A Bayesian Method for Analyzing Lateral Gene Transfer

JOEL SJÖSTRAND1, 2, ALLI TOFGH3, VINCENT DAUBIN4, LAS ARVESJAD1, 2, 5, BENET SENNBLAD1, 6, and JENS LAGERGREN1, 7, ∗

1 Science for Life Laboratory, Tomtebodavägen 23A, 17165 Solna, Sweden, 2 Department of Numerical Analysis and Computer Science, Stockholm University, Sweden, 3 McGill Centre for Bioinformatics, 4th floor, Bellini Building, Life Sciences Complex, 3489 Promenade Sir William Osler, Montreal, Quebec, Canada, H3G 0B1, 4UMR CNRS 5558 - LBBE, “Biometrie et Biologie évolutive”, UCB Lyon 1 - Bât. Grégor Mendel, 43 bd du 11 novembre 1918, 69622 VILLEURBANNE cedex, 5 Department of Mechanics, Osquars Backe 18, KTH, SE-100 44 Stockholm, Sweden, 6 Karolinska University Hospital, CMM LB 03, Solna, SE-171 76 Stockholm, Sweden and 7 The School of Computer Science and Communication, Lindstedtsvägen 3, 5, KTH CSC, SE-100 44 Stockholm, Sweden

∗ Correspondence to be sent to: Science for Life Laboratory, KISPA (Karolinska Institutet Science Park), Tomtebodavägen 23A, 17165 Solna, Sweden, E-mail: jens.lagergren@sci.fh.se

Joel Sjöstrand and Ali Tofigih contributed equally to this article.

Received 1 March 2013; revisions returned 24 May 2013; accepted 10 February 2014
Associate Editor: Peter Foster

Abstract—Lateral gene transfer (LGT)—which transfers DNA between two non-vertically related individuals belonging to the same or different species—is recognized as a major force in prokaryotic evolution, and evidence of its impact on eukaryotic evolution is over increasing. LGT has attracted much public attention for its potential to transfer pathogenic elements and antibiotic resistance in bacteria, and to transfer pesticide resistance from genetically modified crops to other plants. In a wider perspective, there is a growing body of studies highlighting the role of LGT in enabling organisms to occupy new niches or adapt to environmental changes. The challenge LGT poses to the standard tree-based conception of evolution is also being debated. Studies of LGT have, however, been severely limited by a lack of computational tools. The best currently available LGT algorithms are parsimony-based phylogenetic methods, which require a pre-computed gene tree and cannot choose between sometimes wildly differing most parsimonious solutions. Moreover, in many studies, simple heuristics are applied that can only handle putative orthologs and completely disregard gene duplications (GDs). Consequently, proposed LGT among specific gene families, and the rate of LGT in general, remain debated. We present a Bayesian method for analyzing LGT, MrBayes (Ronquist and Huelsenbeck 2003) and BEAST (Drummond and Rambaut 2007), are highly popular and provide a very precise understanding of the evolutionary role of LGT.

Today, it is firmly established that lateral gene transfer (LGT) occurs frequently among prokaryotes (Brown 2003; Dagan and Martin 2007; Andam and Gogarten 2011), and evidence of eukaryotic LGT is established and accumulating (Keeling and Palmer 2008). There are studied examples of laterally transferred material between all domains of life (Boto 2010). Prokaryotic LGT is now known to have transferred genes involved in, for example, antibiotic resistance, nitrogen fixation, and virulence (Swithers et al. 2012), while eukaryotic LGT examples include, for instance, the transfer of pesticide resistance from genetically modified crops to other plants (Brown 2003). Although the reality of LGT is well accepted, its prevalence and ramifications for organismal evolution are still hotly debated; a key question is whether a tree is a meaningful representation of LGT, the focus has typically been on LGT between different species and the potential for innovation this provides, say, with respect to existing pathways (Pal et al. 2005). It has, however, also been argued that among prokaryotes, close-species LGT can—alleged to sexual reproduction in other organisms—combine advantageous point mutations from different cell lineages (Muller 1964) as well as render accumulation of deleterious mutations (Hafner et al. 2000), that is, Muller’s ratchet (Muller 1964). A complementary view is that a large portion of laterally transferred genes is in fact nearly neutral to the recipient, and that prokaryotic pan-genomes may harbor a vast exchangeable gene pool, unless reduced by selective sweeps (Gogarten and Townsend 2005). However, present LGT detection methods capable of handling evolutionarily distant genomes have so far not been able to provide a very precise understanding of the evolutionary role of LGT.

