Molecular markers in plant systematics and population biology

10. Phylogenomics, HybSeq, genome skimming

Tomáš Fér

tomas.fer@natur.cuni.cz

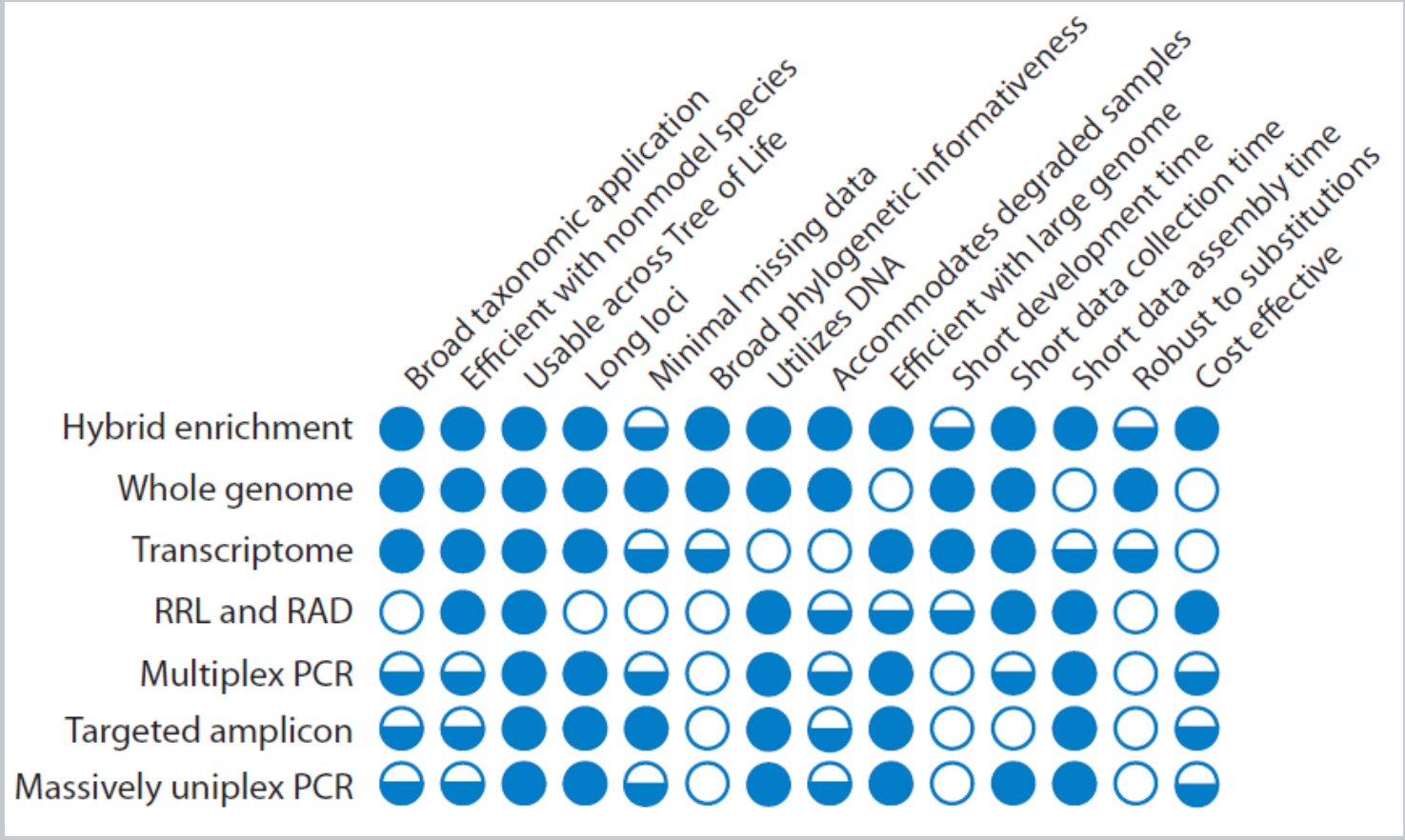
Phylogenomics

- using whole-genome sequences or large portion of the genome to build a phylogeny
  - whole chloroplast sequences
  - hundreds or thousands of genes

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

Phylogenomic data sources

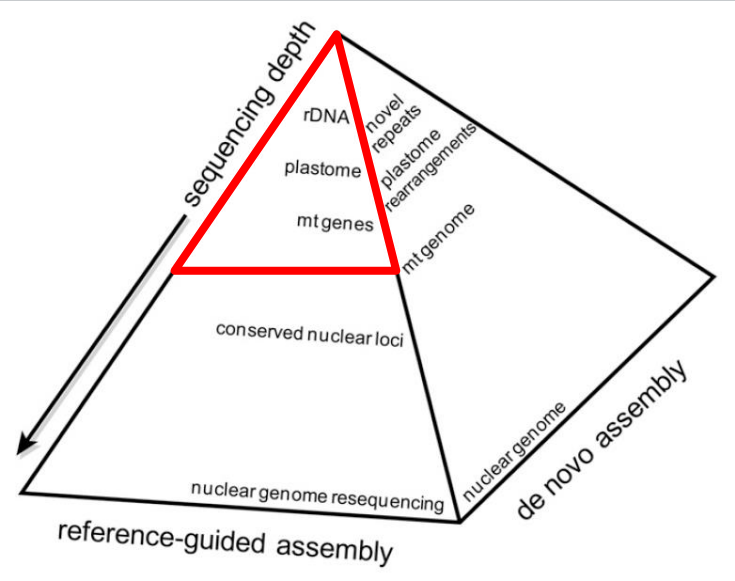
- transcriptomes
- genome skimming
- targeted enrichment



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

Genome-skimming

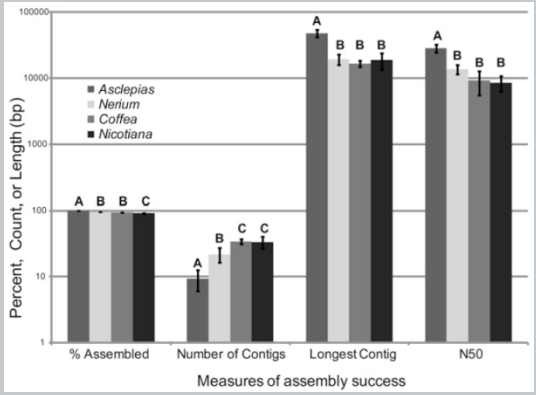
- genome sequencing with low total coverage
- we get enough coverage for assembly
  - whole plastome
  - large portions of mtDNA
  - rDNA cistrone
  - many candidate single-copy genes

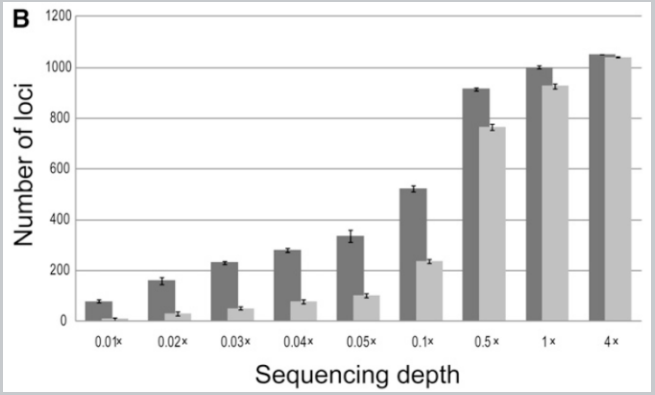


Straub et al. (2012): *Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics*. American Journal of Botany 99: 349–364.
Steel et al. (2012): *Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae*. American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

Genome skimming

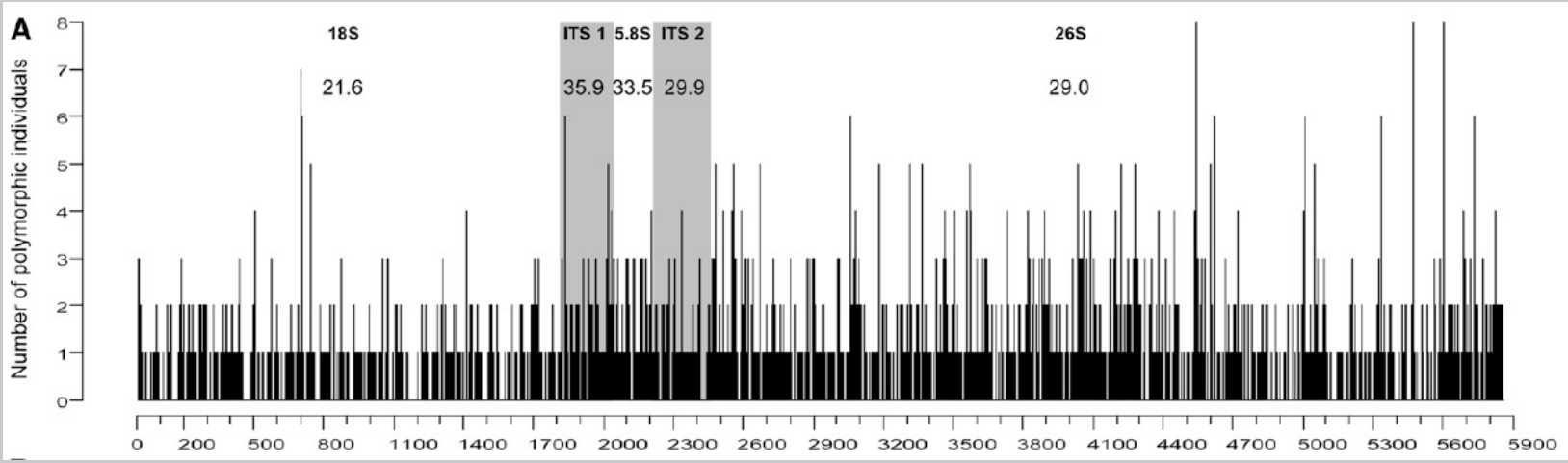
Species	Input DNA amount (ng)	Read count	Nuclear depth	rDNA depth	cpDNA depth	mtDNA depth
A. albicans S. Watson	251	2 194 696	0.19×	124×	101×	9×
A. albicans	2106	1 022 091	0.09×	216×	64×	10×
A. coulteri A. Gray	210 [a]	1 056 844	0.09×	72×	75×	3×
A. cutleri Woodson	570	1 138 762	0.09×	142×	127×	5×
A. cutleri	2260	2 370 822	0.17×	420×	300×	18×
A. leptopus I. M. Johnst.	83	1 041 762	0.09×	134×	66×	13×
A. macrotis Torr.	245	3 475 151	0.30×	636×	185×	21×
A. macrotis	569	1 606 605	0.14×	380×	91×	14×
A. masonii Woodson	714	914 480	0.08×	166×	56×	5×
A. subaphylla Woodson	196	880 844	0.07×	87×	68×	13×
A. subaphylla	173	1 237 517	0.11×	53×	59×	6×
A. subulata Decne.	1185	987 967	0.08×	161×	99×	7×
A. subulata	655	1 037 399	0.08×	158×	109×	11×
A. albicans x *subulata*	448	1 403 961	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 genome – 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.

