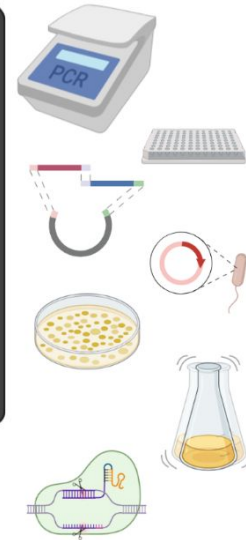
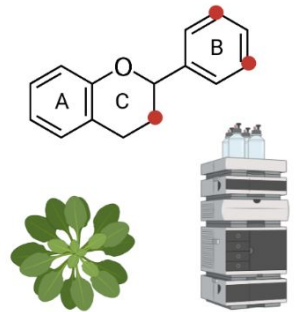




Plant Biotechnology and Bioinformatics

[illegible]

Prof. Dr. Boas Pucker (Plant Biotechnology and Bioinformatics)

Availability of slides

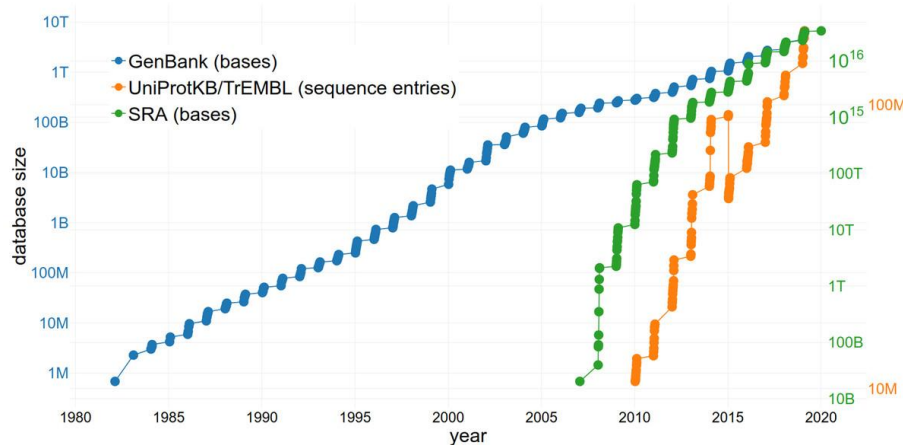
- All materials are freely available (CC BY) - after the lectures:
 - StudIP: [Lecture: Grundlagen der Biochemie und Bioinformatik der Pflanzen \(Bio-MB 09\)](#)
 - GitHub: <https://github.com/bpucker/teaching>
- Questions: Feel free to ask at any time
- Feedback, comments, or questions: [b.pucker\[a\]tu-bs.de](mailto:b.pucker@tu-bs.de)



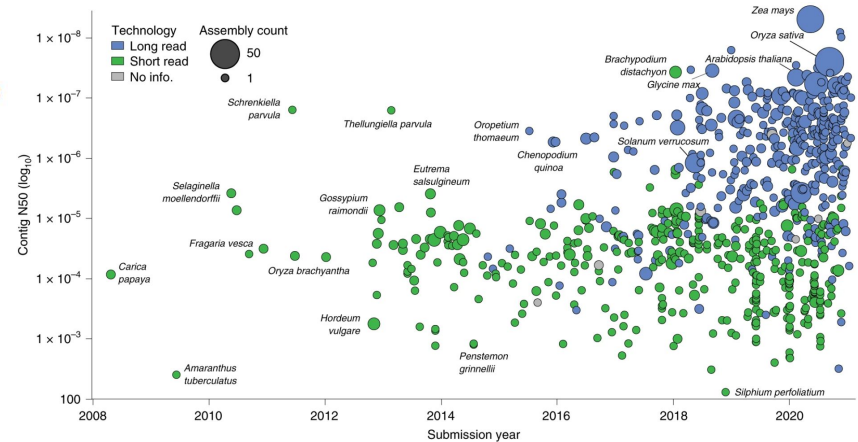
My figures and content can be re-used in accordance with CC-BY 4.0, but this might not apply to all images/logos.

