Research

Author for correspondence:
Bastien Boussau
e-mail: bastien.boussau@univ-lyon1.fr

THE ROYAL SOCIETY
PUBLISHING

Genome-scale phylogenetic analysis finds extensive gene transfer among fungi

Gergely J. Szöllősi[1], Adrián Arellano Davín[2], Eric Tannier[2,3], Vincent Daubin[2,4] and Bastien Boussau[2,4]

[1]ELTE-MTA 'Lendület' Biophysics Research Group, Pázmány P. stny. 1A, 1117 Budapest, Hungary
[2]Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, 69000 Lyon, France
[3]Institut National de Recherche en Informatique et en Automatique Rhône-Alpes, 38334 Montbonnot, France
[4]Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Université Lyon 1, 69622 Villeurbanne, France

Although the role of lateral gene transfer is well recognized in the evolution of bacteria, it is generally assumed that it has had less influence among eukaryotes. To explore this hypothesis, we compare the dynamics of genome evolution in two groups of organisms: cyanobacteria and fungi. Ancestral genomes are inferred in both clades using two types of methods: first, *Count*, a gene tree una-ware method that models gene duplications, gains and losses to explain the observed numbers of genes present in a genome; second, ALE, a more recent gene tree-aware method that reconciles gene trees with a species tree using a model of gene duplication, loss and transfer. We compare their merits and their ability to quantify the role of transfers, and assess the impact of taxonomic sampling on their inferences. We present what we believe is compelling evidence that gene transfer plays a significant role in the evolution of fungi.

1. Introduction

Reconstructing genome evolution and ancestral genomes is instrumental to understanding the diversification of life on Earth. Doing so requires harnessing the information available in complete genome sequences, which is best achieved in a statistical framework. Integrative methods to reconstruct the evolution of genomes and thus ancestral genomes are now able to model particular histories of genes inside a general history of genomes and can integrate many different types of events. They integrate sequence-level events such as substitutions, gene-level events such as duplications (D), losses (L) and exchanges of genes between genomes, modelled by lateral gene transfers (hereafter transfers, T), as well as genome-level events such as speciations (S). This inclusiveness enables them to handle diverse groups of organisms, each with their idiosyncratic way of evolving. It therefore becomes possible to apply a single method to groups from different domains of life, and compare their modes of evolution.

Reconstructing ancestral genomes requires *a minima* two types of data: extant genomes with homology relationships between genome fragments, and a tree along which these genomes are supposed to have evolved. A species tree modelling vertical descent is indispensable, because without it, we cannot differentiate vertical inheritance from lateral transfer, and little can be learned about the processes of genome evolution.

Using a common species tree does not mean that we assume that all homologous fragments have had exactly the same history. Instead, the history of each individual homologous fragment is reconstructed with its own succession of duplications, losses and transfers. For species that have diverged a long time ago, only the protein coding portion of the genomes is analysed, and individual histories are reconstructed for each gene family. These gene histories are subsequently analysed together to gain insights into genome evolution, and infer large-scale patterns of gene duplications, losses or transfers. Both steps, first gene tree reconstruction and second aggregation of gene histories into coherent
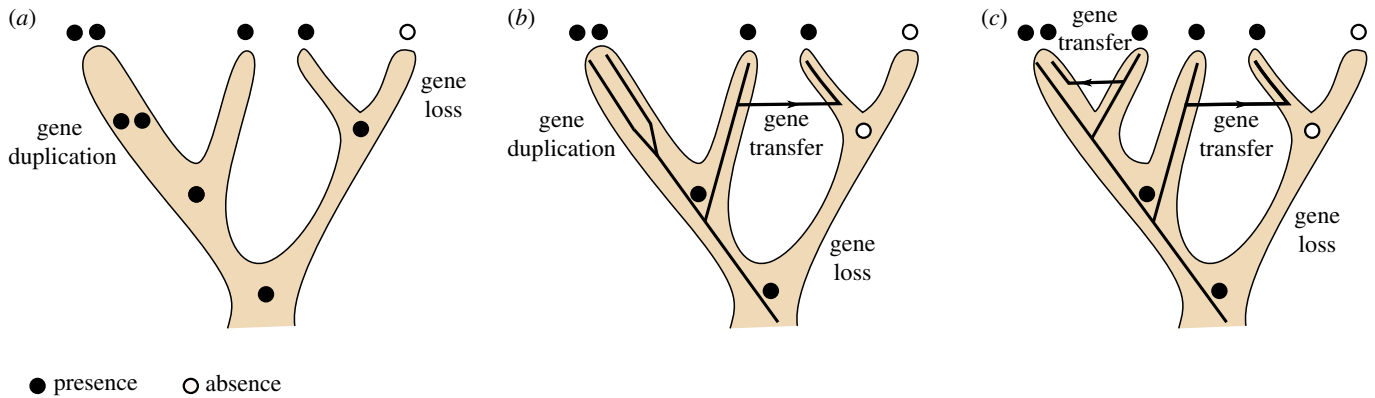
Figure 1. How a gene history can be incorrectly reconstructed if the gene tree is not taken into account, or if taxonomic sampling is incomplete. (*a*) Inference according to the gene content approach. (*b*) Inference according to a gene tree-aware approach. (*c*) Inference according to a gene tree-aware approach, with a more complete taxonomic sampling. Ignoring gene phylogeny and having insufficient species sampling lead to underestimation of gene transfer. In all cases, the same true gene history is assumed, but only with sufficient taxonomic sampling and with a gene tree-aware approach can it be recovered.

patterns, necessitate thoughtful methodologies to overcome possible sources of errors and uncertainties.

(a) Reconstruction of gene histories

Gene sequences are often too short to contain sufficient information for accurate and robust reconstruction of the history of a gene family; worse, even when this information is present, models of sequence evolution may fail to capture it correctly. In general, a gene family's history cannot be reliably inferred, nor interpreted in terms of gene-level events from the set of sequences alone [1,2]. Using additional information coming from the species tree is a way to improve gene tree quality (figure 1). This is the approach taken by 'gene tree-aware approaches'. Alternatively, it is possible to entirely do away with the sequences and avoid gene tree reconstruction: the gene tree unaware 'gene content approach' considers only gene presence/absence patterns, or numbers of genes per species.

(i) Gene content approaches

Gene content approaches work with data in the form of either presence/absence of a gene family inside a given genome, or numbers of genes of a gene family inside a given genome (figure 1). In both cases, parsimony approaches or probabilistic models have been used to reconstruct the evolution of gene families along a species phylogeny.

Among parsimony methods, one can choose between Wagner and Dollo parsimony. Choosing Dollo parsimony amounts to making a strong assumption about the pattern of gene family evolution, as it means that a gene family can be gained only once, on a single branch of the species phylogeny. In short, this means that gene transfers are forbidden. Wagner parsimony can be more moderate in its assumptions, but still requires that costs be defined for all types of events involved in gene family evolution, i.e. duplications, transfers, losses. There is no objective way to set these costs, and users often try a range of costs, eyeball the results, and choose the costs that produce the evolutionary scenarios that seem most reasonable [3]. The most systematic approaches use ancestral genome sizes to pick costs that generate ancestral genomes that are neither too big nor too small, but still lack a proper statistical framework [4,5].

Probabilistic approaches either rely on an ad hoc adaptation of substitution models used to describe sequence evolution [6], or rely on a birth–death model that includes rates of gene duplication, transfer and loss (DTL) [7,8]. They can include corrections for unobserved data, i.e. gene families that are present in none of the sampled species, but that were present in ancestral species [6]. These approaches do not require arbitrary choices of costs: instead rates are estimated from the data. Different models can then be tested against each other, for instance to test whether there is significant support for the presence of gene transfer in the data. These tests rely on the well-known machinery for model testing, and include likelihood ratio tests, Akaike or Bayesian information criterion, or Bayes factors if inference is performed in a Bayesian setting.

