PlantTribes: Galaxy tools for comparative gene family analysis in plant genomics

Eric Wafula, Greg Von Kuster, Joshua Der, Loren Honaas, Saravanaraj Ayyampalayam, Norman Wickett, Jim Leebens-Mack, and Claude dePamphilis.

January 14, 2019 Galaxy Meeting, PAG

Motivation

- Massive genome and transcriptome data for non-model organisms
- Existing gene family resources are static
- A scalable global gene family framework is needed

Objectives

- Generate objective gene family ("orthogroup") classifications ("scaffolds") that leverage diverse genomes
- Optimize circumscription of gene families from both evolutionary and functional perspectives
- Develop modular set of analysis tools for comparative phylogenomics of new datasets
- Make these tools broadly available as a stand-alone package (GitHub) and through the Galaxy framework

Scaffold construction



selected representative proteomes from clades of interest

primary gene family circumscription

orthogroups, paralogous clusters, and singletons

Scaffold construction

Data

- Orthogroups protein and cds fasta
- Orthogroups protein multiple sequence alignment fasta
- Orthogroup hidden Markov model based profiles (HMM-profiles)
- Scaffold database of profile HMMprofiles
- BLAST database of all scaffold proteins

Metadata

- UniProtKB/Swiss-Prot protein functional annotations
- TAIR Arabidopsis thaliana gene functional annotations
- InterPro and Pfam functional domain annotations
- Gene Ontology functional terms
- Gene family sizes (orthogroup gene counts) for scaffold taxa
- Secondary gene family clusters (super-orthogroups)

Scaffold data usage

Genomes

Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.)

ay Ming †🕋, Robert VanBuren [†], Yanling Liu [†], Mei Yang [†], Yuepeng Han , Lei-Ting Li , Qiong Zhang , Min-Jeong Kim , lichael C Schatz, Michael Campbell, Jingping Li, John E Bowers, Haibao Tang, Eric Lyons, Ann A Ferguson, Giusepp avid R Nelson, Crysten E Blaby-Haas, Andrea R Gschwend, Yuannian Jiao, Joshua P Der, Fanchang Zeng, Jennifer Han, lang Jia Min, Karen A Hudson, Ratnesh Singh, Aleel K Grennan, Steven J Karpowicz, Jennifer R Watling, Kikukatsu Ito

haron A Robin Fern genomes elucidate land plant evolution and ancy Chen cvanobacterial symbioses

tephen R Do ^{4ark Yandell, 4} Fay-Wei Li^{12*}, Paul Brouwer³, Lorenzo Carretero-Paulet^{4,5}, Shifeng Cheng⁶, Jan de Vries⁰7, ane Shen-Mille Pierre-Marc Delaux^e, Ariana Eily^e, Nils Koppers¹⁰, Li-Yaung Kuo⁽³⁾, Zheng Li¹¹, Mathew Simenc¹², lan Small^{©13}, Eric Wafula¹⁴, Stephany Angarita¹², Michael S. Barker^{©11}, Andrea Bräutigam^{©15}, Claude dePamphilis¹⁴, Sven Gould^{®16}, Prashant S. Hosmani¹, Yao-Moan Huane¹⁷, Bruno Huettel¹¹⁰, Yoichiro Kato¹⁹, Xin Liu⁰⁶, Steven Maere^{04,5}, Rose McDowell¹³, Lukas A. Mueller¹, Klaas G. J. Nierop²⁰, Stefan A. Rensing⁽⁾²¹, Tanner Robison⁽⁾²², Carl J. Rothfels⁽⁾²³, Erin M. Sigel²⁴ Yue Song⁶, Prakash R. Timilsena¹⁴, Yves Van de Peer^{(04,5,25}, Hongli Wang⁶, Per K. I. Wilhelm Paul G. Wolf²², Xun Xu⁶, Joshua P. Der¹², Henriette Schluepmann³, Gane K.-S. Wong^{6,26} and K

