Phylogenomics: the beginning of incongruence?

Olivier Jeffroy, Henner Brinkmann, Frédéric Delsuc, Herve Philippe

Phylogenomics: the beginning of incongruence?

Olivier Jeffroy, Henner Brinkmann, Frédéric Delsuc and Hervé Philippe

Canadian Institute for Advanced Research. Centre Robert-Cedergren,

Département de Biochimie, Université de Montréal, Succursale Centre-Ville,

Montréal, Québec H3C3J7, Canada

Corresponding author:

Hervé Philippe

Département de Biochimie, Université de Montréal, Succursale Centre-Ville,

Montréal, Québec H3C3J7, Canada

Email: herve.philippe@umontreal.ca

Keywords: systematic error, yeast, inconsistency.

SUMMARY

Until recently, molecular phylogenies based on a single or few orthologous genes often yielded contradictory results. Using multiple genes in a large concatenation was proposed to end these incongruences. Here we show that single gene phylogenies are often incongruent but these observed conflicts mostly lack statistically significant support. In contrast, the use of different tree reconstruction methods on different partitions of the concatenated super-gene leads to well-resolved, but, incongruent phylogenies. Therefore, phylogenomics opens the era of real (i.e. statistically significant) incongruence, instead of ending it. We argue that gathering a large amount of data is not sufficient to obtain a reliable tree because, given the current limitation of tree reconstruction methods, the quality of the input data is also primordial. We propose that selecting only data that contain a minimal amount of non-phylogenetic signal takes full advantage of phylogenomics and seriously reduces incongruence.

Introduction

The use of molecular characters, primarily DNA and derived protein sequences, provides a wealth of new information that sheds light on many parts of the Tree of Life. However, molecular phylogenies based on single genes often lead to apparently conflicting results. To overcome this limitation, it is tempting to apply a genome-scale approach to phylogenetic inference (phylogenomics) by combining a large number of genes. The number of published yeast genomes offers the opportunity to test this proposition. Indeed, using 106 genes from yeast genomes, a fully supported phylogeny has been obtained by the analysis of their concatenation[1,2]. Following from this, it has been anticipated that using large amounts of genomic data will mark the end of incongruence in phylogenetics[3].

The incongruence between two phylogenies can be due to: (1) violations of the orthology assumption generated by mechanisms such as gene duplication, horizontal gene transfer or lineage sorting[4], (2) stochastic error related to the shortness of the genes, and (3) systematic error leading to tree reconstruction artifacts generated by the presence of a non-phylogenetic signal in the data. Adopting a genome-scale approach, theoretically, overcomes incongruences due to the first two reasons: non-orthologous comparisons are gene-specific and will likely be buffered in a multi-gene analysis; and the stochastic error naturally vanishes when more and more genes are considered. In contrast, systematic error is not expected to disappear with the addition of data[5]. Systematic error results from non-phylogenetic signals being present in the data, such as heterogeneity of nucleotide compositions among species (compositional signal),

rate variation across lineages (rate signal), and also within-site rate variation (heterotachous signal)[6]. We assimilate the bias causing systematic error as a signal because, contrary to stochastic noise, it does not average out over a large number of sites. If a bias is strong enough, it can dominate the true phylogenetic signal causing the tree reconstruction method to be inconsistent and lead to an incorrect, but highly supported tree[5,7]. Therefore, phylogenomics, instead of ending incongruence, might open an era of real, statistically significant incongruence resulting from the use of different methods, different taxon samplings, or different character partitions of the same dataset.

To illustrate this paradox, we used the large dataset of 106 genes (120,762 nucleotides) from 14 yeast species assembled by Rokas and Carroll[1]. Phylogenetic trees were inferred by maximum parsimony (MP) from nucleotide sequences as in Ref. [1], and alternatively by probabilistic methods (Bayesian inference (BI)[8] or maximum likelihood (ML)[9-11]), because these methods are generally considered as the most accurate[12,13]. In addition, since divergences among these yeasts are ancient (more than 250 MYa[14]) and amino acid sequences evolve more slowly than nucleotide sequences, the translated protein sequences were also used to construct trees. Phylogenies were inferred from each of the 106 genes and from their concatenation, using two different methods (MP and BI) and two types of characters (nucleotides and amino acids), yielding a total of 428 trees. We estimated the level of incongruence as the number of bipartitions (or splits, i.e. groups of species defined by a branch of a phylogenetic tree), supported by more than a given bootstrap value, that are different between

two trees. Our aim was to compare the level of among-gene incongruence for a given tree reconstruction method with the level of among-method incongruence for a given dataset.

Congruence among phylogenetic markers

The trees inferred from each of the 106 genes are all different (data not shown), yielding an apparent high level of incongruence. However, there are $3x10^{11}$ possible binary trees connecting 14 taxa and it is possible that the different genes recovered different, but very similar trees. Without taking statistical support into account, there are 25.9% and 24.6% different bipartitions when trees are inferred by either MP at the nucleotide level (MP_{nt}) or by BI at the amino acid level (BI_{aa}), respectively.

However, if one restricts the measure of incongruence only to those bipartitions that are supported above a predefined significance level, a statistically significant incongruence (bootstrap support (BS) > 95%) among the 106 individual genes is almost nonexistent. For MP_{nt} and BI_{aa}, only 0.4% or 0.6% of the significantly supported bipartitions are different, respectively. Yet, the non-parametric bootstrap test is often considered as conservative and the use of a p-value of 70% has been suggested[15]. Even with this reduced threshold, only 4.0% and 2.8% of the bipartitions are different, strongly arguing for the absence of statistically significant incongruence among the 106 genes when analyzed with the same method. These results are in line with a similar analysis based on the same 106 genes, although with only eight species[16]. In addition, single gene

phylogenies are not significantly incongruent with concatenation-based trees. Only 1.8% (5.0%) and 3.1% (6.8%) of the bipartitions are different at 95% (70%) bootstrap confidence between the 106 gene trees and the concatenated tree for BI_{aa} and MP_{aa}, respectively.