Whether they are analysed by parsimony or probabilistic approaches, gene content data are limited in their ability to detect events of gene family evolution. Even the approaches that use the numbers of genes and not just their pattern of presence/absence will make mistakes that approaches based on the consideration of gene tree topologies could avoid if the gene trees are accurately reconstructed (figure 1*b*).

(ii) Gene tree-aware approaches

Most gene families share parts of their histories, i.e. have been inherited together from ancestors to descendants during parts of their history (figure 1). If we can reconstruct the parts of their histories where genes have co-evolved, then jointly reconstructing gene histories can be very helpful, because more information is available to reconstruct each gene history. In cases where there is no gene transfer, then all genes share a common pattern of descent along the species tree. When genes can be transferred, they may share only part of their history with other gene families.

Gene tree-aware approaches were first used to deal with incomplete lineage sorting (ILS) through the multispecies coalescent [9]. In that framework, all gene families have evolved within the boundaries of the species history, and heterogeneities among gene histories originate from population-level sorting of alleles only. More recently, similar models have been proposed to deal with other processes of genome evolution, namely gene DTL. For an in-depth review, please see reference [1]. With these models, gene families can have a wider array of histories, and can differ drastically from the species tree. Invariably, whether they deal with ILS or DTL events, gene tree-species tree models have been found to

produce gene trees that are more accurate than competing approaches. This is expected: as more information is used to reconstruct gene trees, stochastic error should diminish.

Much like gene content approaches, gene tree-aware approaches can be based on probabilistic models that include parameters for DTL events [10–13], or on parsimonious models, in which case DTL events are associated with costs [14–16]. Gene tree-species tree approaches, however, are computationally challenging. Interpreting a gene tree in the light of a species tree by placing events of gene DTL, a process called *reconciling* a gene tree, is not difficult provided rates or costs of events are provided. Things get more complicated when the gene tree is not assumed to be known, and needs to be reconstructed. Naturally, if the species tree itself also needs to be reconstructed, then the task becomes extremely difficult; however, in the rest of this article, we will assume the species tree is known without uncertainty.

Methods to reconstruct gene trees using gene tree-aware approaches can use tree exploration heuristics similar to those found in commonly used programs for phylogenetic tree reconstruction [17–21], as in Phyldog [11] or in DLRS [22]. These approaches, however, tend to be slow, which motivated other approaches based on the consideration of a set of candidate gene trees obtained using faster approaches that do not consider a species tree. These approaches include TreefixDTL [15], ALE [23,24] and TERA [16]. The latter two approaches are extensions of an idea initially proposed in [5] and formalized in [23] and are particularly fast and accurate. They are based on the 'amalgamation' idea. Based on a sample of gene trees, amalgamation is a dynamic programming algorithm that allows the exhaustive exploration of a large space of gene trees. In fact, based on a limited set of gene trees, amalgamation allows consideration of a much larger space of gene trees, because it can piece together clades from several trees at a time to generate new trees, not present in the initial sample of gene trees. This technique is found to improve on competing approaches [16,23] in both speed and accuracy.

Probabilistic gene tree-aware approaches can also be used to date trees. In such cases, gene tree-aware models often reconstruct ultrametric gene trees, a model describing the rate of sequence evolution needs to be used, and an ultrametric species tree whose nodes are anchored in time is required [9,13,25–27]. Although these models contain additional parameters that need to be estimated, and are therefore computationally more complex to handle, they provide the ability to date events of gene family evolution along with the ability to estimate rates of events. Rates of events can then be compared across clades, although figure 1 is here to remind us that taxonomic sampling can have a non-trivial impact on the rates of reconstructed events. Another set of approaches avoids modelling the rate of sequence evolution and yet anchors events in time [10,14]. These approaches use a rooted ultrametric species tree in which nodes are ordered relative to each other, and mandate that transfers occur only between contemporaneous lineages. Gene trees, however, do not need to be ultrametric, which makes it possible to avoid using a model describing the rate of sequence evolution. Whether they use models describing the rate of sequence evolution or not, models that use ultrametric species trees are more realistic than models in which the nodes of the species tree are not ordered, because they include the constraint that only contemporaneous lineages

can exchange genes; however, this constraint comes at a high computational cost.

(iii) The impact of incomplete taxonomic sampling

No matter how complex our models of genome evolution, our inferences depend on the sampling of our dataset (figure 1). Although progress in sequencing methods is moving at a fast pace and genome sequences keep accumulating in databases, we will always be missing a clade or species and that will prevent our datasets from being complete. It is unclear how such missing data impacts our inferences. Figure 1 shows that missing species can lead to transfer events being incorrectly interpreted as duplication events, both for gene content and gene tree-aware approaches, but the magnitude of this effect is unknown. Worse, if our sampling of a clade misses a group of species with idiosyncratic characteristics (e.g. larger genomes, larger rates of gene transfer), then our estimate of the parameters of genome evolution for this group will be biased. In the hope of achieving an unbiased estimate of genome evolution, it is important to try to quantify the bias imposed by incomplete taxonomic sampling.

(iv) Comparing gene tree-aware and unaware approaches

Although reconstructing genome evolution is a widely pursued endeavour, there have been few assessments of the inference methods used to reconstruct gene histories along the species tree. In this article, we compare gene-content approaches with gene tree-aware approaches by using publicly available software on two well-known clades in the tree of life. We use a state-of-the-art probabilistic gene-content approach, *Count* [7,28], and a probabilistic gene tree-aware approach, *ALEml_undated* (available at https://github.com/ssolo/ALE), adapted to handle undated species trees. We address the impact of incomplete taxonomic sampling by performing rarefaction studies, whereby species are pruned from our species trees and DTL rates are compared across samples. Our primary aim is to focus on the inferences of the two methods and explain their differences in the light of their strengths and shortcomings. In the process, we will compare genome evolution in cyanobacteria and fungi.

(b) Genome evolution in fungi and cyanobacteria

Fungi and cyanobacteria *a priori* differ in the way their genomes have evolved. For instance, fungi undergo whole genome duplications (WGDs), whereas such events have not been reported in cyanobacteria. While gene transfer has been claimed to occur in both cyanobacteria and fungi, it is unclear how frequent this process has been in these two clades. Another question of interest concerns highways of gene transfers, i.e. pairs of branches or clades that appear to have undergone a high amount of gene transfers. While several highways of gene transfers have been claimed to exist in bacteria, including in cyanobacteria [29], it is unknown whether there are highways of gene transfers also in fungi.

Both cyanobacteria and fungi have been the focus of several studies addressing genome evolution, because they display a wide variety in cell types and genome sizes, and because they have had an important environmental impact throughout their history. In the context of this article, these clades constitute excellent case studies to assess the behaviour of gene content and gene tree-aware approaches because of their wide diversity in genome size, along with the fact that different evolutionary dynamics are expected in eukaryotes and bacteria.

(i) Genome evolution in fungi

Fungi are characterized by two life forms: one, yeast-like, is unicellular. The other is multicellular and includes fungi with macroscopic fruiting bodies as well as filamentous fungi. In this study, we focus on the clade Dikarya, a subkingdom of fungi that account for roughly 98% of described species. This clade is composed of two well-characterized phyla, Basidiomycota and Ascomycota. We use the genome sequences included in the HOGENOM database [30]. These two clades display a wide variety in genome sizes (from 5200 to 10 000 protein coding genes, approx.), and have a phylogeny that can be unambiguously rooted between Basidiomycota and Ascomycota. Studies of genome evolution in these clades have focused, for instance, on the impact of WGDs [31], on the evolution of the yeast (unicellular) form [32], or on the evolution of pathways for the decomposition of plant material [33,34]. Recently, there have been reports of notable amounts of gene transfers in fungi [35–37]. In particular, several examples indicate that the *Aspergillus* genome has been 'sculpted by gene transfer' [38]. This is consistent with reports that lateral gene transfers have been important throughout eukaryotic evolution [39].

