Technische Universität Braunschweig

Plant Biotechnology and Bioinformatics

Summarizing, Visualizing, Publishing - Big Data Analytics in Life Sciences

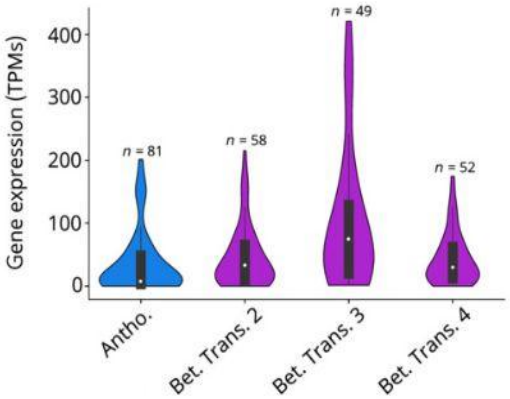
Prof. Dr. Boas Pucker (Plant Biotechnology and Bioinformatics)

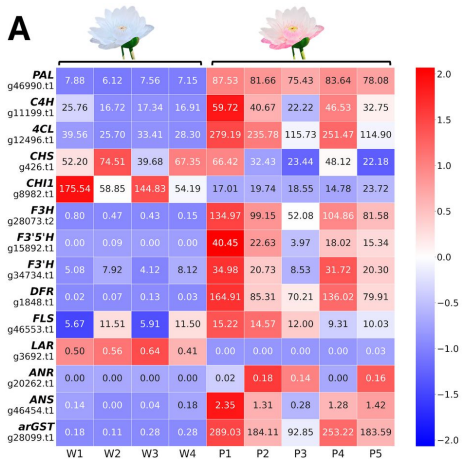
Outline - Summarizing & visualizing

- General advice and file types (Inkscape + BioIcons)
- Automatic figure generation (matplotlib, plotly, ggplot2)
- Visualizing complex networks (Cytoscape)
- Heatmaps, violin plots, KEGG & GO enrichment figures
- Web examples: eFP browser, ePlant

Figure design recommendations

- Careful check for outliers is important
- Comparison of two distributions should be based on boxplots or violin plots (not barplots)
- Violin plots only work for large sample sizes
- Heatmaps should use clustering of data by similarity or biologically informed sorting
- Try different layouts of networks





Pucker et al., 2024: 10.1111/nph.19341
Nowak et al., 2024: 10.1101/2024.06.15.599162
https://github.com/cxli233/FriendsDontLetFriends

Figure file types

- JPG/JPEG (Joint Photographic Experts Group): JFIF, SPIFF, JNG
  - lossy data compression
- PNG (Portable Network Graphics):
  - lossless data compression; transparency possible, 32-bit RGBA colors
- SVG (Scalable Vector Graphics):
  - XML-based; support for interactivity and animations; web-friendly
- PDF (Portable Document Format):
  - contains all information needed for display (fronts, text, graphics)
- TIF/TIFF (Tagged Image File Format)
  - Raster graphic images

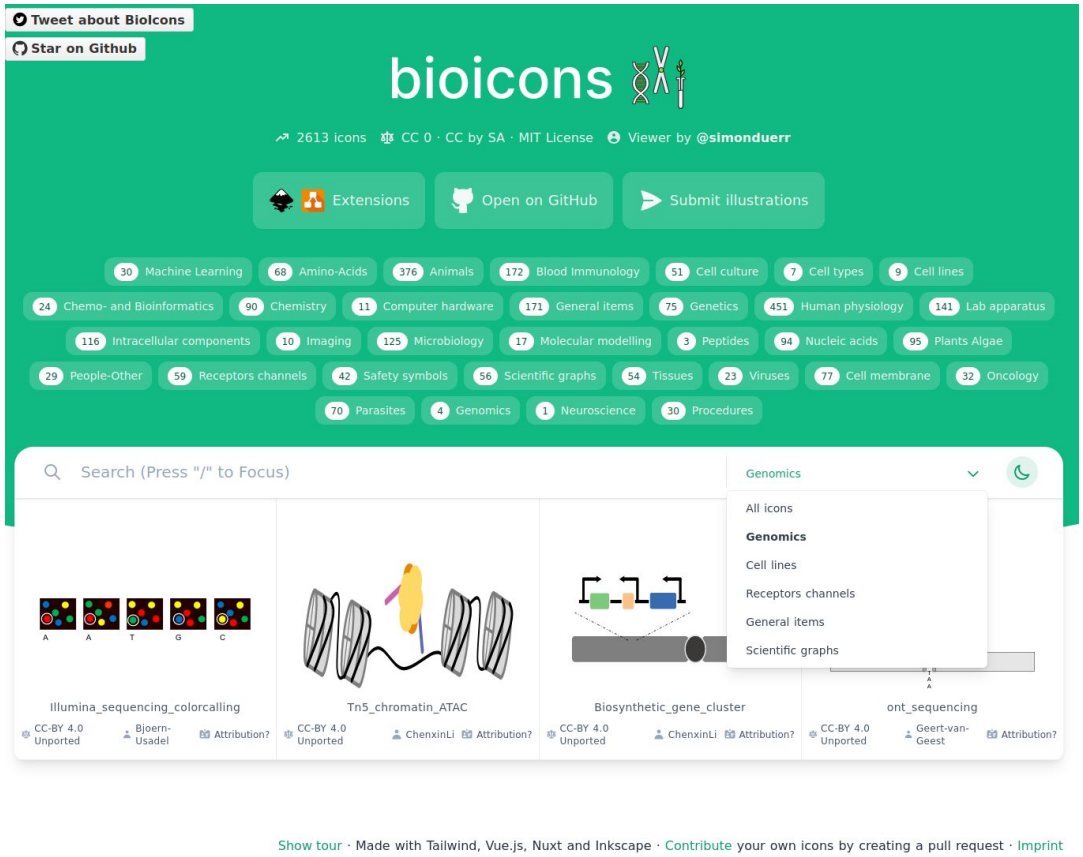
Figure file types II

- EPS (Encapsulated Postscript):
  - Vector image file; resizing without quality loss

- BMP (bitmap)
  - Raster graphics image, data compression, transparency possible

Bioicons

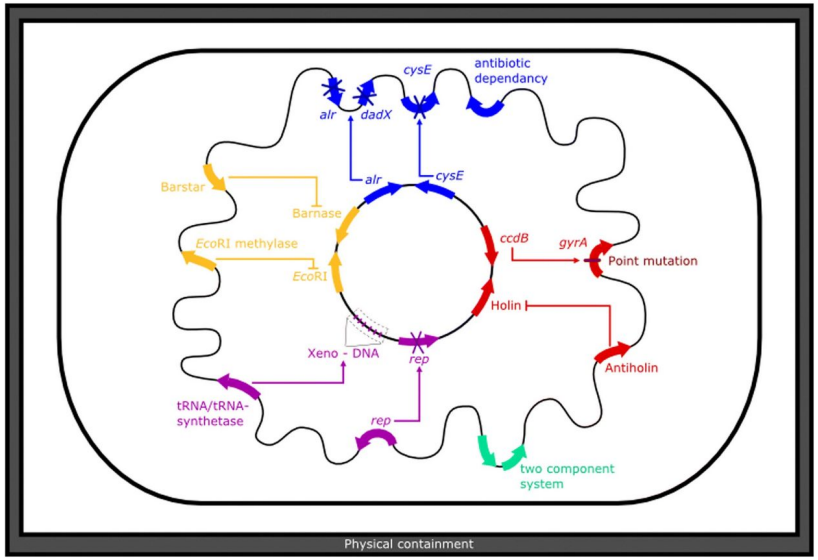
- Collection of free biology-related icons in SVG format
- License is CC0 or CC BY
- Keyword search and filter by topics possible
- Number of figures limited, but growing quickly
- @PuckerLab is actively contributing to this resource

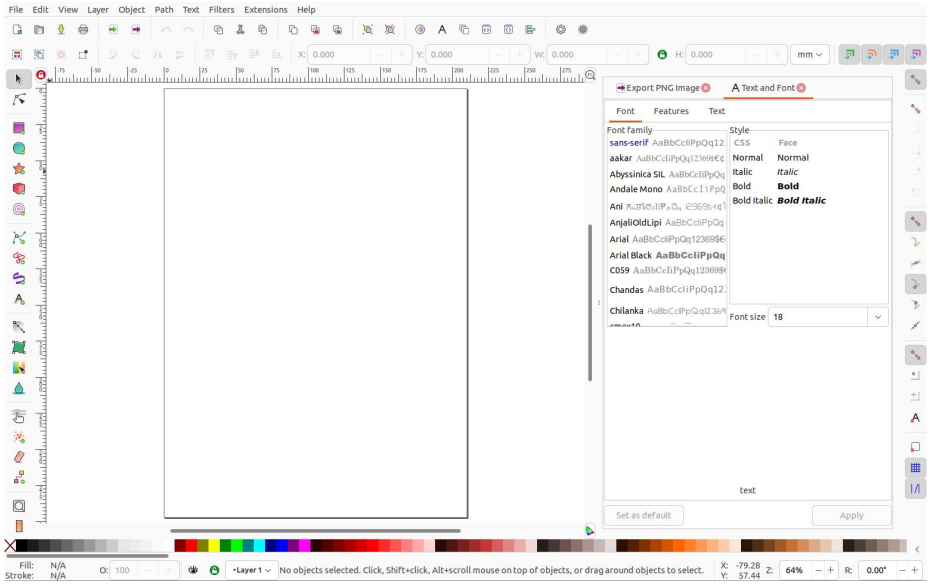


https://bioicons.com/

Manual figure editing with InkScape

- InkScape is a SVG editor
- Freely available for all operating systems
- Export options for various file types



