Designing probes for Sequence capture

Mats Töpel
Department of Marine Sciences
mats.topel@marine.gu.se
@matstopel

Strategy

Identify regions in genomic data that are:

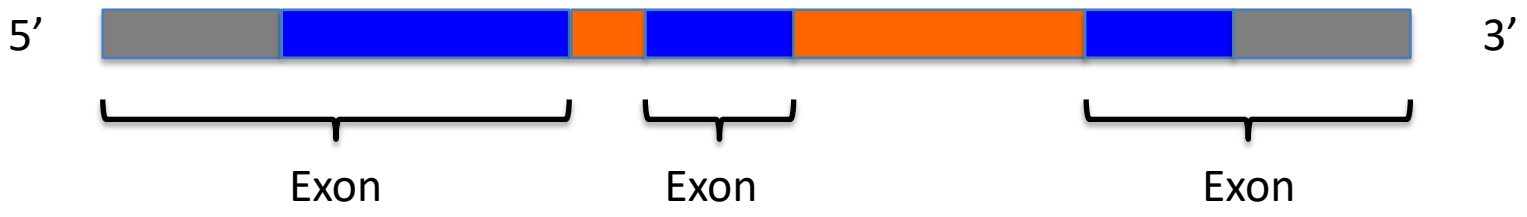
1. Conserved across a group (clade) of organisms.
2. “Uniqe” (single copy).
3. Flanked by variable regions that contain a phylogenetic signal.

The problem

Identify regions in genomic data that are:

1. Conserved across a group (clade) of organisms. – Suitable for probes to bind to.
2. “Uniqe” (single copy). – Assembly and analysis less complicated.
3. Flanked by variable regions that contain a phylogenetic signal.

Gene model



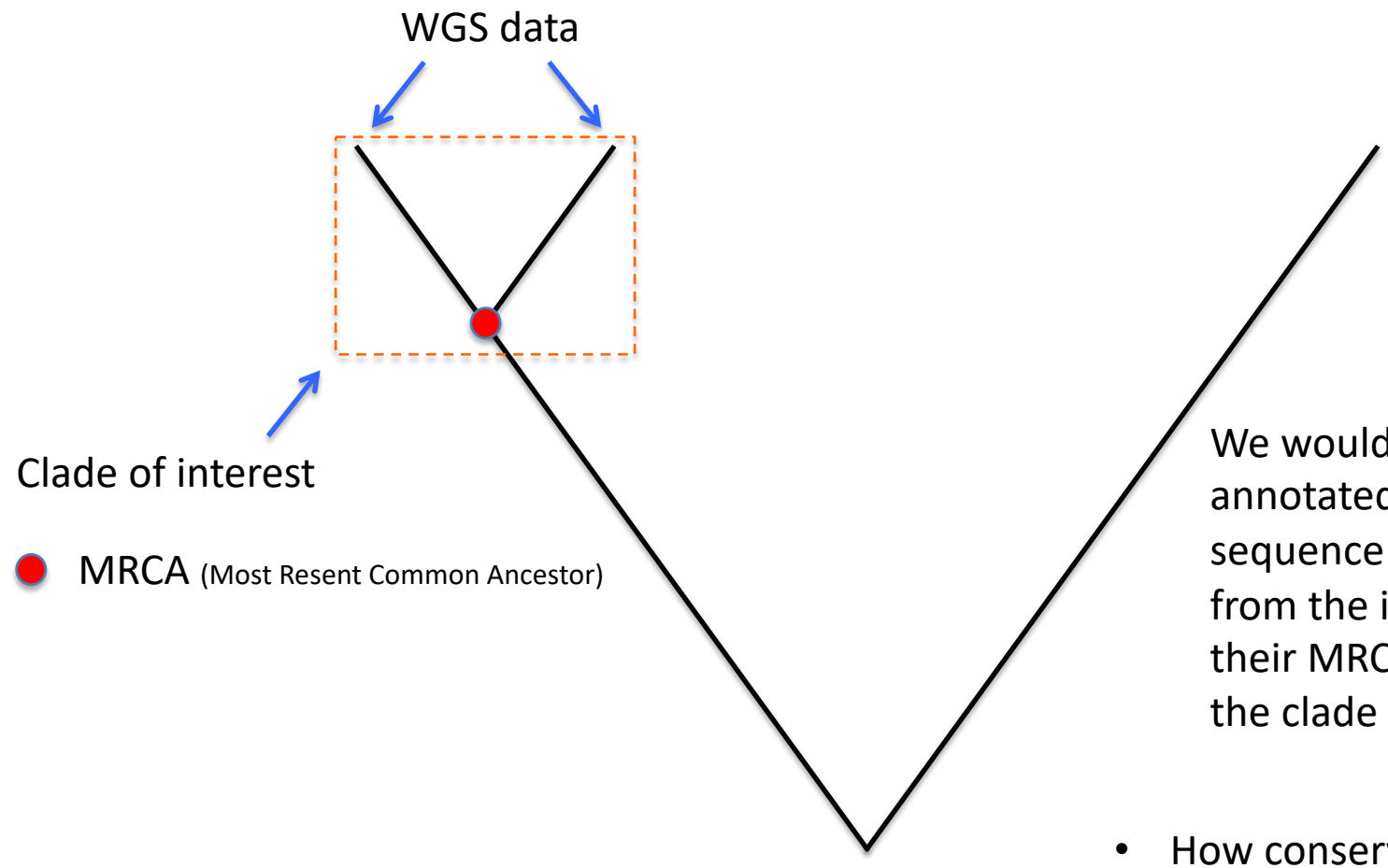
Untranslated region (UTR)

Coding region (CDS)

Intron

- The CDS’s are suitable as templates for probe sequences
- Introns (and exons) will contain phylogenetic signal

