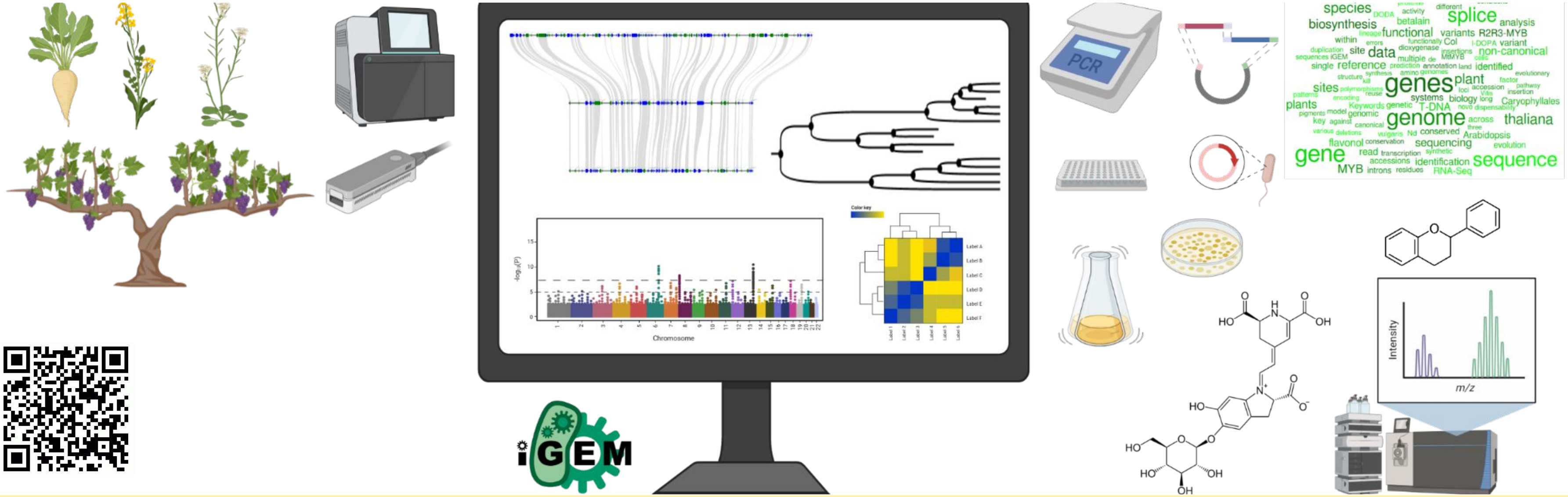
Technische Universität Braunschweig



Introduction to data literacy in genomics

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

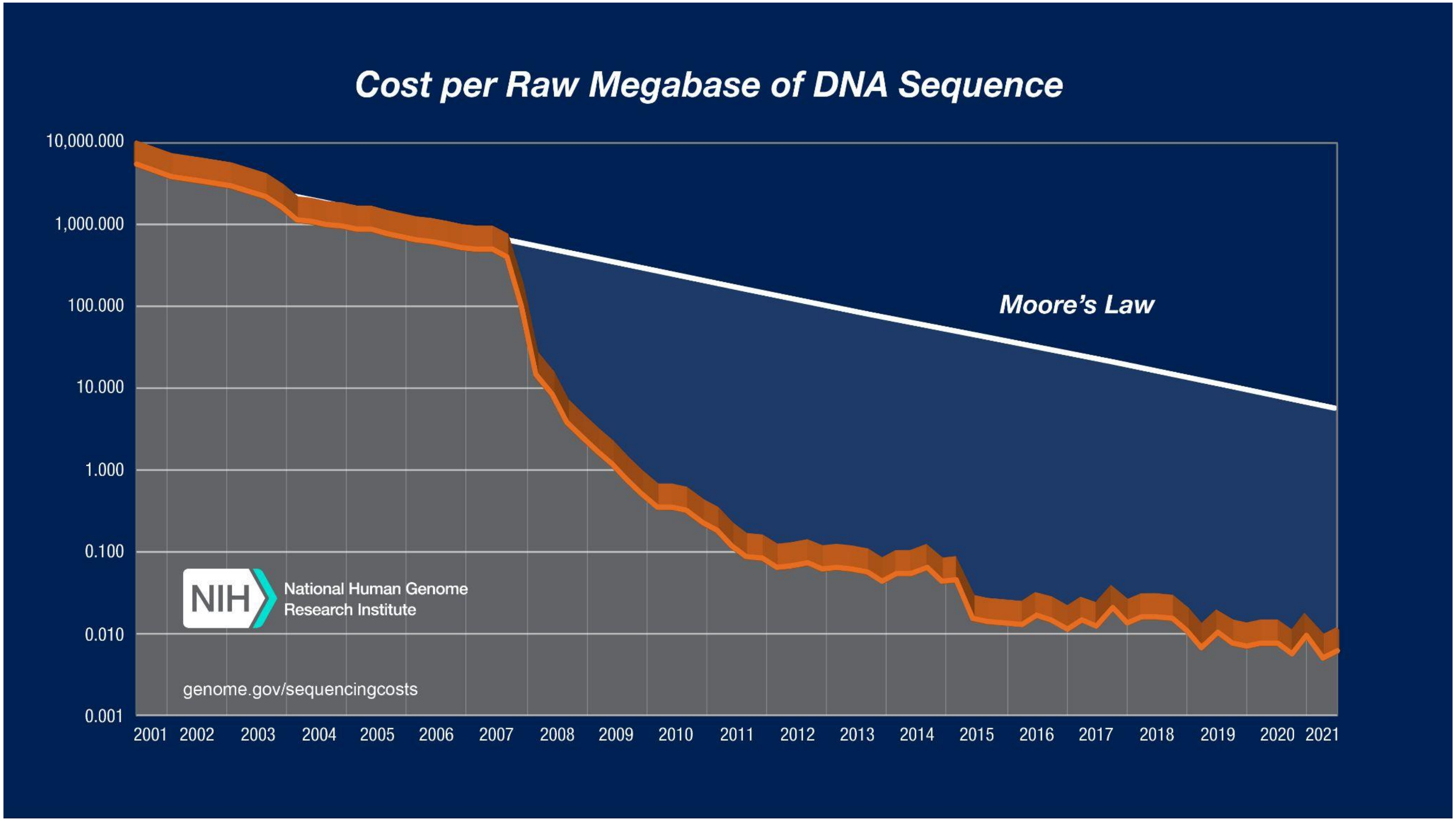
Availability of slides

- 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

‘Big Data’

