

Supplementary Information: A Bayesian Method for Analyzing Lateral Gene Transfer

Joel Sjöstrand, Ali Tofigh, Vincent Daubin,
Lars Arvestad, Bengt Sennblad and Jens Lagergren

Contents

1	Supplementary methods	2
1.1	Computational techniques	2
1.1.1	Approximating the probability of a gene tree with lengths	2
1.1.2	Differential equations	3
1.1.3	Dynamic programming	5
1.1.4	Hyperparameter priors and run-time	7
1.2	Generation of synthetic data	7
1.2.1	Generating species trees	7
1.2.2	Generating gene trees	9
1.2.3	Generating sequence data	10
1.3	MCMC analysis	10
1.4	Estimating DLT-parameters	11
2	Supplementary discussion	11
2.1	Tests on synthetic data	11
2.2	Detecting GD and LGT events	12
2.3	Further convergence tests	14
2.4	MrBayes Robinson-Foulds comparison on synthetic data . . .	15
2.5	LGT highways and genomic islands	16
2.6	Functional analysis	18
2.7	Expanding on functionalization theories of LGT	19
3	Supplementary figures	22

1 Supplementary methods

1.1 Computational techniques

PrIME-DLTRS uses a Metropolis-Hastings MCMC framework for inference. During an MCMC iteration, a complex combination of ordinary differential equations (ODEs) and dynamic programming (DP) is used. These computational techniques are outlined below. Please bear in mind that the species tree S and the gene tree G are rooted, bifurcating, with edge directionality as shown in Figure S1B, and also that the time of S is 0 at the leaves and > 0 at interior vertices. A MCMC state also consists of the gene tree edge lengths l , and the model parameters $\theta = (\delta, \mu, \tau, m, cv)$, representing the GD rate, loss rate, LGT rate, and SE edge rate mean and coefficient of variation, respectively. Additionally, this may be extended with a shape parameter α for modeling gamma site rates.

1.1.1 Approximating the probability of a gene tree with lengths

In order to approximate equation (2) in the main text, discretization vertices with out-degree 1 are introduced on the edges of the species tree to facilitate consideration of only discretized realizations, i.e., realizations mapping each gene tree vertex to a regular species tree vertex or a discretization vertex. The discretization is obtained in two steps as follows. First, a new species tree S' is constructed by subdividing any edge contemporaneous to a speciation, at the time of the speciation. Second, all edges of S' are discretized by introducing new vertices at a set of regularly occurring discretization time points, and the resulting discretized species tree is denoted S'' , see Figure S1. Finally, the approximation is given by

$$\sum_{c \in \mathcal{C}} \int_{a \in A(c)} p[G, l, a | \theta] da \approx \sum_{c \in \mathcal{C}} \sum_{d \in \mathcal{D}(c)} p[G, l, d | \theta] \Delta(d). \quad (1)$$

where $\mathcal{D}(c)$ is the set of discretized realizations that are compatible with the reconciliation c . The factor $\Delta(d)$ is the product of the lengths of the intervals in which the discretization points used by d are found, and accounts for that we are approximating integrals over these intervals.

1.1.2 Differential equations

We will now introduce a probability that turns out to be of crucial importance to our DP algorithm. Let a *mortal* be a gene lineage that may or may not yield descendants among the leaves of the species tree. We define $p_{11}(e, x)$ as the probability of a single gene lineage starting at the beginning of edge $e \in E(S'')$ (i.e., at an infinitesimal distance from its tail) and evolving to exactly one mortal at a point in the species tree while producing an undetermined number of lineages destined to go extinct.

The following two events, and their associated probabilities, are key to the computation of p_{11} :

Extinction: A single gene lineage (and any child lineages) going extinct.

Single mortal: Obtaining a single (pruned) gene lineage between two points of the species tree.

They can be formulated using systems of ODEs, and can be solved numerically for all vertices of S'' , as explained below.

The edges of S' can be partitioned into sets of contemporaneous *edge generations*, see Figure S1. ODE calculations are carried out separately for each edge generation of S' . For an edge $e \in E(S')$, we will use $\mathcal{G}_{S'}(e)$ to denote the edges of e 's generation, excluding e itself.

We now consider the two probabilities in more detail. For the extinction probability, consider an edge generation containing $e \in E(S')$. Let $Q_e(t)$ be the probability that a single lineage starting at time t on edge e will go extinct, i.e., have no descendants which reach the leaves of S' . Using standard techniques for Poisson processes, a system of ODEs for $Q_e(t)$ can be derived as follows. For an infinitesimal interval $[t, t - h]$, the lineage can either be exposed to one of three distinct events or no event at all (main text Figure 1B). Then,

1. *If a duplication occurred in $[t, t - h]$:* Both of the resulting lineages in e at $t - h$ must go extinct.
2. *If a lateral transfer occurred in $[t, t - h]$:* The original lineage in e at $t - h$ as well as the lineage transferred to the uniformly chosen edge f at $t - h$ must go extinct.
3. *If a loss occurred in $[t, t - h]$:* The original lineage has gone extinct.

4. *If no event occurred in $[t, t - h]$:* The original lineage in e at $t - h$ must go extinct.

Applying this argument, performing standard rearrangements, and then taking a limit yield:

$$\begin{aligned} \frac{d}{dt}Q_e(t) &= \delta(Q_e(t))^2 + \tau \left(\sum_{f \in \mathcal{G}_{S'}(e)} \frac{1}{|\mathcal{G}_{S'}(e)|} Q_e(t) Q_f(t) \right) \\ &+ \mu - (\delta + \tau + \mu) Q_e(t) \end{aligned} \quad (2)$$

The initial values of each system depend on the system of the previous generation, enabling the systems to be solved consecutively from the leaves to the root of S' . More specifically, for the initial time t of the considered generation, we have:

$$Q_e(t) = \begin{cases} 0 & \text{if the head of } e \text{ is a leaf at } t = 0, \\ Q_f(t) & \text{if the head of } e \in V(S') \setminus V(S) \text{ and has the} \\ & \text{single outgoing edge } f \text{ at } t > 0, \\ Q_f(t)Q_g(t) & \text{if the head of } e \text{ is a speciation at } t > 0, \text{ with the} \\ & \text{two outgoing edges } f \text{ and } g. \end{cases}$$

The stem generation can be solved analytically [1], because in the model no LGTs can occur there.

Recall, we define $p_{11}(e, x)$ as the probability of a single gene lineage starting at beginning of edge $e \in E(S'')$ (i.e., at the tail) and evolving to exactly one mortal at vertex $x \in V(S'')$ while producing an undetermined number of lineages destined to go extinct. To derive an expression for $p_{11}(e, x)$, first consider two edges $e, f \in E(S')$ of the same edge generation in S' . For times $s > t$, define $Q_{ef}(s, t)$ as the probability of starting on e at time s and having one single mortal on f at time t , i.e., all other descendants are destined to go extinct. Using standard techniques, we obtain

$$\begin{aligned} \frac{d}{ds}Q_{ef}(s, t) &= 2\delta Q_e(s)Q_{ef}(s, t) \\ &+ \tau \left(\sum_{g \in \mathcal{G}_{S'}(e)} \frac{1}{|\mathcal{G}_{S'}(e)|} (Q_{gf}(s, t)Q_e(s) + Q_{ef}(s, t)Q_g(s)) \right) \\ &- (\delta + \tau + \mu)Q_{ef}(s, t), \end{aligned} \quad (3)$$

with initial values according to

$$Q_{ef}(t, t) = \begin{cases} 1 & \text{if } e = f, \\ 0 & \text{otherwise.} \end{cases}$$

This enables us to compute $p_{11}(e, x)$ for any valid pair of vertices in $V(S'')$ that are not, time-wise, separated by a speciation somewhere in S'' .

Now, consider the case with two connected edges $e, f \in E(S')$, such that e is the incoming edge and f is the outgoing edge of $v \in V(S') \setminus V(S)$ at time $t(v) = T$. Let $w \in V(S)$ be the speciation at time T , and let $g \in E(S')$ be the incoming edge of w and let $g', g'' \in E(S')$ be the outgoing edges of w . We obtain

$$\begin{aligned} Q_{ef}(s, t) &= Q_{eg}(s, T) \left(Q_{g'f}(T, t) Q_{g''}(T) + Q_{g''f}(T, t) Q_{g'}(T) \right) \\ &+ \sum_{h \in \mathcal{G}_{S'}(g)} Q_{eh}(s, T) Q_{hf}(T, t). \end{aligned} \quad (4)$$

By recursively applying equations (3) and (4), we can now proceed to compute $p_{11}(e, x)$ for any valid pair e and x of S'' . For $e' = \langle x, y \rangle \in E(S')$ and $e'' \in E(S'')$, we say that e' *captures* e'' if there is a path in S'' from x to y that includes e'' . Let y be the tail of edge $e \in E(S'')$ and let x be the head of edge $f \in E(S'')$. Then,

$$p_{11}(e, x) = Q_{e', f'}(t(y), t(x)),$$

where e', f' are the edges in S' that capture e and f , respectively.

1.1.3 Dynamic programming

We now describe a DP algorithm for computing the probability density of the gene tree G with lengths l . It is based on the approximation (1) derived above.

For a vertex $x \in V(S'')$ and a gene vertex $u \in V(G)$, define $a(x, u)$ to be the probability density of the subtree of G rooted at u given that the event creating u occurred at x . Similarly, for an edge $e \in E(S'')$ and vertex $u \in V(G)$, define $s(e, u)$ to be the probability density that a single lineage starting on e , infinitely close to its tail, generates the incoming edge of u and the subtree of G rooted at u . These two probability densities can be computed using the recursions below.

1. If x is a speciation, then

$$a(x, u) = s(e, v)s(f, w) + s(e, w)s(f, v),$$

where $e, f \in E(S'')$ are the outgoing edges of x and $v, w \in V(G)$ are the children of u .

2. If x is not contemporaneous with any speciation, then

$$\begin{aligned} a(x, u) = & \left(\sum_{f \in \mathcal{G}_{S''}(e)} \frac{\tau(s(e, v)s(f, w) + s(e, w)s(f, v))}{|\mathcal{G}_{S''}(e)|} \right. \\ & \left. + 2\delta s(e, v)s(e, w) \right) \Delta(x), \end{aligned}$$

where $e \in E(S'')$ is the outgoing edge of x ; $v, w \in V(S'')$ are the children of u ; $\mathcal{G}_{S''}(e)$ denotes the edges of the generation of e in S'' excluding e itself; and $\Delta(x)$ is the length of the discretization interval associated with x . The two terms in the equation correspond to a GD event and a LGT event, respectively. We define $a(x, u)$ to be zero in the remaining cases, i.e., (i) when x or u is a leaf and (ii) when x has out-degree 1 and is contemporaneous with a speciation.

3. For an edge $e = \langle x, y \rangle \in E(S'')$ and a leaf $u \in V(G)$, which we assume to belong to the extant species $z \in V(S)$ and have parent $pa(u) \in V(G)$, we have

$$s(e, u) = p_{11}(e, z) \rho \left(\frac{l(pa(u), u)}{t(x)} \right),$$

where $t(x)$ is the time of x and ρ is the density function for our R submodel.

