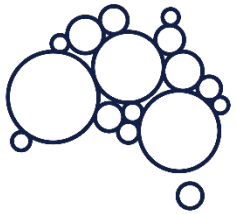
Genomics for Australian Plants **webinar** and **workshop** series for analysing target capture datasets



BIOPLATFORMS
AUSTRALIA

Australian
BioCommons

GAP Webinar and Workshop series

Webinars

Conflict in multi-gene datasets: Why it happens and what to do about it

Dr Alexander Schmidt-Lebuhn, CSIRO

20th May, but recording is available online

Detection and phasing of hybrids in target capture datasets

Dr Lars Nauheimer, Australian Tropical Herbarium

email: lars.nauheimer@jcu.edu.au

recording will be available online

Genomics for Australian Plants

BIOPLATFORMS AUSTRALIA

Australian BioCommons

ATH Australian Tropical Herbarium

GAP Webinar and Workshop series

Workshops (5th – 8th July)

1) Assembly of raw reads using HybPiper

2) Paralogy resolution using Yang and Smith

3) Detection and phasing of hybrids using HybPhaser

Australasian Systematic Botany Society
Annual Conference 2021

12th-16th July

Expressions of interest to participate close this Sunday June 13th!!! https://asbs2021.bablglobal.com/workshop/

Genomics for Australian Plants

BIOPLATFORMS AUSTRALIA

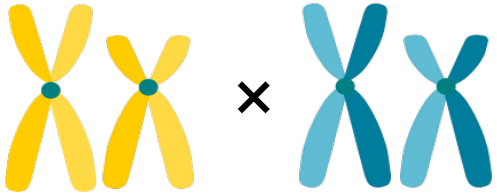
Australian BioCommons

ATH Australian Tropical Herbarium

Definition

Hybridization

Crossing of two species (or genetically divergent populations)



Hybrid

Offspring of a hybridization event, containing genetic material from both parental lineages

Allele

Variant of a gene

Haplotype

Group of alleles / part of the genome that is inherited from a parent

Hybridization

- **Homoploid** hybridization

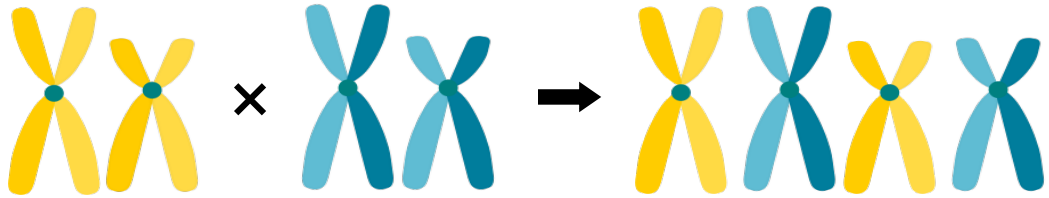
  2n = 2

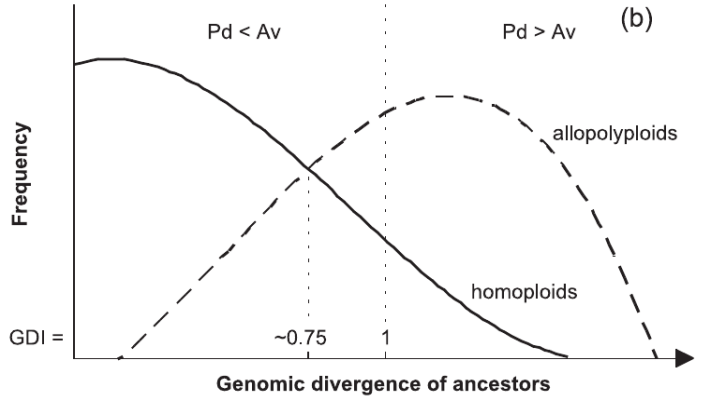
  - More frequent
  - Backcrossing likely

- **Polyploid** hybridization

  2n = 4

  - Reproductive isolation from parent populations through chromosome numbers

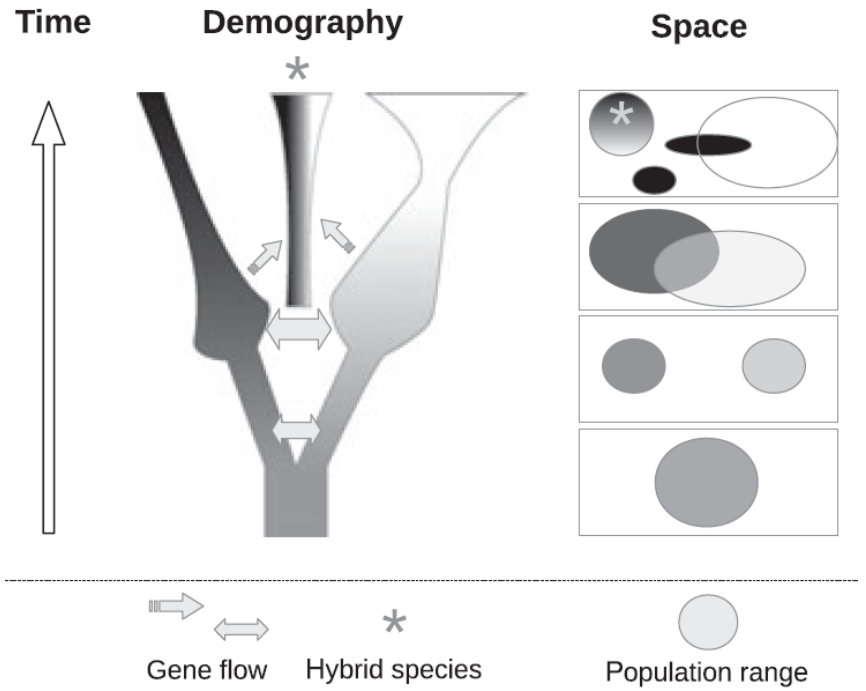




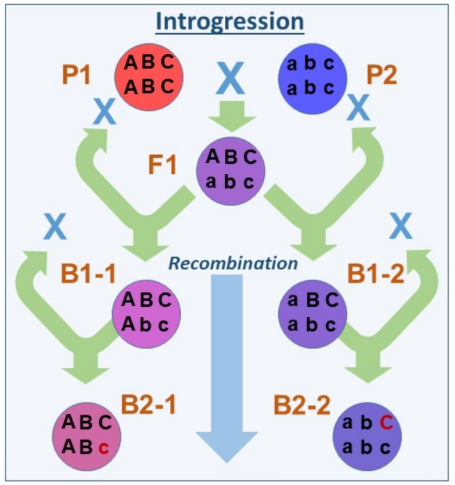
Paun et al. 2009

Hybrid speciation

- Most hybridizations do not lead to speciation
- Speciation often requires reproductive isolation to avoid backcrossing (ingression)
- Introgression can lead to exchange of alleles between species



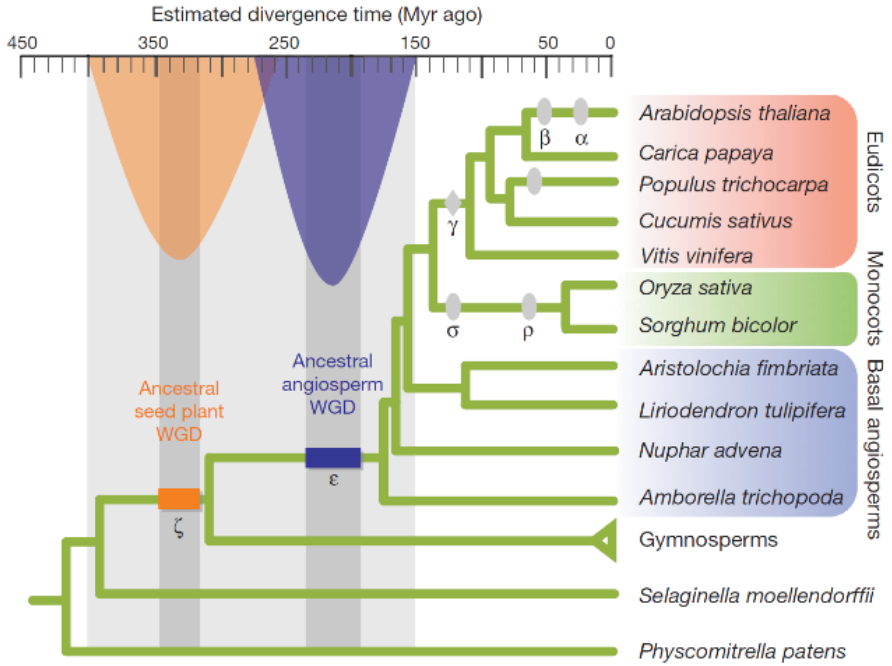
Abbott et al. 2013



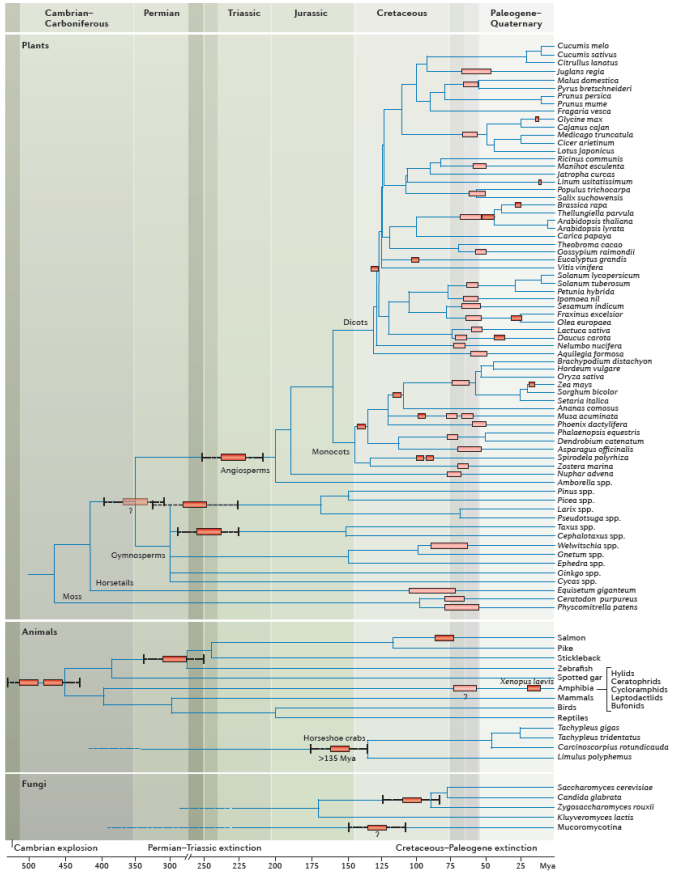
©Mcruzan, Wikipedia.org

Polyploidy / Whole Genome Duplications

Important driver in evolution

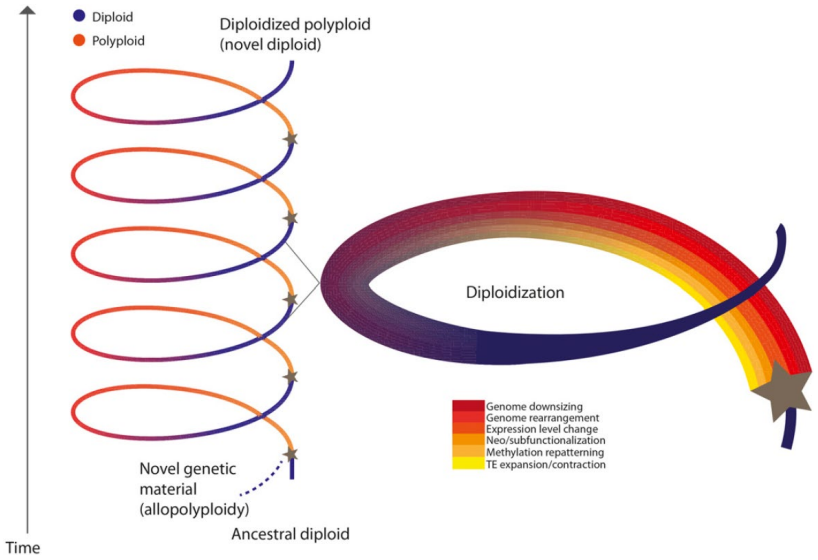


Jiao et al. 2011



Peer et al. 2017

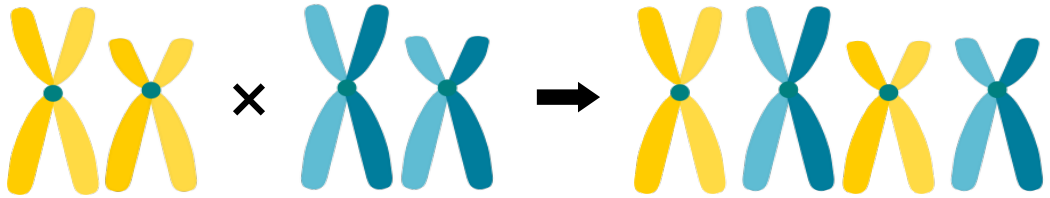
‘slinky of life’