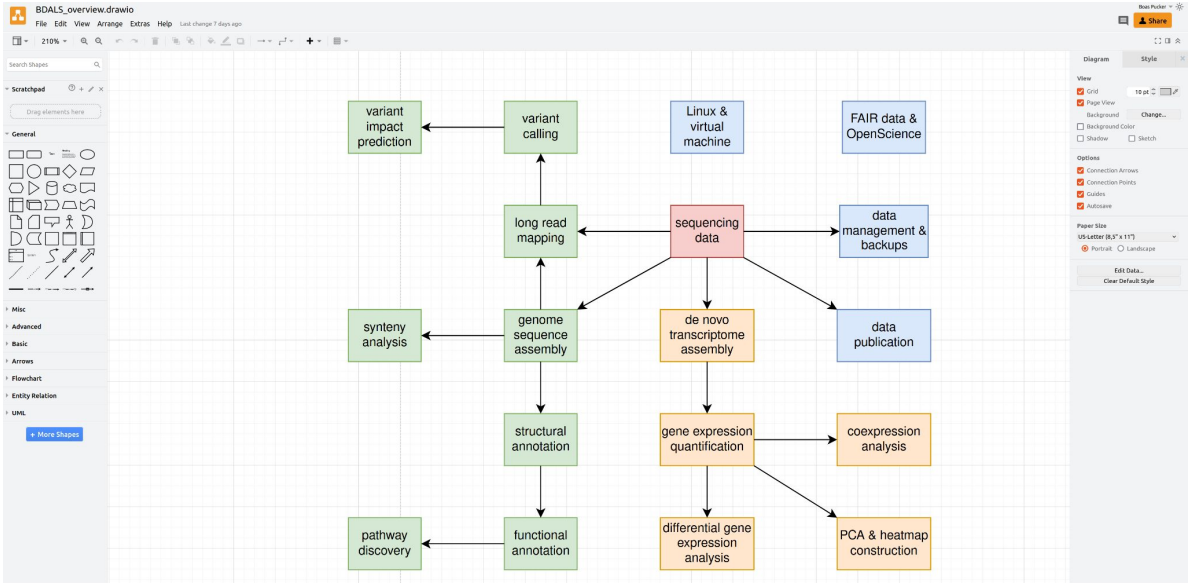


https://inkscape.org/cs/

Whitford et al., 2018: 10.1186/s13036-018-0105-8

Manual figure editing with draw.io

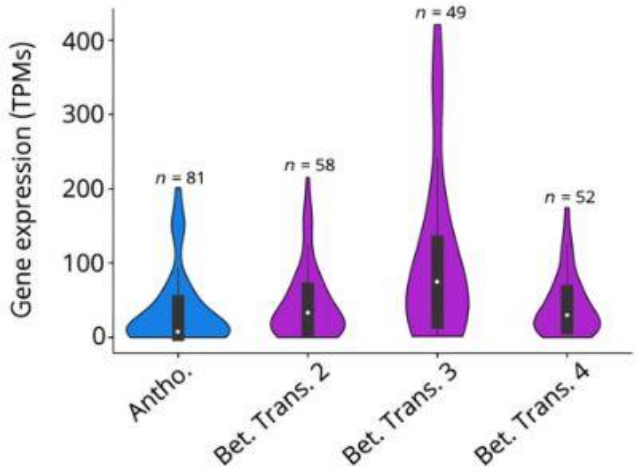
- Online editing of figures
- Export in multiple file formats and sizes
- Figures are stored in gdrive, dropbox, or other clouds
- Freely available



https://app.diagrams.net/

Automatic figure generation - matplotlib

- Well established Python module; often combined with seaborn
- Excellent features and reimplemented in other languages
- Export in various file formats possible (system needs to support)
- ‘Old’ module for figure generation in Python

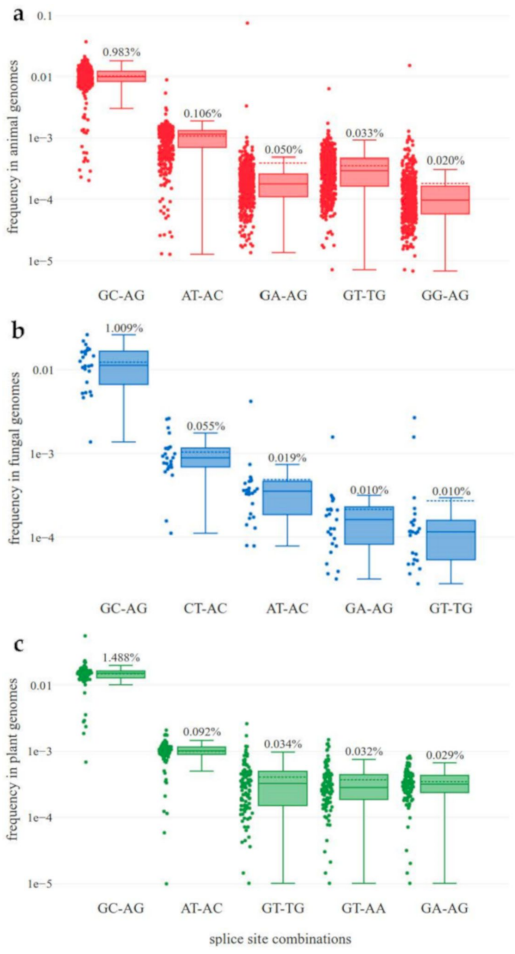


matplotlib

Pucker et al., 2024: 10.1111/nph.19341
https://matplotlib.org/
https://seaborn.pydata.org/

Automatic figure generation - plotly

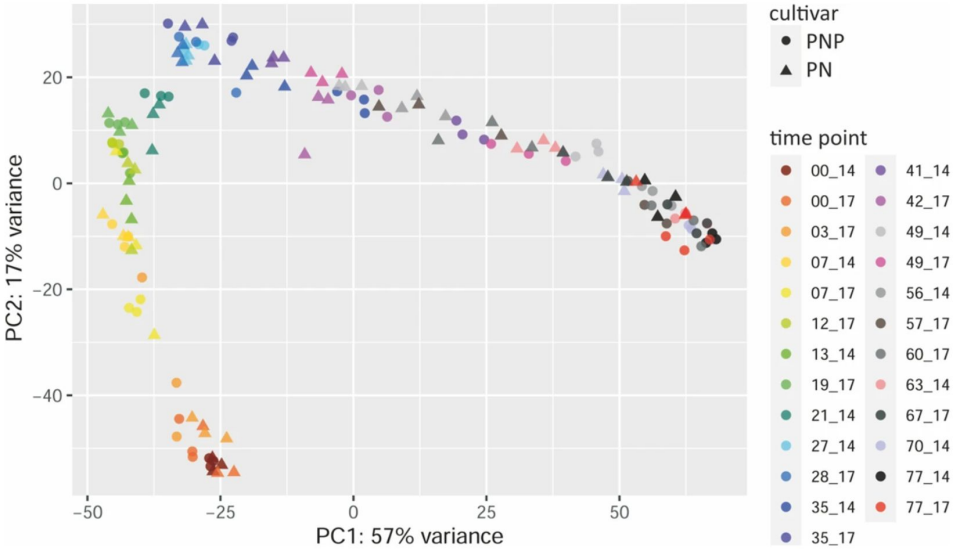
- Modern Python module for the generation of figures
- Integration of interactive figures in websites possible
- Export in different file formats supported



Frey & Pucker, 2020: 10.3390/cells9020458
https://github.com/plotly/plotly.py

Automatic figure generation - ggplot2

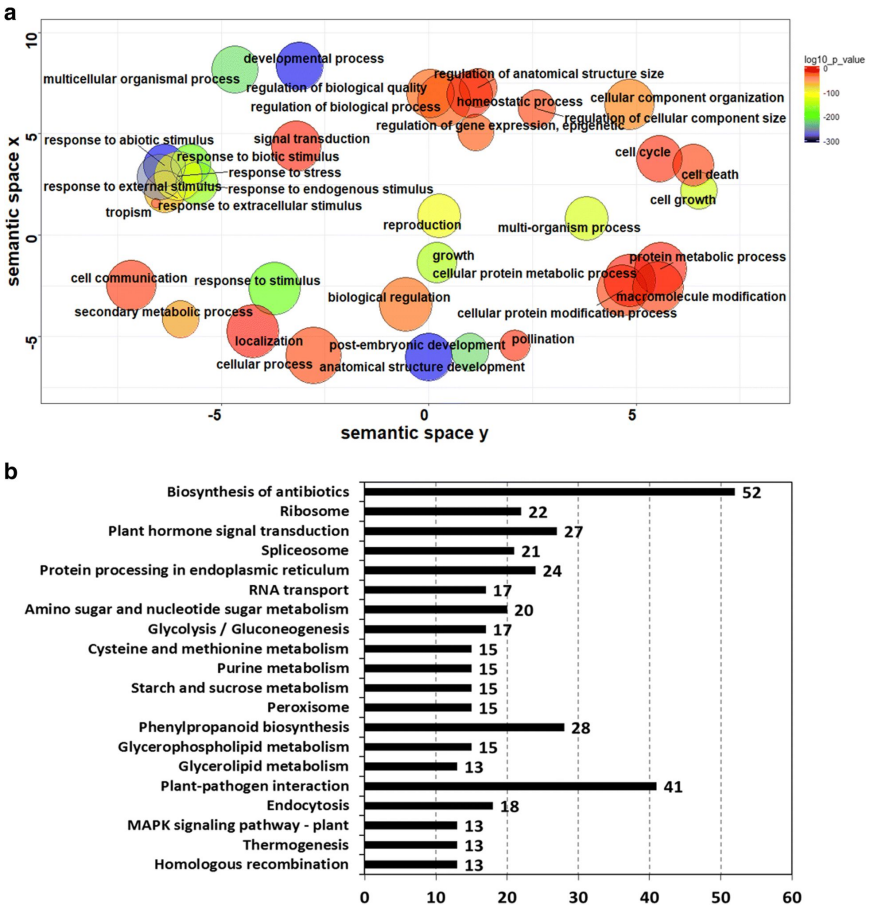
- R package for generation of big data figures
- Helpful instructions available online



