HGTs are not SPRs: In the presence of ghost lineages, series of Horizontal Gene Transfers do not result in series of Subtree Pruning and Regrafting

Eric Tannier[1,2], Théo Tricou[3], Syrine Benali[1] and Damien M. de Vienne[1,*]

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

[2] Inria Lyon Research Center, F-69622 Villeurbanne, France

[3] Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA, Saint-Genès-Champanelle, France

* Corresponding author:

DAMIEN M. DE VIENNE

Université Lyon 1, CNRS

Laboratoire de Biométrie et Biologie Évolutive

Bâtiment Mendel 43 boulevard du 11 Novembre 1918

69622 VILLEURBANNE CEDEX

E-mail: damien.de-vienne@univ-lyon1.fr

Abstract

When a gene is horizontally transferred (HGT), under the "replacement" model where the transferred gene replaces its homolog in the recipient genome, the corresponding gene phylogeny departs from the species phylogeny by a Subtree Prune and Regraft (SPR) operation: the recipient branch is moved from its initial position to attach to the donor branch. Based on this observation, various methods have used SPRs to simulate or infer HGTs. We examine this apparent equivalence in the light of ghost lineages, *i.e.* all related species absent from the phylogeny because they are extinct, unknown or have not been sampled. These absent species branches are much more

numerous than the visible ones under most situations, so HGTs have great chances of originating from ghosts. We show that in this case, there is a high probability that series of HGTs do not correspond to series of SPRs. This biases the simulations under SPRs and prevents the inference under SPRs from retrieving the true transfers. For instance if the studied phylogeny represents 1% of all the species susceptible to exchange genetic material with the 100 sampled species, and 11 transfers occurred, then modelling HGTs with SPRs will be misleading in around 50% of the cases. This applies to a certain extent to other phylogenetic simulation or inference methods of HGT, like reconciliation.

Introduction

Horizontal (or lateral) transfers depict the transmission of genetic material from one organism to another that is not its direct descendant, as opposed to the vertical transmission. This mode of transmission is particularly common in bacteria and archaea, where it is an essential evolutionary process that contributes to the adaptation of lineages to their environment (Arnold et al. 2022; Gophna and Altman-Price 2022). In eukaryotes, it is being increasingly recognized as an important process as well (Keeling 2024).

The horizontal transfer of a gene (HGT) has a theoretical effect on phylogenies, which is commonly exploited for its detection (Ravenhall et al. 2015): if we represent the evolutionary history of genomes by a species phylogeny and the evolutionary history of a considered gene by a gene tree, an HGT can introduce a difference between the two trees. Indeed, consider a simple model where genes do not undergo duplications, losses or other events that can depart their evolutionary history from that of the species, and where an HGT consists of the replacement of a gene by its homolog in another lineage (it is a replacing transfer, Choi et al. 2012). In this model an HGT results in a Subtree Prune and Regraft (SPR) operation (Allen and Steel 2001): the gene tree topology resulting from the HGT can be obtained from the species tree topology by cutting (pruning) the recipient branch of the transfer and reattaching it (regrafting) at the location of the donor branch.

This correspondence between HGT and SPR is the basis of several HGT detection and simulation methods using transformation of trees by SPR operations (Ravenhall et al. 2015; Chan et al. 2017). Various algorithms and methods have been designed to determine sequences of SPR operations (or "edit paths") that transform species trees into gene trees (Hallett and Lagergren 2001; Nakhleh et al. 2005; Suchard 2005; Beiko and Hamilton 2006; Hill et al. 2010; Avni and Snir 2020), such edit paths being interpreted in terms of horizontal transfers (e.g. Beiko 2005; Whidden et al. 2014). Tools have also been developed to simulate gene trees evolving in the presence of horizontal gene transfers by performing SPRs in the species tree following HGTs (e.g., Horiike et al. 2011; Mallo et al. 2016). A founding assumption of all these methods is an extension of the translation of an HGT into an SPR in a phylogeny: a series of HGT should result in a series of SPRs.

