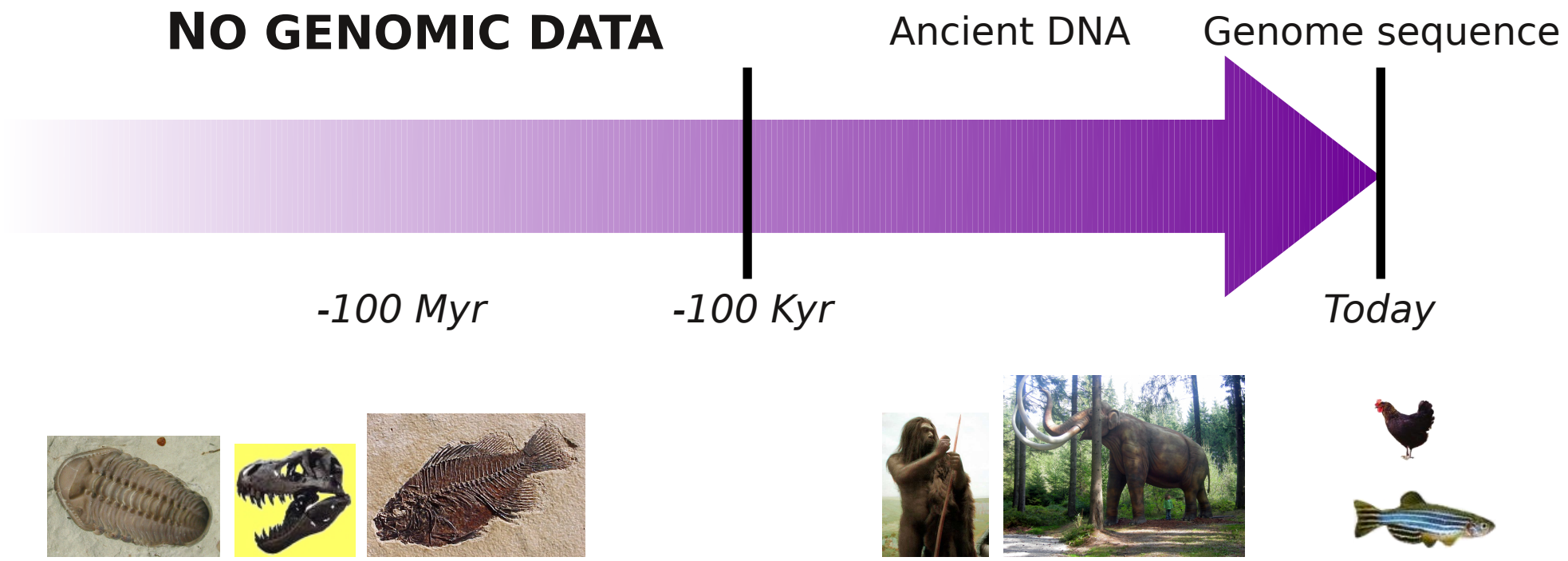
Reconstruction de génomes ancestraux chez les vertébrés

Matthieu Muffato

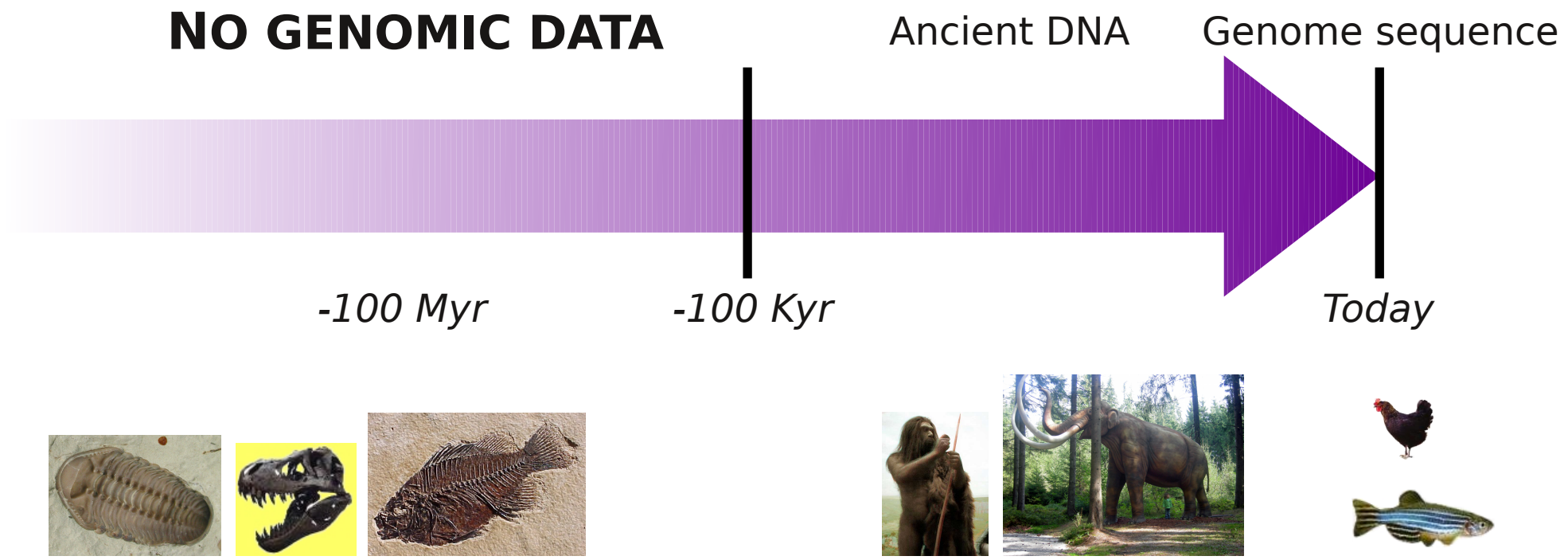
Soutenance de thèse – 15/12/2010

DYOGEN Dynamique et Organisation des Génomes
Ecole Normale Supérieure - 46, Rue d'Ulm 75005 Paris

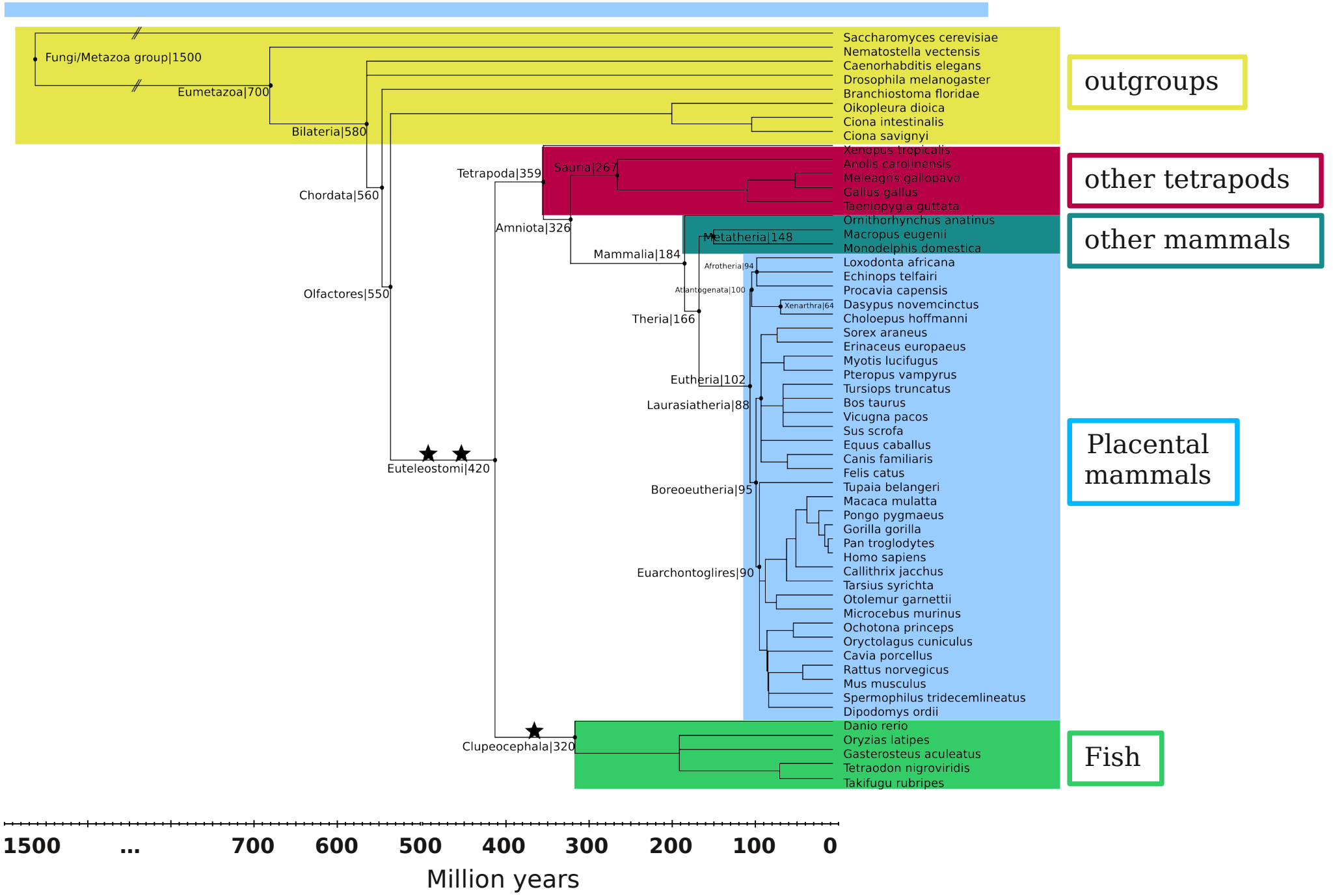
The necessity of ancestral genomes



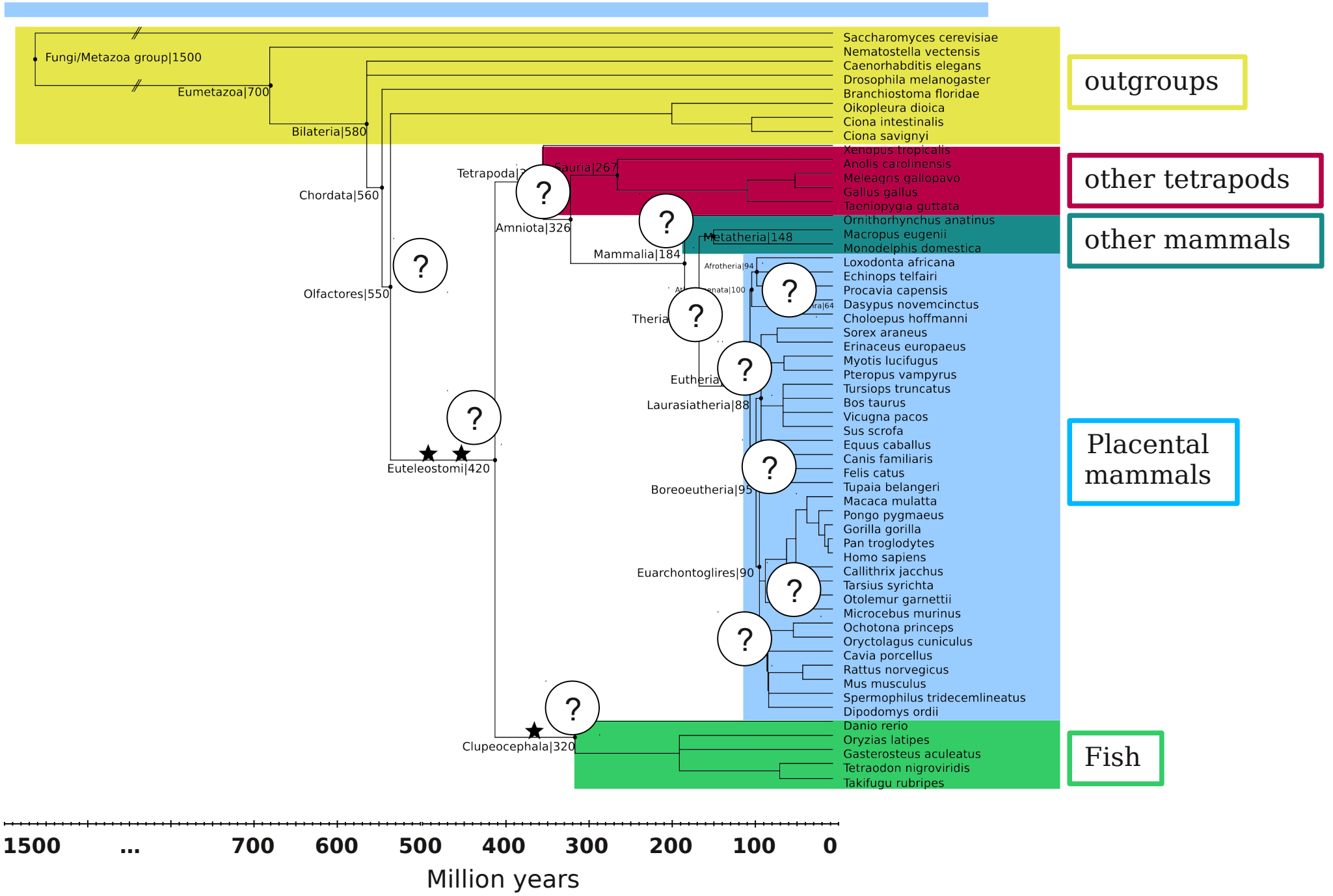
The necessity of ancestral genomes



Phylogenetic tree of vertebrates & model outgroup species



Phylogenetic tree of vertebrates & model outgroup species



Existing tools for ancestral genome reconstruction

Ancestral genome reconstruction has been adressed by:

- Cytogenetics methods (Froenicke et al., Stanyon et al.)
- Minimal rearrangement scenarios (Bourque et al., Pevzner et al.)
- Adjacency analysis (Ma et al., Chauve et al.)

State of the art – Cytogenetics (ZooFISH)

It consists in the hybridization of probes from one chromosome of a reference species, to chromosomes of a target species.

Ancestral genomes are defined as chromosome fragments and associations.
Ex: 2p, 2pq, 16q-19q

If the two species are too divergent (~100 My limit), probes can not hybridize.

Human probes hybridized against chromosomes of the nine-banded armadillo

Svartman et al.

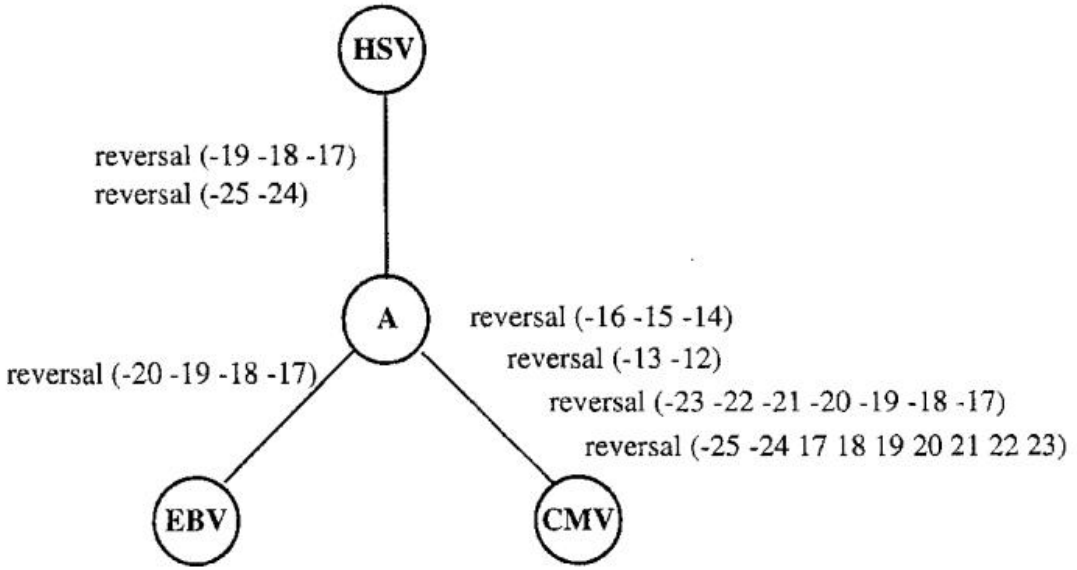
State of the art – Minimal rearrangement scenarios

The goal is to find the minimal set of rearrangements that transform *n* genomes into their last common ancestors.

Authorized rearrangements are a subset of:

inversions (reversals), fusions, fissions, translocations, DCJs

The combinatorial nature of the formulation implies a wide solution space, and current heuristics are too weak to find biologically relevant solutions.



Median genome of three herpes viruses

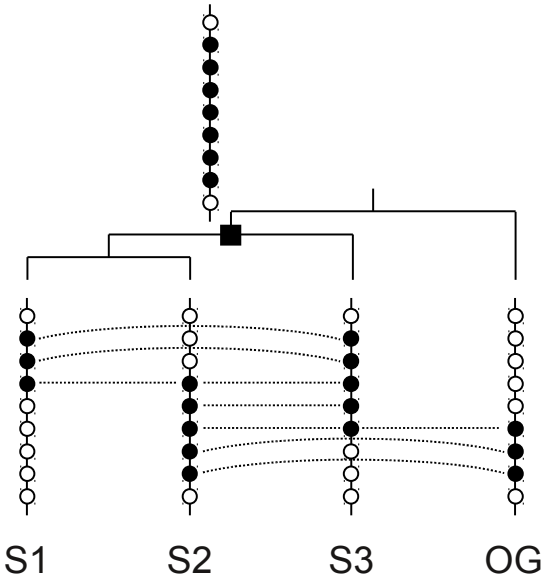
Bourque et al.

State of the art – Adjacency analysis

The brick is here a set of conserved marker adjacencies between extant genomes, from which a set of non-conflicting ancestral adjacencies is extracted,

Ancestral genomes are often composed of chromosome fragments: CARs
Contiguous Ancestral Regions

Current implementations limit to a 1-to-1 mapping between genomes and a single ancestral target.



Definition of an ancestral marker order, based on extant order comparison

Objectives

- Targeting all the ancestors of the vertebrate phylogenetic tree: ~400 Myr
- Handling different content among genomes (duplication, loss)
- Favour specificity over sensitivity (conserved adjacencies)
  i.e. what is reconstructed is reliable

Objectives

- Targeting all the ancestors of the vertebrate phylogenetic tree: ~400 Myr
- Handling different content among genomes (duplication, loss)
- Favour specificity over sensitivity (conserved adjacencies)
  i.e. what is reconstructed is reliable

→ AGORA: Algorithms for Gene Order Reconstruction in Ancestors

AGORA algorithms

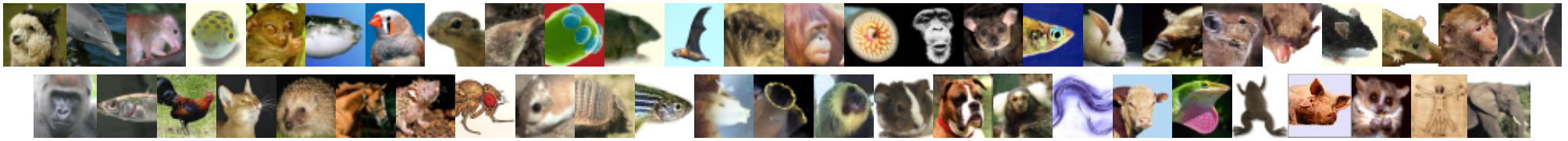
- Ancestral genome content
- Overall presentation of AGORA algorithms
- Gene order reconstruction
- Genomicus
- Comparison to reference *Boreoeutheria*
- Validation by simulations

AGORA algorithms

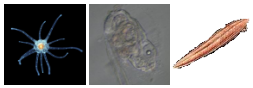
- Ancestral genome content
- Overall presentation of AGORA algorithms
- Gene order reconstruction
- Genomicus
- Comparison to reference *Boreoeutheria*
- Validation by simulations

Extant genomes and gene annotation

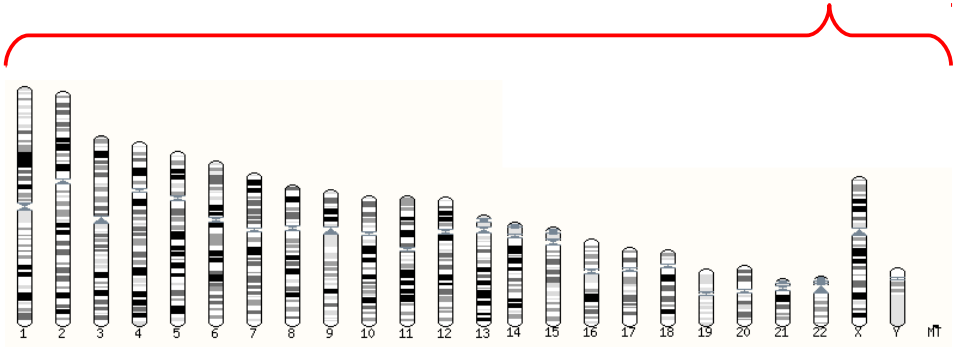
- 50 genomes in the Ensembl database

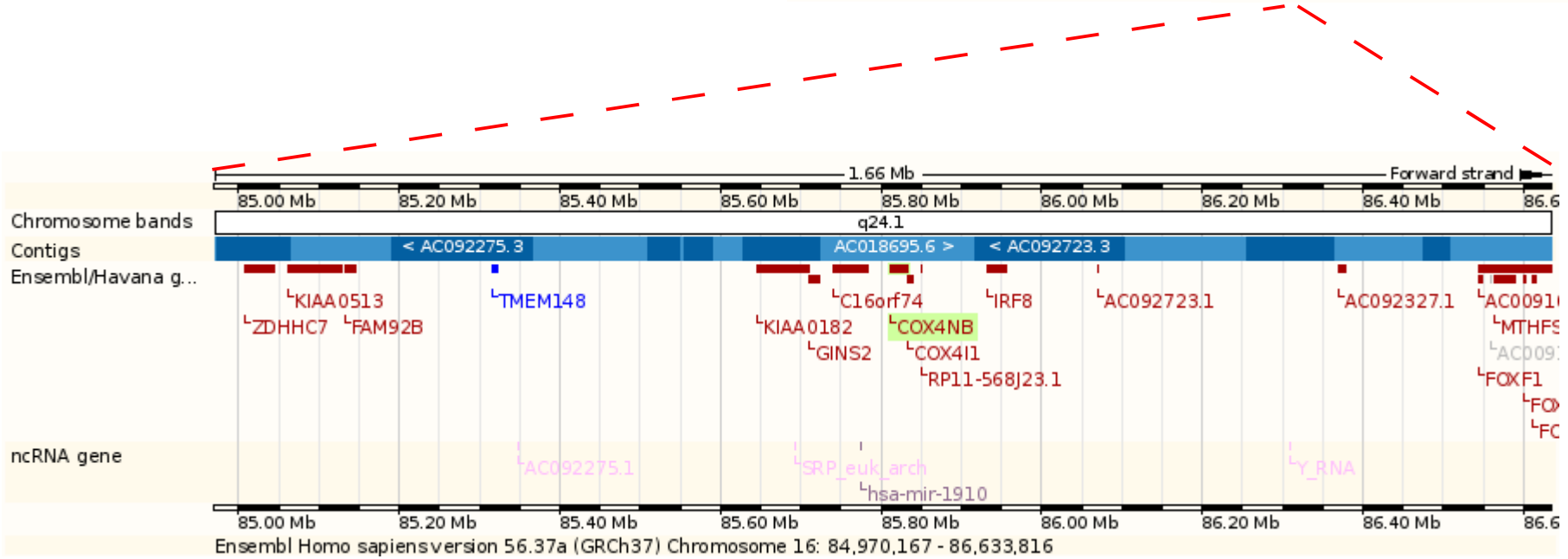


- 3 species from other sources



→ Total: 53 genomes / 961 225 genes





Markers for genome comparison

“Markers” are objects shared by several genomes.
They are used to compare their structure.

