Published in final edited form as:
Nat Ecol Evol. 2018 May ; 2(5): 904–909. doi:10.1038/s41559-018-0525-3.

Gene transfers can date the Tree of Life

Adrián A. Davín[1], **Eric Tannier**[1,2], **Tom A. Williams**[3], **Bastien Boussau**[1], **Vincent Daubin**[1,*], and **Gergely J. Szöllősi**[4,5,*]

[1]Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558, F-69622 Villeurbanne, France

[2]Inria Grenoble Rhône-Alpes, F-38334 Montbonnot, France

[3]School of Biological Sciences, University of Bristol, Life Sciences Building, 24 Tyndall Avenue, Bristol, BS8 1TQ, UK

[4]MTA-ELTE "Lendulet" Evolutionary Genomics Research Group, Budapest, Hungary

[5]Department of Biological Physics, Eotvos Lorand University, Budapest, Hungary

Abstract

Biodiversity has always been predominantly microbial and the scarcity of fossils from Bacteria, Archaea and microbial Eukaryotes has prevented a comprehensive dating of the tree of life. Here we show that patterns of lateral gene transfer deduced from the analysis of modern genomes encode a novel and abundant source of information about the temporal coexistence of lineages throughout the history of life. We use state-of-the-art species tree aware phylogenetic methods to reconstruct the history of thousands of gene families and demonstrate that dates implied by gene transfers are consistent with estimates from relaxed molecular clocks in Bacteria, Archaea and Eukaryotes. We present the order of speciations according to LGT calibrated to geological time for three datasets comprised of 40 genomes for Cyanobacteria, 60 genomes for Archaea and 60 genomes for Fungi. An inspection of discrepancies between transfers and clocks and a comparison with mammal fossils show that gene transfer in microbes is potentially as informative for dating the tree of life as the geological record in macroorganisms.

Until Zuckerkandl and Pauling put forth the “molecular clock”1 hypothesis, the geological record alone provided the timescale for evolutionary history. Their demonstration that distances between amino acid sequences correlate with divergence times estimated from fossils demonstrated that information in DNA can be used to date the Tree of Life. Since

* Vincent.Daubin@univ-lyon1.fr; ssolo@elte.hu.

Author Contributions

E.T., BB, V.D. and Sz.G. conceived of the study. A.D., B.B., E.T. and Sz.G. developed computational tools, A.D., E.T., B.B, V.D. and Sz.G. analyzed data, T.W. and V.D. contributed datasets, A.D., E.T., BB, T.W., V.D. and Sz.G. wrote the manuscript.

Competing Interests

The authors declare no competing interests.

Data Availability

All data used in the study is available used in the study is available in the Supplementary Material or can be downloaded from: ftp://pbil.univ-lyon1.fr/pub/datasets/davin2017/

Europe PMC Funders Author Manuscripts

then, the theory and methodology of the molecular clock have been developed extensively, and inferences from clock analyses (such as the diversification of placentals before the demise of dinosaurs2,3) hotly debated. Despite these controversies, combining information from rocks and clocks is now widely accepted to be indispensable3,4,5: state-of-the-art estimates of divergence times rely on sequence based relaxed molecular clocks anchored by multiple fossil calibrations. This approach provides information on both the absolute timescale and the relative variation of the evolutionary rates across the phylogeny (Fig.1a). Yet, because most life is microbial, and most microbes do not leave discernable fossils, major uncertainties remain about the ages of microbial groups and the timing of some of the earliest and most important events in life's evolutionary history6,7.

In addition to leaving only a faint trail in the geological record, the evolution of microbial life has also left a tangled phylogenetic signal due to extensive lateral gene transfer (LGT). LGT, the acquisition of genetic material potentially from distant relatives, has long been considered an obstacle for reconstructing the history of life8, because different genetic markers can yield conflicting estimates of the species phylogeny. However, it has been previously shown that transfers identified using appropriate phylogenetic methods carry information that can be harnessed to reconstruct species history9–14. This is possible because different hypotheses of species relationships yield different LGT scenarios and can thus be evaluated using phylogenetic models of genome evolution15–19. But in addition to carrying information about the relationships among species, transfers should also carry a record of the timing of species diversification because they have occurred between species that existed at the same time20–22. As a consequence, a transfer event can be used to establish a relative age constraint between nodes in a phylogeny independently of any molecular clock hypothesis: the ancestor node of the donor lineage must predate the descendant node of the receiving lineage (Fig.1b, Fig.S8). Below we show that the dating information carried by transfers is consistent with molecular clock based estimates of relative divergence times in representative groups from the three domains of life.

Results

We examined genome-scale datasets consisting of homologous gene families from complete genomes in Cyanobacteria (40 genomes27), Archaea (60 genomes28) and Fungi (60 genomes29). For each gene family we used the species tree-aware probabilistic gene tree inference method *ALE undated*27,30 to sample evolutionary scenarios involving events of duplication, transfer and loss of genes conditional on a rooted, but undated species phylogeny and the multiple sequence alignment of the family. We recorded the donor and recipient for each transfer, using the frequency with which that transfer was observed in the entire sample to score support. We then used the newly developed optimization method MaxTiC26 (**m**aximal **t**ime **c**onsistency, see Methods also supplementary text) to extract a maximal subset of consistent transfers that specifies a time order of speciation events in the species tree. We found that the maximal subset of transfers implies a time order of speciations that correlates with the distance between amino acid sequences of extant organisms (Spearman's $\rho = 0.741$; $p < 10^{-6}$; Fig. 1d, S9). A similar correlation (Fig. 1c) can be observed if, following Zuckerkandl and Pauling1, we compare fossil dates and sequence

divergence in mammals2 (10 time points, Pearson's R^2=0.664; p = 0.0025 and Spearman's ρ = 0.83; p = 0.0056).