Probabilistic methods for reconstructing phylogenetic trees based on sequence evolution have matured significantly during the last 15 years. Bayesian Markov-chain Monte Carlo (MCMC) approaches, for example, MrBayes (Ronquist and Huelsenbeck 2003) and BEAST (Drummond and Rambaut 2007), are highly popular and considered crucial for resolving complex questions in
evolutionary biology (Hulsenbeck et al. 2001). Recently, gene–species tree reconciliation methods reached a similar maturity when the PrIME-GSR method, which integrates gene duplication (GD), gene loss (GL), and a relaxed molecular clock submodel of sequence evolution, was described (Akerborg et al. 2009). PrIME-GSR effectively reconstructs a gene tree and simultaneously reconciles it with a given dated species tree. Rasmussen and Kellis (2011) have since presented a similar framework that achieves a higher efficiency by using approximations and producing a point estimate rather than the entire distribution. However, their approach requires genome-wide estimation of certain parameters.

By contrast, probabilistic approaches incorporating LGT have been few. Discarding phylogenetic networks—which are perhaps better suited for modeling hybridization (Bloomquist and Suchard 2010)—LGT detection approaches can be broadly classified into sequence composition methods, similarity-based or distance-based methods, and phylogenetic tree methods. Sequence composition methods identify LGT events based on atypical sequence composition and genomic location (Azad and Lawrence 2007), while similarity-based and distance-based methods look for genomic deviations in sequence similarity or divergence rates (e.g., Novichkov et al. 2004). Phylogenetic tree methods infer LGT events from incongruence between a species tree and a gene tree (Suchard et al. 2003; Beiko and Hamilton 2006; Zhaxybayeva et al. 2006; Linz and von Haeseler 2007; Poptsova and Gogarten 2007; Shi and Falkowski 2008; Tofígh et al. 2011). Phylogenetic methods have typically been parsimony-based (Beiko and Hamilton 2006; Tofígh et al. 2011) or heuristic (Hill et al. 2010; Abby et al. 2010, 2012), and have relied on a priori inferred trees; this excludes any opportunity to deliberately balance the trade-off between sequence evolution and tree incongruence. This is unfortunate, since there may be alternative, approximately equiprobable gene trees, which can be reconciled more easily with the species tree, for example, showing no evidence of LGT. In addition, it is often pointed out that GD and GL may also be responsible for gene and species tree incongruences, but with a few exceptions (e.g., David and Alm 2011; Tofígh et al. 2011), most algorithms for LGT detection explain incongruence by primarily invoking LGT events. Another shortcoming is that most phylogenetic methods merely consider the gene tree topology, that is, they disregard the edge lengths implied by sequence evolution (the importance of which is illustrated in Fig. S10 of the Supplementary Material (SM); see http://dx.doi.org/10.5061/dryad.76r2). Finally, most existing phylogenetic methods can only be applied to the subset of all gene families consisting of what we here call monologs, that is, gene families with exactly one member in each extant species under consideration. We note that while incomplete lineage sorting (ILS) constitutes another potential source of incongruence, most LGT methods have been applied on timescales where ILS is believed to be of limited concern.

Recently, Dagan and Martin (2007) described an elegant approach to infer LGT prevalence by analyzing the effect of an assumed upper bound on LGT abundance on ancestral genome sizes. Subsequently, David and Alm (2011) extended this approach, such that parsimony reconciliation weights for GD, GL, and LGT can be obtained and applied in a parsimony analysis of the relative abundance of these events. A fundamental problem with the latter approach is that a single most parsimonious reconciliation is used when analyzing event rates. However, there are often many, equally parsimonious reconciliations that may imply very different rates. Since the reconciliation used in the analysis is selected using a deterministic algorithm rather than sampled randomly, such an approach may suffer from systematic biases. The algorithm in Bansal et al. (2012) constitutes a step toward rectifying this problem. It can be used to sample uniformly from the set of most parsimonious reconciliations. Even so, apart from using the gene tree topology inferred from the gene sequences, this sampling is oblivious to the sequences, despite the important information these provide.