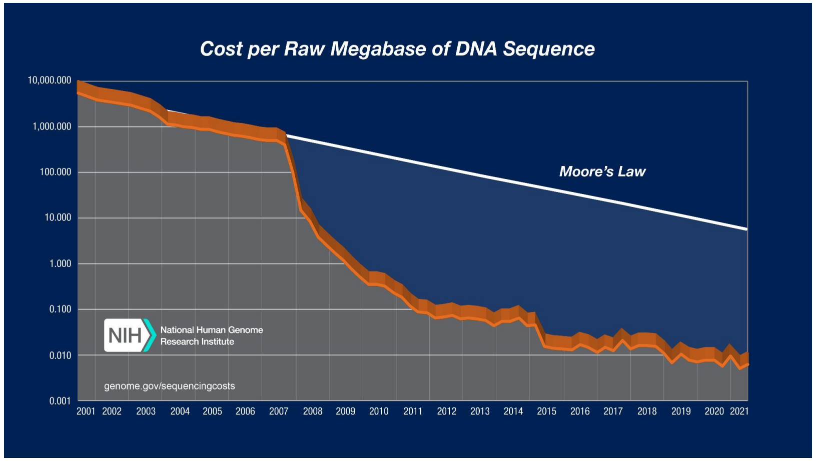

Cost per Raw Megabase of DNA Sequence
10,000.000
1,000.000
100.000
10.000
1.000
0.100
0.010
0.001
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Moore’s Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts

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



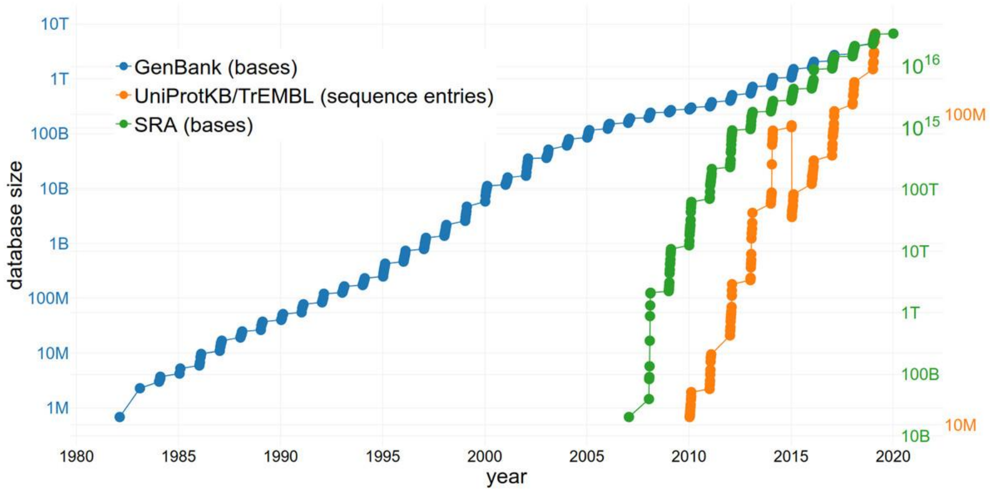
What is Data Literacy?

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



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

Why is Data Literacy important for you?

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

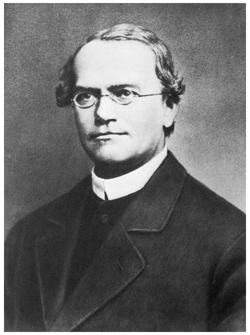
https://doi.org/10.1515/jib-2019-0005

More definitions

- Data: collection of observations
  - 2020: 312 red flowers; 95 white flowers
  - 2021: 298 red flowers; 98 white flowers