The Amborella Genome and the Evolution of Flowering Plants

Amborella Genome Project*,†

Gene and genome duplications



Patrick P. Edger^{a,b,c,1}, Hanna M. Heidel-Fischer^{d,1}, Michaël Bekaert[®], Jadranka Rota^f, Gernot Glöckner^{g,h}, Adrian E. Plattsⁱ, Patrick Y. Edger^{22,1}, Islanna M. Heidel-Hischer^{22,1}, Michael Bekaer²⁷, Jadramka Kota, Gernot Glococne^{27,2}, Adriank z. Patris David G. Hecke¹, Joshua P. De²⁷, Eric K. Wafald, Michelle Tang², Johannes A. Holberger², Ann Smithson^{22,2}, Jocelyn C. Hall⁹, Matthieu Blanchette¹, Thomas E. Bureau², Stephen I. Wright², Claude W. defamphillis, M. Eric Schranz Michael S. Barke^{27,4}, Gavin C. Conner^{47,}, Nikka Wahlberg², Helko Viege^{27,1}, J. Chris Pres^{28,4}, and Christopher W. Whea²⁷



Ecological genomics of tropical trees: how local population size and allelic diversity of resistance genes relate to immune responses, cosusceptibility to pathogens, and negative density dependence

J. H. MARDEN,*† 🌔 S. A. MANGAN,‡ § M. P. PETERSON,*† E. WAFULA,* H. W. FESCEMYER,

P. DER, T. W. DEPAMPHIN Horizontal gene transfer is more frequent with increased heterotrophy and contributes to parasite adaptation

² Zhenzhen Yang^{ab.c.1}, Yeting Zhang^{b.c.d.1,3}, Eric K. Wafula^{b.c.}, Loren A. Honaas^{ab.c.3}, Paula E. Ralph⁹, Sam Jones^{ab}, Christopher R. Clarke⁶, Siming Liu⁷, Chun Su⁹, Huiting Zhang³⁰, Naomi S. Altman¹⁰, Stephan C. Schuster¹¹, Michael P. Timko¹, John I. Ydoer¹, James H. Westwood⁹, and Claude W. dePamphilis^{8,b.c.d.1}

2008 –NAR– 99 citations

Gene family studies

PlantTribes: a gene and gene family resource for comparative genomics in plants

P. Kerr Wall¹, Jim Leebens-Mack^{1,2}, Kai F. Müller^{1,3}, Dawn Field⁴, Naomi S. Altman⁵ and Claude W. dePamphilis^{1,*}

http://bigdata.bx.psu.edu/PlantTribes scaffolds/

Expression shifts

Comparative Transcriptome Analyses Reveal Core Parasitism Genes and Suggest Gene Duplication and Repurposing as Sources of Structural Novelty 👌 Zhenzhen Yang, Eric K. Wafula, Loren A. Honaas, Huiting Zhang, Malay Das, Monica Fernandez-Aparicio,

Kan Huang, Pradeepa C.G. Bandaranayake, Biao Wu, Joshua P. Der, Christopher R. Clarke, Paula E. Ralph, Lena Landherr, Naomi S. Altman, Michael P. Timko, John I. Yoder, James H. Westwood, Claude W. dePamphilis 📼

Functional genomics of a generalist parasitic plant: Laser microdissection of host-parasite interface reveals host-specific patterns of parasite gene expression

Transcriptomes of the Parasitic Plant Family Orobanchaceae Reveal Surprising Conservation of Chlorophyll Synthesis orman J. Wickett ^{1, 5} 条 四, Loren A. Honaas ¹, Eric K. Wafula ¹, Malay Das ^{2, 6}, Kan luang ³, Biao Wu ⁴, Lena Landherr ¹, Michael P. Timko ³, John Yoder ⁴, James H.

Loren A Honaas¹, Eric K Wafula², Zhenzhen Yang¹, Joshua P Der¹², Norman J Wickett¹²⁸, Naomi S Altman Westwood ², Claude W. dePamphilis ¹ A 🗃 Christopher G Taylor⁴, John I Yoder², Michael P Timko⁴, James H Westwood² and Claude W dePamphilis¹²

PlantTribes workflow



Scaffold installation

PlantTribes Scaffolds Download (Galaxy Version 1.1.0)	 Options
Data table entry unique ID	
26Gv2.0	
Description of the data	
26 genome land plants (Embryopyta) gene family scaffold	
Value is optional	
URL for downloading scaffolds	
http://bigdata.bx.psu.edu/PlantTribes_scaffolds/data/26Gv2.0.tar.bz2	
Must be same version as configs	
URL for downloading configs	
http://bigdata.bx.psu.edu/PlantTribes_scaffolds/configs/26Gv2.0.tar.gz	
Must be same version as scaffolds	
✓ Execute	