In summary, all single gene trees are different because of the predominant effect of stochastic error (except one paralogous comparison, Fig. S1), but there is no statistically significant incongruence among these 106 genes when the same tree reconstruction method is used (either MP_{nt} or BI_{aa}). In other words, there is no among-gene incongruence in this data set.

Strong incongruence when different tree reconstruction methods are used

In contrast, a non-negligible statistically significant incongruence exists because of the use of different tree reconstruction methods, and because of the use of nucleotide versus amino acid sequences. On average, 14.2% (23.2%) of the bipartitions are different at the 95% (70%) bootstrap confidence level between the MP_{nt} tree and the BI_{aa} tree, albeit inferred from the very same genes.

Does the phylogenomic approach avoid this incongruence? The answer is no: when phylogenies are inferred from the concatenation of the 106 genes, 36.4% of bipartitions are different between the MP_{nt} and BI_{aa} trees. In fact, four out of 11 nodes, which are all highly supported, are different, indicating that incongruence has in fact increased. Therefore, a large-scale genome approach only ends the statistically insignificant among-gene incongruence but opens the

era of the real statistically significant incongruence among methods and character sets.

Nucleotide composition bias causes most of the incongruence

To better understand the source of this exceptionally high level of incongruence, trees inferred from the concatenation by MP_{nt}, BI_{nt}, MP_{aa} and BI_{aa} were compared (Fig. 1a-d). This allows separating the impact of the type of characters considered from the impact of the reconstruction method used. The topology within the clade containing the five *Saccharomyces* species, *Naumovia castellii* and *Candida glabrata* was identical in all four cases. In addition, *Debaromyces hansenii* invariantly appeared as the sister-group of *Candida albicans*. Incongruences are thus predominant for the most basal nodes. The MP_{nt} tree is the most different from the other three trees (four different bipartitions). The BI_{nt} tree differs from the BI_{aa} tree by three bipartitions and the MP_{aa} tree from the BI_{aa} tree by only two bipartitions. This suggests that, in this case, the method (MP or BI) is less important than the type of characters used (nucleotides or amino acids).

Compositional bias is known to be more prominent in nucleotides than in amino acids[17], because the fast evolving third codon positions accumulate mutational bias due to the degeneracy of the genetic code. The average G+C content at the third codon positions of the 106 genes for a given species is indeed highly heterogeneous, ranging from 27% in *C. albicans* to 68% in *Yarrowia lipolytica* (Fig.S2). Differences in nucleotide or amino acid composition

can render tree reconstruction methods inconsistent if not properly accounted for[7,18]. Strikingly, groupings in the MP_{nt} tree (Fig. 1a) appear to be strongly correlated with G+C content: the GC-rich *Y. lipolytica* (68%) is grouped with *Ashbya gossypii* (66%), then with *Kluyveromyces waltii (*51%) and finally with *Saccharomyces kluyveri* (45%), and on the other hand the relatively GC poor *Kluyveromyces lactis* (39%) is grouped with *D. hansenii* (34%) and *C. albicans* (27%). Therefore, the non-phylogenetic compositional signal is likely dominating over genuine phylogenetic signal in the MP_{nt} tree.

By contrast, in the BI_{aa} tree (Fig. 1d), the species do not appear to be grouped according to their G+C content: *K. lactis* (39%) with *A. gossypii* (66%), *S. kluyveri* (45%) with *K. waltii* (51%), and *Y. lipolytica* (68%) together with *C. albicans* (27%) and *D. hansenii* (34%). Similarly, in the MP_{aa} tree (Fig. 1c), the groups appear no longer determined by nucleotide composition, even if they are slightly different from the ones observed in the BI_{aa} tree (paraphyly of *K. lactis* (39%) + *A. gossypii* (66%)). This argues that this BI_{aa} tree (Fig. 1d) is not, or less, biased by the compositional signal, and is likely to be closer to the correct phylogeny than the likely erroneous MP_{nt} tree. As we will show in the following, this conclusion is supported by the fact that an approach known to increase the impact of G+C bias converges toward the MP_{nt} tree (case 1), and two approaches known to decrease its impact converge toward the BI_{aa} tree (cases 2 and 3).

(1) When only the most biased third codon positions are analyzed by BI, the inferred phylogeny is identical to the MP_{nt} tree (Fig. S3), indicating that for

these fast evolving characters BI is not able to correctly extract the phylogenetic signal that is overwhelmed by the compositional signal.

(2) When the least biased first two codon positions are analyzed by MP, the inferred phylogeny (Fig. S4) is identical to the BI_{nt} tree, indicating that the removal of the third codon positions renders MP less sensitive to the compositional bias.

(3) The use of the slowly evolving transversions is known to avoid artifacts due to the compositional signal[19], because the purine (and pyrimidine) content is generally homogeneous even when the G+C content is not (Fig. S2). As expected, when the transversions at the first two codon positions are analyzed by MP, the same tree as with amino acid sequences is recovered (Fig. S5). Finally, exactly the same result as with BI on amino acids is obtained with BI on transversions from all three codon positions (Fig. S6).

In conclusion, all analyses indicate that the strong statistical incongruence between MP_{nt} and BI_{aa} trees is due to a higher sensitivity of MP to a systematic error related to the compositional bias at the nucleotide level whose effects are attenuated upon translation. In addition, the difference in the type of characters used (nucleotides versus amino acids) explains a greater part of the huge differences observed between the MP_{nt} (Fig. 1a) and the BI_{aa} trees (Fig. 1d) than the use of two different tree reconstruction methods (MP versus BI).

Saturation as an indicator of incongruence

Tree reconstruction artifacts are due to the accumulation of multiple substitutions at the same position over time: convergences and reversions erase the genuine phylogenetic signal. When multiple substitutions are dominating, the dataset is said to be mutationally saturated. Without any bias, a highly saturated dataset will produce an unresolved star-like phylogeny. However, when sequences have been generated by a heterogeneous evolutionary process, saturation will ultimately lead to the accumulation of an erroneous non-phylogenetic signal in the alignments.

