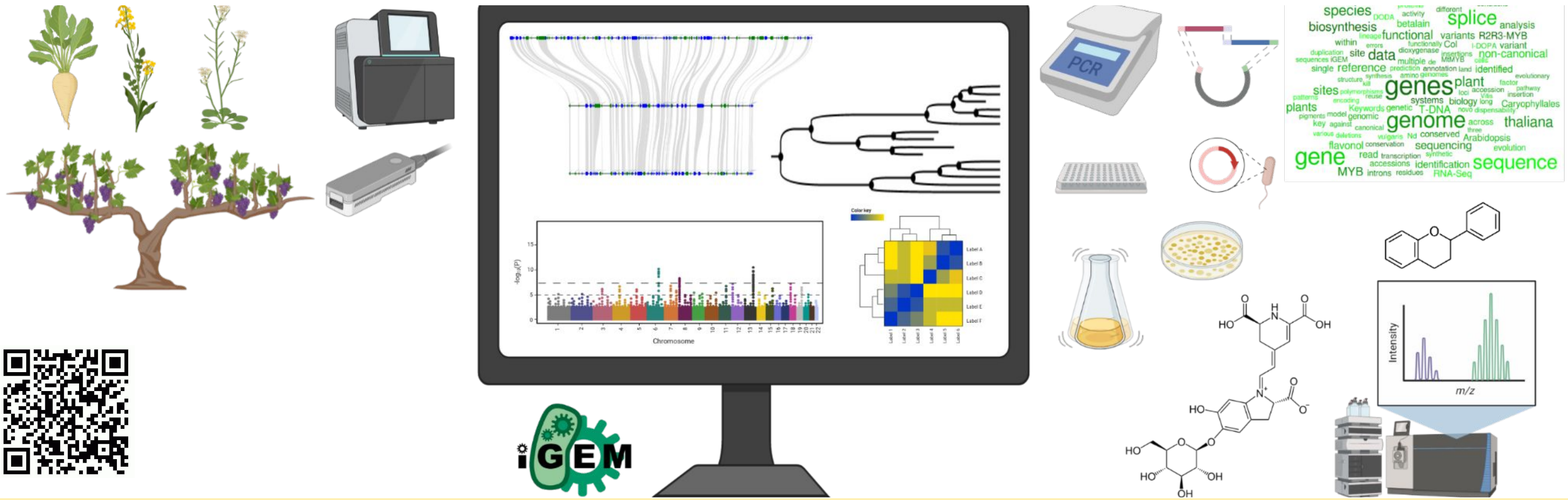
Technische Universität Braunschweig


PCR
Chromosome
-log10(P)
Color key
Intensity
m/z
iGEM
species biosynthesis splice functional variants data genes plant genome gene sequence sequencing thaliana Arabidopsis
HO
OH
N
O

GE32/MM12 - Data Reuse

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

Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - StudIP: **GE32/MM12**
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

- Feedback, comments, or questions: b.pucker[a]tu-braunschweig.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. Some figure were constructed using bioRender.com.

What are the advantages of re-use?

What are the advantages of re-use?

- Cost-effective
- Immediately available
- Extremely large datasets

Challenges with data re-use?

- Lacking metadata
- Unknown details/issues
- Mislabeling possible
- Not perfectly matching needs
- Technology outdated

