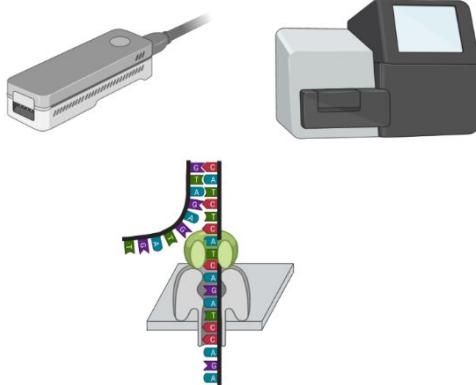
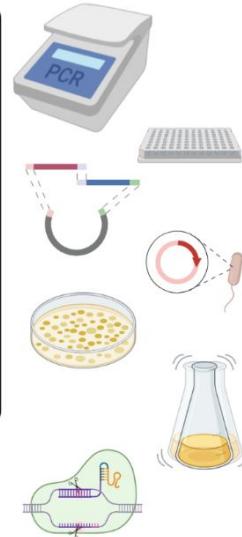
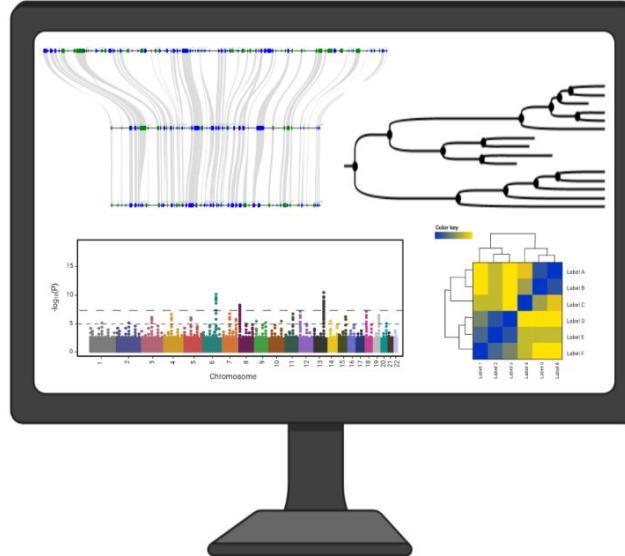




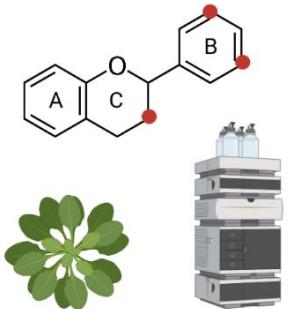
Technische
Universität
Braunschweig



Plant Biotechnology
and Bioinformatics



species biosynthesis proteins analysis different conditions
biosynthesis species activities bellman variants H2R3-MYB
within genes site data functionally Col locora variant
dissolve site data divergent variants non-canonical
sequences KEGG multiple protein annotation level identified
single reference structure synthesis amino acid evolution
sites annotations pathways accessions pathway evolutionary
plants pigments model genome systems biology long insertion
pigments Kewenta genera accessions identification Arabidopsis
key against canonical for conserved free Canophylales
flavonoid conservation sequencing sequencing evolution
plants gene read transcription synthesis MYB introns residues RNA-Seq
genomes across thaliana



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

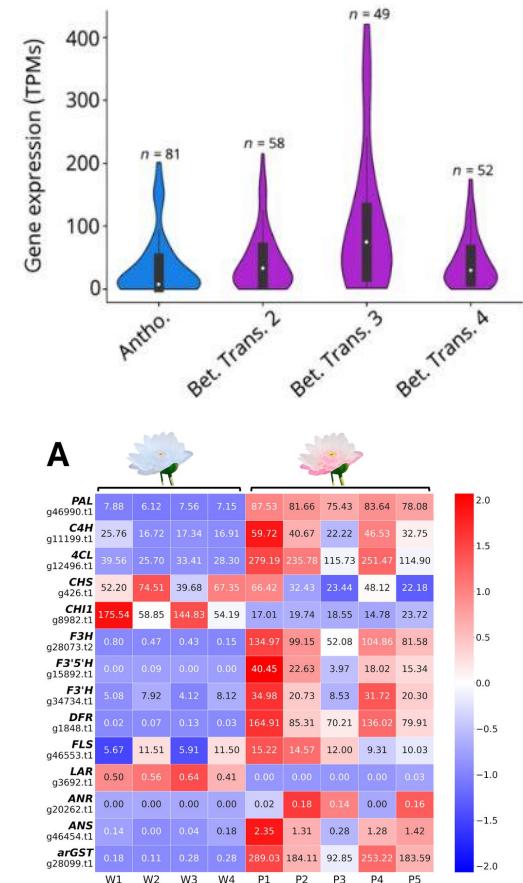
Prof. Dr. Boas Pucker (Plant Biotechnology and Bioinformatics)

Outline - Summarizing & visualizing

- General advice and file types (Inkscape + Biolcons)
- 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

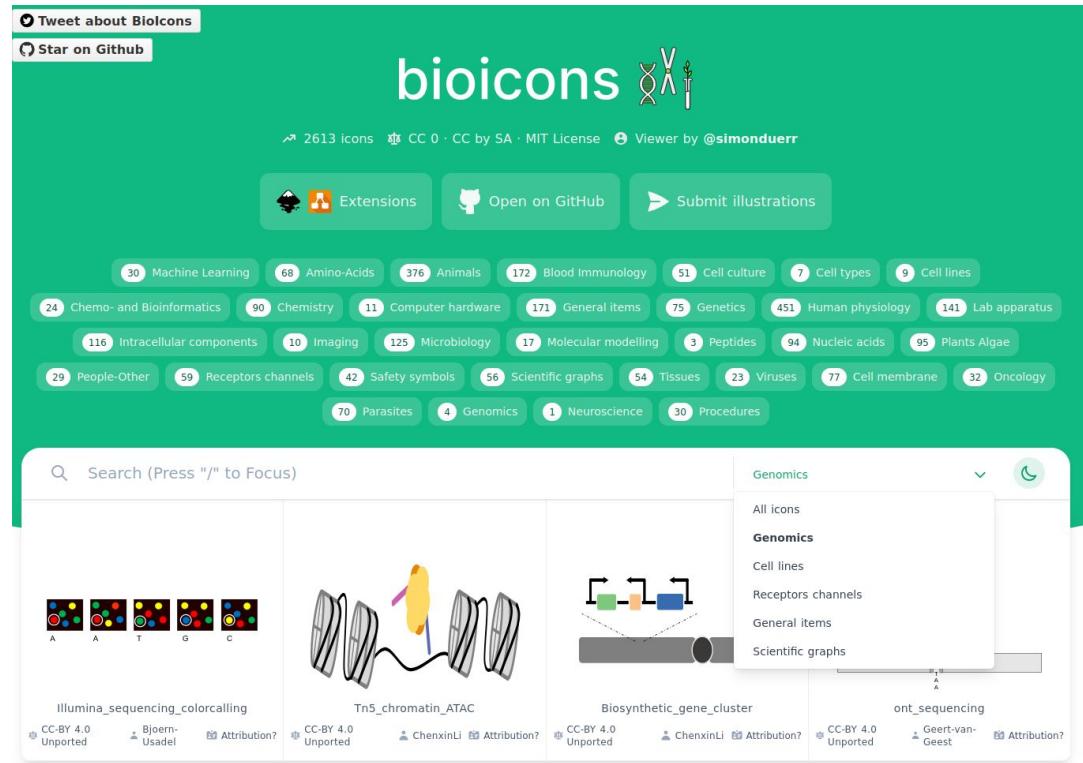
- JPG/JPEG (Joint Photographic Experts Group): JFIF, SPIFF, JNG
 - lossy data compression
- PNG (Portable Network Graphics):
 - lossless data compression; transparency possible, 32-bit RGBA colors
- SVG (Scalable Vector Graphics):
 - XML-based; support for interactivity and animations; web-friendly
- 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

