What I was not allowed to show #2: Networks explaining molecular evolution in wax plants

https://researchinpeace.blogspot.com/2017/12/what-i-was-not-allowed-to-show-2.html

Thanks to the confidentiality of the review system, experimental research that goes beyond the standard applications may be lost to the Impermeable Fog. In this post, I'll will show several reconstructions I made for a paper of a colleague on *Hoya* (Asclepioideae), the wax plants.

The paper (Wanntorp et al. 2014) stumbled quite a time through the Forest of Reviews. The first journal we tried (*J. Biogeogr.*) was uninterested, mainly because we did not include an explicit biogeographic inference and a molecular dating. Which would have been pretty much chasing ghosts and generating house-numbers under the given circumstances. Nevertheless, since chasing ghosts using house-number is not uncommon in (plant) biogeography, they pointed it out as a deficit important enough to reject the paper (among other things, naturally). So, we added those, and the next journal (*BMC Evol. Biol.*) allowed us two rounds, but finally turned down the paper because ... well, under the line ... the added biogeographic inference and molecular dating didn't suit the second journal's reviewers. So, our brain-and-pain child was sent a third time into the Forest, finally finding a journal, *Taxon*, that would publish the study. With a catch. Because going "too far" for their readership, the editor advised us (for non-scientists: editorial advise in confidential peer review means do it or find another journal) to drop all my nice network figures. Figures and reconstructions that were considered "a nice touch" by a reviewer judging our paper before, and a reason that editor allowed a revision. To protest, I was about to drop out, too. I did not, and still regret it to make a point. Had invested too much time, work, and passion into this piece (like the first author); I couldn't. But when they tried this a second time several years later, in case of a paper on another asclepid, *Cynanchum* (Khanum et al. 2016), I retreated from the co-authorship. Which had the positive effect that my co-authors were allowed publishing our bootstrap consensus networks. A sacrifice worth it (compare both papers).