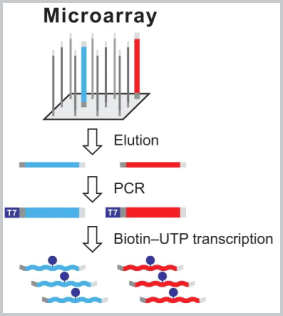
Targeted enrichment

- 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)

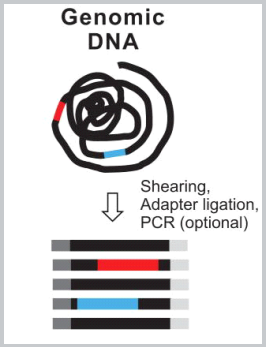
Cronn et al. (2012): *Targeted enrichment strategies for next-generation plant biology*. American Journal of Botany 99: 291-31.

Hyb-Seq overview

probe design



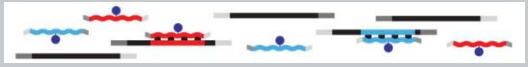
bait design and synthesis
(MYcroarray MYbaits)



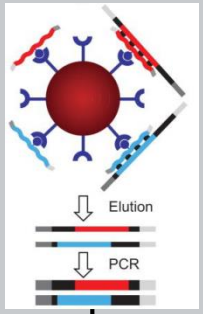
DNA extraction

sonication

library preparation



solution hybridization



Illumina MiSeq
e.g.,2x150 PE

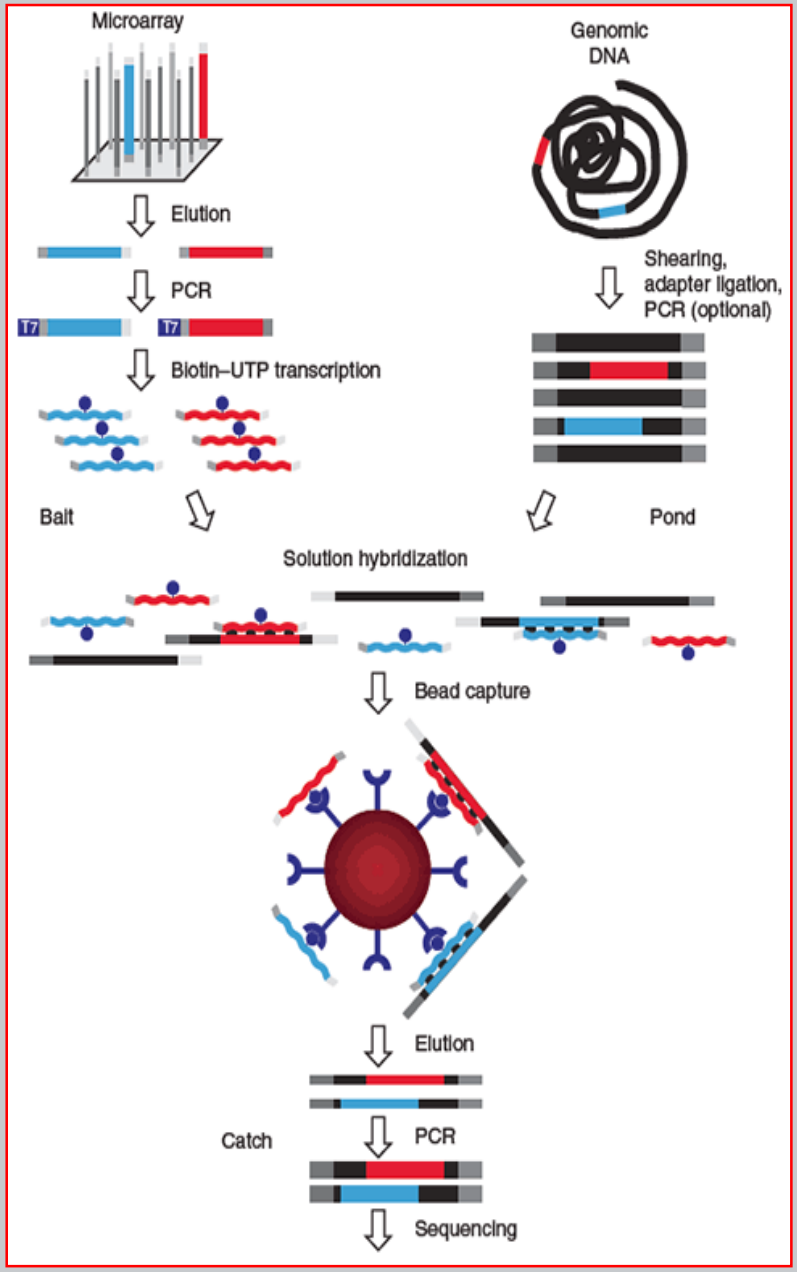
sequencing

data analysis

Gnirke et al. 2009

Hyb-Seq

- solution phase hybridization
- ‘baits’ (short RNA fragments) synthetized on arrays
- hybridization in solution
- immobilization via biotin-streptavidine
  - biotinylated baits
  - streptavidin-coated magnetic beads
- PCR enrichment of target sequences



Mycroarray

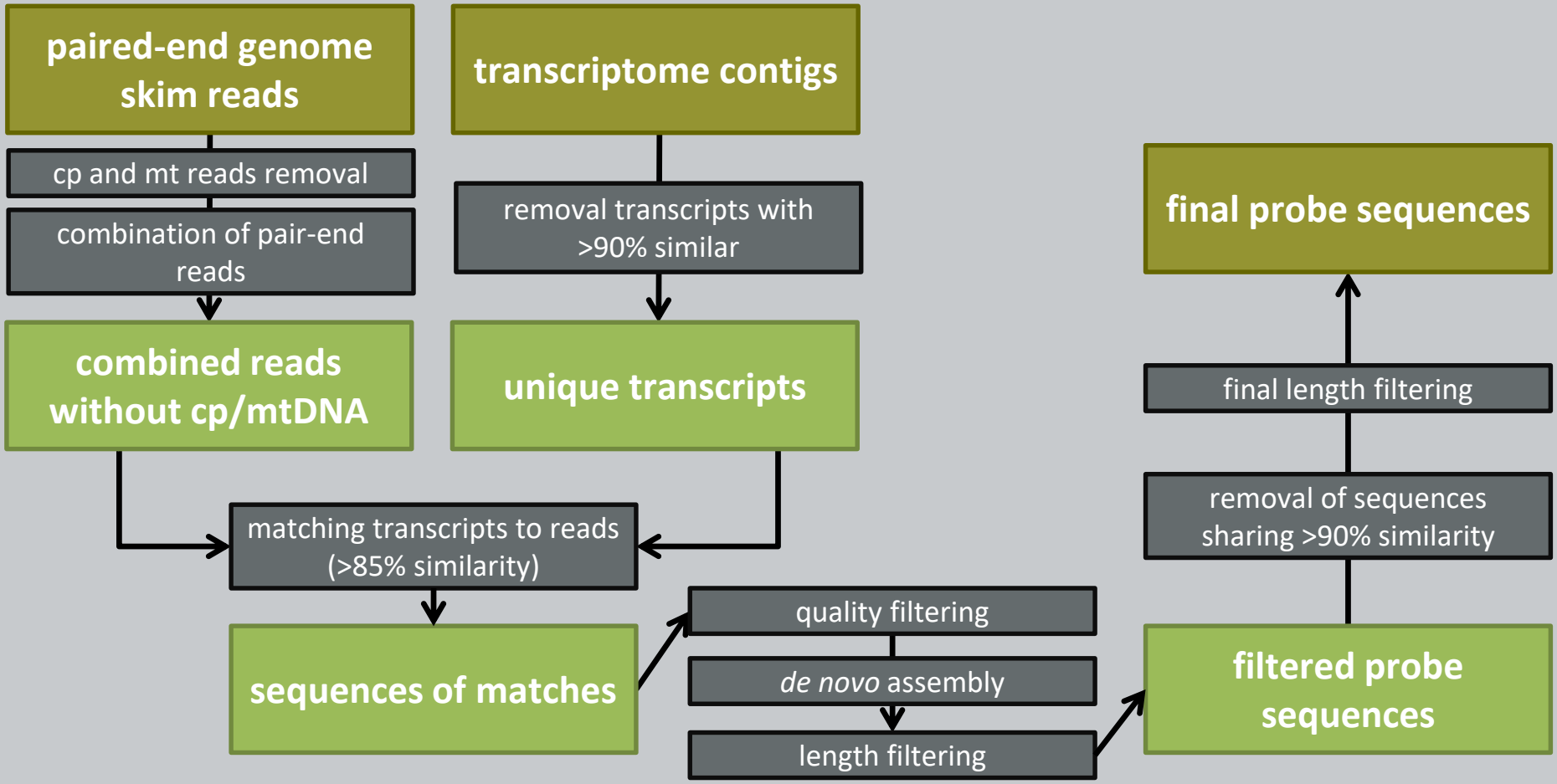
Weitemier et al. (2014) Appl Plant Sci. 2: apps.1400042
Cronn et al. (2012) Amer. J. Bot 99: 291-311
Lemmon et al. (2012) Syst. Biol.