We examine this correspondence between HGT and SPR in the light of an important parameter of the study of life evolution: the fact that we don't have access to most species that live or have lived on earth. Indeed, the vast majority of species are either extinct (Raup 1991) or extant but unknown (Mora et al. 2011; Locey and Lennon 2016), and even for known species, only a tiny fraction is susceptible to be included in any analysis carried out. Phylogenetically speaking, this translates to phylogenetic trees often representing the evolutionary history of a small subset of all the species that descended from their root. Absent species or clades — referred to as ghost lineages — were shown to be of particular importance for the study of horizontal genetic flows (Durvasula and Sankararaman 2020; Ottenburghs 2020; Tricou et al. 2022a, 2022b) because given their overwhelming abondance, it is likely that they are involved, as donors or recipients, in some (or most) of the detected HGTs (Macleod et al. 2005; Galtier and Daubin 2008; Fournier et al. 2009; Szöllősi et al. 2013).

The importance of ghosts when modelling HGTs as SPRs - Illustration

In this context, we propose to investigate the impact of ghost lineages on the robustness of modelling HGTs as SPRs. This requires to differentiate "full trees", *i.e.* species tree and gene tree

including ghosts, and "sampled trees", *i.e.* the species and gene trees from which ghost lineages have been pruned. Of course in an evolutionary study one can only have access to sampled trees ; full trees are theoretical objects with many more species (extinct and extant) and including the sampled ones. HGTs on sampled species trees may differ from HGTs on full species trees depending on whether donors or recipients branches in the full species tree are ghost lineages or not. If the recipient of the transfer is on a ghost branch, then the HGT will become invisible to an evolutionary study (absent from the sampled trees). If the donor and recipient are both on non-ghost branches, then the HGT is unchanged between sampled and full species tree. The most interesting case is when the recipient is on a non-ghost branch but the donor is on a ghost branch (Figure 1), which is the most common case of detectable transfers if we assume that ghost lineages are the most numerous. In this case, the visible result of the HGT (i.e. the HGT in the sampled species tree resulting from the HGT in the full species tree), which we call the "induced HGT", has the same non-ghost recipient, but its donor is the most recent ancestor of the ghost donor in the full species tree that is a non-ghost branch. In Figure 1A for example, the HGT from F to (DE) becomes the "induced HGT" from (ABCDE) to (DE), *i.e*. from the branch in the sampled species tree that was initially supporting the ghost lineage (F in this example).

To evaluate the robustness of treating HGTs as SPRs in the presence of ghosts, we can compare the gene tree that is produced when performing SPRs following the induced HGTs on the sampled species tree (Figure 1A, bottom right), with the gene tree obtained by sampling the full gene tree obtained by performing SPRs following HGTs on the full species tree (Figure 1A, bottom left). When a single HGT is present (Figure 1A), we observe that these two gene trees are topologically identical. In this simple case, SPRs can correctly model HGTs even when the donor branch of the HGT is a ghost. However, the situation is different when two or more transfers emerge from the same ghost branch (Figure 1B).

Translating the two HGTs in the full species tree into "induced HGTs" in the sampled species tree produces the two transfers highlighted in the top right of Figure 1B: the transfers F to (DE) and F to

G have (ABCDE) to (DE) and (ABCDE) to G as induced transfers, because the new donor branch of the transfer is the branch in the sampled tree on which the ghost species F was attached in the full species tree. We observe that performing SPRs on the species tree following these induced transfers produces a gene tree (the induced gene tree in Figure 1B, bottom right) that is topologically different from the correct gene tree (the sampled gene tree in Figure 1B, bottom left). In conclusion, with these toy examples (Figure 1) we see that treating HGTs as SPRs in the presence of ghost lineages may not produce the correct gene tree. For a single HGT, an SPR following the induced transfer has the same effect on the phylogenies as the SPR on the full tree followed by sampling, but this property does not generalise to more than one HGT in the same phylogeny, as soon as two (or more) transfers occur from the same ghost clade.