The background: complex genetic differentiation patterns

The genetic data of my first author (comprising the nuclear-encoded 5'-ETS, ITS1 and ITS2 spacers of the 35S rDNA; and the plastid intergenic *trnH-psbA* and *trnT-trnL* spacer) were complex, why she ended up with me as a co-author.
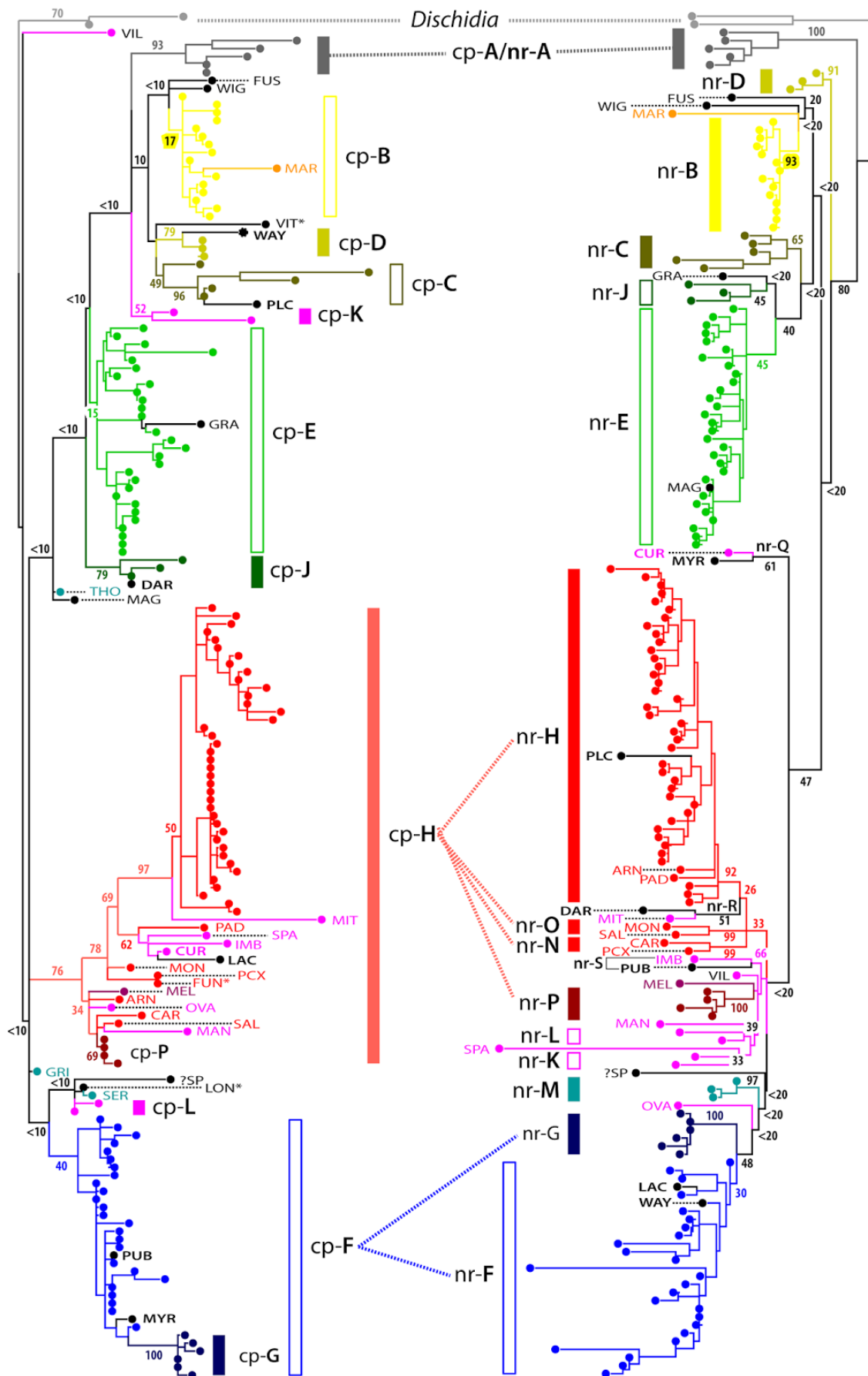
Fig. 1 The plastid (left tree)-nuclear (right tree) simplified tanglegram, a graphic deemed publishable by *Taxon*'s editors. Colours refer to principal intra-generic lineages with lineage-specific nuclear and plastid signatures. Trees are rooted against a few species of *Dischidia*, a closely related (possibly partly co-generic) genus of the Marsdenidae.

But, in the complexity was structure, just waiting to be distilled. Several main clades emerged, supported by nuclear and plastid gene regions, and one could bring them in a quite sensible geographic framework (per hand-and-eye, not really per inference). Because of the signal issues, a full analysis was done: single-gene trees, trees for nuclear vs. plastid data, combined analysis, analysis with restricted taxon set and outgroup, and bootstrap support consensus networks highlighted signal issues, and identified rogues – the taxa jumping across the tree – and the occasional local incongruences. The only figure that passed through final editorial scrutiny was a tanglegram (Fig. 1).

And a (quite ordinary) tree, naturally, based on the combined data and with some critical low supported branches. The electronic supplement at the journal's homepage just includes the voucher information and a tabulation of the figure above. An archive (1.6 MB) including the tree inference and bootstrapping results but not the primary data (first author's call, not mine), too, can be found [here](#).

