Phylogenetic reconstruction based on synteny block and gene adjacencies

Guénola Drillon[1], Raphaël Champeimont[1], Francesco Oteri[1], Gilles Fischer[1], and Alessandra Carbone[1,2]

[1] Sorbonne Universités, UPMC-Univ P6, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative - UMR 7238, 4 Place Jussieu, 75005 Paris, France

[2] Institut Universitaire de France, Paris 75005, France

Alessandra.Carbone@lip6.fr

November 13, 2019

Abstract

Gene order can be used as an informative character to reconstruct phylogenetic relationships-between species independently from the local information present in gene/protein sequences.

PhyChro is a reconstruction method based on chromosomal rearrangements, applicable to a wide range of eukaryotic genomes with different gene contents and levels of synteny conservation. For each synteny breakpoint issued from pairwise genome comparisons, the algorithm defines two disjoint sets of genomes, named partial splits, respectively supporting the two block adjacencies defining the breakpoint. Considering all partial splits issued from all pairwise comparisons, a distance between two genomes is computed from the number of partial splits separating them. Tree reconstruction is achieved through a bottom-up approach by iteratively grouping sister genomes minimizing genome distances. PhyChro estimates branch lengths based on the number of synteny breakpoints and provides confidence scores for the branches.

PhyChro performance isevaluatedon two datasets of 13 vertebrates and 21 yeast genomes by using up to 130 000 and 179 000 breakpoints respectively, a scale of genomic markers that has been out of reach until now. PhyChro reconstructs very accurate tree topologies even at known problematic branching positions. Its robustness has been benchmarked for different synteny block reconstruction methods. On simulated data PhyChro reconstructs phylogenies perfectly in almost all cases, and shows the highest accuracy compared to other existing tools. PhyChro is very fast, reconstructing the vertebrate and yeast phylogenies in less than 15 min.

Availability: PhyChro will be freely available under the BSD license after publication

Contact: alessandra.carbone@lip6.fr

Key words: phylogenetic tree ; chromosomal rearrangement ; synteny block ; adjacency ; breakpoint ; parsimony ; distance ; yeast ; vertebrate ; split.

Introduction

Today, phylogenies of many species can be reconstructed using sequences from numerous proteins, but, despite the availability of a considerable amount of sequence data, reconstructions are not always accurate and can result in incongruent topologies (Philippe et al., 2011). These limitations are partly due to methodological artifacts such as sequence misalignment (different software gives significantly different alignments (Wong et al., 2008)), false-orthologous gene assignment (due to horizontal transfer, gene duplication/loss events (Bapteste et al., 2004)) and homoplasy inherent to the data. These limitations prompted phylogeneticists to explore different types of signal representing rare genomic changes, such as intron indels, retroposon integrations, changes in organelle gene order, gene duplications and genetic code variants (Rokas and Holland, 2000). Although these genomic changes can be useful to validate some topological uncertainties, they have never been used to reconstruct complete phylogenies at the exception of the coherent mitochondrial phylogeny based on gene composition and gene order of mitochondrial genomes (Sankoff et al., 1992). This result offered, for the first time, a strong validation of the hypothesis that the macrostructure of mitochondrial genomes contains quantitatively meaningful information for phylogenetic reconstruction.

Gene order along nuclear chromosomes follows different evolutionary trends than along mitochondrial genomes (Burger et al., 2003), and it has been observed in several occasions that it comprises useful evolutionary information for phylogenetic reconstruction (Boore, 2006; Fertin, 2009). Many methods aiming at exploiting this trait as phylogenetic signal have been developed. They all belong to one of the four classical methodological categories *i.e.* the distance-based methods (Moret et al., 2001a; Wang et al., 2006; Guyon et al., 2009; Luo et al., 2012; Lin et al., 2012), the maximum-parsimony-based methods (Sankoff and Blanchette, 1998; Cosner et al., 2000; Moret et al., 2001b; Bourque and Pevzner, 2002; Tang and Moret, 2003; Bergeron et al., 2004; Zheng and Sankoff, 2011; Xu et al., 2011), the maximum-likelihood-based methods (Larget et al., 2005; Hu et al., 2011; Lin et al., 2013; Feng, 2017) and the quartet-based methods (Liu et al., 2005). Whether they are applied to sequences or to gene orders, these methodological categories harbor a variety of intrinsic limitations: computational complexity, sensitivity to short and long branch attraction (Felsenstein, 1978), requirement for good evolutionary models (Yang and Rannala, 2012), etc. Moreover, gene order-based methods were so far mainly applied to small bacterial or organelle genomes or to highly colinear genomes. The first phylogenetic reconstruction of eukaryotic nuclear genomes harboring different gene contents and different levels of synteny conservation was applied to the very large evolutionary span covered by the super-group of Unikonts and did not assess the performance of the method at known difficult branching positions such as the position of Rodentia relative to Primates and Laurasiatheria (Xu et al., 2011), or the position of *Candida glabrata* in Saccharomycetaceae (Lin et al., 2013; Hu et al., 2014). A recent improvement of this method taking into account balanced rearrangements, insertions, deletions, and duplications into an evolutionary model based on the principle of Double Cut and Join was applied to the phylogenetic reconstruction of 20 yeast species. It achieved accurate phylogeny reconstruction although the tree topology showed a couple of disagreements with previously published phylogenies (Feng, 2017).

We developed PhyChro with the aim of making the most of the evolutionary information derived from chromosome rearrangements. PhyChro is applied to 13 vertebrate and 21 Saccharomycotina yeast genomes and it reconstructs very accurate tree topologies even at known difficult branching positions.

New approaches

PhyChro is a method for phylogenetic reconstruction based on synteny block and gene adjacencies. It relies on two important specificities. First, it uses synteny block adjacencies computed for all possible pairwise combinations of species instead of using synteny blocks universally shared by all the species involved in the reconstruction. This pairwise approach has the advantage to efficiently compare genomes with different levels of synteny conservation, without losing the wealth of synteny information that is shared by most closely related genomes. Second, PhyChro achieves tree reconstruction using the idea that for each synteny breakpoint, (a subset of the) genomes can be split into two disjoint groups depending on whether they support one block adjacency defining the breakpoint or the other. Formally, PhyChro relies on partial splits (Semple & Steel , 2001; Huson et al., 2004; Huber et al., 2005) (**Figure 1**), a generalization of the notion of split used in quartet-based methods. By exploiting partial splits associated to all identified breakpoints, PhyChro defines a distance between genome pairs, called *Partial Split Distance* (PSD), by counting the number of times that two genomes belong to different subsets of a partial split. Note that PSD is a measure defined on a set of n genomes contrary to other previously introduced distance measures based on the comparison of only two genomes at a time. Based on PSD, PhyChro reconstructs tree topologies with a bottom-up approach, by iteratively identifying those sister genomes that minimize the number of times they belong to different subsets of a partial split.

Intuitively, sister genomes are pairs of genomes sharing a high number of gene adjacencies at breakpoint positions. One can think to these pairs of genomes as being located close to each other but also as being located further away from all other genomes. Based on these intuitions, PhyChro (i) focuses on chromosomal rearrangement events supporting internal branches (useful for topology reconstruction) while ignoring all events that occurred on external branches (of no use for topology reconstruction), and (ii) minimises the differences between sister genomes, that is genomes separated by no internal branch, rather than maximising their similarity.

Contrary to distance-based methods, each pairwise distance depends on all genomes (as it depends on breakpoints identified through all genome comparisons) and, at each iteration, PhyChro recomputes distances from scratch between all pairs of genomes not yet included in the reconstruction. This iterated updating, affecting all entries of the distance matrix, is original to PhyChro and absent in distance-based methods. The Neighbor Joining (NJ) algorithm encodes the somewhat similar idea that pairs of genomes need not only be close to each other but also be distant from all others to be considered first in the reconstruction. This second condition is explicitly handled by the NJ algorithm, while PhyChro encodes it directly in its definition of genome distance. In conclusion, PhyChro is an algorithm whose basic data structure is the partial split and whose computational model is a bottom-up iterative reconstruction of the tree based on genome distances. These distances are computed by successive approximations, after the iterative elimination of inconsistencies in the set of partial splits.

PhyChro provides estimations for branch length and branch robustness. Extensive details on the algorithm and on the notions on which it relies are provided in the Materials and Methods section.

Results

Phylogenetic reconstruction of yeast and vertebrate species

We tested PhyChro on two different sets of species comprising 21 yeast and 13 vertebrate genomes. They harbor very different genome characteristics (in terms of genome size, number and density of genes, etc) as well as very different modes of chromosome evolution (number and rates of rearrangements, proportions of inversions versus translocations, whole genome duplication events, etc) (Drillon and Fischer, 2011). Previous analyses using the global level of divergence of orthologous proteins revealed that the evolutionary range covered by the Saccharomycotina subphylum exceeds that of vertebrates and is similar to the span covered by the entire phylum of Chordata (Dujon, 2006). Moreover, for both clades, the level of synteny conservation is highly variable between subclades with only 50% of genes belonging to synteny blocks between Amniota and fishes, or between yeast species from the Protoploid and CUG clades, while more than 95% of genes are conserved in synteny between Primates or between closely related species within the CUG clade (Drillon and Fischer, 2011). Finally, phylogenetic reconstructions in these two groups of species contain some ambiguous branching positions (sometimes controversial in the past), such as the position of Rodentia in the vertebrate tree or the position of *Candida glabrata* in the Saccharomycetaceae family of yeast, that we were interested to test with PhyChro.

We applied PhyChro on the sets of synteny blocks reconstructed with SynChro (Drillon et al., 2013, 2014) (see Methods) that resulted from genome pairwise comparison of the two sets of vertebrate and yeast species. The resulting tree topologies were compared to the reconstructions obtained with existing methods based on protein sequence comparisons, including PhyML, a maximum-likelihood-based method, ProtPars, a maximum-parsimony-based method, and Neighbor, a distance-based method (Guindon and Gascuel, 2003; Felsenstein, 1989) (see Methods).

The tree topology reconstructed by PhyChro for the 13 vertebrate species (**Figure 2a**) is identical to the topology produced by PhyML on 389 families of orthologs (illustrated in **Figure S1**). The position of Rodentia is correctly located, closer to Primates than to the Laurasiatheria. By comparison, ProtPars and Neighbor do not correctly place Rodentia (**Figure S1**). It should be noticed that PhyChro succeeded in correctly placing the rodent branch in the tree despite the fact that no partial split supports the existence of the branch splitting Primates and Rodentia from the other species. This is due to the fact that PhyChro, contrary to the other methods, does not construct the tree by identifying well supporting branches; rather, it avoids creating branches that are contradicted. This strategy allows PhyChro to treat difficult cases generated by small branches and characterised by very few rearrangements. In the specific reconstruction of Rodentia positioning, the detection of the short branch preceding their splitting with Primates, is rendered even more difficult by the important evolutionary history of Rodentia that likely erased the traces of the plausibly few ancestral rearrangements of Primates and Rodentia (see long branches in **Figure 2a**). PhyChro corresponding branch length equals zero and its confidence score cS, which assesses the robustness of the branch, is close to 0 (0.03, **Figure 2a**).

In several ways, the tree topology reconstructed by PhyChro for the 21 yeast species is more accurate than the topologies obtained with either one of the 3 phylogenetic reconstruction methods, based on protein sequence comparison (**Figure 2b** and **Figure S1**). The first difference concerns the position of *Candida glabrata* relatively to *Saccharomyces cerevisiae* and *Naumovozyma castellii* (formely known as *Saccharomyces castellii*). It is known that phylogenies based on protein sequence analysis tend to artefactually place *C. glabrata* outside from *N. castellii* and *S. cerevisiae* (Kurtzman and Robnett, 2003; Hittinger et al., 2004) due to the short/long branch attraction problem (**Figure S1**). Previous studies based on shared patterns of gene losses and rearrangements showed that in fact, *N. castelli* is an outgroup to a clade containing *S. cerevisiae* and *C. glabrata* (Scannell et al., 2006; Gordon et al., 2009). Using the same macro-organisational information, PhyChro correctly recapitulates the phylogeny for these 3 species, despite the very long terminal branch length leading to *C. glabrata* present in its tree (**Figure 2b**). It should be considered that PhyChro reconstruction is automatic while the two previous ancestral gene ordering reconstructions have been manually derived.

In addition, note that PhyML erroneously locates *Pichia pastoris* as an outgroup while *P. pastoris* correctly branches at the root of the CUG clade according to PhyChro, Neighbor and ProtPars. Neighbor erroneously locates *Pichia stipitis* as a sister genome of *Debaryomyces hansenii* while *P. stipitis* is correctly positioned by PhyChro, PhyML and ProtPars. Concerning ProtPars, it erroneously splits the clade containing *Kluyveromyces lactis* and *Eremothecium gossypii* while the clade is correctly reconstructed by PhyChro, PhyML and Neighbor (**Figure 2b** and **Figure S1**). In all these instances, PhyChro outperforms the 3 classical methods based on protein sequence comparison.

The only topological uncertainty that remains corresponds to the position of *Clavispora lusitaniae*. According to PhyChro, this species branches as a sister genome to the clade containing *D. hansenii* and *Pichia guillermondii* (**Figure 2**), while according to PhyML and ProtPars, *C. lusitaniae* branches at the root of the CUG clade. Moreover Neighbor produces a third topology in this region of the tree (**Figure S1**). The confidence scores of the *C. lusitaniae* branch given by PhyChro, PhyML and ProtPars show uncertainties (0.33, 0.96 and 0.97, respectively) demonstrating that the topology associated to this branch remains doubtful.

Branch length estimates provided by PhyChro give interesting information notably for subclades where the synteny conservation is still high. For instance, the terminal branch length leading to the yeast *Lachancea thermotolerans* is computed to be very close to zero (0.33) showing that at most 1 rearrangement (larger than a six genes inversion) occurred in this genome since its divergence from its last common ancestor with *Lachancea waltii*, while long branches such as the ones leading to *C. glabrata*, *Danio rerio* or to Rodentia indicate the accumulation of a large number of chromosomal rearrangements. Note that branch lengths are under estimated for very distant genomes such as *Yarrowia lipolytica* and *P. pastoris* (as they are involved in very few partial splits).

Comparison with MLGO, a gene-order based method for phylogenetic reconstruction

Currently, the only large-scale method to reconstruct gene order phylogenies is Maximum Likelihood for Gene Order Analysis (MLGO) (Lin et al., 2013). The two MLGO trees, issued from the same set of vertebrates and yeasts that we considered, are reported in **Figure S2**. These trees comprise a number of erroneous splits compared to the reference trees. We count two erroneous splits for vertebrates and seven for yeasts, contrary to PhyChro that reconstructs correctly both trees. For vertebrates the errors are due to the misplacements of *M. domestica* and Rodentia. For yeasts, *P. pastoris* is erroneously located closer to the Protoploid clade than to the CUG clade, *L. waltii* and the sister genomes *Torulaspora delbruechii* and *Zygosaccharomyces rouxii* are erroneously located in the Protoploid clade, and finally, *P. stipitis* is erroneously located in the CUG clade. As for PhyChro, *S. cerevisiae* and *C. glabrata* are correctly located.

Robustness of PhyChro

Robustness of PhyChro on different definitions of synteny blocks

To test the sensitivity of PhyChro to different definitions of synteny block, we generated two sets of synteny blocks by using SynChro (Drillon et al., 2014) and i-ADHoRe 3.0 (Proost et al., 2012) and produced the corresponding trees for vertebrate and yeast species. On vertebrates, PhyChro based on i-ADHoRe synteny blocks gives a tree with an erroneous split corresponding to the misplacement of Rodentia (see **Figure S3a**). On yeasts, we count five erroneous splits in the tree reconstruction (**Figure S4a**). These discrepancies are explained by the lower proportion of genomes recovered in the synteny blocks generated by i-ADHoRe than by SynChro, as illustrated in **Figures S3bc** and **S4bc**. A global comparison of block size distributions generated by i-ADHoRe and SynChro over all pairwise comparisons between vertebrate and yeast genomes, is reported in **Figure S5**. We observe that SynChro allows for small blocks made of only 2 genes (noted also in (Drillon et al., 2014)) while i-ADHoRe only allows blocks of at least 3 genes, and that the number of small blocks (< 21 genes) produced by SynChro is systematically larger than for i-ADHoRE. For pairs of genomes that underwent many rearrangements and, in consequence, would have a low synteny conservation, the small blocks detected by SynChro are expected to play a crucial role. This is visually observable in the matrices of **Figures S3bc** and **S4bc** showing higher synteny coverage (lighter blue and darker red colours) for SynChro than for i-ADHoRE for all species pairs. On the other hand, one observes that i-ADHoRE generates a greater number of large blocks (≥ 21) than SynChro (**Figure S5**). This ensures that for pairs of genomes for which synteny blocks allow for more than 60% coverage, SynChro and i-ADHoRE show comparable success, as illustrated by the red coloured cells in the matrices of **Figures S3bc** and **S4bc**. In conclusion, a better synteny coverage reached for all pairs of species allows PhyChro to perform better on SynChro than on i-ADHoRE blocks.