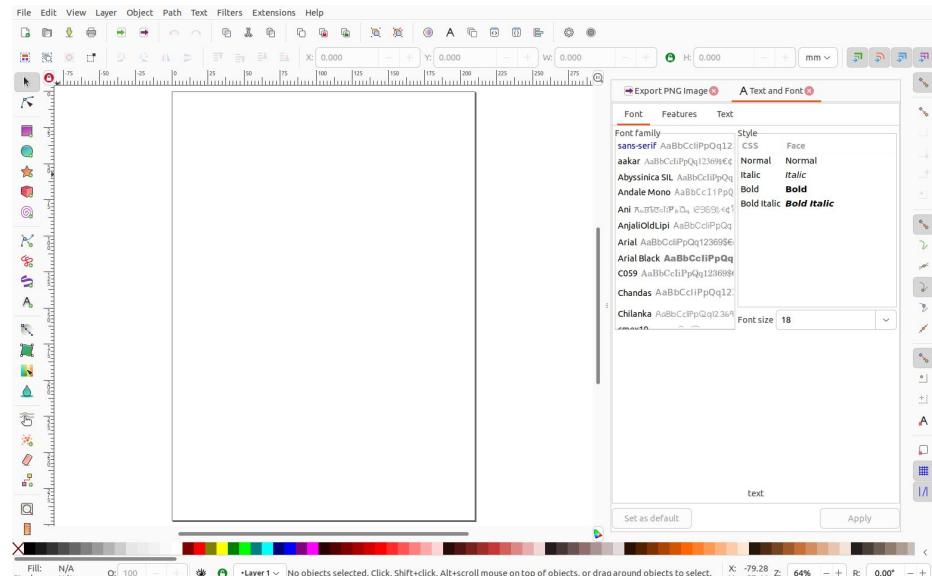
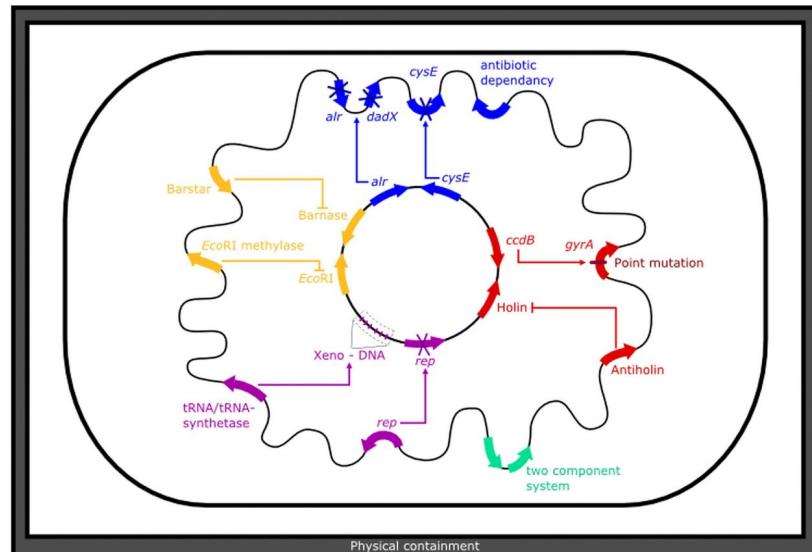
- Collection of free biology-related icons in SVG format
- License is CC0 or CC BY
- Keyword search and filter by topics possible
- Number of figures limited, but growing quickly
- @PuckerLab is actively contributing to this resource



Show tour · Made with Tailwind, Vue.js, Nuxt and Inkscape · Contribute your own icons by creating a pull request · Imprint

Manual figure editing with InkScape

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

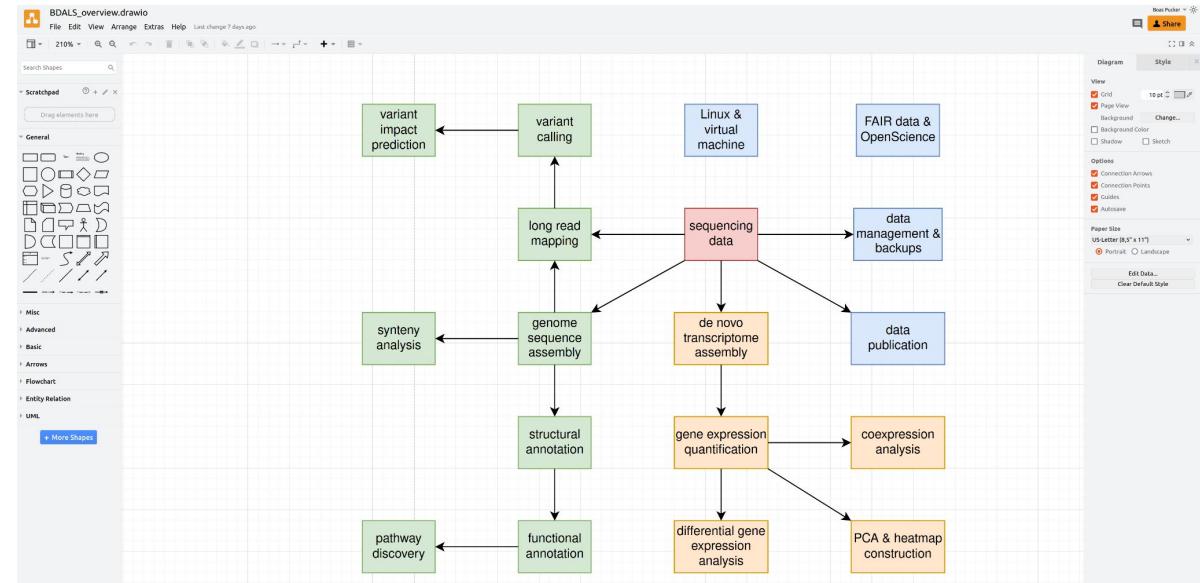


<https://inkscape.org/cs/>

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

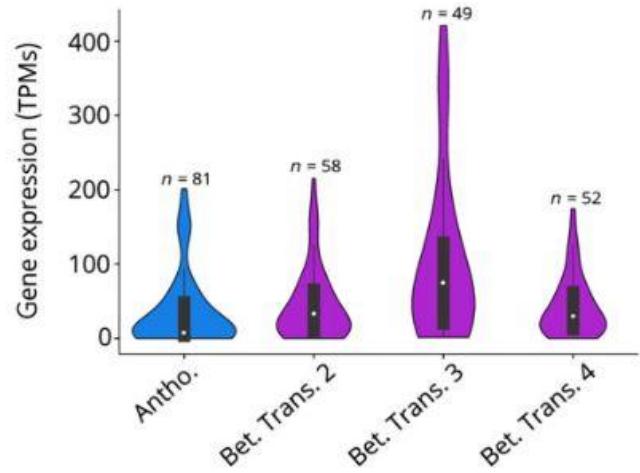
Manual figure editing with draw.io

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



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

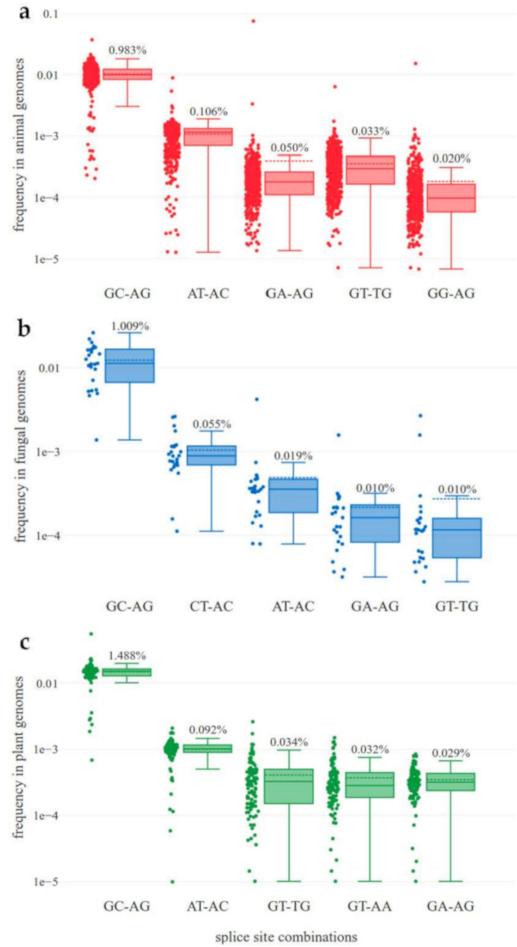


matplotlib

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

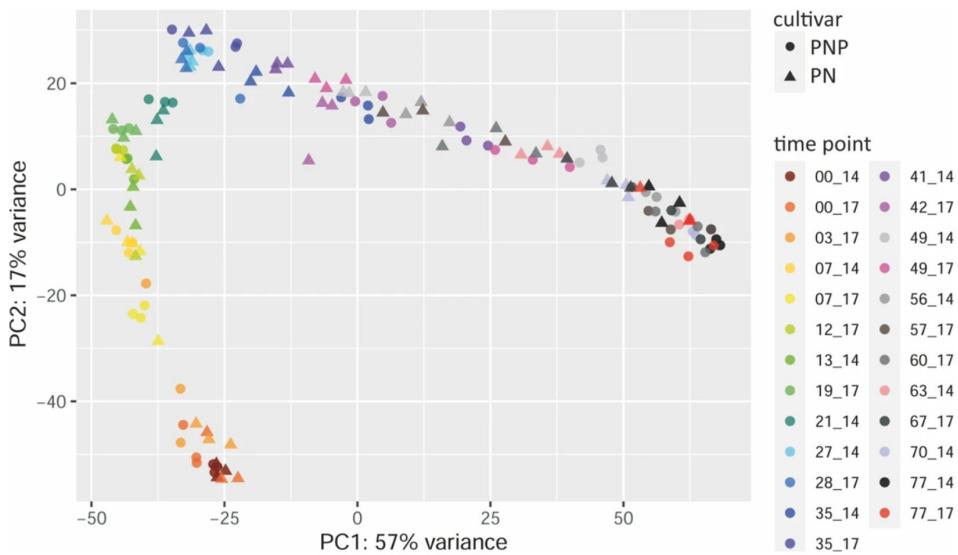


Frey & Pucker, 2020: 10.3390/cells9020458
<https://github.com/plotly/plotly.py>



Automatic figure generation - ggplot2

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



Overview

ggplot2 is a system for declaratively creating graphics, based on [The Grammar of Graphics](#). You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