Why bootstrap consensus networks should be obligatory

In case of complex genetic data, trees may be

- **biased**, showing a topology that prefers branches over better supported, competing alternatives
- **incomprehensive**, showing only the support for the branch in the tree, it cannot be assessed whether a low support is due to a conflicting alternative, or weak but unambiguous signal

So here are the consensus networks (Fig. 2), I prepared for our submitted version(s). It becomes clear that moderate and low support at critical branches regarding inter-lineage relationships relate to generally ambiguous signal. The position of some isolated taxa, collected in low-supported proximal parts of the trees (soft polytomies), cannot be resolved at all, in other cases the data can't decide between two not-so-different alternatives.
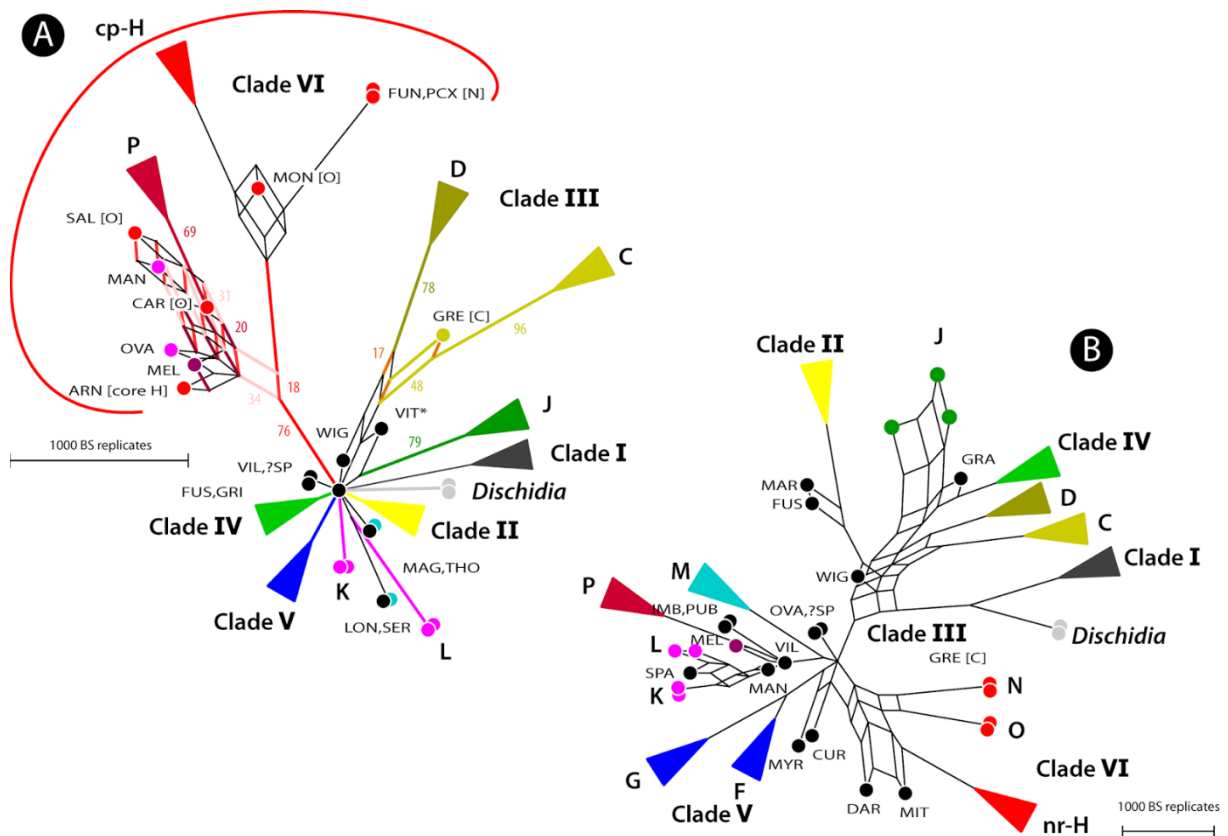
Fig 2 Bootstrap consensus networks for the combined plastid (A; *trnH-psbA + trnT-trnL*) and nuclear (B; 5'-ETS+ITS) data, included in the version for journal #2, and the original version submitted to journal #3 (*Taxon*). And lost to the editorial streamlining. Terminal clades collapsed for better visibility. Labels and colours same as in the (reduced) tanglegram (Fig. 1).

