My request (April 23rd 2018)

send to the two corresponding authors of the paper

Dear colleagues,

I'd be very happy if you could provide me with the data matrix, you used to infer the tree shown in fig. 1, to evaluate some of your claims, especially regarding the (so far unfounded) statement that there can no ingroup-outgroup branching artefacts, because you excluded the extremely long-branched sistergroup of Su et al. (2015, no plastid data available) and keeping only the distinctly long-branched further sister clades. And how incongruent your data are compared to the data set I put together from gene banks (but not Su et al. 2015, as you've written). It says on the journal homepage under *Research Data*: "Data not available / Data will be made available on request". I hope the latter applies, if not, please let me know otherwise.

I think it would make a nice add-on post to the one we had December last year on the Genealogical World of Phylogenetic Networks.

Using consensus networks to understand poor roots http://phylonetworks.blogspot.fr/2017/12/using-consensus-networks-to-understand.html

But I fully understand that you are not interested in looking into signal issues for yourself. You seem to be content with reproducing and discussing in detail our 2017 findings. It's good to see we were not as wrong as originally pointed out to us by another of your co-authors.

An unrelated question (just out of curiosity), given that the senior author is from South America, why are the Psittacanthinae still undersampled? Is is because of the restrictive legislation regarding probing material from the wild as one has in Australia, and particular, New Zealand? If you want to shoot down the alternative *Tupeia*-root, you simply have to sequence the other species allegedly having an A-type pollen, and figuring the pollen from the sequence sample/population.

With best regards, Guido

--
Fight the Fog: Make the peer-review transparent
Sign up on change.org

Guido Grimm
Orléans, France
www.palaeogrimm.org
Res.I.P. blog
Twittering (sort of)

Two days later (April 25th), the answer of one of the corresponding authors

Dear Dr. Grimm,

Thanks for your message and comments through ResearchGate.

I have forwarded your message to our coauthors. Hopefully they will response soon. I will also send you the data matrix once the other corresponding author replies:)

Best regards,

Limin

--

Limin Lu
Assistant Professor
State Key Laboratory of Systematic & Evolutionary Botany
Institute of Botany, Chinese Academy of Sciences
20 Nanxincun, Xiangshan, Beijing, 100093, China
Email: liminlu@ibcas.ac.cn
http://www.lseb.cn/lulimin?id=9
https://www.researchgate.net/profile/Limin_Lu

Note: I naturally thanked him for the quick answer and exchanged a few more mails with Limin including some tips for future undertakings and pointing him to Scotese's beautiful palaeotopographic maps, which can now be found on Research Gate, too (some also available as kmz-File for GoogleEarth). A tip to all biogeographers: don't map ancient ancestral areas and migrations just on the modern globe, but on the one showing how the Earth may have looked like at the time you inferred your dispersals and vicariances (and keep in mind: fossil-constrained node dating estimates are always minima). And it's always a good idea to map the known fossil record (if there is any), too.

Another two days later (April 27th), the reply of the Lord of the Real, and my response to it (in blue font)

> Dear Dr Nickrent
>
> thanks for taking the time to respond, even negatively. I'm just in-between travelling, but wanted to answer you directly to clarify some things.
>
> I don't see why you included my first author in the cc. This has nothing to do with him (you are of course free to forward this mail to anyone you like). My request for the new data matrix stems simply out of my personal, un-professional curiosity about often overlooked (accidently or purposely) signal issues in oligogene matrices that would surely benefit from some out-of-the-box-thinking.
>
> I'm not a paid scientist anymore, free to follow whatever catches my interest (recently, I spend a lot of time with dinosaur datasets and their signals: usually not tree-like but used to infer trees).

Dear Dr. Grimm – Limin Lu shared with me an email you sent him earlier this week. I asked him to wait before responding to you about sharing data, the reasons for which I will share below. First off, I want to tell you that the tone of your communications is unsettling, verging on being rude. For example "I fully understand that you are not interested in looking into signal issues for yourself. You seem to be content with reproducing and discussing in detail our 2017 findings".

