**On building a species-level phylogenetic tree for Neotropical Myrtaceae: conflicting partitions, notes on topology and resources for future studies**

The Neotropical Myrtaceae Group\*

How to cite us: (“ ”)

Authors, in alphabetical order:

[include author name and filiation here]

Thais N. C. Vasconcelos €,±,§,$,£

Department of Biological Sciences, University of Arkansas, Fayetteville, Arkansas 72701, USA

Contributions are as follow:

€ authors responsible for writing

± authors who provided newly available molecular data

§ authors responsible for analysing the data

$ authors who contributed with data cleaning

£ authors who contributed with discussion and comments on the manuscript

# ABSTRACT

Increasingly robust molecular phylogenetic trees underpin studies in several areas of biodiversity. The Myrtaceae family is one of the key components of the exceptionally diverse Neotropical flora and recent massive generation of new molecular data has elevated its status from neglecting to a good model group for eco-evolutionary studies. However, most of this data is either dispersed across different studies or still unpublished and, given its problematic taxonomy, automatic assembling of molecular super-matrices from public databases frequently result in unreliable topologies. Here we build a taxonomically verified molecular super-matrix of Neotropical Myrtaceae by assembling both published and unpublished sequences (3160 and 1024, respectively) from nine molecular markers (the nuclear ITS and ETS and the plastid ndhF, matK, rpl16, trnQ-rps16, rpl32-trnL, trnL-trnF and psbA-trnH). We highlight problems with conflicting topologies inferred from different genes and identify gaps to be filled by future studies. We also provide a time-calibrated tree that can be reliably used to address finer-scale eco-evolutionary questions in the Neotropics. The inferred super-matrix phylogenetic tree covers over 650 species (c. 35% of the total diversity in the clade). The tree inferred from the fully concatenated matrix mostly reflects the topology of the plastid dataset and there is a moderate to strong incongruence between trees inferred from nuclear and plastid partitions separately. Large, species-rich clades are still the poorest sampled within the group. More systematic revisions and field collecting targeting these clades are encouraged, particularly in light of the reported difficulties in extracting sequences from herbarium material.

Key words: *Eugenia*, Myrteae, *Myrcia*, ecology, evolution, super matrix, systematics.

# INTRODUCTION

The exceptionally diverse Neotropical flora accounts for more species of flowering plants than tropical Africa and Asia together (Antonelli and Sanmartin, 2011). One of the dominant components in these environments, especially in the eastern portion of the continent, are the myrtles (Myrtaceae), a family of trees and shrubs of outstanding diversity and abundance in both rainforest and savannas (Oliveira-Filho et al., 2000 (to include more recent references\*)). Myrtaceae species are the base of many Neotropical ecosystems. Besides composing a significant part of the biomass that forms these biomes, they offer flower resources to pollinators and fleshy-fruits to frugivorous, which are produced along all seasons of the year (Staggemeier et al. 2010, 2017).  Thanks to recent systematic knowledge constructed for the family (e.g. Lucas et al., 2007, 2011, 2019; Mazine et al., 2014, 2018), Myrtaceae has been elevated from the status of a neglected group to a model system in addressing broad questions regarding the dynamics of Neotropical biodiversity (e.g. Staggemeier et al., 2015, Lucas and Bunger, 2015, Giaretta et al., 2015, Vasconcelos et al., 2019).

In the centre of this process is the inference of increasingly robust and complete phylogenetic trees for the tribe Myrteae, the most diverse clade in the family (c. 2500 species) and the lineage that comprises all of the Neotropical species. Myrteae has entered the phase of molecular systematics in the 2000s with the work of Wilson et al. (2005) and Lucas et al. (2007) and since then many other studies tackling individual clades (e.g. (cite a bunch of them\*)) or reassessing the tribe’s topology in light of a larger species and molecular sampling (e.g. Staggemeier et al., 2015, Vasconcelos et al., 2017) have been published. These studies have provided data and evidence to justify several taxonomical rearrangements (e.g. (cite a bunch of them\*)) and consistent frameworks to shape and test ecological and evolutionary hypotheses (e.g. (cite a bunch of them\*)).