It is also interesting to note that modulating the size of micro-rearrangements tolerated within synteny blocks with the Δ parameter from SynChro (bigger the Δ, larger the micro-rearrangements tolerated) has an effect on the number of partial splits contradicting a given topology. For example, PhyChro run with $\Delta = 3$ (by default, see Methods) finds 36, 37 and 42 partial splits that contradict the ((Primates, Rodentia), Laurasiatheria), (Primates, (Rodentia, Laurasiatheria)) and ((Primates, Laurasiatheria), Rodent) topologies, respectively. By augmenting Δ to 4 (that is, being more tolerant for larger micro-rearrangements within synteny blocks), PhyChro finds 24, 37 and 53 contradictory partial splits, respectively. These numbers provide confidence in the ((Primates, Rodentia), Laurasiatheria) topology and, since none of the topologies has zero contradictions, they also show that homoplasy is present.

Robustness of the algorithm with respect to simulated genomes

In order to test PhyChro on a large set of simulated data representative of yeast and vertebrate genomes, we used computer simulations based on a realistic evolutionary model. We started with hypothetical ancestral genomes characterized by 5,000 genes distributed along 8 chromosomes for yeasts and by 18,000 genes distributed on 23 chromosomes for vertebrates. In both cases, we simulated random tree topologies with 21 leaves for yeasts and 13 leaves for vertebrates. The method for the construction of a random tree takes genes as building blocks and goes as follows:

1. it generates a random binary tree by defining the branching nodes uniformly over the time scale, with the exception of the first branching which is put at the root. More precisely, for each branching, it selects a leaf to split. It does it by going recursively from the root to the leaf by passing through internal nodes of the tree, with a half probability of choosing the right or the left subtree at an internal node. Once it selects a leaf to split, it attaches to it two new leaves. This construction is repeated until the number of leaves is equal to the expected number of species (21 species for yeasts and 13 for vertebrates).

2. based on the tree produced in step 1, it simulates chromosomal rearrangements along each branch of the tree, following a Poisson distribution, such that the average number of events from the ancestor (located at the root) to the species (located at the leaves) is approximately 500 for yeasts and 1000 for vertebrates. (We note that these values are comparable to those

obtained on actual yeast and vertebrate genomes (Drillon and Fischer, 2011)). Rearrangements were distributed on the tree according to the following proportions: 60% of inversions, 29.79% of reciprocal translocations, 5% of duplications, 5% of deletions, 0.1% of fusions, 0.1% of fissions, 0.01% of whole genome duplications (WGD) (Ma et al., 2006; Drillon and Fischer, 2011). Following a WGD event, one of the two copies of each duplicated gene was deleted with a probability of 80% (Wolfe and Shields, 1997). The number of genes involved in an inversion, duplication and deletion was chosen following a Poisson distribution (where the parameter of the distribution was set to 5 genes for inversions and duplications, and to 1 gene for deletions).

The simulated genomes produced by this approach are consistent with actual yeast and vertebrate genomes in terms of number of genes, number of chromosomes and number of rearrangements along the branches of the trees. For the analysis, the minimum number of rearrangements per branch was set to 1 or to 10 for both yeast and vertebrate trees, and 100 simulations were generated in each case. Synteny blocks were computed between all pairs of simulated genomes (note that here genes are represented by numerical identifiers, not by actual nucleotide or amino-acid sequences) and PhyChro was run on these simulated genomes to compare the predicted topologies with the known (simulated) ones. For determining PhyChro success rate, we counted the number of splits in the trees that were correctly and incorrectly reconstructed by PhyChro. For a minimum number of rearrangements per branch set to 1, the results are reported in **Figure 3**, where one observes that PhyChro is able to reconstruct correct tree topologies without any erroneous split in 79% of the cases for vertebrates and 61% for yeasts, and for the incorrect ones, in most cases (17% for vertebrates and 30% for yeasts), we record just one incorrect split per tree. Over all trees, 97% of the splits are correctly predicted by PhyChro, both for vertebrates and yeasts, and, most importantly, incorrect splits mainly correspond to very short branches, that is branches where only very few rearrangement events took place (see inset plot in **Figure 3**). If we set the number of events in a branch to be at least 10, the number of correct trees for vertebrates increases to 86% and for yeasts to 69%, with 98% and 97% of the splits that are correct over all trees, for vertebrates and yeasts, respectively.

This analysis helps to evaluate a confidence threshold for scores cS. In fact, 99% of correct splits are obtained with a score $cS \geq 0.2$ for the 100 simulated genomes for yeasts, and with a score $cS \geq 0.6$ for the 100 simulated genomes for vertebrates. This means that in the yeast phylogenetic tree reconstructed by PhyChro, the only weakly supported branch (scoring 0.05) is the one locating *E. gossipii* and *K. lactis* within the Protoploid clade, while the branch locating *C. lusitaniae*, displays a sufficiently strong cS score (0.33) to be trusted (**Figure 1b**).

For vertebrates, as discussed above, the position of Rodentia in the tree remains very weakly supported (**Figure 1a**).

A random shuffling of species in the 100 randomly generated trees is reported in **Figure S3**, where we note a shape of the distribution of errors that has a complementary tendency compared to the one obtained for PhyChro, that is the vast majority of events associated to a branch is incorrect and the number of erroneous splits corresponds, most of the times, to the number of internal branches (10 for vertebrates and 18 for yeasts). This corresponds to no correct trees obtained for both vertebrates and yeasts; we note that only the 1% of the splits are correct for yeasts and only the 3% for vertebrates. The same test, based on the same dataset of trees (and the same synteny blocks considered by PhyChro and the random tree analyses), has been realized on MLGO (**Figure S3**). MLGO works much better than the random case but yet is far from PhyChro performance: 3% of trees are correct for vertebrates and 1% for yeasts. Many of the trees that are reconstructed by MLGO have a high number of erroneous splits (57% for yeast and 42% for vertebrates) both for vertebrates and yeasts.

Adding new genomes to PhyChro reconstructions - a case study

The arrival of new sequenced genomes asks for their integration in the phylogenetic tree, and PhyChro can be profitably used to insert these new species. As an example, we considered the vertebrate tree in which, some of the species are known to be difficult to handle. In this respect, the literature contains an open debate because mammalian species positioning appears sensitive to the evolutionary information taken into consideration in phylogenetic reconstruction (Romiguier et al., 2013). We added three recently sequenced genomes, the cow, the pig and the lizard. PhyChro tree reconstruction (**Figure 2a**) correctly placed *Anolis*, the lizard, close to the birds; both are known to be members of Diapsida. It also added *Bos taurus* (cow) and *Sus scrofa* (pig) to the clade including horse and dog, with the nesting (horse, (dog, (cow, pig))). See **Figure S6** for an illustration of the resulting tree, and its legend for an analysis of the dubious position of the cow and the pig with respect to the horse and the dog (this reconstruction being of interest in exemplifying limits and power of PhyChro).

Discussion

PhyChro, a new strategy of phylogenetic recontruction. An important effort was made in this work to identify how chromosomal breaks coming from chromosomal rearrangements could be used as phylogenetically informative characters to perform phylogenetic reconstructions. Phy-

Chro differs in a fundamental way from the classical reconstruction methods. The first difference comes from the pairwise comparison approach between genomes which allows us to make the most out of the synteny information shared between closely and distantly related genomes at the same time. Another difference comes from the definition of 2 functions ($\mathrm{f}_{\mathrm{inc}}$ and $\mathrm{f}_{\mathrm{comp}}$, see Methods) which represent, respectively, the number of times where two genomes are split in two groups of incompatible adjacencies and the number of times where they are grouped together (not split) based on shared adjacencies. The ratio between these two functions is used to identify the least incompatible pairs of species from which sister genomes will be defined. The main originality of PhyChro is that it identifies sister genomes by minimizing the number of incompatible adjacencies rather than by maximizing the number of shared rearrangements. Formally, PhyChro bases its tree reconstruction on the Partial Split Distance. This distance relies on the notion of partial split that allows to record the number of incompatible adjacencies for pairs of genomes among a set of genomes. Hence, PhyChro does not try to combine internal branches into a tree topology, but rather it reconstructs the topology by iteratively identifying genomes and ancestral genomes that are closely related. It uses a bottom-up approach, similarly to what is done in distance-based methods. Note that PSD is a measure defined on a set of n genomes contrary to other previously introduced notions, measuring genome rearrangements, that are based on the comparison of only pairs of genomes. An example is the well known Breakpoint Distance (BD), defined to be the number of breakpoints observable from the comparison between two genomes. The notion was first used in (Nadeau and Taylor, 1984), then formally defined for one (Watterson et al., 1982; Sankoff and Blanchette, 1997) and multiple (Pevzner and Tesler, 2003; Tannier et al., 2009) chromosomes. The direct comparison between PSD and BD is impossible given that for two genomes G, H among n, the distance $BD(G, H)$ depends only on G, H while $PSD(G, H)$ depends on the n genomes. When reconstructing phylogenies, knowledge on the way pairs of genomes split in the tree (recall that the notion of non-trivial split is based on at least four genomes and not on pairs nor triplets) is primordial and one can only gather it through comparisons between all genomes involved in the reconstruction. This is why the intrinsic nature of a measure based on n genomes, like PSD, is expected to bring fundamental information for phylogenetic tree reconstruction. It is important to notice that PSD counts only those breakpoints that are supported by at least a quadruplet of genomes, and associated to rearrangements shared by at least two genomes, while BD counts all breakpoint events including those associated to rearrangements that are specific to a given genome (occurring on the external branches of a tree).

Thanks to this reconstruction strategy, PhyChro is less affected by “short-branch” attraction, which often leads distance-based methods to put genomes having undergone a lot of rearrangements/mutations higher in the tree than they belong. Another originality of PhyChro is that it provides branch length estimates that reflect the level of chromosome plasticity rather than the rates of punctual mutations, as all classical methods of phylogeny reconstruction do. In addition, PhyChro allows estimation of the robustness of branches in a way that is radically different from the bootstrap methods. The advantage here is that computing confidence scores is very fast as it does not involve additional tree reconstructions.

Phylogenetic reconstruction based on chromosomal rearrangements. We showed through the analysis of simulated genomes that PhyChro generates very accurate tree topologies by successfully reconstructing known tree topologies. Applications of PhyChro to real biological datasets comprising different types of genomes (yeasts and vertebrates) and covering different evolutionary ranges shows that chromosomal rearrangements are indeed phylogenetically informative and that accurate phylogenies can be reconstructed solely based on these large scale mutational events. This success demonstrates that the evolutionary signal that derives from chromosome rearrangements comprises at least as much phylogenetic information as the local information present in protein sequences. Moreover, we showed that PhyChro reconstructions are at least as accurate as the best reconstructions deriving from classical methods that use protein sequence comparisons. We also show that at particularly difficult branching positions, such as that of *C. glabrata* relatively to *S. cerevisiae* and *N. castellii*, PhyChro outcompetes all other methods of phylogenetic reconstruction.

Another important application of PhyChro was realized (with the same parameters used for vertebrates and yeasts species) on scleractinian corals, the foundation species of the coral-reef ecosystem. Corallimorpharians had been proposed to originate from a complex scleractinian ancestor that lost the ability to calcify in response to increasing ocean acidification, suggesting the possibility for corals to lose and gain the ability to calcify in response to increasing ocean acidification. A phylogenetic analysis based on 1 421 single-copy orthologs combined with PhyChro phylogenetic reconstruction allowed to disprove this hypothesis contributing evidence for the monophyly of scleractinian corals and the rejection of corallimorpharians as descendants of a complex coral ancestor (Wang et al., 2017).

These results suggest that synteny information should be integrated more broadly in future phylogenetic reconstruction analysis pipelines.

Materials and Methods

The classical notions of synteny blocks, breakpoints, splits and partial splits are recalled. We introduce the notions of "Partial splits associated to breakpoints" and of "Partial Split Distance" that are central in PhyChro.

Synteny blocks

A pairwise genome comparison G/H (or equivalently H/G), between the two genomes G, H, identifies chromosomal segments with conserved orthologous gene order. These segments are called *synteny blocks*, and are also referred to as *blocks*. Without loss of generality, we call B both the occurrences of the synteny block B in G and in H. Different definitions of synteny blocks have been proposed before (Ferretti et al., 1996; Roedelsperger and Dieterich , 2010; Pham and Pevzner , 2010; Proost et al., 2012; Drillon et al., 2014) and they are based on different conditions on the proximity between orthologs. PhyChro works with blocks B that verify the following five conditions:

- B is described by its pairs of homologous genes in G and H, called *anchors* for B. Since a gene can have several homologs, it can be involved in the definition of several anchors (within the same block or in different ones).
- the first and the last genes of B in G (H) have homologs in the corresponding block B in H (G). We say that B in G (H) is delimited by its first and last anchors.
- B is unique, in the sense that duplicated blocks are not explicitly handled and they are defined as independent blocks. For instance, if B is duplicated in G but not in H, the two copies of the block are considered as distinct in G and as overlapping in H.
- B is oriented or signed, and in particular, B can have a different orientation in G and in H. The orientation of B in a genome G maybe fixed in some arbitrary way or might depend on conditions that are specific to the definition of a block, such as the order and the orientation of its genes. When impossible to be established, a block orientation is left undetermined and the block is called "unoriented" or "unsigned". The orientation of a block allows us to differentiate its right and left ends (in order to determine which of its extremities is involved in a breakpoint): the "end" of B corresponds to the "beginning" of $-B$ and reciprocally.
- B, in G or H, can overlap or be included in another block.

A block B is called *telomeric* if it is the first or the last block of a chromosome in G or in H.

Breakpoints

Chromosomal rearrangements generate synteny breakpoints, or analogously, synteny block adjacencies. Given a block B obtained through the comparison G/H, a breakpoint is defined by the pair $[(B\,A)_G\,,\,(B\,C)_H]$ of block adjacencies $(B\,A)$ in G and $(B\,C)$ in H. In I of **Figure 4**, for instance, the right end of block B is contiguous to the left ends of blocks A and C in genomes G and H, respectively. Since blocks are oriented, notice that the same breakpoint might correspond to $[(B\,A)_G, (-C\,-B)_H]$, where $-B$ has $-C$ on its left end instead. Notice also that synteny blocks derived from duplications or chromosome fusions/fissions do not generate pairs of block adjacencies and therefore are not explicitly considered here. Blocks derived from translocations, inversions and transpositions of DNA segments are the only ones that are informative in our analysis. Each block (except the telomeric ones) should, in theory, lead to two breakpoints (one at each end of the block, see I in **Figure 4**). However, complex gene-order configurations might lead to a reconstruction of synteny blocks that overlap, are included in one another or are unoriented (like for blocks reconstructed by SynChro (Drillon et al., 2014)). In the following, we consider as breakpoints only those pairs of regions in G and in H for which preceding and following blocks are unambiguously identified (and ignore the others).

Splits

A split is a bipartition of a set of taxa. **Figure 1a** illustrates an example of a split and of a trivial split, that is, a split induced by an external edge connecting a leaf to the rest of the tree. Splits play an important role in phylogenetic reconstruction (Bandelt and Dress, 1992; Huson et al., 2010) as each edge of an unrooted tree is univocally associated to a split. In fact, an edge splits taxa into the two disjoint subsets S_1, S_2 labeling the leaves of the subtrees rooted at the extremes of the edge. We note that the union of S_1, S_2 covers the full set of taxa. In evolutionary terms, we think of genomes in S_1 (or S_2) as having undergone a number of common ancestral rearrangements, specifically the ones that occurred along the edge, that genomes in S_2 (S_1) did not undergo. Strictly speaking, it cannot be established whether these rearrangements took place for S_1 or for S_2 because the tree is not rooted. Hence, ideally, for the reconstruction of a phylogenetic tree, one could hope (i) to recover rearrangements from genomic data, (ii) to define splits of genomes sharing the rearrangements and (iii) to reconstruct the edges of the tree by combining splits identified from the rearrangements.

Partial splits

For the purpose of tree reconstruction, traces of chromosomal rearrangements may have disappeared in some genomes (due to the accumulation of other rearrangements), and it might become impossible to recover splits. This is why, we shall use a generalisation of the concept of split to the one of partial split. This notion was introduced in (Semple & Steel , 2001; Huson et al., 2004; Huber et al., 2005). Formally, a partial split is a pair of non-empty disjoint sets of taxa. Intuitively, given an unrooted phylogenetic tree whose leaves are labelled by different taxa and given some path c in the tree, we say that c induces a partial split of the sets of genomes S_1, S_2 if: 1. S_1, S_2 are constituted by some (possibly all) of the taxa associated to the subtrees rooted at the extremes of c; 2. in each S_i, for $i = 1, 2$, there are at least two taxa that are connected by a shortest path passing through the root of the corresponding subtree (**Figure 1b**). We note that, by definition, $S_1 \cap S_2 = \emptyset$ and, also, that $S_1 \cup S_2$ does not necessarily correspond to the full set of taxa in the subtrees rooted at the extremes of c. *A fortiori*, $S_1 \cup S_2$ does not necessarily correspond to the full set of taxa in the complete tree, as it is the case for splits. In fact, a split is a partial split where c is an edge, but a partial split induced by an edge need not be a split because of condition 1. As for splits, we think of genomes in S_1 (or S_2) as having undergone a number of common ancestral rearrangements, specifically the ones that occurred along the path c, that genomes in S_2 (S_1) did not undergo.

As for splits, we say that a partial split is *trivial* when one of the two subsets S_1, S_2 is a singleton. Notice that trivial partial splits do not bring information on the topology of the tree (since the set of trivial partial splits is the same for all topologies) and are not used in tree reconstruction. We shall use them to estimate the length of the terminal branches though, that is, branches leading to leaves in the tree.