Bootstrap consensus networks are rarely seen in phylogenetic literature and may be quite alien to reviewers, editors and – later – readers. Hence, the figure above was preceded by a figure that included portions of the (full) bootstrap consensus networks at the relevant branches in the combined tree (Fig. 3).
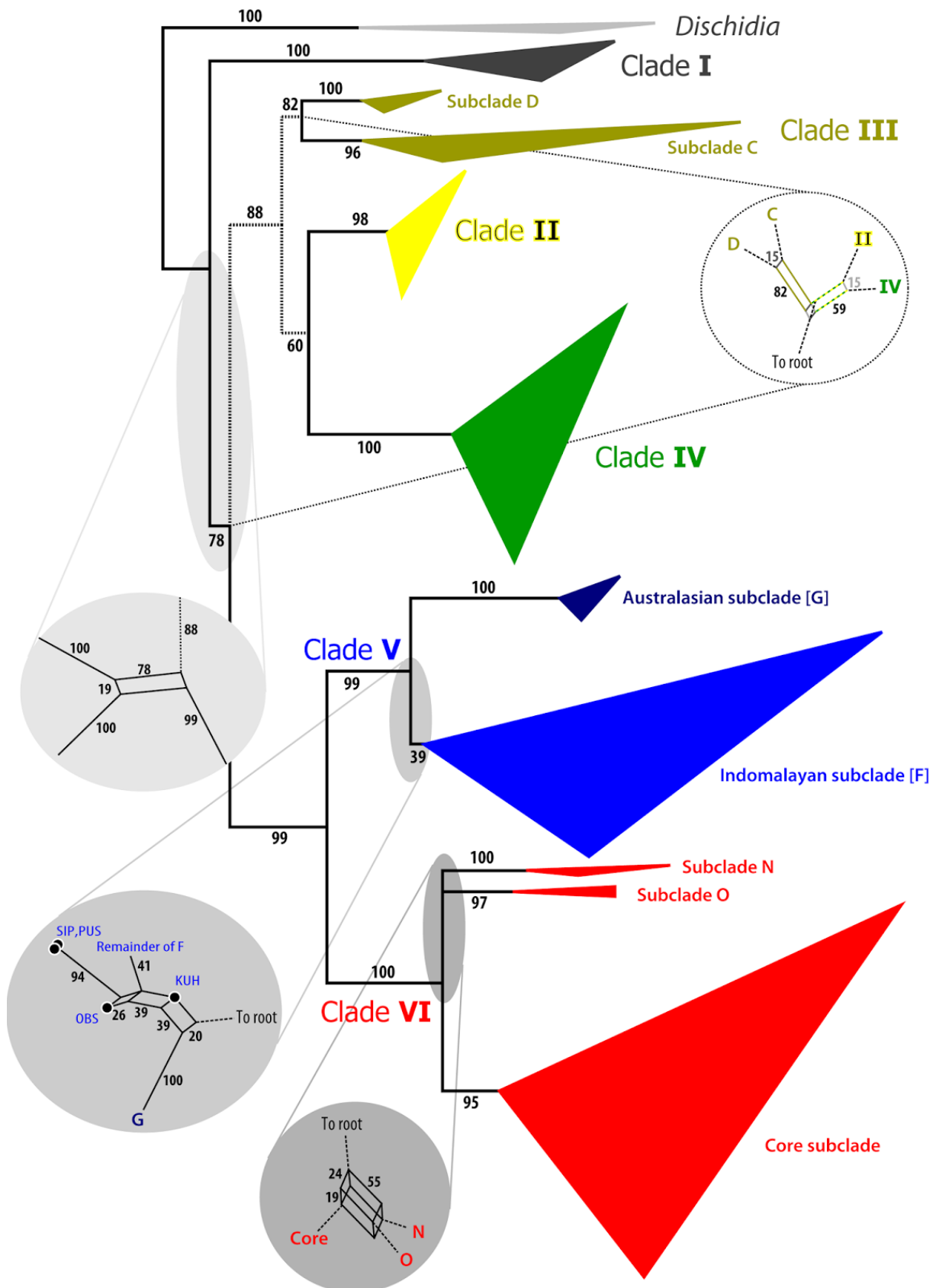
Fig. 3 Combined (plastid + nuclear) tree with phylogenetic "weak spots" addressed using competing bootstrap (BS) support patterns. Colours and labels as before. Values at tree branches and edge bundles give the non-parametric BS support using 1000 pseudoreplicates computed from the combined matrix.

