SUPPLEMENTARY MATERIAL

For "RANGER-DTL 2.0: Rigorous Reconstruction of Gene-Family Evolution by Duplication, Transfer, and Loss", by Mukul S. Bansal, Manolis Kellis, Misagh Kordi, and Soumya Kundu

S1 USING UNDATED SPECIES TREES FOR RECONCILIATION

The primary reason for using *dated* species trees in DTL reconciliation is that optimal reconciliations can sometimes be *temporally infeasible*, in that the inferred transfer events can sometimes impose contradictory temporal constraints. Computing an optimal temporally feasible DTL reconciliation is NP-hard (Tofigh *et al.*, 2011). However, when a fully-dated species tree is used, one can compute an optimal temporally feasible DTL reconciliation in polynomial time by imposing constraints on possible transfer events (Doyon *et al.*, 2010). Still, in RANGER-DTL 2.0 we focus primarily on *undated* species trees for two important reasons: First, it can be difficult to accurately date even small microbial species trees, e.g., Rutschmann (2006); Kuo and Ochman (2009), and nearly impossible for larger species trees with hundreds of microbial taxa. And second, even with an accurately dated species tree, actual transfer events need not always respect the temporal constraints imposed by these dates. This happens because of unsampled and extinct lineages (unavoidable in any dataset), so that any transfers from such lineages appear to originate at the point where the unsampled/extinct lineage attaches to the observed species tree and therefore appear to go forward in time, possibly resulting in temporally infeasible DTL reconciliations. In addition, DTL reconciliation with undated species trees has been extensively tested and shown to be very accurate (Bansal *et al.*, 2015). The accuracy of DTL reconciliation with undated species trees can be further improved by explicitly modeling transfers from extinct/unsampled lineages. However, while there has been progress in handling transfers from extinct lineages when the species tree is fully dated, e.g. (Jacox *et al.*, 2016), more research is needed to assess the impact of allowing transfers from unseen extinct/unsampled lineages on DTL reconciliation accuracy and to identify effective techniques for handling such transfers when the species tree is only partially dated or undated.

S2 COMPARISON WITH EXISTING SOFTWARE

Several software packages exist for DTL reconciliation. These include Notung (Stolzer *et al.*, 2012), ecceTERA (Jacox *et al.*, 2016), Jane (Conow *et al.*, 2010), CoRe-PA (Merkle *et al.*, 2010), and EUCALYPT (Donati *et al.*, 2015). The programs Jane, CoRe-PA, and EUCALYPT are designed for cophylogeny analysis (e.g., comparing evolutionary histories of hosts and parasites), but they can also be used for gene tree-species tree DTL reconciliation. Among the cophylogeny analysis software, EUCALYPT is designed to generate all optimal reconciliations and can also test for temporal infeasibility and output only those optimal reconciliations that are temporally feasible. EUCALYPT cannot handle unrooted or unresolved gene trees, cannot aggregate across reconciliations to compute support values, cannot consider distance dependent transfer costs, and is not scalable beyond a couple of hundred taxa. The software Jane uses a heuristic to compute only temporally feasible reconciliations (using undated species trees) and has several advanced features such as allowing variable transfer costs, computing support values for reconciliations, and heuristically resolving unresolved gene trees. However, Jane cannot handle unrooted gene trees or consider all optimal resolutions of unresolved gene trees, might produce suboptimal solutions due to its heuristic solution for computing only temporally consistent solutions, cannot generate or uniformly sample all optimal reconciliations, and is not scalable beyond a few hundred taxa. Likewise, CoRE-PA offers some advanced features such as automatic estimation of event costs, but cannot work with unrooted gene trees, cannot consider all optimal resolutions of unresolved gene trees, does not perform uniform random sampling of optimal reconciliations or compute support values, and is not scalable beyond a few hundred taxa.

Notung implements a modified version of DTL reconciliation (Tofigh *et al.*, 2011; Bansal *et al.*, 2012), called DTLI reconciliation (Stolzer *et al.*, 2012), which allows Notung to use incomplete lineage sorting to explain and handle any uncertainty in the species tree topology. Notung also checks for temporal feasibility of optimal reconciliations and can generate all optimal reconciliations that are temporally feasible (possibly none). Despite these advanced features, Notung cannot compute DTL reconciliations for non-binary gene trees, cannot simultaneously consider all possible optimal gene tree rootings, and cannot aggregate across multiple optima or across reconciliations computed using different event costs. In terms of features and scalability, the program most comparable to RANGER-DTL 2.0 is ecceTERA. The program ecceTERA offers many useful features, such as computing multiple optima, considering the effect of different event costs, computing support values, etc. The program can also root unrooted gene trees and optimally resolve unresolved gene trees. However, ecceTERA cannot account for all optimal rootings of unrooted gene trees and cannot compute and aggregate across all optimal resolutions of unresolved gene trees. It also cannot consider variable transfer costs.

