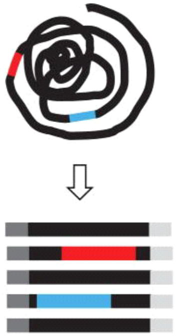
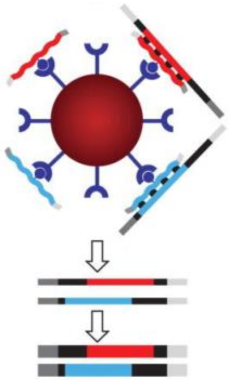


Target enrichment for plant/animal systematics
- methodological workshop -



Introduction

Roswitha Schmickl, Tomáš Fér, Vojtěch Zeisek, Soňa Píšová

Dept. of Botany, Charles University, Prague

9th-17th June 2025

Target enrichment for plant/animal systematics - methodological workshop
9.-17.6.2025

Department of Botany, Charles University, Prague
Roswitha Schmickl, Tomáš Fér, Vojta Zeisek, Soňa Píšová

Monday	Tuesday	Wednesday	Thursday	Friday
9.6.	**10.6.**	**11.6.**	**12.6.**	**13.6.**
Morning **General** introduction **Theory** - library preparation (Rosi, Tomáš) **Theory** - target enrichment principle (Rosi) & discussion	*Morning* **Lab** work demonstration - library preparation, size selection, gel, barcoding… (Soňa) **Paper** presentation & discussion (custom vs. universal probes)	*Morning* **Theory** - approaches for data analysis (Rosi) **Theory** - target enrichment data structure (Vojta) **Computer** work - data cleaning, gene alignments with HybPiper (Vojta)	*Morning* **Theory** - gene trees vs species trees (continue) (Tomáš)	*Morning* **Computer** work - HybPhyloMaker - initial steps (cleaning, mapping, alignment, filtering) (Tomáš)
Afternoon **Lab** work demonstration - DNA conc. (Nanodrop, Qubit), Covaris sonication, gel… (Soňa) **Independent/group** work - preparing presentation of Ufimov et al. (2021) paper – custom vs. universal probes	*Afternoon* **Lab** work demonstration (contin.) *&* **Independent** work on paper of choice about target enrichment in plant/animal systematics	*Afternoon* **Theory** - gene trees vs species trees (Tomáš)	*Afternoon* **Computer** work - gene tree, species tree building (Vojta)	*Afternoon* **Computer** work - HybPhyloMaker - species tree methods, discordance, networks (Tomáš) **Theory&discussion** - discordance, networks, hybridization (Rosi, Tomáš)

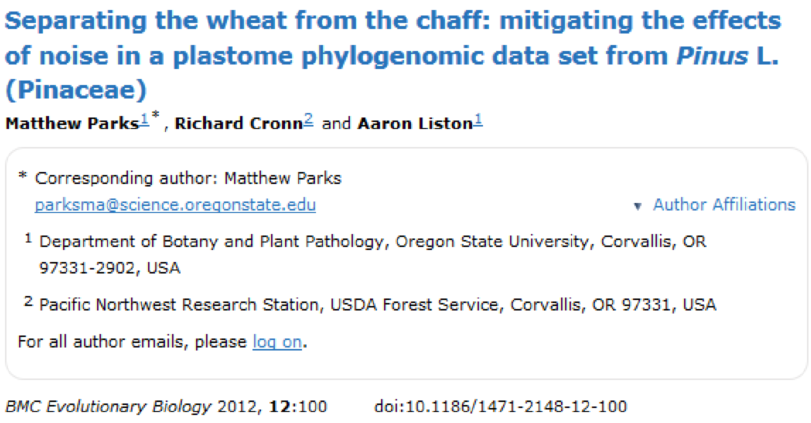
Monday	Tuesday
16.6.	**17.6.**
Morning **Student** presentations - papers of their choice (5mins + 10mins discussion)	*Morning* **Plastome** 'assembly' - HybPhyloMaker, FastPlast (Tomáš)
Afternoon **Group** work/discussion - reading Joyce et al. (2025), discussion of tools&approaches	*Afternoon* **Wrap-up**, varia (Rosi, Tomáš, Vojta) **Hands-on** session with own data etc. (Rosi, Tomáš, Vojta)

What is phylogenomics?

- using whole-genome sequences or large portion of the genome to build a phylogeny
    - Whole organellar (chloroplast/mitochondrial) sequences
    - hundreds or thousands of genes

- gene tree – individual evolutionary history
- species tree – ‘true’ species evolution

Phylogenomics – what is its potential?



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 Cronn**[2] and **Aaron Liston**[1]

* 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

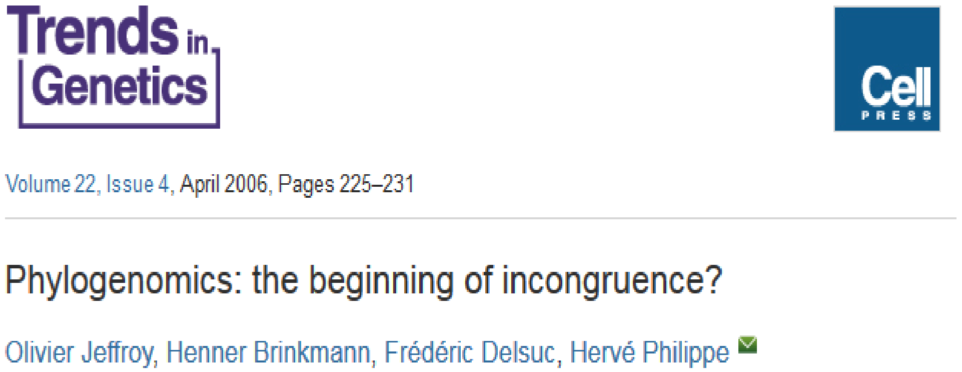
[2] Pacific Northwest Research Station, USDA Forest Service, Corvallis, OR 97331, USA

