

Presentation and implementation of phylogenomics methods

Claire Hoede, PF Bioinfo, Genotoul

Outline

- Build the dataset:
 - What scale for inferring species phylogeny ?
 - Orthology inference
- Phylogenomics analysis
 - Whole genome features methods
 - Sequence based approaches:
 - Supermatrix
 - Supertree
- How to compare trees ?
- Conclusion

Why use more than one gene to reconstruct the evolutionary history of several species of interest ?

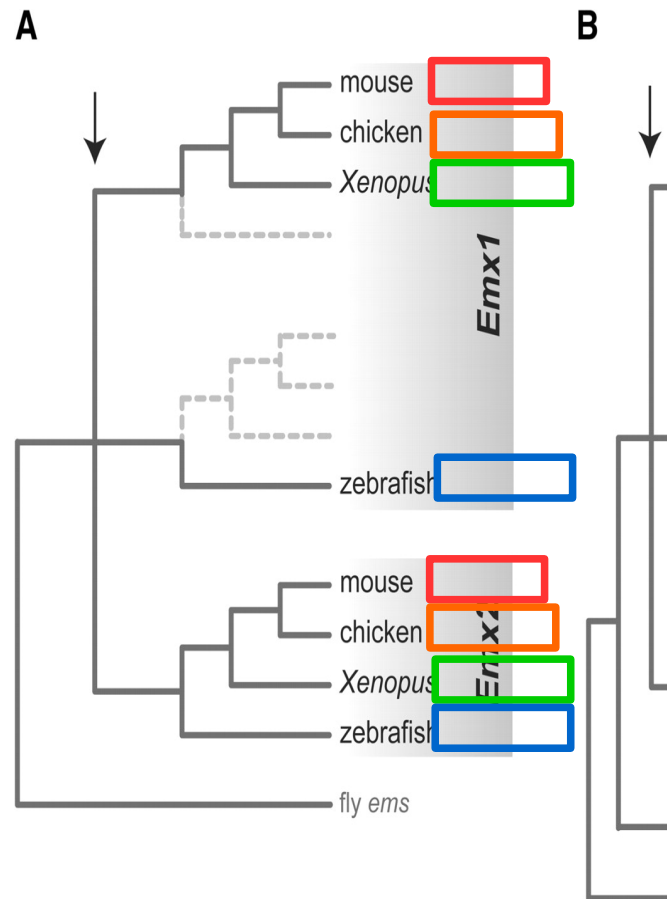
Limits of phylogenies based on a single gene

- Use a single gene allow to reconstruct the evolutionary history of the gene and not specifically of the corresponding OTU.
- The resolution can be poor.
- The evolutionary history of the gene may be different from that of the species because :
 - Hidden paralogy
 - Lateral gene transfer
 - Ancestral polymorphism

Sources of incongruence between the phylogeny of a gene and the evolutionary history of the species

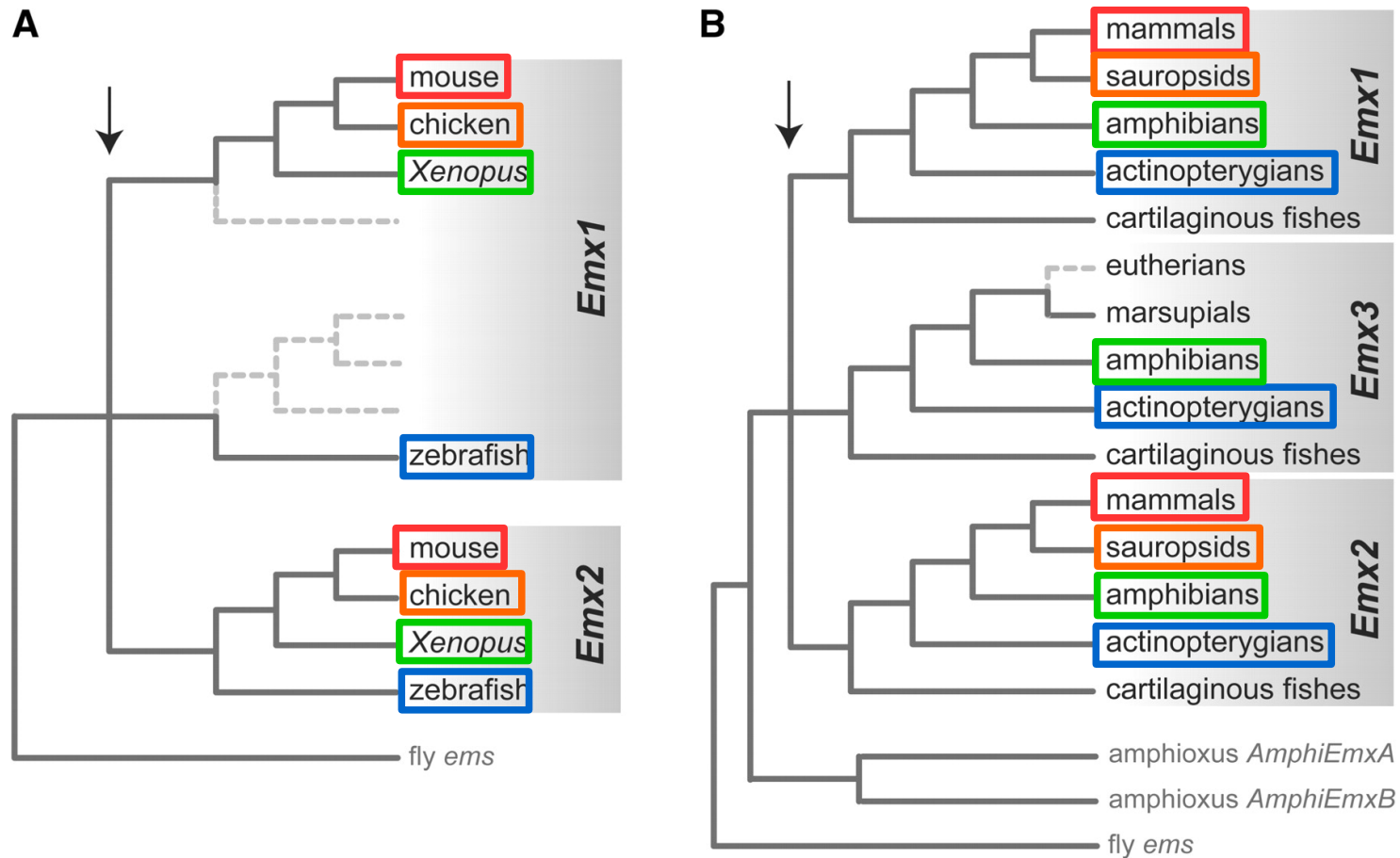
- Hidden paralogy (gene duplication followed by a loss)
- Lateral gene transfer (LGT)
- Ancestral polymorphism :
 - Trans-specific polymorphism (TSP : These alleles have diverged prior to speciation and this diversity is maintained)
 - Incomplete Lineage sorting (ILS : selection or genetic drift may cause alleles to be lost over time in one lineage but not another when two populations diverge)

Sources of incongruence: Hidden paralogy



Hidden paralogy in Emx gene phylogeny. Molecular phylogenetic trees of vertebrate Emx genes before the year 2000 (A) and now (B) are shown. Dotted lines indicate absences of relevant genes (gene loss or incomplete identification). Note that the zebrafish gene, initially recognized as *emx1* in (A) (Morita et al. 1995), was later found orthologous to *emx3* and renamed accordingly as shown in (B) (Kawahara and Dawid 2002). Arrows indicate gene duplications between gnathostome paralogs.

Sources of incongruence: Hidden paralogy



Hidden paralogy in Emx gene phylogeny. Molecular phylogenetic trees of vertebrate Emx genes before the year 2000 (A) and now (B) are shown. Dotted lines indicate absences of relevant genes (gene loss or incomplete identification). Note that the zebrafish gene, initially recognized as emx1 in (A) (Morita et al. 1995), was later found orthologous to emx3 and renamed accordingly as shown in (B) (Kawahara and Dawid 2002). Arrows indicate gene duplications between gnathostome paralogous.

Sources of incongruence: lateral gene transfer

Class 1 Phylogeny of HMG-CoA reductase in several kingdom

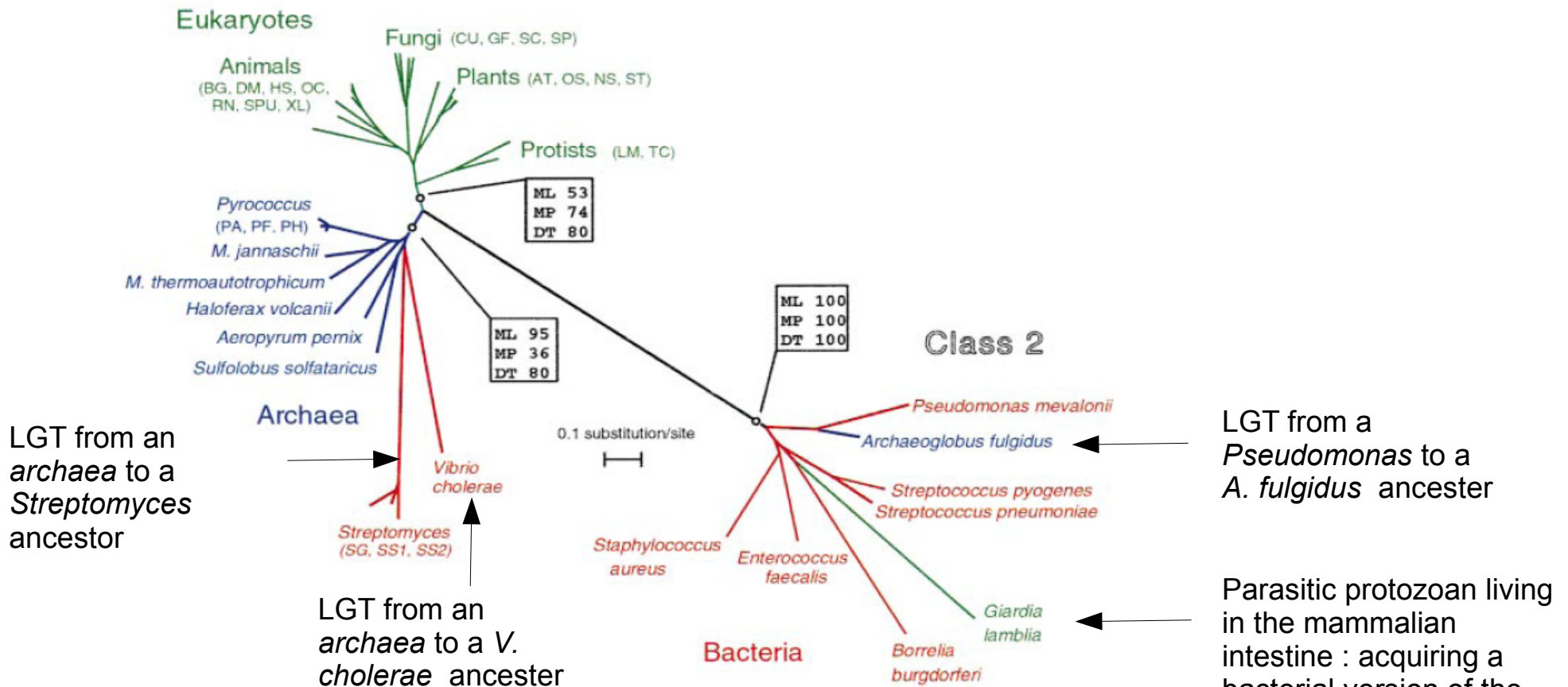
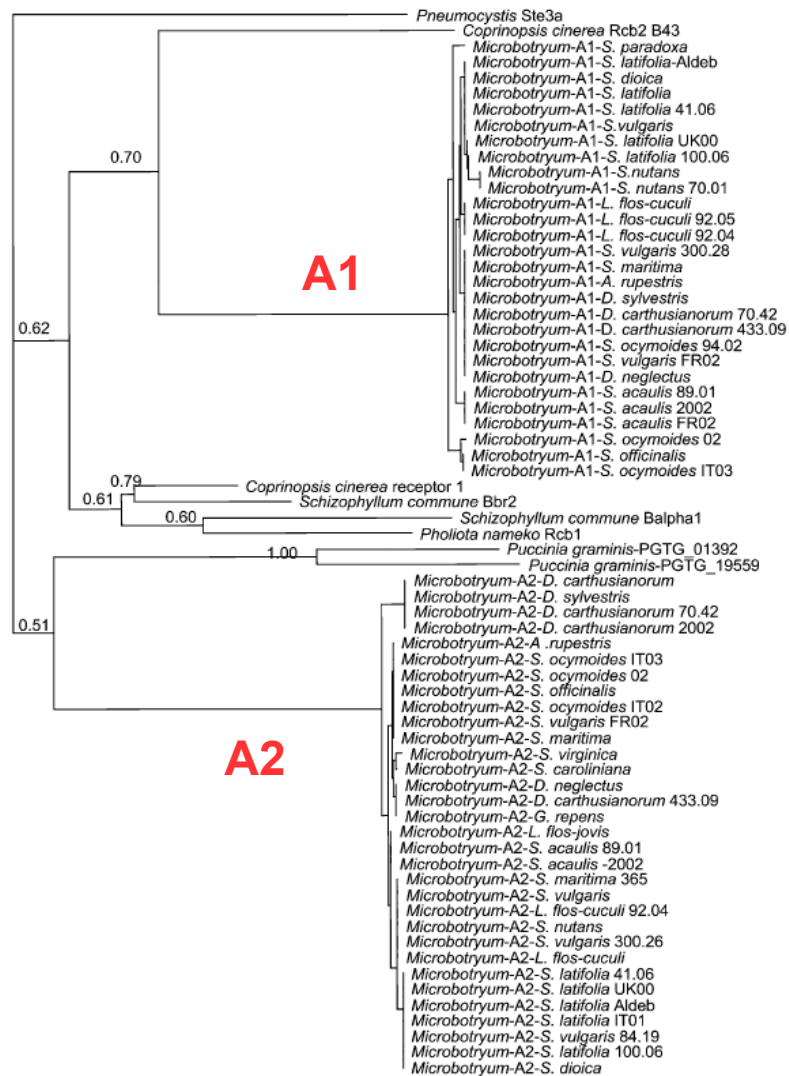


