

plSA-tree

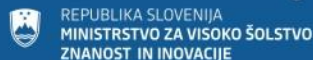
sistem za organizacijo projektnih raziskovalnih podatkov

dr. Marko PETEK, Nacionalni inštitut za biologijo

dr. Maja ZAGORŠČAK, Nacionalni inštitut za biologijo

prof. dr. Andrej BLEJEC, Nacionalni inštitut za biologijo

Ljubljana in online, Centralna tehniška knjižnica Univerze v Ljubljani,
Usposabljanje podatkovnih strokovnjakov, 18. – 21. 11. 2024







Kanza and Knight *BMC Research Notes* (2022) 15:20
<https://doi.org/10.1186/s13104-022-05908-5>

BMC Research Notes

COMMENTARY

Open Access

Behind every great research project is great data management



Samantha Kanza^{*†} and Nicola J. Knight[†]

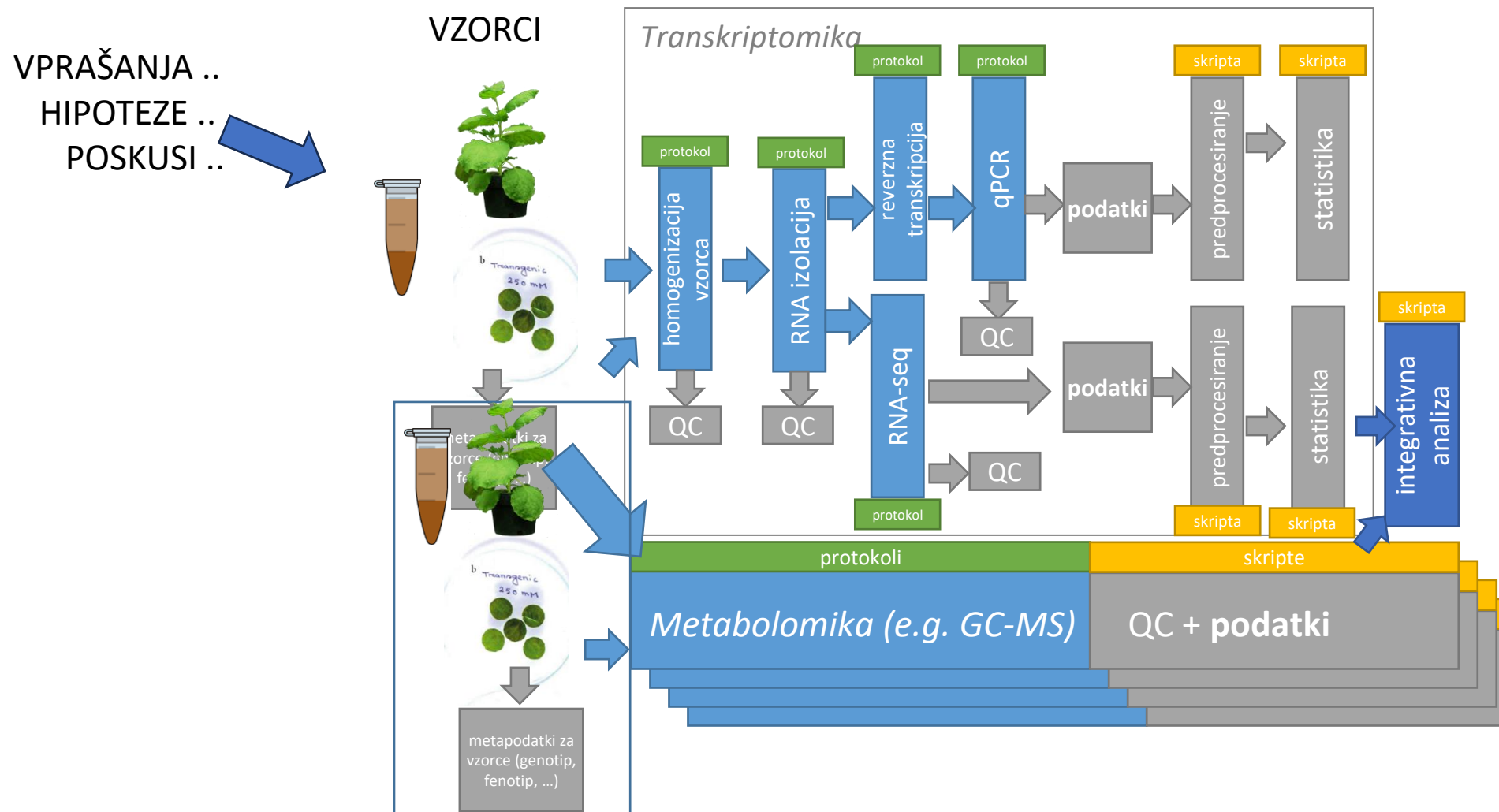
Abstract

Research data management (RDM) is the cornerstone of a successful research project, and yet it often remains an underappreciated art that gets overlooked in the hustle and bustle of everyday project management even when required by funding bodies. If researchers are to strive for reproducible science that adheres to the principles of FAIR, then they need to manage the data associated with their research projects effectively. It is imperative to plan your RDM strategies early on, and setup your project organisation before embarking on the work. There are several different factors to consider: data management plans, data organisation and storage, publishing and sharing your data, ensuring reproducibility and adhering to data standards. Additionally it is important to reflect upon the ethical implications that might need to be planned for, and adverse issues that may need a mitigation strategy. This short article discusses these different areas, noting some best practices and detailing how to incorporate these strategies into your work. Finally, the article ends with a set of top ten tips for effective research data management.

Keywords: Research data management, Data management plans, Data organisation, Data sharing, FAIR data, Data ethics, Reproducibility



Tipične raziskave v sistemski biologiji in biotehnologiji





(project/Investigation/Study/Assay)

- sistem za organizacijo podatkov iz raziskovalnih projektov
- razvili smo ga na Oddelku za biotehnologijo in sistemsko biologijo na NIB
- postavljen na osnovi izkušenj in potreb raziskovalne skupine OMIKE

scientific data

OPEN
ARTICLE

pISA-tree - a data management framework for life science research projects using a standardised directory tree

Marko Petek^{1,2}, Maja Zagorščak^{1,2}, Andrej Blejcek^{1,2}, Živa Ramšak³, Anna Coll⁴, Špela Baebler² & Kristina Gruden¹

We developed pISA-tree, a straightforward and flexible data management solution for organisation of life science project-associated research data and metadata. pISA-tree was initiated by end-user requirements thus its strong points are practicality and low maintenance cost. It enables on-the-fly creation of enriched directory tree structure (project/Investigation/Study/Assay) based on the ISA model, in a standardised manner via consecutive batch files. Templates-based metadata is generated in parallel at each level enabling guided submission of experiment metadata. pISA-tree is complemented by two R packages, *pisar* and *seekr*. *pisar* facilitates integration of pISA-tree datasets into bioinformatic pipelines and generation of ISA-Tab exports. *seekr* enables synchronisation with the FAIRDOMHub repository. Applicability of pISA-tree was demonstrated in several national and international multi-partner projects. The system thus supports findable, accessible, interoperable and reusable (FAIR) research and is in accordance with the Open Science initiative. Source code and documentation of pISA-tree are available at <https://github.com/NIB-SI/pISA-tree>.

