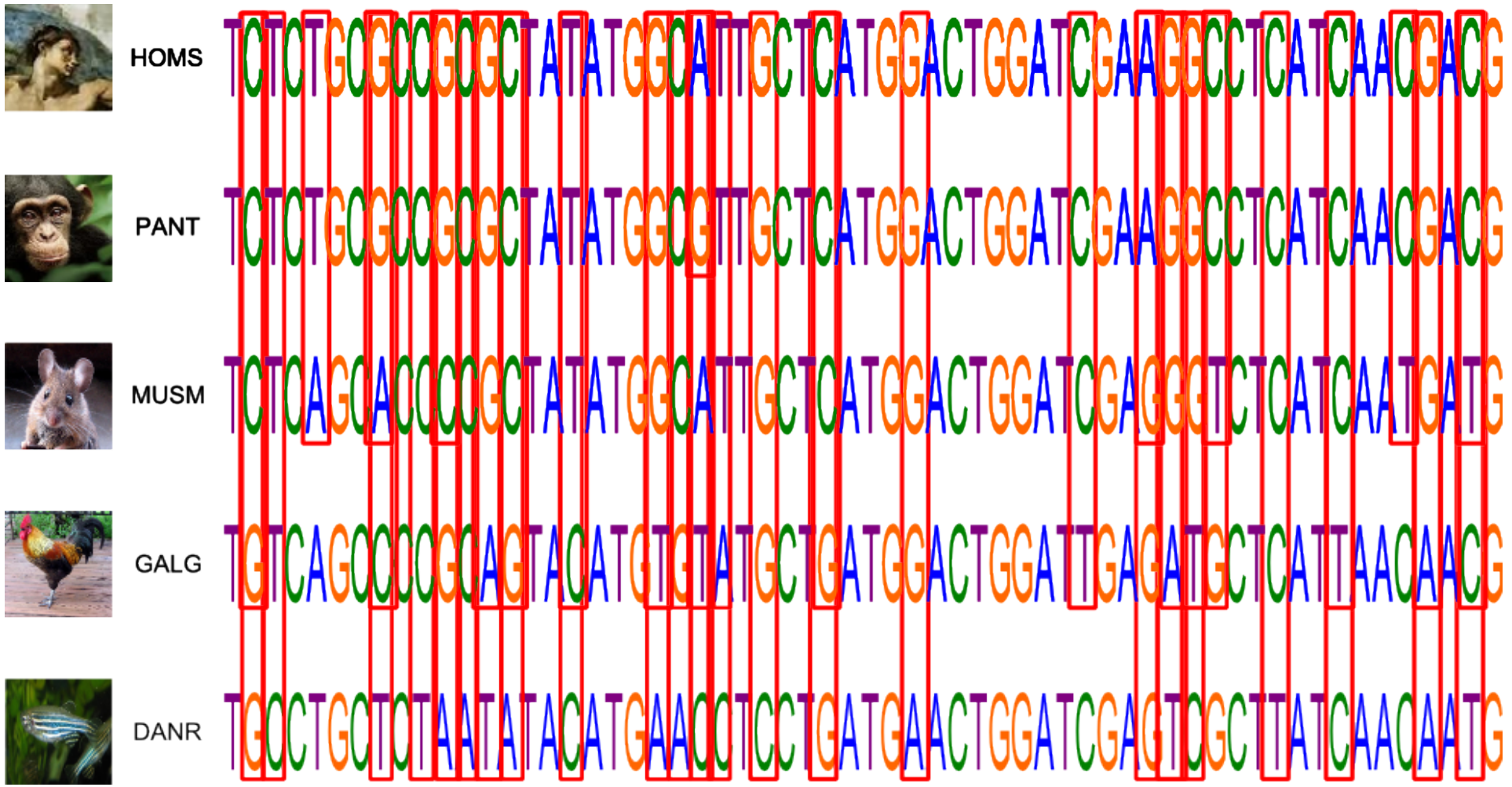
Datasets used in phylogenetic reconstruction – usage of **single genes**

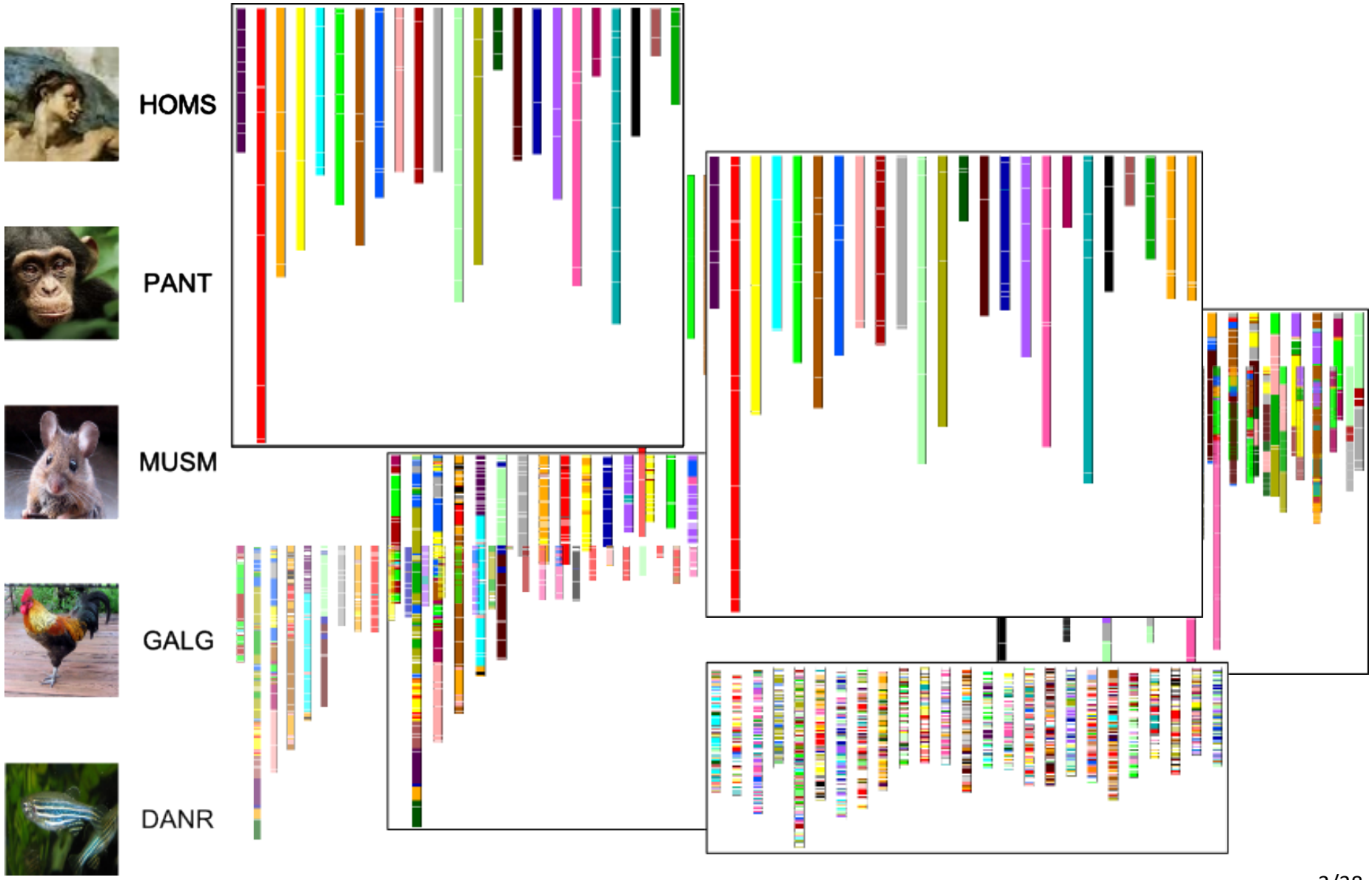
Obtained from various researchers and online databases

- 1322 lsu rRNA of all organisms
- 2000 Eukaryotic rRNA
- 2594 rbcL DNA
- 4583 Actinobacteria 16s rRNA
- 6590 ssu rRNA of all Eukaryotes
- 7180 three-domain rRNA
- 7322 Firmicutes bacteria 16s rRNA
- 8506 three-domain+2org rRNA
- 11361 ssu rRNA of all Bacteria
- 13921 Proteobacteria 16s rRNA

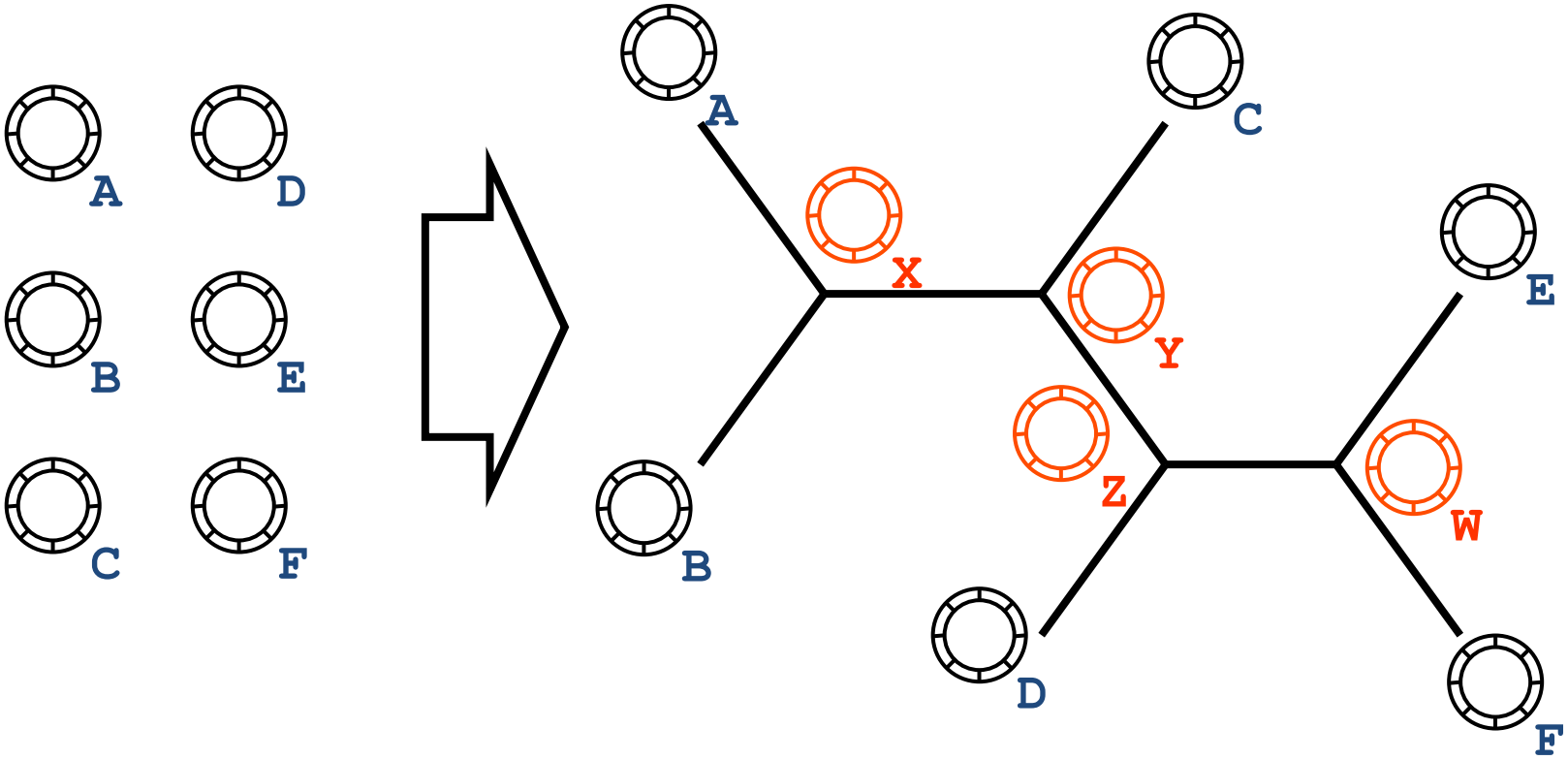
Evolution des espèces et mutations ponctuelles



