Lateral Gene Transfer from the Dead

GERGELY J. SZÖLLŐSI[1,2,*], ERIC TANNIER[1,2,3], NICOLAS LARTILLOT[4,5], AND VINCENT DAUBIN[1,2]

[1]*Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Université Lyon 1, F-69622 Villeurbanne, France;* [2]*Université de Lyon, F-69000 Lyon, France;* [3]*Institut National de Recherche en Informatique et en Automatique Rhône-Alpes, F-38334 Montbonnot, France;* [4]*Département de Biochimie, Centre Robert-Cedergren pour la Bioinformatique, Université de Montréal, Québec, Canada; and* [5]*Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS-Université de Montpellier 2, 34095 Montpellier, France* *Correspondence to be sent to: Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Université Lyon 1, F-69622 Villeurbanne, France; E-mail: ssolo@elte.hu.*

Abstract.—In phylogenetic studies, the evolution of molecular sequences is assumed to have taken place along the phylogeny traced by the ancestors of extant species. In the presence of lateral gene transfer, however, this may not be the case, because the species lineage from which a gene was transferred may have gone extinct or not been sampled. Because it is not feasible to specify or reconstruct the complete phylogeny of all species, we must describe the evolution of genes outside the represented phylogeny by modeling the speciation dynamics that gave rise to the complete phylogeny. We demonstrate that if the number of sampled species is small compared with the total number of existing species, the overwhelming majority of gene transfers involve speciation to and evolution along extinct or unsampled lineages. We show that the evolution of genes along extinct or unsampled lineages can to good approximation be treated as those of independently evolving lineages described by a few global parameters. Using this result, we derive an algorithm to calculate the probability of a gene tree and recover the maximum-likelihood reconciliation given the phylogeny of the sampled species. Examining 473 near-universal gene families from 36 cyanobacteria, we find that nearly a third of transfer events (28%) appear to have topological signatures of evolution along extinct species, but only approximately 6% of transfers trace their ancestry to before the common ancestor of the sampled cyanobacteria. [Gene tree reconciliation; lateral gene transfer; macroevolution; phylogeny.]

From the first growth of the tree, many a limb and branch has decayed and dropped off; and these lost branches of various sizes may represent those whole orders, families, and genera which have now no living representatives, and which are known to us only from having been found in a fossil state.

Charles Darwin, *On the Origins of Species*. London, 1859

Most of the diversity of life that ever existed on earth has gone extinct and can only be glimpsed from the fossil record. Although the comparative approach allows the reconstruction of some morphological and genetical characteristics of ancestral species, it is only informative for species that have founded extant lineages. Yet, the information enclosed in genome sequences is abundant and particularly meaningful for the reconstruction of the descent and evolution of their carriers (Zuckerkandl and Pauling 1965; Boussau and Daubin 2010; David and Alm 2011), so much so that it may have recorded accounts of extinct lineages. This possibility exists because the success of lateral gene transfer (LGT) as an evolutionary process implies that each gene possesses its own, unique history, which is not necessarily confined to the history of those species that have survived (Maddison 1997; Galtier and Daubin 2008; Fournier et al. 2009; Abby et al. 2012).

Several models have recently been developed to reconcile seemingly contradictory gene phylogenies with the species phylogeny by tracing the path on the species phylogeny along which they evolved as a result of a series of speciations, gene duplications, LGT, and losses (Tofigh 2009; Doyon et al. 2010; David and Alm 2011; Szöllősi and Daubin 2012; Szöllősi et al. 2012). None of these models, however, take into consideration the fact that, in the presence of LGT, gene trees record evolutionary paths along the complete species tree, including extinct and unsampled branches, and not only along the phylogeny of the species in which they reside today. This is the case because, as first noted by Maddison (1997) and later elaborated by Gogarten and coworkers (Zhaxybayeva and Gogarten 2004; Fournier et al. 2009), while LGT events imply that the donor and receiver lineages existed at the same time, the donor lineage might have subsequently become extinct, or more generally, might not have been sampled.

Here, we demonstrate that, if the number of species considered in the species phylogeny is small compared with the total number of species, the overwhelming majority of gene transfers involve speciation to and evolution along extinct or unsampled species. Furthermore, we show that, if this condition is met, the evolution of genes along the unrepresented parts of the species phylogeny can to good approximation be treated as those of independently evolving lineages, the behavior of which depends only on the global parameters of the speciation dynamics. This in turn allows us to derive the probability of observing a gene phylogeny by extending the ODT model introduced previously (Szöllősi et al. 2012). Applying our model to a data set derived from 36 cyanobacterial species, we perform a preliminary assessment of the phylogenetic signal for the evolution of transferred genes along extinct species.
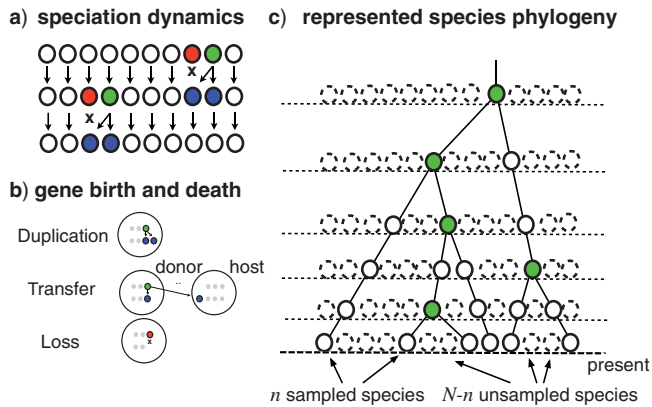
FIGURE 1.   Gene trees are the result of the combination of speciation and gene birth and death. As a minimal description we consider: a) that for each of the N species at a rate σ, a speciation occurs, during which the species is succeeded by two descendants, and a random species suffers extinction; b) at a rate δ per gene, a gene duplicates, that is, it is succeeded by two gene copies in the same genome, at a rate $\tau/(N-1)$ per gene per host species, a gene is transferred, resulting in one copy each in the donor and host species, and finally, with a rate λ per gene, a gene is lost. The represented phylogeny c) corresponds to the tree spanned by the n sampled species. A branch of the represented tree corresponds to a series of speciation events, but only the last of these, the speciation event that gives rise to two represented lineages (filled circles, green online) is explicitly present for internal branches as the speciation node terminating the branch. The number of unrepresented species (dashed circles) is always much larger than the number of represented species (full circles).

A Minimal Model of Speciation and Gene Birth and Death

It is not feasible to specify, much less to reconstruct, the complete phylogeny of all species that ever existed. To describe the evolution of genes outside the represented phylogeny—along lineages that have become extinct or whose descendants have not been sampled—we must resort to modeling the speciation dynamics that gave rise to the complete phylogeny. Modeling the dynamics of speciation provides a stochastic model of the evolution of unrepresented lineages that can be used to describe gene histories given knowledge of the represented phylogeny and a few global parameters.

As a minimal model of speciation, here, we assume that the number of species N is constant, and that the dynamics of speciation is modeled by a continuous time Moran process (Moran 1962). That is, for each species at rate σ, a speciation occurs during which the species gives rise to two descendants and a randomly chosen species goes extinct (cf. Fig. 1a). The central assumption we make is that, of the N species existing at present (i.e., $t=0$), we sample only a small fraction $n \ll N$. In general, the validity of this assumption depends on the phylogenetic problem considered, but should almost always be met for major groups of bacteria and archaea, where the number of species that potentially exchange genes by LGT is inevitably much larger than the number of sampled species, even in large-scale studies (Ochman et al. 2000; Torsvik et al. 2002).

