Estimating Accurate Gene Trees in the presence of Intra-locus Recombination; a Simulation Study

Kevin Ziegler[1], and Lemmon A.[1]

[1]Department of Scientific Computing, Florida State University, Tallahassee, United States

Introduction

Genes trees are estimated for a variety of reasons in phylogenetics: understanding evolutionary history, estimating population size, estimating migration rates, estimating coalescent times, or estimating the historical rate and patterns of dispersal. For several applications, gene trees serving as input are assumed to be free of error[1][2]. Although some summary methods have statistical guarantees in the presence of Gene Tree Estimation Error (GTEE), this does not apply to any of the standard coalescent based methods [3]. The downstream effects of GTEE are also unknown for PhyloMapper and most other methods. In order to avoid unknown effects, maximizing gene tree accuracy is highly desirable.

Sources for gene tree estimation error come from a variety of sources. Insufficient signal is a common problem encountered when the mutation rate is low and locus length is short. Model mis-specification, missing data, and saturation, and method selection can also have a substantial impact on GTEE. Intra-locus recombination can also lead to GTEE when a single history is estimated using data containing a mixture of discordant histories. Unlike the many sources of error mentioned above, the effect of violating the standard assumption of recombination free loci has not well been explored.

This simulation study identifies the method that produces the most accurate gene tree for each region of parameter space. Users aware of their biological system's location in parameter space can use this information to make informed decisions about which methods best suit their needs. This study tests the accuracy of various methods in the presence of recombination. The simulation scenario mirrors human evolutionary history with, population history, recombination rate, population size, and substitution rate.

Methods

This simulation emulated the model of human evolution, the Out of Africa model.

The model's structure contained an ancestral population located in West Africa that branched into six major subgroups as humans dispersed to new continents. I matched tree divergence times and population sizes from a tree used in Huang 2019 [4]. Next gene trees were simulated from the historical population model with recombination.

The program used to simulate trees, Msprime, returns a whole genome genealogical history for a designated number of base pairs represented by a collection of genealogies [5]. We varied the following population history parameters: sequence length, recombination rate, and scaled population size. The values of these parameters varied broadly and centered around values matching human history. All other parameters were assigned as default. Simulations replicated each unique combination of parameters 20 times.

Jukes Cantor was chosen as the model of nucleotide substitution to simulate DNA sequences with differing substitution rates. Each contiguous gene tree output from msprime represents a recombination-free history for a program, Seq-gen, to simulate for a length of base pairs [6]. These contiguous simulated lengths concatenate together to form the overall alignment with many (possibly) different genealogies.

The accuracy of the following programs were compared. A standard maximum likelihood method, FastTree 2 [7], was used on 3 alignment partition strategies: c-genes (recombination free region), estimated c-genes, and subsampled alignments. Three programs which estimate whole genome histories, by simultaneously inferring gene trees and recombination breakpoints: Relate, Tsinfer, Rent+ [8][9][10].

Results

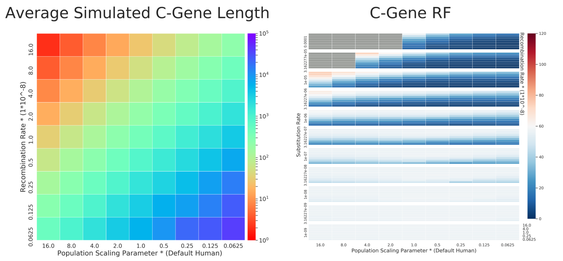


Figure 1. Average c-gene length produced by MS simulations over 20 replicates. The x-axis represents population scaling parameter (relative to humans), the y-axis represents the recombination rate (multiplied by the base human rate). At the extremes c-gene length varies from approximately 100,000 to 1.

Figure 2. Accuracy of gene tree estimates with known recombination breakpoints. Robinson Folds distance between simulated MS trees and c-gene trees estimated by Fast Tree (partition dataset into recombination free regions on simulated breakpoints). Each replicate produces an average RF distance across all c-genes for the whole 100,000 base pairs. Each square is the average of 20 replicates. The x-axis and y-axis are the same as in Figure 1, but each row represents a substitution rate ranging from 0.0001 and 1*10^-9. Regions with higher substitution rate and larger c-gene length have sufficient signal to build accurate trees. Gray boxes represent regions when less than 6 out of 20 replicates succeeded.

