

Comparative Genomics

Isheng Jason Tsai

Introduction to NGS Data and Analysis
Lecture 9

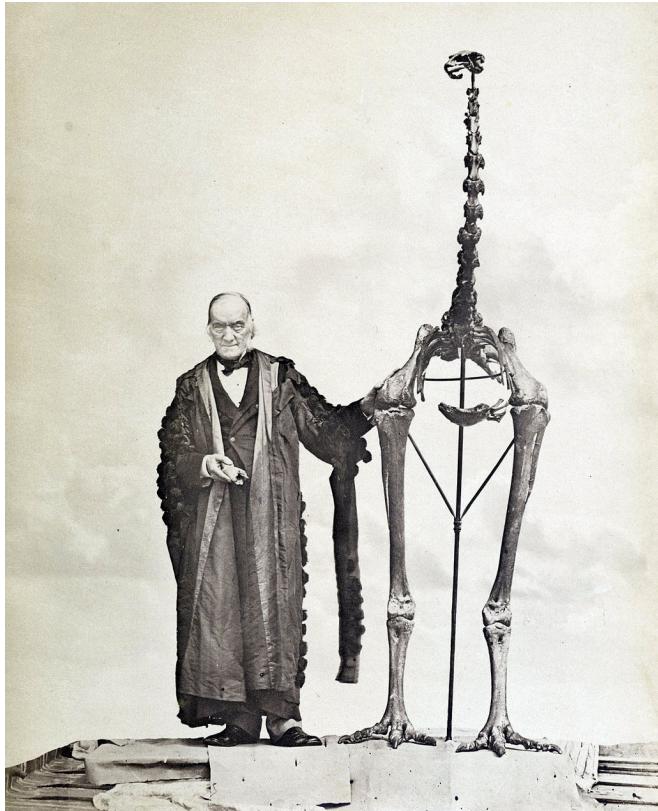


For this lecture, I gathered information from :

- Reviews
- Slideshare (with permission)
- http://webext.pasteur.fr/tekaia/BGA_courses.html
- <http://compbio.mit.edu/lectures.html>
- <http://www.molecularevolution.org/workshops>

Homology

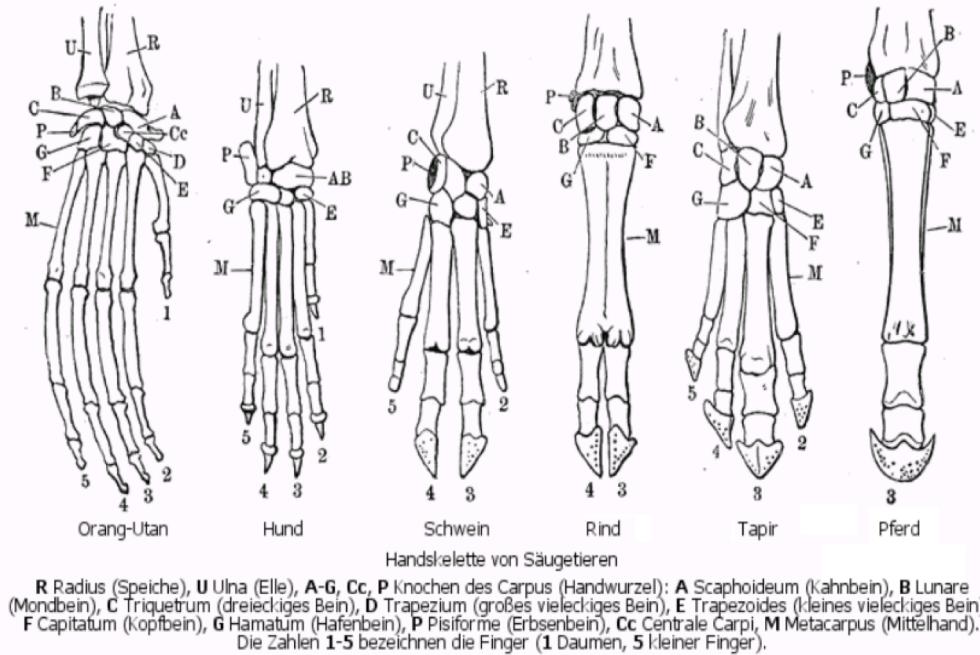
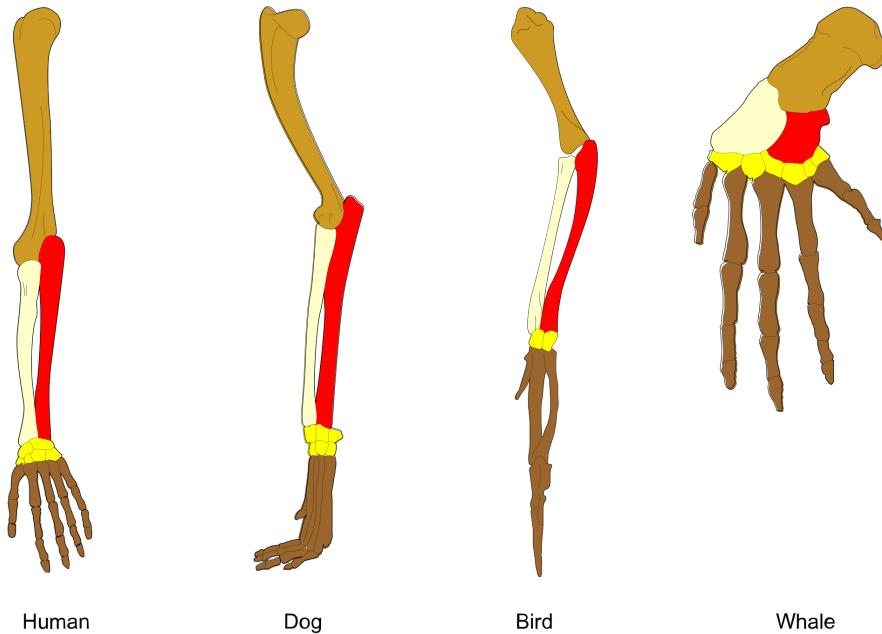
Termed before Darwin's time!



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

Wiki

Homology

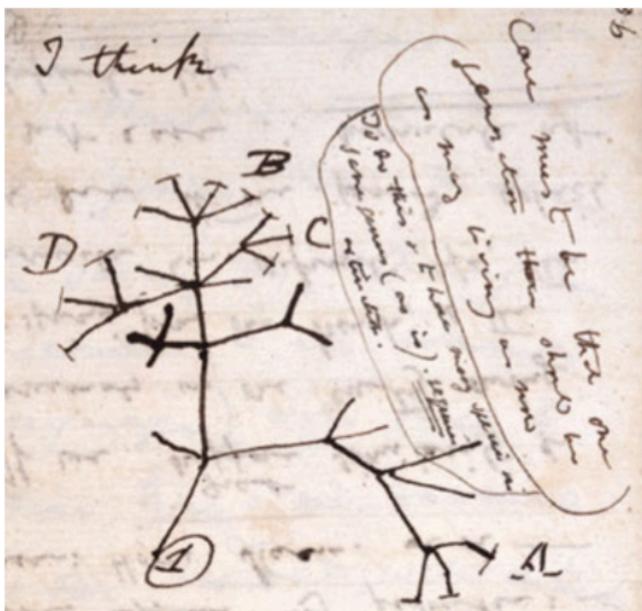


“the same organ in different animals under every variety of form and function” – Richard Owen

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—Classification.

OPEN  ACCESS Freely available online

PLOS BIOLOGY

Historical and Philosophical Perspective

Darwin's Theory of Descent with Modification, versus the Biblical Tree of Life

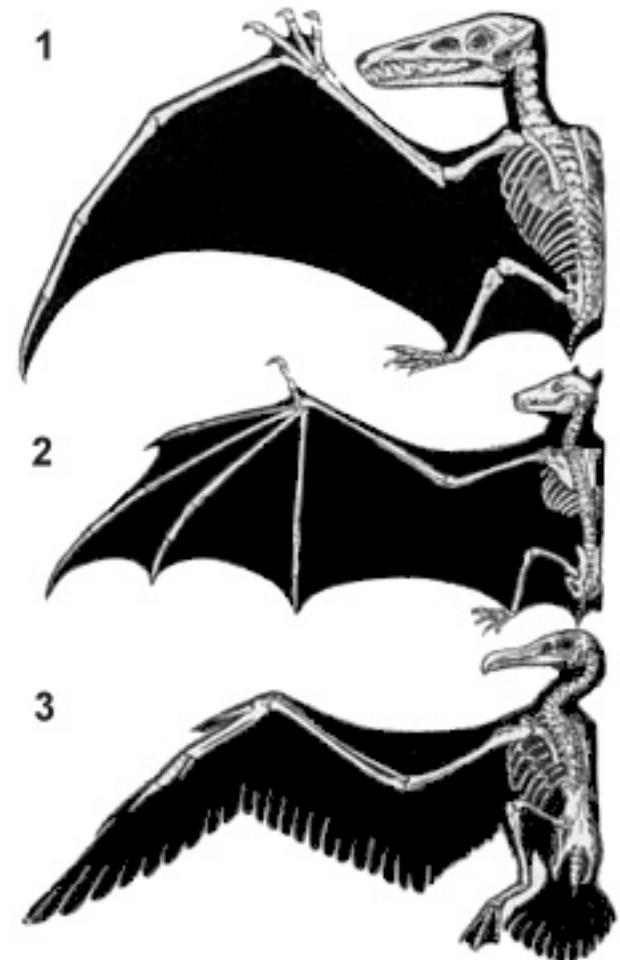
David Penny*

Institute for Molecular BioSciences, Massey University, Palmerston North, New Zealand

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**



DISTINGUISHING HOMOLOGOUS FROM
ANALOGOUS PROTEINS (1970)

WALTER M. FITCH



1929 - 2011

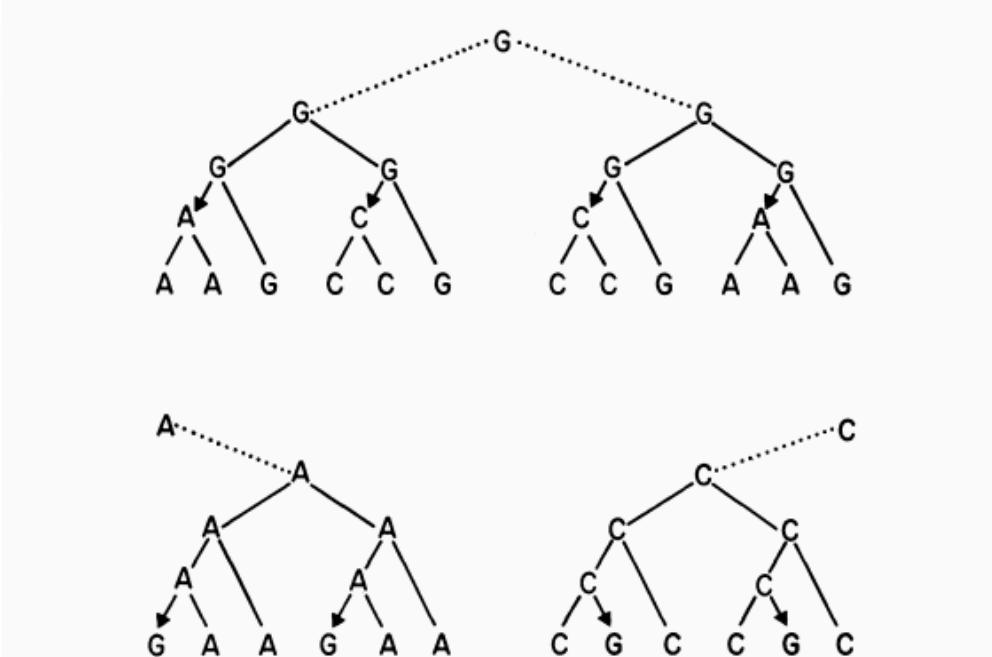
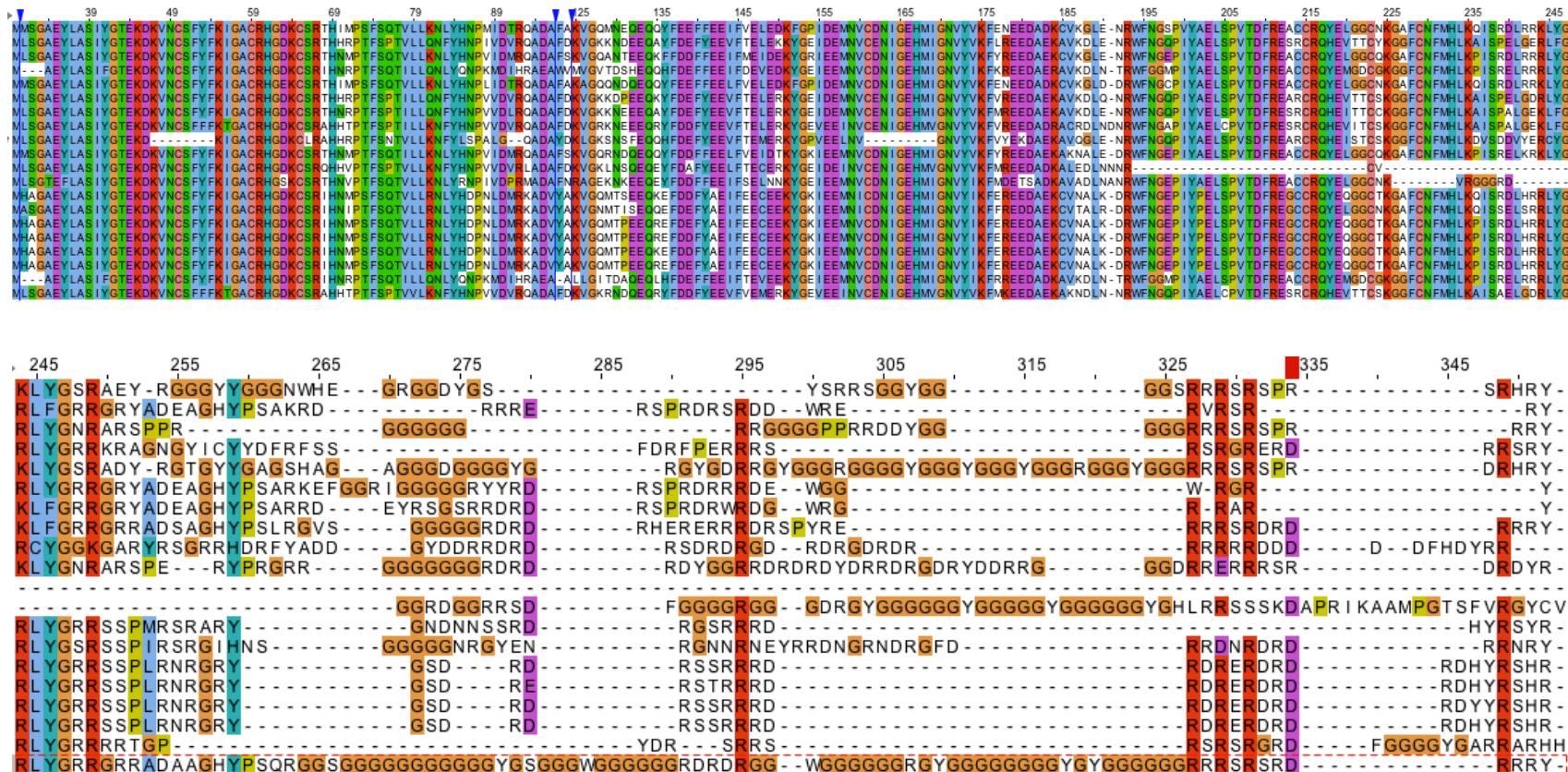


FIG. 1.—Distinguishing convergent from divergent types of nucleotide replacement patterns. Given are two groups of species (related within each group as shown by the solid lines) together with the nucleotide present at a specific position of the gene for each member species as shown at the branch tips.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

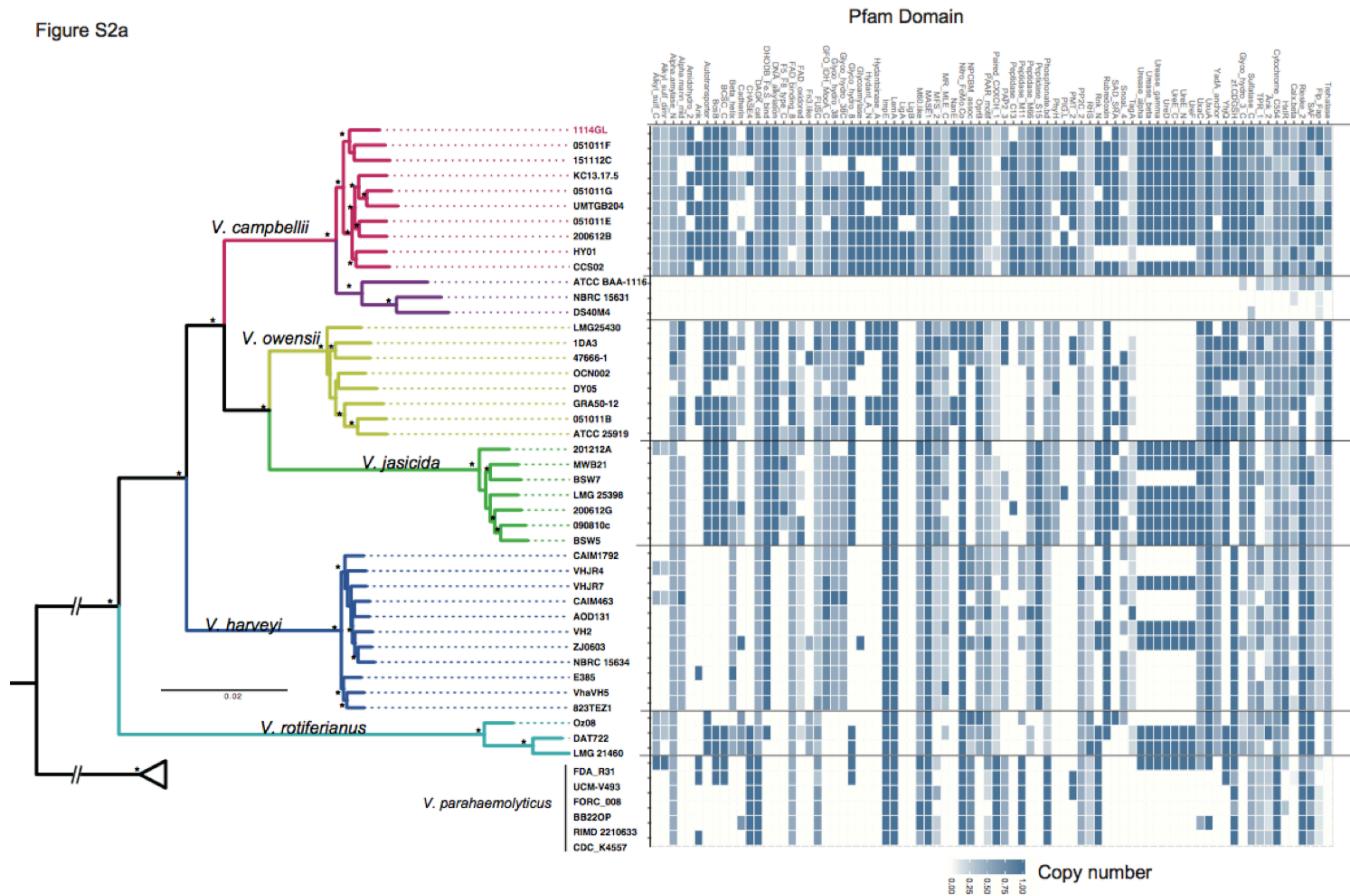
Two sequences are homologous if they share the same a common ancestor



Extension of homology to genomes / species

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

Figure S2a

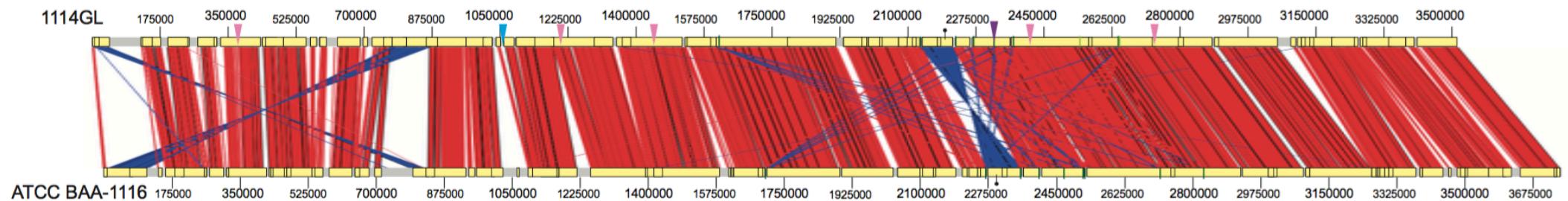


Hueimien Ke

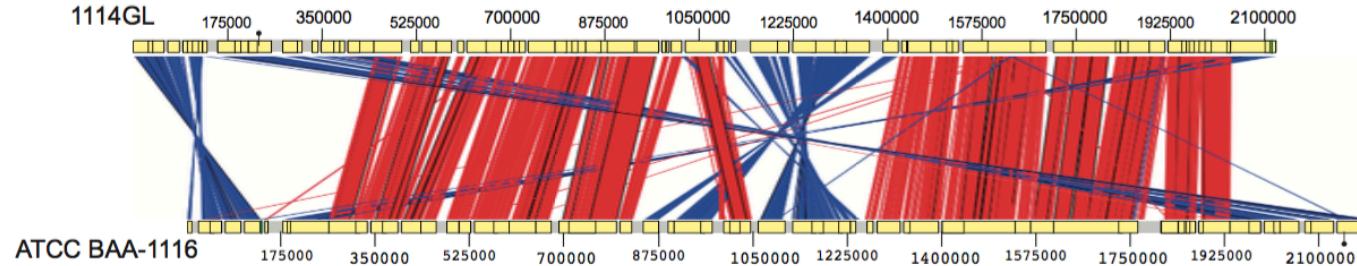
Extension of homology to genomes / species

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

Why comparative genomics?

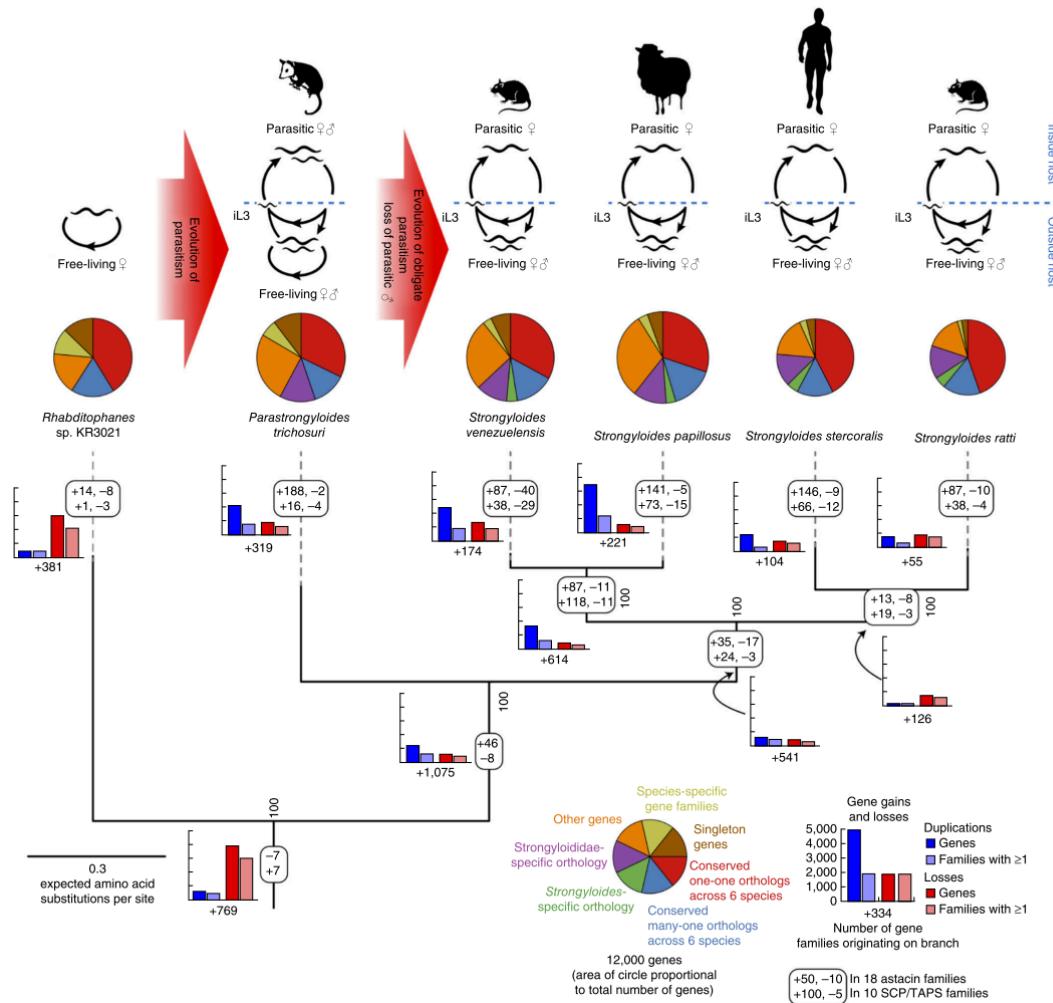
Compare more than 2 genomes

Similarity and differences between genomes

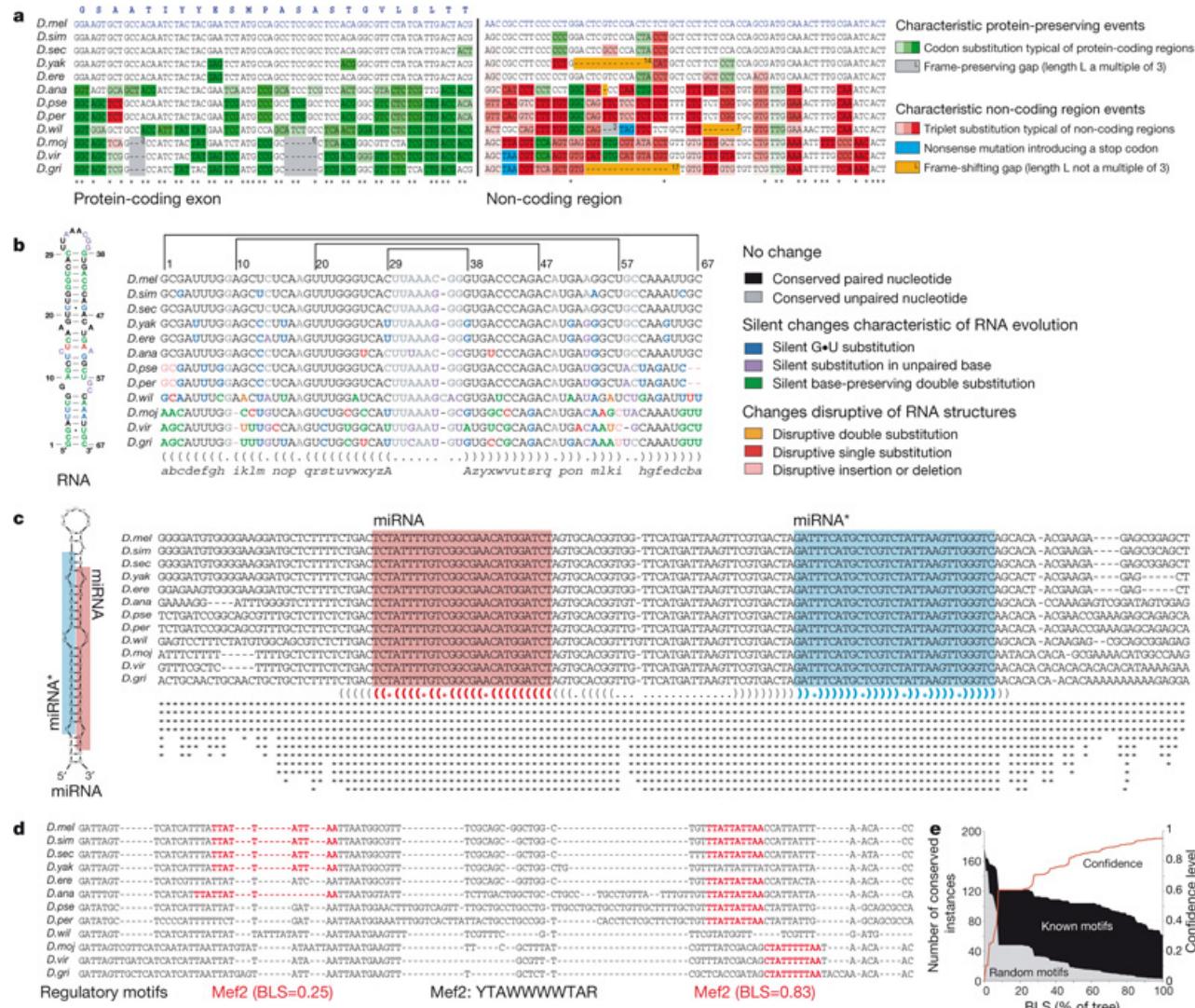
Use genomes to study evolution of these species ;

Use evolution to study the genomes of these species

Reveal the evolutionary relationships among species



Link evolutionary processes with function



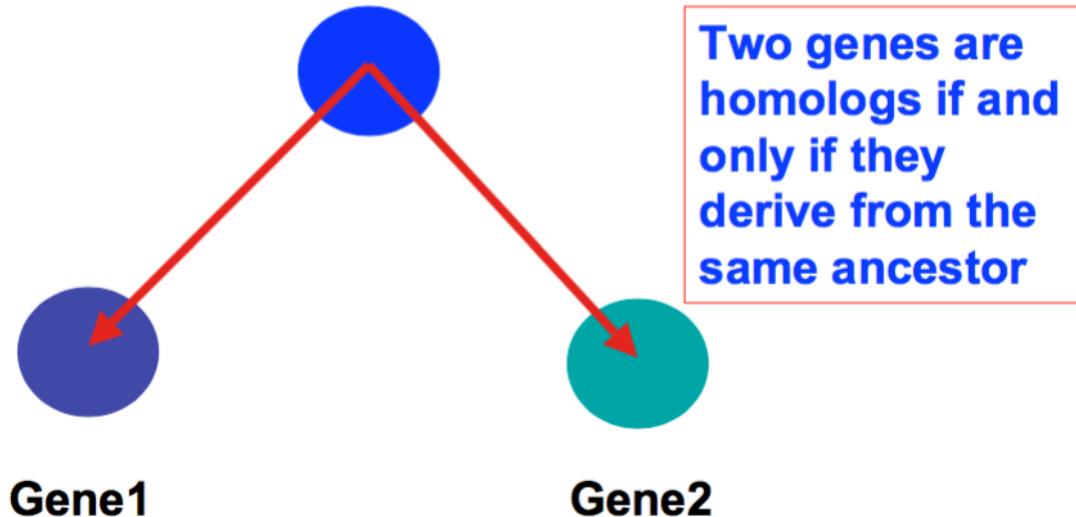
Clark et al (2007)

Search for similarity

Search for similarity

One of the most frequent activity in Bioinformatics

Common ancestor



Gene1

Gene2

Homology is almost uniquely inferred by sequence similarity

Beware ; why?