To describe the evolution of genes within the genomes of species, we assume genes to evolve independently according to a birth-and-death process that consists of gene duplication, transfer, and loss (Tofigh 2009; Szöllősi and Daubin 2012; Szöllősi et al. 2012). As shown in Figure 1b, a gene in the genome of any of the N species can: (i) be duplicated at rate δ; (ii) be transferred from a donor species to any of the other $N-1$ possible host species at a rate $\tau/(N-1)$; or (iii) be lost at a rate λ. Gene copies can also be born and be lost as a result of the speciation dynamics: (iv) at the species level, lineages experience speciation at a rate σ, in which case they are replaced by two copies in the two new species or; (v) suffer extinction at an identical rate σ. A branch e of the represented tree S in general corresponds to a series of speciation events, however, as shown in Figure 1c, only the last one of these, the speciation event that gave rise to two represented lineages, is explicitly present for internal branches as the (green online) speciation node terminating the branch.

Almost All Transfers Involve Speciation

To understand what fraction of transfers involves evolution along unrepresented species, we must compare the relative rate of transfers that are *direct transfers* between branches of the represented phylogeny S and *indirect transfers* that result in a gene returning to S after exiting it via speciation or transfer to unrepresented species.

To compare the contribution of indirect transfers and direct transfers to observed gene histories, we consider first only direct transfers and indirect transfers that involve a speciation to an unrepresented species. To describe the shape of the species tree generated by the Moran process introduced above, we can use the coalescent approach. Here, under Kingman's coalescent, the time to the most recent common ancestor of the n sampled species is of the order of $2N/\sigma(1-1/n) \approx 2N/\sigma$ (Kingman 1982). This implies that the expected number of unrepresented speciation events per branch of the species tree is much larger than one, being of the order of $\sigma \times 2N/\sigma/(2n-2) \approx N/n \gg 1$, as there are $(2n-2)$ branches of S. This suggests that for any pair of coexisting branches of the represented tree, a gene that descends from one of the branches and is transferred to the other, is likely to have experienced a speciation event "away" from the represented phylogeny spanned by the n sampled species before being transferred back to it.

To quantify the above argument, we can compare the expected number of transfers from branch f to branch e of the represented phylogeny, resulting from either a direct transfer or a more complex history involving a speciation event. Clearly, if the branches do not overlap in time, the expected number of direct transfers is zero. To consider overlapping branches, let us consider for simplicity that both e and f are terminal branches—similar results can be derived for any other pair of overlapping branches. The expected branch lengths are then $E(t_e)=E(t_f) \approx N/\sigma n$,
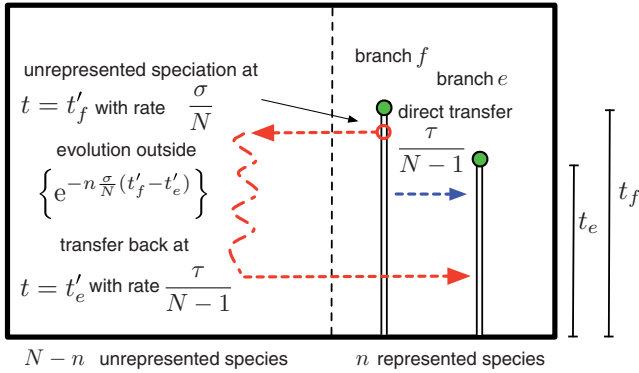
FIGURE 2. The overwhelming majority of transfers involve evolution along unrepresented species. A direct transfers (dark gray, blue online) between two terminal branches of the represented phylogeny occurs with rate $\tau/(N-1)$ and involves a single transfer event. An indirect transfer (light gray, red online) that leaves an indistinguishable record in the gene tree topology. To count indirect transfers, we trace their history backwards in time: transfer back to the host branch on the represented tree (branch e) occur with a rate $\tau/(N-1)$ from each of the $N-n$ unrepresented species, of these we are only concerned with ones which descend from the relevant donor branch (branch f), the number of these can be calculated using the exponential coalescence probability and the rate of unrepresented speciations σ/N from the donor branch (branch f).

with overlap $\min(t_e, t_f) \lesssim N/\sigma n$. Integrating over possible transfer times, the expected number of direct transfers is then

$$T_{\text{direct}} \lesssim \int_0^{\frac{N}{n\sigma}} \frac{\tau}{N-1} \, dt'_e = \frac{\tau}{2} \frac{1}{n} \frac{1}{(N-1)} \left[\frac{2N}{\sigma} \right]. \quad (1)$$

To estimate the expected number of indirect transfers that are topologically indistinguishable from the above direct transfers, we can reason backwards in time as illustrated in Figure 2: (i) the rate at which a transfer occurs from each of the $(N-n)$ unrepresented species to branch e is $\tau/(N-1)$, (ii) the probability of this gene lineage *not* coalescing back to any of the n branches of the represented tree during a time interval t is $\exp(-n\sigma/Nt)$, and (iii) the rate at which it coalesces with branch f is σ/N. Integrating over possible speciation and transfer times gives:

$$T_{\text{indirect}} \simeq \int_0^{\frac{N}{n\sigma}} \int_{t'_e}^{\frac{N}{n\sigma}} \frac{(N-n)\tau}{N-1} \left\{ e^{-n\frac{\sigma}{N}(t'_f - t'_e)} \right\} \frac{\sigma}{N} \, dt'_f \, dt'_e$$

$$= \frac{\tau}{2} \frac{1}{n^2} \frac{(N-n)}{e(N-1)} \left[\frac{2N}{\sigma} \right]. \quad (2)$$

Equations (1) and (2) show that if the number of sampled species is small compared with the total number of species ($n \ll N$), then the expected number of direct transfers is small compared with indistinguishable indirect ones ($T_{\text{direct}} \ll T_{\text{indirect}}$), that is, *the contribution of direct transfers to observed gene histories is negligible.*

To compare the two types of possible indirect transfers back to S—those exiting via speciation and those via transfer—we must contrast the rate σ at which gene copies exit branch f as a result of speciation and the rate

$\tau/(N-1) \times (N-1-n) \approx \tau$ that gene copies exit as a result of transfer. Estimates of τ, and more generally gene birth and death rates, are available from several sources, all of which agree that the expected number of gene birth and death events per branch is below unity. Models that consider the dynamics of the number of homologous gene copies along a species phylogeny (referred to as phylogenetic profiles) (Csűrös and Miklós 2009) have consistently found that birth and death rate is of the same order, with an excess of loss compensated by origination of new families, in agreement with phenomenological models of gene family size distribution (Karev et al. 2002; Szöllősi and Daubin 2012). In a detailed study, Csűrös et al. found for 28 archaea that the expected number of birth events (duplication and gain) is 0.12 and that the expected number of losses is 0.36 (Csűrös and Miklós 2009) per branch per gene. More recently, the ODT model that attempts to explicitly explain the evolution of multicopy gene trees (representative of complete genomes) along an ultrametric species tree has arrived at similar results (Szöllősi et al. 2012), finding for 36 cyanobacterial genomes $\delta \approx \tau \approx 0.2$, $\lambda \approx 1$, in units corresponding to a tree with unit height. Assuming, as above, that the time to the most recent common ancestor of the sampled species is of the order of $2N/\sigma$, that is, the expected number of gene copies (per gene) exiting a branch of S is proportional to N/n, while the number exiting as a result of transfer is less than one. Since the rate at which a gene that has exited the represented phylogeny returns to S as a results of transfer at some point in the future is independent of the mode of exit from S, we can conclude that *indirect transfers are dominated by paths that include a speciation.*

In summary, if the number of sampled species is small compared with the total number of species, transfers in observed gene histories are dominated by paths that include a speciation to an unrepresented species and subsequent transfer back to the represented tree.