Installation

```
# The easiest way to get ggplot2 is to install the whole tidyverse:  
install.packages("tidyverse")  
  
# Alternatively, install just ggplot2:  
install.packages("ggplot2")  
  
# Or the development version from GitHub:  
# install.packages("pak")  
pak::pak("tidyverse/ggplot2")
```

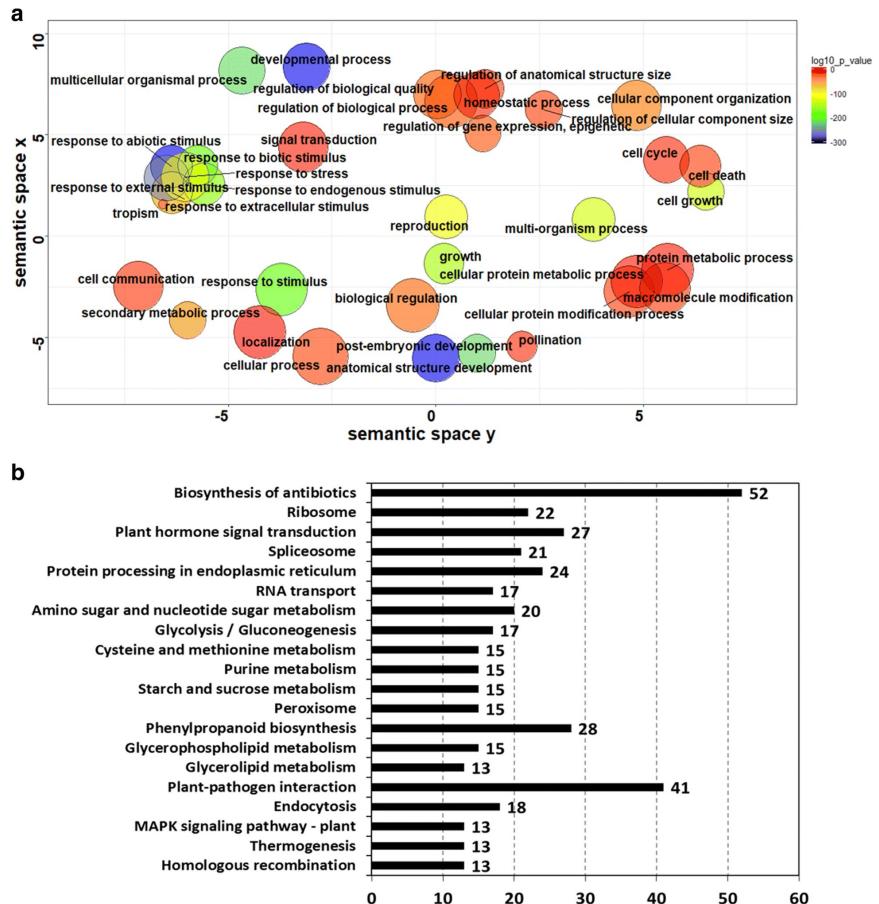
Cheatsheet



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

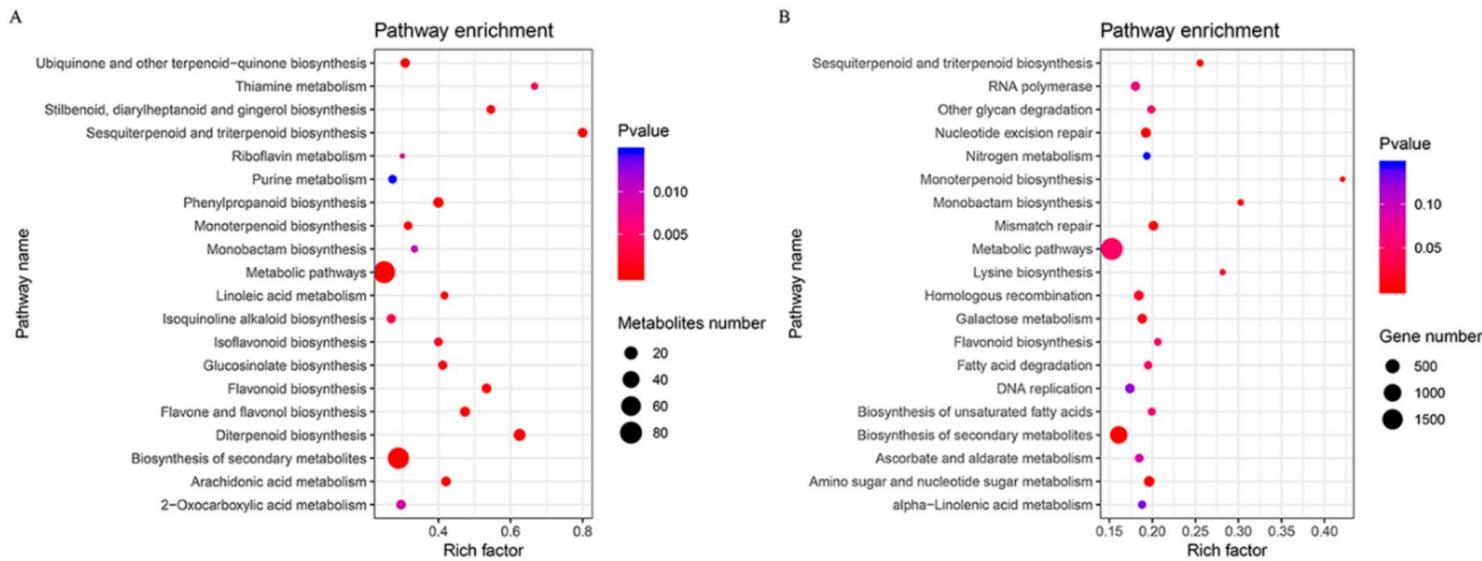
GO enrichment

- 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, ...