Introduction

Compared to a decade ago, present-day life science projects generate vast amounts of data, that are frequently heterogeneous. Both researchers and funders are aware that published results often lack important information that would ensure their reproducibility. Raw data supporting researchers' conclusions are frequently either inaccessible, inadequately organised or not sufficiently annotated to allow straightforward reuse¹⁻³. To alleviate these issues, the FAIR data principles⁴ that advocate policies which ensure findability, accessibility, interoperability and reusability of the generated data were proposed. Thus, reporting, format and semantic standards are of utmost importance⁵. Minimum information metadata standards, providing guidelines for standardised reporting of experimental results, have been developed for a large number of wet-lab assays, e.g. plant phenotyping (MIAPPE⁶), quantitative PCR (MIQE⁷), gene expression microarrays (MIAME⁸), proteomics (MIAPE⁹), as well as dry-lab assays, e.g. quantitative modelling (MIRIAM¹⁰) and simulations in systems biology (MIASE¹¹). To organise and keep track of experimental metadata, the ISA framework¹² implemented the ISA model that captures experimental metadata on three hierarchical levels: Investigation, Study and Assay (abbreviated as ISA levels). Each ISA level contains files describing experimental goals, ontologies used, experimental design, protocols, and experimental conditions. ISA framework developers also provide ISA tools¹³, a software suite that, among others, contains a metadata editor and packages for integration with common programming languages¹⁴⁻¹⁶. ISA tools use a series of graphical interface input forms to generate metadata in the ISA-Tab or ISA-JSON format and use assay-specific templates covering a wide variety of methodologies and community standards. In parallel, based on minimum information standards, several public repositories¹⁷ have formulated metadata templates, that guide the researchers before submitting the data with publications. Besides methodology-specific repositories, there are general repositories that offer long-term research data and metadata storage following FAIR principles. One such service is Zenodo (www.zenodo.org), an open science repository developed by CERN that

¹Department of Biotechnology and Systems Biology, National Institute of Biology, Večna pot 111, SI-1000, Ljubljana, Slovenia. *These authors contributed equally: Marko Petek, Maja Zagorščak, Andrej Blejcek. ✉e-mail: marko.petek@nib.si, maja.zagorscak@nib.si, andrej.blejcek@nib.si

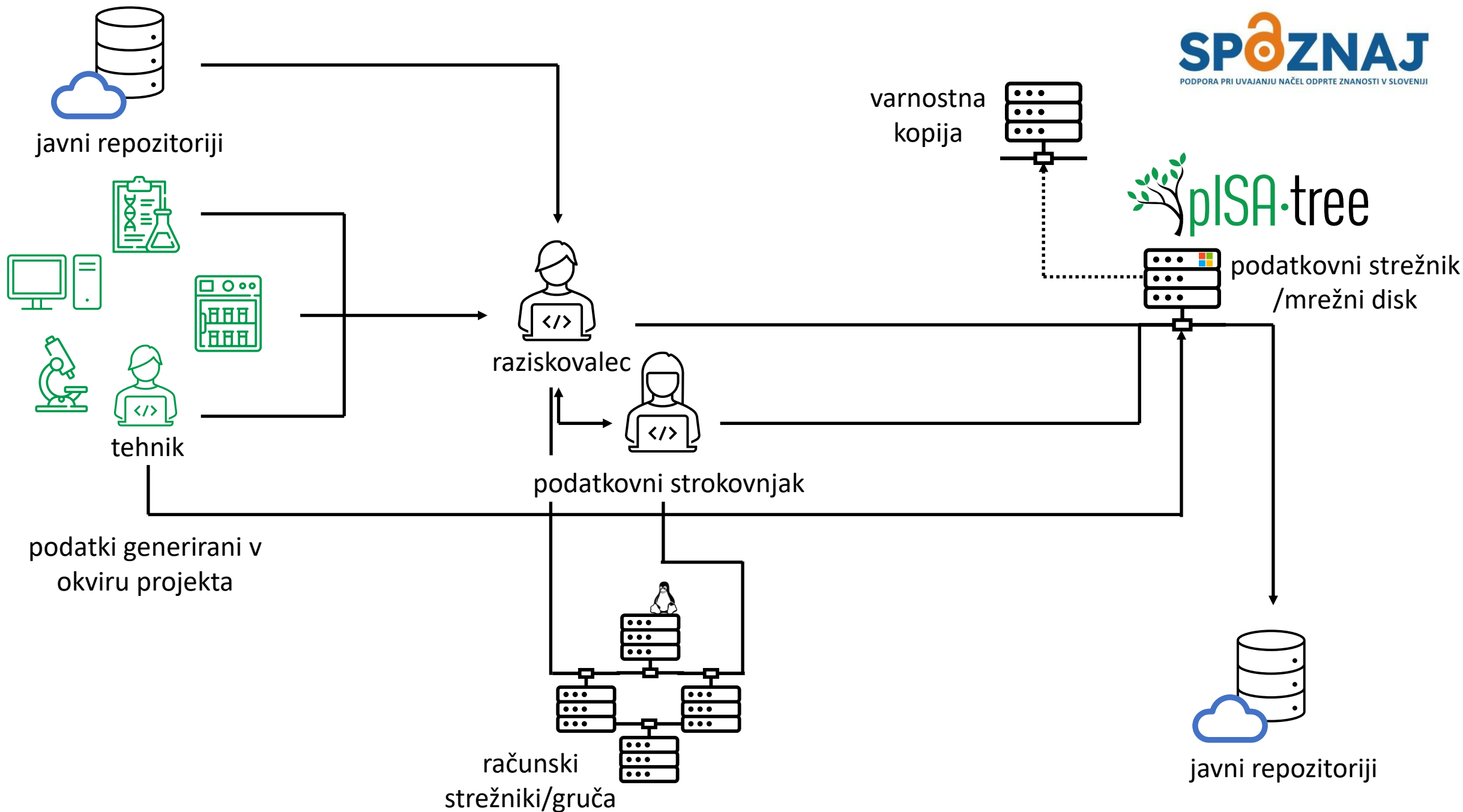


Glavne značilnosti:

- preprost
- osnovan na pravilih, vendar prilagodljiv
- sprotno beleženje metapodatkov
- bolj primeren za raziskave odprtega tipa, manj za diagnostiko
- omogoča izmenjavo med projektnimi partnerji iz drugih organizacij

Raziskovalni podatki:

- neobdelani podatki (meritve fluorescence, DNA zaporedja, mikroskopske fotografije, videoposnetki ...)
- metapodatki (podatki o vzorcih, QC, protokoli, skripte)
- vmesni in končni rezultati obdelav podatkov in modeliranja
- publikacije (članki, poročila, predstavitve, plakati ...)





