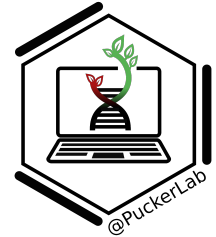
UNIVERSITÄT BONN



Prof. Dr. Boas Pucker

Synteny and Biosynthetic Gene Cluster

Availability of slides

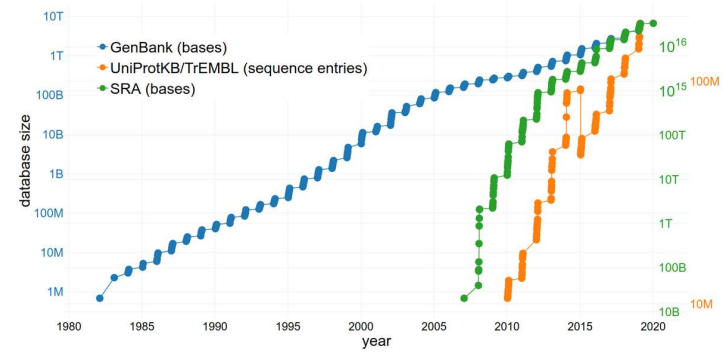
- All materials are freely available (CC BY) - after the lectures:
  - eCampus: WBIO-A-08
  - GitHub: https://github.com/bpucker/teaching

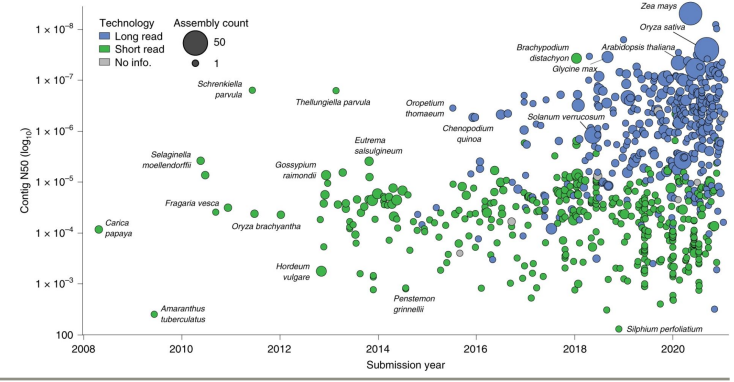
- Questions: Feel free to ask at any time

- Feedback, comments, or questions: pucker[a]uni-bonn.de

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

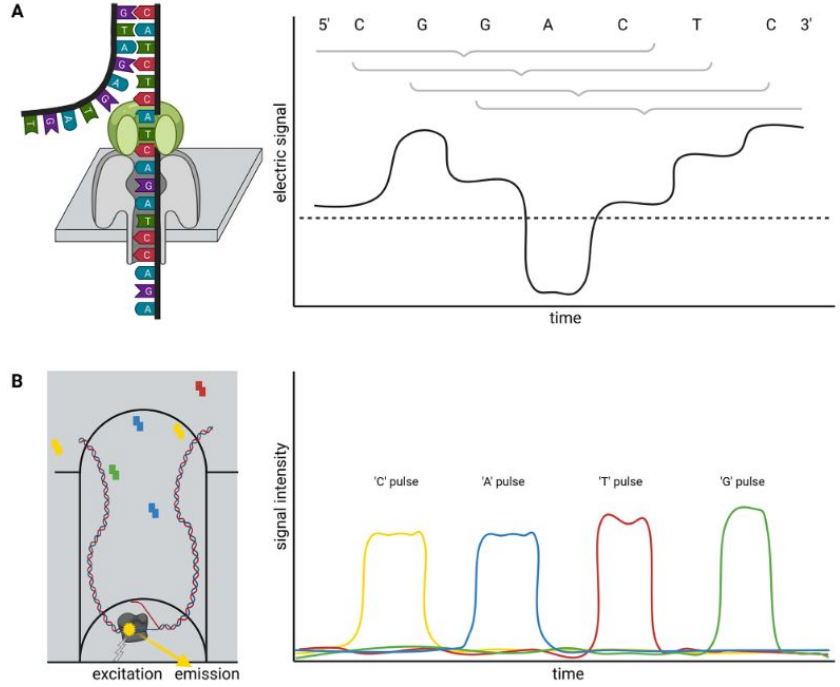




Sielemann et al., 2020: 10.7717/peerj.9954
Marks et al., 2021: 10.1038/s41477-021-01031-8

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)

