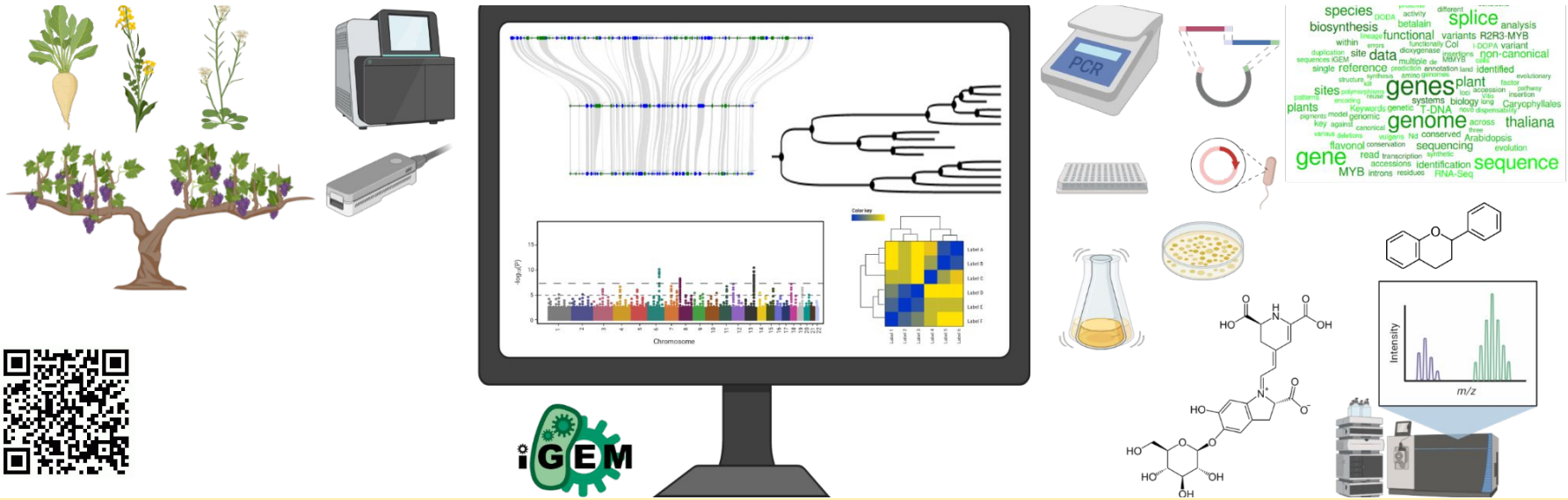
Technische Universität Braunschweig



GE32/MM12 - Data Dissemination & Exchange

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

Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - StudIP: **GE32/MM12**
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

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

My figures and content can be re-used in accordance with CC BY 4.0, but this might not apply to all images/logos. Some figure were constructed using bioRender.com.

Which elements of a study can be shared?

Which elements of a study can be shared?

- Plasmids
- Sequencing data / Sequences
- Specimen
- Seeds / cell lines / strains
- Phenotyping data
- Geographic positions
- ….

Seeds - Stock Centers

- Arabidopsis Biological Resource Center (ABRC)

- Nottingham Arabidopsis Stock Center (NASC)
  - GABI-Kat, SIGnAL

- Other Genetic Stock Centers:
  - Collections in the National Plant Germplasm System (USA)
  - Rice Genetic Resources Stock Collections (Japan)
  - Wheat Precise Genetic Stocks (UK)
  - International Moss Stock Center (Germany)

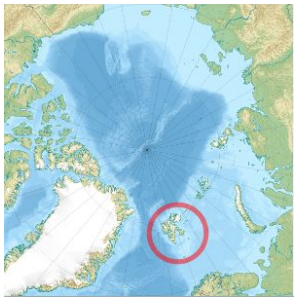
- Directly from authors

https://abrc.osu.edu/
https://arabidopsis.info/
http://maizecoop.cropsci.uiuc.edu/othrcntr.php

Seeds - Global Seed Vault

- Svalbard Global Seed Vault (seed backup only)
- Maintained by Global Crop Diversity Trust
- Located on Spitzbergen in permafrost soil
- Solid infrastructure for accessibility and continuous power supply
- Harboring seeds of the 21 most important crop species
- Storage temperature is -18°C
- Seeds can be stored for decades; old seeds are replaced

https://commons.wikimedia.org/wiki/File:Svalbard_seed_vault_IMG_8894.JPG (CC BY-SA)



https://de.wikipedia.org/wiki/Spitzbergen_(Inselgruppe)

How to submit seeds (to NASC)?