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



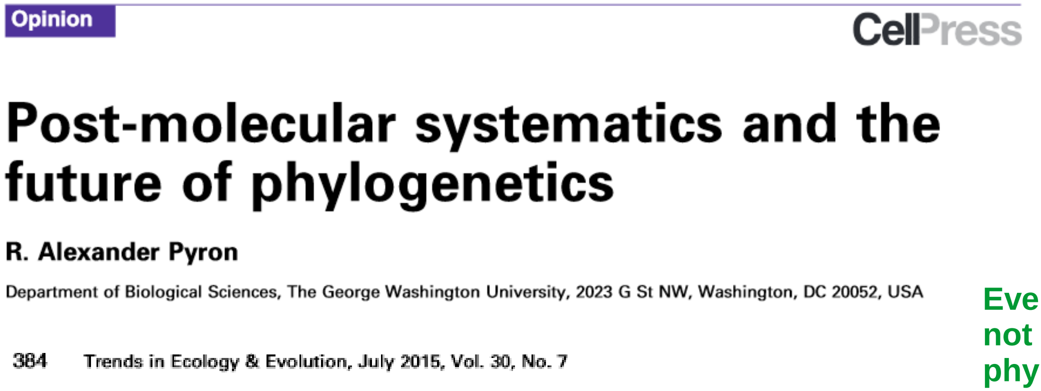
Trends in Genetics

Cell PRESS

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

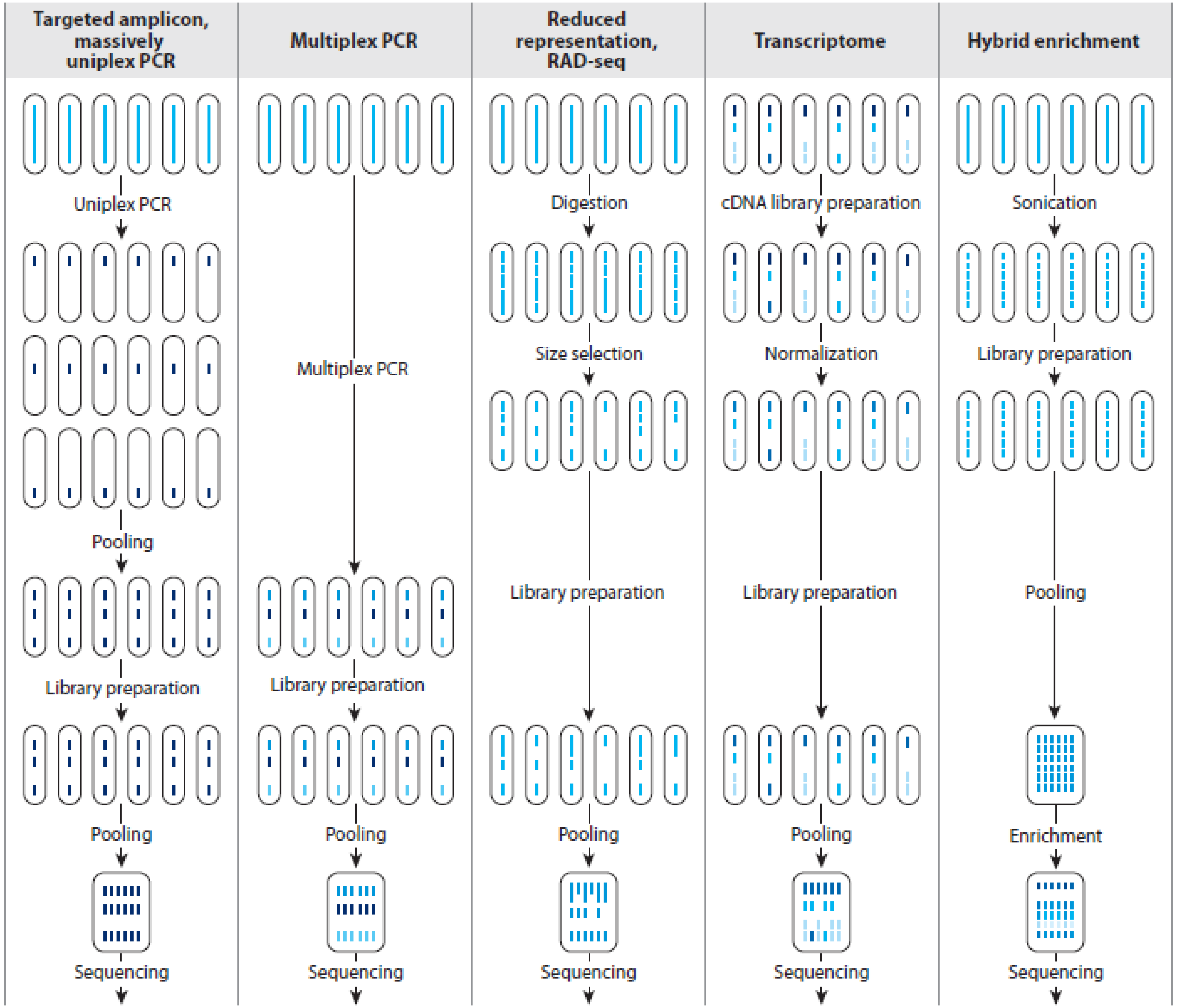
R. Alexander Pyron

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

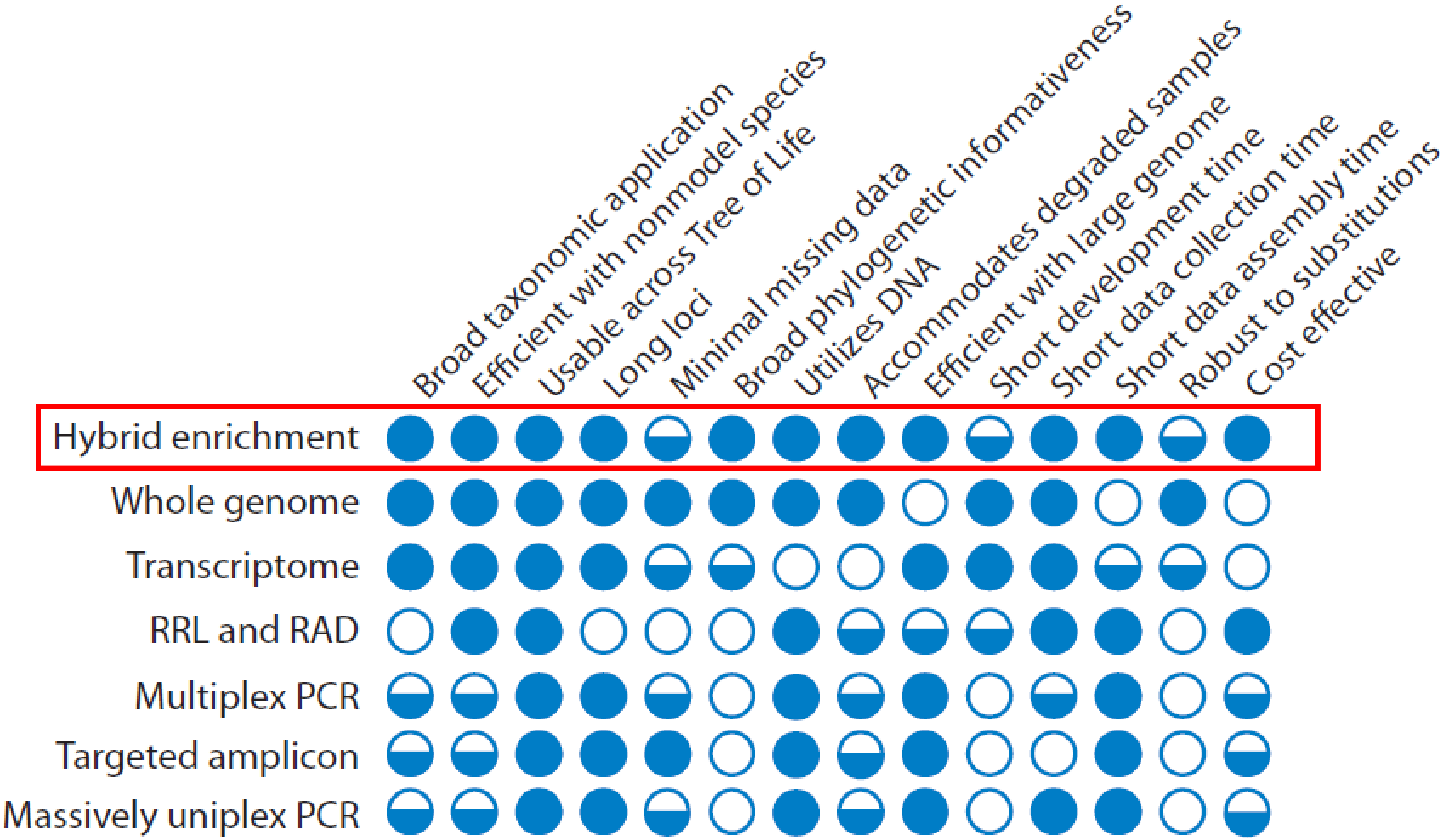
How to generate phylogenomic datasets?



Currently still largely done using Illumina sequencing (short read sequencing), but there is a trend towards PacBio and Oxford Nanopore sequencing (long read sequencing)

Lemmon E.M. & Lemmon A.R. (2013): *High-throughput genomic data in systematics and phylogenetics*. Annu. Rev. Ecol. Evol. Syst, 44, 99–121.

