CEPLAS
Cluster of Excellence on Plant Sciences

Good Data Management Practices

part of M4468 – Plant developmental genetics, evolution and biostatistics in the CEPLAS research program

November 10th, 2024

Vittorio Tracanna, Dominik Brilhaus
CEPLAS Data

Welcome

House-keeping

Pad: https://pad.hhu.de/0NdPKO5LQ5CHBRN2iuG91Q

Materials

Slides will be shared via DataPLANT knowledge base and the Sciebo folder

Tentative agenda

Day 1

Time	Topics
09:30 - 10:45	Intro to RDM and ARC
10:45 - 11:00	*Short break*
11:00 - 12:00	ARC Hands-on
12:30 - 13:30	*Lunch*
13:00 - 15:30	Data storage and sharing
15:30 - 16:00	Wrap-up

Day 2

Time	Topics
09:30 - 10:30	ARC Feedback session
10:30 - 10:45	*Short break*
10:45 - 12:00	ISA and Metadata
12:30 - 13:30	*Lunch*
13:00 - 15:00	Hands-on Swate
15:00 - 15:30	ARC ecosystem: Additional features
15:30 - 16:00	Wrap-up

Goals

- Appreciate FAIR principles
- Learn tools and services for FAIR data management
- Effectively manage your own research data

Introduce yourself

- Used code / programming language before
- Experience with Git / GitLab / GitHub?
- Have an ORCID
- My motivation / expectation
- My favorite lab assay

Let's draw a typical lab workflow 📝

Why Research Data Management (RDM)?

- Increase transparency
- Make data accessible
- Save time (writing, reusing)
- Reduce the risk of data loss
- Optimize the costs
- Facilitate future reuse and sharing
- Improve citations

Your first collaborators are your future selves, be nice to them !

your future self, by Julien Colomb, CC-BY-NC, derived from .NORM Normal File Format, CC-BY-NC, by Randall Munroe

The Research Data Lifecycle

The Research Data Lifecycle

The Research Data Lifecycle

The Research Data Lifecycle

The Research Data Lifecycle

The Research Data Lifecycle

The Research Data Lifecycle

The Research Data Lifecycle *is mutable*

FAIR

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

https://doi.org/10.1038/sdata.2016.18

nature > scientific data > comment > article

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons

— Show fewer authors

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

The FAIR principles

Is your data FAIR?

Findable | **A**ccessible | **I**nteroperable | **R**eusable

- Where do you store your data?
- How do you annotate your data?
- How do you share your data?
- What tools do you use to analyse your data?
- How do you reuse other people's data?

Findable

> The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- F1. (Meta)data are assigned a globally unique and persistent identifier.
- F2. Data are described with rich metadata (defined by R1 below).
- F3. Metadata clearly and explicitly include the identifier of the data they describe.
- F4. (Meta)data are registered or indexed in a searchable resource.

Accessible

> Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

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

Interoperable

> The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles.
- I3. (Meta)data include qualified references to other (meta)data.

Reusable

> The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
- R1.1. (Meta)data are released with a clear and accessible data usage license
- R1.2. (Meta)data are associated with detailed provenance
- R1.3. (Meta)data meet domain-relevant community standards

FAIR on multiple layers

> The principles refer to three types of entities: **data** (or any digital object), **metadata** (information about that digital object), and **infrastructure**.

CEPLAS – One cluster, four locations

Data silos impede collaboration

Data silos impede collaboration

Missing interfaces impede collaboration

📝 Understand your colleague's project 📝

1. Go to the shared folder prepared before class
2. Try to identify one experiment that led to results (e.g. a figure in the thesis)
3. What are the samples (e.g. plants, bacteria)?
4. How were the samples prepared (~ materials)?
5. How was the experiment performed (~ methods)?
6. What is the raw data (~ results)?
7. How was the data analyzed (~ computational methods, statistics)?
8. Collect all of the above in a `README_<YourArbitraryParticipantID>.txt` in the same folder.

Assignment

Participant	looks at project of
Participant01	Participant02
Participant02	Participant03
…	…
Participant n	Participant n+1

DataPLANT – The NFDI4Plants

- NFDI: "Nationale Forschungsdaten Infrastruktur" – www.nfdi.de
- Funded since end of 2020

Data Stewardship between DataPLANT and the community

Community

nfdi4plants

Domain experts
User experience
Training

Service provider
Developers
Tech experts

Annotated Research Context (ARC)

Your entire investigation in a single unified bag

You can store your ARC in the DataHUB

ARCs are versioned

You can invite collaborators

Collaborate and contribute

Reuse data in ARCs

Publish your ARC

Publish your ARC, get a DOI

ARC as single-entry point

Data analysis and workflows

Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

Galaxy integration: Extra value for plant research

e.g. https://plants.usegalaxy.eu

- Full ARC compatibility
- Automated metadata generation
- Specialized tools and workflows for ‘omics processing and analysis
- Public repository compatibility
- Galaxy teaching resource for data analysis

Enabling platforms

- Streamlined exchange of (meta)data
- Communication and project management

Streamlined data exchange

Meet your collaborators in an ARC

The ARC ecosystem

DataPLAN

Write DMP

Metadata Registry

Search ARCs
by metadata

ARCitect
ARC Commander

Create
Add metadata
Sync & track
changes

ARChigator

Publish & Archive

DataHUB

Share & Collaborate
Validate
(Hub to additional
services)

Swate

Annotate
experiments

Knowledge Base

Training & Teaching

What does an ARC look like?

ARCs store experimental data

Computations can be run inside ARCs

ARCs come with comprehensive metadata

ARC builds on standards

- RO-Crate: standardized exchange
- ISA: structured, machine-readable metadata
- CWL: reproducible, re-usable data analysis
- Git: version control

https://isa-tools.org/ | https://www.commonwl.org/
https://www.researchobject.org/ro-crate/ | https://git-scm.com

CEPLAS connection to the NFDI

Data Stewardship between DataPLANT and the community

Community

nfdi4plants

Domain experts
User experience
Training

Service provider
Developers
Tech experts

The DataPLANT DataHUB – a GitLab *Plus*

=

GitLab

+

On-premise DataHUBs

ARC services are available as on-premise option

HHU-DataHUB

intensity
m/z
eLabFTW
elabftw.hhu.de
DataHUB
datahub.hhu.de
Compute-intense
data analysis
HPC
hpc.rz.uni-
duesseldorf.de
Image
Analysis
Data
publication
OMERO
omero-cai.hhu.de
OMERO
Connection
to DataHUB
network
ResearchData
researchdata.hhu.de
DSPACE
Data
PLANT

HHU-DataHUB

- FAIR data along the research data cycle
- Structured exchange of (meta)data
- Local (long-term) storage, backup, version control
- Integration of existing services and infrastructure
- External collaboration

HHU-DataHUB – Node for platforms

Embedded into an RDM network

CEPLAS collaborates with DataPLANT and NFDI4BIOIMAGE

CEPLAS Research Data Policy

CEPLAS
Cluster of Excellence on Plant Sciences

CEPLAS Research Data Policy

Preamble

CEPLAS researchers generate, collect, and compile a multitude of datasets. The interdisciplinary focus of CEPLAS offers and encourages opportunities for collaborative research across individual investigations. To take full advantage of synergies and gain a wider interdisciplinary collaboration, the sharing and collaborative use of research data is crucial. Research data produced by CEPLAS investigations require thorough data documentation, quality control and long-term storage, preservation and accessibility as stipulated by the DFG, as well as fine grained access control. These topics are the subject of this policy. The aim of CEPLAS research data management (RDM) is the implementation of the FAIR (findable, accessible, interoperable, reusable) guiding principles of data stewardship[1]. The long-term aim is to develop tools, pipelines and expertise for integrated data analysis and visualization via web browsers and support the increasing demand to unlock and integrate datasets and increase the compatibility with advanced analysis software and tools to maximise the mining of available datasets.

[1] Wilkinson, M.D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 2018 5 2. 160018–8. https://doi.org/10.1038/sdata.2016.18

1

Hands-on: ARCitect part 1

Follow the **Start Here** guide in the DataPLANT knowledge base.

Until step "add a study"

- 01-create-arc-scaffold.mdx
- 02-investigation.mdx
- 03-study.mdx

Metadata and ISA

What is metadata?