We observed a strong correlation between time estimates from MaxTiC and molecular clocks in all our datasets ($p<10^{-3}$ - Fig S14-S16). This suggests that LGT indeed carries information on the relative age of nodes in all three domains of life. However, it is not conclusive because part of the correlation trivially results from the fact that parent nodes are necessarily both older and more distant to extant sequences than their direct descendants31. To control for this effect, we compared the relative time orders of speciation events inferred from transfers to dates obtained using molecular clocks in the absence of calibrations. As a control for the shape of the tree, we measured the random expectation by sampling chronograms from the prior on divergence times but keeping the species phylogeny fixed (without any sequence information). To compare the dating information from transfers to the information conveyed by fossils, we used the same uncalibrated approach on the same mammalian dataset as above2,34 and derived relative node age constraints from fossil calibrations (see supplementary text). For the prokaryotic and fungi datasets, we derived relative node age constraints from the maximal consistent subsets of transfers obtained using MaxTiC26. For both fossil- and transfer-based constraints, we then measured the fraction of constraints that are in agreement with each chronogram. As Fig. 2 shows, both fossil- and transfer-based constraints agree with uncalibrated molecular clocks significantly more than expected by chance. The observed agreement is robust to the choice of different clock models (Fig. 2), priors on divergence time and models of protein evolution (Figs. S17-S19). This result demonstrates the presence of genuine and substantial dating signal in gene transfers.

Interestingly, molecular clock models show differences in their agreement with relative time constraints. As expected, the strict molecular clock model generally explores a narrow range of dated trees compared to relaxed clocks. However, on average, chronograms based on the strict molecular clock agree less with relative time constraints than those based on relaxed clock models. This is particularly clear in mammals, where the median fraction of satisfied constraints falls within the 95% confidence interval of the random control (Fig. 2a). This is caused, in large part, by the accelerated evolutionary rate in rodents being interpreted (in the absence of fossil calibrations) as evidence for an age older than that implied by fossils (Fig. S4). The lognormal model is best suited to recover such autocorrelated (*e.g.* clade specific) rate variations along the tree, and indeed exhibits a median of 100% agreement with fossil based relative age constraints. The uncorrelated gamma model performs second best, perhaps because it is, in fact, autocorrelated along each branch34. Consistent with this idea, the completely uncorrelated white noise model fares worst (Fig. 2a-d). This is in agreement with previous model comparisons in eukaryotes, vertebrates and mammals34. A similar pattern is apparent when considering LGT-derived relative age constraints in Cyanobacteria, Archaea and Fungi, suggesting strong autocorrelated variation of evolutionary rates in these groups that are best recovered by the lognormal model (Fig. 2b-d).

The motivating principle of the MaxTiC algorithm is that transfers from the maximum consistent set carry a robust and genuine dating signal, while conflicting transfers are likely artefactual. Two lines of evidence suggest that this is indeed the case: first, the agreement of

relative time constraints derived from transfers excluded by MaxTiC with the node ranking inferred by uncalibrated molecular clocks tends to be lower than random (Fig. S12). Second, while the average sequence divergences for donor clades tend to be higher than for corresponding recipient clades in the set of self-consistent transfers ($p < 10^{-8}$ one sided T-test for difference greater than zero, see Fig. 3.), they are lower for those discarded by MaxTiC ($p < 10^{-8}$ one sided T-test for difference lower than zero, cf. Fig. 3).

One obvious difference between fossil- and transfer-based relative ages in Fig. 2 is that the level of agreement is patently lower for the latter. While in mammals approximately half of the chronograms proposed by the lognormal model agree with 100% of relative constraints, for other datasets no model reaches 80% agreement. This means that some relative constraints derived from LGT consistently disagree with uncalibrated molecular clock estimates. These disagreements are difficult to interpret because both molecular clocks and our transfer-based inferences may be subject to error; simulations suggest that spurious gene transfer inferences do occur with ALE, albeit at a low rate (Chauve et al.35, Fig. S23). Nonetheless this low error rate on simulations suggests that at least some transfers contradicting the molecular clocks are genuine. This yields the exciting idea of a new source of dating information, independent of and complementary to the molecular clock.

To gain further insight into the robustness of these transfer-based estimates, we evaluated their statistical support from the data. Since MaxTiC yields a fully ordered species tree, the relative age constraints derived from its output are potentially overspecified and include constraints with relatively low statistical support. To ascertain the extent of overspecification, we evaluated the statistical support of relative constraints by taking random samples of 50% of gene families and reconstructing the corresponding MaxTiC 1000 times (Figs. S20-S22). We then counted the number of times a constraint was observed. In all datasets, a large majority of constraints were highly supported (found in at least 95% of the replicates) and among these, a significant number (between 20% and 32%) consistently disagreed with molecular clock estimates (see Table S2). These strongly-supported transfer-based constraints that disagree with the clocks could result from the inability of uncalibrated molecular clock estimates to recover the correct timing of speciations in groups with large variations in the substitution rate over time.

Specifically, LGTs provide strong support for the relatively recent emergence of the Prochlorococcus - Synechococcus clade in Cyanobacteria (blue clade in Fig. 4a, estimated age 0.86 Gya), irrespective of uncertainty in the root of Cyanobacteria (see Supp. Mat.). Although the Prochlorococcus - Synechococcus clade is inferred to be ancient by three of the four uncalibrated molecular clock models in our study, previous analyses using relaxed molecular clock methods with more extensive species sampling and several fossil calibrations, including fossils dating akinete forming cyanobacteria at up to 2.1 Gya36 (green in Fig. 4.a, estimated age 1.95 Gya) have consistently dated this clade as younger than most of the rest of cyanobacterial diversity37,38. Prochlorococcus have a known history of genome reduction and evolutionary rate acceleration23, which may lead to artifactually ancient age inferences under uncalibrated molecular clock models, as for rodents above. This demonstrates that relative time orders implied by LGT can, like fossils, provide a consistent dating signal that is independent of the rate of sequence evolution.