- Complete seed donation form and send via email
  - Name of the seed stock
  - Seed type (e.g. T-DNA line or transposon line)
  - Name gene/locus if mutant
  - Phenotype
  - Specific growth conditions
  - Publication (reference)
  - Origin of ecotypes

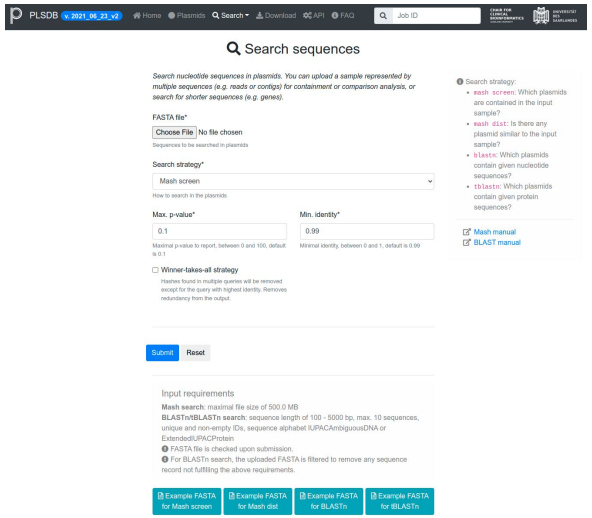
- NASC seed stock numbers will be assigned upon seed arrival

- >20,000 seeds (400mg) needed for initial submissions

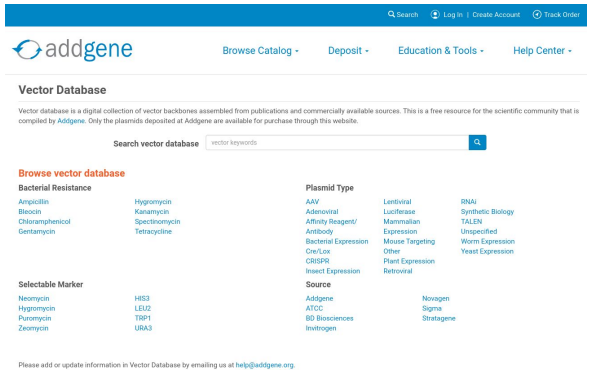
- Submit a picture of the phenotype

Plasmids

- Sequences can be deposited at the NCBI
- PLSDB contains plasmid sequences submitted to the NCBI
- Availability of plasmids is often an issue (not available from authors)
- Addgene maintains a collection of plasmids and ships these upon request
- Gene synthesis makes storage of plasmids less relevant



https://ccb-microbe.cs.uni-saarland.de/plsdb/



https://www.addgene.org/vector-database/

How to submit plasmids (to addgene)?

- Submission to addgene is free and avoids costs for shipping the materials to others
- Submit details about plasmids via email in a spreadsheet
  - Name of organization and PI
  - Related publication / embargo request
- Plasmids are assigned to a preprint or publication
- Plasmid sequences should be submitted as GenBank files

Submit Plasmids from a Published Article

- Look up your article as you would search on PubMed (i.e. title, author, or PMID).
  - No PubMed ID yet? Please use the pre-publication/unpublished submission option below.
- Addgene will provide links from your plasmids to this article.
- Scientists will be asked to cite this article in future publications.
- Must be logged in to start deposit.

Submit Pre-Publication or Unpublished Plasmids

Addgene encourages submitting pre-publication plasmids so they can be available online when the paper is published.

Plasmids can be linked to a peer-reviewed manuscript OR a preprint. In the case of preprint submission, we are happy to update the publication information once the manuscript is published in a peer-reviewed journal.

- Must be logged in to start deposit.