Fig. 3. Phylogeny of HMG-CoA reductase. A subset of 37 taxa from the alignment of all known HMGR protein sequences was used to carry out the analysis. The distance tree shown was determined using PROTDIST with PAM distances and branch lengths calculated with FITCH (PHYLIP 3.57; Felsenstein, 1993). The support values for important nodes of the tree are shown in boxes. (DT) percentage of distance bootstrap replicates supporting this topology using PROTDIST with PAM distances. SEQBOOT was used to generate 1000 bootstrap replicates, and the consensus tree was derived using CONSENSE. (ML) protML REL values obtained using a quick-add search of 1000 trees and the JTT-F substitution model. (MP) bootstrap support for the consensus tree obtained from PROTPARS with 1000 bootstrap replicates. Organism names are

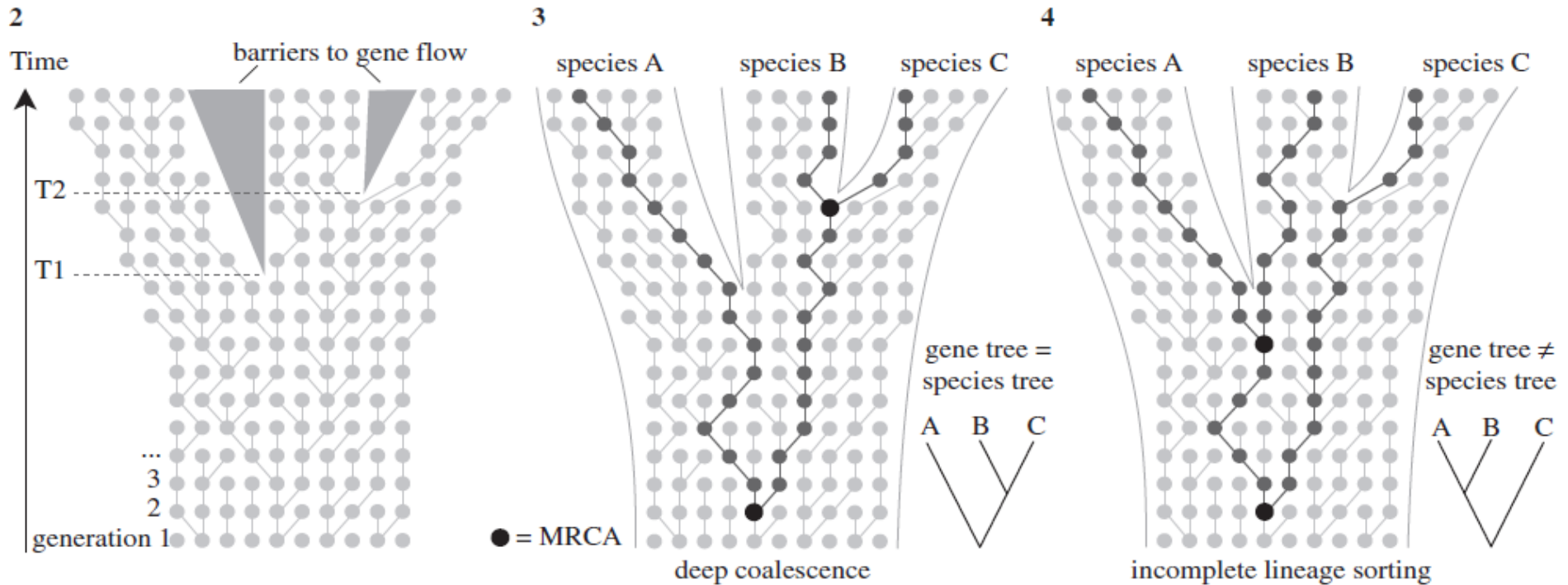
Sources of incongruence: trans-specific polymorphism



Phylogeny based on the pheromone receptor pr-MatA1 and pr-MatA2 of *Microbotryum* and other fungi.

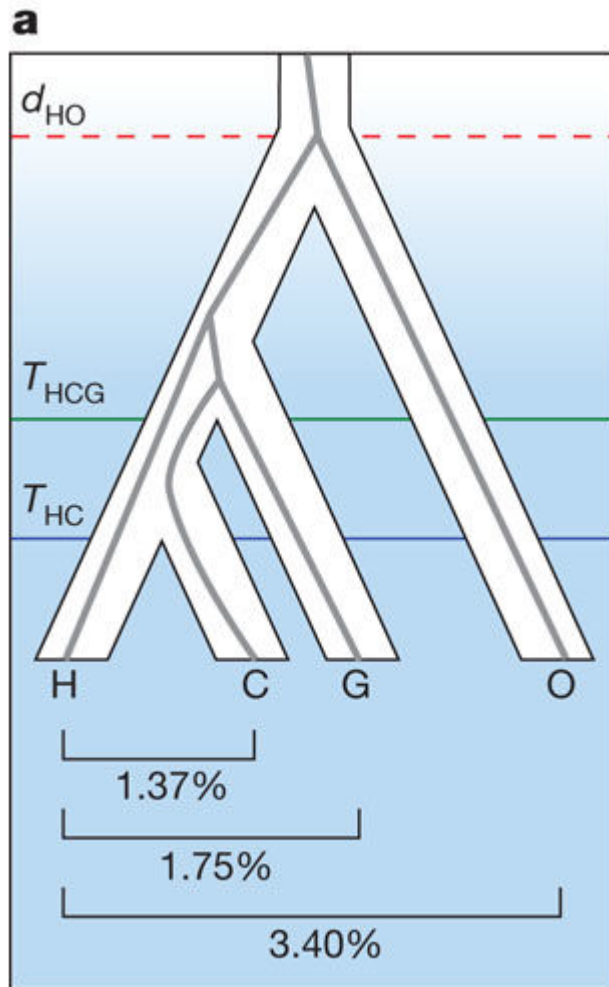
Trans-specific polymorphism: an allele sampled from a particular species can be more related of the same functional allelic class in other species than to members of different allelic classes in the same species (extrem case of balancing selection).

Sources of incongruence: incomplete lineage sorting



- **Incomplete lineage sorting (ILS):** the maintenance of genetic variation within a metapopulation lineage from one speciation event to the next, resulting in deep coalescence and gene tree–species tree incongruence (Baum & Smith, 2012).
- **Lineage sorting:** the process by which alleles are inherited and lost over time
- **Deep coalescence:** coalescence of alleles occurring significantly earlier than the divergence of the species containing those alleles

Sources of incongruence: incomplete lineage sorting



Phylogeny of the great ape family, showing the speciation of human (H), chimpanzee (C), gorilla (G) and orang-utan (O). Horizontal lines indicate speciation times within the hominine subfamily and the sequence divergence time between human and orang-utan. Interior grey lines illustrate an example of incomplete lineage sorting at a particular genetic locus—in this case $((C, G), H), O$ rather than $((H, C), G), O$. Below are mean nucleotide divergences between human and the other great apes from the EPO alignment.

The Chimpanzee and the Human are the most recently speciated. But the Gorilla and the Human are the most recently diverged, in the flow of one particular gene.

Sources of incongruence: incomplete lineage sorting

- ILS: a persistence of polymorphisms across multiple successive speciation events followed by stochastic allele fixation in each descendant lineage.
- Scally *et al.* (Nature, 2012) found 30 % of bases exhibiting ILS between human, chimpanzee and gorilla across the genome.
- When speciation is more rapid than the sorting of genes (in large population for exemple), the sorting along species lines can be incomplete.
- ILS is more likely to occur if the distance between branchings is short (speciation temporally close).

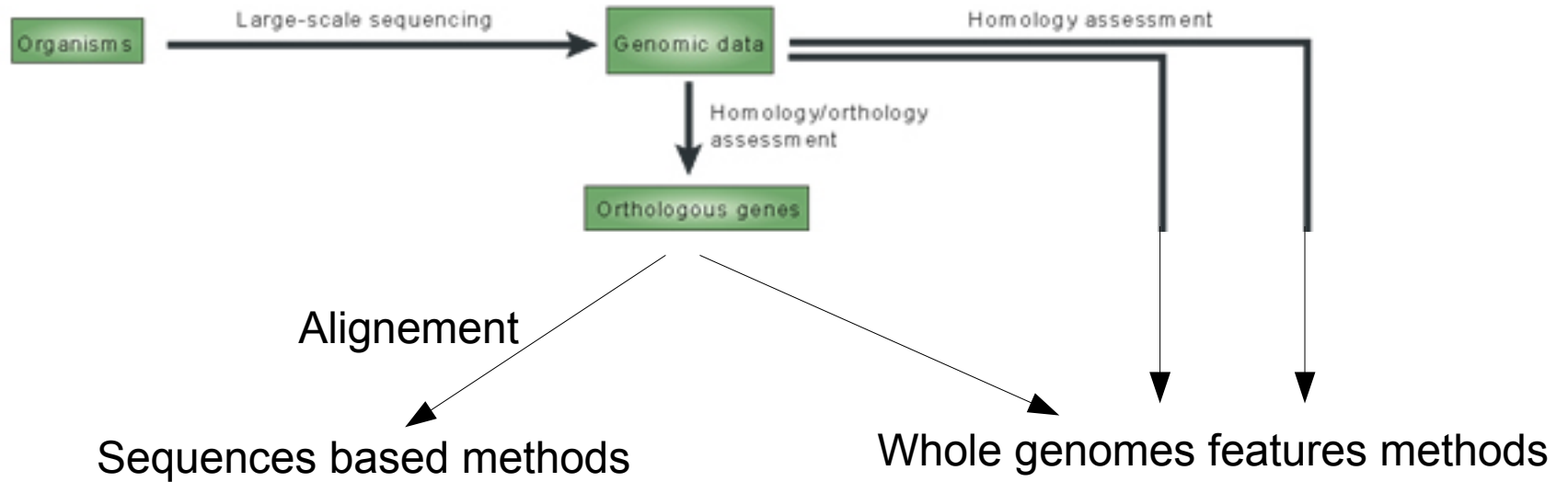
23/10/17 • Balancing selection increases ILS.

