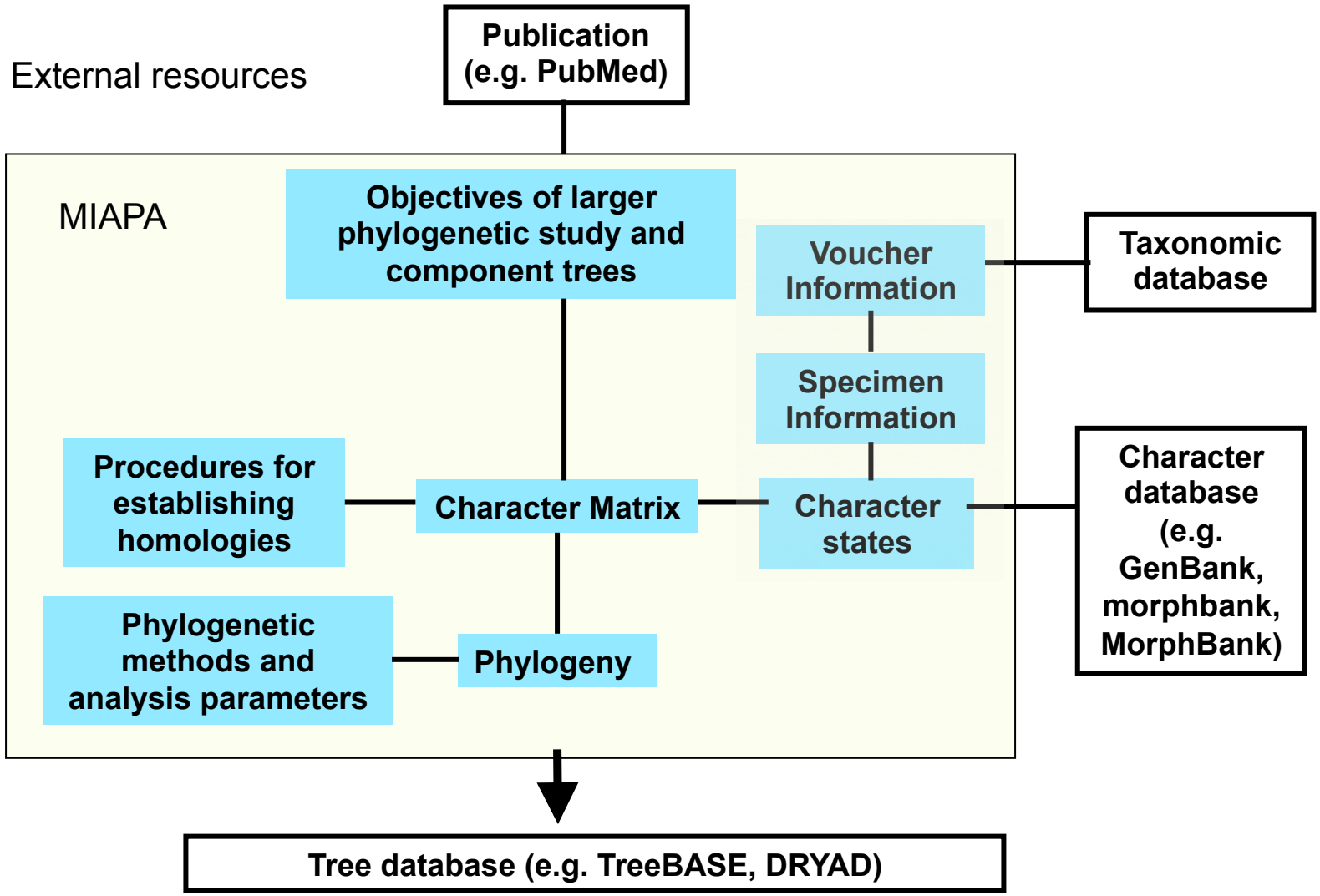
MIAPA (**M**inimum **I**nformation **A**bout a **P**hylogenetic **A**nalysis)



Minimum Information?