4. For an edge $e = \langle x, y \rangle \in E(S'')$ and a vertex $u \in V(G)$ with parent $pa(u) \in V(G)$, we have

$$s(e, u) = \sum_{z \in \mathcal{R}(x)} p_{11}(e, z) \rho \left(\frac{l(pa(u), u)}{t(x) - t(z)} \right) a(z, u)$$

where $t(x)$ is the time of x , $t(z)$ is the time of z , and $\mathcal{R}(x)$ is the set of all vertices $z \in V(S'')$ associated with a more recent time (i.e., closer to the leaves) than x .

Following the computations of all $a(x, u)$ and $s(e, u)$, which are performed from the leaves to the root for both G and S'' , we obtain an approximation of $p[G, l|\theta]$ in $s(f, r)$, where f is the most ancient edge of $E(S'')$ and r the most ancient vertex of $V(G)$. $s(f, r)$ constitutes the last computed element in the DP, and will hold the contribution from all discretized realizations.

1.1.4 Hyperparameter priors and run-time

The parameters in $\theta = (\delta, \mu, \tau, m, cv)$ were each assigned uniform priors over sufficiently large intervals. Although there are rare examples of rate estimates in the literature, these were judged insufficient for crafting informative priors. The uniform intervals used were $(0, 1]$ events per MY for δ, μ and τ ; $(0, 3]$ substitutions per MY and site for m ; and $(0, 10]$ for cv . Note that as empirical substitution models were used throughout, substitution model parameters were not part of the inference (except in some cases α with a uniform prior over $(0, 10]$).

As with most MCMC methods, PrIME-DLTRS run-time is highly reliant on the characteristics of the data and not only the size. The signal conveyed by the data and the posterior landscape impact the time needed for convergence, which should be verified with suitable diagnostics. In any case, assuming an upper bound k on the number of discretization points of edge generations, $|V(S)| = m$ and $|V(G)| = n$ will affect run-time more than alignment length in most cases, yielding a time complexity of roughly $O(k^2 m^4 n)$. As an indication of practical time requirements, running a *Mollucutes* gene family with $m = 14$, $n = 20$ and $k = 5$ for all edge generations for $1 \cdot 10^6$ iterations on a Mac OS X 2,4 GHz Intel Core 2 Duo requires roughly 9 hours, using Oracle's HotSpot Java VM v. 1.6.

1.2 Generation of synthetic data

1.2.1 Generating species trees

To achieve a biologically relevant simulation, without focusing on a few select species trees (which may have specific properties not generalizable to the entire set), we generated synthetic species trees using a birth-death process with parameters drawn from a posterior distribution obtained in an analysis of biological data. We used that of Linder et al. [14], which is a Bayesian estimation of divergence times from sequence data from the plastid gene *rbcL*

on a well-supported tree for 79 species of flowering plants (Angiospermae; cf. Figure S12). The *rbcL* gene tree is assumed to coincide precisely with the species tree, which indeed is a reasonable assumption for plastid genes. An MCMC analysis under an integrated *iid*-gamma model for substitution rates and a birth-death model for the *rbcL* tree evolution including divergence times yielded a posterior distribution over the model parameters. From this distribution, we sampled birth and death parameters for our species tree generation.

Our generation procedure was repeated for two subtrees of the full angiosperm species tree, namely the subtrees rooted at vertex 84 (cf. Supplementary Figure S12), and at vertex 91; we will refer to these subtrees as S_{84} , and S_{91} . The subtrees have 6 and 11 leaves, respectively.

For each subtree, S_i ($i \in \{84, 91\}$), we generated 100 synthetic species trees as follows.

1. We sampled parameters from the posterior distribution by drawing a random iteration from the MCMC output file. The parameters sampled were birth rate λ'_i , death rate μ'_i , and the divergence time, t'_i , of the root of S_i .
2. The sampled parameter values relate to a time scale where the root divergence time of S is 1.0. To obtain values that relate to a time scale measured in million years (MY), we rescaled the sample parameter values using a rough estimate of the root divergence time of S , 130 MY ago, taken from the analysis of Magallon and Castillo [15]. The parameters were rescaled as follows:

$$\begin{aligned}\lambda_i &= \lambda'_i/130 \\ \mu_i &= \mu'_i/130 \\ t_i &= 130t'_i\end{aligned}$$

3. We generated a new tree under a homogeneous birth-death model with the sampled parameters λ_i and μ_i , over a time interval with length t_i . The tree was constrained to have the same number of leaves as S_i .

In this way we obtained two sets, each comprising 100 species trees.

1.2.2 Generating gene trees

We generated synthetic gene trees in accordance with the DLT submodel, using the synthetic species trees created earlier (see Section 1.2.1). The process is described below.

First, for any tree T , let $t(u)$ be the divergence time of vertex $u \in V(T)$; again, using the convention that $t(u) = 0$ at leaves and $t(u) > 0$ at interior vertices. Similarly, let $r(T)$ be the root of T , and let $p(u)$ be the parent of u in T .

Using a species tree, S , we generated an embedded gene tree with divergence times associated to its vertices. To allow for duplication events predating $r(S)$, we set the time interval for the stem edge predating the root to have a length ψ of 10% of the root-to-leaf time. The gene tree was produced with duplication, loss and transfer rates δ, μ and τ , respectively, exactly following our model, as follows:

1. Let $x = r(S)$ and let $t = t(r(S)) + \psi$.
2. A single lineage u at time t evolves down the incoming edge of x .
3. A waiting time ω is drawn from an exponential distribution with parameter $\delta + \tau + \mu$. Four exclusive events may occur, as follows:
 - *Speciation*: If $t - \omega \leq t(x)$, let u evolve down to x . If $x \in L(S)$, stop the process. Otherwise let y, z be the children of x and split u into two new lineages, v and w . Start two new independent processes from (2) with u, x, t set to $v, y, t(x)$ and $w, z, t(x)$ respectively.
 - *Non-speciation event*: If $t - \omega > t(x)$, pick an event type GL, GD, or LGT, with probability $\mu/(\delta + \tau + \mu)$, $\delta/(\delta + \tau + \mu)$, and $\tau/(\delta + \tau + \mu)$, respectively, and given the event proceed as follows:
 - *Loss*: If the event is GL, then stop the process and mark u as lost.
 - *Duplication*: If the event is GD, then split u into two new lineages, v and w . Start two new independent processes from (2) with u, x, t set to $v, x, t - \omega$ and $w, x, t - \omega$ respectively.
 - *Transfer*: If the event is LGT, uniformly pick a random, different edge from the generation of edges in S contemporaneous with t . Split u into two new lineages, v and w . Start two new

independent processes from (2) with u, x, t set to $v, x, t - \omega$ and $w, z, t - \omega$ respectively.

Finally, prune away all lost lineages. For the stem, the above process is corrected so that only duplications and losses are allowed, as there are no edges to which transfers can occur.

1.2.3 Generating sequence data

In accordance with the R submodel, we used *iid* gamma distributed edge-specific substitution rates for each generated gene tree G_i . To obtain relevant values for the model parameters mean m and coefficient of variation cv , we used the substitution rate distribution estimated in Åkerborg et al. [1] for the GSR model for 4,809 yeast gene families. This led us to use

$$\begin{aligned}\hat{m} &= 0.0019 \\ \hat{cv} &= \sqrt{9.0 \times 10^{-7}}/m\end{aligned}$$

as parameters for the gamma distribution of *iid* edge rates.

For each generated gene tree, edge lengths were then computed by multiplying the edge-specific rates and times. Let edge e in gene tree G_i have length, rate and time, $\ell_{i,e}$, $r_{i,e}$ and $t_{i,e}$, respectively, then

$$\ell_{i,e} = r_{i,e} t_{i,e}.$$

Finally, we used the JTT amino acid substitution model as S submodel and generated sequence data accordingly.

1.3 MCMC analysis

Cf. main text Methods for the MCMC framework.

During an iteration, systems of ODEs were solved using a Runge-Kutta numerical solver based on [7]. It makes use of Dormand-Prince parameters of order 4 (order 5 for error estimator), automatic step size control with the initial step size set to an Euler step, and simple stiffness detection. Solver tolerances were set *a priori*, based on its effect on likelihood computation in pilot studies.

The MCMC proposal distributions were tuned so as to achieve roughly 0.2–0.5 acceptance probability for singleton parameter changes. Acceptance probabilities for tree changes are exemplified in Figure S18.

A small number of families were removed based on erratic behaviour due to being too small, having an insignificant sequence length, or having identical sequences. We also excluded families where the most probable posterior gene tree topology, \hat{G}_{MAP} , accounted for less than 5% of the posterior probability.

1.4 Estimating DLT-parameters

As illustrated in Figure S17, the marginal posterior rate distributions of families are peaked, suggesting that the wide uniform prior’s effect is of limited concern. In general, the variance of the GD rate and the LGT rate distributions are similar, whereas the GL rate distribution is slightly more dispersed. For GD and LGT rates, MAP estimates for each family was obtained using kernel density smoothing with the CRAN R package NP (<http://cran.r-project.org/web/packages/np/>) [8] on samples in high density regions. Trials using different kernel types and bandwidth settings had little impact in most cases, and a Gaussian kernel with Scott’s rule-of-thumb bandwidth was ultimately used for computational feasibility. Boundary correction for sparser sampling close to the origin was carried out by sample mirroring. This smoothing technique is conservative w.r.t. rate estimates in that over-smoothing typically results in zero values.

To enhance presentation, we formally define for GL, GD and LGT: *low rate* $< 1/T \leq$ *high rate*, where T is the total time span of the species tree; the intuition is that $1/T$ corresponds to the maximum likelihood rate estimate given one event over the whole tree.

2 Supplementary discussion

2.1 Tests on synthetic data

We performed a number of analyses using our synthetic datasets to evaluate the capacity of PRIME-DLTRS. The accuracy of the GD and LGT rate estimates was tested on synthetic data and compared with the generated rates as well as an informed maximum likelihood (Informed ML) estimation of the GD and LGT rates directly from the generated unpruned realizations (i.e., the unpruned gene trees with all events and their timings known). The Informed ML estimates, favoured by the extra information, provided an upper

bound on how well the parameters can be estimated (we note also that the generated trees were constrained to include at least two species and at least four gene members). The test was carried out with different total birth rates (i.e., $\delta + \tau$) and repeated with δ accounting for different proportions of the total birth rate. The results, shown in Figures 2 and S5, demonstrate that the estimates were very good except in the case of very high total birth rate; both the PRIME-DLTRS MAP and the Informed ML estimates lie close to the generated rates (Figure 2) and the discrepancy between Informed ML and PRIME-DLTRS estimates is very low (Figures S5 and S16). For a very high total birth rate (0.05 events per MY), multiple LGTs affecting the same gene lineage becomes common, and as these may alternatively be explained by fewer GD or LGT events, the PRIME-DLTRS rate estimates tend to be too low. This is also reflected in the posterior distribution, where the space of reasonable candidate topologies becomes vast. The Informed ML estimates, for which the actual realizations were available, had considerably better accuracy. We note that the experimental data share several characteristics with the synthetic data, e.g., the distribution shape over PRIME-DLTRS MAP trees are similar, see Figure S6. Moreover, the rate estimates for the experimental data (see, e.g., Figure S7) lie in the interval close to the lower ones in Figure 2 (in fact, $\delta + \tau$ is always less than, and in the majority of cases much less than, 0.003).