Viola's PhD Project

Exercise: Take 5 minutes to note down the metadata

Viola investigates the effect of the plant circadian clock on sugar metabolism in *W. mirabilis*. For her PhD project, which is part of an EU-funded consortium in Prof. Beetroot's lab, she acquires seeds from a South-African botanical society. Viola grows the plants under different light regimes, harvests leaves from a two-day time series experiment, extracts polar metabolites as well as RNA and submits the samples to nearby core facilities for metabolomics and transcriptomics measurements, respectively. After a few weeks of iterative consultation with the facilities' heads as well as technicians and computational biologists involved, Viola receives back a wealth of raw and processed data. From the data she produces figures and wraps everything up to publish the results in the Journal of Wonderful Plant Sciences.

Metadata everywhere

`Viola` investigates the `effect of the plant circadian clock` on `sugar metabolism` in *`W. mirabilis`*. For her `PhD project`, which is part of an `EU-funded consortium` in `Prof. Beetroot's lab`, she acquires `seeds` from a `South-African botanical society`. Viola `grows the plants` under `different light regimes`, harvests `leaves` from a `two-day time series experiment`, extracts `polar metabolites` as well as `RNA` and submits the samples to nearby `core facilities for metabolomics and transcriptomics` measurements, respectively. `After a few weeks` of iterative consultation with the facilities' heads as well as `technicians` and `computational biologists` involved, Viola receives back a wealth of `raw and processed data`. From the data she `produces figures` and wraps everything up to `publish the results in the Journal of Wonderful Plant Sciences`.

Project metadata

project design

- researcher
- institute and project
- biological context
- research question
- purpose of data collection
- ...

experimental processes

- origin and nature of the biological material
- lab protocols
- instrument model
- ...

data-analytical processes

- algorithms
- tools
- software versions and dependencies employed
- ...

Other types of metadata

bibliographic

- Title
- Publication date and title
- Description
- Author
- Contacts
- Keywords
- ...

legal or administrative

- data origin, ownership, rovenance,
- licensing
- ethical aspects
- ...

technical

- expected data volume
- storage location
- file formats
- ...

Metadata from a FAIR perspective

Findable

- metadata names the content of the data
- basis for search engines
- makes it categorizable for people and machines

Accessible

- information about origin
- location of storage
- access rights

Interoperable

- metadata identifies software and file formats
- required conversions between file formats

Reusable

- obtain and reuse research data according to clear rules described in licenses

Metadata "Standards"

Examples from Minimum Information for Biological and Biomedical Investigations (MIBBI):

- MIAPPE | Minimum Information About a Plant Phenotyping Experiment
  https://www.miappe.org
- MIAME | Minimum Information About a Microarray Experiment
  https://www.fged.org/projects/miame/
- MIAPE | Minimum Information About a Proteomics Experiment
  https://www.psidev.info/miape
- MINSEQE | Minimum Information about a high-throughput SEQuencing Experiment
  https://www.fged.org/projects/minseqe

💡 Check out https://fairsharing.org/ for more examples

Metadata standards ≈ Checklists

- Determine (minimal) required information
- Usually **do not** determine the format (i.e. shape or file type)

A small Interactive detour

-> favorite Movie

How does google "know"?!

Schemas and machine-readability

Structured data and the internet

Schema.org

- create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, ...
- Structured data can be used to ***mark up*** all kinds of items from products to events to recipes
- Communicate with search engines (-> SEO, search engine optimization)
- Enhance findability from search engine results
- Provide context to an ambigous webpage
- Metadata interoperability and standardization across all website using schema.org

Structured data and the internet: Schema.org

https://schema.org/Person

```
<script type="application/ld+json">
{
  "@context": "https://schema.org",
  "@type": "Person",
  "address": {
    "@type": "PostalAddress",
    "addressLocality": "Seattle",
    "addressRegion": "WA",
    "postalCode": "98052",
    "streetAddress": "20341 Whitworth Institute 405 N. Whitworth"
  },
  "colleague": [
    "http://www.xyz.edu/students/alicejones.html",
    "http://www.xyz.edu/students/bobsmith.html"
  ],
  "email": "mailto:jane-doe@xyz.edu",
  "image": "janedoe.jpg",
  "jobTitle": "Professor",
  "name": "Jane Doe",
  "telephone": "(425) 123-4567",
  "url": "http://www.janedoe.com"
}
</script>
```

JSON-LD

JSON-LD = JavaScript Object Notation for Linked Data

```
<script type="application/ld+json">
  {
    "@context": "https://schema.org",
    "@type": "SportsTeam",
    "name": "San Francisco 49ers",
    "member": {
      "@type": "OrganizationRole",
      "member": {
        "@type": "Person",
        "name": "Joe Montana"
      },
      "startDate": "1979",
      "endDate": "1992",
      "roleName": "Quarterback"
    }
  }
</script>
```

RDFa

RDFa = Resource Description Framework in Attributes

```
<div vocab="http://schema.org/" typeof="SportsTeam">
  <span property="name">San Francisco 49ers</span>
  <div property="member" typeof="OrganizationRole">
    <div property="member" typeof="http://schema.org/Person">
      <span property="name">Joe Montana</span>
    </div>
    <span property="startDate">1979</span>
    <span property="endDate">1992</span>
    <span property="roleName">Quarterback</span>
  </div>
</div>
```

Standards

Dublin Core

https://www.dublincore.org/schemas/

DataCite Schema

- Schema: http://schema.datacite.org/meta/kernel-4.3/metadata.xsd
- Full Example: https://schema.datacite.org/meta/kernel-4.3/example/datacite-example-full-v4.xml

DataCite Schema: Simple Example

```
...
  <identifier identifierType="DOI">10.5072/D3P26Q35R-Test</identifier>
  <creators>
    <creator>
      <creatorName nameType="Personal">Fosmire, Michael</creatorName>
      <givenName>Michael</givenName>
      <familyName>Fosmire</familyName>
    </creator>
    <creator>
      <creatorName nameType="Personal">Wertz, Ruth</creatorName>
      <givenName>Ruth</givenName>
      <familyName>Wertz</familyName>
    </creator>
    <creator>
      <creatorName nameType="Personal">Purzer, Senay</creatorName>
      <givenName>Senay</givenName>
      <familyName>Purzer</familyName>
    </creator>
  </creators>
  <titles>
    <title xml:lang="en">Critical Engineering Literacy Test (CELT)</title>
  </titles>
  <publisher xml:lang="en">Purdue University Research Repository (PURR)</publisher>
  <publicationYear>2013</publicationYear>
  <subjects>
    <subject xml:lang="en">Assessment</subject>
    <subject xml:lang="en">Information Literacy</subject>
    <subject xml:lang="en">Engineering</subject>
    <subject xml:lang="en">Undergraduate Students</subject>
    <subject xml:lang="en">CELT</subject>
    <subject xml:lang="en">Purdue University</subject>
  </subjects>
  <language>en</language>
  <resourceType resourceTypeGeneral="Dataset">Dataset</resourceType>
...
```

Ontologies

Ontology

(Sometimes also referred to "semantic model")

An ontology combines features of

- a **dictionary**,
- a **taxonomy**, and
- a **thesaurus**

Dictionary

Alphabetically lists terms and their definitions

Pizza: *"a dish made typically of flattened bread dough spread with a savory mixture usually including tomatoes and cheese and often other toppings and baked"*

Taxonomy

Hierarchy or classification

Thesaurus

Dictionary of synonyms and relations

Pizza ≈ Lahmacun ≈ Focaccia ≈ Flammkuchen

Ontology

- Structures a set of **concepts** in a particular area and the relations between them in a **graph-like manner**
- Can be used in disambiguation, defining hierarchies, a standard to define terms
- Define a common vocabulary of concepts and their relationships to **model** a particular domain while making it **machine understandable**

The semantic triple

Modeling a pizza menu

Modeling a pizza menu

Modeling a pizza menu

Predicates have two directions

Looking at the menu from a different perspective

An object of one triplet can be the subject to another

(Towards) a knowledge graph

Searching the menu

An ontology can be queried:

- *"name all pizzas with topping mushrooms"*

The Pizza Ontology

- Example from protege: https://protege.stanford.edu/ontologies/pizza/pizza.owl
- Visualize via WebVOWL http://vowl.visualdataweb.org/webvowl.html

Example ontologies

EDAM ontology

- Description: http://edamontology.org/page
- Browser: https://edamontology.github.io/edam-browser

PECO ontology

- Human-readble: https://www.ebi.ac.uk/ols/ontologies/peco
- Raw (OWL): http://purl.obolibrary.org/obo/peco.owl

> Explore more examples
>
> - https://www.ebi.ac.uk/ols/
> - https://bioportal.bioontology.org

ARC builds on ISA

Investigation

Overall goals
Scientific context

Study

Experimental design

Assay

Leading to
(raw) data

ARC builds on ISA

Source Name	Characteristics [Organism]	Sample Name
Ath001	*Arabidopsis thaliana*	Lre001
Ath002	*Arabidopsis thaliana*	Lre002
Ath003	*Arabidopsis thaliana*	Lre003
…	…	…

Source Name	Parameter [instrument model]	Raw Data File
Lre001	6130 Quadrupole LC/MS	Lre001.wiff
Lre002	6130 Quadrupole LC/MS	Lre002.wiff
Lre003	6130 Quadrupole LC/MS	Lre003.wiff
…	…	…

Investigation Title	
Investigation Description	
Investigation Person Last Name	

Metadata Annotations

isa.<>.xlsx files within ARCs

isa.investigation.xlsx

	A									
1	ONTOLOGY SOURCE REFERENCE									
2	Term Source Name	OBI	BTO	NEWT	UO	CHEBI	PATO	EFO		
3	Term Source File	http://bioportal.bioo	ArrayExpress Experimental Factor Ontology							
4	Term Source Version	47393	v 1.26	v 1.26	v 1.26	v 1.26	v 1.26	v 1.26		
5	Term Source Description	Ontology for Biomed	BRENDA tissue /	NEWT UniProt Taxo	Unit Ontolog	Chemical Ent	Phenotypic c	ArrayExpress Experimental Factor Ontology		
6	INVESTIGATION									
7	Investigation Identifier	BII-I-1								
8	Investigation Title	Growth control of the eukaryote cell: a systems biology study in yeast								
9	Investigation Description	Background Cell growth underlies many key cellular and developmental processes, yet a limited number of studies have been carried out on cell gr								
10	Investigation Submission Date	30.04.07								
11	Investigation Public Release Date	10.03.09								
12	Comment[Created With Configuration]									
13	Comment[Last Opened With Configuration]	isaconfig-default_v2013-02-13								
14	INVESTIGATION PUBLICATIONS									
15	Investigation PubMed ID	17439666								
16	Investigation Publication DOI	doi:10.1186/jbiol54								
17	Investigation Publication Author List	Castrillo JI, Zeef LA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell MJ, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn WB, Broadhurst								
18	Investigation Publication Title	Growth control of the eukaryote cell: a systems biology study in yeast.								
19	Investigation Publication Status	indexed in Pubmed								
20	Investigation Publication Status Term Accession Number									
21	Investigation Publication Status Term Source REF									
22	INVESTIGATION CONTACTS									
23	Investigation Person Last Name	Stephen	Castrillo	Zeef						
24	Investigation Person First Name	Oliver	Juan	Leo						
25	Investigation Person Mid Initials	G	I	A						
26	Investigation Person Email									
27	Investigation Person Phone									
28	Investigation Person Fax									
29	Investigation Person Address	Oxford Road, Manch	Oxford Road, Ma	Oxford Road, Manchester M13 9PT, UK						
30	Investigation Person Affiliation	Faculty of Life Scienc	Faculty of Life Sc	Faculty of Life Sciences, Michael Smith Building, University of Manchester						
31	Investigation Person Roles	corresponding authc	author	author						
32	Investigation Person Roles Term Accession Number									
33	Investigation Person Roles Term Source REF									
34	Comment[Investigation Person REF]									
35	STUDY									
36	Study Identifier	BII-S-1								
37	Study Title	Study of the impact of changes in flux on the transcriptome, proteome, endometabolome and exometabolome of the yeast Saccharomyces cerevi								
38	Study Description	We wished to study the impact of growth rate on the total complement of mRNA molecules, proteins, and metabolites in S. cerevisiae, independer								
39	Comment[Study Grant Number]									
40	Comment[Study Funding Agency]									
41	Study Submission Date	30.04.07								
42	Study Public Release Date	10.03.09								
43	Study File Name	s_BII-S-1.txt								
44	STUDY DESIGN DESCRIPTORS									
45	Study Design Type	Intervention design								
46	Study Design Type Term Accession Number	http://purl.obolibrary.org/obo/OBI_0000115								
47	Study Design Type Term Source REF	OBI								
48	STUDY PUBLICATIONS									
49	Study PubMed ID	17439666								
50	Study Publication DOI	doi:10.1186/jbiol54								
51	Study Publication Author List	Castrillo JI, Zeef LA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell MJ, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn WB, Broadhurst								
52	Study Publication Title	Growth control of the eukaryote cell: a systems biology study in yeast.								
53	Study Publication Status	published								
54	Study Publication Status Term Accession Number									
55	Study Publication Status Term Source REF									
56	STUDY FACTORS									
57	Study Factor Name	limiting nutrient	rate							
58	Study Factor Type	chemical compound	rate							
59	Study Factor Type Term Accession Number		http://purl.obolibrary.org/obo/PATO_0000161							
60	Study Factor Type Term Source REF		PATO							
61	STUDY ASSAYS									
62	Study Assay Measurement Type	protein expression p	metabolite profil	transcription profiling						
63	Study Assay Measurement Type Term Accession Number	http://purl.obolibrar	http://purl.obolil	424						
64	Study Assay Measurement Type Term Source REF	OBI	OBI	OBI						
65	Study Assay Technology Type	mass spectrometry	mass spectromet	DNA microarray						
66	Study Assay Technology Type Term Accession Number			http://purl.obolibrary.org/obo/OBI_0400148						
67	Study Assay Technology Type Term Source REF	OBI	OBI	OBI						
68	Study Assay Technology Platform	iTRAQ	LC-MS/MS	Affymetrix						
69	Study Assay File Name	a_proteome.txt	a_metabolome.tx	a_transcriptome.txt						
70	STUDY PROTOCOLS									
71	Study Protocol Name	growth protocol	mRNA extraction	protein extraction	biotin labelin	iTRAQ labelin	EukGE-WS4	metabolite extraction		
72	Study Protocol Type	growth	mRNA extraction	protein extraction	labeling	labeling	hybridizatior	extraction		
73	Study Protocol Type Term Accession Number							http://purl.obolibrary.org/obo/OBI_0302884		
74	Study Protocol Type Term Source REF							OBI		
75	Study Protocol Description	1. Biomass samples (	1. Biomass samples (45 ml) were tak		This was done using Enzo		For each target, a hybridisation cocktail was made using the			
76	Study Protocol URI									

isa.study.xlsx

	A	B		I	M
1	Source Name	Characteristics [soluble protein content]	Parameter [Quantification method#2]	Parameter [15N Photosynthesis QconCAT mass#4]	Sample Name
2	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_U1
3	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_U2
4	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_U3
5	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_U4
6	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_5_1
7	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_5_2
8	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_5_3
9	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_5_4
10	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_UF_1
11	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_UF_2
12	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_UF_3
13	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_UF_4
14	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_5F_1
15	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_5F_2
16	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_5F_3
17	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_5F_4
18	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr1_U1
19	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr1_U2
20	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr1_U3
21	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr1_U4
22	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr1_5_1
23	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr1_5_2
24	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr1_5_3

isa.assay.xlsx

	A	B	F		P	T
1	Source Name	Parameter [sample volume]	Parameter [injection volu	ment model]	Parameter [measurement duration#4]	Raw Data File
2	WCGr2_U1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U1.wiff
3	WCGr2_U2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U2.wiff
4	WCGr2_U3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U3.wiff
5	WCGr2_U4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U4.wiff
6	WCGr2_5_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_1.wiff
7	WCGr2_5_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_2.wiff
8	WCGr2_5_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_3.wiff
9	WCGr2_5_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_4.wiff
10	WCGr2_UF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_1.wiff
11	WCGr2_UF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_2.wiff
12	WCGr2_UF_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_3.wiff
13	WCGr2_UF_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_4.wiff
14	WCGr2_5F_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_1.wiff
15	WCGr2_5F_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_2.wiff
16	WCGr2_5F_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_3.wiff
17	WCGr2_5F_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_4.wiff
18	WCGr1_U1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U1.wiff
19	WCGr1_U2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U2.wiff
20	WCGr1_U3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U3.wiff
21	WCGr1_U4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U4.wiff
22	WCGr1_5_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_1.wiff
23	WCGr1_5_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_2.wiff
24	WCGr1_5_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_3.wiff
25	WCGr1_5_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_4.wiff
26	WCGr1_UF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_UF_1.wiff
27	WCGr1_UF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_UF_2.wiff

