Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations.

Gergely J Szöllosi, Bastien Boussau, Sophie S Abby, Eric Tannier, Vincent Daubin

▶ **To cite this version:**

Gergely J Szöllosi, Bastien Boussau, Sophie S Abby, Eric Tannier, Vincent Daubin. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations.. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109 (43), pp.17513-17518. 10.1073/pnas.1202997109 . hal-00740292

HAL Id: hal-00740292
https://hal.science/hal-00740292

Submitted on 29 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations

Gergely J. Szöllősi[a,b], Bastien Boussau[a,b,c], Sophie S. Abby[d,e], Eric Tannier[a,b,f], and Vincent Daubin[a,b,1]

[a]Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Université Lyon 1, F-69622 Villeurbanne, France; [b]Université de Lyon, F-69000 Lyon, France; [c]Department of Integrative Biology, University of California, Berkeley, CA 94720-3140; [d]Microbial Evolutionary Genomics, Département Génomes et Génétique, Institut Pasteur, F-75015 Paris, France; [e]Centre National de la Recherche Scientifique, Unité Mixte de Recherche 3525, F-75015 Paris, France; and [f]Institut National de Recherche en Informatique et en Automatique Rhône-Alpes, F-38334 Montbonnot, France

Edited by Marc A. Suchard, University of California, Los Angeles, CA, and accepted by the Editorial Board September 7, 2012 (received for review February 20, 2012)

The timing of the evolution of microbial life has largely remained elusive due to the scarcity of prokaryotic fossil record and the confounding effects of the exchange of genes among possibly distant species. The history of gene transfer events, however, is not a series of individual oddities; it records which lineages were concurrent and thus provides information on the timing of species diversification. Here, we use a probabilistic model of genome evolution that accounts for differences between gene phylogenies and the species tree as series of duplication, transfer, and loss events to reconstruct chronologically ordered species phylogenies. Using simulations we show that we can robustly recover accurate chronologically ordered species phylogenies in the presence of gene tree reconstruction errors and realistic rates of duplication, transfer, and loss. Using genomic data we demonstrate that we can infer rooted species phylogenies using homologous gene families from complete genomes of 10 bacterial and archaeal groups. Focusing on cyanobacteria, distinguished among prokaryotes by a relative abundance of fossils, we infer the maximum likelihood chronologically ordered species phylogeny based on 36 genomes with 8,332 homologous gene families. We find the order of speciation events to be in full agreement with the fossil record and the inferred phylogeny of cyanobacteria to be consistent with the phylogeny recovered from established phylogenomics methods. Our results demonstrate that lateral gene transfers, detected by probabilistic models of genome evolution, can be used as a source of information on the timing of evolution, providing a valuable complement to the limited prokaryotic fossil record.

molecular dating | gene tree reconciliation | birth-death model

A central aspect of Earth's history is the pattern and timing of diversification of the species that inhabit it. In macroorganisms such as animals or plants, an abundant fossil record, the accumulation of genomic data, and the development of models of molecular evolution accommodating for varying rates of evolutionary changes among lineages are progressively yielding an intelligible picture (1–4). In contrast, the dating of the evolution of microbial life remains largely elusive (5, 6). This situation results from the convergence of two main factors: first, fossils, especially bacterial and archaeal ones, are scarce or cannot be traced to a specific lineage. Therefore, any inference of the timing of microbial evolution must rely almost exclusively on molecular data constrained only by a handful of dates during the course of more than three billion years of evolution. Second, molecular data can be difficult to interpret in terms of patterns of species diversification. Lateral gene transfers (LGTs), the exchange of genes among possibly distant species, have tangled gene phylogenies to the extent that they provide a deeply blurred view of the relationships between lineages. Different approaches [e.g., concatenation, supertrees (7, 8)] have been proposed to overcome this problem, but these methods only combine the inevitably conflicting phylogenetic signal from a limited number of gene families without accounting for the nature of this conflict (9). Although simulations tend to suggest that such average trees can capture true evolutionary relationships, the interpretation of these models remains uncertain and controversial (10, 11) as they offer no explanation for phylogenetic discord and are potentially sensitive to systematic bias (12).

Phylogenetic discord occurs because gene histories do not correspond to the pattern of species diversification but are the end result of a series of speciation, duplication, transfer, and loss events (13). This means that a gene tree (Fig. 1*A*) inferred from homologous sequences found in extant genomes cannot directly be interpreted in terms of patterns of species diversification, as by itself it does not tell us which ancestral gene lineages belonged to which ancestral genomes. To accomplish this, we need to propose an explicit series of events describing the evolution of a gene tree along the species tree that reconciles the relationship between ancestral gene and species lineages (Fig. 1*B*) and explains phylogenetic discord between gene trees. Interestingly, in the presence of transfer, the set of possible reconciliations depends not only on the topology of the species tree, but also on the order of speciation events in time (Fig. 1 *C* and *D*). This suggests that it is possible to use gene tree reconciliation to deduce not only the topology, but also the order of speciations in time.

Here, we demonstrate that a probabilistic model that explicitly accounts for the events that generate differences between gene phylogenies and the species tree—that is, lateral transfer, duplication, and loss of genes—can be used to reconstruct the most likely species tree given an ensemble of gene trees. Because LGT can only occur among contemporary species, this tree describes not only the pattern of speciations but also their order in time. We use simulations to show that the model correctly returns the time-ordered species tree and is robust to errors in gene tree reconstruction. Using data from complete genomes, we also show that our model reconstructs species tree topologies close to those obtained by existing, less sophisticated methods, while at the same time extracting previously overlooked information on the timing of the evolution of prokaryotic life.