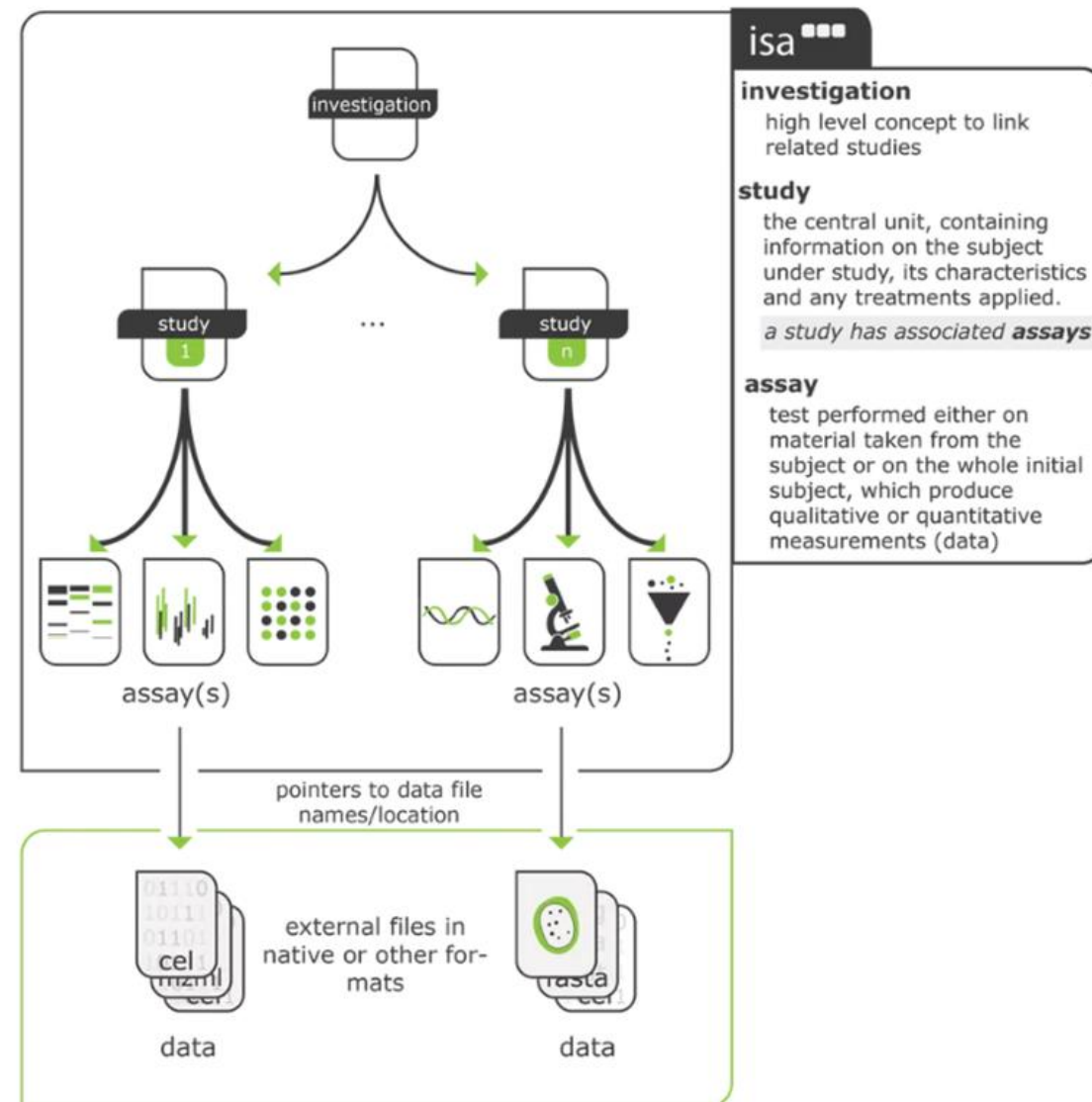
ISA model

Investigation == Raziskava

Study == Študija

Assay == Test

- gre za generičen konceptualni okvir organizacije podatkov in metapodatkov razvijalcev **isatools** (<https://isa-tools.org/>)
- standardna hierarhična struktura
- vsak nivo ima svoje metapodatkovne datoteke
- implementiran v formatih ISA-Tab, ISA-JSON