Progress in plant genomics

- Number of reference quality genome sequences is increasing
- Remaining challenge is separation of haplotypes
- Number of re-sequencing projects is increasing
- Focus is shifting towards investigation of intraspecies variation



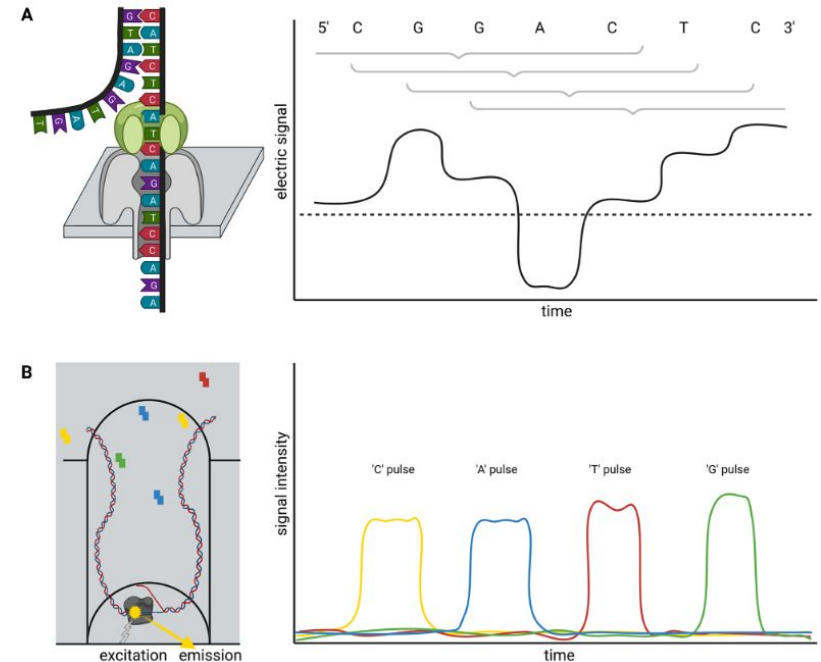
Sielemann et al., 2020: 10.7717/peerj.9954



Marks et al., 2021: 10.1038/s41477-021-01031-8

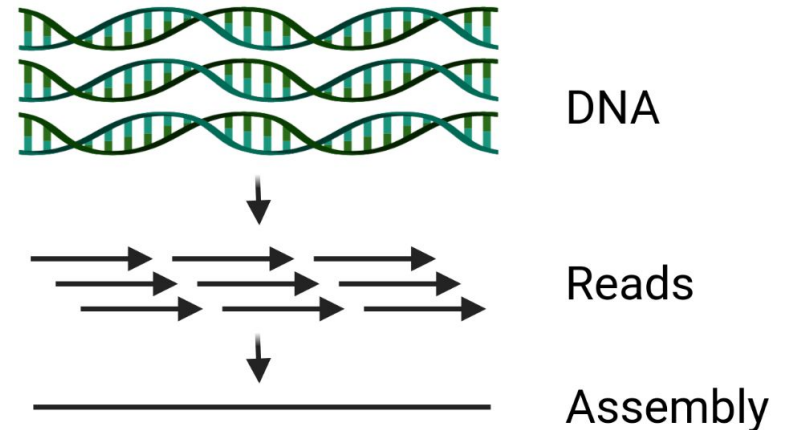
Long read sequencing

- Plant genomics is based on long read sequencing technologies:
 - ONT: Oxford Nanopore Technologies
 - PacBio: Pacific Biosciences
- Read length is >>20kb; up to million bases
- Read quality is continuously increasing (>99% accuracy)



Genome sequence assembly

- Reads are shorter than the chromosome
- even long reads
- Multiple copies of the genome (DNA exist that can be subjected to sequencing
- Assembly = putting sequence pieces together (finding the common string of all substrings)
- Genome = DNA in a cell
- Genome sequence = representation of the DNA in a cell; stored in FASTA files



```
>TRINITY_DN100016_c0_g1_i1
MPKKSSNIQNNQNNRIKRGPRPKKKKYVQQIDSSDEDILSVRASTRPRIISIRRNEIPMRPEIHI
>TRINITY_DN100019_c0_g1_i1
MPPKAADKKPAAPATASKAPEKKDAGKKTAATGDKKKRTKARKETYSSYIYKVLKQVHPDGTGISN
>TRINITY_DN10001_c0_g1_i2
MAKVGNPIVIDTSGSVNEPESSEKNIEVSSSSTQAPESNTTELLVNEKKAFSLATPAVRRVAREH
>TRINITY_DN100025_c0_g1_i1
MVENQDGCFFKPGWKEFVRSNDLEGGDFLVFNLDKISYQVVFIDGTCCPKDLCPFISMNPIFQHLR
>TRINITY_DN10002_c0_g1_i1
MSDELNQVFQRCREQKRPVFVAFVTAGYPDSEETVDILLGLEAGGADIIELGIPFTDPMVDGKTIQD
>TRINITY_DN100061_c0_g1_i1
MQQVAKLKAIIITKNTNENSPVENSSSSTSATSSINNSLHGDLRSRVFDNMELESNVSNSSISSNI
```


