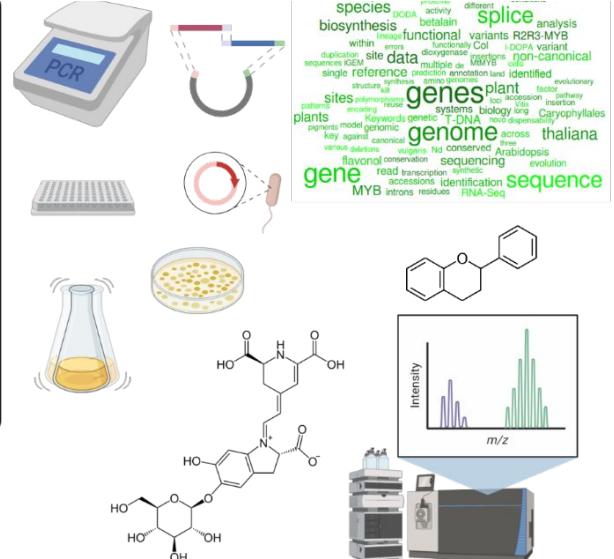
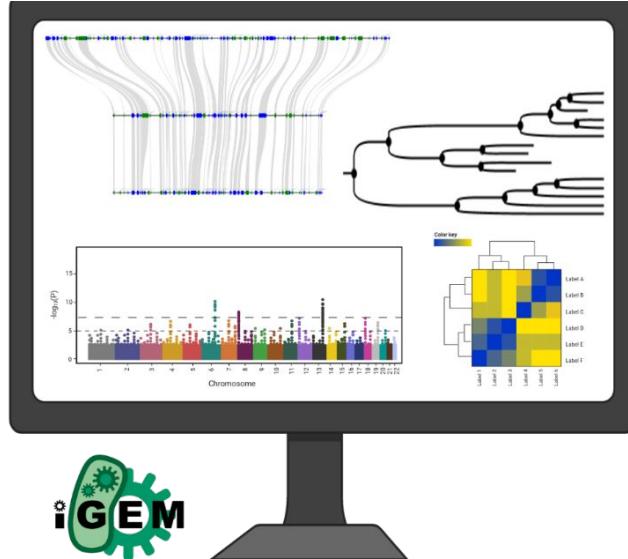
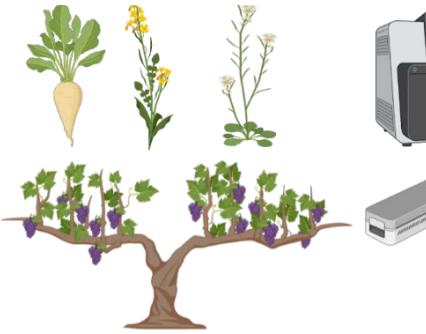




Technische  
Universität  
Braunschweig



# Introduction to data literacy in genomics

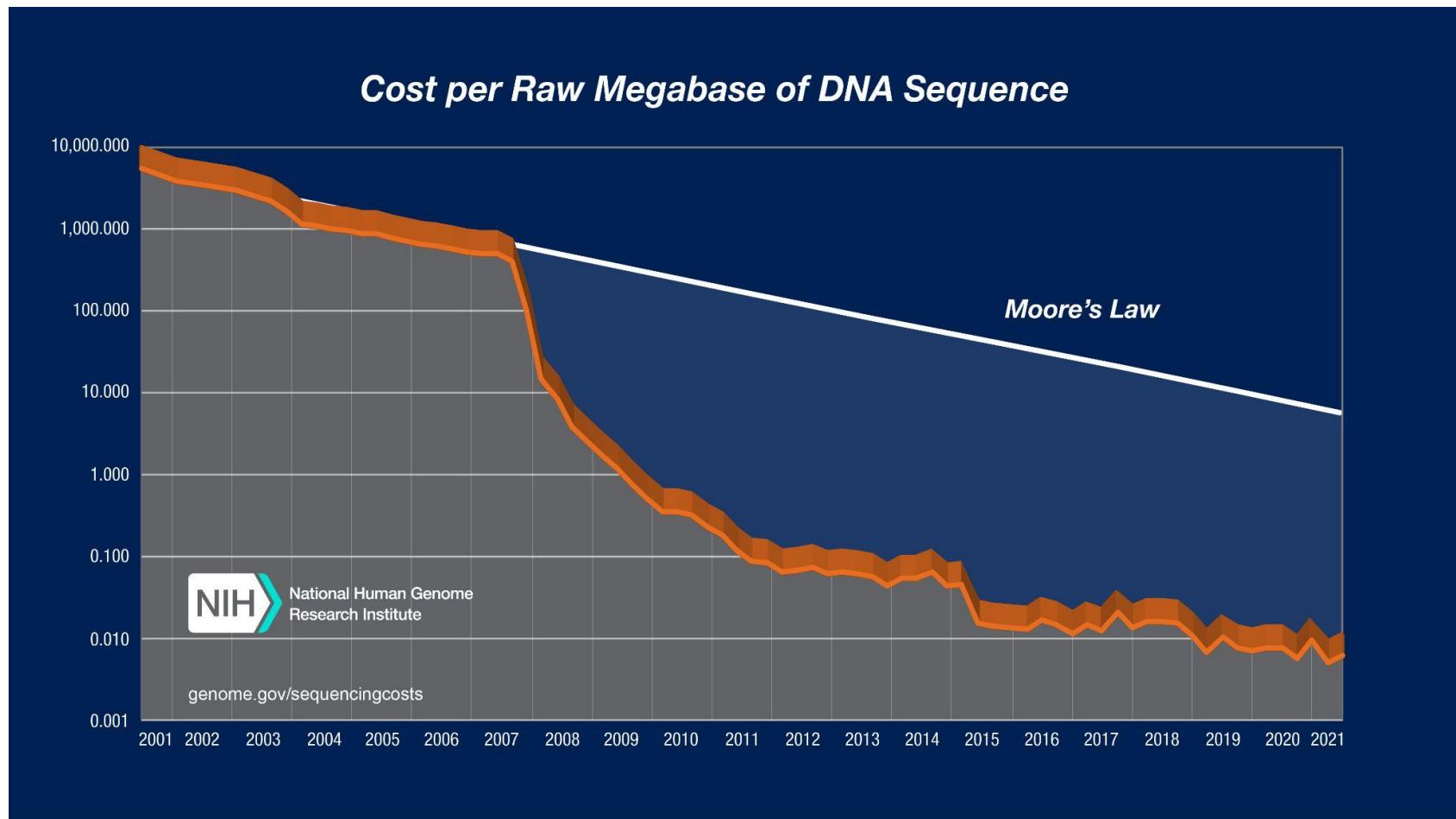
Prof. Dr. Boas Pucker and Katharina Wolff  
(Plant Biotechnology and Bioinformatics)

# Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - StudIP: Data Literacy in Genomics
  - GitHub: <https://github.com/bpucker/teaching>
- Questions: Feel free to ask at any time
- Feedback, comments, or questions: b.pucker[a]tu-bs.de

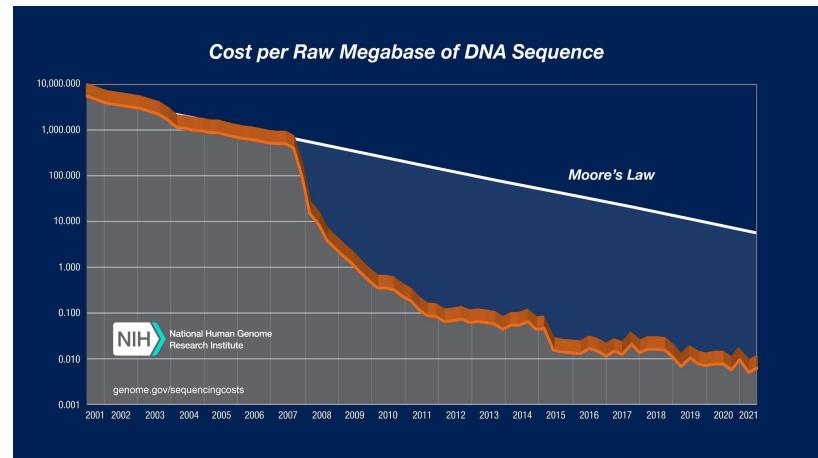
My figures and content can be re-used in accordance with CC BY 4.0, but this might not apply to all images/logos. Some figure were constructed using bioRender.com.

# 'Big Data'



# Reasons for growth of databases

- Data generation costs are dropping
  - Rapid sequencing technology development
  - Resolution of pictures is increasing
  - Robotics supports data acquisition
- Data storage capacities are increasing
- Potential of data reuse is recognized



# What is Data Literacy?

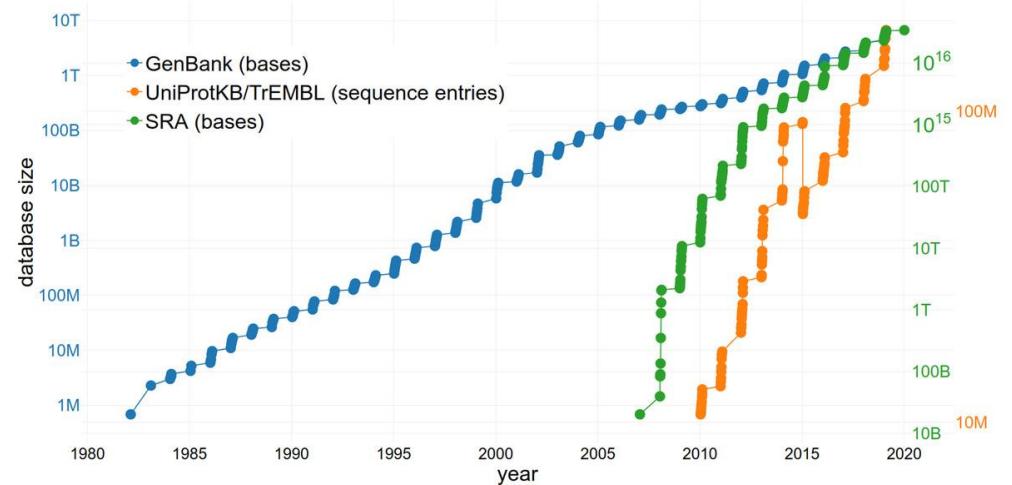


# Data Literacy

- ‘Competency to handle data properly’
- ‘Read, understand, create, communicate data as information’ (wikipedia)

# Why is Data Literacy important?

- We live in a world of data: important for science, business, and society
- Objective: data informed decision making
- Sizes and complexity of data sets are increasing ('BigData')
- Data is accumulating over time

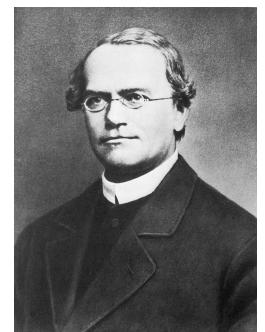


# Why is Data Literacy important for you?

- Data (interpretation) is valuable
- Data scientists are needed
- Combination of different fields is particularly powerful

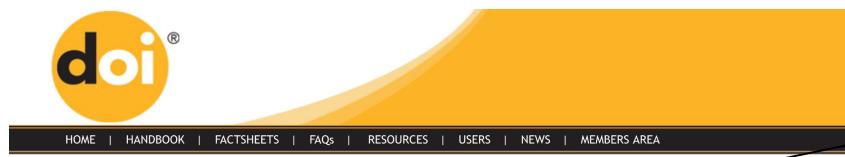
# More definitions

- Data: collection of observations
  - 2020: 312 red flowers; 95 white flowers
  - 2021: 298 red flowers; 98 white flowers
- Knowledge: accumulation of facts and data
  - Ratio of red flowers to white flowers is 3:1
- Insights: grasping the underlying nature of knowledge; understanding general concepts
  - Flower color has a genetic basis
  - Red flower allele dominant over white flower allele



# 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/>



## Resolve a DOI Name

doi: [10.1101/2022.02.16.480750](https://doi.org/10.1101/2022.02.16.480750)

Go

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<sup>®</sup>, DOI<sup>®</sup>, DOI.ORG<sup>®</sup>, and shortDOI<sup>®</sup> are trademarks of the International DOI Foundation.