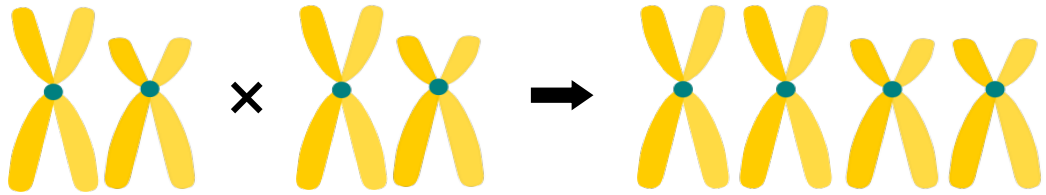
Soltis et al. 2016

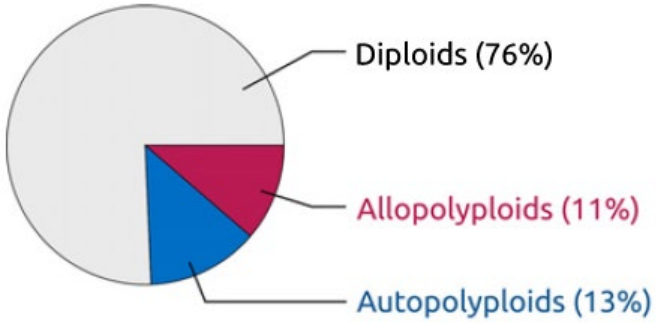
Polyploidy / Whole Genome duplication

- **Allo**polyploidy (hybrid)
  2n = 4



- **Auto**polyploidy
  2n = 4

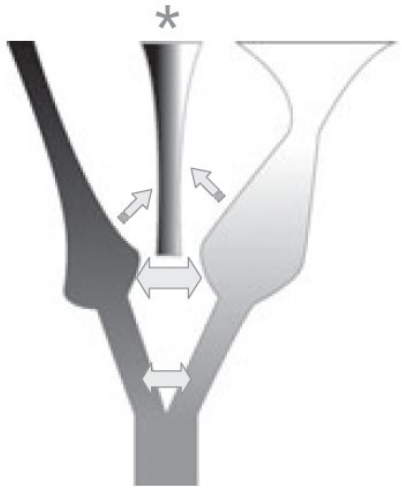


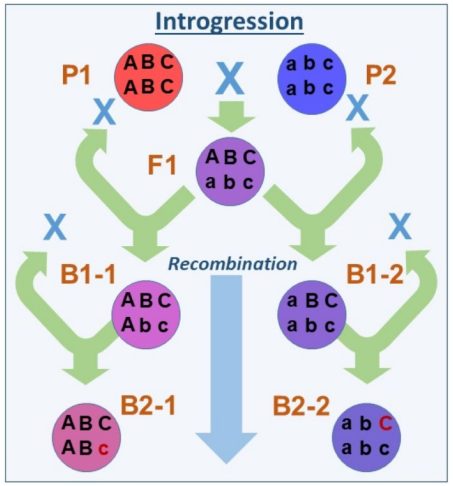


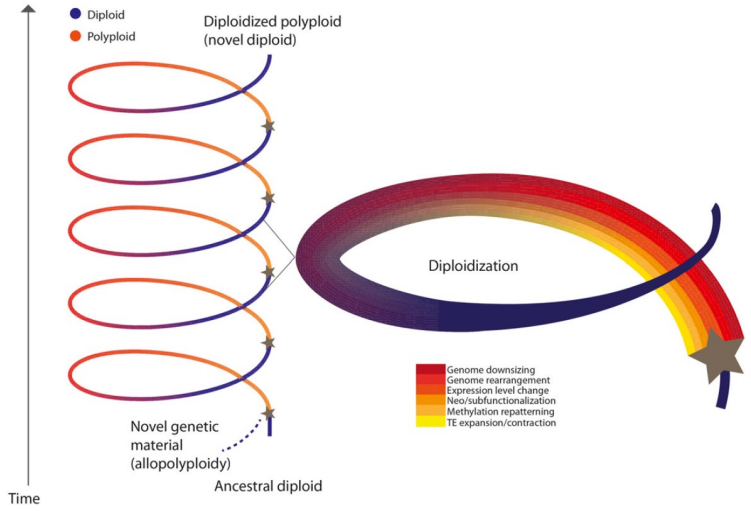
*data from 47 genera / 4003 species
Barker et al. 2016

Polyploids can retain gene copies for a long time that persist as paralogs!

Genome evolution is complex







Hybrid speciation

Introgression

Polyploid-/Diploidisation

+ Autopolyploidy - paralogous genes

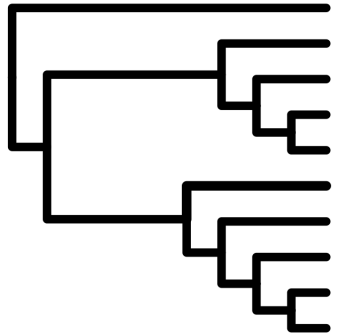
+ Deep coalescence (incomplete lineage sorting) - allele sorting different to speciation

Hybrids in phylogenetics

Hybrids in Phylogenetics

The combination of genetic information from divergent lineages can lead to **conflicting phylogenetic signal**.

=> poor clade support, false reconstructions



Results depend on the methodology!

Assembly

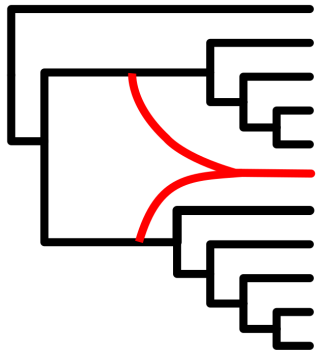
- How are the sequences recovered?
- Is the divergent signal mixed up, separated, removed and noticed at all?

Handling of sequences / dataset

- Are loci concatenated? Plastid / nuclear? Are phased sequences linked?

Phylogenetic analyses

- Bifurcating phylogeny? Network analysis?



Hybrids in Phylogenetics

Sanger sequencing

- Often plastid data only (maternally inherited)
- ITS (nuclear rRNA-array) and plastid data
- Few single copy nuclear markers
- Phasing of sequences with PCR and cloning

- Hybridisation has often been ignored or not noticed

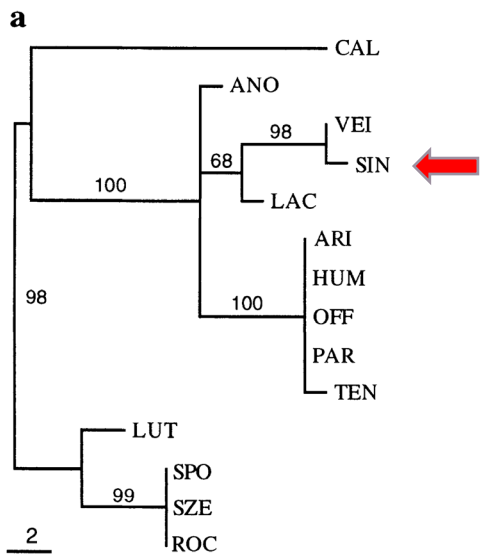
- Conflicting between nuclear and plastid phylogenies – chloroplast capture

- Few studies revealed origin of hybrids through phasing of nuclear loci

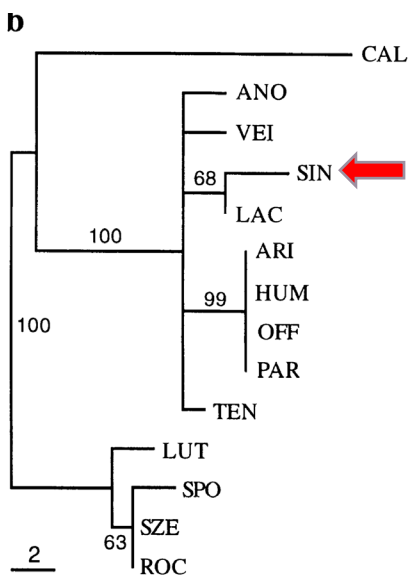
Hybrids in Phylogenetics

Phasing of alleles from nuclear gene can reveal parental clades/lineages

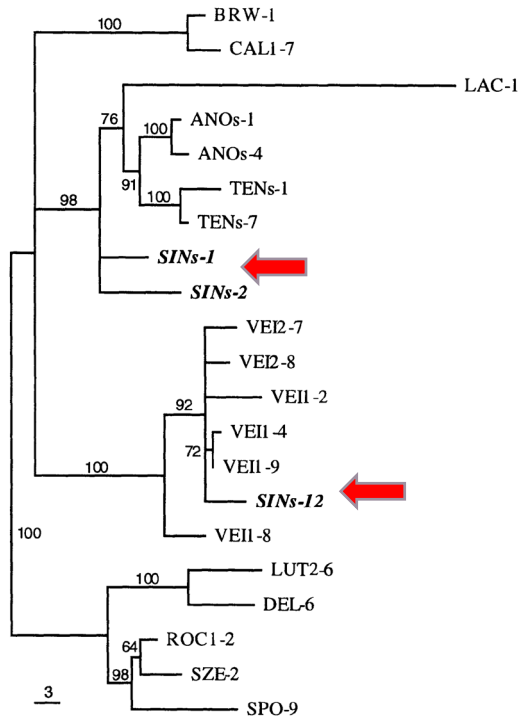
ITS



*mat*K



*adh*2



Sang & Zhang 1999

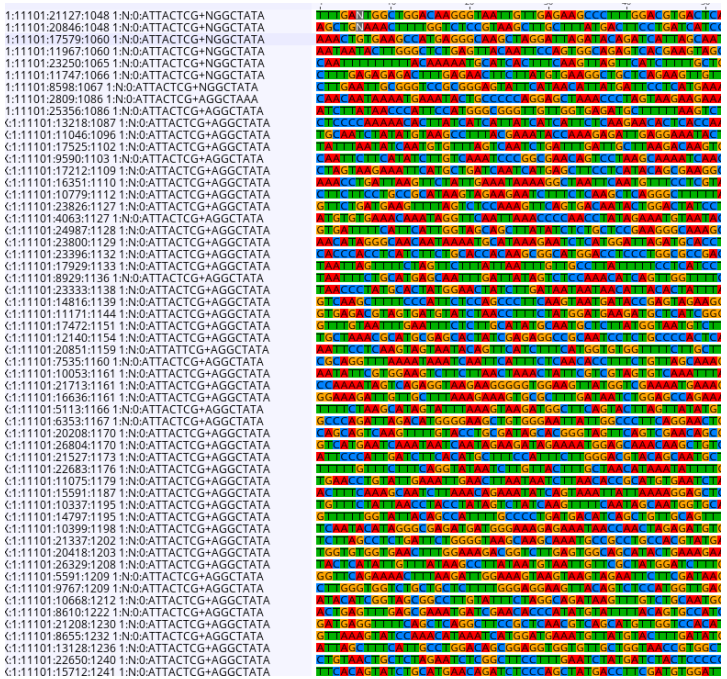
Hybrids in Phylogenomics

Target Capture Sequencing

- Hundreds of nuclear genes
  - Based on transcriptome data (exons)

- Illumina short read sequencing
  - 2-10 Million reads per sample
  - ~125-250bp length
  - Single-end / paired-end

  - Reads from all alleles / haplotypes



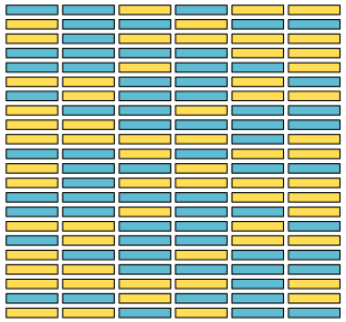
Hybrids in Target Capture Sequencing

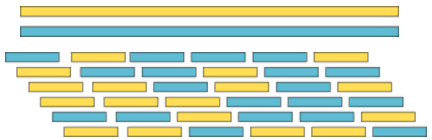
Opportunities

- Nuclear genes (high numbers)
- All variants can be recovered!

Challenges

- Assembly of short reads into alleles
- Phased alleles need to be linked across genes / exons
- Analysis of TC data often intransparent





Target Capture – Assembly

***De novo* assembly**

- Sequence reads are matched to each other
- => contig

Reference based assembly

- Sequence reads are mapped to a reference
- => consensus sequence

Most pipelines use a combination of *de novo* and mapping!

(first mapping reads of a single locus to *de novo* assemble only on-target reads)

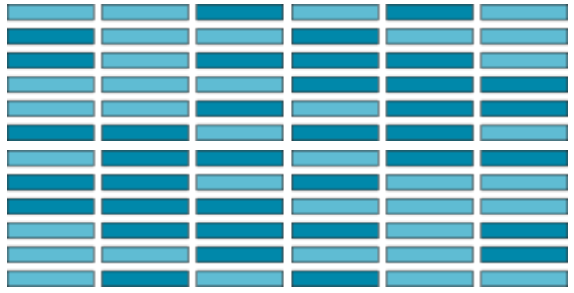
Target Capture – Assembly of hybrids

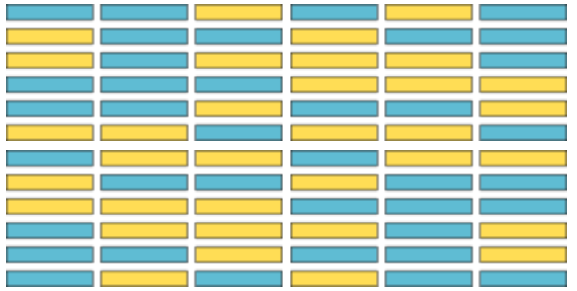
What happens when divergent reads are included?

Diploid non-hybrid

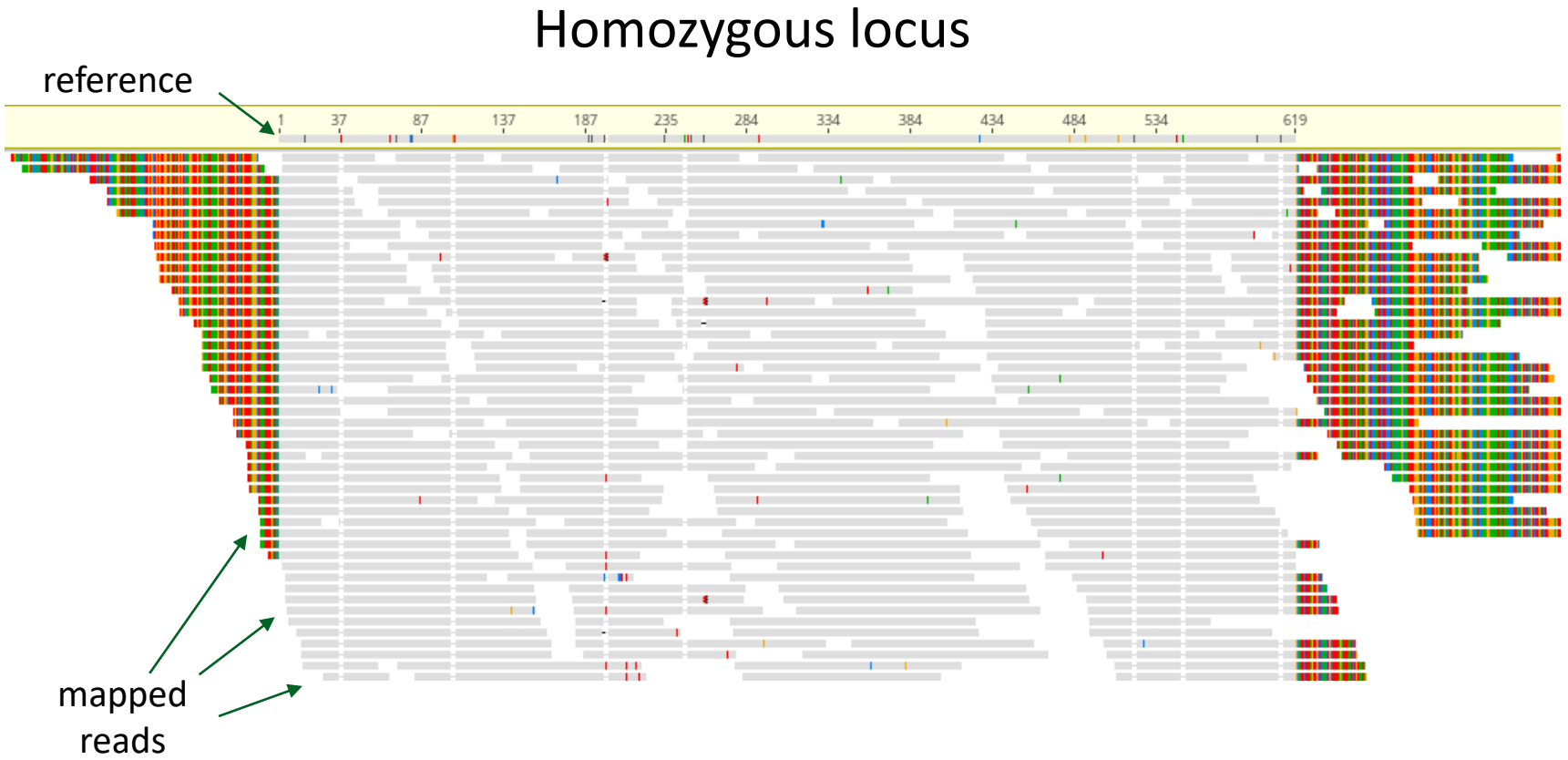
Diploid hybrid / polyploid ‘hybrid’