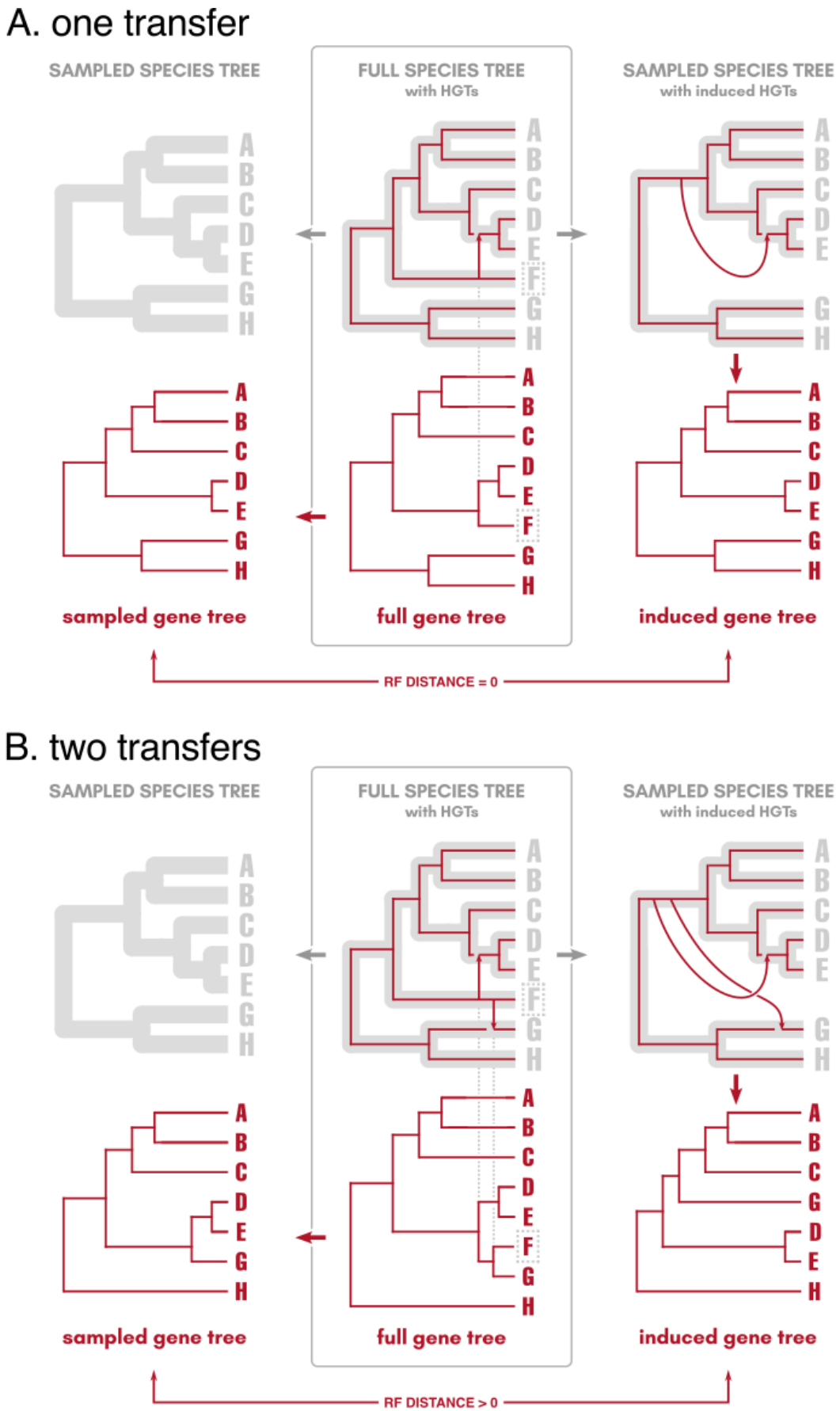

Figure 1. Illustration of the effect of ghosts on the possible disconnection between HGTs and SPRs. Each panel (**A** and **B**) is organised the same way, the only difference is the number of HGTs represented. For each panel, from a full

species tree and a set of HGTs (top of the central box), it is possible to obtain the full gene tree (bottom of the central box) by performing SPRs on the full species tree following the HGTs. From the full species tree and the full gene tree (central box), it is direct to obtain the sampled species tree and the sampled gene tree (left part) by pruning the ghost species (F in this example) in both trees. These sampled trees are the only ones that can be observed in biological studies. From the full species tree and the known HGTs, it is also possible to compute the "induced HGTs" (top right), *i.e.* the HGTs in the sampled tree that correspond to the HGTs in the full tree. "Induced HGTs" simply correspond to mapping HGTs into the sampled species tree by identifying what branches in the sampled species tree are the donors and recipient branches of the HGTs present in the full species tree (see text). The induced gene tree is the gene tree resulting from performing SPRs in the sampled species tree following the induced HGTs. We note that when more than one transfer involves the same ghost donor branch (panel **B**), performing these SPR operations leads to a gene tree that is topologically different from the sampled gene tree (RF distance between the two trees is not null). This is not the case when a single transfer (panel **A**) exists.