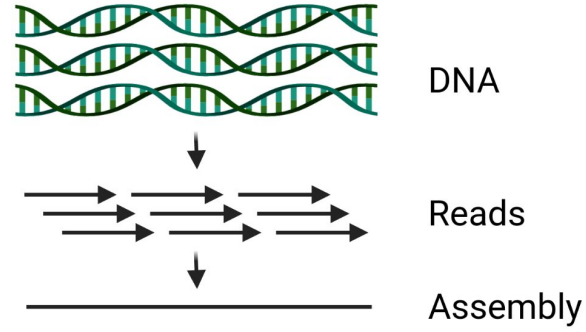


Pucker et al., 2022: 10.1017/qpb.2021.18

UNIVERSITÄT BONN

Genome sequence assembly

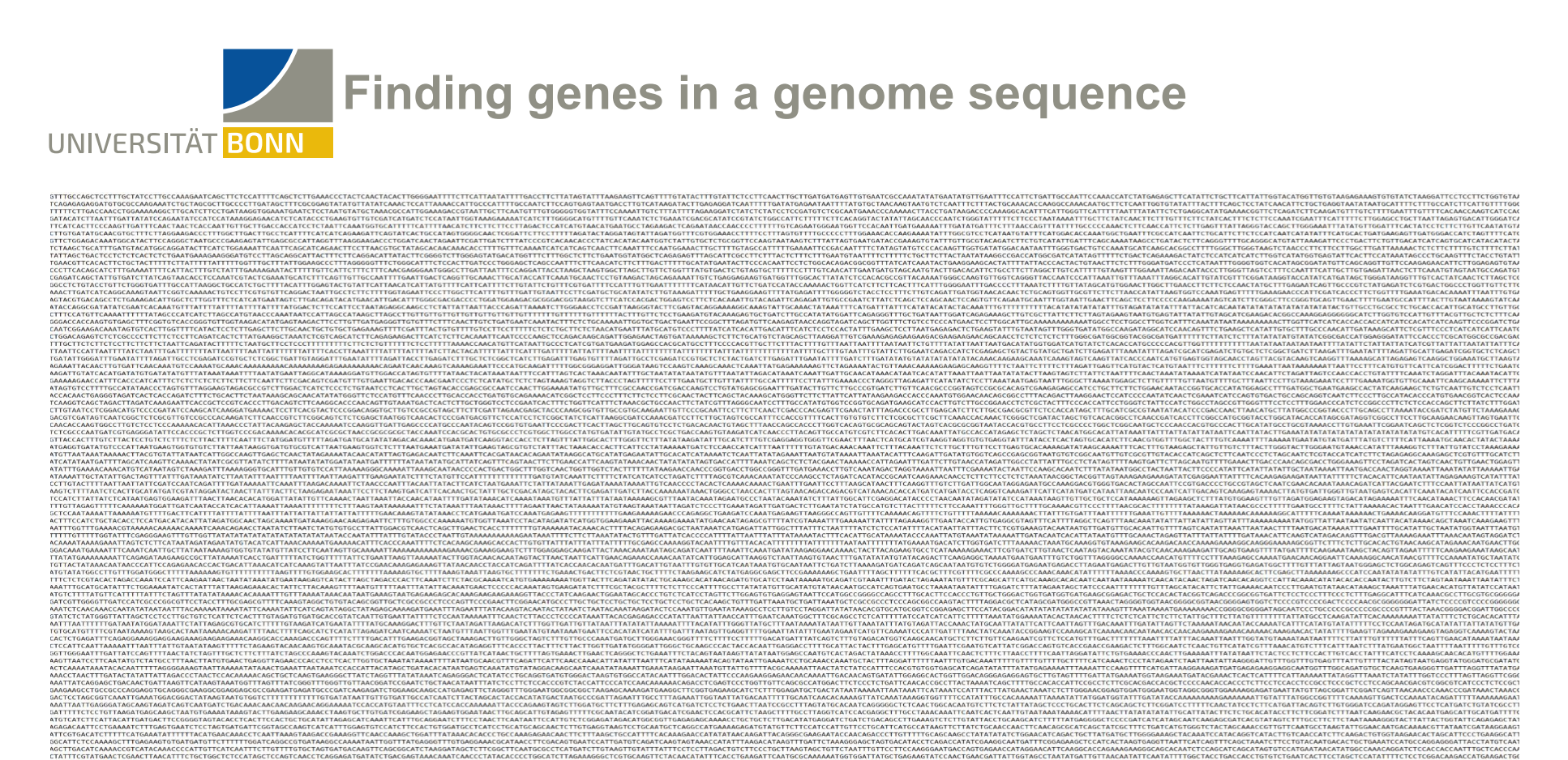
- Reads are shorter than chromosomes - even long reads
- Multiple copies of the genome (DNA exist that can be subjected to sequencing
- Assembly = putting sequence pieces together (finding the common string of all substrings)
- Genome = DNA in a cell
- Genome sequence = representation of the DNA in a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




```
>TRINITY_DN100016_c0_g1_i1
MPKKSSNIQQNNQNRIKRKGRPPKKKKYVQQIDSSDEDILSVRHASTRPRIISIRRNEIPMRPEIHI
>TRINITY_DN100019_c0_g1_i1
MPPKAADKKPAAKAPATASKAPEKKDAGKKTAATGDKKKRTKARKETYSSYIYKVLKQVHPDTGISN
>TRINITY_DN10001_c0_g1_i2
MAKVGNPIVDIETDGSVNEPESSEKNIEVSSSSTQAPESTNTTELLVNEKKAFSLATPAVRRVAREH
>TRINITY_DN100025_c0_g1_i1
MVENQDGCFFKPGWKEFVRSNDLEGGDFLVFNLVDKISYQVVIFDGTCCPKDLCPFSIMNPIFQHLR
>TRINITY_DN10002_c0_g1_i1
MSDELNQVFQRCREQKRPVFVAFVTAGYPDSEETVDILLGLEAGGADIIELGIPFTDPMVDGKTIQD
>TRINITY_DN100061_c0_g1_i1
MQQVVAKLKAIITKTNVTNENSPVENSSSSTSATSSINNSLHGDLSRVFDNMELESNVSNSSISSNI
```