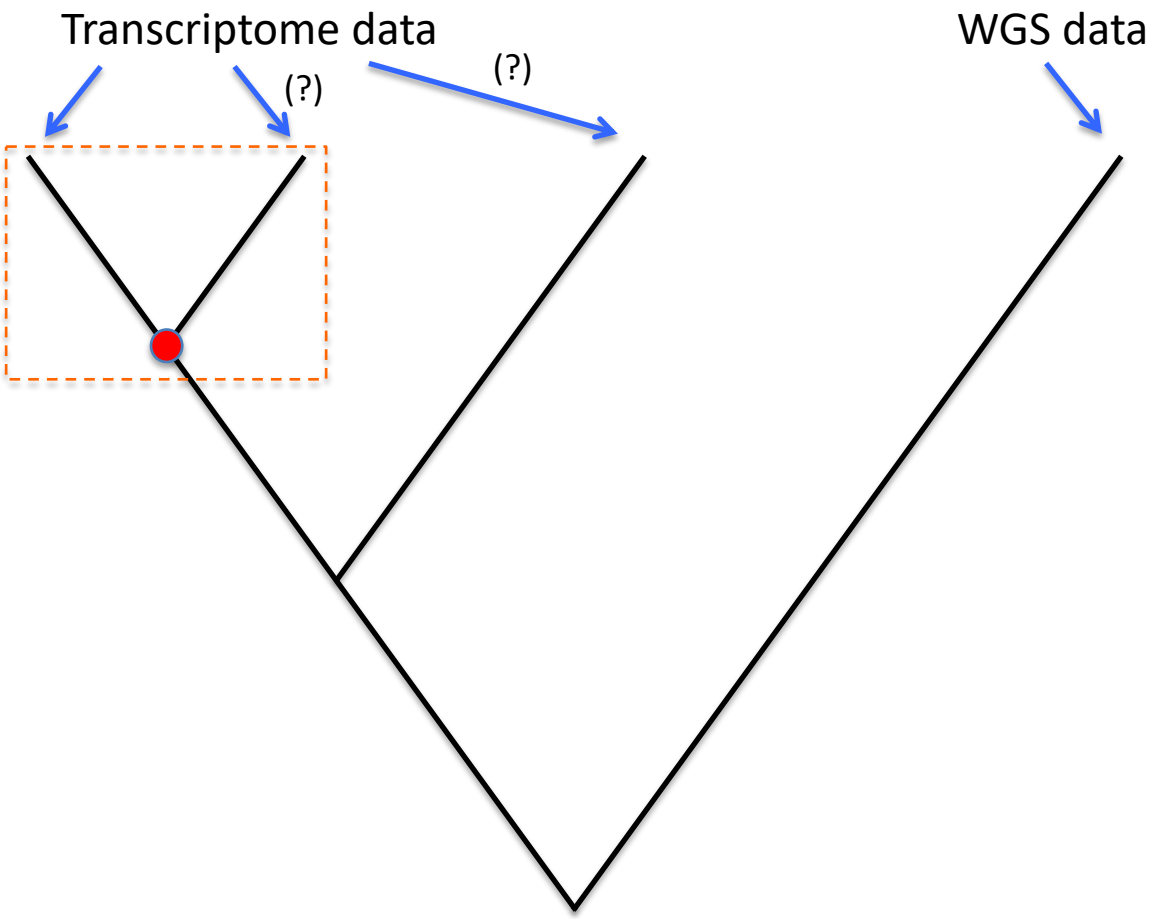
Best case scenario



We would like to have well annotated whole genome sequence (WGS) datasets from the ingroup, with their MRCA being that of the clade of interest.

- How conserved are the CDS's?
- What's the number of introns?
- How long are the introns?

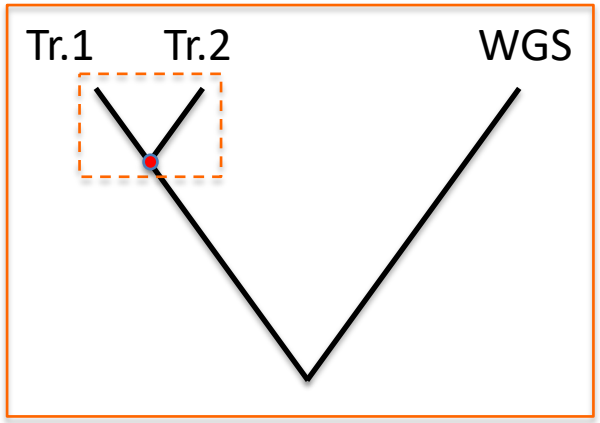
Reality



Most often we only have access to one or a few transcriptome datasets from the ingroup

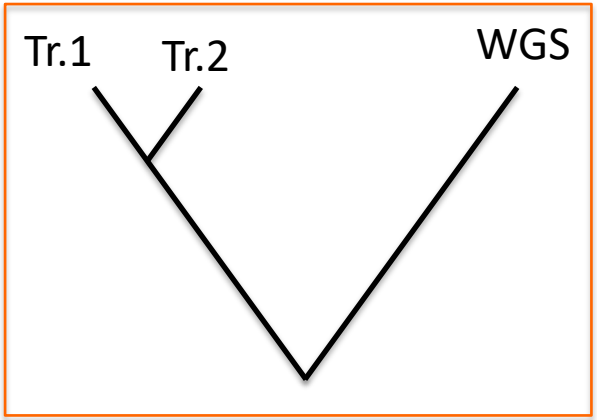
Basic workflow

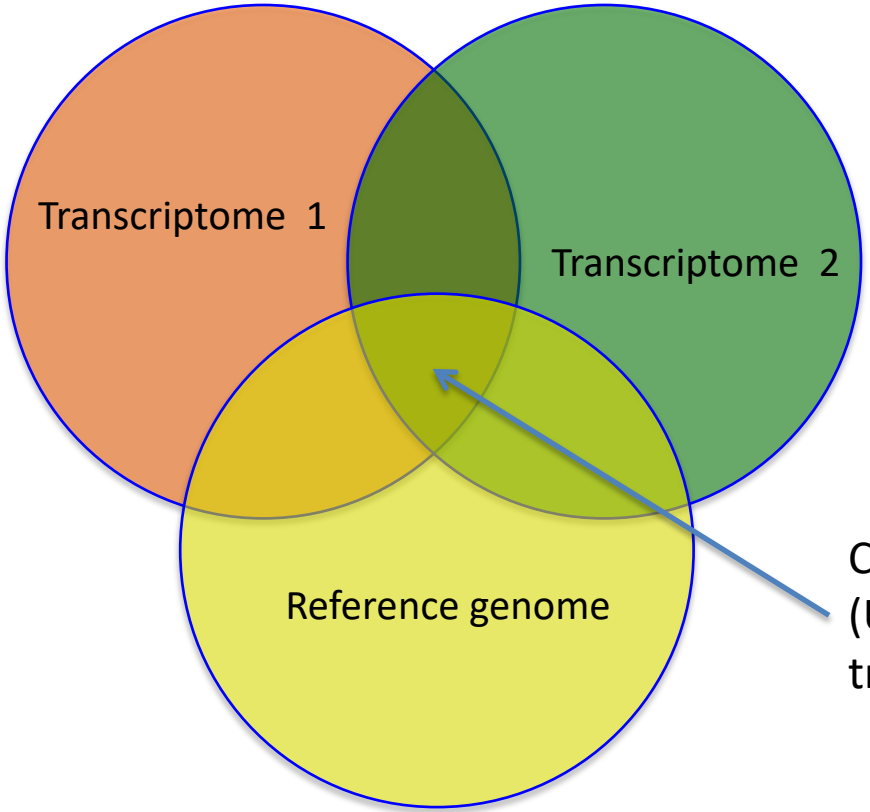
(One way of doing it)



1. Identify low copy gene sequences in transcriptome 1
2. Find the homologous sequences in transcriptome 2
3. [Optional] Extract gene structure data from a WGS dataset (number of exons, length of introns, copy number,…)

