Phylogenomic data reveal how a climatic inversion and glacial refugia shape patterns of diversity in an African rain forest tree species Andrew J. Helmstetter, Biowa E. N. Amoussou, Kevin Bethune, Narcisse G. Kandem, Romain Glèlè Kakaï, Bonaventure Sonké, Thomas L. P. Couvreur
10.1101/807727 version 1

Dear Andrew and coauthors,

Reviewers have responded very positively to your ms. and have made a number of insightful and constructive comments that I am sure you will be able to make good use of. The reviewers' comments are included (presumably) below (R1 & R2), in a separate pdf (R3) plus in an annotated copy of the pdf to which I have added further points here and there.

We thank the editor for their helpful comments

The main points raised:

Hypotheses and tests: It always aids the clarity of this kind of analysis to set out in the introduction all the hypotheses, as well as the results with which they could be rejected. As noted by R2 and R3, those corresponding to flowering times and niche differences are currently neglected. R2 suggests ways in which these might be addressed using the current datasets, and also moots the possibility of formal biogeographic model testing using BioGeoBEARS. These would certainly add considerable value to the paper.

We have now added sections and analyses that address flowering time and ecological niche differences (ln147, 298; Fig. S16; Fig. 4C-F). Ancestral range estimation like BioGeoBEARS is specifically designed for the interspecific level and can perform particularly badly when the tree is very small (see https://onlinelibrary.wiley.com/doi/full/10.1111/jbi.13173) so we do not think it is appropriate for our intraspecific level study. That is why we used methods more appropriate for phylogeography studies. These approaches are used for example when inferring virus dispersion.

Methods and assumptions: I agree with R1 on the use of methods making unrealistic assumptions about gene flow in an analysis within a species using multiple independent markers. A concatenated analysis seems like a bad idea to me in principle, and although I can't compare the ASTRAL tree to the RAxML one (because the tips aren't labelled – I would ask for supplementary tree files/fully labelled trees to represent the information presented in such figures) the network structure in the splitstrees result and the short branch lengths in parts of the tree do nothing to assuage my concern that the single ML tree cannot realistically represent phylogeny here. Both topology and branch lengths may be impacted by the model violation, and the strong support could just be a misleading symptom of that. R1 suggests to replace this with analysis based on multispecies coalescent. Similarly R1 suggests replacing the "mugration" approach with those implementing a structured coalescent.

We have added tip labels to our trees in Fig. S6

We have swapped the positions of ASTRAL and RAxML trees. We also performed a new SNAPP analysis, which makes use of the multispecies coalescent (Fig. S7).

Our spatial diffusion approach is not the “mugration” approach the reviewer is referring to. As we are interested in how lineages moved in a continuous geographic, rather than discrete, manner these approaches are not particularly useful to answer our questions on how *A. affinis* dispersed through Lower Guinea. The structure coalescent focuses on ancestral population sizes and migration rates between populations while we are interested in movement of lineages (before establishment of population) and migration within populations.

Dataset and processing of SNPs R1 asks for a comparison of the datasets resulting from phylogenomic/population-level processing. I agree this would be enlightening: In addition to these comments, I would like to know how within-individual polymorphic sites are treated for the former (I see no sign of phasing; a general weakness of some pipelines in my view). How might these different ways of treating the same data potentially impact the results

We have added a table (Table S2) comparing our datasets. Given the similarities between SNAPP, ASTRAL and raxml and clustering vs phylogenies it appears these different pipelines have little effect on our results.

I would ask that in revision your ms. you copy all these comments into a separate response document and address each individually; ideally I would like to see changes to the ms. in the form of tracking in a word document. Just makes my life easier.

Finally, congratulations on a fine piece of work. I am looking forward to seeing a revised version.