(ii) Genome evolution in cyanobacteria

Cyanobacteria contain both unicellular organisms as well as organisms with two cell types, or that organize in filaments, which makes them unique among prokaryotes for their ability to leave a recognizable trace in the fossil record [40]. They display a wide range in genome size (from 1200 to 4500 protein coding genes, approx.), and have had a lasting impact on the Earth with the release of massive amounts of oxygen in the atmosphere billions of years ago [41]. From a phylogenomics perspective, cyanobacterial genomes share a relatively large core genome that allows the reconstruction of a well-supported species phylogeny despite the antiquity of the phylum. Cyanobacteria have also served as a model system for investigating horizontal gene transfer [10], and have been reported to display highways of gene transfers [29].

2. Methods

(a) Dataset construction
(i) Fungi

First, we selected all the species belonging to fungi in the HOGENOM6 database [30], yielding 32 species. We retrieved the protein sequences clustered into homologous gene families (21 701 families, discarding the very large families HOG100000000, HOG200000000 or HOG300000000, for which no alignment is available in the database). We discarded 8662 families containing only 2 or 1 genes from fungi. Gene trees were constructed for 1791 families containing only three genes (triplets), for which a single topology is possible. We aligned all families with four sequences or more using MUSCLE [42] with default parameters and selected reliably aligned sites using GBLOCKS [43]. The parameters employed were 'minimum number of sequences for a conserved position' b1 = 50, 'minimum number of sequences for a flank position' b2 = 50 and 'allowed gap positions' b5 = a (all). To estimate computing time per family, we measured the time PhyloBayes took [21] to compute 10 trees based on each alignment. We discarded the decile of the slowest families. For each remaining alignment, we ran two chains using PhyloBayes, calculating 5500 gene trees (discarding the first 500 as burn-in), using the LG model of evolution [44]. In the end, we were able to compute at least one chain for 9596 gene families. Combined with the 1791

triplets, our dataset contains in total 11 387 gene families, totalling 135 346 genes, whereas 24 327 genes were discarded during our selection process (not counting the three HOGENOM families without alignments).

Given that the tree of fungi is still unresolved with Microsporidia branching in an undefined place, we decided to use a smaller dataset, comprising the clade of Dikarya (28 species). This has the advantage that this clade can be easily rooted between Ascomycota and Basidiomycota. We pruned the gene trees removing from them two species of Microsporidia as well as *Allomyces macrogynus* and *Spizellomyces punctatus*, which belong to other basal clades of fungi. In total, we used 11 295 gene families. The gene trees are well resolved with an average posterior support of 0.97 (median = 1).

Owing to the uncertain position of *Aspergillus nidulans*, we relied on two species trees: one reconstructed from a concatenate, and one drawn from the literature. For our first tree, which we call tree A, we used a concatenate of 529 near universal single-copy gene family alignments (25 or more species represented out of 28). In total, the alignment contained 221 127 amino acid sites including 24 514 without missing data. Both PhyML [17] using the LG model of evolution [44] and a gamma distribution to account for rate variation [45] and Phylobayes [21] using the CAT model with Poisson exchangeabilities [46] recovered the same topology. We rooted the species tree between Ascomycota and Basidiomycota. The resulting phylogeny identifies the major clades, Pezizomycotina, which groups *Neurospora crassa* and *Aspergillus* fungi, and Saccharomycotina, which notably groups *Yarrowia lipolytica*, *Candida* species and *Saccharomyces*. Our phylogeny is in agreement with the phylogenies of Fitzpatrick *et al.* [47], but in their study the position of *A. nidulans* changes depending on the method: a concatenate based on 153 universal genes places it next to *Aspergillus fumigatus*, as we do, but supertree methods find it at the base of the *Aspergillus* clade. To account for these discrepancies, we reconstructed a second species tree where *A. nidulans* is at the base of the *Aspergillus* clade. We call this tree B, and estimated its branch lengths using PhyML with the same model as above. Tree A and Tree B can be found in the electronic supplementary material.

(ii) Cyanobacteria

We selected all the species belonging to cyanobacteria in the HOGENOM6 database [30], yielding 40 species. We reconstructed an unrooted species phylogeny using PhyML [17] using the LG model of evolution [44] and a gamma distribution to account for rate variation [45] from a concatenate of 470 near universal single-copy gene family alignments (38 or more species represented out of 40). In total, the alignment contained 126 180 amino acid sites including 67 646 without missing data. The resulting tree agreed with our genome-scale reconstruction [10] and other previous phylogenomic results (see discussion in [10], available in the electronic supplementary material). We rooted the species tree according to Szöllosi *et al.* [10]. For 7415 gene families with three or more genes, we employed the alignment procedure and sampling procedure described in §2a(i). The gene trees are well resolved with an average posterior support of 0.96 (median = 1).

(b) Inference methods
(i) Count

Count is a software package for performing studies in gene family evolution. It can perform ancestral genome reconstruction by posterior probabilities in a phylogenetic birth-and-death model [28]. Rates were optimized using a gain−loss−duplication model, with default parameters and allowing different gain−loss and duplication−loss rates for different branches. One hundred rounds of optimization were computed.

(ii) ALEml_undated

ALEml_undated implements a probabilistic approach to exhaustively explore all reconciled gene trees that can be amalgamated as a combination of clades observed in a sample of gene trees [24] in the context of different species tree–gene tree reconciliation models, in particular the model described in [23], which allows for the DTL of genes. ALE can be used to efficiently approximate the sum of the joint likelihood over amalgamations and to find the reconciled gene tree that maximizes the joint likelihood among all such trees or sample the space of possible reconciliations. Here, we use two reconciliation methods, a simplified DTL approach that does not consider the temporal information from the species tree, and a version of this model that only allows duplication and loss (DL). These methods are available as part of the open-source ALE project (https://github.com/ssolo/ALE).

(c) Analyses

Highways were identified between pairs of species that exchange large numbers of genes. The number of genes exchanged was averaged over 100 reconciliations drawn from ALEml_undated using the program ALEsample, and summed across all gene families in our datasets.

Synteny information was extracted from gene positions in the genomes. Pairwise comparisons of genomes between species were performed. Synteny was found to be conserved if a gene had as a neighbour a gene whose orthologue was also its own orthologue's neighbour. For simplicity, only gene families with one gene per species were considered in the synteny analyses. For a given pairwise comparison, a gene was declared as non-transferred if, along the path between two species, no transfer event had affected the gene of each species, and declared as transferred otherwise.

3. Results

(a) General patterns of genome evolution

Figures 2 and 3 show the reconstruction of genome evolution across fungi and cyanobacteria, respectively, using both Count and ALEml_undated. Although Count and ALEml_undated differ in their input data and in the types of events they can detect, their inferences are qualitatively similar, finding comparable genome size dynamics and proportions of events on branches.

In both cyanobacteria and fungi, with both methods, a clade with large genomes (multicellular Aspergillus clade of moulds, and the clade including freshwater and multicellular cyanobacteria such as Nostoc and Cyanothece) and a clade with smaller genomes (unicellular clade of yeasts including Saccharomyces and Candida, and unicellular planktonic cyanobacteria including Prochlorococcus and Synechococcus) can be observed.