Installing gene family scaffold

http://bigdata.bx.psu.edu/PlantTribes_scaffolds/

https://github.com/dePamphilis/PlantTribes

Update PlantTribes scaffold with a new genome ((Galaxy Version 1.0.0.0)	▼ Options
Gene family scaffold		
26Gv2.0		•
Proteins fasta file		
1: Cuscuta_campestris.faa	cds and corresponding proteins	•
Coding sequences fasta file	for the new genome	
1: Cuscuta_campestris.faa	for the new genome	•
Species name		
Cuscuta campestris		
Species code		
Cusca_v1.0		
Species family		
Convolvulaceae		
Species order		
Solanales		
Species group]
Asterids		
Species clade		
Core Eudicate		
Species code for rooting order		
Cusca The new species above will be placed immediately a	ofter this species code in the rooting order configuration file	
Job Resource Parameters	area and species code in the rooting order configuration me	
Use default job resource parameters		•
✓ Execute		

Updating Installed gene family scaffold with a new sequenced genome

Analysis tools

- 1. Post-assembly QC
- 2. Assembly sorting
- 3. Alignment and QC
- **4.** Phylogenetic inference
- 5. Selection and WGD analysis

1. Post-assembly QC Assembly post-processing

AssemblyPostProcessor post-processes de novo transcriptome assembly (Galaxy Version 0.8.0)	options	History	3 ✿ □
Transcriptome assembly fasta file Image: Constraint of the system Image: Constraint of the system </th <th>→]</th> <th>search datasets</th> <th>8</th>	→]	search datasets	8
Coding regions prediction method		12 shown	
TransDecoder	•	15.31 MB	
Options configuration Advanced	•	<u>12: transcripts.cds: Asse</u> mblyPostProcessor on d ata 1	• / ×
Perform targeted gene assembly? primary coding No regions predictions	•	11: transcripts.cleaned.c ds: AssemblyPostProces	• / ×
Strand-specific assembly?		<u>sor on data 1</u>	
No		10: transcripts.cleaned.n r.cds: AssemblyPostProc	
Remove duplicate sequences? validated, filtered		essor on data 1	
Yes coding regions prediction		9: transcripts.cleaned.nr.	• * ×
Minimum sequence length		pep: AssemblyPostProce ssor on data 1	
200		8: transcripts.cleaned.pe	@ # ¥
Job Resource Parameters validated and filtered		p: AssemblyPostProcess or on data 1	C D W
Use default job resource parameters coding regions predictions	_	7: transcripts.pep: Asse	🕘 🖋 🗙
✓ Execute		<u>mblyPostProcessor on d</u> <u>ata 1</u>	

1. Post-assembly QC Targeted gene family assembly

AssemblyPostProcessor post-processes de n	ovo transcriptome assembly (Galaxy Version 0.8.0)	- Options		History	C 🕈 🗆
Transcriptome assembly fasta file					
1: assembly.fasta	de novo transcriptome assembly	•	N .	search datasets	8
Coding regions prediction method	·			PlantTribes test data	
TransDecoder		•		19 shown	
Options configuration			r	15.86 MB	S D
Advanced		•	_		
Perform targeted gene assembly?	primary coding			19: transcripts.cds: Asse	(*)
Yes	regions predictions	_	► ►	ata 6 and data 1	
Targeted gene families			L		
6: targetOrthos.ids	the second s	\ -]		18: transcripts.cleaned.c	(*)
Gene family scaffold	targeted gene family assembly			sor on data 6 and data 1	
22Gv1.1		-			
Protein clustering method				17: transcripts.cleaned.n	• 🖋 🗙
OrthoMCL	and idea down different	-	\backslash	essor on data 6 and data	1
Trim alignments	validated and filtered				
0.1		s	X	16: transcripts.cleaned.n	
Strand-specific assembly?				essor on data 6 and data	1
No		•		1	
Remove duplicate sequences?	validated and filteror	4		15: transcripts.cleaned.p	
Yes				sor on data 6 and data 1	
Minimum sequence length				14 transcripts pape Acco	
200				mblvPostProcessor on d	
Job Resource Parameters				ata 6 and data 1	
Use default job resource parameters	a collection of post process	- h		13: Targeted gene famili	
	a conection of post-processe Targeted going family assembly		► ►	es: AssemblyPostProcess	
✓ Execute	largeted gene family assemble	lies		or on data 6 and data 1	

1. Post-assembly QC Assembly improvement



% of 1,440 universal land plants single-copy orthologs

Assemblies from Parasite Plant Genome Project (PPGP) (Orobancheceae family)