Theine et al., 2021: 10.1186/s12870-021-03110-6
https://ggplot2.tidyverse.org/

GO enrichment

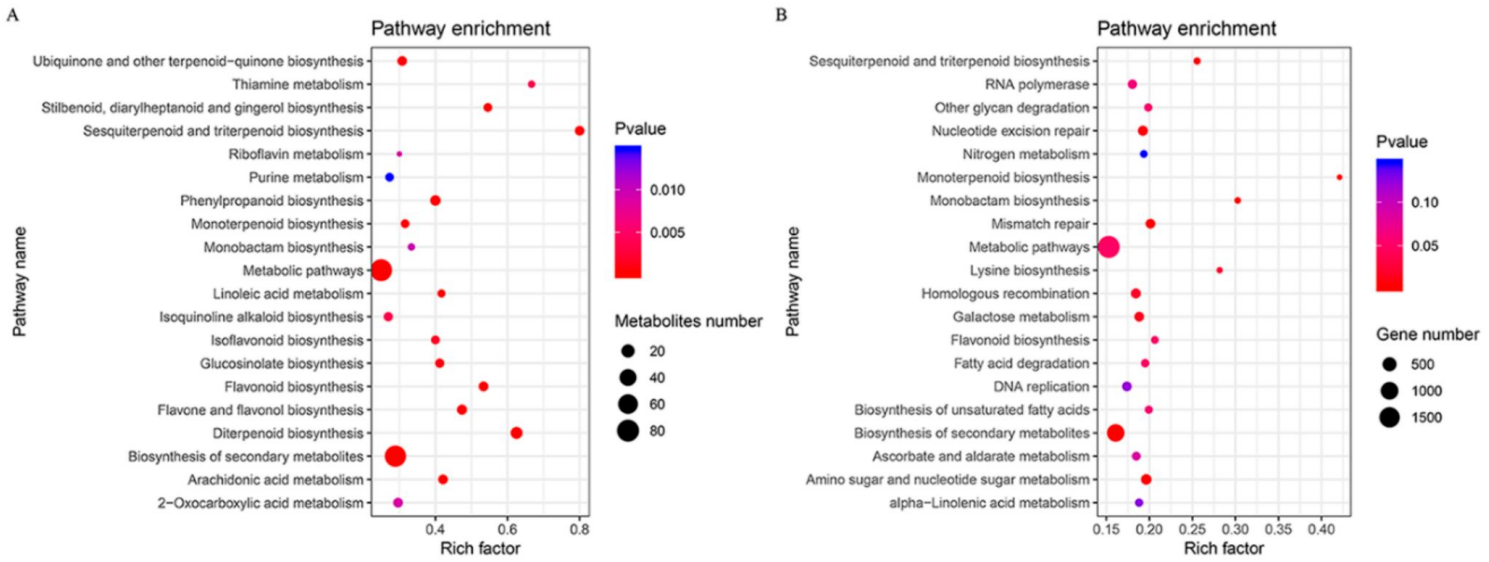
- GO = Gene Ontology
- GO term can be assigned to multiple genes (e.g. ‘flavonoid biosynthesis’)
- Up-regulated GO = larger proportion of genes with this GO term are up-regulated than expected by chance
- Tools: AmiGO, GOrilla, ShinyGO, GOnet, g:Profiler, …



https://doi.org/10.1186/s12864-018-5357-7

KEGG enrichment

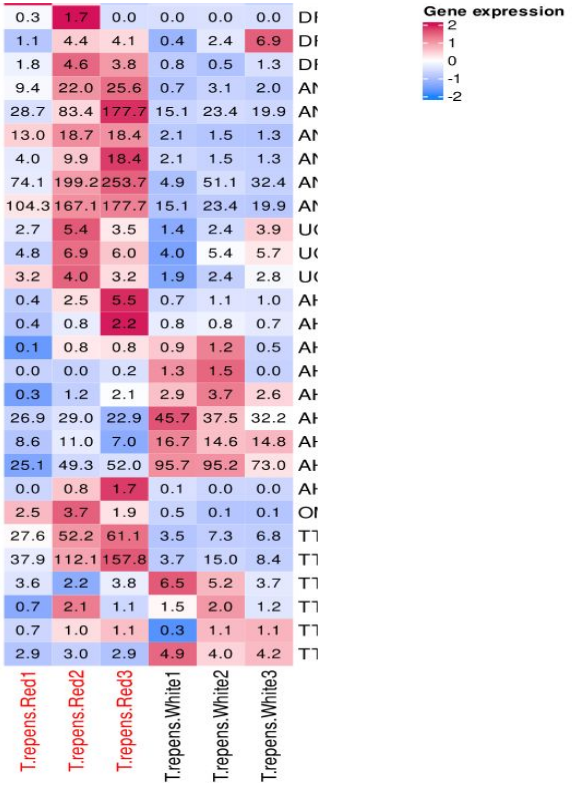
- Enrichment is based on Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways
- Mercator can be used to assign genes to KEGG pathways
- Up-regulated pathway = large proportion of up-regulated genes in pathway



https://doi.org/10.3390/genes11020187

Heatmaps

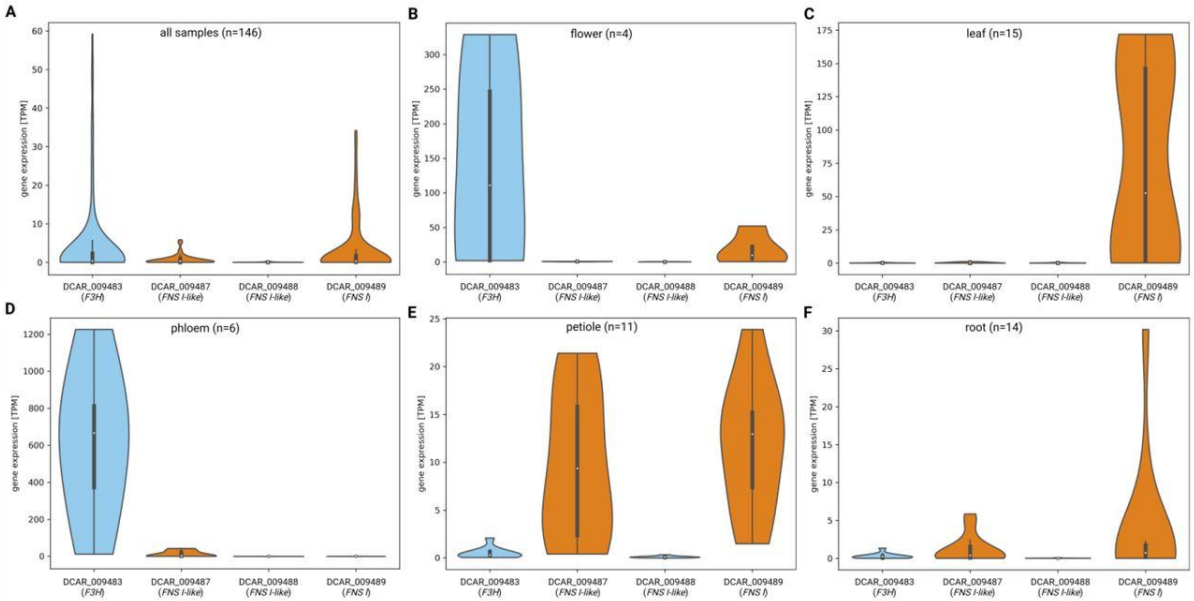
- Visualization of gene expression values based on color/color intensity
- Red (heat) often indication of high expression
- Different color ranges are available in Python/R
- Additional normalization per gene across all samples



https://doi.org/10.1101/2023.06.05.543820

Violin plots

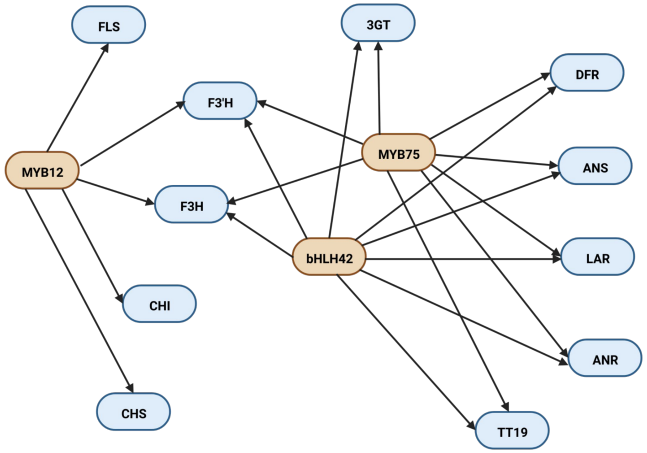
- Summarization of multiple data points
- Indication of data point distribution
- Mean and median can be displayed



https://doi.org/10.1371/journal.pone.0280155

Gene expression networks

- Gene expression in multiple samples can be used to infer regulatory networks
- Edges in the network are based on co-expression
- Transcription factors represent central nodes
- Cytoscape can be used to visualize network files

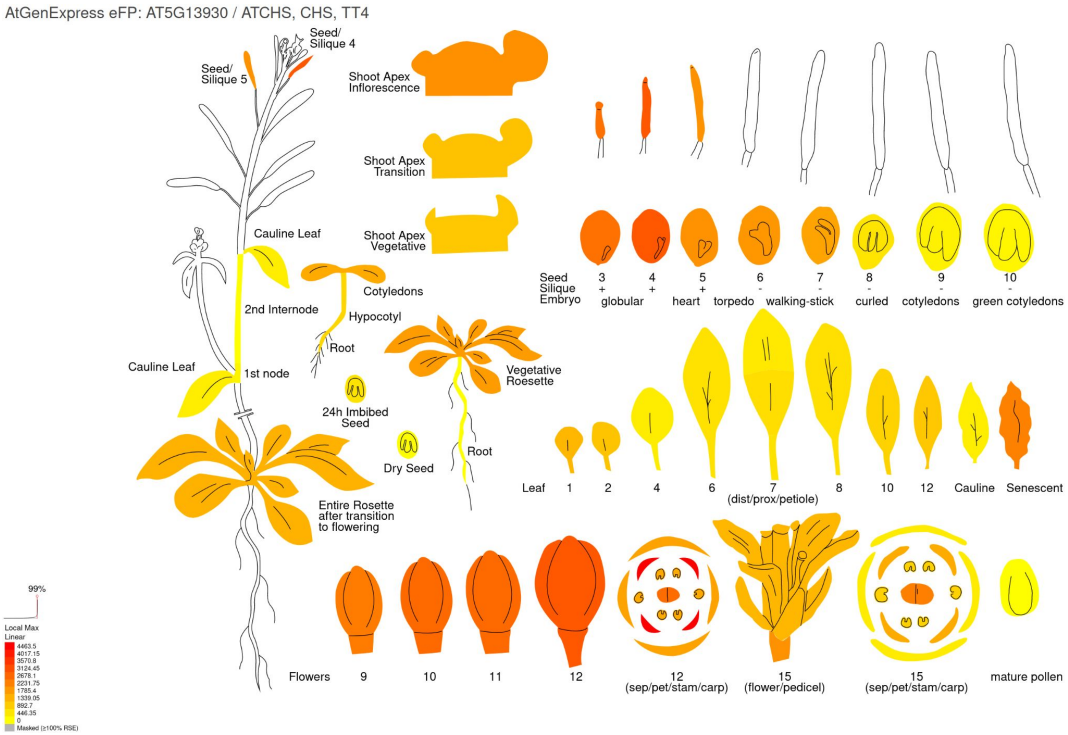


Visualizing complex networks with Cytoscape

- Automatic visualization of complex networks
- Colors, shapes, symbols, and width of elements can indicate values
- Manual rearrangement of elements in network possible
- Availability (v3.10.2): https://cytoscape.org/download.html