Testing the conservation of block adjacencies

Given a breakpoint $[(B\,A)_G, (B\,C)_H]$ in the comparison G/H, we test for the presence of $(B\,A)_G$ in a genome K (by definition, $(B\,A)_G \in G$). The test is similarly stated for $(B\,C)_H$. The test does not directly search for blocks B and A in K because they might not have direct equivalents in G/K. Instead, it infers the presence of the adjacency $(B\,A)_G$ in K at the gene level, by testing whether the genes flanking the $(B\,A)_G$ adjacency in G, that is, the right end of block B and the left end of block A, have syntenic homologs in K. More precisely, the test compares G and K and determines whether there is a synteny block D in K and G such that the following conditions are satisfied (we refer to the notation employed in **Figure 5** - see also **Figure S7**):

(i) the two last anchors (or syntenic homologs) q', q of B and the two first anchors r, r' of A, along G, belong to the same synteny block D in G/K;

(ii) q', r' are preceded and followed along G, respectively, by at least two other anchors in D (possibly including themselves).

(iii) let s be the anchor of D in K whose homolog in G lies in the right most position of B, and let t be the anchor of D in K whose homolog in G lies in the left most position of A. Then, the sum of the number of genes between s and t in K and between their homologs in G (see **Figure 5**) is at most 4.

Conditions (i) and (ii) guarantee block D in G to overlap several anchors of A and B in G, and condition (iii) ensures the genes forming the $(B\,A)_G$ adjacency in G and K to be in physical proximity. Such proximity is computed for a maximum of 4 genes between the two anchors s and t in **Figure 5**. All values from 3 to 6 have been tested to choose the best parameter for yeasts and vertebrates. (Note that value 3 is too strict and value 6 brings noise in the construction.) These three conditions introduce some flexibility in the definition of synteny conservation, without being too permissive. If they are all satisfied, we say that the adjacency belongs to K and write $(B\,A)_G \in K$. If q and r belong to the same block D in G/K but some of the conditions fail, we still say that $(B\,A)_G \in K$ and consider the relation as *weakly* supported. These weak adjacencies can be due to false ortholog assignments or to small inversions. In all other cases, we say that $(B\,A)_G \notin K$.

Partial splits assigned to breakpoints

Given a breakpoint $[(B\,A)_G, (B\,C)_H]$, we define a partial split by identifying two sets of genomes, $S_{(BA)}$ and $S_{(BC)}$, where $S_{(BA)}$ comprises genomes sharing the adjacency $(B\,A)_G$ and $S_{(BC)}$ comprises genomes sharing the adjacency $(B\,C)_H$. For this, we apply the above adjacency test, checking whether the adjacencies $(B\,A)_G$ and $(B\,C)_H$ derived from the G/H comparison are present in a genome K or not, for all $K \neq G, H$. Namely, $K \in S_{(BA)}$ if and only if $(B\,A)_G \in K$, and $K \in S_{(BC)}$ if and only if $(B\,C)_H \in K$.

Notice that a genome K that neither contain $(B\,A)_G$ nor $(B\,C)_H$ belongs to none of the two sets. Also, a genome K may contain, at the same time, the two adjacencies defining a given breakpoint. This ambiguous case might occur either for a breakpoint $[(B\,A)_G, (B\,C)_H]$ when C follows A in G and A is small enough to make condition (iii) true for $(BC)_H$ in K (see **Figure S8a**), or for a breakpoint $[(B\,A)_G, (B\,-A)_H]$ when A is small enough to make $(BA)_G \in K$ and $(B-A)_H \in K$ (see **Figure S8b**).

Intuitively, the coexistence of $(B\,A)_G \in K$ and $(B\,C)_H \in K$, for some K, indicates that $(B\,A)_G$ and $(B\,C)_H$ are too "similar" to claim that they support a split. Therefore, it is only when the two sets of genomes $S_{(B\,A)}, S_{(B\,C)}$ are disjoint that we say that they form a *partial split*, denoted $S_{(B\,A)} \| S_{(B\,C)}$, associated to the breakpoint $[(B\,A)_G, (B\,C)_H]$ (**Figure 1b**).

The Partial Split Distance

Given a set of genomes, genome pairwise distances can be computed by considering the set of partial splits associated to all breakpoints issued from all pairwise genome comparisons. For this, we shall define two functions, $\mathrm{f_{inc}}$ and $\mathrm{f_{comp}}$, on the list of non-trivial partial splits.

The first one, $\mathrm{f_{inc}}(G, H)$ (where "inc" stands for incompatible), counts the number of times that genomes G and H belong to different subsets of a partial split (as for partial splits 1 and 2 in III of **Figure 4**):

$$\mathrm{f_{inc}}(G, H) = |\{S_{(B\,A)} \| S_{(B\,C)} : (G \in S_{(B\,A)} \wedge H \in S_{(B\,C)}) \vee (G \in S_{(B\,C)} \wedge H \in S_{(B\,A)})\}|$$

The second function, $\mathrm{f_{comp}}(G, H)$ (where "comp" stands for compatible), counts the number of times that genomes G and H are found in the same subset of a partial split, *i.e.* sharing a same adjacency (as for the partial split 4 in III of **Figure 4**):

$$\mathrm{f_{comp}}(G, H) = |\{S_{(B\,A)} \| S_{(B\,C)} : (G \in S_{(B\,A)} \wedge H \in S_{(B\,A)}) \vee (G \in S_{(B\,C)} \wedge H \in S_{(B\,C)})\}|$$

The function $\mathrm{f_{inc}}$ represents an "internal" distance between genomes, and we call it *Partial Split Distance*, PSD in short. Intuitively, given two genomes, PSD is proportional to the number of rearrangements that occur along the internal branches separating these two genomes in the phylogenetic tree that we want to reconstruct. The number of these rearrangements is estimated with f_{inc}, by using the number of non-trivial splits separating the two genomes. This means that sister genomes, that is genomes separated by no internal branch, should have a PSD distance equal to zero (independently of the length of their external branches). This property will be used to identify sister genomes and to reconstruct phylogenies bottom-up (see below). In the same way, very close genomes, separated by few and short internal branches, should have a PSD close to zero. However, because f_{inc} is defined from non-trivial splits, very distant genomes, which do not share many adjacencies with other genomes and, therefore, are not involved in many non-trivial splits, have also a PSD very close to zero with all other genomes. To take into account this fact, we consider f_{comp} and use the ratio $\mathcal{R} = (\mathrm{f}_{\mathrm{inc}} + 1)/(\mathrm{f}_{\mathrm{comp}} + 1)$ to discriminate among pairs of genomes that have a very small internal distance ($\mathrm{f}_{\mathrm{inc}}$ close to zero) those that are very closely related (high $\mathrm{f}_{\mathrm{comp}}$ value) from those that are very distantly related ($\mathrm{f}_{\mathrm{comp}}$ close to zero).

Note that, if $\mathrm{f}_{\mathrm{inc}}(G, H) \neq 0$ then there exists at least one non-trivial partial split $S_{(B\,A)} \| S_{(B\,C)}$ that separates G from H. This means that there exist genomes K, L such that $(G, K \in S_{(B\,A)} \wedge H, L \in S_{(B\,C)}) \vee (G, K \in S_{(B\,C)} \wedge H, L \in S_{(B\,A)})$. Ideally, this suggests that in a phylogenetic reconstruction involving genomes G, H, K, L, the two genomes G, H should not be considered as sister genomes. In reality, as mentioned above, it might be difficult to unravel complete information from breakpoints (due either to convergence or to the accumulation of rearrangements) and one might have to treat as sister genomes, those pairs of genomes that display the smallest $\mathrm{f}_{\mathrm{inc}}$ value, even if it is different from 0.

The PhyChro algorithm

Phylogenetic reconstruction based on partial splits is a more delicate problem than tree reconstruction based on splits (Bandelt and Dress, 1992; Huson et al., 2010; Semple & Steel , 2001; Huson et al., 2004; Huber et al., 2005). PhyChro comprises four main parts (I, II, III and IV; see **Figure 4** and **Table 1**) divided into 7 major steps that are detailed below.

Part I - Identification of breakpoints

Step 1: For each pairwise comparison G/H between pairs of genomes among n involved in the reconstruction, PhyChro iteratively identifies the breakpoints associated to each synteny block. See I in **Figure 4**.

Part II - Identification of partial splits

Step 2: For each breakpoint $[(B\,A)_G, (B\,C)_H]$ identified in Step 1 and issued from the comparison G/H and for each genome $K \neq G, H$, PhyChro determines whether $(B\,A)_G$ or $(B\,C)_H$ is present in K (as seen in section "Testing the conservation of block adjacencies").

Step 3: Based on the results from Step 2, PhyChro defines two sets of genomes, $S'_{(BA)}$ and $S'_{(BC)}$, that share one or the other adjacency defining the breakpoint $[(B\,A)_G, (B\,C)_H]$. If $S'_{(BA)}$ and $S'_{(BC)}$ are not disjoint, then the sets are ignored (as seen in section "Partial split assigned to breakpoints"). These partial splits are associated to ambiguous breakpoints, which are themselves due to small blocks. If $S'_{(BA)}$ and $S'_{(BC)}$ are disjoint, then PhyChro removes from the two sets those genomes that support only weakly the adjacency (as seen in section "Testing the conservation of block adjacencies"). Then it checks that both resulting sets $S_{(BA)}$

and $S_{(BC)}$ are not singletons; if so, it adds $S_{(B\,A)}||S_{(B\,C)}$ to the collection of partial splits. Note that $S_{(B\,A)}||S_{(B\,C)}$ may be trivial or not.

At the end of the iteration (steps 2 and 3), PhyChro has identified a collection of partial splits.

Part III - Bottom-up tree reconstruction

Step 4: For each pair of genomes G, H, PhyChro computes their PSD, $\mathrm{f}_{\mathrm{inc}}(G, H)$ and $\mathrm{f}_{\mathrm{comp}}(G, H)$ (as seen in section "The Partial Split Distance").

Step 5: The creation of an internal node $\{KL\}$ of the tree relies on the identification of the two sister genomes K and L (among the n genomes) displaying the smallest $\mathrm{f}_{\mathrm{inc}}$ value. However, as explained above, to avoid considering very distant genomes that could have very small $\mathrm{f}_{\mathrm{inc}}$ values, sister genomes are chosen to be the pair displaying the smallest $\mathrm{f}_{\mathrm{inc}}$ value among the $n/2$ genome pairs that have the smallest ratio $\mathcal{R}$ (III in **Figure 4**). Notice that the maximum number of possible sister genomes in a tree of n species is $n/2$. If there are multiple identical minimal $\mathrm{f}_{\mathrm{inc}}$ values, either they involve different pairs of genomes and they will be treated one after the other in the different and successive iterations, or they involve incompatible pairs of genomes (involving the same genomes; a very unlikely situation that would results into the creation of a node with a low confidence score - see below) and the choice among them is left arbitrary.

Step 6: Once the internal node $\{KL\}$ is created, the list of partial splits identified at Step 3, is updated by replacing all occurrences of K and L by the node $\{KL\}$. Two types of partial splits $S_{(B\,A)}||S_{(B\,C)}$ are deleted: (i) partial splits that are discordant with the new node, that is partial splits where K and L belong to $S_{(B\,A)}$ and $S_{(B\,C)}$, respectively (see partial split 3 in III of **Figure 4**), (ii) partial splits characterized by a set of genomes composed by K and L only, since these partial splits would become trivial carrying no useful information for further topology reconstruction (see partial split 4 in III of **Figure 4**).

The process (steps 4-6) is iterated on the restricted set of genomes, where K, L are replaced by the ancestral genome $\{KL\}$, and on the updated set of partial splits obtained in step 6: all $\mathrm{f}_{\mathrm{inc}}$ and $\mathrm{f}_{\mathrm{comp}}$ values are re-computed from the updated list of partial splits, new internal nodes are created, and the list of partial splits is updated again. The iteration is run until only three genomes remain (exactly one unrooted tree topology is then possible).

Part IV. Estimations on the branches of the phylogenetic tree. PhyChro produces an estimation of the branch length and a confidence score of the reconstructed nodes. The branch length is an indicator of the complexity of the chromosomal structures (that is, of the amount of rearrangements identifiable from the genomes under consideration), and the confidence score indicates how much the reconstruction is supported and/or contradicted by the information contained in the initial non-trivial partial splits.

Step 7: Branch length for internal and terminal branches is estimated by using information contained in non-trivial and trivial partial splits, respectively . Branch length is the sum of a weighted number of partial splits (corresponding to a number of breakpoints, see **Supplementary File**) that support the existence of the branch (**Figure 6**), and therefore it indirectly represents a number of rearrangements. These values are necessarily an underestimation because most partial splits support the existence of a path in the tree rather than a specific branch, and therefore, are not considered for the calculation of branch lengths. In addition, terminal branches of distant genomes and internal branches between distant clades will be even more underestimated as partial splits supporting this kind of branches are rare.

PhyChro also estimates a confidence score for each internal branch by calculating the proportion of non-trivial partial splits that supports its existence over the total number of non-trivial partial splits that either support or contradict it (**Figure 6** and **Supplementary File**). In addition to the confidence score, PhyChro provides the list of all f_{inc} values computed for genome pairs, which can help to know if a node is trustworthy or not.

Description of input data

PhyChro requires as input the list of synteny blocks computed for each pairwise comparison G/H between all pairs of genomes G, H involved in the phylogenetic reconstruction. Anchors must be provided for each pair of synteny blocks issued from a comparison G/H. We recall that synteny blocks handled by PhyChro can overlap and that the same gene can be an anchor for distinct blocks. Duplicated synteny blocks are treated as independent blocks even though their anchors can be shared.

PhyChro accepts synteny blocks that are reconstructed with various tools as long as they are converted into the expected format, described in the README file of the PhyChro package. For the applications to yeast and vertebrate species, synteny blocks were computed with the SynChro software (Drillon et al., 2014), setting the Δ parameter to 3. Δ is a parameter that allows to define synteny blocks by controlling the complexity of internal micro-rearrangements. Intuitively, high values of Δ are more permissive and allow larger micro-rearrangements to be tolerated within synteny blocks while smaller values of Δ are more stringent and split synteny blocks at micro-rearrangement breakpoints. This implies that, for distantly related genomes, increasing the Δ value allows to recover a larger number of synteny blocks. For these genomes, small values of Δ would allow recovering the signal only from small inversions. Notice that when PhyChro reconstructs trees using blocks computed with $\Delta = 2$, the yeast tree contains 3 erroneous splits and the vertebrate tree contains 1, while both trees are correct when blocks are computed with $\Delta = 3$ or 4. SynChro automatically reconstructs pairwise synteny blocks that can be directly read by PhyChro, and it can be downloaded at www.lcqb.upmc.fr/CHROnicle/SynChro.html.

To analyze how sensitive is PhyChro to synteny block reconstruction, we constructed a second set of synteny blocks with the program i-ADHoRe 3.0 (Proost et al., 2012). We followed the protocol used in (Drillon et al., 2014) fixing parameters as follows: prob.cutoff=0.001, gap_size=15, cluster.gap=20, q_value=0.9 and anchor.points=3. The remaining parameters were set with default values. The i-ADHoRe 3.0 software package is available at bioinformatics.psb.ugent.be/software.

PhyChro was tested on 13 vertebrate species and 21 yeast species. The detailed list is given in the **Supplementary Table 1**. The vertebrate genome sequences have been downloaded from NCBI and the yeast species were downloaded from several sites listed in **Supplementary Table 2**.

PhyChro computational time

PhyChro time complexity depends on the number of genomes given in input and on the number of rearrangements that took place among these genomes. Phylogenetic reconstructions with PhyChro ran, on a desk computer, in 15 and 10 minutes for the 13 vertebrate and 21 yeast species, respectively. A total of 130,485 and 179,649 breakpoints and of 17,848 (1,501 different ones) and 20,924 (3,901) partial splits were identified for vertebrate and yeast genomes, respectively.

Comparison with MLGO

PhyChro has been compared with the method of phylogenetic reconstruction Maximum Likelihood for Gene Order Analysis (MLGO). MLGO's input is constituted by chromosomes described as sequences of gene identifiers and these latter can be used multiple times, that is gene duplicates are allowed in MLGO. To prepare the input to MLGO, we used OrthoMCL as suggested in (Lin et al., 2013). Genes have been clustered using OrthoMCL (Li et al., 2003) with 1.5 as inflation value, 30% of similarity cut-off and a E-value of $10e$-5. The same label has been used for genes falling in the same cluster. MLGO analysis was run at geneorder.com/server.php (Lin et al., 2013).

Phylogenetic reconstructions based on protein sequences

We identified 357 families of syntenic homologs (considered as orthologs) sharing more than 90% of similarity between the 13 vertebrate species, and 80 families sharing more than 80% of similarity between the 21 yeast species, using SynChro (Drillon et al., 2014). Orthologous proteins were aligned with MUSCLE (version 3.8.31) (Edgar, 2004) and alignments were cleaned with Gblocks (version 0.91b) (Castresana, 2000). Cleaned concatenated alignments were then provided to PhyML 3.0 (which was run with the LG amino-acid substitution model) and ProtPars. For Neighbor, we computed the distance matrix using ProtDist and ran it with the neighbour joining option. ProtPars, Neighbor and ProtDist are included in the PHYLogeny Inference Package (version 3.67) (Felsenstein, 1989) and have been used online at mobyle.pasteur.fr/cgi-bin/portal.py.

Acknowledgments

We thank Ingrid Lafontaine for critical reading of the manuscript. This work was supported by the Agence Nationale de la Recherche ('GB-3G', ANR-10-BLAN-1606-01) (GF), an ATIP-Avenir Plus grant from the CNRS (GF), a teaching assistantship (ATER) from the Ministère de l'Enseignement Supérieure et de la Recherche (GD), and funds from the Institut Universitaire de France (AC).