Sequence reads





Target Capture – Assembly of hybrids



Target Capture – Assembly of hybrids

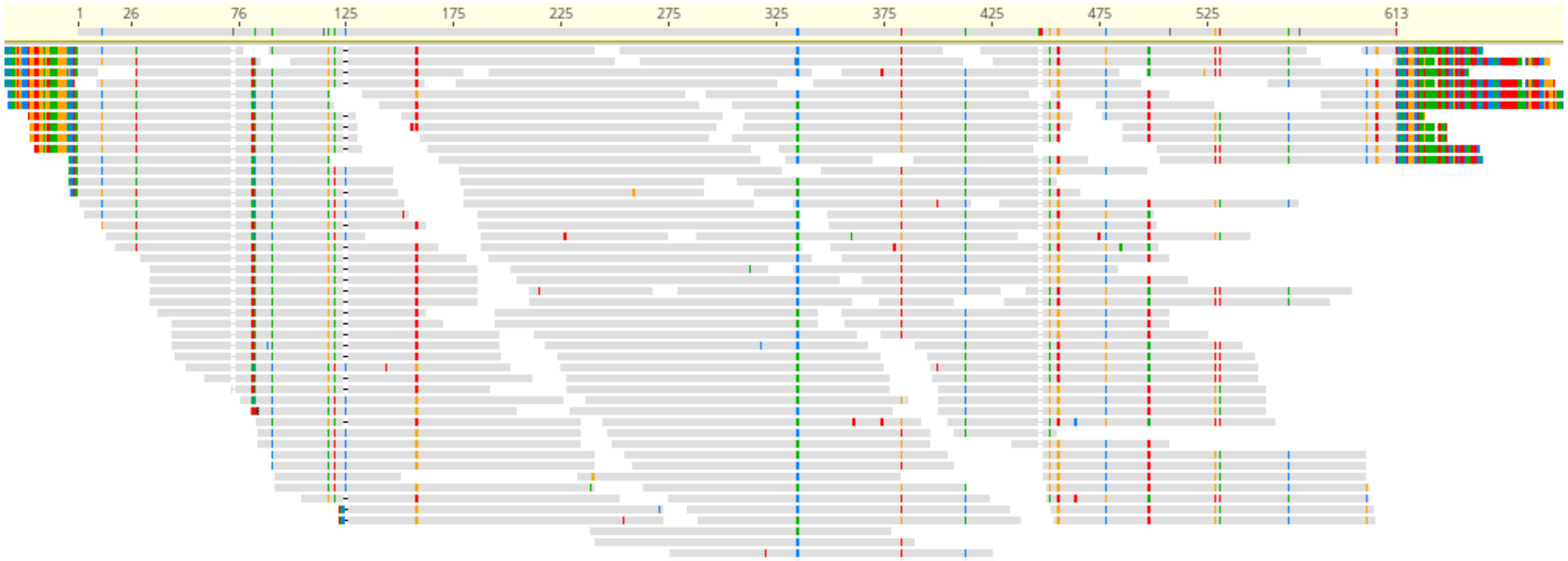
Heterozygous locus

heterozygous site / SNP (single nucleotide polymorphism)

Target Capture – Assembly of hybrids

Reference mapping

Heterozygous locus - hybrid

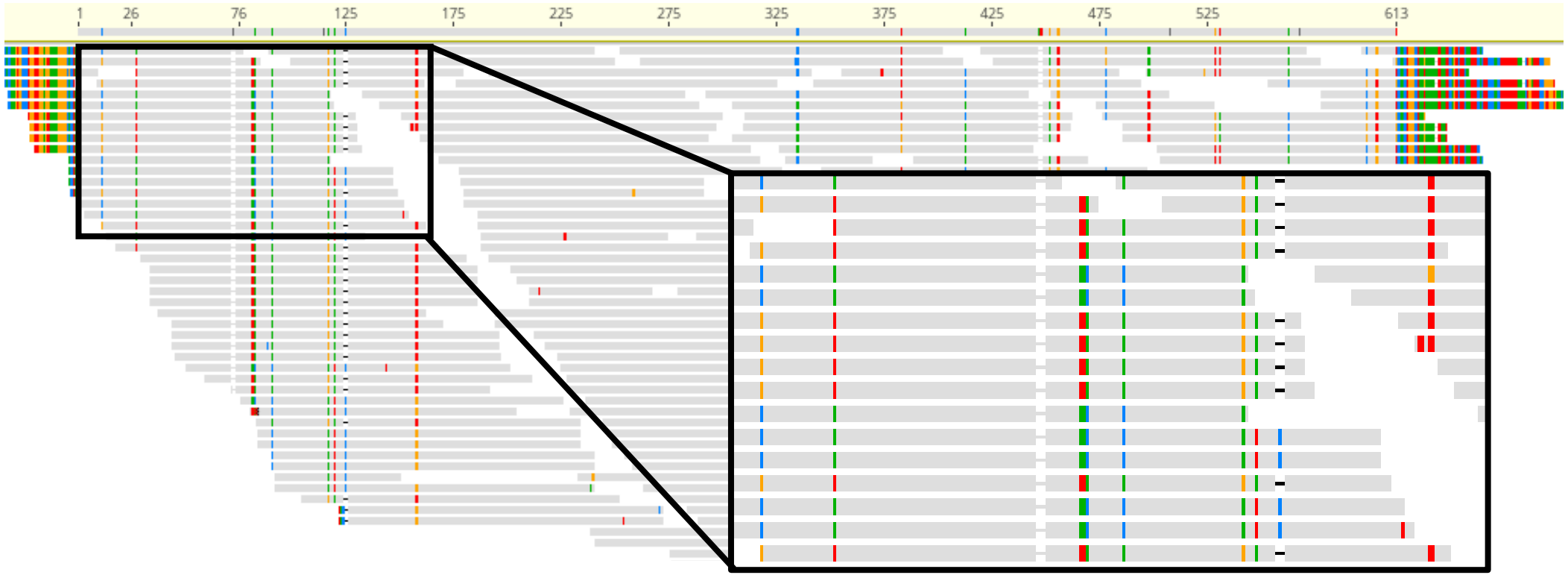


Allele divergence: 21 het. sites/613bp = 3.4%

Target Capture – Assembly of hybrids

Reference mapping

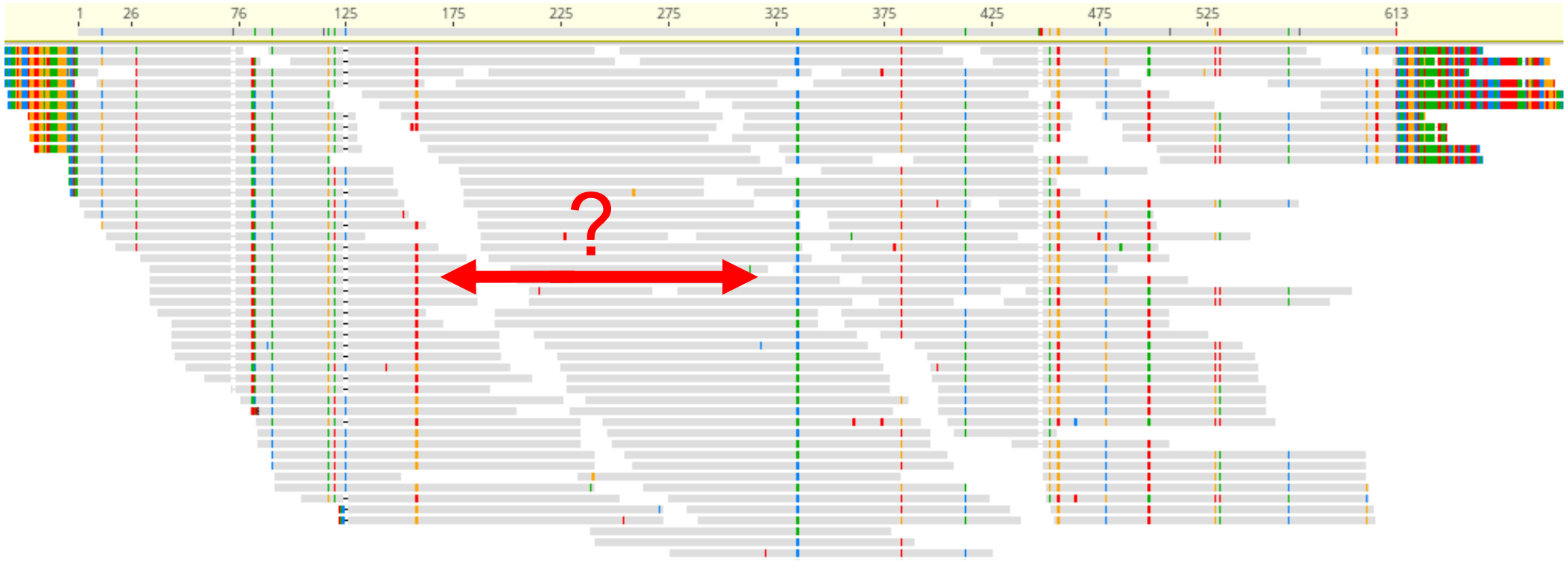
Heterozygous locus - hybrid



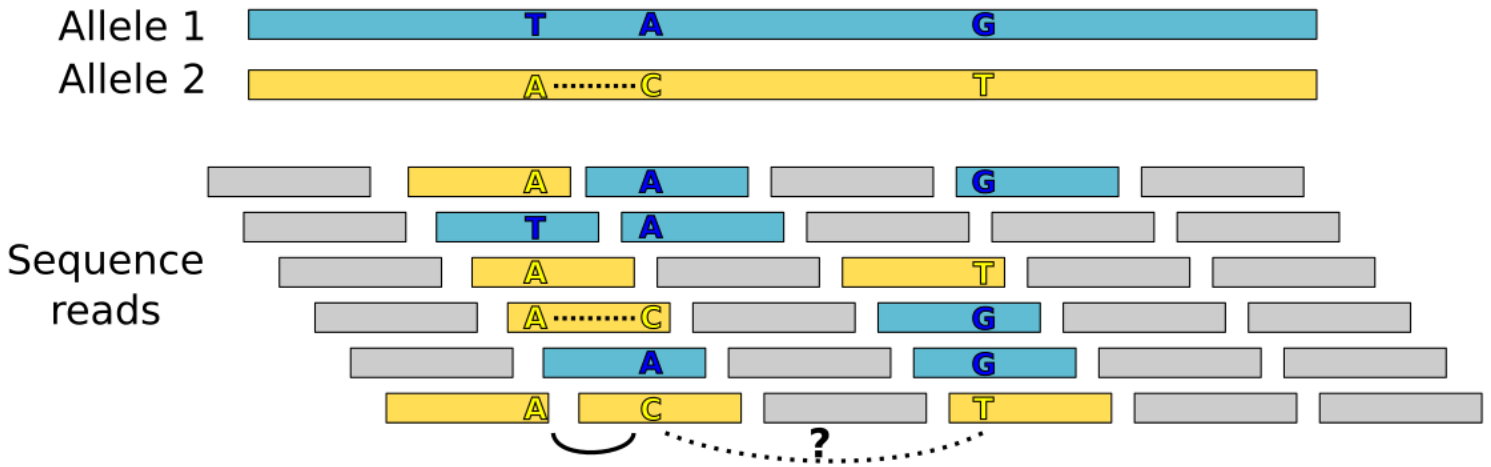
Target Capture – Assembly of hybrids

Reference mapping

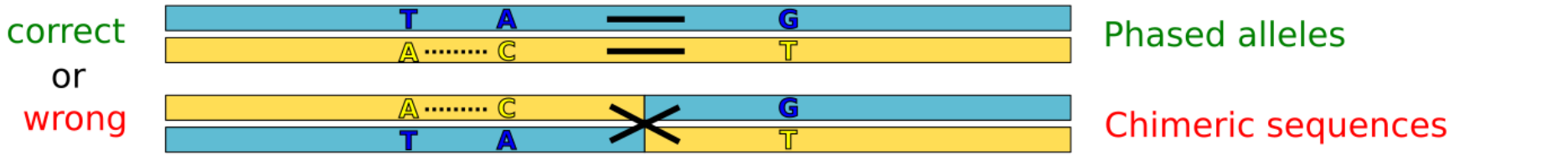
Heterozygous locus - hybrid



Target Capture – Assembly of hybrids



***De novo* assembly**

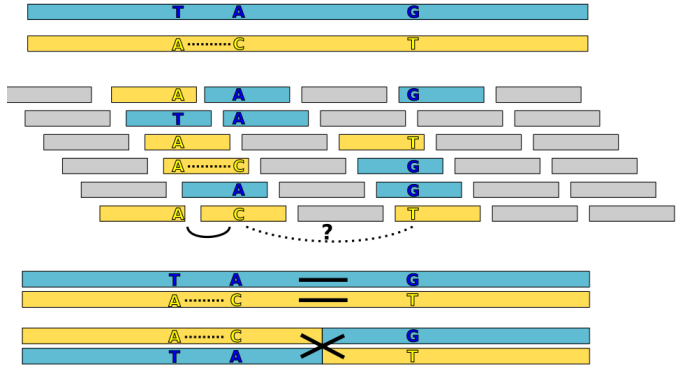


Reference mapping

Target Capture – Assembly of hybrids

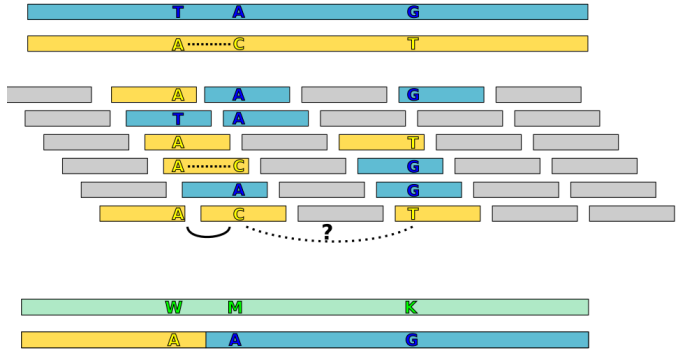
De novo assembly of hybrids

- Can potentially recover chimeric contigs
- How multiple contigs are treated depends on the workflow/assembler.
  - e.g. HybPiper uses the longest contig or the one closest to the target
- Read-backed phasing (actively looking for haplotypes) can improve results especially with paired-end reads and sufficient coverage
- No information on divergence in data or whether a contig can be chimeric



Target Capture – Assembly of hybrids

Reference mapping of hybrids



- Can not phase alleles!
- Majority rule consensus generates chimeric sequence
- Ambiguity coded consensus provides means of ‘removing’ / ‘disarming’ conflicting signal
- Ambiguity coded consensus sequences contain information on the divergence between reads

=> Recording of all SNPs per gene provides information on divergence in the dataset!