2.2 Detecting GD and LGT events

Any analysis can be affected by events not explicitly taken into account. In our case, incomplete lineage sorting (ILS) can under some circumstances be predicted as GD. Moreover, homologous recombination is believed to be a frequent mediator of genetic material within a population, or between closely related populations, and it is relevant to ask how such events affect our analysis. However, ILS and homologous recombination never increase the copy number of a gene in a genome. A majority of the gene families that our analysis predicts to have been exposed to GD has at least two copies in at least one genome (Mollicutes: 58%, Cyanobacteria: 68%), which corroborates our GD rate estimates (Figure S2). Notice, moreover, that if a gene family starts with a single member and the total birth rate equals the loss rate, i.e., $\delta + \tau = \mu$, then we can expect each extant species to have exactly one copy [12]. That is, under these reasonable hypotheses, GDs together with GLs are expected to yield monolog gene families. We conclude that ILS

and homologous recombination are unlikely to have an impact on our results.

As mentioned in the main text, LGT in our setting encapsulates transfer of genes between species (for a different, operational, definition, cf. Treangen and Rocha [20]). A transfer *within* a population appears as a GD in our model, which is sound since it may be conceived as vertical inheritance within the species. One can also note that its effect on gene tree topology is the same as that of a duplication. The above examples concern events taking place only in the species lineages included in the species tree, i.e., species that are ancestral to the analysed extant taxa. LGTs from external species (e.g., species that are ancestral to unsampled or extinct taxa) into the analysed tree may give rise to a variety of scenarios. Some such LGTs may incorrectly be classified as GDs (e.g., Figure S3). Below, we obtain a rough bound on the frequency of misclassification stemming from transfer from unsampled species when the donor-receiver is distantly related. We make use of what we here call *highway-inducing topologies* (HITs). These are 6 gene tree topologies (Figure S4), besides that of the species tree, that were supported by at least 10 Cyanobacteria monolog families (no Mollicute topology was supported by more than 3 monologs). For each such HIT, we considered the gene families supporting it to have been affected by one or more potential LGT *highways* [5] (HWs; see further Section 2.5) from external species.

A HW affects several gene families, and a potential misclassification of LGT events, as illustrated by the scenarios in Figures S3A and S3C, could therefore have a large, systematic impact on our rate estimates. Of the HITs in Figure S4, only HIT 1 shows this pattern with systematically high GD rates and low LGT rates (Table S2). To obtain a bound on the impact on our rate estimates, we assume that (i) all inferred instances of high GD rate in HIT 1 actually are misclassified instances of high LGT rate from external species, (ii) that the frequency of misclassification is not higher in general than among gene families supporting HITs (in fact, a higher frequency of misclassification is expected among HITs), and (iii) that each of the misclassified gene families at most contribute to the mean rate estimates with a rate equalling the mean GD rate (in fact, all contribute a lower rate, cf. Figure S4). Using the results in Table S2 and main text Figure 4, we can then conclude that the adjusted mean LGT rate is at most

$$\frac{8.3 \times 10^{-5} + 21/46 \times 3.6 \times 10^{-4}}{1 + 21/46} = 1.7 \times 10^{-4}.$$

Even with these very generous corrections, the mean GD rate (3.6×10^{-4}) is

substantially higher ($> 2\times$) than the adjusted mean LGT rate (1.7×10^{-4}). This argument is insensitive to the HIT threshold. In other words, this corroborates our finding that the combined rate of true duplications and transfers between closely related species clearly surpass the rate of transfer between distant species.

2.3 Further convergence tests

Although we have shown in tests on synthetic data that the algorithm is well-behaved, any MCMC-based method risks becoming trapped in local optima. This can have an impact on our claim that GD rates are higher than the LGT rates (cf. main text and main text Figure 4). To rule out underestimation of LGT rates and/or overestimation of GD rates, we tested if rates were correctly inferred by the PRIME-DLTRS MCMC algorithm.

We find it unlikely that our MCMC algorithm has systematically exclusively underestimated LGT rates, because low rates imply that few LGT events are needed. If there is a bias towards underestimation of LGT events, then that would be balanced by more GD events and hence overestimated GD rates. Another potential source of high GD rate estimates is if the algorithm has become stuck in a local optimum on an incorrect gene tree, in which case its reconciliation is expected to imply more GD events than necessary, which in turn gives overestimated GD rates. To assess possible rate bias, we wanted to test whether we were overestimating GD rates, with or without underestimating LGT rates. Hence, we focused our investigation on gene families in the third (Q3) and fourth (Q4) quartiles of Figure S7; in which most of the contribution to the average GD rate is found (Q4 has a considerable contribution, and the points found in Q3 and Q4 contributes essentially everything).

For Mollicutes and Cyanobacteria, we randomly selected (i) 40 families with high GD and low LGT rate and (ii) 40 families with low GD and high LGT rate estimates (see supplementary methods). In order to validate the initial rate estimates, each family was reanalysed twice with different constraints on the high rate type. In the first analysis, *high-fixed*, the high rate type was fixed to its estimated high value; in the second analysis, *zero-fixed*, the high rate type was fixed to 0. No constraint was imposed on the low rate type and the gene tree topology. For each family, the unnormalized MAP joint density was then compared between the two constrained analyses to see how often the high-fixed analysis yields a higher value than the zero-fixed

analysis, which confirms that the original high rate estimate is correct. If one of the two analyses fails to converge, its density was considered to be zero, whereas if they both failed to converge, the family was excluded. Observe that if a parameter is fixed to a value not suited for the data, it is more likely that the MCMC will fail to converge. With the exception of the case of low GD rate and high LGT in Mollicutes, the new MCMC analysis confirmed the original estimate for a majority of the cases, while for the remaining cases the improvement was only marginal, indicating a flat posterior for these cases, see Table S1. In addition, upon manual inspection of the original runs, there was considerable agreement of parallel chains in favour of the original rate classification, indicating that a zero-fixed MAP improvement may not necessarily suggest improper mixing in these cases.

Lastly, the sequence length appears to be sufficient for the analysis. That is, longer MSA (multiple sequence alignment) gives a more pronounced mode for smaller families, but otherwise shorter MSA gives approximately the same distribution of MAP values as the longer ones.

2.4 MrBayes Robinson-Foulds comparison on synthetic data

As an additional test of the species tree’s impact on inference accuracy, we compared the Robinson-Foulds (RF) symmetric tree distance of PRIME-DLTRS and MRBAYES with respect to synthetic data produced with rates on par with those of the biological data. Specifically, we generated 100 gene trees on the cyanobacterial species tree with the following parameters (in Myr^{-1}): $\delta = 3 \cdot 10^{-4}$, $\mu = 4 \cdot 10^{-4}$, $\tau = 1 \cdot 10^{-4}$, $m = 4 \cdot 10^{-4}$, $cv = 2.5 \cdot 10^{-4}/m$. 250 amino acids long protein multiple sequence alignments of the trees were created with the JTT substitution model, and this model was also used during inference. PRIME-DLTRS analysis was carried out with $1 \cdot 10^6$ iterations, sampling every 100th iteration, with the first 25% samples removed as burn-in. MRBAYES was run with 4 chains, $4 \cdot 10^6$ iterations, sampling every 200th iteration, with the first 25% samples discarded as burn-in. 9 families failed to be analyzed in MRBAYES, and were discarded, as was 1 family that failed to converge properly in PRIME-DLTRS. The unrooted RF distance between the maximum probability (MAP) inferred gene tree and the true synthetic gene tree was computed for each family. Figure S15 shows the distribution of RF distances, corroborating the results of the RF analysis

on biological monologs, namely that species-unaware methods are more likely to overestimate the true number of duplication and transfer events.

2.5 LGT highways and genomic islands

A LGT highway (HW) [5] can be thought of as some process or event that enhanced LGT events between the involved species. The HITs identified above (Figure S4), which suggest potential HWs in the Cyanobacteria species tree, are based on analysis of monolog gene families only. To further investigate the prevalence of HWs, we extended this analysis to include also non-monolog gene families.

As of yet, PRIME-DLTRS does not produce reconciliations of gene trees, and we therefore adopt the following strategy building on the parsimony-based reconciliation program PHYLTR [19]. We will focus our attention on gene families where GD has had no effect as indicated by a negligible GD rate. This is partly because LGT is the relevant event for HWs and partly to avoid problems with an inflated number of reconciliations due to the trade-off between LGT and GD. This subset of gene families are identified in the following way.

For each gene family, we can obtain reconciliations with LGT only by applying an infinite duplication weight in PHYLTR. Let e_τ denote the number of LGT events as estimated using PHYLTR, let δ^* and τ^* be the MAP GD and LGT rate respectively estimated by PRIME-DLTRS, and let T_G be the total time of the gene tree. It can be shown that e_τ/T_G is the ML estimate of τ^* [11]. Using this result, it can be shown that requiring that $\delta^* \leq 0.1\tau^*/e_\tau$ is a conservative method for identifying gene families where GD has had no effect. Since our conclusion will be based on the gene tree, we also require the posterior probability of the gene tree to be higher than 0.90. We then searched for common LGT patterns in the PHYLTR reconciliations of our selected set of gene families. An LGT shared by at least 10 gene families is considered to be a HW.

The HWs we identified (main text and main text Figure 3B) also correspond well to LGT events in HITs 2, 3, 4 and 5, as predicted using PHYLTR. The GO-terms (Section 2.6) associated with the gene families of these HWs include DNA-repair, amino acid biosynthesis, nucleotide-binding, and ion-binding. Of particular interest is the *ccmM* gene that codes for the carbon dioxide concentrating mechanism protein CcmM, which is an important component of the β -Cyanobacteria photosystem [17, 3]. The LGTs of *ccmM* that

have occurred are $\beta_{hs} \Leftrightarrow \beta_t$ and either $\beta_{hs} \Leftrightarrow \beta_{ff}$ or $\beta_t \Leftrightarrow \beta_{ff}$ (two alternative scenarios).

Earlier quartet-based analyses of LGT in Cyanobacteria monologs [22, 4], have suggested potential LGTs corresponding to $\beta_{ff} \Leftrightarrow \beta_t$ and $\alpha \Leftrightarrow \beta_{ff}$ (although the support appear weaker than in our analysis, e.g., 1–5 monologs displaying LGT for $\beta_{ff} \Leftrightarrow \beta_t$ and only 1 monolog displaying LGT for $\alpha \Leftrightarrow \beta_{ff}$ [22]). While not the focus of our analysis, we found evidence of HWs within α , 12 families, and within β_{ff} , 23 families, corroborating earlier results [4, 22].