~~**Significant homology**~~

~~**Weak homology**~~

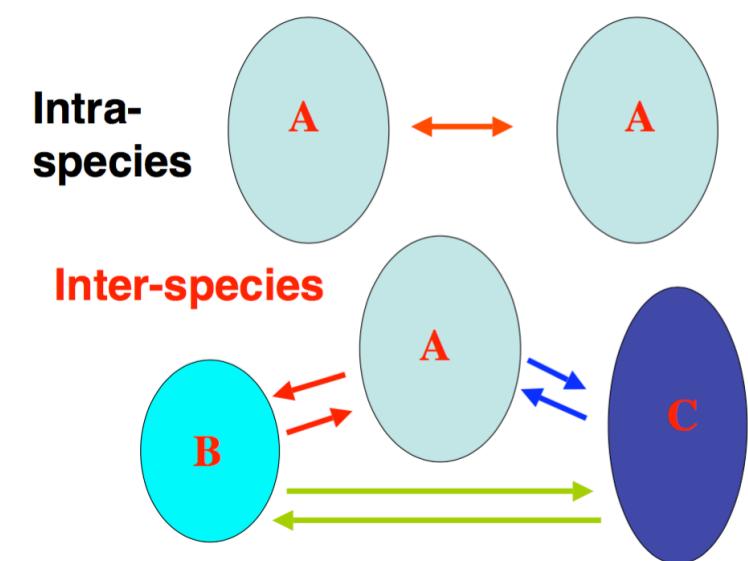
If you think about the
meaning of homology,
**then it really makes no
sense**

Significant similarity

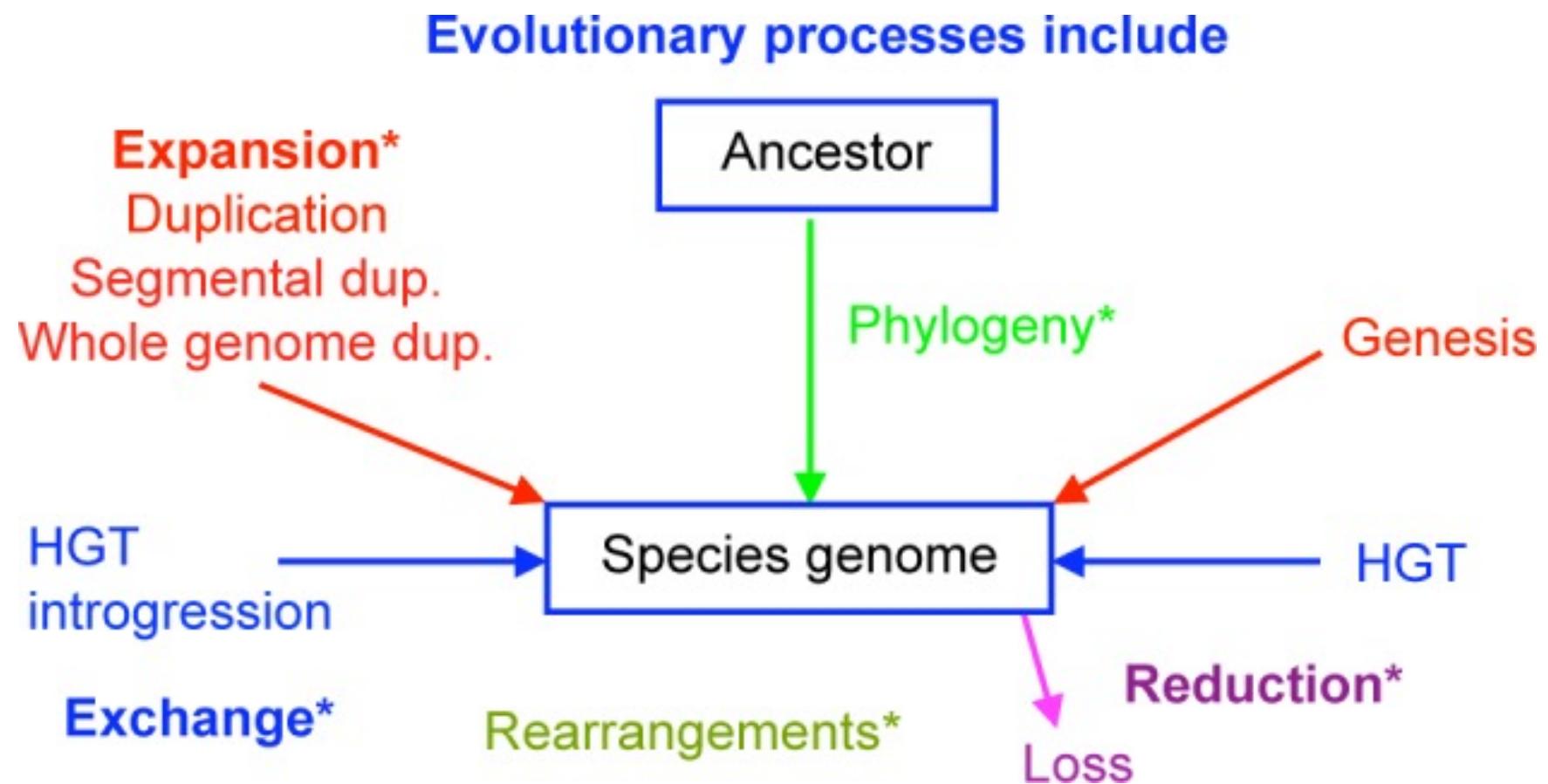
Weak similarity

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)



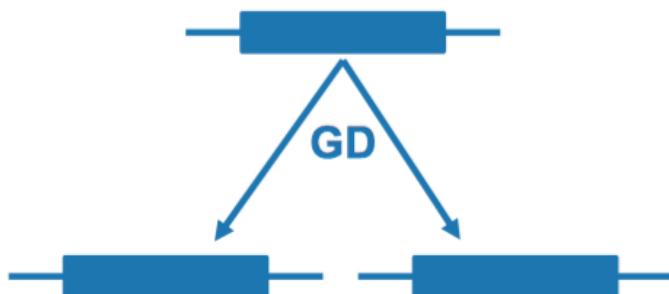
Evolution process of a genome



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

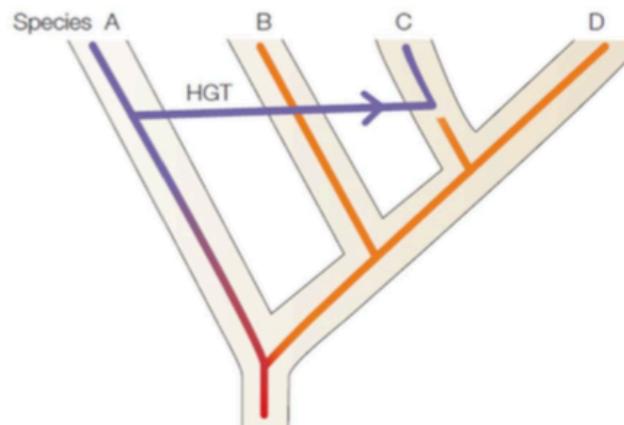
Gene duplication (GD)

Any duplication of a region of DNA that contains a gene



Horizontal gene transfer (HGT)

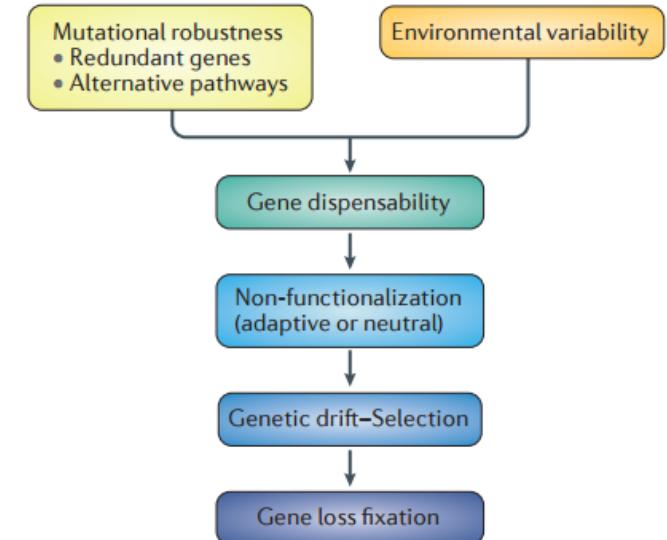
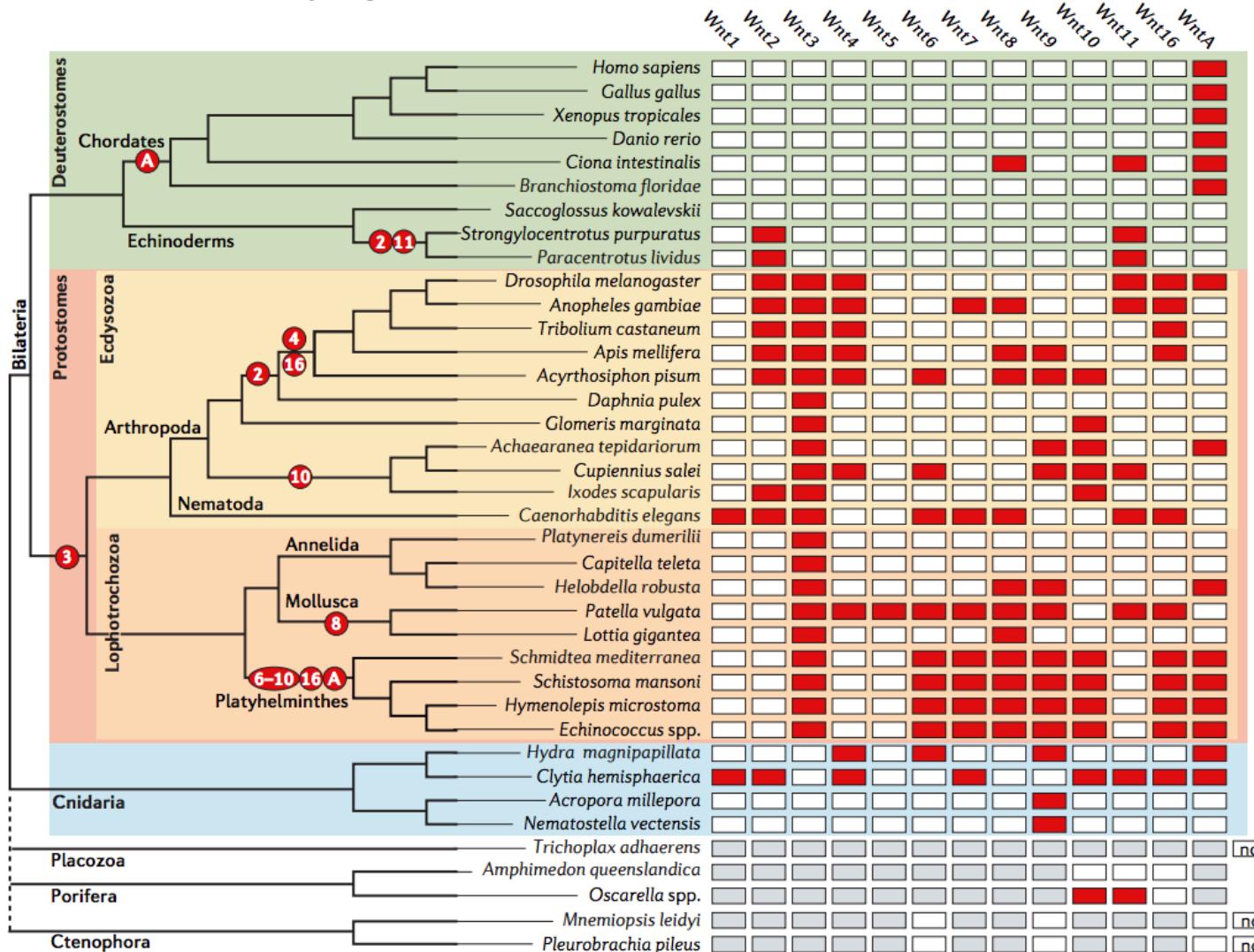
Exchange of genes between organisms other than through reproduction



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

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

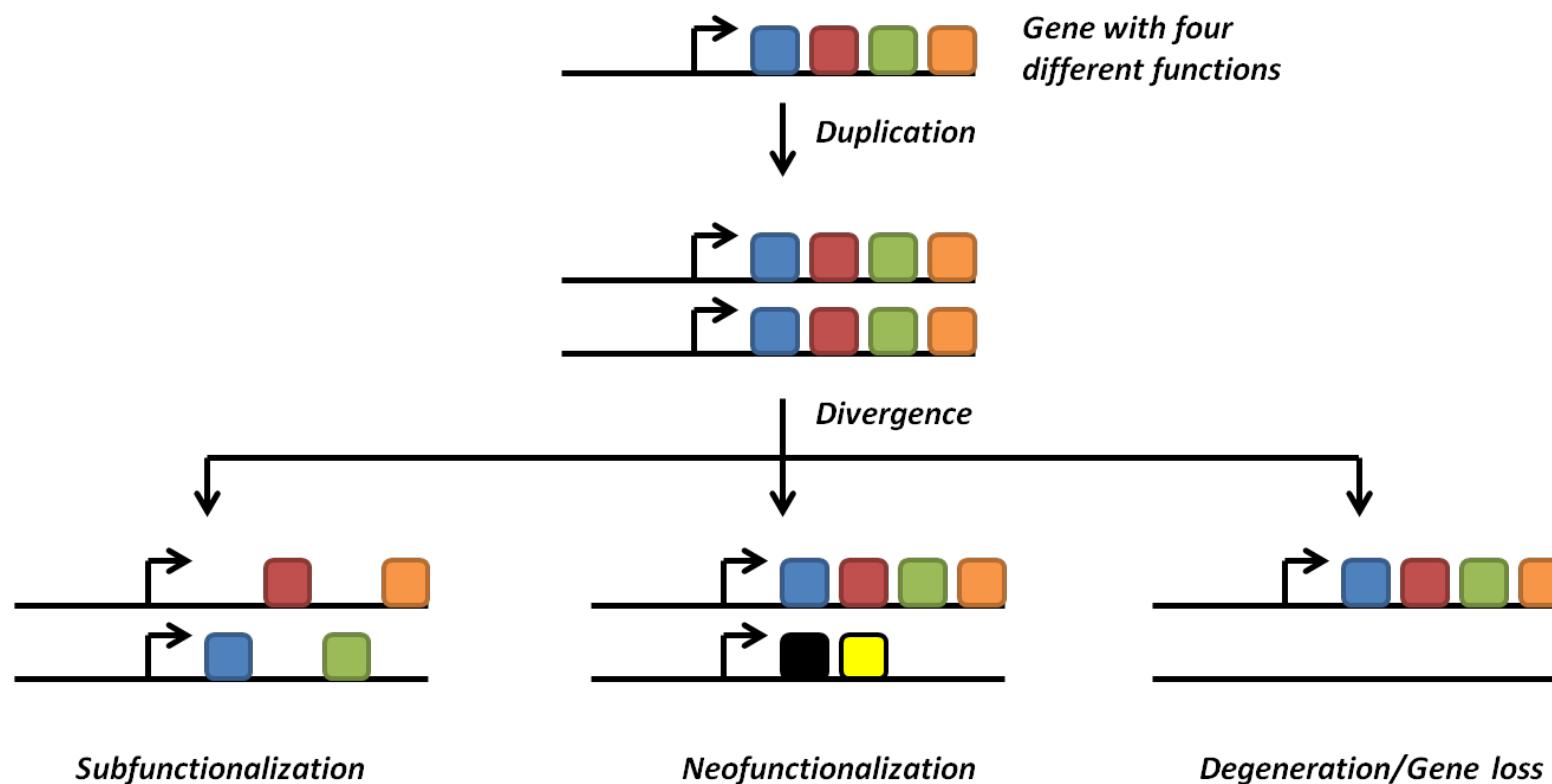
Evolution by gene loss



Albalat and Cañestro (2016)

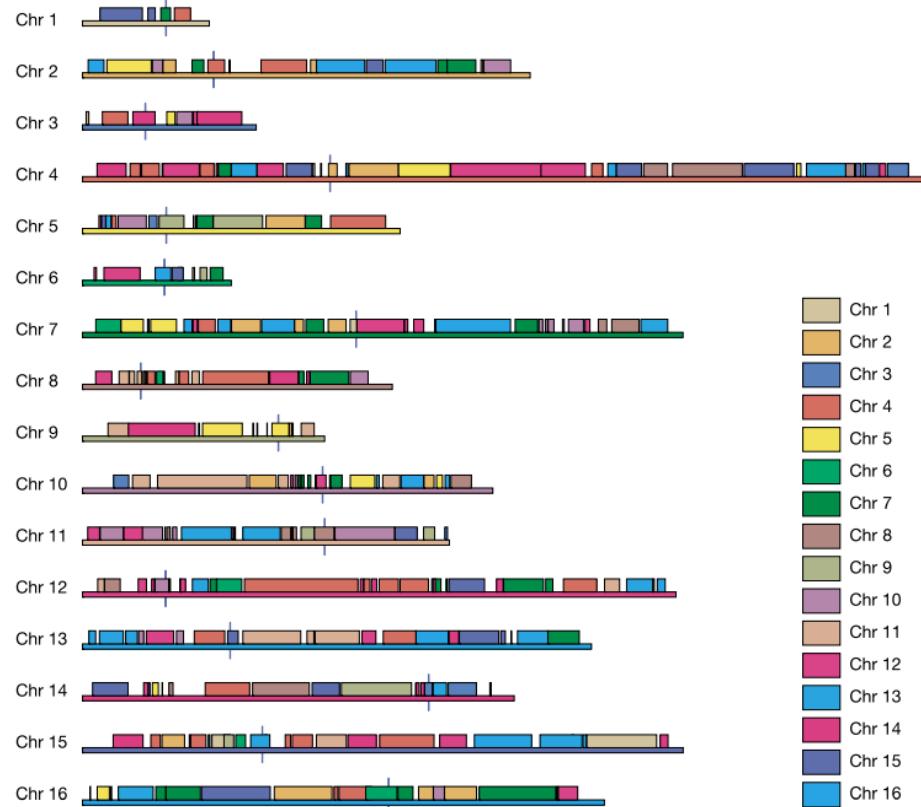
Duplication

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



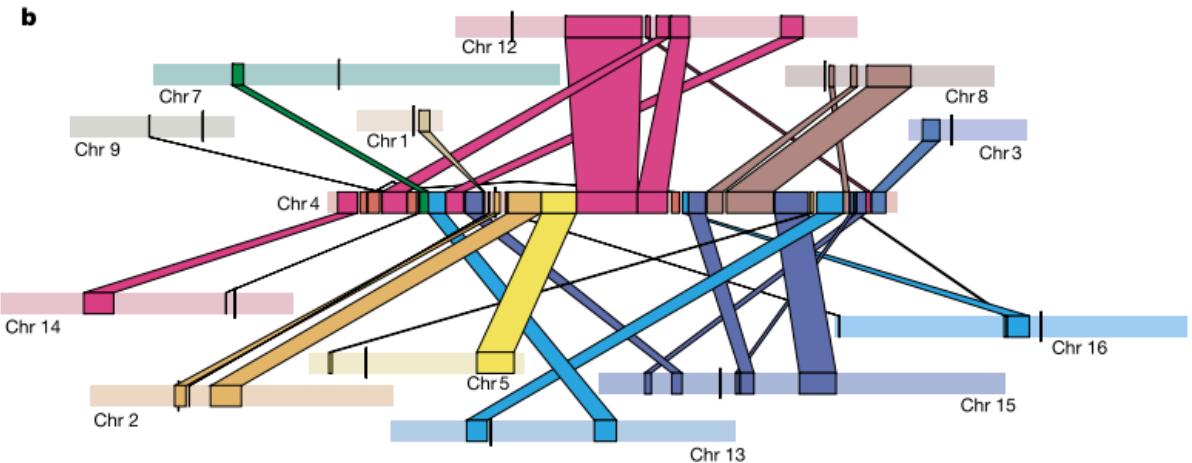
Wiki

Within species



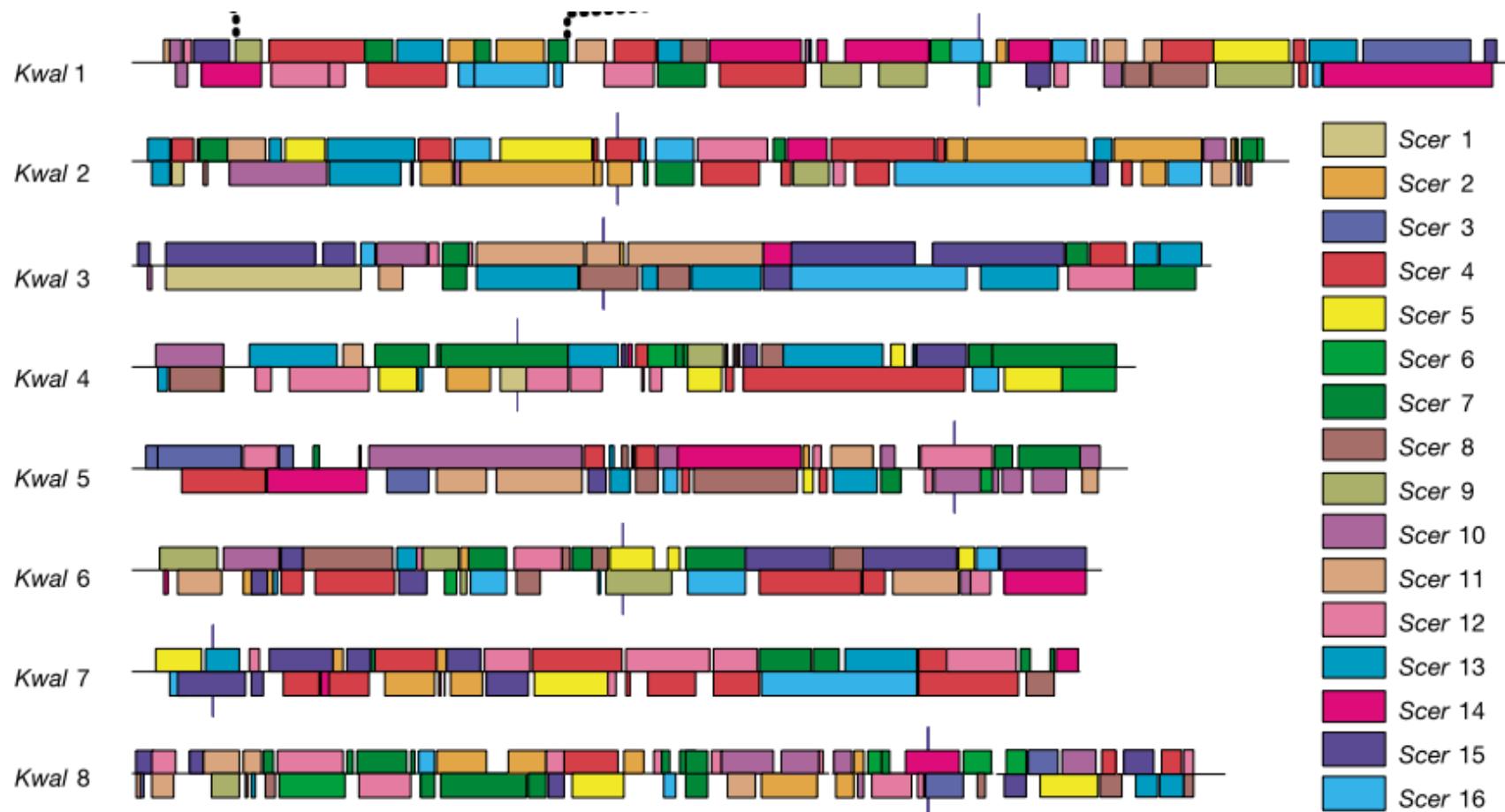
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16

b



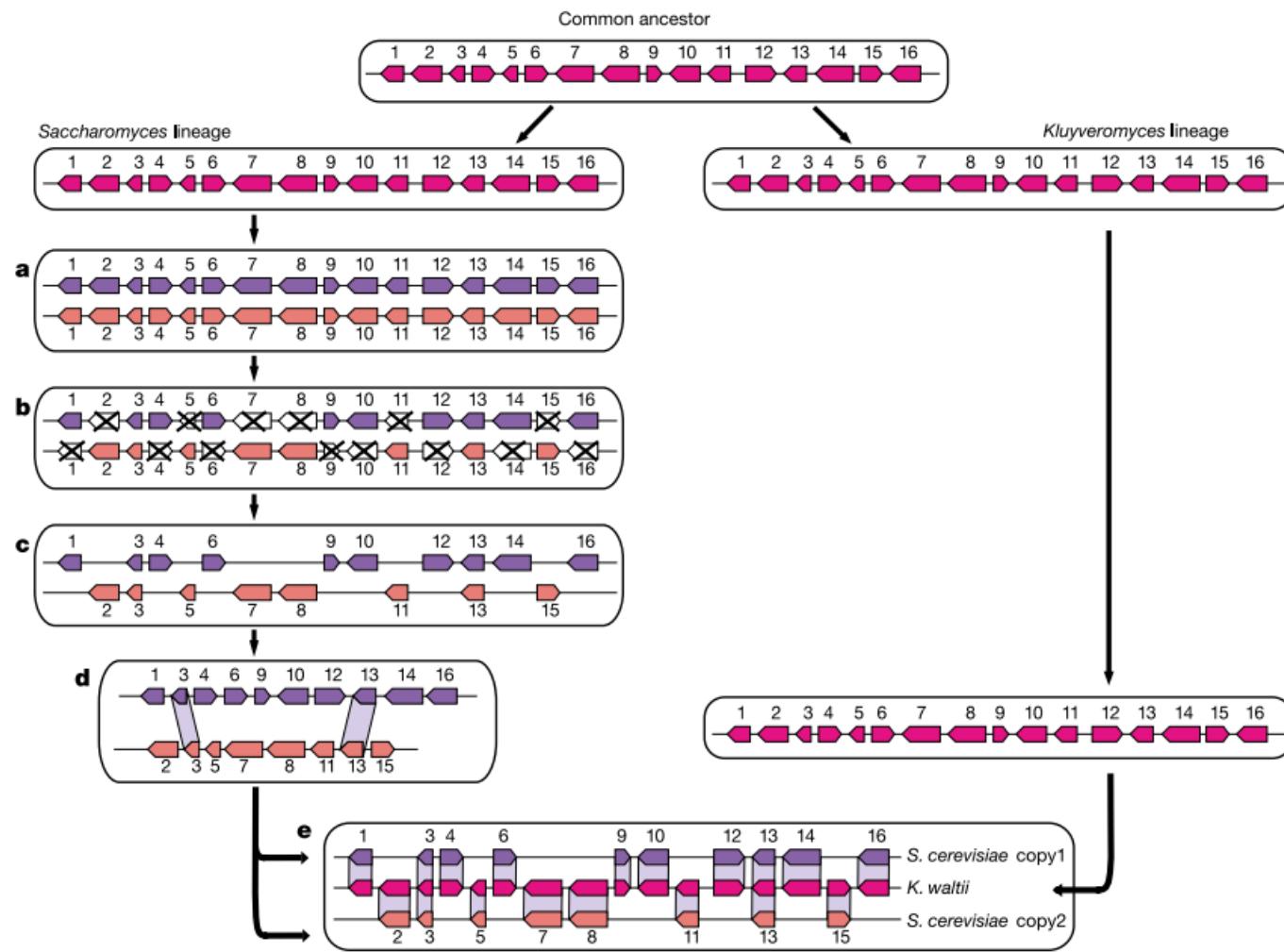
Kellis et al (2004)

Between species



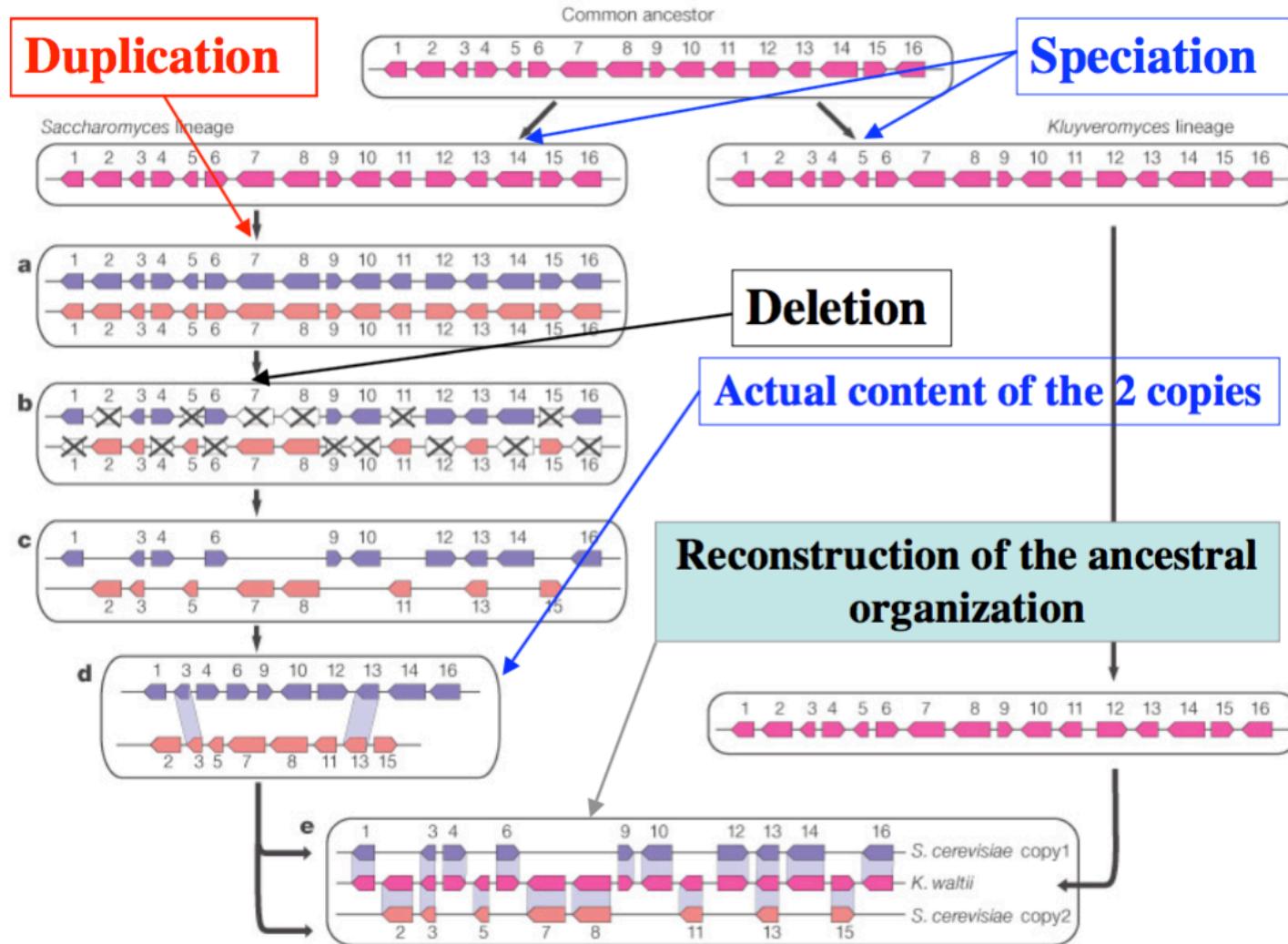
Kellis et al (2004)

Whole genome duplication model



Kellis et al (2004)

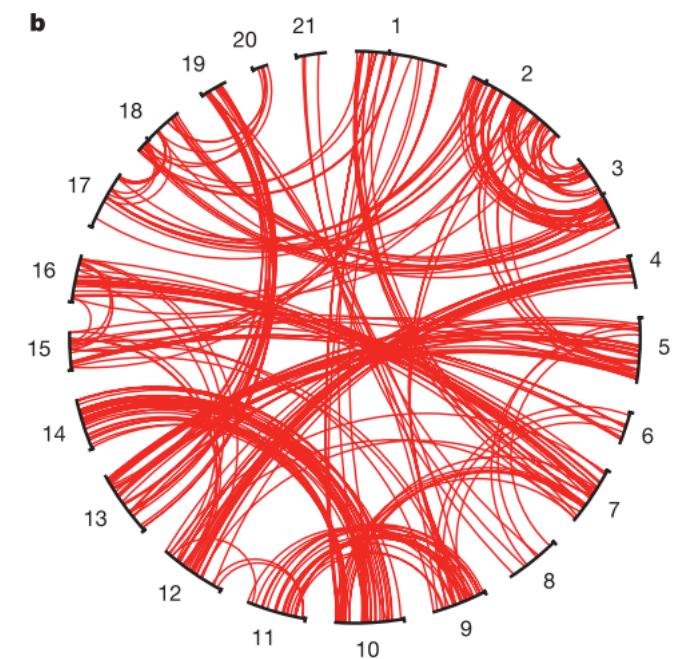
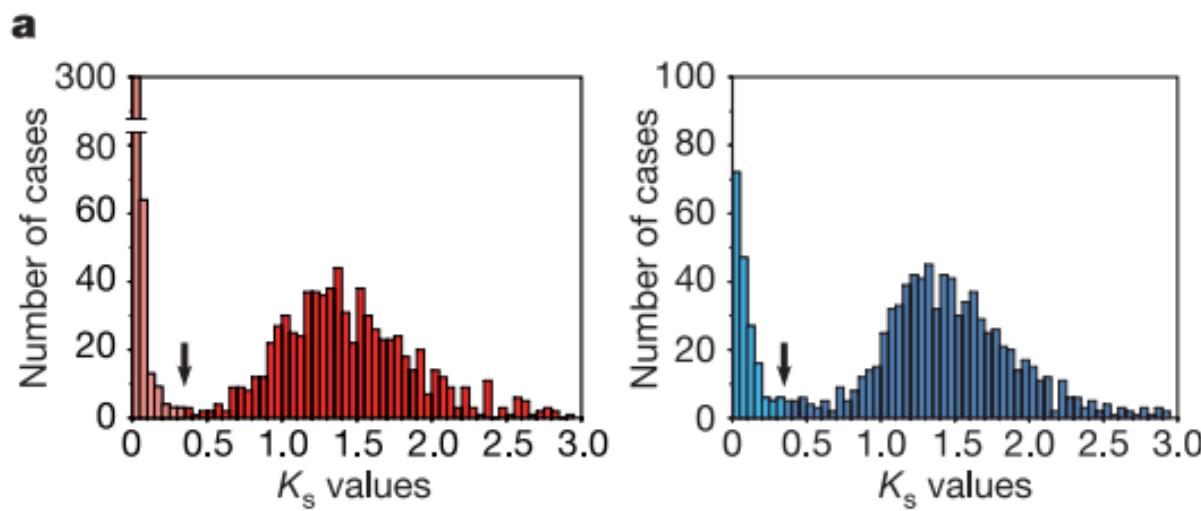
Determining ancestral conservation



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

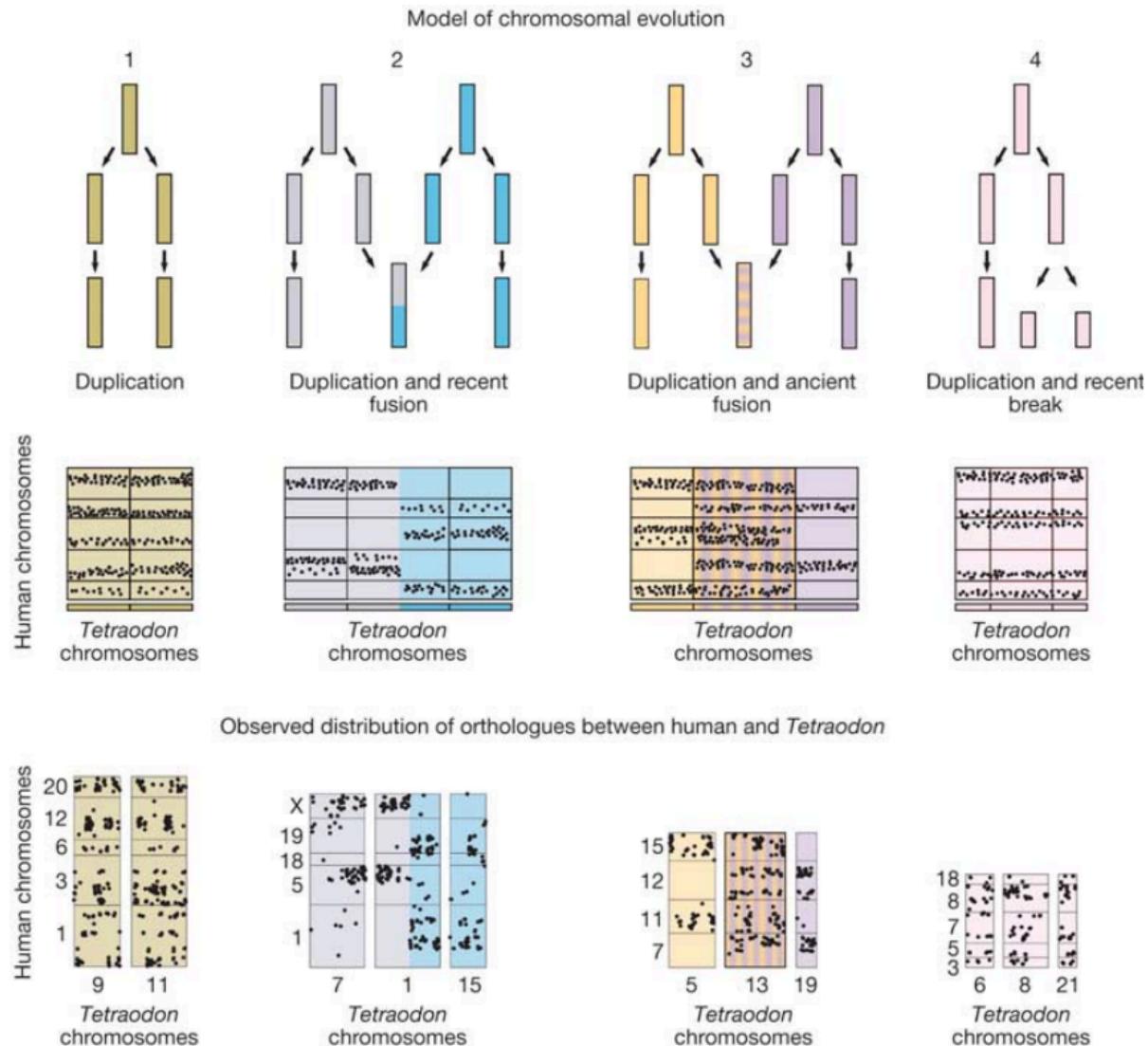
Slide of Fred Tekaia

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

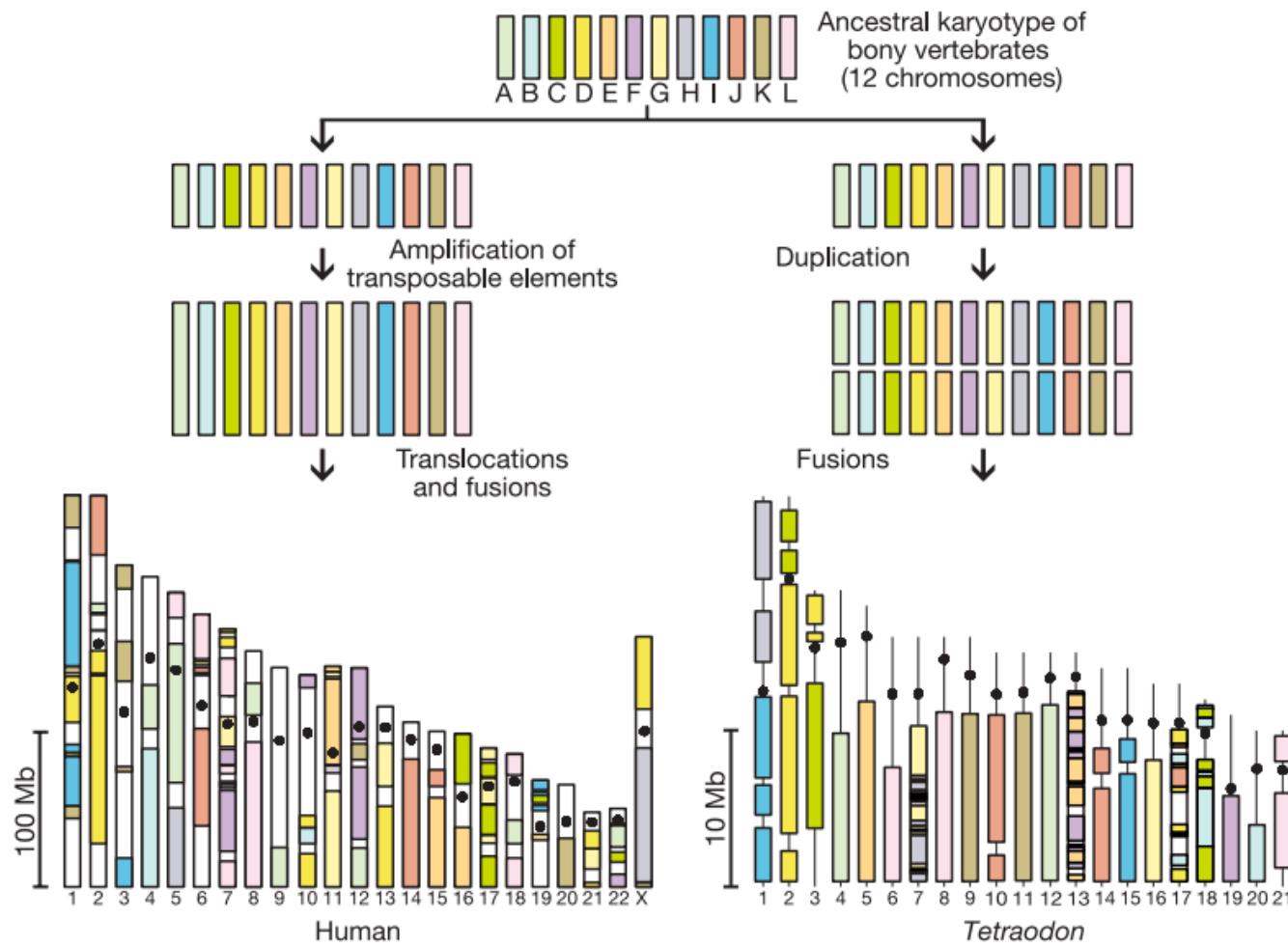


Jaillon et al (2004)