Data PLANT
CEPLAS
CC BY

Study and assay files are registered in the investigation file

isa.investigation.xlsx

	A									
1	ONTOLOGY SOURCE REFERENCE									
2	Term Source Name	OBI	BTO	NEWT	UO	CHEBI	PATO	EFO		
3	Term Source File	http://bioportal.bioo	ArrayExpress Experimental Factor Ontology							
4	Term Source Version	47893	v 1.26	v 1.26	v 1.26	v 1.26	v 1.26	v 1.26		
5	Term Source Description	Ontology for Biomed	BRENDA tissue /	NEWT UniProt Taxi	Unit Ontolog	Chemical Ent	Phenotypic c	ArrayExpress Experimental Fact		
6	INVESTIGATION									
7	Investigation Identifier	BII-I-1								
8	Investigation Title	Growth control of the eukaryote cell: a systems biology study in yeast								
9	Investigation Description	Background Cell growth underlies many key cellular and developmental processes, yet a limited numbe								
10	Investigation Submission Date	30.04.07								
11	Investigation Public Release Date	10.03.09								
12	Comment[Created With Configuration]									
13	Comment[Last Opened With Configuration]	isaconfig default_v2013-02-13								
14	INVESTIGATION PUBLICATIONS									
15	Investigation PubMed ID	17439666								
16	Investigation Publication DOI	doi:10.1186/jbiol54								
17	Investigation Publication Author List	Castrillo JI, Zeef LA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell MJ, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn WB, Broadhurst								
18	Investigation Publication Title	Growth control of the eukaryote cell: a systems biology study in yeast.								
19	Investigation Publication Status	indexed in Pubmed								
20	Investigation Publication Status Term Accession Number									
21	Investigation Publication Status Term Source REF									
22	INVESTIGATION CONTACTS									
23	Investigation Person Last Name	Stephen	Castrillo							
24	Investigation Person First Name	Oliver	Juan							
25	Investigation Person Mid Initials	G								
26	Investigation Person Email									
27	Investigation Person Phone									
28	Investigation Person Fax									
29	Investigation Person Address	Oxford Road, Manch	Oxford Road, Ma	Oxford Road, Manchester M13 9PT, UK						
30	Investigation Person Affiliation	Faculty of Life Scienc	Faculty of Life Sc	Faculty of Life Sciences, Michael Smith Building, University of Manchester						
31	Investigation Person Roles	corresponding author		author						
32	Investigation Person Roles Term Accession Number									
33	Investigation Person Roles Term Source REF									
34	Comment[Investigation Person REF]									
35	STUDY									
36	Study Identifier	BII-S-1								
37	Study Title	Study of the impact of changes in flux on the transcriptome, proteome, endometabolome and exometabolome of the yeast Saccharomyces cerev								
38	Study Description	We wish to the impact of growth rate on the total complement of mRNA molecules, proteins, and metabolites in S. cerevisiae, independe								
39	Comment[Study Grant Number]									
40	Comment[Study Funding Agency]									
41	Study Submission Date									
42	Study Public Release Date									
43	Study File Name									
44	STUDY DESIGN DESCRIPTORS									
45	Study Design Type									
46	Study Design Type Term Accession Number			00115						
47	Study Design Type Term Source REF	OBI								
48	STUDY PUBLICATIONS									
49	Study PubMed ID	17439666								
50	Study Publication DOI	doi:10.1186/jbiol54								
51	Study Publication Author List	Castrillo JI, Zeef LA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell MJ, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn WB, Broadhurst								
52	Study Publication Title	Growth control of the eukaryote cell: a systems biology study in yeast.								
53	Study Publication Status	published								
54	Study Publication Status Term Accession Number									
55	Study Publication Status Term Source REF									
56	STUDY FACTORS									
57	Study Factor Name	limiting nutrient	rate							
58	Study Factor Type	chemical compound	rate							
59	Study Factor Type Term Accession Number		http://purl.obolibrary.org/	PATO_0000161						
60	Study Factor Type Term Source REF		PATO							
61	STUDY ASSAYS									
62	Study Assay Measurement Type	protein expression p	metabolite profil	tr ription profiling						
63	Study Assay Measurement Type Term Accession Number	http://purl.obolibrar	http://purl.o	424						
64	Study Assay Measurement Type Term Source REF	OBI	OBI							
65	Study Assay Technology Type	mass spectrometry	mass spectrom	A microarray						
66	Study Assay Technology Type Term Accession Number			http://purl.obolibrary.org/obo/OBI_0400148						
67	Study Assay Technology Type Term Source REF	OBI								
68	Study Assay Technology Platform	iTRAQ								
69	Study Assay File Name	a_proteome.tx								
70	STUDY PROTOCOLS									
71	Study Protocol Name	growth protoc			labelin	iTRAQ labelin	EukGE-WS4	metabolite extraction		
72	Study Protocol Type	growth	mRNA extraction	protein extraction	labeling	labeling	hybridizatio	extraction		
73	Study Protocol Type Term Accession Number							http://purl.obolibrary.org/obo/OBI_0302884		
74	Study Protocol Type Term Source REF							OBI		
75	Study Protocol Description	1. Biomass samples (	1. Biomass samples (45 ml) were tak		This was done using Enzo		For each target, a hybridisation cocktail was made using the			
76	Study Protocol URI									

isa.study.xlsx

	A	B		I	M
1	Name	Characteristics [soluble protein content]	Parameter [Quantification method#2]	Parameter [15N Photosynthesis QconCAT mass#4]	Sample Name
2	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_U1
3	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_U2
4	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_U3
5	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_U4
6	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_5_1
7	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_5_2
8	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_5_3
9	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_5_4
10	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_UF_1
11	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_UF_2
12	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_UF_3
13	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_UF_4
14	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_5F_1
15	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_5F_2
16	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_5F_3
17	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_5F_4
18	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr1_U1
19	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr1_U2
20	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr1_U3
21	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr1_U4
22	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr1_5_1
23	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr1_5_2
24	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr1_5_3

isa.assay.xlsx

	A				P	T
1	Source Name	Parameter [sample volume]	Parameter [injection volu	ment model]	Parameter [measurement duration#4]	Raw Data File
2	WCGr2_U1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U1.wiff
3	WCGr2_U2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U2.wiff
4	WCGr2_U3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U3.wiff
5	WCGr2_U4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U4.wiff
6	WCGr2_5_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_1.wiff
7	WCGr2_5_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_2.wiff
8	WCGr2_5_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_3.wiff
9	WCGr2_5_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_4.wiff
10	WCGr2_UF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_1.wiff
11	WCGr2_UF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_2.wiff
12	WCGr2_UF_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_3.wiff
13	WCGr2_UF_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_4.wiff
14	WCGr2_5F_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_1.wiff
15	WCGr2_5F_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_2.wiff
16	WCGr2_5F_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_3.wiff
17	WCGr2_5F_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_4.wiff
18	WCGr1_U1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U1.wiff
19	WCGr1_U2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U2.wiff
20	WCGr1_U3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U3.wiff
21	WCGr1_U4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U4.wiff
22	WCGr1_5_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_1.wiff
23	WCGr1_5_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_2.wiff
24	WCGr1_5_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_3.wiff
25	WCGr1_5_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_4.wiff
26	WCGr1_UF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_UF_1.wiff
27	WCGr1_UF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_UF_2.wiff

Study file

Assay file

Data PLANT
CEPLAS
CC BY

The output of a study or assay file can function as input for a new isa.assay.xlsx