I didn't mean any disrespect, I was just stating facts. Let me elaborate:

Systematic botanists (including prominent APG figures and yourself, judging from the molecular papers you co-authored that I have read) have a long record of not realising signal issues in their molecular data sets, for understandable reasons: you can't show the weaknesses of the phylogenetic hypothesis. If you do – as I always did during my career – the typically anonymous peers would not rarely try to take this as an opportunity to turn down our papers (increasingly without success) thanks to the widely applied confidential single-blind peer review. The latter a general problem for science. I collected some examples here: https://researchinpeace.blogspot.fr/search/label/%23FightTheFog

For loranths, the situation is indeed unsettling. In the 2008 paper, you apparently dropped several sequences that inflicted conflicting signals. No offense meant, it's common practise (and often required, see above). My policy was always to include everything we have/find, and deal with the unwanted consequences; something one of our anonymous reviewers [i.e. reviewer #1, likely Dan Nickrent himself, but he didn't sign the review reports; but his 2008 co-author, who was much less appalled by our paper and analytical approach, did after in the 2nd round] of the 2017 *Grana* paper was very upset about (kept referring to unsupported branches in the tree rather than looking at our support consensus networks, which were the basis for the pollen mapping, not the tree). I answered the concerns and explained the background in great detail. My answers were ignored by that very reviewer and the editor. Instead we faced unfounded believes regarding the data used in the just published Su et al. paper brought to our attention. The Su et al. paper reveals (see trees provided in your 2015 supplement) manifold signal issues, which remained unexplored and undiscussed in that paper. Missing data is an issue, but for the loranths it goes much deeper than that (I haven't checked the other groups).

Forced to, I had look into Su et al.'s data to counter the remarks of then two reviewers (a third anonymous "expert on phylogeny" was added by the editor, who, after three months provided a review of such general nature a colleague of mine, also an expert on molecular phylogeny, commented as "takes 3 minutes to write") and demonstrated the basic problems with the 2015 data subset for loranths and their sistergroups (File S6 to Grímsson et al. 2017a; see also this post on the *Genealogical World of Phylogenetic Networks*: http://phylonetworks.blogspot.fr/2017/12/using-consensus-networks-to-understand.html

Known problems that were completely ignored in your new paper as far as I can judge from the publication. Instead, you have been claiming numerous things that may or may not be true (can not be verified due to lack of documentation, again, not uncommon in similar papers, but unfortunate from a scientific point of view). What is clear (2018): many branches in the tree still have ambiguous support, not substantially differing from what was shown in Su et al. (2015), which I demonstrated to be biased in several aspects.

So, from what is shown and written in your new paper, it's obvious that you don't see a problem with your data set, and have little interest in exploratory data analysis.

I do. To show how one can (and should) deal with non-trivial data.

My co-bloggers and I at the *Genealogical World of Phylogenetic Networks* are aware that we won't change the course of main-stream science (people like trees because they are so simply to read, even though many realised evolution is not a strictly dichotomous process), but we want give those people examples what to do who are not happy with just inferring a single tree and face similar problems with their data that we discuss in our posts.
Just check out some posts, and you'll see they always provide solutions, ideas to pursue.

I could provide other examples, but this is sufficient. Your tone is not collegial but adversarial, and for that reason, there is no motivation for any of our team to work with you or respond favorably to your requests.

Authors' reluctance to share phylogenetic data is quite common, it's very rare that requests are fulfilled, often one does not even get an answer. So, I'm (honestly) thankful, you took the time to make your point clear. However, like others dedicated to open data (most of our research as scientists is public-funded, directly or indirectly) I believe, once a phylogenetic paper is published, authors should be able to provide the data matrix, so others can verify their results. We recently had an according post on the *Genealogical World of Phylogenetic Networks* on "Why we want to publish our phylogenetic data": http://phylonetworks.blogspot.fr/2018/02/we-want-to-publish-our-phylogenetic.html

On the issue of Loranthaceae phylogeny, you need to understand several things. First, you criticized us for not sampling enough in Psittacantheae.

> Sorry. In your 2018 paper our pollen-alternative root is criticised by a reference to the pollen variation in this clade and a taxon that has never been included in any of your molecular trees, nor has the pollen ever been confirmed. If you criticise *our* publication, *you* need to provide according data. In the 2018 data set, the sampling of this group is not increased, the support patterns have not changed (judging from fig. 1, compared to my in-depth analysis).
>
> Also, I wrote (and meant) in my mail to Limin "unrelated" and mentioned potential problems in getting the material. The question about Psittacantheae was just to have an idea, why that group is not studied, because it appears to be the most prospective one to start with respect to the papers by Feuer & Kuijt, and the, a bit odd (and using undocumented data), Cairns papers. From the little data that has been so far produced it's clear that this is the most interesting and most understudied (molecularly) group of the entire family, which may establish a basis of how loranth evolution and diversification works in a relatively closed system (South America is a pretty stable, long inhabited by loranths continent, with a probably ancient climate/vegetation gradient/pattern; Andes have always been there)

If you believe adding more taxa from Psittacantheae will change the overall picture for the Loranthaceae phylogeny, then you are perfectly welcome to pay for and conduct that sequencing yourself. As an aside, I will be working with a researcher this summer on exactly this group. Second, am now in the analysis phase of several projects that will greatly increase the amount of sequence data we have for Santalales, including complete chloroplast genomes.

