

Prof. Dr. Boas Pucker

Long Read Genomics

- Submitting Sequencing Data & Reusing Public Data

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](mailto:pucker[a]uni-bonn.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.

Repetition

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

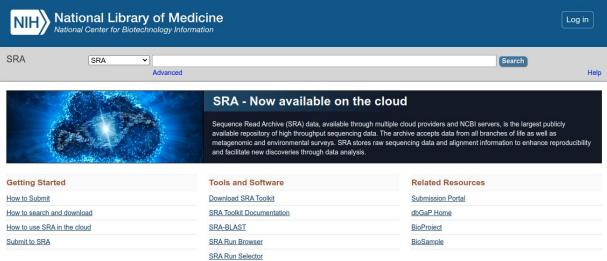
Data Dissemination

Which data of a study can be shared?

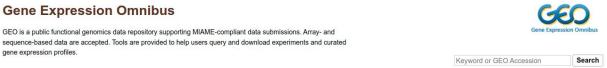
- Sequencing data
- Genome sequence and annotation
- Metadata (phenotyping data, geographic positions, ...)
-
- Remember to conserve the sequenced plant!

Sequencing data sets

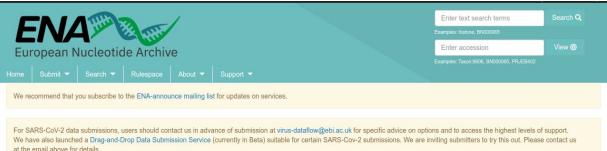
- Sequence Read Archive (SRA)
- Gene Expression Omnibus (GEO)
(also submission to SRA)
- European Nucleotide Archive (ENA)
- Read Selector



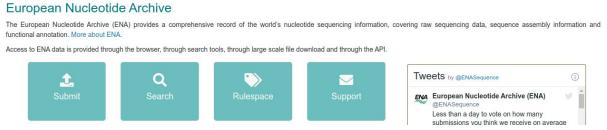
The screenshot shows the SRA homepage with a banner stating "SRA - Now available on the cloud". It includes links for "Getting Started", "Tools and Software", and "Related Resources".



The screenshot shows the GEO homepage with a banner stating "Gene Expression Omnibus". It includes links for "Getting Started", "Tools", and "Browse Content".



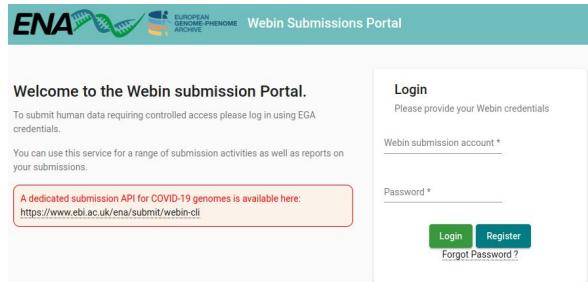
The screenshot shows the ENA homepage with a banner stating "European Nucleotide Archive". It includes links for "Home", "Submit", "Search", "Rulespace", "About", and "Support".



The screenshot shows the ENA footer with a "We recommend that you subscribe to the ENA-announce mailing list for updates on services." message. It also includes a "Tweets by @ENAdatabase" section and links to the ENA Twitter account.

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)



The screenshot shows the ENA Webin Submissions Portal login interface. At the top, there's a logo for ENA (European Nucleotide Archive) with the text "EUROPEAN GENOME-PHENOME ARCHIVE". Below the logo, it says "Webin Submissions Portal". The main area has a teal header with the text "Welcome to the Webin submission Portal." and a note about controlled access using EGA credentials. It also mentions that the service is for submission activities and reports. A red box highlights a link to a dedicated API for COVID-19 genomes. On the right side, there's a "Login" form with fields for "Webin submission account *" and "Password *". Below the password field are "Login" and "Register" buttons, along with a "Forgot Password?" link.

Accession	BioSample	Title
ERS3371290	SAMEA5569268	RNA-Seq Hypertelia bowkeriana flower
ERS3371289	SAMEA5569267	RNA-Seq Hypertelia bowkeriana leaf
ERS3371288	SAMEA5569266	RNA-Seq of Simmondsia chinensis flower
ERS2294062	SAMEA104692579	Corigia litoralis genome sequencing
ERS2294061	SAMEA104692578	Spergula arvensis genome sequencing
ERS2294060	SAMEA104692577	Simmondsia chinensis genome sequencing
ERS2294059	SAMEA104692576	Phaeum exiguum genome sequencing
ERS2294058	SAMEA104692575	Microtes delibis genome sequencing
ERS2294049	SAMEA104692666	Macarthuria australis genome sequencing
ERS2294048	SAMEA104692665	Linum setigerum genome sequencing

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

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

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



The screenshot shows the homepage of the Smithsonian National Museum of Natural History's Botany Collections. It features a header with the Smithsonian logo and navigation links for Visit, Exhibits, Research, Education, Events, About, and Sign Up. Below the header is a search bar titled "Search the Department of Botany Collections". The main content area is titled "Botany Collections" and includes a brief history of the herbarium, stating it began with the acquisition of specimens collected by the U.S. Exploring Expedition (1838-1842). It highlights over 4.2 million specimen records and 115,000 type specimens with images. There are sections for "Search by Keyword", "Search by Field", "Type Register", "Genetic Samples", "Plant Photo Archive", "Botanical Art", "Help", and "Feedback". A large image of a yellow flower is on the right.



The screenshot shows the homepage of the Botanic Garden and Botanical Museum (BGBM) Berlin. It features a header with the BGBM logo and navigation links for Home, Our Work, Membership, News, Resources, About, and Support BGCI. Below the header is a search bar. The main content area is titled "Documentation of specimens and samples" and includes a sub-section "PROJECTS AND CASE STUDIES". It features images of herbarium specimens and a photograph of stacks of specimen boxes. A sidebar on the right provides information about data access policies.

<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>

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

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

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

- 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

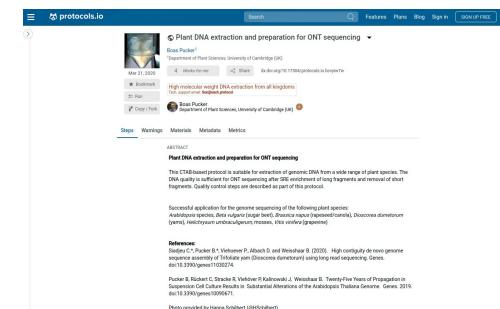
OpenProtocols

- 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



The screenshot shows the main landing page of protocols.io. At the top, there's a navigation bar with links for Features, Plans, Blog, Case study, Sign in, and SIGN UP FREE. Below the header, a large banner says "Bring structure to your research" and "A secure platform for developing and sharing reproducible methods". It has tabs for Biology, Chemistry, Computational workflow, Clinical trials, Operational procedures, Safety checklist, and Instructions/manuals. A prominent orange button says "CREATE A FREE ACCOUNT". Below the banner, there are three main sections: "+ Organize & collaborate" (with sub-points like "Central and secure place to organize up-to-date/reproducible methods with history and concurrent editing", "Explore the editor", and "Edit a protocol"), "Accelerate research" (with sub-points like "Dynamic and interactive methods, notebooks, protocols", "Run a protocol", and "Share reproducible methods"), and "Avoid mistakes" (with sub-points like "Create and discover reproducible experimental and computational methods with video, images, detailed parameters, and units").