Exemples of markers used in reconstructions :

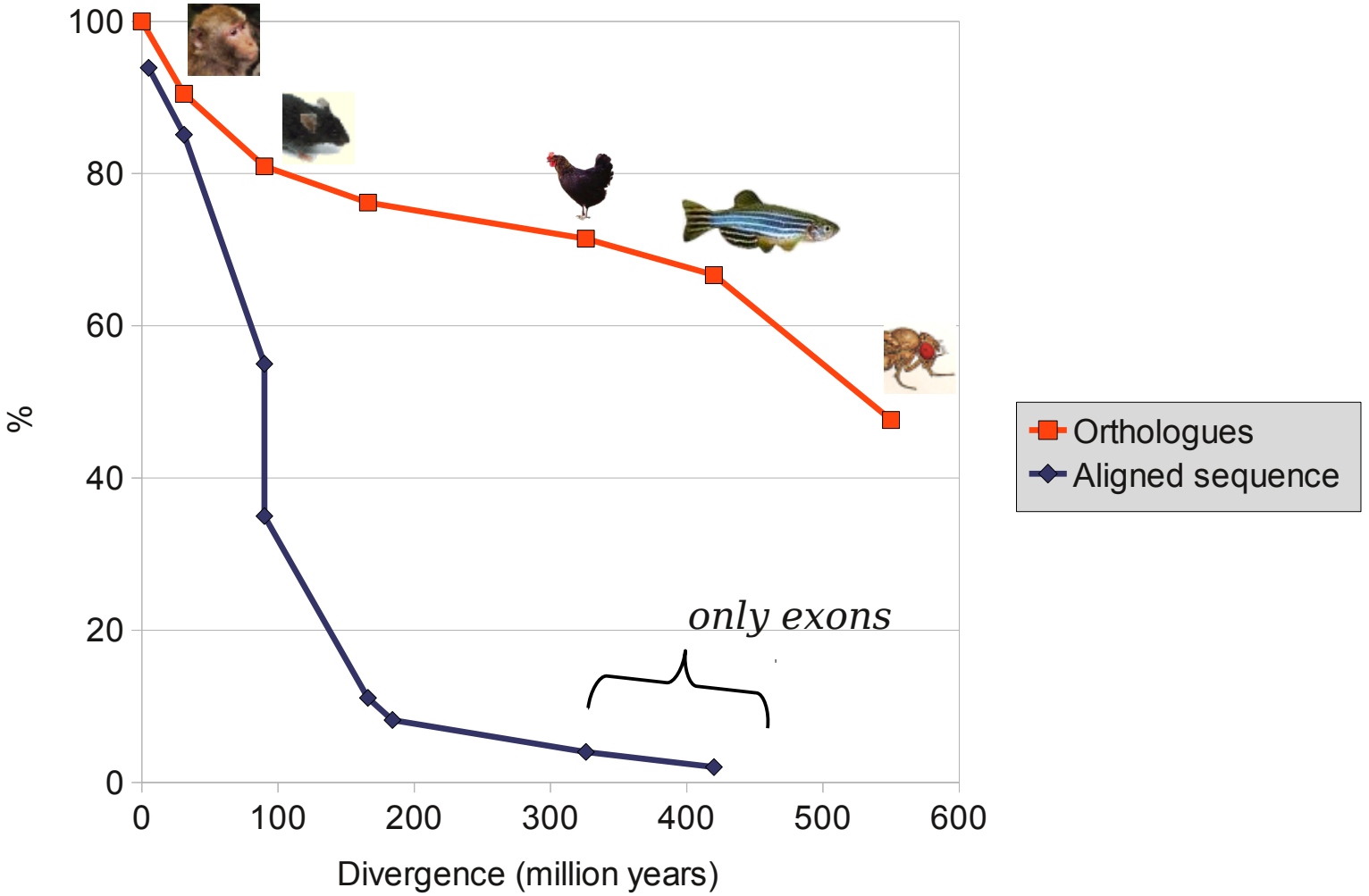
1) Large blocks of nucleic sequence (multiple alignments)

2) Proteic sequence (homologous genes)

What is the best marker for ancestral genome reconstruction ?

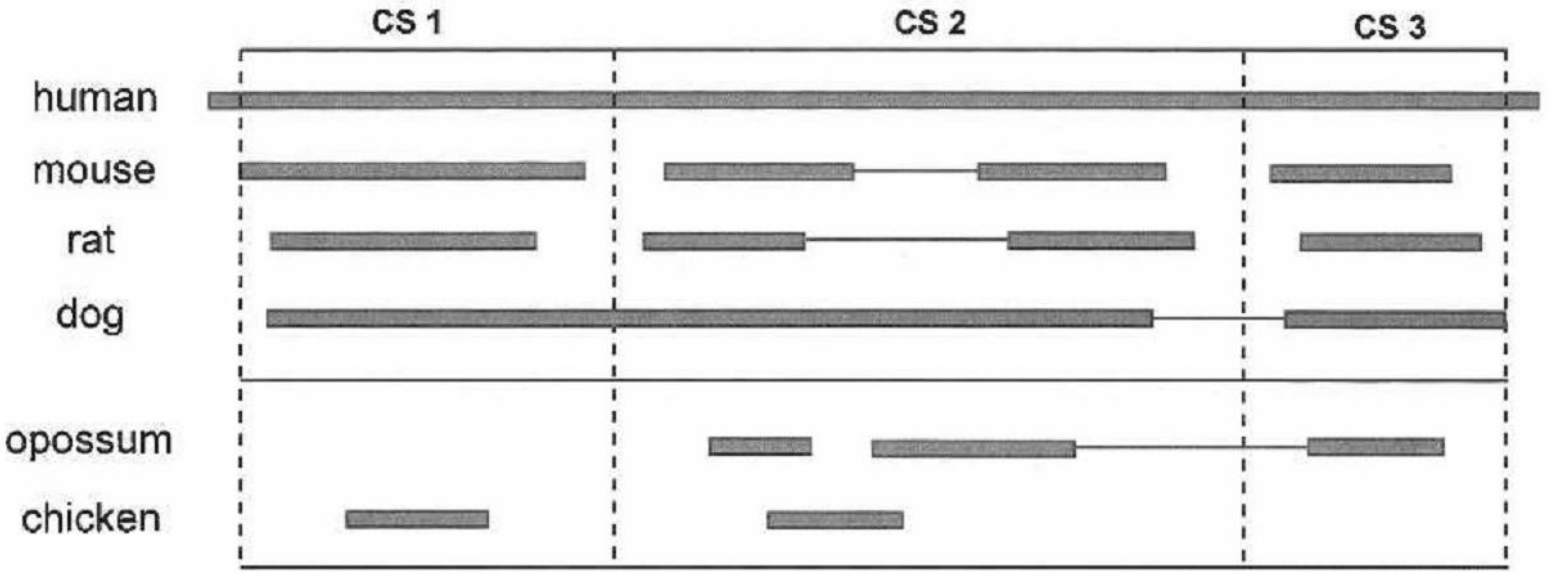
Markers – Example of significant difference

Between tetrapods and fish, only exons can be aligned, whereas nearly 2/3 of genes still have an ortholog.
Homology can be inferred with even older divergence points : *nematode, yeast*



Markers – Multiple alignments

Vertebrate-wide multiple alignments are available in public databases.
ex: UCSC 28-Way vertebrate alignment, Miller et al.

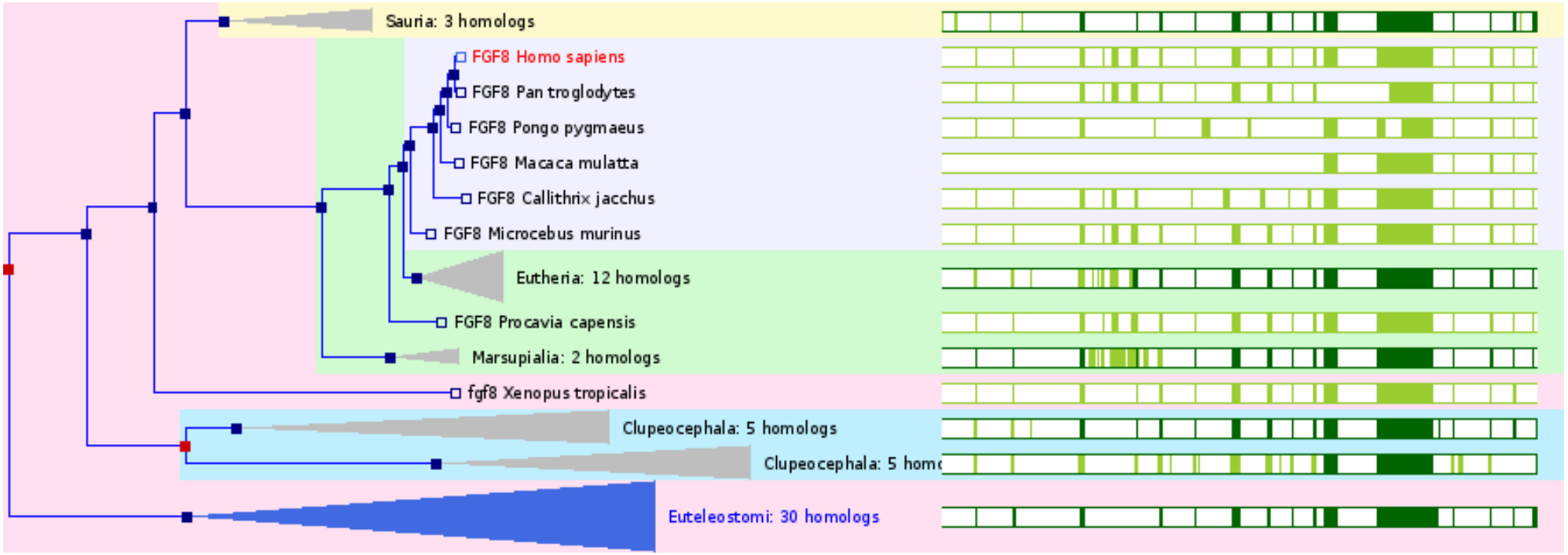


+
- High resolution
- Good continuity along genome
- Sensitive to small rearrangements

-
- Limited to species that can be aligned
- Hard to resolve duplications
- Requires reference genome

Markers – Phylogenetic trees

Ensembl computes systematic phylogenetic trees for all the genes.



+

- Wide phylogenetic range
- Handles duplications well

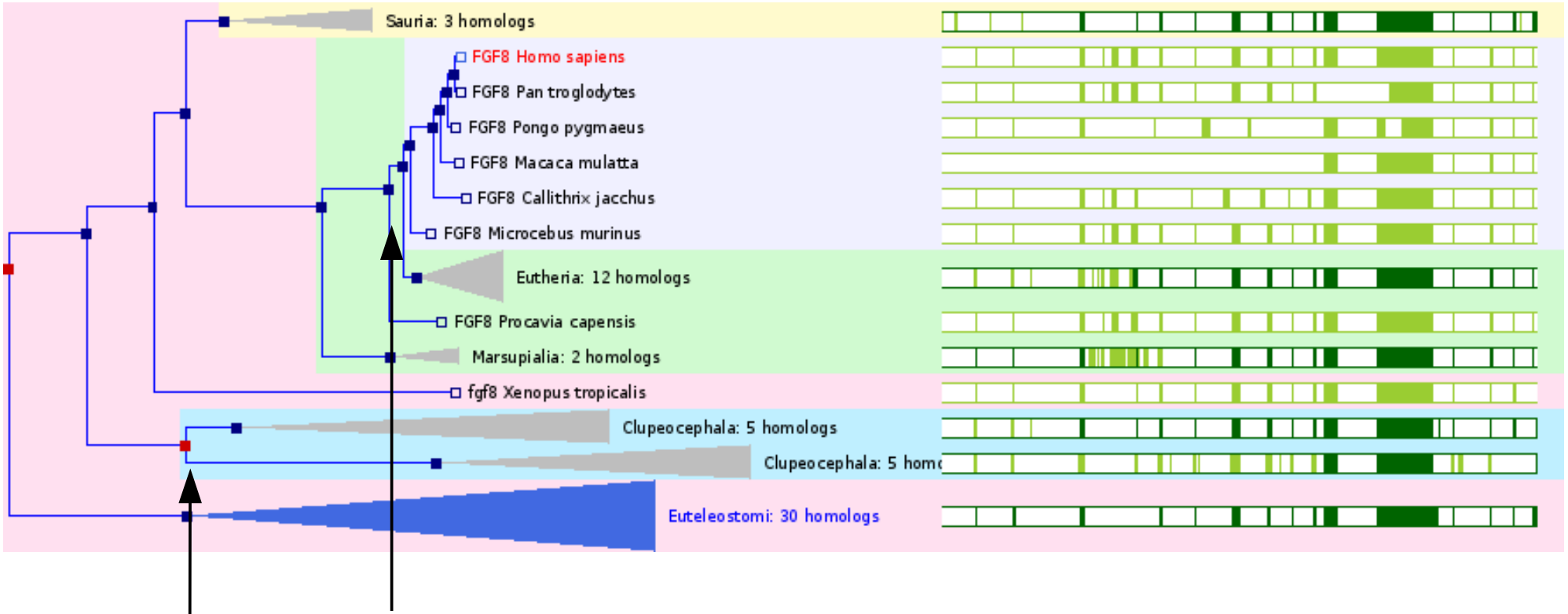
−

- Insensitive to intergenic & intronic rearrangements
- Dependent on gene annotations
- Dependent on tree constructions

Phylogenetic trees and gene history

Phylogenetic trees are used in AGORA to define:

- Gene content at all ancestral nodes
- Orthology / Paralogy relationships



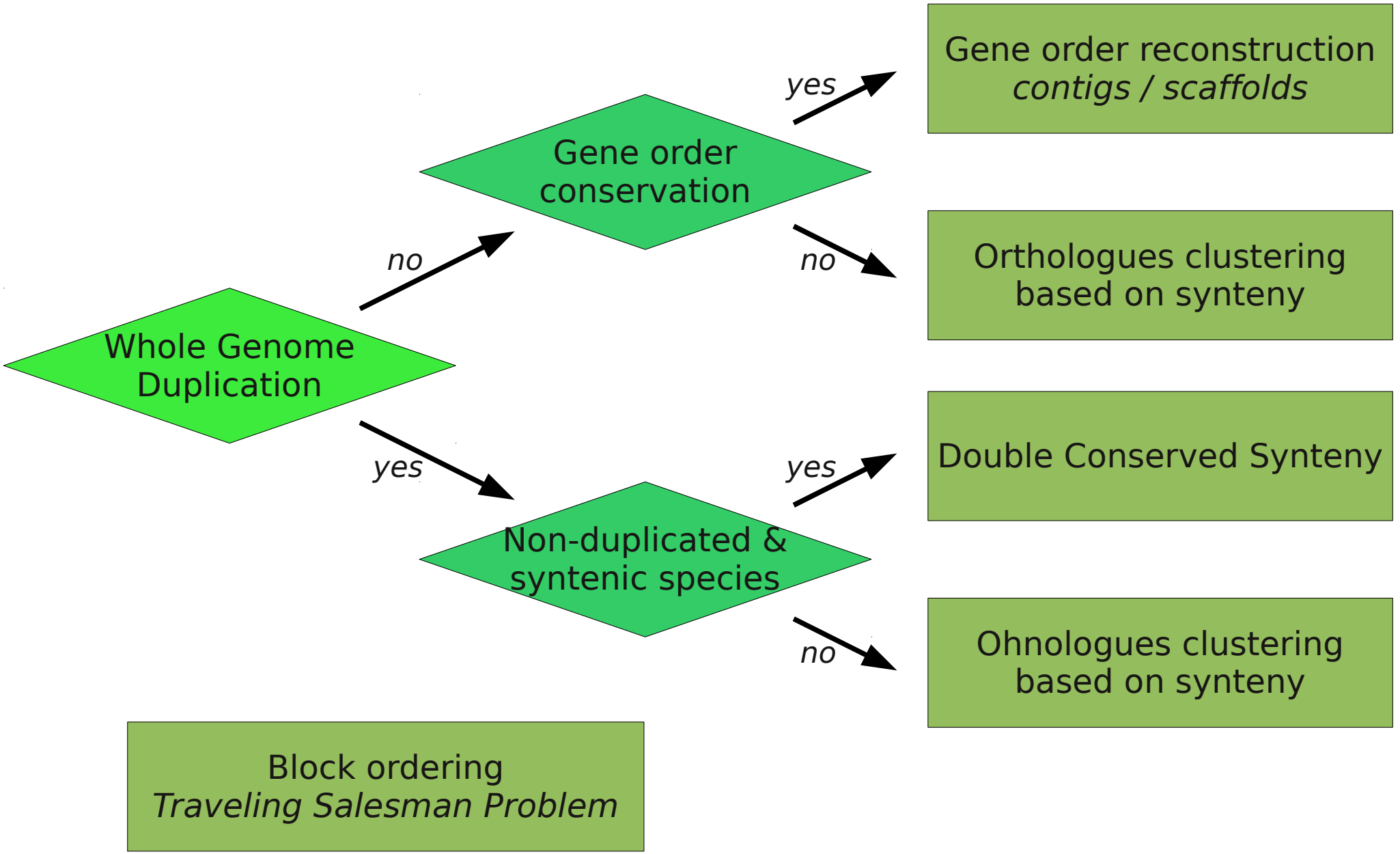
Speciation → 1 copy in the ancestral genome, orthologous pairs