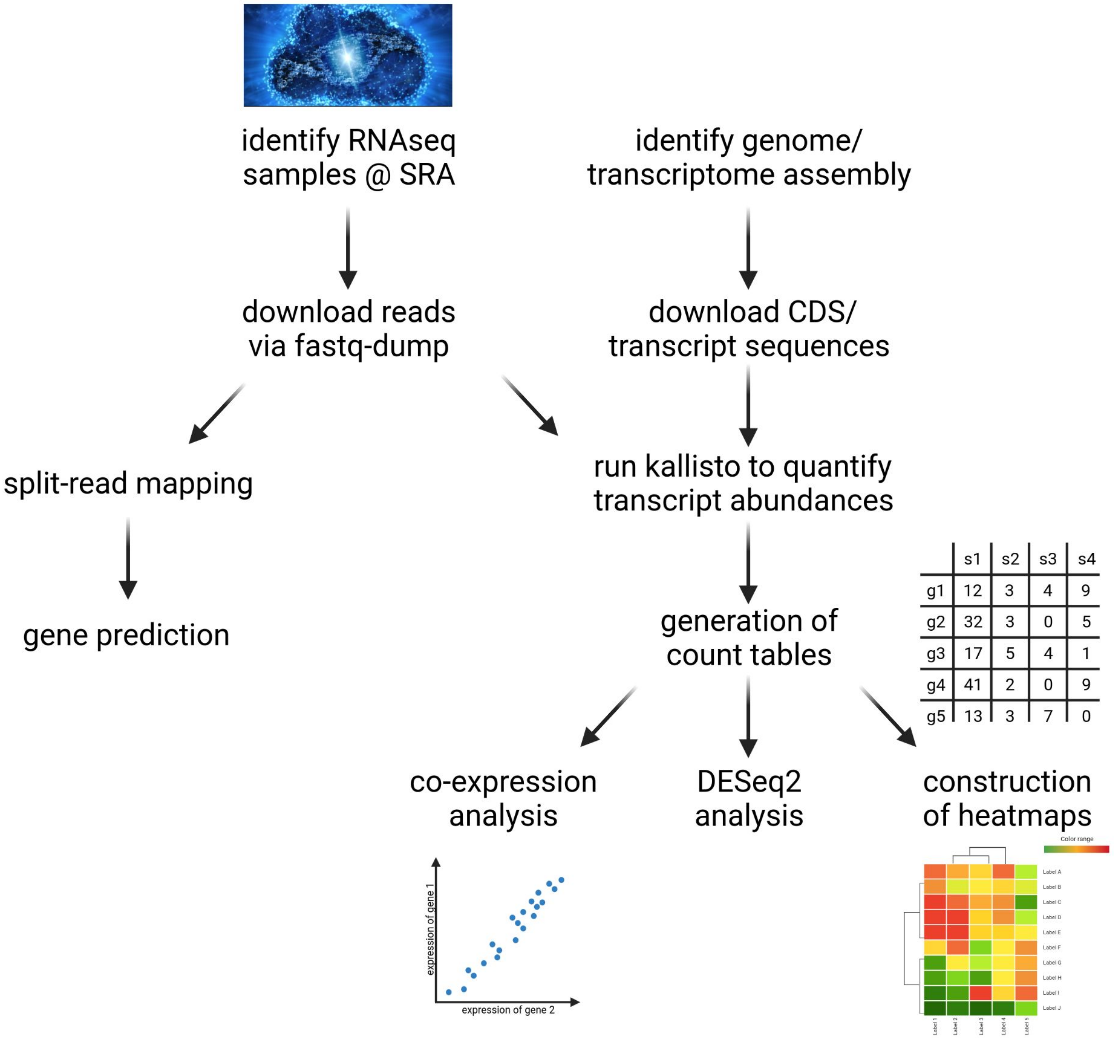
Find gene expression data sets

- SRA read selector
- Gene Expression Omnibus (GEO) search
- Publications: data availability statements & supplementary tables

How to retrieve data?

- Preprocessed data sets (count tables @ GEO)
- Cloud solutions (Galaxy)
- Fastq-dump

Workflow



identify RNAseq samples @ SRA

download reads via fastq-dump

split-read mapping

gene prediction

identify genome/ transcriptome assembly

download CDS/ transcript sequences

run kallisto to quantify transcript abundances

generation of count tables

	s1	s2	s3	s4
g1	12	3	4	9
g2	32	3	0	5
g3	17	5	4	1
g4	41	2	0	9
g5	13	3	7	0

co-expression analysis

DESeq2 analysis

construction of heatmaps

Technische Universität Braunschweig

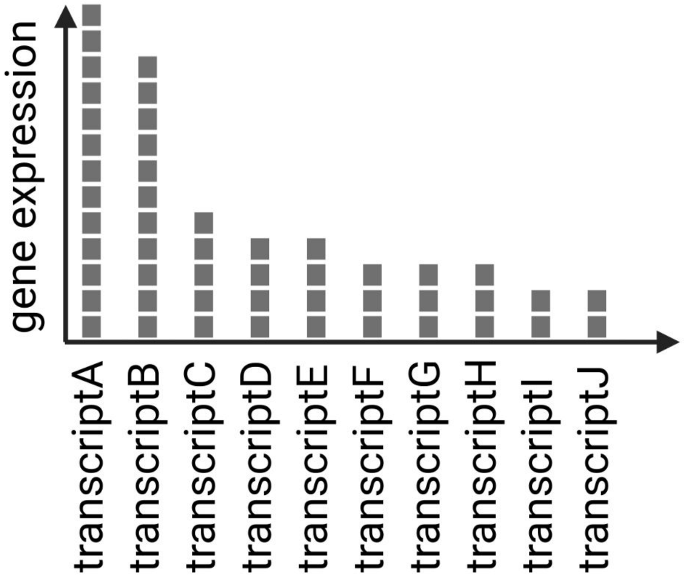
How to check RNA-seq data sets?

