Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

Presentation and implementation of phylogenomics methods

Claire Hoede, PF Bioinfo, Genotoul

Outline

- Build the dataset:
  - What scale for infering 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 ?

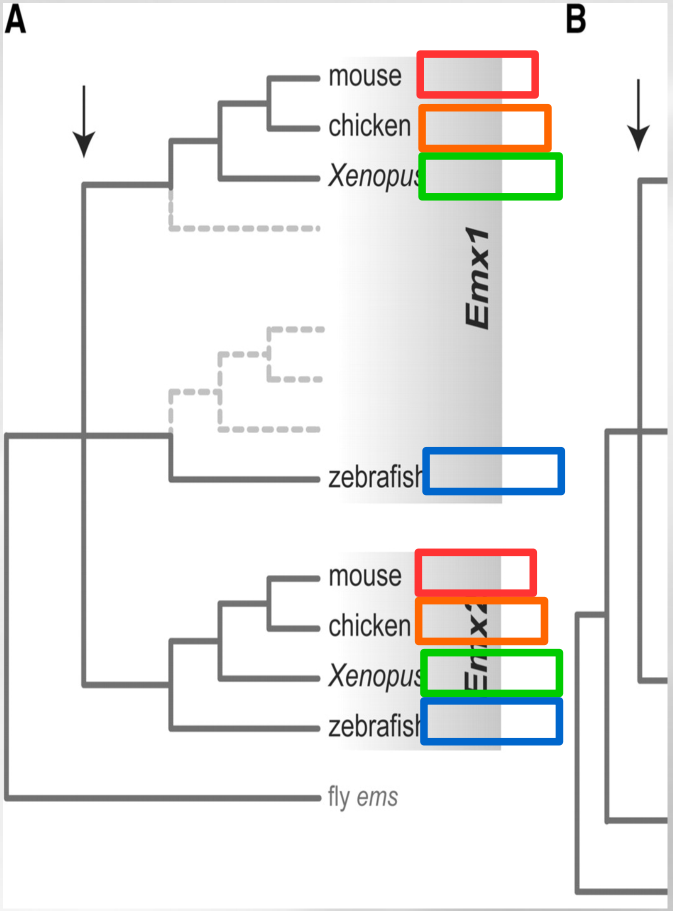
Limits of phylogenies based on a single gene

- 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)

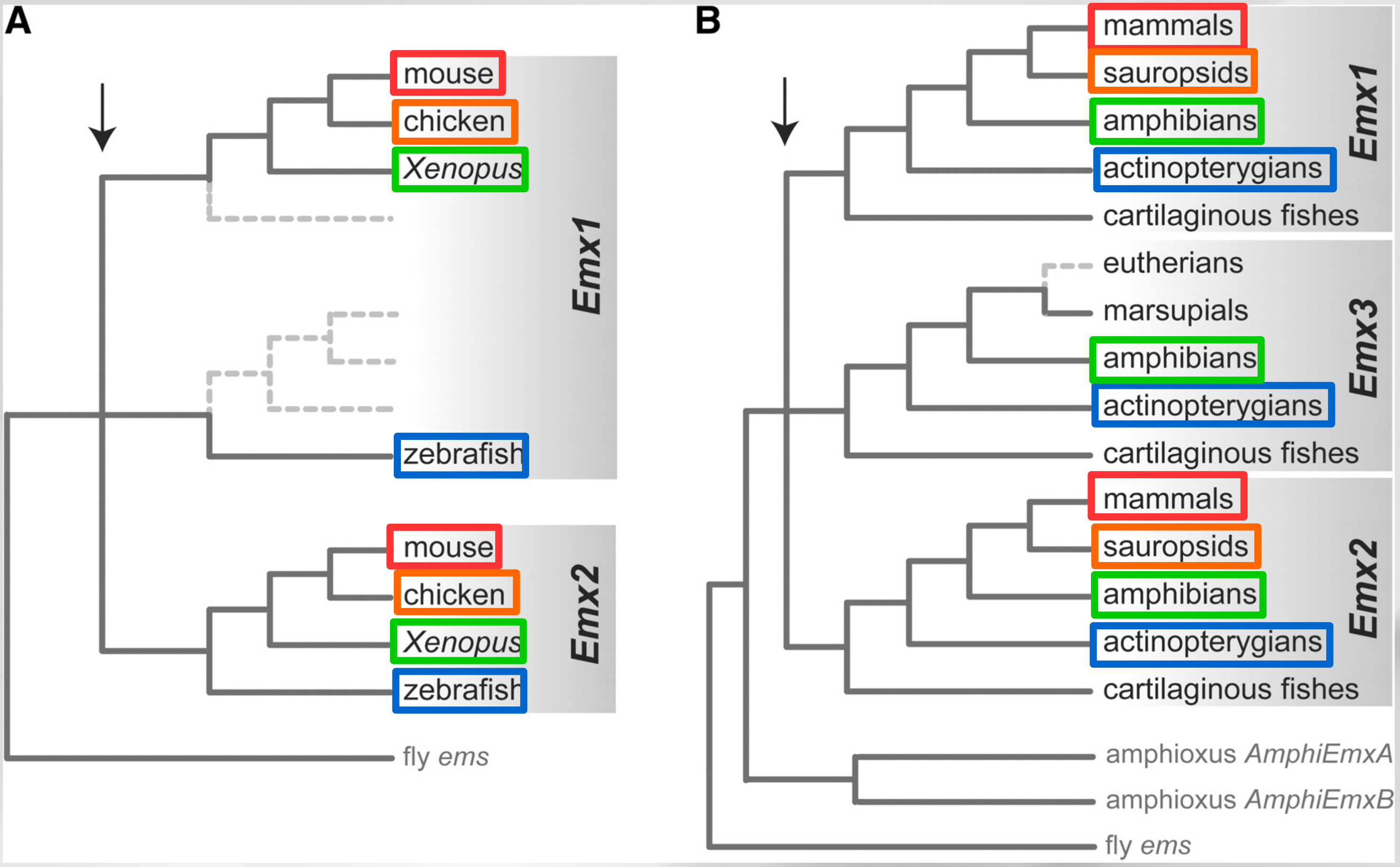
Sources of incongruence: Hidden paralogy



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.

Kuraku (2010) Integr. Comp. Biol. 50 (1): 124-129.

Sources of incongruence: Hidden paralogy

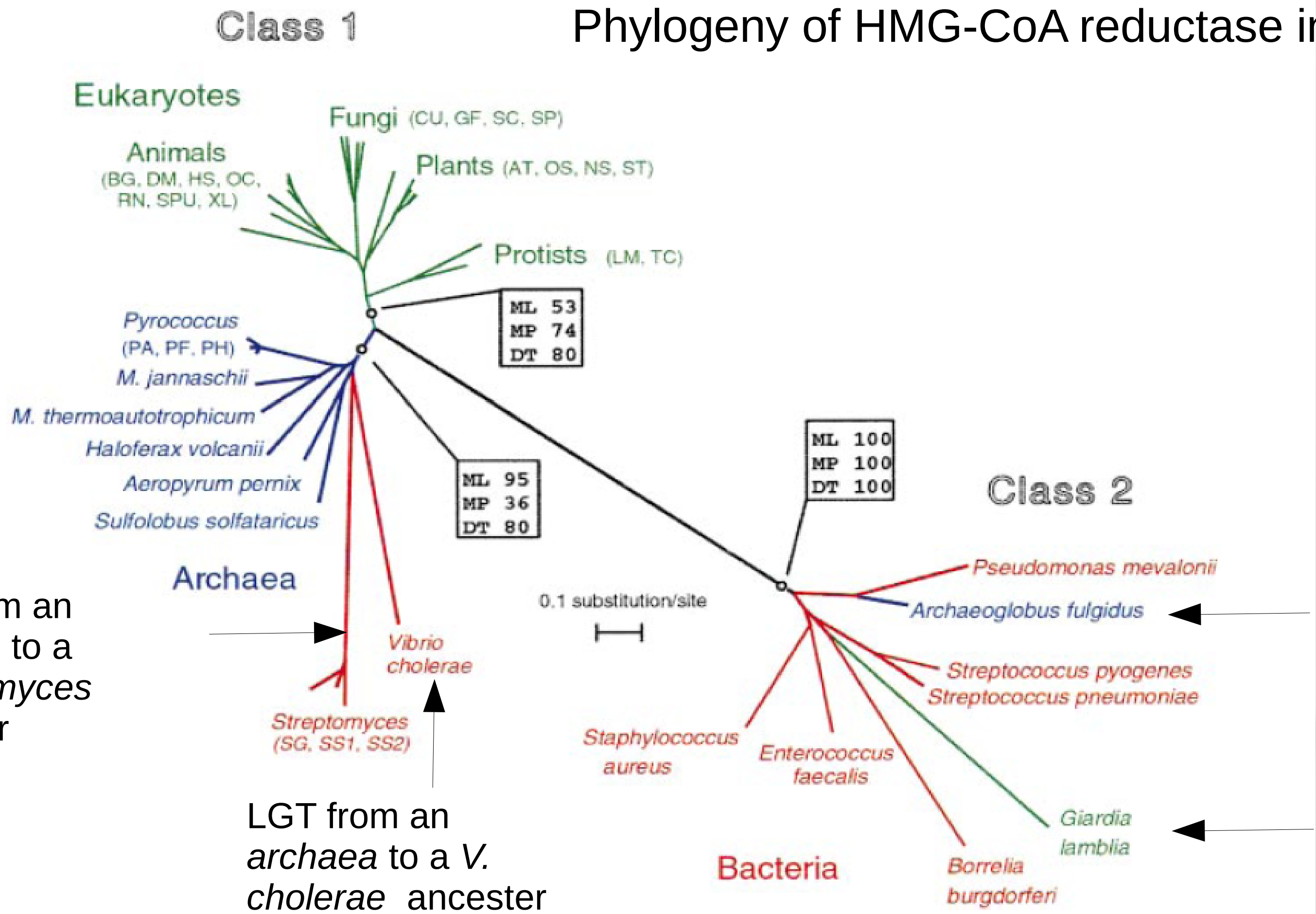


Hidden paralogy in Emx gene phylogeny. Molecular phylogenetic trees of vertebrate Emx genes before the year 2000 (A) and now (B) are shown. Dotted lines indicate absences of relevant genes (gene loss or incomplete identification). Note that the zebrafish gene, initially recognized as emx1 in (A) (Morita 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.

Kuraku (2010) Integr. Comp. Biol. 50 (1): 124-129.

Sources of incongruence: lateral gene transfer

Phylogeny of HMG-CoA reductase in several kingdom



LGT from an *archaea* to a *Streptomyces* ancestor

LGT from an *archaea* to a *V. cholerae* ancester

LGT from a *Pseudomonas* to a *A. fulgidus* ancester

Parasitic protozoan living in the mammalian intestine : acquiring a bacterial version of the gene by LGT

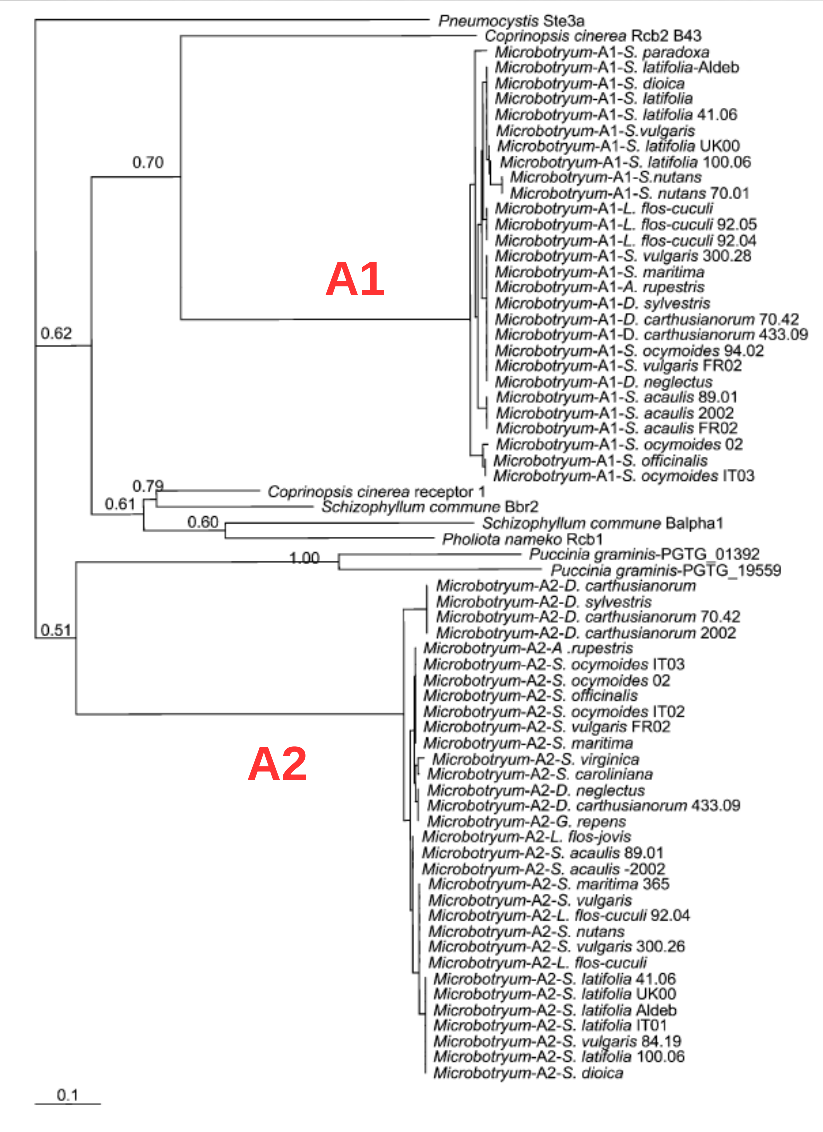
Fig. 3. Phylogeny of HMG-CoA reductase. A subset of 37 taxa from the alignment of all known HMGR protein sequences was used to carry out the analysis. The distance tree shown was determined using PROTDIST with PAM distances and branch lengths calculated with FITCH (PHYLIP 3.57; Felsenstein, 1993). The 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 RELL 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

Boucher and Doolittle (2000) Molecular Microbiology 37 (4): 703-716.