Duplication → 2 copies in the ancestral genome, paralogous pairs

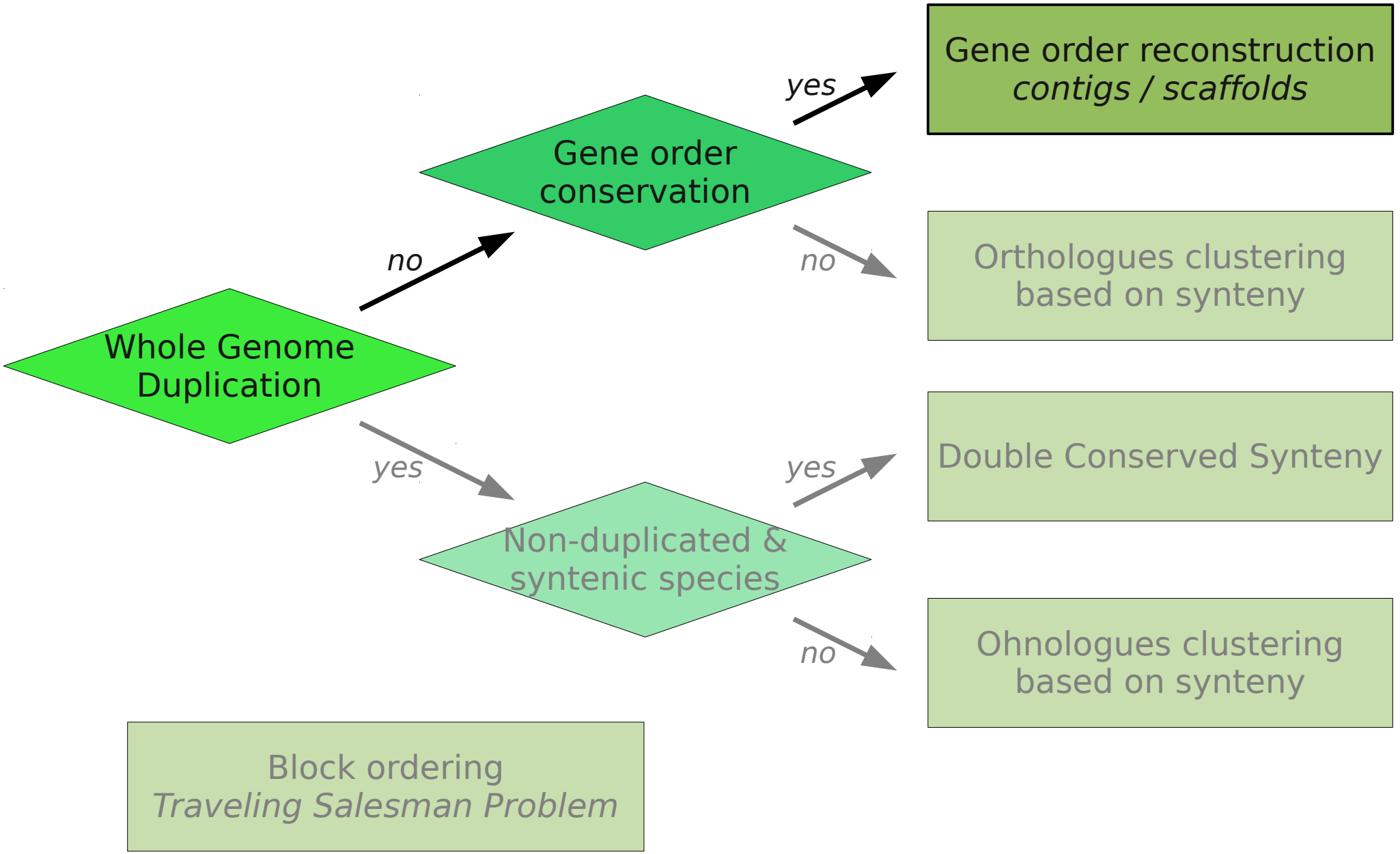
AGORA algorithms

- Ancestral genome content
- Overall presentation of AGORA algorithms
- Gene order reconstruction
- Genomicus
- Comparison to reference *Boreoeutheria*
- Validation by simulations

Algorithms for Gene Order Reconstruction in Ancestors



Algorithms for Gene Order Reconstruction in Ancestors

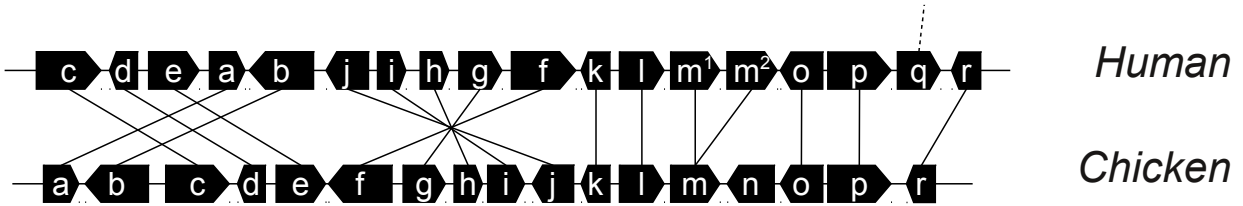


AGORA algorithms

- Ancestral genome content
- Overall presentation of AGORA algorithms
- Gene order reconstruction
- Genomicus
- Comparison to reference *Boreoeutheria*
- Validation by simulations

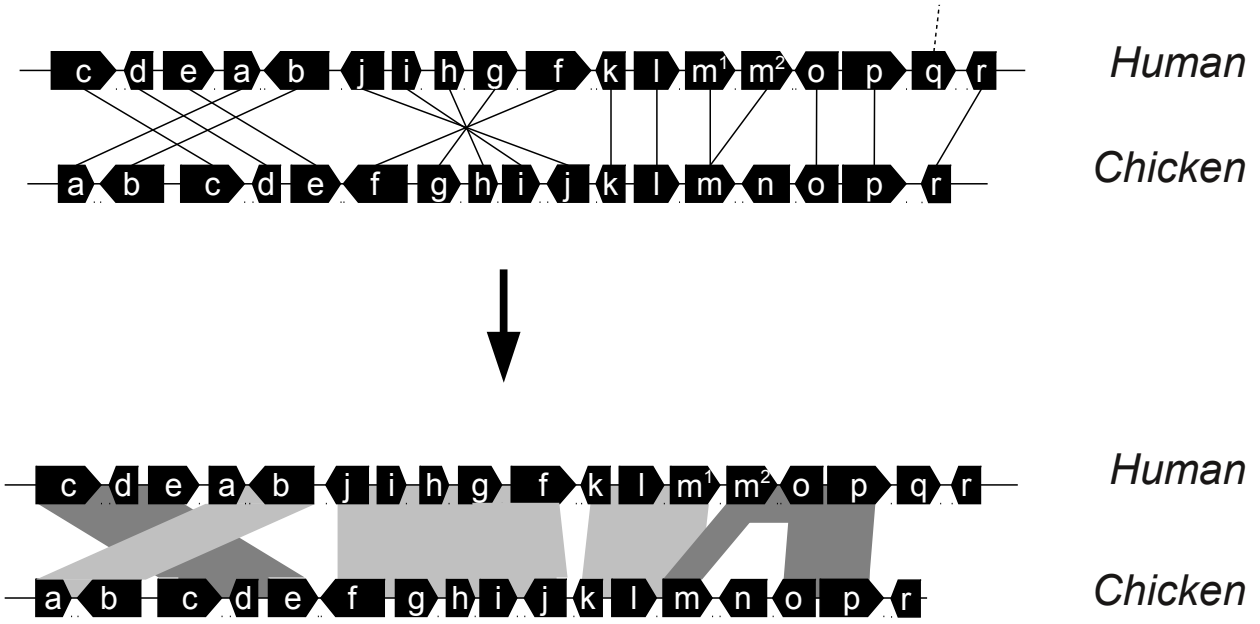
Comparison of two genomes

We search segments of conserved gene order & orientation



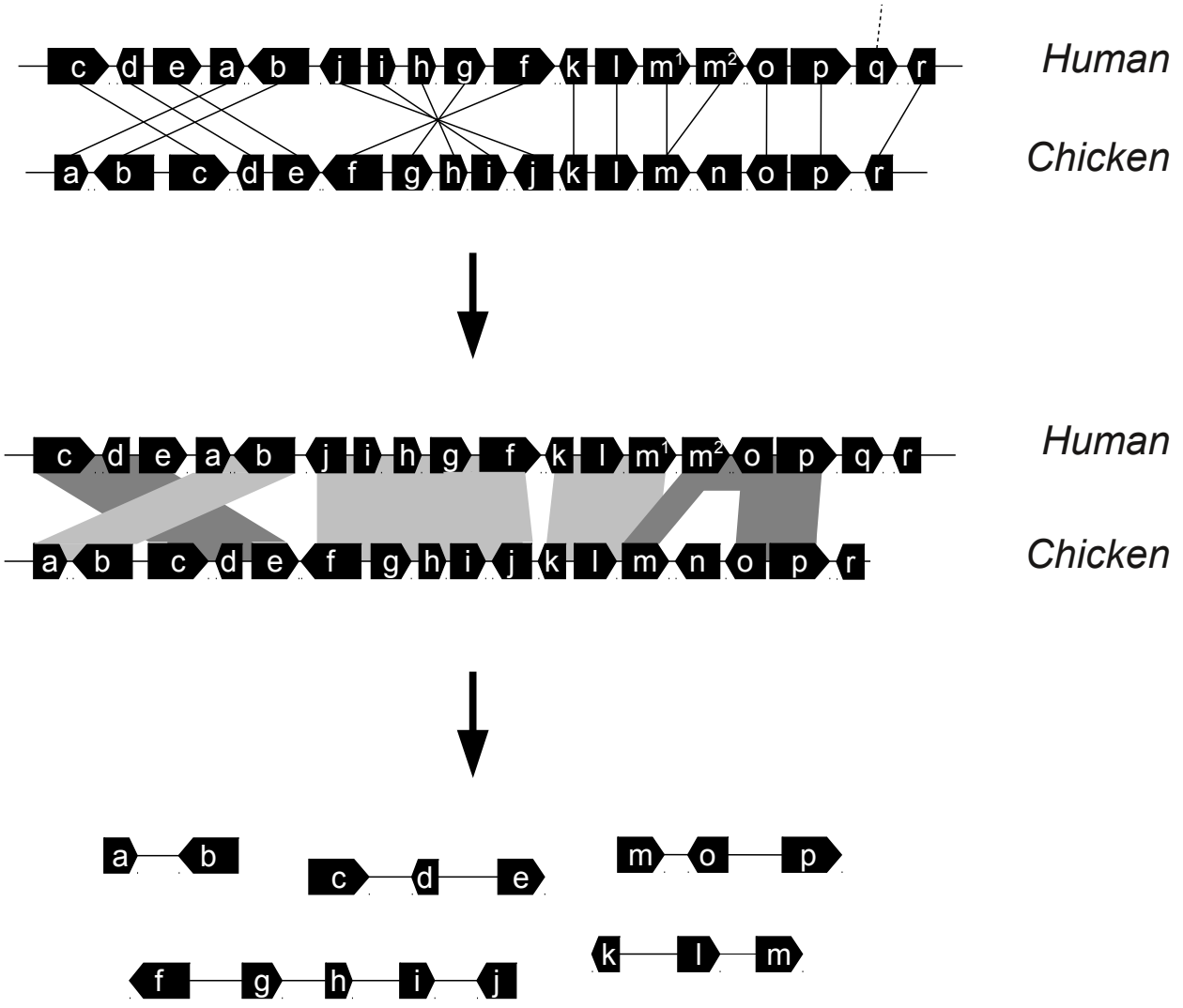
Comparison of two genomes

We search segments of conserved gene order & orientation



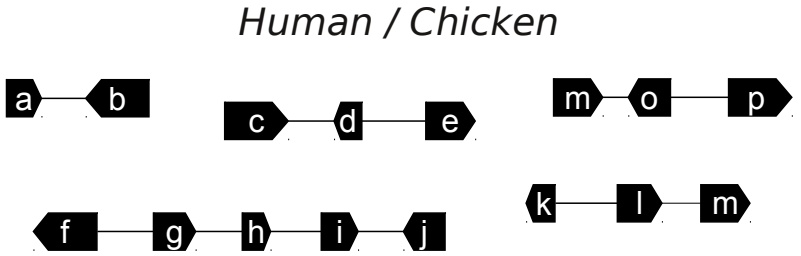
Comparison of two genomes

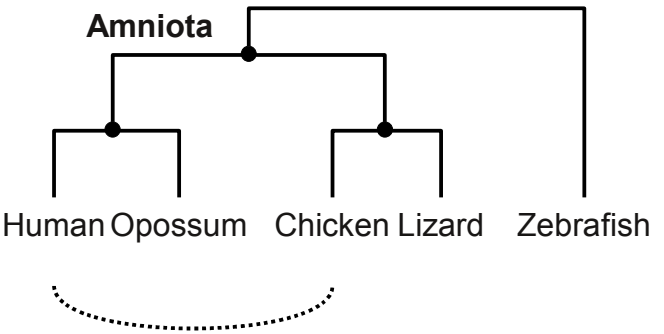
We search segments of conserved gene order & orientation



Comparison of all the genomes

We compare all the informative pair of species

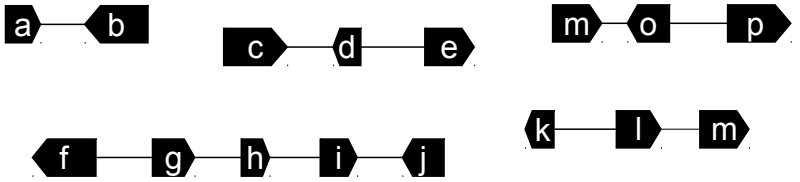




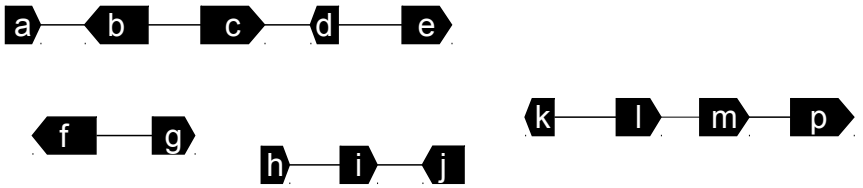
Comparison of all the genomes

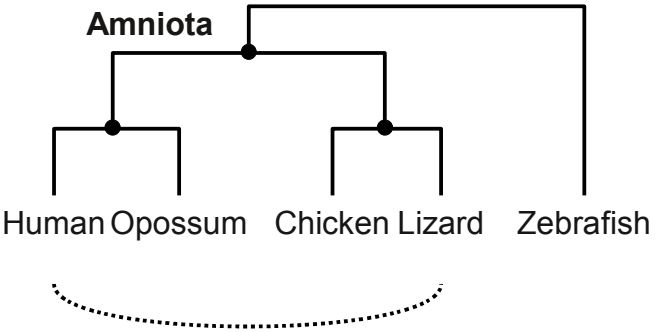
We compare all the informative pair of species