There is a lot of inconsistency sources in individual gene data, so in practice we integrate a lot of informations by assuming that the phylogenetic signal that we want is dominant.

Definition of possible errors

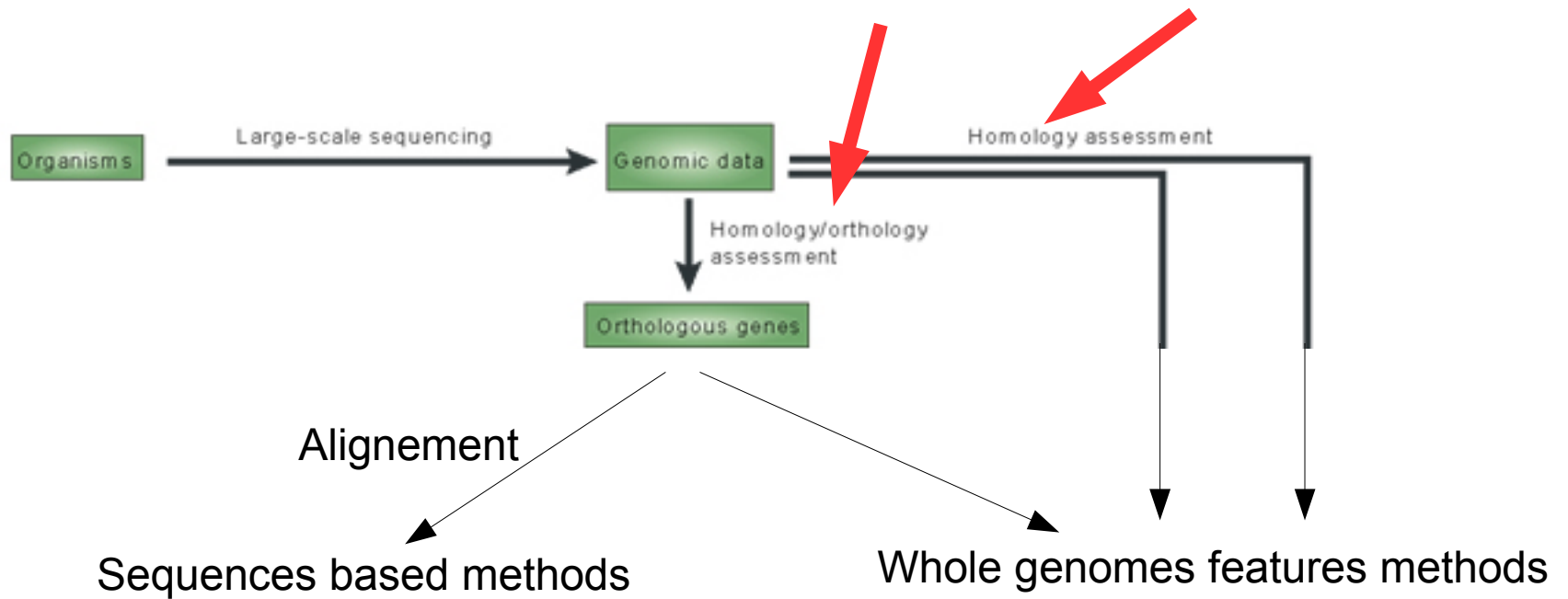
- Stochastic errors are sampling errors caused by a too small sample. To measure it, it's possible to use resampling method bootstrap or jackknife.
- Systematic errors appears when the evolutionary process violates the assumptions of model used for phylogenetic reconstruction.

Phylogenomic analysis : the type of methods

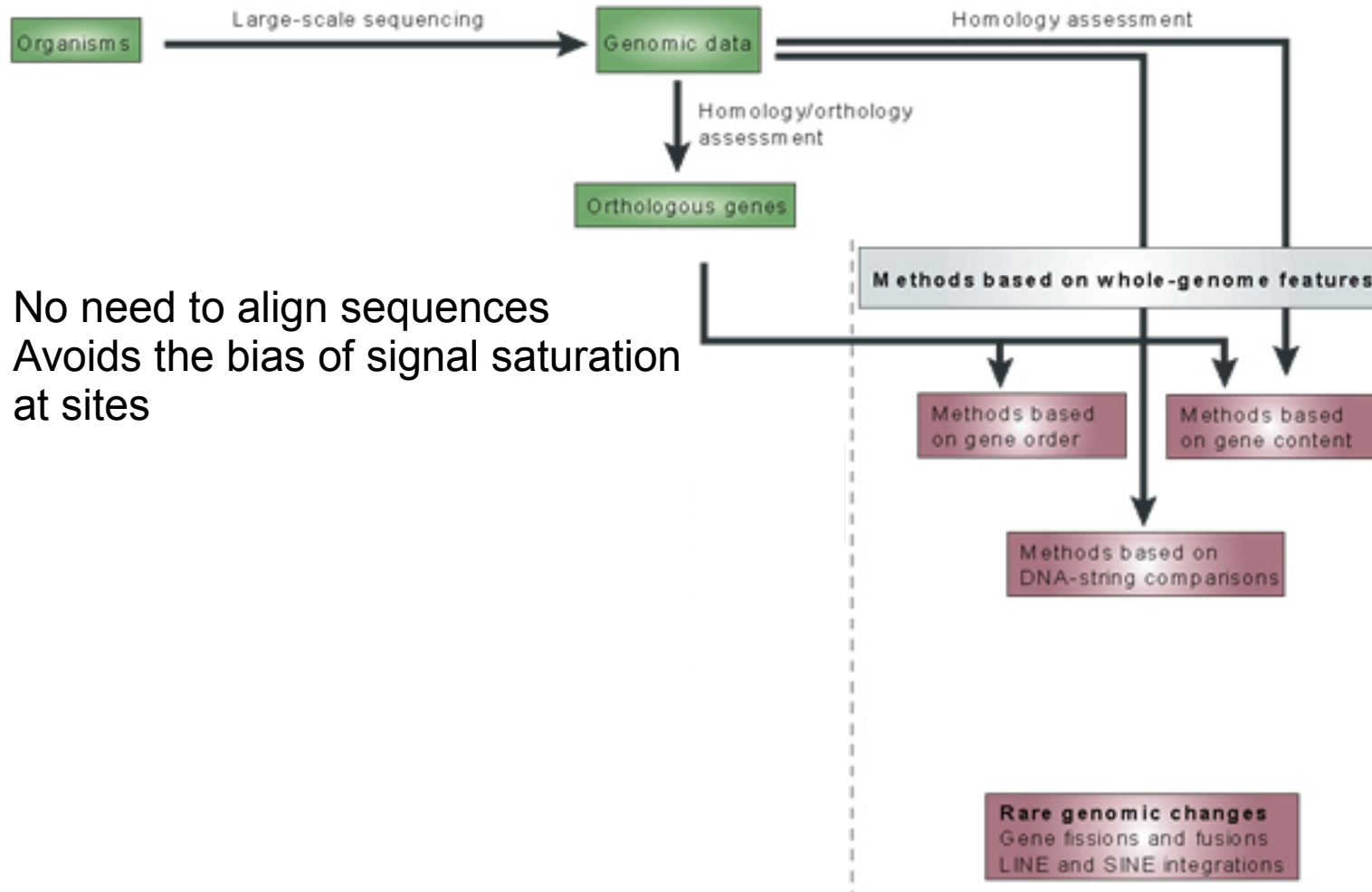


Phylogenomic analysis : the type of methods

Journée de vendredi 20 et lundi 23

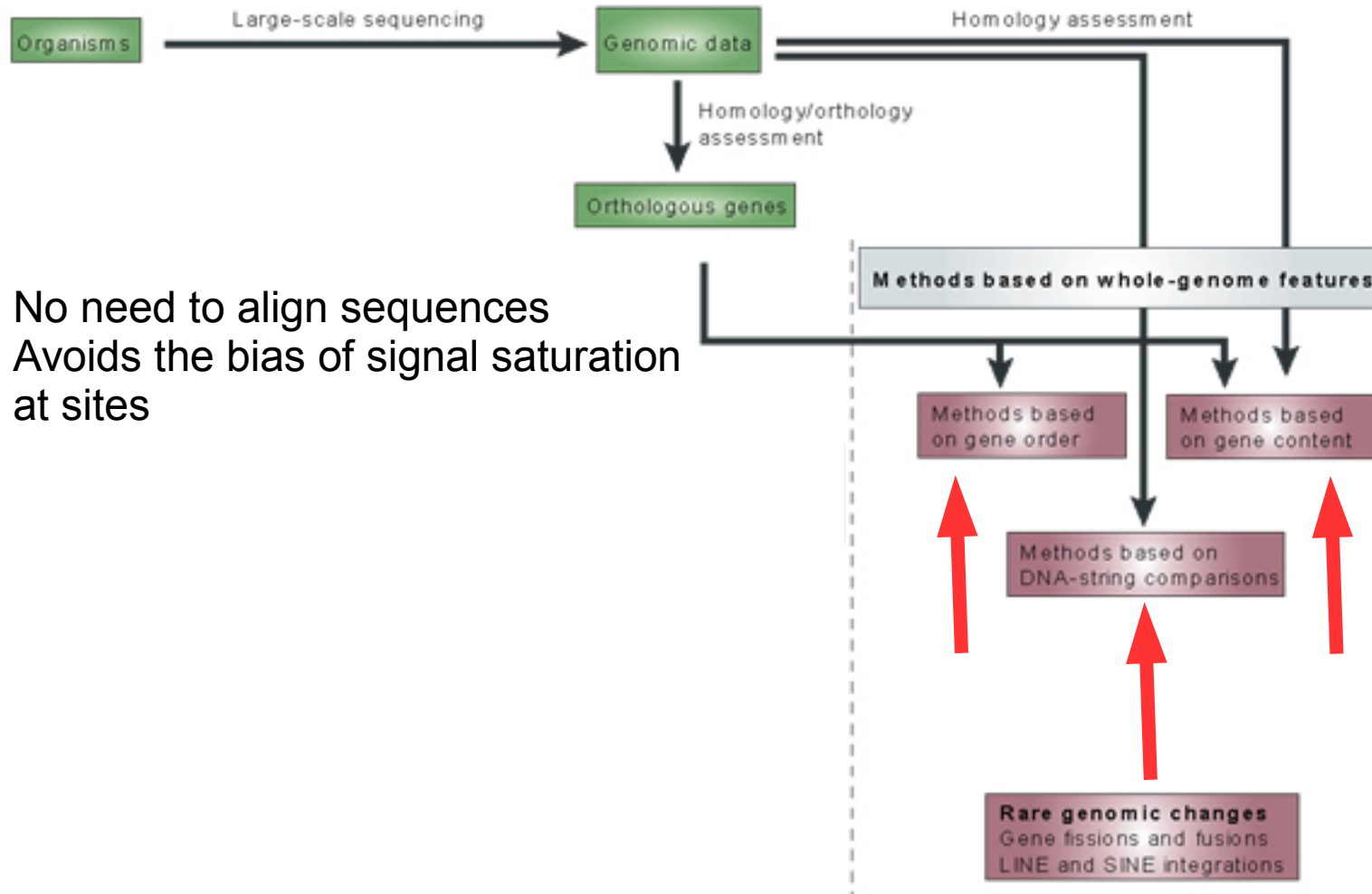


Phylogenomic analysis : the methods



No need to align sequences
Avoids the bias of signal saturation
at sites

Phylogenomic analysis : the methods



No need to align sequences
Avoids the bias of signal saturation
at sites

Whole genome features methods

- Gene content
- Gene order approach
- DNA-string approach

Comparison of gene content

- Find the potential orthologous genes
- Write the presence/absence matrix

	Species 1	Species 2	Species 3	...
Gene 1	0	1	1	
Gene 2	0	0	0	
Gene 3	1	1	0	
...				

