Plant Phylogenomics Workshop

Léo-Paul DAGALLIER

May 8th – 10th

What to expect

- (Some) tools and concepts of phylogenomics
- Based on my personal workflow

- GitHub with scripts

- Feel free to ask or to complete

Phylogenomics

= phylogenetics + genomics

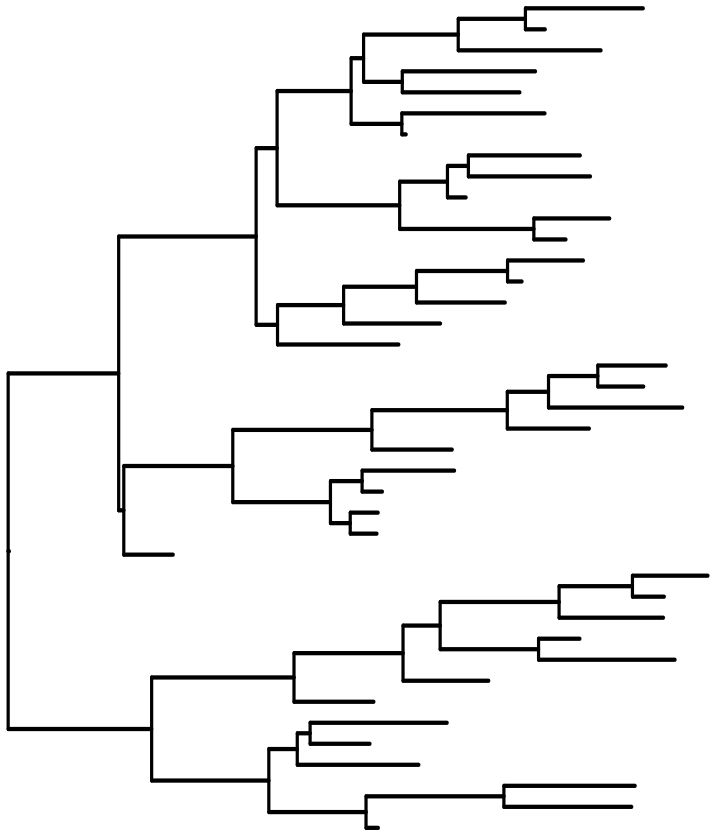
Inference of evolutionary history using genome-scale data

100 – 1000 loci (“genes”)
nucleus, chloroplast, mitochondria

Plant specimens

…

…

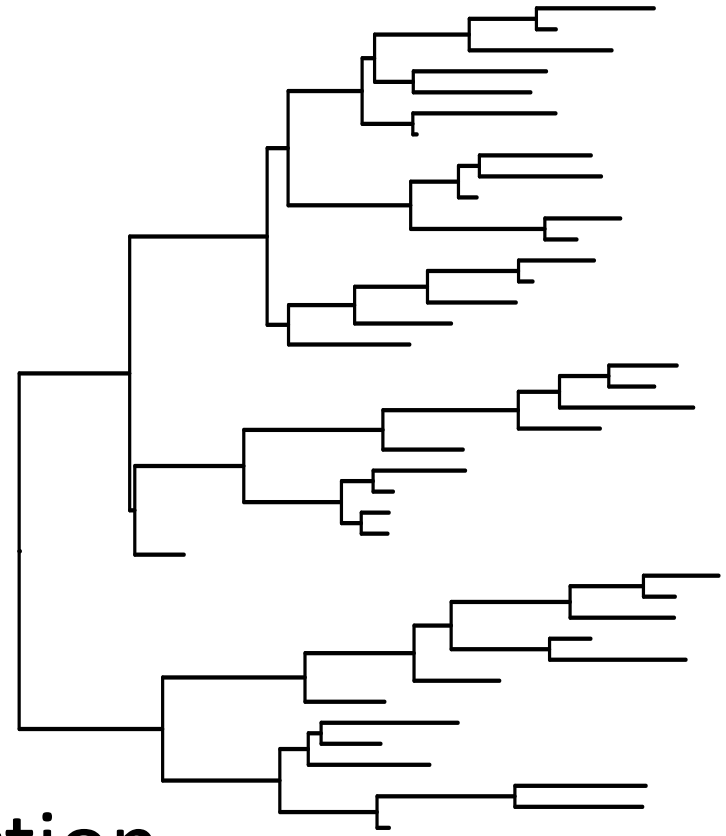


Robust phylogenetic hypothesis

…

…

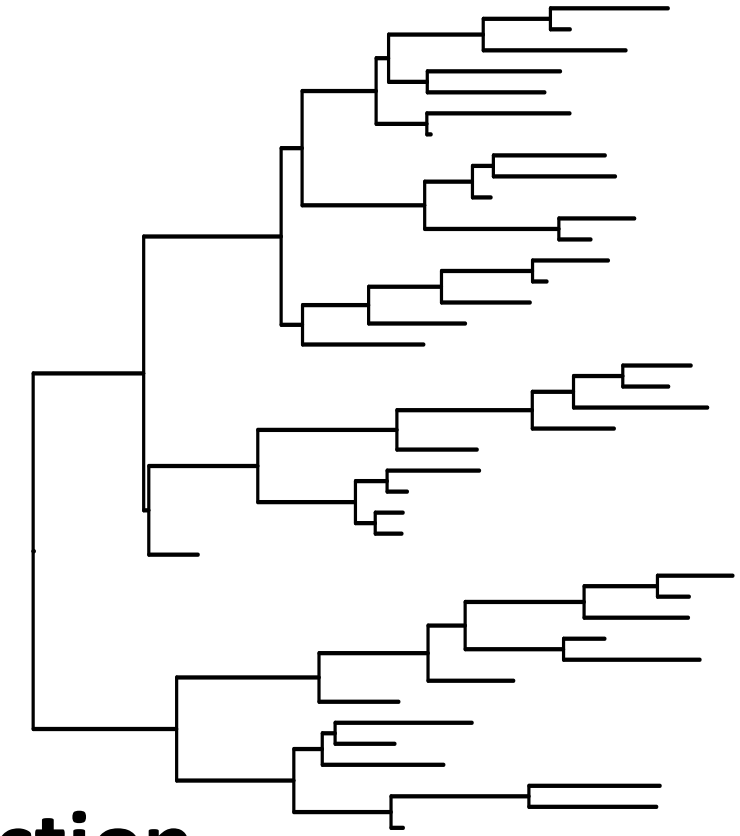
- DNA extraction
- Sequence recovery
- Phylogenetic reconstruction



…

…

- DNA extraction
- **Sequence recovery**
- **Phylogenetic reconstruction**

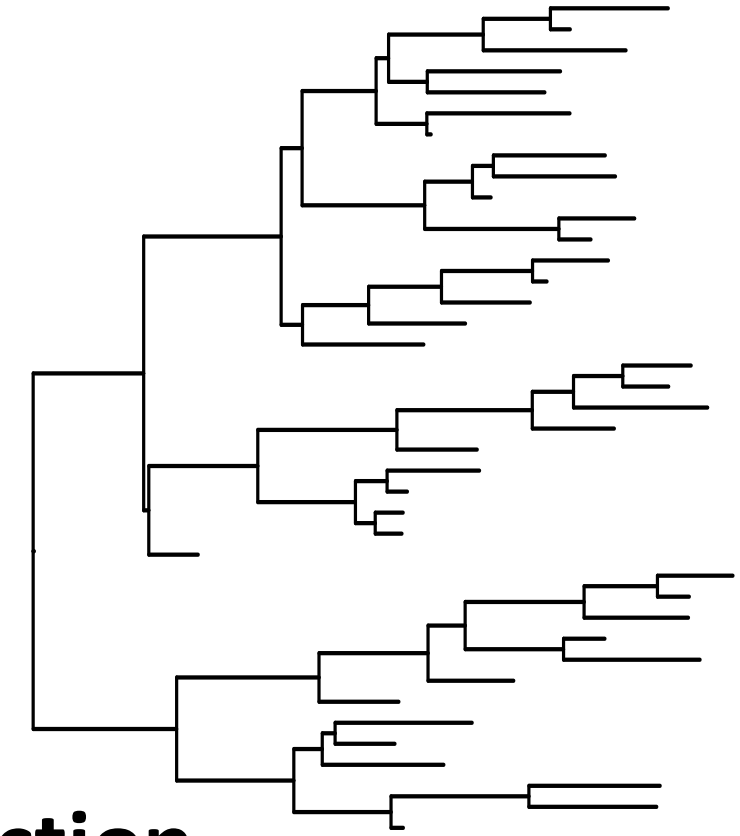


…

…

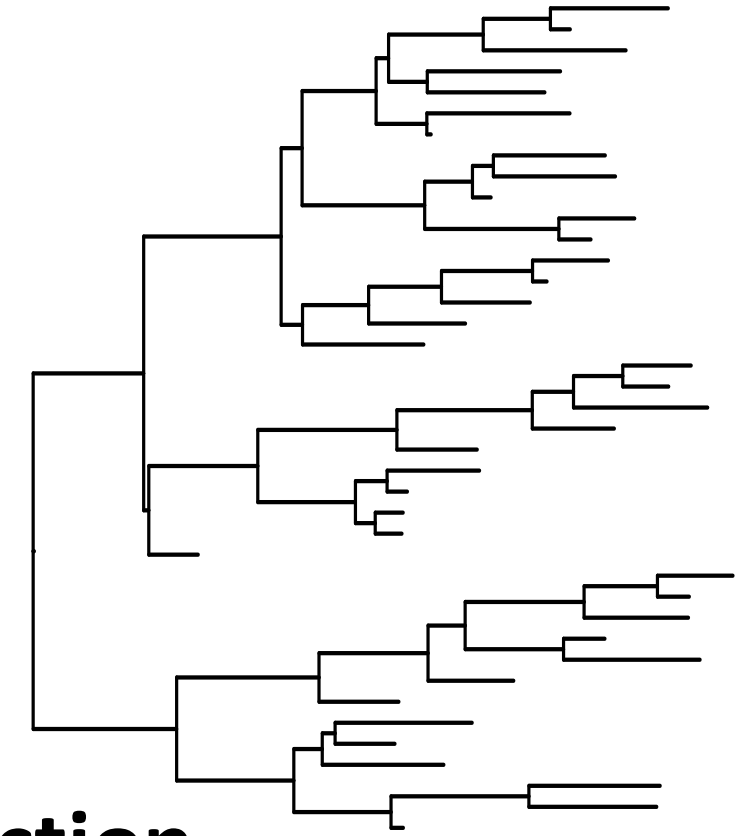
- DNA extraction

- **Sequence recovery**
  - Targeted sequencing
  - Genome skimming
- **Phylogenetic reconstruction**



- DNA extraction

- **Sequence recovery**
  - Targeted sequencing
  - Genome skimming
- **Phylogenetic reconstruction**
  - Gene trees approach
  - Concatenation approach

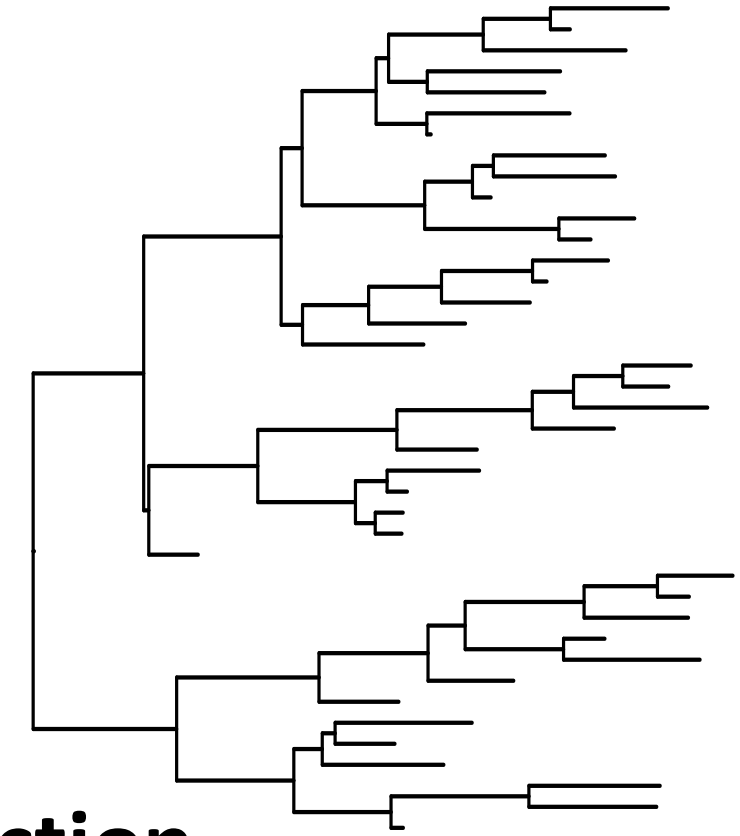


…

…

- DNA extraction

- **Sequence recovery**
  - Targeted sequencing
  - Genome skimming
- **Phylogenetic reconstruction**
  - Gene trees approach
  - Concatenation approach



Sequence recovery

TARGETED SEQUENCING (CAPTURE)

- low-copy elements of the genome
- ideally: single-copy orthologous loci
- nuclear loci (usually)
- require specific probes to be designed:
  - “Universal” e.g. Angiosperms 353
  - Family-specific e.g. Melastomataceae, …

Sequence recovery

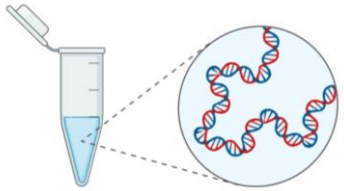
TARGETED SEQUENCING (CAPTURE)

- low-copy elements of the genome
- ideally: single-copy orthologous loci
- nuclear loci (usually)
- require specific probes to be designed:
  - “Universal” e.g. Angiosperms 353
  - Family-specific e.g. Melastomataceae, …

GENOME SKIMMING

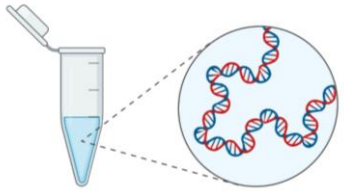
- high-copy elements of the genome
- plastid genes
- mitochondrial genes

Sequence recovery



Extracted DNA

Sequence recovery

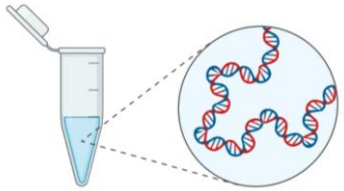


Extracted DNA

Sequence reads

Sequencing

Sequence recovery



Extracted DNA

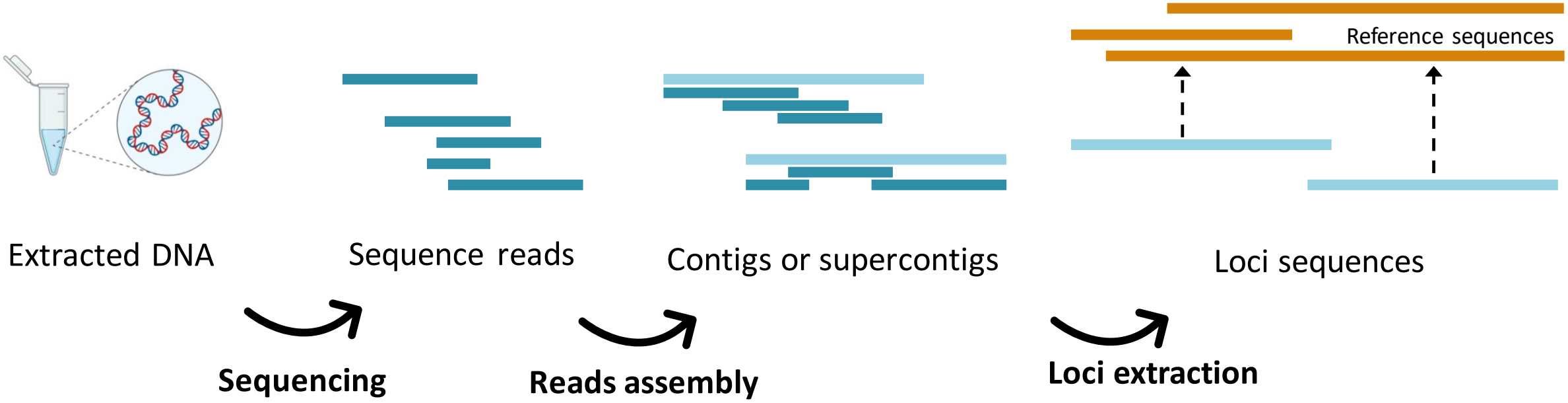
Sequence reads

Contigs or supercontigs

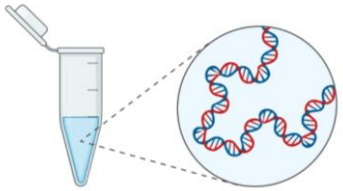
Sequencing

Reads assembly

Sequence recovery



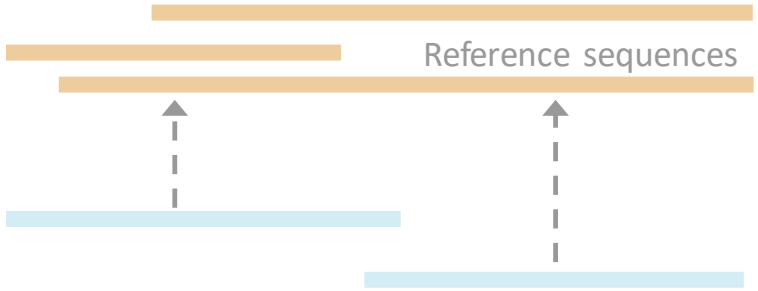
Sequence recovery



Extracted DNA

Sequence reads

Contigs or supercontigs



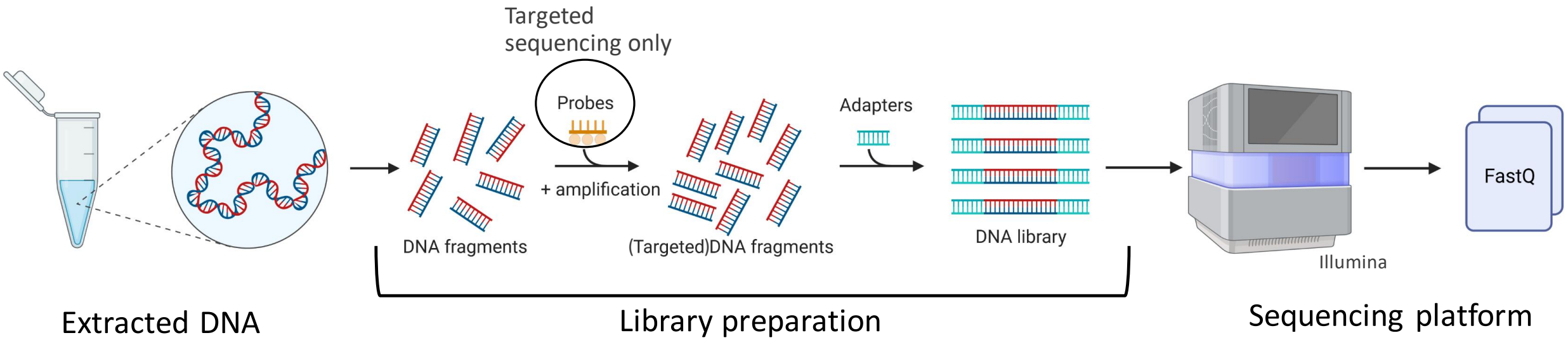
Loci sequences

Sequencing

Reads assembly

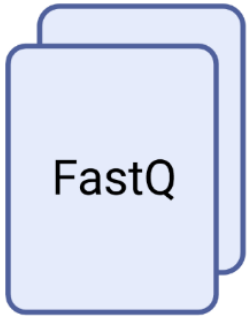
Loci extraction

Sequencing


Targeted
sequencing only
Probes
+ amplification
DNA fragments
(Targeted)DNA fragments
Adapters
DNA library
Illumina
FastQ
Extracted DNA
Library preparation
Sequencing platform

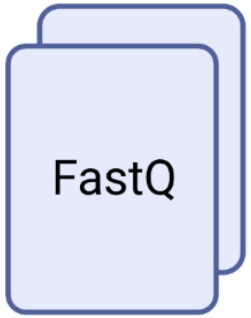
Sequencing

2 .fastq files per specimen: R1 and R2



Sequencing

2 .fastq files per specimen: R1 and R2



Read identifier

Read sequence

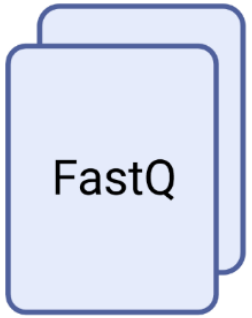
Extra line

Encoded base quality (Phred score)

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Sequencing

2 .fastq files per specimen: R1 and R2



Filter on reads quality:

- Average phred score > 30
- Length > 35 bp
- At least 40% of bases with phred > 15
- Remove low complexity reads