THE PROBABILITY OF OBSERVING A GENE TREE

Reconciling gene trees with the species tree requires iterating over possible paths along which a gene tree may have been generated by a series of speciations, duplications, transfers, and losses (Fig. 3). In existing methods (Tofigh 2009; Doyon et al. 2010; Szöllősi and Daubin 2012; Szöllősi et al. 2012), this is accomplished by only considering paths along the represented phylogeny and using a dynamic programming approach exploiting the independence of gene birth and death events and by extension gene lineages.

Although gene duplication, transfer, and loss can reasonably be modeled as independent birth and death events, speciation and extinction necessarily involve the simultaneous birth and death of many genes. Along the represented phylogeny, speciation events are fully specified and can be explicitly taken into account (Szöllősi et al. 2012). This is not the case, however, for speciation and extinction events that occur
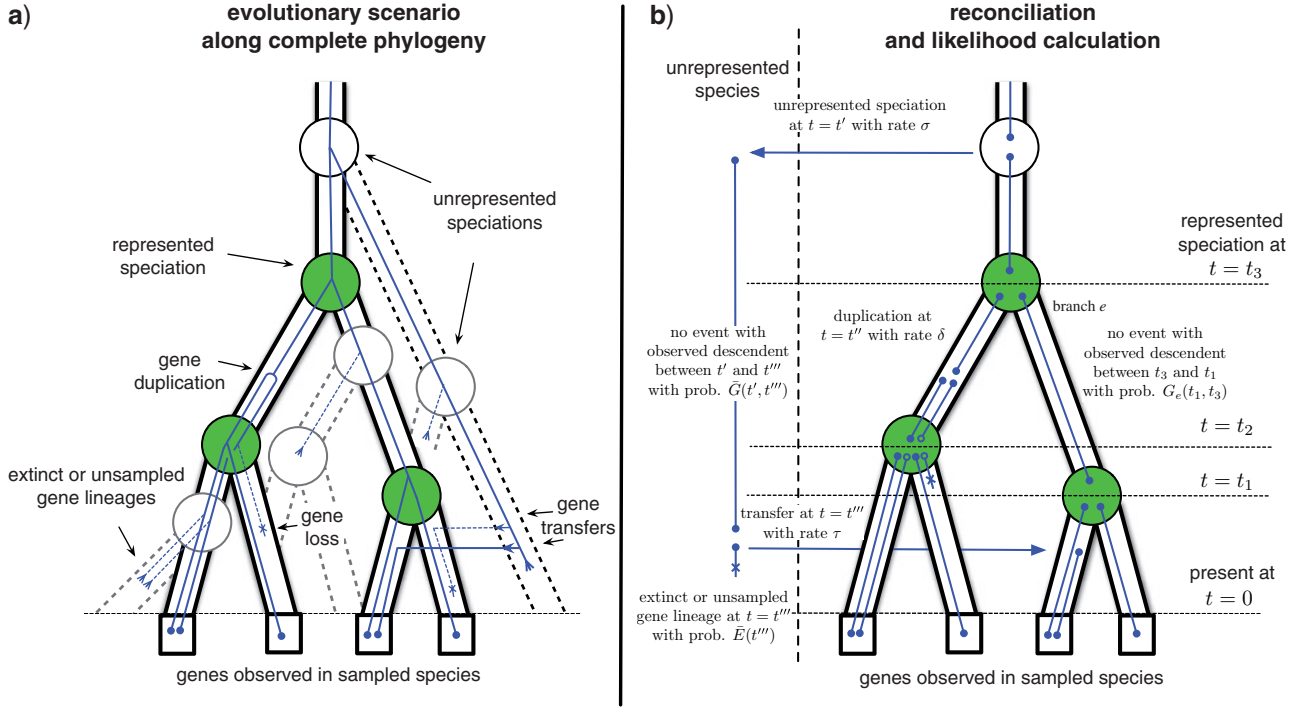
FIGURE 3. Reconciling gene trees with the complete phylogeny. a) An evolutionary scenario that involves a transfer event from an unrepresented species. The represented phylogeny is shown as a solid tube with filled circles (green online) corresponding to represented speciations. The unrepresented phylogeny is indicated by dashed tubes, with white circles corresponding to unrepresented speciations (cf. Fig. 1c). The continuous line traces the gene tree spanned by genes in sampled species that is the result of a series of birth and death events along the complete phylogeny; b) a reconciliation of the gene phylogeny from (a), corresponding to the evolutionary scenario depicted in (a). In general, we do not know the evolutionary scenario that has generated the gene phylogeny. However, we can use the dynamic programming algorithm described in the text to calculate the likelihood of the gene tree by summing over all possible reconciliations, that is, all ways to draw the gene tree into the species using speciation, duplication, transfer, and loss events [cf. Eqs. (4)–(7) and Fig. A1] in the Appendix. The likelihood calculation uses the rate of different events (σ, δ, τ, and λ) together with functions describing the extinction (E_e and $\bar{E}$) and the propagation (G_e and $\bar{G}$) of gene lineages [cf. Eqs. (A.1)–(A.4)].

in the unrepresented part of the phylogeny, or do not correspond to speciation nodes of the represented phylogeny. Therefore, unrepresented speciations result in nonindependence of gene lineages.

Consider for instance the probability $\bar{E}_k(t)$ that k genes present at time t in a species not ancestral to the sample of n extant species leave no observed descendant. Conditional on the complete phylogeny, ϕ including all extinct species lineages, gene lineages are independent, and therefore $\bar{E}_k(t|\phi) = \{\bar{E}(t|\phi)\}^k$. Averaging over all complete phylogenies compatible with the phylogeny reconstructed based on the n species, however, results in $\langle \bar{E}_k(t|\phi) \rangle = \langle \{\bar{E}(t|\phi)\}^k \rangle \neq \langle \bar{E}_k(t|\phi) \rangle^k$, which is not a product of k independent factors.

On the other hand, $n \ll N$ implies that $\bar{E}(t) \approx 1$. Introducing the notation $\bar{E}(t|\phi) = 1 - \epsilon(t|\phi)$ and $\bar{E}(t) = 1 - \epsilon(t)$, and neglecting second- and higher order terms in $\epsilon(t|\phi)$ and $\epsilon(t)$, we have

$$\bar{E}_k(t) = \langle \bar{E}_k(t|\phi) \rangle_\phi = \left\langle \left\{ \bar{E}(t|\phi) \right\}^k \right\rangle_\phi = \left\langle \left\{ 1 - \epsilon(t|\phi) \right\}^k \right\rangle_\phi$$

$$\simeq \left\langle \{1 - k\epsilon(t|\phi)\} \right\rangle_\phi = \left\{ 1 - k \langle \epsilon(t|\phi) \rangle_\phi \right\} = 1 - k\epsilon(t)$$

$$\simeq \{1 - \epsilon(t)\}^k = \left\{ \bar{E}(t) \right\}^k. \tag{3}$$

A similar argument can be derived for k-gene propagator $\bar{G}_k(s, t)$ (see the Appendix). Therefore, if $n \ll N$, then to good approximation, the evolution of two genes observed in the same unrepresented species can be treated as independent without specifying the full phylogeny.

Under the above assumption that unrepresented speciation and extinction events can be considered in a genewise independent manner, we can describe the evolution of gene copies that appear as single gene lineages when observed from the present. We can calculate: (i) the extinction probability $E_e(t)$ that a gene seen at time t on branch e of S leaves no *observed descendant*, that is, no descendant exists at time $t=0$ in the genome of any of the n sampled species; (ii) the extinction probability $\bar{E}(t)$ that a gene seen at time t in an unrepresented species leaves no observed descendant; (iii) the single gene propagation probabilities $G_e(s, t)$ that all observed descendants of a gene seen at time s on branch e descend from a descendant seen at a later time $t < s$ on branch e; and (iv) $\bar{G}(s, t)$ the probability that all observed descendants of a gene seen at time s in an unrepresented species descend from a descendant seen at time $t < s$ in an unrepresented species. Each of the above functions can be expressed as differential

equations describing evolution backwards in time by considering the set of possible events that change the relevant probability. These can be derived analogously to (Tofigh 2009; Stadler 2011; Szöllősi et al. 2012) and can be found in the Appendix.