Moreover, to investigate if there are some genomic traces of LGT for these HWs, we performed additional analyses. We first tested if genes involved in our HWs appeared co-localized on two Cyanobacteria genomes (*Nostoc sp.* PCC 7120 and *Synechocystis sp.* PCC 6803) using the UCSC Microbial Genome Browser (<http://microbes.ucsc.edu/>). There was no apparent major co-localization of highway gene families, cf. Figure S11. This is not unexpected given the great evolutionary age of the LGT events found here, i.e., genome rearrangements are likely to have obscured any co-localization pattern.

As a further attempt to capture traces of LGT in the genomes, we investigated enrichment of *genomic islands* (GI)—genomic regions predicted to be associated with LGT using sequence composition or comparative genomics methods—among our highway monologs. We classified each family as either GI-positive or GI-negative on the basis of having one or more extant member in a putative GI region as classified by ISLANDVIEWER (<http://www.pathogenomics.sfu.ca/islandviewer/>) [13]. We found no significant enrichment of GIs among our highway monologs (Table S3); in fact the odds ratio (OR) of 0.39 rather indicates a depletion of GIs. This again suggests that genomic traces of LGTs in our study have been obscured and that these HWs would be difficult, if not impossible, to detect using sequence composition methods.

We also tested enrichment for GI among gene families with high or low rates, respectively (see supplementary methods for the definition of high/low). We did not find any significant enrichment for GIs related to LGT rates: For Mollicutes there appeared to be a potential trend towards enrichment of GIs in the high LGT class and depletion in the low LGT class, but this did not reach significance, possibly because of insufficient sample size. However, for the Cyanobacteria, there is no enrichment trend at all with respect to LGT rate (Table S3). By contrast, there is a clear trend

towards enrichment of GIs for the high GD rate classes and depletion in the low GD rate class, both in Mollicutes (OR \approx 1.5 and 0.64, respectively) and Cyanobacteria (OR \approx 1.6 and 0.5, respectively). However, again possibly due to insufficient sample size in Mollicutes, this only reached significance in Cyanobacteria (Table S3). Thus, while these results partly corroborates the hypothesis of boosted evolutionary activity at GI regions, in our data this activity appears related to GD rather than LGT.

In conclusion, we found robust evidence for 4 major HWs, that are either new or more strongly supported than with earlier parsimony-based, phylogenetic LGT detection methods, and, moreover, most likely not detectable at all with methods based on synteny or sequence composition.

2.6 Functional analysis

To evaluate potential functional implications of our results, we classified our gene families into six (partly overlapping) classes according to two criteria. The first criteria distinguished four classes that are determined by their relative duplication and transfer rate estimates (see Section 1.4), as follows: *lodup.lotrans* – both duplication and transfer rates are low, *lodup.hitrans* – low duplication rate and high transfer rate, *hidup.hitrans* – high duplication rate and high transfer rate, and *hidup.lotrans* – high duplication rate and low transfer rate. The second criteria distinguished two classes that are determined by their evolution relative the species tree. They are as follows: *sptree* – gene tree coincides with species tree, and *hw* – gene families associated with any of the highways identified above, Table S6 and main text Figure 3B.

We then used two approaches to investigate enrichment of functional terms in these six classes. In the first, we studied enrichment of functional terms from the COG database (<ftp://ftp.ncbi.nih.gov/pub/COG/COG/>) [18]. In the second approach, we studied enrichment of functional terms from GENE ONTOLOGY (GO; <http://www.geneontology.org/>) [2] using the DAVID bioinformatics database web service (<http://david.abcc.ncifcrf.gov/>) [9] (and the GO term subsets GOTERM_BP_FAT, GOTERM_MF_FAT and GOTERM_CC_FAT defined in DAVID). In both cases, we used Fisher’s exact test to evaluate the significance of the enrichment of functional terms in the class gene family set compared to the full gene family set. The reported p-values are corrected for multiple testing, and we used a nominal p-value threshold of 0.05 for significance.

The result of the analyses are shown in Table S7. The COG analysis

showed significance only for the *sptree* class, while enriched GO terms were found for the *lodup.lotrans*, *lodup.hitrans* and *sptree* and classes. The *sptree* class was highly significantly enriched for the COG term “J: TRANSLATION, RIBOSOMAL STRUCTURE AND BIOGENESIS” in the COG class “INFORMATION STORAGE AND PROCESSING” both for Cyanobacteria and Mollicutes. The enrichment of the *sptree* class for translation-related functional terms is also evident in the result from the GO analysis (Table S7), where ribosome- and translation-related terms dominate the significantly enriched terms found for this class, both in Cyanobacteria and Mollicutes. It is also evident that the GO classifications of the *sptree* and *lodup.lotrans* classes are strongly correlated. These results are in line with previously published analyses and lend support to the complexity hypothesis [10], which posits that informational genes, typically members of large complex systems, are less prone to LGT than other, operational genes.

2.7 Expanding on functionalization theories of LGT

In the main text we outlined the need for more elaborate functionalization theories of LGT, similarly to what is already established for duplicated genes. Below, we briefly present some preliminary comments on what such theories need to account for.

Consider first the case of combining advantageous mutations from several lineages. In this case, the effect of intra-/close-species LGT by non-homologous recombination is analogous to the standard functionalization theory for GD, although the biological mechanism yielding the extra copy is a transfer between two individuals genomes rather than a duplication within one individual’s. For the case of eliminating the effect of deleterious mutations, the most interesting scenario starts with a deleteriously mutated gene in the receiver genome, followed by a transfer of a gene with preserved function. If the deleterious mutation is widespread, this may yield a selective advantage for the receiving individual and cause fixation of its genotype in the population; an effect that can be compared to a *selective sweep* in eukaryote allele evolution. This idea needs to be reconciled with the observed exchange of genetic material of the pan-genome of closely related species, whereby a vast genetic diversity can be harboured [6]. It is important to mention also that a large proportion of LGT events occur as the result of homologous recombination that may directly replace an existing region with a potentially advantageous equivalent [21] (from the perspective of our model, this

is a LGT event that induces an immediate obligate loss). As a consequence of homology, such events may be selectively neutral or nearly neutral to the recipient. The functional differences of homologous and non-homologous transfer between closely related species need to be properly accounted for by a proposed theory.

References

- [1] Ö Åkerborg, B Sennblad, L Arvestad, and J Lagergren. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 106(14):5714–9, 2009.
- [2] M Ashburner, CA Ball, and JA Blake. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
- [3] MR Badger, GD Price, BM Long, and FJ Woodger. The environmental plasticity and ecological genomics of the cyanobacterial CO₂ concentrating mechanism. *J. Exp. Bot.*, 57(2):249–265, 2006.
- [4] MS Bansal, G Banay, JP Gogarten, and R Shamir. Detecting highways of horizontal gene transfer. *J. Comput. Biol.*, 18(9):1087–1114, 2011.
- [5] RG Beiko, TJ Harlow, and MA Ragan. Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.*, 102(40):14332–14337, 2005.
- [6] JP Gogarten and JP Townsend. Horizontal gene transfer, genome innovation and evolution. *Nature reviews. Microbiology*, 3(9):679–87, September 2005.
- [7] E Hairer, SP Nørsett, and G Wanner. *Solving ordinary differential equations I: nonstiff problems*. Springer, 1993.
- [8] T Hayfield and JS Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.
- [9] Da Wei Huang, BT Sherman, and RA Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44–57, 2009.

- [10] R Jain, MC. Rivera, and JA Lake. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U.S.A.*, 96(7):3801–3806, 1999.
- [11] N Keiding. Maximum likelihood estimation in the birth-and-death process. *The Annals of Statistics*, 3(2):363–372, 1975.
- [12] DG Kendall. On the generalized ”birth-and-death” process. *Ann. Math. Stat.*, 19:1–15, 1948.
- [13] MGI Langille and FSL Brinkman. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics (Oxford, England)*, 25(5):664–5, March 2009.
- [14] M Linder, T Britton, and B Sennblad. Evaluation of Bayesian models of substitution rate evolution—parental guidance versus mutual independence. *Syst. Biol.*, 60(3):329–42, 2011.
- [15] S Magallon and A Castillo. Angiosperm diversification through time. *Am. J. Bot.*, 96(1):349–365, 2009.
- [16] S Penel, A-M Arigon, J-F Dufayard, A-S Sertier, V Daubin, L Duret, M Gouy, and G Perrière. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10 Suppl 6:S3, 2009.
- [17] BD Rae, B Förster, MR Badger, and G Dean Price. The CO₂-concentrating mechanism of *Synechococcus* WH5701 is composed of native and horizontally-acquired components. *Photosynth. Res.*, 109(1-3):59–72, 2011.
- [18] RL Tatusov, ND Fedorova, JD Jackson, AR Jacobs, B Kiryutin, EV Koonin, DM Krylov, R Mazumder, SL Mekhedov, AN Nikolskaya, BS Rao, S Smirnov, AV Sverdlov, S Vasudevan, YI Wolf, JJ Yin, and DA Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, 2003.
- [19] A Tofigh, M Hallett, and J Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 8(2):517–35, 2011.

- [20] TJ Treangen and EPC Rocha. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.*, 7(1):e1001284, 2011.
- [21] D Williams, JP Gogarten, and RT Papke. Quantifying homologous replacement of loci between haloarchaeal species. *Genome biology and evolution*, 4(12):1223–44, January 2012.
- [22] O Zhaxybayeva, JP Gogarten, RL Charlebois, WF Doolittle, and RT Papke. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res.*, 16(9):1099–1108, 2006.

3 Supplementary figures

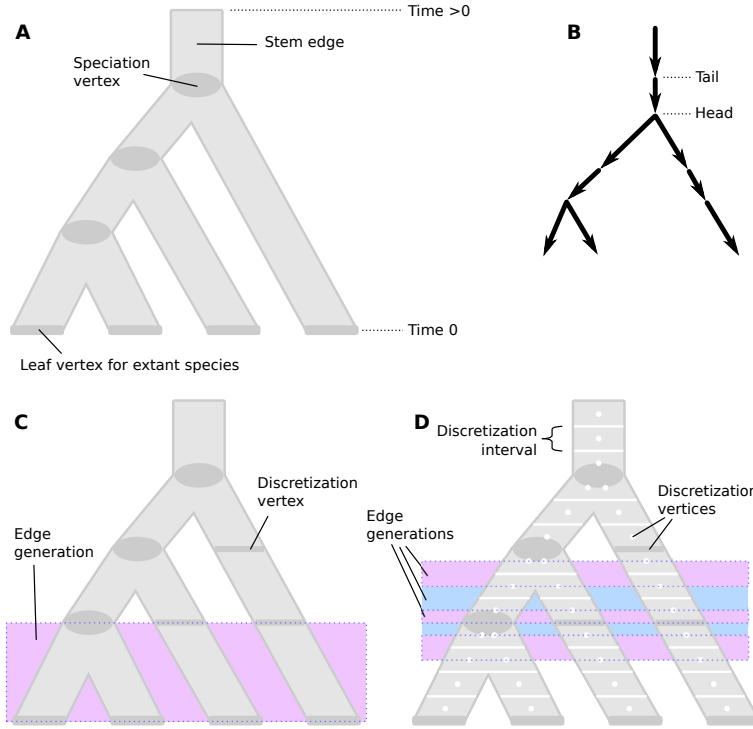


Figure S1: **Discretization of the species tree.** (A) The original tree S . (B) Directionality and notation used for edges. (C) The discretized tree S' , where vertices (grey bars) contemporaneous with speciations have been added. An edge generation is highlighted in purple. (D) The discretized tree S'' , where S' have been extended with vertices (white points) where GD and LGT events are allowed to occur. Vertices unique to S'' are potential placements for GD and LGT events in the DP algorithm and are associated with a discretization interval, while those of S' but not in S merely serve to organize our computations. Some separate edge generations are highlighted in purple and turquoise.