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 variability 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 target genes
  - several thousands of target exons

Probe design for target enrichment

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

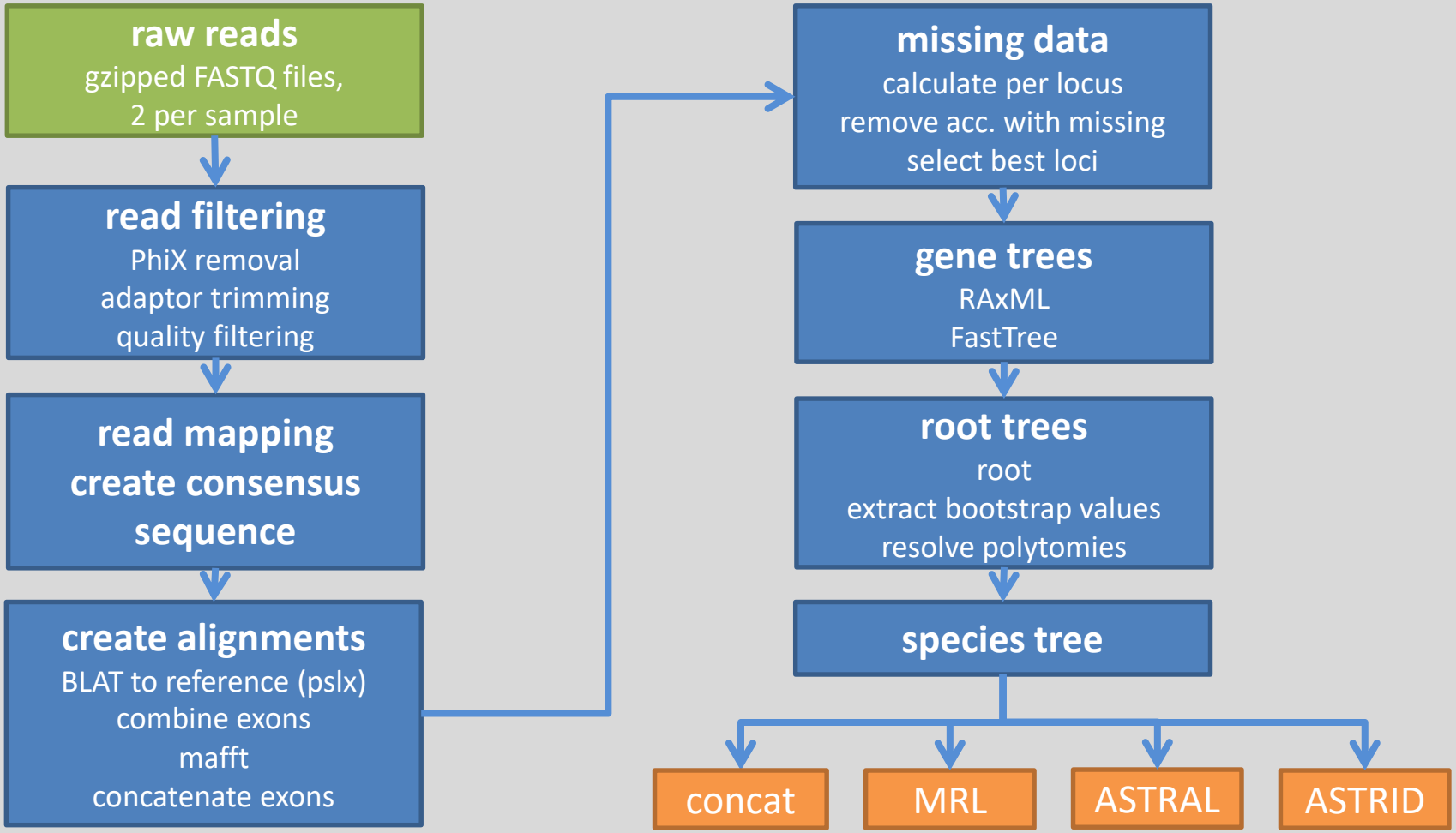


Schmickl et al. (2016): Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). Molecular Ecology Resources 16, 1124–1135.

Hyb-Seq data analysis software

- PHYLUCE
  - software for UCE (and general) phylogenomics
  - UCE – ultraconserved elements (http://ultraconserved.org)
  - Faircloth (2016): *PHYLUCE is a software package for the analysis of conserved genomic loci*. Bioinformatics 32:786-788.
  - https://github.com/faircloth-lab/phyluce
- HybPiper
  - Johnson et al. (2016): HybPiper: *Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment*. Applications in Plant Sciences 4(7): 1600016
  - https://github.com/mossmatters/HybPiper
  - allows analysis of intronic regions, putative paralogs
- HybPhyloMaker
  - Fér & Schmickl (2018): *HybPhyloMaker: target enrichment data analysis from raw reads to species trees*. Evolutionary Bioinformatics 14: 1-9.
  - https://github.com/tomas-fer/HybPhyloMaker
  - complete solution from raw reads to species trees (+networks, other analyses…)

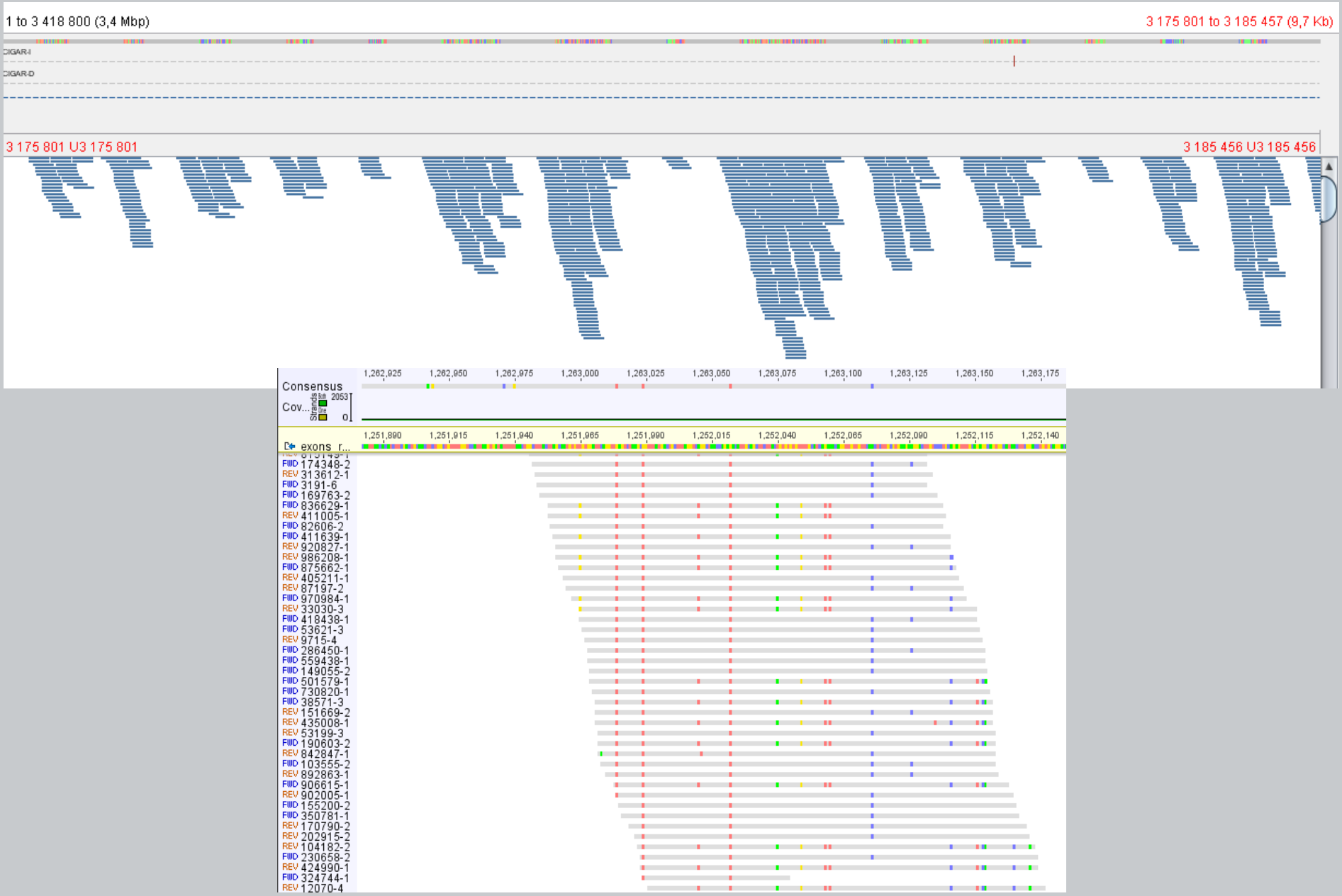
Hyb-Seq data analysis pipeline



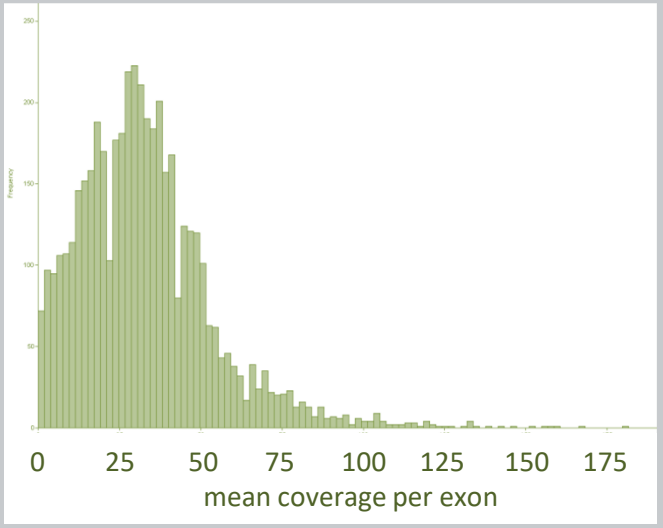
Weitemier et al. (2014) Appl Plant Sci. 2: apps.1400042 - Data Supplement S2