To summarize, RANGER-DTL 2.0 offers important functionality not available in any existing reconciliation program.

S3 ACCURACY OF INFERRED RECONCILIATIONS

Several previous studies have demonstrated the accuracy of parsimony-based models of DTL reconciliation (Doyon *et al.*, 2010; Nguyen *et al.*, 2013; Bansal *et al.*, 2015). Here we report the results of some additional experiments we performed to specifically assess the accuracy of RANGER-DTL 2.0.

For these experiments, we generated simulated datasets of species trees and gene trees using the probabilistic simulation framework PrIME-GenPhyloData (Sjöstrand *et al.*, 2013). Specifically, we generated 100 species trees with unit height, each containing 100 leaves using a birth-death process and, inside each of the species trees, we generated three different gene trees using low, medium, and high rates of duplication, transfer, and loss events, using the probabilistic gene family evolution simulation framework implemented in PrIME-GenPhyloData. This resulted in three sets, low DTL, medium DTL, and high DTL, each with 100 gene-tree/species-tree pairs. To generate the low DTL gene trees, we used duplication, transfer, and loss rates of 0.133, 0.267, and 0.4, respectively; for the

medium DTL gene trees we used rates of 0.3, 0.6, and 0.9, respectively; and for the high DTL gene trees we used rates of 0.6, 1.2, and 1.8 respectively. These rates are based on rates observed in real data and capture both datasets with lower rates of these events and datasets with a very high rate of these events; see, e.g., (Bansal *et al.*, 2015). We point out that the loss rate for each set was assigned to be equal to the rate of duplication plus the rate of loss. This setting results in a large number of losses which obfuscate the evolutionary history, making it even more difficult to infer true reconciliations.

For the low DTL gene trees, the average gene tree leaf set size was 103.3, with an average of 5.5 transfer events and 2.6 duplication events per gene tree. For the medium DTL gene trees, the average gene tree leaf set size was 99.9, with an average of 10.2 transfer events and 4.7 duplication events per gene tree. For the high DTL gene trees, the average gene tree leaf set size was 106.1, with an average of 19.8 transfer events and 9.2 duplication events per gene tree.

We first evaluated the accuracy of RANGER-DTL in inferring the evolutionary event and species tree mapping (i.e., the reconciliation) for each internal node in the simulated gene trees. Specifically, we used the core program *Ranger-DTL* and computed a single optimal reconciliation for each gene-tree/species-tree pair, and compared the computed reconciliation against the true evolutionary history of that gene tree. To compute these reconciliations we used the standard event cost assignment of 1, 2, and 3, for losses, duplications, and transfers, respectively. We observed very high accuracy for inferring the correct event type (speciation, duplication, or transfer) at each gene tree node. For instance, for the low DTL gene trees, 99.7%, 98.5% and 97.4% of the gene tree nodes labeled as speciation, duplication, and transfer, respectively, in the computed reconciliations were inferred correctly. As expected, the accuracy decreases as the DTL rate increases; however, even for the high DTL gene trees, these percentages remained very high at 95%, 85%, and 95%, respectively. These results are shown in Figure 1(a). Overall, we found that RANGER-DTL infers nearly 100% of the events accurately for the low rate of DTL, 98% of the events for the medium rate of DTL, and 94% of the events for the high rate of DTL.

For mapping inference, we found that 99.4%, 97.4%, and 91.2% of all internal nodes were assigned the correct species node mapping for the low, medium, and high DTL gene trees, respectively. In fact, 99.3% 96.8%, and 89.2% of the internal nodes were assigned both the correct event type and correct mapping for the low, medium, and high DTL gene trees, respectively. Broken down by event type, we found that, for low DTL gene trees, 99.7%, 98.5%, and 92.9% of speciation, duplication, and transfer nodes, respectively, had both a correct event assignment and correct mapping assignment. Corresponding percentages for medium DTL gene trees were 98.5%, 92.8%, and 84.2%, respectively, and for high DTL gene trees they were 94.2%, 83.0%, and 72.8%, respectively. These results are shown in Figure 1(b). Thus, while the mapping accuracy for speciations remains very high even for high DTL gene trees, the mapping accuracies for duplications and transfers are more affected as the DTL rate increases. However, even for high DTL gene trees the mapping accuracy for duplications and transfers remains fairly high overall. Furthermore, even when the mapping is assigned incorrectly, it is usually "close" to the correct mapping on the species tree; for instance, even for the high DTL gene trees, only 12% of the nodes were assigned a mapping more than two nodes away from the true mapping and the average distance between the incorrectly mapped transfers and their correct mappings was only 2.8 nodes on average.