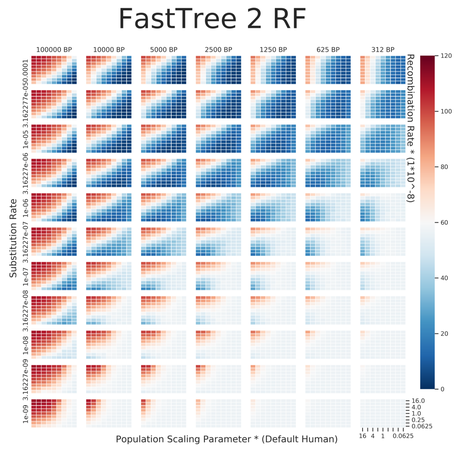


Figure 3. Accuracy of gene tree estimates in the presence of recombination. Robinson Folds distance between simulated MS trees and Tree estimated from concatenated loci of certain size. The axes remain the same as Figure 2, but there is a column for each length of concatenated gene. Large locus sizes struggle with obtaining an accurate gene tree for the whole locus; the trees are too discordant to fit one history. For each combination of recombination rate, substitution rate, and population size there is an optimal length of locus to balance signal from c-gene length and discordant signal from recombination.

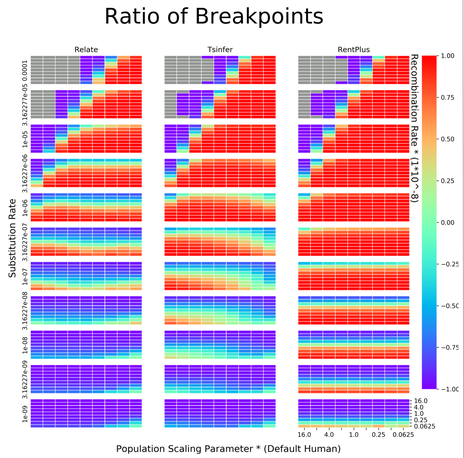
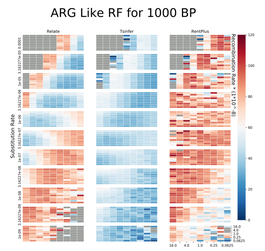


Figure 4. Ability of ARG methods to infer number of recombination breakpoints. Ratio of Estimated breakpoints to simulated break points for each ARG method. The ratio is computed as (Estimated BP – TrueBP)/(larger number); therefore, underestimation is blue and overestimation is red. All three methods do a poor job of capturing the total number of breakpoints. Some recombination events in MS do not impact tree topology or branch lengths, so perfect estimation is almost impossible, yet for most of parameter space the methods are far off. Gray boxes represent regions when less than 6 out of 20 replicates succeeded.



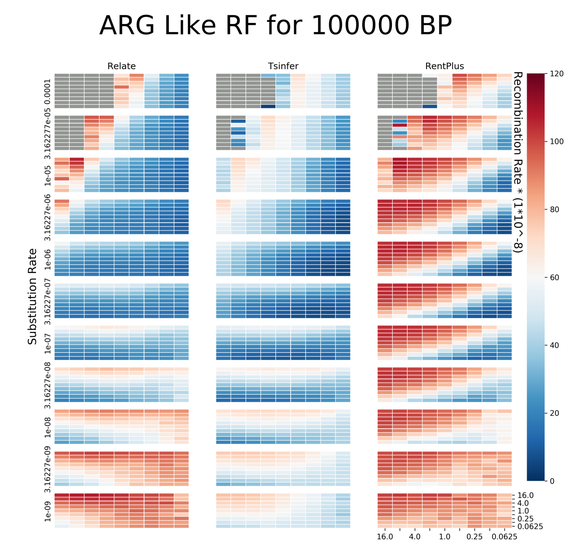


Figure 5. Ability of ARG methods to infer accurate gene trees. Boxes indicate Robinson Foulds distance between simulated MS trees and gene trees estimated by the ARG methods for 1000, 10000, and 100000 base pairs. There is variability between the performance of the methods across parameter space. In particular Tsinfer seems to be much more robust to changing parameters while still obtaining accurate gene trees for regions of "favorable" parameter space. Gray boxes represent regions when less than 6 out of 20 replicates succeeded.

The ability to estimate gene trees accurately varied widely across the broad parameter space explored in this study. Although some of the ARGL methods were fairly robust, producing accurate gene trees over a large proportion of parameter space, other approaches could only produce accurate gene trees in a narrow window of parameter space. The location of parameter space determines which methods or strategies are optimal.

