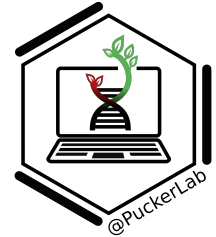
UNIVERSITÄT BONN




Prof. Dr. Boas Pucker

# Long Read Genomics

## - Submitting Sequencing Data & Reusing Public Data

UNIVERSITÄT BONN

# Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - GitHub: https://github.com/bpucker/LRG

- Questions: Feel free to ask at any time

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

UNIVERSITÄT BONN

# Repetition

- Read mapping (short reads vs. long reads)
- Variant calling (short reads vs. long reads)
- Evolutionary genomics
- Comparative genomics & synteny analysis
- Application examples

UNIVERSITÄT BONN

# Data Dissemination

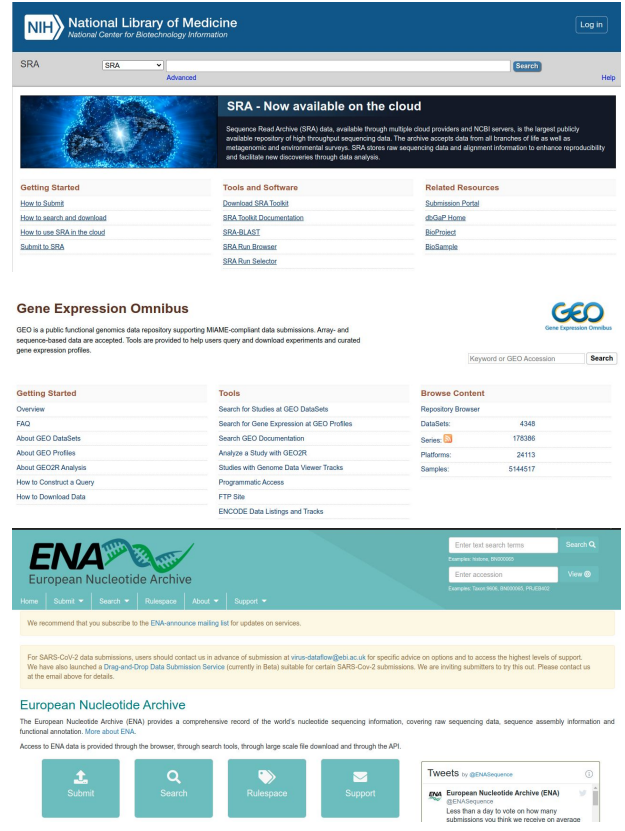
UNIVERSITÄT BONN

# Which data of a study can be shared?

- Sequencing data
- Genome sequence and annotation
- Metadata (phenotyping data, geographic positions, …)
- ….
- Remember to conserve the sequenced plant!

# Sequencing data sets

- Sequence Read Archive (SRA)
- Gene Expression Omnibus (GEO)
  (also submission to SRA)
- European Nucleotide Archive (ENA)
- Read Selector

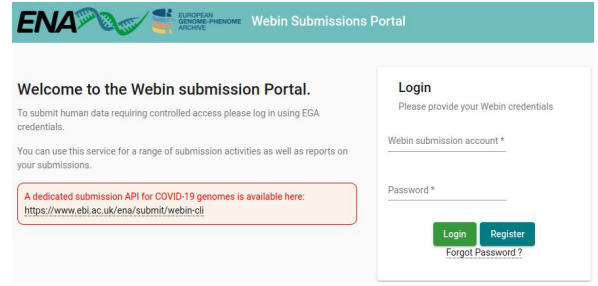





https://www.ncbi.nlm.nih.gov/sra
https://www.ncbi.nlm.nih.gov/geo/
https://www.ebi.ac.uk/ena/browser/home

# How to submit reads (to ENA)?

- Log into the submission portal
- Register study
- Register samples (spreadsheet upload option)
- Prepare and upload read files (via ftp)
- Submit sequence reads (spreadsheet upload option)




| Accession | BioSample | Title |
|---|---|---|
| ERS3371290 | SAMEA5569268 | RNA-Seq Hypertelis bowkeriana flower |
| ERS3371289 | SAMEA5569267 | RNA-Seq Hypertelis bowkeriana leaf |
| ERS3371288 | SAMEA5569266 | RNA-Seq of Simmondsia chinensis flower |
| ERS2294062 | SAMEA104692679 | Corrigiola litoralis genome sequencing |
| ERS2294061 | SAMEA104692678 | Spergula arvensis genome sequencing |
| ERS2294060 | SAMEA104692677 | Simmondsia chinensis genome sequencing |
| ERS2294059 | SAMEA104692676 | Pharnaceum exiguum genome sequencing |
| ERS2294058 | SAMEA104692675 | Microtea debilis genome sequencing |
| ERS2294049 | SAMEA104692666 | Macarthuria australis genome sequencing |
| ERS2294048 | SAMEA104692665 | Limeum aethiopicum genome sequencing |



https://ena-docs.readthedocs.io/en/latest/submit/reads.html

UNIVERSITÄT BONN

# How to submit reads (to ENA)? part I

Submission of reads requires many details:

- Project_accession: accession assigned by ENA
- Project_alias: name assigned by user
- Sample_alias: accession assigned by ENA
- Experiment_alias: accession assigned by ENA
- Run_alias: XXX
- Library_name: User picks this name
- Library_source: GENOMIC
- Library_selection: RANDOM
- Library_strategy: XXX
- Design_description: XXX
- Library_construction_protocol: TrueSeq V2
- Instrument_model: Illumina HiSeq1500

UNIVERSITÄT BONN

# How to submit reads (to ENA)? part II

Submission of reads requires many details:

- File_type: FASTQ
- Library_layout: PAIRED
- Insert_size: 600
- Forward_file_name: fw_file.fastq.gz
- Forward_file_md5: jel9aks5joe8iaj1ie2lfk4jsk6flji
- Forward_file_unencrypted_md5:
- Reverse_file_name: rv_file.fastq.gz
- Reverse_file_md5: k1ea0wi7oji32so45jbae6fo81337xd
- reverse_file_unencrypted_md5

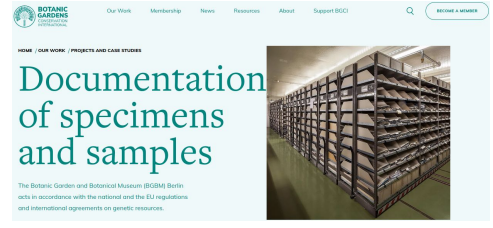
UNIVERSITÄT BONN

# ONT data submission to ENA

- Upload of gzip-compressed FASTQ files to ENA
- Upload of gzip-compressed POD5 + FASTQ files in archive
- Submission of metadata through Webin (web interface)

# Plant collections

- Seed banks: focus on crops or model organisms
- Botanical gardens: living collections
- Museums/herbaria: collections of recent and ancestral plants


https://collections.nmnh.si.edu/search/botany/
https://www.bgci.org/our-work/projects-and-case-studies/documentation-of-specimens-abs-and-nagoya/
https://www.braunschweig.de/english/city/sights/_botanical_garden.php
https://www.museumfuernaturkunde.berlin/en/museum/exhibitions/wet-collection
https://www.botanic.cam.ac.uk/collections/herbarium-2/
https://www.kew.org/science/collections-and-resources/collections/herbarium

UNIVERSITÄT BONN

# How to submit voucher herbarium specimen?

- Voucher herbarium specimen = pressed plant sample with collection data

- Voucher herbarium specimen are helpful to support phylogenetic reclassifications

- Process overview:
  - Initial preparations
  - Pressing and drying
  - Identification
  - Labeling
  - Mounting



https://www.floridamuseum.ufl.edu/herbarium/methods/vouchers/

UNIVERSITÄT BONN

# Part1: Preparation for specimen collection

- Select collection location and date
- Obtain collection permit
- Establish official contact (often required by law)
- Make arrangement with herbarium to deposit specimens
- Purchase collection equipment and supplies



https://www.floridamuseum.ufl.edu/herbarium/methods/vouchers/

UNIVERSITÄT BONN

# Part 2: Processing specimens

- Pressing and drying plant specimens: include all parts; unique collection number; collect replicates; press specimen to 11x16 inches; avoid wilting prior to pressing

- Identification of plant specimens: dichotomous keys; published descriptions

- Label of herbarium specimens (Darwin Core standards): scientific name; determiner; detailed location; habitat; collection number; date of collection, …
  - Label formats vary: https://www.floridamuseum.ufl.edu/herbarium/methods/vouchers/

- Mounting herbarium specimens: most herbaria prefer to do this / consultation needed

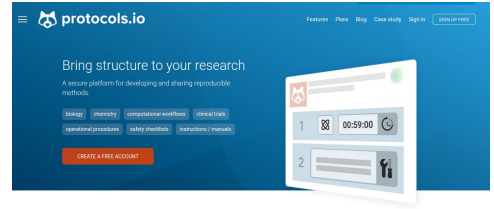
UNIVERSITÄT BONN

# OpenData

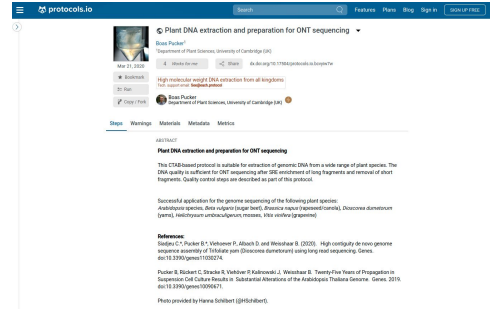
- Everyone can access and re-use these data sets
- Facts cannot be owned by someone
- Huge economic potential through re-use; advantages for society
- Possible restrictions: name author, share-alike
- Related initiative: open source, open content, open access, open education