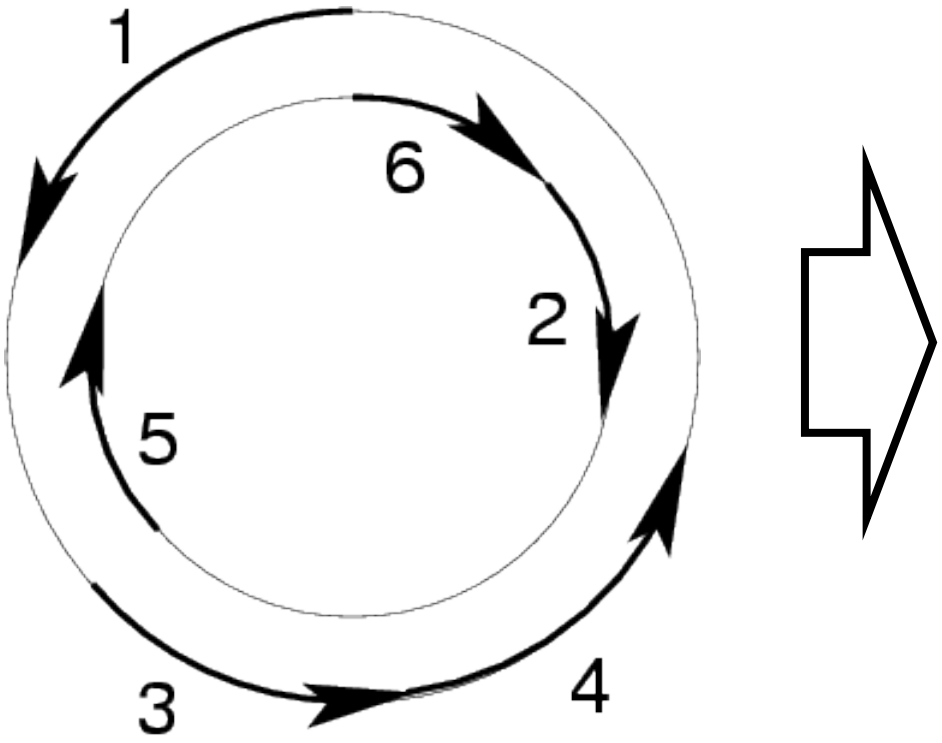
Evolution des espèces et réarrangements chromosomiques



Whole-Genome Phylogenetics



Genomes As Signed Permutations



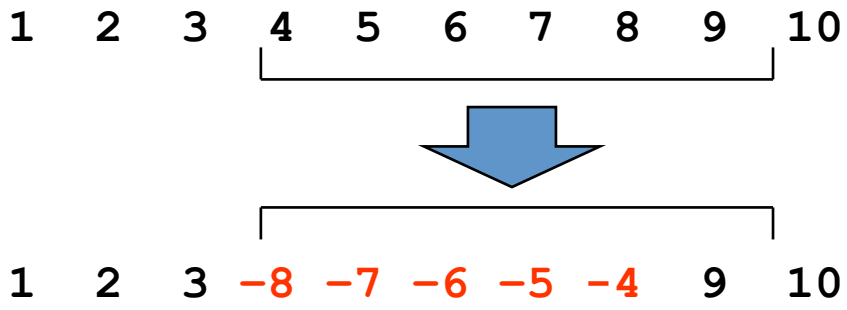
1 –5 3 4 -2 -6

or

6 2 -4 –3 5 –1

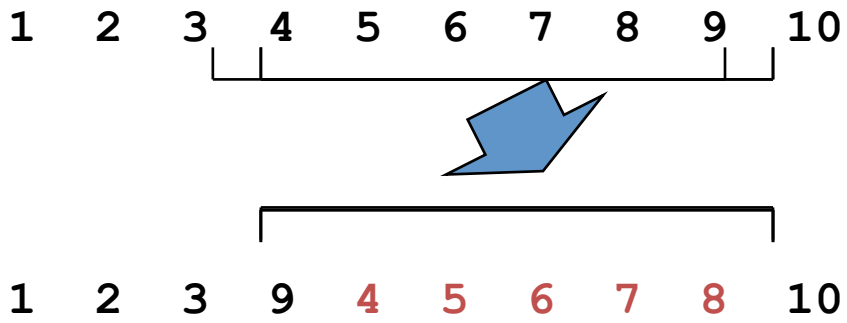
etc.

Genomes Evolve by Rearrangements



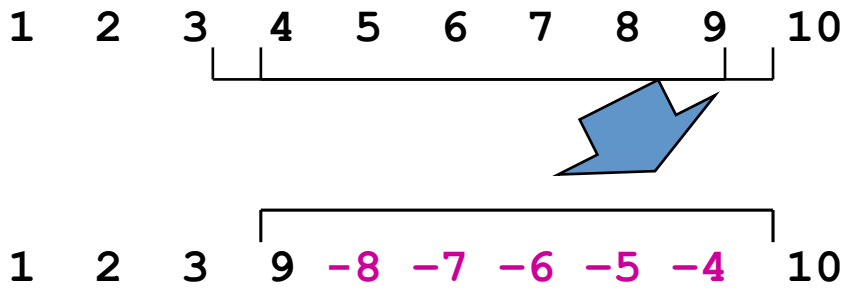
- Inversion (Reversal)

Genomes Evolve by Rearrangements



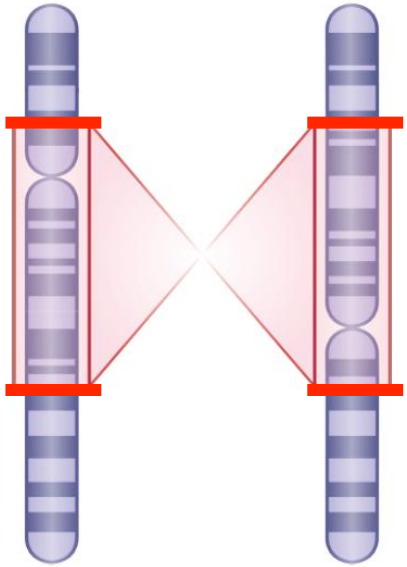
- Transposition

Genomes Evolve by Rearrangements

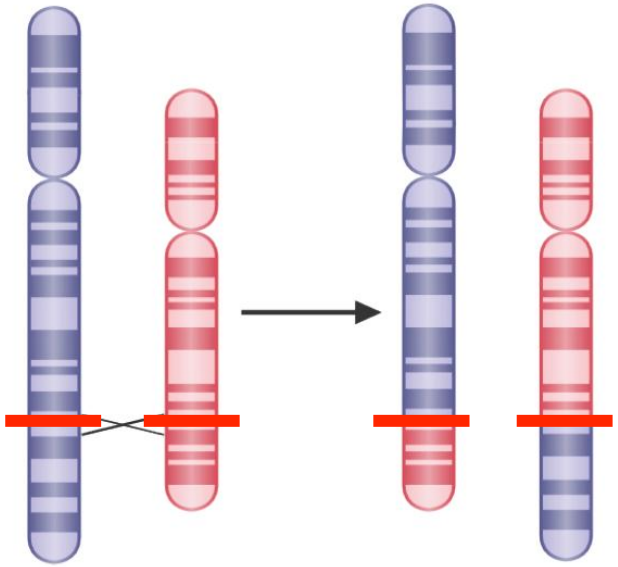


- Inverted Transposition

Les différents réarrangements chromosomiques



(a) Inversion



(b) Translocation.

- Inversion
- Translocation
  - Fusion
  - Fission
- Délétion
- Insertion
- Duplication
- Autres

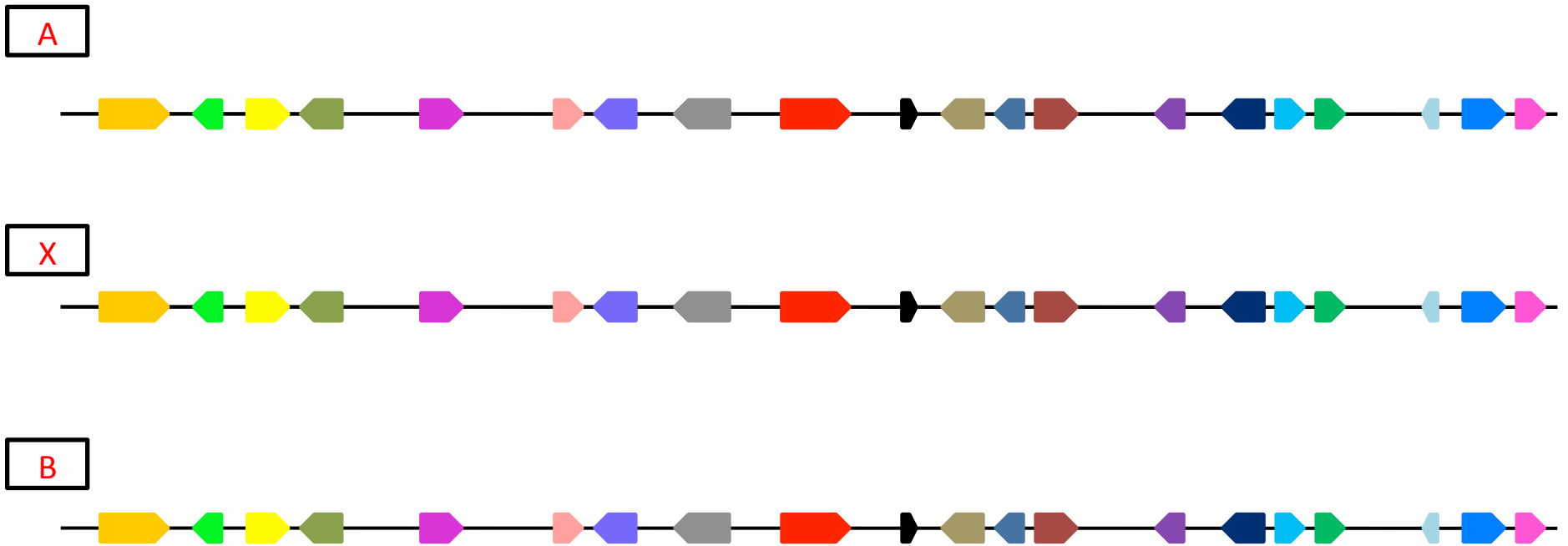
Chaque inversion crée 2 points de cassure dans chacun des génomes

Other types of events

- Duplications, Insertions, and Deletions (changes gene content)
- Fissions and Fusions (for genomes with more than one chromosome)

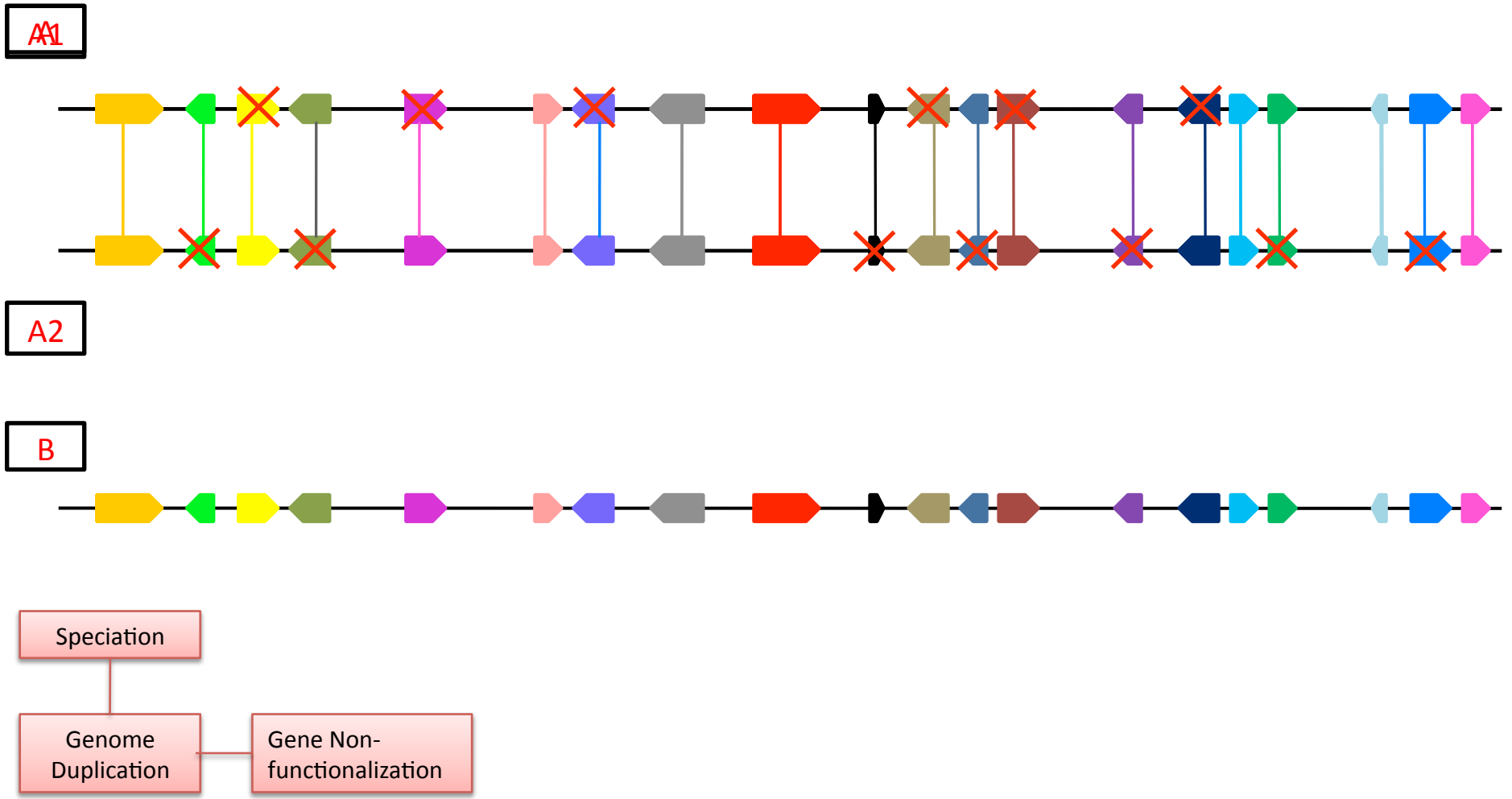
These events change the number of copies of each gene in each genome *("unequal gene content")*