Basic workflow





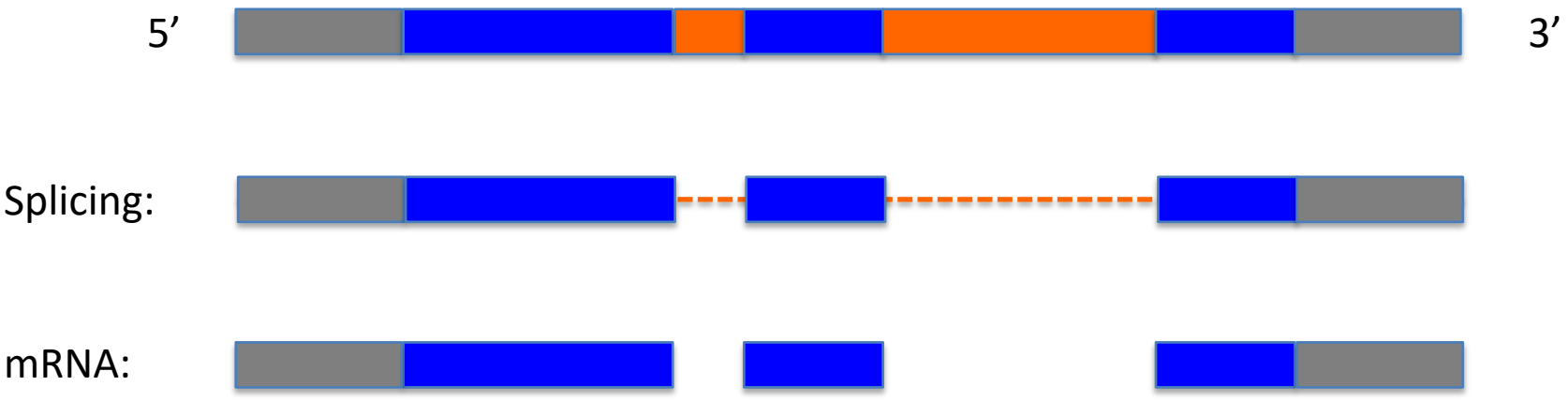
Capture probe template sequences
(Use the sequences from the “best” transcriptome dataset)

FAQ 1

- RNA-seq (transcripts) are often used for probe design
- DNA is then captured

Q: Can this approach cause difficulties?

Gene model

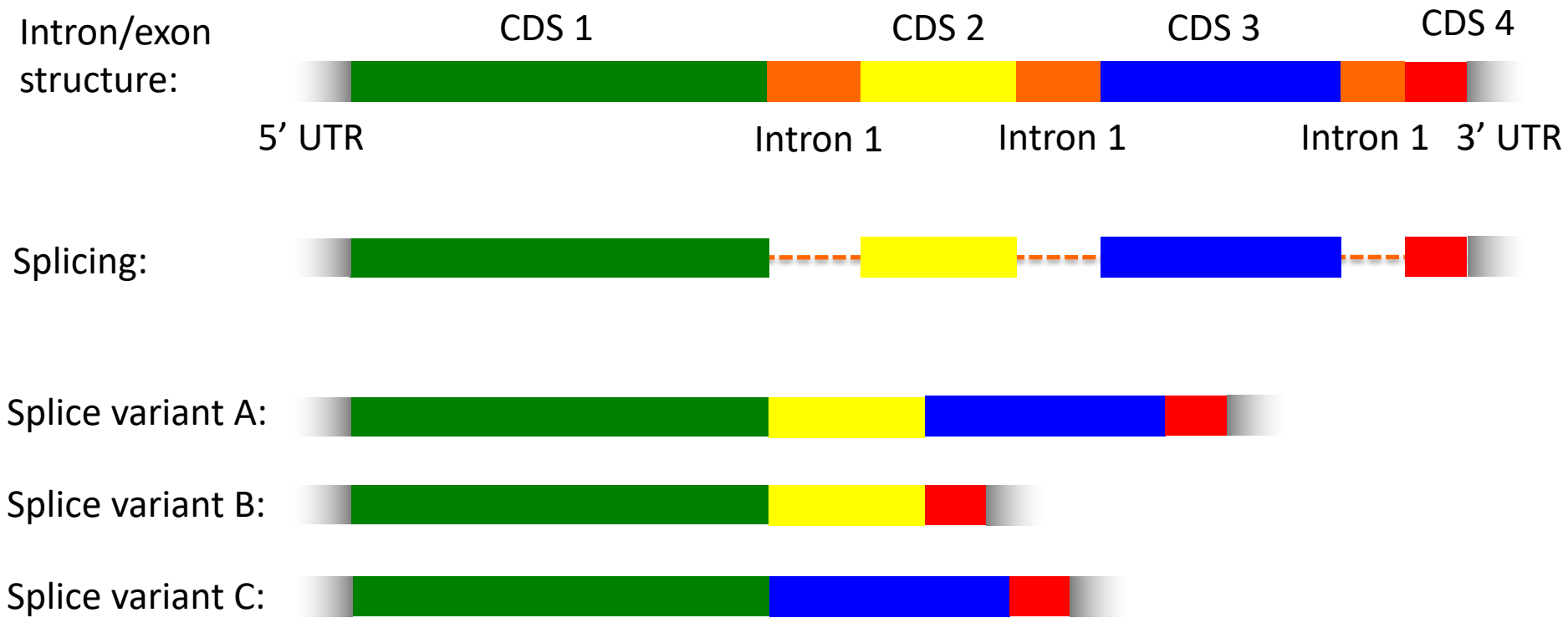


Untranslated region (UTR)

Coding region (CDS)

