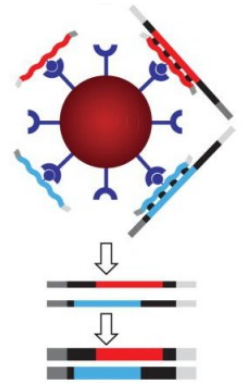
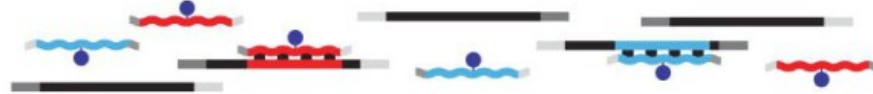


Hyb-Seq course



January 20th-23rd 2020

Vojtěch Zeisek, Roswitha E. Schmickl and Tomáš Fér

Preliminary program of part of day 1

- **Basic introduction to Illumina HTS technology**
 - Library preparation
 - Running on the sequencer
 - Illumina BaseSpace - data access and download
- **Phylogenomic approaches**
 - Overview of available methods
- **Hybridization-based targeted enrichment**
 - What it is (intro to method, obtained datasets)
 - Enrichment probe development (input datasets, important steps)
 - Wet-lab procedure
 - Effect of material preservation on Hyb-Seq success
- **RADseq vs. Hyb-Seq**
 - Pros and cons
 - hyRAD - a compromise
- **Universal vs. group-specific probes**
 - Angiosperm probes
 - Comparisons of probe sets

Basic introduction to Illumina high-throughput sequencing technology

Next generation sequencing (NGS)

- first generation – Sanger sequencing
- second generation – parallel sequencing of many molecules (PCR amplified)
- further generations – single molecule sequencing

Next generation sequencing (NGS)

- **massive** (many sequences, up to hundreds millions per run)
- **parallel** (simultaneous sequencing)
- **several commercial platforms**
 - **pyrosequencing (Roche/454)** – GS FLX, GS Junior
 - long reads (500-700 bp)
 - high error rate in poly-stretches
 - low number of sequences
 - **Illumina (Solexa)** – GA II, MiSeq, NextSeq, HiSeq
 - short reads (75-300 bp)
 - high-throughput
 - **ABI SOLiD, PacBio, Ion Torrent, Oxford Nanopore...**

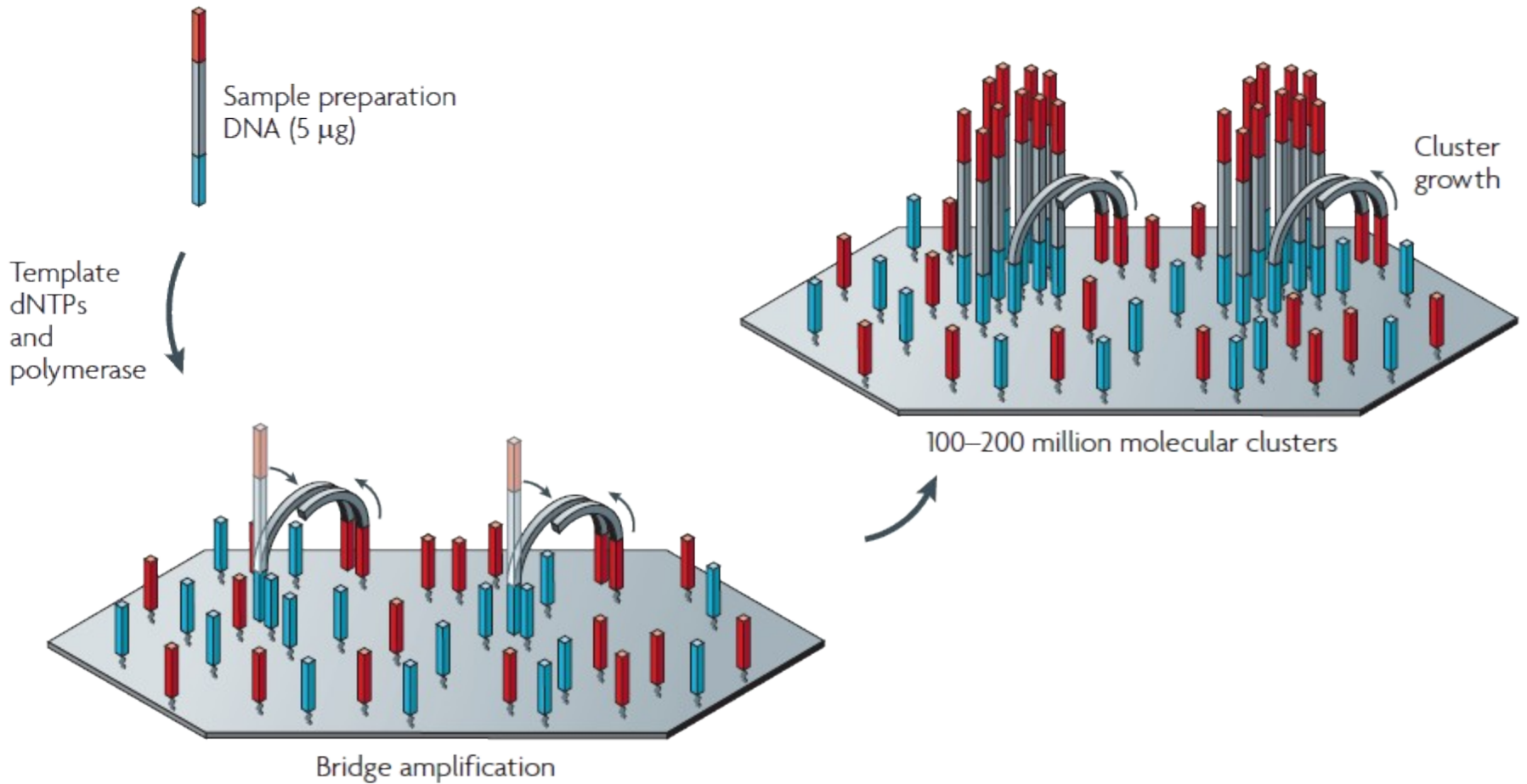
General NGS approach

- library preparation
 - random shearing of genomic DNA to the fragments
 - sequencing adaptor ligation
- spatial separation of individual fragments
- two „basic“ sequencing options
 - sequencing of clonally amplified fragments
 - emulsion PCR (emPCR)
 - solid-phase amplification
 - single molecule sequencing
- immobilization to the surface
- sequencing and data acquisition
 - pyrosequencing (Roche/454)
 - cyclic reversible termination (CRT) (Illumina/Solexa)
 - sequencing by ligation (SOLiD)
- data analysis (analysis of image data, quality control, ...)

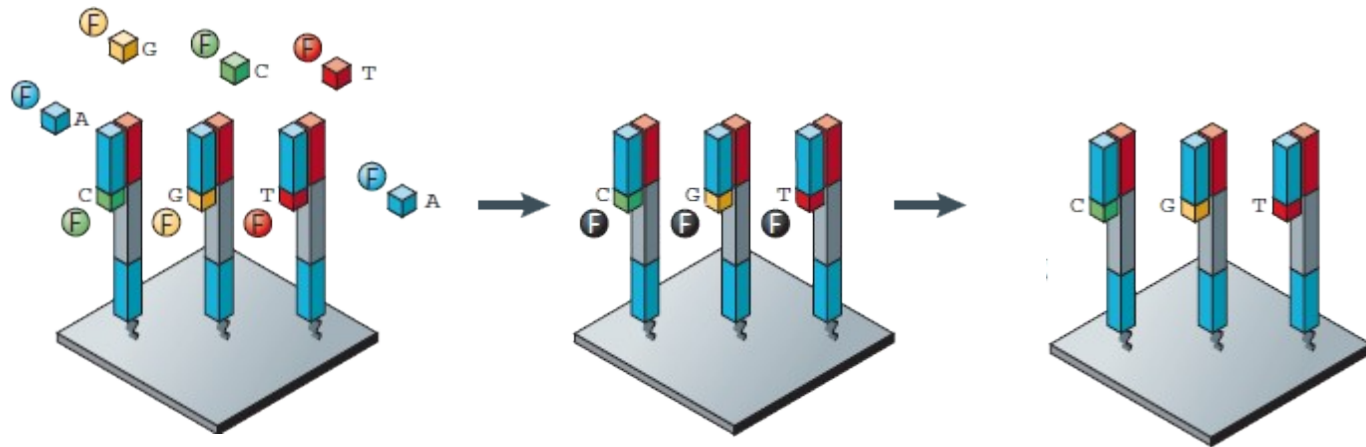
Illumina technology

- library preparation
 - TrueSeq (sonication for DNA fragmentation)
 - Nextera (enzymatic fragmentation – transposase)
- solid-phase amplification (bridge PCR)
- sequencing
 - cyclic reversible termination (CRT)
 - single-end or pair-end sequencing
 - 2x25, 2x75, 2x150, 2x250, 2x300...

Solid-phase amplification (= bridge PCR)



Cyclic reversible termination



incorporate all four nucleotides, each label with a different dye

wash, four-colour imaging

cleave dye and terminating groups, wash



Top: CATCGT
Bottom: CCCCC

Illumina BaseSpace

<https://basespace.illumina.com>

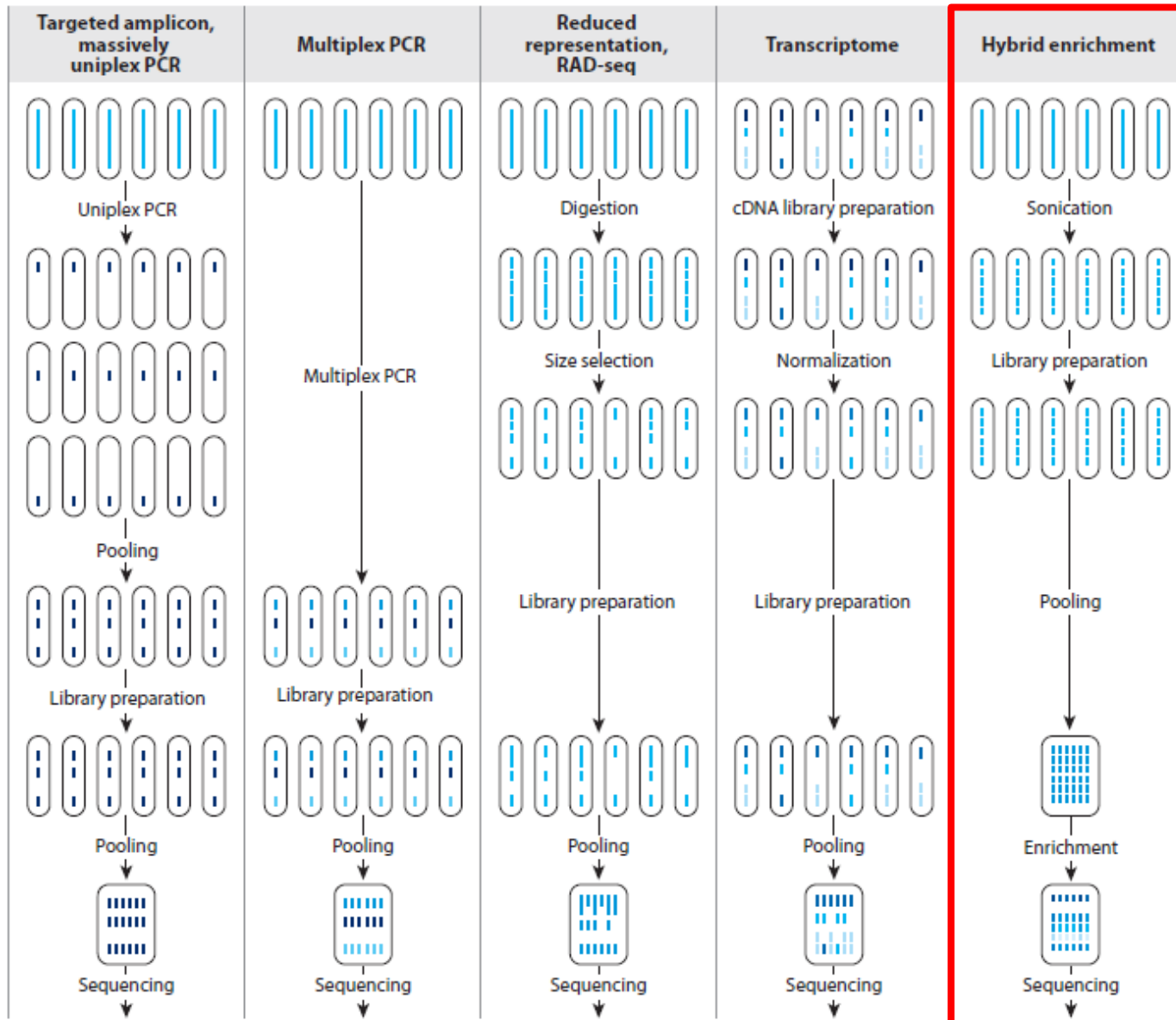
- data from sequencer are send to the cluster
 - general quality filtering of clusters
 - demultiplexing
- information about your runs and projects
 - sample sheet
 - read quality overview
 - number of reads total and per each sample
- samples download
 - FASTQ files (2 per sample if PE)
- data analysis
 - many free as well as paid applications

Phylogenomic approaches

Phylogenomics

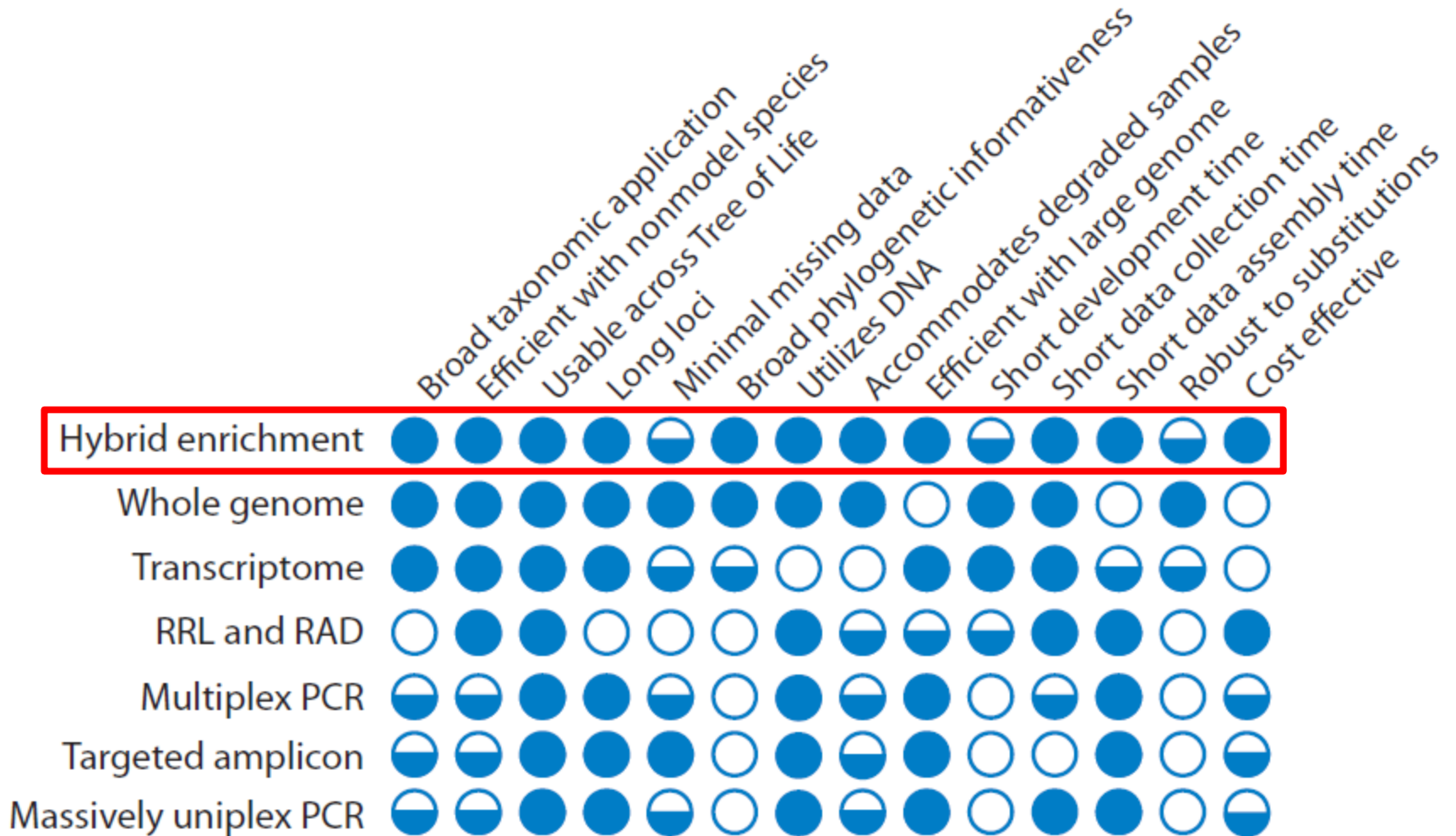
- Using **whole-genome sequences** or a **large portion of the genome** to build a phylogeny, using high-throughput sequencing
 - whole chloroplast sequences
 - hundreds or thousands of genes
- **Gene tree** – individual evolutionary history of a gene
- **Species tree** – ‘true’ species evolution
- **Gene tree/species tree discordance?!**

Common phylogenomic datasets



Lemmon & Lemmon (2013)
Annu. Rev. Ecol. Evol. Syst.

Comparison of phylogenomic approaches



High-throughput sequencing in phylogenetics – potential or not?

Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from *Pinus L.* (Pinaceae)

Matthew Parks^{1*}, Richard Cronn² and Aaron Liston¹

* Corresponding author: Matthew Parks

parksma@science.oregonstate.edu

▼ Author Affiliations

¹ Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331-2902, USA

² Pacific Northwest Research Station, USDA Forest Service, Corvallis, OR 97331, USA

For all author emails, please [log on](#).

BMC Evolutionary Biology 2012, **12**:100 doi:10.1186/1471-2148-12-100

Potential to greatly increase the amount of phylogenetically informative signal in molecular datasets

Opens the era of real incongruence

Trends in
Genetics



Volume 22, Issue 4, April 2006, Pages 225–231

Phylogenomics: the beginning of incongruence?

Olivier Jeffroy, Henner Brinkmann, Frédéric Delsuc, Hervé Philippe

Opinion

CellPress

Post-molecular systematics and the future of phylogenetics

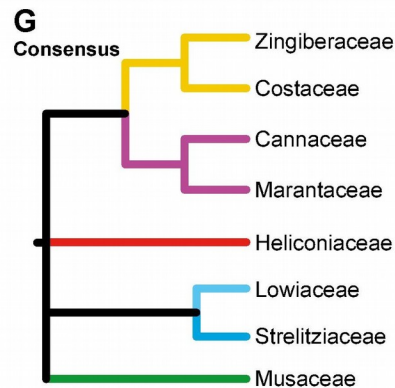
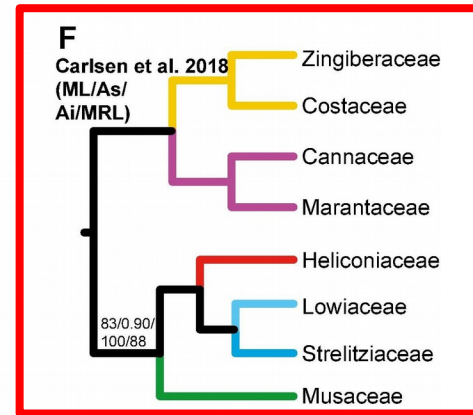
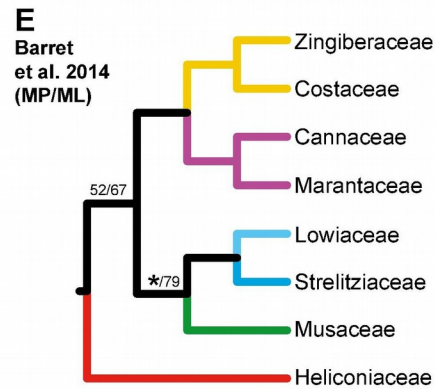
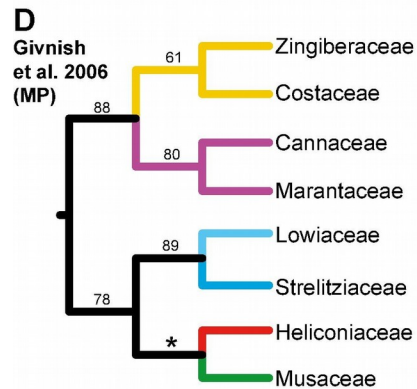
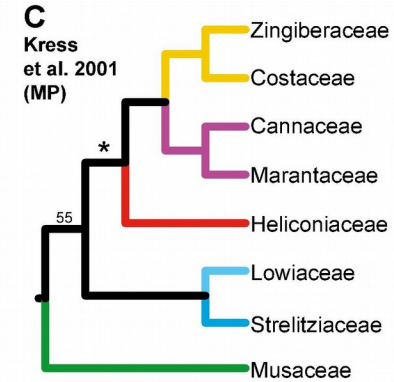
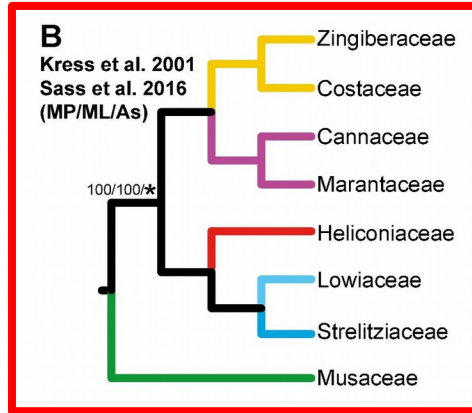
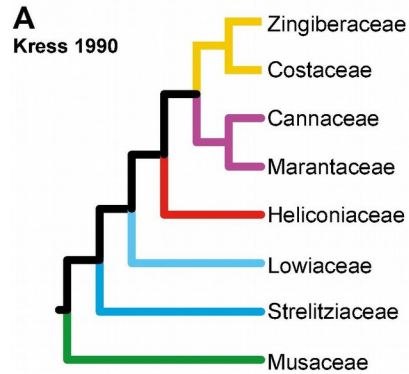
R. Alexander Pyron

Department of Biological Sciences, The George Washington University, 2023 G St NW, Washington, DC 20052, USA

384 Trends in Ecology & Evolution, July 2015, Vol. 30, No. 7

Even massive amounts of sequence data do not always result in strongly resolved phylogenies

Even **high-throughput sequencing data** resolve phylogenies controversially



Carlsen et al. (2018) Mol. Phylogenet. Evol.