HybPhyloMaker: https://github.com/tomas-fer/HybPhyloMaker

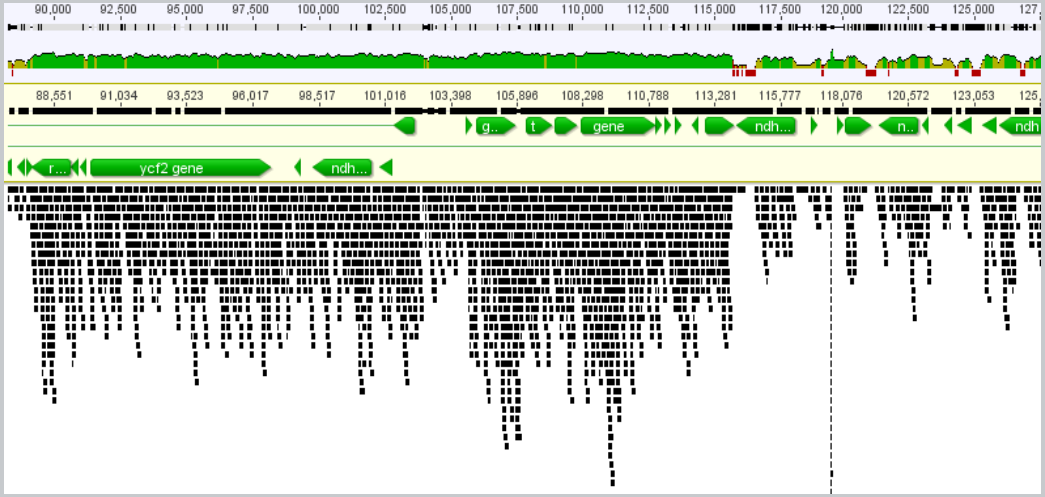
Hyb-Seq – reads mapped to reference

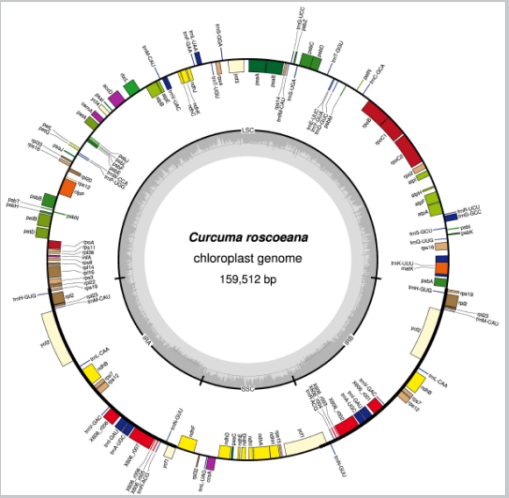


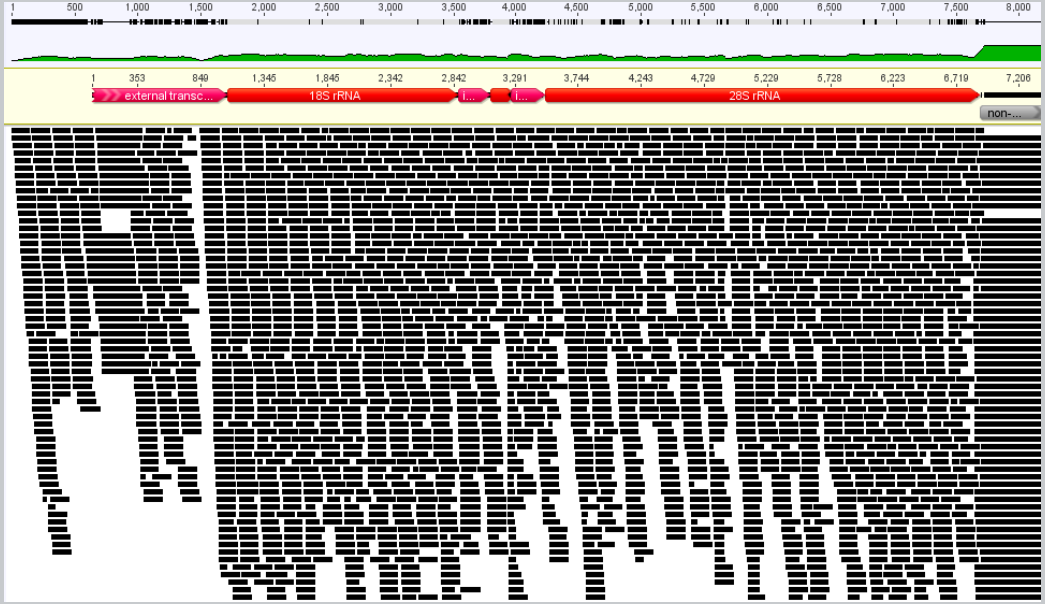
Coverage – nuclear exons



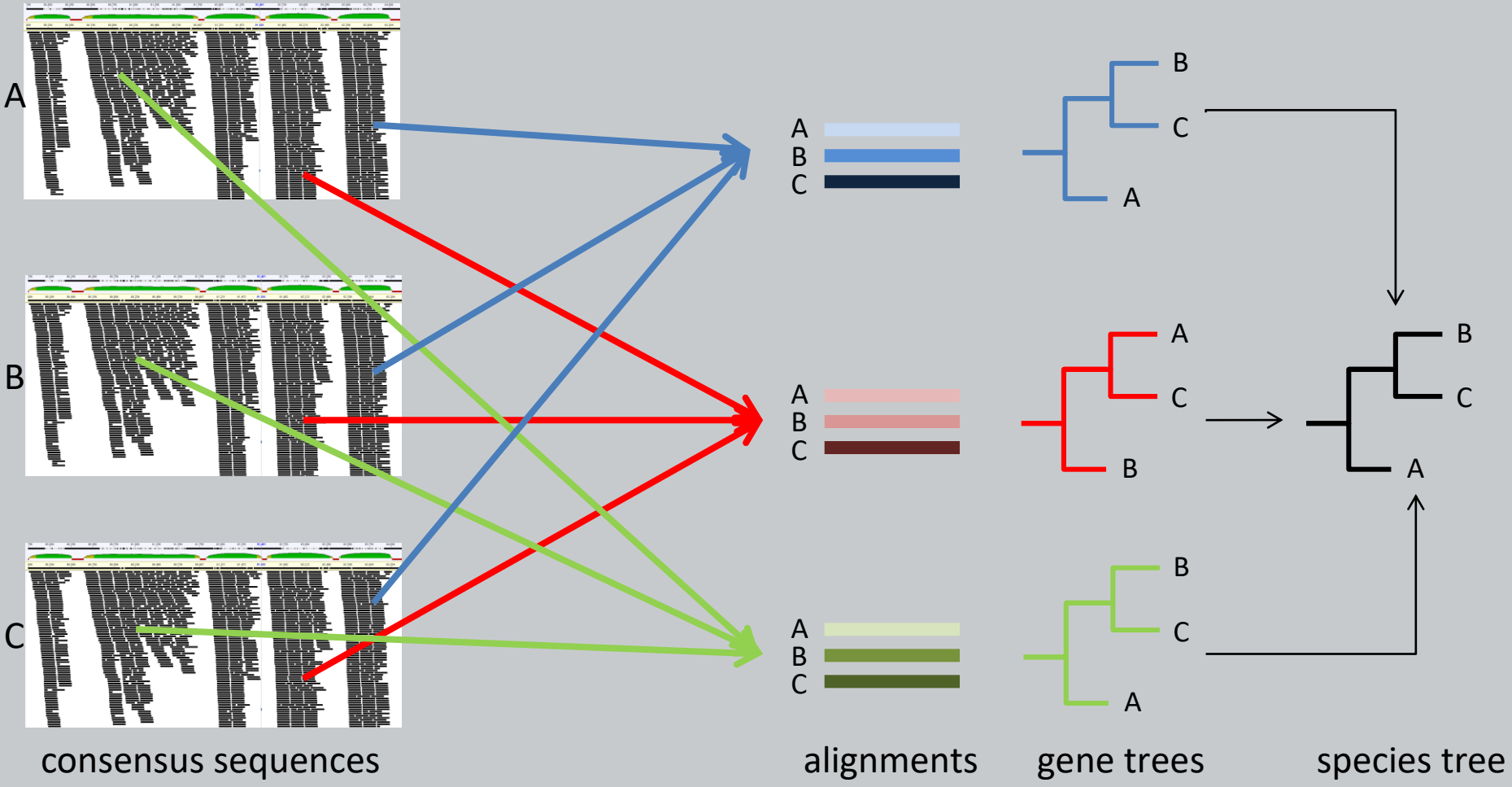
Coverage – cpDNA, rDNA



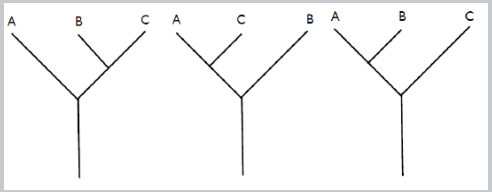




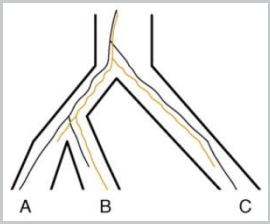
Read processing



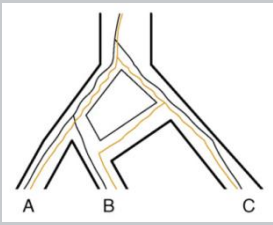
Incongruencies among loci: gene trees vs species tree



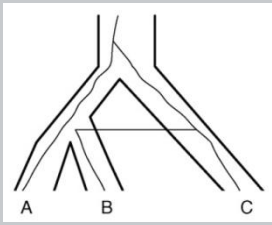
- incomplete lineage sorting (ILS)/deep coalescence
- gene duplications and losses (orthology problem)
- hybridization/polyploidization
  - affects whole genomes
- horizontal gene transfer (HGT)
  - affects small DNA segments
- recombination
  - different histories for neighboring segments in genes



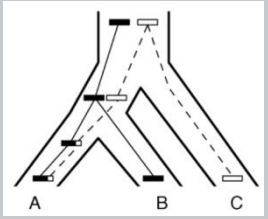
((AC)B)



((AB)C)
(A(BC))