A screenshot of a bioRxiv preprint page. At the top right is the bioRxiv logo with the text 'THE PREPRINT SERVER FOR BIOLOGY'. Below the logo is a yellow banner with the text '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.' and a link 'Follow this preprint'. The main title is 'Apiceace FNS I originated from F3H through tandem gene duplication'. Below the title is the author information: Boas Pucker, Massimo Iorizzo, and the DOI: <https://doi.org/10.1101/2022.02.16.480750>. There is a note that the article has not been certified by peer review. Below the author info are several small icons representing different metrics. At the bottom of the page are links for Abstract, Full Text, Info/History, Metrics, and Preview PDF.

### 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 I is widely distributed in seed plants, one lineage of soluble FNS I appeared to be unique to Apiceace species.

**Results** We show through phylogenetic and comparative genomic analyses that Apiceace 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 Apiceace in a common ancestor of *Daucus carota* and *Apium graveolens*. The results also support previous findings that FNS I in the Apiceace evolved independent of FNS I in other plant species.

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

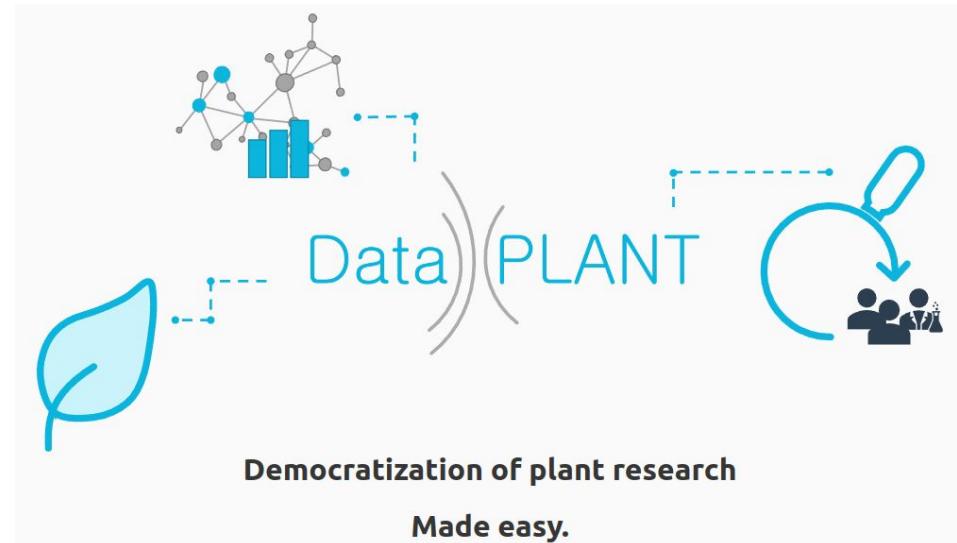
### Competing Interest Statement

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.

# Data management

- Larger data sets require more efficient data management
- Data management plans required in project proposals
- Many services and organizations emerging (NFDI)



<https://nfdi4plants.de/>

# Data protection and data security

- Data protection is becoming a huge issue in the EU
- Avoid any personal data in your research data sets
- Data security is gaining relevance
- German universities are frequently attacked:
  - Gießen (2019): offline for weeks + 1.7 million direct costs
  - HHU Düsseldorf (2020): ransom attack on clinic
  - Bochum, Dresden, Freiburg, Berlin

# Electronic Laboratory Notebooks (ELN)

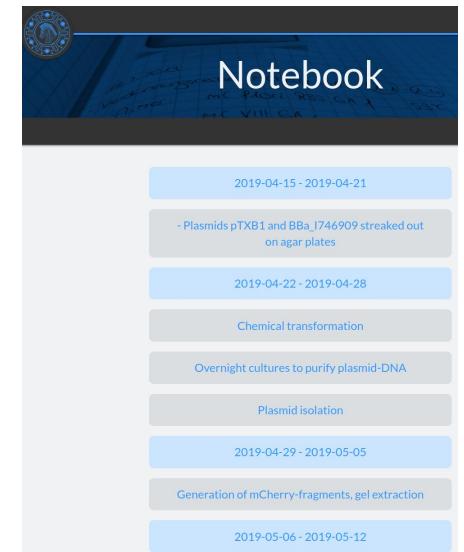
- Links between data sets and documentation
- Avoids repetitive documentation (copy&paste); version control
- Automatic search
- No issues with handwriting
- Quick and regular backups possible
- Accessible from everywhere
- Collaborative

# Technical solutions for electronic lab notebooks

- Simple wiki page
- Dedicated systems developed by research institutions
  - GABI-Kat LIMS
  - Chemotion
  - e!DAL
  - gitlab
- Commercial offers
  - Benchling
  - Dotmatic's
  - Signals Notebook (Perkin Elmer)
  - LabArchives

# Example: wiki

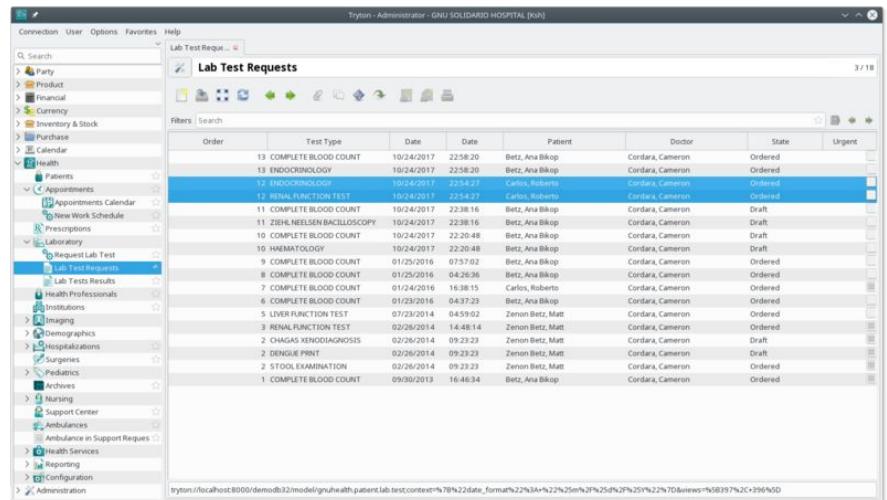
- Electronic lab notebook for collaborative learning
  - iGEM wiki to document team projects
  - Github wiki to document tools



<https://2019.igem.org/Team:Bielefeld-CeBiTec/Notebook>  
<https://github.com/bpucker/MybMonday/wiki>

# Laboratory Information Management System (LIMS)

- Workflows are represented and samples are tracked
- Information are linked between analysis
- Different levels of permissions/access
- Examples:
  - Handle a collection of T-DNA insertion lines (GABI-Kat)
  - Manage all samples submitted for sequencing
  - Manage all oligonucleotide orders



The screenshot shows the GNU LIMS software interface. The left sidebar contains a navigation tree with categories like Patients, Appointments, Laboratory, and Lab Test Requests. The 'Lab Test Requests' node is selected. The main window displays a table titled 'Lab Test Requests' with the following data:

Order	Test Type	Date	Date	Patient	Doctor	State	Urgent
13	COMPLETE BLOOD COUNT	10/24/2017	22:58:20	Betz, Ana Bikop	Cordara, Cameron	Ordered	
13	ENDOCRINOLOGY	10/24/2017	23:58:20	Betz, Ana Bikop	Cordara, Cameron	Ordered	
12	ENDOCRINOLOGY	10/24/2017	23:54:27	Carlos, Roberto	Cordara, Cameron	Ordered	
12	RENAL FUNCTION TEST	10/24/2017	23:54:27	Carlos, Roberto	Cordara, Cameron	Ordered	
11	COMPLETE BLOOD COUNT	10/24/2017	22:38:16	Betz, Ana Bikop	Cordara, Cameron	Draft	
11	ZIEHL-NEELSEN ZYGOLOGY	10/24/2017	22:38:16	Betz, Ana Bikop	Cordara, Cameron	Draft	
10	COMPLETE BLOOD COUNT	10/24/2017	22:20:48	Betz, Ana Bikop	Cordara, Cameron	Draft	
10	HEMATOLOGY	10/24/2017	22:20:48	Betz, Ana Bikop	Cordara, Cameron	Draft	
9	COMPLETE BLOOD COUNT	01/25/2016	07:57:02	Betz, Ana Bikop	Cordara, Cameron	Ordered	
8	COMPLETE BLOOD COUNT	01/25/2016	04:26:36	Betz, Ana Bikop	Cordara, Cameron	Ordered	
7	COMPLETE BLOOD COUNT	01/24/2016	16:38:15	Carlos, Roberto	Cordara, Cameron	Ordered	
6	COMPLETE BLOOD COUNT	01/23/2016	04:37:23	Betz, Ana Bikop	Cordara, Cameron	Ordered	
5	LIVER FUNCTION TEST	07/23/2014	04:59:03	Zenon, Bett, Matt	Cordara, Cameron	Ordered	
3	URINATION TEST	02/26/2014	09:23:14	Zenon, Bett, Matt	Cordara, Cameron	Ordered	
2	CHAGAS VENODIAGNOSIS	02/26/2014	09:23:23	Zenon, Bett, Matt	Cordara, Cameron	Draft	
2	DENGUE PRINT	02/26/2014	09:23:23	Zenon, Bett, Matt	Cordara, Cameron	Draft	
2	STOOL EXAMINATION	02/26/2014	09:23:23	Zenon, Bett, Matt	Cordara, Cameron	Ordered	
1	COMPLETE BLOOD COUNT	09/30/2013	16:46:34	Betz, Ana Bikop	Cordara, Cameron	Ordered	

[https://commons.wikimedia.org/wiki/File:Gnulims\\_lab\\_requests.png](https://commons.wikimedia.org/wiki/File:Gnulims_lab_requests.png)

# Example: Chemotion

- Research data management tool for chemists
- Electronic Laboratory Notebook (ELN)
- Repository for research data (easy transfer from ELN)
- DOIs are assigned to datasets and protocols

The image shows the Chemotion website homepage at the top, featuring the Chemotion logo and the text "Electronic Laboratory Notebook (ELN) & Repository for Research Data". Below the homepage, there is a video thumbnail titled "Chemotion ELN and repository" with a "Watch on YouTube" button. At the bottom, there are four main application modules: "Chemotion ELN" (showing a screenshot of a chromatogram), "Repository" (showing a screenshot of a database interface), "ChemSpectra" (showing a screenshot of a spectrum plot), and "ChemScanner" (showing a screenshot of a document viewer). Each module has a "Demo Version" button.

