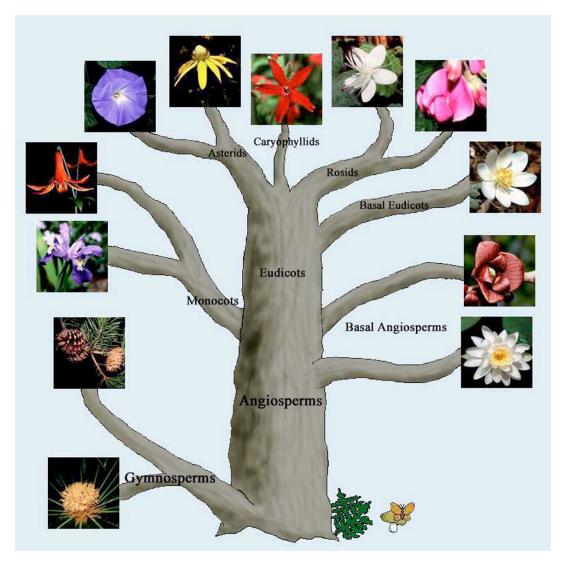
Plant Phylogenomics: Lessons from the 1KP Project



Jim Leebens-Mack

Department of Plant Biology University of Georgia

New Methods for Phylogenomics and Metagenomics Symposium Feb, 2013

Ancestral Angiosperm/ Amborella Genome Vic Albert Raj Ayyampalayam Brad Barbazuk John Bowers Jim Burnette Srikar Chamala Andre Chanderbali Josh Der Claude dePamphils Jamie Estill Hong Ma Doug & Pam Soltis Stephan Schuster Sue Wessler Rod Wing Kerr Wall Norm Wickett Eric Wafula

Collaborators

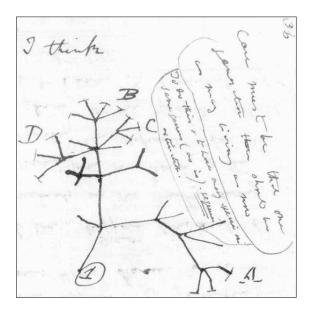
MonAToL

Claude dePamphilis Tom Givnish Cecile Ané Raj Ayyampalayam Sean Graham Dennis Stevenson Jerry Davis Alejandra Gandolfo Chris Pires Norm Wickett Wendy Zomlefer Michael McKain Jill Duarte

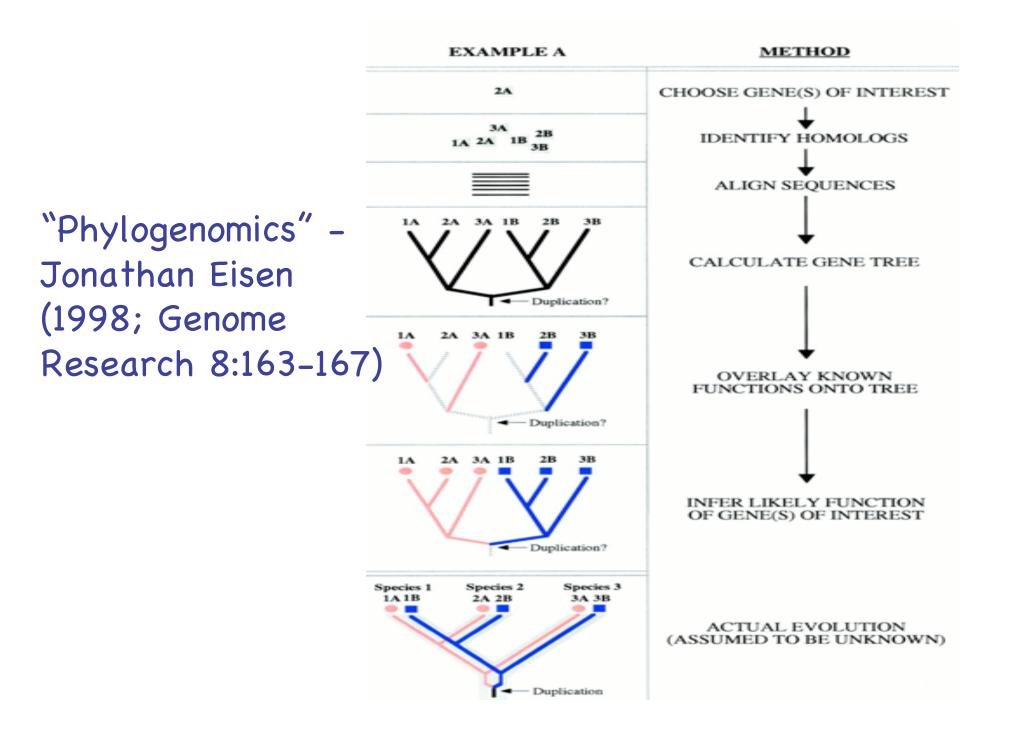
Funding: NSF, iPlant, University of Georgia, OneKP

OneKP/MSA AToL Norman Wickett Nam Nguyen Siavash Mirarab Naim Mataci Gane Ka-Shu Wong BGI Eric Carpenter Brad Ruhfel Herve Philippe. Gordon Burleigh Matt Barker Claude dePamphilis Tandy Warnow Jamie Estill Raj Ayyampalayam Doug & Pam Soltis Sean Graham Dennis Stevenson Michael MelkonianOneKP Consortium "Nothing in biology makes sense except in the light of evolution" (Theodosius Dobzhansky, 1973)

"Nothing in evolution makes sense except in the light of phylogeny"