Results

Likelihood of a Gene Tree. The species phylogeny describes the series of speciations that lead to existing species and can be

Author contributions: G.J.Sz., B.B., E.T., and V.D. designed research; G.J.Sz. performed research; S.S.A. contributed new reagents/analytic tools; G.J.Sz., B.B., E.T., and V.D. analyzed data; and G.J.Sz., B.B., S.S.A., E.T., and V.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. M.A.S. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: vincent.daubin@univ-lyon1.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1202997109/-/DCSupplemental.

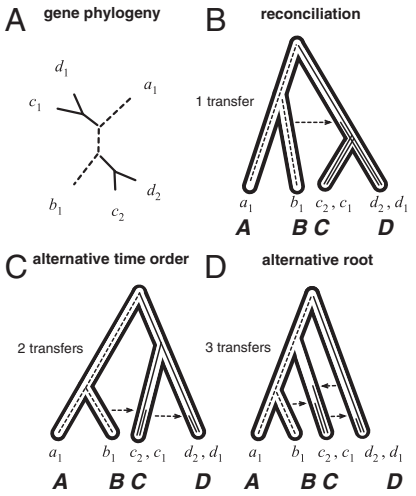

Fig. 1. A gene tree–species tree reconciliation invoking gene transfer and loss. (*A*) A gene tree topology. (*B*) A possible reconciliation of the gene tree in *A* with the species tree invoking one event of transfer. Note that the transfer from the branch leading to ***B*** to above the ancestor of ***C*** and ***D*** implies that the ancestor of ***A*** and ***B*** is older than the ancestor of ***C*** and ***D***. (*C* and *D*) Alternative time orders (*C*) or rootings (*D*) of the species tree violate this condition and only allow reconciliations with a larger number of events.

represented as a rooted bifurcating tree with branch lengths representing time. The leaves of this tree correspond to extant species and their genomes, its internal nodes to ancestral species, and their genomes. Calculating the probability of a gene tree requires summing over all reconciliations—that is, all paths on the species phylogeny along which a series of duplication, transfer, loss, and speciation events could have generated the corresponding unrooted topology. To carry out these calculations, we extend the approach introduced by Tofigh (14), where time along the species tree is coarse grained such that duplication, transfer, and loss events are specified to occur in a given "time slice" (*SI Appendix*, Fig. S1) corresponding to an interval of time between two successive speciations (*SI Appendix*, Fig. S1 gray nodes). In the calculations presented in the article, we transform the species tree S such that all time slices have equal width (*SI Appendix*, Fig. S1*B*).

For a given species tree S along which the order of speciation events is specified together with a set of origination probabilities and duplication, transfer, and loss rates, denoted $\mathcal{M}$, we can calculate the probability of observing a gene topology G as the sum over all reconciliations. We refer to this procedure as the ODT (Origination, Duplication, Transfer, and Loss) model, and it is described in detail in *SI Appendix*, section S1. In the context of the ODT model, the probability of observing a gene topology G is:

$$p(G|S,\mathcal{M}) = \sum_{\vec{R}\in\mathcal{R}(G)} \left(\sum_{x\in\mathcal{N}(S)} p_{\mathrm{O}}(x) P(\vec{R},x) \right), \qquad [1]$$

where $\mathcal{R}(G)$ is the set of all possible roots of G, $\mathcal{N}(S)$ corresponds to all positions along S (*SI Appendix*, Fig. S1), $p_{\mathrm{O}}(x)$ corresponds to the probability of origination at position x along S, and $P(\vec{R},x)$ corresponds to the probability of the gene tree root being at position x. As described in the *SI Appendix*, we use results from Doyon et al. (15) that allow us to improve the speed of the calculation introduced by Tofigh (14).

To include genes from out-group species in a computationally efficient manner, we extend the species phylogeny with an additional "virtual out-group" branch that always branches from above the root of the species tree. On gene trees, each clade of genes containing only genes from out-group species is represented by a single "out-gene" that maps to the out-group branch. Including such an out-group branch also extends the set of possible reconciliations with scenarios, such as transfer from outside and from unsampled and extinct species that are otherwise neglected (*SI Appendix*, Fig. S5). A virtual out-group is used in all of the calculations presented here, but "out-genes"—that is, genes from out-group species—are only included in gene phylogenies when indicated.

Inferring the Most Likely Chronologically Ordered Species Phylogeny. Assuming the evolution of gene families to be independent, the likelihood of S, together with $\mathcal{M}$, the set of origination probabilities and duplication, transfer, and loss rates, is given by the product over the likelihoods of all gene phylogenies $G \in \mathcal{G}$:

$$\mathcal{L}_{\mathrm{ODT}}(S,\mathcal{M}|\mathcal{G}) = \prod_{G\in\mathcal{G}} p(G|S,\mathcal{M}). \qquad [2]$$

During our calculations, we maximize this likelihood by proposing different S, with the parameters of $\mathcal{M}$ estimated from most likely (ML) reconciliations for each proposal.

The proposal of new S proceeds in two steps. In the initial topology exploration phase new topologies are proposed based on reconciled transfers. In the second and final step, local topology and time order changes are used to maximize the likelihood. For details, see *Materials and Methods*.

