EVOMICS — III Workshop on Phylogenomics

31.I.2024

Target Capture Sequencing Approaches

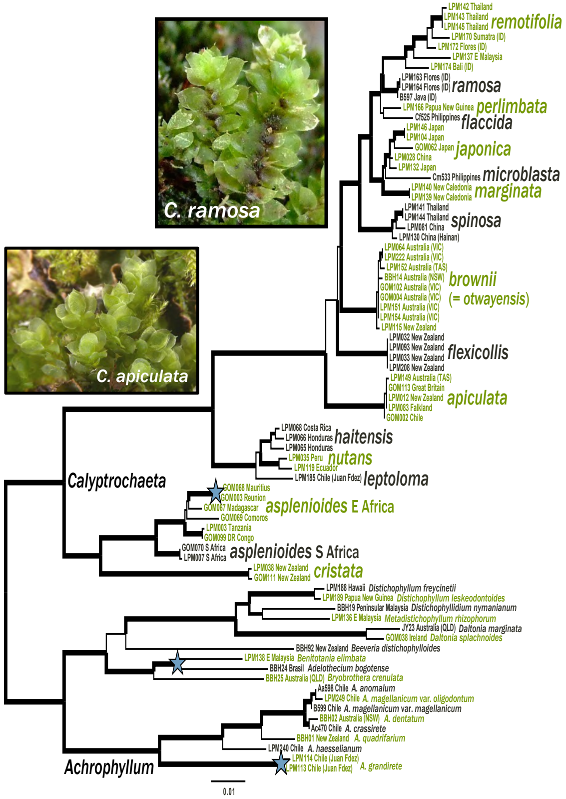
TCS Approaches in Phylogenomics

Lisa Pokorny

Royal Botanic Garden-Spanish National Research Council (RJB-CSIC)

I got my PhD in Biology @ **Duke University** (USA).

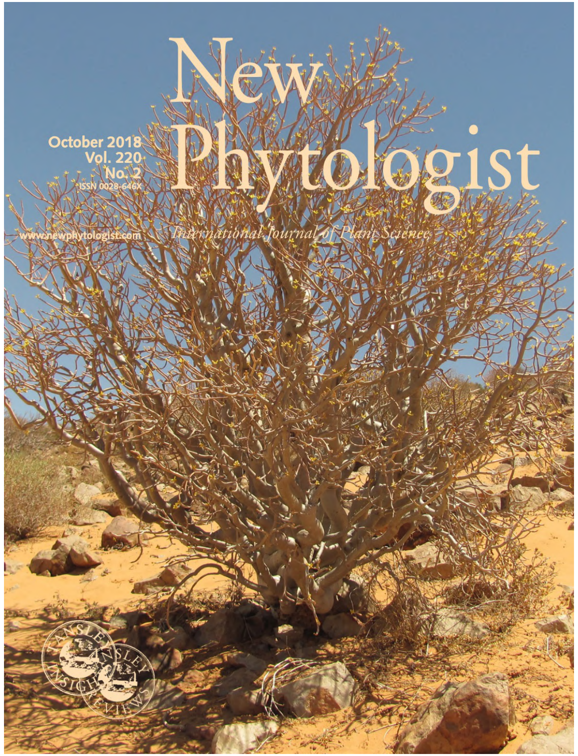
—Lisa Pokorny, PhD



I got my PhD in Biology @ **Duke University** (USA).

My postdoctoral trajectory spans Madrid's Royal Botanic Garden (**RJB-CSIC**, ES)

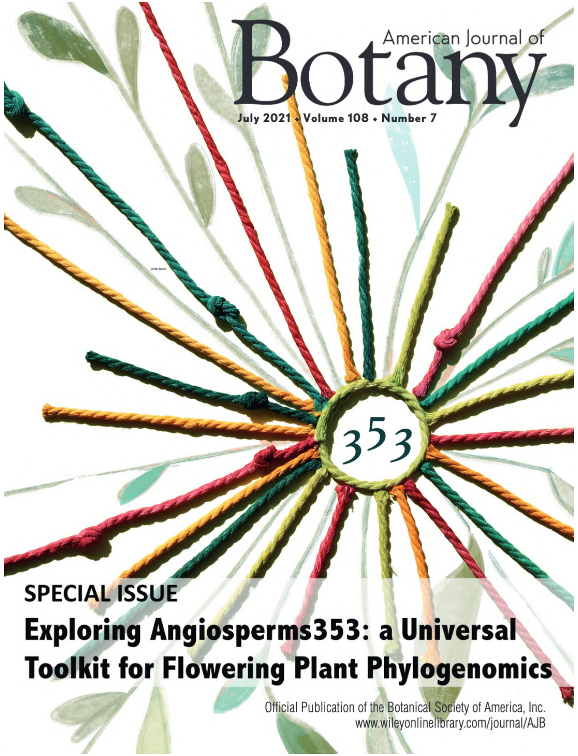
—Lisa Pokorny, PhD



I got my PhD in Biology @ **Duke University** (USA).

My postdoctoral trajectory spans Madrid's Royal Botanic Garden (**RJB-CSIC**, ES), Royal Botanic Gardens, Kew (**RBGK**, UK)

—Lisa Pokorny, PhD



American Journal of Botany

July 2021 • Volume 108 • Number 7

353

SPECIAL ISSUE

Exploring Angiosperms353: a Universal Toolkit for Flowering Plant Phylogenomics

Official Publication of the Botanical Society of America, Inc.
www.wileyonlinelibrary.com/journal/AJB

I got my PhD in Biology @ **Duke University** (USA).

My postdoctoral trajectory spans Madrid's Royal Botanic Garden (**RJB-CSIC**, ES), Royal Botanic Gardens, Kew (**RBGK**, UK), the Centre for Plant Biotechnology & Genomics (**CBGP UPM-INIA**, ES)

—Lisa Pokorny, PhD

I got my PhD in Biology @ **Duke University** (USA).

My postdoctoral trajectory spans Madrid's Royal Botanic Garden (**RJB-CSIC**, ES), Royal Botanic Gardens, Kew (**RBGK**, UK), the Centre for Plant Biotechnology & Genomics (**CBGP UPM-INIA**, ES), and Barcelona's Botanical Institute (**IBB CSIC**, ES).

—Lisa Pokorny, PhD

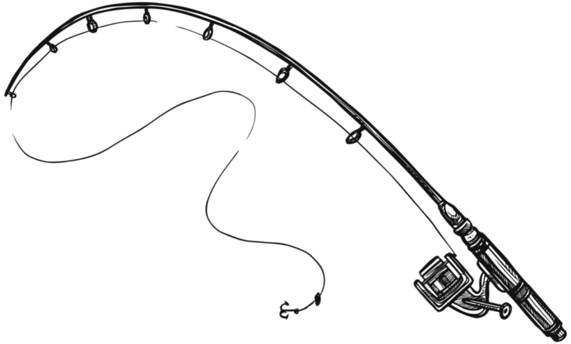
I got my PhD in Biology @ **Duke University** (USA).

My postdoctoral trajectory spans Madrid's Royal Botanic Garden (**RJB-CSIC**, ES), Royal Botanic Gardens, Kew (**RBGK**, UK), the Centre for Plant Biotechnology & Genomics (**CBGP UPM-INIA**, ES), and Barcelona's Botanical Institute (**IBB CSIC**, ES).

Nowadays, I am a **Ramón y Cajal Research Fellow** working for the Spanish National Research Council (CSIC) and based at **RJB** (Madrid, ES).

—Lisa Pokorny, PhD

Table of contents

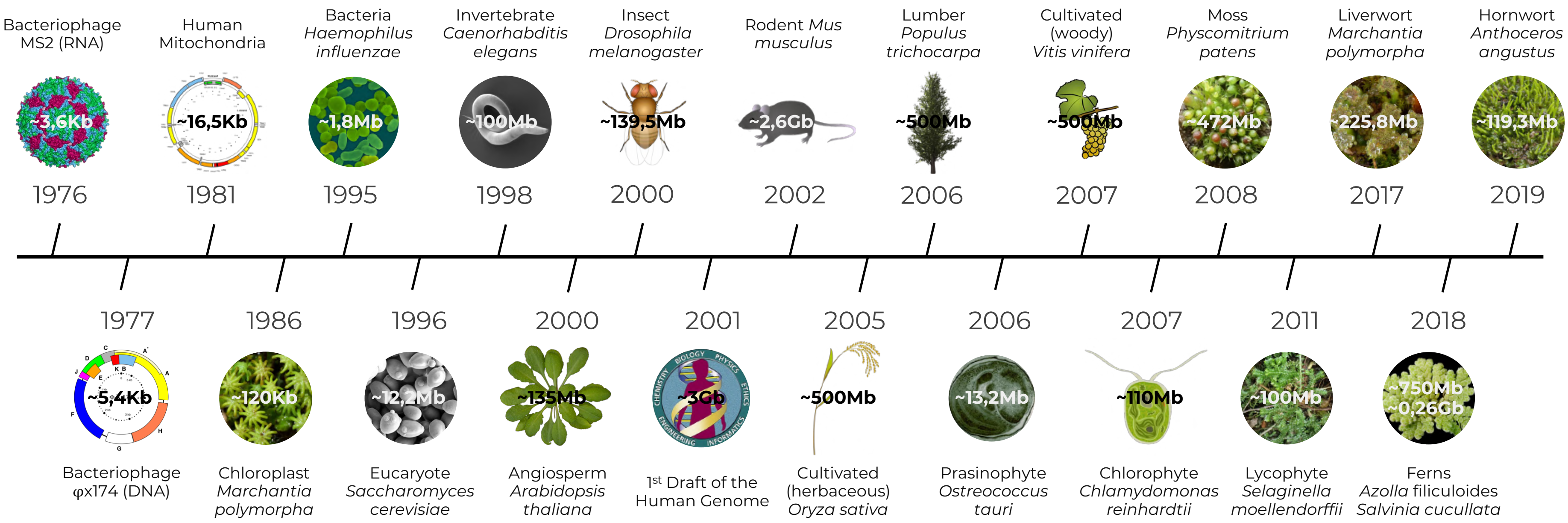


Genome Subsampling

An Introduction to TCS Approaches

01

Whole Genome Sequencing (WGS)



Bacteriophage MS2 (RNA)
~3,6Kb
1976

Human Mitochondria
~16,5Kb
1981

Bacteria *Haemophilus influenzae*
~1,8Mb
1995

Invertebrate *Caenorhabditis elegans*
~100Mb
1998

Insect *Drosophila melanogaster*
~139,5Mb
2000

Rodent *Mus musculus*
~2,6Gb
2002

Lumber *Populus trichocarpa*
~500Mb
2006

Cultivated (woody) *Vitis vinifera*
~500Mb
2007

Moss *Physcomitrium patens*
~472Mb
2008

Liverwort *Marchantia polymorpha*
~225,8Mb
2017

Hornwort *Anthoceros angustus*
~119,3Mb
2019

1977
~5,4Kb
Bacteriophage φx174 (DNA)

1986
~120Kb
Chloroplast *Marchantia polymorpha*

1996
~12,2Mb
Eucaryote *Saccharomyces cerevisiae*

2000
~135Mb
Angiosperm *Arabidopsis thaliana*

2001
~3Gb
1st Draft of the Human Genome

2005
~500Mb
Cultivated (herbaceous) *Oryza sativa*

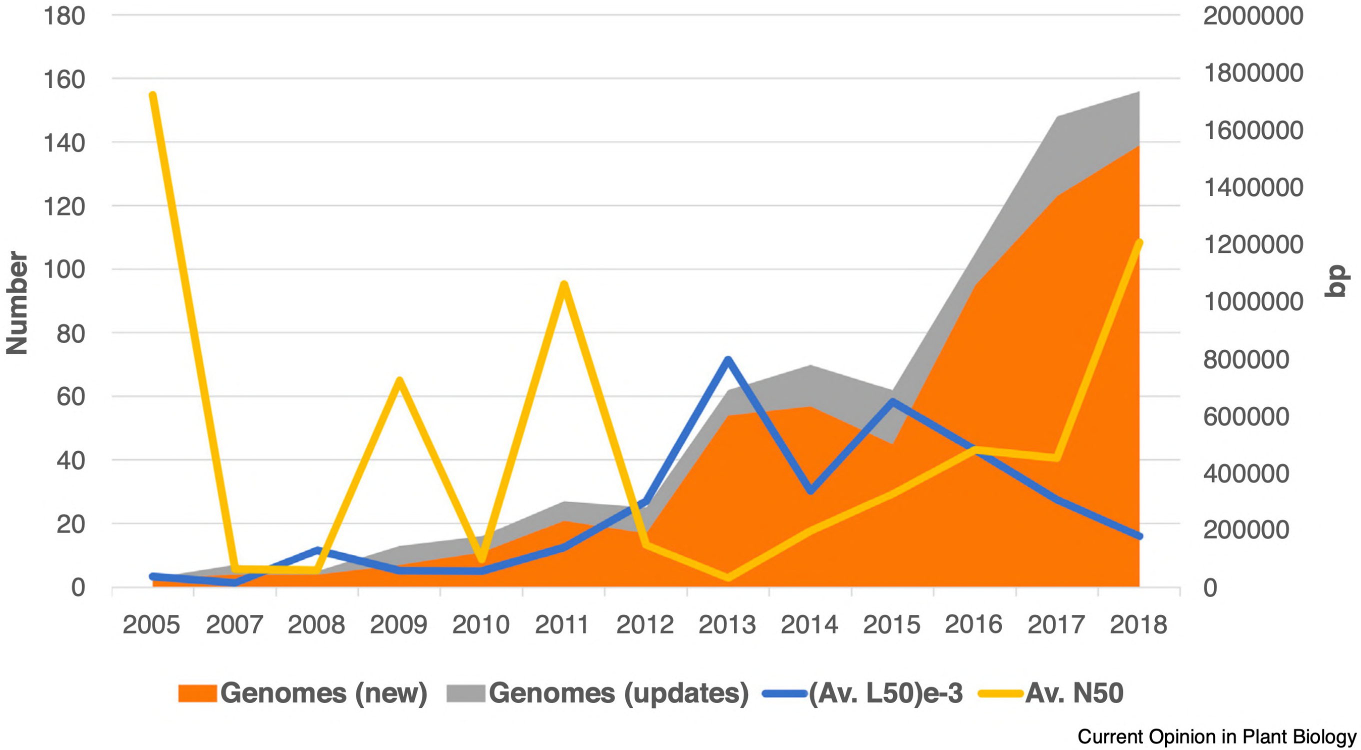
2006
~13,2Mb
Prasinophyte *Ostreococcus tauri*

2007
~110Mb
Chlorophyte *Chlamydomonas reinhardtii*

2011
~100Mb
Lycophyte *Selaginella moellendorffii*

2018
~750Mb
~0,26Gb
Ferns *Azolla* filiculoides *Salvinia cucullata*

Pace of WGS in Plants

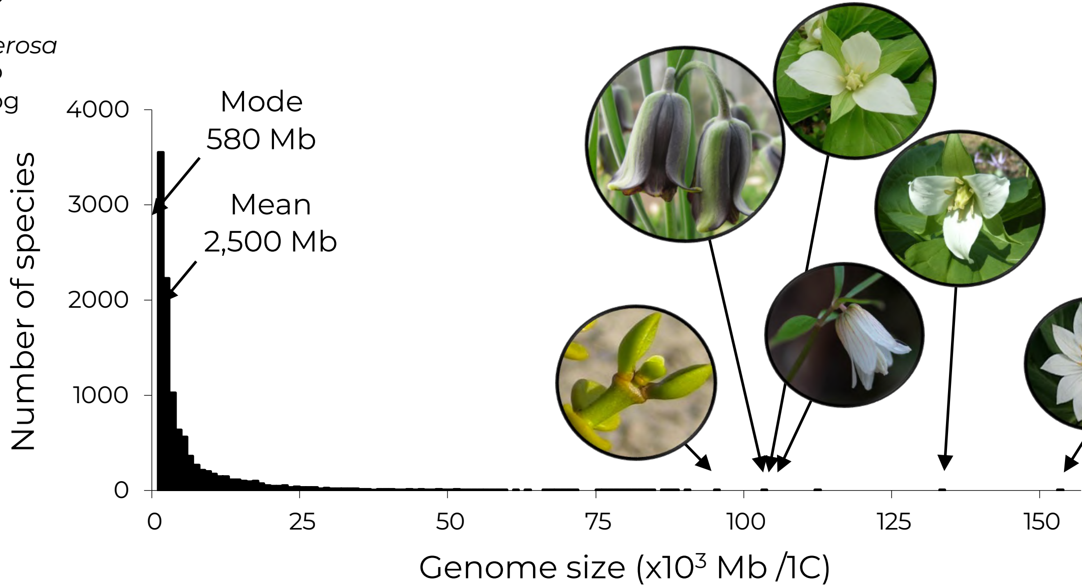


Kersey. 2019. Plant genome sequences: past, present, future. *Curr. Opin. Plant Biol.* 48: 1–8.

Genome Size

x 2400

Genlisea tuberosa
1C=64Mb
2C=0,124 pg



Paris japonica
1C=140Gb
2C=304,46 pg

Pellicer et al. 2010. *Bot. J. Linn. Soc.* | Fleischmann et al. 2014. *Ann. Bot.* | Pellicer & Leitch. 2020. *New Phytol.*

Seq Platforms & Costs

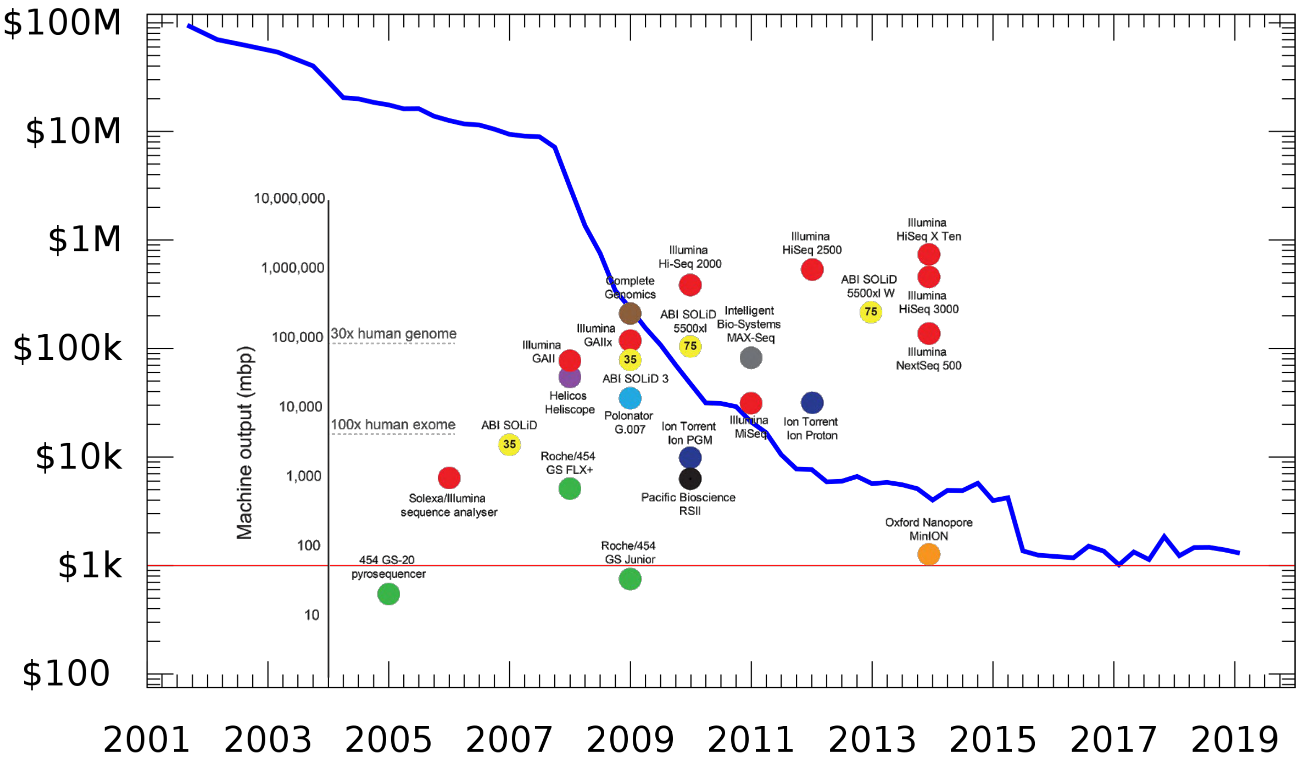
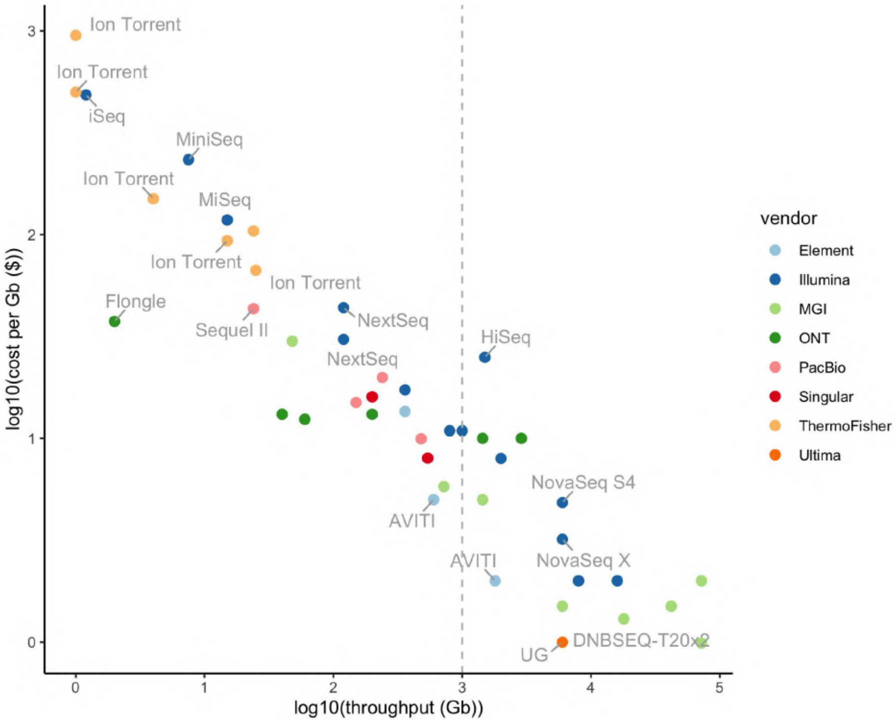


Fig. 1. Reuter et al. 2015. Mol. Cell 58: 586–597 & genomegov/sequencingcosts

Seq Platforms & Performance

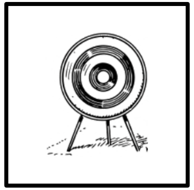


https://twitter.com/geertvangeest/status/1725463615421956343

Genome Subsampling

Anonymous

Genome Skimming
RRS (RAD-seq, GBS)



Targeted

RNA-seq
Target Capture Seq
(TCS)

Mixed

Hyb-Seq

Genome Subsampling

Anonymous

Genome Skimming
RRS (RAD-seq, GBS)

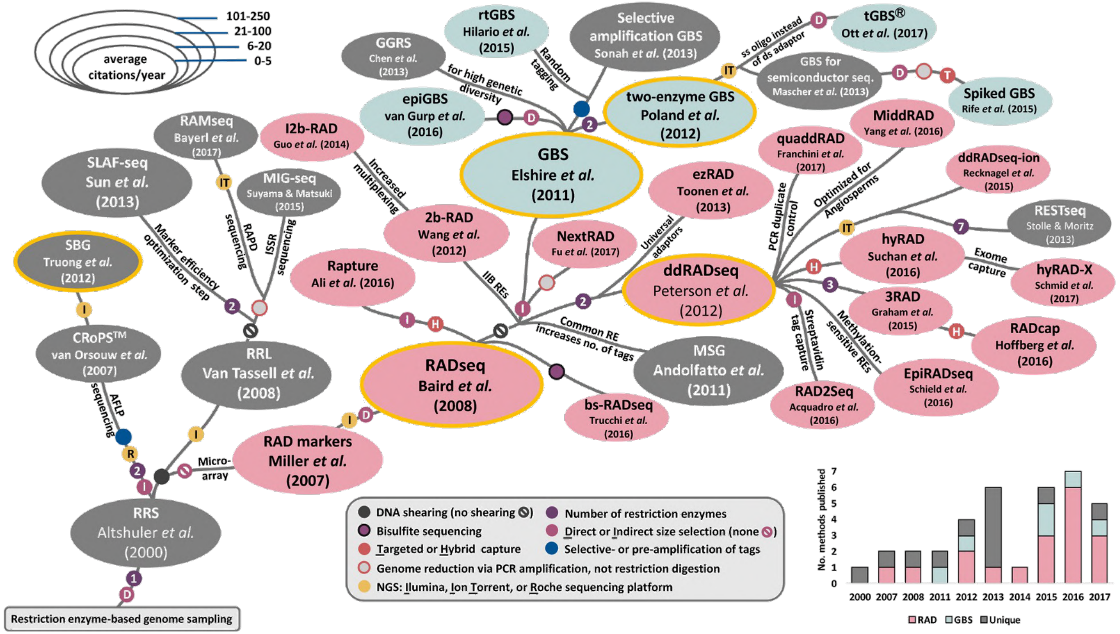
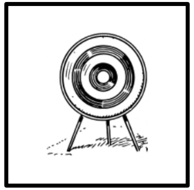


Fig. 1. Campbell et al. 2018. Would an RRS by any other name sound as RAD? *Methods in Ecology and Evolution* 9(9): 1920–1927.

Genome Subsampling

Anonymous

Genome Skimming
RRS (RAD-seq, GBS)



Targeted

RNA-seq
Target Capture Seq
(TCS)

Mixed

Hyb-Seq

Genome Subsampling

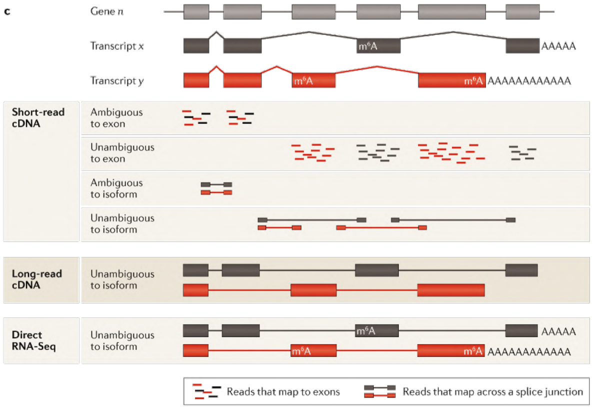
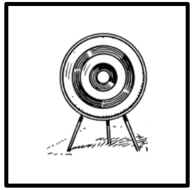


Fig. 1c. Stark et al. 2019. *Nat. Rev. Genet.* 20: 631–656



Targeted

RNA-seq
Target Capture Seq
(TCS)

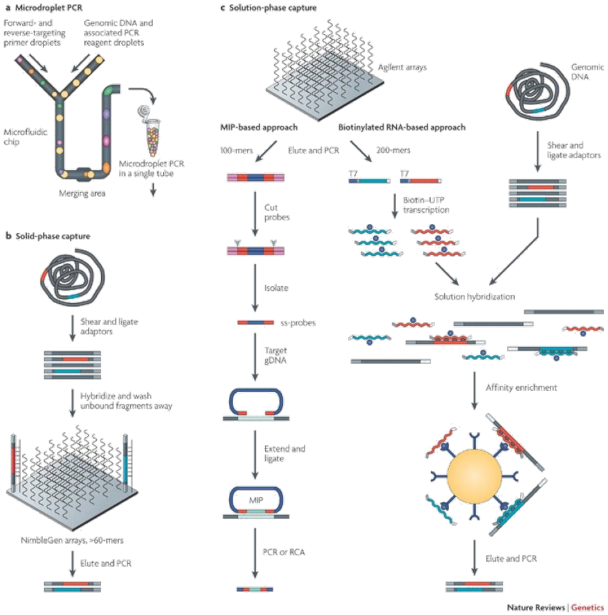
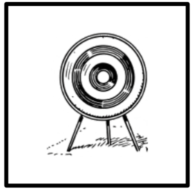


Fig. 5. Metzker. 2010. *Nat. Rev. Genet.* 11: 31–46

Genome Subsampling

Anonymous

Genome Skimming
RRS (RAD-seq, GBS)



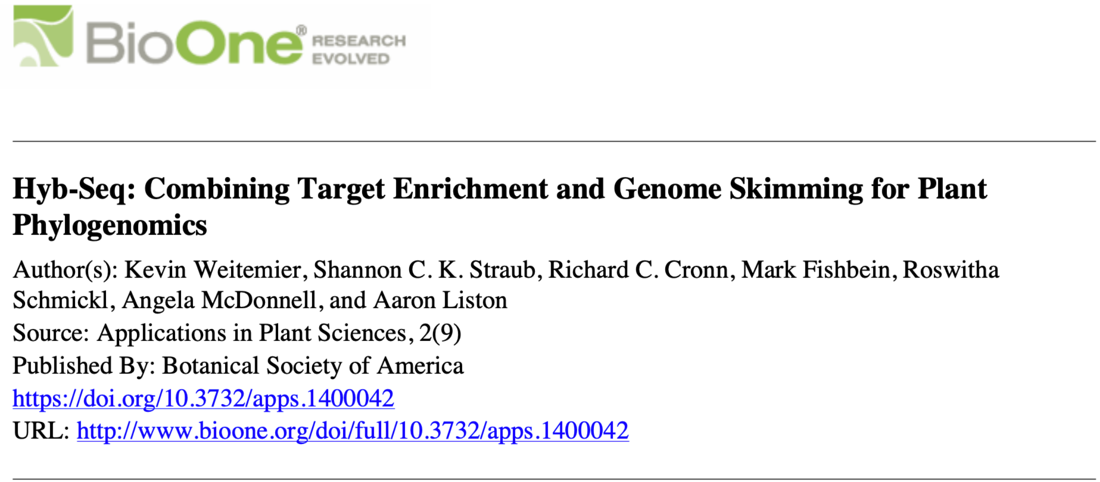
Targeted

RNA-seq
Target Capture Seq
(TCS)

Mixed

Hyb-Seq

Genome Subsampling



BioOne® RESEARCH EVOLVED

Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant Phylogenomics

Author(s): Kevin Weitemier, Shannon C. K. Straub, Richard C. Cronn, Mark Fishbein, Roswitha Schmickl, Angela McDonnell, and Aaron Liston
Source: Applications in Plant Sciences, 2(9)
Published By: Botanical Society of America
https://doi.org/10.3732/apps.1400042
URL: http://www.bioone.org/doi/full/10.3732/apps.1400042

Mixed

Hyb-Seq

Genome Subsampling

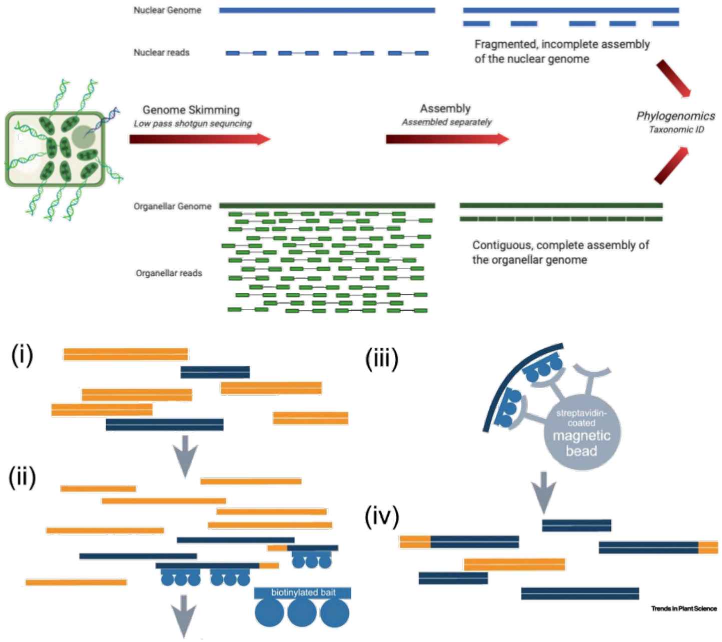


Fig. 1. Dodsworth, Pokorny, Johnson et al. 2019. *Trends Plant Sci.* 24(10): 887–891

TRPLSC 1851 No. of Pages 4

Trends in Plant Science

Forum

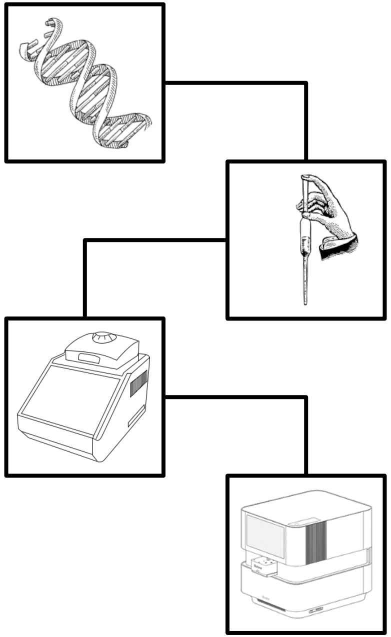
Hyb-Seq for Flowering Plant Systematics

Steven Dodsworth,[1,2,7,*]
Lisa Pokorny,[1,3,7]
Matthew G. Johnson,[4,5,7]
Jan T. Kim,[1] Olivier Maurin,[1]
Norman J. Wickett,[5,6]
Felix Forest,[1] and William J. Baker[1]

Mixed

Hyb-Seq

Hyb-Seq Workflow



DNA extraction

Comercial kits vs.
in-house protocols

Genomic Library Preparation

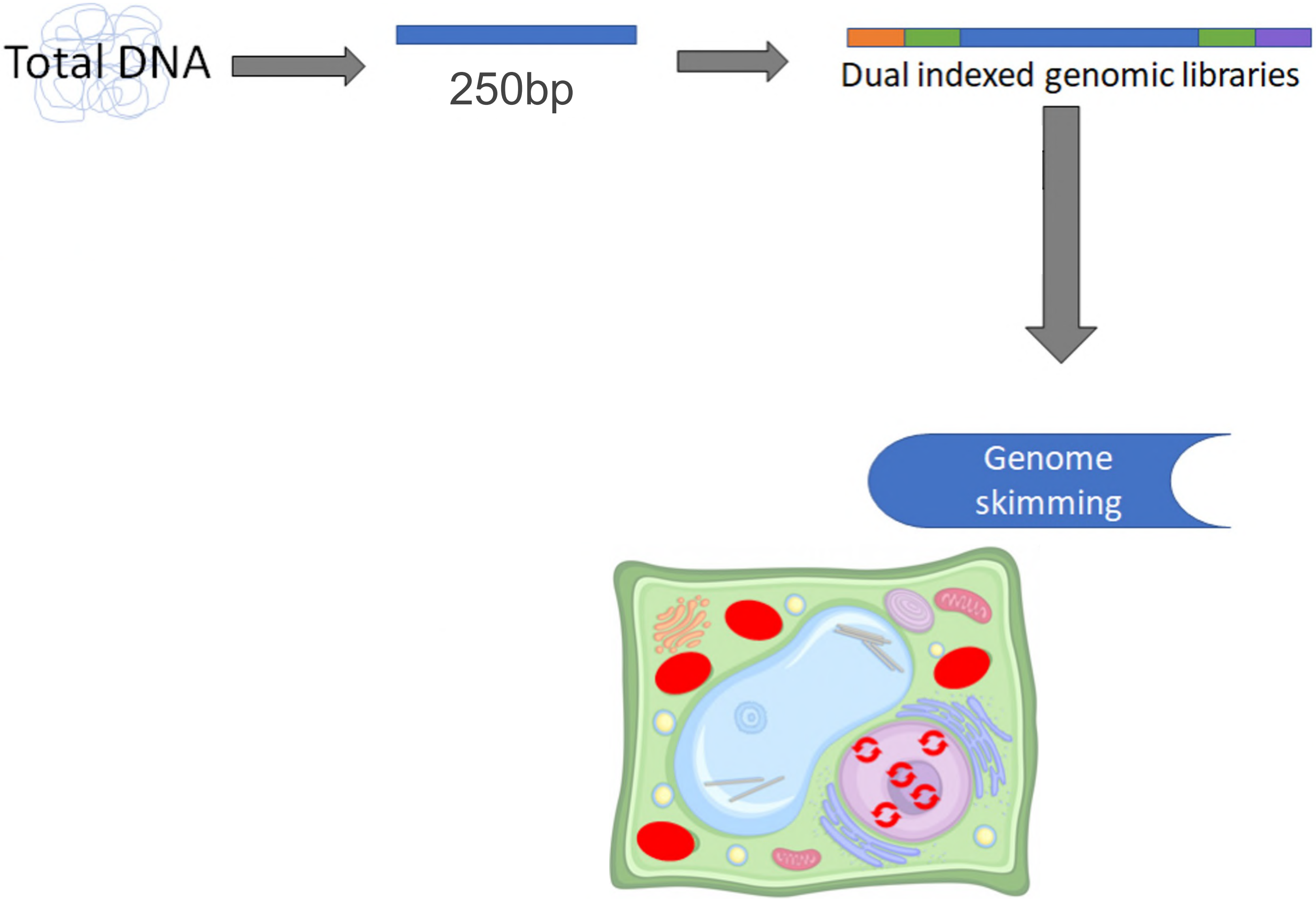
Dual indexing for
multiplexing

Hybridization

Target enrichment with
probe set

HTS

Multiplexed sequencing

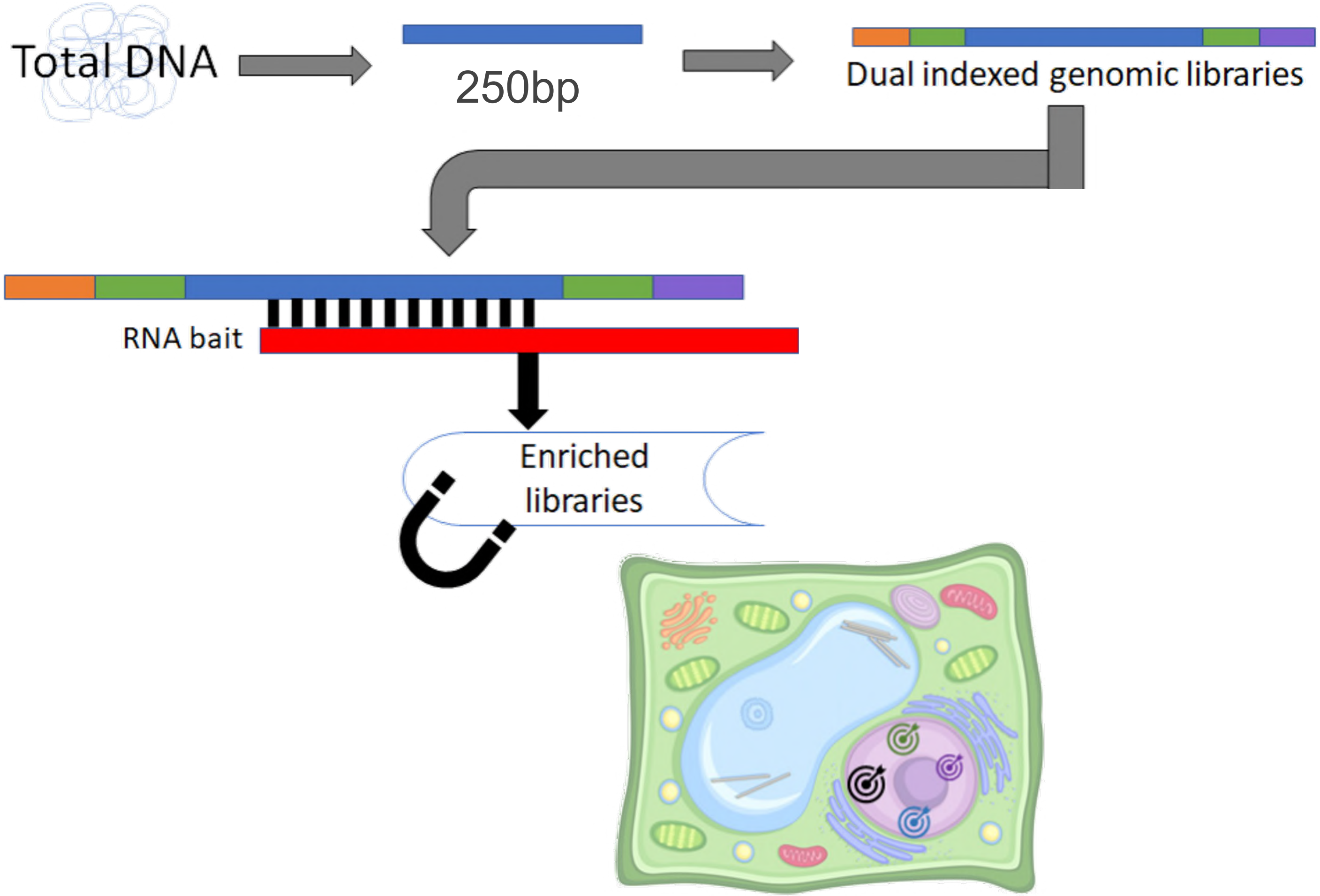


Total DNA

250bp

Dual indexed genomic libraries

Genome skimming



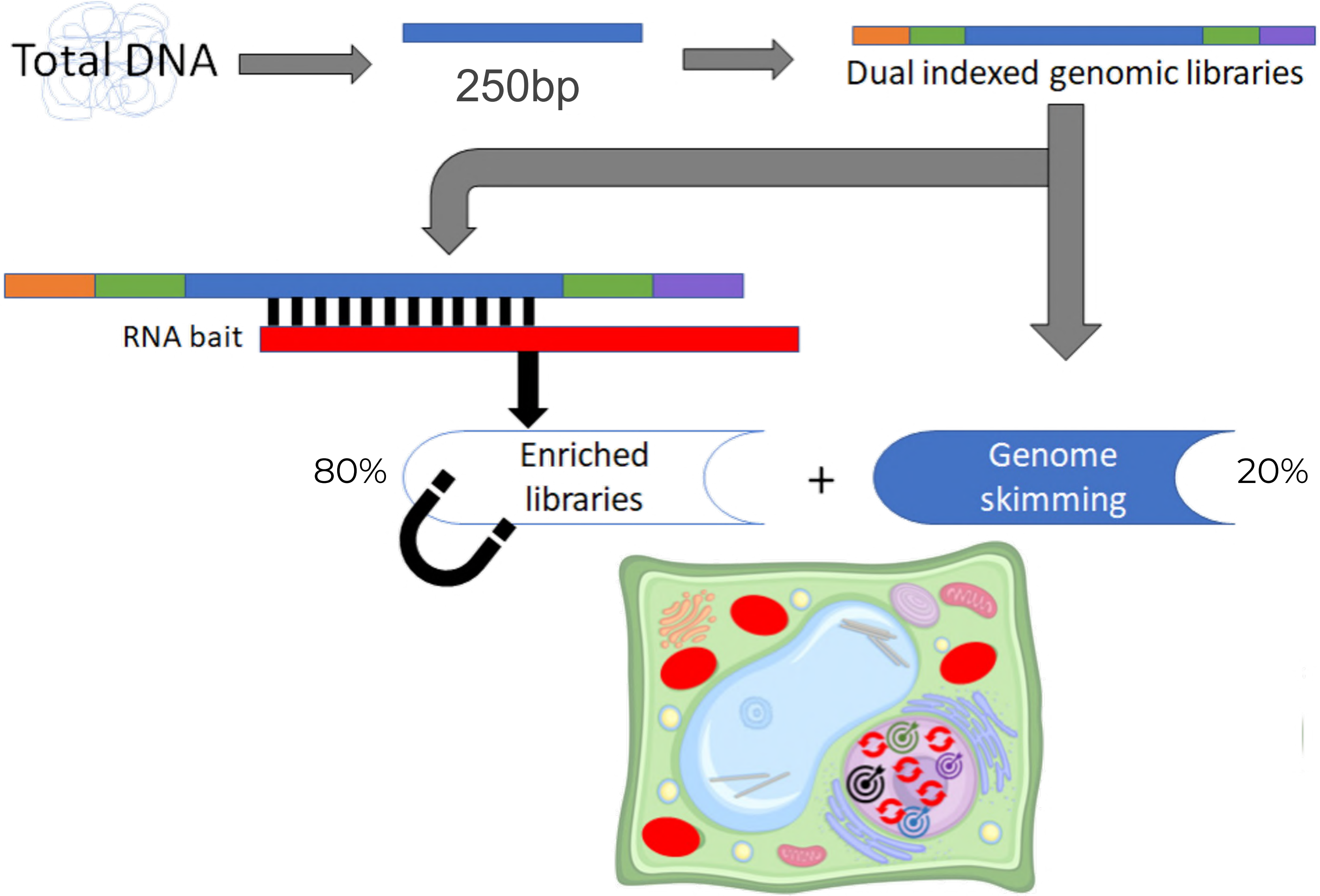
Total DNA

250bp

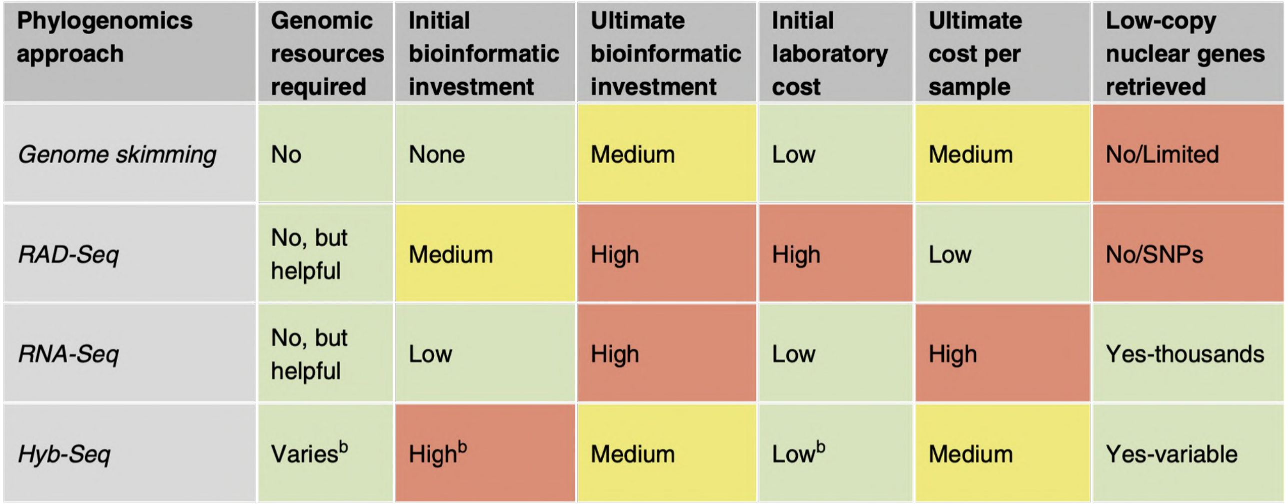
Dual indexed genomic libraries

RNA bait

Enriched libraries


Total DNA
250bp
Dual indexed genomic libraries
RNA bait
80%
Enriched libraries
+
Genome skimming
20%

Advantages & Disadvantages



Phylogenomics approach	Genomic resources required	Initial bioinformatic investment	Ultimate bioinformatic investment	Initial laboratory cost	Ultimate cost per sample	Low-copy nuclear genes retrieved
Genome skimming	No	None	Medium	Low	Medium	No/Limited
RAD-Seq	No, but helpful	Medium	High	High	Low	No/SNPs
RNA-Seq	No, but helpful	Low	High	Low	High	Yes-thousands
Hyb-Seq	Varies[b]	High[b]	Medium	Low[b]	Medium	Yes-variable

[a] Initial costs include the one-time or limited purchase of expensive consumables (e.g., biotinylated baits or adapter sequences). Boxes are highlighted from unfavourable (red) to favourable (green) under each column.