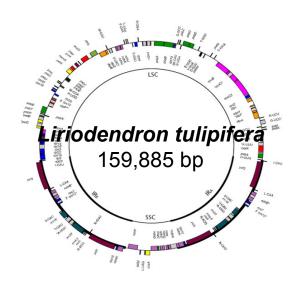
Darwin (1837) First Notebook on Transmutation of Species



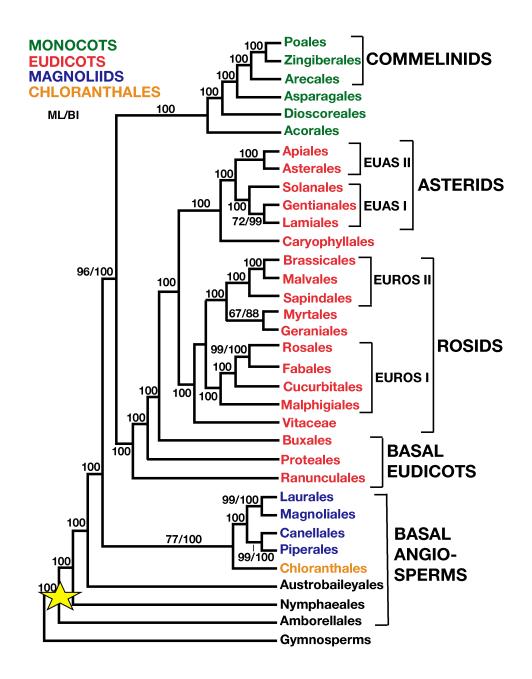
Current Usages

- 1. Using genome-scale data to resolve phylogentic relationships
- 2. Genome-Scale comparisons placed within a phylogenetic context

Phylogenomics¹: Plastid Genome Phylogeny Resolving Many Previously Intractable Questions in Plant Systematics



Jansen et al. 2007 PNAS



What About Nuclear Gene Histories?

Grouped by Phylogeny

Angiosperms

Non Flowering

Green Algae

Grouped by Application

Agricultural

Biochemical

Medicinal

Extremophytes



www.onekp.com

Gane Ka-Shu Wong (Alberta)