The importance of ghosts when modelling HGTs as SPRs - Quantification

We performed simple simulations (code and data available at https://github.com/damiendevienne/hgts-are-not-sprs) to explore and quantify the possible effect of this disconnection between HGTs and SPRs due to ghosts. This approaches the possible error rate of a method using SPRs to detect or simulate HGTs. We estimated the proportion of cases where the sampled gene tree and the gene tree obtained by applying the induced transfers to the sampled species tree produce different topologies. This was done by repeating (100 replicates) the following operations for different numbers of transfers (t=2, 5, 8, 11, 14, 17 or 20) and different numbers of species sampled (n=50, 100, 500 or 1000, corresponding to 0.5%, 1%, 5% or 10% of the total number of species):

- Simulate a tree with 10 000 tips (full species tree T)
- Randomly sample *n* tips as being non-ghosts to produce the sampled species tree S
- Randomly choose donor and recipient branches (contemporaneous in the full tree) of *t* HGTs, ensuring that each transfer has an extant branch as recipient
- Compute the "sampled" gene tree, by

    - Performing SPRs on T following the HGTs
    - Sampling the *n* tips on the obtained tree
- Compute the "induced" gene tree, by
    - Predicting the induced HGTs based on *t*, *n* and N
    - Performing SPRs on S following the induced HGTs
- Compute the topological distance (RF distance, Robinson and Foulds 1981) between the two gene trees

We observe (Figure 2) that the more ghosts there are, the more likely it is that the sampled and the induced gene trees will topologically disagree. It also appears that the proportion of cases where there is an incongruence increases with the number of HGTs, which was expected. If 10% of the species are sampled (90% of the species are ghosts) and there are 11 transfers involving an extant branch as recipient, around 7% of the sampled and induced gene trees will disagree. This percentage increases to around 50% if only 1% of the species are sampled (99% of the species are ghosts) and to 70% if 0.5% are sampled (Figure 2).

Relating this to methods that use SPRs to infer HGTs, this means that if the species tree on which the method is used represents 1% of the total number of species that really populate this tree, and if 11 HGTs really occurred, trying to interpret the SPRs in terms of HGTs will be misleading 50% of the time because SPRs on the tree are not expected to produce the observed gene tree.