Robustness of the Inference to Reconstruction Errors. To assess the robustness of our approach, we performed simulations using realistic rates of duplications, transfers, and losses (Fig. 2 and *Materials and Methods*). To approximate as closely as possible real ensembles of gene trees, we modeled branch-wise variation in duplication, transfer, and loss rates and allowed origination outside the species tree. As shown in Fig. 2*A*, the number of events found per branch in maximum likelihood reconciliations closely matches the number of simulated events.

An important caveat of our approach stems from considering a fixed gene phylogeny for each homologous gene family. Gene trees reconstructed from the multiple sequence alignment of a homologous family inevitably contain phylogenetic relationships that are not supported statistically. These relationships correspond to an arbitrary choice from among a set of relationships supported by the alignment and hence are a source of spurious disagreement between the gene and species phylogeny that lead to an excess of events in reconciliations. The presence of spurious discord resulting from reconstruction errors is problematic, as it has the potential to confound species phylogeny inference by introducing signal for nonexistent transfer events. We therefore performed simulations (Fig. 2 *B* and *C* and *SI Appendix*, Fig. S8) that mimicked random reconstruction errors in gene trees.

We compared these results to the amount of statistically unsupported discord found in real data. We used PRUNIER (16) to estimate the number of transfers among 474 near universal single copy gene families from 36 cyanobacterial genomes. PRUNIER performs statistical reconciliation taking into account branch supports and is capable of recovering the number of transfers for single copy families given a threshold of statistical support. Similar to previous results (13), we find that a significant fraction of apparent transfers are not supported statistically (*SI Appendix*, Table S1): for the moderate support threshold of 0.5, ≈6%, or on average 1.5 transfers per tree, are found to be spurious, whereas this number increases to 66% for the higher support threshold of 0.8, corresponding to 6.3 spurious transfers on average. These events can be attributed to local uncertainties in the gene trees, which are best modeled with random nearest neighbor interchanges (NNIs) (17). Equating the numbers of spurious transfers per gene in the PRUNIER dataset and the simulation results shown in Fig. 2*B*, 1.5 and 6.3 spurious transfers correspond to, respectively, 2.3 and 11.5 random NNIs (*SI Appendix*, Fig. S8). Consequently, we find that even in the presence of comparatively

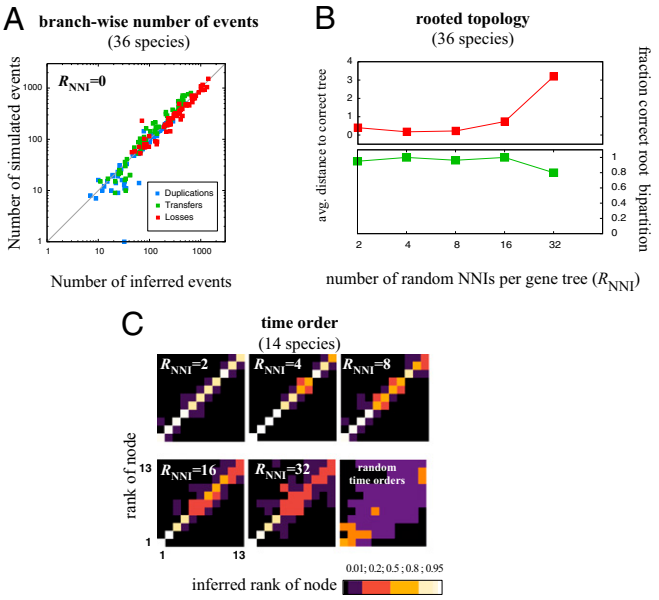

Fig. 2. Inference robustness on simulations. We performed simulations with realistic parameters including branch-wise rate variations, origination outside of the root, and independent reconstruction errors. *A* compares the reconstructed number of branch-wise events for a fixed time order for gene trees with no reconstruction errors. *B* shows the accuracy of recovering a rooted species phylogeny measured as the average Robinson-Foulds distance for different R_{NNI} [number of random NNIs (17)]. Averages over at least 30 simulations are shown, except for R_{NNI} = 32 with eight simulations. *C* shows the accuracy of recovering the time order for a fixed topology and rooting. The color scale gives the fraction of times a rank was inferred for each node in the known tree. Squares along the diagonal indicate the number of times the correct time order was recovered. For each value of R_{NNI} the average over 100 simulations is shown. The number of species in *S* was reduced to 14 for computational tractability (making the R_{NNI} perturbation more severe).

large numbers of simulated reconstruction errors per gene tree, we are able to accurately recover the correct chronologically ordered phylogeny (Fig. 2 *B* and *C*). To further establish the robustness of the inference, we also examined the effect of systematic reconstruction errors (such as long branch attraction) and systematic deviations from the species tree such as highways of gene transfer (18). As shown in *SI Appendix*, Fig. S9, if the fraction of gene families experiencing systematic bias is below 40%, the inference accurately recovers the topology and the rooting.

We then inferred the chronologically ordered phylogenies for 10 prokaryotic phyla from the dataset in ref. 19. Comparing the reconstructed species tree topologies to unrooted trees obtained by established methods (Table 1), we find that our results are generally very similar to phylogenomics methods that rely on single copy genes but differ more markedly from 16S rRNA phylogenies. We are also able to recover the root obtained using concatenates of single copy genes that include genes from the out-group without considering any out-group sequences for most phyla. A notable exception are the Cyanobacteria, for which we do not recover the out-group root even when we include out-group information in our dataset in the form of leaves representing genes from out-group species.