Human / Chicken



Human / Lizard

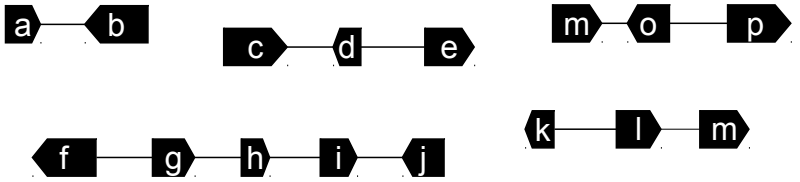




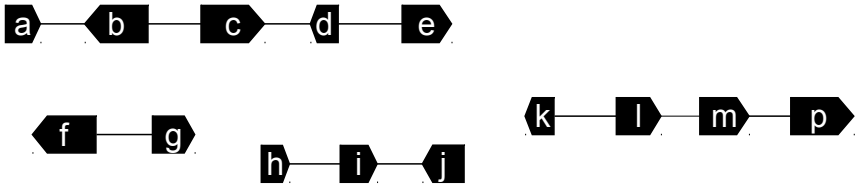
Comparison of all the genomes

We compare all the informative pair of species

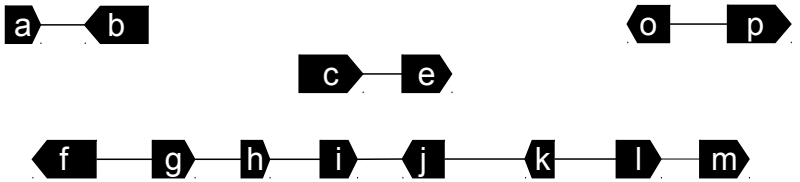
Human / Chicken

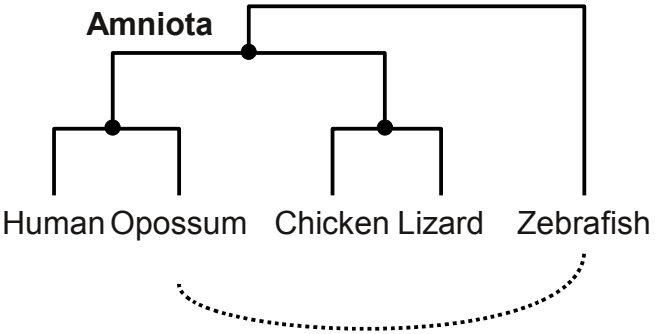


Human / Lizard



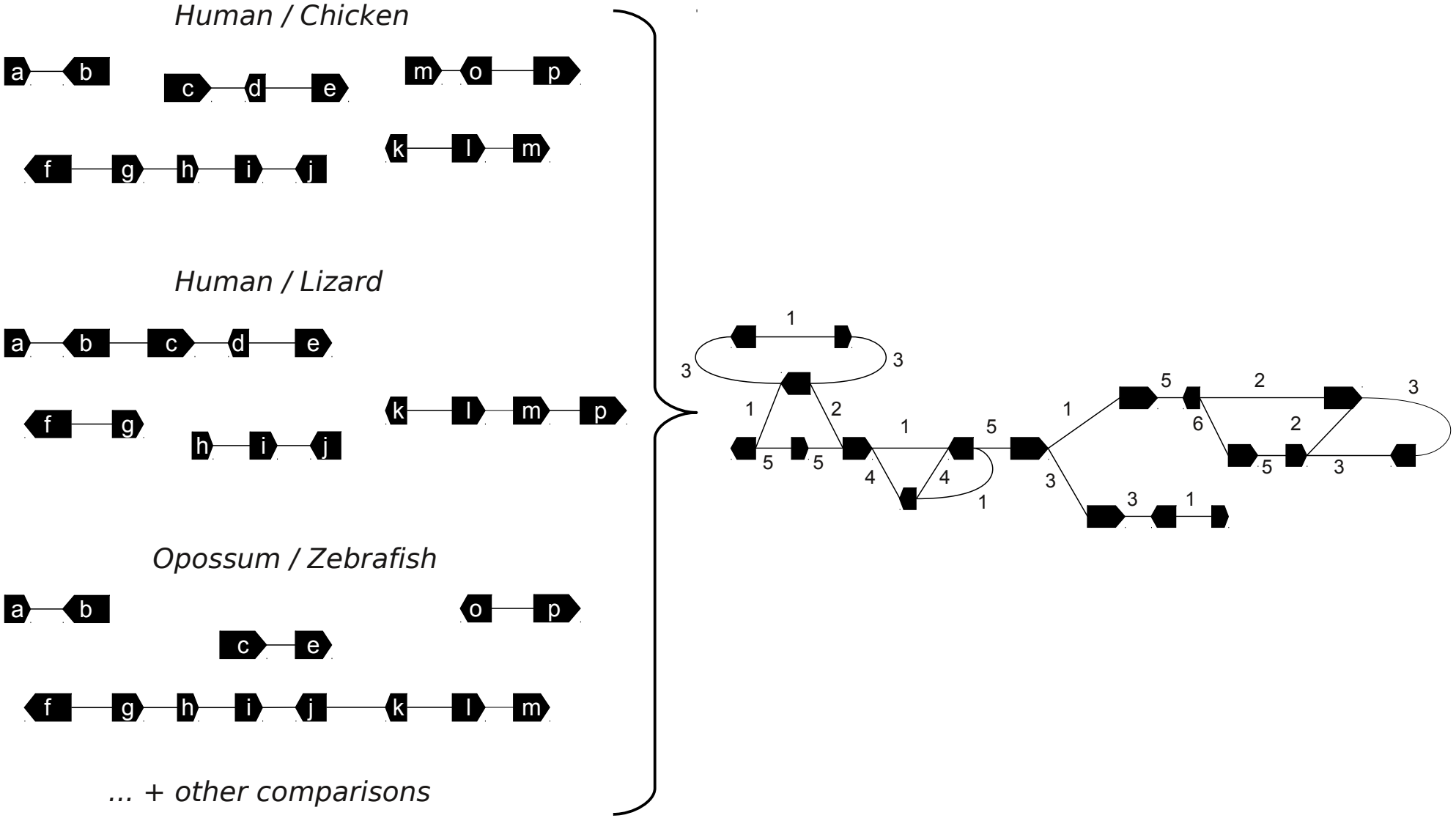
Opossum / Zebrafish





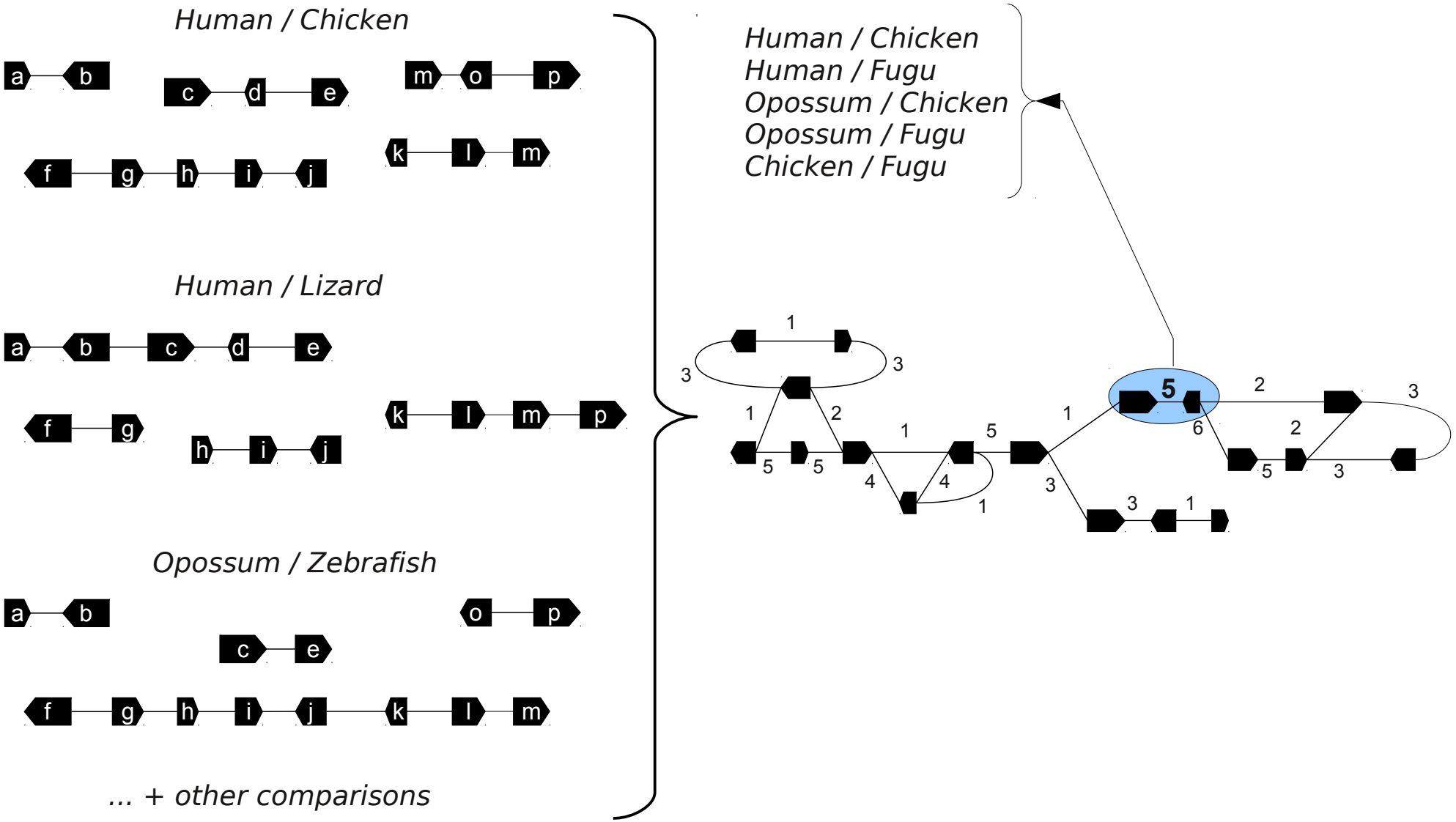
Integration of the comparisons

For each ancestor, data from informative comparisons are merged into a graph



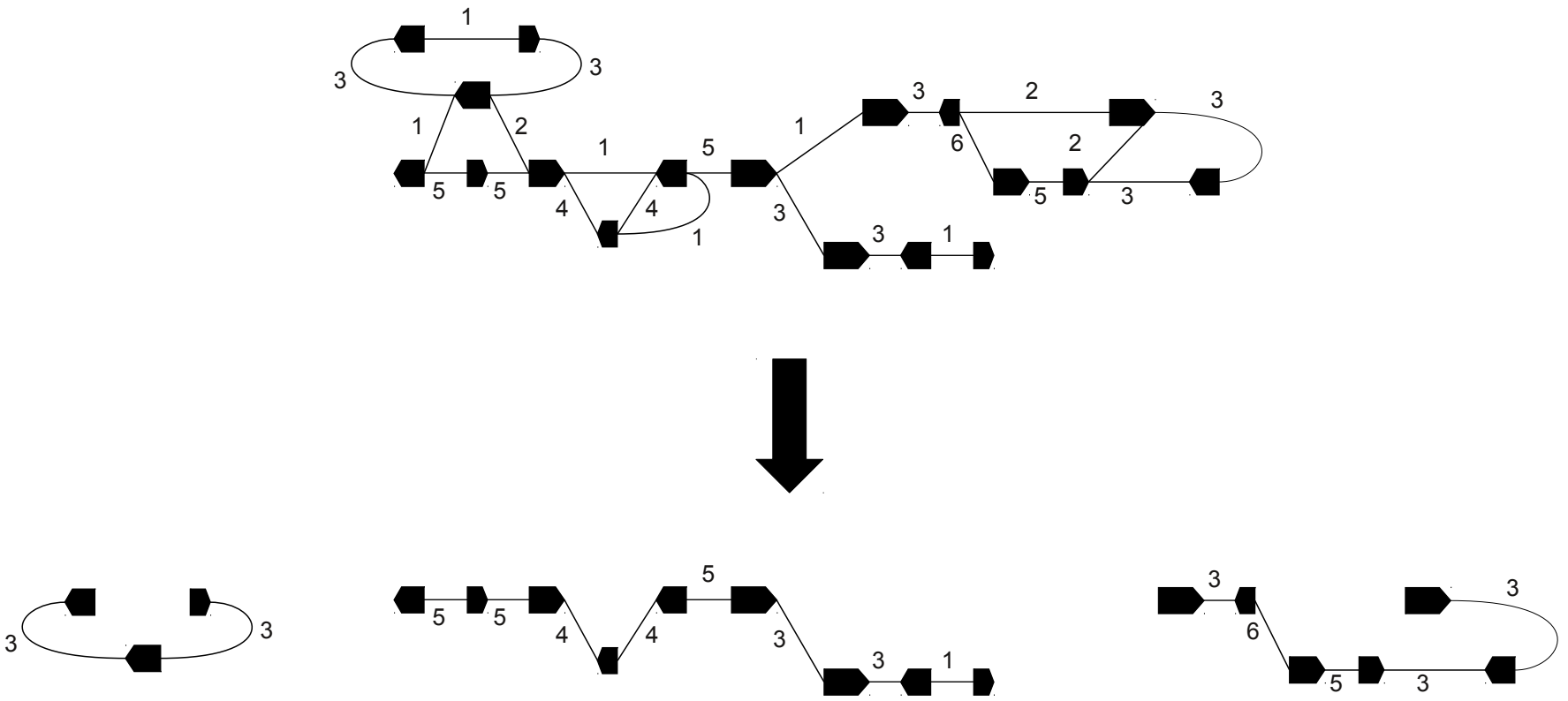
Integration of the comparisons

For each ancestor, data from informative comparisons are merged into a graph



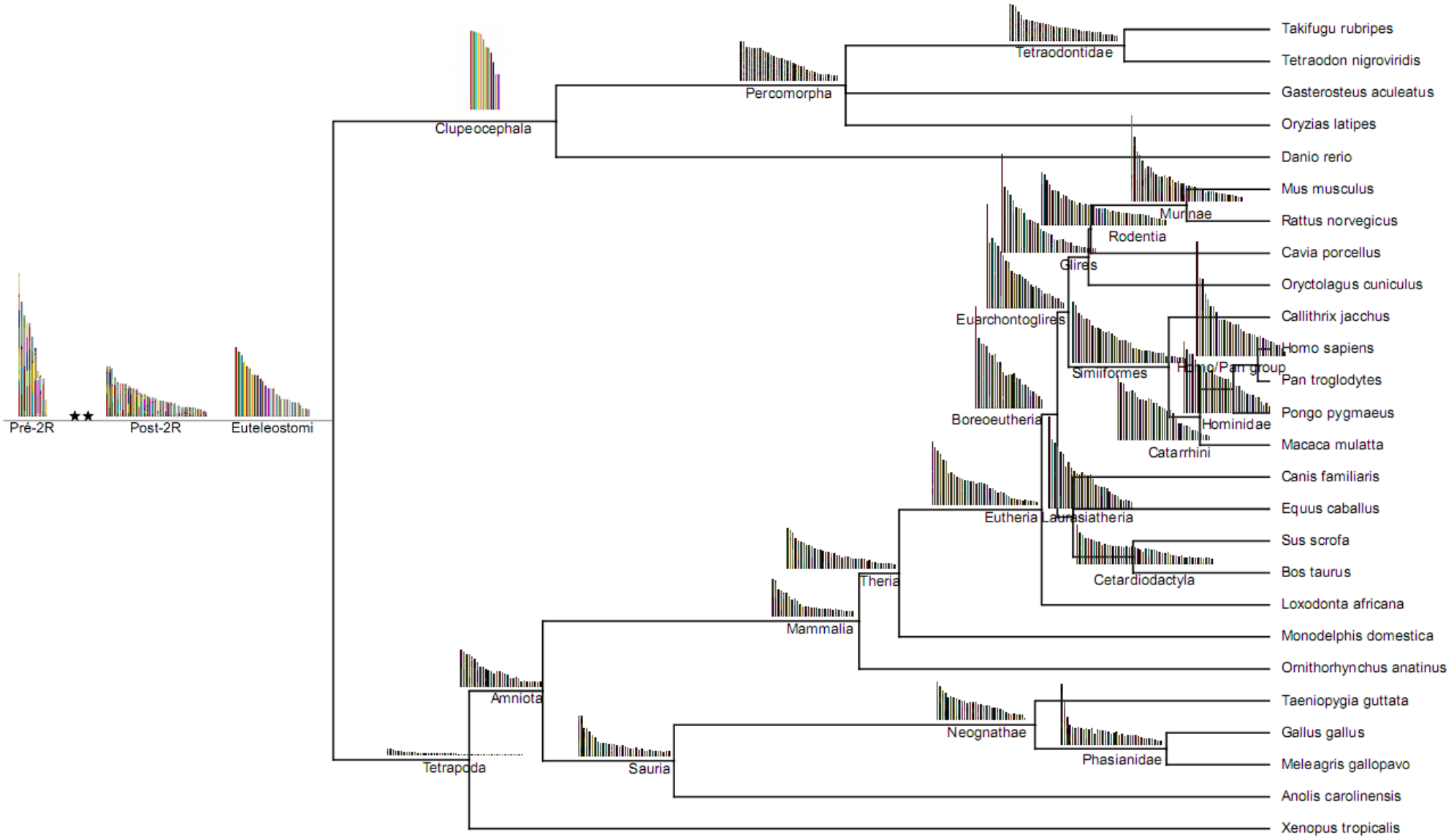
Reconstruction of ancestral gene order

Vertebrate chromosomes are « independant » linear molecules.
AGORA algorithms aim at extracting sets of paths from the adjacency graph.



A typical AGORA algorithm tests edges, by decreasing weight, and selects those that form a coherent set of paths.

Reconstructed ancestral genomes



AGORA algorithms

- Ancestral genome content
- Overall presentation of AGORA algorithms
- Gene order reconstruction
- Genomicus
- Comparison to reference *Boreoeutheria*
- Validation by simulations

Genomicus

A web server (http://www.dyogen.ens.fr/genomicus) provides an eficient tool to gene order comparison (including the ancestral reconstructions).

Genomicus
Genomes in evolution

DYOGEN group

web-code version: 2010-11-26
database version: 60.01

- Help & Documentation
- Examples
- Statistics
- Archives
- Downloads
- Site history

Contact us.

Enter a gene name *(Ensembl nomenclature or approved gene symbol)*
You can restrict the search to one species (ancestral or modern).

-- Select a species --
Default view
Custom view
Go !

Selected examples can be found ***here***

Genomicus is a genome browser that enables users to navigate in genomes in several dimensions: linearly along chromosome axes, transversaly across different species, and chronologicaly along evolutionary time.

Once a query gene has been entered, it is displayed in its genomic context in parallel to the genomic context of all its orthologous and paralogous copies in all the other sequenced metazoan genomes. Moreover, Genomicus stores and displays the predicted ancestral genome structure in all the ancestral species within the phylogenetic range of interest.

All the data on extant species displayed in this browser are from Ensembl, JGI, and Genoscope.

Summary statistics of Genomicus version 60.01:

Number of extant species	54
Number of extant genes	963564
Number of ancestral species	45
Number of ancestral genes	791594
Number of ancestral synteny blocks	31429

What's new in version 60.01 ?

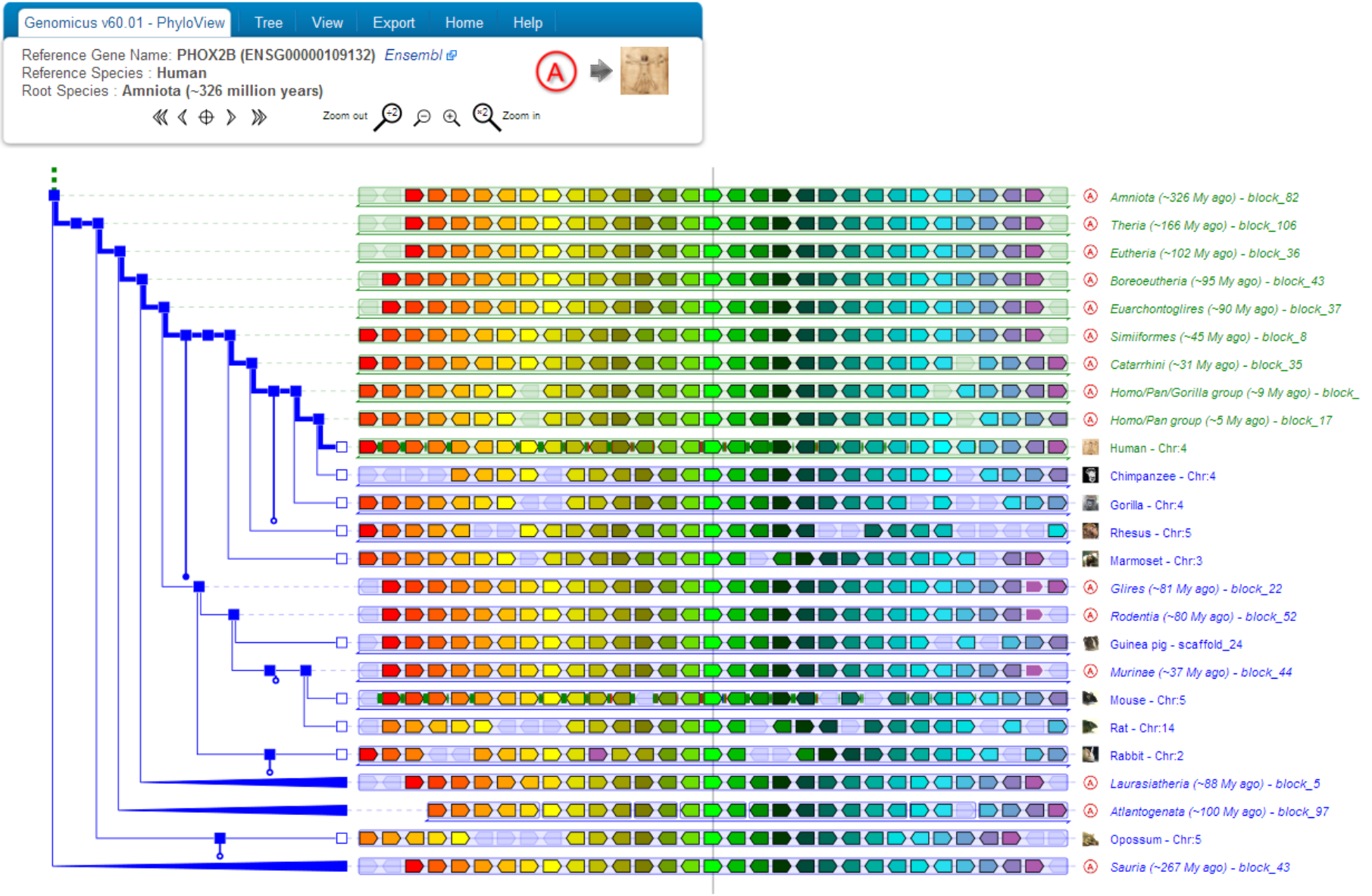
- Database update (Ensembl 60)
- Improved ancestral gene order reconstructions
- Web interface rework & bugfixes

Genomicus — database version: 60.01 / Web-code version: 2010-11-26 — Dyogen Team

Muffato et al., Bioinformatics, 2010

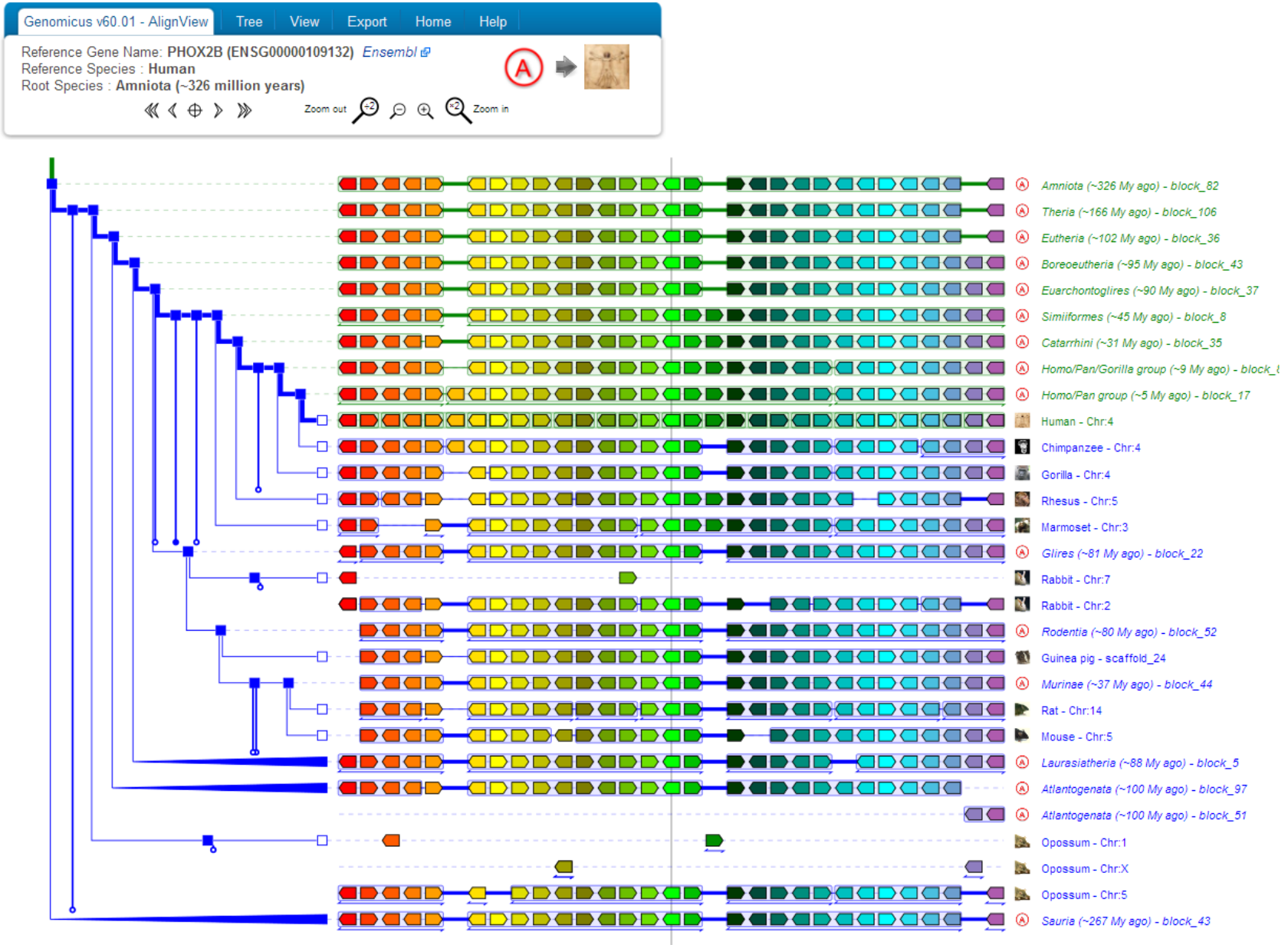
Genomicus – *PhyloView*

PhyloView compares gene order in both sides of a reference gene.



Genomicus – *AlignView*

AlignView aligns a reference genome to other ones.