Sources of incongruence: trans-specific polymorphism

- **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 allelique classes in the same species (extrem case of balancing selection).

Sources of incongruence: trans-specific polymorphism

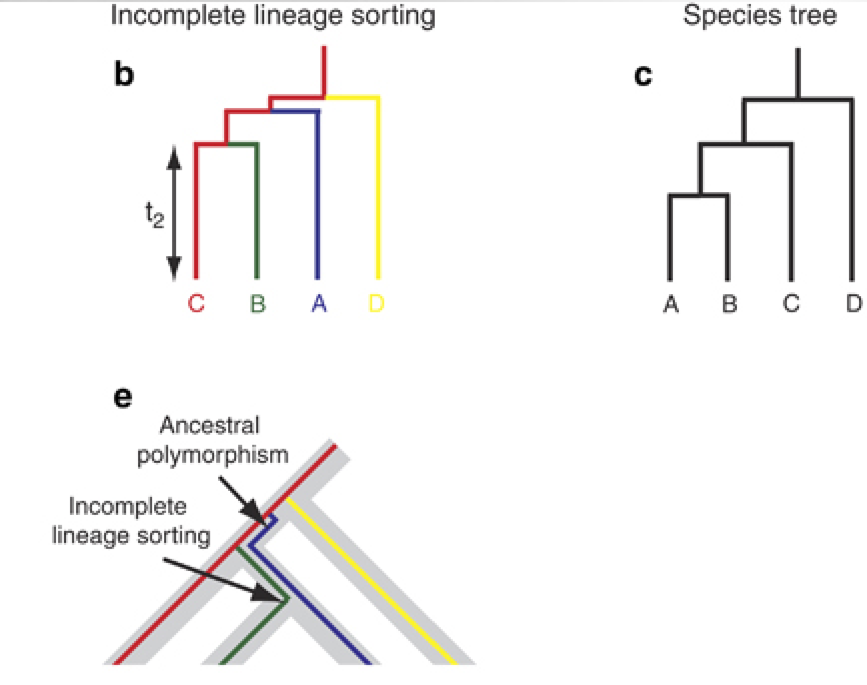

Pneumocystis Ste3a
Coprinopsis cinerea Rcb2 B43
Microbotryum-A1-S. paradoxa
Microbotryum-A1-S. latifolia-Aldeb
Microbotryum-A1-S. dioica
Microbotryum-A1-S. latifolia
Microbotryum-A1-S. latifolia 41.06
Microbotryum-A1-S.vulgaris
Microbotryum-A1-S. latifolia UK00
Microbotryum-A1-S. latifolia 100.06
Microbotryum-A1-S.nutans
Microbotryum-A1-S. nutans 70.01
Microbotryum-A1-L. flos-cuculi
Microbotryum-A1-L. flos-cuculi 92.05
Microbotryum-A1-L. flos-cuculi 92.04
Microbotryum-A1-S. vulgaris 300.28
Microbotryum-A1-S. maritima
Microbotryum-A1-A. rupestris
Microbotryum-A1-D. sylvestris
Microbotryum-A1-D. carthusianorum 70.42
Microbotryum-A1-D. carthusianorum 433.09
Microbotryum-A1-S. ocymoides 94.02
Microbotryum-A1-S. vulgaris FR02
Microbotryum-A1-D. neglectus
Microbotryum-A1-S. acaulis 89.01
Microbotryum-A1-S. acaulis 2002
Microbotryum-A1-S. acaulis FR02
Microbotryum-A1-S. ocymoides 02
Microbotryum-A1-S. officinalis
Microbotryum-A1-S. ocymoides IT03
0.70
0.62
A1
0.79
0.61
Coprinopsis cinerea receptor 1
Schizophyllum commune Bbr2
0.60
Schizophyllum commune Balpha1
Pholiota nameko Rcb1
1.00
Puccinia graminis-PGTG_01392
Puccinia graminis-PGTG_19559
0.51
Microbotryum-A2-D. carthusianorum
Microbotryum-A2-D. sylvestris
Microbotryum-A2-D. carthusianorum 70.42
Microbotryum-A2-D. carthusianorum 2002
Microbotryum-A2-A .rupestris
Microbotryum-A2-S. ocymoides IT03
Microbotryum-A2-S. ocymoides 02
Microbotryum-A2-S. officinalis
Microbotryum-A2-S. ocymoides IT02
Microbotryum-A2-S. vulgaris FR02
Microbotryum-A2-S. maritima
A2
Microbotryum-A2-S. virginica
Microbotryum-A2-S. caroliniana
Microbotryum-A2-D. neglectus
Microbotryum-A2-D. carthusianorum 433.09
Microbotryum-A2-G. repens
Microbotryum-A2-L. flos-jovis
Microbotryum-A2-S. acaulis 89.01
Microbotryum-A2-S. acaulis -2002
Microbotryum-A2-S. maritima 365
Microbotryum-A2-S. vulgaris
Microbotryum-A2-L. flos-cuculi 92.04
Microbotryum-A2-S. nutans
Microbotryum-A2-S. vulgaris 300.26
Microbotryum-A2-L. flos-cuculi
Microbotryum-A2-S. latifolia 41.06
Microbotryum-A2-S. latifolia UK00
Microbotryum-A2-S. latifolia Aldeb
Microbotryum-A2-S. latifolia IT01
Microbotryum-A2-S. vulgaris 84.19
Microbotryum-A2-S. latifolia 100.06
Microbotryum-A2-S. dioica
0.1

Phylogeny based on the pheromone receptor pr-MatA1 and pr-MatA2 of Mycrobotryum 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 allelique classes in the same species (extrem case of balancing selection).

B. Devier *et al*, Genetics 181: 209–223 (2009)

Sources of incongruence: incomplete lineage sorting

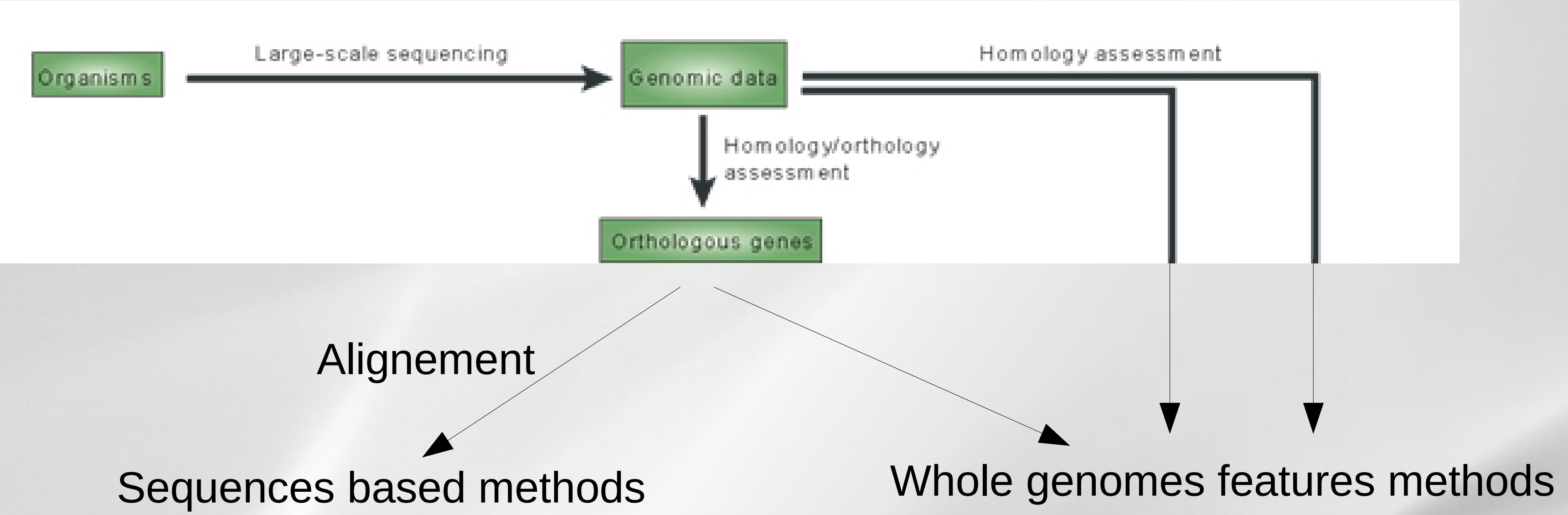


- Lineage sorting always results in coalescence with the other species prior to the speciation event (t_2).
- It can be observed when the speciations are temporally close

Twyford and Ennos, Heredity (2012)

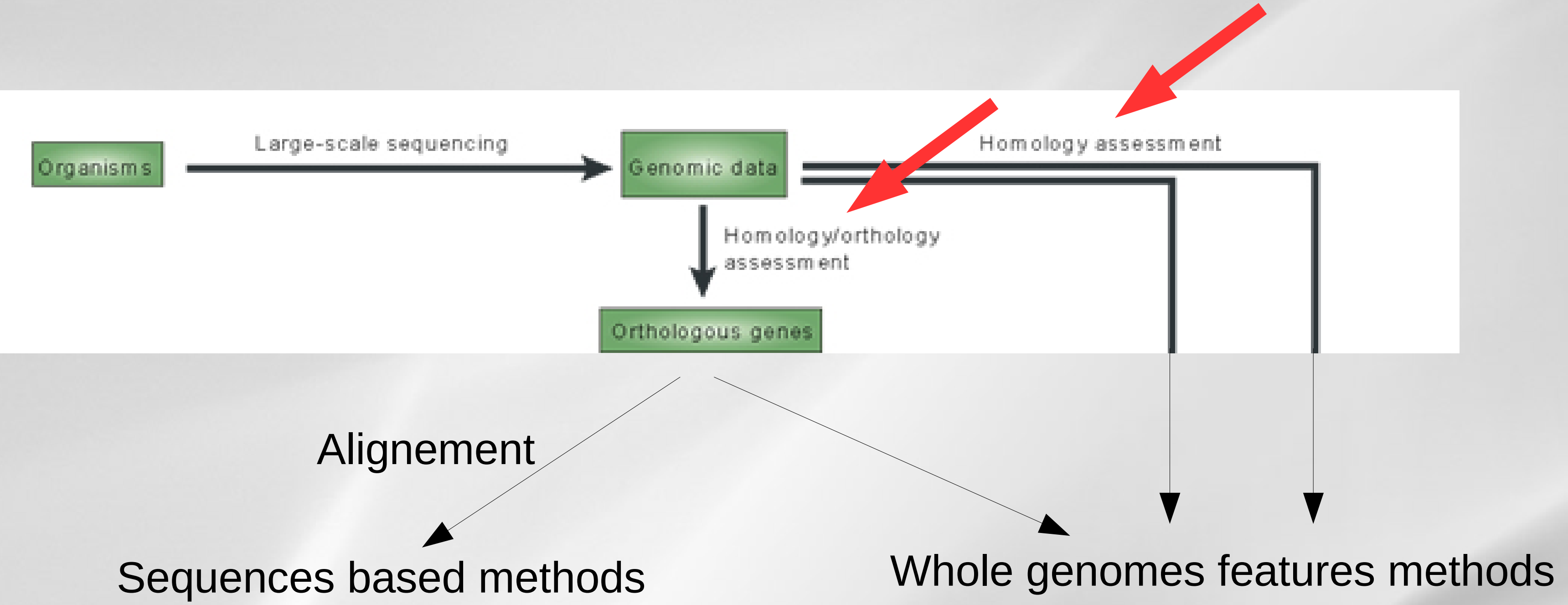
There is a lot of inconsistency sources in individual gene data, so in practice we integrate a lot of informations by assuming that the phylogenetic signal that we want is dominant.