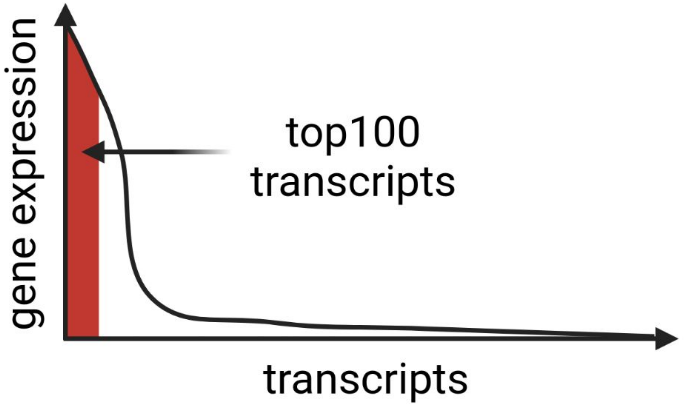
RNA-seq quality control

- Percentage of reads mapped to individual mRNA sequences

- Distribution of abundance across transcripts
  - ○ Substantial coverage of top100 transcripts

- Metadata assessment / marker gene check
  - ○ Highly expressed marker genes like RuBisCO





How to analyze the distribution of species?

GBIF

- Sources have heterogeneous quality
- Extensive filtering required
- What types of issues can be expected?

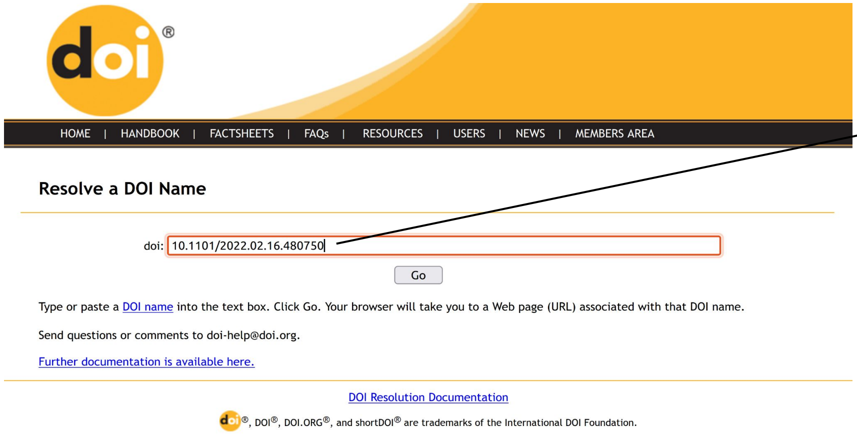
GBIF

- Coordinates located in the sea
- Coordinates located in zoos/botanical gardens
- Coordinates located at the center of grids/in capitals
- Entries of fossils / too old entries
- Entries containing likely typos

How to cite data sets?

Digital Object Identifier (DOI)

- DOI = Digital Object Identifier
- Unique and short way to point to a publication or a data set
- How to resolve a DOI? https://dx.doi.org/


doi®
HOME | HANDBOOK | FACTSHEETS | FAQs | RESOURCES | USERS | NEWS | MEMBERS AREA
Go

Resolve a DOI Name

doi: 10.1101/2022.02.16.480750

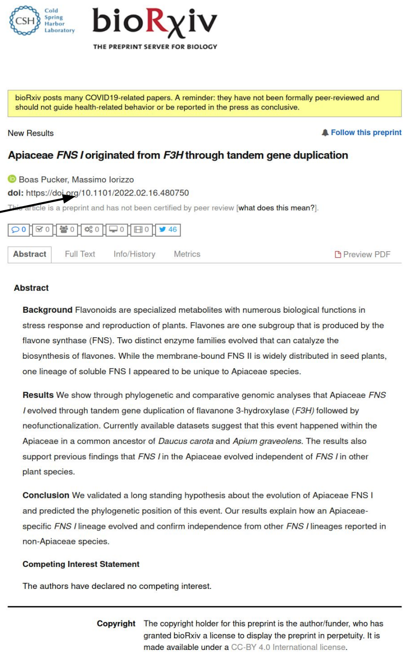
Type or paste a DOI name into the text box. Click Go. Your browser will take you to a Web page (URL) associated with that DOI name.