Reconstructing ancient genome rearrangement

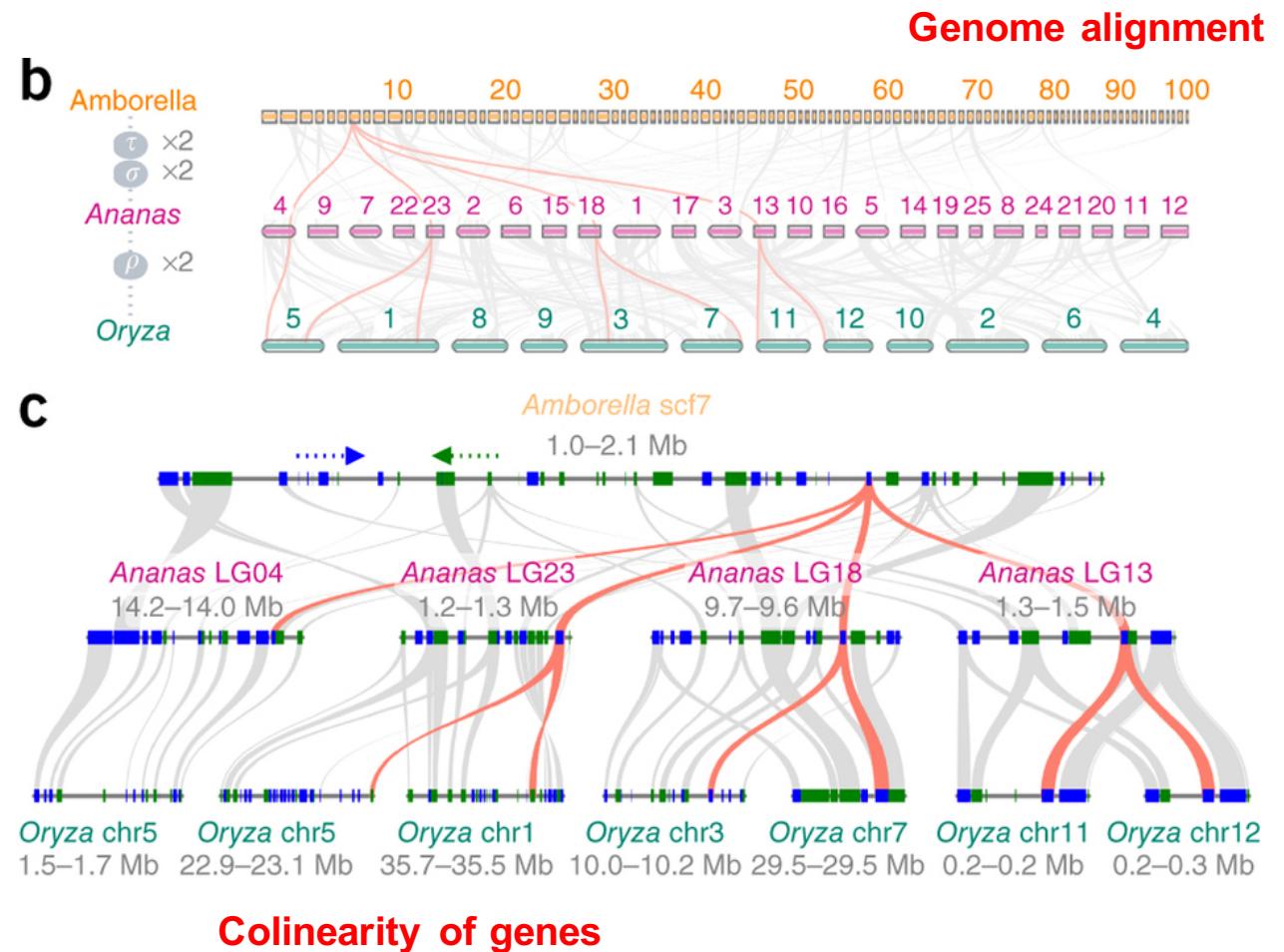
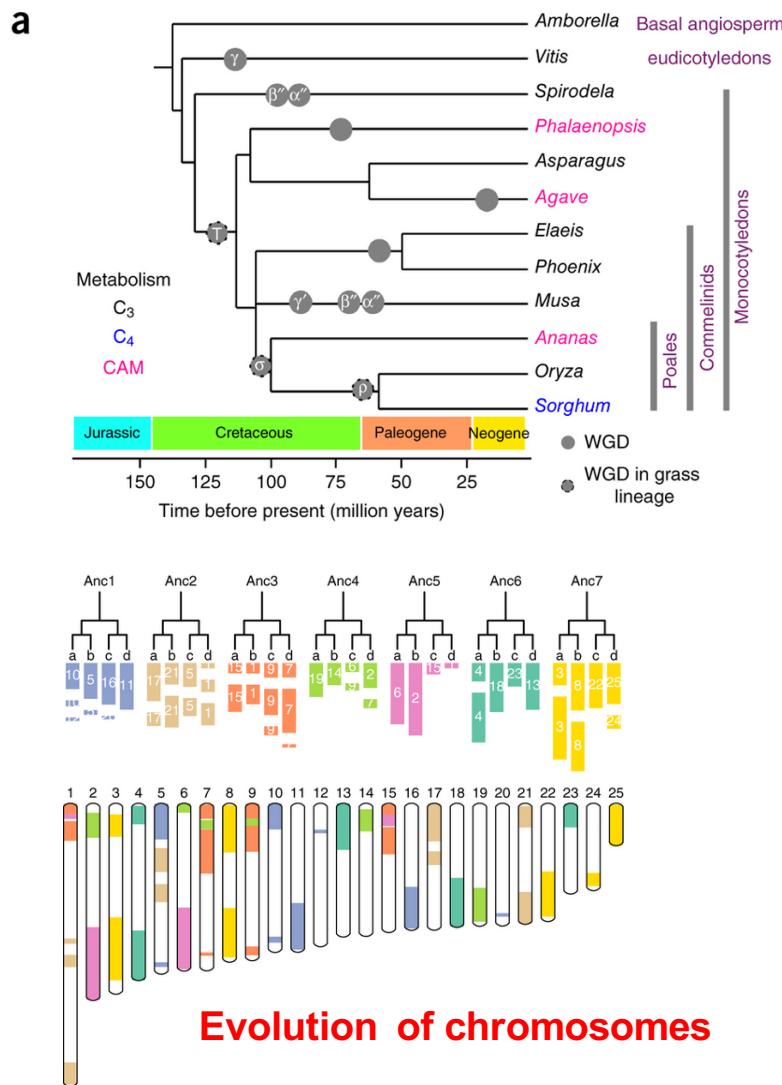


Reconstructing ancient genome rearrangement



Jaillon et al (2004)

Pineapple genome



Ming et al (2015)

We look for

- 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

Orthology

From homology to orthology

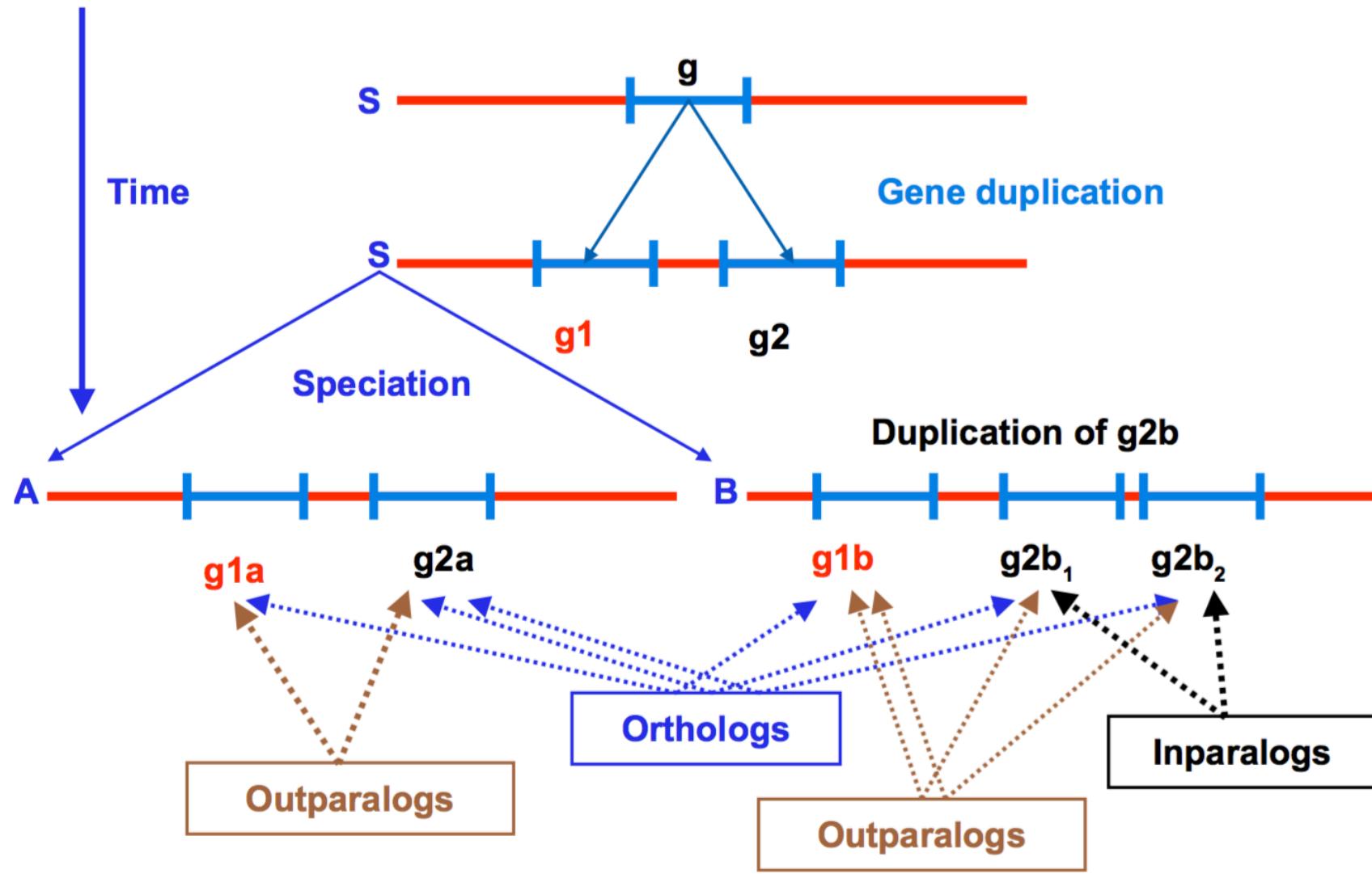
Homologues are sequences derived from a common ancestor...

- 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)."*



Tekaia (2016)

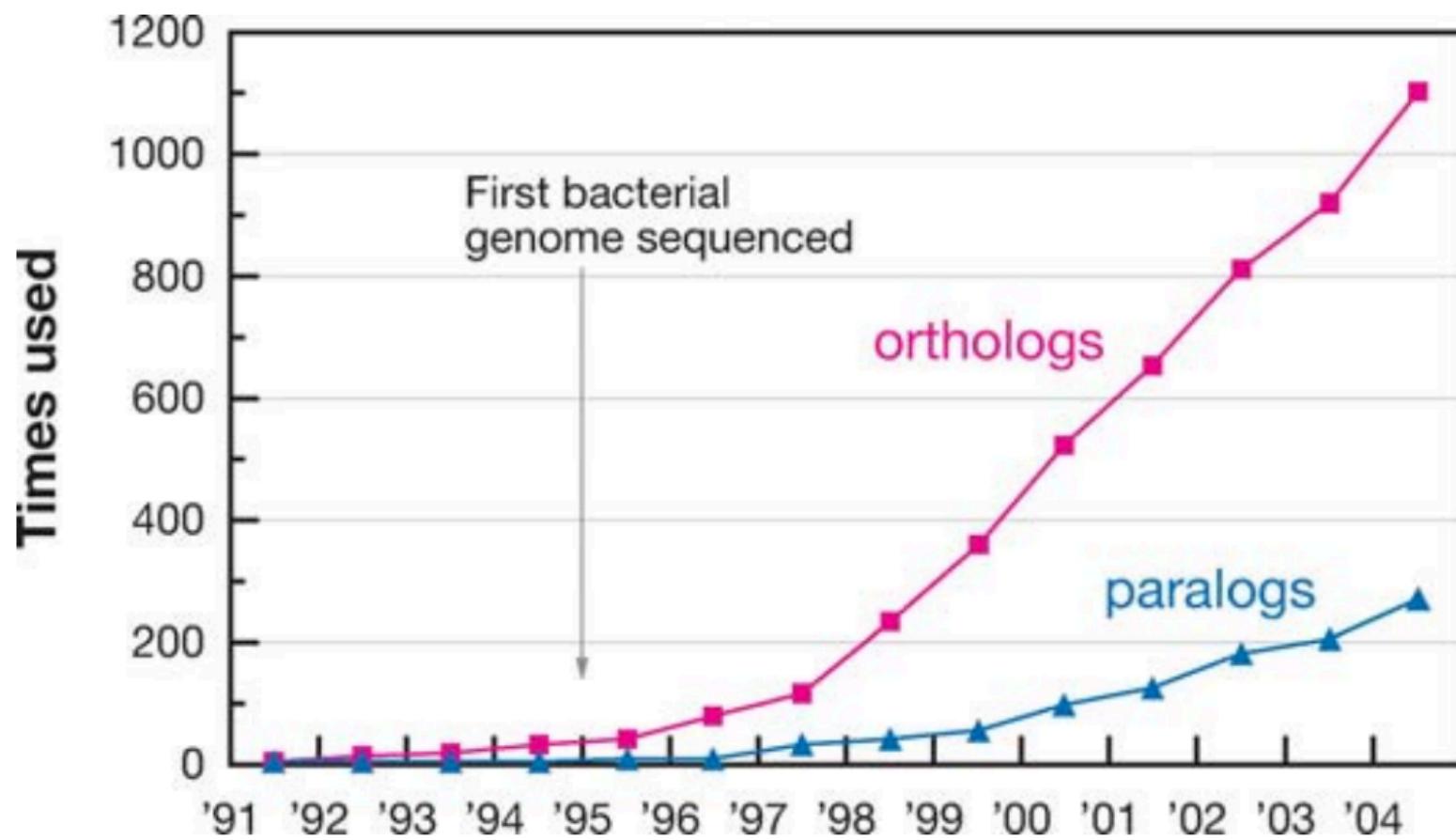
Why is orthology important?

Orthologs detection is of fundamental importance in:

- Reconstruction of the evolution of species and their genomes (Phylogenomics);
- Evolutionary studies of biological systems;
- Annotation of newly sequenced organisms;
- Functional genomics (transfer of functional annotation predicted on “orthology-function conjecture”);
- Gene organization in a given species.

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

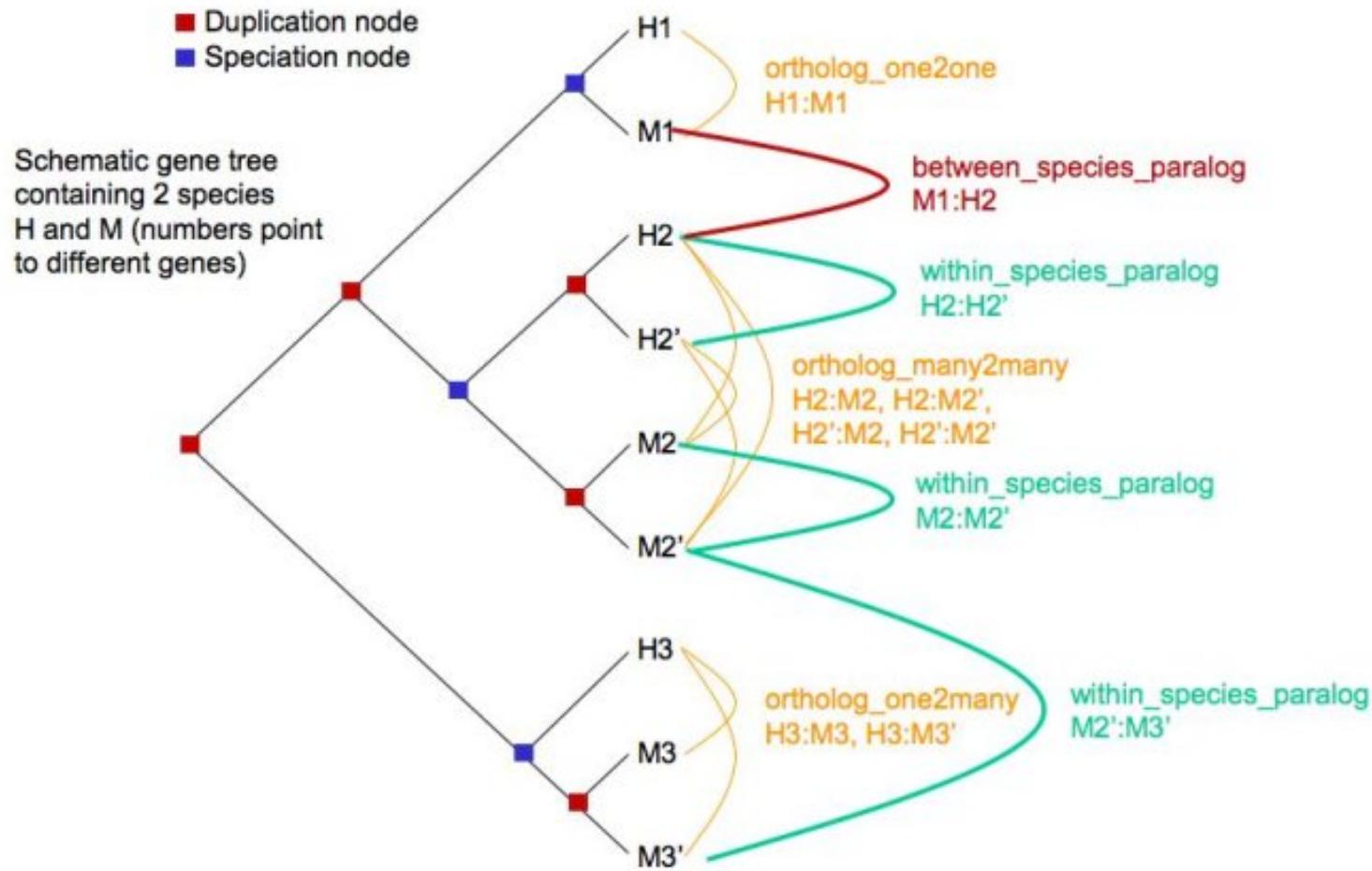
Usage of “ortholog” and “paralog”



Koonin (2005)

Corollary

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



http://public-contigbrowser.sigenae.org:9090/info/data/compara/homology_method.html

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

More precise definitions

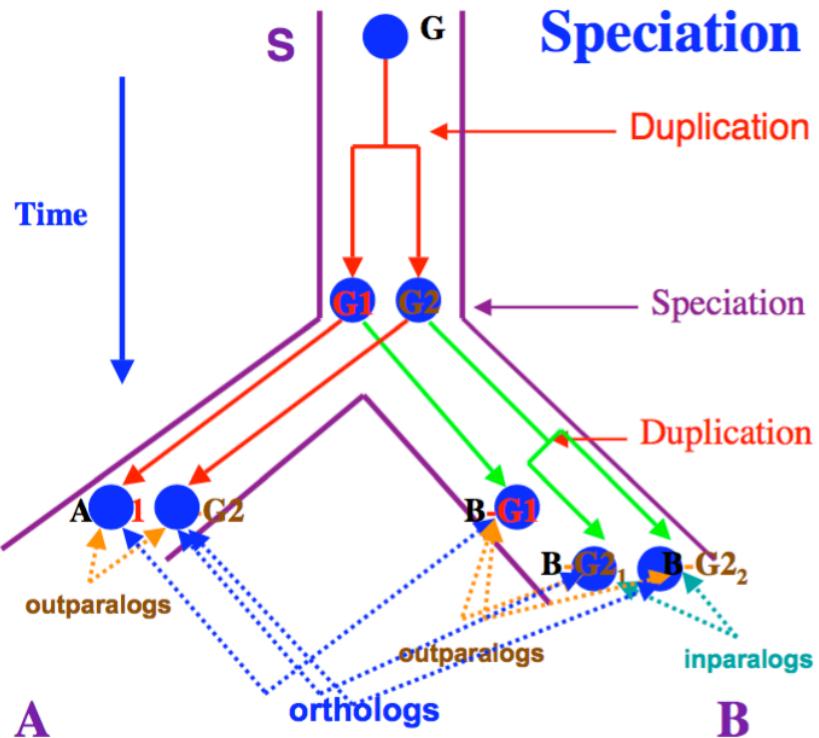


Table 1 Homology: terms and definitions

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.

Ortholog inference methods

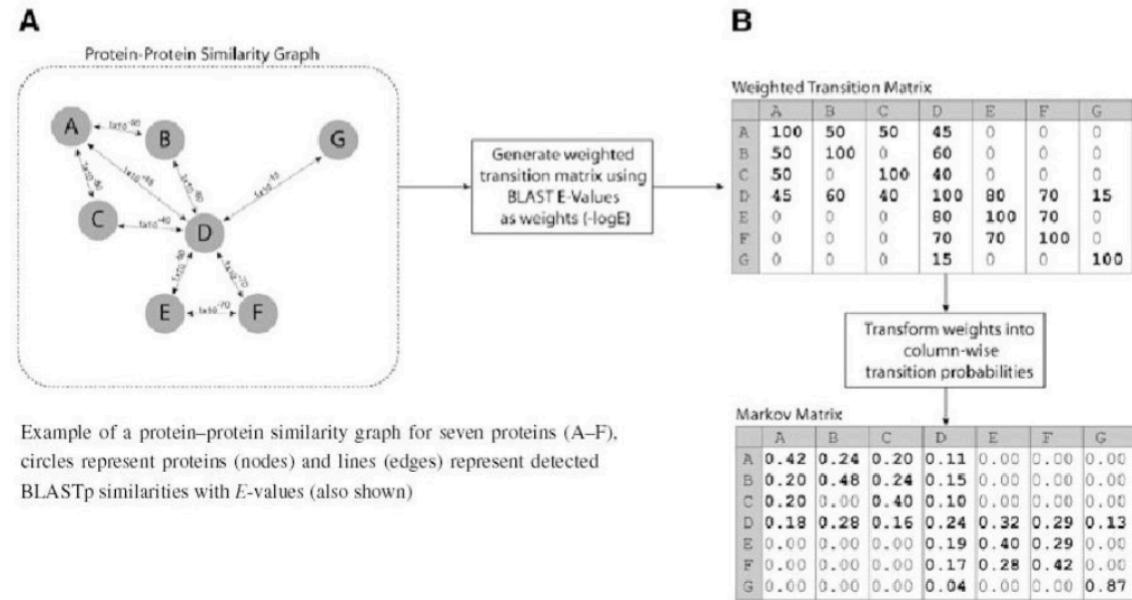
How to detect orthologous genes?

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

 目前無法顯示此影像。

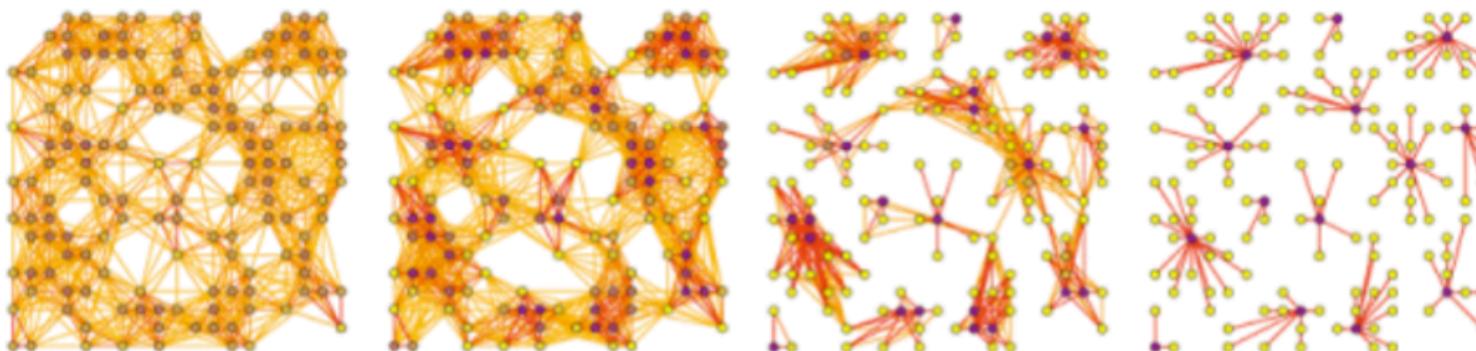
Sequence by clustering

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



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

Produce clusters (gene families) using different inflation parameter

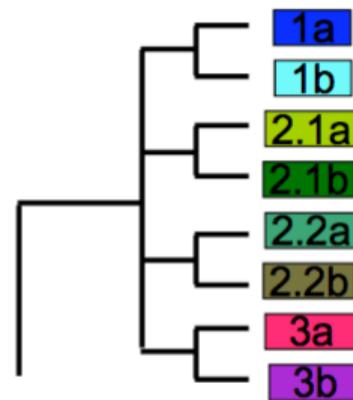


Weighted transition matrix and associated column stochastic Markov matrix for the seven proteins shown in (A).

Slide of Fred Tekaia

How to detect orthologous genes?

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



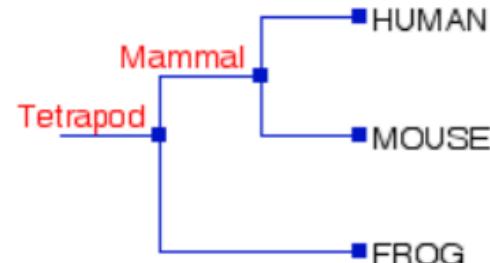
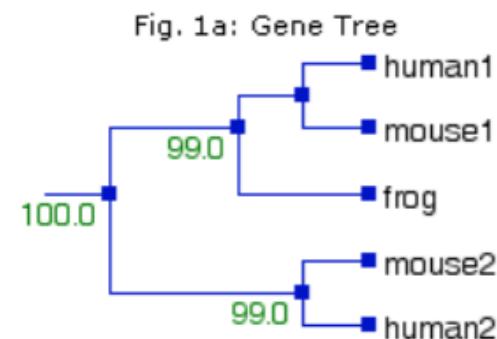
- more rigorous: look at synteny conservation



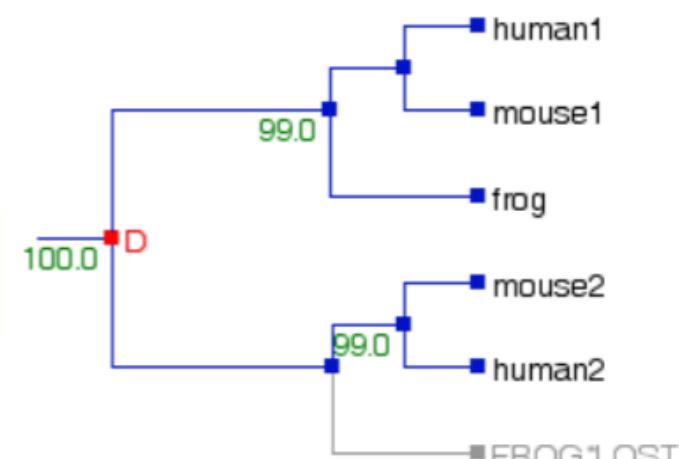
--> In fact inferring orthology is much more complicated particularly when considering more than 2 genomes!

Tree reconciliation

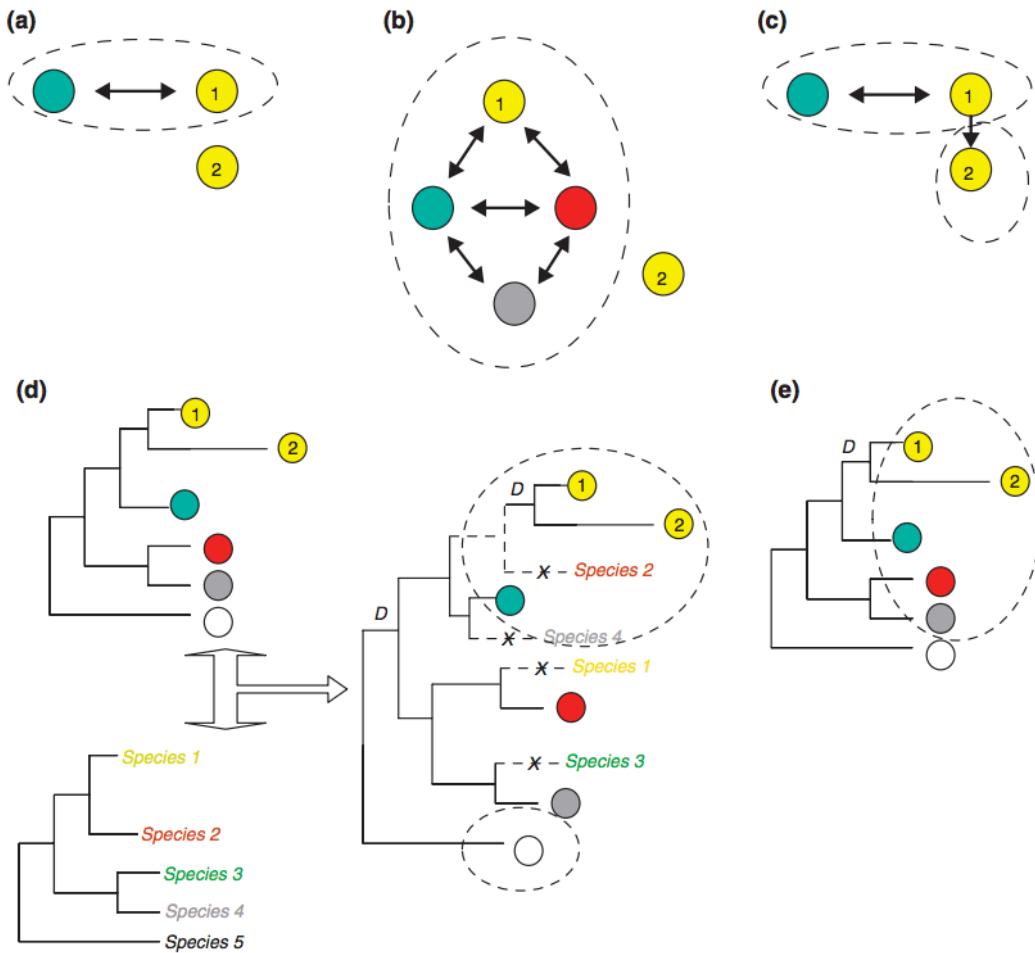
Detection of speciation and duplication events using a species tree and gene family tree



NOTUNG



Orthology prediction methods



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

Gebaldon (2008)

Methods

Similarity

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

Phylogeny

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

(Good and more interpretable for small set of genomes)

Others

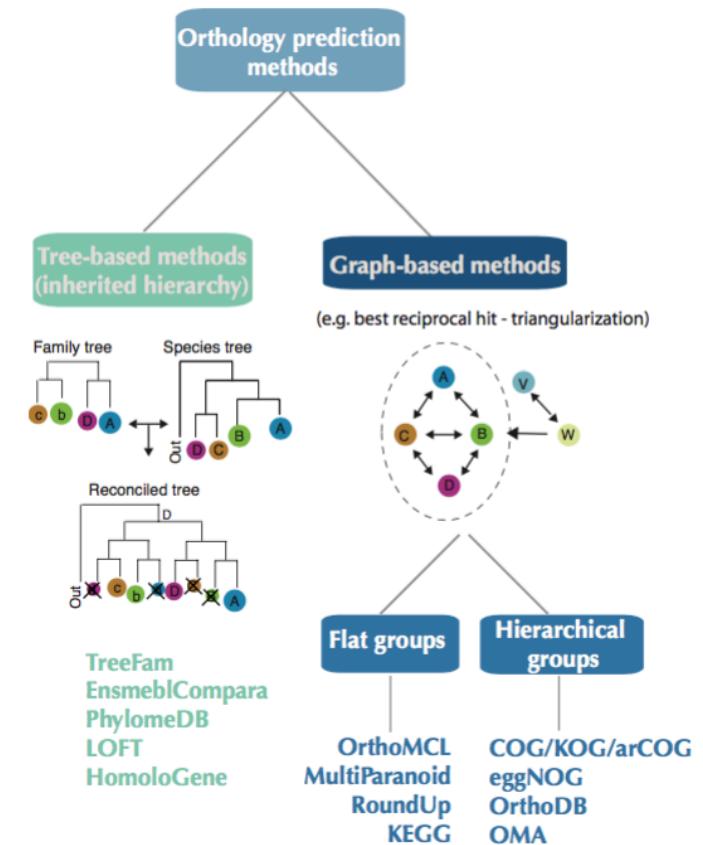
Combination of (1) and (2)

Some uses synteny

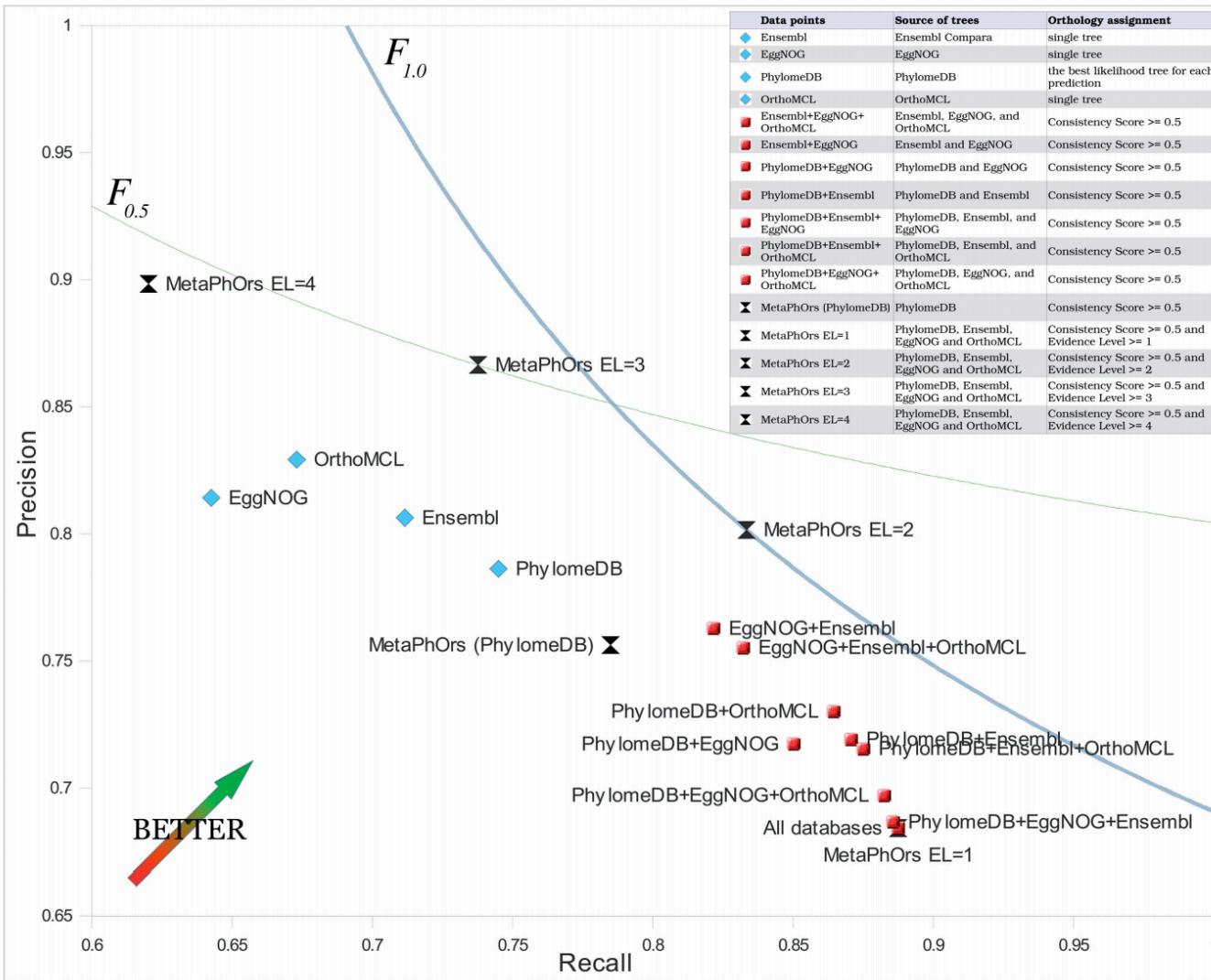
Tools

METHOD	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
Orthotrapp 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.