References

Bandelt HJ, Dress AW. 1992. A canonical decomposition theory for metrics on a finite set. Advances in Mathematics 92:47–105.

Bapteste E, Boucher Y, Leigh J, Doolittle WF. 2004. Phylogenetic reconstruction and lateral gene transfer. Trends in Microbiology 12:406–411.

Bergeron A, Blanchette M, Chateau A, Chauve C. 2004. Reconstructing ancestral gene orders using conserved intervals. In Proceedings of WABI 2004: Algorithms in Bioinformatics, LNCS, volume 3240, 14-25.

Boore JL. 2006. The use of genome-level characters for phylogenetic reconstruction. Trends Ecol. & Evol. 21, 439446.

Bourque G, Pevzner PA. 2002. Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species. Genome Research 12:26–36.

Burger G, Gray MW, Lang BF. 2003. Mitochondrial genomes: anything goes. Trends in Genetics 19:709–716.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17:540–552.

Cosner ME, Jansen RK, Moret BM, Raubeson LA, Wang LS, Warnow T, Wyman S. 2000. A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data. Proc Int Conf Intell Syst Mol Biol 8:104–115.

Drillon G, Fischer G. 2011. Comparative study on synteny between yeasts and vertebrates. Comptes Rendus de Biologie 334:629–638.

Drillon G, Carbone A, Fischer G. 2013. Combinatorics of chromosomal rearrangements based on synteny blocks and synteny packs. J Logic Computation 23 (4): 815-838. doi: 10.1093/logcom/exr047

Drillon G, Carbone A, Fischer G. 2014. SynChro: A fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. PLoS One 9(3).

Dujon B. 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. TRENDS in Genetics 22(7): 375-387.

Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792–1797.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology 27:401–410.

Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (version 3.2). Cladistics 5:164–166.

Feng B, Lin Y, Zhou L, Guo Y, Friedman R, Xia R, Hu F, Liu C, Tang J. 2017. Reconstructing yeasts phylogenies and ancestors from whole genome data. Scientific Reports 7(1):15209.

Ferretti V, Nadeau JH, Sankoff D. 1996. Original Synteny. In Proceedings of the 7th Annual Symposium on Combinatorial Pattern Matching (CPM '96), Daniel S. Hirschberg and Eugene W. Myers (Eds.). Springer-Verlag, London, UK, 159–167.

Fertin G. 2009. Combinatorics of genome rearrangements, MIT press.

Gordon JL, Byrne KP, Wolfe KH. 2009. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. PLoS Genet 5:e1000485.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52:696–704.

Guyon F, Brochier-Armanet C, Guénoche A. 2009. Comparison of alignment free string distances for complete genome phylogeny. Advances in Data Analysis and Classification 3:95–108 10.1007/s11634-009-0041-z.

Hittinger CT, Rokas A, Carroll SB. 2004. Parallel inactivation of multiple gal pathway genes and ecological diversification in yeasts. Proceedings of the National Academy of Sciences 101:14144–14149.

Hu F, Gao N, Zhang M, Tang J. 2011. Maximum likelihood phylogenetic reconstruction using gene order encodings. In Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) 1–6.

Hu F, Lin Y, Tang J. 2014. MLGO: phylogeny reconstruction and ancestral inference from gene-order data. BMC Bioinformatics 15:354.

Huber KT, Moulton V, Semple C, Steel M. 2005. Recovering a phylogenetic tree using pairwise closure operations Applied Mathematics Letters 18(3):361-366.

Huson DH, Dezulian T, Klöpper T, Steel MA. 2004. Phylogenetic super-networks from partial trees. IEEE/ACM Trans Comput Biol Bioinform. 1(4):151-8.

Huson D, Rupp R, Scornavacca C. 2010. Phylogenetic Networks. Concepts, Algorithms and Applications. Cambridge Univ Press.

Kurtzman CP, Robnett CJ. 2003. Phylogenetic relationships among yeasts of the 'saccharomyces complex' determined from multigene sequence analyses. FEMS Yeast Research 3(4):417–432.

Larget B, Simon DL, Kadane JB, Sweet D. 2005. A bayesian analysis of metazoan mitochondrial genome arrangements. Molecular Biology and Evolution 22:486–495.

Li L, Stoeckert CJJr, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13:2178–2189.

Lin Y, Rajan V, Moret B. 2012. Bootstrapping phylogenies inferred from rearrangement data. Algorithms for Molecular Biology, 7:21.

Lin Y, Hu F, Tang J, Moret B. 2013. Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. Pacific Symposium on Biocomputing. 18:285–296.

Liu T, Tang J, Moret BME. 2005. Quartet-based phylogeny reconstruction from gene orders. In Proceedings of the 11th Int" l Conf. Computing and Combinatorics (COCOON" 05). LNCS Volume 3595, Springer-Verlag 63–73.

Luo H, Arndt W, Zhang Y, Shi G, Alekseyev MA, Tang J, Hughes AL, Friedman R. 2012. Phylogenetic analysis of genome rearrangements among five mammalian orders. Mol Phylogenet Evol. 65(3):871–82.

Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, Kent JW, Blanchette M, Haussler D, Miller W. 2006. Reconstructing contiguous regions of an ancestral genome. Genome Res. 16:1557–1565.

Moret BM, Wang LS, Warnow T, Wyman SK. 2001a. New approaches for reconstructing phylogenies from gene order data. Bioinformatics 17 Suppl 1:S165–S173.

Moret BM, Wyman S, Bader DA, Warnow T, Yan M. 2001b. A new implementation and detailed study of breakpoint analysis. In Proceedings of the Pacific Symposium on Biocomputing, 583-94.

Nadeau J, Taylor B. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. Proceedings of the National Academy of Sciences, 81:814–818.

Pevzner P, Tesler G. 2003. Transforming men into mice: the Nadeau-Taylor chromosomal breakage model revisited. In Proceedings of the seventh annual international conference on Research in computational molecular biology (RECOMB'03), ACM, 247–256.

Pham S, Pevzner P. 2010 DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. Bioinformatics, 26:2509–2516.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol 9:e1000602.

Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, et al. 2012. i-ADHoRe 3.0fast and sensitive detection of genomic homology in extremely large data sets. Nucleic Acids Research 40:e11.

Roedelsperger C, Dieterich C. 2010. CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. PLoS One, 5:e8861.

Rokas A, Holland PW. 2000. Rare genomic changes as a tool for phylogenetics. Trends in Ecology & Evolution 15:454–459.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimise tree conflict and unravel the root of placental mammals. Molecular Biology & Evolution 30:2134 – 2144.

Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R. 1992. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc. Natl. Acad. Sci. USA, 89:6575–6579.

Sankoff D, Blanchette M. 1997. The median problem for breakpoints in comparative genomics. In Proceedings of the Third International Computing and Combinatorics Conference COCOON'97, LNCS 1276:251–263.

Sankoff D, Blanchette M. 1998. Multiple genome rearrangement and breakpoint phylogeny. Journal of computational biology, 5:555–570.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440:341–345.

Semple C, Steel M. 2001. Tree Reconstruction via a Closure Operation on Partial Splits. In *Computational Biology: First International Conference on Biology, Informatics, and Mathematics (JOBIM 2000),* O. Gascuel and M.F. Sagot (Eds), LNCS Volume 2066:126–134.

Tang J, Moret BME. 2003. Scaling up accurate phylogenetic reconstruction from gene-order data. Bioinformatics 19 Suppl 1:i305–i312.

Tannier E, Zheng C, Sankoff D. 2009. Multichromosomal median and halving problems under different genomic distances. BMC Bioinformatics,10:120.

Vakirlis N, Sarilar V, Drillon G, Agier N, Blanpain L, Carbone A, Devillers H, Dubois K, Gillet-Markowska A, Huu-Vang N, et al. 2016. Reconstructing genome history in a yeast genus, Genome Research, 26: 918-932.

Wang LS, Warnow T, Moret B, Jansen R, Raubeson L. 2006. Distance-based genome rearrangement phylogeny. Journal of Molecular Evolution 63:473–483.

Wang X, Drillon G, Ryu T, Voolstra CR, Aranda M. 2017. Genome-based analyses of six hexacorallian species reject the naked coral hypothesis. Genome biology and evolution, 9(10):2626–2634.

Watterson G, Ewens W, Hall T, Morgan A. 1982. The chromosome inversion problem. Journal of Theoretical Biology, 99:1–7.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387(6634):708–713.

Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. Science 319:473–476.

Xu AW, Moret BM. 2011. Gasts: Parsimony scoring under rearrangements. In International Workshop on Algorithms in Bioinformatics, 351363 (Springer).

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. Nature Reviews Genetics 13:303–314.

Zheng C, Sankoff D. 2011. On the pathgroups approach to rapid small phylogeny. BMC Bioinformatics 12:S4.

Figure Legends

Figure 1. Splits and partial splits. a) Examples of trivial (orange edge) and non-trivial (red edge) splits. **b)** The two sets of genomes $\{GK\}$ and $\{HL\}$, forming a partial split, uniquely determine a path (in red, in the left tree) or an edge (in red, in the right tree) that join the smallest subtrees including G, K, and H, L.

Figure 2. Phylogenies obtained with PhyChro for 13 vertebrate (a) and 21 yeast (b) species. Confidence scores that range between 0 and 1 are indicated on internal branches. Scale bars provide an estimation of the branch lengths, which correspond to the number of breakpoints, indirectly representing a number of rearrangements. For sake of clarity, internal branches with length smaller than 1 unit are represented in orange with an arbitrary small, but visible, length. Whole Genome Duplication events (WGD) are reported. *Clavispora lusitaniae* location in the tree is dubious and highlighted in dark orange.

Figure 3. PhyChro, MLGO and random reconstructions tested on simulated trees. Simulated phylogenetic trees describing rearrangement events were generated for vertebrate-like (top) and yeast-like (bottom) genomes and used to check whether PhyChro (left) and MLGO (center) could correctly reconstruct the original phylogeny from the corresponding sets of simulated genomes. The simulated trees have been used also to check to which extent a random assignment of rearrangements (right) on the branches could correctly reconstruct the original phylogeny from the corresponding sets of simulated genomes. The histograms report the number of trees with a fixed number of incorrect splits predicted by the three methods. The inset plots represent the distribution of the number of branches with a fixed length (corresponding to a number of simulated rearrangements that were applied to these branches) in the simulated trees, and describe how many of those branches have been reconstructed correctly (white) or incorrectly (black) by a method.

Figure 4. PhyChro algorithm. The four main parts and the seven steps, briefly described here, are detailed in the main text.

Figure 5. Conservation of the adjacency $(B\,A)_G$ in the genome K. Genes are indicated as dots or stars. Stars, in G, are used for the two last (q' and q) and first (r and r') anchors of blocks B and A in the comparison G/H. Red, yellow and green colors are used to highlight anchors associated to the blocks A, B and D, obtained in the comparisons G/H, G/H and G/K, respectively. Genes q', q, r and r' belong to the same block D in G/K. The number of anchors of D lying before q' (after r'), and possibly including it, is indicated above q' (r')

within a square. Gene s (t) is the anchor of D whose homolog in G lies in the right (left) most position of B (A). Homology is indicated by links among genes occurring in different genomes: s is homolog of q and t of r'. Note that, here, the three conditions (i)-(iii) discussed in the text are satisfied and that $K \in (B\,A)_G$.

Figure 6. Examples of partial splits supporting or contradicting the existence of a given branch. Given a branch (red edges in the left and right trees), we consider the sets of genomes $\mathcal{H}, \mathcal{G}, \mathcal{K}, \mathcal{L}$ corresponding to the maximal subtrees associated to the edge by the tree topology. Sets $\mathcal{H}, \mathcal{G}, \mathcal{K}, \mathcal{L}$ contain genomes H, G, K, L, respectively. **a)** Each internal branch is characterized by a double pair of genome sets $[(\mathcal{G}, \mathcal{K}), (\mathcal{H}, \mathcal{L})]$, which allows to define the partial splits that support or contradict this branch. **b)** Each external branch is characterized by one pair of genome sets $(\mathcal{G}, \mathcal{H})$, which allows to define the trivial partial splits that support this branch.

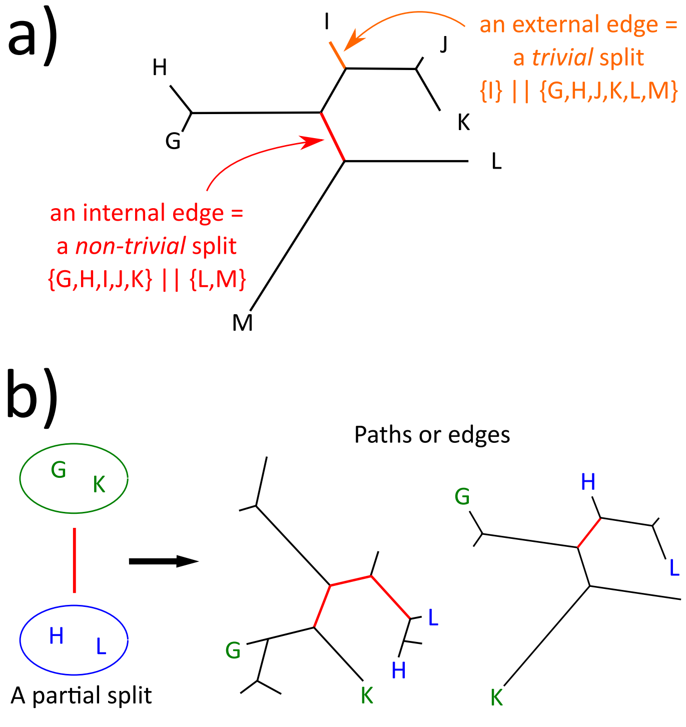


Figure 1: **Splits and partial splits.**

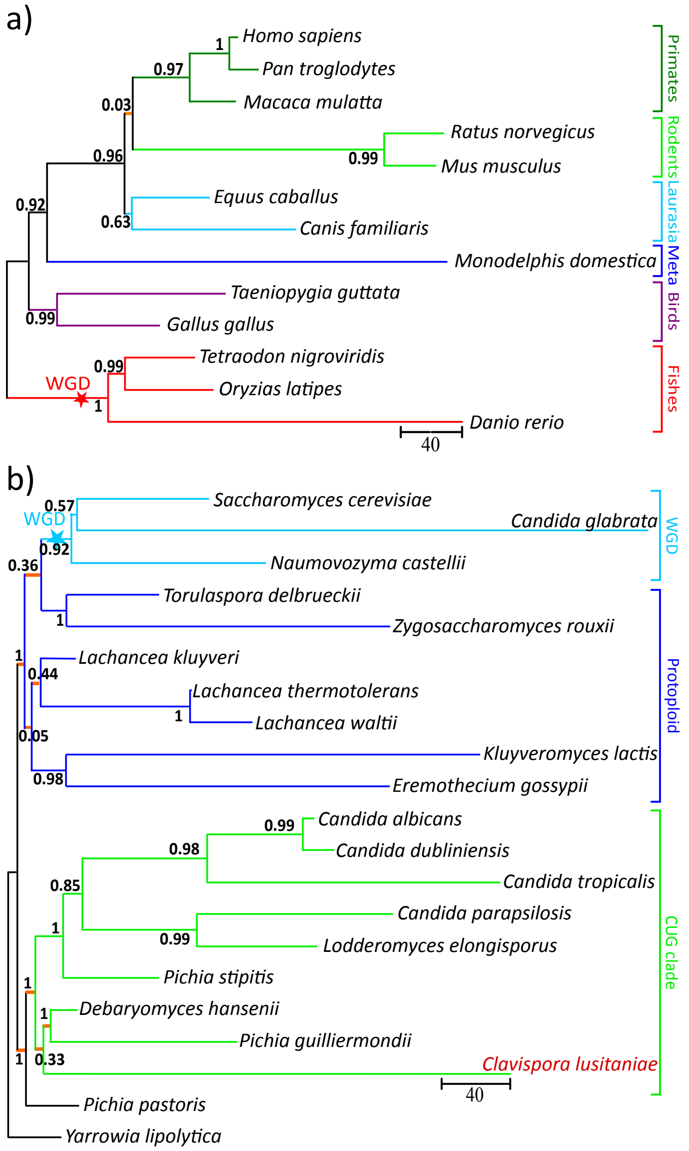


Figure 2: **Phylogenies obtained with PhyChro for 13 vertebrate (a) and 21 yeast (b) species.**