In spite of these advances, it is not uncommon that larger phylogenetic trees built under the purpose of exploring eco-evolutionary questions contain several taxonomic mistakes in this group (e.g. by inferring lots of politomies or using invalid names (e.g. *Myrtus* in the S&B seed plant phylogeny, to cite other examples)). The reason behind this is that automatized or semi-automatized methods of tree inference usually source information from online databases of sequences and names that have not been updated at the same pace as Myrtaceae taxonomy (e.g. NCBI, The Plant List). On the other hand, the frequent lack in communication among systematists working on different lineages means that molecular data produced by them may not overlap completely to infer large phylogenetic trees, hampering more inclusive analyses. A taxonomically verified molecular matrix including all markers that have been most frequently used to infer phylogenies in the group would likely solve these issues and would be an important resource for future finer-scale studies. Much of the molecular data required to produce a satisfactory result in this sense is already produced, but is either scattered throughout several parallel studies or still unpublished. In this sense, coordinating data sharing to build a broad, inclusive and taxonomically reliable molecular matrix is the most effective way to produce a reliable species-level tree for Neotropical Myrtaceae and to make this resource available for future studies.

Here we track back the recent developments in molecular phylogenetics of Neotropical Myrtaceae to combine molecular data generated from studies published in the last 15 years plus c. 1024 unpublished sequences gathered from the authors of this work. We further re-assessed identification of all vouchers to increase taxonomic accuracy and to explore the support and consistency of topologies inferred from different molecular markers. We also identify gaps to suggest guidelines for future studies. Finally, we provide a densely sampled time-calibrated tree that can be reliably used to explore future eco-evolutionary venues in the Neotropics.

In this context, the purposes of this study are: (1) to assemble a super matrix that covers molecular data produced by several distinct studies and infer a large, taxonomically verified species level phylogeny of Myrteae; (2) to investigate how to increase support and consistency of topologies inferred and where to focus resources for future studies; and (3) to provide a trustable calibrated phylogenetic tree that can be used as a resource for studies that require ultrametric trees (e.g. ecophylogenetics, conservation, macroevolution,, biogeography).

# METHODS

## Neotropical myrtaceae

The predominantly Neotropical clade of Myrtaceae encompasses nine subtribes within of tribe Myrteae: Blepharocalycinae (3 species), Myrciinae (c. 750 species), Pliniinae (c. 120 species), Pimentinae (c. 200 species), Ugniinae (12 species), Eugeniinae (c. 1000 species), Lumiinae (c. 50 species), Myrtinae (c. 20 species) (sensu Lucas et al., 2019). This clade includes c. 2200 species in total of which c. 2000 are Neotropical. Most of these species are distributed in two very large genera, *Myrcia* (in the Myrciinae) and *Eugenia* (in the Eugeniinae) and sections within these genera have similar species-richness as most of the other Myrteae genera. For that reason, these sections were also treated here as “major groups” alongside other genera. *Eugenia* sect. *Jossinia* is the only exclusively non-Neotropical clade in this sample, but it is nested within a clade comprising all Neotropical *Eugenia* and thus had to be included to preserve the monophyly of the sample.

## Data mining from Genbank

There are several methods for inferring trees using super matrix approaches of sparse sequences matrices (e.g. ). However, given the problematic nature of relationships within Myrteae and to guarantee that all vouchers can be tracked back for future consultation, several steps of manual cleaning and checking had to be performed during the process of building the matrix. To reconstruct the most species-inclusive phylogenetic hypothesis for Myrteae, DNA sequences were retrieved from two sources: (1) by data-mining the GenBank and (2) by compiling unpublished sequences from the authors of this study. In the first case, we manually searched nucleotide entries for each of the 52 genera recognized by Wilson (2011) as Myrteae in September/2018. This preliminary list was then filtered to remove sequences not belonging to Myrteae (e.g. sequences from pathogens and parasites that were also occasionally recovered by the searching engine). The remaining sequences were classified by subtribe (*sensu* Lucas et al., 2019) and organized in distinct folders in Geneious v. 11 (refe). Sequences belonging to the subtribe Decasperminae, a non-Neotropical clade sister to all of the remaining extant Myrteae, were also removed except for those of a few species selected to be part of the outgroup (i.e. x, y, z). Following this logic, a few species belonging to other Myrtaceae tribes were also searched and included in the dataset to serve as outgroups (i.e. x, y, z).