<https://www.protocols.io/welcome>

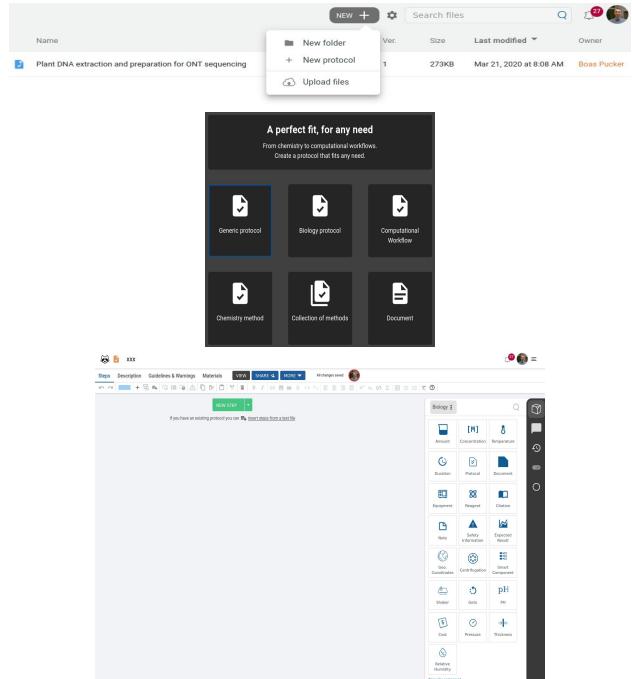


This screenshot shows a detailed view of a protocol titled "Plant DNA extraction and preparation for ONT sequencing" by Boas Pucker. The page includes a preview image, author information (Boas Pucker, Researcher in Plant Sciences, University of Cambridge UK), and a "Copy & Edit" button. The main content area is divided into sections: ABSTRACT, METHODS, and REFERENCES. The ABSTRACT section describes the protocol for extracting genomic DNA from a wide range of plant species. The METHODS section details the workflow, including lysis, proteinase K treatment, phenol-chloroform extraction, and ethanol precipitation. The REFERENCES section lists publications by Boas Pucker and others, such as "High-throughput DNA sequencing after ONT extraction from all teleosts" and "Twenty Five Years of Preparation in Suspension Cell Culture Results in Sustainable Advances of the Arabidopsis Thaliana Genome".

[doi:10.17504/protocols.io.bcvyiw7w](https://doi.org/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)



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

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?

- 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

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)

```
1 {  
2   "authors":  
3   [  
4     {  
5       "firstName": "Boas",  
6       "lastName": "Pucker"  
7     },  
8     {  
9       "firstName": "Iker",  
10      "lastName": "Irisarri"  
11    },  
12    {  
13      "firstName": "Jan",  
14      "lastName": "de Vries"  
15    },  
16    {  
17      "firstName": "Bo",  
18      "lastName": "Xu"  
19    }  
20  ],  
21  "title": "Plant genome sequence assembly in the era of long reads",  
22  "url": "https://doi.org/10.1017/qpb.2021.18",  
23  "doi": "10.1017/qpb.2021.18"  
24 }
```

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 now be completed by a single group 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 long plants. Increases in 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

haplotyping 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
DOI: <https://doi.org/10.1017/qpb.2021.18>

Creative Commons

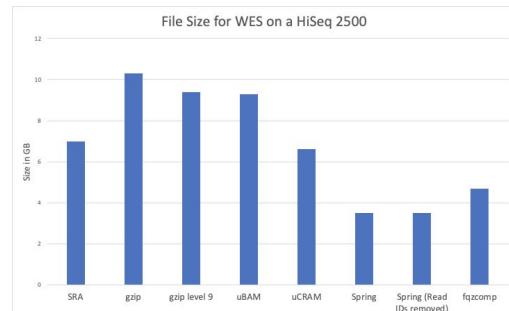
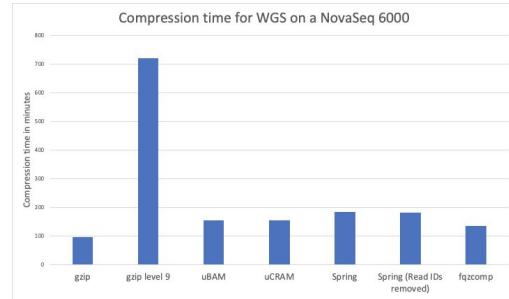
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence
<http://creativecommons.org/licenses/by/4.0/>, which permits unrestricted re-use, distribution, and
reproduction in any medium, provided the original work is properly cited.

Copyright

© The Author(s), 2022. Published by Cambridge University Press in association with The
John Innes Centre

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



- 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`

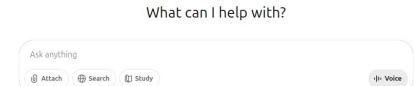
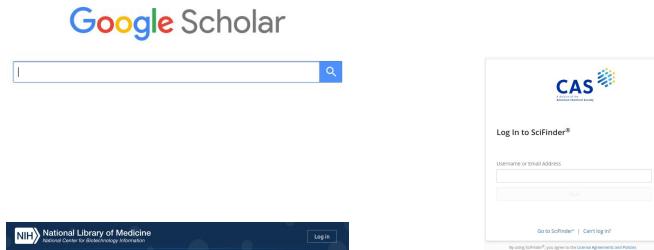
sha256sum & md5sums

- 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

Databases & Services

How to find information in the literature?

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



Which sequence databases do you know?

Sequence databases

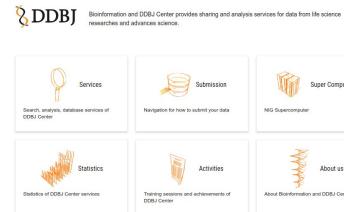
- Sequence Read Archive (SRA)
- European Nucleotide Archive (ENA)
- DNA Data Bank of Japan (DDBJ)
- GenBank (NCBI)
- Joint Genome Institute (JGI) / Phytozome
- PLAZA
- Ensembl genomes



The screenshot shows the SRA section of the NIH homepage. It features a large blue cloud icon with a DNA helix. Below it, there's a "Getting Started" section with links for "How to Submit", "Data Submission and Download", "How to Use SRA in the Cloud", and "Submit to SRA". To the right, there are sections for "Tools and Software", "Related Resources", and "Help". A sidebar on the left lists "DNA Data Bank of Japan", "ENA", and "Biosamples".