# OpenProtocols

- Enables others to reproduce experiments
- Protocols are precisely described and freely available to everyone
- DOIs can be assigned to protocols
- Protocols.io is a platform to support the exchange of protocols
- Example: gDNA extraction protocol




https://www.protocols.io/welcome

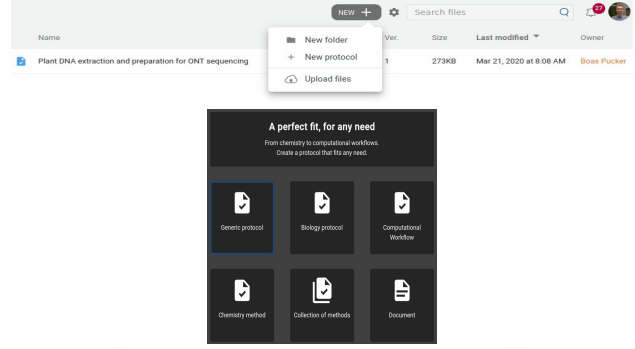


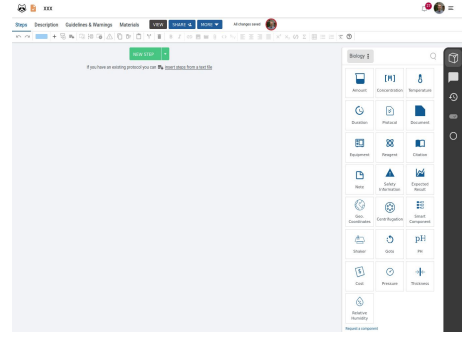

doi:10.17504/protocols.io.bcvyiw7w

# How to submit a protocol?

- Protocol submission:
  - Create a new protocol (can remain private; 5 in free version)
  - Assign a DOI to make it citable
  - Make protocol public

- PDF submission and conversion for $50 (service fee)





https://www.protocols.io/

UNIVERSITÄT BONN

# Creative Commons Licenses

- CC0: creative commons (no restrictions)
- CC BY: no restrictions, but name authors
- CC BY-SA: name authors and share results under same license
- CC BY-NC: name authors; only non-commercial use
- CC BY-NC-SA: name authors; only non-commercial use; share under same license

UNIVERSITÄT BONN

# Which license would you select and why?

- A script to analyze a gene family?
- A table with species observed in a particular forest?
- FASTQ files of a RNA-seq project?
- Genome sequence and the corresponding annotation?
- KM value and Vmax value of an enzyme?
- A protocol for efficient transformation of a plant species?

UNIVERSITÄT BONN

# FAIR data

- Findable:
  - Globally unique and persistent identifier
  - Metadata must be available in connection to the identifier
- Accessible:
  - Retrieval based on the identifier
  - Protocol is open, free, and universally applicable
  - Authentication is possible where needed
  - Metadata are available even if data are restricted
- Interoperable:
  - Metadata use a formal, accessible, broadly accessible language
  - Vocabulary need to follow FAIR standards
- Re-usable:
  - Clear and accessible data usage license
  - Meet domain-relevant community standards



https://www.go-fair.org/fair-principles/

UNIVERSITÄT BONN

# JSON

- JSON = JavaScript Object Notation
- File format for exchange between different tools
- File structure readable by many different tools & human-readable
- Attribute-value pairs (dictionary)

**Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions**

Published online by Cambridge University Press: **11 March 2022**

Boas Pucker, Iker Irisarri, Jan de Vries and Bo Xu

Article | Figures | Peer reviews | Metrics

Save PDF | Share | Cite

**Abstract**

Third-generation long-read sequencing is transforming plant genomics. Oxford Nanopore Technologies and Pacific Biosciences are offering competing long-read sequencing technologies and enable plant scientists to investigate even large and complex plant genomes. Sequencing projects can be conducted by single research groups and sequences of smaller plant genomes can be completed within days. This also resulted in an increased investigation of genomes from multiple species in large scale to address fundamental questions associated with the origin and evolution of land plants. Increased accessibility of sequencing devices and user-friendly software allows more researchers to get involved in genomics. Current challenges are accurately resolving diploid or polyploid genome sequences and better accounting for the intra-specific diversity by switching from the use of single reference genome sequences to a pangenome graph.

**Keywords**

haplophasing | long read sequencing | Oxford Nanopore Technologies (ONT) | Pacific Biosciences (PacBio) | plant genome assembly | plant genomics

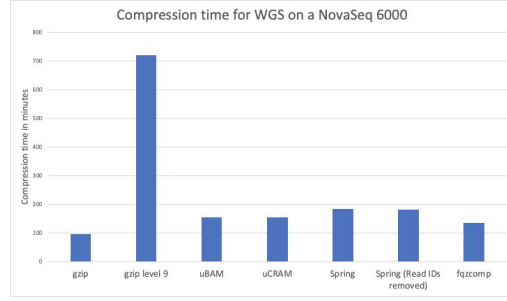
| | |
|---|---|
| Type | Review |
| Information | Quantitative Plant Biology, Volume 3, 2022, e5<br>DOI: https://doi.org/10.1017/qpb.2021.18 |
| Creative Commons | This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited. |
| Copyright | © The Author(s), 2022. Published by Cambridge University Press in association with The John Innes Centre |

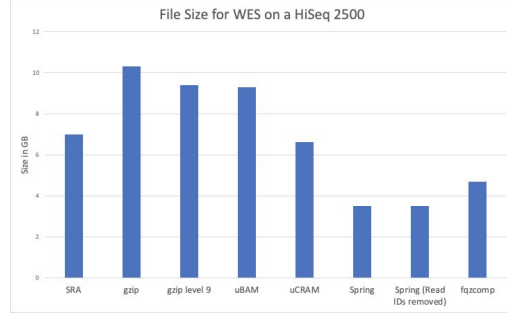
```
{
"authors":
[
{
"firstName": "Boas",
"lastName": "Pucker"
},
{
"firstName": "Iker",
"lastName": "Irisarri"
},
{
"firstName": "Jan",
"lastName": "de Vries"
},
{
"firstName": "Bo",
"lastName": "Xu"
}
],
"title": "Plant genome sequence assembly in the era of long reads",
"url": "https://doi.org/10.1017/qpb.2021.18",
doi: "10.1017/qpb.2021.18"
}
```

UNIVERSITÄT BONN

# Data compression

- Gzip is most frequently applied tool
- Different compression levels (default=6)
  - Level 1 = fast, but small size reduction
  - Level 9 = slow, but substantial size reduction
- File sizes can be reduced by 75%
- Gzip should always be used to reduce disk space requirements






https://github.com/godotgildor/fastq_compression_comparison

UNIVERSITÄT BONN

# tar

- Transfer to large numbers of files is a challenge
- Tar can be used to merge many files into a tar ball
- '.tar.gz' and '.tgz' are extensions of tarballs
- Construct: `tar -cavf archive.tar.gz content`
- Extract: `tar -xzvf archive.tar.gz`



http://www.fibel.org/linux/lfo-0.6.0/node370.html

UNIVERSITÄT BONN

# sha256sum & md5sums

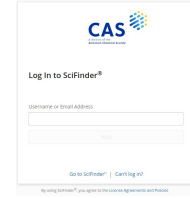
- Generate a digital fingerprint of a given file
- Md5sum is a 128-bit hash
- Md5sum is the standard in many bioinformatic workflows to ensure files have been transferred completely
- Sha256sum is recommended for security relevant purposes when malicious intent is expected

UNIVERSITÄT BONN

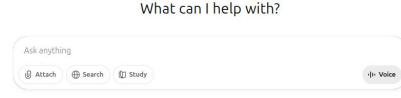
# Databases & Services

# How to find information in the literature?

- GoogleScholar
- SciFinder
- PubMed
- WebOfScience/WebOfKnowledge
- PubPharm
- ChatGPT



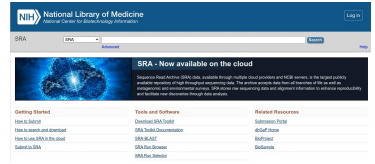



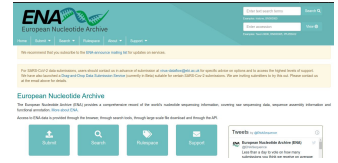
UNIVERSITÄT BONN

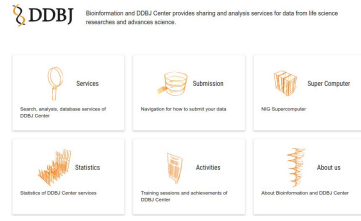
# Which sequence databases do you know?

# Sequence databases

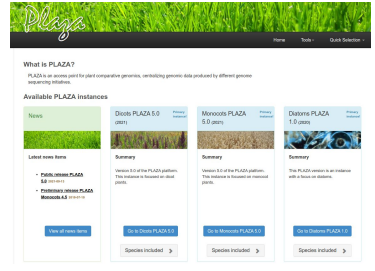
- Sequence Read Archive (SRA)
- European Nucleotide Archive (ENA)
- DNA Data Bank of Japan (DDBJ)
- GenBank (NCBI)
- Joint Genome Institute (JGI) / Phytozome
- PLAZA
- Ensembl genomes





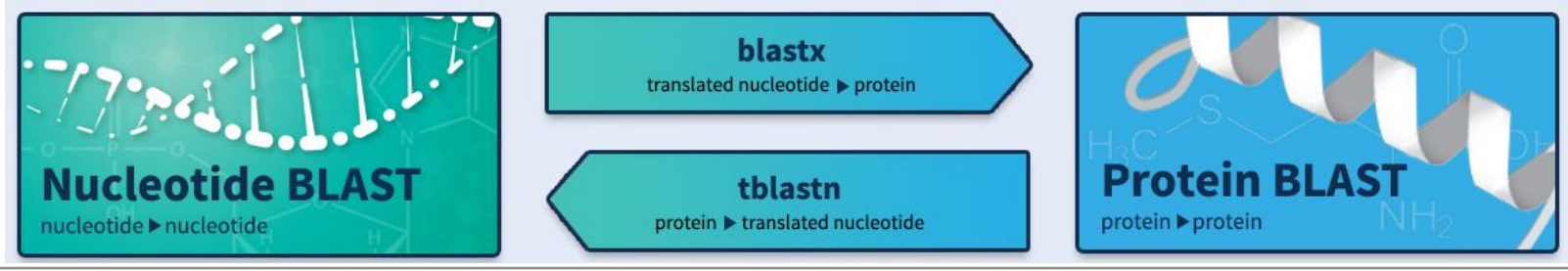


UNIVERSITÄT BONN

# How to search for sequences?

# Basic Local Alignment Search Tool (BLAST)

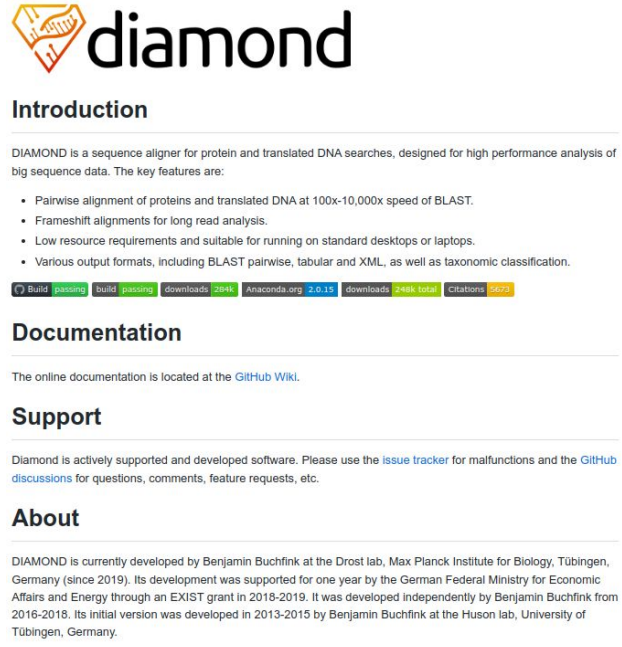
- BLASTn = nucleotide query against a nucleotide database
- BLASTp = protein query against a protein database
- BLASTx = (translated) nucleotide query against a protein database
- tBLASTn = protein query against (translated) nucleotide database



