# reads2trees-hybpiper Manual

This pipeline is adapted from the "reads2trees" workflow from Dr. Karolina Heyduk. Here we use newer programs and some computational shortcuts to decrease wait time. Many of the scripts in this pipeline are hardcoded for use on the GACRC Sapelo2 HPC.

Pipeline workflow:
Trim reads
Assemble genes
Align genes
Infer gene trees using a maximum likelihood approach
Infer species tree using coalescence-based model

**Trim Reads- Trimmomatic**

Usage:
**perl trim.pl </directory/for/trimmed/reads> <readindex.txt>**

This script is designed to remove adapters and low-quality bases at the ends of reads with program Trimmomatic. Run this script in your raw reads directory.

</directory/for/trimmed/reads> is the path to a new directory for the trimmed reads to be stored. This must be different than the raw reads directory.
<readindex.txt> is a tab delimited text file that matches shorthand IDs to the two read files (forward and reverse) and should be set up like this:

| #ID | #readfile |
|---|---|
| A01 | A1_S23_R1_001.fastq.gz |
| A01 | A1_S23_R2_001.fastq.gz |
| A02 | A2_S23_R1_001.fastq.gz |
| A02 | A2_S23_R2_001.fastq.gz |
| … | … |

(Don't include #ID and #readfile in your text file)

This script uses the default Trimmomatic settings. In order to change the trim settings, you can edit the "trim.pl" script.

Outputs:
ID_R1_P.fastq.gz
ID_R2_P.fastq.gz
ID_R1_U.fastq.gz
ID_R2_U.fastq.gz

The unpaired reads can be deleted, unless you want to include them in your analysis.

Trimmomatic Citation:
Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

**Assembly- HybPiper**

Usage:
Create a directory for hybpiper in your working directory.
Place 353_targets.fa in this directory
Run:

```
perl batchhybpiper.pl </hyb/dir/> </directory/for/trimmed/reads> 8 <IDs.txt>
```

This script is designed to run hybpiper for each sample, and then run the intronerate.

</hyb/dir/> is the path to your hybpiper directory.
</directory/for/trimmed/reads> is the path to the trimmed reads.
<cpus> is the number of cpus you want to use. I recommend 8.
<IDs.txt> is a text file that contains the shorthand IDs (one per line) used as the prefix for the trimmed reads. Like this:

```
#ID
A01
A02
…
```

(don’t include #ID in your text file)

This script only uses the paired reads for assembly. If you would like to change the settings, consult the Hybpiper manual and edit the script.

Outputs:
The Hybpiper directory will contain individual folders (named with shorthand IDs) which contain many other subdirectories. So, we will run one more script to retrieve those sequences and rename the fasta headers.

Make a directory for your supercontig genes

Run this script in supercontig directory:

```
bash hybretseq.sh </hyb/dir/>
```

This script pulls genes from the various hybseq directories and puts them into a file

</hyb/dir/> is the path to your hybpiper directory.

Output:
geneid_supercontig.gene

Hybpiper Citation:
Matthew G. Johnson, Elliot M. Gardner, Yang Liu, Rafael Medina, Bernard Goffinet, A. Jonathan Shaw, Nyree J. C. Zerega, Norman J. Wickett "HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment," Applications in Plant Sciences, 4(7), (12 July 2016)

**Align- PASTA**

Usage:
Put pasta.pl in the supercontig genes directory. Then run:

```
perl pasta.pl </supercontig/dir/> <listofgenes.txt>
```

</supercontig/dir/> Path to supercontig genes directory
<listofgenes.txt> the list of genes you are using. For Angiosperm353 use 353genelist.txt

Runs PASTA alignment algorithm on each gene.

Output:
geneid.marker###.geneid_supercontig.gene.aln files for each gene.

To rename the files to “geneid_supercontig.gene.aln”, we use this command:

```
for f in *.aln; do mv $f ${f: -25}; done
```

PASTA Citation:

Mirarab S, Nguyen N, Warnow T. PASTA: ultra-large multiple sequence alignment. Sharan R, ed. Res Comput Mol Biol. 2014:177-191.

Mirarab S, Nguyen N, Guo S, Wang L-S, Kim J, Warnow T. PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. J Comput Biol. 2015;22(5):377-386. doi:10.1089/cmb.2014.0156.

Balaban, Metin, Niema Moshiri, Uyen Mai, and Siavash Mirarab. “TreeCluster : Clustering Biological Sequences Using Phylogenetic Trees.” BioRxiv, 2019, 591388. doi:10.1101/591388.

**Make Gene Trees- IQTREE**

Usage:
Make a new directory for iqtree.
Move alignments from supercontig directory to iqtree directory:

```
mv </supercontig/dir/>*.aln </iqtree/dir>
```

Put iq_sub.sh in /iqtree/dir
Run it:

```
bash iq_sub.sh
```

iq_sub.sh runs each alignment through the iqtree software. With the default settings it runs a modelfinder test for each alignment and creates a ML gene tree with 1000 bootstraps

Outputs:
IQTREE generates many files, but the ones we want are:
geneID.treefile for each gene. This is the ML gene tree.
geneID.ufboot for each gene. This contains the bootstrap trees for a given gene

IQTREE citation:
D.T. Hoang, O. Chernomor, A. von Haeseler, B.Q. Minh, L.S. Vinh (2018) UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol., 35:518–522. https://doi.org/10.1093/molbev/msx281
L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies.. Mol. Biol. Evol., 32:268-274. https://doi.org/10.1093/molbev/msu300
S. Kalyaanamoorthy, B.Q. Minh, T.K.F. Wong, A. von Haeseler, L.S. Jermiin (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods, 14:587-589. https://doi.org/10.1038/nmeth.4285

**Species Tree Estimation- ASTRAL**

Usage:
Start in </iqtree/dir>
Concatenate bipartition trees into one file:

```
cat *.treefile > biparts.tre
```

Make a text file that contains the path to each bootstrap file, one per line:

```
realpath *.ufboot > boots.txt
```

Make an astral directory and put biparts.tre, boots.txt and astral.sh in it

```
qsub astral.sh
```

This script runs ASTRAL, which will use an algorithm statistically consistent with the multispecies coalescent model to estimate a species tree from the gene trees made in IQTREE.

In the astral.sh file, when running astral, there should be 4 flags: -i, -b , -r and -o
-i is bipartitions.tre
-b is boots.txt
-r is the number of bootstrap replicates to consider (needs to be less than number of bootstraps generated i.e. if 1000 bootstraps generated in IQTREE use 900 in ASTRAL.
-o is output.tre or whatever you want to name the output file.

Output:

output.tre: The first n lines are bootstrapped replicate trees with n being the number specified in the -r flag. The next line is the greedy consensus tree. The last line is the main astral tree with bootstrap support values. Last lines is your maintree.

Using different settings in ASTRAL you can see posterior probabilities, quartet support values and test polytomies. These may be useful in your analysis. Refer to the ASTRAL manual for use.

ASTRAL citation:
Zhang, Chao, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. 2018. “ASTRAL-III: Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees.” BMC Bioinformatics 19 (S6): 153. doi:10.1186/s12859-018-2129-y.