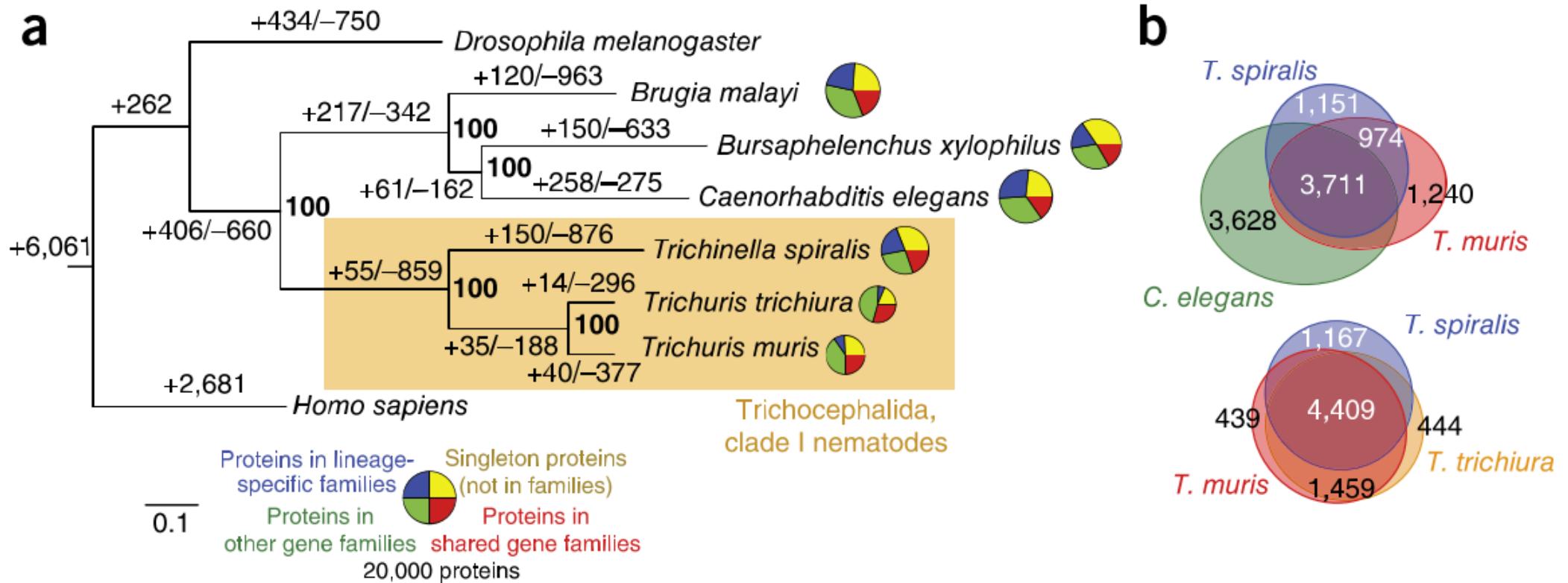
Every tool kind of disagrees...



Pryszzcz et al (2011)

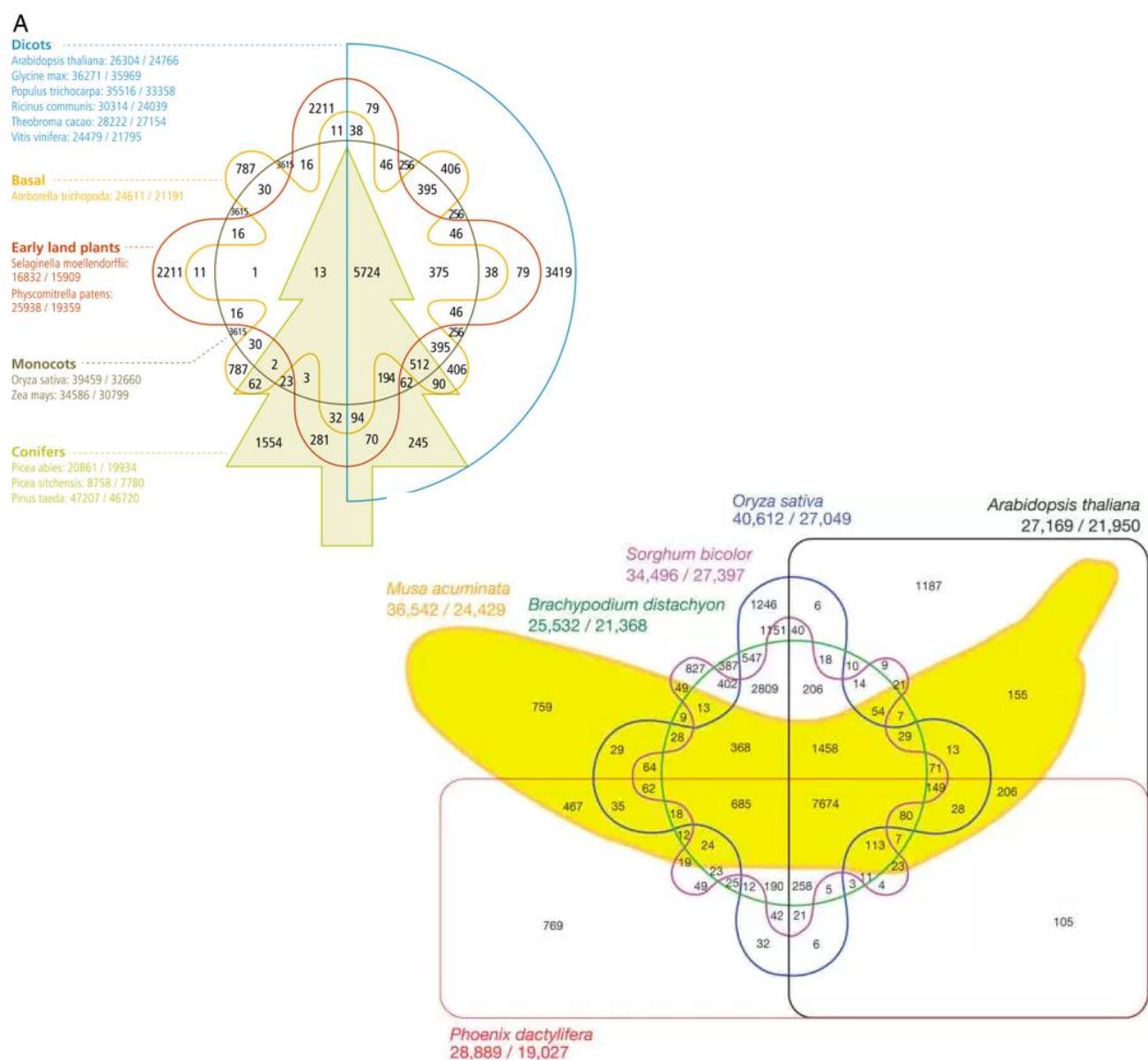
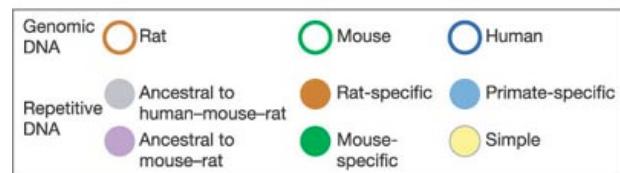
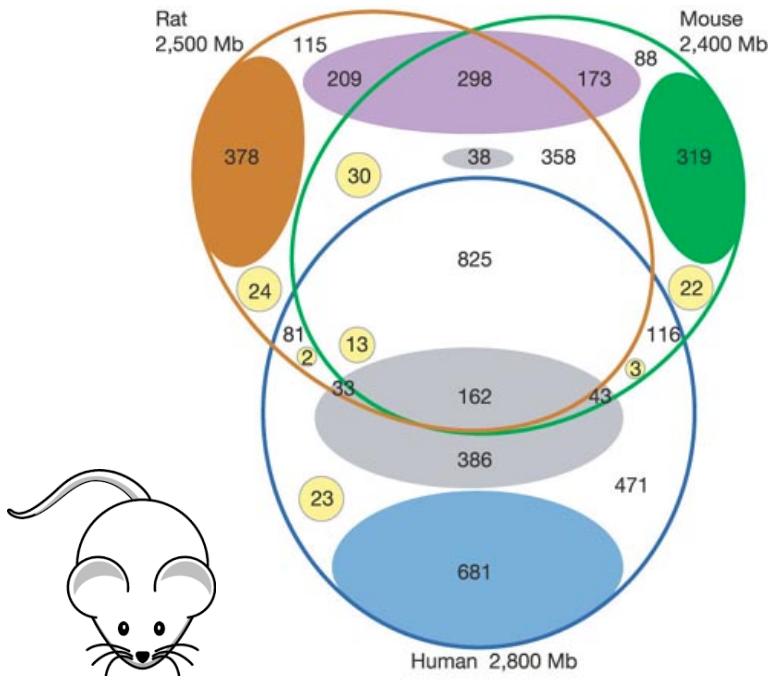
Visualisation of gene families

Phylogeny + Venn diagram to show expansion/loss



Foth (2014)

Trend of venn diagram...



Caveats

Evolution of multi-domain proteins

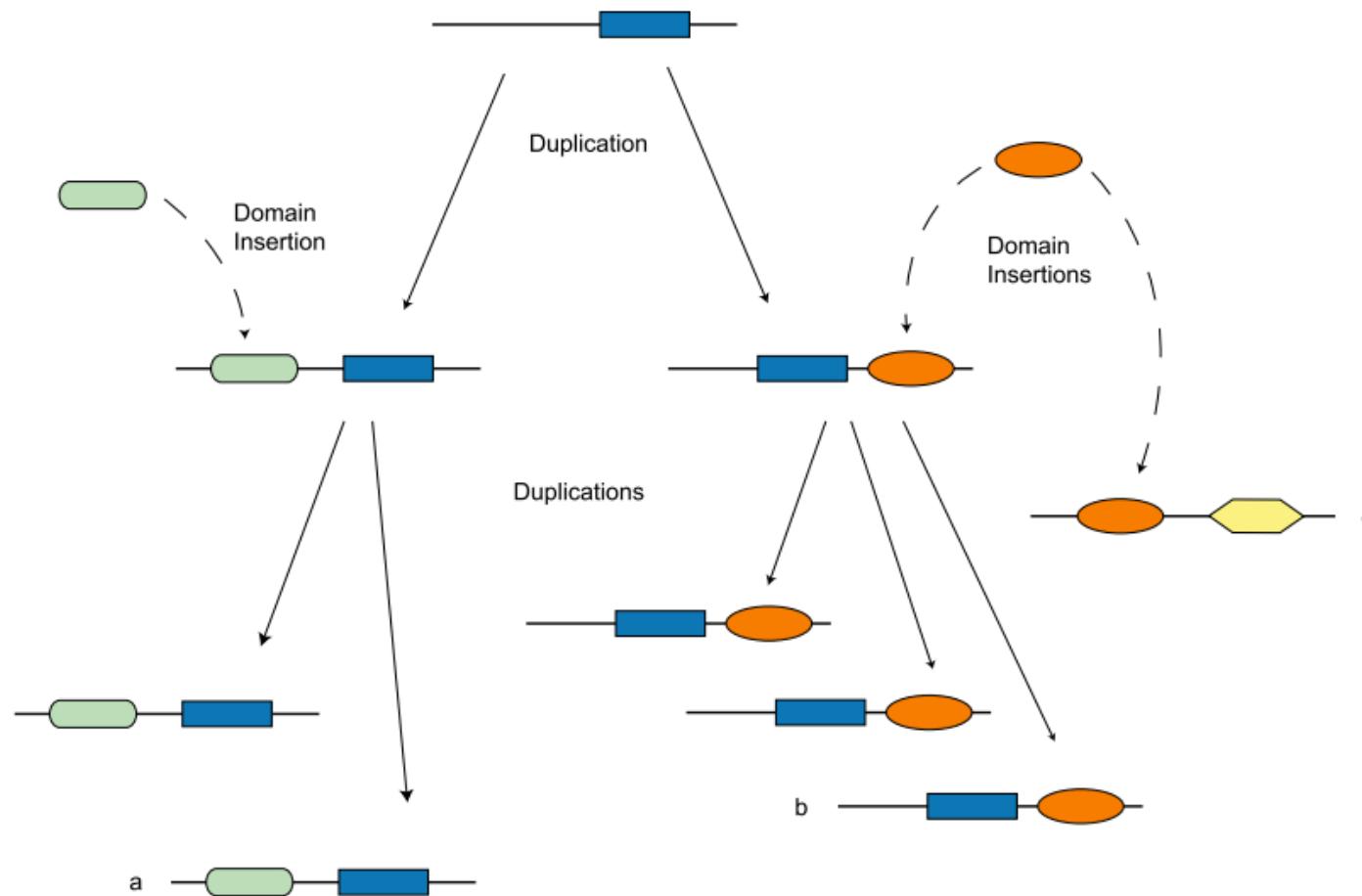
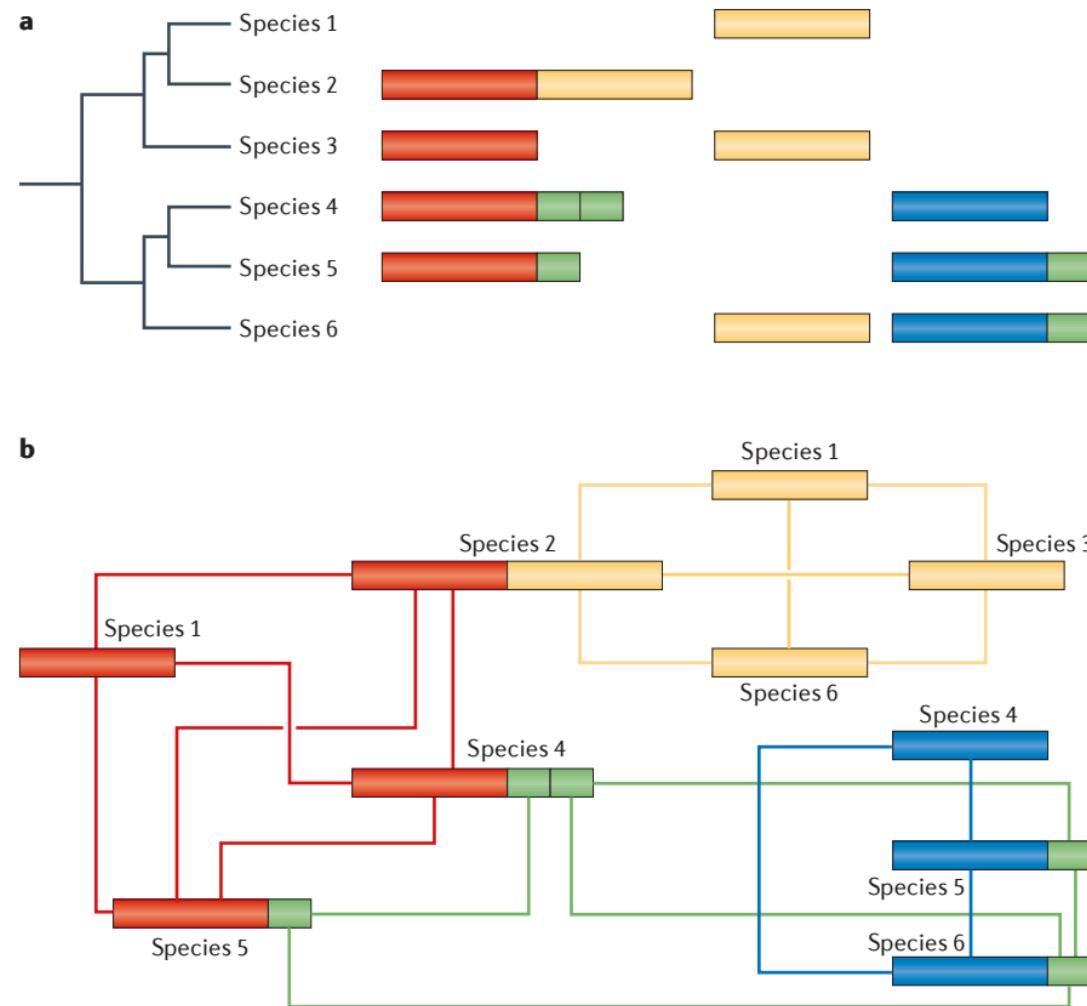


Figure 1. The evolution of a hypothetical multidomain family by gene duplication and domain insertion. Genes in the *a* and *b* subfamilies share a common ancestor but do not have identical domain composition. Gene *c* shares a homologous domain with genes in the *b* subfamily, but there is no gene that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

doi:10.1371/journal.pcbi.1000063.g001

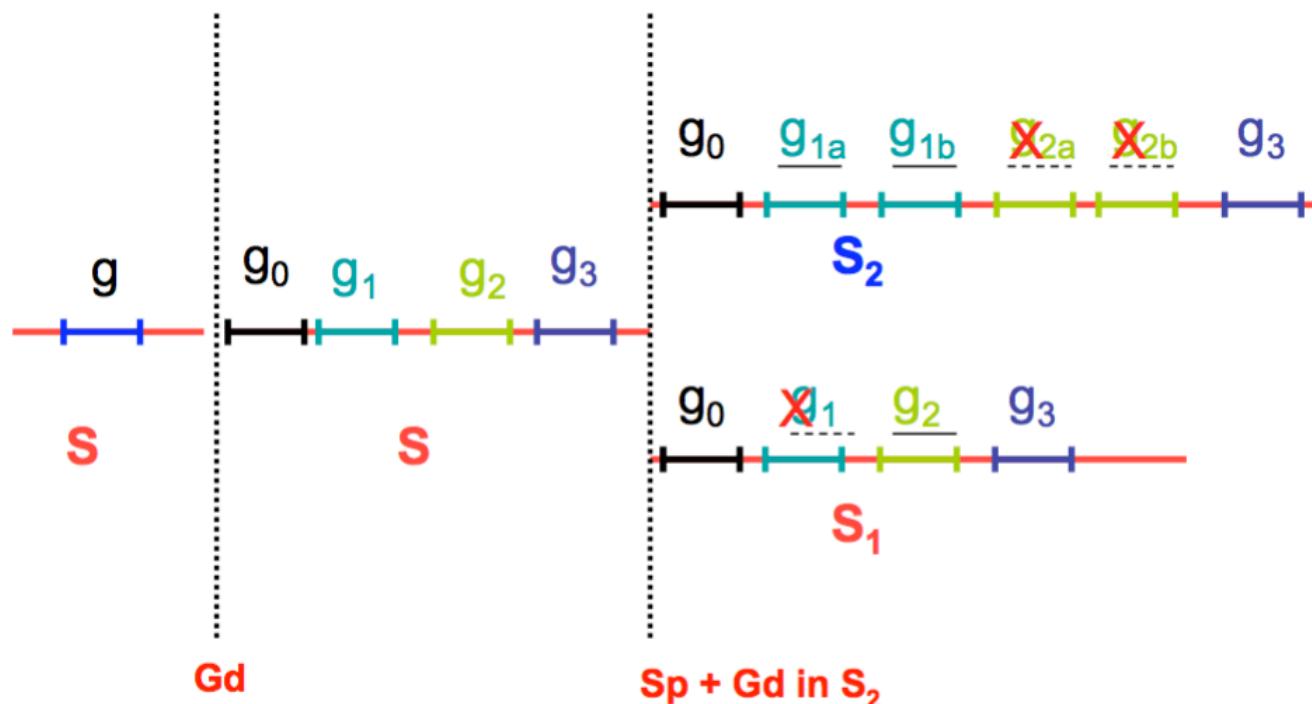
Song et al (2008)

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



Gabaldon and Koonin (2013)

Detection can go wrong: Example of an orthology misleading situation



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

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

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

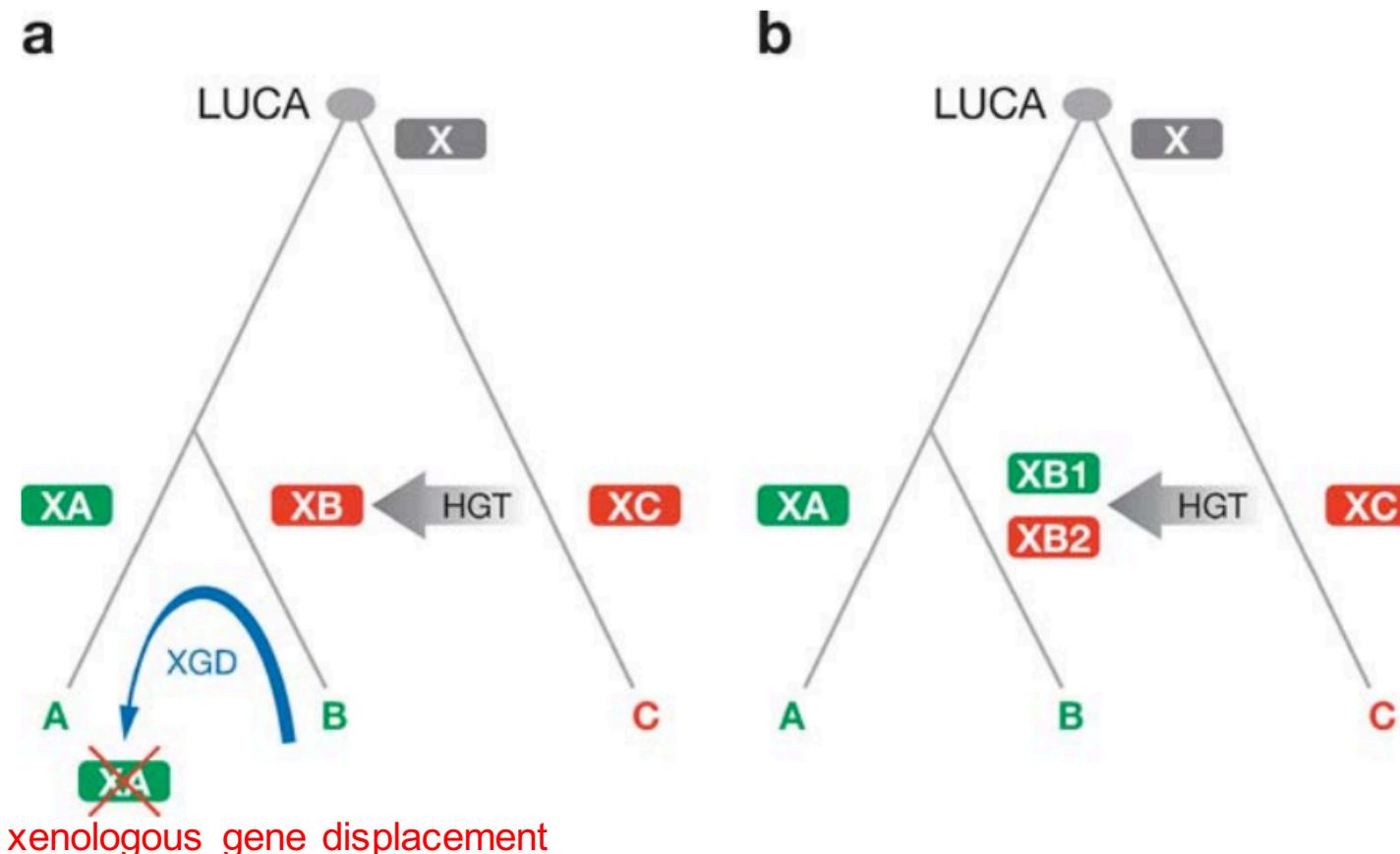


Figure 4

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

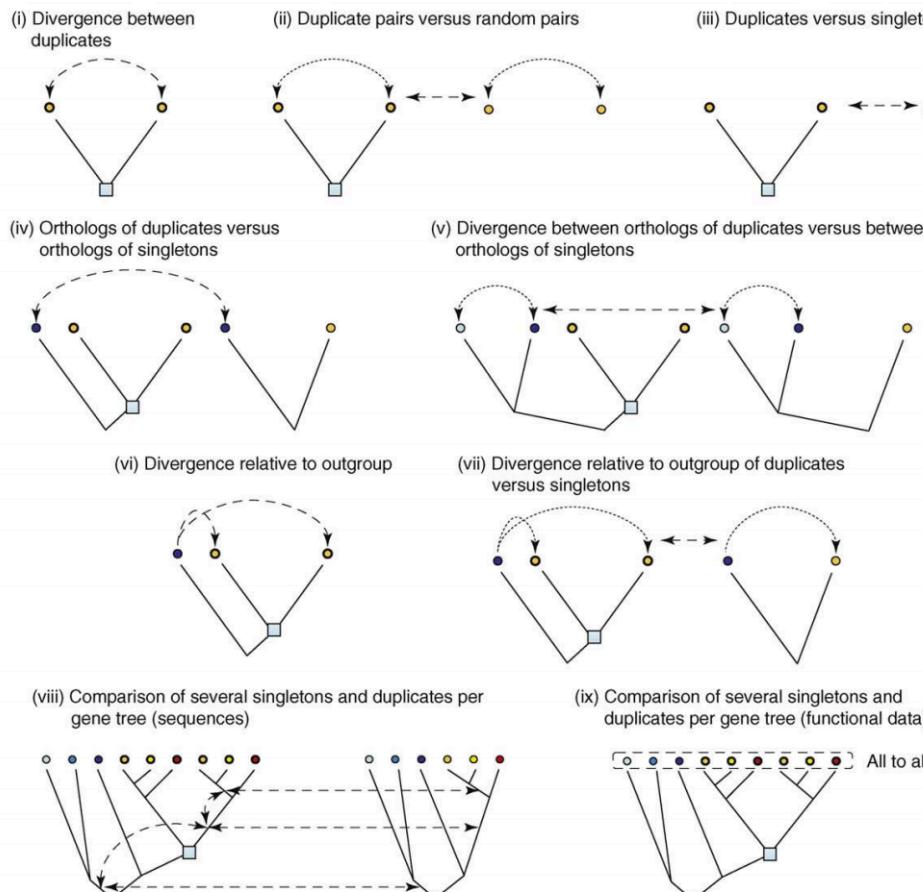
PLOS COMPUTATIONAL BIOLOGY

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 singlettons	Functional	Measure of retention bias, confused by evolution after duplication			[11,19,25]
(iv) Orthologs of duplicates versus orthologs of singlettons	Functional	Measure of retention bias			[12]
(v) Divergence between orthologs of duplicates versus between orthologs of singlettons	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 ^d	[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 singlettons	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 singlettons 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) idem	Functional	Conservation of pattern among singlettons; 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

Studer and Robinson-Rechavi (2009)

Testing duplication combining transcriptome dataset

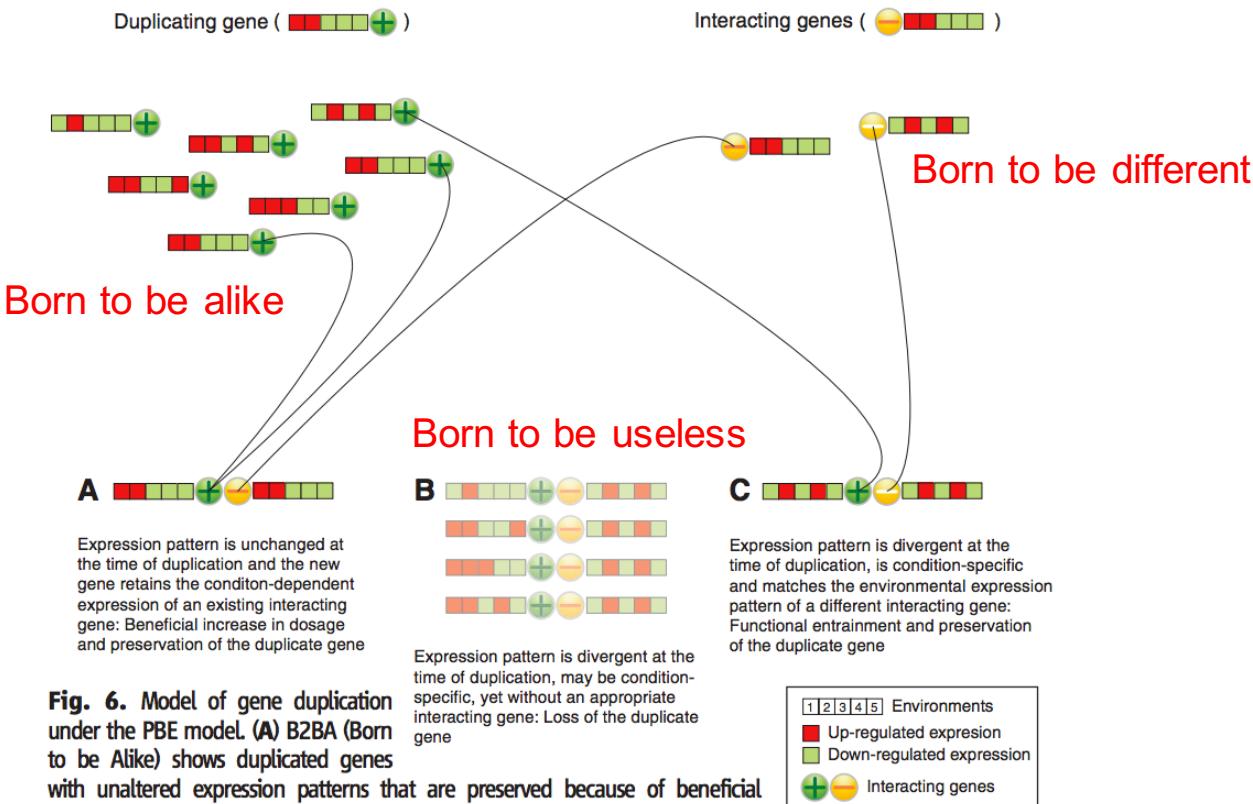


Fig. 6. Model of gene duplication under the PBE model. (A) B2BA (Born to be Alike) shows duplicated genes with unaltered expression patterns that are preserved because of beneficial increase in dosage (20) in association with the condition-dependent expression of an interacting gene. (B) B2BU (Born to 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.

Colbourne et al (2011)

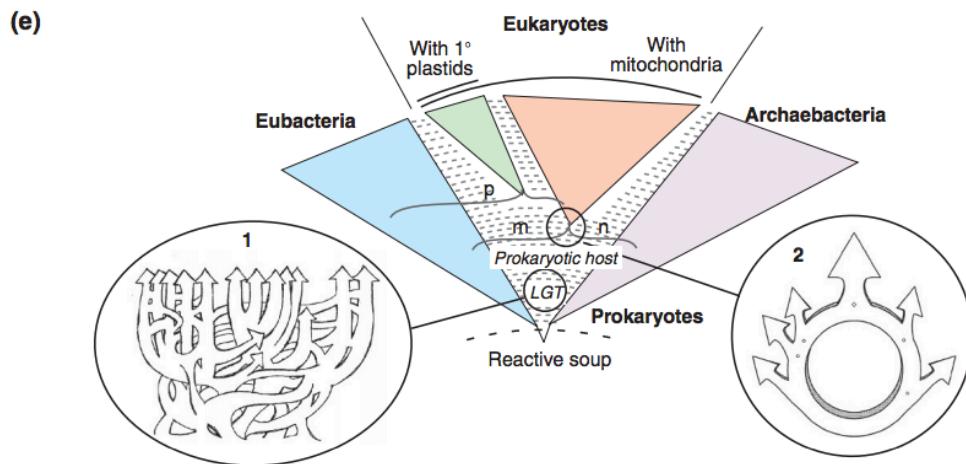
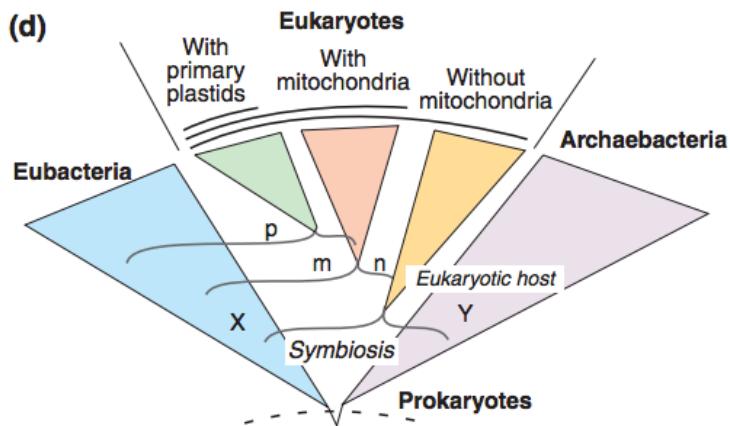
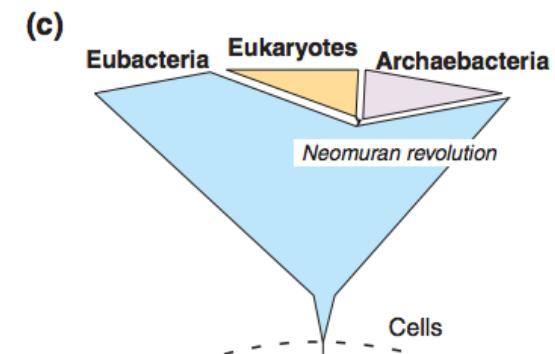
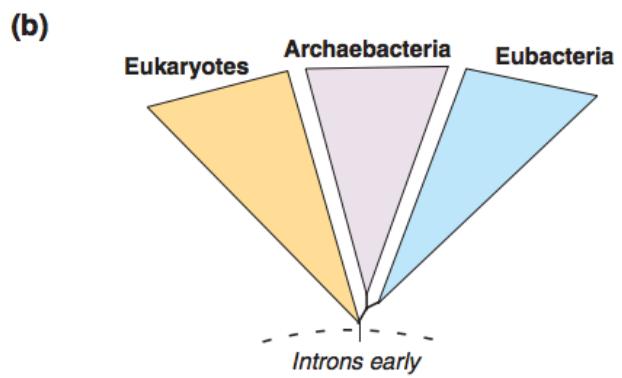
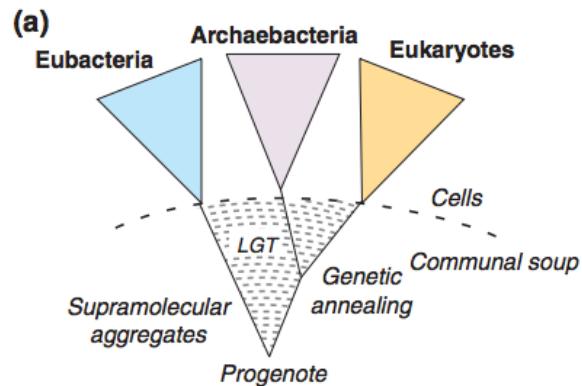
Summary point

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.