Output building blocks:

- Sample Name
- Raw Data File
- Derived Data File

	A	B	I	M
1	Source Name	Characteristics [soluble protein content]	Parameter [Quantification method#2]	Parameter [15N Photosynthesis QconCAT mass#4]	Sample Name
2	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_U1
3	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_U2
4	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_U3
5	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_U4
6	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_5_1
7	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_5_2
8	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_5_3
9	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_5_4
10	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_UF_1
11	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_UF_2
12	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_UF_3
13	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_UF_4
14	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_5F_1
15	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_5F_2
16	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_5F_3
17	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_5F_4
18	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr1_U1
19	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr1_U2
20	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr1_U3
21	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr1_U4
22	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr1_5_1
23	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr1_5_2
24	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr1_5_3

	A	B	F		P	T
1	Source Name	Parameter [sample volume]	Parameter [injection volume]	ment model]	Parameter [measurement duration#4]	Raw Data File
2	WCGr2_U1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U1.wiff
3	WCGr2_U2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U2.wiff
4	WCGr2_U3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U3.wiff
5	WCGr2_U4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U4.wiff
6	WCGr2_5_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_1.wiff
7	WCGr2_5_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_2.wiff
8	WCGr2_5_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_3.wiff
9	WCGr2_5_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_4.wiff
10	WCGr2_UF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_1.wiff
11	WCGr2_UF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_2.wiff
12	WCGr2_UF_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_3.wiff
13	WCGr2_UF_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_UF_4.wiff
14	WCGr2_5F_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_1.wiff
15	WCGr2_5F_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_2.wiff
16	WCGr2_5F_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_3.wiff
17	WCGr2_5F_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_4.wiff
18	WCGr1_U1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U1.wiff
19	WCGr1_U2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U2.wiff
20	WCGr1_U3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U3.wiff
21	WCGr1_U4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_U4.wiff
22	WCGr1_5_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_1.wiff
23	WCGr1_5_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_2.wiff
24	WCGr1_5_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_3.wiff
25	WCGr1_5_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_5_4.wiff
26	WCGr1_UF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_UF_1.wiff
27	WCGr1_UF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr1_UF_2.wiff

isa.study.xlsx

	A	B	I	M
1	Source Name	Characteristics [soluble protein content]	Parameter [Quantification method#2]	Parameter [15N Photosynthesis QconCAT mass#4]	Sample Name
2	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_U1
3	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_U2
4	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_U3
5	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_U4
6	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_5_1
7	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_5_2
8	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_5_3
9	G2_532_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_5_4
10	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_UF_1
11	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_UF_2
12	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_UF_3
13	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_UF_4
14	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_5F_1
15	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr2_5F_2
16	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr2_5F_3
17	G2_532_F3_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr2_5F_4
18	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr1_U1
19	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr1_U2
20	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr1_U3
21	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WCGr1_U4
22	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr1_5_1
23	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WCGr1_5_2
24	G1_532_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WCGr1_5_3

isa.assay.xlsx

	A	B	F		P	T
1	Source Name	Parameter [sample volume]	Parameter [injection volume]	ment model]	Parameter [measurement duration#4]	Raw Data File
2	WCGr2_U1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U1.wiff
3	WCGr2_U2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U2.wiff
4	WCGr2_U3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U3.wiff
5	WCGr2_U4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_U4.wiff
6	WCGr2_5_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	_5_1.wiff
7	WCGr2_5_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	_5_2.wiff
8	WCGr2_5_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_3.wiff
9	WCGr2_5_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5_4.wiff
10	WCGr2_UF_1			1 TripleTOF 6600	87,00 minute	WCGr2_UF_1.wiff
11	WCGr2_UF_2			1 TripleTOF 6600	87 minute	WCGr2_UF_2.wiff
12	WCGr2_UF_3			1 TripleTOF 6600	87,00 minute	WCGr2_UF_3.wiff
13	WCGr2_UF_4			1 TripleTOF 6600	87,00 minute	WCGr2_UF_4.wiff
14	WCGr2_5F_1			1 TripleTOF 6600	87,00 minute	WCGr2_5F_1.wiff
15	WCGr2_5F_2			1 TripleTOF 6600	minute	WCGr2_5F_2.wiff
16	WCGr2_5F_3			1 TripleTOF 6600	minute	WCGr2_5F_3.wiff
17	WCGr2_5F_4			1 TripleTOF 6600	minute	WCGr2_5F_4.wiff
18	WCGr1_U1			1 TripleTOF 6600	minute	WCGr1_U1.wiff
19	WCGr1_U2			1 TripleTOF 6600	minute	WCGr1_U2.wiff
20	WCGr1_U3			1 TripleTOF 6600	87,00 minute	WCGr1_U3.wiff
21	WCGr1_U4			1 TripleTOF	87,00 minute	WCGr1_U4.wiff
22	WCGr1_5_1			TripleTOF 6600	87,00 minute	WCGr1_5_1.wiff
23	WCGr1_5_2			1 TripleTOF 6600	87,00 minute	WCGr1_5_2.wiff
24	WCGr1_5_3			1 TripleTOF 6600	87,00 minute	WCGr1_5_3.wiff
25	WCGr1_5_4			1 TripleTOF 6600	87,00 minute	WCGr1_5_4.wiff
26	WCGr1_UF_1			1 TripleTOF 6600	87,00 minute	WCGr1_UF_1.wiff
27	WCGr1_UF_2			1 TripleTOF 6600	87,00 minute	WCGr1_UF_2.wiff

WCGr2_5_1.wiff	WIFF File
WCGr2_5_2.wiff	WIFF File
WCGr2_5_3.wiff	WIFF File
WCGr2_5_4.wiff	WIFF File
WCGr2_U1.wiff	WIFF File
WCGr2_U2.wiff	WIFF File
WCGr2_U3.wiff	WIFF File
WCGr2_U4.wiff	WIFF File
WCGr2_UF_1.wiff	WIFF File
WCGr2_UF_2.wiff	WIFF File
WCGr2_UF_3.wiff	WIFF File
WCGr2_UF_4.wiff	WIFF File

Raw data

Data PLANT
CEPLAS
CC BY

Swate

Annotation by flattening the knowledge graph

- Low-friction metadata annotation
- Familiar spreadsheet, row/column-based environment

Annotation principle

- Low-friction metadata annotation
- Familiar spreadsheet, row/column-based environment

Adding new building blocks (columns)

Annotation Building Block types

- Source Name (Input)
- Protocol Columns
  - Protocol Type, Protocol Ref
- Characteristic
- Parameter
- Factor
- Component
- Output Columns
  - Sample Name, Raw Data File, Derived Data File

Let's take a detour on Annotation Principles | slides

Ontology term search

Enable **related term directed search** to directly fill cells with child terms

Fill your table with ontology terms

Hierarchical combination of ontologies

isa.study.xlsx or isa.assay.xlsx

Swate templates

Checklists and Templates

=

Metadata standards or repository requirements can be represented as templates

Realization of lab-specific metadata templates

Directly import templates via Swate

- DataPLANT curated
- Community templates

Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
  github: https://github.com/brilator
  orcid: https://orcid.org/0000-0001-9021-3197
- name: Martin Kuhl
  github: https://github.com/Martin-Kuhl
  orcid: https://orcid.org/0000-0002-8493-1077

Code Notebooks and IDEs

- Visual Studio Code: https://code.visualstudio.com/
- RMarkdown: http://rmarkdown.rstudio.com
- Jupyter Notebooks: https://jupyter.org/

- 🟢 Interactive (good start for non-coders)
- 🟢 Document and comment code
- 🟢 Often offer integrated version control (e.g. git plugin)
- 🟠 Executable code + "result preview"
- 🟠 Reproducibility (package / library dependencies)

Computational workflow languages

- Describe analysis workflows and tools
- Make them portable and scalable
- Across a variety of environments (software and hardware)

https://www.commonwl.org

https://www.nextflow.io

https://nf-co.re/

https://snakemake.github.io

https://galaxyproject.eu/

Hands-on part 2: ARCitect (and Swate)

Follow the **Start Here** guide in the DataPLANT knowledge base.