UNIVERSITÄT BONN

# DIAMOND

- BLAST equivalent with respect to result quality and format
- About 50x faster sequence search
- RAM requirements are substantially higher than for BLAST



diamond

## Introduction

DIAMOND is a sequence aligner for protein and translated DNA searches, designed for high performance analysis of big sequence data. The key features are:

- Pairwise alignment of proteins and translated DNA at 100x-10,000x speed of BLAST.
- Frameshift alignments for long read analysis.
- Low resource requirements and suitable for running on standard desktops or laptops.
- Various output formats, including BLAST pairwise, tabular and XML, as well as taxonomic classification.

Build passing | build passing | downloads 284k | Anaconda.org 2.0.15 | downloads 248k total | Citations 5623

## Documentation

The online documentation is located at the GitHub Wiki.

## Support

Diamond is actively supported and developed software. Please use the issue tracker for malfunctions and the GitHub discussions for questions, comments, feature requests, etc.

## About

DIAMOND is currently developed by Benjamin Buchfink at the Drost lab, Max Planck Institute for Biology, Tübingen, Germany (since 2019). Its development was supported for one year by the German Federal Ministry for Economic Affairs and Energy through an EXIST grant in 2018-2019. It was developed independently by Benjamin Buchfink from 2016-2018. Its initial version was developed in 2013-2015 by Benjamin Buchfink at the Huson lab, University of Tübingen, Germany.



https://github.com/bbuchfink/diamond
Buchfink et al., 2021: 10.1038/s41592-021-01101-x

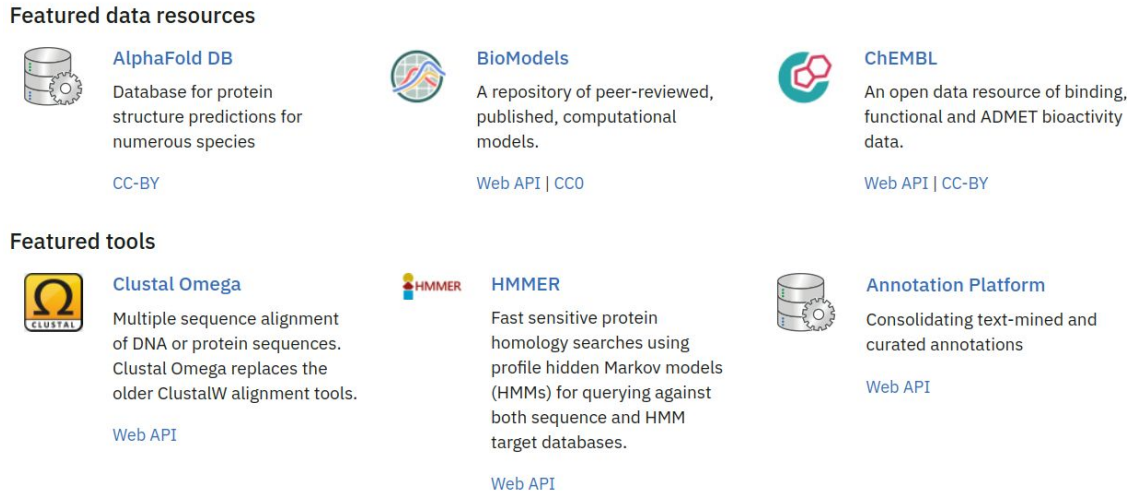
UNIVERSITÄT BONN

# Do you know websites with bioinformatic tools?

# EBI services

- EBI = European Bioinformatics Institute

- Broad range of different services:
  - DNA&RNA
  - Gene expression
  - Literature
  - Structures

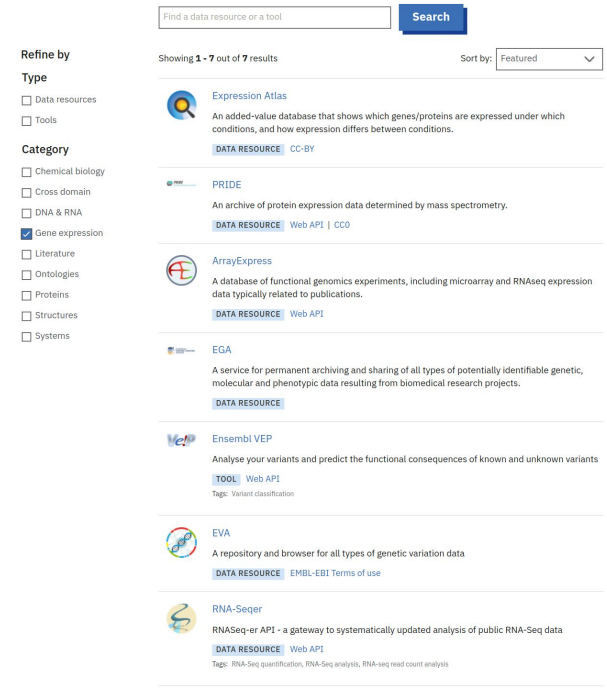





https://www.ebi.ac.uk/services/data-resources-and-tools

# EBI services (gene expression)

- Expression Atlas = Visualization of gene expression
- ArrayExpress = collection of various gene expression analysis data sets
- PRIDE = database of protein abundances (MS-based)




https://www.ebi.ac.uk/services/data-resources-and-tools

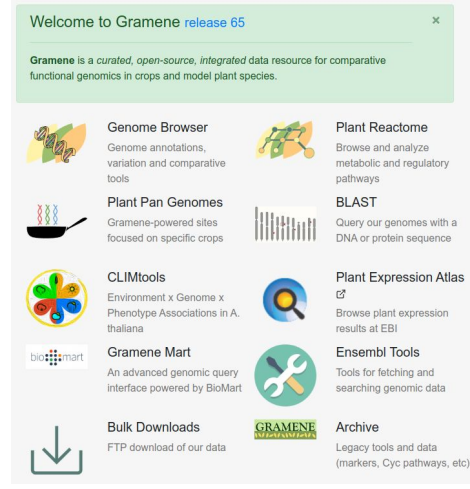
UNIVERSITÄT BONN

# EBI services (DNA&RNA)

- Ensembl = genome databases
- Clustal Omega = sequence alignment tool
- ENA = European Nucleotide Archive (sequence read submission)
- Gramene = comparative plant biology
- MAFFT = sequence alignment tool
- Rfam = database of RNA structures
- T-coffee = sequence alignment tool
- TreeFam = database of phylogenetic trees

# EBI services - Gramene

- Genome browsers allow inspection of plant annotations
- Pangenomic data sets = enables comparative genomics
- Bulk download options = enables users to conduct more customized analysis
- BLAST = sequence comparison
- Expression Atlas (dedicated to plants)



UNIVERSITÄT BONN

# EMBOSS

- EMBOSS = European Molecular Biology Open Software Suite
- Collection of (small) bioinformatic tools on a website
- Link: https://www.bioinformatics.nl/emboss-explorer/
- Homepage: http://emboss.sourceforge.net/



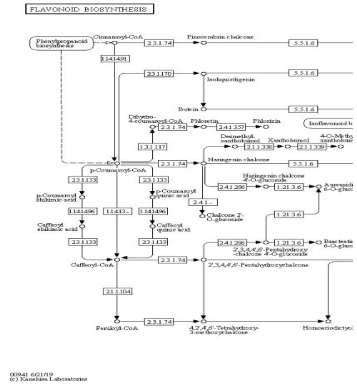
Rice et al., 2000: 10.1016/S0168-9525(00)02024-2

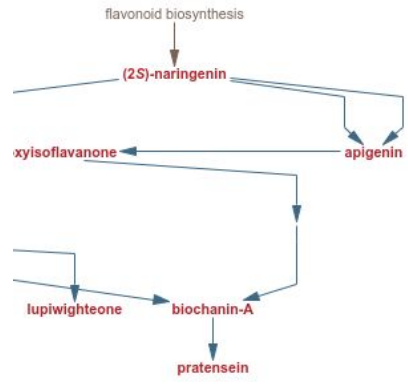
UNIVERSITÄT BONN

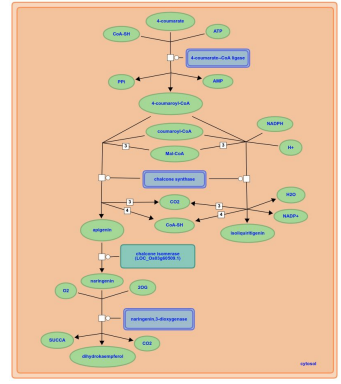
# Where would you look for pathways?