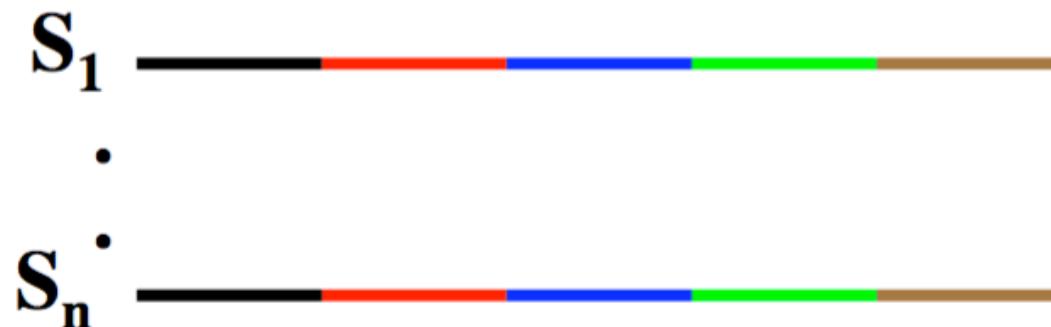
Phylogenomics

Five models models of tree of life



Dagan and Martin (2006)

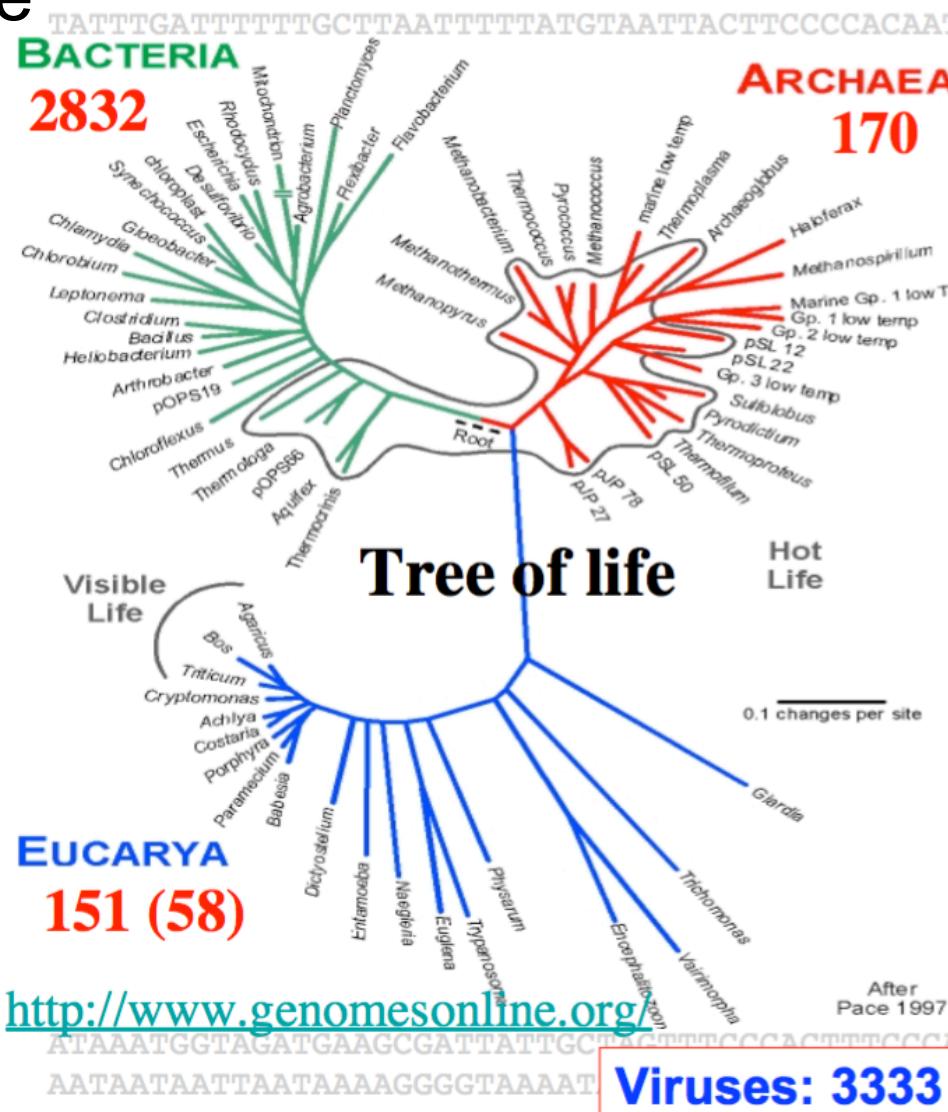
Probably the most common (easy) way to construct alignment of concatenated gene shared across all species



Limitations:

- genes don't evolve at the same rate nor in the same way;
- a limited number of genes are shared among all species;

Tree of life



Complete finished genomes: 3060

(04/09/14)

- 2832 Bacteria
- 170 Archaea
- 58 eukaryotes

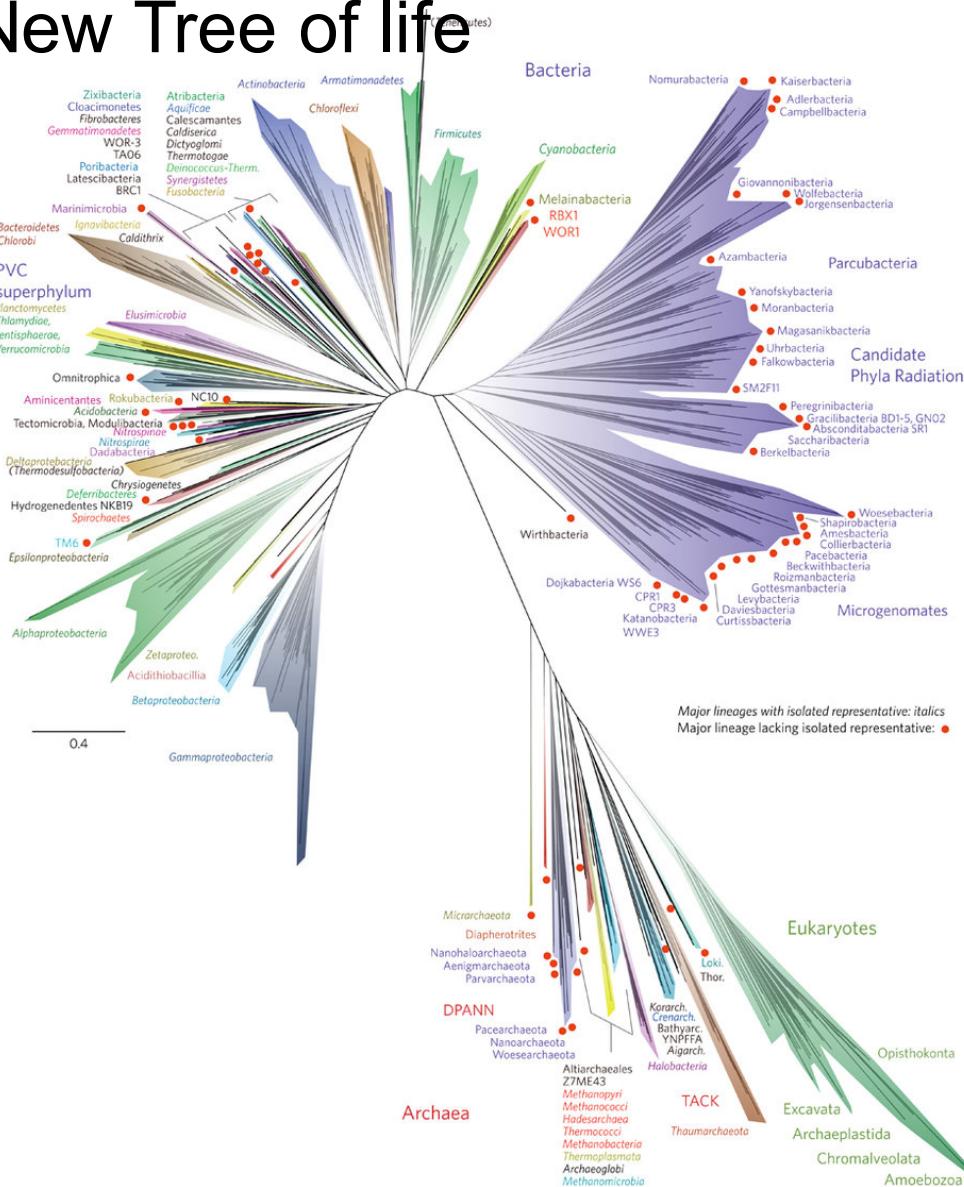
Incomplete genomes projects: 38262

- 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.”

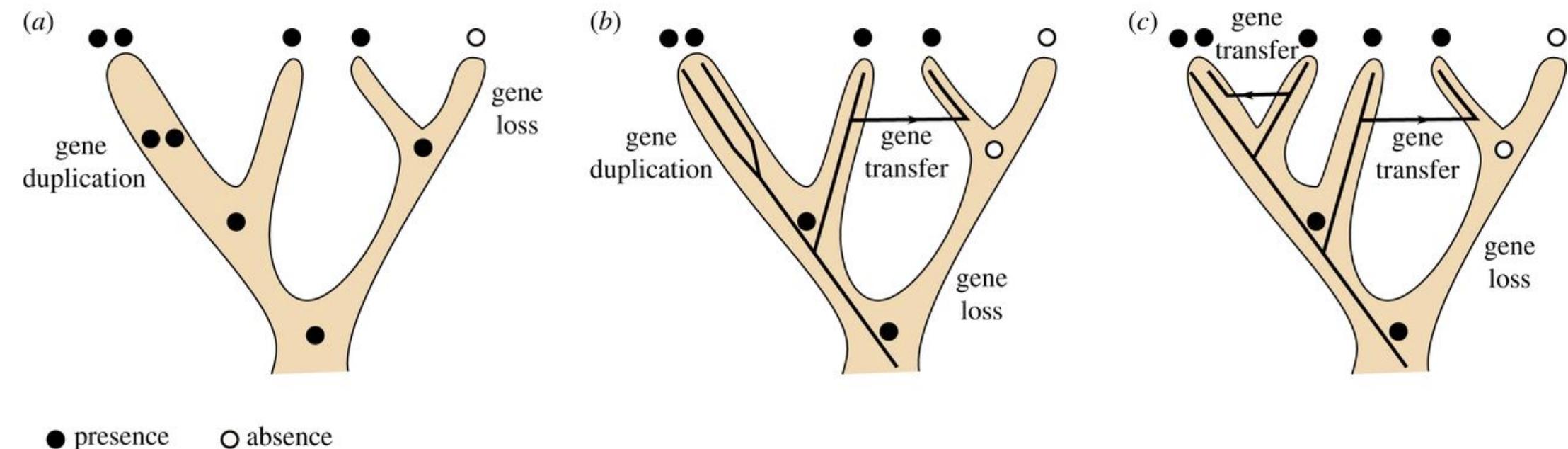
Hug et al (2016)

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

Horizontal gene transfer (HGT)

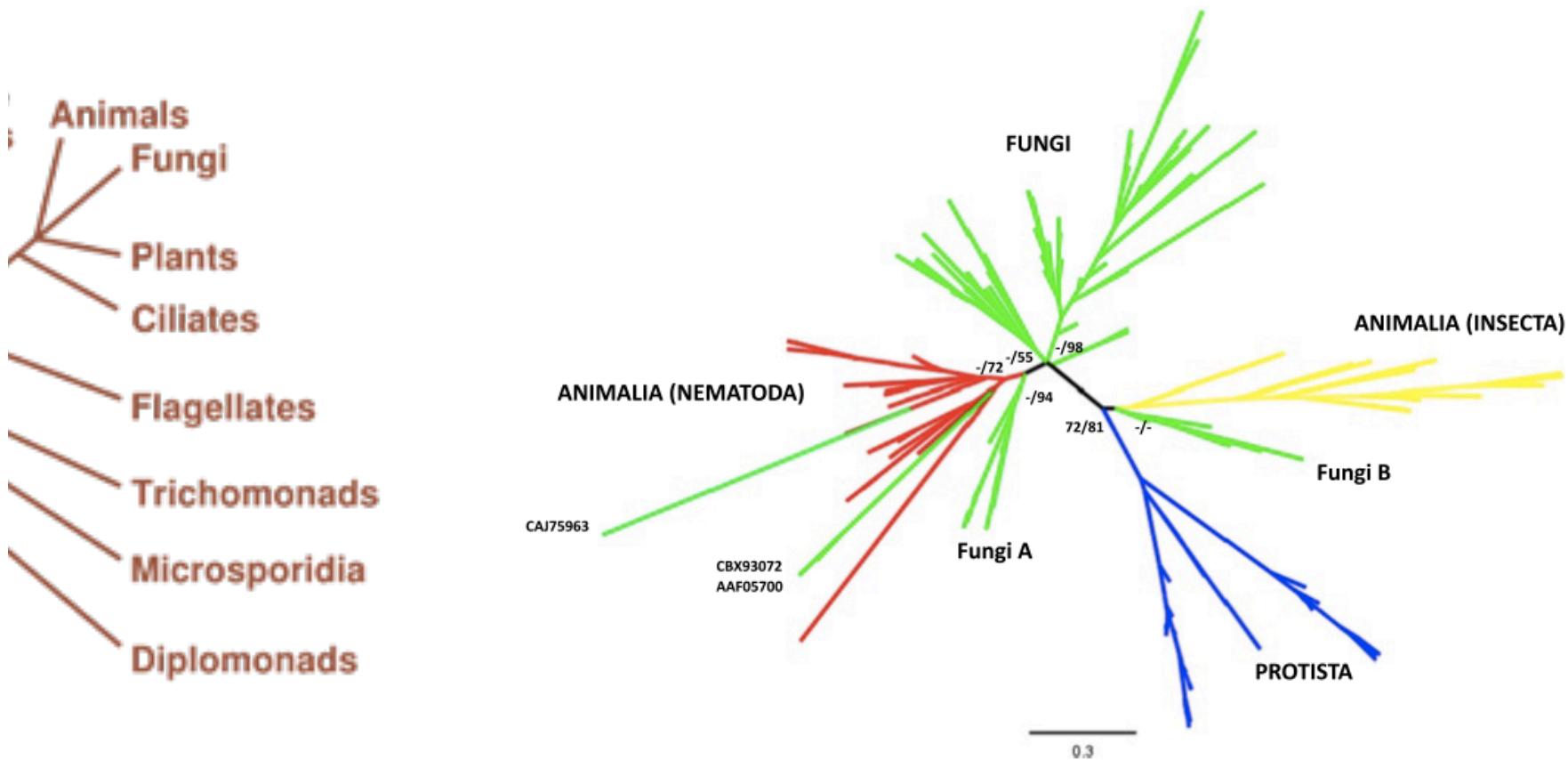
Inferring HGT require

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



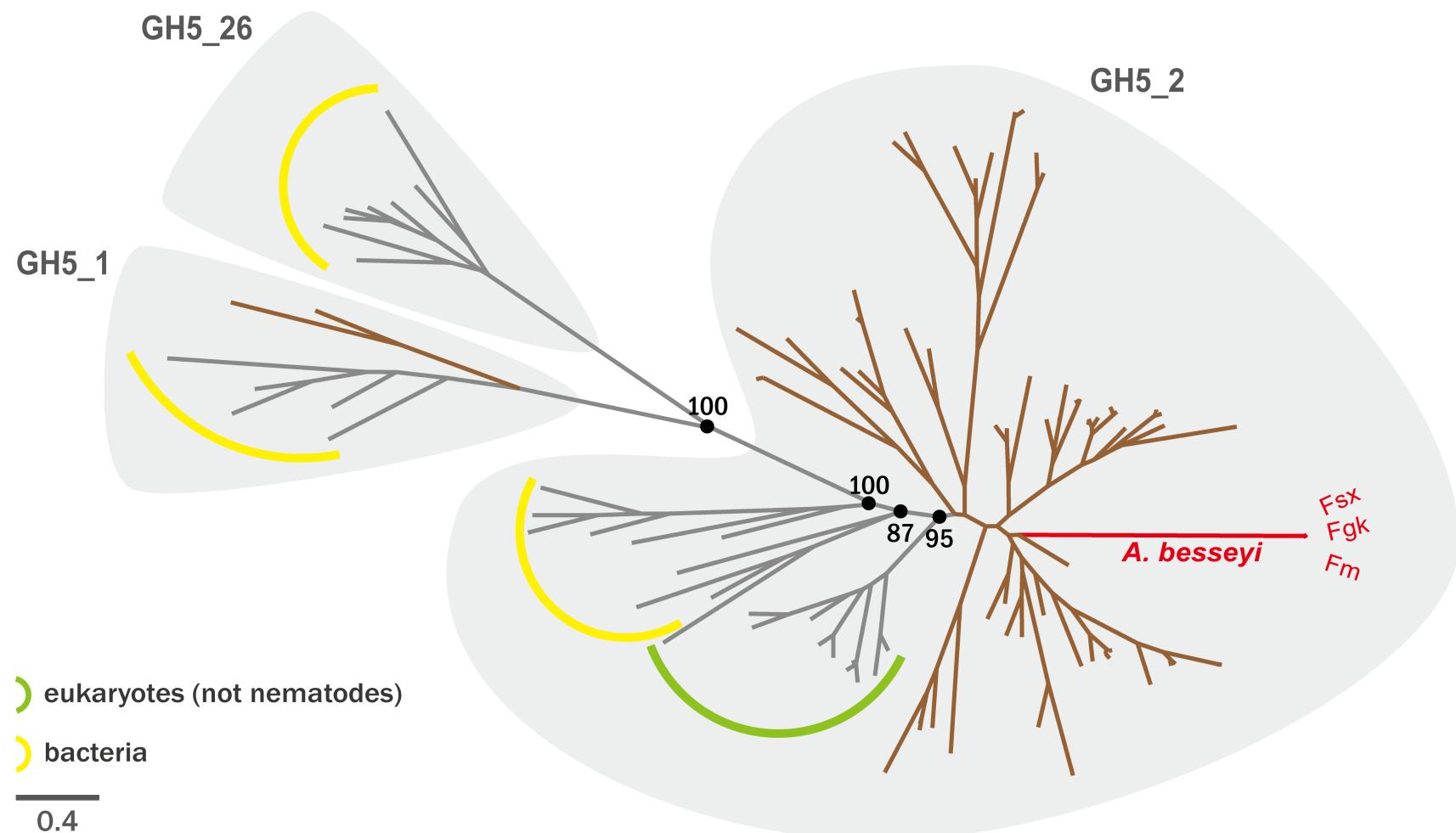
Szöllősi et al (2015)

HGT of GH45 enables plant parasitism in nematodes



Juan E Palomares-Rius, et al., BMC Evo. Biology (2014)

HGT of GH5 from bacteria in nematodes



Complicated history of genes: dig into finer details

Gene fusion



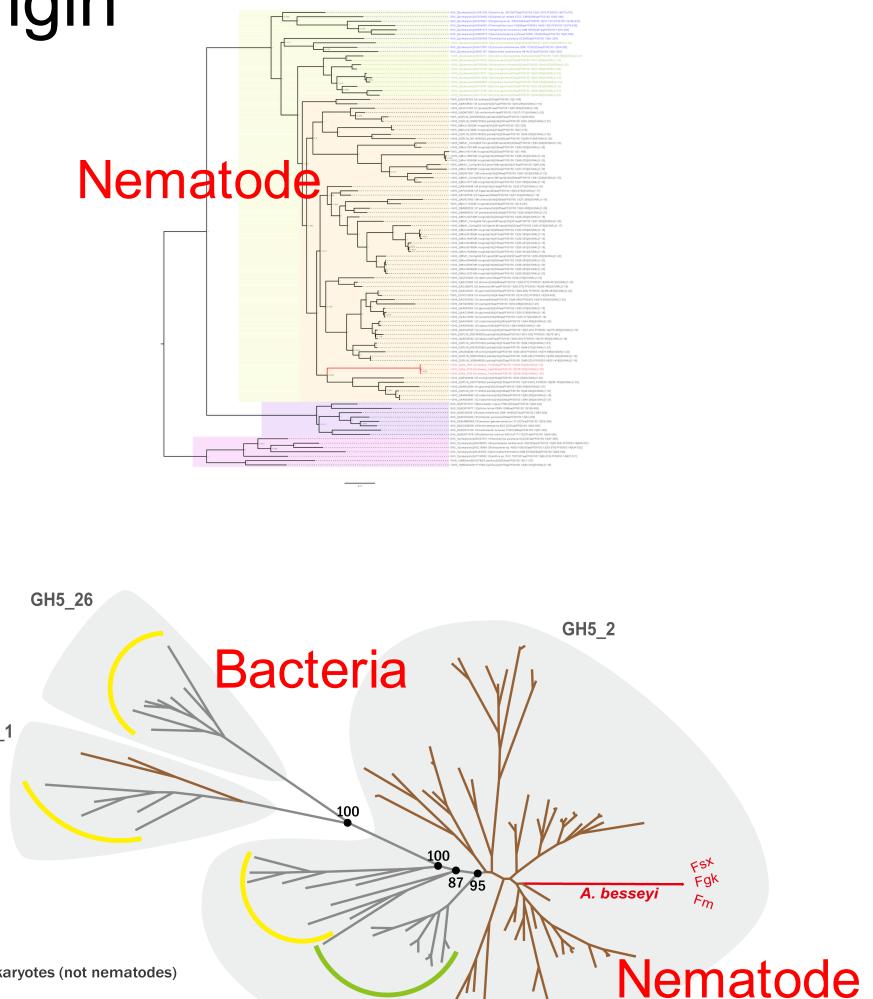
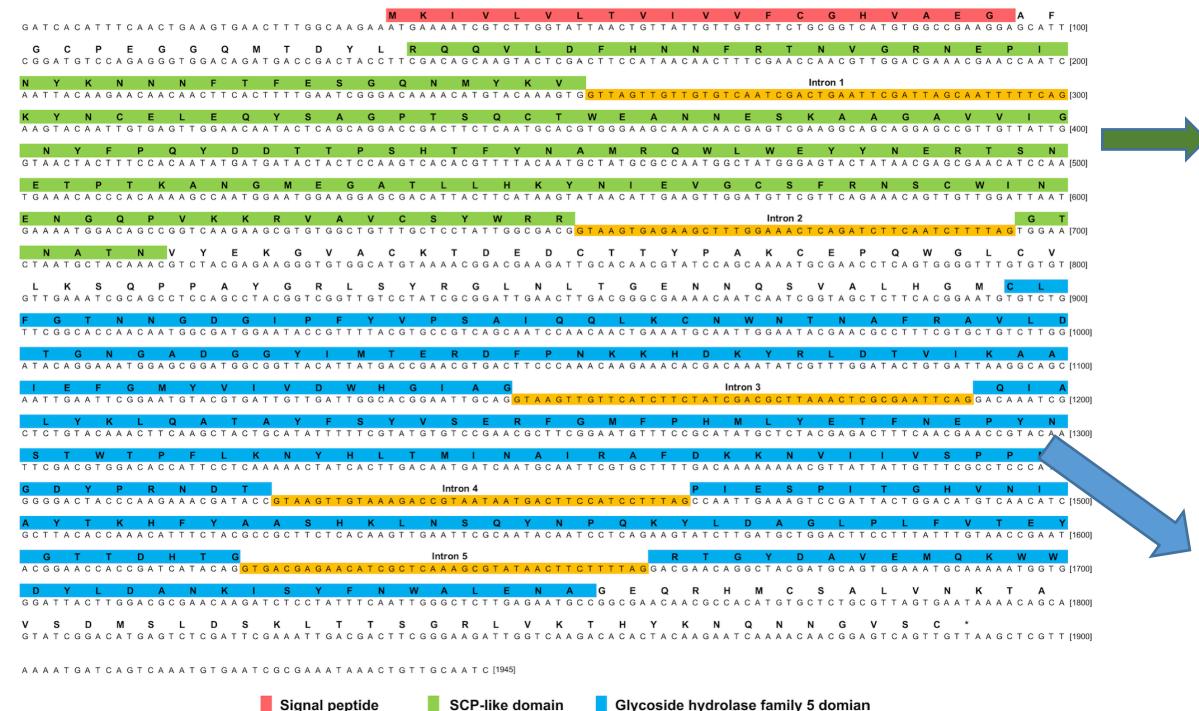
Gene fission



Domains shuffling



A gene with domains of different history one HGT + one nematode ancestor origin



Wu and Guo et al (in press)

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

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

*MAPS of LINKAGE and
SYNTENY HOMOLOGIES
between MOUSE and MAN*

JOSEPH H. NADEAU

1989

Synteny

conservation of gene content

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:



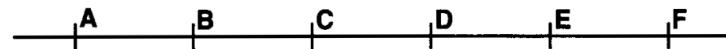
Count:

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.

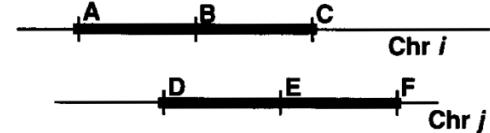
A. Genetic map in reference species



Each unit is gene

Conserved synteny, disrupted synteny, conserved linkage, disrupted linkage

Gene arrangement:



Count:

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

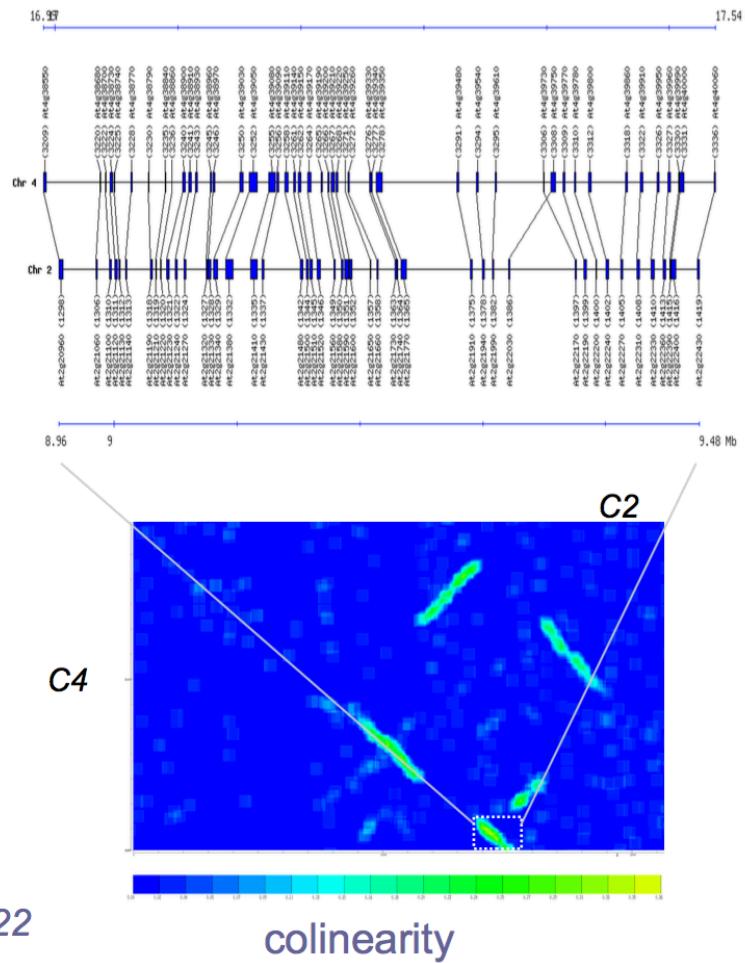
Possible causes:

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

Ehrlich (1997)

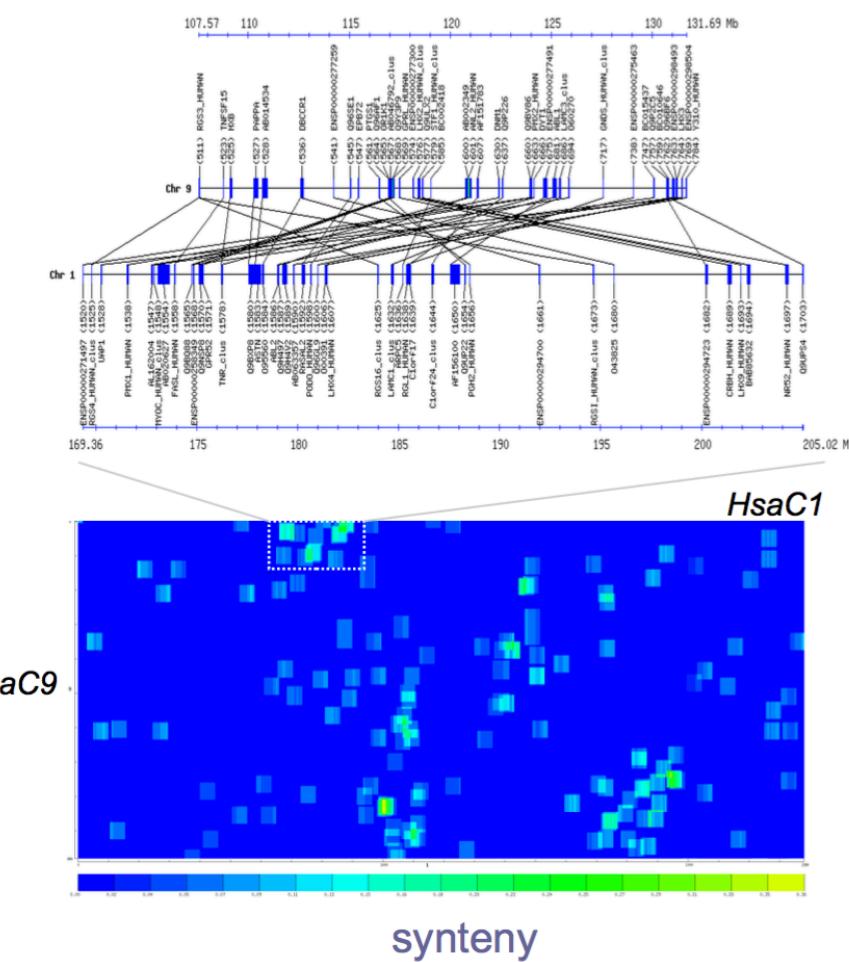
Synteny and colinearity

recent duplication



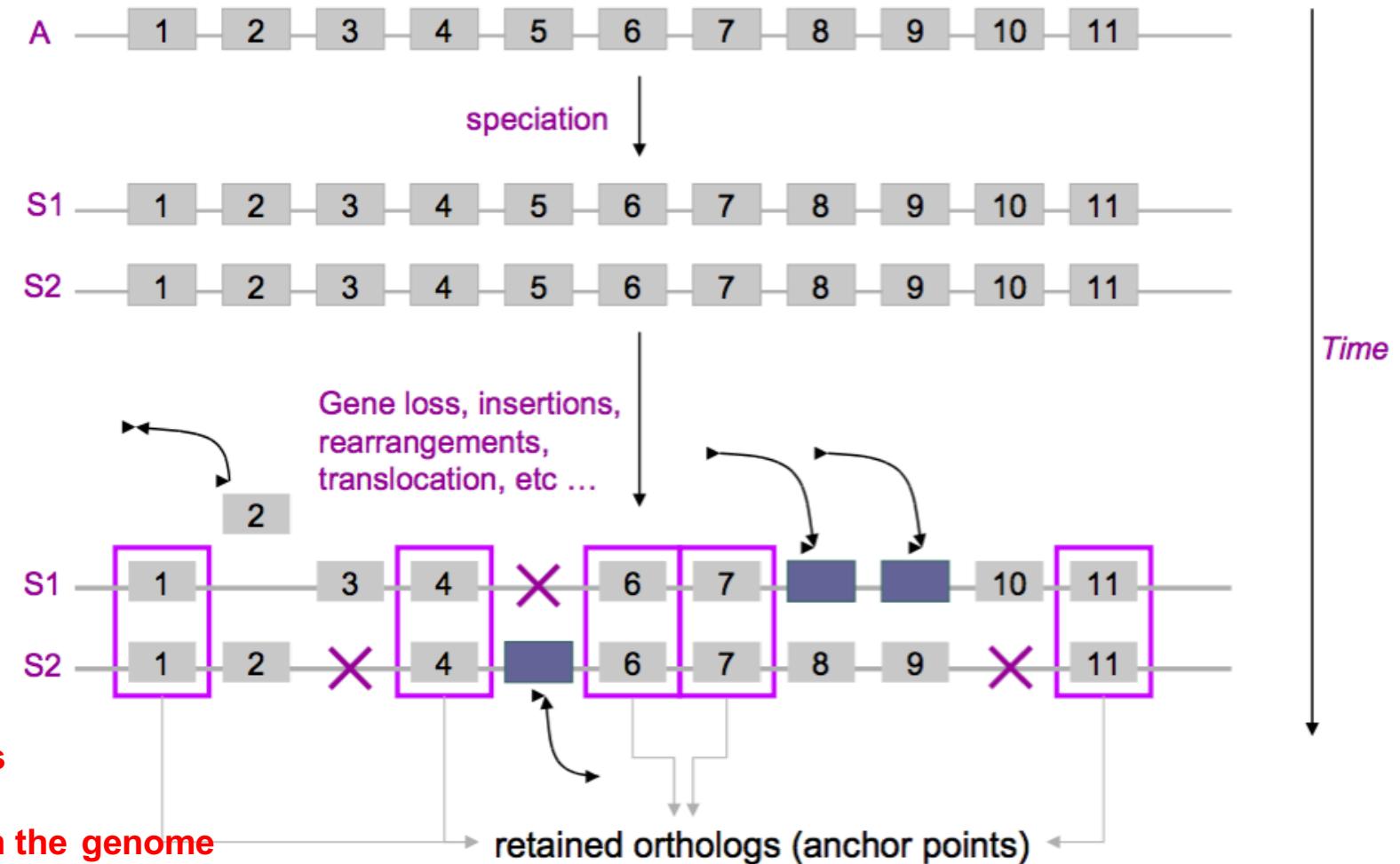
22

ancient duplication



Slide by Klaas Vandepoele

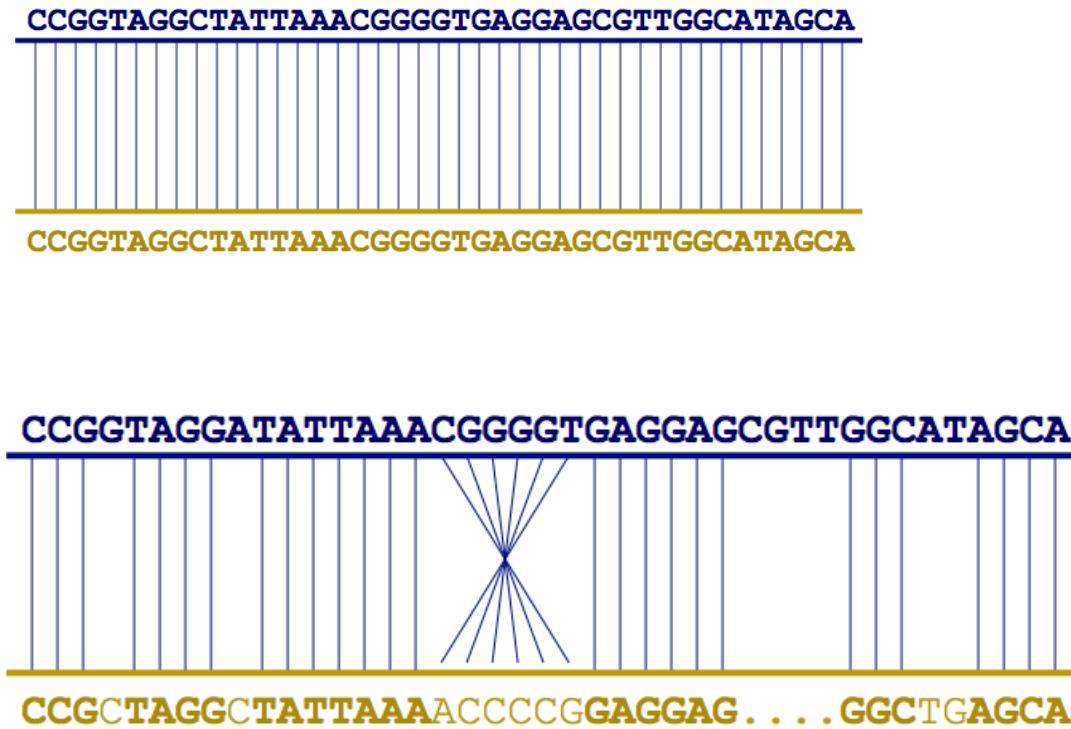
Inferring gene collinearity



Whole genome alignment

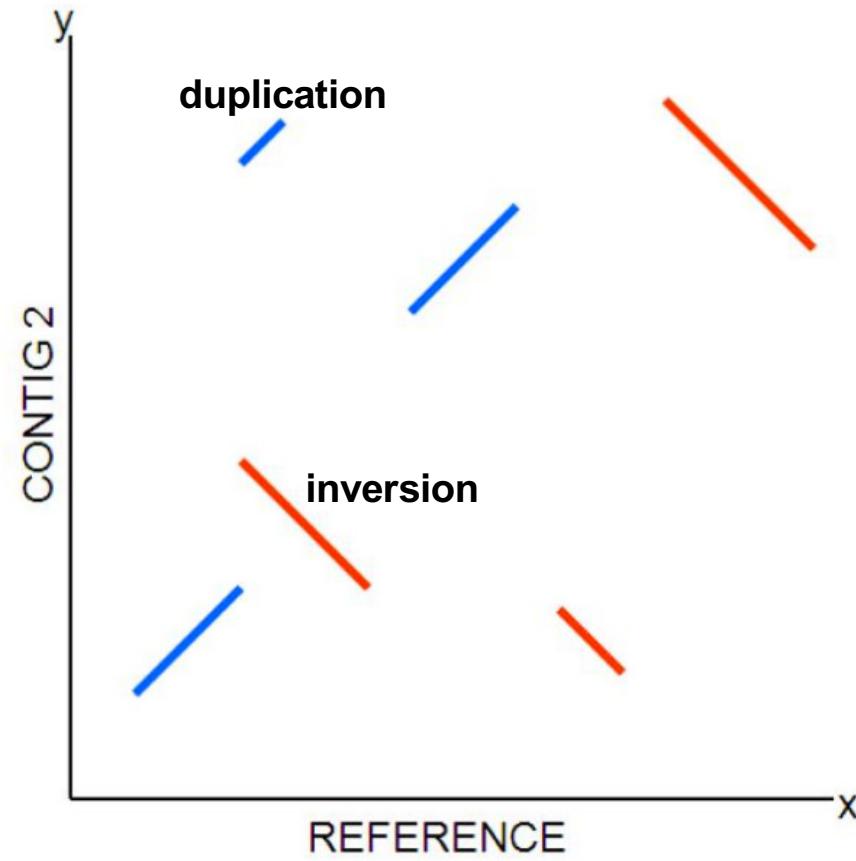
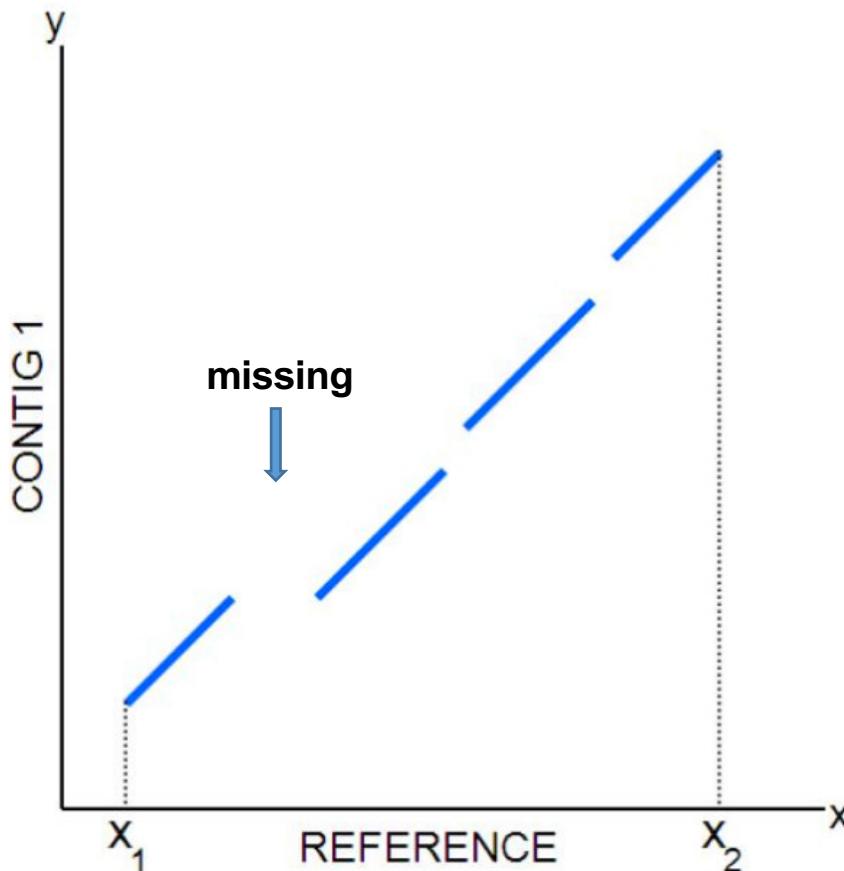
For two genomes, A and B,
find a mapping from each
position in A to its
corresponding position in B