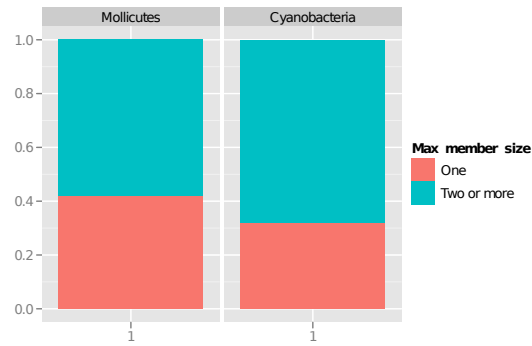


Figure S2: **Gene copy number for bacterial families.** Gene copy number for bacterial families with high GD rate estimates and low LGT rate estimates (see Supplementary methods). For Mollicutes, 58% of such families have at least one species with two or more family members, whereas for Cyanobacteria, 68% have this property (indicated by red bars). The result indicates that GD is a more common event than ILS and homologous recombination in the data.

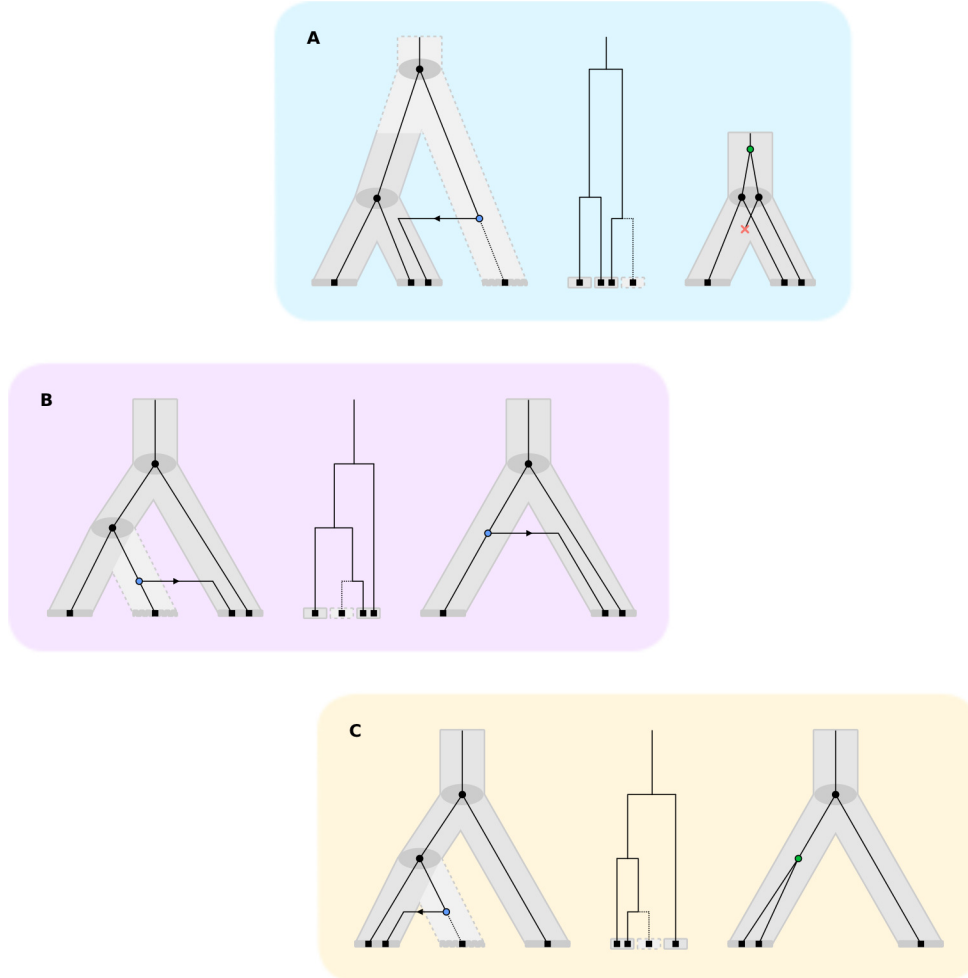


Figure S3: **LGT from unsampled species.** Three examples of LGT from unsampled, potentially extinct, species (indicated by dashed outline) that may confound analysis. (A) LGT from an outgroup species. This may be interpreted as a GD above the LCA in the sampled part of the species tree. (B) LGT from an ingroup species causing an LGT event to move. (C) LGT from an ingroup species causing an LGT event to be misclassified as a GD event. For each case, the actual condition is shown on the left, a corresponding ultrametric gene tree is shown in the middle, and a possible realization arising in analysis is shown on the right.

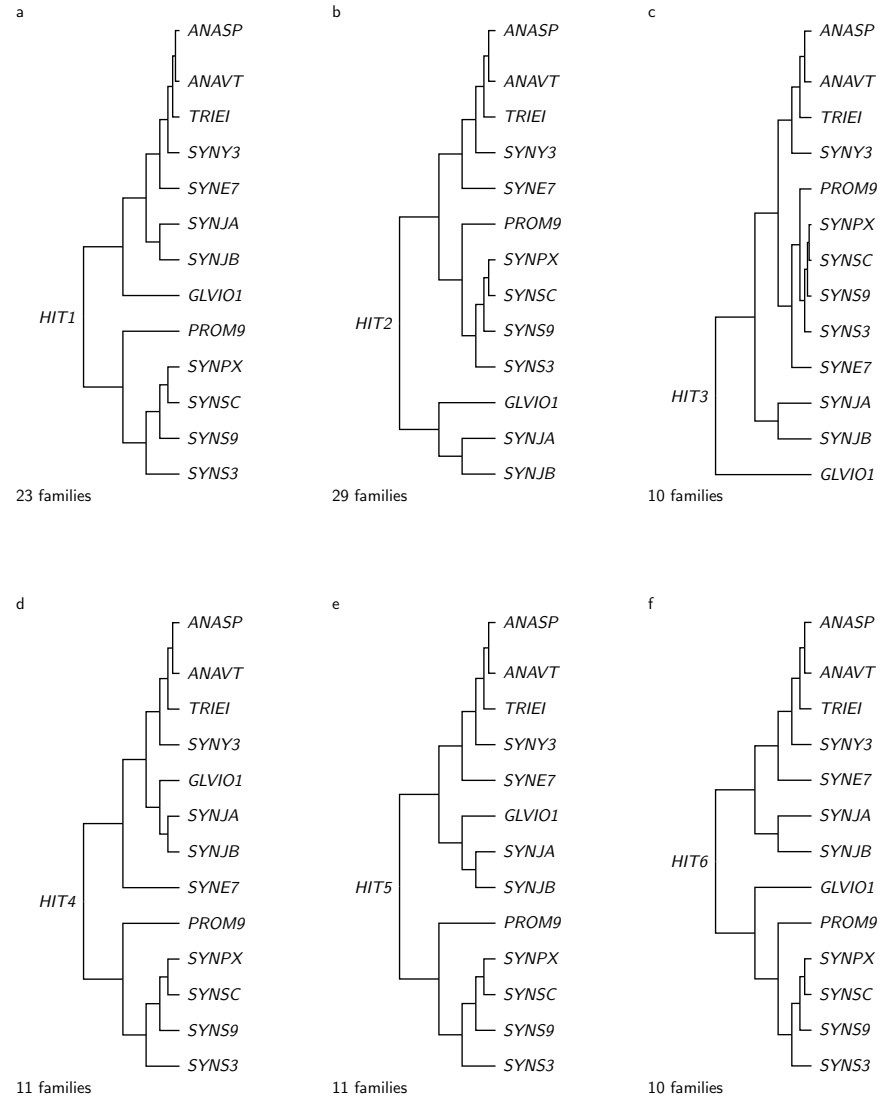


Figure S4: **Cyanobacteria highway-inducing topologies (HITs).** (A)–(F) illustrate the six different HITs among Cyanobacteria monolog families. The name of the HIT is shown to the left of the tree, while the number of gene families supporting it is shown below. Full species names can be found in Table S5.

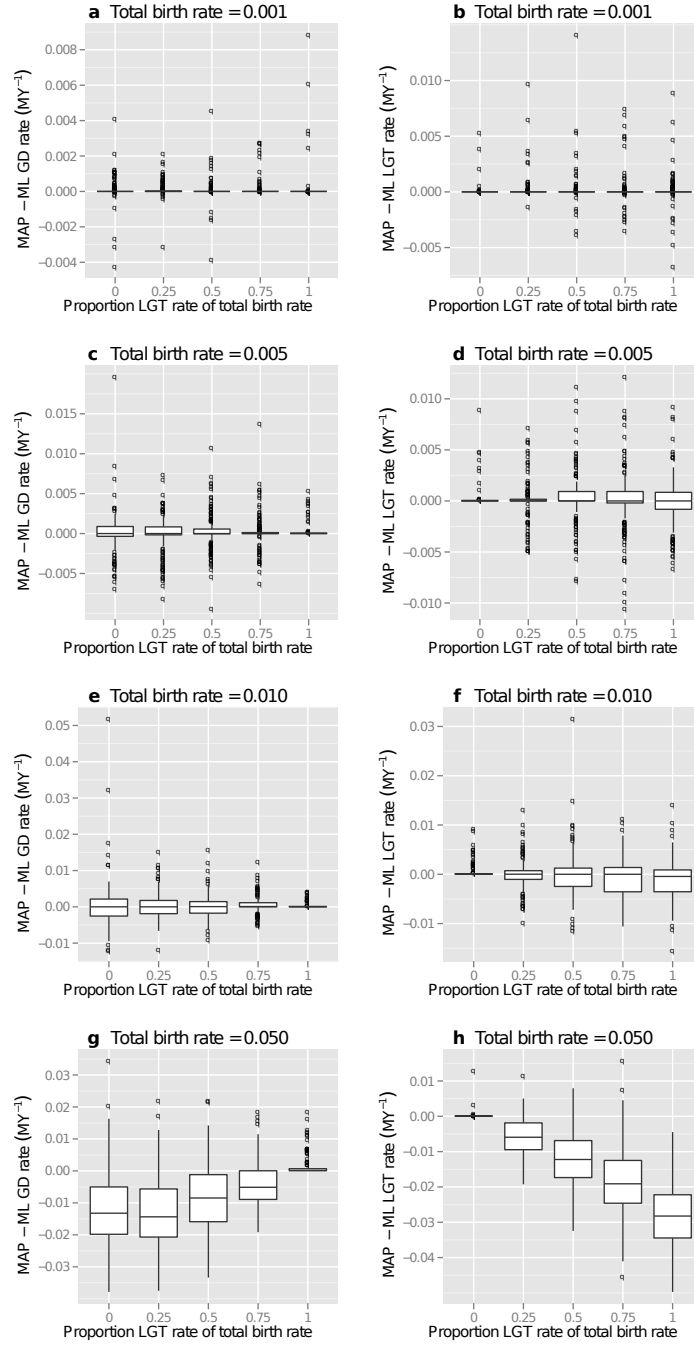


Figure S5: **Inference performance on synthetic data.** (A)–(H) show boxplots of the discrepancy between inferred PRIME-DLTRS MAP rate estimates and Informed ML estimates based on complete information of the true unpruned trees. Each distribution is based on estimates from 100 families. Except for very high total birth rate (0.05 MY⁻¹), PRIME-DLTRS estimates are very close to the Informed ML estimates.