- And build the tree with maximum parsimony
- Or compute the distance matrix (normalized by the number of genes in each genome involved)
 - And build the tree with NJ

23/10/17 • Disadvantages: big/small genome attraction

Comparison of gene content

Table 1• Common gene content in genomes

	AF	MT	MJ	PH	AQ	SY	BS	MG	BB	EC	HI	HP	SC
AF	2,407	48.1	50.1	40.2	38.2	26.3	26.8	33.3	25.2	28.1	26.4	23.6	23.1
MT	900	1,871	55.7	37.4	35.3	31.1	30.9	30.3	24.8	32.0	24.2	22.3	27.9
MJ	870	966	1,735	43.7	32.7	29.2	28.1	31.2	22.2	31.1	22.4	22.3	27.8
PH	829	699	759	2,061	30.9	23.8	27.2	31.4	24.0	26.1	21.7	20.1	23.7
AQ	582	537	497	471	1,522	52.5	53.8	54.5	44.6	59.0	44.0	43.7	31.1
SY	632	581	506	491	799	3,168	30.5	58.8	48.1	35.9	44.6	41.0	19.1
BS	645	578	488	561	819	967	4,100	70.7	56.5	33.6	51.3	42.0	16.1
MG	156	142	146	147	255	275	331	468	50.4	62.2	57.5	52.1	40.4
BB	214	211	189	204	379	409	480	236	850	52.2	46.2	43.8	29.4
EC	676	598	539	538	898	1,138	1,376	291	444	4,290	77.8	49.9	17.1
HI	453	416	384	372	669	766	880	269	393	1335	1,717	41.1	28.8
HP	375	355	354	320	665	652	668	244	372	793	653	1,590	22.2
SC	555	522	482	488	474	606	659	189	250	735	494	353	6,296

The numbers of genes shared (see Methods) between genomes (lower left triangle), the percentage of genes shared between genomes (the total number divided by the number of genes in the smallest genome; upper right triangle) and the numbers of genes per genome (bold). HI, *H. influenzae*¹⁶; MG, *M. genitalium*¹⁷; SY, *Synechocystis* sp. PCC 6803 (ref. 18); MJ, *M. jannaschii*¹⁹; EC, *E. coli*²⁰; MT, *M. thermoautotrophicum*²¹; HP, *H. pylori*²²; AF, *A. fulgidus*²³; BS, *B. subtilis*²⁴; BB, *B. burgdorferi*²⁵; SC, *S. cerevisiae*²⁶; AQ, *A. aeolicus*²⁷; PH, *P. horikoshii*²⁸.

(Snel B. *et al.*, Nature genetics, 1999)

Comparison of gene content

- Used for large evolutive scale, no problem with:
 - => LGT
 - => Duplication
 - => Sites saturation
- Other distances have been proposed:
 - SHOT distance (Korbel et al., 2002)
 - Huson and Steel's model (Huson and Steel, 2004)
 - Gu and Zhang's method (Gu and Zhang, 2004)

Whole genome features methods

- Gene content
- **Gene order approach**
- DNA-string approach

Comparison of gene order

- Find the genes families (homologies).
- Compute distance matrix based on breakpoint between genomes (inversions, transpositions, deletion, duplications).
- Software example : GRAPPA, DCM-GRAPPA (Tang & Moret, 2003)

Comparison of gene order

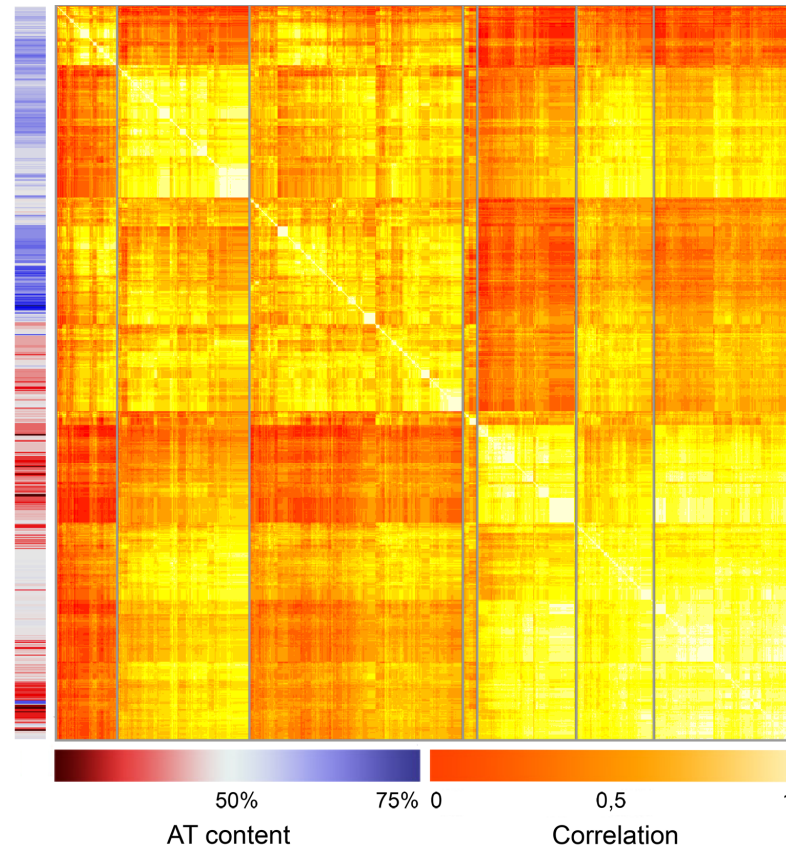
- Used for mitochondries and chloroplasts genomes
- Low error rate
- Rare events in eucaryotes genomes (large evolutionary scale)
- Problems :
 - Very limited data (mostly organelles)
 - Mathematics complex
 - Evolutionary models not well known

Whole genome features methods

- Gene content
- Gene order approach
- **DNA-string approach**

DNA string approach

- No need to orthology / homology
- Frequency matrix of words in sequences.
- Compute distance matrix (difference in the use of words).

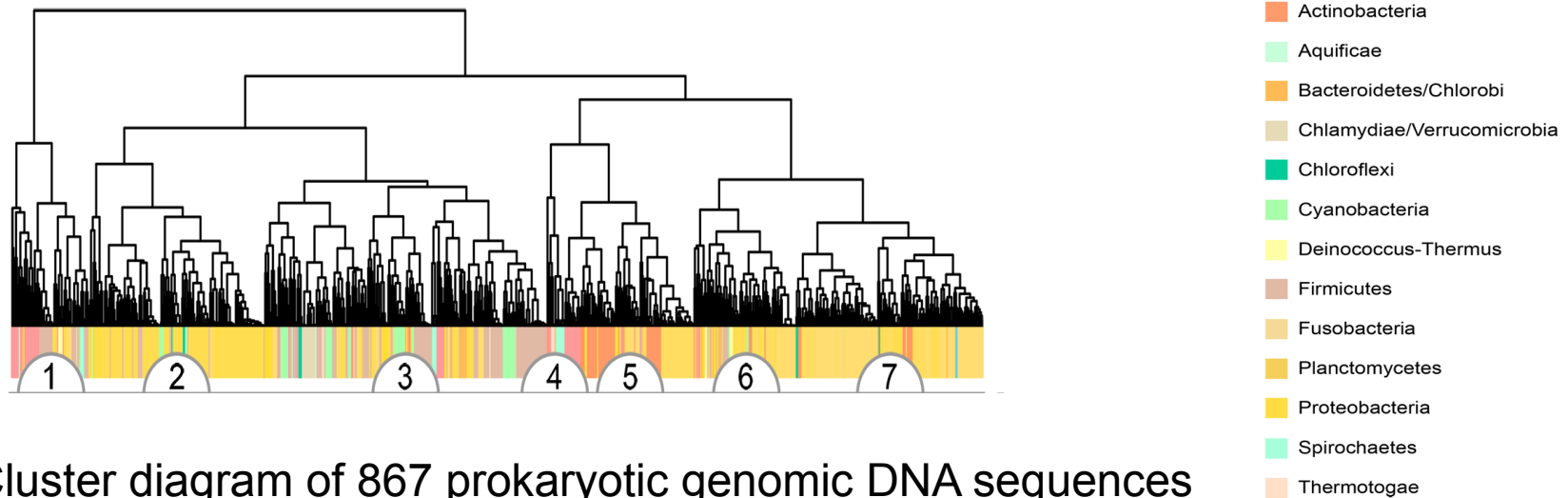


867 prokaryotic genomic DNA sequences compared pair-wise using hexanucleotide-based genomic signatures.

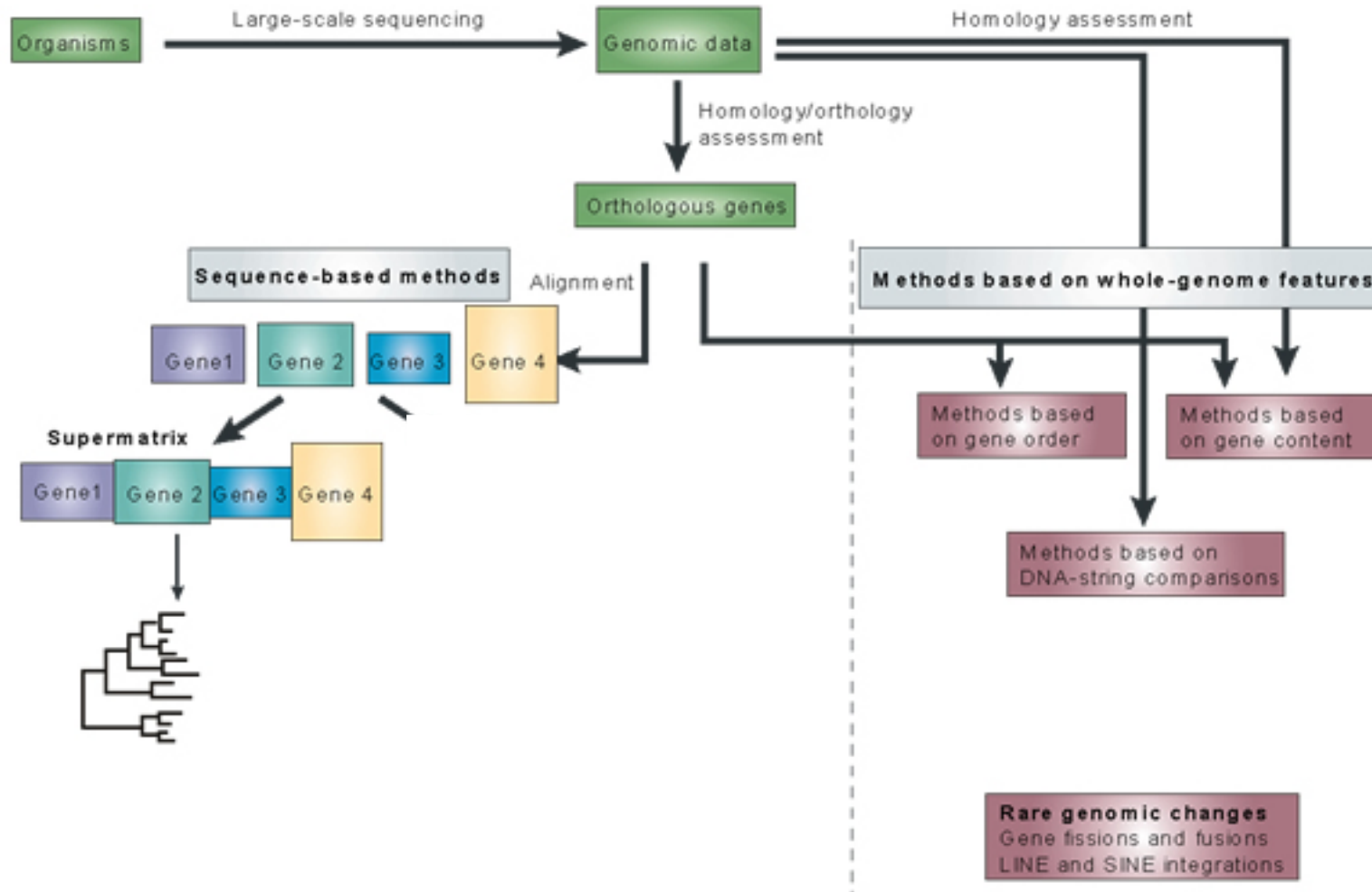
(Bohlin J. *et al.*, BMC genomics, 2009)