Send questions or comments to doi-help@doi.org.

Further documentation is available here.

DOI Resolution Documentation

®, DOI®, DOI.ORG®, and shortDOI® are trademarks of the International DOI Foundation.


CSH Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY
Follow this preprint
0 0 0 0 0 0 46
Abstract Full Text Info/History Metrics
Preview PDF

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

Apiaceae *FNS I* originated from *F3H* through tandem gene duplication

Boas Pucker, Massimo Iorizzo

doi: https://doi.org/10.1101/2022.02.16.480750

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract

Background Flavonoids are specialized metabolites with numerous biological functions in stress response and reproduction of plants. Flavones are one subgroup that is produced by the flavone synthase (FNS). Two distinct enzyme families evolved that can catalyze the biosynthesis of flavones. While the membrane-bound FNS II is widely distributed in seed plants, one lineage of soluble FNS I appeared to be unique to Apiaceae species.

Results We show through phylogenetic and comparative genomic analyses that Apiaceae *FNS I* evolved through tandem gene duplication of flavanone 3-hydroxylase (*F3H)* followed by neofunctionalization. Currently available datasets suggest that this event happened within the Apiaceae in a common ancestor of *Daucus carota* and *Apium graveolens*. The results also support previous findings that *FNS I* in the Apiaceae evolved independent of *FNS I* in other plant species.

Conclusion We validated a long standing hypothesis about the evolution of Apiaceae FNS I and predicted the phylogenetic position of this event. Our results explain how an Apiaceae-specific *FNS I* lineage evolved and confirm independence from other *FNS I* lineages reported in non-Apiaceae species.

Competing Interest Statement

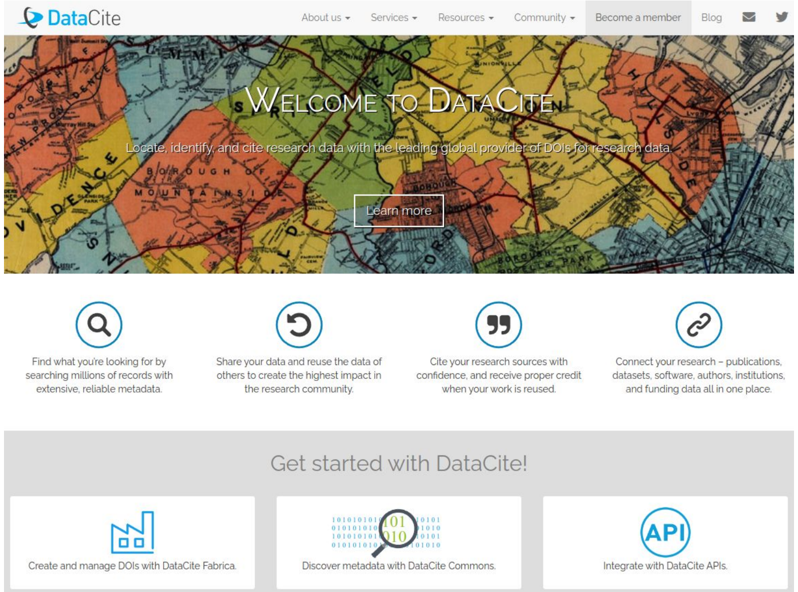
The authors have declared no competing interest.

Copyright The copyright holder for this preprint is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

DataCite

- Central service for generation and management of DOIs for research data sets

- Enable connection and reuse of data sets



WELCOME TO DATACITE

Locate, identify, and cite research data with the leading global provider of DOIs for research data.

Find what you're looking for by searching millions of records with extensive, reliable metadata.

Share your data and reuse the data of others to create the highest impact in the research community.

Cite your research sources with confidence, and receive proper credit when your work is reused.

Connect your research – publications, datasets, software, authors, institutions, and funding data all in one place.

Get started with DataCite!

Create and manage DOIs with DataCite Fabrica.

Discover metadata with DataCite Commons.