Given a rooted gene tree topology G, we can now calculate the probability $p(G|S,\mathcal{M})$ of observing G, where $\mathcal{M}$ denotes the parameters of the model, by summing over all possible paths along S and over all complete phylogenies compatible with the species tree spanning the n species of the sample. We can sum over all paths by recursively mapping the branches of G onto branches of S generalizing the ODT models algorithm (Szöllősi et al. 2012) to include evolution along unrepresented species (cf. Figs 3 and A1 in the Appendix).

A branch of G represents the evolution of a gene copy for which (i) if the branch is nonterminal, all observed descendants descend from one of the two daughter gene lineages which emerge from the gene tree node in which the branch terminates or (ii) if the branch is terminal, a gene is observed in one of the genomes mapping to a leaf of S. To describe possible paths along S that this gene copy may take before arriving at the gene tree node in which it terminates, we must consider five events: (i) single-copy evolution along branch e of S described by G_e, (ii) single-copy evolution outside S described by $\bar{G}$, (iii) speciation from a branch of S to an unrepresented species such that only descendants of this copy are observed, (iv) transfer such that only descendants of the transferred copy are observed, and (v) speciation represented in S such that only one of the descending copies leaves an observed descendant. Each of these events leads to a single gene copy with observed descendants. The gene tree node in which the branch terminates can correspond to three possible events: (i) a duplication, a speciation represented in S; (ii) a speciation not represented in S; or (iii) a transfer. Each of these events leads to two gene copies with observed descendants.

To derive the recursion expressing the probability of G as the sum over possible paths along S, we discretize time along S keeping track of speciation times t_i along S. Speciations represented in S define the time intervals $[0,t_1),\ldots,[t_i,t_{i+1}),\ldots[t_{n-1},t_{n-1})$ referred to as *time slices* (Tofigh 2009; Doyon et al. 2010) with indices $0,\ldots,i,\ldots n$. We further divide each time slice into D equal time intervals of height $\Delta t_i = (t_{i+1} - t_i)/D$.

The probability of the gene lineage leading to node u of G being seen on branch e of S at time $t+\Delta t$ given the probabilities at time $t = t_i + \Delta t_i$ is

$$P_e(u,t+\Delta t_i) = G_e(t+\Delta t,t)P_e(u,t) \qquad (4)$$
$$+\{\delta \Delta t_i\}P_e(v,t)P_e(w,t)$$
$$+\{\sigma \Delta t_i\}\bar{P}(v,t)P_e(w,t)$$
$$+\{\sigma \Delta t_i\}P_e(v,t)\bar{P}(w,t)$$
$$+\{\sigma \Delta t_i\}\bar{P}(u,t)E_e(t),$$

where $\bar{P}(u,t)$ denotes the probability of the gene lineage leading to node u of G being seen in an unrepresented species at time t, v and w descend from u in G. As shown in Figure A1a in the Appendix, the terms correspond to (i) no event with an observed descendent; (ii) birth of two gene linages by duplication, such that both leave observed descendants; (iii) and (iv) birth of two gene linages with observed descendants as a result of an unrepresented speciation; and finally, (v) unrepresented speciation followed by the loss of the copy in branch e such that only the copy in the unrepresented phylogeny leaves an observed descendant. In the above expression, we only consider indirect transfers that involve a speciation, see the Appendix for the full expression.

The probability of being seen in such an unrepresented species is

$$\bar{P}(u,t+\Delta t_i) = \bar{G}(t+\Delta t_i,t)\bar{P}(u,t) \qquad (5)$$
$$+\left\{\left(2\sigma+\delta+\frac{N-n_i}{N-1}\tau\right)\Delta t_i\right\}\bar{P}(v,t)\bar{P}(w,t)$$
$$+\sum_{e\in\mathcal{E}_i}\left\{\frac{\tau\Delta t_i}{N-1}\right\}\bar{P}(v,t)P_e(w,t)$$
$$+\sum_{e\in\mathcal{E}_i}\left\{\frac{\tau\Delta t_i}{N-1}\right\}P_e(v,t)\bar{P}(w,t)$$
$$+\sum_{e\in\mathcal{E}_i}\left\{\frac{\tau\Delta t_i}{N-1}\right\}\bar{E}(t)P_e(u,t),$$

where $\mathcal{E}_i(S)$ denotes the set of branches of S in time slice i. As shown in Figure A1b, the terms correspond to (i) no event with an observed descendent; (ii) birth of two gene linages by speciation, duplication, or transfer, such that both leave observed descendants; (iii) and (iv) birth of two gene linages with observed descendants as a result of transfer back to the represented phylogeny; and finally, (v) transfer back to the represented phylogeny following which the copy in the unrepresented donor linage does not leave an observed descendant. Terms involving gene lineages v, w are zero if u is a leaf of G in both the above expressions.

At speciation times $t=t_i$ where branches f and g descend from e in S, a represented speciation takes place that may be followed by a loss:

$$P_e(u,t) = P_f(v,t)P_g(w,t) + P_f(w,t)P_g(v,t) \qquad (6)$$
$$+P_f(u,t)E_g(t) + E_f(t)P_g(u,t).$$

The terms (cf. Fig. A1c) correspond to (i) and (ii) represented speciation such that both resulting gene lineages lead to observed descendants; and (iii) and (iv) represented speciation such that only one of them do. Finally, at time $t=0$ on each terminal branch e of S, the presence of observed genes is expressed as:

$$P_e(u,0) = \begin{cases} 1 \text{ if } u \text{ is a leaf of } G \text{ found in } e \\ 0 \text{ otherwise} \end{cases}. \qquad (7)$$

As illustrated in Figures 3b and A1, each term in Equations (4)–(7) above corresponds to a series of speciation, duplication, and transfer events that recursively draw the gene phylogeny into the species tree. The recursion calculates the probability of a gene tree with m genes in $O(Dn^2m)$ steps, as there are fewer than n branches in each time slice and n time slices. Summing over roots of G can be accomplished with identical complexity using double recursion. The most likely reconciliation can be recovered by tracing back along the sum choosing at each step the event with the highest probability.

Calculating the probability of a gene tree requires knowledge of the ultrametric species tree S, with branch lengths corresponding to time, the rate of duplication δ, transfer τ, and loss λ, as well as the parameters of the speciation dynamics, the species replacement rate σ, and the total number of species N. The number of parameters is reduced, if we assume the time to the common ancestor of the sampled species to correspond to its expected value under speciation dynamics. Choosing units such that S is of unit height, this corresponds to the choice σ = 2N. Furthermore, under the present choice of parameters and time scale, the probability of a gene tree and its maximum-likelihood reconciliation depends only very weakly on N, as long as the condition $n \ll N$ is satisfied. This is the case because the expected number of transfers between branches of S is nearly independent of N. In particular, if we assume that a gene lineage returns at most once to S we arrive at the result derived in Equation (2) according to which the number of transfers is independent of N.

ROUTES TO CYANOBACTERIAL GENOMES

To carry out a preliminary analysis of the signal for evolution outside the represented phylogeny in real data, we considered a set of 473 single-copy gene families present in the genome of at least 34 of 36 cyanobacteria and use the dated species tree reconstructed in Szöllősi et al. (2012). We choose single-copy near-universal gene families as they are expected to be (i) relatively slowly evolving and hence to harbor a strong signal of homology and yield high-quality alignments, and (ii) they can be assumed to be well described by a single set of uniform duplication, transfer, and loss rates, at least in contrast to more complex data sets composed of multicopy families. For each family, gene tree topologies and duplication, transfer, and loss rates that maximize the joint likelihood (Maddison 1997; Szöllősi and Daubin 2012) were inferred as described in the Appendix. Using these results, 1000 reconciliations per family were sampled by stochastic backtracking along the sum over reconciliations.