# Example: e!DAL

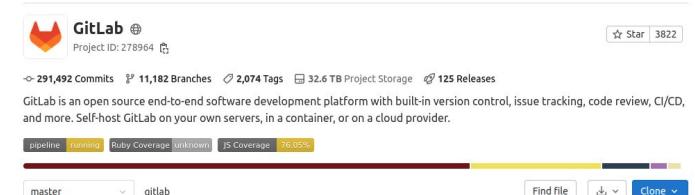
- Maintained at IPK; supported by de.NBI
- DOIs are assigned to submitted data sets
- Reporting about access statistics
- Version tracking
- FAIR compatible and data linkage



<https://edal.ipk-gatersleben.de/>

# Example: gitlab

- Free version control solution
- Can be used for software development, but also suitable as ELN
- GITZ offers a central service at TUBS:  
<https://doku.rz.tu-bs.de/doku.php?id=server:gitlab>
- Non-commercial alternative to github



The screenshot shows a GitLab project page for 'gitlab-org/gitlab'. Key statistics displayed include 291,492 Commits, 11,182 Branches, 2,074 Tags, 32.6 TB Project Storage, and 125 Releases. A summary bar indicates pipeline status (running), Ruby Coverage (unknown), JS Coverage (76.95%), and a file search bar with 'master' and 'gitlab' dropdowns.

<https://gitlab.com/gitlab-org/gitlab>



The screenshot shows the central GitLab instance for TU Braunschweig. It features a header with the TU Braunschweig logo and the text 'Gitlab'. Below the header, it states: 'Es existiert eine zentrale Gitlab Installation für die Nutzung durch alle TU Mitarbeiter und Studierenden, erreichbar unter der URL [git.rz.tu-bs.de](https://git.rz.tu-bs.de)'. It includes links to 'Zentrales Gitlab der TU Braunschweig' and 'Gitlab User Documentation'. The main area shows the 'Gitlab TU Braunschweig' interface with a sign-in form for 'SSO TU Braunschweig' and a 'Remember me' checkbox.

<https://doku.rz.tu-bs.de/doku.php?id=server:gitlab>

# Example: Benchling

- Cloud-based platform for biotechnology
- Templates for documentations
- Convenient to use
- Compatible with various platforms
- Suitable for certified processes



# Lab 4.0: digital lab for higher efficiency

- Connection of lab and analysis processes
- Data from different instruments are synchronized through a cloud
  - Values and pictures are directly inserted into the lab book
- Automation of processes (robotics)
  - Pipetting robot for large scale sample processing
  - Automatic phenotyping facilities
- Samples are labeled with barcodes
  - Barcodes and scanning avoid human errors
  - Can also prevent fraud

# Summary: data is everywhere

- Size of data sets is growing
- More sophisticated data collection methods
- Databases enable dissemination/reuse
- Electronic documentation
- Digitalization makes processes more efficient

# Course challenge



[add pictures and motivation of specific course challenge here]

# Previous knowledge?

- Are there previous reports of similar observations? (internet search)
- Are there publications about possible explanations? (literature search)
- Are there data sets that could be helpful for a study? (re-use)

[Add some screenshots of specific datasets here]

# Hypotheses

- Research should be driven by hypothesis (hypothesis testing)
- Hypothesis must be specific and falsifiable
- Proofing hypothesis right is not possible
- Rejecting a hypothesis if possible based on contradicting evidence
- Different hypothesis must not be mutually exclusive



# What is your hypothesis?



# Example: hypotheses

- Activation of anthocyanin biosynthesis gene *DFR* explains color appearance
- pH change of the soil/water determines the color
- Transcriptional activators of the anthocyanin biosynthesis are triggered
- Competing branches of the flavonoid biosynthesis are inactivated
- Intracellular transport of anthocyanins is changing

# What do we need to test a hypothesis?

- Example: activation of anthocyanin biosynthesis gene *DFR* explains color emergence
- Genome sequence (long read sequencing & assembly)
- Structural annotation of the genome sequence (gene prediction)
- Functional annotation of the genes (function prediction)
- Analysis of gene activity (RNA-seq, data re-use)

# Disseminate findings and data

- Submit data sets to appropriate repository
- Share documentation and scripts developed for analyses
- Share findings through talks, posters, or publications

# Data management tipps



# Check the literature

- Was the research question already answered?
- Have others proposed hypotheses to explain the observed phenomenon?
- Reading numerous publications can avoid a lot of unnecessary work

# How to find publications?



# How to find publications?

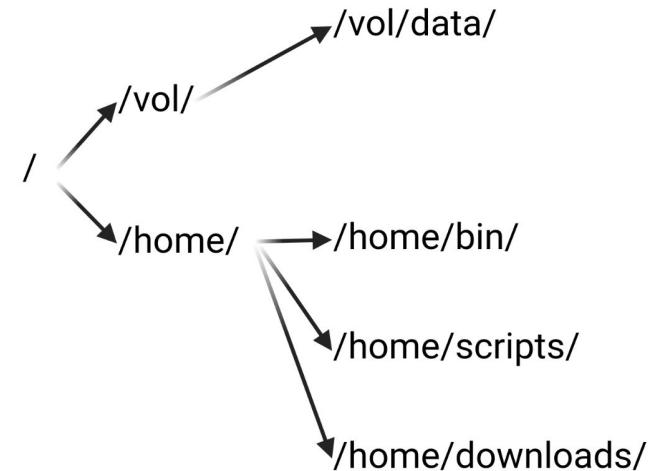
- Where can you search?
  - GoogleScholar, PubMed, SciFinder, PubPharm, WebOfScience
  - Various AI-supported search engines (e.g. Bing)
    - Warning: some AIs make up references that do not exist
- What keywords should you use?
  - Try synonyms to find the specific terms people in the field use
  - Try different keyword combinations and orders
- Which publications should you read?
  - Review articles could be a good start
  - Recent & highly cited articles seem promising
  - Walk through network of citations

# How to structure your data?

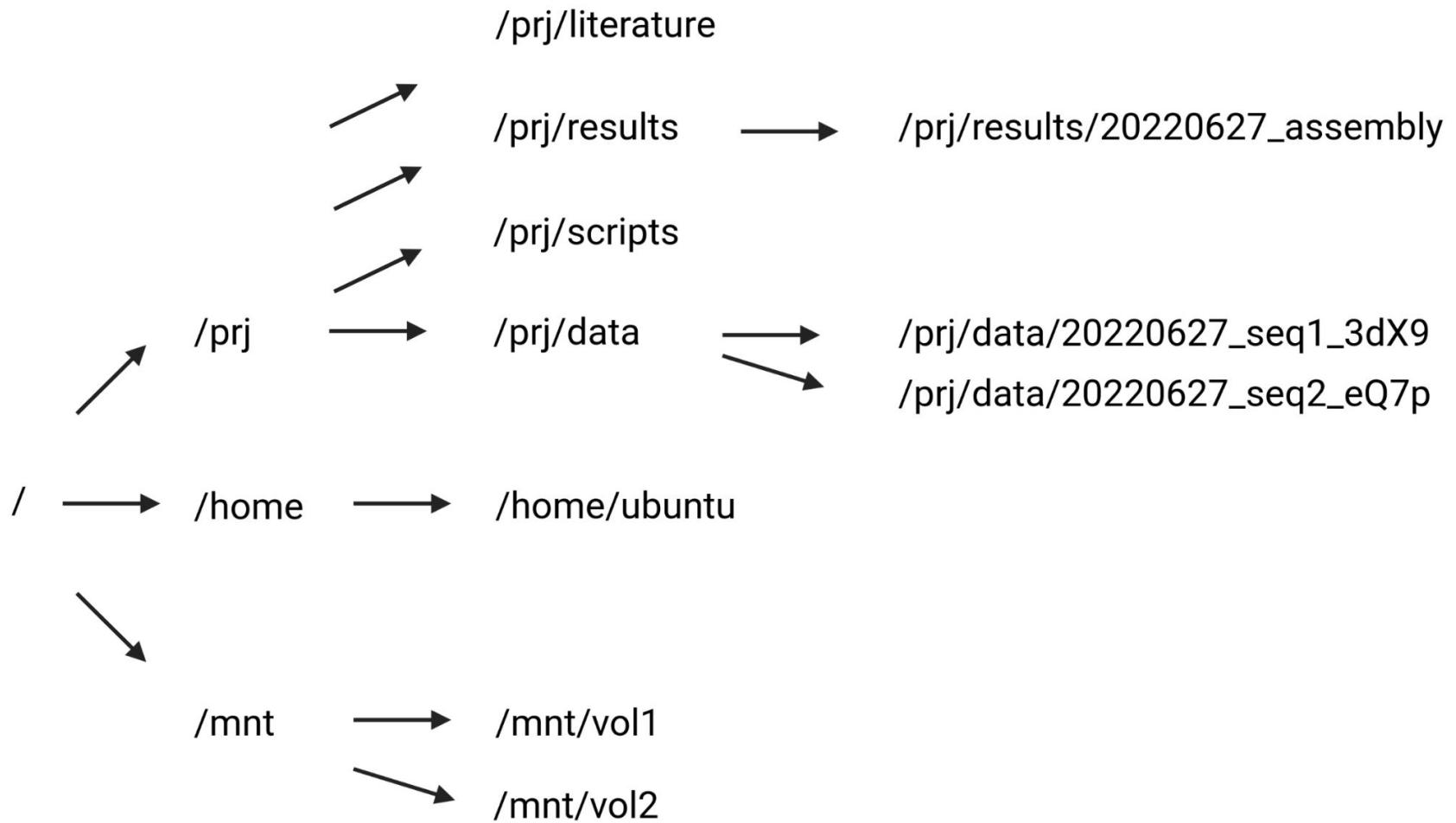
- Document every step (e.g. in a README)
  - Origin of data sets; versions of tools, parameters of analyses
- Keep raw data sets separated from scripts and results
- Sort data by project
- Structure analysis related data sets/results by date

# Linux file system structure

- Linux (Ubuntu) is operation system of choice for bioinformatics
- Hierarchical structure of directory ('/' is basis)
- Separation of tool installation and data sets



# Example: Linux file system



# What do you know about file names?



# Naming files

- File names should be informative
- Never use ‘new’ or ‘final’; use version numbers instead
- Use date as file name prefix (e.g. 2022-06-27 or 20220627)
- Never use spaces in file or folder names (underscore or minus as replacement)
- Include your initials as suffix in collaborative projects

# Documentation

- Document as much as possible
- Others must be able to repeat your experiments/analyses
- Document dates of data acquisition and processing
- Document all steps of data processing
- Document versions and parameters of applied tools

# de.NBI cloud

- Virtual machine (VM) for data analysis
- Accounts are required for access (ORCID or TUBS SSO for login)
- Addition to project required for access
- User create pair of private and public keys for authentication



de.NBI ELIXIR-DE Services Training de.NBI Cloud News



## Compute Power for Your Project

In life sciences today, the handling, analysis and storage of enormous amounts of data is a challenging issue. For example, new sequencing and imaging technologies result in the generation of large scale genomic and image data. Hence, an appropriate IT infrastructure is crucial to perform analyses with such large datasets and to ensure secure data access and storage. In addition, it is difficult to directly compare result data that have been processed at different sites, due to a lack in standardization of workflows. The de.NBI cloud is an excellent solution to enable integrative analyses for the entire life sciences community in Germany and the efficient use of data in research and application.

To a large extent, de.NBI will close the gap of the missing computational resources for researchers in Germany. A federated de.NBI Cloud concept and infrastructure leads to the reduction in overall infrastructure and operational costs.

[Click here to enter the de.NBI Cloud Portal](#)

Cloud Access Portal
de.NBI Cloud Flyer
de.NBI Cloud Poster
Cloud Training Courses

### Get access to the Cloud

1.) Register for an [ELIXIR Account](#) and apply for membership in the de.NBI virtual organisation.