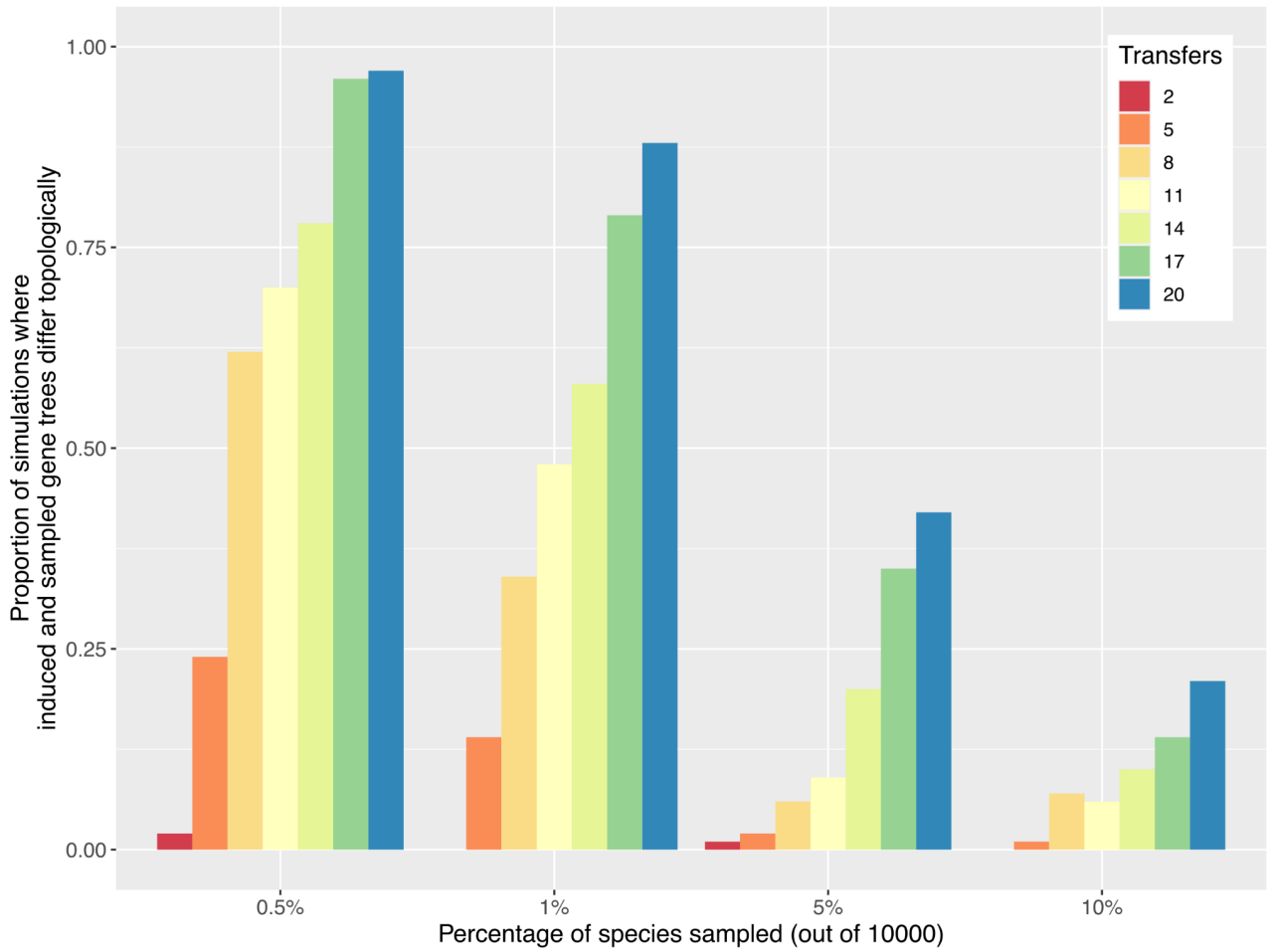

Figure 2. Effect of the amount of ghost lineages (x-axis) and of the number of horizontal gene transfers (HGTs) on the proportion of cases where the sampled gene trees and the induced gene trees (see text) are topologically different. This proportion approximates the possible error rate of a method using SPRs to detect or simulate HGTs.

Conclusions

Our results show that in the presence of ghosts, no series of SPRs transforming the extant species tree into the extant gene tree can be interpreted in terms of the effective HGT. As a consequence, no method transforming phylogenetic trees by the means of SPR can detect the true HGTs (the induced ones) if the species tree is not exhaustive (contains no ghosts), which is highly unlikely. This calls for great care in the interpretation of SPR edit paths in terms of HGTs in future studies.

To obtain our results, we had to define the concept of "induced HGTs", i.e. horizontal transfers in a given sampled species tree corresponding to the horizontal transfers occurring in the full species tree. Note that this concept, that should be useful for future research on the impact of ghosts for the

study of gene flow, is present — although defined conversely, from the sampled tree instead of the full tree — in the HGT inference literature. There, it is termed “transfers to the future” and depicts the combination of a speciation to an unsampled lineage followed by a transfer back to a sampled one (see Menet et al. 2022 for a recent review).

Finally, We want to emphasise, as a more positive note, that recent methods for reconciling gene and species trees, such as DTL methods (that infer Duplications, Transfers and Losses from the comparison between gene and species trees)(Szöllősi et al. 2015; Jacox et al. 2016) may not be subject to the issue demonstrated here, as they theoretically allow the diversification of genes outside the given species tree. However few software seem to have implemented this possibility, so in their present form might be subject to the same error risks.

Funding

This work was supported by the French Agence Nationale de la Recherche under grant ANR-19-CE45-0010 (Evoluthon).

References

Allen B.L., Steel M. 2001. Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees. *Annals of Combinatorics*. 5:1–15.

Arnold B.J., Huang I.-T., Hanage W.P. 2022. Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology*. 20:206–218.

Avni E., Snir S. 2020. A New Phylogenomic Approach For Quantifying Horizontal Gene Transfer Trends in Prokaryotes. *Scientific Reports*. 10:12425.

Beiko R.G. 2005. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences*. 102:14332–14337.

Beiko R.G., Hamilton N. 2006. Phylogenetic identification of lateral genetic transfer events. *BMC Evolutionary Biology*. 6:15.

Chan C.X., Beiko R.G., Ragan M.A. 2017. Scaling Up the Phylogenetic Detection of Lateral Gene Transfer Events. In: Keith J.M., editor. Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution. New York, NY: Springer. p. 421–432.