Target Capture – Assembly of hybrids

Assembly workflows (e.g., HybPiper, PHYLLUCE, SECAPR, custom…)

- Read selection per gene
- De novo assembly of reads

Allele phasing workflows

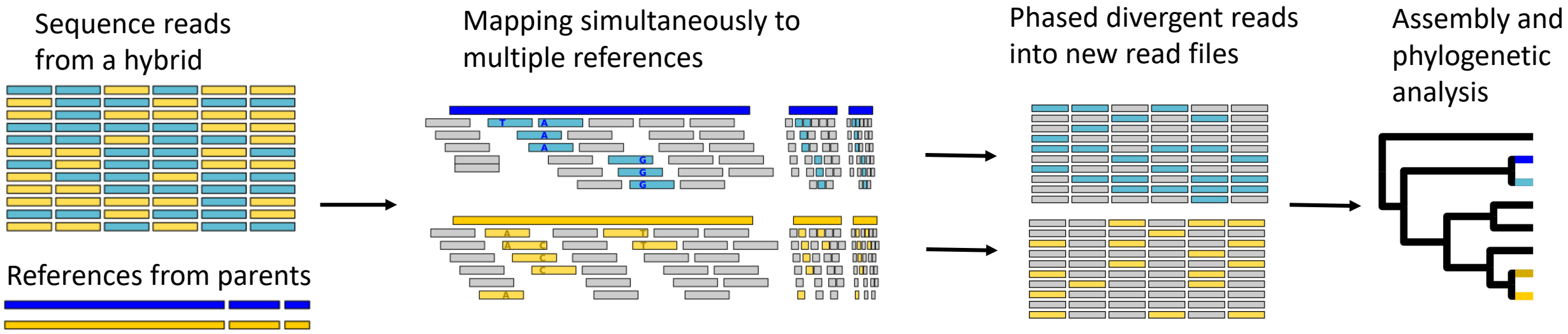
Kates et al. 2018, Andermann et al. 2019, Tiley et al. 2021

- *De novo* assembly with read backed phasing
- Only phased largest block (exon in gene)
- Phased alleles were not linked across haplotypes
- Mixed results

A different approach to phasing!

Phasing reads instead of alleles!

Mapping reads to multiple references simultaneously and separating reads into different read files.



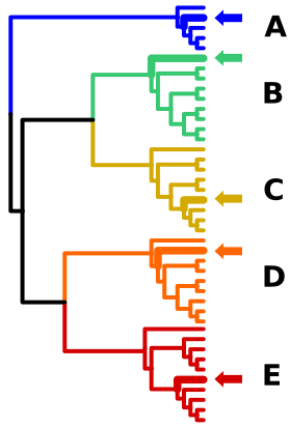
A different approach to phasing!

What if we do not know who the parents are?

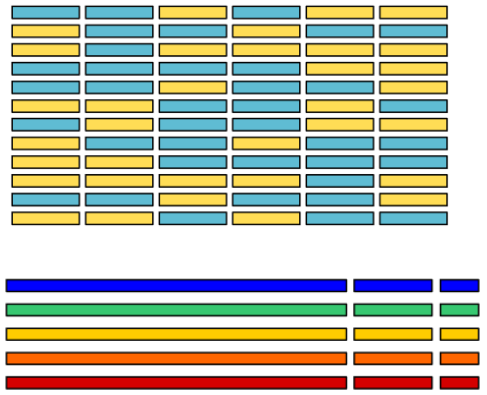
1. Map reads to multiple references covering all available clades!
2. Record to which clades the reads match.

Hybrids should have reads matching to the clades of their parents

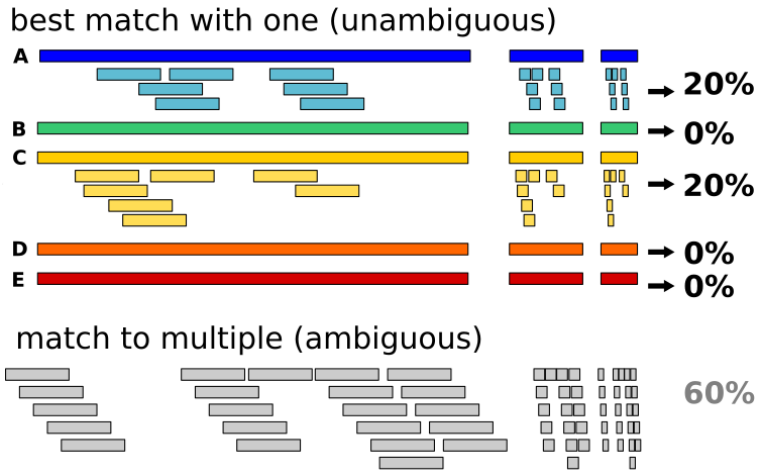
Standard assembly and phylogenetic analysis for reference selection



Mapping sequence reads to references from all clades



Analyse proportions of reads mapped to each reference



	A	B	C	D	E
acc1	20%	0	20%	0	0
acc2	...	...	...	...	...
acc3	...	...	...	...	...

A different approach to phasing!

Some advantages:

- Read phasing **before** the assembly
  - Avoids linking of heterozygous sites
  - Avoids linking of phased alleles
- Works with any amount of divergence (*depends on references)
- Works with higher level ploidy

Some disadvantages:

- Requires a phylogenetic framework to find suitable references
- There might be no suitable references available

HybPhaser

HybPhaser

Workflow for detection and phasing of hybrid accessions in target capture datasets

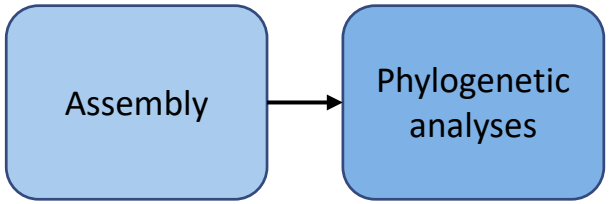
Utilizes mapping of reads to multiple references (simultaneously)

- Assess clade association of samples
- Phase read files

+ Analysis of SNPs in consensus sequences to detect hybrids and remove paralogs

- Builds on the assembly pipeline HybPiper
- Command line (bash) and R-scripts
- Uses freely available tools (Samtools, BBMap, …)
- Linux

Simple Workflow



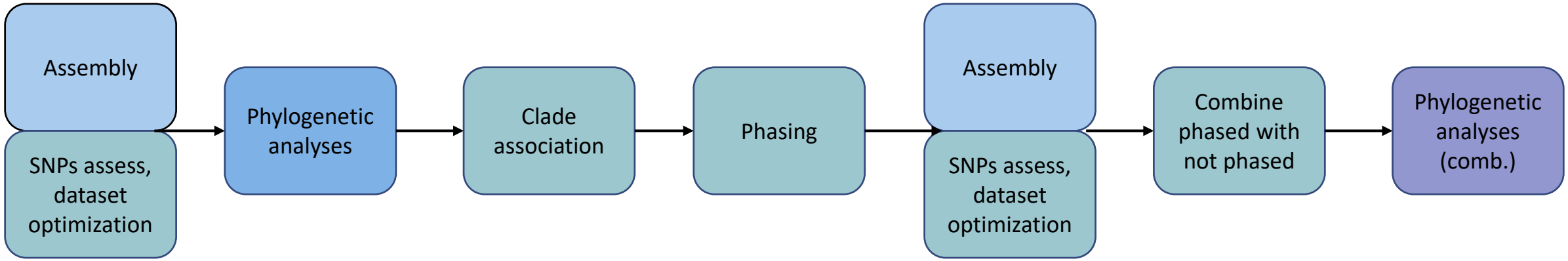
e.g. HybPiper

e.g. MAFFT, IQTREE

Contigs

Alignments, Phylogenies

HybPhaser Workflow



HybPiper + HybPhaser	e.g. MAFFT, IQTREE	HybPhaser	HybPhaser	HybPiper + HybPhaser	HybPhaser	e.g. MAFFT, IQTREE
Contigs, Consensus seqs. *Hybrid detection Paralog removal Reduction of missing data*	*Alignments, Framework phylogeny for reference selection*	*Relevant references for hybrid phasing*	*Phased accessions (read files)*	*Contigs, Consensus seqs.* *Hybrid detection Paralog removal Reduction of missing data*	*Combined dataset*	*Alignments, Phylogenies of combined dataset*

What type of study is HybPhaser for?

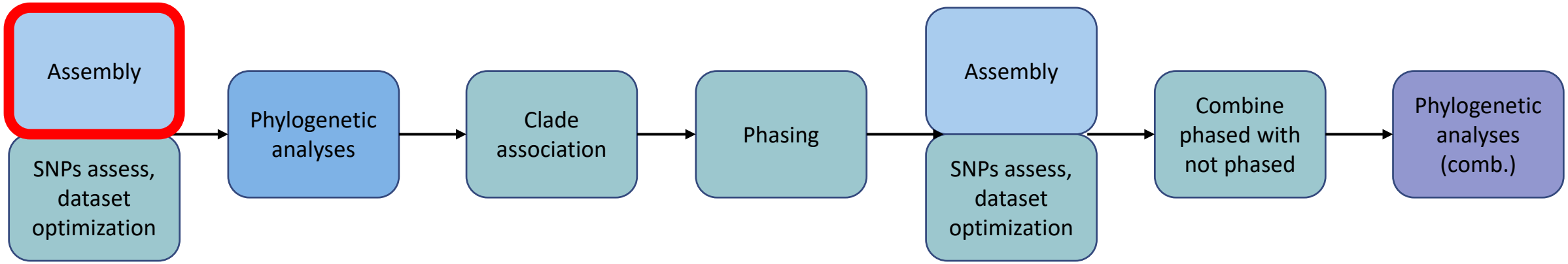
Phylogenetic study

- Reconstruction of relationships in a closely related clade, e.g. a genus, subgenus, tribe, family
- Relatively complete sampling (e.g., all major clades, most species)
- Might contain hybrids or allopolyploids

Dataset

- Target capture sequencing (e.g. Angiosperm353 baits)
- Nuclear loci
- Input: sequence reads and target sequence file

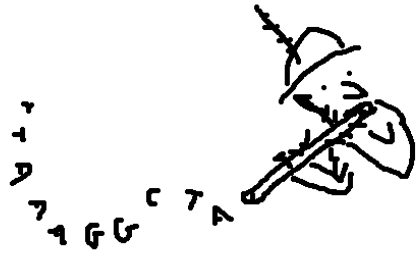
HybPhaser Workflow



HybPiper

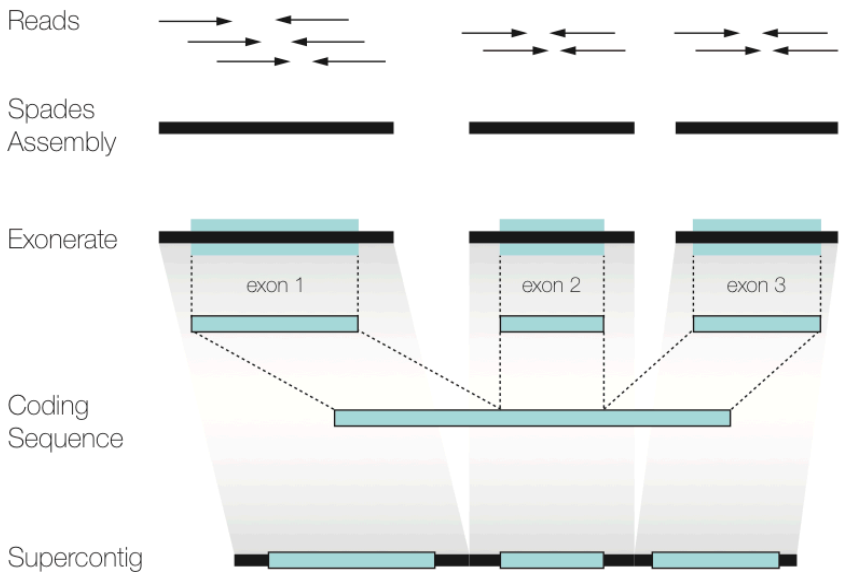
Contigs

Assembly with HybPiper



Assembly pipeline from Johnson et al. (2016)

- Contigs for each exon (*de novo*)
- Concatenated exons to genes
- Optionally include introns



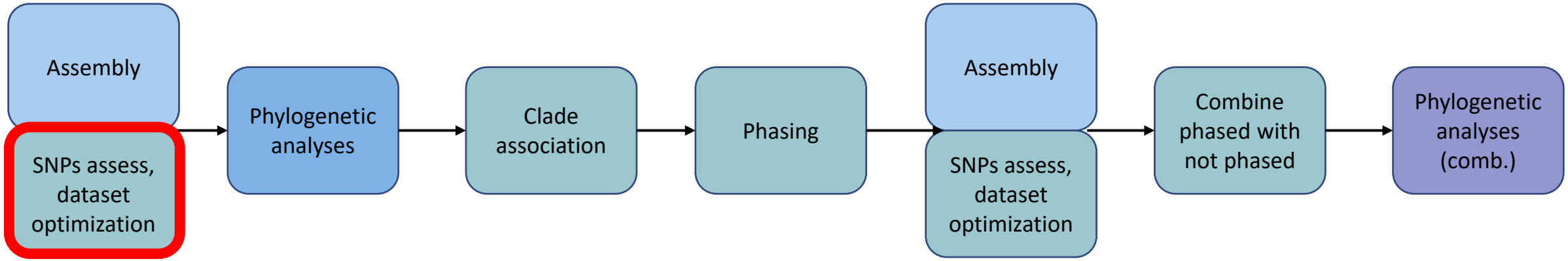
Improved version HybPiper-RBGV by Chris Jackson featured in the workshops!

