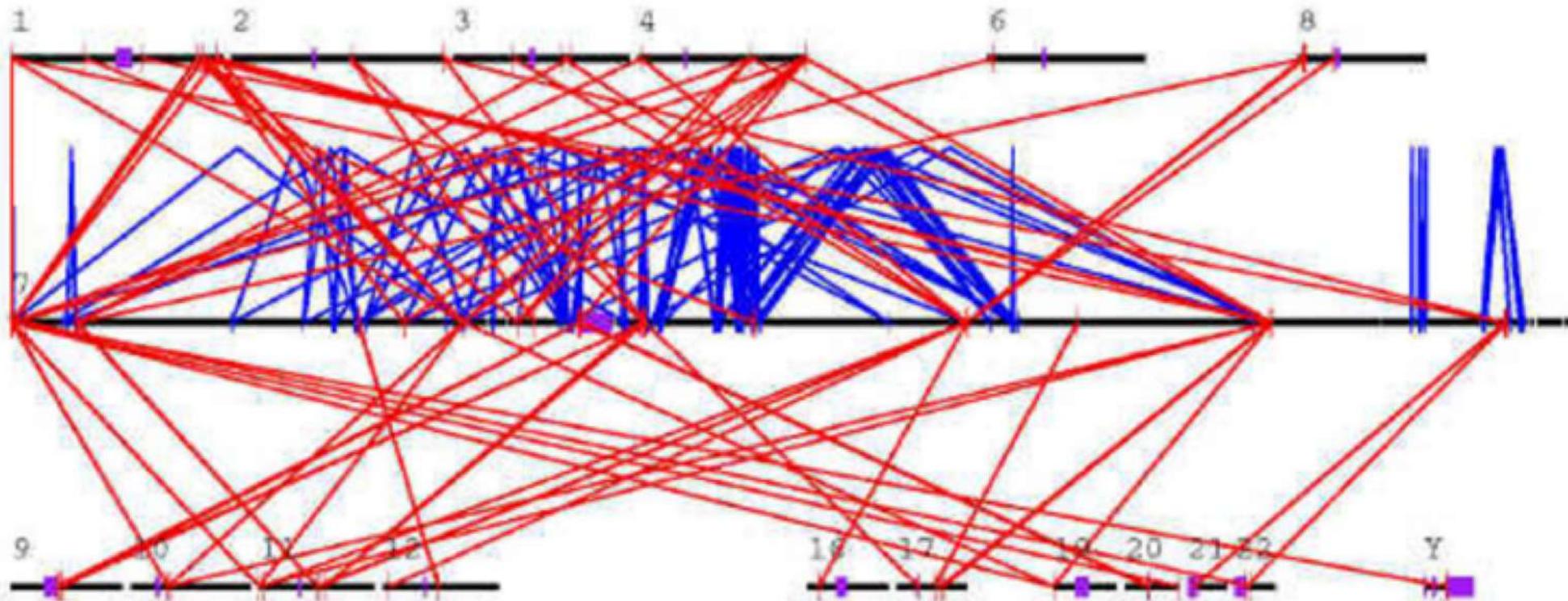
Genome Res. 2005 May; 15(5):629-40

- when complexity is increased, the figure starts to lose cohesion
 - routing becomes difficult to follow
 - there is no focus point for the eye – your eye wanders over the figure



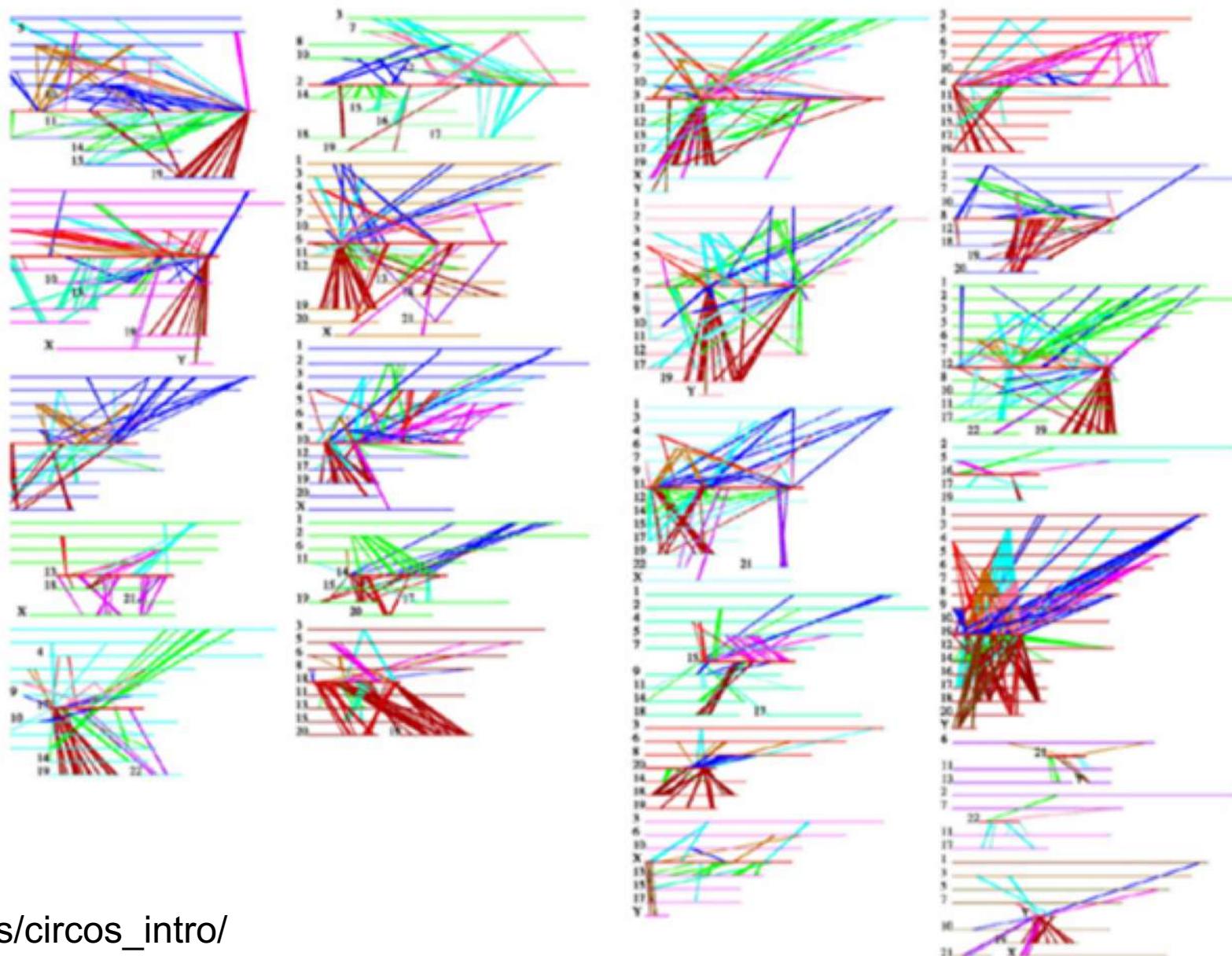
Genome Res. 2003 Jan;13(1):37-45

- things get worse and worse when mappings that link both neighbouring (blue) and distant (red) positions are shown



- you can try to fix things by partitioning your data set (somehow)

- mileage varies
 - generally poor



http://circos.ca/presentations/talks/circos_intro/

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." *Science* 291(5507): 1304-51.

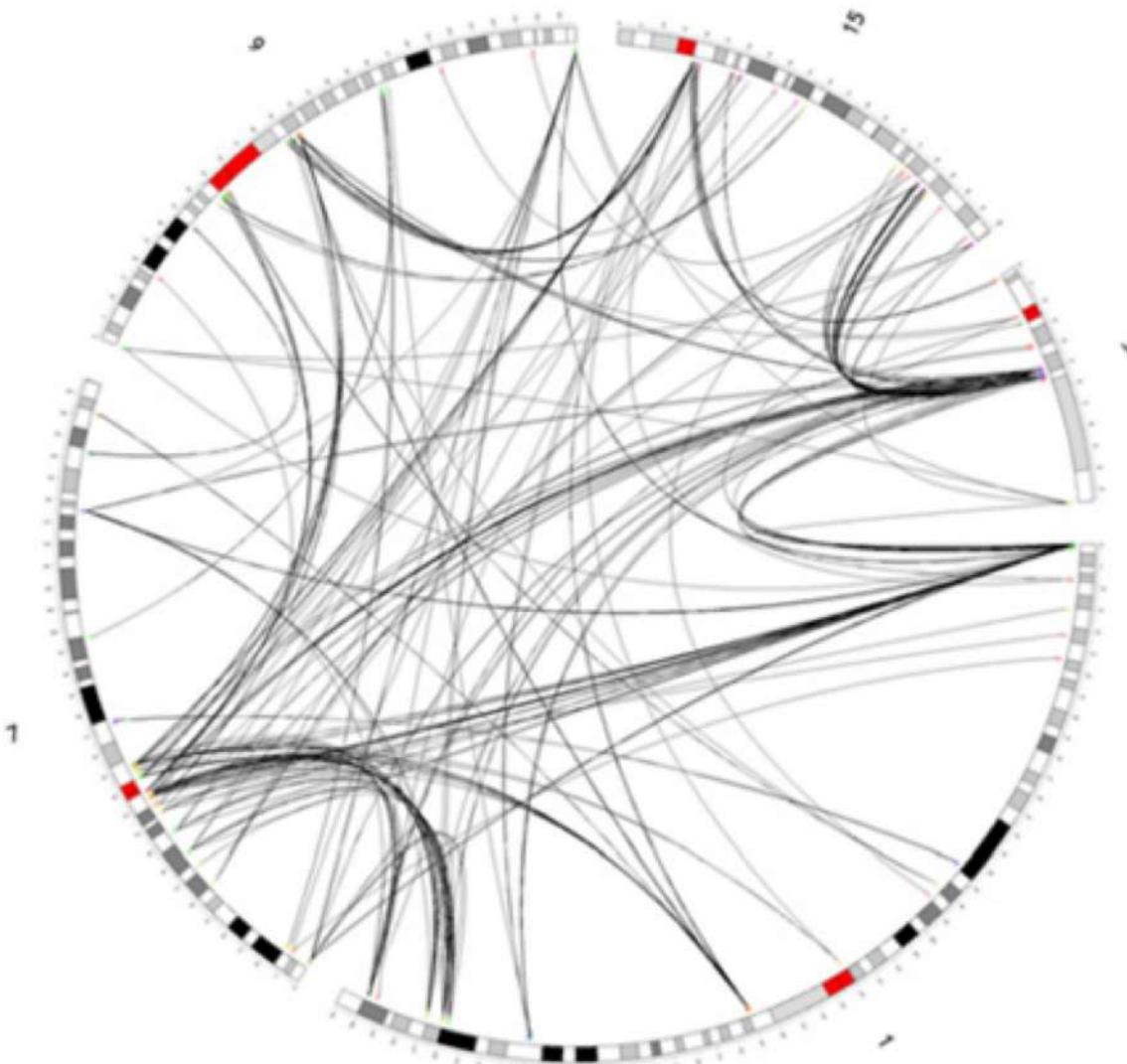
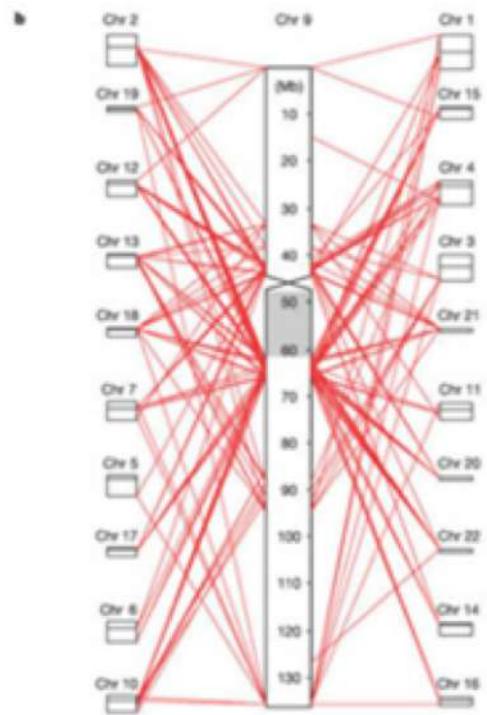
- finally, you descend into data overload and information hell
 - this is not an informative plot, although a pretty one