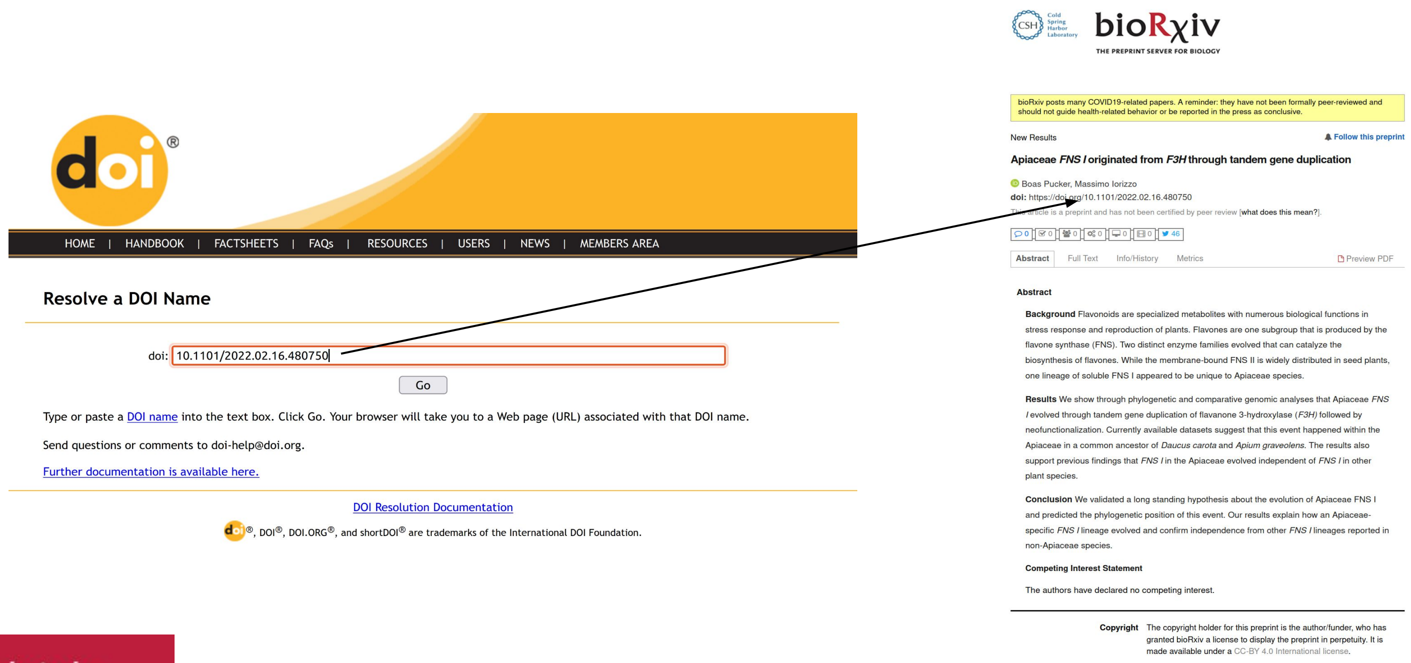
- Knowledge: accumulation of facts and data
  - Ratio of red flowers to white flowers is 3:1

- 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/


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

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.

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.

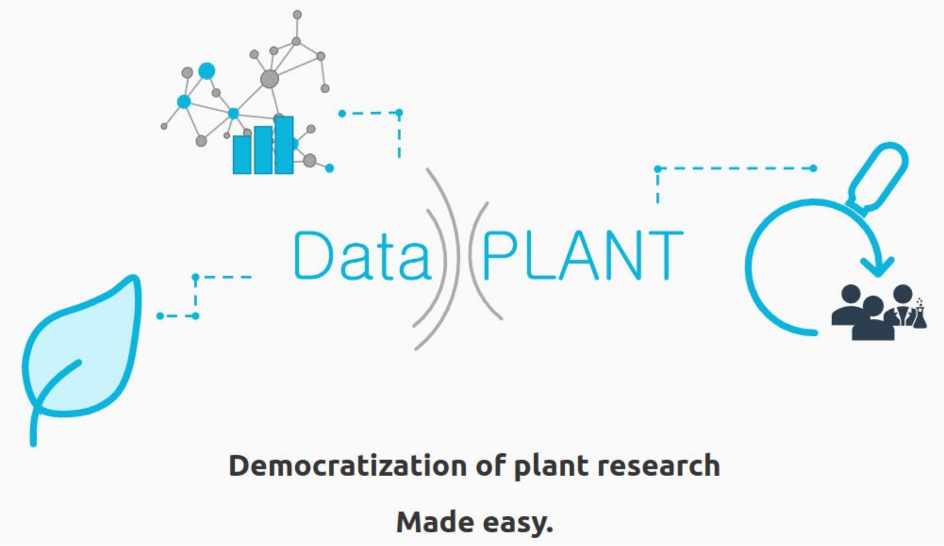
Competing Interest Statement

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.