[illegible]

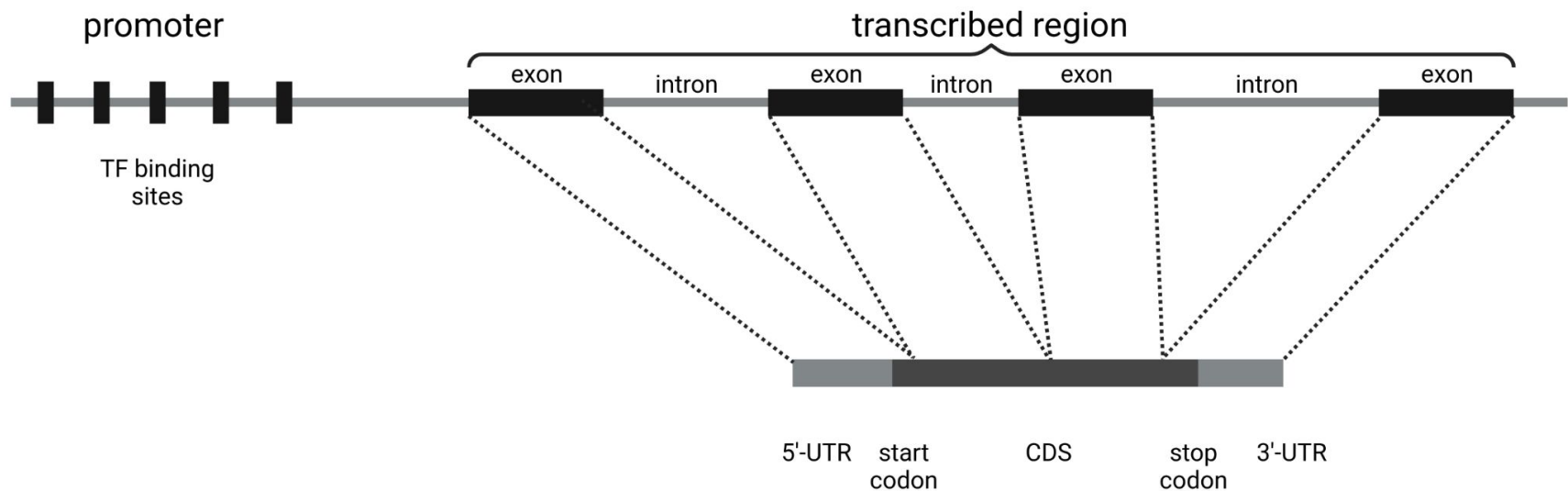
Plant gene structure

CDS = (Protein) Coding Sequence

ORF = Open Reading Frame

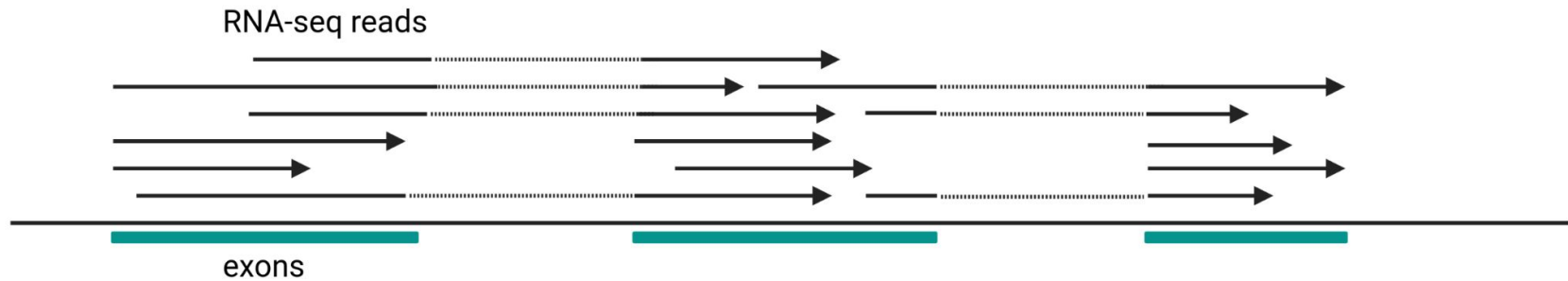
UTR = UnTranslated Region

TF = Transcription Factor



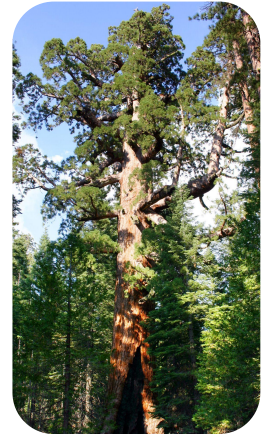
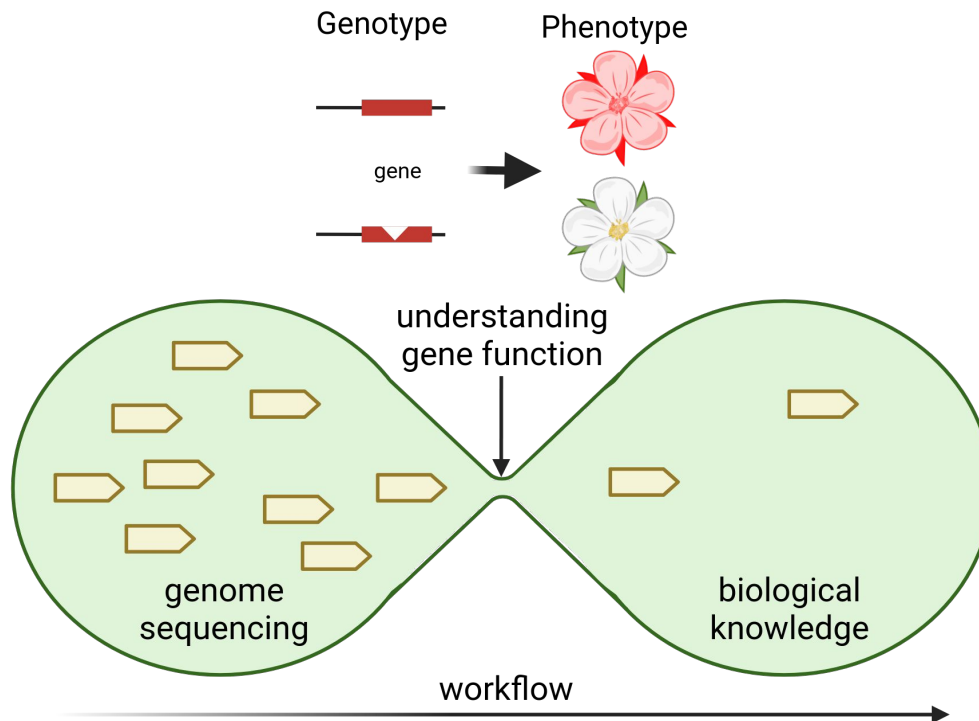
Gene prediction (structural annotation)

- Aligned RNA-seq (cDNA-derived) reads indicate exon positions
- Splitting of reads indicates intron positions
- CDS can be identified as ORF within the covered regions



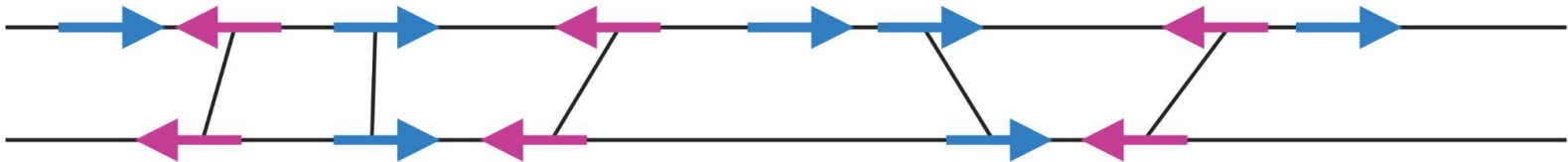
Gene function prediction (functional annotation)

- Thousands of genes per plant species; only a minority of *Arabidopsis thaliana* genes studied
- Not all species are genetically accessible