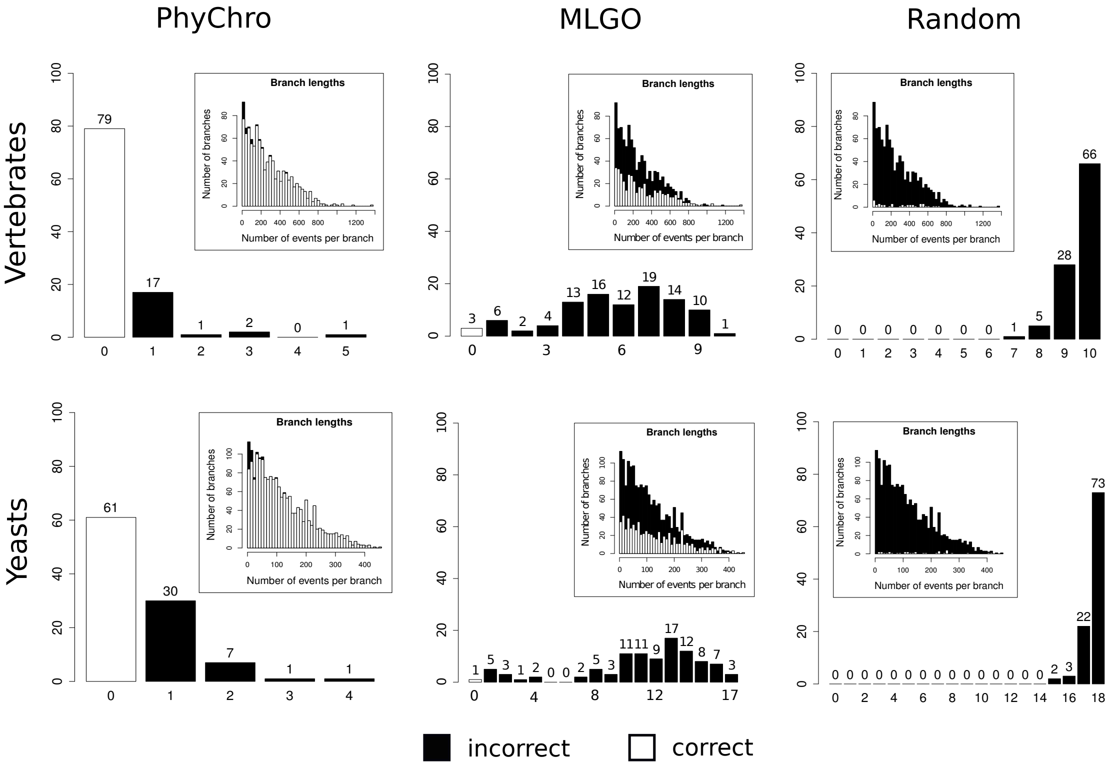
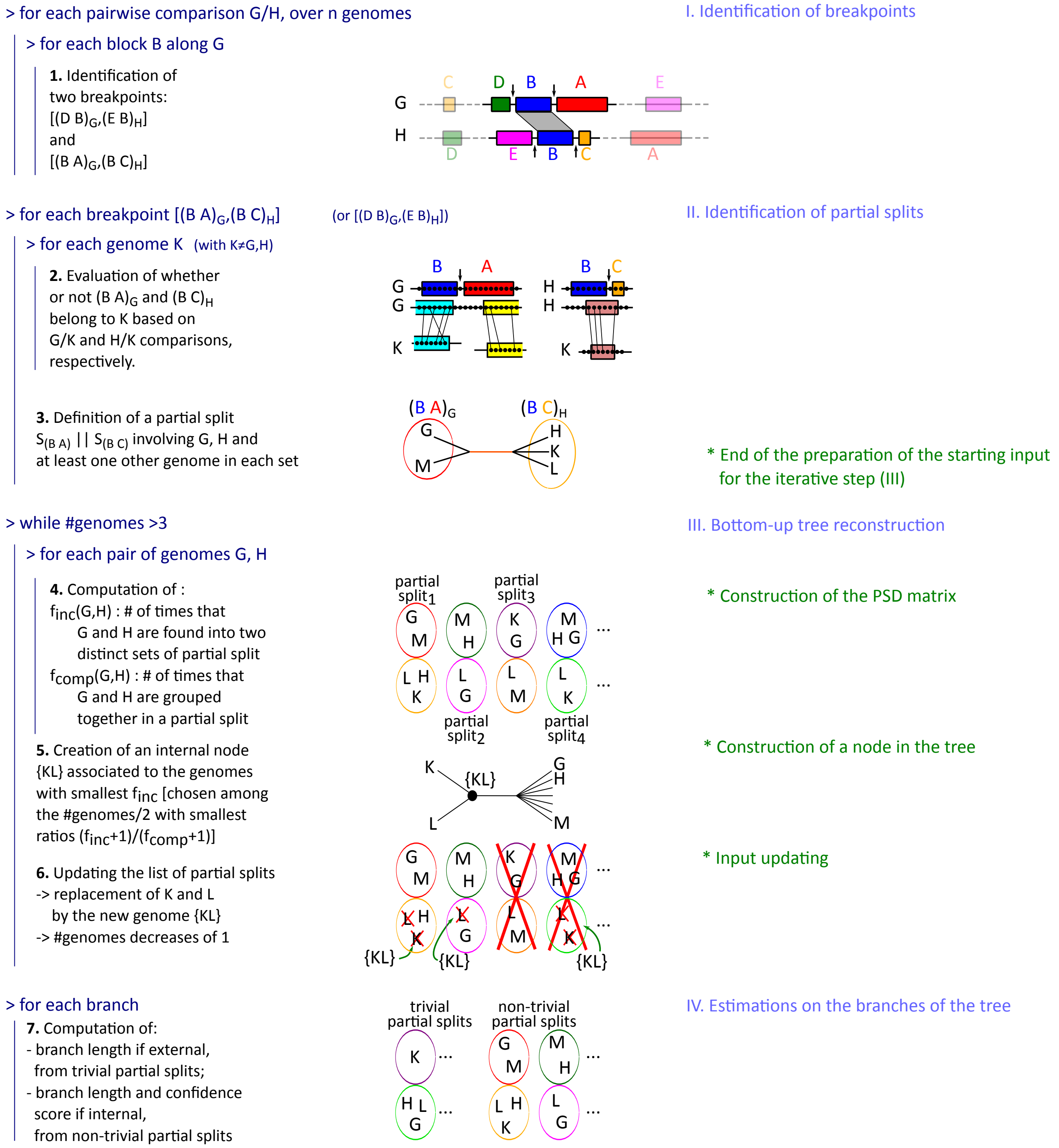


Figure 3: **PhyChro, MLGO and random reconstructions tested on simulated trees.**



```
> for each pairwise comparison G/H, over n genomes                          I. Identification of breakpoints
    > for each block B along G
        1. Identification of two breakpoints: [(D B)$_G$,(E B)$_H$] and [(B A)$_G$,(B C)$_H$]

> for each breakpoint [(B A)$_G$,(B C)$_H$]   (or [(D B)$_G$,(E B)$_H$])   II. Identification of partial splits
    > 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
                                  * End of the preparation of the starting input for the iterative step (III)

> while #genomes >3                                                         III. Bottom-up tree reconstruction
    > for each pair of genomes G, H
        4. Computation of :                                                 * Construction of the PSD matrix
           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          * Construction of a node in the tree
       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                                  * Input updating
       -> replacement of K and L
          by the new genome {KL}
       -> #genomes decreases of 1

> for each branch                                                           IV. Estimations on the branches of the tree
    7. Computation of:
    - branch length if external,
      from trivial partial splits;
    - branch length and confidence
      score if internal,
      from non-trivial partial splits
```

Figure 4: **PhyChro algorithm.**

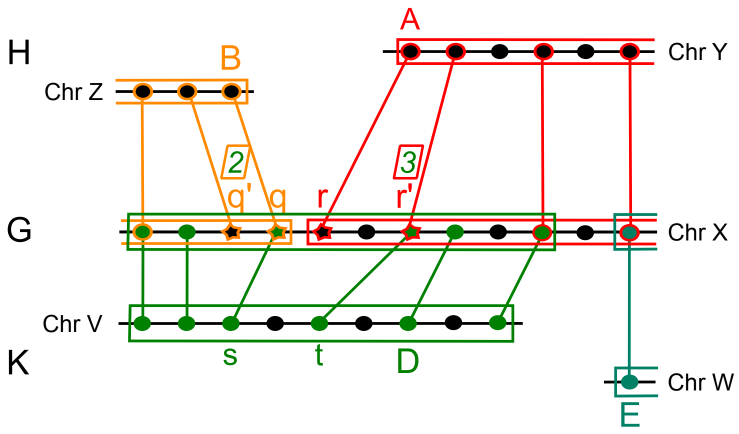


Figure 5: **Localization of the adjacency** $(B\ A)_G$ **in the genome** K**.**

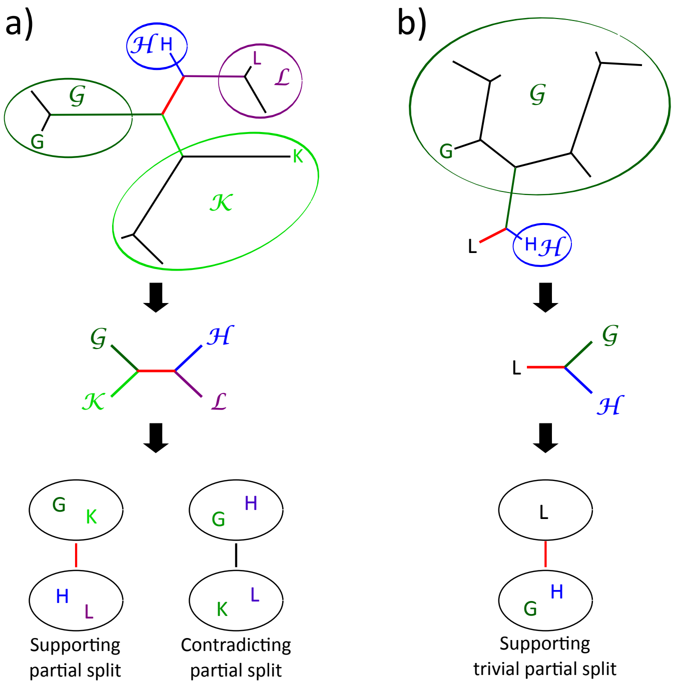


Figure 6: **Examples of partial splits supporting or contradicting the existence of a given branch.**

Table 1: **PhyChro algorithm**

Parts I and II

- Preparation of the input for the iterative step

```
for each synteny block do
    identify the associated pair of breakpoints
end
for each pair of breakpoints do
    construct the associated partial splits
end
```

Part III

- Tree reconstruction

```
while # genomes > 3 do
    • construct the Partial Split Distance matrix based on all partial splits;
    • chose the pair G,H based on minimal distance;
    • update the partial split list, where G,H are excluded;
end
```

Part IV

- Estimation of the branch length

```
for each branch do
    if the branch is external then
        compute branch length from trivial partial splits;
    else
        compute branch length and confidence score from non trivial partial splits;
    end
end
```

Supplementary Material

How to compute branch length and confidence score.

For internal branches

Each internal branch b is flanked by four branches leading to four sets of genomes, two for each extremity of b (**Figure 6a**). This means that we can describe b by a double pair of genome sets $[(\mathcal{G},\mathcal{K}),(\mathcal{H},\mathcal{Q})]$, where $\mathcal{G}$ and $\mathcal{K}$ are two sets of genomes lying in the maximal subtree excluding b and containing one of its extremities, and where $\mathcal{H}$ and $\mathcal{Q}$ are the two sets of genomes lying in the maximal subtree excluding b and containing the other extremity of b. Depending on these two pairs of sets, we can determine whether a given partial split $S_{(B\,A)} \| S_{(B\,C)}$ supports the branch b or not. To do so, we define the predicate $\mathrm{p}_{\mathrm{support}}$ as follows:

$$\mathrm{p}_{\mathrm{support}}\left(S_{(B\,A)} \| S_{(B\,C)}, b\right) = \mathrm{p}_{\mathrm{support}}\left(S_{(B\,A)} \| S_{(B\,C)}, \left[(\mathcal{G},\mathcal{K}),(\mathcal{H},\mathcal{Q})\right]\right) =$$
$$\begin{aligned}
&\exists G.\,(G{\in}\mathcal{G}{\wedge}G{\in}S_{(B\,A)}){\wedge}\exists G.\,(G{\in}\mathcal{K}{\wedge}G{\in}S_{(B\,A)}){\wedge}\\
&\nexists G.\,(G{\in}\mathcal{H}{\wedge}G{\in}S_{(B\,A)}){\wedge}\nexists G.\,(G{\in}\mathcal{Q}{\wedge}G{\in}S_{(B\,A)}){\wedge}\\
&\exists G.\,(G{\in}\mathcal{H}{\wedge}G{\in}S_{(B\,C)}){\wedge}\exists G.\,(G{\in}\mathcal{Q}{\wedge}G{\in}S_{(B\,C)}){\wedge}\\
&\nexists G.\,(G{\in}\mathcal{G}{\wedge}G{\in}S_{(B\,C)}){\wedge}\nexists G.\,(G{\in}\mathcal{K}{\wedge}G{\in}S_{(B\,C)})
\end{aligned}$$

and say that $S_{(B\,A)} \| S_{(B\,C)}$ supports b if the predicate is true. The left panel of **Figure 6a** gives an example of a partial split that supports a given branch.

Similarly, we can determine when a given partial split contradicts a given internal branch b, meaning that this partial split would imply at least two rearrangements occurring in two different branches of the tree and involving the same block extremities. To do so, we define the predicate $\mathrm{p}_{\mathrm{contradict}}$ as follows:

$$\mathrm{p}_{\mathrm{contradict}}\left(S_{(B\,A)} \| S_{(B\,C)}, \left[(\mathcal{G},\mathcal{K}),(\mathcal{H},\mathcal{Q})\right]\right) =$$
$$\begin{aligned}
&(\exists G.\,((G{\in}\mathcal{G}{\vee}G{\in}\mathcal{K}){\wedge}G{\in}S_{(B\,A)}){\wedge}\exists G.\,((G{\in}\mathcal{H}{\vee}G{\in}\mathcal{Q}){\wedge}G{\in}S_{(B\,A)})){\vee}\\
&\quad(\exists G.\,((G{\in}\mathcal{G}{\vee}G{\in}\mathcal{K}){\wedge}G{\in}S_{(B\,C)}){\wedge}\exists G.\,((G{\in}\mathcal{H}{\vee}G{\in}\mathcal{Q}){\wedge}G{\in}S_{(B\,C)}))
\end{aligned}$$

and say that $S_{(B\,A)} \| S_{(B\,C)}$ contradicts b if the predicate is true. The right panel of **Figure 6a** gives an example of a partial split that contradicts a given branch.

For each internal branch b, we define the confidence score of b as:

$$\mathrm{cS}(b) = \frac{\#\mathrm{Support}(b) + 1}{\#\mathrm{Support(b)} + \#\mathrm{Contradict(b)} + 1}$$

$$\text{where } \#\mathrm{Support}(b) = |\{S_{(B\,A)} \| S_{(B\,C)} : \mathrm{p}_{\mathrm{support}}(S_{(B\,A)} \| S_{(B\,C)}, b)\}|$$
$$\text{and } \#\mathrm{Contradict}(b) = |\{S_{(B\,A)} \| S_{(B\,C)} : \mathrm{p}_{\mathrm{contradict}}(S_{(B\,A)} \| S_{(B\,C)}, b)\}|$$

Also, we would like to compute the length of an internal branch as the number of breakpoints that would be identified if the two ancestral genomes associated to the extremities of b were compared. To estimate this number, we use the predicate $\mathrm{p}_{\mathrm{support}}$, since every partial split supporting b is indeed associated to one of these breakpoints. However, the calculation is not that simple, since each breakpoint is associated to several partial splits rather than to a single one. For instance, given two sets of genomes, $S_{(B\,A)}$ and $S_{(B\,C)}$, regrouping the actual genomes that have conserved the two adjacencies of a given breakpoint $[(B\,A)_G, (B\,C)_H]$ respectively, each comparison between two genomes from these two sets should lead to the identification of a partial split associated to this breakpoint. There are $|S_{(B\,A)}| * |S_{(B\,C)}|$ such comparisons. Therefore, the number of breakpoints is estimated from the number of partial splits by weighting them according to the maximal number of partial splits that could be associated to the same breakpoint:

$$\mathrm{length}(b) = \sum_{\substack{S_{(B\,A)} \| S_{(B\,C)} \text{ s.t.} \\ \mathrm{p}_{\mathrm{support}}(S_{(B\,A)} \| S_{(B\,C)}, b)}} \frac{1}{|S_{(B\,A)}| * |S_{(B\,C)}|}$$

For terminal branches

Each terminal branch, leading to a genome K, is characterized by two branches leading to two sets of genomes (**Figure 6b**), and therefore, it can be described by a pair $[K, (\mathcal{G}, \mathcal{H})]$, where $\mathcal{G}$ and $\mathcal{H}$ are the two sets of genomes. Depending on these sets, we can determine whether a given trivial partial split, $S_{(B\,A)} \| S_{(B\,C)}$, where $S_{(B\,A)}$ is constituted by an unique genome, supports this branch or not. This is done with the predicate $\mathrm{t}_{\mathrm{support}}$ defined as follows:

$$\mathrm{t}_{\mathrm{support}}\Big(S_{(B\,A)} \| S_{(B\,C)}, \big[K, (\mathcal{G}, \mathcal{H})\big]\Big) = $$
$$K {\in} S_{(B\,A)} {\wedge} (\exists G.\,(G{\in}\mathcal{G} {\wedge} G{\in}S_{(B\,C)})) {\wedge} (\exists G.\,(G{\in}\mathcal{H} {\wedge} G{\in}S_{(B\,C)}))$$

We say that $S_{(B\,A)} \| S_{(B\,C)}$ supports the external branch b if the predicate is true. **Figure 6b** gives an example of trivial partial split that supports a given external branch.

From this predicate, we compute the length of an external branch b as follows:

$$\text{length}(b) = \sum_{\substack{S_{(B\,A)} \| S_{(B\,C)} \text{ s.t.} \\ \text{t}_{\text{support}}(S_{(B\,A)} \| S_{(B\,C)}, b)}} \frac{1}{|S_{(B\,C)}|}$$

Supplementary Tables

Table S1: List of 21 yeasts and 16 vertebrates with a high quality assembled genome.