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

https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en

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

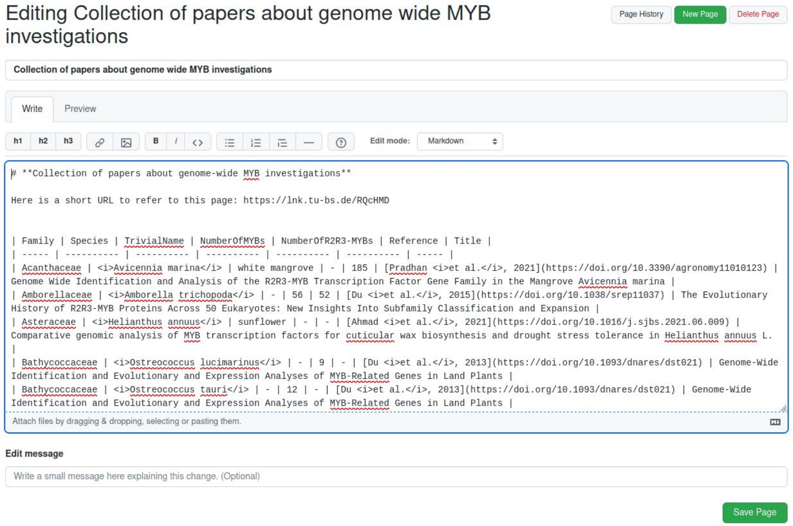
- Simple wiki page

- 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



Editing Collection of papers about genome wide MYB investigations

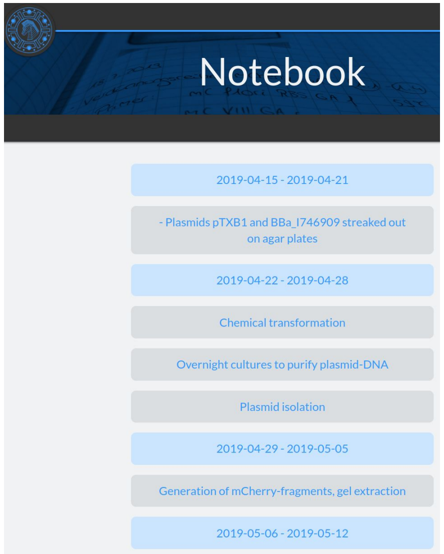
Collection of papers about genome wide MYB investigations

```
# **Collection of papers about genome-wide MYB investigations**

Here is a short URL to refer to this page: https://lnk.tu-bs.de/RQcHMD


| Family | Species | TrivialName | NumberOfMYBs | NumberOfR2R3-MYBs | Reference | Title |
| ----- | ---------- | ---------- | ---------- | ---------- | ---------- | ----- |
| Acanthaceae | <i>Avicennia marina</i> | white mangrove | - | 185 | [Pradhan <i>et al.</i>, 2021](https://doi.org/10.3390/agronomy11010123) |
Genome Wide Identification and Analysis of the R2R3-MYB Transcription Factor Gene Family in the Mangrove Avicennia marina |
| Amborellaceae | <i>Amborella trichopoda</i> | - | 56 | 52 | [Du <i>et al.</i>, 2015](https://doi.org/10.1038/srep11037) | The Evolutionary
History of R2R3-MYB Proteins Across 50 Eukaryotes: New Insights Into Subfamily Classification and Expansion |
| Asteraceae | <i>Helianthus annuus</i> | sunflower | - | - | [Ahmad <i>et al.</i>, 2021](https://doi.org/10.1016/j.sjbs.2021.06.009) |
Comparative genomic analysis of MYB transcription factors for cuticular wax biosynthesis and drought stress tolerance in Helianthus annuus L.
|
| Bathycoccaceae | <i>Ostreococcus lucimarinus</i> | - | 9 | - | [Du <i>et al.</i>, 2013](https://doi.org/10.1093/dnares/dst021) | Genome-Wide
Identification and Evolutionary and Expression Analyses of MYB-Related Genes in Land Plants |
| Bathycoccaceae | <i>Ostreococcus tauri</i> | - | 12 | - | [Du <i>et al.</i>, 2013](https://doi.org/10.1093/dnares/dst021) | Genome-Wide
Identification and Evolutionary and Expression Analyses of MYB-Related Genes in Land Plants |
```

Attach files by dragging & dropping, selecting or pasting them.

Edit message

Write a small message here explaining this change. (Optional)



Notebook

2019-04-15 - 2019-04-21

- Plasmids pTXB1 and BBa_I746909 streaked out on agar plates

2019-04-22 - 2019-04-28

Chemical transformation

Overnight cultures to purify plasmid-DNA

Plasmid isolation

2019-04-29 - 2019-05-05

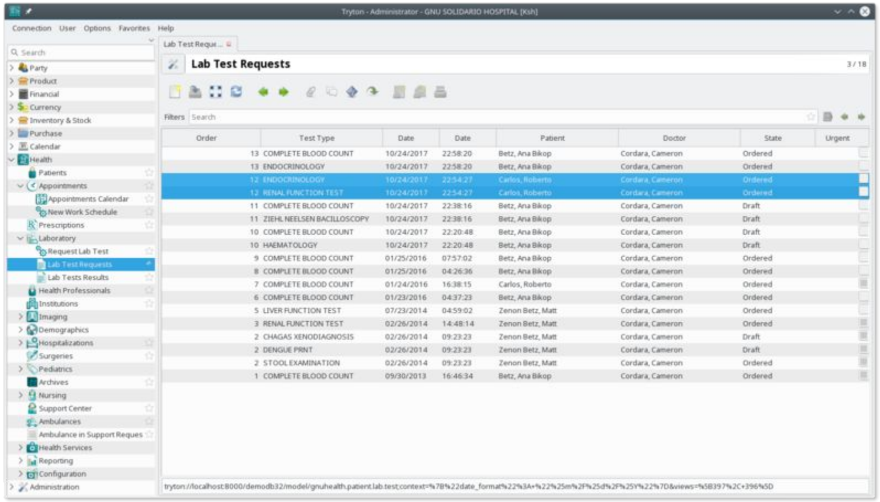
Generation of mCherry-fragments, gel extraction

2019-05-06 - 2019-05-12

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



Lab Test Requests

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	22:58:20	Betz, Ana Bikop	Cordara, Cameron	Ordered	□
12	ENDOCRINOLOGY	10/24/2017	22:54:27	Carlos, Roberto	Cordara, Cameron	Ordered	□
12	RENAL FUNCTION TEST	10/24/2017	22:54:27	Carlos, Roberto	Cordara, Cameron	Ordered	□
11	COMPLETE BLOOD COUNT	10/24/2017	22:38:16	Betz, Ana Bikop	Cordara, Cameron	Draft	□
11	ZIEHL NEELSEN BACILLOSCOPY	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	HAEMATOLOGY	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	04:59:02	Zenon Betz, Matt	Cordara, Cameron	Ordered	□
3	RENAL FUNCTION TEST	02/26/2014	14:48:14	Zenon Betz, Matt	Cordara, Cameron	Ordered	■
2	CHAGAS XENODIAGNOSIS	02/26/2014	09:23:23	Zenon Betz, Matt	Cordara, Cameron	Draft	■
2	DENGUE PRNT	02/26/2014	09:23:23	Zenon Betz, Matt	Cordara, Cameron	Draft	■
2	STOOL EXAMINATION	02/26/2014	09:23:23	Zenon Betz, Matt	Cordara, Cameron	Ordered	■
1	COMPLETE BLOOD COUNT	09/30/2013	16:46:34	Betz, Ana Bikop	Cordara, Cameron	Ordered	■