In Archaea, patterns of LGT suggest that several nodes within the Euryarchaeota including cluster 1 and 2 methanogens (blue and purple clades in Fig. 4 with estimated ages of 3.0 Gya and 2.8 Gya) are older than both the TACK+*Lokiarchaeum* clade (red clade in Fig. 4, the clade uniting Thaumarchaeota, Crenarchaeota, Aigarchaeota, Korarchaeota and *Lokiarchaeum,* estimated age 2.3 Gya) and the DPANN Archaea (grey in Fig. 4, a genomically diverse group with small cells and genomes, with reduced metabolism suggestive of symbiont or parasite lifestyles, estimated age 1.8 Gya). The relative antiquity of methanogens is consistent with evidence of biogenic methane at a very early stage of the geological record (~3.5 Gya39), and with another recent analysis that used a single LGT to place the origin of methanogens before the radiation of Cyanobacteria14. These relationships are not recovered by any of the molecular clock models, and suggest that LGT-derived constraints may be highly informative for future dating studies.

The relative order of appearance of archaeal energy metabolisms corresponds to increasing energy yield, with methanogenesis evolving before sulphate reduction, and the oxidative metabolisms of Thaumarchaeota and Haloarchaea evolving most recently. In addition, we find that *Ignicoccus hospitalis* branches before its obligate parasite *Nanoarchaeum* (cf. Fig. S2), despite the early divergence of the DPANN clade from other Archaea.

In Fungi, we recover LGTs that provide information on the order of some of the deepest splits. In particular, among crown groups, LGTs indicate that Zoopagomycota40 (blue in Fig. 4, estimated age 0.71 Gya) diverged earlier than Mucoromycotina, Basidiomycota and Ascomycota (purple, grey and green in Fig. 4, estimated ages 0.24 Gya, 0.64 Gya and 0.53 Gya respectively). Note that some inferred LGTs could result from processes such as hybridisation or allopolyploidisation, and that these processes contribute dating information that can be treated in the same way as LGTs. On a wider scale, between Eukaryotic groups, LGTs suggest that Amoebozoa (the outgroup, yellow in Fig. 4, estimated age 0.85 Gya) diversified earlier than Opisthokonta and Apusozoa (the ingroup). This indicates that LGTs could strongly reduce the uncertainty associated with the divergence of the major eukaryotic clades41.

Discussion

Our demonstration that clocks and transfers contain complementary and compatible dating signals casts the phylogenetic discord of LGT in a new light, and calls for the development of new methods to combine these two types of dating information. Relaxed molecular clock models are fitted in a Bayesian framework, but current MCMC proposal mechanisms can handle absolute, but not relative time constraints. Calibrating a molecular clock in a consistent probabilistic framework with both fossil-based and transfer-based time information will require modelling the effects of dependencies between separate partsof the tree, which current methods consider as independent. In the meantime, it is possible to partially take relative constraints intoaccount in a typical relaxed clock analysis by two means: first, when fossil calibrations are available for some nodes, by propagating theirminimum age to all nodes constrained by transfers to be older, and symmetrically by propagating their maximum age to all nodes constrained by transfers to be younger; second, by rejection sampling, i.e. discarding posterior samples that fall below a threshold level of

agreement with transfer-based constraints. These approaches however do not guarantee that all strongly supported relative constraints will be respected. To produce time-calibrated chronograms that respect all constraints (cf. Fig. 4), we used a heuristic approach that indirectly estimates the age of nodes that are incompatible with constraints by interpolating between nodes whose ages do not violate the constraints.

The geological record of microbial life is sparse, and its interpretation is fraught with difficulty. Our results show that there is abundant information in extant genomes on dating the Tree of Life waiting to be harvested from the reconstruction of genome evolution. This signal mostly contains information on the relative timing of diversification of groups that have exchanged genes through LGT, but we foresee several strategies to relate this relative timing to the broader history of life on Earth. First, gene transfers between bacteria and multicellular organisms that have left a trace in the fossil record will allow the propagation of absolute time calibrations to the microbial part of the Tree of Life42. Similarly, the signal of coevolution between hosts and their symbionts, such as in the gut microbiome of mammals43, could also be used to propagate absolute dating information from the host to the symbiont phylogeny. Finally, geochemistry can provide major constraints on early evolution44,45: for example, LGT events to the ancestors of bacteria capable of oxygenic photosynthesis, *i.e.* Oxyphotobacteria46, imply that the donor lineages must be older than the oxygenation of Earth's atmosphere at approximately 2.3 Gya44,45. Phylogenetic models of genome evolution have the potential to turn the phylogenetic discord caused by gene transfer into an invaluable source of information on dating the tree of life.

Methods

We considered genome-scale datasets of homologous gene families from complete genomes in Cyanobacteria (40 genomes27), Archaea (60 genomes28) and Fungi (60 genomes29). For each gene family we used the species tree-aware probabilistic gene tree inference method *ALE undated*27,30 to sample evolutionary scenarios involving events of duplication, transfer and loss of genes conditional on a rooted species phylogeny and the multiple sequence alignment of the family. The undated reconciliation method ignores tree branch lengths and does not impose any constraint on possible donor-recipient branch pairs aside from forbidding transfers to go from descendants to parents (Fig. S9). For putative gene transfer events, we recorded the donor and recipient branches and used the frequency with which they occurred among the sampled scenarios to filter transfers and weigh the relative age information they imply. Because the reference species tree is not dated, individual transfers can imply conflicting information about the relative age of speciation nodes (Fig. S11). To extract a maximal subset of transfers consistent with each other, we used the newly developed optimization method MaxTiC26 (**m**aximal **t**ime **c**onsistency, see also supplementary text). A maximal subset of consistent transfers specifies a time order of speciation events in the species tree. For instance, using MaxTiC on the 4816 transfers that correspond to relative age constraints (see Figs 1b, S8, S10) in the 5322 gene families considered for Cyanobacteria, we identified a maximal subset of 3322 (69%) transfers that are consistent (Table S1). This maximal subset of transfers implies a time order of speciations that correlates with the distance between amino acid sequences of extant organisms (Spearman's $\rho = 0.741$; $p < 10^{-6}$; Fig. 1d, S9). A similar correlation (Fig. 1c) can