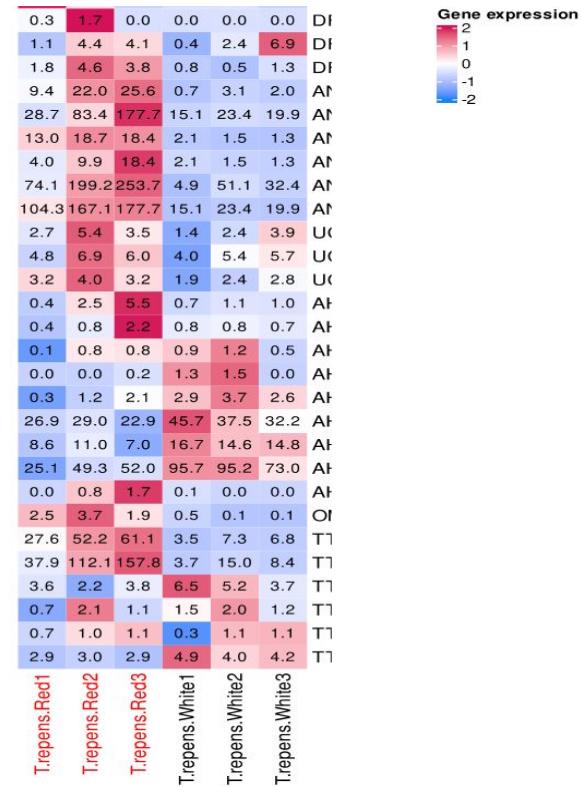
KEGG enrichment

- 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



Heatmaps

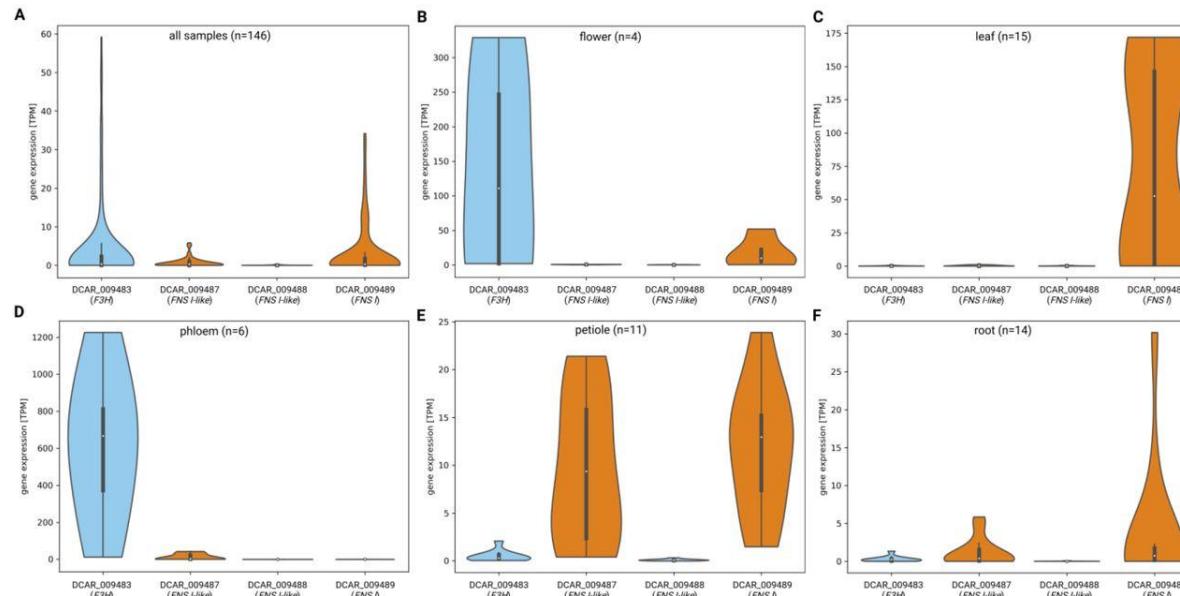
- Visualization of gene expression values based on color/color intensity
- Red (heat) often indication of high expression
- Different color ranges are available in Python/R
- Additional normalization per gene across all samples



<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

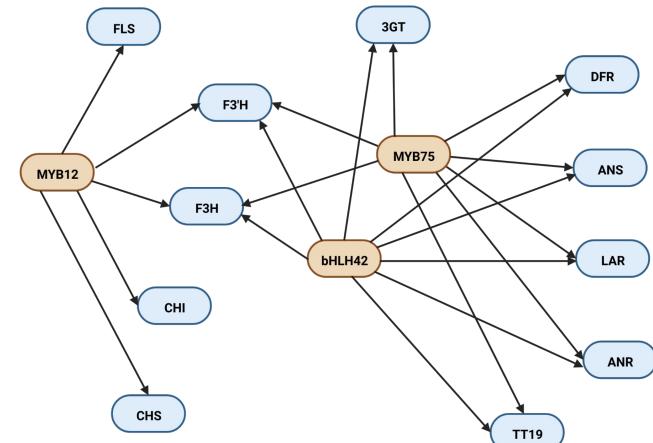


<https://doi.org/10.1371/journal.pone.0280155>



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

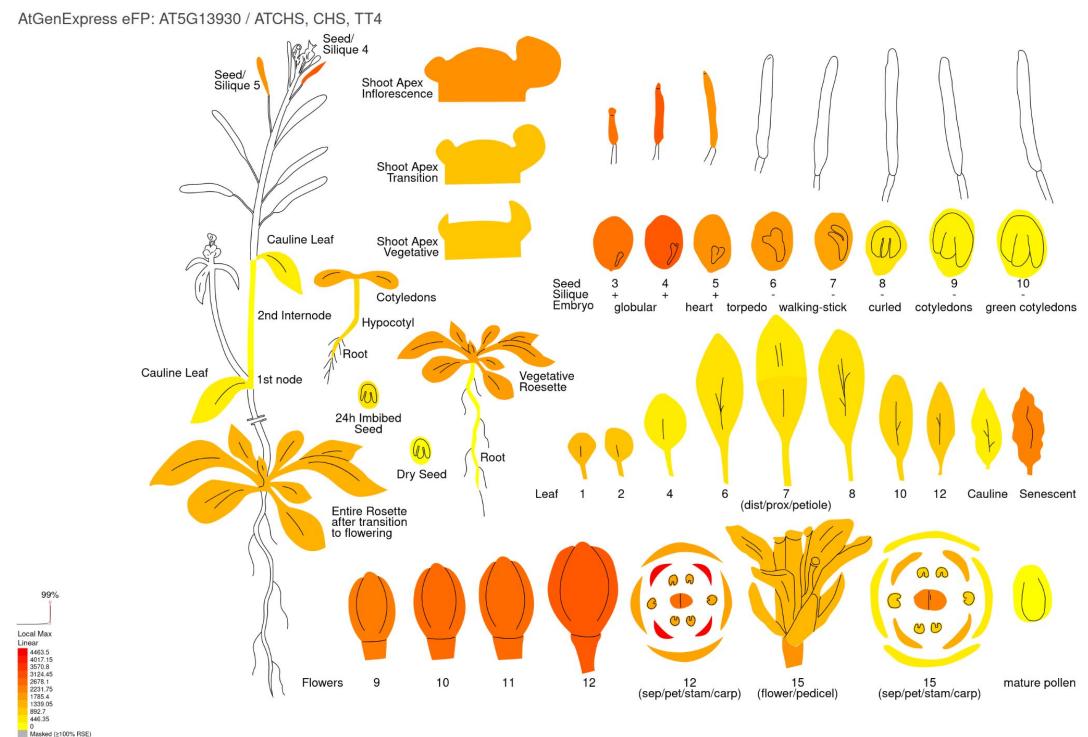
- Automatic visualization of complex networks
- Colors, shapes, symbols, and width of elements can indicate values
- Manual rearrangement of elements in network possible
- Availability (v3.10.2):
<https://cytoscape.org/download.html>



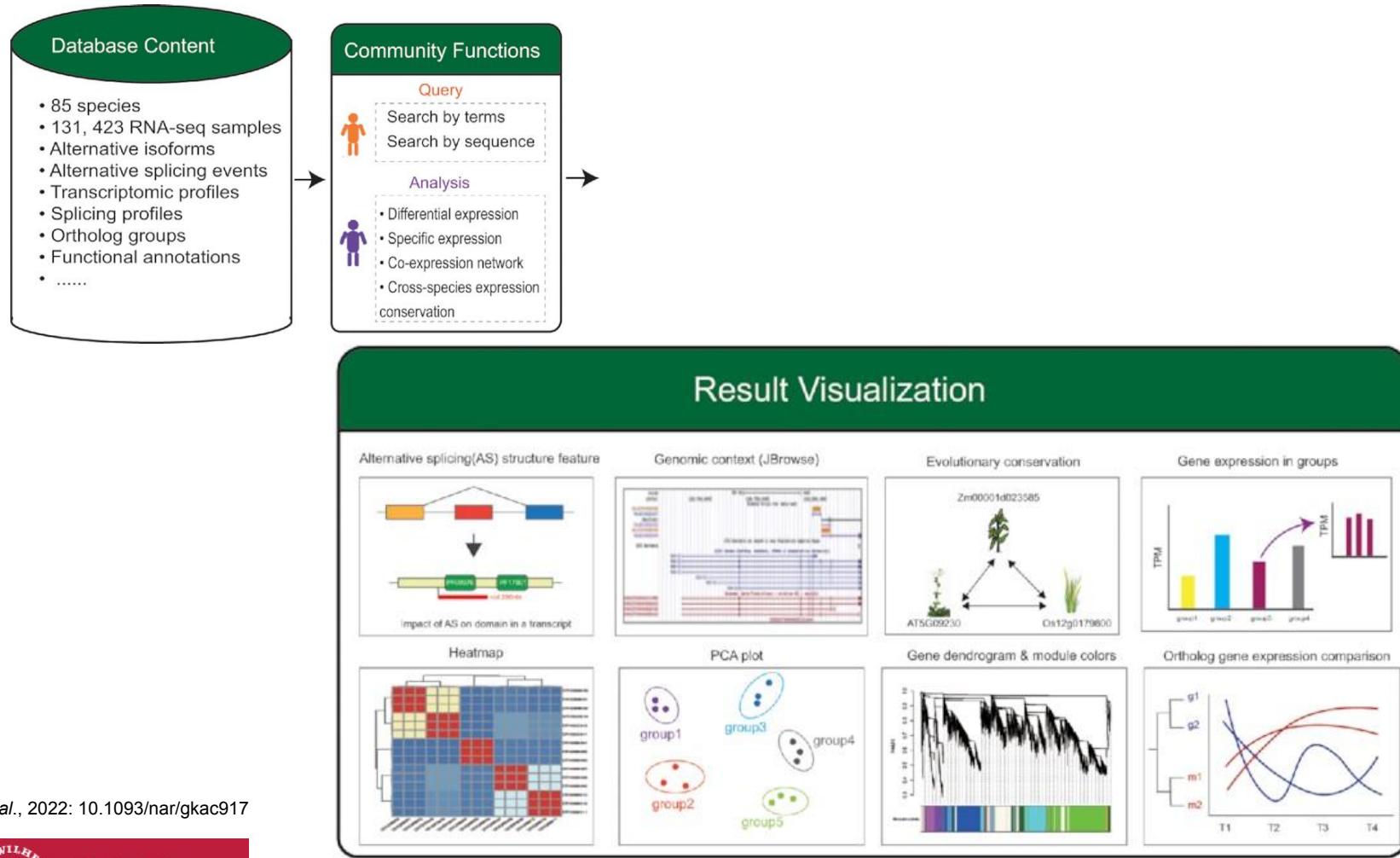
Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data. A lot of [Apps](#) are available for various kinds of problem domains, including bioinformatics, social network analysis, and semantic web.
[Learn more...](#)