All the best, Mike Pirie

Download recommender's annotations (PDF)

Reviews

Reviewed by Lars Chatrou, 2020-01-13 06:15

Helmstetter et al. – Phylogenomic data reveal how a climatic inversion and glacial refugia shape patterns of diversity in an African rain forest tree species
This paper present exciting data on the phylogeography of an African species of rain forest trees, Annickia affinis. The sampling (112 individuals) shows a long-term dedication to collecting this species, and is impressive. The use of recently developed baiting kit, and the application of several analytical tools that produce coherent results have resulted in a paper that will draw much attention, and will be a forerunner of similar studies in the future.

We thank the reviewer for their considered and helpful comments.

- It would be good to add references to the second and third hypotheses, which according to the authors have been suggested. I'd be especially interested in seeing a reference for the second hypotheses, because here I feel the metaphor of a hinge may be taken to far by suggesting that flowering times flipped along the North-South axis around 0-3 degrees N. Is there any evidence in other papers than Hardy et al. (2013) to suggest this pattern?

To our knowledge there is no published work that could be added to the second and third hypotheses, so we cannot add references here. We have seen some unpublished work on flowering times across the north south boundary in the species of the genus *Barteria* by R. Blatrix (CNRS) et al. but this was not published.

- "If glacial refugia have played an important role in CAR plant dynamics we would expect to find evidence of dispersal inland because most putative CAR refugia are located in the Atlantic Guineo-Congolian region". This seems a very strongly phrased hypothesis to me, given the uncertainty surrounding the location and importance of Pleistocene refugia. The authors have indicated this uncertainty themselves (lines 59-61). Also, several papers, e.g. Piñeiro et al. (2017) have only demonstrated partial overlap at best between Maley's refugia and contemporary genetic clusters. Furthermore, one of the refugia Maley suggested is located in the Congo Basin, to the east, which overlaps with the eastern part of the distribution of Annickia affinis. I appreciate the beauty of clearly phrased and unambiguous hypotheses, but in this case I wonder if the clarity of the hypothesis is not disguising suggestions in the literature that would suggest a differently phrased hypothesis.

Although the nature of Anhuf et al's and Maley's refugia is different, both agree that most were close to the Atlantic coast, which is why we are able to phrase our hypothesis around dispersal inland. We started the sentence with "if" to try to highlight our uncertainty about the situation but as it was unclear we have attempted to rephrase this sentence to make the uncertainty behind and what we are referring to clearer (ln103).

- The generation time of 15 years is likely to be a serious underestimation, and it would be interesting to see what the effect on the results would be if a longer generation time had been used in the analyses. The generation time is based on a paper on Annona crassiflora, a savanna species from the Neotropics. Looking at mortality rates of Annonaceae species in Baker et al. (2014; Ecology Letters 17: 527–536), and assuming that generation times can be approximated by the mortality rate -1 , the generation times of Neotropical tree species of Annonaceae vary roughly between 40 and 100 years. These species are better comparable to Annickia affinis in terms of habit and habitat, and would therefore probably reflect the generation time of the latter species more accurately. I appreciate that the authors are cautious and avoid interpreting the timing of demographic events. Having said that, the patterns that are disclosed in this paper did happen in real time, and a temporal framework is pivotal for linking up this study with other work. Moreover, Fig. 3 has an axis indicating absolute time, and species distribution modelling was done using LGM climatic data – both cases are explicit about absolute time. This will be picked up by the readers regardless of the authors evading to draw strong conclusions on time. So, I think it would be good if the authors could provide more insight into the effect of the short generation time of 15 years on the results of their analyses.

We have added further results modelling demography using a generation time of 50 years (Fig. S15) and discussed this in the text (ln399). This resulted in a minor change in the timing of events, with a slight increase in dates that was still fairly close to the LGM. However we did find this caused a fairly significant reduction in effective population size.