We also evaluated the accuracy of gene tree rooting through RANGER-DTL 2.0. Specifically, we suppressed the true root of each gene tree, resulting in unrooted gene trees, and used the core program *OptRoot* to compute all optimal rootings for each gene tree. We then checked how often the true rooting was inferred among these optimal rootings. We again used the standard event cost assignment of 1, 2, and 3, for losses, duplications, and transfers, respectively, to compute optimal rootings. We found that RANGER-DTL outputs the correct rooting of the gene tree among its reported optimal rootings for 91% of the gene trees generated with a low rate of DTL, for 81% of the gene trees generated with a medium rate of DTL, and for 52% of the gene trees generated with a high rate of DTL. As expected the accuracy of correct rooting inference falls with increasing DTL rate. It is worth noting that the average number of optimal rootings inferred per gene tree remained small at 1.2 for the low DTL gene tree, 1.5 for the medium DTL gene trees, and 2.6 for the high DTL gene trees. Crucially, we also found that the average Robinson-Foulds distance between each of the optimal rootings generated by RANGER-DTL and the actual rooting for the gene tree is only 0.4 for the low DTL gene trees, 0.9 for the medium DTL gene trees, and 2.5 for the high DTL gene trees, indicating that even when the correct rooting is not recovered, the inferred rooting is usually very close to the correct one.

S4 RANGER-DTL 2.0 SOFTWARE DETAILS

The RANGER-DTL software package consists of ten related programs designed to work together to support various reconciliation analyses. These ten programs are organized into (i) three *core programs*, which define the core functionality of RANGER-DTL 2.0, (ii) five *supplementary programs* that provide additional functionality such as handling of dated species trees or reconciliation of unresolved (non-binary) gene trees, and (iii) two *summary scripts* that use the core and supplementary programs to implement two particularly useful reconciliation analyses.

The three core programs are designed to be used together sequentially and consist of *OptRoot*, which computes all optimal roots for unrooted gene trees, *Ranger-DTL*, which computes a single optimal reconciliation, sampled uniformly at random among all optimal reconciliations, of a rooted gene tree with the given species tree, and *AggregateRanger*, which takes as input multiple randomly sampled optimal reconciliations (computed using *Ranger-DTL*) and combines them into a single 'aggregate' reconciliation showing support values for inferred events and mappings.

The five supplementary programs provide additional functionality and are designed to be used either instead of or in addition to the core programs. Specifically, *OptRoot-Dated* and *OptRoot-Fast* can be used instead of the core *OptRoot* program to use dated reconciliation or use a much faster (but not fully-featured) version of the program, respectively. Likewise, the programs *Ranger-DTL-Dated* and *Ranger-DTL-Fast* can be used instead of the core *Ranger-DTL* program. The fifth program, *OptResolutions*, provides additional functionality for handling gene tree topological uncertainty and can be used to collapse weakly supported branches in the gene tree and generate all optimal resolutions (minimizing the reconciliation cost