The bootstrap consensus networks can identify how rogues induce topological ambiguity. Moreover, they show whether low branch support (e.g. BS <50) relates

- only to the lack of decisive signal – all alternatives with very low to diminishing support, or
- conflicting signal – two, three competing equally valid alternatives sharing ample (low) support (usually the case here, when it comes to inter-lineage relationships).

They also inform about the reason of non-ambiguous but moderate to high support. For instance, is a BS = 66 due to

- just two-third of the data (segregating sites) supporting this split, the remainder being uninformative, or
- one third of the data support a conflicting split?

Why using median-networks, a population genetic method, for interspecies relationships

The level of *intra*-lineage divergence in each gene region *Hoya* was quite low. Furthermore, I tabulated the sequences (included in the original supplement, but not in the final "edited" version) and checked their mutation patterns. Mutation patterns that revealed another level of ambiguous signal: ancestral sequence variants coexisting with obviously derived ones.

Lack of divergence can be a problem for probabilistic tree inferences, since the likelihood surface of the tree space – the magical plane that decides whether one tree is better than another – is quite flat, more like a lake than a mountain chain. Under these conditions Bayesian inference falters (posterior probabilities, PP << 1.00), and ML tree-inferences struggle, but parsimony can work. The problem with parsimony tree-building was in case of our data the amount of stochastic signal, its complexity. And all tree inferences fall short in the face of actual ancestor-descendant relationships: an ancestor would need to be placed at an internal node, not at a tip (further reading: first phylogenetic trees, stacking neighbour-nets, why networks are inevitable, also in cladistics).

Countering both issues, median networks are a brilliant tool:

1. They include all parsimony trees, the parsimonious solutions that can explain the data.
2. In contrast to ordinary parsimony trees, which treat all taxa as tips, median networks allow placing a taxon at an internal node, the median.

Thus, again in contrast to any traditionally optimised tree, they can infer and depict actual ancestor-descendant relationships. The problem with median networks is that they are