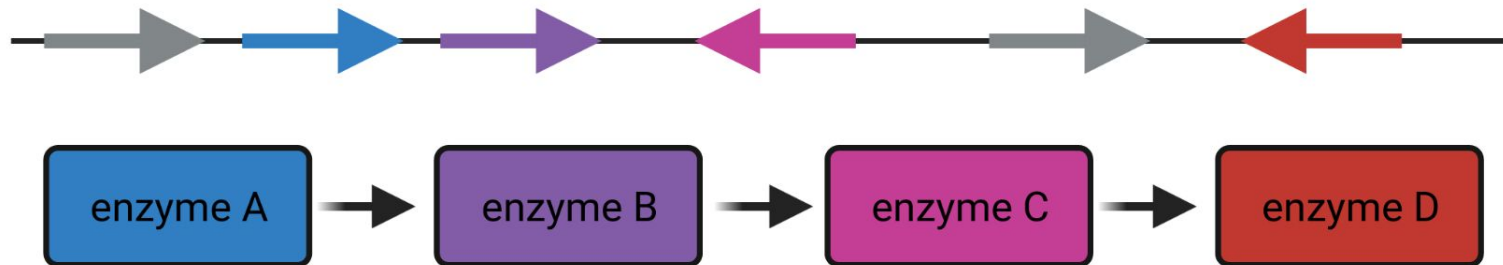
Synteny

- Synteny = same order of genes in different species
- Shared ancestry is responsible for synteny
- Changes during evolutionary times reduce synteny



Biosynthetic gene cluster

- Genes involved in one biosynthetic pathway
- Genes located in the same gene cluster
- Genes belong to different gene families
- Evolutionary benefits of clustering?



Biosynthetic gene cluster - example 1

- Thalianol gene cluster

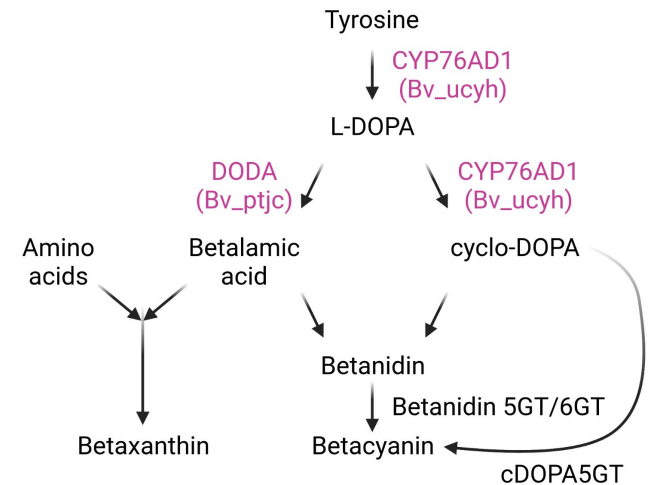


- Marneral gene cluster



Biosynthetic gene cluster - example 2

- Betalain biosynthesis gene cluster in sugar beet
- Evolution of betalain and anthocyanin biosynthesis is a model system for evolutionary questions



Brockington et al., 2015: 10.1111/nph.13441
Timoneda et al., 2019: 10.1111/nph.15980

Biosynthetic gene clusters vs. gene arrays

- Biosynthetic gene clusters are characterized by genes encoding enzymes with different functions
- Gene arrays comprise copies of the same gene (result of tandem duplications)

Biosynthetic gene cluster:



Gene array:



Gene array examples

- Gene arrays are frequent, but less interesting
- CHS gene array in sugar beet (*Beta vulgaris*)
- FLS gene array in *Arabidopsis thaliana*

Beta vulgaris:



Arabidopsis thaliana:

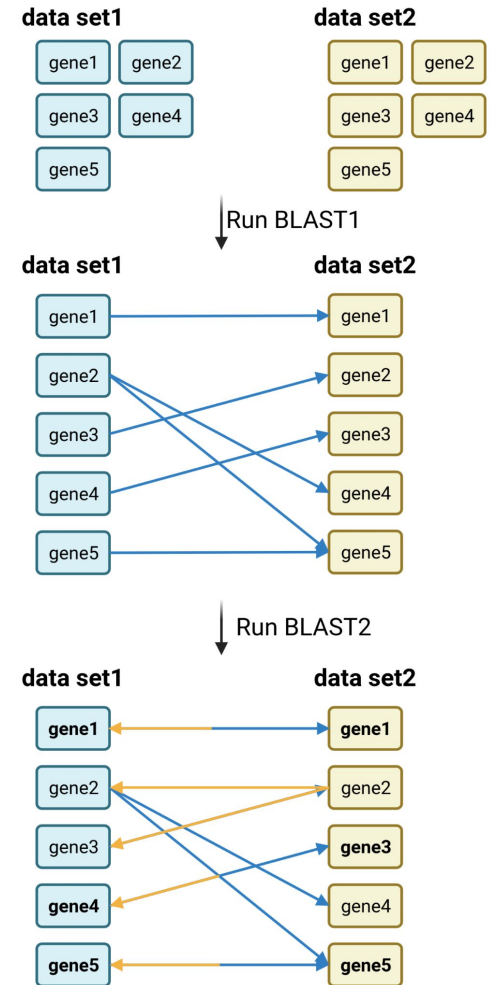


Identification of biosynthetic gene clusters

- Neighbouring genes must have different functions
 - Assign functional annotation (InterProScan, Mercator, RBHs)
- Neighbouring genes must be involved in the same pathway
 - Assign pathway to genes (KEGG, GO, MetaCyc)
- Co-expression of clustered genes is additional evidence
 - Analyse gene expression across numerous samples
- Conservation across some species can be additional evidence
 - Synteny analysis (JCVI)

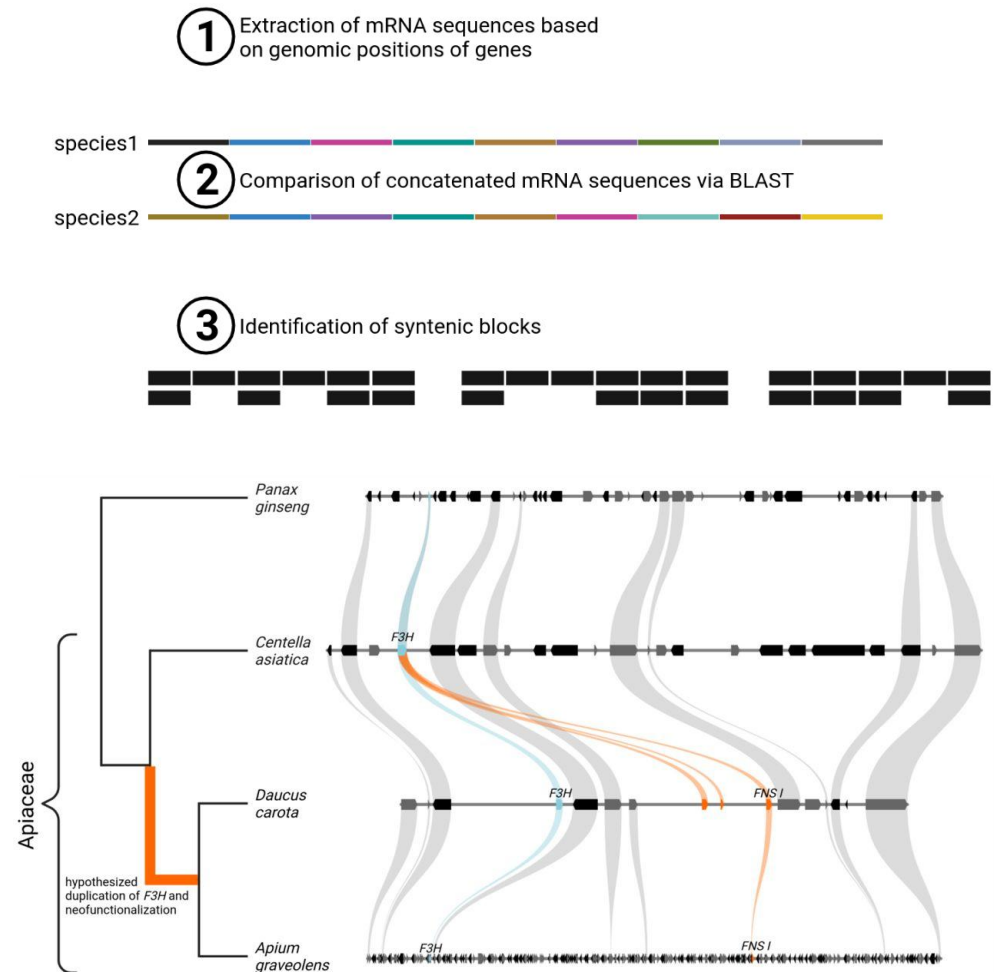
Reciprocal Best BLAST hits (RBHs)

- Run BLAST of data set 1 against data set 2
- Run BLAST in opposite direction
- Check for bidirectional (reciprocal) hits
- Reciprocal hits add reliability to the pairing
- Functional annotation can be transferred from data set 1 to data set 2
- Computationally efficient, but not perfect method



