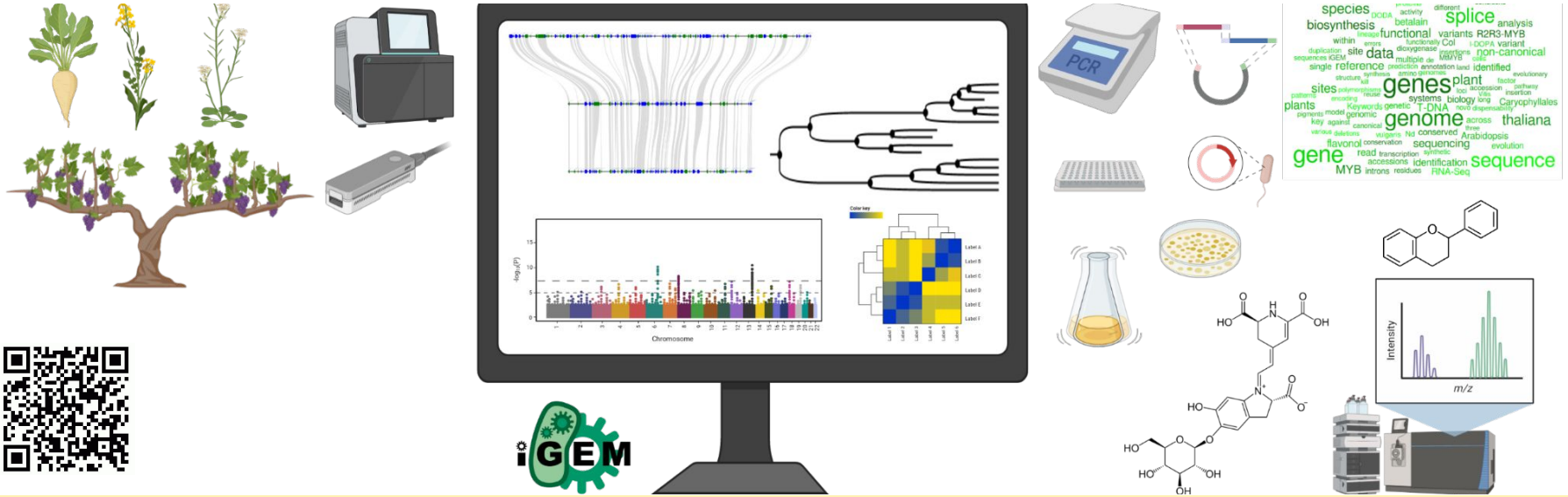
Technische Universität Braunschweig



Submitting, publishing, and re-using data

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

Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - StudIP: **Data Literacy in Genomics**
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

- Feedback, comments, or questions: b.pucker[a]tu-bs.de

Sharing data

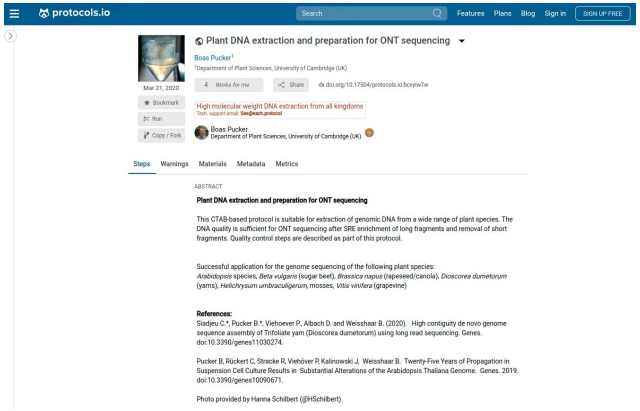
OpenData

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

OpenProtocols

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

https://www.protocols.io/welcome

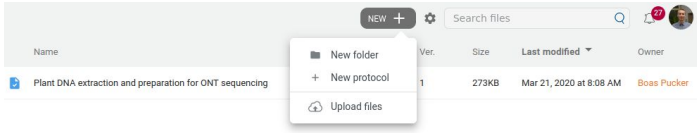


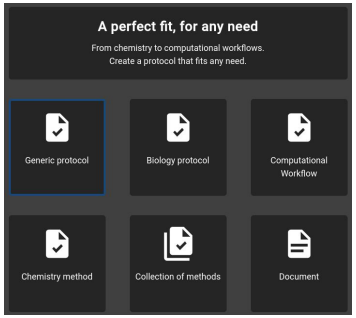
doi:10.17504/protocols.io.bcvyiw7w

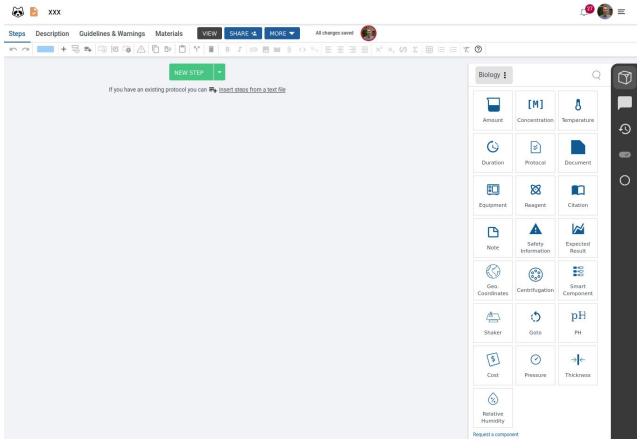
How to submit a protocol?

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

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







https://www.protocols.io/

Creative Commons Licenses

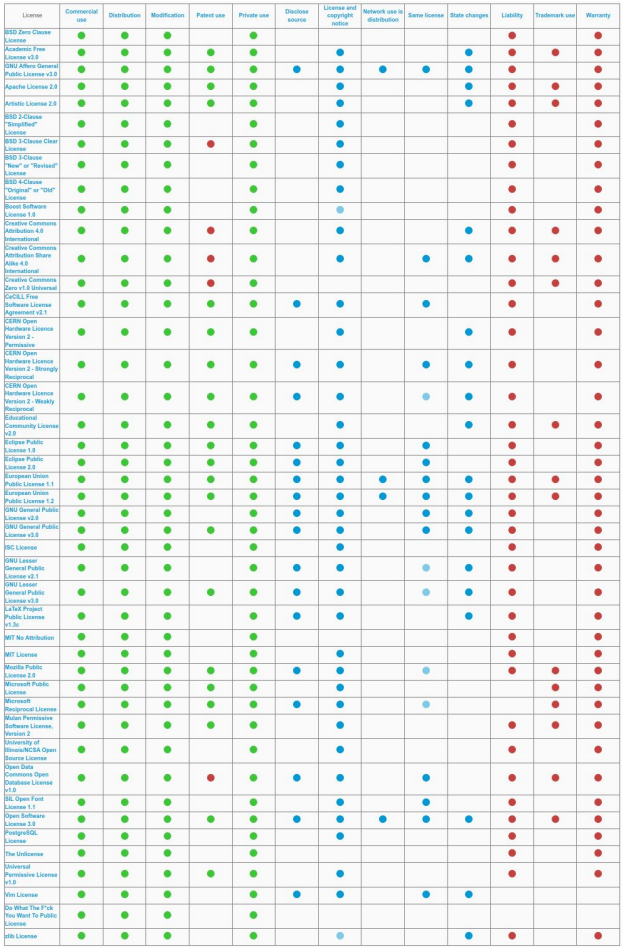
- CC0: creative commons (no restrictions)
- CC BY: no restrictions, but name authors
- CC BY-SA: name authors and share results under same license
- CC BY-NC: name authors; only non-commercial use
- CC BY-NC-SA: name authors; only non-commercial use; share under same license

License stacking

- What happens if we combine different data sets?
- CC0 + CC BY
- CC0 + CC NC-SA
- CC BY + CC BY-NC
- …

Software licenses

- MIT: leanest licence (everything is possible)
- Apache: similar to MIT, but lengthy
- GPL (General Public License): ensures that derived work remains open
- BSD (Berkeley Software Distribution): similar to MIT, but more cases specified



License	Commercial use	Distribution	Modification	Patent use	Private use	Disclose source	License and copyright notice	Network use is distribution	Same license	State changes	Liability	Trademark use	Warranty
BSD Zero Clause License	●	●	●		●						●		●
Academic Free License v3.0	●	●	●	●	●		●			●	●	●	●
GNU Affero General Public License v3.0	●	●	●	●	●	●	●	●	●	●	●		●
Apache License 2.0	●	●	●	●	●		●			●	●	●	●
Artistic License 2.0	●	●	●	●	●		●			●	●	●	●
BSD 2-Clause "Simplified" License	●	●	●		●		●				●		●
BSD 3-Clause Clear License	●	●	●	●	●		●				●		●
BSD 3-Clause "New" or "Revised" License	●	●	●		●		●				●		●
BSD 4-Clause "Original" or "Old" License	●	●	●		●		●				●		●
Boost Software License 1.0	●	●	●		●		○				●		●
Creative Commons Attribution 4.0 International	●	●	●	●	●		●			●	●	●	●
Creative Commons Attribution Share Alike 4.0 International	●	●	●	●	●		●		●	●	●	●	●
Creative Commons Zero v1.0 Universal	●	●	●	●	●						●	●	●
CeCILL Free Software License Agreement v2.1	●	●	●	●	●	●	●		●		●		●
CERN Open Hardware Licence Version 2 - Permissive	●	●	●	●	●		●			●	●		●
CERN Open Hardware Licence Version 2 - Strongly Reciprocal	●	●	●	●	●	●	●		●	●	●		●
CERN Open Hardware Licence Version 2 - Weakly Reciprocal	●	●	●	●	●	●	●		○	●	●		●
Educational Community License v2.0	●	●	●	●	●		●			●	●	●	●
Eclipse Public License 1.0	●	●	●	●	●	●	●		●		●		●
Eclipse Public License 2.0	●	●	●	●	●	●	●		●		●		●
European Union Public License 1.1	●	●	●	●	●	●	●	●	●	●	●	●	●
European Union Public License 1.2	●	●	●	●	●	●	●	●	●	●	●	●	●
GNU General Public License v2.0	●	●	●		●	●	●		●	●	●		●
GNU General Public License v3.0	●	●	●	●	●	●	●		●	●	●		●
ISC License	●	●	●		●		●				●		●
GNU Lesser General Public License v2.1	●	●	●		●	●	●		○	●	●		●
GNU Lesser General Public License v3.0	●	●	●	●	●	●	●		○	●	●		●
LaTeX Project Public License v1.3c	●	●	●		●	●	●			●	●		●
MIT No Attribution	●	●	●		●						●		●
MIT License	●	●	●		●		●				●		●
Mozilla Public License 2.0	●	●	●	●	●	●	●		○		●	●	●
Microsoft Public License	●	●	●	●	●		●					●	●
Microsoft Reciprocal License	●	●	●	●	●	●	●		○			●	●
Mulan Permissive Software License, Version 2	●	●	●	●	●		●				●	●	●
University of Illinois/NCSA Open Source License	●	●	●		●		●				●		●
Open Data Commons Open Database License v1.0	●	●	●	●	●	●	●		●		●	●	●
SIL Open Font License 1.1	●	●	●		●		●		●		●		●
Open Software License 3.0	●	●	●	●	●	●	●	●	●	●	●	●	●
PostgreSQL License	●	●	●		●		●				●		●
The Unlicense	●	●	●		●						●		●
Universal Permissive License v1.0	●	●	●	●	●		●				●		●
Vim License	●	●	●		●	●	●		●	●			
Do What The F*ck You Want To Public License	●	●	●		●								
zlib License	●	●	●		●		○			●	●		●

https://choosealicense.com/appendix/

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?