susceptible to convergence (which is like in case of any data) and missing data (which is a problem for not few data sets); and can get easily very multi-dimensional (and messy). At this point, the median-joining networks can step in (Fig. 4).
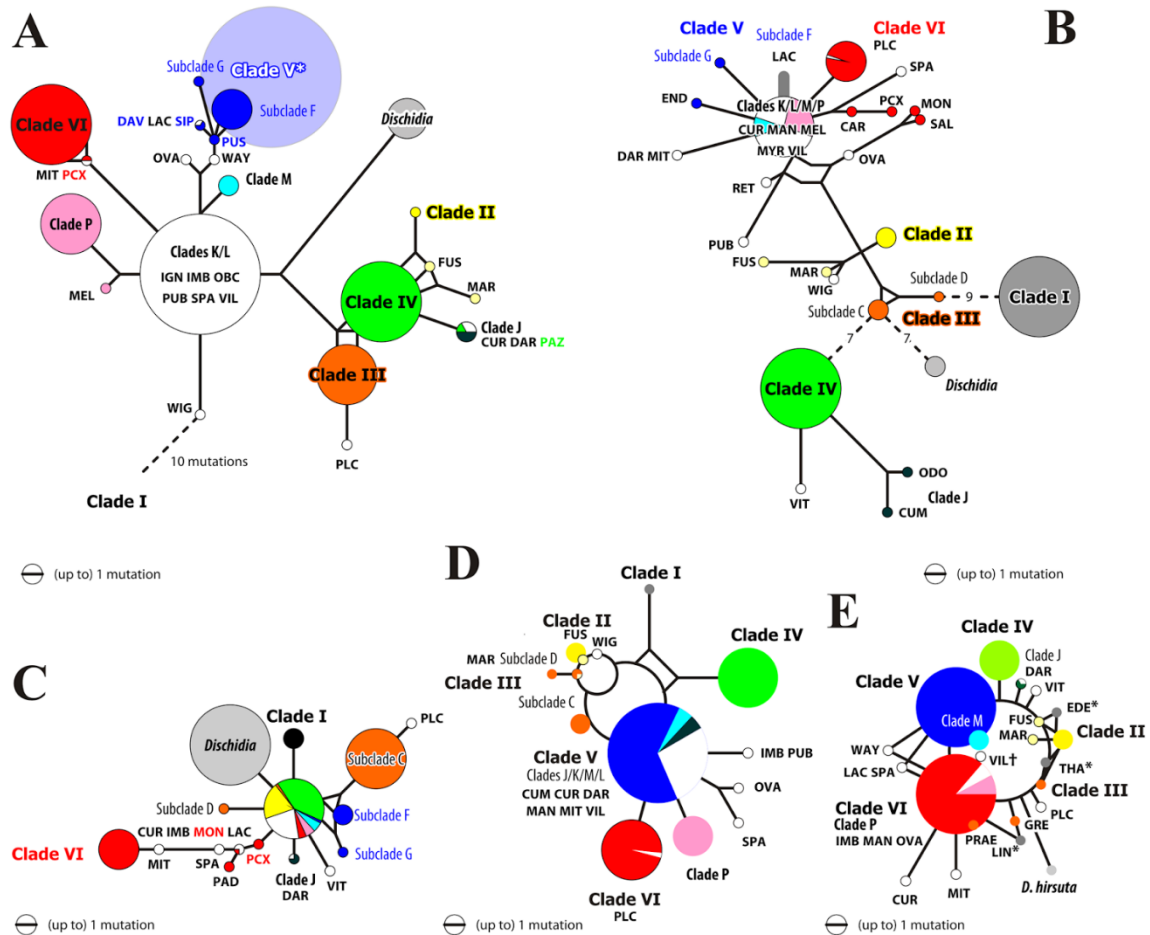

Fig. 4 Simplified median-joining networks based on group-informative sites. A–C. Relatively variable/discriminative spacer regions: nuclear-encoded 5'-ETS (A), ITS1 (B), and plastid *trnT-trnL* (C). D–E. Conserved/rather indiscriminative regions: nuclear-encoded ITS2 (currently advocated as barcode in equally or worse differentiated groups) and plastid *trnH-psbA* (length-heterogenous spacer, only unambiguously alignable nucleotide were considered). Note that the isolated micro-lineages (cyan, pink, white) are often ancestral in their basic sequence compared to the diversified and likely more recently radiated clades V (blue) and VI (red).

Noticing the resolution issues visible from the single-gene bootstrap consensus networks (towards the tips of the trees) and the quite obvious mutation patterns emerging from the data tabulation, I decided using median-joining networks as vehicle to

- make a call about ancestry and derivedness of sequence variants, and see
- how retention of ancestral variants in one or several gene regions affects the tree inference and explains conflicting bootstrap support patterns.

To filter intra-clade noise and reduce the level of (stochastic) homoplasy, I identified and included only group-informative sites in the analysis, i.e. mutation patterns consistent within one of the intrageneric groups (our main clades).