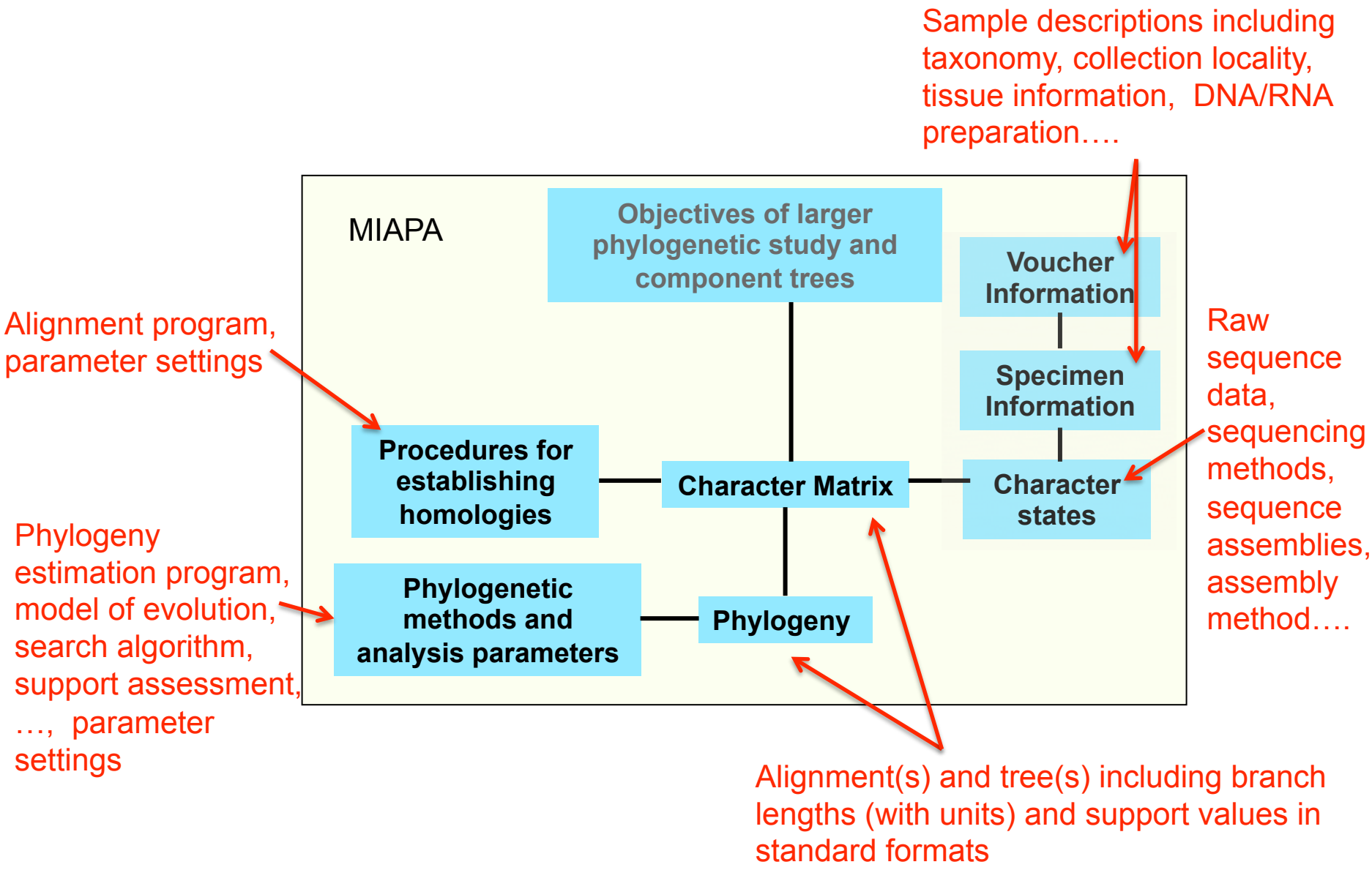
Objectives of a MIAPA standard?