[b] If designing new kit(s) genome or transcriptome resources are required, otherwise readily available kits exist for different groups of plants as well as angiosperms as a whole (Angiosperms-353) and are much cheaper than designing a new custom bait set.

Table 1. Dodsworth, Pokorny, Johnson et al. 2019. *Trends Plant Sci.* 24(10): 887–891

Table of contents

01

Genome Subsampling

An Introduction to TCS Approaches

02

Probe Set Design

Custom & Universal Kits

03

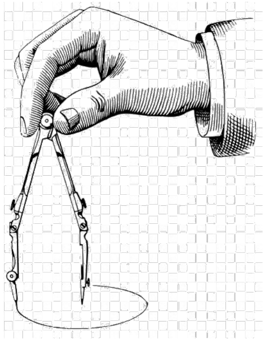
Implementing TCS

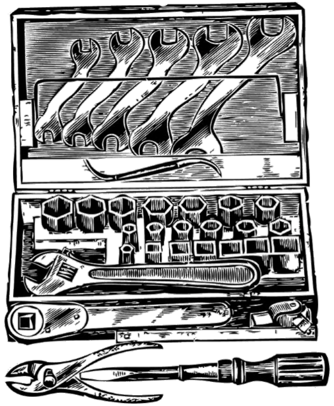
Wet Lab Workflow & Dry Lab Pipeline

04

Biodiversity Genomics

Natural History Collections



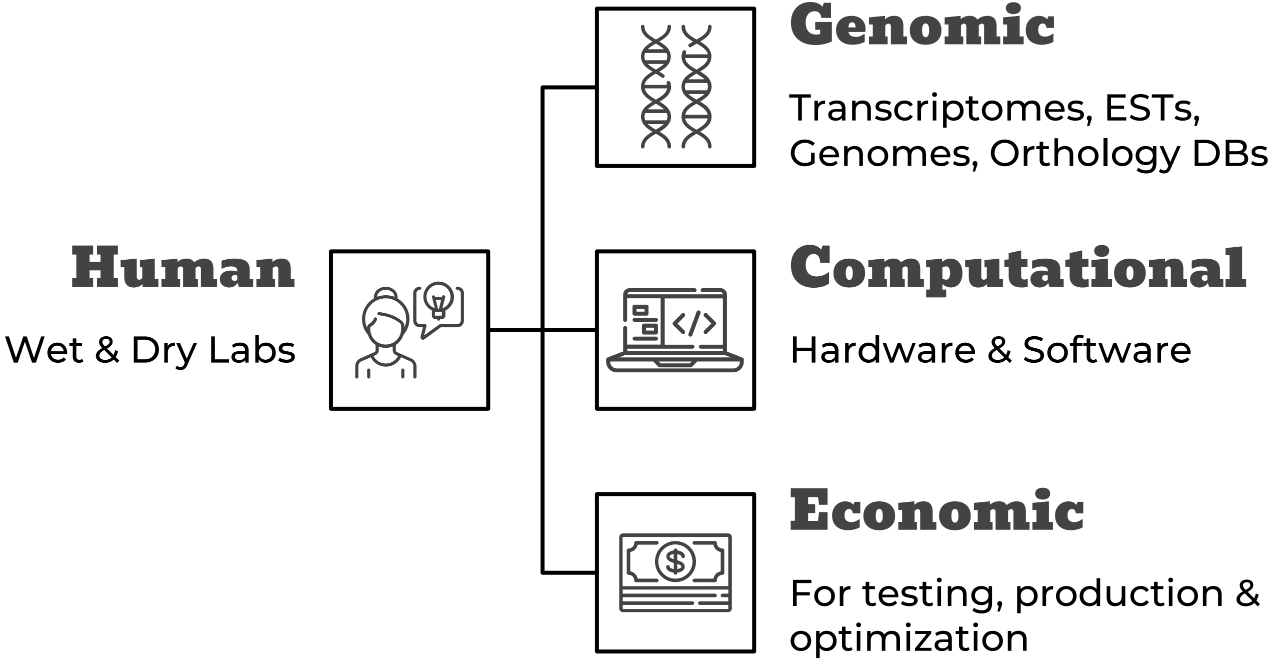


Probe Set Design

Custom Kits

02

Resources needed



Human

Wet & Dry Labs

Genomic

Transcriptomes, ESTs, Genomes, Orthology DBs

Computational

Hardware & Software

Economic

For testing, production & optimization

Resources needed

Human

Wet & Dry Labs

Human Resources

Wet Lab

REAL JARDÍN BOTÁNICO

CSIC

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Sociedad de amigos del

REAL JARDÍN BOTÁNICO

Dry Lab

Resources needed

Human

Wet & Dry Labs

Genomic

Transcriptomes, ESTs, Genomes, Orthology DBs

Computational

Hardware & Software

Economic

For testing, production & optimization

Resources needed

Genomic

Transcriptomes, ESTs, Genomes, Orthology DBs

Transcriptomes

The international journal of science / 31 October 2019

nature

outline
Prostate cancer

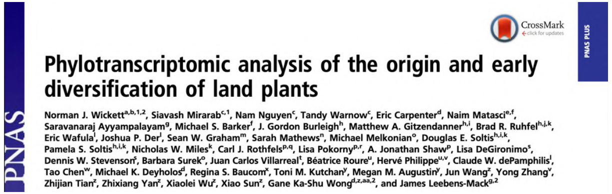
ROOTS OF DIVERSITY

Transcriptome analysis illuminates evolution of the world's green plants

A history of ethics The long and bumpy road to responsible research

Ancient climate A snapshot of CO_2 in the atmosphere more than 1 million years ago

Insects in decline Ten-year survey offers strong evidence of falling numbers



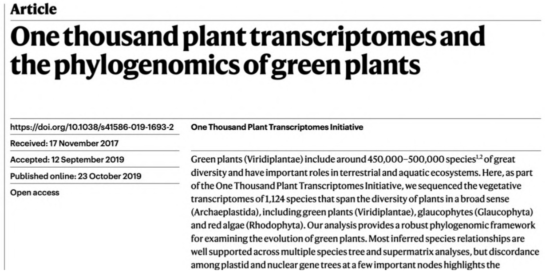
PNAS

CrossMark

PNAS PLUS

Phylotranscriptomic analysis of the origin and early diversification of land plants

Norman J. Wickett[a,b,1,2], Siavash Mirarab[c,1], Nam Nguyen[c], Tandy Warnow[c], Eric Carpenter[d], Naim Matasci[e,f], Saravanaraj Ayyampalayam[g], Michael S. Barker[f], J. Gordon Burleigh[h], Matthew A. Gitzendanner[h,i], Brad R. Ruhfel[h,j,k], Eric Wafula[l], Joshua P. Der[l], Sean W. Graham[m], Sarah Mathews[n], Michael Melkonian[o], Douglas E. Soltis[h,i,k], Pamela S. Soltis[h,i,k], Nicholas W. Miles[k], Carl J. Rothfels[p,q], Lisa Pokorny[p,r], A. Jonathan Shaw[p], Lisa DeGironimo[s], Dennis W. Stevenson[s], Barbara Surek[o], Juan Carlos Villarreal[t], Béatrice Roure[u], Hervé Philippe[u,v], Claude W. dePamphilis[l], Tao Chen[w], Michael K. Deyholos[d], Regina S. Baucom[x], Toni M. Kutchan[y], Megan M. Augustin[y], Jun Wang[z], Yong Zhang[v], Zhijian Tian[z], Zhixiang Yan[z], Xiaolei Wu[z], Xiao Sun[z], Gane Ka-Shu Wong[d,z,aa,2], and James Leebens-Mack[g,2]



Article

One thousand plant transcriptomes and the phylogenomics of green plants

https://doi.org/10.1038/s41586-019-1693-2
Received: 17 November 2017
Accepted: 12 September 2019
Published online: 23 October 2019
Open access

One Thousand Plant Transcriptomes Initiative

Green plants (Viridiplantae) include around 450,000–500,000 species[1,2] of great diversity and have important roles in terrestrial and aquatic ecosystems. Here, as part of the One Thousand Plant Transcriptomes Initiative, we sequenced the vegetative transcriptomes of 1,124 species that span the diversity of plants in a broad sense (Archaeplastida), including green plants (Viridiplantae), glaucophytes (Glaucophyta) and red algae (Rhodophyta). Our analysis provides a robust phylogenomic framework for examining the evolution of green plants. Most inferred species relationships are well supported across multiple species tree and supermatrix analyses, but discordance among plastid and nuclear gene trees at a few important nodes highlights the

1KP

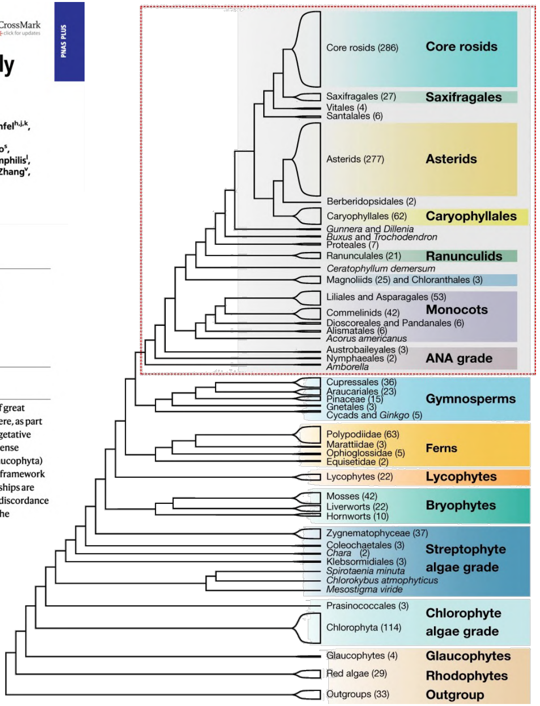
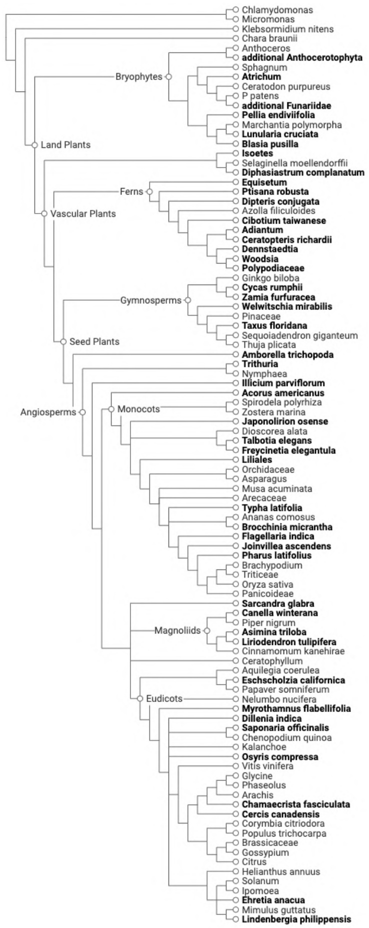
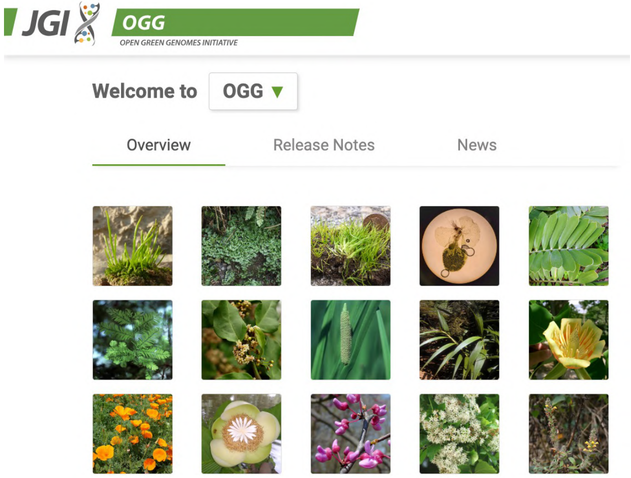


Fig. 1b. OTPTI. 2019. Nature 574: 679–685.

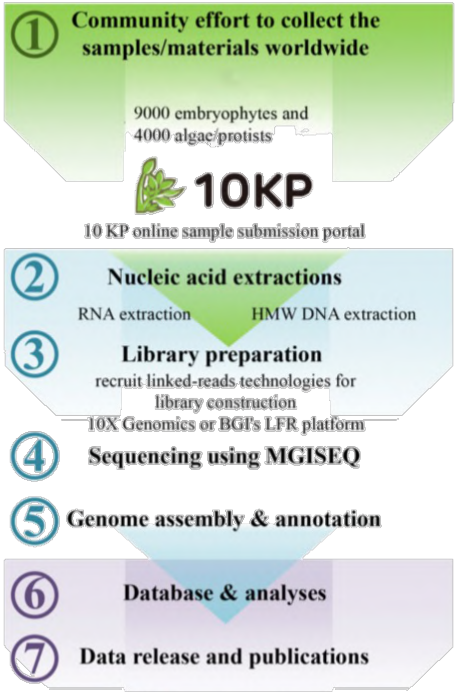


Genomes



The **Open Green Genomes Initiative** (**OGG**, a Joint Genome Institute community science program) seeks to improve the phylogenetic representation of reference genome sequences across land plants by generating at least thirty-five high-quality genome assemblies and annotations across all major evolutionary lineages in the land plant tree of life (target genomes are labelled in **bold** below). This work will greatly improve comparative analyses of the genes, regulatory networks and metabolic pathways influencing plant growth, responses to environmental stress, and production of valuable plant products. The resulting comparative framework will drive translation of genome science into predictive understanding and inform engineering of efficient production of plant-based biofuels and bioproducts.

Sequence a member of every embryophyte genus and a phylodiverse collection of algae/protists



① Community effort to collect the samples/materials worldwide

9000 embryophytes and 4000 algae/protists

10KP

10 KP online sample submission portal

② Nucleic acid extractions

RNA extraction

HMW DNA extraction

③ Library preparation

recruit linked-reads technologies for library construction

10X Genomics or BGI's LFR platform

④ Sequencing using MGISEQ

⑤ Genome assembly & annotation

⑥ Database & analyses

⑦ Data release and publications

CNGB open access database

Biodiversity data release

Orthology Databases

Ensembl

http://www.ensembl.org/

EggNOG

http://eggnog5.embl.de/

OrthoDB

https://www.orthodb.org/

PLAZA

bioinformatics.psb.ugent.be/plaza/

MetaPhOrs

orthology.phylomedb.org/

OMA

https://omabrowser.org/

KO (KEGG)

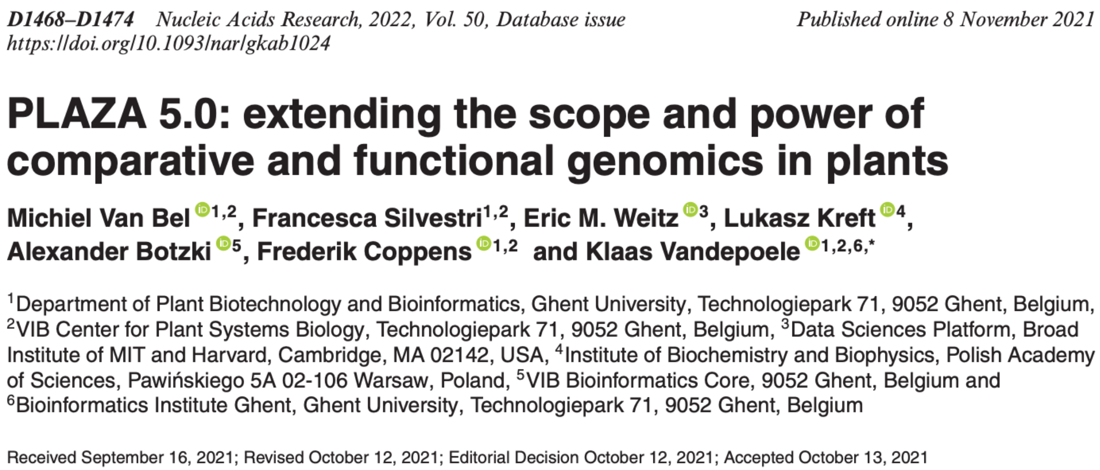
www.genome.jp/kegg/ko.html

FungiDB

https://fungidb.org/

https://questforthorthologs.org/orthology_databases

Orthology Databases



D1468–D1474 *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 8 November 2021*
https://doi.org/10.1093/nar/gkab1024

PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants

Michiel Van Bel [1,2], Francesca Silvestri[1,2], Eric M. Weitz [3], Lukasz Kreft [4], Alexander Botzki [5], Frederik Coppens [1,2] and Klaas Vandepoele [1,2,6,*]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 71, 9052 Ghent, Belgium, [2]VIB Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium, [3]Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA, [4]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawińskiego 5A 02-106 Warsaw, Poland, [5]VIB Bioinformatics Core, 9052 Ghent, Belgium and [6]Bioinformatics Institute Ghent, Ghent University, Technologiepark 71, 9052 Ghent, Belgium

Received September 16, 2021; Revised October 12, 2021; Editorial Decision October 12, 2021; Accepted October 13, 2021

PLAZA

bioinformatics.psb.ugent.be/plaza/

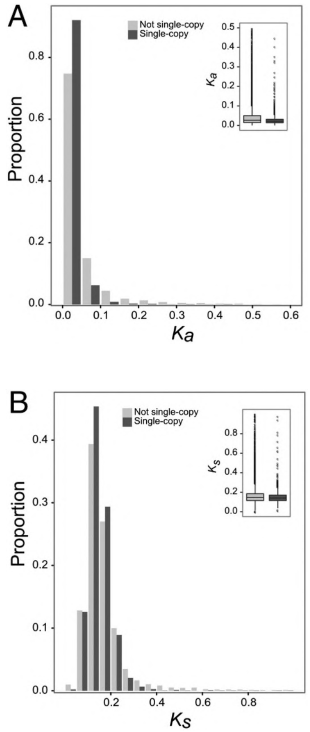


Fig. 3. De Smet et al. 2013. *PNAS* 110: 2898–2903

Resources needed

Human

Wet & Dry Labs

Genomic

Transcriptomes, ESTs, Genomes, Orthology DBs

Computational

Hardware & Software

Economic

For testing, production & optimization

Resources needed

Computational

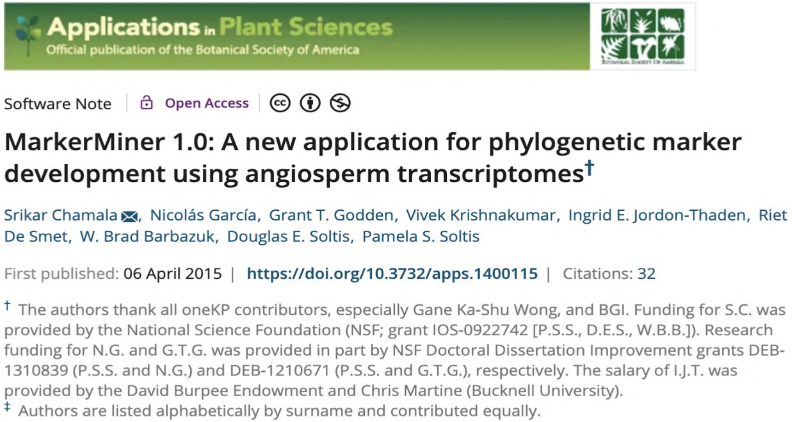
Hardware & Software

Hardware

BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

CESGA
GALICIA SUPERCOMPUTING CENTER

Software


Applications in Plant Sciences
Official publication of the Botanical Society of America

Software Note | Open Access

MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes†

Srikar Chamala, Nicolás García, Grant T. Godden, Vivek Krishnakumar, Ingrid E. Jordon-Thaden, Riet De Smet, W. Brad Barbazuk, Douglas E. Soltis, Pamela S. Soltis

First published: 06 April 2015 | https://doi.org/10.3732/apps.1400115 | Citations: 32

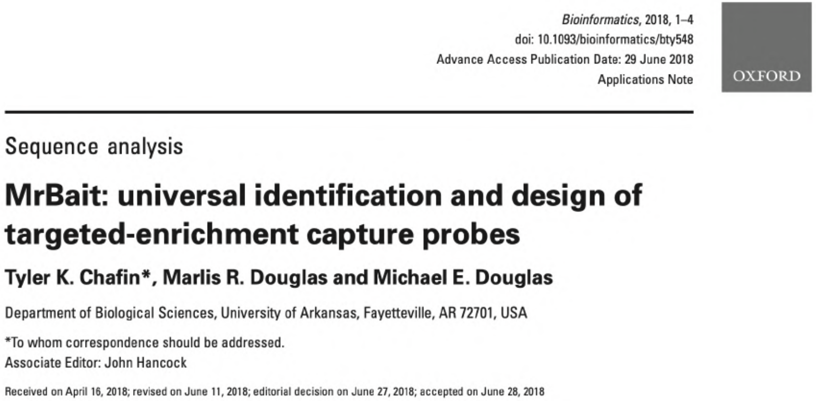
† The authors thank all oneKP contributors, especially Gane Ka-Shu Wong, and BGI. Funding for S.C. was provided by the National Science Foundation (NSF; grant IOS-0922742 [P.S.S., D.E.S., W.B.B.]). Research funding for N.G. and G.T.G. was provided in part by NSF Doctoral Dissertation Improvement grants DEB-1310839 (P.S.S. and N.G.) and DEB-1210671 (P.S.S. and G.T.G.), respectively. The salary of I.J.T. was provided by the David Burpee Endowment and Chris Martine (Bucknell University).
‡ Authors are listed alphabetically by surname and contributed equally.


BaitFisher: A Software Package for Multispecies Target DNA Enrichment Probe Design FREE

Christoph Mayer, Manuela Sann, Alexander Donath, Martin Meixner, Lars Podsiadlowski, Ralph S. Peters, Malte Petersen, Karen Meusemann, Karsten Liere, Johann-Wolfgang Wägele ... Show more

Author Notes

Molecular Biology and Evolution, Volume 33, Issue 7, July 2016, Pages 1875–1886,
https://doi.org/10.1093/molbev/msw056
Published: 23 March 2016


Bioinformatics, 2018, 1–4
doi: 10.1093/bioinformatics/bty548
Advance Access Publication Date: 29 June 2018
Applications Note

OXFORD

Sequence analysis

MrBait: universal identification and design of targeted-enrichment capture probes

Tyler K. Chafin*, Marlis R. Douglas and Michael E. Douglas

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

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on April 16, 2018; revised on June 11, 2018; editorial decision on June 27, 2018; accepted on June 28, 2018


Kuchinski *et al. BMC Genomics* (2022) 23:579
https://doi.org/10.1186/s12864-022-08790-4

BMC Genomics

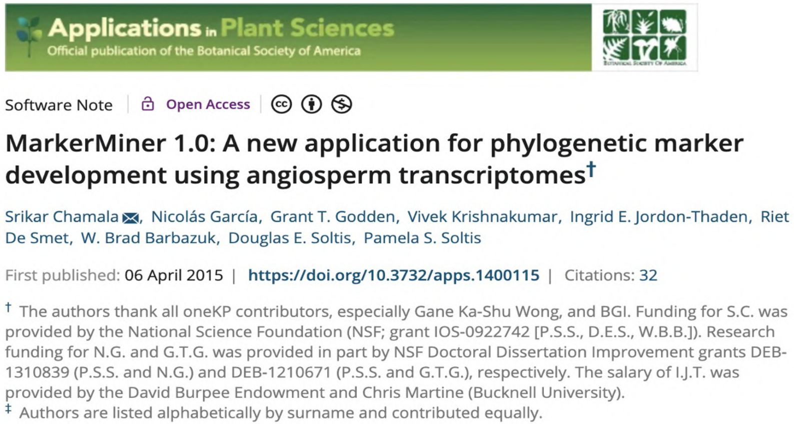
RESEARCH Open Access

Check for updates

ProbeTools: designing hybridization probes for targeted genomic sequencing of diverse and hypervariable viral taxa

Kevin S. Kuchinski[1,2*], Jun Duan[1], Chelsea Himsworth[3,4], William Hsiao[1,5] and Natalie A. Prystajecky[1,6]

Software



Applications in Plant Sciences
Official publication of the Botanical Society of America

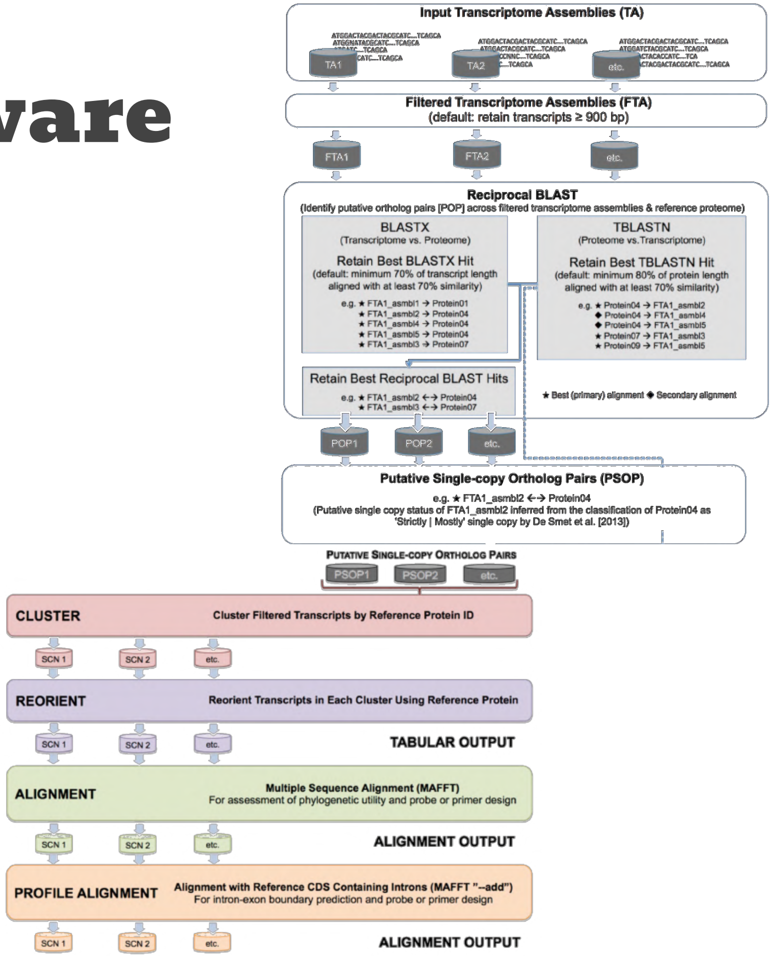
Software Note | Open Access | 🅭 🅯 🄏

MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes†

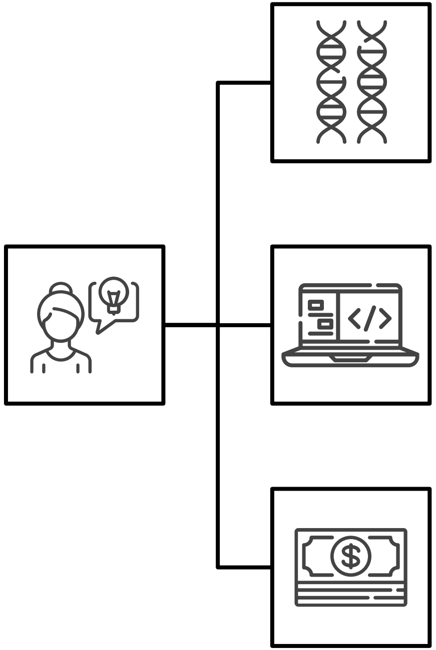
Srikar Chamala, Nicolás García, Grant T. Godden, Vivek Krishnakumar, Ingrid E. Jordon-Thaden, Riet De Smet, W. Brad Barbazuk, Douglas E. Soltis, Pamela S. Soltis

First published: 06 April 2015 | https://doi.org/10.3732/apps.1400115 | Citations: 32

† The authors thank all oneKP contributors, especially Gane Ka-Shu Wong, and BGI. Funding for S.C. was provided by the National Science Foundation (NSF; grant IOS-0922742 [P.S.S., D.E.S., W.B.B.]). Research funding for N.G. and G.T.G. was provided in part by NSF Doctoral Dissertation Improvement grants DEB-1310839 (P.S.S. and N.G.) and DEB-1210671 (P.S.S. and G.T.G.), respectively. The salary of I.J.T. was provided by the David Burpee Endowment and Chris Martine (Bucknell University).
‡ Authors are listed alphabetically by surname and contributed equally.



Resources needed



Human

Wet & Dry Labs

Genomic

Transcriptomes, ESTs, Genomes, Orthology DBs

Computational

Hardware & Software

Economic

For testing, production & optimization

Resources needed

Economic

For testing, production & optimization

Economic Resources

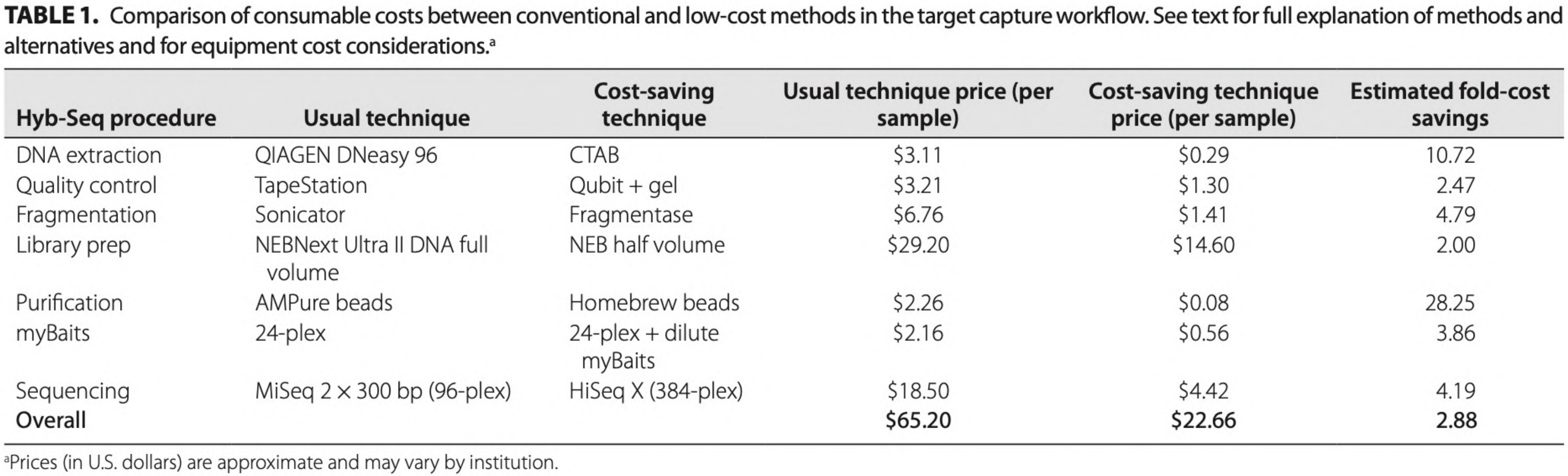


TABLE 1. Comparison of consumable costs between conventional and low-cost methods in the target capture workflow. See text for full explanation of methods and alternatives and for equipment cost considerations.[a]

Hyb-Seq procedure	Usual technique	Cost-saving technique	Usual technique price (per sample)	Cost-saving technique price (per sample)	Estimated fold-cost savings
DNA extraction	QIAGEN DNeasy 96	CTAB	\$3.11	\$0.29	10.72
Quality control	TapeStation	Qubit + gel	\$3.21	\$1.30	2.47
Fragmentation	Sonicator	Fragmentase	\$6.76	\$1.41	4.79
Library prep	NEBNext Ultra II DNA full volume	NEB half volume	\$29.20	\$14.60	2.00
Purification	AMPure beads	Homebrew beads	\$2.26	\$0.08	28.25
myBaits	24-plex	24-plex + dilute myBaits	\$2.16	\$0.56	3.86
Sequencing	MiSeq 2 × 300 bp (96-plex)	HiSeq X (384-plex)	\$18.50	\$4.42	4.19
Overall			**\$65.20**	**\$22.66**	**2.88**

[a] Prices (in U.S. dollars) are approximate and may vary by institution.

Table 1. Hale et al. 2019. *Apps. Plant Sci.* 8(4): e11337.

Economic Resources

Duke University
One Thousand Plant Transcriptomes (1KP, Alberta, Musea Ventures, BGI, etc.)

RJB-CSIC
AFFLORA & BAYESNEXT (State Research Agency, AEI)
BOREOTROP (Community of Madrid)

RBGK
Global Tree Seed Bank (GTSB) Programme (Garfield Weston Foundation)
Plant and Fungal Trees of Life (PAFTOL) Programme (Calleva & GW Foundations)

CBGP UPM-INIA/CSIC
Severo Ochoa Excellence Program (AEI)

IBB CSIC-MCNB
GIANTS & CANTHAROMICS (AEI)

Probe Set Design

Custom Kits

Euphorbiaceae431
Dioscoreaceae303
Cortinariaceae193
Etc.

Universal Kits

Angiosperms353
ProkaryoKit
Etc.

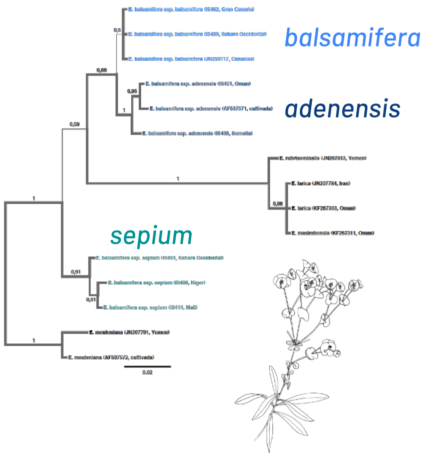
Euphorbiaceae431

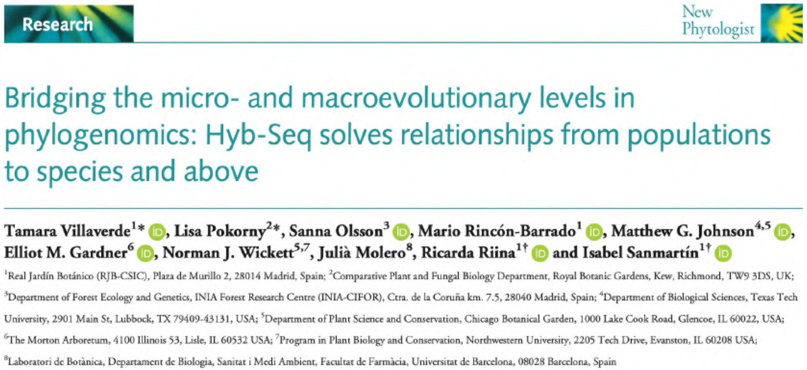
Classification: order Malpighiales (Eudicot), ~7.5K species (~300 genera)

Importance: rubber (latex), cassava (edible tuber), castor oil, etc.

Available resources:
2 *Euphorbia* transcriptomes 1KP
1 *Ricinus communis* genome
MarkerMiner Locus development made easy & Plaza DB
Daicel Arbor Biosciences myBaits®

Resulting kit:
431 ~single copy nuclear orthologs
~700 Kbp matrix
~20K 120-mer tiled probes (variable tiling: T 2x y G 1,5x)




Research

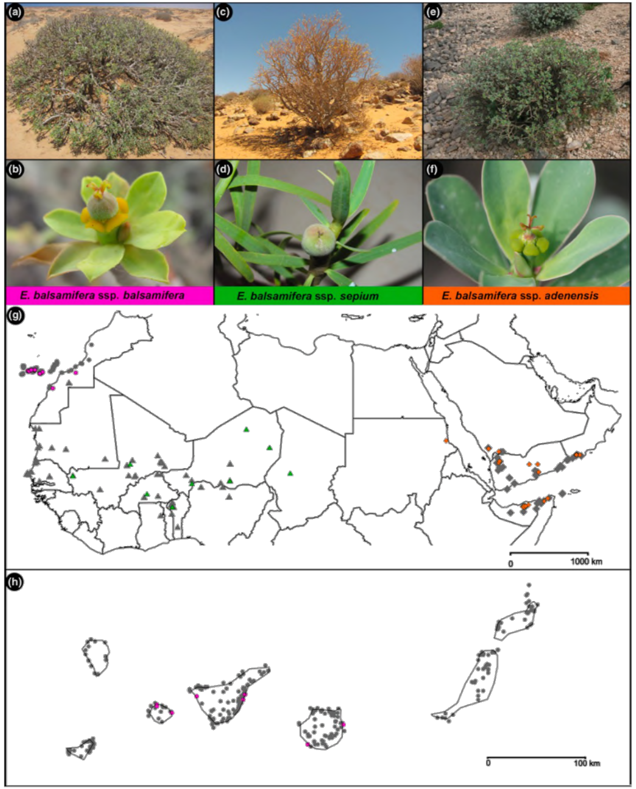
New Phytologist

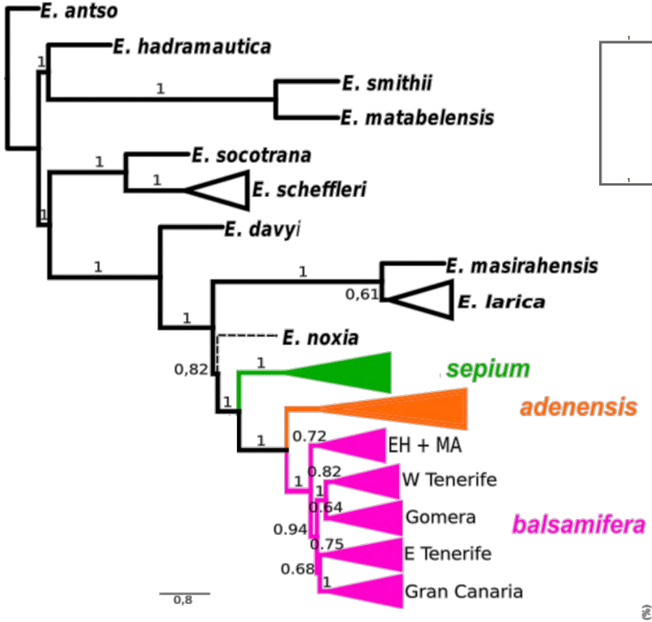
Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above

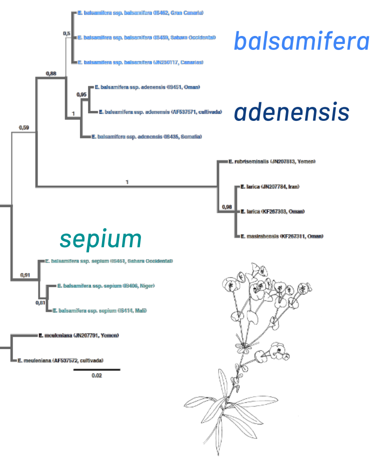
Tamara Villaverde[1]*, Lisa Pokorny[2]*, Sanna Olsson[3], Mario Rincón-Barrado[1], Matthew G. Johnson[4,5], Elliot M. Gardner[6], Norman J. Wickett[5,7], Julià Molero[8], Ricarda Riina[1†] and Isabel Sanmartín[1†]

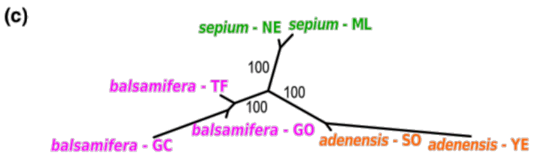
[1]Real Jardín Botánico (RJB-CSIC), Plaza de Murillo 2, 28014 Madrid, Spain; [2]Comparative Plant and Fungal Biology Department, Royal Botanic Gardens, Kew, Richmond, TW9 3DS, UK; [3]Department of Forest Ecology and Genetics, INIA Forest Research Centre (INIA-CIFOR), Ctra. de la Coruña km. 7.5, 28040 Madrid, Spain; [4]Department of Biological Sciences, Texas Tech University, 2901 Main St, Lubbock, TX 79409-43131, USA; [5]Department of Plant Science and Conservation, Chicago Botanical Garden, 1000 Lake Cook Road, Glencoe, IL 60022, USA; [6]The Morton Arboretum, 4100 Illinois 53, Lisle, IL 60532 USA; [7]Program in Plant Biology and Conservation, Northwestern University, 2205 Tech Drive, Evanston, IL 60208 USA; [8]Laboratori de Botànica, Departament de Biologia, Sanitat i Medi Ambient, Facultat de Farmàcia, Universitat de Barcelona, 08028 Barcelona, Spain

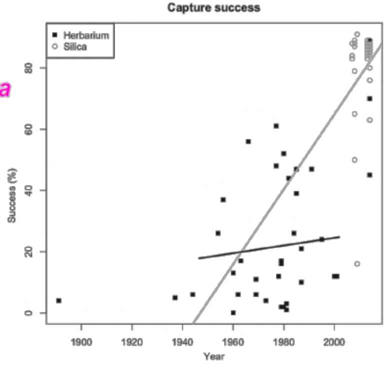
Euphorbiaceae431











Dioscoreaceae303

Classification: order Dioscoreales (Monocot), ~300 species (~4 genera)

Importance: yams (edible tuber), steroidal saponins (contraceptives), etc.