2.) Log in to our [de.NBI Cloud portal](#) to manage your projects and project members.

# Transferring files

- Filezilla: graphical user interface for file transfer protocols
  - <https://filezilla-project.org/>
- Scp (secure copy): command line file transfer method
- Wget: command line file transfer method
  - [https://ftp.gnu.org/old-gnu/Manuals/wget-1.8.1/html\\_mono/wget.html](https://ftp.gnu.org/old-gnu/Manuals/wget-1.8.1/html_mono/wget.html)
- Rsync: sophisticated file transfer method that avoids redundant transfers
  - <https://wiki.ubuntuusers.de/rsync/>

# Permissions

- Files can have different permissions:
  - Read (r)
  - Write (w)
  - Execute (x)
- Users have full permissions to edit their own files
- Downloaded files are usually not executable without adjustment
- `chmod XXX <FILE_NAME> ...` can be run to change file permissions

# Linux introduction (1)

Connection to virtual machine (VM):

```
$ ssh -i /path/to/private_key ubuntu@123.133.7.49 -p 1234  
(base) ubuntu@agilezuse-10552:~$
```

'\$' is used to indicate that following text needs to go into terminal

# indicates comment (should not be transferred into terminal)

Frequent issues:

- 1) Path to private key file not correct
- 2) Private key file too public

# Linux introduction (2)

- Moving through the folder structure:
  - `$ cd /full/path/to/folder` #change into specific folder
  - `$ cd subfolder` #change into subfolder
  - `$ cd ..` #change into parent directory
- Checking content of a folder:
  - `$ ls` #shows content of current folder
  - `$ ls -lh` #shows more details
  - `-l` triggers display of additional details
  - `-h` human readable
  - `-a` show also hidden files

# How to install bioinformatics tools

- Install tools from package management (apt-get install)

Installing mafft package on Ubuntu is as easy as running the following command on terminal:

```
sudo apt-get update  
sudo apt-get install mafft
```

- Download a precompiled binary

## Releases

The kallisto GitHub repository is [here](#).

Version	Date	Mac	Linux	Windows	Rock64	Source
Release notes: v0.46.1	October 04, 2019					
Release notes: v0.46.0	June 12, 2019					
Release notes: v0.45.0	November 17, 2018					
Release notes: v0.44.0	January 29, 2018					

- Compile from source

## COMPILING FROM SOURCE

Download the latest [release from](#) and uncompress it

```
# Get latest STAR source from releases  
wget https://github.com/alexdobin/STAR/archive/2.7.10b.tar.gz  
tar -xzf 2.7.10b.tar.gz  
cd STAR-2.7.10b
```

# Alternatively, get STAR source using git  
git clone https://github.com/alexdobin/STAR.git

## Compile under Linux

```
# Compile  
cd STAR/source  
make STAR
```

For processors that do not support AVX extensions, specify the target SIMD architecture, e.g.

```
make STAR CXXFLAGS SIMD=sse
```

- Install via conda

## Installers



conda install [?](#)

To install this package run one of the following:

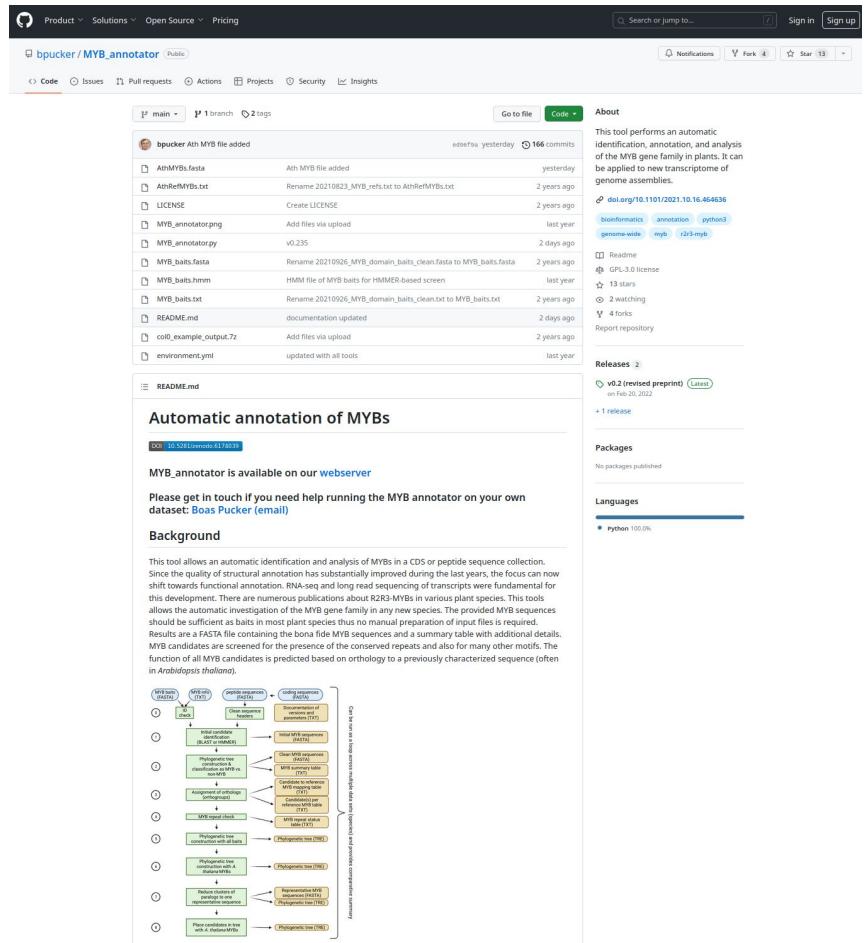
```
conda install -c bioconda shasta
```

<https://howtoinstall.co/en/mafft>  
<http://pederlelab.github.io/kallisto/download>  
<https://github.com/alexdobin/STAR>  
<https://anaconda.org/bioconda/shasta>



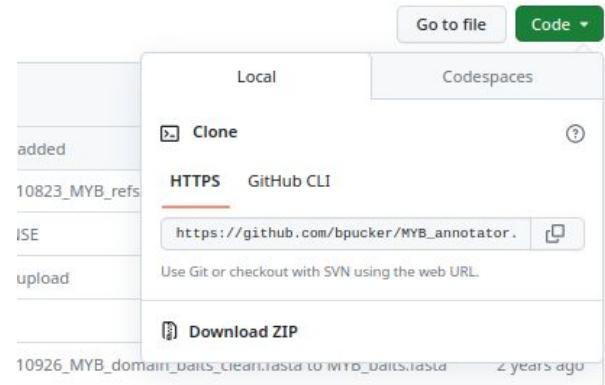
# Retrieving tools from GitHub

- Find GitHub repository corresponding to tool of interest
  - Read the installation instructions
  - Download single file via “Raw”, “Save as”
  - Download everything via “Code” as ZIP archive
  - Use git clone to get the content



# Cloning a GitHub repository

1. Go to GitHub repository website
2. Click “code”
3. Copy URL (HTTPS)
4. Open a terminal
5. Change working directory to the location where you like to have the tool
6. Clone the repository: `git clone <URL>`



# Virtual environments

- Python tool (venv) for dependency management and project isolation
- Conda keeps packages separated and works without administrator rights
- Increase reproducibility of research
- Configure virtual environment:
  - `python3 -m pip install --user --upgrade pip`
  - `python3 -m pip --version`
  - `python3 -m pip install --user virtualenv`
  - `python3 -m venv env`
  - `source env/bin/activate`
  - [here you can do things inside the virtual environment]
  - `deactivate` #to leave the virtual environment again

# Running Python scripts

- Run script or open argument infos:
  - \$ python <SCRIPT\_NAME>
- Run script with arguments:
  - \$ python <SCRIPT\_NAME> --argument1\_name <ARGUMENT1>  
--argument2\_name= <ARGUMENT2>
- Scripts show help message if started with insufficient arguments

```
python3 ./KIPES3.py --baits ./flavonoid_baits/ --positions  
.flavonoid_residues/ --out ./ --subject ./croton_red.fasta --seqtype pep --scoreratio 0.3 --simcut  
40.0 --minsim 0.4 --minres 0.0 --minreg 0.0 --possibilities 3 --cpus 1
```

# Running other tools

- Show help message:
  - `$ <NAME_OF_TOOL>`
  - `$ <NAME_OF_TOOL> -h`
  - `$ <NAME_OF_TOOL> -- help`
- Providing arguments is different for each tool:
- Most tools show help message if provided with insufficient/wrong arguments

# Giving presentations



# Presentation structure

- Title slide with picture
- Problem/challenge
- Results
- Summary & conclusions
- Future directions
- Acknowledgements



Evolutionary blocks to anthocyanin accumulation  
in betalain-pigmented Caryophyllales

Boas Pucker

# Story telling

- Introduction should be easy to follow for non-experts
- Clearly present motivation
- Questions can catch attention and provide structure
- Connect different elements of the presentation

Why are there no anthocyanins in  
betalain-pigmented species?



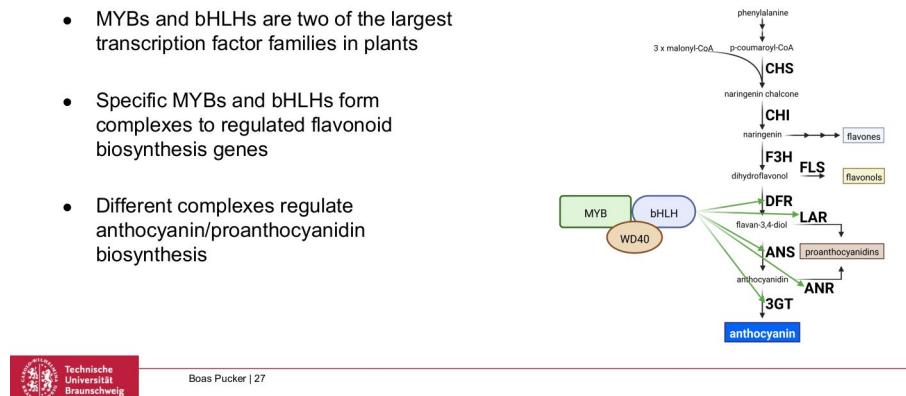
Boas Pucker | 7

# Clear slides

- Reduce text
- Replace text with figures
- One clear message per slide
- Large font size

## Contribution of transcription factors

- MYBs and bHLHs are two of the largest transcription factor families in plants
- Specific MYBs and bHLHs form complexes to regulate flavonoid biosynthesis genes
- Different complexes regulate anthocyanin/proanthocyanidin biosynthesis



# Tools for presentation creation

- PowerPoint
- GoogleSlides (“free”)
- LaTeX (e.g. overleaf)
- Templafy
- Prezi



An open-source online real-time collaborative LaTeX editor.

[Wiki](#) • [Server Pro](#) • [Contributing](#) • [Mailing List](#) • [Authors](#) • [License](#)



Figure 1: A screenshot of a project being edited in Overleaf Community Edition.