UNIVERSITÄT BONN

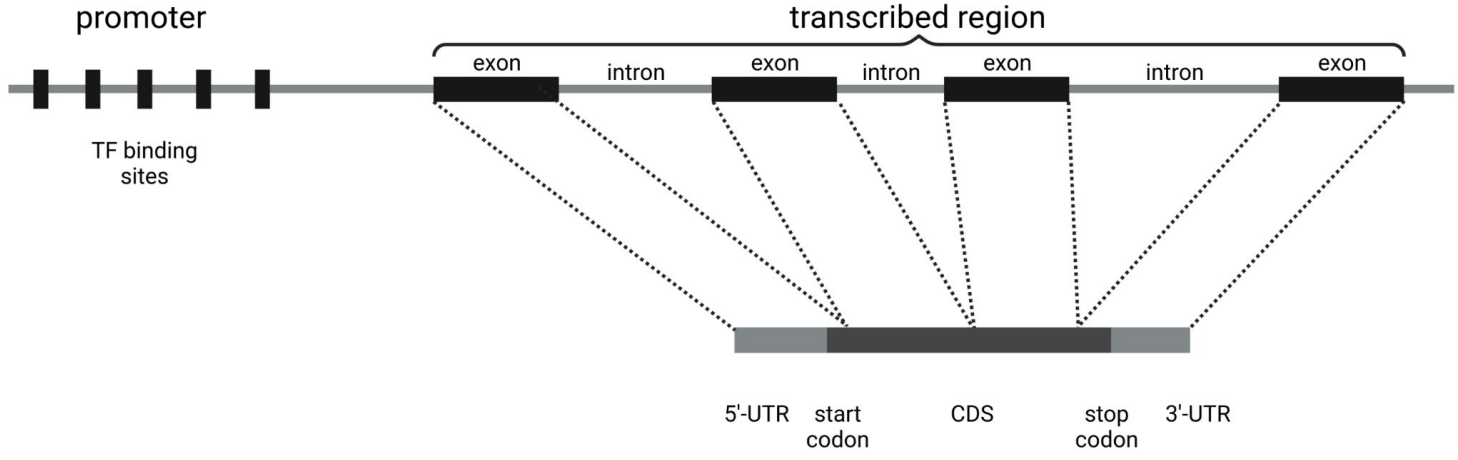
Finding genes in a genome sequence



Siadjeu & Pucker et al., 2020: 10.3390/genes11030274

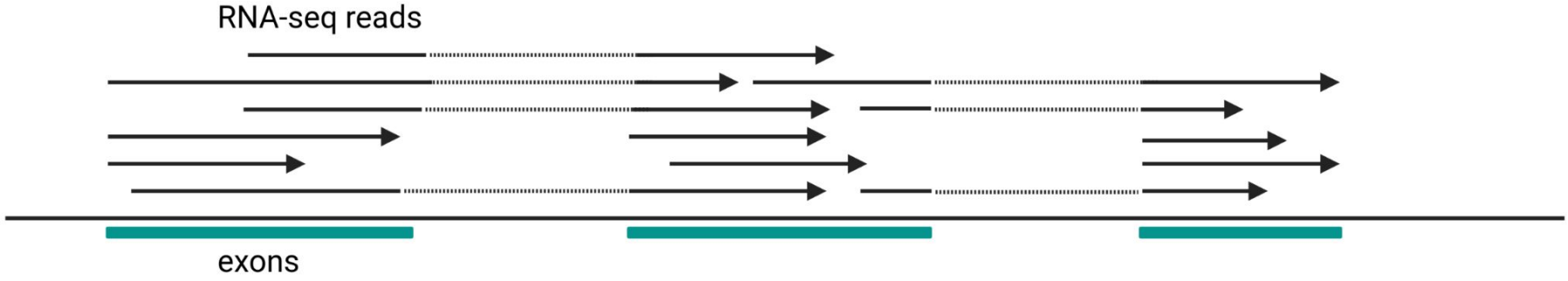
Plant gene structure

- 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

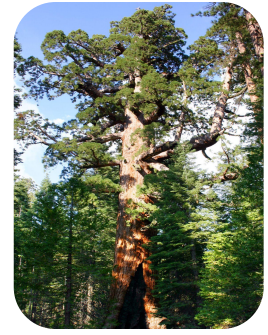


Gene function prediction (functional annotation)