Submit Plasmids Using a Spreadsheet

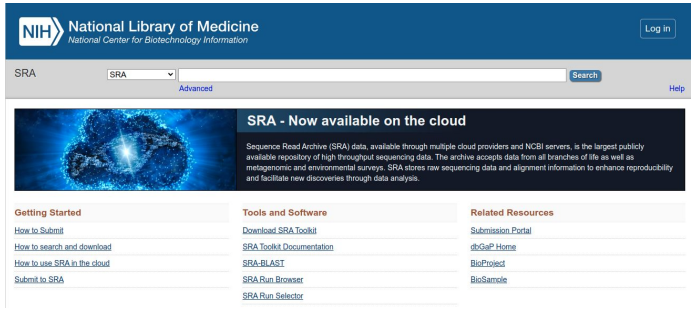
- Recommended for depositing 10 or more similar plasmids.
- Copy and paste your plasmid data directly into our file.
- Email the spreadsheet back to us at deposit@addgene.org along with:
  - Addgene account username. In order for the plasmids to appear in your Addgene account under **My Plasmids**, we need to know your Addgene account username.
  - Plasmid sequences or GenBank files. We can accept sequence files in any format. We encourage submission of QUEEN-generated GenBank files (Mori and Yachie, 2021).
  - Distribution status for your plasmids - Hold for Publication or Distribute pending QC.
  - The name of the Principal Investigator and Organization where these constructs were first created.

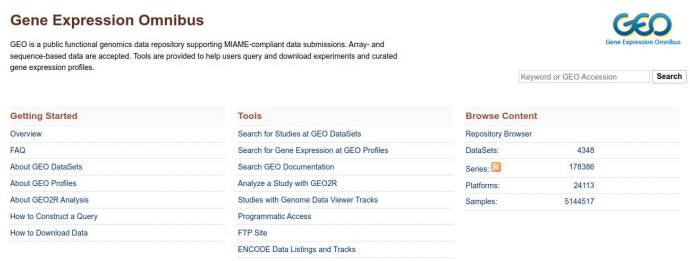
Download Deposit Spreadsheet

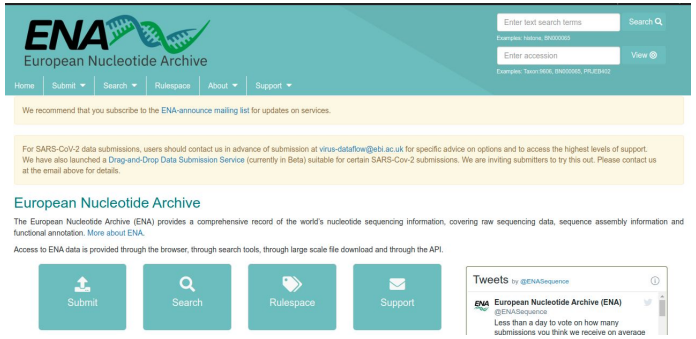
https://www.addgene.org/depositing/start-deposit/

Sequencing data sets

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



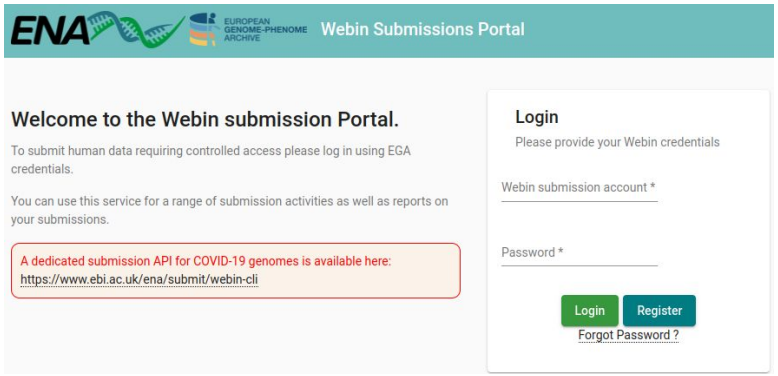




https://www.ncbi.nlm.nih.gov/sra
https://www.ncbi.nlm.nih.gov/geo/
https://www.ebi.ac.uk/ena/browser/home

How to submit reads (to ENA)?

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



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

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

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:

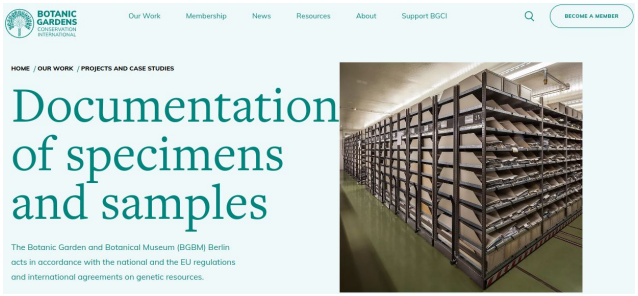
- File_type: FASTQ
- Library_layout: PAIRED
- Insert_size: 600
- Forward_file_name: fw_file.fastq.gz
- Forward_file_md5: jel9aks5joe8iaj1ie2lfk4jsk6flji
- Forward_file_unencrypted_md5:
- Reverse_file_name: rv_file.fastq.gz
- Reverse_file_md5: k1ea0wi7oji32so45jbae6fo81337xd
- reverse_file_unencrypted_md5