Hybridization-based target enrichment

Plant phylogenetics: the advent of HTS from a historical perspective

ASSEMBLING THE TREE OF THE MONOCOTYLEDONS: PLASTOME SEQUENCE PHYLOGENY AND EVOLUTION OF POALES¹

Thomas J. Givnish,² Mercedes Ames,² Joel R. McNeal,³ Michael R. McKain,³ P. Roxanne Steele,⁴ Claude W. dePamphilis,⁵ Sean W. Graham,⁶ J. Chris Pires,⁴ Dennis W. Stevenson,⁷ Wendy B. Zomlefer,³ Barbara G. Briggs,⁸ Melvin R. Duval,⁹ Michael J. Moore,¹⁰ J. Michael Heaney,¹¹ Douglas E. Soltis,¹¹ Pamela S. Soltis,¹² Kevin Thiele,¹³ and James H. Leebens-Mack³

ANN. MISSOURI BOT. GARD. 97: 584–616. PUBLISHED ON 27 DECEMBER 2010.

Plastid genomes



High-copy fractions of genomes (genome skimming)

[Am J Bot.](#) 2012 Feb;99(2):349-64. doi: 10.3732/ajb.1100335. Epub 2011 Dec 14.

Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics.

[Straub SC](#)¹, [Parks M](#), [Weitemier K](#), [Fishbein M](#), [Cronn RC](#), [Liston A](#).

[Appl Plant Sci.](#) 2014 Sep; 2(9): apps.1400042.

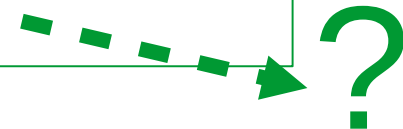
Published online 2014 Aug 29. doi: [10.3732/apps.1400042](https://doi.org/10.3732/apps.1400042)

PMCID: PMC4162667

Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics¹

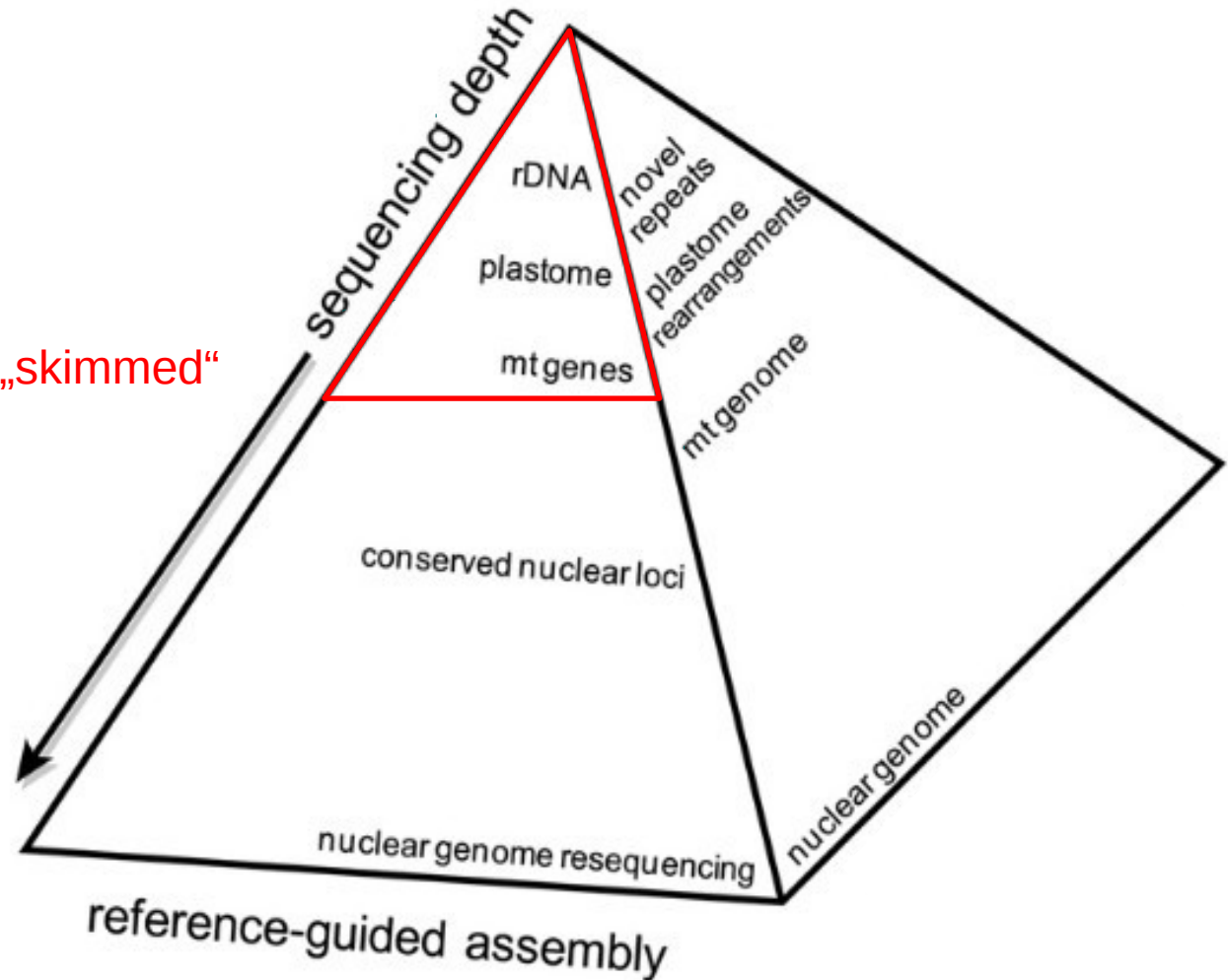
[Kevin Weitemier](#),^{2,7} [Shannon C. K. Straub](#),^{2,7} [Richard C. Cronn](#),³ [Mark Fishbein](#),⁴ [Roswitha Schmickl](#),⁵ [Angela McDonnell](#),⁴ and [Aaron Liston](#)^{2,6}

Combination of genome skimming with target enrichment



What are genome skim data?

Mainly this portion is „skimmed“



Plant phylogenetics: the advent of HTS from a historical perspective

ASSEMBLING THE TREE OF THE MONOCOTYLEDONS: PLASTOME SEQUENCE PHYLOGENY AND EVOLUTION OF POALES¹

Thomas J. Givnish,² Mercedes Ames,² Joel R. McNeal,³ Michael R. McKain,³ P. Roxanne Steele,⁴ Claude W. dePamphilis,⁵ Sean W. Graham,⁶ J. Chris Pires,⁴ Dennis W. Stevenson,⁷ Wendy B. Zomlefer,³ Barbara G. Briggs,⁸ Melvin R. Duval,⁹ Michael J. Moore,¹⁰ J. Michael Heaney,¹¹ Douglas E. Soltis,¹¹ Pamela S. Soltis,¹² Kevin Thiele,¹³ and James H. Leebens-Mack³

ANN. MISSOURI BOT. GARD. 97: 584–616. PUBLISHED ON 27 DECEMBER 2010.

Plastid genomes

High-copy fractions of genomes (genome skimming)

[Am J Bot.](#) 2012 Feb;99(2):349-64. doi: 10.3732/ajb.1100335. Epub 2011 Dec 14.

Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics.

[Straub SC](#)¹, [Parks M](#), [Weitemier K](#), [Fishbein M](#), [Cronn RC](#), [Liston A](#).

[Appl Plant Sci.](#) 2014 Sep; 2(9): apps.1400042.

Published online 2014 Aug 29. doi: [10.3732/apps.1400042](https://doi.org/10.3732/apps.1400042)

PMCID: PMC4162667

Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics¹

[Kevin Weitemier](#),^{2,7} [Shannon C. K. Straub](#),^{2,7} [Richard C. Cronn](#),³ [Mark Fishbein](#),⁴ [Roswitha Schmickl](#),⁵ [Angela McDonnell](#),⁴ and [Aaron Liston](#)^{2,6}

Combination of genome skimming with target enrichment



Hyb-Seq (combination of target enrichment and genome skimming): general workflow

Custom probe design

Bait synthesis
(usually outsourced to company)

Genomic library preparation
(5 ng – 1 μg DNA,
partly degraded DNA also works)

In this step target enrichment is combined with genome skimming.

Hybridization of baits to genomic library
(100-500 ng DNA of genomic library,
tested with a minimum of 9 ng per sample in
a 24plex reaction)

Sequencing of enriched targets (e.g., nuclear exons)
and off-target sequences (mainly plastome and
nrDNA cistron)

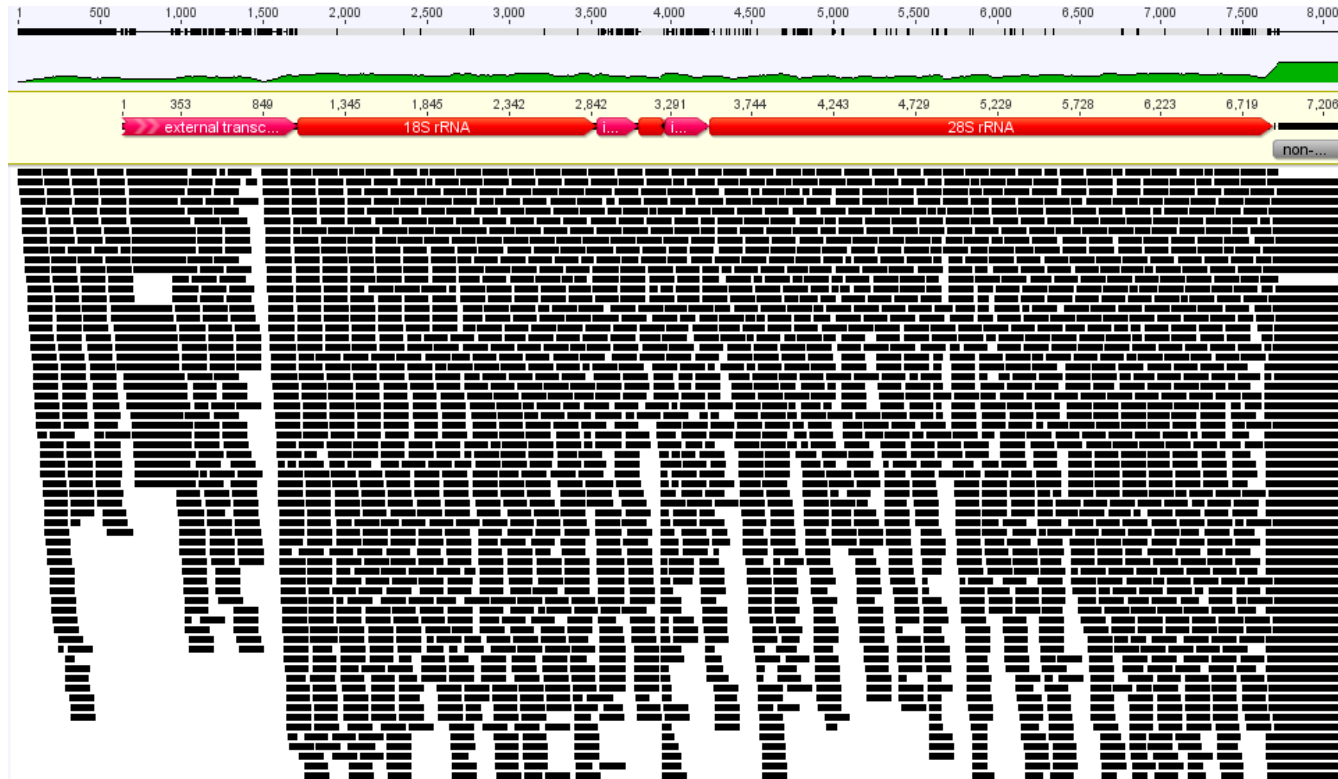
Data analysis

Comparison Hyb-Seq with genome skimming (for the same accession of an *Oxalis obtusa*)

	# on-target, quality-filtered reads after duplicate removal	# plastid reads after duplicate removal	Mean sequencing depth of LCN genes	Mean sequencing depth of plastome
Hyb-Seq	408,559 (57%)	183,972 (26%)	14	166
Genome skimming	659,726 (8%)	1,114,157 (14%)	11	825

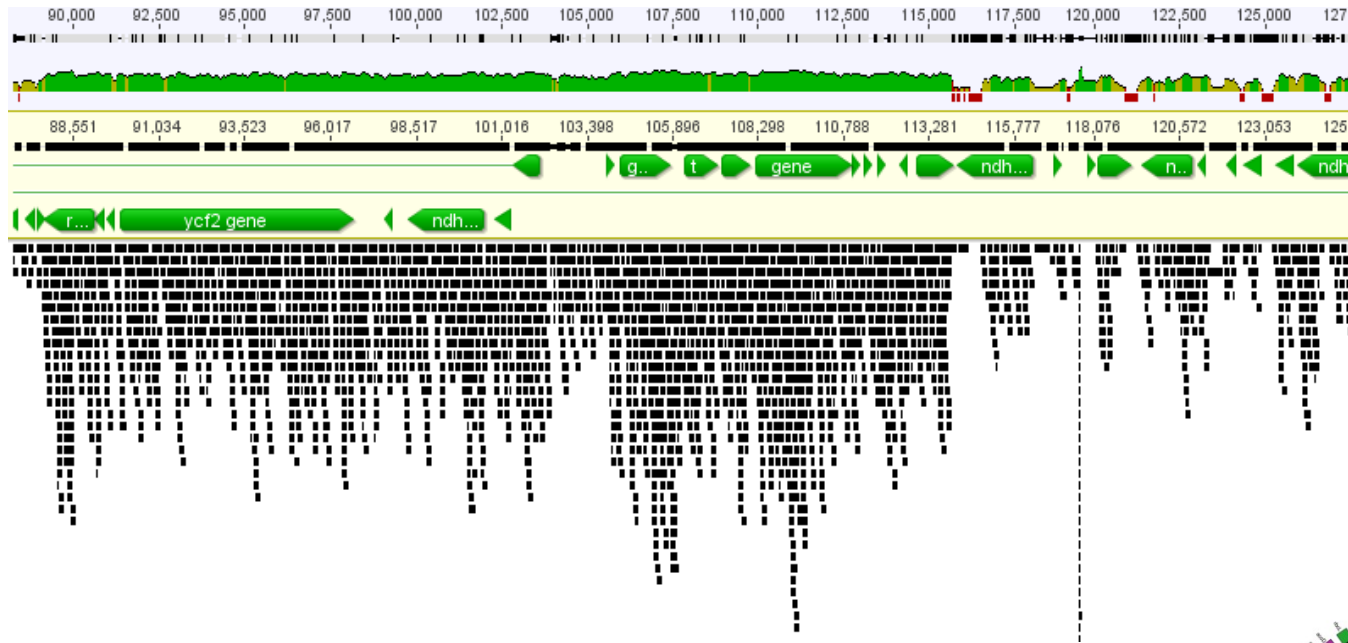


Read depth - nrDNA

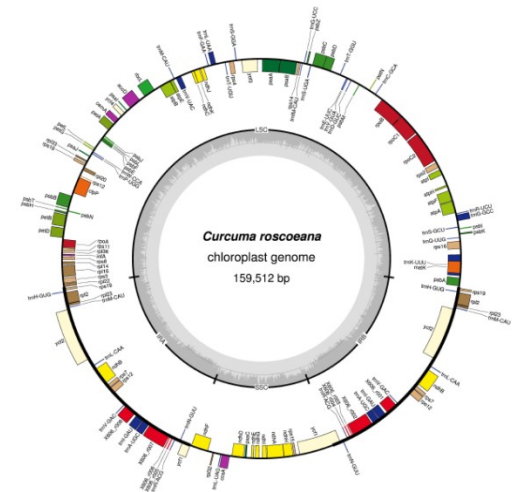


	Mean	
	Number of reads	Coverage
Zingiberaceae	12 462	166
<i>Globba</i>	6 850	142

Read depth - chloroplast



	Mean	
	Number of reads	Coverage
ALL data	26 049	23.41
Zingiberaceae	23 604	21.34
<i>Globba</i>	29 877	26.55
other Zingiberales	42 926	38.20

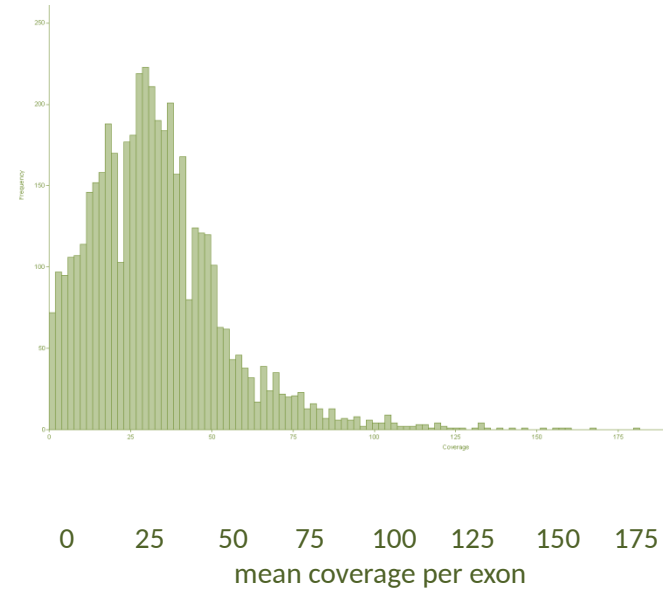


Hyb-Seq studies often do not make use of plastid data

Table 1. Recent phylogenetic or population genetic studies that used hybrid-enrichment methods, including plastid-oriented studies. Studies targeting only nuclear regions but that utilized plastid sequences found in bycatch are denoted by an asterisk (*). # Ref. Taxa refers to the number of reference taxa included in a probe design. # Sp. Copy Assessment indicates how many species the study included in their assessment of copy number for each locus included in the probe kit developed.

Study	Basal Group	Major Group	Family	Genus	# Ref. Taxa	# Nuclear Loci Targeted	Probes Target Plastids	# Sp. Copy Assessment
Parks et al. (2012)	Gymnosperms	-	Pinaceae	<i>Pinus</i>	2	0	yes	2
Stall et al. (2013)	Angiosperms	eudicots	-	-	22	0	yes	0
Mandel et al. (2014)	Angiosperms	eudicots	Asteraceae	-	1	1061	no	4
Weitmier et al. (2014)	Angiosperms	eudicots	Asteraceae	<i>Asclepias</i>	1	3385	no*	6
de Sousa et al. (2014)	Angiosperms	eudicots	Leguminosae	<i>Medicago</i>	1	319	no	3
Grover et al. (2015)	Angiosperms	eudicots	Malvaceae	<i>Gossypium</i>	1	500	no	1
Stephens et al. (2015a)	Angiosperms	eudicots	Sarraceniacae	<i>Sarracenia</i>	2	646	no*	2
Stephens et al. (2015b)	Angiosperms	eudicots	Asteraceae	<i>Helianthus</i>	4	598	no*	2
Schmickl et al. (2016)	Angiosperms	eudicots	Oxalidaceae	<i>Oxalis</i>	1	1164	no*	3
Heyduk et al. (2016)	Angiosperms	monocots	Arecaceae	<i>Sabal</i>	15	837	yes	15
Syring et al. (2016)	Gymnosperms	-	Pinaceae	<i>Pinus albicaulis</i>	1	7849	no	1
Sass et al. (2016)	Angiosperms	monocots	Zingiberaceae	-	8	494	yes	8
This Study	Angiosperms	All groups	-	-	25	517	no*	43

Read depth - exons



min - max	0 - 1032
mean	33.8
50% within	19 - 43

Hyb-Seq (combination of target enrichment and genome skimming): general workflow

Custom **probe design**

Bait synthesis
(usually outsourced to company)

Genomic library preparation
(5 ng – 1 µg DNA,
partly degraded DNA also works)

In this step target
enrichment is
combined with
genome skimming.