be observed if, following Zuckerkandl and Pauling1, we compare fossil dates and sequence divergence in mammals2 (10 time points, Pearson's R^2=0.664; p = 0.0025 and Spearman's ρ = 0.83; p = 0.0056).

We used Phylobayes32 on a concatenate of nearly universal gene family alignments to sample chronograms (*i.e.*, dated trees) under four different uncalibrated molecular clock models33 (the strict molecular clock, the autocorrelated lognormal, the uncorrelated gamma and the white-noise model). Chronograms were sampled using different calibration schemes described in the Supplementary Materials and in the main text.

To estimate trees calibrated to geological time that obey transfer-based relative age constraints presented in Figure 4 we followed a three-step approach: First, for each dataset we sampled 5000 time orders compatible with LGT-based constraints obtained from MaxTiC. Second, for each dataset we sampled chronograms calibrated to geological time with fossil calibrations using Phylobayes as described above (see also supporting text and Table S4) and assigned to each node of the phylogeny a direct age estimate corresponding to the median of the node ages in chronograms with top 5% agreement with LGT-based constraints obtained from MaxTiC (cf. Figures S25-27). Finally, we calibrated each of the 5000 time orders to geological time by removing conflicting node age estimates until we obtained a set of node ages compatible with the time order. Nodes left without node age estimates were assigned an indirect age corresponding to a random date distributed uniformly between the nearest existing dates such that the time order was obeyed. For each sampled time order conflicting age estimates were removed in a fixed order corresponding to decreasing conflict calculated over all 5000 sampled time orders, so that the ages that conflicted with the largest number of time orders were removed first.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Sz.G. received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 714774. This project was supported by the French Agence Nationale de la Recherche (ANR) through grant no. ANR-10-BINF-01–01 'Ancestrome'. Computations were made using the Curie supercomputer thanks to PRACE project 2013081661 and the computing facilities of the CC LBBE/PRABI. T.A.W. is supported by a Royal Society University Research Fellowship. We thank N. Lartillot, T. Warnow, M. Paris, I. Derényi, L. Nagy and J. Miguel Blanca Postigo for useful discussions, comments on the manuscript and further computing resources.

References

1. Zuckerkandl, E., Pauling. Molecular Disease, Evolution, and Genic Heterogeneity. 1962.
2. dos Reis M, et al. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc Biol Sci. 2012; 279:3491–3500. [PubMed: 22628470]
3. O'Leary MA, et al. The placental mammal ancestor and the post-K-Pg radiation of placentals. Science. 2013; 339:662–667. [PubMed: 23393258]
4. Donoghue PCJ, Benton MJ. Rocks and clocks: calibrating the Tree of Life using fossils and molecules. Trends Ecol Evol. 2007; 22:424–431. [PubMed: 17573149]

5. Yang Z, Donoghue PCJ. Dating species divergences using rocks and clocks. Philos Trans R Soc Lond B Biol Sci. 2016; 371 20150126.
6. Knoll, AH. The Fossil Record of Microbial Life. Fundamentals of Geobiology. 2012. p. 297-314.
7. Knoll AH. Paleobiological perspectives on early eukaryotic evolution. Cold Spring Harb Perspect Biol. 2014; 6
8. Doolittle WF. Phylogenetic Classification and the Universal Tree. Science. 1999; 284:2124–2128. [PubMed: 10381871]
9. Abby SS, Tannier E, Gouy M, Daubin V. Lateral gene transfer as a support for the tree of life. Proceedings of the National Academy of Sciences. 2012; 109:4962–4967.
10. Szöllősi GJ, Boussau B, Abby SS, Tannier E, Daubin V. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. PNAS. 2012; 109:17513–17518. [PubMed: 23043116]
11. Williams TA, et al. Integrative modelling of gene and genome evolution roots the archaeal tree of life. PNAS. 2017
12. Huang J, Gogarten JP. Ancient gene transfer as a tool in phylogenetic reconstruction. Methods Mol Biol. 2009; 532:127–139. [PubMed: 19271182]
13. Huang J, Xu Y, Gogarten JP. The presence of a haloarchaeal type tyrosyl-tRNA synthetase marks the opisthokonts as monophyletic. Mol Biol Evol. 2005; 22:2142–2146. [PubMed: 16049196]
14. Wolfe JM, Fournier GP. Tunneling through time: Horizontal gene transfer constrains the timing of methanogen evolution. 2017; doi: 10.1101/129494
15. Maddison WP. Gene Trees in Species Trees. Syst Biol. 1997; 46:523.
16. Szöllősi, GJ., Daubin, V. Modeling Gene Family Evolution and Reconciling Phylogenetic Discord. Methods in Molecular Biology. 2012. p. 29-51.
17. Sjöstrand J, et al. A Bayesian method for analyzing lateral gene transfer. Syst Biol. 2014; 63:409–420. [PubMed: 24562812]
18. Szöllősi GJ, Tannier E, Daubin V, Boussau B. The Inference of Gene Trees with Species Trees. Syst Biol. 2015; 64(1):e42–e62. [PubMed: 25070970]
19. Daubin V, Szöllősi GJ. Horizontal Gene Transfer and the History of Life. Cold Spring Harb Perspect Biol. 2016; 8:a018036. [PubMed: 26801681]
20. Gogarten JP, Murphey RD, Olendzenski L. Horizontal gene transfer: pitfalls and promises. Biol Bull. 1999; 196:359–61. discussion 361-2. [PubMed: 10390834]
21. Szöllősi GJ, Boussau B, Abby SS, Tannier E, Daubin V. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. PNAS. 2012; 109:17513–17518. [PubMed: 23043116]
22. Szöllősi GJ, Tannier E, Lartillot N, Daubin V. Lateral Gene Transfer from the Dead. Syst Biol. 2013; 62:386–397. [PubMed: 23355531]
23. Dufresne A, Garczarek L, Partensky F. Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol. 2005; doi: 10.1186/gb-2005-6-2-r14
24. Benton MJ, Donoghue PCJ. Paleontological evidence to date the tree of life. Mol Biol Evol. 2007; 24:26–53. [PubMed: 17047029]
25. dos Reis M, Donoghue PCJ, Yang Z. Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet. 2016; 17:71–80. [PubMed: 26688196]
26. Chauve C, et al. MaxTiC: Fast Ranking Of A Phylogenetic Tree By Maximum Time Consistency With Lateral Gene Transfers. bioRxiv. 2017; doi: 10.1101/127548
27. Szöllősi GJ, Davín AA, Tannier E, Daubin V, Boussau B. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. Philos Trans R Soc Lond B Biol Sci. 2015; 370 20140335.
28. Williams TA, et al. Integrative modelling of gene and genome evolution roots the archaeal tree of life. PNAS. 2017
29. Nagy LG, et al. Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. Nat Commun. 2014; 5:4471. [PubMed: 25034666]
30. Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. Efficient exploration of the space of reconciled gene trees. Syst Biol. 2013; 62:901–912. [PubMed: 23925510]