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



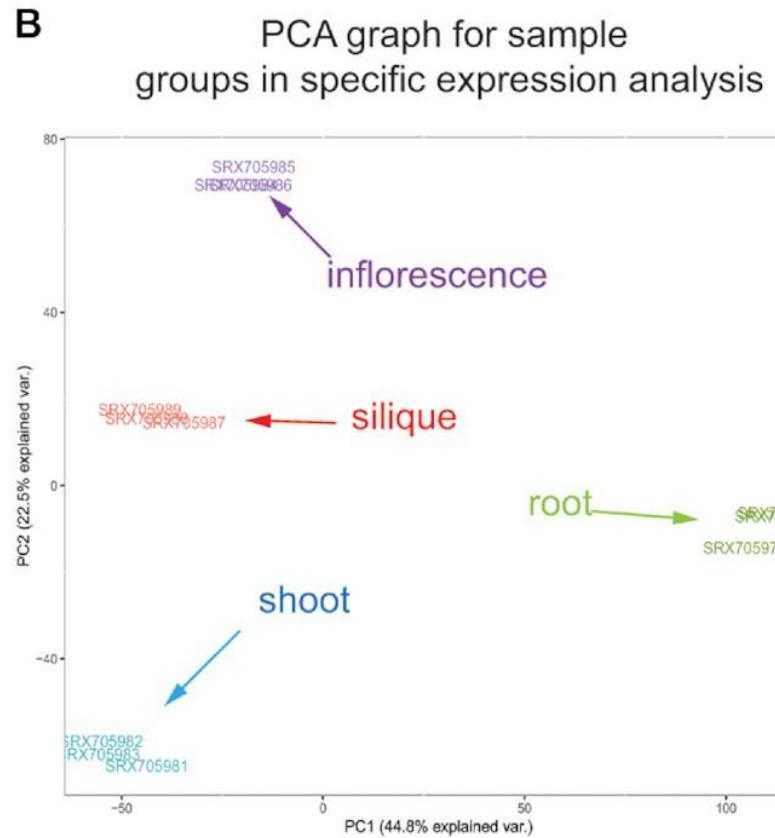
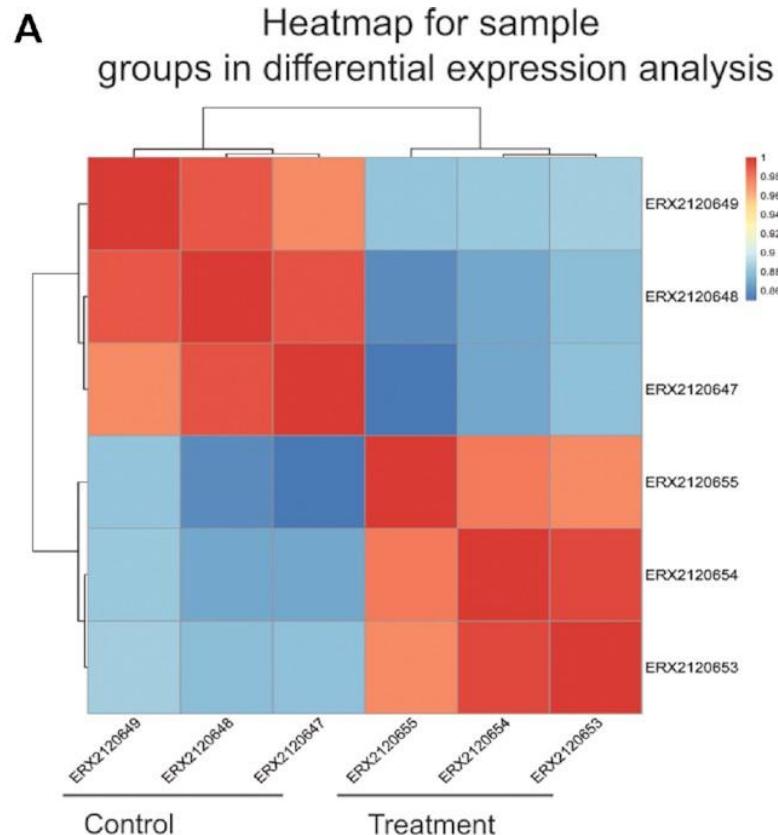
PlantExp: gene expression & alternative splicing



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

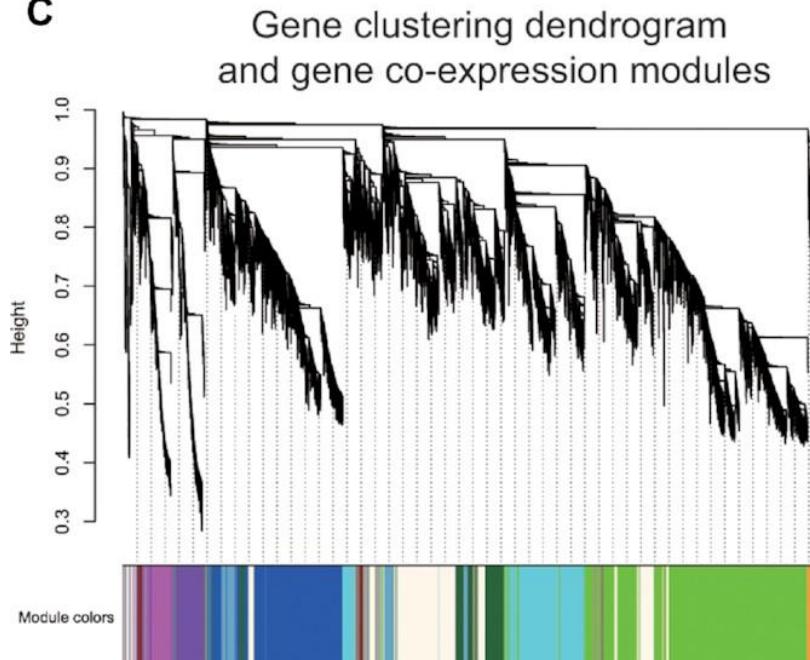


PlantExp: example plots (1)



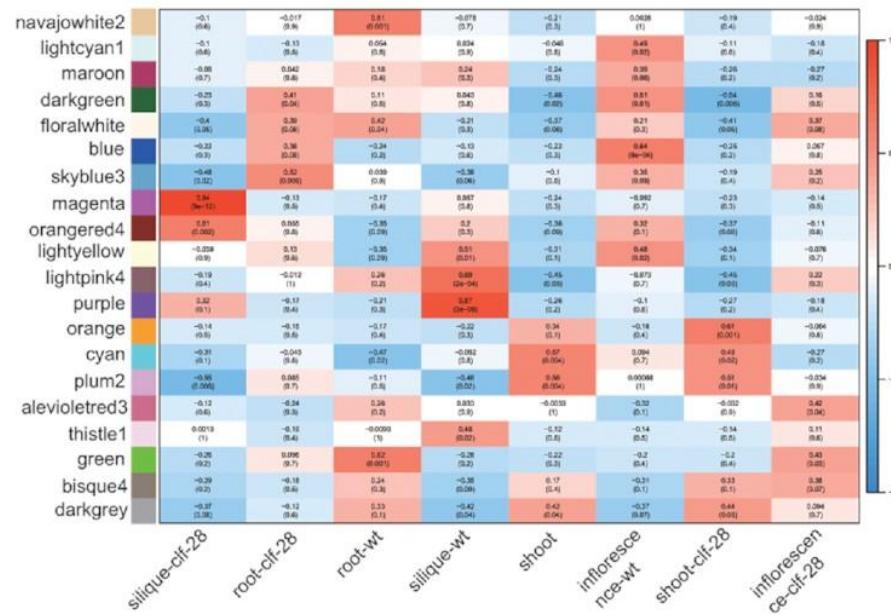
PlantExp: example plots (2)