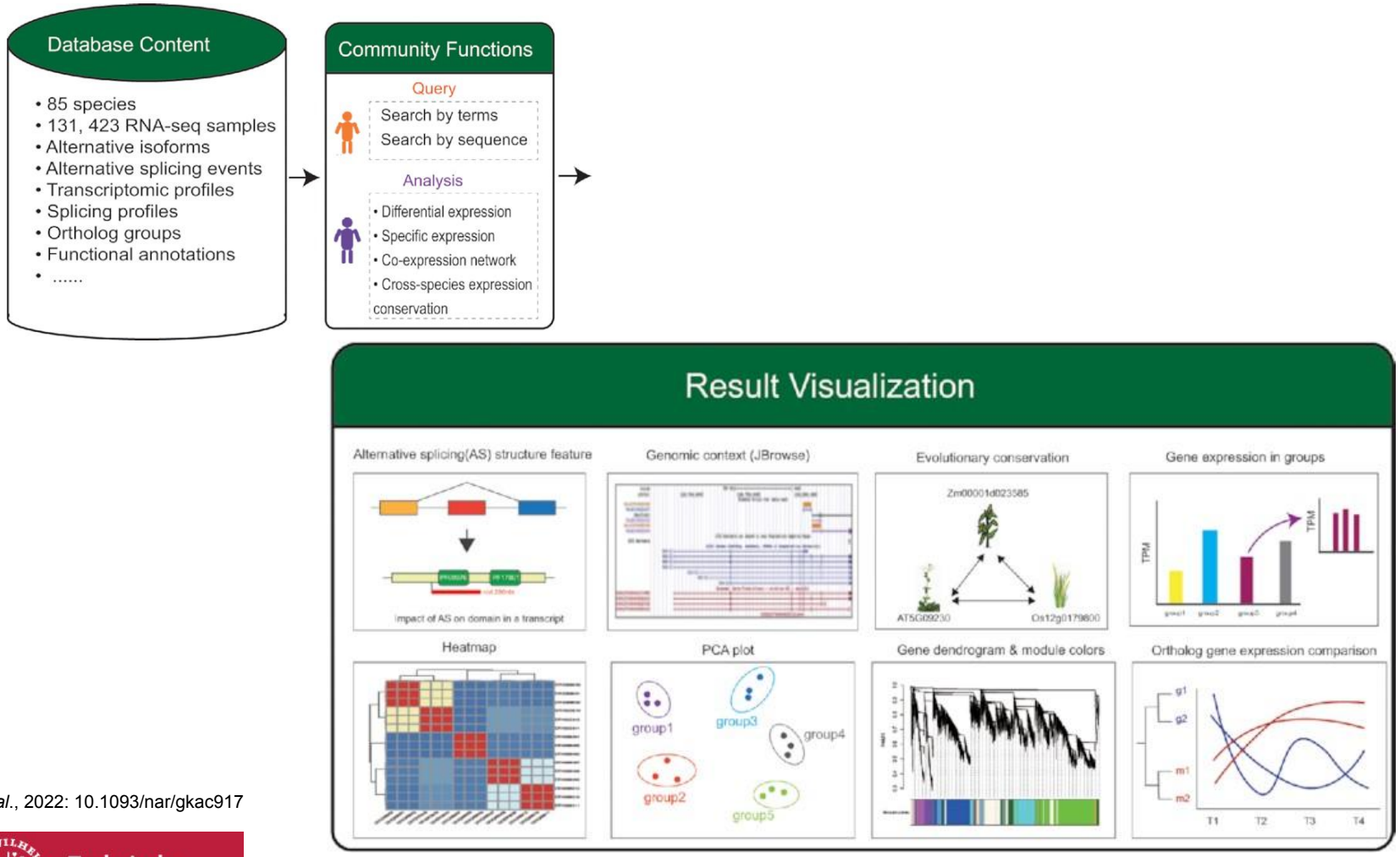
eFP browser / ePlant

- Gene expression visualization in pictures
- Inclusion of other qualitative plant data
- Mapping to individual plant parts or conditions
- Many other connected data can be visualized