Chronologically Ordered Phylogeny of Cyanobacteria. We reconstructed the chronologically ordered phylogeny of Cyanobacteria using a dataset consisting of 77,678 genes represented by the gene phylogenies of 8,332 homologous gene families for 36 cyanobacterial genomes from the HOGENOM database (20). The maximum likelihood tree obtained is shown in Fig. 3*A*. The topology of the reconstructed tree is consistent with previous phylogenomics results (21, 22) and is identical to the unrooted topology obtained for the concatenate of 474 near universal single copy genes (PHYML, LG+Γ8+I) and the consensus tree of 364 universal single copy trees (using both PHYML, LG+Γ8+I and TreeFinder, WAG+Γ8+I trees). The topology differs from two recent studies with different (23) or larger (24) species sampling than ours: in ref. 23 the positions of *Synechocystis* and *Thermosynechococcus* and in ref. 24 the positions of *Trichodesmium* and *Synechocystis* are different, in both cases being moderately supported and also differing between each other.

Phylogenetic studies relying on out-group sequences unanimously agree on an early branching of *Gloeobacter violaceus* at the root of the Cyanobacteria (e.g., refs. 19, 21, and 24). This rooting, however, is sensitive to the choice of out-group species (25). In contrast to these results, we find *Gloeobacter violaceus* emerging relatively late (node 16 in Fig. 3*A*). To validate our rooting, we performed two tests summarized in Table 2: first, we tested the statistical significance of the three root positions with the highest likelihood and the *Gloeobacter* root accounting for several factors, such as the effect of the presence of out-group species in gene trees and the degree of spurious phylogenetic conflict (*Materials and Methods*); second, as independent validation, we compared the number of transfers inferred by PRUNIER in 474 near universal single copy gene families using these four different positions of the root. Rows correspond to the four root positions indicated by the coloring in Fig. 3*A*, and root positions that could not be rejected are marked in bold. Although the support for the ML root is only statistically significant with the ODT analysis in the absence of out-genes, all tests strongly reject the early emergence of *Gloeobacter*. Consistent with this observation, reconciliation scenarios show that the late emergence of the "violet" clade, comprised of *Gloeobacter* and two species of cyanobacteria recovered from Yellowstone hot springs (*Synechococcus* sp. JA-3–3AB and JA-2–3B′A), appears to be supported by a large number of transfers from the species of the "green" clade (Fig. 3*C*). The presence of spurious transfers associated with reconstruction errors make it difficult to interpret individual ODT reconciliations, but a late emergence of *Gloeobacter violaceus* is also supported by PRUNIER, which only reports well-supported transfer events (Table 2 and Table S1 in the *SI Appendix*). A late emerging violet clade is consistent with the inclusion of *Synechococcus* sp. JA-2–3B′A and JA-3–3AB in the order Gloeobacterales based on sequences from a recently cultured early branching cyanobacteria (26).

Table 1. Inferences for 10 prokaryotic phyla

		Distance*			Rooting†	
Phylum	#sp.	16S	Cc.	Cs.		Out-genes
Bacillales	16	7	3	2	Same	—
Bacteroidetes/Chlorobi	10	1	0	0	Same	—
Chlamydiae/Verrucomicrobia	7	0	0	0	Same	—
Crenarcheota	10	3	1	1	Same	—
Cyanobacteria	14	1	1	0	Far	Far
δ-Proteobacteria	13	3	1	1	Near	Same
ε-Proteobacteria	7	0	0	0	Near	Same
Euryarcheota	25	8	1	3	Other	Same
Lactobacillales	21	6	4	3	Same	—
Mollicutes	14	2	2	2	Same	—

#sp., number of genomes.
*Robinson-Foulds distance to the ML tree obtained from 16S rRNA (16S), concatenate (Cc.), and the consensus tree ML trees of universal families (Cs.).
†"Same" indicates a root on the same bipartition as (19), "near" and "far" indicate neighboring or more distant bipartition, and "other" that the root bipartition in (19) was not present.