Different phylogenomic approaches

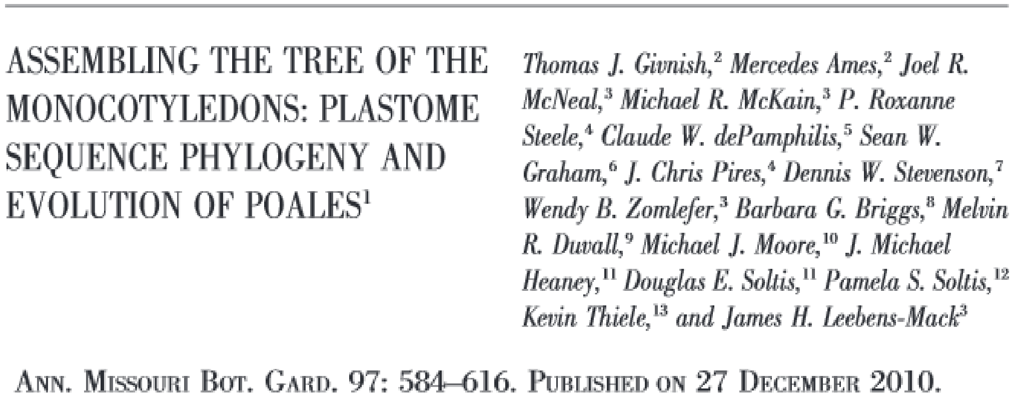


	Broad taxonomic application	Efficient with nonmodel species	Usable across Tree of Life	Long loci	Minimal missing data	Broad phylogenetic informativeness	Utilizes DNA	Accommodates degraded samples	Efficient with large genome	Short development time	Short data collection time	Short data assembly time	Robust to substitutions	Cost effective
Hybrid enrichment	●	●	●	●	◒	●	●	●	●	◒	●	●	◒	●
Whole genome	●	●	●	●	●	●	●	●	○	●	●	○	●	○
Transcriptome	●	●	●	●	◒	◒	○	○	●	●	●	◒	◒	○
RRL and RAD	○	●	●	○	○	○	●	◒	◒	◒	●	●	○	●
Multiplex PCR	◒	◒	●	●	◒	○	●	◒	●	○	◒	●	○	◒
Targeted amplicon	◒	◒	●	●	●	○	●	◒	●	○	○	●	○	◒
Massively uniplex PCR	◒	◒	●	●	◒	○	●	◒	●	○	●	●	○	◒

Lemmon E.M. & Lemmon A.R. (2013): *High-throughput genomic data in systematics and phylogenetics*. Annu. Rev. Ecol. Evol. Syst, 44, 99–121.

Target(ed) enrichment, target capture, hybrid capture, Hyb-Seq

Plant phylogenomics: a historical perspective



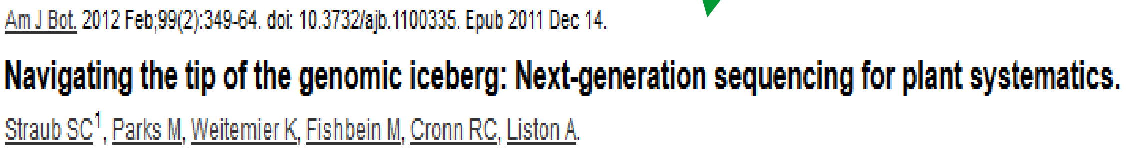
ASSEMBLING THE TREE OF THE MONOCOTYLEDONS: PLASTOME SEQUENCE PHYLOGENY AND EVOLUTION OF POALES[1]

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

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

Plastid genomes

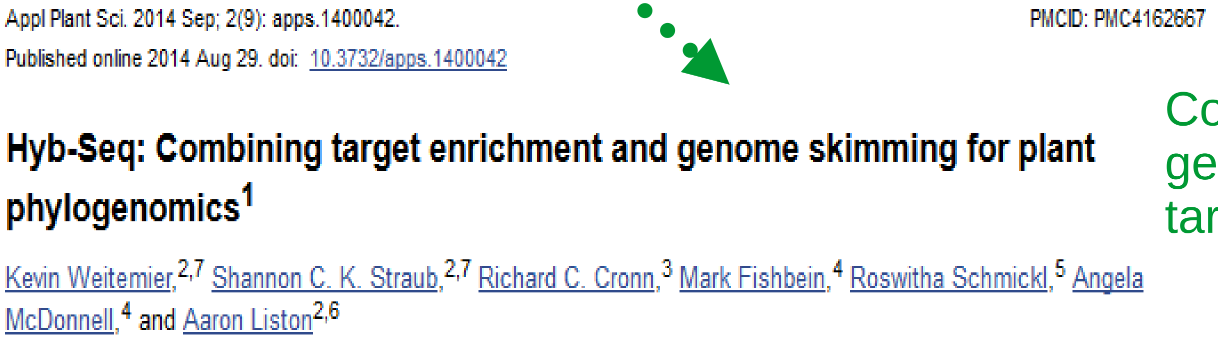
High-copy fractions of genomes (genome skimming)



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. PMCID: PMC4162667

Published online 2014 Aug 29. doi: 10.3732/apps.1400042

Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics[1]

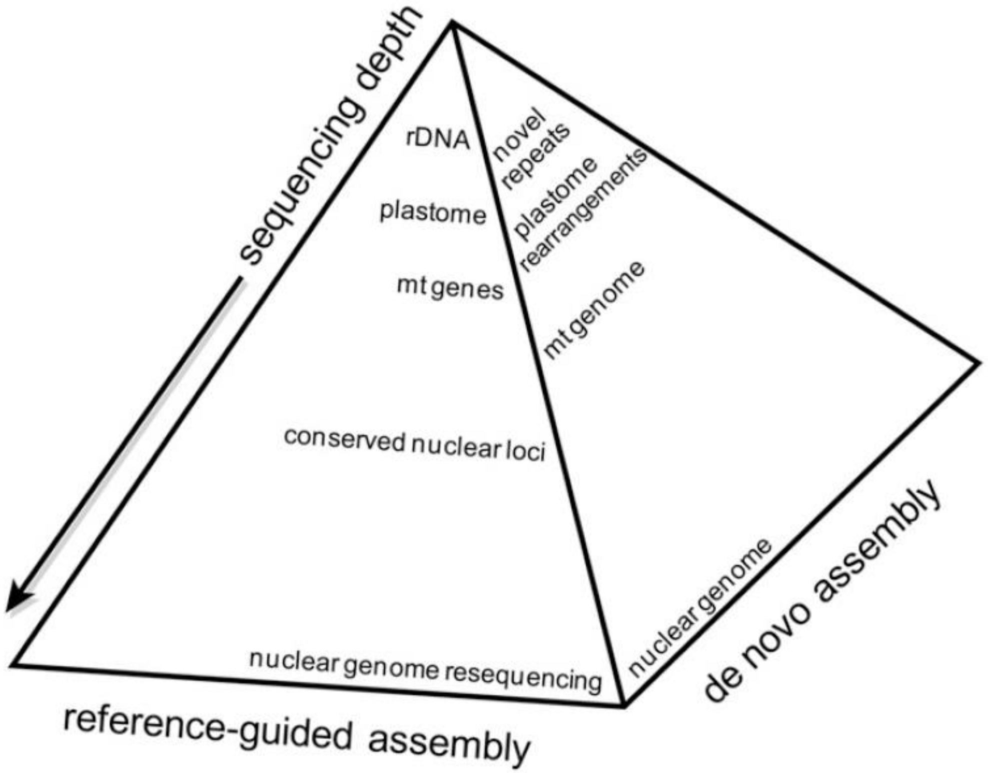
Kevin Weitemier,[2,7] Shannon C. K. Straub,[2,7] Richard C. Cronn,[3] Mark Fishbein,[4] Roswitha Schmickl,[5] Angela McDonnell,[4] and Aaron Liston[2,6]