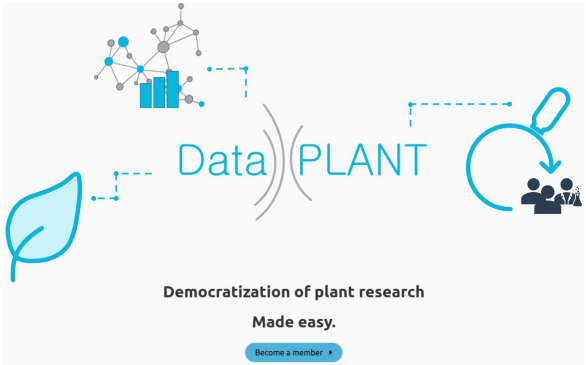
FAIR data

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

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

nfdi4plants

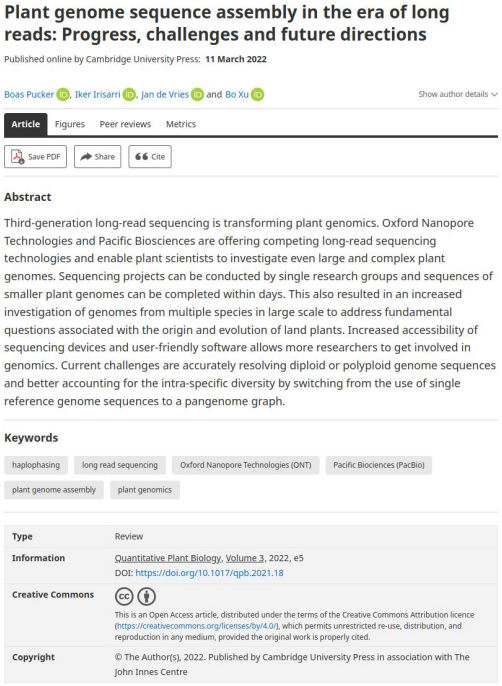
- NFDI = Nationale Forschungsdaten Infrastruktur e.V.
- Provide sustainable, annotated data management platform
- Pave way to pure data publications with research context
- Omics and imaging at petabyte size
- ARC = annotated research context
- Add user-oriented services to existing IT infrastructure (make submissions easy)



https://nfdi4plants.de/

JSON

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



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

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

Boas Pucker, Iker Irisarri, Jan de Vries and Bo Xu — Show author details

Article | Figures | Peer reviews | Metrics

Save PDF | Share | Cite

Abstract

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

Keywords

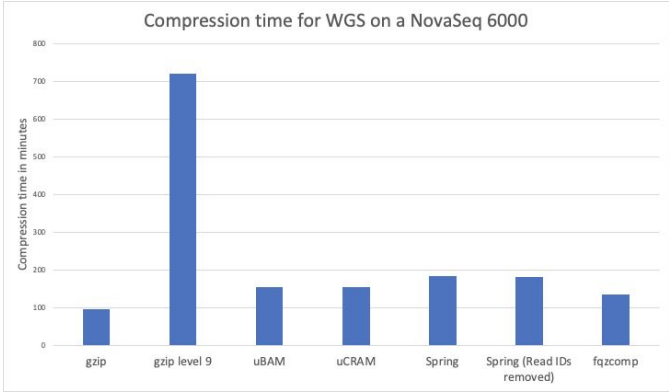
haplophasing | long read sequencing | Oxford Nanopore Technologies (ONT) | Pacific Biociences (PacBio) | plant genome assembly | plant genomics

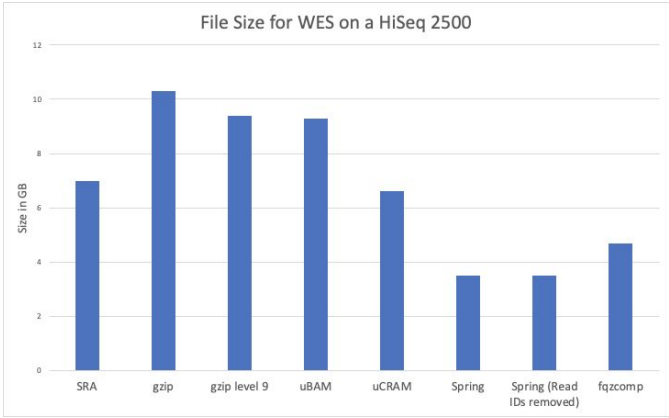
Type	Review
Information	Quantitative Plant Biology, Volume 3, 2022, e5 DOI: https://doi.org/10.1017/qpb.2021.18
Creative Commons	This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted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

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

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`

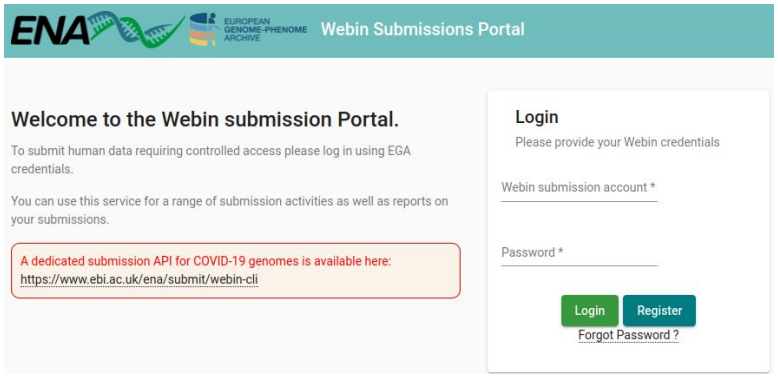
http://www.fibel.org/linux/lfo-0.6.0/node370.html

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

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:

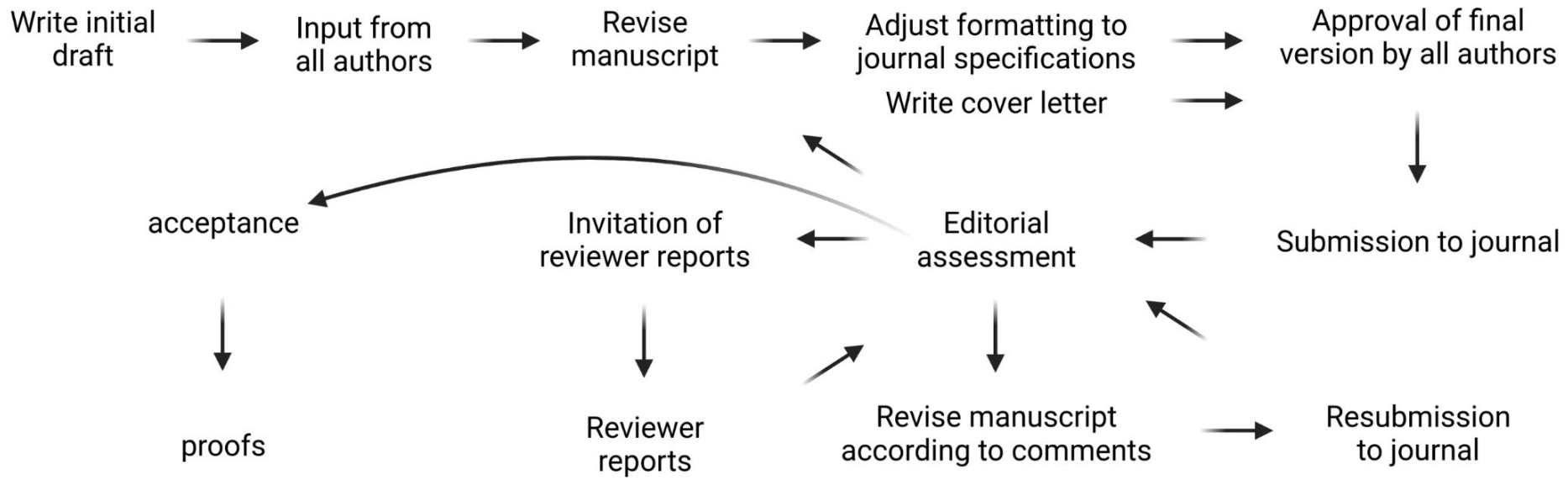
- File_type: FASTQ
- Library_layout: PAIRED
- Insert_size: 600
- Forward_file_name: fw_file.fastq.gz
- Forward_file_md5: jel9aks5joe8iaj1ie2lfk4jsk6flji
- Forward_file_unencrypted_md5:
- Reverse_file_name: rv_file.fastq.gz
- Reverse_file_md5: k1ea0wi7oji32so45jbae6fo81337xd
- reverse_file_unencrypted_md5

Publishing a research paper

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

Technische Universität Braunschweig

Institute for Plant Biology
Technische Universität Braunschweig

Plants
Editorial Office

Prof. Dr. Boas Pucker
Mendelssohnstr. 4
38106 Braunschweig
Germany

09. March 2022

Biochemistry and molecular basis of intracellular flavonoid transport in plants

Dear Editor, dear editorial team,

I am submitting the manuscript 'Biochemistry and molecular basis of intracellular flavonoid transport' by Boas Pucker and Dirk Selmar for publication in Plants as a review article in the Special Issue 'Evolution of Specialized Metabolism in Plants'.

Different subclasses of flavonoids play numerous important roles in plants. Examples are the coloration of flowers by anthocyanins, the pigmentation of seeds by proanthocyanins, and the protection against the consequences of UV radiation by flavonols. While the biosynthesis pathway is well understood due to investigations in many different plant species, the knowledge about the intracellular transport of flavonoids is sparse. In our review, we summarize the current state of knowledge and point out some questions for future studies.

We hope that you will find our manuscript suitable for publication and hereby confirm that it has not been submitted for publication elsewhere. However, there is a preprint available on Preprints.org (https://www.preprints.org/manuscript/202203.0124/v1) that already received some attention.

I have been in contact with [redacted] about this submission. She confirmed that no publication fees should be charged for the publication, because of my support of the journal as a guest editor.

Sincerely (on behalf of both authors),

[redacted]

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?

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?

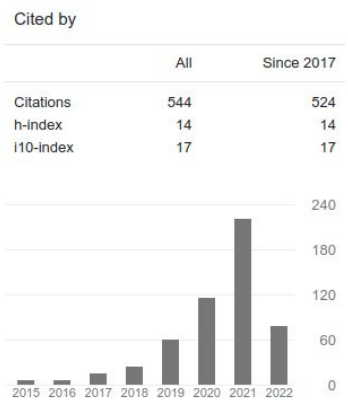
i10-index

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

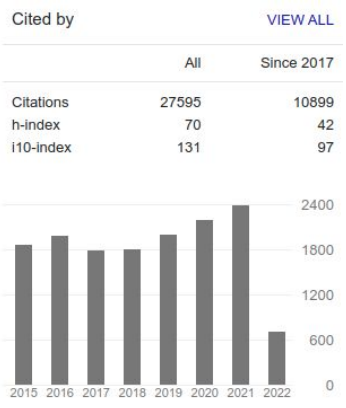
Citations are career stage and field-dependent

Doctoral student

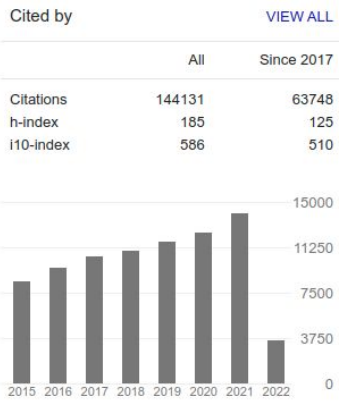
Young PI



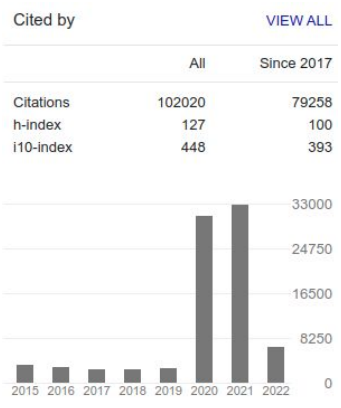
Highly cited PI



Pioneer and geneious



Researcher with ‘hot topic’