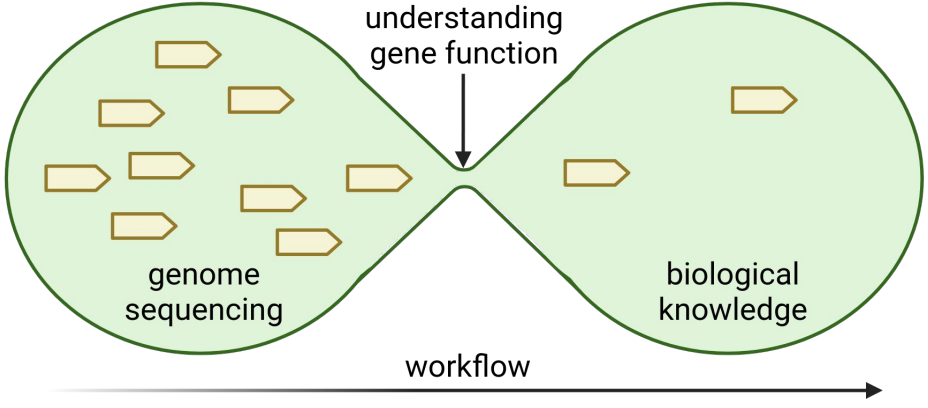
- Thousands of genes per plant species; only a minority of *Arabidopsis thaliana* genes studied

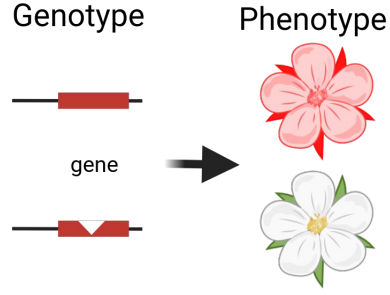
- Not all species are genetically accessible