1. Post-assembly QC Tool documentation



2. Assembly sorting Gene family classification

GeneFamilyClassifier classifies gene sequences into pre-computed orthologous gene family clusters (Galaxy Version 0.8.0)		search datasets (8)	
Proteins fasta file			
Image: Construction of the second		PlantTribes test data	
Gene family scaffold		21 shown	History C & T
22Gv1.1 •	\neg	19 33 MB	
Protein clustering method			Rack to PlantTribes test data
OrthoMCL		21: GeneFamilyClassifier 🛛 🝙 💉 🖌	Dack to Hantmbes test data
Protein classifier a collection of classified		(gene family clusters) on	GeneFamilyClassifier on data 10 and
Both blastp and hmmscan Orthogroup protein and cds fasta		data 10 and data 9	data 9
Save hmmscan log?			a list of datasets
No		20: GeneFamilyClassifier on dat	A did to an
Options configuration		a 10 and data 9	Add tags
Advanced		a list of datasets	
Super orthogroups configuration			proteins.blastp.22Gv1.1 ()
No			
Single copy orthogroups configuration			proteins.blastp.22Gv1.1.best
No		Classification metadata	Orthos
Orthogroups fasta configuration	• b	lasto results	proteins.both.22Gv1.1.best0
Yes	• h	mmscan results	rthos
Orthogroups coding sequences	• b	est scoring orthogroup (blast) —	
Yes	• b	est scoring orthogroup (hmm)	proteins.both.22Gv1.1.bestO 💿 🖋
Coding sequences fasta file post-assembly predicted cds	• \$	elected best scoring orthogroup	rthos.summary
10: transcripts.cleaned.nr.cds: AssemblyPostPro	• 0	rthogroup annotation summary	proteins hmmscan 22Gv1 1
Job Resource Parameters		· · · · · · · · · · · · · · · · · · ·	
Use default job resource parameters			proteins hmmscan 22Gv1 1
✓ Execute			bestOrthos
			<u>wester (1105</u>

2. Assembly sorting Advanced analyses

GeneFamilyClassifier classifies gene sequences into pre-computed orthologous gene family clusters	Options	GeneFamilyClassifier classifies gene sequences into pre-computed orthologous gene family clusters (Galaxy Version 1.0.3.0)
(Galaxy Version 1.0.3.0)		Proteins fasta file
Proteins fasta file		□ 4 □ 1437: assembly.fasta
ப 42 D 1437: assembly.fasta	•	Gene family scaffold
Gene family scaffold		12Gv1.0
12Gv1.0	•	arthamel
Protein clustering method		Protein classifier
orthomcl	•	• blastp
Protein classifier		Options configuration
blastp	•	Advanced 🗸
Options configuration		Super orthogroups configuration
Advanced	-	No *
Super orthogroups configuration		Single copy orthogroups configuration
Yes super-orthogroups clusters	•	selection of low-copy orthogroups
Clustering distance measure (dofault min_ovalue)		Selection criterion (no default)
average e-value	•	
Si	٩	Global selection
N minimum e-value		Custom selection
Or average e-value		1438: Arabidopsis_thaliana.smat
No		Orthogroups fasta configuration
Job Resource Parameters		No
Use default job resource parameters	•	Job Kesource Parameters
✓ Execute		✓ Execute

2. Assembly sorting Selection low-copy gene families

GeneFamilyClassifier classifies gene sequences into pre-computed orthologous gene family clusters (Galaxy Version 1.0.3.0)	▼ Options
Proteins fasta file	
🖸 🕫 🗀 1437: assembly.fasta	•]
Gene family scaffold	
12Gv1.0	•
Protein clustering method	
orthomcl	•
Protein classifier	
blastp	•
Options configuration	
Advanced	•
Super orthogroups configuration	
No	•
Single copy orthogroups configuration	
Yes	•
Selection criterion	
Global selection	•
Minimum single copy taxa	
global selection	
Minimum taxa present	
Orthogroups fasta configuration	
No	•
lob Pasource Parameters	
lice default ich recourse parameters	
ose delaur jou resource parameters]
✓ Execute	