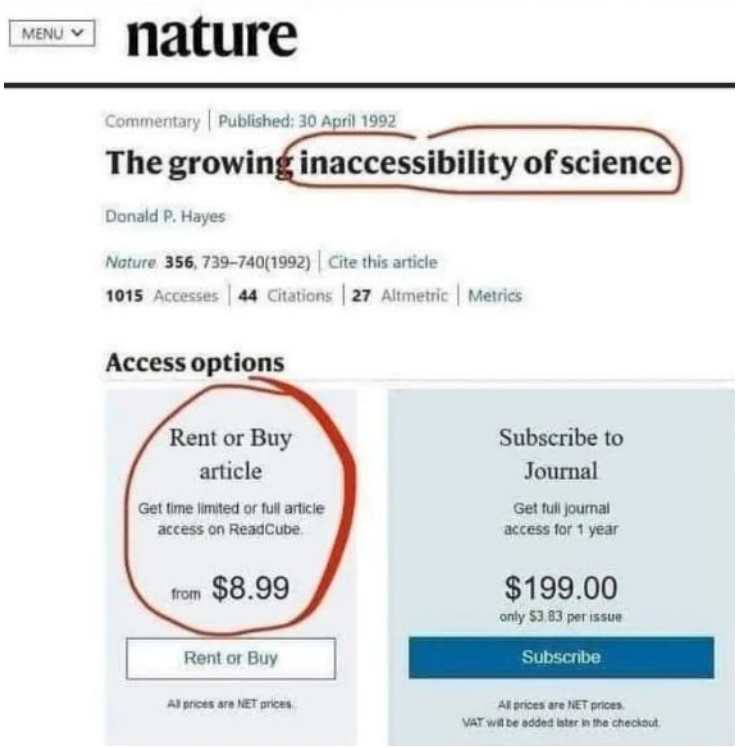


Journal impact factor (IF)

- 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)
- Solid journals: 3-8 (technically solid)

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

(Molecular) Plant Science Journals

- Nature/Science
- Nature Genetics / Nature Plants / Nature Communication
- New Phytologist
- Journal of Experimental Botany

- BMC Genome Biology / Genomics / Plant Biology / Plant Methods
- PLOS ONE / Genetics / Biology / Bioinformatics
- MDPI Plants / Genes / Agriculture / International Journal of Molecular Sciences
- Frontiers in Plant Sciences / Genetics / Molecular Sciences
- PeerJ
- Molecular Biology and Evolution, Genome Biology and Evolution
- Nucleic Acid Research
- Bioinformatics
- GigaScience
- The Plant Journal / Plant Cell

- …

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

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

- 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

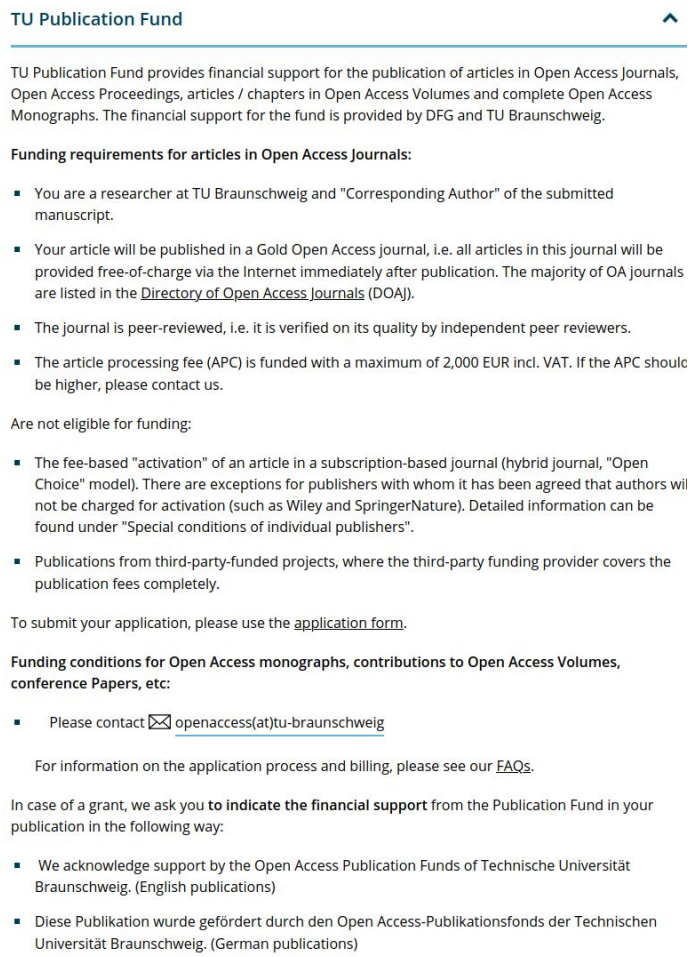
- Summary of the submitted manuscript
- Evaluation of novelty and overall quality
- Major issues
  - Constructive comments about invalid methods
  - List missing controls
  - Point out cases of overstated conclusions
  - …
- Minor issues (include line numbers)
  - Typos
  - Language issues

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

Article Processing Charges

- DEAL: funding for all scholars at German research institutes for Springer Nature journals (except NatureComms + ScientificReports)
- OpenAccess Publication Funds of TU Braunschweig (only supports <2k € fees)
- Statement required: ‘We acknowledge support by the Open Access Publication Funds of Technische Universität Braunschweig. ’
- Funding for publications through DFG and other third party funded projects



TU Publication Fund

TU Publication Fund provides financial support for the publication of articles in Open Access Journals, Open Access Proceedings, articles / chapters in Open Access Volumes and complete Open Access Monographs. The financial support for the fund is provided by DFG and TU Braunschweig.

Funding requirements for articles in Open Access Journals:

- You are a researcher at TU Braunschweig and "Corresponding Author" of the submitted manuscript.
- Your article will be published in a Gold Open Access journal, i.e. all articles in this journal will be provided free-of-charge via the Internet immediately after publication. The majority of OA journals are listed in the Directory of Open Access Journals (DOAJ).
- The journal is peer-reviewed, i.e. it is verified on its quality by independent peer reviewers.
- The article processing fee (APC) is funded with a maximum of 2,000 EUR incl. VAT. If the APC should be higher, please contact us.

Are not eligible for funding:

- The fee-based "activation" of an article in a subscription-based journal (hybrid journal, "Open Choice" model). There are exceptions for publishers with whom it has been agreed that authors will not be charged for activation (such as Wiley and SpringerNature). Detailed information can be found under "Special conditions of individual publishers".
- Publications from third-party-funded projects, where the third-party funding provider covers the publication fees completely.

To submit your application, please use the application form.

Funding conditions for Open Access monographs, contributions to Open Access Volumes, conference Papers, etc:

- Please contact ✉ openaccess(at)tu-braunschweig

For information on the application process and billing, please see our FAQs.

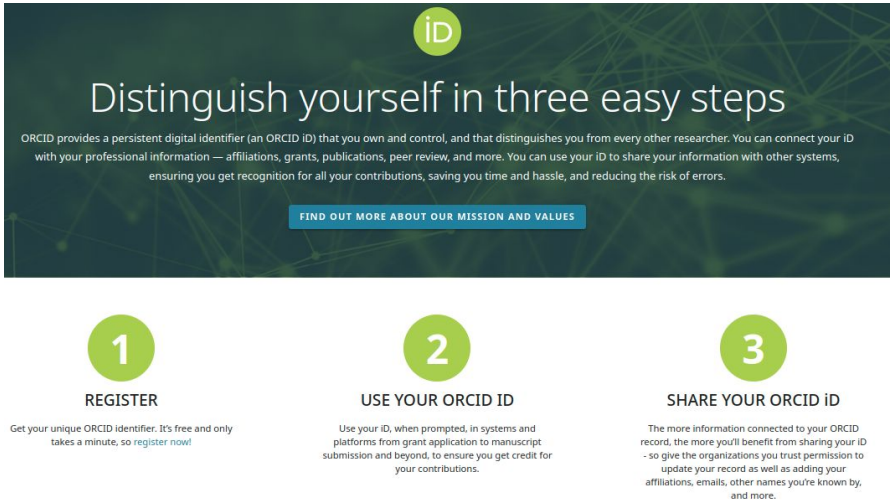
In case of a grant, we ask you **to indicate the financial support** from the Publication Fund in your publication in the following way:

- We acknowledge support by the Open Access Publication Funds of Technische Universität Braunschweig. (English publications)
- Diese Publikation wurde gefördert durch den Open Access-Publikationsfonds der Technischen Universität Braunschweig. (German publications)

https://www.tu-braunschweig.de/en/ub/publishing-open-access/open-access-publishing

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/

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

Flawed papers

- Typos e.g. comma misplaced
- Honest mistakes in analyses
- Scientifically wrong conclusions
- Fraud e.g. fabricated data sets or manipulated images

Retracted article
See the retraction notice

> BMC Plant Biol. 2020 Jul 31;20(1):361. doi: 10.1186/s12870-020-02566-2.

Contribution of anthocyanin pathways to fruit flesh coloration in pitayas

Ruiyi Fan [1], Qingming Sun [1], Jiwu Zeng [1], Xinxin Zhang [2]

Affiliations + expand

PMID: 32736527 PMCID: PMC7394676 DOI: 10.1186/s12870-020-02566-2

Free PMC article

Retraction in

Retraction Note: Contribution of anthocyanin pathways to fruit flesh coloration in pitayas.

Fan R, Sun Q, Zeng J, Zhang X.

BMC Plant Biol. 2021 May 20;21(1):225. doi: 10.1186/s12870-021-03005-6.

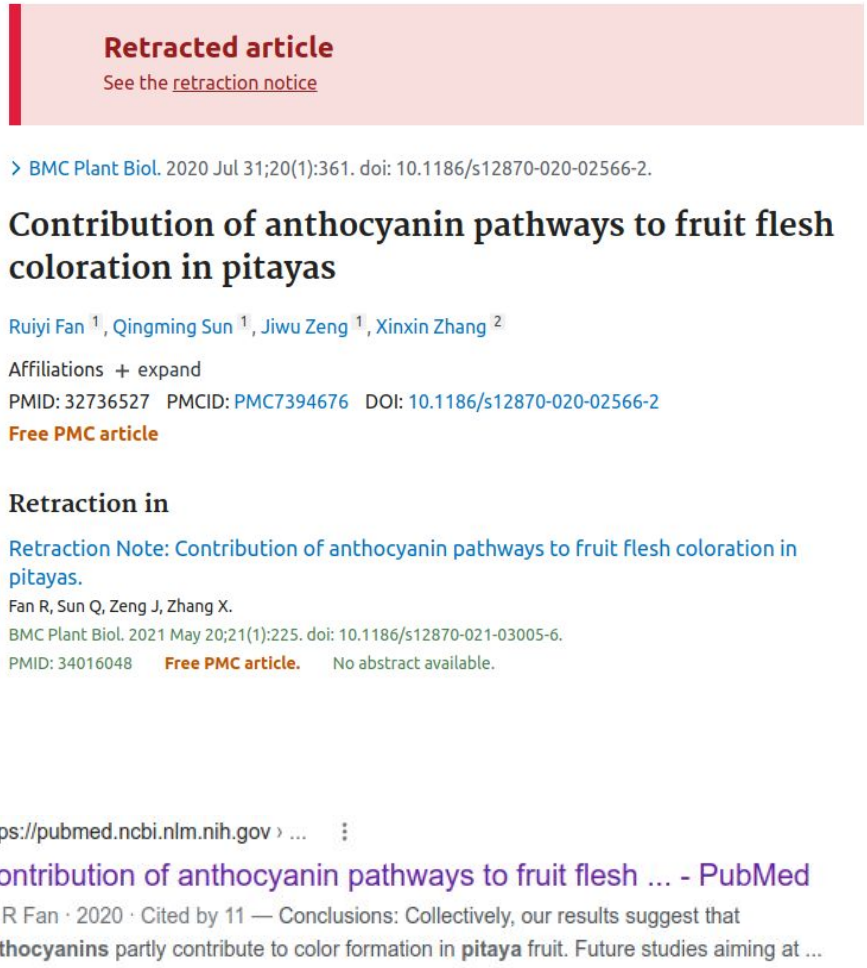
PMID: 34016048 **Free PMC article.** No abstract available.

Correction

- Minor mistakes in articles can be solved by publishing a correction
- Mistakes must not affect any of the major conclusions of the article
- Examples: typos in numbers or mislabeled figures

Retraction

- Retraction is the removal of an article from the body of (valid) literature
- Retraction is done if a correction is not possible
- Retracted articles should not be cited (tools like Zotero can warn you)
- Errors can be propagated through the literature



Retracted article
See the retraction notice

› BMC Plant Biol. 2020 Jul 31;20(1):361. doi: 10.1186/s12870-020-02566-2.