Evolution after Whole Genome Duplication



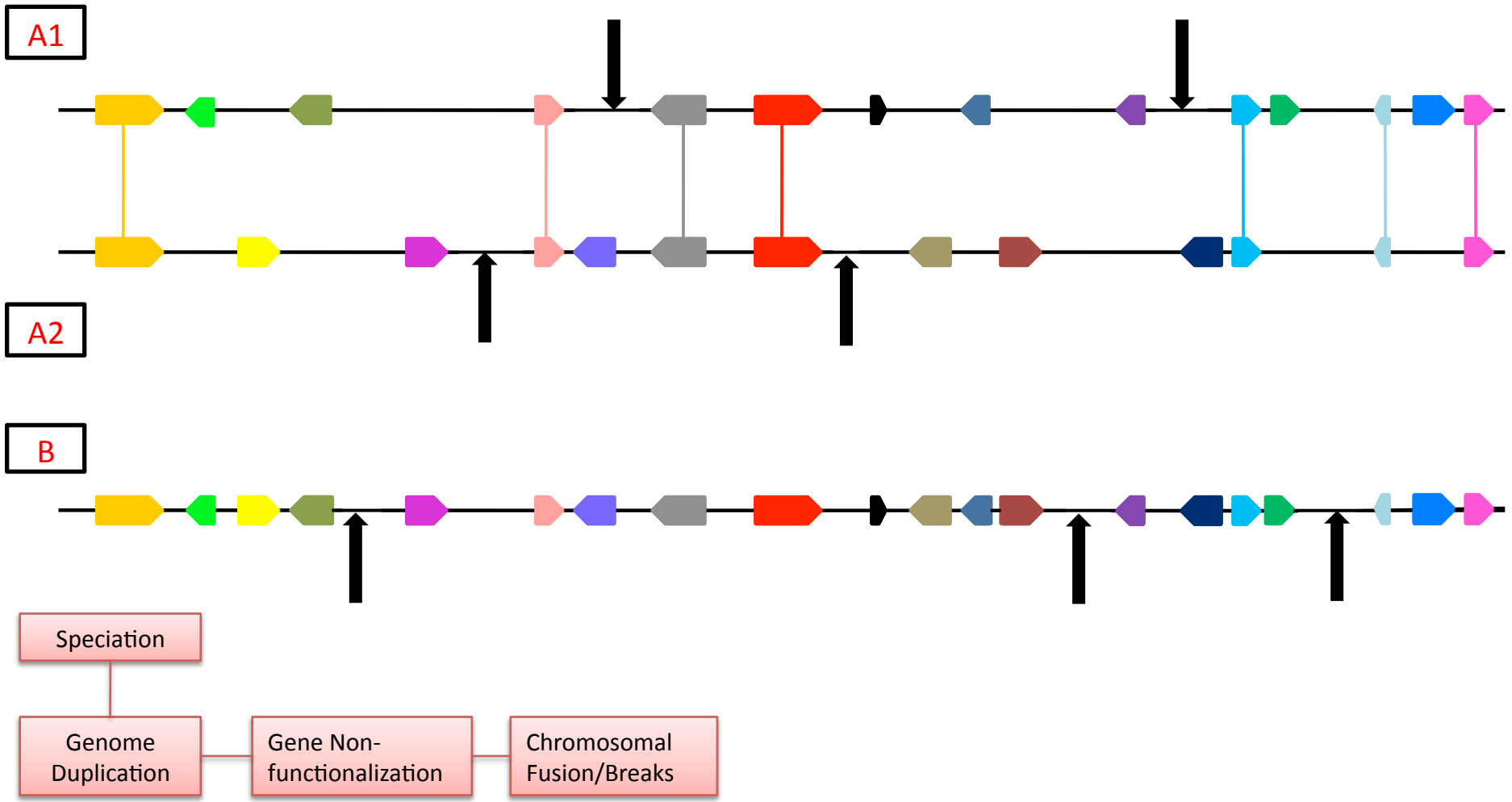
Courtesy of Param-Priva Singh

Evolution after Whole Genome Duplication



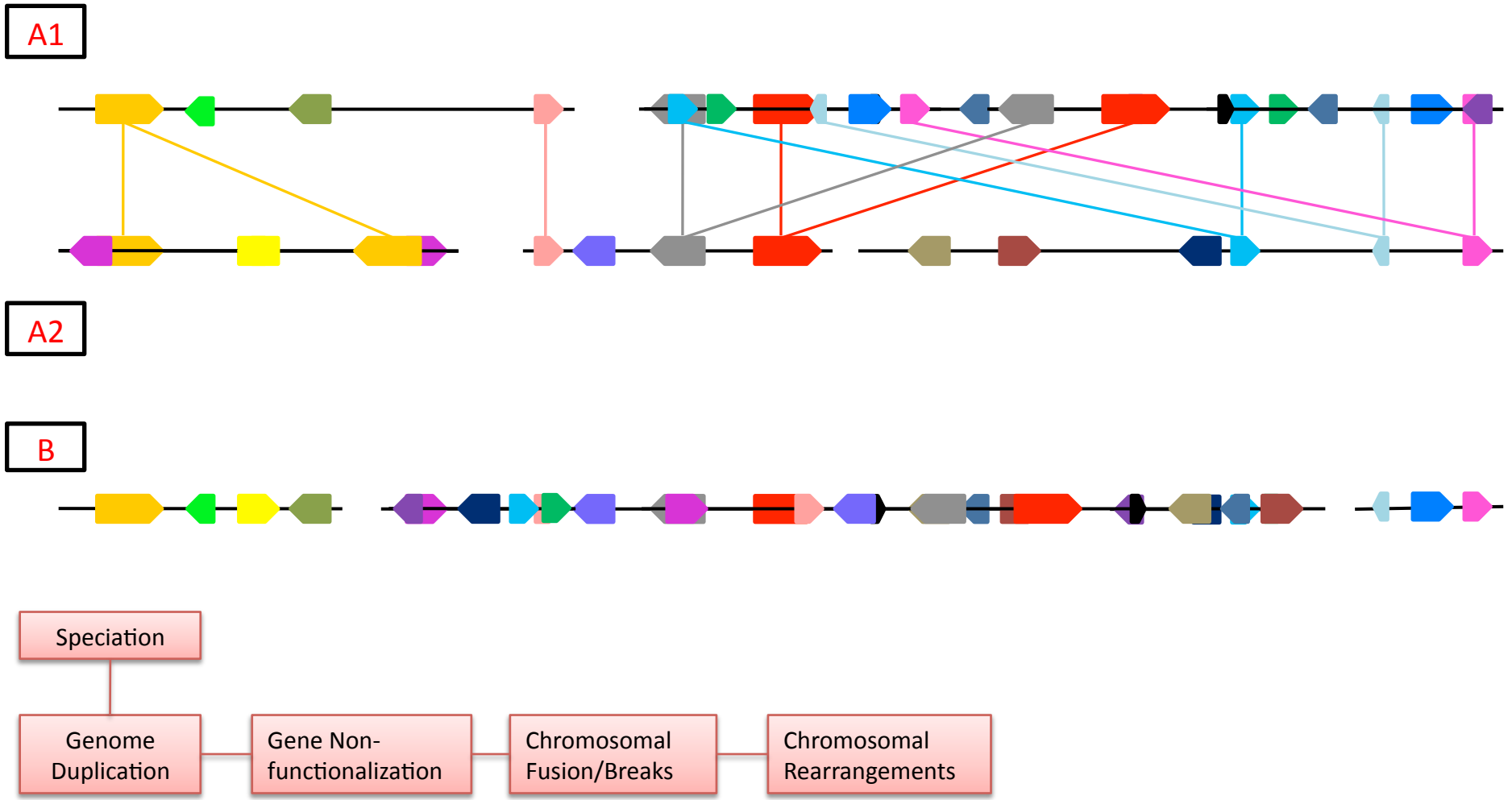
Courtesy of Param-Priva Singh

Evolution after Whole Genome Duplication



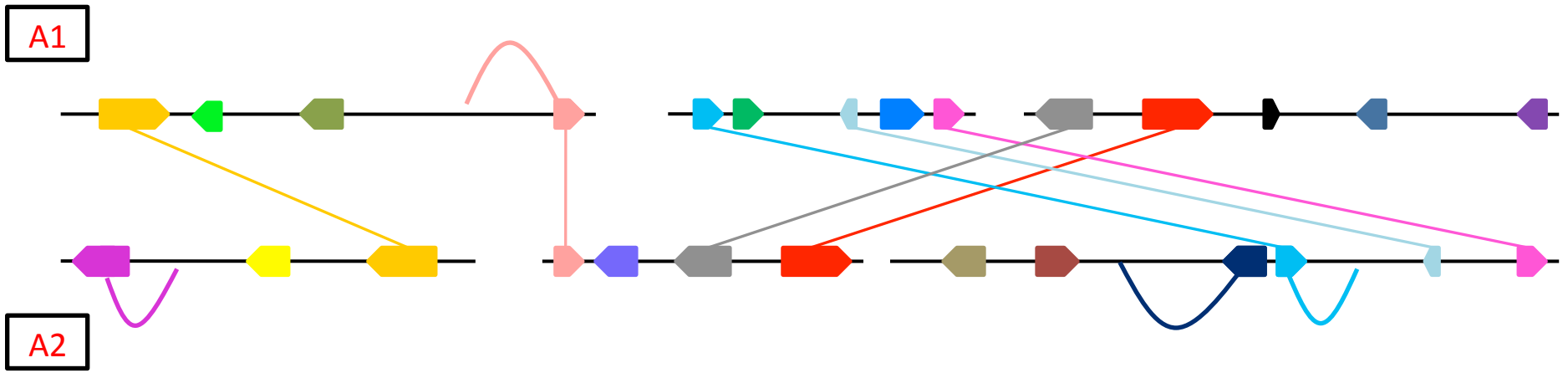
Courtesy of Param-Priva Singh

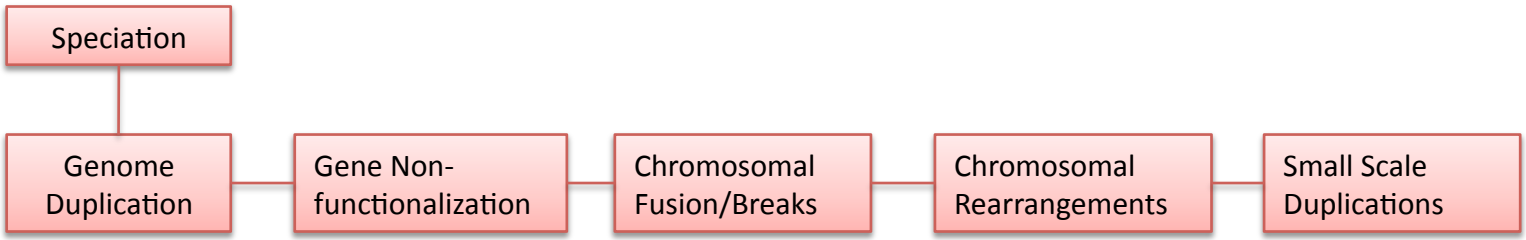
Evolution after Whole Genome Duplication



Courtesy of Param-Priva Singh

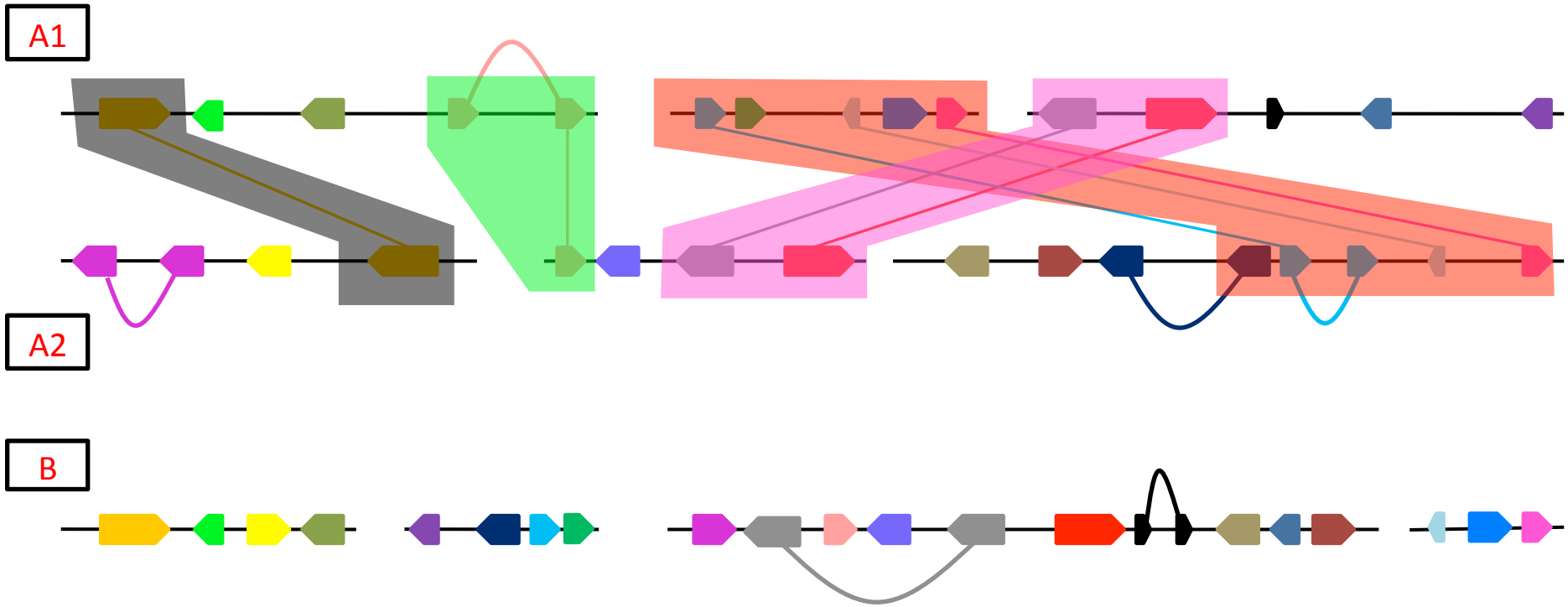
Evolution after Whole Genome Duplication