On average, we found 0 duplication, 2.15 transfers, and 2.56 losses per family. The distribution in time of transfer events and the preceding speciations to unrepresented species are shown in Figure 4a. The majority of transfers occur between branches of S that overlap in time,

hence the resulting gene tree carries no topological signature of the length of time spent evolving along unrepresented lineages. Transfers between branches that do not overlap in time, for which the gene tree topologies explicitly record evolution outside the represented tree, correspond to 27.8% of all transfers. About a fifth of these (5.9% of all transfers) branch above the root indicating transfer from outside the sampled diversity of cyanobacteria. The median interval of time spent evolving in unrepresented lineages is 0.083 (or 222 million years, henceforth myr) for transfers between overlapping branches and 0.39 (or 1000 myr) for transfers between nonoverlapping branches. Similar values are obtained if we consider only the maximum-likelihood reconciliations, except for the median interval of time spent evolving in unrepresented lineage for transfers between overlapping branches which is only 0.0028 (or 8.1 myr corresponding to the minimum length allowed by time discretization). The corresponding value for transfers between nonoverlapping branches, 0.36 (or 990) myr, is nearly identical to the value above.

We emphasize that an important caveat of these results is that the accuracy of our method to infer correct reconciliations and gene topologies has not been assessed. This could be accomplished by explicit simulations of gene family evolution along the complete phylogeny. Such simulations are, however, outside the scope of the current publication, as they are technically challenging due to the large number of species in the complete phylogeny, and because they must address a potentially long list of possible questions. In lieu of simulation, it is possible to examine the posterior support of individual transfer events, which, as described in the Appendix, can be calculated as the fraction of times we find a given transfer event among the sampled reconciliations for each family. Using this measure, we find that transfers are well supported with 66.8% of transfer events having support over 0.95.

It is also important to discuss to what extent we can expect observed transfers between nonoverlapping branches to be robust to increasing the number of sampled species. Consider the extreme case that all N extant species are sampled. It is clear that transfers between overlapping branches of S (red in Fig. 4b) may correspond to transfer between nonoverlapping branches of the full phylogeny spanned by all N extant species. To ascertain how often we expect the opposite to occur, to have a transfer between nonoverlapping branches of S correspond to transfers between overlapping branches of the full phylogeny spanned by all N extant species, we need to estimate how often we expect to sample an extant descendant of the unrepresented donor lineage involved in a transfer between nonoverlapping branches of S (light bars, yellow online, in Fig. 4b). Assuming a tree with unit height, the total branch length of the full phylogeny under Kingman's coalescent is of the order of log(N), while the total branch length including extinct species is of the order N. Thus, we expect
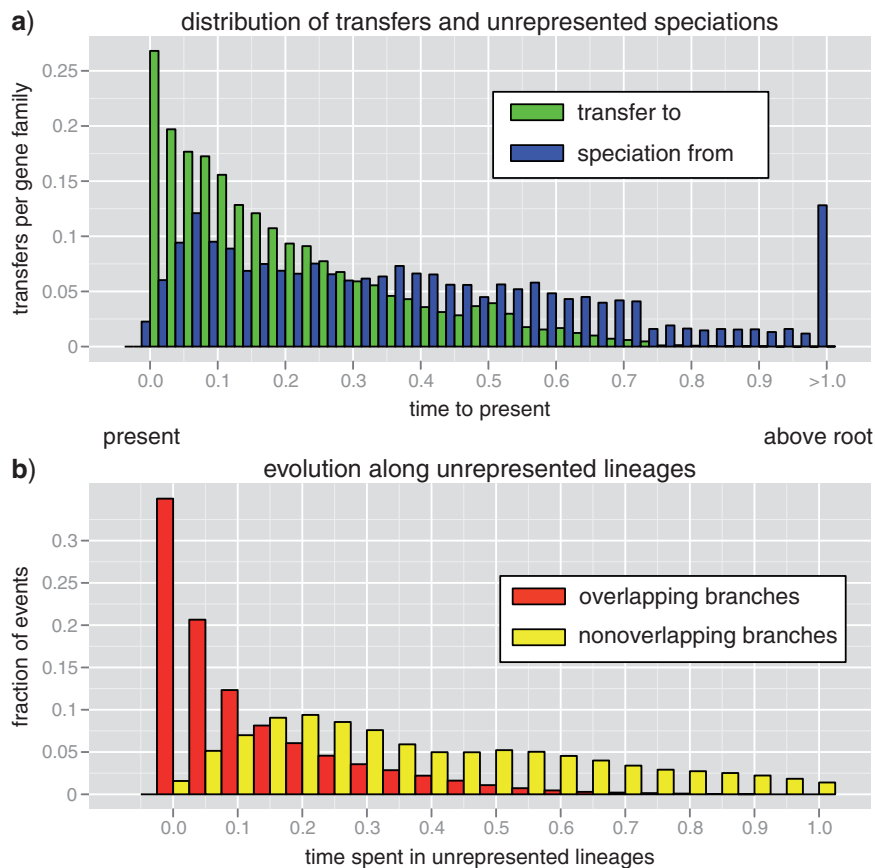
FIGURE 4. LGT events for 36 cyanobacteria. For 473 near universal single-copy families from 36 cyanobacterial genomes gene trees that maximize the joint likelihood were reconstructed. For the trees obtained 1000 reconciliations were sampled. a) The distribution of transfer events (light bars, green online) and the preceding speciation events (dark bars, blue online). The final bin summarizes all events occurring above the root of S. b) The distribution of the time spent by transferred genes evolving along unrepresented species for transfers between overlapping branches (dark bars, red online, 72.2% of transfers) and transfers between nonoverlapping branches (light bars, yellow online, 27.8% of all transfers). Both sets of bins sum to unity. Time units are chosen such that the height of the root of S is 1.0. The age of the root falls in the 3500–2700 My interval (Falcón et al. 2010; Szöllősi et al. 2012). Data are available from Dryad under doi:10.5061/dryad.27d0g.

that only a vanishing fraction of the order $\log(N)/N$ of donor lineages have left extant descendants. This implies that not only do most transfers involve speciation to, and evolution along branches of the complete phylogeny, but also the majority of these donor lineages have gone extinct. Consequently, most transfers between nonoverlapping branches of S correspond to transfers between nonoverlapping branches of the full phylogeny where the donor lineage has gone extinct.

In summary, we find that nearly a third (27.8%) of transfers evolve on average a billion years along lineages unrepresented in the phylogeny—most often, in fact, along extinct lineages, and only a moderate fraction of transfers originates from outside the cyanobacteria. Furthermore, both of these estimates are conservative, as increasing the number of sampled species is expected to lead to an increase in the ratio of transfers between nonoverlapping branches, and to a decrease in the fraction of transfers from outside of cyanobacteria. The first of the above results, however, apply only to transfers

between branches of S, that is, transfers observed for the $n = 36$ cyanobacteria considered. For the complete set of transfers between branches of the full phylogeny, the fraction of transfers evolving along extinct lineages is potentially different, for example, a macroscopic fraction of transfers are expected to correspond to direct transfers between its branches.

DISCUSSION

The results developed above are conditional on two crucial assumptions: (i) that the number of sampled species is small compared with the total number of species, and (ii) the evolution of gene lineages can be treated as independent, both in the represented and in the unrepresented part of the phylogeny. As we argue above, if genes are duplicated, transferred, and lost independently, the former assumption (i.e., $n \ll N$) implies that the evolution of genes outside the represented phylogeny can also be treated as

independent, even if the complete phylogeny is not specified.