The accuracy splitting an alignment into C-Genes performs best for regions of parameter space with higher substitution rates, lower recombination rates, and smaller population sizes. This strategy works provided there is enough genetic signal.

The accuracy of the second partitioning strategy, subdividing an alignment into loci of specified length, does well within the same region as the first strategy; high substitution rate, low recombination rate and low population size. Accuracy largely depends on the tradeoff between genetic signal and the set of discordant trees making up the locus. Both of these are affected by growing the locus length. For each region of parameter space there exists a locus length which performs the best in balancing these two factors. Overall, for most of parameter space at least one of the subsampled lengths performs well.

None of the ARG-like methods tested accurately capture the correct quantity of recombination breakpoints. There is a gradient of underestimation to overestimation. There are small regions in which the number of estimated and simulated breakpoints match, but since there is a gradient, they inevitably overlap.

ARGL methods were run on alignment lengths of 1,000 10,000 and 100,000 base pairs; results are program specific. Tsinfer and Relate perform well over medium substitution rates but perform slightly worse than Fast Tree 2 at higher substitution rates. At low substitution rates and low recombination rates, Relate and Tsinfer can harness limited information to build partially correct gene trees. Tsinfer performs the best out of the three ARGL programs, over most of parameter space, and Tsinfer seems to be the most robust to the effects of any of the simulation parameters. At medium and high substitution rates Relate is also robust to recombination rate and population size, but as substitution rate decreases Relate starts to struggle in regions with high recombination rate and low population size.

The accuracy of the third partitioning strategy, subdividing an alignment into estimated c-genes, works poorly for most of parameter space due to a lack of genetic signal (either the substitution rate is too low or there is an overestimation of breakpoints).

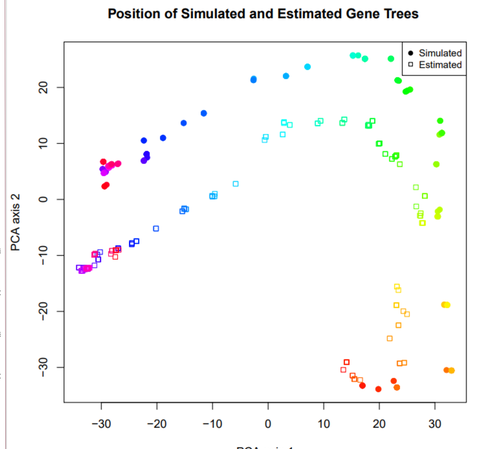


Figure 6. MDS plot of simulated trees and estimated Tsinfer trees for 33,000 base pairs of an example replicate (recombination scaling 1.0, population scaling 1.0, and substitution rate 0.0000001). Filled circles represent simulated trees, and open squares represent estimated trees. Color represents the location of the tree in the region of the 33,000 base pairs. Tsinfer's estimated trees are clearly near the simulated trees, but for this replicate there is some bias.

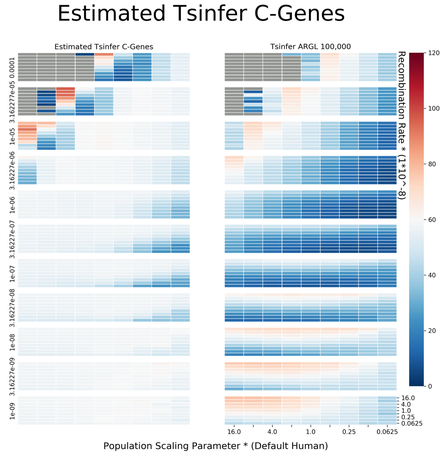


Figure 7. Accuracy of gene trees estimated on predicted c-genes in isolation. Similar to Figure 2, but simulated c-genes are replaced by c-genes estimated by Tsinfer. Robinson Foulds distance between simulated gene trees and c-gene (estimated by Tsinfer) trees estimated by Fast Tree (partition dataset into regions estimated to be recombination free). For most of parameter space there is insufficient signal to obtain an accurate gene tree. The ARG methods are using neighboring information to estimate smaller c-gene regions.

Discussion

Higher recombination rates and population sizes make estimating evolutionary histories more difficult. Recombination brings together different evolutionary histories on the same physical copy of DNA; this impacts gene tree accuracy because the DNA supports multiple evolutionary histories, yet the model tries to force these different histories under one evolutionary history.