Contribution of anthocyanin pathways to fruit flesh coloration in pitayas

Ruiyi Fan [1], Qingming Sun [1], Jiwu Zeng [1], Xinxin Zhang [2]

Affiliations + expand

PMID: 32736527 PMCID: PMC7394676 DOI: 10.1186/s12870-020-02566-2

Free PMC article

Retraction in

Retraction Note: Contribution of anthocyanin pathways to fruit flesh coloration in pitayas.

Fan R, Sun Q, Zeng J, Zhang X.

BMC Plant Biol. 2021 May 20;21(1):225. doi: 10.1186/s12870-021-03005-6.

PMID: 34016048 **Free PMC article.** No abstract available.

https://pubmed.ncbi.nlm.nih.gov › ...

Contribution of anthocyanin pathways to fruit flesh ... - PubMed

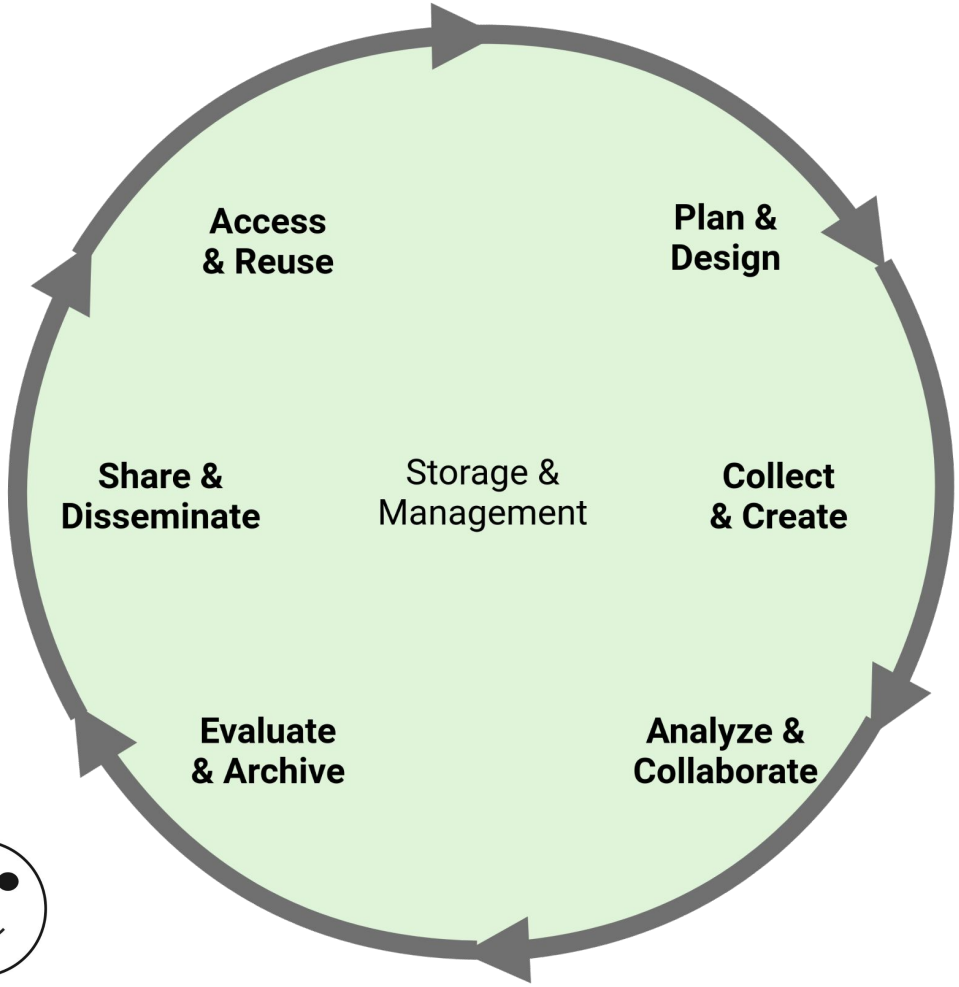
by R Fan · 2020 · Cited by 11 — Conclusions: Collectively, our results suggest that **anthocyanins** partly contribute to color formation in **pitaya** fruit. Future studies aiming at ...

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
  - Blogs & news outlets
  - Reddit
  - CitationTools

Data life cycle

Data life cycle



PLAN

https://commons.wikimedia.org/wiki/File:RDM_02_Collect-and-Capture_frame01.svg
https://www.picpedia.org/keyboard/a/analyze.html

Plan & design

- What are the research questions/objectives?
- Plan analysis/experiment that will generate data OR use existing data sets
- How many replicates? Which statistical tests (power)?
- Data management plan
- How to ensure data safety (backups)?

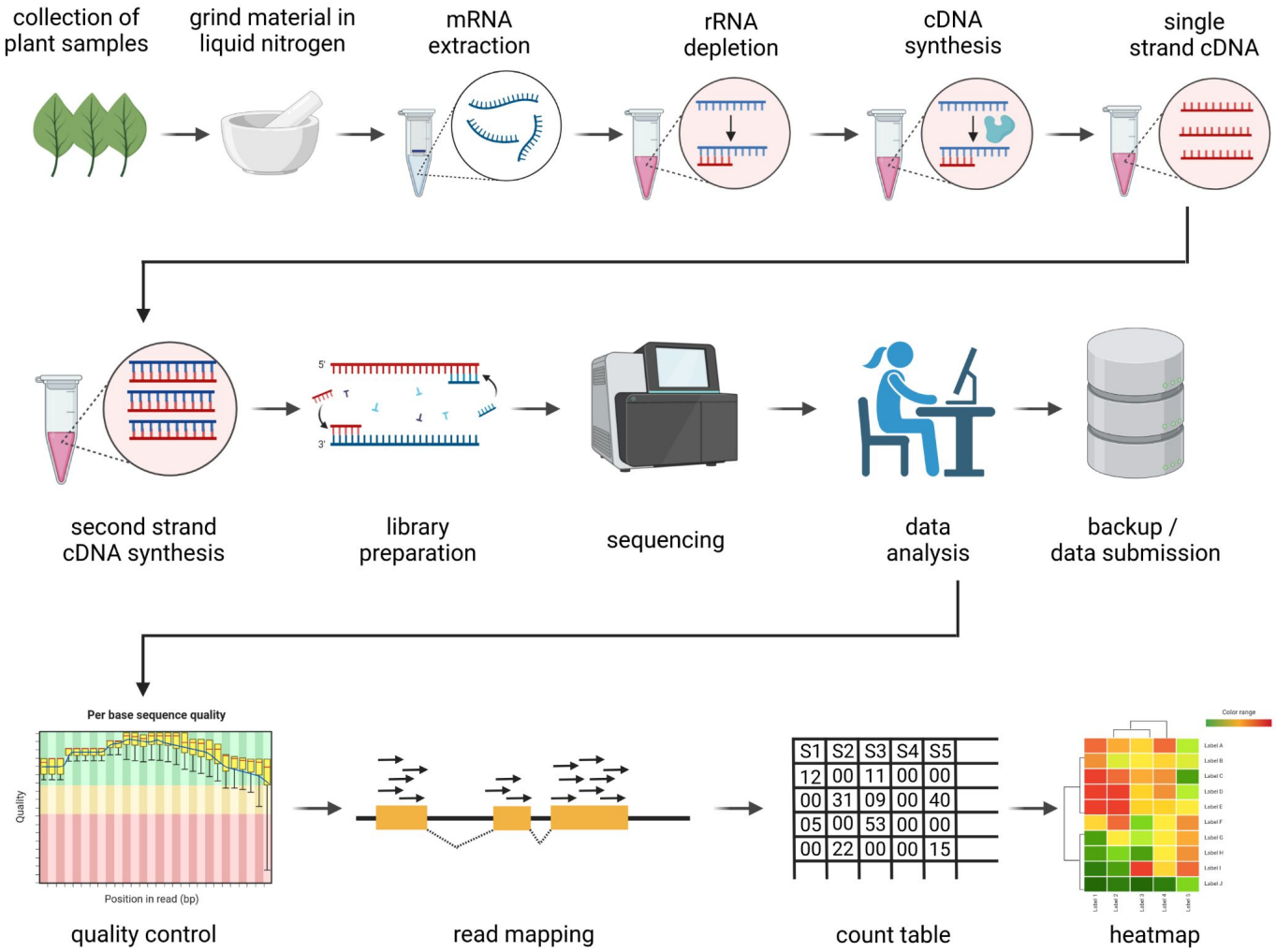
EXAMPLE: Plan & design

- Research question: What are the molecular mechanisms explaining white and red pigmentation of flowers?

- Experiment:
  - RNA-Seq
  - >=3 replicates per morphotype
  - paired-end sequencing
  - 30 million tags

- Data storage:
  - on a local hard drive
  - backup in cloud
  - cold storage

RNA-Seq


collection of plant samples
grind material in liquid nitrogen
mRNA extraction
rRNA depletion
cDNA synthesis
single strand cDNA
second strand cDNA synthesis
library preparation
sequencing
data analysis
backup / data submission
Per base sequence quality
Quality
Position in read (bp)
quality control
read mapping
S1 S2 S3 S4 S5
12 00 11 00 00
00 31 09 00 40
05 00 53 00 00
00 22 00 00 15
count table
heatmap

Data management plan

- Tools are available to generate data management plans
- Funder-specific differences in expectations
- Content:
  - Information about data and data format
    - method/time of collection; version control; backup
  - Metadata content and format
  - Policies for access, sharing, and reuse
  - Long-term storage
  - Budget: considerable costs might arise

Collect & create

- Perform planned experiment with replicates
- Document all steps
- Collect the results (data)
- Synchronize results across all storage/backup locations

EXAMPLE: Collect & create

- Grow plants under precisely controlled conditions
- Harvest material in a reproducible manner; replicates are important!
- Extract RNA and subject to RNA-seq experiment
- Store resulting FASTQ files

Synchronize & backup

- Synchronization of data between different locations:
  - $ rsync <SOURCE> <TARGET>

- Backup solutions:
  - Desktop computers & laptops
  - External hard drives
  - Network drives
  - Central storage options (e.g. cold storage; tape storage)
  - Cloud storage
  - Optical storage

- 3 copies of data are recommended; one off-site

EXAMPLE: Synchronize & backup

- Sequencing data will be stored on local hard drive
- Copy will be uploaded to de.NBI cloud for analysis
- Backup of sequencing data in secure place after run completion

Analyze & collaborate

- Quality control
- Plausibility checks
- Sample identity checks
- Perform actual analysis alone/in collaboration
- Interdependence of collaborators possible

EXAMPLE: Analyze & collaborate

- Quality control via principal component analysis (PCA)
- Differential gene expression analysis (DESeq2)
- Pathway enrichment analysis (KEGG, GO)

- How to exchange files with collaborators
  - cloud
  - hard drives

- How to document collaboration
  - Documentation of meetings
  - Documentation of contributions to project
  - Documentation of progress
  - Writing reports for funding agencies

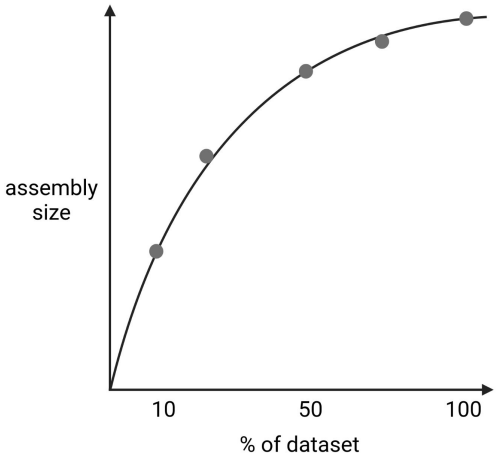
https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html

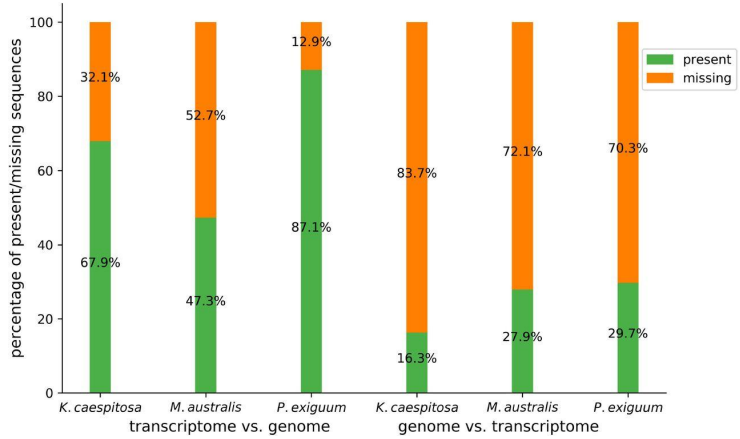
Evaluate & archive

- Are the data sets large enough?
- What could have been done better?
- Was the number of replicates sufficient?
- Are there clear differences between sample groups?
- Low variation between biological replicates?
- Archive all research data on tape storage for at least 10 years
- Off-site backups; Rsync to transfer only modified files
- Commercial clouds (Dryad, GigaDB); tape storage @ TUBS (contact GITZ)

Enough data?