Alberta Innovates





musea ventures

AAGP Ancestral Angiosperm Genome Project

ancangio.uga.edu





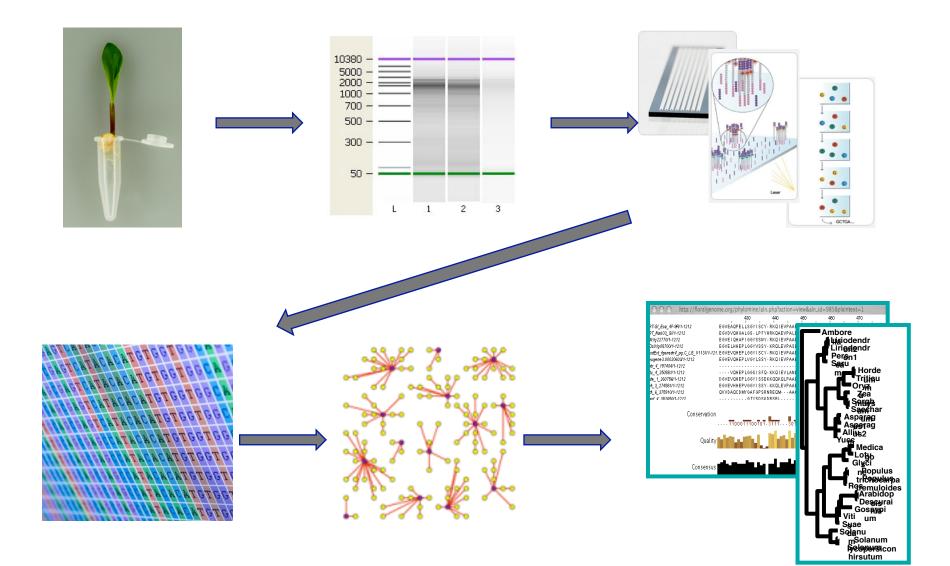
monatol.uga.edu



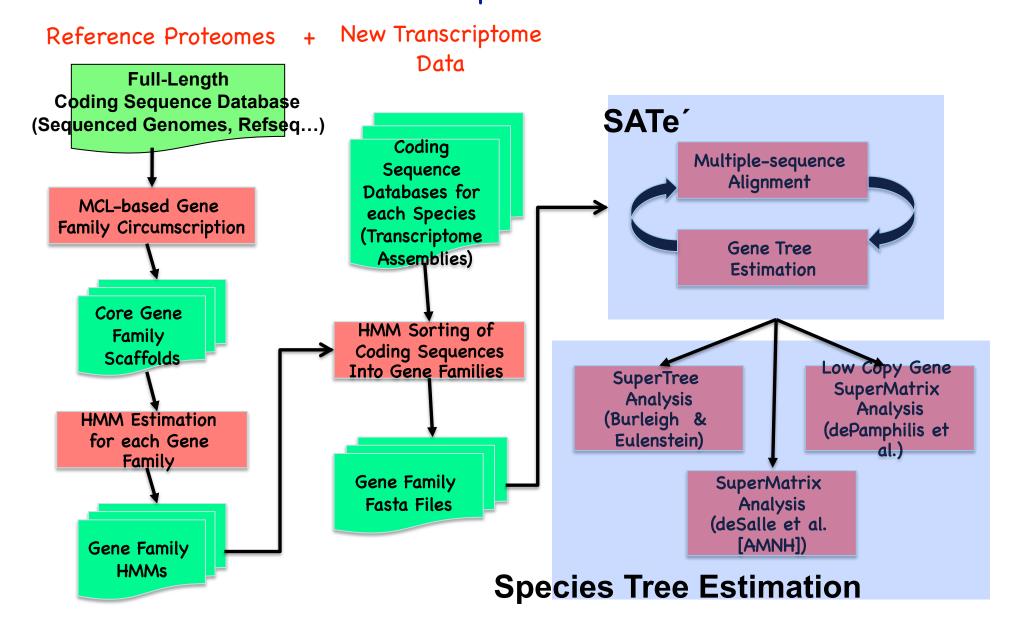




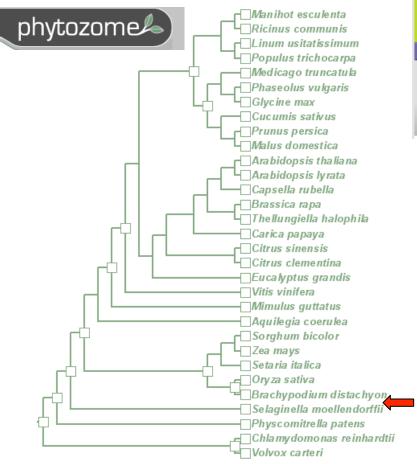
Phylo-Transcriptomics...



Gene Family Circumscription, Sequence Alignment, Gene Tree Estimation and Species Tree Estimation

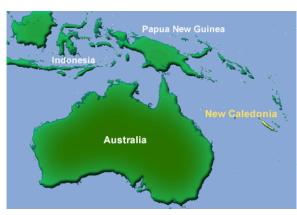


Gene Family Classification for Full-Length Genes from Sequenced Genomes – orthoMCL clustering

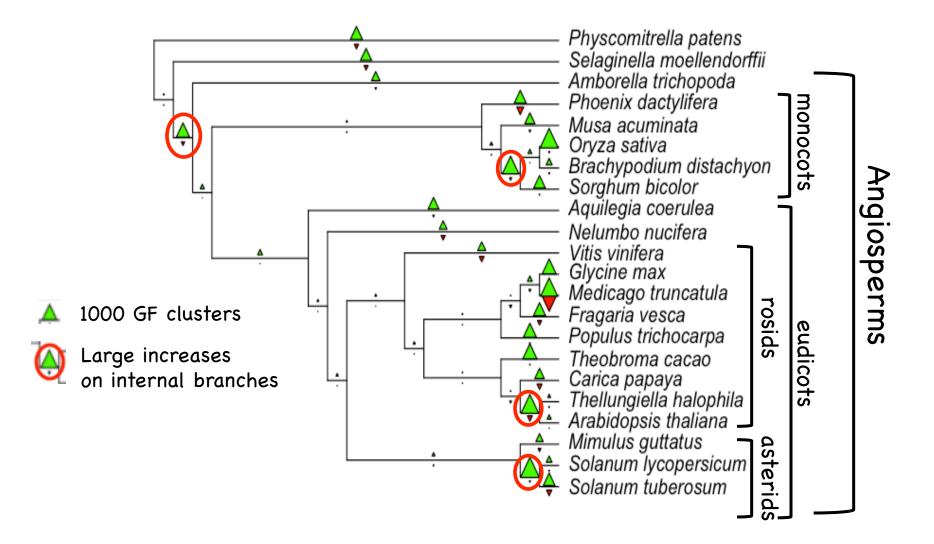


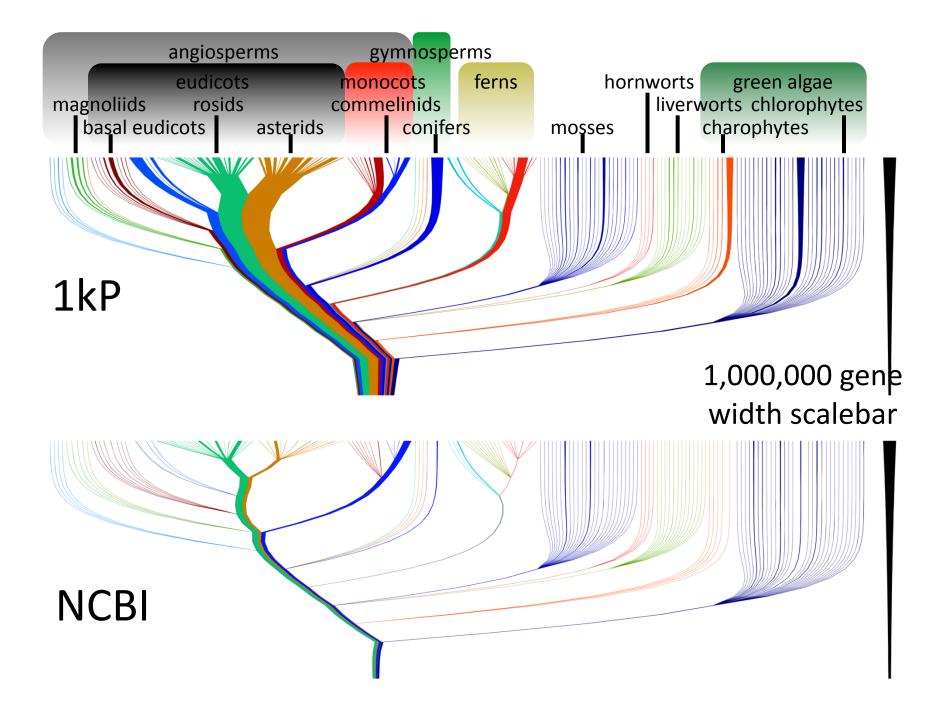






Gain and Loss of "Gene Families"





1kp -Questions to be Addressed through Estimation of Gene Trees and Species Relationships

- What are the relationships among lineages across the green plant tree of life?
- What was the nature and gene composition of the likely ancestor of the Viridiplantae?
- What was the impact of lateral gene transfer (from bacteria) on the early evolution of the Viridiplantae and the Chlorophyta/ Streptophytra?
- Were gene and/or genome duplication events associated with innovations in green plant evolution including the origin of the flower, the seed, the vascular cambium (wood), alternation of generations, the shift from a life history dominated by the haploid phase to a dominant diploid life stage, colonization of land and shifts from single cells to multicellularity,
- Has polyploidy played a role in the diversification of angiosperms?