Integrate with DataCite APIs.

https://datacite.org/index.html

Managing references

- Entries about publications / data sets are stored in a local database
- Convenient citation when writing manuscripts
- Support for comments and key words assigned to entries
- Examples:
    - Zotero: free
    - Mendeley: free, but belongs to Elsevier
    - Citavi: commercial, but campus license
    - EndNote: commercial

Data reuse examples

Benchmarking of NOVOPlasty

- Public sequence read data sets are used for plastome assemblies
- Data sets can be selected based on specific criteria
- Pure bioinformatics groups can work with real data sets
- No costs for generation of data sets

Dierckxsens et al., 2016: 10.1093/nar/gkw955

Benchmarking SANPolyA

- SANPolyA detects Poly(A) signals through deep learning
- Comparison of SANPolyA results against results of other tools
- Freely available output of tools is required for systematic comparison

Yu & Dai, 2020: 10.1093/bioinformatics/btz970

Pangenome of hexaploid bread wheat

- Pangenome = combination of all genomes of a species
- Integration of public data sets to identify presence/absence of genes in wheat cultivars
- Power of pangenomes increases with number of analyzed

Montenegro et al., 2017: 10.1111/tpj.13515

Single plant GWAS + bulk segregant analysis

- spGWAS = single plant genome-wide association studies
- Bulk segregant analysis = comparisons of plant groups with contrasting phenotypes
- Comparison of different approaches to identify genetic basis of a trait (plant height)
- Evaluation of findings against previous reports

Gyawali et al., 2019: 10.1186/s12870-019-2000-y

Co-expression gene network analysis

- Genes belonging to the same pathway show similar expression patterns
- Bamboo genome sequence was integrated with RNA-seq data sets
- Identification of co-expression modules associated with development

Ma et al., 2018: 10.3389/fgene.2018.00574

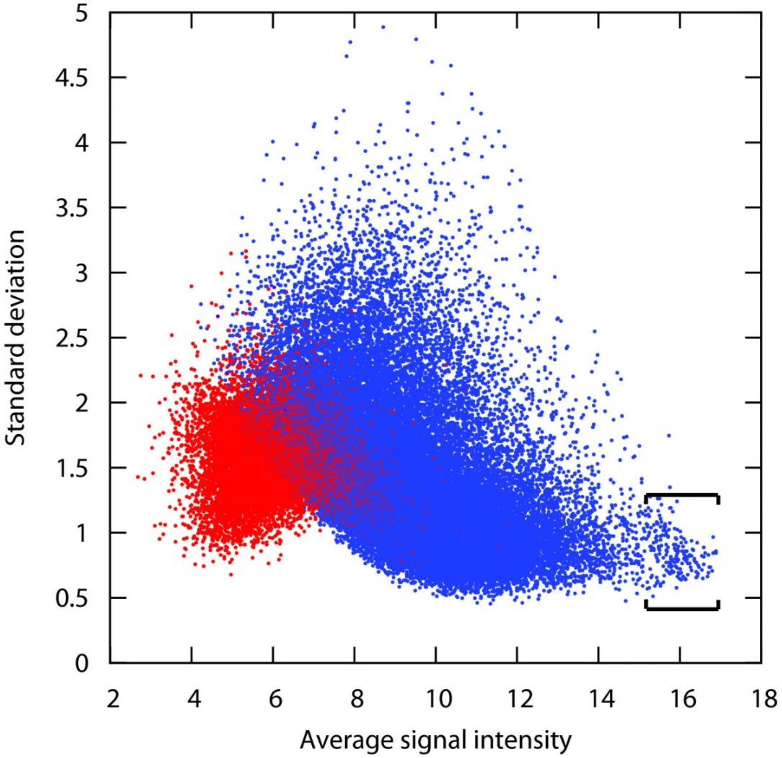
Integration of GWAS and co-expression analysis

- GWAS allows the identification of genomic regions associated with a trait
- Co-expression analysis can help to identify individual genes in these regions
- RNA-seq data sets of most species are available for free

Schaefer et al., 2018: 10.1105/tpc.18.00299

Identification of RT-qPCR reference genes

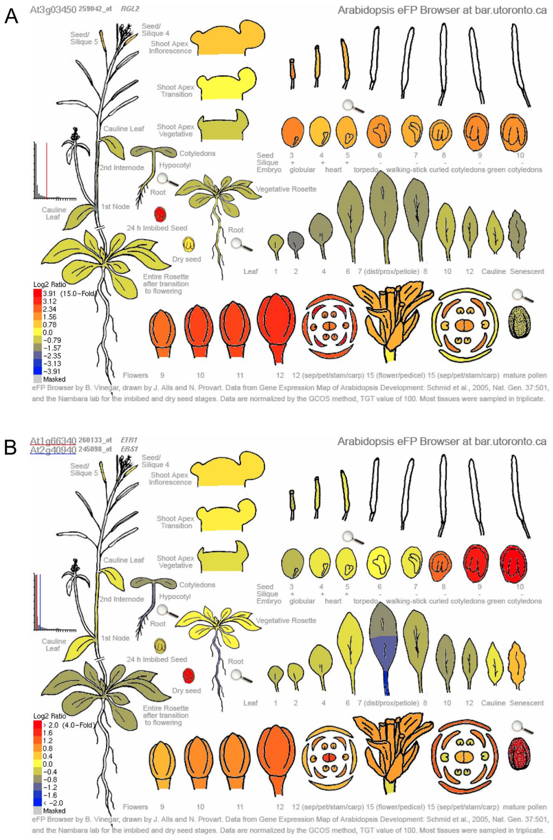
- RNA-seq data sets can be analyzed to identify constantly expressed genes
- Reference genes with constant expression across samples is crucial for RT-qPCRs
- Reference genes can be specific for certain conditions/tissues