Until "Data analysis"

Q&A and Wrap-up Day1

Preparation for next day

- Please try to prepare your own ARC
- Please install SWATE

Resources

DataPLANT (nfdi4plants)

Website: https://nfdi4plants.org/

Knowledge Base: https://nfdi4plants.org/nfdi4plants.knowledgebase/

DataHUB: https://git.nfdi4plants.org

GitHub: https://github.com/nfdi4plants

HelpDesk: https://helpdesk.nfdi4plants.org

You can help us by raising issues, bugs, ideas...

Overview of Institutional services at UoC and HHU

UoC

- C3RDM: https://fdm.uni-koeln.de/en/home
- Data storage and sharing: https://rrzk.uni-koeln.de/daten-speichern-teilen
- HPC: https://rrzk.uni-koeln.de/hpc-projekte
- service overview: https://fdm.uni-koeln.de/en/rdm-services/service-catalogue

HHU

- RDM Competence Center: https://www.fdm.hhu.de
- Support for research including HPC: https://www.zim.hhu.de/servicekatalog/forschungsunterstuetzung
- Processing & storing data: https://www.zim.hhu.de/servicekatalog/rechnen-und-speichern

Five-Finger-Feedback

"I did not like at all...!"

"This idea or advice was good: ..."

"I'm happy with ..."

"... was too short."

"I really liked ..."

CEPLAS
Cluster of Excellence on Plant Sciences

Good Data Management Practices

part of M4468 – Plant developmental genetics, evolution and biostatistics in the CEPLAS research program

November 12th, 2024

Vittorio Tracanna, Dominik Brilhaus
CEPLAS Data

House-keeping

Pad: https://pad.hhu.de/0NdPKO5LQ5CHBRN2iuG91Q

Points to discuss from and since day 1

Data Storage and Versioning

Data stores

Institute server

University server

Local hard disks

Cloud services

sciebo
dieCampuscloud

Chat

Your data

Wiki, Project management

Mail

eLabFTW

labfolder

Electronic lab notebooks

Git repositories

Data PLANT
CEPLAS
CC BY

Backup vs. Archive

	Backup	Archive
Storage type	Short-, mid-term	Long-term
Purpose	Disaster recovery	Long-term storage, compliance
Reason	Duplication	Migration
Usage	Work in progress	Cold, Unused data
Changes	Short-term updates	No updates
Trend	Cyclic, Replacement	Growing
Latency	Short/Costly	High/Cheaper

3-2-1 backup rule

3 copies of data

2 storage media

1 copy off-site

Version control and track changes

It’s good practice to document:

- What was changed?
- Who is responsible?
- When did it happen?
- Why the changes?

Types of Version Control

- by file name (_v1, _v2)
- cloud services
  - dropbox, icloud, gdrive
- distributed version control system
  - e.g. Git

Which files need to be "versioned"? 📝

- paper manuscript (.docx)
- single-cell RNASeq reads (.fastq.gz)
- spread sheet with photometer measurements (.xlsx)
- calendar invitation (.ical)
- photo of SDS-PAGE (.jpeg)
- excel workbook with calculations (.xlsx)
- presentation for a conference (.pdf)
- data analysis script (.py)

Concept of Git and git-based platforms

Cloud Services

✓ Documents

✓ Small data

✓ Presentations

X Code

X Data analytical projects

X Big (“raw”) data

Git and git platforms

~ Documents

✓ Small data

~ Presentations

✓✓ Code

✓✓ Data analytical projects

~ Big (“raw”) data

Why git? ≈> Why code?

- Save time
- Avoid doing repetitive tasks “by hand”
- Reuse scripts, analyses, pipelines
- Reproduce results

A simple example: RNASeq project

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

A simple example: RNASeq project

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

Input(s)

A simple example: RNASeq project

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

Input(s) → Computation

A simple example: RNASeq project

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

A simple example: RNASeq project

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
  01_input_data/barley_sample02.fastq.gz \
  01_input_data/barley_sample03.fastq.gz -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
  -o 03_results/RNASeq_counts.txt
```

*

Input(s) → Computation → Output

** Disclaimer: this is not a good example for reusable code*

Take snapshots of your code work…

(… as long as it works)

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
  01_input_data/barley_sample02.fastq.gz \
  01_input_data/barley_sample03.fastq.gz -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
  -o 03_results/RNASeq_counts.txt
```

Input(s) → Computation → Output ✓

Take snapshots of your code work…

(… as long as it works)

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
  01_input_data/barley_sample02.fastq.gz \
  01_input_data/barley_sample03.fastq.gz -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
  -o 03_results/RNASeq_counts.txt
```

Scenario 1: More data

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
```

```
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
  01_input_data/barley_sample02.fastq.gz \
  01_input_data/barley_sample03.fastq.gz -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
  -o 03_results/RNASeq_counts.txt
```

Scenario 1: More data

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
```

```bash
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
  01_input_data/barley_sample02.fastq.gz \
  01_input_data/barley_sample03.fastq.gz -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
  -o 03_results/RNASeq_counts.txt
```

```bash
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
  01_input_data/barley_sample02.fastq.gz \
  01_input_data/barley_sample03.fastq.gz \
  01_input_data/barley_sample04.fastq.gz \
  01_input_data/barley_sample05.fastq.gz \
  01_input_data/barley_sample06.fastq.gz -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
  -o 03_results/RNASeq_counts.txt
```

Scenario 1: More data

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
├── 02_analysis
│   ├── RNASeq_quant.sh
│   ├── RNASeq_quant_first_samples.sh
│   ├── RNASeq_quant_including_all_samples.sh
│   ├── RNASeq_quant_including_all_samples_updated.sh
│   └── RNASeq_quant_including_all_samples_updated_v2.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
  01_input_data/barley_sample02.fastq.gz \
  01_input_data/barley_sample03.fastq.gz -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
  -o 03_results/RNASeq_counts.txt
```

```
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
  01_input_data/barley_sample02.fastq.gz \
  01_input_data/barley_sample03.fastq.gz \
  01_input_data/barley_sample04.fastq.gz \
  01_input_data/barley_sample05.fastq.gz \
  01_input_data/barley_sample06.fastq.gz -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
  -o 03_results/RNASeq_counts.txt
```

Let git track changes and keep things clean

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

Project_barley > 02_analysis > $ RNASeq_quant.sh

```
1   # RNASeq quantification
2
3   read_trimming -i 01_input_data/barley_sample01.fastq.gz \
4     01_input_data/barley_sample02.fastq.gz \
5-    01_input_data/barley_sample03.fastq.gz -o trimmed_reads/
6
7   rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
8     -o 03_results/RNASeq_counts.txt
9
10
11
```

“version 1”

```
1   # RNASeq quantification
2
3   read_trimming -i 01_input_data/barley_sample01.fastq.gz \
4     01_input_data/barley_sample02.fastq.gz \
5+    01_input_data/barley_sample03.fastq.gz \
6+    01_input_data/barley_sample04.fastq.gz \
7+    01_input_data/barley_sample05.fastq.gz \
8+    01_input_data/barley_sample06.fastq.gz -o trimmed_reads/
9
10  rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
11    -o 03_results/RNASeq_counts.txt
12
13
14
```

“version 2”

Scenario 2: Pipeline breaks

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

Revert to snapshot

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

"version xy"

Scenario 3: New project, same type of data and analysis

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt

Project_maize
├── 01_input_data
│   ├── maize_genome_ref.fasta
│   ├── maize_sample01.fastq.gz
│   ├── maize_sample02.fastq.gz
│   └── maize_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

Scenario 3: New project, same type of data and analysis

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt

Project_maize
├── 01_input_data
│   ├── maize_genome_ref.fasta
│   ├── maize_sample01.fastq.gz
│   ├── maize_sample02.fastq.gz
│   └── maize_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
  01_input_data/barley_sample02.fastq.gz \
  01_input_data/barley_sample03.fastq.gz -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
  -o 03_results/RNASeq_counts.txt
```

```
# RNASeq quantification

read_trimming -i 01_input_data/maize_sample01.fastq.gz \
  01_input_data/maize_sample02.fastq.gz 01_input_data/maize_sample03.fastq.gz \
  -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/ maize_genome_ref.fasta \
  -o 03_results/RNASeq_counts.txt
```

Re-use code

Project_barley
Input(s)
Computation
Output
Project_maize
Input(s)
Computation
Output

Re-use code