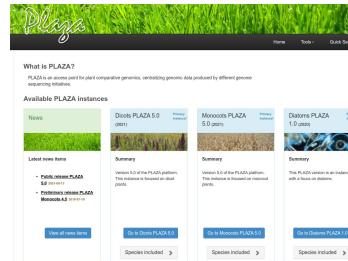
The screenshot shows the ENA homepage. It has a green header with the ENA logo and a search bar. Below the header, there's a "Welcome to ENA" message and a "My account" link. The main content area includes sections for "Data submission", "Search", "Browse", and "Help". A sidebar on the left provides information about ENA's mission and its role in the European Bioinformatics Institute.



The screenshot shows the DDBJ homepage. It features a green header with the DDBJ logo and a search bar. Below the header, there are sections for "Services", "Submission", "Super Computer", "Statistics", "Activities", and "About us". A sidebar on the left lists "Bioinformatics and DDBJ Center", "DDJB", and "DDJB Center".



The screenshot shows the Phytozome homepage. It has a green header with the JGI logo and a search bar. Below the header, there are sections for "Welcome to Phytozome", "Overview", "Release Notes", and "News". A sidebar on the left lists "Joint Genome Institute Archive (JGI)", "Phytozome", and "JGI Data Portal".

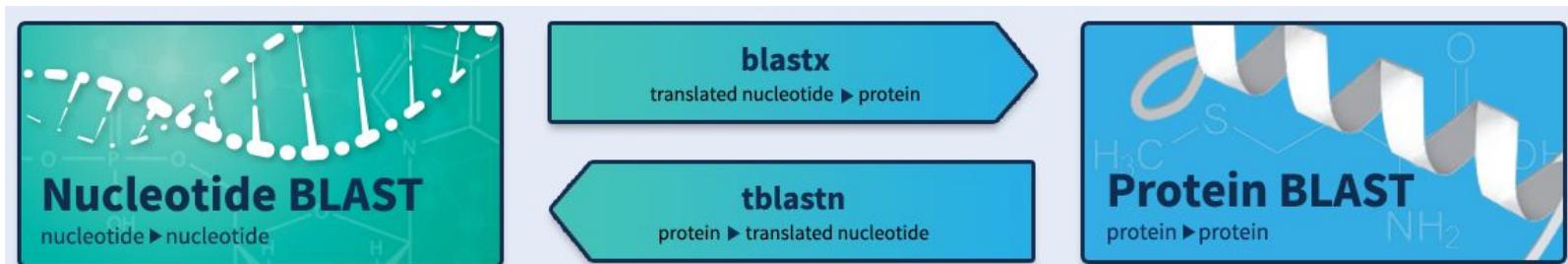


The screenshot shows the PLAZA homepage. It features a green header with the PLAZA logo and a search bar. Below the header, there are sections for "What Is PLAZA?", "Available PLAZA Instances", and "Recent news items". The "Available PLAZA Instances" section shows four instances: "Neuro" (Summary), "Dicots PLAZA 5.0 (2011)" (Summary), "Monocots PLAZA 5.0 (2011)" (Summary), and "Duniova PLAZA 1.0 (2011)" (Summary). Each instance has a "View details" button.

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



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



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 264K](#) [Anaconda.org 2.6.19](#) [downloads 249K total](#) [Citations 5673](#)

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.

Do you know websites with bioinformatic tools?

EBI services

- EBI = European Bioinformatics Institute
- Broad range of different services:
 - DNA&RNA
 - Gene expression
 - Literature
 - Structures

Featured data resources



AlphaFold DB

Database for protein structure predictions for numerous species

CC-BY



BioModels

A repository of peer-reviewed, published, computational models.

[Web API](#) | [CCO](#)



ChEMBL

An open data resource of binding, functional and ADMET bioactivity data.

[Web API](#) | [CC-BY](#)

Featured tools



Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

[Web API](#)



HMMER

Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.

[Web API](#)



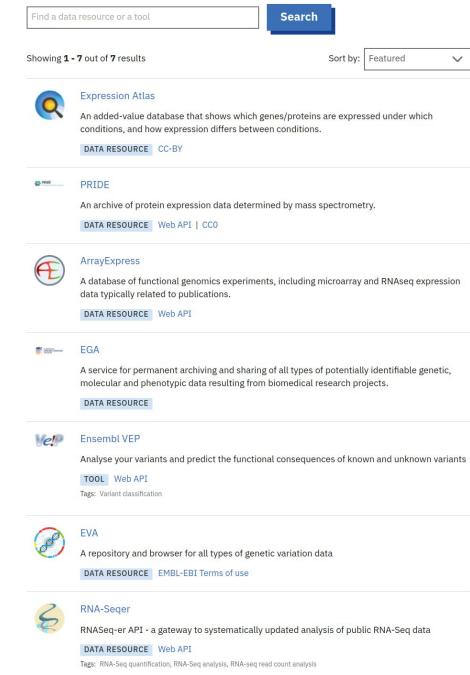
Annotation Platform

Consolidating text-mined and curated annotations

[Web API](#)

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)



The screenshot shows a search results page for 'gene expression' on the EBI website. The search bar at the top contains the query 'gene expression'. Below the search bar, there are filters for 'Refine by Type' (Data resources, Tools) and 'Category' (Chemical biology, Cross domain, DNA & RNA, Gene expression, Ontologies, Proteins, Structures, Systems). The results section displays seven entries:

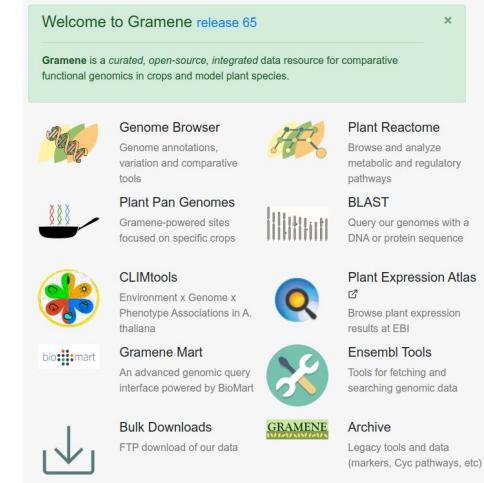
- Expression Atlas**: An added-value database that shows which genes/proteins are expressed under which conditions, and how expression differs between conditions. Status: DATA RESOURCE CC-BY.
- PRIDE**: An archive of protein expression data determined by mass spectrometry. Status: DATA RESOURCE Web API | CCO.
- ArrayExpress**: A database of functional genomics experiments, including microarray and RNAseq expression data typically related to publications. Status: DATA RESOURCE Web API.
- EGA**: A service for permanent archiving and sharing of all types of potentially identifiable genetic, molecular and phenotypic data resulting from biomedical research projects. Status: DATA RESOURCE.
- veP**: Ensembl VEP. Analyse your variants and predict the functional consequences of known and unknown variants. Status: TOOL Web API. Tags: Variant classification.
- EVA**: A repository and browser for all types of genetic variation data. Status: DATA RESOURCE EMBL-EBI Terms of use.
- RNA-Seqr**: RNASeq-er API - a gateway to systematically updated analysis of public RNA-Seq data. Status: DATA RESOURCE Web API. Tags: RNA-Seq quantification, RNA-Seq analysis, RNA-seq read count analysis.