31. Felsenstein J. Phylogenies and the Comparative Method. Am Nat. 1985; 125:1–15.
32. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004; 21:1095–1109. [PubMed: 15014145]
33. Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 2009; 25:2286–2288. [PubMed: 19535536]
34. Lepage T, Bryant D, Philippe H, Lartillot N. A General Comparison of Relaxed Molecular Clock Models. Mol Biol Evol. 2007; 24:2669–2680. [PubMed: 17890241]
35. Chauve C, et al. MaxTiC: Fast Ranking Of A Phylogenetic Tree By Maximum Time Consistency With Lateral Gene Transfers. 2017; doi: 10.1101/127548
36. Tomitani A, Knoll AH, Cavanaugh CM, Ohno T. The evolutionary diversification of cyanobacteria: Molecular-phylogenetic and paleontological perspectives. Proceedings of the National Academy of Sciences. 2006; 103:5442–5447.
37. Blank CE, Sánchez-baracaldo P. Timing of morphological and ecological innovations in the cyanobacteria--a key to understanding the rise in atmospheric oxygen. Geobiology. 2010; 8:1–23. [PubMed: 19863595]
38. Shih PM, Hemp J, Ward LM, Matzke NJ, Fischer WW. Crown group Oxyphotobacteria postdate the rise of oxygen. Geobiology. 2017; 15:19–29. [PubMed: 27392323]
39. Ueno Y, Yamada K, Yoshida N, Maruyama S, Isozaki Y. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. Nature. 2006; 440:516–519. [PubMed: 16554816]
40. Spatafora JW, et al. A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. Mycologia. 2016; 108:1028–1046. [PubMed: 27738200]
41. Eme L, Sharpe SC, Brown MW, Roger AJ. On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. Cold Spring Harb Perspect Biol. 2014; 6
42. Wybouw N, et al. A gene horizontally transferred from bacteria protects arthropods from host plant cyanide poisoning. Elife. 2014; 3:e02365. [PubMed: 24843024]
43. Groussin M, et al. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nat Commun. 2017; 8:14319. [PubMed: 28230052]
44. Knoll AH, Bergmann KD, Strauss JV. Life: the first two billion years. Philos Trans R Soc Lond. B Biol Sci. 2016; 371 20150493.
45. Wolfe JM, Fournier GP. Tunneling through time: Horizontal gene transfer constrains the timing of methanogen evolution. bioRxiv. 2017; :129494.doi: 10.1101/129494
46. Soo RM, Hemp J, Parks DH, Fischer WW, Hugenholtz P. On the origins of oxygenic photosynthesis and aerobic respiration in Cyanobacteria. Science. 2017; 355:1436–1440. [PubMed: 28360330]
47. Szöllősi GJ, Tannier E, Daubin V, Boussau B. The inference of gene trees with species trees. Syst Biol. 2015; 64:e42–62. [PubMed: 25070970]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