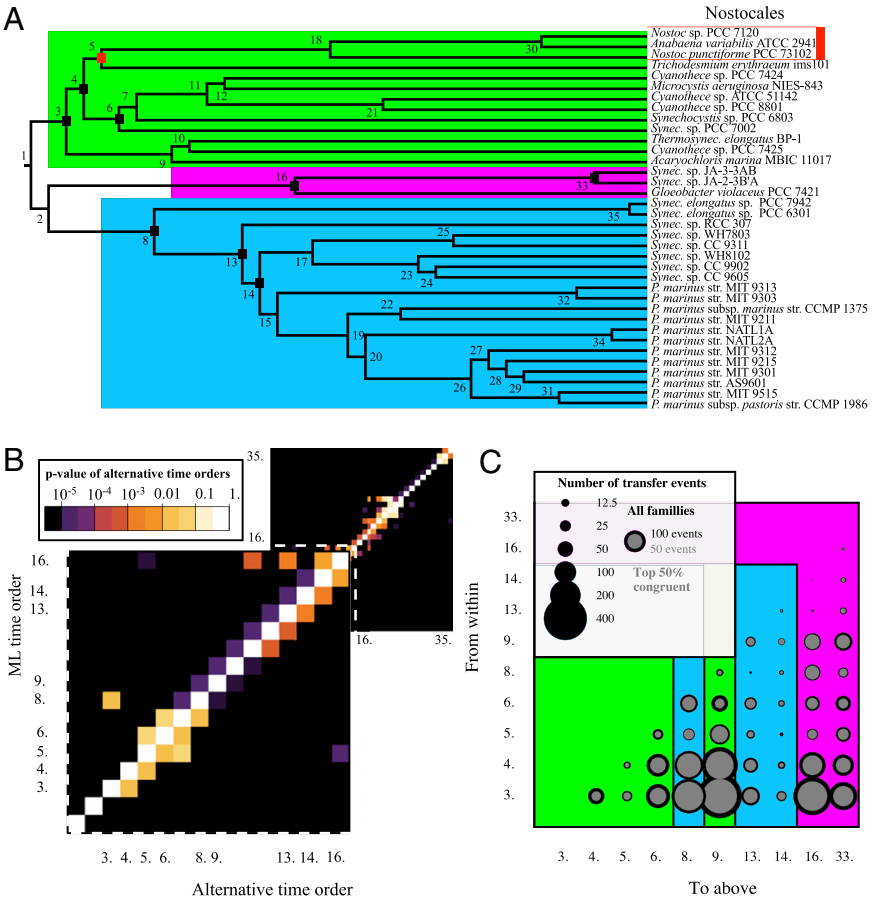

Fig. 3. Chronologically ordered phylogeny reconstructed from 36 cyanobacterial genomes. (*A*) The maximum likelihood time orders are indicated as node labels. Squares correspond to major diversification events discussed in the text. The tree shown was calculated using a uniform model; using a branch-specific DTL rate model gives an identical topology and very similar time orders (*SI Appendix*, Fig. S10). Root positions discussed in the text and Table 2 are indicated by the coloring: green, node 3; blue, node 8; violet, node 16 rooting the tree. The green rooting is shown. Branch lengths are derived from the time order assuming time slices of equal width. (*B*) The *P* value of alternative time orders were calculated using 328 candidate time orders corresponding to all deep time order moves around the ML solution (combined AU test with all families and 50% most congruent). (*C*) The number of transfers supporting different time orders in the ML solution are shown for the nodes discussed in the text. The area of each bubble is proportional to the number of transfers from branches descending from a speciation to above a speciation occurring later in time. In general a transfer belongs to multiple fields; for example, some of the transfers from within 4 to above 16 may be from within 5 to above 16.

Using maximum likelihood reconciliations averaged over gene tree roots and origination positions, we also estimated the number of genes in ancestral genomes of cyanobacteria (Fig. 4). The reconstructed ancestral gene numbers indicate the ancestor of Cyanobacteria had relatively few genes. The number of genes, however, expanded early in the diversification of the phylum in the lineage leading to node 3, from which all extant Cyanobacteria with a larger number of genes (3,000–4,000 genes) descend, while remaining relatively constant for its sister lineage, the descendants of which generally have had a smaller number of genes (1,200–2,200 genes). The increase in gene number is predominantly due to origination with ~13% of all gene families originating in the lineage leading to node 3, with a second burst of origination involving ~15% of all gene families occurring in the lineage leading to Nostocales (node 5).

Discussion

We present a method to reconstruct a chronologically ordered tree of species by explicitly modeling the evolution of the genes present in their genomes. Using simulations, we show that our method is robust to realistic levels of reconstruction errors and sources of systematic bias. Examining a diverse set of prokaryotic phyla, we find that the topology of the reconstructed species tree is very similar to results obtained with classic phylogenomics methods, consistent with the observation (13, 19, 27) that concatenate and supertree approaches are able to extract a robust signal even in the presence of substantial LGT. However, aggregate approaches may be sensitive to biased LGT (12), a caveat which is especially relevant in light of recent empirical evidence of strong habitat drivers in LGT (28). In contrast, simulations indicate that our approach of explicitly modeling duplication, transfer, and loss events is robust to significant amounts of systematic bias (*SI Appendix*, Fig. S12). We also neglect the fact that multiple mechanisms (e.g., being part of the same operon) can cause genes to be duplicated, transferred, or lost in a coordinated fashion. Such coevolution is certainly of great interest for individual gene histories, but similar to systematic bias, we can expect genome scale reconstructions, such as ours, to be robust to its presence.

To validate the ability of our approach to reconstruct the chronological order of speciation events, we reconstructed the phylogenetic history of 36 cyanobacterial genomes. Cyanobacteria are distinguished among prokaryotes by a relative abundance of microfossil and biomarker records (29). However, even this relative abundance reduces to a handful calibration points for the timing of cyanobacterial evolution: (*i*) the presence of free oxygen in late Archean oceans (30) and the rapid rise in atmospheric oxygen 2.45–2.32 billion years ago (Ga) (31) is attributed to cyanobacteria (32), but many details of these associations continue to be investigated (33, 34); (*ii*) unambiguous microfossil evidence dates to ~2.0 Ga, by which time diversification was advanced with fossils of filamentous forms dating from 1.9 Ga (35) and akinete forming Cyanobacteria from 2.1 Ga (29); and (*iii*) by 1.4 Ga diversification of major extant morphologically distinguishable lineages was complete (22, 36).