C



D

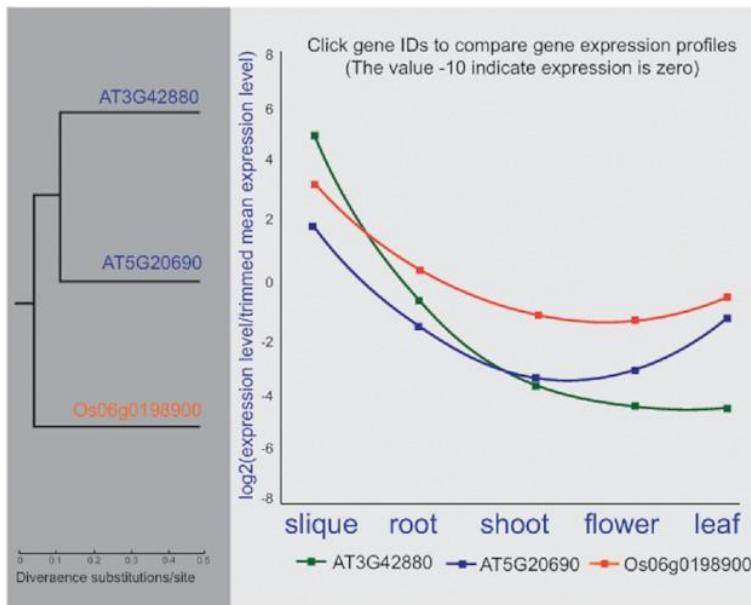
Heatmap for relationships between gene co-expression modules and sample groups



PlantExp: example plots (3)

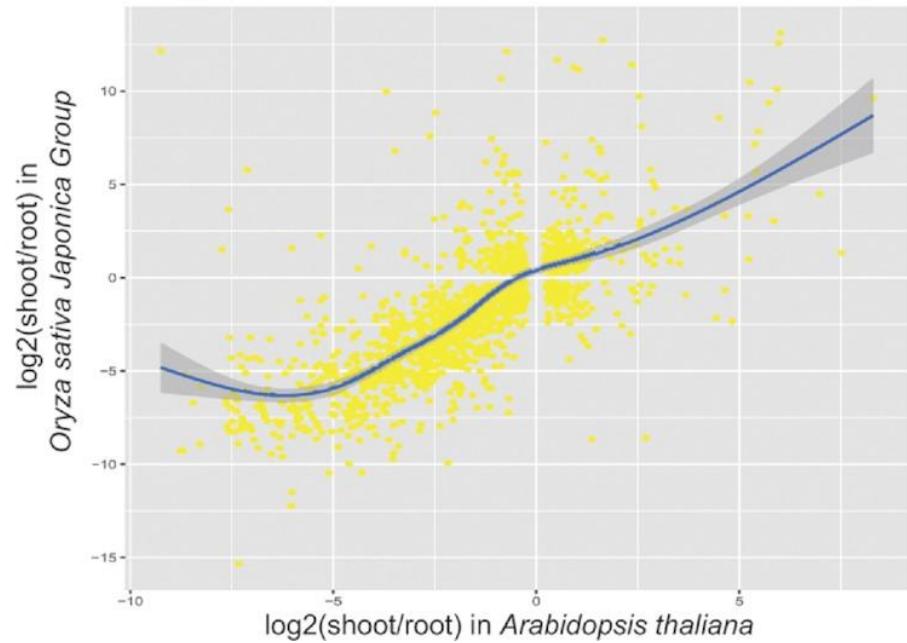
E

Ortholog phylogenetic tree
and expression profile comparison



F

Scatter plot of 1:1 ortholog expression
changes for shoot Vs. root in rice and arabidopsis



Outline - Publishing

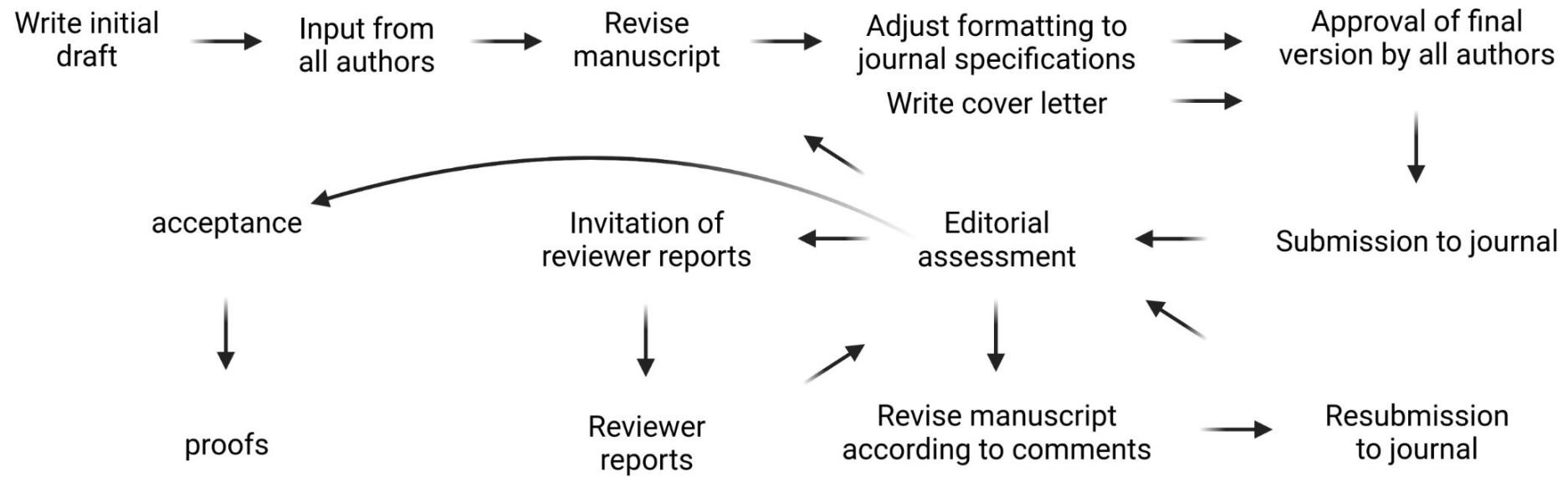
- Introduction to scientific publishing business
- Submission of sequencing data to ENA
- Data publications (LeoPARD, Dryad)
- OpenMethods & protocols.io
- Depositing scripts in GitHub and Zenodo