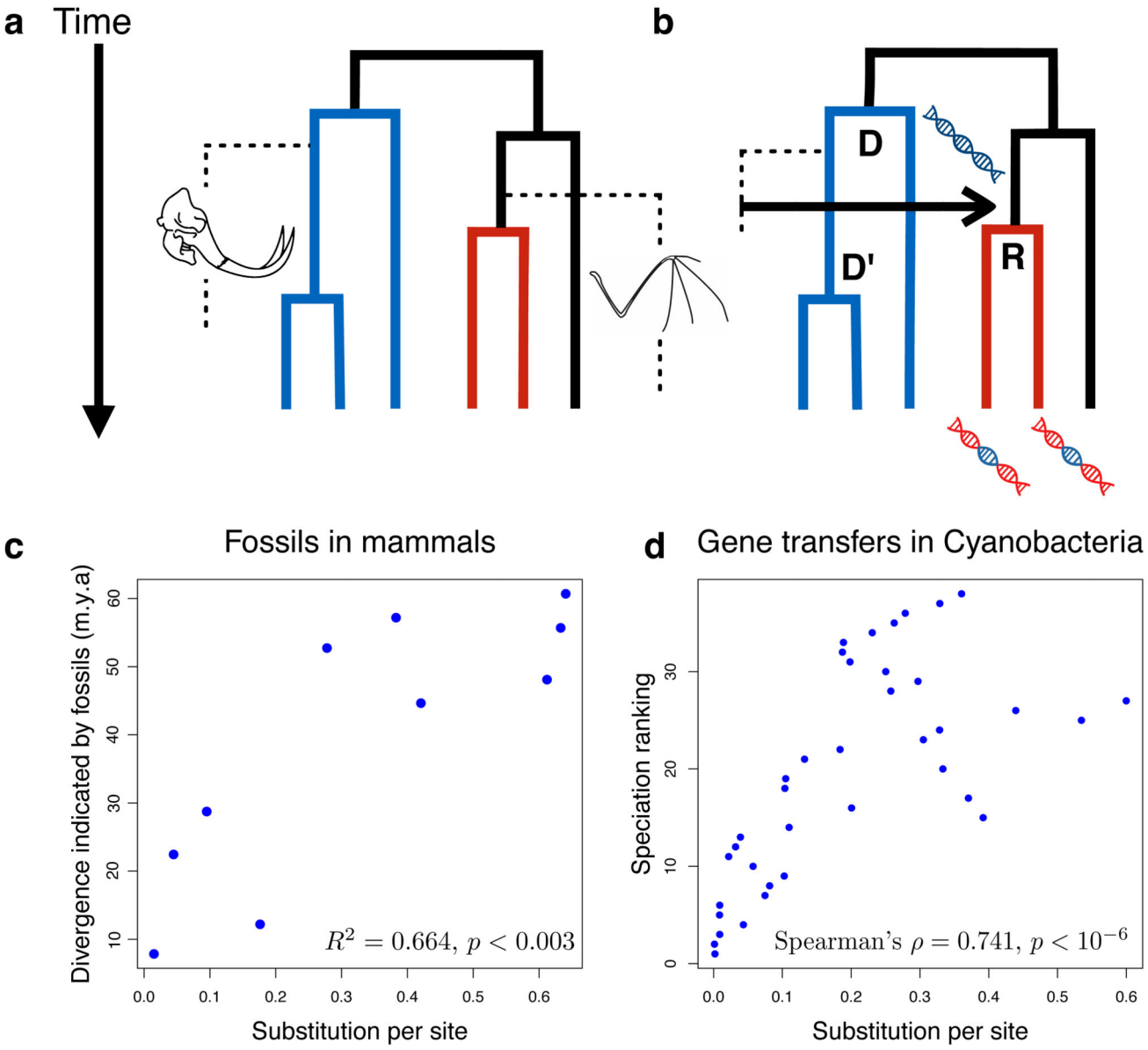

Figure 1. Gene transfers, like fossils, carry information on the timing of species divergence

a) The geological record provides the only source of information concerning absolute time: the age of the oldest fossil representative of a clade provides direct evidence on its minimum age, but inferring maximum age constraints (e.g. dashed line for the red clade), and by extension the relative age of speciation nodes, must rely on indirect evidence on the absence of fossils in the geological record5,23–25. **b)** Gene transfers, in contrast, do not carry information on absolute time, but they do define relative node age constraints by providing direct evidence for the relative age of speciation events: the gene transfer depicted by the black arrow implies that the diversification of the blue donor clade predates the diversification of the red clade (*i.e.* node D is necessarily older than node R). Note, however, that the depicted transfer is not informative about the relative age of nodes D' and R. **c)**

Sequence divergence (here measured in units of expected number of nucleotide substitutions along a strict molecular clock time tree, see supplementary materials) for 36 mammals2 is correlated (Pearson's R^2=0.664, $p<10^{-2}$) with age estimates based on the fossil record (ages corresponding to the time of divergence in million years). **d)** A similar relationship can be seen for gene transfer based relative ages by plotting the sequence divergence (measured similar to part c) against the relative age of ancestral nodes for 40 cyanobacterial genomes (Spearman's rank correlation rho=0.741, $p<10^{-6}$) inferred by the MaxTiC (**max**imal **ti**me **c**onsistency) algorithm26.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