```
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
  01_input_data/barley_sample02.fastq.gz \
  01_input_data/barley_sample03.fastq.gz \
  -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
  -o 03_results/RNASeq_counts.txt
```

“version barley”

```
# RNASeq quantification

read_trimming -i 01_input_data/maize_sample01.fastq.gz \
  01_input_data/maize_sample02.fastq.gz 01_input_data/maize_sample03.fastq.gz

  -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/ maize_genome_ref.fasta -o 03_results/RNASeq_counts.txt
```

“version maize”

Re-use code – People have done this

Re-use code – People have done this

Re-use code – Link and contribute

Git: summary

- Version control system
- Git “repository” = a central data package (directory)
- Allows to track changes to any file in the repository
  - **What** was changed
  - **When** was it changed
  - By **whom** was it changed
  - **Why** was it changed?

GitHub and GitLab

- A well-documented cloud environment
- Active syncing
- Not automatically synced
- Non-automated version control
- You have the control what changes to track and what to sync
- Time machine to go back to older versions

GitHub and Gitlab team projects

Simplifies concurrent work & merging changes

- Online service to host our projects
- Share code with other developers
- Others can download our projects, work on and contribute to them
- They can upload their changes and merge them with the main project

Cloud vs. Git

	Cloud services	Git / GitHub / GitLab
Track changes	✓ Documents ✓ Small data ✓ Presentations	✓ Code ✓ Data analytical projects
Collaboration		issue tracker, tracked contribution
Versioning	Automated	Well-documented (commit history)
Syncing	Automated	Active / controlled by user
Access	Oftentimes only within organization / institution	Easily collaborate across institutions
Data security	Private / commercial	GitLab: on-premise and custom solutions

DataPLANT DataHUB

ARC builds on standards + Git

git

The DataPLANT DataHUB – a GitLab *Plus*

=

GitLab

+

DataHUB
Storage & Backup

Versioning

Collaboration &
Access Management

Multiple
Contribution

Reference
& Reuse
Multiple
Contribution

DataHUB
Publish
Receive a DOI
GEO
ENA
MetaboLights
PRIDE Archive
Data PLANT
INVENIO RDM

FINDABLE
ACCESSIBLE
INTEROPERABLE
REUSABLE
Metadata Templates
Controlled Vocabularies
Standards
Collaboration & Access Management
Versioning
Storage & Backup
Multiple Contribution
Reference & Reuse
Publish
Receive a DOI
GEO
ENA
MetaboLights
PRIDE Archive
Data PLANT
INVENIO RDM

Mutable data life cycle

Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

Project management

Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

📝 Structure your project as an ARC 📝

1. Follow the Start Here guide to build an ARC for your own project
2. Go back to the shared folder prepared before class
3. Add data from your project to your ARC

Data publication and repositories

Peristent Identifiers (PIDs)

Spot the PIDs

doi:10.1093/plcell/koab243

THE PLANT CELL

THE PLANT CELL 2021: 00: 1–24

Research Article

Interactions between SQUAMOSA and SHORT VEGETATIVE PHASE MADS-box proteins regulate meristem transitions during wheat spike development

Kun Li [1,2,†] Juan M. Debernardi [1,2,*,†,‡] Chengxia Li [1,2] Huiqiong Lin [1,2] Chaozhong Zhang [1] Judy Jernstedt [1] Maria von Korff [3,4] Jinshun Zhong [3] and Jorge Dubcovsky [1,2,*,‡]

1 Department of Plant Sciences, University of California, Davis, California 95616, USA
2 Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA
3 Institute for Plant Genetics, Heinrich Heine University, Düsseldorf 40225, Germany
4 Cluster of Excellence on Plant Sciences "SMART Plants for Tomorrow's Needs", Heinrich Heine University, Düsseldorf 40225, Germany

*Author for correspondence: jmdebernardi@ucdavis.edu (J.M.D), jdubcovsky@ucdavis.edu (J.D.)
†These authors contributed equally (K.L and J.M.D.)
‡Senior authors

C.L., J.M.D., and J.D. designed the research. K.L. performed most of the experimental work. J.M.D., C.L., H.L., and C.Z. performed research, J.J. contributed the SEM images, M.V.K. and J.Z. contributed *in situ* hybridizations, C.L., H.L., J.M.D., K.L., and J.D. analyzed the data. C.L., J.M.D., K.L., H.L., and J.D. wrote the article.
The authors responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plcell) are: Jorge Dubcovsky (jdubcovsky@ucdavis.edu) and Juan Manuel Debernardi (jmdebernardi@ucdavis.edu).

Abstract

Inflorescence architecture is an important determinant of crop productivity. The number of spikelets produced by the wheat inflorescence meristem (IM) before its transition to a terminal spikelet (TS) influences the maximum number of grains per spike. Wheat MADS-box genes *VERNALIZATION 1* (*VRN1*) and *FRUITFULL 2* (*FUL2*) (in the *SQUAMOSA*-clade) are essential to promote the transition from IM to TS and for spikelet development. Here we show that SQUAMOSA genes contribute to

Downloaded from https://academic.oup.com/plcell/advance-article/doi/10.1093/plcell/koab243/6415951 b

Data PLANT
CEPLAS
CC BY

https://doi.org/10.1093/plcell/koab243

Globally unique, stable, persistent identifiers (PIDs)

- Long-term findability
- Make data, digital objects, people, … uniquely identifiable
- Diminish “dead links”
- Cope with name changes

Open
Researcher and Contributor ID
https://orcid.org/

Digital
Object Identifier
https://www.doi.org

Research
Resource Identifiers

https:// www.rrids.org

ePIC consortium

https ://www.pidconsortium. net

Research
Organization Registry

https://ror.org

Global
Research Identifier Database

https://grid.ac

Properties of a PID

Ideally, PIDs are

- Stable and permanent
- Location-independent
- Globally unique and valid
- Addressable (citable)
- Clickable (resolvable)

Adapted from https://www.ebi.ac.uk/rdf/documentation/good_practice_uri/

Additional resources

- https://www.doi.org
- https://www.orcid.org
- https://pidservices.org/
- https://datacite.org
- https://www.project-freya.eu/en

Institutional publication guidelines

HHU Düsseldorf recommends use of ORCID and other PIDs

Publikationsrichtlinie der Heinrich-Heine-Universität Düsseldorf vom 09.11.2023:

https://www.hhu.de/die-hhu/kontakt-und-services/zentrale-und-amtliche-bekanntmachungen/nr-34-2023

Domain-specific data repositories

Good

- Assign PIDs / DOIs
- Long-term accessible
- Data type specific
- Apply metadata standards
- Usually recommended / required by journals
- Mostly accepted by the community

Intermediate

- User-friendliness
- Different metadata schema
- Complex and versatile submission routines

Domain-specific data repositories

Repository	Description	Biological data domain
EBI-ENA	European Nucleotide Archive	genome / transcriptome sequences
EBI-ArrayExpress	Archive of Functional Genomics Data	transcriptome
EBI-MetaboLights	Database of Metabolomics	metabolome
EBI-PRIDE	PRoteomics IDEntifications Database	proteome
EBI-BioImage Archive	Stores and distributes biological images	imaging, microscopy
e!DAL-PGP	Plant Genomics & Phenomics Research Data Repository	phenome

Choosing a data repository

Domain-specific >> Generic >> Institutional

Find repositories at:

- https://www.re3data.org
- https://fairsharing.org

Generic data repositories

Good

- Allow publication of any kind of data Assign PIDs / DOIs
- Long-term accessible
- Very simple to use

Intermediate

- Only generic / high-level metadata schema
- Limited reusability

https://zenodo.org

https://datadryad.org/

https://figshare.com

From ARC to repositories

Moving from paper to data publications

Publish your ARC, get a DOI

Publish your ARC with a few clicks

Validation

adapted from Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

Validate & publish

Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

Receive a DOI