- **Facilitate reuse of trees and data**
  - Ensure digital accessibility of trees and underlying data
  - Aid database search
  - Enable assessment of tree(s) to be reused
- **Facilitate reuse of analysis workflow**
  - Enable running of automated workflow
- **Facilitate replication of analysis**
  - Enable running of automated workflow

A Standard Workflow for Phylogenetic Inference

1. Pose question(s)
2. Design a sampling scheme (taxa, specimens, genes)
3. Collect (and voucher) specimens with meta-information (e.g. locality, conditions….) – this might be done AFTER data collection when data are obtained from databases.
4. Collect data to be used to estimate tree (e.g. generate and/or reuse sequence data)
5. Infer homologies and construct data matrix (e.g. align sequences)
6. Estimate tree(s), typically with support values
7. Interpret tree(s)
8. Publish trees and interpretations with respect to specific question(s).

Typical Molecular Phylogenetic Study

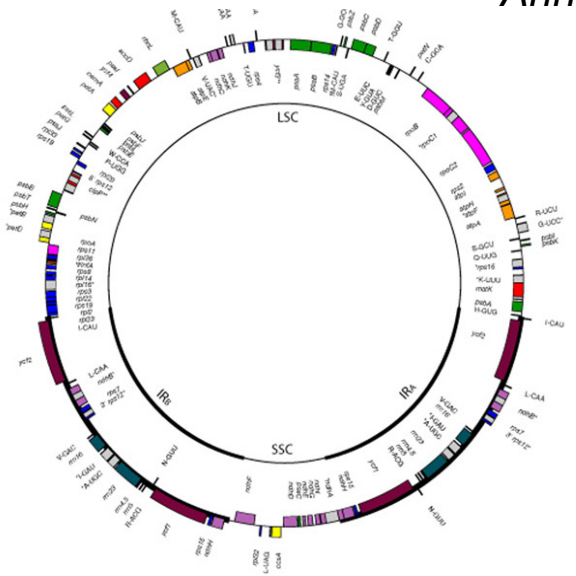


A “simple” example

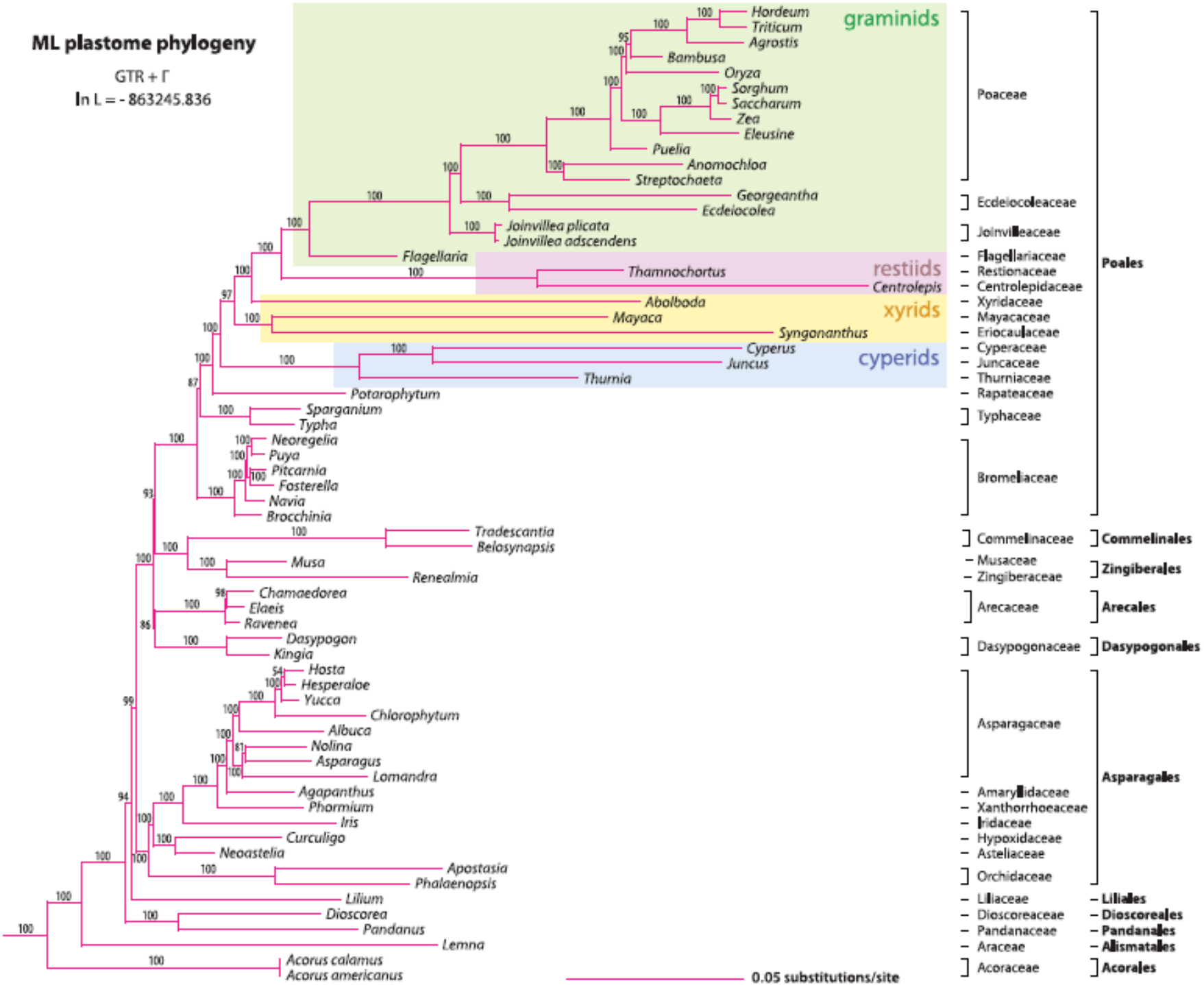
ASSEMBLING THE TREE OF THE MONOCOTYLEDONS: PLASTOME SEQUENCE PHYLOGENY AND EVOLUTION OF POALES[1]

Thomas J. Givnish,[2] Mercedes Ames,[2] Joel R. McNeal,[3] Michael R. McKain,[3] P. Roxanne Steele,[4] Claude W. dePamphilis,[5] Sean W. Graham,[6] J. Chris Pires,[4] Dennis W. Stevenson,[7] Wendy B. Zomlefer,[3] Barbara G. Briggs,[8] Melvin R. Duvall,[9] Michael J. Moore,[10] J. Michael Heaney,[11] Douglas E. Soltis,[11] Pamela S. Soltis,[12] Kevin Thiele,[13] and James H. Leebens-Mack[3]

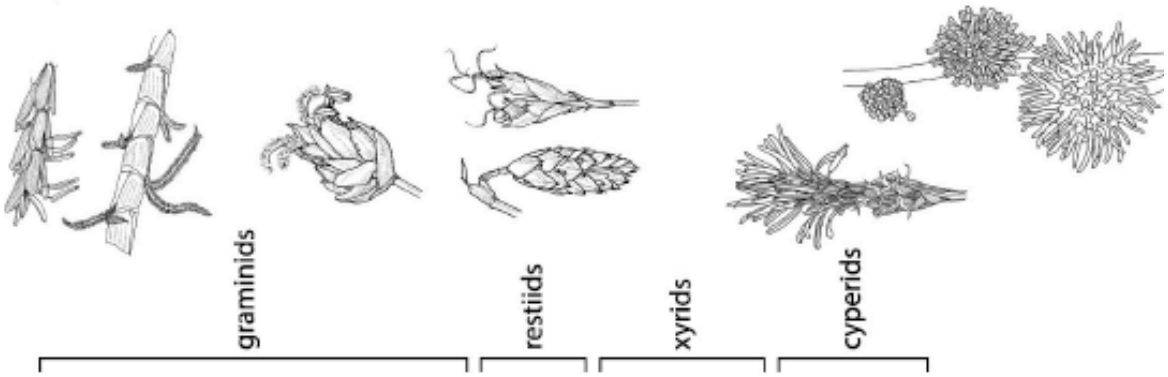
Annals of the Missouri Botanical Garden, 97(4):584-616. 2010