Genomicus – Usage statistics

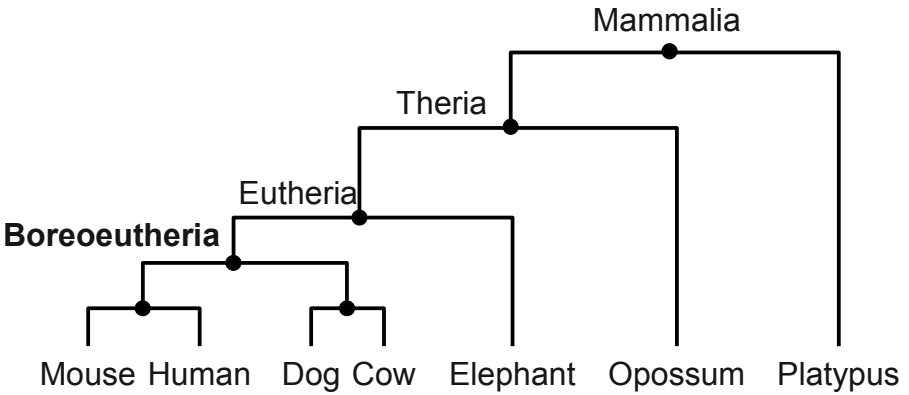
Month	Unique visitors	Visits	Hits	Bandwith
Oct 2009	734	1299	17334	1.92 Gb
Nov 2009	803	1320	12525	1.55 Gb
Déc 2009	629	974	7916	1.17 Gb
Jan 2010	750	1186	8633	3.29 Gb
Fév 2010	794	1207	20601	4.76 Gb
Mar 2010	1421	2577	27383	13.45 Gb
Avr 2010	967	1657	22767	8.02 Gb
Mai 2010	773	1314	20931	4.22 Gb
Juin 2010	654	1180	15638	4.18 Gb
Juil 2010	587	1134	15959	3.68 Gb
Aoû 2010	573	1405	14381	3.24 Gb
Sep 2010	592	1241	16225	5.52 Gb

Country	Nb pages	Bandwith
Europe	67336	4.25 Gb
USA	53103	4.59 Gb
France	33698	4.87 Gb
Norway	10036	2.16 Gb
Japan	6896	2.43 Gb
United Kingdom	4345	1.10 Gb
Singapor	4081	0.67 Gb
Canada	3808	1.04 Gb
China	3034	0.98 Gb
Germany	2484	1.00 Gb
Portugal	2327	0.78 Gb

AGORA algorithms

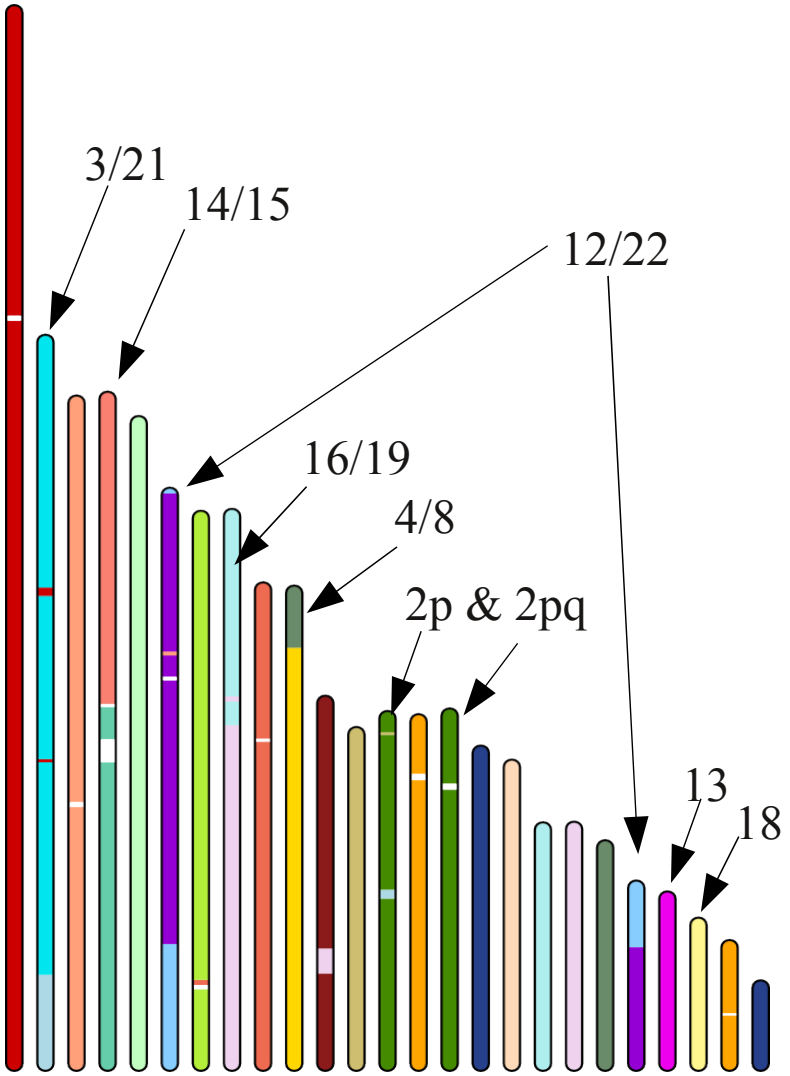
- Ancestral genome content
- Overall presentation of AGORA algorithms
- Gene order reconstruction
- Genomicus
- Comparison to reference *Boreoeutheria*
- Validation by simulations

Boreoeutheria : Comparison between cytogenetics & AGORA

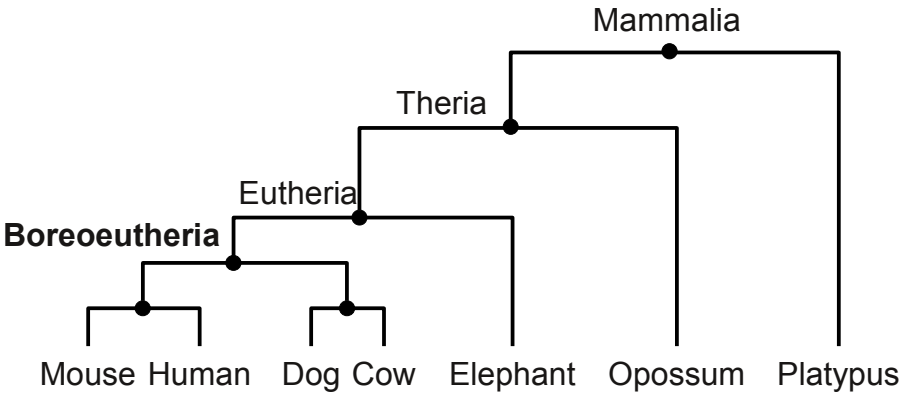


Species phylogenetic tree

Boreoeutheria : Comparison between cytogenetics & AGORA



Karyotype of *Boreoeutheria* reconstructed by AGORA, with a human color code



Species phylogenetic tree

We selected in the AGORA reconstruction the 25 longest CARs (minimum size of 150 genes) for a total gene count of 17827 genes.

The only difference with the cytogenetics reference reconstruction is a missing 7/16 association.

7-16 association

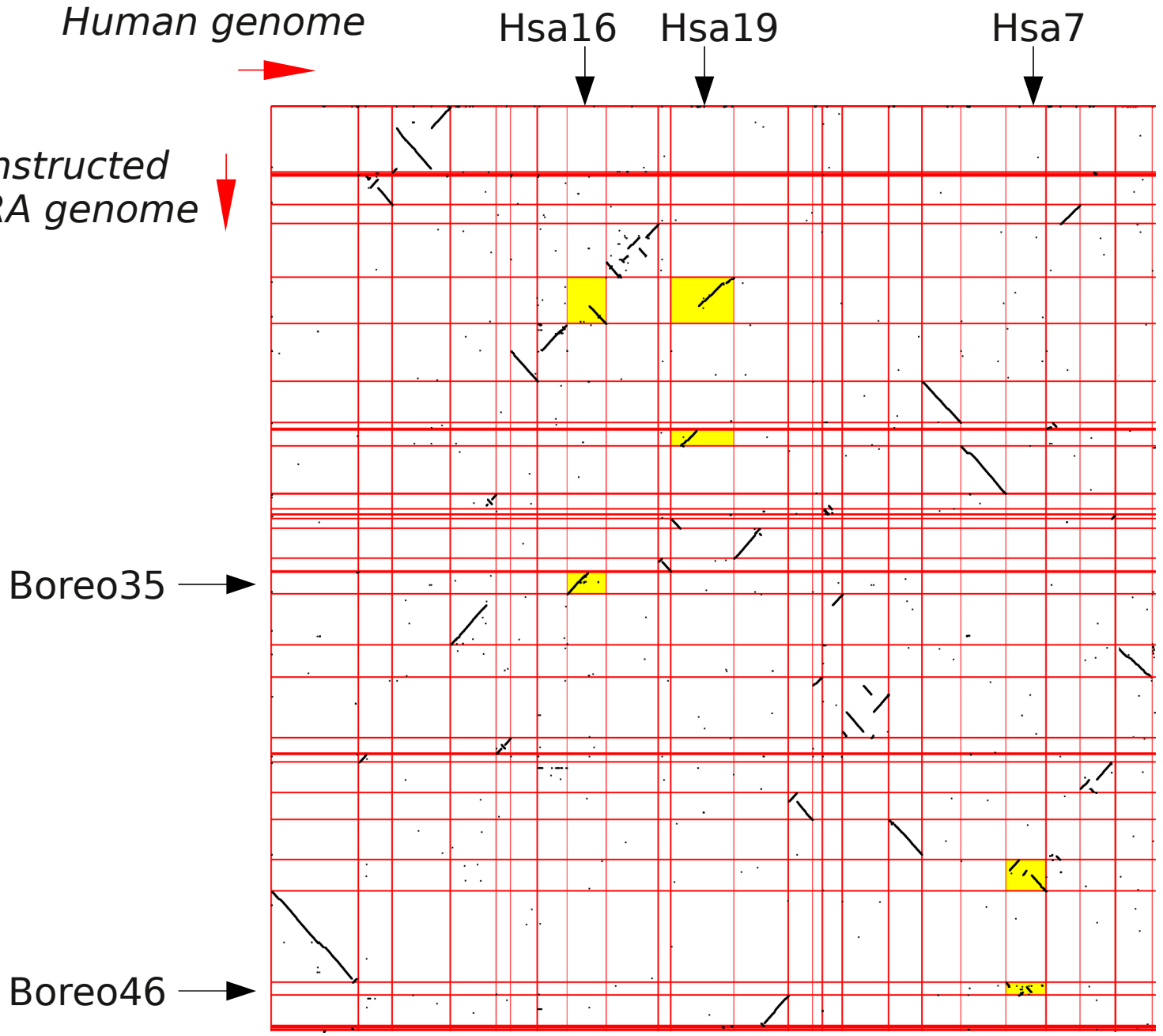
Cytogenetics studies claim that :

- Human chromosomes 7, 16, and 19 were each in two parts
- Parts 7b & 16p were associated in one ancestral chromosome
- Parts 16q & 19q in another one

Do we have this repartition in the reconstructed Boreoeutheria genome ?

Froenicke et al., Ferguson-Smith et al.

7b-16p adjacency – Validation of chromosome parts



7b-16p association – supporting evidence (cytogenetics)

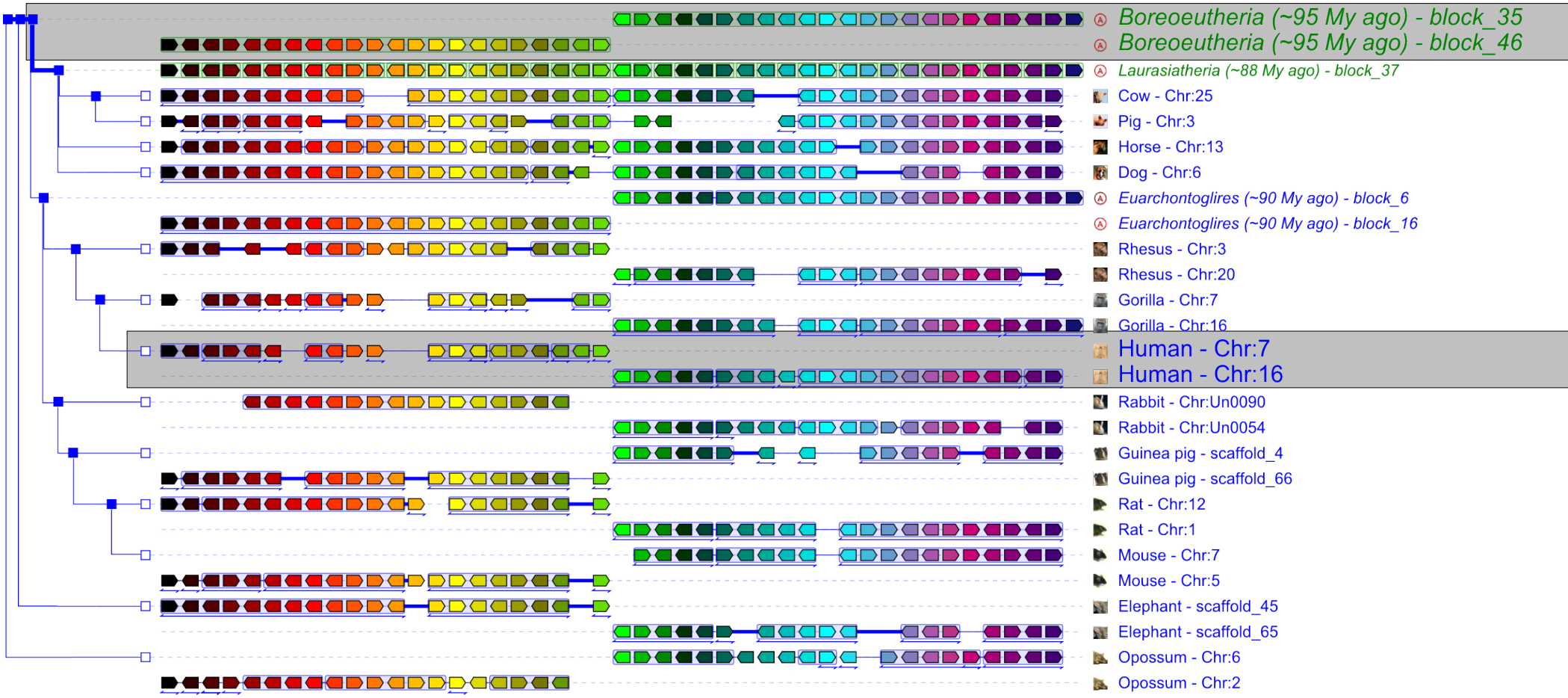
Among sequenced genomes, the 7b-16p association should be supported by :

- *Laurasiatheria :* dog, pig, cow, horse
- *Euarchontoglires :* rabbit, mouse (**by projection*)
- Outgroups : elephant

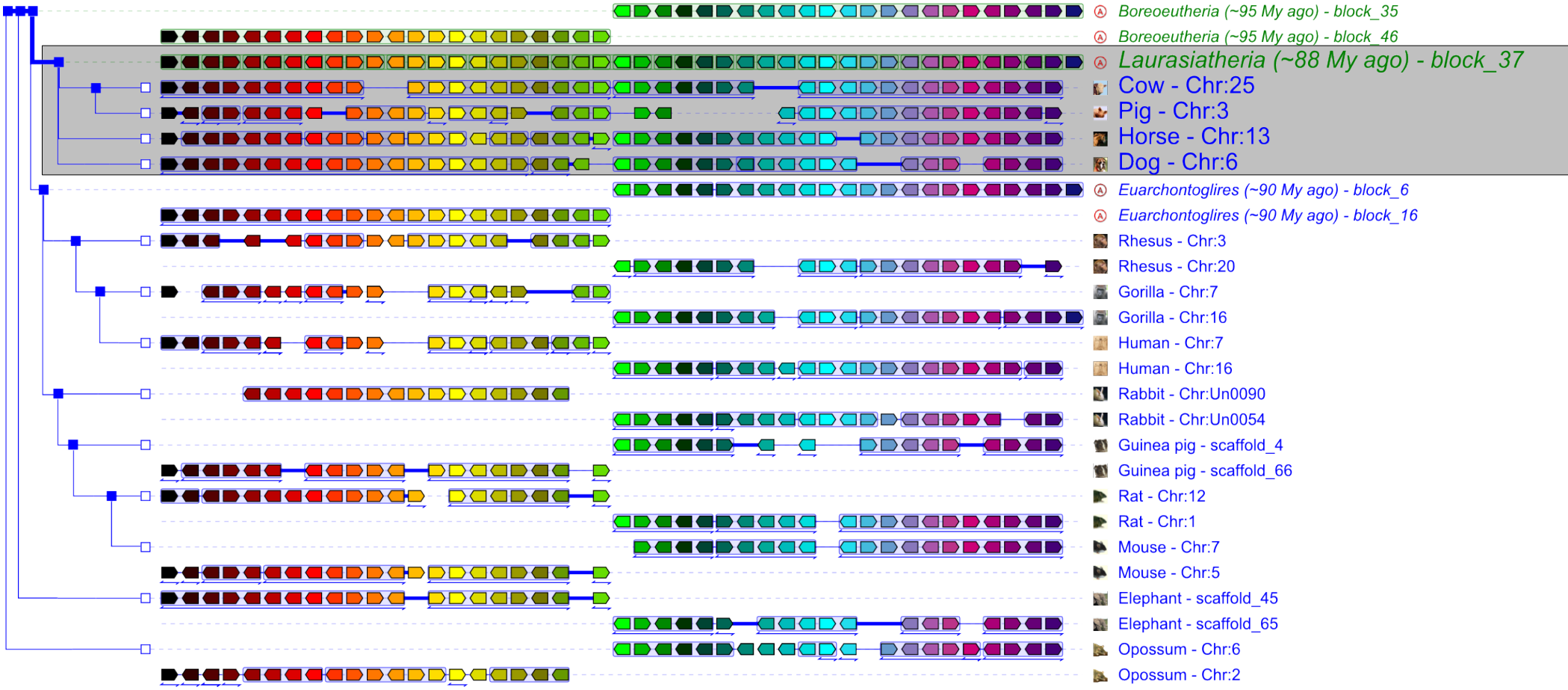
Does the genomic data (sequence, assembly, gene annotation) agree ?

Froenicke et al., Ferguson-Smith et al.

7b-16p adjacency in Genomicus

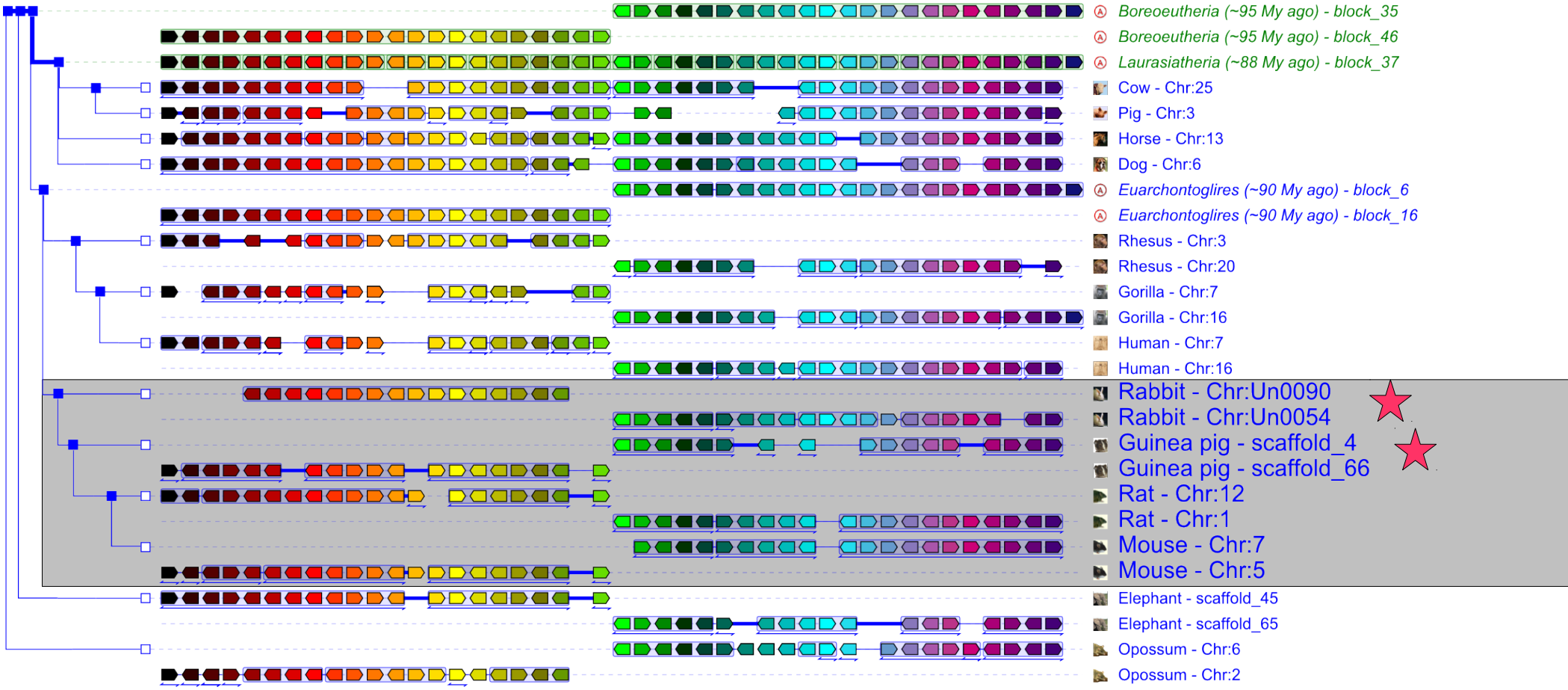


7b-16p adjacency – Presence in *Laurasiatheria*



→ Synteny & adjacency in laurasiatherians, as predicted by cytogenetics

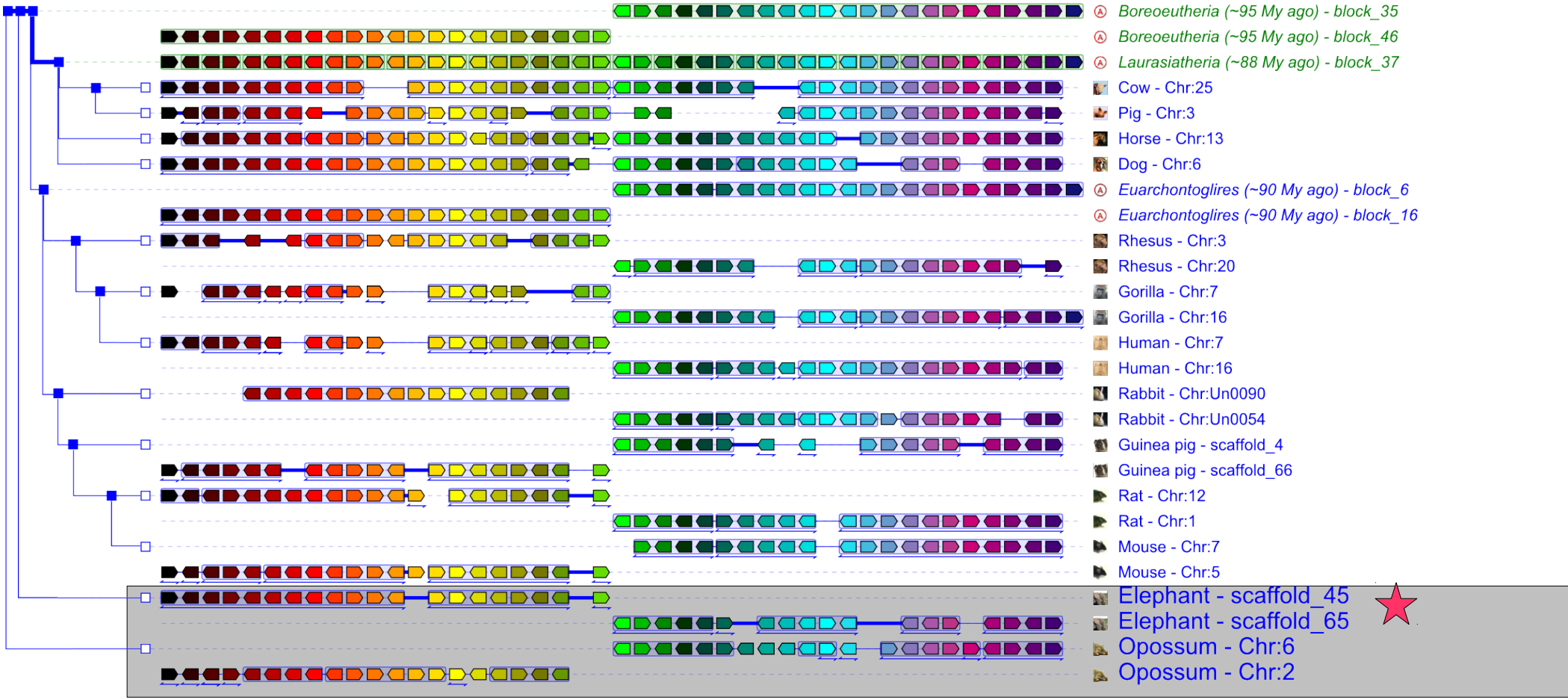
7b-16p adjacency – Absence in *Glires*



→ No synteny in any rodent / lagomorph
(cytogenetics predicts some in mouse & rabbit)

