CC BY 4.0

Data publishing and archival

FAIRDOM

HITS

Alexey Kolodkin and Olga Krebs

Training on "Best practices in research data management and stewardship"

17 June 2021

FAIR publishing and archival (outline)

- Data Publishing in Open Science
- How to choose right discipline-specific data repository (examples)
- CoreTrustSeal concept
- Introduction to Zenodo and FAIRDOMHub (and interoperable publishing)
- Practicum using FAIRDOMHub and Zenodo for FAIR data and results publishing





Transformation to Open Science

ELIXIR All Hands meeting, June 9th, 2021

"Driving the paradigm shift towards Open Science", Kostas Glinos, Head of unit for Open Science / Directorate-General for Research and Innovation European Commission:

Current System (dominant)		Open Science			
Excellence defined largely on the basis of <i>where</i> scientists publish		Composite definition of excellence			
Incentivises researchers to produce specific outputs (<i>mainly</i> <i>publications</i>) and to publish as much and as fast as possible (<i>publish or</i> <i>perish!</i>)	Use of quantitative metrics	Incentivises researchers to share knowledge/data early and openly, to collaborate, and to increase quality and impact; While considering diversity of outputs and research cultures	Use of qualitative and quantitative metrics		
Rewarding individual competing scientists - gaining scientific prestige		Rewarding team work, collaboration and sharing to achieve societal impact (e.g. Covid-19)			





Publishing for Open Science and Reproducibility

- no papers without data
- preferences for open access papers, open access code, etc.
- FAIR data publishing





Still, many researches are not familiar with FAIR \otimes

Open Data Survey, 2019



European commission published that we waste €26bn/yr in Europe by not having FAIR research data:

You are exception ③

Please select the top 3 approaches you adopt to data availability when publishing your research



Most popular platforms (number of people, from 18 participants)

 Zenodo (13, 76.5%)
 FairdomHub (3, 17.6%)

 European Nucleotide Archive (ENA) (6, 35.3%)
 Metabolights (3, 17.6%)

 Gene Expression Omnibus (GEO) (5, 29.4%)
 Genbank (3, 17.6%)

 European Genome-Phenome Archive (EGA) (4, 23.5%)
 PRIDE (2, 11.8%)

 Dryad (6, 35.3%)
 Figshare (2, 11.8%)

 DNA Data Bank of Japan (DDBJ) (2, 11.8%)
 GitHub (2, 11.8%)







Our survey 2021 results (thanks for your participation!)

Discipline-specific repositories. How to choose the right one? How to do the right thigs?

Nucleic acid sequence data: European Nucleotide Archive (ENA) •GenBank •dbSNP European Variation Archive (EVA) DNA DataBank of Japan (DDBJ) •dbVar MGnify NCBI Trace Archive NCBI Sequence Read Archive (SRA) NCBI Assembly

Mathematical & modelling resources: BioModels Database Kinetic Models of Biological Systems (KiMoSys) The Network Data Exchange (NDEx)

HITS

Protein sequence: UniProtKB

Molecular & supramolecular structure: Protein Circular Dichroism Data Bank (PCDDB) Crystallography Open Database (COD) Coherent X-ray Imaging Data Bank (CXIDB) Biological Magnetic Resonance Data Bank (BMRB) Electron Microscopy Data Bank (EMDB) Worldwide Protein Data Bank (wwPDB) Structural Biology Data Grid

PRIDE (required for publishing structural data) **Neuroscience:** NeuroMorpho.org OpenNeuro (formerly OpenfMRI) •EBRAINS **Metabolomics & Proteomics**

 MassIVE MetaboLights

- •Gene Expression Omnibus (GEO)
- ArrayExpress
- GenomeRNAi
- •dbGAP
- The European Genome-phenome Archive (EGA)
- Database of Interacting Proteins (DIP)
- IntAct
- Japanese Genotype-phenotype Archive (JGA)
- Biological General Repository for Interaction Datasets
- Genomic Expression Archive (GEA)
- NCBI PubChem BioAssay



https://www.nature.com/sdata/policies/repositories

Discipline-specific repositories. How to choose the right one? How to do the right thigs?