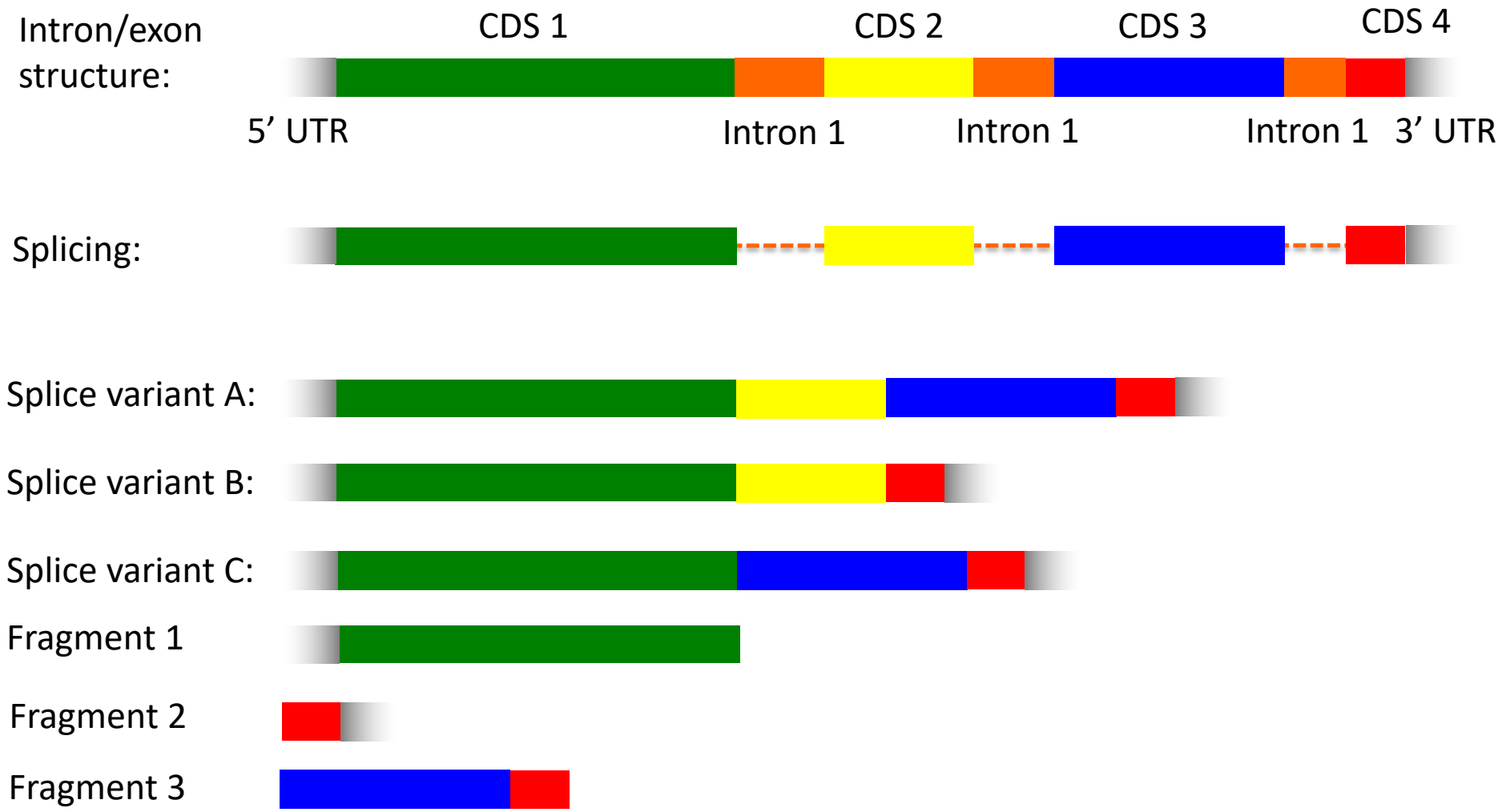
Intron

Alternative splicing



This phenomenon will greatly complicate the assembly problem.

Assembly problems



Transcriptome assembly

https://github.com/trinityrnaseq/trinityrnaseq/wiki

Trinity workflow

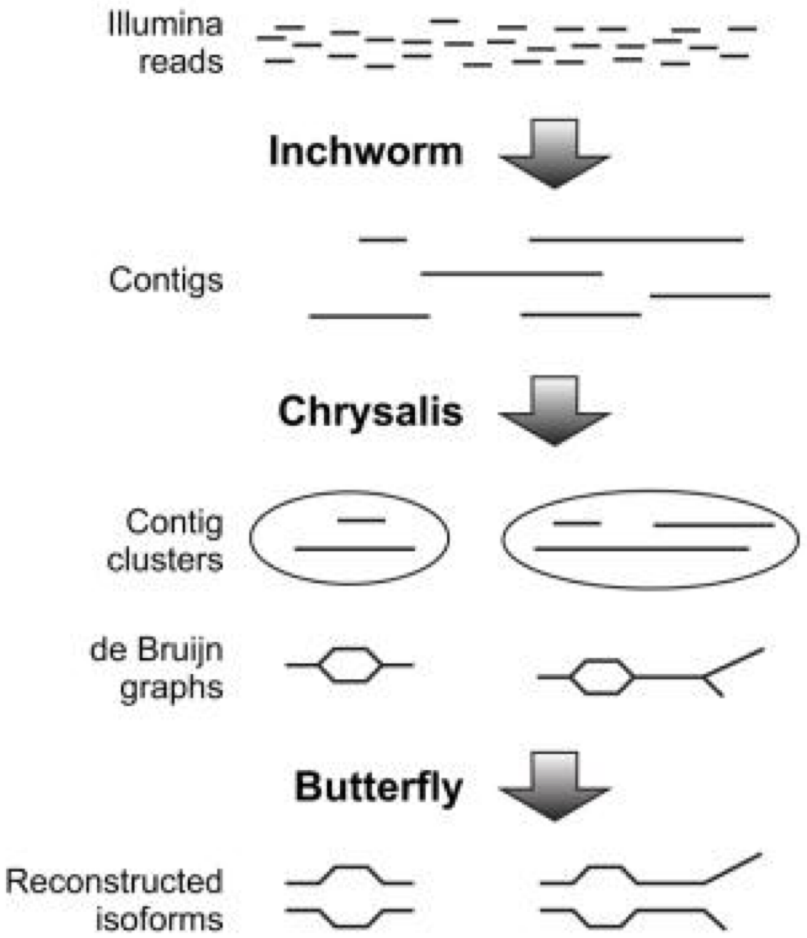
Inchworm assembles the data into unique transcripts, often generating full-length transcripts for the dominant isoform, but only reports unique portions of alternatively spliced transcripts.

Chrysalis clusters the contigs into clusters and constructs complete de Bruijn graphs for each cluster.