- New sequences assembled from Illumina Genome Shotgun Sequence (GSS) data
- 83 taxa, 39 reused from previous studies
- 81 plastid genes
- ML and MP analyses
- Assessments of trait evolution

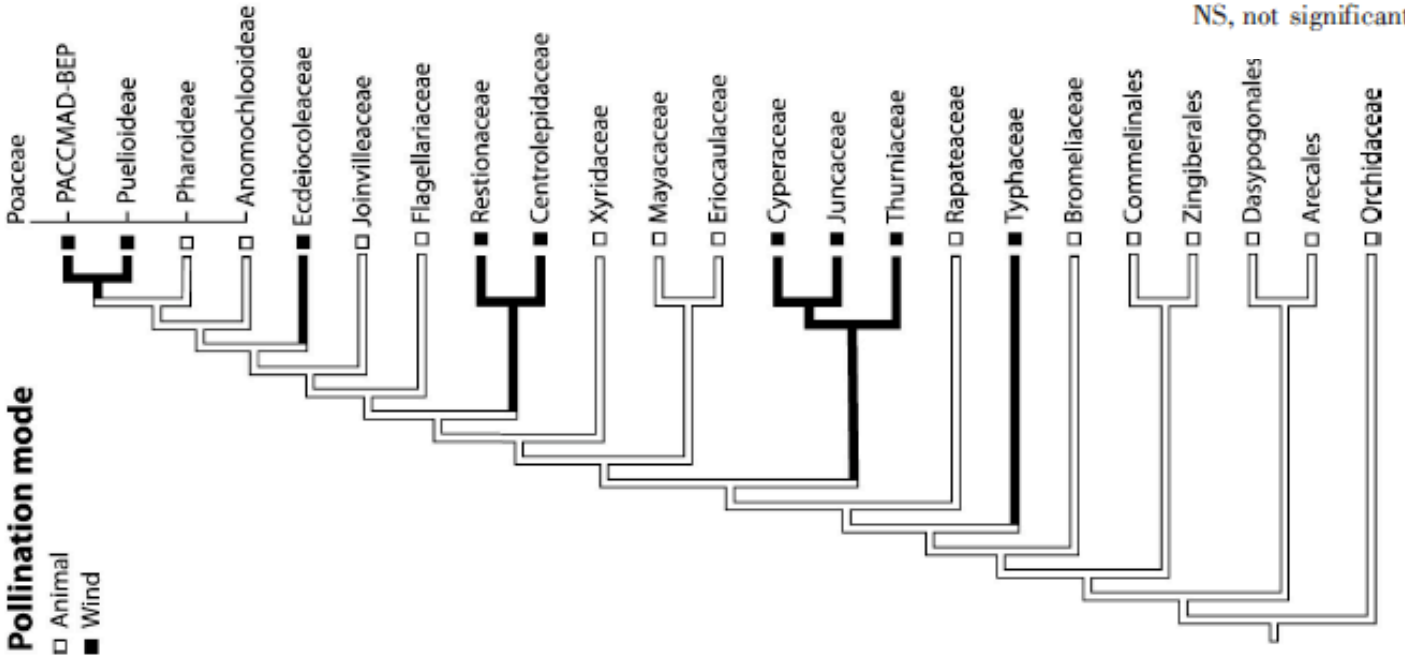

ML plastome phylogeny
GTR + Γ
ln L = - 863245.836
graminids
restiids
xyrids
cyperids
Hordeum
Triticum
Agrostis
Bambusa
Oryza
Sorghum
Saccharum
Zea
Eleusine
Puelia
Anomochloa
Streptochaeta
Georgeantha
Ecdeiocolea
Joinvillea plicata
Joinvillea adscendens
Flagellaria
Thamnochortus
Centrolepis
Abolboda
Mayaca
Synganthus
Cyperus
Juncus
Thurnia
Potarophytum
Sparganium
Typha
Neoregelia
Puya
Pitcarnia
Fosterella
Navia
Brocchinia
Tradescantia
Belosynapsis
Musa
Renealmia
Chamaedorea
Elaeis
Ravenea
Dasypogon
Kingia
Hosta
Hesperaloe
Yucca
Chlorophytum
Albuca
Nolina
Asparagus
Lomandra
Agapanthus
Phormium
Iris
Curculigo
Neoastelia
Apostasia
Phalaenopsis
Lilium
Dioscorea
Pandanus
Lemna
Acorus calamus
Acorus americanus
0.05 substitutions/site
Poaceae
Ecdeiocoleaceae
Joinvilleaceae
Flagellariaceae
Restionaceae
Centrolepidaceae
Xyridaceae
Mayacaceae
Eriocaulaceae
Cyperaceae
Juncaceae
Thurniaceae
Rapateaceae
Typhaceae
Bromeliaceae
Poales
Commelinaceae
Commelinales
Musaceae
Zingiberaceae
Zingiberales
Arecaceae
Arecales
Dasypogonaceae
Dasypogonales
Asparagaceae
Amaryllidaceae
Xanthorrhoeaceae
Iridaceae
Hypoxidaceae
Asteliaceae
Orchidaceae
Asparagales
Liliaceae
Dioscoreaceae
Pandanaceae
Araceae
Liliales
Dioscoreales
Pandanales
Alismatales
Acoraceae
Acorales

Trait Mapping



Trait	Significance with $\kappa = 1$	Significance with optimal κ
Flower size	$P < 0.012$	$P < 0.027$
Floral showiness	$P < 0.0034$	$P < 0.013$
Floral sexuality	$P < 0.0002$	$P < 0.0005$
Habitat openness	$P < 0.013$	$P < 0.02$
Nectar production	$P < 0.035$	$P < 0.038$
Dioecy	NS	NS
Ovule number	NS	NS