Arabidopsis thaliana



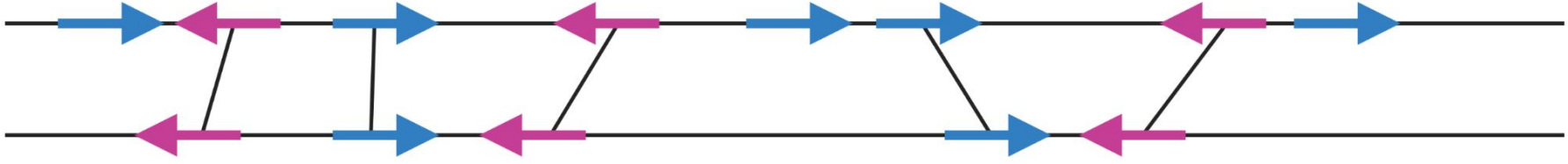
Sequoiadendron giganteum





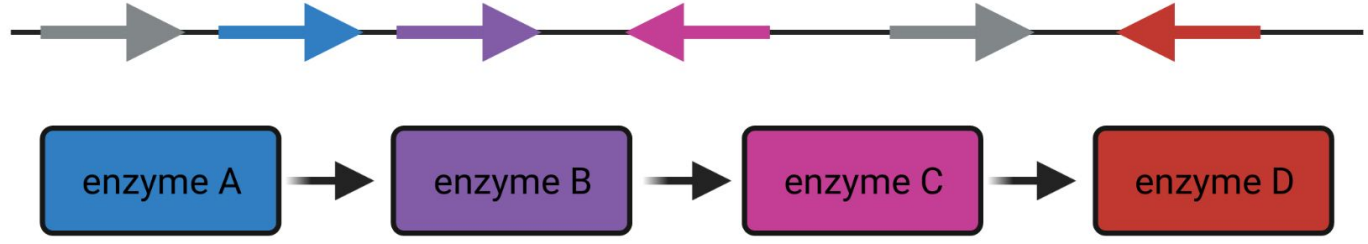
Synteny

- 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

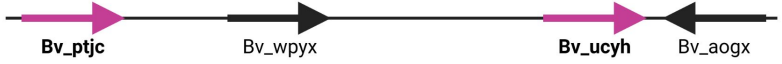
- Thalianol gene cluster

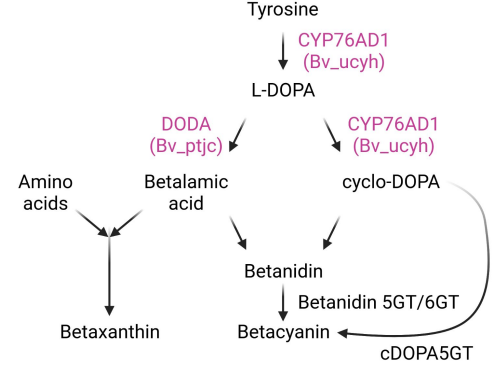
- Marneral gene cluster

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



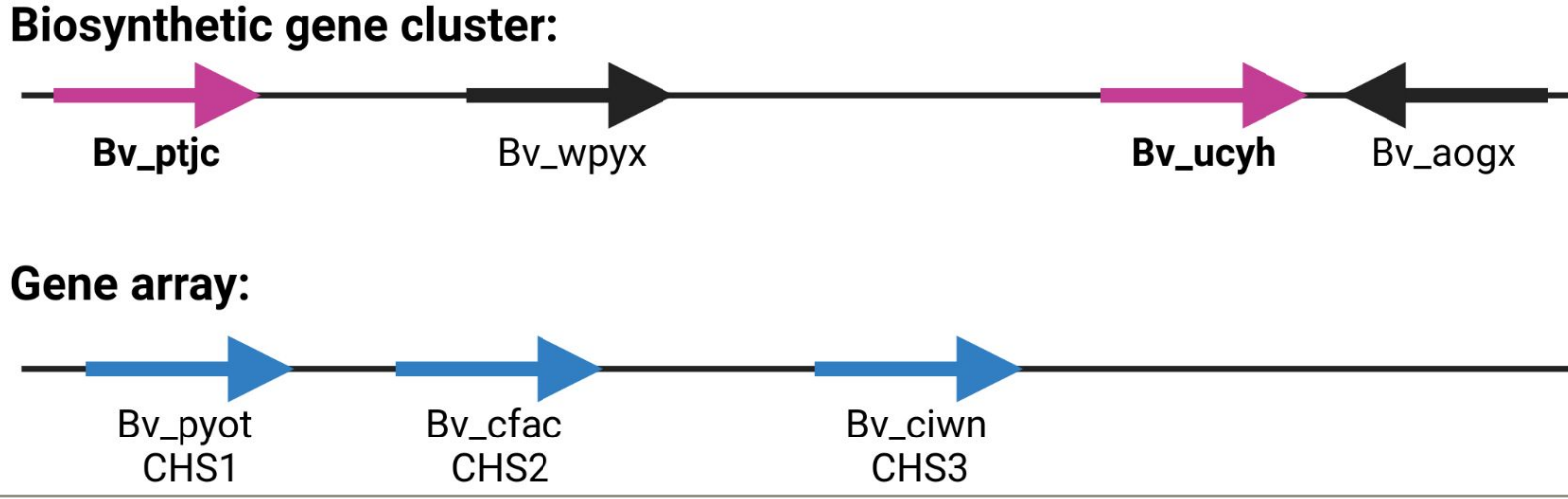


Brockington et al., 2015: 10.1111/nph.13441
Timoneda et al., 2019: 10.1111/nph.15980

Biosynthetic gene clusters vs. gene arrays

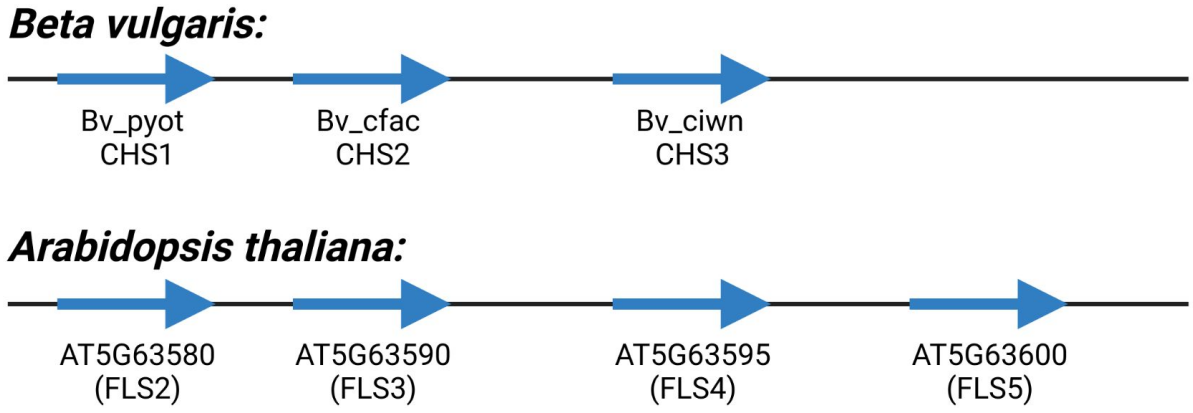
- Biosynthetic gene clusters are characterized by genes encoding enzymes with different functions

- Gene arrays comprise copies of the same gene (result of tandem duplications)



Gene array examples

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



UNIVERSITÄT BONN

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)