We evaluated the saturation level of the yeast phylogenomic dataset by comparing the number of substitutions inferred by ML with the number of observed differences for each pair of species[20], for the complete alignment (Fig. 2). The lower the slope of the linear regression, the higher the level of saturation is, and therefore the higher the probability of tree reconstruction artifacts (hence of incongruence) is. As expected, nucleotides (slope = 0.31, Fig. 2d) are more saturated than amino acids (slope = 0.51, Fig. 2a). However, the saturation of nucleotides is highly concentrated in third codon positions (slope = 0.16, Fig. 2b), and much less pronounced for the first two codon positions (slope = 0.47, Fig. 2c).

Multiple substitutions are so frequent at the third codon positions (up to 10 inferred substitutions for only one observed difference) that the BI method, albeit efficient in detecting multiple substitutions[12,13], is seriously misled by the compositional signal (Fig. S2). The high level of saturation suggests that third codon positions should not be used for inferring ancient phylogenies. However,

the sparse taxon sampling here considered (only 14 species) likely aggravates the case, and this conclusion might have to be reevaluated with a denser sampling of species.

Interestingly, Fig. 2f shows a negative correlation between the level of saturation of a given dataset and the number of differences to the least biased tree (BI_{aa} or BI_{tv123}, Fig. 1d). In other words, when fast evolving positions are removed from the analysis, the inferred phylogeny is less biased by non-phylogenetic (in particular compositional) signal, even if the tree reconstruction method is not very accurate. As a result, a more reliable phylogeny is obtained with a poorly performing method (MP) and a relatively unsaturated dataset (amino acids) than with a more accurate method (BI or ML with a complex model) and a highly saturated dataset (third codon positions only). The quality of the dataset is therefore as important as, if not more, the accuracy of the tree reconstruction method.

A greater sensitivity of MP to mutational saturation as compared to BI could explain why the MP_{aa} and BI_{aa} trees are slightly different. In fact, *K. lactis* and *A. gossypii* evolve faster than *K. waltii* and *S. kluyveri* (Fig. 1d) and are likely to be attracted by the long unbroken branch of the outgroup in the MP tree (Fig. 1c) because of a long branch attraction (LBA) artifact[5]. To obtain a less saturated dataset than the complete amino acid alignment, we removed the 18,075 positions that display at least two different amino acids in the outgroup species (*D. hansenii*, *C. albicans* and *Y. lipolytica*). This approach has the additional advantage to efficiently shorten the branch length of the outgroup, thus reducing

the impact of the LBA. The MP tree inferred from the remaining 22,179 slowly evolving positions (Fig. 3) is identical to the BI_{aa} tree, albeit with a reduced bootstrap support. When the saturation is much reduced (slope = 0.58 for these 22,179 amino acid positions, Fig. 2e), MP recovers exactly the same tree as BI, strongly arguing that this tree is the best current working hypothesis for the phylogeny of these 14 yeast species in the light of which yeast genomic evolution should be interpreted.

Conclusion and recommendation

We of course do not argue against the use of a large number of genes for phylogenetic inference[1,2,21], as it is generally required to solve difficult phylogenetic questions[22,23]. However, contrary to some current opinions[1,2], obtaining a highly supported tree from the analysis of a concatenation of multiple genes does not guarantee that "it accurately represents the historical relationships"[2]. Highly supported groupings can prove to be incorrect because of the inconsistency of the tree reconstruction method (Fig. 1a). Since these errors are due to systematic biases that generally become apparent when using large datasets, phylogenomic trees should be regarded with greater caution than single gene trees for possible tree reconstruction artifacts[6,7,23-27].

We stress that phylogenomics should not only emphasize the quantity of data under study but also their quality (i.e. their degree of saturation). Since current tree reconstruction methods are not always able to correctly handle the presence of multiple substitutions, efforts should be made to reduce their

potentially misleading effects. First, because using a large number of taxa allows a better detection of multiple substitutions, increasing the taxon sampling is particularly important. All recent empirical phylogenomic studies[23,25,28-31] but one[1] supports this conclusion. This latter study[1] should nevertheless be treated with caution since the tree used as reference (the MP_{nt} tree of Fig. 1a) is almost certainly incorrect. Second, probabilistic methods should be used with models of sequence evolution that handle the most flagrant aspects of real substitution patterns in order to reduce the inconsistency of current methods due to model misspecification[32]. Non-stationary models for dealing with heterogeneous G+C content and mixture models certainly represent steps in the right direction.

Finally, we believe that an efficient way to take advantage of the wealth of genomic data currently produced is to voluntary discard a part of the data from phylogenetic analyses. This is already a common practice, as demonstrated by the removal of ambiguously aligned regions or of odd species (e.g. the fast-evolving microsporidia are never used to represent fungi) or by the use of amino acid instead of nucleotides for ancient divergences. Extensive data removal is often unpractical in single gene analyses because too few positions remain available, producing a poorly resolved tree[33]. This limitation becomes negligible in phylogenomics, and highly supported trees cleared up from tree reconstruction artifacts can be recovered when more than half of the data have been discarded[23,28,30]. We therefore suggest putting the emphasis on the development and refinement of objective methods aimed at detecting and removing the part of the data containing a high level of non-phylogenetic signal[6]. As we showed in the

case of yeasts, the application of these guidelines will hopefully avoid that incongruence dominates the phylogenomic era.

Acknowledgements

We wish to thank Antonis Rokas for providing his alignment and Franz Lang, Nicolas Lartillot, Nicolas Rodrigue and Naiara Rodriguez-Ezpeleta for critical readings of the manuscript. This work was supported by operating funds from Génome Québec. H.P. is a member of the Program in Evolutionary Biology of the Canadian Institute for Advanced Research (CIAR), which is acknowledged for salary and interaction support. H.P. is also grateful to the Canada Research Chairs Program and the Canadian Foundation for Innovation (CFI) for salary and equipment support.