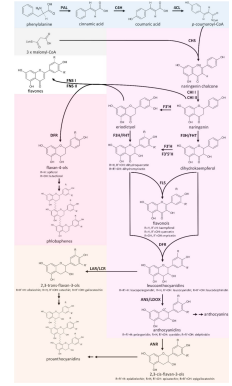
# Pathway databases

- KEGG: https://www.genome.jp
- MetaCyc: https://metacyc.org
- Plant Reactome: https://plantreactome.gramene.org/
- Publications



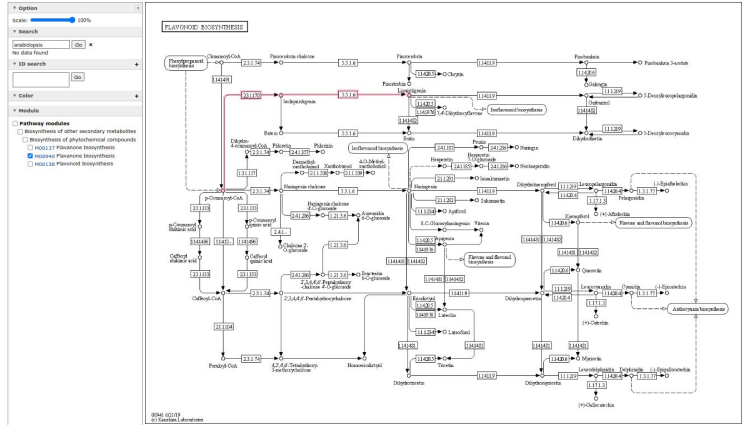


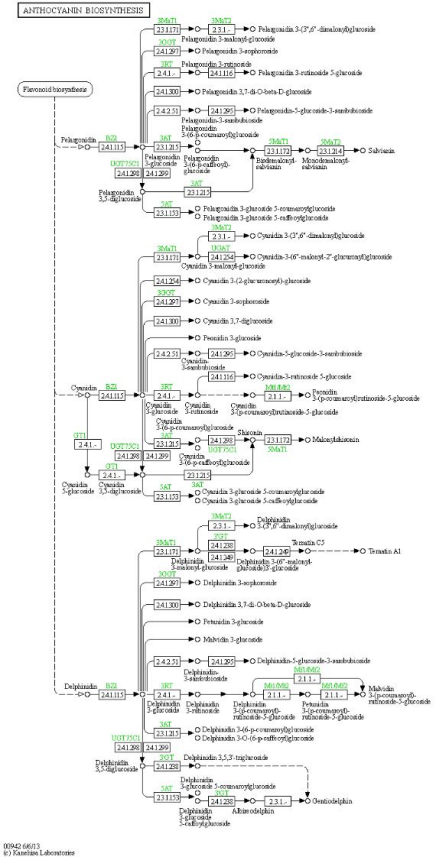


Kanehisa et al., 2015: 10.1093/nar/gkv1070
Pucker et al., 2020: 10.3390/plants9091103

# KEGG - Flavonoid biosynthesis

- Selection of one particular species
- Selection of individual branches in pathways
- Gene names are given in addition to EC numbers






https://www.genome.jp/kegg-bin/show_pathway?map00942

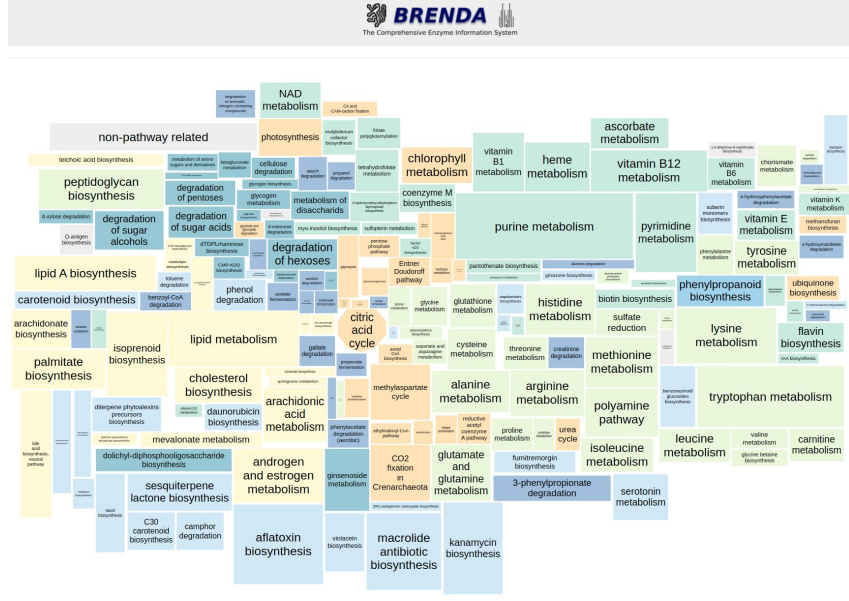
UNIVERSITÄT BONN

# Details about enzymes?

UNIVERSITÄT BONN

# BRENDA: BRaunschweig ENzyme DAtabase

- Enzyme database hosted at TU Braunschweig (BRICS)
- Text and structure-based queries
- Visualization of pathways
- Manual curation of datasets
- Many details about enzyme properties (substrates, kinetics, mutants, …)

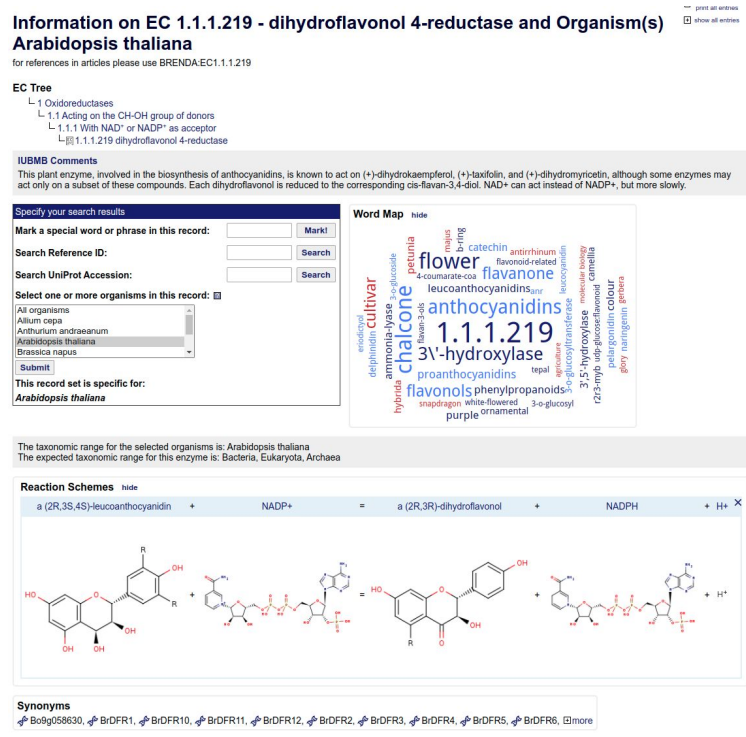





www.brenda-enzymes.org
Chang et al., 2021: 10.1093/nar/gkaa1025

# BRENDA - example (1)

- DFR (dihydroflavonol 4-reductase) is a central enzyme in the anthocyanin biosynthesis
- BRENDA provides information about reaction mechanism and species-specific parameters
- Enzymatic properties of multiple species allow quick comparison

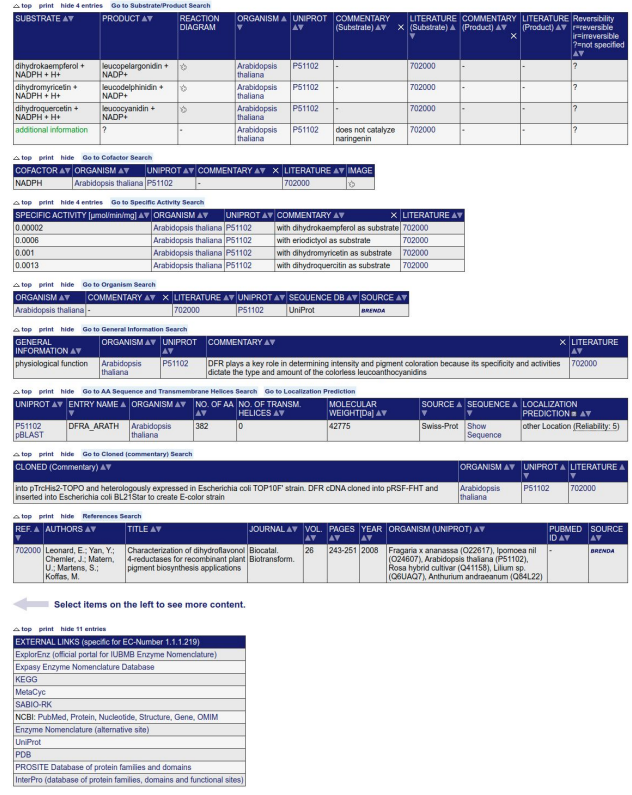





https://www.brenda-enzymes.org/

# BRENDA - example (2)

- Specific substrates and products are listed (reaction details)
- $K_M$ and $V_{max}$ values are included
- References point to publications and other databases; synonyms
- EC number
- Biochemical details: pH optimum, pI value, molecular weight



△ top print hide 4 entries Go to Substrate/Product Search

| SUBSTRATE | PRODUCT | REACTION DIAGRAM | ORGANISM | UNIPROT | COMMENTARY (Substrate) | LITERATURE (Substrate) | COMMENTARY (Product) | LITERATURE (Product) | Reversibility r=reversible ir=irreversible ?=not specified |
|---|---|---|---|---|---|---|---|---|---|
| dihydrokaempferol + NADPH + H+ | leucopelargonidin + NADP+ | | Arabidopsis thaliana | P51102 | - | 702000 | - | - | ? |
| dihydromyricetin + NADPH + H+ | leucodelphinidin + NADP+ | | Arabidopsis thaliana | P51102 | - | 702000 | - | - | ? |
| dihydroquercetin + NADPH + H+ | leucocyanidin + NADP+ | | Arabidopsis thaliana | P51102 | - | 702000 | - | - | ? |
| additional information | ? | - | Arabidopsis thaliana | P51102 | does not catalyze naringenin | 702000 | - | - | ? |

△ top print hide Go to Cofactor Search

| COFACTOR | ORGANISM | UNIPROT | COMMENTARY | LITERATURE | IMAGE |
|---|---|---|---|---|---|
| NADPH | Arabidopsis thaliana | P51102 | - | 702000 | |

△ top print hide 4 entries Go to Specific Activity Search

| SPECIFIC ACTIVITY [µmol/min/mg] | ORGANISM | UNIPROT | COMMENTARY | LITERATURE |
|---|---|---|---|---|
| 0.00002 | Arabidopsis thaliana | P51102 | with dihydrokaempferol as substrate | 702000 |
| 0.0006 | Arabidopsis thaliana | P51102 | with eriodictyol as substrate | 702000 |
| 0.001 | Arabidopsis thaliana | P51102 | with dihydromyricetin as substrate | 702000 |
| 0.0013 | Arabidopsis thaliana | P51102 | with dihydroquercitin as substrate | 702000 |