Segmental Duplications in *Arabidopsis* Genome. Alexander Kozik and Richard Michalewski, UC Davis, California

Image created with GenomePixelizer

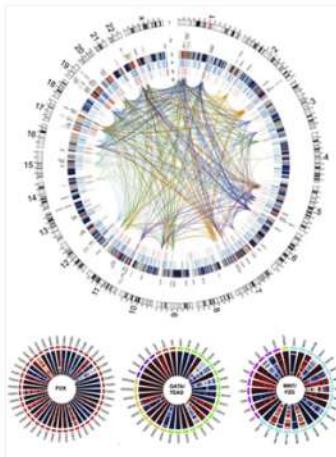
Circos



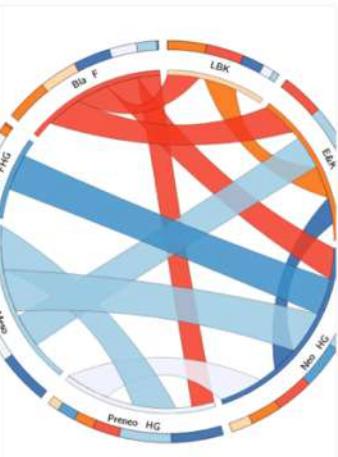
Circos image

Humphray, S. J., K. Oliver, et al. (2004).
"DNA sequence and analysis of human chromosome 9."
Nature 429(6990): 369-74.

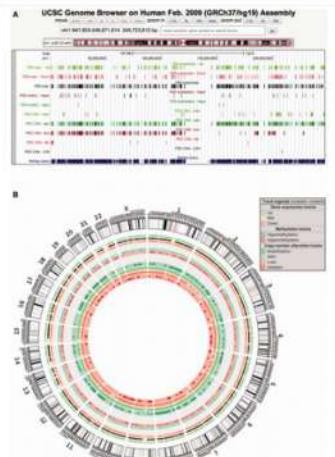
Circos



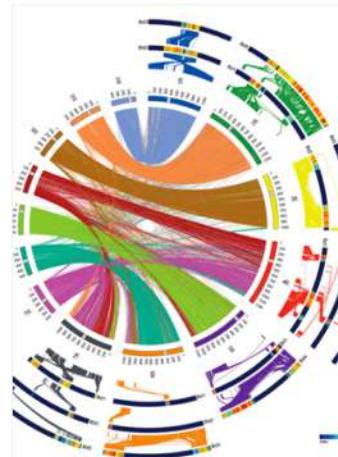
▲ 1 · 1 Dec 2013 | Saben J, Zhong Y, McKelvey S et al. (2014) A comprehensive analysis of the human placenta transcriptome *Placenta* 35:125-131.



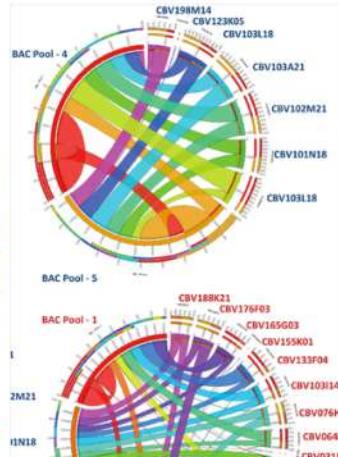
▲ 2 · 25 Oct 2013 | Bollongino R, Nehlich O, Richards MP et al. (2013) 2000 years of parallel societies in Stone Age Central Europe *Science* 342:479-481.



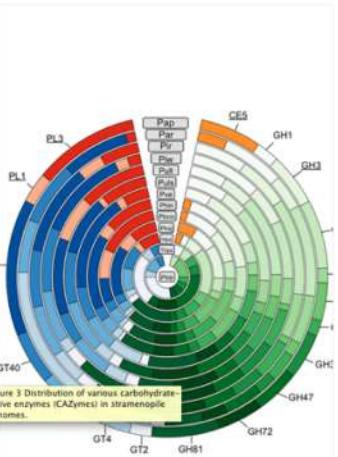
▲ 3 · 25 Oct 2013 | Dayem Ullah AZ, Cutts RJ, Ghetia M et al. (2013) The pancreatic expression database: recent extensions and updates *Nucleic Acids Res*



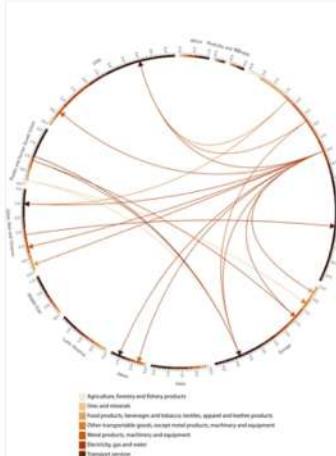
▲ 13 · 8 Oct 2013 | Martis MM, Zhou R, Haseneyer G et al. (2013) Reticulate Evolution of the Rye Genome *Plant Cell*



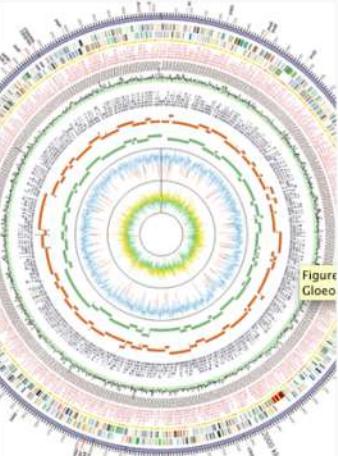
▲ 14 · 8 Oct 2013 | Buyyaparu R, Kantety RV, Yu JZ et al. (2013) BAC-Pool Sequencing and Analysis of Large Segments of A12 and D12 Homologous Chromosomes in Upland Cotton *PLoS One* 8:e76757.



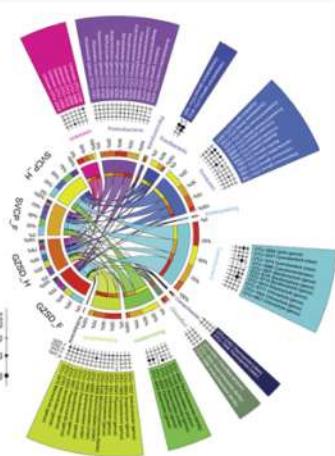
▲ 15 · 4 Oct 2013 | Adhikari BN, Hamilton JP, Zerillo MM et al. (2013) Comparative Genomics Reveals Insight into Virulence Strategies of Plant Pathogenic Oomycetes *PLoS One* 8:e75072.



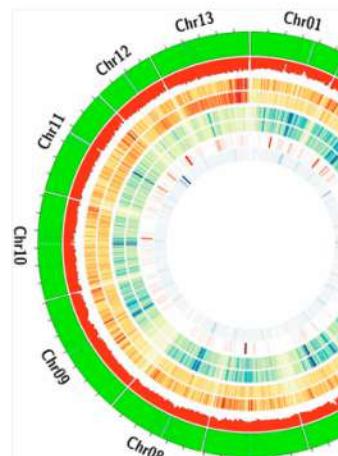
▲ 4 · 23 Oct 2013 | Kanemoto K, Moran D, Lenzen M et al. (2013) International trade undermines national emission reduction targets: New evidence from air pollution *Global Environmental Change*



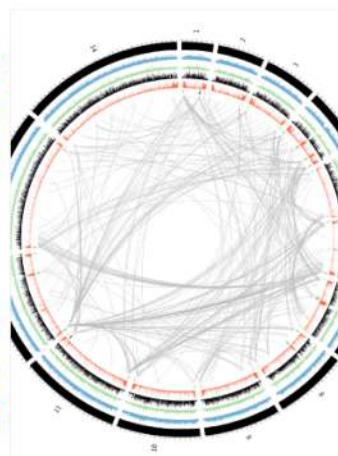
▲ 5 · 23 October 2013 | Saw JHW, Schatz M, Brown MV et al. (2013) Cultivation and Complete Genome Sequencing of *Gloeo bacter kilaueensis* sp. nov., from a Lava Cave in Kilauea Caldera, Hawaii *PLoS One* 8:e76376.



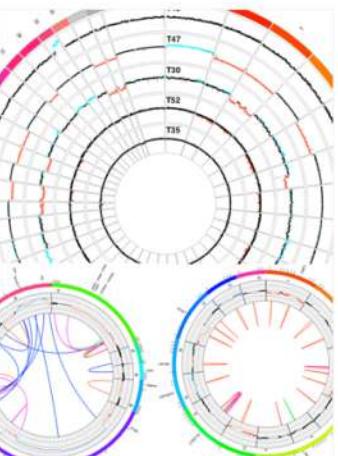
▲ 6 · 17 Oct 2013 | Ye L, Amberg I, Chapman D et al. (2013) Fish gut microbiota analysis differentiates physiology and behavior of invasive Asian carp and indigenous American fish *The ISME Journal*



▲ 16 · 1 Oct 2013 | Page JT, Huynh MD, Liechty ZS et al. (2013) Insights into the Evolution of Cotton Diploids and Polyploids from Whole-Genome Re-sequencing *Genes Genomes Genetics* 3:1809-1818.