tryton://localhost:8000/demodb32/model/gnuhealth.patient.lab.test;context=%7B%22date_format%22%3A+%22%25m%2F%25d%2F%25Y%22%7D&views=%5B397%2C+396%5D

https://commons.wikimedia.org/wiki/File:Gnulims_lab_requests.png

Example: Chemotion

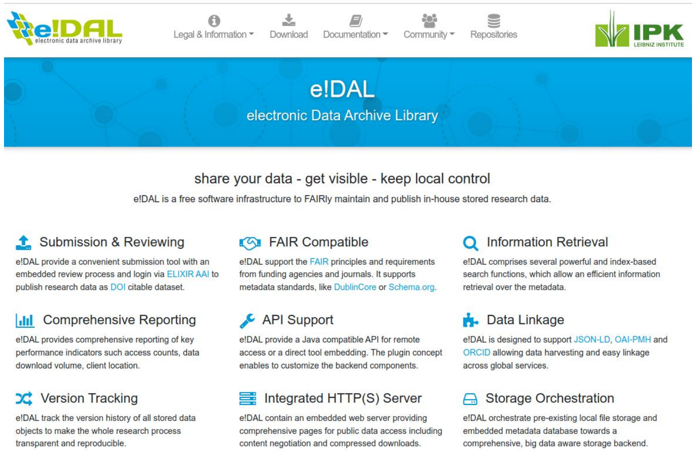
- Research data management tool for chemists
- Electronic Laboratory Notebook (ELN)
- Repository for research data (easy transfer from ELN)
- DOIs are assigned to datasets and protocols

Electronic Laboratory Notebook (ELN) & Repository for Research Data

https://www.chemotion.net/chemotionsaurus/index.html; © Chemotion - KIT

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



e!DAL

electronic Data Archive Library

share your data - get visible - keep local control

e!DAL is a free software infrastructure to FAIRly maintain and publish in-house stored research data.

Submission & Reviewing
e!DAL provide a convenient submission tool with an embedded review process and login via ELIXIR AAI to publish research data as DOI citable dataset.

FAIR Compatible
e!DAL support the FAIR principles and requirements from funding agencies and journals. It supports metadata standards, like DublinCore or Schema.org.

Information Retrieval
e!DAL comprises several powerful and index-based search functions, which allow an efficient information retrieval over the metadata.

Comprehensive Reporting
e!DAL provides comprehensive reporting of key performance indicators such access counts, data download volume, client location.

API Support
e!DAL provide a Java compatible API for remote access or a direct tool embedding. The plugin concept enables to customize the backend components.

Data Linkage
e!DAL is designed to support JSON-LD, OAI-PMH and ORCID allowing data harvesting and easy linkage across global services.

Version Tracking
e!DAL track the version history of all stored data objects to make the whole research process transparent and reproducible.

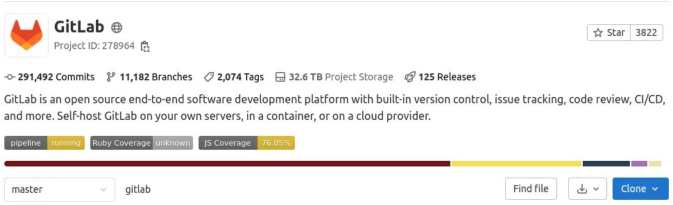
Integrated HTTP(S) Server
e!DAL contain an embedded web server providing comprehensive pages for public data access including content negotiation and compressed downloads.

Storage Orchestration
e!DAL orchestrate pre-existing local file storage and embedded metadata database towards a comprehensive, big data aware storage backend.

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



GitLab

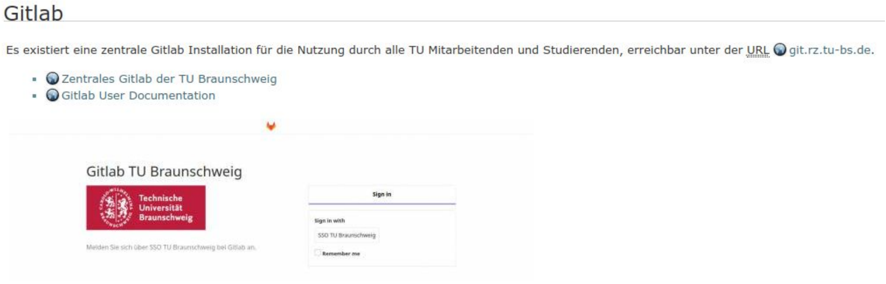
Project ID: 278964

☆ Star 3822

291,492 Commits · 11,182 Branches · 2,074 Tags · 32.6 TB Project Storage · 125 Releases

GitLab is an open source end-to-end software development platform with built-in version control, issue tracking, code review, CI/CD, and more. Self-host GitLab on your own servers, in a container, or on a cloud provider.

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



Gitlab

Es existiert eine zentrale Gitlab Installation für die Nutzung durch alle TU Mitarbeitenden und Studierenden, erreichbar unter der URL git.rz.tu-bs.de.

- Zentrales Gitlab der TU Braunschweig
- Gitlab User Documentation

Gitlab TU Braunschweig

Melden Sie sich über SSO TU Braunschweig bei Gitlab an.