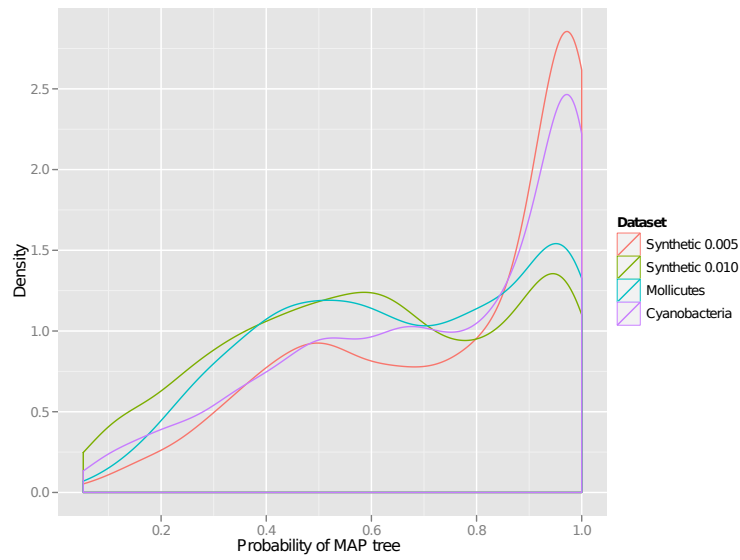


Figure S6: **Distribution of the posterior probabilities of inferred MAP trees for synthetic and experimental data.** For the synthetic data, the simulation results using a total birth rate of 0.005 and 0.01 is shown (the distributions using a total birth rate of 0.001 and 0.05 have a similar overall shape). The distribution shapes for synthetic data resemble those of biological datasets.

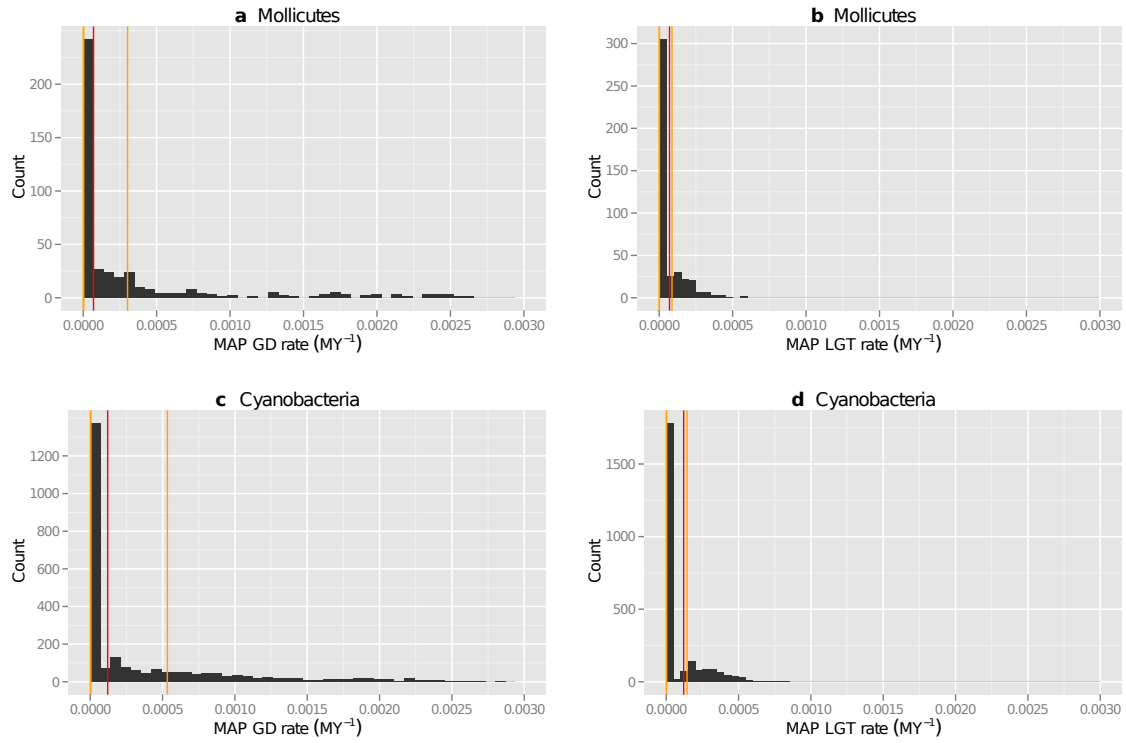


Figure S7: **Histograms showing the distributions of MAP rate estimates among bacterial gene families.** Orange bars indicate the quartiles of the distributions, while red bars indicate the limits for the classification into high and low rates (cf. Supplementary methods). The orange bars for the two first quartiles (25% and 50%) are both very close to zero and overlap in the figure.

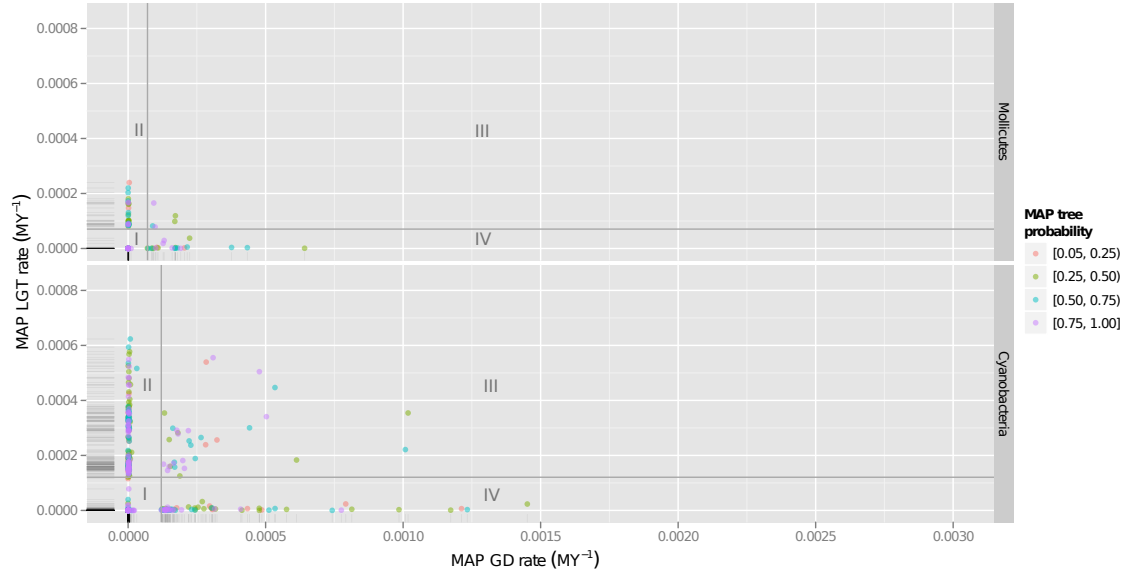


Figure S8: **MAP rate estimates for individual monolog gene families.** For Mollicutes, $98/425 = 23\%$ are monologs, while for Cyanobacteria $469/2467 = 19\%$ are monologs. For reference, horizontal and vertical grey lines indicate limits for the classification into high and low rates (cf. Supplementary methods). These define the following quadrants (marked by roman numerals in figure), I: with low GD and LGT rates, II: with low GD and high LGT rates, III: with high GD and LGT rates, and IV: with high GD and low LGT rates; the percentage of families in each quadrant are, in order, 42%, 21%, 5%, and 32% for Mollicutes and 40%, 15%, 7%, and 38% for Cyanobacteria. Markers are color-coded to represent the posterior probability of the MAP estimate. Average GD/LGT rates per MY are $4.9 \times 10^{-5}/4.7 \times 10^{-5}$ for Mollicutes and $7.1 \times 10^{-5}/1.2 \times 10^{-4}$ for Cyanobacteria. By comparing with main text Figure 4, it is evident that restricting the analysis to monologs can affect rate estimates substantially.

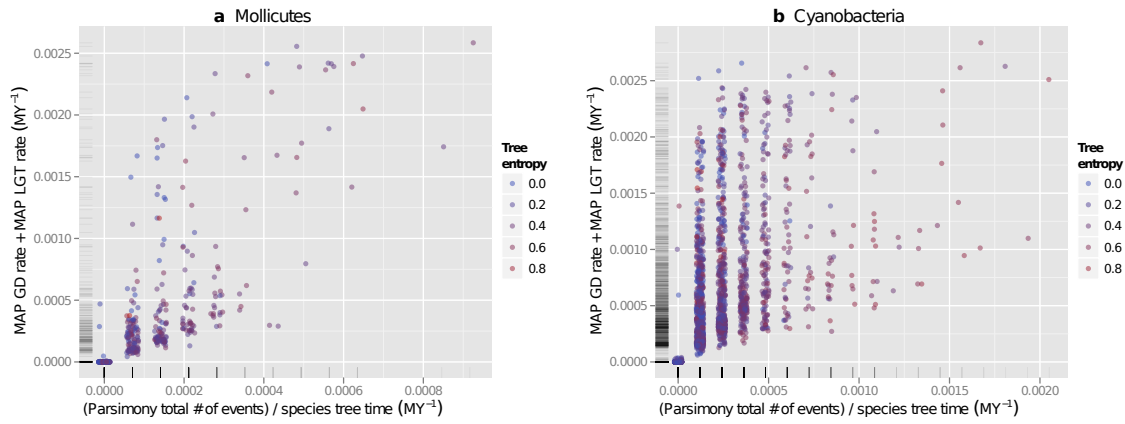


Figure S9: **Parsimony-probabilistic estimate correlation.** Scatter plots showing the correlation between parsimony minimum number of events and PRIME-DLTRS MAP estimates of total birth rate (i.e., GD + LGT rate). (A) Mollicutes and (B) Cyanobacteria. While there is an overall correlation (correlation coefficients, Pearson's/Spearman's rank/Kendall's, are as follows: Mollicutes, 0.75/0.84/0.72; Cyanobacteria, 0.59/0.83/0.69), it is clear that the PRIME-DLTRS probabilistic results refine the parsimony results. Marker's color indicates variance of the MAP tree estimate, as measured by Shannon's entropy.