Phylogenomic analysis : the type of methods



(From Delsuc *et al*, Nature reviews, 2005)

Phylogenomic analysis : the type of methods



Organisms

Large-scale sequencing

Genomic data

Homology assessment

Homology/orthology assessment

Orthologous genes

Alignement

Sequences based methods

Whole genomes features methods

(From Delsuc *et al*, Nature reviews, 2005)

BBH the most widely used method to infer potential orthology

- Synteny can be used to improve it.

- But there are some difficulties for example when the genomes have undergone duplications this method misses many orthologs. (Dalquen and Dessimoz, Genome biology and evolution, 2013)

BBH : advantages and limitations

Similar evolutionary rates

Detect only 1 to 1 orthology

Orthologs losses

Orthologs evolve faster

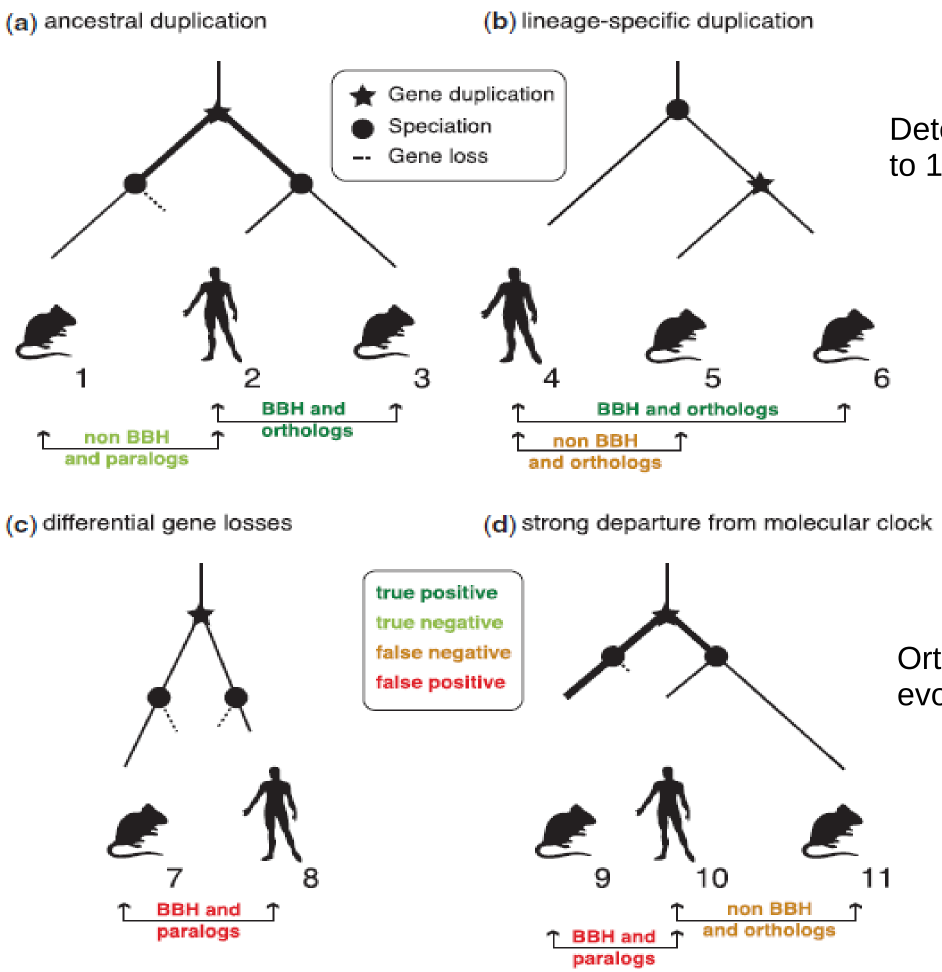


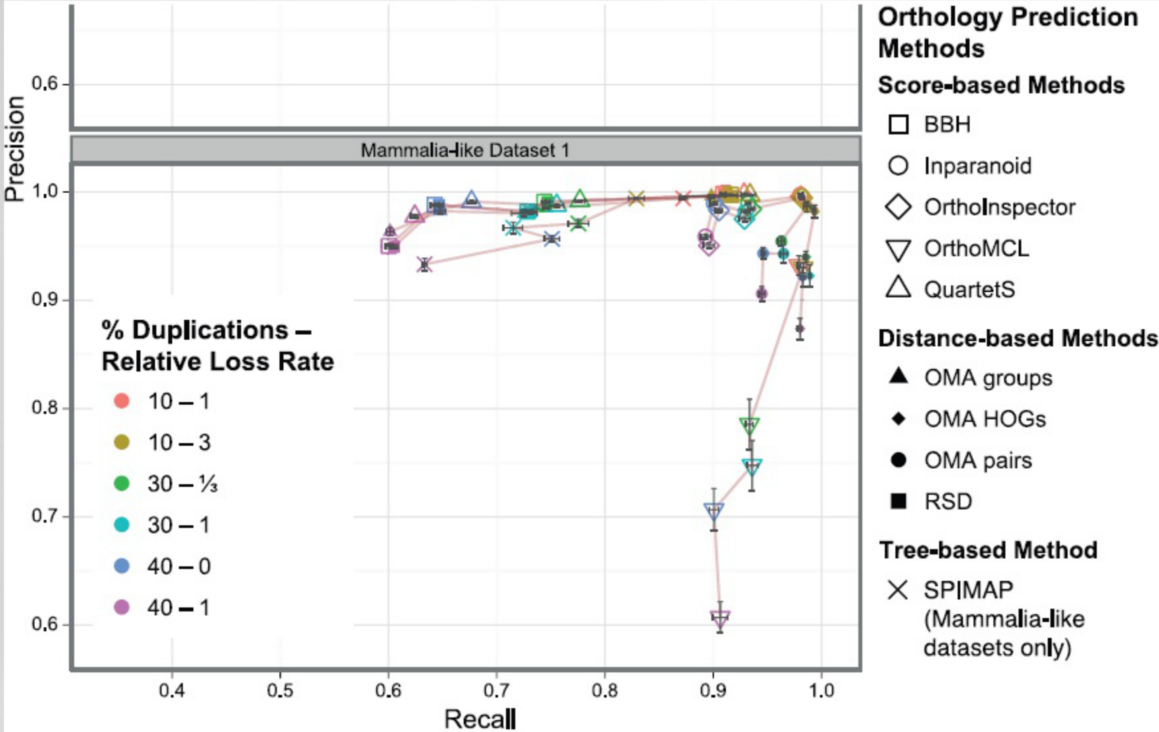
Fig. 1.—Performance of BBH in conceptual examples. (*a*) BBH recovers the orthologous pair, because the orthologous pair is closer than the paralogous pair due to evolution accrued between the duplication and speciation events (highlighted in bold). (*b*) BBH only identifies one of the two orthologous pairs, namely the one with higher score. This scenario is common if duplication occurs after speciations of interest. (*c*) BBH identifies paralogs if the orthologous counterpart is missing in both species. This might happen if the rate of gene losses is high (e.g., following a whole genome duplication). (*d*) BBH identifies paralogs if the departure from the molecular clock is so strong that paralogs are closer in sequence despite having started diverging before the orthologs.

(Dalquen and Dessimoz, 2013)

Potential orthology inference : many tools

- Score based methods:
  - BBH
  - OrthoMCL (Li et al., 2003)
- Distance-based methods:
  - OMA (Roth at al., 2008 ; Altenhoff et al., 2011)
- Tree-based methods:
  - SPIMAP (Rasmussen et al., 2011)

Potential orthology inference : many tools to be chosen according to the characteristics of the data



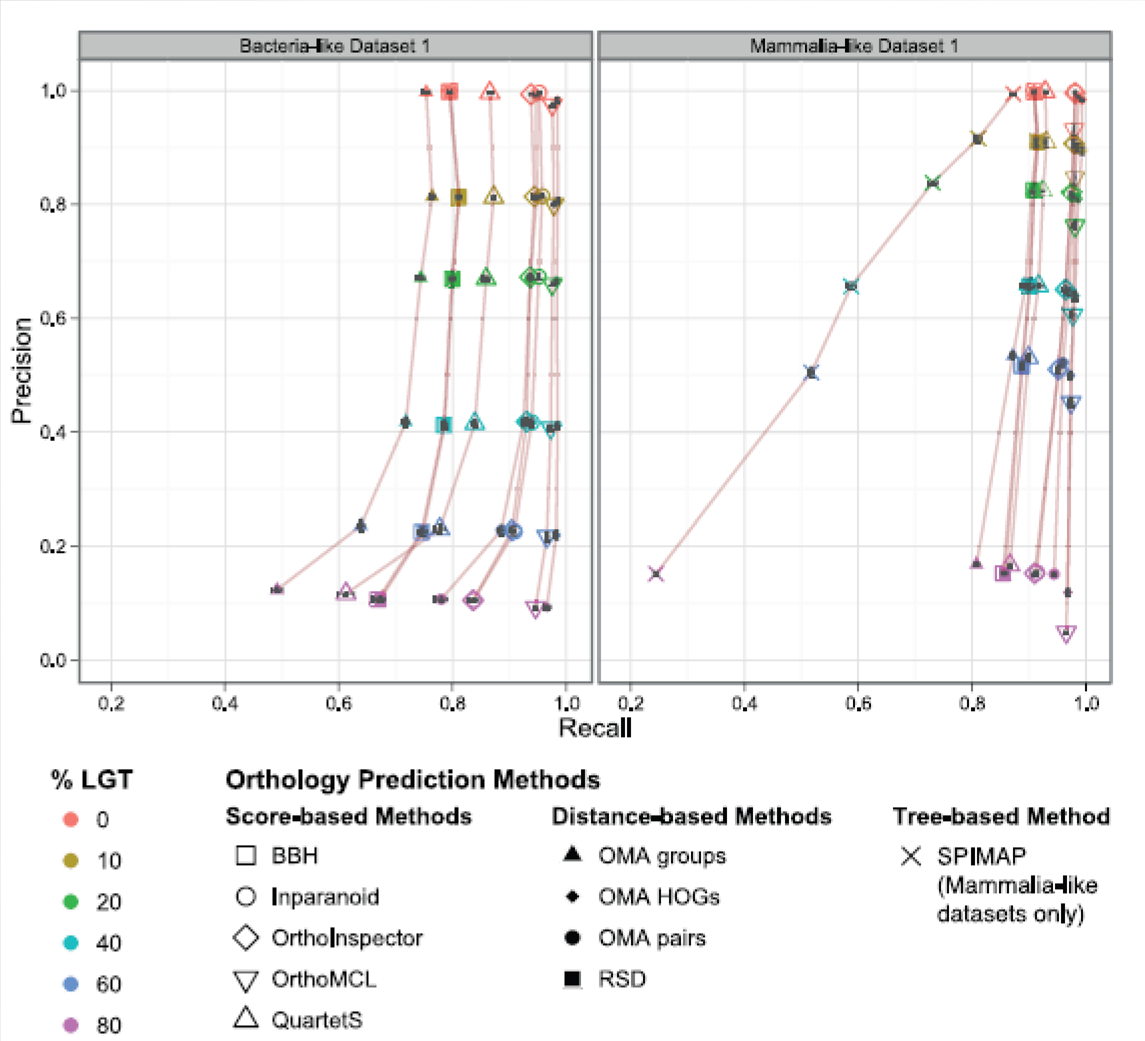
Sum TP / Sum test outcome P

Sensitivity = Sum TP / Sum P

(Dalquen *et al*, PLOS one, 2013)

- When the gene duplication rate is high, BBH misses a large proportion of orthologs. But in experiment that only require few but trusted orthologs, the performance of BBH is sufficient.
- Best : Inparanoid, orthoInspector, OMA HOGS, OMA pairs