Remove possibly remaining adapters sequence

Fastp
Trimmomatic
Cutadapt
bbduk

Read identifier

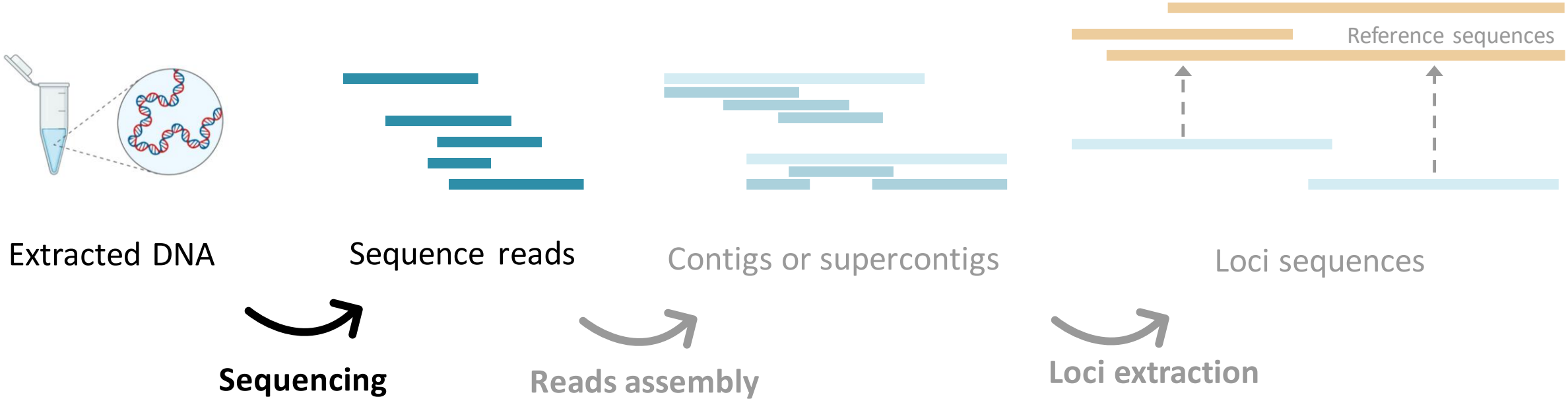
Read sequence

Extra line

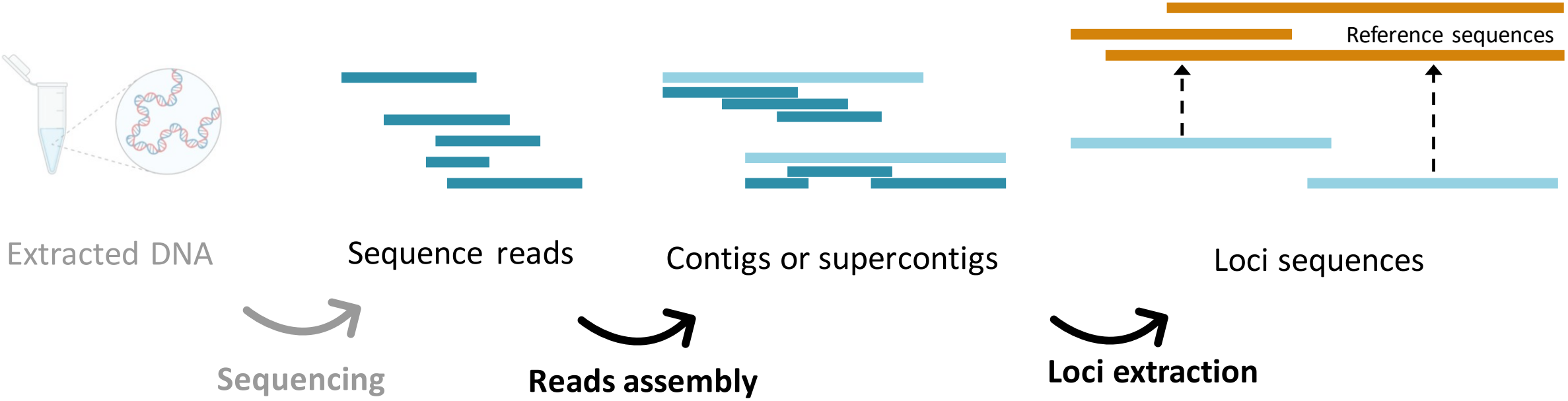
Encoded base quality (Phred score)

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

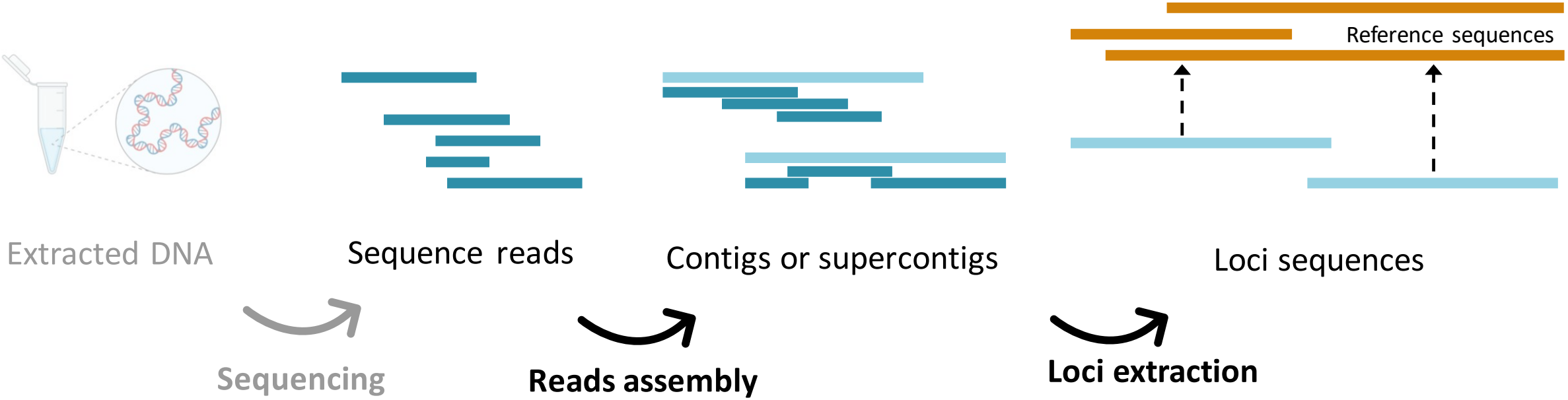
Sequence recovery



Sequence recovery



Sequence recovery



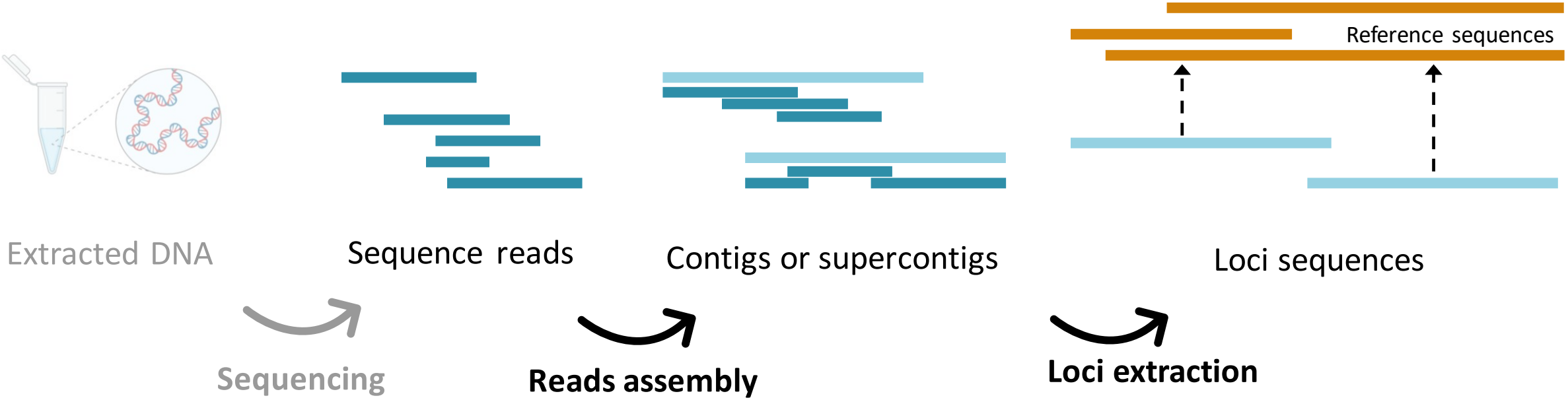
GENOME SKIMMING:

Plastome assembly and annotation

NovoPlasty
GetOrganelle
Fast-Plast
Geneious (GUI but $$$)

Freudenthal JA, et al. (2020)
A systematic comparison of chloroplast genome assembly tools
https://doi.org/10.1186/s13059-020-02153-6

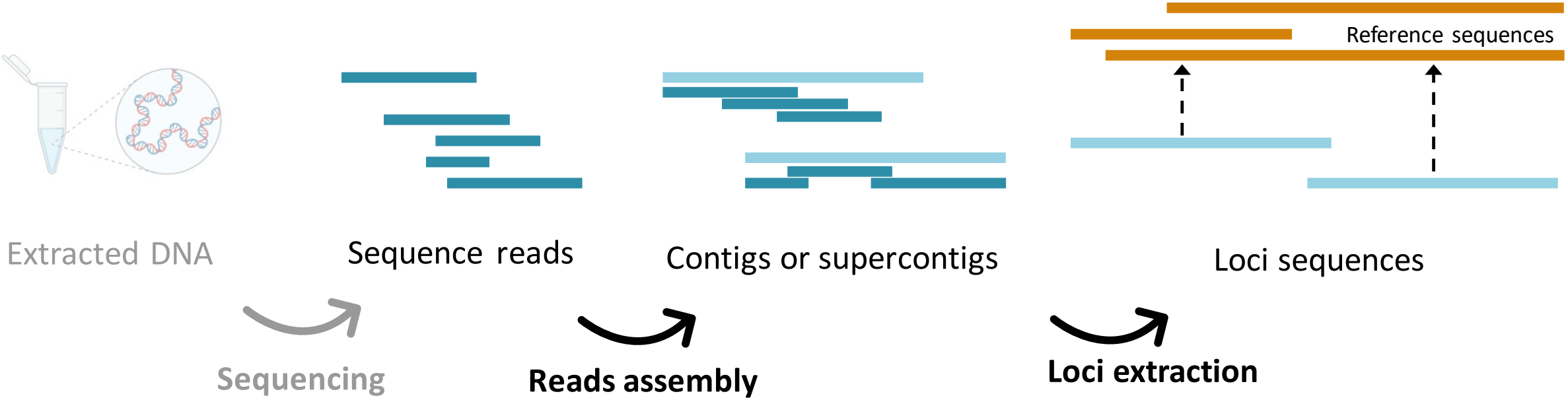
Sequence recovery



TARGET CAPTURE:

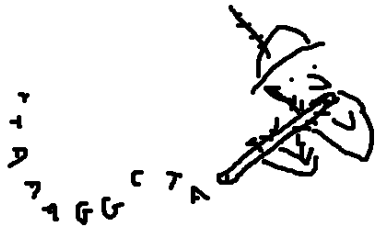
HybPiper (Johnson et al. 2016)
Captus (Edgardo M. Ortiz, in prep.)
SECAPR (Andermann et al. 2018)

Sequence recovery



TARGET CAPTURE:

HybPiper (Johnson et al. 2016)
Captus (Edgardo M. Ortiz, in prep.)
SECAPR (Andermann et al. 2018)



HybPiper - assembly and loci extraction

Command line tool (bash)
Linux and MacOS (and computation clusters)
Uses Python scripts wrapping other programs: SPAdes assembler, BLAST aligner, Exonerate, …

Different subcommands:

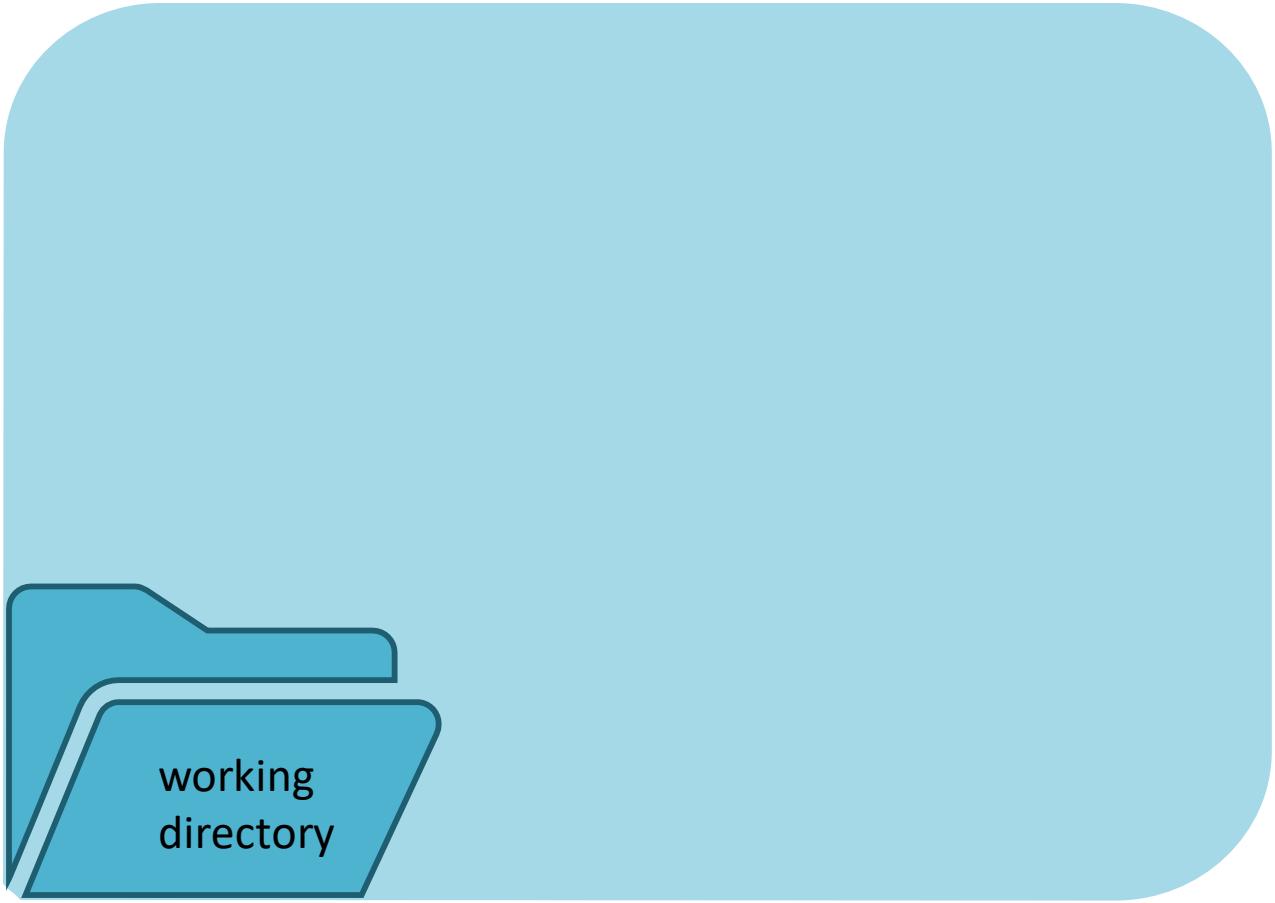
```
hybpiper assemble
hybpiper retrieve_sequences
```

…

+ computes recovery statistics, identifies putative paralogous loci, …

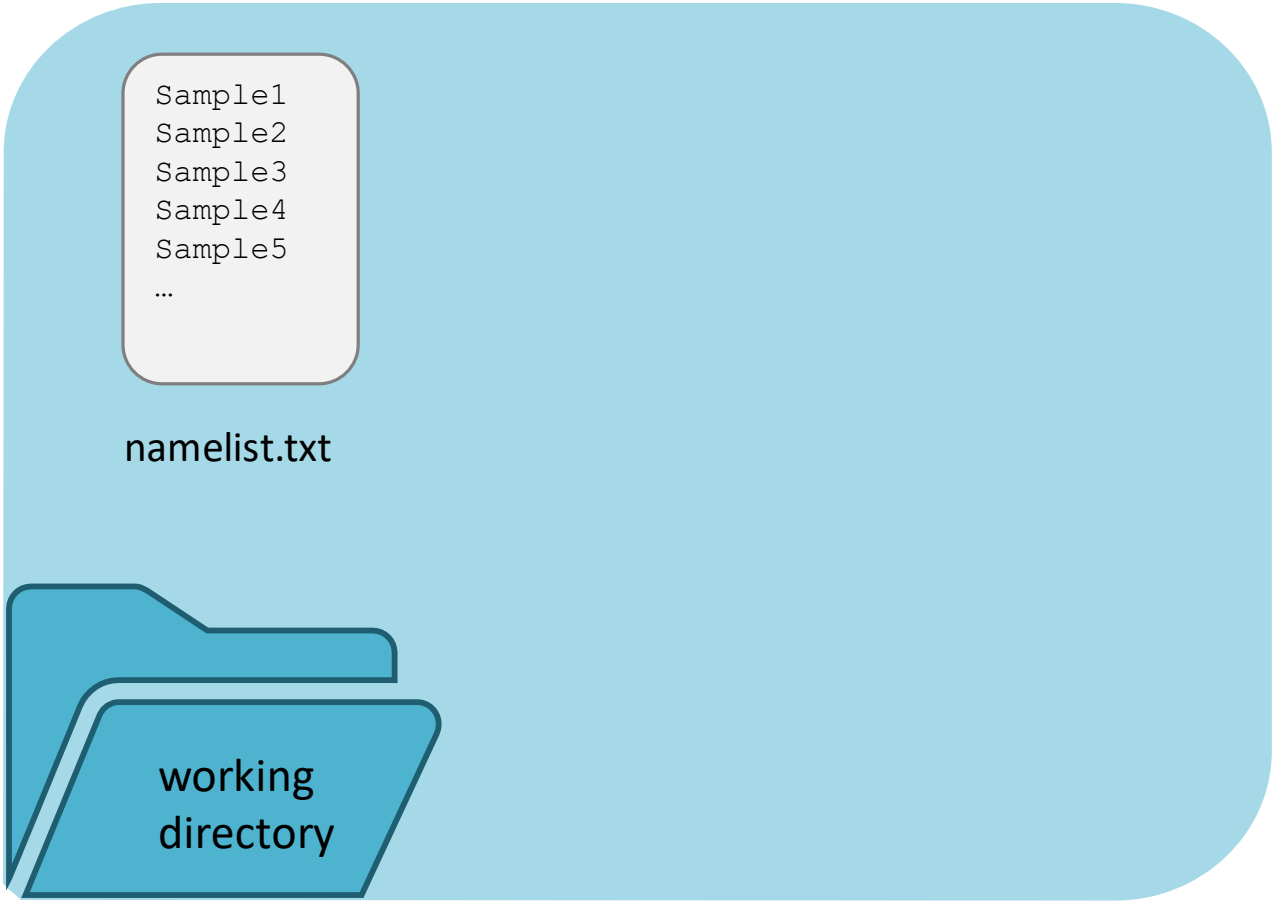
Very good tutorials: https://github.com/mossmatters/HybPiper/wiki
(and responsive developer Chris Jackson)

HybPiper - assembly and loci extraction



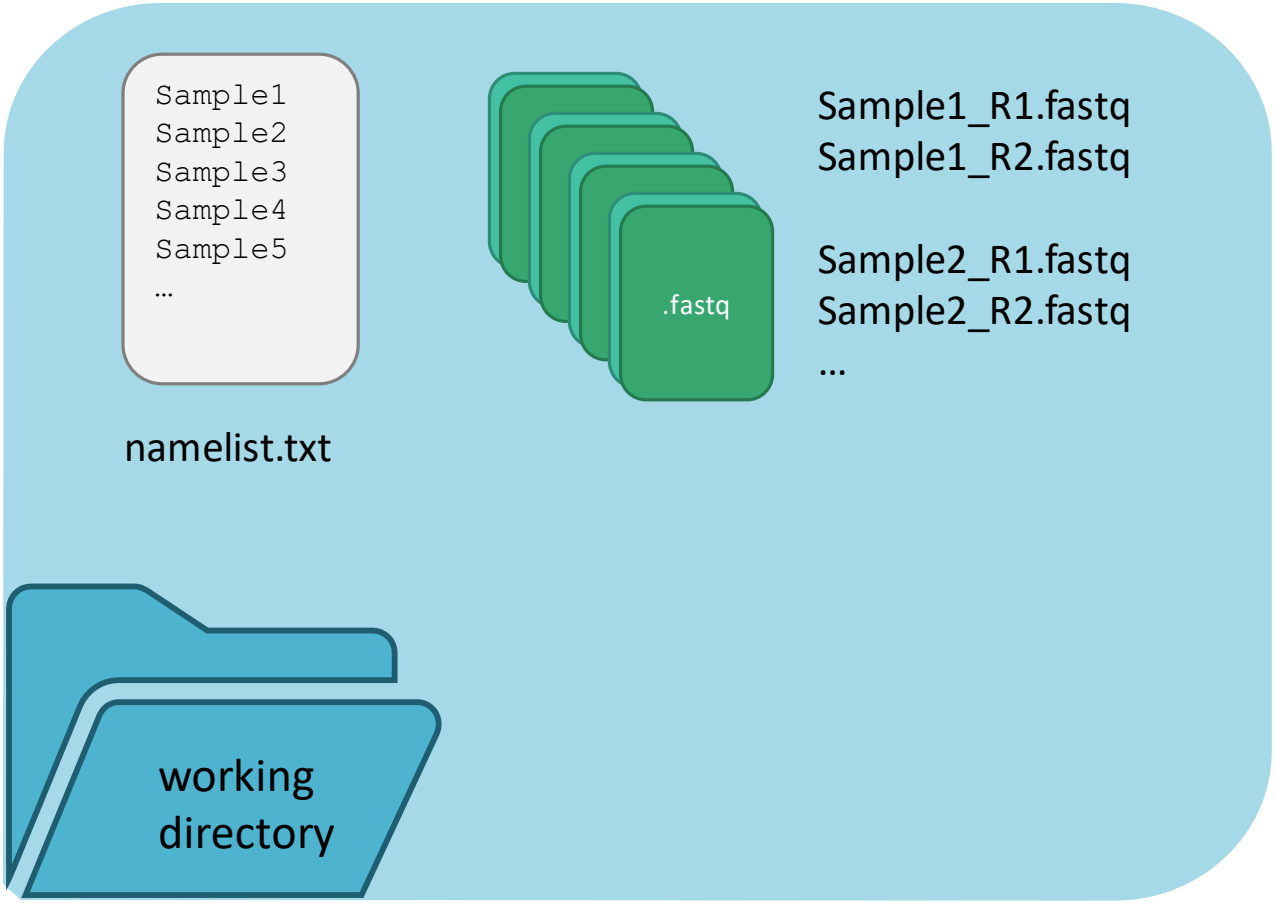
-
-
-

HybPiper - assembly and loci extraction



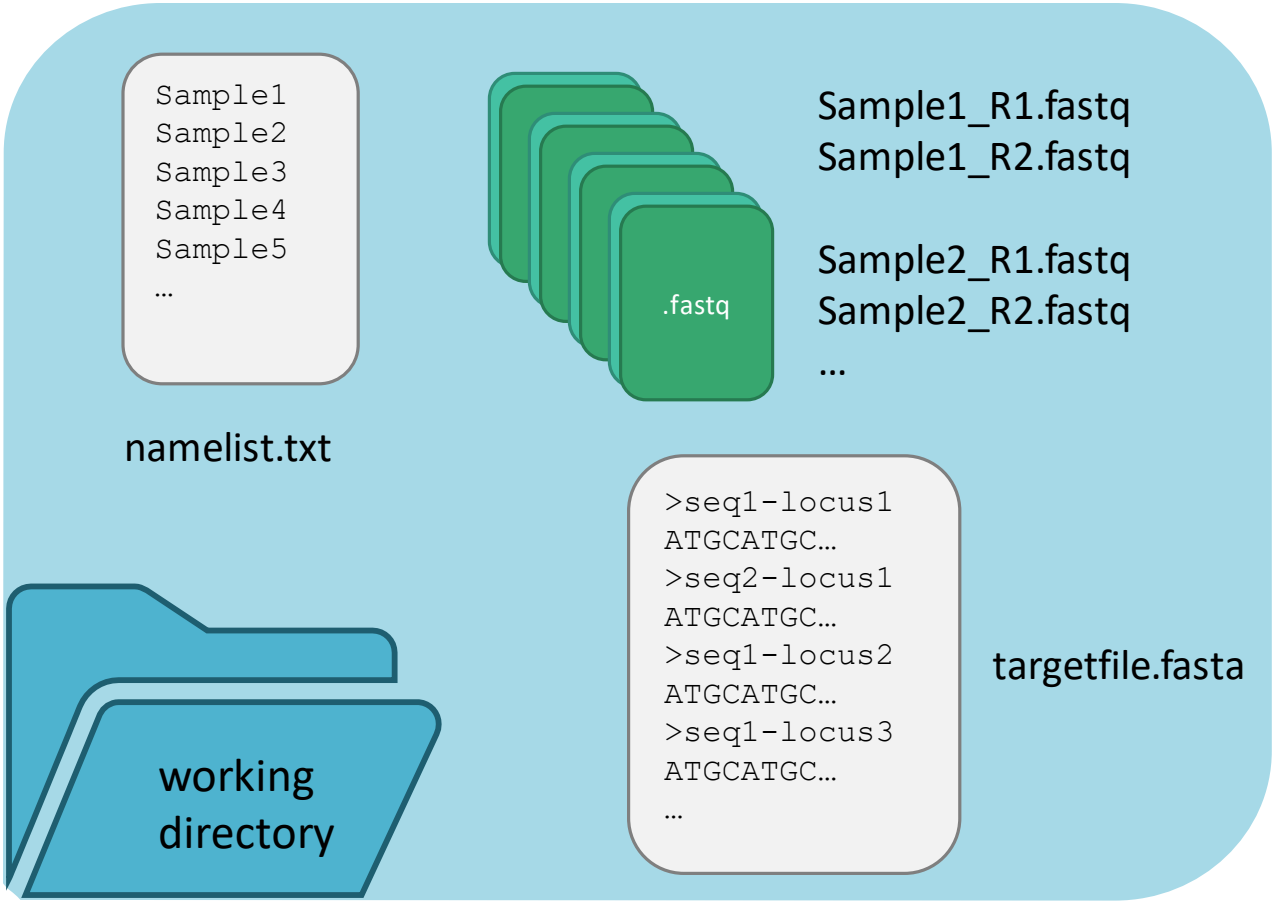
- list of the samples (text file)
-
-

HybPiper - assembly and loci extraction



- list of the samples (text file)
- clean reads files (R1 and R2.fastq) for each sample
-