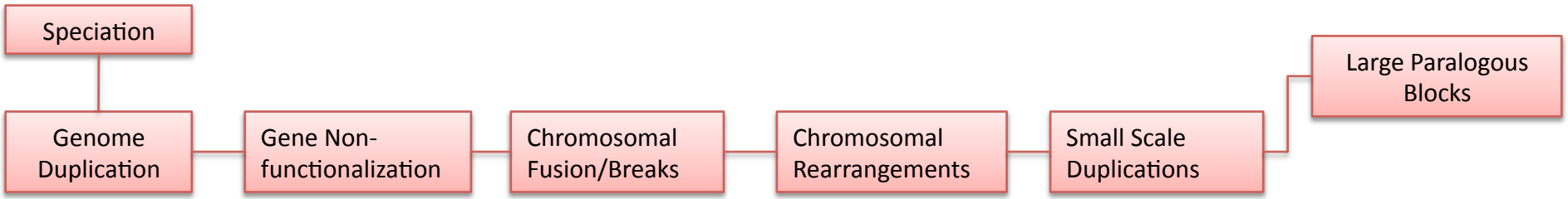




Courtesy of Param-Priva Singh

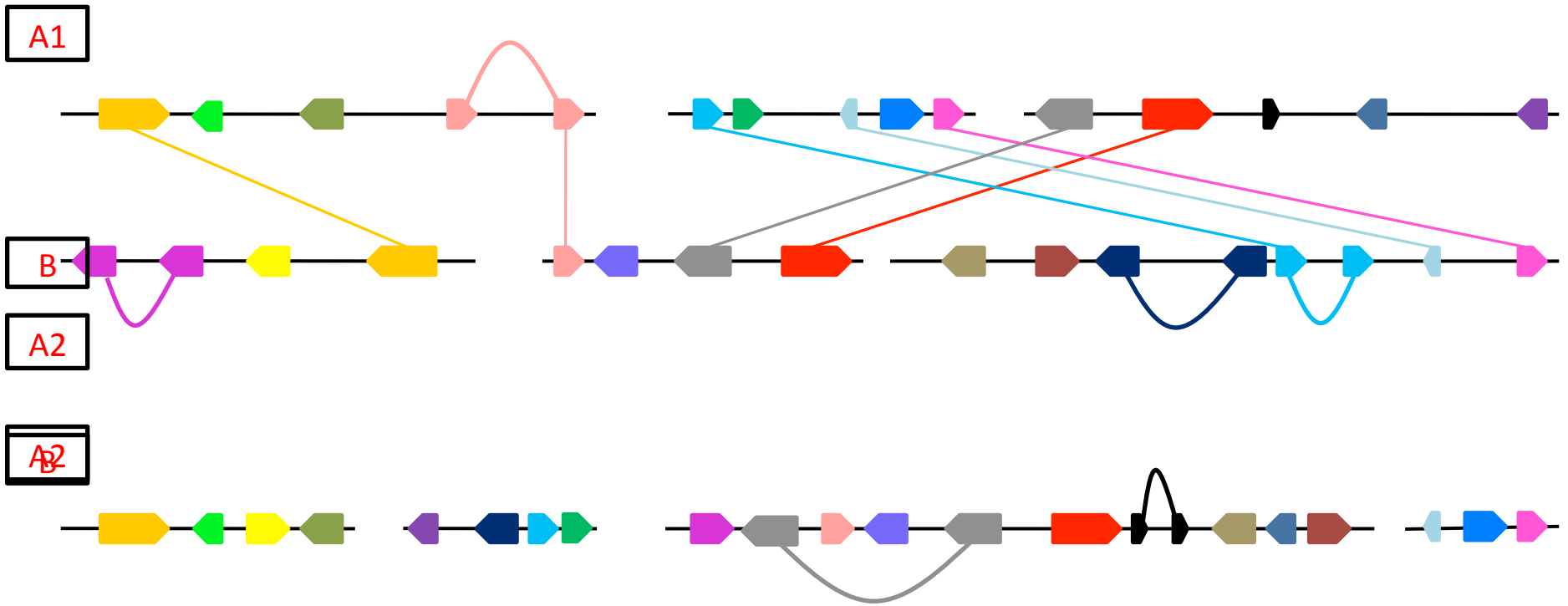
Evolution after Whole Genome Duplication

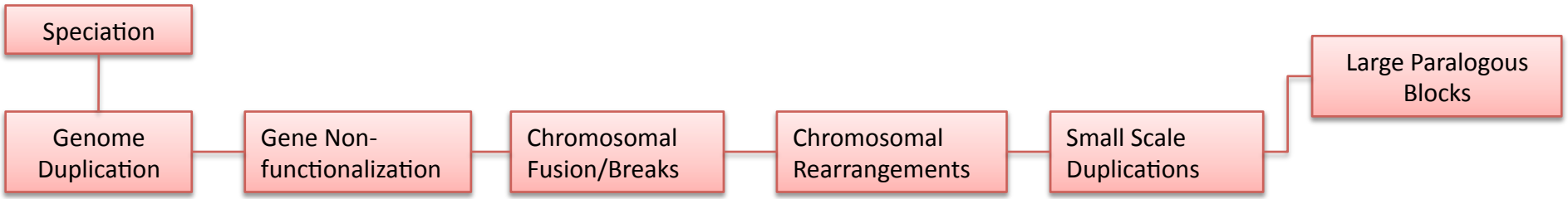




Courtesy of Param-Priva Singh

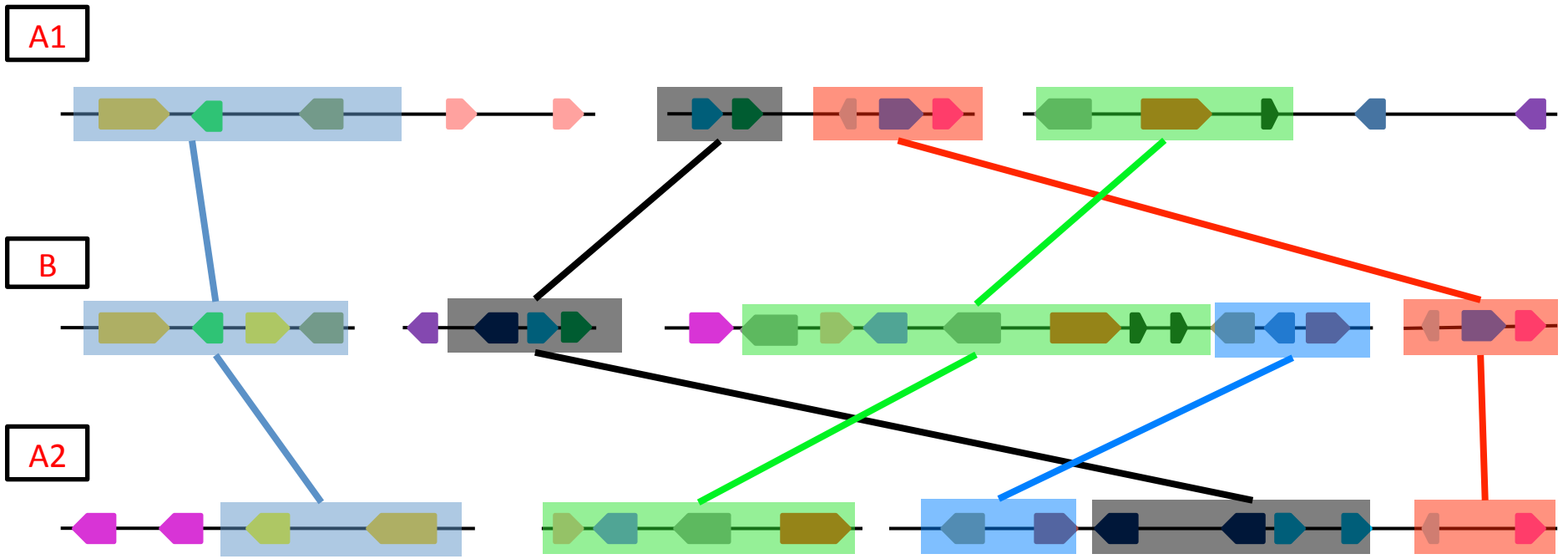
Evolution after Whole Genome Duplication

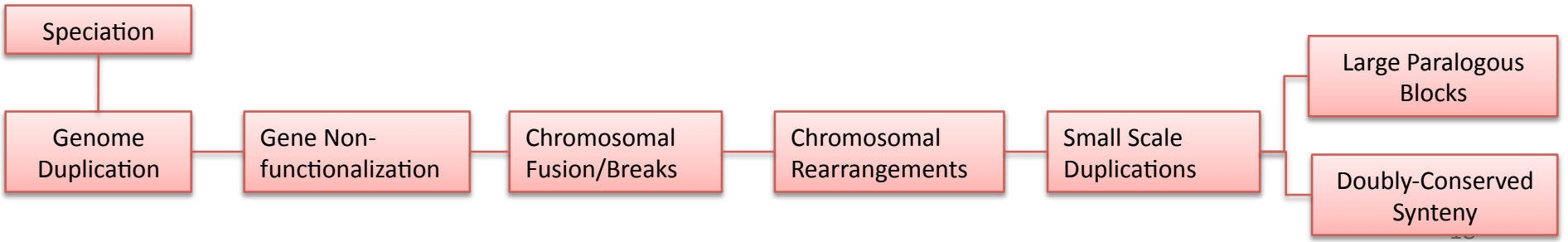




Courtesy of Param-Priva Singh

Evolution after Whole Genome Duplication





Courtesy of Param-Priva Singh

Genome rearrangement has a huge state space

- DNA sequences : 4 states per site
- Signed circular genomes with n genes:

$$2^{n-1}(n-1)!\text{states, 1 site}$$

- Circular genomes (1 site)
  - with 37 genes (mitochondria): 2.56×10^{52} states
  - with 120 genes (chloroplasts): 3.70×10^{232} states

Why use gene orders?

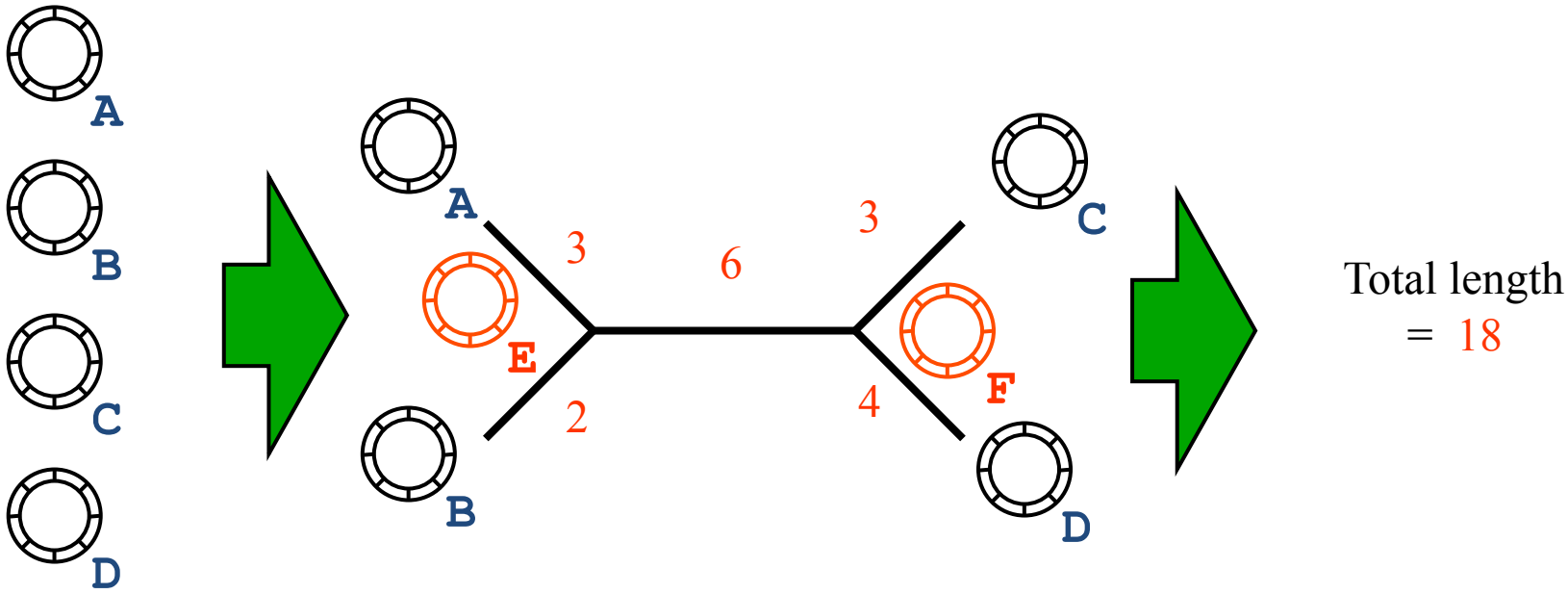
- “Rare genomic changes”: huge state space and relative infrequency of events (compared to site substitutions) could make the inference of deep evolution easier, or more accurate.
- Our research shows this is true, but accurate analysis of gene order data is computationally very intensive!