- Containerized, bug fixes, more features

=> Workshop for TC assembly using HybPiper!!!

ASBS 2021 *virtually*

HybPhaser Workflow



HybPiper +
HybPhaser

Contigs,
Consensus seqs.

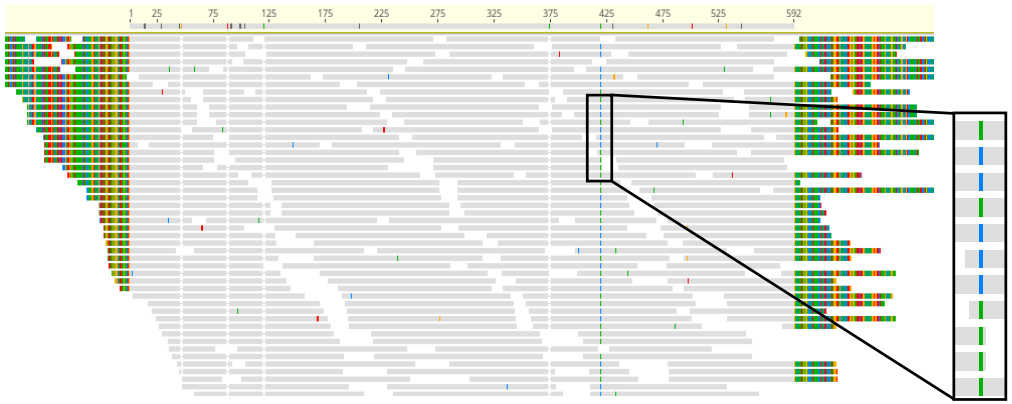
Hybrid detection
Paralog removal
Reduction of missing data

HybPhaser – SNPs assessment

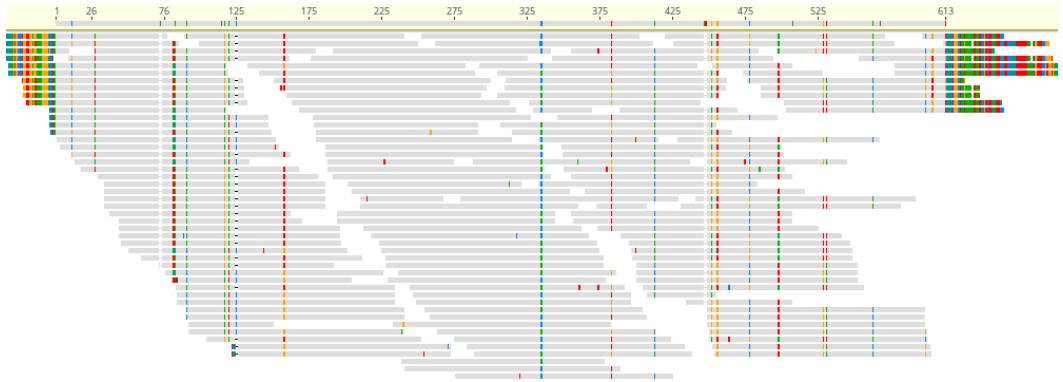
Assessment of SNPs across the dataset provides valuable insights:

- Detect hybrid accessions
- Estimate how close or distant related the parent of a putative hybrid
- Detect and remove paralogous genes

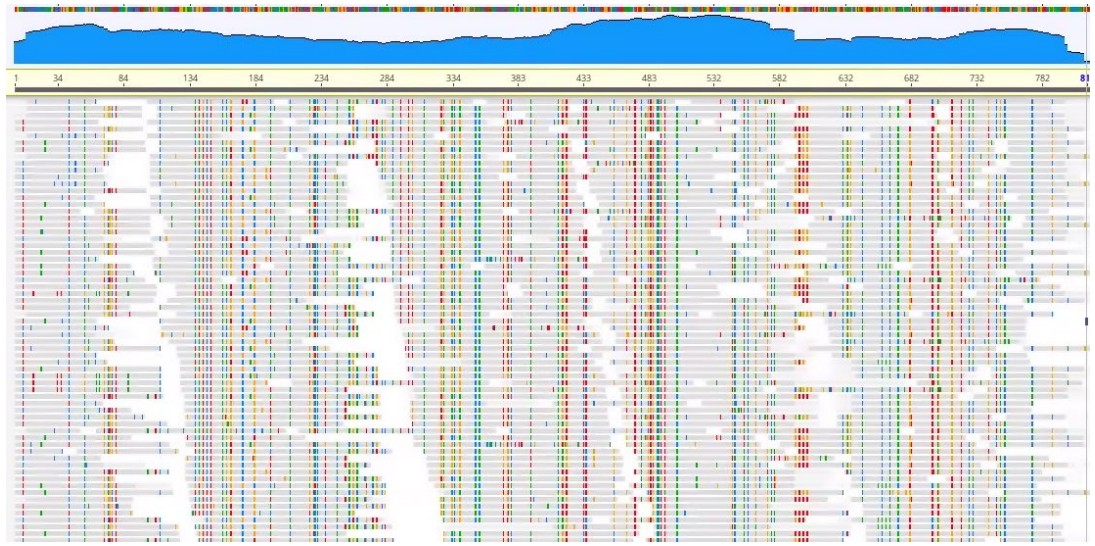
Allele Divergence (AD): % of SNPs



Non-hybrid alleles: 1 SNPs / 592 bp = 0.17%



Hybrid alleles : 21 SNPs / 613 bp = 3.4%



Highly paralogous / erroneous: 133 SNPs / 816 bp = 18.3%

Paralogous exon: 21 SNPs / 990 bp = 2.1 %

HybPhaser – SNPs assessment

Allele divergence - % of SNPs

- **Hybrids** should have high **AD** across all genes
  - amount correlates with the divergence between parents

    => compare **AD** between species

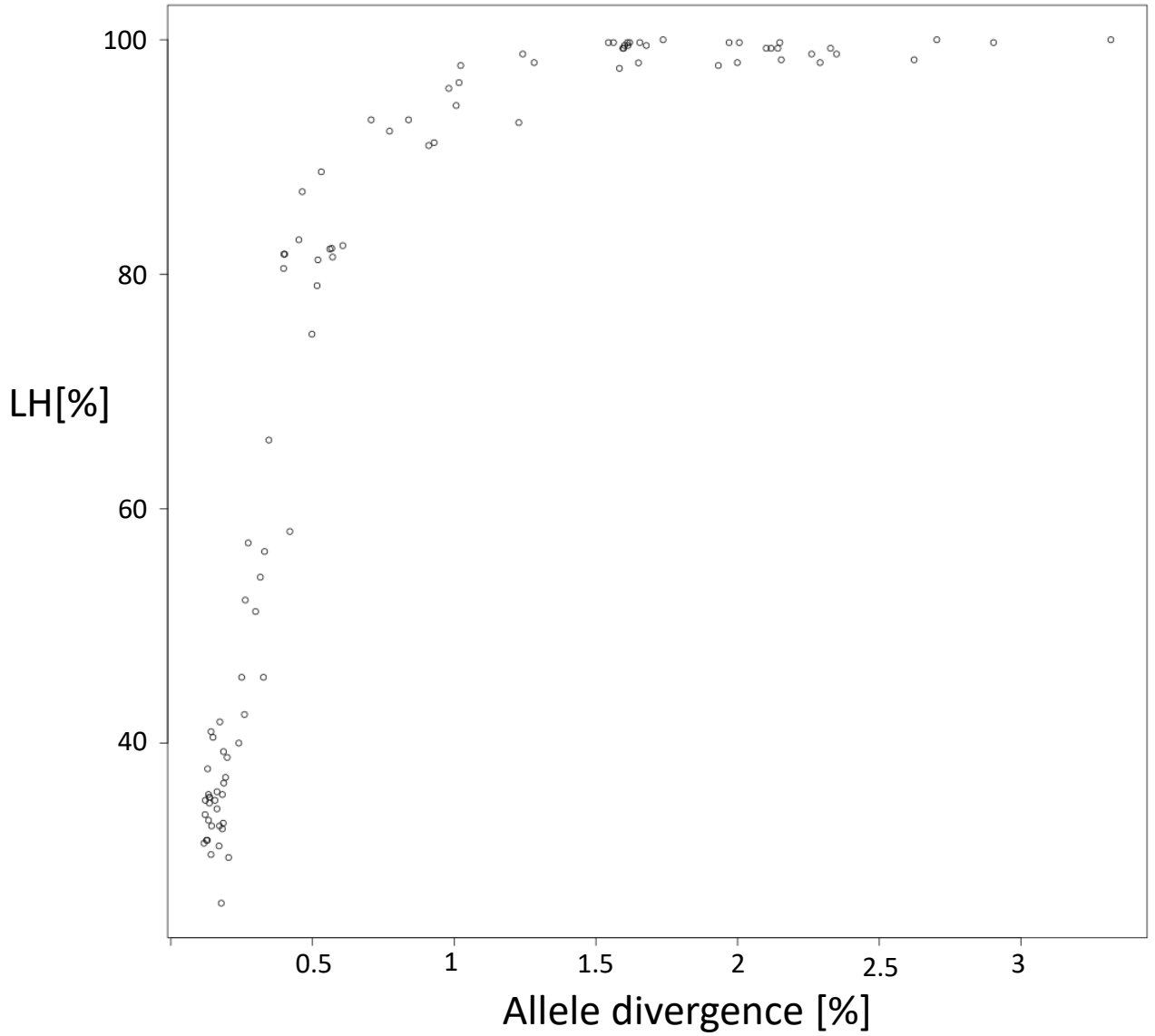
- **Paralogs** should have higher **AD** compared to normal genes

    => compare **AD** between genes

HybPhaser – SNPs assessment

Proportion of SNPs per **sample** to detect **hybrids**

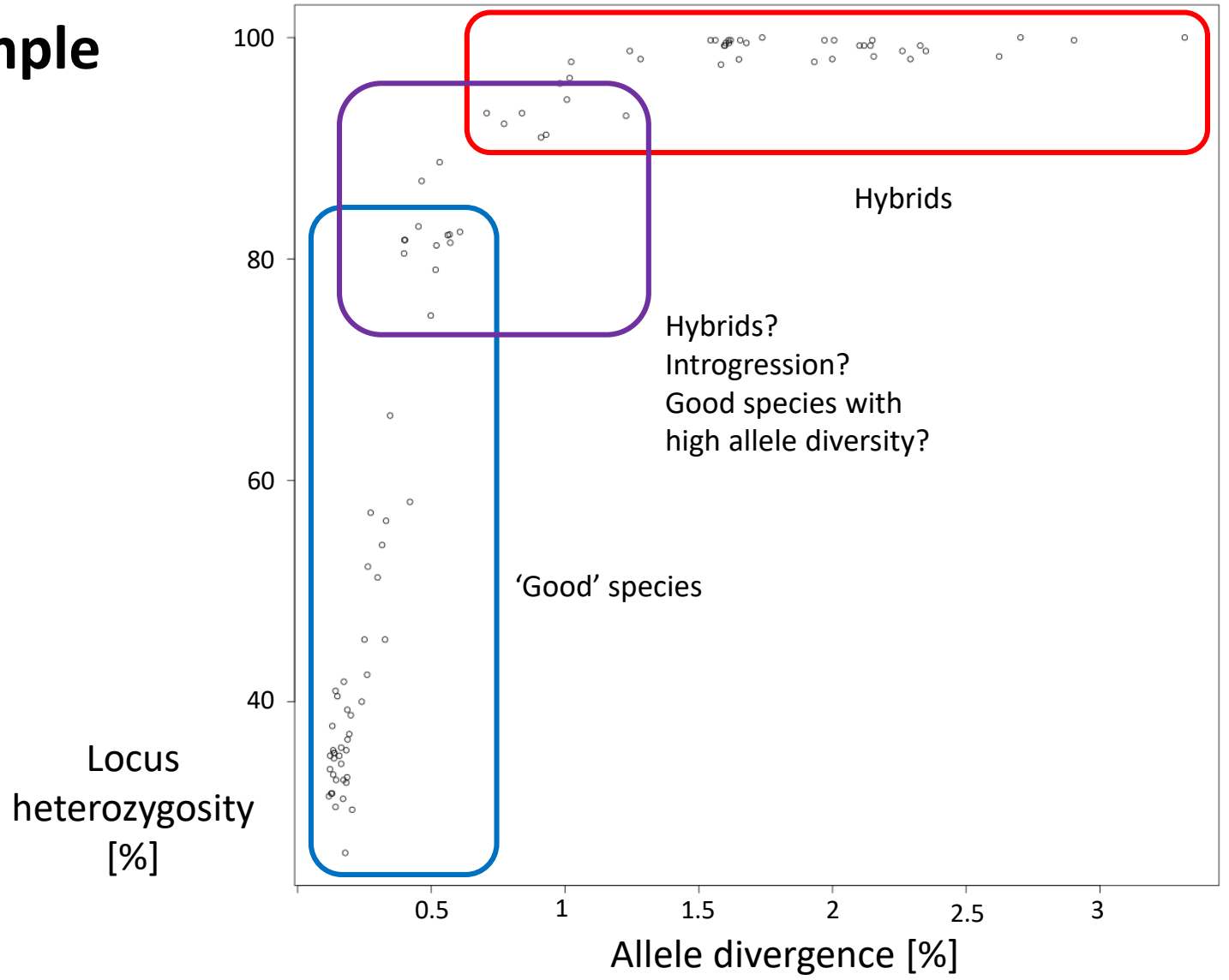
- Mean allele divergence
  (% of SNPs)
- Locus heterozygosity
  (% of loci with SNPs)



