Dept. of Anthropology
260 Central Campus Dr., Rm 4428
Univ. of Utah
Salt Lake City, UT 84112

May 3, 2021

Prof. Matteo Fumagalli
PCI Mathematical and Computational Biology

Dear Prof. Fumagalli:

Thank you for handling this manuscript, and thanks to all of the reviewers. The reviews were thorough and helpful, and I think the manuscript is much improved.

The reviews asked for clarifications, fuller descriptions, and for more background on the Legofit package, all of which I have provided. In addition, one reviewer asked for an analysis of real data. In response, I have used the new algorithm to replicate the analysis from our 2020 paper in *Science Advances*. This strengthed some of the results from that paper and replicated the rest.

In the detailed responses below, reviewers' comments are in *italics*, and my responses are in roman type. I begin with your own detailed comments:

1. *clarifications are sought on the simulations performed.*
   The section on simulations has been greatly expanded.
2. *Additionally, either a small application on real data or explicit comparison of run times between different versions of legofit will improve the appealing of this study.*
   Figure 5 compares the run times of legofit's stochastic and deterministic algorithms. The stochastic algorithm was the only algorithm available in previous versions of legofit, so this amounts to a comparison of the new version to the old.
   In addition, the revised manuscript now includes an analysis of real data, as discussed above.
3. *Please also fix any typos and inconsistencies (e.g. numbering of references).*
   See my response to item 25 below.
4. *Additionally, I would encourage you to provide more details on the context of legofit for readers who are not familiar with the original paper.*
   I have expanded the introduction to provide more background about Legofit.
5. *provide a more comprehensive abstract.*
   Done.

Thanks again for your help with this, and let me know if you need anything further from me.

Responses to reviewers are in the postscript.

Yours

Alan R. Rogers

P.S. Detailed responses. (Reviewers' comments in *italic type*, my responses in roman.)

Reviewer 1

6. *It is stated that the simulation uses the gene genealogy of Fig. 1. Does this imply that individuals sampled from population Y can exclusively be migrants from populations N or D?*

   No, there are three potential sources for a nucleotide in Y: it may come from N, from D, or from the common ancestor (XY) of X and Y. The text is now explicit about this.

7. *What is the assumed sample size (i.e. n) in the simulations?*

   The simulations assume that a single nucleotide is sampled from each of 4 populations: X, Y, N, and D.

   Within legofit (used for estimation, not simulation) $n = 1$ in each of the sampled segments (x, y, n, and d in Fig. 2) but may be larger in ancestral segments. It may be as large as 4 in the root segment, xynd. In these ancestral segments, the program iterates across all possible values of n, weighting by the probability of each value.

8. *It is mentioned that there are ten times as many points as free parameters (p. 7), implying a sample size of around ten based on a statement in ref. 11. Why was this sample size chosen?*

   This remark does not refer to the sample size; it refers to the size of the swarm of points used by the "differential evolution" algorithm. (This is the algorithm responsible for minimizing KL divergence.) The authors of that algorithm recommend using $10\times$ as many points as there are free parameters, and I have followed their recommendation. The text is now clear on this point.

9. *In Fig. 2, it is stated that the number of estimated parameters is 11, whereas, according to the supplemental file, the model has a total of 16 parameters that concern demography (i.e. population sizes, split times, migration rates). Are the five remaining parameters assumed to be known? Could they be estimated as well (e.g. time of Neanderthal admixture)? Or are they fixed to avoid unidentifiability of other parameters like e.g. the migration rates? Does the need to fix these important parameters (including also Nxy and Txynd) limit the applicability of the approach to relevant real-world scenarios?*

   Section 2.6 now details which parameters are free, which are fixed, and why. There are 5 fixed parameters: T_{XYND} (separation time of the modern and archaic lineages) is fixed in order to calibrate the molecular clock. Its value is taken from the literature.

   Two other fixed parameters, T_α and T_ϵ, are the times of the two migration events. These parameters cannot be estimated, because their values have no effect at all on the data—site pattern frequencies—used by legofit. The length of the genealogical branch connecting a nucleotide from Y to (say) the common ancestor of N and D is exactly the same, irrespective of whether the migration from $D \to Y$ was early or late. And the length of this branch is all that matters.