Comparing the reconstructed relative chronology with two recent molecular dating studies by Falcon et al. (23) and Blank et al. (22), we find good agreement for early chronological relationships: both find that node 4, the ancestor of nodes 5 (clade containing *Trichodesmium* and Nostocales) and 6 (clade containing *Synchocystis* and *Microcystis*) is the first major group to emerge; both also support the order of nodes 5, 6, and 8, as well as the late diversification of the *Synechococcus-Prochlorococcus* group (nodes 13–15 and 17); this correspondence with our results is remarkable given that both studies rely on a relaxed molecular clock and calibration constraints based on the geological and

Table 2. Support for alternative hypotheses for the root of cyanobacteria

Root	PRUNIER* (No out-genes)	ODT† (No out-genes)	ODT† (Out-genes)
Green	**1,893**	**1st**	**1st**
Blue	**1,893**	2nd	**3rd**
Violet	**1,905**	3rd	**2nd**
Gloeobacter	1,915	4th	4th

*Number of transfers per family for 474 near universal single copy families; bold roots could not be rejected with $P < 0.001$.
†The order of root positions according to likelihood is given; bold roots could not be rejected with $P < 0.001$.

fossil records (three calibration points in ref. 23 and five in ref. 22), whereas our model uses neither.

Direct comparison with the fossil record is only possible for node 5, the speciation leading to all akinete-forming Nostocales. Node 5 occurs at the earliest position in the time order considering topology and the time order of its ancestor and has a time order between 5 and 7 with statistical support (Fig. 3*B*) based on a large number of transfer events (Fig. 3*C*). Placing the evolution of genome size in this timeline, the large increase in gene number in the lineage leading to node 3 is placed in the middle Archaean, suggesting a parallel with the results of an episode of rapid evolutionary innovation during the middle Archaean (37).

In the future, it should be possible to provide direct date estimates by using the results of the ODT inference to inform a relaxed molecular clock analysis. This can be accomplished by using the ODT inference to provide a set of relative time constraints complementary to any molecular fossil calibrations. Relaxed molecular clock analyses incorporating these combined constraints have the potential to significantly better resolve our picture of the timing of prokaryotic evolution. The list of constraints obtained from the cyanobacterial dataset can be found in *SI Appendix*, Table S3.

Probabilistic models of genome evolution that consider information from complete genomes are important as they lay the foundations for the parallel reconstruction of the relative chronology of the diversification and the oddities of individual gene histories. Here we have only considered large-scale observables that are robust to uncertainties in the reconstructed gene histories, such as time order and the number of genes in ancestral genomes. To go further, we must reduce the amount of phylogenetic reconstruction error that limits the accuracy of reconciliations. This caveat is especially pertinent in the case of systematic reconstruction biases (resulting from, e.g., compositional bias and long branch attraction), which are very difficult to handle for individual families and particularly serious in deep phylogenies, where we know full well that our models are far from the truth. A path to a potential solution lies in integrating probabilistic models, such as the ODT model presented here, with traditional phylogenetic methods to refine gene trees based on reconciliation with the species tree (9, 38). For fixed species trees with models using only duplications and loss, but not transfer, this has already been demonstrated (39). Our results show that extending such integrative methods to include species tree inference and to model LGT has the potential to access information on the evolution of prokaryotic genomes that is at present overlooked.

Materials and Methods

Extending Possible Reconciliations Using a Virtual Out-Group. We extended the species tree *S* with a virtual out-group species that branches above the root regardless of any other changes to its topology and in which all "out-genes" reside. Such a virtual out-group permits reconciliations that allow us to consider in an approximate manner (*i*) genes from out-group genomes, (*ii*) transfer from extinct lineages, and (*iii*) transfer from outside (*SI Appendix*, Fig. S5).

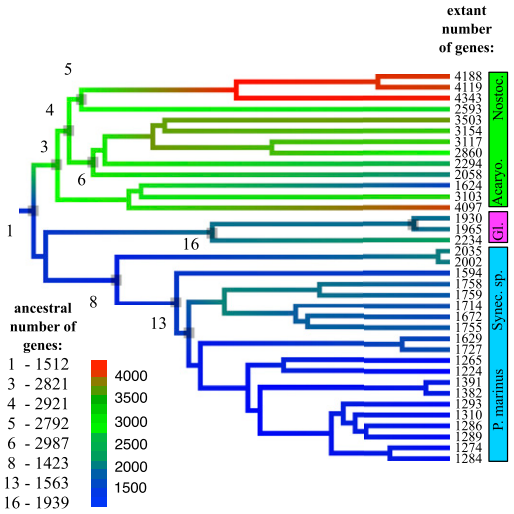

Fig. 4. Number of genes in ancestral genomes. Color scale shows the number of genes in ancestral genomes on the tree presented in Fig. 3*A*. Estimates were obtained by averaging maximum likelihood reconciliations over gene tree roots and origination positions and compensating for extinct gene lineages. Squares correspond to major diversification events discussed in the text. Color bars show correspondence with species names in Fig. 3*A*.

Exploration of the Species Phylogeny and Time Orders. In the initial phase of the ML exploration, we aim to efficiently propose and evaluate new topologies before a more detailed and precise search is undertaken. We count the number of transfers between all pairs of branches in *S* that share a time slice and attempt changes to the species tree such that they resolve the highest number of transfers as speciations (*SI Appendix*, Fig. S6). Following this initial search, we proceed by trying local topology rearrangements and local time order rearrangements until no topology or time order move is found that improves the likelihood. Local topology rearrangements correspond to all NNIs. Local time order moves are achieved by either exchanging the time order of a node with a node that has an adjacent time order (shallow moves, e.g., exchanging in *SI Appendix*, Fig. S1 node 3 with nodes 2 or 4) or alternatively moving any node such that the resulting time orders remain compatible with the rooted tree topology (deep moves, e.g., changing the time order in *SI Appendix*, Fig. S1 of node 8 such that it has an order 3, 5, 6, 7). The likelihood calculations and the general reconciliation algorithm were implemented in a parallel framework using MPI that relies on components from the Bio++ (40), Boost (41), and BLAS (42) libraries. See also *SI Appendix*, section S2.