References

1 Rokas, A. and Carroll, S.B. (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol* 22 (5), 1337-1344

2 Rokas, A. *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425 (6960), 798-804.

3 Gee, H. (2003) Evolution: ending incongruence. *Nature* 425 (6960), 782

4 Maddison, W.P. (1997) Gene trees in species. *Systematic Biology* 46 (3), 523-536

5 Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401-410

6 Philippe, H. *et al.* (2005) Phylogenomics. *Annu Rev Ecol Evol Syst* 36, 541-562

7 Phillips, M.J. *et al.* (2004) Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution* 21, 1455-1458

8 Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic

inference under mixed models. *Bioinformatics* 19 (12), 1572-1574.

9 Swofford, D.L. (2000) PAUP*: Phylogenetic Analysis Using Parsimony and other methods. (4b10 edn), Sinauer, Sunderland, MA

10 Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52 (5), 696-704.

11 Jobb, G. *et al.* (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol. Biol.* 4 (1), 18

12 Whelan, S. *et al.* (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet* 17 (5), 262-272.

13 Felsenstein, J. (2004) *Inferring phylogenies*, Sinauer Associates, Inc.

14 Douzery, E.J. *et al.* (2004) The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci U S A*

15 Hillis, D.M. and Bull, J.J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst biol* 42 (2), 182-192

16 Taylor, D.J. and Piel, W.H. (2004) An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. *Mol Biol Evol* 21 (8), 1534-1537

17 Hasegawa, M. and Hashimoto, T. (1993) Ribosomal RNA trees misleading? *Nature* 361 (6407), 23

18 Lockhart, P. *et al.* (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution* 11 (4), 605-612

19 Woese, C.R. *et al.* (1991) Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts. *Syst Appl Microbiol* 14 (4), 364-371.

20 Philippe, H. *et al.* (1994) Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. *Journal of Evolutionary Biology* 7, 247-265

21 Rosenberg, M.S. and Kumar, S. (2003) Taxon sampling, bioinformatics, and phylogenomics. *Syst Biol* 52 (1), 119-124

22 Philippe, H. *et al.* (1994) Can the cambrian explosion be inferred through molecular phylogeny ? *Development* 120, S15-S25

23 Delsuc, F. *et al.* (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6 (5), 361-375

24 Soltis, D.E. *et al.* (2004) Genome-scale data, angiosperm relationships, and

"ending incongruence": a cautionary tale in phylogenetics. *Trends Plant Sci* 9 (10), 477-483

25 Stefanovic, S. *et al.* (2004) Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? *BMC Evol Biol* 4 (1), 35

26 Lockhart, P.J. and Penny, D. (2005) The place of Amborella within the radiation of angiosperms. *Trends Plant Sci* 10 (5), 201-202

27 Holland, B.R. *et al.* (2005) Improved Consensus Network Techniques for Genome-Scale Phylogeny. *Mol Biol Evol*

28 Brinkmann, H. *et al.* (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol* 54 (5), 743-757

29 Philippe, H. (1997) Rodent monophyly: pitfalls of molecular phylogenies. *Journal of Molecular Evolution* 45 (6), 712-715

30 Philippe, H. *et al.* (2005) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol* 22 (5), 1246-1253

31 Leebens-Mack, J. *et al.* (2005) Identifying the Basal Angiosperm Node in Chloroplast Genome Phylogenies: Sampling One's Way Out of the Felsenstein Zone. *Mol Biol Evol*

32 Steel, M. (2005) Should phylogenetic models be trying to 'fit an elephant'? *Trends Genet* 21 (6), 307-309

33 Philippe, H. *et al.* (2000) Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proceedings of the Royal Society of London* 267 (1449), 1213-1221

34 Douady, C.J. *et al.* (2003) Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol Biol Evol* 20 (2), 248-254.

35 Philippe, H. (1993) MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res.* 21 (22), 5264-5272

Figure Legends

Figure 1: Phylogenies based on the concatenation of 106 genes.

Trees were inferred using MP (1a,c) and BI (1b,d) with both nucleotide (1a,b; 120,762 positions) and deduced amino acid (1c,d; 40,254 positions) sequences. They are highly supported and only bootstrap values below 100% are indicated to the left of the corresponding node; 1000 replicates for MP using PAUP*[9] (10 random species addition with TBR branch swapping) and 100 replicates for BI using MrBayes[8] (GTR+Γ model for nucleotides and WAG+Γ model for amino acids) following Douady *et al.*[34] were performed. ML trees and bootstrap supports inferred using PAUP*, PHYML[10] and Treefinder[11] are virtually identical. The G+C content at the third codon position is indicated in brackets and the color of the species name varies from purple to red with increasing G+C content. *Y. lipolytica* is used as outgroup in all tree representations. Parts of phylogenies that are not identical among the four trees are shown in bold. Scale bar indicates the number of substitutions (MP) or the number of substitutions per position (BI). When a different approach gives the same topology as the one indicated in bold, its name is indicated using normal font (e.g. BI_{nt3}). The complete species names are *Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces kudriazevii*, *Saccharomyces bayanus*, *Naumovia castellii*, *Candida glabrata*, *Debaromyces hansenii*, *Candida albicans*, *Kluyveromyces lactis*, *Saccharomyces kluyveri*, *Kluyveromyces waltii*, *Ashbya gossypii* and *Yarrowia lipolytica*. The phylogeny that is most likely to be the correct tree is outlined in red.

Figure 2: Mutational saturation of the concatenation of 106 genes.