We also make the assumptions that (iii) transfer occurs with identical rate between any two species and (iv) that the time to the last common ancestor of the sampled species corresponds to its expected value under the speciation dynamics. These conditions serve to simplify the development of the above arguments and can be relaxed without affecting our conclusion that the majority of transfers involve evolution along extinct or unsampled species. Relaxing Condition (iv) is straightforward. Concerning Assumption (iii), if, for example, transfer occurs preferentially between species that are more closely related (Andam and Gogarten 2011), the scenarios shown in Figure 2 are affected to an identical extent because the last common ancestor of branch e and either branch f (the donor lineage for dark gray paths, blue online) or any extinct species that descends from an unrepresented speciation along f (a donor lineage along light gray paths, red online) is the same. Conversely, there are known cases, for example, the transfer of thermostable enzymes from thermophilic archaea to thermophilic bacteria (Nelson et al. 1999; Nesbo et al. 2001; Brochier-Armanet and Forterre 2007), of preferential transfer between distantly related taxa due to shared ecology. In this second case, we expect to observe genes preferentially transferred from phylogenetically distant taxa to lead to an excess of transfers descending from above the root of the sampled species for which topologically equivalent direct transfers do not exist. On a more practical ground, however, relaxing the assumption of homogeneous rates of transfer between lineages might seriously complicate the computation of the likelihood, as it would require modeling the distribution of the rates of transfers from and to unrepresented lineages.

More importantly, as long as these conditions are met, it is possible to extend the above results to more general models of speciation. Modeling variation in N, the total number of species, over geological times, could be of particular interest. Indeed, a corollary of the observation that LGT events record evolutionary paths along the complete species tree is that the phylogenies of genes from a limited sample of extant species carry information about extinct lineages, and therefore about the size and dynamics of ancient biodiversity. In fact, patterns of gene transfer may be even more informative about past biodiversity than the species tree itself. Drawing an analogy with population genetics, inferring biodiversity dynamics based on species trees (Nee 2001; Morlon et al. 2010; Stadler 2011) is similar to inferring past demography based on single-locus data. Single-locus inference is limited by the intrinsic stochasticity of Kingman's coalescent, in particular in the deep part of the genealogy. LGTs, on the other hand, are analogous to multiple loci (Heled and Drummond 2008), and as such, have the potential to increase the statistical power for inferring past biodiversity.

APPENDIX

The dynamic programming algorithm described in Equations (4)–(7) calculates the likelihood of a gene tree given the species tree S and the rates of speciation, duplication, transfer, and loss. As illustrated in Figure 3, the likelihood is calculated in a piecewise independent manner, from the evolution of gene copies that appear as single gene lineages when observed from the present. Because in contrast to Figure 3, we do not know the exact evolutionary scenario we must sum over all reconciliations. This process can be represented as summing over all possible ways to draw the gene tree into the species tree using the set of events shown in Figure A1. The diagrams in Figure A1, and the corresponding terms in Equations (4)–(7), are expressed using two types of functions describing the evolution of single gene lineages: (i) the extinction probabilities E_e and $\bar{E}$ that give the probability of gene present on, respectively, branch e of S or an unrepresented species having no descendant at time $t=0$ in the genome of any of the n sampled species; (ii) the single gene propagators $G_e(s,t)$ and $\bar{G}(s,t)$ corresponding to the probability that all sampled descendants of the gene seen at time s, respectively, on branch e of S, or in an unrepresented species, descend from the gene present at a later time t in the same species.

Below we provide the expressions for each of these functions that can be derived using the theory of birth-and-death processes. We also discuss the independence assumption in relation to the single gene propagators, write down the complete form of Equation (4) and describe the details of the data analysis presented in the main text.
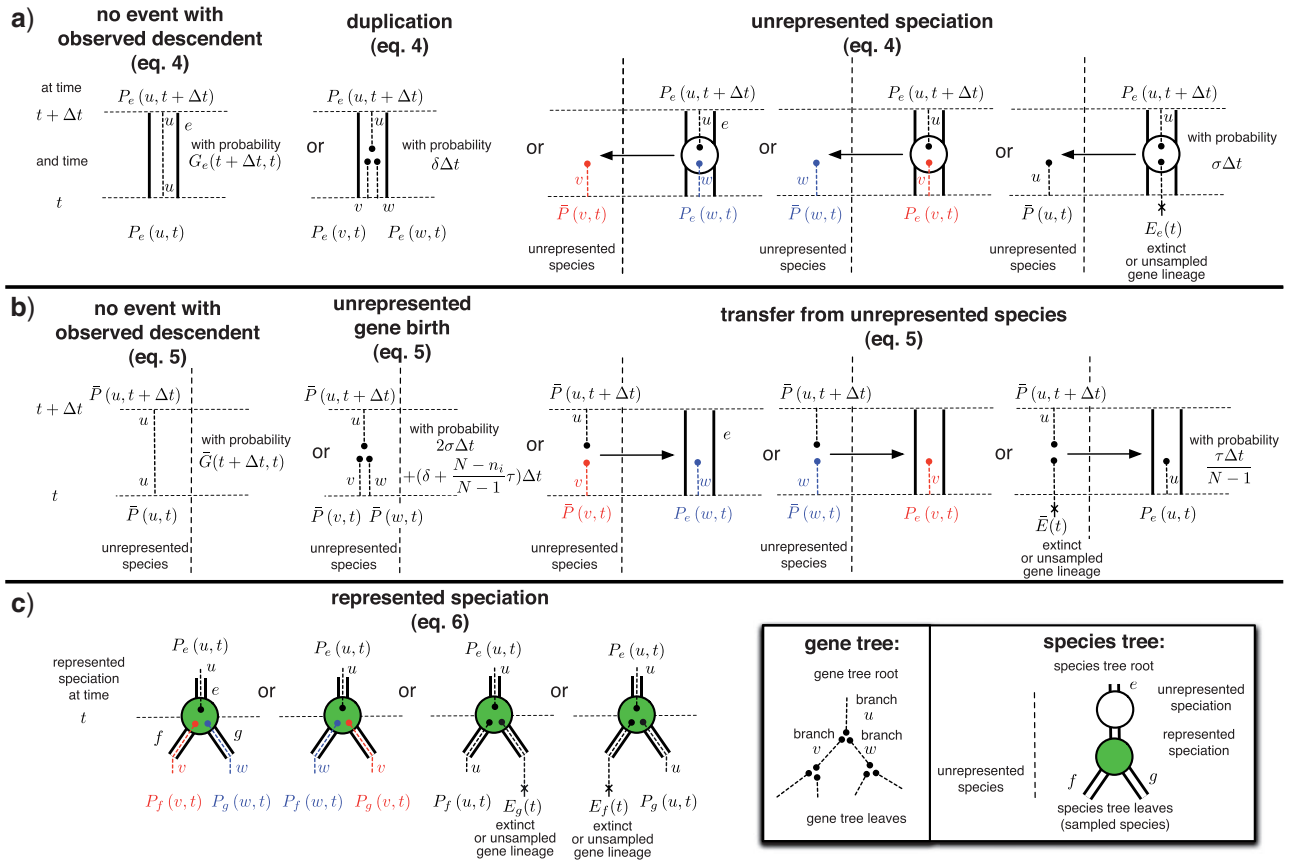
FIGURE A1. Diagrams corresponding to reconciliation events. Each diagram corresponds to a term in Equations (4)–(7), with diagrams following each other in the same order as terms in the indicated equation. a) Depicts events that start with a gene lineage u in represented branch e of S at time $t+\Delta t$; b) events that start with a gene lineage u in an unrepresented species at time $t+\Delta t$; and finally, c) corresponds to represented speciation events in S. To illustrate the correspondence between terms and equations, consider the third diagram in the top row (a) depicting an unrepresented speciation and the corresponding (third) term in Equation (4). This term, $P_e(u,t+\Delta t)=\cdots+\{\sigma\Delta t\}P_e(v,t)\bar{P}(w,t)+\cdots$, describes the probability that gene lineage u seen at time $t+\Delta t$ is succeeded as a result of an unrepresented speciation by two gene linages (v and w) one of which (w) is present in the same branch e as u while the other (v) resides in an unrepresented species.