▲ 17 · 30 Sep 2013 | Lemieux JE, Kyes SA, Otto TD et al. (2013) Genome-wide profiling of chromosome interactions in *Plasmodium falciparum* characterizes nuclear architecture and reconfigurations associated with antigenic variation *Molecular microbiology*

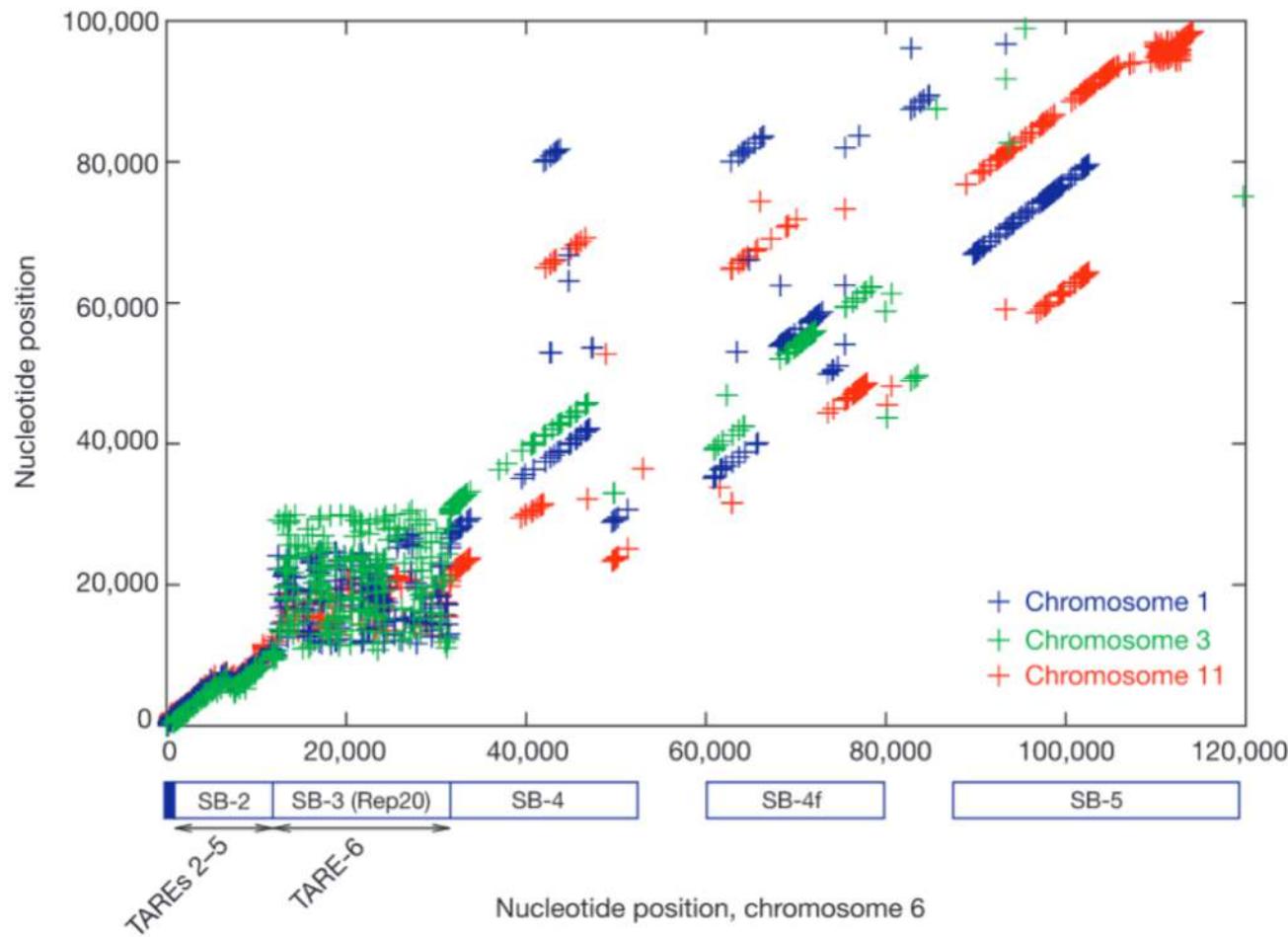


▲ 18 · 30 Sep 2013 | Beck J, Hennecke S, Bornemann-Kolatzki K et al. (2013) Genome Aberrations in Canine Mammary Carcinomas and Their Detection in Cell-Free Plasma DNA *PLoS One* 8:e75485.

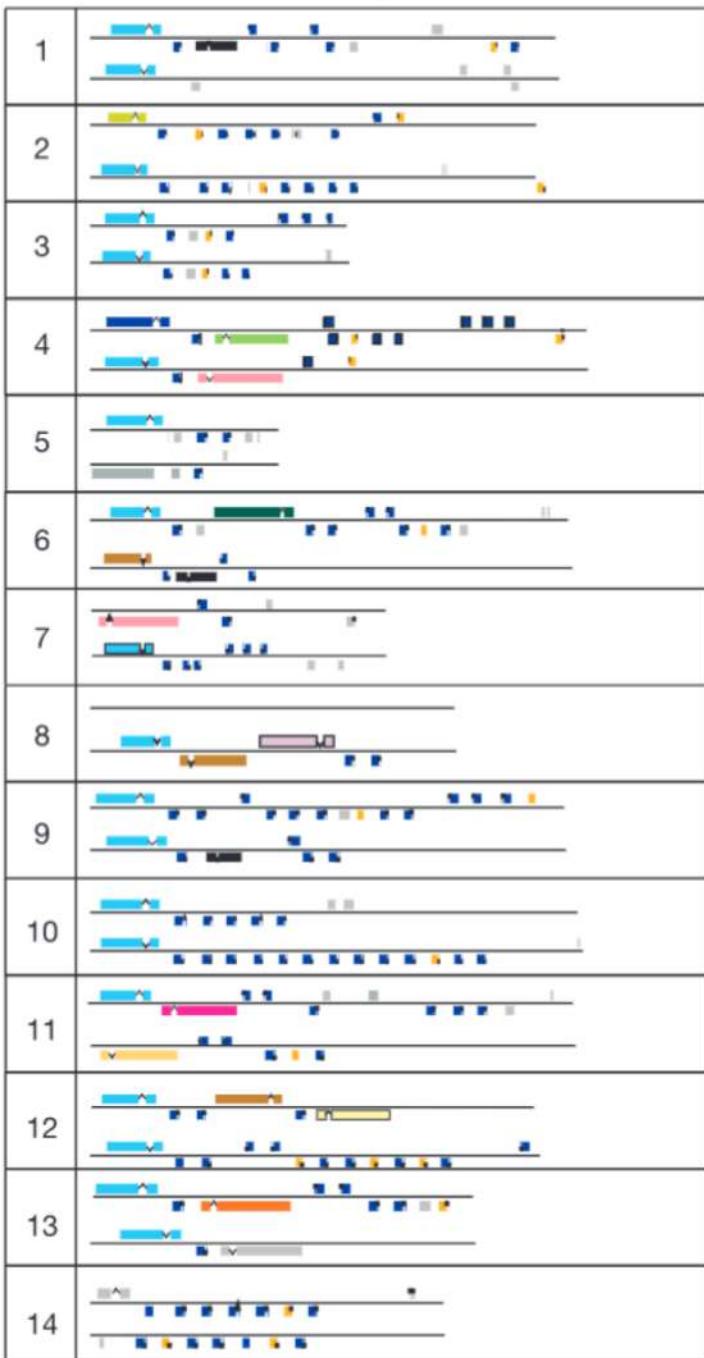
More case studies:

Genome sequence of the human malaria parasite *Plasmodium falciparum*

Malcolm J. Gardner¹, Neil Hall², Eula Fung³, Owen White¹, Matthew Berriman², Richard W. Hyman³, Jane M. Carlton¹, Arnab Pain², Karen E. Nelson¹, Sharen Bowman^{2*}, Ian T. Paulsen¹, Keith James², Jonathan A. Eisen¹, Kim Rutherford², Steven L. Salzberg¹, Alister Craig⁴, Sue Kyes⁵, Man-Suen Chan⁵, Vishvanath Nene¹, Shamira J. Shallom¹, Bernard Suh¹, Jeremy Peterson¹, Sam Angiuoli¹, Mihaela Pertea¹, Jonathan Allen¹, Jeremy Selengut¹, Daniel Haft¹, Michael W. Mather⁶, Akhil B. Vaidya⁶, David M. A. Martin⁷, Alan H. Fairlamb⁷, Martin J. Fraunholz⁸, David S. Roos⁸, Stuart A. Ralph⁹, Geoffrey I. McFadden⁹, Leda M. Cummings¹, G. Mani Subramanian¹⁰, Chris Mungall¹¹, J. Craig Venter¹², Daniel J. Carucci¹³, Stephen L. Hoffman^{13*}, Chris Newbold⁵, Ronald W. Davis³, Claire M. Fraser¹ & Bart Barrell²



....The conserved regions fall into five large subtelomeric, contains the 7-bp telomeric repeat in a variable number of near-exact copies

a Telomeric organization

Genome sequence of the human malaria parasite *Plasmodium falciparum*

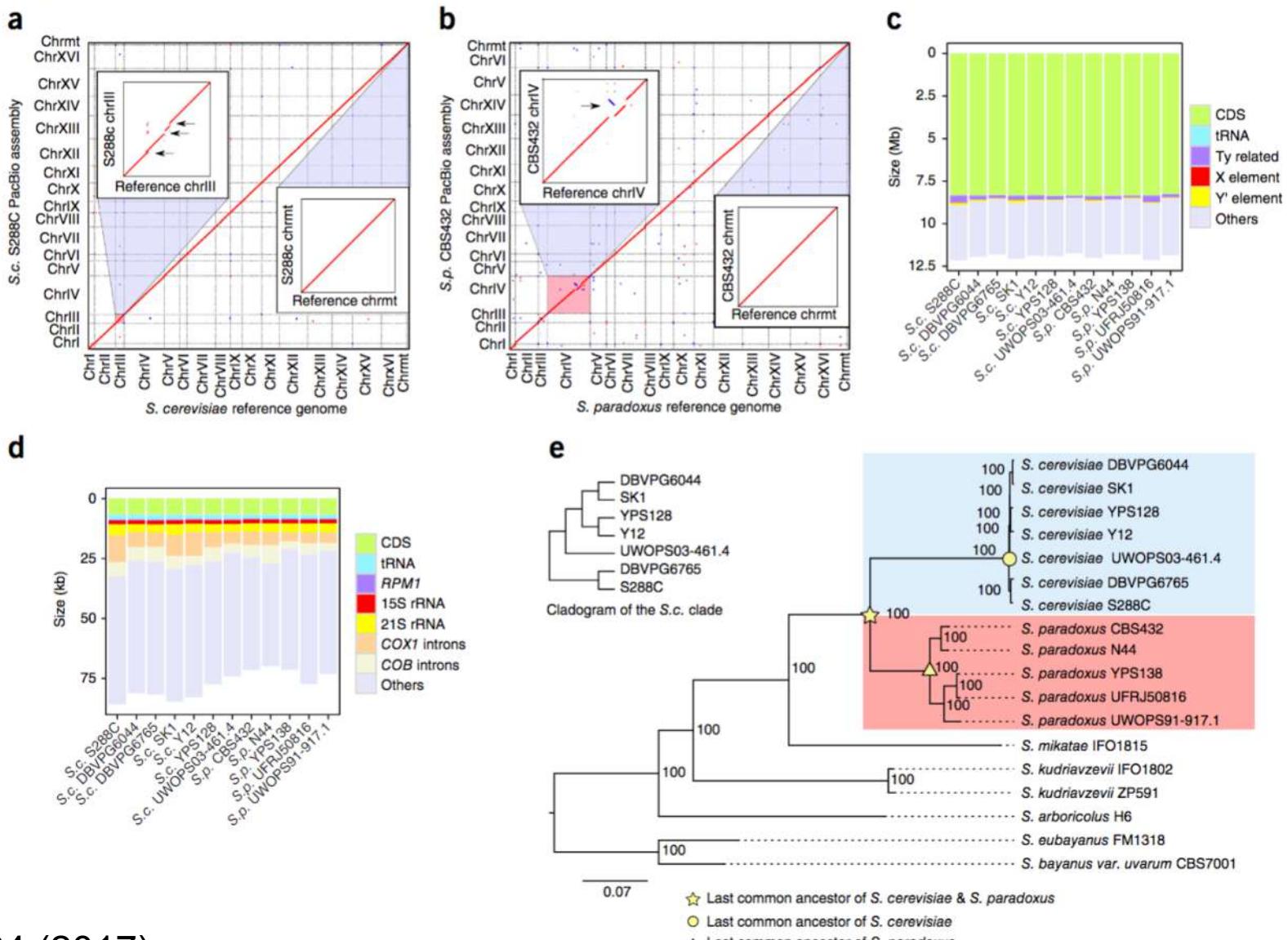
Malcolm J. Gardner¹, Neil Hall², Eula Fung³, Owen White¹, Matthew Berriman², Richard W. Hyman³, Jane M. Carlton¹, Arnab Pain², Karen E. Nelson¹, Sharen Bowman^{2*}, Ian T. Paulsen¹, Keith James², Jonathan A. Eisen¹, Kim Rutherford², Steven L. Salzberg¹, Alister Craig⁴, Sue Kyes⁵, Man-Suen Chan⁵, Vishvanath Nene¹, Shamira J. Shallom¹, Bernard Suh¹, Jeremy Peterson¹, Sam Angiuoli¹, Mihaela Pertea¹, Jonathan Allen¹, Jeremy Selengut¹, Daniel Haft¹, Michael W. Mather⁶, Akhil B. Vaidya⁶, David M. A. Martin⁷, Alan H. Fairlamb⁷, Martin J. Fraunholz⁸, David S. Roos⁸, Stuart A. Ralph⁹, Geoffrey I. McFadden⁹, Leda M. Cummings¹, G. Mani Subramanian¹⁰, Chris Mungall¹¹, J. Craig Venter¹², Daniel J. Carucci¹³, Stephen L. Hoffman^{13*}, Chris Newbold⁵, Ronald W. Davis³, Claire M. Fraser¹ & Bart Barrell²