Reviewed by Oscar Vargas, 2019-12-16

Review for PCH EVOL BIOL of the manuscript titled: “Phylogenomic data reveal how a climatic inversion and glacial refugia shape patterns of diversity in an African rain forest tree species“

The manuscript mentioned above present a phylogenomic study using targeted sequencing. Authors study one plant species distributed in the tropical rainforest of Africa trying to elucidate if its populations have genetic structure and the tentative reason for such. Authors found a sticking pattern of structure dividing northern and southern populations in accordance with previous studies; they conclude that there is some evidence supporting Pleistocene changes in forest coverage as the cause for the demographic history of the species’ populations.

This manuscript presents a pioneering effort to study historical demography in the tropical rainforest of Africa. I praise authors efforts along with their selection of methods to analyze the data. Writing is grammatically correct and clear. I believe this study is worth of being published after some adjustment to the writing, the framing of the study, and perhaps some additional analyses. With these editions/additions I believe this study will be a beautiful and exciting contribution to the field Main concerns:

We thank the reviewer for their inspiring and helpful comments.

1. Hypothesis testing. In the introduction, authors clearly stated that there are three hypotheses to explain genetic structure. Yet, they focused mainly in the Pleistocene hypothesis. A clear example is how in the introduction they stated what are the expectations under the Pleistocene hypothesis, without stating potential ways to test the other two hypotheses. Similarly, in the discussion, authors seem to solely focus on the Pleistocene hypothesis. Authors, I believe, do have the data to test the other hypotheses. For flowering times, they can simply look at herbarium records looking for differences in flowering times between populations. For the third hypothesis, using their climate data, they can test whether there are differences among the niches of the different populations–if climatic niches are different, then there is an indication for habitat filtering.

We have added two new sets of analyses for testing these hypotheses, though we stress that further work is needed to explore them sufficiently (particularly in the case of flowering time).

First we assembled fruiting and flowering time data currently available for herbarium specimens of *A. affinis*. This was extracted from the BRAHMS database of the Naturalis herbarium in which most *Annickia* specimens were entered during the revision of the genus (Vertseegh & Sosef 2007). There were relatively few records of flowering individuals (15) and results did not show any divergent patterns among individuals north and south of the climatic inversion. However this does not mean that we have disproved flowering time as we lack the appropriate sampling/power. Nevertheless we have added a figure (Fig S16) and a section to the text to highlight what we did (ln413).

To further assess whether differences in niches among the different populations exist, we performed a set of analyses using ENMTools (Warren et al. 2010). Given that we were testing for differences on either side of the climatic inversion, we grouped individuals into two groups by their location – north or south of the equator. We then built climate models using GLM and measured overlap among them (ln298; Fig 4C-F). We also used a new phylogenetic tree built using the coalescent approach SNAPP to assess divergence in niche among populations, indicating that the two large populations north and south of the inversion were most dissimilar (ln420; Fig S7).

2. Biogeography Authors use their mapping of the specific location on the phylogeny S. fig 7, specifically the location of the sister taxa to the rest, as a historical biogeographic reconstruction and draw conclusions based on this, e.g. lines 302–313. Simply looking at the sample that is sister to the rest is not enough to draw conclusions about historical biogeography and dispersal. I suggest authors to make a bioregionalization of the area and perform a formal historical biogeographical analysis on the whole phylogeny, BIOGEOBEARS is one option.

Bioregionalization is difficult over this relatively small area as it is hard to distinguish discrete regions. BioGeoBEARS is an approach designed to estimation the ancestral ranges at species rather than within species population levels. In addition, BioGeoBEARS can perform particularly badly when the tree is very small (see https://onlinelibrary.wiley.com/doi/full/10.1111/jbi.13173) as it would be in our study.

We also examine patterns within populations (previously ln442-444). We did attempt a structured coalescent approach, as was suggested by another reviewer and useful for population level analyses, though we were not able to make it converge with our dataset (see Fig. 1 below).

Our diffusion analyses (Fig. 2) should be used as the primary source of information on dispersal in the species. We have rephrased these lines (previously 302-313) to avoid drawing conclusions on historical biogeography from this figure (ln367).