The level of saturation was estimated using the method described in Philippe *et al.*[20], as implemented in MUST[35]. The X-axis corresponds to the number of substitutions inferred from the ML tree while the Y-axis corresponds to the number of differences observed in a pairwise comparaison, for the same pair of species. A linear regression is performed and the slope of the dotted line starting from the origin is used as an indicator of the saturation level (e. g. multiple substitutions at the same position), the steeper the slope the less saturated the dataset is. Analyses were performed for amino acids (a; 40,254 positions), third codon positions (b; 40,254 positions), first two codon positions (c; 80,508 positions), all three codon positions (d; 120,762 positions), and a dataset where variable positions in the outgroup species (*D. hansenii*, *C. albicans* and *Y. lipolytica*) have been removed (e; 22,179 positions). The diagonal, which corresponds to the case where no multiple substitutions occurred, is indicated by a bold line. Finally, Fig. 2f shows the relation of the number of bipartitions different to the least biased tree (BI_{aa} or BI_{tv123} see Fig. 1d) as a function of the saturation level expressed as the slope of the regression line. The three data sets used are nt_3, nt_{123}, and nt_{12}.

Figure 3: Removal of fast-evolving positions.

The 18,075 amino acid positions that are variable in outgroup species (*D. hansenii*, *C. albicans* and *Y. lipolytica*) were eliminated and a MP analysis as in Fig. 1c was performed on the remaining 22,179 slowly evolving positions. This

removal reduces the impact of the long-branch attraction artifact and the resulting MP tree is identical to the one obtained using BI based on the complete amino acid concatenation (Fig. 1d). Color code and species names are the same as in Figure 1.

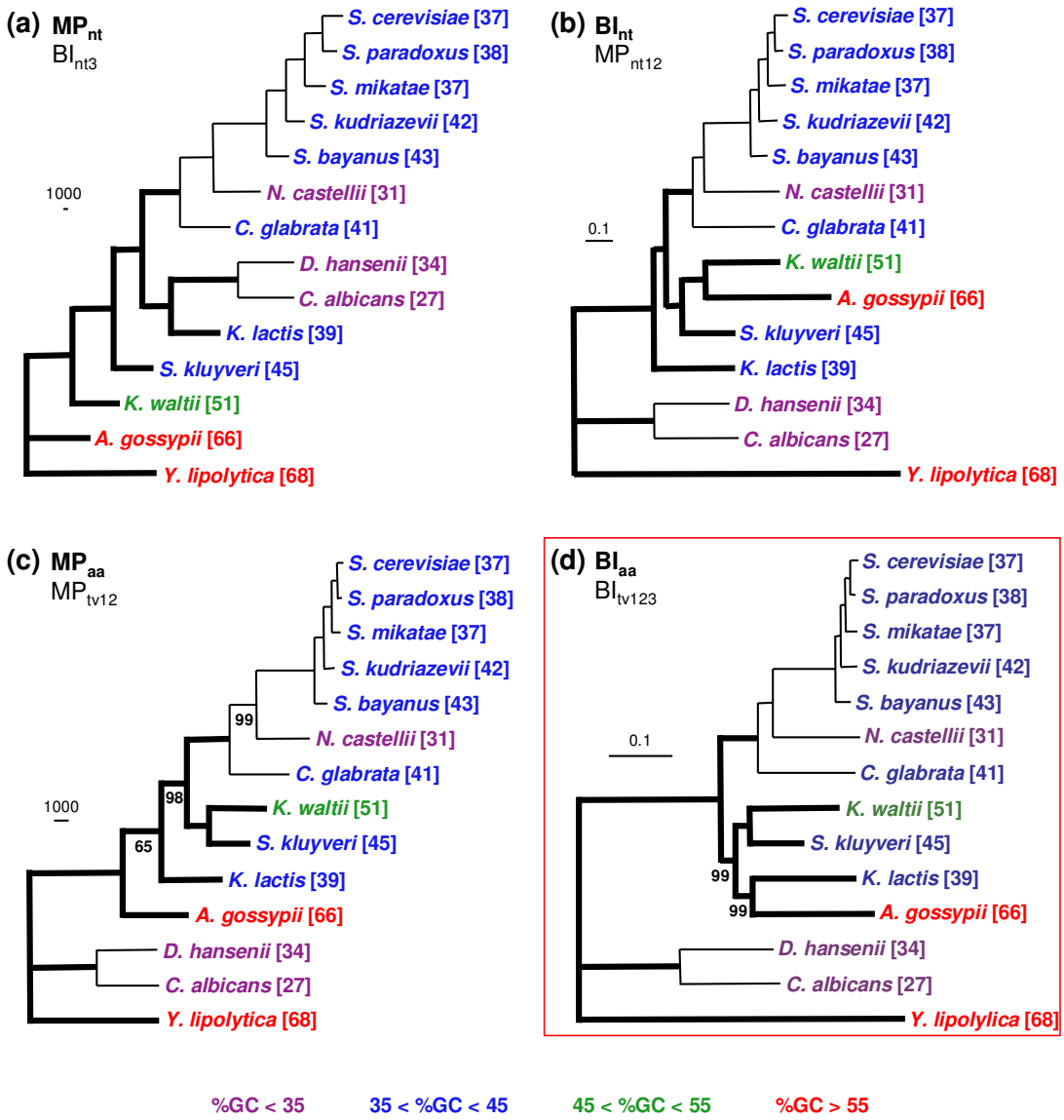
(a) MP nt
BI nt3
1000
S. cerevisiae [37]
S. paradoxus [38]
S. mikatae [37]
S. kudriazevii [42]
S. bayanus [43]
N. castellii [31]
C. glabrata [41]
D. hansenii [34]
C. albicans [27]
K. lactis [39]
S. kluyveri [45]
K. waltii [51]
A. gossypii [66]
Y. lipolytica [68]
(b) BI nt
MP nt12
0.1
S. cerevisiae [37]
S. paradoxus [38]
S. mikatae [37]
S. kudriazevii [42]
S. bayanus [43]
N. castellii [31]
C. glabrata [41]
K. waltii [51]
A. gossypii [66]
S. kluyveri [45]
K. lactis [39]
D. hansenii [34]
C. albicans [27]
Y. lipolytica [68]
(c) MP aa
MP tv12
1000
99
98
65
S. cerevisiae [37]
S. paradoxus [38]
S. mikatae [37]
S. kudriazevii [42]
S. bayanus [43]
N. castellii [31]
C. glabrata [41]
K. waltii [51]
S. kluyveri [45]
K. lactis [39]
A. gossypii [66]
D. hansenii [34]
C. albicans [27]
Y. lipolytica [68]
(d) BI aa
BI tv123
0.1
99
99
S. cerevisiae [37]
S. paradoxus [38]
S. mikatae [37]
S. kudriazevii [42]
S. bayanus [43]
N. castellii [31]
C. glabrata [41]
K. waltii [51]
S. kluyveri [45]
K. lactis [39]
A. gossypii [66]
D. hansenii [34]
C. albicans [27]
Y. lipolylica [68]
%GC < 35
35 < %GC < 45
45 < %GC < 55
%GC > 55