The var genes code for proteins which are exported to the surface of infected red blood cells where they mediate adherence to host endothelial receptors, resulting in the sequestration of infected cells in a variety of organs. These and other adherence properties are important virulence factors that contribute to the development of severe disease

Contrasting evolutionary genome dynamics between domesticated and wild yeasts

Jia-Xing Yue¹ , Jing Li¹, Louise Aigrain², Johan Hallin¹ , Karl Persson³ , Karen Oliver², Anders Bergström², Paul Coupland^{2,5}, Jonas Warringer³ , Marco Cosentino Lagomarsino⁴, Gilles Fischer⁴, Richard Durbin² & Gianni Liti¹

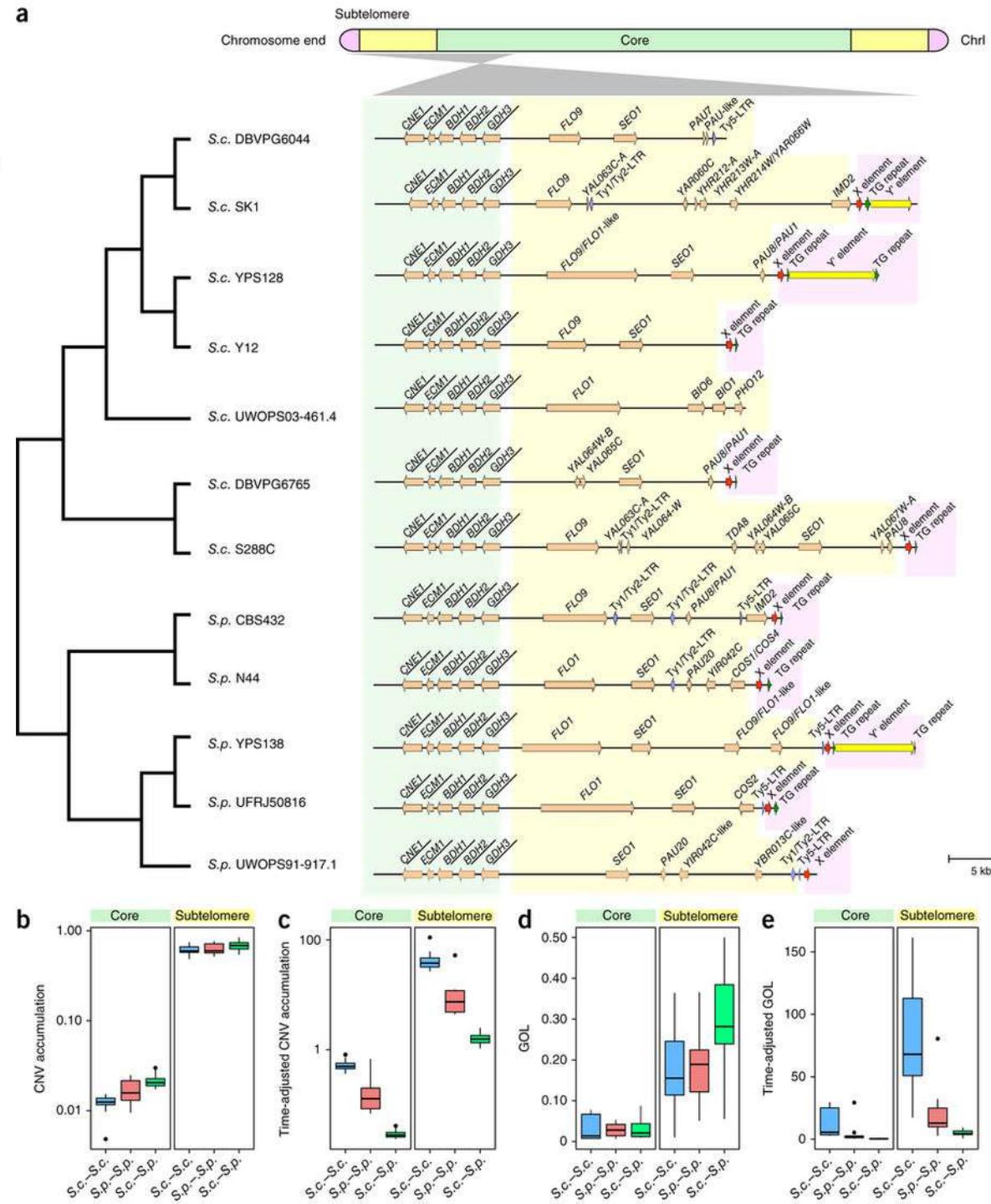
- long-read sequencing to generate **end-to-end genome assemblies** for **12 strains** representing major subpopulations of the partially domesticated yeast *Saccharomyces cerevisiae* and its wild relative *S. paradoxus*.



Contrasting evolutionary genome dynamics between domesticated and wild yeasts

Jia-Xing Yue¹ , Jing Li¹, Louise Aigrain², Johan Hallin¹ , Karl Persson³ , Karen Oliver², Anders Bergström², Paul Coupland^{2,5}, Jonas Warringer³ , Marco Cosentino Lagomarsino⁴, Gilles Fischer⁴, Richard Durbin² & Gianni Liti¹

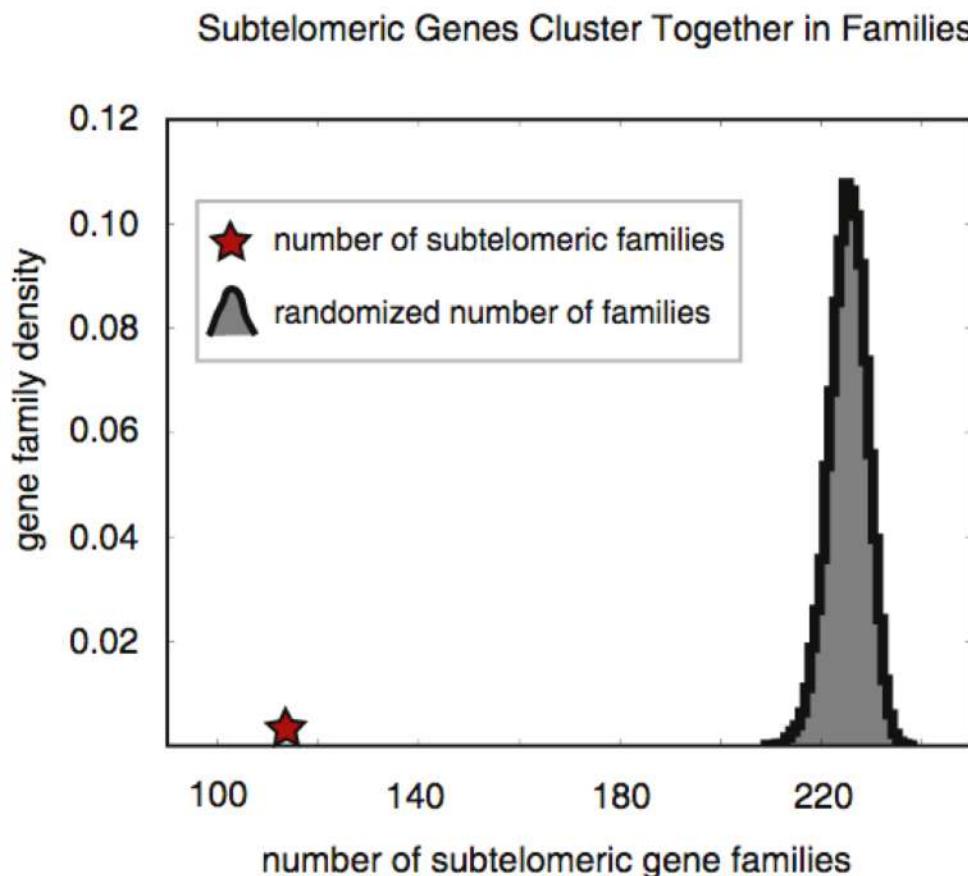
- enable precise definition of chromosomal boundaries between cores and subtelomeres
- S. paradoxus* shows faster accumulation of balanced rearrangements (inversions, reciprocal translocations and transpositions), *S. cerevisiae* accumulates unbalanced rearrangements (novel insertions, deletions and duplications) more rapidly.
- Such striking contrasts between wild and domesticated yeasts are likely to reflect the influence of human activities on structural genome evolution.



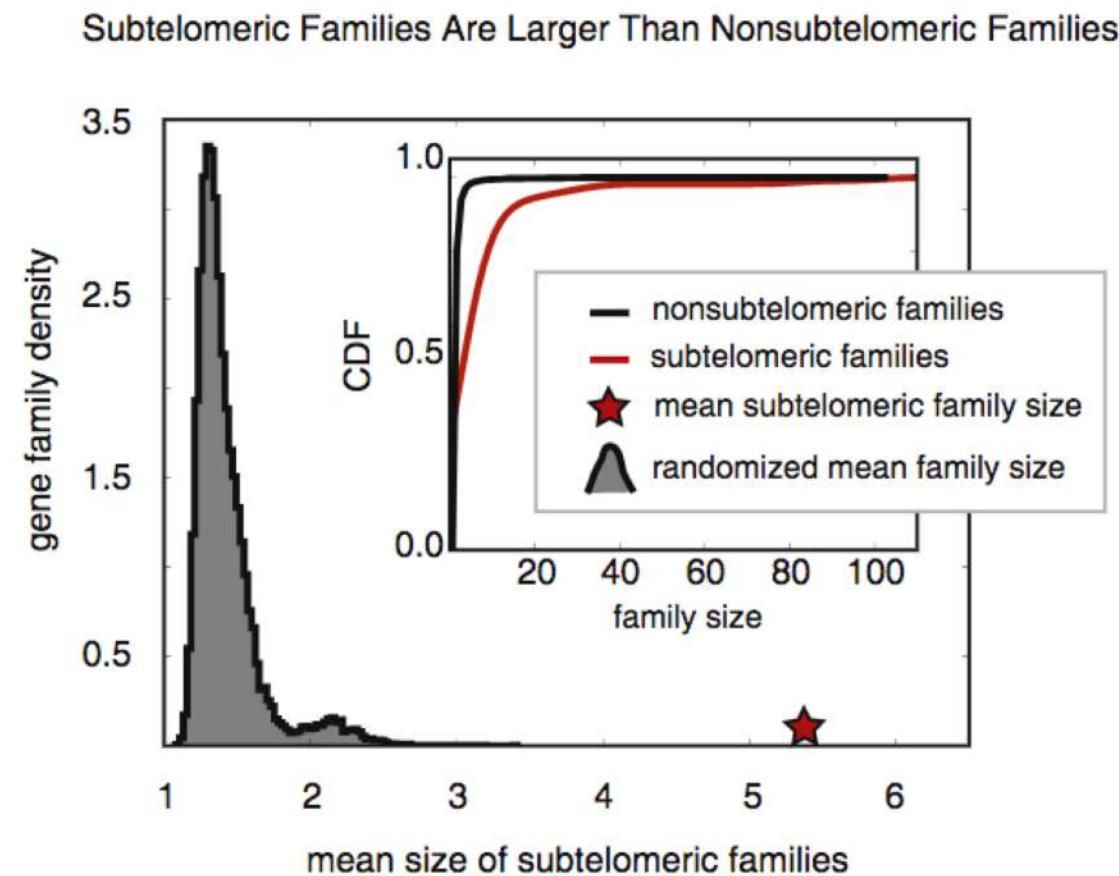
Rapid Expansion and Functional Divergence of Subtelomeric Gene Families in Yeasts