Hybridization of baits to genomic library
(100-500 ng DNA of genomic library,
tested with a minimum of 9 ng per sample in
a 24plex reaction)

Sequencing of enriched targets (e.g., nuclear exons)
and off-target sequences (mainly plastome and
nrDNA cistron)

Data analysis

Hyb-Seq probe design

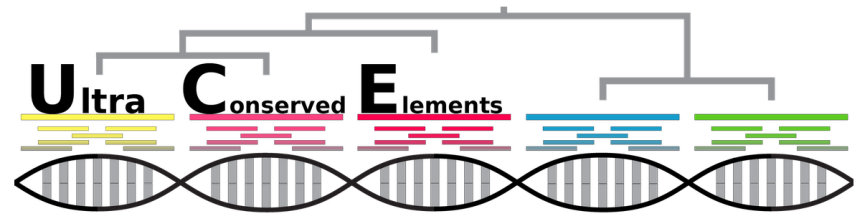
Probe design:

- Exons of low-copy nuclear genes
- Intronic regions

Bait synthesis:

- RNA baits

Commonly used in animal phylogenomics:



Alternatives (without bait synthesis):

- DNA or cDNA (in hyRAD)
- PCR products

<http://www.ultraconserved.org/>

<p>What are UCEs?</p> <p>As their name implies, ultraconserved elements (UCEs) are highly conserved regions of organismal genomes shared among evolutionary distant taxa - for instance, birds share many UCEs with humans. UCEs were first described in a wonderful manuscript by Gil Bejerano et al (2004) from David Haussler's group and subsequently identified in several classes of organisms outside the group of original taxa (Siepel et al. 2005) used to identify these genomic elements. The 27-way vertebrate genome alignment (Miller et al. 2007) identified additional regions of high conservation.</p>	<p>Why are UCEs useful?</p> <p>We have discovered (see Citations) that we can collect data from UCEs and the DNA adjacent to UCE locations (flanking DNA), and that these data are useful for reconstructing the evolutionary history and population-level relationships of many organisms. Because UCEs are conserved across disparate taxa, UCEs are also universal genetic markers in the sense that the locations (or loci) that we can target in humans are identical, in many cases, to the loci that we can target in ducks or snakes or lizards.</p>	<p>What do UCEs do?</p> <p>That's an extremely good question, and one to which we do not entirely know the answer (Dermitzakos et al. 2005). UCEs have been associated with gene regulation (Pernachio et al. 2006) and development (Sandelin et al 2004, Woolfe et al. 2004) and we generally assume that UCEs must be important by the very nature of their near-universal conservation across extremely divergent taxa. However, gene knockouts of UCE loci in mice resulted in viable, fertile offspring (Ahituv et al. 2007), suggesting that their role in the biology of the genome may be cryptic.</p>
<p>How do I identify UCEs?</p> <p>You can identify UCEs in organismal genome sequences by aligning several genomes to each other, scanning the resulting genome alignments for areas of very high (95-100%) sequence conservation, and filtering on user-defined criteria, such as length (e.g., Bejerano et al. 2004). If you want to use these regions as genetic markers, it is best to remove UCEs that appear to be duplicates of one another which we loosely define as being in more than one spot within each genome that you aligned. The resulting loci are the highly conserved that we target for use as molecular markers.</p>	<p>How do I collect UCE data?</p> <p>From the resulting set of UCEs shared among a taxonomic group, we design sequence capture (AKA solution hybrid selection sensu Girnie et al. 2009) probes that are similar in sequence to the UCE loci we are targeting. These probe sets differ in number and composition, depending on the types of questions we are asking and the taxa with which we are working. Once we design a probe set, we follow sequence capture protocols to enrich DNA libraries for the target UCEs, usually in multiplex. Following enrichment, we sequence the DNA enriched for UCEs using massively parallel sequencing.</p> <p>Get protocols +</p>	<p>How do I analyze UCE data?</p> <p>The most complex part of using UCEs to understand evolutionary relationships, population structure, and population relationships is analyzing the DNA sequence data. We have created several software packages and we're working on tutorials to help get you started. Many of the steps, at this point, require that you are comfortable working with computer software on the command line. We encourage everyone interested to get the software and contribute to the effort of documenting, improving, and extending our computer code.</p> <p>Get computer software ></p>

Hyb-Seq (target enrichment) starts with the probe design

Appl Plant Sci. 2014 Feb; 2(2): apps.1300085.
Published online 2014 Feb 6. doi: [10.3732/apps.1300085](https://doi.org/10.3732/apps.1300085)

PMCID: PMC4103609

A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae¹

[Jennifer R. Mandel](#),^{2,9} [Rebecca B. Dikow](#),³ [Vicki A. Funk](#),⁴ [Rishi R. Masalia](#),⁵ [S. Evan Staton](#),⁶ [Alex Kozik](#),⁷ [Richard W. Michelmore](#),⁷ [Loren H. Rieseberg](#),⁸ and [John M. Burke](#)⁵

Appl Plant Sci. 2014 Sep; 2(9): apps.1400042.
Published online 2014 Aug 29. doi: [10.3732/apps.1400042](https://doi.org/10.3732/apps.1400042)

PMCID: PMC4162667

Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics¹

[Kevin Weitemier](#),^{2,7} [Shannon C. K. Straub](#),^{2,7} [Richard C. Cronn](#),³ [Mark Fishbein](#),⁴ [Roswitha Schmickl](#),⁵ [Angela McDonnell](#),⁴ and [Aaron Liston](#)^{2,6}

[Mol Phylogenet Evol.](#) 2015 Apr;85:76-87. doi: [10.1016/j.ympev.2015.01.015](https://doi.org/10.1016/j.ympev.2015.01.015). Epub 2015 Feb 14.

Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment.

[Stephens JD](#)¹, [Rogers WL](#)², [Heyduk K](#)³, [Cruse-Sanders JM](#)⁴, [Determann RO](#)⁵, [Glenn TC](#)⁶, [Malmberg RL](#)⁷.

 frontiers
in Plant Science

ORIGINAL RESEARCH
published: 17 September 2015
doi: [10.3389/fpls.2015.00710](https://doi.org/10.3389/fpls.2015.00710)

Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae)

[James A. Nicholls](#)^{1,2}, [R. Toby Pennington](#)², [Erik J. M. Koenen](#)², [Colin E. Hughes](#)², [Jack Hearn](#)¹, [Lynsey Bunnefeld](#)¹, [Kyle G. Dexter](#)¹, [Graham N. Stone](#)¹ and [Catherine A. Kidner](#)^{1,4*}

AMERICAN JOURNAL OF BOTANY

RESEARCH ARTICLE

SPECIES TREE ESTIMATION OF DIPLOID *HELIANTHUS* (ASTERACEAE) USING TARGET ENRICHMENT¹

JESSICA D. STEPHENS², WILLIE L. ROGERS, CHASE M. MASON, LISA A. DONOVAN,
AND RUSSELL L. MALMBERG

Department of Plant Biology, University of Georgia, Athens, Georgia 30602 United States

From transcriptomes/genomes, gene expression studies, the literature, or a combination of these sources.

A probe design for a non-model plant group as example

<> Code 🔔 Issues 0 🔗 Pull requests 0 📖 Wiki ➦ Pulse 📊 Graphs GitHub

Home

rschmickl edited this page 3 days ago · 21 revisions

Sondovač is a script to create orthologous low-copy nuclear probes from transcriptome and genome skim data for target enrichment.

When using Sondovač, please cite Schmickl et al. 2016: Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae) in Molecular Ecology Resources; DOI: [10.1111/1755-0998.12487](https://doi.org/10.1111/1755-0998.12487).

Sondovač is written primarily in BASH, so that it is portable among operating systems (any UNIX-based operating system - Linux, Mac OS X and more - is equipped with BASH, and it can be installed into Windows), although only selected Linux distributions and Mac OS X are fully tested and supported.

How to obtain Sondovač

- GitHub repository contains only basic scripts, not all files needed to run Sondovač - **get the latest release containing also all needed binaries and 3rd party software.**
 - Currently, the latest release is [v0.99-rc](#).
- Get separate package with [sample data](#).
- See [documentation](#) for information about installation and usage of Sondovač.


Notes


- In case of problems, requests, questions, feel free to open [new issue](#).
- Sondovač is currently tested on major Linux distributions and Mac OS X (see [README](#) or [manual](#) for details). Anyway, you can run it on any UNIX-based operating system.


▶ Pages 3

- See [basic information](#)
- Check [documentation](#)
- Download [latest release](#)
- Download [sample data](#)
- [Report problem, ask question and/or have some feature request](#)
- [Read paper introducing Sondovač](#)

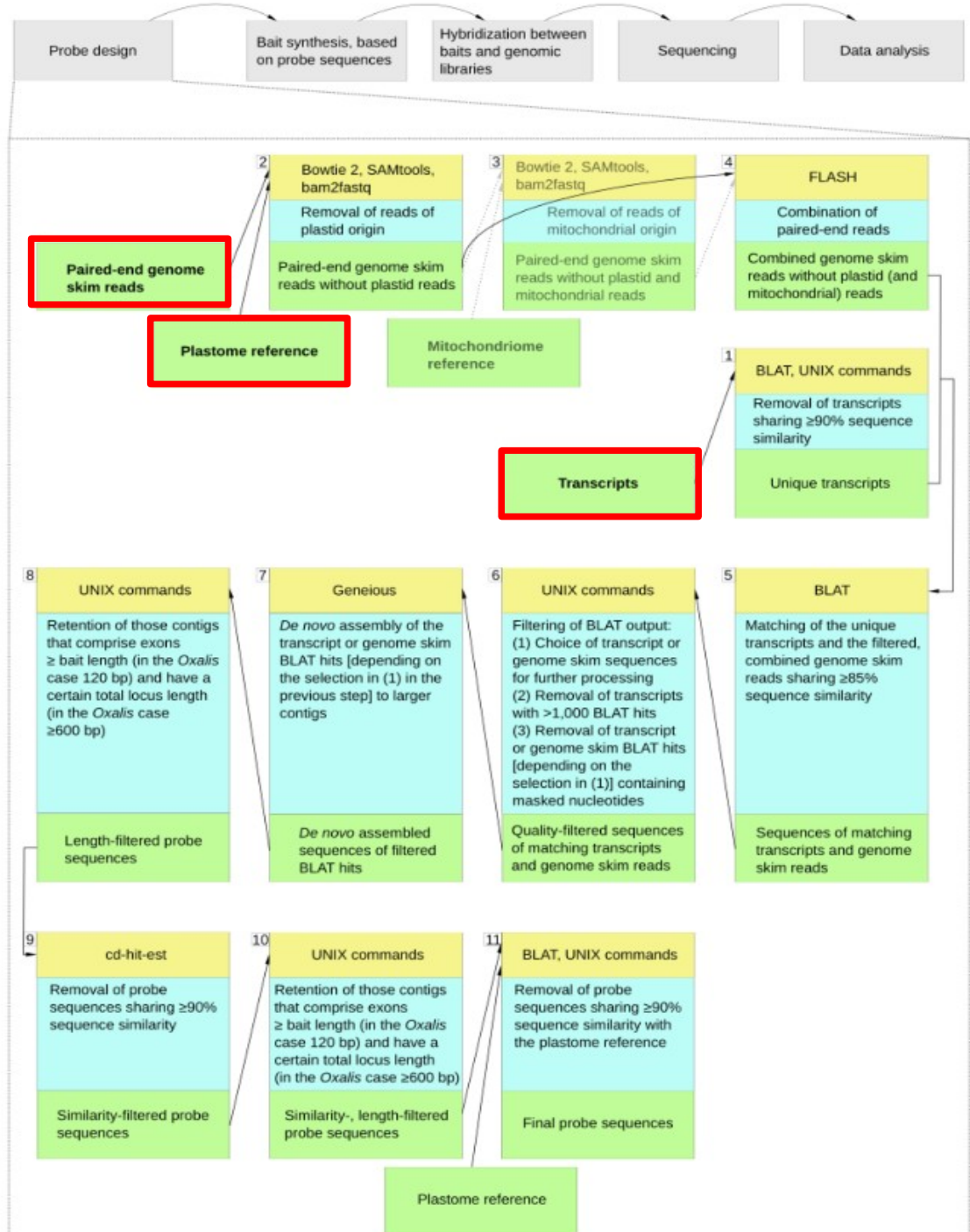
Clone this wiki locally

<https://github.com/V-Z/soi> 

 Clone in Desktop



Workflow of Sondovač



Input files for Sondovač: 1) transcriptome

The 1KP initiative as source for plant transcriptomes



1000 Plants

[HOME](#)

[CONTACT INFO](#)

[GREEN PLANTS](#)

[MEDIA](#)

▼ [SUB-PROJECTS](#)

[AGRICULTURE](#)

[ANGIOSPERMS](#)

[BIOCHEMISTRY](#)

[EXTREMOPHYTES](#)

[GREEN ALGAE](#)

[MEDICINES](#)

[NON-FLOWERING](#)

[SITEMAP](#)

Home

The 1000 plants (oneKP or 1KP) initiative is an international multi-disciplinary consortium generating large-scale gene sequencing data for over 1000 species of plants. Major supporters include Alberta Enterprise and Advanced Education, Musea Ventures (Somekh Family Foundation), Beijing Genomics Institute in Shenzhen (BGI-Shenzhen), Alberta Innovates Technology Futures (AITF-iCORE Strategic Chair), iPlant Tree-of-Life (iPToL) Grand Challenge, and WestGrid Compute-Calcul. Sample collection was determined by a series of overlapping sub-projects with scientific objectives that could be addressed by sequencing multiple plant species (see links to left). As more collaborators joined 1KP, however, the objectives evolved and are now exemplified by the diverse collection of papers described in the links below.

Many companion papers have already been published and a final capstone paper is planned for 2014.

[Catalog of manuscripts in progress.](#)

[Description of final capstone paper.](#)

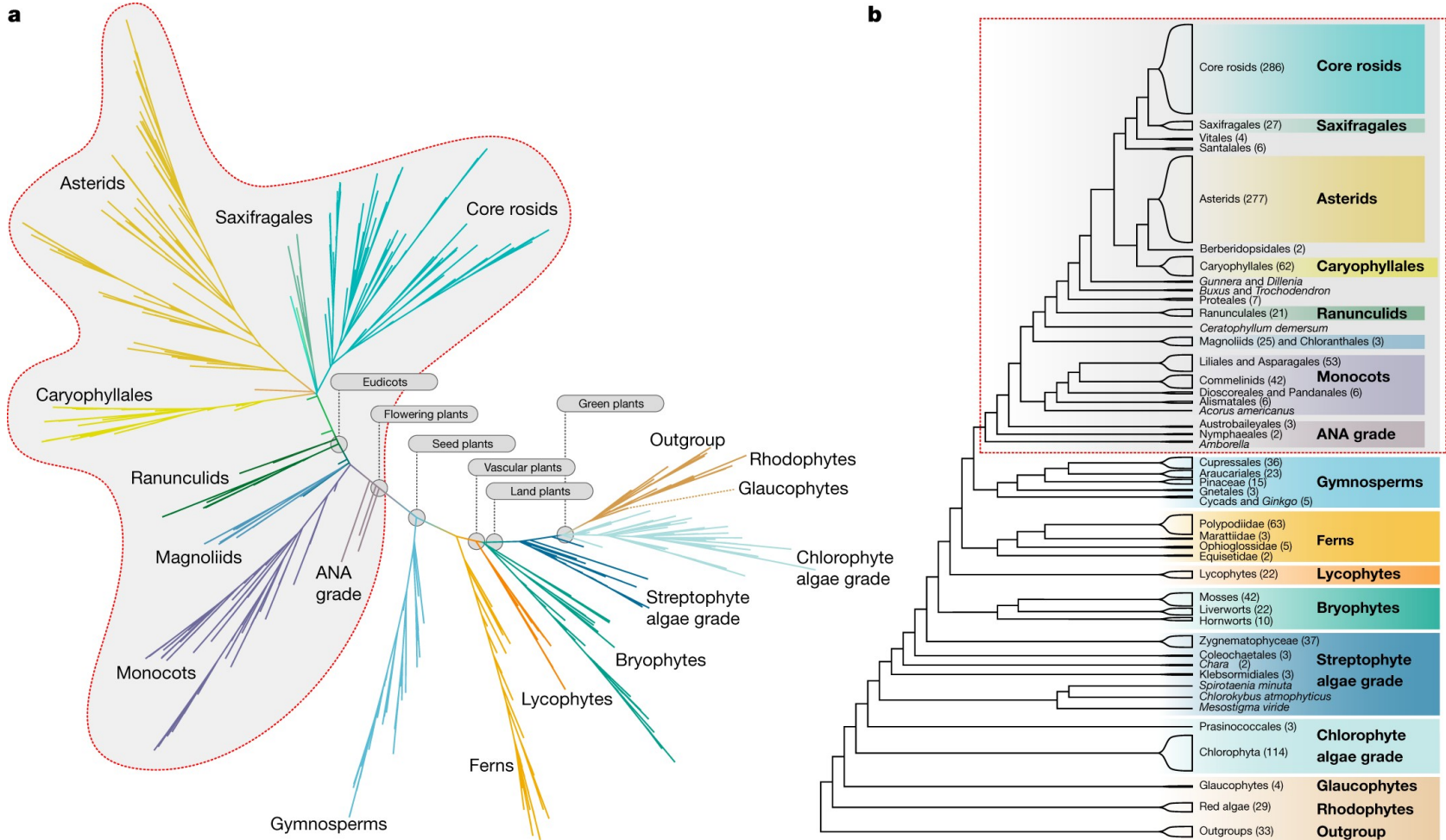
Limited access to the sequence is provided, in advance of publication, through a BLAST search portal.

[BLAST access into transcriptomes](#)

[Table of sequenced plant samples.](#)

Transcriptomes from the 1KP initiative as a source for phylogenies

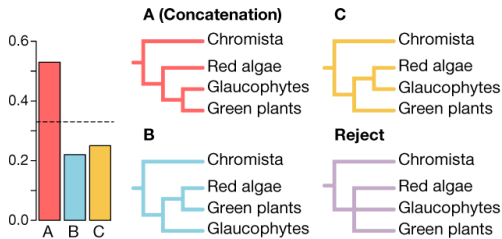
Plant phylogeny based on transcriptomes from 1KP was recently published!!



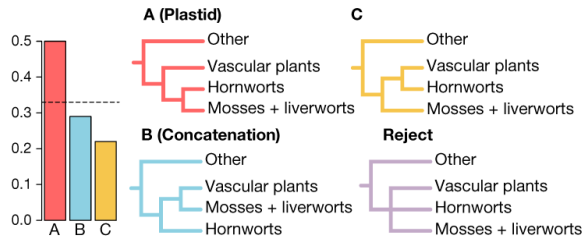
One thousand plant transcriptomes initiative (2019) Nature

Still, this does not mean that there is THE plant phylogeny ...