Plant collections

- Seed banks: focus on crops or model organisms

- Botanical gardens: living collections

- Museums/herbaria: collections of recent and ancestral plants



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

How to submit voucher herbarium specimen?

- Voucher herbarium specimen = pressed plant sample with collection data
- Voucher herbarium specimen are helpful to support phylogenetic reclassifications
- Process overview:
  - Initial preparations
  - Pressing and drying
  - Identification
  - Labeling
  - Mounting

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

Part1: Preparation for specimen collection

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

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

Part 2: Processing specimens

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

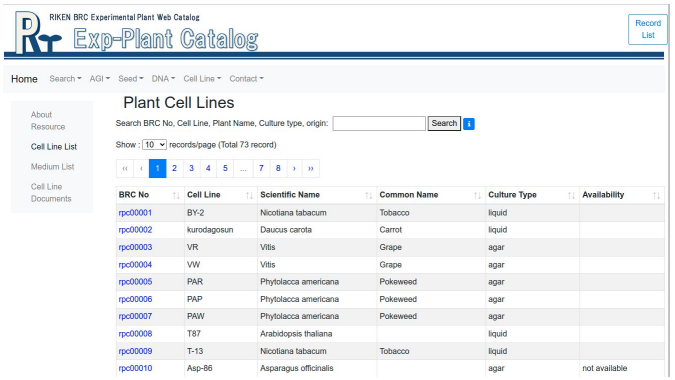
- Identification of plant specimens: dichotomous keys; published descriptions

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

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

Cell lines / strains

- RIKEN BRC Experimental Plant Web Catalog
- Maintenance of plant cell cultures is complicated
- Direct exchange between groups
- Example: At7



BRC No	Cell Line	Scientific Name	Common Name	Culture Type	Availability
rpc00001	BY-2	Nicotiana tabacum	Tobacco	liquid	
rpc00002	kurodagosun	Daucus carota	Carrot	liquid	
rpc00003	VR	Vitis	Grape	agar	
rpc00004	VW	Vitis	Grape	agar	
rpc00005	PAR	Phytolacca americana	Pokeweed	agar	
rpc00006	PAP	Phytolacca americana	Pokeweed	agar	
rpc00007	PAW	Phytolacca americana	Pokeweed	agar	
rpc00008	T87	Arabidopsis thaliana		liquid	
rpc00009	T-13	Nicotiana tabacum	Tobacco	liquid	
rpc00010	Asp-86	Asparagus officinalis		agar	not available

https://www.gmp-creativebiolabs.com/plant-cell-lines_62.htm

How to ship a cell line?

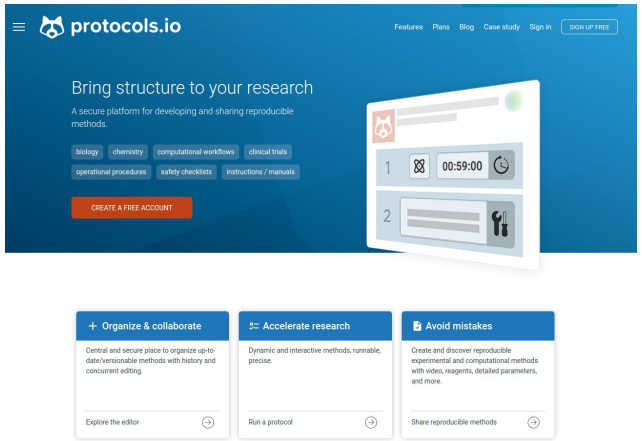
- Temperature needs to be kept constant
- Careful transport
- Sterile conditions
- Shipping liquid culture in plastic reaction tubes in package can work
- Shipping on MS agar plate