Next, we filtered these results to exclude all but nine molecular markers that showed the greatest coverage among the species: the nuclear ITS and ETS and the plastid matK, ndhF, psbA-trnH, rpl16, rpl32-trnL, trnQ-rps16 and trnL-trnF. The remaining sequences were renamed as “species” plus “voucher information” and reorganized in different folders for individual markers. This process was performed manually to avoid combining sequences from different vouchers. Sequences for which no voucher information was found or could not be easily tracked to a herbarium were also excluded in this step. When there were two sequences for the same molecular region and voucher, we kept the one with the longest length. We kept duplicate entries from the same species as long as they belonged to different vouchers. In this way, the dataset compiled from GenBank sequences encompasses xxx sequences and is restricted to specimens for which the vouchers are known and can be tracked back to a herbarium.

## Previously unpublished data – extraction and sequencing protocols

Previously unpublished molecular data gathered from the authors account for 1024 sequences of the nine molecular markers described above. Overall DNA extraction was carried out from silica-dried leaf material using the CTAB extraction protocol (Doyle & Doyle 1987) and left for precipitation under conditions of -18°C in 100% ethanol followed by a purification by equilibrium centrifugation in CsCl-ethidium bromide gradients (1.55g × ml-1). Butanol extraction followed by dialysis were employed to remove the ethidium bromide and caesium chloride. The target regions were amplified on the GeneAmp PCR System 9700 (Applied Biosystems, Foster City, California, U.S.A.) and Mastercycler nexus (Eppendorf, Hamburg, Germany) following the protocols and the corresponding primer described in the Table x. Sequencing reactions were carried out with the Taq DyeDeoxy™ Terminator Cycle Sequencing Kit (Applied Biosystems, Inc). The PCR products were purified using QIAGEN® QIAquick™ PCR Purification Kit according to the manufacturer`s protocol. Sequences were read on an ABI PRISM 3100 Genetic Analyzer (Applied Biosystems, Inc). Primers, PCR and sequencing conditions are detailed in the Supplementary Information 1.

## Final data cleaning and alignment

 In the next step, all names and vouchers were circulated among authors to have their taxonomic identities checked and, when required, to update their identifications prior to tree inference analyses. That also comprised vouchers marked as “sp.” (unidentified species) from both GenBank and unpublished data, enabling new identifications to be performed and included. This procedure of manual cleaning and re-assessment of all data was time-consuming, but also is still the most effective way to address the problem of taxonomic inaccuracy that is widespread in Neotropical Myrteae.

We then ran alignments for each region separately using the Muscle algorithm (refe) implemented in Geneious (refe), with default settings. Alignments were visually inspected and adjusted for issues with sequence in different direction. Sequences that aligned poorly or represented possible contaminations were excluded from the matrix. At the end of this process, the resultant dataset for the nine markers (henceforward “the super-matrix”) was considered cleaned and used in all the subsequent analyses (vouchers information and GenBank number for all sequences are in Supplementary Information 1; newly available molecular data is marked as “(unpublished)”).

## Tree inference and notes on support and resolution

We first inferred a phylogenetic tree including all of the molecular data available, comprising a concatenate matrix of all nine molecular markers and all tips (henceforward “the super matrix tree”). To this end, we used a maximum likelihood (ML) algorithm implemented in RAxML through Cipres, setting it to 1000 bootstrap replications. This analysis was useful to contrast species positions against previously published trees using smaller datasets and to observe general patterns in support and topology among the eight subtribes focused in here.

We then filtered three subsets matrices that were complete for all molecular markers: (1) a subset of 52 tips for which all the nine markers were available (henceforward “the all52” set); (2) a subset of 295 tips for which all the five markers frequently used in studies of the large genus *Myrcia* were sampled (i.e. nuclear ITS and the plastid trnL-trnF, ndhF, psbA-trnH and trnQ-rps16; see Lucas et al., 2011; Staggemeier et al., 2015; Wilson et al., 2016; Santos et al., 2016, 2017; Amorim et al., 2019; L.L. Santos in prep.); and (3) a subset of 314 tips for which all the five markers frequently used in studies of the large genus *Eugenia* were sampled (the nuclear ITS and the plastid trnQ-rps16, psbA-trnH, rpl16 and rpl32-trnL; see Mazine et al., 2014, 2018; Bunger et al., 2016, Faria 2014; Giaretta et al.; 2019) (see also Table 1). Using these subsets, we ran separate ML analyses for each different marker in each subset. We also ran ML analyses for combining partitions from different organelles (“nuclear” and “plastid”) and one analysis for the full-concatenated matrix in each subset (i.e. “full” analysis). All ML analyses were run using the same methods and settings as described above.