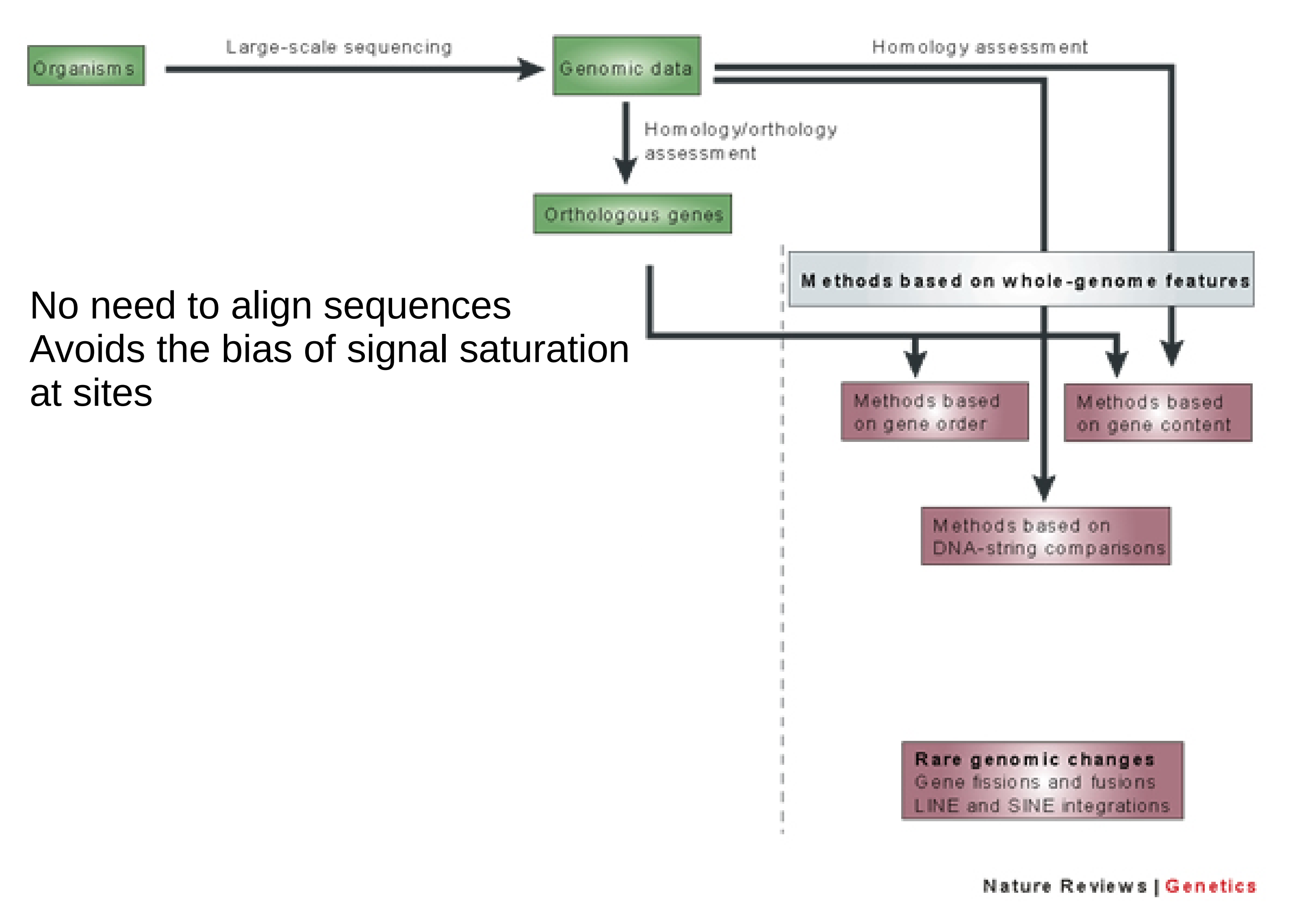
Orthology inference : many tools to be chosen according to the characteristics of the data



- All methods are very sensitive to LGT.

(Dalquen *et al*, PLOS one, 2013)

Phylogenomic analysis : the methods


Organisms
Large-scale sequencing
Genomic data
Homology assessment
Homology/orthology assessment
Orthologous genes
Methods based on whole-genome features
Methods based on gene order
Methods based on gene content
Methods based on DNA-string comparisons
Rare genomic changes
Gene fissions and fusions
LINE and SINE integrations
Nature Reviews | Genetics

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

(From Delsuc *et al*, Nature reviews, 2005)

Whole genome features methods

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

Plateforme Bioinformatique Midi-Pyrénées

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
- Disadvantages: big/small genome attraction

Comparison of gene content

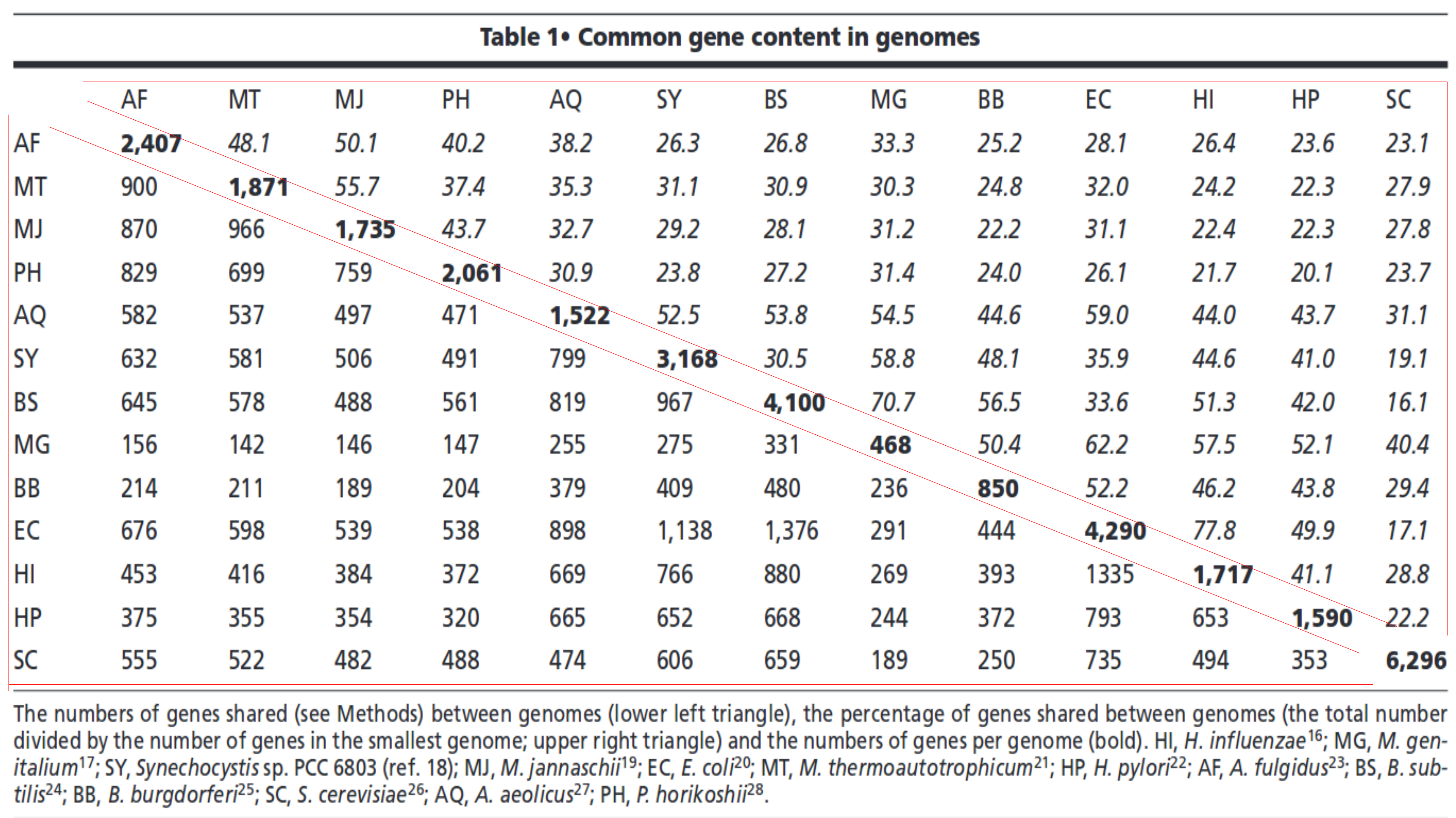


Table 1• Common gene content in genomes

	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 stell, 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

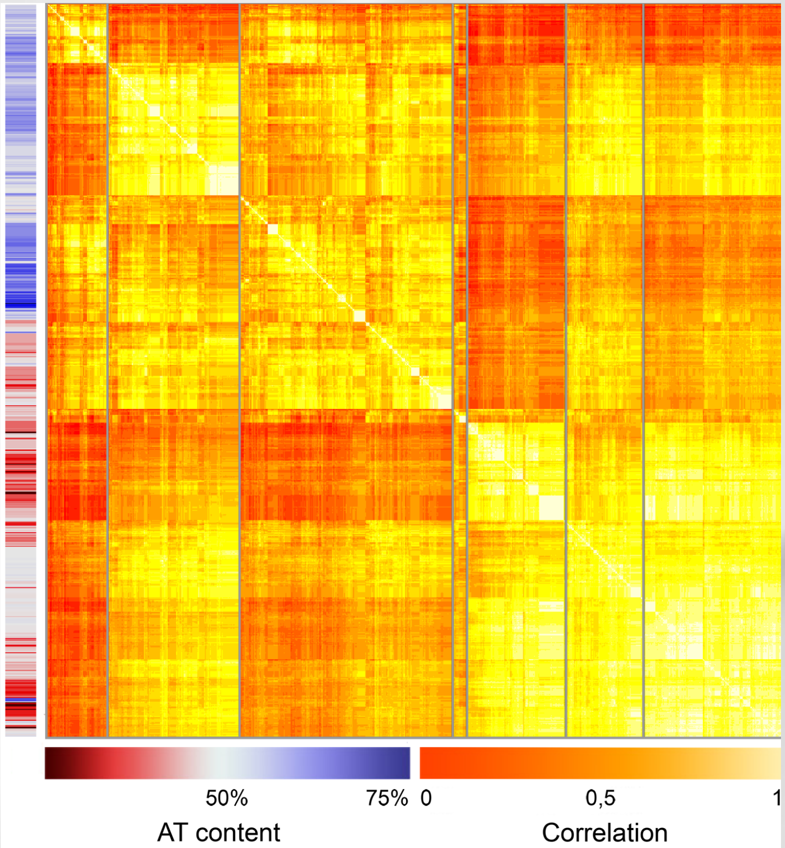
- Used for mitochondries and chloroplasts genomes
- Low error rate
- Rare events in eucaryotes genomes (large evolutionary scale)
- Problems :
  - Very limited data (mostly organelles)
  - Mathematics complex
  - Evolutionary models not well known

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).



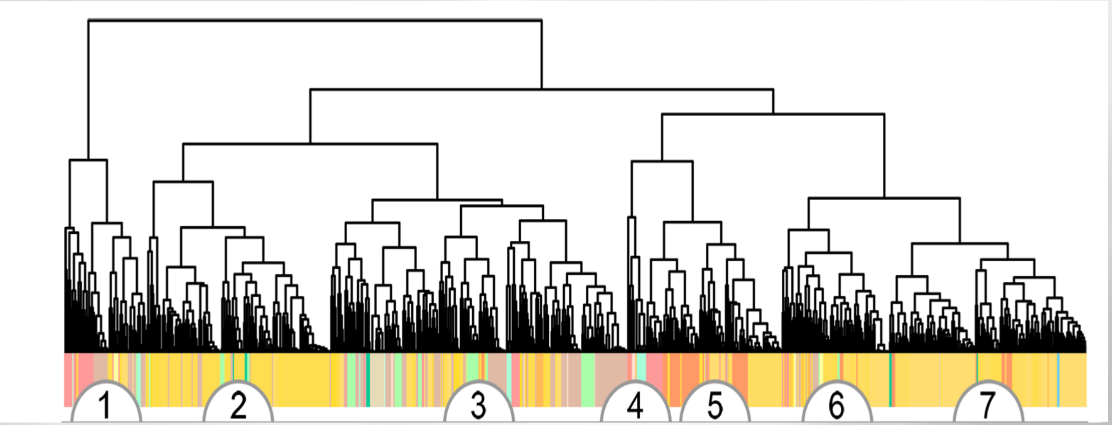
867 prokaryotic genomic DNA sequences compared pair-wise using hexanucleotide-based genomic signatures

.