OpenData

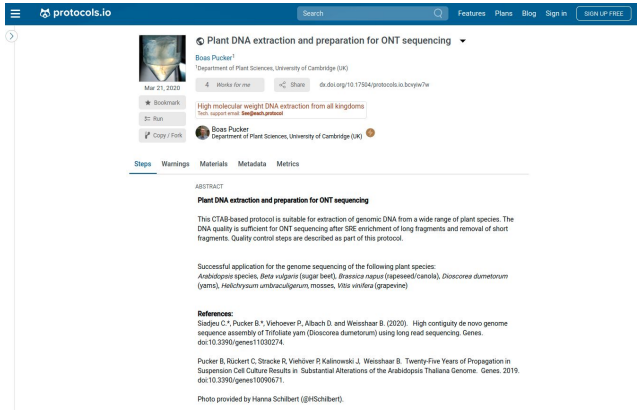
- Everyone can access and re-use these data sets
- Facts cannot be owned by someone
- Huge economic potential through re-use; advantages for society
- Possible restrictions: name author, share-alike
- Related initiative: open source, open content, open access, open education

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

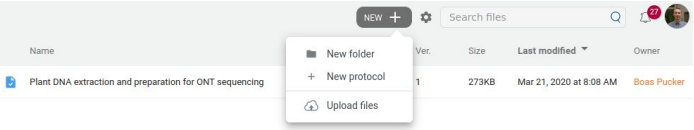


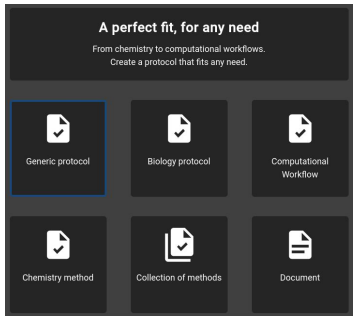
doi:10.17504/protocols.io.bcvyiw7w

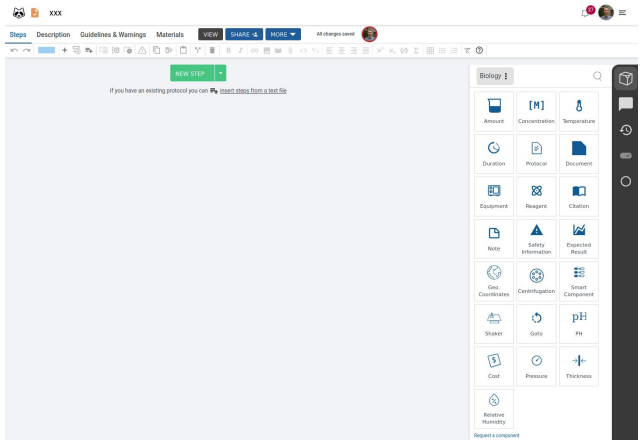
How to submit a protocol?

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

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







https://www.protocols.io/

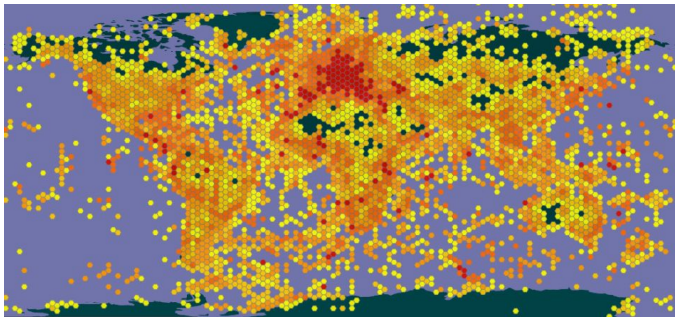
e!DAL - PGP phenotypic data sets

- PGP Repository = Plant Genomics & Phenomics Research Data Repository
- Central storage of large data sets avoids backup issues
- Documentation of data sets avoids metadata issues
- Data sets become citeable with DOIs
- See introduction video for additional details

https://edal-pgp.ipk-gatersleben.de/

GBIF - geographical positions

- GBIF: Global Biodiversity Information Facility
- Network of countries and organizations
- Species occurrence records
  - Species observed in the wild
  - Species observed in botanical gardens/zoos
  - Specimen in other collections



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?