<https://github.com/overleaf/overleaf>

# Tools for figure generation

- PowerPoint/GoogleSlides
- bioRender
- InkScape

# Social media/networks

# Why are scientists using social media?

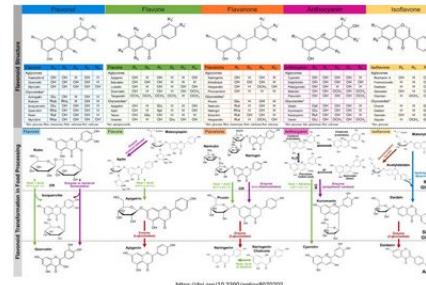
- Scientists share publications, open positions, and other news
- Important platform for science communication
- Platform for scientific discussions
- Learn about latest publications and trends

# SciComm posts

- Statements need to be scientifically correct
- Explanations need to be simple
- Connect information to emotion (cute pictures)
- Engage in discussions

## (42) Biochemical diversity of flavonoids

**TWEET:** Biochemical diversity of flavonoids is based on hydroxylation, glycosylation, methylation, acylation, and many other modifications #FlavonoidFriday (details: <https://bit.ly/3dKUQnN>)



Hydroxylation patterns and other modifications of the aglycon distinguish between different types of flavonoids. Aglycons can be modified through the successive addition of sugar moieties. The 3 and 5 position of the C ring are most frequently modified. Different sugar groups and can be attached by a large set of different glycosyltransferases as recently reviewed by [Tohge et al., 2017](#).

<https://github.com/bpucker/FlavonoidFriday>

# Twitter

- Important platform for science communication
- Scientists share publications, open positions, and other news
- Length of messages is limited (280 characters)
- Links will trigger generation of preview
- Pictures and videos can boost views
- Hashtags can be used as key words (avoid excessive use)

**Boas Pucker**  
@boas\_pucker

Prof. @tuBraunschweig | Plant Biotechnology & Bioinformatics (@PuckerLab) | @iGEM.TULBS | 🌱 #plant #genomics #evolution #Python #DataScience | my views  
Wissenschaft und Technik 🌱 Braunschweig, Deutschland  
tu-braunschweig.de/en/ipp/pbb  
Seit Juli 2017 bei Twitter

2.756 Folge ich 3.987 Follower

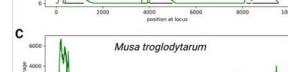
Tweets Antworten Medien Gefällt mir

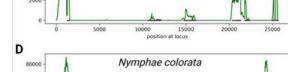
Angehörteter Tweet

Boas Pucker @boas\_pucker - 22. März  
Annotation is THE big challenge in #Genomics. Happy to share our latest preprint about some annotation issues and how to find them:  
  
Identification of annotation artifacts concerning the CHALCONE SYNTHASE (CHS)  
doi.org/10.1101/2023.0...  
  
#Bioinformatics #DataScience

A   
Macadamia integrifolia

B   
Musa balbisiana

C   
Musa troglodytarum

D   
Nymphaea colorata

ALT

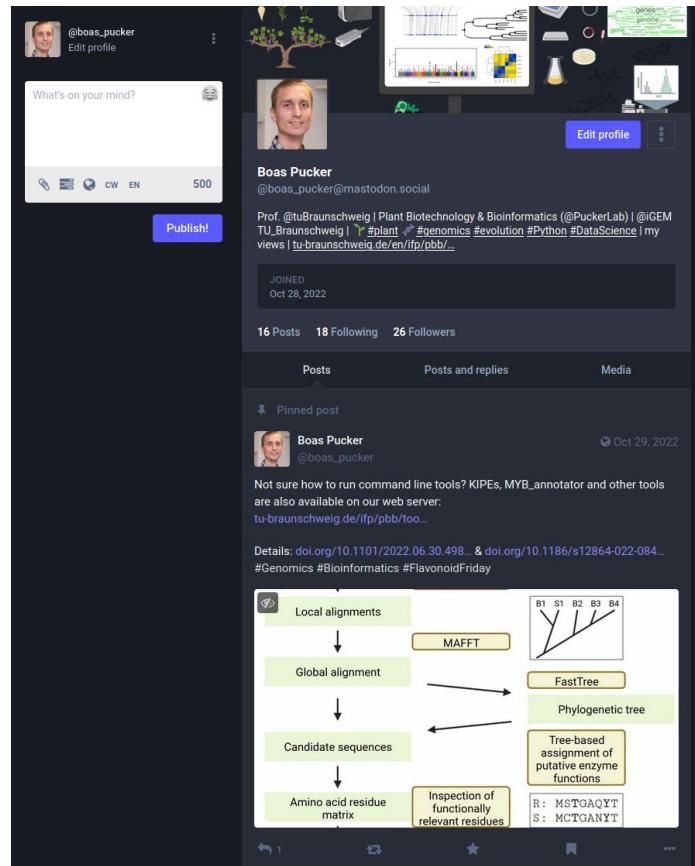
Plant Biotechnology and Bioinformatics und 2 weitere Personen

Sponsoren

3 25 95 18.982

# Mastodon

- Similar to Twitter and presented as replacement
- User own their data
- Running on multiple different servers
- Messages can be 500 characters



# Instagram

- Excellent for sharing pictures and videos
- Communicating science to public
- Dedicated to young audience



# ResearchGate

- Sharing full texts of publications (not always permitted by publisher)
- Promoting publications
- Presenting research interests
- Advertising open positions
- Option for questions and discussions

The screenshot shows Boas Pucker's profile on ResearchGate. At the top, there is a circular profile picture with a camera icon, followed by the name "Boas Pucker" and "Edit". Below this, it says "Dr. rer. nat. - Professor (Assistant) at Technische Universität Braunschweig, Germany | Website". A bio section states: "Unlocking the specialized metabolism of plants with genomics, transcriptomics, and bioinformatics". Below the profile picture, there are tabs for "Profile", "Research (137)", "Stats", "Following", and "Saved list".  
  
The "Business card" section contains a photo of Boas Pucker, his title "Boas Pucker", his affiliation "Dr. rer. nat. - Professor (Assistant)", and his bio "Unlocking the specialized metabolism of plants with genomics, transcriptomics, and bioinformatics". It also lists his institution "Technische Universität Braunschweig - Institute of plant biology" and his skills "Beta vulgaris · Plant Genomics · DNA Marker Technology + 27 others".  
  
The "About me" section includes an "Edit" button. It features a "Introduction" section with a photo, mentioning "Evolution of specialised metabolism in plants, Flavonoid and betalain biosynthesis, R2R3 MYB transcription factors in plants, Long read sequencing of plant genomes, Bioinformatics with focus on genomics and transcriptomics; Molecular Biology; Synthetic Biology (iGEM)". It also lists "Disciplines" like "Evolutionary Biology · Agricultural Plant Science · Bioinformatics · Genetics · Biochemistry" and "Skills and expertise" such as "Beta vulgaris · Plant Genomics · DNA Marker Technology · Plant Physiology · Plant Breeding · Plant Genetics · Plant Molecular Biology · Plant DNA Extraction · PCR · Gel Electrophoresis · Genetic Engineering · Metabolic Engineering · Cell Culture Techniques · Python · Synthetic Biology · Bacterial Cloning · Arabidopsis thaliana · Bioinformatics · R Programming · Bioinformatics and Computational Biology · Agrobacterium Mediated Plant Transformation · iGEM · Plant Biology · DNA Isolation · Genome Assembly · Arabidopsis · Bioinformatic Tools · RNA-Seq · Copy Number Variation · Sequencing".  
  
The "Languages" section shows "German · English". The "Contact information" section includes a "Twitter" link. The "Activity on ResearchGate" section shows "137 Research Items · 0 Questions · 155 Answers".  
  
At the bottom, there is a "Research Spotlight" section with a "Create a Spotlight" button and a note: "Showcase your recent work in a Spotlight to get 4x more reads on average. Learn more".

# Loop (Frontiers)

- Social network platform operated by Frontiers
- Handles review process at Frontiers journals
- Journals try to catch researchers and to collect information?



Boas Pucker

Doctorate  
Assistant Professor  
Technical University of Braunschweig  
Braunschweig, Germany

[Overview](#) [Bio](#) [37 Network](#) [82 Publications](#) [9 Editorial Contributions](#) [Impact](#)

**Brief Bio**

Genome Research, Bioinformatics, Synthetic Biology, Evolution

[View Full Bio and Expertise](#)

**82 Publications**

**Apiaceae FNS I originated from F3H through tandem gene duplication**  
Boas Pucker · Massimo Iorizzo  
 PLOS ONE  
Published on 19 Jan 2023

**The evidence for anthocyanins in the betalain-pigmented genus *Hylocereus* is weak**  
Boas Pucker · Samuel F. Brockington  
 BMC Genomics  
Published on 09 Nov 2022

**Evolutionary blocks to anthocyanin accumulation and the loss of an anthocyanin carrier protein in betalain-pigmented Caryophyllales**  
Boas Pucker · Nathanael Walker-Hale · Won C. Yim · John Cushman · Alexandra Crum · Ya Yang · Samuel Brockington  
 bioRxiv  
Published on 21 Oct 2022

**9 Editorial Contributions**

2 Edited Research Topics  
 2 Edited Publications  
 5 Reviewed Publications

[View Editorial Contributions](#)

**Editorial Roles**

You do not have an active role on a Frontiers editorial board. You can apply to join or refer a colleague to one of our editorial boards [here](#).

Frontiers Topic Editor

Genomics and Gene Editing of Orphan Plants  
[Open for Submissions](#)

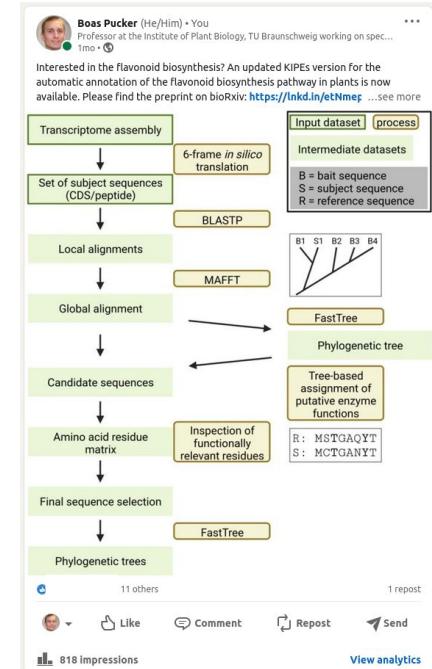
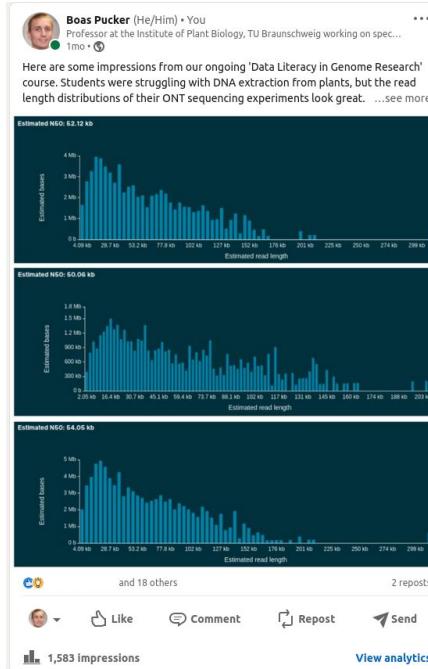
Development and Regulation of Flower Color in Ornamental Plants  
[Open for Submissions](#)

[View All Topics](#)

**21 Followers** **27 Following**

# LinkedIN

- Important for exchange with industry
  - Students can leave academia
  - Scientists can find funding options
- Excellent platform to learn about job opportunities
- Scientists share publications and other news