Below, we go beyond earlier parsimony frameworks and introduce a fully integrated probabilistic model and inference tool, PrIME-DLTRS, and show that it performs self-consistently on biologically realistic synthetic data. We then analyze two bacterial data sets subject to LGT: Mollicutes and Cyanobacteria. Earlier reports on prokaryotic LGT prevalence have varied significantly (Dagan and Martin 2007), although studies including both Cyanobacteria and Mollicutes have typically reported moderate and similar LGT rates (Dagan and Martin 2007; Zhaxybayeva et al. 2006; Abby et al. 2012). Both Shi and Falkowski (2008) and Abby et al. (2012) found support for resistance to LGT within “core” gene families in Cyanobacteria (e.g., those involved in macromolecular interactions in complex protein structures), whereas Zhaxybayeva et al. (2006) found genes from all functional categories to be subject to transfer. However, these differences may be due to sampling at different taxonomic depths. We extend on these studies, and perform a sound quantitative and functional investigation of LGT and GD using our tool where—unlike earlier attempts—a multitude of evolutionary mechanisms are simultaneously accounted for.

**MATERIALS AND METHODS**

**DLTRS Inference Algorithms**

PrIME-DLTRS makes use of an MCMC framework, discretizes the species tree, solves systems of ordinary differential equations (ODEs), and performs sophisticated dynamic programming (DP) computations (Fig. 1b–d). See SM Methods for a detailed description of the PrIME-DLTRS inference algorithm constituents.
FIGURE 1. Various aspects of the DLTRS model and inference. The species tree $S$ is shown in gray and the gene tree $G$ in black. a) The events that can affect a gene lineage. Substitution rates across gene tree edges are iid gamma, resulting in a gene tree with relaxed-clock edge lengths. b) Discretization of $S$. Probabilities are approximated by considering discretized realizations only (see also Supplementary Fig. S1). c) ODEs are derived and solved for lineage-related probabilities, here illustrated with $Q_e(t)$, the extinction probability of a gene lineage at time $t$ on edge $e$. d) Dynamic programming employed to sum the probability contributions of all discretized realizations. Here we illustrate the summation of all realizations mapping the internal vertices of the subtree rooted at $v$ to the discretization vertices of the dark gray area.

Generating Synthetic Data

To achieve a biologically relevant simulation, without focusing on a few selected 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. (2011); an MCMC analysis under an integrated independent and identically distributed (iid) gamma model for substitution rates and a birth–death model for the rbcl tree evolution for 79 species of flowering plants. The rbcl gene tree is assumed to coincide precisely with the species tree, which indeed is a reasonable assumption for plastid genes (see SM Methods for details).
Prokaryotic Species Tree

We used the species trees obtained by Abby et al. (2012); see Figure 3. Our Mollicutes species tree has also been obtained from maximum likelihood (ML) reconstruction using selected concatenated coding sequences (Vasconcelos et al. 2005). In the cyanobacterial tree, the uncertain placement of *Synechococcus elongates* has posed a particular problem (Shi and Falkowski 2008). However, recently, Gupta and Mathews (2010) gave compelling arguments for placing *S. elongatus* as in the species tree reconstructed by Abby et al. (2012). Relative divergence times were estimated with MAP-DP (Åkerborg et al. 2008) using monolog gene families assuming a birth–death divergence prior, the JTT substitution model (Jones et al. 1992), iid gamma substitution rates over edges, and a discrete gamma model for rate categories over sites. Calibration of the divergence times was then made using estimates from Battistuzzi and Hedges (2009); see Figure 3.