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

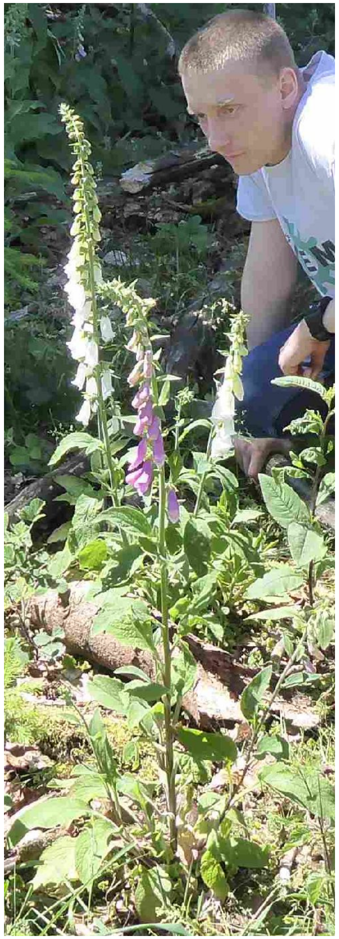
- Samples are labeled with barcodes
  - Barcodes and scanning avoid human errors
  - Can also prevent fraud

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

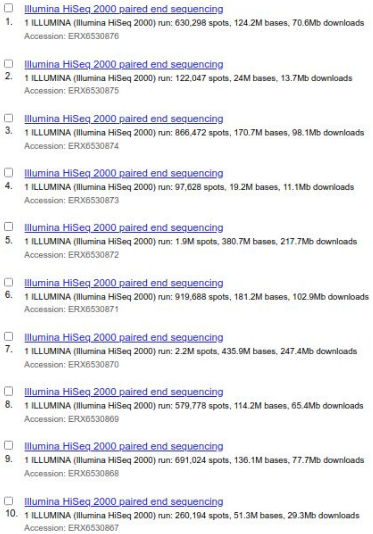
What causes the color difference?



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)

(Mateus S. Figueiredo, https://commons.wikimedia.org/wiki/File:Digitalis_purpurea_phenotypes.png, CC BY-SA 3.0)



□ Illumina HiSeq 2000 paired end sequencing
1. 1 ILLUMINA (Illumina HiSeq 2000) run: 630,298 spots, 124.2M bases, 70.6Mb downloads
Accession: ERX6530876

□ Illumina HiSeq 2000 paired end sequencing
2. 1 ILLUMINA (Illumina HiSeq 2000) run: 122,047 spots, 24M bases, 13.7Mb downloads
Accession: ERX6530875

□ Illumina HiSeq 2000 paired end sequencing
3. 1 ILLUMINA (Illumina HiSeq 2000) run: 866,472 spots, 170.7M bases, 98.1Mb downloads
Accession: ERX6530874

□ Illumina HiSeq 2000 paired end sequencing
4. 1 ILLUMINA (Illumina HiSeq 2000) run: 97,628 spots, 19.2M bases, 11.1Mb downloads
Accession: ERX6530873

□ Illumina HiSeq 2000 paired end sequencing
5. 1 ILLUMINA (Illumina HiSeq 2000) run: 1.9M spots, 380.7M bases, 217.7Mb downloads
Accession: ERX6530872

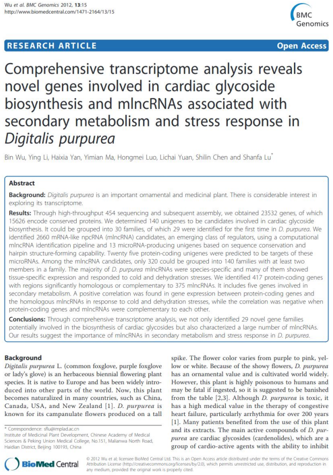
□ Illumina HiSeq 2000 paired end sequencing
6. 1 ILLUMINA (Illumina HiSeq 2000) run: 919,688 spots, 181.2M bases, 102.9Mb downloads
Accession: ERX6530871

□ Illumina HiSeq 2000 paired end sequencing
7. 1 ILLUMINA (Illumina HiSeq 2000) run: 2.2M spots, 435.9M bases, 247.4Mb downloads
Accession: ERX6530870

□ Illumina HiSeq 2000 paired end sequencing
8. 1 ILLUMINA (Illumina HiSeq 2000) run: 579,778 spots, 114.2M bases, 65.4Mb downloads
Accession: ERX6530869

□ Illumina HiSeq 2000 paired end sequencing
9. 1 ILLUMINA (Illumina HiSeq 2000) run: 691,024 spots, 136.1M bases, 77.7Mb downloads
Accession: ERX6530868

□ Illumina HiSeq 2000 paired end sequencing
10. 1 ILLUMINA (Illumina HiSeq 2000) run: 260,194 spots, 51.3M bases, 29.3Mb downloads
Accession: ERX6530867



Wu *et al. BMC Genomics* 2012, **13**:15
http://www.biomedcentral.com/1471-2164/13/15

BMC Genomics

RESEARCH ARTICLE Open Access

Comprehensive transcriptome analysis reveals novel genes involved in cardiac glycoside biosynthesis and mlncRNAs associated with secondary metabolism and stress response in *Digitalis purpurea*

Bin Wu, Ying Li, Haixia Yan, Yimian Ma, Hongmei Luo, Lichai Yuan, Shilin Chen and Shanfa Lu*

Abstract

Background: *Digitalis purpurea* is an important ornamental and medicinal plant. There is considerable interest in exploring its transcriptome.

Results: Through high-throughput 454 sequencing and subsequent assembly, we obtained 23532 genes, of which 15626 encode conserved proteins. We determined 140 unigenes to be candidates involved in cardiac glycoside biosynthesis. It could be grouped into 30 families, of which 29 were identified for the first time in *D. purpurea*. We identified 2660 mRNA-like npcRNA (mlncRNA) candidates, an emerging class of regulators, using a computational mlncRNA identification pipeline and 13 microRNA-producing unigenes based on sequence conservation and hairpin structure-forming capability. Twenty five protein-coding unigenes were predicted to be targets of these microRNAs. Among the mlncRNA candidates, only 320 could be grouped into 140 families with at least two members in a family. The majority of *D. purpurea* mlncRNAs were species-specific and many of them showed tissue-specific expression and responded to cold and dehydration stresses. We identified 417 protein-coding genes with regions significantly homologous or complementary to 375 mlncRNAs. It includes five genes involved in secondary metabolism. A positive correlation was found in gene expression between protein-coding genes and the homologous mlncRNAs in response to cold and dehydration stresses, while the correlation was negative when protein-coding genes and mlncRNAs were complementary to each other.