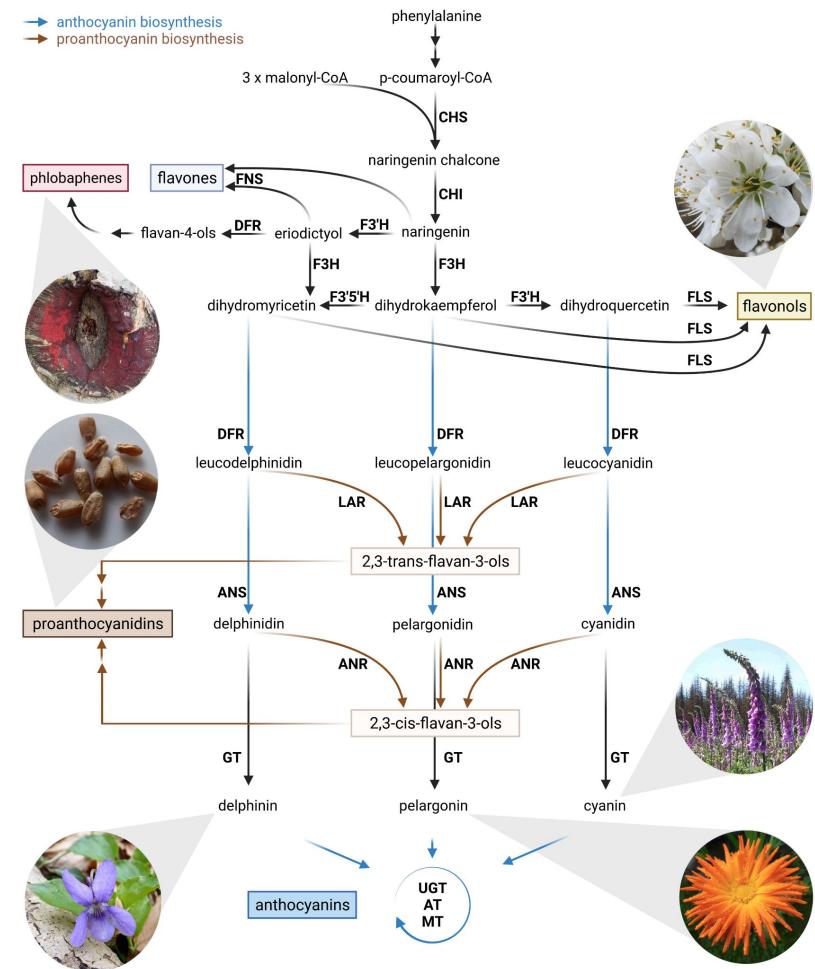


# Flavonoid biosynthesis



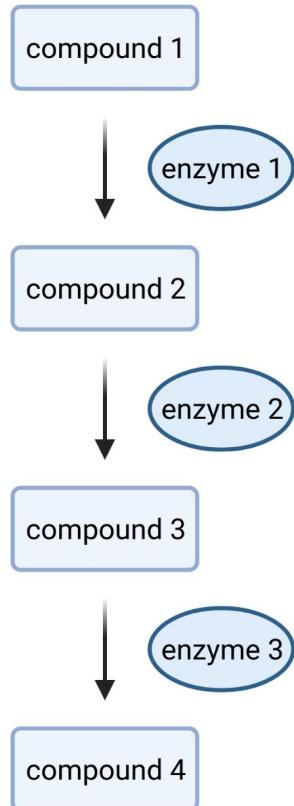
# Flavonoid biosynthesis

- Phenylalanine is common substrate
- Flavones, flavonols, phlobaphenes, proanthocyanidins, anthocyanins are the major groups
- enzymes catalyze different reactions: flavonoid 3'-hydroxylase (F3'H), flavonol synthase (FLS), dihydroflavonol 4-reductase (DFR), ...

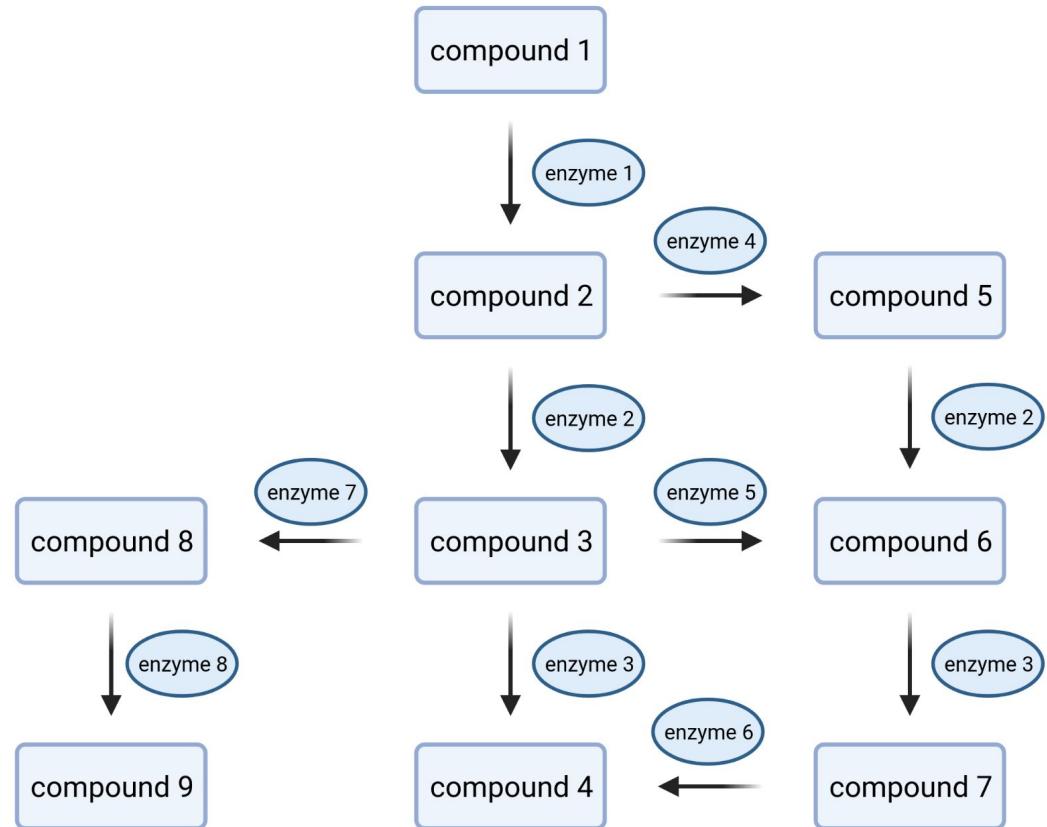


# Pathways are networks

Text book

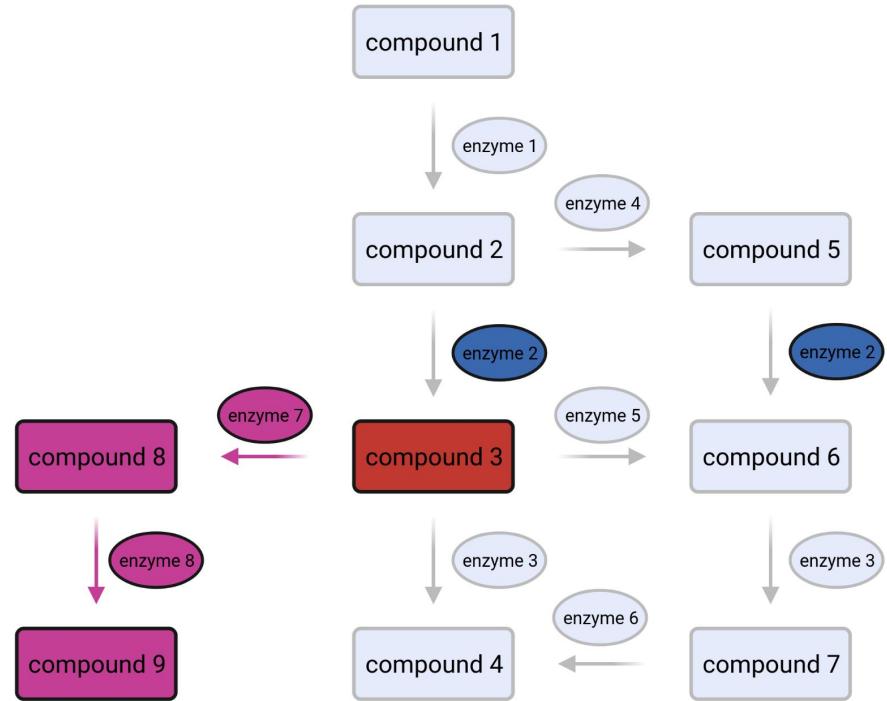


Reality



# Metabolic networks

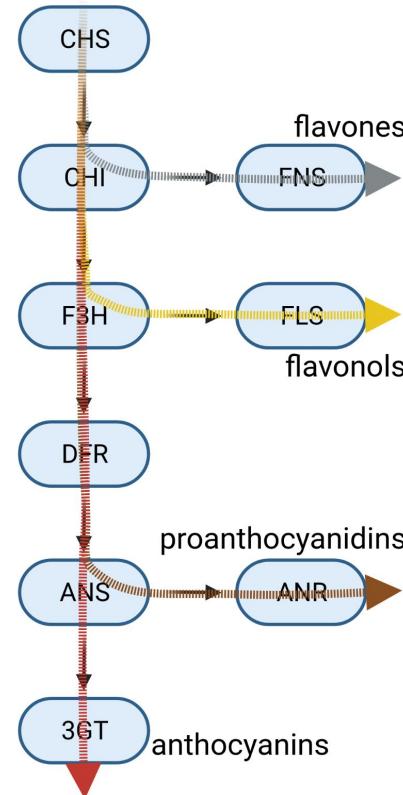
- Metabolic networks are characterized by hubs (central intermediates)
- Many enzymes show promiscuity i.e. catalyze different reactions
- Branches could be considered linear pathways



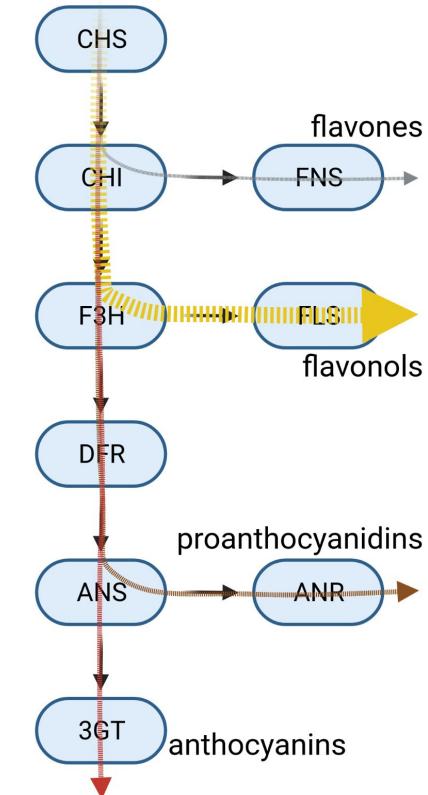
# Competing branches

- Different branches of a pathway can compete for substrate
- Activity of different branches can vary between conditions/tissues
- Metabolic modeling can be applied to optimize flux through pathway

Equal flux to all branches



Realistic flux of metabolites

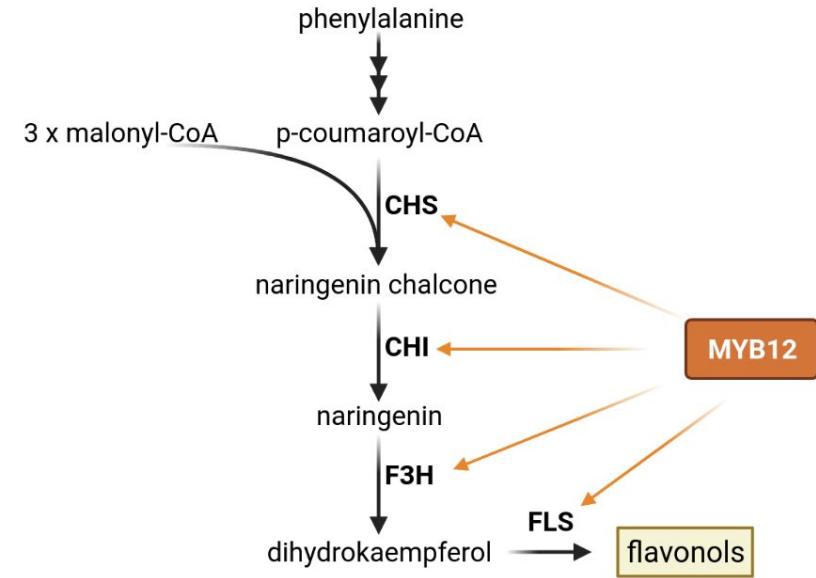


# Model system - pathway

- Easily detectable product
- Stable intermediates that can be quantified
- Gene knock-outs must not be lethal
- Conserved across many plant species
- Community working on the pathway
- Different branches
- ...