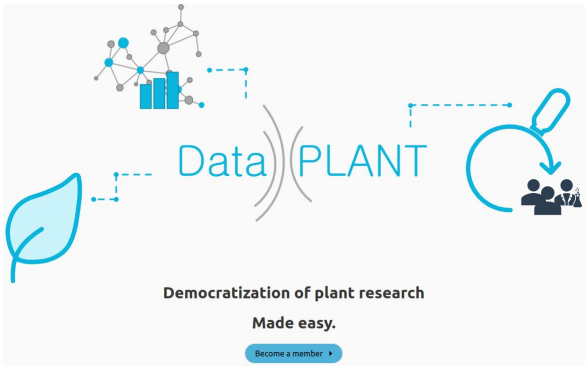
FAIR data

- 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

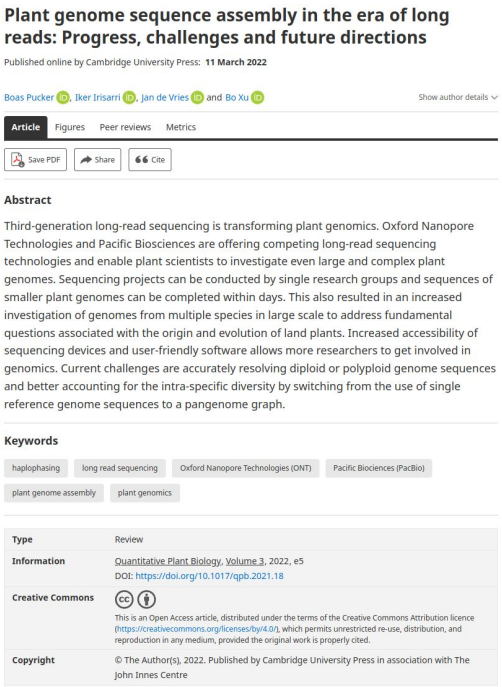
- NFDI = Nationale Forschungsdaten Infrastruktur e.V.
- Provide sustainable, annotated data management platform
- Pave way to pure data publications with research context
- Omics and imaging at petabyte size
- ARC = annotated research context
- Add user-oriented services to existing IT infrastructure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

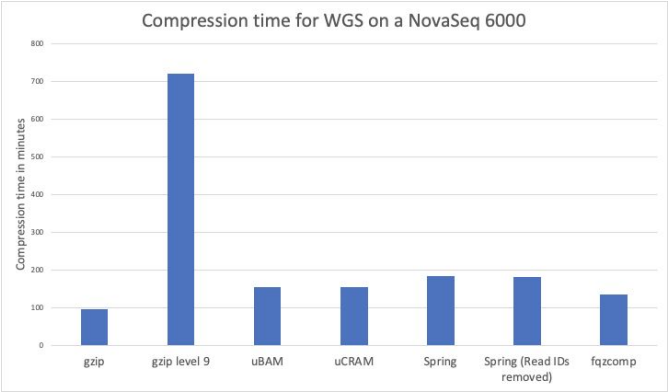
haplophasing | long read sequencing | Oxford Nanopore Technologies (ONT) | Pacific Biosciences (PacBio) | plant genome assembly | plant genomics

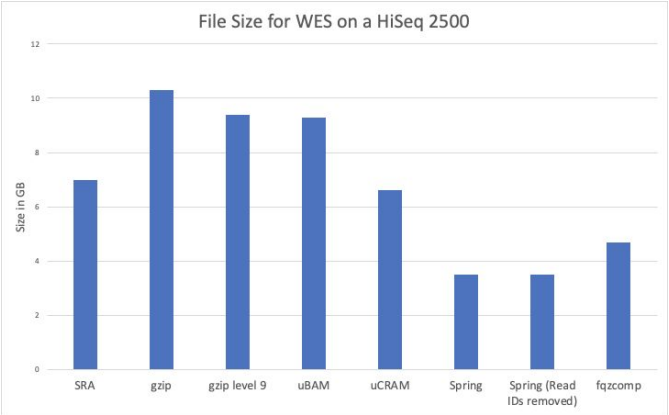
Type	Review
Information	Quantitative Plant Biology, Volume 3, 2022, e5 DOI: https://doi.org/10.1017/qpb.2021.18
Creative Commons	This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted 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 -xvf 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

Time for questions!

Questions

1. Which elements of a study can be shared?
2. How are seeds stored and distributed?
3. Where are plasmid data stored and shared?
4. Where can you deposit sequencing data sets?
5. How can you share a protocol for effective re-use by others?
6. Which license can be assigned to a data set?
7. Which license can be assigned to software?
8. What is FAIR data?
9. What is nfdi4plants?
10. What are the objectives of nfdi4plants?