Conclusions: Through comprehensive transcriptome analysis, we not only identified 29 novel gene families potentially involved in the biosynthesis of cardiac glycosides but also characterized a large number of mlncRNAs. Our results suggest the importance of mlncRNAs in secondary metabolism and stress response in *D. purpurea*.

Background

Digitalis purpurea L. (common foxglove, purple foxglove or lady's glove) is an herbaceous biennial flowering plant species. It is native to Europe and has been widely introduced into other parts of the world. Now, this plant becomes naturalized in many countries, such as China, Canada, USA, and New Zealand [1]. *D. purpurea* is known for its campanulate flowers produced on a tall spike. The flower color varies from purple to pink, yellow or white. Because of the showy flowers, *D. purpurea* has an ornamental value and is cultivated world widely. However, this plant is highly poisonous to humans and may be fatal if ingested, so it is suggested to be banished from the table [2,3]. Although *D. purpurea* is toxic, it has a high medical value in the therapy of congestive heart failure, particularly arrhythmia for over 200 years [1]. Many patients benefited from the use of this plant and its extracts. The main active compounds of *D. purpurea* are cardiac glycosides (cardenolides), which are a group of cardio-active agents with the ability to inhibit

* Correspondence: sflu@implad.ac.cn
Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, No.151, Malianwa North Road, Haidian District, Beijing 100193, China

BioMed Central © 2012 Wu et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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

Example: hypotheses

- pH value of the soil is responsible for the color
- Presence of metal ions in the soil is responsible
- Genetic factors determines the color
- Pollination status determines the color
- Plants belong to different species

What do we need to test a hypothesis?

- Example: genetic factor determines the color difference
- Genome sequence (long read sequencing & assembly)
- Structural annotation of the genome sequence (gene prediction)
- Functional annotation of the genes (function prediction)
- Data sets to compare white and red flowering plants (read mapping & variant calling)
- 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

How to structure your data?

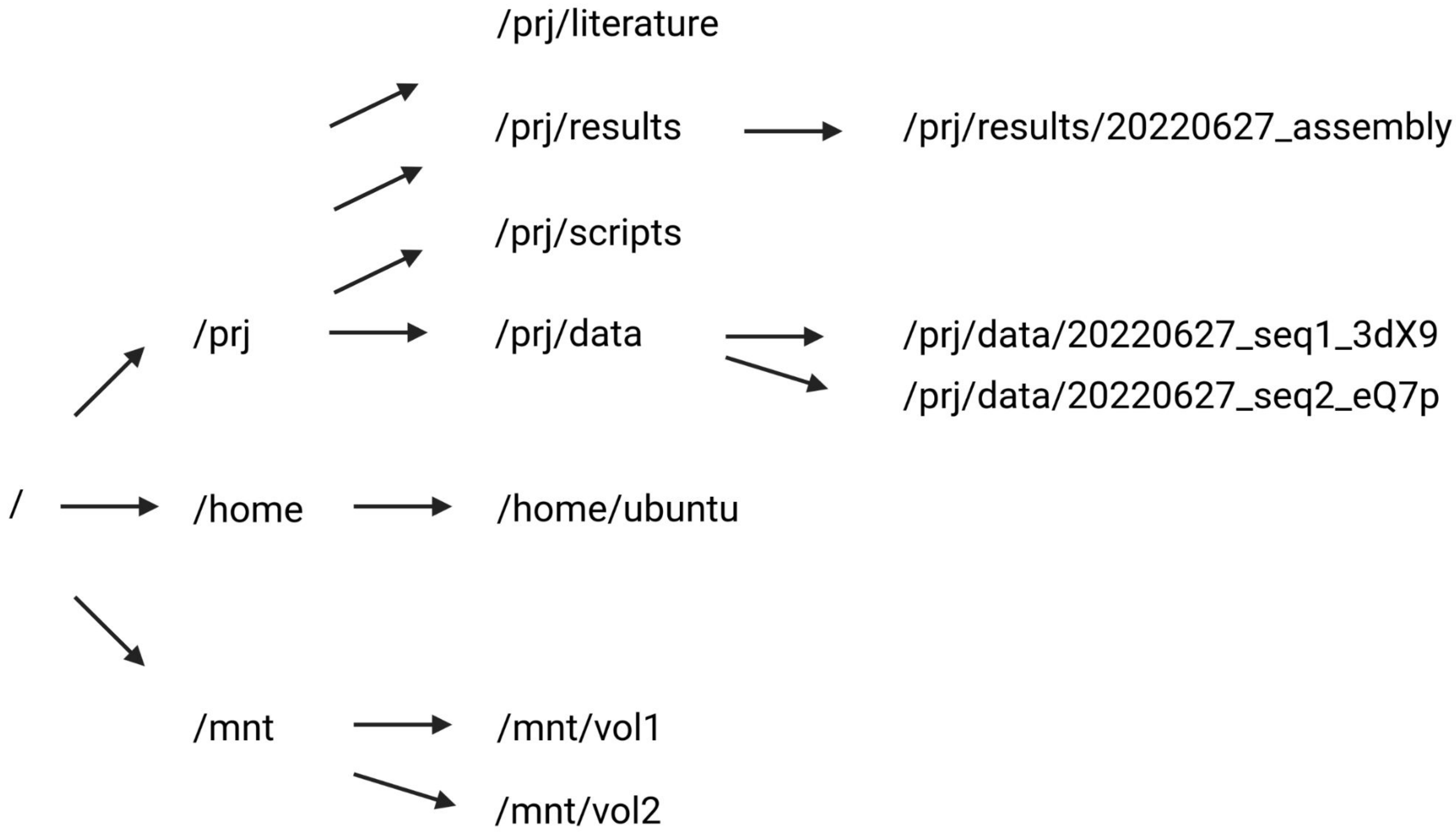
- Document every step (e.g. in a README)
    - Origin of data sets; versions of tools, parameters of analyses