HybPhaser – SNPs assessment

Proportion of SNPs per **sample** to detect **hybrids**

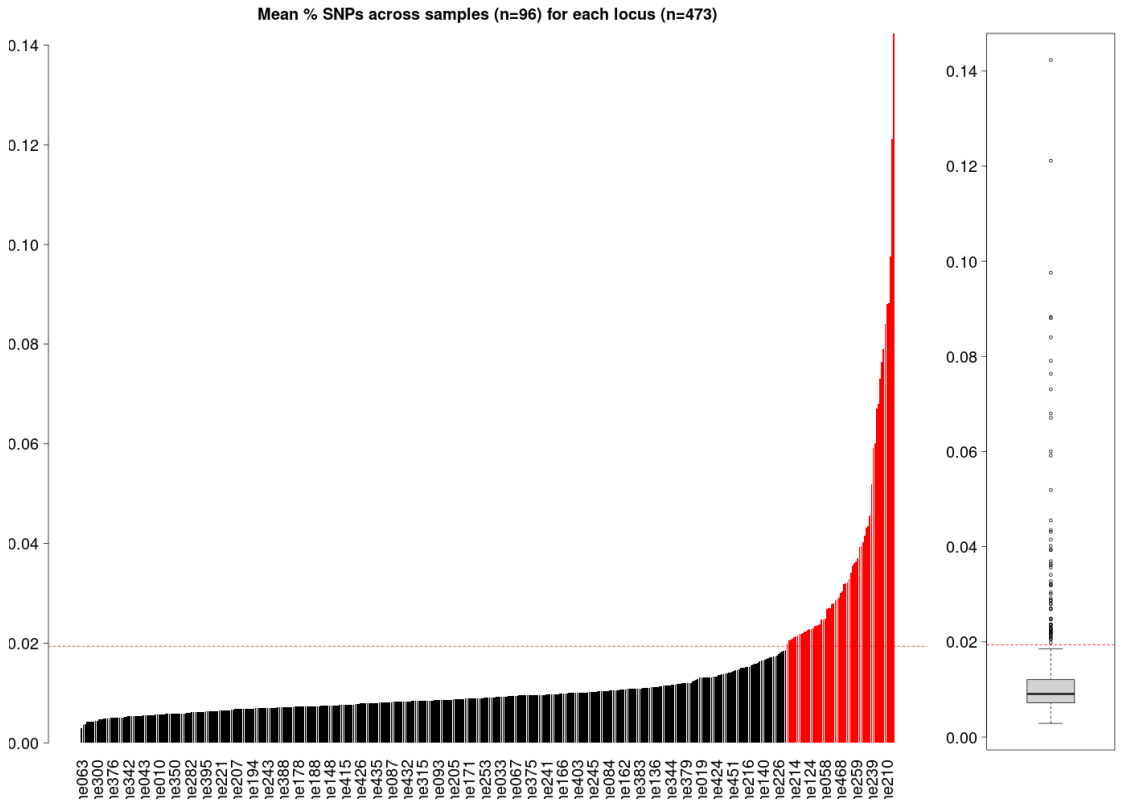
- Mean allele divergence
  (% of SNPs)
- Locus heterozygosity
  (% of loci with SNPs)



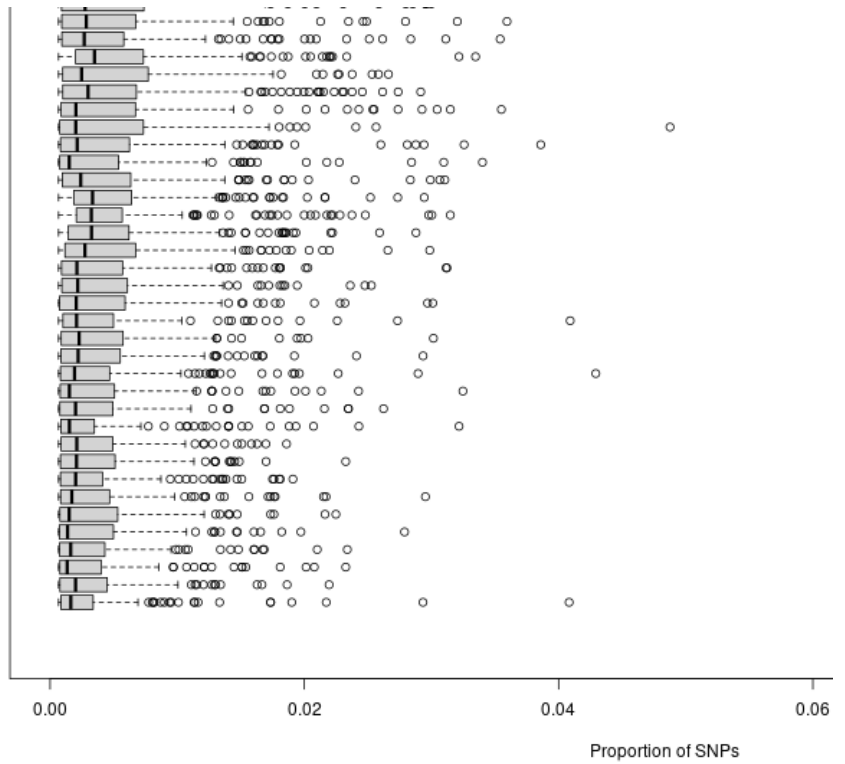
HybPhaser – SNPs assessment

Proportion of SNPs per **gene** to detect **paralogs**

average across all samples



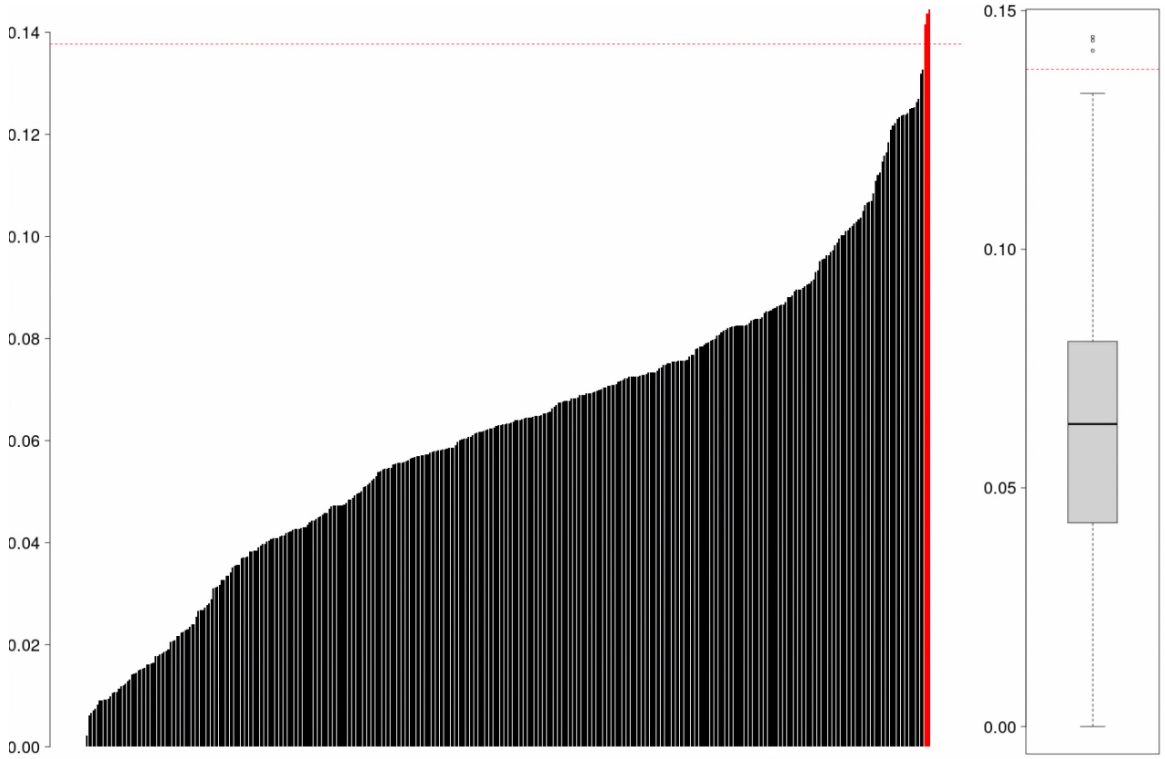
for each sample



HybPhaser – SNPs assessment

Proportion of SNPs per **gene** to detect **paralogs**

Too many paralogs!



=> Workshop for paralogy resolution !!!

ASBS 2021 virtually

Expressions of interest to participate close this Sunday June 13th!!!
https://asbs2021.bablglobal.com/workshop/

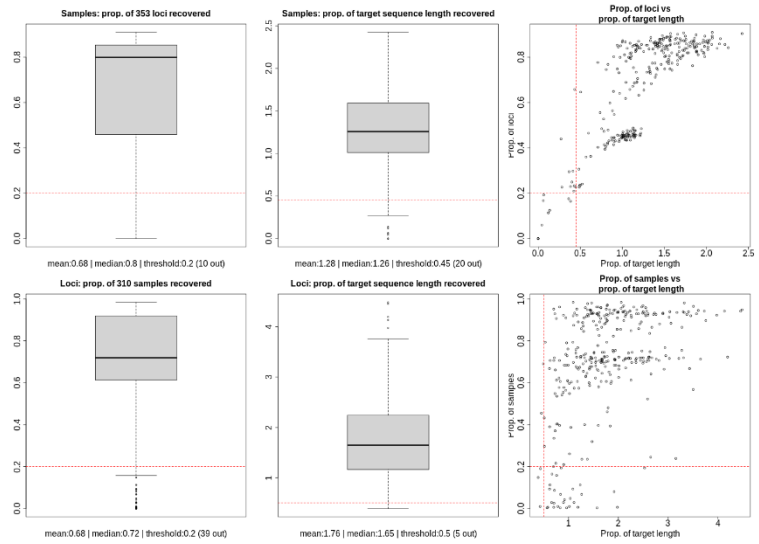
HybPhaser – Dataset optimisation

Reducing missing data

Removal of poorly recovered loci / samples

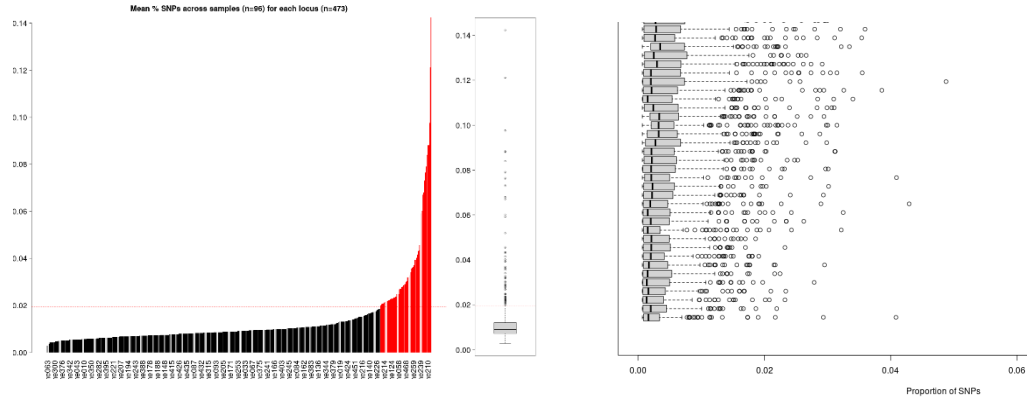
Paralogs

Removal of outliers across all samples / for each sample

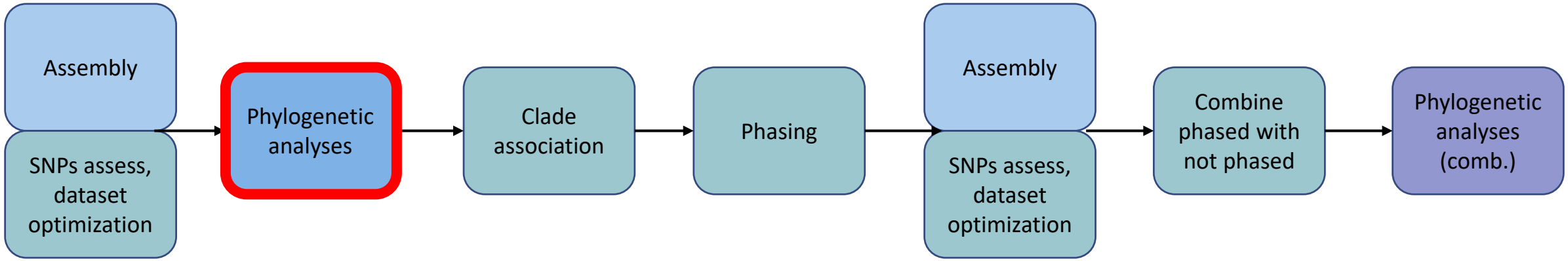


Output:

- contigs/consensus
- raw / optimized
- per locus/sample



HybPhaser Workflow



HybPiper + HybPhaser

e.g. MAFFT, IQTREE

Contigs, Consensus seqs.

Hybrid detection
Paralog removal
Reduction of missing data

Alignments, Framework phylogeny for reference selection

Tree Reconstruction for Target Capture

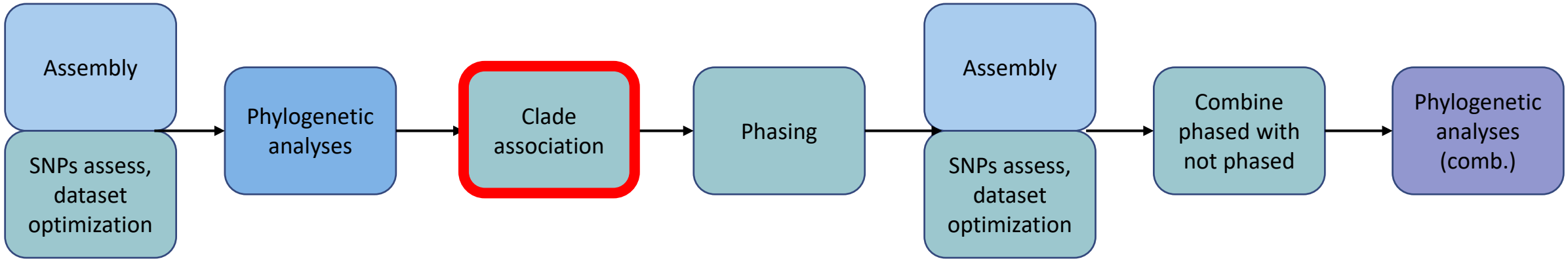
Input:

- **Contigs** or **consensus** sequences (with ambiguity codes)
- **Cleaned** from paralogs and reduced missing data

Quick recommendation for phylogenetics with TC data:

- Supermatrix and gene tree summary
- Bootstrap and concordance factors
- Networks are useful too

HybPhaser Workflow



HybPiper + HybPhaser

Contigs, Consensus seqs.