...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



B

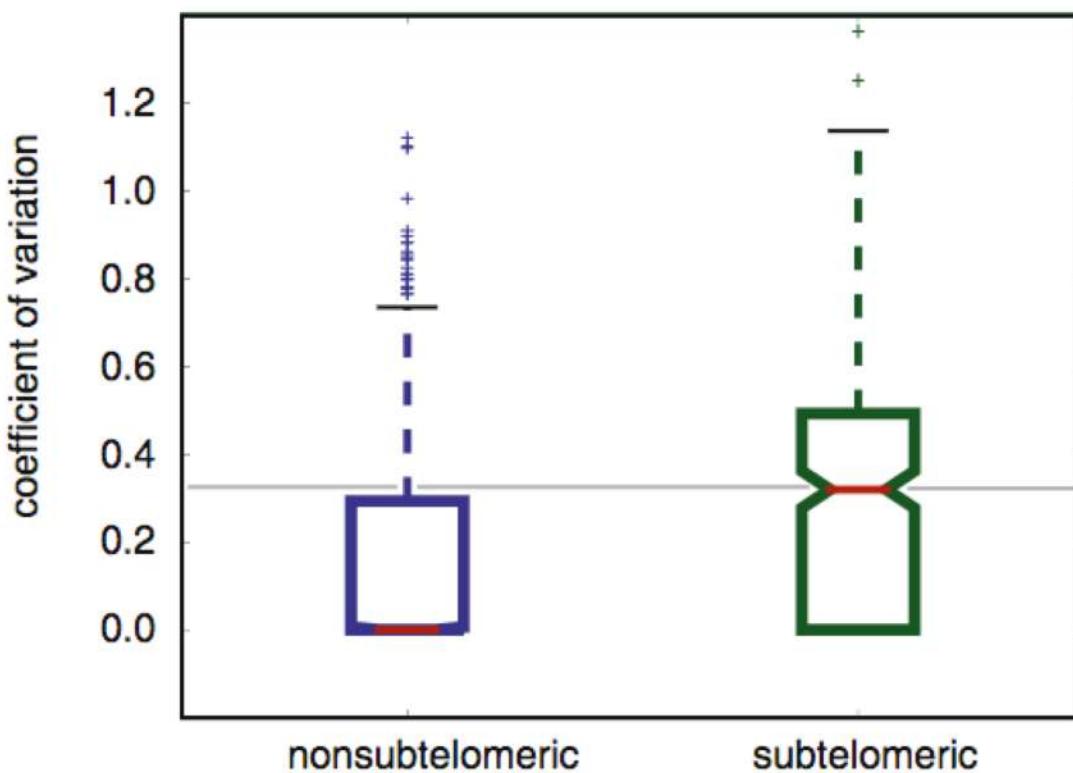


Rapid Expansion and Functional Divergence of Subtelomeric Gene Families in Yeasts