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)



Welcome to Gramene release 65

Gramene is a curated, open-source, integrated data resource for comparative functional genomics in crops and model plant species.

Genome Browser
Genome annotations, variation and comparative tools

Plant Pan Genomes
Gramene-powered sites focused on specific crops

BLAST
Query our genomes with a DNA or protein sequence

CLIMtools
Environment x Genome x Phenotype Associations in *A. thaliana*

bioMart

Gramene Mart
An advanced genomic query interface powered by BioMart

Bulk Downloads
FTP download of our data

Plant Reactome
Browse and analyze metabolic and regulatory pathways

Plant Expression Atlas
Browse plant expression results at EBI

Ensembl Tools
Tools for fetching and searching genomic data

GRAMENE

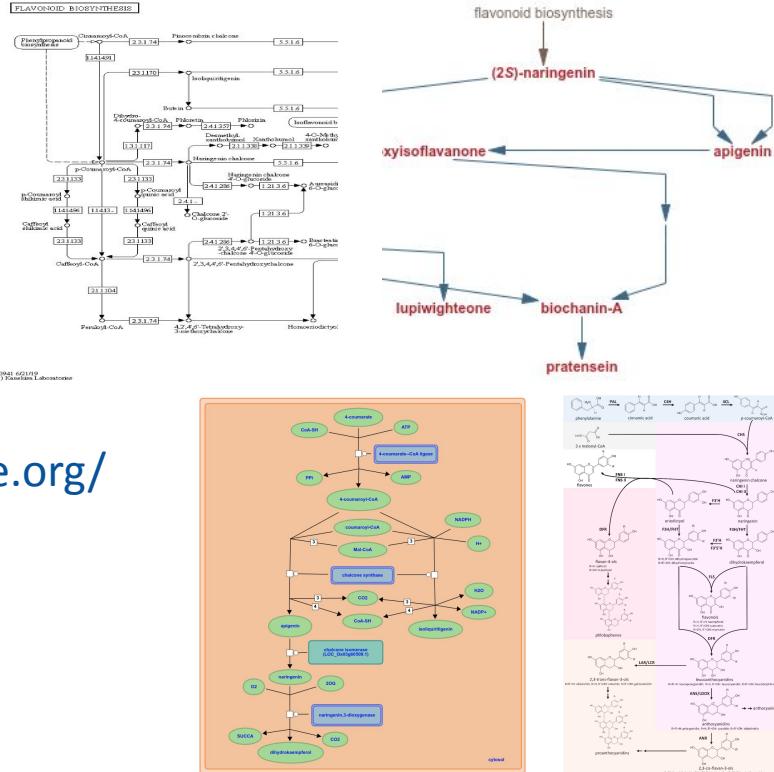
Archive
Legacy tools and data (markers, Cyc pathways, etc)

- 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/>

Where would you look for pathways?

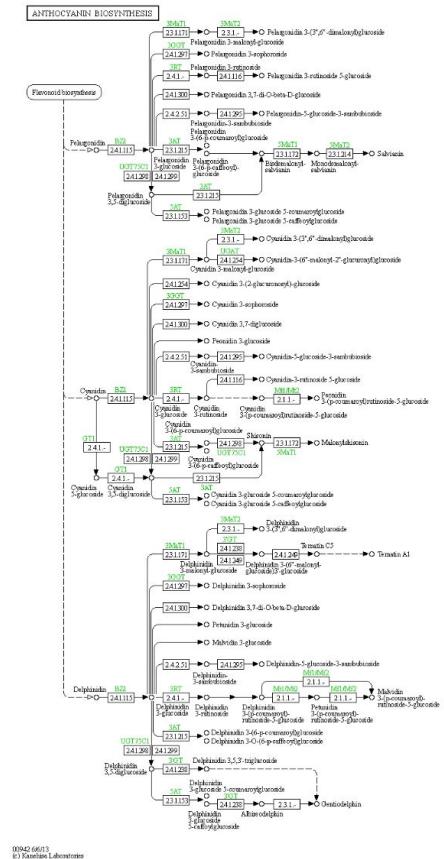
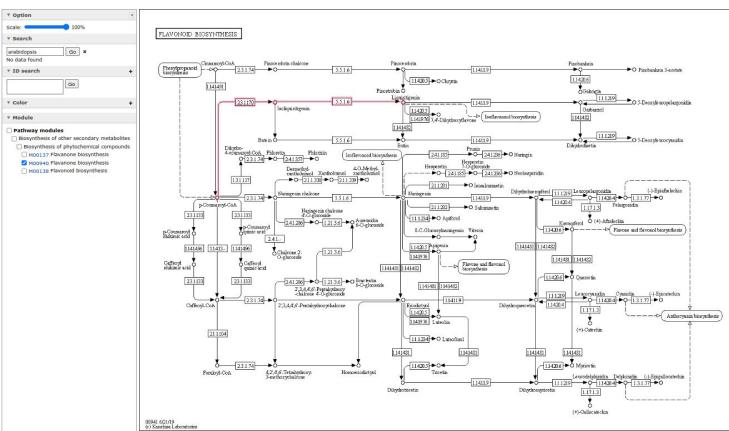
Pathway databases

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



KEGG - Flavonoid biosynthesis

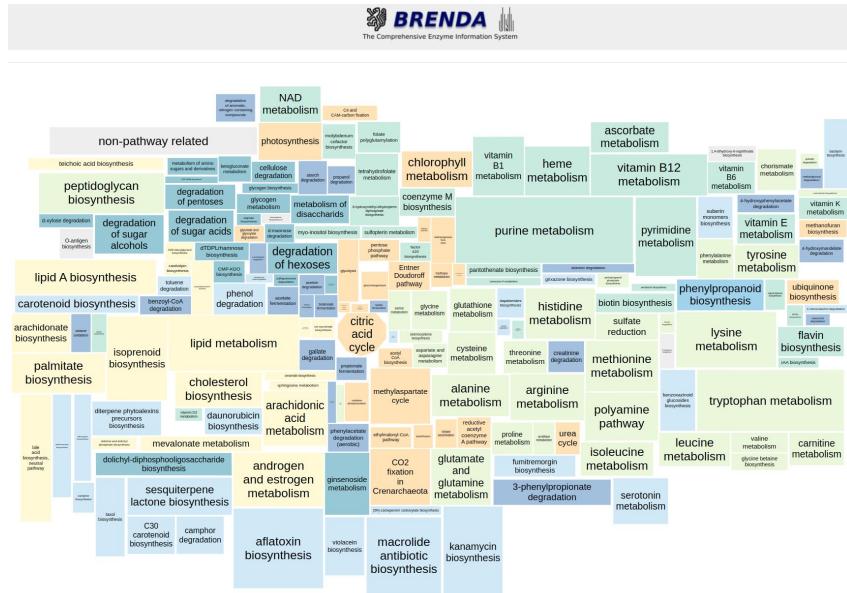
- Selection of one particular species
- Selection of individual branches in pathways
- Gene names are given in addition to EC numbers



Details about enzymes?