HybPiper - assembly and loci extraction



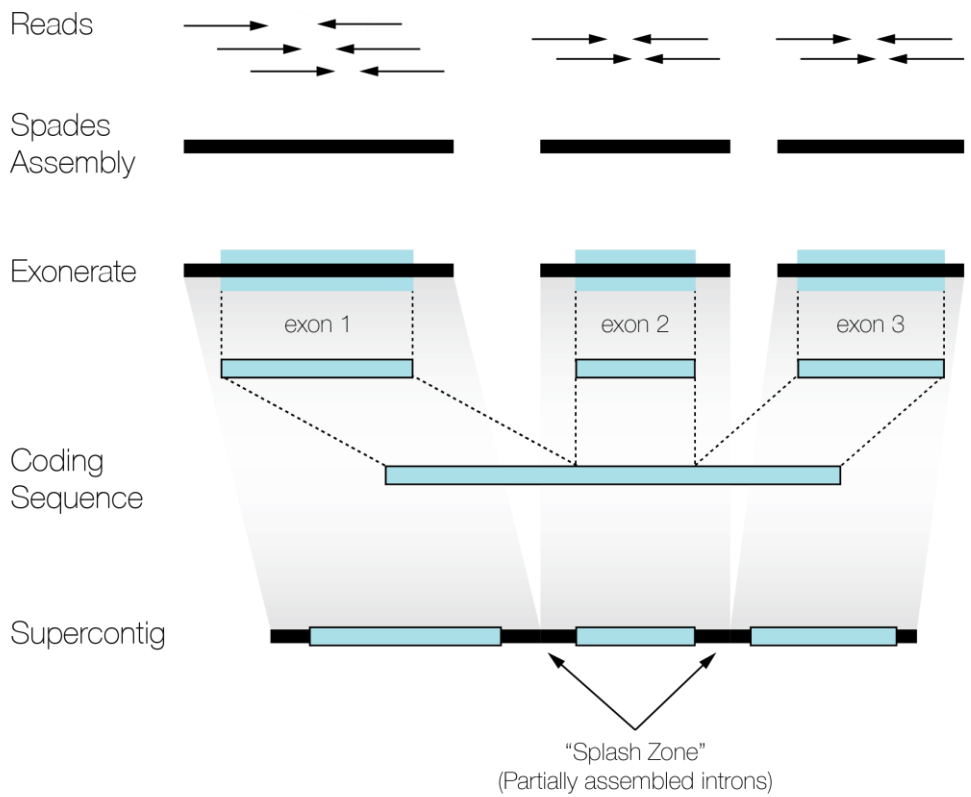
- list of the samples (text file)
- clean reads files (R1 and R2.fastq) for each sample
- target file (.fasta): contains sequence(s) of the targeted loci

HybPiper - **assembly**

```
hybpiper assemble –t_dna targetfile.fasta –r Sample1*.fastq --run_intronerate
```

HybPiper - **assembly**

```
hybpiper assemble –t_dna targetfile.fasta –r Sample1*.fastq --run_intronerate
```



For Sample1:

1. Reads are searched against the target file and sorted according to the target loci (`BWA`, `BLAST`, or `Diamond`); then for each locus:
2. The reads are assembled into contigs (`SPAdes`),
3. Contigs are aligned to the translated reference sequence (target locus); slightly overlapping contigs are scaffolded (i.e. concatenated) into supercontigs
4. Supercontigs are translated to identify exons and introns sequences (`Exonerate`)
5. Exons, introns, and supercontigs (exons+introns) are generated

Note: HybPiper assumes target sequences are exon only

HybPiper - **assembly**

Looping over all samples:

```
Sample1
Sample2
Sample3
Sample4
Sample5
…
```

namelist.txt

HybPiper - **assembly**

Looping over all samples:

```
Sample1
Sample2
Sample3
Sample4
Sample5
…
```

namelist.txt

```
while read LINE;
do
    COMMAND TO EXECUTE ON $LINE
done < TEXTFILE
```

HybPiper - **assembly**

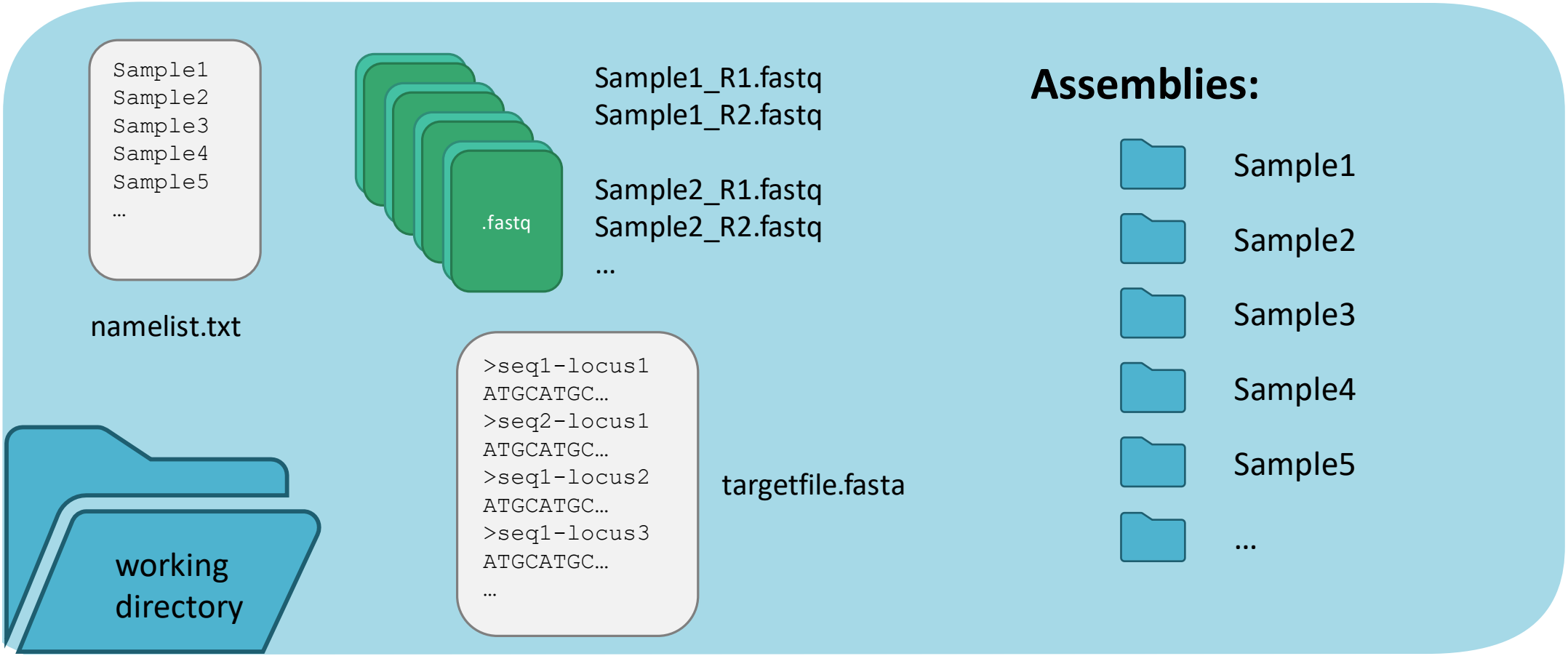
Looping over all samples:

```
Sample1
Sample2
Sample3
Sample4
Sample5
…
```

namelist.txt

```
while read name;
do
    hybpiper assemble -t_dna targetfile.fasta -r $name*.fastq --prefix $name --run_intronerate ;
done < namelist.txt
```

HybPiper - **assembly**



HybPiper - **assembly**

Compute assembly statistics:

```
hybpiper stats -t_dna targetfile.fasta --seq_lengths_filename genes_sequences_lengths --stats_filename
hybpiper_genes_statistics gene namelist.txt
```

→ genes_sequences_lengths.tsv

Species	locus1	locus 2	locus3
MeanLength	1548.0	468.75	953.16
Sample1	1833	270	0
Sample2	1833	357	120
Sample3	1833	468	120

→ hybpiper_genes_statistics.tsv

Name	NumReads	ReadsMapped	PctOnTarget	…
Sample1	3057338	1359471	44.5	
Sample2	7809750	3468403	45.0	
Sample3	6214972	2784358	20.7	

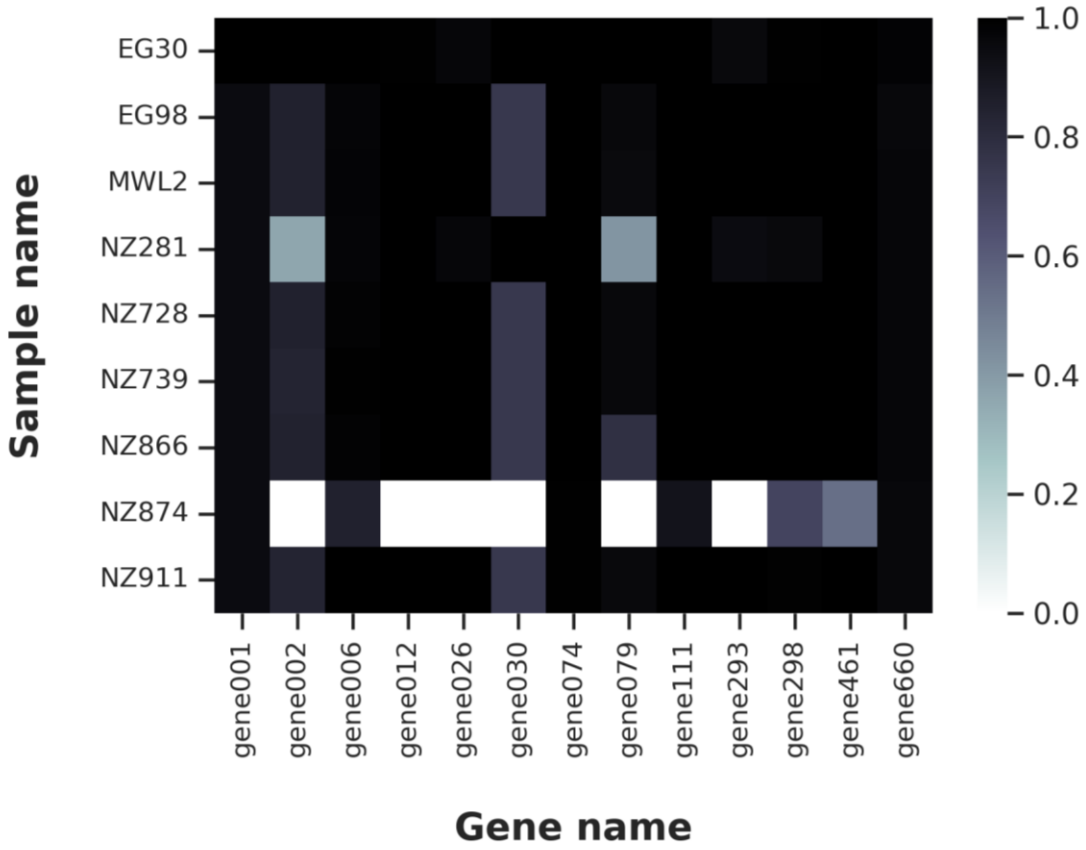
HybPiper - **assembly**

Visualizing the results:

```
hybpiper recovery_heatmap --heatmap_dpi 300
--heatmap_filetype pdf --heatmap_filename
recovery_heatmap_exons
genes_sequences_lengths.tsv
```

→ recovery_heatmap_exons.pdf

Percentage length recovered for each gene



HybPiper - **assembly**

```
Sample1
Sample2
Sample3
Sample4
Sample5
…
```

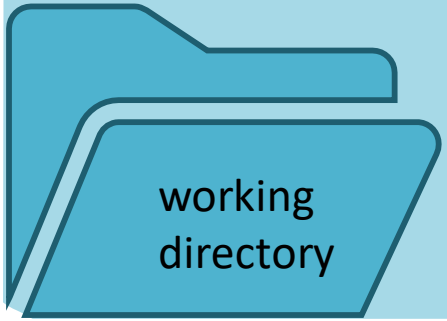
namelist.txt

Sample1_R1.fastq
Sample1_R2.fastq

Sample2_R1.fastq
Sample2_R2.fastq
…

```
>seq1-locus1
ATGCATGC…
>seq2-locus1
ATGCATGC…
>seq1-locus2
ATGCATGC…
>seq1-locus3
ATGCATGC…
…
```

targetfile.fasta



working directory

Assemblies:

Sample1
Sample2
Sample3
Sample4
Sample5
…

genes_sequences_lengths.tsv
recovery_heatmap_exons.pdf
hybpiper_genes_statistics.tsv

HybPiper

General recommendations for `hybpiper assemble`:

- BWA, BLAST, Diamond: default uses BWA (Burrow Wheeler Aligner) to distribute reads to target. I would recommend to use Diamond, which uses protein alignment (i.e `assemble --t_aa` instead of `assemble --t_dna`)

```
hybpiper assemble –t_aa targetfile.fasta –diamond –r Sample1*.fastq --run_intronerate
```

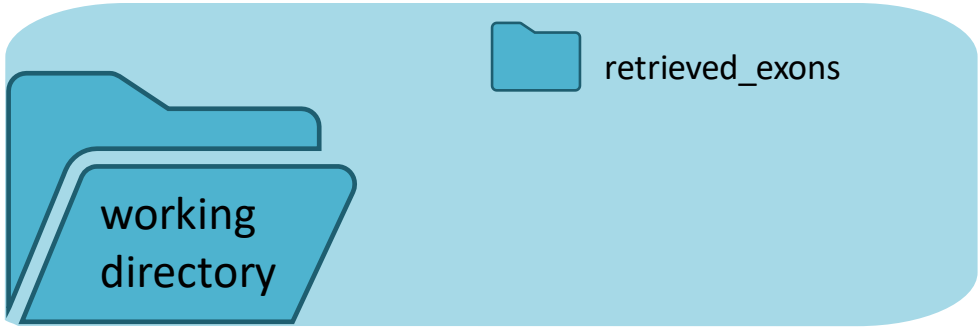
- Use no more than 8 CPUs (--cpu 8): too many instances of hybpiper running in parallel can cause issues on some HPC systems

HybPiper - loci extraction

```
hybpiper retrieve_sequences
```

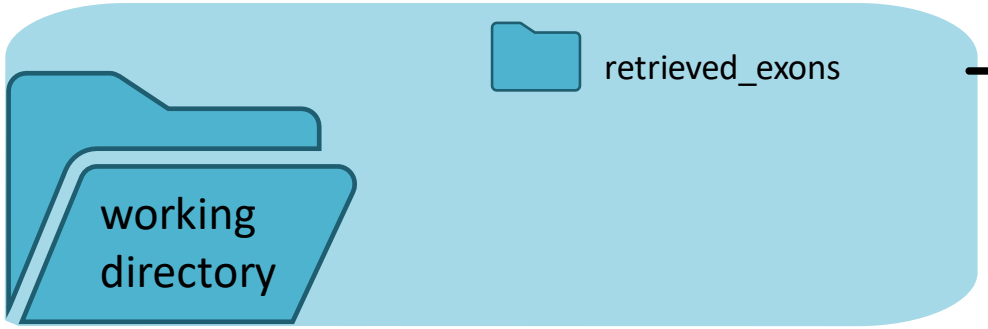
HybPiper - loci extraction

```
mkdir retrieved_exons
hybpiper retrieve_sequences
```



HybPiper - loci extraction

```
mkdir retrieved_exons
hybpiper retrieve_sequences dna -t_dna targetfile.fasta --sample_names namelist.txt --fasta_dir retrieved_exons
```



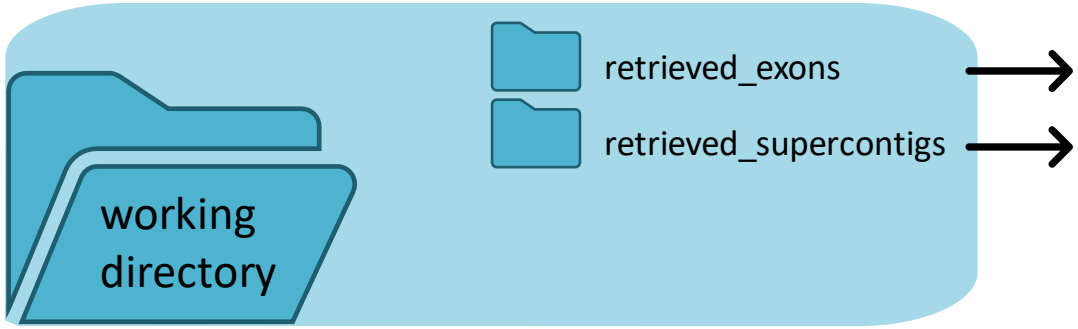
1 file/locus, with 1 sequence per recovered sample

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
…
```

locus1.FNA
locus2.FNA
…

HybPiper - loci extraction

```
mkdir retrieved_exons
hybpiper retrieve_sequences dna -t_dna targetfile.fasta --sample_names namelist.txt --fasta_dir retrieved_exons

mkdir retrieved_supercontigs
hybpiper retrieve_sequences supercontig -t_dna targetfile.fasta --sample_names namelist.txt --fasta_dir
retrieved_supercontigs
```



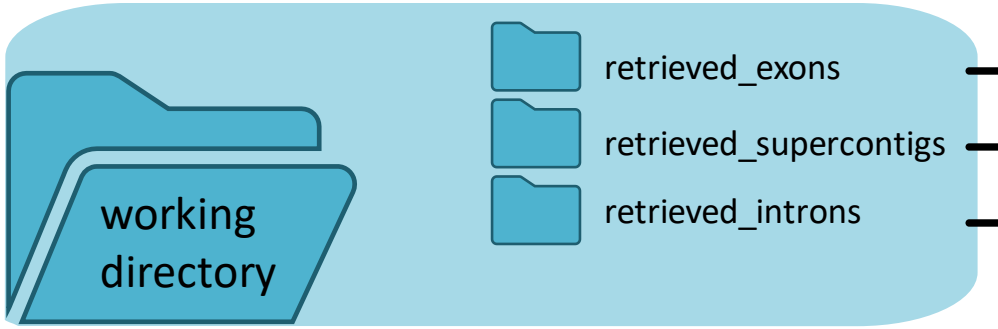
1 file/locus, with 1 sequence per recovered sample

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
…
```

locus1.FNA
locus2.FNA
…

HybPiper - loci extraction

```
mkdir retrieved_exons
hybpiper retrieve_sequences dna -t_dna targetfile.fasta --sample_names namelist.txt --fasta_dir retrieved_exons

mkdir retrieved_supercontigs
hybpiper retrieve_sequences supercontig -t_dna targetfile.fasta --sample_names namelist.txt --fasta_dir
retrieved_supercontigs

mkdir retrieved_introns
hybpiper retrieve_sequences intron -t_dna targetfile.fasta --sample_names namelist.txt --fasta_dir
retrieved_introns
```



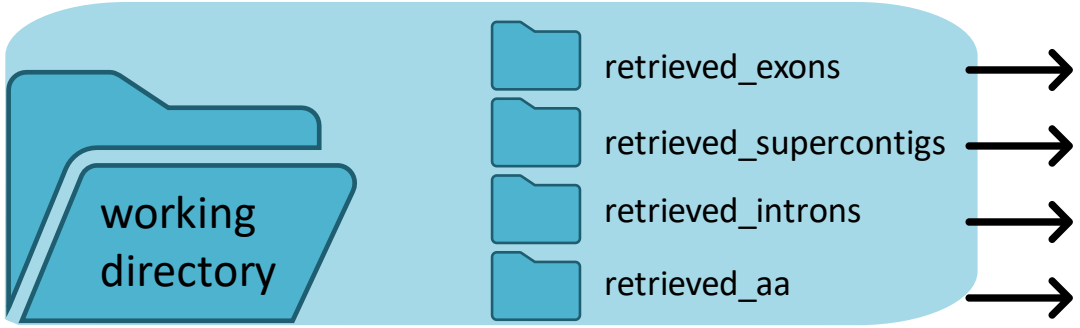
1 file/locus, with 1 sequence per recovered sample

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
…
```

locus1.FNA
locus2.FNA
…

HybPiper - loci extraction

```
mkdir retrieved_exons
hybpiper retrieve_sequences dna -t_dna targetfile.fasta --sample_names namelist.txt --fasta_dir retrieved_exons

mkdir retrieved_supercontigs
hybpiper retrieve_sequences supercontig -t_dna targetfile.fasta --sample_names namelist.txt --fasta_dir
retrieved_supercontigs

mkdir retrieved_introns
hybpiper retrieve_sequences intron -t_dna targetfile.fasta --sample_names namelist.txt --fasta_dir
retrieved_introns

mkdir retrieved_aa
hybpiper retrieve_sequences aa -t_dna targetfile.fasta --sample_names namelist.txt --fasta_dir retrieved_aa
```



1 file/locus, with 1 sequence per recovered sample

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
…
```