> That will be straightforward, but not solve any loranth questions.

This is in collaboration with US and international colleagues. Over the last few weeks I have assembled a 5-gene dataset for all Santalales - 146 of the 163 genera in the order.

> Sounds like fun data. My feeling is that there is no big incongruence but you still may want to infer a separate plastid and nuclear tree, too, and show the phylograms, so readers can depict the amount of change in each clade per genome.
> For the visualisation of the plastid vs. nuclear tree, I can recommend using the "link" option implemented in Dendroscope
> http://ab.inf.uni-tuebingen.de/software/dendroscope/
> If you do want to make an exploratory data analysis of the data, here's a post about how I get to my graphics:
> https://researchinpeace.blogspot.fr/2018/02/how-did-i-do-it-short-guide-to-nice.html
> For such a still small data set (regarding the capacities of RAxML) you can easily run a full analysis as described in the post (the batch/shell-file code-lines for RAxML are provided, too). If you want to compare supports across partitions directly (e.g. using RAxML support mapping option; RAxML can read in MrBayes .t outputs for a ML-BS vs. PP plot) you have to prune the taxon set accessions covering all five genes as the read-in trees have to have the same set of leaves.

Of these, 40 have Illumina HiSeq genome skimming data (but at present only two of these are from Loranthaceae). All nodes in the tree are resolved, except some along the spine of Loranthaceae.

> Exactly. My guess is that it's because of a fast ancient (Eocene or older as we know now, our 2017, *PeerJ*, and your 2018 paper) radiation, the current genes cannot solve the open questions in the loranths. At least not by relying on mere tree inference (or NGS-SNPs from a few samples).

And during this process, I discovered uncorrected sequences from my lab that were submitted to Genbank years ago. These will be corrected shortly.

> This is one reason, why one should document the used matrix. Upload errors happen to everyone, and the more data we have in gene banks, the more difficult it is to keep track of updates.

Therefore, there is no reason for you to be focusing so heavily upon sequences obtained from Genbank – we have more and better quality sequences in hand now and many of these will be submitted within the year when the manuscripts are submitted for publication. You intense scrutiny of the data is all fine and good, but this effort would be best directed at a dataset that does not have missing sequence and missing taxa.

First of all, I already analysed a data set with no missing sequences (our dated tree in the *PeerJ* paper). “Missing taxa” will probably take some time, given the number of species. But that'd be no job for me, but for a paid scientist.

Second, let me again stress something (I already wrote this to Limin [in one of the follow-up mails]). I think the loranths need to be studied further using such a data set. It would be great for e.g. FBD dating. I think one could get great insights, not only for the group, but for plant evolution in general. Because there are so many independent data to compare/connect the molecular results with: the pollen who seem to be lineage-specific (but is not properly studied in the important hemisphere, the southern), the host availability, the pollinators. Maybe it would also allow mapping morphological traits considered to be diagnostic on the molecular trees similar to the approach of Sauquet, Schönenberger et al., *Nature Comm.*, https://www.nature.com/articles/ncomms16047, have done to reconstruct the ancient angiosperm flower. Is morphological and genetic derivation correlated in Loranthaceae?

In this general context, Köppen profiles (http://dx.doi.org/10.1111/jbi.13154) may be very interesting to establish for the main lineages, too. Do the exclusively warm temperate (subtropical) lineages evolve at different speed from the tropical lineages? There is a group in France that has coarse, should be ok for loranths, Palaeoköppen maps for the past world (we used some of them for the supplement figure in the JBi paper; note they tend to be too cold towards highest latitudes and the deeper one goes back in time). I think Scotese has also something now in this respect. Köppen characterisation/profiles could help to establish potential past corridors and compensate for the lack of fossils to test inferred biogeographic scenarios (I still find it odd, that no loranths have been described from Antarctica, but it may just be a sample/determination artefact, because people are just not looking for them? Alternative may be that Antarctica was already too temperate in the Cainozoic, and the basic pattern ones sees is really Cretaceous; I still fancy Romina's and yours 2007 hypothesis, and like to stress that yours and our age estimates provide minima)

Just filling the blanks won't however not get away the potential ingroup-outgroup attraction problem. *Nuytsia* is fully covered and the most distinct Loranthaceae. Replacing the current set by highly conserved genes may, possibly. Some of the additional markers in Su et al. show interesting features (e.g. a closeness of *Gaiadendron* to the South American aerials), but they would need to be done for at least some representatives of the main loranth clades before they are of any use.