GeneFamilyClassifier classifies gene sequences into pre-computed of	orthologous gene family clusters (Galaxy Version 1.0.3.0)	▼ Options
Proteins fasta file		
🗅 🕙 🗀 1437: assembly.fasta		•
Gene family scaffold		
22Gv1.1		•
Protein clustering method		
orthomcl		•
Protein classifier		
blastp		•
Options configuration		
Advanced		•
Super orthogroups configuration		
No		•
Single copy orthogroups configuration		
Yes		•
Selection criterion		
Custom selection		-
Custom selection configuration	- custom soloction -	
Yes	custom selection	•
Custom selection file		
1439: 22Gv1.1.singleCopy.config		•
Orthogroups fasta configuration		
No		-
Job Resource Parameters		
Use default job resource parameters		•
✓ Execute		

2. Assembly sorting Gene family integration

GeneFamilyIntegrator integrates gene models in pre-computed orthologous gene family clusters with classified gene coding		History	€ 🕈 🗆
sequences (Galaxy Version 0.8.0)		<u> </u>	
Classified orthogroup fasta files		search datasets	8
Protein and coding sequences orthogroup fasta files		PlantTribes test data	
Protein and coding sequences orthogroup fasta files a collection of classified		22 shown, 6 <u>hidden</u>	
C 21: GeneFamilyClassifier (gene family clusters) on data 10 and data 9 Orthogroup protein and cds fasta	Ľ	44.66 MB	
Gene family scaffold		28: GeneFamilyIntegrato	
22Gv1.1		r (integrated gene family	
Protein clustering method		clusters) on data 21	
OrthoMCL		Î	
Orthogroups coding sequences			
Yes	a collectio	on merged classified an	d scaffold
Job Resource Parameters	ortho	ogroup protein and cds	fasta
Use default job resource parameters			
✓ Execute			

3. Alignment and QC Gene family alignments

GeneFamilyAligner aligns integrated orthologous gene famil	y clusters (Galaxy Version 1.0.3.0)		▼ Options		History	C 🌣 🗆
Integrated orthogroup fasta files						
1443: GeneFamilyIntegrator (integrated gene family clusters) o	n data 1436, data 1435, and others		•		search datasets	8
Multiple sequence alignment method	a collection merrod a	lessified and soffeld				
MAFFT	a collection merged c	assined and scanold	•		PlantTribes test	data
	or mogroup proc		٩		39 shown, 9 <u>deleted</u> , 2	233 <u>hidden</u>
MAFFT					113 39 MB	
PASTA					115.55 MB	
Yes			*		1048: GeneFamilvA	igner (trim 🛛 🖌
Trimming method					med orthogroup co	don alignme
Gap score based trimming			-		nts) on data 1046. c	lata 1045, and ot
Gap score	gap trimming				hors	<u>utu 1015, utu 01</u>
0.1					a list with 164 items	
Remove sequences					1047: ConoEamilyA	ignor (trim
No			•		med orthogroup pr	atoin alignm
Output primary and intermediate alignments?					ants) on data 1046	data 1045 and o
⊘ No					thore	<u>uata 1045, anu 0</u>
OYes					a list with 164 itoms	
In addition to trimmed/filtered alignments					a list with 104 items	
Job Resource Parameters					^	
Use default job resource parameters			•			
✓ Execute				colle	ection of estimated of protein multiple sequ	rthogroup codon and ence alignments

3. Alignment and QC Alignment post-processing

GeneFamilyAligner aligns integrated orthologous gene family clusters (Galaxy Version 1.0.3.0)	▼ Options
Integrated orthogroup fasta files	
No fasta dataset collection available.	•
Multiple sequence alignment method	
MAFFT	-
Codon alignments	
Yes	-
Alignment post-processing configuration	
Yes	-
Trimming method	
Gap score based trimming automated trimming recommend	-
for near full-length transcripts	٩
Gap score based trimming	
Automated heuristic trimming	
No	•
Output primary and intermediate alignments?	
© №	
OYes	
In addition to trimmed/filtered alignments	
Job Resource Parameters	
Use default job resource parameters	•
✓ Execute	

eneFamilyAligner aligns integrated orthologous g	gene family clusters (Galaxy Version 1.0.3.0)	▼ Options
tegrated orthogroup fasta files		
No fasta dataset collection available.		•
ultiple sequence alignment method		
/AFFT		•
odon alignments		
/es		-
lignment post-processing configuration		
/es		•
Trimming method		
Gap score based trimming		-
Gap score		
0.1	iterative trimming.	
D	filtoring and realignment	
kemove sequences	intering, and realignment	
Yes		
Coverage score		
0.5		
Realignment iteration limit		
Output primary and intermediate alignments?		
ON0		
© Yes		
In addition to trimmed/filtered alignments		
b Resource Parameters		
Ise default job resource parameters		-