Combination of genome skimming with target enrichment

?

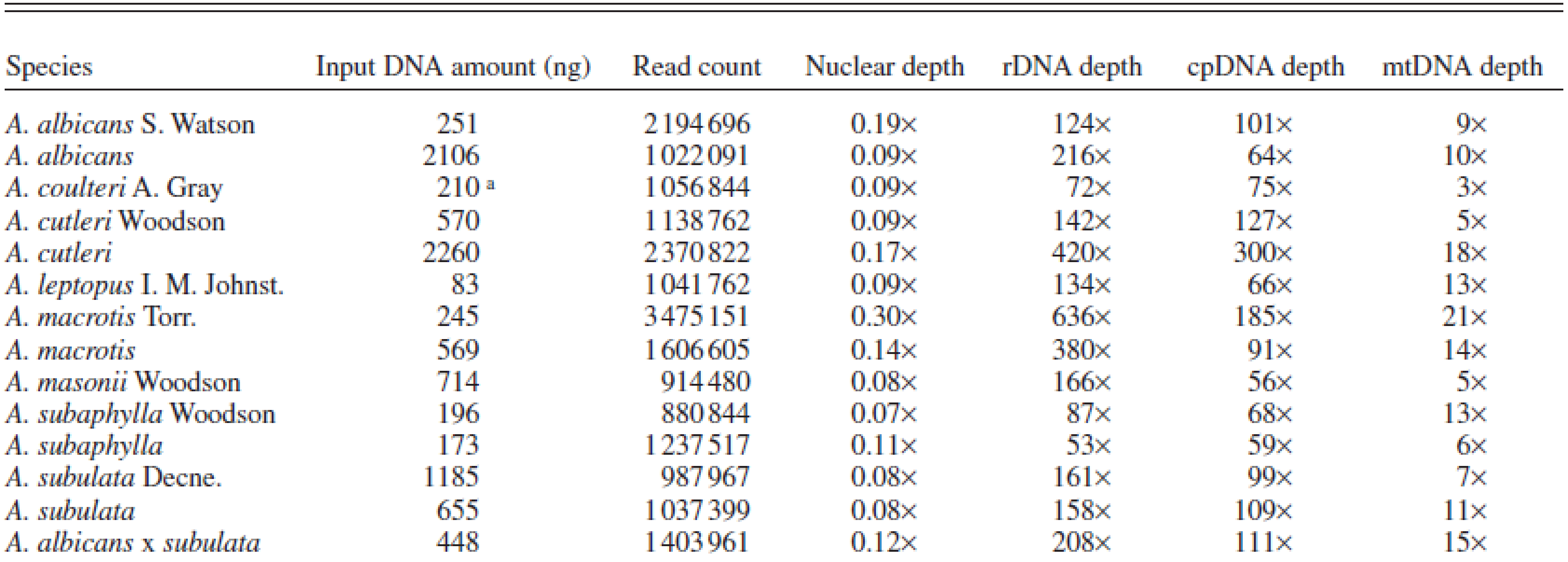
Genome-skimming

- genome sequencing with low total coverage
- we get enough coverage for assembly
  - whole plastome
  - large portions of mtDNA
  - rDNA cistrone
  - many candidate single-copy genes
  - microsatellite regions

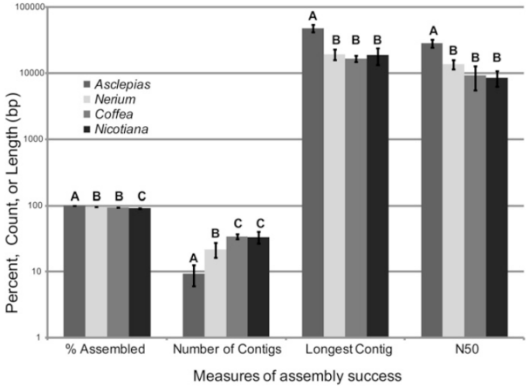


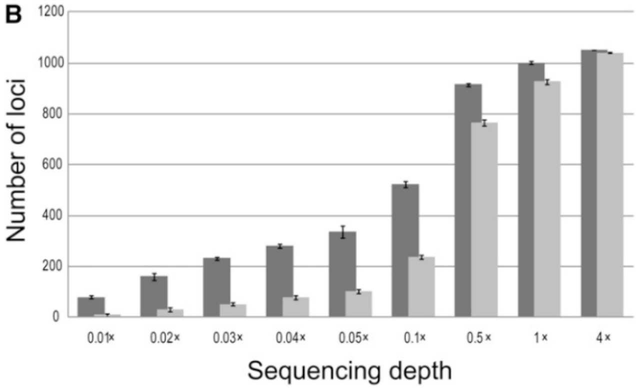
Straub et al. (2012): *Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics*. American Journal of Botany 99: 349–364.
Steel et al. (2012): *Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae*. American Journal of Botany 99: 330-348.