Prokaryotic Gene Families

Genome-wide protein families were extracted from HOGENOM release 05 (Penel et al. 2009), which constitutes a publicly available resource. We created multiple sequence alignments using MUSCLE (Edgar 2004), and removed uninformative positions using Gblocks (Castresana 2000). Gene families were removed if they had fewer than 4 member sequences, belonged to only 1 of our selected species, or had alignment length less than 50 amino acids. A total of 444 and 2542 gene families were retained after filtering for Mollicutes and Cyanobacteria, respectively. Of these, 98 and 469 gene families constituted monologs for Mollicutes and Cyanobacteria, respectively.

Gene Tree Reconstruction

Gene tree topologies obtained from PrIME-DLTRS and MrBayes (Ronquist and Huelsenbeck 2003) correspond to the topologies with the highest marginal posterior probabilities (MAP trees). We evaluated how well PrIME-DLTRS and MrBayes reconstruct gene trees using the parsimony method PHYLTR (Tofilgh et al. 2011), which simultaneously considers GD and LGT events.

MCMC

Each MCMC run consisted of four parallel chains of $4 \times 10^6$ iterations and a thinning factor of 400 (for the experimental data, a thinning factor of 200 was used when estimating MAP GD, GL, and LGT rates, see SM Methods). Burn-in was set to $10^6$ iterations, chosen based on a safety margin obtained by inspection of pilot analyses. Gene tree topologies were perturbed using nearest-neighbor interchange, subtree pruning and regrafting, and rerooting. Lengths were heuristically altered during branch-swapping. The initial trees were uniformly selected from the set of all leaf-labeled topologies. Parameters of the rate model and edge lengths were perturbed using truncated normal proposals around the current value, with tuning parameters handcrafted with respect to acceptance ratios. Substitution rate parameters were changed by either perturbing the distribution mean or the coefficient of variation (CV). Gelman–Rubin diagnostics, inspection of trace plots, and various outlier detection methods were used to determine convergence.

RESULTS

The DLTRS Model

Our duplication-loss-transfer model with iid rates across gene tree edges and sequence evolution, denoted DLTRS, is the first relaxed molecular clock model capturing GD, GL, LGT, and sequence evolution. The model gives rise to significantly harder computational problems than when LGT is not included (Åkerborg et al. 2009). DLTRS incorporates the three probabilistic submodels: DL, R, and S, which are described below. The species tree, denoted S (not to be confused with submodel S), is a rooted bifurcating clock-like tree, extended with an additional stem edge predating the root (Fig. 1). We will consider S and its divergence times to be given and omit them from our notation, that is, $P[S]$ will be written $P[\cdot]$. Although it is possible to include leaves in S that represent extinct species, we will describe the case where the leaves represent extant species at time 0, with interior vertices occurring at time $> 0$.

The DL submodel describes the evolution of a gene tree G through GD, GL, and LGT events as an augmented birth-death process over S (Fig. 1a). A gene lineage on a species tree edge will be exposed to GD, GL, and LGT events at rates $b, \mu$, and $\tau$, respectively. A GD or LGT event affecting a gene lineage evolving over a species tree edge instantaneously results in two child lineages. Subsequent to a GD event, both child lineages continue to evolve over the same edge as did the parental gene lineage. An LGT event, on the other hand, instantaneously transfers one of the child lineages to another (uniformly selected) contemporaneous species tree edge, while the other child continues to evolve on the same edge as did the parental gene lineage. In contrast, a GL event removes the affected lineage from the process. When a gene lineage reaches a species tree vertex, it splits into two independent lineages evolving down the different outgoing edges of the bifurcation. The process continues recursively down toward the leaves of S, resulting in a gene tree G. Finally, extinct lineages of G are pruned away.

We note that the model makes no distinction between whether a gene duplicate was received through a lateral transfer from another individual within the same species or if it stems from, for example, a tandem
duplication. That is, such events will appear as a GD in the species under consideration, since from a species lineage perspective, both events constitute a duplication in the population gene pool (see further below). A similar effect can be expected for transfer events between evolutionarily closely related species, in particular when the donor species is not part of the species tree.