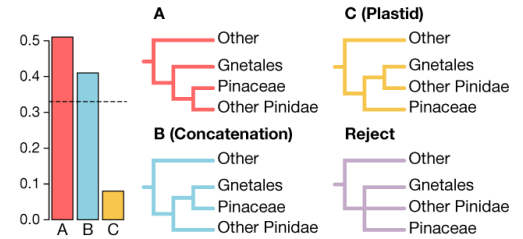
a Early Archaeplastida diversification



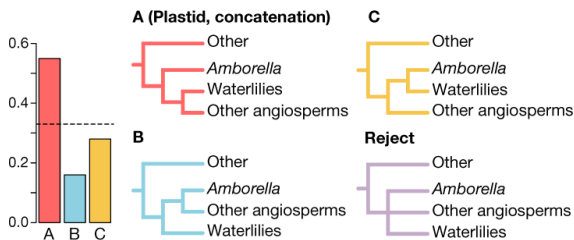
b Early embryophyte diversification



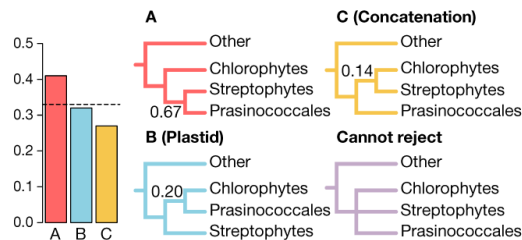
c Gymnosperms



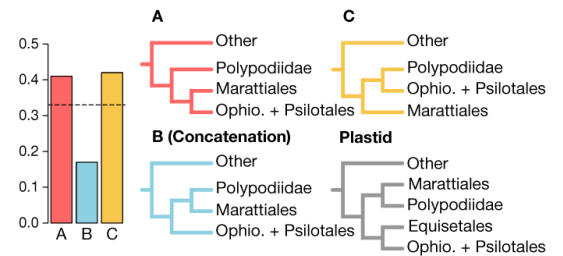
d Early angiosperm diversification



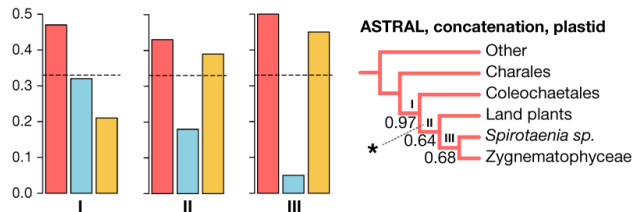
e Early Viridiplantae diversification



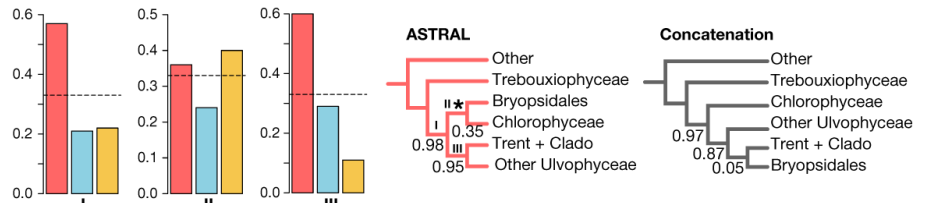
f Early fern diversification



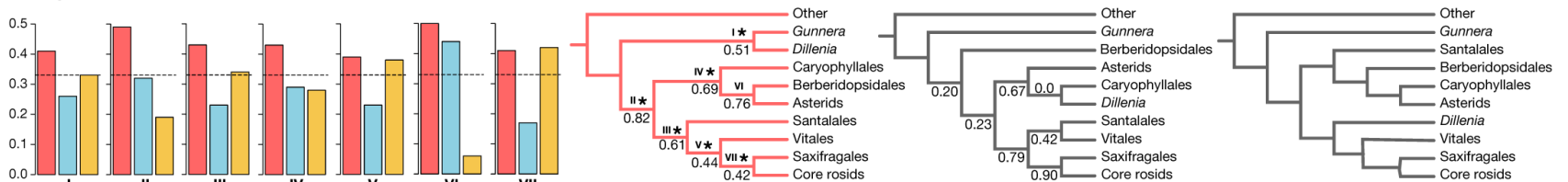
g Sister group to land plants



h TUC clade



i Early core eudicot diversification



Input files for Sondovač: 1) transcriptome

The 1KP initiative as source for plant transcriptomes

Transcripts in FASTA

```
>1
CAGCATCATGTACACCAACAACAAGGCAATTCTTACCAAATCAAACCTGCCATACCTCAAACGAGTTTTTCAGGGAGCAAACATTGCATATGCTGGCCAGATCAAATTTTCCCTGGATCTTCTAAGTTCTCTGGAGAAACGCAACCAAGTACTCAAACACCTGGTTTTAGTGCTCC
TTTTGCTAATCCAGCTCCCTACATGACTTCTGCAAATCCGTTTTATCCGCCAACCCACAGGCAGCTGGCTTCTATCCTCAGTATGCTCAGTGGACCATATCCTCTGAATTCAAATGTCGTTCCGCCGTTTTATTGCTGGCTATCCTCCTCAAATCCTTTTCCACTGTTTCCACTGG
TTGAAGGAAATCCTAGCACAAACCAAAATGTCGGGATCCGGGGTTCCAGCTGGTGGAAATCTGGGACTGGAGCTGATATGCAGCATTATAATAAGGTTTACGGACAGTTTGGCTATGTACAACAACCTCTTTTACTGATCCTATTTATATGCAGTACTATCAACAGCCTTTC
GAGGCATATAACATTTACCTCAGTTTGTATCCTCAGCTCCGAGAAGTACTGTTTTCCGGGAGCCAGATAGTACCCTTGACCCTCGGAAAGCATCTGTCAATGCTGTTCCCATGGATGATCAGAAATACCAGCAGCAGAGTACTGGTAGAGGCATAATGACAAGTCTTACTTTGG
TCAAGAAAACCTGGGCATGTTGACACAATATTCTGCTTCCACCGCTCCCGAGTCCAATATTACCAGCATACCAGCTGGCGGGATGGGAGGGAGAAATGAGGTGAGGTTTTCCGGTAGGTTCTGGTAGGTACACAGGTGTTTACTCC

>4
TGTGAAGTTTGTATTGCTGGTGGTAAAAATGTCGTCATCTGCAATTAATTGATTCTTCTAACACGATGAGTTGCTATTTTTATCCAAGATATCACCTTGATCAAAGCTCGCTGCTTTGTCGAAACCATGCTGCAAGCCGATGGTGTGGATGTATATCAAGATTATATACT
CACTTATCGTCTTTTGTATGACACATTCTCCACGTAATAATCTGGGGAGTTGACGCCTCTAGTACTCCAGATCTCCAGCTTCTACAGTGGGGAACCTCAATCATGACTGCAAGAGCCATCCTGACGCAATCGGTTTATCCCTGATCAAGTCCCAGGGAGTCTACCT
GAAAACGGGTTCTTCTTATCAGACTTGTGGTTCAAGAACCTACAAGGTGTTAATAATTGAGAGGCAATGGAGAGCTTTCAGTTCTCGATTTAGATGATGGAAGGGAAAGGGAGCTTACAGATTTCTGTTGAATTTATCTGGTTACCTGTGGCCCAACAGAGGAGAAAACAAT
GCTCCCGTTCCCTTCAATTGAAGATGTTTCATGGCTAGATTAAGGCCACCGAGGAATGCAGGTTTTGGTATCCATCTCCTGGTGTGACTCTTCAAGCAGGAAGATTTCTTACAGTTGGACCCAGAACTGGAATTTGATAGGGAGGTTTTATCCTTTGGGCTCCTTCCAAATGCTGG
TGTAGTTGGTGTTCCTTCAAGAGTTATCTTTTTCAGCATGCAACGAGTTCCCATGTTTTGAGCAACTCTCAAGCACAACATCTTACATTTGCTTACTTTCGACACCTTCTACAGAGAAAACAAAATGAAGAGGCTCTACGATTGGCCCAACTATCTACTGAAAAGCCTATTTTT
ATTGATGGAATGGCTTCTTTTCACTGATTTTGGGAGACATATCTAGGCAAGTGAACAAGAACCAGAATACAGCTTCTAACCTGTTGCTAAATTTCACTTTTGGACAAGACTTGTAAATTTGCTCCGAACTTCCGGAATCCTCGATATTGTTGTGAGTGTGCAAGA
ACGGATGCCCGACATTGGGCTGATTTGTTCTCTGCTGCTGGAAGATCCACTGAGTTATTTGAAGATTGCTTCAACGAAGATGGTACCGCCTGACGCTGCTATATTTCTGTGATAGCAAACCTGAAGGGCTGCTGTCACTGAGTACTGTGCTTACGTTTATTACAGGCTAC
TGACCAATCTTATATGAGTCTGACGAGAGCTGGTGGGTTTTACTGAGATCTGGAAGGGATTATGAACAAGCATCTACGGATTCCGCAACAACATCTCCTAGATTTCTGGGCTATTTTATTTTTGGATCTAGTTACAAAAGACCATCCTTGGATAAGAACCCCTCAATGAAA
AGAGTGGTCATATTGGCTCAGTGAAGAACATCTTGAAGAGTCAATGCAACTATCTAATGCTGGGAAAGAACTCTCGAAGCTTTGTCATTTGTCGAAAGGCACTCAGTTTGAATTTAGTTGAATATCTTCAACGCGAAAGATATGGGCTAGCTCGCTTGGAAAATTTTGGCTTCCGGA
GAGCTGATCGGAGAAAAGCTCCAAATGGGATTTGCTGACAGCCGTTAGATGCAAGAATTCCTGCTGCTCATATGTGCTCTGTCAAGTTCAAAGAGTGGGTTGATGCTTAGCCACACTTTTGGAGCCTTCTGAGGTCCTTTTGTATCTTTCAGACATGATGTTTCTGTTATGGAA
ATATGACATTACATTCAGCTCACCCGTCATTTCTGAGTACCACGATCTGCTATATGACTTGAAGAGAAAGCTTTCA

>5
CCAAATGAGTTGGTGGCTCCGGCAATGGAGATAATCCAATTTAGTATGCTTTTCAATGGAAGACCTGACCAGCATGATAGCTTCCAAGTTGCGGGAATGCATATTTGTTGGTGCAGAAATGAGTGGAAACAAGGGTGACAGTGTACACTCATGCGAGAATATTAAGAGTTCTTGGAGA
TGCCGAGGGAAATGCAATGAGTTGTCATGGCTTGCAGGAAATGAAGAAAACAAGAGAACTGCTGTGCTCTGTAAAAAAGATCTGATATGGAAAAGCTGCTTATACTCGGAGCATTTTGAAGCTTTTTATTTCGAGAATCTACATACAGAAGTTTTCTTGGGTGCTTCATGAAA
TAGCACGAAGACATGGTATCTCGAGGGACCTTCTGAAAGATTAACAAGTAAAGTTTTCTACCAAGACGGAGTGTGGTTGCTAAAAAGATTAAAGAGGAAAAAATAACAGGACATAGGAGAATCTAGGAAGTCAAACAAGAAAGAAAGAAAGATTCTCAGAAACCTGAAGGATCCATG
AATGCTGCCAAAGATGAAGAGGAAGACATCAAGTATCCAATTGATGACCTGTGGTGCAGCCTGGTTATGATGATCCAGTTTTCTACTGAGCTCCTTCTCCTTCAAAGGATTTCAATGATTCCAATGGATTGTGTGGGGATCTTTTAAATGGTCTGGGATTTTTGTACTTCTTT
CAAGCTACTACAATTTGCTGCTATTTTCTTGAAGATTTTGAATAATGCTATATGCCACAAGGACAGTAATTCGCTCAATTTTGAAGTCTCACTCGGCTCTTTTTCGGGTTCTCTTAAAAAACAACACTGAGTACAGATCCTTTGTAGCAAAAAGAAAGAGAAAGCAAGATTA
TACTCAATTTGGCAGAAATCTGTGTGATTTCTTGAAGATCGTAAACATTCCTGAATTTATGCACTCATGCGGCTACTATTAGGCGAGGCACTATGGCCTTGTAGATATCAGTGCTAAAGTCTGGTATCTTACGAGAATTTGTAATCAAGTCTGGAGACGAGTACTTTCAAGGTG
TTGGATGAACATATTGATCAGCGGCAAGCATTAGGAGCAACGAGAAGAGAAGAGGCAATTGGCAGAAAGTAAAAAGAAAAGAGAAGAGAAGGACCGCTAAAGTCTGAATCTGGAATCAATGGAGTAATTTAGTGACCATCATTTTGGACGAAGCTGAAGGGGACAGCCTATGTTGGC
TGATAATCATATCCAAGAAGATGAAAGTGTGGAAGAGCAAGAAATGGGGAAGTCACTTGTGCTCCGAAAGAACTGCAACAAGAAAGACAGAGAGATCAACACTCACTCACTGGATGTAGCAAAAGTTGAGAAGCCAAACATGAATGGAAAGGACTCGACAGAATTTGAAAATGATA
GAAACGATTCAGCATCTAACAGAAATCTCAGAAGCAGATTGATGGGAAAATGAACAGTGGAGCAAGGAACAAGATCAAAGAAAGAGAGTATTATGAACCGGAGATGGAAAATTTTTGTAAGCACCAATACCTTGGGAAAAGACAGAGACTATAATCGGTACTGGTGGTTT
CGAGATCAGAGAGTGTGTTGAAAGTTATGACATGATGGAGTGGGTTACTACAATACCAAGGAGGAGGTTGATGCTGATGAGGTTCCCTGAATATCAAAGTGAAGAGAGAGA
```


Input files for Sondovač: 3) organellar genomes NCBI organellar database as source

www.ncbi.nlm.nih.gov/genome/browse/?report=5

NCBI Resources How To Sign in to NCBI

Genome Genome Search

Genome Information by organism

Search by organism Clear

Download Reports from FTP site

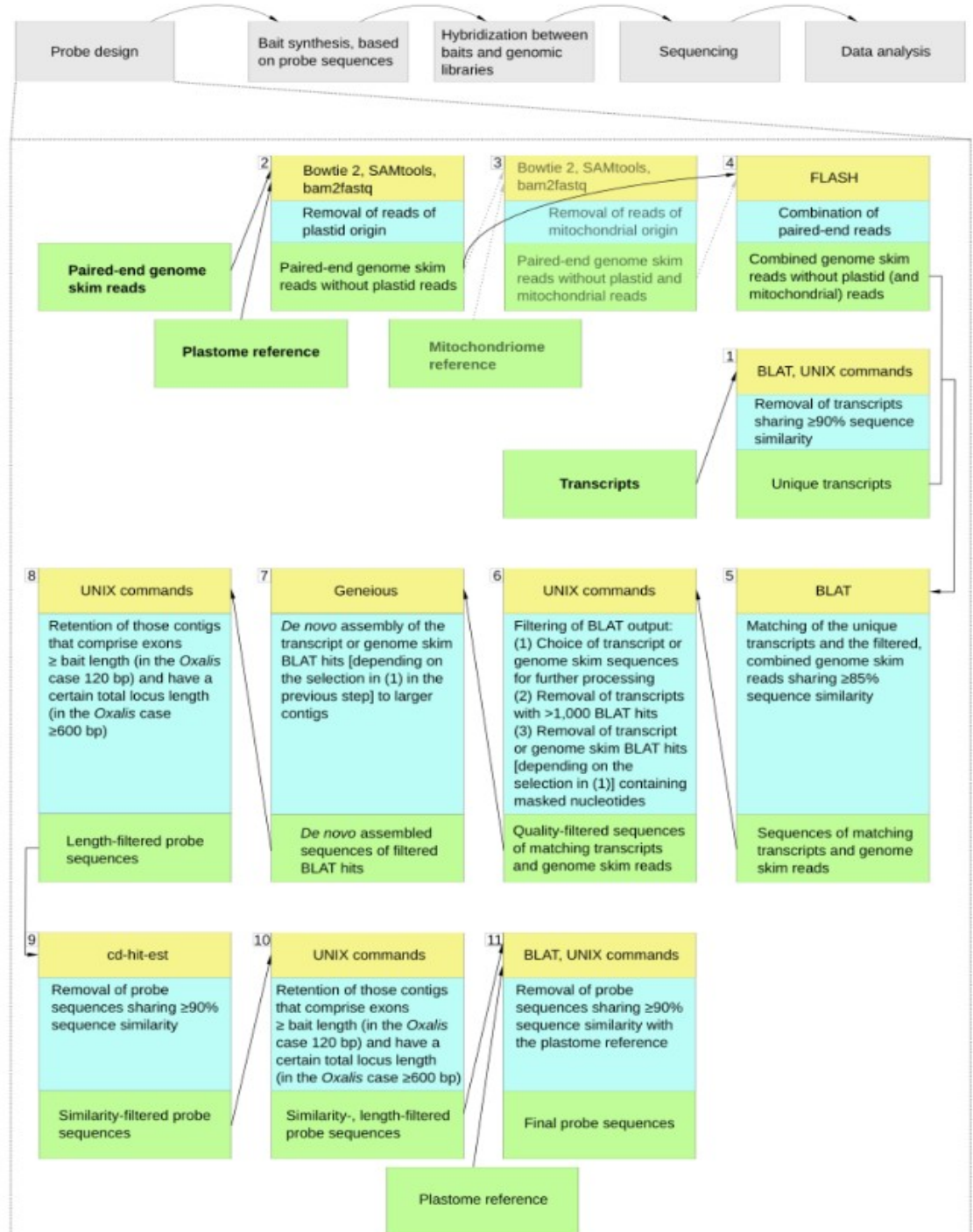
Overview [14508] Eukaryotes [2557] Prokaryotes [56235] Viruses [5052] Plasmids [6361] Organelles [7615]

Download selected records

Items 1 - 100 of 7616 << First < Prev Page 1 of 77 Next > Last >>

Organism/Name	Group	SubGroup	Type	RefSeq	INSDC	Size (Kb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene	Release Date	Modify Date
	-- All Eukaryota --	-- All Eukaryota --	All												
<i>Abacion magnum</i>	Animals	Other Animals	mitochondrion	NC_021932.1	JX437082	15.160	33.40	13	2	22	-	35	-	2013/08/06	2013/08/06
<i>Abalistes stellaris</i>	Animals	Fishes	mitochondrion	NC_011943.1	AP009202	16.502	44.42	13	2	22	-	13	-	2009/01/22	2009/02/06
<i>Abbottina obtusirostris</i>	Animals	Fishes	mitochondrion	NC_026900.1	KP727758	16.599	44.08	13	2	22	-	37	-	2015/04/22	2015/06/02
<i>Abbottina rivularis</i>	Animals	Fishes	mitochondrion	NC_023781.1	KF577979	16.597	44.29	13	2	22	-	13	-	2014/03/13	2014/03/13
<i>Abidama producta</i>	Animals	Insects	mitochondrion	NC_015799.1	GQ337955	15.277	22.65	13	2	22	-	13	-	2011/07/14	2014/12/15
<i>Abies koreana</i>	Plants	Land Plants	chloroplast	NC_026892.1	KP742350	121.373	38.25	74	4	35	-	113	-	2015/04/22	2015/04/22
<i>Abisara fylloides</i>	Animals	Insects	mitochondrion	NC_021746.1	HQ269069	15.301	18.83	13	2	22	-	13	-	2013/07/18	2013/07/31
<i>Abispa ephippium</i>	Animals	Insects	mitochondrion	NC_011520.1	EU302588	16.953	19.39	13	2	26	-	13	-	2008/11/04	2008/11/19
<i>Abiennus hiars</i>	Animals	Fishes	mitochondrion	NC_011180.1	AB373007	16.825	41.74	13	2	22	-	13	-	2008/08/27	2008/12/05
<i>Abramis brama</i>	Animals	Fishes	mitochondrion	NC_020356.1	AP009305	16.607	43.04	13	2	22	-	13	-	2013/02/25	2013/03/05
<i>Abronia graminea</i>	Animals	Reptiles	mitochondrion	NC_005956.1	AB080273	16.016	39.70	13	2	22	-	13	-	2004/07/02	2010/02/01
<i>Abronia inornata</i>	Animals	Birds	mitochondrion	NC_024726.1	KF742677	16.875	46.52	13	2	22	-	13	-	2014/08/26	2014/08/29
<i>Abrota ganga</i>	Animals	Insects	mitochondrion	NC_024404.1	KF590536	15.356	18.79	13	2	22	-	13	-	2014/06/24	2014/07/02
<i>Abudedefduf vaigiensis</i>	Animals	Fishes	mitochondrion	NC_009064.1	AP006016	16.703	45.41	13	2	22	-	13	-	2007/02/27	2007/02/28
<i>Acacia ligulata</i>	Plants	Land Plants	chloroplast	NC_026134.1	LN555649	158.724	36.21	82	8	36	-	126	-	2015/01/05	2015/01/13
<i>Acanella eburnea</i>	Animals	Other Animals	mitochondrion	NC_011016.2	EF672731	18.616	37.25	14	2	1	-	14	-	2008/06/23	2010/08/20
<i>Acanthacorydalis orientalis</i>	Animals	Insects	mitochondrion	NC_023482.1	KF840564	15.753	23.23	13	2	22	-	13	-	2014/02/13	2014/02/27
<i>Acanthaluteres brownii</i>	Animals	Fishes	mitochondrion	NC_011947.1	AP009212	16.441	47.29	13	2	22	-	13	-	2009/01/22	2009/02/06
<i>Acanthamoeba castellanii</i> Neff (ATCC 30010)	Protists	Other Protists	mitochondrion	NC_001637.1	U12386	41.591	29.40	40	2	15	-	57	-	1994/10/31	2015/10/01
<i>Acanthaster brevispinus</i>	Animals	Other Animals	mitochondrion	NC_007789.1	AB231476	16.254	43.63	13	2	22	-	13	-	2006/02/08	2009/04/15
<i>Acanthaster planci</i>	Animals	Other Animals	mitochondrion	NC_007788.1	AB231475	16.234	43.66	13	2	22	-	13	-	2006/02/08	2010/02/01
<i>Acanthis flammea</i>	Animals	Birds	mitochondrion	NC_027285.1	KR422696	16.820	45.49	13	2	22	-	37	-	2015/06/18	2015/06/24
<i>Acanthocardia tuberculata</i>	Animals	Other Animals	mitochondrion	NC_008452.1	DQ832743	16.104	40.10	12	2	23	-	12	-	2006/10/05	2006/10/13
<i>Acanthocheilonema viteae</i>	Animals	Roundworms	mitochondrion	NC_016197.1	HQ186249	13.724	26.46	12	2	22	-	12	-	2011/11/15	2012/09/14

Workflow of Sondovač



Final probe sequences in FASTA

```
>Assembly_10081.Contig_10_329
AGCTATGCTTTTACTGTGGCTCTTCAAGAGCTAATCTGAGTGTGATGAATCCTGGTCCACTAATTACGCCACCAAAAAAGGAGAAGGCTTACCATGGTGGGATGAAATGAGAATTACATTCATGGACACATCAGCTTATTCTTTGCTGAAACTAATTGGCATATTCTTGCAGCTAC
TAATCCTTATKAAAAGCTTGACAAACTTCAAATTTTAACTGGATCTATGGAGATTCAGCAGTCGGACGGCGTATTTATGTCTCTGCAAAGGATTTCAAGGCTTCTACTGAGCAGTTTGGAGAGCTTACTAAATCGGCCAACTTAAAAAC
>Assembly_10081.Contig_1_398
TTGAATGGTTGCACAGAAAAGCATCGTGACGATGGGTTTCTATTATCATCTGATTACTTCAACAATAAGAAAAGCAGGCTCAAAGGCTGACCCAGCTAGATTATTGGCTTGGSAAGATGCGYGGGAGAAAAATATCGAGATGACATATGTCAGGCTGAGTGTGAGCATGCAAGTGAGAG
TGATGAGCACACGAGATCAGATCCAAGTGAYGATGATGGATACAATGTGGTTATTGCTGACAATTGCCAGCGTGTTTTTATTTATGGMCTTAARCTTTTATGGMCTCTTGAAAACAGAGATGCTGTTTGGTCTTTGTTGGTGGACTGTCAAAGCATTGCGCCCTCAAACCTCAC
CTTCGAGGCAATATGCACAAAAGAAATTAATGAAGAGAA
>Assembly_10081.Contig_2_316
GTTGGCCTCGATTCCGAGTTCCTCGAATCCAGATCAGGCAATTTGTCAATGGACAGGGTGTGACAGAATTCAGTCTTCTGATAGATGCTACGCCATCTCTATAAAGCACGTACCTTTAGATGACAGATGATCCTGCCAAAGGGTTGACCTTTACTATGAATAAGTTGAAATATGAA
ATGTATTTTAGTCGAGGTAAGCAAAGTTACCTTTGATTGCCAGCGTCAGACTTTGATCTTGTATATCAGGGTGTGACCTTACATGCCAAAGATTTTTATAGATAAAGAAAGCTTGACCAGCTCCACAAAAGT
>Assembly_10081.Contig_3_294
GGTAGATTTCTGCTGCTGCTGTAGTGGGCGTGTTTAGCCCGGTCATTTCACTCAGTCTTCAAGTTGGATACGAGTTATAGAACAAGCGCTTGGCACAGAAAATGTTCAAATCCTGAATGTGAACCTGAAATGACATGGAACGTATGGAGTCTCTGTTATGTTGGAGCATGT
TCAGGCTCATGTTGCTCAAATGATGTTGATCCAGGTGCAGGAATCCAGTGGCTTCGAAAAATTTCTAGAAGCTTCCAAAAGTGAAAGCGCACTGGTGTCTTCTTGAAGAGKG
>Assembly_10081.Contig_7_122
AGGTGAAGCCTTTAAAAGAGCTCACTTTTAATTCACATAACATAACGGCAACAATGACATCTCGCCAGTTCAGGTCATGCTAGATGTTGACTAATCTTCTGTTTGCACGGCTCCCAAAG
>Assembly_10081.Contig_8_121
TCTTTACTCTAGAAGTTACAATGAATGGGGATGTGATTCTGGTAATCCATTGAATCATTATTTATTTGCACTTCTAATGAAGGCAAGCCCCGTGAATTTGTTTTGATCCTTTTCGATC
AGAGAACTTATAAATGCACAAAATCTAAGRAGGCGACTCTGCATTATAAAGGTTGGCTCTTCAGAAAGCTCTCAGCAACGAATCATGAAAAGGAAAAGAACAAGTCCATCATATGCAATGCGTATTTCTCTGCAAAATACACAAAAGTTGTTGGAGCATGCTTTAGATGGTAA
ATCCTTTGCTGAGGCTGAGATAAATGACATG
>Assembly_10133.Contig_10_134
AGAGAGGCTTGCATTCATATTGACGTTGTTGCTGATATATTGATCTCGCAAAGTCTTGAAGGGAACACATTTTCATATCATTAGGGATAATGTGTAATTCCTCGAGGTCACACATTAGCAGGATCTT
>Assembly_10133.Contig_1_140
CAGTGTGCTTAATAATTGGTGCATATTCGAAATGGYTTGATACTGCATCGATGGTGTATCCATATTACCCTCAGTCTCGATATTCTAACAAATGGCATGGGCAATRTCTGAAGAYTCTGCTGCAGCCGCTGCTTTGGCT
>Assembly_10133.Contig_2_125
GATTGCGGAAAAGCTATGTGGATTCTGTGATGGTCTCTTCTCTYATATACCAGACAGCAGTMAGTGGGGAAGGCAGATTTAAAGTCTCTCTGAGGACTCTCTTATCTTGTGAGGCATTGAG
>Assembly_10133.Contig_4_130
GTTATTTCTGAATTGATCCATTACTCAGCAACTGGAAGTCCGGTGGCATTCTGTGCAATGCCACTGATTCAAGTAATGTGCCACAAGTGATGAATCTAAAACCTCAGCTAAGGGATTCTTCAAAGG
>Assembly_10133.Contig_6_136
AGGAATTCCTGACGACCTRACATAGCTGATGATTGCTTCTGCTAGCATCGAGATGATACGCTATTGTCTCAGCTATTCATACCATCTCCTGATTTTCCATGTCTAGTTGATTGCTATGAGTGGGATCAC
>Assembly_10133.Contig_7_125
ATGGAGCTGCAAAACGACGCTCAAAGAAGCTCTCAACGCACTATCATCATCCGAAGATATGGTTCGTGTTCAAGCCGATCGTTGGCTTCAAGACTTCAACGCACGATCGACGCTTGGCAGGT
>Assembly_10133.Contig_8_120
GTGAGAATCTTGGAAAGGTTAATGGGGATCACAATTTGGAGCGATGCTAGAAGAAATCAAGGCCCTTATCAACACCACCTCCAGCCATGCTTCTCTATCTCTCAAGTGAAGTTATAAAG
>Assembly_10133.Contig_9_136
AGGAAGTGTCACTACAAGATAGCAGCTCGCTCTGAAAGACGTCGCCAATTTGAAAAGGAGCTTACTTCTCAAATGGAGATAGCTCTTAATATATTAACATCTTGCTGAGCATTAAATGGACTGGCAGAGCAGGT
>Assembly_10176.Contig_1_431
ATGGAGATGTATCAAAGAATGCAAGTGGCGGTTTGAATCCCGACACATTCTCTTACAGTGTGATGATAAATGCCTTGGTAAAGCTGGACACTTGCCTGTGCGCAGAAAATTTCTTGGAGATGATTGATCATGGTTGTGTCCTAATCTAGTCACATACAATATCATGATTGCTCT
VCAGGCCAAAGCAAGAAATACCAAAGTCCCTGCAAATTTACCGTGACATAGAAGATCAGGATTTAAACACAGAAAAGTGACTTACAGATTTGTGATGGAAGTRCTTGGCATTGYGGTYAYCTWAGTAGAAGMGAAGCGGTTTTTACCGAAATGAMGCAGAAAYACTGGGTCGCCG
ATGAGCCTGTGTACGGTCTTTTAGTAGATTTATGGGGTAAAGCTGGTAAAGTTGAAAAGCATGGGAATGATGA
>Assembly_10176.Contig_3_364
TCGCCATATGATATGGGGTTCTGTCTGTGAGCTTATGGCAATCTCAGGCCACCCAGCATATACATTCTTACTATCCATGCCAGCAGCGGGGCCAGCGTAAAAATGTCGGGATCATGTGAGCAATTTCTTAGATTTGATGACAGTGGAGACCGGGAGAGCAAAAAGAGGTTTAATTGA
TGCCGTGGTTGATTTCTCACAATCCGGGCTCAAGGAAGAAGCTAGCTCAGTTTGGGAGTGGCTGCACGGAAAAATGTTTATCCAGACGCTGTTAAAGAAAAAGCTCATCTATTGGTTGATTAATCTTACGTTATGTCCGATGTCAGCTGTGACTGCACCTCTCGAGAACAC
TAGCTT
```

