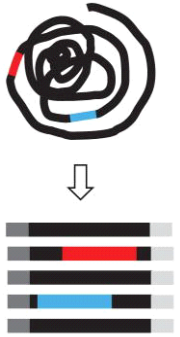
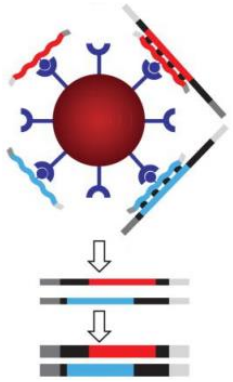


Target enrichment for plant/animal systematics - methodological workshop -

Data analysis pipelines



Roswitha Schmickl, Tomáš Fér, Vojtěch Zeisek, Soňa Píšová

Dept. of Botany, Charles University, Prague

9th-17th June 2025

Hyb-Seq data analysis pipelines

- Reference-based assembly

- Read mapping against a **reference (library)**; all reads should be orthologous to a given reference sequence, while at the same time avoiding reads from paralogous genomic regions

- **User-defined sequence similarity thresholds**, based on how similar the reads are expected to match the reference sequence
-
- Reference library can consist of a collection of individual reference sequences for the targeted loci (usually exons) or of a complete reference genome (plastid genome)

- Read mapping to be done with BWA, Bowtie, Minimap

Hyb-Seq data analysis pipelines

- De novo assembly

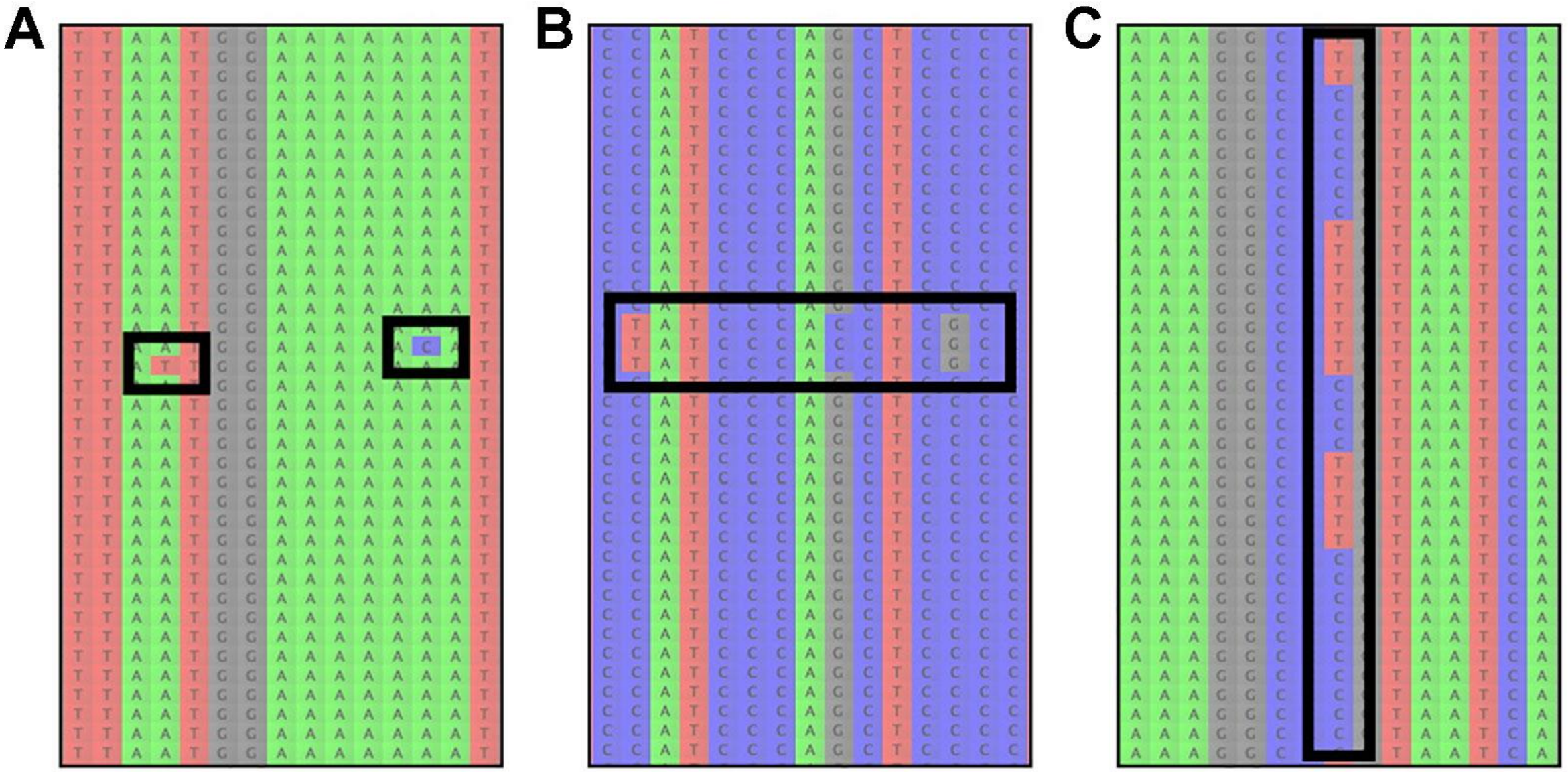
- **Reads with sequence overlap are assembled** into continuously growing clusters of reads (contigs) which are then collapsed into a single contig consensus sequence for each cluster