Check out OneKP data



Blast for OneKP Project

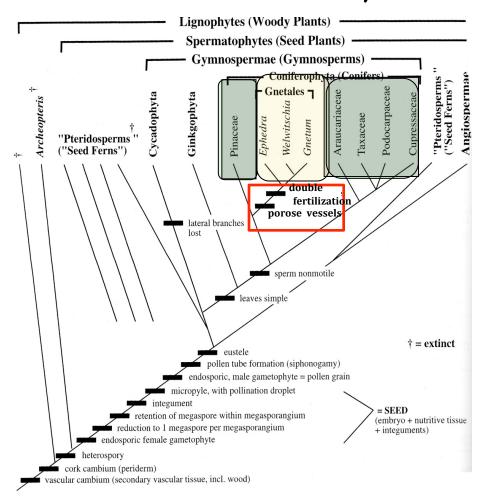
Document

Terms of use: These sequences are being released in advance of publication as a service to the community. We only ask that you follow the spirit of the <u>Fort Lauderdale agreement</u> and refrain from doing the kinds of analyses for which these data were generated, as described on the <u>1KP project website</u>. More generally we ask that you refrain from using these data in any studies that involve multiple genes across multiple species (e.g. studies of a biological process in a particular taxon). Analyses of one or two genes across multiple species, or multiple genes in one species, are generally not in conflict. If you wish to publish your findings, we ask that you credit 1KP with minor authorship. The individuals who must be credited will depend on the particular data that you use. Please contact <u>gane@ualberta.ca</u> or <u>ejc@ualberta.ca</u> for additional clarification.

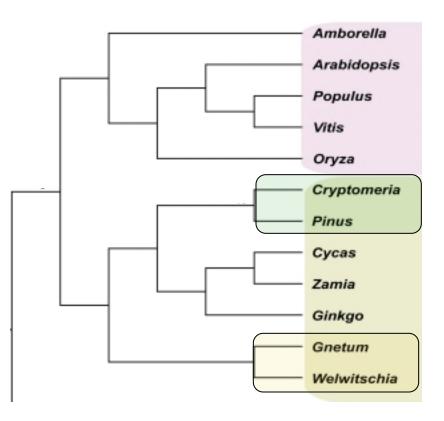
<mark>ộ</mark> Enter Query	sequence
Enter FASTA sequence	re
clear 🛛 Or,	upload file: Choose File no file selected
<mark>ಿ</mark> Basic Param	
Blast Tools	blastn 🔷
DataBase	Class: 1059-All_Sample 🗘 Sample: All_Sample 🛟

Surprises from Phylogenomics: Monophyly of Conifers?