★ Rabbit & guinea pig genomes are not totally assembled

7b-16p adjacency – Absence in outgroups



→ No synteny in opossum / elephant
(cytogenetics predict some in elephant)

★ Elephant genome is not totally assembled

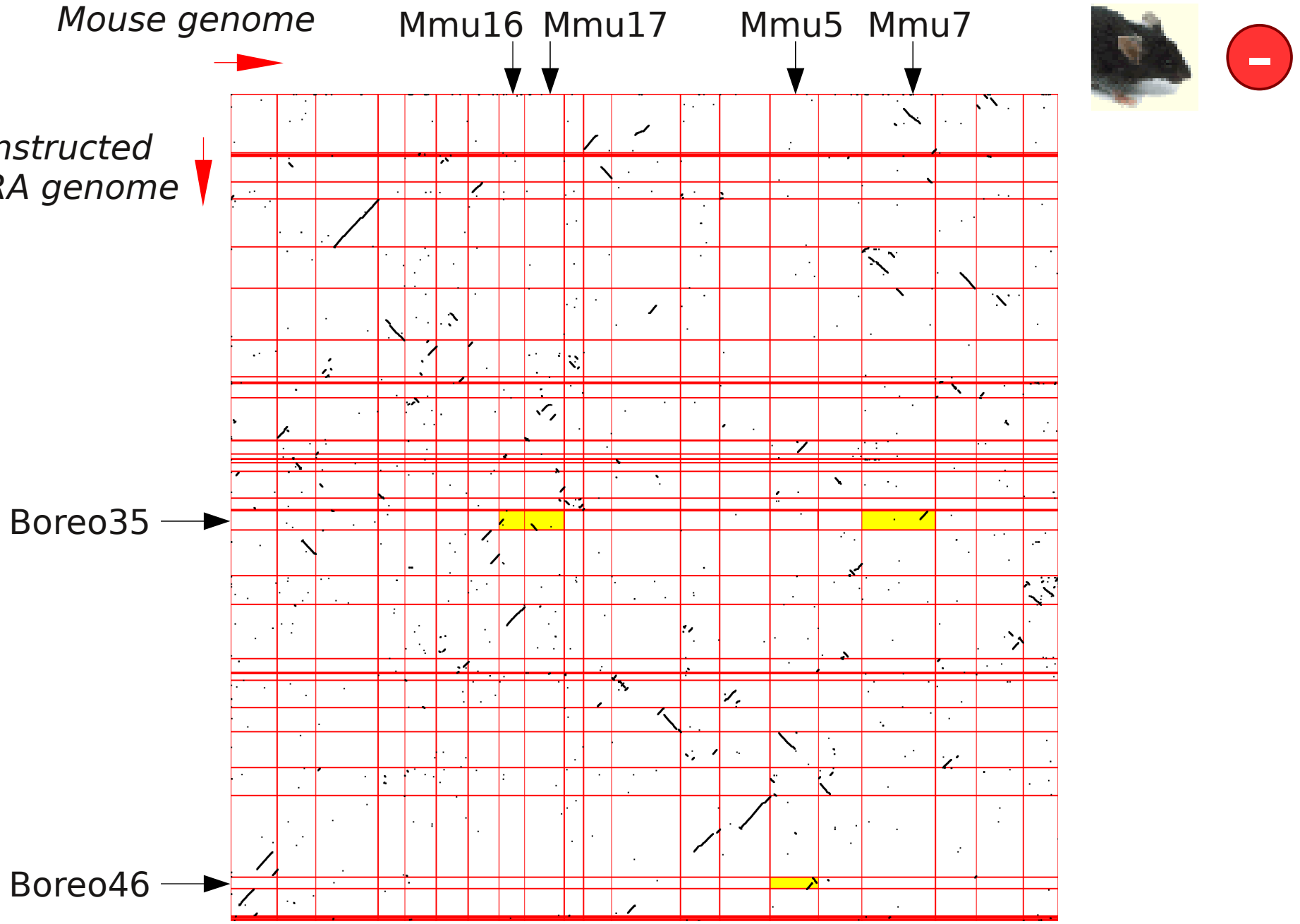
7b-16p adjacency – supporting evidence (AGORA)

In our data, adjacency between 7b & 16p is only seen in laurasiatherians, and therefore only reconstructed in *Laurasiatheria*.

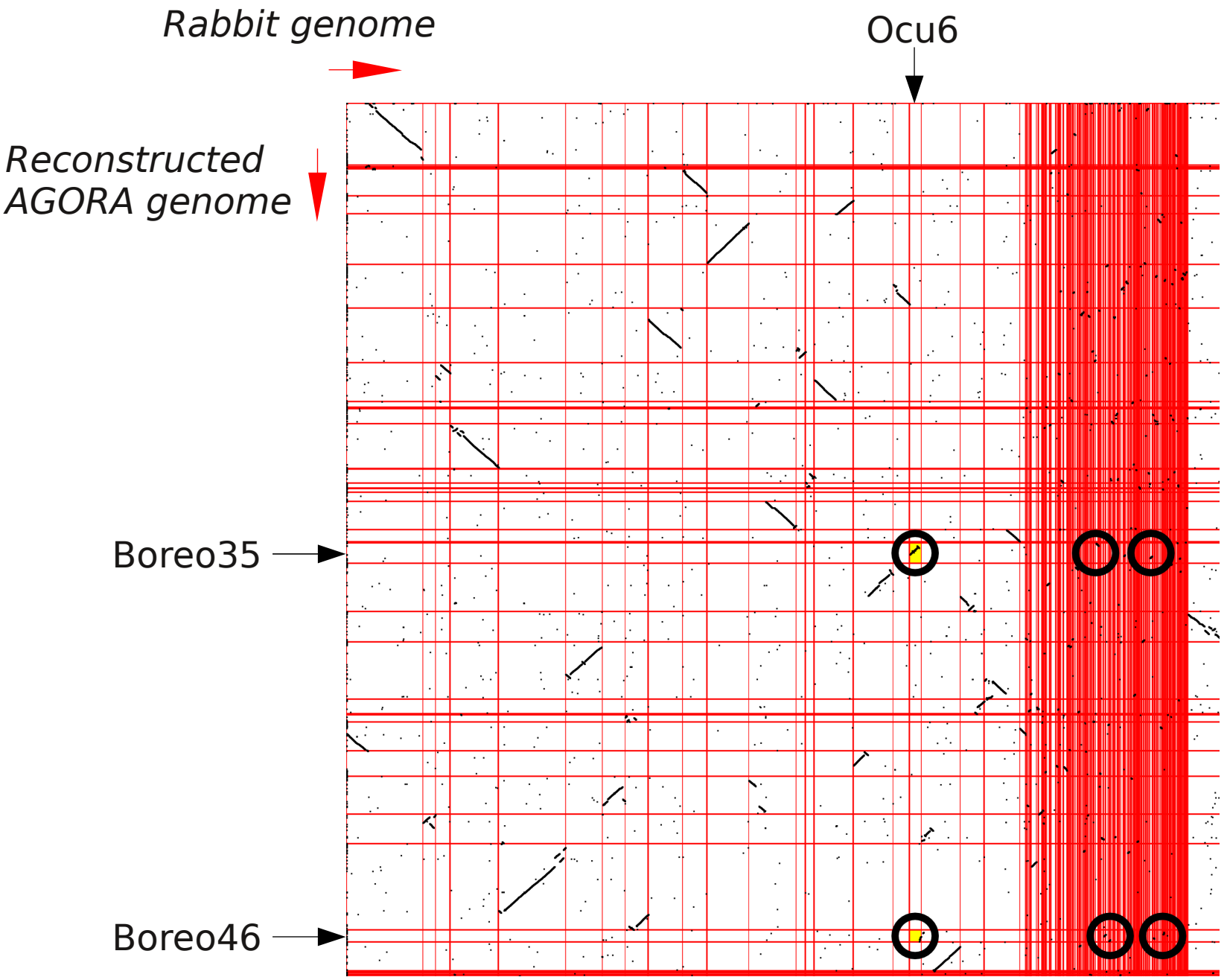
→ As AGORA uses adjacency signal, the association can not be retrieved in *Boreoeutheria*.

Is the 7b-16p association valid ?

7b-16p association – Absence in mouse



7b-16p association – Presence in rabbit



7b-16p association – supporting evidence revisited

In our data, synteny between 7b & 16p exists in :

- Laurasiatherians
- Rabbit (with some unassembled parts)
- Outgroups (opossum, chicken with rearrangements)

Adjacency is seen in :

- Laurasiatherians

→ The 7b-16p association seems correct, but is explained only by a *synteny* signal. Therefore, AGORA gene order algorithms can not reconstruct it.

AGORA algorithms

- Ancestral genome content
- Overall presentation of AGORA algorithms
- Gene order reconstruction
- Genomicus
- Comparison to reference *Boreoeutheria*
- Validation by simulations

Validation of ancestral reconstructions

Reconstructions quality can be assessed by :

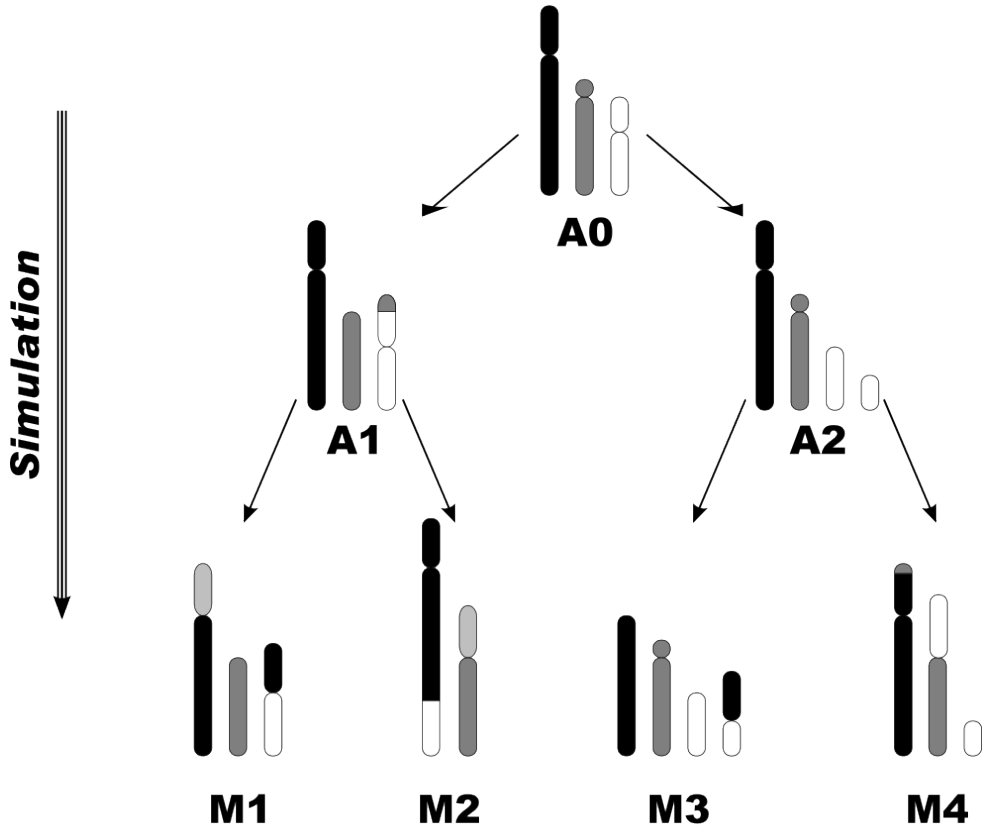
- Comparison with other, independent, reconstructions
  available for some key ancestral species, such as *Boreoeutheria*

- Genome simulations
  no framework available

We have developped a new tool for genome rearrangements simulation:
MagSim: Modern and Ancestral Genome SIMulator

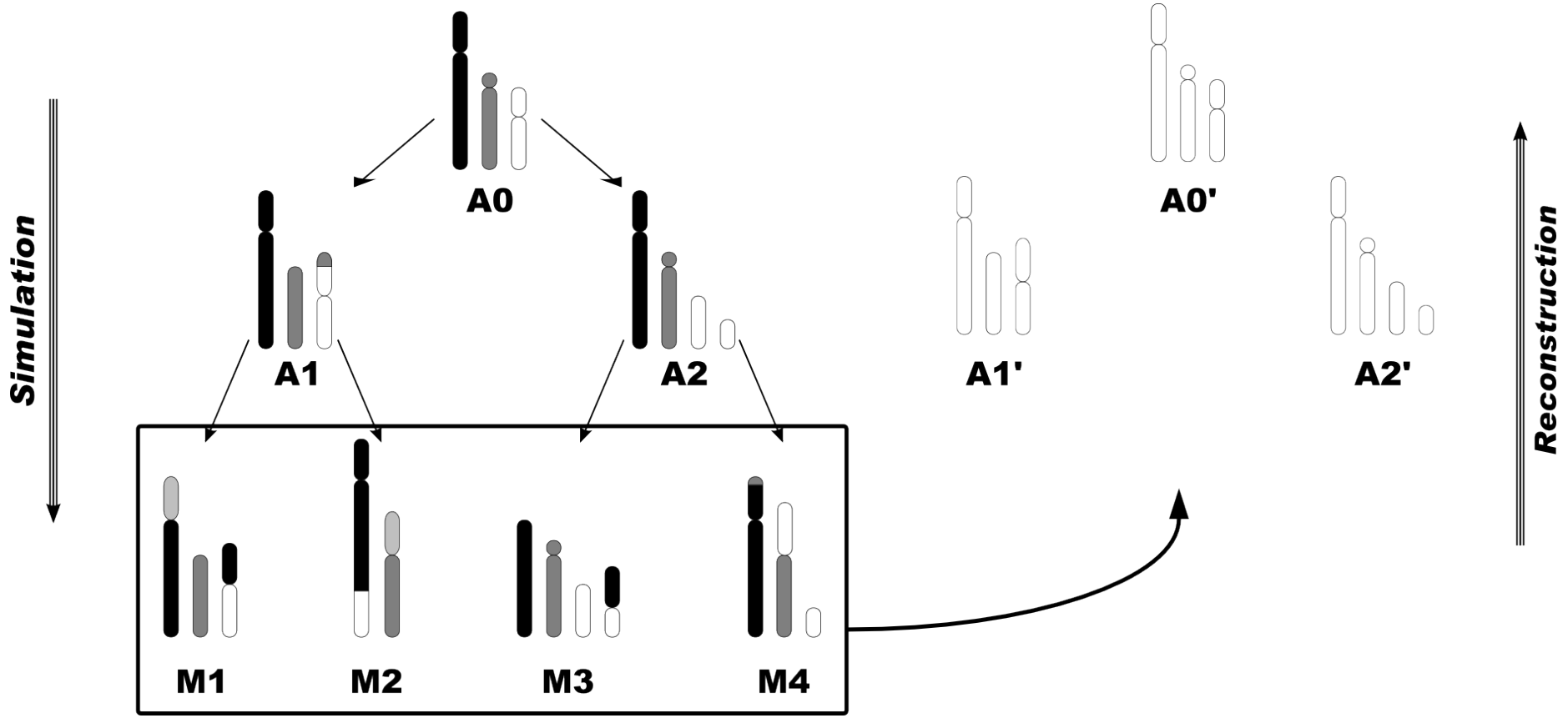
Simulation of genome rearrangements

Starting from a random ancestral genome A0, ancestral (A1, A2) and modern genomes (M1, M2, M3, M4) are simulated through chromosome rearrangements.



Simulation of genome rearrangements

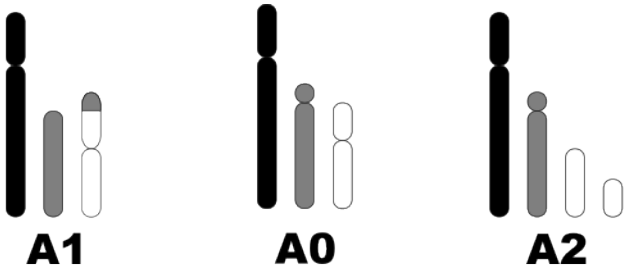
Starting from a random ancestral genome A0, ancestral (A1, A2) and modern genomes (M1, M2, M3, M4) are simulated through chromosome rearrangements. The reconstruction program is called using only modern genes.



Simulation of genome rearrangements

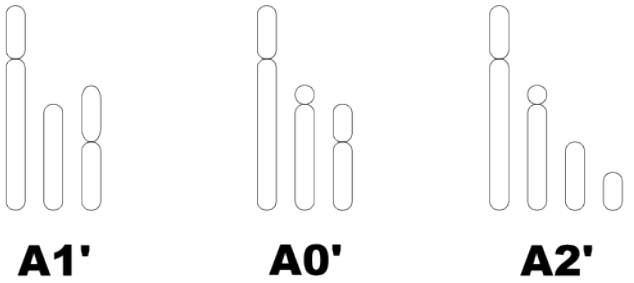
Reconstructed ancestral genomes are compared to simulated ones.
Measures such as sensitivity & specificity can be derived.

Simulated ancestral genomes



Comparison

Reconstructed ancestral genomes



Simulations – Benchmark

Benchmark :

- From 100 to 20000 markers (genes)
- With a rearrangement rate from 0.2x to 3x
  reference rate is taken from Zhao et al.
- With a constant gene content in the benchmark (other methods' limit)

Simulations – Benchmark

MGR – Bourque et al., 2004

	100	500	1000	5000	10000	20000
0.2x						
0.5x						
1x						
2x						
3x						

MGRA – Alekseyev et al., 2009

	100	500	1000	5000	10000	20000
0.2x						
0.5x						
1x						
2x						
3x						

InferCARs – Ma et al., 2006

	100	500	1000	5000	10000	20000
0.2x						
0.5x						
1x						
2x						
3x						

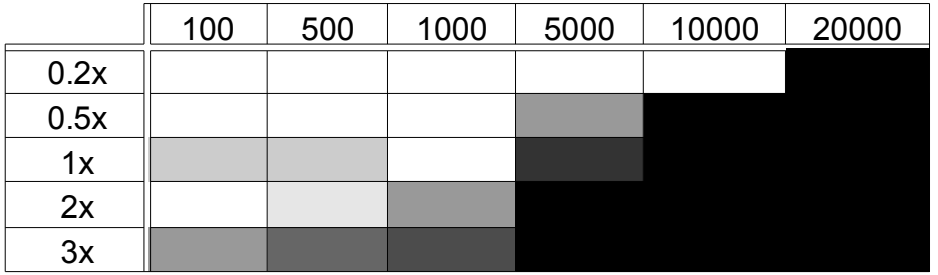
AGORA

	100	500	1000	5000	10000	20000
0.2x						
0.5x						
1x						
2x						
3x						

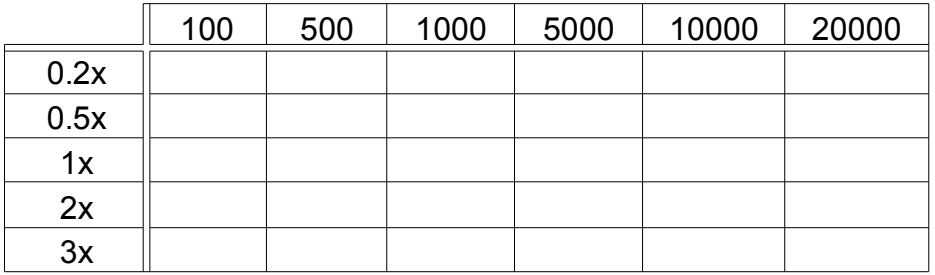
Simulations – Benchmark

MGR – Bourque et al., 2004

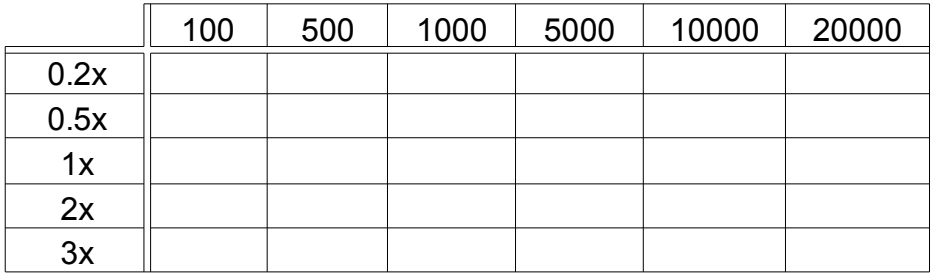
MGRA – Alekseyev et al., 2009



InferCARs – Ma et al., 2006



AGORA



With large / highly rearranged genomes, MGR & MGRA fail to produce any result.