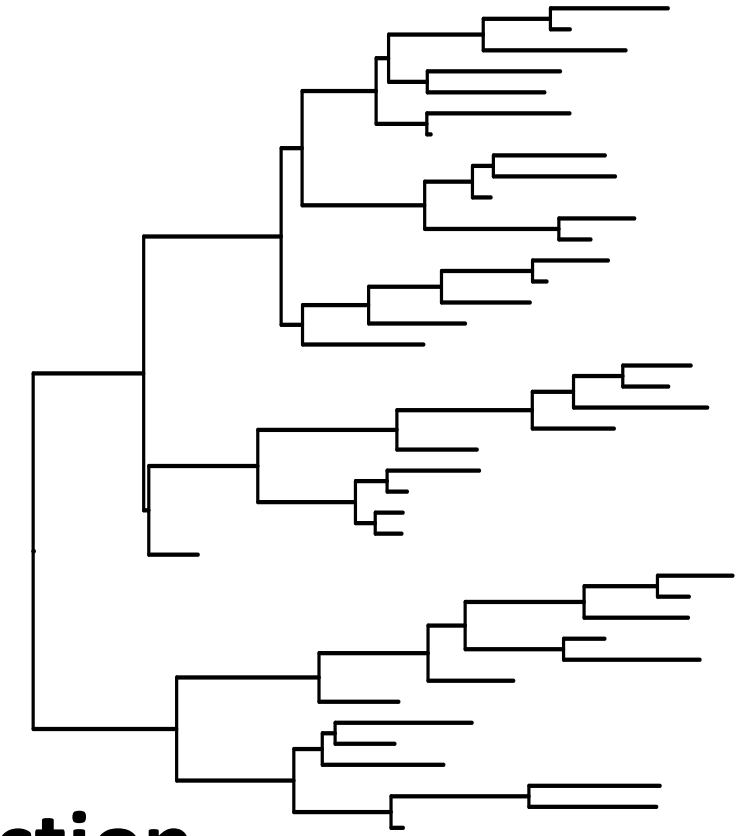
locus1.FNA
locus2.FNA
…

…

…

- DNA extraction

- **Sequence recovery**
  - Targeted sequencing
  - Genome skimming
- **Phylogenetic reconstruction**
  - Gene trees approach
  - Concatenation approach

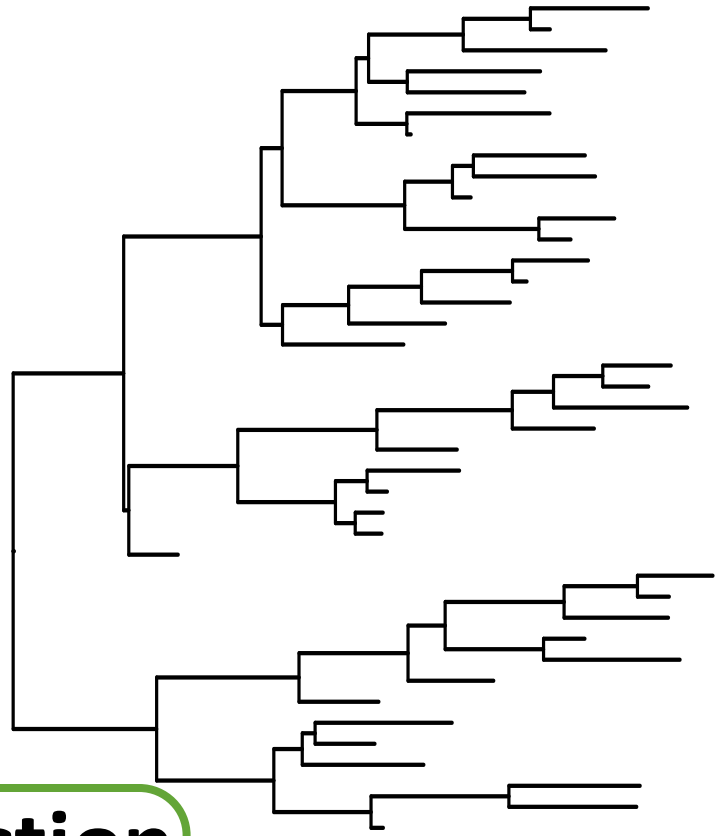


…

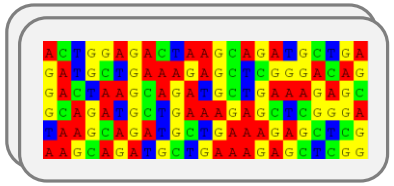
…

- DNA extraction

- **Sequence recovery**
  - Targeted sequencing
  - Genome skimming
- **Phylogenetic reconstruction**
  - Gene trees approach
  - Concatenation approach

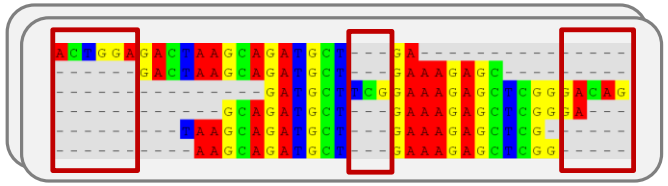


Phylogenetic reconstruction



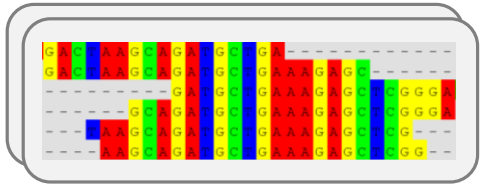
Extracted sequences

Alignment



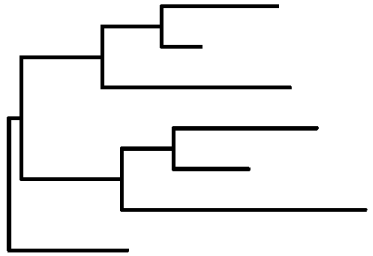
Multi-samples alignment

Alignment trimming

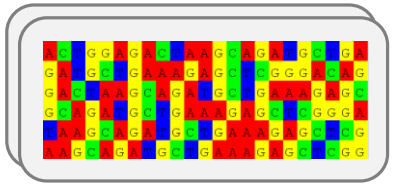


Cleaned alignment

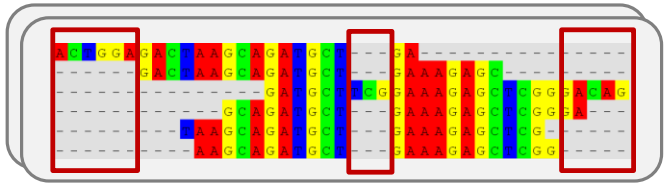
Tree estimation



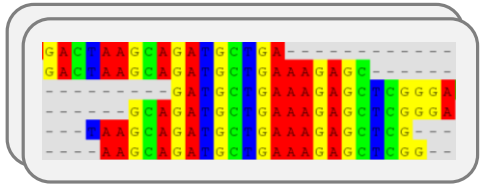
Phylogenetic reconstruction



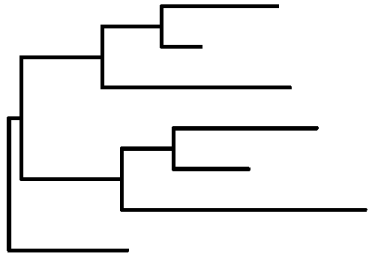
Extracted sequences



Multi-samples alignment



Cleaned alignment



Alignment

Alignment trimming

Tree estimation

+ loci filtering

Paralogs

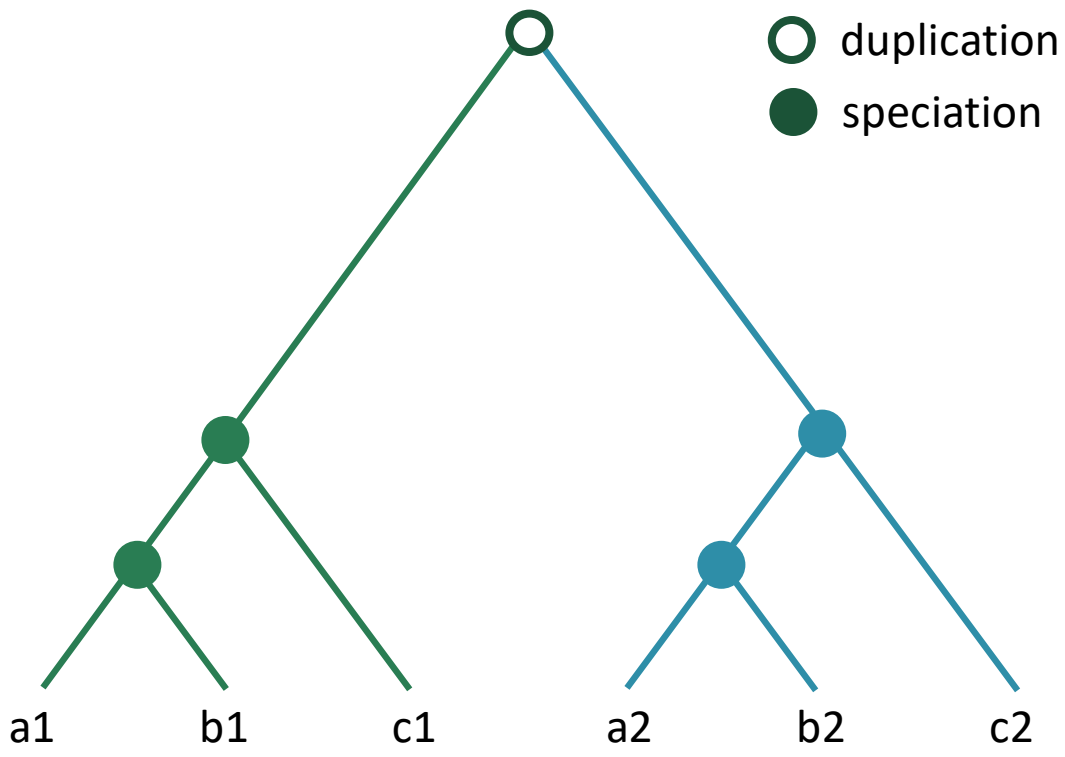
Homologous genes = inherited from an ancestral gene

- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event

Paralogs

Homologous genes = inherited from an ancestral gene

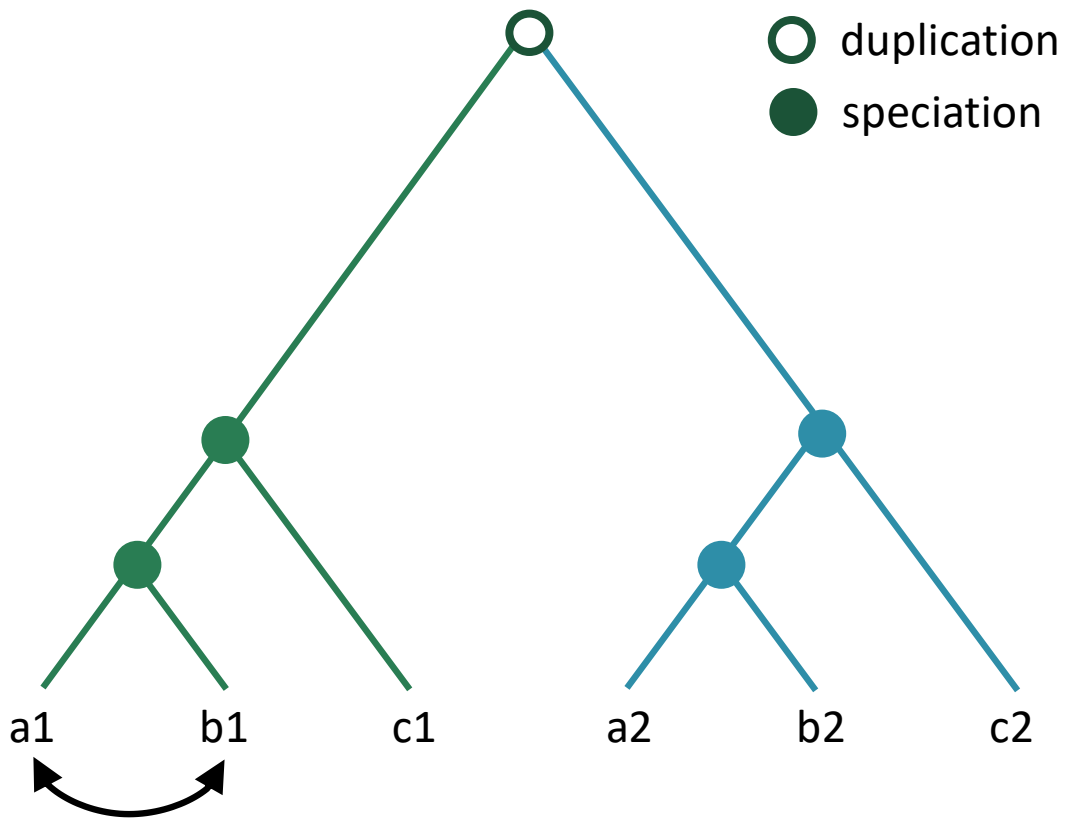
- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event



Paralogs

Homologous genes = inherited from an ancestral gene

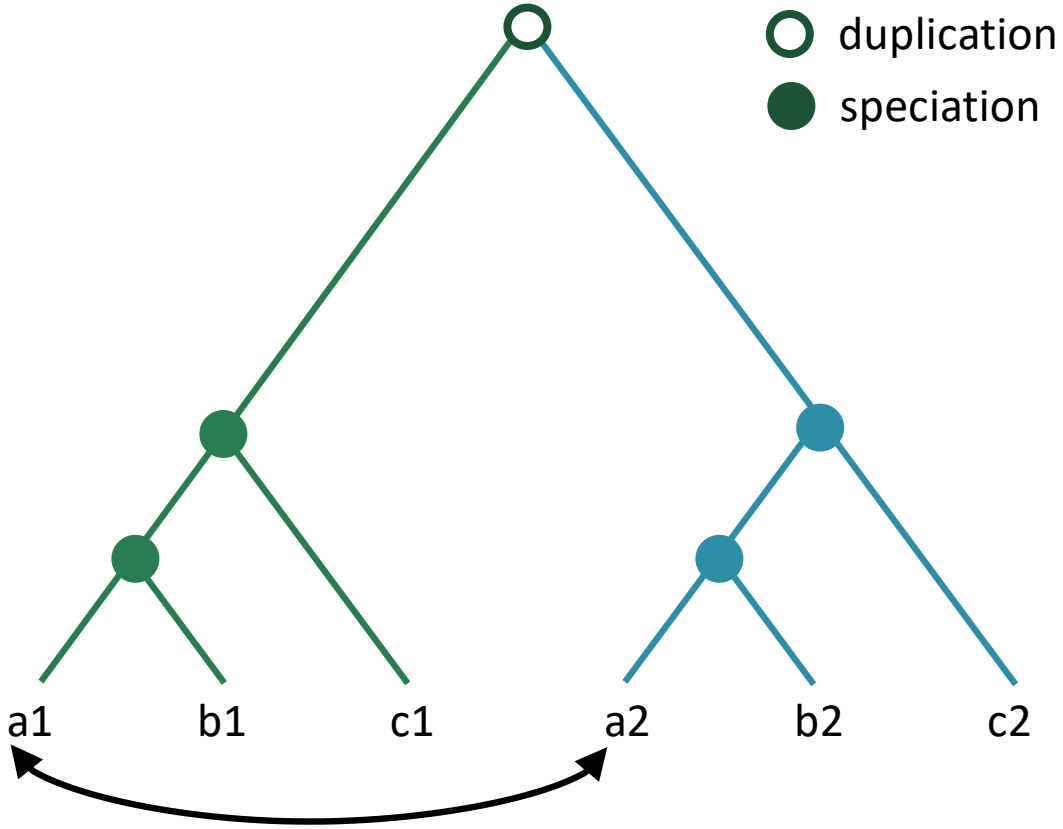
- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event



Paralogs

Homologous genes = inherited from an ancestral gene

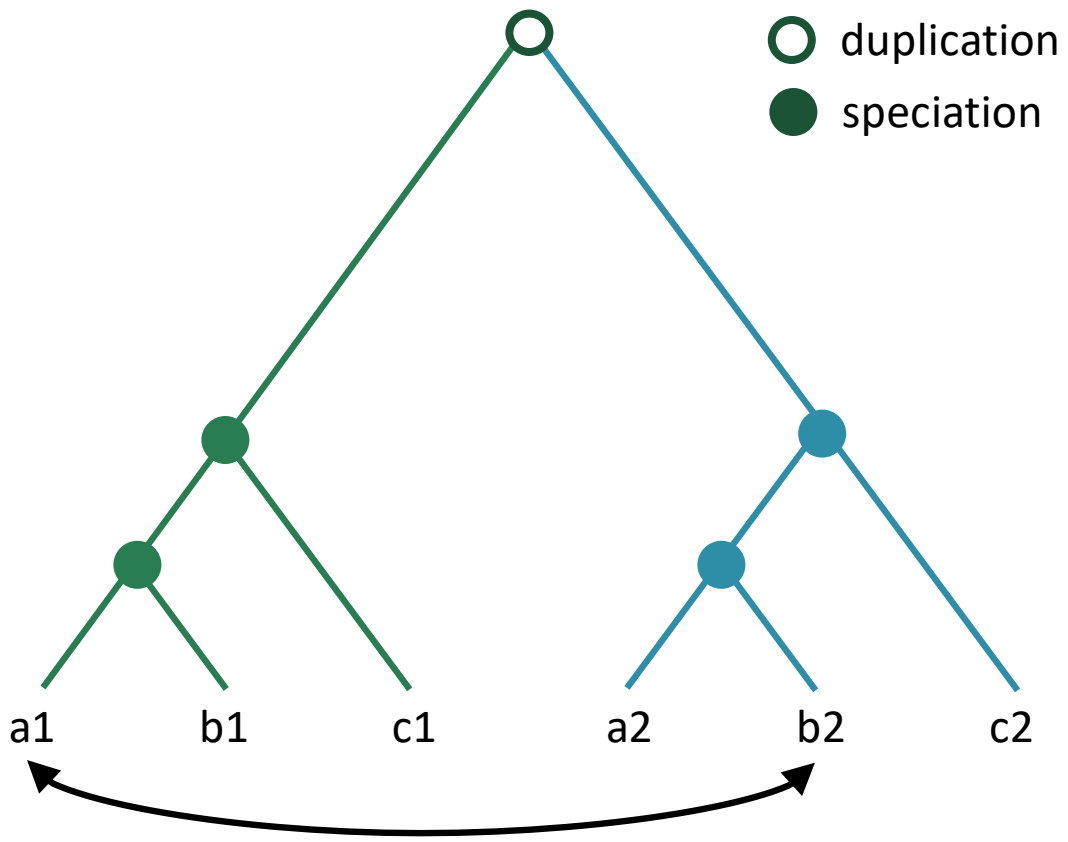
- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event



Paralogs

Homologous genes = inherited from an ancestral gene

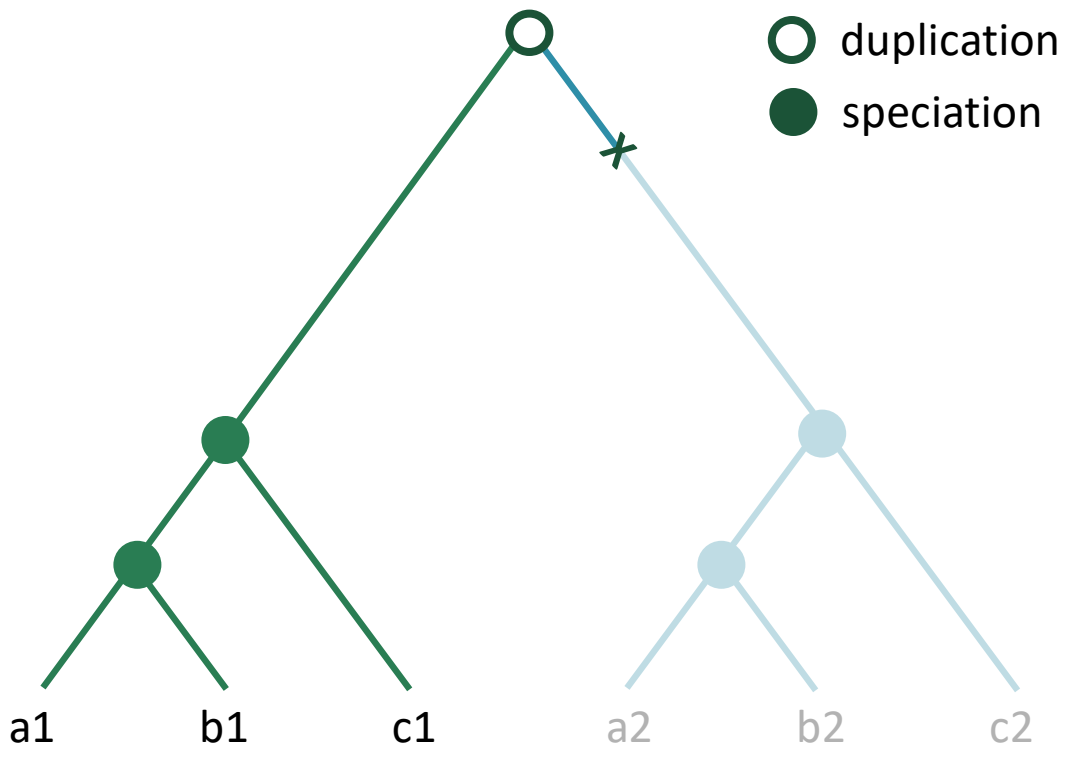
- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event