To compare topologies and supports of each marker and combined partition against the full analysis, we used three approaches. First, we extracted all bootstrap values in each tree and compared their median, highlighting which trees had highest general support. Then, we estimated the similarity between the topology of each marker against the full analysis by extracting a matrix of distances among tips and calculating Mantel statistics between pairs of matrices. The posthoc significance values were estimated following the Pearson’s method and 1000 permutations, using functions of the R packages *ape* and *vegan* (refes). Values range from 0 (topologies are completely different) to 1 (topologies are completely similar). Lastly, we used functions of the R package *treespace* (refe) to simultaneously explore resolution and topology inferred from different markers and partitions, using a sample of 100 trees randomly extracted from the set of bootstrap replications. This analysis is similar to a multidimensional scaling and visually depicts how different topologies are similar (i.e. how much they overlap in the multidimensional space) and how robust is the support wield from each inference (i.e. the larger the area, the lower the support).

## Identifying taxonomical and geographical gaps

To visually inspect how balanced is the data in terms of taxonomic coverage, we estimated these proportions for each major clade of the super matrix tree. This information is useful to establish where to invest resources in clades that are in need of further studies. To this end, we annotated total species estimate per section within the large *Eugenia* and *Myrcia* from the literature (Mazine et al., 2018 and Lucas et al., 2018, respectively) and per genus from the WCSP (WCSP, 2020).

## Tree calibration

Finally, we produced a time-calibrated tree using the full-concatenated dataset. The fossil dataset used in calibration correspond to that of the approach B in Vasconcelos et al. 2019

# RESULTS

## Super matrix tree inference

The final cleaned super matrix of nine molecular markers contains a total of 4184 sequences, of which 3160 (76%) were retrieved from the GenBank and 1024 (24%) are newly published here. The concatenated alignment has a length of 10942 sites, of which 46% represent missing data. All of the 34 accepted genera of Neotropical clade of Myrteae are represented. ITS and psba-trnH are the markers most commonly represented throughout the super matrix (see Table 1). The tree inferred from this super matrix encompasses 864 tips, representing 681 species, 48 genera (including outgroups) and c. of 34% of the species diversity in the clade. Among the xx previously unidentified species (i.e. those marked as “sp.” on GenBank), xx (xx%) could be newly identified to species level. Xx remain unidentified to species level. 32 changes

## Contrasting resolution and topology of each marker

Analyses using the three subsets with complete molecular matrices highlight topological conflicts between partitions. In general, full-concatenated matrices constantly present higher support values, followed by a concatenated plastid partition and a concatenated nuclear partition. All the nine Neotropical subtribes of Myrteae were recovered as monophyletic, although some with low support (table? Bs values?). The same is observed for several genera and sections of the gigantic *Myrcia* and *Eugenia* (table? Bs?). Eugeniinae is exceptional for presenting the tree inferred using solely the trnQ marker with a higher support than the nuclear partition (see Figure 2). In general, Mantel test posthoc significance values were congruent to the analyses of support; i.e. phylogenies with lower support also had topologies that were more distinct from the fully concatenated one, but some exceptions are also observed. Surprisingly, some of the most recurrently used markers wield very low support and were highly inconsistent with the full matrix, e.g. rpl16, trnlf, ndhf and psba-trnh. Moreover, the most informative individual markers in the all52 analyses are the ETS and the trnQ, two regions that are often sequenced.

The Eugeniinae subset presented the lowest conflict between nuclear and plastidial datasets. In both Myrciinae and all52 subsets, however, there is a strong conflict in topologies inferred from nuclear and plastid matrices, as they seldom overlap in the treespace. In both of these occasions, the topology of the full matrix reflects the topology inferred from the plastid partition, possibly because the plastid partition is more data-rich (i.e. longer) than the nuclear one.