Plastid Genome Analyses

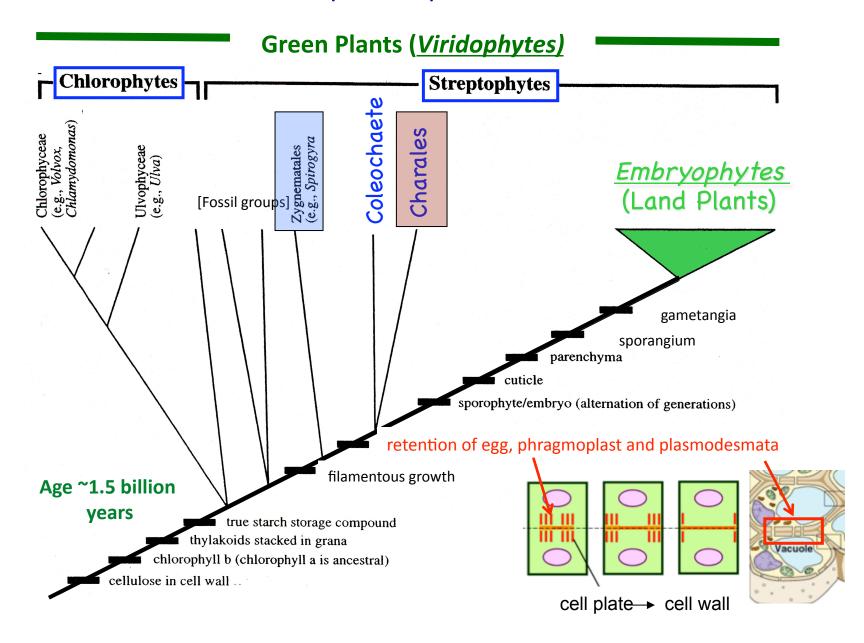


Large Nuclear Gene Analyses

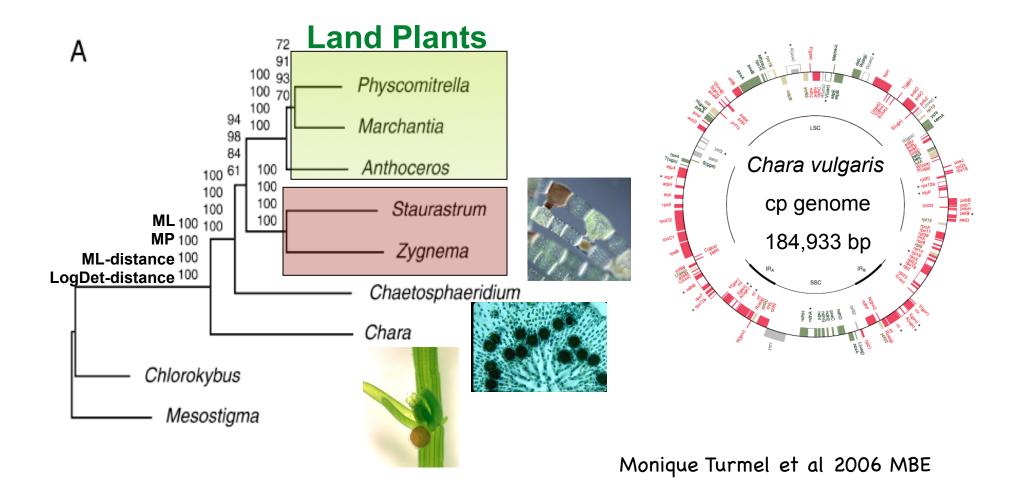


Cibrián et al. 2010, Lee et al 2011, OneKp unpublished

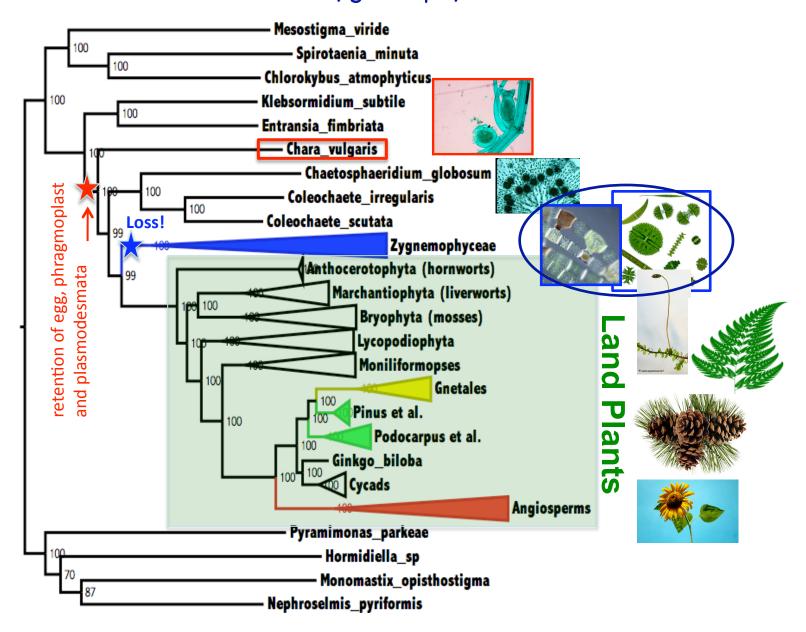
Origin of Land Plants: Retention of egg, phragmoplast and plasmodesmata preadaptations for colonization of land



Plastid phylogenomic analyses yielding surprising results for the origin of land plants – filamentous rather than more complex algal lineages inferred sister to Land Plants!

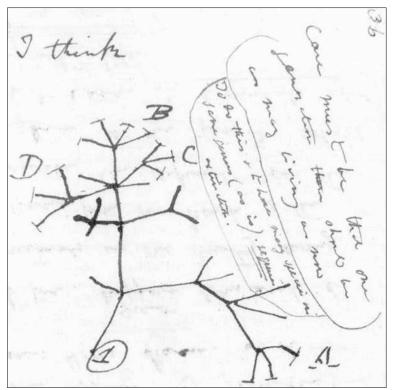


Analysis of 604 single-copy nuclear genes agrees with plastome analysis suggesting that retention of egg, plasmodesmata, and phragmoplast lost in diverse Zygnemophyceae



Why should you believe these results?

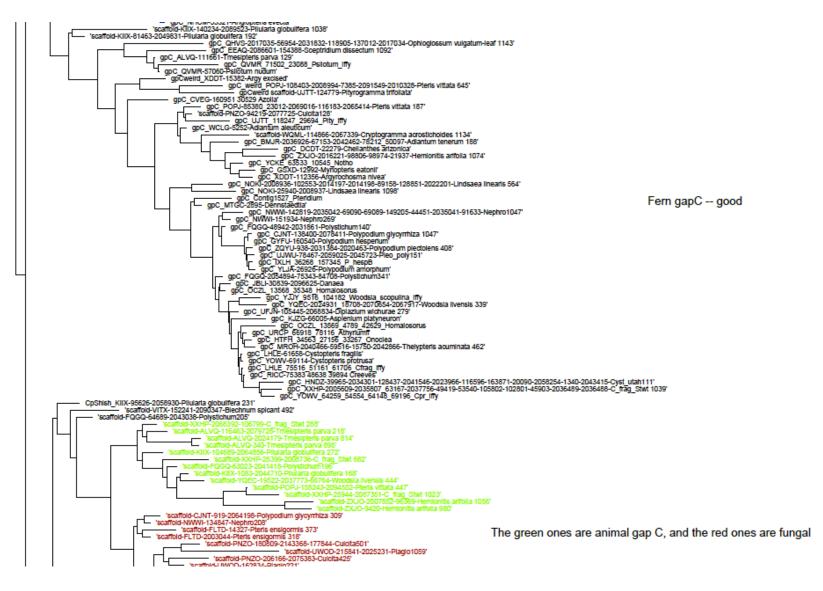
 You shouldn't.... Inferred trees are hypotheses!



Consider Possible Complications

- Lots of missing data!
 - Remove genes/taxa with lots of missing data
- Ortholog identification
 - Focus on genes/genomes that tend not to be duplicated
- Contamination
 - BLAST, Sequence placement on reference tree (SEPP?), Long branch trimming
- Model misspecification
 - Model validation simulations
 - Test robustness of hypothesis among inferences derived from alternative models

Contamination is an issue!