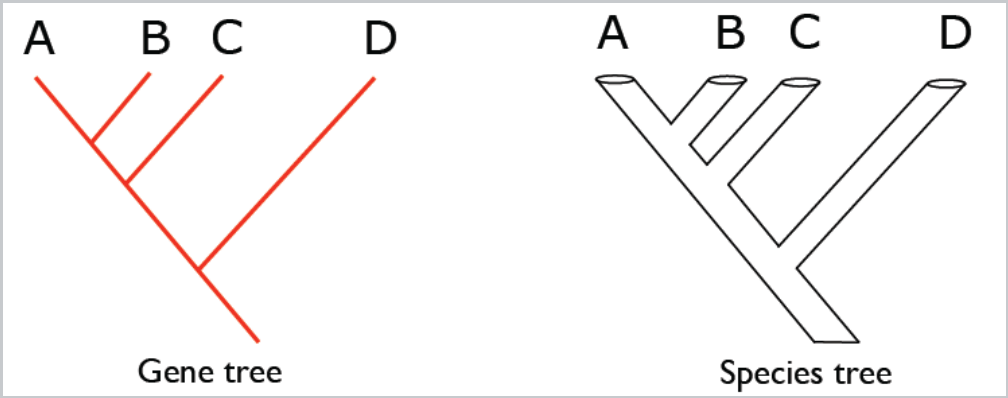
(A(BC))

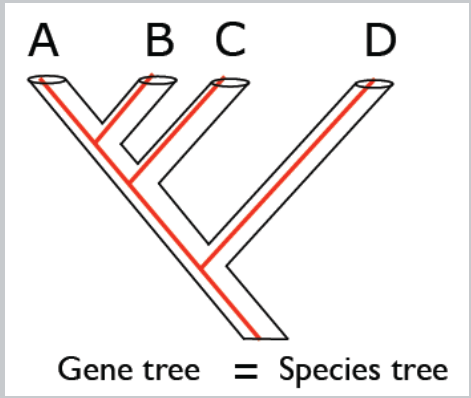


((AB)C)
((AC)B)

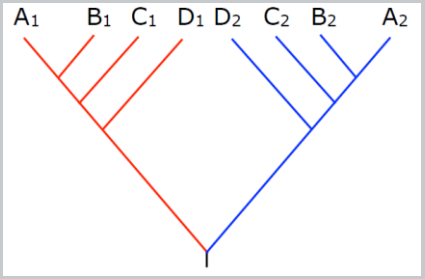
Degnan & Rosenberg, 2009

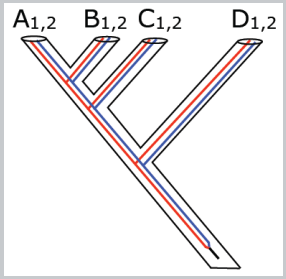
Gene trees vs species tree

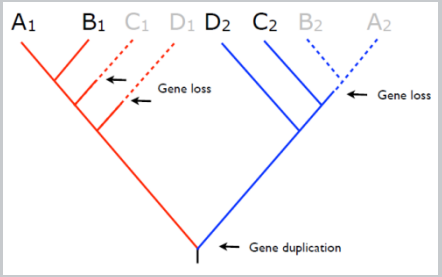




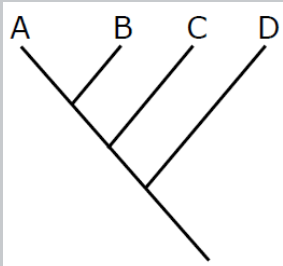
gene duplications and losses (GDL)



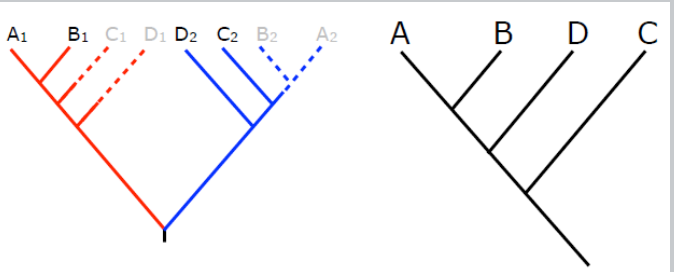




true species tree

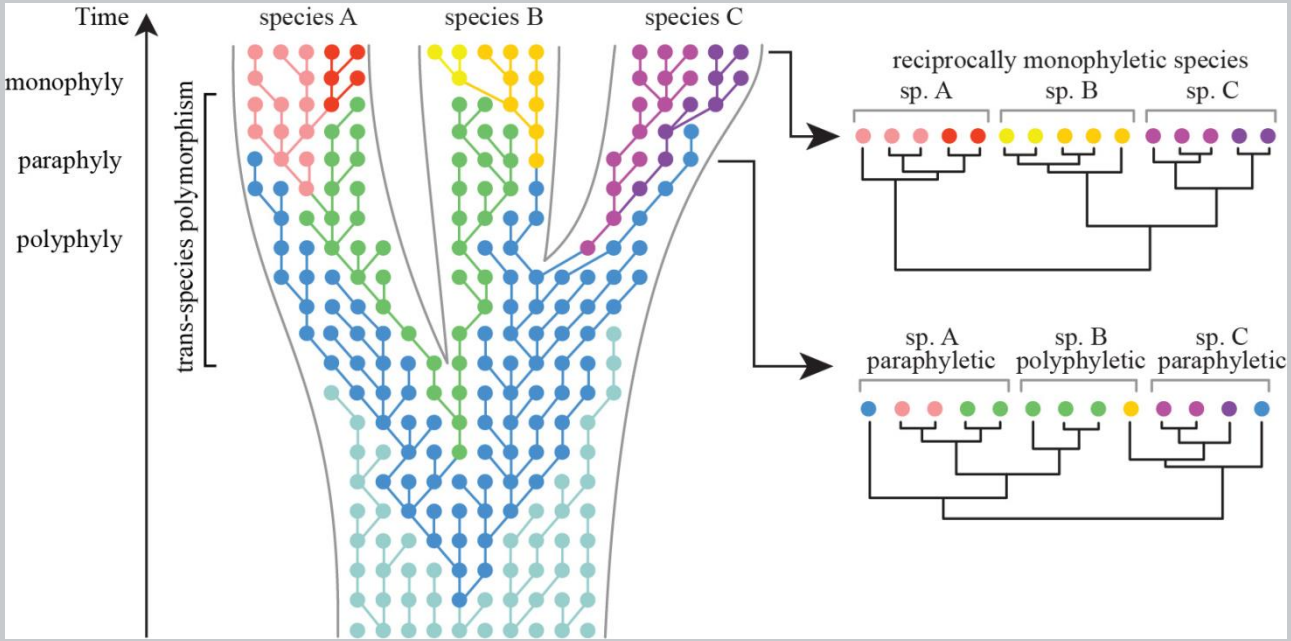


inferred species tree



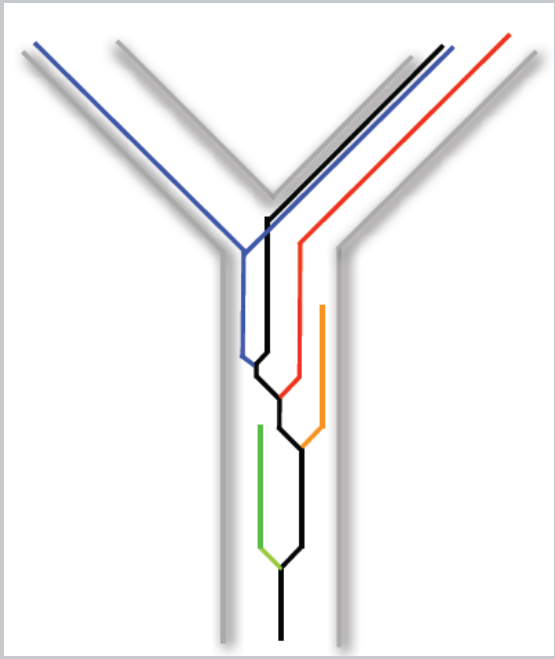
M. Popp, Oslo

Coalescence processes



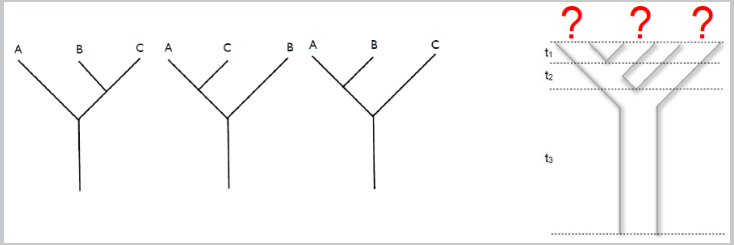
https://frederikleliaert.wordpress.com/green-algae/dna-based-species-delimitation-in-algae/

incomplete lineage sorting



M. Popp, Oslo

Species tree estimation



- **concatenation** – good unless strong ILS
  - single partition model (e.g. MP)
  - multiple partitions model (ML or Bayesian)
- **consensual methods** using MP – minimizes deep coalescences (MDC)
- multispecies coalescence (all incongruences due to differences in coalescence processes, no hybridization)
  - **coestimation** of gene trees and species tree – *BEAST – Bayesian analysis
  - **summary methods**
    - supertree methods – MRL (maximum representation using likelihood)
    - MP-EST – maximum likelihood estimation of rooted species tree
    - ASTRAL, ASTRID, STAR, STEAC – very fast and accurate
- Bayesian concordance analysis (BUCKy) – quartet-based Bayesian species tree estimation – uses concordance factor to build dominant history
- **site-based** methods
  - SNAPP, SVDquartets