Minor comments in pdf It was pleasure and honor to review this paper

We have addressed the comments in the PDF, thank you.

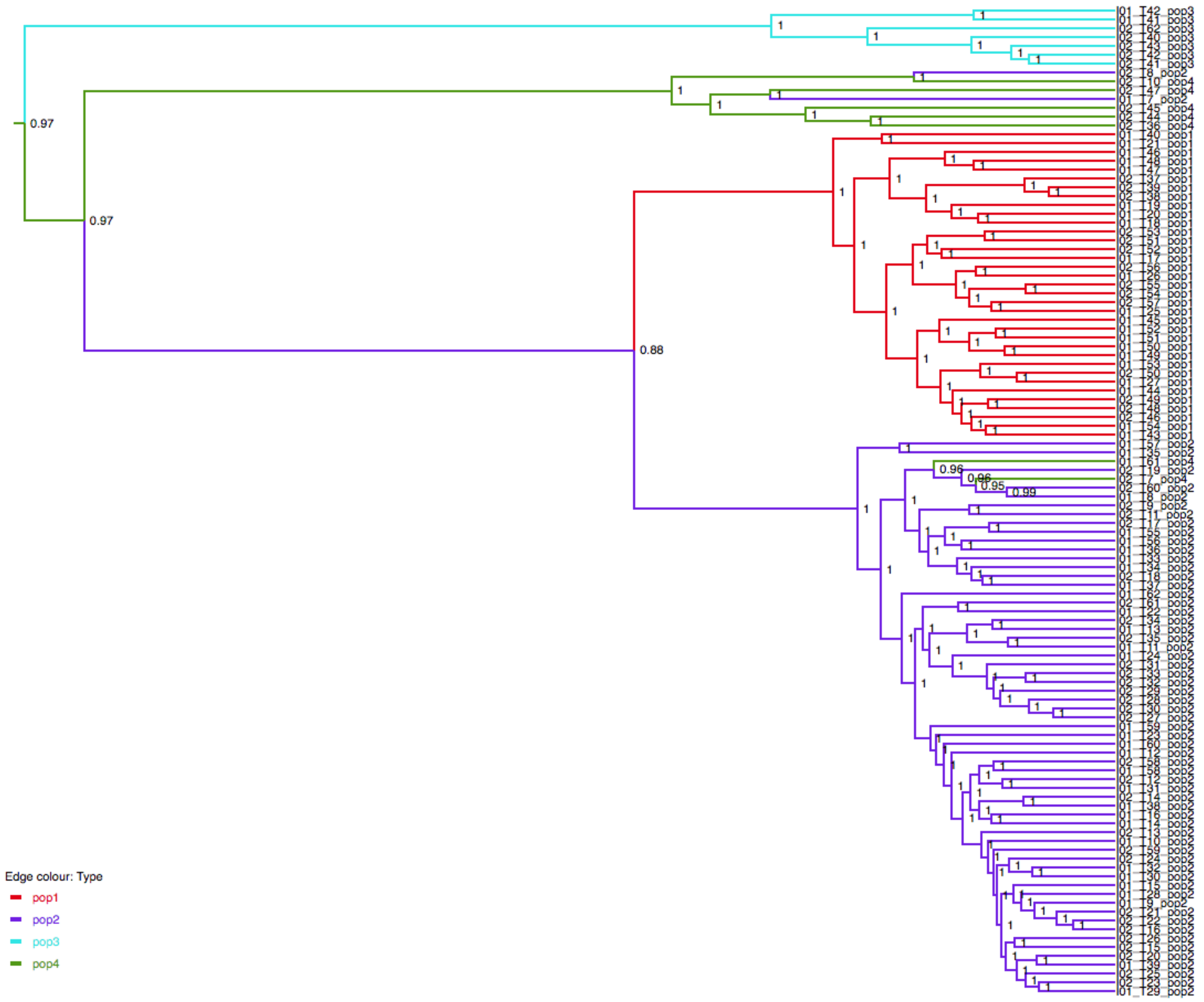

Figure 1 Example of multi type tree structured coalescent analysis. Probability of population state is shown at nodes. This analysis did not converge after 100m generations in >5 attempts.