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

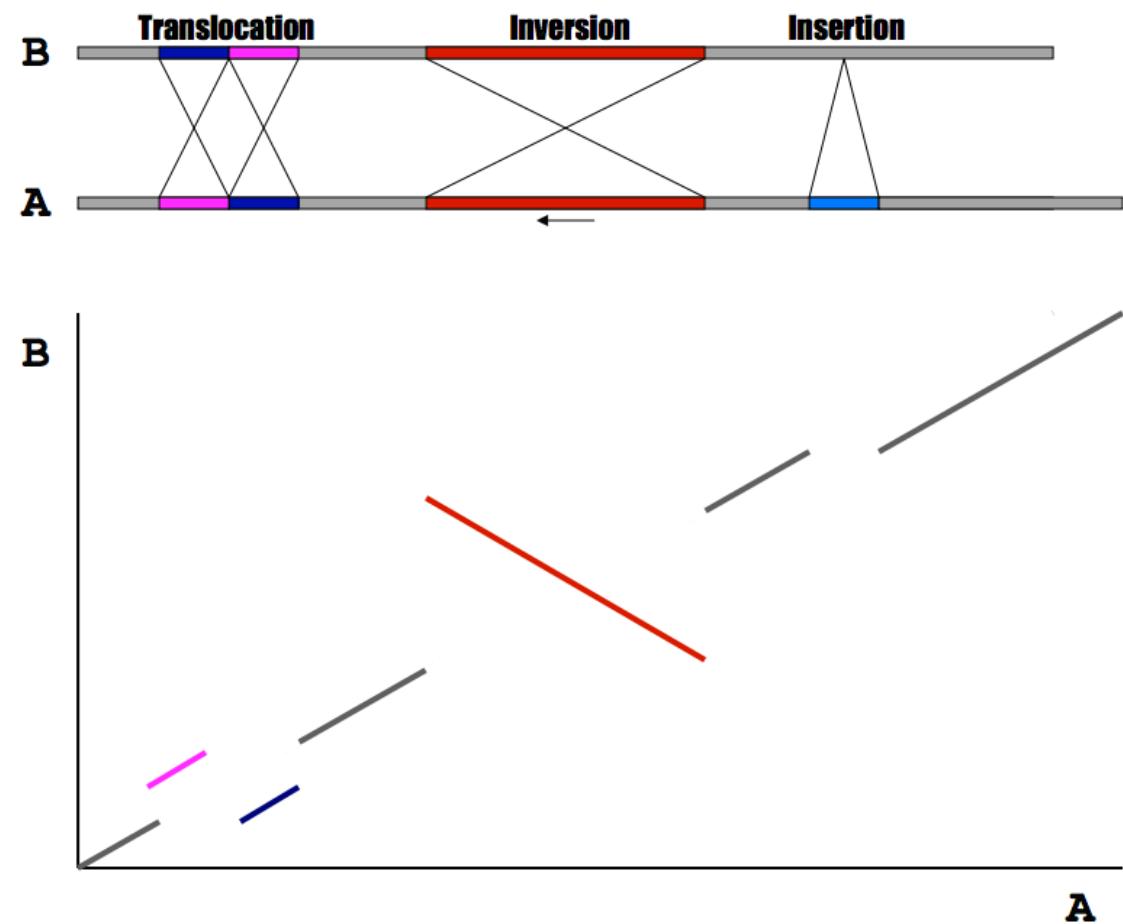


Aligning genome at nucleotide / amino acid level

Visualise through **dotplot**



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



Available tools

Synteny inference:

i-ADHoRe 3.0

DAGchainer

Mercator

MCscanX

Genome alignment:

MUMMER (nucmer and promer; <http://mummer.sourceforge.net/>)

LASTZ (<http://www.bx.psu.edu/~rsharris/lastz/>)

BLAT (<http://genome.ucsc.edu/goldenPath>)

Mugsy (<http://mugsy.sourceforge.net/>)

megaBLAST (<http://www.ncbi.nlm.nih.gov/blast/>) • MUMmer

LAGAN (<http://lagan.stanford.edu/lagan> web/index.shtml)

Mummer usage

```
nucmer -maxmatch C092.fasta KIM.fasta
```

-maxmatch Find maximal exact matches (MEMs)

```
delta-filter -m out.delta > out.filter.m
```

-m Many-to-many mapping

```
show-coords -r out.delta.m > out.coords
```

-r Sort alignments by reference position

```
dnadiff out.delta.m
```

Construct catalog of sequence variations

```
mummerplot --large --layout out.delta.m
```

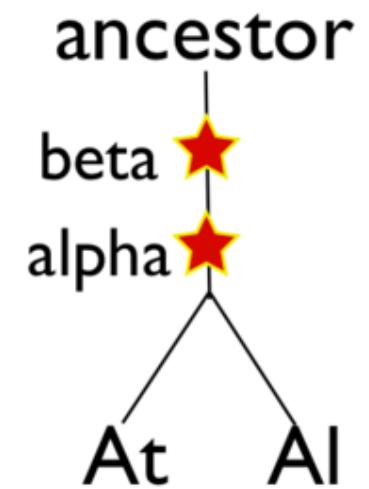
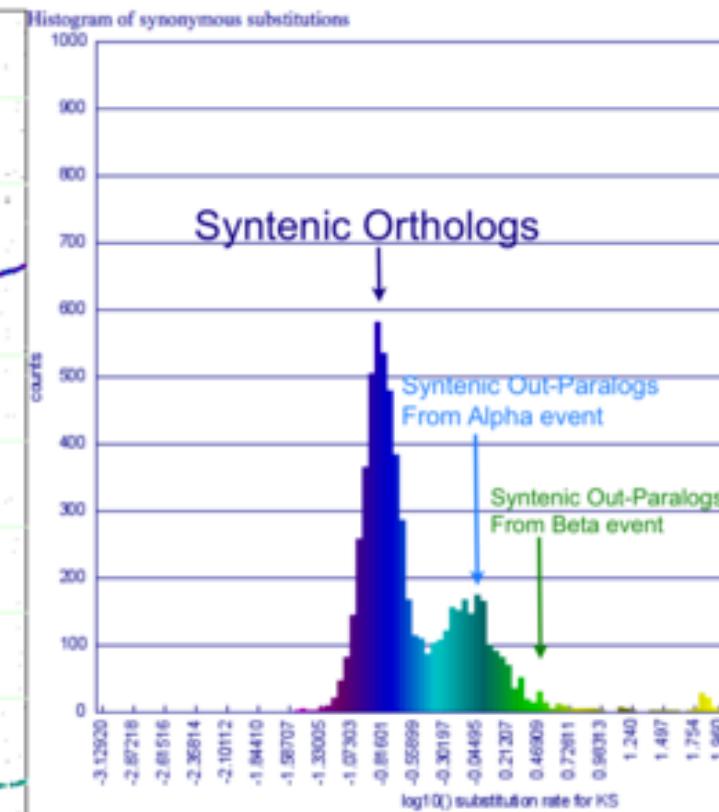
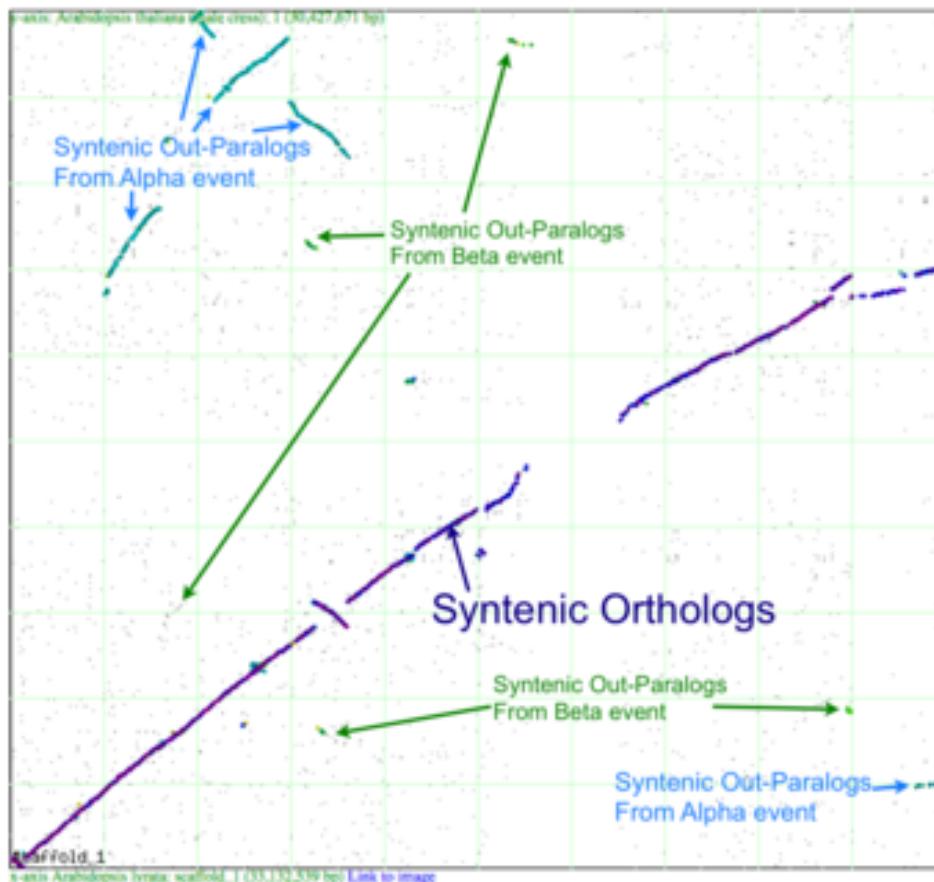
--large Large plot

--layout Nice layout for multi-fasta files

--x11 Default, draw using x11 (--postscript, --png)

*requires gnuplot

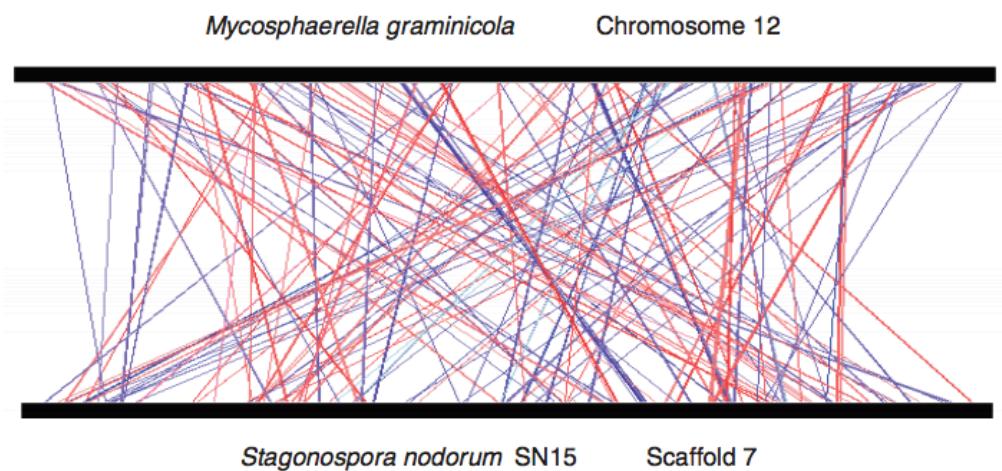
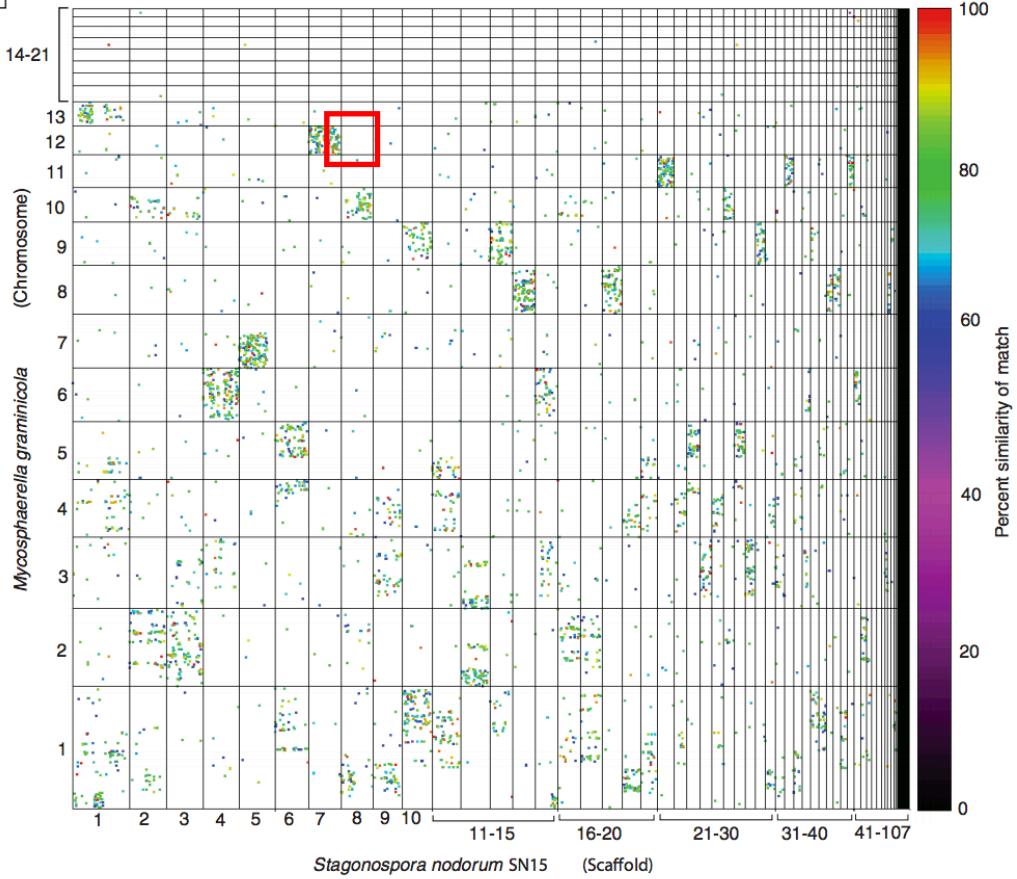
Relationship between genome synteny, syntenic orthologs and duplications



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

Relationship between genome synteny, syntenic orthologs and duplications

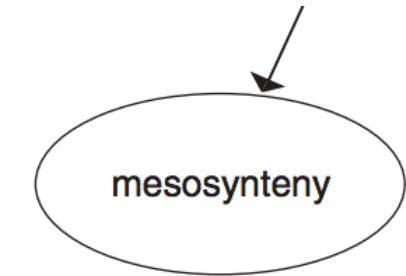
(a)



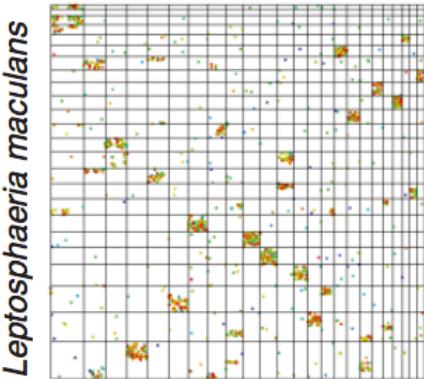
Hane et al (2011)

Different kinds of genome synteny

mesosynteny

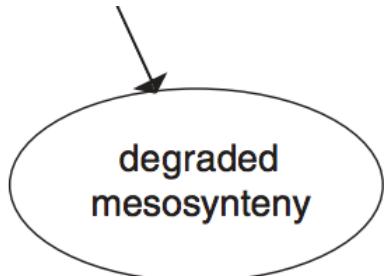


Phaeosphaeria nodorum

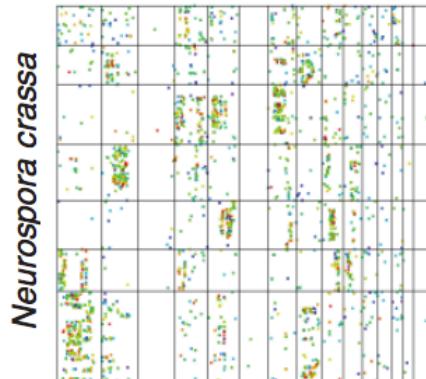


genes are conserved within homologous chromosomes,
but with randomized orders and orientations

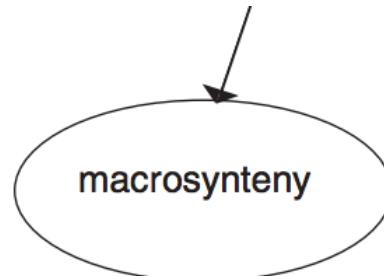
degraded
mesosynteny



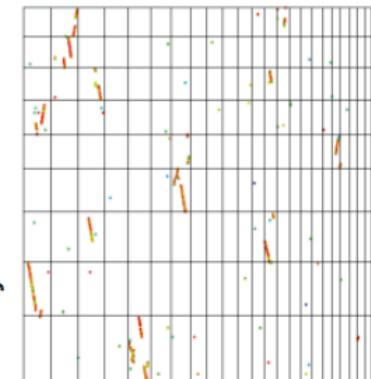
Fusarium oxysporum



macrosynteny

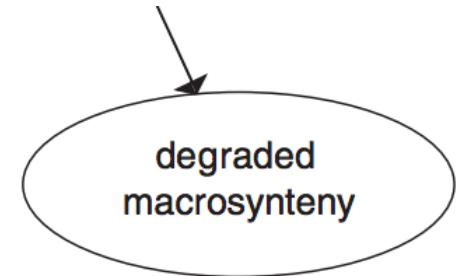


Sclerotinia sclerotiorum

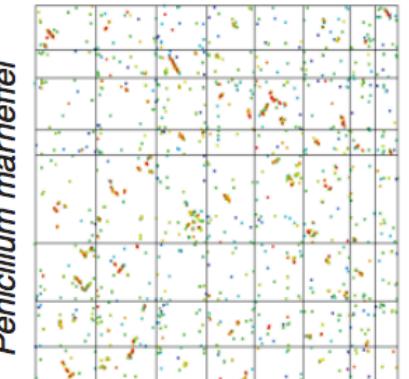


genes are conserved within homologous chromosomes,
and with colinear gene regions

degraded
macrosynteny

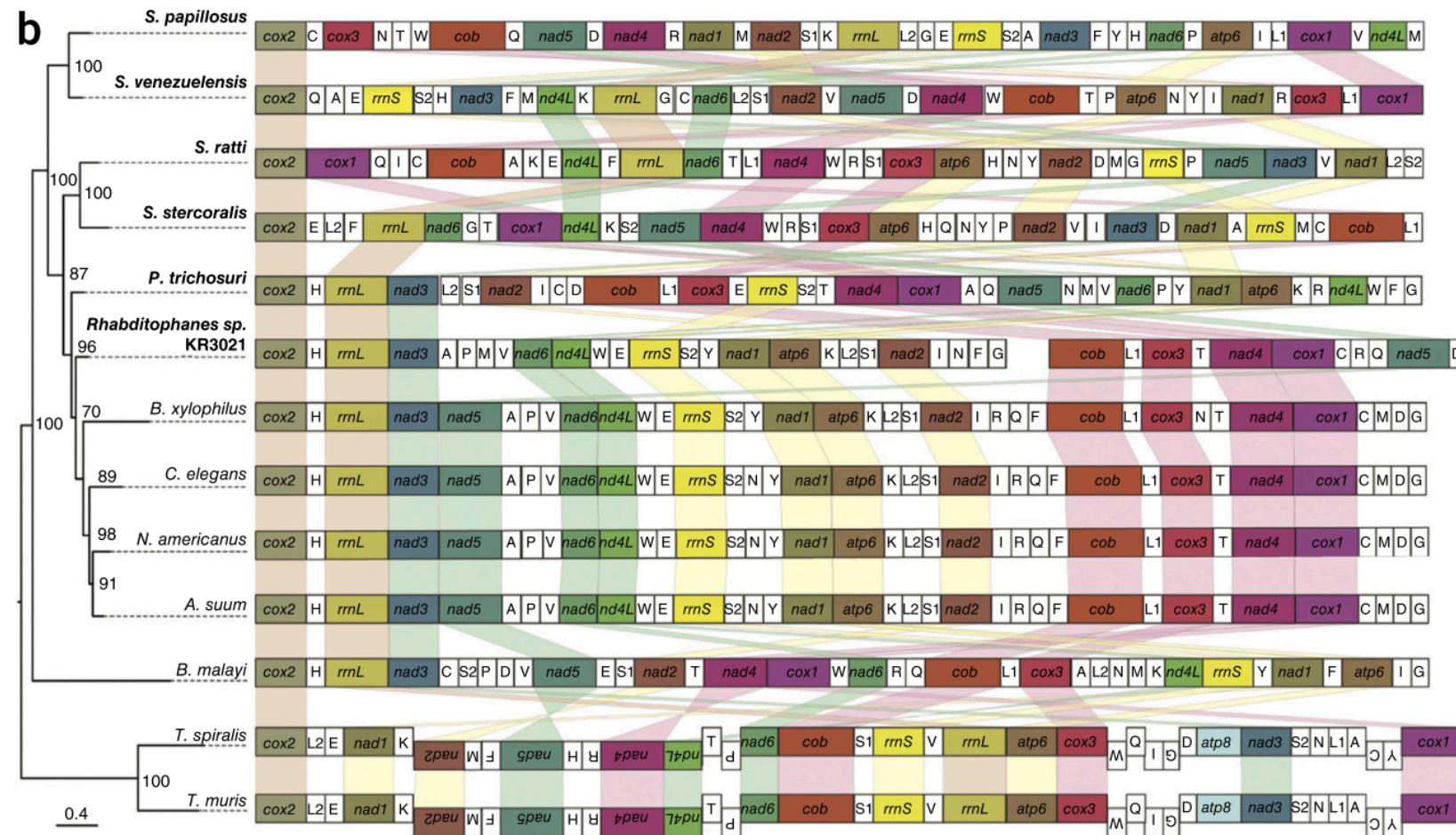


Aspergillus fumigatus



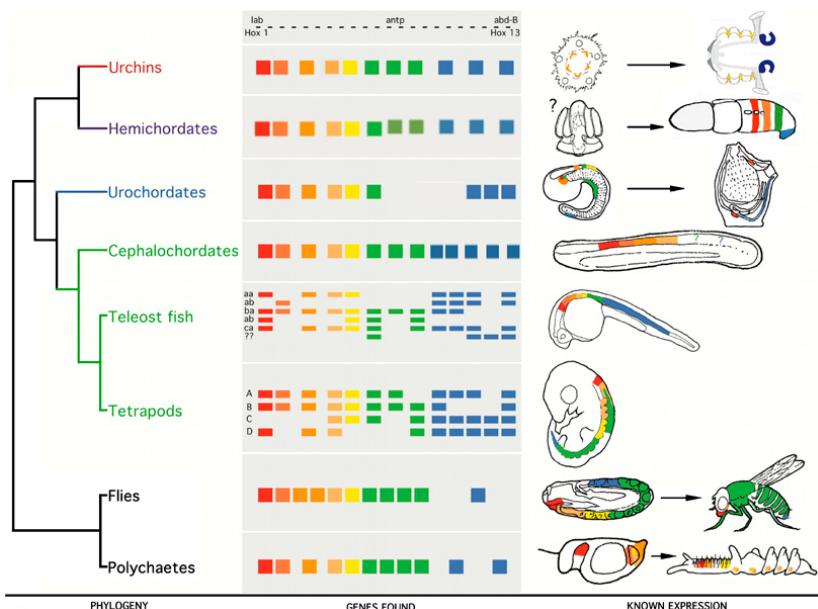
Why are we interested in synteny and collinearity?

Establish relationship between species

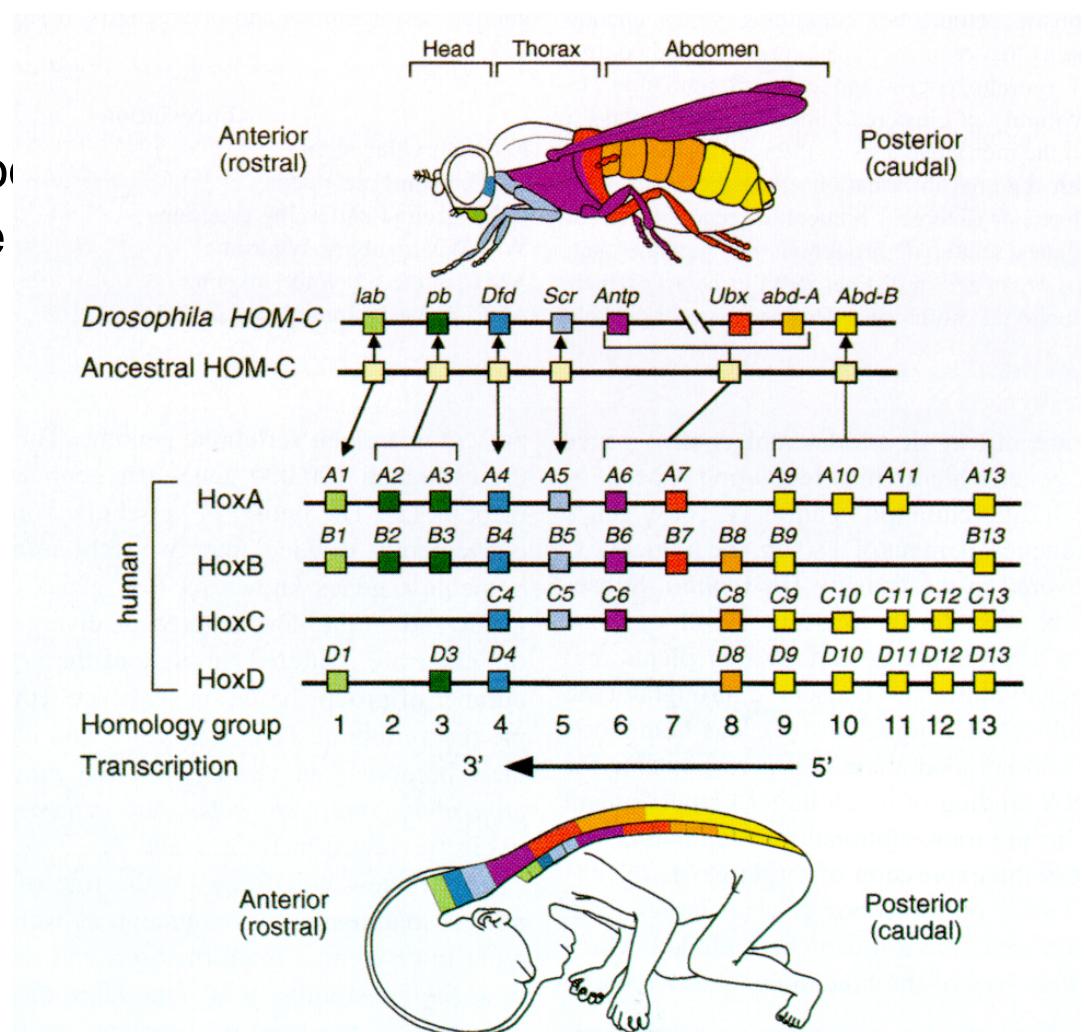


Why are we interested in synteny and collinearity?

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

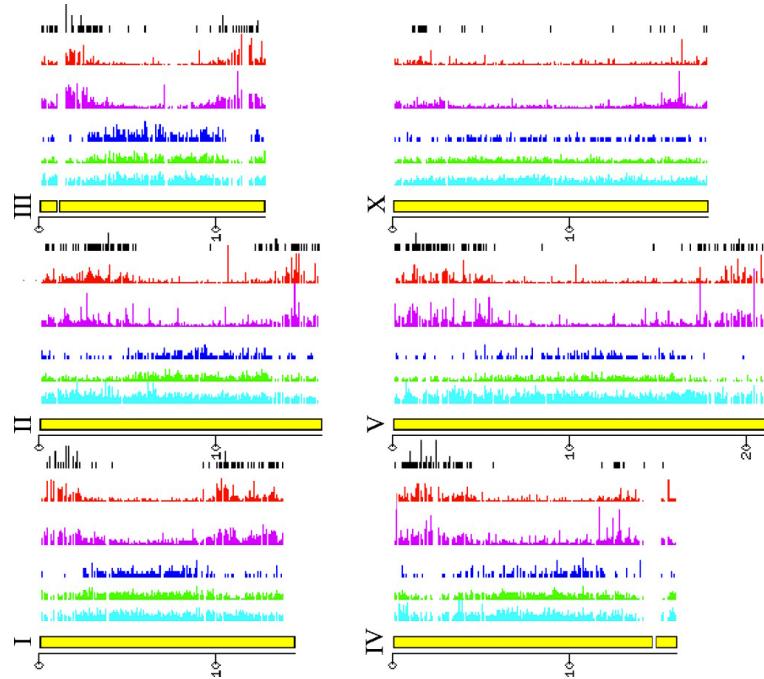


Swalla (2006) ; Mark et al (1997)



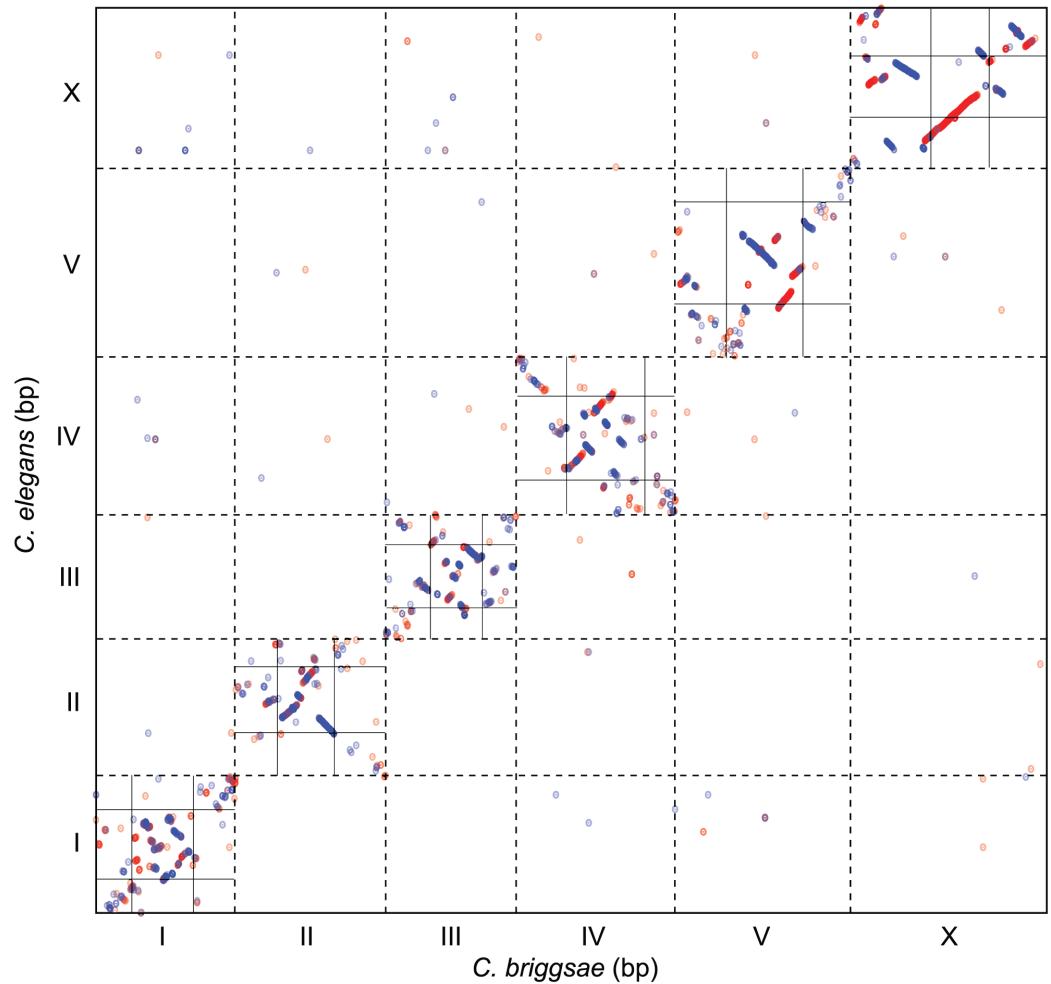
Why are we interested in synteny and collinearity?