DNA string approach

- Build trees with clustering or NJ.
- Using of species known to have benchmarks to locate the analyzed species



Phylogenomic analysis : the methods



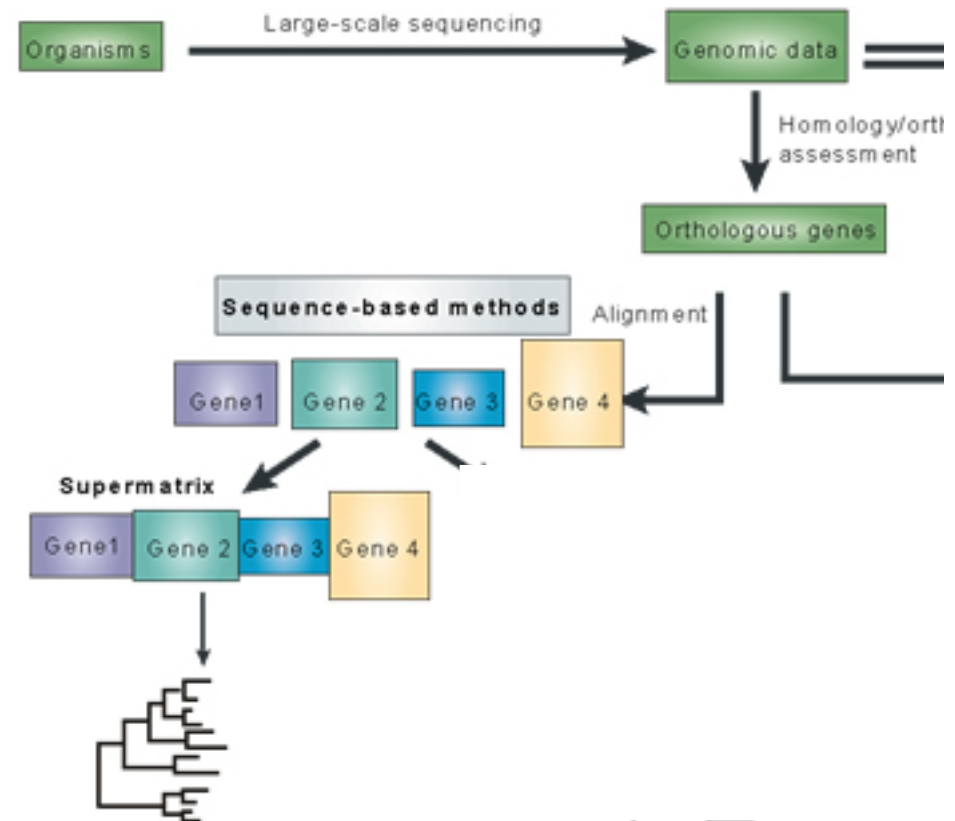
Nature Reviews | Genetics

Sequence-based methods

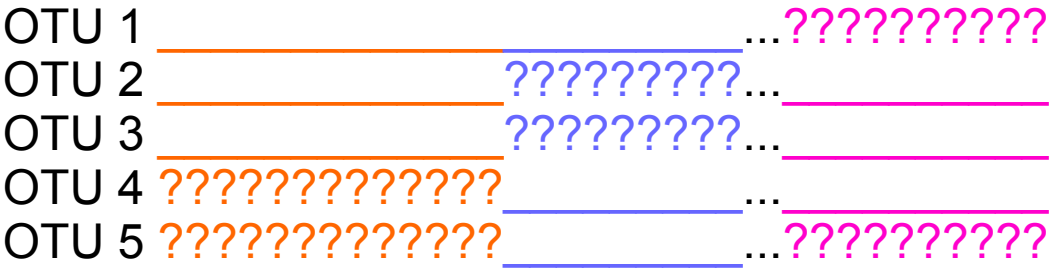
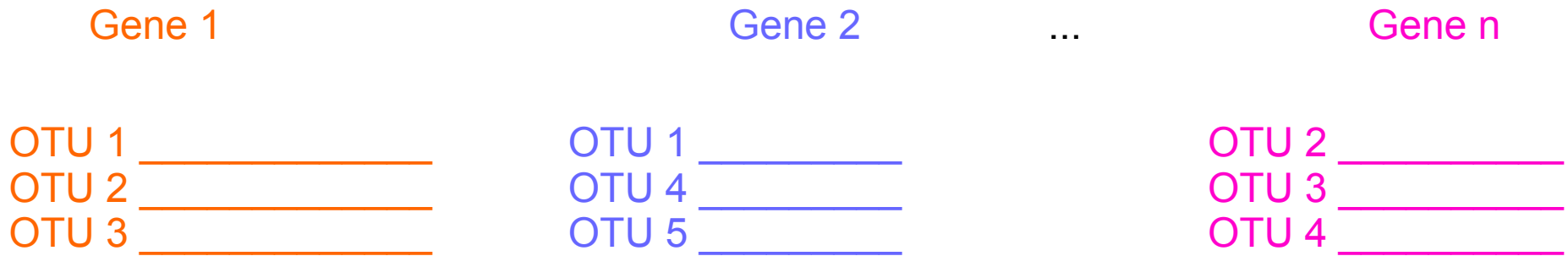
- **Supermatrix approach**
- Consensus
- Supertree approach

The supermatrix approach

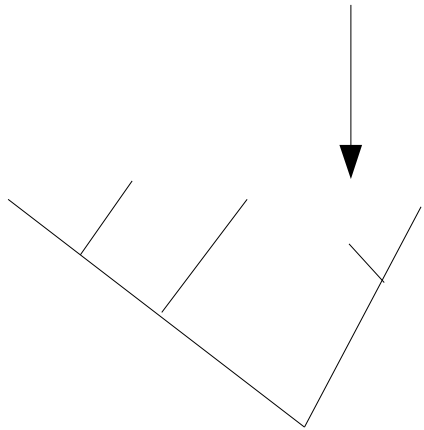
- The basic assumption is that the desired phylogenetic signal is dominant.
- Super alignment: concatenation of individual genes alignment
- Using « standard » methods of phylogeny (ML and bayesian if it's possible).



The supermatrix approach (2)



1 model fixed
 1 set of parameters inferred
 ML or bayesian methods



The supermatrix approach (3)

- May mix phylogenetic signal from different evolutionary histories
- Will require an evolutionary model with a lot of parameters (+ heterogeneity of sub. rate: gamma law + pInv) or a mixture model (explained later) (Lartillot & Philippe, *Mol Biol Evol*, 2004)
- Missing data are represented with ????? => The impact of missing data is relatively low if the alignment is sufficiently large (Roure *et al*, *Mol Biol Evol*, 2013)
- Works relatively fine when the sampling (genes and species) is good.

The supermatrix approach (4)

- Advantages/disadvantages :
 - (+) Minimize stochastic errors
 - (-) Long computation time and high memory usage for very large datasets
 - (-) It only sets a model and parameters for this model for all the superalignment
 - (-) Even the most complex model of sequence evolution cannot yet account for the complexity in superalignments (increases the systematic bias)
 - (-) Sensible to the relative sizes of datasets. For instance, if two data sets conflict, the supermatrix is dominated by the signal of the biggest one

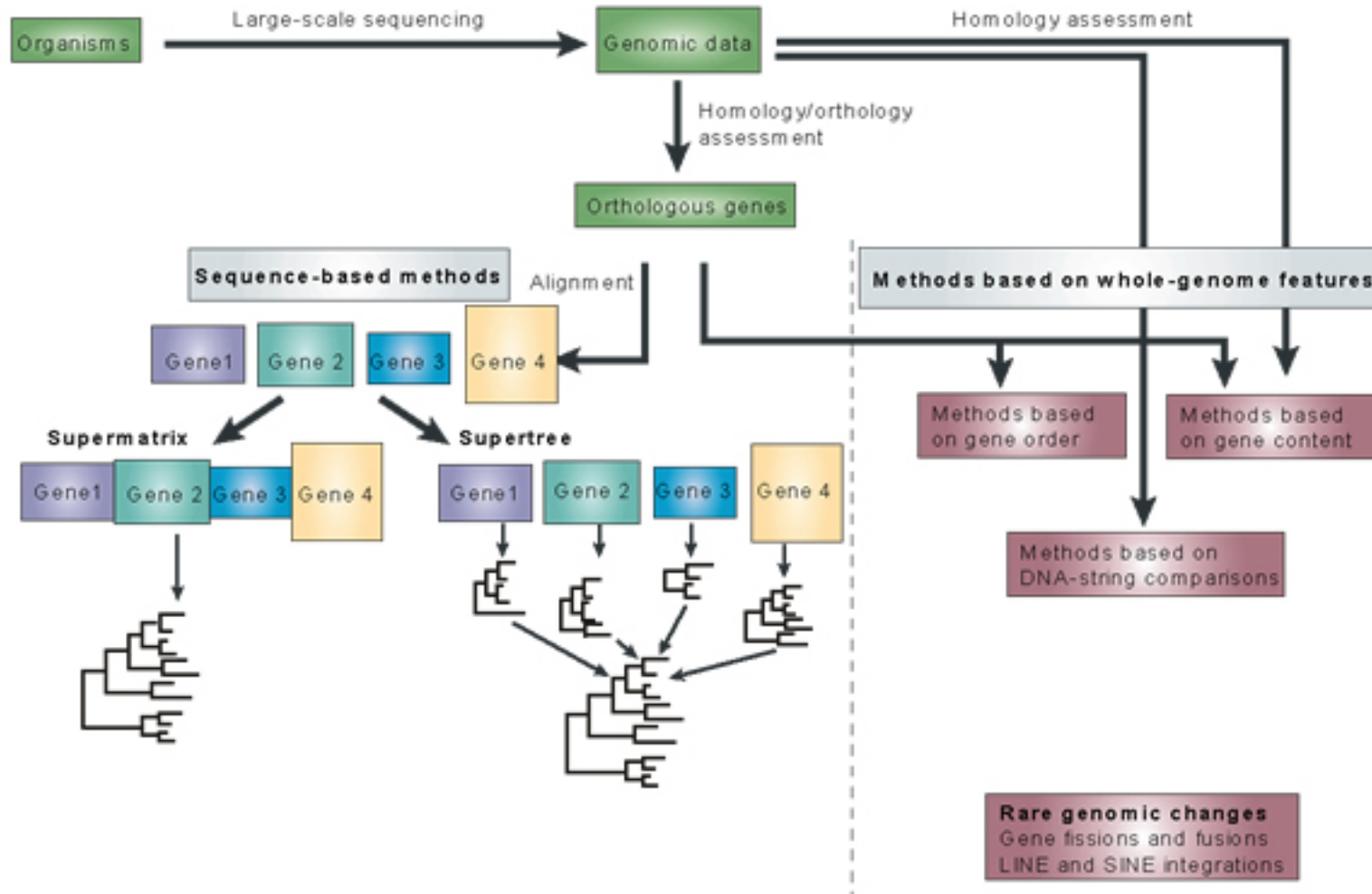
CAT Model (Lartillot & Philippe, 2004)

- Bayesian mixture model allowing that amino-acid replacement pattern at different site of a protein alignment to be described by distinct substitution processes.
- Distinct classes (categories) differing by their equilibrium frequencies over the 20 residues.
- The number of classes, their respective amino-acids profiles and the affiliation of each site to a given class are variables in the models.
- CAT model is designed to better capture the heterogeneity in the substitution pattern

Partitioned / mixed models

- Partitioning the supermatrix, applying appropriate models and their specified parameter estimates to each data partition and subsequently incorporate this into a single tree search.
- But they introduce a huge number of parameters and this may result in over-parametrized models as unadapted as the under-parametrized “concatenate” one.
- ⇒ implemented in MrBayes 3
- Bayesian analysis is able to deal with higher dimensional models than ML.