A reconciliation is a mapping that associates each gene tree vertex with a species tree vertex or a species tree edge. A realization is a constrained reconciliation, in which every gene tree vertex mapped to a species tree edge is also pinpointed to a time on this edge. Both reconciliations and realizations map the gene tree vertices in a manner consistent with the gene tree; a gene tree vertex is never mapped closer to the root in the species tree than its parent. In addition, a realization never maps a child vertex and its parent to the same time (see Supplementary Fig. S13 for an illustration). The DLT submodel effectively generates a gene tree together with a realization; the latter implies divergence times for all gene tree vertices and time intervals for all gene tree edges.

The D submodel describes sequence evolution rate variation across gene tree edges. We use iid gamma-distributed sequence evolution rates (Åkerborg et al. 2008; Linder et al. 2011) in order to obtain a relaxed molecular clock, which allows for more biological realism (Thorne et al. 1998; Drummond et al. 2006; Lepage et al. 2007; Åkerborg et al. 2008; Linder et al. 2011). The product of an edge time and sequence evolution rate yields the edge length conventionally used in standard sequence evolution models. Finally, the S submodel can be any substitution model (Felsenstein 2004), and it describes how sequences are generated by molecular evolution over the gene tree with edge lengths.

Bayesian Inference Using MCMC

We employ an MCMC framework using the Metropolis–Hastings algorithm for inference. A naive, straightforward MCMC adaptation of the DLT-R model would use a state space encompassing a realization and a species tree, the latter implied by divergence times for all gene tree vertices and time intervals for all gene tree edges. The only factor for which the computation remains to be specified is the joint likelihood for the DLT and S submodels: p[G,θ|l]. This, however, is highly non-trivial, and we compute it using a complex combination of DP and multiple systems of ODEs. Below, we highlight some of the key concepts underlying the algorithm.

First, a crucial aspect of the algorithm concerns probabilities related to the lineages of a realization of G onto S. When a vertex v of G and its parent w have been fixed temporally onto specific locations in S, the presence of LGT enables myriad possible trajectories between v and w, and there may have been intermediary events whose lineages were subsequently lost. In the MCMC, we compute such lineage probabilities by numerically solving consecutive systems of ODEs (an example is illustrated in Fig. 1c) and applying DP; see SM for more details.

Second, let C be the set of all possible reconciliations of G and S. For any c ∈ C, a realization a is said to be compatible with c if for any vertex v of G either: (i) a(v) and c(v) is the same species tree vertex; or (ii) a(v) is a point on the species tree edge c(v). Let A(c) be the set of realizations that are compatible with c. Note also that C is finite, whereas a reconciliation c spans an infinite number of realizations A(c) (as long as c classifies some gene tree vertex as a GD or LGT event). Since the vertices of S create discontinuities in the density of realizations, we express the probability density p[G,θ] as follows:

\[ p[G, θ] = \sum_{c ∈ C} p[G, l, c | θ] = \sum_{c ∈ C} \int p[G, l, a | θ] da, \]  

where the integration is over all realizations a compatible with the reconciliation c. By discretizing S (Fig. 1b), we can approximate the sum of integrals as a sum of sums, which then becomes amenable for DP (Fig. 1d). The details of this are outlined in the SM.
PrIME-DLTRS is free software and is distributed in a Java JAR file available on http://code.google.com/p/jprime/. An older C++ version can be obtained from http://prime.sbc.su.se.

Analysis of Synthetic Data

We performed extensive tests on synthetic data sets in order to validate the capacity of PrIME-DLTRS to correctly estimate LGT and GD rates. Biologically sound synthetic data were generated with varying GD and LGT rates. 100 species trees with 6 leaves and 100 species trees with 11 leaves were created, and for each of these, we generated $4 \times 5 = 20$ gene trees, with varying birth and loss rates, together with multiple sequence alignments according to the DLTRS model, see Methods. Specifically, we used four birth rates, $b + c$, of 0.001, 0.005, 0.010, and 0.050 Myr$^{-1}$, respectively; the latter may be considered to be unrealistically high. For these total birth rates, we varied the proportion constituted by LGT rate as 0%, 25%, 50%, 75%, and 100%. In all cases, the loss rate was set to equal the total birth rate. In all, this resulted in $2 \times 100 \times 4 \times 5 = 4000$ individual gene trees, with the number of leaves ranging from 3 to 62. Our tests on synthetic data show that PrIME-DLTRS, except when the LGT rate is exceptionally high, reconstructs rates with high accuracy; see Figure 2, Supplementary Figure S5, and SM Discussion.