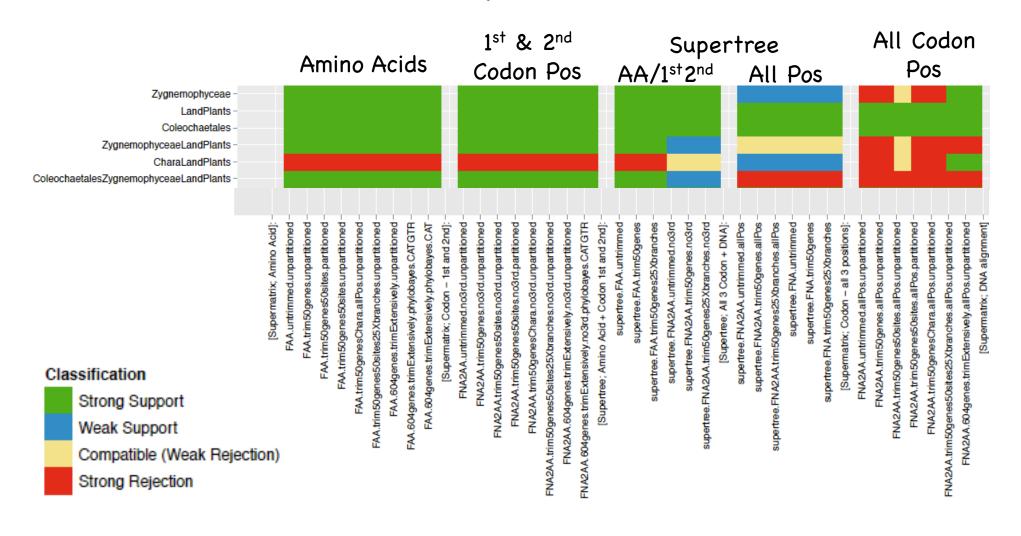
Data Matrix/Analysis Perturbation

- DNA vs Amino Acids
- All codon positions vs remove third position
- Remove gappy genes or sites
- Remove genes on long branches in gene trees
- Remove genes where contamination seems to persist.
- Supermatrix and Supertree estimation

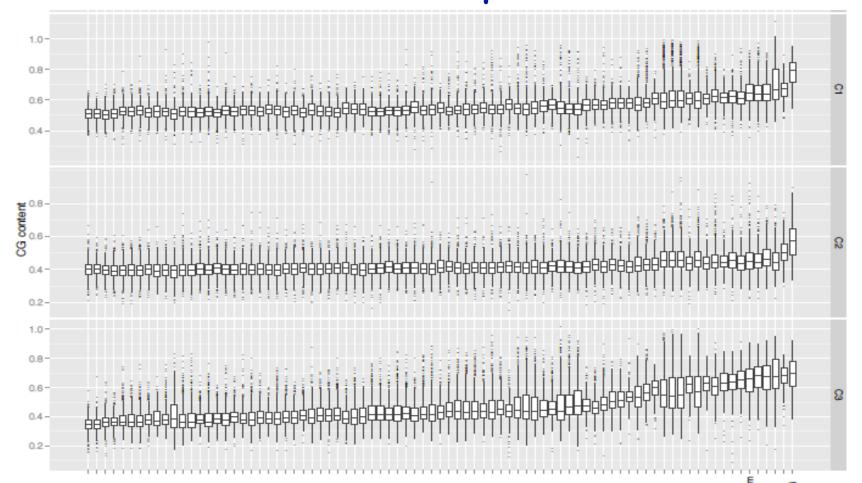
Good News: Some results seem to be very robust

	AngioSperms -																																							
AngioSpe	rmsNoAmborela -																																							
AngioSpermsNo/	AmborelaNuphar -																																							
		[Supermatrix; Amino Acid]: -	FAA.untrimmed.unpartitioned	FAA.trim50genes.unpartitioned - FAA trim50nenes50sites nartitioned -	FAA.trim50genes50sites.unpartitioned -	.trim50genesChara.allPos.unpartitioned -	-AA.trim50genes50sites25Xbranches.unpartitioned -	FAA.ou4genes.trimExtensively.unpartitioned = 604.eenes trimExtensively ohvlobaves CATGTB =	FAA.604genes.trimExtensively.phylobayes.CAT	[Supermatrix; Codon - 1st and 2nd]: -	FNA2AA.untrimmed.no3rd.unpartitioned -	FNA2AA.trim50genes.no3rd.unpartitioned - A trim50renes50sites po3rd unpartitioned -	ENA2AA trim50geressosies.nooru.unparunored - ENA2AA trim50genes50sites.no3rd.partitioned -	FNA2AA. trim50genesChara. no3rd, unpartitioned -	A.trim50genes50sites25Xbranches.no3rd.unpartitioned -	FNA2AA.604genes.trimExtensively.no3rd.unpartitioned	.no3rd.phylobayes	[Supertree; Amino Acid + Codon 1st and 2nd]: -	supertree.FAA.untrimmed - supertree.FAA.trim50genes -	supertree. FAA. trim50genes25 Xbranches -	supertree. FNA2AA.untrimmed.no3rd -	supertree.FNA2AA.trim50genes.no3rd -	s25Xbranches	[Supertree; All 3 Codon + DNA]: -	supertree.FNAZAA.untrimmed.aliPos - supertree FNA2AA trim50nenes allPos -	supertree FNA2AA.trim50genes25Xbranches.allPos -	supertree.FNA.untrimmed -	supertree.FNA.trim50genes -	supertree.FNA. trim50genes25Xbranches -	[Supermatrix; Codon - all 3 positions]: -	FNA2AA.untrimmed.allPos.unpartitioned	FNA2AA.trim50genes.allPos.unpartitioned -	-NA2AA.trim50genes50sites.allPos.unpartitioned	FNA2AA.trim50genes50sites.allPos.partitioned	.trim50genesChara.allPos.unpartitioned -	A.trim50genes50sites25 Xbranches.allPos.unpartitioned	OS.UN	ENA trim50cenes50sites unpartitioned -	FNA.trim50genes50sites25Xbranches.unpartitioned	
Classificat	ion					FAA.t	n50ge	TAA.0	A.604					2AA.1	nes50	04gen	trimE	npert		ns			e.FN			e.FN			ิส		Ē	Ĩ.	AA.tr	A2AA	FNA2AA.t	les50	14gen		n50ge	
Strong	Support						AA.trir	- FAA.604				ENADA		ENA 1	n50ge.	2AA.6(genes	S					upertir			Ipertre	-						ENAS	Ē	ENA 2	150gel	SAA.00		NA.trir	
Weak	Support						Ш.	Ē							AA.trin	FNA	VA. 604						3			SL										AA.trin	FNAZ		Œ	
Compatible (Weak Rejection)																														FNAZ										
Strong	Rejection																-																							