Class	*Species*	Genome size (Mb)	# of Chr.	# of Scaf.	# of Gen.	Reference
Saccharomycetes	*Candida albicans*	14.3	8	8	6182	(Jones *et al.*, 2004)
Saccharomycetes	*Candida dubliniensis*	14.6	8^{1}	8^{1}	5858	(Jackson *et al.*, 2009)
Saccharomycetes	*Candida glabrata*	12.3	13	13	5202	(Dujon *et al.*, 2004)
Saccharomycetes	*Candida parapsilosis*	13.1	7	14	5608	(Butler *et al.*, 2009)
Saccharomycetes	*Candida tropicalis*	14.6	8	20	6253	(Butler *et al.*, 2009)
Saccharomycetes	*Clavispora lusitaniae*	12.1	8	8	5936	(Butler *et al.*, 2009)
Saccharomycetes	*Debaryomyces hansenii*	12.2	7	7	6272	(Dujon *et al.*, 2004)
Saccharomycetes	*Eremothecium gossypii*	8.7	7	7	4768	(Dietrich *et al.*, 2004)
Saccharomycetes	*Kluyveromyces lactis*	10.7	6	6	5076	(Dujon *et al.*, 2004)
Saccharomycetes	*Lachancea kluyveri*	11.3	8	8	5321	(Souciet *et al.*, 2009)
Saccharomycetes	*Lachancea thermotolerans*	10.4	8	8	5092	(Souciet *et al.*, 2009)
Saccharomycetes	*Lachancea waltii*	10.7	8	10	6614	(Kellis *et al.*, 2004)
Saccharomycetes	*Lodderomycese elongisporus*	15.5	8	22	5795	(Butler *et al.*, 2009)
Saccharomycetes	*Naumovozyma castelii*	11.2	9	10	5648	(Gordon *et al.*, 2011)
Saccharomycetes	*Pichia guilliermondii*	10.6	8	9	5920	(Butler *et al.*, 2009)
Saccharomycetes	*Pichia pastoris*	9.4	4	6	5077	(De Schutter *et al.*, 2009)
Saccharomycetes	*Pichia stipitis*	15.4	8	9	5818	(Jeffries *et al.*, 2007)
Saccharomycetes	*Saccharomyces cerevisiae*	12.1	16	16	6664	(Goffeau *et al.*, 1996)
Saccharomycetes	*Torulaspora delbrueckii*	9.2	6	8	4972	(Gomez-Angulo *et al.*, 2015)
Saccharomycetes	*Yarrowia lipolytica*	20.5	6	6	6448	(Dujon *et al.*, 2004)
Saccharomycetes	*Zygosaccharomyces rouxii*	9.8	7	7	4991	(Souciet *et al.*, 2009)
Sauropsida	*Anolis carolinensis*	1780	18	18	18595	(Alföldi *et al.*, 2011)
Mammalia	*Bos taurus*	2500	30	30	24293	(Zimin *et al.*, 2009)
Mammalia	*Canis familiaris*	2400	39	39	19014	(Lindblad-Toh *et al.*, 2005)
Actinopterygii	*Danio rerio*	1700	25	25	22940	(Howe *et al.*, 2013)
Mammalia	*Equus caballus*	2689	32	32	20257	(Wade *et al.*, 2009)
Aves	*Gallus gallus*	1000	40^{2}	29	15308	(The International Chicken Genome Sequencing Consortium, 2004)
Mammalia	*Homo sapiens*	3080	23	23	19439	(The International Human Genome Sequencing Consortium, 2001)
Mammalia	*Macaca mulatta*	2871	22	21	21023	(The Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007)
Marsupialia	*Monodelphis domestica*	3475	9	9	18640	(Mikkelsen *et al.*, 2007)
Mammalia	*Mus musculus*	2644	20	20	21923	(The Mouse Genome Sequencing Consortium, 2002)
Actinopterygii	*Oryzias latipes*	800	24	24	17445	(Kasahara *et al.*, 2007)
Mammalia	*Pan troglodytes*	3100	24	24	19125	(The Chimpanzee Sequencing and Analysis Consortium, 2005)
Mammalia	*Rattus Norvegicus*	3000	21	21	22925	(The Rat Genome Sequencing Project Consortium, 2004)
Mammalia	*Sus scrofa*	2800	19	19	21630	(Wernersson *et al.*, 2005)
Aves	*Taeniopygia guttata*	2644	28	28	12337	(Warren *et al.*, 2010)
Actinopterygii	*Tetraodon nigroviridis*	350	21	21	13580	(Jaillon *et al.*, 2004)

[a] Pseudochromosomes obtained by mapping onto *C. albicans* chromosomes (Jackson *et al.*, 2009).

[b] Including microchromosomes that were not assembled.

Table S2: 21 Yeast Species Summary

	Species Name	Short name	# chr	# scaffolds	References (* : centromere data)	Strain	Downloaded from	Issue Date (version)
[1]	*Candida albicans*	CAAL	8	8	(Butler *et al.*, 2009)*	SC5314	http://www.candidagen ome.org/	08-Jan-2012
[2]	*Candida dubliniensis*	CADU	8	8	(Jackson *et al.*, 2009; Padmanabhan *et al.*, 2008)*	CD36	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)
[3]	*Candida glabrata*	CAGL	13	13	(Dujon *et al.*, 2004)*	CDS138	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)
[4]	*Candida parapsilosis*	CAPA	7	14	(Butler *et al.*, 2009)	CDC317	http://www.broadinstitute.org http://www.ebi.ac.uk/	26-May-2011
[5]	*Candida tropicalis*	CATR	8	20	(Butler *et al.*, 2009)	MYA-3404	http://www.ebi.ac.uk/	26-May-2011 (Release 108)
[6]	*Clavispora lusitaniae*	CLLU	8	8	(Butler *et al.*, 2009; Lynch *et al.*, 2010)*	ATCC 42720	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)
[7]	*Debaryomyces hansenii*	DEHA	7	7	(Dujon *et al.*, 2004; Lynch *et al.*, 2010)*	CBS767	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)
[8]	*Eremothecium Gossypii*	ERGO	7	7	(Dietrich *et al.*, 2004)*	ATCC 10895	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)
[9]	*Kluyveromyces lactis*	KLLA	6	6	(Dujon *et al.*, 2004)*	NRRL Y-1140	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)
[10]	*Lachancea kluyveri*	LAKL	8	8	(Souciet *et al.*, 2009)*	CBS3082	http://www.genolevures. ac.uk/	05-Dec-2007
[11]	*Lachancea thermotolerans*	LATH	8	8	(Souciet *et al.*, 2009)*	CBS6340	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)
[12]	*Lachancea waltii*	LAWA	8	10	(Kellis *et al.*, 2004)*	NCYC2644	http://www.nature.com	2004
[13]	*Lodderomyces elongisporus*	LOEL	8	22	(Butler *et al.*, 2009)	NRRL YB-4239	http://www.ebi.ac.uk/	26-May-2011 (Release 108)
[14]	*Naumovozyma castelii*	NACA	9	10	(Gordon *et al.*, 2011)	CBS 4309	https://www.ncbi.nlm.nih.gov/	27-Feb-2015
[15]	*Pichia guilliermondii*	PIGU	8	9	(Butler *et al.*, 2009)	ATCC 6260	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)
[16]	*Pichia pastoris*	PIPA	4	6	(De Schutter *et al.*, 2009)	GS115	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)
[17]	*Pichia stipitis*	PIST	8	9	(Jeffries *et al.*, 2007; Lynch *et al.*, 2010)*	CBS 6054	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)
[18]	*Saccharomyces cerevisiae*	SACE	16	16	(Goffeau *et al.*, 1996)*		http://www.ensembl.org/	22-Nov-2011 (EF4)
[19]	*Torulaspora delbrueckii*	TODE	6	8	(Gomez-Angulo *et al.*, 2015)	NRRL Y-50541	https://www.ncbi.nlm.nih.gov/	30-Jul-2015
[20]	*Yarrowia lipolytica*	YALI	6	6	(Dujon *et al.*, 2004)*	CLIB122	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)
[21]	*Zygosaccharomyces rouxii*	ZYRO	7	7	(Souciet *et al.*, 2009)*	CBS732	http://www.ebi.ac.uk/	01-Dec-2011 (Release 110)

References

Alföldi J, Di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, et al. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, 477(7366):587.

Butler G, Rasmussen MD, Lin MF, Santos MAS, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, et al. 2009. Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature*, 459(7247):657–662.

De Schutter K, Lin YC, Tiels P, Van Hecke A, Glinka S, Weber-Lehmann J, Rouze P, Van de Peer Y, Callewaert N. 2009. Genome sequence of the recombinant protein production host Pichia pastoris. *Nature Biotechnology*, 27(6):561–566.

Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pöhlmann R, Luedi P, Choi S, et al. 2004. The *Ashbya gossypii* Genome as a Tool for Mapping the Ancient Saccharomyces cerevisiae Genome. *Science*, 304(5668):304–307.

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuvéglise C, Talla E, et al. 2004. Genome evolution in yeasts. *Nature*, 430(6995):35–44.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. 1996. Life with 6000 genes. *Science*, 274(5287):546–567.

Gomez-Angulo J, Vega-Alvarado L, Escalante-Garca Z, Grande R, Gschaedler-Mathis A, Amaya-Delgado L, et al. 2015. Genome sequence of *Torulaspora delbrueckii NRRL Y-50541*, isolated from mezcal fermentation. In *Genome announcements*, 3(4), e00438-15.

Gordon JL, Armisén D, Proux-Wéra E, Óhéigeartaigh SS, Byrne KP, Wolfe KH. 2011. Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. *Proceedings of the National Academy of Sciences, USA*, 108(50), 20024-20029.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, 496(7446):498.

Jackson AP, Gamble JA, Yeomans T, Moran GP, Saunders D, Harris D, Aslett M, Barrell JF, Butler G, Citiulo F, et al. 2009. Comparative genomics of the fungal pathogens candida dubliniensis and candida albicans. *Genome Research*, 19(12):2231–2244.

Jaillon O, Aury J.-M, Brunet F, Petit J.-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957.

Jeffries TW, Grigoriev IV, Grimwood J, Laplaza JM, Aerts A, Salamov A, Schmutz J, Lindquist E, Dehal P, Shapiro H, et al. 2007. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis. *Nature Biotechnology*, 25(3):319–326.

Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT, et al. 2004. The diploid genome sequence of *Candida albicans*. *Proceedings of the National Academy of Sciences, USA*, 101(19):7329–7334.

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447(7145):714.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, 428(6983):617–624.

Lindblad-Toh, K, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–819.

Lynch DB, Logue ME, Butler G, Wolfe KH. 2010. Chromosomal G+ C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. *Genome biology and evolution*, 2, 572-583.

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, et al. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature*, 447(7141):167–177.

Padmanabhan S, Thakur J, Siddharthan R, Sanyal K. 2008. Rapid evolution of Cse4p-rich centromeric DNA sequences in closely related pathogenic yeasts, *Candida albicans* and *Candida dubliniensis*. *Proceedings of the National Academy of Sciences, USA*, 105(50), 19797-19802.

Souciet JL, Dujon B, Gaillardin C, Johnston M, Baret PV, Cliften P, Sherman DJ, Weissenbach J, Westhof E, Wincker P, et al. 2009. Comparative genomics of protoploid Saccharomycetaceae. *Genome Researchearch*, 19:1696–1709.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.

The International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716.

The International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

The Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

The Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521.

The Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316(5822):222–234.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, et al. 2009. Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse. *Science*, 326(5954):865–867.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S, et al. 2010. The genome of a songbird. *Nature*, 464(7289):757–762.

Wernersson R, Schierup MH, Jorgensen FG, Gorodkin J, Panitz F, Saerfeldt HH, et al. 2005. Pigs in sequence space: a 0.66 X coverage pig genome survey based on shotgun sequencing. *BMC genomics*, 6(1):70.

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. 2009. A whole-genome assembly of the domestic cow, Bos taurus. *Genome biology*, 10(4):R42.

Supplementary Figures

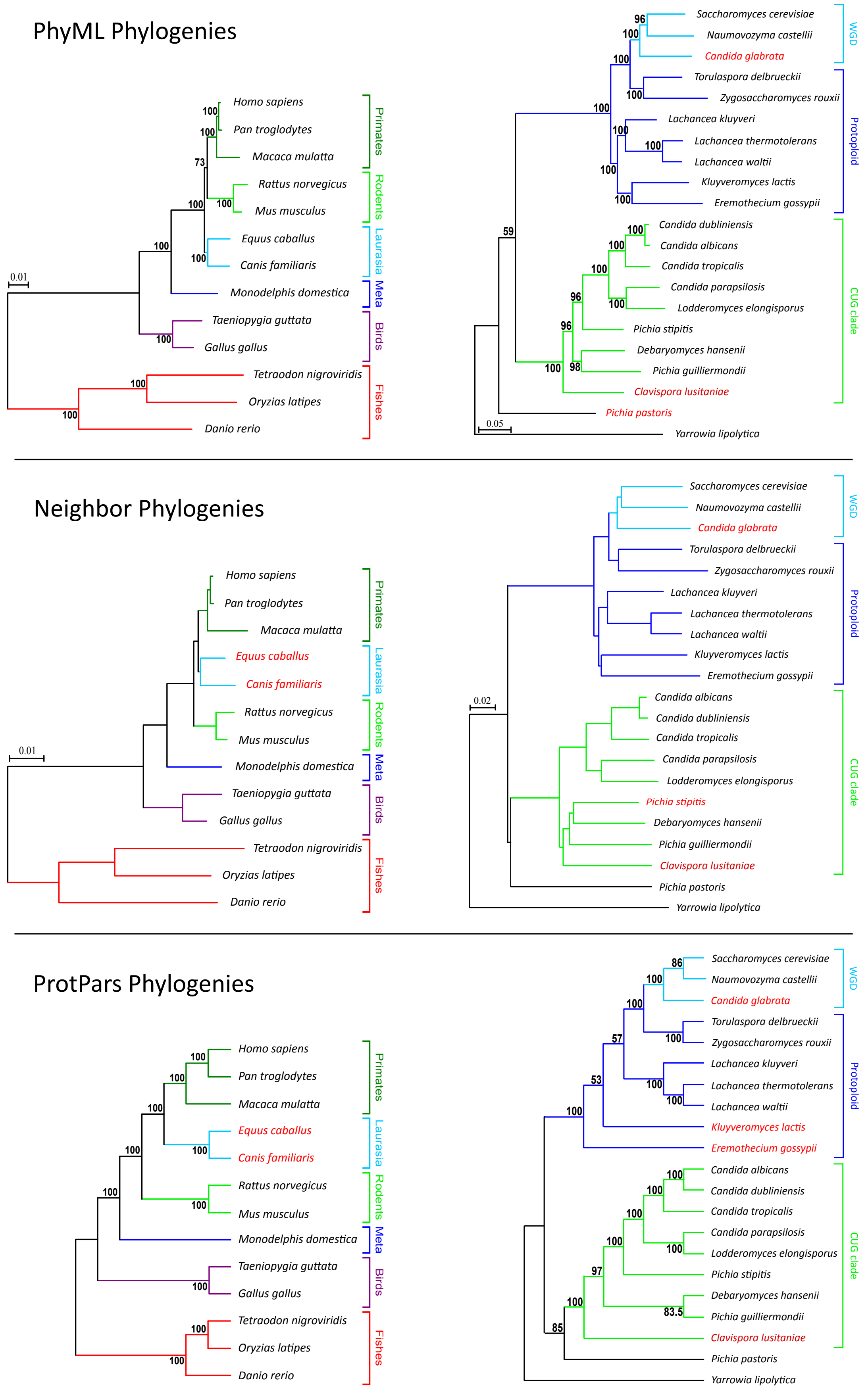


Figure S1: **Phylogenies obtained with PhyML, Neighbor and ProtPars methods.** Bootstrap values are given for the PhyML and the ProtPars phylogenies (for 100 resampled data sets). For PhyML and Neighbor reconstructions, branch length scales are indicated on the left of the trees.