△ top print hide Go to Organism Search

| ORGANISM | COMMENTARY | LITERATURE | UNIPROT | SEQUENCE DB | SOURCE |
|---|---|---|---|---|---|
| Arabidopsis thaliana | - | 702000 | P51102 | UniProt | BRENDA |

△ top print hide Go to General Information Search

| GENERAL INFORMATION | ORGANISM | UNIPROT | COMMENTARY | LITERATURE |
|---|---|---|---|---|
| physiological function | Arabidopsis thaliana | P51102 | DFR plays a key role in determining intensity and pigment coloration because its specificity and activities dictate the type and amount of the colorless leucoanthocyanidins | 702000 |

△ top print hide Go to AA Sequence and Transmembrane Helices Search Go to Localization Prediction

| UNIPROT | ENTRY NAME | ORGANISM | NO. OF AA | NO. OF TRANSM. HELICES | MOLECULAR WEIGHT [Da] | SOURCE | SEQUENCE | LOCALIZATION PREDICTION ■ |
|---|---|---|---|---|---|---|---|---|
| P51102 pBLAST | DFRA_ARATH | Arabidopsis thaliana | 382 | 0 | 42775 | Swiss-Prot | Show Sequence | other Location (Reliability: 5) |

△ top print hide Go to Cloned (commentary) Search

| CLONED (Commentary) | ORGANISM | UNIPROT | LITERATURE |
|---|---|---|---|
| into pTrcHis2-TOPO and heterologously expressed in Escherichia coli TOP10F' strain. DFR cDNA cloned into pRSF-FHT and inserted into Escherichia coli BL21Star to create E-color strain | Arabidopsis thaliana | P51102 | 702000 |

△ top print hide References Search

| REF. | AUTHORS | TITLE | JOURNAL | VOL. | PAGES | YEAR | ORGANISM (UNIPROT) | PUBMED ID | SOURCE |
|---|---|---|---|---|---|---|---|---|---|
| 702000 | Leonard, E.; Yan, Y.; Chemler, J.; Matern, U.; Martens, S.; Koffas, M. | Characterization of dihydroflavonol 4-reductases for recombinant plant pigment biosynthesis applications | Biocatal. Biotransform. | 26 | 243-251 | 2008 | Fragaria x ananassa (O22617), Ipomoea nil (O24607), Arabidopsis thaliana (P51102), Rosa hybrid cultivar (Q41158), Lilium sp. (Q6UAQ7), Anthurium andraeanum (Q84LZ2) | - | BRENDA |

Select items on the left to see more content.

△ top print hide 11 entries

| EXTERNAL LINKS (specific for EC-Number 1.1.1.219) |
|---|
| ExplorEnz (official portal for IUBMB Enzyme Nomenclature) |
| Expasy Enzyme Nomenclature Database |
| KEGG |
| MetaCyc |
| SABIO-RK |
| NCBI: PubMed, Protein, Nucleotide, Structure, Gene, OMIM |
| Enzyme Nomenclature (alternative site) |
| UniProt |
| PDB |
| PROSITE Database of protein families and domains |
| InterPro (database of protein families, domains and functional sites) |



https://www.brenda-enzymes.org/

UNIVERSITÄT BONN

# What if you need more details about a gene?

- Publications:
  - PubMed
  - GoogleScholar

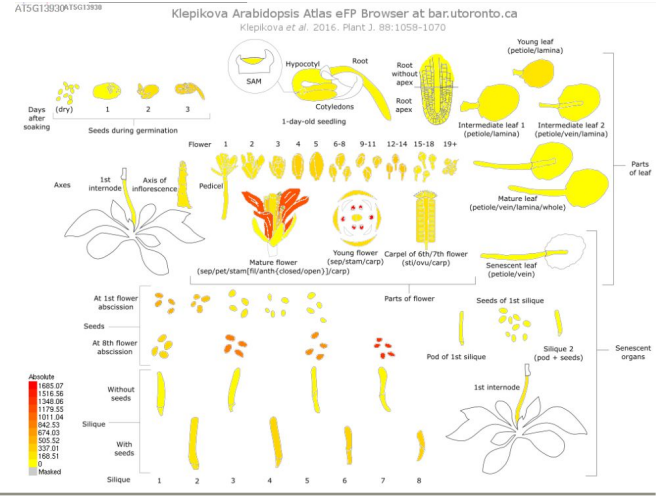
- Organism specific databases:
  - TAIR (*Arabidopsis thaliana*)
  - BananaGenomeHub (*Musa acuminata*)
  - TomatoGenome (*Solanum lycopersicum*)

- Genome browsers

UNIVERSITÄT BONN

# TAIR

- Information about the gene function (annotation text, GO terms, …)
- Information about gene structure including splice variants
- Information about gene expression (eFP browser)
- Knock-out lines of a gene of interest
- External links to various databases & references to corresponding publications

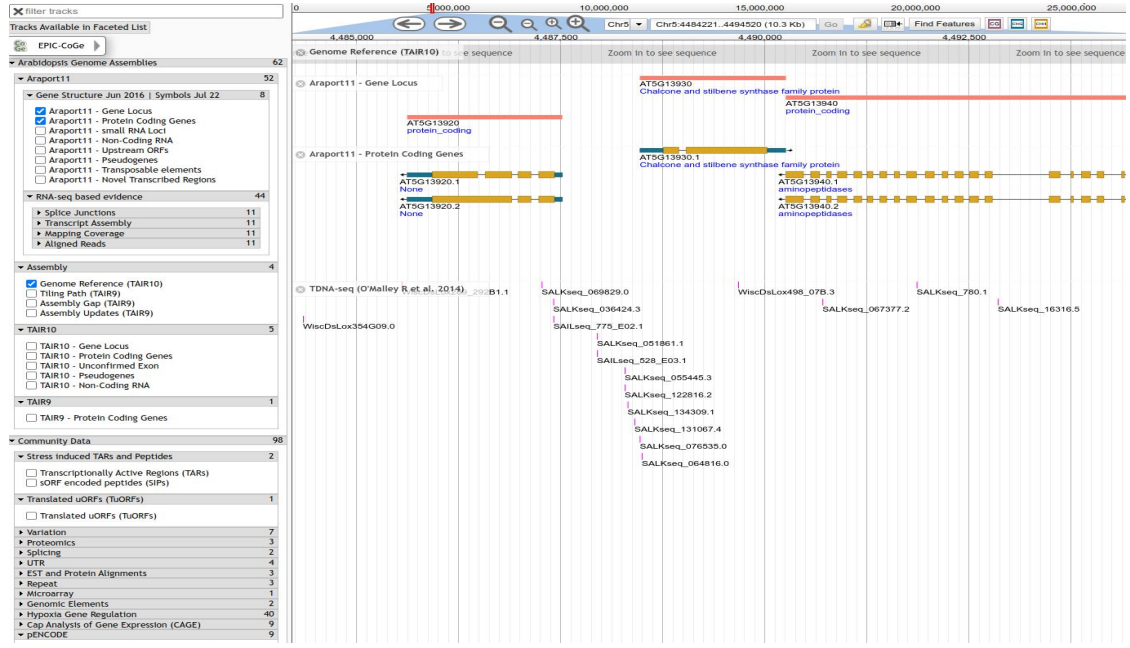





https://www.arabidopsis.org/servlets/TairObject?type=locus&name=At5g13930

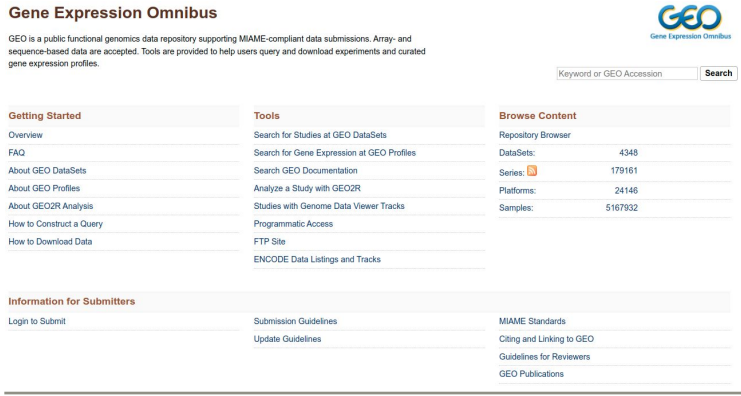
# Genome browser (jbrowse)

- Genome browsers enable interactive inspection of gene structures
- Feature tracks can be selected/deselected
- Functional annotation and related information can be displayed



# Where would you look for gene expression data?

- Electronic Fluorescent Pictograph (eFP)
- Gene Expression Omnibus (GEO)
- ArrayExpress







http://bar.utoronto.ca/; https://www.ncbi.nlm.nih.gov/geo/; https://www.ebi.ac.uk/arrayexpress/

# Large scale gene expression analysis

identify RNAseq samples @ SRA

identify genome/ transcriptome assembly

download reads via fastq-dump

download CDS/ transcript sequences

split-read mapping

run kallisto to quantify transcript abundances

gene prediction

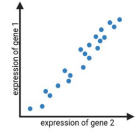
generation of count tables

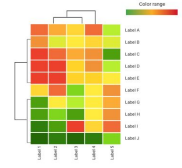
| | s1 | s2 | s3 | s4 |
|---|---|---|---|---|
| g1 | 12 | 3 | 4 | 9 |
| g2 | 32 | 3 | 0 | 5 |
| g3 | 17 | 5 | 4 | 1 |
| g4 | 41 | 2 | 0 | 9 |
| g5 | 13 | 3 | 7 | 0 |

co-expression analysis

DESeq2 analysis

construction of heatmaps



UNIVERSITÄT BONN

# Where to deposit large datasets?

UNIVERSITÄT BONN

# Dryad

- Any file format / data type is accepted
- Support for best practice by curators
- Compliance with funder or publisher requirements
- Nonprofit membership organization
- Dryad's Data Publishing Charges (DPCs): $120 per submission; additional $50/10GB exceeding 50GB
- Content published under CC0 to facilitate reuse