Available resources: 1KP
3 *Dioscorea* & 1 Tacca transcriptomes
1 *Oryza sativa* & 1 *Xerophyta viscosa*
MarkerMiner Locus development made easy & Plaza DB
Daicel Arbor Biosciences myBaits®

Resulting kit:
260 ~single copy nuclear orthologs
43 genes of agronomic interest
~442 Kbp matrix
~17K 120-mer 3x tiled probes

Applications in Plant Sciences

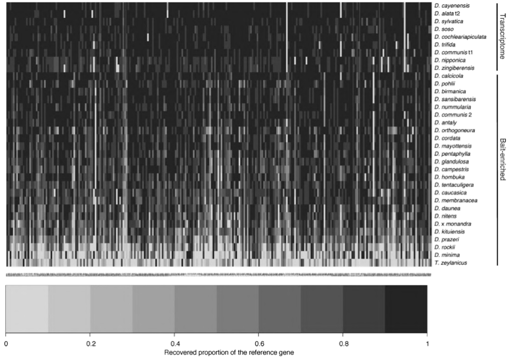
APPLICATION ARTICLE

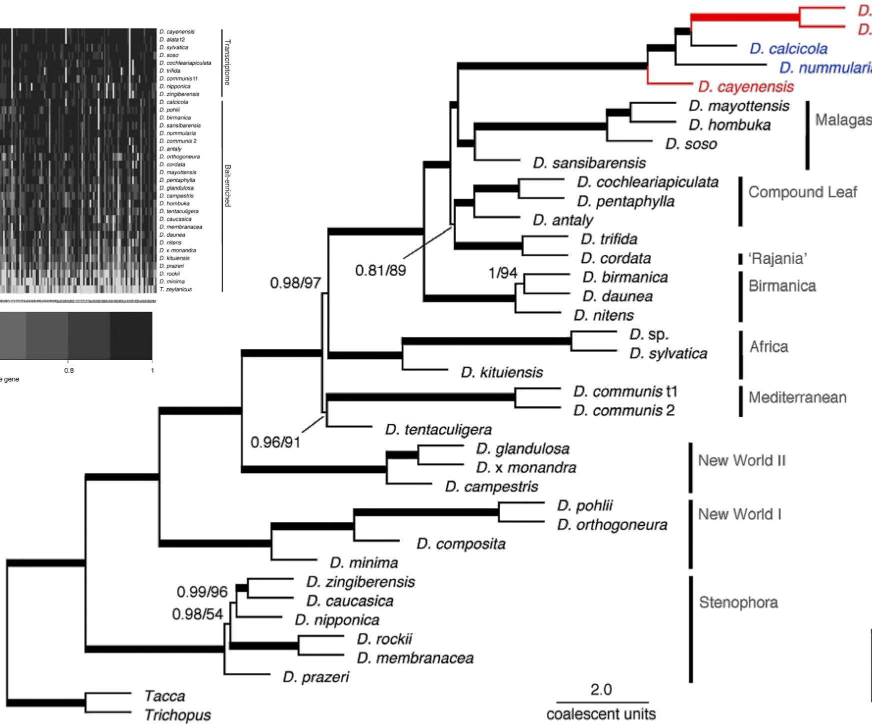
Check for updates

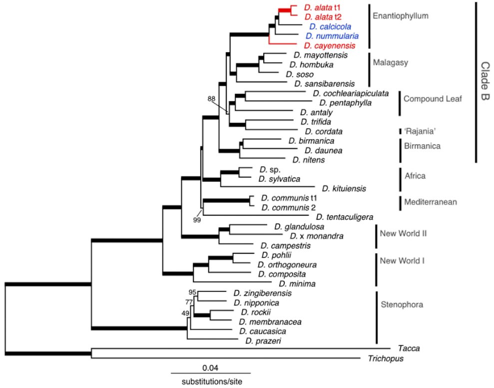
A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* yams (Dioscoreaceae)

Marybel Soto Gomez[1,2,8], Lisa Pokorny[3], Michael B. Kantar[4], Félix Forest[3], Ilia J. Leitch[3], Barbara Gravendeel[5,6,7], Paul Wilkin[3], Sean W. Graham[1,2], and Juan Viruel[3]

Dioscoreaceae303







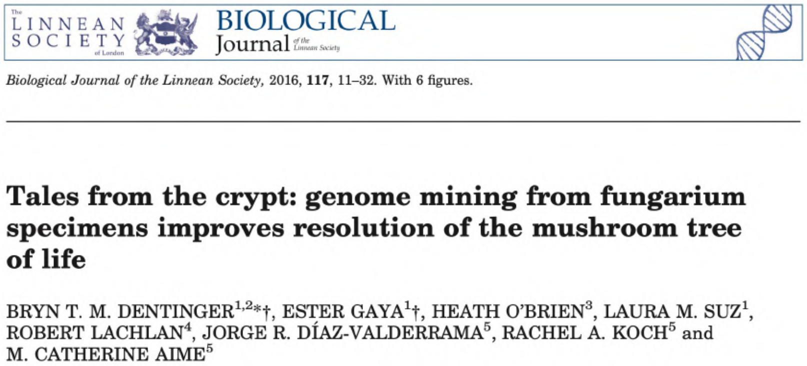
Cortinariaceae193

Classification: order Agaricales (Basidiomycota), ~3K species (1 genus)

Importance: ectomycorrhizal fungi

Available resources:
4 *Cortinarius* assemblies (ABySS)
208 Agaricales single-copy orthologs genes (SCOGs) & exonerate
Daicel Arbor Biosciences myBaits

Resulting kit:
188 nuclear SCOGs
RPB1, RPB2, MCM7, GPD, & TEF1
~67.5 Kbp matrix
~20K 120-mer 2x tiled probes

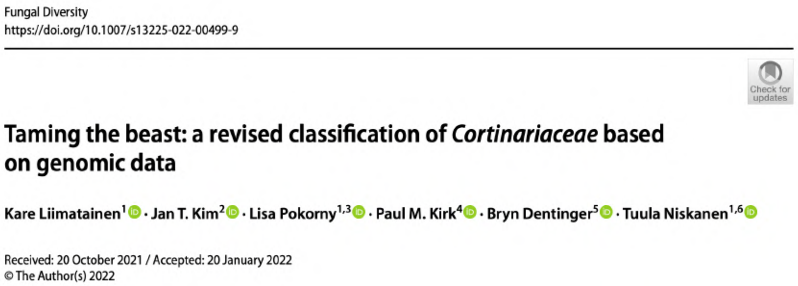


The LINNEAN SOCIETY of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2016, **117**, 11–32. With 6 figures.

Tales from the crypt: genome mining from fungarium specimens improves resolution of the mushroom tree of life

BRYN T. M. DENTINGER[1,2]*†, ESTER GAYA[1]†, HEATH O'BRIEN[3], LAURA M. SUZ[1], ROBERT LACHLAN[4], JORGE R. DÍAZ-VALDERRAMA[5], RACHEL A. KOCH[5] and M. CATHERINE AIME[5]



Fungal Diversity
https://doi.org/10.1007/s13225-022-00499-9

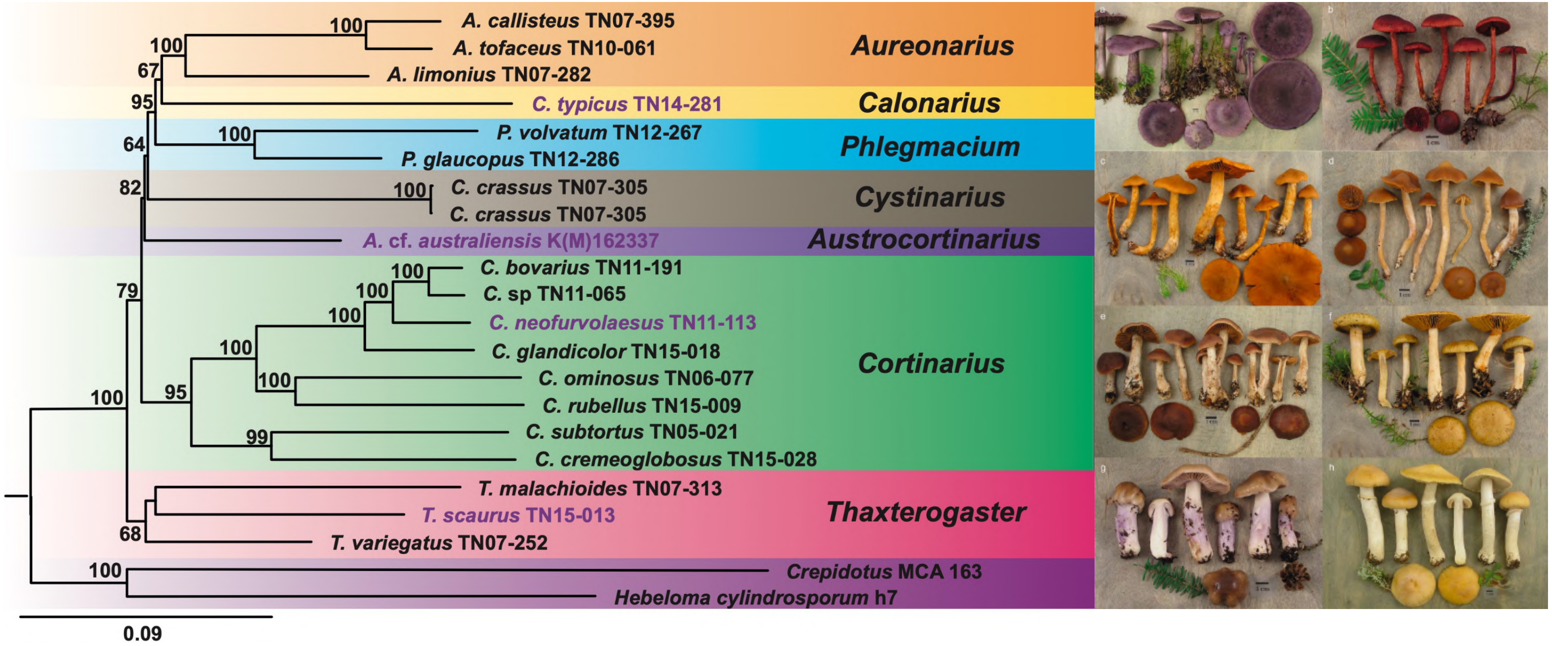
Check for updates

Taming the beast: a revised classification of *Cortinariaceae* based on genomic data

Kare Liimatainen[1] · Jan T. Kim[2] · Lisa Pokorny[1,3] · Paul M. Kirk[4] · Bryn Dentinger[5] · Tuula Niskanen[1,6]

Received: 20 October 2021 / Accepted: 20 January 2022
© The Author(s) 2022

Cortinariaceae193



Angiosperms353

Classification: flowering plants, ~ 370K spp. (APG VI: 416 fams. & 64 orders)

Importance: agronomic, edible, timber, pulp, fiber, meds, ornamental, fuel, etc.

Available resources:
655 transcriptomes 1KP
20 genomes Phytozome
Hamming dist., k-medoids clustering
Daicel Arbor Biosciences myBaits

Resulting kit:
353 nuclear SCOGs
~260,8 Kbp matrix
~75,2K 120-mer 3x tiled probes

Article

One thousand plant transcriptomes and the phylogenomics of green plants

https://doi.org/10.1038/s41586-019-1693-2

One Thousand Plant Transcriptomes Initiative

Syst. Biol. 68(4):594–606, 2019
© The Author(s) 2018. Published by Oxford University Press on behalf of the Society of Systematic Biologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
DOI:10.1093/sysbio/syy086
Advance Access publication December 10, 2018

A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering

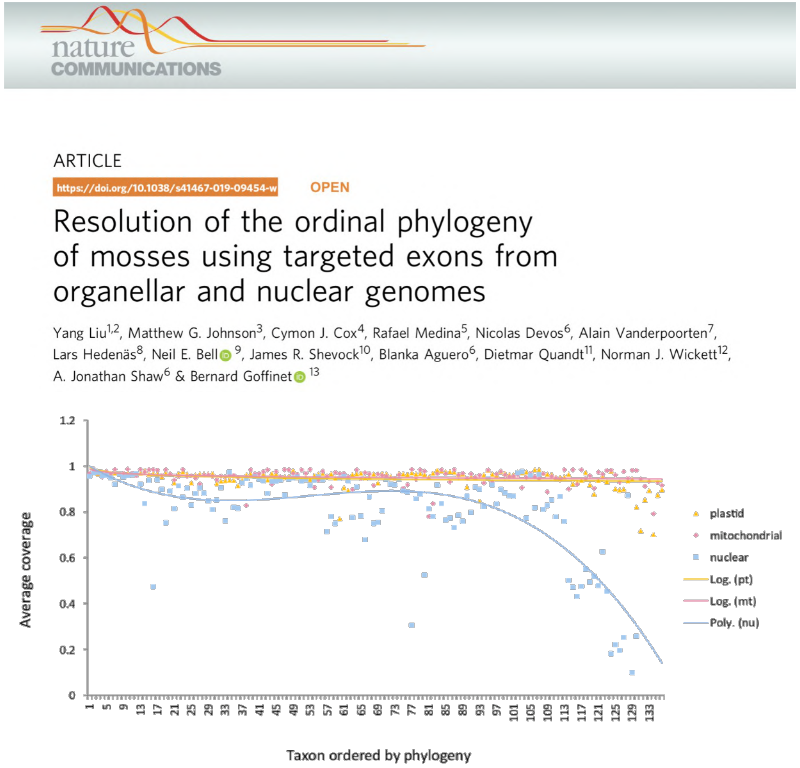
Matthew G. Johnson[1,2,*], Lisa Pokorny[3], Steven Dodsworth[3,4], Laura R. Botigué[3,5], Robyn S. Cowan[3], Alison Devault[6], Wolf L. Eiserhardt[3,7], Niroshini Epitawalage[3], Félix Forest[3], Jan T. Kim[3], James H. Leebens-Mack[8], Ilia J. Leitch[3], Olivier Maurin[3], Douglas E. Soltis[9,10], Pamela S. Soltis[9,10], Gane Ka-Shu Wong[11,12,13], William J. Baker[3], and Norman J. Wickett[2,14]

[1]*Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA;* [2]*Plant Science and Conservation, Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, USA;* [3]*Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, UK;* [4]*School of Life Sciences, University of Bedfordshire, University Square, Luton LU1 3JU, UK;* [5]*Centre for Research in Agricultural Genomics, Campus UAB, Edifici CRAG, Bellaterra Cerdanyola del Vallès, 08193 Barcelona, Spain;* [6]*Arbor Biosciences, 5840 Interface Dr, Suite 101, Ann Arbor, MI 48103, USA;* [7]*Department of Bioscience, Aarhus University, 8000 Aarhus C, Denmark;* [8]*Department of Plant Biology, University of Georgia, 2502 Miller Plant Sciences, Athens, GA 30602, USA;* [9]*Department of Biology, University of Florida, 220 Bartram Hall, Gainesville, FL 32611-8525, USA;* [10]*Florida Museum of Natural History, University of Florida, 3215 Hull Road, Gainesville, FL 32611-2710, USA;* [11]*BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China;* [12]*Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada;* [13]*Department of Medicine, University of Alberta, Edmonton, AB T6G 2E1, Canada; and* [14]*Program in Plant Biology and Conservation, Northwestern University, 2205 Tech Drive, Evanston, IL 60208, USA*

**Correspondence to be sent to: Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA;*
E-mail: matt.johnson@ttu.edu.

Matthew G. Johnson, Lisa Pokorny, Steven Dodsworth contributed equally to this article.

Angiosperms353



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09454-w OPEN

Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes

Yang Liu[1,2], Matthew G. Johnson[3], Cymon J. Cox[4], Rafael Medina[5], Nicolas Devos[6], Alain Vanderpoorten[7], Lars Hedenäs[8], Neil E. Bell[9], James R. Shevock[10], Blanka Aguero[6], Dietmar Quandt[11], Norman J. Wickett[12], A. Jonathan Shaw[6] & Bernard Goffinet[13]

Article

One thousand plant transcriptomes and the phylogenomics of green plants

https://doi.org/10.1038/s41586-019-1693-2

One Thousand Plant Transcriptomes Initiative

Syst. Biol. 68(4):594–606, 2019

© The Author(s) 2018. Published by Oxford University Press on behalf of the Society of Systematic Biologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

DOI:10.1093/sysbio/syy086
Advance Access publication December 10, 2018

A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering

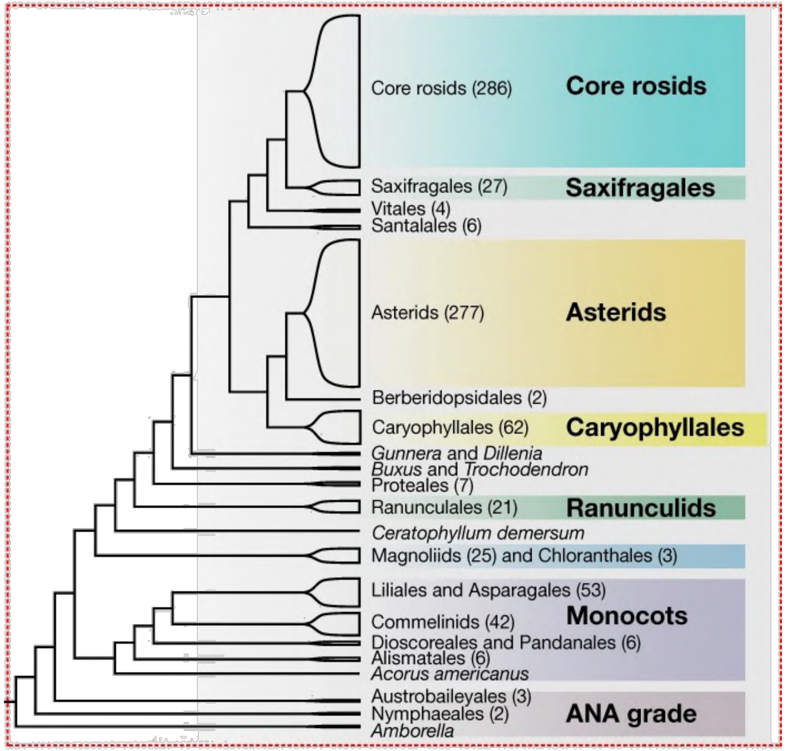
MATTHEW G. JOHNSON[1,2,*], LISA POKORNY[3], STEVEN DODSWORTH[3,4], LAURA R. BOTIGUÉ[3,5], ROBYN S. COWAN[3], ALISON DEVAULT[6], WOLF L. EISERHARDT[3,7], NIROSHINI EPITAWALAGE[3], FÉLIX FOREST[3], JAN T. KIM[3], JAMES H. LEEBENS-MACK[8], ILIA J. LEITCH[3], OLIVIER MAURIN[3], DOUGLAS E. SOLTIS[9,10], PAMELA S. SOLTIS[9,10], GANE KA-SHU WONG[11,12,13], WILLIAM J. BAKER[3], AND NORMAN J. WICKETT[2,14]

[1]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA; [2]Plant Science and Conservation, Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, USA; [3]Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, UK; [4]School of Life Sciences, University of Bedfordshire, University Square, Luton LU1 3JU, UK; [5]Centre for Research in Agricultural Genomics, Campus UAB, Edifici CRAG, Bellaterra Cerdanyola del Vallès, 08193 Barcelona, Spain; [6]Arbor Biosciences, 5840 Interface Dr, Suite 101, Ann Arbor, MI 48103, USA; [7]Department of Bioscience, Aarhus University, 8000 Aarhus C, Denmark; [8]Department of Plant Biology, University of Georgia, 2502 Miller Plant Sciences, Athens, GA 30602, USA; [9]Department of Biology, University of Florida, 220 Bartram Hall, Gainesville, FL 32611-8525, USA; [10]Florida Museum of Natural History, University of Florida, 3215 Hull Road, Gainesville, FL 32611-2710, USA; [11]BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China; [12]Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada; [13]Department of Medicine, University of Alberta, Edmonton, AB T6G 2E1, Canada; and [14]Program in Plant Biology and Conservation, Northwestern University, 2205 Tech Drive, Evanston, IL 60208, USA

**Correspondence to be sent to: Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA; E-mail: matt.johnson@ttu.edu.*

Matthew G. Johnson, Lisa Pokorny, Steven Dodsworth contributed equally to this article.

Angiosperms353



Article

One thousand plant transcriptomes and the phylogenomics of green plants

https://doi.org/10.1038/s41586-019-1693-2

One Thousand Plant Transcriptomes Initiative

Syst. Biol. 68(4):594–606, 2019
© The Author(s) 2018. Published by Oxford University Press on behalf of the Society of Systematic Biologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
DOI:10.1093/sysbio/syy086
Advance Access publication December 10, 2018

A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering

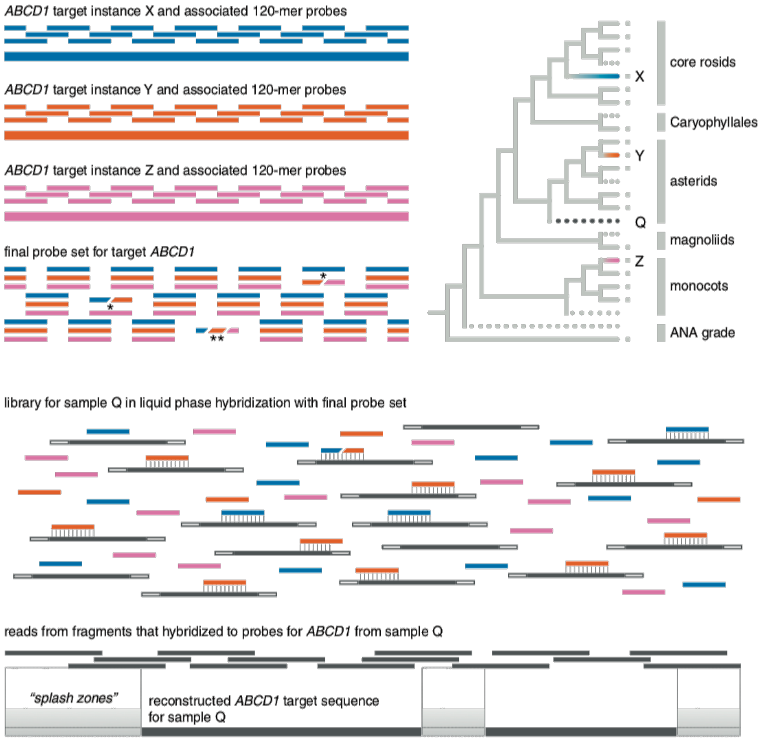
Matthew G. Johnson[1,2,*], Lisa Pokorny[3], Steven Dodsworth[3,4], Laura R. Botigué[3,5], Robyn S. Cowan[3], Alison Devault[6], Wolf L. Eiserhardt[3,7], Niroshini Epitawalage[3], Félix Forest[3], Jan T. Kim[3], James H. Leebens-Mack[8], Ilia J. Leitch[3], Olivier Maurin[3], Douglas E. Soltis[9,10], Pamela S. Soltis[9,10], Gane Ka-Shu Wong[11,12,13], William J. Baker[3], and Norman J. Wickett[2,14]

[1]*Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA;* [2]*Plant Science and Conservation, Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, USA;* [3]*Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, UK;* [4]*School of Life Sciences, University of Bedfordshire, University Square, Luton LU1 3JU, UK;* [5]*Centre for Research in Agricultural Genomics, Campus UAB, Edifici CRAG, Bellaterra Cerdanyola del Vallès, 08193 Barcelona, Spain;* [6]*Arbor Biosciences, 5840 Interface Dr, Suite 101, Ann Arbor, MI 48103, USA;* [7]*Department of Bioscience, Aarhus University, 8000 Aarhus C, Denmark;* [8]*Department of Plant Biology, University of Georgia, 2502 Miller Plant Sciences, Athens, GA 30602, USA;* [9]*Department of Biology, University of Florida, 220 Bartram Hall, Gainesville, FL 32611-8525, USA;* [10]*Florida Museum of Natural History, University of Florida, 3215 Hull Road, Gainesville, FL 32611-2710, USA;* [11]*BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China;* [12]*Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada;* [13]*Department of Medicine, University of Alberta, Edmonton, AB T6G 2E1, Canada; and* [14]*Program in Plant Biology and Conservation, Northwestern University, 2205 Tech Drive, Evanston, IL 60208, USA*

**Correspondence to be sent to: Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA;*
E-mail: matt.johnson@ttu.edu.

Matthew G. Johnson, Lisa Pokorny, Steven Dodsworth contributed equally to this article.

Angiosperms353



Syst. Biol. 68(4):594–606, 2019

© The Author(s) 2018. Published by Oxford University Press on behalf of the Society of Systematic Biologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

DOI:10.1093/sysbio/syy086
Advance Access publication December 10, 2018

A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering

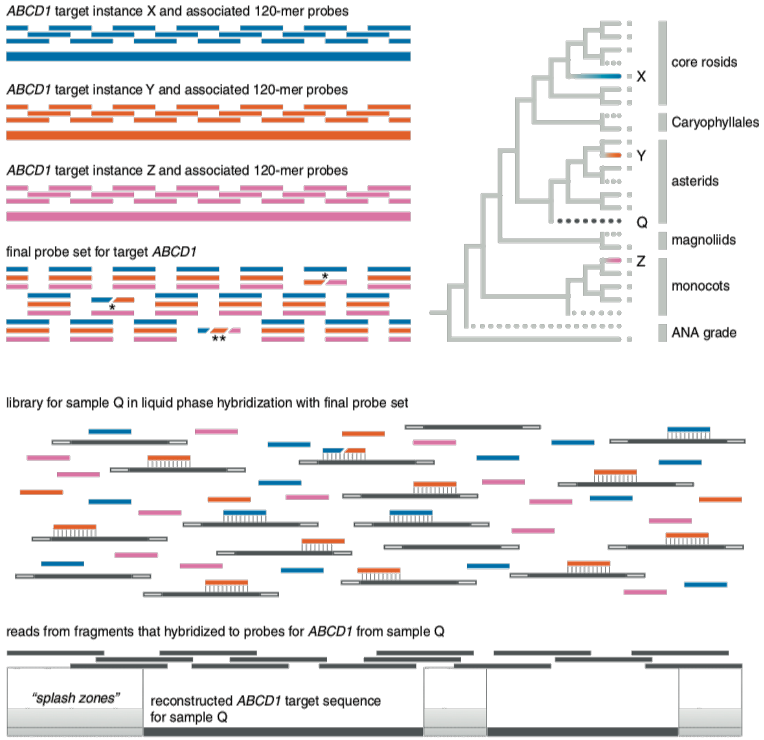
MATTHEW G. JOHNSON[1,2,*], LISA POKORNY[3], STEVEN DODSWORTH[3,4], LAURA R. BOTIGUÉ[3,5], ROBYN S. COWAN[3], ALISON DEVAULT[6], WOLF L. EISERHARDT[3,7], NIROSHINI EPITAWALAGE[3], FÉLIX FOREST[3], JAN T. KIM[3], JAMES H. LEEBENS-MACK[8], ILIA J. LEITCH[3], OLIVIER MAURIN[3], DOUGLAS E. SOLTIS[9,10], PAMELA S. SOLTIS[9,10], GANE KA-SHU WONG[11,12,13], WILLIAM J. BAKER[3], AND NORMAN J. WICKETT[2,14]

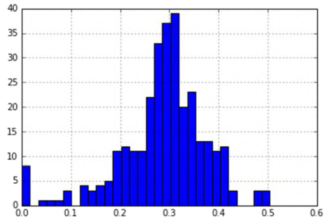
[1]*Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA;* [2]*Plant Science and Conservation, Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, USA;* [3]*Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, UK;* [4]*School of Life Sciences, University of Bedfordshire, University Square, Luton LU1 3JU, UK;* [5]*Centre for Research in Agricultural Genomics, Campus UAB, Edifici CRAG, Bellaterra Cerdanyola del Vallès, 08193 Barcelona, Spain;* [6]*Arbor Biosciences, 5840 Interface Dr, Suite 101, Ann Arbor, MI 48103, USA;* [7]*Department of Bioscience, Aarhus University, 8000 Aarhus C, Denmark;* [8]*Department of Plant Biology, University of Georgia, 2502 Miller Plant Sciences, Athens, GA 30602, USA;* [9]*Department of Biology, University of Florida, 220 Bartram Hall, Gainesville, FL 32611-8525, USA;* [10]*Florida Museum of Natural History, University of Florida, 3215 Hull Road, Gainesville, FL 32611-2710, USA;* [11]*BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China;* [12]*Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada;* [13]*Department of Medicine, University of Alberta, Edmonton, AB T6G 2E1, Canada; and* [14]*Program in Plant Biology and Conservation, Northwestern University, 2205 Tech Drive, Evanston, IL 60208, USA*

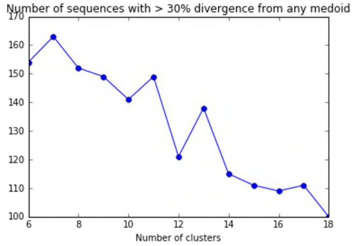
**Correspondence to be sent to: Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA;*
E-mail: matt.johnson@ttu.edu.

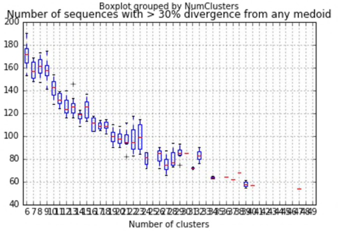
Matthew G. Johnson, Lisa Pokorny, Steven Dodsworth contributed equally to this article.

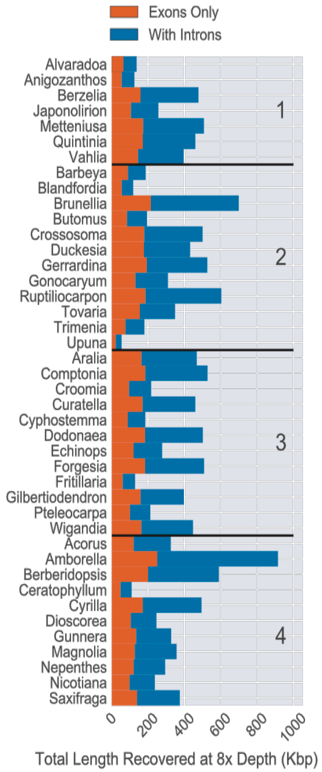
Angiosperms353











Angiosperms353

myBaits Expert — Predesigned Panels

myBaits Expert – Plant – Angiosperms-353

Home » Genomics » Targeted Sequencing » myBaits – Hyb Capture Kits » myBaits Expert — Predesigned Panels » myBaits Expert – Plant – Angiosperms-353

This new panel enriches 100's of single-copy genes orthologous across all angiosperms (flowering plants).

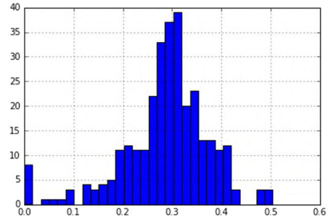
Combining the flexible hybridization power of in-solution target capture with an expertly selected set of orthologous locus sequences, this new probe set has been demonstrated to enrich hundreds of putatively single-copy protein-coding genes across a broad range of angiosperms (flowering plants). Probes were designed from 353 loci, each with 5-15 representative sequences from across all angiosperms which were selected using a novel "k-medoids clustering approach" to maximize taxonomic breadth of the design (**Johnson, Pokorny, Dodsworth et al 2018, Systematic Biology**).

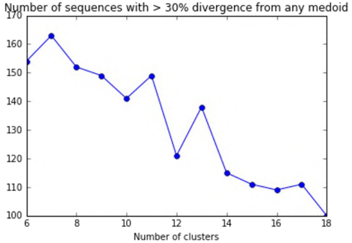
These probes are broadly applicable for phylogenetic research across all flowering plants, and the kit is available from Arbor Biosciences as an in-stock catalog kit available for immediate shipment at a low per-reaction cost. As with all myBaits kits, the Angiosperms353 panel is provided as a complete solution target capture kit, including buffers, blockers, and baits, along with an easy-to-use protocol. Or if you would prefer to outsource the work, our **myReads NGS service** expert scientists are available to perform library preparation, target capture, and sequencing for your entire project.

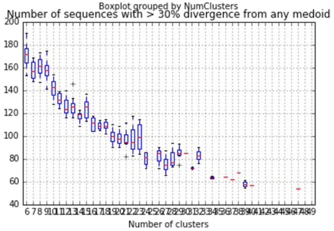
- Predesigned Panel – Designed by expert plant geneticists
- Proven results – Demonstrated utility across broad taxonomic range
- Freely accessible pipelines – Simple, user-friendly data analysis
- Cost-effective – Use one bait set across multiple distantly-related taxa

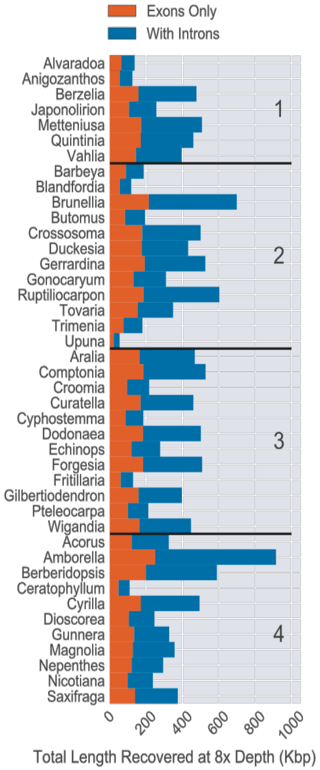
Ordering Information	Size	Cat#	Qty	Price	
myBaits Expert Angiosperms 353 v1	8 Reactions	308108	1	**$1000**	ADD TO CART

Available in-stock for immediate shipment.

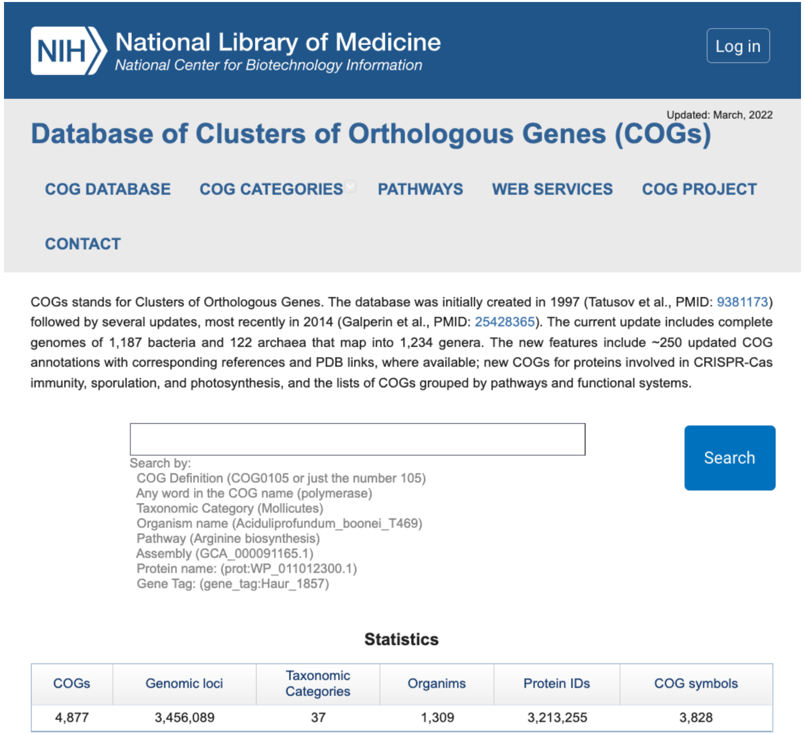








ProkarioKit



Updated: March, 2022

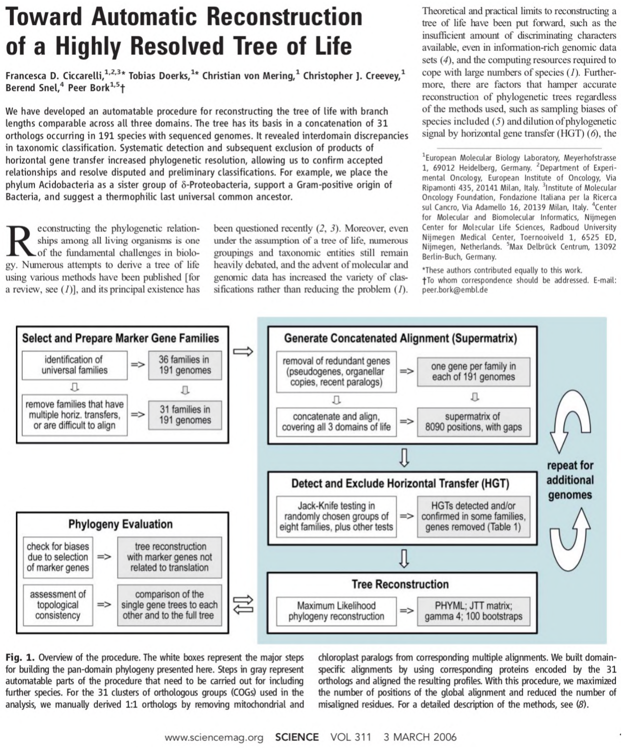
Database of Clusters of Orthologous Genes (COGs)

COGs stands for Clusters of Orthologous Genes. The database was initially created in 1997 (Tatusov et al., PMID: 9381173) followed by several updates, most recently in 2014 (Galperin et al., PMID: 25428365). The current update includes complete genomes of 1,187 bacteria and 122 archaea that map into 1,234 genera. The new features include ~250 updated COG annotations with corresponding references and PDB links, where available; new COGs for proteins involved in CRISPR-Cas immunity, sporulation, and photosynthesis, and the lists of COGs grouped by pathways and functional systems.

Search by:
- COG Definition (COG0105 or just the number 105)
- Any word in the COG name (polymerase)
- Taxonomic Category (Mollicutes)
- Organism name (Aciduliprofundum_boonei_T469)
- Pathway (Arginine biosynthesis)
- Assembly (GCA_000091165.1)
- Protein name: (prot:WP_011012300.1)
- Gene Tag: (gene_tag:Haur_1857)

Statistics

COGs	Genomic loci	Taxonomic Categories	Organims	Protein IDs	COG symbols
4,877	3,456,089	37	1,309	3,213,255	3,828



Toward Automatic Reconstruction of a Highly Resolved Tree of Life

Francesca D. Ciccarelli,[1,2,3]* Tobias Doerks,[1]* Christian von Mering,[1] Christopher J. Creevey,[1] Berend Snel,[4] Peer Bork[1,5]†

We have developed an automatable procedure for reconstructing the tree of life with branch lengths comparable across all three domains. The tree has its basis in a concatenation of 31 orthologs occurring in 191 species with sequenced genomes. It revealed interdomain discrepancies in taxonomic classification. Systematic detection and subsequent exclusion of products of horizontal gene transfer increased phylogenetic resolution, allowing us to confirm accepted relationships and resolve disputed and preliminary classifications. For example, we place the phylum Acidobacteria as a sister group of δ-Proteobacteria, support a Gram-positive origin of Bacteria, and suggest a thermophilic last universal common ancestor.

Reconstructing the phylogenetic relationships among all living organisms is one of the fundamental challenges in biology. Numerous attempts to derive a tree of life using various methods have been published [for a review, see (*1*)], and its principal existence has been questioned recently (*2*, *3*). Moreover, even under the assumption of a tree of life, numerous groupings and taxonomic entities still remain heavily debated, and the advent of molecular and genomic data has increased the variety of classifications rather than reducing the problem (*1*). Theoretical and practical limits to reconstructing a tree of life have been put forward, such as the insufficient amount of discriminating characters available, even in information-rich genomic data sets (*4*), and the computing resources required to cope with large numbers of species (*1*). Furthermore, there are factors that hamper accurate reconstruction of phylogenetic trees regardless of the methods used, such as sampling biases of species included (*5*) and dilution of phylogenetic signal by horizontal gene transfer (HGT) (*6*), the

---

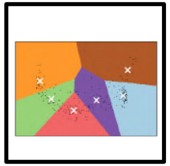
[1] European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany. [2] Department of Experimental Oncology, European Institute of Oncology, Via Ripamonti 435, 20141 Milan, Italy. [3] Institute of Molecular Oncology Foundation, Fondazione Italiana per la Ricerca sul Cancro, Via Adamello 16, 20139 Milan, Italy. [4] Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, Radboud University Nijmegen Medical Center, Toernooiveld 1, 6525 ED, Nijmegen, Netherlands. [5] Max Delbrück Centrum, 13092 Berlin-Buch, Germany.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: peer.bork@embl.de

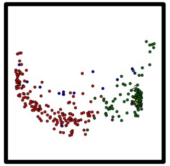
Fig. 1. Overview of the procedure. The white boxes represent the major steps for building the pan-domain phylogeny presented here. Steps in gray represent automatable parts of the procedure that need to be carried out for including further species. For the 31 clusters of orthologous groups (COGs) used in the analysis, we manually derived 1:1 orthologs by removing mitochondrial and chloroplast paralogs from corresponding multiple alignments. We built domain-specific alignments by using corresponding proteins encoded by the 31 orthologs and aligned the resulting profiles. With this procedure, we maximized the number of positions of the global alignment and reduced the number of misaligned residues. For a detailed description of the methods, see (*8*).

Clustering Algorithms



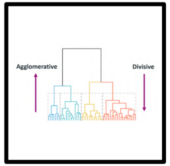
Centroid

Proximity, e.g., K-means, K-medoids



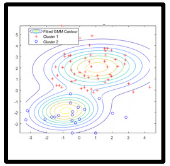
Graph

Distance, e.g., spectral (affinity propagation)



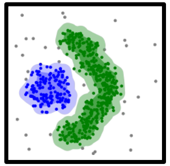
Connectivity

Agglomerative & Divisive (Hierarchical)



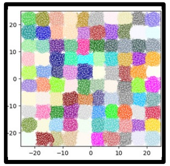
Distribution

Belonging, Gaussian Mixture Models



Density

DBSCAN, OPTICS, etc.