Paralogs

Homologous genes = inherited from an ancestral gene

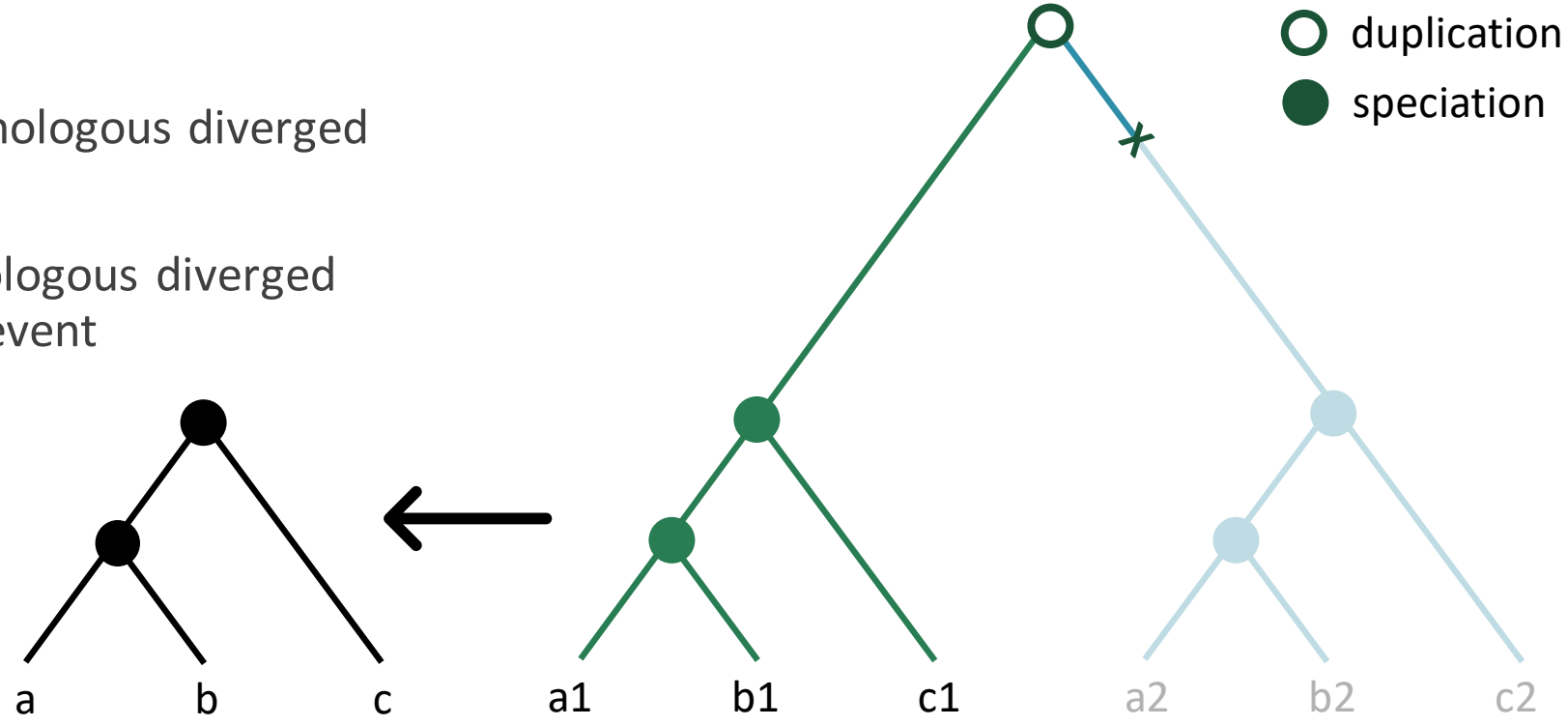
- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event



Paralogs

Homologous genes = inherited from an ancestral gene

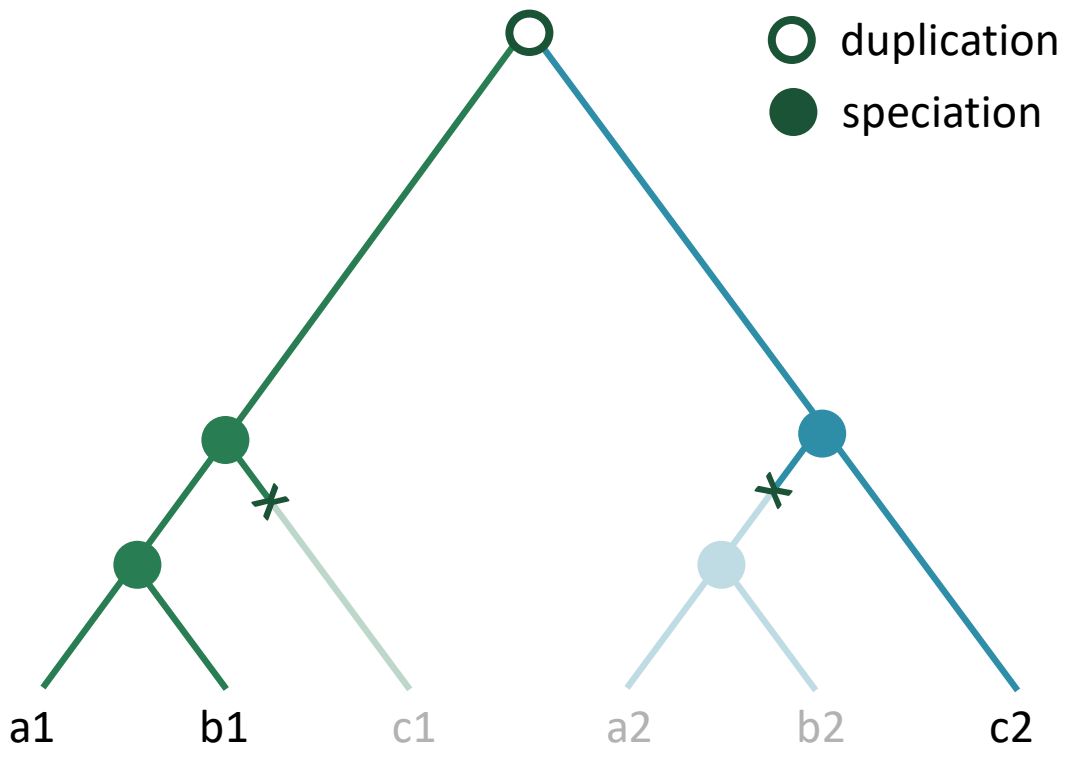
- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event



Paralogs

Homologous genes = inherited from an ancestral gene

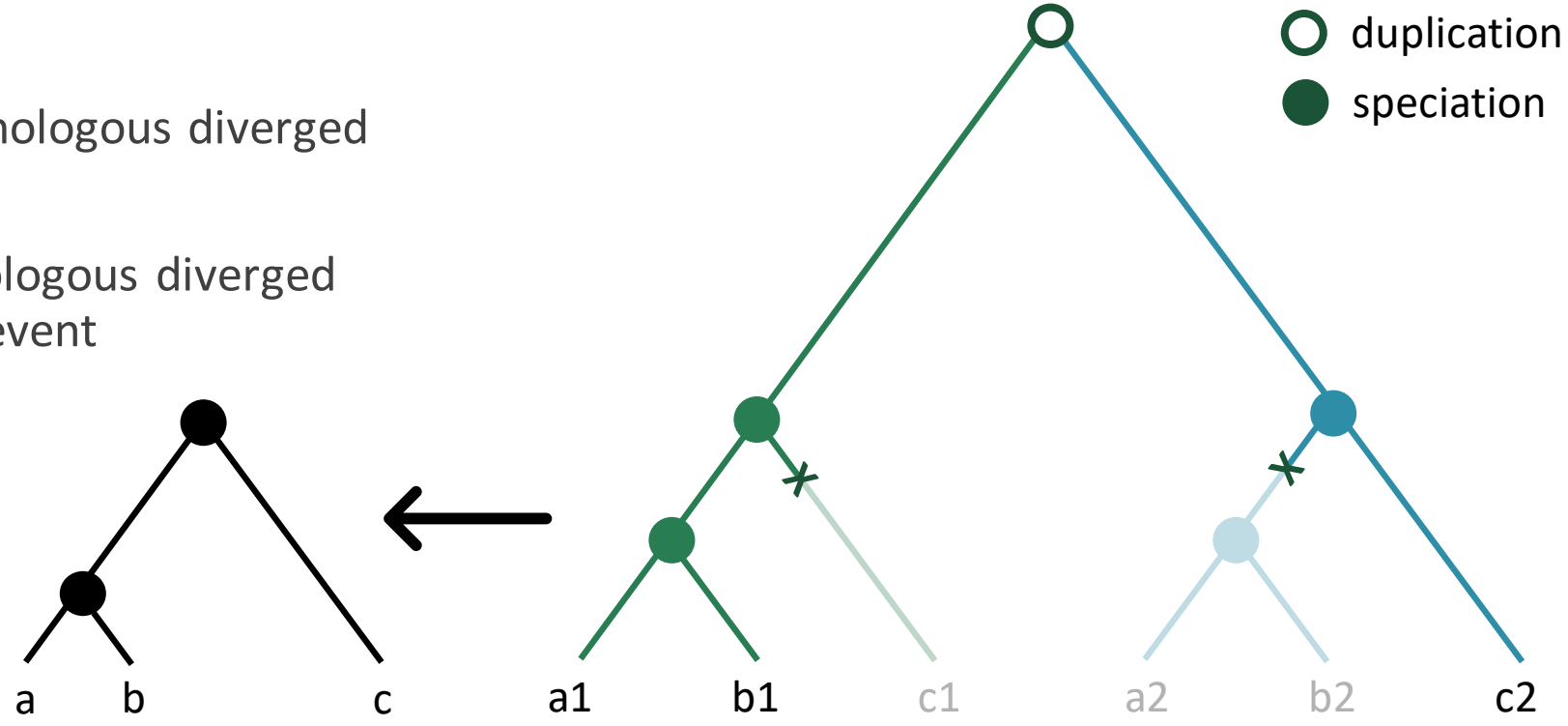
- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event



Paralogs

Homologous genes = inherited from an ancestral gene

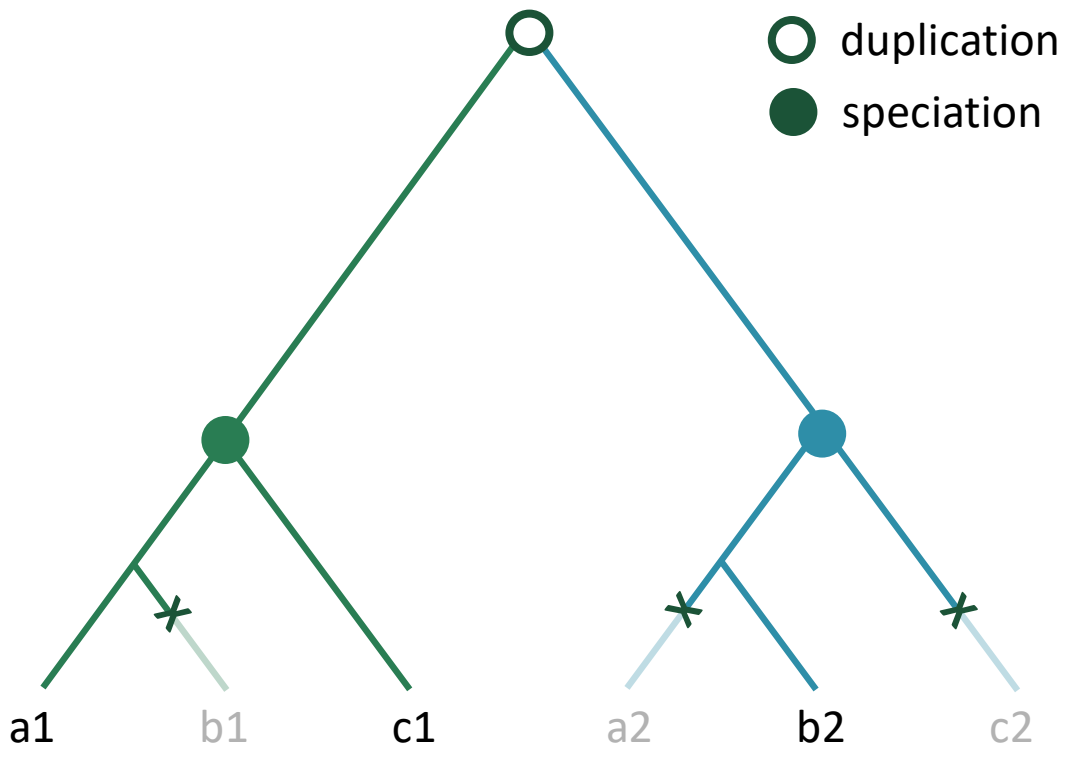
- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event



Paralogs

Homologous genes = inherited from an ancestral gene

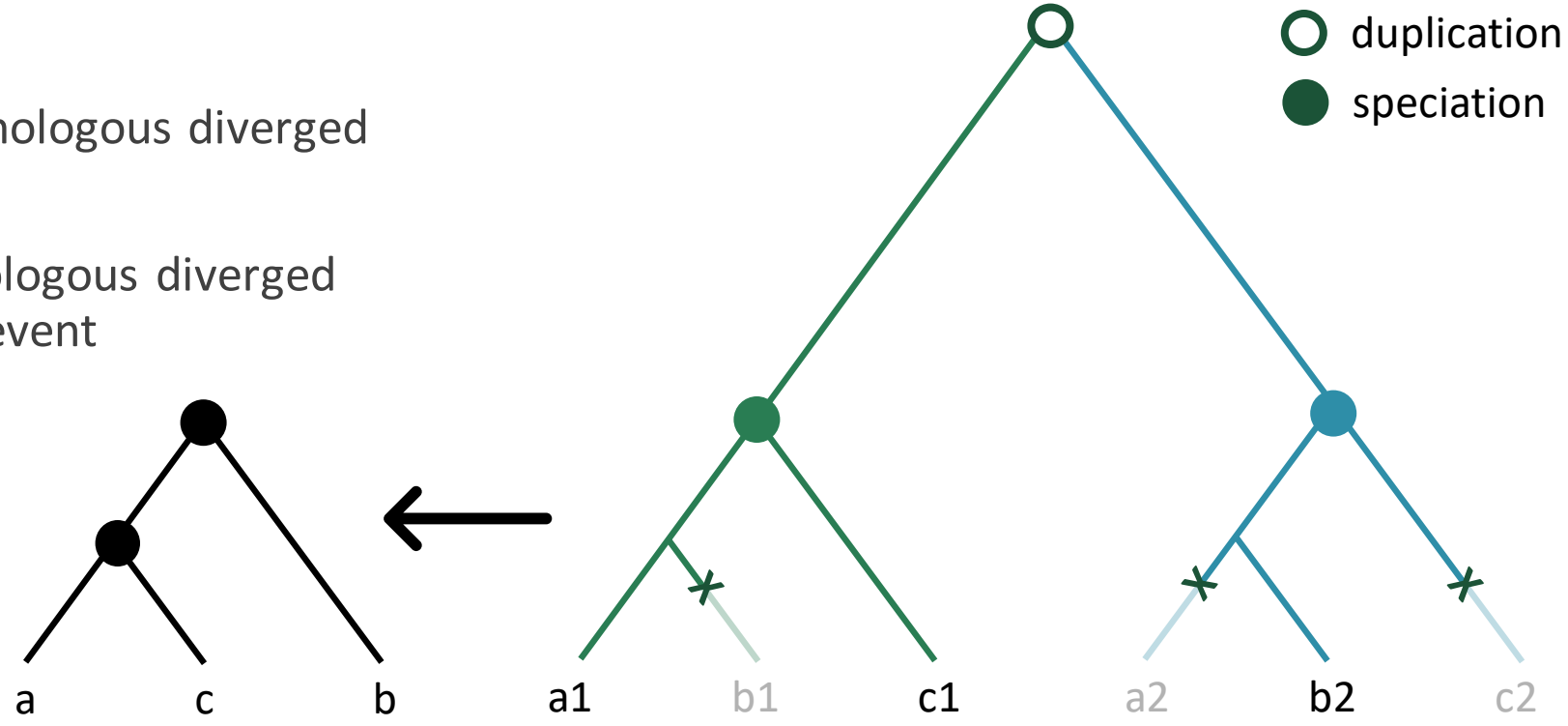
- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event



Paralogs

Homologous genes = inherited from an ancestral gene

- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event

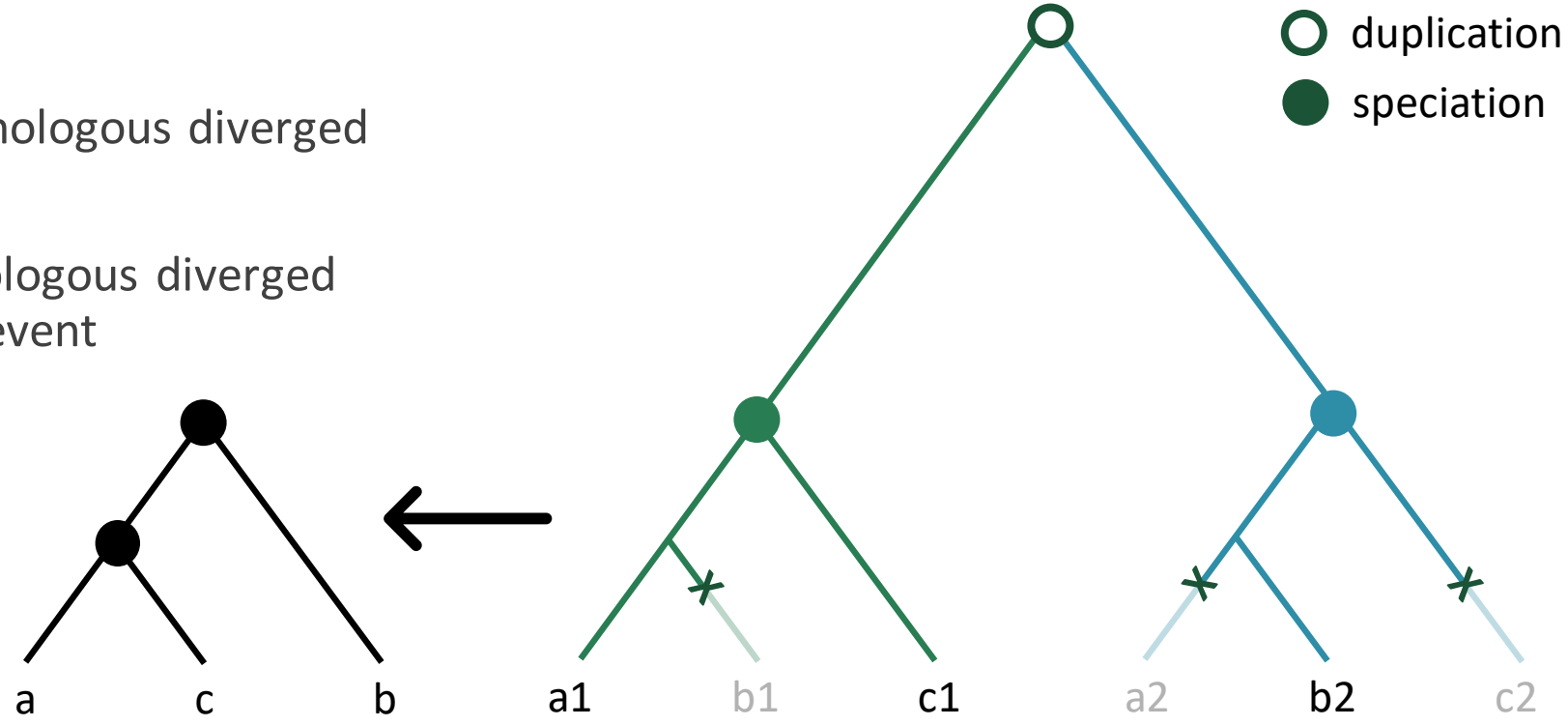


Paralogs

Homologous genes = inherited from an ancestral gene

- **Orthologous** genes = homologous diverged from a **speciation** event
- **Paralogous** genes = homologous diverged from a gene **duplication** event

<u>Need to filter out paralogs</u>



HybPiper – paralogs identification

```
hybpiper paralog_retriever
```

During the assembly (`hybpiper assembly`):
- ideally, 1 single long contig aligns to the reference sequence
- but sometimes (often!) **multiple long contigs** align to the reference sequence

HybPiper – paralogs identification

```
hybpiper paralog_retriever
```

During the assembly (`hybpiper assembly`):
- ideally, 1 single long contig aligns to the reference sequence
- but sometimes (often!) **multiple long contigs** align to the reference sequence

In this case (multiple long contigs aligned to at least 75% of the length), HybPiper will:

- generate a **paralog warning** for this locus and sample
- choose among the multiple contigs:
  - the contig that has a coverage depth >10x the other contigs, or if coverage depth similar:
  - the contig that has the greatest percent identity with the reference

HybPiper – paralogs identification

```
hybpiper paralog_retriever namelist.txt -t_dna targetfile.fasta --heatmap_filetype pdf --heatmap_dpi 300
```

HybPiper – paralogs identification

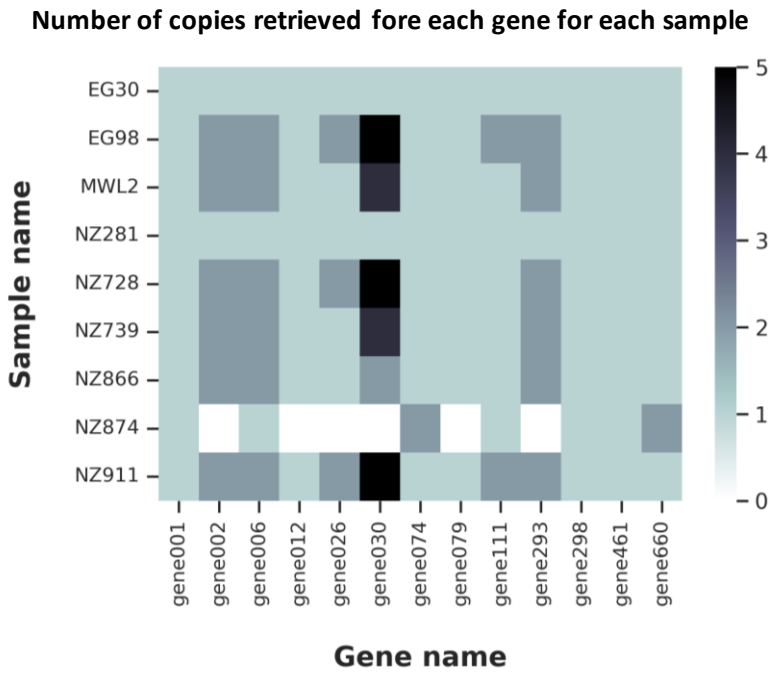
```
hybpiper paralog_retriever namelist.txt -t_dna targetfile.fasta --heatmap_filetype pdf --heatmap_dpi 300
```

→ paralog_report.tsv

Species	locus1	locus 2	locus3
Sample1	1	1	1
Sample2	5	2	2
Sample3	4	1	2

→

→ paralog_heatmap.pdf



HybPiper – paralogs identification

```
hybpiper paralog_retriever namelist.txt -t_dna targetfile.fasta --heatmap_filetype pdf --heatmap_dpi 300
```

→ paralog_report.tsv

Species	locus1	locus 2	locus3
Sample1	1	1	1
Sample2	5	2	2
Sample3	4	1	2

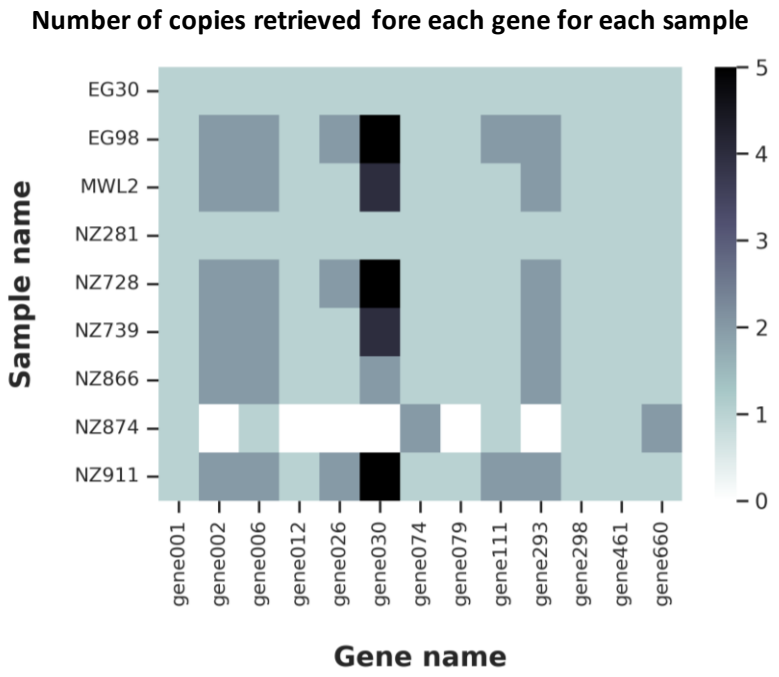
→ paralogs_all

locus1_paralogs_all.fasta
locus2_paralogs_all.fasta
…

```
>Sample1.main
ATGCATGCATGCATGCAT
>Sample1.0
ATGCATGCATGCTT
>Sample1.1
ATGCATGCATGTTT
```