(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

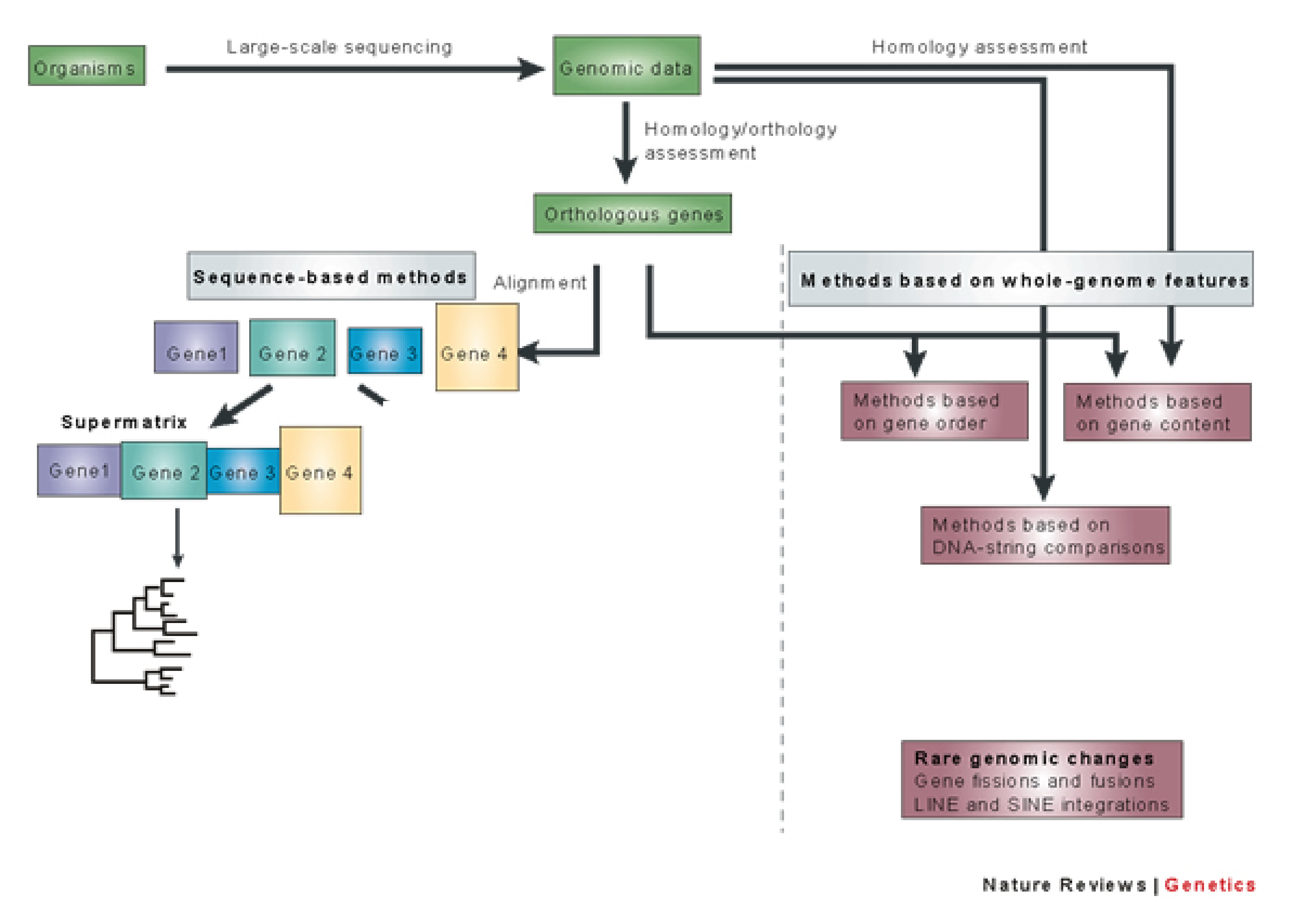


- ■ Archaea
- ■ Acidobacteria
- ■ Actinobacteria
- ■ Aquificae
- ■ Bacteroidetes/Chlorobi
- ■ Chlamydiae/Verrucomicrobia
- ■ Chloroflexi
- ■ Cyanobacteria
- ■ Deinococcus-Thermus
- ■ Firmicutes
- ■ Fusobacteria
- ■ Planctomycetes
- ■ Proteobacteria
- ■ Spirochaetes
- ■ Thermotogae

Cluster diagram of 867 prokaryotic genomic DNA sequences compared pair-wise using hexanucleotide-based genomic signatures

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

Phylogenomic analysis : the methods


Organisms
Large-scale sequencing
Genomic data
Homology assessment
Homology/orthology assessment
Orthologous genes
Sequence-based methods
Alignment
Methods based on whole-genome features
Gene1
Gene 2
Gene 3
Gene 4
Supermatrix
Gene1
Gene 2
Gene 3
Gene 4
Methods based on gene order
Methods based on gene content
Methods based on DNA-string comparisons
Rare genomic changes
Gene fissions and fusions
LINE and SINE integrations
Nature Reviews | Genetics

(From Delsuc *et al*, Nature reviews, 2005)

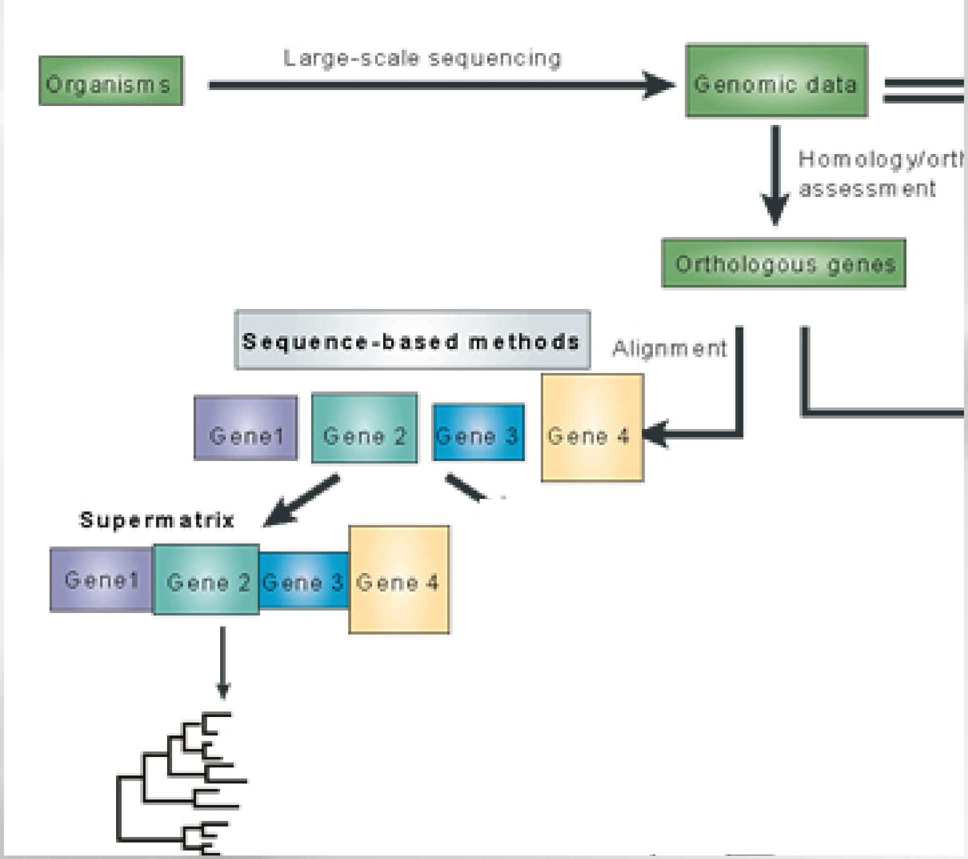
Plateforme Bioinformatique Midi-Pyrénées

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)

Gene 1

OTU 1 ______________
OTU 2 ______________
OTU 3 ______________

Gene 2

OTU 1 __________
OTU 4 __________
OTU 5 __________

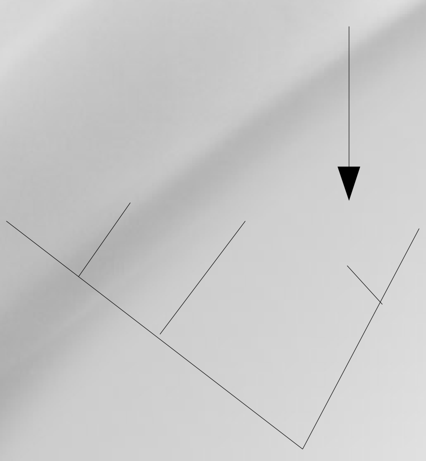
...

Gene n

OTU 2 ___________
OTU 3 ___________
OTU 4 ___________

OTU 1 ______________________________...???????????
OTU 2 ______________??????????..._________
OTU 3 ______________??????????..._________
OTU 4 ??????????????__________..._________
OTU 5 ??????????????__________...???????????

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 (ex: CAT : by categories, heterogeneity of evolutionary process)
- 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.

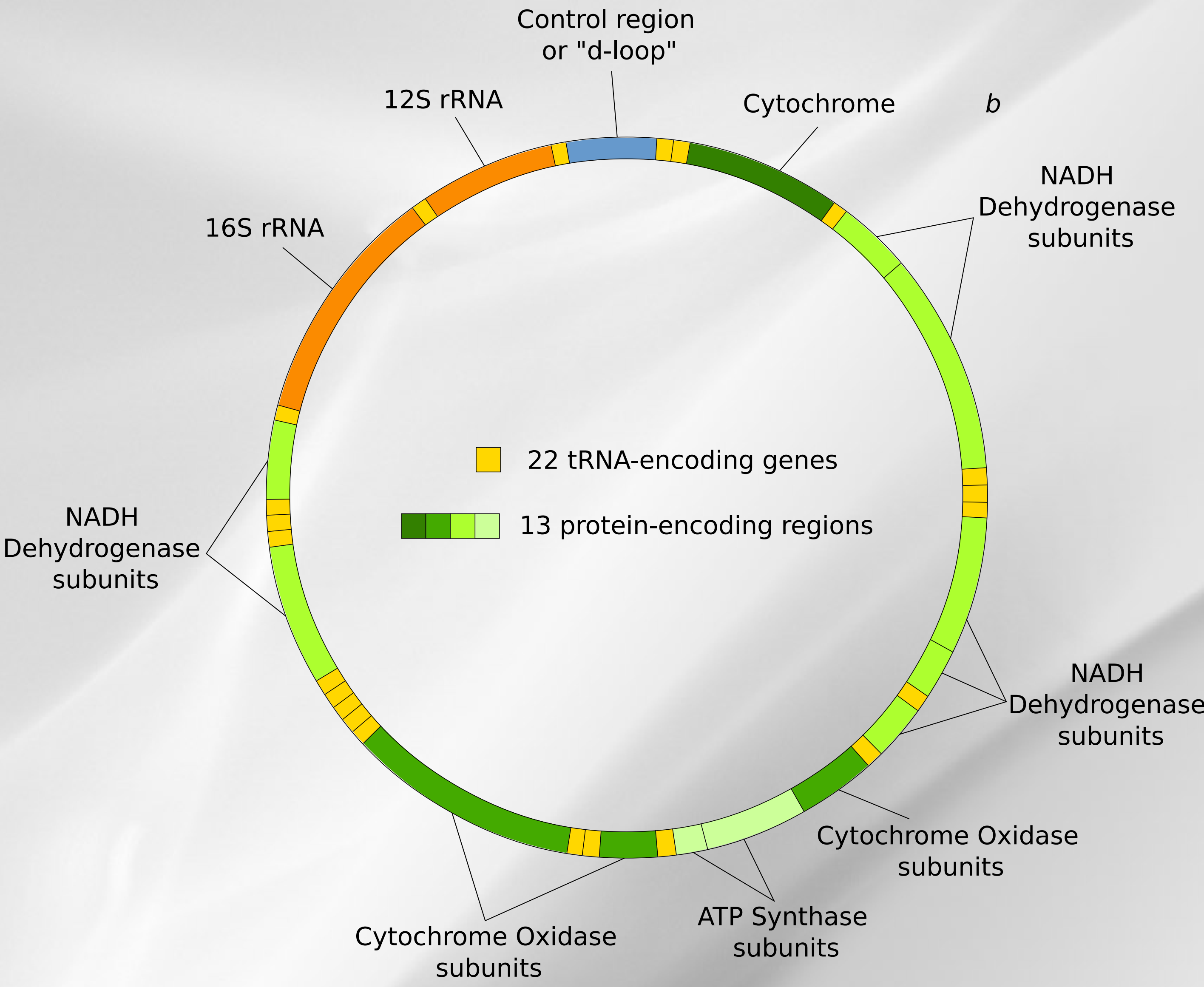
The supermatrix approach (4)

- 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)

TP 1:

- get started with the mitochondrial genes dataset
- build several trees from the concatenation file and comparing them visually

The mitochondrial genome of mammalian



http://en.wikipedia.org/wiki/Mitochondrial_DNA#mediaviewer/File:Mitochondrial_DNA_en.svg

TP1 Supermatrix method

- Superalignment of the 13 genes encoding proteins in the motochondrial genome of 66 primates
  - Transeq, Clustalo, catfasta2phyml.pl, Gblocks, script to recode in codon used and fasta2phylip : *prot_nt.concat.phy*
  - Transeq, Clustalo, catfasta2phyml.pl, Gblocks, and fasta2phylip : *prot_aa.concat.phy*
- Superalignment of the 25 genes RNA of 66 primates
  - Mafft with -qinsi, catfasta2phyml.pl, Gblocks : *RNA.concat.phy*