3. Alignment and QC Alignment visualization

•



collection of estimated orthogroup protein multiple sequence alignments

History	€ ♥ 🗆				
< <u>Back to PlantTribes test data</u> GeneFamilyAligner (trimmed orthogroup protein alignments) on data 1046, data 1045, and others					
Add tags					
<u>10002.faa.aln</u>	۲				
<u>10350.faa.aln</u>	۲				
21 sequences format: fasta , database: <u>j</u>	2				
🖹 🛈 🎜 📖 ?	>				
display with IGV <u>local</u>					
>gnl_Glyma1.01_PACid_16288422					
YSYDEAEDEAFRRAFGQRDDGDDSSSNPK					
LHEELDPLIAEKKNPYDFPRDPEEWTEQDLREFWADGP					
EQIADGEEPPIAPFYLPYRKHYPPIPDNHYDIATPKDA					
<u>10742.faa.aln</u>	۲				

10783.faa.aln





4. Phylogenetic inference Gene family phylogenetic trees

GeneFamilyPhylogenyBuilder builds phylogenetic trees of aligned orth	ologous gene family clusters (Galaxy Version 1.0.3.1)	 Options 		History	C 🌣 🗆
Proteins orthogroup alignments					
1048: GeneFamilyAligner (trimmed orthogroup codon alignments) on da	a 1046, data 1045, and others	•		search datasets	8
Phylogenetic inference method			Ν	Search databets	
RAxML		•		PlantTribes test data	
Gene family scaffold				24 shown, 8 deleted, 1866 hig	lden
22Gv1.1		•		<u> </u>	
Protein clustering method				83.55 MB	
orthomcl	ML orthogroup phylogenetic	-			
Rooting order configuration	inference with RAxML			1530: GeneFamilyPhyloge	nyBuil 🗙
Yes		-		der (phylogenetic trees) o	n dat
Rooting order configuration				a 2. data 1376. and others	5
2: 22Gv1.1.rootingOrder.config		•		a list with 150 items	-
Bootstrap replicates					
100				↑	
Maximum orthogroup size					
100					
Minimum orthogroup size					
4				collection of inferred or	thogroup
Job Resource Parameters				Phylogenetic tre	es
Use default job resource parameters		•			
✓ Execute					

4. Phylogenetic inference Phylogenetic tree visualization

History	C 🌣 🗆			
search datasets	8			
PlantTribes test data 24 shown, 8 <u>deleted</u> , 1866 <u>hidd</u>	en			
83.55 MB				
1530: GeneFamilyPhylogenyBuil der (phylogenetic trees) on dat <u>a 2, data 1376, and others</u> a list with 150 items				

collection of inferred orthogroup Phylogenetic trees

History	€ ✿ 🗆	
Seck to PlantTribes test of GeneFamilyPhylogenyBuil (phylogenetic trees) on de 1376, and others a list with 150 items	data Ider ata 2, data	
Add tags		
10002.fna.aln.tree	۲	
10350.fna.aln.tree	۲	
data format: newick , database:	2	
8 8 2 11 ?	۲	
10742.fna.aln.tree	۲	
10783.fna.aln.tree	۲	



5. Selection and WGD analysis Nucleotide substitution rates

KaKsAnalysis estimates paralogous and orthologous pairwise (Galaxy Version 1.0.3.0)	synonymous (Ks) and non-synonymous (Ka) subs	titution rates		History	3 ♥ □
Coding sequences for the first species		•	N	search datasets	8
Protein sequences for the first species 5: species1.faa Cds and corr for the	esponding proteins			PlantTribes test data 31 shown, 9 <u>deleted</u> , 2233 <u>his</u>	dden
Type of sequence comparison			V	109.76 MB	S D
Coding sequences for the second species		species1 non-synon	ymous	2272: KaKsAnalysis (KaK	● / ×
Protein sequences for the second species for the	rresponding proteins	(Ka) and synonymou substitution rat	us (<i>Ks</i>) es	and data 6	
C C 7: species2.faa Determine for cross-species orthologs using		species1 paralogous	s pairs>	2271: KaKsAnalysis (par alogous pairs) on data 5	• / ×
 ○ reciprocal best BLAST ● conditional reciprocal best BLAST 				and data 6	
Options Configuration		species1 self blastn	results →	2270: KaKsAnalysis (bla stn results species1 vs s	• / ×
I certify that I am not using this tool for commercial purposes.]		pecies1) on data 5 and da	<u>ata 6</u>
Job Resource Parameters					
Use default job resource parameters					
✓ Execute					