In fungi, the clade with large genomes (fungi from the multicellular Aspergillus clade of moulds) shows a large portion of gene transfers on several of its branches, whereas gene transfers appear much less prevalent in the clade with small genomes (containing the unicellular yeasts). This result confirms earlier reports based on smaller datasets of larger amounts of gene transfers in the Aspergillus clade than in the yeast clade [37]. Several branches show an excess of gene duplications compared to gene transfers. Although a WGD occurred in the ancestor of Saccharomyces cerevisiae and Candida glabrata [31], both models fail to pick an increased amount of gene duplications on the relevant branch. They recover an increased amount of duplications on the branch leading to Saccharomyces cerevisiae alone, possibly because Candida glabrata has lost a

large number of genes, which, in the absence of synteny (which was used by Wapinski et al. [31] to detect the WGD), may have erased a large part of the signal supporting the WGD. The ancestor of all Dikarya is predicted to have a very small genome, which is likely the consequence of our unbalanced taxonomic sampling with only three Ascomycota. Owing to this design, families present only in two Ascomycota have been discarded from our dataset, and therefore cannot be inferred at the root.

In cyanobacteria, both the clades with large and small genomes appear to have similar genome dynamics, with more gene transfers than gene duplications. The ancestor of cyanobacteria is predicted to have intermediate genome content in between that of the clade with small genomes (containing Prochlorococcus species) and that of the clade with larger genomes.

(b) Gene tree-aware approaches are more sensitive than gene content approaches

By design, gene tree-aware approaches can detect more events than gene-content approaches (figure 1). Consistently, ALEml_undated finds significantly more transfers than Count, with ALEml_undated finding an average of 0.16 and 0.07 transfers per gene in, respectively, cyanobacteria and fungi, in contrast to Count, which finds 0.14 and 0.06. It is difficult to determine how many of the additional transfers are due to ALEml_undated finding true transfer events that Count failed to detect and how many result from errors in reconstructed gene trees. Simulations do indicate that ALE recovers an unbiased estimate of the number of transfers [23], and in the case of cyanobacteria reduces the number of inferred transfers by approximately two-thirds compared with gene trees reconstructed without the species tree (by PhyML [17]). Furthermore, figure 4a shows that for cyanobacterial families represented in eight or fewer genomes the ALEml_undated and Count estimates closely agree. Regardless of potential overestimation of the number of transfers, figure 4a also highlights that the number of transfer events per gene family is more-or-less homogeneous with respect to the number of species represented in the gene family for ALEml_undated, but systematically decreases for Count as more complete taxonomic distributions are approached. We see no biological reason why, for example, families with a complete taxonomic distribution would undergo much fewer transfers compared with families with slightly incomplete taxonomic sampling. Instead, we believe this effect results from a shortcoming of gene tree unaware methods, such as Count, whereby they are not able to infer transfer among families with complete taxonomic distribution, and progressively lose signal as complete taxonomic distribution is approached.

(c) Duplication and loss methods systematically overestimate ancestral gene content