Phylogenomic analysis : the methods



Nature Reviews | Genetics

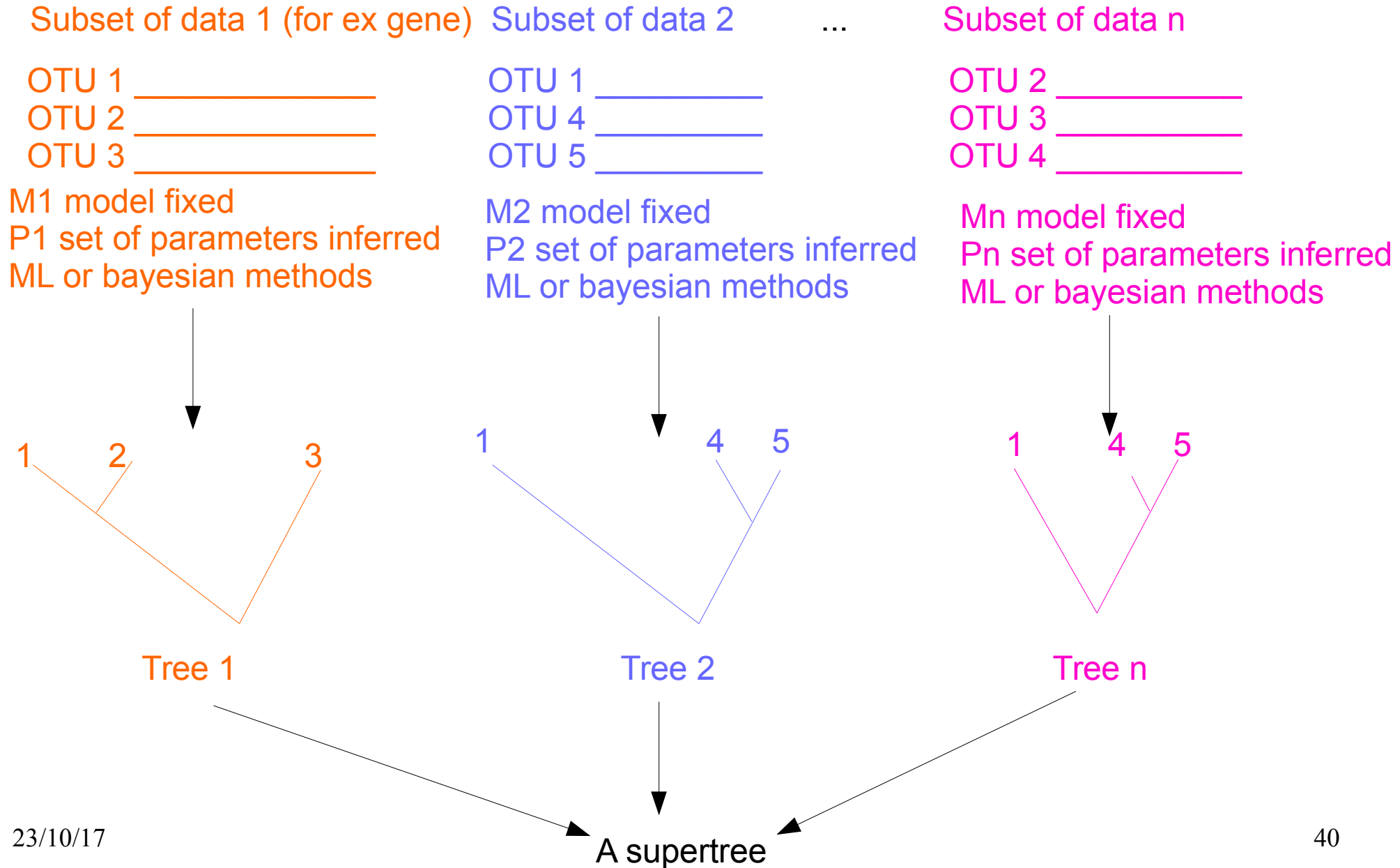
Sequence-based methods

- Supermatrix approach
- **Supertree approach**
 - **Consensus**
 - Other supertree approach

Some characteristics of supertrees

- Meta-analysis: analyses of smaller datasets are combined
- Can be used to build very large phylogenies for partially overlapping analyses
- Input trees can be based on different kinds of data (e.g. morphology, DNA-DNA hybridization) and they can be obtained by different methodologies

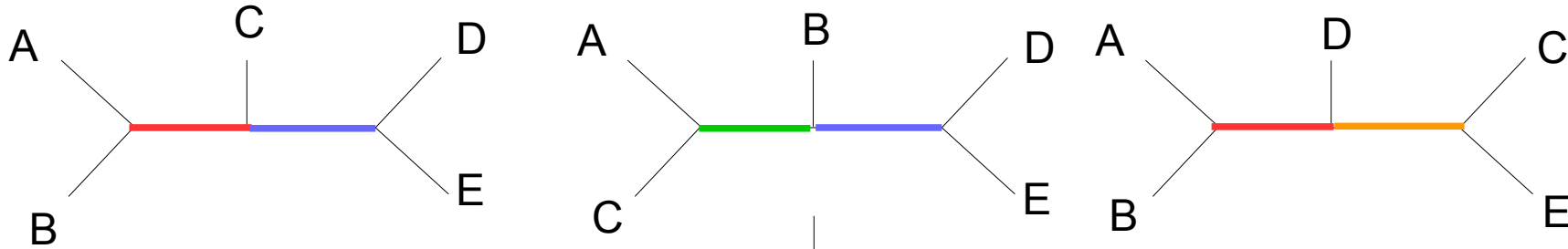
The supertree approach



Consensus Tree

- Used to test the tree robustness and for the bootstrap
- For example :
 - Strict consensus tree: a bipartition will be included if it's present in all input trees (cannot handle incompatible source trees)
 - Majority consensus tree: a bipartition will be included if it's present in more than half of the input trees (conflict resolved by vote method)

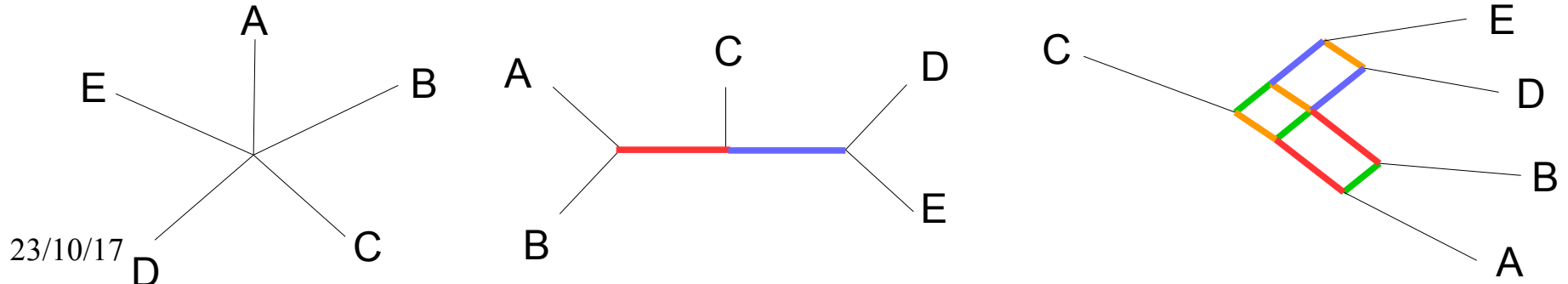
Consensus Tree (2)



Weighted bipartitions

A, B C, D, E	2
A, B, C D, E	2
A, C B, D, E	1
A, B, D C, E	1

Strict consensus (100%) Majority consensus (50%) Consensus networks ($\geq 33\%$)



(Holland & Moulton, Algorithms in bioinformatics, 2003)

Network Tree

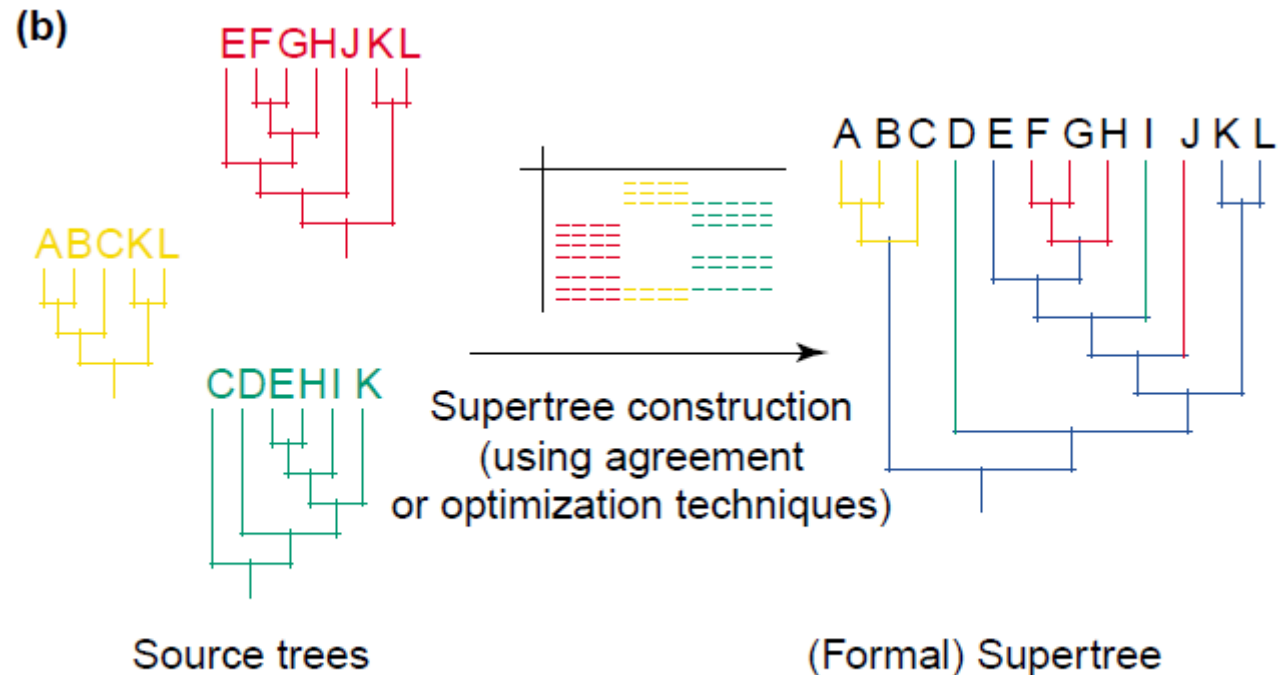
- Consensus network is one method to build network tree.
- Splitstree, for example, is a program for computing unrooted phylogenetic networks from molecular sequence data <http://www.splitstree.org/>, (Huson & Bryant, 2006).
- Phylogenetic networks should be looked when hybridization, horizontal gene transfer, recombination or gene duplication and losses are involved.