- Check if subsets lead to complete assemblies (saturation)
- Comparison of genome and transcriptome assembly
- Genome sequence assemblies cover lowly expressed genes
- Transcriptome assemblies can be advantageous for genes with large introns

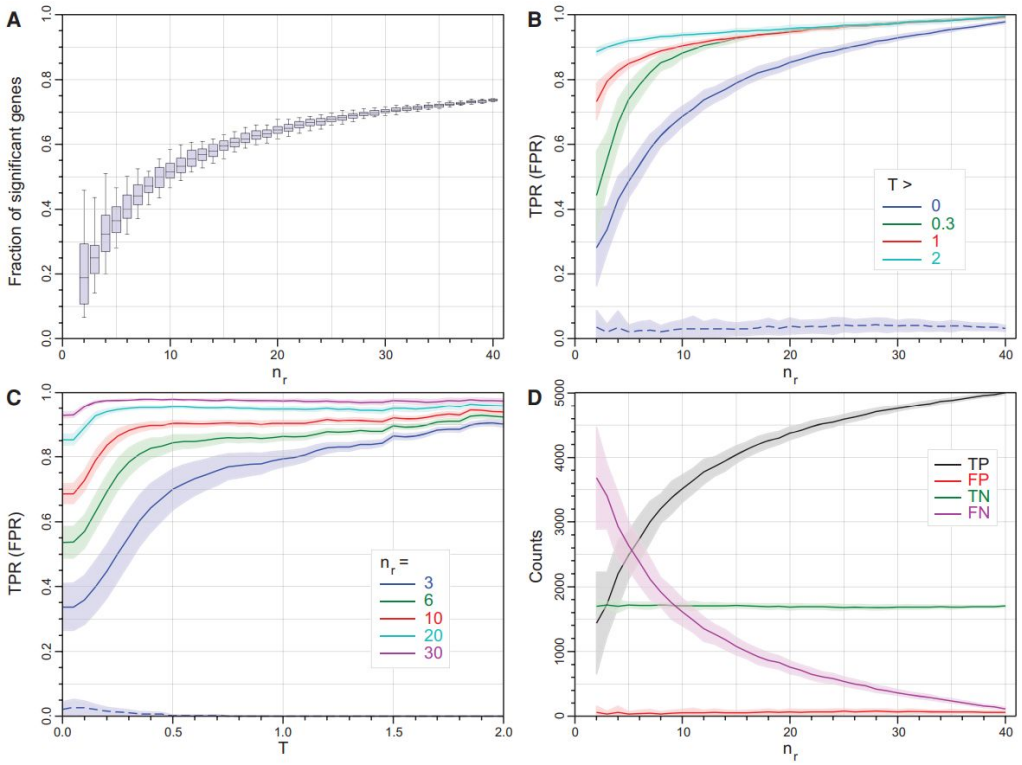




Haak et al., 2018: 10.3389/fmolb.2018.00062
Pucker et al., 2019: 10.1101/646133

Sufficient replicates?

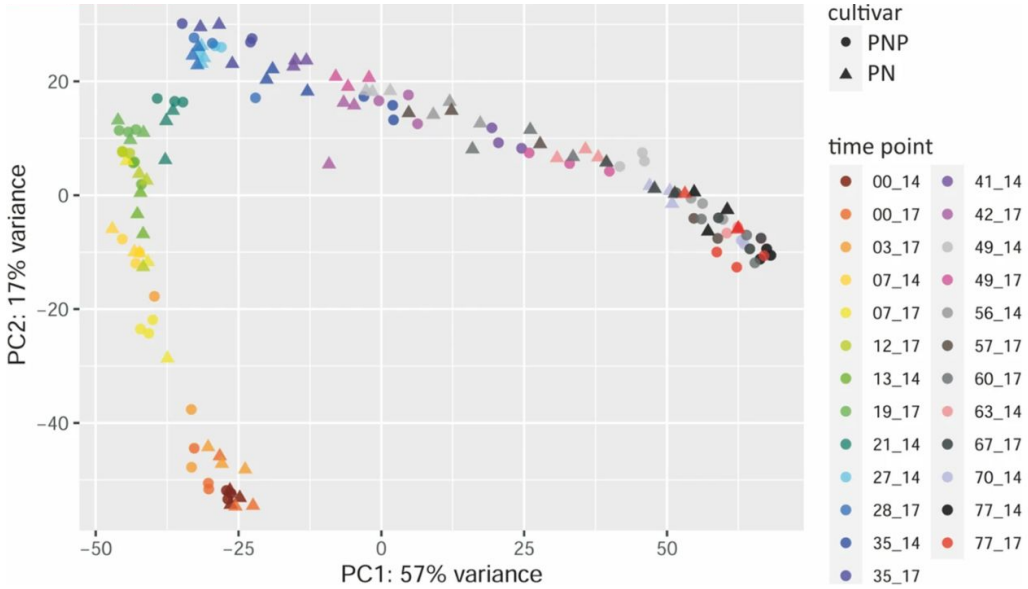
- More replicates allow the identification of a larger number of DEGs
- Reliability of gene classification increases with number of replicates
- Up to 40 replicates can boost signal strength



Schurch et al., 2016: 10.1261/rna.053959.115

Systematic differences between groups?

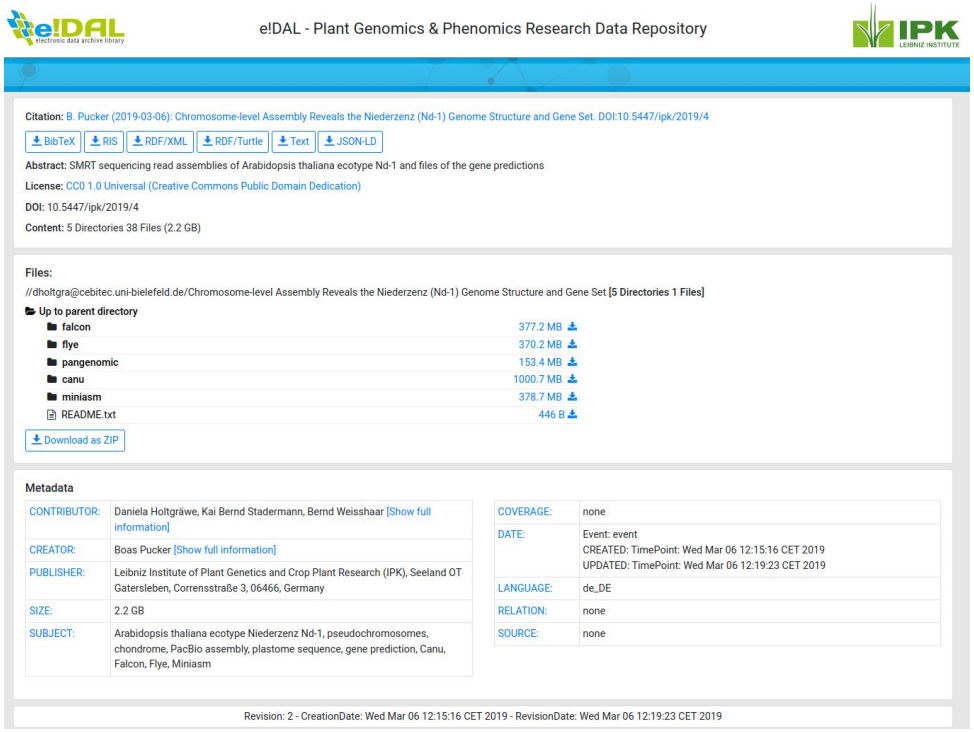
- Principal Component Analysis (PCA) separates RNA-seq samples based on gene expression patterns
- Similar samples are grouped together
- Principal Components (axes) are artificial axes



Theine et al., 2021: 10.1186/s12870-021-03110-6

EXAMPLE: archive

- Submission of Nd-1 data sets to e!DAL
- Citable in corresponding publication via DOI
- Data are also archived on tape storage

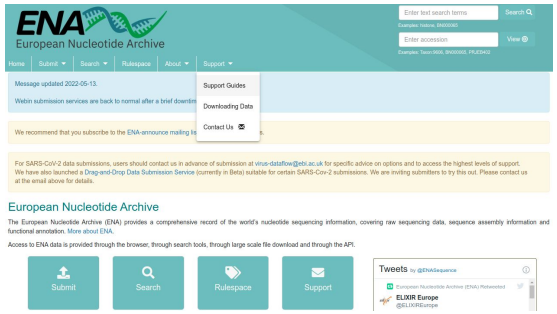


Share & disseminate

- All data sets underlying a publication should be released
- Clearly refer to published data sets in data availability statement
- Submission of large data sets to suitable databases
- Select proper license for data sets
- Include all ‘customized scripts’ in publication

EXAMPLE: Share & disseminate

- Options to submit sequencing data:
  - European Nucleotide Archive (ENA)
  - Sequence Read Archive (SRA)
  - Gene Expression Omnibus (GEO)



- Submission of individual sequences: NCBI

- Options to share data sets (github, dryad, GigaDB, PUB, TUBS?)

- Scripts can be shared via github, bitbucket, gitlab, codeberg

EXAMPLE: data availability statement

- Sequence read datasets generated and analyzed during this study were made available at ENA under the accession PRJEB35658. Individual run IDs are included in Additional file 1. The Col-0 genome sequence assembly of the GABI-Kat Col-0 genetic background (Col-0_GKat-wt) is available at ENA under the accession GCA_905067165.

- **Availability of supporting code and requirements**
  - Project name: KIPEs3
  - Project home page: https://github.com/bpucker/KIPEs
  - Operating system(s): Linux (website is platform independent)
  - Programming language: Python3
  - Other requirements: BLAST, MAFFT, FastTree2, dendropy, scipy
  - License: GNU General Public License v3.0
  - RRID: SCR_022370

- All data sets analyzed in this study are publicly available. Data sets generated as part of this study are shared via GitHub (https://github.com/bpucker/KIPEs). A docker image is available via DockerHub (https://hub.docker.com/r/bpucker/kipes).

Pucker et al., 2021: 10.1186/s12864-021-07877-8
Rempel & Pucker, 2022: 10.1101/2022.06.30.498365

EXAMPLE: data publication

PUB - Publications at Bielefeld University

Arabidopsis thaliana methylation pattern analysis based on ONT sequence reads

Schilbert H, Kleinbölting N, Weisshaar B, Pucker B (2021)
Bielefeld University.