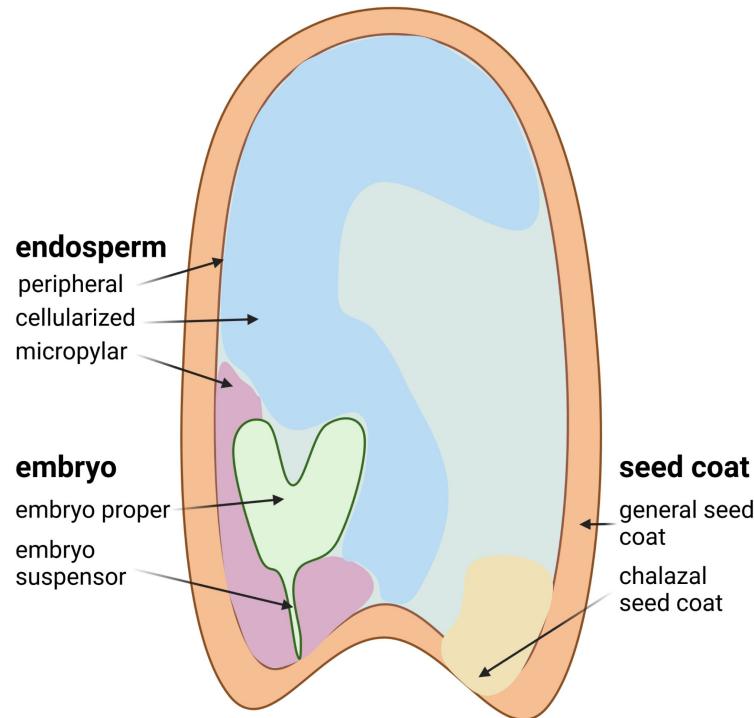
# Flavonols - Coexpression of pathway genes

- Shared transcription factor can explain similar expression patterns
- Example: MYB12 controls the flavonol biosynthesis through activation of *CHS*, *CHI*, *F3H*, and *FLS*
- Expectation: flavonol biosynthesis genes *CHS*, *CHI*, *F3H*, and *FLS* should show a similar expression pattern



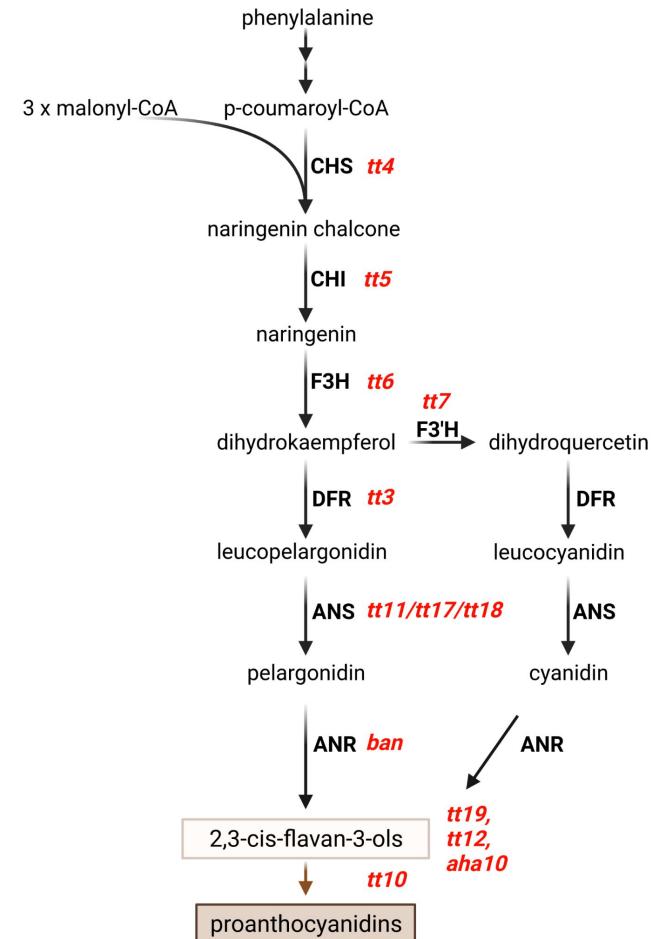
# Proanthocyanidins - Transparent testa

- Brown proanthocyanidins are located in the testa
- Transparent testa = lack of proanthocyanidins (no pigmentation)



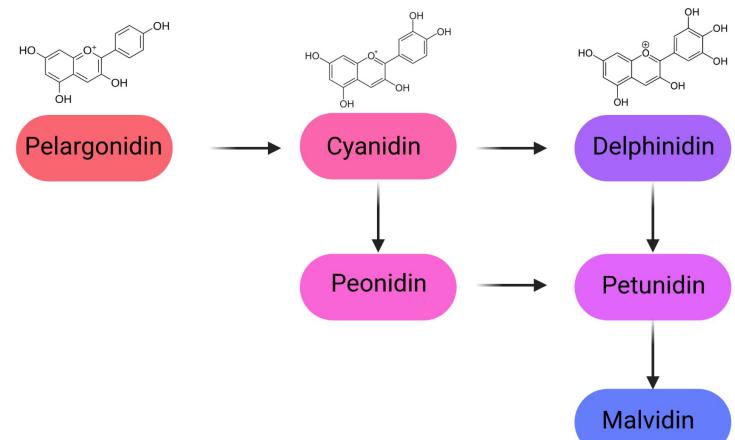
# ***Transparent testa* mutants**

- Large collection of different tt mutants
- Numbering is based on order of discovery
- Checking if new *transparent testa* (tt) mutants are allelic with existing ones



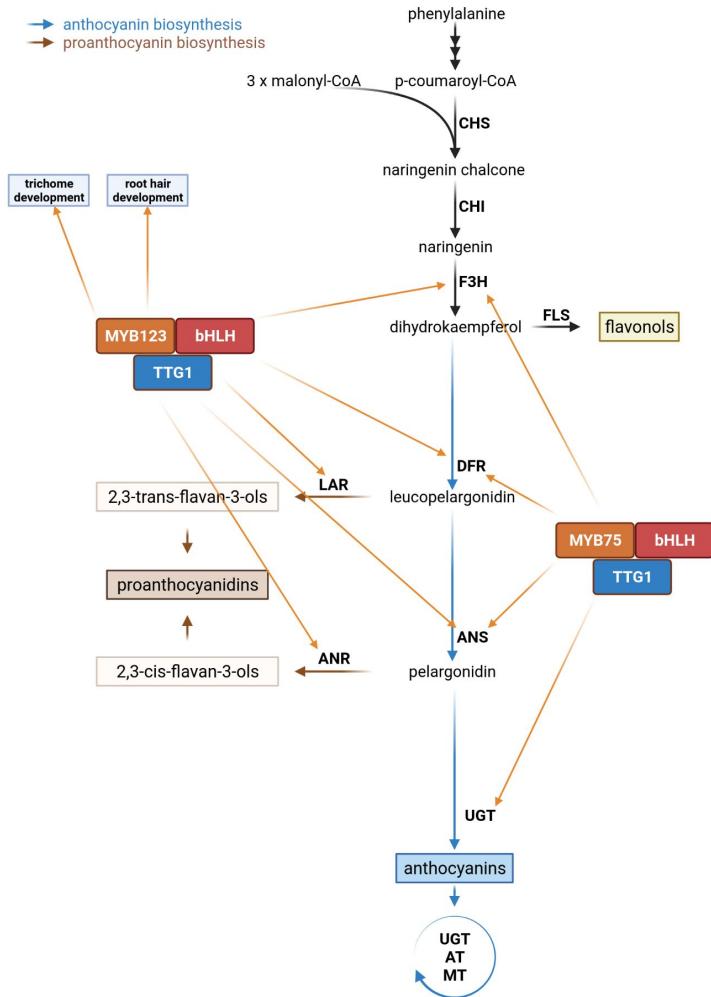
# Anthocyanin loss

- Anthocyanins are responsible for red to blue pigmentation of flowers
- Anthocyanins are derivatives from phenylalanine through the flavonoid biosynthesis
- *Digitalis purpurea* is an ornamental plant with biomedical potential
- Different varieties of *D. purpurea* show pigmentation differences



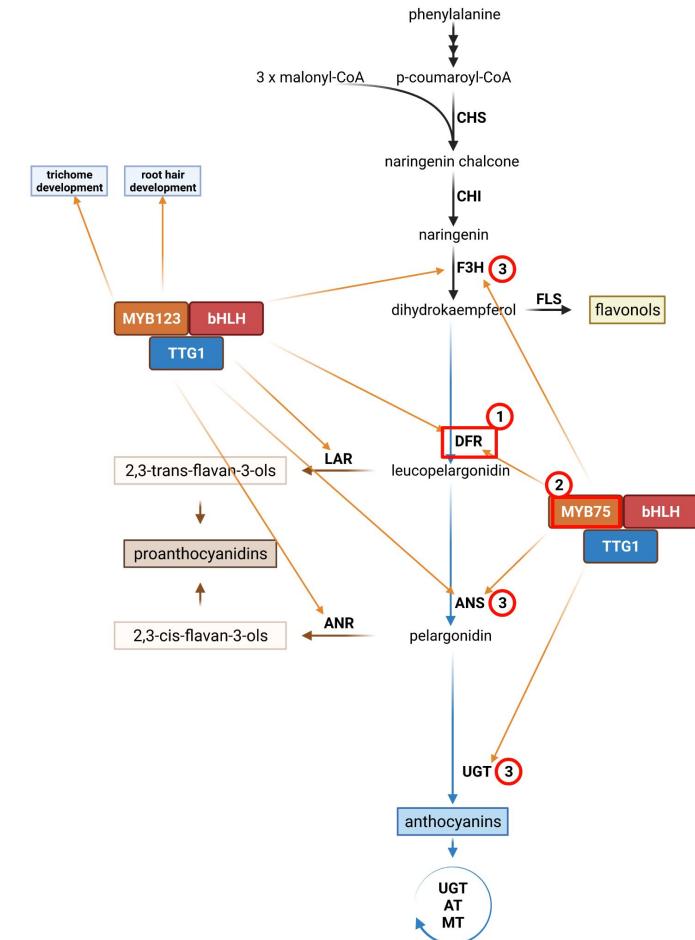
# MBW complex: MYB+bHLH+WD40

- Genes can be regulated by multiple TFs (e.g. *DFR* by MYB123 and MYB75)
- TFs can control different processes (e.g. proanthocyanidins, trichome development, root hair development)
- Co-expression of TFs and structural genes in pathways is not perfect