Sequence-based methods

- Supermatrix approach
- **Supertree approach**
 - Consensus
 - **Other supertree approaches**

Supertree methods

- Identical taxons sets are not needed (# consensus).
- Start with a set of trees constructed independently and not with an alignment (# super matrix method)

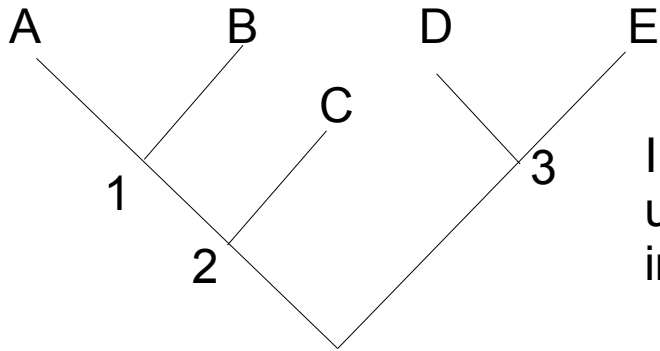


TRENDS in Ecology & Evolution

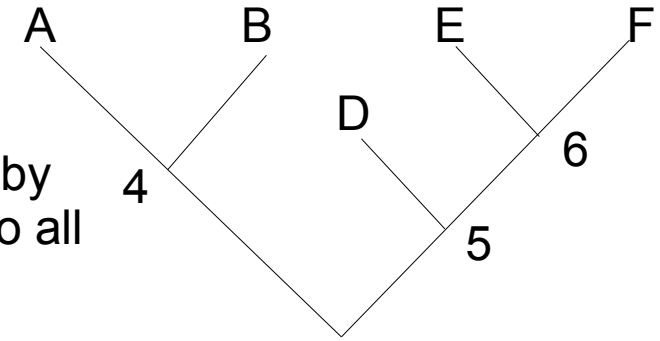
Matrix representation using parsimony (MRP)

- This is the most common method
- It's a vote method :
 - The hope is that each taxon is erroneously placed in only few source
 - Trees are highly resolved and accurate, but can lead to propose supertrees containing clades that contradict all source trees
- MRP needs a matrix representation

Build a super-tree MRP



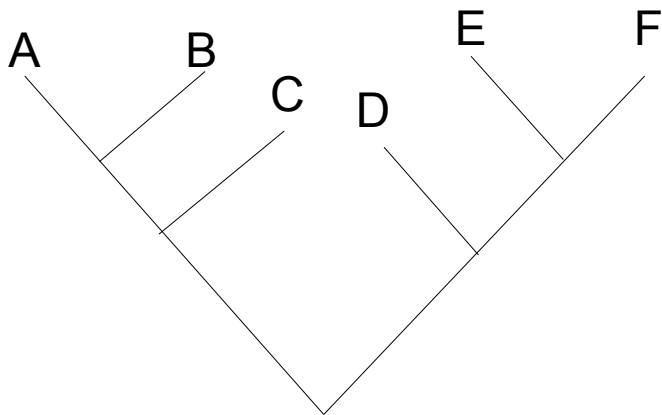
I) Input trees are rooted by using a taxon common to all input trees



II) Binary matrix representation (Baum and Ragan, 1992)

	1	2	3	4	5	6
A	1	1	0	1	0	0
B	1	1	0	1	0	0
C	0	1	0	?	?	?
D	0	0	1	0	1	0
E	0	0	1	0	0	1
F	?	?	?	0	1	1

III) Super-trees MRP (Parsimony)



1: species share a common node
 0: species do not share a common node
 ?: species not present in tree

PhySIC & PhySIC_IST

- It's a veto method :
 - the phylogenetic information of every source topology is to be respected,
 - and the supertree is not allowed to contain clades that a source tree would vote against
 - these methods remove conflicts either proposing multifurcations in the supertree or pruning rogue taxa

PhySIC & PhySIC_IST

- **Induction property:** every piece of phylogenetic information displayed in the supertree is present in one or several source topologies, or induced by their interaction (PI)
- **Non-contradiction property:** supertrees must not containing clusters that conflict either directly with a source tree or indirectly with a combination of them (PC)
- A consensus tree problem satisfies the Pareto property on clusters if every cluster that is present in every input tree appears in the consensus tree.

PhySIC & PhySIC_IST

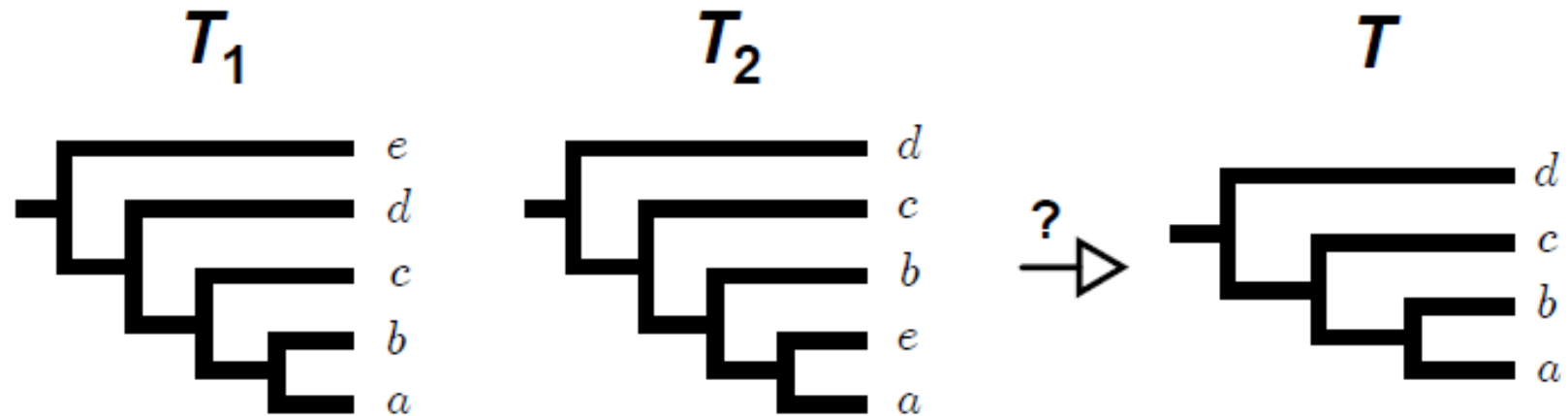


Figure 4.1: An example of informative non plenary supertree for a forest of two rooted trees - Excluding rogue taxa from the analysis can lead to more informative supertrees.

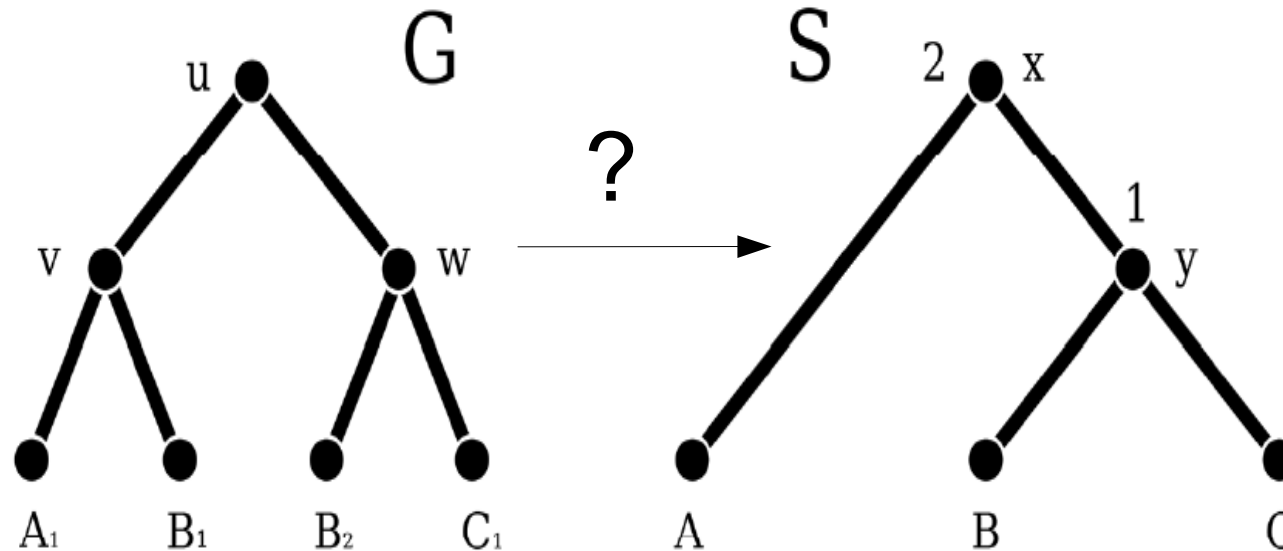
Phylogenetic Signal with Induction and non-Contradiction

- The aim of PhySIC is to infer supertrees that satisfy PI and PC and that resolve as many triplets as possible. It consists in two steps:
 - given a forest of rooted trees F , a supertree T_{PC} satisfying PC for F is computed by the PhySIC_{PC} algorithm.
 - some branches of T_{PC} are eventually collapsed by the PhySIC_{PI} algorithm until the so-modified T_{PC} satisfies also property PI.

Super Tree methods: advantage / disadvantage

- (-) The length of branches are not directly interpretable in terms of evolutionary distance
- (-) Most methods weigh poorly-supported and well-supported input trees equally
- (-) Input trees must be rooted properly
- (-) If input trees are clashing in their topologies \Rightarrow supertree resolution is too low
- (+) It's faster for very large dataset than super matrix approach
- (+) Phylogeny of each gene is made with the appropriate model and parameters and/or methods

Reconciliation



Gene duplications, gene losses, and/or lateral gene transfers are taken explicitly into account to explain the observed incongruency between a gene tree and a corresponding species tree.

Reconciliation methods

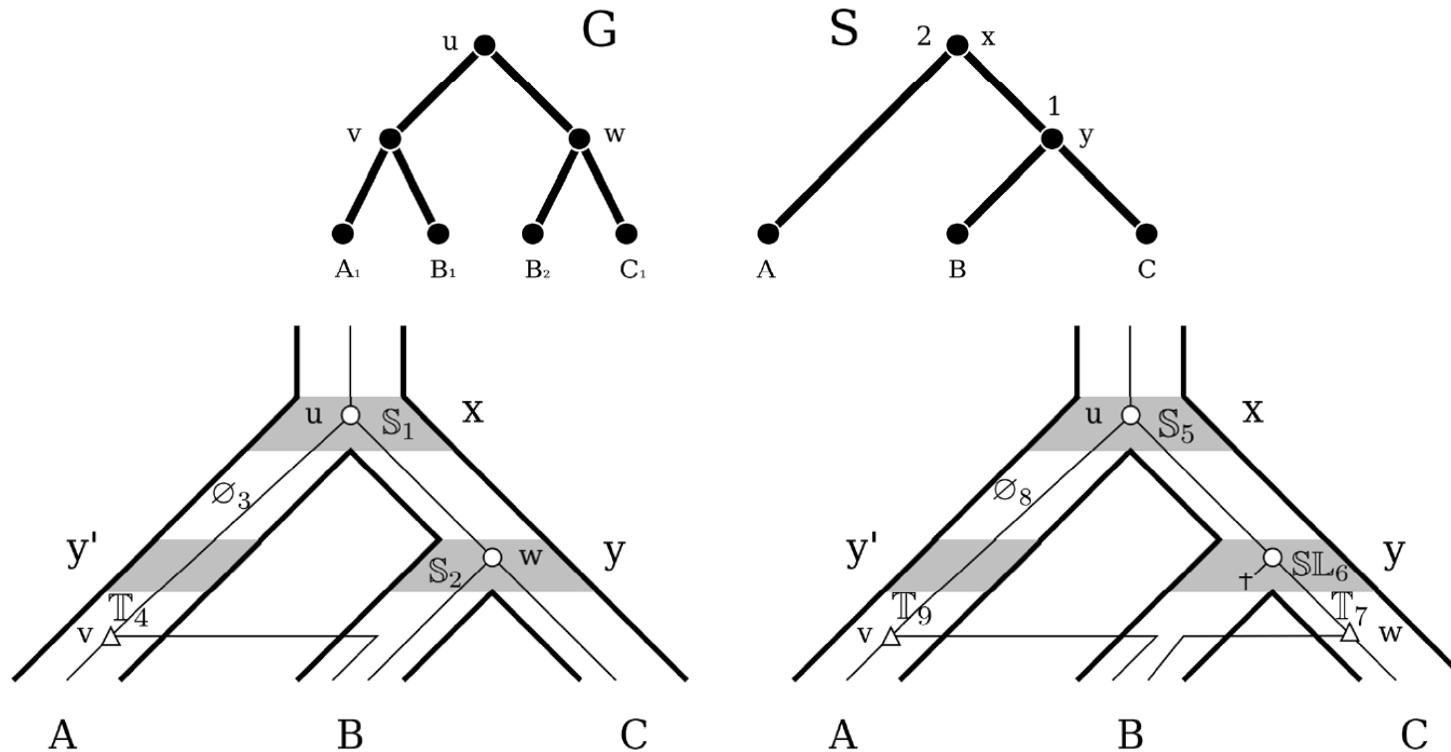


Figure 2. An example of reconciliation. Two valid reconciliations for the trees depicted in Figure 1 (C events are not indicated). The reconciliation on the left contains two S events, one O event, a T event and four C events, while the one on the right contains one S event, one SL, one O event, two T events and four C events.
doi:10.1371/journal.pone.0073667.g002

S speciation, D duplication, T transfer, TL a transfer followed by loss of the non-transferred child, SL a speciation followed by loss of one of the two resulting children, O no event indicating that a gene lineage has crossed a time boundary, and C contemporary event associating an extant gene copy with its corresponding species.