Analysis of Mollicutes and Cyanobacteria

We used PrIME-DLTRS to analyze GD and LGT events in two bacterial data sets: 444 gene families from 14 strains from the class Mollicutes, and 2542 gene families from 13 strains from the phylum Cyanobacteria, see Figure 3, (elsewhere in this article, we refer to strains as species). We note that the impact of ILS can be expected to be limited given the extensive timespan separating the speciations, which is also supported by the gene copy-number distributions in the data; see SM Discussion. The increase in copy-number also distinguishes our observation from what would be expected if intra-species or close-species transfer events from closely related species in the GD event in 15% of the cases. Accounting for potential transfers event from closely related species in the GD number, this gives an estimate of 38–53% of monolog families subject to LGT, which is in accordance with the findings of Shi and Falkowski (2008). This result also shows that restricting analysis to monologs may not provide representative estimates of LGT and GD for the whole gamut of gene family sizes (compare Fig. 4 and Supplementary Fig. S8).

Rates of LGT and GD.—Our analysis shows several common features for both data sets (Fig. 4). First, the estimated GD rates are on average clearly higher than the LGT rates. Second, the estimated average rate for GD, as well as for LGT, are approximately the same in Mollicutes and Cyanobacteria, which is in agreement with the results of Dagan and Martin (2007) and Abby et al. (2012). Third, for more than a third of all families, the gene tree follows the species tree and exhibits no GD and no LGT. These conclusions are supported by tests on synthetic data, showing that PrIME-DLTRS reconstructs rates with high precision. At a glance, our results appear to contradict those of Treangen and Rocha (2011), who use a heuristic based on sequence similarity and co-localization to differentiate between LGT and GD. They conclude that LGT has been responsible for 88–98% of gene family expansions in bacterial genomes. It is, therefore, essential to consider differences between the two analyses. First, there are methodological dissimilarities: Treangen and Rocha use pairwise similarity comparisons, while we implement a full Bayesian phylogenetic model for GD, GL, LGT, and sequence evolution. Moreover, Treangen and Rocha tailored their method to detect recent GD and LGT events. It is very likely that LGT's have a higher initial loss rate than GDs, since LGTs in general are not adapted to the receiving cell. Hence, we would expect a relatively higher ratio of LGT to GD among recent events than after the fixation of such events in the genome. Lastly, it is vital to recognize that their definition of GD differs from ours. Our definition counts intra-species LGT as GD, which is sound viewed from the perspective of a gene pool being vertically inherited in the species lineage. That is, while such an event mechanistically might be an LGT event between individuals of the same species, in the context of long-term species evolution, it is functionally equivalent to a GD event; this is analogous to GD caused by recombination between chromosome pairs in eukaryotes. The definition of Treangen and Rocha, on the other hand, counts intra-species transfer as LGT and, in fact, may also count GD followed by rearrangements, or caused by large tandem duplications followed by differential loss, as LGT. In our PrIME-DLTRS analysis, the proportions of all gene families where LGT is dominating are 23% and 19% for Mollicutes and Cyanobacteria, respectively, while the corresponding proportions where GD dominates are 38% and 35%, respectively (Fig. 4). In another study of LGT among Cyanobacteria limited to monolog gene families, Shi and Falkowski (2008) investigated 682 families, using visual cluster identification in a principal coordinates analysis of topology distances. They report a core cluster, comprising 48% of the monologs, with the remaining proportion was ascribed to LGT. To allow comparison, we performed a corresponding PrIME-DLTRS analysis of our Cyanobacteria monolog families. It shows that a slightly lower proportion, 40%, displays the species topology (Supplementary Fig. S8). However, the proportion where LGT is actually dominating is 38% of the monologs, while GD is the dominating event in 15% of the cases. Accounting for potential transfer events from closely related species in the GD number, this gives an estimate of 38–53% of monolog families subject to LGT, which is in accordance with the findings of Shi and Falkowski (2008). This result also shows that restricting analysis to monologs may not provide representative estimates of LGT and GD for the whole gamut of gene family sizes (compare Fig. 4 and Supplementary Fig. S8).
The complexity hypothesis for LGT (Jain et al. 1999), which posits that informational genes—typically members of large complex systems—are less prone to LGT than operational genes, finds support in our results, similarly to what was reported by Shi and Falkowski (2008) and Abby et al. (2012). For both Mollicutes and Cyanobacteria, there is a clear enrichment of translation-related functional terms among gene families with low GD and LGT rates, for example, those with a gene tree resembling the species tree (SM Discussion and Supplementary Table S7).