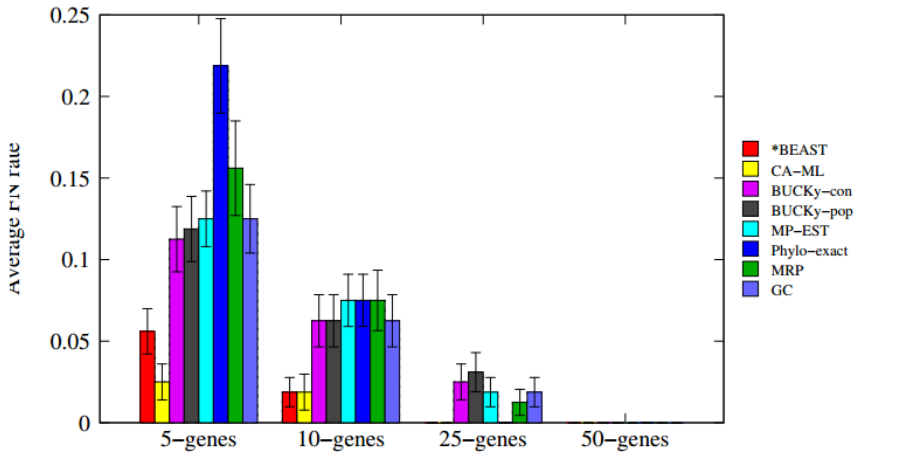
Summary methods

Species tree estimation

- supertree
  - MRP – matrix representation using parsimony
  - MRL – matrix representation using likelihood
- MP-EST – maximum pseudo-likelihood approach for estimating species trees
- ASTRAL – Accurate Species Tree Reconstruction ALgorithm
- ASTRID – Accurate Species TRees from Internode Distances (reimplementation of NJ_{st} method)
- …

Methods comparison

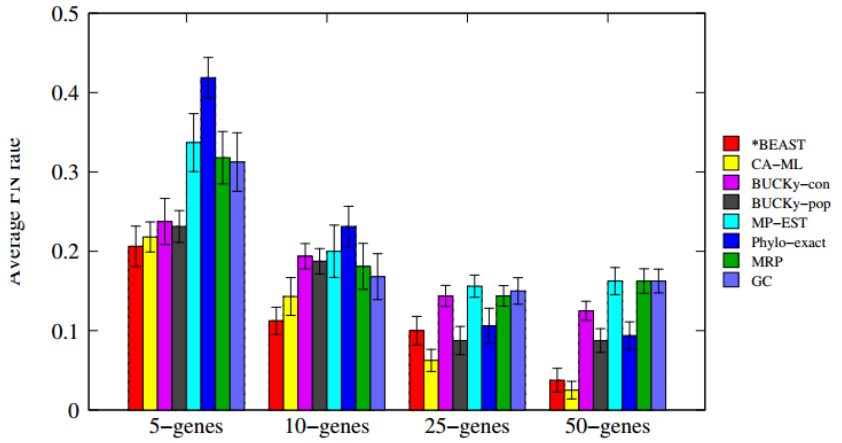
Results on 11-taxon datasets with weak ILS



*BEAST more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: concatenated analysis) most accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013

Results on 11-taxon datasets with strongILS

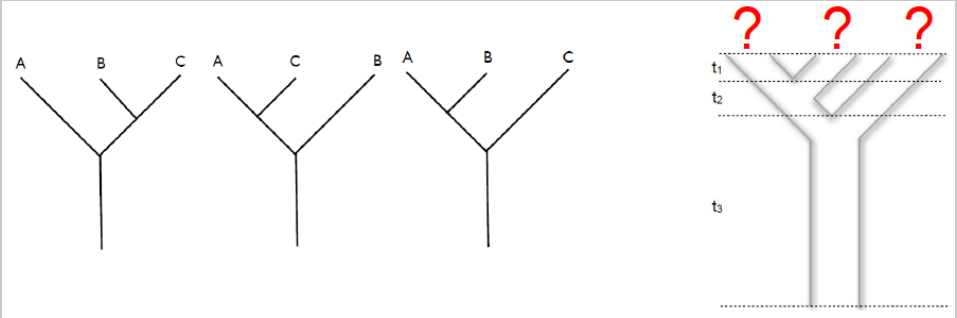


*BEAST more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: (concatenated analysis) also very accurate

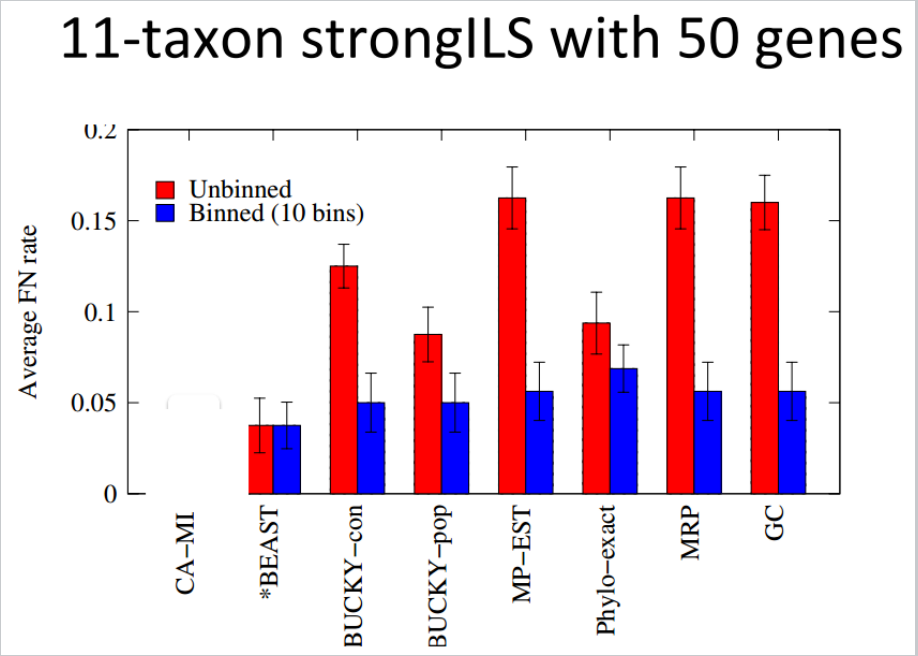
Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013

T. Warnow, The University of Texas
https://www.cs.utexas.edu/users/tandy/394C-nov20-2013.pdf

Gene binning



- wrong species tree if poor gene trees
  - shorter alignments usually give poorly supported trees
- improve gene trees
  - collapse unsupported branches
  - **binning** – assign gene to bins
    – create supergene alignments
    - **naïve binning** – random
    - **statistical binning**
      – no incompatibility among gene trees in the same set



T. Warnow, The University of Texas
https://www.cs.utexas.edu/users/tandy/394C-nov20-2013.pdf

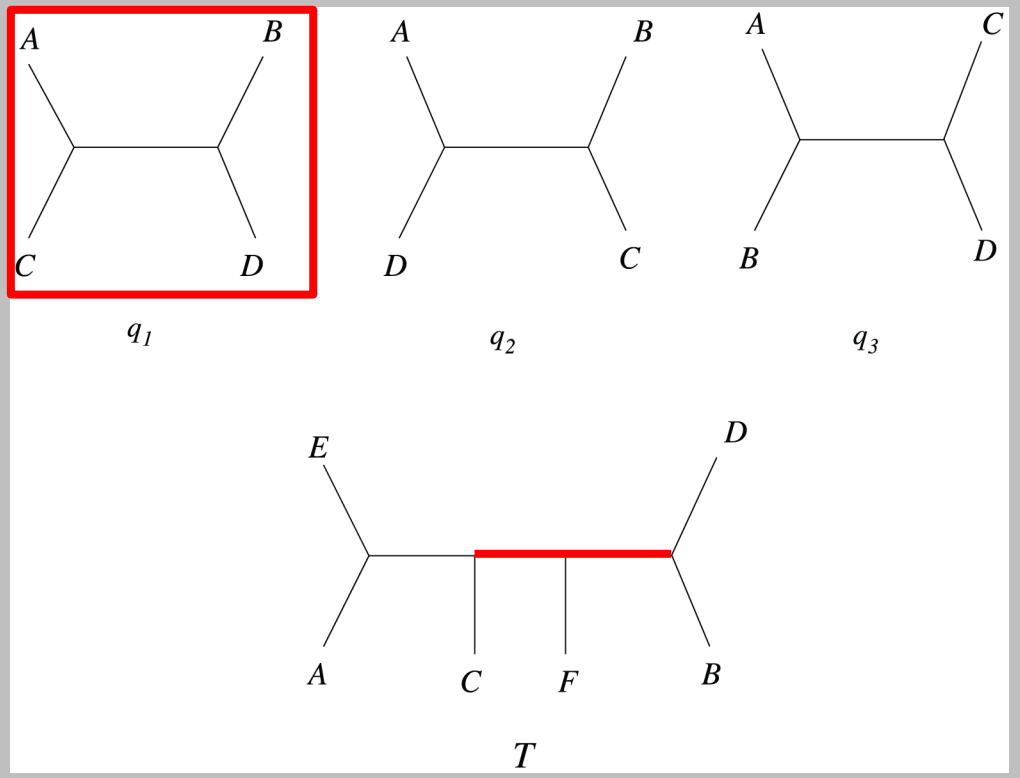
ASTRAL

Accurate Species Tree Reconstruction ALgorithm

- unrooted gene trees
- species tree that agrees with the largest number of quartet trees induced by the set of gene trees
- weighting all three alternative quartet topologies according to their relative frequencies within gene trees
  - much more frequent topology – trees without this topology are penalized
  - similar frequencies (i.e., close to 0.33) – the quartet has little impact to optimization
- final species tree with
  - local posterior probability that the branch is in the species tree
  - the length of internal branches in coalescent units

Tree reconstruction from quartets

- quartet – unrooted tree over 4 taxa
- three possible quartets
- only one quartet *q* is consistent with final tree **T**



Reaz et al. (2015): *Accurate Phylogenetic Tree Reconstruction from Quartets: A Heuristic Approach*. PLoS ONE 9, e104008.