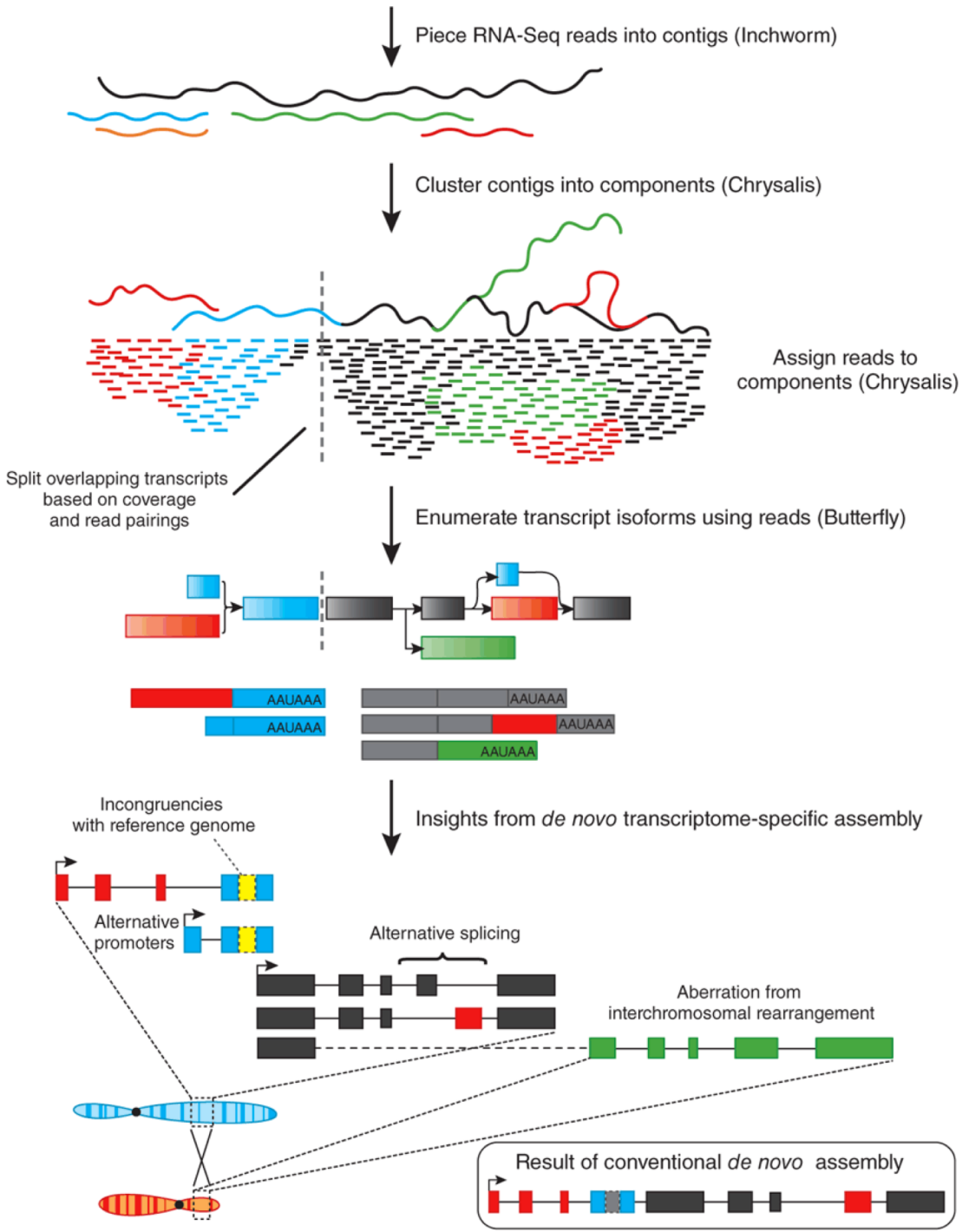
Butterfly processes the individual graphs in parallel, ultimately reporting full-length transcripts for alternatively spliced isoforms.

https://github.com/trinityrnaseq/trinityrnaseq/wiki

Trinity workflow



https://github.com/trinityrnaseq/trinityrnaseq/wiki


Piece RNA-Seq reads into contigs (Inchworm)
Cluster contigs into components (Chrysalis)
Assign reads to components (Chrysalis)
Split overlapping transcripts based on coverage and read pairings
Enumerate transcript isoforms using reads (Butterfly)
AAUAAA
AAUAAA
AAUAAA
AAUAAA
AAUAAA
Insights from de novo transcriptome-specific assembly
Incongruencies with reference genome
Alternative promoters
Alternative splicing
Aberration from interchromosomal rearrangement
Result of conventional de novo assembly

Trinity output

>TRINITY_DN1000|c115_g5_i1 len=247 path=[31015:0-148 23018:149-246]
AATCTTTTTTGGTATTGGCAGTACTGTGCTCTGGGTAGTGATTAGGGCAAAAGAAGACAC
ACAATAAAGAACCAGGTGTTAGACGTCAGCAAGTCAAGGCCTTGGTTCTCAGCAGACAGA
AGACAGCCCTTCTCAATCCTCATCCCTTCCCTGAACAGACATGTCTTCTGCAAGCTTCTC
CAAGTCAGTTGTTCACAGGAACATCATCAGAATAAATTTGAAATTATGATTAGTATCTGA
TAAAGCA

TRINITY_DN1000|c115: Read cluster ("Gene family")

g5: "Gene"

i1: Isoform

https://github.com/trinityrnaseq/trinityrnaseq/wiki/Output-of-Trinity-Assembly

Trinity is memory intensive

Requires ~1G of RAM per 1M reads

- One MiSeq run generate ~170M reads => 170G RAM required
- HiSeq 2500 can generate 4’000G paired end reads (but samples are usually pooled and each dataset hence smaller).

You can use the digital normalisation function of trinity if you have very large datasets, to overcome RAM limitations.

https://github.com/trinityrnaseq/trinityrnaseq/wiki

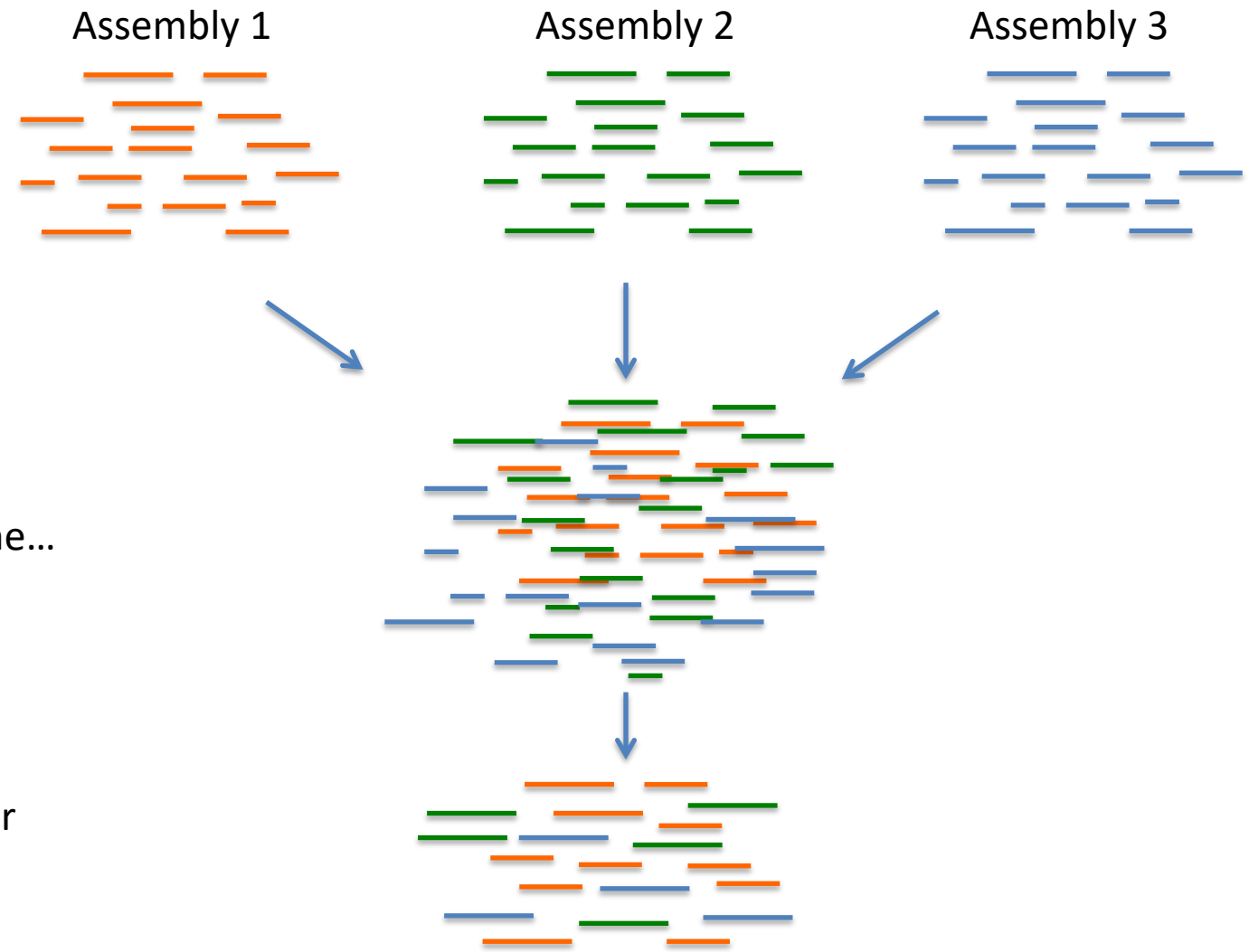
FAQ:

- Q: “I have RNA-seq data from several individuals of the same species. Should I assemble these reads as one dataset?”

FAQ:

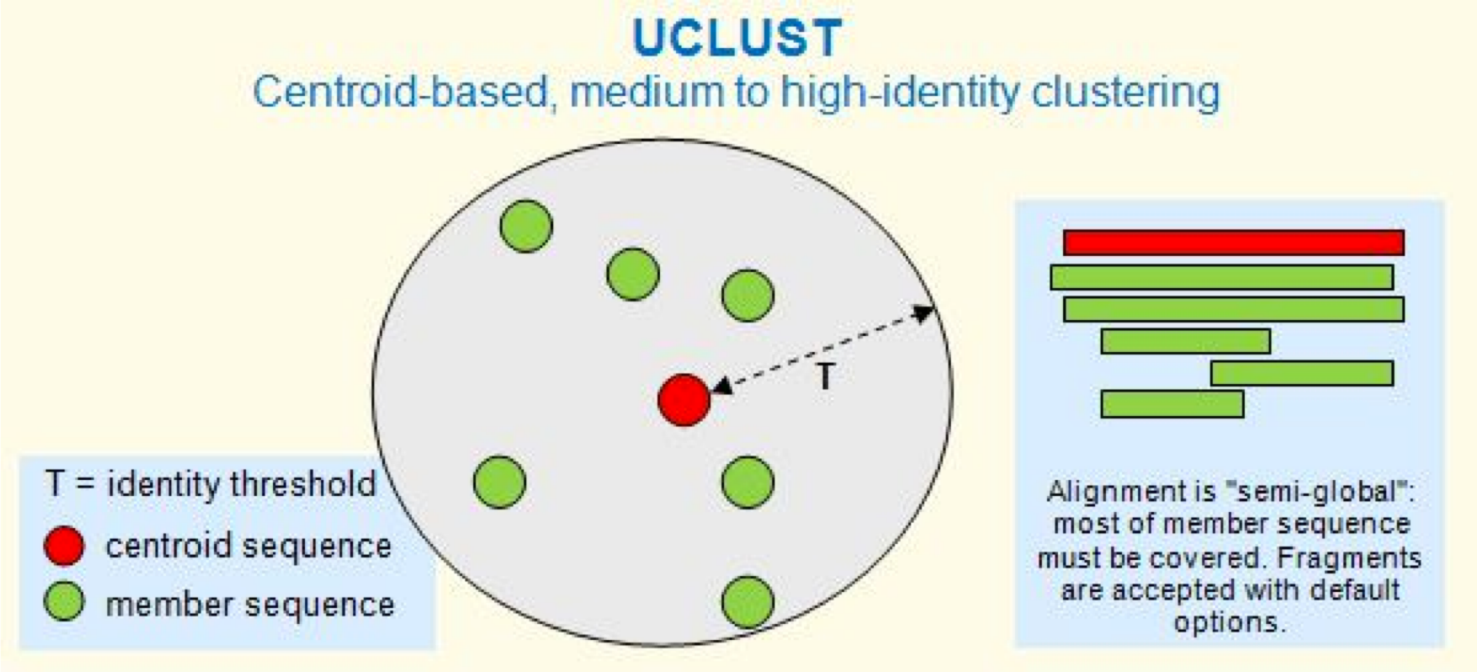
- Q: “I have RNA-seq data from several individuals of the same species. Should I assemble these reads as one dataset?”
- A: “No. It is most likely better to assemble the individuals separately and then identify unique sequences afterwards using sequence clustering.

Combining RNA-seq datasets



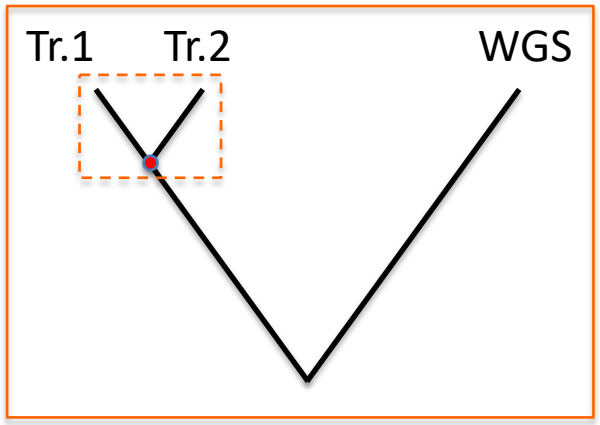
http://weizhongli-lab.org/cd-hit/
http://drive5.com/usearch/manual/uclust_algo.html

Uclust



http://drive5.com/usearch/manual/uclust_algo.html

Basic workflow



1. Identify low copy gene sequences in transcriptome 1
2. Find the homologous sequences in transcriptome 2
3. [Optional] Extract gene structure data from a WGS dataset (number of exons, length of introns, copy number,…)

Identify low copy gene sequences in transcriptome 1

- Using different clustering methods (UCLUST, CD-HIT)
  - Fast
  - Can be a bit of a black box – Read the manual.
  - Order of sequences in input file can effect the result

- Sequence comparison using BLAST
  - Computationally intensive
  - Less of a black box – but still, read the manual
  - Many different output formats. Custom formats possible.
  - Allows for comparison on nucleotide or amino acid sequence level (useful when comparing transcriptomes to gene models of WGS data in later steps)