...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

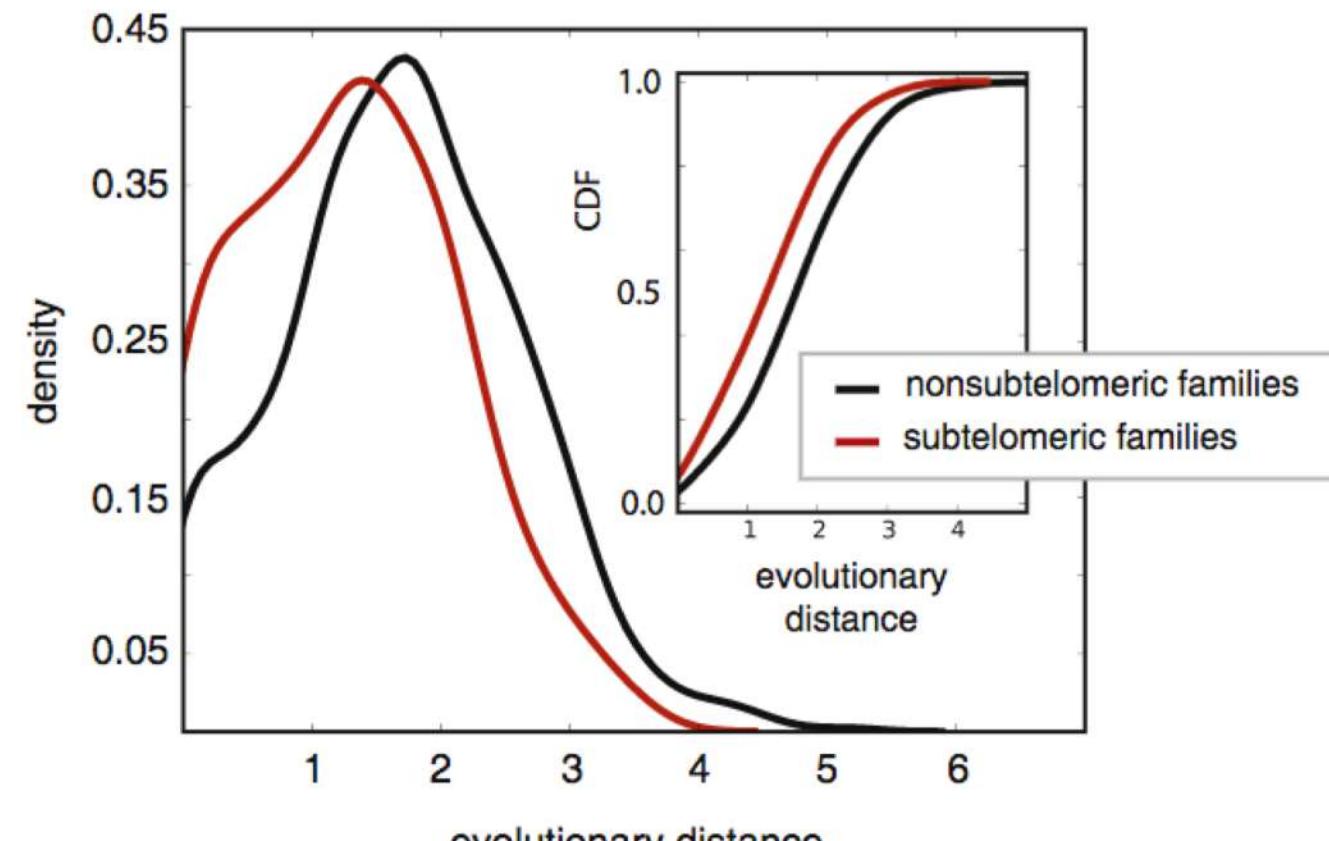
C

Subtelomeric Families Show More Copy Number Variation Between Species



D

Subtelomeric Families Show More Recent Duplications

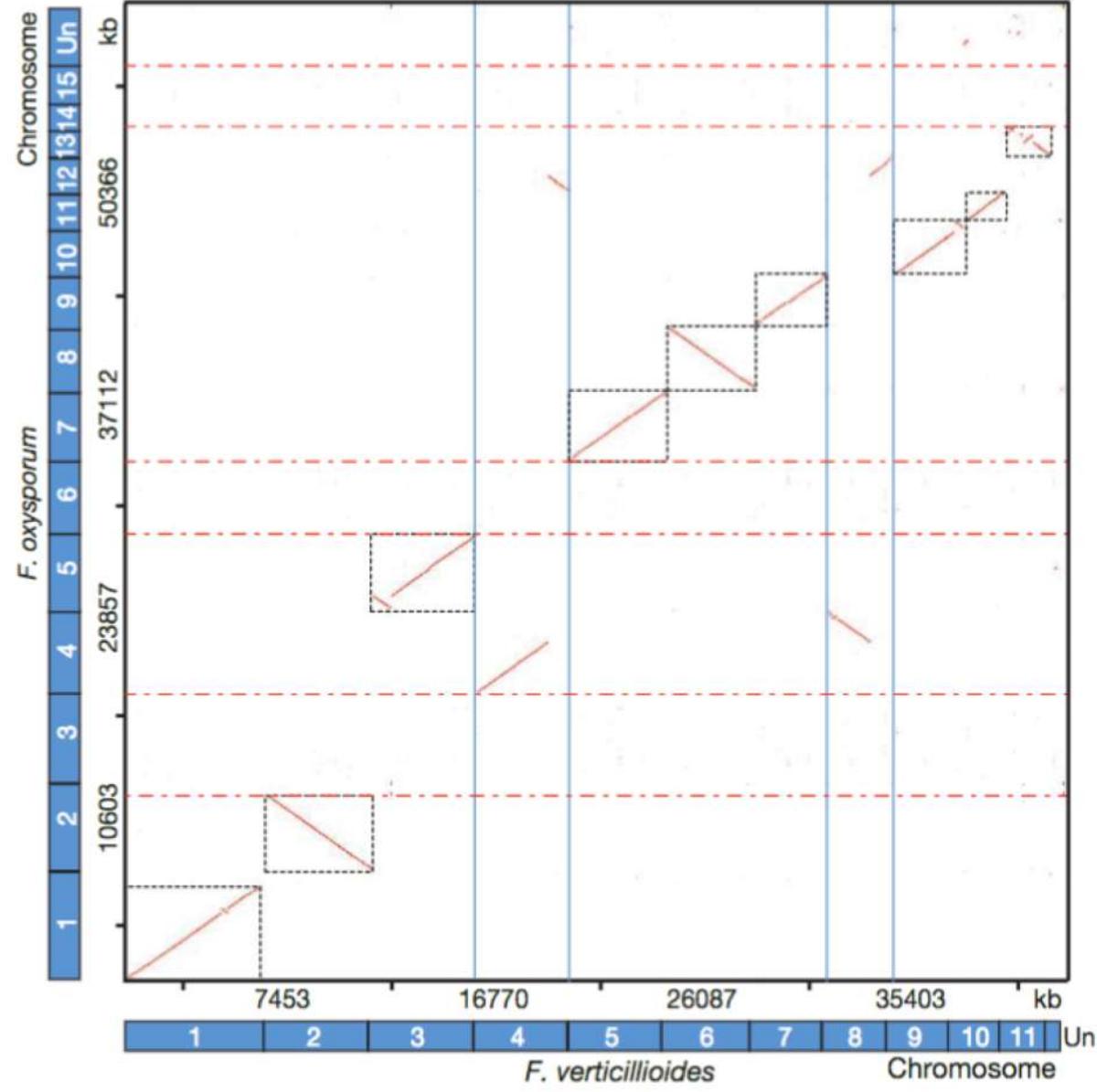
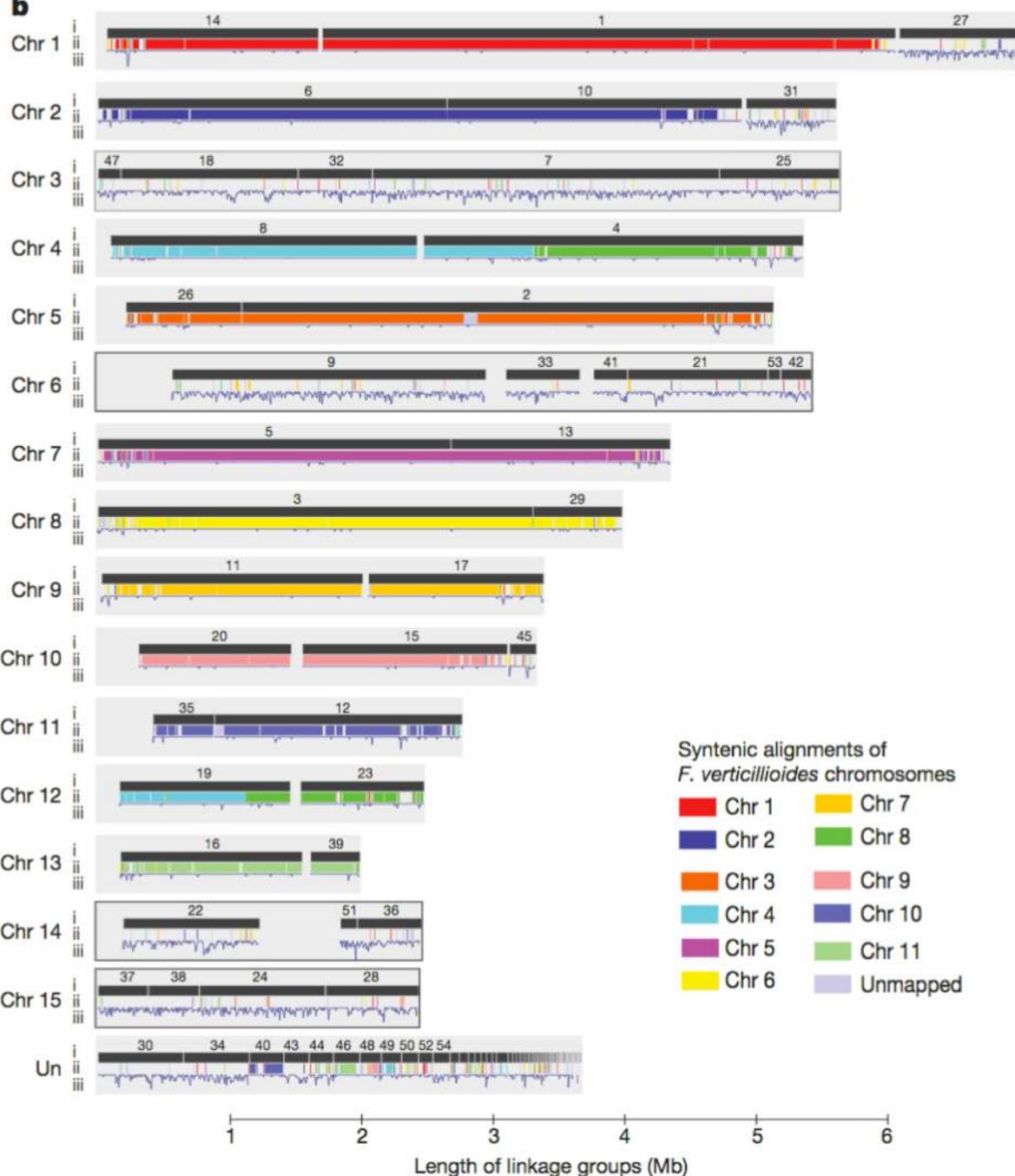


ARTICLES

Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*

Li-Jun Ma^{1*}, H. Charlotte van der Does^{2*}, Katherine A. Borkovich³, Jeffrey J. Coleman⁴, Marie-Josée Daboussi⁵, Antonio Di Pietro⁶, Marie Dufresne⁵, Michael Freitag⁷, Manfred Grabherr¹, Bernard Henrissat⁸, Petra M. Houterman², Seogchan Kang⁹, Won-Bo Shim¹⁰, Charles Woloshuk¹¹, Xiaohui Xie¹², Jin-Rong Xu¹¹, John Antoniw¹³, Scott E. Baker¹⁴, Burton H. Bluhm¹¹, Andrew Breakspear¹⁵, Daren W. Brown¹⁶, Robert A. E. Butchko¹⁶, Sinead Chapman¹, Richard Coulson¹⁷, Pedro M. Coutinho⁸, Etienne G. J. Danchin^{8†}, Andrew Diener¹⁸, Liane R. Gale¹⁵, Donald M. Gardiner¹⁹, Stephen Goff²⁰, Kim E. Hammond-Kosack¹³, Karen Hilburn¹⁵, Aurélie Hua-Van⁵, Wilfried Jonkers², Kemal Kazan¹⁹, Chinnappa D. Kodira^{1†}, Michael Koehrsen¹, Lokesh Kumar¹, Yong-Hwan Lee²¹, Liande Li³, John M. Manners¹⁹, Diego Miranda-Saavedra²², Mala Mukherjee¹⁰, Gyungsoon Park³, Jongsun Park²¹, Sook-Young Park^{9†}, Robert H. Proctor¹⁶, Aviv Regev¹, M. Carmen Ruiz-Roldan⁶, Divya Sain³, Sharadha Sakthikumar¹, Sean Sykes¹, David C. Schwartz²³, B. Gillian Turgeon²⁴, Ilan Wapinski¹, Olen Yoder²⁵, Sarah Young¹, Qiandong Zeng¹, Shiguo Zhou²³, James Galagan¹, Christina A. Cuomo¹, H. Corby Kistler¹⁵ & Martijn Rep²

Our analysis revealed lineage-specific (LS) genomic regions in *F. oxysporum* that include four entire chromosomes and account for more than one-quarter of the genome. LS regions are rich in transposons and genes with distinct evolutionary profiles but related to pathogenicity, indicative of horizontal acquisition. Experimentally, we demonstrate the transfer of two LS chromosomes between strains of *F. oxysporum*, converting a non-pathogenic strain into a pathogen.

a**b**

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

Beatriz Vicoso¹ & Doris Bachtrog¹

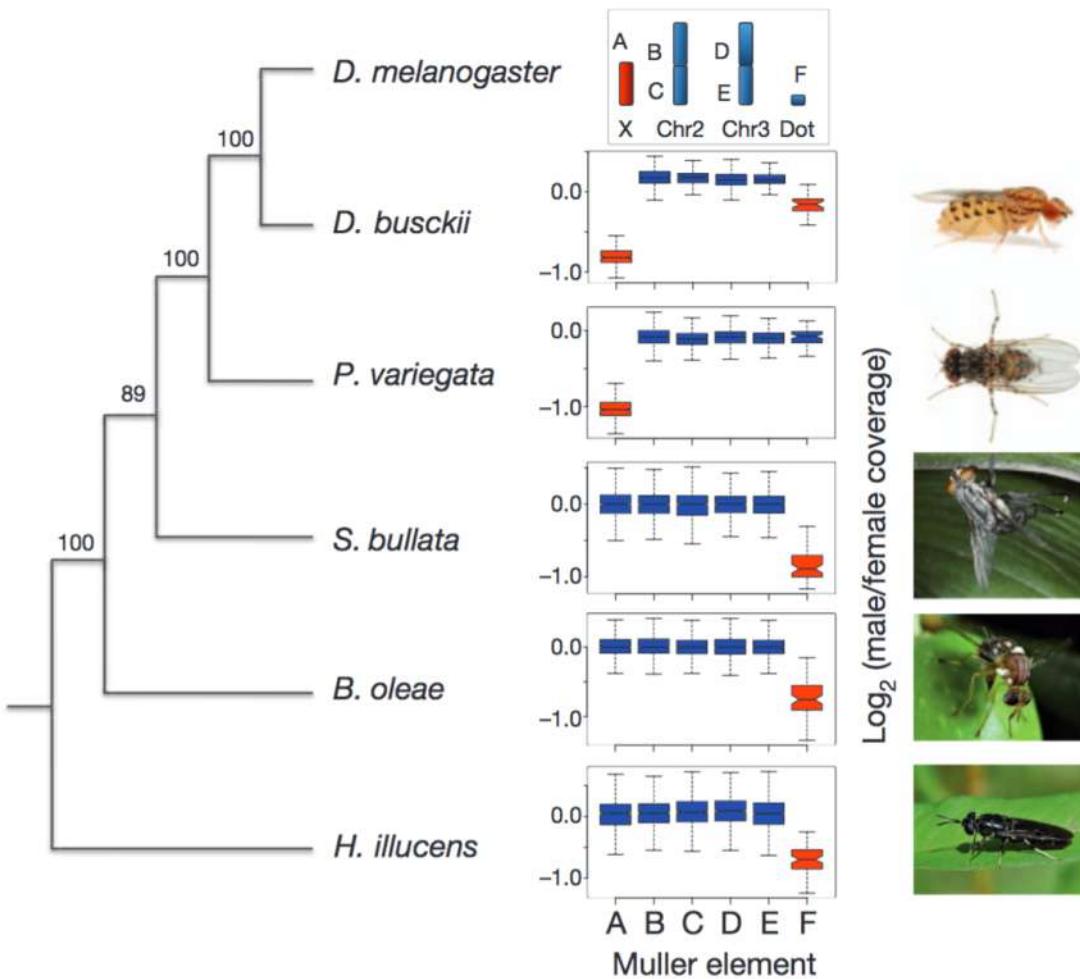


Figure 1 | Sex chromosomes in higher Diptera revealed by genome analysis. Evolutionary relationship inferred from 185 conserved protein-coding genes (93,134 amino acids) using PhyML (with bootstrap values indicated at the nodes), and male-to-female coverage ratio across chromosome elements (Muller elements 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 interquartile range.