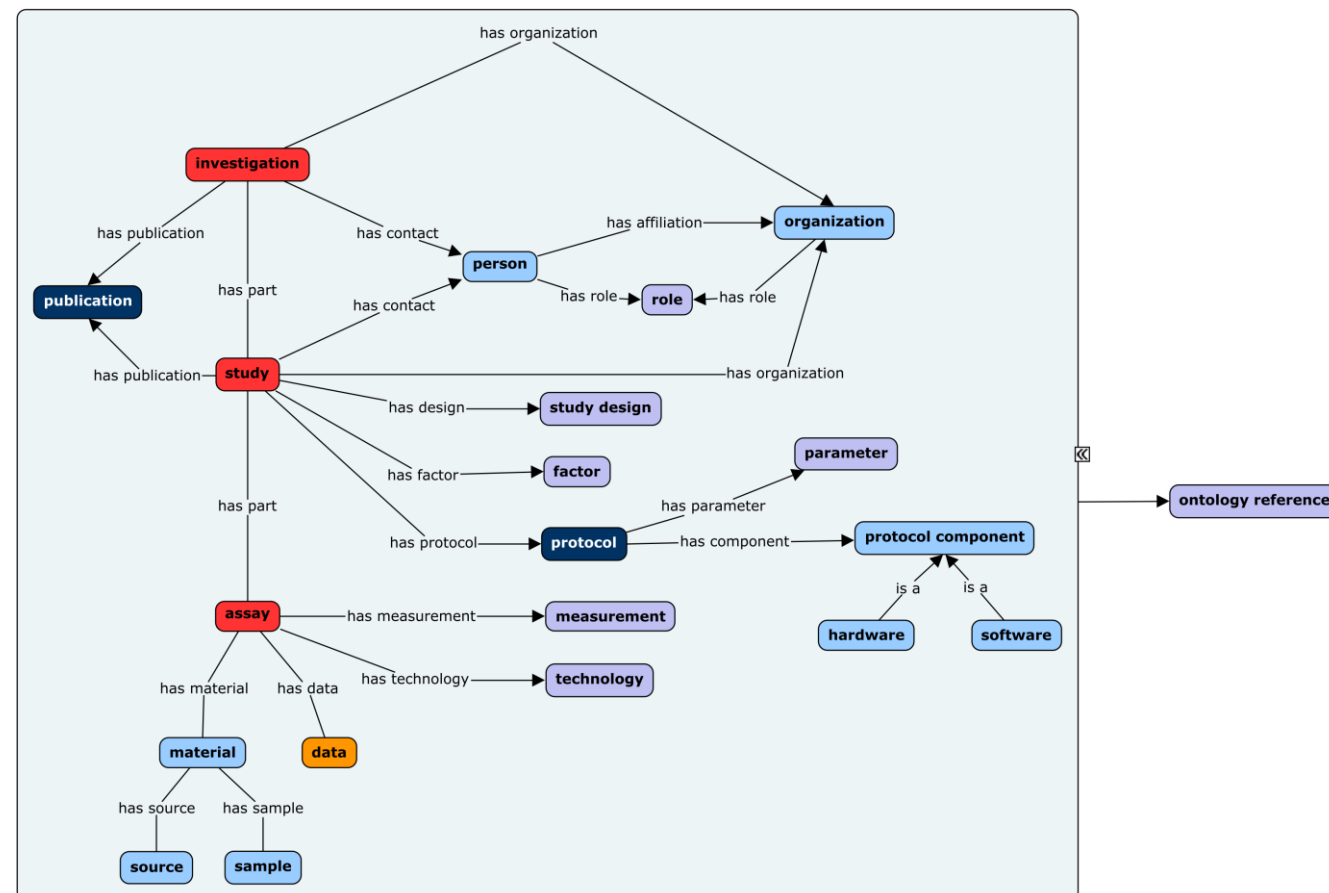




ISA model

<https://isa-specs.readthedocs.io/en/latest/isamodel.html>

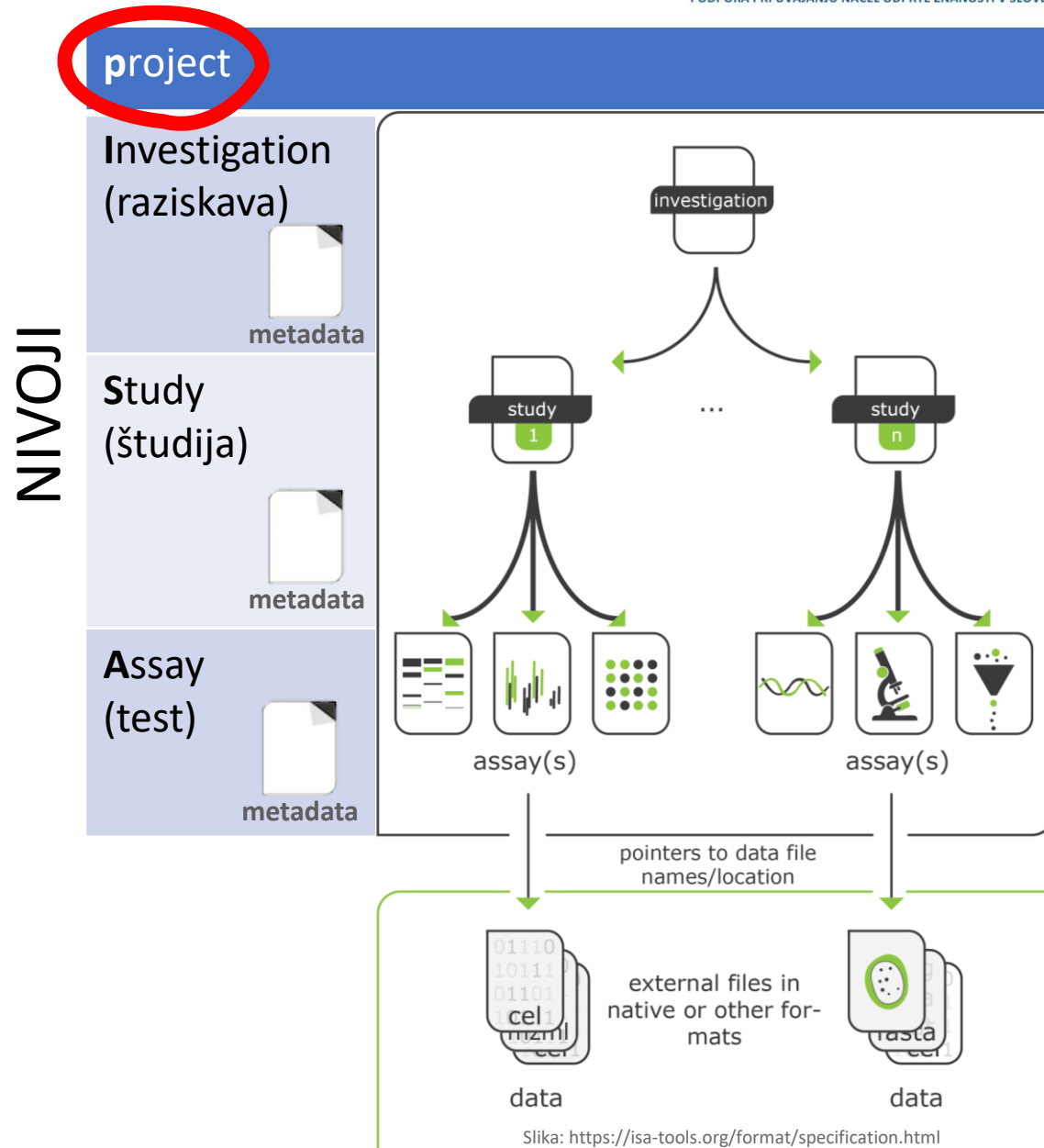
- isatools omogoča hierarhičen opis raziskovalnih projektov
- *generiranje nivojev je možno šele ko imamo rezultate testov*