Hruz et al., 2011: 10.1186/1471-2164-12-156

Electronic Fluorescent Pictograph browser

- Integration of existing microarray datasets
- Option to integrate other large scale datasets
- Basis for hypothesis generation
- Visualization of gene expression in many different plant tissues/conditions



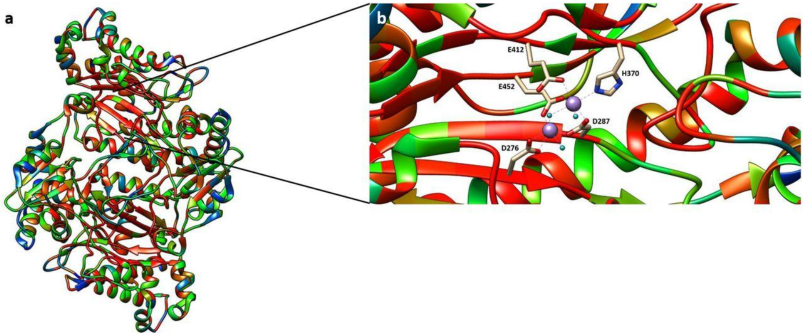
eFP Browser by B. Vinegar, drawn by J. Alls and N. Provart. Data from Gene Expression Map of Arabidopsis Development: Schmid et al., 2005, Nat. Gen. 37:501, and the Nambara lab for the imbibed and dry seed stages. Data are normalized by the GCOS method, TGT value of 100. Most tissues were sampled in triplicate.

eFP Browser by B. Vinegar, drawn by J. Alls and N. Provart. Data from Gene Expression Map of Arabidopsis Development: Schmid et al., 2005, Nat. Gen. 37:501, and the Nambara lab for the imbibed and dry seed stages. Data are normalized by the GCOS method, TGT value of 100. Most tissues were sampled in triplicate.

Winter et al., 2007: 10.1371/journal.pone.0000718

Identification of conserved amino acid residues

- Identification of orthologous sequences in hundreds of species
- Comparison of sequences to identify highly conserved amino acid residues
- 3D modeling based on known structures of homologs