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, ...)



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

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, pl value, molecular weight

Go to Substrate/Product Search

SUBSTRATE ▾▼	PRODUCT ▾▼	REACTION DIAGRAM ▾▼	ORGANISM ▾▼	UNIPROT ▾▼	COMMENTARY (Substrate) ▾▼	LITERATURE (Substrate) ▾▼	COMMENTARY (Product) ▾▼	LITERATURE (Product) ▾▼	Reversibility (Product) ▾▼
dihydrokaempferol + NADPH + H ⁺	leucopelargonidin + NADP ⁺	○	Arabidopsis thaliana	P51102	-	702000	-	-	?mod reversible?
dihydromyricetin + NADPH + H ⁺	leucodaphnidin + NADP ⁺	○	Arabidopsis thaliana	P51102	-	702000	-	-	?
dihydroquercetin + NADPH + H ⁺	quercetin + NADP ⁺	○	Arabidopsis thaliana	P51102	-	702000	-	-	?
additional information:	-	-	Arabidopsis thaliana	P51102	does not catalyze naringenin	702000	-	-	?

Go to Cofactor Search

COFACTOR ▾▼	ORGANISM ▾▼	UNIPROT ▾▼	COMMENTARY ▾▼	LITERATURE ▾▼	IMAGE ▾▼
NADPH	Arabidopsis thaliana	P51102	-	702000	○

Go to Specific Activity Search

SPECIFIC ACTIVITY (μmol/min/mg) ▾▼	ORGANISM ▾▼	UNIPROT ▾▼	COMMENTARY ▾▼	LITERATURE ▾▼
0.0002	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 dihydroquercetin as substrate	702000

Go to Organism Search

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

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 resulting leucoanthocyanidins	702000

Go to AA Sequence and Transmembrane Helices Search

UNIPROT ▾▼	ENTRY NAME ▾▼	ORGANISM ▾▼	AA OF AA	NO. OF TRANSM.	AA HELICES ▾▼	MOLECULAR WEIGHT(DA) ▾▼	SOURCE ▾▼	SEQUENCE ▾▼	LOCALIZATION PREDICTION ▾▼	LITERATURE ▾▼
P51102	DFRA_ARATH	Arabidopsis thaliana	382	0	42775	Swiss-Prot	Show Sequence	other Location (Reliability: 5)	702000	

Go to Cloned (Commentary) Search

CLONED (Commentary) ▾▼	ORGANISM ▾▼	LITERATURE ▾▼
into pTrcHisZ-TOPO and heterologously expressed in Escherichia coli TOP10F strain. DFR cDNA cloned into pRSF-HFT and inserted into Escherichia coli BL21 Star to create E-color strain.	Arabidopsis thaliana	P51102 702000

Reference Search

REF. ▾▼	AUTHORS ▾▼	TITLE ▾▼	JOURNAL ▾▼	VOL ▾▼	PAGES ▾▼	YEAR ▾▼	ORGANISM (UNIPROT) ▾▼	PUBMED SOURCE ▾▼
702000	Leonard, E.; Yan, Y.; Chmelir, J.; Maken, J.; Ljung, R.; Koffas, M.	Characterization of dihydroflavonol 4-reductases for recombinant plant pigment biosynthetic approaches	BioCat. Biotransform.	28	243-251	2008	Fragaria x ananassa (O22817), Ipomoea nil (O24697), Arabidopsis thaliana (P51102), Ricinus communis (Q6UQ7), Solanum tuberosum (Q6UAQ7), Anthurium andreanum (Q84L22)	BRENDA

Select Items on the left to see more content.

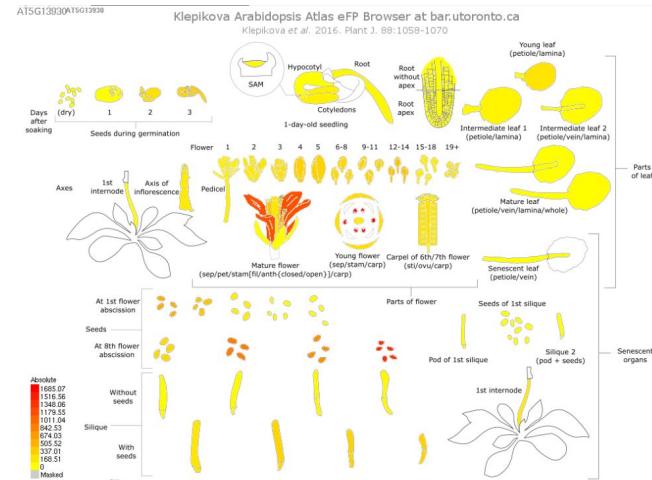
External Links (specific for EC-Number 1.1.1.219)

- ECB (official portal for IUBMB Enzyme Nomenclature)
- Enzyme 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)

What if you need more details about a gene?

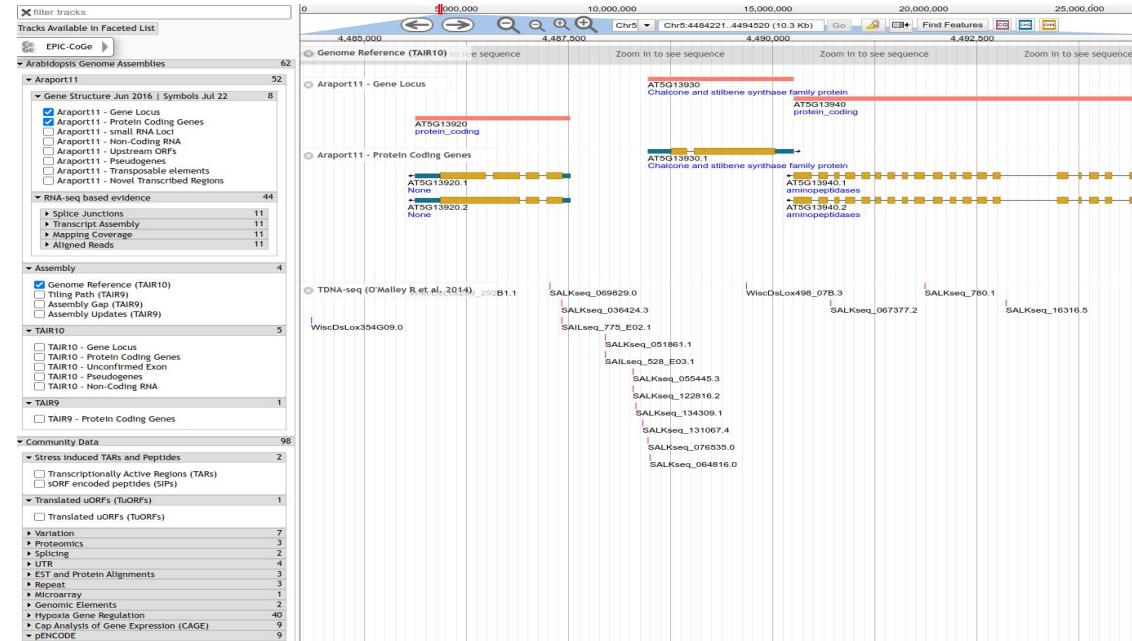
- Publications:
 - PubMed
 - GoogleScholar
- Organism specific databases:
 - TAIR (*Arabidopsis thaliana*)
 - BananaGenomeHub (*Musa acuminata*)
 - TomatoGenome (*Solanum lycopersicum*)
- Genome browsers

- 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



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

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.