Organizacija podatkov

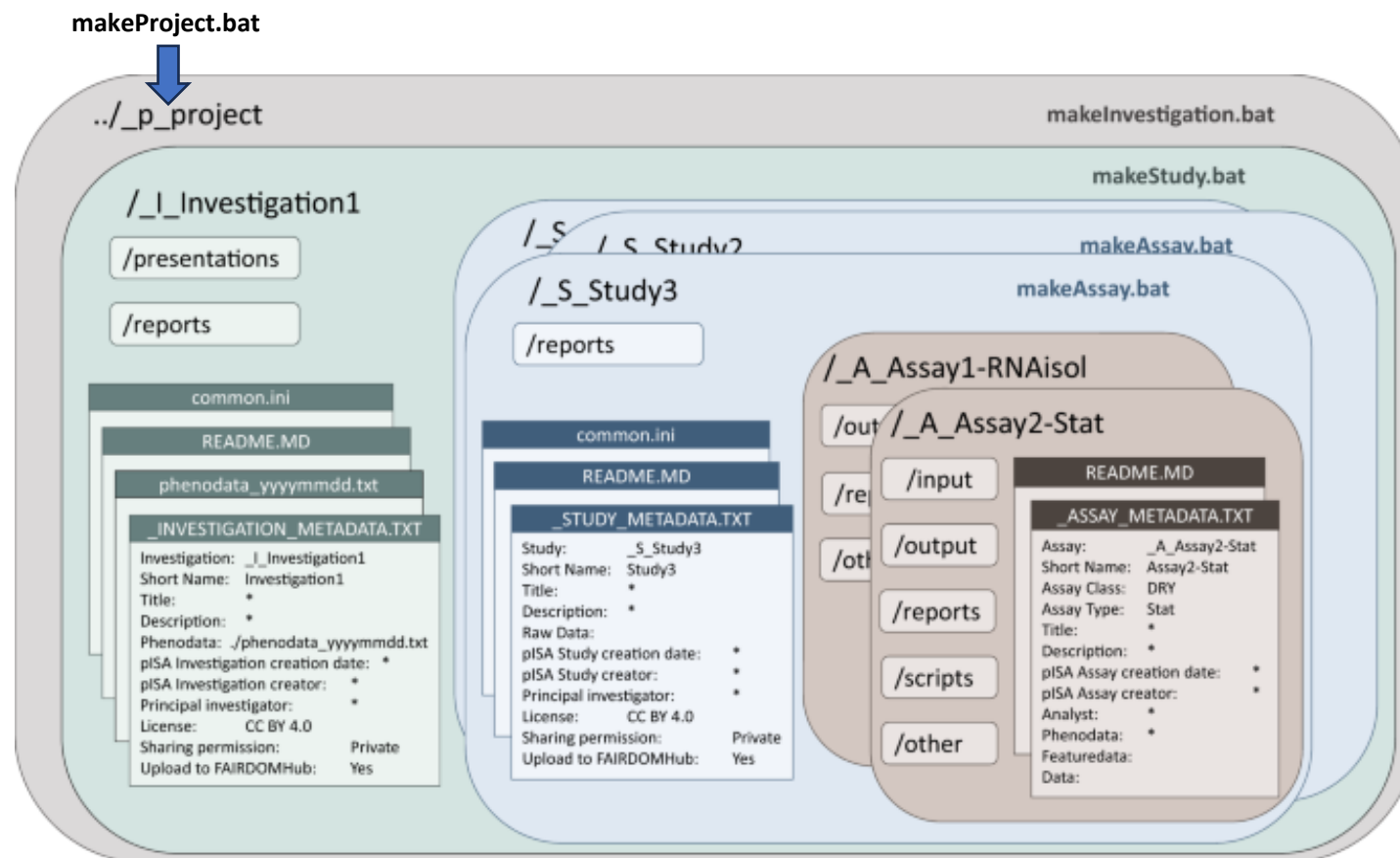
- *.bat skripte (delujejo v OS Windows)
 avtomatično ↓ generiramo
- drevesna struktura map ("directories") po nivojih ISA modela
- metapodatkovne *.txt datoteke v formatu kompatibilnim z ISA-Tab
- *sprotno generiranje nivojev z metapodatki (preden imamo rezultate testov) in kasnejše dopolnjevanje*





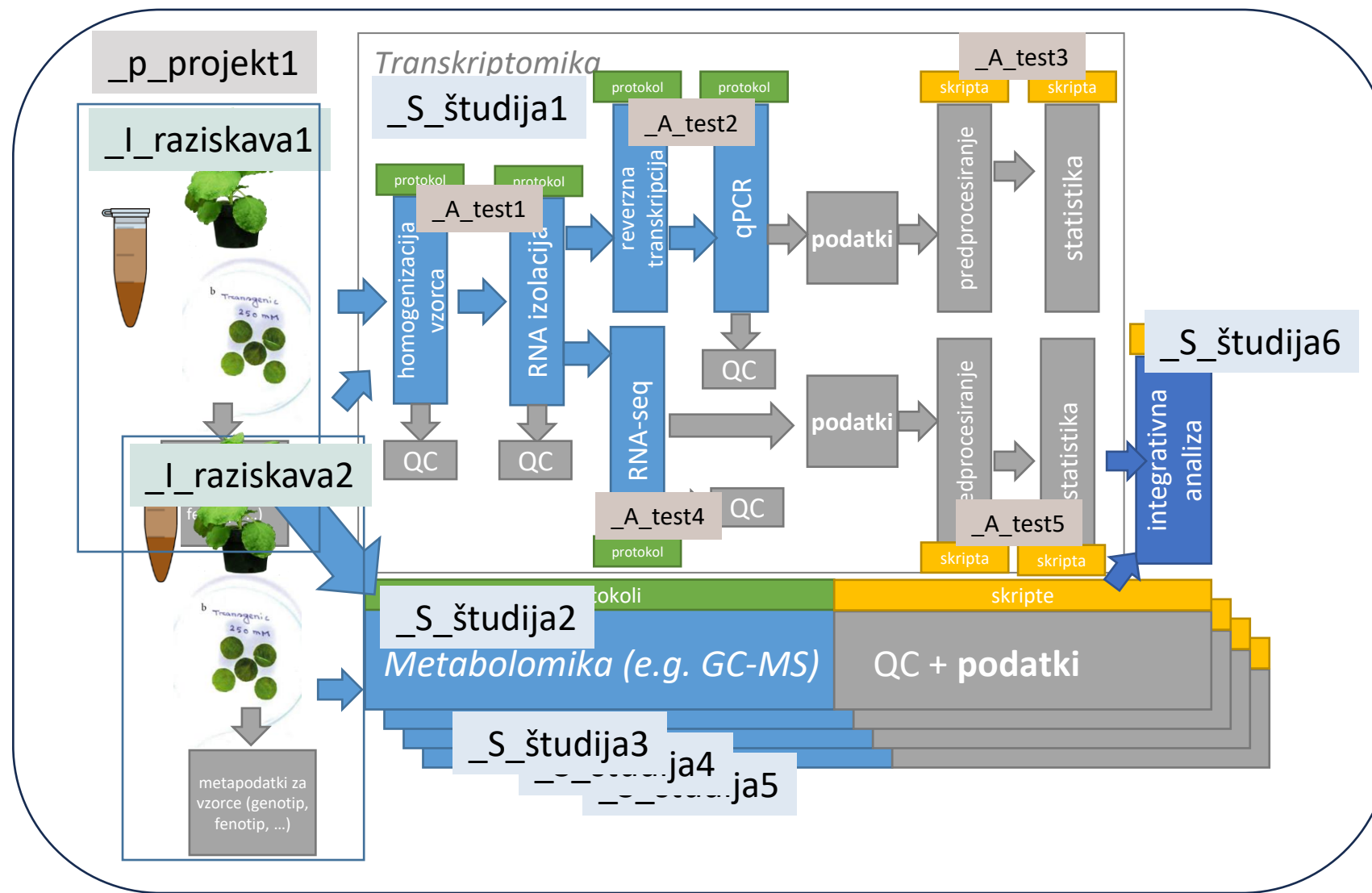
Nivoji, ki jih generiramo s pISA-tree skriptami:

- standardizirane podmape “_<nivo>_ime” (katere podmape bodo ustvarjene, je odvisno od nivoja)
- standardizirane metapodatkovne datoteke *.txt, *.MD
- najbolj (meta)podatkovno “bogat” nivo je nivo testov, ki so ločeni na:
 - wet-lab
 - dry-lab





Tipične raziskave v sistemski biologiji in biotehnologiji





Ustvarjanje nivojev – primer za nivo testa (Assay) RNAisol

```
pISA-tree v.3.0.3

=====
pISA-tree: make ASSAY
-----
Assay Class:      WET
Assay Type:       RNAisol
Assay ID:         02_jun2016-RNAisol
Study:            tequr
Title:            RNA isolation and pre-qPCR sample prep experiment jun 2016
Description:      *
RNA ID           $_RNA

=====

Homogenisation protocol

1 fastPrep
2 TissueLyser
3 mortar
4 Other

Select [1,2,3,4]?
```