If you just want high supported interclade relationships, just reduce the taxon set to the longest-branching members of each subtribe with full gene coverage, or the shortest-branching ones. This eliminates intraclade phylogenetic noise.

And frankly, it would serve all of us better if you acted in a collegial, collaborative way instead of being adversarial.

My comment on RG and the matrix request was triggered by the pretty unfounded statements in your 2018 paper, which can be seen in many other papers not providing access to their primary data, hence, verified. Since published, they will also never be questioned.
My critiques may be unusually blunt for your taste and experience (I don't need to be political correct), but I comment to make people think (beyond the addressed authors), rectify potentially wrong statements and maybe trigger a discussion (very uncommon in our scientific fields). And when it's easy for me (knowing the data first-hand). And my critiques, like my RG comment, always have a constructive touch and include suggestions for the future. Hence, this long email (also, I have the time).

As I have seen recently, many phylogeny issues are resolved once one has accurate and complete sequences.

It depends. It helps as long as you stay at the level of one placeholder per genus (and may be wrong, attached an old paper, Delsuc et al 2005, showing the backside of increased gene sampling), and stop there.

As soon as one starts to fill the leaves with species or even sampling geographically distant populations, you will face signal issues again, increasingly so with every sample you add. Compared to what you will probably face within the (sub)tribes (not all, but some), the eventually resolved Loranthaceae spine will be an easy task. That's why so many (have to) stay in the light places, keep adding genes or now NGS data but not samples. Understandably, because it's publish-or-perish. A fully supported tree is a sure publication, pointing out signal conflict, can be a dead-end.

Exploratory data analysis using networks will be the only way out, also for the loranths, if you want to go wide not only deep. Not really surprising: do we really want to believe that speciation is a strictly dichotomous process in such volatile plants?

Bear in mind our Santalales sequences generated in the 1990s were from Sanger sequencing from PCR products and in most cases those are of lower quality than ones obtained using automated methods. This is not a fair arena for criticism – we did the best we could with the available technology at the time.

I'm well aware of this, I sequenced my first plant in 1999, and was trained by people who started sequencing even *before* the PCR was invented. And I have harvested quite a lot of sequences of various plant groups, looked at many data sets (including e.g. the famous Soltis et al matrices, which have some issues due to old accessions).

The signal issues in the loranths are not due to quality. There are obvious sequencing artefacts, but the divergence within the Santalales in general is so high that even using usually conserved gene regions (like the 18S) the genuine mutations outnumber the potential sequencing/editing artefacts. Cleaning out the old data may give you some higher BS, but will not eliminate all signal issues.

What may be more an issue is oversaturation. This makes 3rd codon positions and noncoding regions, but also parts of the rDNAs (they fulfil very stringent structural constraints, stems are hindered to evolve, terminals loops and D1-D4 regions are pretty free to) a double-edged sword, they will give the inference more potency to resolve the leaves but will also invite branching artefacts. And incomplete lineage sorting.

Why I think it is necessary to
a) do exploratory data analysis, and your new 2018 data set would have been optimal for showing this given its most comprehensive gene sample (a presumption based on the branch support in your 2018 cladogram). With respect to what I showed in our 2017 papers, it would have been necessary to ***demonstrate*** where your new matrix outperforms the old data, and not just claim it. For instance, you will probably find that the low BS for the first and second branch is (still) because of the alternative of an *Atkinsonia* + *Gaiadendron* clade, and maybe a South American clade including Psittacantheae and *Gaiadendron.* If the latter is mainly a plastid signal, this would reflect not necessarily monophyly (in a strict, Hennigian sense), but that the both have a similar point of origin (note Bayesian inference will tilt into one alternative, even if part of the data objects it). Open question is: how strongly is plastid divergence and geographic distance correlated in Loranthaceae? I mean within the great continental realms.
For the nuclear tree, one likely needs to add a single-copy gene. Being rRNA genes, the 18S and 25S underly very particular mutational constraints. Regarding the deepest relationships, also 160-nt of the 5.8S rDNA, being the most conserved of the three genes in the 35S rDNA cistron, may hold interesting clues regarding the loranth root in comparison to the putative sister clades, although it may be very difficult to amplify (I imagine the flanking ITS1 and ITS2 could be a nightmare in the loranths, and generally in the fast-evolving Santalales)

b) always do an ingroup-only analysis to establish the ingroup relationships (preferred and alternatives); without such analysis you cannot assess whether there may be any ingroup-outgroup branching artefacts: no outgroup sample should ever change ingroup relationships (my guess is, with your new matrix, the splits involving the root parasites will have higher BS support when only using the ingroup).