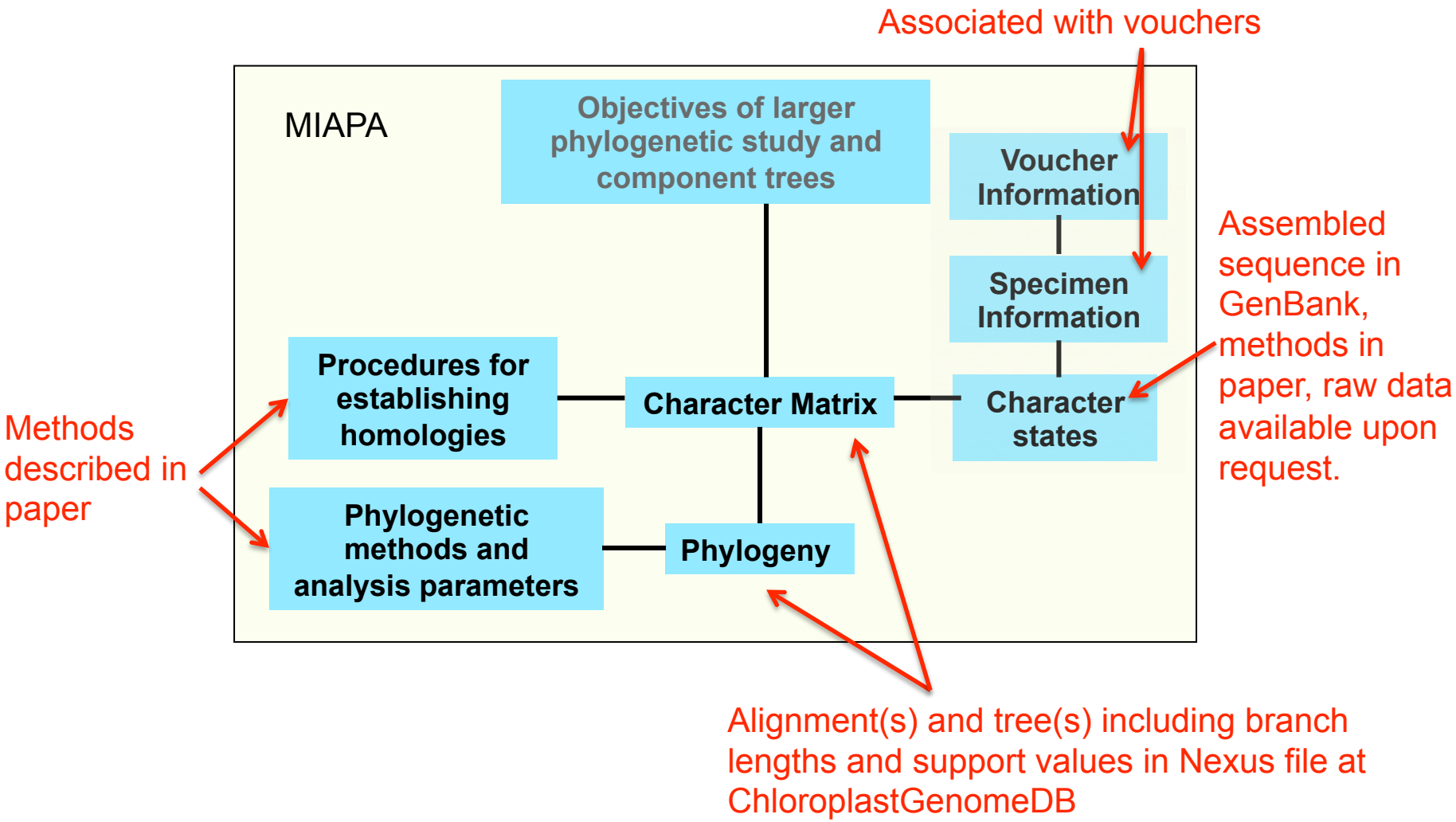
NS, not significant.



Major clade	Order	Family	Species	GenBank accession numbers*	Voucher data†
		Centrolepidaceae	*Centrolepis monogyna* Benth.	this study*	*McKain 116* (GA)
		Cyperaceae	*Cyperus alternifolius* L.	this study*	*Leebens-Mack 1002-2010* (GA)
		Ecdeiocoleaceae	*Ecdeiocolea monostachya* F. Muell.	this study*	KRT3786 (PERTH)
			Georgeantha hexandra B. G. Briggs & L. A. S. Johnson	this study*	KRT3775 (PERTH)
		Eriocaulaceae	*Syngonanthus chrysanthus* Ruhland	this study*	*M. Ames 10/15/2009* (WIS)
		Flagellariaceae	*Flagellaria indica* L.	this study*	*K. Hansen 77-394* (BH)
		Joinvilleaceae	*Joinvillea ascendens* Gaudich. ex Brongn. & Gris	this study*	*Lorence 9066* (NTBG), *800379* (NTBG)
			Joinvillea plicata (Hook. f.) Newell & B. C. Stone	FJ486219–FJ486269	Leseberg & Duvall, 2009

* GenBank accession numbers for plastid genes newly sequenced in this study are HQ180399–HQ183709. A spreadsheet listing individual accession number for each region and species is available at <http://chloroplast.cbio.psu.edu/supplement.html>.

† Voucher specimen (collector and number, with acronym for herbarium of deposit, or citation for sequences previously published elsewhere).



Trait mapping data, methods and results (trees) described in paper text and figures

Minimum Information?

What are the requirements of a MIAPA standard?

- **Facilitate reuse of trees and data**
  - Ensure digital accessibility of trees and underlying data – database submission (accessibility though web-services?), standard data models
  - Aid database search – controlled vocabluaries
  - Enable assessment of tree(s) to be reused – standard data models, controlled vocabluar (ontologies)
- **Facilitate reuse of analysis workflow**
  - Provide access to plug-and-play workflows(?) – data models/ ontologies for workflows (including parameter settings)
- **Facilitate replication of analysis**
  - Enable execution of workflow with data used in study – all of the above

MIAPA (**M**inimum **I**nformation **A**bout a **P**hylogenetic **A**nalysis)

Where do we start?