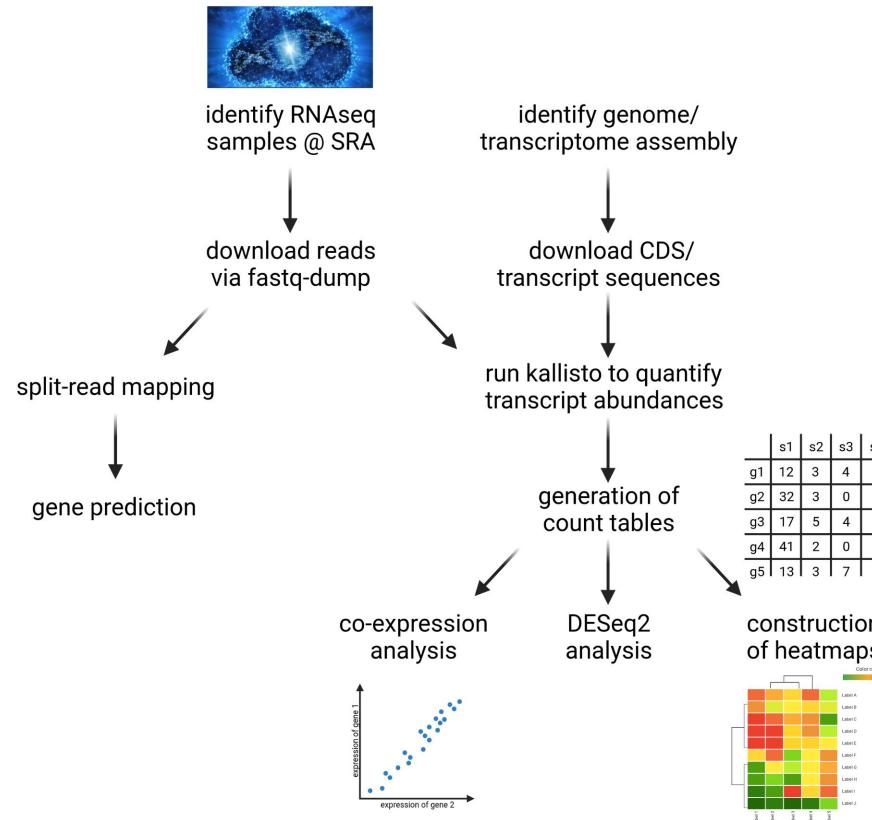
Keyword or GEO Accession Search

Getting Started	Tools	Browse Content
Overview	Search for Studies at GEO DataSets	Repository Browser
FAQ	Search for Gene Expression at GEO Profiles	DataSets: 4348
About GEO DataSets	Search GEO Documentation	Series: 179161
About GEO Profiles	Analyze a Study with GEO2R	Platforms: 24146
About GEO2R Analysis	Studies with Genome Data Viewer Tracks	Samples: 5167932
How to Construct a Query	Programmatic Access	
How to Download Data	FTP Site	
	ENCODE Data Listings and Tracks	

Information for Submitters		
Login to Submit	Submission Guidelines	MIAME Standards
	Update Guidelines	Citing and Linking to GEO
		Guidelines for Reviewers
		GEO Publications



Large scale gene expression analysis



Where to deposit large datasets?

- 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

How it works

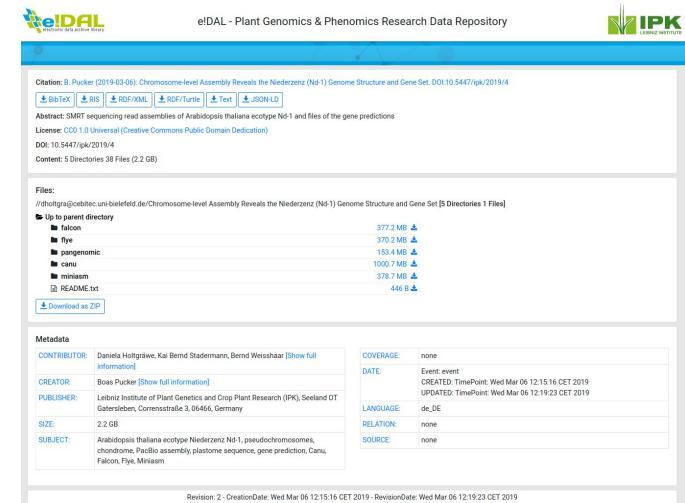
 Login Use your ORCID. If your institution is a Dryad member , connect to your existing credentials.	 Submit Whether or not your data are related to an article, upload your data files and receive a citable DOI.	 Review Our curators will check through your submission to ensure the data are usable. They may contact you with advice or questions.	 Cite Cite and promote your data publication!
--	---	---	---

Why use Dryad?

-  **Any field. Any format.** Submit data in any file format from any field of research. Share all of the data from a project in one place.
-  **Quality control and assistance.** Our curators will check your files before they are released, and help you follow best practices.
-  **Straightforward compliance.** Submit your data to satisfy publisher and funder requirements for preservation and availability with a minimum of effort. We work directly with many publishers -- including Wiley, The Royal Society, and PLOS -- to streamline the process.
-  **Community-led.** Dryad is a nonprofit membership organization that is committed to making data available for research and educational reuse now and into the future. Modest Data Publishing Charges help ensure our sustainability.

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



The screenshot shows the e!DAL data repository interface for the Nederenz N6-1 genome assembly. At the top, there is a header with the e!DAL logo and the text "e!DAL - Plant Genomics & Phenomics Research Data Repository". Below the header, there is a navigation bar with links for BibTeX, RIS, RDF/XML, RDF/Turtle, Text, and JSON-LD. The main content area displays the following information:

- Citation:** B. Pucker (2019-03-06): Chromosome-level Assembly Reveals the Nederenz (N6-1) Genome Structure and Gene Set. DOI:10.5447/ipk/2019/4
- Abstract:** SMRT sequencing read assemblies of *Arabidopsis thaliana* ecotype N6-1 and files of the gene predictions.
- License:** CC0 1.0 Universal (Creative Commons Public Domain Dedication)
- DOI:** 10.5447/ipk/2019/4
- Content:** 5 Directories 38 Files (2.2 GB)

Below this, there is a table of files with their sizes:

File:	Size
Up to parent directory	377.2 MB
falcon	3.7 MB
pangenomic	153.4 MB
canu	1000.7 MB
miniasm	378.7 MB
README.txt	446 B

At the bottom of the page, there is a link to "Download as ZIP".