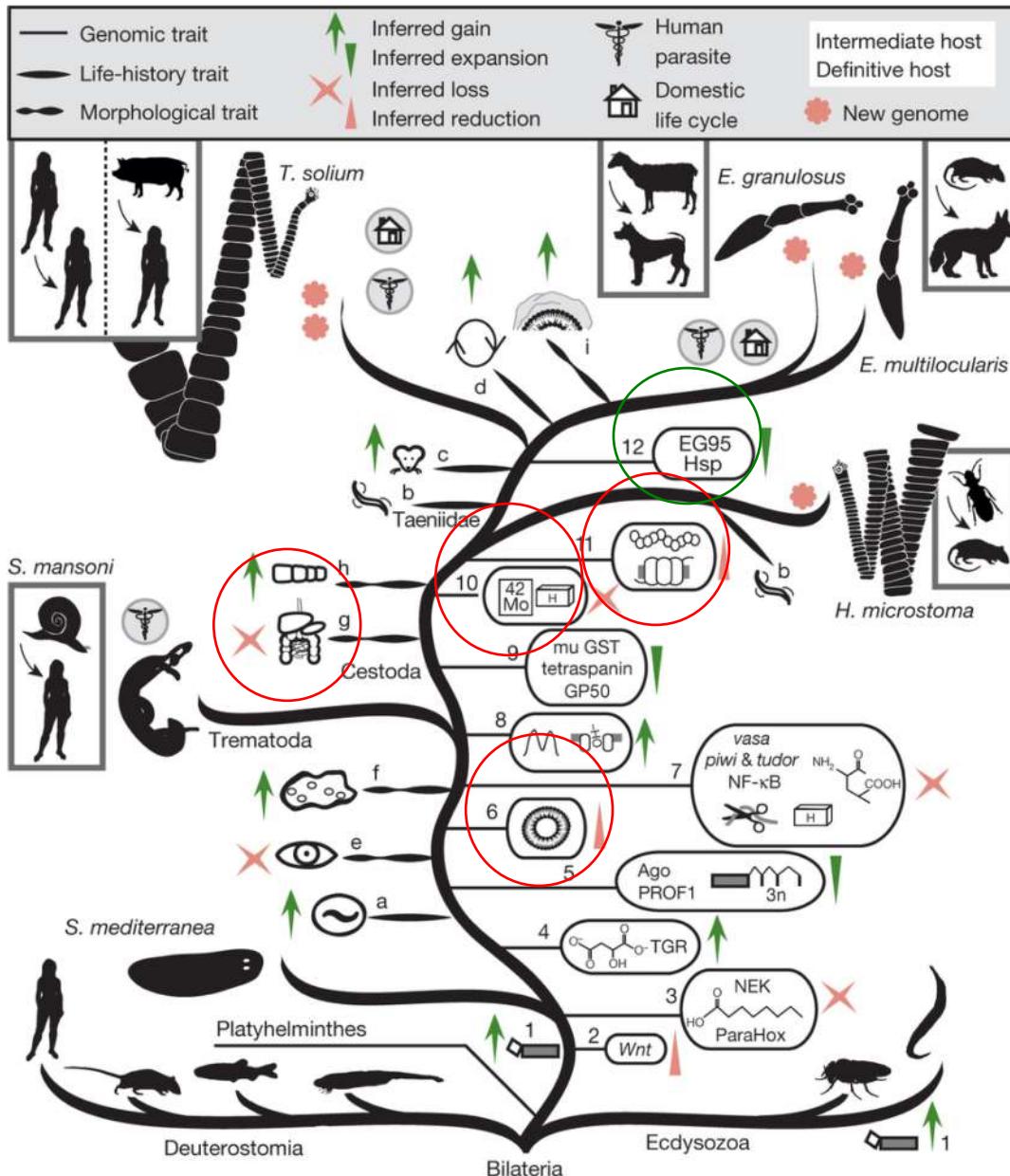
Comparative genomics of tapeworms

tapeworms

Blood fluke

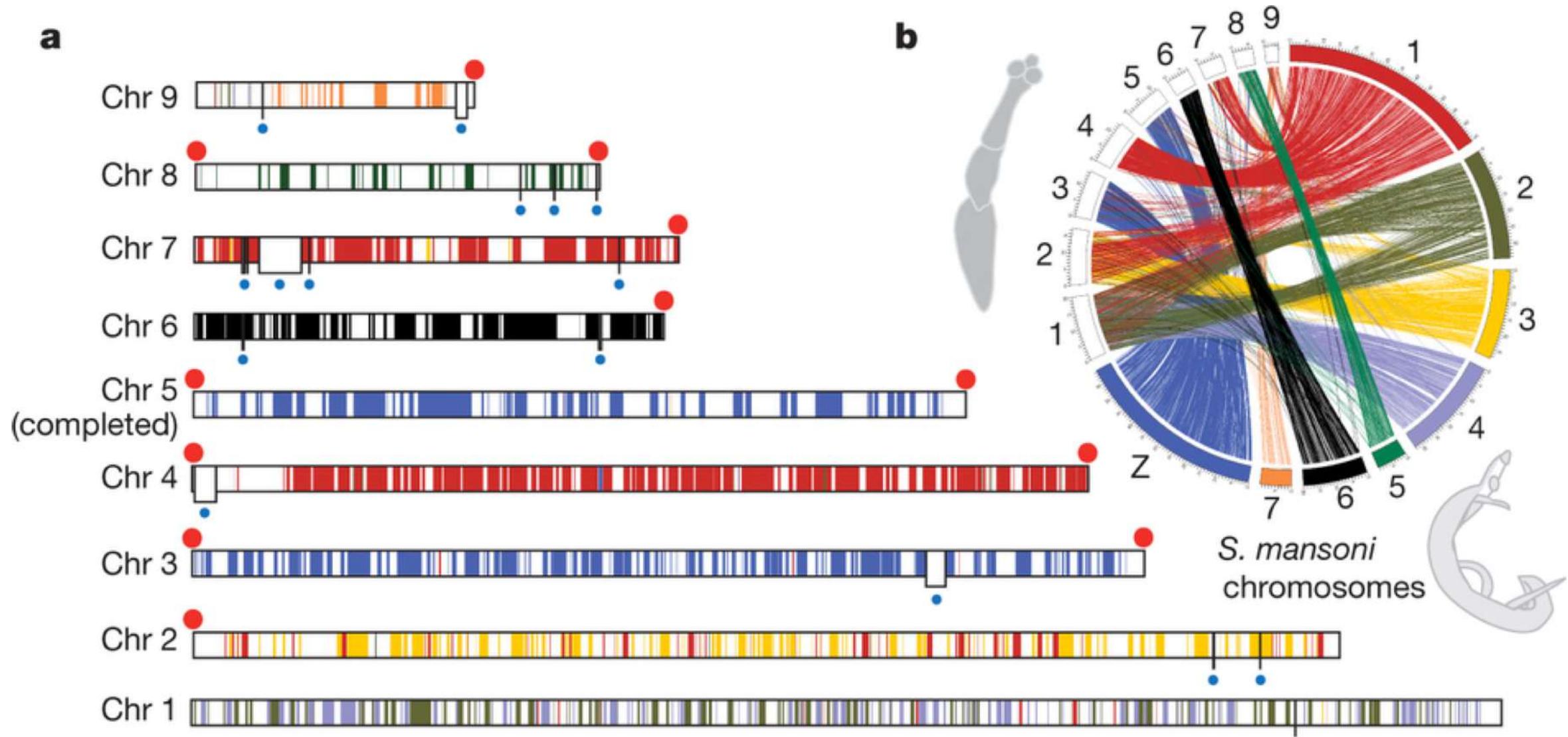
Free-living

Model

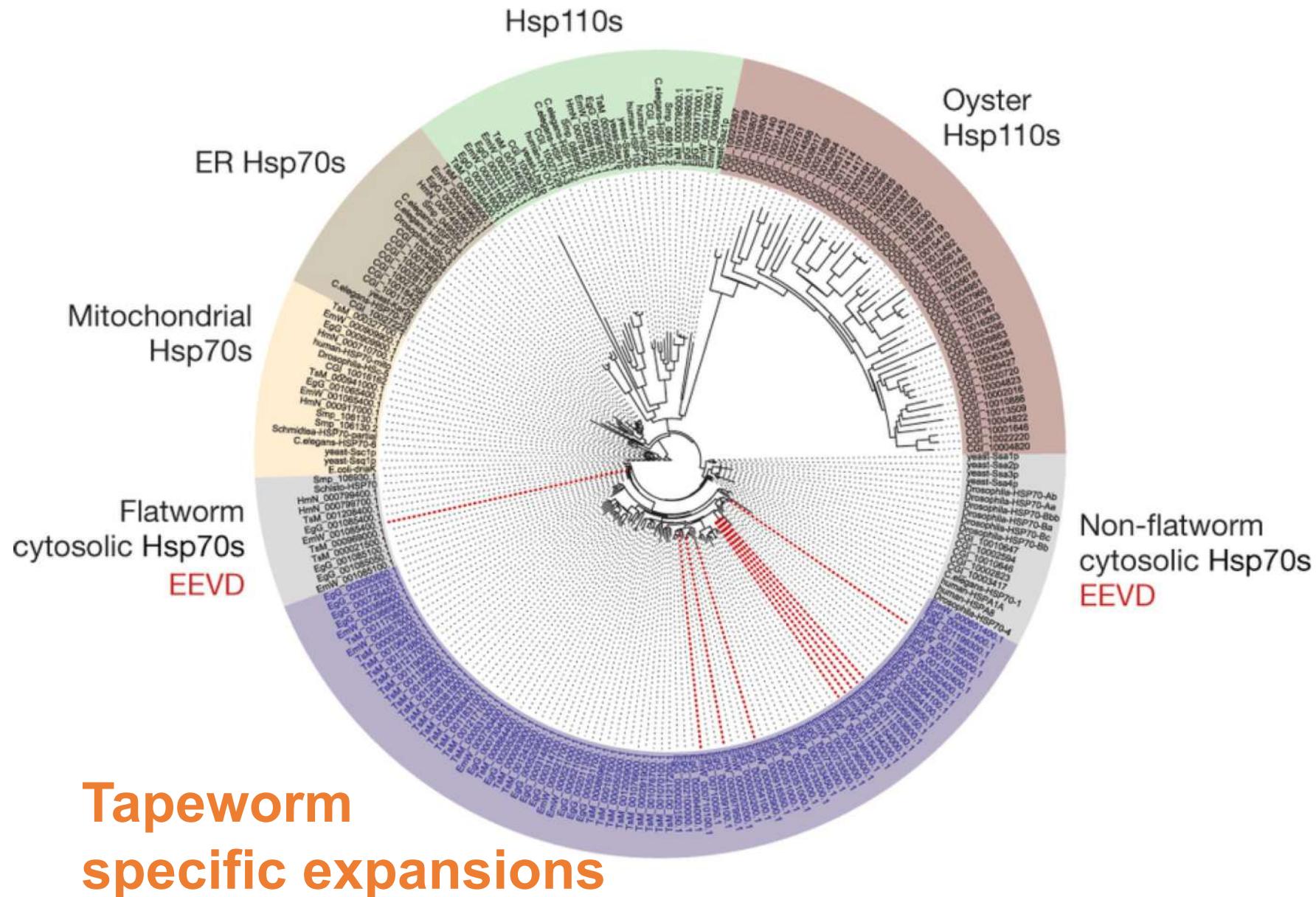


- A total of four tapeworm genomes were sequenced
- We compare with free-living and other parasite genomes
- ‘A route’ to complete parasitism