Research Data

Download
GK_040A12.vcf.gz 176.58 MB
GK_050B11.vcf.gz 183.65 MB
GK_082G09.vcf.gz 185.46 MB
+ All

DOI https://doi.org/10.4119/unibi/2956654

Details | Files | Links

Creator: Schilbert, Hanna[UniBi]; Kleinbölting, Nils[UniBi]; Weisshaar, Bernd[UniBi]; Pucker, Boas[UniBi]

Department: Fakultät für Biologie > Genetik und Genomik der Pflanzen
Centrum für Biotechnologie > Research Group B. Weisshaar

Abstract / Notes:

We sequenced the genomes of 14 Arabidopsis thaliana GABI-Kat T-DNA insertion lines (Col-0 background), which eluded flanking sequence tag-based attempts to fully characterize their insertion alleles, with Oxford Nanopore Technologies (ONT) long reads. Detailed information about each line, e.g. their T-DNA insertion site(s), have been described by Pucker et al. (BMC Genomics (2021) 22:599; https://doi.org/10.1186/s12864-021-07877-8). The DNA that has been sequenced is derived from a mixtures of tissues as expected for DNA extracted from young plantlets. For 11 of the datasets from these lines, we called 5mC methylation patterns with the tool Megalodon v.2.2.9 provided by ONT (https://github.com/nanoporetech/megalodon/). Results of this analysis are available per line as individual VCF files. We are sharing unfiltered output files to grant full control over the down-stream analysis steps and to accelerate the research in epigenomics. We identified between 706,254 and 1,020,654 positions per line where at least 80% of 5 or more reads support a methylation site. The following datasets are available and the file names reveal the respective GK line:

1. GK_040A12.vcf
2. GK_050B11.vcf
3. GK_082G09.vcf
4. GK_290G05.vcf
5. GK_399C06.vcf
6. GK_410B07.vcf
7. GK_430F05.vcf
8. GK_433E06.vcf
9. GK_654A12.vcf
10. GK_767D12.vcf
11. GK_909H04.vcf

Keywords: long read sequencing; ONT; methylation; plant epigenomics; GABI-Kat; Megalodon

Year of Publication: 2021

Copyright and Licenses: Creative Commons Attribution 4.0 International Public License (CC-BY 4.0)

Page URI: https://pub.uni-bielefeld.de/record/2956654

Cite this

AMA | APA (6th ed.) | Frontiers | Harvard | IEEE | LNCS | MLA

Schilbert H, Kleinbölting N, Weisshaar B, Pucker B. *Arabidopsis thaliana methylation pattern analysis based on ONT sequence reads*. Bielefeld University; 2021.

PUB - Publications at Bielefeld University

Arabidopsis thaliana methylation pattern analysis based on ONT sequence reads

Schilbert H, Kleinbölting N, Weisshaar B, Pucker B (2021)
Bielefeld University.

Research Data

Download
GK_040A12.vcf.gz 176.58 MB
GK_050B11.vcf.gz 183.65 MB
GK_082G09.vcf.gz 185.46 MB
+ All

DOI https://doi.org/10.4119/unibi/2956654

Details | Files | Links

All files available under the following license(s):

Creative Commons Attribution 4.0 International Public License (CC-BY 4.0):
https://creativecommons.org/licenses/by/4.0/
https://creativecommons.org/licenses/by/4.0/legalcode

Main File(s)

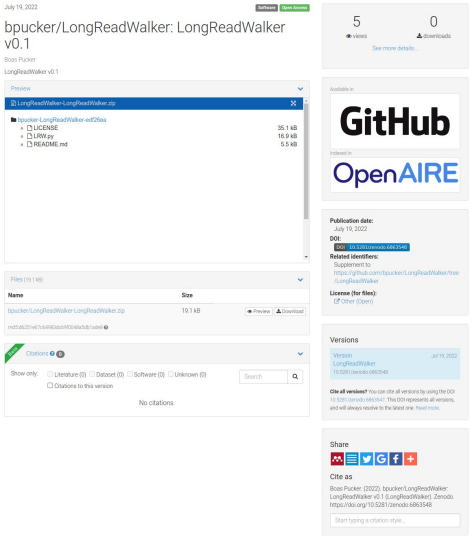
File Name	GK_040A12.vcf.gz 176.58 MB
Access Level	Open Access
Last Uploaded	2021-08-09T16:28:06Z
MD5 Checksum	5639514d2396b649688a1c08c9bb7ca7
File Name	GK_050B11.vcf.gz 183.65 MB
Access Level	Open Access
Last Uploaded	2021-08-09T16:28:07Z
MD5 Checksum	8ad8a957b3a7c48fa3f8bcc2eaff4178
File Name	GK_082G09.vcf.gz 185.46 MB
Access Level	Open Access
Last Uploaded	2021-08-09T16:28:09Z
MD5 Checksum	ae122833858f8e826f51d465c377443c
File Name	GK_290G05.vcf.gz 199.73 MB
Access Level	Open Access
Last Uploaded	2021-08-09T16:28:09Z
MD5 Checksum	df135692cab92ae1d4e1a0ff852c0691
File Name	GK_399C06.vcf.gz 181.61 MB
Access Level	Open Access
Last Uploaded	2021-08-09T16:28:12Z
MD5 Checksum	59202cb9168b9c5a9b5d72282ee17a66
File Name	GK_410B07.vcf.gz 200.89 MB
Access Level	Open Access
Last Uploaded	2021-08-09T16:28:13Z
MD5 Checksum	608d96b3879e79170b51cab0d78c01d9
File Name	GK_430F05.vcf.gz 197.97 MB
Access Level	Open Access
Last Uploaded	2021-08-09T16:28:15Z
MD5 Checksum	eb9e3d4f957f33fb623d9e6d44cb00f3
File Name	GK_433E06.vcf.gz 192.92 MB
Access Level	Open Access
Last Uploaded	2021-08-09T16:28:16Z
MD5 Checksum	4ecfa41f5aa385e5c8b851f700e1dc1f
File Name	GK_654A12.vcf.gz 187.28 MB
Access Level	Open Access
Last Uploaded	2021-08-09T16:28:19Z
MD5 Checksum	29822d0433e5178e0110573213d76886

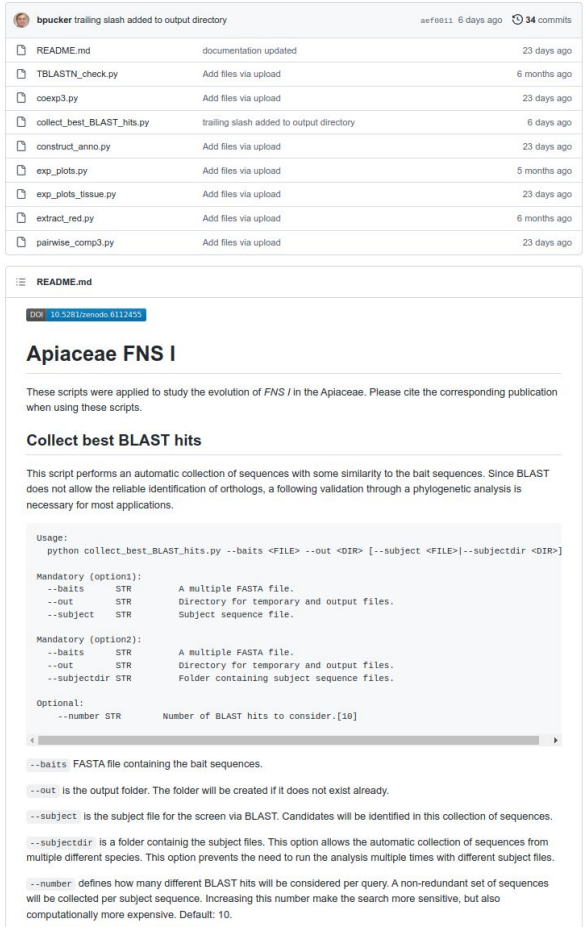
Schilbert et al., 2021: 10.4119/unibi/2956654

EXAMPLE: publication of scripts

- Scripts should be shared through repositories:
  - Github
  - Bitbucket
  - Codeberg
  - Gitlab

- Archive repositories in Zenodo (DOI assignment)





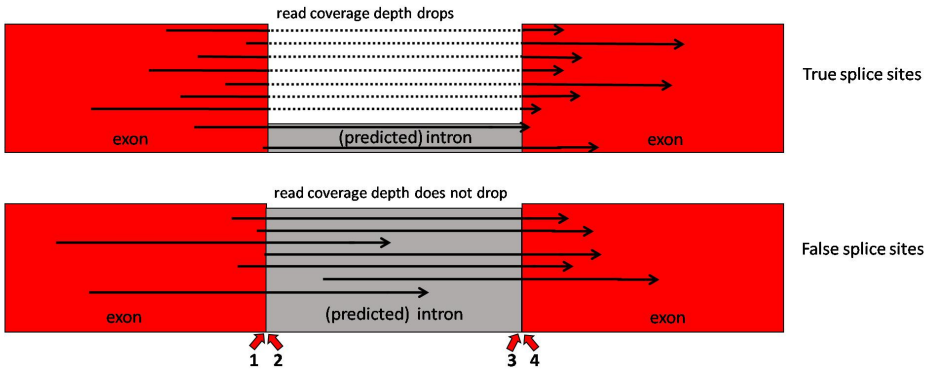
https://github.com/bpucker/ApiaceaeFNS1

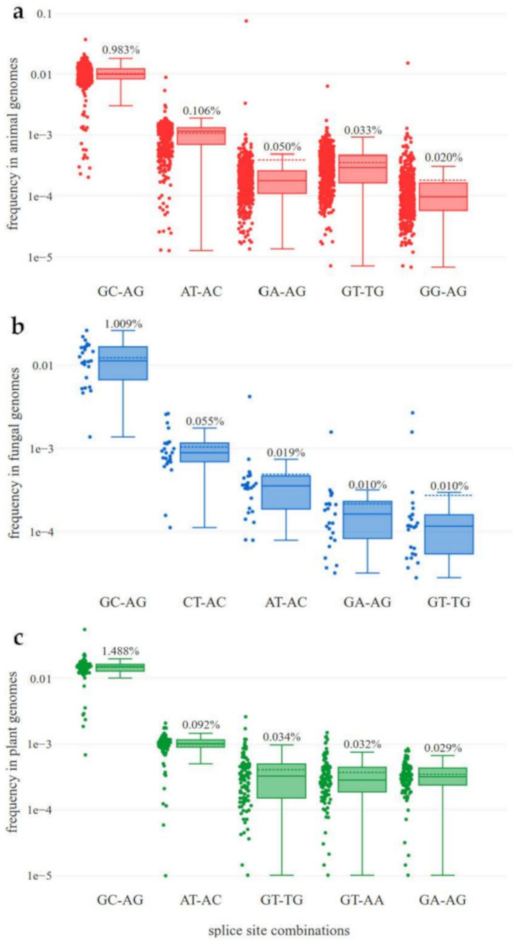
Access & reuse

- Data sets should be freely available to enable validation of findings
  - NOT: ‘Available upon reasonable request from the corresponding author’
- Public data sets are an excellent resource for re-use
- No costs for data generation
- Massive datasets can enable identification of small effect sizes
- Generation of novel hypothesis

EXAMPLE: access & reuse

- Investigation of non-canonical splice sites in plants, animals, and fungi
- Screening genome sequences+annotations for splice sites
- Harnessing RNA-seq to quantify usage of (non-canonical) splice sites

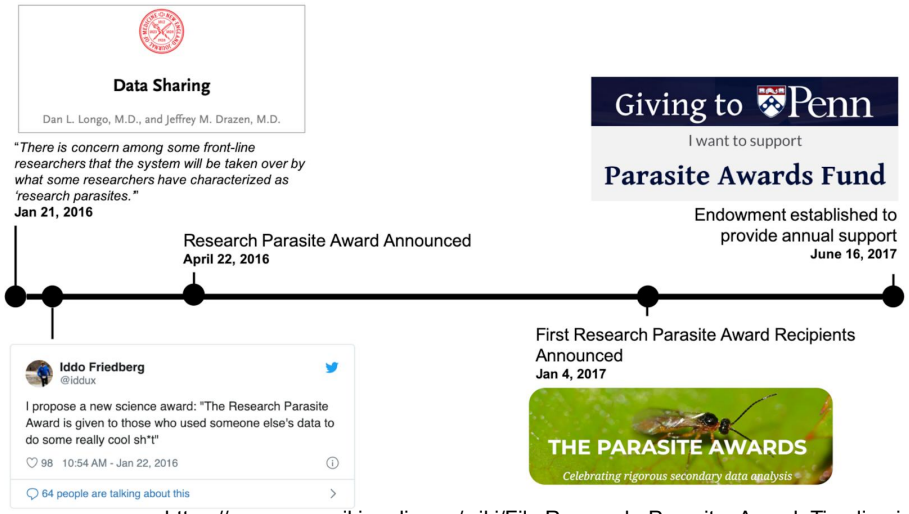




Pucker & Brockington, 2018: 10.1186/s12864-018-5360-z
Frey & Pucker, 2020: 10.3390/cells9020458