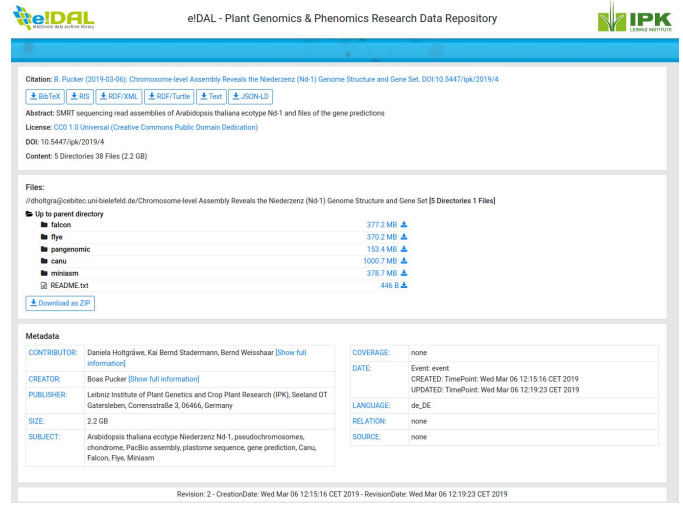



https://datadryad.org/stash

UNIVERSITÄT BONN

# e!DAL (IPK database)

- Hosted at the IPK
- Data sets are released under a CC0 license to enable reuse
- No restrictions with respect to file formats
- Support for review process



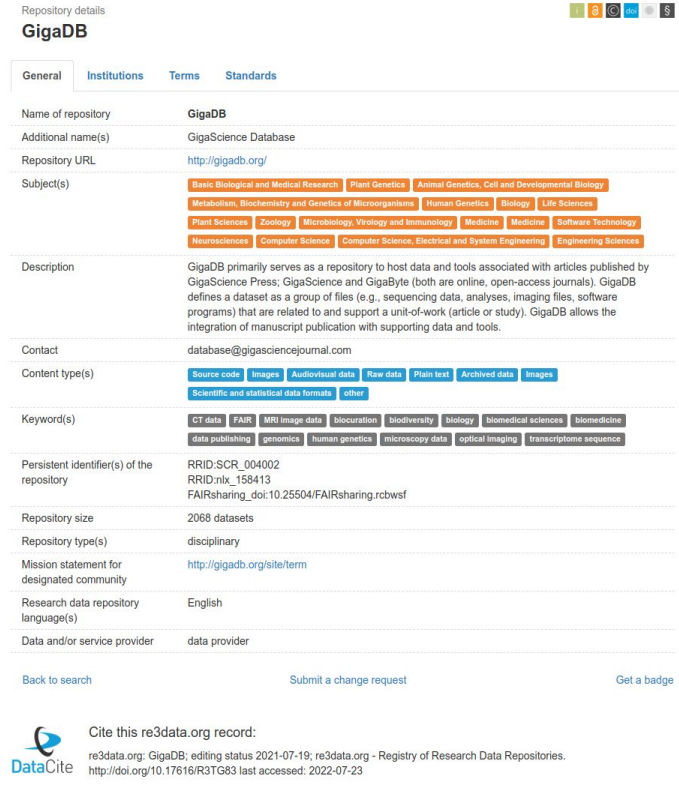



Pucker et al., 2019: 10.5447/ipk/2019/4

UNIVERSITÄT BONN

# GigaDB

- Publication of data sets belonging to GigaScience articles
- CC0 to enable future studies
- Data sets are citeable by DOIs



Repository details

**GigaDB**

General | Institutions | Terms | Standards

| | |
|---|---|
| Name of repository | **GigaDB** |
| Additional name(s) | GigaScience Database |
| Repository URL | http://gigadb.org/ |
| Subject(s) | Basic Biological and Medical Research; Plant Genetics; Animal Genetics, Cell and Developmental Biology; Metabolism, Biochemistry and Genetics of Microorganisms; Human Genetics; Biology; Life Sciences; Plant Sciences; Zoology; Microbiology, Virology and Immunology; Medicine; Medicine; Software Technology; Neurosciences; Computer Science; Computer Science, Electrical and System Engineering; Engineering Sciences |
| Description | GigaDB primarily serves as a repository to host data and tools associated with articles published by GigaScience Press; GigaScience and GigaByte (both are online, open-access journals). GigaDB defines a dataset as a group of files (e.g., sequencing data, analyses, imaging files, software programs) that are related to and support a unit-of-work (article or study). GigaDB allows the integration of manuscript publication with supporting data and tools. |
| Contact | database@gigasciencejournal.com |
| Content type(s) | Source code; Images; Audiovisual data; Raw data; Plain text; Archived data; Images; Scientific and statistical data formats; other |
| Keyword(s) | CT data; FAIR; MRI image data; biocuration; biodiversity; biology; biomedical sciences; biomedicine; data publishing; genomics; human genetics; microscopy data; optical imaging; transcriptome sequence |
| Persistent identifier(s) of the repository | RRID:SCR_004002<br>RRID:nlx_158413<br>FAIRsharing_doi:10.25504/FAIRsharing.rcbwsf |
| Repository size | 2068 datasets |
| Repository type(s) | disciplinary |
| Mission statement for designated community | http://gigadb.org/site/term |
| Research data repository language(s) | English |
| Data and/or service provider | data provider |

Back to search | Submit a change request | Get a badge

DataCite

Cite this re3data.org record:

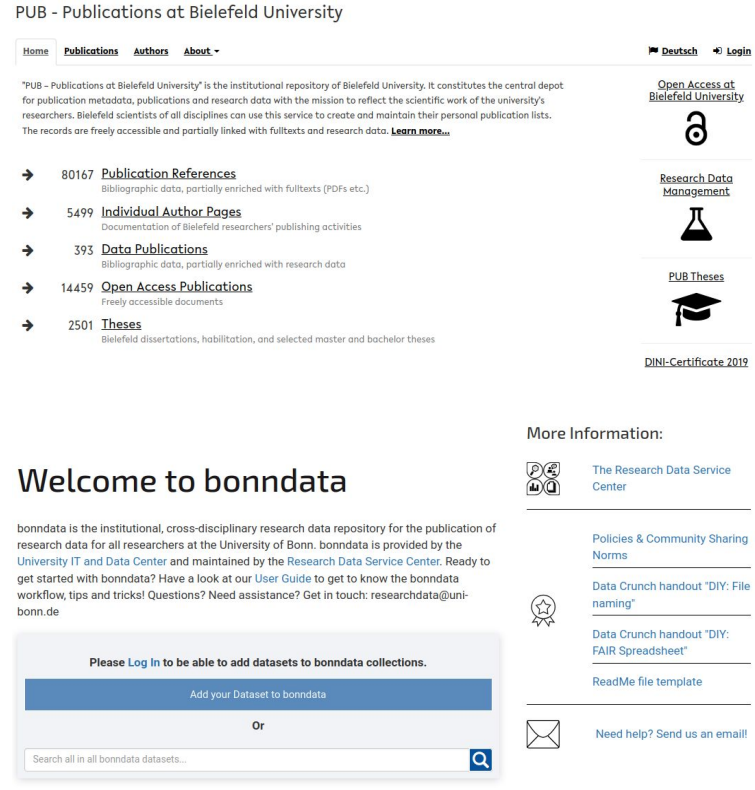
re3data.org: GigaDB; editing status 2021-07-19; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R3TG83 last accessed: 2022-07-23



https://www.re3data.org/repository/r3d100010478

# Institute-specific data publications

- Data publications can be hosted a libraries
- Assignment of DOIs
- PUB at Bielefeld University and bonndata at Uni Bonn