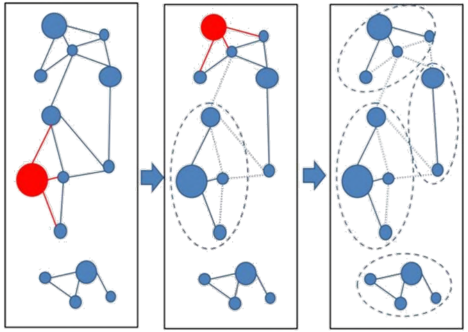
Compression

Dim. Transformation, e.g., BIRCH

Biological Sequence Clustering

CD-HIT

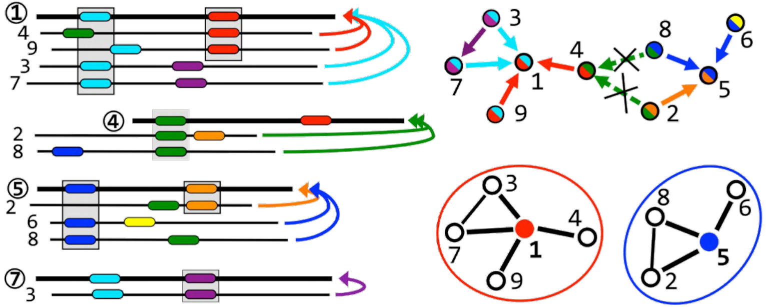
Greedy Incremental Clustering



https://github.com/soedinglab/mmseqs2/wiki#clustering-modes

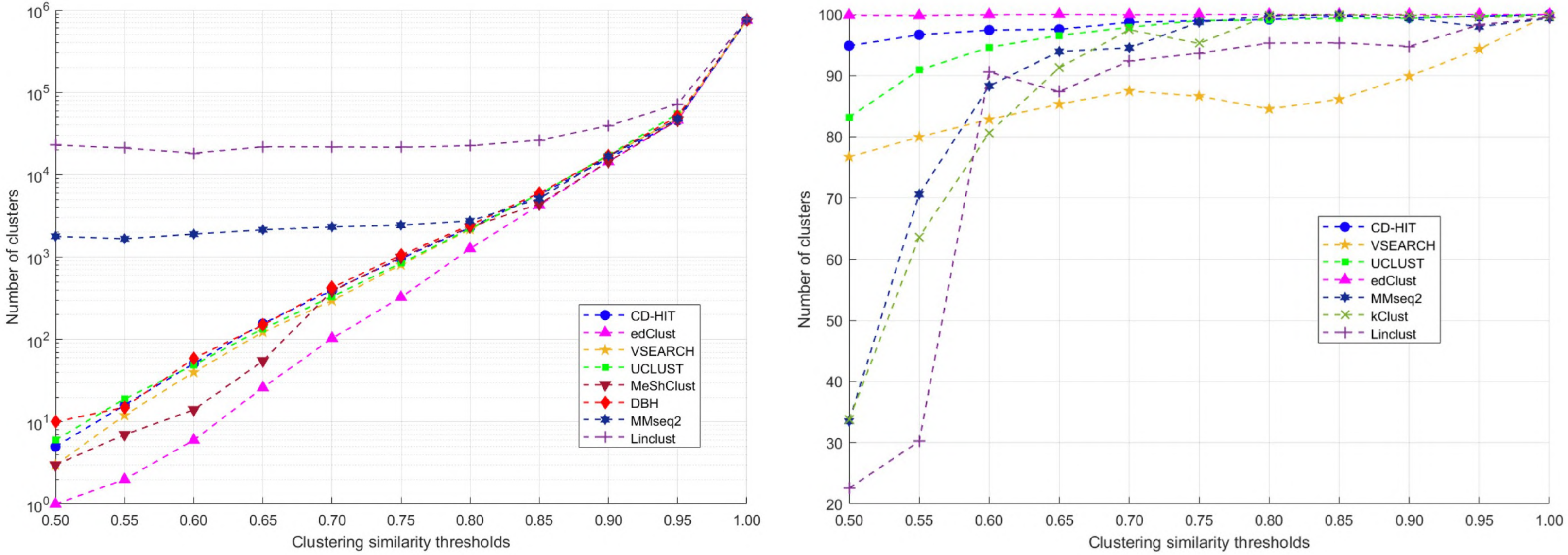
MMseqs2

Linear Time Clustering



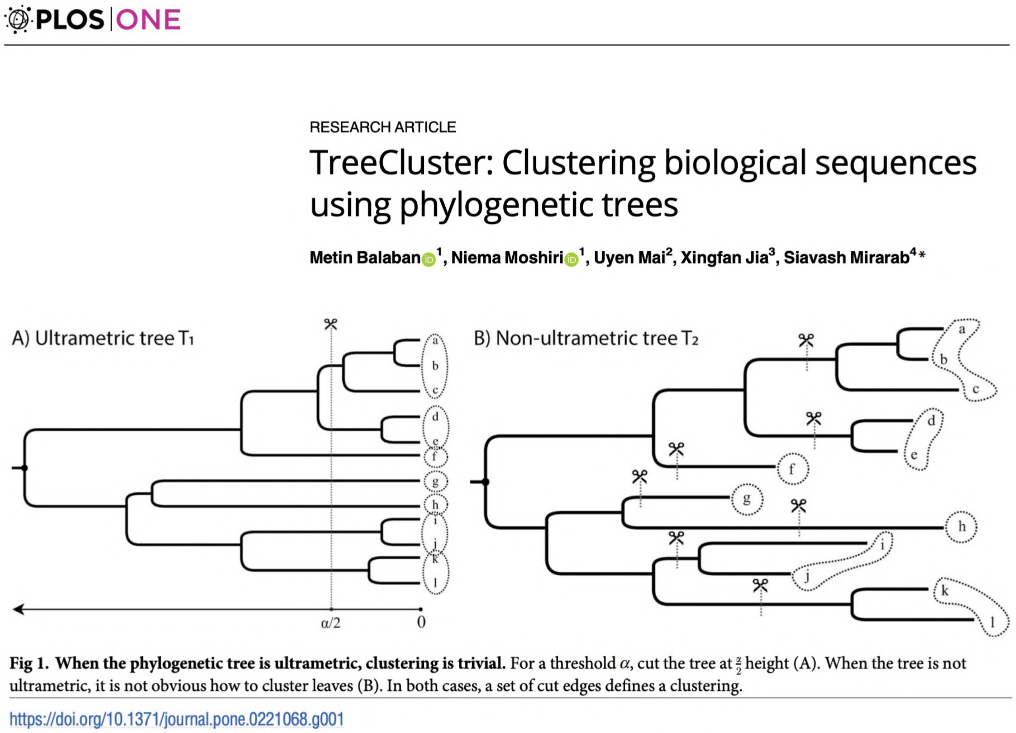
https://github.com/soedinglab/mmseqs2/wiki#linear-time-clustering-using-mmseqs-linclust

Biological Sequence Clustering



Figs. 3 & 7. Wei Z.-G. et al. 2023. Comparison of Methods for Biological Sequence Clustering. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 20(5): 2874–2888.

Tree-Based Clustering



PLOS | ONE

RESEARCH ARTICLE

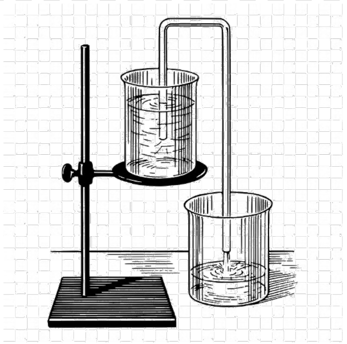
TreeCluster: Clustering biological sequences using phylogenetic trees

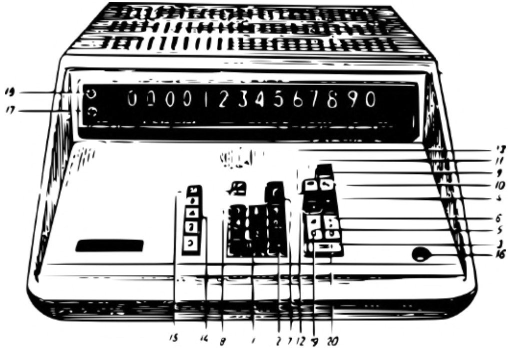
Metin Balaban[1], Niema Moshiri[1], Uyen Mai[2], Xingfan Jia[3], Siavash Mirarab[4]*

Fig 1. When the phylogenetic tree is ultrametric, clustering is trivial. For a threshold α, cut the tree at $\frac{\alpha}{2}$ height (A). When the tree is not ultrametric, it is not obvious how to cluster leaves (B). In both cases, a set of cut edges defines a clustering.

https://doi.org/10.1371/journal.pone.0221068.g001

Table of contents





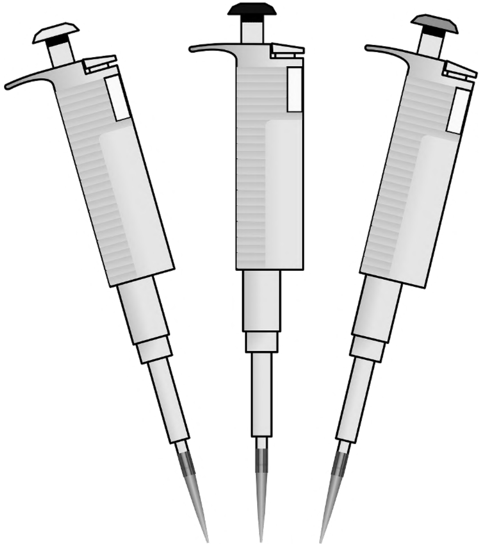
Implementing TCS

Wet Lab Workflow & Dry Lab Pipeline

03

Wet Lab Workflow

Generating TCS data

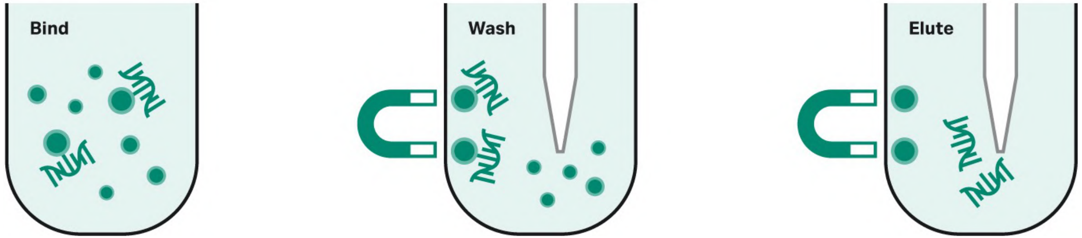


Wet Lab Workflow

DNA Extraction (modified CTAB) & Purification
(cleaning w/ magnetic beads)

Tissue Template

Purified DNA Template



Magnetic particles are added to sample and bind to target molecule

Magnetic particles are captured and remainder of sample is washed away

Target molecule is released from magnetic particles for further analysis

Dioscorea tomentosa, 1796

Wet Lab Workflow

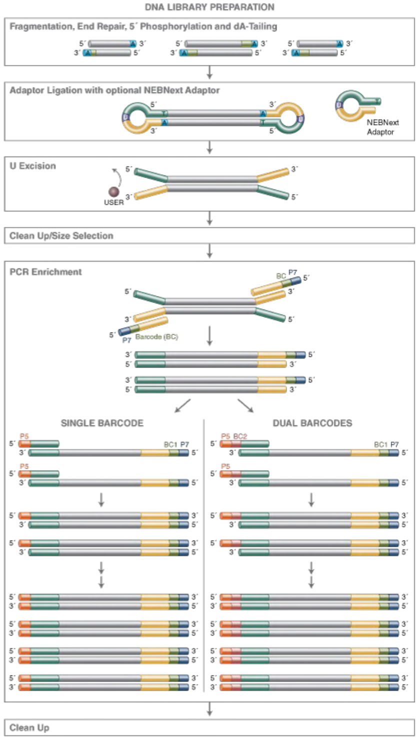
DNA Extraction (modified CTAB) & Purification (cleaning w/ magnetic beads)

Genomic Library Preparation (commercially available kits for dual-indexed sequencing)

Tissue Template

Purified DNA Template

Dual-indexed Genomic Libraries



Wet Lab Workflow

DNA Extraction (modified CTAB) & Purification (cleaning w/ magnetic beads)

Genomic Library Preparation (commercially available kits for dual-indexed sequencing)

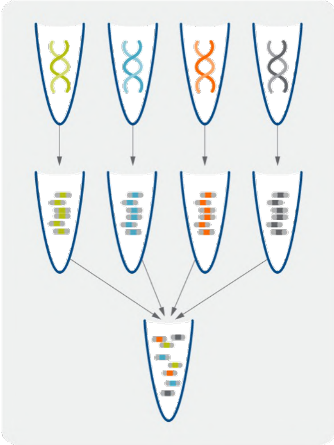
Hybridization (biotinylated probes), Capture (streptavidin-coated beads) & Enrichment (PCR)

Tissue Template

Purified DNA Template

Dual-indexed Genomic Libraries

Target-enriched Genomic Libraries



Wet Lab Workflow

DNA Extraction (modified CTAB) & Purification (cleaning w/ magnetic beads)

Genomic Library Preparation (commercially available kits for dual-indexed sequencing)

Hybridization (biotinylated probes), Capture (streptavidin-coated beads) & Enrichment (PCR)

High-Throughput Sequencing (HTS, multiplexed enriched dual-indexed libraries)

Tissue Template

Purified DNA Template

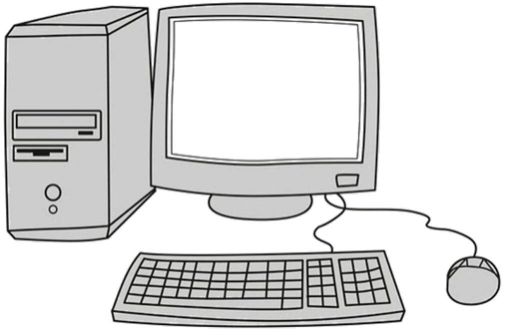
Dual-indexed Genomic Libraries

Target-enriched Genomic Libraries

Raw Reads

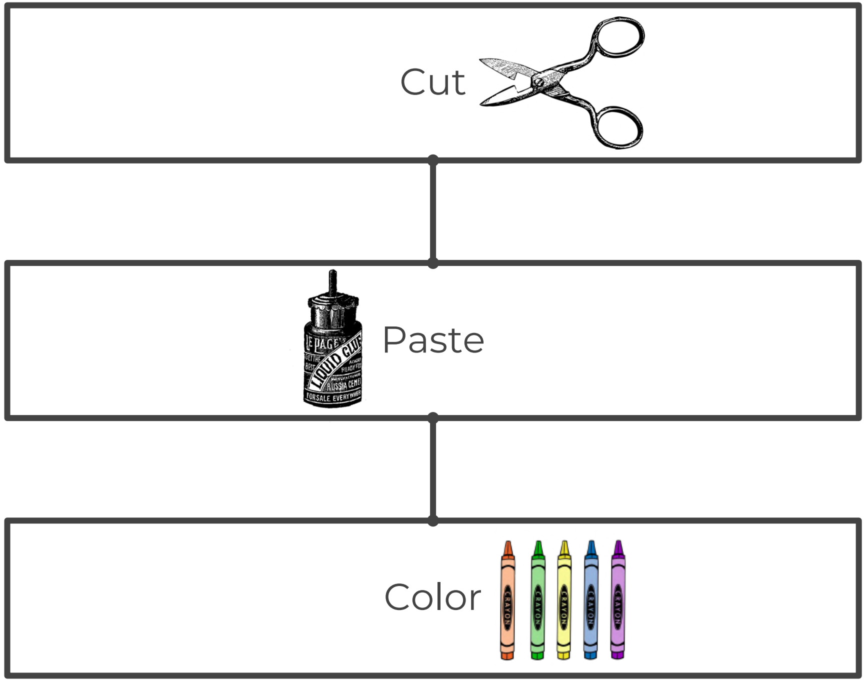
Dry Lab Pipeline

Processing TCS data



Dry Lab Pipeline

Raw
Reads

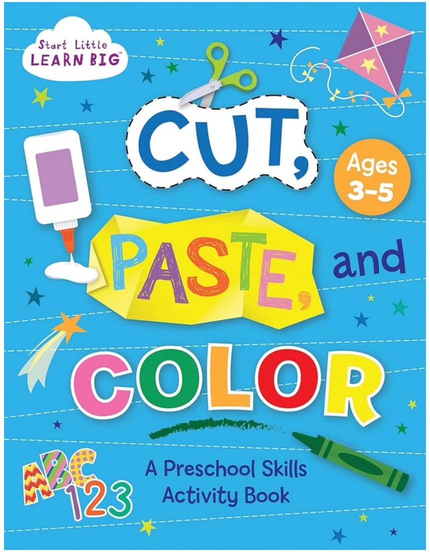


Cut

Paste

Color

Target
sequences



Dry Lab Pipeline

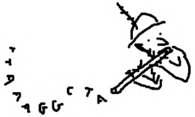
Raw Reads

FastQC | MultiQC & Trimmomatic | fastp

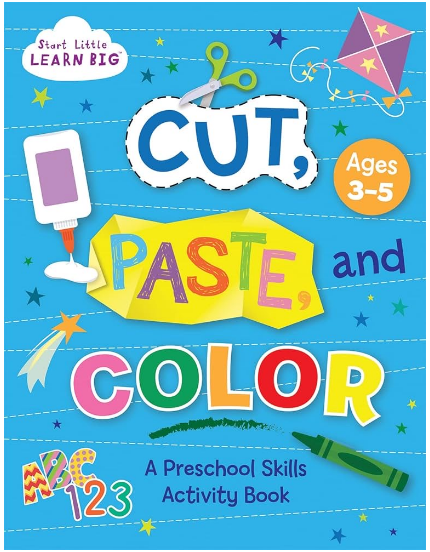
BLAST | BWA | DIAMOND

SPAdes | exon(intron)erate

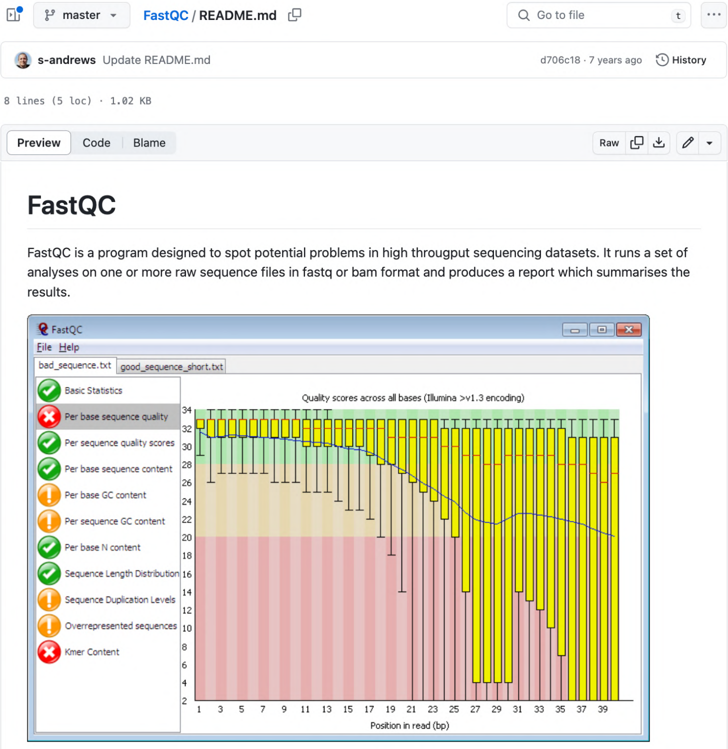
Target sequences



HybPiper



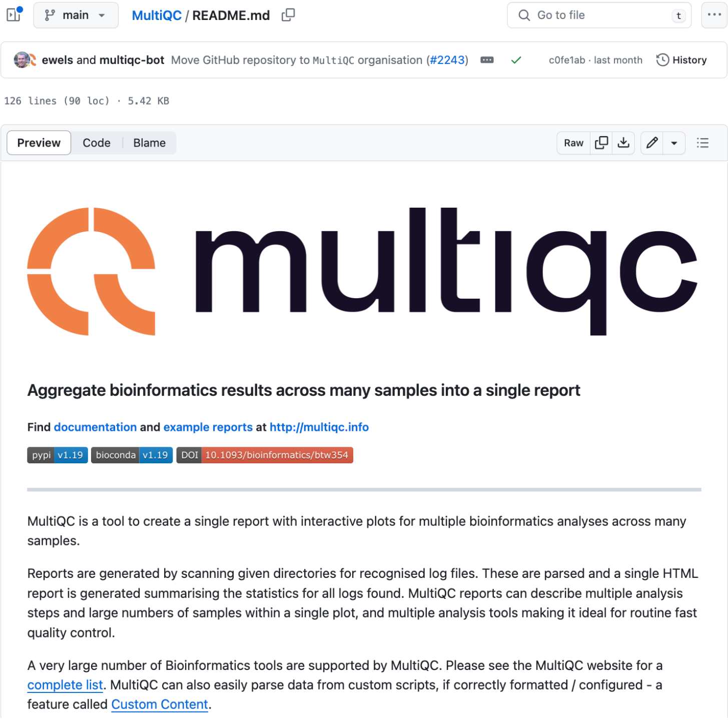
FastQC



FastQC

FastQC is a program designed to spot potential problems in high througput sequencing datasets. It runs a set of analyses on one or more raw sequence files in fastq or bam format and produces a report which summarises the results.

MultiQC



Aggregate bioinformatics results across many samples into a single report

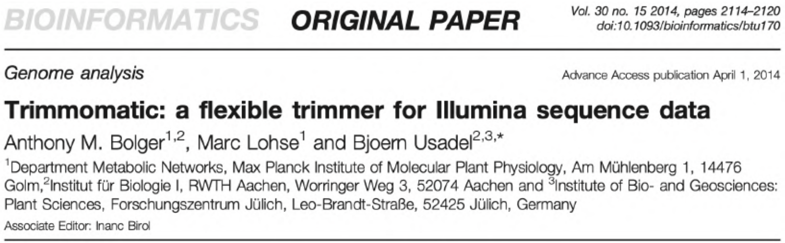
Find documentation and example reports at http://multiqc.info

MultiQC is a tool to create a single report with interactive plots for multiple bioinformatics analyses across many samples.

Reports are generated by scanning given directories for recognised log files. These are parsed and a single HTML report is generated summarising the statistics for all logs found. MultiQC reports can describe multiple analysis steps and large numbers of samples within a single plot, and multiple analysis tools making it ideal for routine fast quality control.

A very large number of Bioinformatics tools are supported by MultiQC. Please see the MultiQC website for a complete list. MultiQC can also easily parse data from custom scripts, if correctly formatted / configured - a feature called Custom Content.

trimmomatic



BIOINFORMATICS **ORIGINAL PAPER**

Vol. 30 no. 15 2014, pages 2114–2120
doi:10.1093/bioinformatics/btu170

Genome analysis

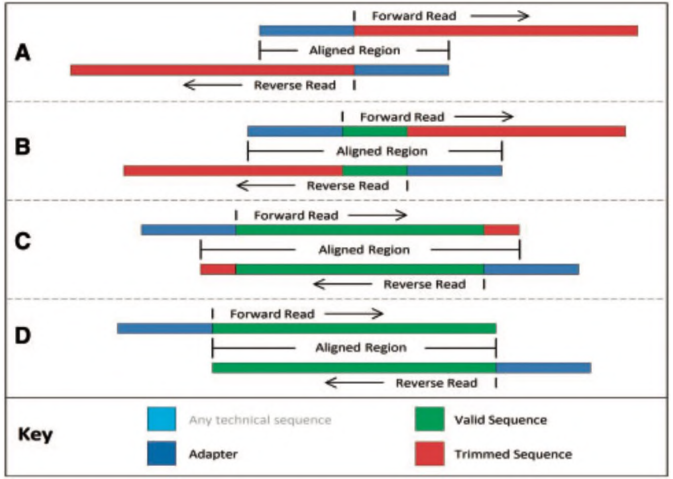
Advance Access publication April 1, 2014

Trimmomatic: a flexible trimmer for Illumina sequence data

Anthony M. Bolger[1,2], Marc Lohse[1] and Bjoern Usadel[2,3,*]

[1]Department Metabolic Networks, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm,[2]Institut für Biologie I, RWTH Aachen, Worringer Weg 3, 52074 Aachen and [3]Institute of Bio- and Geosciences: Plant Sciences, Forschungszentrum Jülich, Leo-Brandt-Straße, 52425 Jülich, Germany

Associate Editor: Inanc Birol

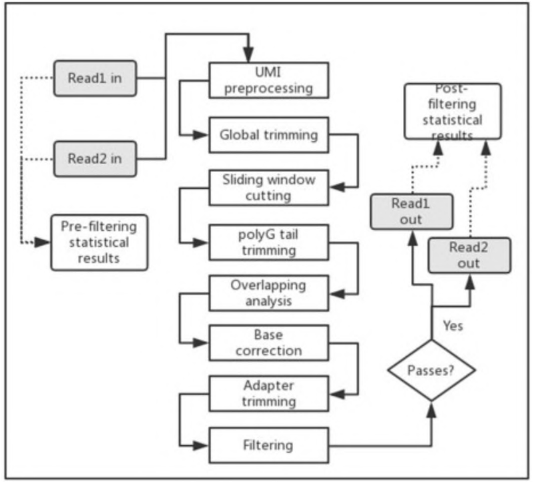


fastp

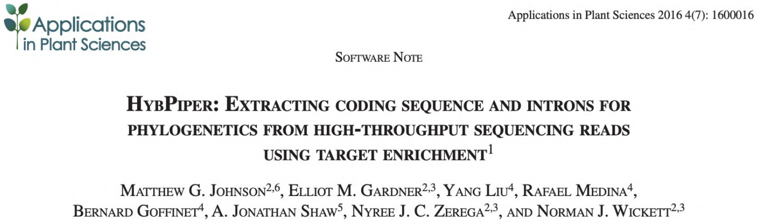
Bioinformatics, 34, 2018, i884–i890
doi: 10.1093/bioinformatics/bty560
ECCB 2018

fastp: an ultra-fast all-in-one FASTQ preprocessor

Shifu Chen[1,2,*], Yanqing Zhou[1], Yaru Chen[1] and Jia Gu[2]



HybPiper

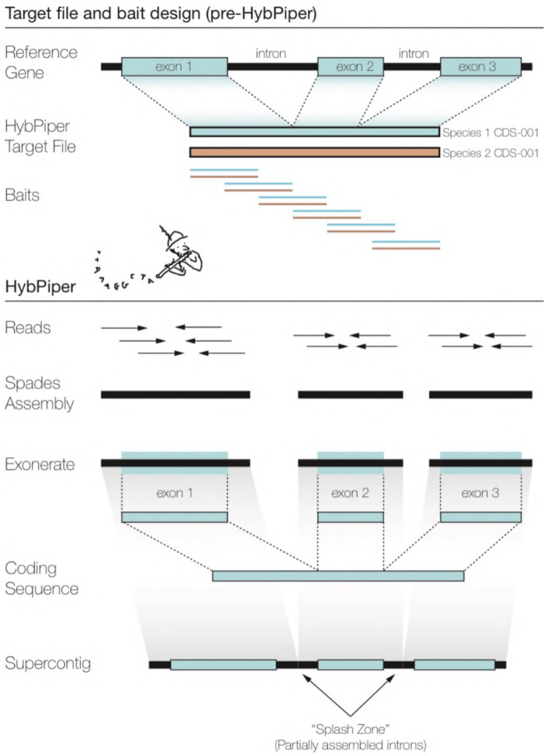


Applications in Plant Sciences 2016 4(7): 1600016

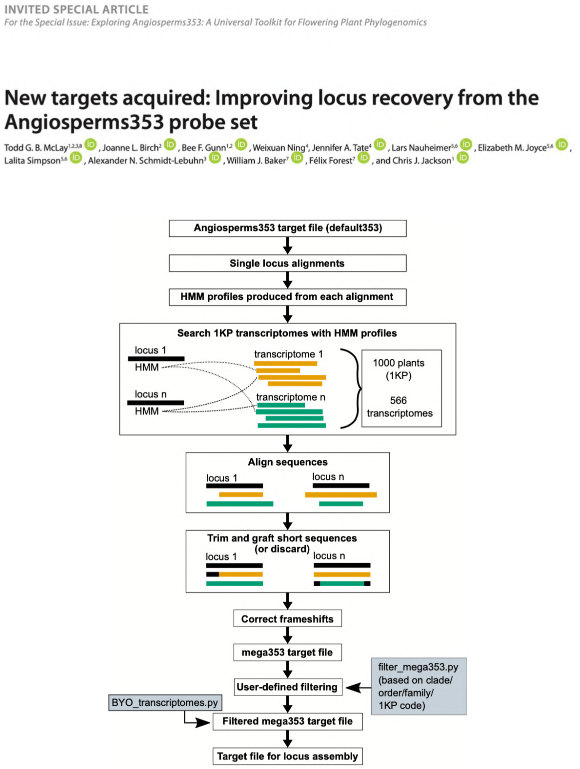
SOFTWARE NOTE

HYBPIPER: EXTRACTING CODING SEQUENCE AND INTRONS FOR PHYLOGENETICS FROM HIGH-THROUGHPUT SEQUENCING READS USING TARGET ENRICHMENT[1]

MATTHEW G. JOHNSON[2,6], ELLIOT M. GARDNER[2,3], YANG LIU[4], RAFAEL MEDINA[4],
BERNARD GOFFINET[4], A. JONATHAN SHAW[5], NYREE J. C. ZEREGA[2,3], AND NORMAN J. WICKETT[2,3]



HybPiper

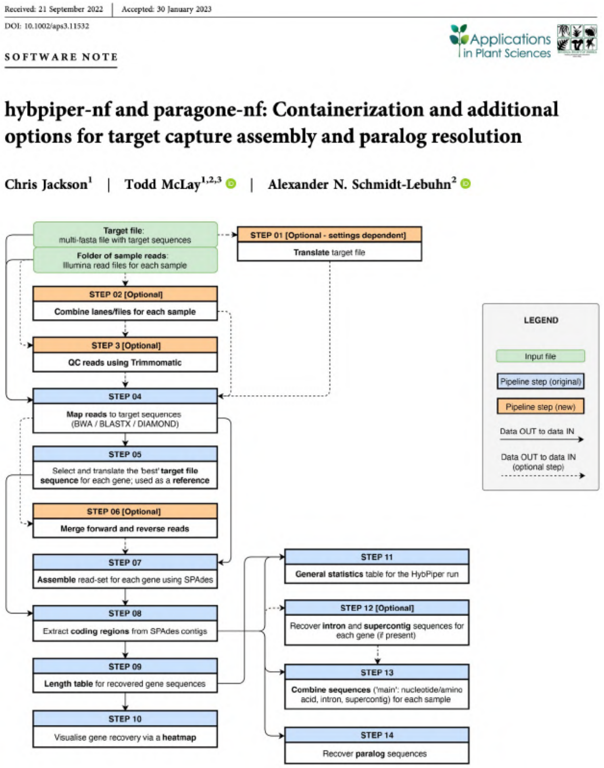


INVITED SPECIAL ARTICLE

For the Special Issue: Exploring Angiosperms353: A Universal Toolkit for Flowering Plant Phylogenomics

New targets acquired: Improving locus recovery from the Angiosperms353 probe set

Todd G. B. McLay[1,2,3,8], Joanne L. Birch[2], Bee F. Gunn[1,2], Weixuan Ning[4], Jennifer A. Tate[4], Lars Nauheimer[5,6], Elizabeth M. Joyce[5,6], Lalita Simpson[5,6], Alexander N. Schmidt-Lebuhn[3], William J. Baker[7], Félix Forest[7], and Chris J. Jackson[1]



Received: 21 September 2022 | Accepted: 30 January 2023

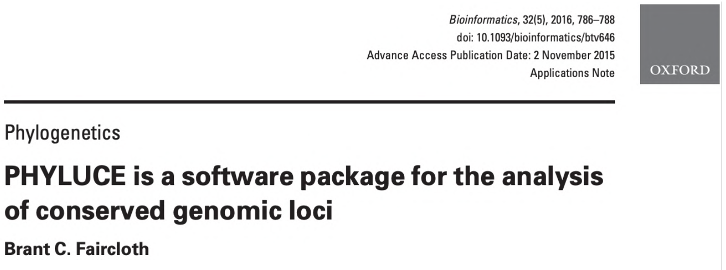
DOI: 10.1002/aps3.11532

SOFTWARE NOTE

hybpiper-nf and paragone-nf: Containerization and additional options for target capture assembly and paralog resolution

Chris Jackson[1] | Todd McLay[1,2,3] | Alexander N. Schmidt-Lebuhn[2]

PHYLUCE



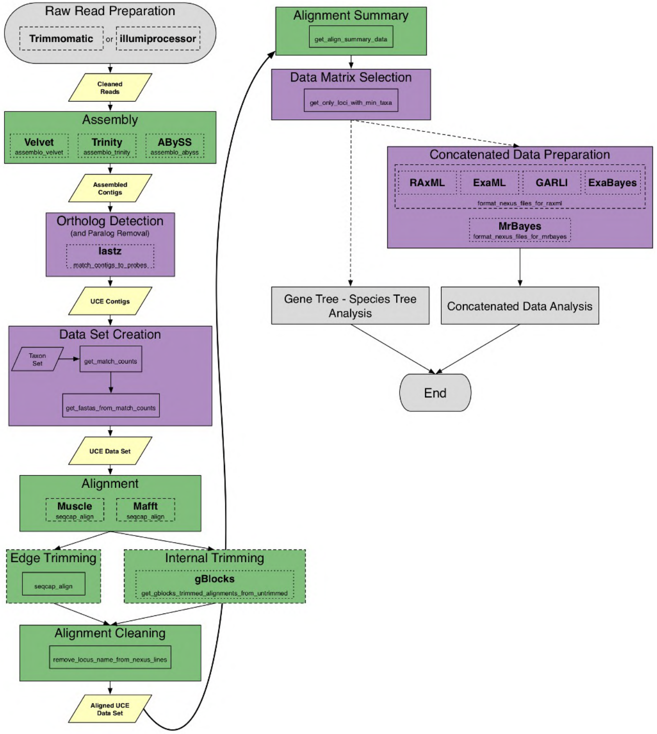
Bioinformatics, 32(5), 2016, 786–788
doi: 10.1093/bioinformatics/btv646
Advance Access Publication Date: 2 November 2015
Applications Note

OXFORD

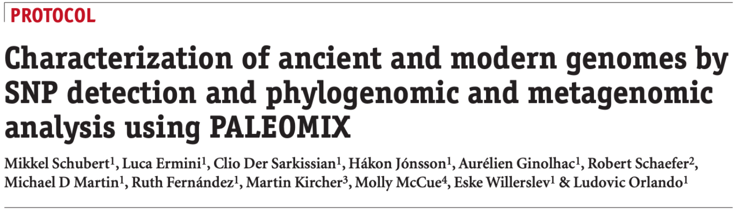
Phylogenetics

PHYLUCE is a software package for the analysis of conserved genomic loci

Brant C. Faircloth



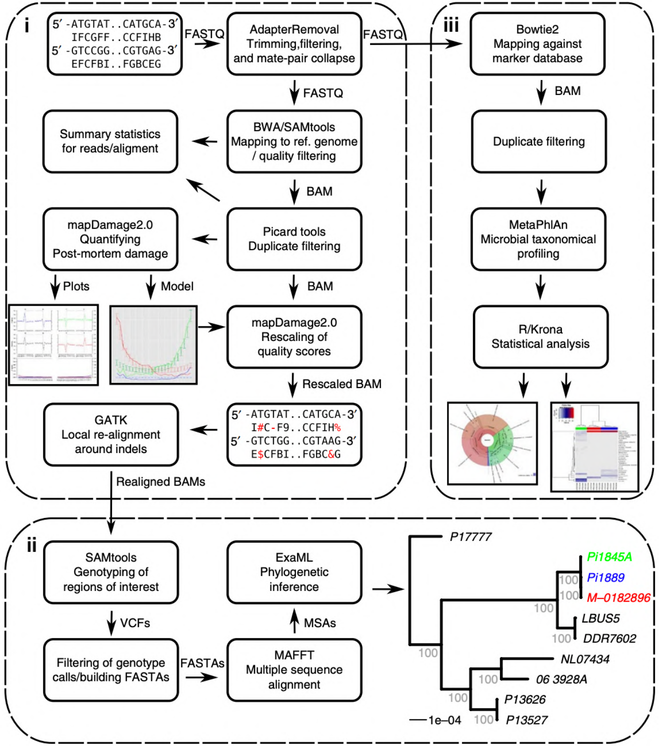
PALEOMIX



PROTOCOL

Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX

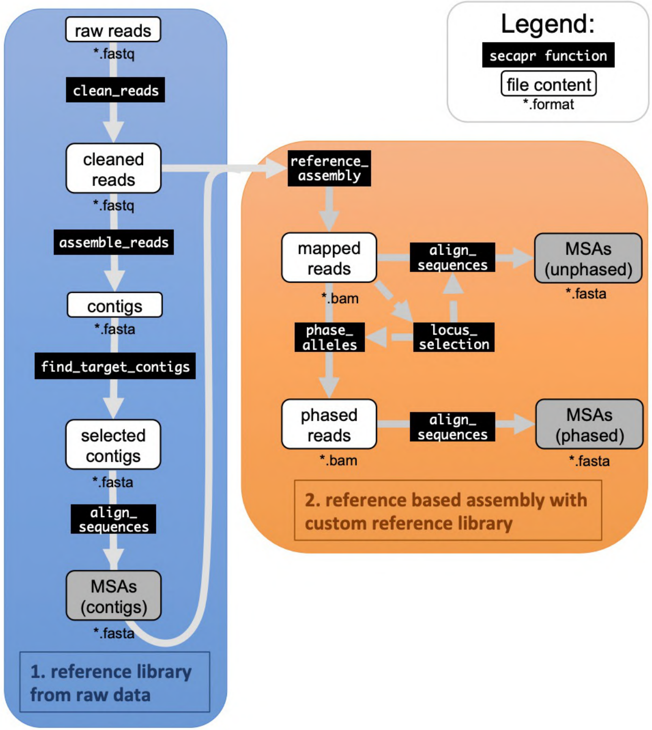
Mikkel Schubert[1], Luca Ermini[1], Clio Der Sarkissian[1], Hákon Jónsson[1], Aurélien Ginolhac[1], Robert Schaefer[2], Michael D Martin[1], Ruth Fernández[1], Martin Kircher[3], Molly McCue[4], Eske Willerslev[1] & Ludovic Orlando[1]



SECAPR

SECAPR—a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments

Tobias Andermann[1,2], Ángela Cano[2,3], Alexander Zizka[1,2],
Christine Bacon[1,2] and Alexandre Antonelli[1,2,4,5]



HybPhyloMaker

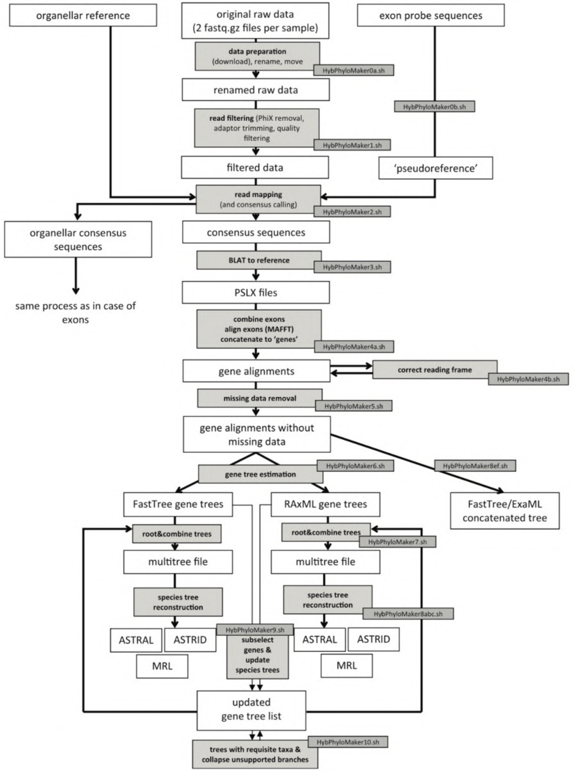
Evolutionary Bioinformatics
Volume 14, January-December 2018
© The Author(s) 2018, Article Reuse Guidelines
https://doi.org/10.1177/1176934317742613

SAGE journals

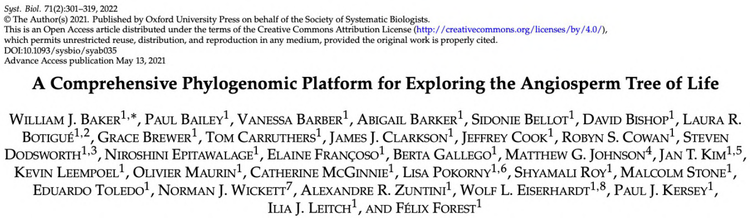
Short Report

HybPhyloMaker: Target Enrichment Data Analysis From Raw Reads to Species Trees

Tomáš Fér[1] and Roswitha E Schmickl[2]



PAFTools



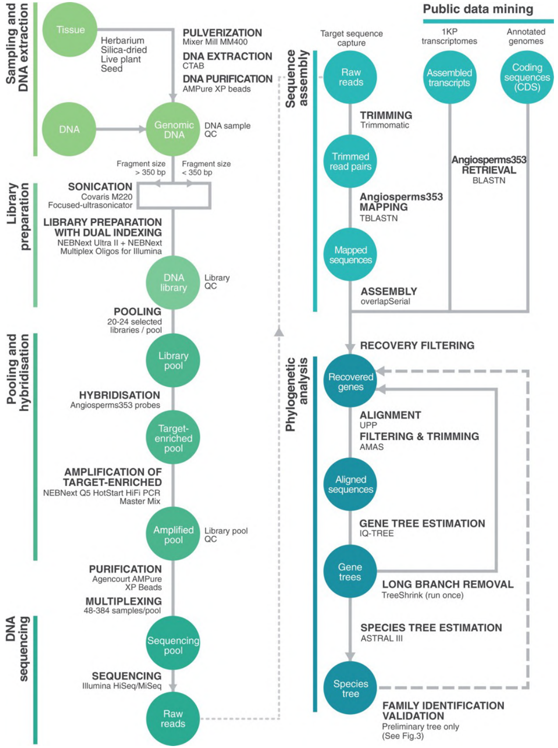
Syst. Biol. 71(2):301–319, 2022

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society of Systematic Biologists.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

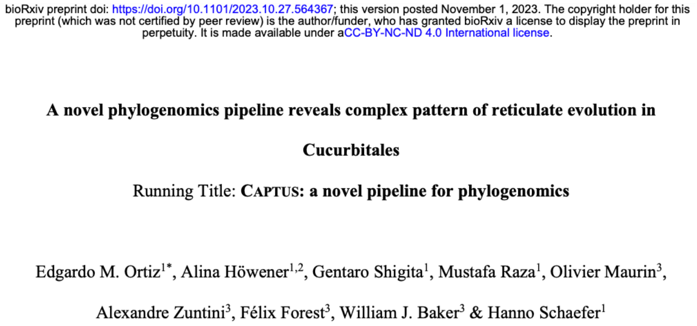
DOI:10.1093/sysbio/syab035
Advance Access publication May 13, 2021

A Comprehensive Phylogenomic Platform for Exploring the Angiosperm Tree of Life

William J. Baker[1,*], Paul Bailey[1], Vanessa Barber[1], Abigail Barker[1], Sidonie Bellot[1], David Bishop[1], Laura R. Botigué[1,2], Grace Brewer[1], Tom Carruthers[1], James J. Clarkson[1], Jeffrey Cook[1], Robyn S. Cowan[1], Steven Dodsworth[1,3], Niroshini Epitawalage[1], Elaine Françoso[1], Berta Gallego[1], Matthew G. Johnson[4], Jan T. Kim[1,5], Kevin Leempoel[1], Olivier Maurin[1], Catherine McGinnie[1], Lisa Pokorny[1,6], Shyamali Roy[1], Malcolm Stone[1], Eduardo Toledo[1], Norman J. Wickett[7], Alexandre R. Zuntini[1], Wolf L. Eiserhardt[1,8], Paul J. Kersey[1], Ilia J. Leitch[1], and Félix Forest[1]



CAPTUS

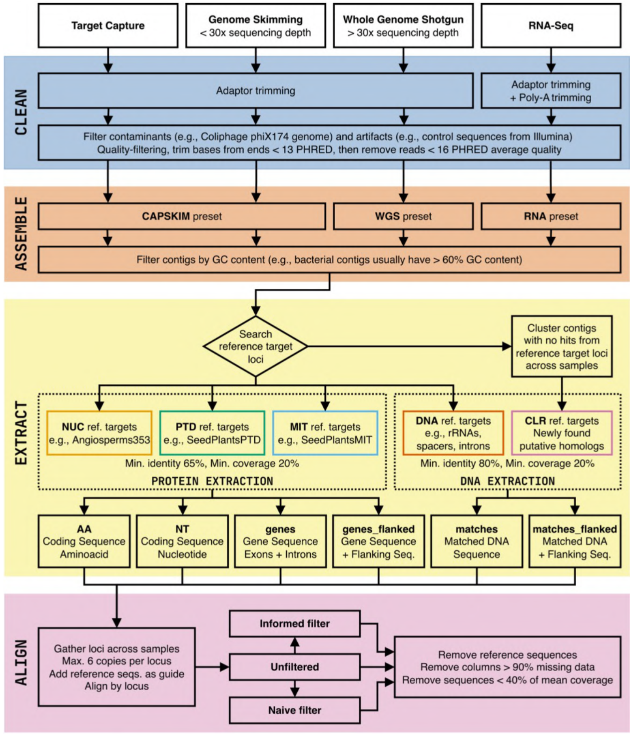


bioRxiv preprint doi: https://doi.org/10.1101/2023.10.27.564367; this version posted November 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A novel phylogenomics pipeline reveals complex pattern of reticulate evolution in Cucurbitales

Running Title: **CAPTUS: a novel pipeline for phylogenomics**

Edgardo M. Ortiz[1*], Alina Höwener[1,2], Gentaro Shigita[1], Mustafa Raza[1], Olivier Maurin[3], Alexandre Zuntini[3], Félix Forest[3], William J. Baker[3] & Hanno Schaefer[1]



Dry Lab Pipeline

Raw Reads

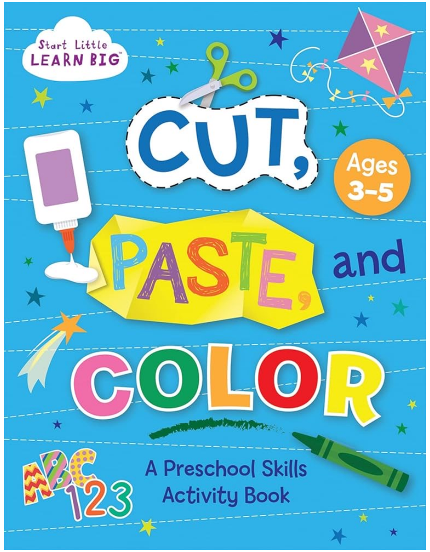
FastQC | MultiQC & Trimmomatic | fastp

BLAST | BWA | DIAMOND

SPAdes | exon(intron)erate

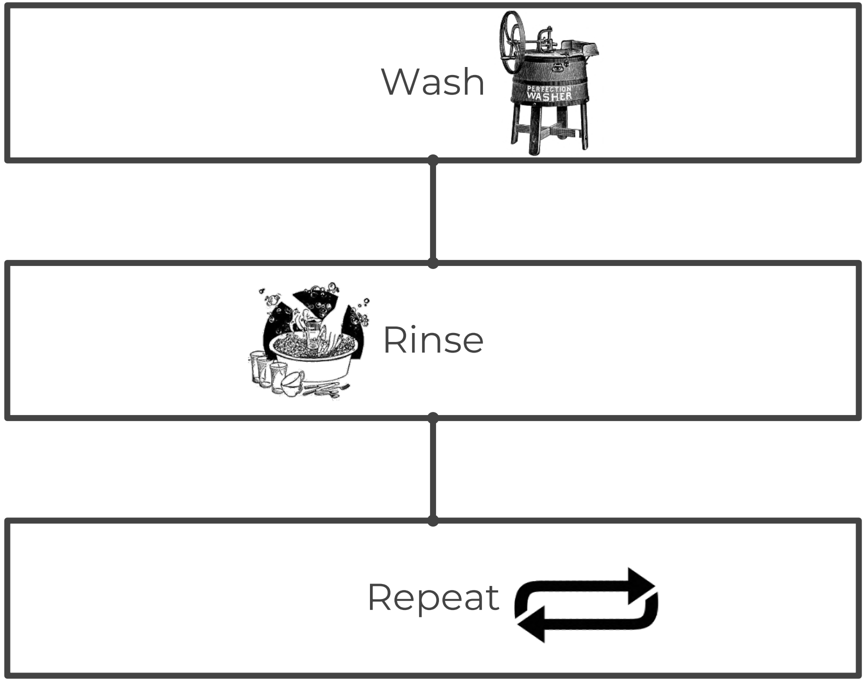
HybPiper

Target sequences



Dry Lab Pipeline

Target Sequences



Wash

Rinse

Repeat

Aligned Matrices

Dry Lab Pipeline

Target Sequences

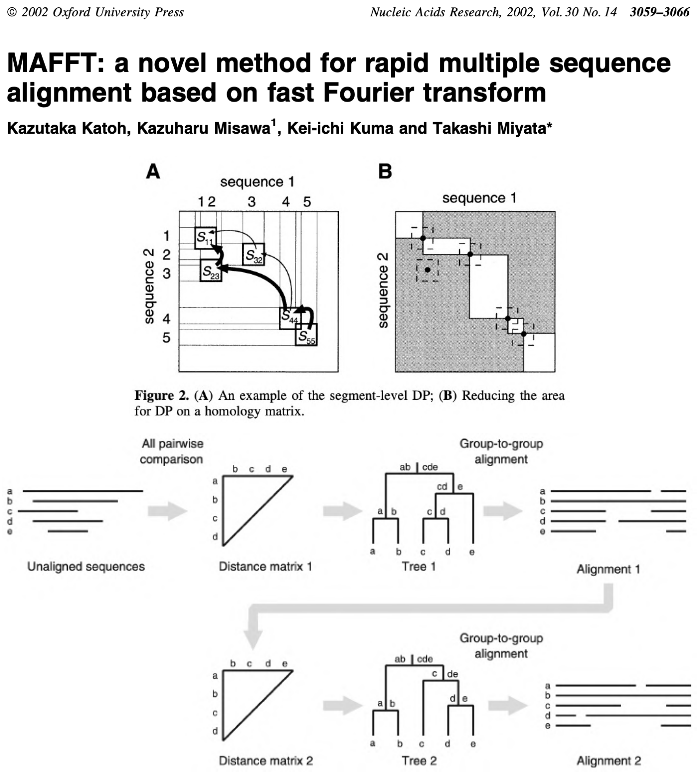
MAFFT | MUSCLE | PASTA | UPP

AMAS | PhyKIT | trimAl | ClipKIT

TreeShrink | PhylteR

Aligned Matrices

MAFFT



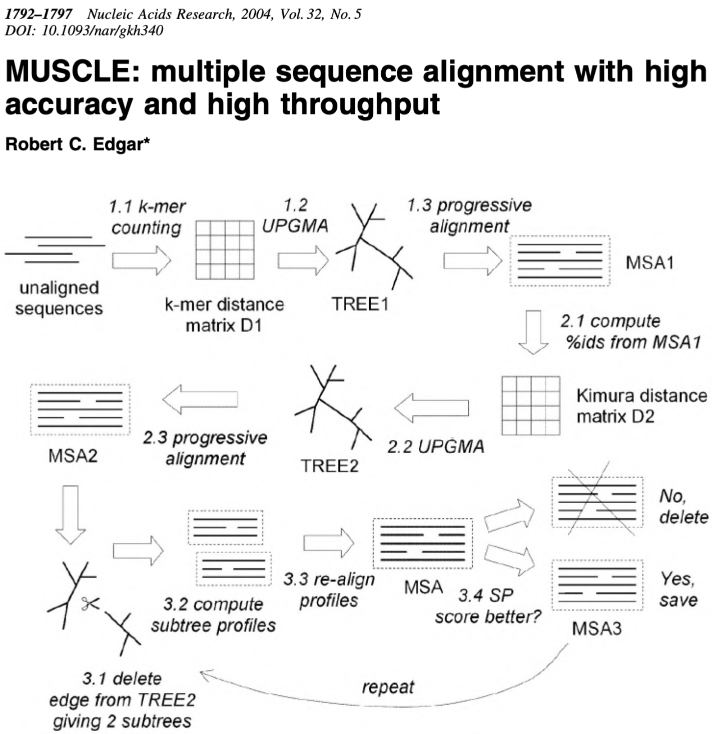
© 2002 Oxford University Press Nucleic Acids Research, 2002, Vol. 30 No. 14 **3059–3066**

MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform

Kazutaka Katoh, Kazuharu Misawa[1], Kei-ichi Kuma and Takashi Miyata*

Figure 2. (**A**) An example of the segment-level DP; (**B**) Reducing the area for DP on a homology matrix.