Identified highways of LGT—Cyanobacteria can be functionally divided into two major groups based on their photosynthesis system (Jain et al. 1999; Rae et al. 2011). The α-Cyanobacteria, which all are marine organisms, form a monophyletic subtree in
our species tree, while the sampled β-Cyanobacteria, which comprise terrestrial, hot spring, and freshwater and filamentous colonial species, are dispersed on the three remaining subtrees of our Cyanobacteria species tree; henceforth these subtrees are referred to by the mnemonics (α, β₁, β₂₀, and β₂₁) listed in Supplementary Table S4, see also Figure 3b. LGT is proposed to have been instrumental in the evolution of the Cyanobacteria photosynthesis system, but also to have enhanced the adaptation of β-Cyanobacteria to extreme habitats, including freshwater, hot spring, and terrestrial domains (Jain et al. 1999; Rae et al. 2011). We will for this reason focus on LGT between the subtrees depicted in Figure 3b (we will indicate a highway between two subtrees with ⇔, e.g., α ⇔ β₁).

Our analysis, see SM Discussion, identified four Cyanobacteria LGT highways (Beiko et al. 2005) representing transfer of several genes between Cyanobacteria subtrees. Highways have occurred between all pairs of β-Cyanobacteria subtrees, and additionally between α and β₁ (Supplementary Table S6). Interestingly, the 153 gene families affected by these highways include a fundamental gene in the Cyanobacteria photosynthesis system (Rae et al. 2011). We also use, what we here call, highway-inducing topologies in a technical argument that limits how
much our rate estimates may be affected by LGT from extinct or unsampled distant species (see SM Discussion). Moreover, we show that the highways identified by PrIME-DLTRS are unlikely to be recovered by either synteny or sequence composition methods (SM Discussion).

Comparison with MrBayes.—To see how a well-proven phylogenetic approach uninformed of the species tree would perform, we compared PrIME-DLTRS with MrBayes (Ronquist and Huelsenbeck 2003) on gene tree inference, using identical sequence evolution models. Using the maximum a posteriori (MAP) gene trees from this analysis, we made the following observations.

First, for MrBayes, a monolog gene tree with a topology identical to that of the species tree was identified in only one Mollicutes family, and not at all among the Cyanobacteria monologs. In contrast, PrIME-DLTRS identified the species tree in 32 (33%) and 137 (29%) families for Mollicutes and Cyanobacteria, respectively. Moreover, the MrBayes tree distribution across the monolog families was very wide: among the 98 Mollicutes and 469 Cyanobacteria families, there were 83 and 255 different MAP trees, respectively, with the most common tree appearing a mere 12 times. The corresponding numbers for PrIME-DLTRS were substantially smaller: 58 and 178 different trees for Mollicutes and Cyanobacteria, respectively. It seems likely that conflicting phylogenetic signals reported by, for example, Shi and Falkowski (2008) and Zhaxybayeva et al. (2006) also stem from not accounting for the species tree already in the inference stage. Second, we applied the parsimony-based LGT estimation algorithm PHYLTR (Tollig et al. 2011) to both the MrBayes and the PrIME-DLTRS MAP gene trees. The results are shown in Figure 5. It is evident that the PrIME-DLTRS trees yield considerably more parsimonious results. For Mollicutes, PrIME-DLTRS gives on average 1.3 events and MrBayes 2.4 events, while for Cyanobacteria, PrIME-DLTRS gives on average 1.1 events and MrBayes 2.5 events. That is, MrBayes produces on average more than one additional event. Thus, by integrating sequence evolution and duplication-loss-transfer, and so accounting for constraints implied by the species tree and its edge lengths, PrIME-DLTRS provides more accurate gene trees and better estimates of the number of events.