1 file/locus, with all copies recovered

→ paralog_heatmap.pdf



HybPiper – paralogs identification

Quick phylogenetic inference of the paralogs to inspect the trees:

paralogs_all

```
cat locus1_paralogs_all.fasta | mafft --auto | FastTree -nt -gtr > locus1_paralogs_all.tre
```

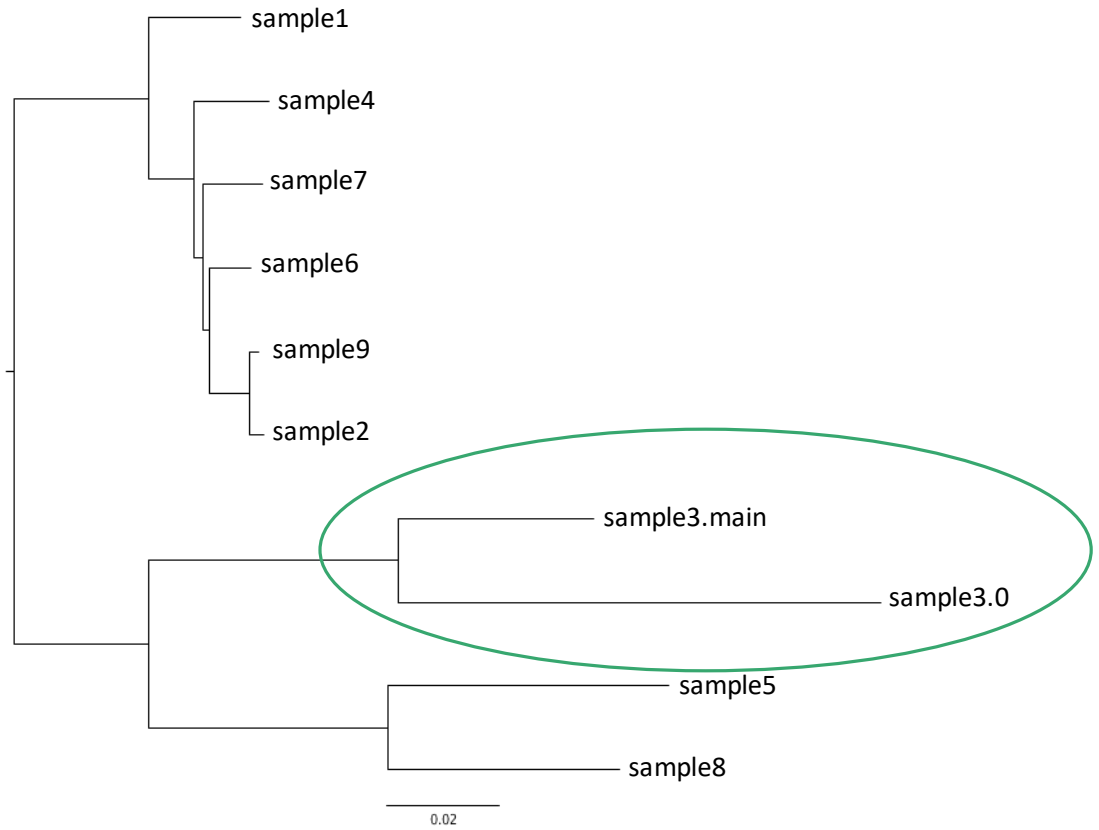
alignment

phylo inference

(possible to loop over all loci)

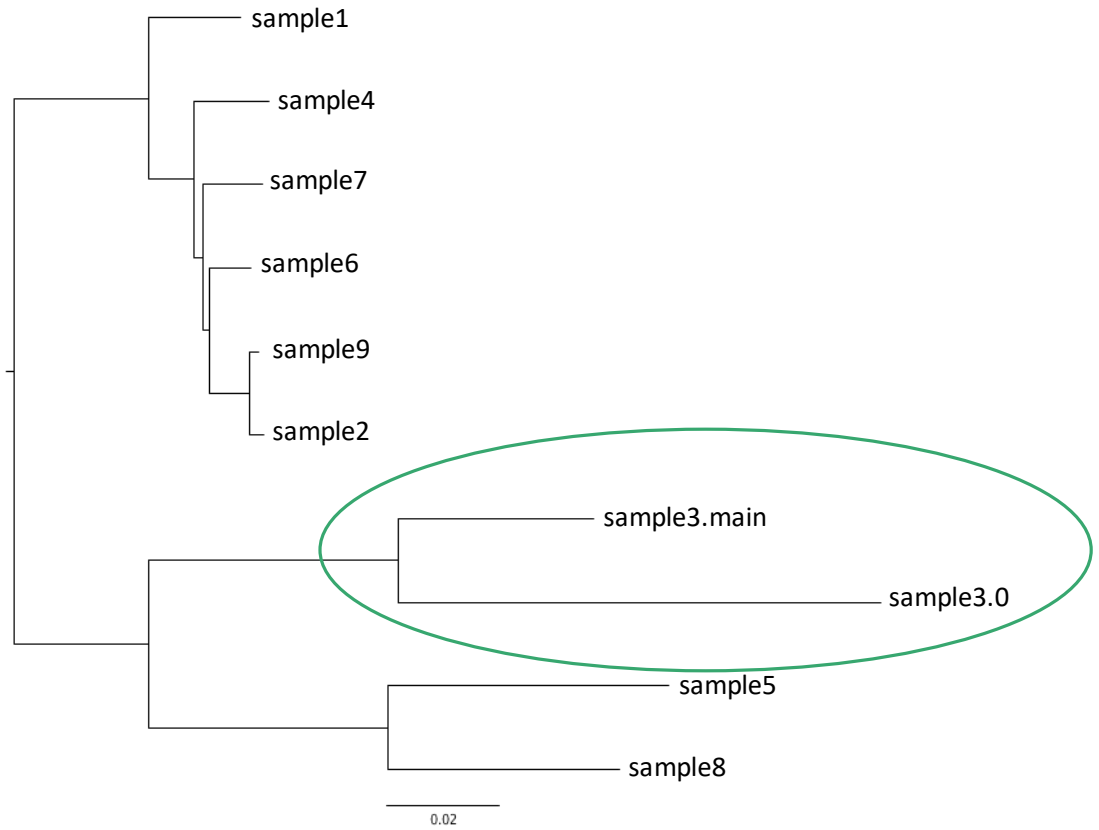
HybPiper – paralogs identification

locus1

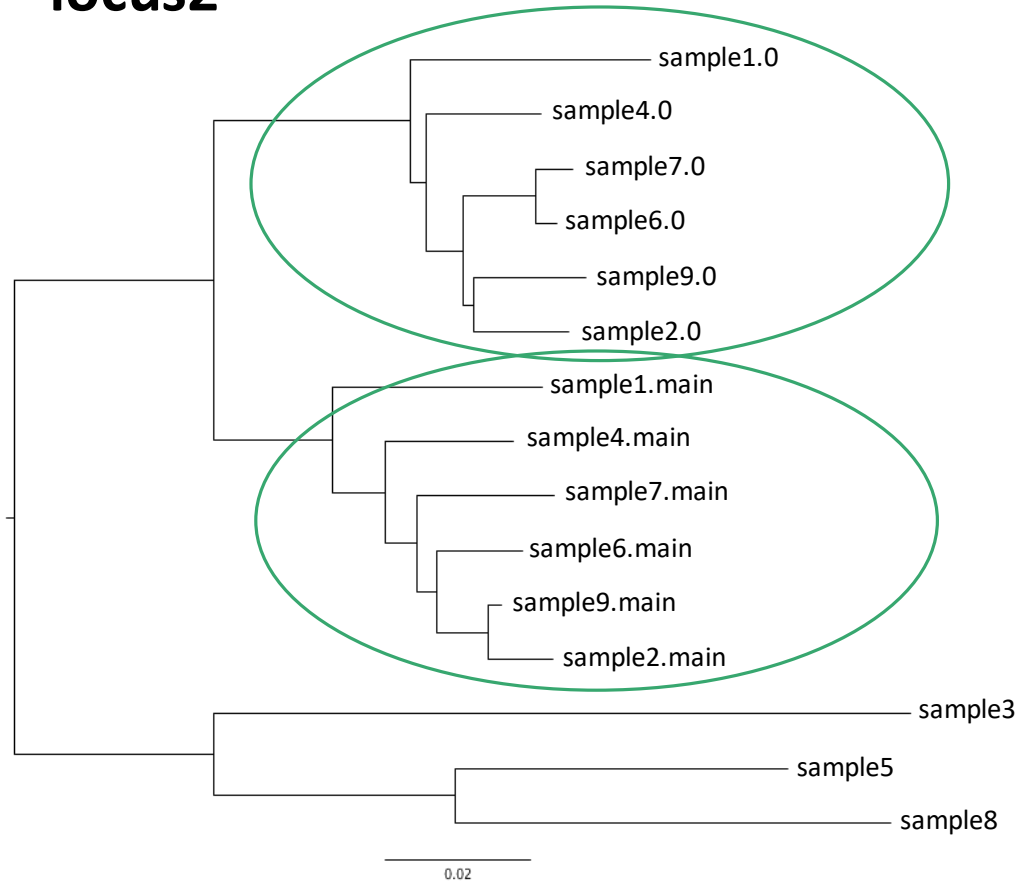


HybPiper – paralogs identification

locus1



locus2



HybPiper – paralogs identification

all .tre files in working directory

```
plot_hybpiper_paralog_trees.R
```

paralog_trees.pdf

For each locus:

- trees with all the samples (including the samples with single copy)
- trees with only the samples that have more than 1 copy for this locus

selected copy (“.main”)

other copies (.0, .1, etc.)

HybPiper – paralogs identification

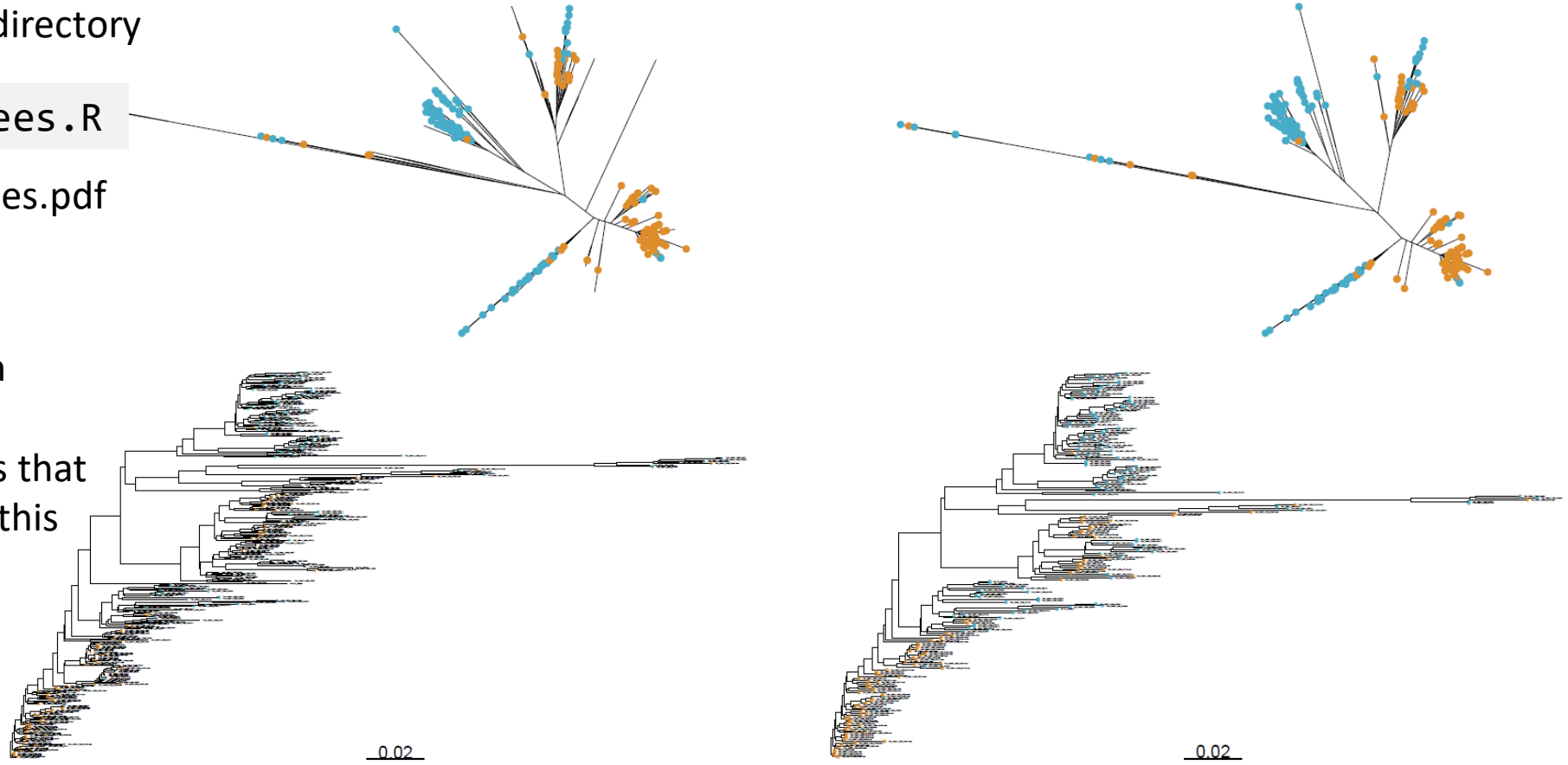
all .tre files in working directory

```
plot_hybpiper_paralog_trees.R
```

paralog_trees.pdf

For each locus:

- trees with all the samples (including the samples with single copy)
- trees with only the samples that have more than 1 copy for this locus



selected copy (".main")

other copies (.0, .1, etc.)

HybPiper – paralogs identification

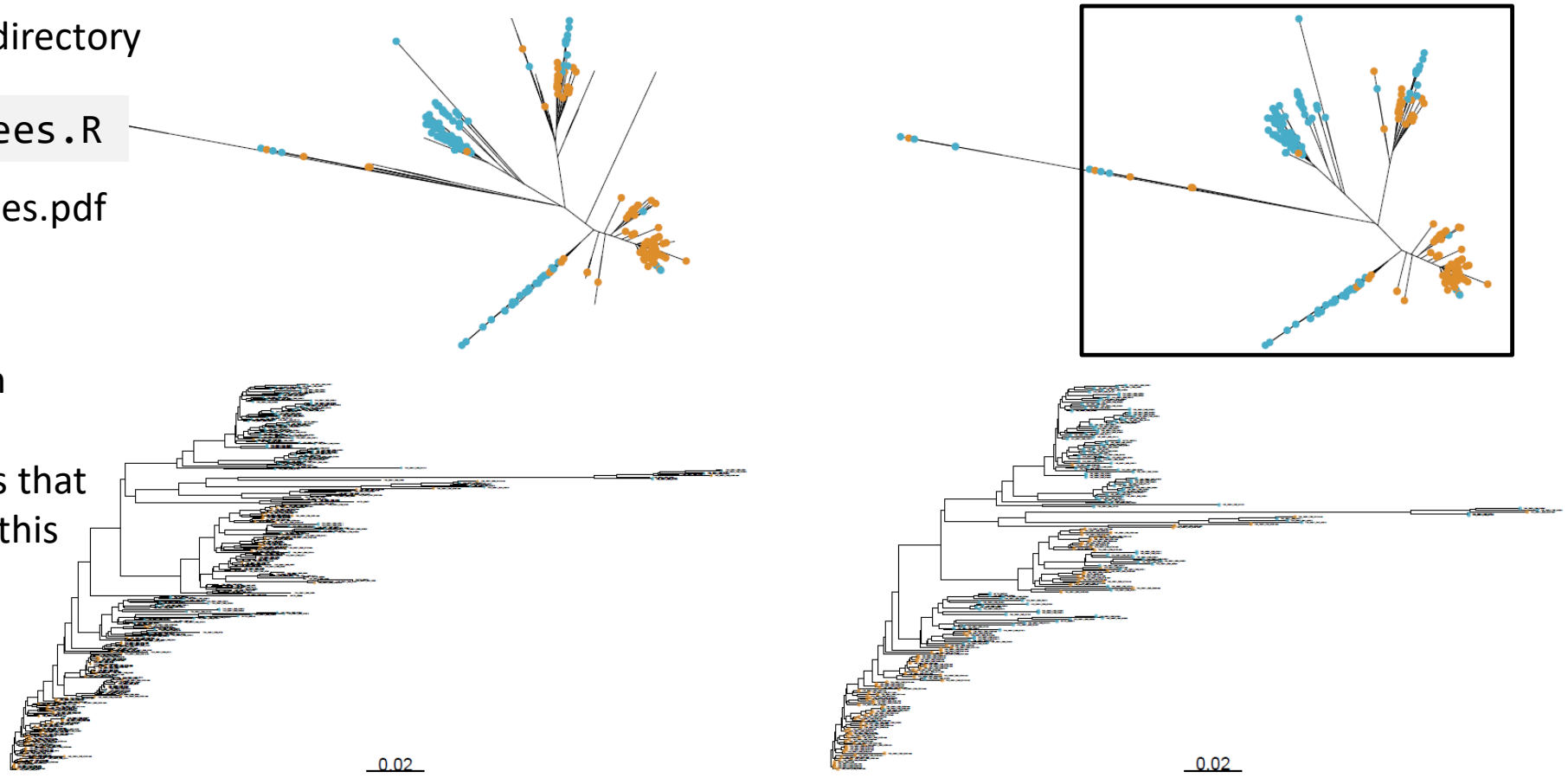
all .tre files in working directory

```
plot_hybpiper_paralog_trees.R
```

paralog_trees.pdf

For each locus:

- trees with all the samples (including the samples with single copy)
- trees with only the samples that have more than 1 copy for this locus



selected copy (".main")

other copies (.0, .1, etc.)

Hyb

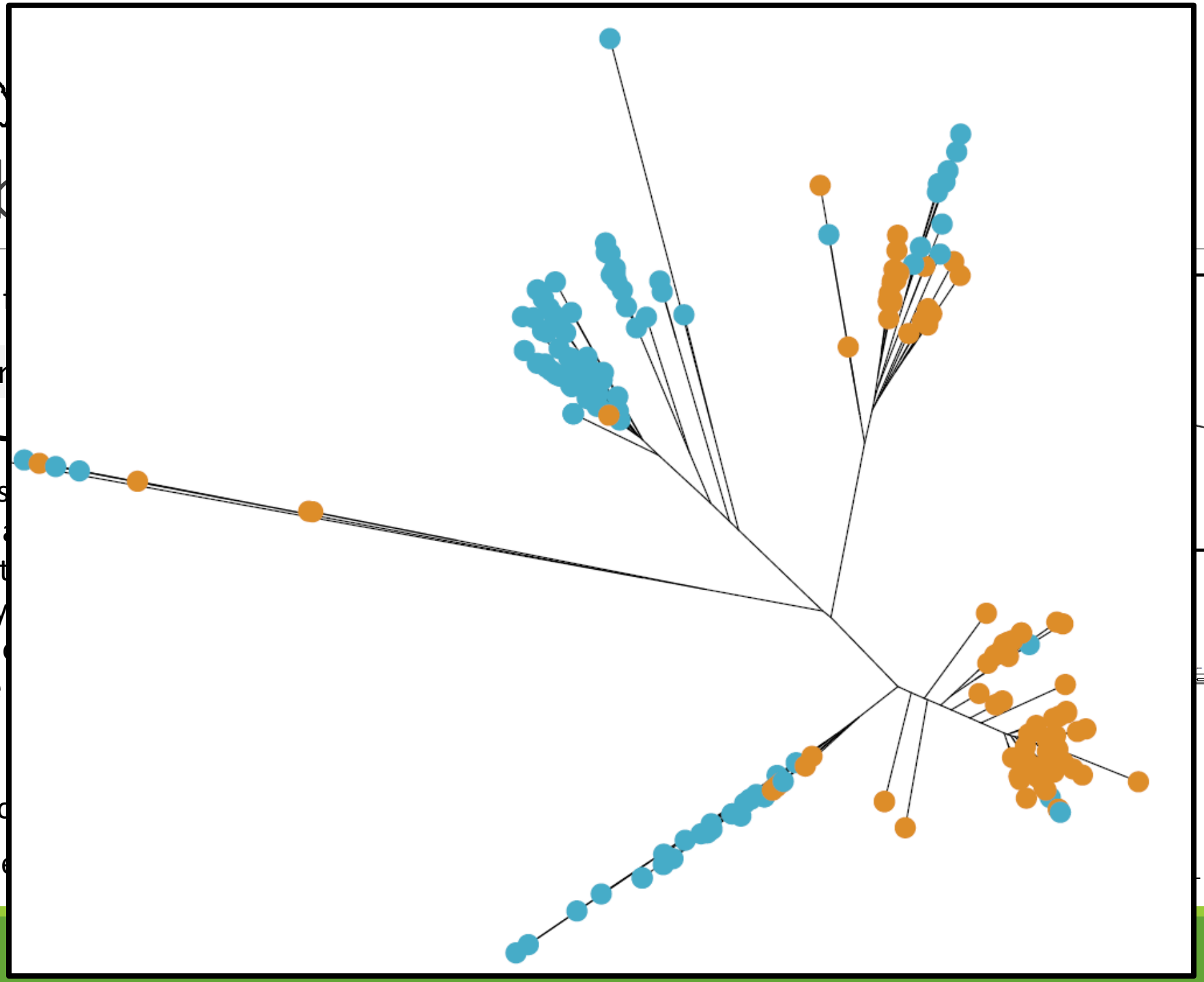
all .tre

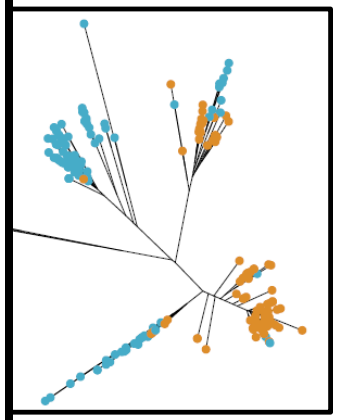
```
plot_hybpiper
```

For each locus
- trees with
(including t
single copy
- trees with
have more
locus