c) for outgroup-rooting, you have to prove your claim (no ingroup-outgroup branch attraction) by tests. The classic way is what I had to show in File S6 for the Su et al. data, another likely profitable and very quick way would be what I would have done with your new 2018 data set given what you wanted to demonstrate, namely that there is no issue with the outgroup affecting the ingroup: using the evolutionary placement algorithm (EPA) implemented in RAxML (see attached Hubert et al. 2014 for an example) on a tree including one and all three surviving root parasites. The EPA is easy to do (see also RAxML google group threads on the topic). You read in a alignment including all or potential outgroup taxa (e.g. your new

> comprehensive Santalales data set) and the ingroup-optimised tree, and the EPA will place each outgroup taxa at a internode (branch) of the ingroup-only reference tree, providing a probability for the placement. If you have a good outgroup (monophyletic), EPA will place them always at the same internode, in your case that should be the *Nuytsia* terminal branch. The EPA is, however, not invulnerable to LBA (ML has a 50% chance to escape LBA in the "Felsenstein zone", MP will ***always*** get it wrong), so the crucial question would be here: when one takes *Nuytsia* out of the equation, will all the outgroup taxa insert at the *Atkinsonia* branch, and when only *Gaiadendron* is left (being the shortest-branching of the three, it will be the most critical one), will they all insert at the *Gaiadendron* branch? If not, you cannot exclude ingroup-outgroup branching attraction.

So to put this in perspective, if you can come up with $5000 for my lab, I will conduct genome skimming of the remaining 53 genera of Loranthaceae DNAs that I have in my freezer. If you are not willing or able to do this, you will have to simply wait for more work from my lab and collaborators to appear in print.

> See above. Please realise that I'm a privateer, non-profit blogger on science and other things. For my networks blogposts, I can only take actual data into consideration (*Papier ist geduldig* as we say in German). Which you didn't want to share with the public.
>
> Let me make clear what I'm planning to do.
>
> Independent of whether you will provide me with your 2018 matrix or not (I surely will not harvest gene bank again), I may take the 2018 paper as an opportunity to write a post on my Res.I.P. blog on what bugs me with many published plant biogeographic papers. Just one of many examples showing that top-down reconstructions can show little beyond the trivial, even if done in the best-possible way (using a rather sensible chronogram and several methods via the biogeoBEARS script). To do so I'll by mapping the result of the inference on according palaeogeographic maps, i.e. visualise the result in its temporal context. Something rarely done but what I regard obvious to do for such a paper, especially with respect to our 2017 paper in *PeerJ*. In order to show people how easy it would be to put biogeographic inference results into context.
>
> And – provided you reconsider and send me the matrix used for the 2018 paper – there could be a network-advertising post on the *Genealogical World of Phylogenetic Networks*, a quick exploratory data analysis starting where you stopped in the 2018 paper: showing what's behind the BS << 100, where it matters and where not, and why loranths would be a fine and interesting model group that should get as much attention as possible. Including by networks.
>
> Since I don't need any academic acclaim (i.e. peer reviewed papers), I'm fine with pointing out problems, providing ideas how they could be avoided or countered in future for those that are interested in such. This is the reason why I joined the *Genealogical World of Phylogenetic Networks*. Because we usually can make more out of the data than we see in standard phylogenetic publications. We shouldn't fear (or ignore) "low" support.
>
> Cheers, Guido

Sincerely,

Dan Nickrent

* *

Dr. Daniel L. Nickrent, Professor Emeritus
Department of Plant Biology
1125 Lincoln Drive, LSII 420
Southern Illinois University
Carbondale, IL 62901-6509

(618) 453-3223 - office
(618) 453-3823 - lab
(618) 536-2331 - main PLB office
(618) 453-3441 – fax
Work E-mail: nickrent@plant.siu.edu

My professional web site:
http://www.nickrentlab.siu.edu
The Parasitic Plant Connection:
http://www.parasiticplants.siu.edu/
PhytoImages:
http://www.phytoimages.siu.edu/
Co's Digital Flora of the Philippines:
http://www.philippineplants.org/
Illinois Plants
http://www.inhs.illinois.edu/data/plantdb

"In the case of the mistletoe, which draws its nourishment from certain trees, which has seeds which must be transported by certain birds, and which has flowers with separate sexes absolutely requiring the agency of certain insects to bring pollen from one flower to another, it is especially preposterous to account for the structure of this parasite, with its relations to several distinct organic beings, by the effects of external conditions or of habit, or of the volition of the plant itself... It is therefore, of the highest importance to gain a clear insight into the means of modification and coadaptation." Darwin (1859)

* *

[End of story, probably]