Third, while there is a very good correlation between the total birth rate (i.e., GD rate + LGT rate) and the minimum number of GD and LGT events needed to reconcile a MAP gene tree and its corresponding species tree (Supplementary Fig. S9), it is clear that in reporting a posterior distribution, PrIME-DLTRS gives a more comprehensive analysis. This is in line with previously reported results for probabilistic orthology methods (Sennblad and Lagergren 2009), and also with phylogenetics folklore.

We also conducted a comparison with MrBayes on synthetic data created with rates similar to those estimated for the bacterial data sets (SM Discussion). The analysis found that the symmetric Robinson–Foulds tree distance (Robinson and Foulds 1981) to the true topology was significantly lower on average for trees inferred with PrIME-DLTRS than MrBayes (Supplementary Fig. S15), corroborating the bacterial monolog results (Fig. 5).

**DISCUSSION**

From a technical perspective, our comparative tests show a clear beneficial effect from using the integrated,
cannot merely consist of one deleterious mutated gene event, a biological model respecting our observation copy-number of the gene families exposed to this. Since we show that GD typically has increased the well as inhibit accumulation of deleterious mutations. that such events, analogously to sexual reproduction colonization of new habitats and diversification (Jain 1999; Pal et al. 2005; Fournier and Gogarten 2008; Dieterich and Sommer 2009; Rae et al. 2011; for a fuller review, see Swithers et al. 2012). We identify several highways between the major clades of Cyanobacteria, indicating that numerous LGTs have occurred, for example, of genes in the photosynthesis system. This enables a Bayesian analysis of GD, GL, LGT, and sequence evolution. PriME-DLTRS can be applied on single gene families or genome wide. Our work also paves the way for models that further intertwine GD and LGT with sequence evolution in order to accommodate neo- and subfunctionalization, and loss of function, as well as models that extend our DLTS submodel, say, by introducing different loss rates following GD or LGT events. Models for GD, GL, and LGT may also be applied in other situations, for example, evolution of parasite species with respect to host species (Haftner et al. 2000; Huelsenbeck et al. 2000).

Arguably, uncertainties in species evolution pose a problem for our phylogenetic LGT method, as well as for others that use a species tree as a backbone. The effect of incorrect species edges for inferring LGT activity warrants further investigation. In our setting, while we use results from recent studies (Gupta and Mathews 2010; Abby et al. 2012), the exact resolution of Prochlorococcus and marine Synechococcus strains in Cyanobacteria is debated and remains to be established. Also, the dates of the \( \tau_{p} \) and the \( \text{c} \) cyanobacterial clade root branches deviate from current fossil calibration estimates, possibly due to differences in evolutionary rates. We note, however, that the effect of rescaling the mentioned branches would most likely have a minor effect on the identified cyanobacterial highways. Ultimately, probabilistic LGT-aware multi-locus species tree reconstruction seems like the way to resolve such issues, and we believe DLTRS provides a major step down that path.

SUPPLEMENTARY MATERIAL
Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.78r2.

FUNDING
This work was supported by the Swedish Research Council [2010-4757 to J.L., and 2010-4634 to L.A.]. B.S.'s position is supported by a Karolinska Institutet
**ACKNOWLEDGMENTS**

J.L. and A.T. devised the algorithmic underpinnings; J.S. developed the software; J.L., J.S., B.S., and L.A. performed analyses; V.D. provided the data; all authors wrote, read, and approved the final manuscript.

**REFERENCES**