**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

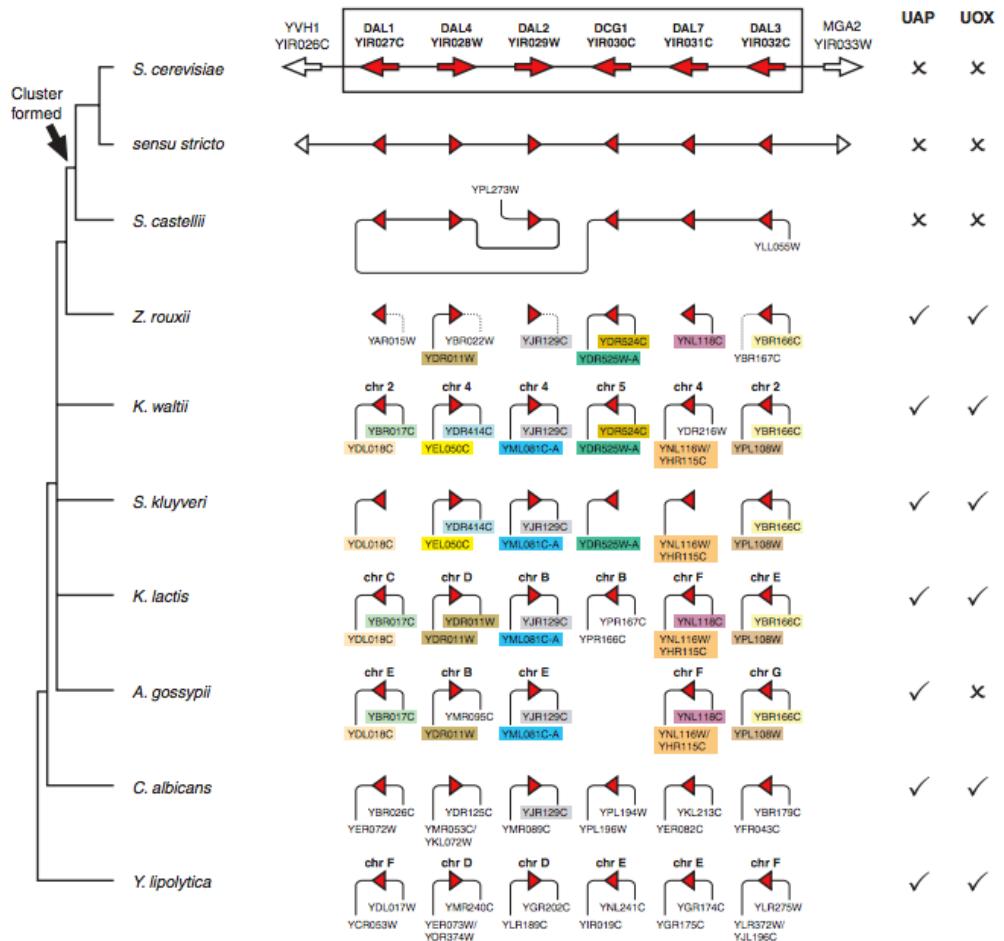
The *C. elegans* Sequencing Consortium Science 1998

Why are we interested in synteny and collinearity?

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

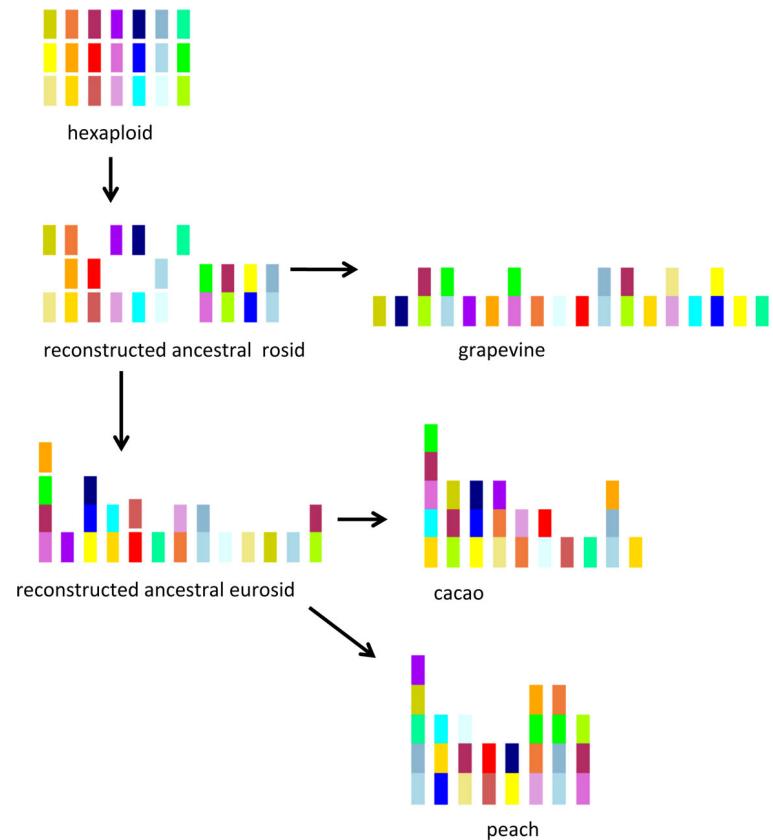
Birth of a metabolic gene cluster in yeast by adaptive gene relocation

Simon Wong & Kenneth H Wolfe

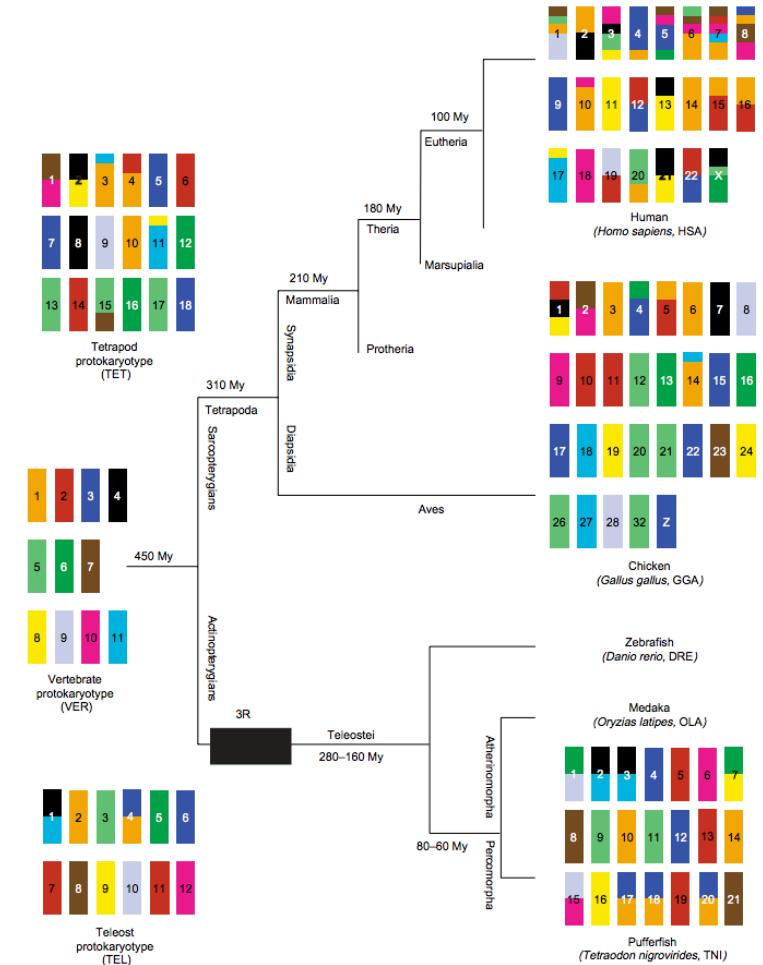


Why are we interested in synteny and collinearity?

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



Zheng et al (2013)

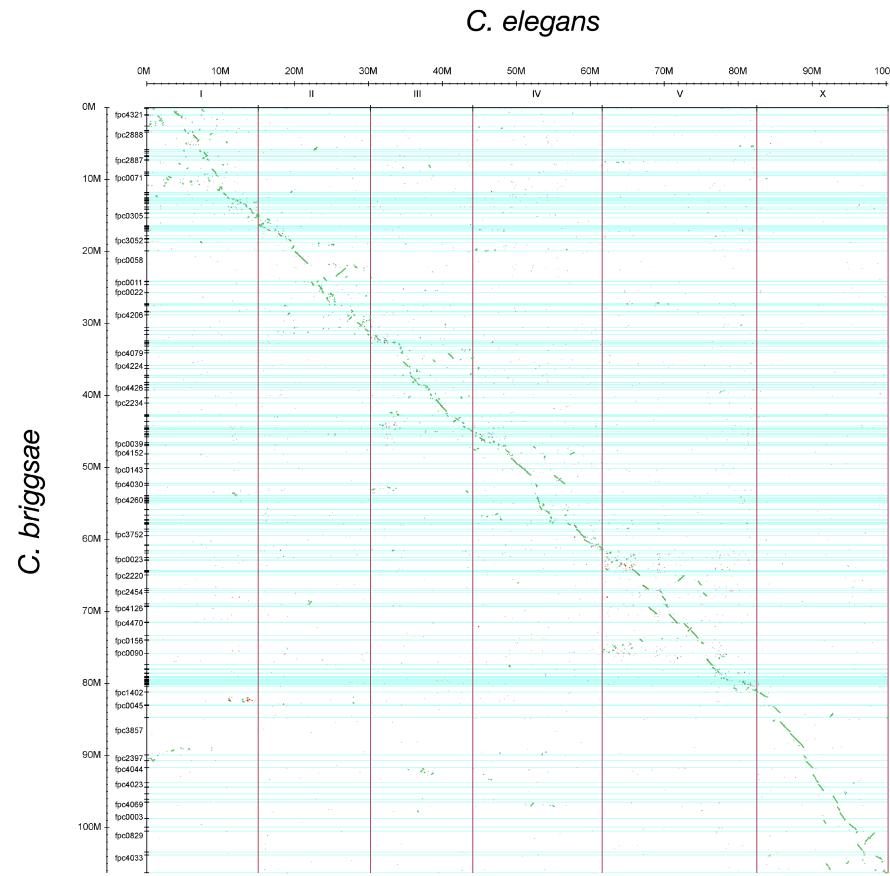


Kohn et al (2006)

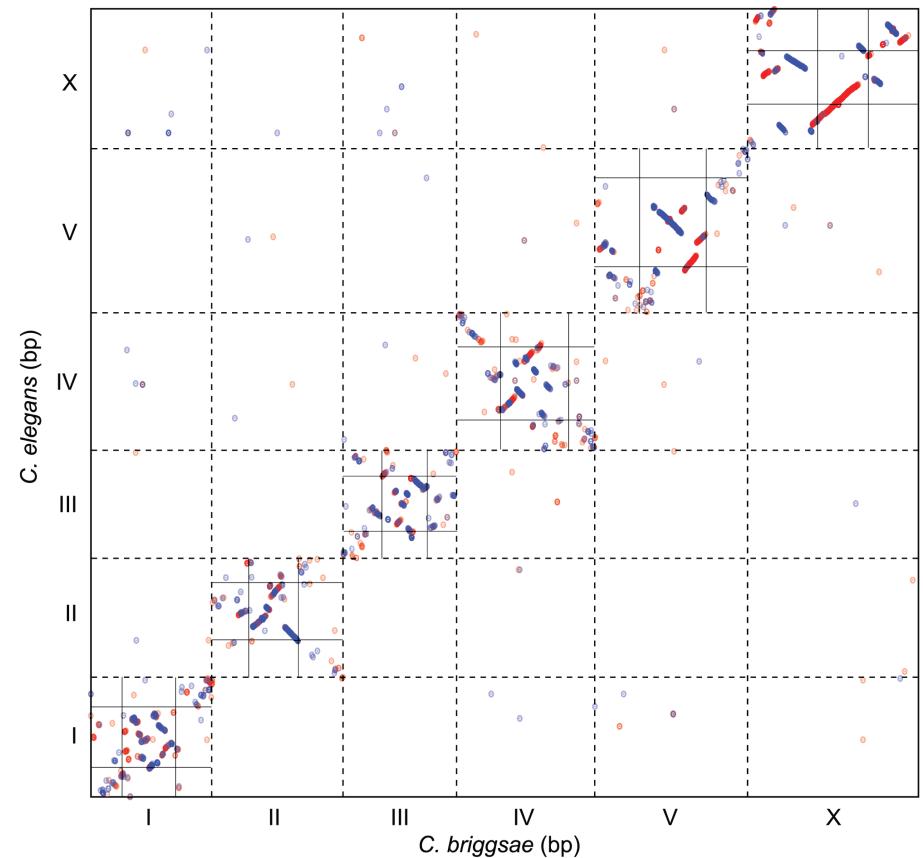
TRENDS in Genetics

Some caveats

Assembly quality likely to influence synteny observation



Stein et al., PLOS Genetics (2003)



Ross et al., PLOS Genetics (2011)

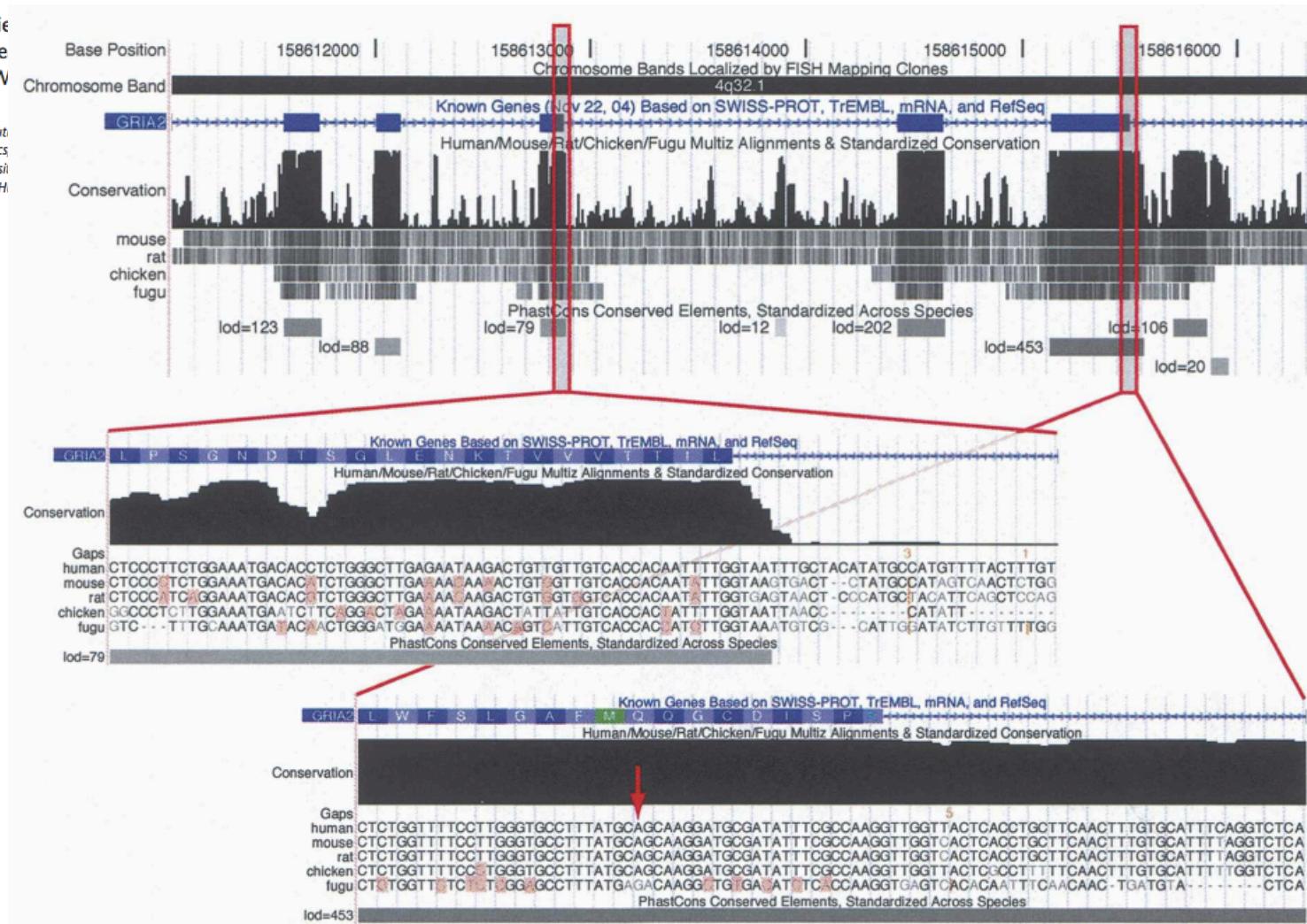
Comparing genomes beyond gene level

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

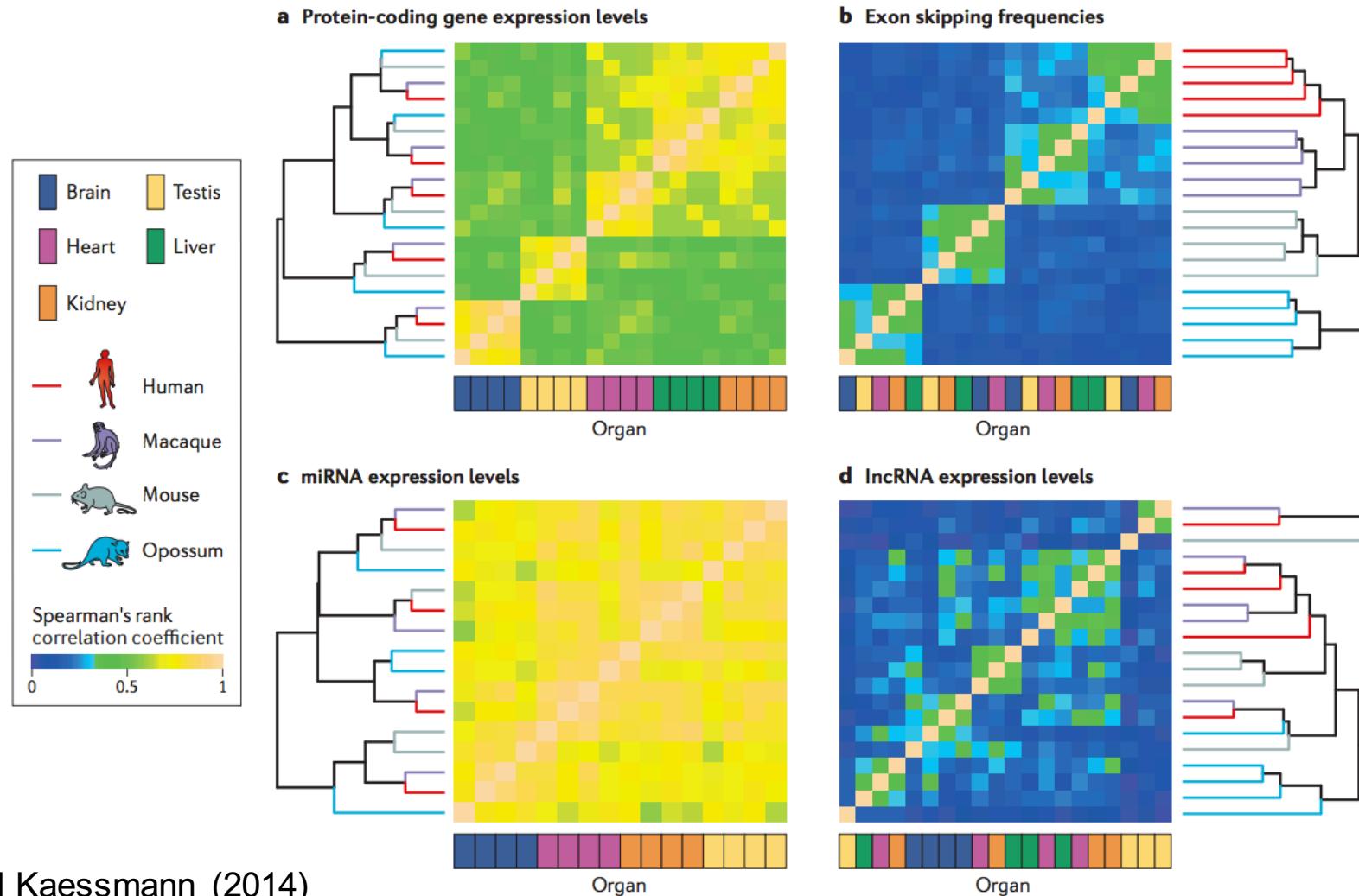
Adam Siepel,^{1,6} Gill Bejerano,¹ Jakob S. Pedersen,¹ Angie Kate Rosenbloom,¹ Hiram Clawson,¹ John Spieth,⁴ LaDe Stephen Richards,⁵ George M. Weinstock,⁵ Richard K. W. James Kent,¹ Webb Miller,³ and David Haussler^{1,2}

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

PhastCons

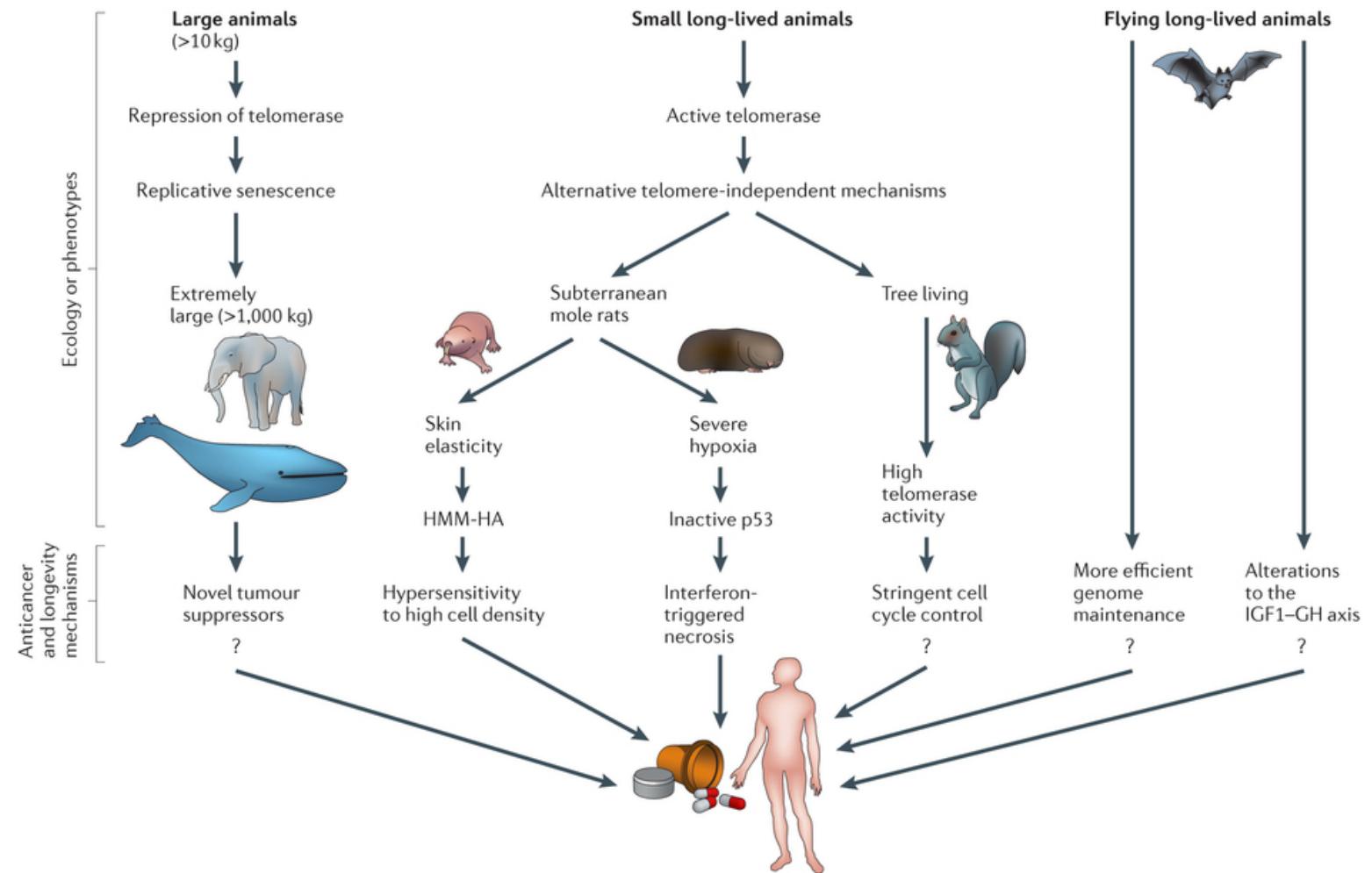


Global patterns of evolution for different aspects of the transcriptome



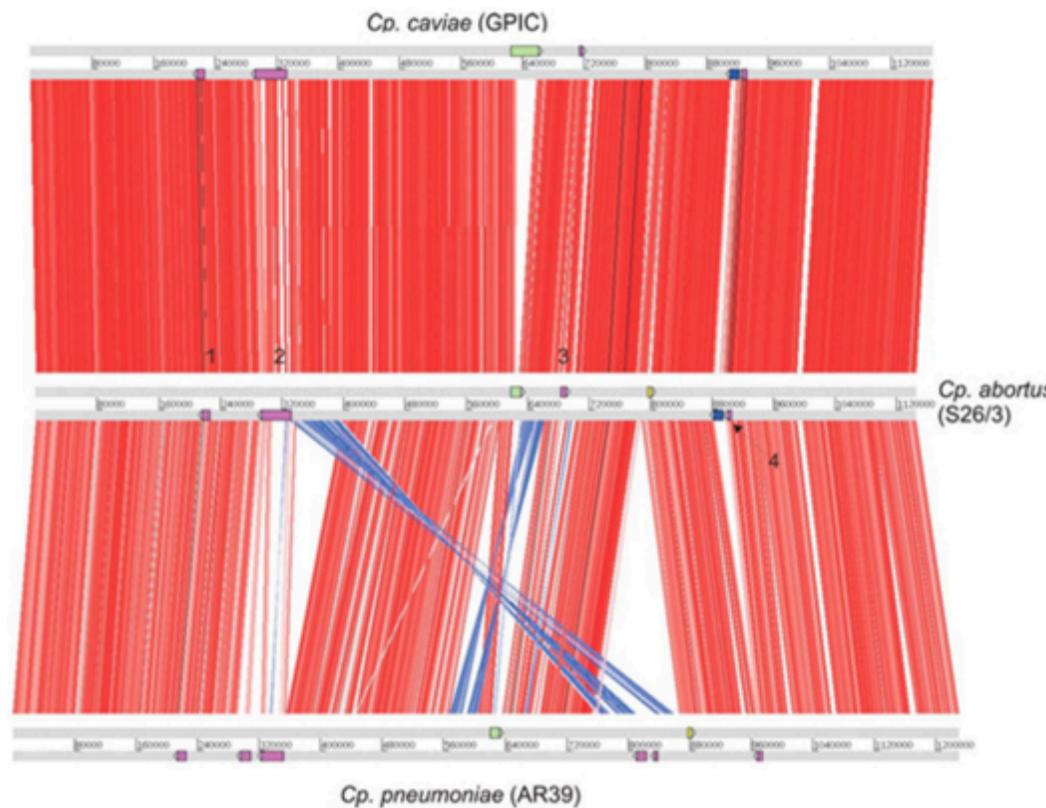
Necsulea and Kaessmann (2014)

Comparative genomics of longevity ageing (with focus)



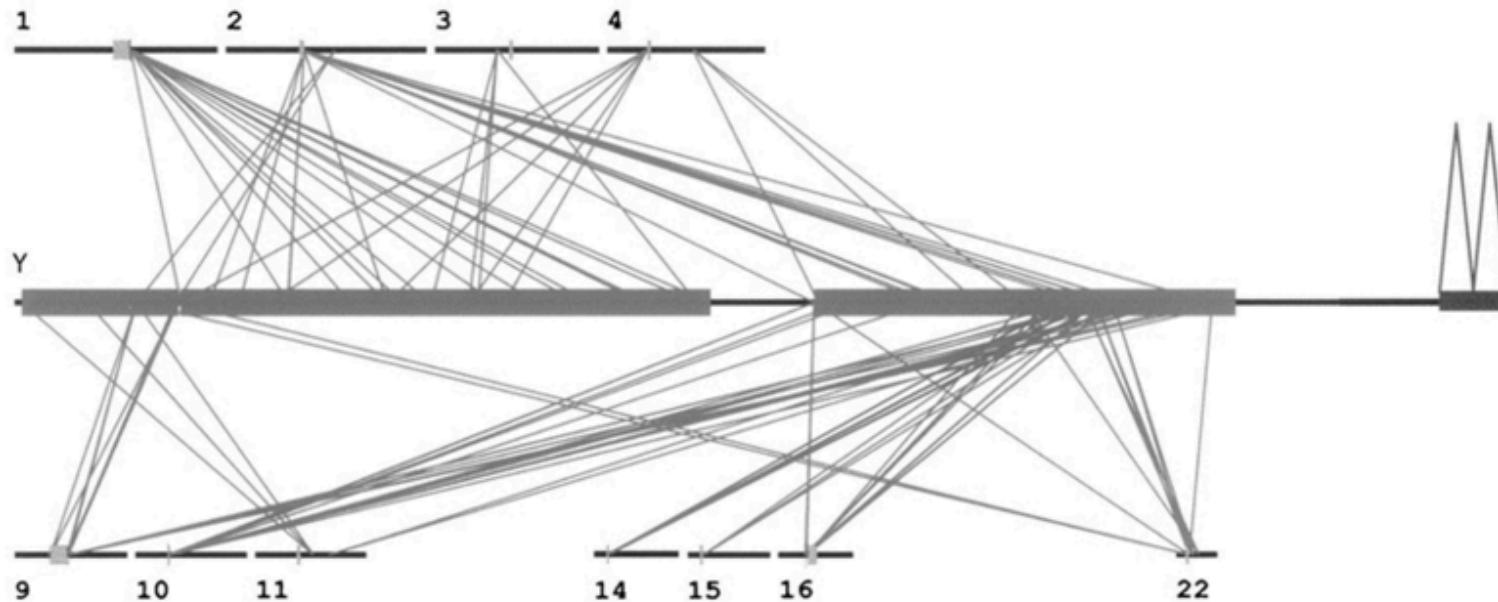
Genome visualisation

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



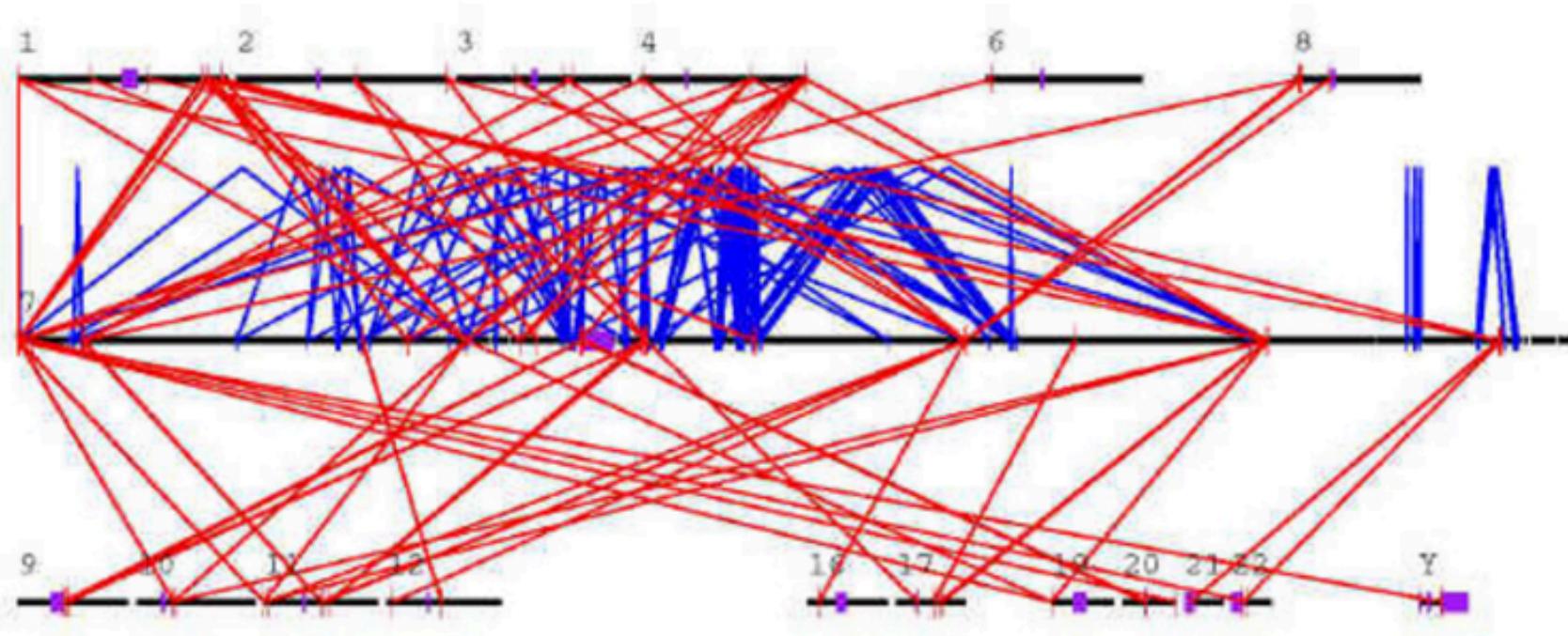
Genome Res. 2005 May;15(5):629-40

- 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

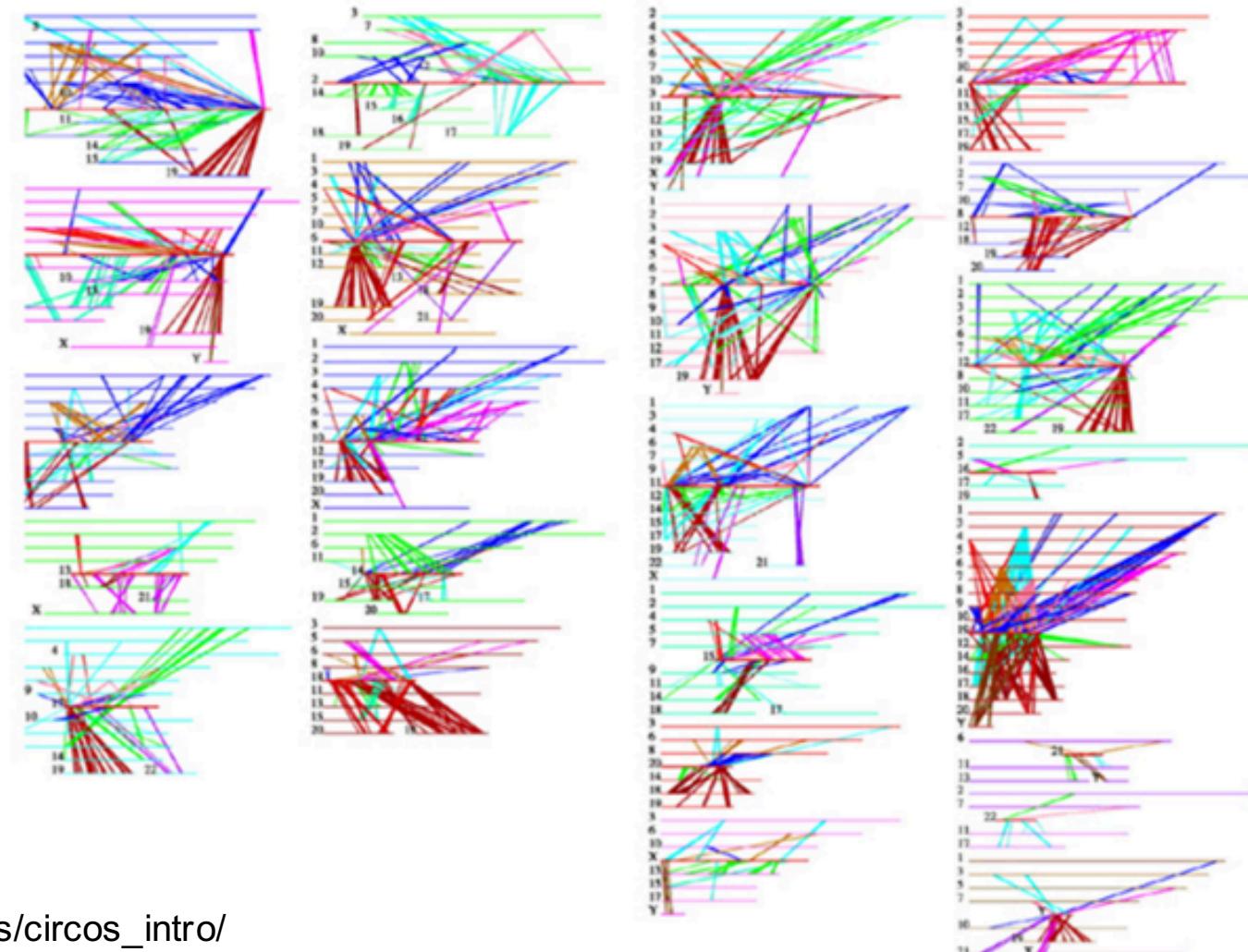
- things get worse and worse when mappings that link both neighbouring (blue) and distant (red) positions are shown