Reviewed by Miguel Navascués, 2019-12-16 13:31

Helmstetter and collaborators present a study of the genetic diversity of Annickia affinis, an African rain forest tree. They study the geographic structure of its genetic diversity and they infer its demographic history. The results are discussed in relation to the climatic inversion in Central Africa, the glacial refugia and the inferred potential distribution in the past via climatic niche modelling. This study adds to a body of work on the phylogeography of Central African rain forest plants that try to shed light on the biogeographical processes in the region. Cumulative evidence from different species is very valuable to understand these processes and the present work will be a good contribution. An additional merit over previous works is the use of a larger set of molecular markers thanks to the use of high throughput sequencing technologies. However, I would not go as far as saying that this work is an exemplary study (i.e. "proof-of-concept for future work") because the analytical methods are not particularly novel and some of them are flawed. Some of these analyses need to be revised before this work can be recommended.

We thank the reviewer for their insightful and helpful comments.

1) My first concern is with the analysis of spatial diffusion based on using the evolution of a trait along the genealogies as an approximation for migration (an approach sometimes called "mugration", i.e. "mutation as migration"). In such analysis, branch length and topology of genealogies are modelled by a panmictic coalescent model, which makes little biological sense in an analysis targeting structured populations. The justifications for the use of such an artificial, yet mechanistic, model are an easier implementation and a lower computational cost. That could be reasonable if the results were meaningful regarding the true migration dynamics. However, an evaluation of the "mugration" approach by De Maio et al. (2015, doi:10.1371/journal.pgen. et al. (2015) results, I can only recommend to remove completely this analysis from the manuscript. As an alternative, authors might explore alternative phylogeographic analysis based on the structured coalescent, for which recent methodological advances have been done by different research groups (e.g. Müller et al. 2017, doi:10.1093/molbev/msx186; Flouris et al. 2019, doi:10.1093/molbev/msz296).

Spatial diffusion, as we used in this study and the "mugration" and structured coalescent methods the reviewer refers to here are two different approaches (doi:10.1016/j.tree.2010.08.010). As we are interested in how lineages moved in a continuous geographic, rather than discrete, manner these approaches are not particularly useful to answer our questions on how *A. affinis* moved through Lower Guinea. The structure coalescent focuses on ancestral population sizes and migration rates between populations while we are interested in movement of lineages (before establishment of population) and migration within populations (spatial diffusion). Bioregionalization is difficult over this relatively small area as there are no obvious regions to distinguish, and even more difficult when attempting to examine patterns within populations.

Nevertheless we were inspired by the reviewers comment and attempted to use the structured coalescent, but this resulted in convergence failure and was not more informative than our diffusion approach. It told us little about migration inland from the coast, and dispersal within genetic clusters. We have attached a figure of our analysis below (Fig. 1).