Issues with inclusion of 3rd codon position?

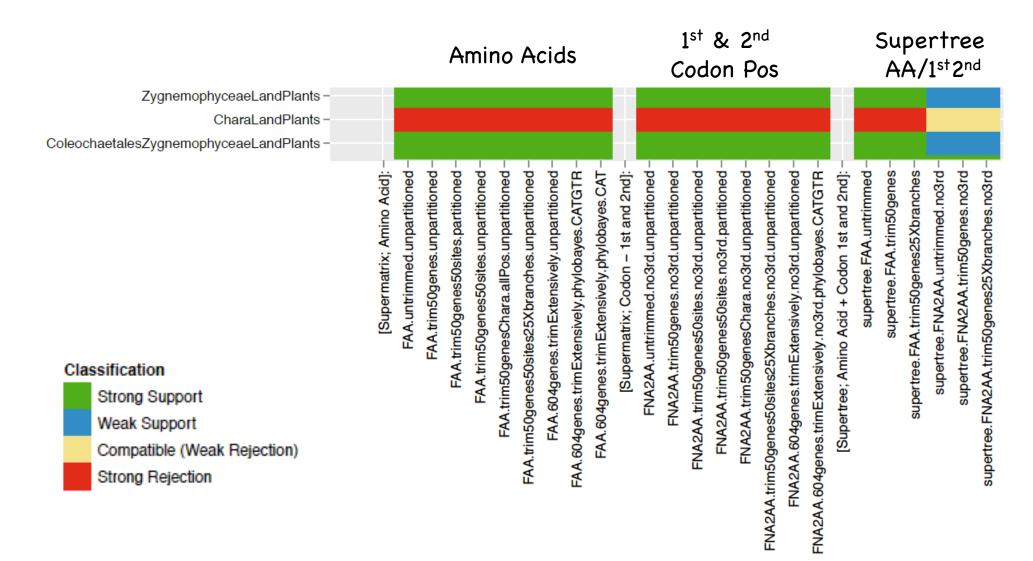


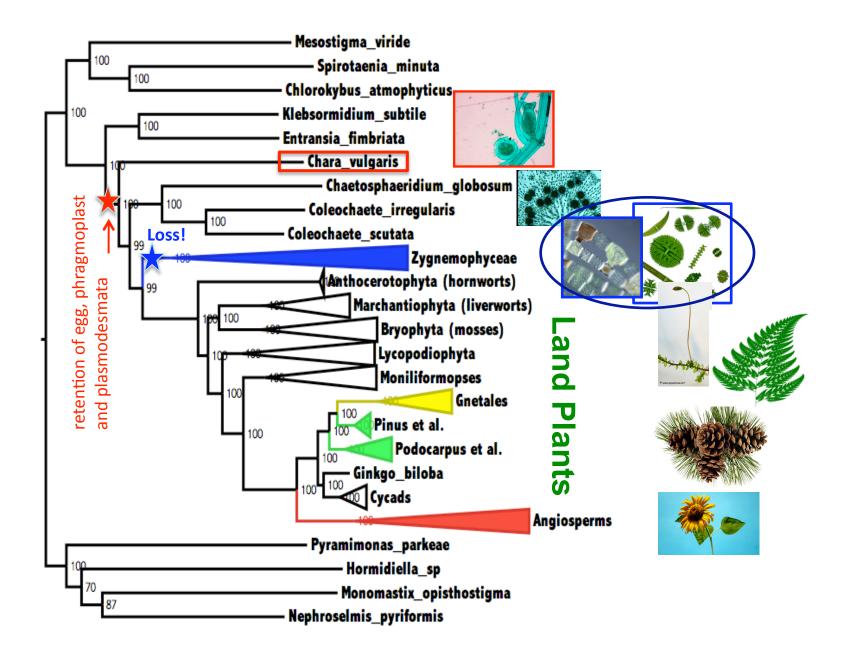
Problems due to heterogeneous substitution process?