Reconciliation methods

- Parsimony or probabilistic criteria have been proposed.
- Most reconciliation tools need a dated species tree.
- For a review, see : Doyon *et al*, briefings in Bioinformatics, 2011 => ATGC : Montpellier bioinformatic platform
- Softwares: Notung (Durand *et al.*, J. Comput. Biol, 2006) (DL model), Mowgli (Doyon *et al.*, RECOMB-CG, 2010) (DTL model)

Super Tree methods: the future

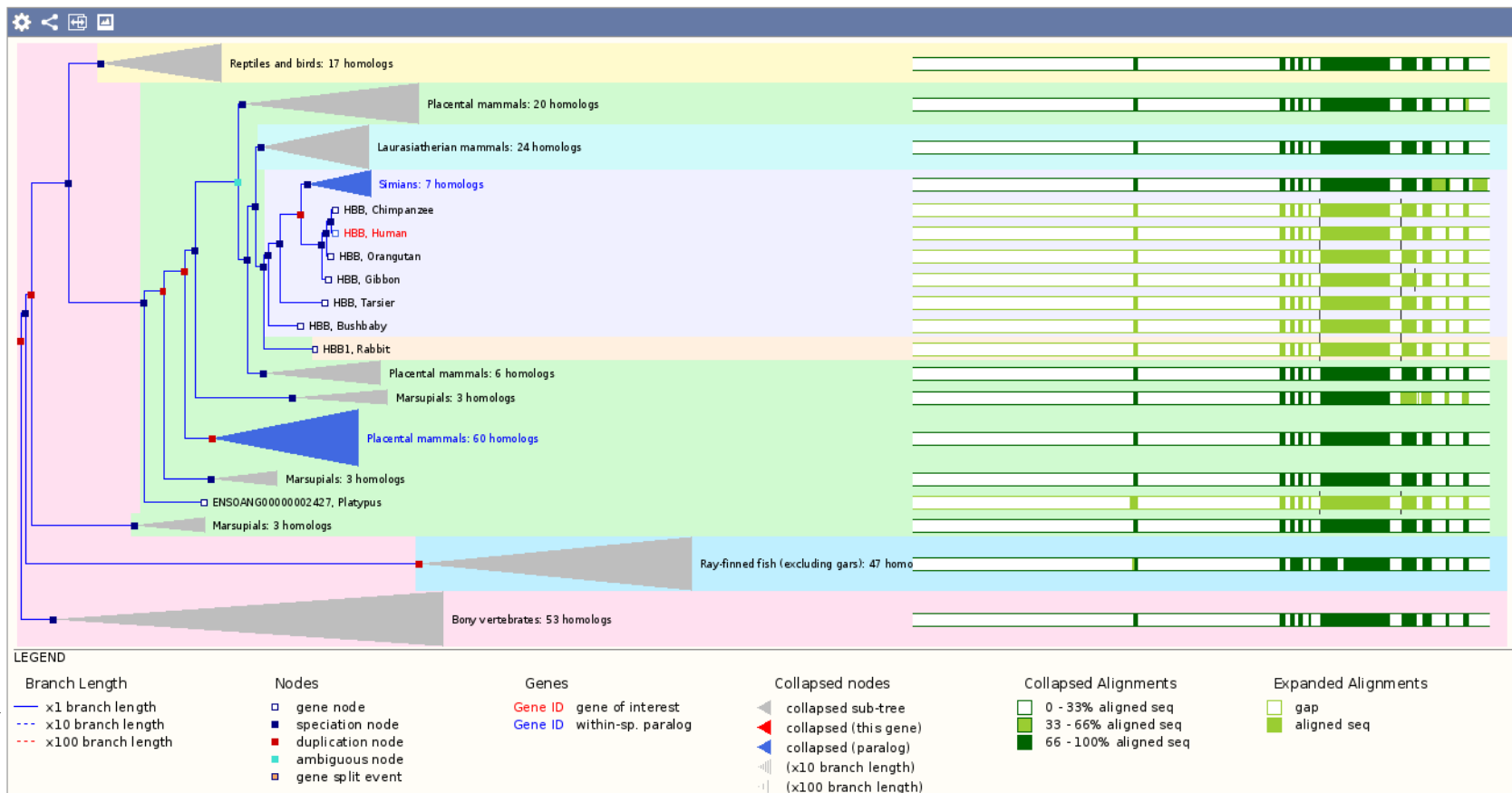
- Few methods allow to create a super tree from individual multigene families considering the events of duplication, horizontal transfer ...
 - Finding the species tree that minimizes the reconciliation cost
 - SPR (Subtree Prune-and-Regraft) distance (Whidden *et al*, Syst. Biol., 2014) => LGT
 - iGTP (Gene Tree Parsimony) (Chaudhary *et al.*, BMC Bioinformatics, 2010) => gene Duplication and Loss, or Incomplete lineage sorting.
 - Using Hierarchical Bayesian model: very computationally extensive (Martins *et al.*, Syst. Biol, 2014) « guenomu » => D,L, ILS

Compare trees with metrics

- Robinson & Foulds (symmetric difference metric): Sum of the specific bipartitions for each two trees (treedist)
- Branch score distance: using the branch length (treedist)
- In a likelihood framework (tree-puzzle, RaxML, CONSEL) :
 - The SH test (Shimodaira and Hasegawa, 1999)
 - Two-sided KH test (Kishino and Hasegawa, 1989), the one-sided KH test (Goldman et al., 2000)
 - Expected likelihood weights (Strimmer and Rambaut 2002)

Ensembl compara

- Use a reconciliation method to call duplication events.
- Allow to extract orthologs and paralog sequences.

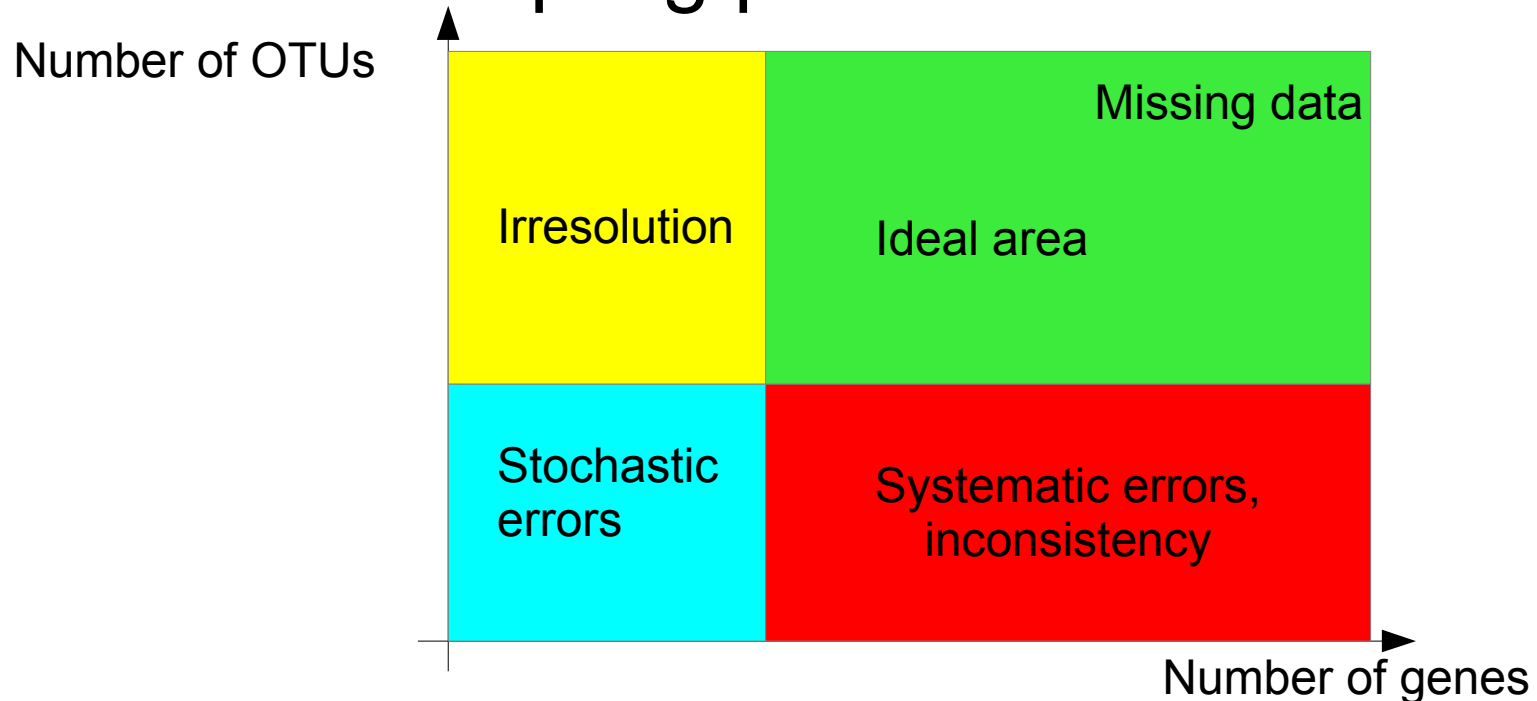


Ensembl compara

- Go to <http://www.ensembl.org>
- Select Chimpanzee genome
- Search ND1 gene and click on the appropriate result
- Click on Gene Tree (Image) and explore it to find ambiguous nodes concerning primates and duplication nodes in the tree.
- Click on Orthologues and explore the result table.
- Retrieve one fasta sequence of one 1:1 orthologs of this gene

To conclude

- The phylogenomic is still a research domain (methods and analysis)
- Test several models and methods for testing the robustness of the tree produced (computationally intensive)
- Be aware of sampling problems



Stochastic and systematic errors

- Stochastic errors are sampling errors caused by a too small sample. To measure it, it's possible to use resampling method bootstrap or jackknife.
- Systematic errors appears when the evolutionary process violates the assumptions of model used for phylogenetic reconstruction.
 - ⇒ To reduce it we need to reduce the non-phylogenetic signal : eliminate species with rapid evolution, remove positions saturate with multiple substitutions, make a recoding ...

Methods and use cases

Class Methods	Methods	Use Case
Based on whole genome features => No need to align sequences => Avoid the signal saturation at sites	Genome signature	Unknown species
	Gene Content	Large evolutionary scale Doesn't need orthology inference
	Gene Content	Large evolutionary scale in Eucaryotes Used for organelles
Based on sequences => need to align sequences	Supermatrix	Individual genes have not enough signal Phylogenetic signal is assumed majority
	Supertree	Individual genes have enough signal Heterogeneous dataset Very big dataset if you're using simple methods

References

- Scientific articles cited in the slides
- Presentation :
 - M2 – Phylogénomique. Frédéric Delsuc : Equipe de Phylogénie et Evolution Moléculaire, Institut des Sciences de l'Evolution de Montpellier
- Thèses :
 - Béatrice Roure soutenue en 2011 : « Amélioration de l'exactitude de l'inférence phylogénomique »
 - Céline Sconavacca soutenue en 2009 : « Supertree methods for phylogenomics »