5. Selection and WGD analysis Genome duplication inference

KaKsAnalysis estimates paralogous and orthologous pairwise synonymous (Ks) and non-synonymous (Ka) substitution rates	▼ Options		a + 17
Options Configuration			History 😺 🕸 📖
Advanced	•		
Alignment coverage configuration			Caranah data anta
Yes	-		search datasets
match score			
0.5			PlantTribes test data
Species rates recalibration configuration			31 shown, 9 deleted, 2233 hidden
Yes	•		
Recalibration rate			109.76 MB 🗹 🍽 🗩
0			
PAML codeml configuration		anapiast actimated significant	2273: KaKsAnalysis (sig 🗶 🖋 🗙
Yes	•	species resultated significant	nificant components in t
PAML codemi control file		components in the distribution of	he Kake distribution) on data 5 and
2267: codeml.ctl.args	•	synonymous (Ks) substitution rates	ne kaks distribution) on data 5 and
Rates clustering configuration			<u>data 6</u>
Yes	•		
Number of components		species1 non-synonymous (Ka)	2272: KaKsAnalysis (KaK 🗶 🖋 🗶
3		and synonymous (Ks)	s distribution) on data 5
Lower limit synonymous subsitution rates configuration		substitution rates	and data 6
Yes	•	Substitution rates	and data o
Minimum rate			2271: KakeAnalysis (nar and a se
0.02			2271. KaksAnarysis (par
Upper limit synonymous subsitution rates configuration		species i paralogous pairs —	alogous pairs) on data 5
Yes	•		and data 6
Maximum rate			
2			2270: KaKsAnalysis (bla 🕐 🖋 🗙
I certify that I am not using this tool for commercial purposes.		species1 self blastn results	stn results species1 vs s
Yes No			pecies1) on data 5 and data 6
Job Resource Parameters			1
Use default job resource parameters	•		
✓ Execute			

5. Selection and WGD analysis Genome duplication synonymous substitution rates (Ks) distribution plot

KsDistribution plots the distribution of synonymous substitution (Ks) rates and fits significant component(s) (Galaxy Version 1.0.3.0)	 Options
KaKsAnalysis tabular file	
🖸 🖗 🗅 2272: KaKsAnalysis (KaKs distribution) on data 5 and data 6	•
Significant components	
2273: KaKsAnalysis (significant components in the KaKs distribution) on data 5 and data 6	•
Choose colors for significant components	
Yes	•
Component colors	
1: Component colors	Ē
Color	
green	•
2: Component colors	圃
Color	
blue	•
+ Insert Component colors	
In the component colors	
Use default iob resource parameters	•
✓ Execute	



Applications

- Improving genome annotation quality
- Transcriptome coverage and functional annotation
- Species tree inference using single copy genes
- Gene tree species tree reconciliation
- Timing of gene duplications and polyploidy
- Ancestral gene content reconstruction
- Gene family expansions/contractions
- Timing of new gene function evolution
- Studies of horizontal transfers among species
- Many others...

Availability

Tools:

- Stand-alone: https://github.com/dePamphilis/PlantTribes
- Galaxy: https://usegalaxy.org/
- Galaxy Tool Shed: https://toolshed.g2.bx.psu.edu

Tutorials:

- Stand-alone:
 https://github.com/dePamphilis/PlantTribes/blob/master/docs/Tutorial.md
- Galaxy: https://galaxyproject.org/tutorials/pt_gfam/

Acknowledgments



Claude dePamphilis Penn State



James Leebens-Mack UGA



Raj Ayyampalayam UGA



Kerr Wall BASF



Greg Von Kuster Penn State



Norman Wickett Chicago Botanic Garden



Joshua Der CSU Fullerton



Loren Honaas USDA ARS

Funding and support:

Funding for this work was provided by the NSF Plant Genome Research Program through the Amborella Genome Sequencing Project (NSF Grant 0922742), Parasitic Plant Genome Project (NSF Grant 0701748), and Discovery and Functional Characterization of Genes Regulating Plant Immunity in Perennial Crops (NSF Grant 1546863) with additional support from the One Thousand Plants (1KP) transcriptome project.