JCVI (tool for synteny analysis)

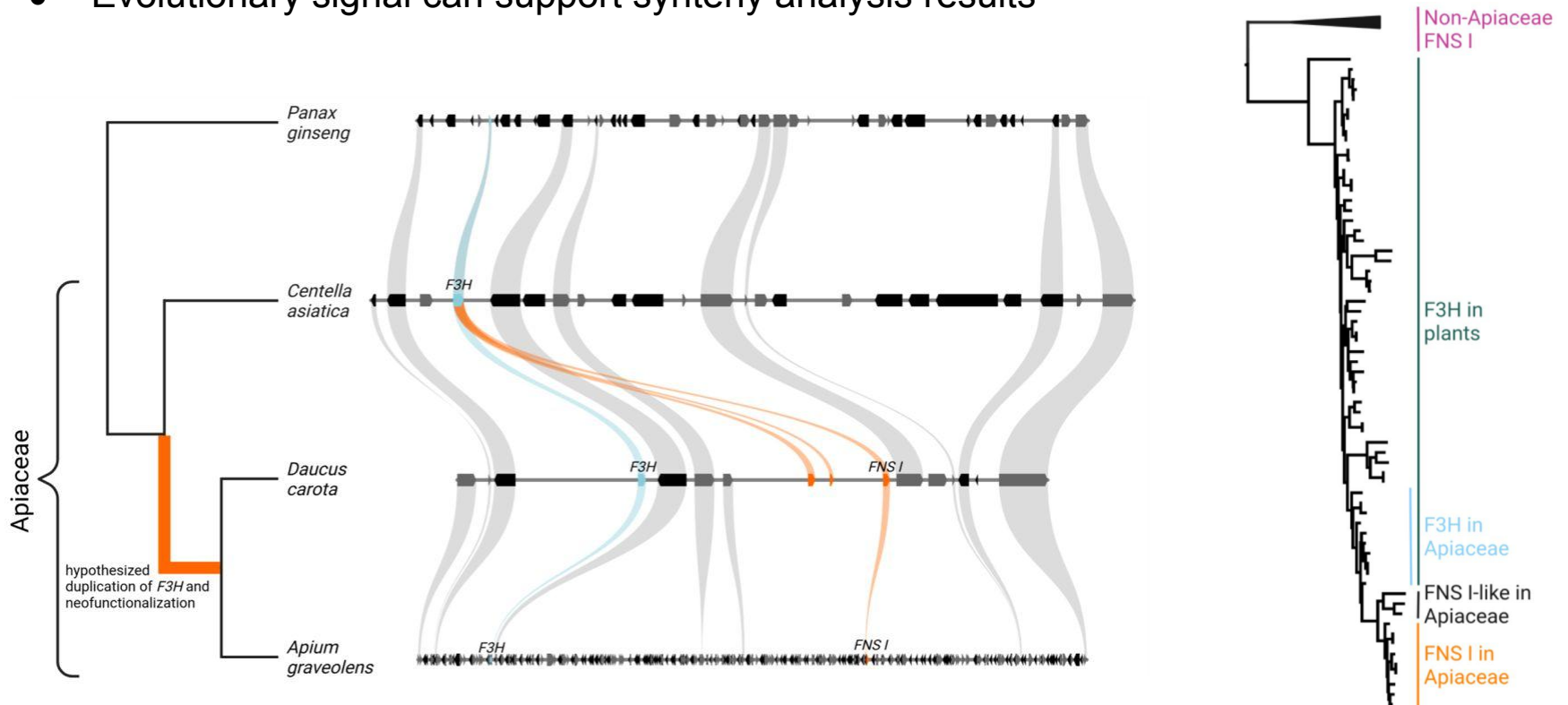
- Required input: FASTA file with mRNA sequences and GFF3 file with gene positions
- BLAST of concatenated mRNA sequences is used to connect gene across species
- Identification of syntenic gene blocks
- Extension of syntenic blocks



Pucker & Iorizzo, 2022: 10.1101/2022.02.16.480750

Ortholog assignment confirmation

- Phylogenetic tree for ortholog assignment confirmation
- Evolutionary signal can support synteny analysis results



Pucker & Iorizzo, 2022: 10.1101/2022.02.16.480750

Chromatin structure influences gene expression

- Chromatin structure can determine gene expression
- Shared chromatin structure in gene cluster
- Shared regulation of all genes in cluster
- Genes acting in the same pathway need co-expression