## Taxonomical, biogeographical and molecular coverage

Not surprisingly, the major groups with best taxonomic coverage are the species-poorest ones. Among the xx genera and major groups with over 30 species, xx are represented by less than half of their species-richness. This is the case of the gigantic clades *Eugenia* sect. *Umbellatae* (xx out of xx species sampled), *Myrcia* sect. *Calyptranthes* (xx out of xx species sampled) and (more examples, etc). account for the poorest coverage relative to their total estimated species-richness. In *Myrcia*, sections with best coverage are xx, with xx% of species coverage. In *Eugenia*, that is xx with xx% of species coverage.

## Tree calibration

The dated tree estimates an mid-Eocene origin for the Neotropical Myrteae (c. 41Ma)

# DISCUSSION

## General comments on the topology

Our inferred topology represents the first Myrtaceae phylogenetic tree to include all of the 34 genera of the hyper-diverse Neotropical clade of Myrteae (sensu Vasconcelos et al. 2017b). The broader sampling also highlights several non-monophyletic species (e.g. examples) pointing to future required nomenclatural adjustments in several instances. Most of the resultant suprageneric relationships are also congruent with the recent phylogenetic analyses performed using different samples within tribe Myrteae (see abovementioned studies) and estimated ages for clade divergence are in accordance with previous calibrations. Furthermore, by enhancing collaboration and reassessing data already available on public repositories, we managed to improve. This tree (supplementary material 4) is now in good shape to be used by large scale studies such as Diaz et al 2019, Antonelli et al. 2018, and many others and can be downloaded from the Supplementary material 4 and we forecast multiple usage for this resource.

At the same time, some general comments on topology are necessary. The recently re-circumscribed subtribes based on the informal groups of Lucas et al. (2007) and Vasconcelos et al. (2017) are all recovered as monophyletic, except for Pimentiinae, which was found paraphyletic here (see un-calibrated RaxML tree on the Supplementary Material) and in the latter study (i.e. divided into the informal *Pimenta* and *Psidium* groups). This subtribe was forced to be monophyletic in the calibration analysis to concur to current subtribe classification and because the non-monophyletism of this subtribe may be just an artefact resulting from the sparse sampling. Pimentiinae is found along the whole of the Myrteae geographic distribution in the Neotropics and include genera with wide variation in embryo (refe), floral (Vasconcelos et al. 2019b) and fruit traits (refe). The matter of recovering Pimentiinae as monophyletic has been discussed before (i.e. in Lucas et al., 2019) and is an issue that definitely deserves more attention from future studies, especially given the economical importance as either food (e.g. *Pimenta dioica*, *Psidium guajava*) or invasive (*Psidium cattleianum*) of some members in this clade.

Some general comments and warnings are also required also regarding the topologies of the super diverse Eugeniinae and Myrciinae. The reconstructed phylogenetic hypothesis of Eugeninae mostly confirms the general topology proposed by Mazine et al. (2014, 2018) and Vasconcelos et al. (2017). Myrcianthes is supported as sister to the.. Due to the size of the molecular matrix, relationships among sections within the large *Eugenia* and *Myrcia* also presented generally lower support in the full dataset than those recovered by recently published studies focused on more restricted samples. The relationships among *Myrcia* sections, for instance, received higher support in the super matrix analysis of Amorim et al. (2019), while the same is true for the analysis of *Eugenia* in Mazine et al. (2018). This is probably a result of using a sparse and patchy molecular matrix in this study, which contain several gaps that occasionally reduce branch support. An evidence of this is that the support again increases when just the full subsets focusing on each of these clades is considered (i.e. Fig. 2). In several cases, increased sample of molecules and species are necessary to improve resolution, but care must be taken to interpret the conflict between nuclear and plastid (see section below).