- Spades used for de novo assembly
- In order to **extract and annotate the contigs that represent targeted loci**, a common approach is to run a BLAST search between the contig sequences on the one hand and the probe sequences or some other collection of reference sequences on the other hand

- Pipelines aTRAM, HYBPIPER, PHYLUCE, and SECAPR all contain functions that employ some BLAST algorithm to match the assembled contigs

- **Optimally, de novo and reference-assembly approaches are used in conjunction, iteratively** (such as a first reference-based mapping, followed by de novo assembly)

Most common sources of read variation



Sequencing error

Paralogy

Allelic variation

HybPiper

(reference-based mapping, de novo assembly)

▸ Appl Plant Sci. 2016 Jul 12;4(7):apps.1600016. doi: 10.3732/apps.1600016

HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment[1]

Matthew G Johnson [2,6], Elliot M Gardner [2,3], Yang Liu [4], Rafael Medina [4], Bernard Goffinet [4], A Jonathan Shaw [5], Nyree J C Zerega [2,3], Norman J Wickett [2,3]

▸ Author information ▸ Article notes ▸ Copyright and License information

PMCID: PMC4948903 PMID: 27437175

Abstract

Premise of the study:

Using sequence data generated via target enrichment for phylogenetics requires reassembly of high-throughput sequence reads into loci, presenting a number of bioinformatics challenges. We developed HybPiper as a user-friendly platform for assembly of gene regions, extraction of exon and intron sequences, and identification of paralogous gene copies. We test HybPiper using baits designed to target 333 phylogenetic markers and 125 genes of functional significance in *Artocarpus* (Moraceae).

Methods and Results:

HybPiper implements parallel execution of sequence assembly in three phases: read mapping, contig assembly, and target sequence extraction. The pipeline was able to recover nearly complete gene sequences for all genes in 22 species of *Artocarpus*. HybPiper also recovered more than 500 bp of nontargeted intron sequence in over half of the phylogenetic markers and identified paralogous gene copies in *Artocarpus*.

HybPhyloMaker

(reference-based mapping)

▸ Evol Bioinform Online. 2018 Jan 12;14:1176934317742613. doi: 10.1177/1176934317742613

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

Tomáš Fér [1,✉], Roswitha E Schmickl [2]

▸ Author information ▸ Article notes ▸ Copyright and License information

PMCID: PMC5768271 PMID: 29348708

Abstract

Summary:

Hybridization-based target enrichment in combination with genome skimming (Hyb-Seq) is becoming a standard method of phylogenomics. We developed HybPhyloMaker, a bioinformatics pipeline that performs target enrichment data analysis from raw reads to supermatrix-, supertree-, and multispecies coalescent-based species tree reconstruction. HybPhyloMaker is written in BASH and integrates common bioinformatics tools. It can be launched both locally and on a high-performance computer cluster. Compared with existing target enrichment data analysis pipelines, HybPhyloMaker offers the following main advantages: implementation of all steps of data analysis from raw reads to species tree reconstruction, calculation and summary of alignment and gene tree properties that assist the user in the selection of "quality-filtered" genes, implementation of several species tree reconstruction methods, and analysis of the coding regions of organellar genomes.

ParalogWizard

(reference-based assembly,
de novo assembly,
reference-based assembly using a
paralog-prone reference)

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE | Open Access

Utilizing paralogues for phylogenetic reconstruction has the potential to increase species tree support and reduce gene tree discordance in target enrichment data

Roman Ufimov, Juan Manuel Gorospe, Tomáš Fér, Martha Kandziora, Luciana Salomon, Marcela van Loo, Roswitha Schmickl

First published: 07 July 2022 | https://doi.org/10.1111/1755-0998.13684 | Citations: 9

Handling Editor: Suhua Shi

Abstract

The analysis of target enrichment data in phylogenetics lacks optimization toward using paralogues for phylogenetic reconstruction. We developed a novel approach of detecting paralogues and utilizing them for phylogenetic tree inference, by retrieving both ortho- and paralogous copies and creating orthologous alignments, from which the gene trees are built. We implemented this approach in ParalogWizard and demonstrate its performance in plant groups that underwent a whole genome duplication relatively recently: the subtribe Malinae (family Rosaceae), using Angiosperms353 as well as Malinae481 probes, the genus *Oritrophium* (family Asteraceae), using Compositae1061 probes, and the genus *Amomum* (family Zingiberaceae), using Zingiberaceae1180 probes. Discriminating between orthologues and paralogues reduced gene tree discordance and increased the species tree support in the case of the Malinae, but not for *Oritrophium* and *Amomum*. This may relate to the difference in the proportion of paralogous loci between the data sets, which was highest for the Malinae. Overall, retrieving paralogues for phylogenetic reconstruction following ParalogWizard has the potential to increase the species tree support and reduce gene tree discordance in target enrichment data, particularly if the proportion of paralogous loci is high.