Agristication in the second second and a second second and an anticology second a second and an anticology second and an anticology second and an anticology and a second and any anticology and a second and an anticology and a second and any anticology and a second and any anticology and an anticology and and any anticology and an anticology and an anticology and and any anticology and any anticology and and any anticology and and an

Filamentous Algae + Land Plant Hypothesis Seems to be Robust

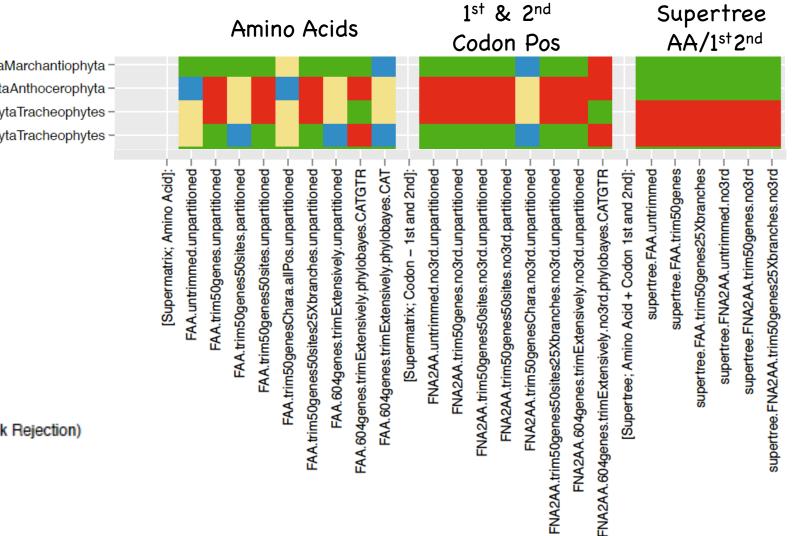




Supertree and Supermatrix Analyses Provide Different Inferences Concerning the Monophyly of Conifers

	A	mir	וסו	Ac	ids						2 ⁿ 1 Pa		Supertree AA/1 st 2 nd									
GnetalesPinaceae -																						
Coniferales -																						
ConiferalesGnetales -																						
Classification Strong Support Weak Support Compatible (Weak Rejection) Strong Rejection	[Supermatrix; Amino Acid]: - FAA.untrimmed.unpartitioned	FAA.trim50genes.unpartitioned -	FAA.trim50genes50sites.partitioned – FAA.trim50genes50sites.unpartitioned –	FAA.trim50genesChara.allPos.unpartitioned -	FAA.trim50genes50sites25Xbranches.unpartitioned -	FAA.604genes.trimExtensively.unpartitioned	FAA.604genes.trimExtensively.phylobayes.CATGTR -	FAA.604genes.trimExtensively.phylobayes.CAT -	[Supermatrix; Codon – 1st and 2nd]: – FNA2AA.untrimmed.no3rd unpartitioned –	FNA2AA.trim50genes.no3rd.unpartitioned	FNA2AA.trim50genes50sites.no3rd.unpartitioned	FNA2AA.trim50genes50sites.no3rd.partitioned -	FNA2AA.trim50genesChara.no3rd.unpartitioned - ENA2AA trim50genes50sites25Xhranches.no3rd.unpartitioned -	FNA2AA.604genes.trimExtensively.no3rd.unpartitioned -	FNA2AA.604genes.trimExtensively.no3rd.phylobayes.CATGTR -	[Supertree; Amino Acid + Codon 1st and 2nd]: -	supertree.FAA.untrimmed	supertree.FAA.trim50genes -	supertree.FAA.trim50genes25Xbranches -	supertree.FNA2AA.untrimmed.no3rd	supertree.FNA2AA.trim50genes.no3rd -	supertree.FNA2AA.trim50genes25Xbranches.no3rd -

Mosses Sister + Liverworts Clade Sister to Vascular Plants?



BryophytaMarchantiophyta – BryophytaMarchantiophytaAnthocerophyta – AnthocerophytaTracheophytes – BryophytaMarchantiophytaTracheophytes –

Classification

Strong Support Weak Support Compatible (Weak Rejection) Strong Rejection

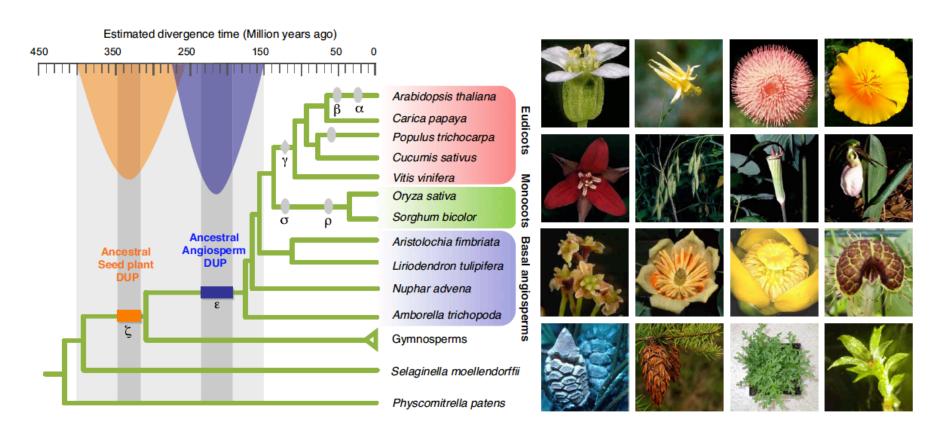
Conclusions

- Genome-scale phylogenetic inference is complex!
- Model mis-specification can results in statistical inconsistency: more data -> stronger support for the wrong answer.
- Proposal: Apply multiple analysis strategies and explore/understand basis of conflicts among resulting trees

Phylogenomics???

- 1. Using genome-scale data to resolve phylogentic relationships
- 2. Genome-Scale comparisons placed in a phylogenetic context

Phylogenetic Analysis and Molecular Dating of 100's of Gene Families Implicates Genome Duplications Associated with Origin of Angiosperms and Seed Plants



Jiao et al. 2011

Thank You!