- metadatotečne predloge za laboratorijske teste, ki jih izvajamo v naši raziskovalni skupini, so dostopne na GitHub-u

(<https://github.com/NIB-SI/pISA-tree-assay-types>)

- enostavno ustvarjanje lastnih predlog
- GitHub omogoča deljenje in sledenje sprememb predlog

NIB-SI / [pISA-tree-assay-types](https://github.com/NIB-SI/pISA-tree-assay-types) Public

<> Code Issues 1 Pull requests Actions

Files

main

Go to file

- Assay-types
 - DRY
 - R
 - Stat
 - WET
 - GCMS
 - IlluminaSeq
 - LUM
 - ONTSeq
 - RNAisol
 - qPCR

pISA-tree-assay-types

markopetek Upd

Name
..
GCMS
IlluminaSeq
LUM
ONTSeq
RNAisol
qPCR
README.MD



predloga za test RNAisol

```
#Key name      Key Value
# Lines starting with + will be silently added to metadata
+Measurement Type      sample preprocessing with quality control
+Technology Type       RNA isolation and reverse transcription
#
Lab manager            *
RNA ID  $_RNA
ng/ul  Blank
260/280 Blank
260/230 Blank
Homogenisation protocol fastPrep/TissueLyser/mortar
Date Homogenisation    %today%
Isolation Protocol     Rneasy_Plant/ZymoRNA
Date Isolation         %today%
Storage RNA            CU0369
Dnase treatment protocol *
Dnase ID               $_DNase
Date DNase_treatment   %today%
Storage_DNase_treated   CU0370
Operator              *
cDNA ID  $_cDNA
DateRT  %today%
Notes
```

NIB-SI / [pISA-tree-assay-types](#)

Public

<> Code

Issues 1

Pull requests

Actions

+

Files

main

Go to file

Assay-types

DRY

R

Stat

WET

GCMS

IlluminaSeq

LUM

ONTSeq

RNAisol

qPCR

pISA-tree-assay-types



markopetek Upd

Name

..

GCMS

IlluminaSeq

LUM

ONTSeq

RNAisol

qPCR

README.MD



predloga za test qPCR

```
#Key name      Key value
#Creation date 2021-03-05
#MIQE precis minimal standard was used for preparation of template:
#https://bmcmolbiol.biomedcentral.com/articles/10.1186/1471-2199-11-74/tables/1
# Lines starting with + will be silently added to metadata
+Measurement Type      transcriptional profiling
+Technology Type        quantitative PCR
#
Lab manager      *
#
#Assay optimisation/validation
Assay chemistry TaqMan Probe/MGB Probe/SYBRGreen/mixed
Assay optimisation protocol (file path, link or citation) *
Pipetting        liquid handling station/manual
Reaction volume *
#
#qPCR
qPCR platform    HT7900 Fast, ABI (SDS)/Viia7, ABI (QuantStudio)/QuantStudio7 Flex , ABI (QuantStudio)/QuantStudio7 Pro, ABI (Design and Analysis) /LightCycler 480, Roche (LightCycler 480)
qPCR protocol (file path, link or citation) *
One-step RT-PCR (RNA as input) No/Yes
Well format      384/96/tubes
MasterMix        FastStart Universal Probe Master (ROX), Roche (4914058001)/TaqMan Universal PCR Master mix, ABI (4318157)/Power SYBR® Green PCR Master Mix, ABI (4367659)
No. sample technical replicates *
No. sample dilutions *
PCR controls used      NTC only/NTC and PC/only PC/None
No. standard curve dilutions (if using)
Operator          *
#
#Data analysis
Cq value determination software (if different from PCR platform)
Quantification software quantGenius/Excel
qPCR quantification method      relative standard curve/absolute standard curve/ddCq
Notes
```

NIB-SI / [pISA-tree-assay-types](#) Public

<> Code

Issues 1

Pull requests

Actions

Grid icon

Files

main

Search icon

Go to file

pISA-tree-assay-types



markopetek Upd



Projekti organizirani s pISA-tree



- SUSPHIRE - Sustainable Bioproduction of Pheromones for Insect Pest Control in Agriculture

<https://fairdomhub.org/projects/106>



- INDIE - Biotechnological production of sustainable indole

<https://fairdomhub.org/projects/120>



- Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato

<https://fairdomhub.org/projects/161>



- Using RNA interference and systems biology approaches for validation of insecticide targets in Colorado potato beetle guts

<https://fairdomhub.org/projects/252>



Projekti organizirani s pISA-tree



- SUSPHIRE - Sustainable Bioproduction of Pheromones for Insect Pest Control in Agriculture