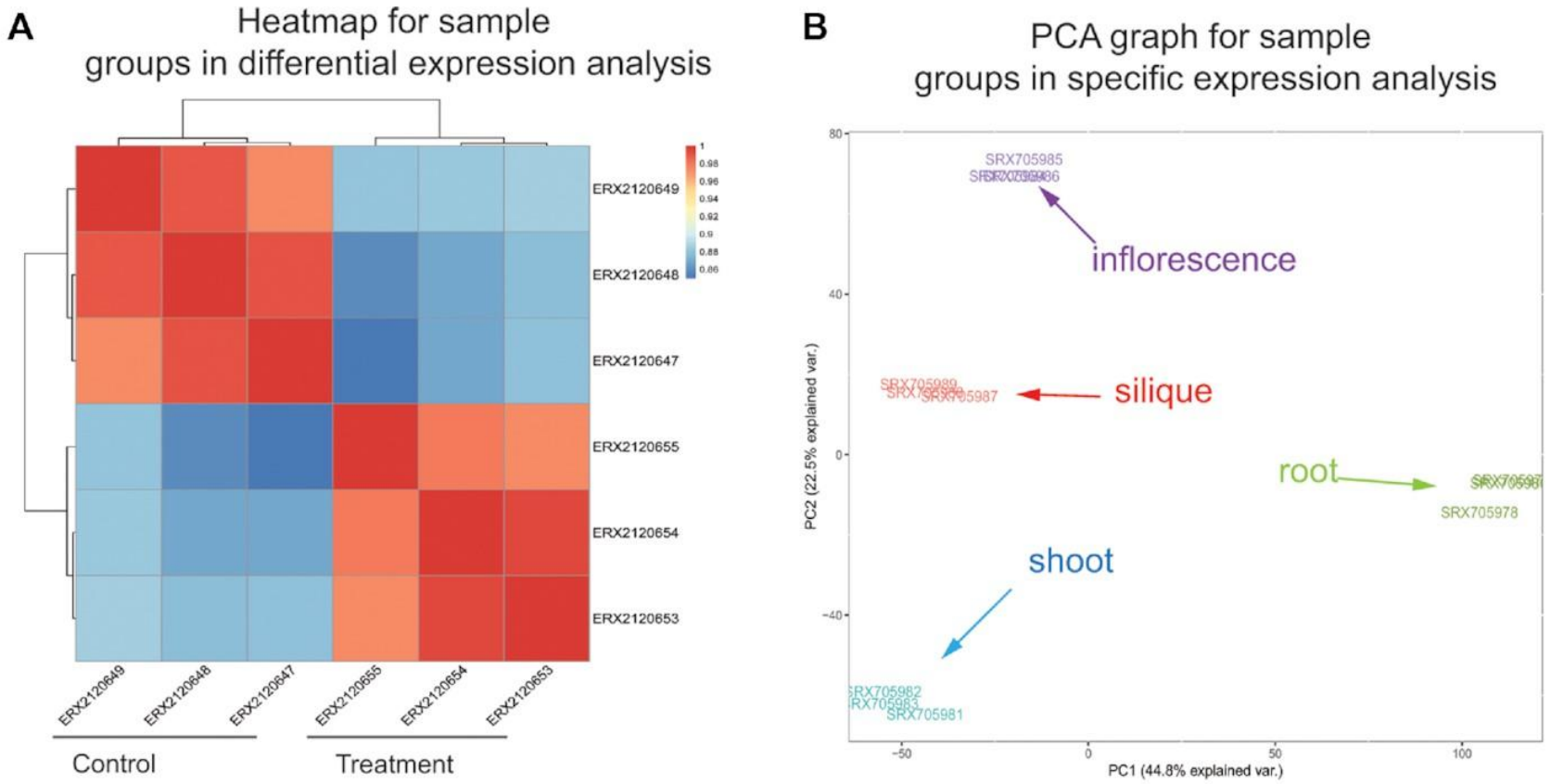
https://bar.utoronto.ca/eplant/

PlantExp: gene expression & alternative splicing



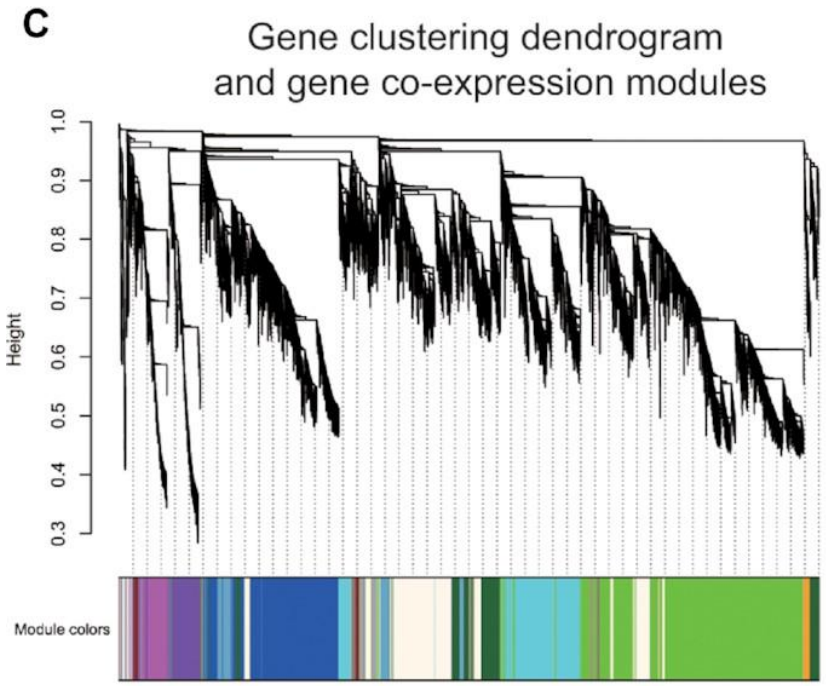
Liu *et al*., 2022: 10.1093/nar/gkac917

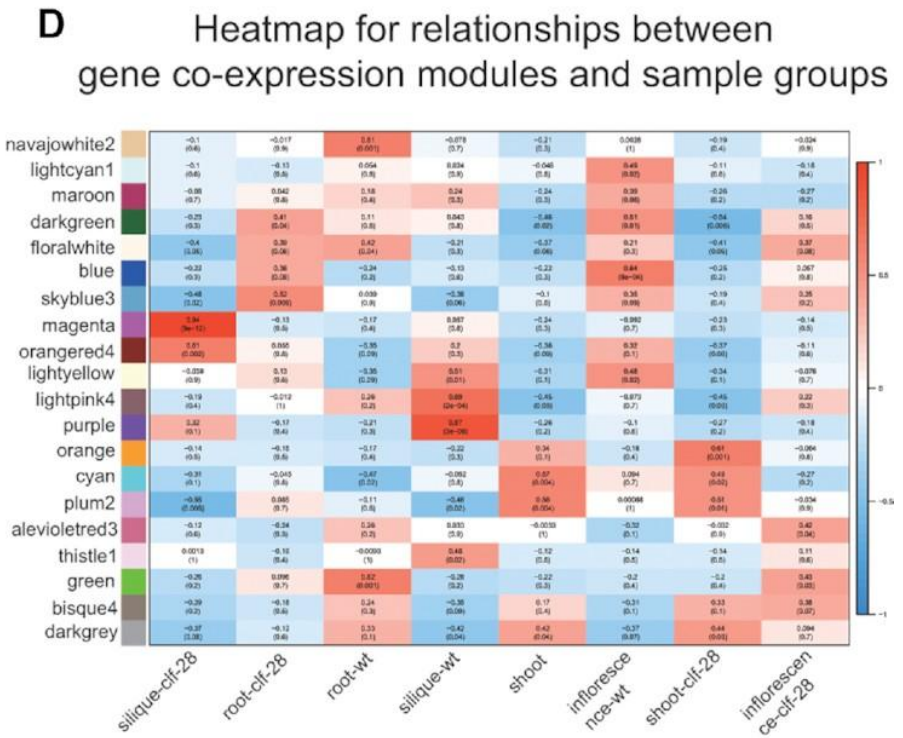
PlantExp: example plots (1)



Liu *et al.*, 2022: 10.1093/nar/gkac917

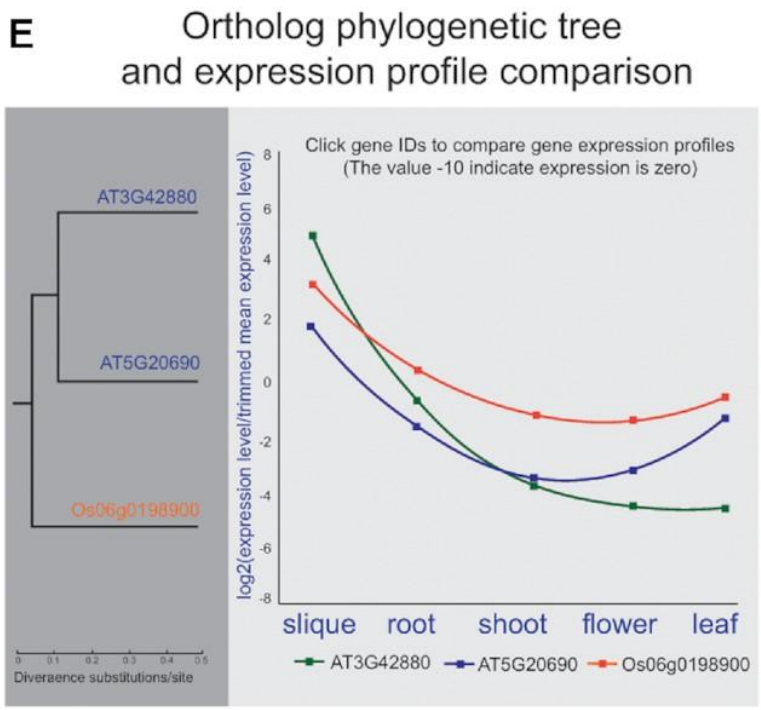
PlantExp: example plots (2)

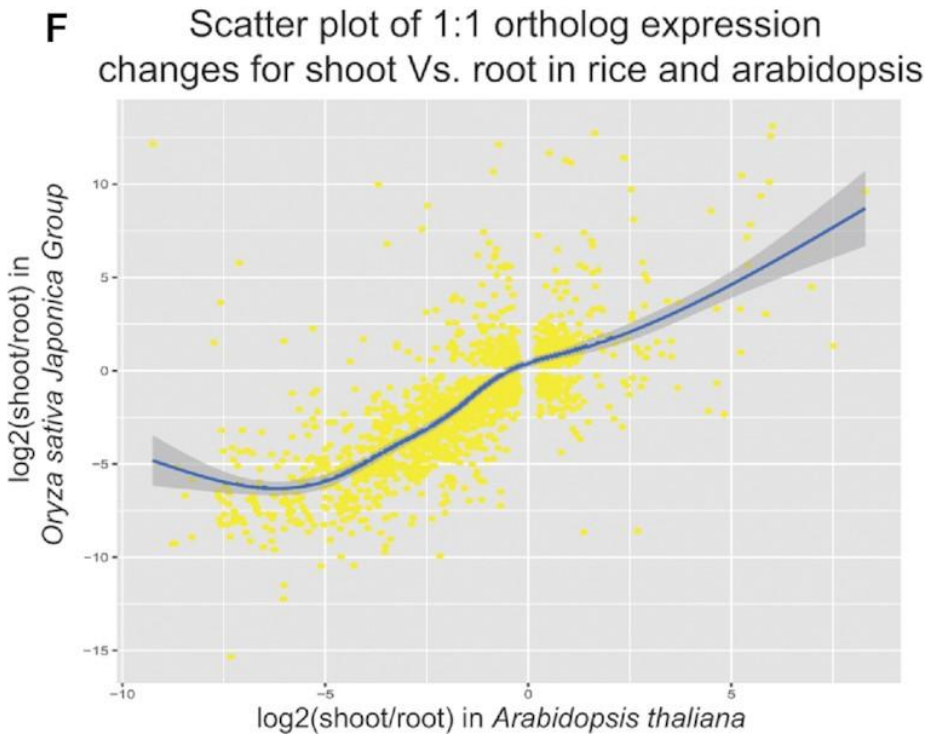




Liu *et al*., 2022: 10.1093/nar/gkac917

PlantExp: example plots (3)





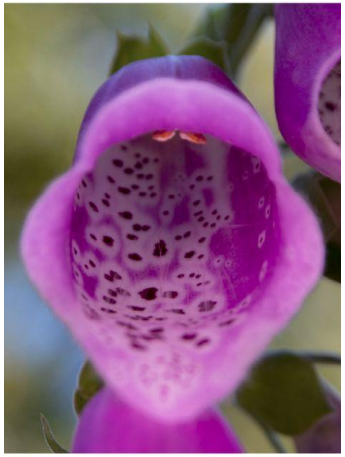
Liu *et al*., 2022: 10.1093/nar/gkac917

Outline - Publishing

- Introduction to scientific publishing business
- Submission of sequencing data to ENA
- Data publications (LeoPARD, Dryad)
- OpenMethods & protocols.io
- Depositing scripts in GitHub and Zenodo

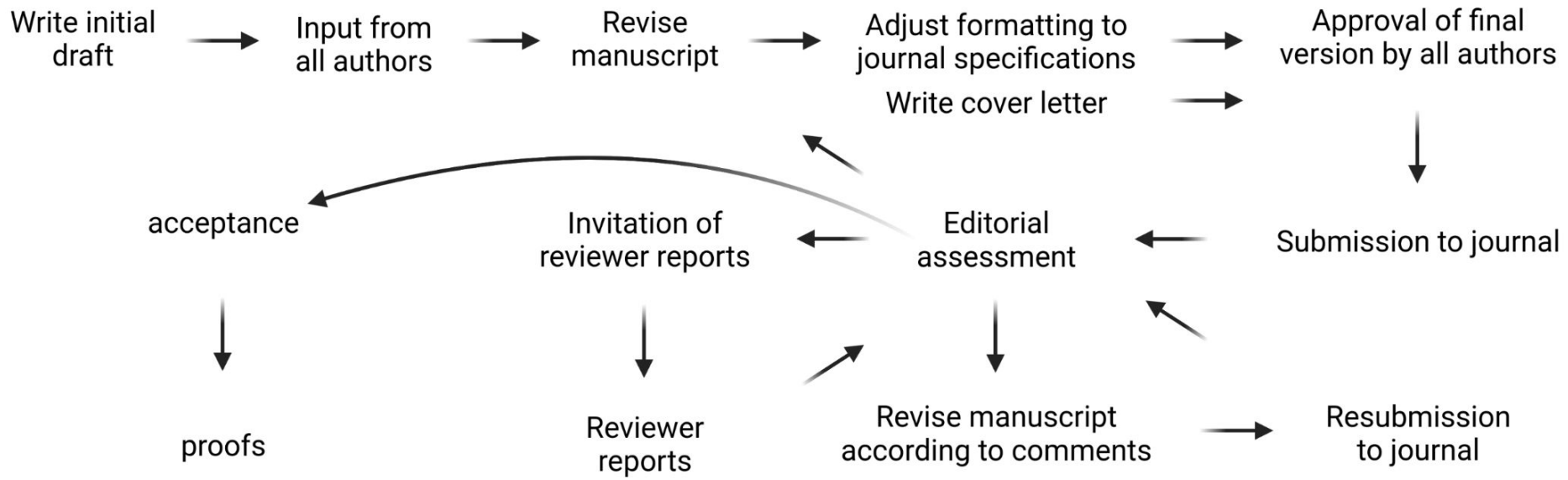
Motivation for publishing a scientific paper

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



https://doi.org/10.1101/2024.02.14.580303

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

Value of citations

- Currency of science: ‘Publish or perish’
- Important for scientific career: job applications, grant applications
- Quantity vs. quality
- Applications often require 3, 5, or 10 ‘most important’ publications

What is the h-index?