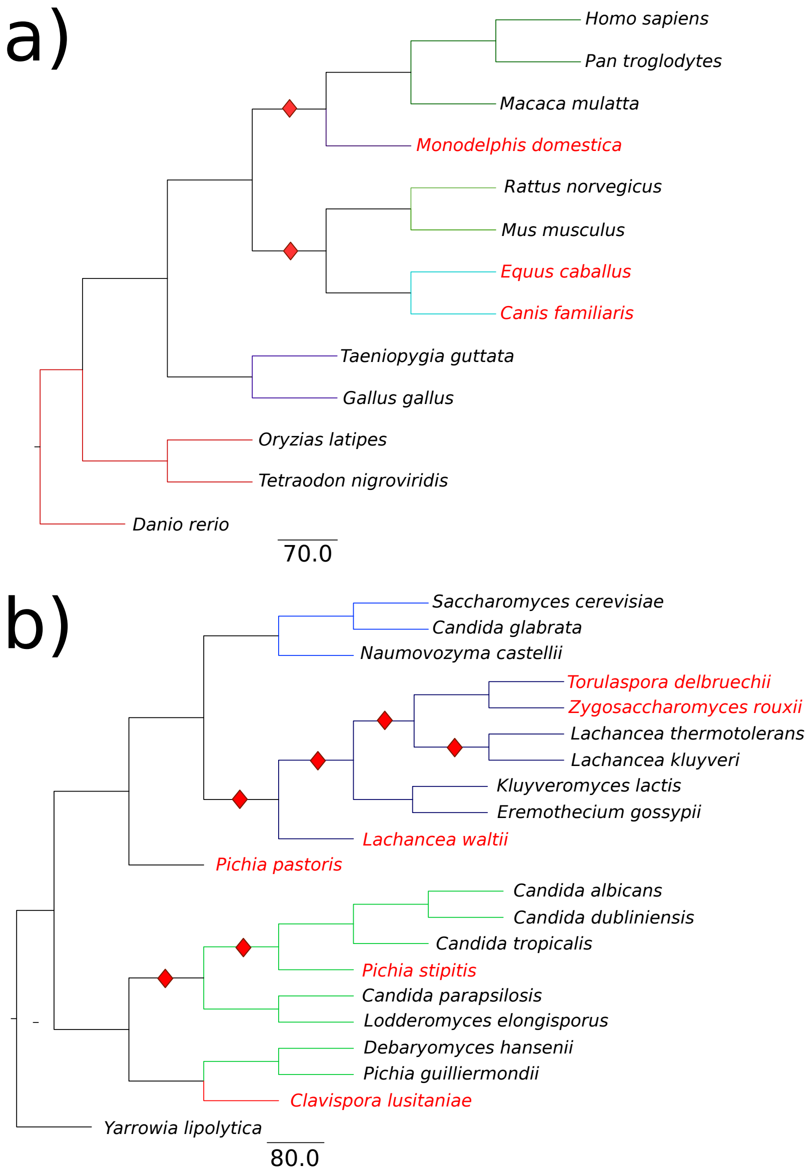
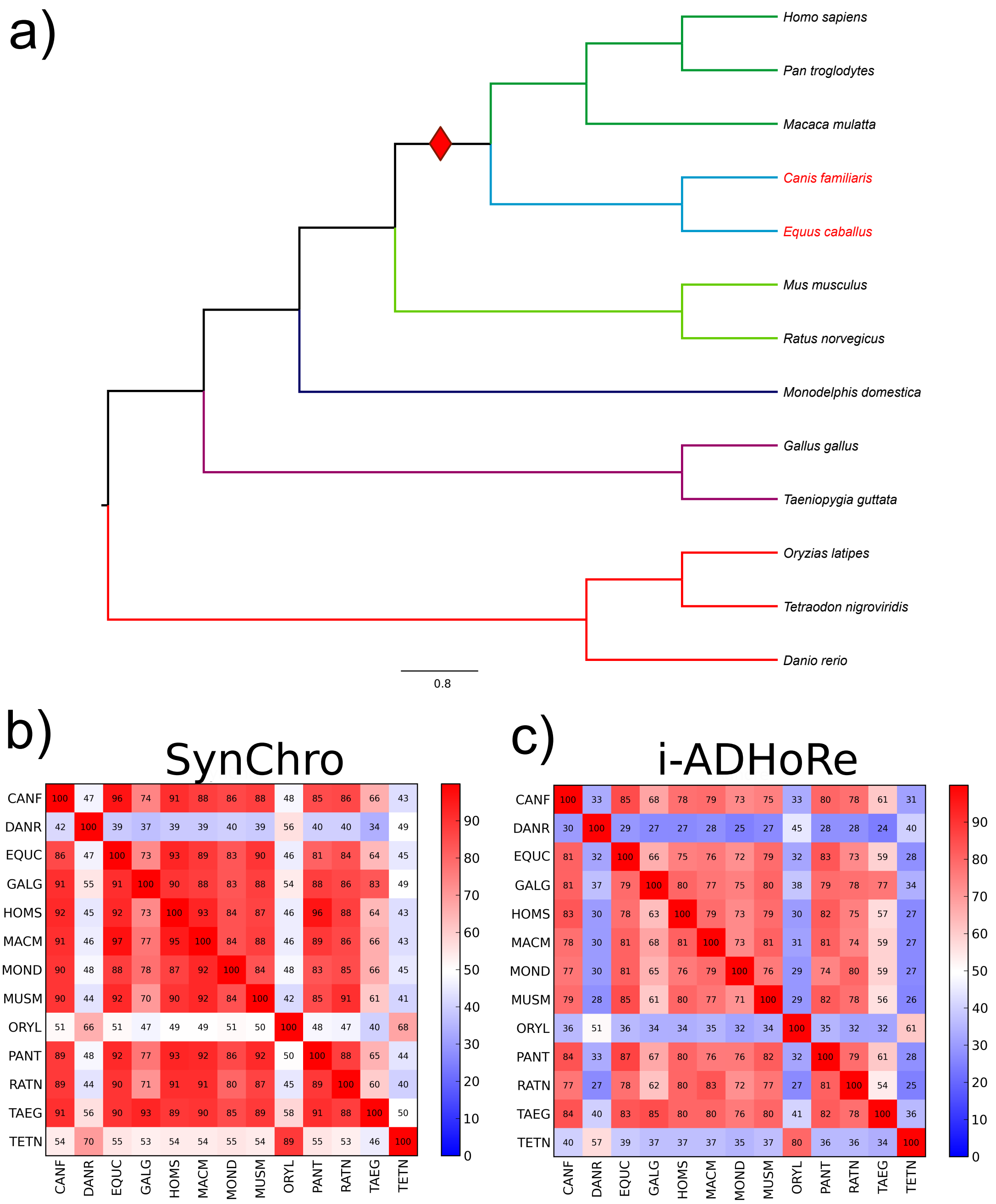


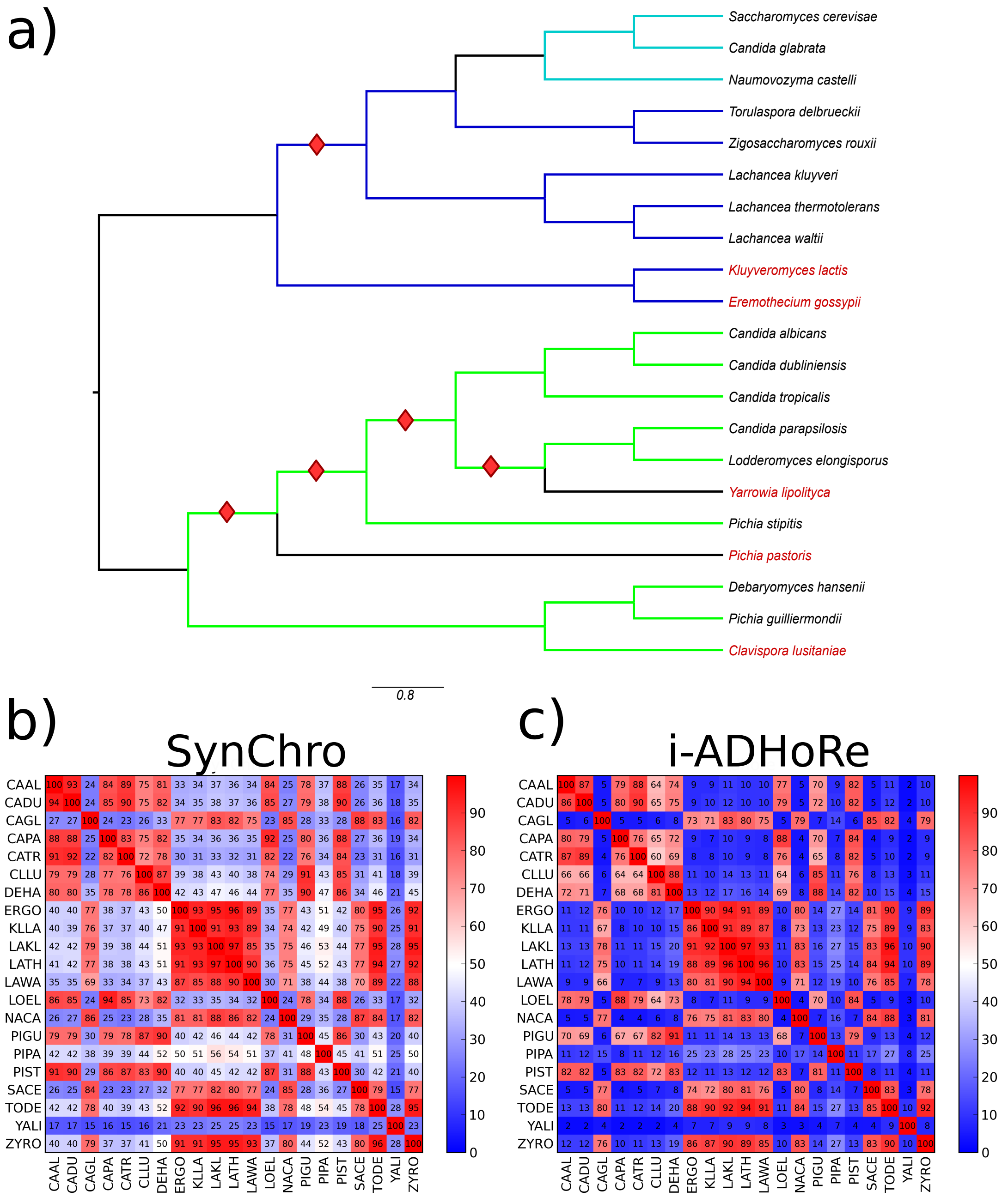
Figure S2: **Phylogenies obtained with MLGO method.** Tree reconstructions realized with MLGO, for vertebrates (**a**) and yeasts (**b**). Red losanges indicate erroneous splitting of the tree compared to the known one. Compare to **Figure 2**, where the colors assigned to branches are the same.



	CANF	DANR	EQUC	GALG	HOMS	MACM	MOND	MUSM	ORYL	PANT	RATN	TAEG	TETN
CANF	100	47	96	74	91	88	86	88	48	85	86	66	43
DANR	42	100	39	37	39	39	40	39	56	40	40	34	49
EQUC	86	47	100	73	93	89	83	90	46	81	84	64	45
GALG	91	55	91	100	90	88	83	88	54	88	86	83	49
HOMS	92	45	92	73	100	93	84	87	46	96	88	64	43
MACM	91	46	97	77	95	100	84	88	46	89	86	66	43
MOND	90	48	88	78	87	92	100	84	48	83	85	66	45
MUSM	90	44	92	70	90	92	84	100	42	85	91	61	41
ORYL	51	66	51	47	49	49	51	50	100	48	47	40	68
PANT	89	48	92	77	93	92	86	92	50	100	88	65	44
RATN	89	44	90	71	91	91	80	87	45	89	100	60	40
TAEG	91	56	90	93	89	90	85	89	58	91	88	100	50
TETN	54	70	55	53	54	54	55	54	89	55	53	46	100

	CANF	DANR	EQUC	GALG	HOMS	MACM	MOND	MUSM	ORYL	PANT	RATN	TAEG	TETN
CANF	100	33	85	68	78	79	73	75	33	80	78	61	31
DANR	30	100	29	27	27	28	25	27	45	28	28	24	40
EQUC	81	32	100	66	75	76	72	79	32	83	73	59	28
GALG	81	37	79	100	80	77	75	80	38	79	78	77	34
HOMS	83	30	78	63	100	79	73	79	30	82	75	57	27
MACM	78	30	81	68	81	100	73	81	31	81	74	59	27
MOND	77	30	81	65	76	79	100	76	29	74	80	59	27
MUSM	79	28	85	61	80	77	71	100	29	82	78	56	26
ORYL	36	51	36	34	34	35	32	34	100	35	32	32	61
PANT	84	33	87	67	80	76	76	82	32	100	79	61	28
RATN	77	27	78	62	80	83	72	77	27	81	100	54	25
TAEG	84	40	83	85	80	80	76	80	41	82	78	100	36
TETN	40	57	39	37	37	37	35	37	80	36	36	34	100

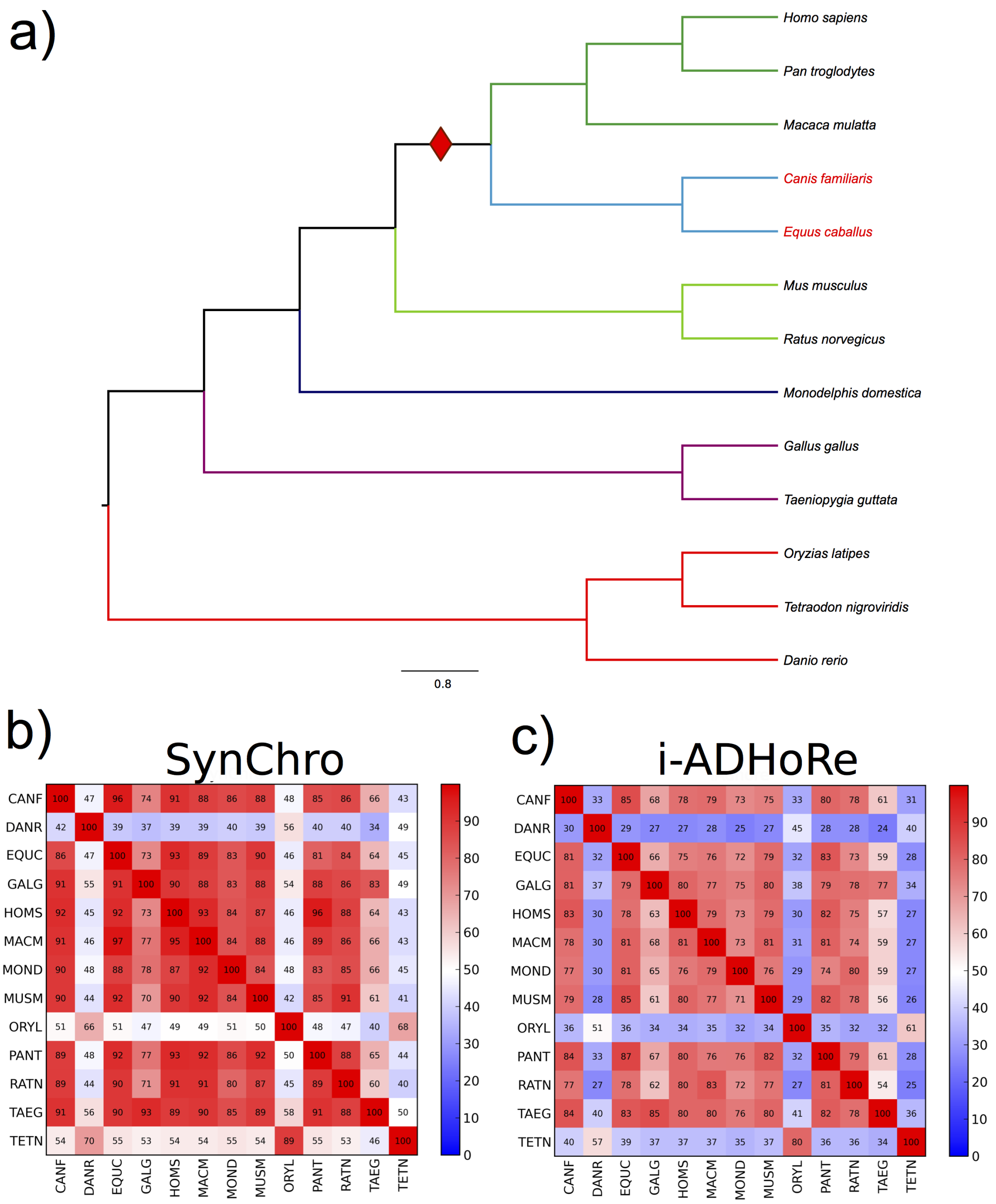
Figure S3: **Vertebrate phylogeny obtained with PhyChro based on i-ADHoRe synteny blocks. a)** Tree topology reconstruction realized with PhyChro and based on i-ADHoRe synteny blocks. The red losange indicates an error in the phylogenetic reconstruction. Colors correspond to the ones used in **Figure 2a**. Compare it to **Figure 2a**. **b)** Matrix representing the synteny blocks coverage of the vertebrate genomes after pairwise comparison, where synteny blocks are obtained with SynChro (run with $\Delta = 3$). At row X and column Y, the number in the cell of the matrix corresponds to the coverage of genome Y after comparison with genome X, that is the percentage of the number of genes in the genome that belong to synteny blocks. The scale indicates the coverage level and goes from low (blue) to high (red). **c)** Matrix representing the synteny blocks coverage of the vertebrate genomes after pairwise comparison, where synteny blocks are obtained with i-ADHoRe.