Genome-skimming



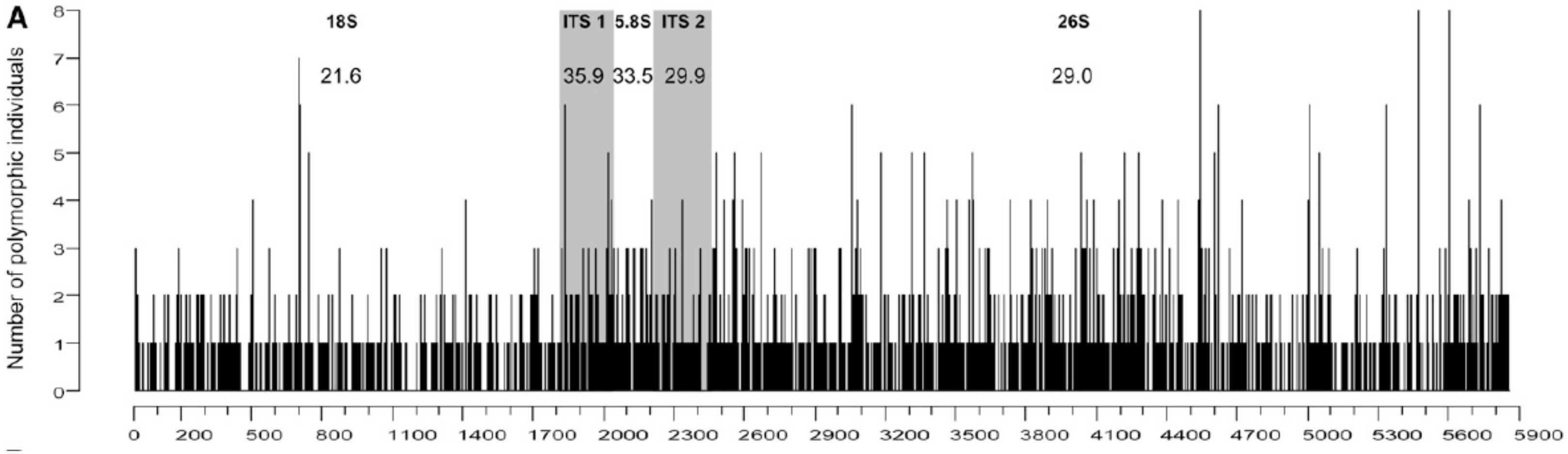
Species	Input DNA amount (ng)	Read count	Nuclear depth	rDNA depth	cpDNA depth	mtDNA depth
A. albicans S. Watson	251	2194696	0.19×	124×	101×	9×
A. albicans	2106	1022091	0.09×	216×	64×	10×
A. coulteri A. Gray	210 [a]	1056844	0.09×	72×	75×	3×
A. cutleri Woodson	570	1138762	0.09×	142×	127×	5×
A. cutleri	2260	2370822	0.17×	420×	300×	18×
A. leptopus I. M. Johnst.	83	1041762	0.09×	134×	66×	13×
A. macrotis Torr.	245	3475151	0.30×	636×	185×	21×
A. macrotis	569	1606605	0.14×	380×	91×	14×
A. masonii Woodson	714	914480	0.08×	166×	56×	5×
A. subaphylla Woodson	196	880844	0.07×	87×	68×	13×
A. subaphylla	173	1237517	0.11×	53×	59×	6×
A. subulata Decne.	1185	987967	0.08×	161×	99×	7×
A. subulata	655	1037399	0.08×	158×	109×	11×
A. albicans x *subulata*	448	1403961	0.12×	208×	111×	15×





Straub et al. (2012): *Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics*. American Journal of Botany 99: 349–364.

Genome-skimming



rDNA cistron

nearly complete cpDNA genom - reference-guided assembly

- distantly related reference (~ 10%) – more than 90%
- conspecific reference – more than 99%

Straub et al. (2012): *Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics*. American Journal of Botany 99: 349–364.