Hybrid detection
Paralog removal
Reduction of missing data

e.g. MAFFT, IQTREE

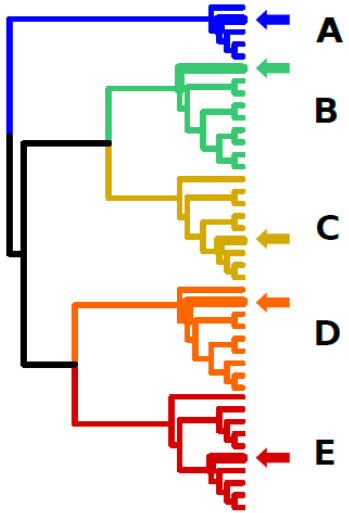
Alignments, Framework phylogeny for reference selection

HybPhaser

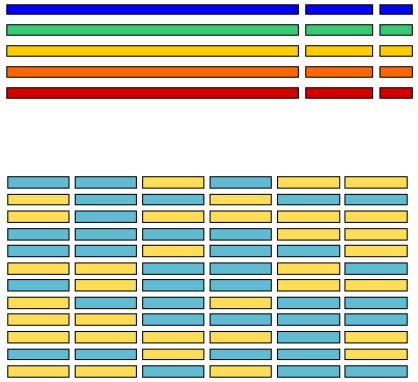
Relevant references for hybrid phasing

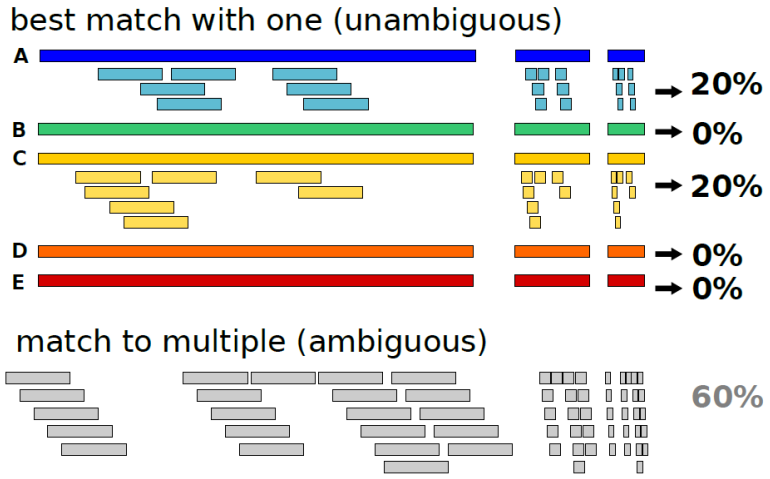
HybPhaser – Clade association

Selection of clade references



Mapping of reads to all clade references simultaneously (BBSplit)





Output

Clade association table

	A	B	C	D	E
acc1	20%	0	20%	0	0
acc2	...	...	...	...	...
acc3	...	...	...	...	...

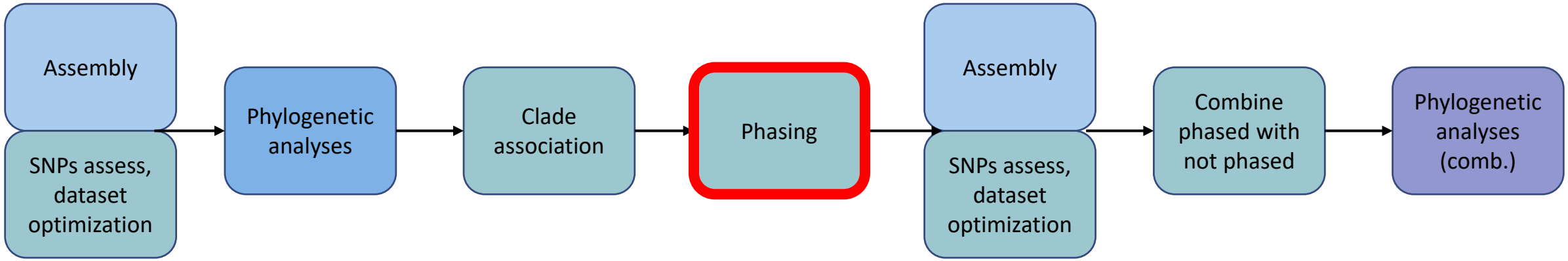
HybPhaser – Clade association

Output:

Proportions of reads that match **unambiguously** to a single reference!

samples	LH	AD	*Ref1*	*Ref2*	*Ref3*	*Ref4*	*Ref5*	*Ref6*	*Ref7*	*Ref8*	*Ref9*	*Ref10*	*Ref11*	*Ref12*	*Ref13*	*Ref14*	*Ref15*	*Ref16*
Sample-01	99%	2.00%	0.5	0.7	0.2	0.2	0.2	0.4	0.4	0.7	10.3	0.4	0.4	0.5	0.3	0.6	0.5	6.8
Sample-02	99%	2.12%	0.8	1.0	0.2	0.3	0.2	0.6	0.5	0.5	0.7	4.9	0.5	0.5	0.3	0.6	0.5	6.6
Sample-03	29%	0.14%	1.1	1.2	0.3	0.4	0.3	0.8	0.6	0.6	0.9	12.4	0.6	0.4	0.3	0.3	0.2	0.2
Sample-04	50%	0.30%	0.5	0.7	0.2	0.2	0.2	0.4	0.4	0.3	0.5	0.4	0.5	1.2	11.1	0.9	1.0	1.6
Sample-05	100%	1.55%	0.4	0.6	0.2	0.2	0.1	0.4	0.3	0.3	0.4	0.2	0.4	11.0	0.9	0.6	6.1	0.7
Sample-06	45%	0.27%	0.4	0.7	0.1	0.2	0.2	0.3	0.4	0.3	0.5	0.3	0.4	0.8	17.2	0.6	0.4	0.2
Sample-07	99%	2.20%	14.4	0.7	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4	0.3	0.4	0.2	0.5	0.4	7.0
Sample-08	79%	0.51%	0.5	0.7	0.2	0.2	0.2	0.4	0.4	0.3	0.5	0.4	0.5	1.2	10.1	0.8	0.8	1.4
Sample-09	28%	0.12%	0.5	0.7	0.2	0.2	0.1	0.3	0.3	0.3	0.5	0.3	0.3	0.6	0.4	0.9	0.7	9.9
Sample-10	99%	1.22%	0.8	1.1	14.1	2.0	1.6	0.6	0.6	0.5	0.7	0.7	0.8	0.5	0.3	0.3	0.2	0.2
Sample-11	92%	1.09%	1.0	1.2	1.9	2.2	8.6	0.7	0.6	0.6	0.8	0.7	0.8	0.5	0.3	0.3	0.2	0.2
Sample-12	100%	1.53%	0.4	0.6	0.2	0.2	0.1	0.3	0.4	0.3	0.4	0.2	0.4	11.2	0.4	0.6	6.8	0.6
Sample-13	99%	1.06%	1.2	1.3	0.3	0.4	0.3	3.4	0.7	0.7	1.0	8.7	0.7	0.5	0.4	0.4	0.1	0.2
Sample-14	25%	0.10%	0.5	0.8	0.2	0.2	0.2	0.4	0.4	0.3	0.6	0.4	0.6	0.9	1.2	2.1	2.8	3.4
Sample-15	89%	0.86%	1.2	1.5	0.5	0.7	0.4	0.9	0.8	1.5	1.6	2.7	1.1	0.6	0.6	0.4	0.3	0.3
Sample-16	22%	0.35%	1.0	1.4	0.5	0.7	0.5	1.0	0.7	1.4	1.5	2.6	1.0	0.7	0.6	0.4	0.3	0.3
Sample-17	26%	0.11%	0.4	0.7	0.2	0.2	0.2	0.4	0.4	0.3	0.5	0.4	25.3	0.4	0.2	0.2	0.1	0.1

HybPhaser Workflow



HybPiper + HybPhaser

Contigs, Consensus seqs.

Hybrid detection
Paralog removal
Reduction of missing data

e.g. MAFFT, IQTREE

Alignments, Framework phylogeny for reference selection

HybPhaser

Relevant references for hybrid phasing

HybPhaser

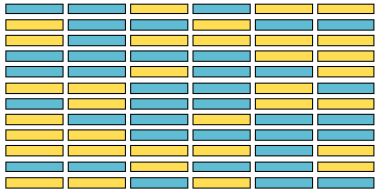
Phased accessions (read files)

HybPhaser – Phasing

Input

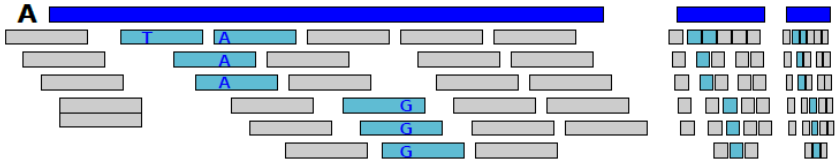
Selected clade references

Sequence reads

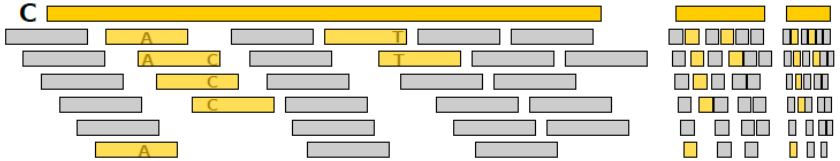


Mapping of sequence reads to only the relevant references

match only to A + match multiple

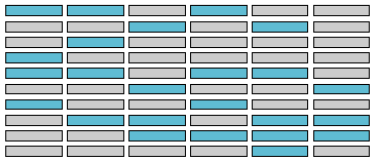


match only to C + match multiple

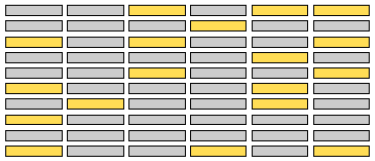


Generating new read files that contain phased reads

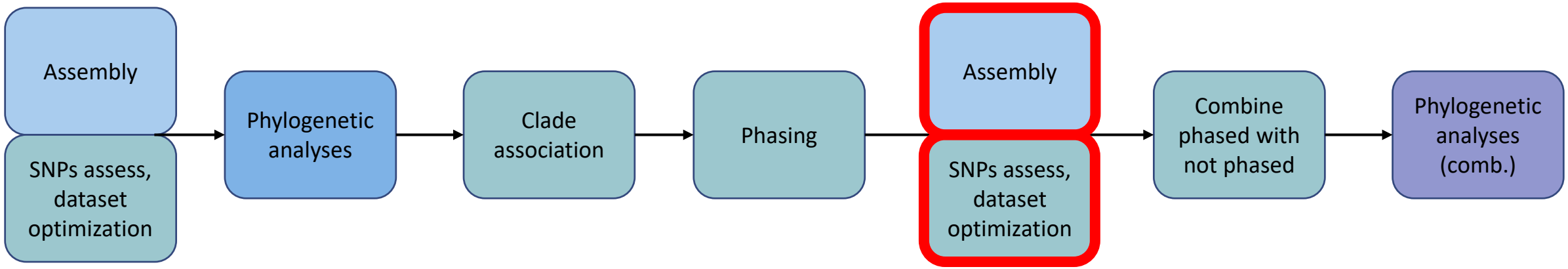
Phased accession 1



Phased accession 2



HybPhaser Workflow



HybPiper + HybPhaser

Contigs, Consensus seqs.

Hybrid detection
Paralog removal
Reduction of missing data

e.g. MAFFT, IQTREE

Alignments, Framework phylogeny for reference selection

HybPhaser

Relevant references for hybrid phasing

HybPhaser

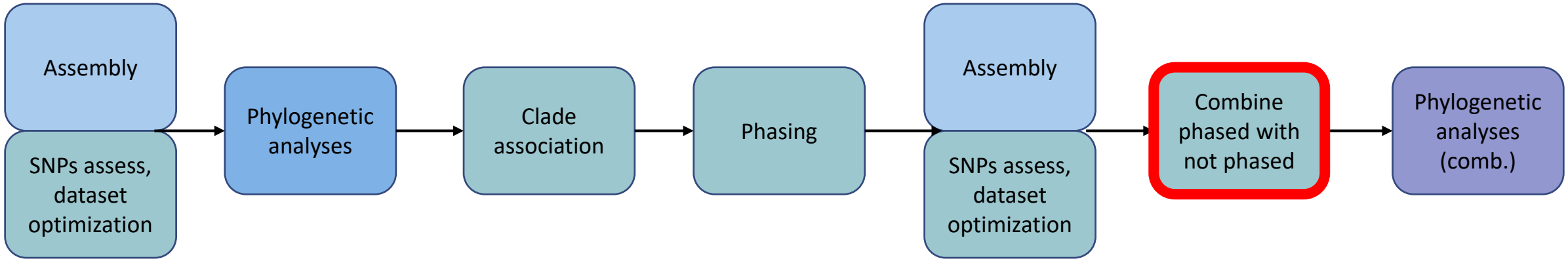
Phased accessions (read files)

HybPiper + HybPhaser

Contigs, Consensus seqs.

Hybrid detection
Paralog removal
Reduction of missing data

HybPhaser Workflow



HybPiper + HybPhaser

Contigs, Consensus seqs.

Hybrid detection
Paralog removal
Reduction of missing data

e.g. MAFFT, IQTREE

Alignments, Framework phylogeny for reference selection

HybPhaser

Relevant references for hybrid phasing

HybPhaser

Phased accessions (read files)

HybPiper + HybPhaser

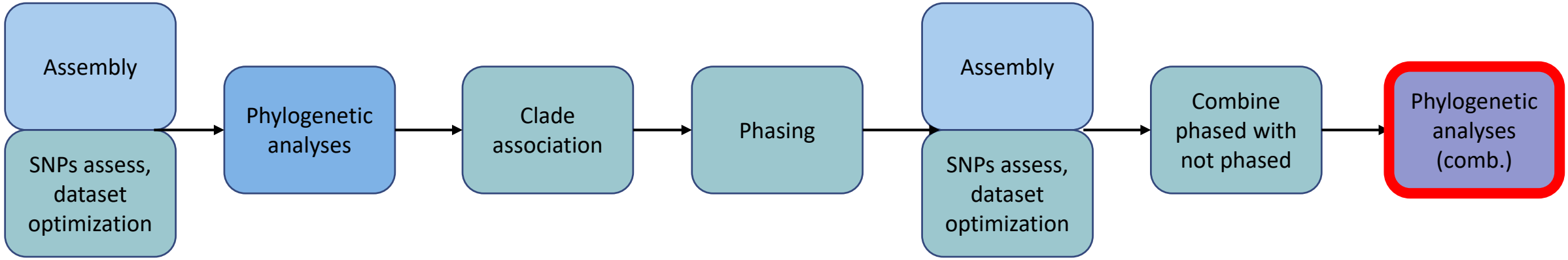
Contigs, Consensus seqs.