- Keep raw data sets separated from scripts and results

- Sort data by project

- Structure analysis related data sets/results by date

Example: Linux file system



/

/prj

/prj/literature

/prj/results

/prj/results/20220627_assembly

/prj/scripts

/prj/data

/prj/data/20220627_seq1_3dX9

/prj/data/20220627_seq2_eQ7p

/home

/home/ubuntu

/mnt

/mnt/vol1

/mnt/vol2

Naming files

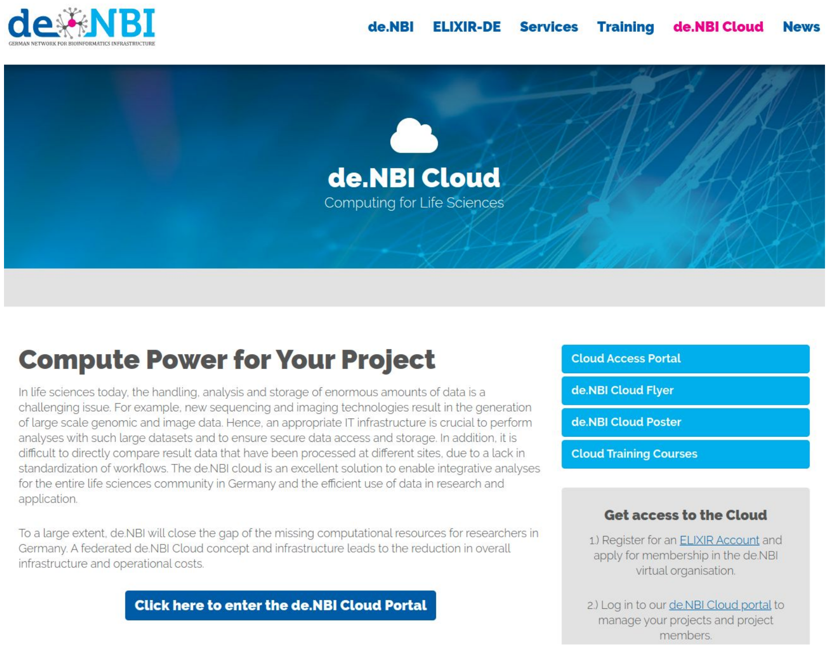
- File names should be informative
- Never use ‘new‘ or ‘final’; use version numbers instead
- Use date as file name prefix (e.g. 2022-06-27 or 20220627)
- Never use spaces in file or folder names (underscore or minus as replacement)
- Include your initials as suffix in collaborative projects

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



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.

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/

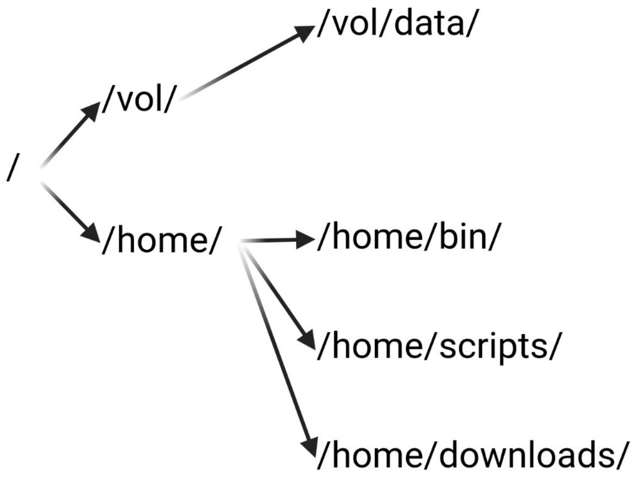
- Scp (secure copy): command line file transfer method

- Wget: command line file transfer method
  - https://ftp.gnu.org/old-gnu/Manuals/wget-1.8.1/html_mono/wget.html

- Rsync: sophisticated file transfer method that avoids redundant transfers
  - https://wiki.ubuntuusers.de/rsync/

Linux system structure

- Linux (Ubuntu) is operation system of choice for bioinformatics
- Hierarchical structure of directory ('/' is basis)
- Separation of tool installation and data sets
- File naming conventions:
  - Never use spaces in file or directory names
  - Include dates in file names (year-month-day)
- Indicating commands with prefix '$'



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

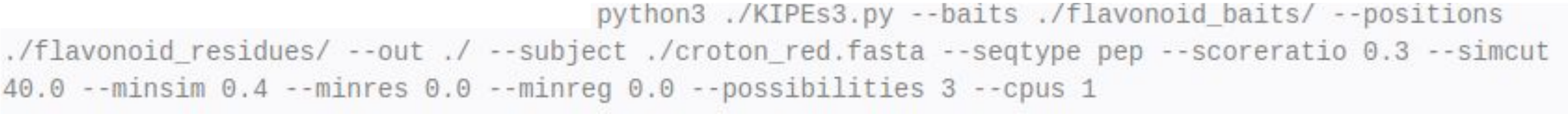
- Checking content of a folder:
  - $ ls #shows content of current folder
  - $ ls -lh   #shows more details
  - -l triggers display of additional details
  - -h human readable
  - -a show also hidden files

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

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?