selected co

other copie





HybPiper – paralogs identification

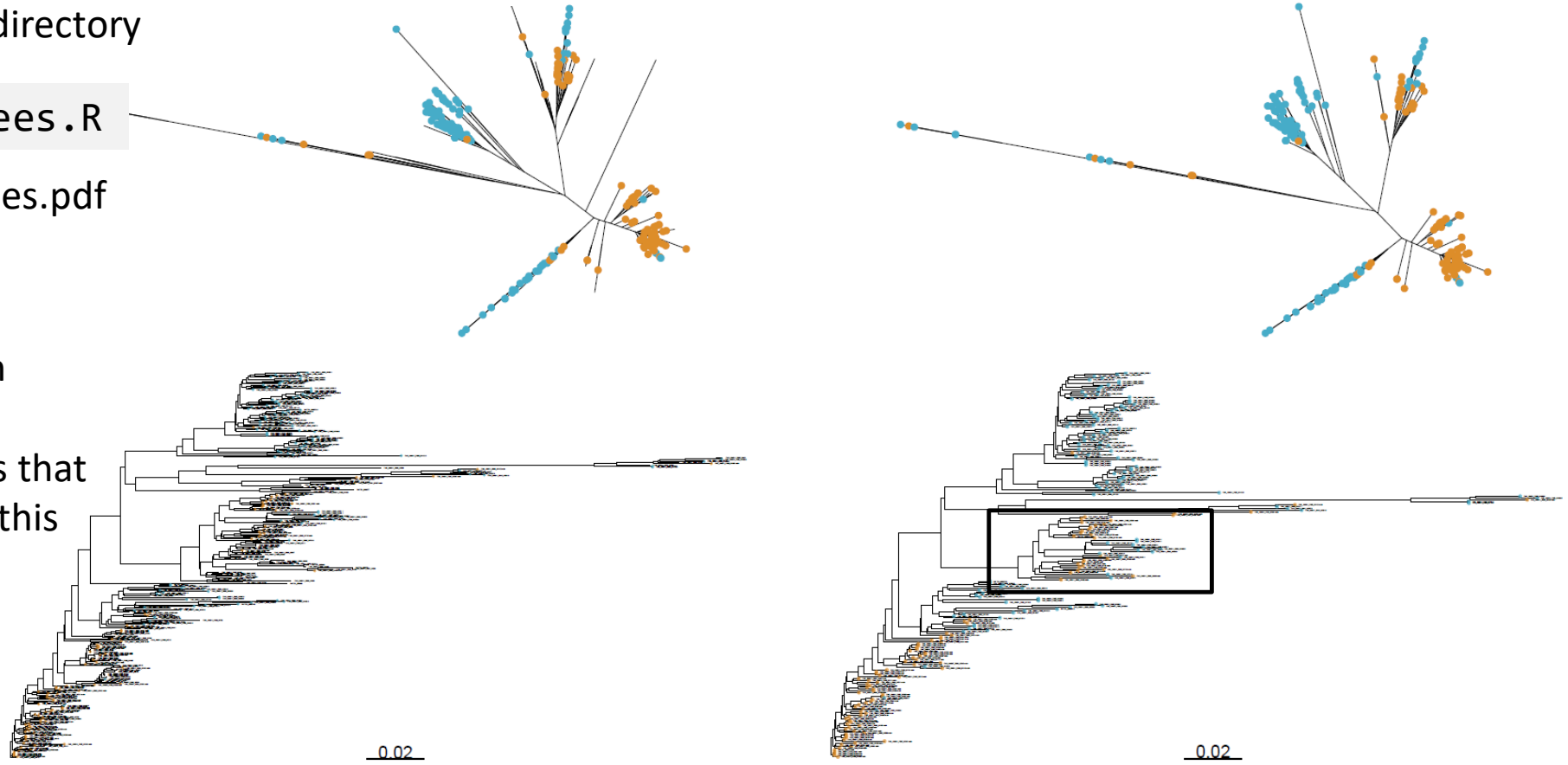
all .tre files in working directory

```
plot_hybpiper_paralog_trees.R
```

→ paralog_trees.pdf

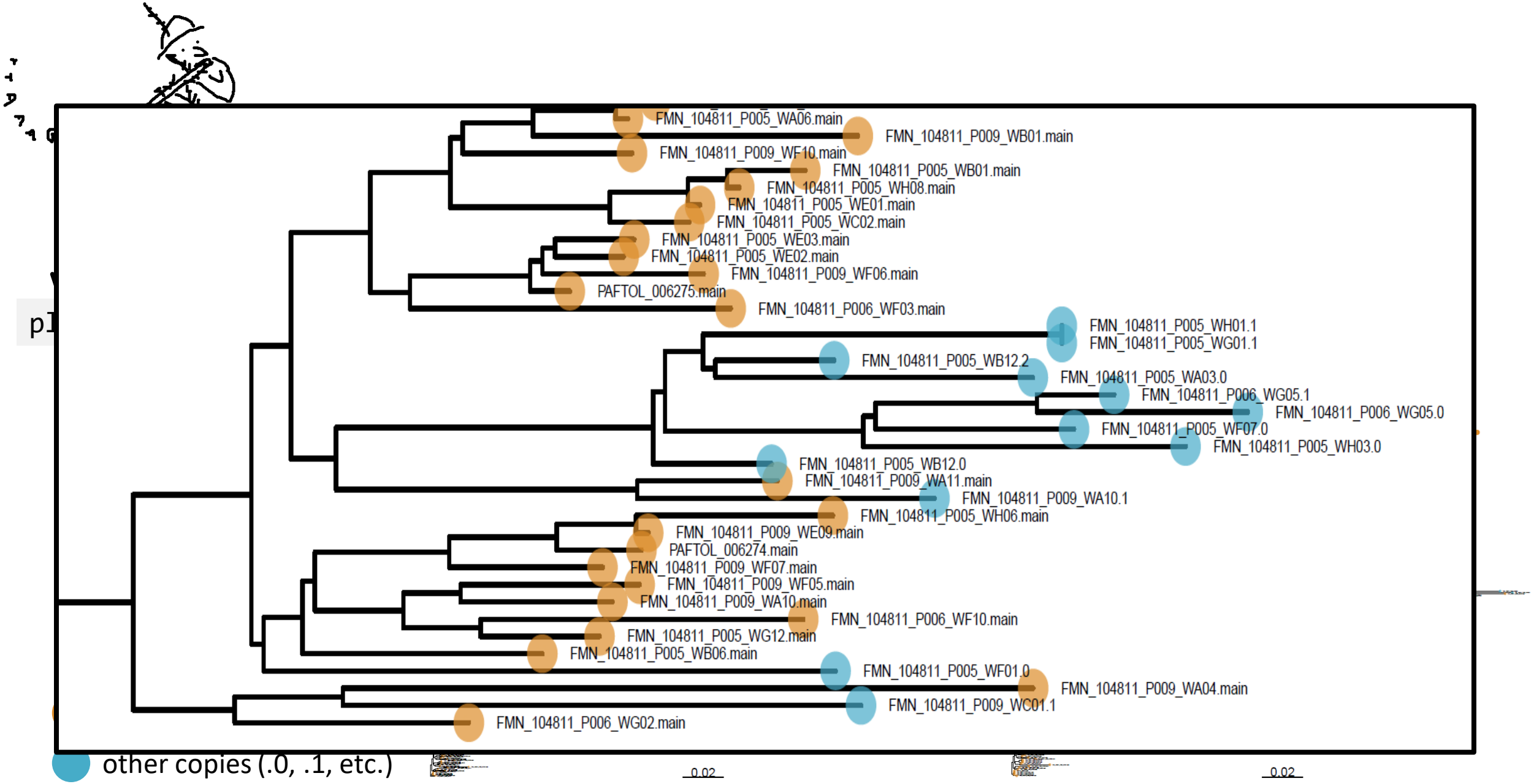
For each locus:

- trees with all the samples (including the samples with single copy)
- trees with only the samples that have more than 1 copy for this locus



selected copy (".main")

other copies (.0, .1, etc.)



other copies (.0, .1, etc.)

Loci filtering

- remove paralogs

Loci filtering

- remove paralogs
- filter on recovery (“L_N filter”)
  L = minimum length recovered
  N = minimum number of samples

  e.g. a 75_75 filter keeps only those loci for which at least 75% of the targeted sequence was recovered in at least 75% of the samples

Loci filtering

- remove paralogs
- filter on recovery ("L_N filter")
  L = minimum length recovered
  N = minimum number of samples

  e.g. a 75_75 filter keeps only those loci for which at least 75% of the targeted sequence was recovered in at least 75% of the samples

Use `loci_filtering.R` script and `genes_sequences_lengths.tsv`

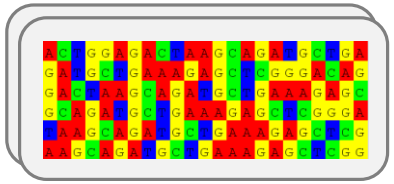
```
genes_sequences_lengths_raw <- read.table("genes_sequences_lengths.tsv",
                                          header = T, row.names = 1, sep = "\t", check.names = F)
limit_perc_length_wanted = c(0.5, 0.75)
limit_perc_nb_wanted = c(0.5, 0.75)
source("loci_filtering.R")
```

→ Lists of filtered loci
list_50_50.txt
list_75_75.txt

→ Bash commands to move the files
move_50_50.txt
move_75_75.txt

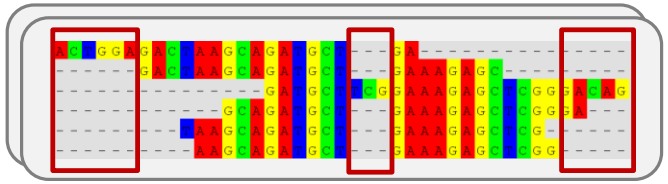
→ + additional files

Loci filtering



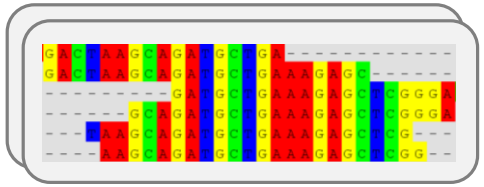
Extracted sequences

Alignment



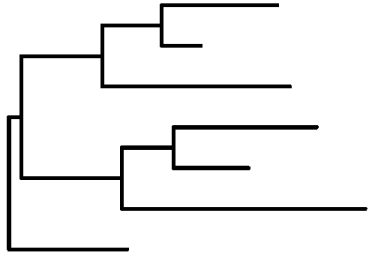
Multi-samples alignment

Alignment trimming

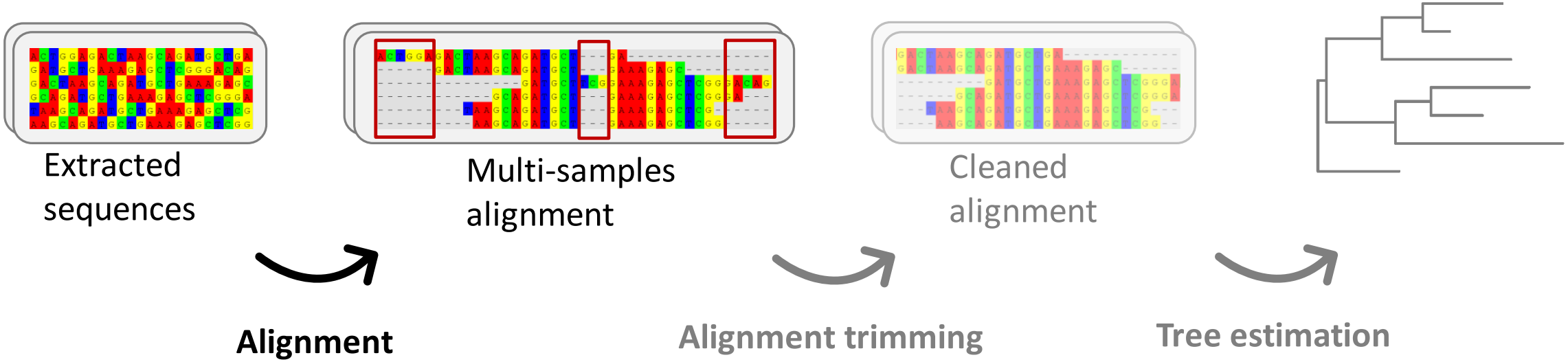


Cleaned alignment

Tree estimation



Phylogenetic reconstruction



MAFFT v7 (Katoh and Standley 2013)
Muscle5 (Edgar 2022)
Clustal (Sievers and Higgins 2018)
…

Alignment

retrieved_exons

```
mafft --thread 2 --auto locus1.FNA > aligned.locus1.FNA
```

automatic selection of the best alignment algorithm
(possible to choose specifically which algorithm to use)

Alignment

retrieved_exons

```
mafft --thread 2 --auto locus1.FNA > aligned.locus1.FNA
```

automatic selection of the best alignment algorithm
(possible to choose specifically which algorithm to use)

Looping over all loci present in working directory (locus1.FNA, locus2.FNA, …):

```
ls -1 ./ | \
while read file; do
    mafft --thread 2 --auto $file > aligned.$file
done
```

Alignment

retrieved_exons

```
mafft --thread 2 --auto locus1.FNA > aligned.locus1.FNA
```

automatic selection of the best alignment algorithm
(possible to choose specifically which algorithm to use)

Looping over all loci present in working directory (locus1.FNA, locus2.FNA, …), running in parallel :

```
ls -1 ./ | \
while read file; do
    mafft --thread 2 --auto $file > aligned.$file
done | parallel -j16
```

Alignment with reference sequence(s)

Running the alignment with the reference sequences (for target capture data) can lead to more accurate alignments, especially if aligning supercontigs (exons + introns)

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
…
```

locus1.FNA

```
>seq1-locus1
ATGCATGC…
>seq2-locus1
ATGCATGC…
>seq1-locus2
ATGCATGC…
…
```

targetfile.fasta

Alignment with reference sequence(s)

Running the alignment with the reference sequences (for target capture data) can lead to more accurate alignments, especially if aligning supercontigs (exons + introns)

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
…
```

locus1.FNA

```
>seq1-locus1
ATGCATGC…
>seq2-locus1
ATGCATGC…
>seq1-locus2
ATGCATGC…
…
```

targetfile.fasta

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
>seq1-locus1
ATGCATGC
>seq2-locus1
ATGCATGC
…
```

locus1_w_refs.FNA

Alignment with reference sequence(s)

Running the alignment with the reference sequences (for target capture data) can lead to more accurate alignments, especially if aligning supercontigs (exons + introns)

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
…
```

locus1.FNA

```
>seq1-locus1
ATGCATGC…
>seq2-locus1
ATGCATGC…
>seq1-locus2
ATGCATGC…
…
```

targetfile.fasta

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
>seq1-locus1
ATGCATGC
>seq2-locus1
ATGCATGC
…
```

locus1_w_refs.FNA

```
cat locus1.FNA > locus1_w_refs.FNA
seqkit grep -w0 –nrp locus1 targetfile.fasta >> locus1_w_refs.FNA
```

Alignment with reference sequence(s)

Running the alignment with the reference sequences (for target capture data) can lead to more accurate alignments, especially if aligning supercontigs (exons + introns)

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
…
```

locus1.FNA

```
>seq1-locus1
ATGCATGC…
>seq2-locus1
ATGCATGC…
>seq1-locus2
ATGCATGC…
…
```

targetfile.fasta

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
>seq1-locus1
ATGCATGC
>seq2-locus1
ATGCATGC
…
```

locus1_w_refs.FNA

Alignment with reference sequence(s)

Running the alignment with the reference sequences (for target capture data) can lead to more accurate alignments, especially if aligning supercontigs (exons + introns)

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
…
```

locus1.FNA

```
>seq1-locus1
ATGCATGC…
>seq2-locus1
ATGCATGC…
>seq1-locus2
ATGCATGC…
…
```

targetfile.fasta

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
>seq1-locus1
ATGCATGC
>seq2-locus1
ATGCATGC
…
```

locus1_w_refs.FNA

MAFFT

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATG----CATGCAT
>seq1-locus1
ATGC--------ATGC--
>seq2-locus1
ATGC--------ATGC--
…
```

aligned.locus1_w_refs.FNA

Alignment with reference sequence(s)

Running the alignment with the reference sequences (for target capture data) can lead to more accurate alignments, especially if aligning supercontigs (exons + introns)

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
…
```

locus1.FNA

```
>seq1-locus1
ATGCATGC…
>seq2-locus1
ATGCATGC…
>seq1-locus2
ATGCATGC…
…
```

targetfile.fasta

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
>seq1-locus1
ATGCATGC
>seq2-locus1
ATGCATGC
…
```

locus1_w_refs.FNA

MAFFT

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATG----CATGCAT
>seq1-locus1
ATGC--------ATGC--
>seq2-locus1
ATGC--------ATGC--
…
```

aligned.locus1_w_refs.FNA

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATG----CATGCAT
…
```

aligned.locus1.FNA

Alignment with reference sequence(s)

Running the alignment with the reference sequences (for target capture data) can lead to more accurate alignments, especially if aligning supercontigs (exons + introns)

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
…
```

locus1.FNA

```
>seq1-locus1
ATGCATGC…
>seq2-locus1
ATGCATGC…
>seq1-locus2
ATGCATGC…
…
```

targetfile.fasta

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATGCATGCAT
>seq1-locus1
ATGCATGC
>seq2-locus1
ATGCATGC
…
```

locus1_w_refs.FNA

MAFFT

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATG----CATGCAT
>seq1-locus1
ATGC--------ATGC--
>seq2-locus1
ATGC--------ATGC--
…
```

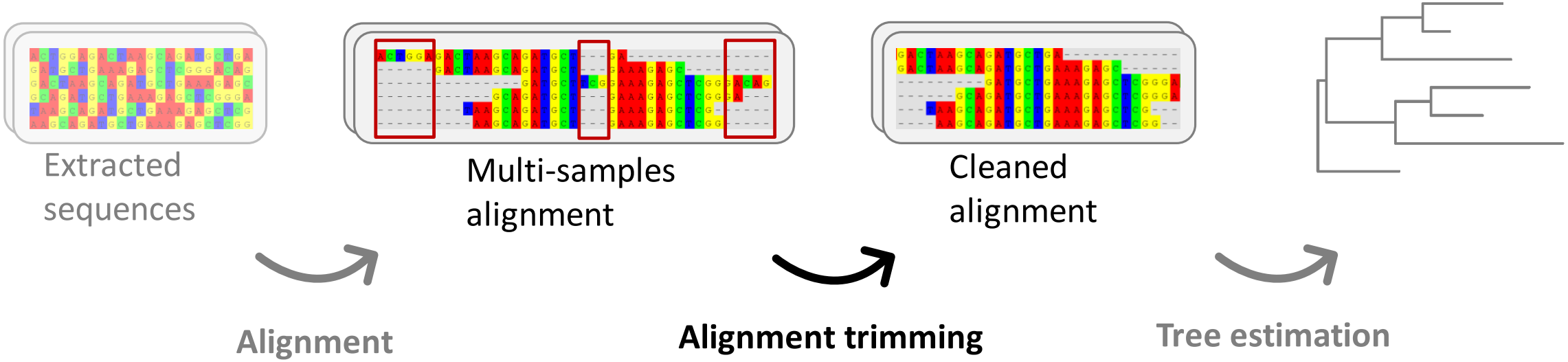
aligned.locus1_w_refs.FNA

```
>Sample1
ATGCATGCATGCATGCAT
>Sample3
ATGCATG----CATGCAT
…
```

aligned.locus1.FNA

```
seqkit grep -v -nrp locus1 aligned.locus1_w_refs.FNA > aligned.locus1.FNA
```

Phylogenetic reconstruction



ClipKIT (Steenwyk et al. 2020) Retain phylogenetically informative sites

TrimAl (Capella-Gutiérrez et al. 2009)
Gblocks (Talavera and Castresana 2007)
Remove sites ambiguously aligned sites, or with high levels of missing data

Alignment trimming

2 options (but see other options in ClipKIT documentation):