The effect of gene transfers on gene phylogenies can be mimicked by a combination of gene duplications and losses. Therefore, gene duplications and losses may be sufficient to account for genome dynamics in our two clades, and it is legitimate to ask about the need to incorporate transfers. As shown in figure 4b comparison of the gene content of extant genomes and reconstructions based on gene-tree aware reconstruction that considers transfer (DTL) shows that these methods reconstruct ancestral gene contents that are similar to those observed for extant genomes. In stark contrast, gene-tree aware
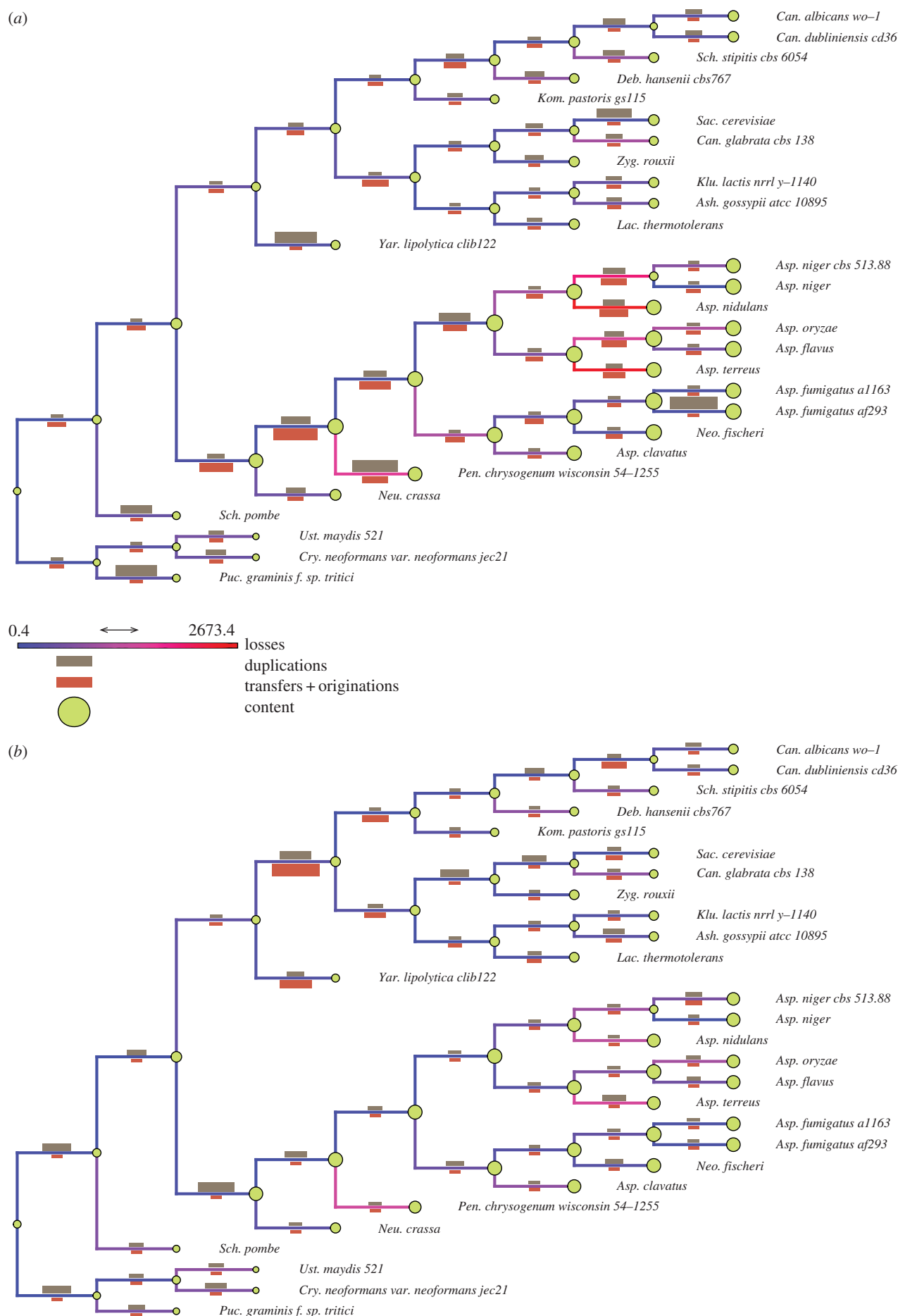
Figure 2. Genome evolution in Ascomycota and Basidiomycota (tree A). Edges are colour-coded according to the inferred numbers of losses along the branches. Crimson bars represent numbers of gene gains (transfers + originations) arriving on the branch; taupe bars represent numbers of duplications happening on the branch. At each node, genome content size is represented as a green disc. (*a*) Inferences from *ALEml_undated*. (*b*) Inferences from *Count*. The corresponding graph for tree B is available in the supplementary material.
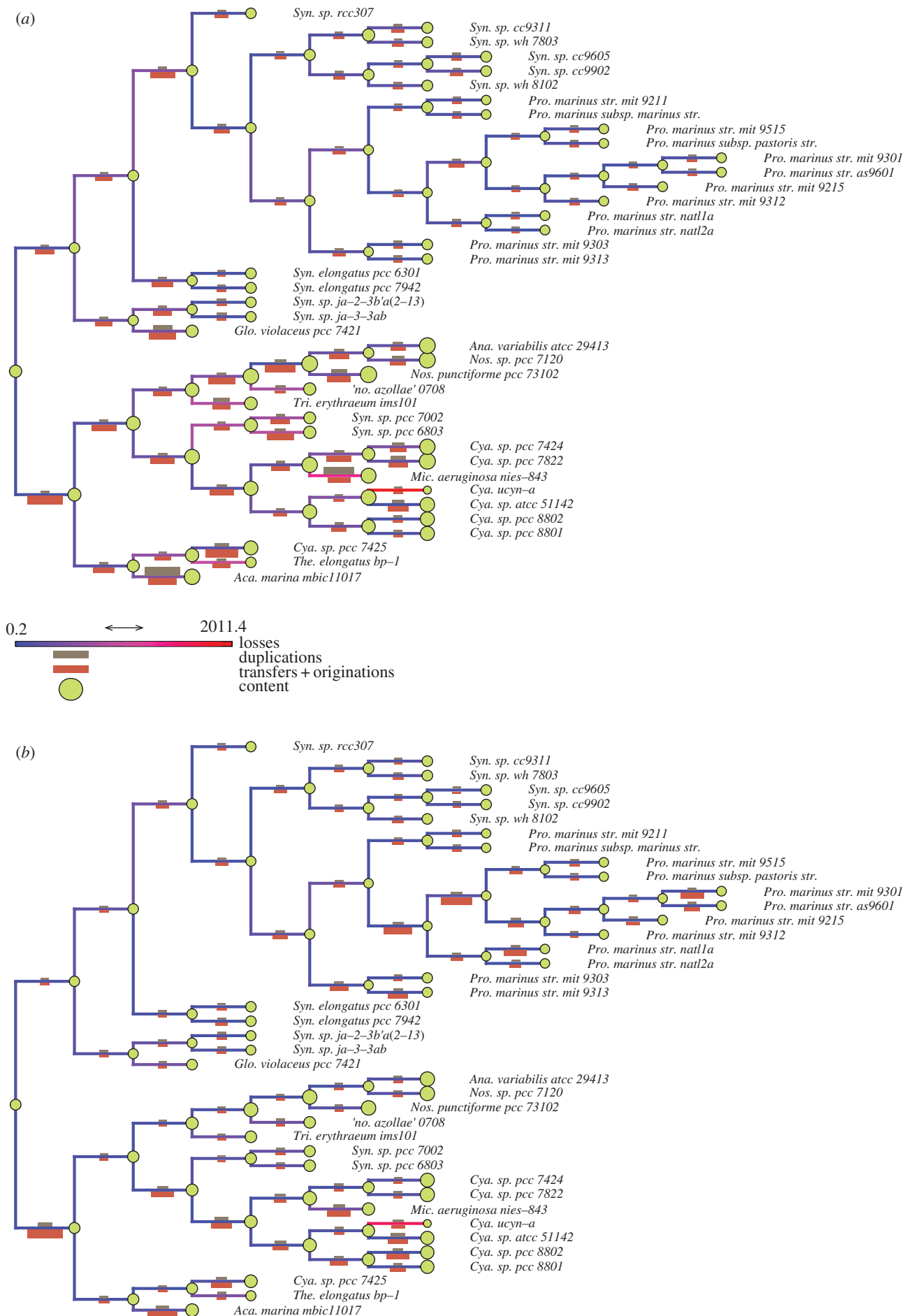
Figure 3. Genome evolution in cyanobacteria. Edges are colour-coded according to the inferred numbers of losses along the branches. Crimson bars represent numbers of gene gains (transfers + originations) arriving on the branch; taupe bars represent numbers of duplications happening on the branch. At each node, genome content size is represented as a green disc. (*a*) Inferences from *ALEml_undated*. (*b*) Inferences from *Count*.
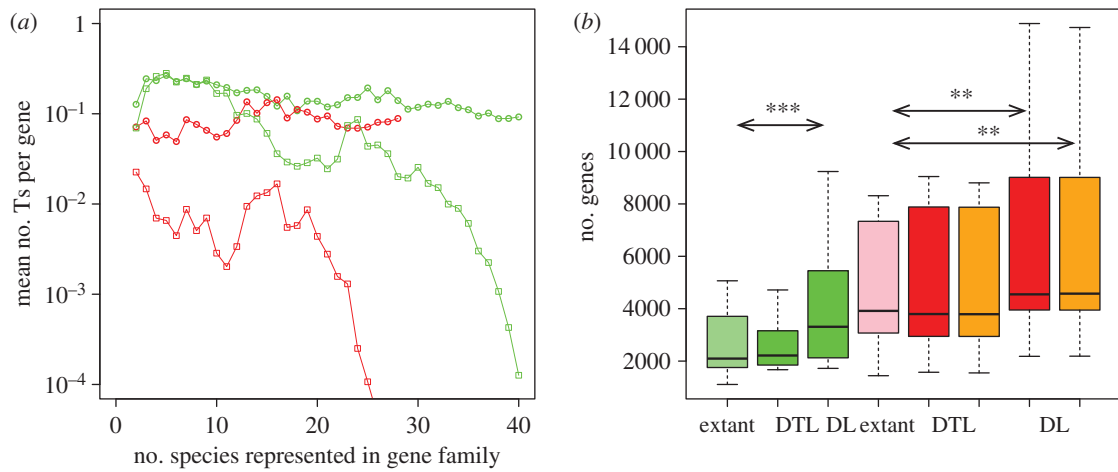
Figure 4. Comparison between alternative genome evolution reconstruction methods. (*a*) Comparing methods that consider transfers alongside duplications and losses, we find that the gene-tree aware method (*ALEml_undated*, circles) infers an order of magnitude more transfers for fungi. The gene-tree unaware method (*Count*, squares) loses its ability to detect transfers as gene families become universal or near universal (28 species for fungi, 40 for cyanobacteria), whereas the gene-tree aware method does not. (*b*) To differentiate gene-tree aware reconstruction that considers transfer (DTL) to one that only accounts for duplications and losses, but does not consider transfer (DL), we compare the number of genes in ancestral genomes to the number of genes in extant genomes. The DTL method infers ancestral genome sizes that fall within the distribution for extant genomes ($p > 0.5$ two-sided Wilcoxon test) for both fungi and cyanobacteria. In contrast, gene numbers based on the DL method are systematically larger than extant ones ($p < 10^{-3}$ for cyanobacteria and $p = 0.02$ for fungi with a one-sided Wilcoxon test for both tree A and B). Fungi are in red (tree A) and orange (tree B), and cyanobacteria in green throughout.

reconstructions that only account for duplications and losses, but do not consider transfer (DL), infer systematically more genes to have been present in ancestral genomes than in extant ones. The largest ancestral genomes are inferred by DL methods in the common ancestors of clades where DTL methods predict the most transfers: the deepest node in the *Aspergillus* genus has ancestral gene contents of 14 244 genes according to the DL estimate, in comparison with 8238 genes for the DTL-based estimate. The extant gene content in the genus in our sample is between 7797 (*A. fumigatus* A1163) and 8891 (*Aspergillus terreus*).