<https://fairdomhub.org/projects/106>



- INDIE - Biotechnological production of sustainable indole

<https://fairdomhub.org/projects/120>



- Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato

<https://fairdomhub.org/projects/161>

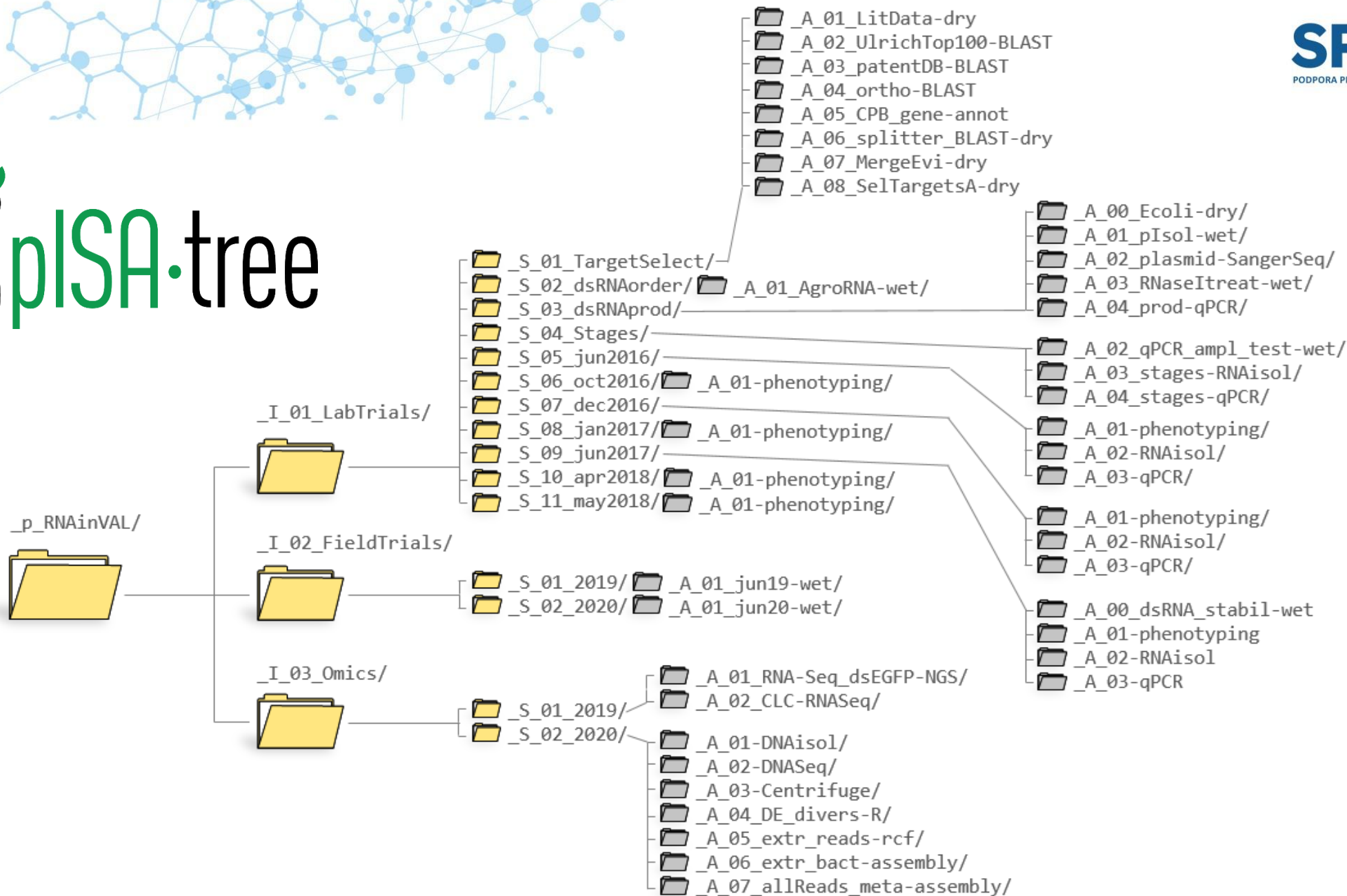


- Using RNA interference and systems biology approaches for validation of insecticide targets in Colorado potato beetle guts

<https://fairdomhub.org/projects/252>



pISA-tree





→ ▾ ↑ << _p_RNAinVAL > _I_01_LabTrials > _S_05_jun2016 > _A_02_jun2016-RNAisol > output ▾

p

I

S

A

- > _p_RNAinVAL
 - > _I_01_LabTrials
 - > _S_01_TargetSelect
 - > _S_02_dsRNAorder
 - > _S_03_dsRNAprod
 - > _S_04_Stages
 - > _S_05_jun2016
 - > _A_01_jun2016-phenotyping
 - > _A_02_jun2016-RNAisol
 - other
 - output**
 - reports
 - > _A_03_jun2016-qPCR
 - reports
 - > _S_06_oct2016
 - > _S_07_dec2016

Name ^

- raw
- README.MD
- totRNA_DirectZol_jun2016_23-05-2017.ndv
- totRNA_DirectZol_jun2016_23-05-2017.txt
- totRNA_DirectZol_jun2016_23-05-2017.xlsx**
- totRNA_DirectZol_jun2016_24-05-2017.ndv
- totRNA_DirectZol_jun2016_24-05-2017.txt
- totRNA_DirectZol_jun2016_24-05-2017.xlsx

PODATKI



File Explorer path: `_p_RNAinVAL > _I_01_LabTrials > _S_05_jun2016 > _A_02_jun2016-RNAisol`

Search: `_A_02_jun2016-RNAisol`

Name	Date modified	Type	Size
other	13. 05. 2021 17:22	File folder	
output	13. 05. 2021 17:16	File folder	
reports	10. 05. 2021 16:27	File folder	
_ASSAY_METADATA.TXT	12. 11. 2021 15:33	Text Document	1 KB
analytes.xlsx	17. 05. 2021 10:34	Microsoft Excel Work...	16 KB
README.MD	10. 05. 2021 16:27	MD File	1 KB
showMetadata.bat	1. 04. 2020 14:26	Windows Batch File	1 KB
showTree.bat	1. 04. 2020 14:26	Windows Batch File	1 KB

Notepad window: `_ASSAY_METADATA.TXT - Notepad`

```
File Edit Format View Help
Assay: _A_02_jun2016-RNAisol
Short Name: 02_jun2016-RNAisol
Assay Class: WET
Assay Type: RNAisol
Title: RNA isolation and pre-qPCR sample prep
Description: RNA isolation and sample preparation for qPCR
pISA Assay creation date: 2021-05-10
pISA Assay creator: Marko Petek
Lab manager: Marko Petek
Phenodata: ../../phenodata_20210113.txt
Featuredata:
RNA ID: $_RNA
Homogenisation protocol: handheld electric rotor-stator homogenization in tubes with Trizol
Date Homogenisation: 2017-04-26
Isolation Protocol: ZymoRNA
Date Isolation: 2017-04-26
Storage RNA: CU470
Dnase treatment protocol: Zymo DNase I
Dnase ID: $_DNase
Date DNase_treatment: 2017-05-10
Storage_DNase_treated: NA
Operator: Marko Petek, Katja Stare
cDNA ID: $_cDNA
DateRT: 2017-05-10
Notes:
Data:
```

1 item selected 789 bytes

METAPODATKI

Ln 1, Col 1 | 100% | Windows (CRLF) | UTF-8



File Explorer window showing the directory structure of a project. The path is: O:\DEJAVNOSTI\OMIKE\pISA-Projects_p_RNAinVAL_I_03_Omics_S_02_metagenome_resp_A_04_DE_divers.

The left pane shows the directory tree, with the 'scripts' folder selected under '_A_04_DE_divers-R'.

The right pane displays the contents of the 'scripts' folder:

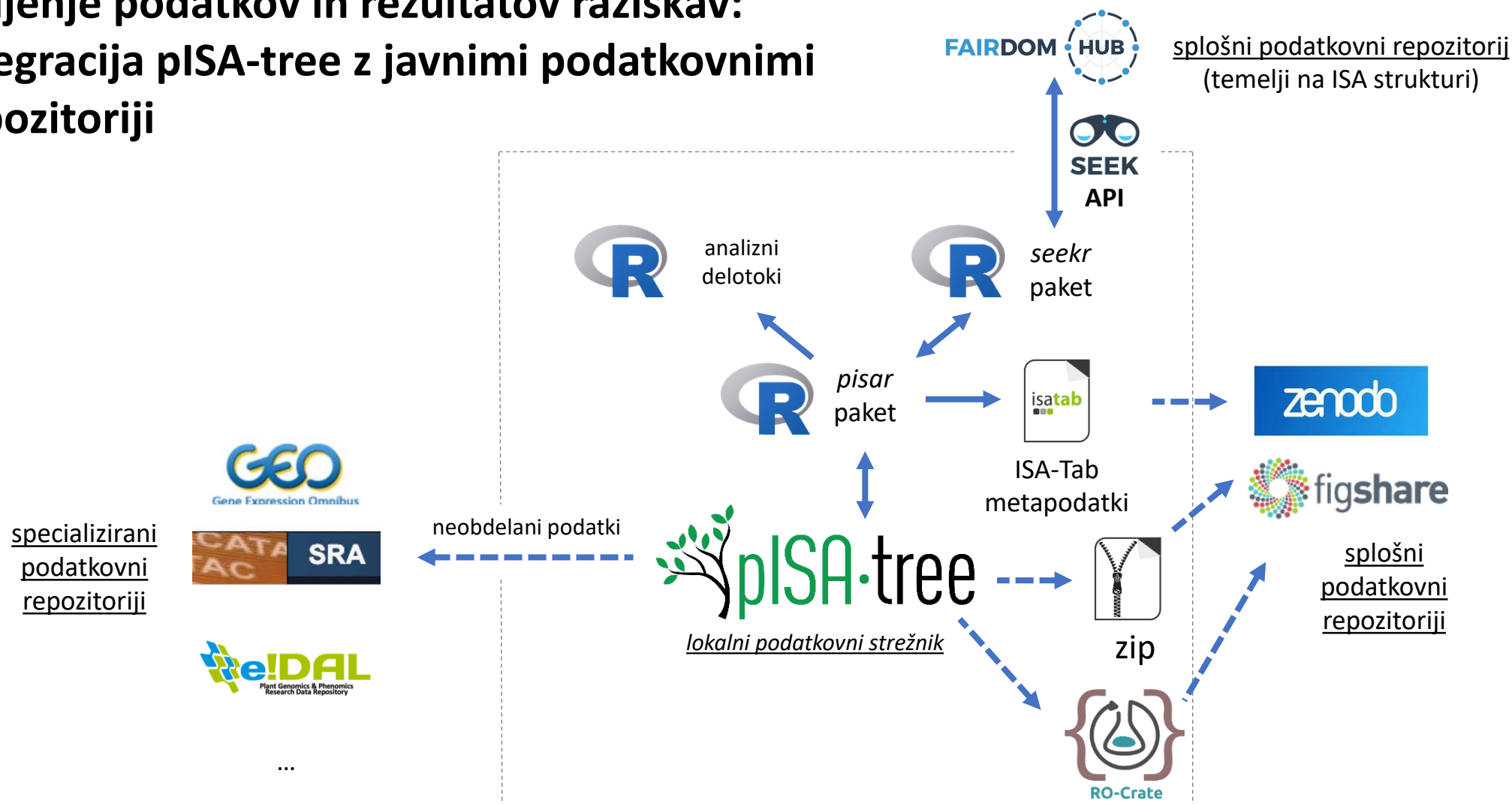
Name	Date modified	Type	Size	Owner
kraken_PCoA_Abundance&Diversity.R	30. 04. 2019 08:31	R File	11 KB	NIBSQL\MarkoP
metagenomics_plots.R	30. 04. 2019 08:31	R File	14 KB	NIBSQL\MarkoP
phyloseq_DESeq2.Rmd	30. 04. 2019 08:31	RMD File	6 KB	NIBSQL\MarkoP
README.MD	8. 04. 2021 18:00	MD File	1 KB	NIBSQL\MarkoP

4 items

skripte ("dry lab") in protokoli ("wet lab") zagotavljajo boljšo reproducibilnost raziskav



Deljenje podatkov in rezultatov raziskav: integracija pISA-tree z javnimi podatkovnimi repozitoriji





Integracija pISA-tree in FAIRDOMhub

→ ▾ ↑ « _p_RNAinVAL > _I_01_LabTrials > _S_05_jun2016 > _A_02_jun2016-RNAisol > output

p

I

S

A

- ▾ _p_RNAinVAL
 - ▾ _I_01_LabTrials
 - > _S_01_TargetSelect
 - > _S_02_dsRNAorder
 - > _S_03_dsRNAprod
 - > _S_04_Stages
 - ▾ _S_05_jun2016
 - > _A_01_jun2016-phenotyping
 - ▾ _A_02_jun2016-RNAisol
 - other
 - output**
 - reports
 - > _A_03_jun2016-qPCR
 - reports
 - > _S_06_oct2016
 - > _S_07_dec2016

Name

- raw
- README.MD
- totRNA_DirectZol_jun2016_23-05-2017.ndv
- totRNA_DirectZol_jun2016_23-05-2017.txt
- totRNA_DirectZol_jun2016_23-05-2017.xlsx**
- totRNA_DirectZol_jun2016_24-05-2017.ndv
- totRNA_DirectZol_jun2016_24-05-2017.txt
- totRNA_DirectZol_jun2016_24-05-2017.xlsx



fairdomhub.org/data_files/4574?graph_view=tree 🔍 ☆

Selected: output/totRNA_DirectZol_jun2016_23-05-2017.xlsx (Data file)
Description: _p_RNAinVAL/_I_01_LabTrials/_S_05_jun2016/_A_02_jun2016-RNAisol/
SEEK ID: https://fairdomhub.org/data_files/4574

I

S

A

- _I_01_LabTrials
 - Investigation files
 - _S_01_TargetSelect
 - _S_02_dsRNAorder
 - _S_03_dsRNAprod
 - _S_04_Stages
 - ▾ _S_05_jun2016
 - _S_05_jun2016-files
 - _A_01_jun2016-phenotyping
 - ▾ _A_02_jun2016-RNAisol
 - _ASSAY_METADATA.TXT
 - analytes.xlsx
 - other/nalepke_june2016_target_selection_totRNA.xlsx
 - output/totRNA_DirectZol_jun2016_23-05-2017.txt
 - output/totRNA_DirectZol_jun2016_23-05-2017.xlsx**
 - output/totRNA_DirectZol_jun2016_24-05-2017.txt
 - output/totRNA_DirectZol_jun2016_24-05-2017.xlsx
 - other/nalepke_june2016_target_selection_totRNA.l6f
 - output/totRNA_DirectZol_jun2016_23-05-2017.ndv
 - output/totRNA_DirectZol_jun2016_24-05-2017.ndv
 - _A_03_jun2016-qPCR
 - _S_06_oct2016
 - _S_07_dec2016
 - _S_08_jan2017
 - _S_09_jun2017
 - _S_10_apr2018
 - _S_11_may2018

Nacionalni inštitut za biologijo (NIB)
Oddelek za biotehnologijo in sistemsko biologijo
Večna pot 121, 1000 Ljubljana
Slovenija



dr. Marko Petek



doc. dr. Anže Županič



prof. dr. Kristina Gruden



doc. dr. Špela Baebler



dr. Tjaša Lukan



prof. dr. Andrej Blejec



dr. Živa Ramšak



Katja Stare



dr. Maja Zagorščak



dr. Anna Coll Rius



Valentina Levak

A photograph of a classroom from a rear perspective. Several students are seated at wooden desks, with their hands raised in the air. The focus is on the hands and arms in the foreground, while the faces and bodies of the students are blurred. In the background, a male teacher in a blue shirt is visible, also looking towards the front. A brown speech bubble with a white border is positioned in the upper center of the frame, containing the white text '?!'.

?!