## The future of this data in light of high-throughput sequencing approaches.

Perhaps one of the most important messages of comparing supports and topologies resulting from analyses of different markers is that topologies inferred from nuclear and plastid datasets are strongly incongruent and the full dataset, the topology that has been used to infer the current classification of tribe Myrteae, overlaps better with the plastid dataset (Fig. 2). In any case, we preferred to provide the calibrated tree inferred with the full concatenated matrix because we think it is more biologically realistic to consider both nuclear and plastid data altogether. However, this may a consequence of …

Recent advancements in sequencings techniques and phylogenomics analyses using different techniques sample either more of the plastome or more of the nuclear (?). However, results from nuclear markers can be also problematic due to problems in paralogy with the ITS maker (e.g. refe). In addition to that, the higher support from ITS and ETS regions can be merely due to biases related to and saturation (i.e. very fast mutation rates) in these sequences (e.g. refes). Attention must be paid for interpretation in these instances in the near future, as this can continue to generate chaos in this already unstable group. One example: Algrizea. That means that future studies using broadly used target-capturing 353 genes, which sample only from the nucleous, will possibly increase taxonomic instability in this group. Care must be given to insure that evidence from both descendence (nuclear and plstid) can be combined for a robust classification.

Well, the fact is that for many tropical angiosperms, we are still in this step of knowing the phylogenies. So we get constrained in making more complex assumptions based on the evolutionary history of these species because we don’t know much information about this evolutionary history. On the other hand, phylogenomic approaches are not always helping to solve these relationships. They are still expensive, methodological unstable and usually applied in a small number of species or in a genus level. Local researchers in these countries still don’t have access to these methods. To wait for a phylogenomic solution for 2000 species … So estimations of extinction risk, ecological correlations, microevolutionary processes and macroevolutionary pattern are all jeopardize. Making information in taxonomic literature available for other analyses and nomenclatural stabilisation for posterior data mining.

## Gaps to be filled by future studies

Some of the clades are more urgently in need of revision than others, and our results can help guiding future studies. For instance, the gigantic Eugenia sect. Umbellatae (c. 500 species) has only c. 12% of its species-richness sampled in our phylogeny. In relation to the only non-Neotropical group, sect Jossinia, this lack in sampling may result from the extremely broad distribution in difficult to access areas. A more focused study on this group would be very profitable, as the clades within it present strong geographical structure. Vasconcelos et al., (2017) cites that this is one of the few groups in Myrteae where there is strong evidence for cross ocean long distance dispersal and establishment, so it can be a good model group to studies in this sense. The large tropical clades are still very poorly sampled.

Even though not directly treated in this study, there seem to be also a bias in collecting effort. Areas such as the AF are much better represented than the western Amazon. The predominantly higher latitude subtribes are better sampled, maybe because they are also species poorer. Given the reported difficulties in extracting sequences from herbarium material even using target sequencing approaches (e.g. bla), More field collections are needed, especially focusing on groups that lack sampling and in poorly collected areas. This can be particularly hard as a good part of this work was possible through funding postgraduate students that now have no funding anymore.

## Challenges in producing a manually cleaned super matrix in species-rich groups

Manually assembled datasets are troublesome and time-consuming, but still the best approach for.. Our resulting tree will certainly contribute to finer-scale ecological and evolutionary studies in the Neotropics, but producing a matrix of about 850 specimens and 11000 sites is perhaps the limit for what can be realistically done manually. In here, we chose this approach given the context of the problem: the complex taxonomy of an important model group that required identification of each specimen sampled in the phylogeny to be confirmed by specialists. Nevertheless, choosing the best approach to reconstruct super-matrices certainly depends on the purpose of the study. In the era of big-data, the ever-increasing availability of molecular data is accompanied by the development of tools that enable automatize tree inference and calibration (e.g. Supersmart) or massive data mining from online and free repositories (i.e. the GenBank).

These analytical resources certainly speed up the process of reconstructing trees that are large in terms of both tip number and matrix length and are very useful in the inference of large-scale patterns of biodiversity (e.g. bla, ble, bli, blo, blu). However, had we chosen one of these approaches, we would certainly have added a high amount of unreliable data into our matrix. For instance, data originated from barcoding projects or population genetics were often of little use in our approach and were among the most frequently removed in data cleaning due to the absence of voucher information [incluir exemplos de filogenias cagadas de myrtaceae usadas por aí? Tem um exemplo na arvore do Smith&brown publicada em 2018 onde eles plugam um monte de especie de Myrtus porque retiraram os dados de nomes de uma fonte nao atualizada]. Our manual reassessment of sequences deposited on GenBank also revealed wrongly identified data, and this is probably recurrent in families with challenging taxonomy such as Myrtaceae.