Figure 1

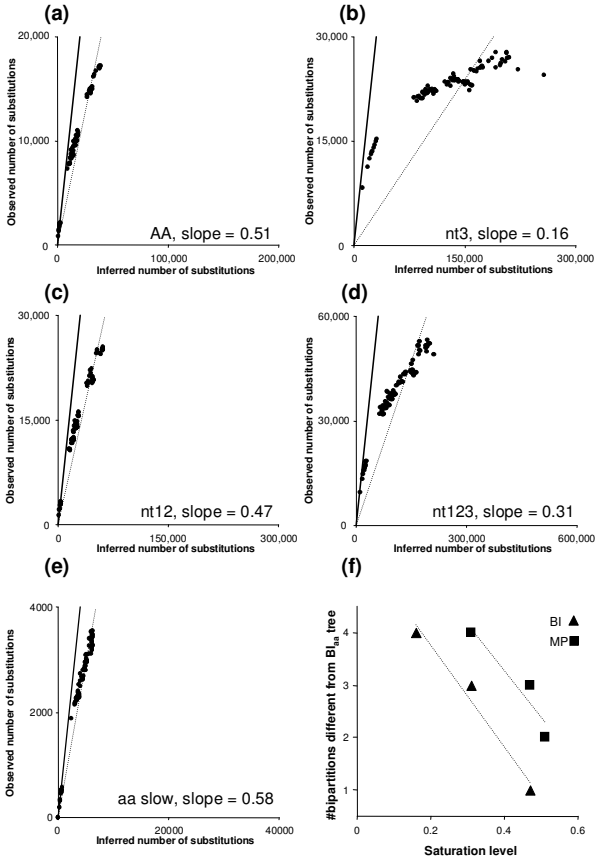
(a)
Observed number of substitutions
Inferred number of substitutions
AA, slope = 0.51
(b)
nt3, slope = 0.16
(c)
nt12, slope = 0.47
(d)
nt123, slope = 0.31
(e)
aa slow, slope = 0.58
(f)
#bipartitions different from BI$_{aa}$ tree
Saturation level
BI
MP

Figure 2

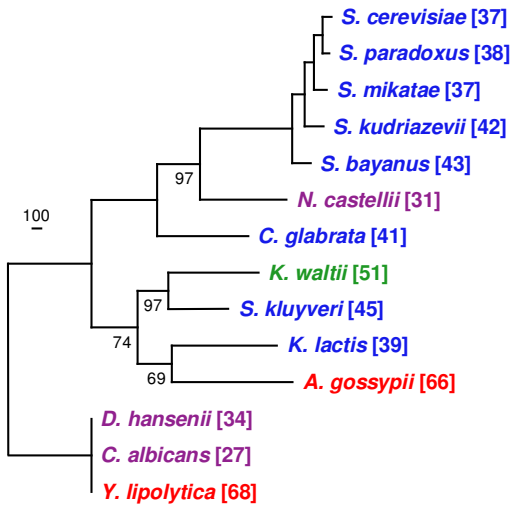
S. cerevisiae [37]
S. paradoxus [38]
S. mikatae [37]
S. kudriazevii [42]
S. bayanus [43]
N. castellii [31]
C. glabrata [41]
K. waltii [51]
S. kluyveri [45]
K. lactis [39]
A. gossypii [66]
D. hansenii [34]
C. albicans [27]
Y. lipolytica [68]
97
100
97
74
69