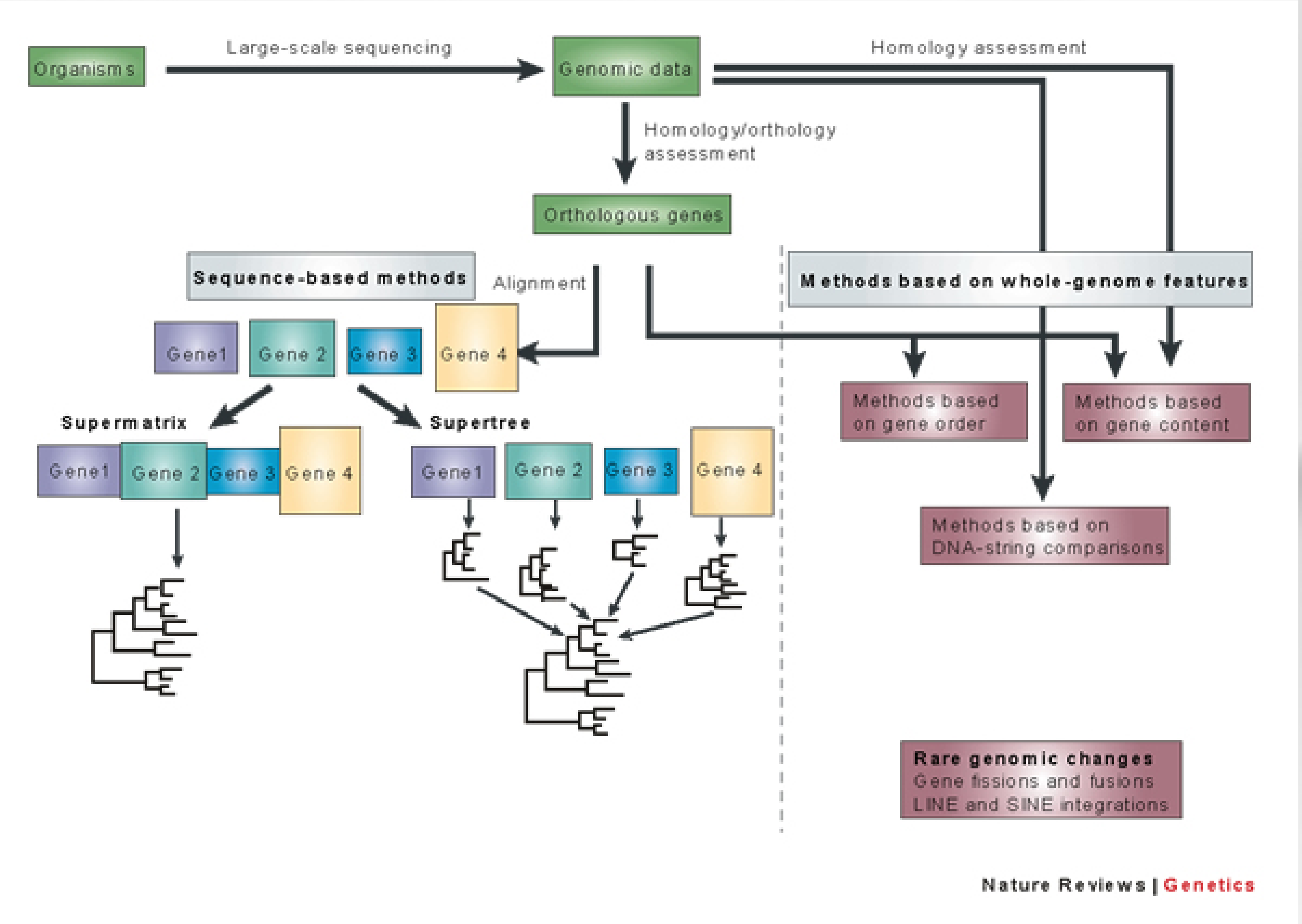
TP1 Supermatrix method

- qsub -V -b Y -N phymlRNA -l h_vmem=10G -l mem=8G "phyml -i RNA.concat.phy -n 1 -b -1 -m GTR -v e -c 4 -a e -o tlr --quiet"
- Look at the result files of the analysis abose and these files:
  - prot_nt.concat.phy_phyml_stats et prot_aa.concat.phy_phyml_stats.
  - Combien de temps l'analyse a-t-elle duré dans le premier cas ? Quel est la longueur de l'alignement utilisé ?
  - Quelle commande a-t-on lancée dans le dernier cas ?

TP1 Supermatrix method

- Compare the three trees obtained visually with figtree
  - Can you see the differences ? And with the tree in the paper : Menezes *et al.*, 2013 ? (Il manque 4 espèces)
  - The species name are in the file: SpeciesNames.txt
  - There is a paper about the phylogeny of primates (Perelman *et al*, 2011)
  - Globalement que pensez-vous de ces arbres ? Les supports, leur congruence ? Les grands groupes sont-t-ils retrouvés ?

Phylogenomic analysis : the methods


Organisms
Large-scale sequencing
Genomic data
Homology assessment
Homology/orthology assessment
Orthologous genes
Sequence-based methods
Alignment
Methods based on whole-genome features
Gene1
Gene 2
Gene 3
Gene 4
Supermatrix
Supertree
Gene1
Gene 2
Gene 3
Gene 4
Gene1
Gene 2
Gene 3
Gene 4
Methods based on gene order
Methods based on gene content
Methods based on DNA-string comparisons
Rare genomic changes
Gene fissions and fusions
LINE and SINE integrations
Nature Reviews | Genetics

(From Delsuc *et al*, Nature reviews, 2005)

Sequence-based methods

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

The supertree approach

Gene 1

OTU 1 ______________
OTU 2 ______________
OTU 3 ______________

M1 model fixed
P1 set of parameters inferred
ML or bayesian methods

Gene 2

OTU 1 __________
OTU 4 __________
OTU 5 __________

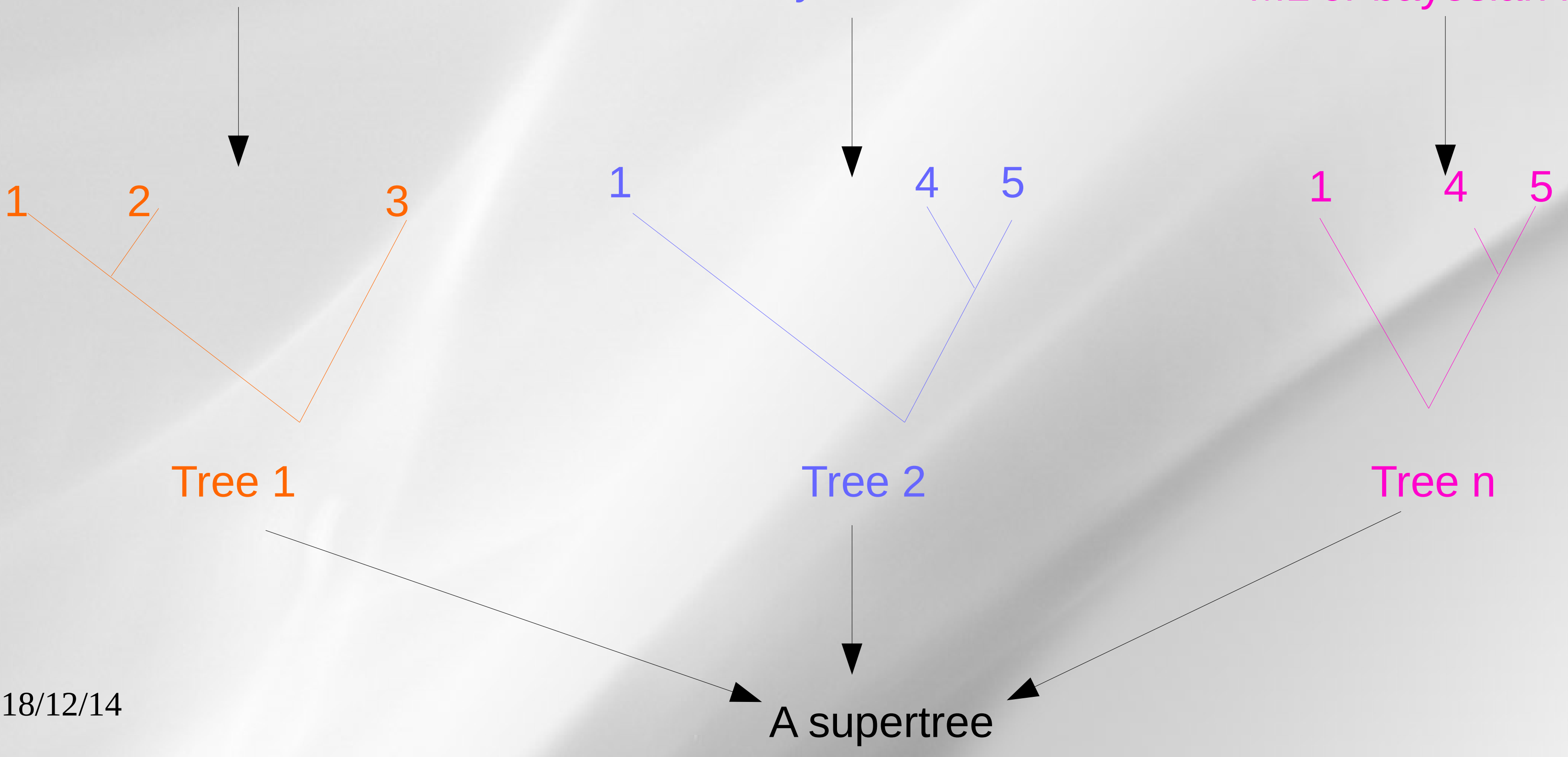
M2 model fixed
P2 set of parameters inferred
ML or bayesian methods

...

Gene n

OTU 2 ___________
OTU 3 ___________
OTU 4 ___________

Mn model fixed
Pn set of parameters inferred
ML or bayesian methods



Tree 1

Tree 2

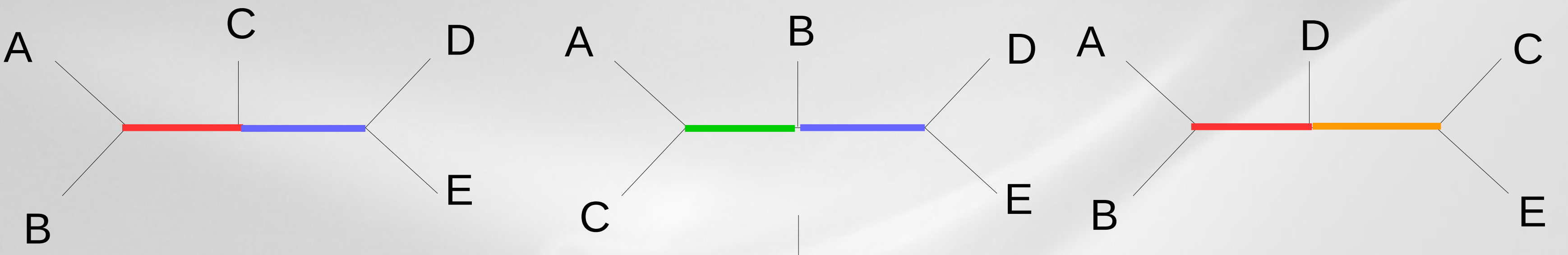
Tree n

A supertree

Consensus Tree

- Used to test the tree robustness and for the bootstrap
- Strict consensus tree: a bipartition will be included if it's present in all input trees
- Majority consensus tree: a bipartition will be included if it's present in more than half of the input trees

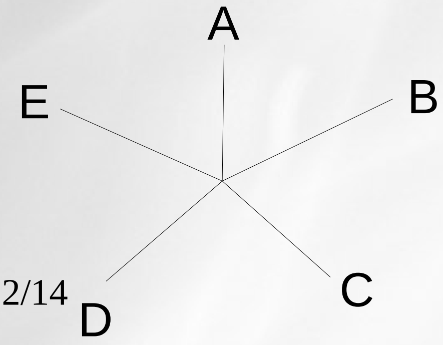
Consensus Tree (2)