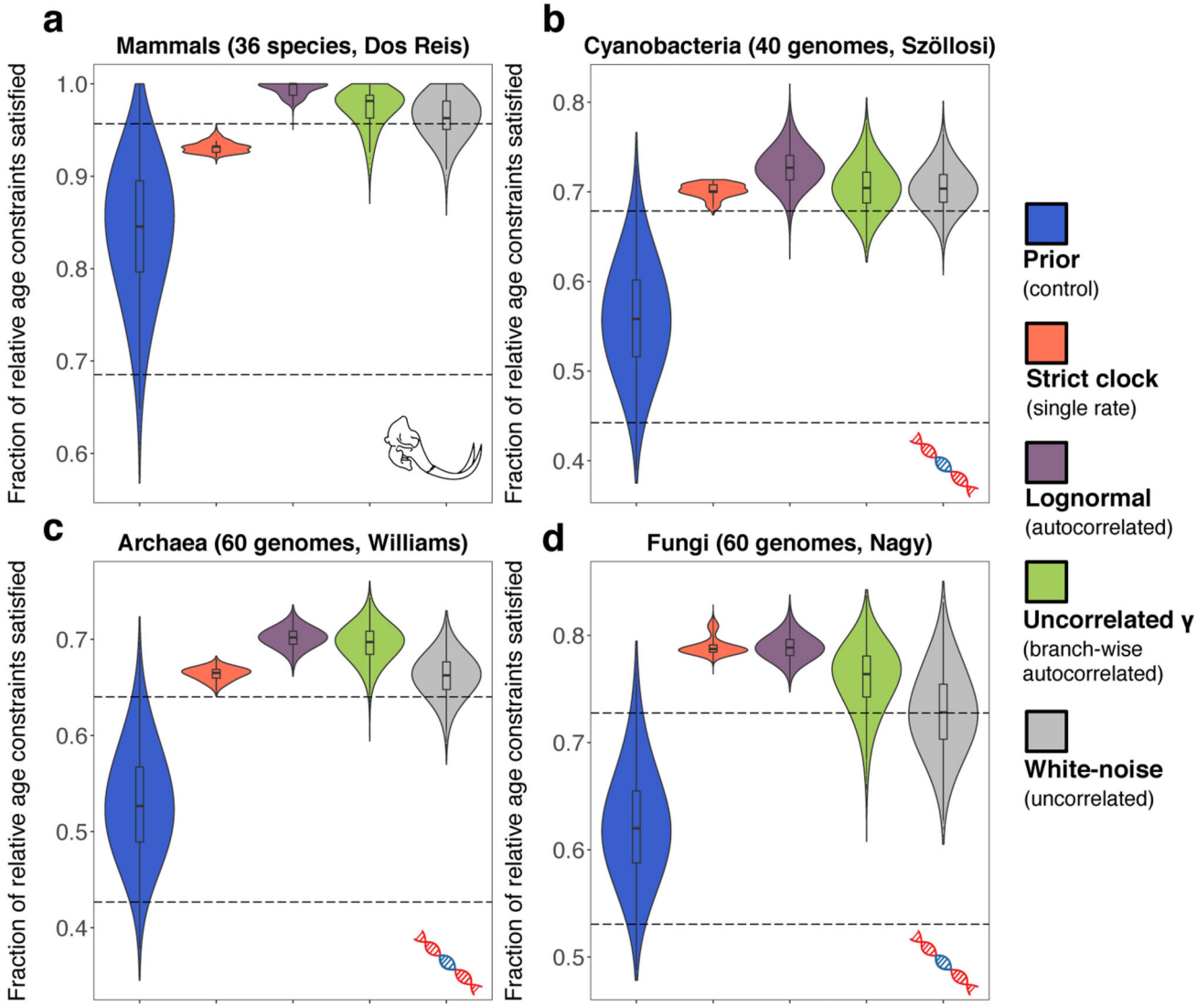

Figure 2. Agreement between transfer based relative ages and molecular clocks

a) Relative ages derived from 12 fossil calibrations from a phylogeny of 36 extant mammals were compared with node ages sampled from four different relaxed molecular clock models implemented in Phylobayes and with node ages derived from random chronograms, keeping the species phylogeny fixed. **b-d)** Relative ages derived from gene transfers using the MaxTiC algorithm were compared with estimates from the same 5 models as in a). For each model and each sampled chronogram we calculated the fraction of relative age constraints that are satisfied. Each violin plot shows the distribution of the fraction of relative age constraints satisfied by 5000 sampled chronograms; inside the violins, boxes correspond to the first and third quartiles of the distribution, a thick horizontal line to the median, and the whiskers extend to extrema no further than 1.5 times the interquartile range. The blue distribution corresponds to random chronograms drawn from the prior with the 95% confidence interval denoted by dashed lines, orange to the strict molecular clock, purple to

the autocorrelated lognormal, green to the uncorrelated gamma and grey to the white-noise models.