(d) Rates of transfers are similar in fungi and cyanobacteria

Rates of transfer in fungi and in cyanobacteria appear to be very similar, as shown by the *ALEml_undated* inferences (figure 5*a*). This finding does not come from differences in the age of the clades as we compare ratios of numbers of events, which controls for age. It does not appear to come from incomplete sampling either, as figure 5*b* shows that predictions based on subsampling the species in each dataset still converge to similar ratios of numbers of events for fungi and cyanobacteria. To extrapolate the $T/(T + D)$ values, we fit an ad hoc curve that reaches saturation exponentially starting from an initial value for zero species. Using all subsampled replicates a least-squares Marquardt–Levenberg algorithm yielded the similar asymptotic values of $T/(T + D)$, with 0.8 ± 0.1 (fungi assuming tree A), 0.7 ± 0.03 (fungi assuming tree B) and 0.74 ± 0.01 in cyanobacteria. The same procedure for $L/(T + D + L)$ produced the slightly higher asymptotic value for fungi of 0.582 ± 0.01 for tree A and 0.595 ± 0.01 for tree B, compared with 0.52 ± 0.01 compared with cyanobacteria.

These genome-wide inferences confirm earlier reports based on manual analyses of smaller datasets that significant numbers of transfers occurred in fungi, in particular in the *Aspergillus* clade [35–38]. Overall, these data show that genomes in prokaryotes and eukaryotes are not undergoing fundamentally different dynamics. We consider that additional analyses of datasets for different clades of both prokaryotes and eukaryotes, using gene tree-aware approaches as in this work, would provide a more fine-grained, quantitative view of the dynamics of genome evolution across the entire tree of life.

(e) There are highways of gene transfer in fungi

One feature of genome evolution in prokaryotes that has received considerable attention is the concept of highways of gene transfers [29,48]. According to this model, some pairs of species or clades have exchanged large numbers of genes throughout their history, possibly because of a shared ecological niche. *ALEml_undated* inferences provide us with an opportunity to look for such highways in cyanobacteria and fungi. Figure 6 shows the distribution of the number of transfers per pairs of branches of the species tree in both cyanobacteria and fungi. Both distributions show a long tail, with many transfers occurring between branches that otherwise have exchanged little genetic material. However, some pairs of branches show very high numbers of gene transfer events. The heterogeneity is strongest in fungi, where some pairs of branches are predicted to have undergone more than 150 gene transfers, or even 300 transfers on tree B (figure 7). These transfers do not seem to be due to hybridization, as most of them are not replacement transfers, whereby a gene in a species is replaced by another gene coming from another species (the median branch-wise fraction of gene transfers that are compensated by loss on the same branch, i.e. replacement transfers, is 31% for fungi on tree A, 34% on tree B and 49% for cyanobacteria). For the same reason, these transfers cannot be misinterpreted events of ILS. In fact, among genes that have only one orthologue per species, genes that have undergone a gene transfer tend to change position on the chromosome more often than genes that have not undergone a gene transfer (see figure 7, right, for the *Aspergillus* clade). The pairs of branches with the largest numbers of transfers belong to the *Aspergillus* clade, in agreement with the overall larger amount of transfers
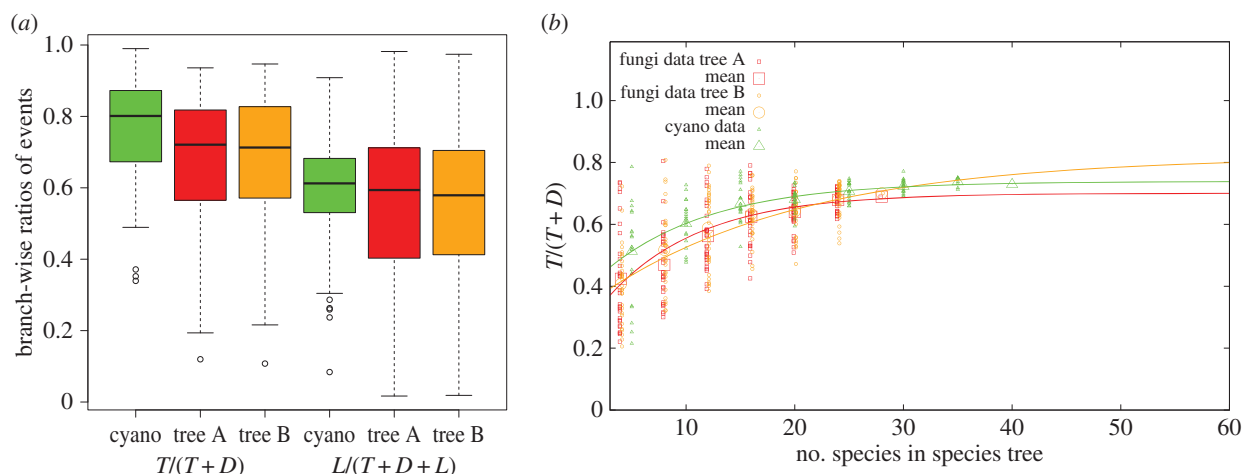
Figure 5. Similar rate of transfers in fungi and cyanobacteria. The cyanobacterial and fungal datasets we considered differ both in the number of genomes (40 versus 28) and in their age (nearly three billion years versus. less than three quarters of a billion years). (a) In order to account for differences in age, we consider the ratio $T/(T+D)$, i.e. the fraction of transfer events among all gene birth events (duplications plus transfers) and $L/(T+D+L)$ the ratio of loss events to all events. (b) To ascertain the effect of the differences in the number of genomes, we constructed replicates with random subsamples of genomes of varying numbers. Extrapolation of the results suggests that $T/(D+T)$ for fungi is as large or larger than for cyanobacteria (for details see §3d). Fungi are shown with squares for tree A (red), and circles for tree B (orange) and cyanobacteria with triangles (green).
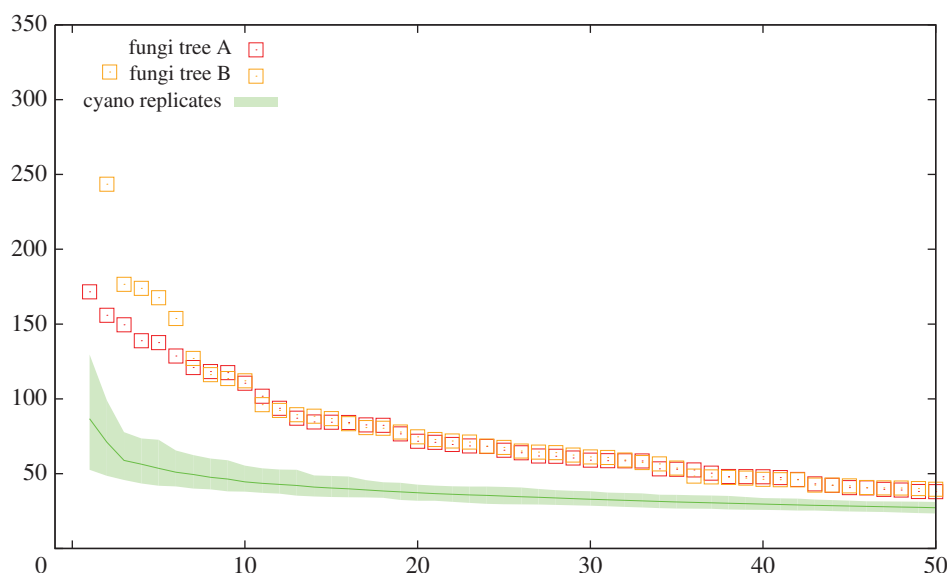

Figure 6. Stronger highways in fungi. Data points, red for tree A and orange for tree B, correspond to numbers of transfers between pairs of branches ('highways of transfers') in either fungi phylogeny plotted in decreasing order. The continuous line (green online) shows the mean number of transfers between pairs of branches among 25 replicates where a random set of 28 cyanobacterial genomes were chosen as in figure 5b. The shaded area shows the 95% CI. Fungi are in red and cyanobacteria in green throughout.

detected in this clade and in agreement with previous reports [38]. The species involved in the largest number of transfers in either tree A or tree B is *A. nidulans*, precisely the species whose position is contentious. This suggests that lateral gene transfers in fungi may be significant enough to make reconstruction of the species phylogeny difficult. Although deeper sampling could change the numbers of gene transfers found on each pair of branches, for instance by breaking branches involved in a highway, it seems unlikely that the conclusion that there are branch pairs or group of branches exchanging large numbers of genes in fungi would change.