The Parasite Awards

- Celebrates comprehensive secondary data analysis
- Novel insights inferred from existing (underutilized) data sets
- Supported by GigaScience/GigaByte



https://commons.wikimedia.org/wiki/File:Research_Parasite_Award_Timeline.jpg

https://researchparasite.com/

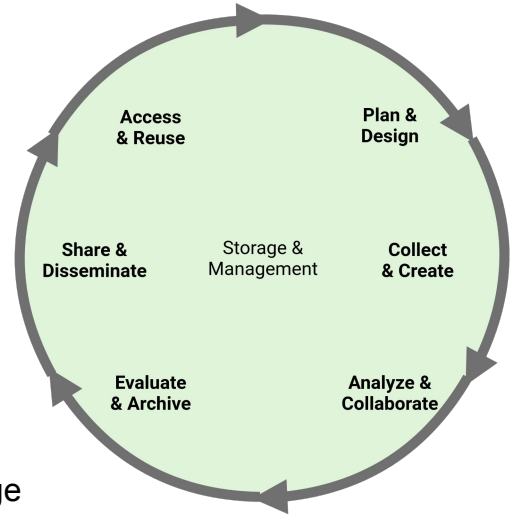
Full cycle: Nd-1

Data sets reused for variant caller benchmarking study

Publication about findings

data sets published at SRA

NGS not sufficient for complete genome sequence; on tape storage



Sequence Nd-1 genome, because it is a parent of a mapping population

Sequencing on Roche 454, GAIIx, & HiSeq1500

fastQC analysis, trimming with Trimmomatic, and assembly+annotation

Full cycle: *Croton tiglium*

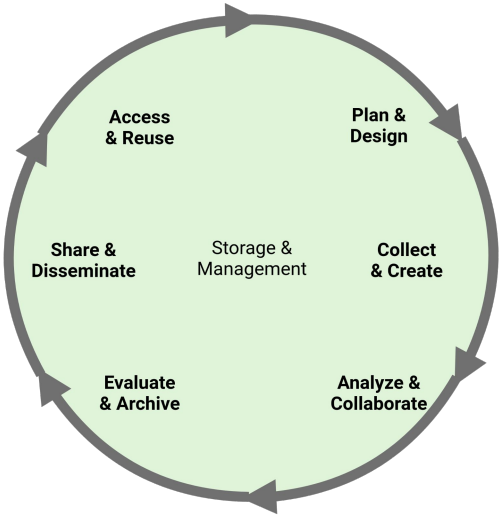
Data reuse for validation of KIPEs

Publication about findings

data sets published at SRA

Evaluation to identify optimal parameters

Data stored via tape storage



Identify specific genes in *Croton tiglium* via transcriptome assembly

RNA-seq with samples of different tissues & normalized library

Transcriptome assembly and characterization of candidate genes

Re-use

What are the advantages and challenges of re-use?

Advantages and challenges of data re-use

Advantage	Challenge
Cost-effective	Lacking metadata
Immediately accessible	Unknown details/issues
Extremely large data sets	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

Workflow

identify RNAseq samples @ SRA

identify genome/ transcriptome assembly

download reads via fastq-dump

download CDS/ transcript sequences

split-read mapping

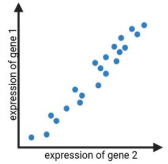
run kallisto to quantify transcript abundances

gene prediction

generation of count tables

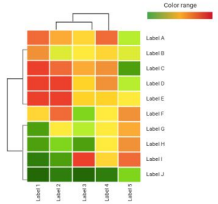
	s1	s2	s3	s4
g1	12	3	4	9
g2	32	3	0	5
g3	17	5	4	1
g4	41	2	0	9
g5	13	3	7	0

co-expression analysis



DESeq2 analysis

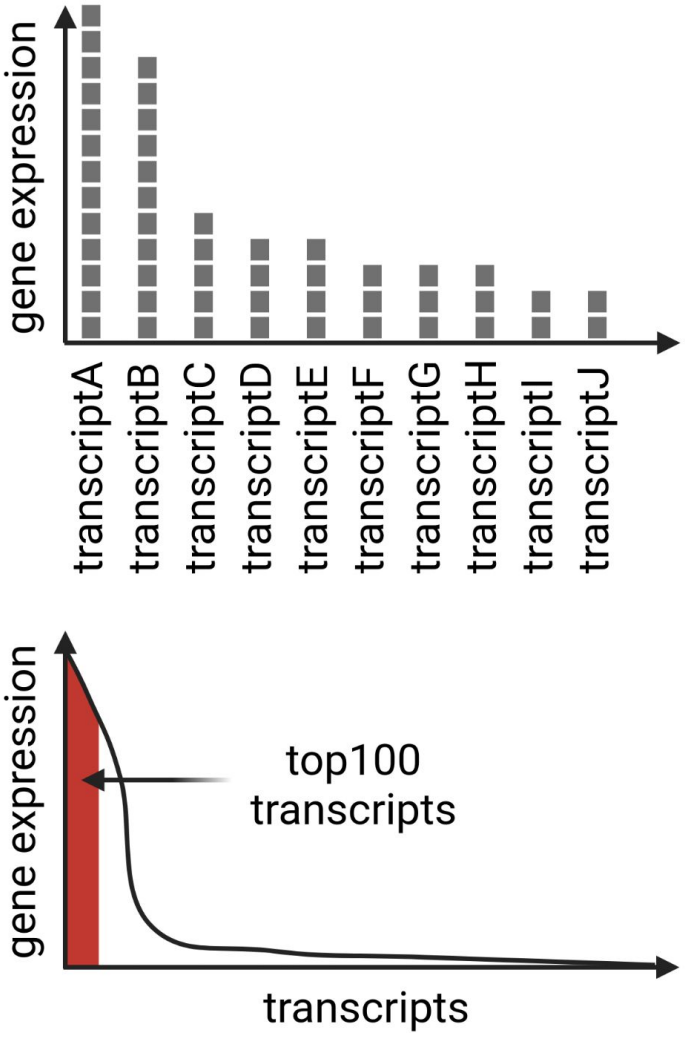
construction of heatmaps



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



Managing references

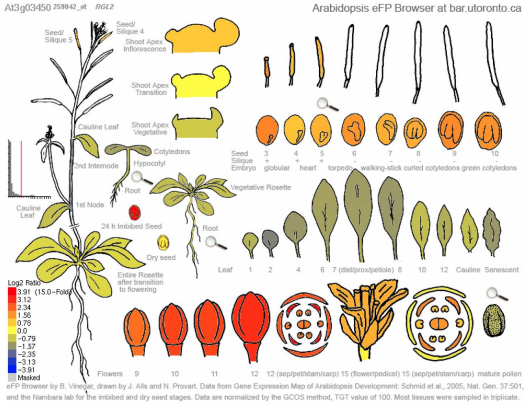
- Entries about publications / data sets are stored in a local database
- Convenient citation when writing manuscripts
- Support for comments and key words assigned to entries
- Examples:
  - Zotero: free
  - Mendeley: free, but belongs to Elsevier
  - Citavi: commercial, but campus license
  - EndNote: commercial

Data re-use examples

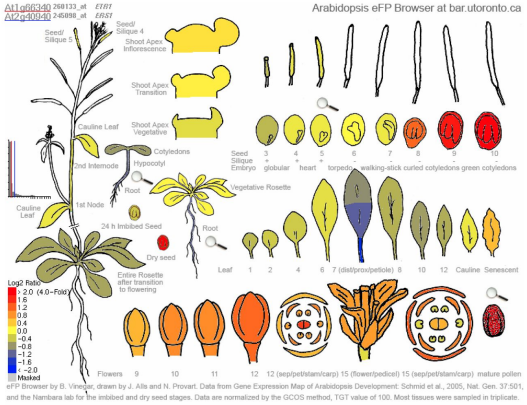
Electronic Fluorescent Pictograph browser

- Integration of existing microarray datasets
- Option to integrate other large scale datasets
- Basis for hypothesis generation
- Visualization of gene expression in many different plant tissues/conditions

A



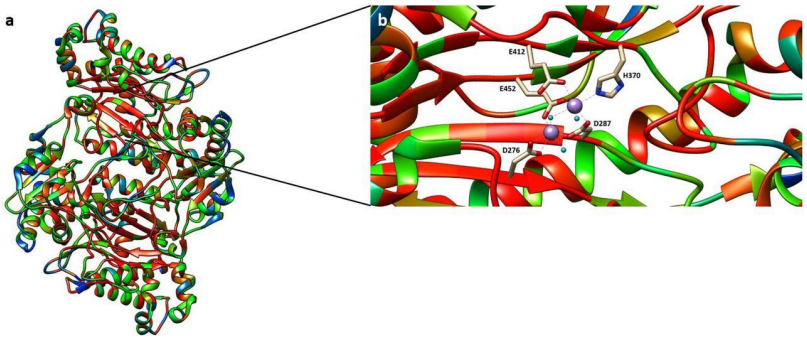
B



Winter et al., 2007: 10.1371/journal.pone.0000718

Identification of conserved amino acid residues

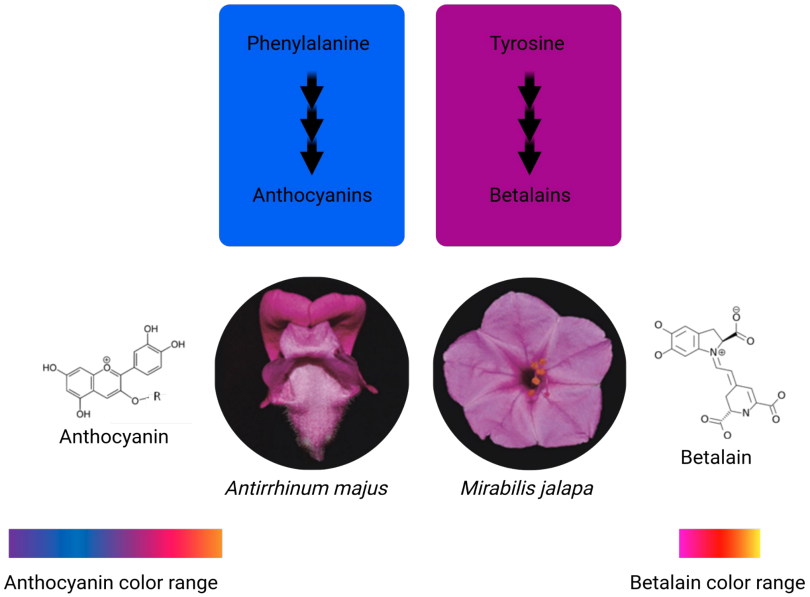
- Identification of orthologous sequences in hundreds of species
- Comparison of sequences to identify highly conserved amino acid residues
- 3D modeling based on known structures of homologs