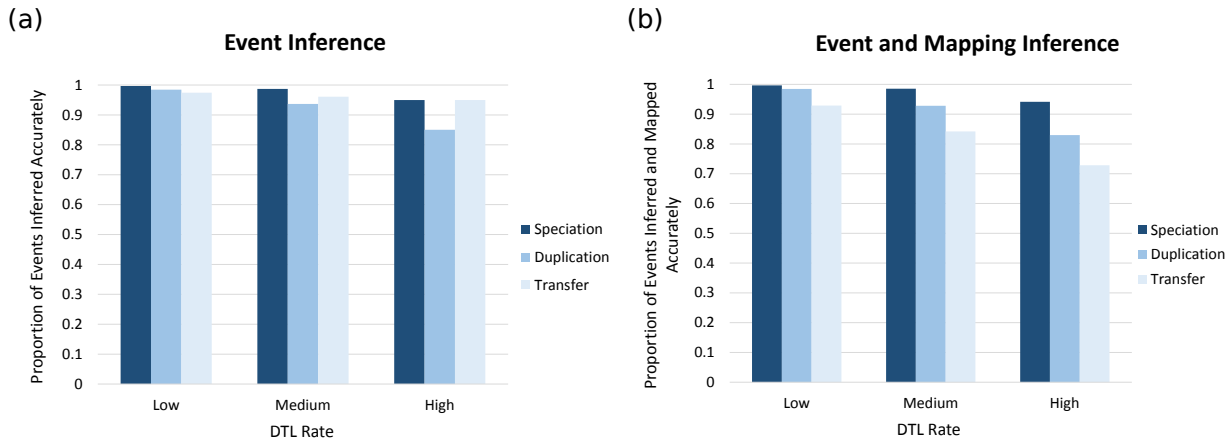

Fig. 1. Inference accuracy of RANGER-DTL 2.0**.** Part (a) shows the fraction of internal nodes across all low DTL, medium, DTL, and high DTL gene trees, whose event types, speciation, duplication, or transfer, are inferred correctly. Part (b) shows the fraction of internal nodes across all low DTL, medium, DTL, and high DTL gene trees, whose event types and mappings are both inferred correctly.

with the species tree). These resolutions can then be reconciled using the core *Ranger-DTL* program or its supplemental variants.

The two summary scripts *SummarizeOptRootings* and *SummarizeOptResolutions* are designed to simplify and automate those analyses that involve gene trees with multiple optimal roots and unresolved gene trees, respectively. The *SummarizeOptRootings* script uses *OptRoot* and *Ranger-DTL* (or their dated variants), along with *AggregateRanger*, to compute a "consensus reconciliation" across all optimal gene tree rootings (Kundu and Bansal, 2017). Similarly, the *SummarizeOptResolutions* script uses *OptResolutions* and *Ranger-DTL* to compute a summary reconciliation that aggregates reconciliations across all optimal resolutions.

REFERENCES

Bansal, M. S., Alm, E. J., and Kellis, M. (2012). Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, **28**(12), 283–291.

Bansal, M. S., Wu, Y.-C., Alm, E. J., and Kellis, M. (2015). Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics*, **31**(8), 1211–1218.

Conow, C., Fielder, D., Ovadia, Y., and Libeskind-Hadas, R. (2010). Jane: a new tool for the cophylogeny reconstruction problem. *Algorithm. Mol. Biol.*, **5**(1), 16.

Donati, B., Baudet, C., Sinaimeri, B., Crescenzi, P., and Sagot, M.-F. (2015). Eucalypt: efficient tree reconciliation enumerator. *Algorithms for Molecular Biology*, **10**(1), 3.

Doyon, J.-P., Scornavacca, C., Gorbunov, K. Y., Szöllosi, G. J., Ranwez, V., and Berry, V. (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In E. Tannier, editor, *RECOMB-CG*, volume 6398 of *Lecture Notes in Computer Science*, pages 93–108. Springer.

Jacox, E., Chauve, C., Szollosi, G. J., Ponty, Y., and Scornavacca, C. (2016). eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, **32**(13), 2056.

Kundu, S. and Bansal, M. S. (2017). On the impact of uncertain gene tree rooting on duplication-transfer-loss reconciliation. In *Bioinformatics Research and Applications - 13th International Symposium, ISBRA 2017*, page In press.

Kuo, C.-H. and Ochman, H. (2009). Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. *Biology Direct*, **4**(1), 35.

Merkle, D., Middendorf, M., and Wieseke, N. (2010). A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinf.*, **11**(Suppl 1), S60.

Nguyen, T.-H., Ranwez, V., Berry, V., and Scornavacca, C. (2013). Support measures to estimate the reliability of evolutionary events predicted by reconciliation methods. *PloS one*, **8**(10), e73667.

Rutschmann, F. (2006). Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. *Divers. Distrib.*, **12**(1), 35–48.

Sjöstrand, J., Arvestad, L., Lagergren, J., and Sennblad, B. (2013). Genphylodata: realistic simulation of gene family evolution. *BMC Bioinformatics*, **14**(1), 209.

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., and Durand, D. (2012). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, **28**(18), 409–415.

Tofigh, A., Hallett, M. T., and Lagergren, J. (2011). Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans. Comput. Biology Bioinform.*, **8**(2), 517–535.