```
clipkit aligned.locus1.FNA -m smart-gap -o aligned.locus1.FNA.clipkit
```

Remove gappy sites

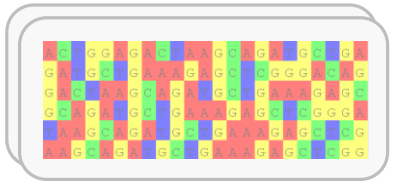
```
clipkit aligned.locus1.FNA -m kpic-smart-gap -o aligned.locus1.FNA.clipkit
```

Remove gappy sites and keep only parsimony informative and constant sites

↳ sites that contain at least 2 character states that occur in at least 2 samples

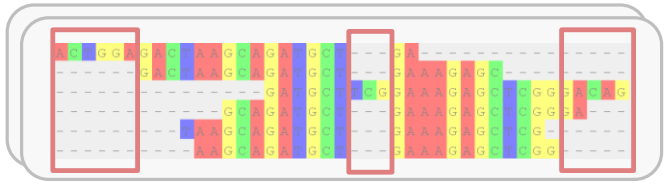
Rename the files to remove the ".clipkit" suffix: `rename -v '.FNA.clipkit' '.FNA' *`

Phylogenetic reconstruction



Extracted sequences

Alignment

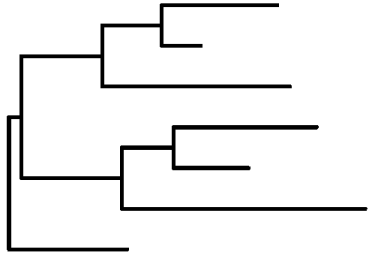


Multi-samples alignment

Alignment trimming

Cleaned alignment

Tree estimation



Tree estimation

- Distance-based methods: neighbor joining (NJ), UPGMA, …
- Character-based methods:

Tree estimation

- Distance-based methods: neighbor joining (NJ), UPGMA, …
- Character-based methods:
  - **Maximum parsimony** (MP): inferred tree = tree that requires the minimum number of changes in characters (nucleotides, amino-acids) to explain the data (the alignment)

Tree estimation

- Distance-based methods: neighbor joining (NJ), UPGMA, …
- Character-based methods:
  - **Maximum parsimony** (MP): inferred tree = tree that requires the minimum number of changes in characters (nucleotides, amino-acids) to explain the data (the alignment)
  - Explicit models of sequence evolution (substitution models, branch lengths):

Tree estimation

- Distance-based methods: neighbor joining (NJ), UPGMA, …
- Character-based methods:
  - **Maximum parsimony** (MP): inferred tree = tree that requires the minimum number of changes in characters (nucleotides, amino-acids) to explain the data (the alignment)
  - Explicit models of sequence evolution (substitution models, branch lengths):

Maximum likelihood (ML)

- ML tree = (single) tree that best explains the data given the model(s)

Bayesian inference (BI)

- Search for sets of plausible trees and average a "best" tree over the set of plausible trees
- The space of the search is limited by **prior** information and by the data
- Incorporates **uncertainty** around the parameters in the models (prior information)
- **MCMC** algorithm searches the space of parameters and tree topologies
- Computes posterior probabilities (= probability that a tree is true given the data and prior)
- Summarizes a "best" tree and **uncertainty** around the clades

Tree estimation

- Distance-based methods: neighbor joining (NJ), UPGMA, …
- Character-based methods:
  - **Maximum parsimony** (MP): inferred tree = tree that requires the minimum number of changes in characters (nucleotides, amino-acids) to explain the data (the alignment)
  - Explicit models of sequence evolution (substitution models, branch lengths):

Maximum likelihood (ML)

- ML tree = (single) tree that best explains the data given the model(s)

Bayesian inference (BI)

- Search for sets of plausible trees and average a "best" tree over the set of plausible trees
- The space of the search is limited by **prior** information and by the data
- Incorporates **uncertainty** around the parameters in the models (prior information)
- **MCMC** algorithm searches the space of parameters and tree topologies
- Computes posterior probabilities (= probability that a tree is true given the data and prior)
- Summarizes a "best" tree and **uncertainty** around the clades

IQTREE
RAxML
PAML, PhyML, FastTree, …

BEAST2
RevBayes
MrBayes, …

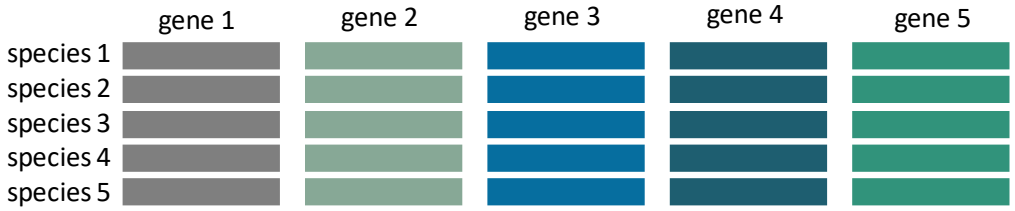
Species tree estimation

GENE TREES APPROACH

CONCATENATION APPROACH

Species tree estimation

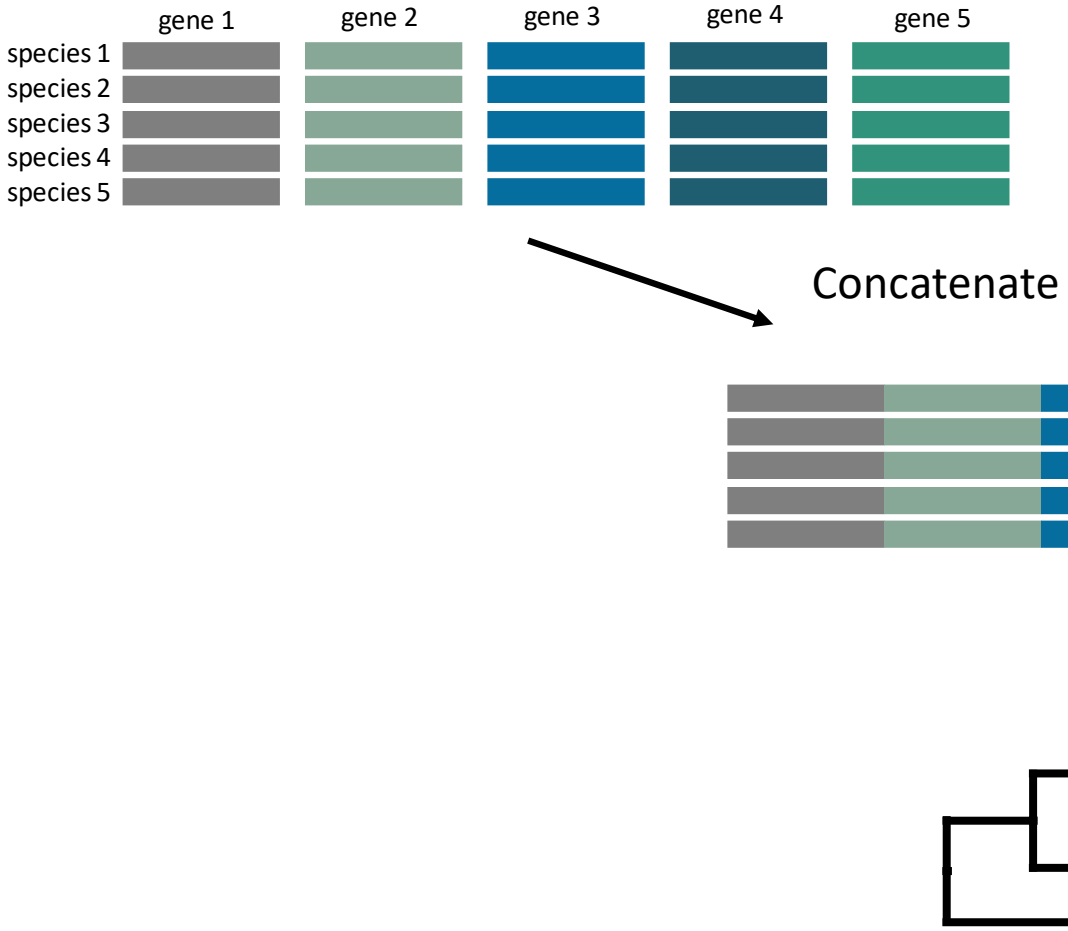
GENE TREES APPROACH



CONCATENATION APPROACH

Species tree estimation

GENE TREES APPROACH

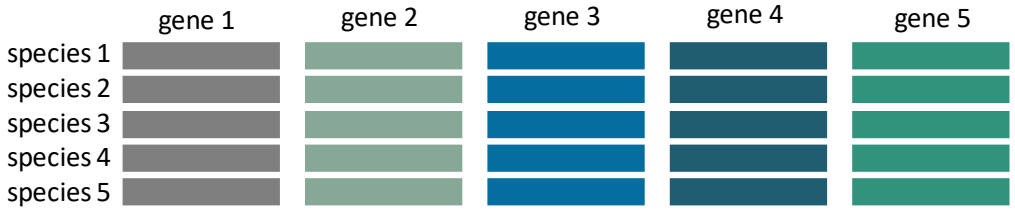


CONCATENATION APPROACH

Species tree estimation

GENE TREES APPROACH

CONCATENATION APPROACH

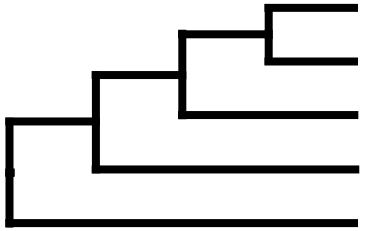


Estimate (IQTREE, …)

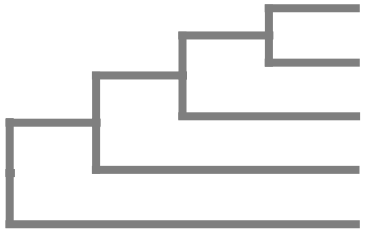
Concatenate

Supermatrix

Summarize (ASTRAL, …)



Estimate (IQTREE, …)



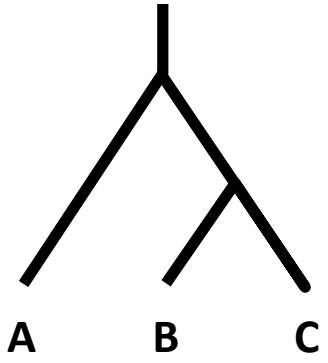
Species tree estimation

CONCATENATION APPROACH

Assumption: all genes have the same evolutionary history.

Assumption infringed by various biological processes such as horizontal gene transfer, hybridization, or incomplete lineage sorting.

Species tree

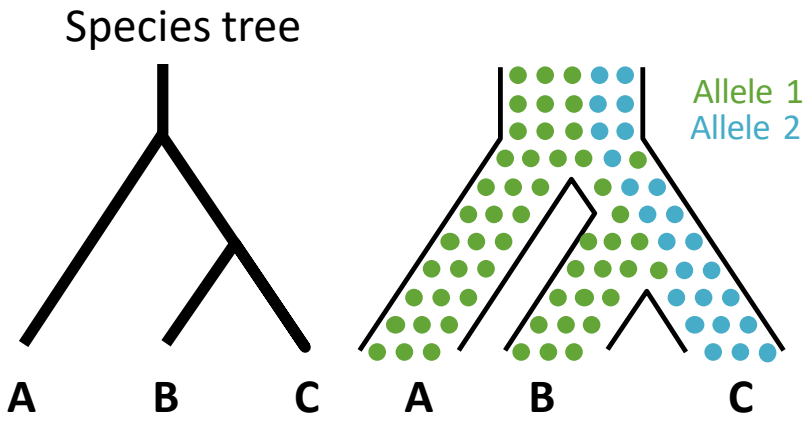


Species tree estimation

CONCATENATION APPROACH

Assumption: all genes have the same evolutionary history.

Assumption infringed by various biological processes such as horizontal gene transfer, hybridization, or incomplete lineage sorting.

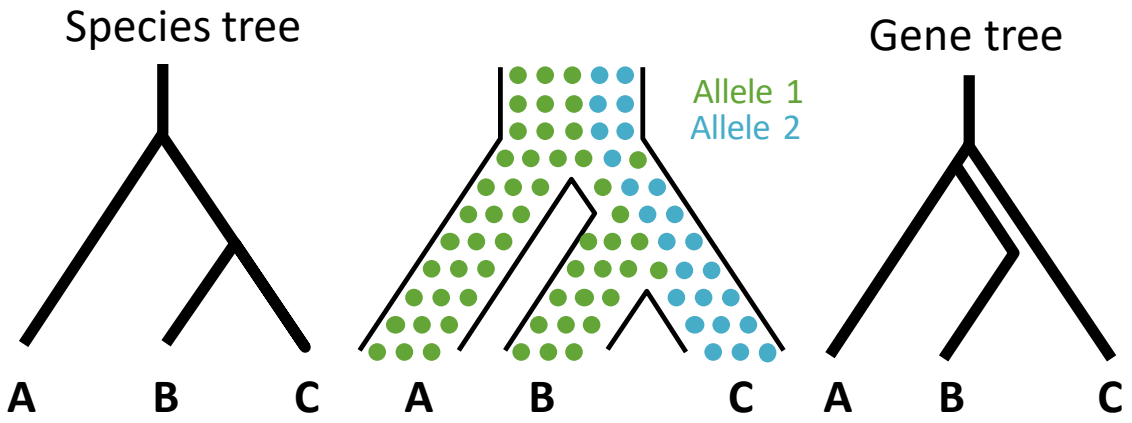


Species tree estimation

CONCATENATION APPROACH

Assumption: all genes have the same evolutionary history.

Assumption infringed by various biological processes such as horizontal gene transfer, hybridization, or incomplete lineage sorting.



Species tree estimation

GENE TREES APPROACH

Accounts for incomplete lineage sorting (ILS), under the Multi-Species Coalescent (MSC) model.

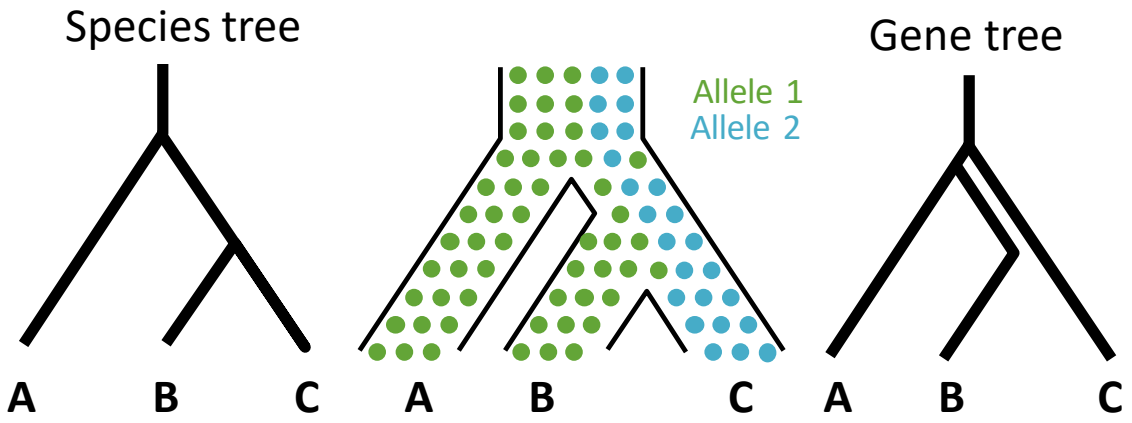
Assumptions: genes trees are estimated accurately.

ASTRAL (Accurate Species TRee Algorithm)
(Mirarab et al. 2014, Zhang et al. 2018)

CONCATENATION APPROACH

Assumption: all genes have the same evolutionary history.

Assumption infringed by various biological processes such as horizontal gene transfer, hybridization, or incomplete lineage sorting.



Species tree estimation

GENE TREES APPROACH

Accounts for incomplete lineage sorting (ILS), under the Multi-Species Coalescent (MSC) model.

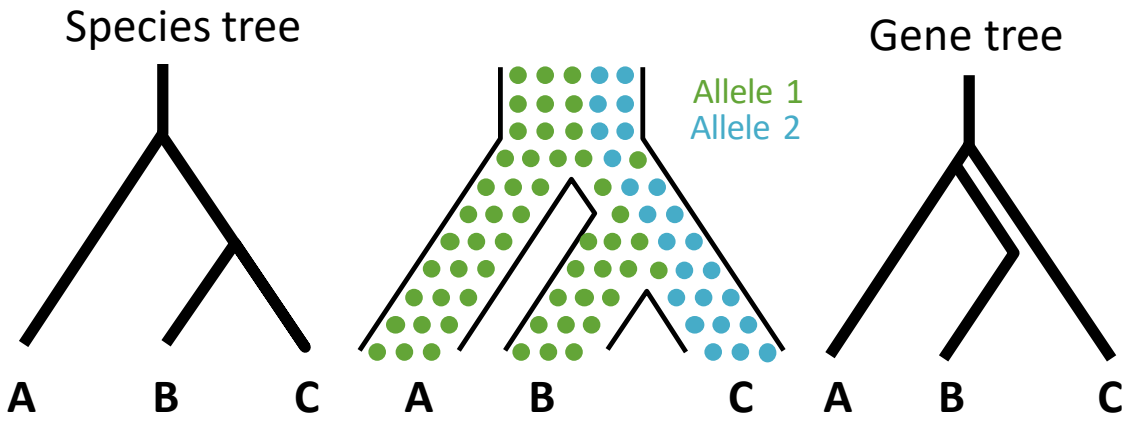
Assumptions: genes trees are estimated accurately.

ASTRAL (Accurate Species TRee Algorithm)
(Mirarab et al. 2014, Zhang et al. 2018)

CONCATENATION APPROACH

Assumption: all genes have the same evolutionary history.

Assumption infringed by various biological processes such as horizontal gene transfer, hybridization, or incomplete lineage sorting.



use both approaches cautiously and compare topologies to detect incongruences

Concatenation approach

- Make sure all the alignments have the same length:

```
bash ~/scripts/fill_fasta.sh namelist.txt
```

`fill_fasta.sh`: ensures all lines in the .FNA files are the same length (fills with gaps if not)

Concatenation approach

- Make sure all the alignments have the same length:

```
bash ~/scripts/fill_fasta.sh namelist.txt
```

`fill_fasta.sh`: ensures all lines in the .FNA files are the same length (fills with gaps if not)

- Concatenate the trimmed alignments (`aligned.locus1.FNA`, …):

```
pxcat -s *.FNA -o concat_all.fasta -p concat_all.partitions
```

`pxcat` from phyx tool (Brown et al. 2017)

Concatenation approach

- Infer tree with **RAxML**:

Number of threads (CPU) to use

Substitution model

```
raxmlHPC-PTHREADS -f a -x 12345 -p 12345 -T 2 -# 100 -m GTRGAMMA -o outgroup_sample ...
```

Select the algorithm (cf. manual)

Random seed numbers

Number of bootstrap replicates

Name of outgroup sample

```
... -O -q ./ concat_all.partitions -s ./concat_all.fasta -n my_analysis
```

Partition specification file
(generated by phyx before)

Input alignment file

Name of output file

Concatenation approach

- Infer tree with **IQTREE**: no need to concatenate alignment in a single file beforehand

Input alignment folder (here current directory)

Automatic model selection for each partition
Merge partitions with the same model

Number of Ultra-Fast Bootstrap replicates

Specification of number of threads (CPU) to use

```
iqtree2 -p ./ -pre my_analysis -m MFP+MERGE -msub nuclear -B 1000 -bnni -T AUTO -ntmax 16
```

Prefix for output file names

Restrict model selection to model designed for nuclear data

Perform an additional step to further optimize UFBoot trees by nearest neighbor interchange (NNI)

See IQTREE Documentation to tweak the different parameters according to what you want to do.

Gene trees approach

- Infer gene trees with **IQTREE** (also possible with RAxML or other program)

```
iqtree2 -s aligned.locus1.FNA -m MFP+MERGE -B 1000 -bnni -T AUTO -ntmax 2 -mem 8G
```

Gene trees approach

- Infer gene trees with **IQTREE** (also possible with RAxML or other program)

```
iqtree2 -s aligned.locus1.FNA -m MFP+MERGE -B 1000 -bnni -T AUTO -ntmax 2 -mem 8G
```

- In a loop:

```
FILES=*.FNA

for f in $FILES
do
  iqtree2 -s $f -m MFP+MERGE -B 1000 -bnni -T AUTO -ntmax 2 -mem 8G
done
```

Gene trees approach

- Infer gene trees with **IQTREE** (also possible with RAxML or other program)

```
iqtree2 -s aligned.locus1.FNA -m MFP+MERGE -B 1000 -bnni -T AUTO -ntmax 2 -mem 8G
```

- In a loop, in parallel:

```
FILES=*.FNA
touch iqtree_parallel.txt
for f in $FILES
do
  echo "iqtree2 -s $f -m MFP+MERGE -B 1000 -bnni -T AUTO -ntmax 2 -mem 8G" >> iqtree_parallel.txt
done

parallel -j 8 < iqtree_parallel.txt
```

Gene trees approach

- Infer gene trees with **IQTREE** (also possible with RAxML or other program)

```
iqtree2 -s aligned.locus1.FNA -m MFP+MERGE -B 1000 -bnni -T AUTO -ntmax 2 -mem 8G
```

- In a loop, in parallel:

```
FILES=*.FNA
touch iqtree_parallel.txt
for f in $FILES
do
  echo "iqtree2 -s $f -m MFP+MERGE -B 1000 -bnni -T AUTO -ntmax 2 -mem 8G" >> iqtree_parallel.txt
done

parallel -j 8 < iqtree_parallel.txt
```

Number of CPU and RAM: in this example, we need **8** x **2** = 16 CPU threads and **8** x **8** = 64 Gb RAM

Gene trees approach

- Filter gene trees (remove putative paralogs, 75_75 filter, ...)
- Group the filtered gene trees into a single file

```
cat *.FNA.treefile > all.trees
```

- Collapse branches with bootstrap support below 10 (Zhang et al. 2018) using newick utils

```
nw_ed all.trees 'i & b<=10' o > all_bs10.trees
```

- Gene trees to species tree with ASTRAL

```
astral -t 2 -i all_bs10.trees -o ASTRAL_all_bs10.tree 2> ASTRAL_all_bs10.log
```

Branch annotations options | Input trees | Output tree | Output log file

- Root the tree (using phyx)

```
pxrr -t ASTRAL_all_bs10.tree -g outgroup_sample_name > rooted.ASTRAL_all_bs10.tree
```

Gene trees approach

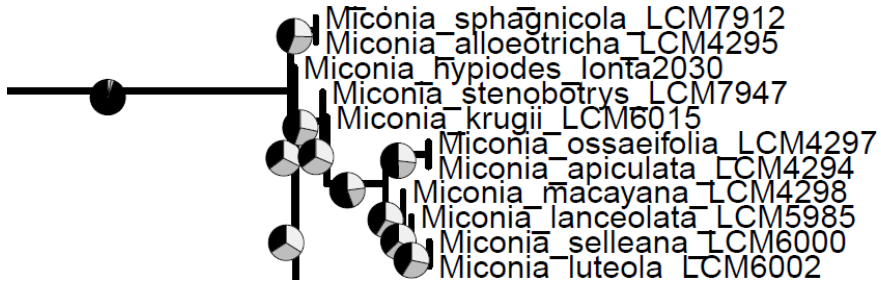
- Plot the tree using R custom script `plot_astral_tree.R`

```
tree_file <- c("rooted.ASTRAL_all_bs10.tree")
tiplabels <- read.table(file = "tips_rename.txt", col.names = c("OLD", "NEW"))


source(plot_astral_tree.R)

plot_astral(tree_file, rename = T, tiplabels = tiplabels, annotations = c("LPP", "QS"))
```

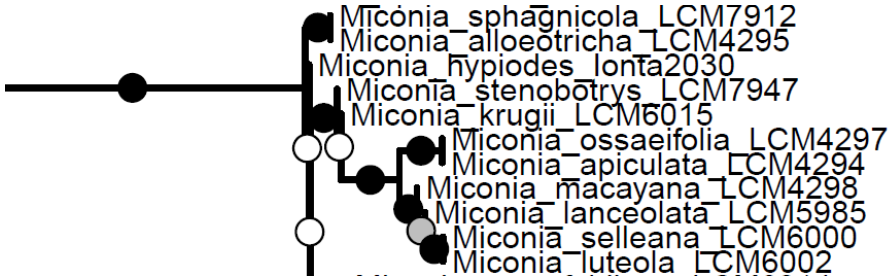
```
Sample1    New_name1
Sample2    New_name2
…
```

tips_rename.txt

rooted.ASTRAL_all_bs10.tree_QS.pdf



rooted.ASTRAL_all_bs10.tree_PP.pdf



LPP = Local Posterior Probabilities
QS = Quartet Support

Gene trees approach

- ASTRAL accuracy might be hindered by high levels of missing data (e.g. gene sequence recovered in a few samples only)

- **Other programs**:
- ASTRAL-Pro2 (ASTRAL for PaRalogs and Orthologs) (Zhang et al. 2020, Zhang and Mirarab 2022)
- ASTEROID (Accurate Species Tree Estimation RObust to Incomplete Data sampling) (Morel et al. 2022)

Miscellaneous

- RStudio as working station (R console + bash terminal)
- Reproducibility via Rmarkdown (= similar to lab notebook)
- GitHub (and link GitHub with RStudio)
- Presented simplified workflow, room to adapt to your own way of working

- Alignment viewers: AliView, Seaview, Mesquite, …
- Phylogenetic tree viewers: FigTree, Dendroscope, …

- Parallel processing via the “parallel” command (linux)
- Be careful with over-threading

Going further

- Time calibration (BEAST2, TreePL, LSD2 (IQTREE), RevBayes), see also:
  Sauquet H (2013) **A practical guide to molecular dating**. Comptes Rendus Palevol 12: 355–367.
  https://doi.org/10.1016/j.crpv.2013.07.003

- CAPTUS (alternative to HybPiper), excellent documentation and tutorial
  https://edgardomortiz.github.io/captus.docs/basics/index.html