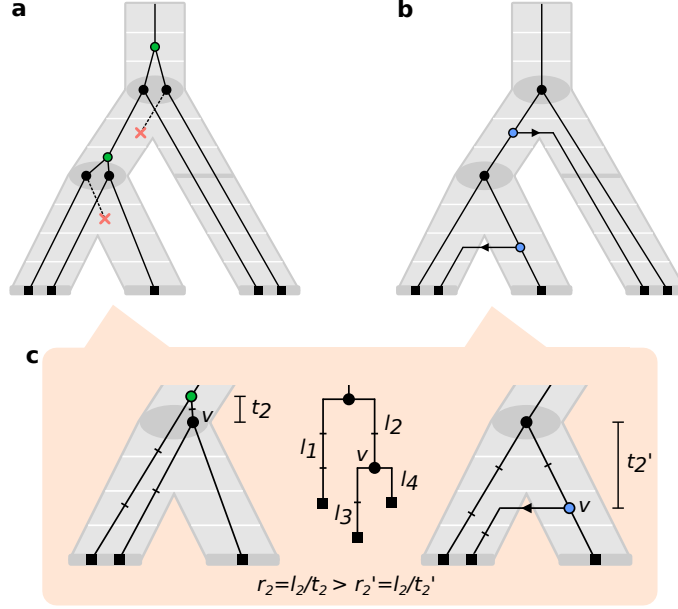
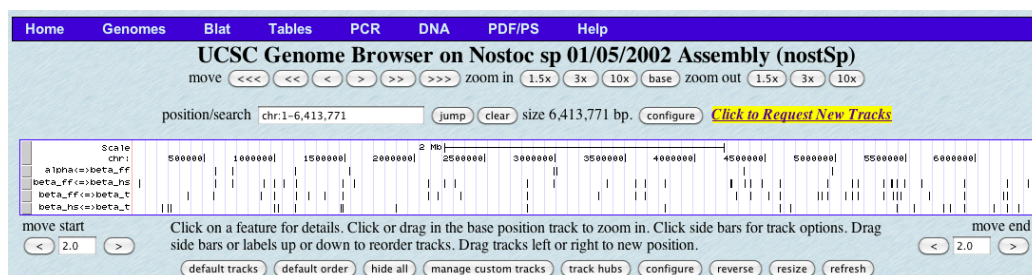


Figure S10: **Example of SE constraints on reconciliations.** Consider two different realizations of the same gene tree to the species tree, S , e.g., with (A) only GD and GL events, and (B) only LGT events. (C) The two realizations both need one GD/LGT event. However, the edge lengths (here indicated by tick bars) induced by sequence divergence favor the right scenario due to the more sound rates this would yield. Consequently, as the probabilities of the two realizations under the DLTRS model will be different, the edge lengths assist in weighting between different scenarios. During an MCMC iteration in PRIME-DLTRS, we integrate the probability contributions from all possible reconciliations and realizations of a gene tree to the species tree.

a



b

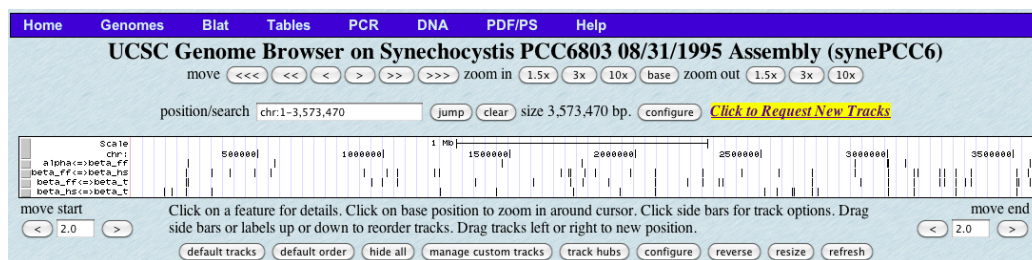


Figure S11: **Genomic co-localization of LGT highways (HWs).** The genomic position of HW gene families using the UCSC Microbial Genomic Browser on (A) the *Nostoc sp. PCC 7120* genome, and (B) the *Synechocystis sp. PCC 6803* genome. In both cases the full length of the chromosome is shown in the white window, with genomic positions shown at the top. For each highway (see Table S6), the genomic position of each gene family of the HW is indicated by a vertical bar. There is no apparent clustering of HW gene families to specific genomic regions, however, some gene families participate in several HWs.

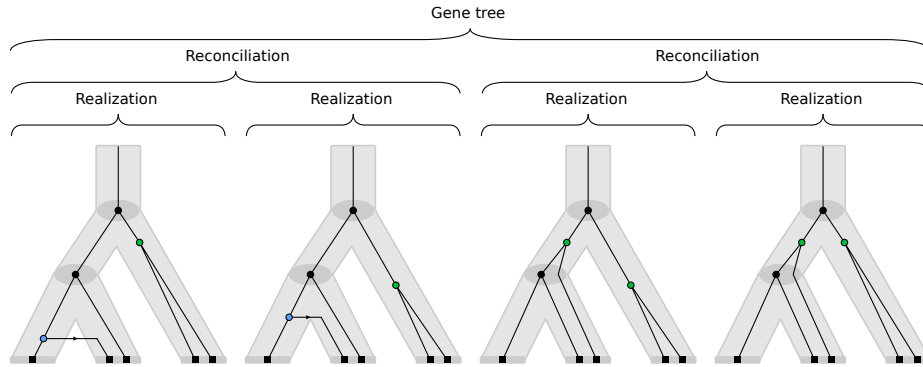


Figure S13: **Reconciliations and realizations.** Two different reconciliations of the same gene tree, each of which is illustrated with two realizations.

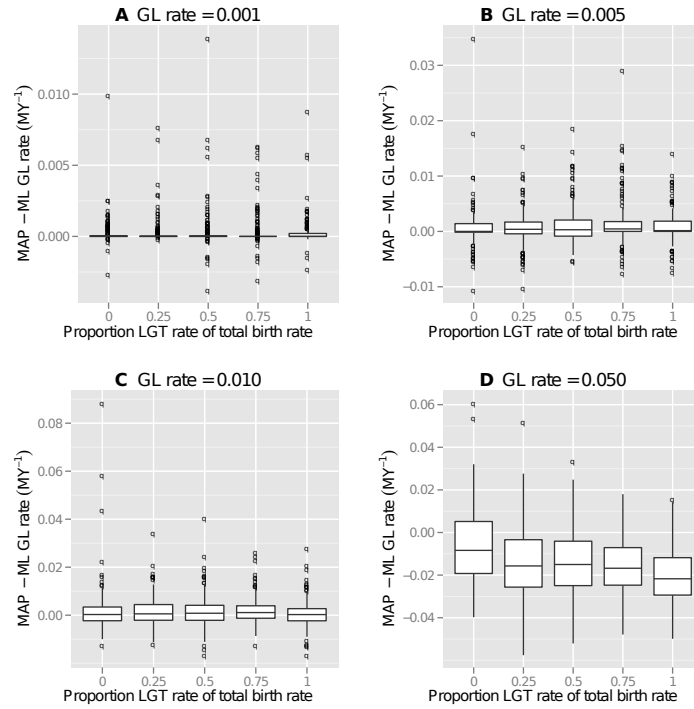


Figure S14: **Inference performance on synthetic data for GL rate.** (A)–(D) show boxplots of the discrepancy between inferred PRIME-DLTRS MAP rate estimates and Informed ML estimates based on complete information of the true unpruned trees. Each distribution is based on estimates from 100 families.

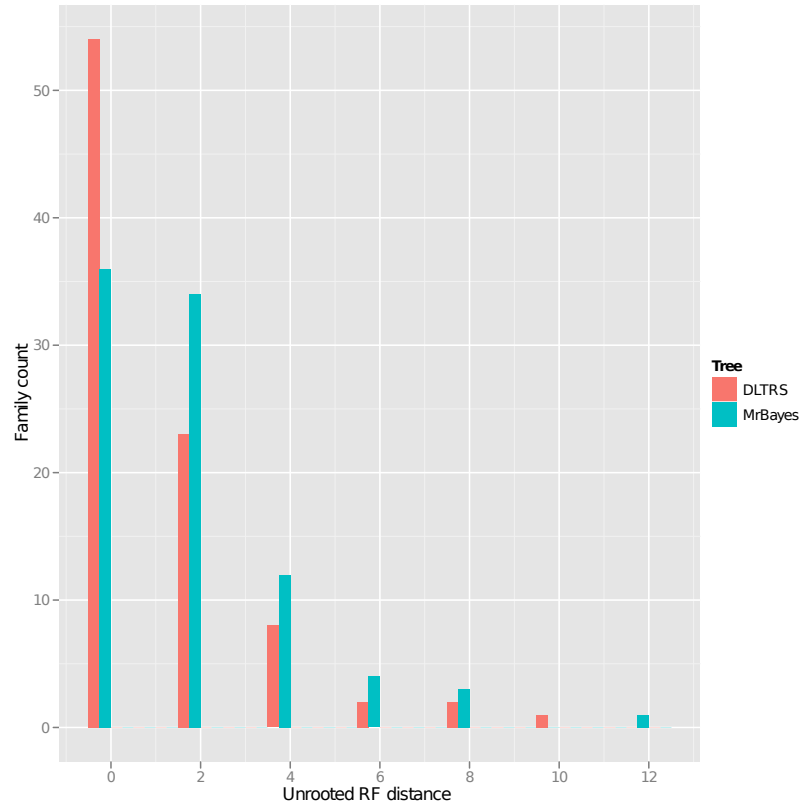


Figure S15: **MrBayes comparison on synthetic data.** The symmetric unrooted Robinson-Foulds (RF) distance between 90 true topologies and MAP trees inferred with PRIME-DLTRS and MRBAYES. The synthetic gene families were created using the DLTRS model with parameters similar to those estimated from the bacterial datasets, and using the cyanobacterial species tree. 9 families were discarded due to failure of MRBAYES summary analyses, and 1 family due to PRIME-DLTRS insufficient convergence.

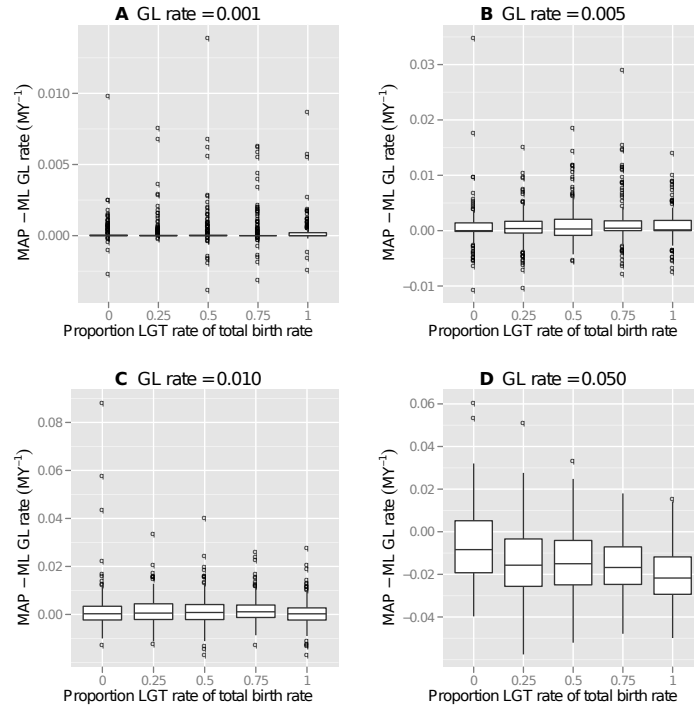


Figure S16: **Inference performance on synthetic data for GL rate.** (A)–(D) show boxplots of the discrepancy between inferred PRIME-DLTRS MAP rate estimates and Informed ML estimates based on complete information of the true unpruned trees. Each distribution is based on estimates from 100 families.