Phylogeny reconstruction from gene orders

- Distance-based reconstruction: estimate pairwise distances, and apply methods like Neighbor-Joining
- "Maximum Parsimony": find tree with the minimum length (inversions, transpositions, or other edit distances)
- Maximum Likelihood: find tree and parameters of evolution most likely to generate the observed data

Maximum Parsimony on Rearranged Genomes

- The leaves are rearranged genomes.
- Find the tree that minimizes the total number of rearrangement events (e.g., inversion phylogeny minimizes the number of inversions)

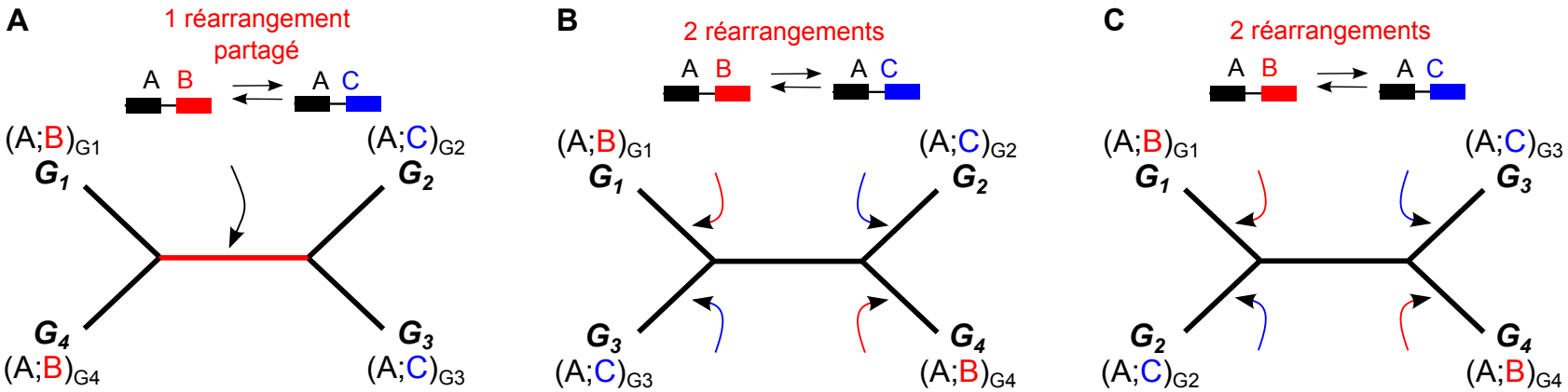


Limitations and ongoing research

- Current methods are mostly limited to single chromosomes with equal gene content (or very small amounts of deletions and duplications).
- Guénola Drillon (2012) developed a reliable distance based method for multiple chromosomes with unequal gene content (tests on real and simulated data show high accuracy)
- Handling multiple chromosome case is hard

Idée de base: regarder les adjacences des bloques

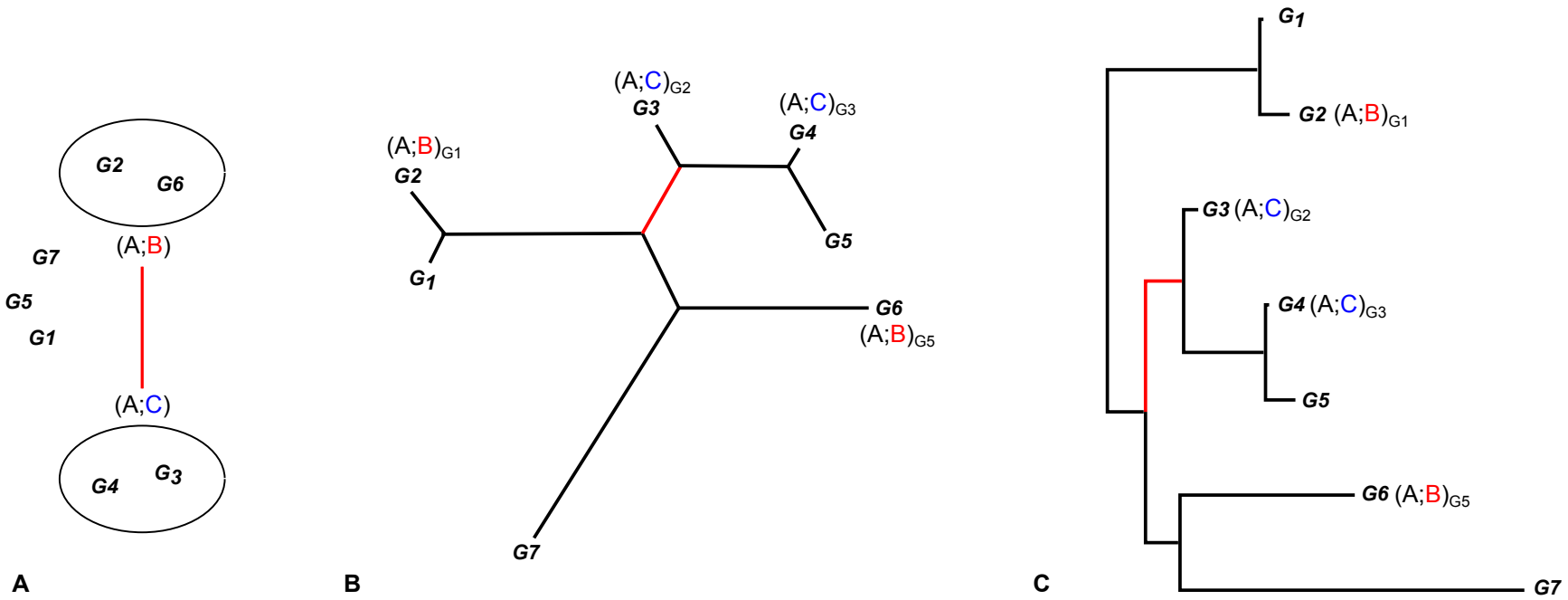
Etant donné 4 génomes qui partagent les mêmes blocs, et deux types d'adjacences au bloc A, une par B et l'autre par C, on cherche la topologie de l'arbre que induit le nombre **minimale** de re-arrangements.



G_1, G_3 and G_2, G_4 sont des groupes **de génomes incompatibles**

Etant donnés les génomes $(A;B)_{G1}$ $(A;B)_{G2}$ $(A;C)_{G3}$ $(A;D)_{G4}$, n'importe quel topologie aurait induit au moins 2 re-arrangements et donc l'impossibilité d'en « choisir » une.

N génomes et m blocs communs

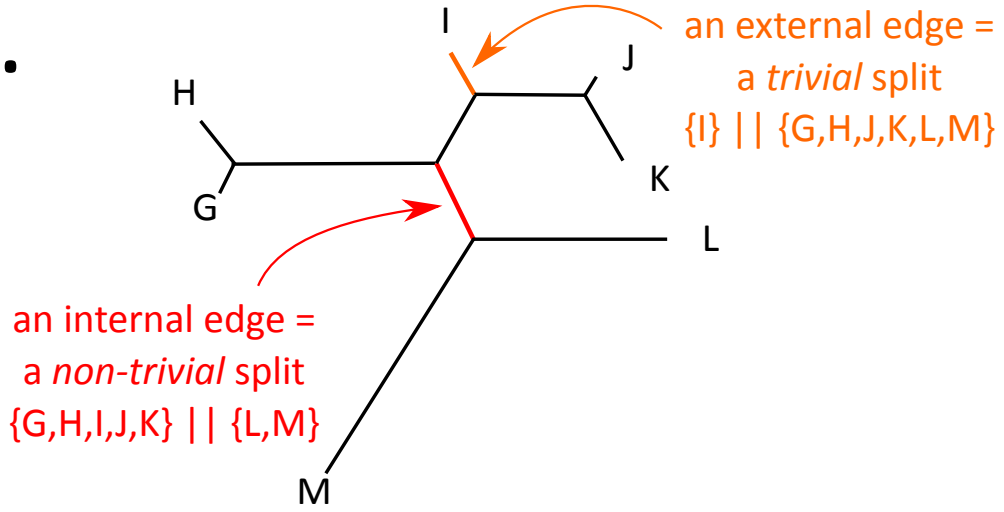


Dans la reconstruction de la phylogénie, **seulement un sous-ensemble** des génomes sera concerné: celui qui partage les adjacences (A;B) et (A;C).

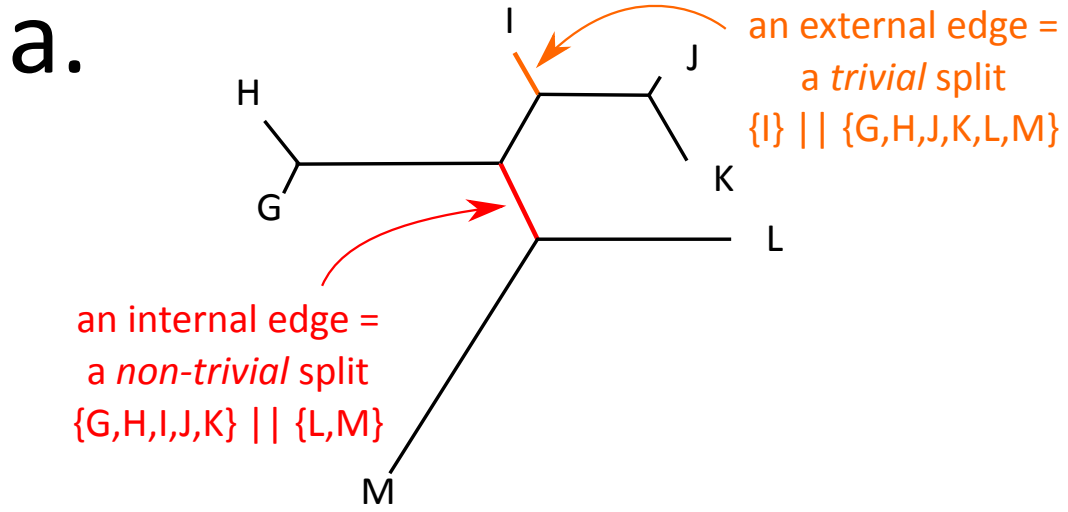
Pour G_1 G_5 et G_5 G_7 on doit supposer qu'il n'y a pas de (A;X) associé, avec un même X pour les deux. Ceci serait en contradiction avec la topologie de l'arbre choisie.
Aussi, il faut que jamais G_6 et G_7, par exemple, se retrouvent dans des ensembles incompatibles par rapport à d'autres adjacences (Y;Z).

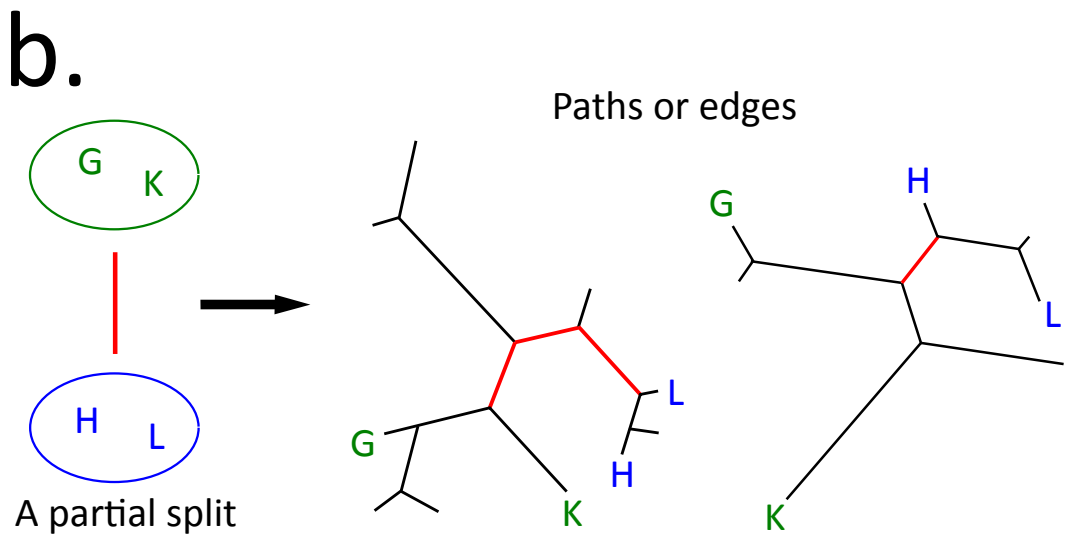
From splits to partial splits

a.