Published September 7, 2023 | Version v1

Dataset Open

Systems-wide investigation of responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii.

Zhang, Ningning[1]; Mattoon, Erin[1]; McHargue, Will[1]; Venn, Benedikt[2]; Zimmer, David[2]; Pecani, Kresti[3]; Jeong, Jooyeon[1]; Anderson, Cheyenne[1]; Chen, Chen[4]; Berry, Jeffrey[1]; Xia, Ming[1]; Tzeng, Shin-Cheng[1]; Becker, Eric[1]; Pazouki, Leila[1]; Evans, Bradley[1]; Cross, Fred[3]; Cheng, Jianlin[4]; Czymmek, Kirk[1]; Schroda, Michael[5]; Mühlhaus, Timo[2]; Zhang, Ru[1]

Show affiliations

1

Citation

Style APA

Zhang, N., Mattoon, E., McHargue, W., Venn, B., Zimmer, D., Pecani, K., Jeong, J., Anderson, C., Chen, C., Berry, J., Xia, M., Tzeng, S.-C., Becker, E., Pazouki, L., Evans, B., Cross, F., Cheng, J., Czymmek, K., Schroda, M., … Zhang, R. (2023). Systems-wide investigation of responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii. [Data set]. DataPLANT. https://doi.org/10.60534/9e5jx-75d83

Description

hosted on: https://git.nfdi4plants.org/projects/122

Files

arc-summary.md

[Data set] Systems-wide investigation of responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii.

File contents:

- root
  - isa.investigation.xlsx
  - README.md
  - runs

Edit

New version

Share

Versions

Version v1 Sep 7, 2023

10.60534/9e5jx-75d83

2

Details

DOI

DOI 10.60534/9e5jx-75d83

Resource type

Dataset

Publisher

DataPLANT

Export

JSON Export

Legal aspects of RDM

Different laws touched by RDM

Hartmann, Thomas. (2019). Rechtsfragen: Institutioneller Rahmen und Handlungsoptionen für universitäres FDM. Zenodo. https://doi.org/10.5281/zenodo.2654306

Open Access (OA) categories

- Gold: Published in an open-access journal that is indexed by the DOAJ.
- Green: Toll-access on the publisher page, but there is a free copy in an OA repository.
- Hybrid: Free under an open license in a toll-access journal.
- Bronze: Free to read on the publisher page, but without a clearly identifiable license.
- Closed: All other articles, including those shared only on an Academic Social Network or in Sci-Hub.

Piwowar H et al. (2018), PeerJ https://doi.org/10.7717/peerj.4375

Open Science is more than Open Access

Okafor et al. (2022) https://doi.org/10.3389/frma.2022.855198, Figure 1

Creative commons

Check out: https://creativecommons.org/about/cclicenses/

adapted from https://wiki.creativecommons.org../images/0/01/6licenses-folded.pdf

Data protection

GDPR: General Data Protection Regulation

DS-GVO (german): Datenschutz-Grundverordnung

Use of biological materials

- MTA: material transfer agreement
- Nagoya Protocol: https://www.cbd.int/abs/about/
- DSI: Digital sequence information

FAIR and CARE

https://www.gida-global.org/care

CARE principles

https://datascience.codata.org/articles/10.5334/dsj-2020-043/

Research Data policies

Abbildung 2: FD-Policy-Schema: Die sechs Kategorien einer FD-Policy mit ihren inhaltlichen Bestandteilen

Hiemenz, Bea & Kuberek, Monika (2018) http://dx.doi.org/10.14279/depositonce-7521

CEPLAS relevant data handling guidelines & policies

- Deutsche Forschungsgemeinschaft (2015): DFG Guidelines on the Handling of Research Data
- Amtliche Mitteilungen der Universität zu Köln AM 07/2018: Leitlinie zum Umgang mit Forschungsdaten
- Amtliche Bekanntmachung der Heinrich-Heine-Universität Nr. 43/2022: Forschungsdaten-Richtlinie
- Leitlinie zum Umgang mit Forschungsdaten im Forschungszentrum Jülich 05/2019
- Senat der Max-Planck-Gesellschaft (2009): Regeln zur Sicherung guter wissenschaftlicher Praxis

The Data Management Plan (DMP)

- Covers the full research data lifecycle
- Frequently updated as your project develops
- Required to different extents by funding agencies (e.g. DFG, Horizon Europe, BMBF, BMEL, ...)

DMP tools

- Data Stewardship Wizard https://ds-wizard.org/
- RDMO https://rdmorganiser.github.io/ (e.g. https://rdmo.hhu.de)
- Dataplan: https://dmpg.nfdi4plants.org

> Check out the Elixir RDMkit for more

📝 Share your ARC 📝

1. Follow the next slides to learn how to share your ARC with a colleague
2. Go to your colleague's ARC and try to understand it

📝 Understand your colleague's ARC 📝

1. Go to your colleague's ARC
2. Try to identify one experiment that led to results (e.g. a figure in the thesis)
3. What are the samples (e.g. plants, bacteria)?
4. How were the samples prepared (~ materials)?
5. How was the experiment performed (~ methods)?
6. What is the raw data (~ results)?
7. How was the data analyzed (~ computational methods, statistics)?
8. Collect all of the above in a `README_<YourArbitraryParticipantID>.md` in the same folder.

Assignment

Participant	looks at ARC of
Participant02	Participant01
Participant03	Participant02
…	…
Participant n	Participant n-1

Electronic Lab Notebooks (ELNs)

ELN ≈ Digital Protocol Editors

- Documenting daily lab routine
- Lab methods & protocols
- Lab inventory (biologicals, chemicals, instruments)
- Local sharing & collaboration
- Backup (locally)

✅ ELNs help to digitalize research documentation

Are ELNs FAIR?

FAIR indicator*	ELN
Findable	
F1. (Meta)data are assigned a globally unique and persistent identifier.	
F2. Data are described with rich metadata (defined by R1 below).	
F3. Metadata clearly and explicitly include the identifier of the data they describe.	
F4. (Meta)data are registered or indexed in a searchable resource.	
Accessible	
A1. (Meta)data are retrievable by their identifier using a standardised communications protocol	
A1.1 The protocol is open, free, and universally implementable	

*FAIR classified by https://www.go-fair.org/fair-principles/

Findable

FAIR indicator*	elabFTW
F1. (Meta)data are assigned a globally unique and persistent identifier.	🟠 / 🔴
F2. Data are described with rich metadata (defined by R1 below).	🔴
F3. Metadata clearly and explicitly include the identifier of the data they describe.	🟠 / 🔴
F4. (Meta)data are registered or indexed in a searchable resource.	🔴

*FAIR classified by https://www.go-fair.org/fair-principles/

Accessible

FAIR indicator*	elabFTW
A1. (Meta)data are retrievable by their identifier using a standardised communications protocol	🔴
A1.1 The protocol is open, free, and universally implementable	🟢 / 🟠
A1.2 The protocol allows for an authentication and authorisation procedure, where necessary	?
A2. Metadata are accessible, even when the data are no longer available	?

*FAIR classified by https://www.go-fair.org/fair-principles/

Interoperable

FAIR indicator*	elabFTW
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	●
I2. (Meta)data use vocabularies that follow FAIR principles.	●
I3. (Meta)data include qualified references to other (meta)data.	●

*FAIR classified by https://www.go-fair.org/fair-principles/

Reusable

FAIR indicator*	elabFTW
R1. (Meta)data are richly described with a plurality of accurate and relevant attributes	●
R1.1. (Meta)data are released with a clear and accessible data usage license	●
R1.2. (Meta)data are associated with detailed provenance	●
R1.3. (Meta)data meet domain-relevant community standards	●

*FAIR classified by https://www.go-fair.org/fair-principles/

Contributors

If not referenced otherwise, figures and slides presented here were created by members of DataPLANT (https://nfdi4plants.org). Additional slides were contributed by

- name: Dominik Brilhaus
  github: https://github.com/brilator
  orcid: https://orcid.org/0000-0001-9021-3197
- name: Cristina Martins Rodrigues
  github: https://github.com/CMR248
  orcid: https://orcid.org/0000-0002-4849-1537
- name: Hajira Jabeen
  github: https://github.com/HajiraJabeen
  orcid: https://orcid.org/0000-0003-1476-2121
- name: Kevin Frey
  github: https://github.com/Freymaurer
  orcid: https://orcid.org/0000-0002-8493-1077
- name: Sabrina Zander
  orcid: https://orcid.org/0009-0000-4569-6126
- name: Martin Kuhl