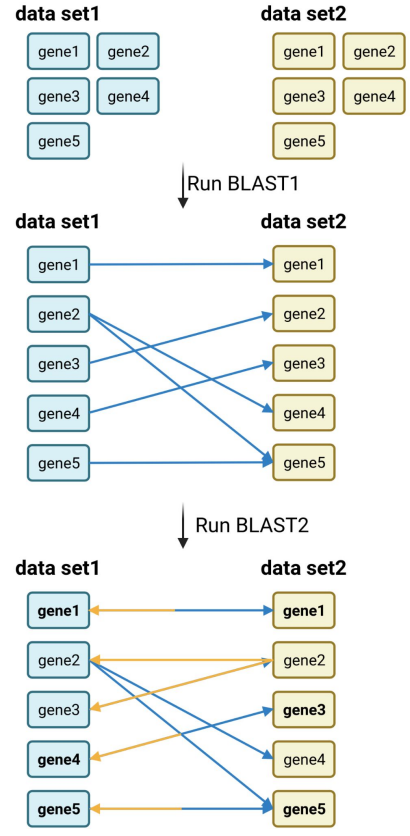
- Co-expression of clustered genes is additional evidence
  - Analyse gene expression across numerous samples

- Conservation across some species can be additional evidence
  - Synteny analysis (JCVI)

UNIVERSITÄT BONN

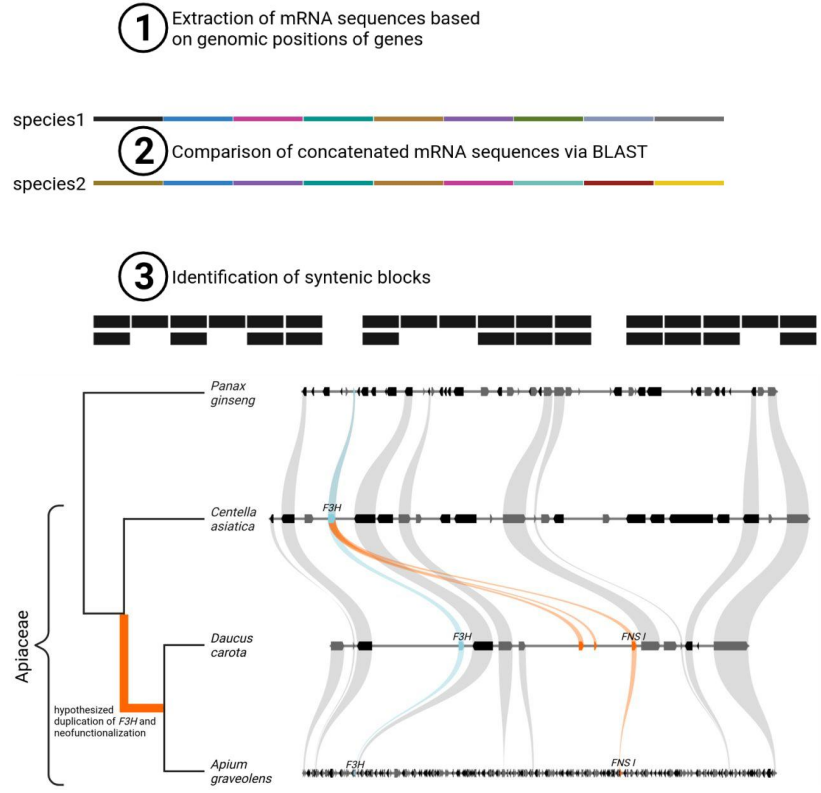
Reciprocal Best BLAST hits (RBHs)