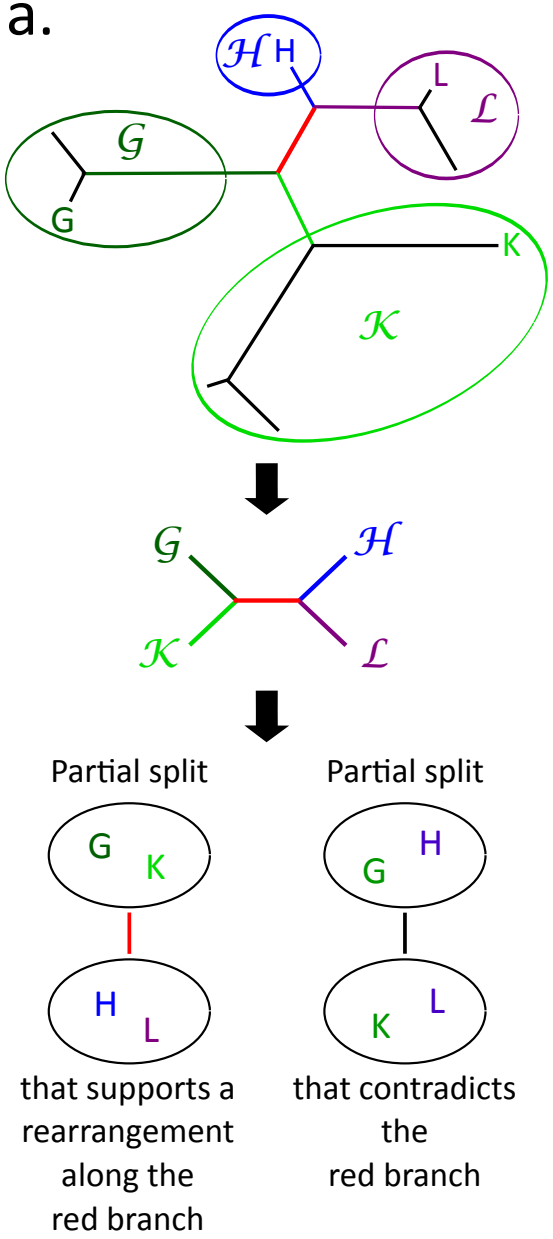


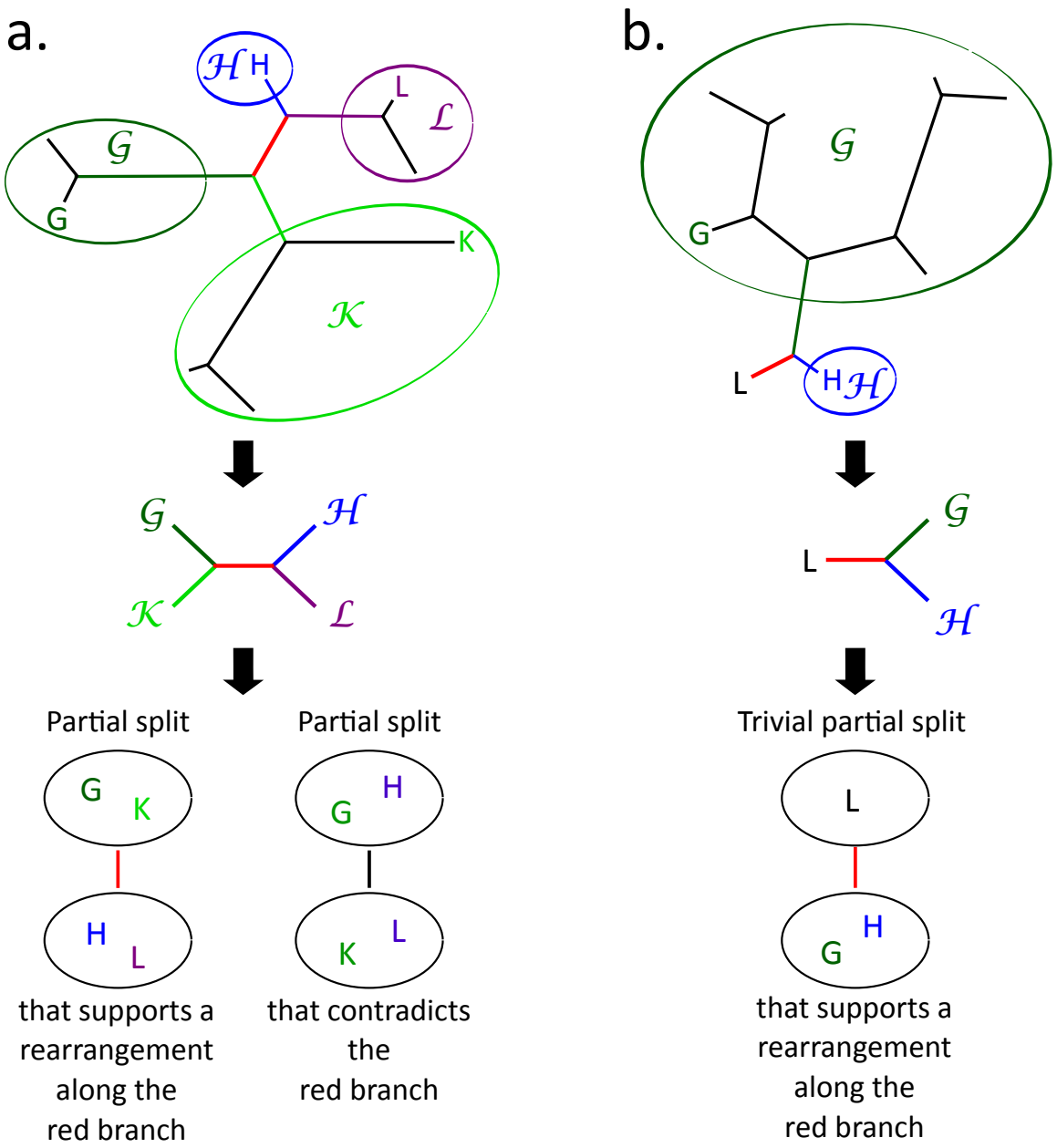
From splits to partial splits





a.


H
L
G
K
Partial split
Partial split
that supports a rearrangement along the red branch
that contradicts the red branch


a.
b.
Partial split
Partial split
Trivial partial split
that supports a rearrangement along the red branch
that contradicts the red branch
that supports a rearrangement along the red branch

Idée de l'algorithme

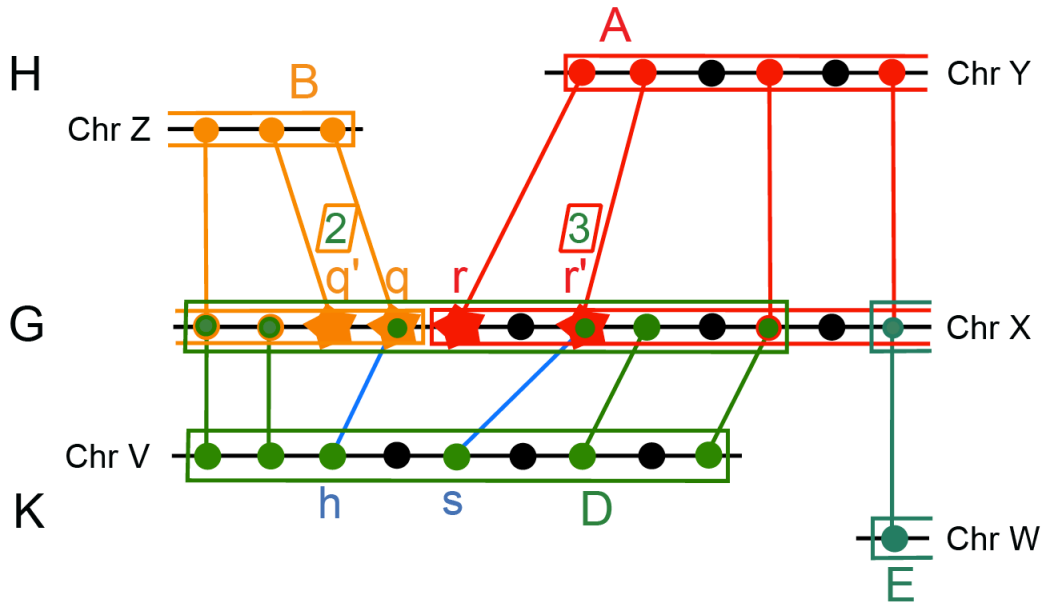
1. Regrouper les génomes qui se sont retrouvés le moins de fois au sein du même groupe de génomes incompatibles, par rapport aux différentes adjacences.
2. Definir les distances par paires de genomes en utilisant les partial splits.
3. Utiliser ces distances pour reconstruire iterativement et bottom-up l'arbre. **Attention**: on recalcule le nombre des incompatibilités en supprimant les groupes qui n'apportent plus d'information et ceux qui sont contradictoires avec le noeud fraîchement reconstruit.

L'algorithme minimize le nombre d'incompatibilités.

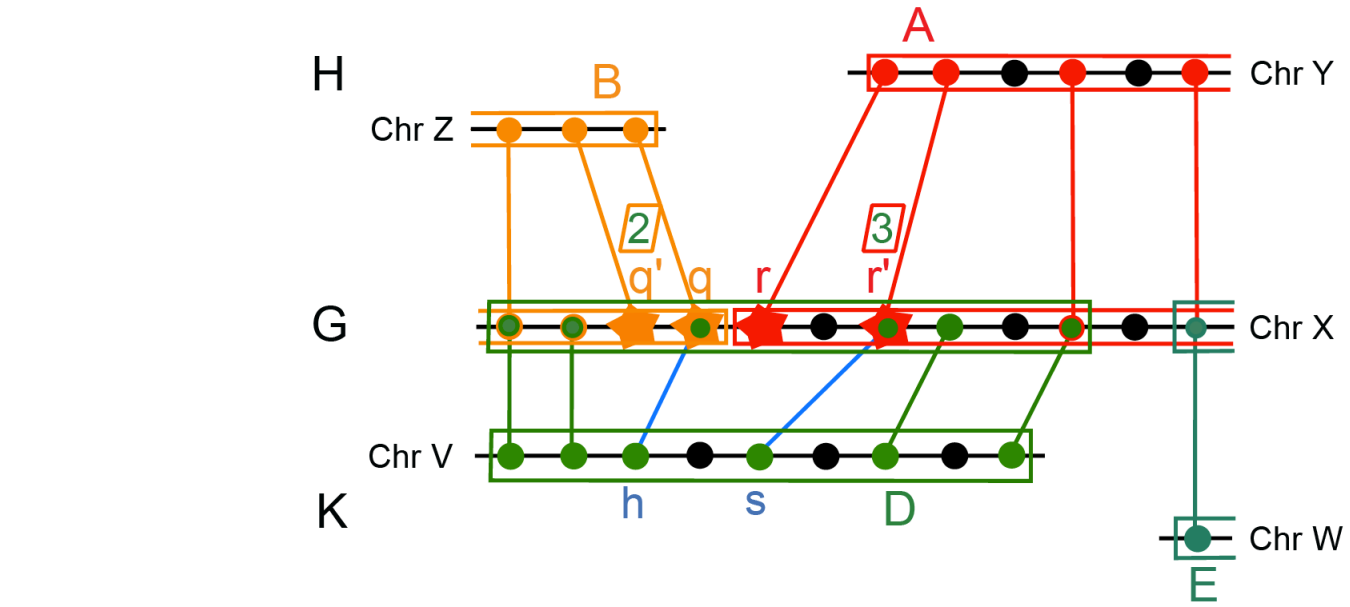
Les génomes utilisés jusqu'ici partagent les même ensembles de blocs de synténie.

Si on garde cette contrainte, on ne pourra reconstruire que des arbres phylogénétiques associés à des espèces **assez proches** entre elles.