Europe PMC Funders Author Manuscripts

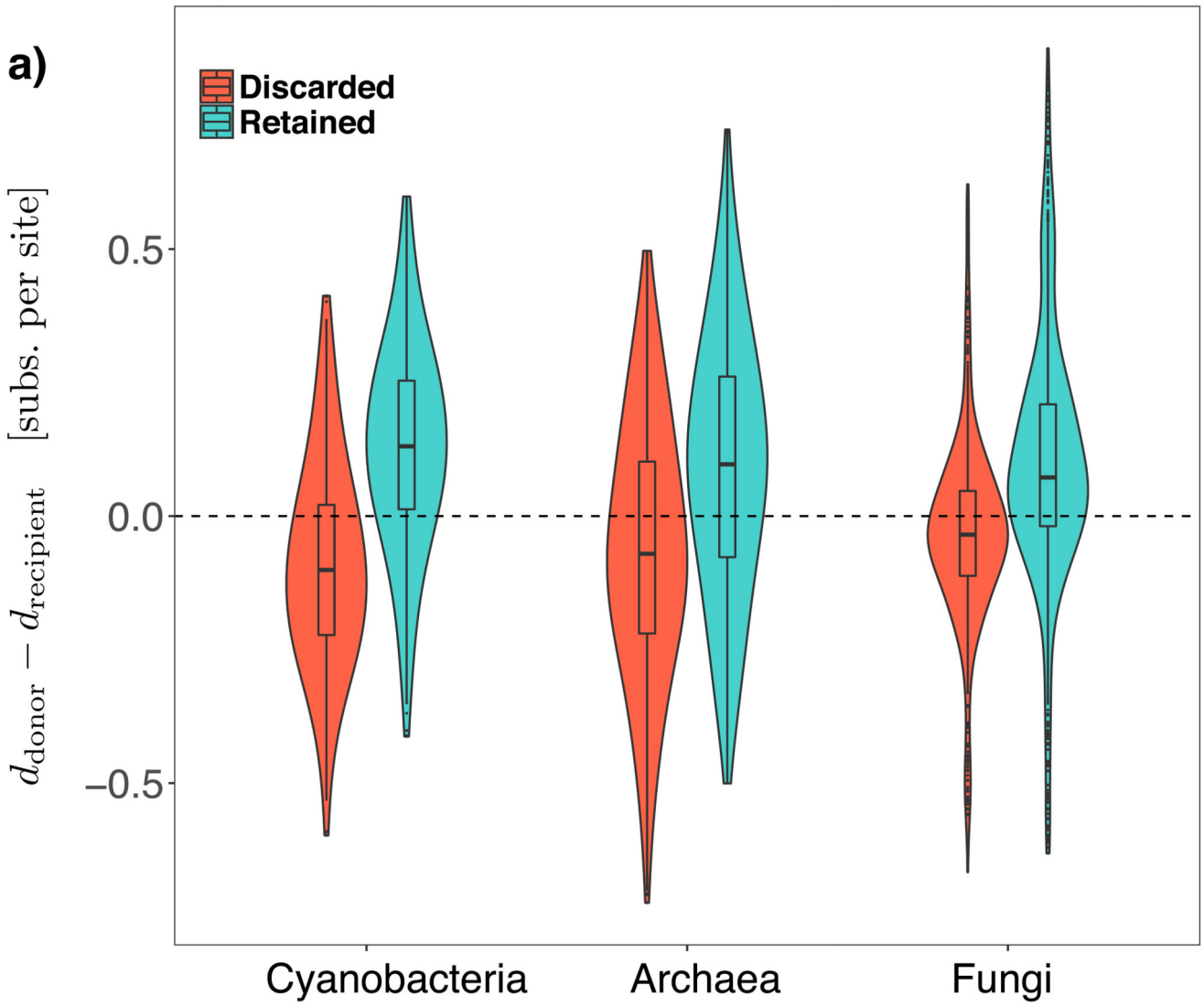

Figure 3. Donor clades appear older than recipient clades in LGTs retained by MaxTiC.

For genuine LGTs, the donor lineage must be at least as old as the recipient. As one proxy to investigate whether this is the case for transfers retained by our MaxTiC algorithm, we calculated clade-to-tip distances (see supporting text for details) for the inferred donor and recipient clades for LGTs that were retained and discarded by MaxTiC. (a) In all three datasets, transfers retained by MaxTiC (in red) have the property that donor clades are further from the tips of the tree than recipient clades, but the opposite pattern is observed for conflicting transfers rejected by MaxTiC (green), consistent with the idea that MaxTiC identifies genuine LGTs.

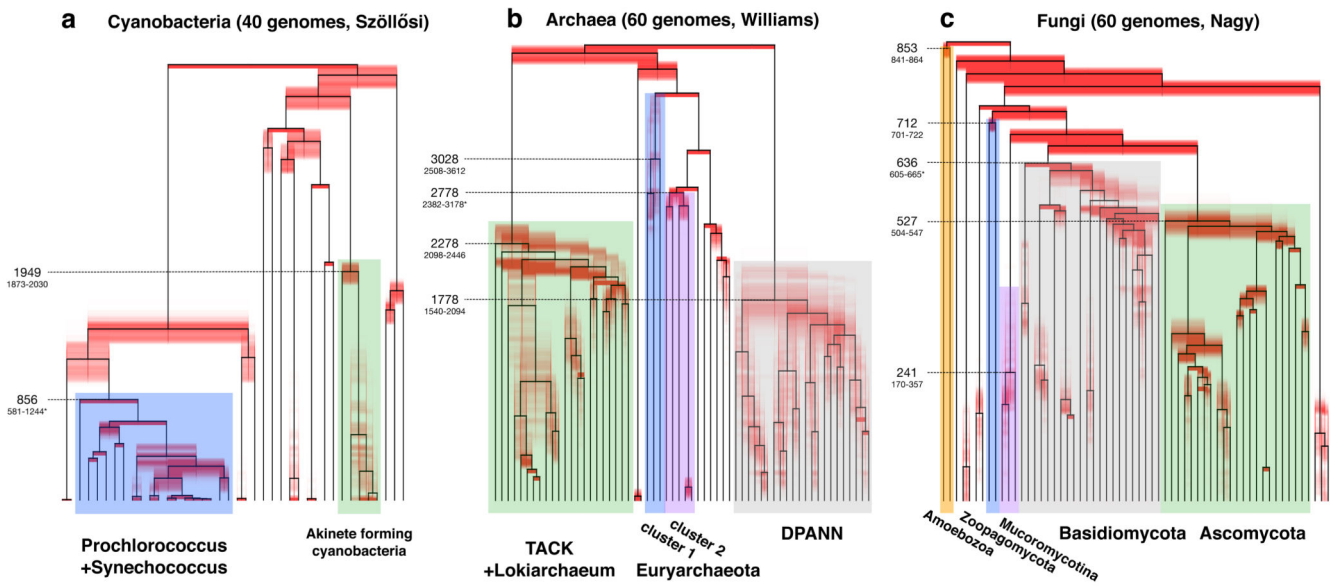

Figure 4. The order of speciations according to LGT calibrated to geological time.

5000 chronograms with a speciation time order compatible with LGT-based constraints were sampled per data set and calibrated to geological time (a: Cyanobacteria, b: Archaea, c: Fungi, for details see Methods). The black line corresponds to the consensus chronogram. Red shading represents the spread of node orders within the sample: nodes are in bright red if there is little or no uncertainty on their order according to LGT, in a light red smear if there is high uncertainty on their order. Dates in units of millions of years ago are provided for clades discussed in the text, which are labeled and shaded. Confidence intervals indicate 95% HPD of the time calibrated time orders with the exception of nodes, indicated with an asterisk, that had unambiguous calibrated time orders for which the 95% HPD of the corresponding node from Figures S25-27 is given. Supplementary Figures S1, S2 and S3 provide the same consensus chronograms with species names at the tips.

Europe PMC Funders Author Manuscripts