Choi S.C., Rasmussen M.D., Hubisz M.J., Gronau I., Stanhope M.J., Siepel A. 2012. Replacing and Additive Horizontal Gene Transfer in Streptococcus. *Molecular Biology and Evolution*. 29:3309–3320.

Durvasula A., Sankararaman S. 2020. Recovering signals of ghost archaic introgression in African populations. *Science Advances*. 6:eaax5097.

Fournier G.P., Huang J., Gogarten J.P. 2009. Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 364:2229–2239.

Galtier N., Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. *Philos Trans R Soc Lond, B, Biol Sci*. 363:4023–9.

Gophna U., Altman-Price N. 2022. Horizontal Gene Transfer in Archaea—From Mechanisms to Genome Evolution. *Annual Review of Microbiology*. 76:481–502.

Hallett M.T., Lagergren J. 2001. Efficient algorithms for lateral gene transfer problems. *Proceedings of the fifth annual international conference on Computational biology*.:149–156.

Hein J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences*. 98:185–200.

Hickey G., Dehne F., Rau-Chaplin A., Blouin C. 2008. SPR Distance Computation for Unrooted Trees. *Evolutionary Bioinformatics Online*. 4:17–27.

Hill T., Nordstrom K.J., Thollesson M., Safstrom T.M., Vernersson A.K., Fredriksson R., Schioth H.B. 2010. SPRIT: Identifying horizontal gene transfer in rooted phylogenetic trees. *BMC Evol Biol*. 10:42.

Horiike T., Miyata D., Tateno Y., Minai R. 2011. HGT-Gen: a tool for generating a phylogenetic

tree with horizontal gene transfer. *Bioinformation*. 7:211–213.

Jacox E., Chauve C., Szöllősi G.J., Ponty Y., Scornavacca C. 2016. ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*. 32:2056–2058.

Keeling P.J. 2024. Horizontal gene transfer in eukaryotes: aligning theory with data. *Nature Reviews Genetics*. 25:416–430.

Kordi M., Kundu S., Bansal M.S. 2019. On Inferring Additive and Replacing Horizontal Gene Transfers Through Phylogenetic Reconciliation. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*.:514–523.

Locey K.J., Lennon J.T. 2016. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*. 113:5970–5975.

Macleod D., Charlebois R.L., Doolittle F., Bapteste E. 2005. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evol Biol*. 5:27.

Mallo D., De Oliveira Martins L., Posada D. 2016. SimPhy : Phylogenomic Simulation of Gene, Locus, and Species Trees. *Systematic Biology*. 65:334–344.

Menet H., Daubin V., Tannier E. 2022. Phylogenetic reconciliation. *PLOS Computational Biology*. 18:e1010621.

Mora C., Tittensor D.P., Adl S., Simpson A.G.B., Worm B. 2011. How Many Species Are There on Earth and in the Ocean? *PLoS Biology*. 9:e1001127.

Nakhleh L., Ruths D., Wang L.-S. 2005. RIATA-HGT: A Fast and Accurate Heuristic for Reconstructing Horizontal Gene Transfer. *Computing and Combinatorics*.:84–93.

Ottenburghs J. 2020. Ghost Introgression: Spooky Gene Flow in the Distant Past. *BioEssays*. 42:2000012.

Raup D.M. 1991. Extinction: bad genes or bad luck? New York: W.W. Norton.

Ravenhall M., Škunca N., Lassalle F., Dessimoz C. 2015. Inferring horizontal gene transfer. *PLoS computational biology*. 11:e1004095.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*. 53:131–147.

Suchard M.A. 2005. Stochastic Models for Horizontal Gene Transfer: Taking a Random Walk Through Tree Space. *Genetics*. 170:419–431.

Szöllősi G.J., Davín A.A., Tannier E., Daubin V., Boussau B. 2015. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 370:20140335.

Szöllősi G.J., Tannier E., Lartillot N., Daubin V. 2013. Lateral Gene Transfer from the Dead. *Systematic Biology*. 62:386–397.

Tricou T., Tannier E., de Vienne D.M. 2022a. Ghost lineages highly influence the interpretation of introgression tests. *Systematic Biology*. 71:1147–1158.

Tricou T., Tannier E., de Vienne D.M. 2022b. Ghost lineages can invalidate or even reverse findings regarding gene flow. *PLOS Biology*. 20:1–16.

Whidden C., Zeh N., Beiko R.G. 2014. Supertrees Based on the Subtree Prune-and-Regraft Distance. *Systematic Biology*. 63:566–581.