Position		Citations	
1	Genome-wide identification and characterisation of R2R3-MYB genes in sugar beet (Beta vulgaris) R Stracke, D Holtgräwe, J Schneider, B Pucker, T Rosleff Sörensen, ... BMC Plant Biology 14 (1), 1-17	77	2014
2	The evolution of betalain biosynthesis in Caryophyllales A Timoneda, T Feng, H Sheehan, N Walker-Hale, B Pucker, ... New Phytologist 224 (1), 71-85	56	2019
3	The negative regulator SMAX1 controls mycorrhizal symbiosis and strigolactone biosynthesis in rice J Choi, T Lee, J Cho, EK Servante, B Pucker, W Summers, S Bowden, ... Nature communications 11 (1), 1-13	44	2020
4	A *De Novo* Genome Sequence Assembly of the *Arabidopsis thaliana* Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny B Pucker, D Holtgräwe, T Rosleff Sörensen, R Stracke, P Viehöver, ... PLoS One 11 (10), e0164321	34	2016
5	Evolution of L-DOPA 4,5-dioxygenase activity allows for recurrent specialisation to betalain pigmentation in Caryophyllales H Sheehan, T Feng, N Walker-Hale, S Lopez-Nieves, B Pucker, R Guo, ... New Phytologist 227 (3), 914-929	29	2020
6	A chromosome-level sequence assembly reveals the structure of the *Arabidopsis thaliana* Nd-1 genome and its gene set B Pucker, D Holtgräwe, KB Stadermann, K Frey, B Huettel, R Reinhardt, ... PloS one 14 (5), e0216233	29	2019
7	Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes B Pucker, SF Brockington BMC genomics 19 (1), 1-13	24	2018
8	High contiguity de novo genome sequence assembly of trifoliate yam (Dioscorea dumetorum) using long read sequencing C Siadjeu, B Pucker, P Viehöver, DC Albach, B Weisshaar Genes 11 (3), 274	23	2020
9	High quality de novo transcriptome assembly of Croton tiglium M Haak, S Vinke, W Keller, J Droste, C Rückert, J Kalinowski, B Pucker Frontiers in molecular biosciences 5, 62	23	2018
10	Auxotrophy to Xeno-DNA: an exploration of combinatorial mechanisms for a high-fidelity biosafety system for synthetic biology applications CM Whitford, S Dymek, D Kerkhoff, C März, O Schmidt, M Edich, J Droste, ... Journal of Biological Engineering 12 (1), 1-28	18	2018
11	Consideration of non-canonical splice sites improves gene prediction on the Arabidopsis thaliana Niederzenz-1 genome sequence B Pucker, D Holtgräwe, B Weisshaar BMC Research Notes 10 (1), 1-6	17	2017
12	Automatic identification of players in the flavonoid biosynthesis with application on the biomedicinal plant Croton tiglium B Pucker, F Reiher, HM Schilbert Plants 9 (9), 1103	16	2020
13	Comparison of read mapping and variant calling tools for the analysis of plant NGS data HM Schilbert, A Rempel, B Pucker Plants 9 (4), 439	15	2020
14	The *R2R3-MYB* gene family in banana *(Musa acuminata)*: Genome-wide identification, classification and expression patterns B Pucker, A Pandey, B Weisshaar, R Stracke PloS one 15 (10), e0239275	14	2020
15	Animal, fungi, and plant genome sequences harbor different non-canonical splice sites K Frey, B Pucker Cells 9 (2), 458	14 *	2020
16	The land plant-specific MIXTA-MYB lineage is implicated in the early evolution of the plant cuticle and the colonization of land B Xu, L Taylor, B Pucker, T Feng, BJ Glover, SF Brockington New Phytologist 229 (4), 2324-2338	11	2021
17	Integrating molecular biology and bioinformatics education B Pucker, HM Schilbert, SF Schumacher Journal of integrative bioinformatics 16 (3)	11	2019
18	The reuse of public datasets in the life sciences: potential risks and rewards K Sielemann, A Hafner, B Pucker PeerJ 8, e9954	9	2020

h=14

What is the i10-index?

Position		Citations	
1	Genome-wide identification and characterisation of R2R3-MYB genes in sugar beet (Beta vulgaris) R Stracke, D Holtgräwe, J Schneider, B Pucker, T Rosleff Sörensen, ... BMC Plant Biology 14 (1), 1-17	77	2014
2	The evolution of betalain biosynthesis in Caryophyllales A Timoneda, T Feng, H Sheehan, N Walker-Hale, B Pucker, ... New Phytologist 224 (1), 71-85	56	2019
3	The negative regulator SMAX1 controls mycorrhizal symbiosis and strigolactone biosynthesis in rice J Choi, T Lee, J Cho, EK Servante, B Pucker, W Summers, S Bowden, ... Nature communications 11 (1), 1-13	44	2020
4	A *De Novo* Genome Sequence Assembly of the *Arabidopsis thaliana* Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny B Pucker, D Holtgräwe, T Rosleff Sörensen, R Stracke, P Viehöver, ... PLoS One 11 (10), e0164321	34	2016
5	Evolution of L-DOPA 4,5-dioxygenase activity allows for recurrent specialisation to betalain pigmentation in Caryophyllales H Sheehan, T Feng, N Walker-Hale, S Lopez-Nieves, B Pucker, R Guo, ... New Phytologist 227 (3), 914-929	29	2020
6	A chromosome-level sequence assembly reveals the structure of the *Arabidopsis thaliana* Nd-1 genome and its gene set B Pucker, D Holtgräwe, KB Stadermann, K Frey, B Huettel, R Reinhardt, ... PloS one 14 (5), e0216233	29	2019
7	Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes B Pucker, SF Brockington BMC genomics 19 (1), 1-13	24	2018
8	High contiguity de novo genome sequence assembly of trifoliate yam (Dioscorea dumetorum) using long read sequencing C Sladjeu, B Pucker, P Viehöver, DC Albach, B Weisshaar Genes 11 (3), 274	23	2020
9	High quality de novo transcriptome assembly of Croton tiglium M Haak, S Vinke, W Keller, J Droste, C Rückert, J Kalinowski, B Pucker Frontiers in molecular biosciences 5, 62	23	2018
10	Auxotrophy to Xeno-DNA: an exploration of combinatorial mechanisms for a high-fidelity biosafety system for synthetic biology applications CM Whitford, S Dymek, D Kerkhoff, C März, O Schmidt, M Edich, J Droste, ... Journal of Biological Engineering 12 (1), 1-28	18	2018
11	Consideration of non-canonical splice sites improves gene prediction on the Arabidopsis thaliana Niederzenz-1 genome sequence B Pucker, D Holtgräwe, B Weisshaar BMC Research Notes 10 (1), 1-6	17	2017
12	Automatic identification of players in the flavonoid biosynthesis with application on the biomedicinal plant Croton tiglium B Pucker, F Reher, HM Schilbert Plants 9 (9), 1103	16	2020
13	Comparison of read mapping and variant calling tools for the analysis of plant NGS data HM Schilbert, A Rempel, B Pucker Plants 9 (4), 439	15	2020
14	The *R2R3-MYB* gene family in banana *(Musa acuminata)*: Genome-wide identification, classification and expression patterns B Pucker, A Pandey, B Weisshaar, R Stracke PloS one 15 (10), e0239275	14	2020
15	Animal, fungi, and plant genome sequences harbor different non-canonical splice sites K Frey, B Pucker Cells 9 (2), 456	14 *	2020
16	The land plant-specific MIXTA-MYB lineage is implicated in the early evolution of the plant cuticle and the colonization of land B Xu, L Taylor, B Pucker, T Feng, BJ Glover, SF Brockington New Phytologist 229 (4), 2324-2338	11	2021
17	Integrating molecular biology and bioinformatics education B Pucker, HM Schilbert, SF Schumacher Journal of integrative bioinformatics 16 (3)	11	2019
18	The reuse of public datasets in the life sciences: potential risks and rewards K Sielemann, A Hafner, B Pucker PeerJ 8, e9954	9	2020

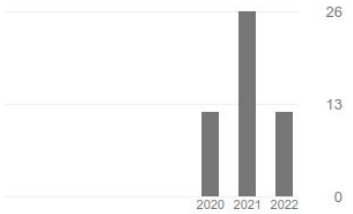
i10=17

Citations are career stage and field-dependent

Doctoral student

Zitiert von

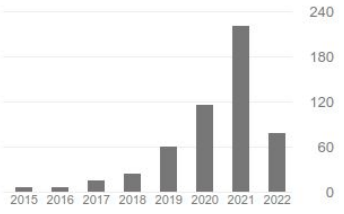
	Alle	Seit 2017
Zitate	53	53
h-index	4	4
i10-index	3	3



Young PI

Cited by

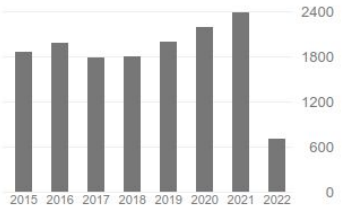
	All	Since 2017
Citations	544	524
h-index	14	14
i10-index	17	17



Highly cited PI

Cited by VIEW ALL

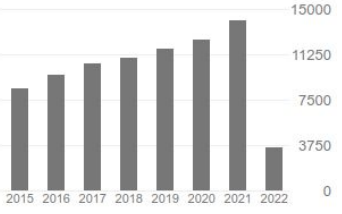
	All	Since 2017
Citations	27595	10899
h-index	70	42
i10-index	131	97



Pioneer and geneious

Cited by VIEW ALL

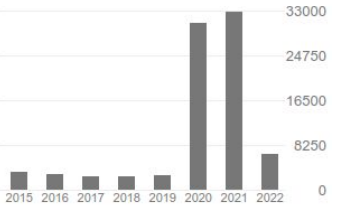
	All	Since 2017
Citations	144131	63748
h-index	185	125
i10-index	586	510



Researcher with ‘hot topic’