Increasing population size increases the occurrence of incomplete lineage sorting. Incomplete lineage sorting occurs when a gene fails to coalesce before the most recent speciation event. The result of this can be that the gene history does not match the population history. Recombination can mix these different histories into parts of the genome for all individuals in the population.

ML methods perform very poorly in regions with high recombination rate and high population size. High recombination rate and population size impose clear limitations on locus length for obtaining an accurate gene tree. Longer loci have a greater chance to contain different evolutionary histories. As a result, the single estimated gene tree no longer reflects the evolutionary history of the whole locus. With a high recombination rate this issue negatively impacts accuracy more than increasing genetic signal positively impacts accuracy. For regions of low substitution rate and high recombination rate or population size, there is no locus length which returns an accurate gene tree. The optimal locus length to choose depends on the location in parameter space.

Within the partitioning schemes for ML methods, dividing on c-genes (first strategy) performs the best for regions in which ML methods are better than the ARG methods (regions with high sub and low pop and rec). These regions have long c-genes, so there is enough genetic signal within each c-gene and dividing on c-genes ensures there is no mixing of discordant histories. Partitioning on estimated c-genes (third strategy) does not perform well. In regions where the strategy could perform well Tsinfer overestimates breakpoints; this leads to a lack of genetic signal due to short alignments. Subdividing the alignment into loci of various sizes (second strategy) works better than the ARG like methods for regions of parameter space with high substitution, low recombination rate, and low population size. However, dividing on c-genes is still more accurate for the same regions of parameter space. For regions with high substitution rate, high recombination rate and high population size, creating gene trees from small partitions outperforms all other methods. Gene trees estimated from loci of size, 1250, 625, and 312 from this region of parameter space are more robust to recombination due to the small locus size, yet if the substitution rate is high enough a reasonable tree can still be reconstructed. C-gene size for this region is too small; it does not contain enough genetic signal to estimate an accurate gene tree.

Tsinfer is the best choice of obtaining accurate gene trees for an unspecified region of parameter space; it is the most accurate across much of parameter space and does not perform poorly in any region of parameter space.

Although ARGL methods divide the given locus into different sections, they do not accurately capture the number of recombination breakpoints; therefore, ARGL methods still break the assumption of no intra locus recombination but to a lesser degree. Simulated breakpoints might not change the gene tree, change the branch lengths, or change the topology; this makes accurately determining the location and number of recombination breakpoints a challenging problem. This simulation study shows that too many differing histories concatenated into a single locus will cause serious problems for a standard ML method like Fast Tree 2. However, trying to uphold the assumption of no intra locus recombination is also not currently feasible. The problem each recombination event presents falls on a spectrum. If the evolutionary histories combined during a recombination event are the same, the evolutionary history of the whole sequence is still the same; detecting this event and modeling it has no impact. If a recombination event combines two histories which are almost identical but have slightly differing branch lengths, how important is it to identify and model separately?

It is possible the ARG like methods are in fact identifying recombination spots where the history of the trees significantly changes. This would explain why the Tsinfer seems robust to recombination although the number of estimated breakpoints is very different from the simulated number of breakpoints (Figure 6). More work needs to be done to verify if this is the case.

References

[1] Posada, D., & Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *Journal of molecular evolution*, *54*(3), 396-402.

[1] Lemmon, A. R., & Lemmon, E. M. (2008). A likelihood framework for estimating phylogeographic history on a continuous landscape. *Systematic biology*, *57*(4), 544-561.

[2] Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC bioinformatics*, *19*(6), 15-30.

[3] Roch, S., & Warnow, T. (2015). On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Systematic Biology*, *64*(4), 663-676.

[4] Huang, X., Wang, S., Jin, L., & He, Y. (2021). Dissecting dynamics and differences of selective pressures in the evolution of human pigmentation. *Biology open*, *10*(2).

[5] Jerome Kelleher, Alison M Etheridge and Gilean McVean (2016), Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes, PLOS Comput Biol 12(5): e1004842. doi: 10.1371/journal.pcbi.1004842

[6] Rambaut, A., & Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, *13*(3), 235-238.

[7] Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, *5*(3), e9490.

[8] Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature genetics*, *51*(9), 1321-1329.

[9] Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., & McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature genetics*, *51*(9), 1330-1338.

[10] Mirzaei, S., & Wu, Y. (2017). RENT+: an improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics*, *33*(7), 1021-1030.