https://pub.uni-bielefeld.de/
https://bonndata.uni-bonn.de/

UNIVERSITÄT BONN

# Data Reuse

UNIVERSITÄT BONN

# Pros and cons of re-use?

- Cost-effective
- Immediately available
- Extremely large datasets

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

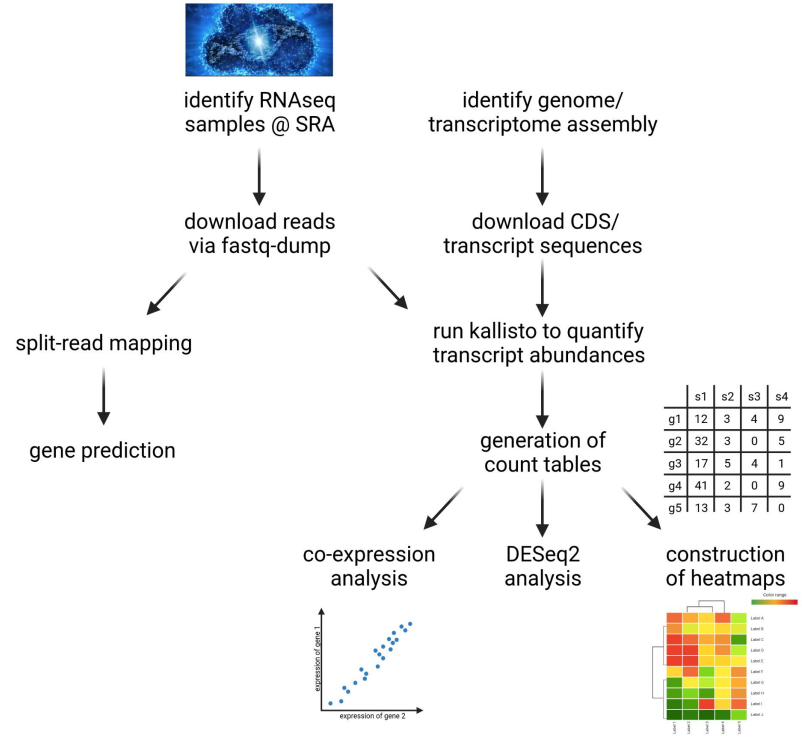
UNIVERSITÄT BONN

# 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

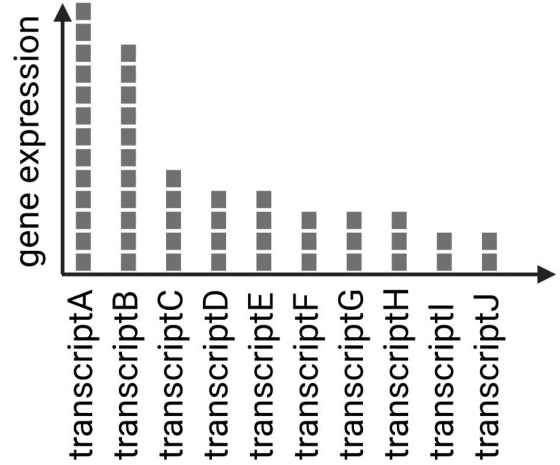


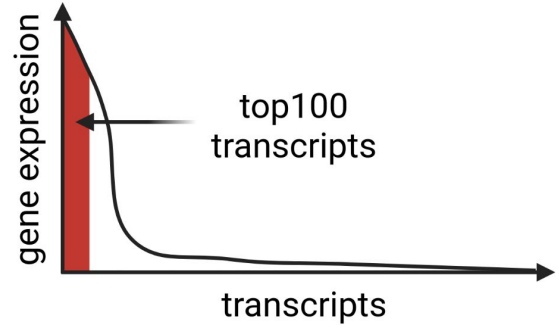
UNIVERSITÄT BONN

# 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



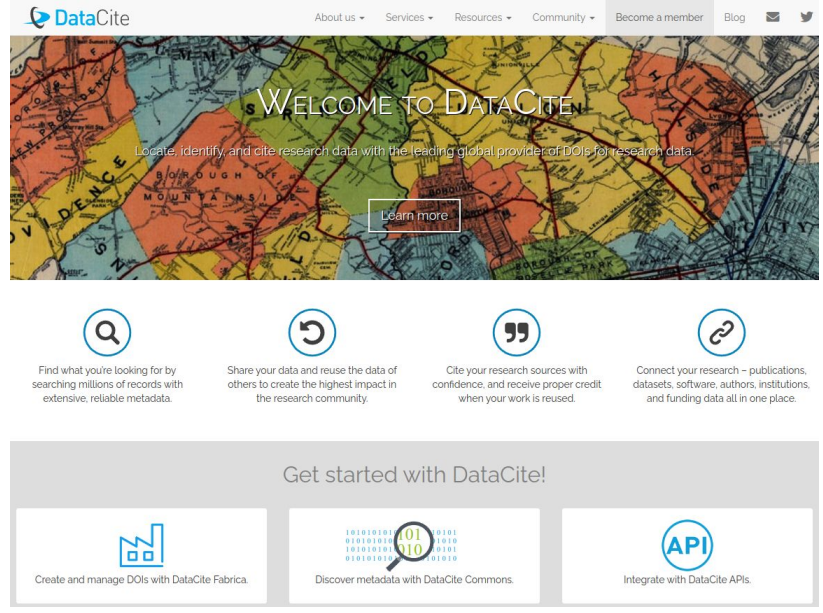



UNIVERSITÄT BONN

# DataCite

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

- Enable connection and reuse of data sets






https://datacite.org/index.html

UNIVERSITÄT BONN

# 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
  - EndNote: commercial

UNIVERSITÄT BONN

# Data reuse examples

UNIVERSITÄT BONN

# 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

UNIVERSITÄT BONN

# 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

UNIVERSITÄT BONN

# Pangenome of hexaploid bread wheat

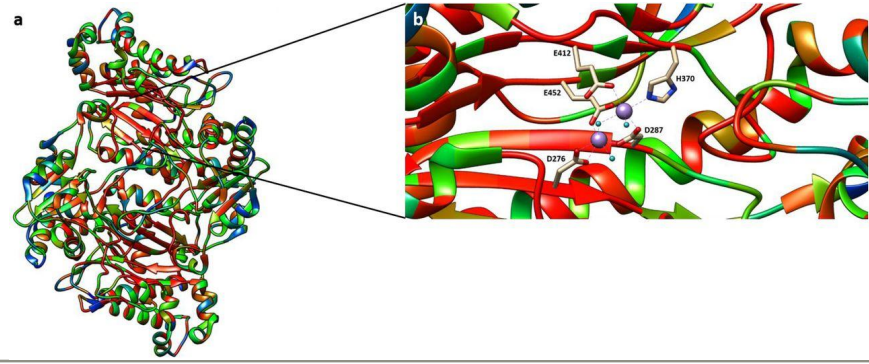
- Pangenome = combination of all genomes of a species
- Integration of public data sets to identify presence/absence of genes in wheat cultivars
- Power of pangenomes increases with number of analyzed species



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

# 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

UNIVERSITÄT BONN

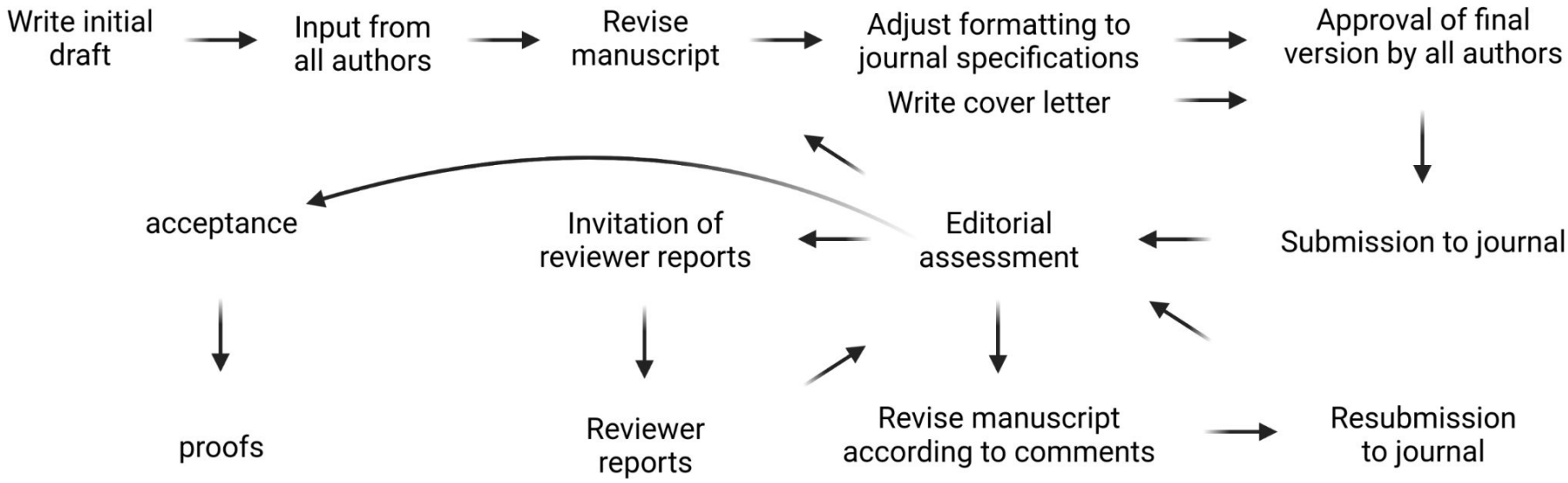
# Scientific Publication Business

UNIVERSITÄT BONN

# Motivation

You have exciting findings that you would like to share with the scientific community or the whole world!

# How to publish a paper - publication process



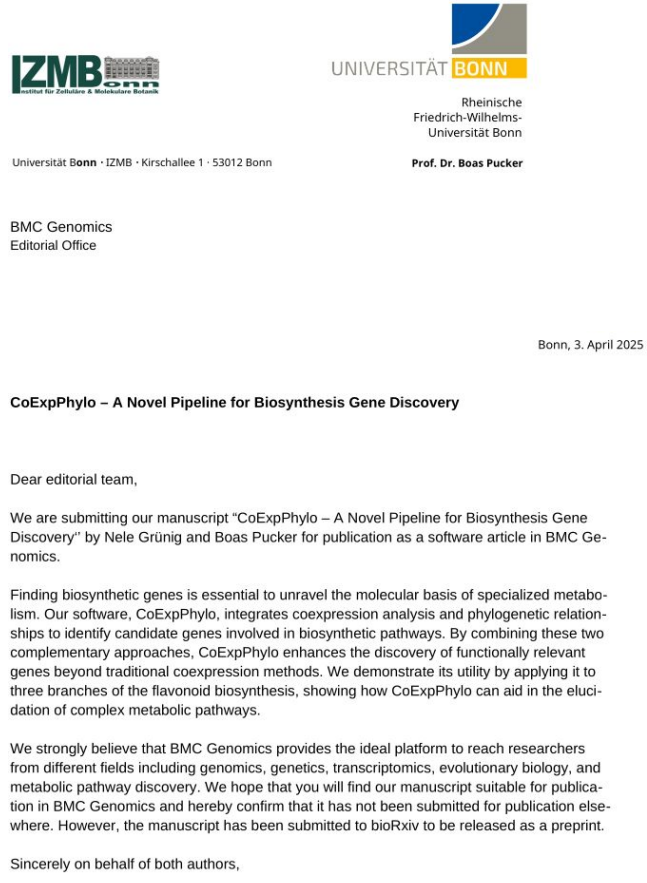
UNIVERSITÄT BONN

# Authorship

- COPE: specifies criteria for authors (https://publicationethics.org/)
- Substantial contribution to
  - (1) data generation / analysis and
  - (2) data interpretation, and
  - (3) manuscript writing
- Order of authors indicates contribution to study:
  - First is best
  - Last is supervision of work + correspondence
- Example 1:
  - Head of an institute does not qualify (ghost authorship)
- Example 2:
  - Technician/undergrad contributing without scientific input does not qualify
- Example 3:
  - Student developing novel method, generating useful data, and contributing to the manuscript does qualify

# Cover letter

- State title of the submission
- State article type (and special issue)
- Summarize the content of the submission
- Convince the editor of relevance
- State that the work has not been published (include link to preprint)
- Indicate previous interactions with journal



IZMB Bonn
Institut für Zelluläre & Molekulare Botanik

UNIVERSITÄT BONN

Rheinische
Friedrich-Wilhelms-
Universität Bonn

Universität Bonn · IZMB · Kirschallee 1 · 53012 Bonn

**Prof. Dr. Boas Pucker**

BMC Genomics
Editorial Office

Bonn, 3. April 2025

**CoExpPhylo – A Novel Pipeline for Biosynthesis Gene Discovery**

Dear editorial team,

We are submitting our manuscript "CoExpPhylo – A Novel Pipeline for Biosynthesis Gene Discovery" by Nele Grünig and Boas Pucker for publication as a software article in BMC Genomics.