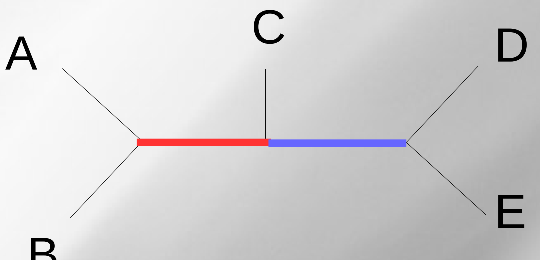
Weighted bipartitions

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

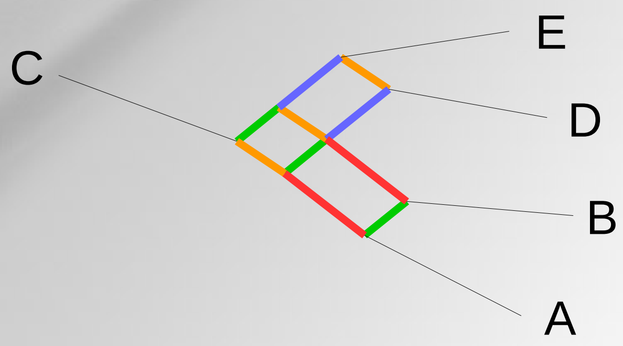
Strict consensus (100%)



Majority consensus (50%)



Consensus networks (≥ 33%)



(Holland & Moulton, Algorithms in bioinformatics, 2003)

Network Tree

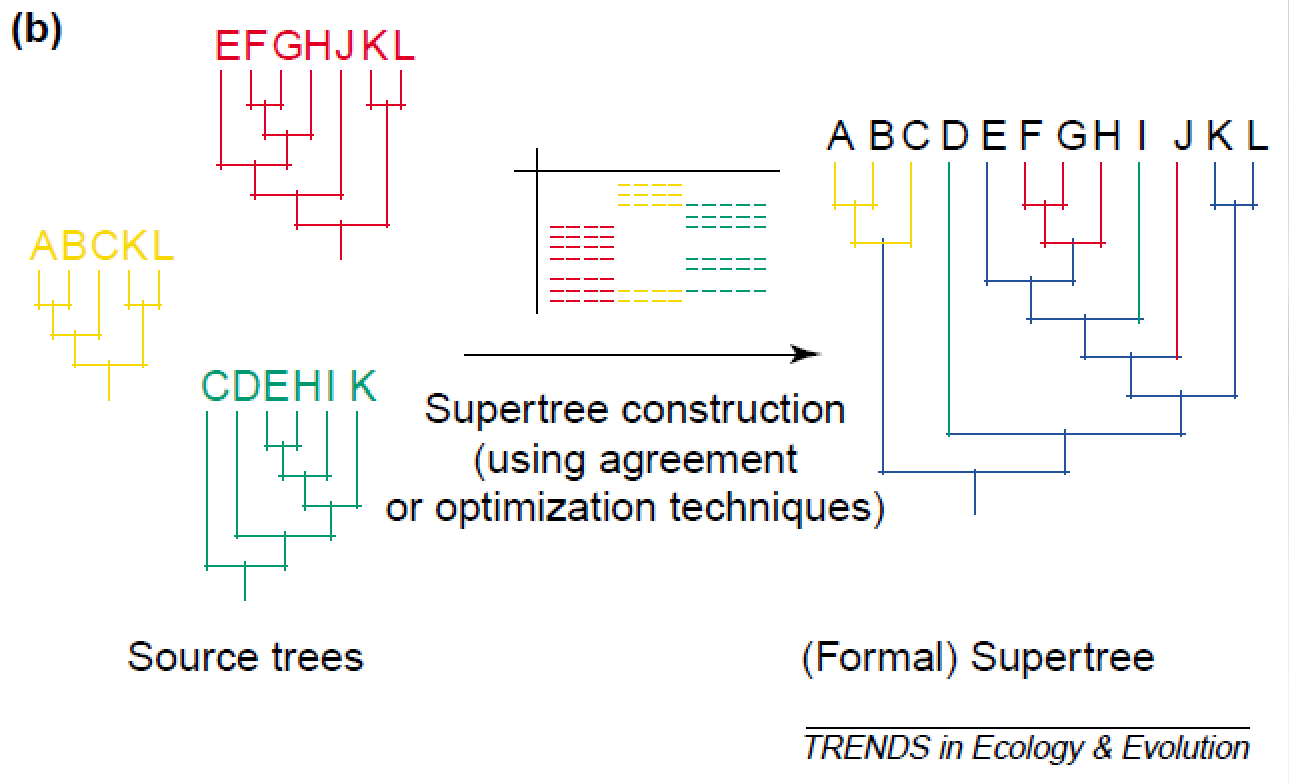
- Consensus network is one method to build network tree.
- Splitstree, for example, is a program for computing unrooted phylogenetic networks from molecular sequence data http://www.splitstree.org/, (Huson & Bryant, 2006).
- Phylogenetic networks should be looked when hybridization, horizontal gene transfer,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



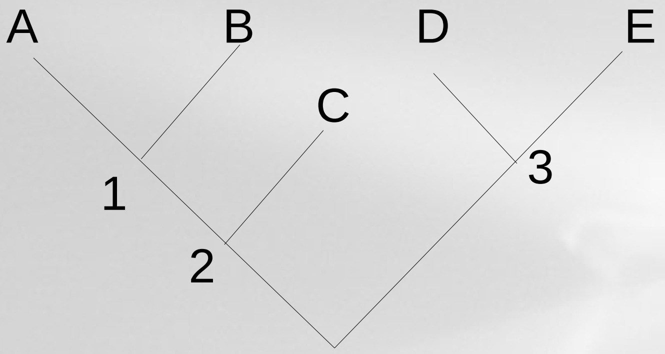
(Bininda-Emonds O., 2004)

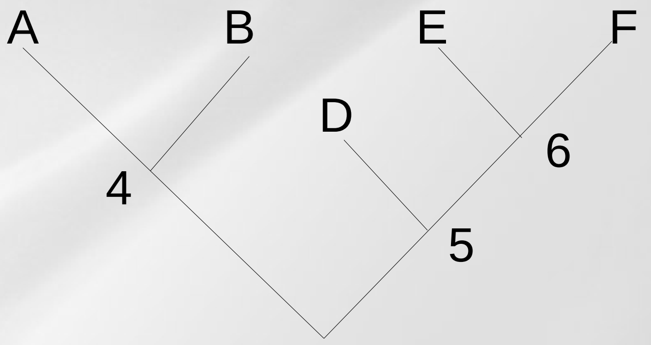
Matrix representation using parsimony

- This is the most common method
- MRP needs a matrix representation

(Bininda-Emonds O., Trends in ecology and Evolution, 2004)

Build a super-tree MRP



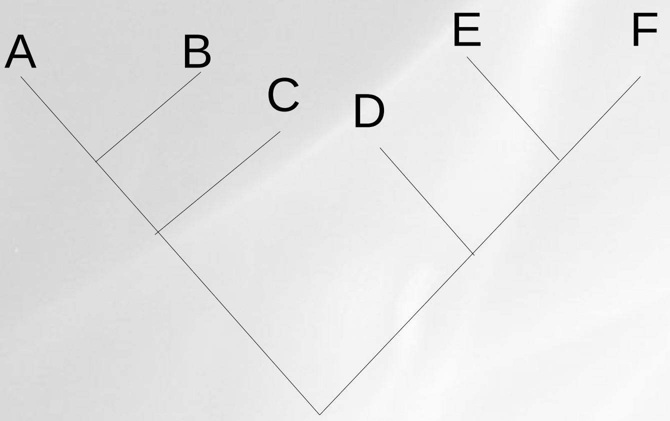


Binary matrix representation
(Baum and Ragan, 1992)

	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

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

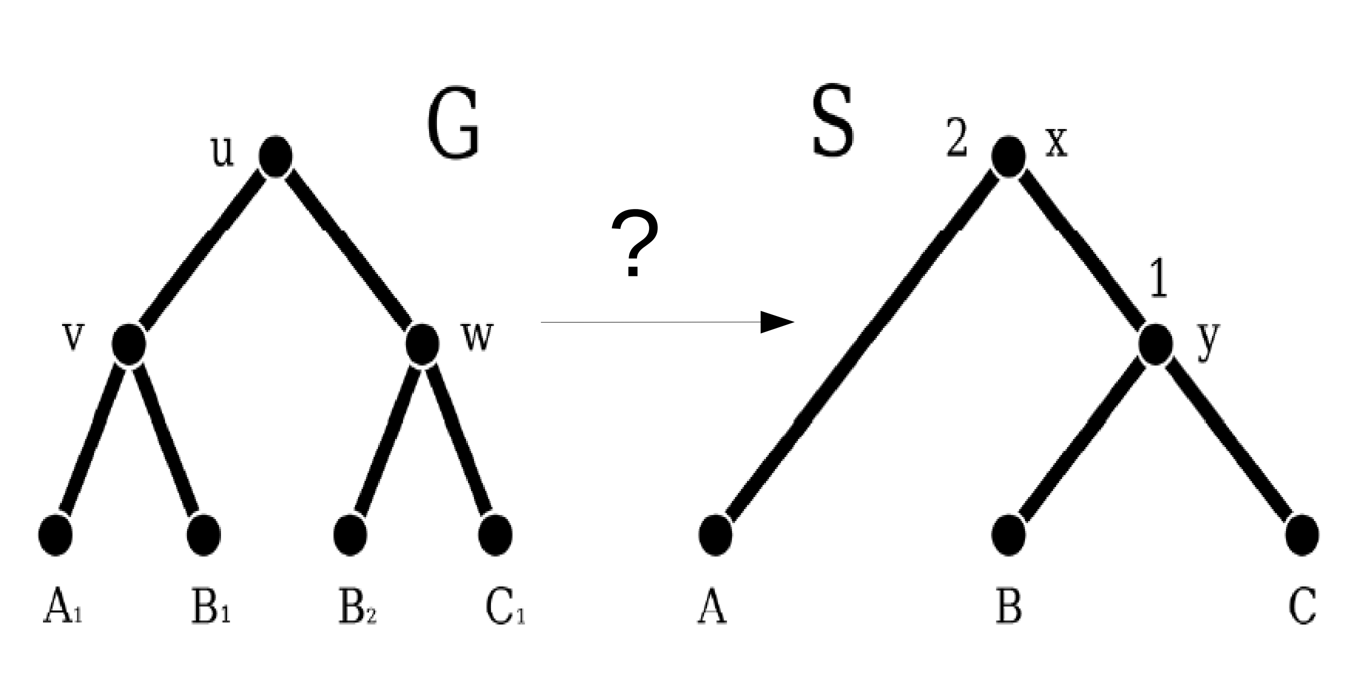
Super-tree
MRP



Super Tree methods: advantage / disadvantage

- (-) The length of branches are not directly interpretable in terms of evolutionary distance
- (+) It's faster for very large dataset that super matrix approach
- (+) Phylogeny of each gene is made with the appropriate model and parameters and/or methods
- (-) When the input alignments are too small or don't have enough phylogenetic signal this can become problematic, because most methods weigh poorly-supported and well-supported input trees equally

Reconciliation



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

(Nguyen T-H *et al.*, PLOS one 2013)