Figure 3

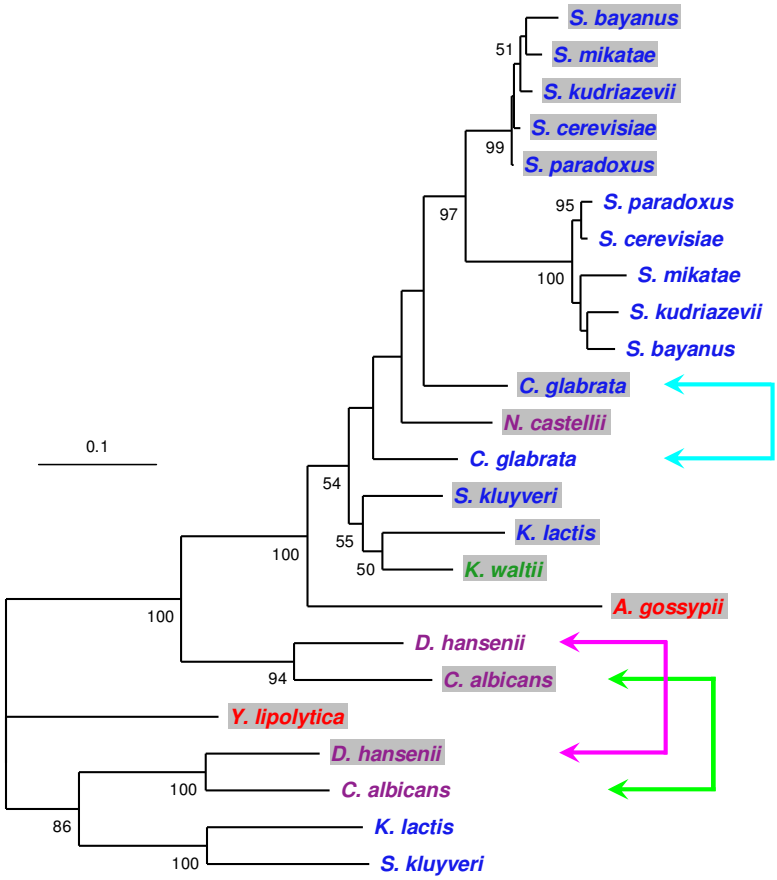

Figure S1: Paralogous comparison. For the 106 genes, we performed tblastx on the 14 yeast and retrieve all the sequences with an e-value below 1e-10. Amino acid alignments were performed with muscle, unambiguously aligned regions were selected using Gblocks with default options, and trees were inferred using TreeFinder with a WAG+F+Γ model. Only one gene, YNL104C, showed evidence for paralogous comparison and is displayed here. For *D. hansenii*, an erroneous paralogous copy was incorporated by Rokas and Carroll (2005), breaking the monophyly of the clade *D. hansenii* / *C. albicans*. For *C. glabrata*, there are also two paralogous copies, but it is more difficult to know which copy should be considered as the ortholog. In any case, this non-orthologous comparison artificially increases among gene incongruence. For instance, when YNL104C is discarded from the analysis, the frequency of different bipartitions drops from 2.8% to 2.6% (trees inferred with BI_{aa} and a bootstrap threshold of 70%). Interestingly, the inclusion of this non-orthologous gene in the concatenation does not prevent to recover the monophyly of the clade *D. hansenii* / *C. albicans*, indicating the robustness of phylogenomics versus this type of error.

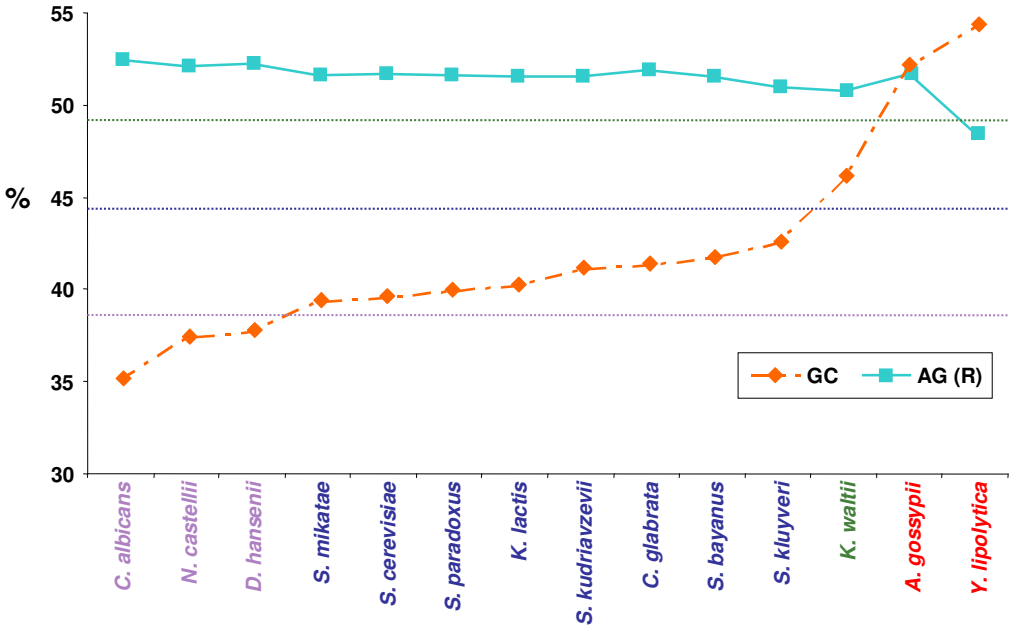

Figure S2: Compositional bias analysis. This figure illustrates GC and purine (R) contents of the 14 species for the 106 genes under study. The GC content is highly heterogeneous ranging from 35 to 54%. According to the GC content, four categories of species can be arbitrarily distinguished as shown by the different colors on the figure. In contrast, the purine / pyrimidine contents are almost homogenous among species with the exception of *Y. lipolytica*.

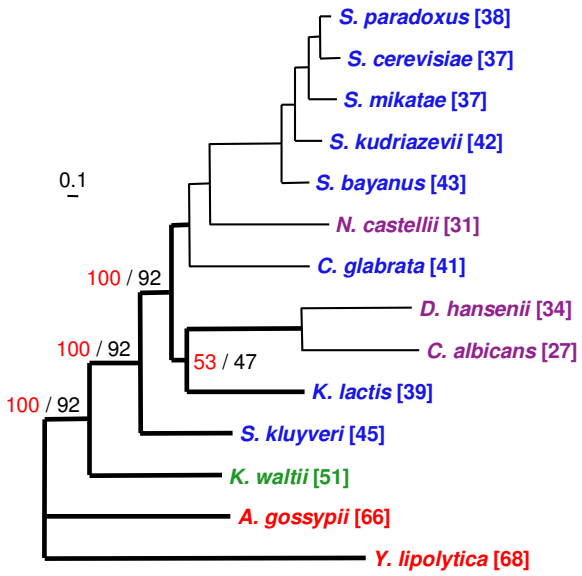

Figure S3: Bayesian analysis of concatenated third codon positions (40,254 sites) under the GTR model and a gamma distribution with four rate categories using MrBayes (Ronquist & Huelsenbeck, 2003). Maximum likelihood tree obtained from the analysis of the same dataset using Treefinder (Jobb et al., 2004) under the same model leads to the same topology. All nodes received 100% bootstrap support except where indicated at the corresponding nodes. These values were computed using 100 replicates drawn by SeqBoot (Felsenstein, 2001) and subsequently analyzed with Treefinder (red) and MrBayes (black) following Douady *et al.* (2003). The probabilistic analysis of third codon positions, which are the most biased, therefore recovers the same topology as maximum parsimony on the complete nucleotide dataset (see Figure 1a).