MUSCLE

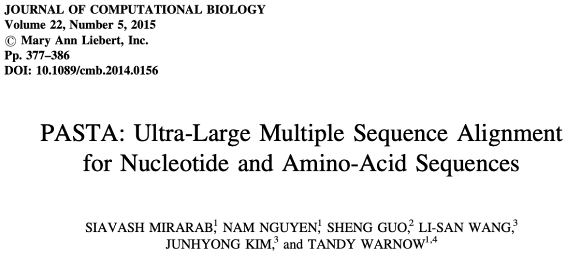


1792–1797 Nucleic Acids Research, 2004, Vol. 32, No. 5
DOI: 10.1093/nar/gkh340

MUSCLE: multiple sequence alignment with high accuracy and high throughput

Robert C. Edgar*

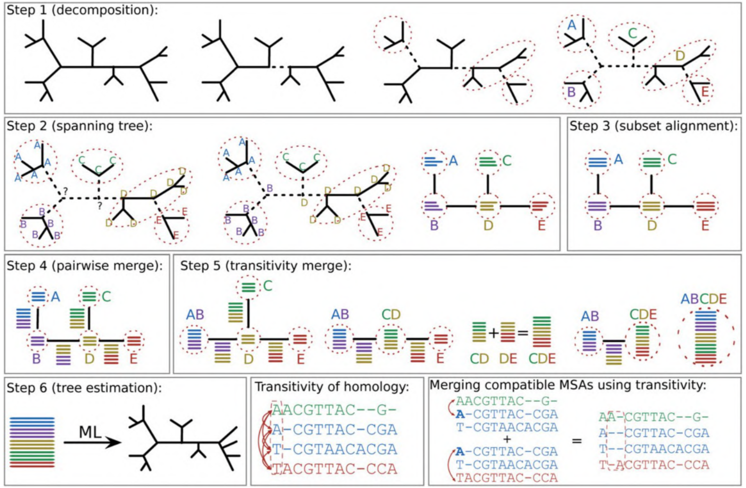
PASTA



JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 22, Number 5, 2015
© Mary Ann Liebert, Inc.
Pp. 377–386
DOI: 10.1089/cmb.2014.0156

PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences

SIAVASH MIRARAB[1] NAM NGUYEN[1] SHENG GUO[2] LI-SAN WANG[3]
JUNHYONG KIM[3] and TANDY WARNOW[1,4]



UPP

Nguyen *et al. Genome Biology* (2015) 16:124
DOI 10.1186/s13059-015-0688-z

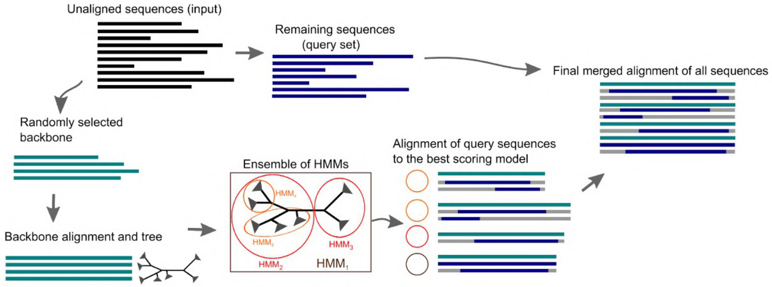
Genome **Biology**

METHOD **Open Access**

CrossMark

Ultra-large alignments using phylogeny-aware profiles

Nam-phuong D. Nguyen[1], Siavash Mirarab[2], Keerthana Kumar[2] and Tandy Warnow[1,3,4*]



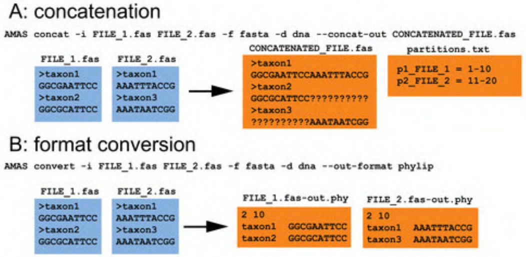
AMAS

PhyKIT

PeerJ

AMAS: a fast tool for alignment manipulation and computing of summary statistics

Marek L. Borowiec



A: concatenation

```
AMAS concat -i FILE_1.fas FILE_2.fas -f fasta -d dna --concat-out CONCATENATED_FILE.fas
```

B: format conversion

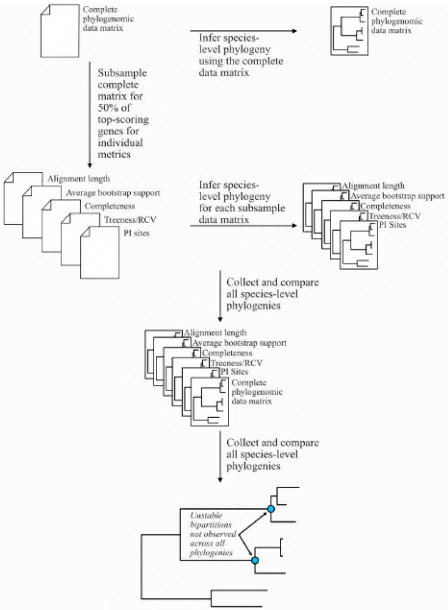
```
AMAS convert -i FILE_1.fas FILE_2.fas -f fasta -d dna --out-format phylip
```

Function	AMAS	FASconCAT-G	Phyutility
Input formats	fasta phylip nexus	clustal fasta phylip nexus	fasta nexus
Concatenation	yes	yes	yes
Splitting or site extraction	yes	yes	yes (gaps only)
Summary statistics	yes	yes	no
Replicate alignments	yes	no	no
Taxon removal	yes	no	no
Translation	no	yes	no
RY coding	no	yes	no
Consensus sequences	no	yes	no
NCBI interactions	no	no	yes
Tree manipulations	no	no	yes

Phylogenetics

PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data

Jacob L. Steenwyk[1,*], Thomas J. Buida III[2], Abigail L. Labella[1], Yuanning Li[1], Xing-Xing Shen[3], & Antonis Rokas[1,*]



phyutility

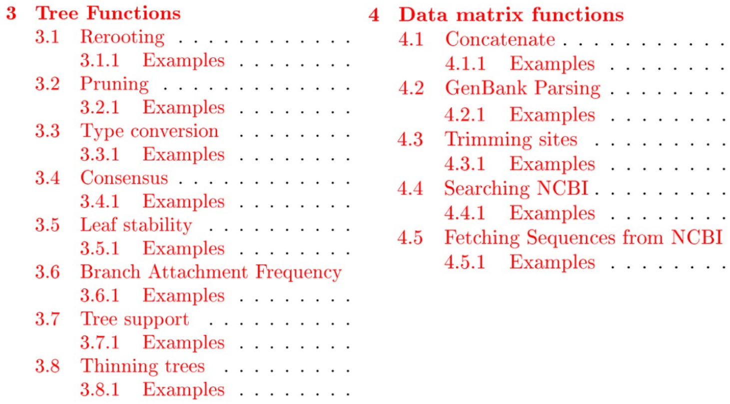
FASconCAT-G

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 24 no. 5 2008, pages 715–716*
doi:10.1093/bioinformatics/btm619

Phylogenetics

Phyutility: a phyloinformatics tool for trees, alignments and molecular data

Stephen A. Smith[1,*] and Casey W. Dunn[2]



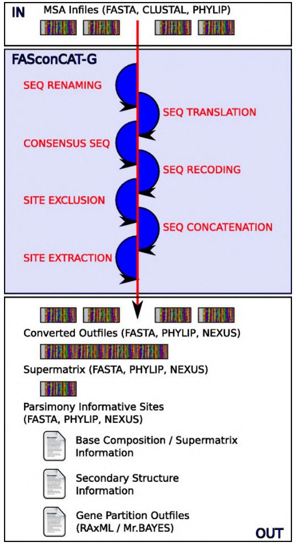
Kück and Longo *Frontiers in Zoology* 2014, **11**:81
http://www.frontiersinzoology.com/content/11/1/81

FRONTIERS IN ZOOLOGY

METHODOLOGY **Open Access**

FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies

Patrick Kück[1*] and Gary C Longo[2]



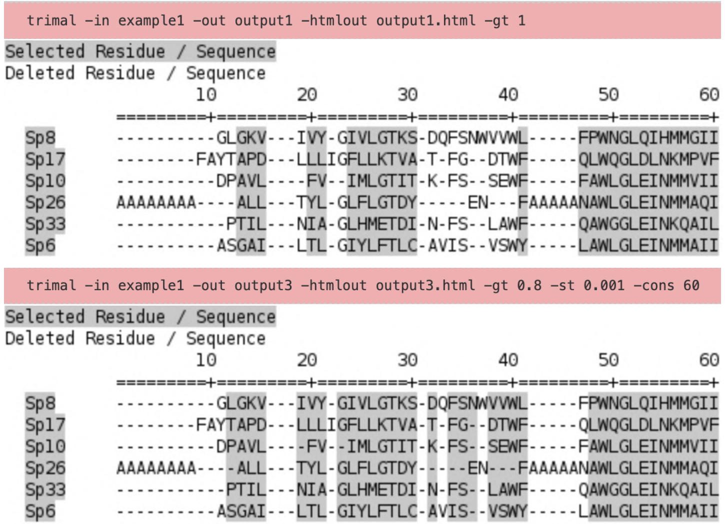
trimAl

BIOINFORMATICS **APPLICATIONS NOTE**
Vol. 25 no. 15 2009, pages 1972–1973
doi:10.1093/bioinformatics/btp348

Phylogenetics

trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses

Salvador Capella-Gutiérrez, José M. Silla-Martínez and Toni Gabaldón*



```
trimal -in example1 -out output1 -htmlout output1.html -gt 1
```

```
Selected Residue / Sequence
Deleted Residue / Sequence
                   10        20        30        40        50        60
           =========+=========+=========+=========+=========+=========+
  Sp8      ----------GLGKV---IVY-GIVLGTKS-DQFSNWVVWL-----FPWNGLQIHMMGII
  Sp17     --------FAYTAPD---LLLIGFLLKTVA-T-FG--DTWF-----QLWQGLDLNKMPVF
  Sp10     ----------DPAVL----FV--IMLGTIT-K-FS--SEWF-----FAWLGLEINMMVII
  Sp26     AAAAAAAA----ALL---TYL-GLFLGTDY-----EN---FAAAAANAWLGLEINMMAQI
  Sp33     -----------PTIL---NIA-GLHMETDI-N-FS--LAWF-----QAWGGLEINKQAIL
  Sp6      ----------ASGAI---LTL-GIYLFTLC-AVIS--VSWY-----LAWLGLEINMMAII
```

```
trimal -in example1 -out output3 -htmlout output3.html -gt 0.8 -st 0.001 -cons 60
```

```
Selected Residue / Sequence
Deleted Residue / Sequence
                   10        20        30        40        50        60
           =========+=========+=========+=========+=========+=========+
  Sp8      ----------GLGKV---IVY-GIVLGTKS-DQFSNWVVWL-----FPWNGLQIHMMGII
  Sp17     --------FAYTAPD---LLLIGFLLKTVA-T-FG--DTWF-----QLWQGLDLNKMPVF
  Sp10     ----------DPAVL----FV--IMLGTIT-K-FS--SEWF-----FAWLGLEINMMVII
  Sp26     AAAAAAAA----ALL---TYL-GLFLGTDY-----EN---FAAAAANAWLGLEINMMAQI
  Sp33     -----------PTIL---NIA-GLHMETDI-N-FS--LAWF-----QAWGGLEINKQAIL
  Sp6      ----------ASGAI---LTL-GIYLFTLC-AVIS--VSWY-----LAWLGLEINMMAII
```

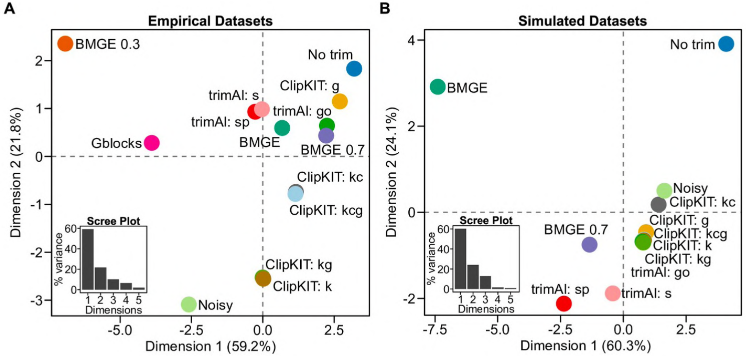
ClipKIT

PLOS BIOLOGY

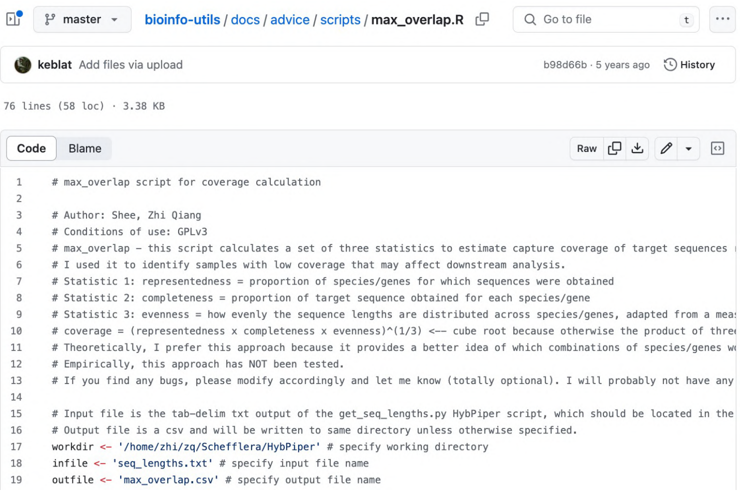
METHODS AND RESOURCES

ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference

Jacob L. Steenwyk[1]*, **Thomas J. Buida, III**[2], **Yuanning Li**[1], **Xing-Xing Shen**[3], **Antonis Rokas**[1]*



max_overlap

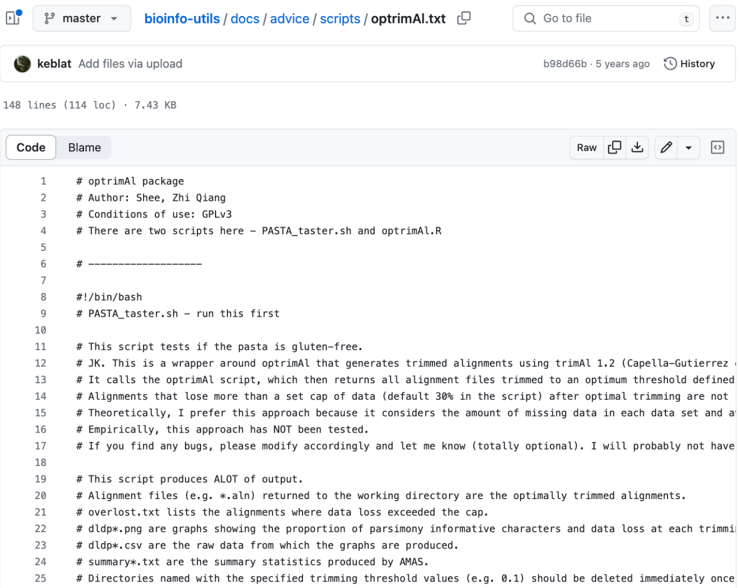


```r
# max_overlap script for coverage calculation

# Author: Shee, Zhi Qiang
# Conditions of use: GPLv3
# max_overlap – this script calculates a set of three statistics to estimate capture coverage of target sequences
# I used it to identify samples with low coverage that may affect downstream analysis.
# Statistic 1: representedness = proportion of species/genes for which sequences were obtained
# Statistic 2: completeness = proportion of target sequence obtained for each species/gene
# Statistic 3: evenness = how evenly the sequence lengths are distributed across species/genes, adapted from a mea
# coverage = (representedness x completeness x evenness)^(1/3) <–– cube root because otherwise the product of thre
# Theoretically, I prefer this approach because it provides a better idea of which combinations of species/genes w
# Empirically, this approach has NOT been tested.
# If you find any bugs, please modify accordingly and let me know (totally optional). I will probably not have any

# Input file is the tab–delim txt output of the get_seq_lengths.py HybPiper script, which should be located in the
# Output file is a csv and will be written to same directory unless otherwise specified.
workdir <- '/home/zhi/zq/Schefflera/HybPiper' # specify working directory
infile <- 'seq_lengths.txt' # specify input file name
outfile <- 'max_overlap.csv' # specify output file name
```

optrimAl



```
# optrimAl package
# Author: Shee, Zhi Qiang
# Conditions of use: GPLv3
# There are two scripts here – PASTA_taster.sh and optrimAl.R

# -------------------

#!/bin/bash
# PASTA_taster.sh – run this first

# This script tests if the pasta is gluten–free.
# JK. This is a wrapper around optrimAl that generates trimmed alignments using trimAl 1.2 (Capella–Gutierrez
# It calls the optrimAl script, which then returns all alignment files trimmed to an optimum threshold defined
# Alignments that lose more than a set cap of data (default 30% in the script) after optimal trimming are not
# Theoretically, I prefer this approach because it considers the amount of missing data in each data set and
# Empirically, this approach has NOT been tested.
# If you find any bugs, please modify accordingly and let me know (totally optional). I will probably not have

# This script produces ALOT of output.
# Alignment files (e.g. *.aln) returned to the working directory are the optimally trimmed alignments.
# overlost.txt lists the alignments where data loss exceeded the cap.
# dldp*.png are graphs showing the proportion of parsimony informative characters and data loss at each trimmi
# dldp*.csv are the raw data from which the graphs are produced.
# summary*.txt are the summary statistics produced by AMAS.
# Directories named with the specified trimming threshold values (e.g. 0.1) should be deleted immediately once
```

TreeShrink

Mai and Mirarab *BMC Genomics* 2018, **19**(Suppl 5):272
https://doi.org/10.1186/s12864-018-4620-2

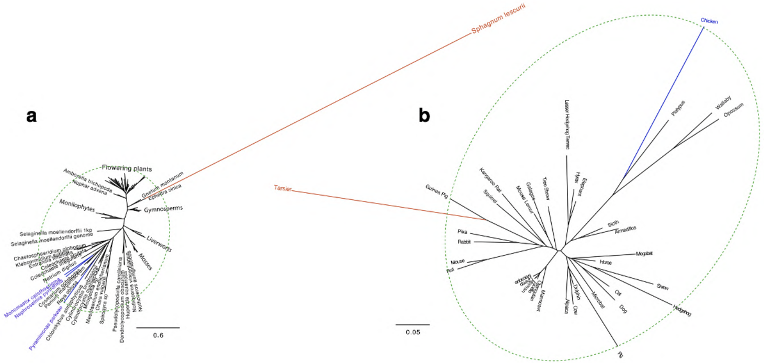
BMC Genomics

RESEARCH Open Access

TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees

Uyen Mai[1] and Siavash Mirarab[2*]

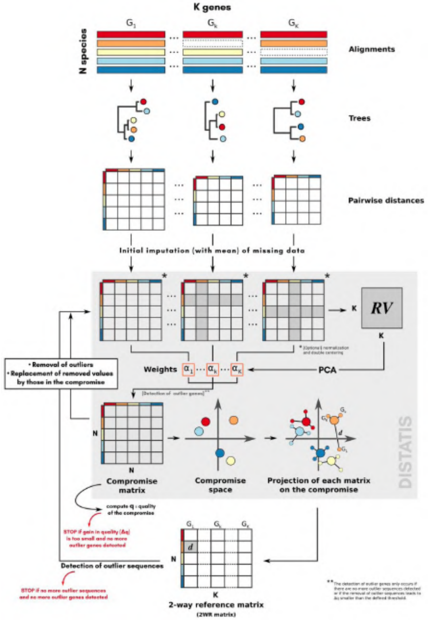
From RECOMB-CG - 2017 : The Fifteenth RECOMB Comparative Genomics Satellite Conference
Barcelona, Spain. 04-06 October 2017



PhylteR

PhylteR: Efficient Identification of Outlier Sequences in Phylogenomic Datasets

Aurore Comte,[1,2,†] Théo Tricou,[3,†] Eric Tannier,[3,4] Julien Joseph,[3] Aurélie Siberchicot,[3] Simon Penel,[3] Rémi Allio,[5] Frédéric Delsuc,[6] Stéphane Dray,[3] and Damien M. de Vienne[3,*]



Dry Lab Pipeline

Target Sequences

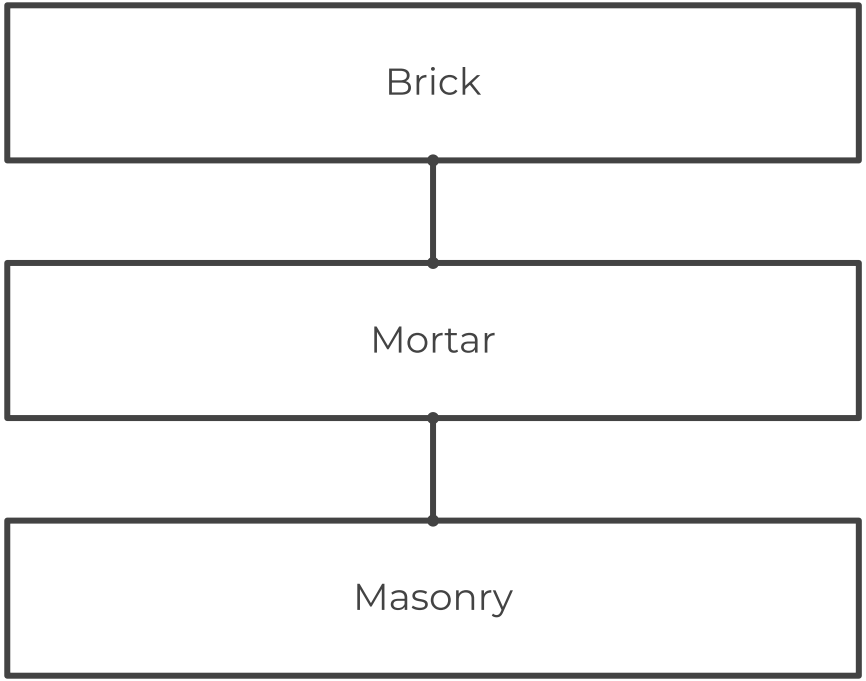
MAFFT | MUSCLE | PASTA | UPP

AMAS | PhyKIT | trimAl | ClipKIT

TreeShrink | PhylteR

Aligned Matrices

Dry Lab Pipeline



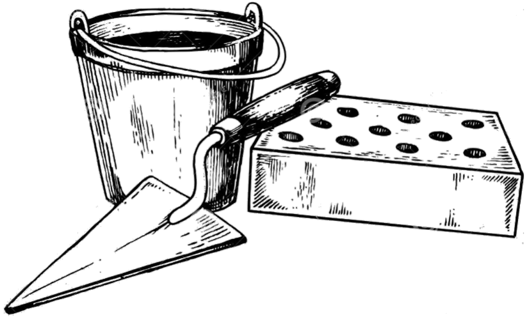
Aligned Matrices

Brick

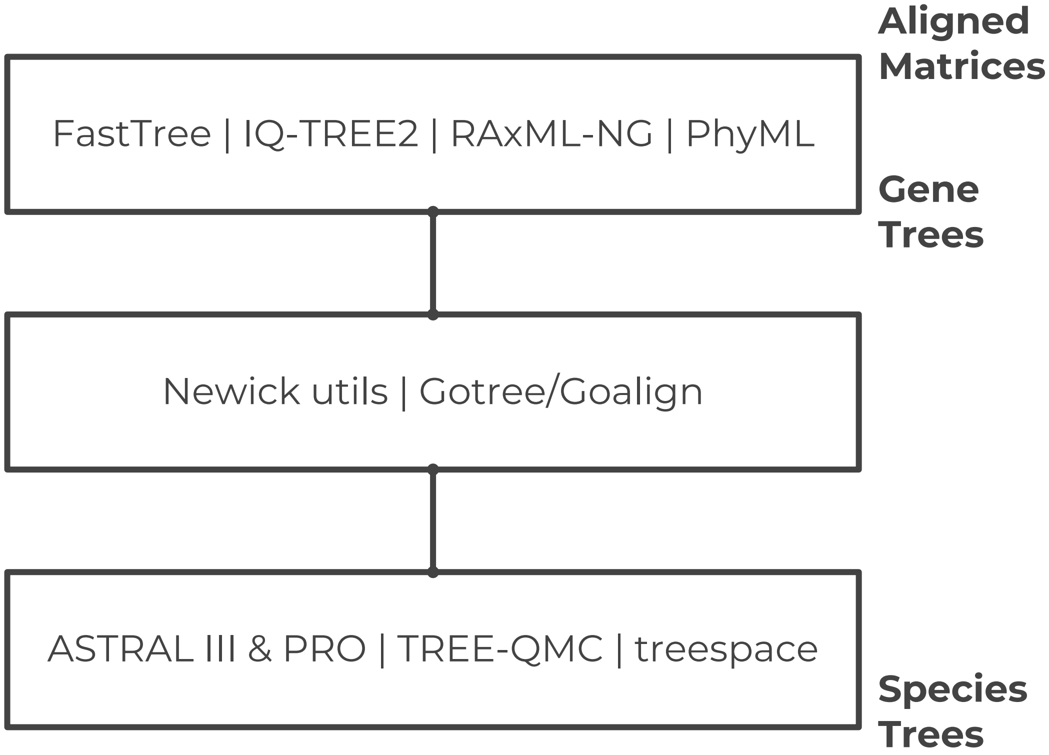
Mortar

Masonry

Species Trees



Dry Lab Pipeline



Aligned Matrices

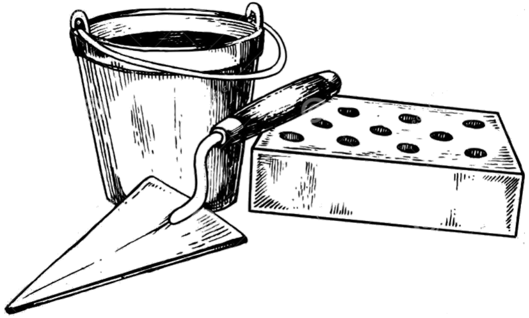
FastTree | IQ-TREE2 | RAxML-NG | PhyML

Gene Trees

Newick utils | Gotree/Goalign

ASTRAL III & PRO | TREE-QMC | treespace

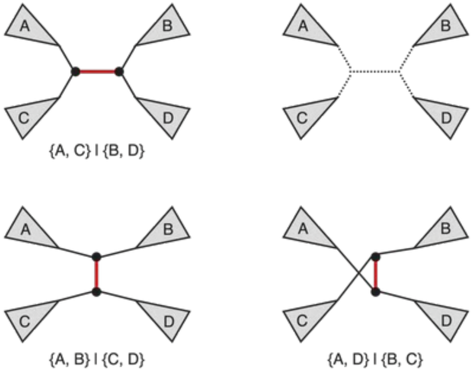
Species Trees



IQ-TREE2

IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era

Bui Quang Minh ,*[1,2] Heiko A. Schmidt,[3] Olga Chernomor,[3] Dominik Schrempf,[3,4] Michael D. Woodhams,[5] Arndt von Haeseler,[†,3,6] and Robert Lanfear[†,2]



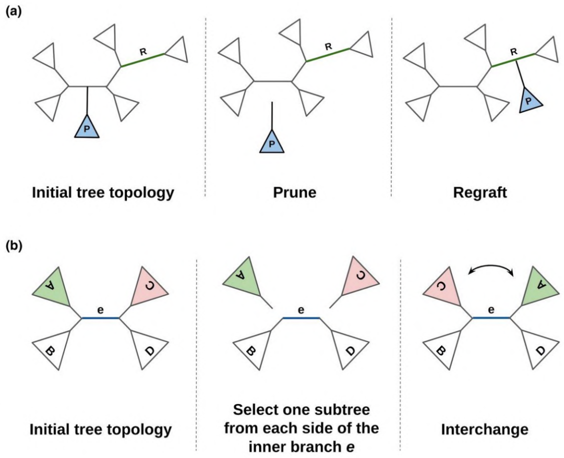
RAxML-NG

Bioinformatics, 2019, 1–3
doi: 10.1093/bioinformatics/btz305
Advance Access Publication Date: 9 May 2019
Application Note

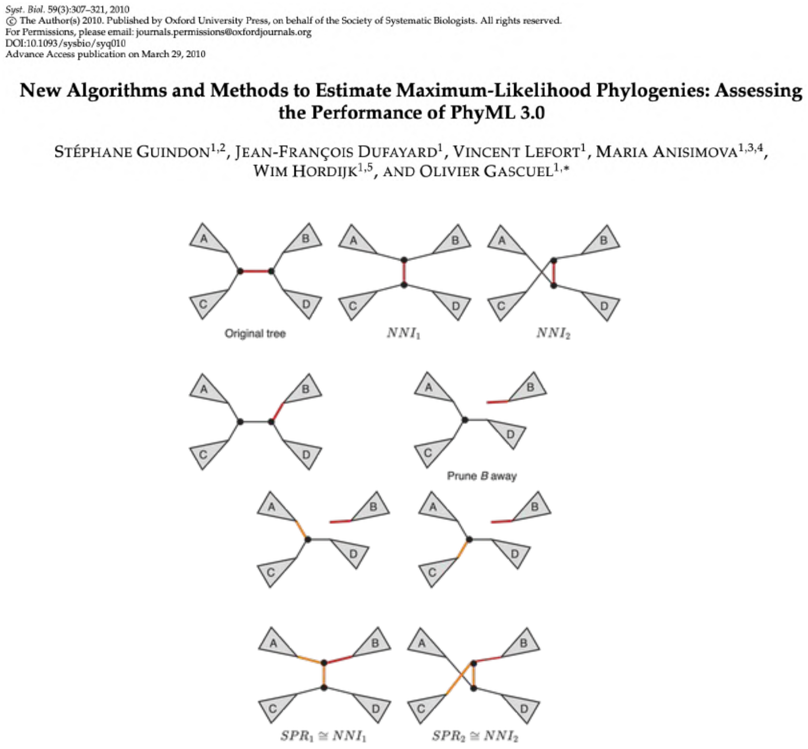
Phylogenetics

RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference

Alexey M. Kozlov [1,]*, Diego Darriba [1], Tomáš Flouri[1], Benoit Morel[1] and Alexandros Stamatakis[1,2]



PhyML

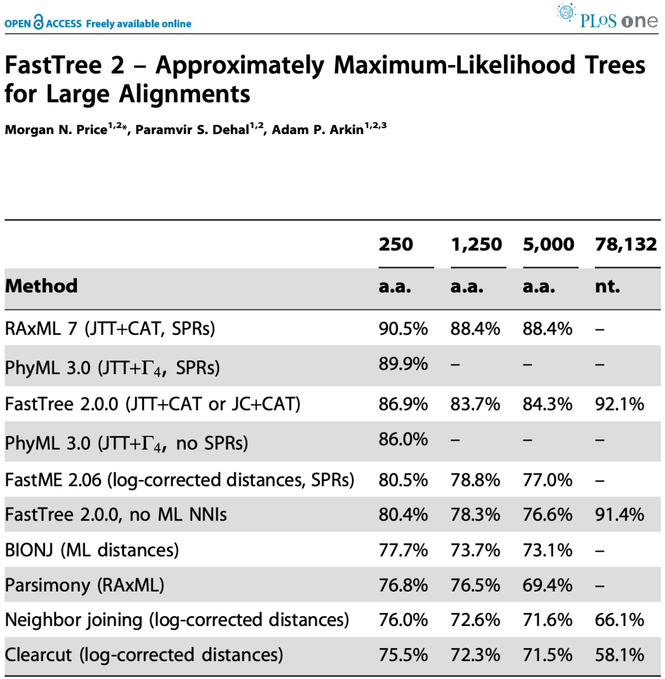


Syst. Biol. 59(3):307–321, 2010
© The Author(s) 2010. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oxfordjournals.org
DOI:10.1093/sysbio/syq010
Advance Access publication on March 29, 2010

New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0

STÉPHANE GUINDON[1,2], JEAN-FRANÇOIS DUFAYARD[1], VINCENT LEFORT[1], MARIA ANISIMOVA[1,3,4],
WIM HORDIJK[1,5], AND OLIVIER GASCUEL[1,*]

FastTree2



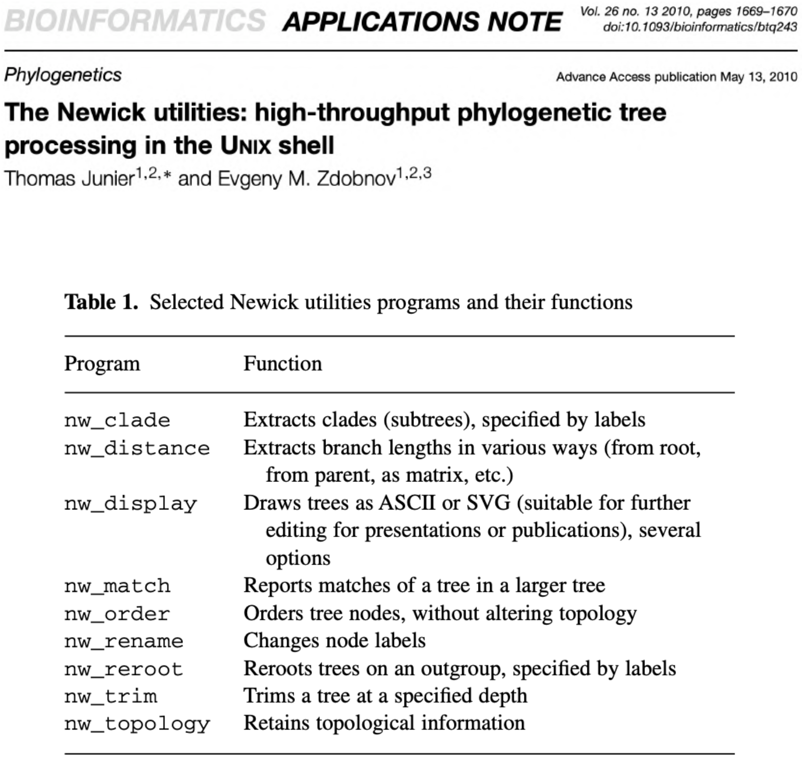
OPEN ACCESS Freely available online — PLoS one

FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments

Morgan N. Price[1,2*], Paramvir S. Dehal[1,2], Adam P. Arkin[1,2,3]

	250	1,250	5,000	78,132
Method	**a.a.**	**a.a.**	**a.a.**	**nt.**
RAxML 7 (JTT+CAT, SPRs)	90.5%	88.4%	88.4%	–
PhyML 3.0 (JTT+Γ_4, SPRs)	89.9%	–	–	–
FastTree 2.0.0 (JTT+CAT or JC+CAT)	86.9%	83.7%	84.3%	92.1%
PhyML 3.0 (JTT+Γ_4, no SPRs)	86.0%	–	–	–
FastME 2.06 (log-corrected distances, SPRs)	80.5%	78.8%	77.0%	–
FastTree 2.0.0, no ML NNIs	80.4%	78.3%	76.6%	91.4%
BIONJ (ML distances)	77.7%	73.7%	73.1%	–
Parsimony (RAxML)	76.8%	76.5%	69.4%	–
Neighbor joining (log-corrected distances)	76.0%	72.6%	71.6%	66.1%
Clearcut (log-corrected distances)	75.5%	72.3%	71.5%	58.1%

Newick Utils



BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 13 2010, pages 1669–1670
doi:10.1093/bioinformatics/btq243

Phylogenetics Advance Access publication May 13, 2010

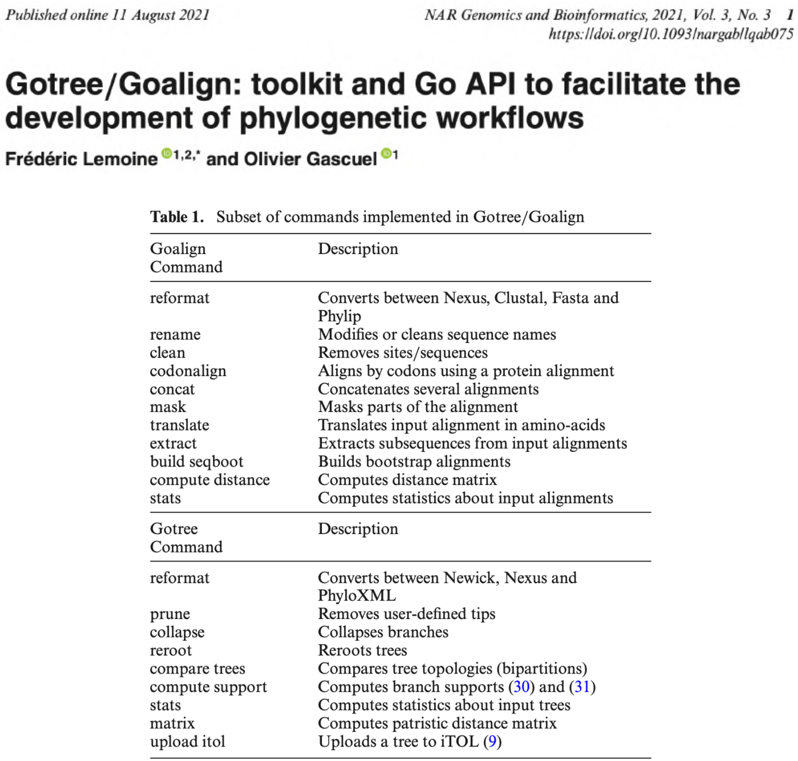
The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell

Thomas Junier[1,2,*] and Evgeny M. Zdobnov[1,2,3]

Table 1. Selected Newick utilities programs and their functions

Program	Function
`nw_clade`	Extracts clades (subtrees), specified by labels
`nw_distance`	Extracts branch lengths in various ways (from root, from parent, as matrix, etc.)
`nw_display`	Draws trees as ASCII or SVG (suitable for further editing for presentations or publications), several options
`nw_match`	Reports matches of a tree in a larger tree
`nw_order`	Orders tree nodes, without altering topology
`nw_rename`	Changes node labels
`nw_reroot`	Reroots trees on an outgroup, specified by labels
`nw_trim`	Trims a tree at a specified depth
`nw_topology`	Retains topological information

Gotree/Goalign



Published online 11 August 2021 *NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 3* **1**
https://doi.org/10.1093/nargab/lqab075

Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows

Frédéric Lemoine[1,2,*] and Olivier Gascuel[1]

Table 1. Subset of commands implemented in Gotree/Goalign

Goalign Command	Description
reformat	Converts between Nexus, Clustal, Fasta and Phylip
rename	Modifies or cleans sequence names
clean	Removes sites/sequences
codonalign	Aligns by codons using a protein alignment
concat	Concatenates several alignments
mask	Masks parts of the alignment
translate	Translates input alignment in amino-acids
extract	Extracts subsequences from input alignments
build seqboot	Builds bootstrap alignments
compute distance	Computes distance matrix
stats	Computes statistics about input alignments
Gotree Command	**Description**
reformat	Converts between Newick, Nexus and PhyloXML
prune	Removes user-defined tips
collapse	Collapses branches
reroot	Reroots trees
compare trees	Compares tree topologies (bipartitions)
compute support	Computes branch supports (30) and (31)
stats	Computes statistics about input trees
matrix	Computes patristic distance matrix
upload itol	Uploads a tree to iTOL (9)

ASTRAL-III

ASTRAL-Pro

Zhang *et al. BMC Bioinformatics* 2018, **19**(Suppl 6):153
https://doi.org/10.1186/s12859-018-2129-y