Evolution of Single Genes

The forward Kolmogorov equations describing single gene extinction and propagation can be derived analogously to Tofigh (2009), Stadler (2011), and Szöllősi et al. (2012). The main differences is that here we also consider the speciation dynamics.

The extinction probability for branch e of S:

$$\frac{d}{dt}E_e = +\{\lambda\}(1-E_e) \qquad (A.1)$$
$$-\{\delta(1-E_e)\}E_e$$
$$-\left\{\sum_{f\in\mathcal{E}_i(S)}\frac{\tau}{N-1}(1-E_f)\right\}E_e$$
$$-\left\{\left(\sigma+\frac{N-n_i}{N-1}\tau\right)(1-\bar{E})\right\}E_e$$
$$E_e(t_e^{\text{end}})=\begin{cases}0 & \text{if } t_e^{\text{end}}=0,\\ E_f(t_g^{\text{begin}})E_g(t_f^{\text{begin}}) & \text{otherwise}\end{cases},$$

where $\mathcal{E}_i(S)$ denotes the set of branches of S in time slice i, and n_i their number. The terms correspond to (i) loss (ii) the rate duplications, and (iii) transfers to represented hosts, both conditional on survival, and finally, (iv) the rate of unrepresented speciations and transfers to unrepresented hosts, again conditional on survival. The initial conditions specify that at the end of branch e the probability of extinction is 0 if we are at time $t=0$, that is, e is a terminal branch of S, and the product of the extinction probability of the descendants of e in S otherwise.

The extinction probability in an unrepresented species:

$$\frac{d}{dt}\bar{E} = +\{\sigma+\lambda\}(1-\bar{E}) \qquad (A.2)$$
$$-\left\{\left(\sigma+\delta+\frac{N-n_i}{N-1}\tau\right)(1-\bar{E})\right\}\bar{E}$$
$$-\left\{\sum_{f\in\mathcal{E}_i(S)}\frac{\tau}{N-1}(1-E_f)\right\}\bar{E}$$
$$\bar{E}(0)=1.$$

Note that the term corresponding to transfer back to S acts as an inhomogeneity in the absence of which the only solution is $\bar{E}(t) = 1$.

The single observed lineage propagator along branch e of S:

$$\frac{\mathrm{d}}{\mathrm{d}t} G_e = -\{\lambda + \delta(1 - 2E_e)\} G_e \tag{A.3}$$

$$-\left\{ \sum_{f \in \mathcal{E}_i(S)} \frac{\tau}{N-1}(1 - E_f) \right\} G_e$$

$$-\left\{ \frac{N - n_i}{N-1} \tau (1 - \bar{E}) \right\} G_e$$

$$-\left\{ \sigma(1 - \bar{E}) \right\} G_e$$

$$G_e(t, t) = 1.$$

The single observed lineage propagator in an unrepresented species:

$$\frac{\mathrm{d}}{\mathrm{d}t} \bar{G} = -\{\sigma + \lambda\} \bar{G} \tag{A.4}$$

$$-\left\{ \left(\sigma + \delta + \frac{N - n_i}{N-1} \tau \right)(1 - 2\bar{E}) \right\} \bar{G}$$

$$-\left\{ \sum_{f \in \mathcal{E}_i(S)} \frac{\tau}{N-1}(1 - E_f) \right\} \bar{G}$$

$$\bar{G}(t, t) = 1.$$

Note that if we set $\bar{E}(t) = 1$ and neglect gene birth and death, which is much slower than the speciation dynamics, that is, $\delta + \tau + \lambda \ll \sigma$, we recover the exponential probability of coalescence with the represented tree assumed in Equation (2).

The propagator describing the evolution of k gene copies can be expressed using the single gene copy propagator. Consider the expression for $\bar{G}_k(s, t)$, the probability that k genes seen at time s in an unrepresented species all leave a single descendant descending from the copy seen at time $t < s$ in an unrepresented species:

$$\frac{\mathrm{d}}{\mathrm{d}t} \bar{G}_k = -\{\sigma + k\lambda\} \bar{G}_k \tag{A.5}$$

$$-\left\{ \sigma(1 - 2\bar{E}_k) + k \left(\delta + \frac{N - n_i}{N-1} \tau \right)(1 - 2\bar{E}) \right\} \bar{G}_k$$

$$-\left\{ \sum_{f \in \mathcal{E}_i(S)} \frac{k\tau}{N-1}(1 - E_f) \right\} \bar{G}_k$$

$$\bar{G}(t, t)_k = 1.$$

Since Equation (3) implies $-\sigma \bar{G}_k - \sigma(1 - 2\bar{E}_k) \bar{G}_k \simeq \sigma 2(1 - \{\bar{E}\}^k) \bar{G}_k$ and neglecting second- and higher order terms in $1 - \bar{E}$ gives $2(1 - \{\bar{E}\}^k) \simeq 2k(1 - \bar{E})$ and the above can be

written as

$$\frac{\mathrm{d}}{\mathrm{d}t} \bar{G}_k \simeq -k\{\sigma + \lambda\} \bar{G}_k \tag{A.6}$$

$$-k \left\{ \sigma(1 - 2\bar{E}) + \left(\delta + \frac{N - n_i}{N-1} \tau \right)(1 - 2\bar{E}) \right\} \bar{G}_k$$

$$-k \left\{ \sum_{f \in \mathcal{E}_i(S)} \frac{\tau}{N-1}(1 - E_f) \right\} \bar{G}_k,$$

which has the solution $\bar{G}_k = \{\bar{G}\}^k$. Analogous reasoning can be used to show that $E_e(t)$ and $G_e(s, t)$ can be used to factor the respective functions describing the evolution of multiple gene copies.

The Probability of a Gene Tree

The expressions for $P_e(u, t)$ and $\bar{P}(u, t)$ are derived under the approximation that unrepresented speciation and extinction is independent per gene, as discussed above. The full expression for $P_e(u, t + \Delta t_i)$ including terms corresponding to direct transfers and indirect transfers that depart S via a preceding transfer events, which are neglected in Equation (4), is:

$$P_e(u, t + \Delta t_i) = G_e(t + \Delta t, t) P_e(u, t) \tag{A.7}$$

$$+ \{\delta \Delta t_i\} P_e(v, t) P_e(w, t)$$

$$+ \sum_{f \in \mathcal{E}_i} \left\{ \frac{\tau \Delta t_i}{N-1} \right\} P_e(v, t) P_f(w, t)$$

$$+ \sum_{f \in \mathcal{E}_i} \left\{ \frac{\tau \Delta t_i}{N-1} \right\} P_f(v, t) P_e(w, t)$$

$$+ \sum_{f \in \mathcal{E}_i} \left\{ \frac{\tau \Delta t_i}{N-1} \right\} P_f(u, t) E_e(t)$$

$$+ \left\{ \left(\sigma + \frac{N - n_i}{N-1} \tau \right) \Delta t_i \right\} \bar{P}(v, t) P_e(w, t)$$

$$+ \left\{ \left(\sigma + \frac{N - n_i}{N-1} \tau \right) \Delta t_i \right\} P_e(v, t) \bar{P}(v, t)$$

$$+ \left\{ \left(\sigma + \frac{N - n_i}{N-1} \tau \right) \Delta t_i \right\} E_e(t) \bar{P}(u, t),$$