Metadata

CONTRIBUTOR	Daniela Holzgrafe, Kai Bernd Städemann, Bernd Weißhaar [Show full information]
CREATOR	Boas Pucker [Show full information]
PUBLISHER	Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland OT Gatersleben, Corrensstraße 3, 06466, Germany
SIZE	2.2 GB
SUBJECT	<i>Arabidopsis thaliana</i> ecotype Nederenz N6-1, pseudochromosomes, whole-genome Falcon assembly, plastome sequence, gene prediction, Canu, Falcon, Flye, Miniasm

On the right side, there is a table of metadata fields:

COVERAGE	none
DATE	Event event CREATED: TimePoint: Wed Mar 06 12:15:16 CET 2019 UPDATED: TimePoint: Wed Mar 06 12:19:23 CET 2019
LANGUAGE	de_DE
RELATION	none
SOURCE	none

At the very bottom, there is a note: "Revision: 2 - CreationDate: Wed Mar 06 12:15:16 CET 2019 - RevisionDate: Wed Mar 06 12:19:23 CET 2019"

- 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 | Neuroscience | 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
RRID:nix_158413
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



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

Institute-specific data publications

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

PUB - Publications at Bielefeld University

[Home](#) [Publications](#) [Authors](#) [About](#) [Log in](#)

"PUB - Publications at Bielefeld University" is the institutional repository of Bielefeld University. It constitutes the central depot for publication metadata, publications and research data with the mission to reflect the scientific work of the university's researchers. Bielefeld scientists of all disciplines can use this service to create and maintain their personal publication lists. The records are freely accessible and partially linked with fulltexts and research data. [Learn more...](#)

- 80167 [Publication References](#)
Bibliographic data, partially enriched with fulltexts (PDFs etc.)
- 5499 [Individual Author Pages](#)
Documentation of Bielefeld researchers' publishing activities
- 393 [Data Publications](#)
Bibliographic data, partially enriched with research data
- 14459 [Open Access Publications](#)
Freely accessible documents
- 2501 [Theses](#)
Bielefeld dissertations, habilitation, and selected master and bachelor theses

[Open Access at Bielefeld University](#)



[Research Data Management](#)



[PUB Theses](#)



[DINI-Certificate 2019](#)

More Information:

 The Research Data Service Center

 Policies & Community Sharing Norms

 Data Crunch handbook "DIY: File naming"

 Data Crunch handbook "DIY: FAIR Spreadsheet"

 ReadMe file template

 Need help? Send us an email!

Welcome to bonndata

bonndata is the institutional, cross-disciplinary research data repository for the publication of research data for all researchers at the University of Bonn. bonndata is provided by the University IT and Data Center and maintained by the Research Data Service Center. Ready to get started with bonndata? Have a look at our [User Guide](#) to get to know the bonndata workflow, tips and tricks! Questions? Need assistance? Get in touch: researchdata@uni-bonn.de

Please [Log In](#) to be able to add datasets to bonndata collections.

Add your Dataset to bonndata

Or

Search all in all bonndata datasets... 

Data Reuse

Pros and cons of re-use?

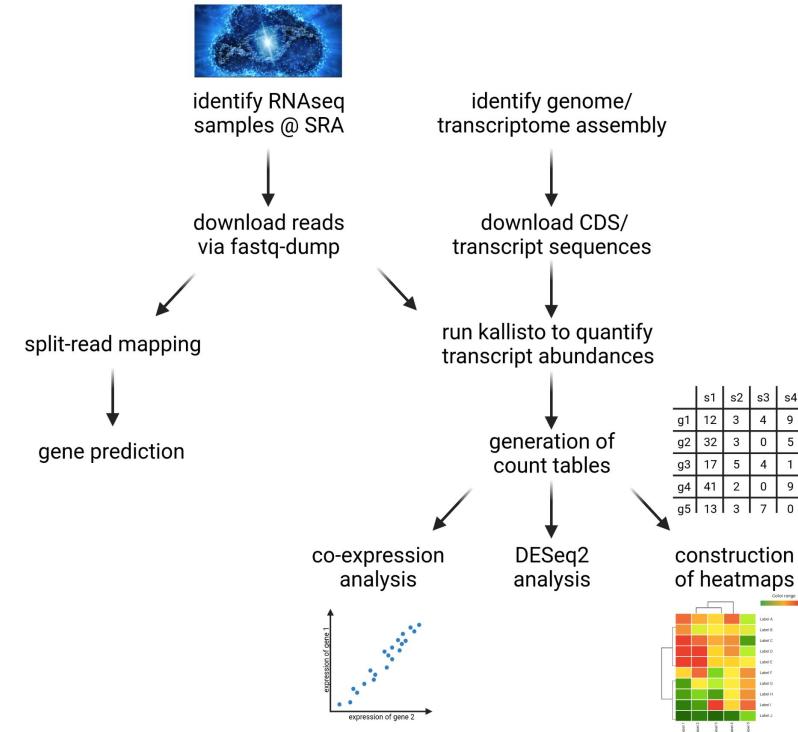
- Cost-effective
- Immediately available
- Extremely large datasets
- Lacking metadata
- Unknown details/issues
- Mislabeling possible
- Not perfectly matching needs
- Technology outdated

Find gene expression data sets

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

How to retrieve data?

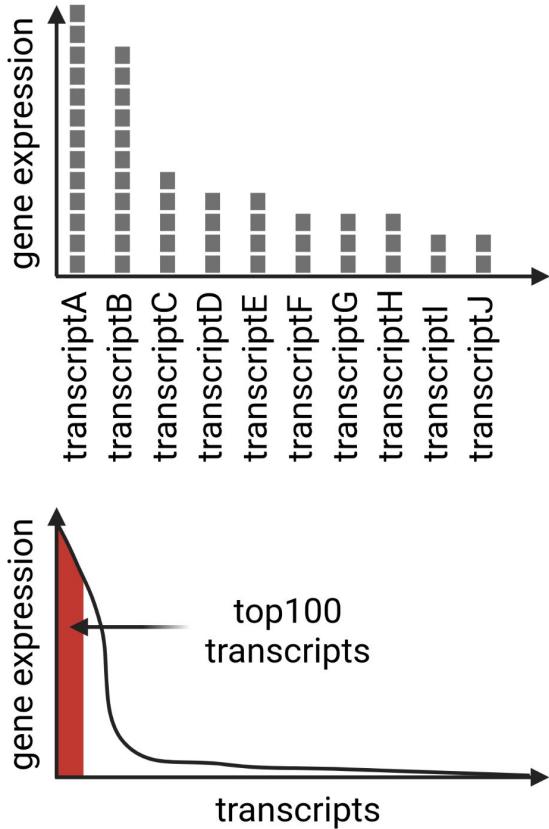
- Preprocessed data sets (count tables @ GEO)
- Cloud solutions (Galaxy)
- Fastq-dump



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



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



Get started with DataCite!