Motivation for publishing a scientific paper

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



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 (<i>Beta vulgaris</i>) R Stracke, D Holzgräfe, J Schneider, B Pucker, T Rosleff Sørensen, ... BMC Plant Biology 14 (1), 1-17	77 2014
2	The evolution of betalain biosynthesis in Caryophyllales A Timoreira, 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 <i>Arabidopsis thaliana</i> Accession Niederenz-1 Displays Presence/Absence Variation and Strong Synteny B Pucker, D Holzgräfe, T Rosleff Sørensen, R Stracke, P Venhöver, ... 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 <i>Arabidopsis thaliana</i> Nd-1 genome and its gene set B Pucker, D Holzgräfe, KB Städemann, 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 18 (1), 1-13	24 2018
8	High contiguity de novo genome assembly of trifoliate yam (<i>Dioscorea dumetorum</i>) using long read sequencing C Sladjeu, B Pucker, P Venhöver, DC Albach, B Weisshaar Genes 11 (3), 274	23 2020
9	High quality de novo transcriptome assembly of <i>Croton tiglium</i> M Haak, S Vinke, W Keller, J Drost, 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 Whifford, S Dymek, O Kerhoff, C März, O Schmidt, M Edich, J Drost, ... Journal of Biological Engineering 12 (1), 1-28	18 2018
11	Consideration of non-canonical splice sites improves gene prediction on the <i>Arabidopsis thaliana</i> Niederenz-1 genome sequence B Pucker, D Holzgräfe, 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 <i>Croton tiglium</i> B Pucker, F Reiter, 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 (<i>Musa acuminata</i>): Genome-wide identification, classification and expression patterns B Pucker, A Pandey, B Weisshaar, R Stracke GigaScience 10 (1), 1-10	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), 2324-2330	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 Rathen, 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 (<i>Beta vulgaris</i>) R Stracke, D Holzgräfe, J Schneider, B Pucker, T Rosleff Sørensen, ... 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 <i>Arabidopsis thaliana</i> Accession Niederenz-1 Displays Presence/Absence Variation and Strong Synteny B Pucker, D Holzgräfe, T Rosleff Sørensen, R Stracke, P Vienöver, ... 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-Nievaes, B Pucker, B Guo, ... New Phytologist 227 (3), 914-929	29	2020
6 A chromosome-level sequence assembly reveals the structure of the <i>Arabidopsis thaliana</i> Nd-1 genome and its gene set B Pucker, D Holzgräfe, KB Städemann, K Frey, B Huetten, 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 (<i>Dioscorea dumetorum</i>) using long read sequencing C Stadje, B Pucker, P Vienöver, DC Albach, B Weisshaar Genes 11 (3), 274	23	2020
9 High quality de novo transcriptome assembly of <i>Croton tiglium</i> M Haak, S Vinke, W Keller, J Drost, 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 Whifford, S Dymek, O Kerhoff, C März, O Schmidt, M Edich, J Drost, ... Journal of Biological Engineering 12 (1), 1-28	18	2018
11 Consideration of non-canonical splice sites improves gene prediction on the <i>Arabidopsis thaliana</i> Niederenz-1 genome sequence B Pucker, D Holzgräfe, 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 <i>Croton tiglium</i> B Pucker, F Rehner, HM Schiltbert Plants 9 (9), 1103	16	2020
13 Comparison of read mapping and variant calling tools for the analysis of plant NGS data HM Schiltbert, A Rempel, B Pucker Plants 9 (4), 439	15	2020
14 The <i>R2R3-MYB</i> gene family in banana (<i>Musa acuminata</i>): 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, SJ Glover, SF Brockington New Phytologist 229 (4), 2324-2336	11	2021
17 Integrating molecular biology and bioinformatics education B Pucker, HM Schiltbert, SF Schumacher Journal of molecular bioinformatics 10 (3)	11	2019
18 The reuse of public datasets in the life sciences: potential risks and rewards K Städemann, A Hafner, B Pucker PeerJ 8, e9954	9	2020

i10=17

Citations are career stage and field-dependent

Doctoral
student

Young
PI

Highly
cited PI

Pioneer and
genious

Researcher
with 'hot topic'

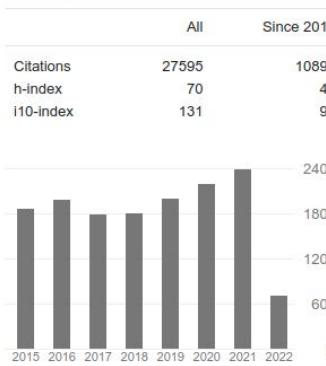
Zitiert von



Cited by

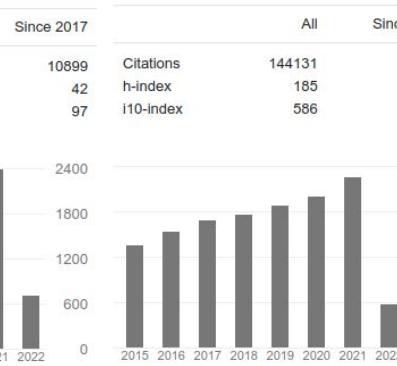


Cited by



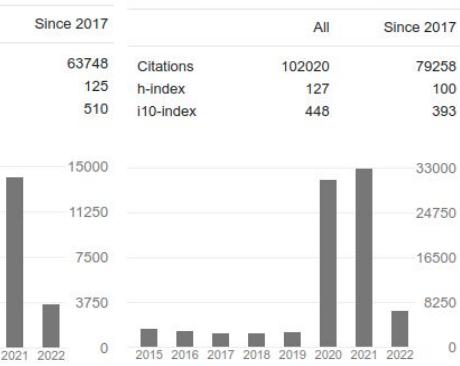
[VIEW ALL](#)

Cited by



[VIEW ALL](#)

Cited by



[VIEW ALL](#)

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

The screenshot shows a journal article page from the **nature** website. The article is titled "The growing inaccessibility of science" by Donald P. Hayes, published on 30 April 1992. The page includes metrics: 1015 accesses, 44 citations, and 27 Altmetric. Below the article title, there is a section titled "Access options" with two main choices: "Rent or Buy article" and "Subscribe to Journal". The "Rent or Buy article" option is highlighted with a red circle and shows a price of \$8.99. The "Subscribe to Journal" option shows a price of \$199.00 per year. Both options include a "Rent or Buy" button and a note that VAT will be added later in the checkout.

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



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.

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 measured 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



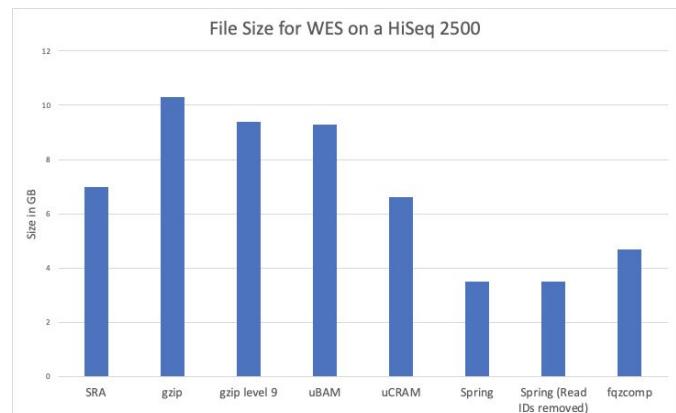
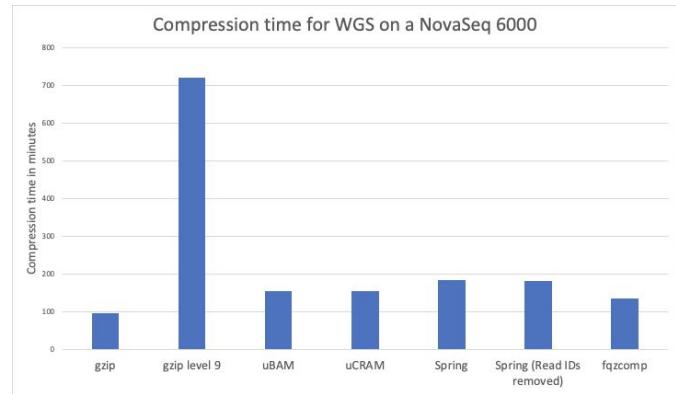
Hashes (sha256sum & md5sums)

- 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

```
bp@bp-:~/Downloads$ md5sum SRA.png
18a5a74af965752b7b716162202f745 SRA.png
bp@bp-:~/Downloads$ █
```

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



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`

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 is a logo for 'ENA' with a DNA helix icon, followed by 'EUROPEAN GENOME-PHENOME ARCHIVE' and 'Webin Submissions Portal'. Below this, a heading says 'Welcome to the Webin submission Portal.' A note indicates that for human data requiring controlled access, users should log in using EGA credentials. It also mentions that the service supports various submission activities and generates reports. A red-bordered box highlights a link to a dedicated COVID-19 submission API: <https://www.ebi.ac.uk/ena/submit/webin-cli>. To the right, there is a 'Login' form with fields for 'Webin submission account *' and 'Password *', and buttons for 'Login', 'Register', and 'Forgot Password ?'.

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

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

The screenshot shows a search result for "Gene expression data sets of selected plant species". The results list various plant species with their corresponding file names and sizes. The interface includes navigation buttons, citation style options (Elsevier - Harvard), access statistics (Downloads, Abstractviews), rights information, and export options (BibTeX, Endnote, MODS, MARCXML, RIS, PICA, DC, CSV).

Search in 27,257 documents

LeoPARD
TU Braunschweig Publications And Research Data

Gene expression data sets of selected plant species

Pucker, Boas | Choudhary, Nancy

RNA-seq data sets of various plant species were retrieved from the Sequence Read Archive. Gene expression was analyzed in all species via kallisto v0.44 based on annotated coding sequences.