(f) Genes tend to be transferred together
The distribution of transferred genes along chromosomes appears to be consistent with the transfer of chromosomal

segments that can include more than one gene. Counting only transfers to the terminal branches, transferred genes appear preferentially next to another transferred gene: on average 4.7 times more often in fungi (on tree A; 6.1 times more often on tree B), and 5.3 times more often in cyanobacteria. Given that genes transferred on terminal branches make up a minority of the genomes, this means that transfers tend to affect blocks of several genes at a time.

(g) *ALEml_undated* reconstructs accurate gene trees
In [23], we found using realistic simulations that amalgamation of gene trees using a DTL model produced accurate gene trees: the number of duplications and transfers needed to reconcile our reconstructed trees was statistically indistinguishable from the corresponding number of events needed
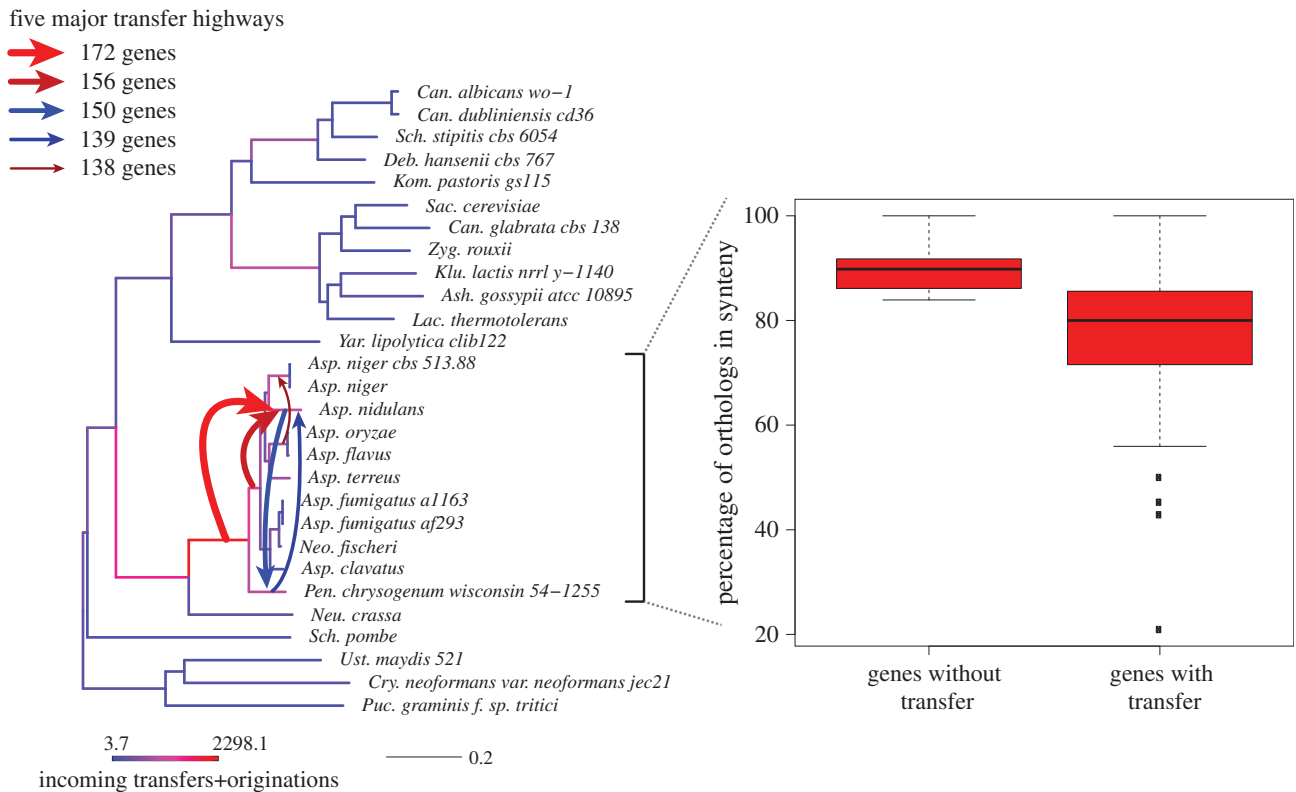
Figure 7. Position of the five largest highways in fungi on tree A. Branch lengths are in expected numbers of amino acid substitutions, and coloured according to the number of incoming gene transfers and originations. The five largest highways of transfers are represented. On the right, boxplots show pairwise genome synteny comparisons. Transferred genes are found to change neighbours more often than non-transferred genes. On tree B, the five largest highways concern 329, 244, 177, 174, 168 genes, the three largest ones of which involve *Aspergillus nidulans*.

to reconcile the 'real' trees that had been used to simulate gene alignments. In this study, empirical results also show that gene trees reconstructed by *ALEml_undated* are accurate. First, the fact that the reconstructions of ancestral genome sizes based on our reconciled gene trees are not significantly different from extant genome sizes suggests that our gene trees do not contain large numbers of incorrect bipartitions. Second, the over-representation of transferred genes in tandem cannot be explained by random errors in gene trees, but shows that bona fide information can be retrieved from gene trees reconstructed by *ALEml_undated*.

4. Conclusion

Our genome-scale phylogenetic analysis of genome evolution in cyanobacteria and fungi shows that fungi exhibit similar rates of transfers as cyanobacteria, and display apparent highways of gene transfers. Whether these highways of gene transfers correspond to shared ecological niches or to particular mechanisms to incorporate foreign DNA remains to be investigated. In both clades, gene transfers appear to occur in blocks, not just one gene at a time. Further investigation of those transferred blocks of genes may prove useful for functional annotation, as co-transferred genes may be functionally related.

This study also allows the comparative study of different methodologies for reconstructing genome evolution. We show that the recent developments provide a framework adapted to different domains of life, and that gene tree-aware methods show more precision in the quantification of gene transfers.

Our results suggest that further analyses of datasets for other clades of prokaryotes and eukaryotes, using gene tree-aware approaches, will provide a more fine-grained, quantitative view of the dynamics of genome evolution across the tree of life.