Our pipeline for phylogenetic marker development for target enrichment of low-copy nuclear genes in southern African *Oxalis*

utilizing a transcriptome and genome skim data, resulted in

- ca. 5,000 exons ≥ 120 bp
- >1,000 genes of 600-4,125 bp (mean 968 bp) length



The *Curcuma* bait design in numbers

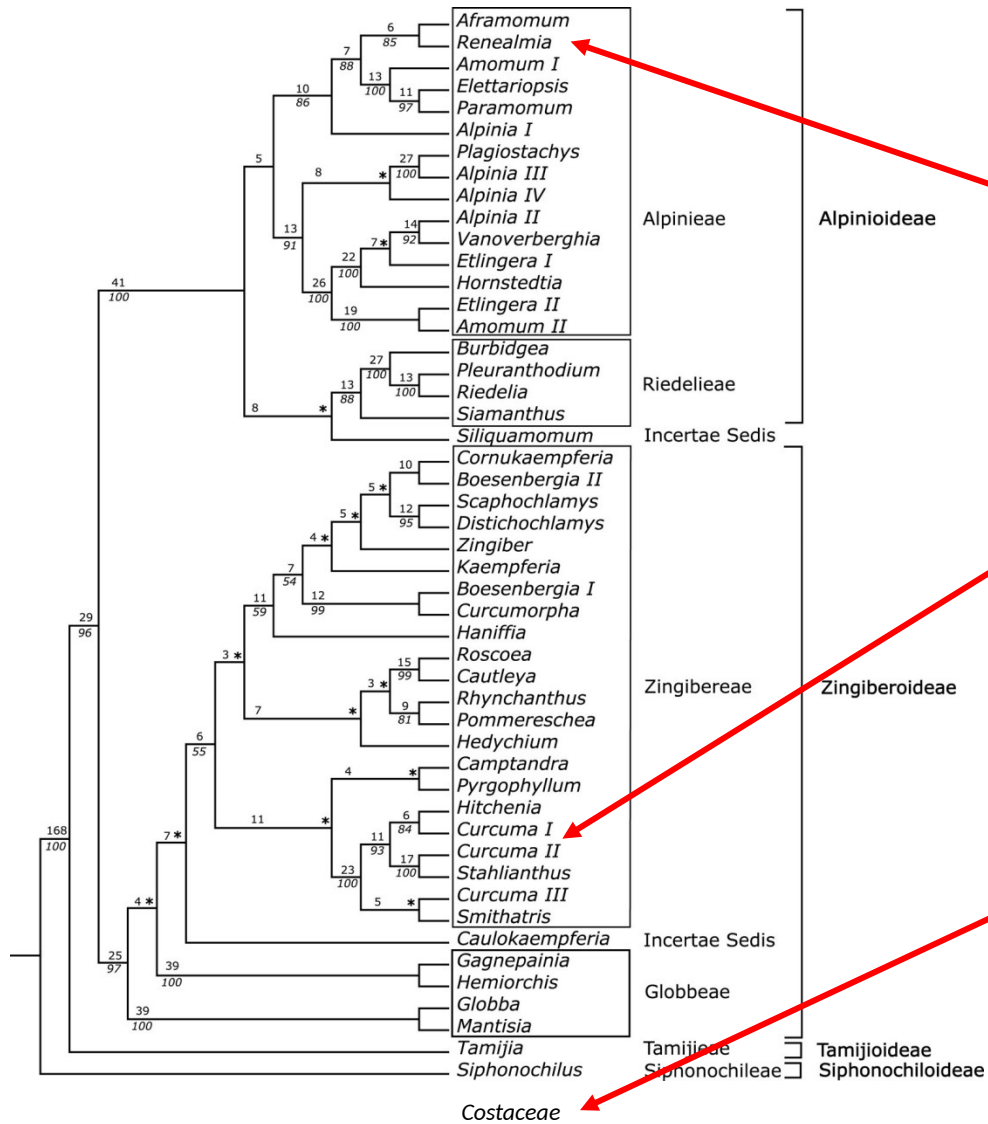
Transcriptome	<i>Curcuma longa</i>
Genome skimming data	<i>Curcuma ecomata</i>
Reference genome to remove chloroplast reads from the genome skimming data	<i>Zingiber spectabile</i>
Reference genome to remove mitochondrial reads from the genome skimming data	<i>Oryza sativa</i> var. <i>indica</i>
# BLAT scores: transcriptome vs. transcriptome	33,667
# Unique BLAT scores: transcriptome vs. transcriptome	17,203
# of exons ≥ 120 bp detected with a BLAT search between the unique transcripts and the genome skimming data	4,618
# of genes ≥ 960 bp	1,180
# of bp covered by genes of ≥ 960 bp	1,571,800



Is one bait sequence sufficient for the family Zingiberaceae?



Bait to genomic library hybridization efficient in case of <15% sequence divergence between baits and library

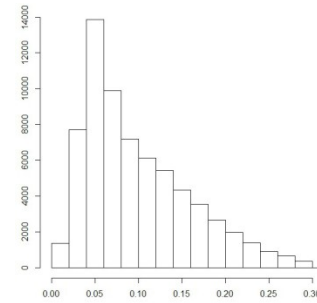


Curcuma longa versus



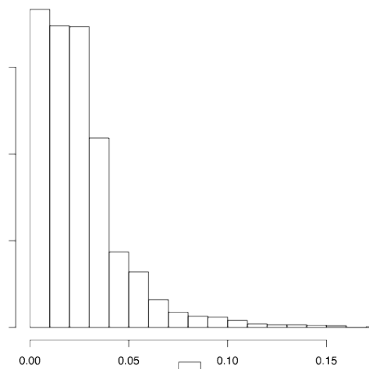
Renealmia nicolaioides

~ 5%



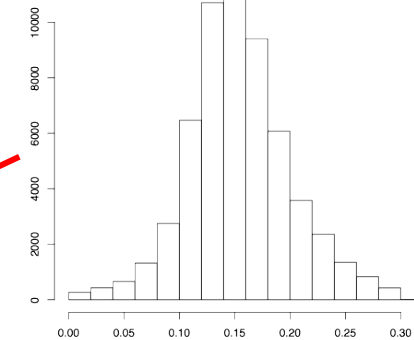
Curcuma ecomata

~ 0-3%



Costus pictus

~ 12-16%



Sequence divergence between transcriptomes

**Introduction to hybrid/target
enrichment/capture and Hyb-Seq
(wet lab)**

Hyb-Seq (combination of target enrichment and genome skimming): general workflow

Custom probe design

Bait synthesis
(usually outsourced to company)

Genomic library preparation
(5 ng – 1 µg DNA,
partly degraded DNA also works)

In this step target
enrichment is
combined with
genome skimming.

Hybridization of baits to genomic library

(100-500 ng DNA of genomic library,
tested with a minimum of 9 ng per sample in
a 24plex reaction)

Sequencing of enriched targets (e.g., nuclear exons)
and off-target sequences (mainly plastome and
nrDNA cistron)

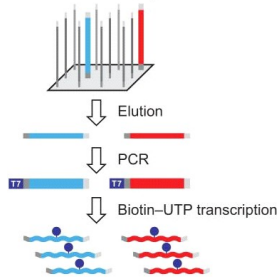
Data analysis

Hyb-Seq wet lab overview

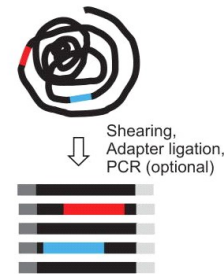
probe design

bait design and synthesis
(MYcroarray MYbaits)

Microarray



Genomic DNA

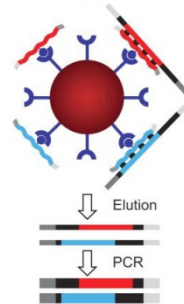


DNA extraction

sonication

library preparation

solution hybridization



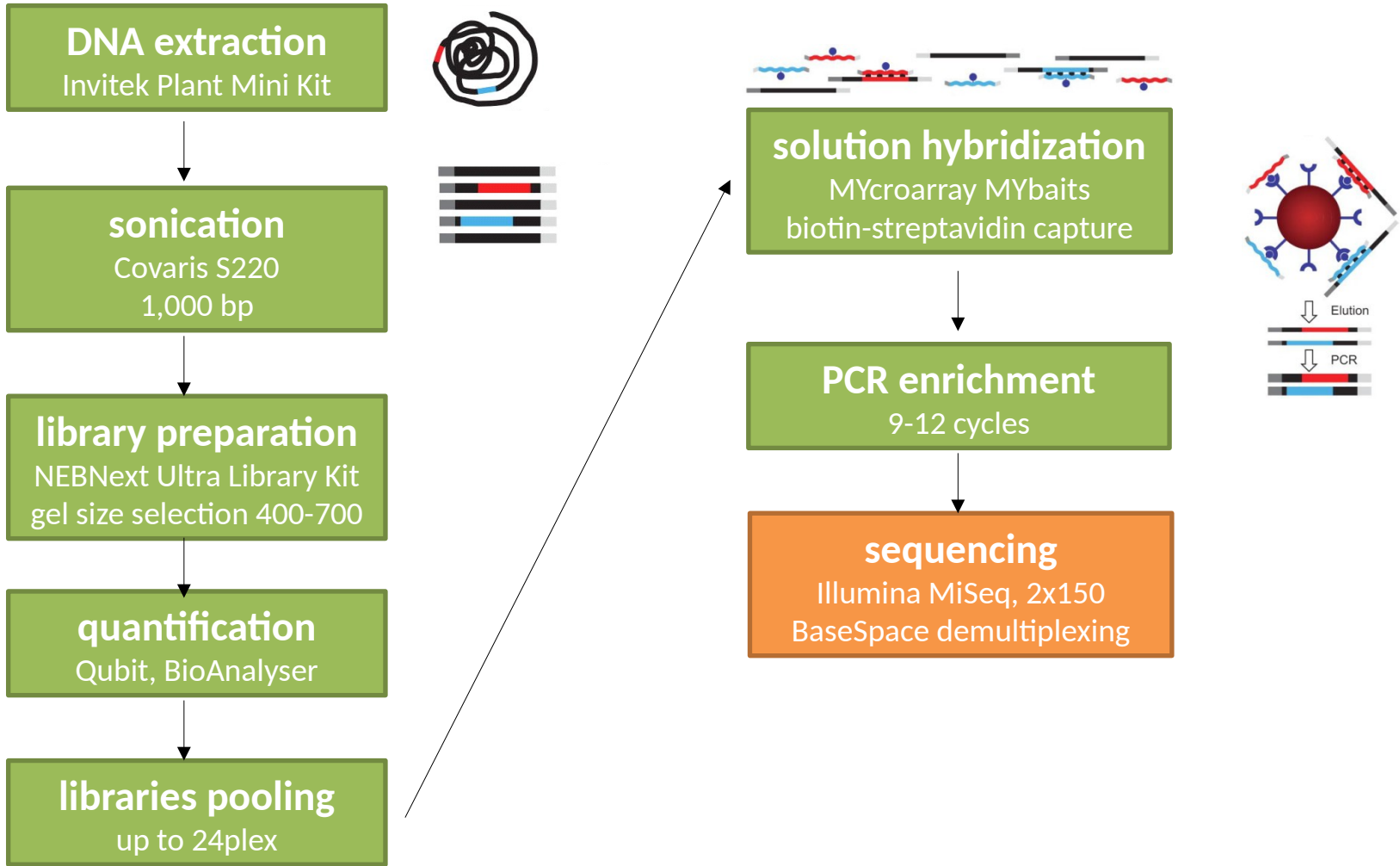
Gnirke et al. (2009) Nature Biotechnol.

Illumina MiSeq
2x150 PE

sequencing

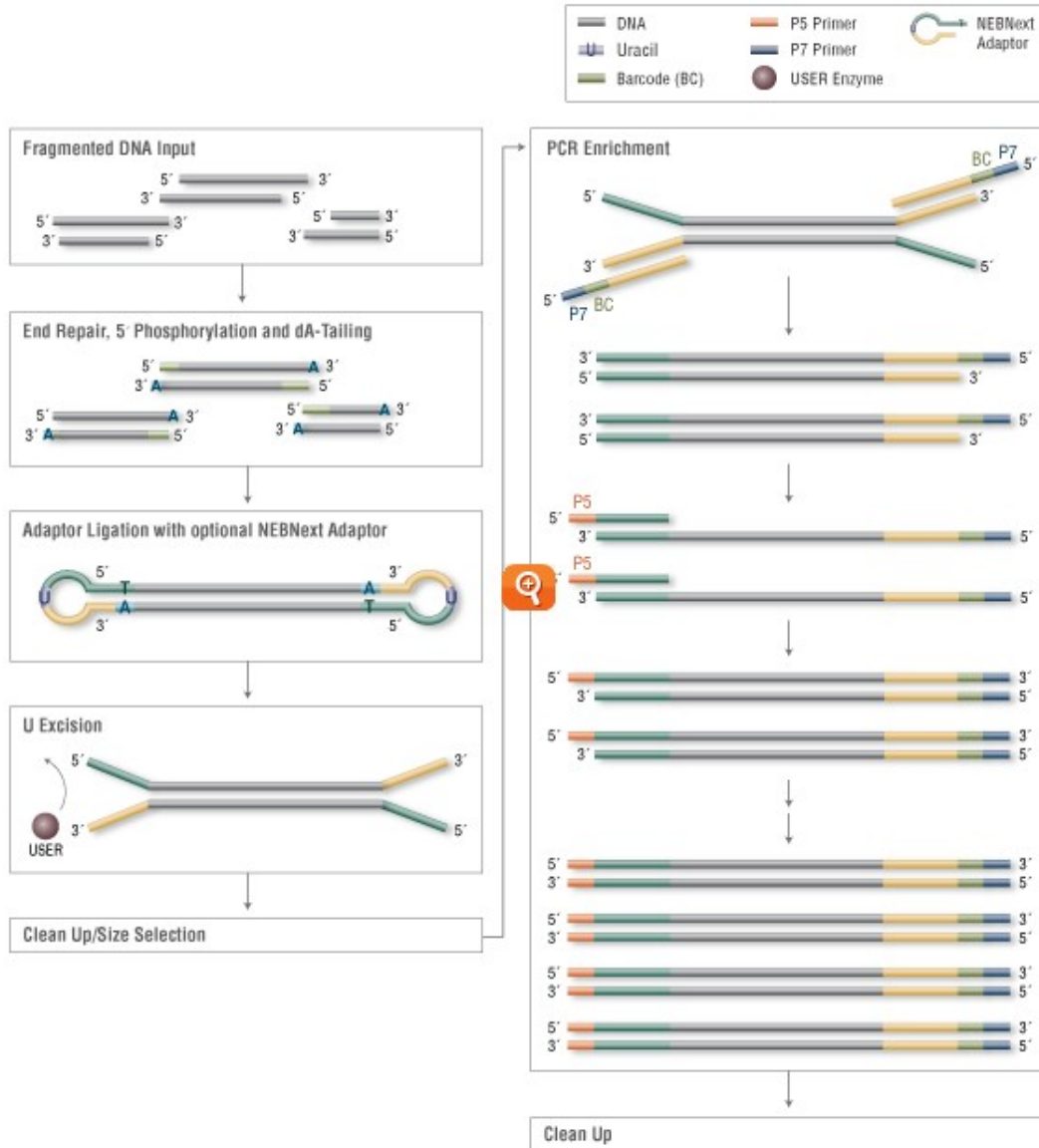
data analysis

Hyb-Seq wet lab overview in more detail

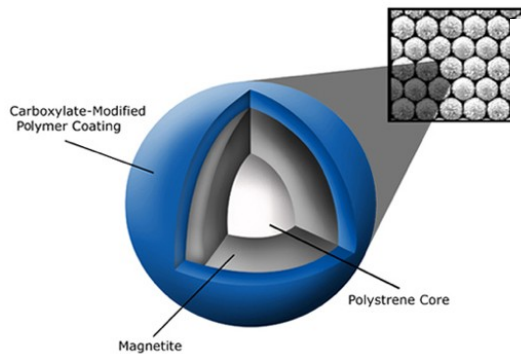


Genomic library preparation (example)

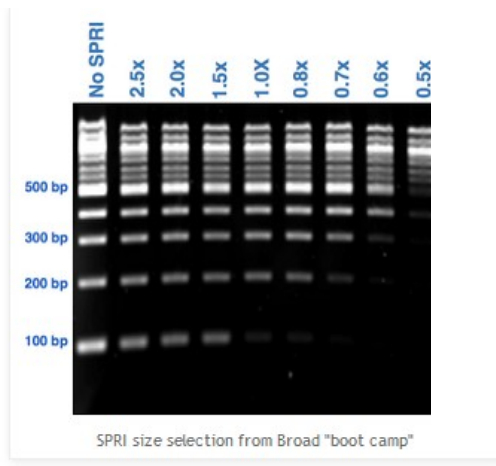
Ultra DNA Library Preparation Workflow for Illumina



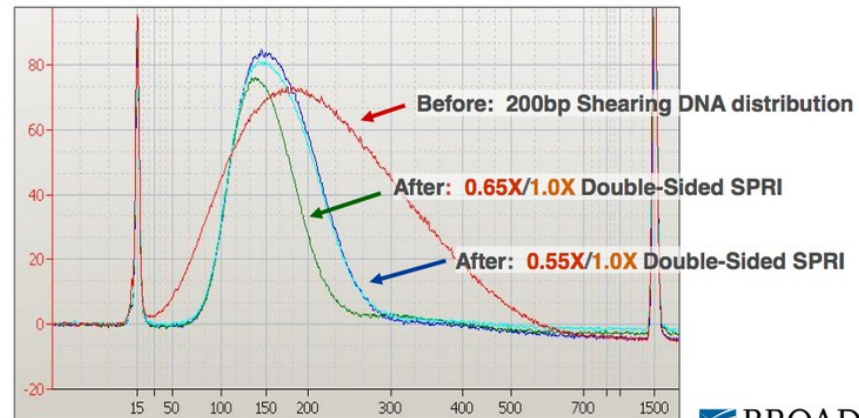
Size selection - with AMPure XP beads (example)