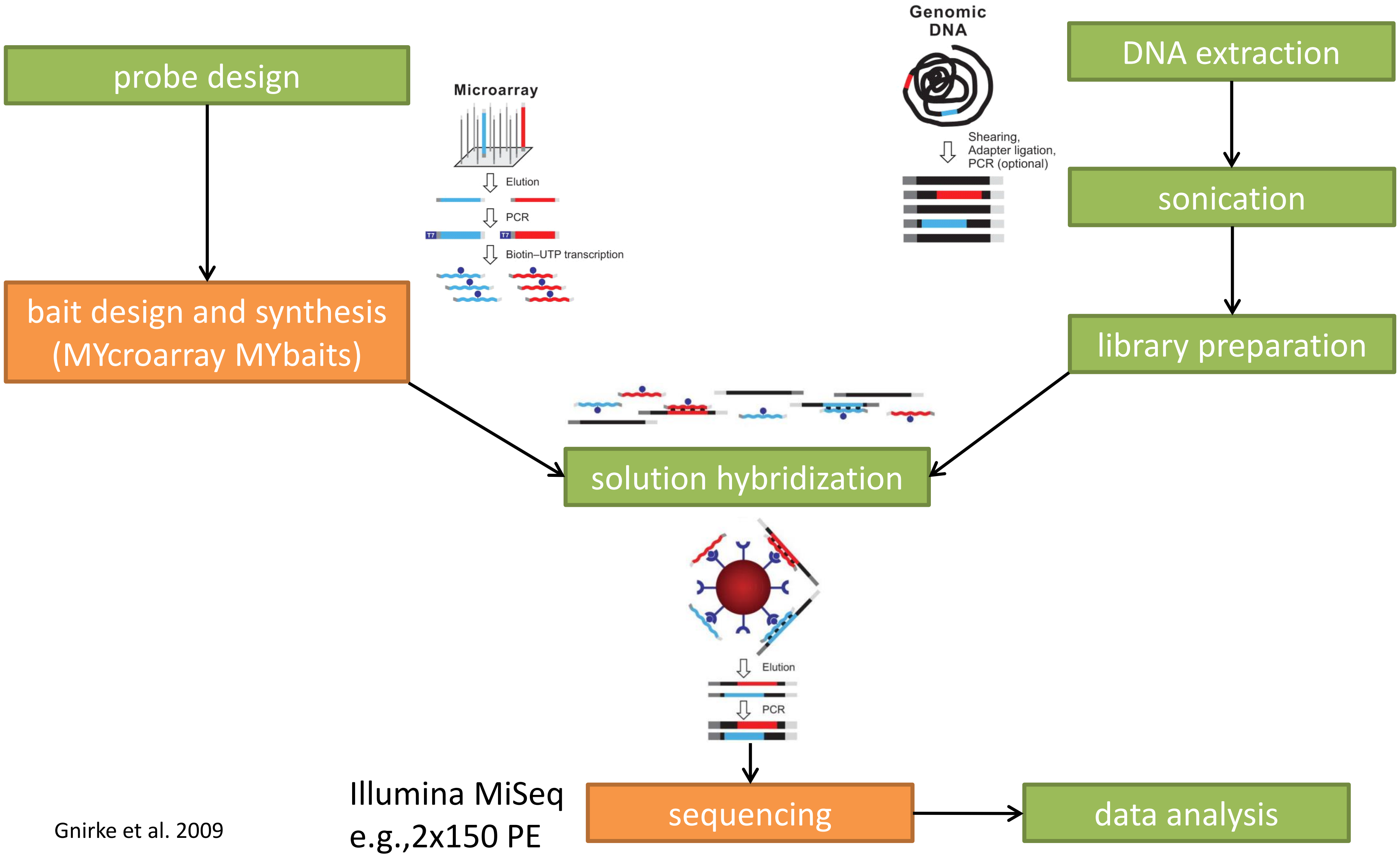
Genome subselection methods

- reduction of the complexity of sequenced parts
- enzyme restriction of the genome
  - sequencing only the part of the genome associated with restriction sites
  - searching for SNPs -> binary data
  - RAD-sequencing
  - GBS (genotyping-by-sequencing)

- Hyb-Seq
  - hybridization based enrichment
  - selection of specific sequences (thousands of exons)

Cronn et al. (2012): *Targeted enrichment strategies for next-generation plant biology*. American Journal of Botany 99: 291-31.

Hyb-Seq overview



probe design

bait design and synthesis (MYcroarray MYbaits)

DNA extraction

sonication

library preparation

solution hybridization

Illumina MiSeq e.g.,2x150 PE

sequencing

data analysis

Gnirke et al. 2009

Target enrichment starts with the choice of the probe set

Probe design:

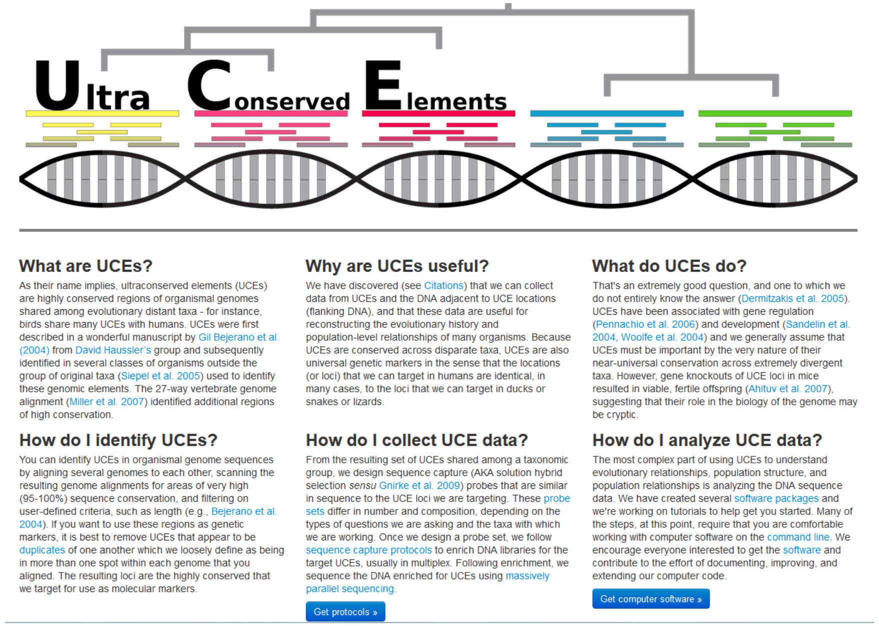
- Exons, low-copy nuclear genes
- Intronic regions less common

Bait synthesis:

- RNA baits
- DNA baits less common

Alternatives (without bait synthesis):

- PCR products (amplicon sequencing)



What are UCEs?

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.

Why are UCEs useful?

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.

What do UCEs do?