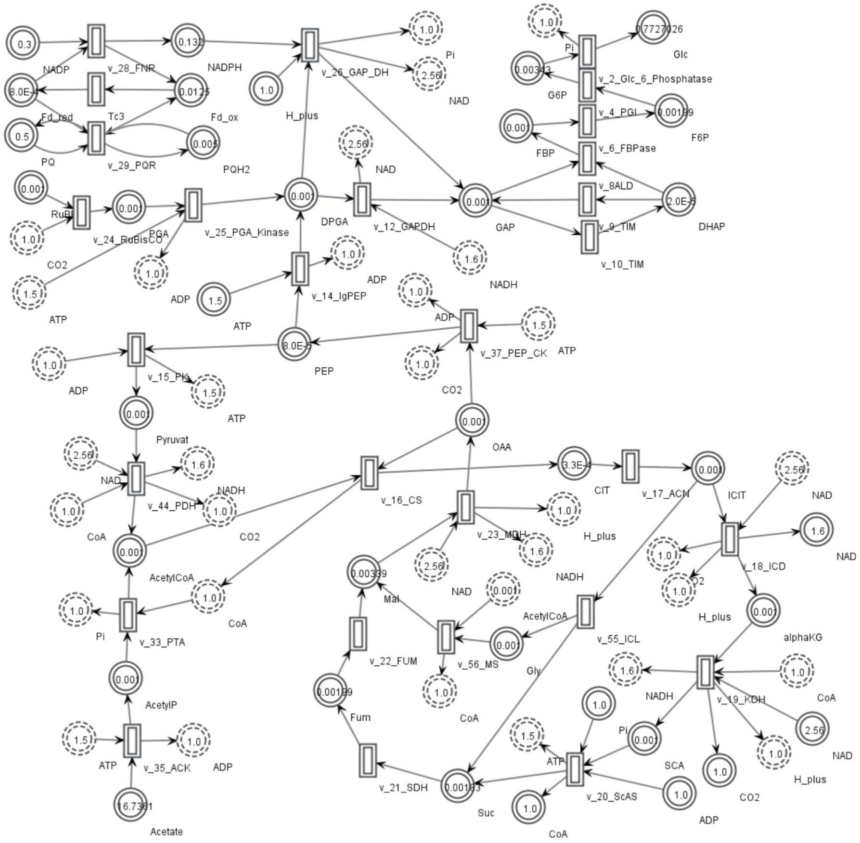


	Animals	Plants	Fungi	Archaea	Bacteria
D276	94	99	100	100	100
D287	94	98	100	100	100
H370	94	98	100	100	100
E412	94	96	100	100	100
E452	91	97	100	100	100
T289	94	97	100	100	97
T410	93	96	100	79	100
H377	94	98	100	100	97
R398	93	98	89	10	57
W107	88	98	96	0	96
Y241	94	96	100	2	90
I244	93	98	97	88	100
H255	94	98	100	100	100
V376	89	1	38	81	94
C58	58	64	0	0	0
C158	40	1	0	0	0

Schilbert et al., 2018: 10.1101/423475

VANESA: enzyme properties for model

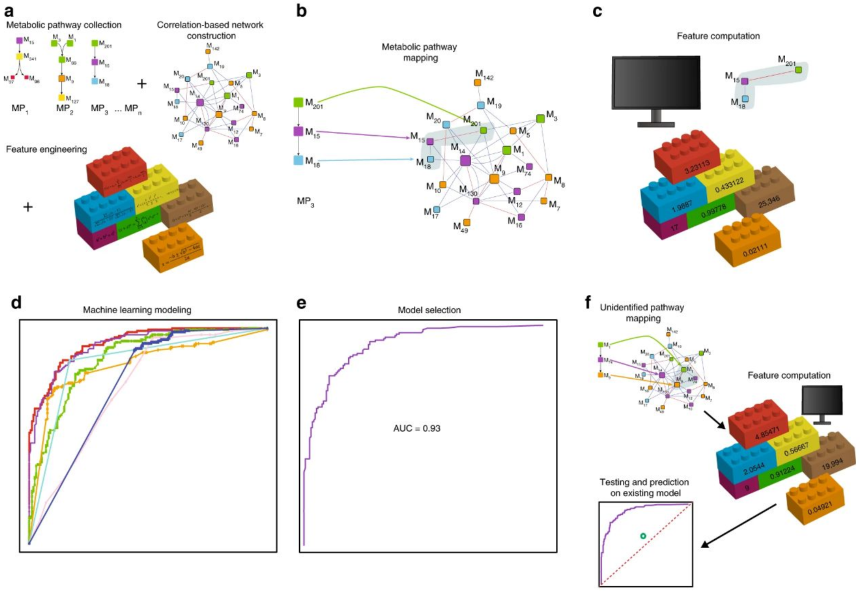
- Pathway structure taken from KEGG
- Integration of enzymatic properties (K_M, V_{max}) from BRENDA
- Freely accessible metabolic modeling solution



Brinkrolf et al., 2018: 10.1515/jib-2018-0018

Inference of metabolic pathways from metabolite data

- Correlation-based inference of metabolic pathways from metabolite data sets
- Machine learning was deployed
- Identification of the most important features