- ▶ By implementing a combination of good shearing with SPRI and “reverse” SPRI, one can select a fairly tight size range *with no gel*:



Results:



The perfect genomic libraries for Hyb-Seq

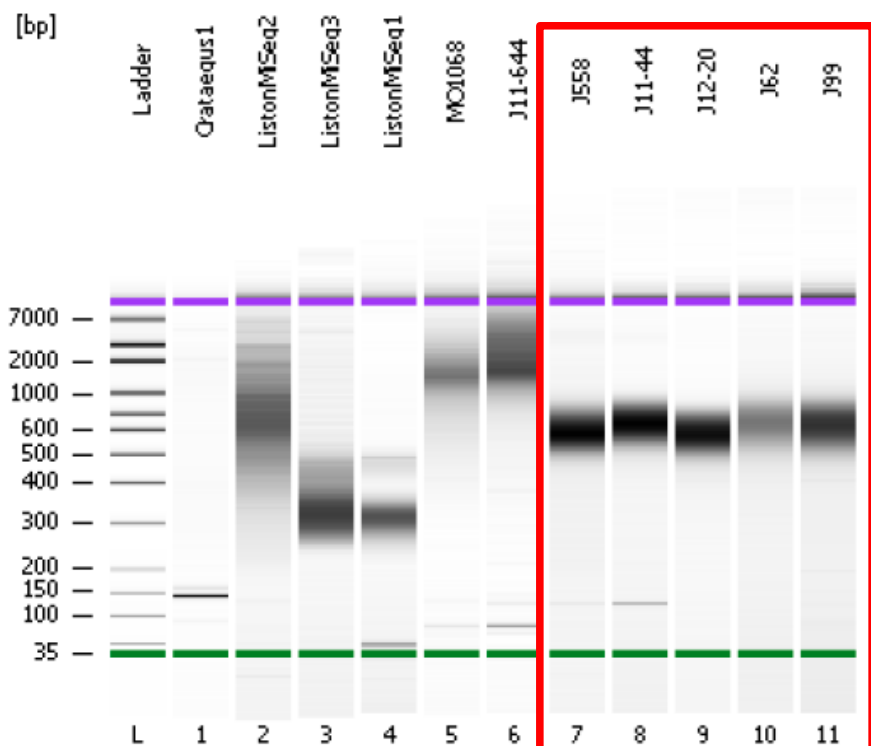
H2770_Schmickl_2014-05-12_12-06-28.xad

Page 1 of 19

Assay Class: High Sensitivity DNA Assay
Data Path: C:\AgilentData\2014-05-12\H2770 Schmickl 2014-05-12 12-06-28.xad

Created: 5/12/2014 12:06:27 PM
Modified: 5/12/2014 12:47:57 PM

Electrophoresis File Run Summary



Instrument Information:

Instrument Name: DE24802282 Firmware: C.01.069
Serial#: DE24802282 Type: G2938B

Assay Information:

Assay Origin Path: C:\Program Files\Agilent\2100 bioanalyzer\2100 expert\assays\dsDNA\High Sensitivity DNA.xsy
Assay Class: High Sensitivity DNA Assay
Version: 1.03
Assay Comments: Copyright © 2003-2010 Agilent Technologies

Chip Information:

Chip Lot #: sa28bk50
Reagent Kit Lot #: set a
Chip Comments: HS-DNA | High Sensitivity DNA | Roswitha Schmickl | Jan Suda | 420 271 015 490 | roswitha.schmickl@ibot.cas.cz | Plants | Leaves |

Bait hybridization (example)

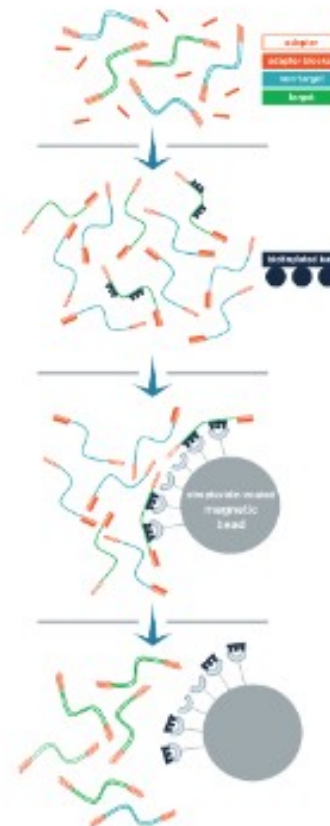


INTRODUCTION

myBaits[®] is an in-solution NGS library target enrichment system, compatible with Illumina[®], Ion Torrent[®], and many other sequencing library types. We use a versatile nucleic acid synthesis technology to make biotinylated RNA "baits" that are complementary to your sequence targets. Baits and other reagents for NGS target enrichment are supplied with the myBaits kit.

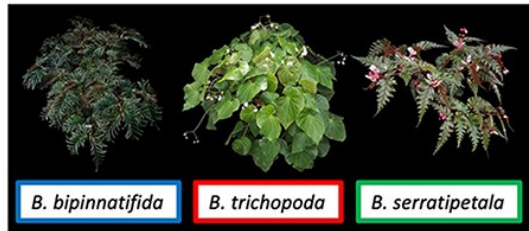
Procedure overview

1. Sequencing library, adapter blockers, and other hybridization reagents are combined
2. Libraries are denatured and cooled to allow blockers to hybridize to adapters, and then baits are introduced and allowed to hybridize to targets for several hours
3. Bait-target hybrids are bound to streptavidin-coated magnetic beads and sequestered with a magnet
4. Most non-target DNA is washed away, and the remaining library is amplified



Impact of preservation techniques on target enrichment success

A

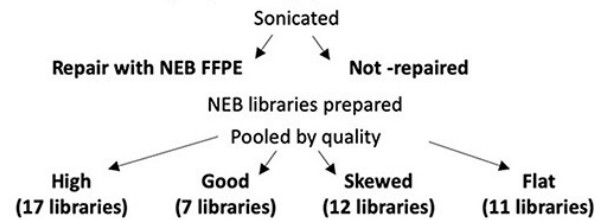


Leaves treated with 7 preservation methods

Alcohol_drying_room Ambient Hairdryer Drying_room Pickled RNAlater Silica

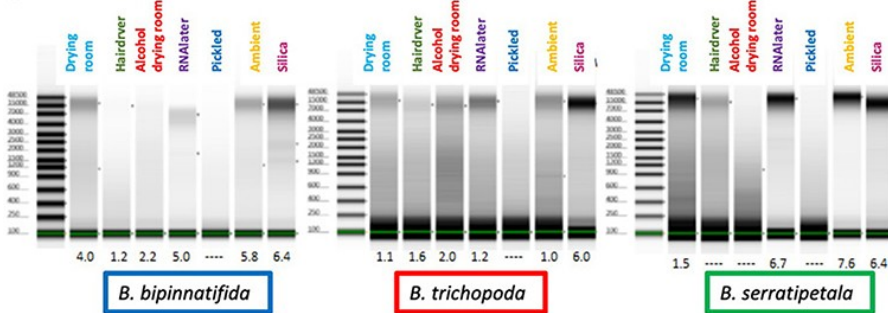
Four 1cm² of leaf samples of each species-treatment for DNA extraction

Samples pooled by species-treatment



Hybrid Capture and Sequencing

B



C

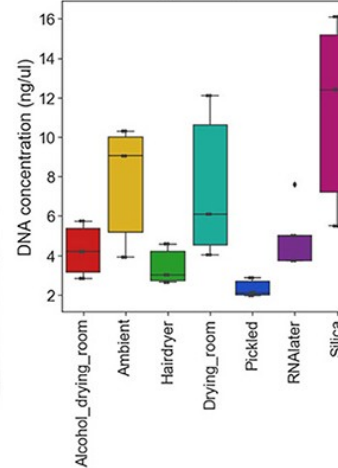
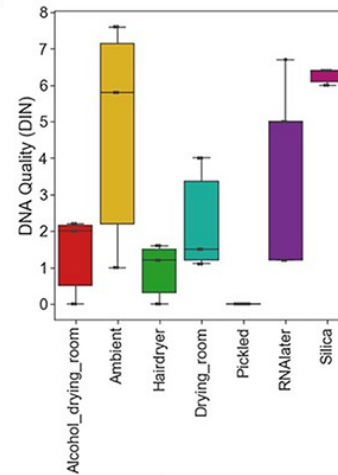


Figure 1. DNA quality and quantity.

(A) Individuals sampled.

(B) Tapestation images of DNA from 4 pooled extractions, with DIN values listed at base.

(C) Effect of drying method on DNA quality and concentration by treatment.

DNA quantity and quality does not affect read number

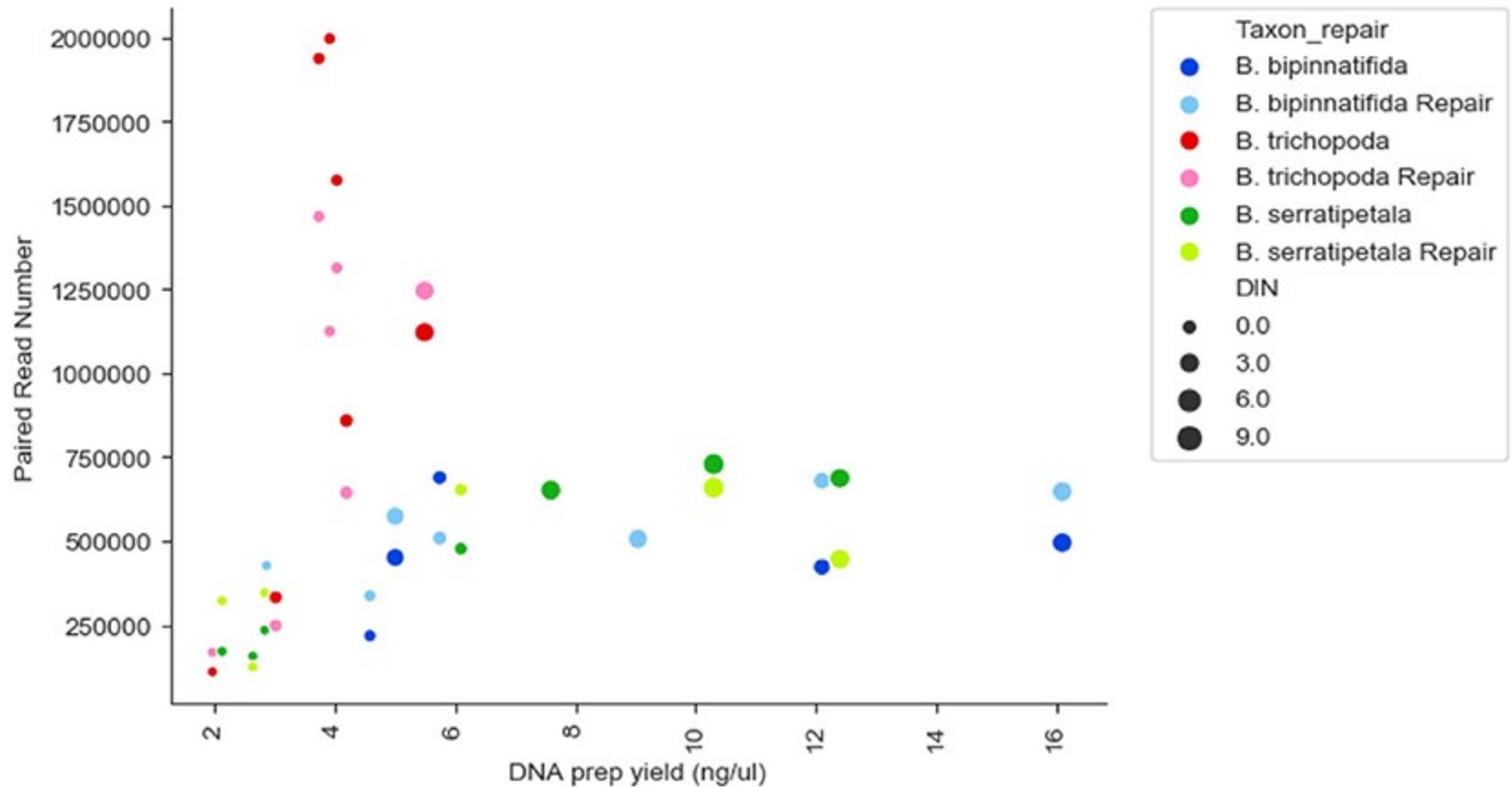


Figure 3. Reads generated per sample by DNA prep yield, DNA quality, and species. DIN is represented by size of marker, species by color of marker.

In **aging specimens**: **accumulation of thymine bases** due to the deamination of cytosine increases with time, leading to an excess of C to T substitutions toward both ends of the DNA fragments.

However, sampling **herbarium** and fresh material of the same individuals separated by 40–120 years did **not find the types of nucleotide misincorporation that are associated with ancient DNA**.

Enrichment efficiency using differently preserved material

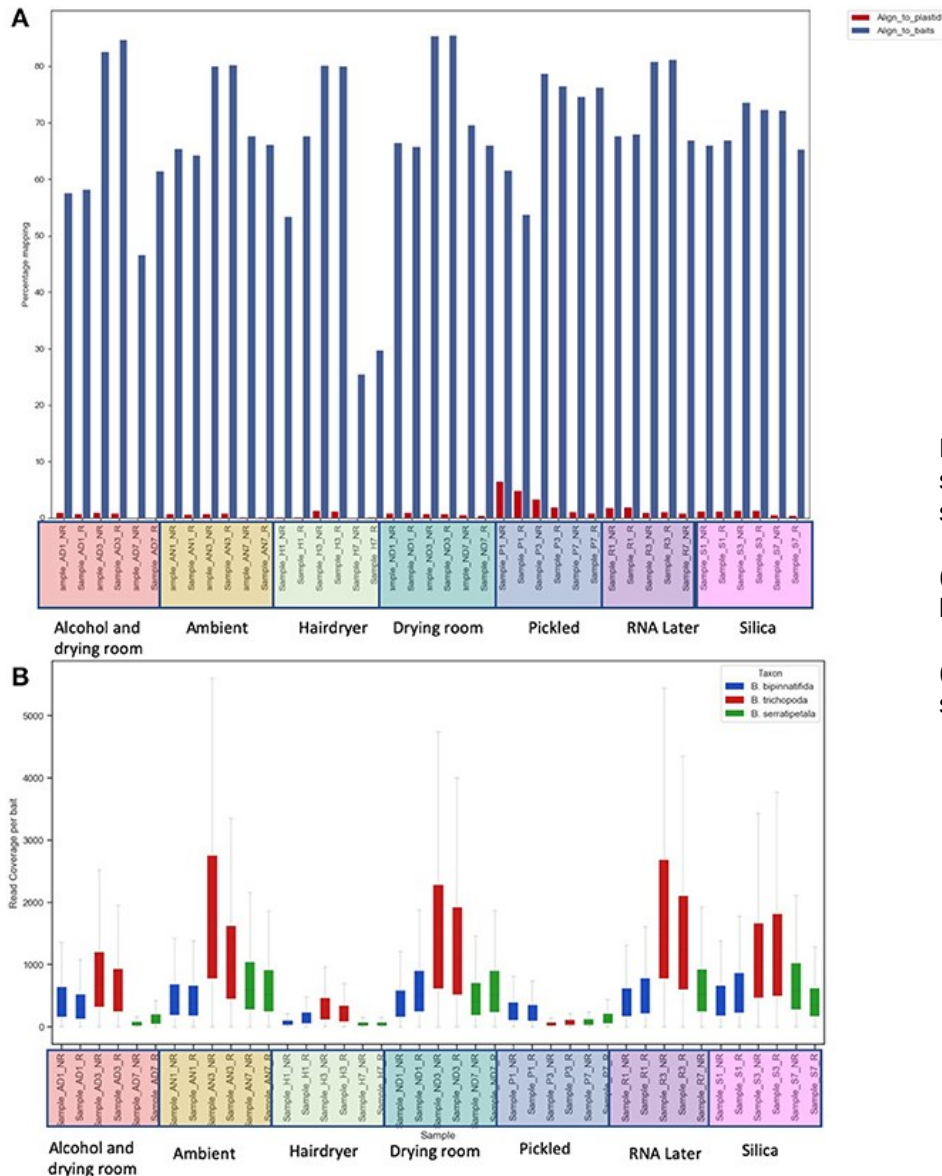


Figure 4. The effect of treatment on recovery of bait sequence, coverage of baits and length of consensus sequence called.

(A) Effect of treatment on percentage of reads mapping to baits (blue) and to plasmid sequence (red) using BWA.

(B) Read coverage per bait by treatment (x-axis) and by species (color of bar).

SNP quality using differently preserved material

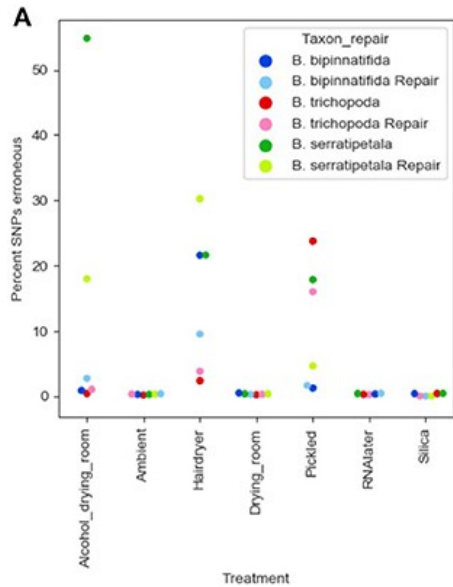
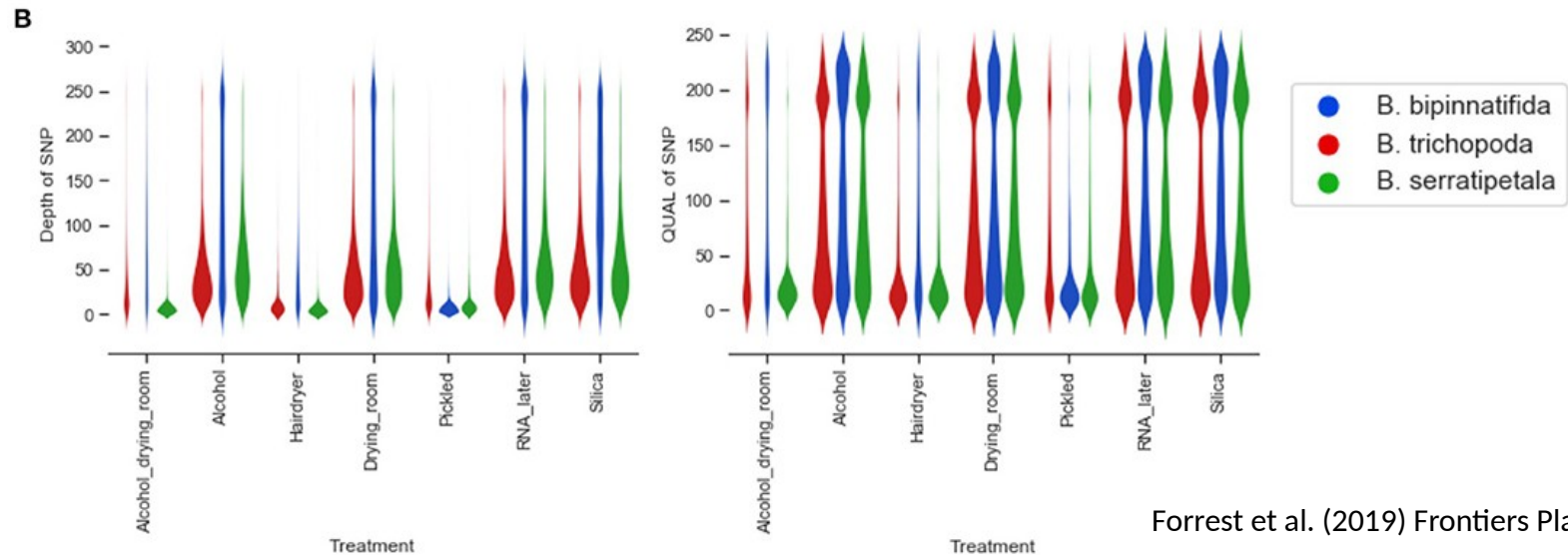


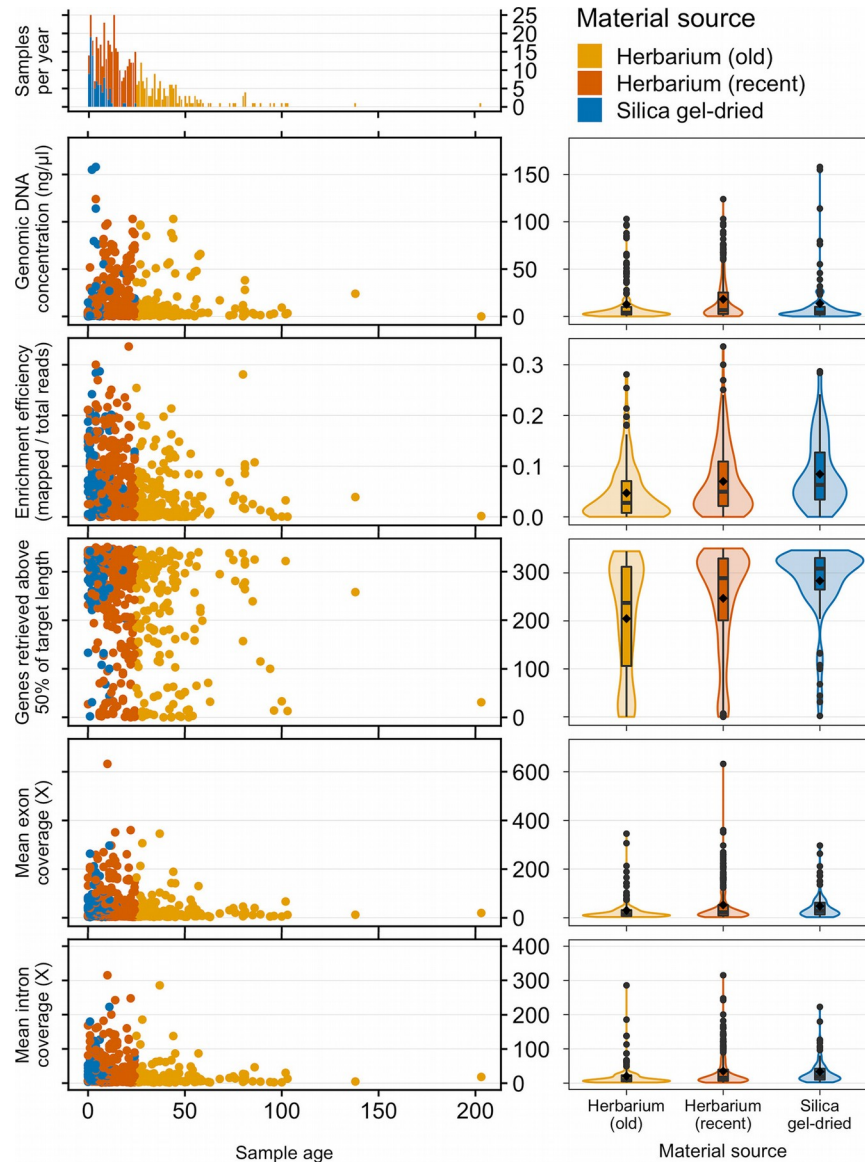
Figure 7. Erroneous SNPs by species and treatment.