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

<http://www.genome.wustl.edu/projects/human/chr7paper/chr7data/o30113/segmental/index.php>

- 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.v

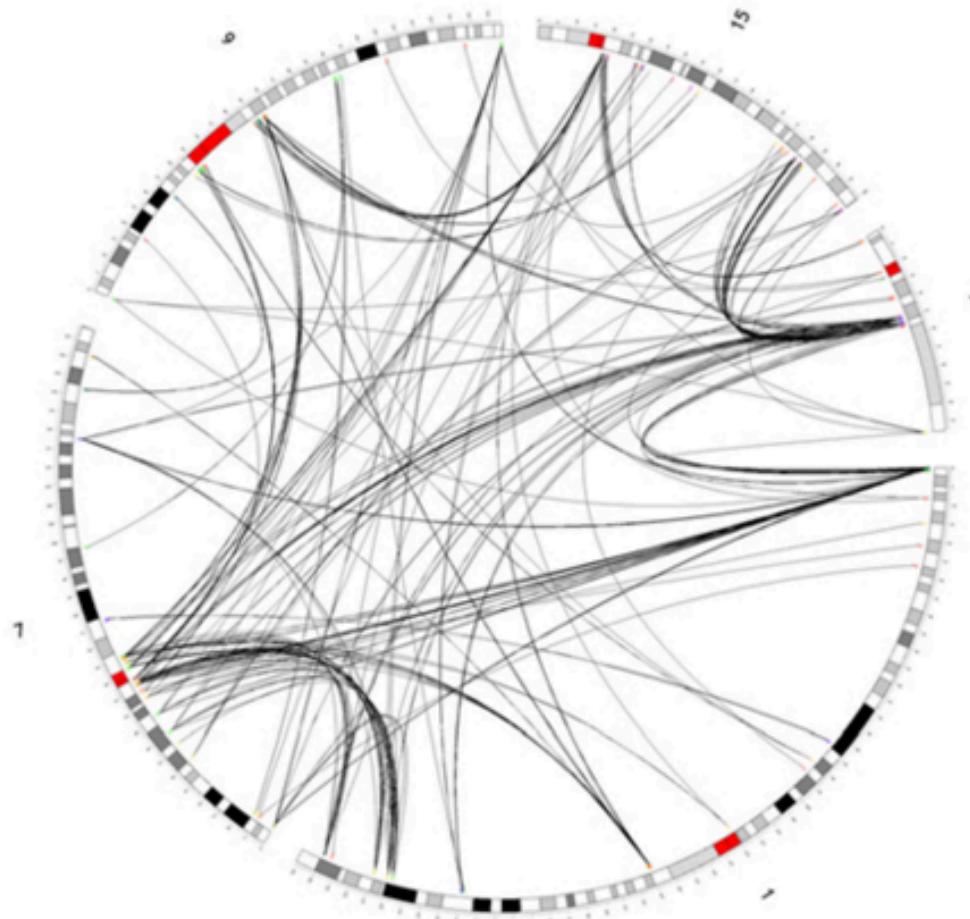
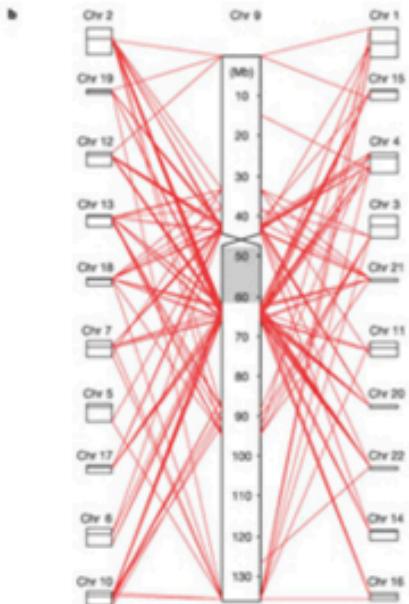
- finally, you descend into data overload and information hell
 - this is not an informative plot, although a pretty one



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

Image created with GenomePixelizer

Circos

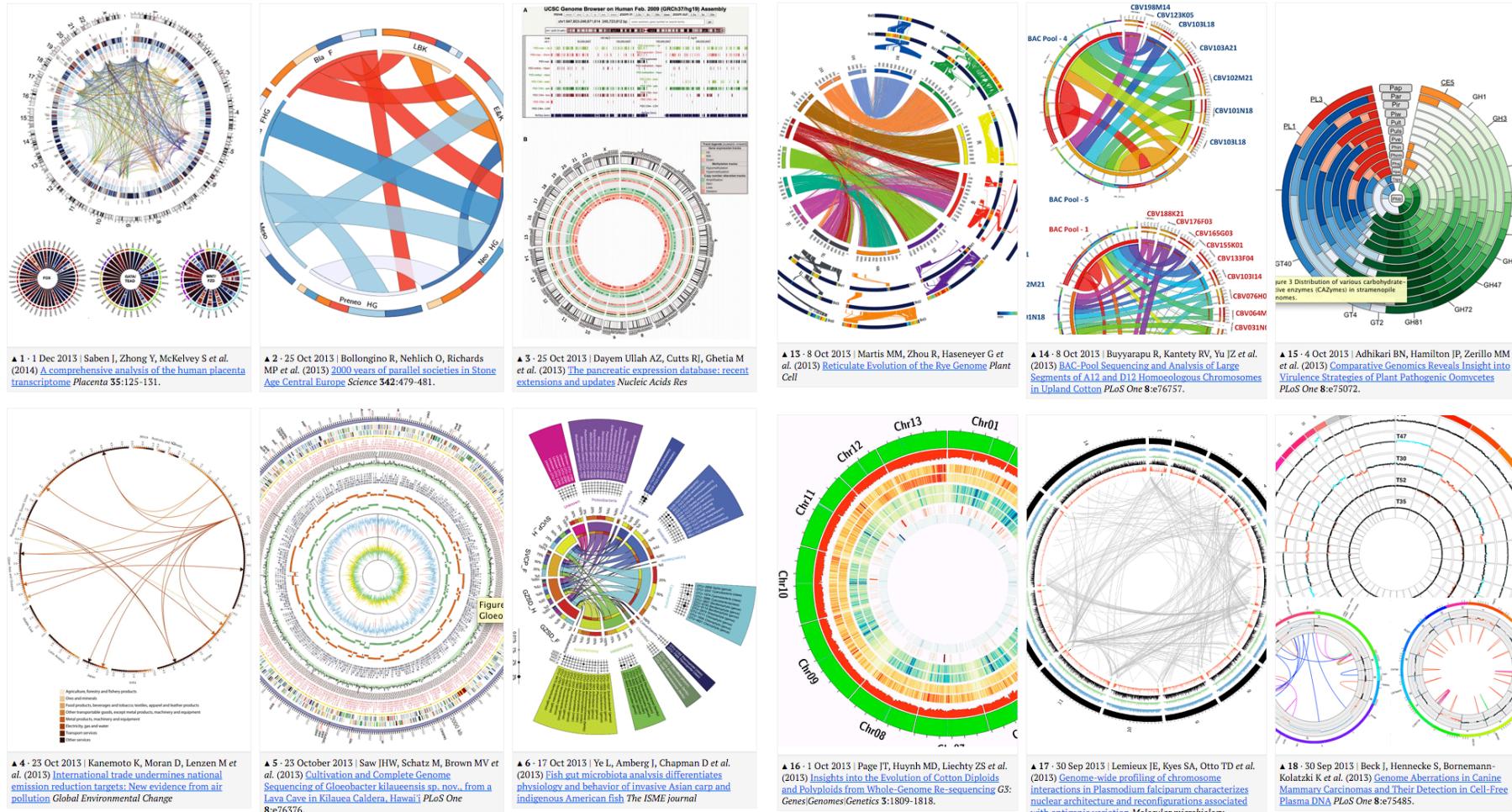


Circos image

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

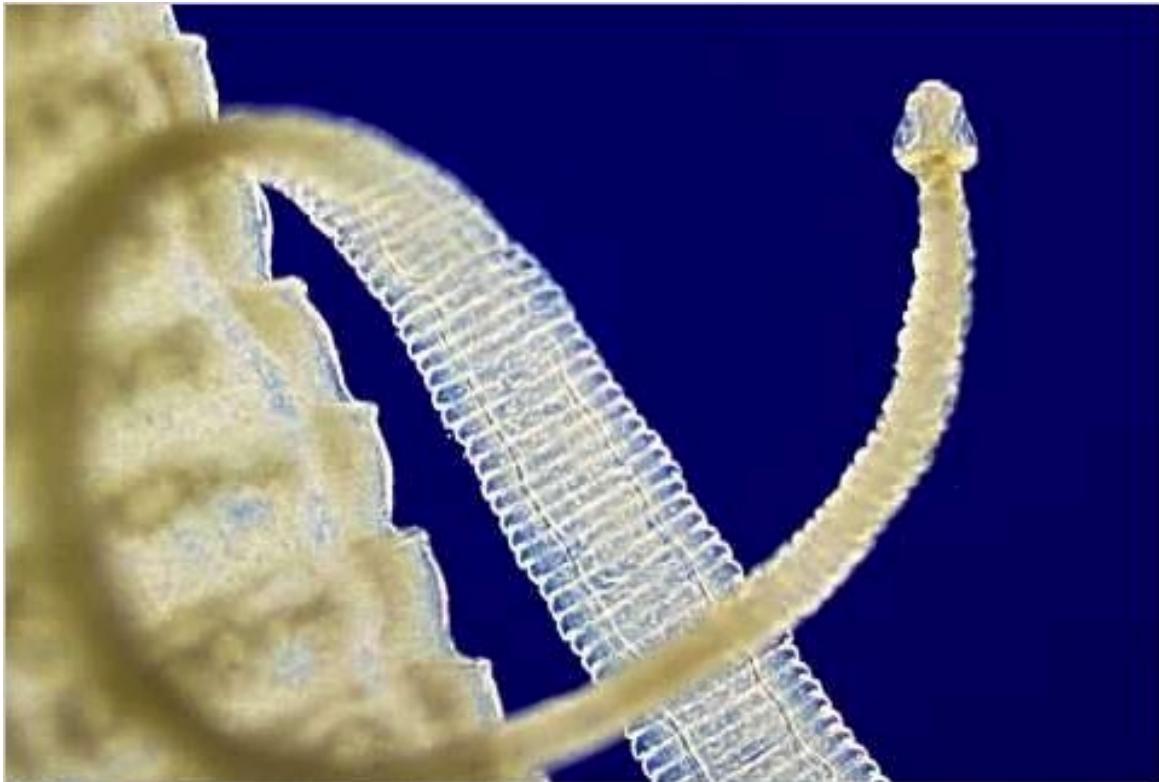
Humphray, S. J., K. Oliver, et al. (2004).
"DNA sequence and analysis of human chromosome 9." Nature 429(6990): 369-74.

Circos

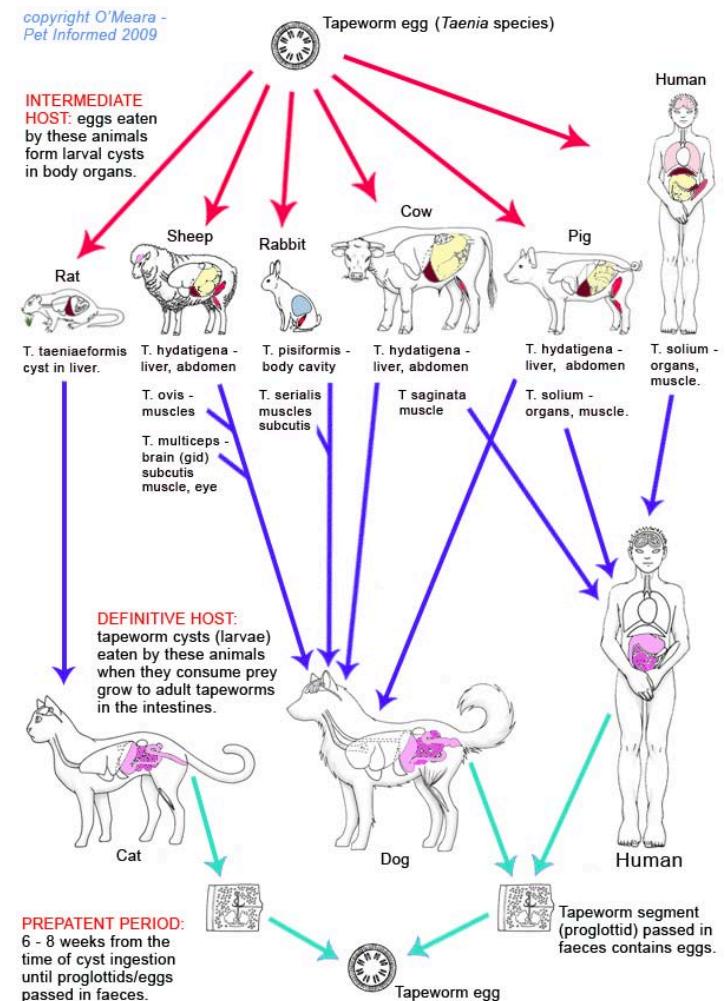


Case study: tapeworm genomes

Tapeworms

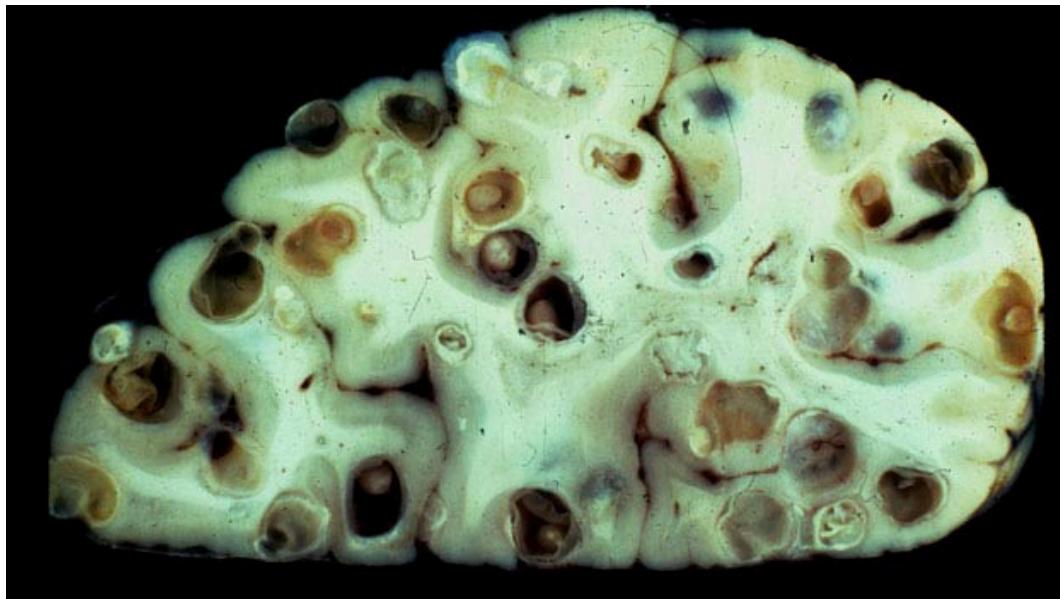


- Affects 10 million people worldwide
- Annual loss of US\$2 billion in livestock
- Genome size ~100Mb



Life cycle involving many hosts

Tapeworms can be deadly

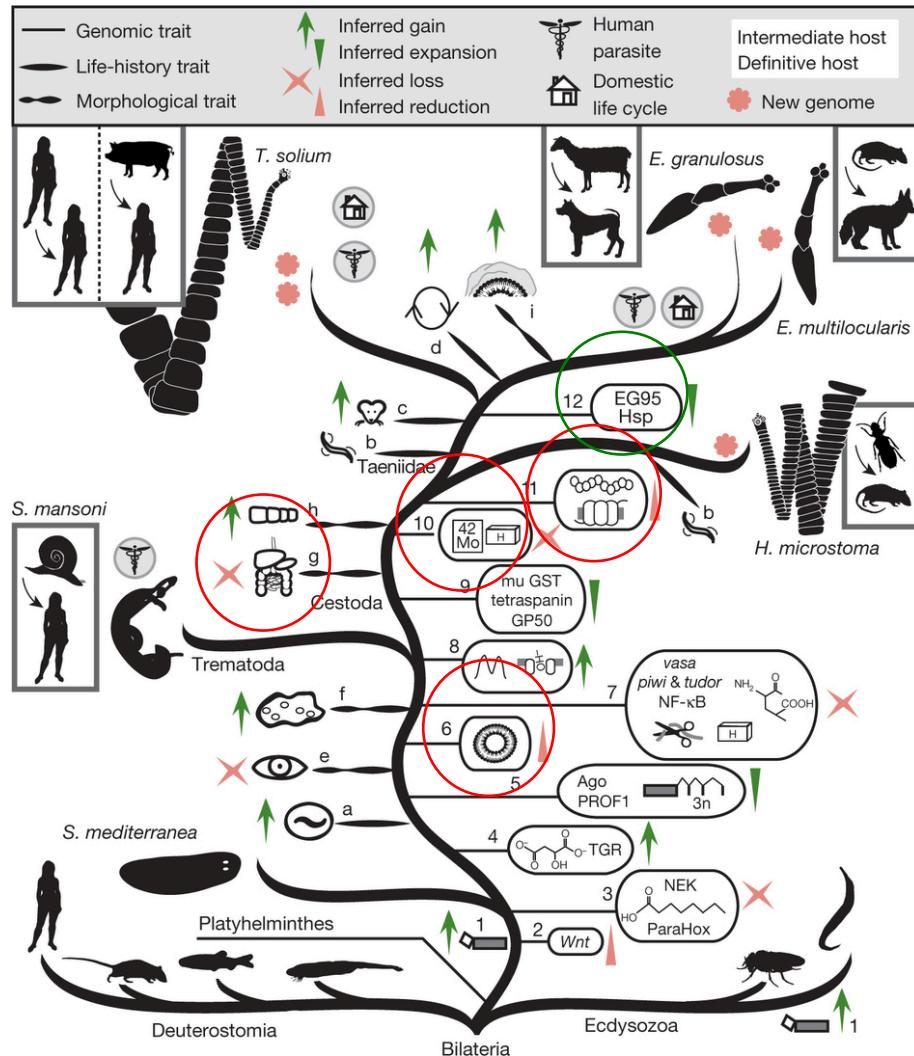


(c) 2012, Richard M. Jakowski, DVM, PhD, DACVP

cyst in different organs

Comparative genomics of tapeworms

tapeworms



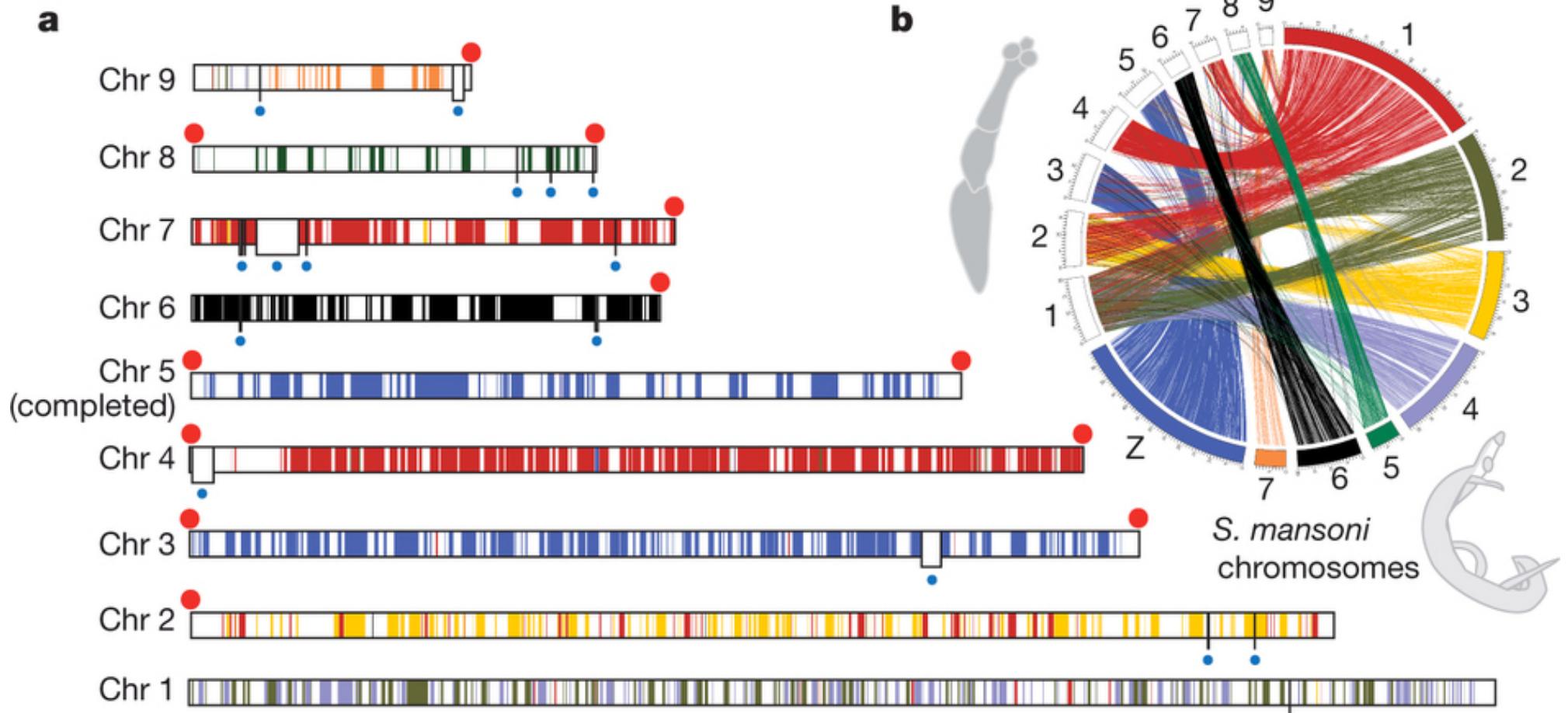
Blood fluke

Free-living

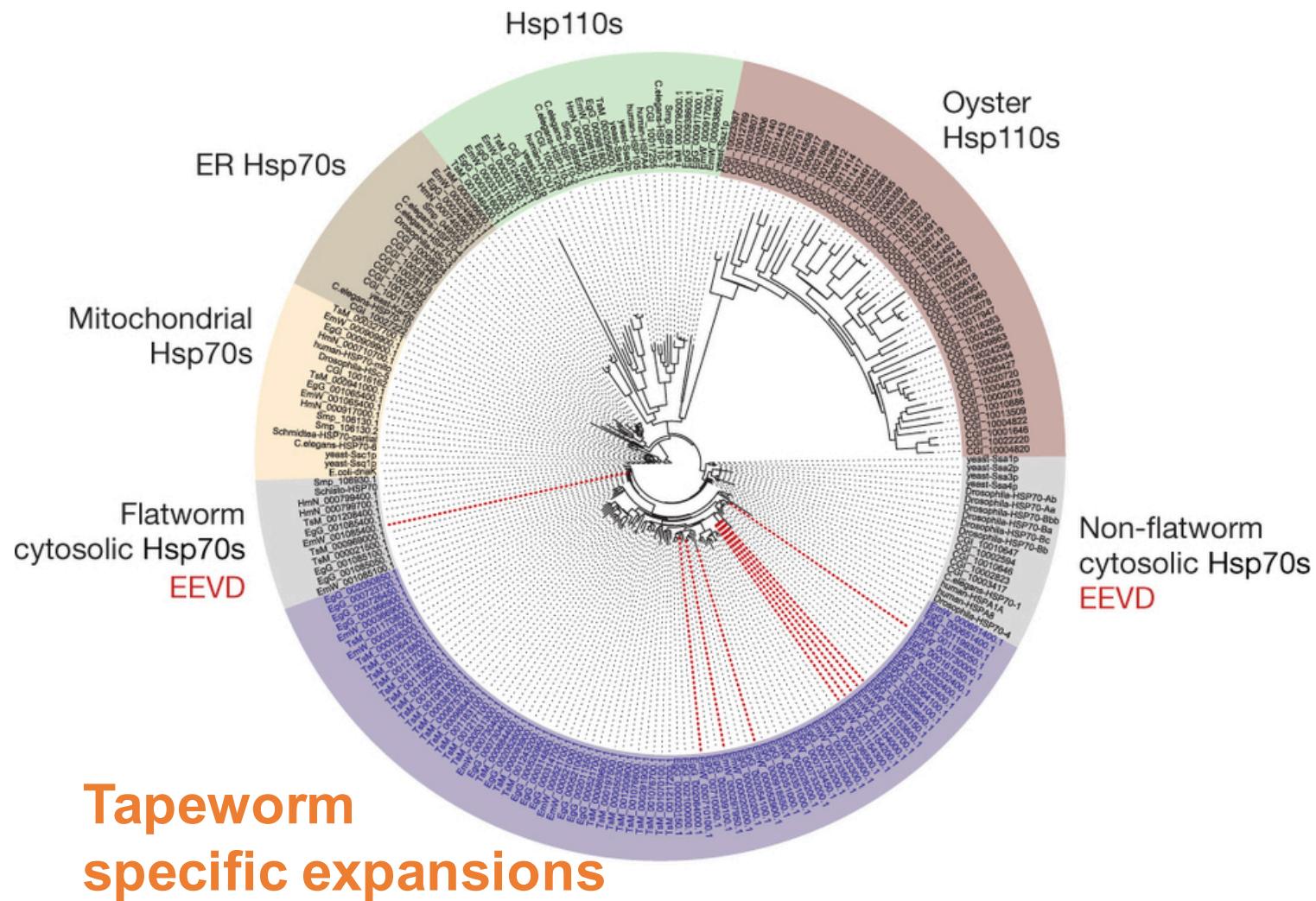
Model

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

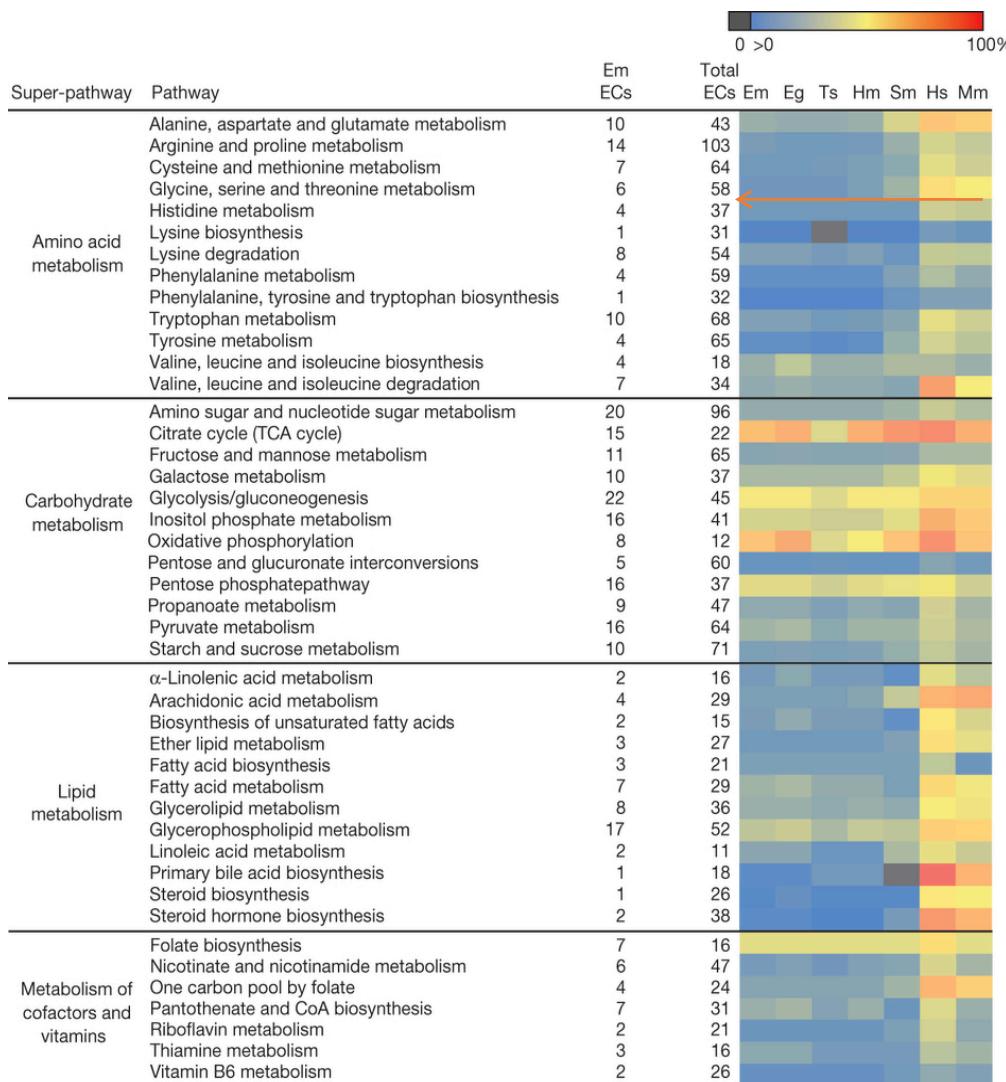
Genome of *E. multilocularis*



Heat shock protein expansion in tapeworms



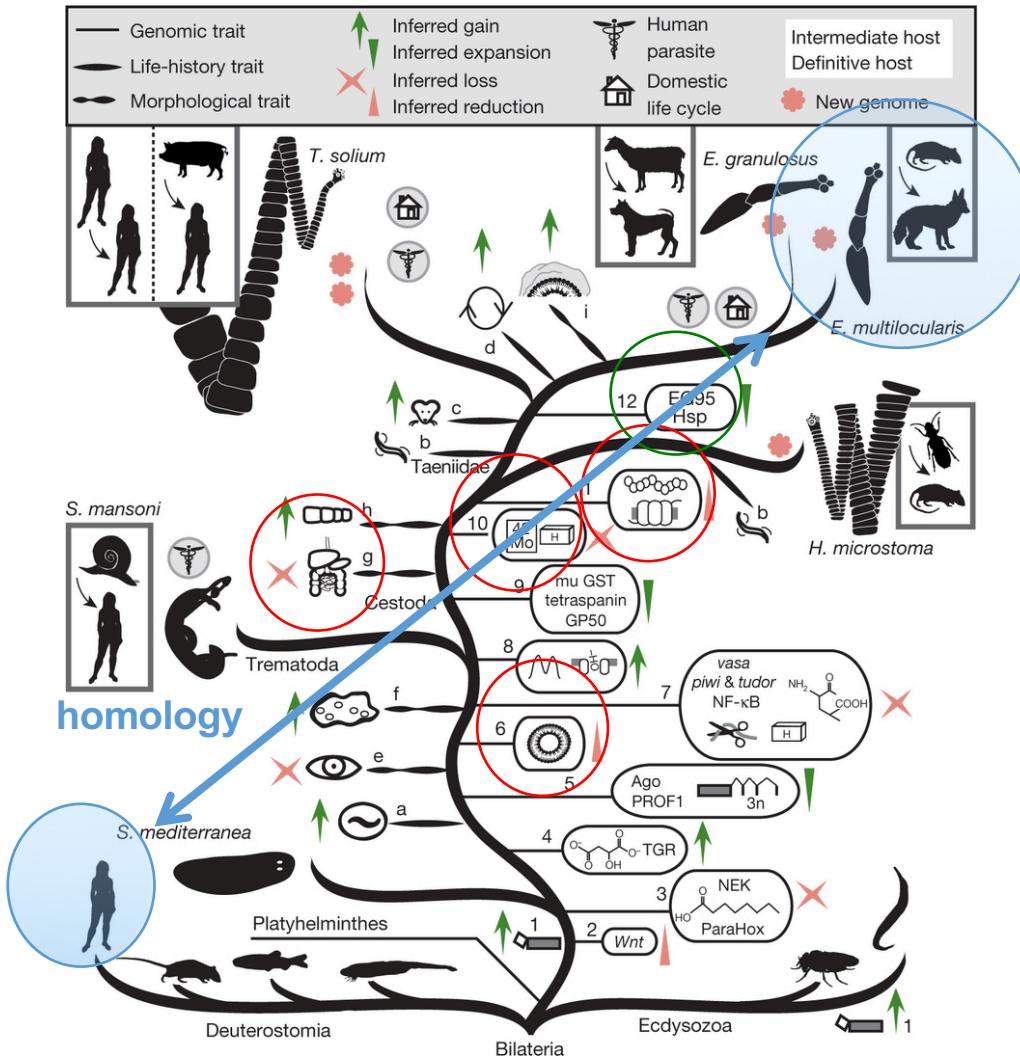
Reduced metabolism in tapeworms



Reduced metabolism

Tapeworm's road to parasitism

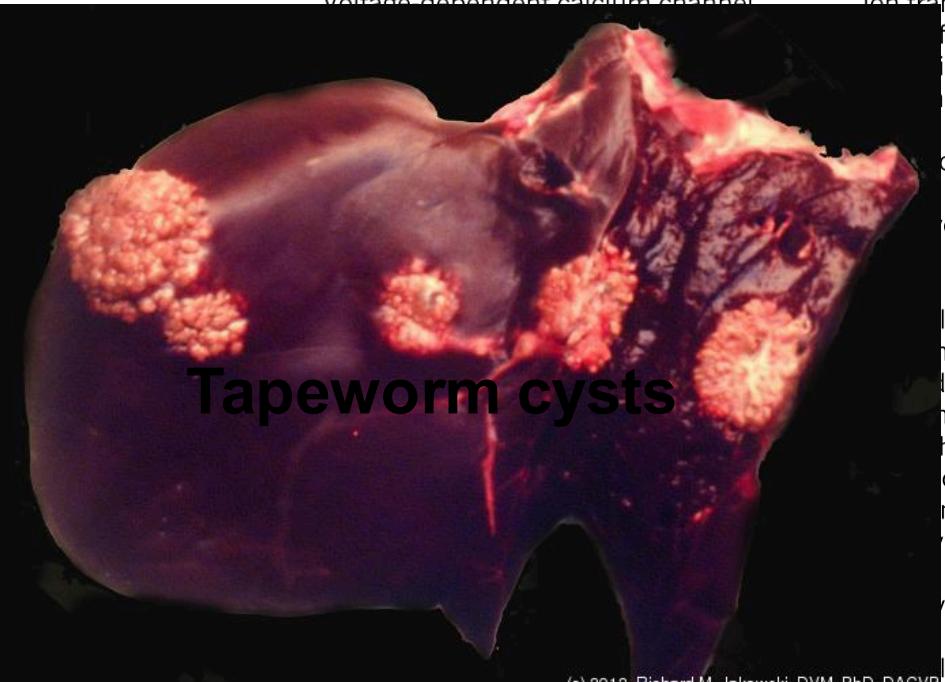
tapeworms



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
					
	(c) 2012, Richard M. Jakowski, DVM, PhD, DACVP				
	Elongation factor 2 Cathepsin B Dual-specificity mitogen activated protein Purine nucleoside phosphorylase	Translation Protease Signalling, activation of p38 Purine metabolism	M,A M M M,A	Lorazepam) Experimental compounds Experimental compounds Experimental compounds Didanosine	54 55 56 63

<http://en.wikipedia.org/wiki/Metastasis>

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

What we didn't discuss today

Complexities behind whole genome alignment

Detailed complexities behind orthology assignment