MRL

Maximum Representation with Likelihood

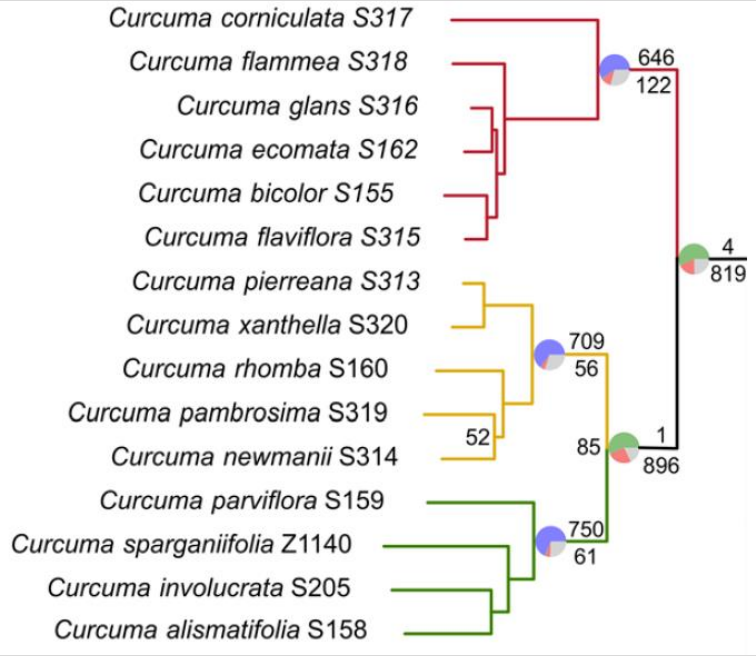
- supertree methods – estimates species tree on full taxon sets from sets of smaller trees (i.e., with missing species)
- encodes a set of gene trees by a large randomized matrix
- each edge (branch) in each gene tree
  - '0' for the taxa that are on one side of the edge
  - '1' for the taxa on the other side
  - '?' for all the remaining taxa (i.e., the ones that do not appear in the tree)
- MRL matrix is analyzed using heuristics for a symmetric 2-state Maximum Likelihood
  - in RAxML as 'BINCAT' model

Evaluating gene tree discordance

PhyParts

https://bitbucket.org/blackrim/phyparts

proportion of trees concordant with species tree



blue: support the shown topology

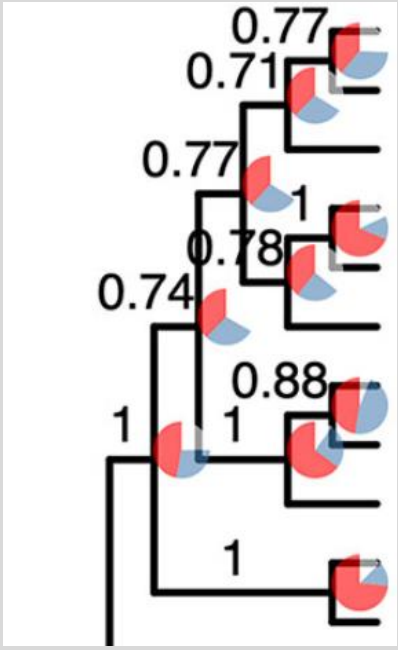
green: conflict with the shown topology (most common conflicting bipartition)

red: conflict with the shown topology (all other supported conflicting bipartitions)

gray: have no support for conflicting bipartion

ASTRAL scoring

proportion of quartets concordant with node



Concatenation vs. coalescence

- concatenation
  - in favor: longer datasets allow for hidden support to appear
  - against: could be misleading under strong ILS

- coalescence (i.e., “shortcut coalescence” or summary methods)
  - in favor: addresses ILS
  - against:
    - short genes give poor gene trees (big problem!)
    - definition of coalescence-gene (segments with no internal recombining) debatable
    - concatenating coalescence-genes to longer alignments (“concatalescence”) not recommended?

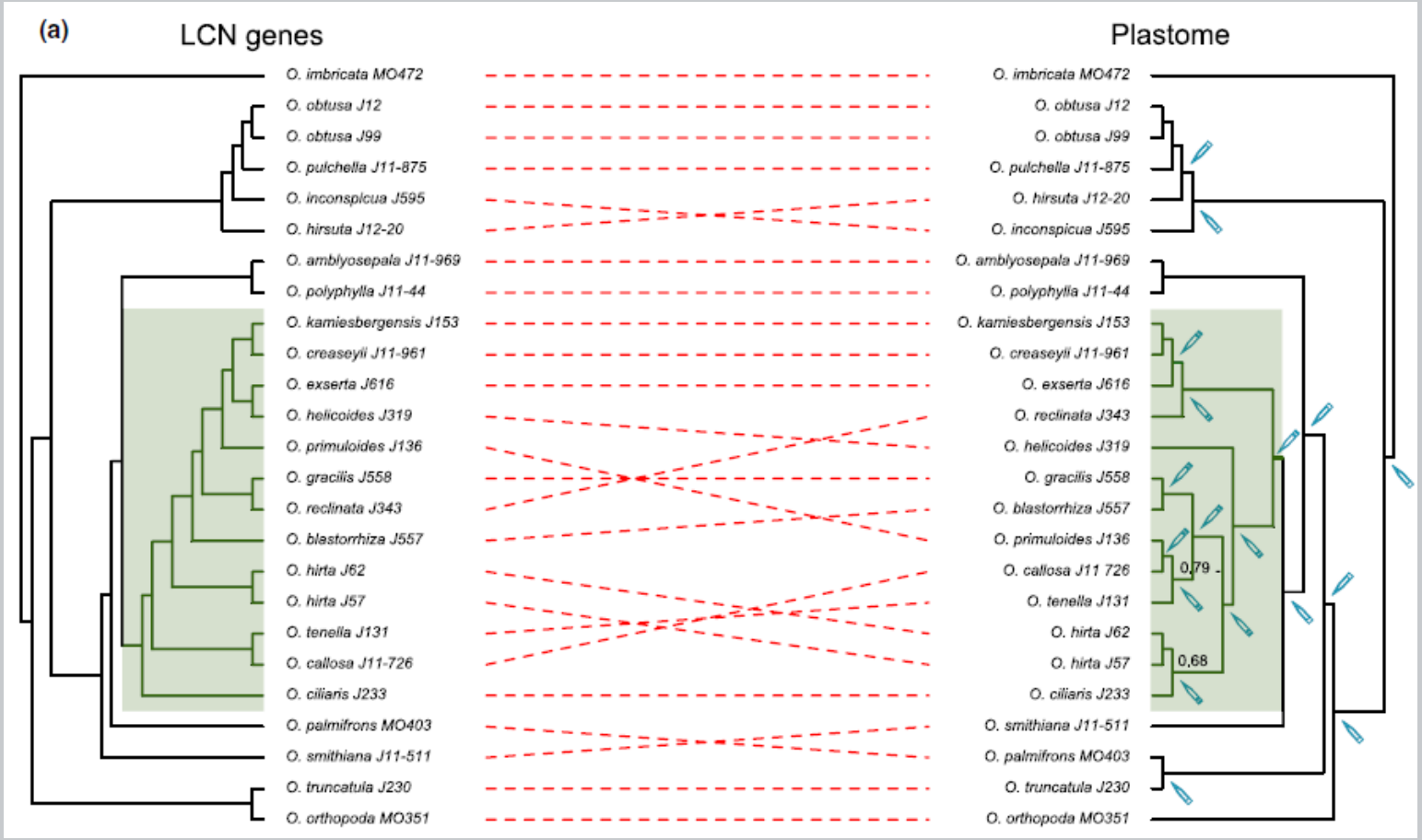
see also:
Gatesy & Springer (2014): Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Molecular Phylogenetics and Evolution 80: 231–266.

HybSeq applications

- family/order phylogeny
  - Asteraceae (Mandel et al. 2014)
  - Zingiberales (Sass et al. 2016, Carlsen et al. 2018)
  - Brasicaceae (Nikolov et al. 2019)
  - Costaceae (Böhmová et al. 2023)
  - …
- genus level radiation
  - *Sarracenia* (Stephens et al. 2015)
  - *Inga* (Nicholls et al. 2015)
  - *Helianthus* (Stephens et al. 2015)
  - *Sabal* (Heyduk et al. 2016)
  - *Oxalis* (Schmickl et al. 2016)
  - *Amomum* (Hlavatá et al. 2023)
  - …
- phylogeny of polyploids
  - *Gossypium* (Grover et al. 2015)
- within species variability
  - *Pinus albicaulis* (Syring et al. 2016)
  - *Euphorbia balsamifera* (Villaverde et al. 2018)