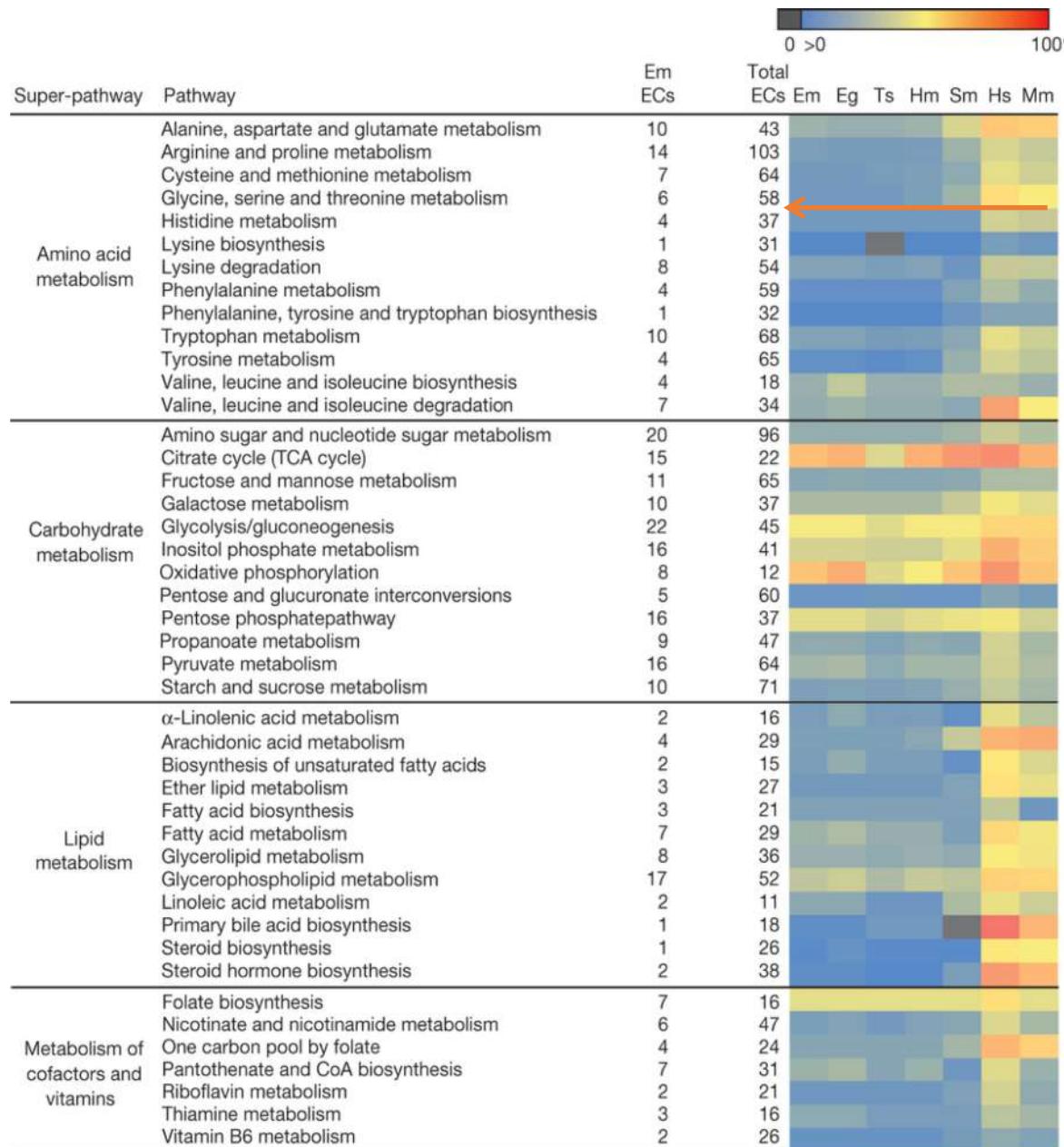
Genome of *E. multilocularis*



Heat shock protein expansion in tapeworms



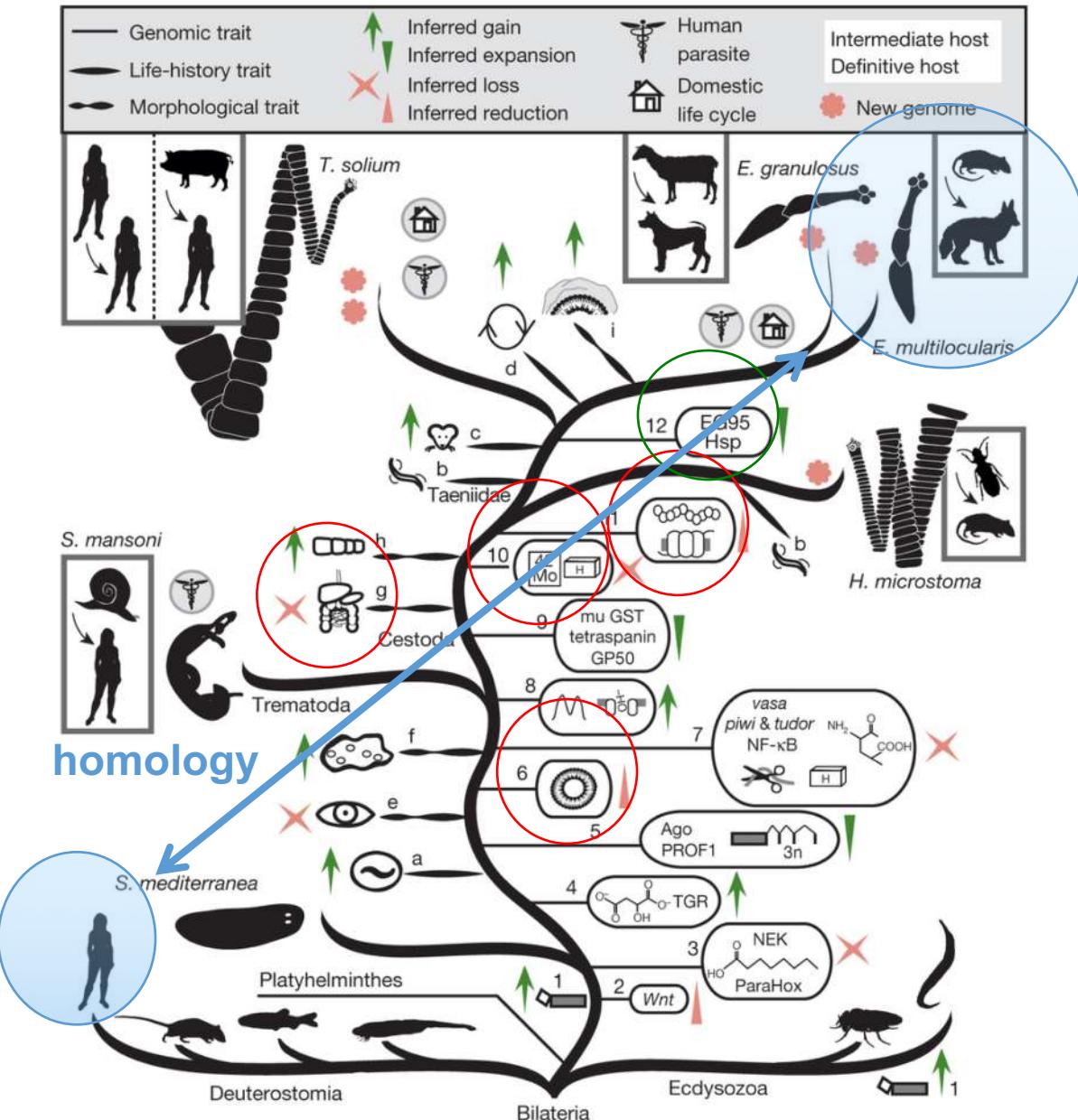
Reduced metabolism in tapeworms



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 Voltage-dependent calcium channel	Cytoskeleton Ion transport	M,A	Albendazole Praziquantel	406 277
	 Tapeworm cysts				
	 Second metastasis				
	(c) 2012, Richard M. Jakowski, DVM, PhD, DACVP				
	Elongation factor 2	Translation	M,A	Lorazepam)	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>

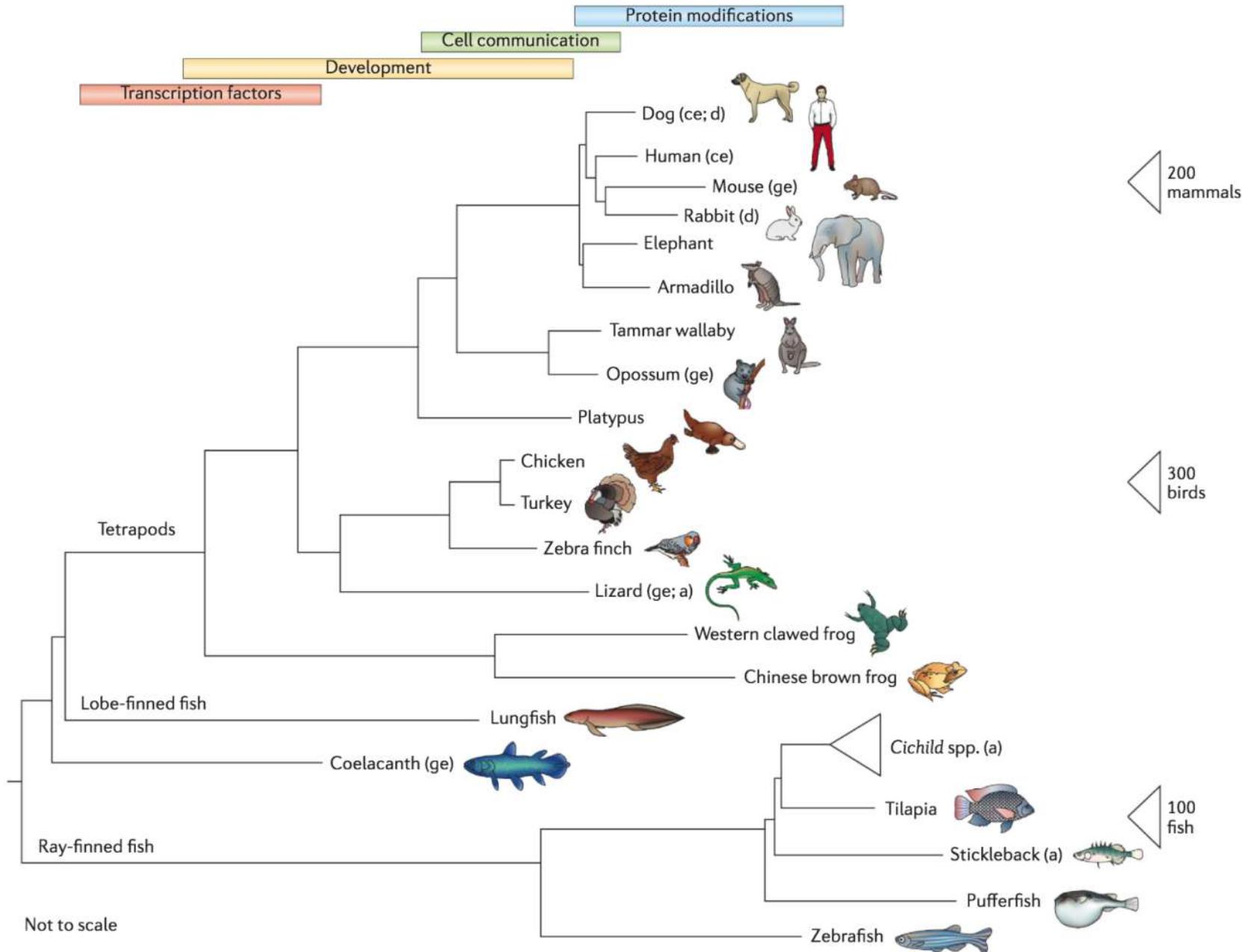
<http://ocw.tufts.edu/data/>

Good review – recent update

Dissecting evolution and disease using comparative vertebrate genomics

Jennifer R. S. Meadows¹ 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.



Looking back in 2003

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

Why comparative genomics? – A summary

- Duplication (genes, chromosome, segments, whole genome)
- Conservation (genes, chromosomes, segments);
- Specificity (species-specific genes);
- Inferring Paralogs, orthologs;
- Families (clusters) of paralogs, of orthologs;
- Shared motifs in clusters of paralogs, orthologs;
- Protein conservation profiles;
- Gene Transfer, introgression between species;

How genome evolved;
How genome functions

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

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