- FAIR-suitable data repository
- standardized file format
- minimal required information (metadata), in standardized form, machine readable



https://www.nature.com/sdata/policies/repositories

DNA DataBank of Japan (DDBJ)	view FAIRsharing entry
European Nucleotide Archive (ENA)	view FAIRsharing entry
GenBank	view FAIRsharing entry
dbSNP	view FAIRsharing entry
European Variation Archive (EVA)	view FAIRsharing entry
dbVar	view FAIRsharing entry
MGnify	view FAIRsharing entry
NCBI Trace Archive	view FAIRsharing entry
NCBI Sequence Read Archive (SRA)	view FAIRsharing entry
NCBI Assembly	

Protein sequence 🧈

UniProtKB

view FAIRsharing entry





Examples: ENA (European Nucleotide Archive)







Examples: PRIDE (Proteomics IDEntification Database)



Project PXD026560

PRIDE Assigned Tags: Sars-cov-2 Covid-19	Name 💠			Download	
	checksum.txt	🗋 OTHER	732 bit	FTP	
Summary	LZ-20200501-Chen-S-CC-TASVc-AV5pmol.raw_20200505(1).byrsit	Q SEARCH	516	FTP	
	LZ-20200501-Chen-S-CC-TASVC-AV5pmol.raw.byspec2	Q SEARCH	753	FTP	
Title Analysis of SARS-CoV-2 S protein by LC-MS	LZ-20200501-Chen-S-CC-TASVc-AV5pmol.raw	RAW	4139	FTP	
	LZ-20200501-Chen-S-CC-KREc-AV5pmol.raw_20200505(1).byrslt	Q SEARCH	602	FTP	

Description

Analysis of disulfide bonds for SARS-Cov-2 S protein by LC-MS

Sample Processing Protocol

Cross-linked peptides generated by digesting SARS-Cov-2 S protein using alpha lytic protease, or a combination of trypsin,

Data Processing Protocol

Byonic (v3.8.13, Protein Metrics Inc.) was used for database searches with mass tolerance set as 20 ppm for both precursc

Contact

Peng Zhao, University of Georgia Lance Wells, University of Georgia (lab head)

Submission Date 08/06/2021

Publication Date 08/06/2021



Examples: GEO (Gene Expression Omnibus)

Series GSE1754	50	Query DataSets for GSE175450
Status Title	Public on Jun 08, 2021 Complement activation induces 6	excessive T cell cytotoxicity in severe COVID-
Organiem	19 Homo conione	
Experiment type	Expression profiling by high throu	about sequencing
Summary	We performed single cell transc versus severe COVID-19 subject like-illness and HBV infection to und function.	iptonics in 13 acute and convalescent mild s, in healthy controls and in sujects with flu- assess COVID-19-specific T cell populations
Overall design	Blood from 13 COVID-19 patient healthy controls as well as from with chronic HBV infection was or PBMC CD38+ CD4+ an CD8+ we CD38+ T cells were subjected to For many samples, both all sequenced. Samples were po demultiplexed. based on both Hat	ts at acute and convalescent phases, from 5 5 patients with flu-like disease and 4 patients 5 patients with flu-like disease and 4 patients 5 patients of the re purified and both, total PBMC and enriched 5' GEX scRNA- and 5'-VDJ library preparation. PBMCs and enriched CD38+ T-cells were soled with Antibody-based Hashtags and shtags and SNPs.
Contributor(s)	Sawitzki B, Astaburuaga-García R	, Blüthgen N, Obermaver-Wasserscheid B,

HITS

	N4 04 0004
Submission date	May 24, 2021
Last update date	Jun 10, 2021
Contact name	Nils Blüthgen
E-mail(s)	nils.bluethgen@charite.de
Organization name	Charite Universitätsmedizin Berlin
Department	Institute of Pathology
Street address	Chariteplatz 1
City	Berlin
ZIP/Postal code	10117
Country	Germany
Platforms (1)	GPL24676 Illumina NovaSeq 6000 (Homo sapiens)
Samples (57)	GSM5333722 C19-CB-0029_acute_CD38-T-cells
± More	GSM5333723 C19-CB-0029_acute_PBMC
	GSM5333724 C19-CB-0029_convalescent_CD38-T-cells
Relations	
BioProject	PRJNA732416