   The final 2 fixed parameters are the sizes, $2N_X$ and $2N_Y$, of the African and European

populations. Their values cannot be estimated either. Under this model of history, it is not possible for either of these populations to have more than a single lineage, so no coalescent events are possible, and population size doesn't matter. This illustrates the insensitivity of legofit to recent changes in population size, which reduces parameter count and allows legofit to focus on deep population history.

10. *How does the accuracy of the estimation of population sizes and split times depend on the migration rates (i.e. how robust are they under variation of migration rates)?*

    To answer this question, look at the pairwise scatter plots in Fig. 3. Some pairs (such as T_{XY} and $2N_{XY}$) are tightly correlated, so sampling error in one parameter estimate has a large effect on the other. This results in large uncertainties in the estimates of these parameters, as seen in Fig. 6. The two migration rates (m_α and m_ϵ) have modest correlations with each other and with two population sizes ($2N_N$ and $2N_D$) but are essentially uncorrelated with separation times. This undoubtedly causes some of the bias that m_α, m_ϵ, $2N_N$, and $2N_D$ exhibit in Fig. 6. These problems, however, seem modest in the current analysis, and they are discussed in the text.

11. *The described procedure to compute partitions among ancestors is carried out within segments of constant population size $2N$. However, in the case of migration as depicted in Fig. 1, when looking backward in time a lineage leaves population Y and appears in population N. It has therefore evolved in a segment with population size $2N_Y$ until the time of admixture and "after" that it evolves in a segment of size $2N_N$. It is not clear to me how this dynamics is accounted for in the algorithm to compute the branch lengths of site patterns.*

    Figure 2 of the revised manuscript shows how the network of segments relates to the population history in Fig. 1. I hope that the text discussing this figure will clear up this point of confusion.

    In case it doesn't, let me trace a bit of the history of a nucleotide sampled from population Y, using the segments defined in Fig. 2. We begin in segment y. Because that segment has only 1 lineage, no coalescent events can occur. We therefore add the length of this segment to the branch length associated with site pattern y. At the ancient end of segment y, the lineage derives either from segment d0 or from y1. Let us assume that it came from y1, and before that from n2. In n2 there are two lineages—in addition to the lineage we are following, there is also the lineage sampled in segment d. These two lineages may coalesce within segment d2, or they may remain distinct. The probabilities of these outcomes are given by the theory in sec. 2.3 and depend on the population size within d2. The expected branch lengths within d2 are given by the same theory and also depend on population size within d2.

12. *On p. 9, it is discussed how the observed correlations of estimated parameters (leading to non-identifiability) can be attempted to be ameliorated using PCA. However, it is mentioned that PCA introduces further bias and that therefore it is omitted here. I do not understand how, then, the correlations are corrected for instead. Also, what are stages 4 and 5 of the algorithm doing, as apparently there is still some kind of dimensionality reduction carried out in them.*

    Stages 1 and 2 operate on the original variables. In stages 3 and 4, the original variables are re-expressed in terms of PCs. There is no dimensional reduction in the present analysis, but stages 3–4 improve the fit nevertheless, because the optimization algorithm works better when the variables are uncorrelated. I do not claim that this procedure is a full solution to identifiability problems—these problems still account for much of the spread in parameter estimates seen in Fig. 6. It does, however, reduce that spread to some extent. I have clarified the text on this point.

13. *In Fig. 3, since absolute differences in pattern frequencies are plotted, it is hard to assess the accuracy of the algorithm (besides the direct comparison between the two modes, stochastic vs. deterministic). I would recommend plotting relative errors instead.*

    This comment refers to what is now Fig. 4. This is a good point—the site patterns with substantial spread in this residual plot are also those with large means, so if I plotted relative error, the spreads would be more similar. Nonetheless, I've decided against this suggestion for two reasons. First, legofit is minimizing KL divergence, and the components of KL divergence are approximately equal to the residuals as I have plotted them. The existing graph therefore illustrates the contributions to KL divergence. Second, residual plots are standard in statistics, and it is conventional to plot them in this way.