Hybrid detection
Paralog removal
Reduction of missing data

HybPhaser

Combined dataset

HybPhaser Workflow



HybPiper + HybPhaser

Contigs, Consensus seqs.

Hybrid detection
Paralog removal
Reduction of missing data

e.g. MAFFT, IQTREE

Alignments, Framework phylogeny for reference selection

HybPhaser

Relevant references for hybrid phasing

HybPhaser

Phased accessions (read files)

HybPiper + HybPhaser

Contigs, Consensus seqs.

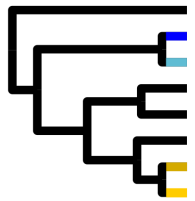
Hybrid detection
Paralog removal
Reduction of missing data

HybPhaser

Combined dataset

e.g. MAFFT, IQTREE

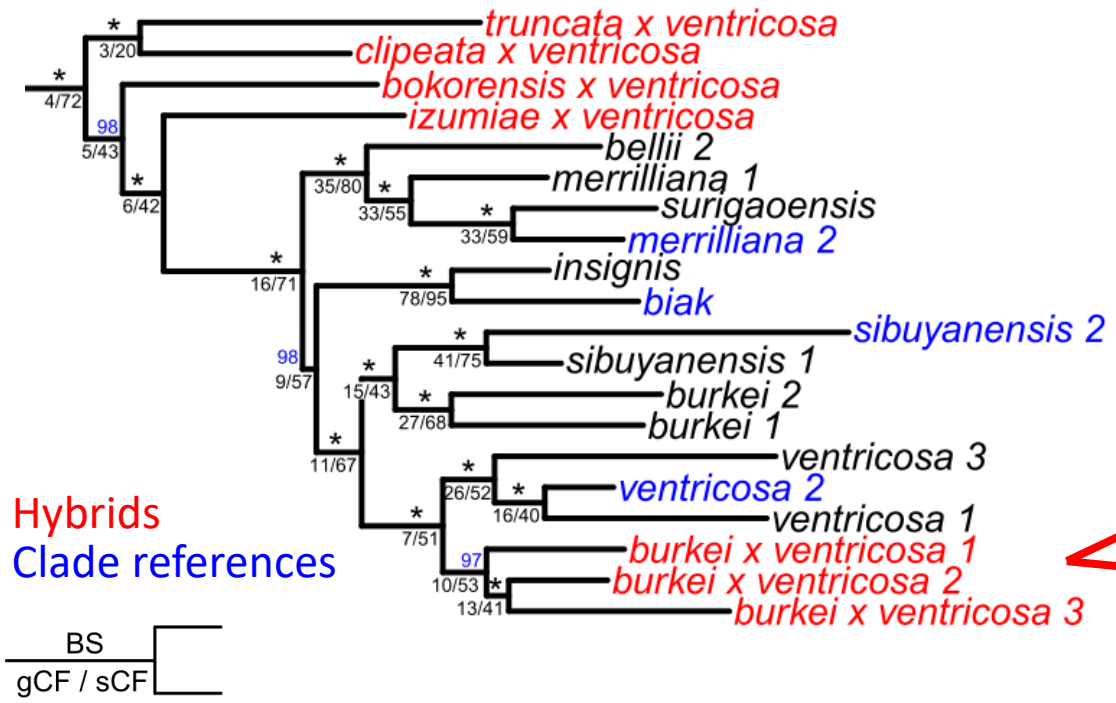
Alignments, Phylogenies of combined dataset



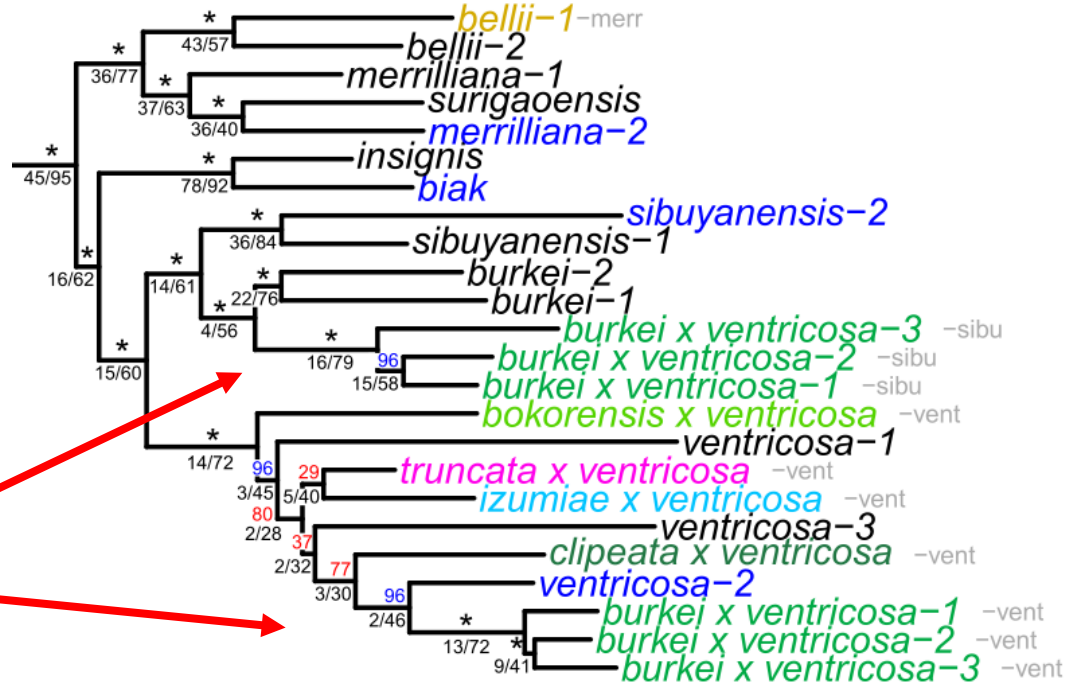
HybPhaser – Resulting phylogenies

- Phased accessions group with parental lineages

Original



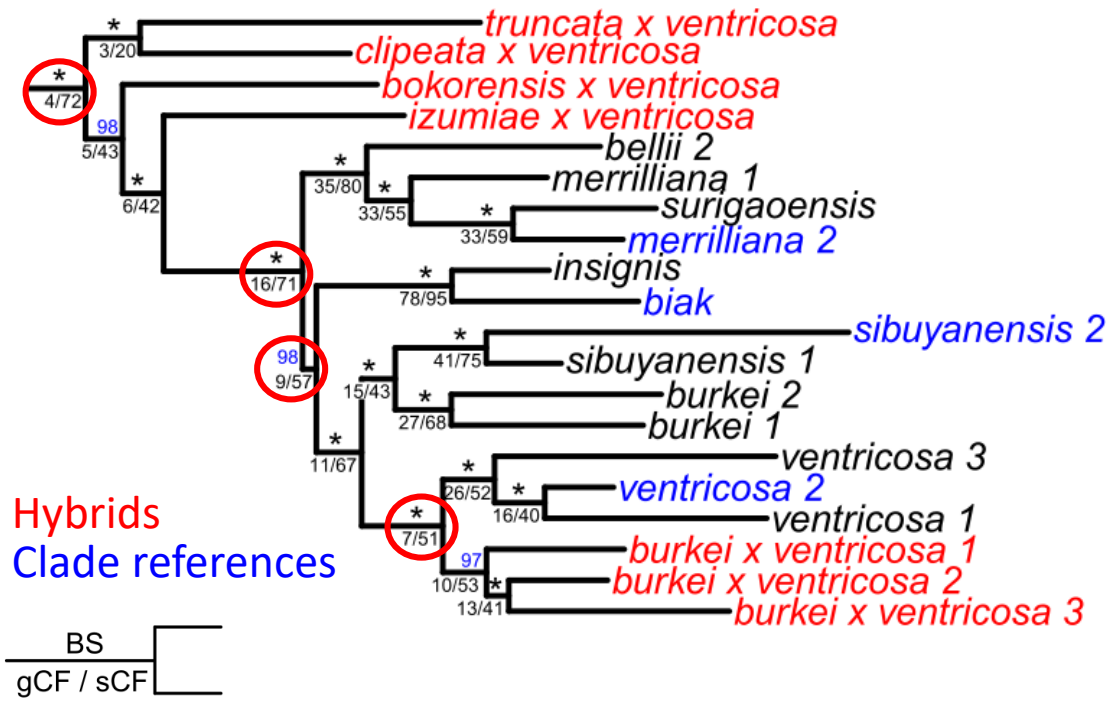
Phased accessions included



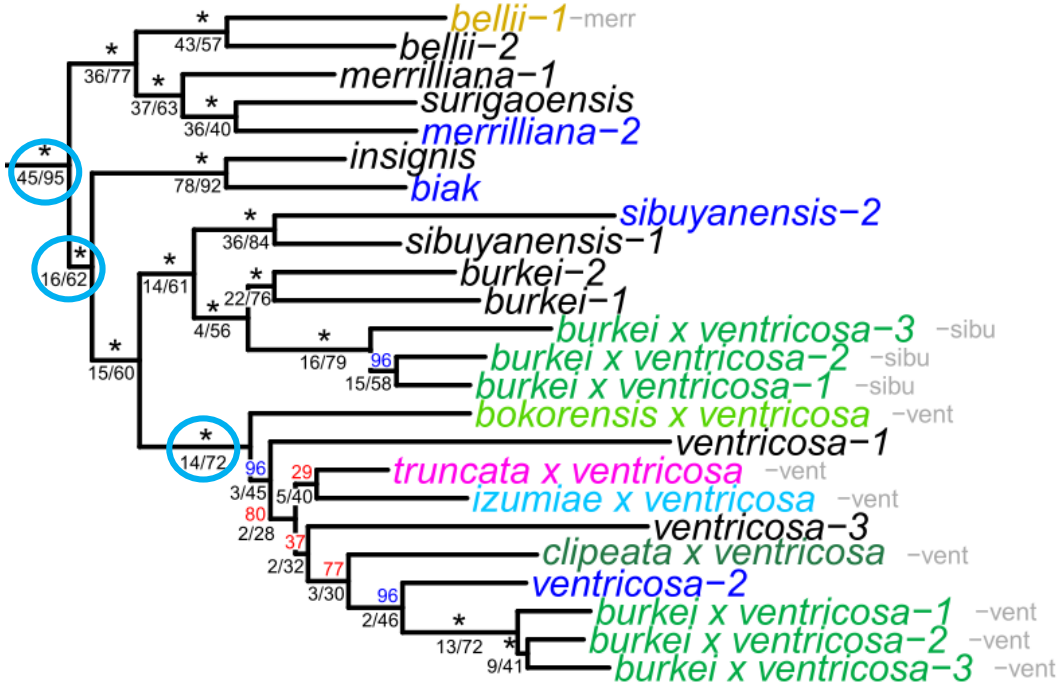
HybPhaser – Resulting phylogenies

- Improved clade support of relevant nodes

Original



Phased accessions included



More information HybPhaser

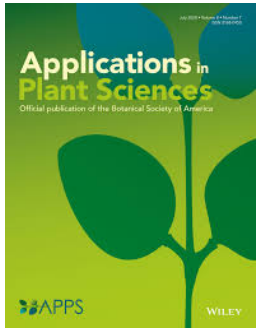
More detailed information is available on GitHub

https://github.com/LarsNauheimer/HybPhaser

Publication is available as preprint on bioRxiv,

https://www.biorxiv.org/content/10.1101/2020.10.27.354589v2

and soon also in Applications for Plant Sciences



Please email me, if you have any questions!

Lars.Nauheimer@jcu.edu.au

References

- Abbott, R., D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, N. Bierne, J. Boughman, et al. 2013. Hybridization and speciation. *Journal of Evolutionary Biology* 26: 229–246.
- Andermann, T., A. M. Fernandes, U. Olsson, M. Töpel, B. Pfeil, B. Oxelman, A. Aleixo, et al. 2019. Allele phasing greatly improves the phylogenetic utility of ultraconserved elements. *Systematic Biology* 68: 32–46.
- Barker, M. S., N. Arrigo, A. E. Baniaga, Z. Li, and D. A. Levin. 2016. On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* 210: 391–398.
- Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.
- Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and N. J. Wickett. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4: 1600016.
- Kates, H. R., M. G. Johnson, E. M. Gardner, N. J. C. Zerega, and N. J. Wickett. 2018. Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of Artocarpus. *American Journal of Botany* 105: 404–416.
- Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution* 20: 229–237.
- Paun, O., F. Forest, M. F. Fay, and M. W. Chase. 2009. Hybrid speciation in angiosperms: parental divergence drives ploidy. *New Phytologist* 182: 507–518.
- Peer, Y. V. de, E. Mizrachi, and K. Marchal. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18: 411–424.
- Sang, T., and D. Zhang. 1999. Reconstructing hybrid speciation using sequences of low copy nuclear genes: hybrid origins of five Paeonia species based on adh gene phylogenies. *Systematic Botany* 24: 148–163.
- Soltis, D. E., C. J. Visger, D. B. Marchant, and P. S. Soltis. 2016. Polyploidy: Pitfalls and paths to a paradigm. *American Journal of Botany* 103: 1146–1166.
- Tiley, G. P., A. A. Crowl, P. S. Manos, E. B. Sessa, C. Solís-Lemus, A. D. Yoder, and J. G. Burleigh. 2021. Phasing Alleles Improves Network Inference with Allopolyploids. Evolutionary Biology.

Thank you very much!!!

Don’t forget about to check out the workshops
(5th – 8th July)

1) Assembly of raw reads using HybPiper

2) Paralogy resolution using Yang and Smith

3) Detection and phasing of hybrids using HybPhaser

Australasian Systematic Botany Society
Annual Conference 2021

12th-16th July

Expressions of interest to participate close this Sunday June 13th!!! https://asbs2021.bablglobal.com/workshop/

Genomics for Australian Plants

BIOPLATFORMS AUSTRALIA

Australian BioCommons

ATH Australian Tropical Herbarium