Cited by VIEW ALL

	All	Since 2017
Citations	102020	79258
h-index	127	100
i10-index	448	393

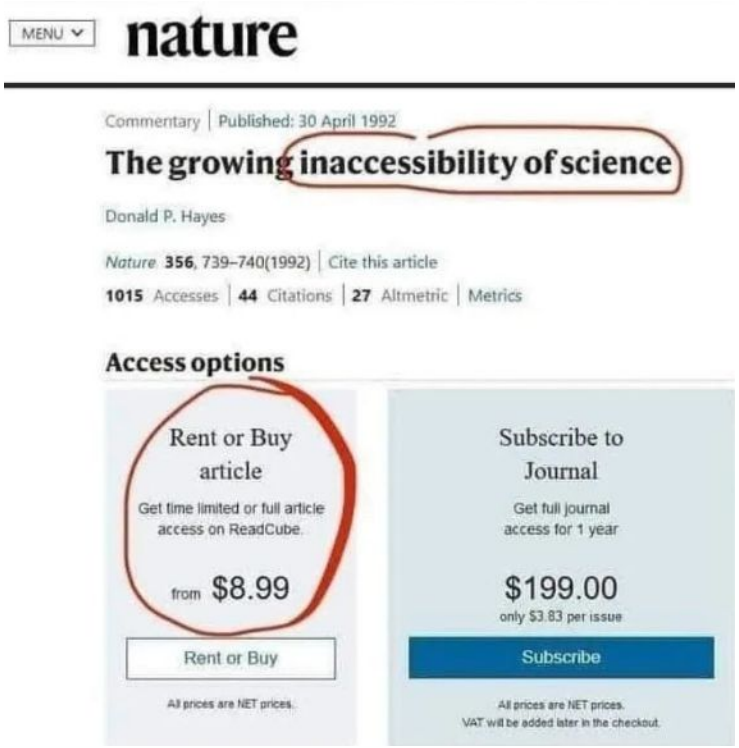


Journal impact factor (IF)

- Journals like to present a journal impact factor (JIF or IF)
  2-year-impact factor and 5-year impact factor
- IF = Average number of citations that each publication receives
- Number of citations per publication is extremely heterogeneous i.e. not an accurate reflection of article quality; publication-based metrics are more accurate
- Top journals: Nature, Science, Cell (30-50) (exciting story, ‘broad interest’)
- Good journals: New Phytologist, Genome Biology, Nucleic Acid Research … (>8) (novel findings)
- Technically solid journals: 3-8

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

Data set publications (description of data publication)

- 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

Preprints

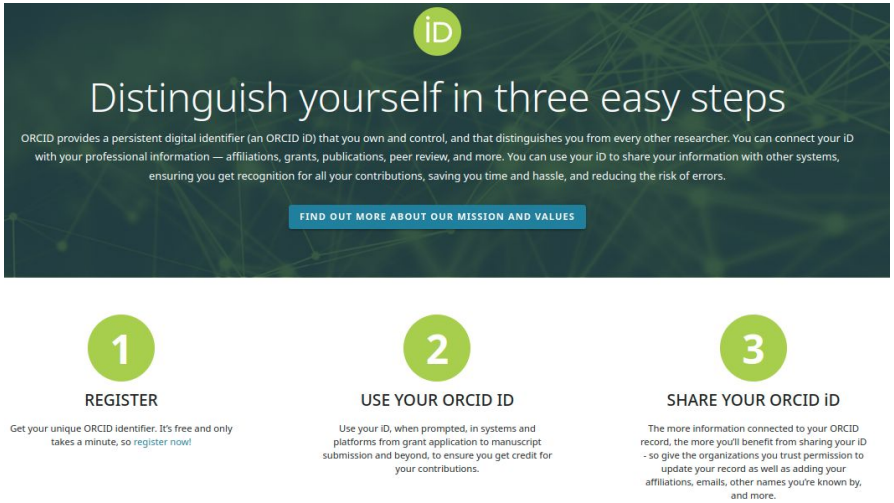
- bioRxiv: https://www.biorxiv.org/ (Cold Spring Harbor Laboratory)
  - Recommended for original research
  - Does not accept reviews

- Preprints.org: https://www.preprints.org/ (operated by MDPI)
  - Best place for reviews
  - Should be avoided; suggested by criticized MDPI journals

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

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/

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):
  - X/Twitter
  - Blogs & news outlets
  - Reddit
  - CitationTools

Data availability

- Open access journals require freely available data sets and scripts/methods

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

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

Sequencing data submission to ENA

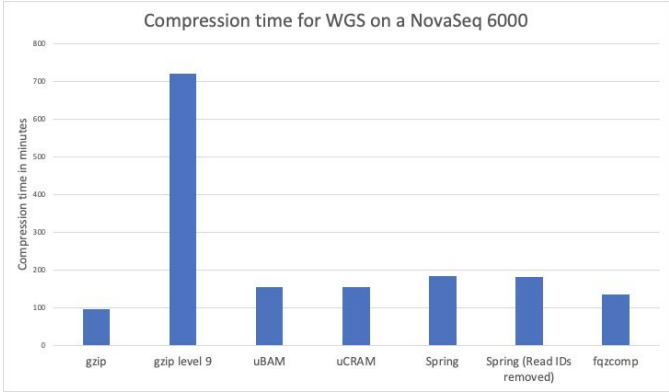
- Databases: GEO/SRA/ENA (ENA for Europeans)
- Hashes for submission check: md5sum, sha256
- Filezilla for transfer
- Specify accessions (IDs) in data availability statement
- Sharing metadata is crucial for efficient reuse

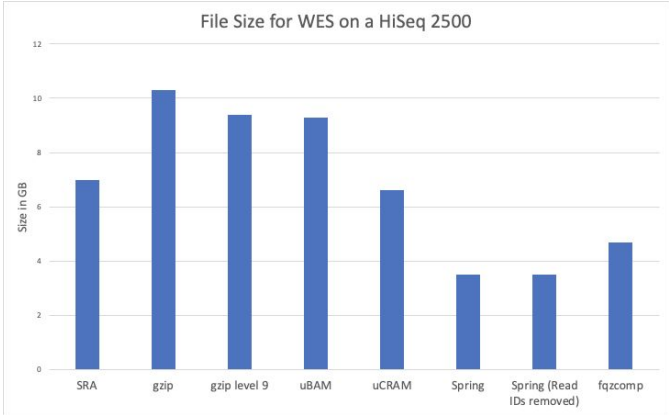
Hashes (sha256sum & md5sums)

- Hash = digital fingerprint of a file
- Hashes are useful to check completeness of file transfer
- Md5sum (128 bit): frequently used hash to validate file completeness
- Sha256sum (256 bit): more secure hash to exclude file manipulation

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

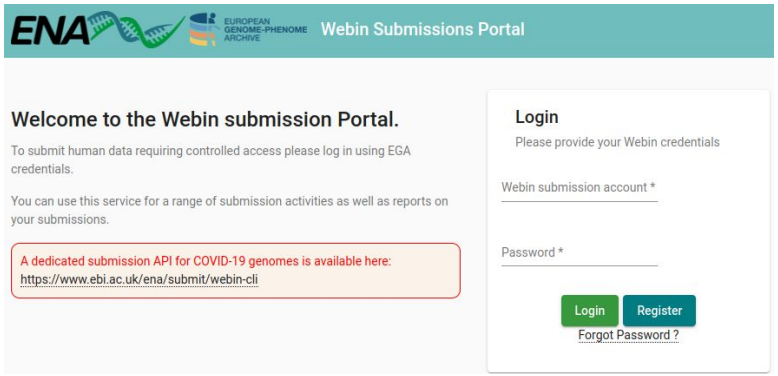
tar

- 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

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

How to submit reads (to ENA)? - read infos part 1

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

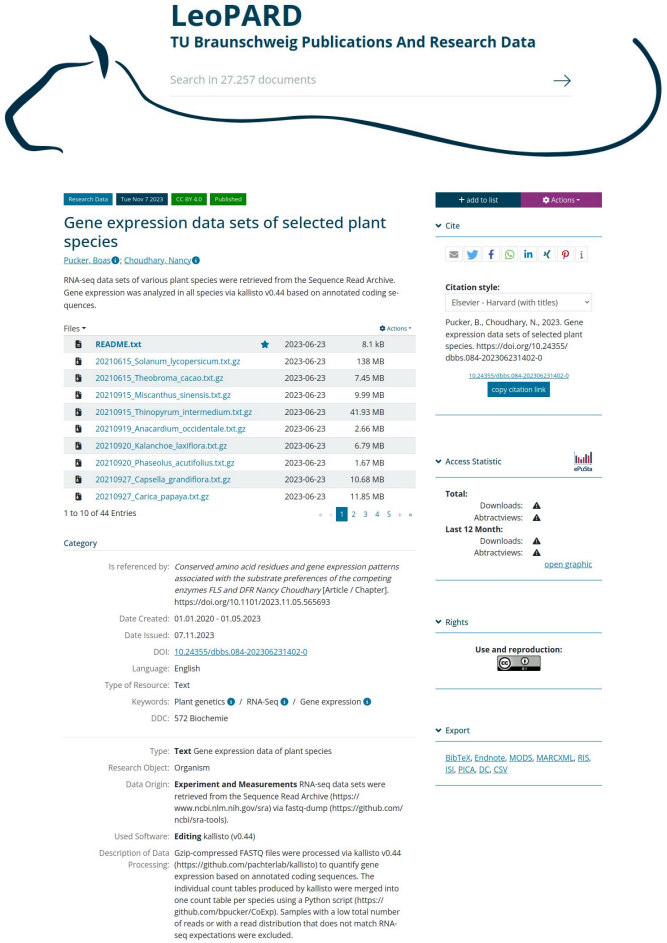
How to submit reads (to ENA)? - read infos part 2

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

Data publications: LeoPARD

- LeoPARD is operated by the library at TU Braunschweig
- Members of TU Braunschweig can publish data sets which get a DOI
- Data submission via NextCloud folder
- REAME contains metadata
- Publication agreement to transfer permission
- OpenAccess publishing