File	Description	Date	Size
README.txt		2023-06-23	8.1 kB
20210615_Solanum_lycopersicum.txt.gz		2023-06-23	138 MB
20210615_Theobroma_cacao.txt.gz		2023-06-23	7.45 MB
20210915_Miscanthus_sinensis.txt.gz		2023-06-23	9.99 MB
20210915_Thinopyrum_intermedium.txt.gz		2023-06-23	41.93 MB
20210919_Amaranthus_occidentale.txt.gz		2023-06-23	2.66 MB
20210920_Kalanchoe_laxiflora.txt.gz		2023-06-23	6.79 MB
20210920_Phaseolus_acutifolius.txt.gz		2023-06-23	1.67 MB
20210927_Capsella_grandiflora.txt.gz		2023-06-23	10.68 MB
20210927_Carica_papaya.txt.gz		2023-06-23	11.85 MB

1 to 10 of 44 Entries

Category

is referenced by: Conserved amino acid residues and gene expression patterns associated with the substrate preferences of the competing enzymes FLS and DFR Nancy Choudhary [Article / Chapter]. <https://doi.org/10.1101/2023.11.05.565693>

Date Created: 01.01.2020 - 01.05.2023
Date Issued: 07.11.2023
DOI: 10.24355/dbbs.084-202306231402-0
Language: English
Type of Resource: Text
Keywords: Plant genetics / RNA-Seq / Gene expression
DOC: 572 Biochemie

Type: Text Gene expression data of plant species
Research Object: Organism
Data Origin: Experiment and Measurements RNA-seq data sets were retrieved from the Sequence Read Archive (<https://www.ncbi.nlm.nih.gov/sra>) via fastq-dump (<https://github.com/ncbi/sra-tools>).
Used Software: Editing kallisto (v0.44)
Description of Data: Gzip-compressed FASTQ files were processed via kallisto v0.44. Processing included quality filtering, removal of low-quality reads, and gene expression based on annotated coding sequences. The individual count tables produced by kallisto were merged into one count table per species using a Python script (<https://github.com/pucker/CofExp>). Samples with a low total number of reads or with a read distribution that does not match RNA-seq expectations were excluded.

Citation style:
Elsevier - Harvard (with titles)

Pucker, B., Choudhary, N., 2023. Gene expression data sets of selected plant species. <https://doi.org/10.24355/dbbs.084-202306231402-0> [copy citation link](#)

Access Statistics

Total: Downloads: ▲ Abstractviews: ▲
Last 12 Month: Downloads: ▲ Abstractviews: ▲ [open graphic](#)

Rights

Use and reproduction: [open](#)

Export

BibTeX, Endnote, MODS, MARCXML, RIS, PICA, DC, CSV

Data publications: Dryad

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



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



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

The screenshot shows the main interface of protocols.io. At the top, there's a navigation bar with 'protocol.io', 'FEATURES', 'PLANS', 'BLOG', and 'CASE STUDY'. Below the navigation is a search bar with a magnifying glass icon and a sign-in/sign-up button. The main content area features a large banner with the text 'Bring structure to your research' and 'A secure platform for developing and sharing reproducible methods.' Below this are several categories: 'clinical trials', 'operational procedures', 'safety checklists', 'Instructions / manuals', 'biology', and 'chemistry'. Further down are three call-to-action boxes: 'Organize & collaborate' (Central and secure place to organize up-to-date/versionable methods with history and concurrent editing.), 'Accelerate research' (Dynamic and interactive methods, runnable, precise.), and 'Avoid mistakes' (Create and discover reproducible experimental and computational methods with video, reagents, detailed parameters, and more.). At the bottom, there are links to 'EXPLORE THE EDITOR', 'RUN A PROTOCOL', and 'SHARE REPRODUCIBLE METHODS'.

Mar 21, 2020

Plant DNA extraction and preparation for ONT sequencing

DOI

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



Boas Pucker¹

¹Plant Biotechnology and Bioinformatics, Institute of Plant Biology, TU Braunschweig

High molecular weight DNA extraction from all kingdoms
Tech. support email: See@each.protocol

Boas Pucker

Plant Biotechnology and Bioinformatics, Institute of Plant B...

4 19

RUN

COPY / FORK

Steps Warnings Metadata Materials Metrics

Abstract

Plant DNA extraction and preparation for ONT sequencing

This CTAB-based protocol is suitable for extraction of genomic DNA from a wide range of plant species. The DNA quality is sufficient for ONT sequencing after SRE enrichment of long fragments and removal of short fragments. Quality control steps are described as part of this protocol.

Successful application for the genome sequencing of the following plant species:

Arabidopsis species, *Beta vulgaris* (sugar beet), *Brassica napus* (rapeseed/canola), *Dioscorea dumetorum* (yams), *Helichrysum umbraculigerum*, mosses, *Vitis vinifera* (grapevine)

References:

Siadjeu C., Pucker B., Viehöver P., Albach D. and Weisshaar B. (2020). High contiguity de novo genome sequence assembly of Trifoliate yam (*Dioscorea dumetorum*) using long read sequencing. *Genes*. doi:10.3390/genes1030274.

Pucker B., Rückert C., Stracke R., Viehöver P., Kalinowski J., Weisshaar B. Twenty-Five Years of Propagation in Suspension Cell Culture Results in Substantial Alterations of the *Arabidopsis* Thaliana Genome. *Genes*. 2019. doi:10.3390/genes10090671.

Photo provided by Hanna Schilbert (@HSchilbert).

CTAB-based DNA extraction

1 Preheat 5 ml CTAB1 + 300 µl 8-ME in 50 ml tube to 75 C
(The amount of 8-ME is working, but was not optimized)

2 Homogenize fresh material with morta and pestle in liquid N2
(It is important to use fresh material which was not frozen previously. Less than 1g of young leaves is usually sufficient. Keeping plants in the dark for a few days prior to DNA extraction might reduce the amount of starch and the number of plastids.)

3 Add homogenized material (powder) to the preheated CTAB1 buffer and resuspend carefully
(It is crucial to completely resuspend the powder. Avoid transfer of ice.)

4 Incubation for at least 30 minutes (up to 2 hours possible). Invert tubes frequently to carefully mix the solution.

5 Re-cool to room temperature

6 Add 5 ml dichloromethane and mix gently by inverting the tube.

<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

The screenshot displays a GitHub profile for Boas Pucker (@bpucker). The profile includes a large circular photo of Boas Pucker, his name, handle, and a brief bio mentioning his affiliation with TU Braunschweig and his research interests in plant genomics and metabolism. Below the bio, there are sections for pinned repositories, recent contributions (a heatmap showing activity over the last year), contribution history (a timeline of commits and repository creation), and organizations he is part of.

Pinned

- bpucker** (Public) Overview Repositories 90 Projects Packages Stars 48
- MYB_annotator** (Public) Overview Public This tool performs an automatic identification, annotation, and analysis of the MYB gene family in plants. It can be applied to new transcriptome or genome assemblies. Python 16 5
- MGSE** (Public) Overview Public Mapping-based Genome Size Estimation (MGSE) performs an estimation of a genome size based on a read mapping to an existing genome sequence assembly. Python 28 3
- KIPEs** (Public) Overview Public Knowledge-based Identification of Pathway Enzymes (KIPEs) performs an automatic annotation of the flavonoid biosynthesis steps in a new transcriptome of genome sequence assembly. Python 10 4
- PBBtools** (Public) Overview Public collection of mini tools for the BioinToolServer Python 1 1

Contributions in the last year

364 contributions in the last year Contribution settings ▾

Contribution activity Year: 2024 ▾

June 2024

- Created 19 commits in 7 repositories
- Created 1 repository bpucker/vici

Show more activity

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

The screenshot shows the Zenodo user interface for Boas Pucker. The top navigation bar includes links for 'Search', 'Communities', 'My dashboard', and a user profile. A 'New upload' button is visible in the top right. The main area displays a list of uploaded datasets, each with a title, description, file type, and download statistics. The datasets listed are:

- February 3, 2023 (v1.1) Software: bpucker/variant_calling: v1.1
- February 3, 2023 (v1.1) Software: bpucker/NAIP: v1.1
- June 27, 2023 (v0.1) Software: bpucker/GenomeAssembly: v0.1 release for DOI generation
- June 19, 2023 (v0.1) Software: bpucker/beetresmabs: v0.1 release for DOI generation
- May 7, 2023 (v0.1) Software: bpucker/bHLH_annotator: bHLH_annotator
- March 11, 2023 (v0.1) Software: bpucker/PBTools: initial release for zenodo
- March 9, 2023 (v0.1) Software: bpucker/lycocytc: initial release for zenodo
- October 31, 2022 (v0.1) Software: bpucker/ApiaceaeFNS1: v0.16 (updated preprint)

<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!