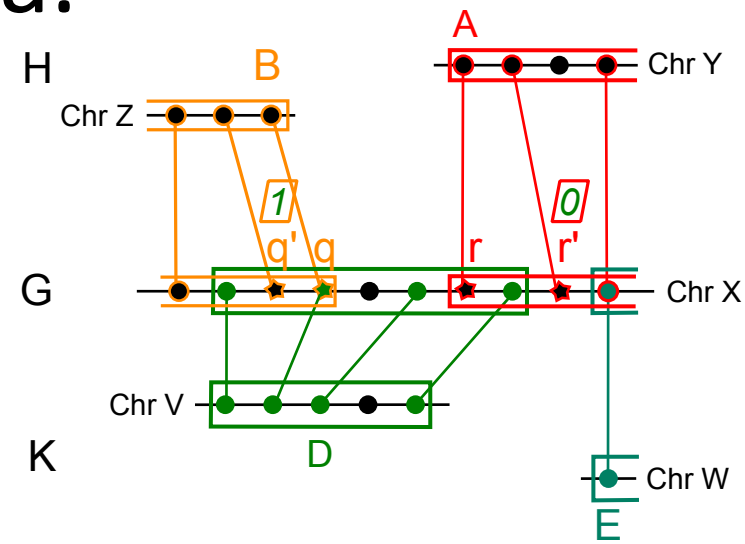
Etant donnés G_1, G_2, G_3 et la comparaison G_1/G_2, il est possible de comparer G_1/G_2 à G_3 en utilisant G_1/G_3 et G_2/G_3. Plus précisement on veut arriver à dire si une adjacence définie sur G1, par deux blocs de synténie issus de la comparaison G_1/G_2 est également présente dans d'autres génomes G_3, G_4, G_5….



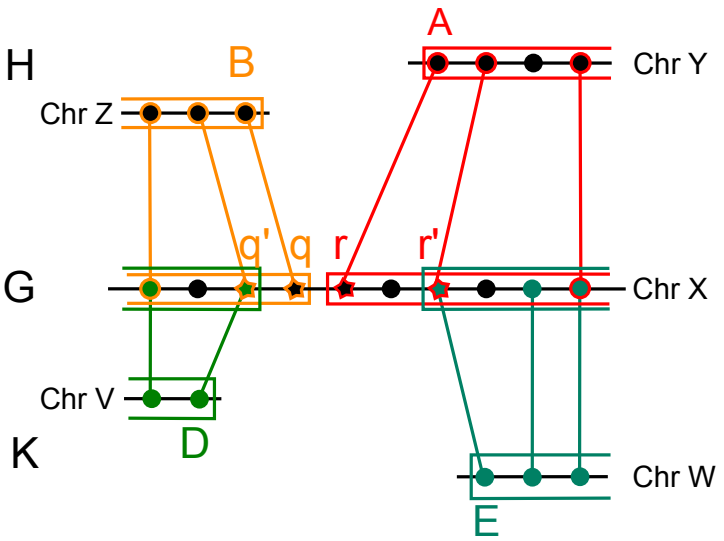
Do we have signals in K of the junction (B;A)?


H
B
A
Chr Y
Chr Z
2
3
q'
q
r
r'
G
Chr X
Chr V
h
s
D
K
Chr W
E

a.


H
B
A
Chr Y
Chr Z
1
0
q'
q
r
r'
G
Chr X
Chr V
D
K
Chr W
E

b.

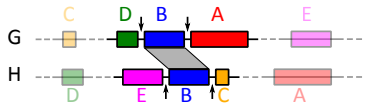

H
B
A
Chr Y
Chr Z
q'
q
r
r'
G
Chr X
Chr V
D
K
Chr W
E

I. Identification of breakpoints

> for each pairwise comparison G/H, over n genomes

> for each block B along G

1. Identification of two breakpoints: $[(D\ B)_G,(E\ B)_H]$ and $[(B\ A)_G,(B\ C)_H]$

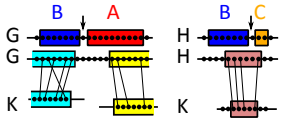


II. Identification of partial splits

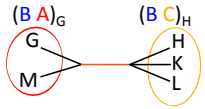
> for each breakpoint $[(B\ A)_G,(B\ C)_H]$ (or $[(D\ B)_G,(E\ B)_H]$)

> for each genome K (with K≠G,H)

2. Evaluation of whether or not $(B\ A)_G$ and $(B\ C)_H$ belong to K based on G/K and H/K comparisons, respectively.



3. Definition of a partial split $S_{(B\ A)}\ ||\ S_{(B\ C)}$ involving G, H and at least one other genome in each set

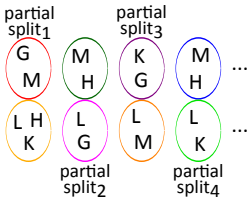


III. Bottom-up tree reconstruction

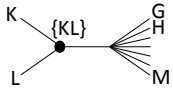
> while #genomes >3

> for each pair of genomes G, H

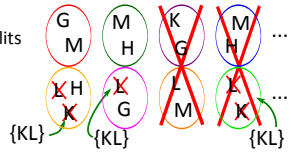
4. Computation of :
$f_{inc}(G,H)$: # of times that G and H are found into two distinct sets of partial split
$f_{comp}(G,H)$: # of times that G and H are grouped together in a partial split



5. Creation of an internal node {KL} associated to the genomes with smallest f_{inc} [chosen among the #genomes/2 with smallest ratios $(f_{inc}+1)/(f_{comp}+1)$]



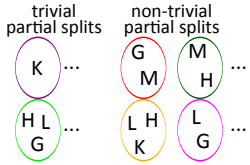
6. Updating the list of partial splits
-> replacement of K and L by the new genome {KL}
-> #genomes decreases of 1

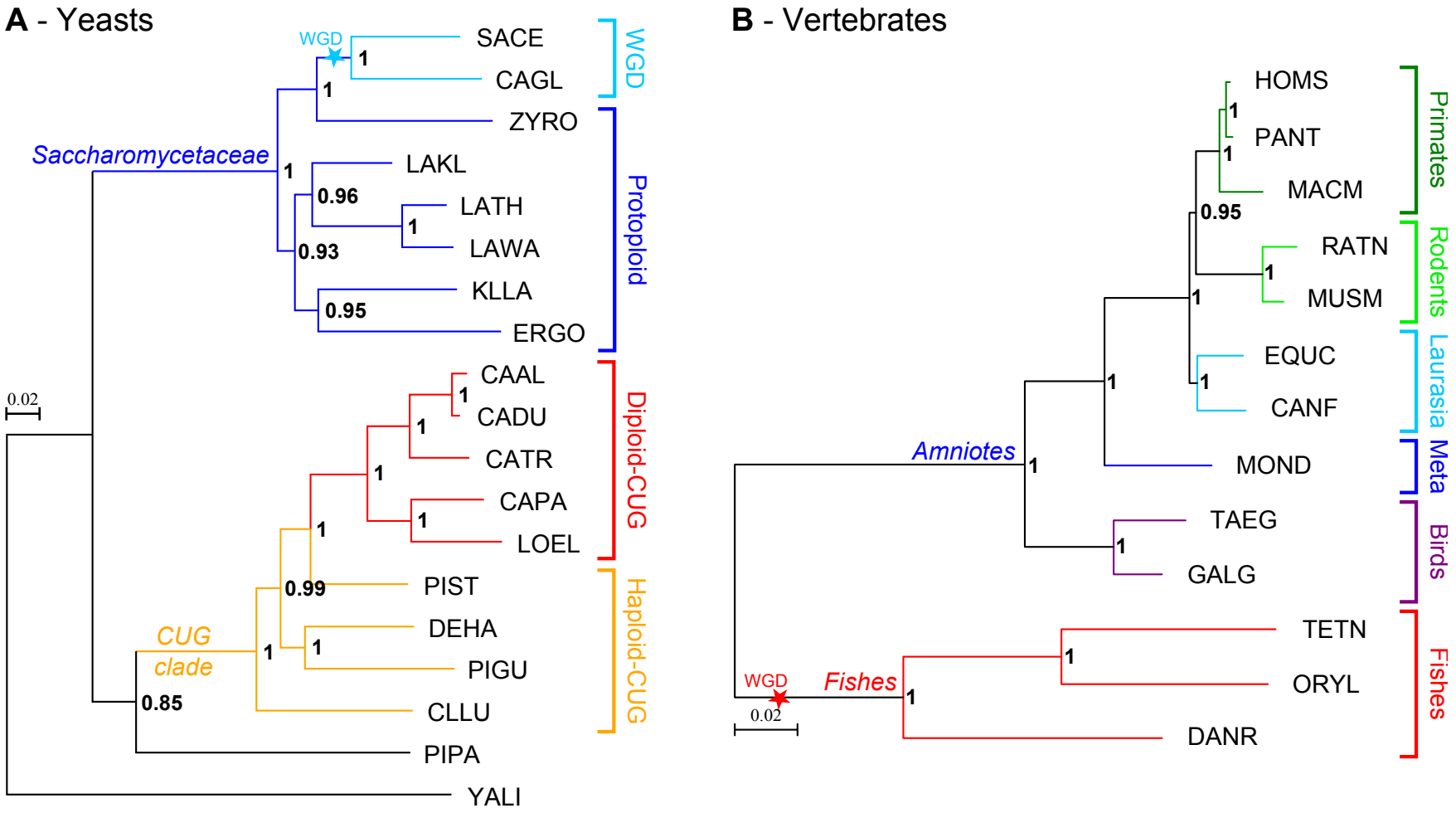


IV. Estimations on the branches of the tree

> for each branch

7. Computation of:
- branch length if external, from trivial partial splits;
- branch length and confidence score if internal, from non-trivial partial splits

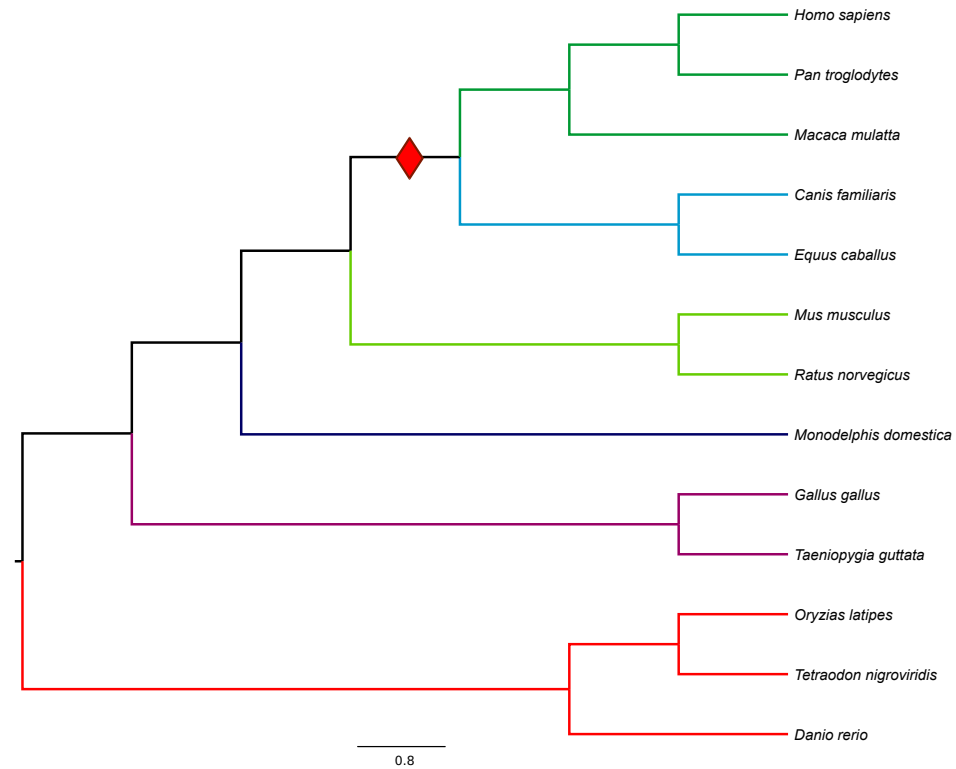



A - Yeasts
WGD
SACE
CAGL
ZYRO
Saccharomycetaceae
LAKL
LATH
LAWA
KLLA
ERGO
CAAL
CADU
CATR
CAPA
LOEL
PIST
DEHA
PIGU
CLLU
PIPA
YALI
CUG
clade
0.02
WGD
Protoploid
Diploid-CUG
Haploid-CUG
B - Vertebrates
HOMS
PANT
MACM
RATN
MUSM
EQUC
CANF
MOND
TAEG
GALG
TETN
ORYL
DANR
Amniotes
Fishes
WGD
0.02
Primates
Rodents
Laurasia
Meta
Birds
Fishes

Robustness of the construction wrt synteny blocks definition

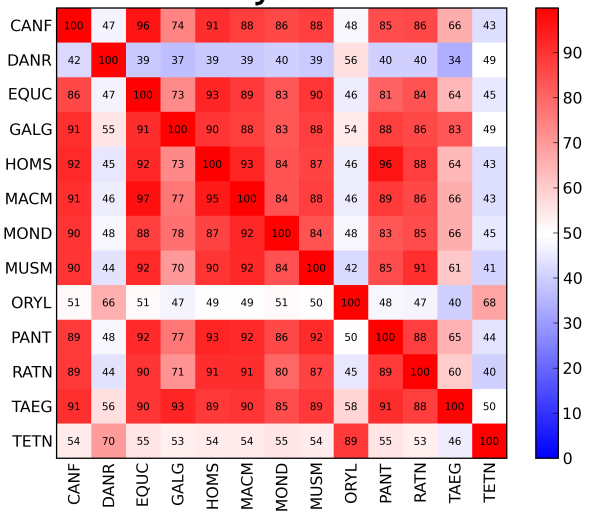
Synteny blocks coverage of the genomes after pairwise comparison

a.