- **Define stakeholder needs/objectives**
  - Minimum requirements will follow from objectives
- **Publish checklist**
  - Encourage compliance
- **Develop/use data model standards (NeXML, phyloXML)**
- **Develop/use controlled vocabularies/ontologies (e.g. CDAO, PhylOnt)**
- **Develop/use databases with APIs and support of web-services for database entry and extraction**
- **Develop database entry forms**
  - Faciltiation of data entry
  - Implementation of controlled vocabularies
  - Enforcement of compliance??

Possible MIAPA Checklist for Promoting Reuse (Stoltzfus et al.)

- **Informatics context.** Each record (including tree, data matrix and metadata) should draw on accepted practices to make clear how it is to be processed (formats, languages). Each record will include a globally unique and stable identifier (GUID).
- **Scientific context.** The scientific context for the analysis will be provided, including the identities of responsible experimenters, links to any associated publication, and the purpose of the analysis.
- **Reusable trees.** Phylogenetic results will be represented in an electronic form that can be processed without loss of information (e.g., Newick, NeXML; graphics files are not sufficient).
- **Identifiable OTUs.** The OTU objects represented in electronic files will have identifiers that are externally meaningful (e.g., LSIDs) or, if local names are used, will be linked to such identifiers.
- **Support values.** As inferences, phylogenetic results are estimates, therefore a best practice is to present each result with a measure of uncertainty. Support for clades or bipartitions, if reported in an associated publication, will be included.
- **Inputs**. The record will represent, either by inclusion, or by external reference, the data on which the phylogenetic inference is based (typically a character matrix or alignment; mapped traits and/or fossil calibration points may be relevant).
- **Methods**. The method by which the phylogenetic result is inferred from the inputs will be described, ideally by drawing on a controlled vocabulary.

Acknowledgements

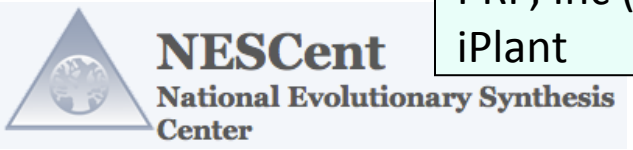
Community Support (Initial Publication)

Brent D. Mishler
Christian Zmasek
Claude Depamphilis
Clifford W. Cunningham
Dennis W. Stevenson
Douglas E. Soltis
Elizabeth A. Kellogg
Eric Brenner
Eugene V. Koonin
Herveé Philippe
J. Chris Pires
Jeff J. Doyle
Jim Leebens-Mack
John E. Bowers
John Harshman
Jonathan A. Eisen
Kerr Wall
Kimmen Sjölander
Mark J. Clement
Pamela S. Soltis
Rob DeSalle
Robert K. Jansen
Seung Y. Rhee
Steven Cannon
Tandy Warnow
Todd Vision
Xun Gu
Yin-Long Qiu

Organizations

NESCent
CDAO
NeXML
DRYAD
pPOD
TDWG
PRF, Inc (TreeBASE)
iPlant

Arlin Stoltzfus
Enrico Pontelli
Maryam Panahiazar
Jamie Estill
Nico Cellinese
William Piel
Val Tannen
Rutger Vos
Hilmar Lapp
Dawn Field
Chris Taylor…



Fine Scale Provenance Relationships for Molecular Studies

1. An estimated tree is derived from a specific analysis of aligned sequences
2. An alignment is derived from a set of contig sequences using an MSA procedure
3. Contigs are derived from assembly of sequences
4. "Called" sequences are interpreted from raw data from sequencer
5. Sequenced tamplate derived DNA/RNA extractions
6. DNA/RNA extractions derived from (vouchered) specimens
7. (vouchered) specimens identified (given taxon ID) by an expert
8. Specimens collected at a specific locality