2) Another issue in the analyses is the use of phylogenetic methods on concatenated sequences for intra-specific data. Concatenation is widely used in phylogenetics sensu stricto (i.e. inference of species trees). In some cases, it can be a good strategy to deal gene tree heterogeneity and large (genomic) data sets. An alternative way to address gene tree heterogeneity is the use of multispecies coalescent methods (equivalent to the structured coalescent mentioned above) which has the advantage to explicitly acknowledge the biological reality of recombination among loci. Multispecies coalescent methods have also been shown to be more robust to the presence of gene flow, taxon sampling, long branch attraction and anomalous gene trees. A recent review by Liu et al. (2015, doi:10.1111/nyas.12747) suggests that the more biologically relevant multispecies coalescent should be preferred to concatenation, which can be biased and have overinflated bootstrap values. I do not have a position on the debate on whether concatenated and coalescent approaches are more appropriate for phylogenetics, because it is not my field of research. However, for population genetic analysis, I find the use of concatenated approach unjustified. The problems that coalescent approaches addresses in phylogenetics come from the analysis of species that have dynamics closer to populations: incomplete lineage sorting, anomalous gene trees, gene flow, low divergence. Population structure analysis such as those implemented in DAPC or fastSTRUCTURE allow to uncover how genetic diversity is distributed in clusters, without imposing a hierarchical structure. The use of phylogenetic approaches forces a hierarchical structure (tree) for the data. This tree structure might be relevant if it is related to the possible population divergence processes within a species. The statistical model used to reveal that hierarchical structure is crucial to obtain relevant results and concatenation seems to force a rather unrealistic model (same gene genealogy for all loci among individuals of the same species). To me, the tree presented in figure 1D is more likely showing a mixture of true biological features (already reveled by, for instance, DAPC) and artefactual structure, supported by some misleading bootstrap values. I think figure S6 shows a more relevant result which reveals, for instance, the low confidence between the “phylogenetic” relationships between clusters EG, GC and (WG+CA). To sum up, I think that the analysis of concatenated sequences does not add any further insight to this data and can potentially be misleading.

The debate rages on about concatenation vs coalescence but we agree with the reviewer that coalescent approaches are likely to be more informative. We have swapped figure S6 and 1D because of this.

Additionally, we have conducted an analysis using the SNAPP approach and a subset of our SNP dataset (Fig. S7; ln228,360). This approach makes use of the coalescent and produces results at the population level. This tree had a topology of genetic cluster divergence that matched our other results.

In addition to these two main points I have some minor suggestions for the authors, concerning mainly the presentation of their work:

3) Line 92: Substitute “phylogenomic data” for “genomic data”

This has been changed.

4) Materials and methods: Data for “phylogenetic” and population genetic methods have followed a slightly different bioinformatic process for selecting the loci/polymorphic sites to be analyzed. I think it would be useful to describe how different are this two subsets of data (from the same raw data). How many loci and polymorphic sites are presented in each subsets? How much do they overlap?

We have included a table in the supplement with various statistics relating to the reviewers question (Table S2) and mentioned level of overlap (ln323) in the results section.

5) Line 201: A description of the cross-validation procedure for the DAPC analysis is missing. The current revision of the text does not allow the reader to understand how this procedure was performed nor how they should interpret the results presented in figure S1. In addition, this figure needs also a better description: what are the solid and dashed lines? What are the black squares? What is the meaning of the blue shadows? I do not see any maximum over the value of 40 PCs; it looks like the same results were obtained for any number of PCs.

We have explained the cross-validation approach more thoroughly in the methods (ln190) and removed the corresponding figure as it is not very informative and difficult to interpret.

6) Lines 381-390. I am not sure of the relevance of discussing the presence of potentially admixed individuals as “hybrids”. Is there any evidence that points towards an incipient speciation among clusters of this species? Is there evidence for local adaptation? The presence of few admixed individuals can be attributed to low gene flow or recent secondary contact, I do not see the need to invoke selection (nor to reject selection). Also “The existence of hybrids in the absence of gene flow...” seems to be a contradiction, do you mean “absence of historical gene flow” or “absence of introgression”? I am not sure you have evidence of any of this two alternatives, though.

Blatrix et al. (2017) described the relevant individuals as "hybrids" between the Northern and Southern groups as they used the newhybrids approach to assign individuals in categories (Parents, F1s etc.). We have removed the term hybrid and the sentence reading "The existence of hybrids in the absence of gene flow..."

7) Label x-axis in figures 1B and S5 in some way that the results can be compared, i.e. individuals (or groups of individuals) need to be identifiable.

We have added labels to clustering barplots.