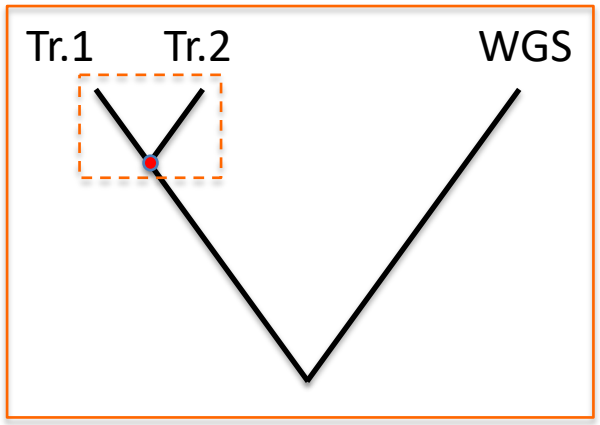
Identify low copy gene sequences in transcriptome 1

Possible workflow using BLAST
(Cluster the reads from your transcriptome assemblies and remove redundant and fragmented sequences first)

1. Blast the whole transcriptome dataset to itself.
2. Disregard the best match as it will be the query sequence
3. Analyse the second best match as it will be the closest paralog (good to decent score) or a random sequence from the dataset (indicated by a poor blast score). Look for the latter.
4. Extract the sequences using e.g “fp.py”

https://github.com/mtop/ngs/fp.py
https://github.com/mtop/sequence_capture

Basic workflow

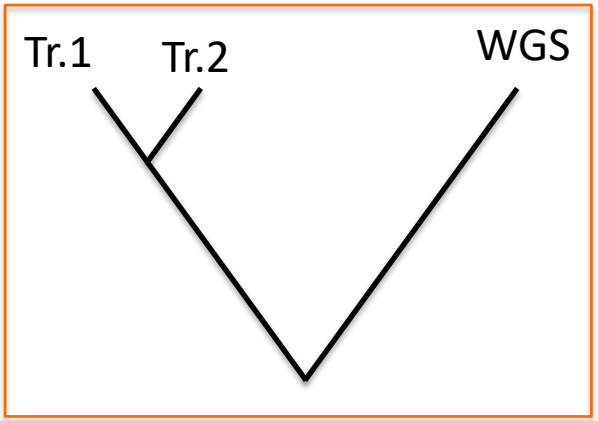


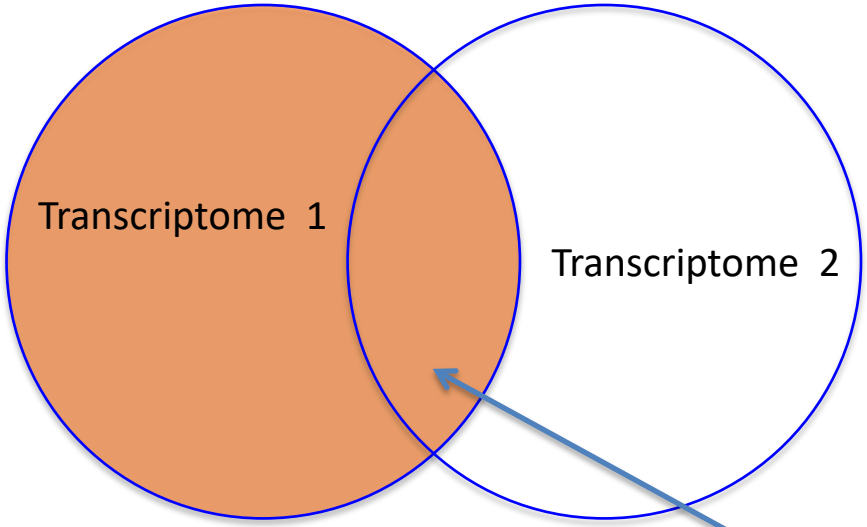
1. Identify low copy gene sequences in transcriptome 1
2. Find the homologous sequences in transcriptome 2
3. [Optional] Extract gene structure data from a WGS dataset (number of exons, length of introns, copy number,…)

Find the homologous sequences in transcriptome 2

1. Blast the sequences identified in the previous step to the sequences in transcriptome 2.
2. Identify the query sequences that have:
   - Good enough Blast score to transcriptome 2
   - High enough identity…
   - … over a long enough region
   - And what ever other criteria you like to use and that blast can report (e.g. query or target length).

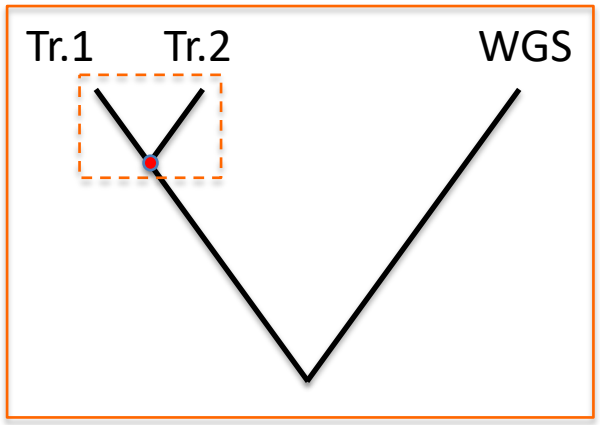
Basic workflow





We have now identified this part of the dataset.

Basic workflow



1. Identify low copy gene sequences in transcriptome 1
2. Find the homologous sequences in transcriptome 2
3. [Optional] Extract gene structure data from a WGS dataset (number of exons, length of introns, copy number,…)

Sources for WGS data

- http://www.phytozome.net/
- http://jgi.doe.gov/
- http://www.ensembl.org/
- ftp://ftp.ncbi.nlm.nih.gov/genomes/