- 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

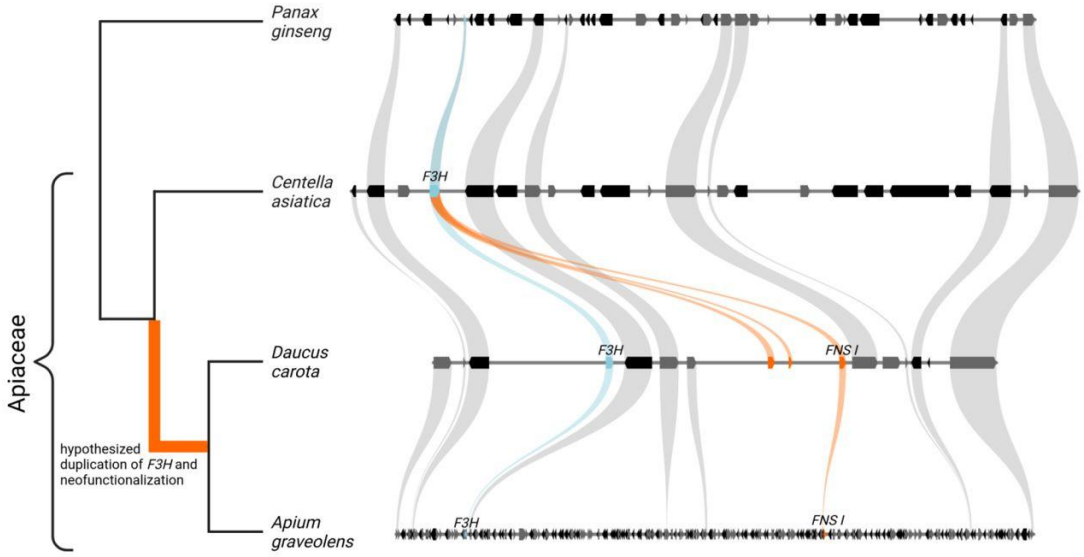


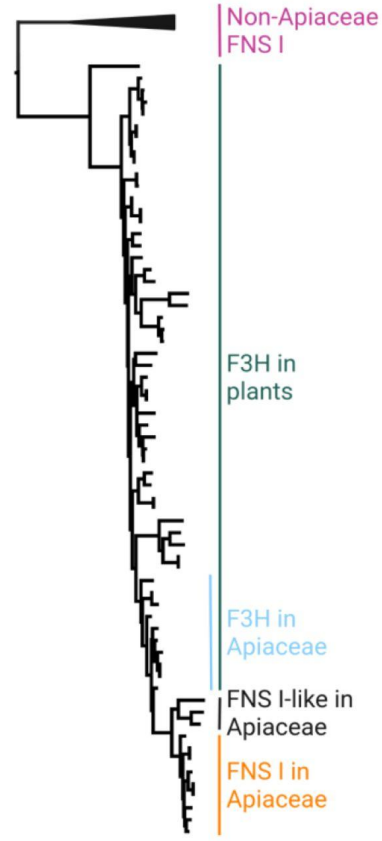
Pucker & Iorizzo, 2022: 10.1101/2022.02.16.480750

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

UNIVERSITÄT BONN

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

UNIVERSITÄT BONN

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

UNIVERSITÄT BONN

Biosynthetic gene cluster - examples

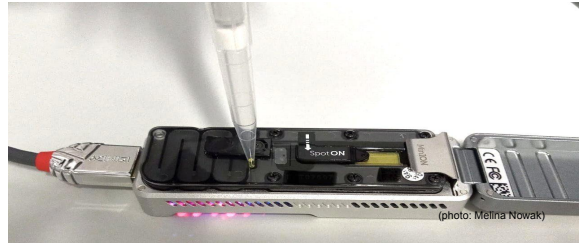
S.no	Species	Metabolite	Compound class	Chromosome (core genes)	Function	Tissue of expression	References
1	*Arabidopsis thaliana*	Thalianol	Triterpenes	5	Unknown physiological function, unregulated expression of cluster genes leads to dwarfing of plant. Modulate root microbiome content	Roots	Field and Osbourn, 2008; Field et al., 2011; Chen et al., 2019a
2	*A. thaliana*	Marneral	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	*Avena strigosa*	Avenacins	Triterpenes	1	Defense against pathogens	Roots	Li et al., 2021a
4	*Cucumis sativus*	Cucurbitacins	Triterpenes	6	Insect deterrent properties and possess medicinal value	Leaves and fruits	Shang et al., 2014
5	*Ricinus communis and Jatropha curcas*	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	*Zea mays*	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	*Oryza sativa*	Momilactones	Diterpenes	4	Insect deterring properties and anti-fungal properties	Induced expression during pathogen attack	Shimura et al., 2007; Wang et al., 2011
8	*O. sativa*	Phytocassanes Oryzalides	Diterpenes	2	Defense related activities	Induced expression during pathogen attack	Swaminathan et al., 2009; Wu et al., 2011
9	*Lotus japonicus*	Linamarin Lotuastralin	Cyanogenic glucosides	3	Herbivore deterrent activities	Above ground plant parts	Takos et al., 2011
10	*Sorghum bicolor*	Dhurrin	Cyanogenic glucosides	1	Herbivore deterrent activities	Above ground plant parts	Takos et al., 2011
11	*L. japonicus*	20-hydroxy-betulinic acid	Triterpene	3	Possible role in plant development and nodule formation	Elevated expression in roots and nodules	Krokida et al., 2013

Bharadwaj et al., 2021: 10.3389/fpls.2021.697318

EXAMPLE: Withanolides

- Antiproliferative activity (potential cancer drugs)
- Biosynthesis is largely unknown except for first step