Create and manage DOIs with DataCite Fabrica.



Discover metadata with DataCite Commons.



Integrate with DataCite APIs.

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

Data reuse examples

Benchmarking of NOVOPlasty

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

Benchmarking SANPolyA

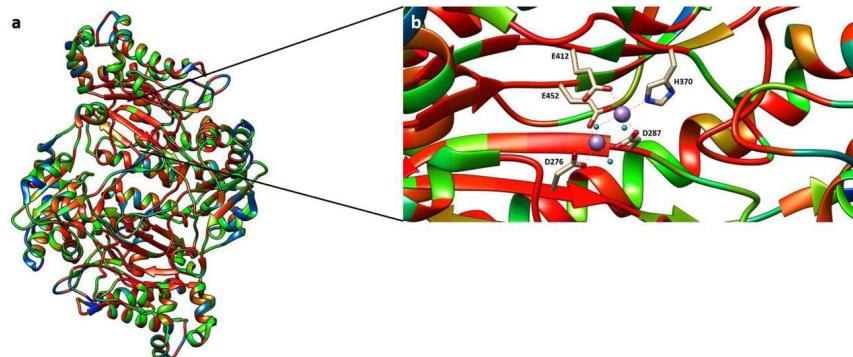
- SANPolyA detects Poly(A) signals through deep learning
- Comparison of SANPolyA results against results of other tools
- Freely available output of tools is required for systematic comparison

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

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



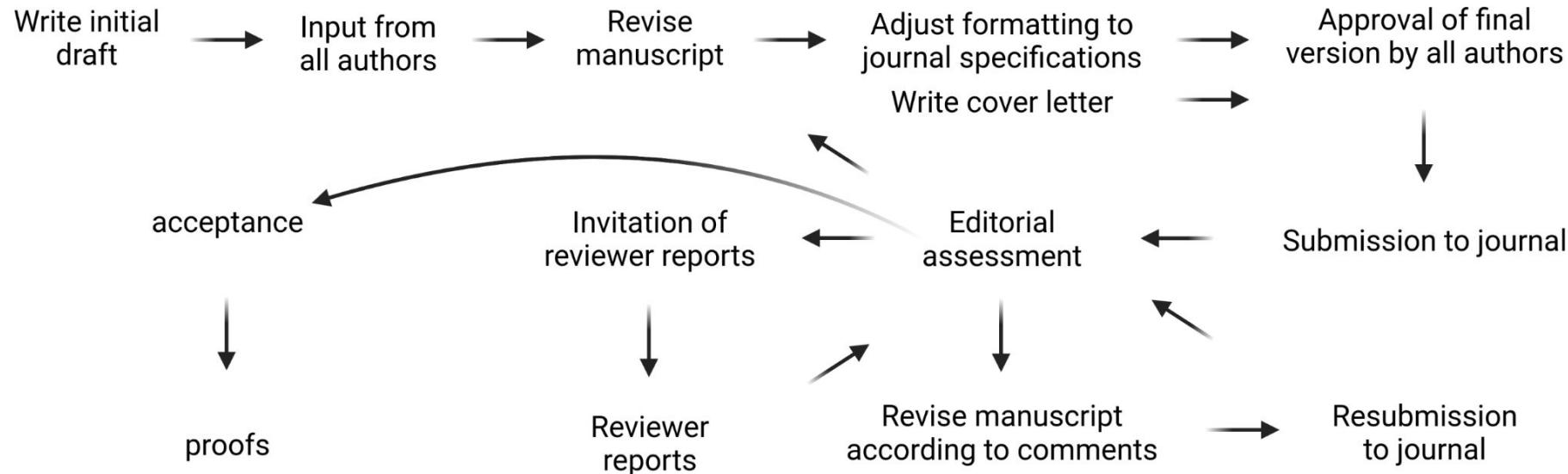
	Animals	Plants	Fungi	Archaea	Bacteria
D276	94	99	100	100	100
D287	94	98	100	100	100
H370	94	98	100	100	100
E412	94	96	100	100	100
E452	91	97	100	100	100
T289	94	97	100	100	97
T410	93	96	100	79	100
H377	94	98	100	100	97
R398	93	98	89	10	57
W107	88	98	96	0	96
Y241	94	96	100	2	90
I244	93	98	97	88	100
H255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

Scientific Publication Business

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



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



Universität Bonn · IZMB · Kirschallee 1 · 53012 Bonn



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.

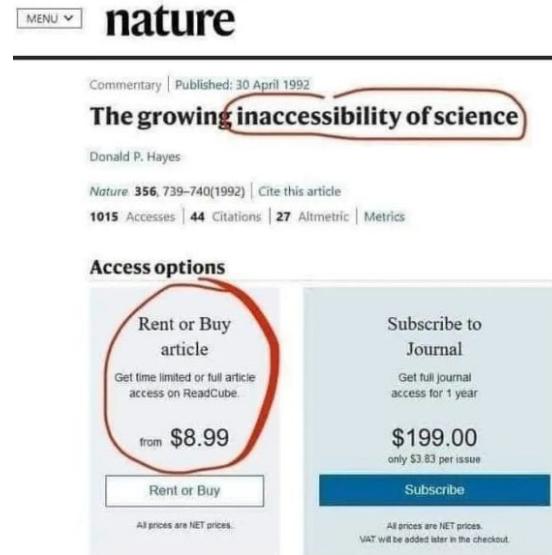
Finding biosynthetic genes is essential to unravel the molecular basis of specialized metabolism. Our software, CoExpPhylo, integrates coexpression analysis and phylogenetic relationships to identify candidate genes involved in biosynthetic pathways. By combining these two complementary approaches, CoExpPhylo 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



The screenshot shows a journal article from *Nature* (Volume 356, page 739, published on April 30, 1992). The title of the article is "The growing inaccessibility of science" by Donald P. Hayes. Below the title, there are links for "Cite this article", "1015 Accesses", "44 Citations", "27 Altmetric", and "Metrics". The "Access options" section is highlighted with a red oval. It offers two main choices: "Rent or Buy article" (from \$8.99) and "Subscribe to Journal" (\$199.00, only \$3.83 per issue). Both options mention "Get full journal access for 1 year". A note at the bottom states "All prices are NET prices. VAT will be added later in the checkout."

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

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

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

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

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

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 APC.

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

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



1

REGISTER

Get your unique ORCID identifier. It's free and only takes a minute, so [register now!](#)

2

USE YOUR ORCID ID

Use your ID, when prompted, in systems and platforms from grant application to manuscript submission and beyond, to ensure you get credit for your contributions.

3

SHARE YOUR ORCID ID

The more information connected to your ORCID record, the more you'll benefit from sharing your ID - so give the organizations you trust permission to update your record as well as adding your affiliations, emails, other names you're known by, and more.

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



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

Time for questions!

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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1017/qpb.2021.18).