BMC Bioinformatics

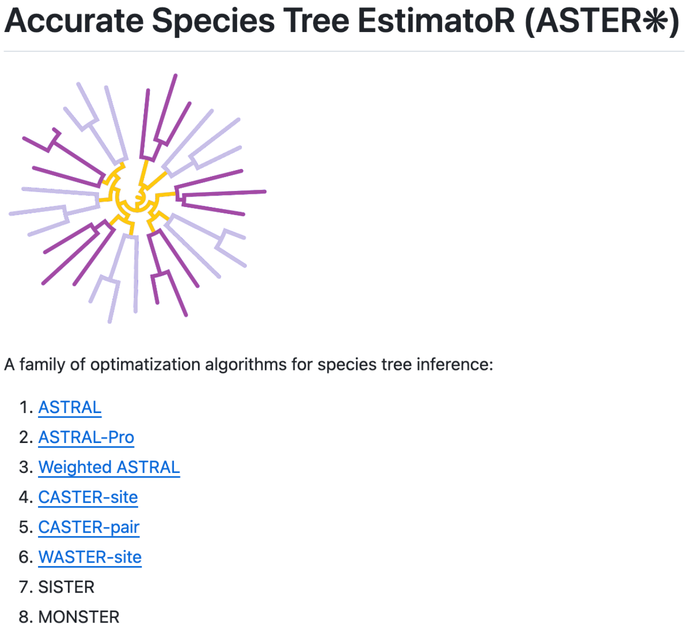
RESEARCH Open Access

ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees

Chao Zhang[3], Maryam Rabiee[2], Erfan Sayyari[1] and Siavash Mirarab[1*]

ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy

Chao Zhang,[1] Celine Scornavacca,[2] Erin K Molloy,[3] and Siavash Mirarab *,[4]



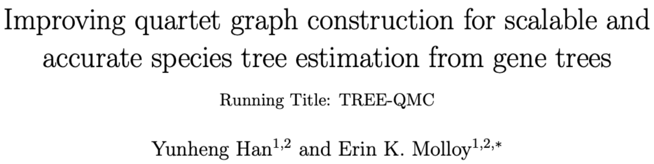
Accurate Species Tree EstimatoR (ASTER✻)

A family of optimatization algorithms for species tree inference:

1. ASTRAL
2. ASTRAL-Pro
3. Weighted ASTRAL
4. CASTER-site
5. CASTER-pair
6. WASTER-site
7. SISTER
8. MONSTER

TREE-QMC

treespace



Improving quartet graph construction for scalable and accurate species tree estimation from gene trees

Running Title: TREE-QMC

Yunheng Han[1,2] and Erin K. Molloy[1,2,*]

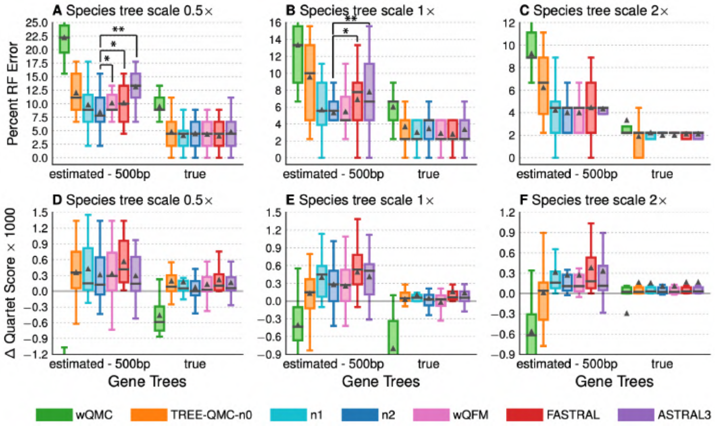
Received: 22 November 2016 | Revised: 17 March 2017 | Accepted: 21 March 2017
DOI: 10.1111/1755-0998.12676

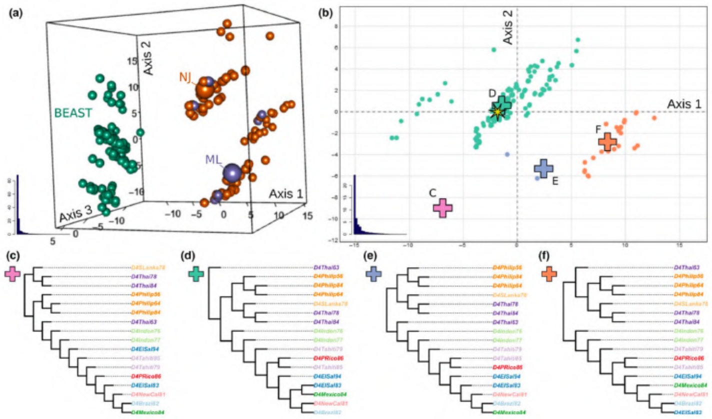
RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

TREESPACE: Statistical exploration of landscapes of phylogenetic trees

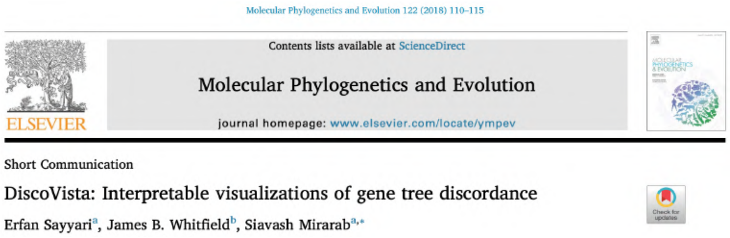
Thibaut Jombart[1,*] | Michelle Kendall[2,*] | Jacob Almagro-Garcia[3] | Caroline Colijn[2]





DiscoVista

AstralPlane



Molecular Phylogenetics and Evolution 122 (2018) 110–115

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

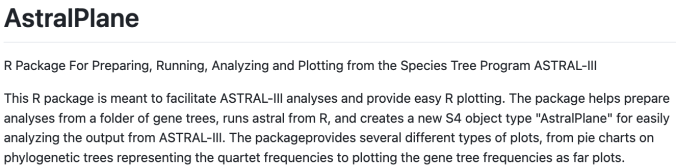
journal homepage: www.elsevier.com/locate/ympev

ELSEVIER

Short Communication

DiscoVista: Interpretable visualizations of gene tree discordance

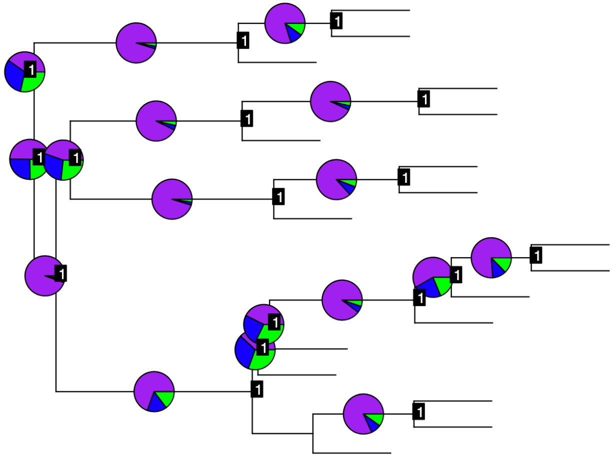
Erfan Sayyari[a], James B. Whitfield[b], Siavash Mirarab[a,*]



AstralPlane

R Package For Preparing, Running, Analyzing and Plotting from the Species Tree Program ASTRAL-III

This R package is meant to facilitate ASTRAL-III analyses and provide easy R plotting. The package helps prepare analyses from a folder of gene trees, runs astral from R, and creates a new S4 object type "AstralPlane" for easily analyzing the output from ASTRAL-III. The packageprovides several different types of plots, from pie charts on phylogenetic trees representing the quartet frequencies to plotting the gene tree frequencies as far plots.



TreePL

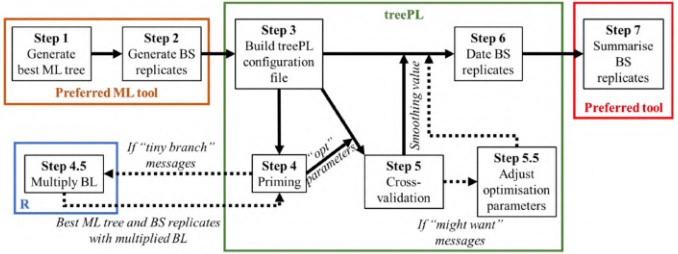
BIOINFORMATICS **APPLICATIONS NOTE** Vol. 28 no. 20 2012, pages 2689–2690
doi:10.1093/bioinformatics/bts492

Phylogenetics

Advance Access publication August 20, 2012

treePL: divergence time estimation using penalized likelihood for large phylogenies

Stephen A. Smith[1,*] and Brian C. O'Meara[2]



Copyright *Kévin J. L. Maurin.* Creative Commons Attribution License (CC BY 4.0).

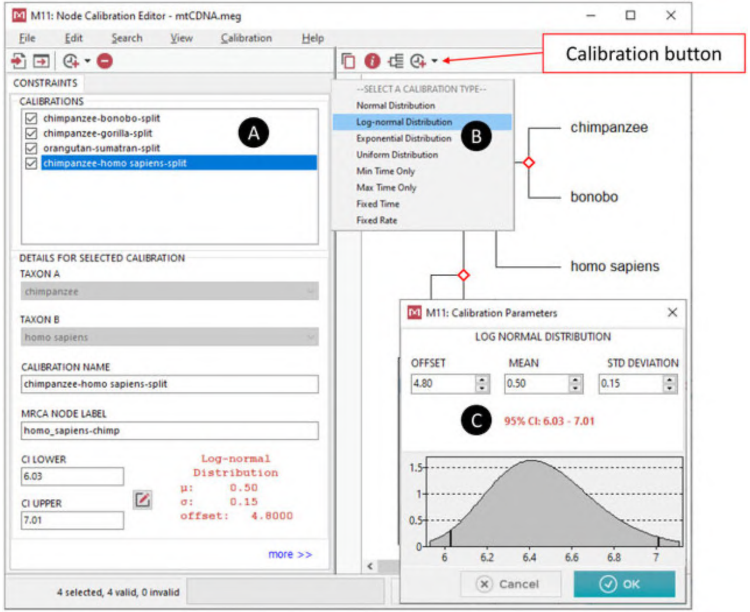
An empirical guide for producing a dated phylogeny with treePL in a maximum likelihood framework

Kévin J. L. Maurin

RelTime

Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates

Koichiro Tamura,[1,2] Qiqing Tao,[3,4] and Sudhir Kumar*[,3,4,5]



☑ chimpanzee-bonobo-split
☑ chimpanzee-gorilla-split
☑ orangutan-sumatran-split
☑ chimpanzee-homo sapiens-split

DETAILS FOR SELECTED CALIBRATION
TAXON A: chimpanzee
TAXON B: homo sapiens
CALIBRATION NAME: chimpanzee-homo sapiens-split
MRCA NODE LABEL: homo_sapiens-chimp
CI LOWER: 6.03
CI UPPER: 7.01

```
Log-normal
Distribution
µ:      0.50
σ:      0.15
offset:   4.8000
```

4 selected, 4 valid, 0 invalid

OFFSET: 4.80
MEAN: 0.50
STD DEVIATION: 0.15

95% CI: 6.03 - 7.01

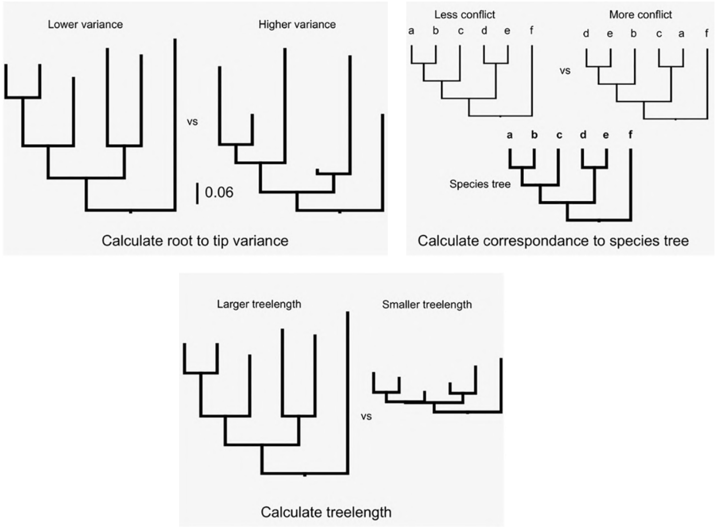
SortaDate

PLOS | ONE

RESEARCH ARTICLE

So many genes, so little time: A practical approach to divergence-time estimation in the genomic era

Stephen A. Smith[1]*, Joseph W. Brown[2☯], Joseph F. Walker[1☯]



genesortR

Phylogenomic Subsampling and the Search for Phylogenetically Reliable Loci

Nicolás Mongiardino Koch*

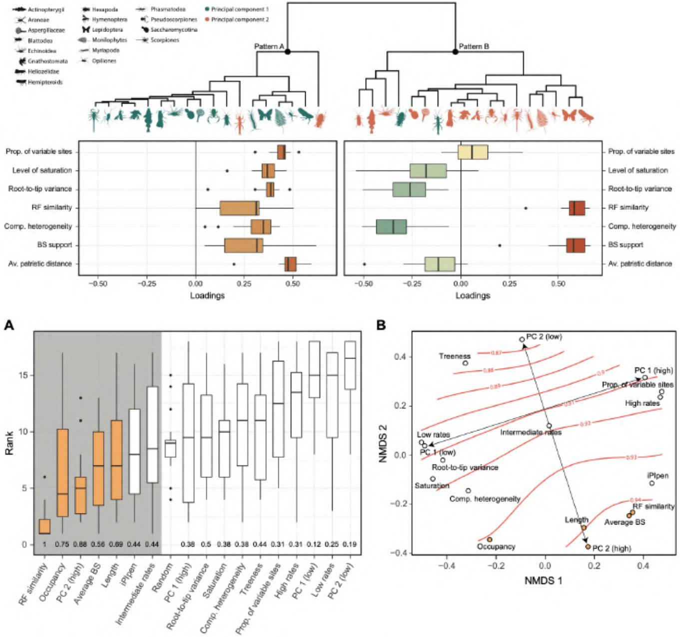
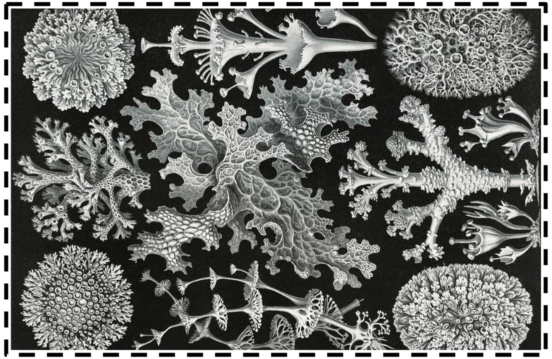
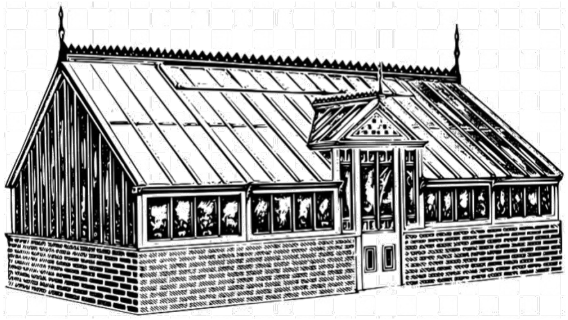


Table of contents





Biodiversity Genomics

Natural History Collections

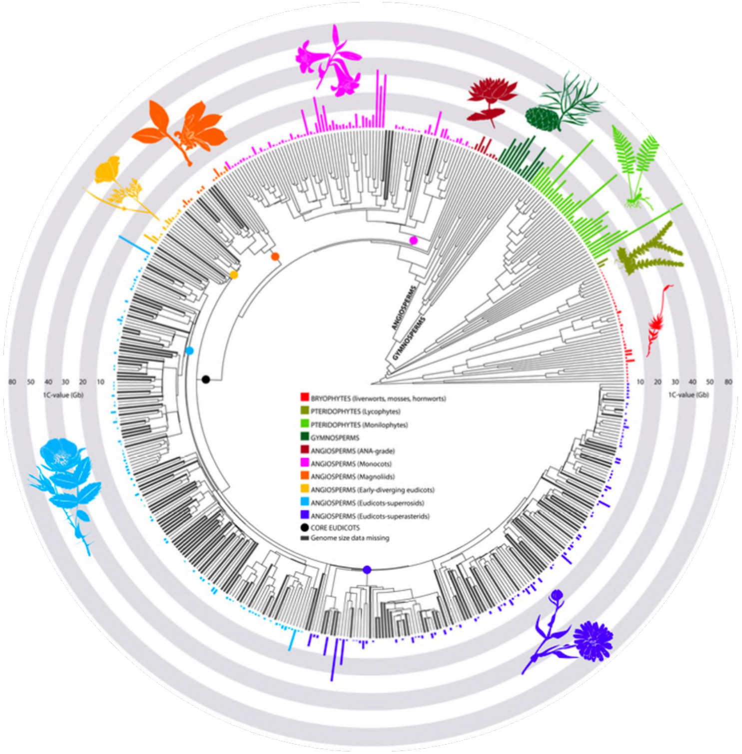
04

Explaining the fate of diversity

Understanding the eco-evolutionary processes that have shaped the traits leading to its origin and maintenance

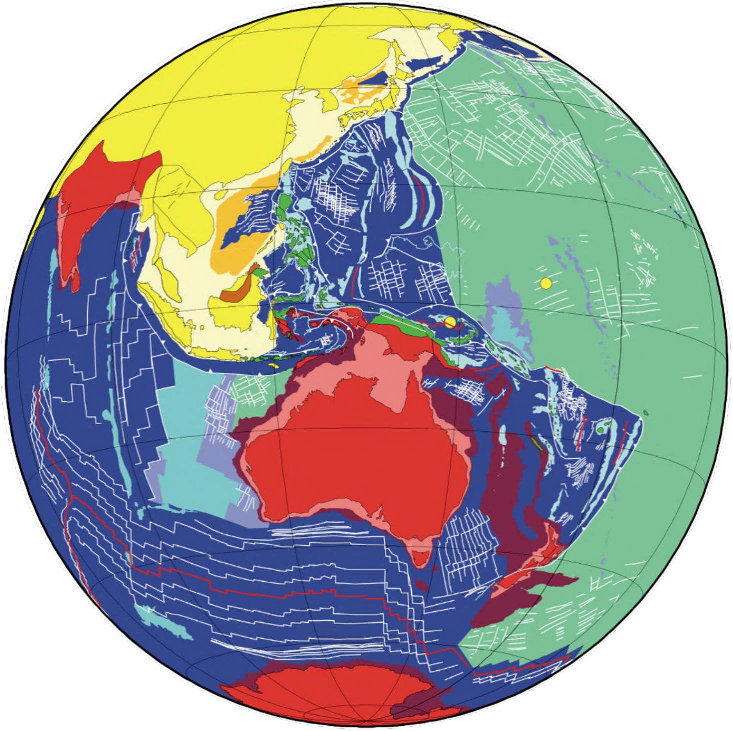
Understanding eco-evolutionary processes

Taking into account intrinsic factors...



Understanding eco-evolutionary processes

...and extrinsic factors
that, in turn, can
be abiotic...



Understanding eco-evolutionary processes

…or biotic

Understanding eco-evolutionary processes

The information stored in genomes is a record of past events…

Understanding eco-evolutionary processes

...also registered
in the fossil
record

Angiosperms353

A universal tool for flowering plant systematics

Angiosperms353

Classification: flowering plants, ~ 370K spp. (APG VI: 416 fams. & 64 orders)

Importance: agronomic, edible, timber, pulp, fiber, meds, ornamental, fuel, etc.

Available resources:
655 transcriptomes 1KP
20 genomes Phytozome
Hamming dist., k-medoids clustering
Daicel Arbor Biosciences myBaits

Resulting kit:
353 nuclear SCOGs
~260,8 Kbp matrix
~75,2K 120-mer 3x tiled probes

Syst. Biol. 68(4):594–606, 2019

© The Author(s) 2018. Published by Oxford University Press on behalf of the Society of Systematic Biologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

DOI:10.1093/sysbio/syy086
Advance Access publication December 10, 2018

A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering

MATTHEW G. JOHNSON[1,2,*], LISA POKORNY[3], STEVEN DODSWORTH[3,4], LAURA R. BOTIGUÉ[3,5], ROBYN S. COWAN[3], ALISON DEVAULT[6], WOLF L. EISERHARDT[3,7], NIROSHINI EPITAWALAGE[3], FÉLIX FOREST[3], JAN T. KIM[3], JAMES H. LEEBENS-MACK[8], ILIA J. LEITCH[3], OLIVIER MAURIN[3], DOUGLAS E. SOLTIS[9,10], PAMELA S. SOLTIS[9,10], GANE KA-SHU WONG[11,12,13], WILLIAM J. BAKER[3], AND NORMAN J. WICKETT[2,14]

[1]*Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA;* [2]*Plant Science and Conservation, Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, USA;* [3]*Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, UK;* [4]*School of Life Sciences, University of Bedfordshire, University Square, Luton LU1 3JU, UK;* [5]*Centre for Research in Agricultural Genomics, Campus UAB, Edifici CRAG, Bellaterra Cerdanyola del Vallès, 08193 Barcelona, Spain;* [6]*Arbor Biosciences, 5840 Interface Dr, Suite 101, Ann Arbor, MI 48103, USA;* [7]*Department of Bioscience, Aarhus University, 8000 Aarhus C, Denmark;* [8]*Department of Plant Biology, University of Georgia, 2502 Miller Plant Sciences, Athens, GA 30602, USA;* [9]*Department of Biology, University of Florida, 220 Bartram Hall, Gainesville, FL 32611-8525, USA;* [10]*Florida Museum of Natural History, University of Florida, 3215 Hull Road, Gainesville, FL 32611-2710, USA;* [11]*BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China;* [12]*Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada;* [13]*Department of Medicine, University of Alberta, Edmonton, AB T6G 2E1, Canada; and* [14]*Program in Plant Biology and Conservation, Northwestern University, 2205 Tech Drive, Evanston, IL 60208, USA*

**Correspondence to be sent to: Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA;
E-mail: matt.johnson@ttu.edu.*

Matthew G. Johnson, Lisa Pokorny, Steven Dodsworth contributed equally to this article.

Flowering plants (Angiospermae)

APG VI classification (2016): ~370K species, ~13.6 genera, 416 families, & 64 orders
Importance: 90% terrestrial spp., nutrition, fuels, lumber, meds, cultural, etc.

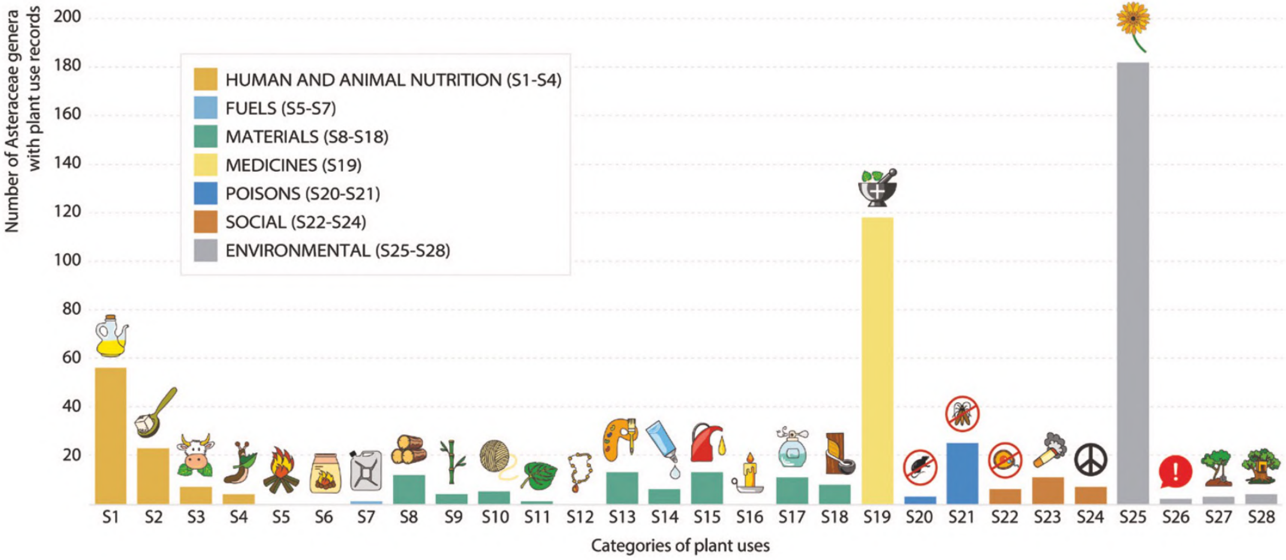
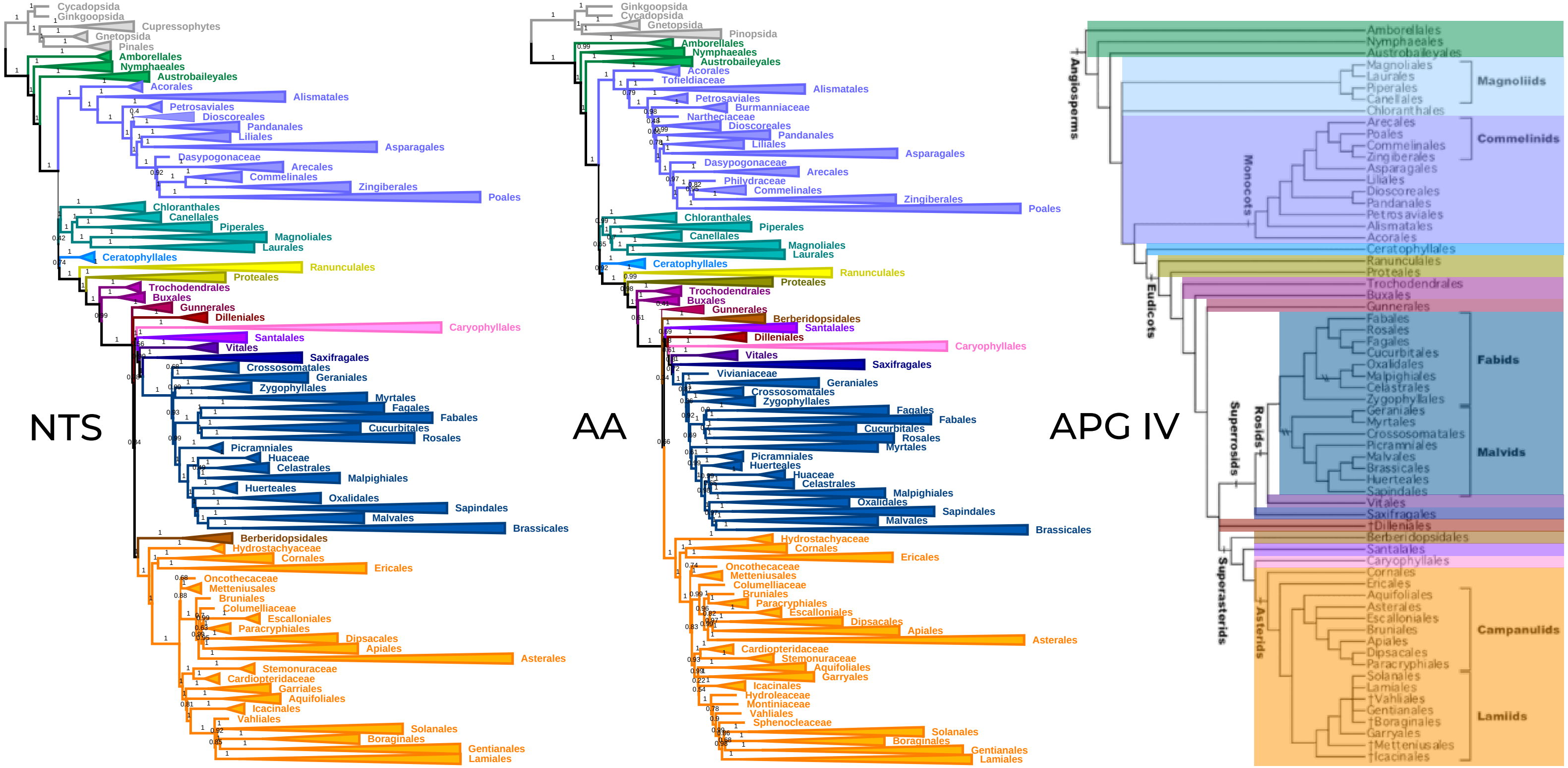


Fig. 2. Palazzesi et al. (2022). *Bot. J. Linn. Soc.* 200(2): 143–164.

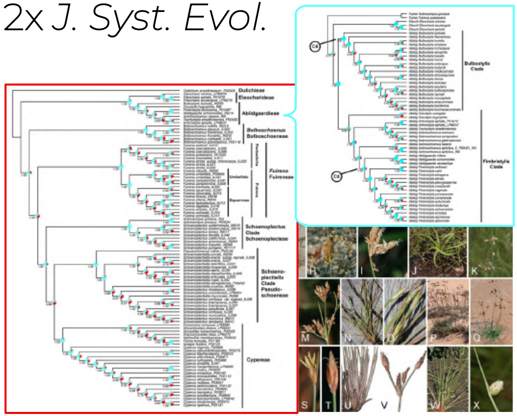
Angiosperms353



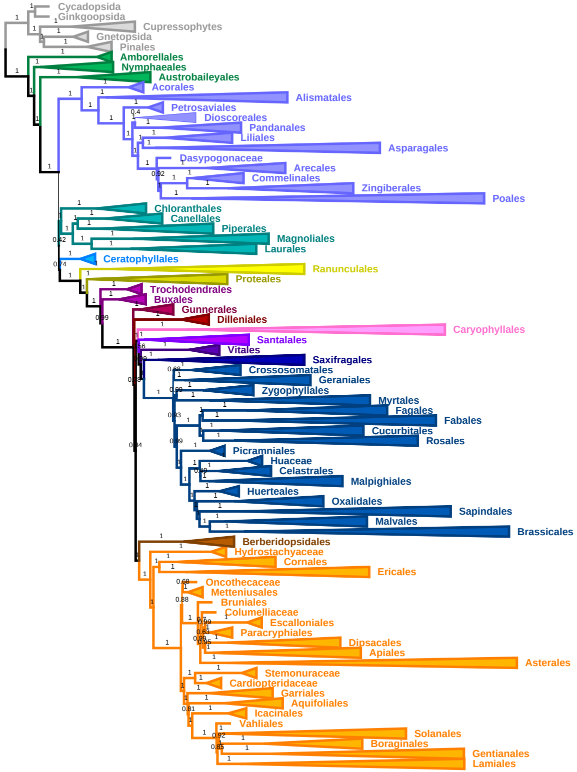
Angiosperms353

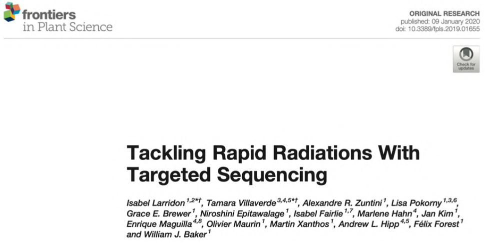
Commelinales: ~885 spp. (70 gen., 5 fams.)
Interest: ornamental (kangaroo paws, spiderworts), invasive (water hyacinth)
Publications: *Am. J. Bot.*

Cyperaceae (Poales): ~5,6K spp. (~90 gen.)
Interest: edible (chufa), pulp (papyrus)
Publications: *Front. Plant Sci.*, *Bot. J. Linn. Soc.*, & 2x *J. Syst. Evol.*



Angiosperms353



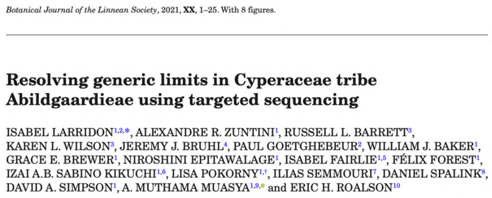

frontiers in Plant Science

ORIGINAL RESEARCH
published: 09 January 2020
doi: 10.3389/fpls.2019.01655

Check for updates

Tackling Rapid Radiations With Targeted Sequencing

Isabel Larridon[1,2†], Tamara Villaverde[3,4,5*†], Alexandre R. Zuntini[1], Lisa Pokorny[1,3,6], Grace E. Brewer[1], Niroshini Epitawalage[1], Isabel Fairlie[1,7], Marlene Hahn[4], Jan Kim[1], Enrique Maguilla[4,8], Olivier Maurin[1], Martin Xanthos[1], Andrew L. Hipp[4,5], Félix Forest[1] and William J. Baker[1]*


Botanical Journal of the Linnean Society, 2021, **XX**, 1–25. With 8 figures.

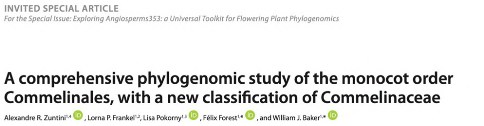
Resolving generic limits in Cyperaceae tribe Abildgaardieae using targeted sequencing

ISABEL LARRIDON[1,2,*], ALEXANDRE R. ZUNTINI[1], RUSSELL L. BARRETT[3], KAREN L. WILSON[3], JEREMY J. BRUHL[4], PAUL GOETGHEBEUR[2], WILLIAM J. BAKER[1], GRACE E. BREWER[1], NIROSHINI EPITAWALAGE[1], ISABEL FAIRLIE[1,5], FÉLIX FOREST[1], IZAI A.B. SABINO KIKUCHI[1,6], LISA POKORNY[1,†], ILIAS SEMMOURI[7], DANIEL SPALINK[8], DAVID A. SIMPSON[1], A. MUTHAMA MUASYA[1,9] and ERIC H. ROALSON[10]


American Journal of Botany

Check for updates

RESEARCH ARTICLE


INVITED SPECIAL ARTICLE

For the Special Issue: Exploring Angiosperms353: a Universal Toolkit for Flowering Plant Phylogenomics

A comprehensive phylogenomic study of the monocot order Commelinales, with a new classification of Commelinaceae

Alexandre R. Zuntini[1,4], Lorna P. Frankel[1,2], Lisa Pokorny[1,3], Félix Forest[1,*], and William J. Baker[1,*]

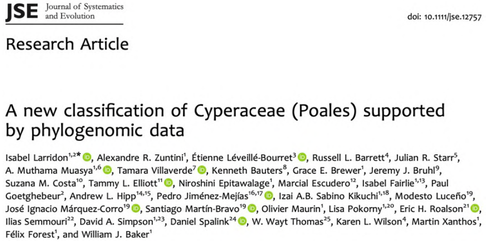

JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12721

Research Article

Targeted sequencing supports morphology and embryo features in resolving the classification of Cyperaceae tribe Fuireneae s.l.

Julian R. Starr[1*], Pedro Jiménez-Mejías[2,3], Alexandre R. Zuntini[4], Étienne Léveillé-Bourret[5], Ilias Semmouri[6], Muthama Muasya[7], William J. Baker[4], Grace E. Brewer[4], Niroshini Epitawalage[4], Isabel Fairlie[4,8], Félix Forest[4], Izai A. B. Sabino Kikuchi[4,9], Lisa Pokorny[4], and Isabel Larridon[4,10]


JSE Journal of Systematics and Evolution

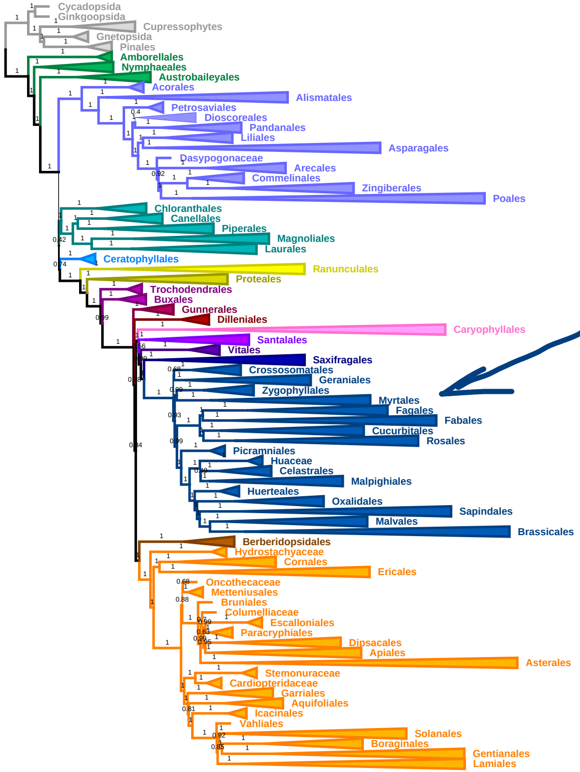
doi: 10.1111/jse.12757

Research Article

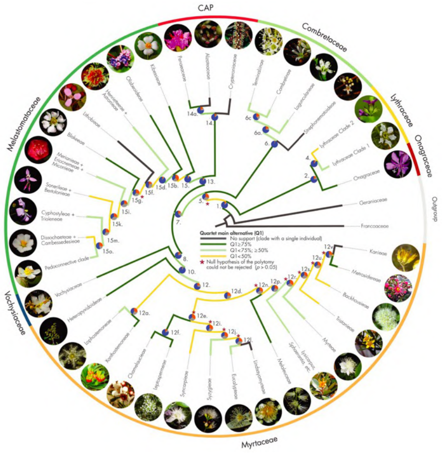
A new classification of Cyperaceae (Poales) supported by phylogenomic data

Isabel Larridon[1,2*], Alexandre R. Zuntini[1], Étienne Léveillé-Bourret[3], Russell L. Barrett[4], Julian R. Starr[5], A. Muthama Muasya[1,6], Tamara Villaverde[7], Kenneth Bauters[8], Grace E. Brewer[1], Jeremy J. Bruhl[9], Suzana M. Costa[10], Tammy L. Elliott[11], Niroshini Epitawalage[1], Marcial Escudero[12], Isabel Fairlie[1,13], Paul Goetghebeur[2], Andrew L. Hipp[14,15], Pedro Jiménez-Mejías[16,17], Izai A.B. Sabino Kikuchi[1,18], Modesto Luceño[19], José Ignacio Márquez-Corro[19], Santiago Martín-Bravo[19], Olivier Maurin[1], Lisa Pokorny[1,20], Eric H. Roalson[21], Ilias Semmouri[22], David A. Simpson[1,23], Daniel Spalink[24], W. Wayt Thomas[25], Karen L. Wilson[4], Martin Xanthos[1], Félix Forest[1], and William J. Baker[1]

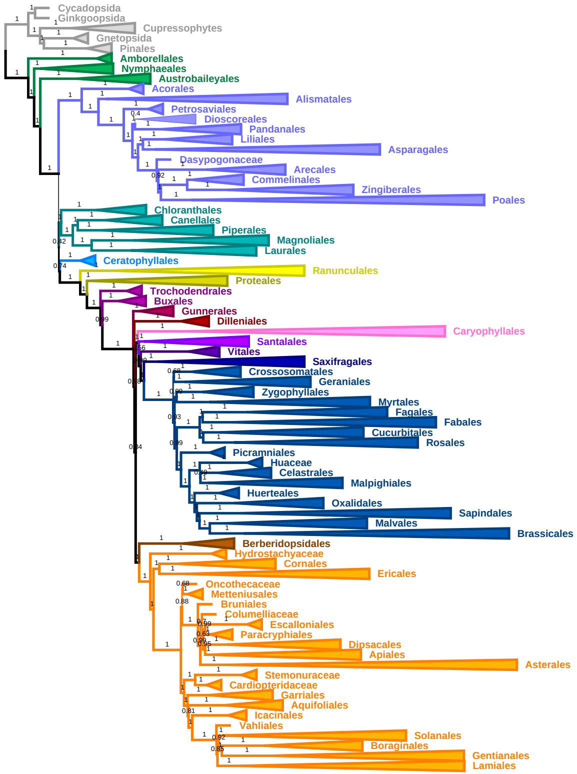
Angiosperms353

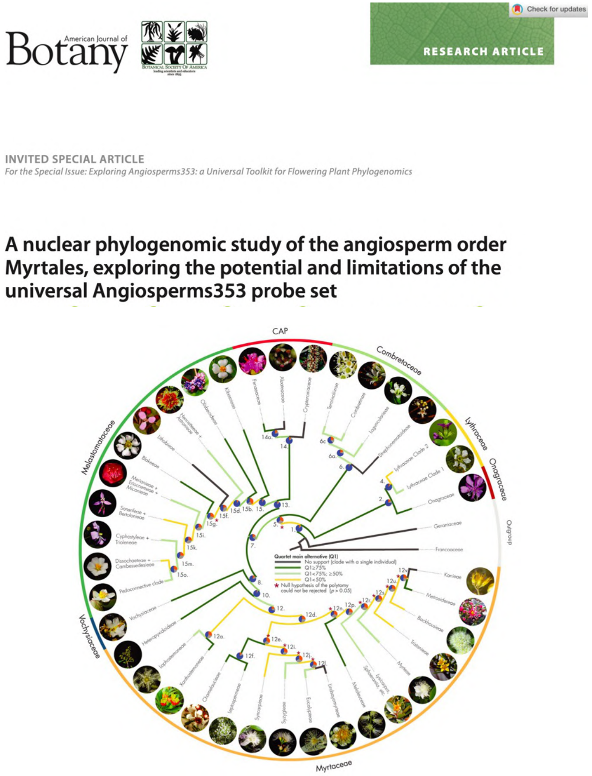


Myrtales: ~13K spp. (~380 gen. 9 fams.)
Interest: edible (fodder), edible (pomegranate, guava, cloves), ornamental (fuchsia, evening primroses), pulp (eucalypts), essential oils (sundrops)
Publications: *Am. J. Bot.*



Angiosperms353




American Journal of Botany

Check for updates

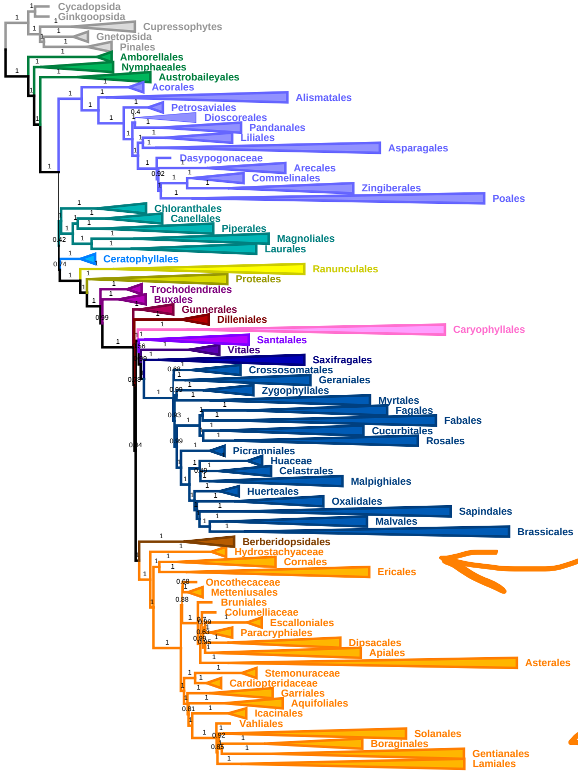
RESEARCH ARTICLE

INVITED SPECIAL ARTICLE

For the Special Issue: Exploring Angiosperms353: a Universal Toolkit for Flowering Plant Phylogenomics

A nuclear phylogenomic study of the angiosperm order Myrtales, exploring the potential and limitations of the universal Angiosperms353 probe set

Angiosperms353



Cornales: ~600 spp. (~50 gen., 6 fams.)