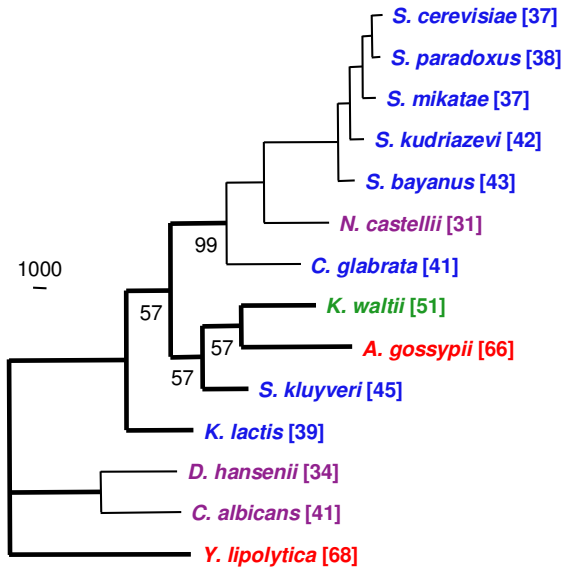

Figure S4: Most parsimonious tree obtained using PAUP* (Swofford, 2000) with 10 random sequence addition replicates and TBR branch swapping on the first two codon positions (80,508 sites). All nodes received 100% bootstrap support based on 1,000 replicates except where indicated for the corresponding nodes. By removing third codon positions, the MP topology is the same as the one obtained by using probabilistic methods on the complete nucleotide dataset (see Figure 1b).

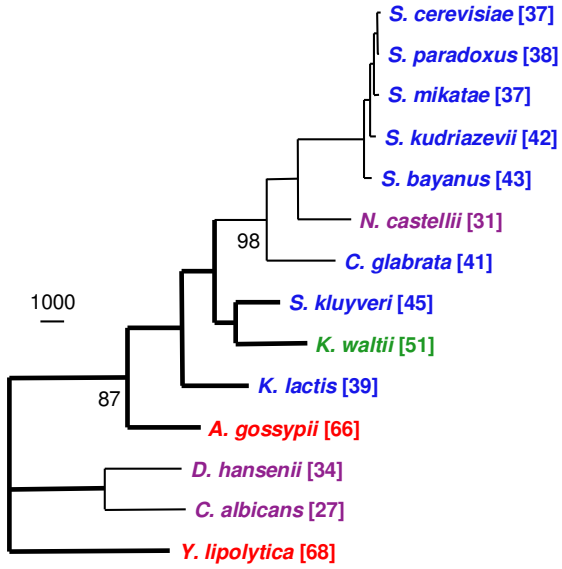

Figure S5: Most parsimonious tree obtained using PAUP* (Swofford, 2000) with 10 random sequence addition replicates and TBR branch swapping on the first two codon positions recoded using RY-coding (AG=>R and CT=>Y) (80,508 sites). All nodes received 100% bootstrap support based on 1,000 replicates except where indicated for the corresponding nodes. Eliminating the transition bias therefore makes the nucleotide tree topology the same as the one obtained by using maximum parsimony on the amino acid dataset (see Figure 1c).

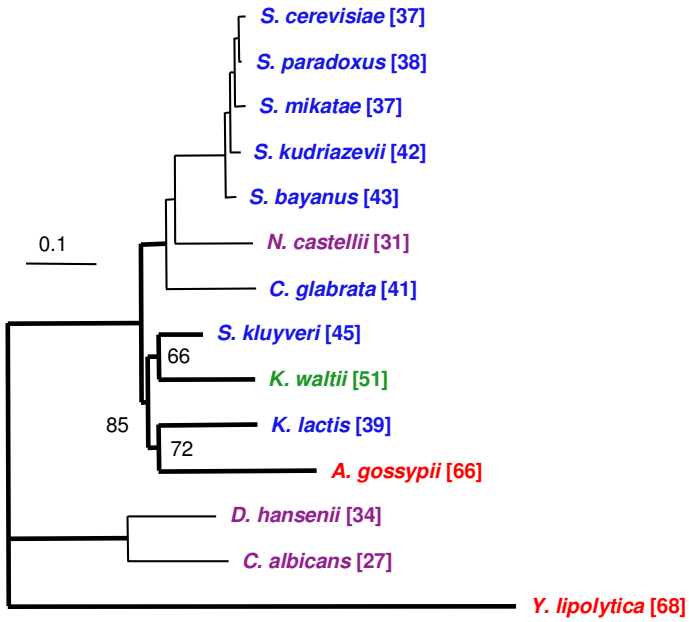

Figure S6: Bayesian analysis of the complete nucleotide dataset (120,762 sites) recoded as RY (AG and CT are each merged into a single state) under a two-state model and a gamma distribution with four categories using MrBayes (Ronquist and Huelsenbeck, 2003). Maximum likelihood tree inferred under the same conditions (GTR2 model) using Treefinder (Jobb et al., 2004) is the same. All nodes received 100% bootstrap support based on 100 replicates except where indicated for the corresponding nodes. As with maximum parsimony, eliminating the transition bias makes the nucleotide based tree topology the same as the one obtained by using maximum likelihood on the amino acid dataset (see Figure 1d).

References (Supplementary Material)

1. Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*5, 113
2. Castresana, J. (2000) Selection of conserved blocks from multiplealignments for their use in phylogenetic analysis. *Mol.Biol.Evol.*17, 540-552
3. Jobb, G. *et al.*(2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol. Biol.*4, 18
4. Rokas, A. and Carroll, S.B. (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy.*Mol.Biol.Evol.*22, 1337-1344
5. Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesianphylogenetic inference under mixed models. *Bioinformatics*19, 1572-1574
6. Felsenstein, J. (2001) PHYLIP (Phylogene Inference Package). (3.6 edn), Distributed by the author, Department of Genetics, University of Washington
7. Douady, C.J. *et al.*(2003) Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol.Biol.Evol.*20, 248-254
8. Swofford, D.L. (2000) PAUP*: Phylogenetic analysis using parsimony and other methods. (4b10 edn), Sinauer, Sunderland, MA