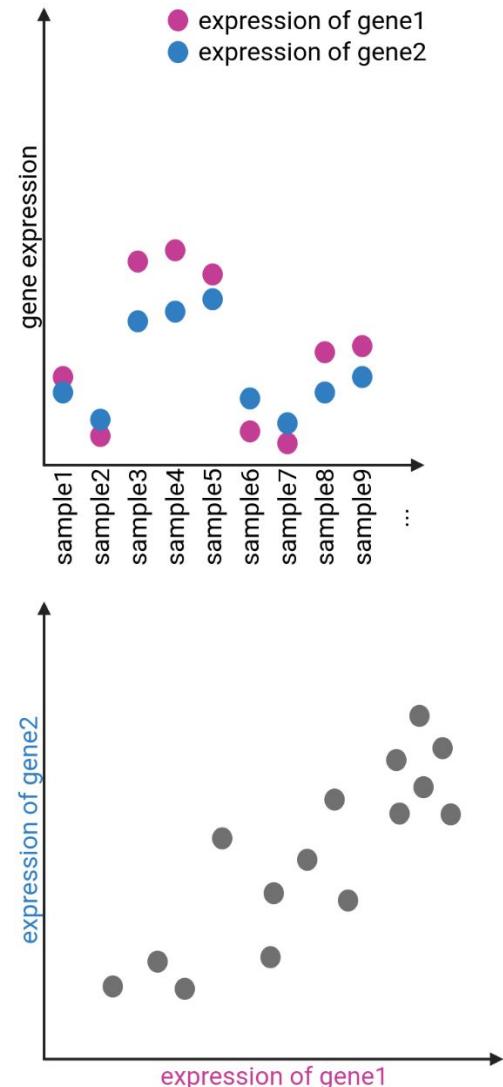
# Revealing pathways through coexpression analysis

- Bait genes are previously characterized genes with a function of interest e.g. encode an enzyme in the same biosynthesis pathway
- Shared transcription factors of a pathway can be helpful to identify all structural genes of a pathway
- Knowledge from other species can be applied in this step (details in later section)



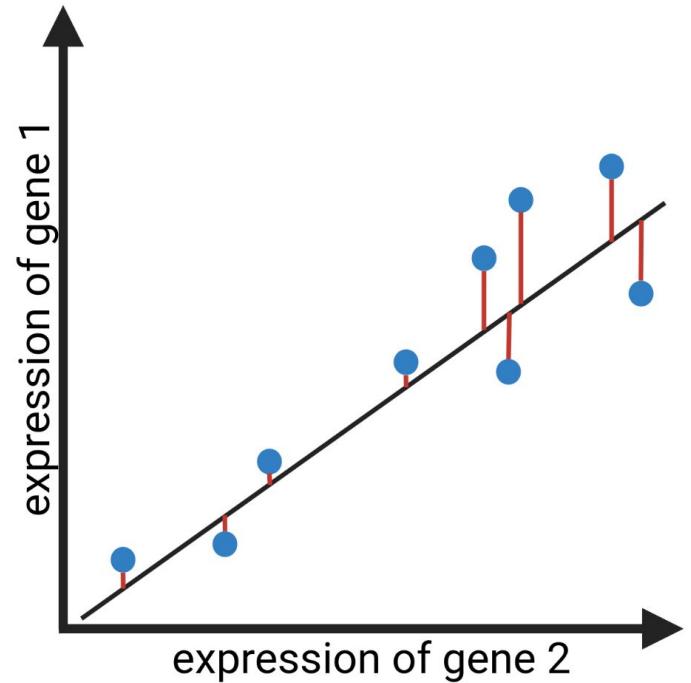
# Concept of coexpression

- Genes can show similar expression values across numerous samples
- Reality usually results in similar, but not identical patterns
- Different samples could be different plant parts of plants cultivated under different conditions



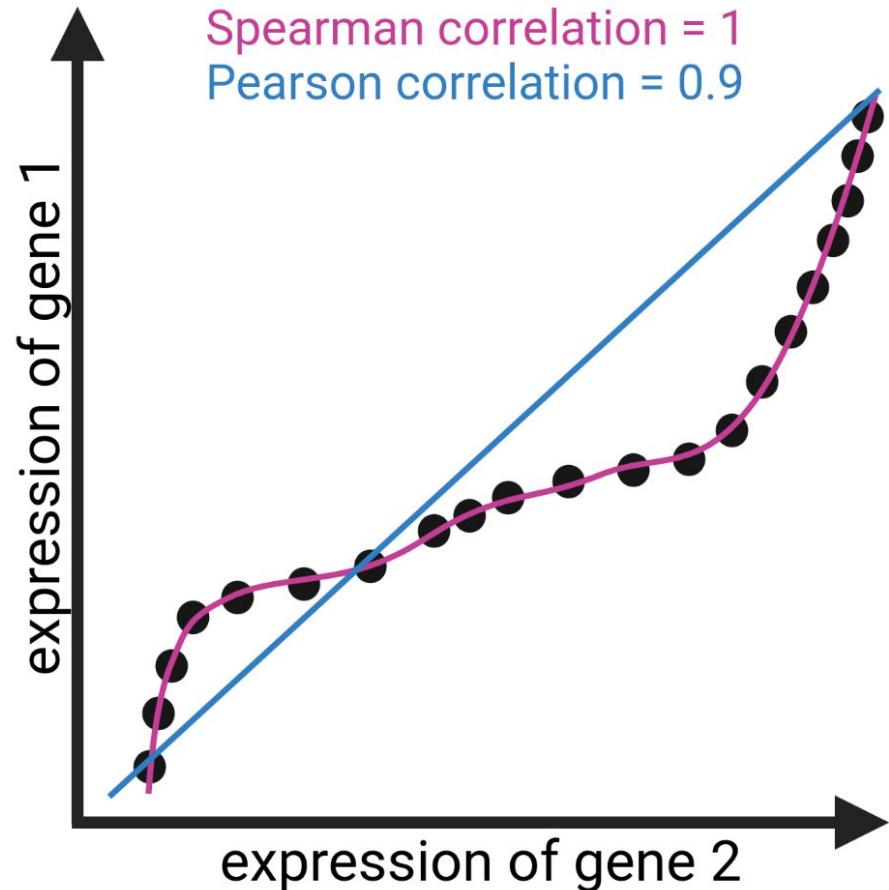
# Pearson correlation coefficient

- Line is fitted to achieve minimal distance of all data points to the line
- Only good for linear correlation

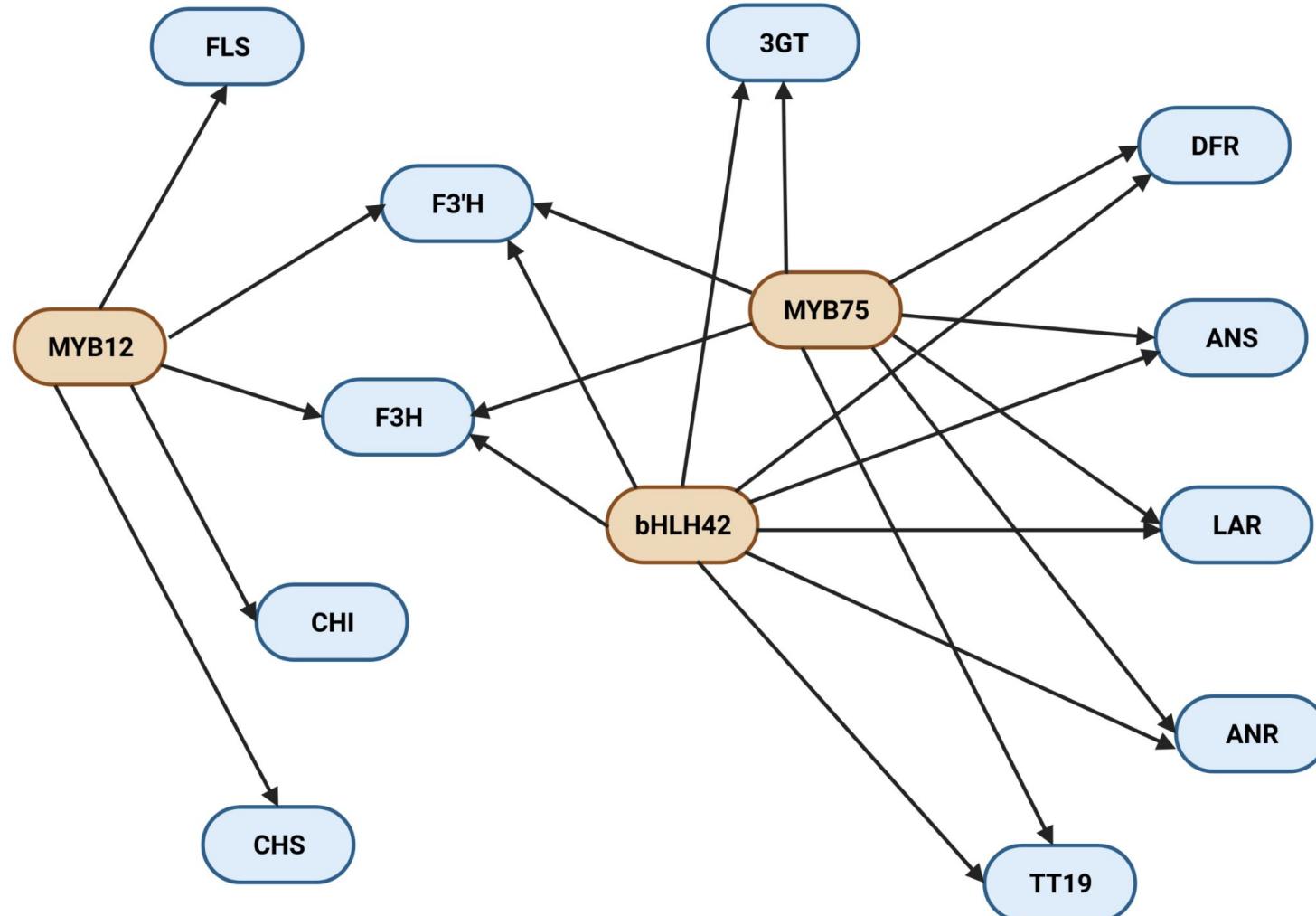


# Spearman correlation coefficient

- Rank-based correlation coefficient
- Not restricted to linear correlation
- More appropriate for gene expression which might not show linear correlation



# Coexpression network example



# Summary

- Importance of data literacy
- Electronic lab books & repositories
- Hypothesis-driven research
- Hands-on data management tipps
- Linux introduction
- Giving presentations
- Social media/networks
- Flavonoid biosynthesis and its regulation

# Time for questions!



# Questions

1. What do you know about the growth of (sequence) databases?
2. What is Data Literacy?
3. Why is Data Literacy important?
4. Which properties has a hypothesis?
5. What are important considerations when naming files?
6. What needs to be documented?
7. What are advantages of an ELN?
8. What is LIMS?
9. What characterizes lab 4.0?
10. What is a good structure for a scientific talk?
11. Which social media have a relevance for scientists?
12. Which factors can cause the lack of anthocyanins?
13. Which branches belong to the flavonoid biosynthesis?
14. What is the MBW complex?