Reconciliation methods

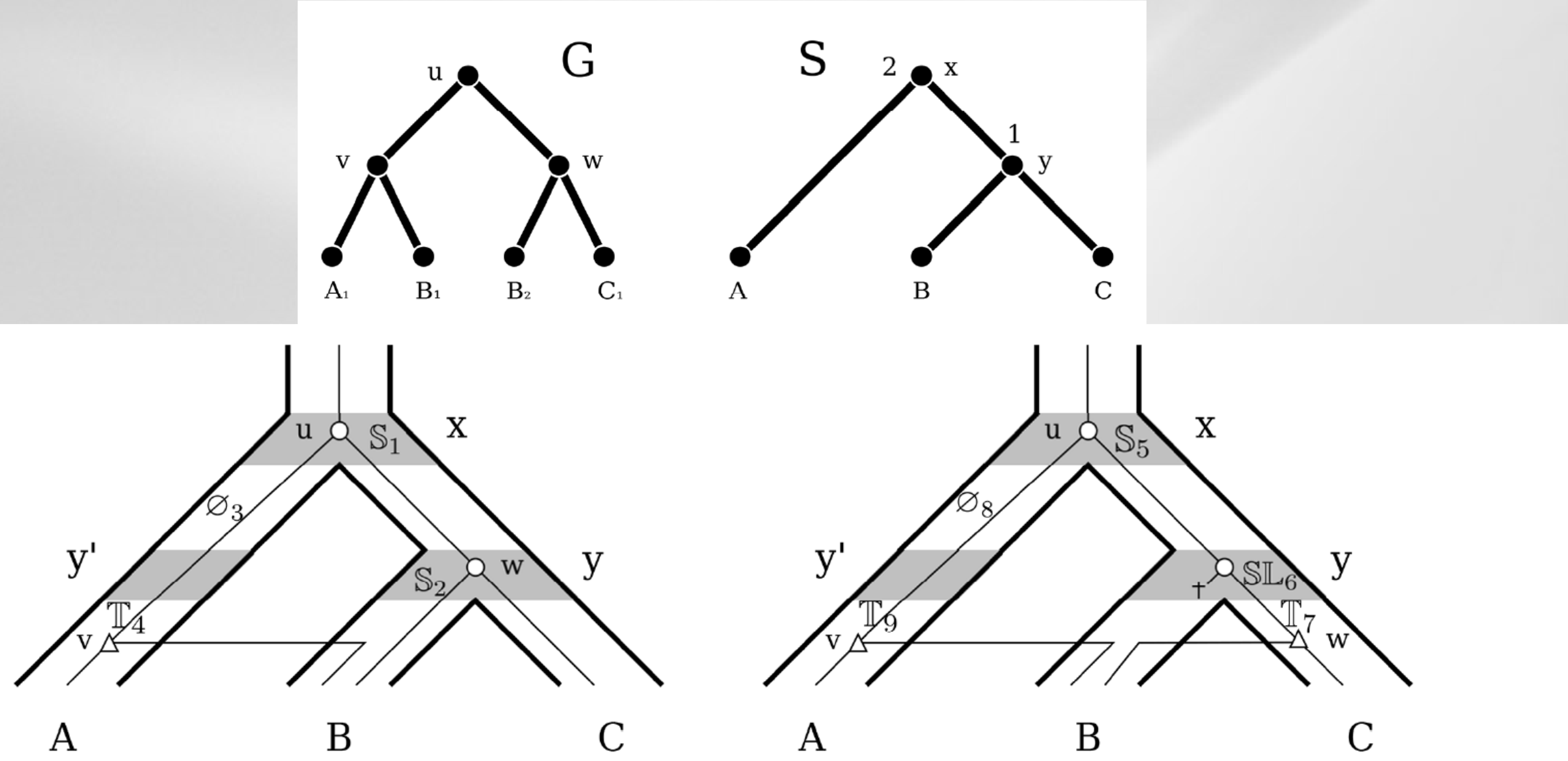


Figure 2. An example of reconciliation. Two valid reconciliations for the trees depicted in Figure 1 ($\mathbb{C}$ events are not indicated). The reconciliation on the left contains two $\mathbb{S}$ events, one $\varnothing$ event, a $\mathbb{T}$ event and four $\mathbb{C}$ events, while the one $\mathbb{S}$ on the right contains one $\mathbb{S}$ event, one $\mathbb{SL}$, one $\varnothing$ event, two $\mathbb{T}$ events and four $\mathbb{C}$ 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, Ø 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.

(Nguyen T-H *et al.*, PLOS one 2013)

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

TP 2:

- building a super-tree
- compare with other trees visually and with metrics if we have time

TP2 Supertree method

- Chacun prend un des alignements de gène protéique
  - Fichiers : Name.idx.2.fa.align.rename
  - Lancer le Gblocks avec les paramètres par défaut sauf mettre codons pour le type d'alignement.
  - Lancer perl fasta2phylip.pl pour convertir en format phylip
  - Puis lancer phyML avec un modèle GTR+I+G, 4 catégories si vous ne voulez pas vérifier le modèle
  - Copier votre arbre avec un nom explicite dans /tmp

TP2 Supertree method

- Concaténer tous les arbres dans un même fichier (commande cat)
- Tapez qrsh dans un autre terminal connecté à genotoul
- Puis appelez R
- library(phytools)
- trees = read.tree("Mon Path/allTree.txt")
- supertree<-mrp.supertree(trees,rearrangements="SPR", start="NJ")

TP2 Supertree method

- Vous avez obtenu les super arbres les plus parcimonieux. Sauvez-les.
- Ex pour le premier arbre : write.tree(supertree[[1]],file = "/home/choede/work/formation_phylo/superTree1")
- Faire un consensus de ces superarbres.
- Renommer la sortie : mv outtree supertree.cons
- Faire le consensus avec consense des 13 arbres de gènes.
- Renommer la sortie : mv outtree consensus.tree

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)

TP2 trees distance

Utilisez treedist pour déterminer les distances topologiques (symmetric difference) entre les différents arbres obtenus précédemment : prot_nt, prot_aa, RNA.concat, le consensus des arbres MRP et le consensus des 13 arbres de gènes.

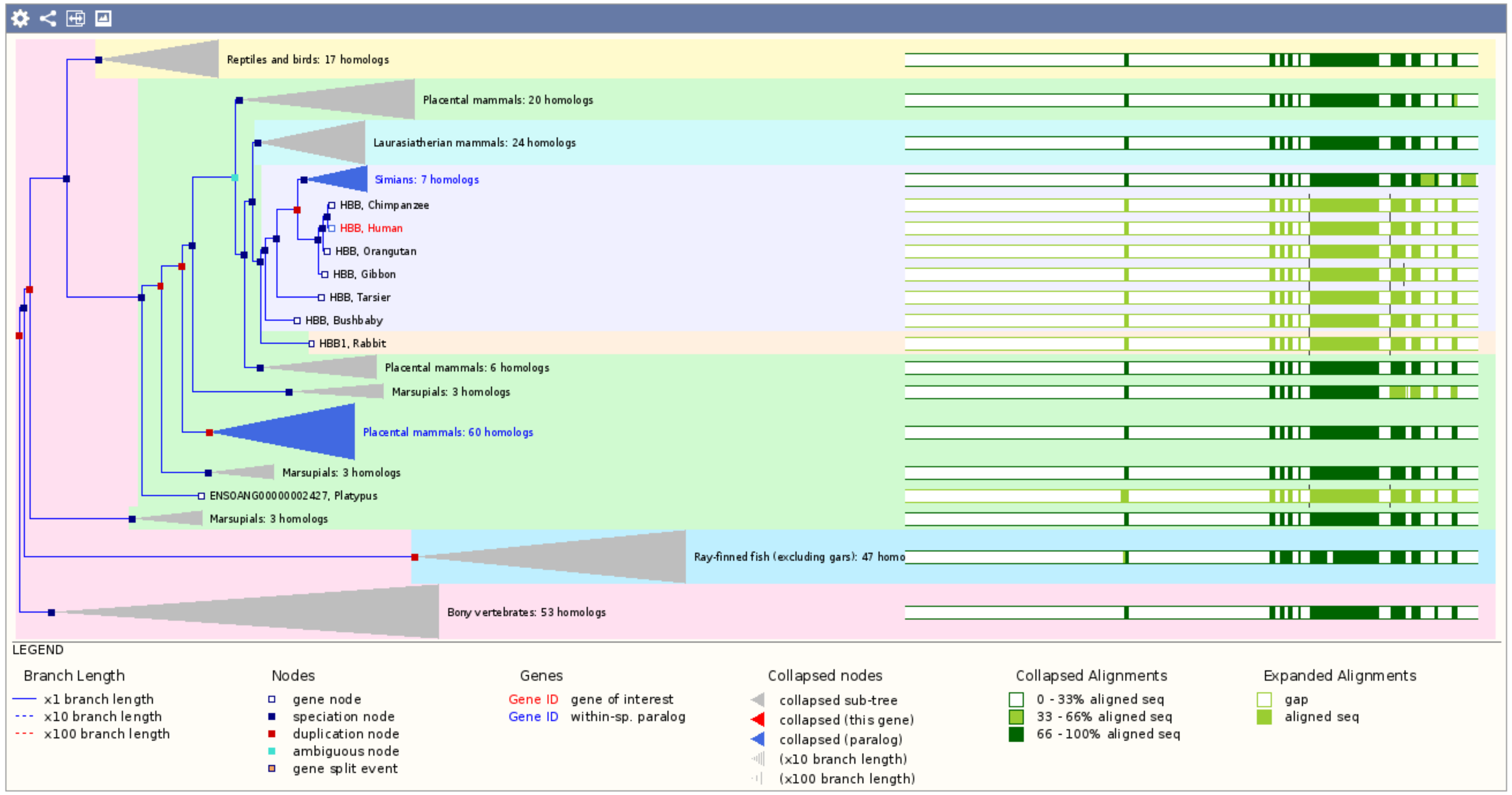
- Pour cela faire un fichier qui les concatène (se souvenir de l'ordre)
- cat consensus.tree supertree.cons prot_aa.concat.phy_phyml_tree.txt prot_nt.concat.phy_phyml_tree.txt RNA.concat.phy_phyml_tree.txt > trees.final

TP2 Supertree method

- Quels sont les arbres les plus proches topologiquement ?
- Quels sont les arbres les plus éloignés topologiquement ? Est-ce attendu ?
- Dans ce jeu de données avait-on besoin de faire un super-arbre ?
- Dans le consensus, comme dans le super arbre, est-ce que NC_010299 Daubentonia madagascariensis est bien placé ? (Il doit être plus proche des Propithecus que des Lorisidae)
- Qu'en pensez-vous ?

Ensembl compara

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

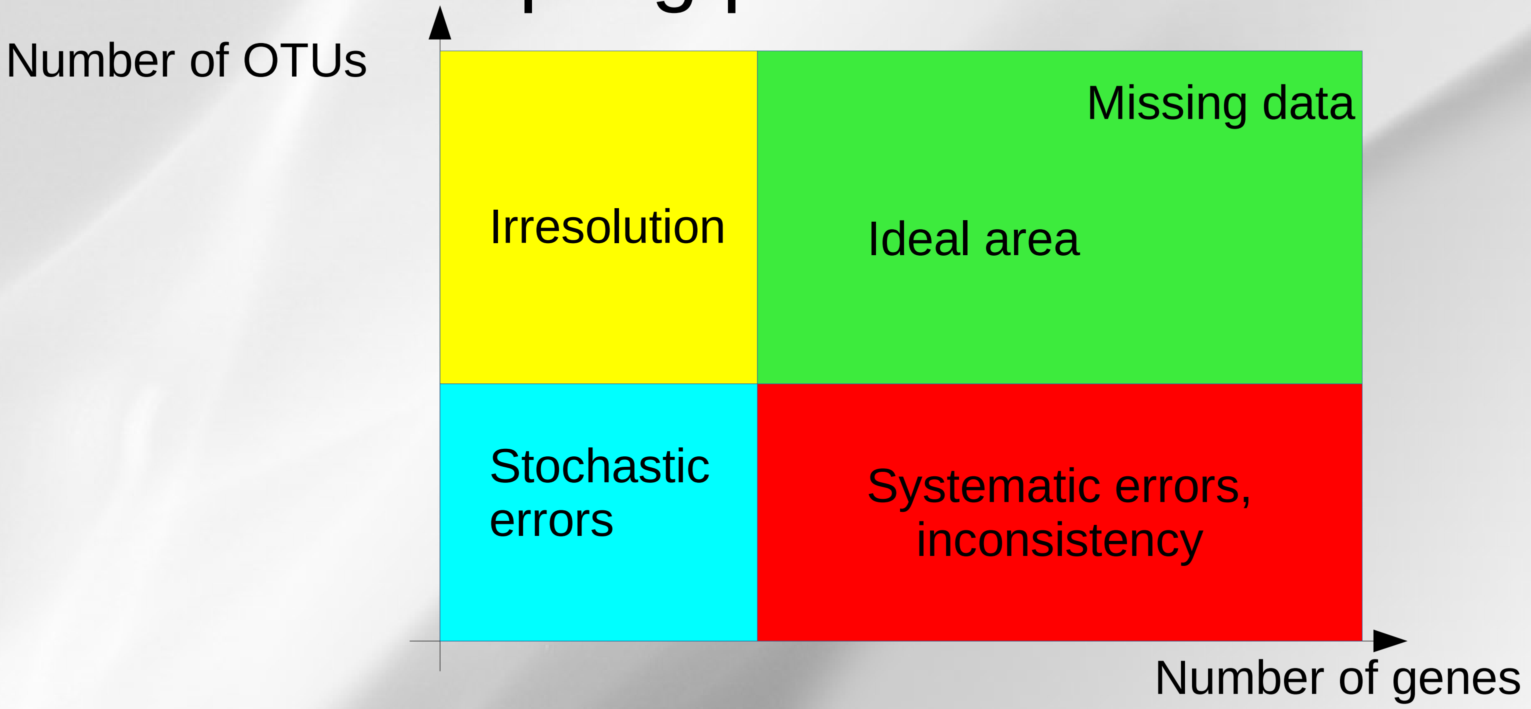


Ensembl compara

- 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



Number of OTUs

Irresolution

Missing data

Ideal area

Stochastic errors

Systematic errors, inconsistency

Number of genes

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

- Scientifique articles cited in the slides
- Presentation :
  - M2 – Phylogénomique. Frédéric Delsuc : Equipe de Phylogénie et Evolution Moléculaire, Institut des Sciences de l'Evolution de Montpellier
- Thèse :
  - Béatrice Roure soutenue en 2011 : « Amélioration de l'exactitude de l'inférence phylogénomique »