That's an extremely good question, and one to which we do not entirely know the answer (Dermitzakis et al. 2005). UCEs have been associated with gene regulation (Pennachio 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.

How do I identify UCEs?

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.

How do I collect UCE data?

From the resulting set of UCEs shared among a taxonomic group, we design sequence capture (AKA solution hybrid selection *sensu* Gnirke 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.

How do I analyze UCE data?

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.

http://www.ultraconserved.org/

Commonly used in animal phylogenomics:

Probe design for target enrichment

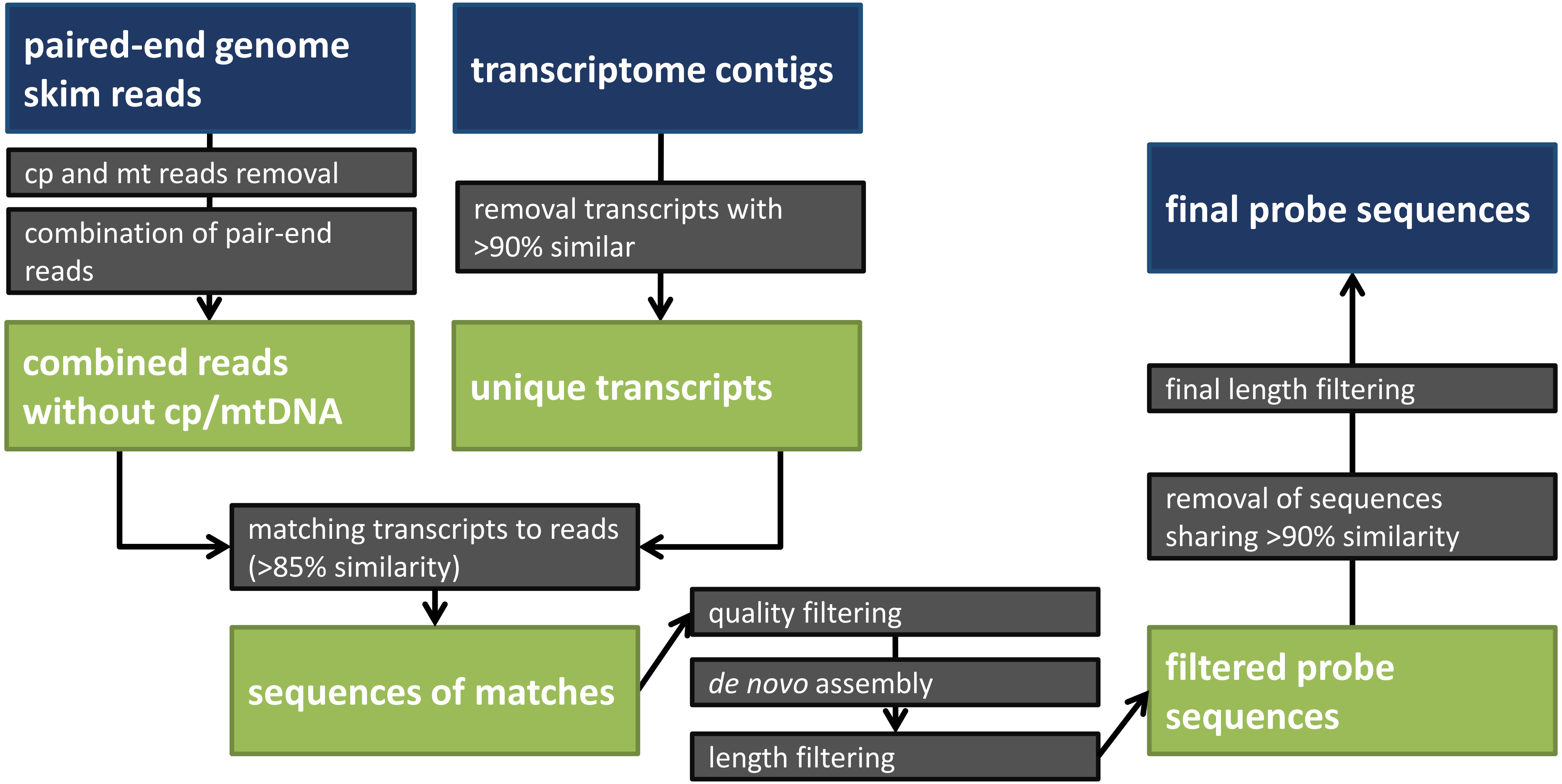
- targets
  - single/low-copy genes, orthologous genes
  - <15% sequence pairwise divergence across the genomes/transcriptomes (*otherwise putative paralogues captured*)
  - >10% divergence when compared genome vs. transcriptome (*otherwise loci with low variabilty captured*)
  - longer genes (i.e., longer than ca. 600 bp) (*otherwise poor gene trees*)

- comparison of
  - transcriptome (from, e.g., oneKP project)
  - genome or genome skimming data (e.g., half of Illumina MiSeq capacity, 2x250 bp)
  - *ability to define exon/intron boundaries*

- result
  - several hundreds of targeted genes
  - several thousands of targeted exons

Probe design for target enrichment

e.g., automatic pipeline – Sondovač (https://github.com/V-Z/sondovac/)



paired-end genome skim reads

cp and mt reads removal

combination of pair-end reads

combined reads without cp/mtDNA

transcriptome contigs

removal transcripts with >90% similar

unique transcripts

matching transcripts to reads (>85% similarity)

sequences of matches

quality filtering

de novo assembly

length filtering

filtered probe sequences

removal of sequences sharing >90% similarity

final length filtering

final probe sequences

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). Molecular Ecology Resources 16, 1124–1135.