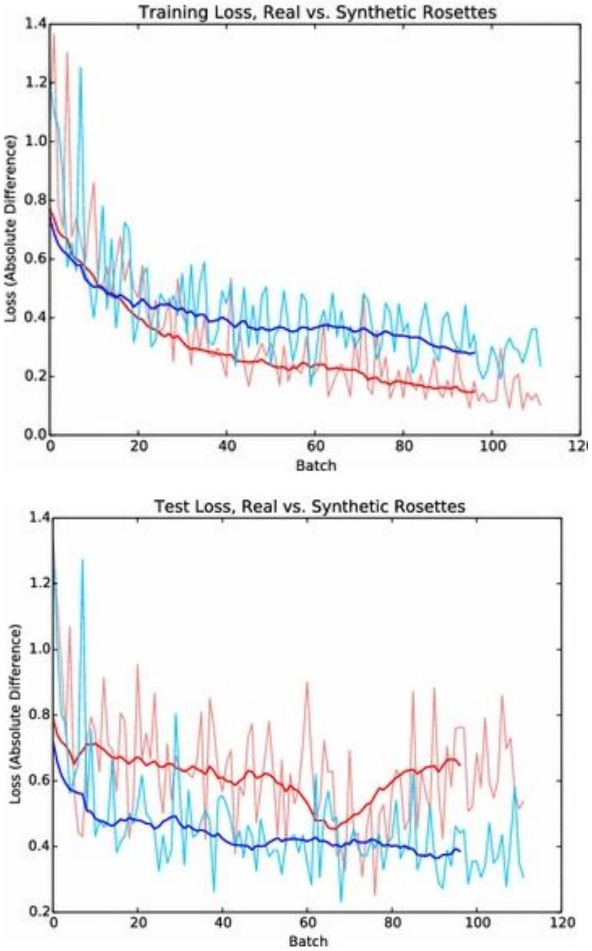

a
Metabolic pathway collection
Correlation-based network construction
MP1 MP2 MP3 ... MPn
Feature engineering
b
Metabolic pathway mapping
MP3
c
Feature computation
d
Machine learning modeling
e
Model selection
AUC = 0.93
f
Unidentified pathway mapping
Feature computation
Testing and prediction on existing model

Combined correlation-based network analysis and machine learning workflow. The workflow of the current study: **a** Metabolic pathways were gathered from existing repositories. In parallel, correlation-based networks of metabolites were constructed for the tissue of the organism of interest (here, the tomato pericarp). In addition, a vector of features was engineered based on network properties. **b** Metabolic pathways with partial to full coverage in the correlation networks were mapped to the networks. Each pathway was considered as a single instance. Training and test sets were proposed based on the existence of the pathways in the tomato. **c** A set of features was computed for each instance in the training set (for the current study 148 * 3 networks = 444 features in total). **d** The training set was used to generate different ML models. **e** The model that generated the best performance measures (the AUC) was selected. The ML model was validated in silico using cross-validation. **f** Test set instances were mapped onto the networks with subsequent feature computation. The proposed ML model was used to predict the potential existence of unidentified pathways in the tomato pericarp

Toubiana et al., 2019: 10.1038/s42003-019-0440-4

Leaf counting

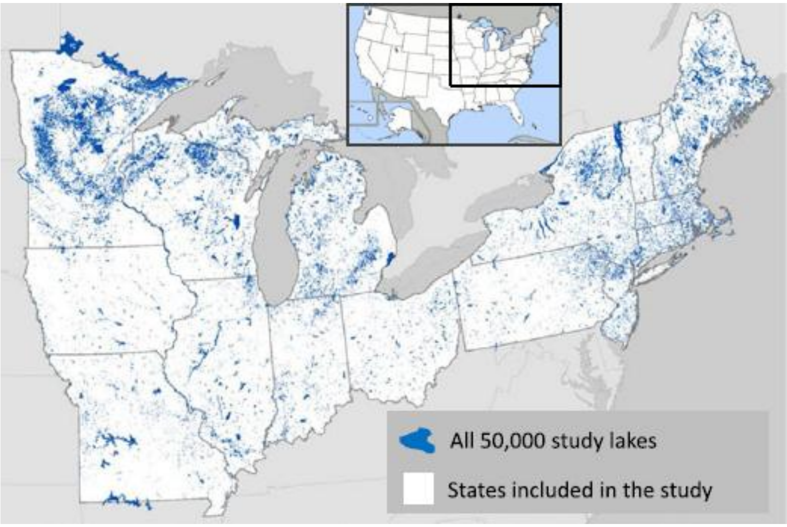
- Automatic leaf counting in rosette plants
- Deep convolutional neural networks require large and diverse data sets for optimization
- Synthetic plants are proposed as an alternative



Ubbens, et al., 2018: 10.1186/s13007-018-0273-z

LAGOS: Ecology database

- LAGOS = LAke multi-scaled GeOSpatial and temporal database
- Combination of site-based ecosystem datasets with national geospatial datasets
- Best practice example for lake related data sets covering about 50,000 lakes
- 100 individual datasets about water quality



Soranno et al., 2015: 10.1186/s13742-015-0067-4

Time for questions!

Questions

1. What are the advantages/disadvantages of reuse?
2. How can you assess the quality of RNA-seq data sets?
3. How to filter GBIF data sets?
4. How can you cite data sets?
5. Which studies used existing data sets to gain novel insights?