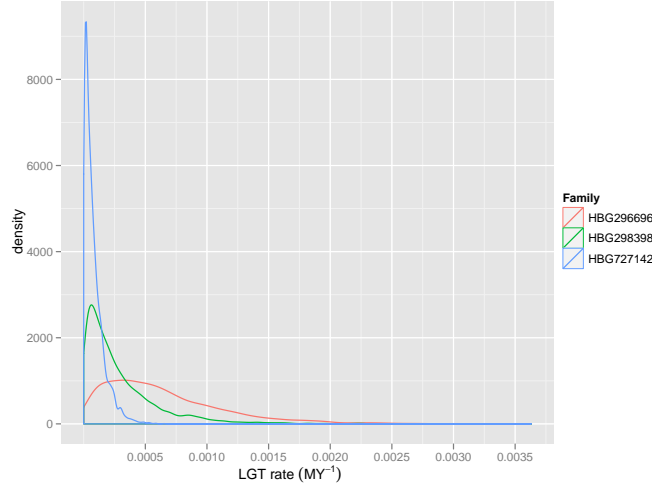


Figure S17: **LGT rate posterior.** Shows the LGT rate marginal posterior distribution for three Mollicutes families, corresponding to low, intermediate, and high LGT rate, respectively. The prior was set to $(0, 1]$ in all cases.

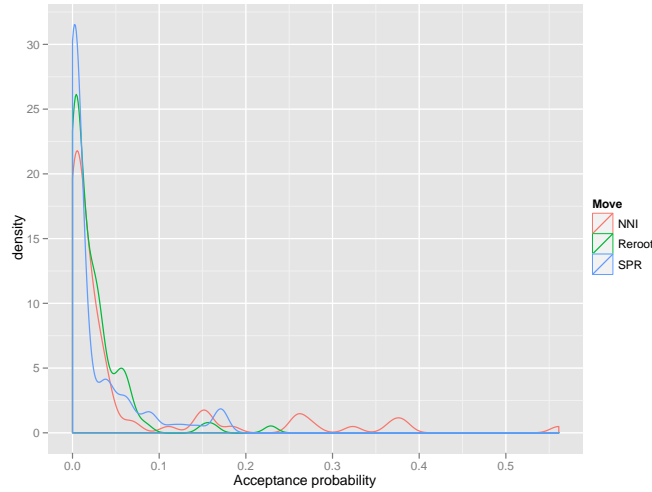


Figure S18: **Tree perturbation acceptance probabilities.** For 100 uniformly drawn Mollicutes families, shows the MCMC acceptance probability for various tree operations. Particularly high NNI values stem from small gene families with as few as 4 extant sequences. Average acceptance probabilities were as follows: NNI: 0.058; SPR: 0.024; Reroot: 0.024.

Table S1: **Results of validations of GD and LGT rate estimates.**

In order to test the initial rate estimates, each family was reanalysed twice with different constraints: *high-fixed*, the high rate type was fixed to its estimated high value, and *zero-fixed*, the high rate type was fixed to 0. The right column contains the number of the reanalysed instances for which the high-fixed gave a higher MAP joint and which, consequently, are unlikely to have been misclassified. Furthermore, manual inspection of original parallel runs for potentially misclassified families strongly suggested that the original estimates were sound.

Original rate estimates	Number of families where high-fixed analysis better
Mollicutes, high GD, low LGT	19 of 29
Mollicutes, low GD, high LGT	22 of 38
Cyanobacteria, high GD, low LGT	19 of 27
Cyanobacteria, low GD, high LGT	29 of 37

Table S2: **HIT-associated gene families.** The distribution of number of gene families, associated with each HIT, over the quadrants implied by high and low GD and LGT rates. The quadrants are I: low GD and LGT rates, II: high GD and low LGT rates, III: high GD and LGT rates, and IV: with low GD and high LGT rates (cf. Supplementary methods and main text Figure 4).

HIT	I	II	III	IV
1	2	21	0	0
2	12	1	0	16
3	0	0	0	10
4	0	0	2	9
5	4	1	0	6
6	4	1	0	5
All	22	24	2	46

Table S3: **Enrichment for genomic islands (GIs).** The following abbreviations are used: nq^+ and nbg^+ indicates number of GI-positive monologs in query and background, respectively, nq and nbg indicates the total number of monologs in query and background sets, respectively, OR and p-value indicates odds ratio and Bonferroni corrected p-value from the Fisher's exact test.

Taxon	Query subset	nq^+	nq	nbg^+	nbg	OR	p-value
Cyanobacteria	HWs	3	82	205	2467	0.39	0.56
Cyanobacteria	High GD rate	141	1055	205	2467	1.6	0.00027
Cyanobacteria	Low GD rate	64	1412	205	2467	0.55	0.00011
Cyanobacteria	High LGT rate	50	655	205	2467	0.92	1.0
Cyanobacteria	Low LGT rate	155	1812	205	2467	1.0	1.0
Mollicutes	High GD rate	12	182	19	425	1.5	1.0
Mollicutes	Low GD rate	7	243	19	425	0.64	1.0
Mollicutes	High LGT rate	7	119	19	425	1.3	1.0
Mollicutes	Low LGT rate	12	306	19	425	0.88	1.0

Table S4: **Cyanobacteria subtree division.** Functional classification, based on photo-system [17, 3], and mnemonics for the Cyanobacteria subtrees in Figure 2B.

Functional class	Habitat	Subtree mnemonic
α -Cyanobacteria	Marine	α
β -Cyanobacteria	Freshwater and filamentous colonies	β_{ff}
	Hot springs	β_{hs}
	Terrestrial	β_t

Table S5: Mnemonics [16] and full names for the investigated Mollicute and Cyanobacteria taxa.

Mnemonic	Full name
Mollicutes	
AYWBP	<i>Aster yellows witches'-broom phytoplasma</i> AYWB
ONYEL	<i>Onion yellows phytoplasma</i> OY-M
MEFLO	<i>Mesoplasma florum</i> l1
MYCCT	<i>Mycoplasma capricolum</i> subsp. <i>capricolum</i> ATCC 27343
MYMYC	<i>Mycoplasma mycoides</i> subsp. <i>mycoides</i> SC PG1
MYGAL	<i>Mycoplasma gallisepticum</i> R
MYGEN	<i>Mycoplasma genitalium</i> G 37
MYPNE	<i>Mycoplasma pneumoniae</i> M 129
MYPEN	<i>Mycoplasma penetrans</i> HF-2
URPAR	<i>Ureaplasma parvum</i> serovar 3 str. ATCC 700970
MYCH2	<i>Mycoplasma hyopneumoniae</i> 232
MYMOB	<i>Mycoplasma mobile</i> 163 K
MYPUL	<i>Mycoplasma pulmonis</i> UAB CTIP
MYCS5	<i>Mycoplasma synoviae</i> 53
Cyanobacteria	
ANAVT	<i>Anabaena variabilis</i> ATCC 29413
ANASP	<i>Nostoc</i> sp. PCC 7120
TRIEI	<i>Trichodesmium erythraeum</i> IMS 101
SYNY3	<i>Synechocystis</i> sp. PCC 6803
SYNE7	<i>Synechococcus elongatus</i> PCC 7942
PROM9	<i>Prochlorococcus marinus</i> str. MIT 9312
SYNS3	<i>Synechococcus</i> sp. CC 9311
SYNSC	<i>Synechococcus</i> sp. CC 9605
SYNPX	<i>Synechococcus</i> sp. WH 8102
SYNS9	<i>Synechococcus</i> sp. CC 9902
SYNJB	<i>Synechococcus</i> sp. JA-2-3B'a(2-13)
SYNJA	<i>Synechococcus</i> sp. JA-3-3Ab
GLVIO	<i>Gloeobacter violaceus</i> PCC 7421

Table S6: **Highways between Cyanobacteria subtrees of Figure 3B.**
(see Table S4 for subtree mnemonics).

HW between subtrees	# affected gene families	Corresponding HITs
$\alpha \Leftrightarrow \beta_{ff}$	13	3
$\beta_{ff} \Leftrightarrow \beta_{hs}$	40	
$\beta_{ff} \Leftrightarrow \beta_t$	40	4,5
$\beta_{hs} \Leftrightarrow \beta_t$	21	2,4,5

Table S7: **Enriched functional terms.** The table gives Bonferroni-corrected p-values for COG- and GO-terms in different Cyanobacteria gene family classes based on (i) GD and LGT rates and (ii) topology in relation to the species tree; only terms with p-value < 0.05 are shown. P-values for Mollicutes, when relevant, are given in brackets.

Functional terms	<i>lodup.lotrans</i>	<i>lodup.hitrans</i>	<i>sptree</i>
COG terms			
J: Translation, ribosomal structure and biogenesis (INFORMATION STORAGE AND PROCESSING)	-	-	1.02E-7 [0.0035]
D: Cell cycle control, cell division, chromosome partitioning (CELLULAR PROCESSES AND SIGNALLING)	-	-	-
S: Function unknown (POORLY CHARACTERIZED)	-	-	-
go terms			
GOTERM_MF_FAT:GO:0003735 structural constituent of ribosome	2.9E-5	-	9.5E-10 [2.5E-6]
GOTERM_CC_FAT:GO:0005840 ribosome	2.4E-5	-	1.3E-8 [2.3E-7]
GOTERM_CC_FAT:GO:0030529 ribonucleoprotein complex	6.1E-5	-	2.9E-8 [2.3E-7]
GOTERM_MF_FAT:GO:0005198 structural molecule activity	0.00018	-	4.7E-9 [2.5E-7]
GOTERM_BP_FAT:GO:0006412 translation	0.00034	-	2.0E-7 [7.4E-7]
GOTERM_CC_FAT:GO:0043232 intracellular non-membrane-bounded organelle	0.0013	-	2.8E-6 [1.3E-6]
GOTERM_CC_FAT:GO:0043228 non-membrane-bounded organelle	0.0013	-	2.8E-6 [1.3E-7]
GOTERM_MF_FAT:GO:0019843 rRNA binding	0.029	-	0.00015 [7.1E-5]
GOTERM_MF_FAT:GO:0003723 RNA binding	0.019	-	0.0002 [0.028]
GOTERM_BP_FAT:GO:0044271 nitrogen compound biosynthetic process	-	0.028	-
GOTERM_MF_FAT:GO:0032553 ribonucleotide binding	-	-	-
GOTERM_MF_FAT:GO:0032555 purine ribonucleotide binding	-	-	-