Despite not welcomed by the editors, the network-based findings (Figs 2–4; largely undocumented so far) still play a role in the published paper. They were the main basis for the discussed hypothesis(-es).

Confidential peer review hinders out-of-the-box-thinking

Median-joining networks are parsimony-based graphs; and that I had to omit them did not lack irony. One of the recurrent reviewer critiques was that we did not include a parsimony analysis to back up our maximum likelihood-based analysis framework. Grace to the Impermeable Fog shrouding the Forest of Reviewers, peers (rarer editors) judging the quality of a paper are free to act like imbeciles (on a case-to-case basis). Only authors and editors read their reports. Authors, who are not only bound by confidentiality, but have to be friendly and wilful to not lose the chance to publish their paper. Facing editors who may think very highly of their (anonymous to the authors) catholic (all-encompassing and obviously infallible) peers, and not rarely lack competence in the same fields. Naturally, none of *Taxon*'s editors, a systematic-botanical journal, or our final reviewers, can be considered an expert of phylogenetic inference and comprehensive analysis of non-trivial data patterns. Editing a low/mid-tier journal can be tedious, the editor may be just happy having found *anyone* reporting on the submitted work (which can be difficult enough, in particular for drafts roaming in the Forest of Reviews for quite some time). So that person should be kept happy for the next time(s) his/her services are needed. But since the review process is confidential, the reader (or any 3rd party) cannot know about the fights and fiddles surrounding the publication of a paper. And thanks to *Taxon*'s editorial wisdom, which shields their readers from reconstructions going "too far", the more generally interested reader may find it hard to understand where the ideas of the authors come

from. Because the main reconstructions were stripped from the main paper (and supplement) during review. Like in this case.

Open data and graphics (free to re-use)

The lost figures and supplementary data tables have been uploaded to *figshare* (including the primary data files) and can be referenced as:

- Grimm GW. 2017. Over-the-edge tables and reconstructions linked to the slimmed-down paper of Wanntorp et al. (2014), published in *Taxon.* figshare https://doi.org/10.6084/m9.figshare.5688181

Included in the upload are the used data matrices (NEXUS-formatted) and the raw bootstrap consensus networks (Splits-NEXUS-formatted).

Further links and references

- Grimm, GW. Using consensus networks to understand poor roots. *The Genealogical World of Phylogenetic Networks* (ed. by D. Morrison). http://phylonetworks.blogspot.fr/2017/12/using-consensus-networks-to-understand.html
- Khanum R, Surveswaran S, Meve U, Liede-Schumann S. 2016. *Cynanchum* (Apocynaceae: Asclepiadoideae): A pantropical Asclepiadoid genus revisited. *Taxon* **65**:467–486. — Official link; — electronic supplement archive @ palaeogrimm.org/data (may differ slightly from the finally accepted version).
- Morrison D. 2014. Some things you probably don't know about the bootstrap. *Genealogical World of Networks.* http://phylonetworks.blogspot.fr/2014/04/some-things-you-probably-dont-know.html
- Wanntorp L, Grudinski M, Forster PI, Muellner-Riehl AN, Grimm GW. 2014. Wax plants (*Hoya*, Apocynaceae) evolution: epiphytism drives successful radiation. *Taxon* 63:89–102. — Official link; — electronic supplement archive @ palaegrimm.org/data (1.6MB); — matrix and raw trees @ TreeBase — PDF with the original set of figures (including full legends)

Useful software

- NETWORK 5 – free network software by Fluxus to handle data and compute median, median-joining and reduced median networks: http://www.fluxus-engineering.com/sharenet.htm
- RAxML 8 – the current standard of maximum likelihood analysis, ideal to run a full set of analyses (tree inference + bootstrapping using a batch/shell file) on small data sets like this one (also standalone; no need for a processor grid)
- SPLITSTREE 4 – the free and classic software by Huson and collaborators to generate consensus networks and prepare them for publication