Download family		F	ormat
SOFT formatted family file(s) SOFT 🛛			
MINiML formatted family file(s)	MINIML		
Series Matrix File(s) TXT 🛽			XT 🕐
Supplementary file	Size	Download	File type/resource
GSE175450_RAW.tar	235.5 Mb	(http)(custom)	TAR (of TSV)
GSE175450_cell_metadata_table.tsv.gz	1.5 Mb	(ftp)(http)	TSV
Raw data not provided for this record			
Processed data provided as supplementary file			



CoreTrastSeal future



COAR – Confederation of Open Access Repositories , CoreTrustSeal, the European University Association, Science Europe and the World Data System have published a joint position statement on "Data Repository Selection – Criteria That Matter".

DSA (Data Seal Approval) and WDS (World Data System)

Core Trustworthy Data Repositories Requirements and procedures:

- universal catalogue of requirements that reflects the core characteristics of trustworthy data repositories
- certification to any interested data repository





Generalist repositories



https://about.zenodo.org



http://www.fairdomhub.org http://www.fair-dom.org Data repository that allows to deposit research papers, data sets, research software, reports, and any other research related digital artifacts (provides DOI)

Open-source software and services that allows to deposit and catalogue data, SOPs, models, visualizations (provides DOI)

Not just repositories, but FAIR project-oriented platforms to integrate data, SOPs, codes, visualizations, publications...and research communities..

FAIRDOMHub and Zenodo can talk to each other using same language (Interoperable)

...you will see it on the example





Practicum using FAIRDOMHub and Zenodo for FAIR data and results publishing



Practicum using FAIRDOMHub and Zenodo for FAIR data and results publishing



One can publish various things in Zenodo, easily search and organize

File Type	Keywords	Туре
□ Pdf (922741)	□ Taxonomy (658312)	Publication (1053896)
🗆 Jpg (367879)	Biodiversity (657049)	🗆 Image (612242) +
🗆 Png (250679)	🗆 Animalia (377029)	🗆 Dataset (90605)
🗆 Html (109493)	Arthropoda (290929)	□ Software (56151)
□ Zip (89258)	🗆 Terrestrial (270732)	Presentation (22134)
🗆 Xml (16484)	🗆 Herbarium (270653)	🗆 Poster (8853)
□ Hdf5 (15142)	🗆 Insecta (196335)	🗆 Other (7383)
Docx (14456)	Coleoptera (54916)	🗆 Video (3679)
□ Xlsx (14403)	🗆 Arachnida (46335)	🗆 Lesson (2705)
🗆 Txt (13597)	🗆 Chordata (44654)	Physicalobject (69)







Entering metadata when publishing in Zenodo

					Publication date *	2021-06-17
Zenodo Search	Q Uplo	ad Communities	alex	xey.kolodkin@uni.lu 🗸		Required. Forr
會 Delete				ave 🗸 Publish	🗐 Title *	
New upload						Required.
Instructions: (i) Upload minimum one file or fill-in requ	uired fields (marked with a red star). (ii) Pi	ess "Save" to save your upload for editing later. (iii)	When ready, press "Publish" to finalize and m	ake your upload public.	💄 Authors *	Family nan
Filename (1 files)		Size	Progress	Delete		
Data_Publishing_AK_2021-06-14_v10.ppt		6.1 MB		â		+ Add anoth
Upload type				required 💙	🖋 Description *	
Publication Poster	Presentation Dataset	Video/Audio Software	Lesson Physical object	Other	and so on	
Publication type	Journal article			~		
	- univation type.	https://sandbox	.zenodo.org/			



Findable Accessible Interoperable Reusable

Data S O Ps Models

F

Α

R

D

Μ







Many Objects are the Outcomes of Research

All are first class citizens and are required to make research FAIR+R Each object has its own metadata and repositories



FAIRDOM Platform: Catalogue, Commons, Collections, Project-centric Data Management



Organising and Linking Assets sainfrastructure



 Use national or institutional data storage infrastructure

Project centric approach

<u>Home</u> > <u>Projects Index</u> > MOSES

MOSES (Micro Organism Systems biology: Energy and Saccharomyces cerevisiae) develops a new Systems Biology approach, which is called 'domino systems biology'. It uses this to unravel the role of cellular free energy ('ATP') in the control and regulation of cell function. MOSES operates though continuous iterations between partner groups through a new systems-biology driven data-management workflow. MOSES also tries to serve as a substrate for three or more other SYSMO programs.