The problem with these approaches is that Myrtaceae is also one of the most diverse tree lineages in the Neotropics, so a reliable tree to which can be used to infer and connect both raw/large general patterns (i.e. signal from macroevolutionary patterns that allow a certain level of data dirtiness) and micro-processes (i.e. microevolution, solving species complexes, speciation mechanisms, etc) is desirable. People talk about the possibility of reconstructing a “as real as possible” tree of life, but this is only truth if taxonomic knowledge is used to clean data toward the tips, and not just stabilizing relationships at deeper nodes. Ideally, there should be a way to involve taxonomists more actively in cleaning these datasets, even if a posteriori (i.e. after the data has been deposited) as the constant curation of online datasets is a must in the era of big-data. Large phylogenies do not combine quality in both levels, which is the quality that will really make us understand from processes to patterns in biodiversity, especially where this is the richest as in the Neotropics.

**Conclusions**

Here we manually assembled and cleaned a densely sampled molecular super matrix for Neotropical Myrtaceae. We showed that reassessing the identification of vouchers used in other studies improved identification accuracy of currently available data, which is of extremely importance to diminish biases in analyses using taxonomically complex groups such as this. We also provided over 1000 new sequences representing one fourth of the currently available molecular information that can be trustily used for ecological and evolutionary inferences in this important tropical group. This process resulted in a broadly sampled calibrated tree certified by systematists and in good shape to be used by ecologists and evolutionary biologists to reliably test hypotheses in these areas, avoiding data of dubious identifications. Gaps for future studies are highlighted. Filling them will involve solving the incongruence between plastid and nuclear partitions, including a broader sampling of species in mega-diverse groups and resolutions for some taxonomic issues, such as the monophyly of the Pimentiinae group.

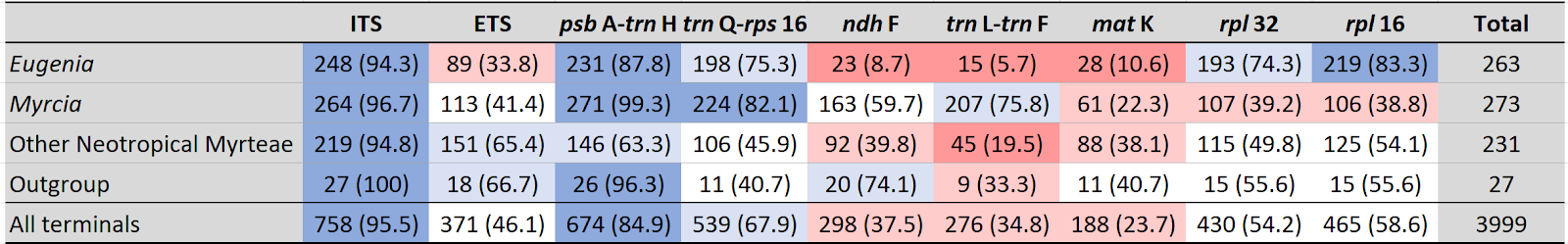
Given the difficulty in collecting proper samples for sequencing, taxonomic chaos and large size, compiling large phylogenies of tropical groups is a challenge. However, these phylogenies can represent a window to understand diversity in tropical forests and savannas where they are best represented and environments under strong anthropogenic pressure. Additionally, taxonomic validated phylogenies and nomenclatural trait standardizing are increasingly important to provide state of art databases to test hypothesis related to ecology, evolution and systematics, the key-areas of biodiversity research.

**ACKNOWLEDGEMENTS**

# References

A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree

**Table 1:** Molecular sampling for 794 Myrtaceae specimens: number of sequences (% coverage). Reddish indicates lower molecular coverage and bluish higher coverage



**Figure 3:** A ML phylogenetic tree inferred from the concatenated super matrix of nine markers, xx tips and eight predominantly Neotropical subtribes of tribe Myrteae (Myrtaceae). Bootstraps and pps? – could go in the figure to the major clades?

Supplementary information (SI):

SI1: Primers and PCR conditions

SI2: List of vouchers and GenBank accession numbers

SI3: Concatenated alignment for the nine regions used to infer the ML and the calibrated tree.

SI4: Calibrated tree in newick format.

SI5: Figure 1 in high resolution.