Evolutionary benefits of gene clusters

- Biosynthetic gene clusters are often evolutionary young pathways
- Many biosynthetic gene clusters lead to pathogene resistances
- Clustered genes ensure joint inheritance
- Clustering might prevent toxic intermediates
- Clustering could support shared regulation

Biosynthetic gene cluster - examples

S.no	Species	Metabolite	Compound class	Chromosome (core genes)	Function	Tissue of expression	References
1	<i>Arabidopsis thaliana</i>	Thalianol	Triterpenes	5	Unknown physiological function, unregulated expression of cluster genes leads to dwarfing of plant. Modulate root microbiome content	Roots	Field and Osbourn, 2008; Field et al., 2011; Chen et al., 2019a
2	<i>A. thaliana</i>	Marmeral	Triterpenes	5	Unknown physiological function, unregulated expression of cluster genes leads to dwarfing of plant	Roots	Field and Osbourn, 2008; Field et al., 2011
3	<i>Avena strigosa</i>	Avenacins	Triterpenes	1	Defense against pathogens	Roots	Li et al., 2021a
4	<i>Cucumis sativus</i>	Cucurbitacins	Triterpenes	6	Insect deterrent properties and possess medicinal value	Leaves and fruits	Shang et al., 2014
5	<i>Ricinus communis</i> and <i>Jatropha curcas</i>	Casbenes	Diterpenes	1	Possess medicinal value and used in treating cancers and HIV infection	Constitutive expression in leaves, roots and stems	King et al., 2014, 2016
6	<i>Zea mays</i>	2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA)	Hydroxamic acids	4	Defense related activities	Mainly expressed during seedling stages and in roots.	Frey et al., 1997
7	<i>Oryza sativa</i>	Momilactones	Diterpenes	4	Insect deterring properties and anti-fungal properties	Induced expression during pathogen attack	Shimura et al., 2007; Wang et al., 2011
8	<i>O. sativa</i>	Phytocassanes Oryzalides	Diterpenes	2	Defense related activities	Induced expression during pathogen attack	Swaminathan et al., 2009; Wu et al., 2011
9	<i>Lotus japonicus</i>	Linamarin Lotustralin	Cyanogenic glucosides	3	Herbivore deterrent activities	Above ground plant parts	Takos et al., 2011
10	<i>Sorghum bicolor</i>	Dhurrin	Cyanogenic glucosides	1	Herbivore deterrent activities	Above ground plant parts	Takos et al., 2011
11	<i>L. japonicus</i>	20-hydroxy-betulinic acid	Triterpene	3	Possible role in plant development and nodule formation	Elevated expression in roots and nodules	Krokida et al., 2013

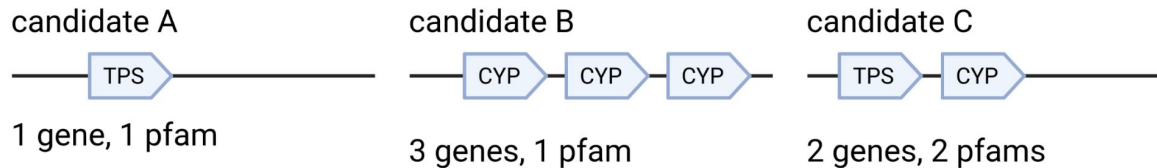




Tools for gene cluster identification - plantiSMASH

- Required input: FASTA file and GFF3 file
- Web server available: <http://plantismash.secondarymetabolites.org/>
- plantiSMASH is plant version of antiSMASH (bacteria tool)
- Many integrated tools: NCBI BLAST+, Diamond, HMMer3, GlimmerHMM, CD-HIT

Biosynthetic gene cluster detection:



Biosynthetic gene cluster extension:



Kautsar et al., 2017: 10.1093/nar/gkx305

Summary

- Long read sequencing technologies
- Genome sequence assembly and annotation
- Synteny
- Biosynthetic gene clusters in plants

Time for questions!

Questions

1. Which sequencing technologies are used to analyze plant genomes?
2. What is synteny?
3. What are the characteristics of a biosynthetic gene cluster?
4. What are evolutionary benefits/forces that cause clustering of genes?
5. Which plant biosynthesis pathways are clustered? (examples)
6. You read about a specific metabolite that is only produced by 10 species of a plant family. How would you approach the identification of underlying genes?