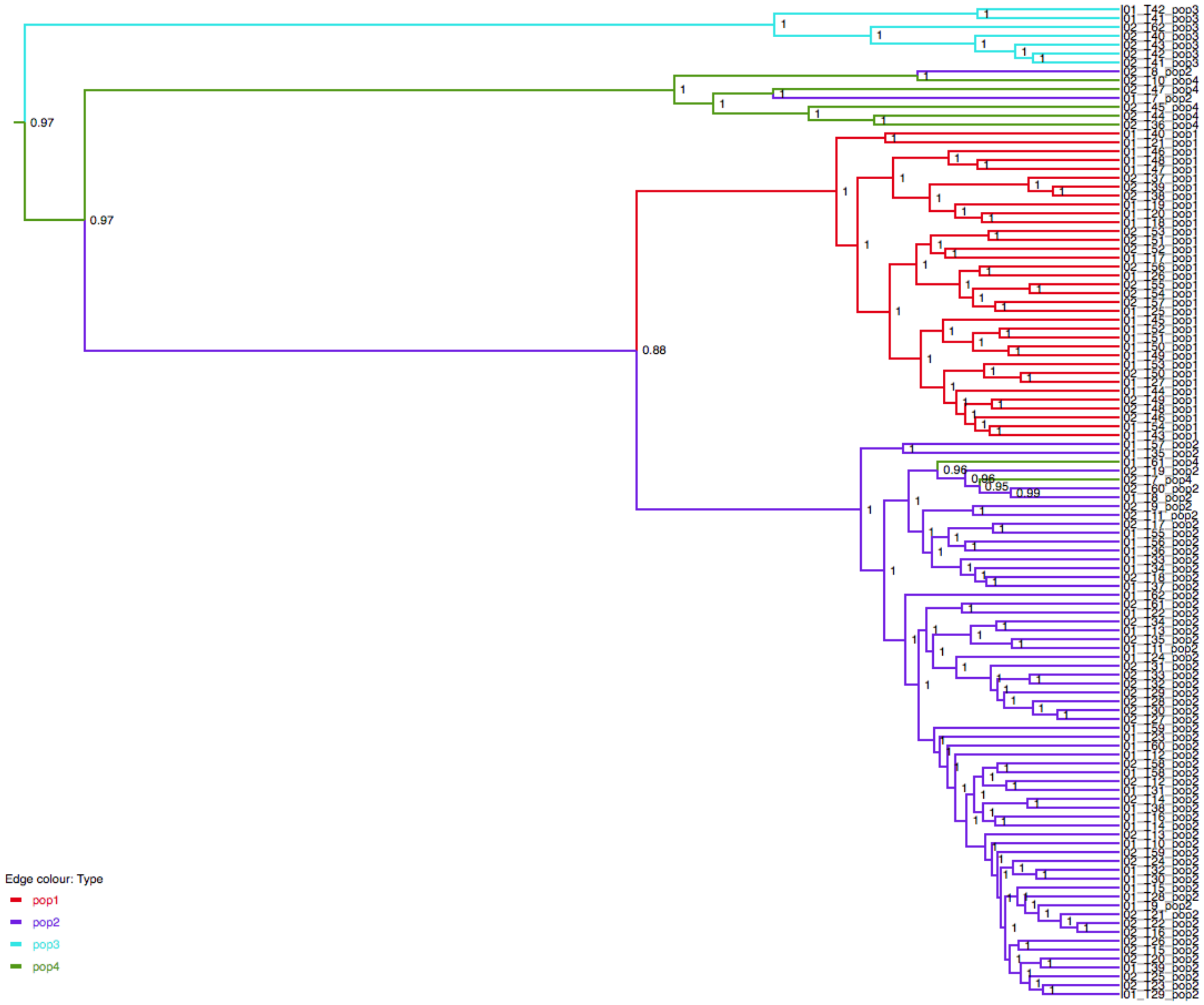

Figure 1 Example of multi type tree structured coalescent analysis. Probability of population state is shown at nodes. This analysis did not converge after 100m generations in >5 attempts.