Estimating Parameters. We estimate observed rates recursively using ML reconciliations. In the case of the uniform model, the mean of observed branch-wise rates are used, whereas for the branch-specific DTL rate model, categories are derived from a gamma distribution parametrized by the branch-wise mean and variance of the corresponding observed rate. The number of rate categories was chosen using a Bayesian Information Criterion. For origination probabilities, we estimate the full set branch-wise origination probabilities using the sum over all reconciliations. See also *SI Appendix*, sections S2.3–S2.5.

Datasets. For the results presented in Table 1, we extracted all families with trees from version 4 of the HOGENOM database (20) for 10 prokaryotic phyla using the species selection of ref. 19. We retained all families with at least two genes in the set of species considered. In calculations involving out-genes, clades of genes from out-group phyla were replaced with a single virtual out-gene mapping to the virtual out-group. In the case of the 14 cyanobacteria, eukaryotic genes were neglected. We constructed a second independent dataset representing all 36 cyanobacterial genomes found in version 5 of the HOGENOM database comprised of 8,332 families with 77,678 genes (90% of all families with at least two genes). Sequences were extracted for each family including genes from actinobacteria and chloroflexi, aligned using MUSCLE (43), and cleaned using GBLOCKS (44). Alignments with less than 75 sites were discarded, and trees were inferred using PhyMLwith LG+Γ8+I model (45). Clades of genes from out-group phyla were

EVOLUTION

replaced with a single virtual out-gene mapping to the virtual out-group present above the root of the species tree described above. We also extracted 474 near universal single copy families, each present in at least 34 genomes. We separately aligned and cleaned these families, considering only cyanobacterial sequences. See also *SI Appendix*, section S3. The results and the data used in the PRUNIER analysis are available from: ftp://pbil.univ-lyon1.fr/pub/datasets/DAUBIN/SZOLLOSI2012/. Inferring ML phylogenies with, for example, PhyML requires up to several hours for individual gene families and a few days for the larger concatenates, using individual processors The calculations presented in Fig. 3 require several days of calculation on 80 processors.

Simulations. We simulated gene tree ensembles with rates similar to those inferred for the cyanobacterial datasets. We modeled branch-wise rate variation by assigning species tree branches randomly from an exponential distribution. We chose a mean duplication and transfer rate of $\Delta_{mean} = T_{mean} = 0.2 \pm 0.2$ and a mean loss rate of $\Lambda_{mean} = 0.6 \pm 0.6$. In comparison, the mean and the standard variation of observed branch-wise rates were $\Delta_{mean} = 0.20 \pm 0.18$, $T_{mean} = 0.31 \pm 0.18$, and $\Lambda_{mean} = 0.39 \pm 0.37$ for the smaller dataset with 14 genomes and $\Delta_{mean} = 0.20 \pm 0.21$, $T_{mean} = 0.17 \pm 0.08$, and $\Lambda_{mean} = 1.04 \pm 0.64$ for the larger data set with 36 genomes. We simulated gene trees by choosing a random origination point along *S*, with 50% of families originating above the root of *S*. We modeled reconstruction errors by introducing a number of random NNIs drawn from a Poisson distribution with mean $R_{NNI} = 2\ldots32$. Fig. 2*A* shows the number of simulated and reconstructed events for 8,000 gene trees. Data points and heat maps in Fig. 2 *B* and *C* are averages over independent simulation runs with 800 simulated gene families per run. See also *SI Appendix*, sections S2.6 and S2.7.

Statistical Analysis. Statistical analyses were performed with Mathematica and the CONSEL software package (46). In Table 2, for each root, we performed a sign test against the blue and green roots (tied) with the minimum number of transfers. For both ODT rows for each root, two approximately unbiased (AU) tests were performed using the likelihood of families given the ML time orders conditional on the root. The first test used the likelihood of the complete set of gene trees, and the second used a subset of these corresponding to 50% of trees with the highest congruence. A root was rejected if it had a P value < 0.001 according to both tests. In Fig. 3*B*, an AU test was performed considering all 328 time orders corresponding to all deep moves around the ML solution for the set of all trees and 50% most congruent, with the larger P value displayed. The minimal set of constraints representing 29 time orders with $P > 0.001$ can be found in *SI Appendix*, Table S3. We choose the 50% most congruent trees, acting on the hypothesis that gene trees that are more congruent with the species tree contain less reconstruction errors. We defined congruence for maximum likelihood reconciliations in a root-independent manner, as the maximum overall roots of the fraction of events that are not a transfer or a loss to all events.

ACKNOWLEDGMENTS. We thank all participants in the "Phylariane" and "Ancestrome" projects. We also thank Imre Derényi for computing resources. G.J.Sz. is supported by the Marie Curie Fellowship 253642 Geneforest. B.B. is supported by a Human Frontier Science Program fellowship. This work was granted access to the computing centre of the French National Institute of Nuclear Physics and Particle Physics and to the resources of the Centre Informatique National de l'enseignement supérieur under Allocation 2010-076436. This project was supported by the French Agence Nationale de la Recherche (ANR) through Grants ANR-10-BINF-01-01 "Ancestrome" and ANR-08-EMER-011-03 "Phylariane".