References

1. Szollosi GJ, Tannier E, Daubin V, Boussau B. 2015 The inference of gene trees with species trees. *Syst. Biol.* **64**, 1063–5157. (doi:10.1093/sysbio/syu048)

2. Groussin M, Hobbs JK, Szöllősi GJ, Gribaldo S, Arcus VL, Gouy M. 2015 Toward more accurate ancestral protein genotype-phenotype reconstructions with the use of species tree-aware gene trees. *Mol. Biol. Evol.* **32**, 13–22. (doi:10.1093/molbev/msu305)

11

rstb.royalsocietypublishing.org *Phil. Trans. R. Soc. B* **370**: 20140335

3. Boussau B, Karlberg EO, Frank AC, Legault B-AA, Andersson SGE. 2004 Computational inference of scenarios for α-proteobacterial genome evolution. *Proc. Natl Acad. Sci. USA* **101**, 9722–9727. (doi:10.1073/pnas.0400975101)

4. Dagan T, Martin W. 2007 Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl Acad. Sci. USA* **104**, 870–875. (doi:10.1073/pnas.0606318104)

5. David LA, Alm EJ. 2010 Rapid evolutionary innovation during an archaean genetic expansion. *Nature* **469**, 93–96. (doi:10.1038/nature09649)

6. Cohen O, Pupko T. 2010 Inference and characterization of horizontally transferred gene families using stochastic mapping. *Mol. Biol. Evol.* **27**, 703–713. (doi:10.1093/molbev/msp240)

7. Csűrös M, Miklós I. 2006 A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. *Res. Comput. Mol. Biol.* **3909**, 206–220 (doi:10.1007/11732990_18)

8. Szöllosi GJ, Daubin V. 2012 Modeling gene family evolution and reconciling phylogenetic discord. *Methods Mol. Biol.* **856**, 29–51. (doi:10.1007/978-1-61779-585-5_2)

9. Rannala B, Yang Z. 2003 Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656.

10. Szöllosi GJ, Boussau B, Abby SS, Tannier E, Daubin V. 2012 Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl Acad. Sci. USA* **109**, 17 513–17 518. (doi:10.1073/pnas.1202997109)

11. Boussau B, Szöllosi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013 Genome-scale coestimation of species and gene trees. *Genome Res.* **23**, 323–330. (doi:10.1101/gr.141978.112)

12. Akerborg O, Sennblad B, Arvestad L, Lagergren J. 2009 Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl Acad. Sci. USA* **106**, 5714–5719. (doi:10.1073/pnas.0806251106)

13. Rasmussen MD, Kellis M. 2010 A Bayesian approach for fast and accurate gene tree reconstruction. *Mol. Biol. Evol.* **28**, 273–290. (doi:10.1093/molbev/msq189)

14. Doyon JP, Scornavacca C, Gorbunov KY, Szöllősi GJ, Ranwez V, Berry V. 2010 An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications, and transfers. In *Comparative genomics*, vol. 6398 (ed. E Tannier), pp. 93–108. Berlin, Germany: Springer.

15. Bansal MS, Wu Y-C, Alm EJ, Kellis M. 2014 Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics* **31**, 1211–1218. (doi:10.1093/bioinformatics/btu806)

16. Scornavacca C, Jacox E, Szöllősi GJ. 2015 Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics* **31**, 841–848. (doi:10.1093/bioinformatics/btu728)

17. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321. (doi:10.1093/sysbio/syq010)

18. Stamatakis A, Ludwig T, Meier H. 2005 RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **15**, 456–463. (doi:10.1093/bioinformatics/bti191)

19. Ronquist F *et al.* 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542. (doi:10.1093/sysbio/sys029)

20. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006 Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88. (doi:10.1371/journal.pbio.0040088)

21. Lartillot N, Lepage T, Blanquart S. 2009 PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288. (doi:10.1093/bioinformatics/btp368)

22. Sjöstrand J, Sennblad B, Arvestad L, Lagergren J. 2012 DLRS: gene tree evolution in light of a species tree. *Bioinformatics* **28**, 2994–2995. (doi:10.1093/bioinformatics/bts548).

23. Szöllösi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013 Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912. (doi:10.1093/sysbio/syt054)

24. Szöllösi GJ, Tannier E, Lartillot N, Daubin V. 2013 Lateral gene transfer from the dead. *Syst. Biol.* **62**, 386–397. (doi:10.1093/sysbio/syt003)

25. Heled J, Drummond AJ. 2010 Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27**, 570–580. (doi:10.1093/molbev/msp274)

26. Liu L, Pearl DK. 2007 Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* **56**, 504–514. (doi:10.1080/10635150701429982)

27. Arvestad L, Berglund A-C, Lagergren J, Sennblad B. 2003 Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* **19**, i7–i15. (doi:10.1093/bioinformatics/btg1000)

28. Csurös M. 2010 Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912. (doi:10.1093/bioinformatics/btq315)

29. Beiko RG, Harlow TJ, Ragan MA. 2005 Highways of gene sharing in prokaryotes. *Proc. Natl Acad. Sci. USA* **102**, 14 332–14 337. (doi:10.1073/pnas.0504068102)

30. Penel S, Arigon A-M, Dufayard J-F, Sertier A-S, Daubin V, Duret L, Gouy M, Perrière G. 2009 Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* **10**(Suppl. 6), S3. (doi:10.1186/1471-2105-10-S6-S3)

31. Wapinski I, Pfeffer A, Friedman N, Regev A. 2007 Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**, 54–61. (doi:10.1038/nature06107)

32. Nagy LG *et al.* 2014 Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nat. Commun.* **5**, 4471. (doi:10.1038/ncomms5471)

33. Eastwood DC *et al.* 2011 The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. *Science* **333**, 762–765. (doi:10.1126/science.1205411)

34. Floudas D *et al.* 2012 The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* **336**, 1715–1719. (doi:10.1126/science.1221748)

35. Slot JC, Rokas A. 2011 Horizontal transfer of a large and highly toxic secondary metabolic gene cluster between fungi. *Curr. Biol.* **21**, 134–139. (doi:10.1016/j.cub.2010.12.020)

36. Richards TA, Leonard G, Soanes DM, Talbot NJ. 2011 Gene transfer into the fungi. *Fungal Biol. Rev.* **25**, 98–110. (doi:10.1016/j.fbr.2011.04.003)

37. Richards TA. 2011 Genome evolution: horizontal movements in the fungi. *Curr. Biol.* **21**, R166–R168. (doi:10.1016/j.cub.2011.01.028)

38. Gibbons JG, Rokas A. 2013 The function and evolution of the *Aspergillus* genome. *Trends Microbiol.* **21**, 14–22. (doi:10.1016/j.tim.2012.09.005)

39. Hirt RP, Alsmark C, Embley TM. 2015 Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites. *Curr. Opin. Microbiol.* **23**, 155–162. (doi:10.1016/j.mib.2014.11.018)

40. Tomitani A, Knoll AH, Cavanaugh CM, Ohno T. 2006 The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. *Proc. Natl Acad. Sci. USA* **103**, 5442–5447. (doi:10.1073/pnas.0600999103)

41. Kaufman AJ *et al.* 2007 Late Archean biospheric oxygenation and atmospheric evolution. *Science* **317**, 1900–1903. (doi:10.1126/science.1138700)

42. Edgar RC. 2004 MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113. (doi:10.1186/1471-2105-5-113)

43. Castresana J. 2000 Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552. (doi:10.1093/oxfordjournals.molbev.a026334)

44. Le SQ, Gascuel O. 2008 An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320. (doi:10.1093/molbev/msn067)

45. Yang Z. 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314. (doi:10.1007/BF00160154)

46. Lartillot N, Philippe H. 2004 A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109. (doi:10.1093/molbev/msh112)

47. Fitzpatrick DA, Logue ME, Stajich JE, Butler G. 2006 A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol. Biol.* **6**, 99. (doi:10.1186/1471-2148-6-99)

48. Bansal MS, Banay G, Harlow TJ, Gogarten JP, Shamir R. 2013 Systematic inference of highways of horizontal gene transfer in prokaryotes. *Bioinformatics* **29**, 571–579. (doi:10.1093/bioinformatics/btt021)