14. *Why is* $2N$ *repeatedly named the "haploid population size" (instead of* N*)? Conversely, in the supplement* N *is confusingly called the "di[p]loid population size".*

    I now define "haploid population size," the first time I use the term. I couldn't find "diploid" or "diloid" in current version the supplement. I must have fixed this problem in some previous edit.

15. *What is the following parameter, mentioned in the supplemental file: "c = 1e-8 # rate per base pair per generation"?*

    It's the recombination rate. It and all other parameters are now defined in the main text.

16. *Note that migration rates are denoted differently in Figs. 1 and 2.*

    Fixed.

Reviewer 2 (Fernando Racimo)

17. *While individual subsections of the method section were clearly explained, I feel the connections between the different sections were a bit hard to follow. For example, the transition from section 2.2 to 2.3 was very disjointed: it was unclear what role the matrix coalescent was playing in the calculation of the expected branch length, which was brought up right above it.*

    I have added a transitional paragraph to the end of the section on the matrix coalescent.

18. *I would recommend adding a schematic that describes how the different steps of the algorithm relate to each other, and where the key improvements are, perhaps using an example drawing of a segment with a certain number of descendant and ancestral lineages, embedded in a larger population graph.*

    I have added a new figure (Fig. 2) which shows the network of segments.

19. *Another useful schematic could be used to describe the different steps in the data analysis pipeline.*

    I was unable to see how to clarify this with a figure, so I have tried to do it in prose.

20. *What is the size (in physical distance) of each simulation? What are their mutation and recombination rates? Are these supposed to represent 50 uncorrelated windows from the same genome, which are used to infer the same demographic history? Does this imply that a real-world application would involve partitioning the genome a priori somehow? It was unclear why the author was using 50 simulations, it would be be good to explain the reason for that, especially for readers who are not familiar with the original legofit program.*

    I have greatly expanded the section (2.5) on simulations in order to clarify these details. There are 50 independent data sets, each consisting of 4 genomes. Each data set is comparable to a real data set consisting of multiple genomes. Each data set is used to estimate parameters, and the variation among these 50 estimates measures uncertainty.

I got into the habit of using 50 because the stochastic algorithm was so slow. I could easily increase that number now and will probably do so in the next publication that uses only the deterministic algorithm.

21. *It would be nice to see a small application to a real-world data problem, to motivate the use of the algorithm, and a comparison of run times between the old and new versions under this scenario, for an end user to have a realistic assessment of the speed improvement. This could involve, for example, inference in a dataset involving a small number of admixture and population split times on a real human genome, like the super-archaic scenario from Prufer et al. (2014), used by the author in Rogers (2019).*

    This has been done and is discussed above under item 2.

22. *Related to the above comments, are the results on the super-archaic scenario in any way affected by the improvements (in terms of greater parameter accuracy) of the new algorithm?*

    The new algorithm replicates all the results of our 2020 paper and strengthens two of them. This is now discussed in the manuscript.

23. *It would be helpful to add parameter names to the schematic in Figure 1: TA, TXYND, NXY, etc. to guide the reader in subsequent figures.*

    I used to do that when my models were simpler. But with the present model, the figure becomes extremely cluttered when you add all the parameters. I think the revised text is much clearer, because the parameters are all described fully in section 2.5, on simulations.

24. *It was unclear why the author was focusing on both the sum and the difference in the 2 migration parameters. Are these sums and differences calculated post-hoc after inference of the individual migration parameters? If so, what is the information provided by the accuracy plots of different linear combinations of them?*

    This refers to the paragraph just before "Conclusions," which discusses the biases that can be seen in Fig. 6. I have expanded this paragraph, and I hope it is now clear. The sums and differences are calculated post-hoc, as the reviewer suspected.

25. *Numerical references are out of order (e.g. first reference: 11-13).*

    In the list of references, the references are listed and numbered in alphabetical order. The first reference cited is number 11 because it is not at the beginning of the alphabet. This is a widely-used convention—it is the "plain" style of BibTeX—which I prefer to the alternative of numbering references in order of their appearance in the manuscript.