Extract gene structure data from a WGS dataset

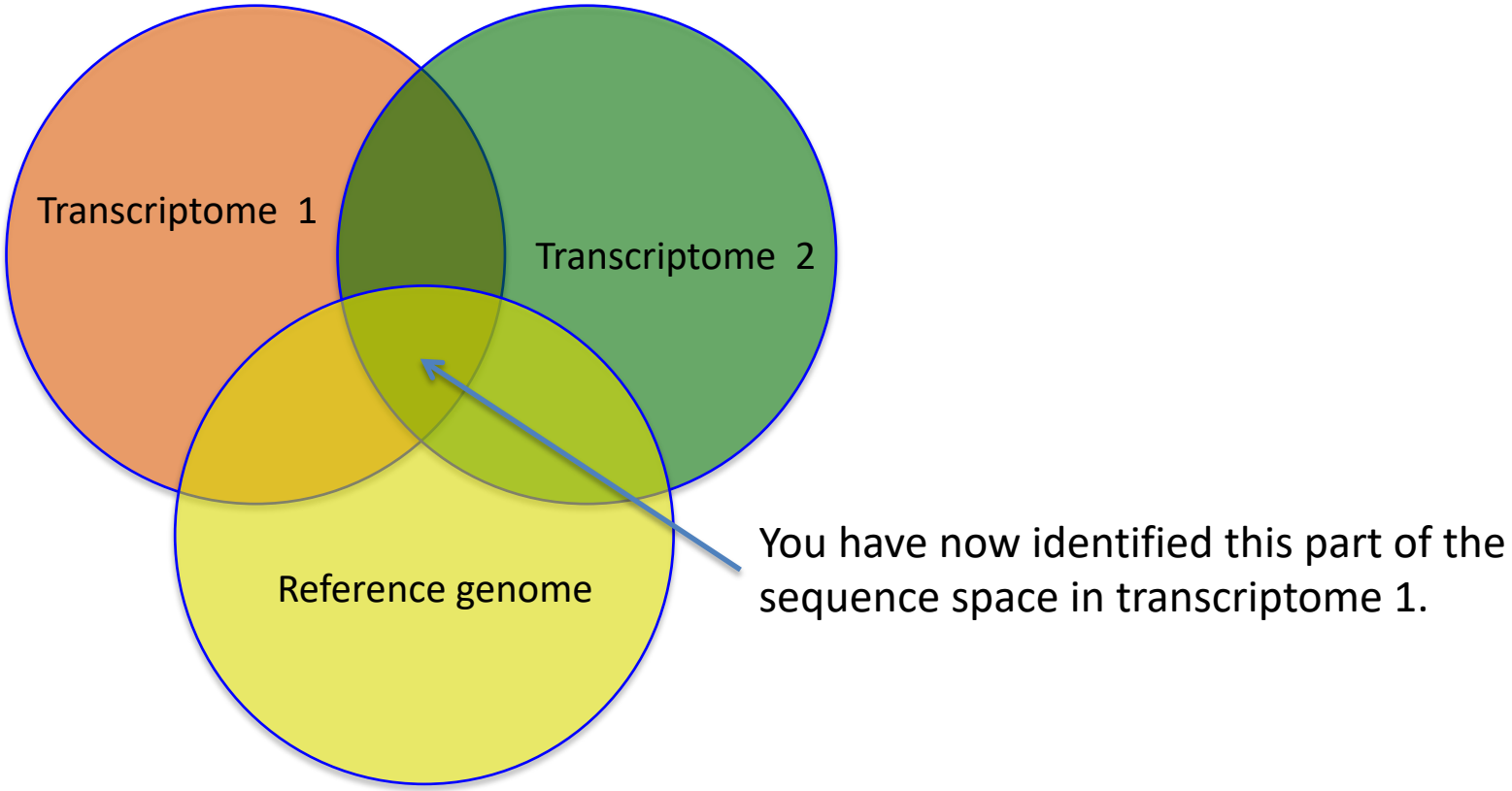
This type of information is stored in gff format, and distributed with an annotated WGS dataset. Can be extracted using e.g “parseGff3.py”.

```
##gff-version  3
# Sequence Data: seqnum=1;seqlen=3612930;seqhdr="unitig_0"
# Model Data: version=Prodigal.v2.6.2;run_type=Single;model="Ab initio";…
unitig_0    Prodigal_v2.6.2 CDS 78  323 11.3    -   0   ID=1_1;partial=00;…
unitig_0    Prodigal_v2.6.2 CDS 400 5154    833.8   -   0   ID=1_2;partial=00;…
unitig_0    Prodigal_v2.6.2 CDS 5266    5880    69.0    -   0   ID=1_3;partial=00;…
unitig_0    Prodigal_v2.6.2 CDS 5877    7313    208.9   -   0   ID=1_4;partial=00;…
unitig_0    Prodigal_v2.6.2 CDS 7310    9439    261.7   -   0   ID=1_5;partial=00;…
unitig_0    Prodigal_v2.6.2 CDS 9698    10855   112.1   -   0   ID=1_6;partial=00;…
```

https://github.com/mtop/sequence_capture/parseGff3.py

Extract gene structure data from a WGS dataset

1. Use one of the Blast algorithms (blastx, tblastn) for the last homology search between Tr. 1 and the reference sequences .



At this stage you probably want to start over and refine your search

- How many sequences did you end up with?
- Is that enough for what you want to do?
- Is one of the search steps (sequence length, identity, number of exons) more important for a successful result?

Additional approaches to solve the probe design problem

MarkerMiner

Journal List › Appl Plant Sci › v.3(4); 2015 Apr › PMC4406834

This Article | Current Issue | About *APPS* | About the BSA

Applications in Plant Sciences

Appl Plant Sci. 2015 Apr; 3(4): apps.1400115. PMCID: PMC4406834
Published online 2015 Apr 6. doi: 10.3732/apps.1400115

MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes[1]

Srikar Chamala,[2,12] Nicolás García,[2,3,4,*] Grant T. Godden,[2,3,5,*] Vivek Krishnakumar,[6] Ingrid E. Jordon-Thaden,[7,8] Riet De Smet,[9,10] W. Brad Barbazuk,[2,11] Douglas E. Soltis,[2,3,11] and Pamela S. Soltis[3,11]

Author information ▶ Article notes ▶ Copyright and License information ▶

Abstract

Go to:

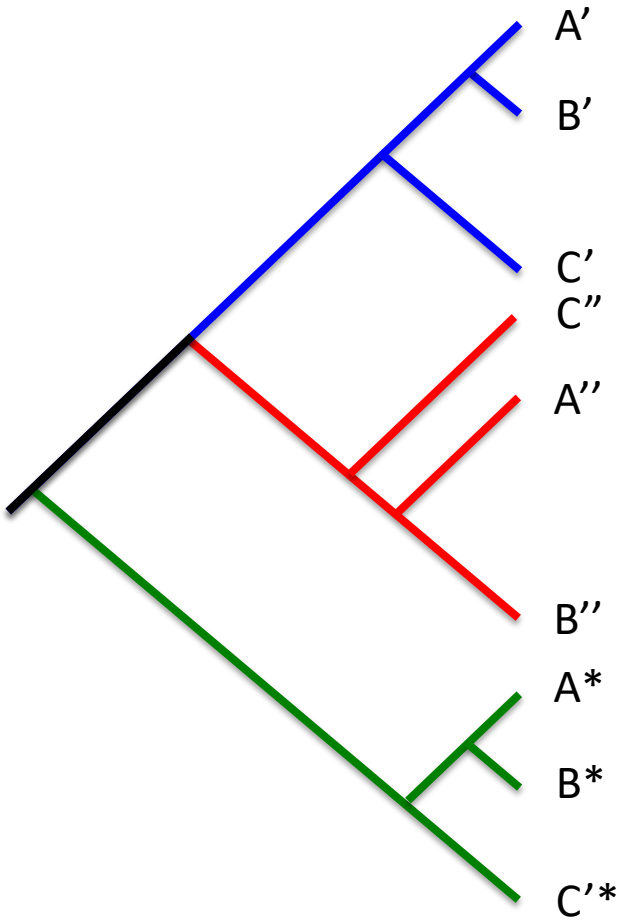
Premise of the study:

Targeted sequencing using next-generation sequencing (NGS) platforms offers enormous potential for plant systematics by enabling economical acquisition of multilocus data sets that can resolve difficult phylogenetic problems. However, because discovery of single-copy nuclear (SCN) loci from NGS data requires both bioinformatics skills and access to high-performance computing resources, the application of NGS data has been limited.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4406834/

Multi copy genes

Gene tree



Species tree