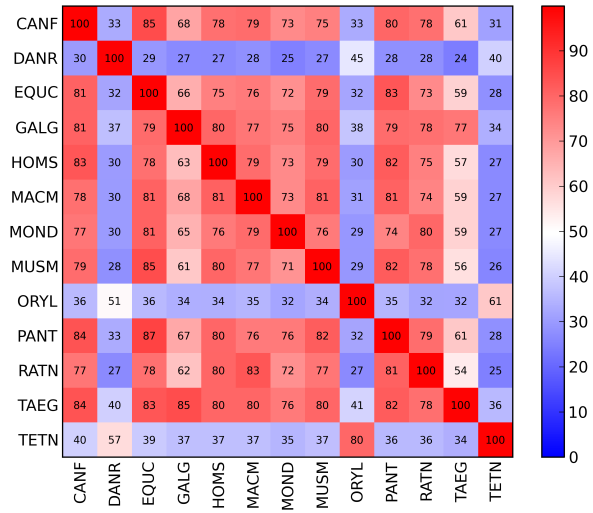
b.

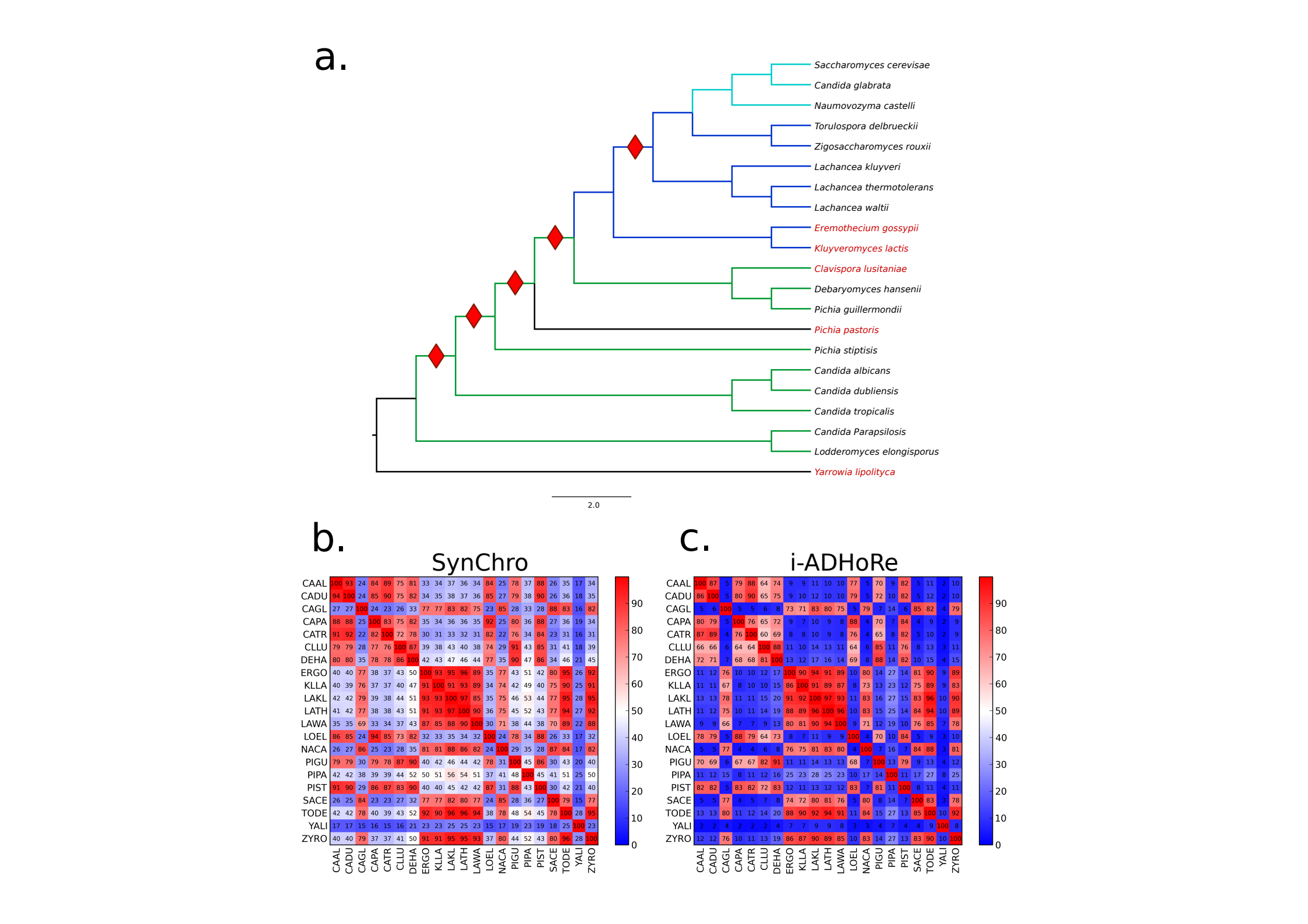
SynChro

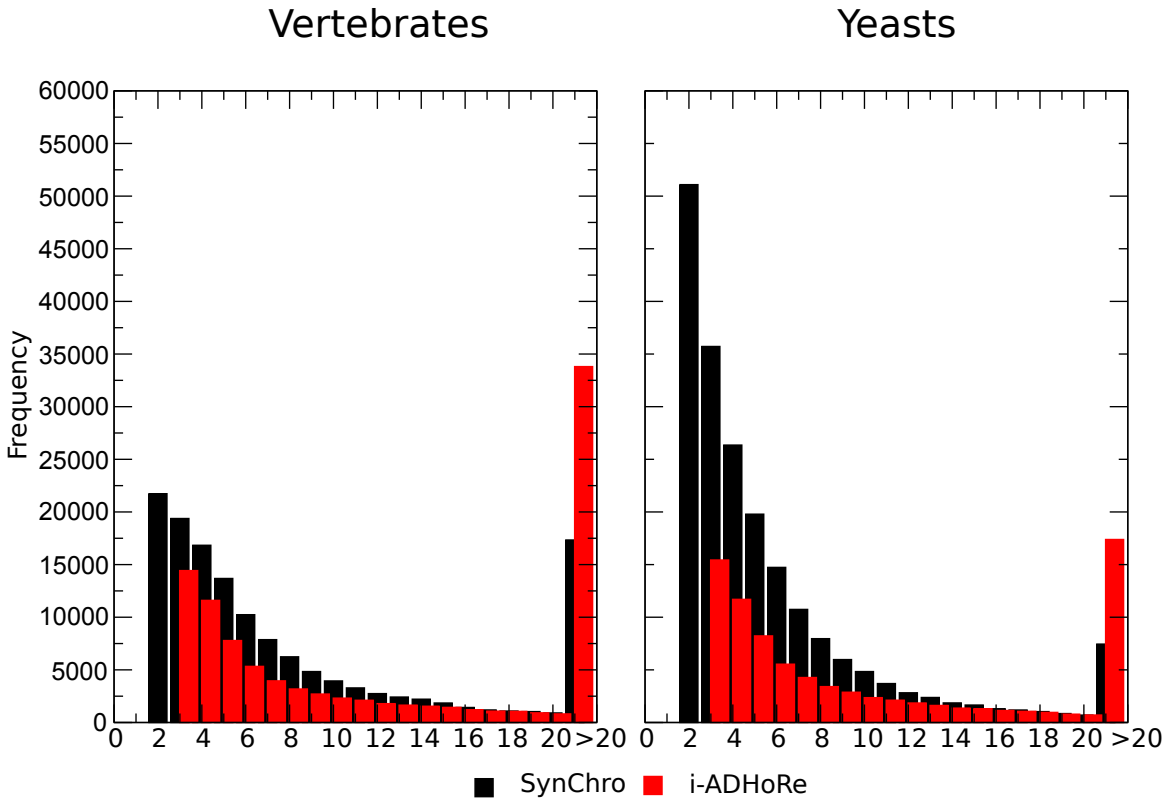


c.

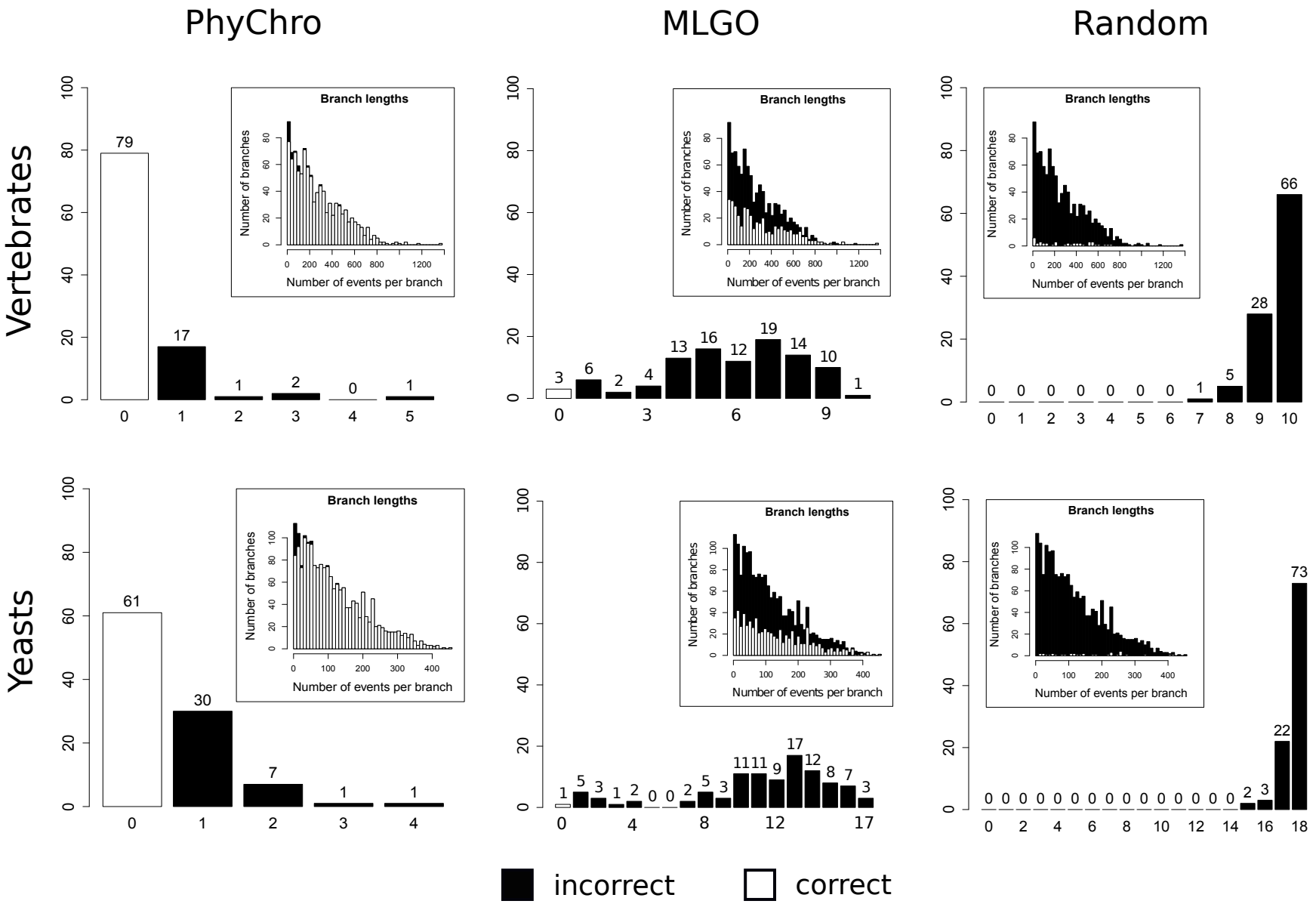
i-ADHoRe




a.
Saccharomyces cerevisae
Candida glabrata
Naumovozyma castelli
Torulospora delbrueckii
Zigosaccharomyces rouxii
Lachancea kluyveri
Lachancea thermotolerans
Lachancea waltii
Eremothecium gossypii
Kluyveromyces lactis
Clavispora lusitaniae
Debaryomyces hansenii
Pichia guillermondii
Pichia pastoris
Pichia stiptisis
Candida albicans
Candida dubliensis
Candida tropicalis
Candida Parapsilosis
Lodderomyces elongisporus
Yarrowia lipolityca
2.0
b.
SynChro
c.
i-ADHoRe
CAAL
CADU
CAGL
CAPA
CATR
CLLU
DEHA
ERGO
KLLA
LAKL
LATH
LAWA
LOEL
NACA
PIGU
PIPA
PIST
SACE
TODE
YALI
ZYRO
0
10
20
30
40
50
60
70
80
90


Vertebrates
Yeasts
Frequency
60000
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0 2 4 6 8 10 12 14 16 18 20 >20
0 2 4 6 8 10 12 14 16 18 20 >20
SynChro
i-ADHoRe

Generation of 100+100 simulated trees



Phylogenetic reconstruction based on synteny blocks and gene adjacencies
G.Drillon, R.Champeimont, F.Oteri, G.Fischer, A.Carbone
Submitted, 2014.