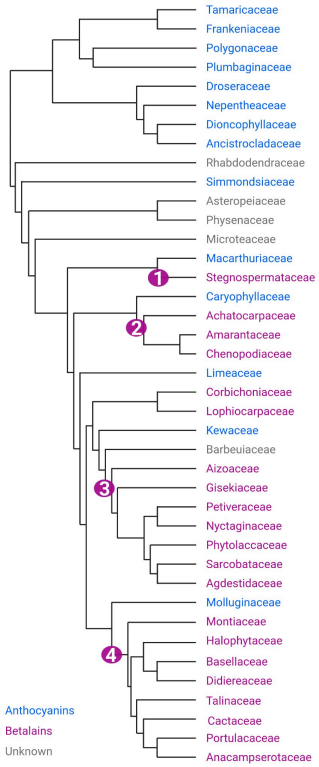


	Animals	Plants	Fungi	Archaea	Bacteria
D276	94	99	100	100	100
D287	94	98	100	100	100
H370	94	98	100	100	100
E412	94	96	100	100	100
E452	91	97	100	100	100
T289	94	97	100	100	97
T410	93	96	100	79	100
H377	94	98	100	100	97
R398	93	98	89	10	57
W107	88	98	96	0	96
Y241	94	96	100	2	90
I244	93	98	97	88	100
H255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

Schilbert et al., 2018: 10.1101/423475

Complex pigment evolution in the Caryophyllales

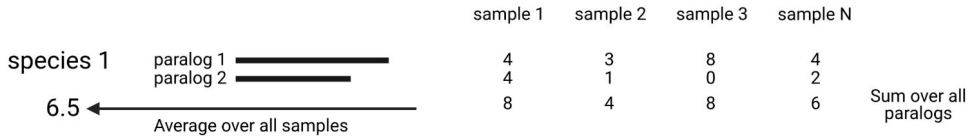


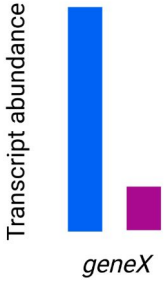


- Functional redundancy of anthocyanins and betalains
- At least four independent origins of betalain biosynthesis
- Mutual exclusion: anthocyanins and betalains were never observed in same (natural) plants

Timoneda *et al.*, 2019: 10.1111/nph.15980
Sheehan *et al.*, 2020: 10.1111/nph.16089

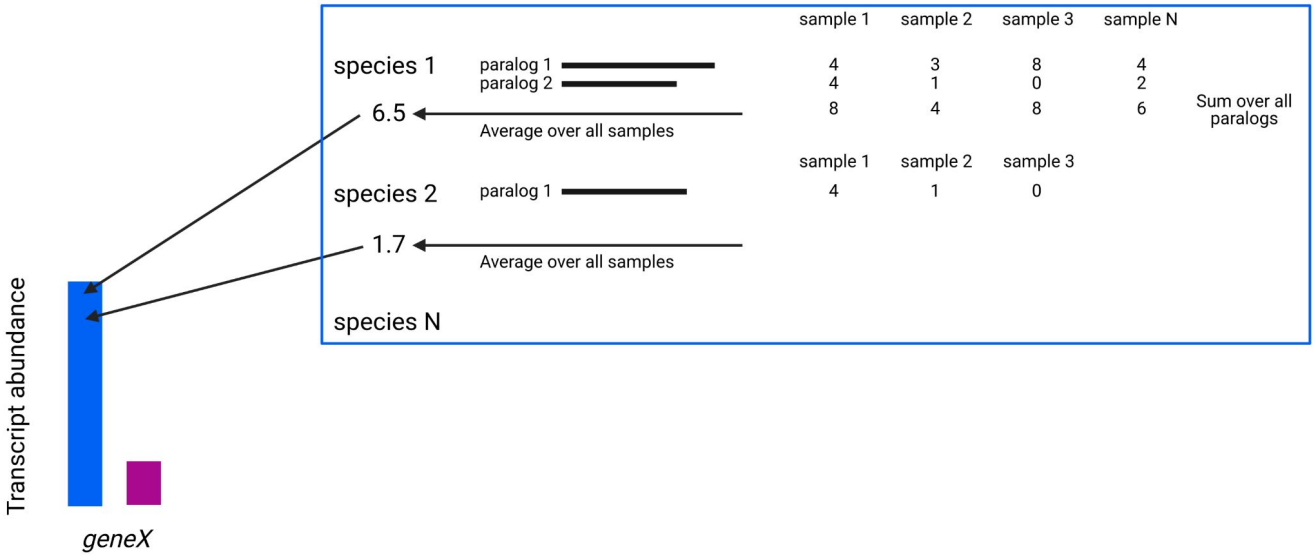
Cross species transcriptomics





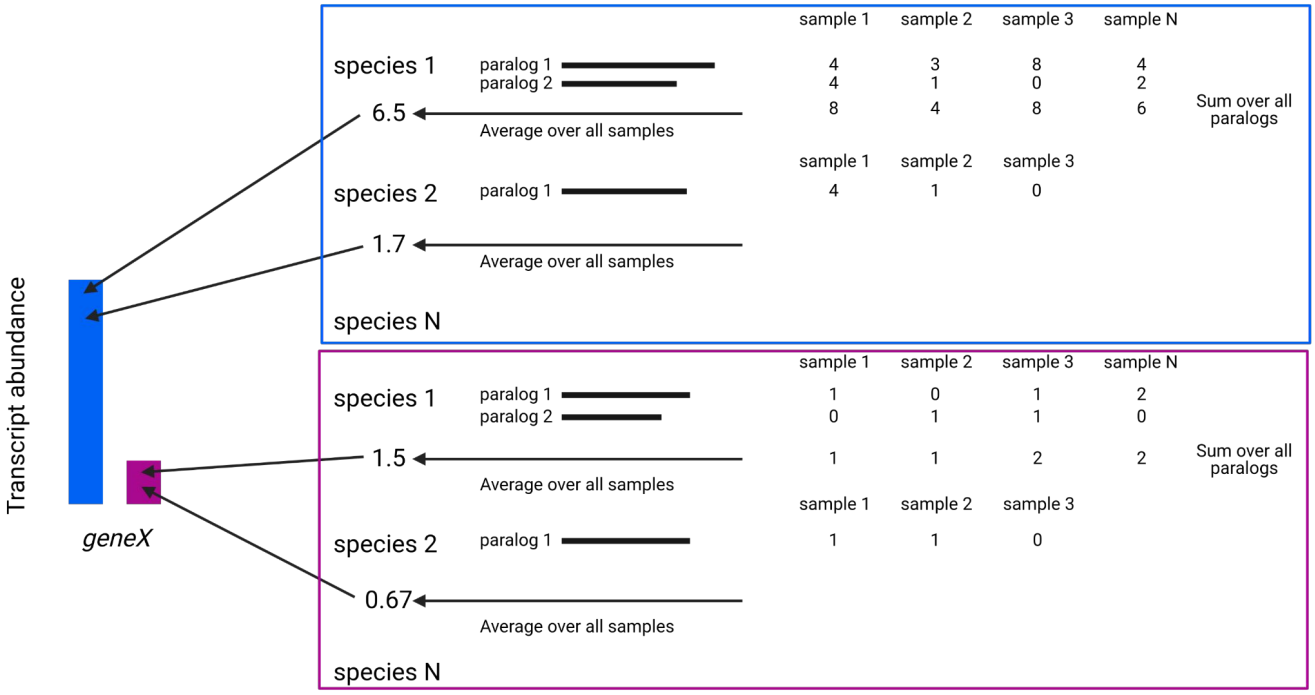
Pucker *et al.*, 2022: 10.1101/2022.10.19.512958

Cross species transcriptomics



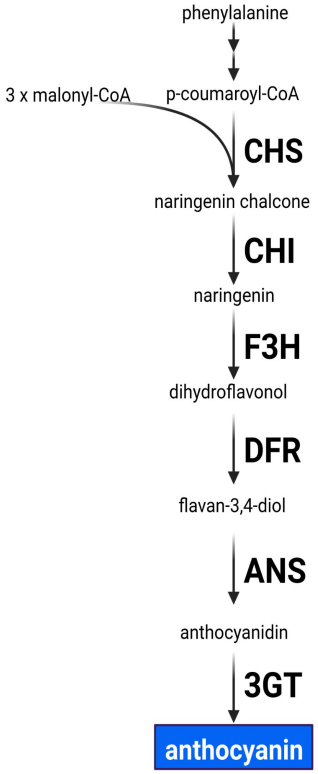
Pucker *et al.*, 2022: 10.1101/2022.10.19.512958

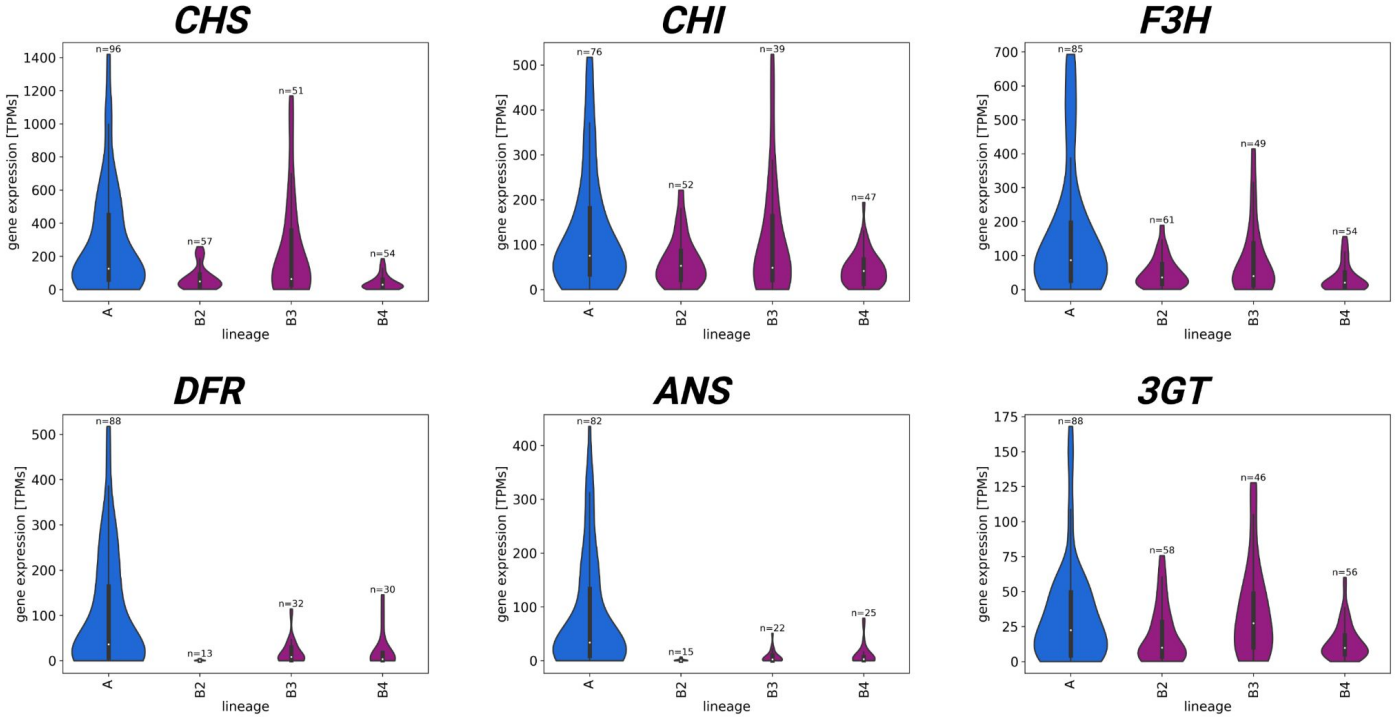
Cross species transcriptomics



Pucker *et al.*, 2022: 10.1101/2022.10.19.512958

Low transcriptional activity of anthocyanin biosynthesis genes





Pucker *et al.*, 2022: 10.1101/2022.10.19.512958

Summary

- Sharing protocols
- Licenses
- FAIR data
- ENA submission
- Scientific publication process
- Data life cycle
- Re-use

Time for questions!

Questions

1. Which licenses are available to share data sets / software?
2. What is license stacking and which issues result from it?
3. What is FAIR in data science?
4. How does a submission to ENA work?
5. Which steps and loops are part of a paper publishing process?
6. What elements are required in a cover letter?
7. What are h-index, i10-index, and (J)IF?
8. What are the advantages of preprints?
9. Which elements should be part of a review report?
10. Where would you deposit scripts/data sets?
11. What are the steps of a data life cycle?
12. What are the advantages and challenges of re-use?
13. Which options for QC and filtering of public RNA-seq data sets do you know?