Interest: ornamental (hortensias, mock-oranges, dogwoods), lumber (tupelos)

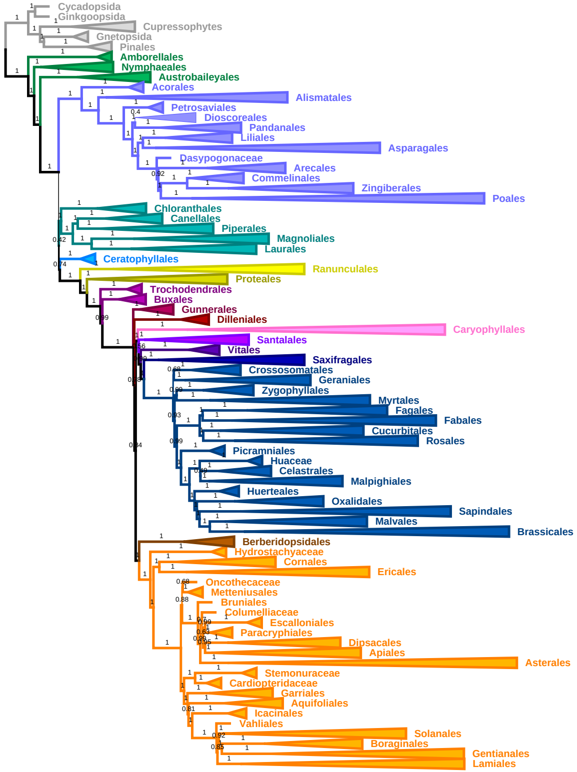
Articles: *Am. J. Bot.*

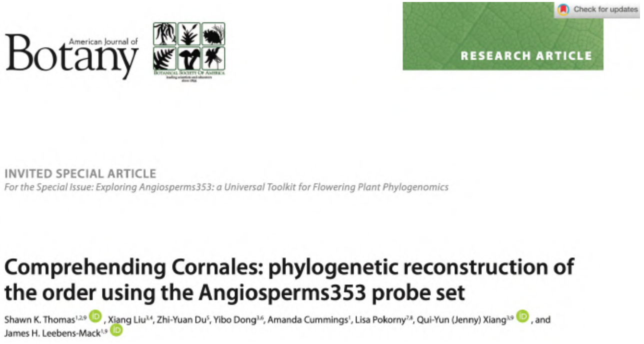
Gentianales: ~22,5K spp. (1,1K gen., 5 fams.)

Interest: drinks (coffee), food (soursop), meds (quinine, strychnine, vincristine), ornamental (gardenia, oleander), dyes.

Articles: *Am. J. Bot.*

Angiosperms353





American Journal of Botany

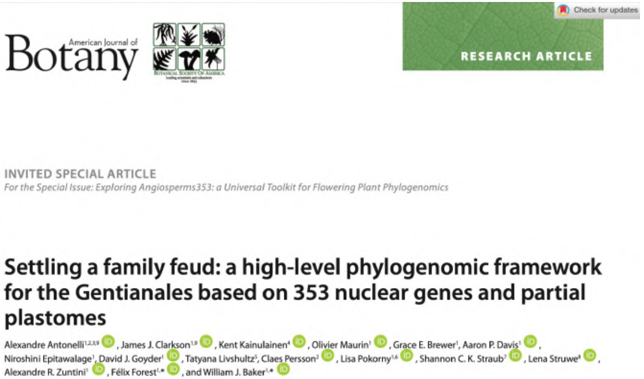
RESEARCH ARTICLE

INVITED SPECIAL ARTICLE

For the Special Issue: Exploring Angiosperms353: a Universal Toolkit for Flowering Plant Phylogenomics

Comprehending Cornales: phylogenetic reconstruction of the order using the Angiosperms353 probe set

Shawn K. Thomas[1,2,9], Xiang Liu[3,4], Zhi-Yuan Du[5], Yibo Dong[3,6], Amanda Cummings[1], Lisa Pokorny[7,8], Qui-Yun (Jenny) Xiang[3,9], and James H. Leebens-Mack[1,9]



American Journal of Botany

RESEARCH ARTICLE

INVITED SPECIAL ARTICLE

For the Special Issue: Exploring Angiosperms353: a Universal Toolkit for Flowering Plant Phylogenomics

Settling a family feud: a high-level phylogenomic framework for the Gentianales based on 353 nuclear genes and partial plastomes

Alexandre Antonelli[1,2,3,9], James J. Clarkson[1,9], Kent Kainulainen[4], Olivier Maurin[1], Grace E. Brewer[1], Aaron P. Davis[1], Niroshini Epitawalage[1], David J. Goyder[1], Tatyana Livshultz[5], Claes Persson[2], Lisa Pokorny[1,6], Shannon C. K. Straub[7], Lena Struwe[8], Alexandre R. Zuntini[1], Félix Forest[1,*], and William J. Baker[1,*]

Flowering Plant Quipu

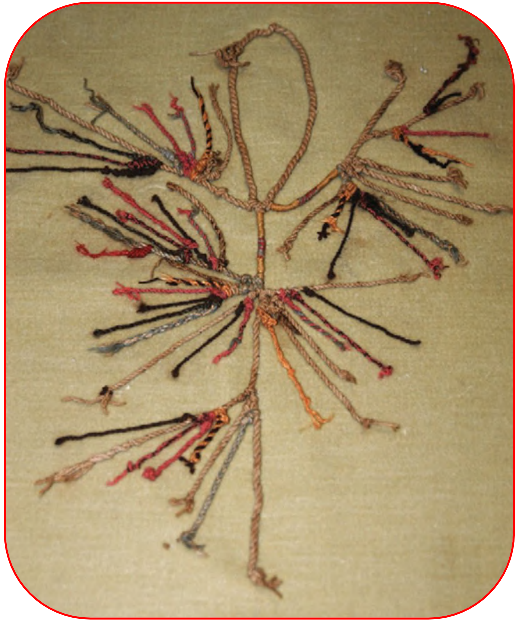
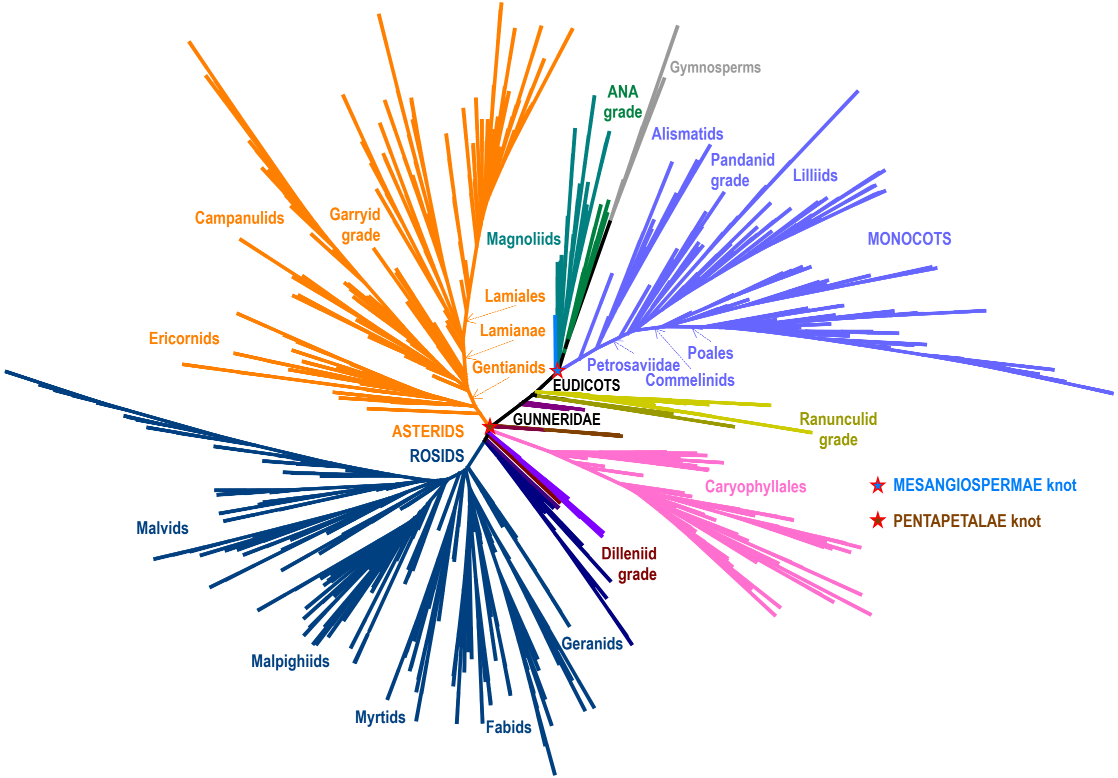


Fig. 8. Loop-and-branch-type khipu. Urton. 2014. *Antiquity.*



Population Genomics

Applications in Plant Sciences

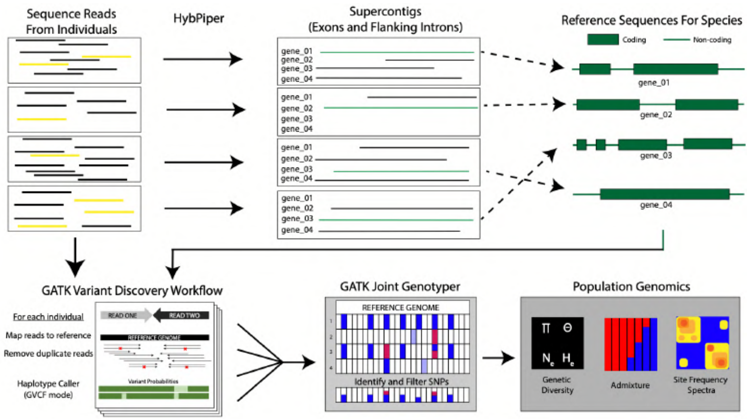
BOTANICAL SOCIETY OF AMERICA
leading scientists and educators
since 1893

APPLICATION ARTICLE

INVITED SPECIAL ARTICLE
For the Special Issue: Exploring Angiosperms353: A Universal Toolkit for Flowering Plant Phylogenomics

On the potential of Angiosperms353 for population genomic studies

Madeline Slimp[1], Lindsay D. Williams[1], Haley Hale[1], and Matthew G. Johnson[1,2]



Systematic Botany (2021), 46(4): pp. 1107–1113
© Copyright 2021 by the American Society of Plant Taxonomists
DOI 10.1600/036364421X16370109698740
Date of publication December 21, 2021

Are Palmer's Elm-Leaf Goldenrod and the Smooth Elm-Leaf Goldenrod Real? The Angiosperms353 Kit Provides Within-Species Signal in *Solidago ulmifolia* s. l.

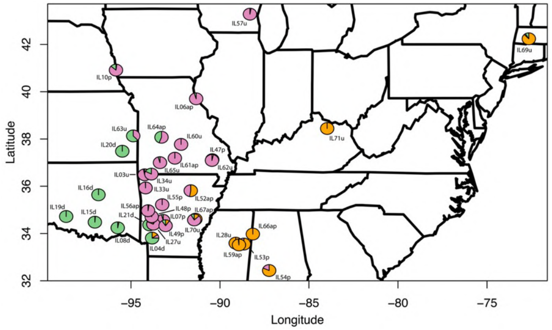
James B. Beck,[1,2,5] Morgan L. Markley,[1] Mackenzie G. Zielke,[1] Justin R. Thomas,[3] Haley J. Hale,[4] Lindsay D. Williams[4], and Matthew G. Johnson[4]

[1]Department of Biological Sciences, Wichita State University, Wichita, Kansas 67260, USA
[2]Botanical Research Institute of Texas, Fort Worth, Texas 76107, USA
[3]NatureCITE, Springfield, Missouri 65803, USA
[4]Department of Biological Sciences, Texas Tech University, Lubbock, Texas 79409, USA
[5]Author for correspondence (james.beck@wichita.edu)

Communicating Editor: Ashley B. Morris

Abstract—The genus *Solidago* represents a taxonomically challenging group due to its sheer number of species, putative hybridization, polyploidy, and shallow genetic divergence among species. Here we use a dataset obtained exclusively from herbarium specimens to evaluate the status of *Solidago ulmifolia* var. *palmeri*, a morphologically subtle taxon potentially confined to Alabama, Arkansas, Mississippi, and Missouri. A multivariate analysis of both discrete and continuous morphological data revealed no clear distinction between *S. ulmifolia* var. *palmeri* and *Solidago ulmifolia* var. *ulmifolia*. *Solidago ulmifolia* var. *palmeri*'s status was also assessed with a phylogenomic and SNP clustering analysis of data generated with the "Angiosperms353" probe kit. Neither analysis supported *Solidago ulmifolia* var. *palmeri* as a distinct taxon, and we suggest that this name should be discarded. The status of *Solidago delicatula* (formerly known as *Solidago ulmifolia* var. *microphylla*) was also assessed. Both morphological and phylogenetic analyses supported the species status of *S. delicatula* and we suggest maintaining this species at its current rank. These results highlight the utility of the Angiosperms353 probe kit, both with herbarium tissue and at lower taxonomic levels. Indeed, this is the first study to utilize this kit to identify genetic groups within a species.

Keywords—hyb-seq, North America, Ozark Mountains, *Solidago delicatula*, *S. ulmifolia* var. *palmeri*, species delimitation.

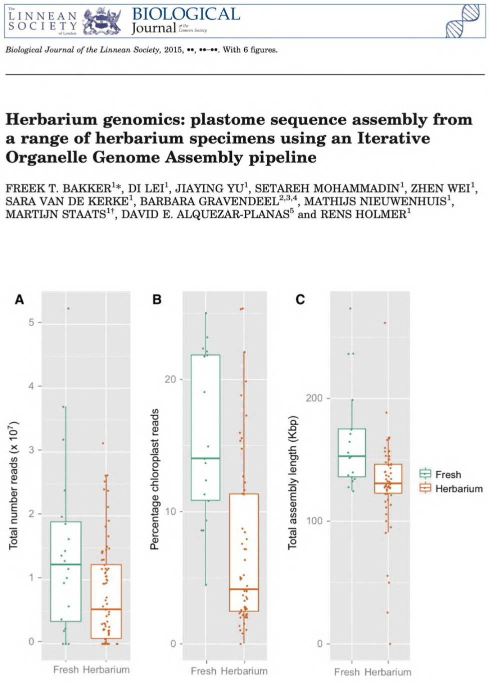


Museomics

Herbariomics / Fungariomics

Integrating Natural History Collections into Biodiversity Research

Herbariomics / Fungariomics

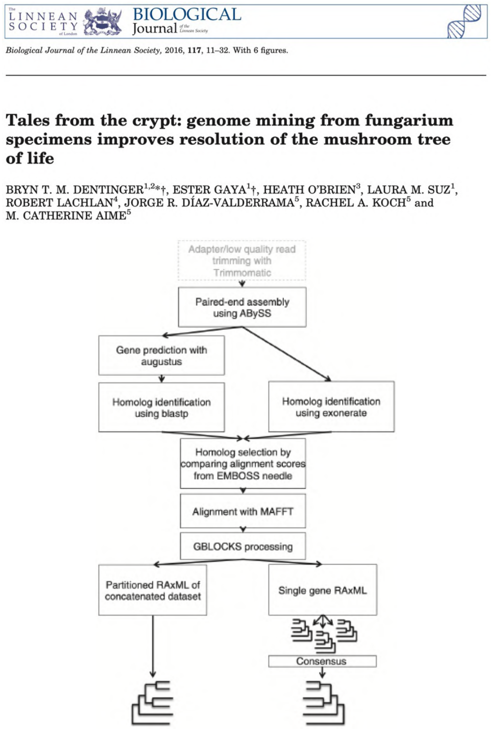


The LINNEAN SOCIETY of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2015, ••, ••–••. With 6 figures.

Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline

FREEK T. BAKKER[1]*, DI LEI[1], JIAYING YU[1], SETAREH MOHAMMADIN[1], ZHEN WEI[1], SARA VAN DE KERKE[1], BARBARA GRAVENDEEL[2,3,4], MATHIJS NIEUWENHUIS[1], MARTIJN STAATS[1†], DAVID E. ALQUEZAR-PLANAS[5] and RENS HOLMER[1]



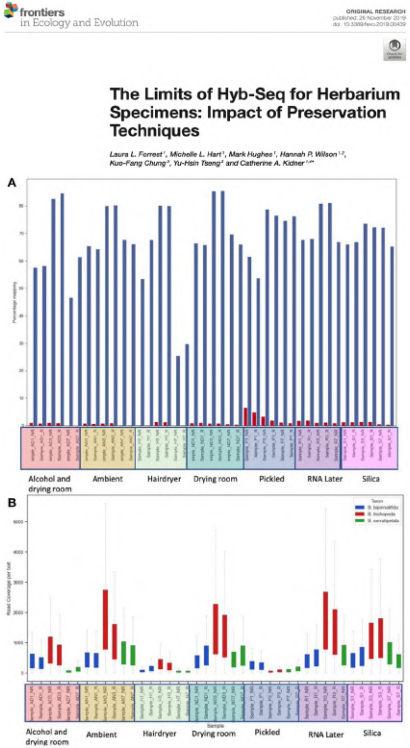
The LINNEAN SOCIETY of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2016, **117**, 11–32. With 6 figures.

Tales from the crypt: genome mining from fungarium specimens improves resolution of the mushroom tree of life

BRYN T. M. DENTINGER[1,2]*†, ESTER GAYA[1]†, HEATH O'BRIEN[3], LAURA M. SUZ[1], ROBERT LACHLAN[4], JORGE R. DÍAZ-VALDERRAMA[5], RACHEL A. KOCH[5] and M. CATHERINE AIME[5]

Herbariomics

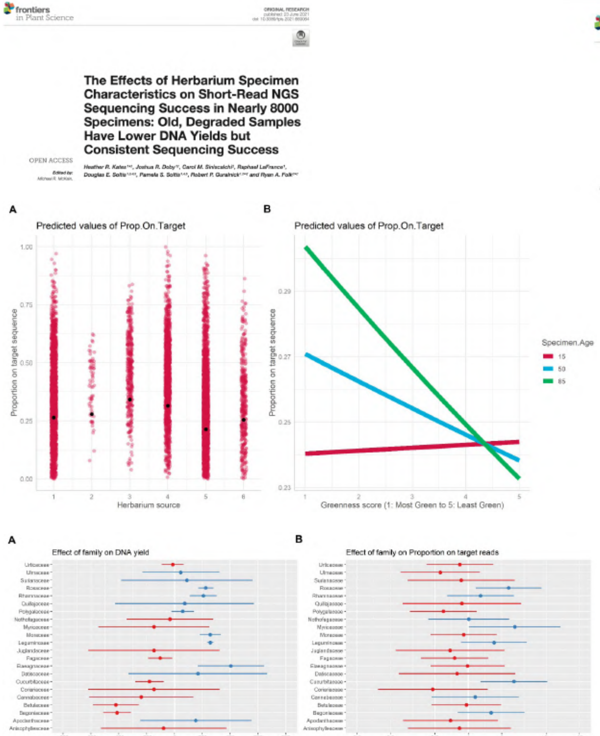


frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 26 November 2019
doi: 10.3389/fevo.2019.00439

The Limits of Hyb-Seq for Herbarium Specimens: Impact of Preservation Techniques

*Laura L. Forrest¹, Michelle L. Hart¹, Mark Hughes¹, Hannah P. Wilson¹,², Kuo-Fang Chung³, Yu-Hsin Tseng³ and Catherine A. Kidner¹,⁴**

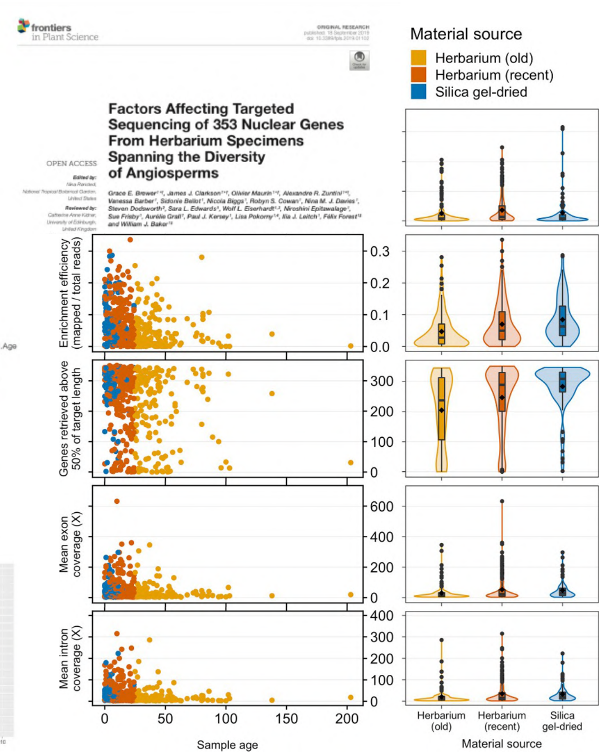


frontiers in Plant Science

ORIGINAL RESEARCH
published: 23 June 2021

The Effects of Herbarium Specimen Characteristics on Short-Read NGS Sequencing Success in Nearly 8000 Specimens: Old, Degraded Samples Have Lower DNA Yields but Consistent Sequencing Success

OPEN ACCESS



frontiers in Plant Science

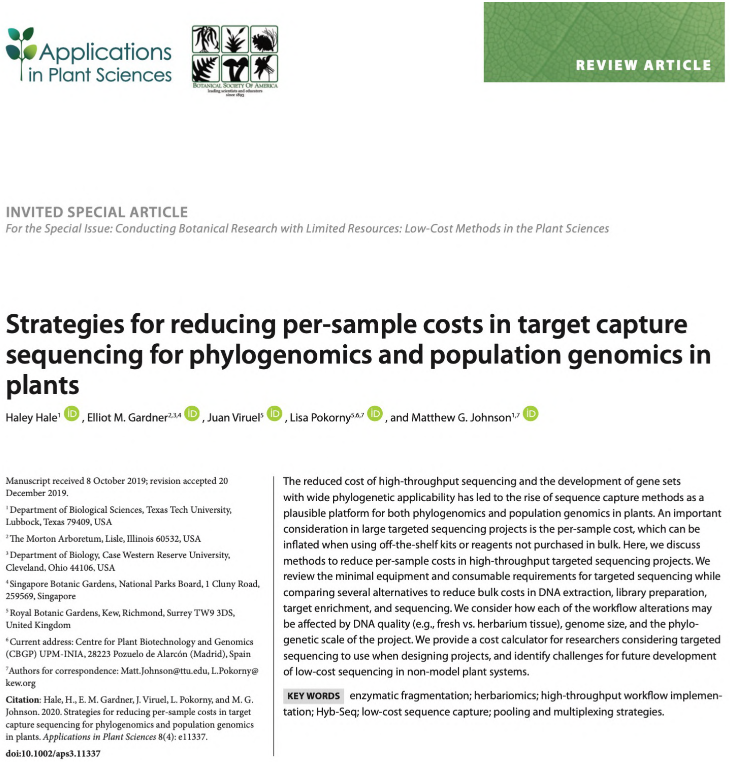
ORIGINAL RESEARCH
published: 18 September 2019

Factors Affecting Targeted Sequencing of 353 Nuclear Genes From Herbarium Specimens Spanning the Diversity of Angiosperms

OPEN ACCESS

Grace E. Brower¹†, James J. Clarkson¹*†, Olivier Maurin¹*†, Alexandre R. Zuntini¹*†, Vanessa Barber¹, Sidonie Bellot¹, Nicola Biggs¹, Robyn S. Cowan¹, Nina M. J. Davies¹, Steven Dodsworth², Sara L. Edwards¹, Wolf L. Eiserhardt¹,³, Niroshini Epitawalage¹, Sue Frisby¹, Aurélie Grall¹, Paul J. Kersey¹, Lisa Pokorny¹,⁴, Ilia J. Leitch¹, Félix Forest¹† and William J. Baker¹†*

Herbariomics



Applications in Plant Sciences

BOTANICAL SOCIETY OF AMERICA

REVIEW ARTICLE

INVITED SPECIAL ARTICLE

For the Special Issue: Conducting Botanical Research with Limited Resources: Low-Cost Methods in the Plant Sciences

Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants

Haley Hale[1], Elliot M. Gardner[2,3,4], Juan Viruel[5], Lisa Pokorny[5,6,7], and Matthew G. Johnson[1,7]

Manuscript received 8 October 2019; revision accepted 20 December 2019.

[1] Department of Biological Sciences, Texas Tech University, Lubbock, Texas 79409, USA

[2] The Morton Arboretum, Lisle, Illinois 60532, USA

[3] Department of Biology, Case Western Reserve University, Cleveland, Ohio 44106, USA

[4] Singapore Botanic Gardens, National Parks Board, 1 Cluny Road, 259569, Singapore

[5] Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3DS, United Kingdom

[6] Current address: Centre for Plant Biotechnology and Genomics (CBGP) UPM-INIA, 28223 Pozuelo de Alarcón (Madrid), Spain

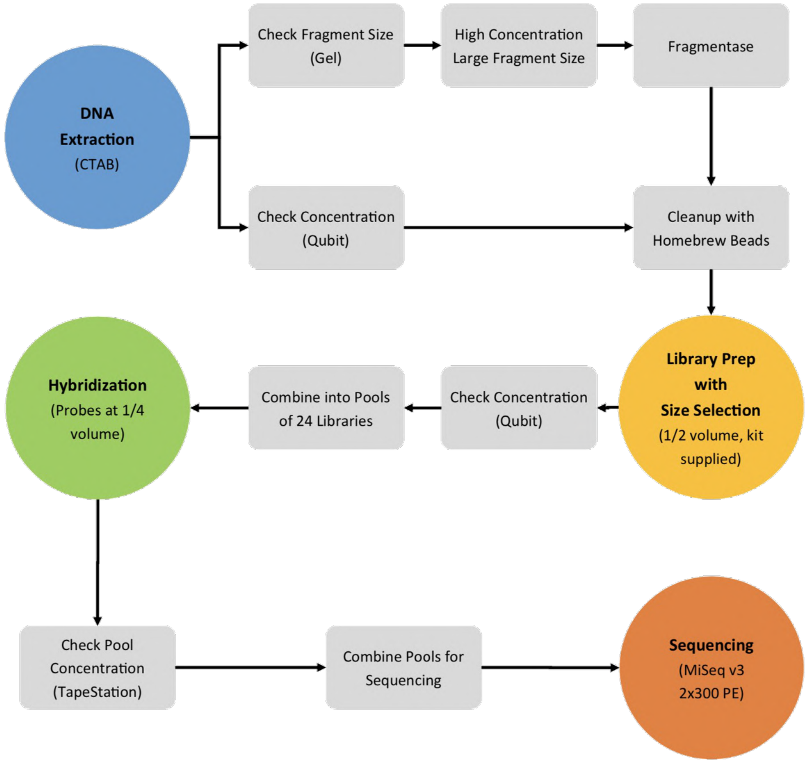
[7] Authors for correspondence: Matt.Johnson@ttu.edu, L.Pokorny@kew.org

Citation: Hale, H., E. M. Gardner, J. Viruel, L. Pokorny, and M. G. Johnson. 2020. Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. *Applications in Plant Sciences* 8(4): e11337.

doi:10.1002/aps3.11337

The reduced cost of high-throughput sequencing and the development of gene sets with wide phylogenetic applicability has led to the rise of sequence capture methods as a plausible platform for both phylogenomics and population genomics in plants. An important consideration in large targeted sequencing projects is the per-sample cost, which can be inflated when using off-the-shelf kits or reagents not purchased in bulk. Here, we discuss methods to reduce per-sample costs in high-throughput targeted sequencing projects. We review the minimal equipment and consumable requirements for targeted sequencing while comparing several alternatives to reduce bulk costs in DNA extraction, library preparation, target enrichment, and sequencing. We consider how each of the workflow alterations may be affected by DNA quality (e.g., fresh vs. herbarium tissue), genome size, and the phylogenetic scale of the project. We provide a cost calculator for researchers considering targeted sequencing to use when designing projects, and identify challenges for future development of low-cost sequencing in non-model plant systems.

KEY WORDS enzymatic fragmentation; herbariomics; high-throughput workflow implementation; Hyb-Seq; low-cost sequence capture; pooling and multiplexing strategies.



DNA Extraction (CTAB)

Check Fragment Size (Gel)

High Concentration Large Fragment Size

Fragmentase

Check Concentration (Qubit)

Cleanup with Homebrew Beads

Library Prep with Size Selection (1/2 volume, kit supplied)

Check Concentration (Qubit)

Combine into Pools of 24 Libraries

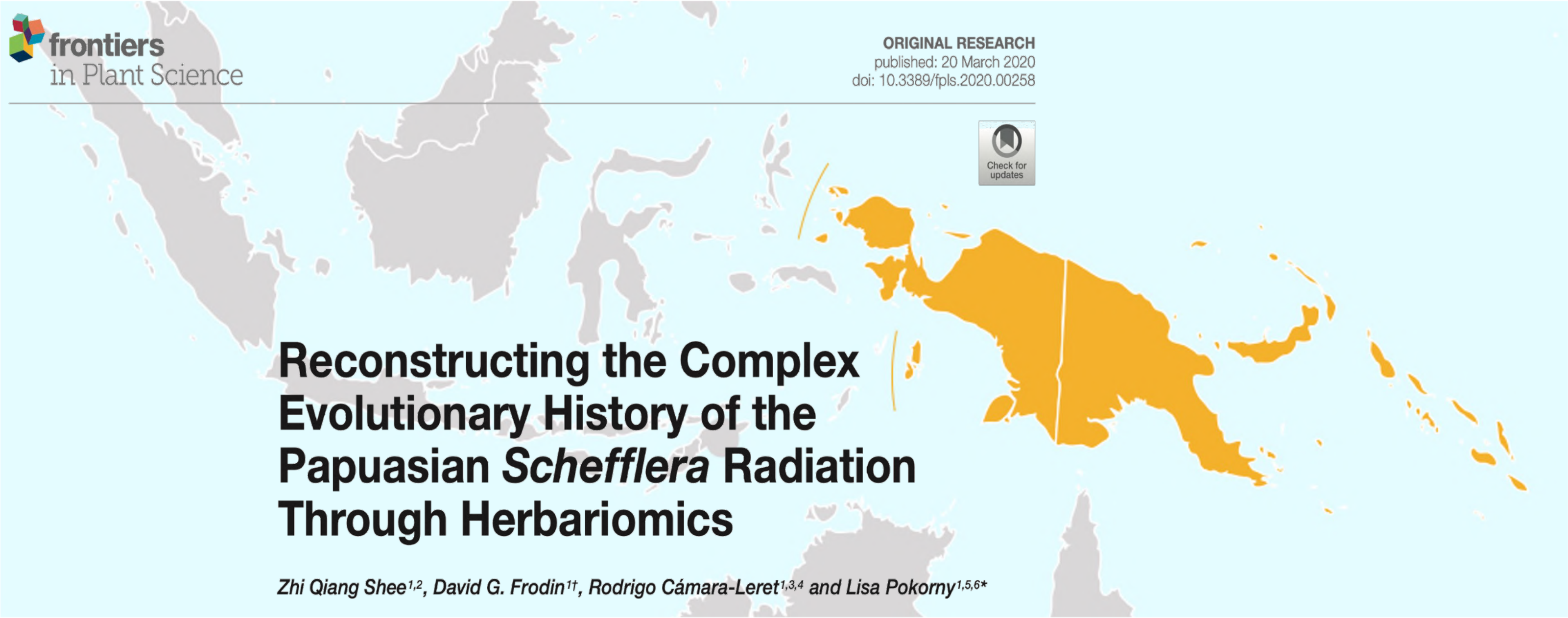
Hybridization (Probes at 1/4 volume)

Check Pool Concentration (TapeStation)

Combine Pools for Sequencing

Sequencing (MiSeq v3 2x300 PE)

Anatomy of a Plant Radiation



frontiers
in Plant Science

ORIGINAL RESEARCH
published: 20 March 2020
doi: 10.3389/fpls.2020.00258

Reconstructing the Complex Evolutionary History of the Papuasian *Schefflera* Radiation Through Herbariomics

Zhi Qiang Shee[1,2], David G. Frodin[1†], Rodrigo Cámara-Leret[1,3,4] and Lisa Pokorny[1,5,6]*

Assembling the Richest Flora

Article

New Guinea has the world's richest island flora

https://doi.org/10.1038/s41586-020-2549-5

Received: 2 October 2019

Accepted: 29 June 2020

Published online: 05 August 2020

Check for updates

A list of authors and their affiliations appears at the end of the paper.

New Guinea is the world's largest tropical island and has fascinated naturalists for centuries[1,2]. Home to some of the best-preserved ecosystems on the planet[3] and to intact ecological gradients—from mangroves to tropical alpine grasslands—that are unmatched in the Asia-Pacific region[4,5], it is a globally recognized centre of biological and cultural diversity[6,7]. So far, however, there has been no attempt to critically catalogue the entire vascular plant diversity of New Guinea. Here we present the first, to our knowledge, expert-verified checklist of the vascular plants of mainland New Guinea and surrounding islands. Our publicly available checklist includes 13,634 species (68% endemic), 1,742 genera and 264 families—suggesting that New Guinea is the most floristically diverse island in the world. Expert knowledge is essential for building checklists in the digital era: reliance on online taxonomic resources alone would have inflated species counts by 22%. Species discovery shows no sign of levelling off, and we discuss steps to accelerate botanical research in the 'Last Unknown'[8].

Great uncertainty remains as to the number of New Guinea plant species known to science, with conflicting estimates ranging from 9,000 to 25,000 species[9,10]. To narrow this range, here we catalogue the entire known vascular flora (angiosperms, gymnosperms, ferns and lycophytes) of mainland New Guinea and its surrounding islands (hereafter 'New Guinea'; Fig. 1a, Extended Data Fig. 1). We do so through a large-scale collaborative effort in which 99 plant experts verified the identity of 23,381 taxonomic names derived from 704,724 specimens (see Methods). Overall, we find that New Guinea supports 13,634 described species, 1,742 genera and 264 families of vascular plants (Supplementary Tables 1, 2). This suggests that New Guinea is the world's most floristically diverse island, with a known vascular plant flora 19% larger than the 11,488 species recorded in Madagascar[11] and 22% larger than the 11,165 species recorded in Borneo (http://www.plantsoftheworldonline.org, accessed 27 April 2019). New Guinea contains almost three times the 4,598 spermatophyte species of Java[12] and 1.4 times the 9,432 vascular plant species of the Philippines[13]—the only Malesian island regions for which Floras have been published. The vascular plant flora of New Guinea is divided between two political entities (Fig. 1a): Papua New Guinea, with 10,973 species, has 44% more species than Indonesian New Guinea (Papua Barat and Papua provinces), which has 7,616. Papua New Guinea also has more genera (1,654 versus 1,511) and families (260 versus 248). These differences partly arise from the lower collecting density in Indonesian New Guinea[1,2] (Fig. 1a). Nevertheless, the order of country rankings in plant diversity is unlikely to change with further collections because Papua New Guinea has a larger area, and surface area is the strongest predictor of island plant diversity[14]. Our species total for Papua New Guinea differs markedly from the 29,756 species that were presented in an unverified list of the Global Biodiversity Information Facility[15] and our total number of genera for New Guinea is 28% lower than the 2,437 unverified genera reported in a previous macroecological study[16]. Together, these differences underscore the need for expert validation in the digital era, which we discuss below.

Floristic patterns

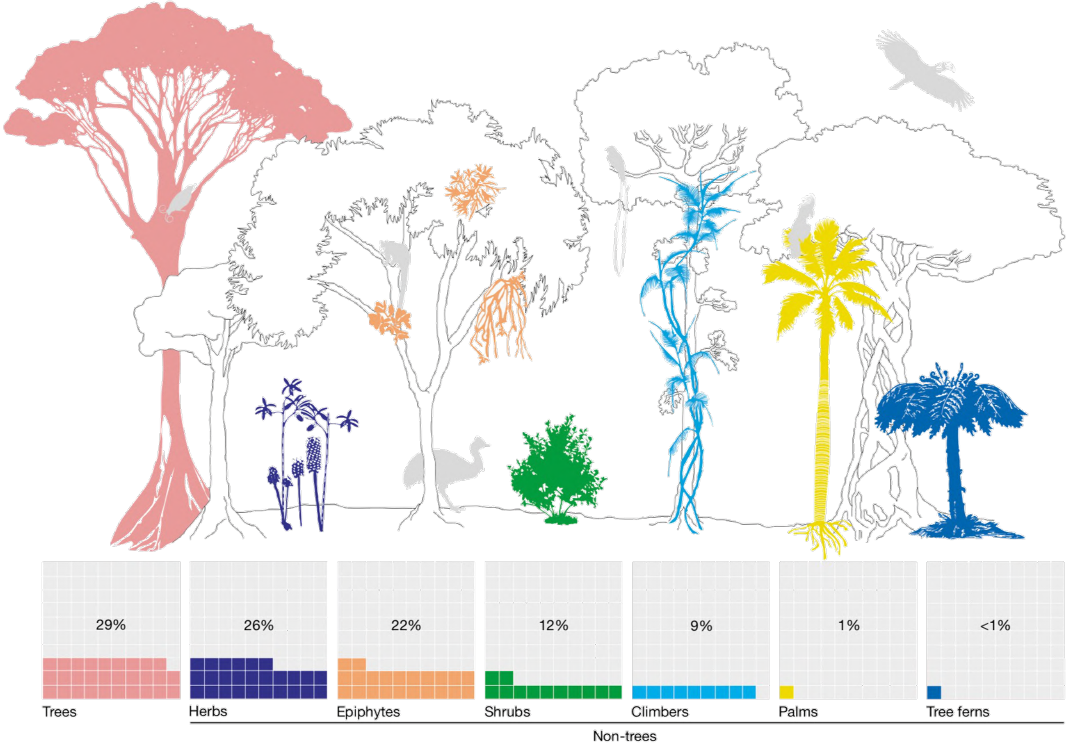
Five species-rich families make up 35% of the flora of New Guinea: Orchidaceae (2,856 species), Rubiaceae (784), Ericaceae (438), Poaceae (376) and Myrtaceae (352) (Fig. 1b, Extended Data Table 1). Orchidaceae account for 20% and 17% of the flora of Papua New Guinea and Indonesian New Guinea, respectively. The floristic importance of orchids is comparable to that in other megadiverse countries such as Ecuador (23% of total flora) and Colombia (15%)[17]. The five largest genera of vascular plants in New Guinea are *Bulbophyllum* (658 species; Orchidaceae), *Dendrobium* (614 species; Orchidaceae), *Syzygium* (207 species; Myrtaceae), *Ficus* (179 species; Moraceae) and *Rhododendron* (171 species; Ericaceae) (Fig. 2, Extended Data Table 2). Of the 1,742 genera found, 13 have more than 100 species and make up 21% of all species, whereas 692 genera are represented by a single species in New Guinea.

Endemism

Plant endemism in New Guinea is remarkably high: it is the only Malesian island group with more endemic than non-endemic species (9,301 endemic species; 68% of the total). This preponderance of endemic species was noted in earlier studies, although these were based on smaller floristic samples[9,18]. The uniqueness of New Guinea within Malesia may be explained by its greater land surface area and habitat diversity[5]; its location, marking the junction between Malesia, Australia and the Pacific; and its highly complex tectonic history[19]. Geographically, 53% of the endemic species have been found only in Papua New Guinea and 24% occur only in Indonesian New Guinea. Of the total species from Papua New Guinea, 64% are endemic, and 58% of the total species from Indonesian New Guinea are endemic. Such high richness of endemic species means that both countries have a unique responsibility for the survival of this irreplaceable biodiversity. Given the general trend of plant endemism to increase with elevation[20], the conservation of ecosystems along altitudinal gradients is particularly critical.

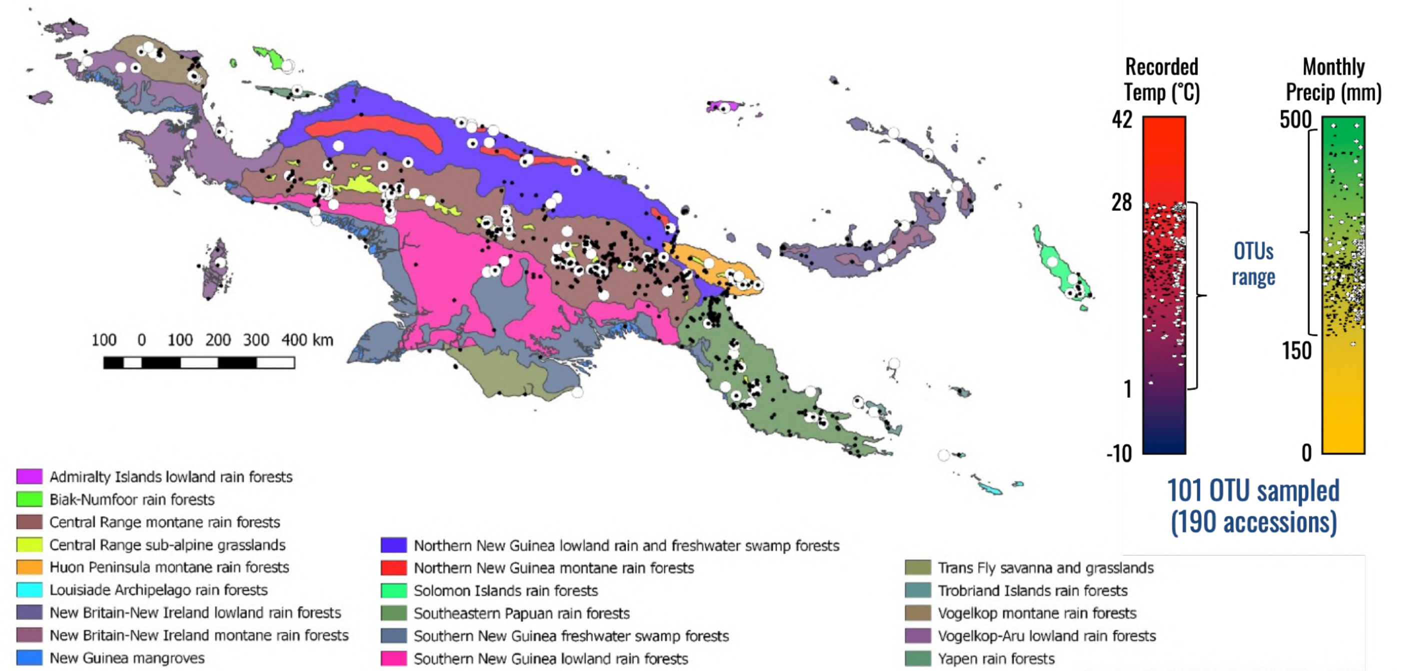
Angiosperms have higher species endemism (71%) than ferns and lycophytes (46%) or gymnosperms (41%). Endemism within families is highly uneven, with just eight angiosperm families comprising 50% of all endemics: Orchidaceae (2,464 endemic species), Rubiaceae (669), Ericaceae (431), Arecaceae (257), Myrtaceae (255), Gesneriaceae (218), Apocynaceae (196) and Lauraceae (195) (Fig. 1c). The families with the highest proportions of endemism are Ericaceae (98% of species

Nature | www.nature.com | 1



Umbrella Tree Herbariomics

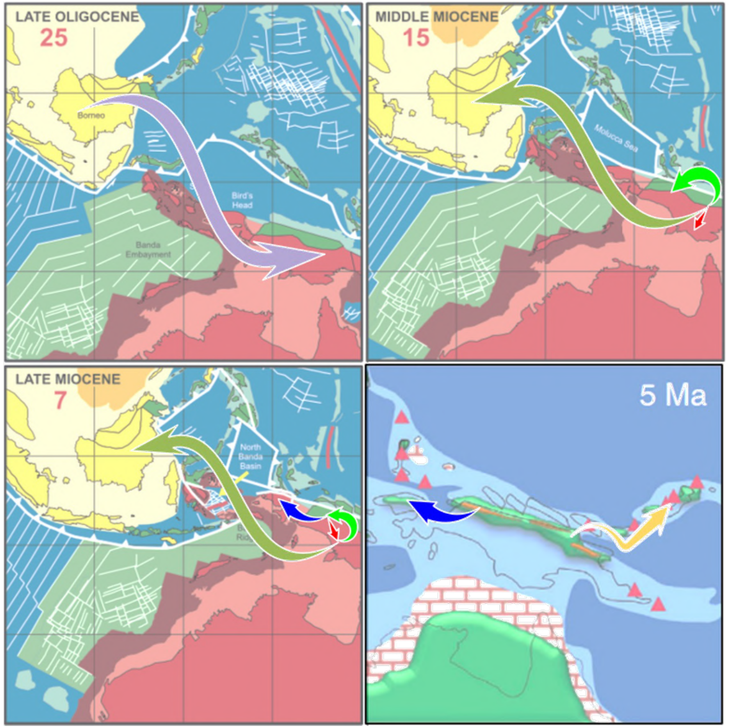
Umbrella Tree Herbariomics

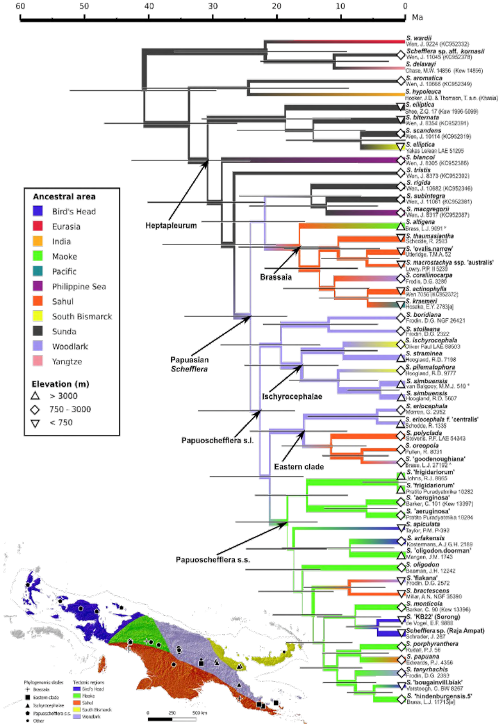


101 OTU sampled
(190 accessions)

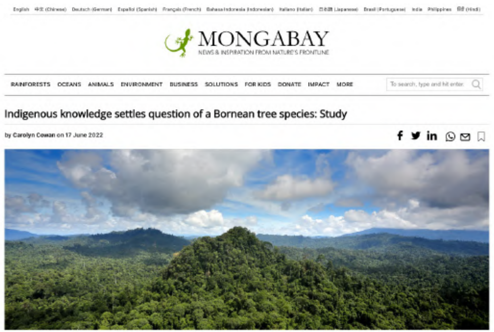
Data from World Wildlife Fund, WorldClim & DG Frodin

Umbrella Tree Geogenomics





Jackfruit Herbariomics



Indigenous knowledge settles question of a Bornean tree species: Study

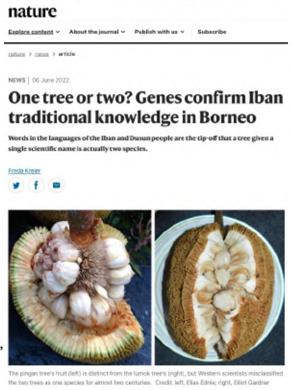
by Carolyn Cowan on 17 June 2022

Current Biology
Magazine

Correspondence

Engagement with indigenous people preserves local knowledge and biodiversity alike

Elliot M. Gardner[1,2,10,*], Aida Shafreena Ahmad Puad[3], Joan T. Pereira[4], Jugah anak Tagi[5], Salang anak Nyegang[6], Postar Miun[4], Jeisin Jumian[4], Lisa Pokorny[7], and Nyree J.C. Zerega[8,9]



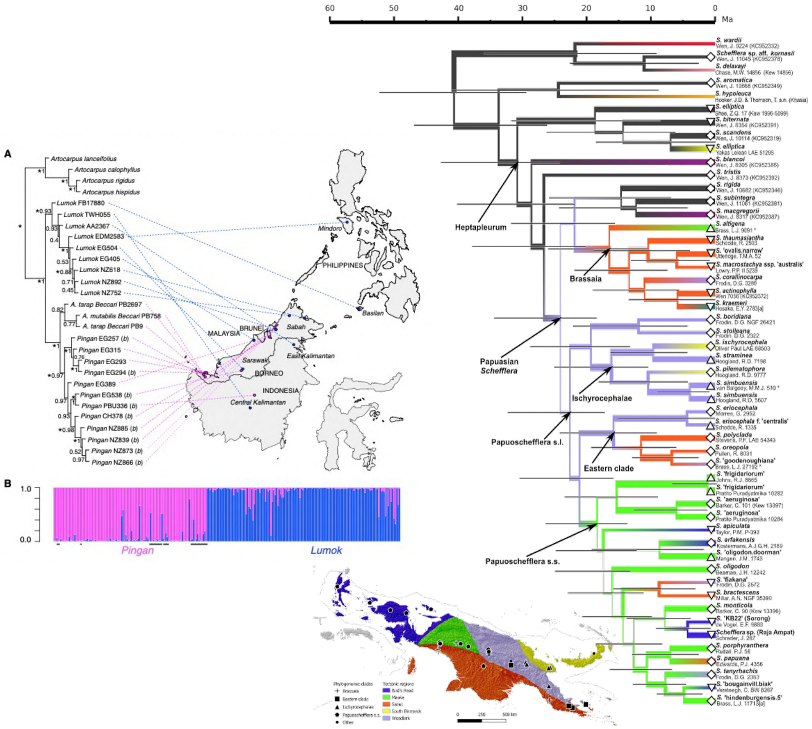
NEWS | 06 June 2022

One tree or two? Genes confirm Iban traditional knowledge in Borneo

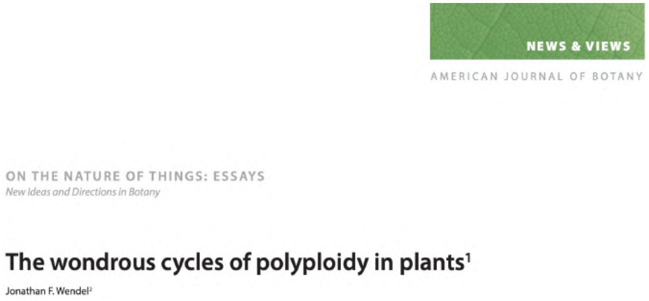
Words in the languages of the Iban and Dusun people are the tip-off that a tree given a single scientific name is actually two species.