(A) Percentage of erroneous SNPs by treatment (x axis) and by species (marker color).

(B) Depth and quality of all SNPs called by treatment (x axis) and by species (plot color).



Silica-dried material is the best!!



RADseq vs. Hyb-Seq

For which phylogenetic depth is Hyb-Seq optimal?

A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs)

Brant C. Faircloth¹, Laurie Sorenson, Francesco Santini, Michael E. Alfaro^{1*}

Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, California, United States of America

Abstract

Ray-finned fishes constitute the dominant radiation of vertebrates with over 32,000 species. Although molecular phylogenetics has begun to disentangle major evolutionary relationships no widely available approach for efficiently collecting phylogenomic data potential of massively parallel sequencing technologies for resolving major we provide a genomic perspective on longstanding questions regarding fishes through targeted enrichment of ultraconserved nuclear DNA elements workflow efficiently and economically generates data sets that are orders traditional approaches and is well-suited to working with museum specimens supported phylogeny at both shallow and deep time-scales that support *Lepisosteus* (Holostei) and reveals elopomorphs and then osteoglossomorphs. Our approach additionally reveals that sequence capture of UCE regions potential for resolving phylogenetic relationships within ray-finned fishes

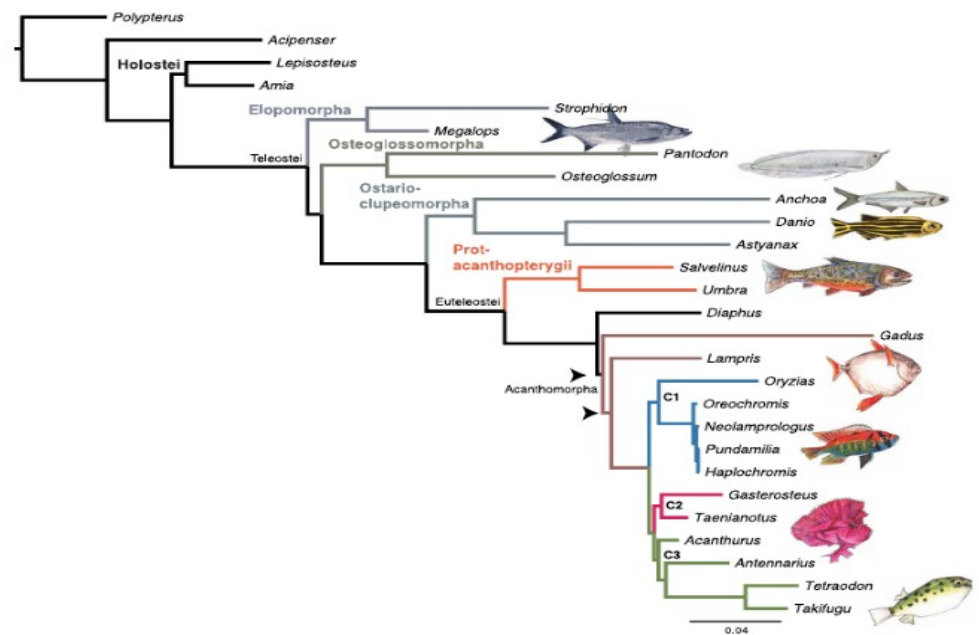


Figure 1. Maximum likelihood phylogram of ray-finned fish relationships based upon UCE sequences. All nodes except for two (indicated by arrows) supported by bootstrap proportions and Bayesian posterior probabilities >0.99. Our analysis supports a monophyletic Holostei and reveals the elopomorphs to be the earliest diverging lineage of teleosts. C1, C2, and C3 indicate clades within acanthomorphs consistent with other recent molecular studies (see Discussion). doi:10.1371/journal.pone.0065923.g001

For which phylogenetic depth is Hyb-Seq optimal?

Syst. Biol. 63(1):83–95, 2014
 © The Author(s) 2013. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
 For Permissions, please email: journals.permissions@oup.com
 DOI:10.1093/sysbio/syt062
 Advance Access publication September 10, 2013

Target Capture and Massively Parallel Sequencing of Ultraconserved Elements for Comparative Studies at Shallow Evolutionary Time Scales

BRIAN TILSTON SMITH^{1,*}, MICHAEL G. HARVEY^{1,2}, BRANT C. FAIRCLOTH³, TRAVIS C. GLENN⁴,
 AND ROBB T. BRUMFIELD^{1,2}

¹Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA; ²Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA; ³Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA; and ⁴Department of Environmental Health Science, University of Georgia, Athens, GA 30602, USA

*Correspondence to be sent to: 119 Foster Hall, Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA; E-mail: briantilstonsmith@gmail.com.

Brian Tilston Smith and Michael G. Harvey contributed equally to this article.

Received 7 January 2013; reviews returned 1 April 2013; accepted 30 August 2013

Associate Editor: Michael Charleston

Abstract.—Comparative genetic studies of non-model organisms are transforming rapidly due to major advances in sequencing technology. A limiting factor in these studies has been the identification and screening of orthologous loci across an evolutionarily distant set of taxa. Here, we evaluate the efficacy of genomic markers targeting ultraconserved DNA elements (UCEs) for analyses at shallow evolutionary timescales. Using sequence capture and massively parallel sequencing to generate UCE data for five co-distributed Neotropical rainforest bird species, we recovered 776–1516 UCE loci across the five species. Across species, 53–77% of the loci were polymorphic, containing between 2.0 and 3.2 variable sites per polymorphic locus, on average. We performed species tree construction, coalescent modeling, and species delimitation, and we found that the five co-distributed species exhibited discordant phylogeographic histories. We also found that species trees and divergence times estimated from UCEs were similar to the parameters obtained from mtDNA. The species that inhabit the understory had older divergence times across barriers, contained a higher number of cryptic species, and exhibited larger effective population sizes relative to the species inhabiting the canopy. Because orthologous UCEs can be obtained from a wide array of taxa, are polymorphic at shallow evolutionary timescales, and can be generated rapidly at low cost, they are an effective genetic marker for studies investigating evolutionary patterns and processes at shallow timescales. [Birds; coalescent theory; isolation-with-migration; massively parallel sequencing; Neotropics; next-generation sequencing; phylogeography; SNPs.]



FIGURE 1. Map of the areas of endemism for lowland Neotropical birds that we used to define populations for this study. For *C. lineatus*, *X. minutus*, and *Q. purpurata*, Inambari (SA1) and Rondônia (SA2) were collapsed as SA.

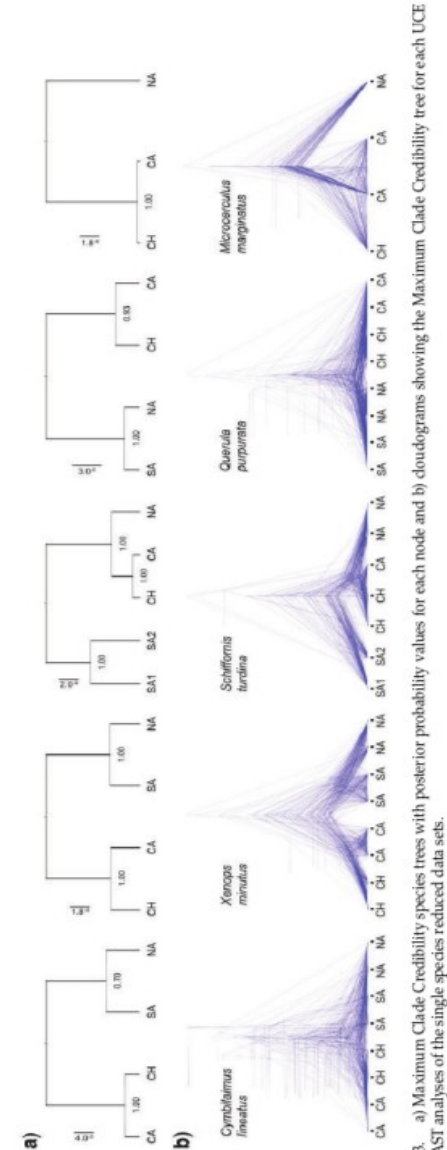
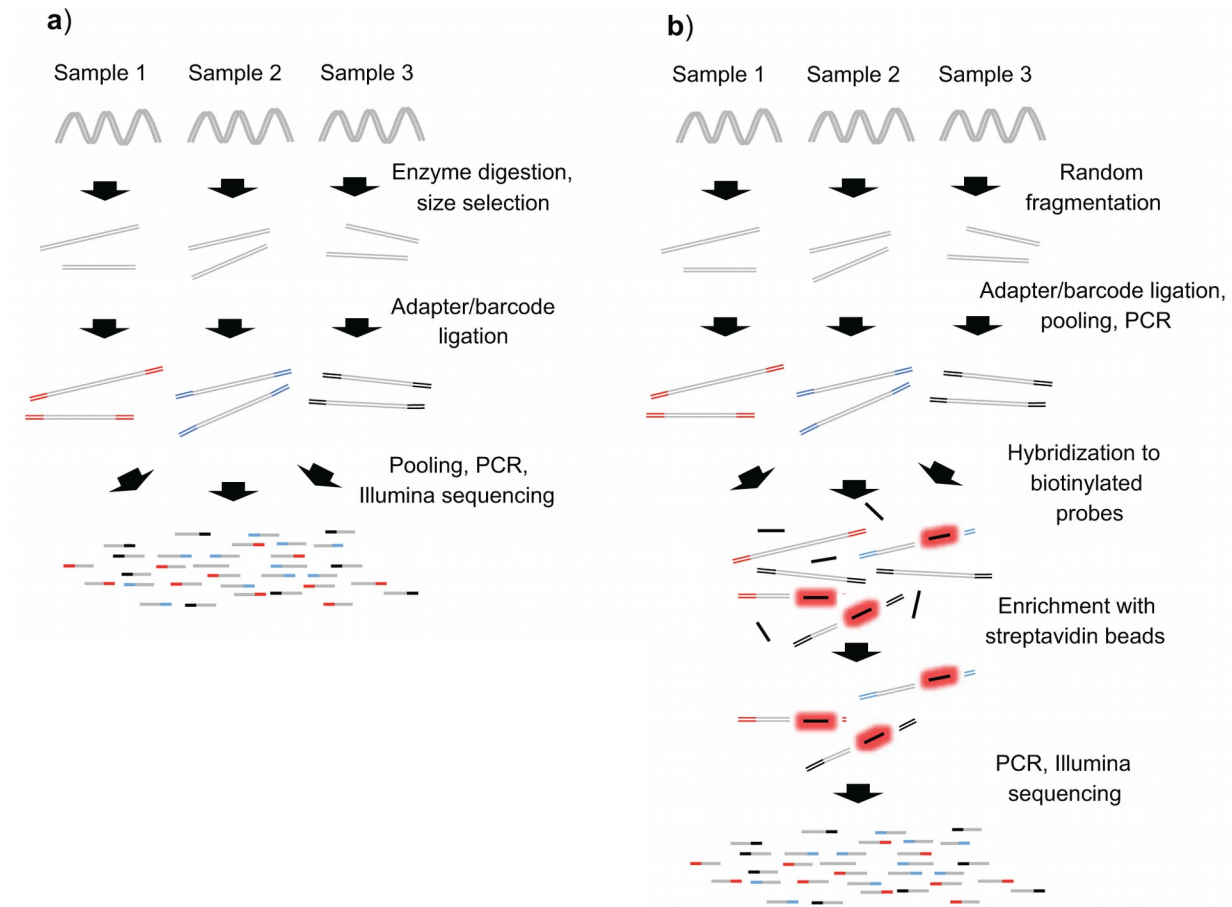


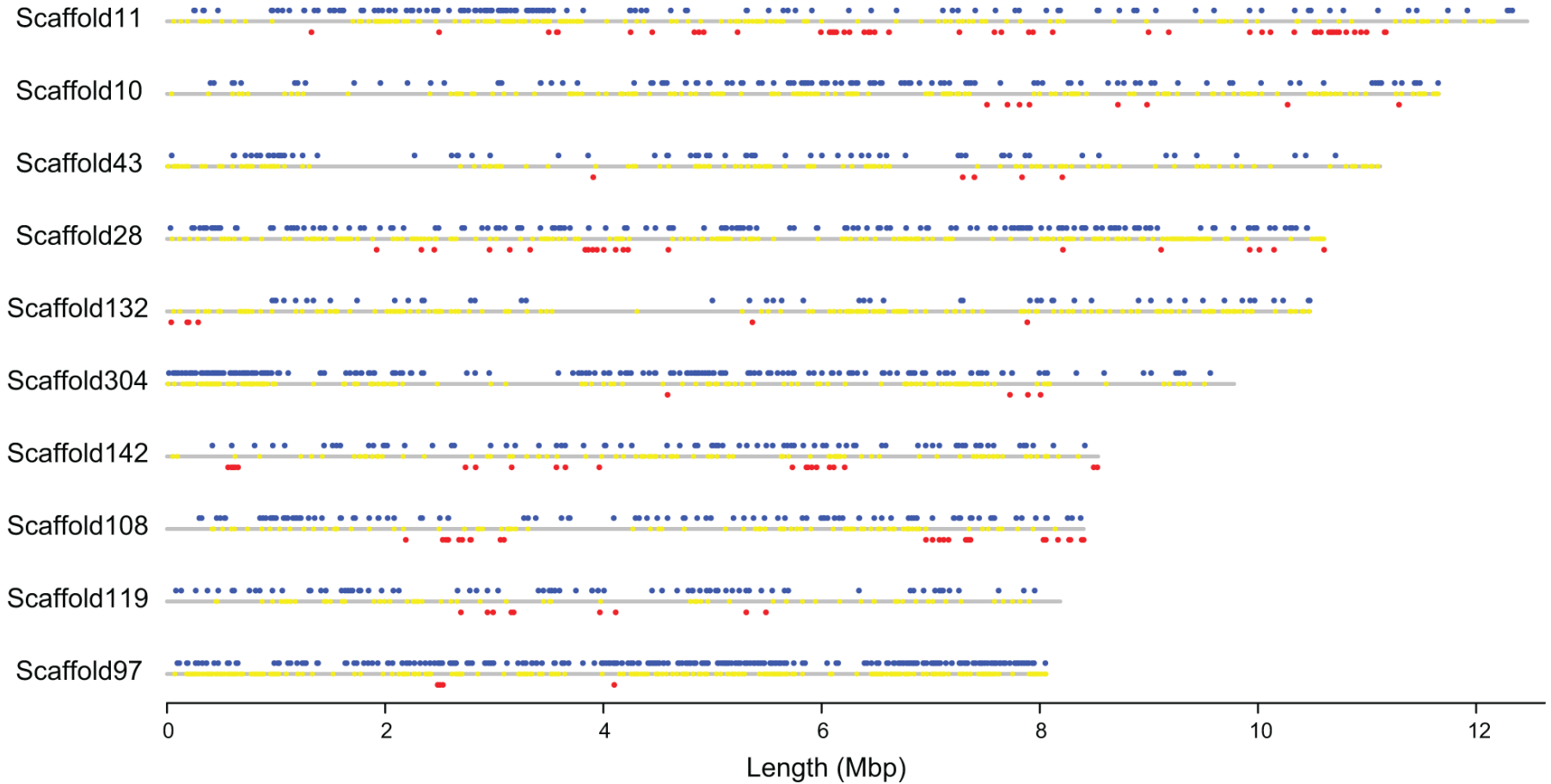
FIGURE 3. a) Maximum Clade Credibility trees with posterior probability values for each node and b) cloudograms showing the Maximum Clade Credibility tree for each UCE from *BEAST analyses of the single species reduced data sets.

RADseq or Hyb-Seq for phylogenies/phylogeographies?

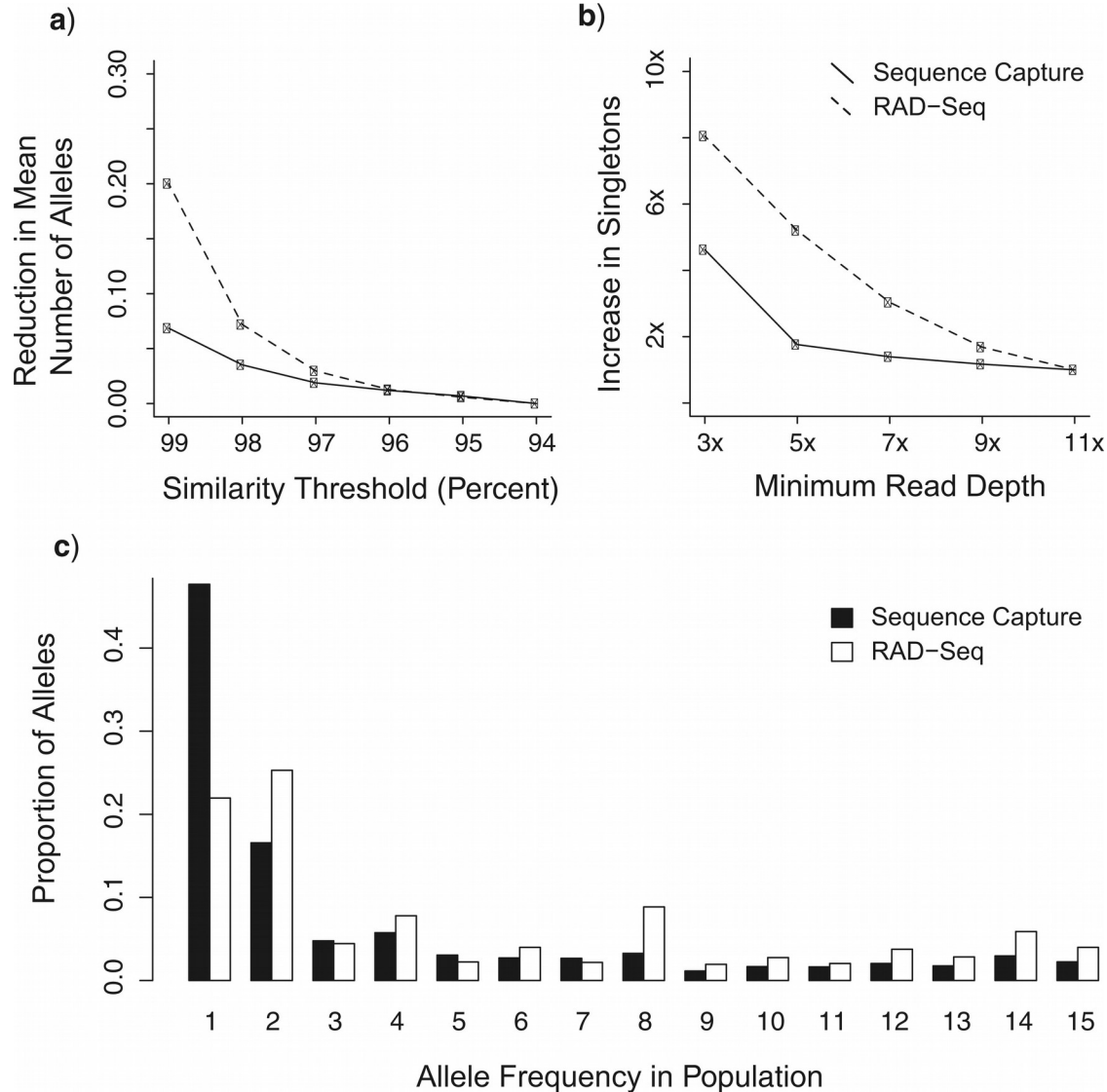


Harvey et al. (2016) Syst. Biol.

Genomic distributions of RADseq loci (blue dots above the line) and UCEs (red dots below)



Reduction of alleles in target capture and RADseq datasets when using stringent sequence similarity



Pros and cons: RADseq or Hyb-Seq for phylogenies/phylogeographies?