Pucker & Choudhary, 2023: 10.24355/dbbs.084-202306231402-0

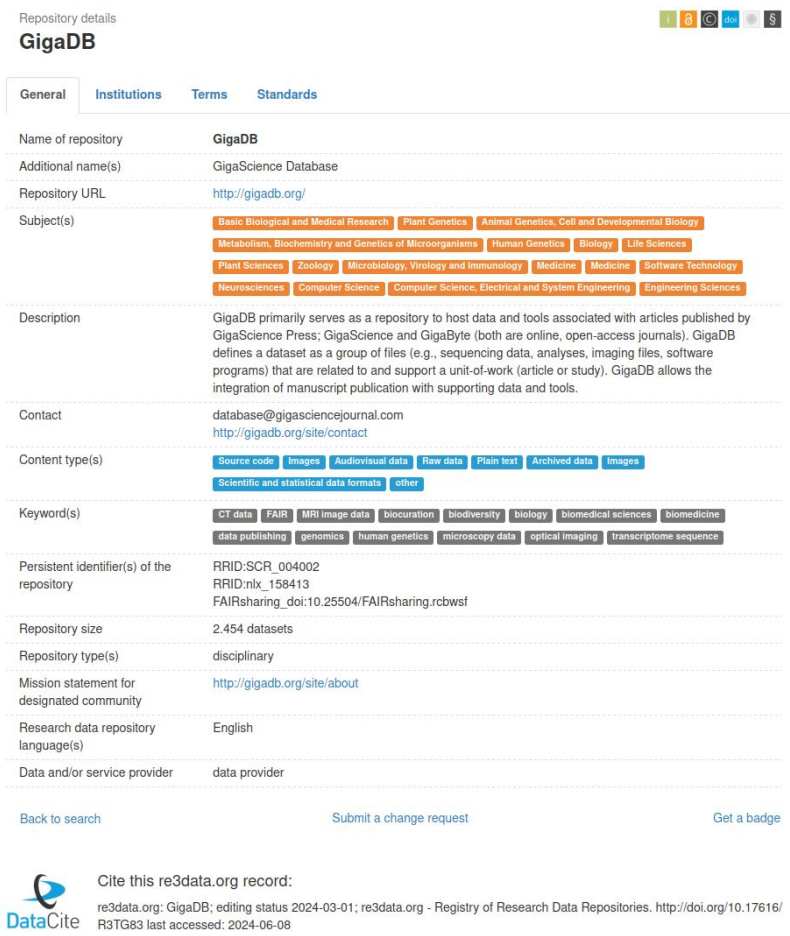
Data publications: Dryad

- Suitable for all types of data
- Accepted by open access journals
- Nonprofit membership organization
- Data publishing charges to cover costs

https://datadryad.org/stash

Data publications: GigaDB

- Publishing large datasets under CC0 license
- Connected to journal GigaScience
- Offers data publication service for other journals
- Fees apply to publishing datasets in GigaDB



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 http://gigadb.org/site/contact
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:nlx_158413 FAIRsharing_doi:10.25504/FAIRsharing.rcbwsf
Repository size	2.454 datasets
Repository type(s)	disciplinary
Mission statement for designated community	http://gigadb.org/site/about
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 2024-03-01; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R3TG83 last accessed: 2024-06-08

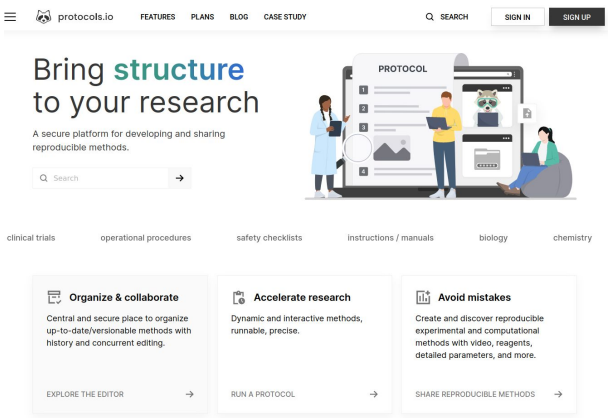
https://www.re3data.org/repository/r3d100010478

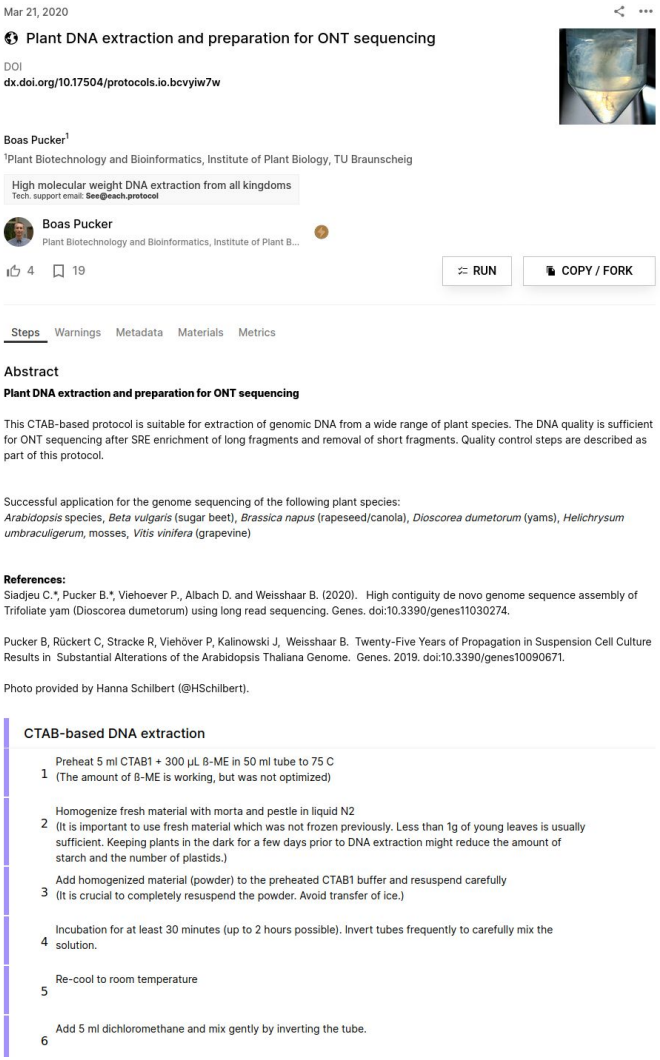
OpenAccess/OpenData/OpenMethods

- OpenAccess is making scientific publication accessible to everyone
- OpenData makes data underlying an open access publication accessible
- OpenMethods is important to enable reproduction of research
- Scientific publications will gain reliability by being easily reproducible

protocols.io

- Freely accessible protocols and ability to publish protocols for free (OpenResearch)
- DOI assignment to protocols
- Support digital lab books and LIMS integration
- Subscription model for private protocols

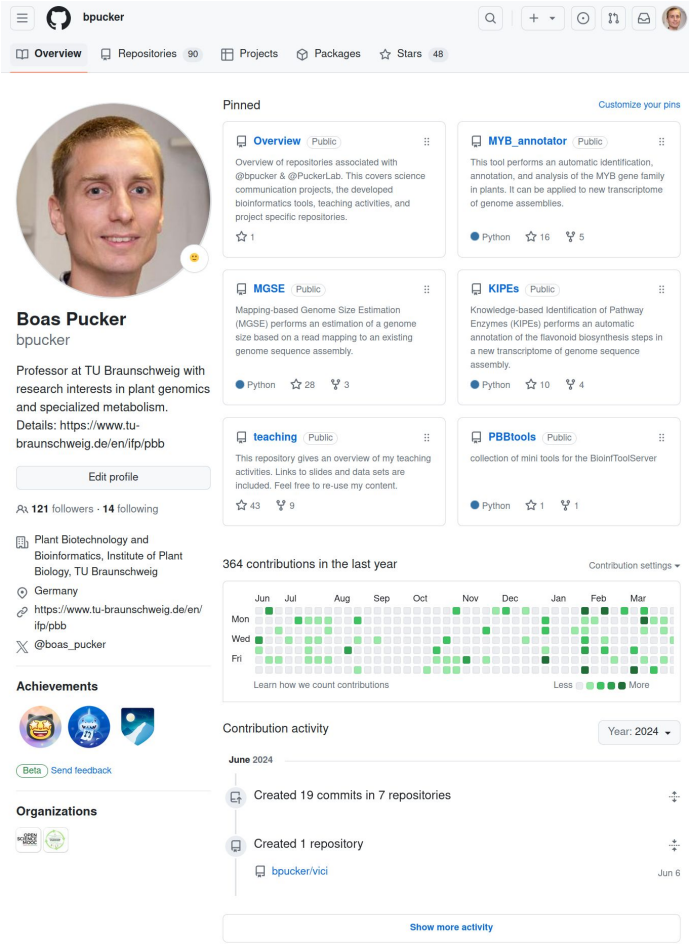




https://dx.doi.org/10.17504/protocols.io.bcvyiw7w

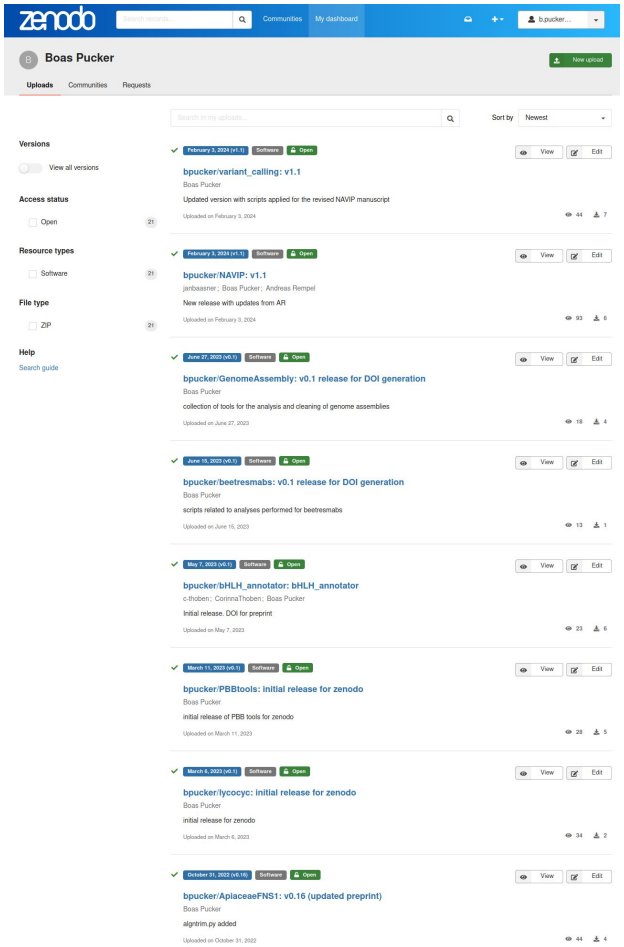
Sharing scripts: GitHub

- Platform for version control of scripts
- Activity and all changes are tracked
- Content organized into repositories
- Established for publicly sharing scripts
- Other data sets could be shared through GitHub (<25MB per file)
- Commercial platform that might not be continue to be freely available forever



Sharing scripts: Zenodo

- Backup solution for GitHub repositories to reference a repository snapshot in publications
- Zenodo is operated by CERN's Data Centre
- DOI generation enables citation
- Automatic connection to GitHub to generate archive for each repository release
- Statistics about access to data sets available



https://zenodo.org

Summary

- Introduction to scientific publishing business
- Submission of sequencing data to ENA
- Data publications (LeoPARD, Dryad)
- OpenMethods & protocols.io
- Depositing scripts in GitHub and Zenodo

Time for questions!