Differences among nuclear and cpDNA

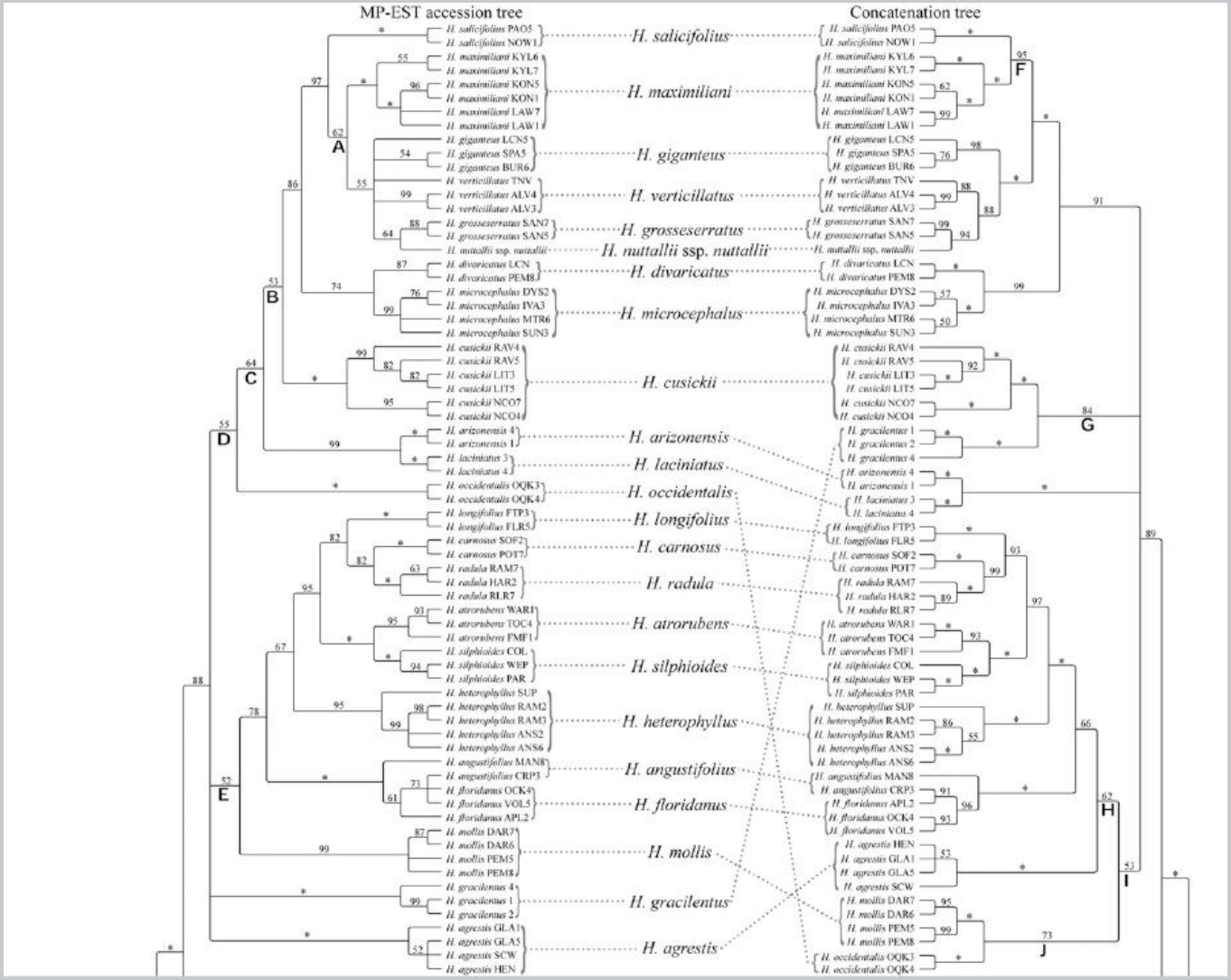
Oxalis
727 genes



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.

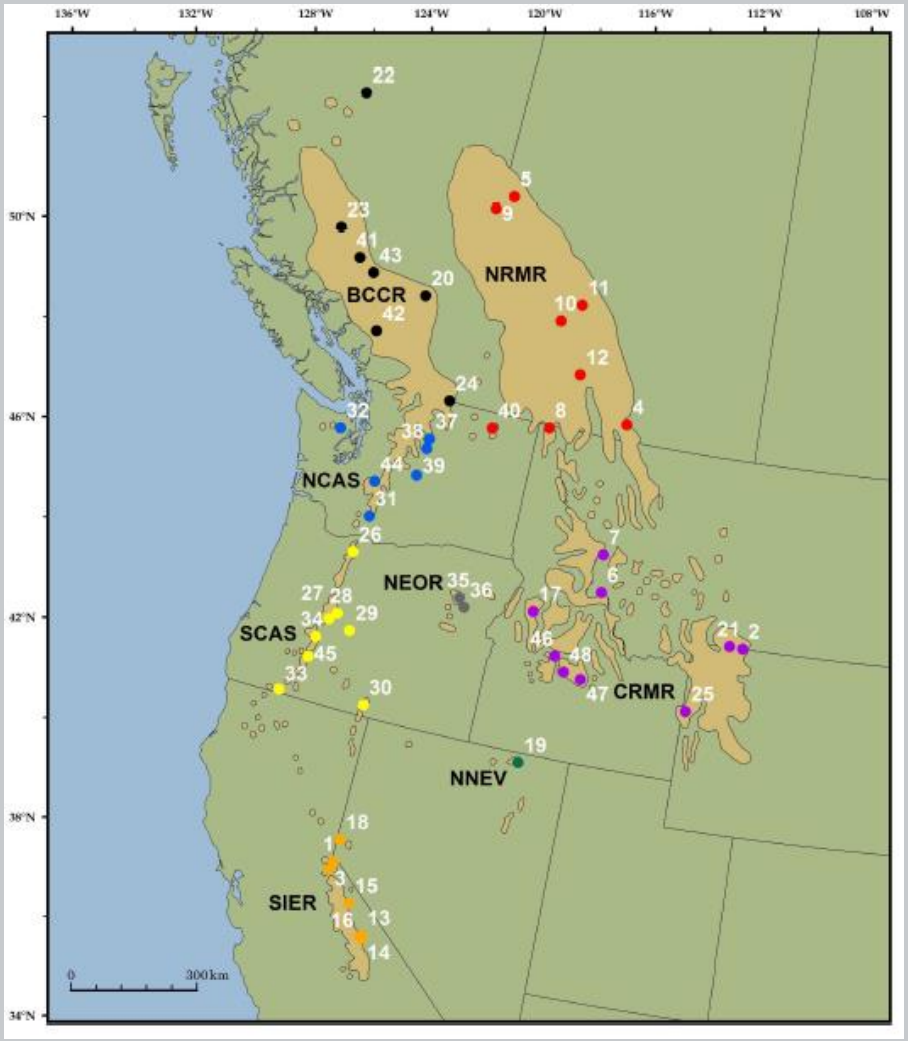
Differences among concatenation and coalescence-based trees

Helianthus
170 genes
106,862 sites



Stephens et al. (2015): *Species tree estimation of diploid Helianthus (Asteraceae) using target enrichment*. American Journal of Botany 102, 1–11.

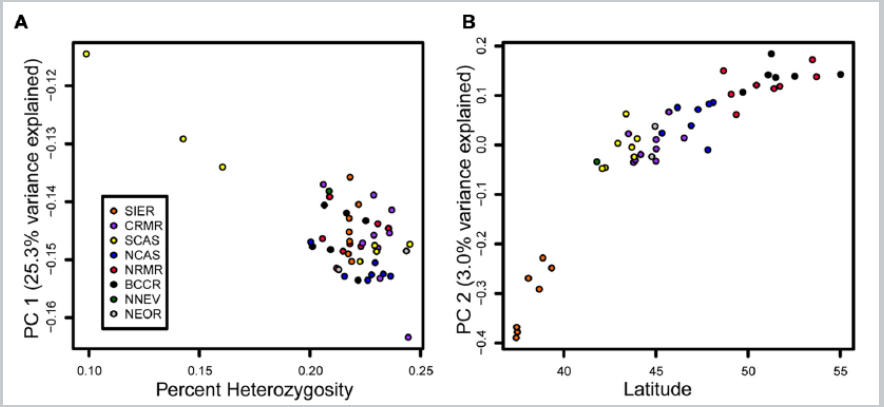
Within-species relationships



Pinus albicaulis

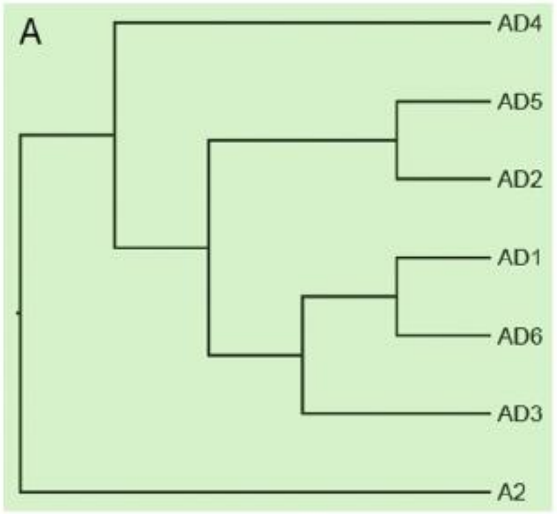
4,452 genes

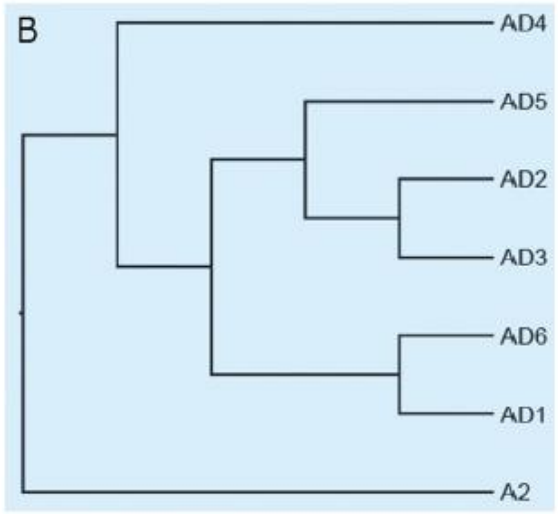
528,873 sites



Syring et al. (2015): T*argeted capture sequencing in Whitebark Pine reveals range-wide demographic and adaptive patterns despite challenges of a large, repetitive genome*. Frontiers in Plant Science 7:484.

Diploid-polyploid relationships





A homoeologs

	CF	CI	
[1]	0.950	(0.868-0.974)	AD1/AD6
[2]	0.926	(0.816-0.974)	AD2/AD5
[3]	0.914	(0.737-1.000)	A2/AD4
[4]	0.844	(0.553-0.947)	AD1/AD3/AD6
[5]	0.106	(0.026-0.395)	AD2/AD3/AD5

D homoeologs

	CF	CI	
[1]	0.981	(0.912-1.000)	AD2/AD5
[2]	0.778	(0.618-0.912)	AD1/AD6
[3]	0.727	(0.500-0.882)	D5/AD4
[4]	0.704	(0.559-0.794)	AD1/AD3/AD6
[5]	0.214	(0.147-0.324)	AD2/AD3/AD5
[6]	0.151	(0.059-0.382)	D5/AD2/AD5
[7]	0.133	(0.029-0.235)	D5/AD1/AD4
[8]	0.077	(0.029-0.206)	AD1/AD4/AD6
[9]	0.058	(0.029-0.176)	D5/AD3/AD4

Gossypium

Grover et al. (2015): *Re-evaluating the phylogeny of allopolyploid Gossypium L.* Molecular Phylogenetics and Evolution 92: 45–52.

Systematic study

Stephens et al. (2015): Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. *Molecular Phylogenetics and Evolution* 85, 76–87.

Literature

Weitemier K. et al. (2014): *Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics*. Appl. Plant Sci. 2: apps.1400042

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.

McCormack J.E. et al. (2011): *Applications of next-generation sequencing to phylogeography and phylogenetics*. Mol. Phylogenet.Evol.

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

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

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