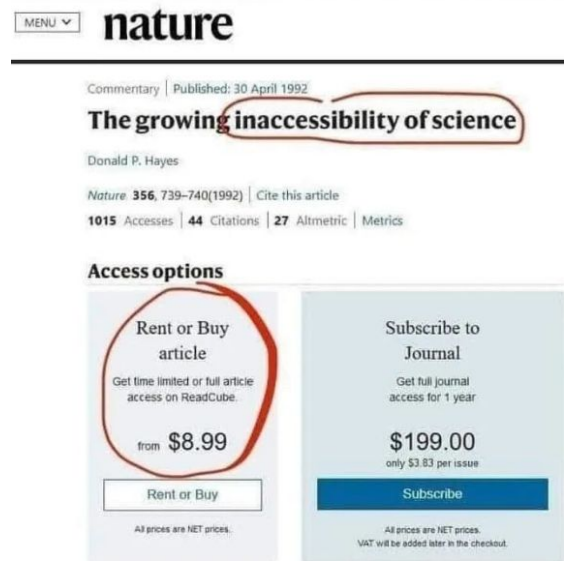
Finding biosynthetic genes is essential to unravel the molecular basis of specialized metabolism. Our software, CoExpPhylo, integrates coexpression analysis and phylogenetic relationships to identify candidate genes involved in biosynthetic pathways. By combining these two complementary approaches, CoExpPhylo enhances the discovery of functionally relevant genes beyond traditional coexpression methods. We demonstrate its utility by applying it to three branches of the flavonoid biosynthesis, showing how CoExpPhylo can aid in the elucidation of complex metabolic pathways.

We strongly believe that BMC Genomics provides the ideal platform to reach researchers from different fields including genomics, genetics, transcriptomics, evolutionary biology, and metabolic pathway discovery. We hope that you will find our manuscript suitable for publication in BMC Genomics and hereby confirm that it has not been submitted for publication elsewhere. However, the manuscript has been submitted to bioRxiv to be released as a preprint.

Sincerely on behalf of both authors,

# Journal types

- Subscription: articles are paywalled
  - Readers have to pay for access
- Hybrid:
  - Authors can pay to get article open
  - Readers have to pay for most articles
- Open Access:
  - Articles are freely accessible
  - Author pay publication fees






http://people.soc.cornell.edu/dph1/lexical-analysis/schoolbooks/Papers/HayesGrowingInaccessibility1992.pdf

UNIVERSITÄT BONN

# Data set publications

- Publications describing data sets without any novel insights
- Dedicated journals were established years ago, but many other are now happy to take everything ‘scientifically solid’
- Classic journals:
  - BMC Research Notes
  - Genome Announcements in many journals
- Institute specific repositories

UNIVERSITÄT BONN

# Preprints

- bioRxiv: https://www.biorxiv.org/ (Cold Spring Harbor Laboratory)
  - Recommended for original research
  - Does not accept reviews

- Preprints.org: https://www.preprints.org/ (operated by MDPI)
  - Best place for reviews
  - Suggested by MDPI journals (and connected to them)

- ResearchSquare: https://www.researchsquare.com/ (supported by Springer Nature)
  - Should be avoided; suggested by Springer Nature journals

UNIVERSITÄT BONN

# Reviewers

- Reviewers are picked based on qualifications and availability
- Qualifications = previous publications on a similar topic
- Lower quality journals have trouble finding reviewers (unqualified reviewers accepted)

- Review activity is documented via ORCID or Publons
- Some journals list reviewers and editors on publications (e.g. Frontiers)

- Getting review invitations is challenging for young scientists
- International visibility is important
- MDPI allows researchers to apply as reviewer for submissions
- Alternative reviewers can be suggested when rejecting an invitation

UNIVERSITÄT BONN

# Review structure

- Summary of the submitted manuscript
- Evaluation of novelty and overall quality
- Major issues
  - Important topics/references missing in the introduction
  - Constructive comments about invalid methods
  - List missing controls
  - Point out cases of overstated conclusions
  - …
- Minor issues (include line numbers)
  - Typos
  - Language issues

UNIVERSITÄT BONN

# Rebuttal letter

- Authors write response to editorial decision
- Response to individual reviewer comments
- Authors' response should include references to the revised manuscript with line numbers
- Good rebuttal letter keeps reviewers from reading the manuscript again

**#Reviewer 1**

Thank you very much for assessing our work.

Hard to understand. Significance of statement not clear to me. How is the roles of plants in complex specialized metabolism [I don't understand this part of the sentence already] dependent on information about the biochemical function of a gene product?

Thank you for pointing out this unclear section. We agree that the sentence was unclear due to a miswording. We corrected "roles of plants" to "roles of genes", which was the intended meaning. The correct sentence now reads: "In plants, understanding the roles of genes in the complex specialized metabolism is a particular challenge and heavily dependent on information about the biochemical functions of gene products."

Ok. I think now I get it. You start with a product in mind, then try to identify the genes involved in production of the product. I suggest to mention this in the beginning, e.g. "A typical approach of …"

Thank you for your helpful comment. We included this clarification in the Background section: "Starting from minimal prior knowledge about some genes involved in the biosynthesis of a metabolite of interest, CoExpPhylo enables the identification of candidate genes associated with the biosynthesis of that metabolite."

There are several others. Check the citations of one of the cluster finding tools that exist that do not use transcriptomics.

We have added further examples of known biosynthesis gene clusters in plants and revised the sentence for improved clarity. It now reads: "Notable exceptions are biosynthesis gene clusters, including those reported for withanolide biosynthesis in *Solanaceae*, noscapine biosynthesis in *Papaver somniferum,* or terpenoid biosynthesis pathways such as thalianol, marneral, tirucalladienol, and arabidol in *Arabidopsis thaliana*, among a limited number of other known cases [11-14]."

Unclear why this decision was made. Typically there is a primary transcript and sometimes several other, secondary transcripts. It is not the case, that the primary transcript is the longest one in general.

For many species, there is no clear definition of the primary transcript among all provided transcript isoforms in the provided files. To reduce redundancy and computational complexity, we selected the longest coding sequence (CDS), not the longest transcript. The longest transcript could be the result of an artificially long UTR, but it is very unlikely that the longest



Full peer review history: https://doi.org/10.1186/s12864-025-12061-3

# Article Processing Charges

- DEAL: funding for all scholars at German research institutes for Springer Nature (except NatureComms + ScientificReports) and Wiley
- Acknowledgement statement should be included about Project DEAL
- Funding for publications through DFG and other third party funded projects

### What do I as an author need to do?

- ☑ Log into the authors' portal for the respective publisher, Wiley, Springer Nature, or starting on 1st January 2024 also Elsevier. Submit your article. Enter "University of Bonn" or "University Hospital Bonn" as your **affiliation** (also on the PDF of your article, see also the guidelines of the University of Bonn for the standardized indication of affiliation in scientific publications).

  Whenever possible, use your institutional email address (@uni-bonn.de or @ukbonn.de) for correspondence and in the publisher's submission platform as the "submitting corresponding author."

- ☑ You will be notified that your institution is a signatory to an agreement with the publisher. You can now make your article open access.

- ☑ We check all articles for eligibility and the affiliation to the University of Bonn is confirmed or rejected by us as an institution. Please take note of the **exceptions** that apply (see below).

- ☑ If you qualify, your publication fees will be billed to us, so **you should not receive an invoice from the publisher** for any APCs yourself.



https://www.open-access.uni-bonn.de/en/funding/project-deal

UNIVERSITÄT BONN

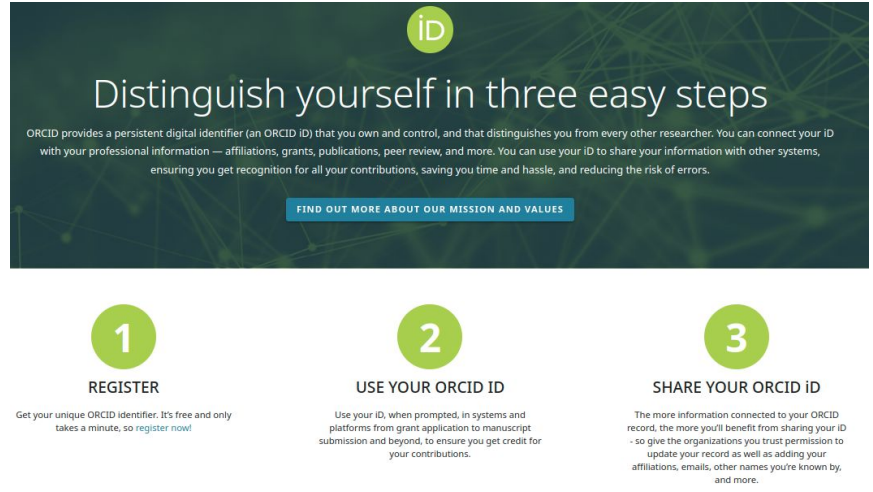
# Do you know licences?

- CC0
- CC BY
- CC BY-SA
- CC BY-NC
- CC BY-ND
- CC BY-NC-SA
- CC BY-NC-ND-SA

UNIVERSITÄT BONN

# ORCID

- ORCID = Open Researcher and Contributor ID
- Unique identifier for researchers (should be included in publications)
- ORCID can be used as SSO on many websites






https://orcid.org/

UNIVERSITÄT BONN

# Data availability

- Open access journals require freely available data sets

- Established data repositories need to be used
  - Dryad
  - Zenodo

- Scripts have to be shared through suitable repositories
  - GitHub (codeberg)
  - Bitbucket

# Are journals still relevant?

- Original function of journals is to share research with the community

- Research can be shared via preprints and through social media

- Article impact is measure by Altmetrics (alternative metrics):
  - (Twitter)
  - BlueSky
  - Blogs & news outlets
  - Reddit
  - CitationTools

UNIVERSITÄT BONN

# Summary

- Data dissemination & ENA submission
- OpenScience & FAIR data
- Compression and transfer
- Databases & services
- Depositing large data sets
- Data reuse
- Examples of large scale data reuse

UNIVERSITÄT BONN

# Time for questions!

UNIVERSITÄT BONN

# Literature


- de Oliveira, J. A. V. S.; Choudhary, N.; Meckoni, S. N.; Nowak, M. S.; Hagedorn, M.; Pucker, B. (2025). Cookbook for Plant Genome Sequences. doi: 10.20944/preprints202508.1176.v2.

- Wolff, K.; Friedhoff, R.; Schwarzer, F.; Pucker, B. (2023). Data Literacy in Genome Research. Journal of Integrative Bioinformatics, 2023, pp. 20230033. doi: 10.1515/jib-2023-0033.

- Pucker B, Irisarri I, de Vries J and Xu B (2022). Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. Quantitative Plant Biology, 3, E5. doi: 10.1017/qpb.2021.18.