	CAAL	CADU	CAGL	CAPA	CATR	CLLU	DEHA	ERGO	KLLA	LAKL	LATH	LAWA	LOEL	NACA	PIGU	PIPA	PIST	SACE	TODE	YALI	ZYRO
CAAL	100	93	24	84	89	75	81	33	34	37	36	34	84	25	78	37	88	26	35	17	34
CADU	94	100	24	85	90	75	82	34	35	38	37	36	85	27	79	38	90	26	36	18	35
CAGL	27	27	100	24	23	26	33	77	77	83	82	75	23	85	28	33	28	88	83	16	82
CAPA	88	88	25	100	83	75	82	35	34	36	36	35	92	25	80	36	88	27	36	19	34
CATR	91	92	22	82	100	72	78	30	31	33	32	31	82	22	76	34	84	23	31	16	31
CLLU	79	79	28	77	76	100	87	39	38	43	40	38	74	29	91	43	85	31	41	18	39
DEHA	80	80	35	78	78	86	100	42	43	47	46	44	77	35	90	47	86	34	46	21	45
ERGO	40	40	77	38	37	43	50	100	93	95	96	89	35	77	43	51	42	80	95	26	92
KLLA	40	39	76	37	37	40	47	91	100	91	93	89	34	74	42	49	40	75	90	25	91
LAKL	42	42	79	39	38	44	51	93	93	100	97	85	35	75	46	53	44	77	95	28	95
LATH	41	42	77	38	38	43	51	91	93	97	100	90	36	75	45	52	43	77	94	27	92
LAWA	35	35	69	33	34	37	43	87	85	88	90	100	30	71	38	44	38	70	89	22	88
LOEL	86	85	24	94	85	73	82	32	33	35	34	32	100	24	78	34	88	26	33	17	32
NACA	26	27	86	25	23	28	35	81	81	88	86	82	24	100	29	35	28	87	84	17	82
PIGU	79	79	30	79	78	87	90	40	42	46	44	42	78	31	100	45	86	30	43	20	40
PIPA	42	42	38	39	39	44	52	50	51	56	54	51	37	41	48	100	45	41	51	25	50
PIST	91	90	29	86	87	83	90	40	40	45	42	42	87	31	88	43	100	30	42	21	40
SACE	26	25	84	23	23	27	32	77	77	82	80	77	24	85	28	36	27	100	79	15	77
TODE	42	42	78	40	39	43	52	92	90	96	96	94	38	78	48	54	45	78	100	28	95
YALI	17	17	15	16	15	16	21	23	23	25	25	23	15	17	19	23	19	18	25	100	23
ZYRO	40	40	79	37	37	41	50	91	91	95	95	93	37	80	44	52	43	80	96	28	100

	CAAL	CADU	CAGL	CAPA	CATR	CLLU	DEHA	ERGO	KLLA	LAKL	LATH	LAWA	LOEL	NACA	PIGU	PIPA	PIST	SACE	TODE	YALI	ZYRO
CAAL	100	87	5	79	88	64	74	9	9	11	10	10	77	5	70	9	82	5	11	2	10
CADU	86	100	5	80	90	65	75	9	10	12	10	10	79	5	72	10	82	5	12	2	10
CAGL	5	6	100	5	5	6	8	73	71	83	80	75	5	79	7	14	6	85	82	4	79
CAPA	80	79	5	100	76	65	72	9	7	10	9	8	88	4	70	7	84	4	9	2	9
CATR	87	89	4	76	100	60	69	8	8	10	9	8	76	4	65	8	82	5	10	2	9
CLLU	66	66	6	64	64	100	88	11	10	14	13	11	64	6	85	11	76	8	13	3	11
DEHA	72	71	7	68	68	81	100	13	12	17	16	14	69	8	88	14	82	10	15	4	15
ERGO	11	12	76	10	10	12	17	100	90	94	91	89	10	80	14	27	14	81	90	9	89
KLLA	11	11	67	8	10	10	15	86	100	91	89	87	8	73	13	23	12	75	89	9	83
LAKL	13	13	78	11	11	15	20	91	92	100	97	93	11	83	16	27	15	83	96	10	90
LATH	11	12	75	10	11	14	19	88	89	96	100	96	10	83	15	25	14	84	94	10	89
LAWA	9	9	66	7	7	9	13	80	81	90	94	100	9	71	12	19	10	76	85	7	78
LOEL	78	79	5	88	79	64	73	8	7	11	9	9	100	4	70	10	84	5	9	3	10
NACA	5	5	77	4	4	6	8	76	75	81	83	80	4	100	7	16	7	84	88	3	81
PIGU	70	69	6	67	67	82	91	11	11	14	13	13	68	7	100	13	79	9	13	4	12
PIPA	11	12	15	8	11	12	16	25	23	28	25	23	10	17	14	100	11	17	27	8	25
PIST	82	82	5	83	82	72	83	12	11	13	12	12	83	7	81	11	100	8	11	4	11
SACE	5	5	77	4	5	7	8	74	72	80	81	76	5	80	8	14	7	100	83	3	78
TODE	13	13	80	11	12	14	20	88	90	92	94	91	11	84	15	27	13	85	100	10	92
YALI	2	2	4	2	2	2	4	7	7	9	9	8	3	3	4	7	4	4	9	100	8
ZYRO	12	12	76	10	11	13	19	86	87	90	89	85	10	83	14	27	13	83	90	10	100

Figure S4: **Yeasts phylogeny obtained with PhyChro based on i-ADHoRe synteny blocks. a)** Tree topology reconstruction realized with PhyChro and based on i-ADHoRe synteny blocks. Red losanges indicate errors in the phylogenetic reconstruction. Colors correspond to the ones used in **Figure 2b**. Compare it to **Figure 2b**. **b)** Matrix representing the synteny blocks coverage of the yeast genomes after pairwise comparison, where synteny blocks are obtained with SynChro (run with $\Delta = 3$). At row X and column Y, the number in the cell of the matrix corresponds to the coverage of genome Y after comparison with genome X, that is the percentage of the number of genes in the genome that belong to synteny blocks. The scale indicates the coverage level and goes from low (blue) to high (red). **c)** Matrix representing the synteny blocks coverage of the yeast genomes after pairwise comparison, where synteny blocks are obtained with i-ADHoRe.



SynChro

	CANF	DANR	EQUC	GALG	HOMS	MACM	MOND	MUSM	ORYL	PANT	RATN	TAEG	TETN
CANF	100	47	96	74	91	88	86	88	48	85	86	66	43
DANR	42	100	39	37	39	39	40	39	56	40	40	34	49
EQUC	86	47	100	73	93	89	83	90	46	81	84	64	45
GALG	91	55	91	100	90	88	83	88	54	88	86	83	49
HOMS	92	45	92	73	100	93	84	87	46	96	88	64	43
MACM	91	46	97	77	95	100	84	88	46	89	86	66	43
MOND	90	48	88	78	87	92	100	84	48	83	85	66	45
MUSM	90	44	92	70	90	92	84	100	42	85	91	61	41
ORYL	51	66	51	47	49	49	51	50	100	48	47	40	68
PANT	89	48	92	77	93	92	86	92	50	100	88	65	44
RATN	89	44	90	71	91	91	80	87	45	89	100	60	40
TAEG	91	56	90	93	89	90	85	89	58	91	88	100	50
TETN	54	70	55	53	54	54	55	54	89	55	53	46	100

i-ADHoRe

	CANF	DANR	EQUC	GALG	HOMS	MACM	MOND	MUSM	ORYL	PANT	RATN	TAEG	TETN
CANF	100	33	85	68	78	79	73	75	33	80	78	61	31
DANR	30	100	29	27	27	28	25	27	45	28	28	24	40
EQUC	81	32	100	66	75	76	72	79	32	83	73	59	28
GALG	81	37	79	100	80	77	75	80	38	79	78	77	34
HOMS	83	30	78	63	100	79	73	79	30	82	75	57	27
MACM	78	30	81	68	81	100	73	81	31	81	74	59	27
MOND	77	30	81	65	76	79	100	76	29	74	80	59	27
MUSM	79	28	85	61	80	77	71	100	29	82	78	56	26
ORYL	36	51	36	34	34	35	32	34	100	35	32	32	61
PANT	84	33	87	67	80	76	76	82	32	100	79	61	28
RATN	77	27	78	62	80	83	72	77	27	81	100	54	25
TAEG	84	40	83	85	80	80	76	80	41	82	78	100	36
TETN	40	57	39	37	37	37	35	37	80	36	36	34	100

Figure S5: **Distribution of blocks obtained by i-ADHoRe and SynChro on vertebrate and yeast species.** The x-axis reports the number of genes in a block (block size) and the y-axis reports the number of blocks of a given size found in all pairwise comparisons between vertebrate (left) and yeast (right) species. Blocks of size ≥ 21 are added up in the last columns for both i-ADHoRe (red) and SynChro (black).

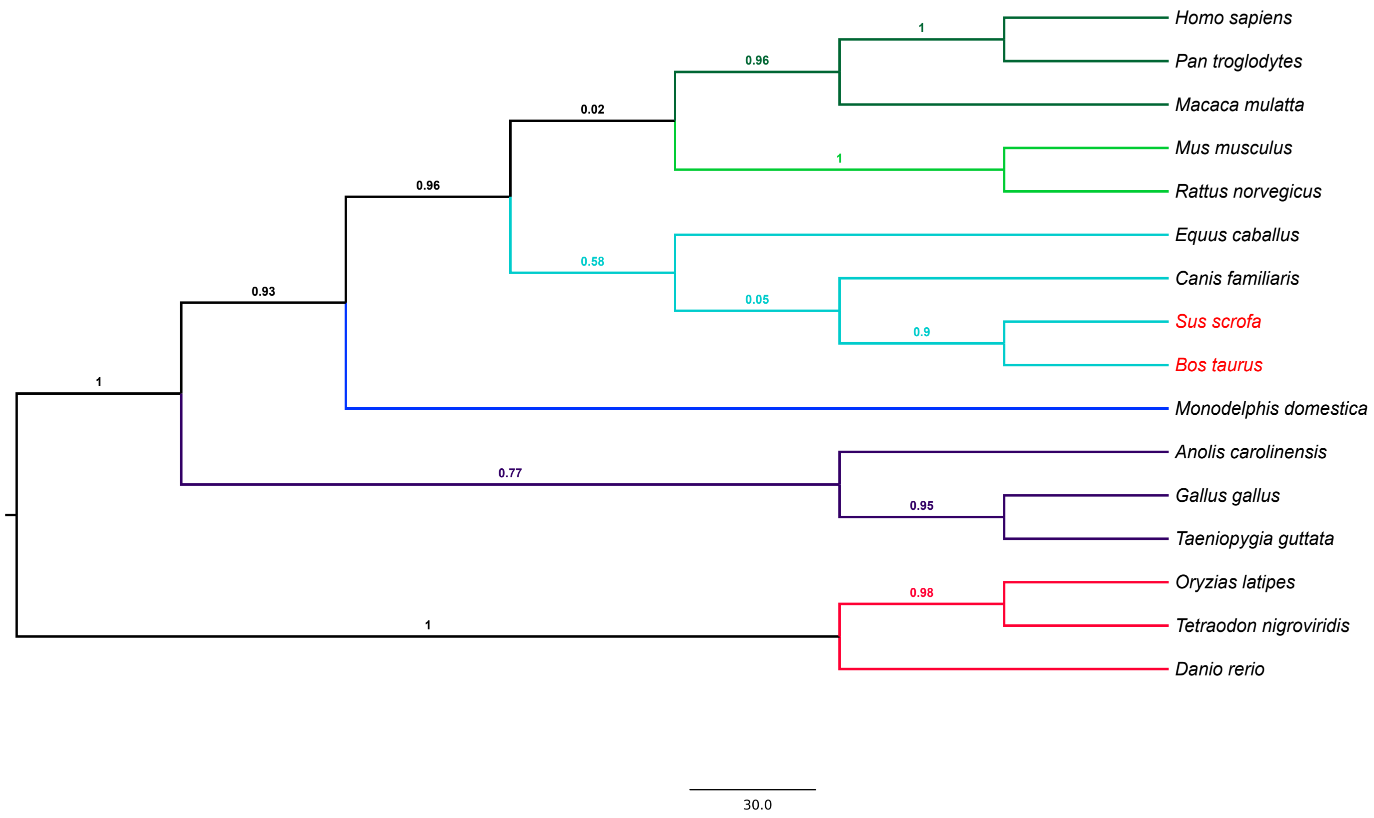


Figure S6: **Phylogenies obtained with PhyChro for the 13 vertebrates plus the cow, the pig and the lizard.** See legend in **Figure 2**. In this reconstruction, the position of horse and dog is not well supported by PhyChro confidence score (0.05) and this suggests the possibility of alternative topological nestings. By looking at the data produced by PhyChro partial split analysis for the reconstruction of the tree in **Figure 2a**, note that horse and dog are sister genomes with f_{inc}(dog,horse)= 38, a value that is much larger than what is ideally expected for sister genomes, that is 0. It corresponds to a large number of partial splits separating them, suggesting that the dog and the horse might have been particularly sensitive to convergent rearrangements (homoplasy) or to the accumulation of small inversions. Therefore, even if we assume that they actually are sister genomes in the tree including the cow and the pig, they are likely to be found further away from each other than they are from the cow and the pig. Indeed, this is what PhyChro finds: f_{inc}(dog,horse)= 38, f_{inc}((cow,pig),dog)= 10 and f_{inc}((cow,pig),horse)= 21. Note that in (Romiguier et al., *Molecular Biology and Evolution*, 2013), the position of ((cow,pig),(dog,horse)) is well supported. This case illustrates well, on the one hand, the limits of PhyChro, which is sensitive to rearrangement convergences when such events exist (note that from the very good quality of PhyChro tree reconstructions, it appears that such rearrangement convergences are few), and on the other hand, its power, since its output can help to build a good understanding of the tree topology, and most of all, whether one should have confidence or not in the reconstructed topology.

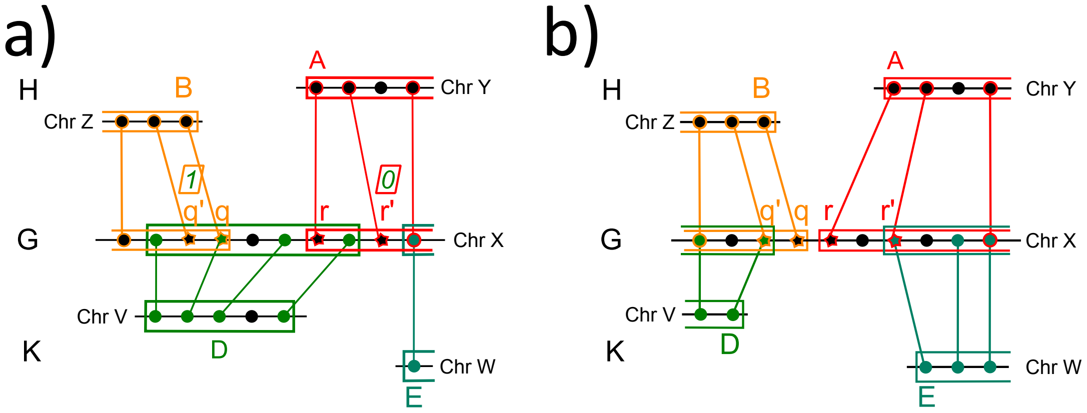


Figure S7: **Illustration of two possible localizations of an adjacency** $(B\,A)_G$ **in a genome** K**.** Genes are indicated by dots or stars. Stars, in G, are used for the two first (q' and q) and last (r and r') anchors of blocks B and A respectively, in the comparison G/H. Red, yellow and green colors are used to highlight anchors associated to the blocks A, B and D, obtained in the comparisons G/H, G/H and G/K, respectively. **a)** $(BA)_G \in K$ but K supports only weakly $(BA)_G$: genes q and r belong to the same block D, along G, in G/K but the list of expected conditions is not completely fulfilled (see text). The number of anchors of D lying before q' (after r'), and possibly including it, is indicated above q' (r') within a square. **b)** $(BA)_G \notin K$: genes q and r do not belong to the same block in G/K.

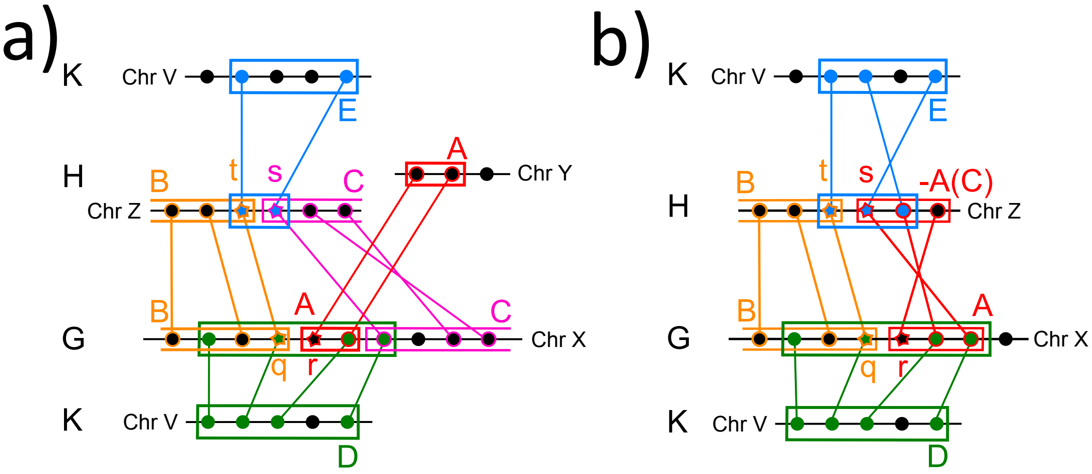


Figure S8: **Illustration of two cases where** K **belongs to both sets** $\boldsymbol{S_{(B\,A)}}$ **and** $\boldsymbol{S_{(B\,C)}}$. For sake of clarity, the 5 genes of chromosome V in genome K, which are involved in the adjacencies $(B\,A)_G$ and $(B\,C)_H$, are represented twice to illustrate both the G/K comparison (green) and the H/K comparison (blue). In the G/K (resp. H/K) comparison, genes q, r (resp. t, s) characterizing $(B\,A)_G$ (resp. $(B\,C)_H$), are included in the same block D (resp. E) along G (resp. H). **a)** Illustration of the ambiguous breakpoint $[(B\,A)_G, (B\,C)_H]$, where C follows A in G and where A is a small block. **b)** Illustration of the ambiguous breakpoint $[(B\,A)_G, (B\ -A)_H]$, where A is a small block.