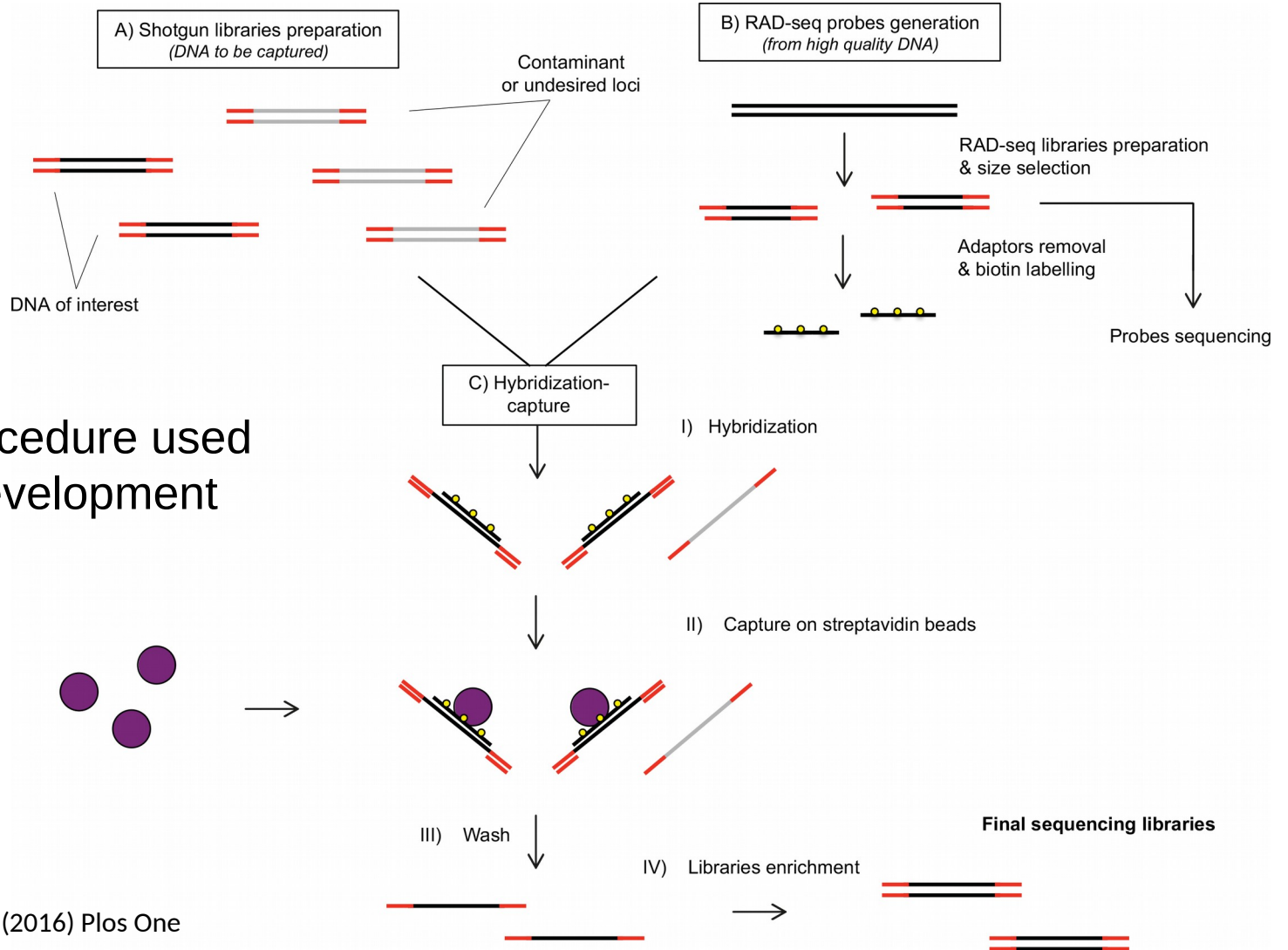
Table 2.

Pros, Cons, and Applications of RAD-Seq and Sequence Capture Datasets.

Category	RAD-Seq	Sequence capture
Marker distribution and genomic context	Pro: Widely dispersed across genome	Pro: Can be tailored using new genomic information
	Con: Anonymous, evolutionary processes largely unknown	Con: Purifying selection impacts allele frequencies
Practical considerations	Pro: Less expensive, faster	Pro: Works with low-quality and highly contaminated samples
Assembly and orthology identification	Pro: Deep coverage, high read overlap	Pro: Over-splitting less problematic
Variant-calling and genotyping	Pro: Fewer rare alleles may make errors easier to distinguish, phasing more straightforward	Pro: Fewer low-coverage rare alleles, no allele dropout
Information content	Pro: More overall information	Pro: More information per locus
Applications	Genome scans, rapid and inexpensive analyses, analyses using species in clades without genomic information, extremely shallow divergences and otherwise intractable relationships.	Comparisons across species, calibrating parameter estimates, targeting loci of known utility or interest, studies using poor-quality samples, studies requiring resolved gene trees, deeper phylogenetic studies.

Harvey et al. (2016) Syst. Biol.

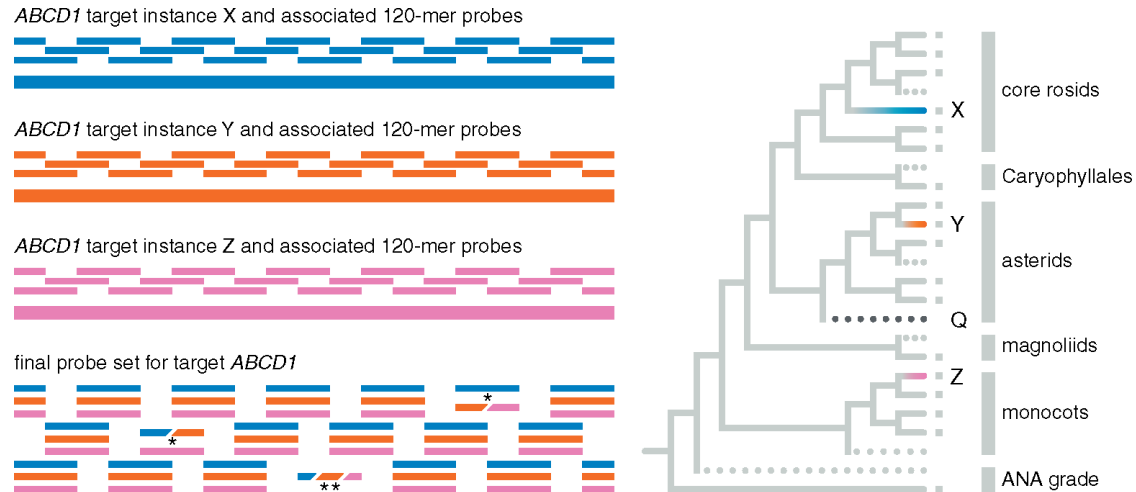
Combining the pros and cons of RADseq and Hyb-Seq: hyRAD



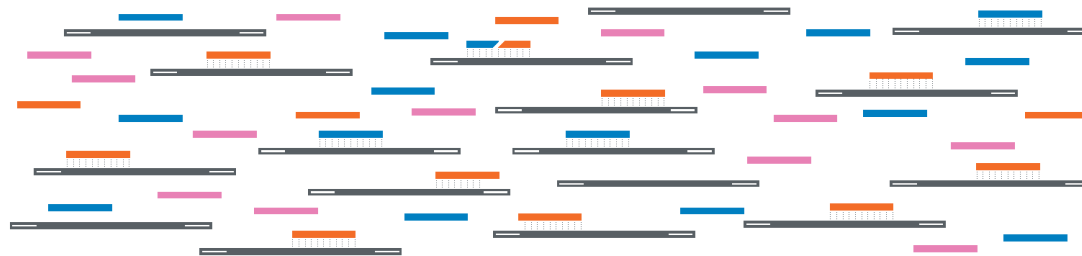
Lab-work procedure used for hyRAD development

Universal vs. group-specific probes

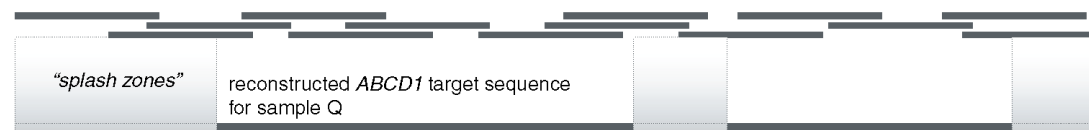
Group-specific versus universal probe sets: a universal angiosperm probe set



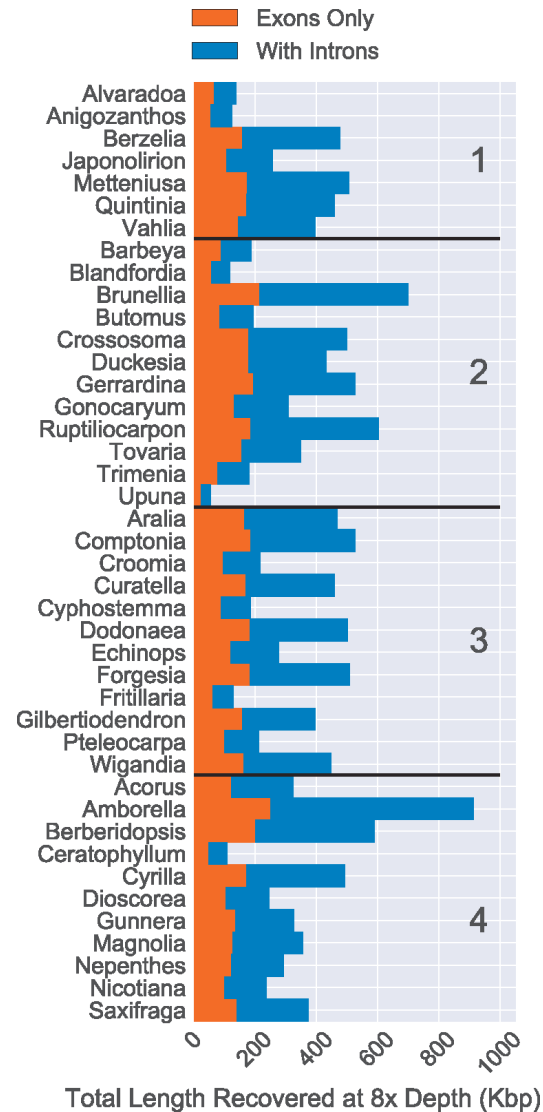
library for sample Q in liquid phase hybridization with final probe set



reads from fragments that hybridized to probes for *ABCD1* from sample Q



Sequence recovery for coding and non-coding regions across 353 loci for 42 angiosperms



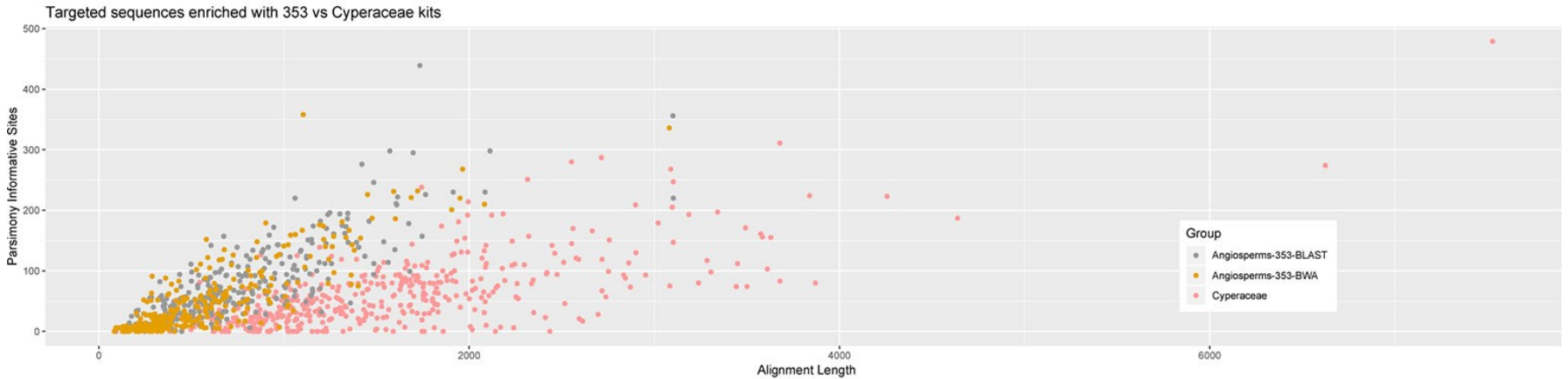
Johnson et al. (2019) Syst. Biol.

Length and informativeness of angiosperm vs. group-specific (here Cyperaceae) probes

		Contig	PIS
Angiosperms-353 (38 accessions) <i>Dataset 1</i>	Mean	751	75
	SD	438	65
	Min	87	0
	Max	3,103	439
	Total	233,429	23,217
Cyperaceae-specific (9 accessions) <i>Dataset 2</i>	Mean	1,608	63
	SD	830	59
	Min	93	0
	Max	7,527	479
	Total	683,427	26,630
Angiosperms-353 (subset of 8 accessions) <i>Dataset 3</i>	Mean	717	25
	SD	411	28
	Min	150	0
	Max	2,826	147
	Total	221,564	7,613
Cyperaceae-specific (8 accessions) <i>Dataset 4</i>	Mean	1,471	50
	SD	818	51
	Min	162	0
	Max	7,524	400
	Total	667,945	22,767

Length and informativeness of angiosperm vs. group-specific (here Cyperaceae) probes

A



Scatter plot of aligned contig length versus number of parsimony informative sites.

Larridon et al. (2020) *Frontiers Plant Sci.*

Species tree robustness using various datasets for species tree building

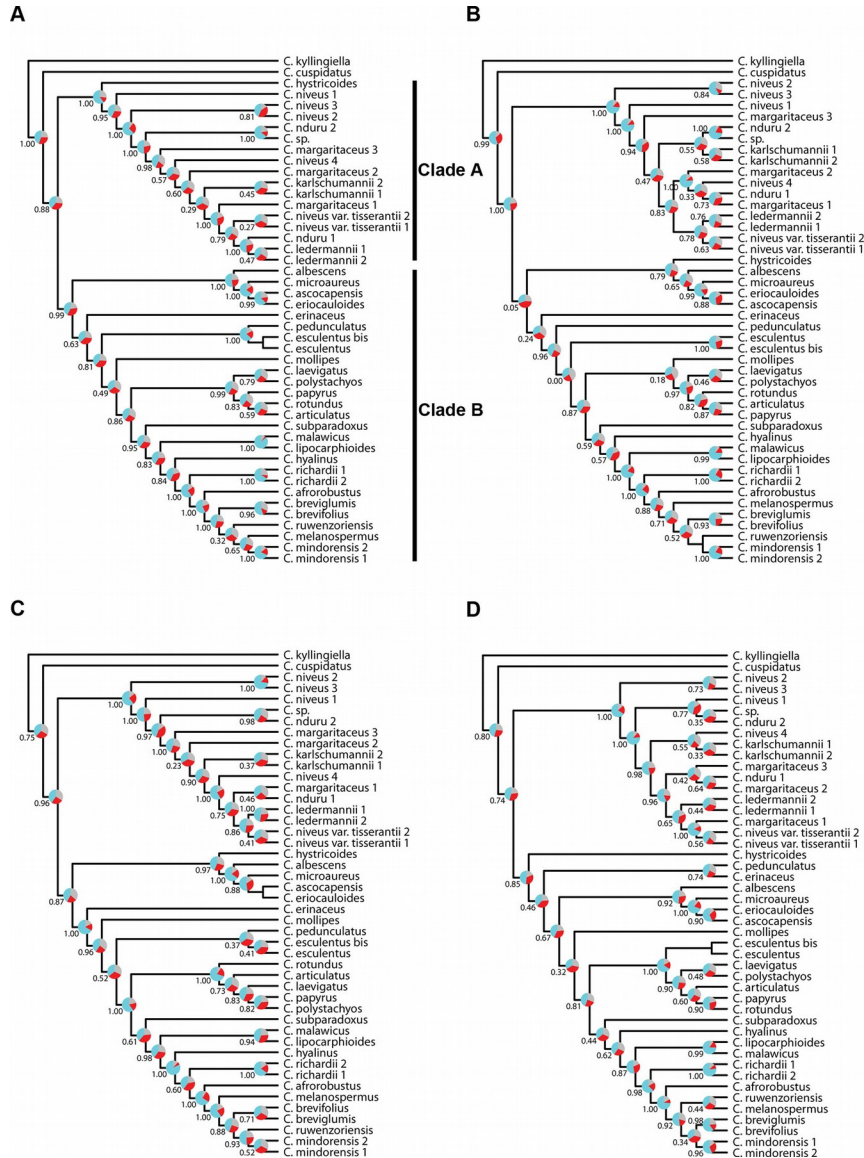


Figure 6 Phylogenetic reconstructions using ASTRAL of the relationships in the C4 Cyperus clade inferred for all accessions from aligned contigs of

- (A) dataset 5, i.e., the loci targeted with the Angiosperms-353 probes,
- (B) dataset 6, i.e., the loci targeted with the Cyperaceae-specific kit,
- (C) dataset 7, i.e., all targeted loci, and
- (D) dataset 8, i.e., the overlapping loci targeted by both kits.

The trees show local posterior probability values and pie charts visualizing quartet support values at the nodes (blue = agreeing loci; red = disagreeing loci; gray = uninformative).

Another examples of group-specific versus universal probes

TABLE 1. Characteristics of target locus sets for probe design.

Locus sets	Total target sequences ^a	Total target loci ^b	Total target length (bp)	Average target sequence length (bp)	Average target locus length (bp)
Taxon-specific	1880	708	580,437	309	820
General	1034	344	431,226	419	1260
COSII	280	67	74,988	268	1119
APVO SSC	572	162	174,848	306	1079
PPR	174	112	179,755	1033	1605
Total	2906	1049	1,010,028	348	963

^aA target sequence is a single consensus sequence from reads mapped to a target locus. There may be multiple target sequences for a target locus if sequences do not overlap.

^bA target locus is from a single transcript (taxon-specific approach) or a single gene (general approach).

TABLE 2. Average performance of locus sets in assembly for 44 *Buddleja* samples (excludes two *Buddleja* samples with failed sequencing) and four outgroup samples.^a

Locus sets	<i>Buddleja</i>		Outgroup	
	No. of sequences	Total length	No. of sequences	Total length
Taxon-specific	1845 (98%) ^A	567,161 (98%) ^F	992 (53%)	287,344 (50%)
General	984 (95%)	421,307 (98%)	733 (71%)	312,062 (72%)
COSII	264 (94%) ^C	72,207 (96%) ^F	184 (66%)	48,390 (65%)
APVO SSC	549 (96%) ^B	170,224 (97%) ^F	425 (74%)	130,148 (74%)
PPR	171 (98%) ^A	178,876 (100%) ^X	124 (71%)	133,524 (74%)
Total	2829 (97%)	988,468 (98%)	1724 (59%)	599,405 (59%)

^aShown are the average number of target sequences with assembled coding sequence and the average total length of assembled coding sequences. In parentheses are the percentages of total target sequences used for probe design. Superscript letters show significant differences in averages at the 0.05 level among locus sets for *Buddleja* samples from Tukey multiple comparison tests with blocking by sample.

TABLE 3. Characteristics of assembled sequence data sets used for phylogenetic analyses.^a

Locus sets	Total sequences ^b	Total loci	Average sequence length (bp)	Average locus length (bp)	Average total length: unaligned (bp)	Total length: aligned, trimmed (bp)	Average % variable sites
Taxon-specific	800 (43%)	511 (72%)	336	526	268,710 (46%)	268,603	36.07% ^A
General	400 (39%)	261 (76%)	605	928	242,161 (56%)	242,359	30.55%
COSII	82 (29%)	50 (75%)	346	567	28,332 (38%)	28,380	27.96% ^B
APVO SSC	217 (38%)	128 (79%)	429	728	93,194 (53%)	93,253	28.56% ^B
PPR	101 (58%)	83 (74%)	1194	1453	120,635 (67%)	120,726	35.17% ^A
Total	1200 (41%)	772 (74%)	425	661	510,579 (51%)	510,962	34.20%

^aSequences with missing data or paralogous sequences in any sample out of 44 *Buddleja* samples and two outgroups used were removed from data sets. In parentheses are the percentages of total target sequences used for probe design. Superscript letters in the last column show significant differences in averages at the 0.05 level among locus sets from a Tukey multiple comparison test.

^bA sequence is assembled to a single target sequence. There may be multiple target sequences for a target locus if target sequences do not overlap.

A target capture-based method to estimate ploidy from herbarium specimens

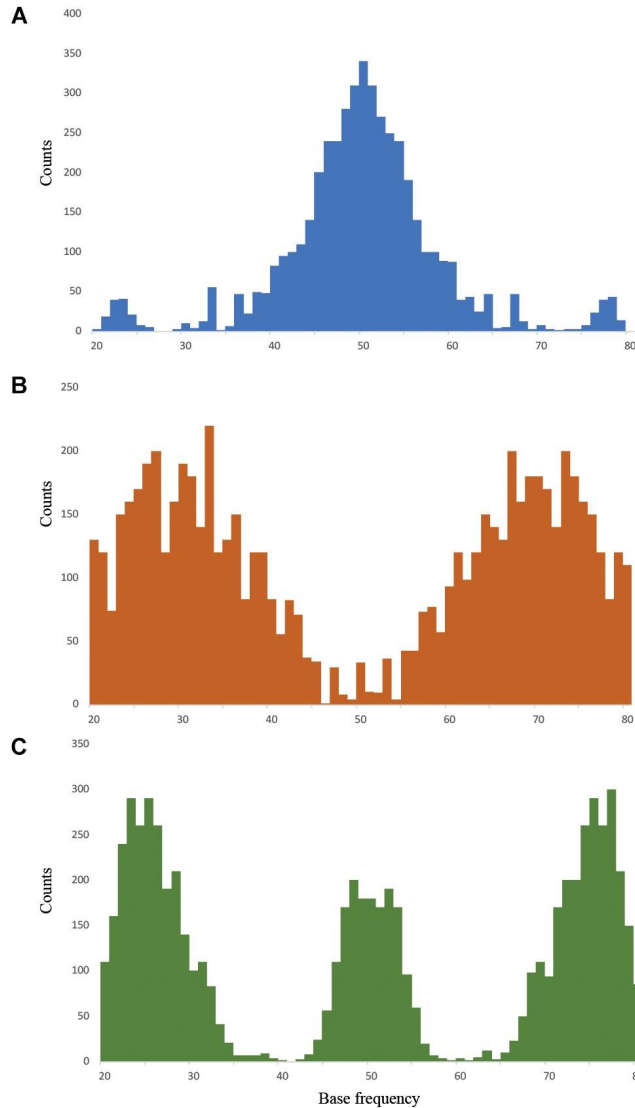


Figure 1. Allelic frequency patterns found in (A) diploid, in blue (*Dioscorea sylvatica* R104), (B) triploid, in orange (*D. alata* T38), and (C) tetraploid, in green (*D. communis* P06), models using nQuire.

Viruel et al. (2019) *Frontiers Plant Sci.*