1. Benton MJ, Ayala FJ (2003) Dating the tree of life. *Science* 300:1698–1700.
2. Peterson KJ, et al. (2004) Estimating metazoan divergence times with a molecular clock. *Proc Natl Acad Sci USA* 101:6536–6541.
3. Douzery EJP, Snell EA, Bapteste E, Delsuc F, Philippe H (2004) The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci USA* 101:15386–15391.
4. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88.
5. Knoll AH (2003) *Life on a Young Planet: The First Three Billion Years of Evolution on Earth* (Princeton Univ Press, Princeton, NJ).
6. Hedges SB, Kumar S (2009) *The Timetree of Life* (Oxford Univ Press, Oxford, UK).
7. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6:361–375.
8. Ciccarelli FD, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.
9. Boussau B, Daubin V (2010) Genomes as documents of evolutionary history. *Trends Ecol Evol* 25:224–232.
10. Dagan T, Martin W (2006) The tree of one percent. *Genome Biol* 7:118.
11. Doolittle WF, Bapteste E (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci USA* 104:2043–2049.
12. Beiko RG, Doolittle WF, Charlebois RL (2008) The impact of reticulate evolution on genome phylogeny. *Syst Biol* 57:844–856.
13. Galtier N, Daubin V (2008) Dealing with incongruence in phylogenomic analyses. *Philos Trans R Soc Lond B Biol Sci* 363:4023–4029.
14. Tofigh A (2009) *Using Trees to Capture Reticulate Evolution Lateral Gene Transfers and Cancer Progression* (PhD thesis, The Royal Institute of Technology, Stockholm, Sweden).
15. Doyon JP, et al. (2010) An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. *Comparative Genomics, LNCS* 6398:93–108.
16. Abby SS, Tannier E, Gouy M, Daubin V (2010) Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics* 11:324.
17. Felsenstein J (2004) *Inferring Phylogenies* (Sinauer Associates, Sunderland, MA).
18. Beiko RG, Harlow TJ, Ragan MA (2005) Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA* 102:14332–14337.
19. Abby SS, Tannier E, Gouy M, Daubin V (2012) Lateral gene transfer as a support for the tree of life. *Proc Natl Acad Sci USA* 109:4962–4967.
20. Penel S, et al. (2009) Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* 10(Suppl 6):S3.
21. Swingley WD, Blankenship RE, Raymond J (2008) Integrating Markov clustering and molecular phylogenetics to reconstruct the cyanobacterial species tree from conserved protein families. *Mol Biol Evol* 25:643–654.
22. Blank CE, Sánchez-Baracaldo P (2010) Timing of morphological and ecological innovations in the cyanobacteria—A key to understanding the rise in atmospheric oxygen. *Geobiology* 8:1–23.
23. Falcón LI, Magallón S, Castillo A (2010) Dating the cyanobacterial ancestor of the chloroplast. *ISME J* 4:777–783.
24. Criscuolo A, Gribaldo S (2011) Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. *Mol Biol Evol* 28:3019–3032.
25. Schirrmeister BE, Antonelli A, Bagheri HC (2011) The origin of multicellularity in cyanobacteria. *BMC Evol Biol* 11:45.
26. Couradeau E, et al. (2012) An early-branching microbialite cyanobacterium forms intracellular carbonates. *Science* 336:459–462.
27. Galtier N (2007) A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst Biol* 56:633–642.
28. Smillie CS, et al. (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480:241–244.
29. 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.
30. Waldbauer JR, Newman DK, Summons RE (2011) Microaerobic steroid biosynthesis and the molecular fossil record of Archean life. *Proc Natl Acad Sci USA* 108:13409–13414.
31. Bekker A, et al. (2004) Dating the rise of atmospheric oxygen. *Nature* 427:117–120.
32. Herrero A, Flores E (2008) *The Cyanobacteria: Molecular Biology, Genomics, and Evolution* (Caister Academic Press, Norfolk, UK).
33. Lyons TW, Reinhard CT (2011) Earth science: Sea change for the rise of oxygen. *Nature* 478:194–195.
34. Gaillard F, Scaillet B, Arndt NT (2011) Atmospheric oxygenation caused by a change in volcanic degassing pressure. *Nature* 478:229–232.
35. Golubic S, Hofmann HJ (1976) Comparison of holocene and mid-precambrian entophysalidaceae (cyanophyta) in stromatolitic algal mats: Cell division and degradation. *J Paleontol* 50:1074–1082.
36. Golubic S, Seong-Joo L (1999) Early cyanobacterial fossil record: Preservation, palaeoenvironments and identification. *Eur J Phycol* 34:339–348.
37. David LA, Alm EJ (2011) Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* 469:93–96.
38. Szöllősi GJ, Daubin V (2012) Modeling gene family evolution and reconciling phylogenetic discord. *Methods Mol Biol* 856:29–51.
39. 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.
40. Dutheil J, et al. (2006) Bio++: A set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics* 7:188.
41. Dave Abrahams (2011) BOOST, Peer-reviewed portable c++ source libraries. (The Boost Community) Version 1.48.
42. Blackford LS, et al. (2002) An Updated Set of Basic Linear Algebra Subprograms (BLAS). ACM Trans. Math. Soft., 28-2:135–151.
43. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
44. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 56:564–577.
45. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
46. Shimodaira H, Hasegawa M (2001) CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247.