Withania somnifera



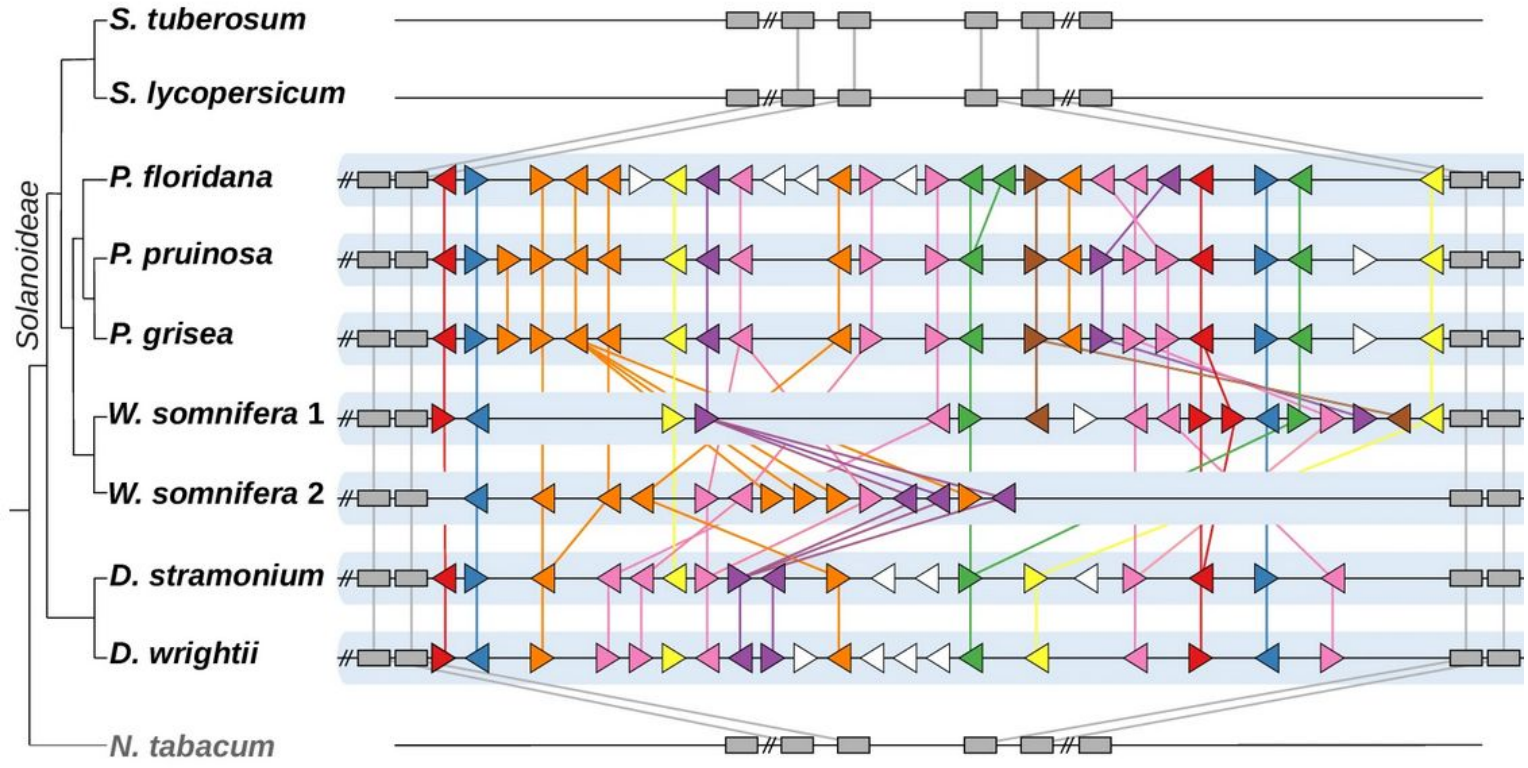
- ONT sequencing, genome sequence construction, annotation

Withania somnifera
(Wowbobwow12, CC BY-SA 3.0)

Datura innoxia
(Flobbadob, CC BY-SA 4.0)

Nicandra physalodes
(Flobbadob, CC BY-SA 4.0)

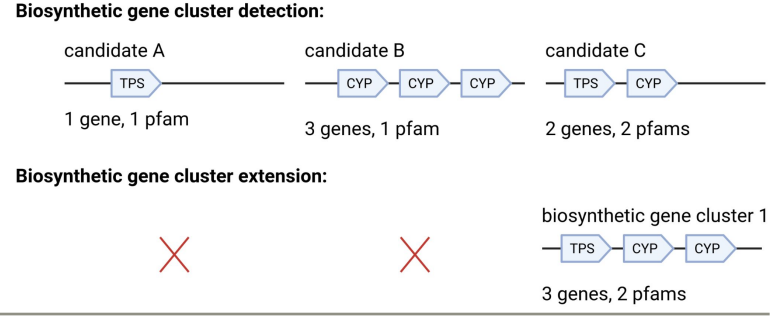
EXAMPLE: Withanolide biosynthesis gene cluster



Hakim, Choudhary, Malhotra, Peng et al., 2024: 10.1038/s41467-025-61686-1

Tools for gene cluster identification - plantiSMASH

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



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

Summary

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

UNIVERSITÄT BONN

Time for questions!

UNIVERSITÄT BONN

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?