Routes to Cyanobacterial Genomes

The data set was constructed for near universal single-copy genes from all 36 cyanobacterial genomes found in Version 5 the HOGENOM database (Penel et al. 2009). Amino acid sequences were extracted for each family that had a single copy in at least 34 of the 36 cyanobacterial genomes. For each family, sequences were aligned using MUSCLE (Edgar 2004) with default parameters. The multiple alignment was subsequently cleaned using GBLOCKS (Talavera and Castresana 2007)

with the options:

$$\text{“-t=p -b1 50 -b2 50 -b5=a -t=p”.} \quad (A.8)$$

Subsequently, we inferred a gene topology G that maximizes the joint likelihood

$$p_{\text{exODT}}(G|S, \delta, \tau, \lambda, \sigma, N) \times p_{\text{Felsenstein}}(\text{alignment}|G), \quad (A.9)$$

where the first term corresponds to the likelihood of observing the unrooted gene tree topology G according to the exODT model developed above [Eqs. (4) and (5)), while the second term corresponds to the classic Felsenstein likelihood (Felsenstein 1981) of the alignment. For the exODT model, we fixed the parameter values $N = 10^6$ and $\sigma = 2N$, used the dated phylogeny from Szöllősi et al. (2012) and estimated global gene birth and death rates as described below. To calculate the Felsenstein likelihood, we used the Bio++ library (Dutheil et al. 2006) with an LG+Γ4+I model. Alignments and reconstructed gene trees are available from Dryad under doi:10.5061/dryad.27d0g.

Gene trees inference was performed in a two-step approach:

Initial estimate

1. Using the DTL rates $\delta = 1.0 \times 10^{-2}$, $\tau = 1.0 \times 10^{-2}$, $\lambda = 2.0 \times 10^{-2}$ for each family, the joint likelihood was calculated for all nearest neighbor interchanges (NNIs) (Felsenstein 2004) and a move was accepted if it improved the joint likelihood.

2. For the set of trees obtained global DTL parameters were estimated that maximize the product of the joint likelihood of all 473 gene families.

Final estimate

1. Using the obtained DTL rates for each family, the joint likelihood was calculated for all NNIs (Felsenstein 2004) and a move was accepted if it improved the joint likelihood.

2. For the set of trees obtained global DTL parameters were again estimated with the results: $\delta = 1.010 \times 10^{-5}$, $\tau = 4.438 \times 10^{-3}$, and $\lambda = 1.015 \times 10^{-1}$.

Before performing NNIs starting gene tree topologies were estimated using an amalgamation approach (David and Alm 2011) wherein the Felsenstein likelihood was approximated using conditional clade probabilities (Höhna and Drummond 2012) based on posterior sample of 10 000 tree topologies obtained using PhyloBayes using an LG+Γ4+I substitution model.

The support of transfer events was measured based on a posterior sample of 1000 reconciliations per family. For each family, we assessed the support of all transfer events in the reconciliation that was seen the largest number of times. Two transfers were considered equivalent if they involved the transfer of the same gene linage between identical branches of the species tree.

REFERENCES

Abby S.S., 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.

Andam C.P., Gogarten J.P. 2011. Biased gene transfer in microbial evolution. Nat. Rev. Microbiol. 9:543–555.

Boussau B., Daubin V. 2010. Genomes as documents of evolutionary history. Trends Ecol. Evol. 25:224–232.

Brochier-Armanet C., Forterre P. 2007. Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers. Archaea 2:83–93.

Csűrös M., Miklós I. 2009. Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. Mol. Biol. Evol. 26:2087–2095.

David L.A., Alm E.J. 2011. Rapid evolutionary innovation during an archaean genetic expansion. Nature 469:93–96.

Doyon J.-P., Scornavacca C., Gorbunov K., Szöllősi G.J., Ranwez V., Berry V. 2010. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. Berlin (Germany): Springer. p. 93–108 (Comparative genomics; vol. 6398).

Dutheil J., Gaillard S., Bazin E., Glémin S., Ranwez V., Galtier N., Belkhir K. 2006. Bio++: a set of c++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. BMC Bioinform. 7:188.

Edgar R.C. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Falcón L.I., Magallón S., Castillo A. 2010. Dating the cyanobacterial ancestor of the chloroplast. ISME J. 4:777–783.

Felsenstein J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Fournier G.P., Huang J., Gogarten J.P. 2009. Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life. Phil. Trans. R. Soc. Lond. B Biol. Sci. 364:2229–2239.

Galtier N., Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Phil. Trans. R. Soc. Lond. B Biol. Sci. 363:4023–4029.

Heled J., Drummond A.J. 2008. Bayesian inference of population size history from multiple loci. BMC Evol. Biol. 8:289.

Höhna S., Drummond A.J. 2012. Guided tree topology proposals for bayesian phylogenetic inference. Syst. Biol. 61:1–11.

Karev G.P., Wolf Y.I., Rzhetsky A.Y., Berezovskaya F.S., Koonin E.V. 2002. Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol. Biol. 2:18.

Kingman J. 1982. On the genealogy of large populations. J. Appl. Probab. 19:27–43.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46: 523–536.

Moran P.A.P. 1962. The statistical processes of evolutionary theory. Oxford: Clarendon Press.

Morlon H., Potts M.D., Plotkin J.B. 2010. Inferring the dynamics of diversification: a coalescent approach. PLoS Biol. 8:e1000493.

Nee S. 2001. Inferring speciation rates from phylogenies. Evolution 55:661–668.

Nelson K.E., Clayton R.A., Gill S.R., Gwinn M.L., Dodson R.J., Haft D.H., Hickey E.K., Peterson J.D., Nelson W.C., Ketchum K.A., McDonald L., Utterback T.R., Malek J.A., Linher K.D., Garrett M.M., Stewart A.M., Cotton M.D., Pratt M.S., Phillips C.A., Richardson D., Heidelberg J., Sutton G.G., Fleischmann R.D., Eisen J.A., White O., Salzberg S.L., Smith H.O., Venter J.C., Fraser C.M. 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of thermotoga maritima. Nature 399:323–329.

Nesbo C.L., L'Haridon S., Stetter K.O., Doolittle W.F. 2001. Phylogenetic analyses of two "archaeal" genes in thermotoga maritima reveal multiple transfers between archaea and bacteria. Mol. Biol. Evol. 18:362–375.

Ochman H., Lawrence J.G., Groisman E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.

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 Bioinform. 10 (Suppl 6):S3.

Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Natl Acad. Sci. USA 108:6187–6192.

Szöllősi G.J., Daubin V. 2012. Modeling gene family evolution and reconciling phylogenetic discord. Methods Mol. Biol. 856:29–51.

Szöllősi G.J., Boussau B., Abby S.S., Tannier E., Daubin V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc. Natl Acad. Sci. USA. 109:17513–17518.

Talavera G., Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564–577.

Tofigh A. 2009. Using trees to capture reticulate evolution: lateral gene transfers and cancer progression [dissertation]. [Stockholm (Sweden)]: KTH, School of Computer Science and Communication.

Torsvik V., Øvreås L., Thingstad T.F. 2002. Prokaryotic diversity– magnitude, dynamics, and controlling factors. Science 296: 1064–1066.

Zhaxybayeva O., Gogarten J.P. 2004. Cladogenesis, coalescence and the evolution of the three domains of life. Trends Genet. 20:182–187.

Zuckerkandl E., Pauling L. 1965. Molecules as documents of evolutionary history. J. Theor. Biol. 8:357–366.