Freda Kreier

The pingan tree's fruit (left) is distinct from the lumok tree's (right), but Western scientists misclassified the two trees as one species for almost two centuries. Credit: left, Elias Ednie; right, Elliot Gardner



Polyploidy in Plants



NEWS & VIEWS

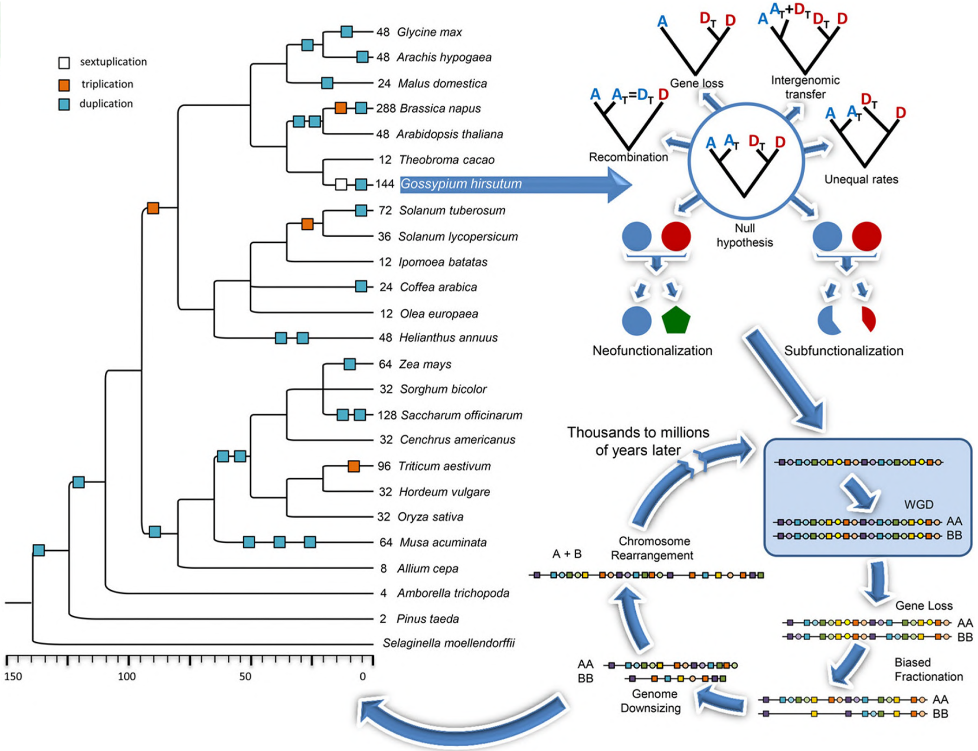
AMERICAN JOURNAL OF BOTANY

ON THE NATURE OF THINGS: ESSAYS

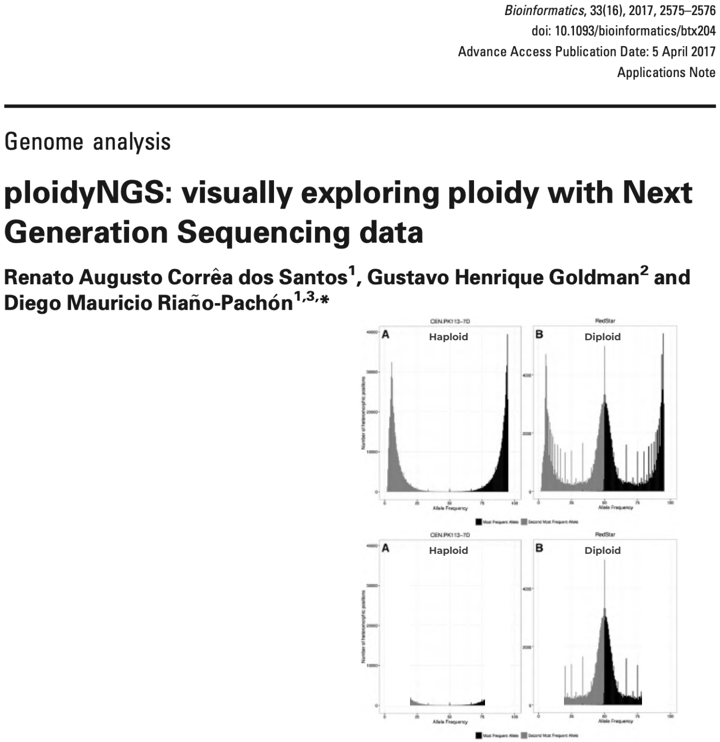
New Ideas and Directions in Botany

The wondrous cycles of polyploidy in plants[1]

Jonathan F. Wendel[2]



Ploidy Level Estimation

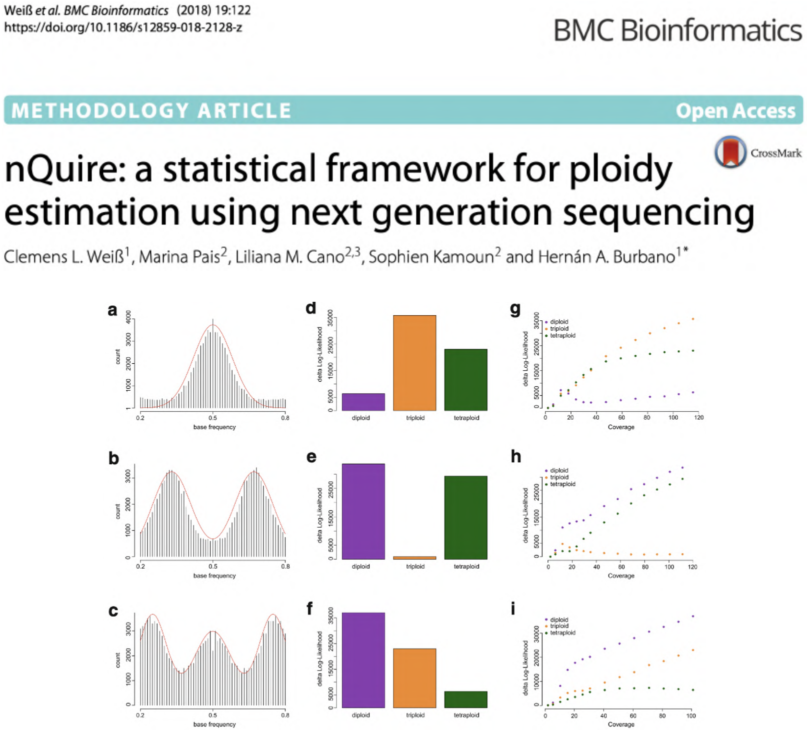


Bioinformatics, 33(16), 2017, 2575–2576
doi: 10.1093/bioinformatics/btx204
Advance Access Publication Date: 5 April 2017
Applications Note

Genome analysis

ploidyNGS: visually exploring ploidy with Next Generation Sequencing data

Renato Augusto Corrêa dos Santos[1], Gustavo Henrique Goldman[2] and Diego Mauricio Riaño-Pachón[1,3,*]



Weiß *et al. BMC Bioinformatics* (2018) 19:122
https://doi.org/10.1186/s12859-018-2128-z

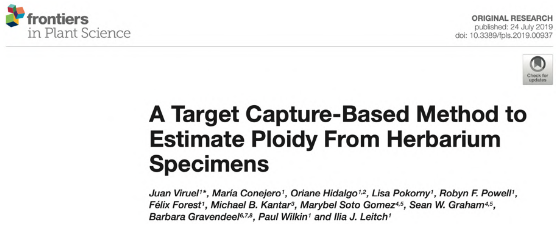
BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

nQuire: a statistical framework for ploidy estimation using next generation sequencing

Clemens L. Weiß[1], Marina Pais[2], Liliana M. Cano[2,3], Sophien Kamoun[2] and Hernán A. Burbano[1*]

Ploidy Level from TCS Data

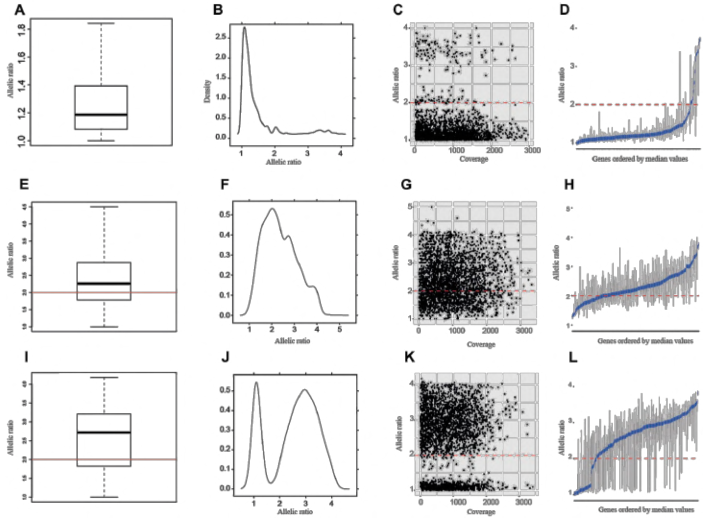


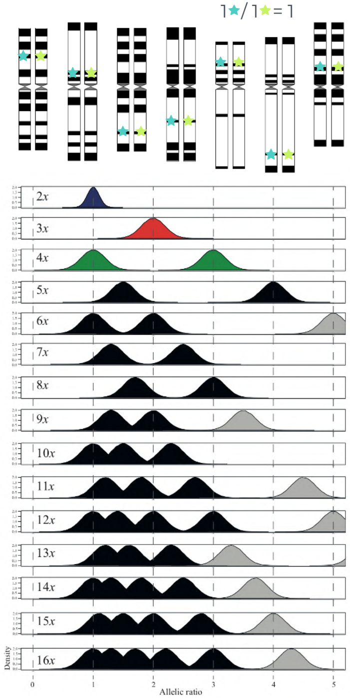
frontiers in Plant Science

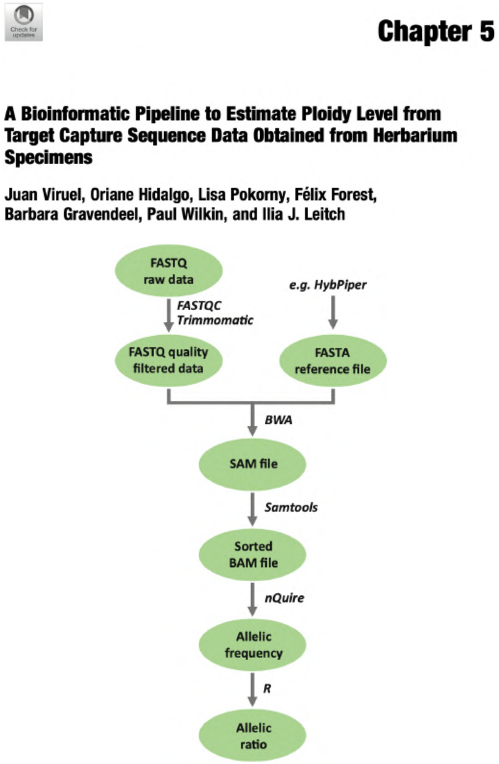
ORIGINAL RESEARCH
published: 24 July 2019
doi: 10.3389/fpls.2019.00937

A Target Capture-Based Method to Estimate Ploidy From Herbarium Specimens

Juan Viruel[1]*, María Conejero[1], Oriane Hidalgo[1,2], Lisa Pokorny[1], Robyn F. Powell[1], Félix Forest[1], Michael B. Kantar[3], Marybel Soto Gomez[4,5], Sean W. Graham[4,5], Barbara Gravendeel[6,7,8], Paul Wilkin[1] and Ilia J. Leitch[1]







Chapter 5

A Bioinformatic Pipeline to Estimate Ploidy Level from Target Capture Sequence Data Obtained from Herbarium Specimens

Juan Viruel, Oriane Hidalgo, Lisa Pokorny, Félix Forest, Barbara Gravendeel, Paul Wilkin, and Ilia J. Leitch

Dioscoreaceae303

Classification: order Dioscoreales (Monocot), ~300 species (~4 genera)

Importance: yams (edible tuber), steroidal saponins (contraceptives), etc.

Available resources: 1KP
3 *Dioscorea* & 1 Tacca transcriptomes
1 *Oryza sativa* & 1 *Xerophyta viscosa*
MarkerMiner Locus development made easy & Plaza DB
Daicel Arbor Biosciences myBaits®

Resulting kit:
260 ~single copy nuclear orthologs
43 genes of agronomic interest
~442 Kbp matrix
~17K 120-mer 3x tiled probes

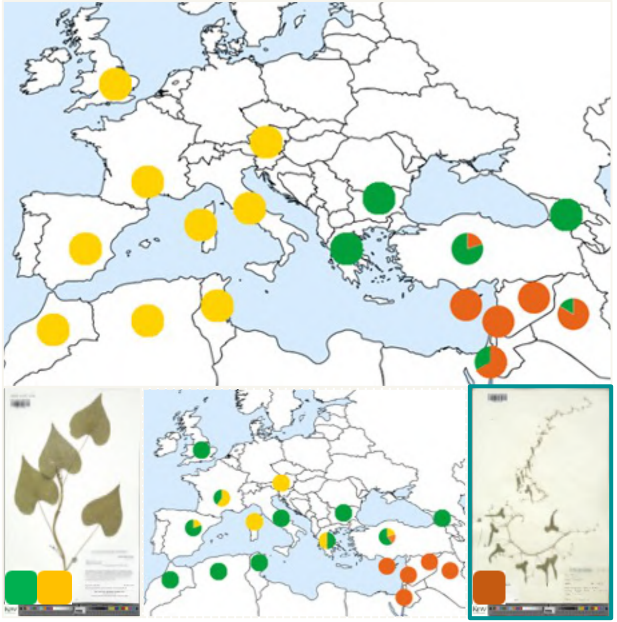
Applications in Plant Sciences

APPLICATION ARTICLE

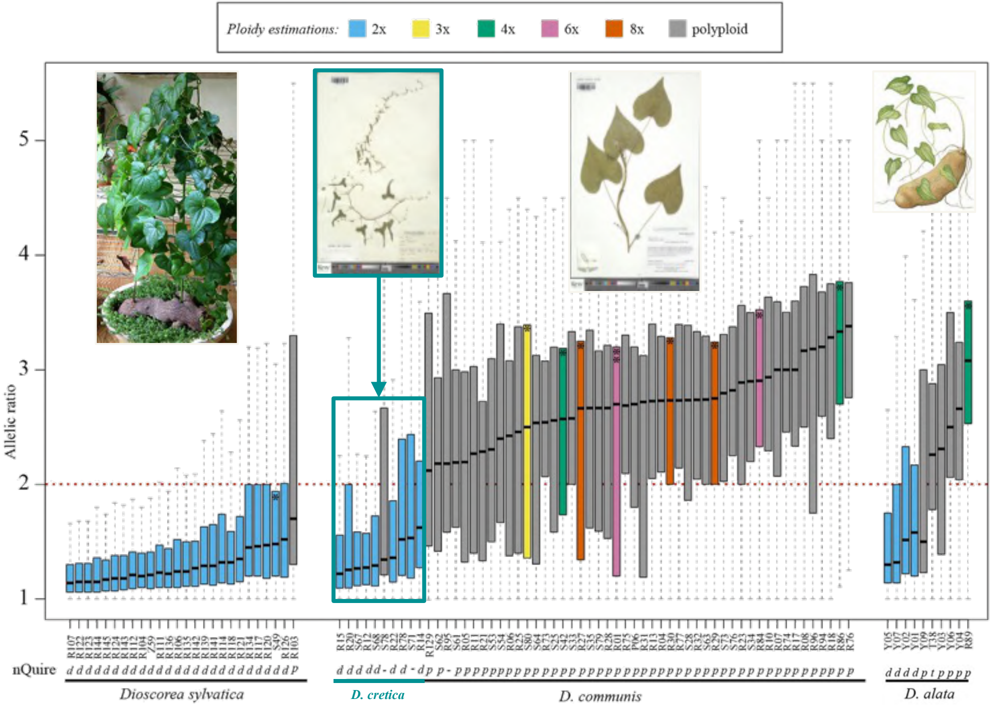
A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* yams (Dioscoreaceae)

Marybel Soto Gomez[1,2,8], Lisa Pokorny[3], Michael B. Kantar[4], Félix Forest[3], Ilia J. Leitch[3], Barbara Gravendeel[5,6,7], Paul Wilkin[3], Sean W. Graham[1,2], and Juan Viruel[3]

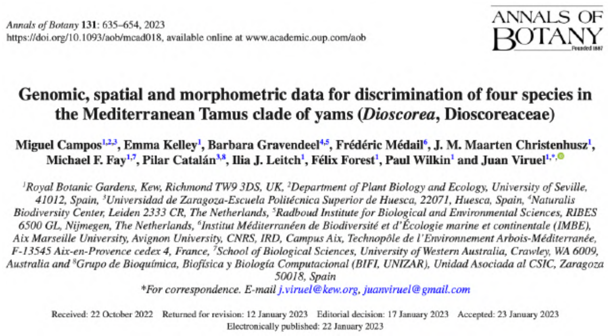
Ploidy Level in Yams



D. cretica



Ploidy Level in Yams



Annals of Botany **131**: 635–654, 2023
https://doi.org/10.1093/aob/mcad018, available online at www.academic.oup.com/aob

ANNALS OF BOTANY
Founded 1887

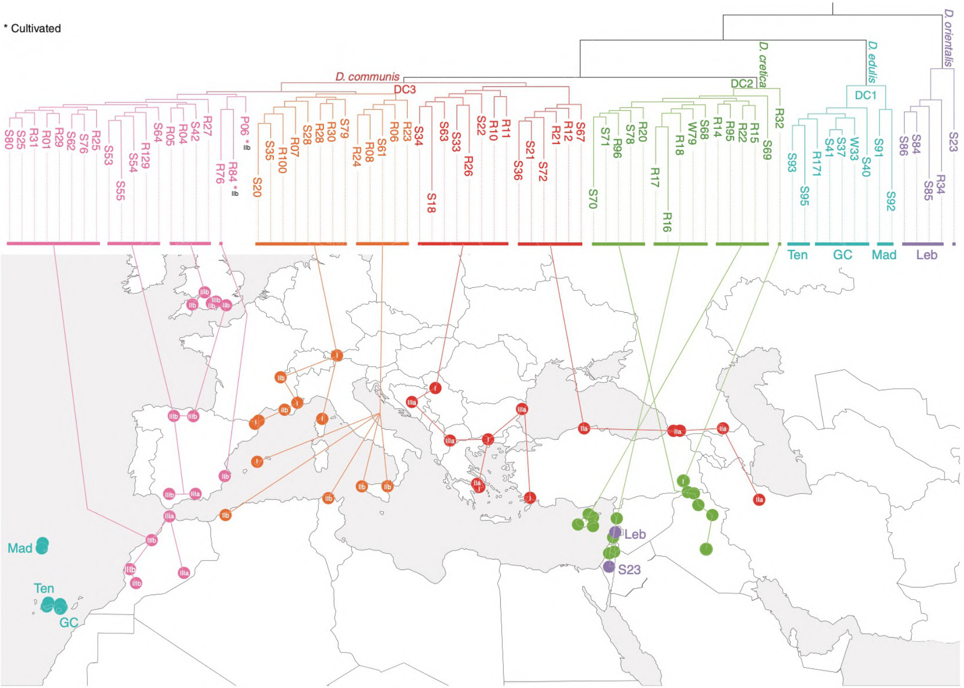
Genomic, spatial and morphometric data for discrimination of four species in the Mediterranean Tamus clade of yams (*Dioscorea*, Dioscoreaceae)

Miguel Campos[1,2,3], Emma Kelley[1], Barbara Gravendeel[4,5], Frédéric Médail[6], J. M. Maarten Christenhusz[1], Michael F. Fay[1,7], Pilar Catalán[3,8], Ilia J. Leitch[1], Félix Forest[1], Paul Wilkin[1] and Juan Viruel[1,*]

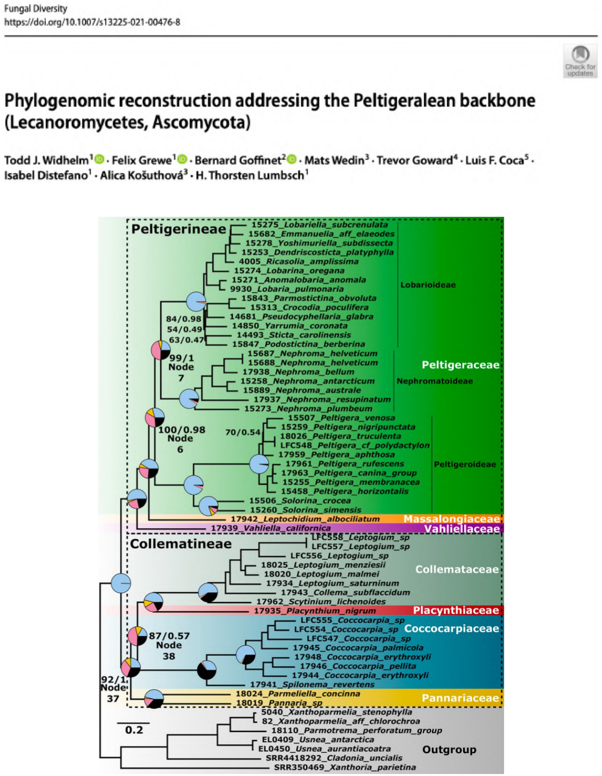
[1]Royal Botanic Gardens, Kew, Richmond TW9 3DS, UK, [2]Department of Plant Biology and Ecology, University of Seville, 41012, Spain, [3]Universidad de Zaragoza-Escuela Politécnica Superior de Huesca, 22071, Huesca, Spain, [4]Naturalis Biodiversity Center, Leiden 2333 CR, The Netherlands, [5]Radboud Institute for Biological and Environmental Sciences, RIBES 6500 GL, Nijmegen, The Netherlands, [6]Institut Méditerranéen de Biodiversité et d'Écologie marine et continentale (IMBE), Aix Marseille University, Avignon University, CNRS, IRD, Campus Aix, Technopôle de l'Environnement Arbois-Méditerranée, F-13545 Aix-en-Provence cedex 4, France, [7]School of Biological Sciences, University of Western Australia, Crawley, WA 6009, Australia and [8]Grupo de Bioquímica, Biofísica y Biología Computacional (BIFI, UNIZAR), Unidad Asociada al CSIC, Zaragoza 50018, Spain

**For correspondence. E-mail j.viruel@kew.org, juanviruel@gmail.com*

Received: 22 October 2022 Returned for revision: 12 January 2023 Editorial decision: 17 January 2023 Accepted: 23 January 2023
Electronically published: 22 January 2023



Fungariomics

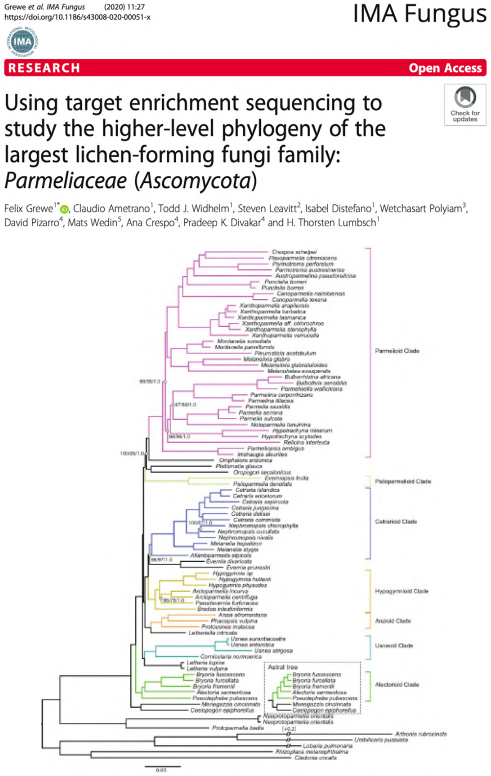


Fungal Diversity
https://doi.org/10.1007/s13225-021-00476-8

Check for updates

Phylogenomic reconstruction addressing the Peltigeralean backbone (Lecanoromycetes, Ascomycota)

Todd J. Widhelm[1] · Felix Grewe[1] · Bernard Goffinet[2] · Mats Wedin[3] · Trevor Goward[4] · Luis F. Coca[5] · Isabel Distefano[1] · Alica Košuthová[3] · H. Thorsten Lumbsch[1]



Grewe et al. IMA Fungus (2020) 11:27
https://doi.org/10.1186/s43008-020-00051-x

IMA Fungus

RESEARCH

Open Access

Check for updates

Using target enrichment sequencing to study the higher-level phylogeny of the largest lichen-forming fungi family: *Parmeliaceae* (*Ascomycota*)

Felix Grewe[1*], Claudio Ametrano[1], Todd J. Widhelm[1], Steven Leavitt[2], Isabel Distefano[1], Wetchasart Polyiam[3], David Pizarro[4], Mats Wedin[5], Ana Crespo[4], Pradeep K. Divakar[4] and H. Thorsten Lumbsch[1]

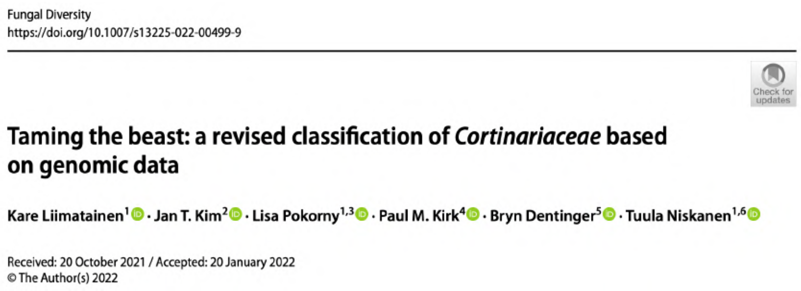
Cortinariaceae193

Classification: order Agaricales (Basidiomycota), ~3K species (1 genus)

Importance: ectomycorrhizal fungi

Available resources:
4 *Cortinarius* assemblies (ABySS)
208 Agaricales single-copy orthologs genes (SCOGs) & exonerate
Daicel Arbor Biosciences myBaits

Resulting kit:
188 nuclear SCOGs
RPB1, RPB2, MCM7, GPD, & TEF1
~67.5 Kbp matrix
~20K 120-mer 2x tiled probes



Fungal Diversity
https://doi.org/10.1007/s13225-022-00499-9

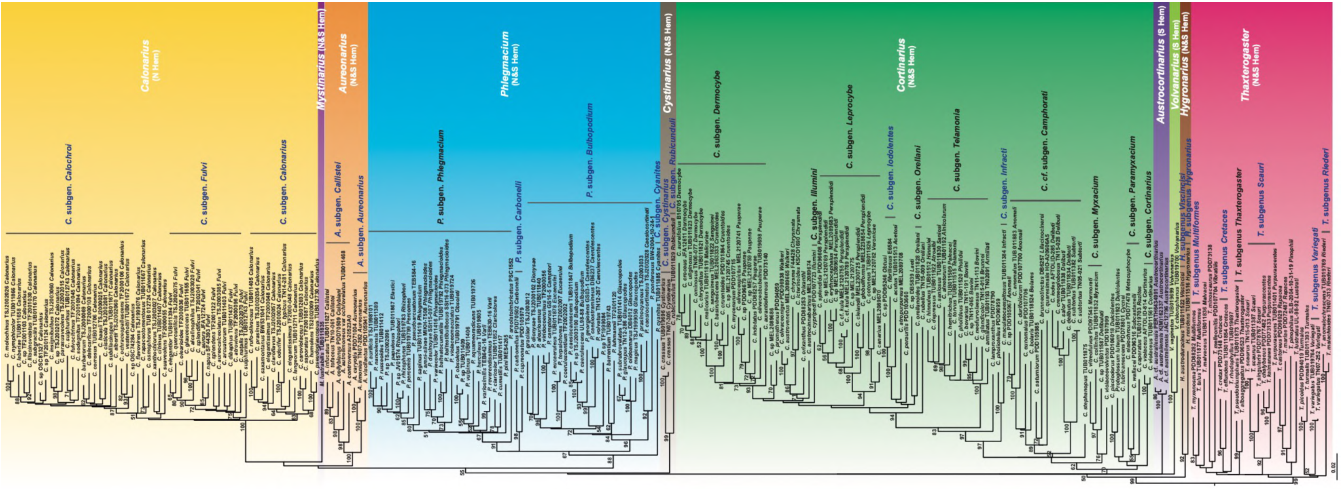
Taming the beast: a revised classification of *Cortinariaceae* based on genomic data

Kare Liimatainen[1] · Jan T. Kim[2] · Lisa Pokorny[1,3] · Paul M. Kirk[4] · Bryn Dentinger[5] · Tuula Niskanen[1,6]

Received: 20 October 2021 / Accepted: 20 January 2022
© The Author(s) 2022

Table 3 Species diversity, distribution, and summary of the selected morphological characteristics of the 10 genera of *Cortinariaceae*

Genus	Generic synonyms	Estimated no. of species	Subgen./section no	Distribution	Cystidia	Stipitocarpic (S)/Pileocarpic (P)	Pileipellis simplex/duplex	Sequestrate sp	Appearance of the agaricoid basidiomata				Rozitoid/cuphocyboid
									Myxacioid	Phlegmac	Telamonioid	Cortinarioid	
Cortinarius	11	> 2000	11/130*	N + S	(X)	S/(P)	Duplex (+ simpl.)	X	X	(X)	X	X	X
Phlegmacium	3	> 200	4/23	N (+ S)		P/S	Duplex (+ simpl.)	X		X	(X)		
Thaxterogaster	3	> 200	6/22	(N +) S		S/P	Duplex	X	X	X			
Calonarius	–	~ 200	3/14	N		P	Simplex	X		X			
Aureonarius	–	~ 25	2/3	(N +) S		S	Duplex					X	
Cystinarius	–	~ 10	2/2	N + S	X	S	Duplex			X		(X)	
Volvanarius	–	~ 10	1/1	S	(X)	P	Duplex	X		X			
Hygronarius	–	~ 10	2/2	(N +) S		S	Duplex				X		
Austrocortinarius	–	< 5	1/1	S		S	Duplex			X			
Mystinarius	–	< 5	1/1	N + S		S	Duplex		X	X			



Cantharellales526

Classification: Agaricomycetes (Basidiomycota), ~630 spp. (~40 gen.)

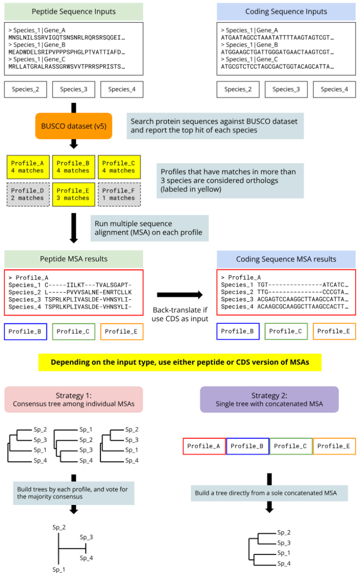
Importance: ectomycorrhizal & edible

Available resources:

10 DOE-JGI genomes (1KFG)
PHYling & UFCG SCOGs
get_homologues, HybPiper, AMAS
Daicel Arbor Biosciences myBaits®

Resulting kit:

523 nuclear SCOGs
RPB1&2, EF, & MCM7
~7.5 Mbp matrix
~55K 120-mer 1x tiled probes



Search protein sequences against BUSCO dataset and report the top hit of each species

Profiles that have matches in more than 3 species are considered orthologs (labeled in yellow)

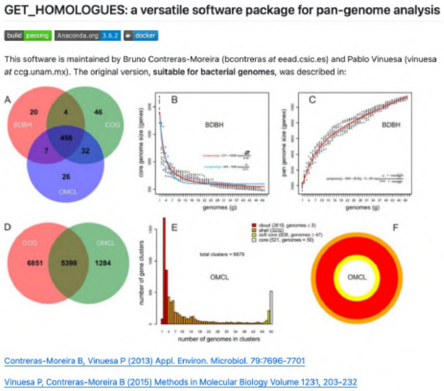
Run multiple sequence alignment (MSA) on each profile

Back-translate if use CDS as input

Depending on the input type, use either peptide or CDS version of MSAs

Build trees by each profile, and vote for the majority consensus

Build a tree directly from a sole concatenated MSA

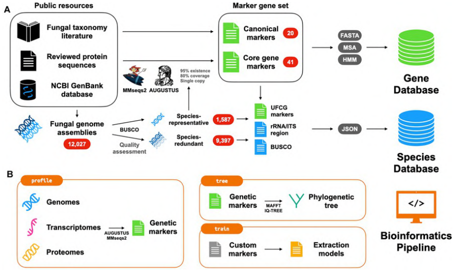


GET_HOMOLOGUES: a versatile software package for pan-genome analysis

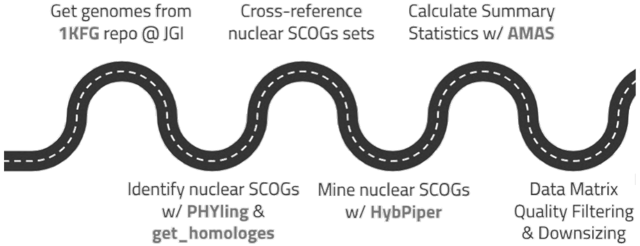
This software is maintained by Bruno Contreras-Moreira (bcontreras *at* eead.csic.es) and Pablo Vinuesa (vinuesa *at* ccg.unam.mx). The original version, **suitable for bacterial genomes**, was described in:

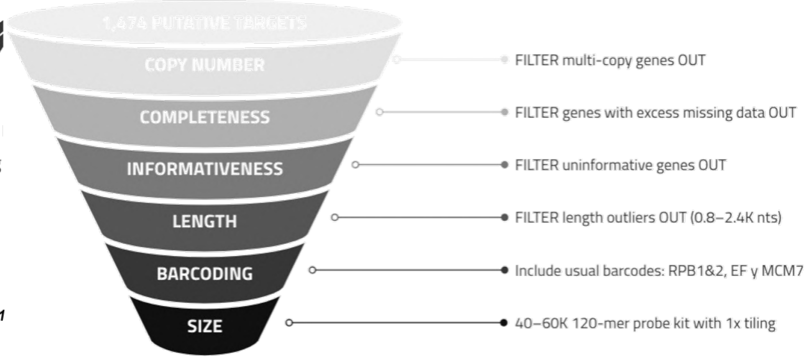
Contreras-Moreira B, Vinuesa P (2013) Appl. Environ. Microbiol. 79:7696-7701

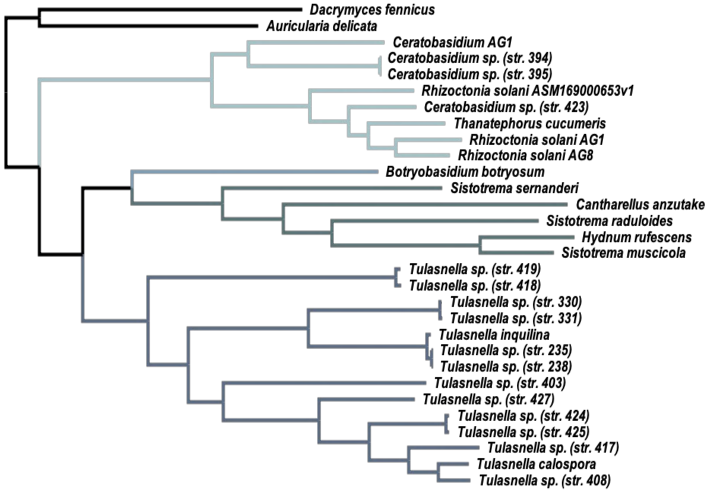
Vinuesa P, Contreras-Moreira B (2015) Methods in Molecular Biology Volume 1231, 203-232



Cantharellales526







7,459,405 nucleotides

Target data matrix size

54,979 baits

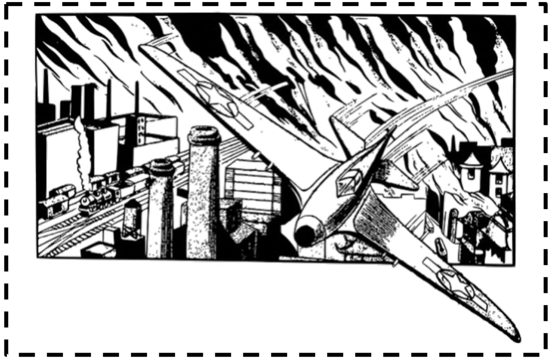
Probe set size

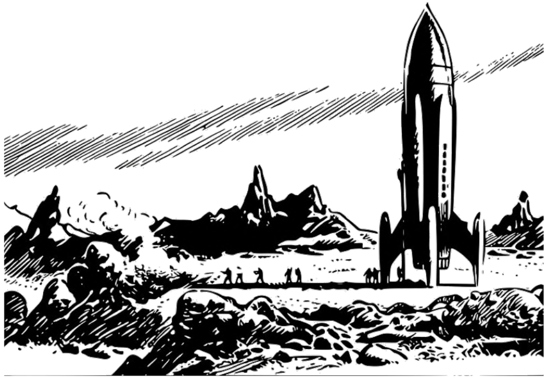
526 genes

Single-copy orthologous genes (SCOGs)

Table of contents

Table of contents





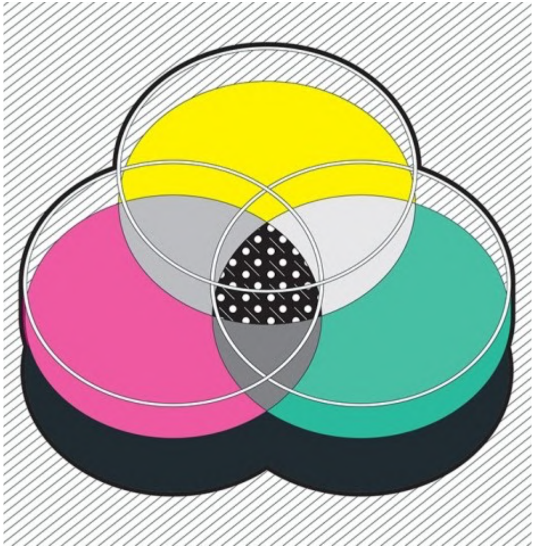
Future Perspectives

From Strains to Biomes

05

Biodiversity Prospection

Exploring the Known Unknowns and the Unknown Unknowns (Microbial Dark Matter)



Mesoevolutionary Approaches

Bridging Micro- and Macro-Evolutionary Scales

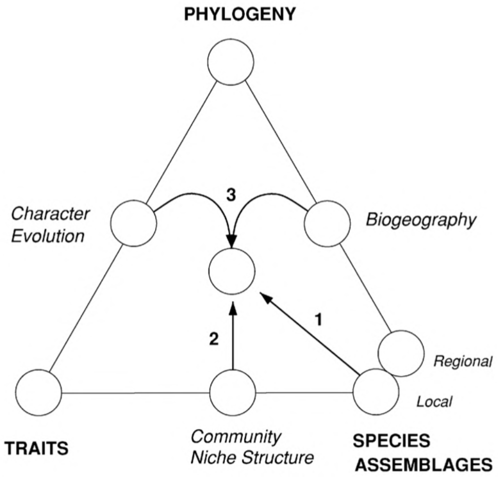


Fig. 1. Webb et al. 2002. *Annu. Rev. Ecol. Syst. 33:475–505.*

Community Phylogenomics

At the Intersection of Biogeography and Ecology

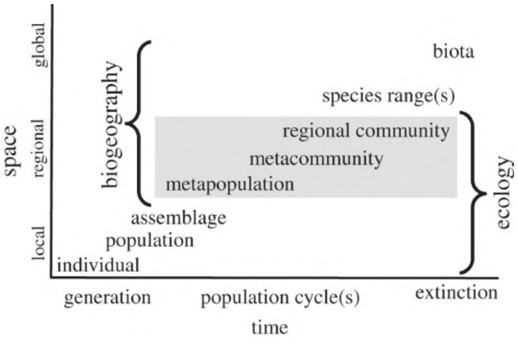


Fig. 1. Jenkins & Ricklefs. 2011. *Phil. Trans. R. Soc. B*: 366, 2331–35

Thanks!

Questions?

pokorny@rjb.csic.es
@Calyptrochaeta

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**