= 0% failure = 100% failure

InferCARs & AGORA don't have this limit.

Simulations – Benchmark

MGR – Bourque et al., 2004

	100	500	1000	5000	10000	20000
0.2x	100%	100%	100%	100%		
0.5x	100%	100%	100%	100%		
1x	100%	99,86%	100%			
2x	100%	99,96%	100%			
3x	99,77%	100%				

MGRA – Alekseyev et al., 2009

	100	500	1000	5000	10000	20000
0.2x	100%	100%	100%	100%	100%	
0.5x	100%	100%	100%	100%		
1x	100%	100%	100%	100%		
2x	100%	100%	100%			
3x	100%	100%	100%			

InferCARs – Ma et al., 2006

	100	500	1000	5000	10000	20000
0.2x	100%	100%	100%	100%	100%	100%
0.5x	100%	100%	100%	100%	100%	100%
1x	100%	100%	100%	100%	100%	100%
2x	100%	100%	100%	100%	99,99%	99,99%
3x	100%	100%	100%	99,99%	99,98%	99,98%

AGORA

	100	500	1000	5000	10000	20000
0.2x	100%	100%	100%	100%	100%	100%
0.5x	100%	100%	100%	100%	99,99%	99,99%
1x	100%	100%	100%	99,99%	99,99%	99,99%
2x	100%	100%	99,99%	99,98%	99,98%	99,99%
3x	100%	100%	99,99%	99,98%	99,98%	99,99%

Specificity is above 99.7% in all reconstructions (equivalent figures for sensitivity).

Even at high rearrangement rate / large genome size, InferCARs & AGORA show more than 99.98% of specificity.

Simulations – AGORA real performances

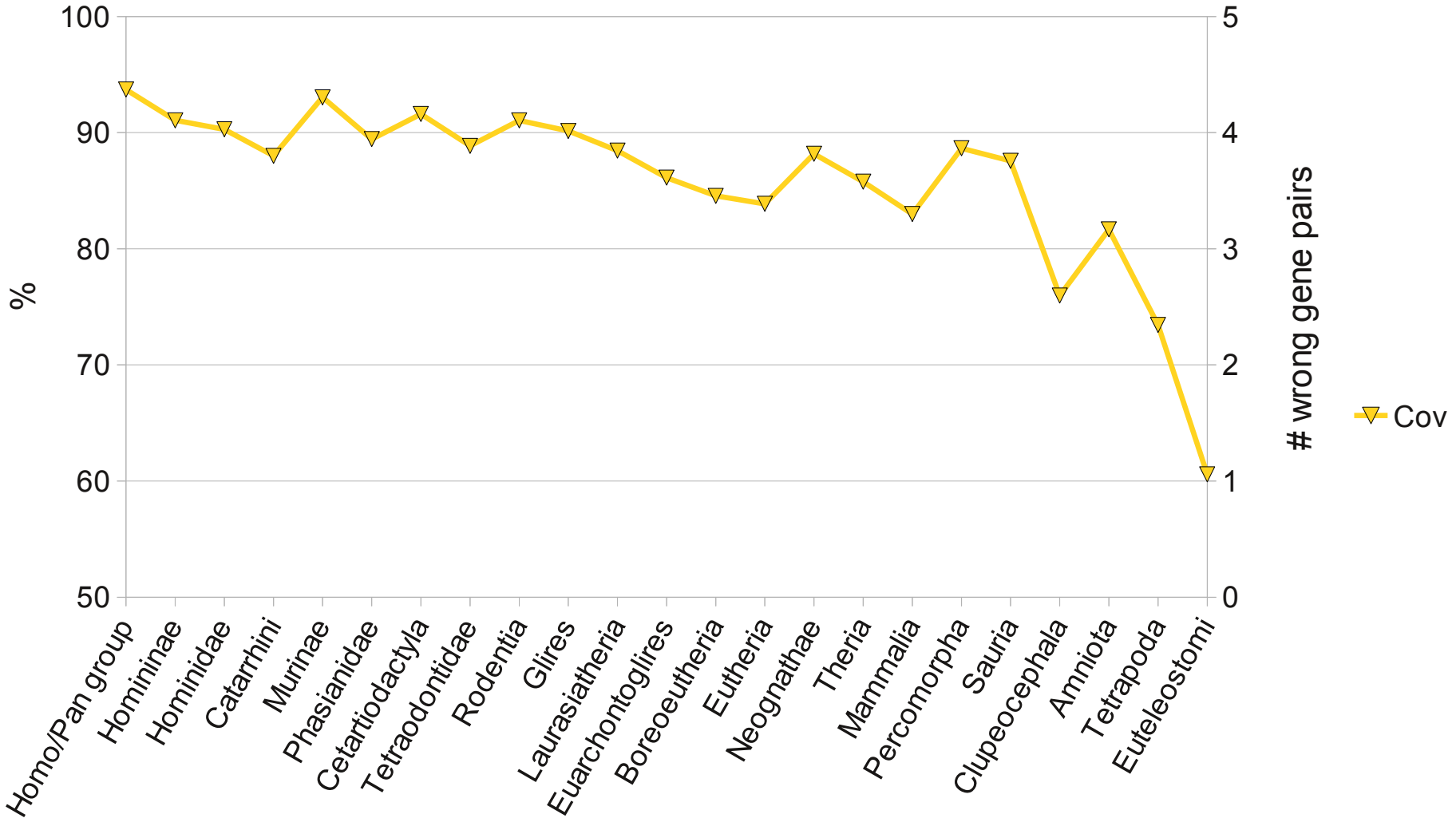
AGORA performances :

- With gene content copied from Ensembl phylogenetic trees (AGORA alone)
- 1x rearrangement rate

Simulations – AGORA real performances

The performances are evaluated by :

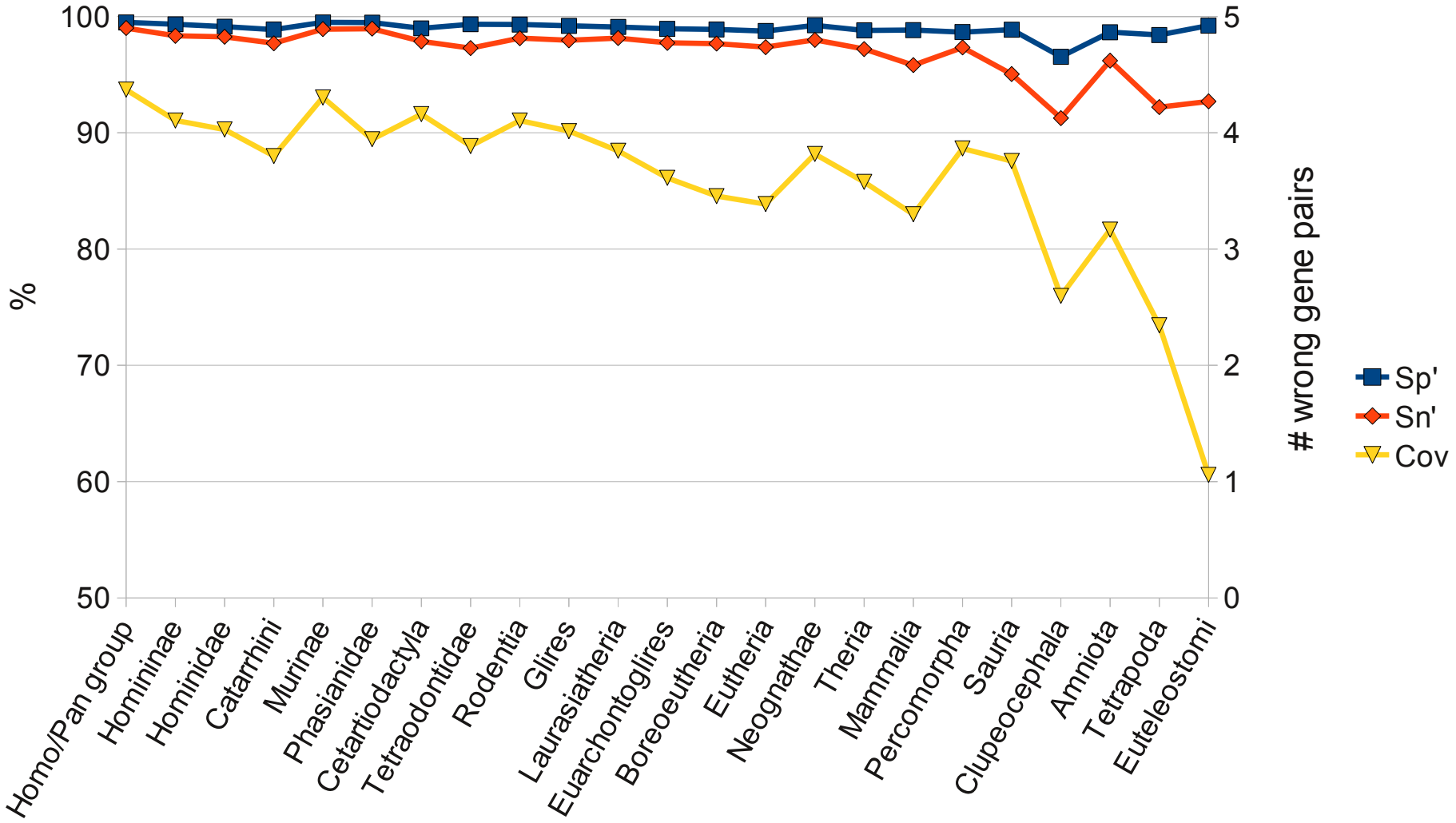
- coverage (proportion of genes included in contigs)



Simulations – AGORA real performances

The performances are evaluated by :

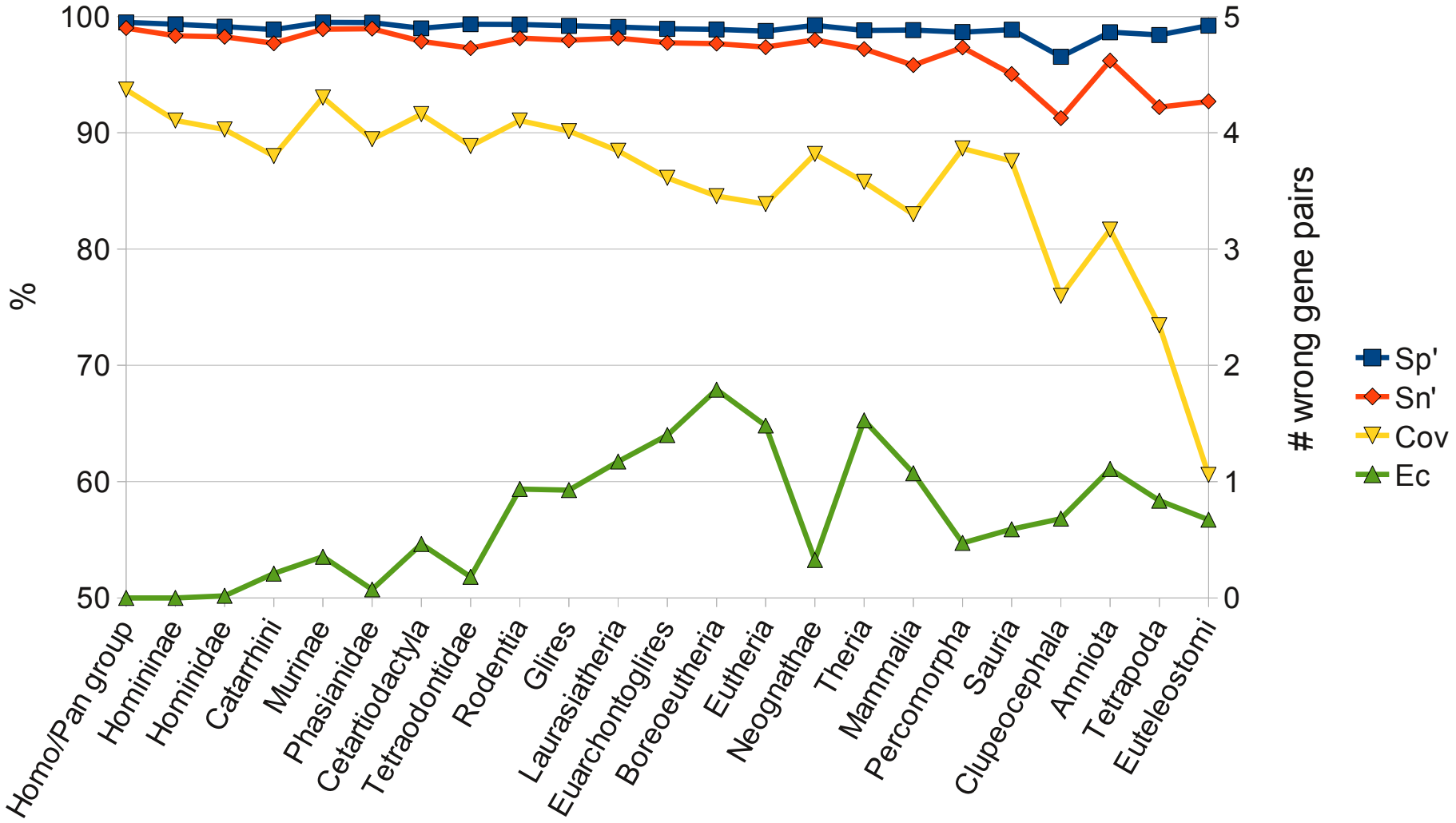
- coverage (proportion of genes included in contigs)
- specifitivity & sensitivity , on gene pairs



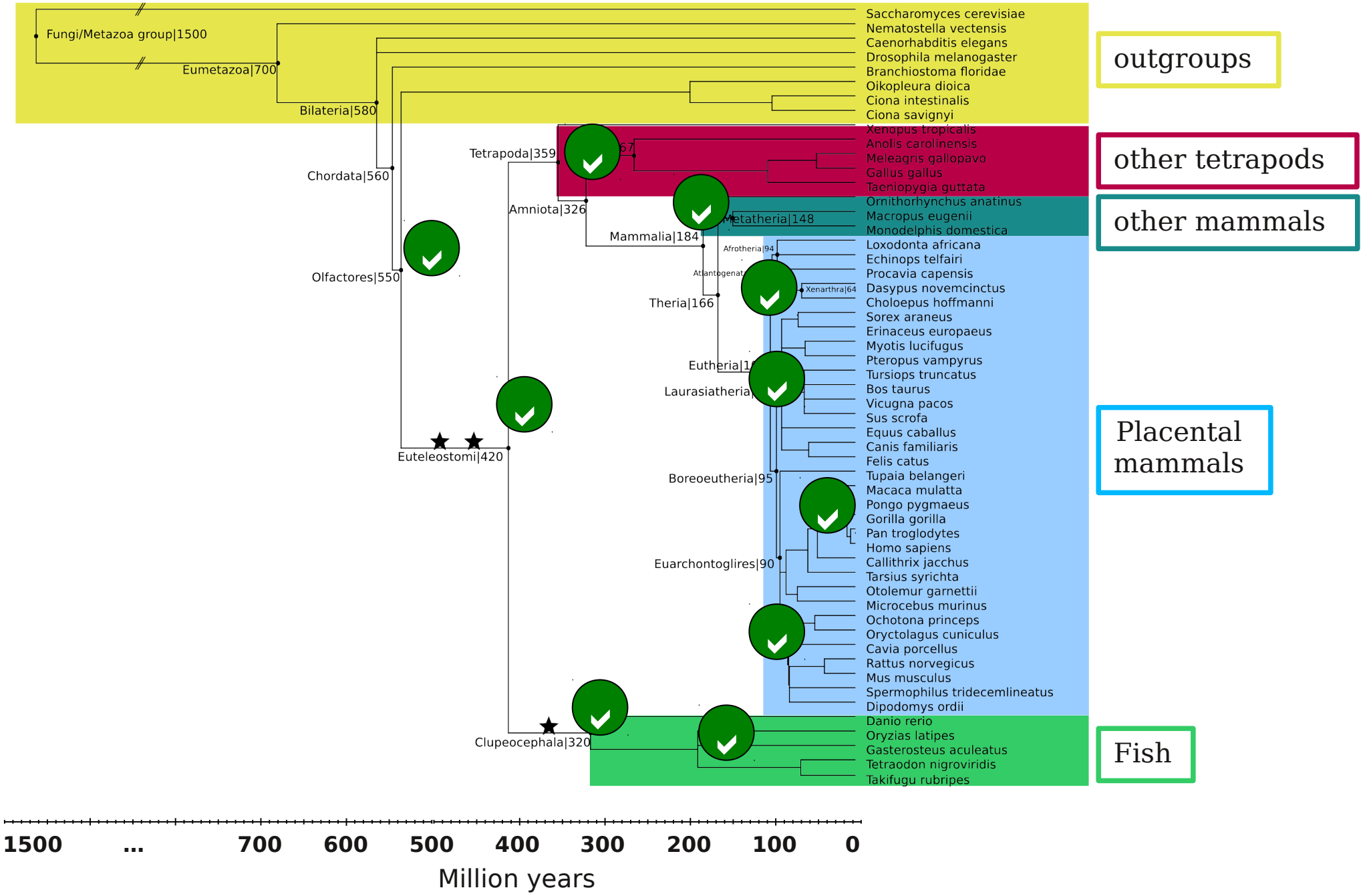
Simulations – AGORA real performances

The performances are evaluated by :

- coverage (proportion of genes included in contigs)
- specifitivity & sensitivity , on gene pairs
- average number of wrong gene pairs per ancestral genome



Phylogenetic tree of vertebrates & model outgroup species



Conclusions

AGORA accurately reconstruct gene order in vertebrate genomes.

Reconstructions are available on the web server Genomicus
http://www.dyogen.ens.fr/genomicus

AGORA validated method is « limited » to adjacency signal
→ Reconstructions may miss ancestral associations

Ongoing work : validation of synteny-based, WGD-based methods

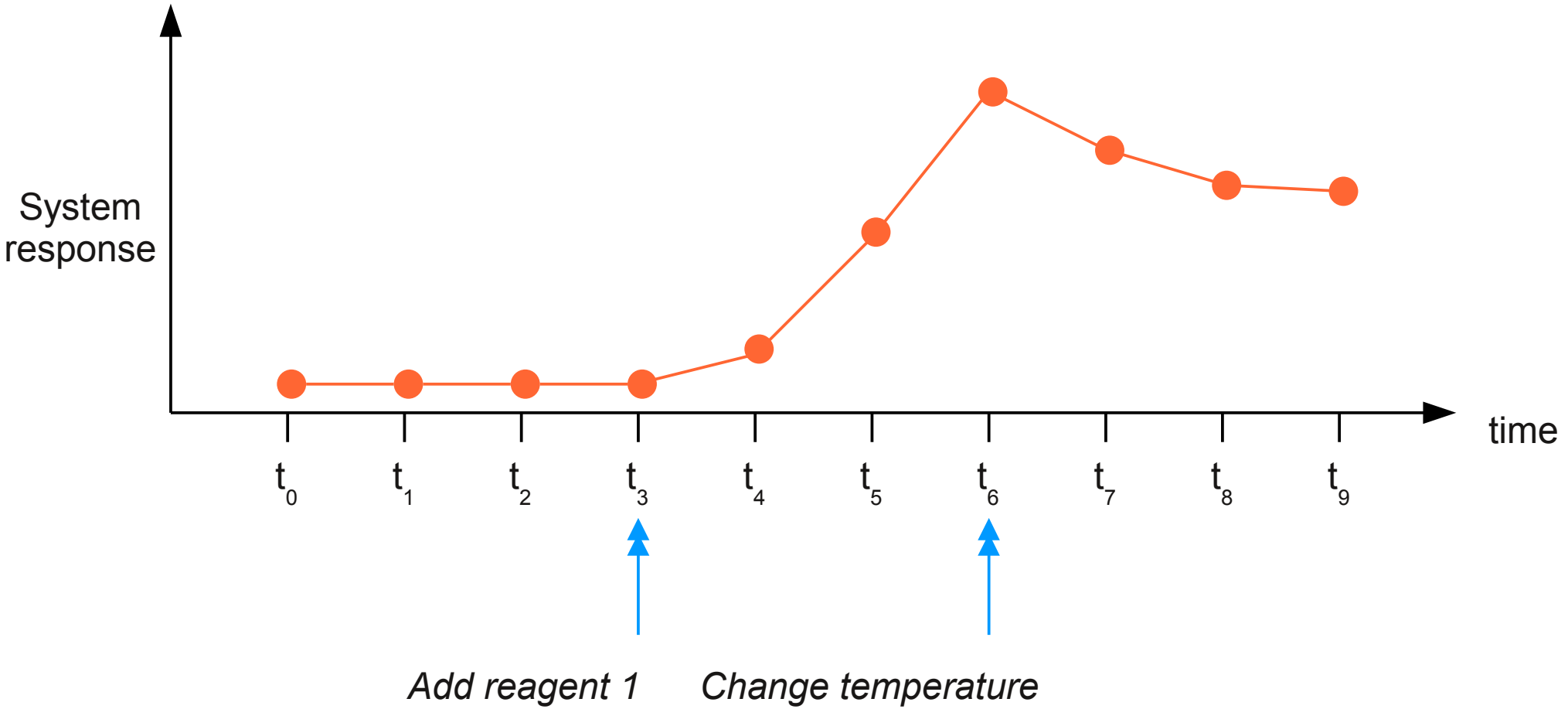
Enhancements & Future work

- Functionnal annotation
- Rearrangement model
- Other organisms

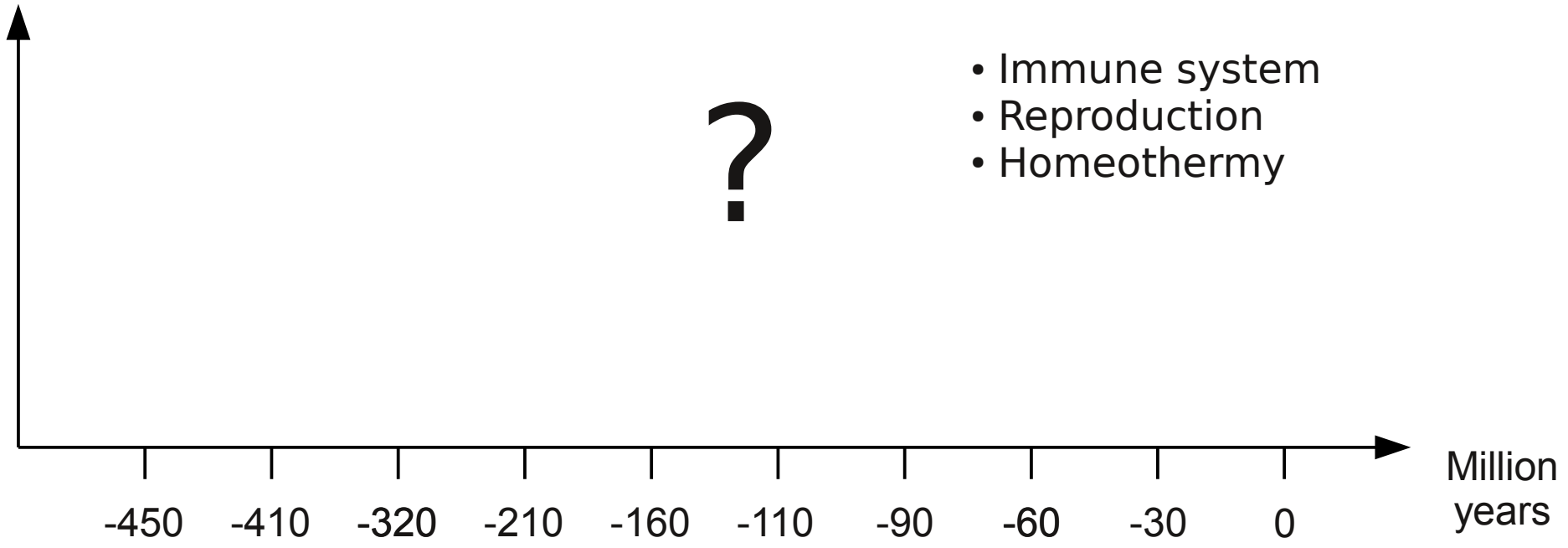
Enhancements & Future work

- Functionnal annotation
- Rearrangement model
- Other organisms

New framework to study evolution



New framework to study evolution



Functional annotation of ancestral genomes

Current AGORA reconstructions are limited to (protein coding) gene order, and lack functional annotations such as :

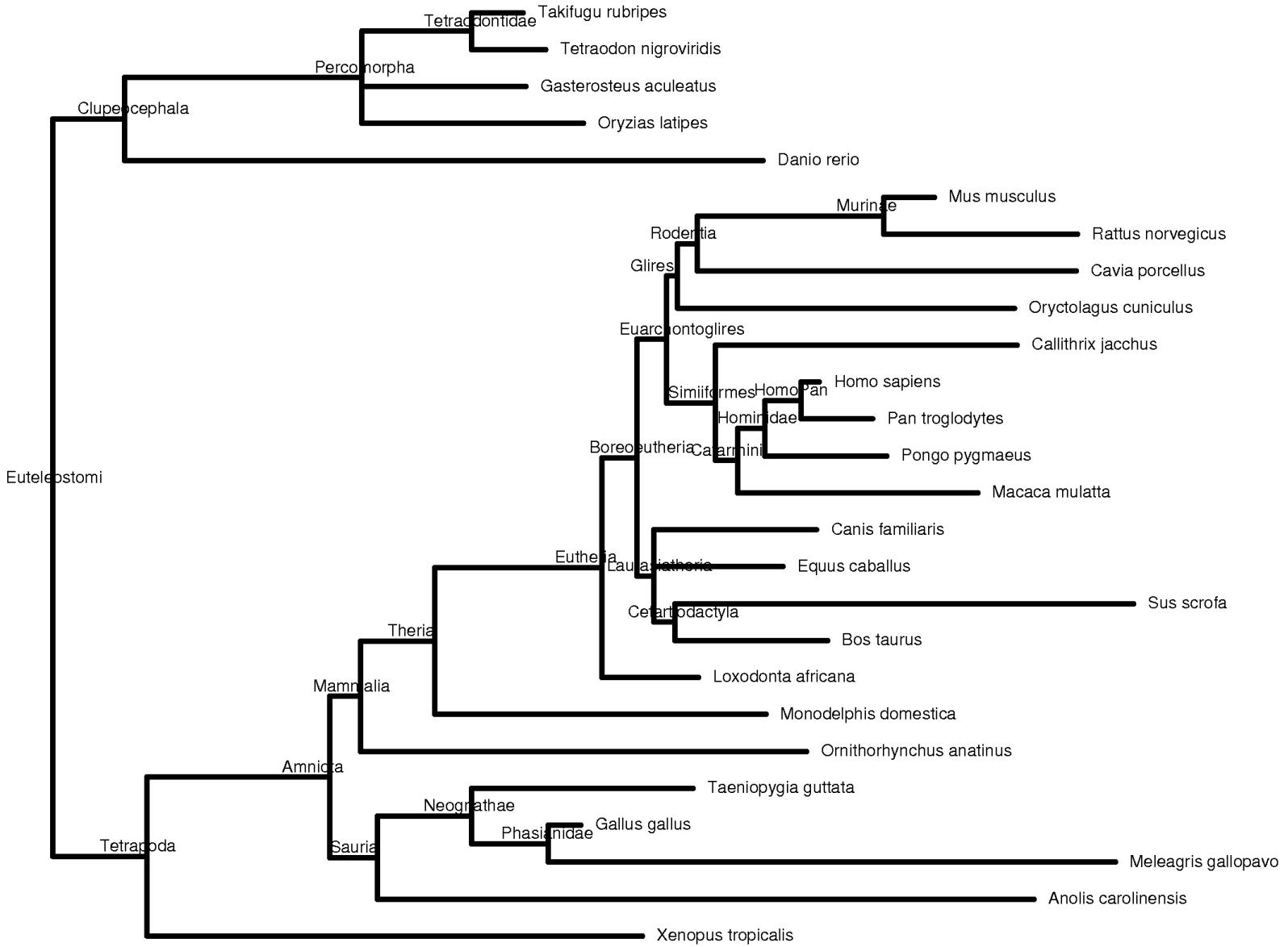
- ncRNAs and CNEs (Conserved Non-coding Elements) position
- Ancestral regulatory circuits
- Ancestral sequence
- Gene functions (ex: Gene Ontology)

Enhancements & Future work

- Functionnal annotation
- Rearrangement model
- Other organisms

Rearrangement rate

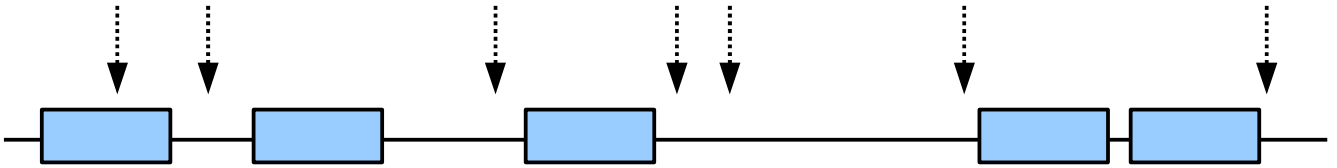
Pairwise comparisons between successive ancestors define the list of all the rearrangements in vertebrates.



Rearrangement model

The model of chromosome breakage is still debated. The three hypotheses are:

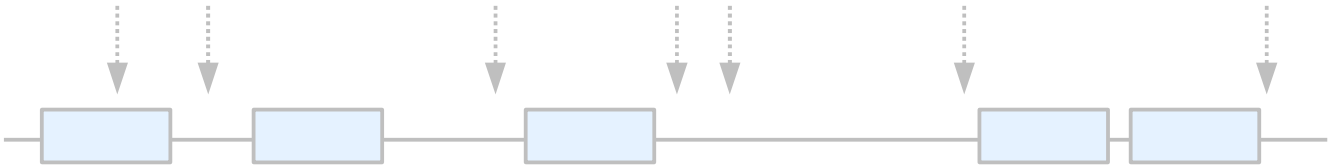
- Random Breakage Model (Nadeau et al., Ma et al.)



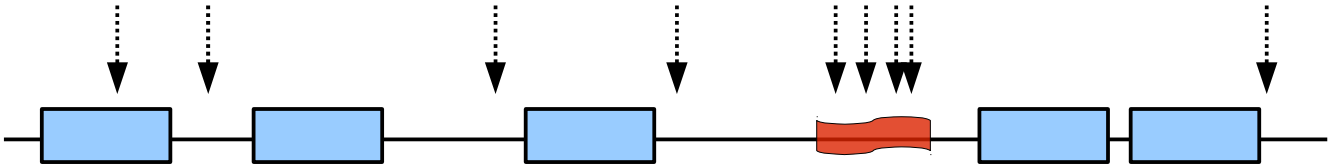
Rearrangement model

The model of chromosome breakage is still debated. The three hypotheses are:

- Random Breakage Model (Nadeau et al., Ma et al.)



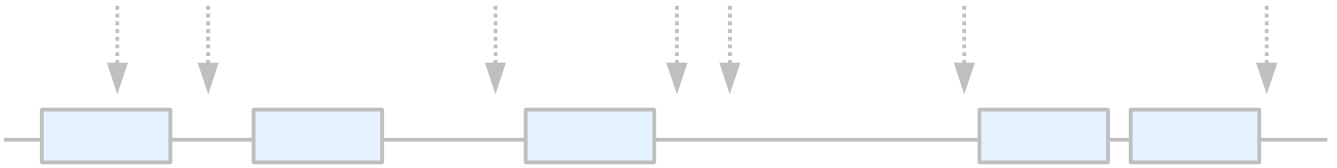
- Fragile Sites Model (Pevzner et al.)



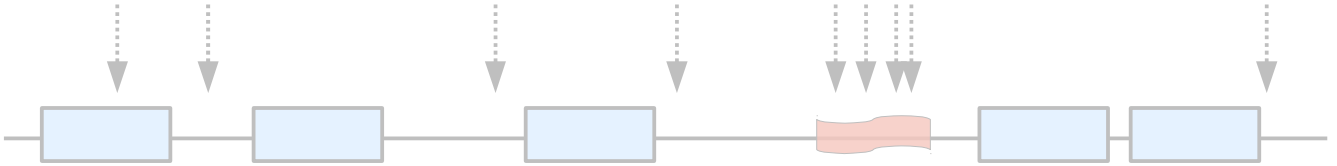
Rearrangement model

The model of chromosome breakage is still debated. The three hypotheses are:

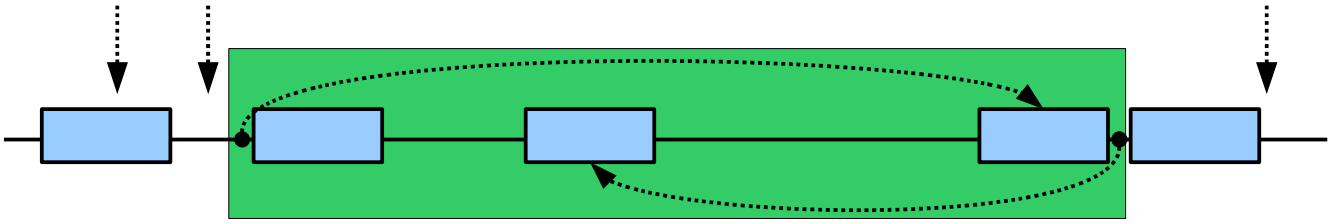
- Random Breakage Model (Nadeau et al., Ma et al.)



- Fragile Sites Model (Pevzner et al.)



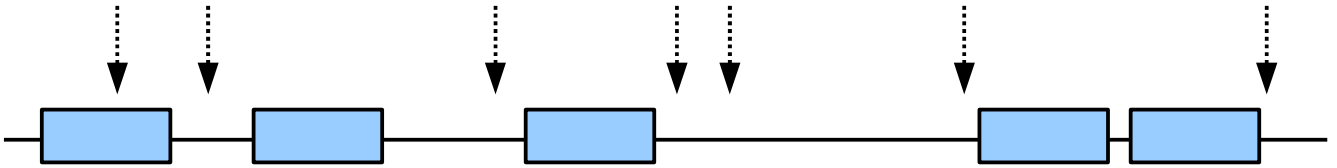
- Genomic Regulatory Blocks (Kikuta et al.)



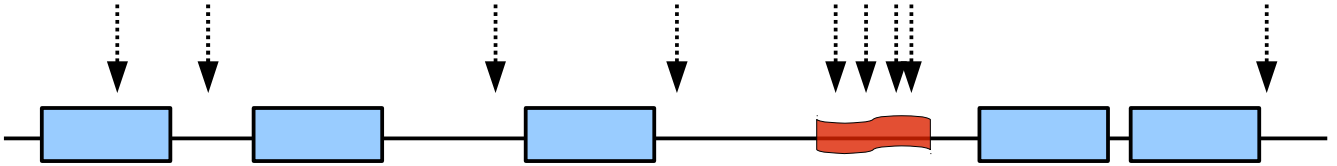
Rearrangement model

The model of chromosome breakage is still debated. The three hypotheses are:

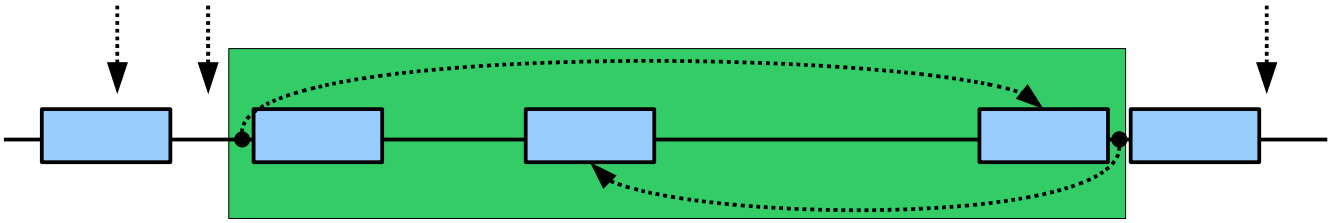
- Random Breakage Model (Nadeau et al., Ma et al.)



- Fragile Sites Model (Pevzner et al.)



- Genomic Regulatory Blocks (Kikuta et al.)



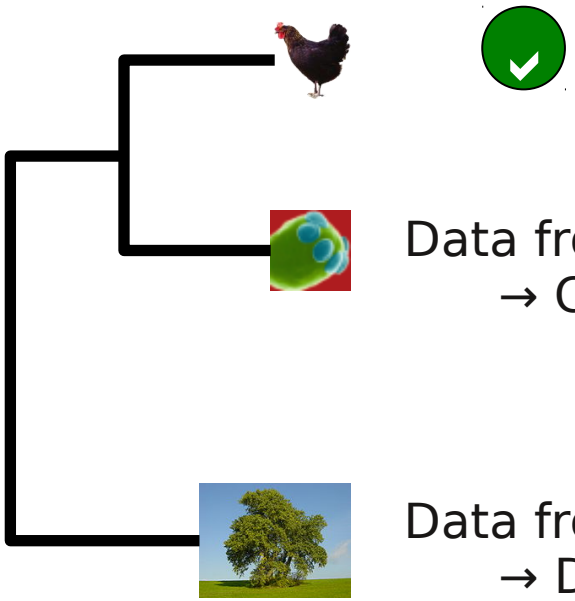
Ongoing work: analysis of mammalian breakpoints (PhD student)

Enhancements & Future work

- Functionnal annotation
- Rearrangement model
- Other organisms

AGORA in other organisms

Genome sequence data is available for many clades



Data from BroadInstitute & EnsemblGenomes
→ Computation of all the phylogenetic trees with TreeBest
Vilella et al.

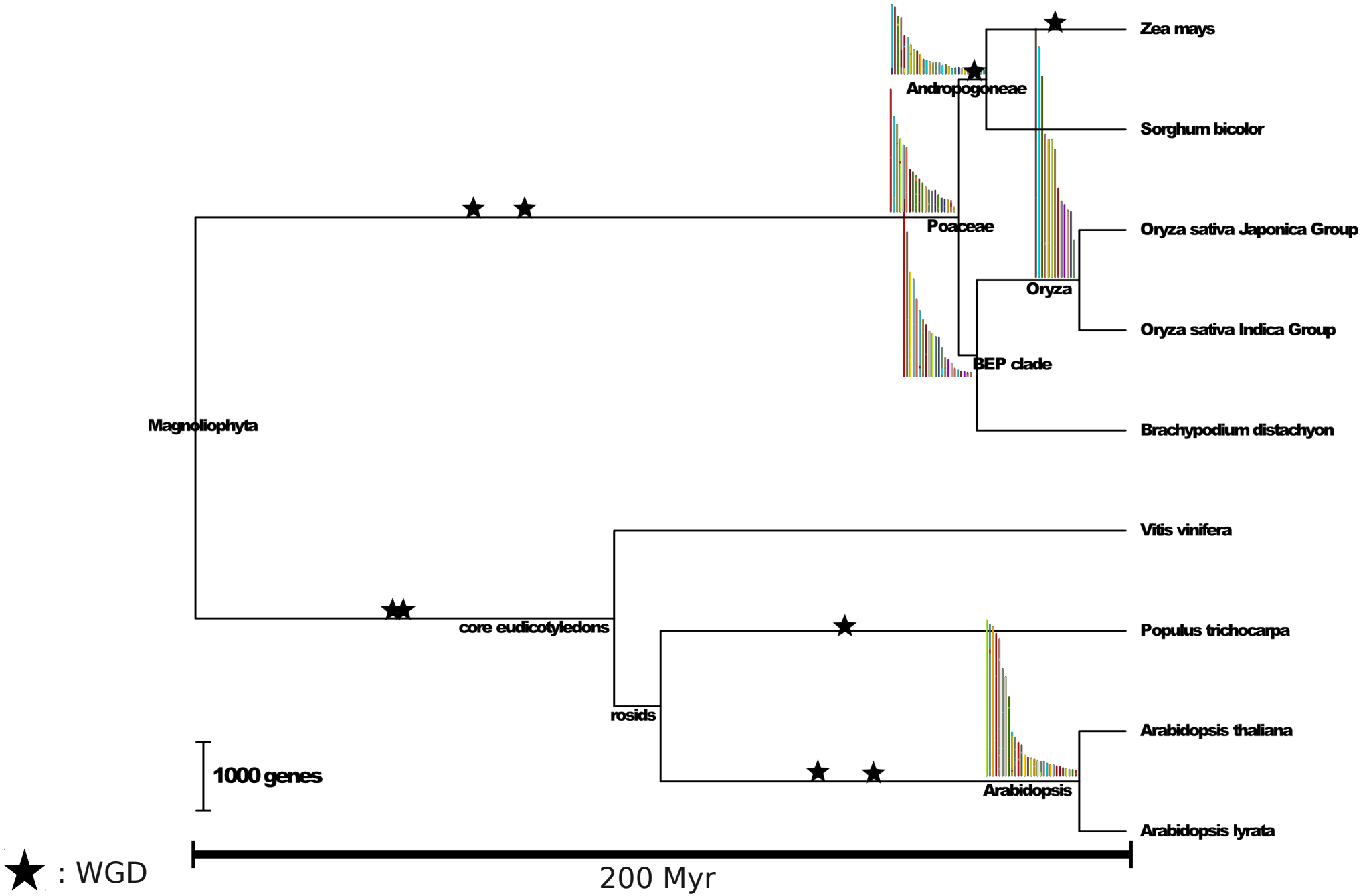
Data from EnsemblGenomes & collaborators (INRA Clermont)
→ Direct application of AGORA + optimisation of gene families

Ongoing work: optimisation of AGORA pipeline

AGORA in other organisms – Fungi

★ : WGD

AGORA in other organisms – Plants



Conclusions

AGORA reconstructions define a new, generic, framework to study genome evolution.

Questions that were not answered due to the lack of ancestral data can now be tackled.

This will need a lot more work !

Acknowledgements

The Dyogen Team

Pierre Vincens & his team

« Monday cake » team

Alumni: Sarah, Emmanuel, Charles