Programme: SysMO

Public web page: http://www.moses.sys-bio.net/

Organisms: Saccharomyces cerevisiae

FAIRDOM PALs: Femke Mensonides, Maksim Zakhartsev, Walter Glaser

Related items

 People (20)
 Institutions (6)
 Investigations (3)
 Studies (8)
 Assays (12+1)
 Samples (0+4)
 Cell cultures (0+4)
 Strains (1)
 Data files (9+1)
 Models (1+2)

 SOPs (7+4)
 Publications (5)
 Publications (5)
 Publications (5)
 Publications (5)



FAIRDOMHub: Landing page



Data Generation



Standard Operating Procedures (SOP):

Quality Control

Home > SOPs Index > Introduction of shRNAs, miRNAs or anti-microRNAs into primary human hepatocytes with lentivirus

Introduction of shRNAs, miRNAs or anti-microRNAs into primary human hepatocytes with lentivirus

Q View content Download SOP

Here we used VSV-G-pseudotyped, EGFP-expressing lentiviral vectors to develop an efficient gene transfer protocol to modify gene expression in primary human hepatocytes (by RNAi). The protocol comprises the production of recombinant viruses as well as the steps for efficient delivery of short-hairpin RNA (shRNAs), microRNAs or anti-microRNAs to human hepatocytes. On average infection efficiencies of over 95% are achieved at relatively low multiplicity of infection (MOI), which effectively reduces the amount of preparative work required per experiment. Depending on the laboratory equipment available, we provide here two alternative workflows, which can be easily adapted in the lab. The procedure of virus production with subsequent titer determination takes approx. 6 to 10 working days. The procedure of viral infection of hepatocytes until effects can be measured takes approx. 3 to 5 days. This protocol should be helpful to study many aspects of functional genomics in primary human hepatocytes.



Filename: Lentiviral production and infection_SOP_04042011.pdf

Format: PDF document



Fig.1: Schematic presentation of designed template sequences which are processed intracellularly into short hairpin RNAs, microRNAs or anti-microRNAs. The stem-loop structures consisting of both the sense and anti-sense strands of the targeted sequence are separated by a loop sequence.

MATERIALS

REAGENTS/KITS

BLOCK-iT[™] Lentiviral RNAi Expression Kit (Invitrogen#49-4400) ViraPower[™] Lentiviral Gateway Expression Kit (Invitrogen#K49-6000) miRZip™ Lentivector-based Anti- MicroRNAs (System Biosciences#MZIPxxxPA/AA-1) d microRNAs (System Biosciences#PMIRHxxxPA/AA-1)

PROCEDURE

NOTE: all the steps marked with "S" should be performed following recommended guidelines for working with BL-2 organisms (Germany: S2 lab).

1. Preparation of HEK293FT cells.

For cultivating HEK293FT cells, add G148 (Geneticin, final concentration 500 µg/ml) to the DMEM culture medium with components (see Reagent Setup). The cells should be passaged at least 1-2 times after thawing to adapt to the culture conditions. Three days prior to transfection, plate out the cells at a density of approximately 3,5*106 cells/per 1 T175 flask in 30 ml of medium with components and G148 to achieve optimal phase of cellular

Reproducible models in FAIRDOM

metadata annotation against standards validation, comparison and simulation

JWS Online



Deletions are coloured in red and insertions are coloured in blue					
SBML Diffe	Model versioning				
Parameters					
VappSPSSPP	Attribute value has changed: 797 → 500				
Compartments					
default_compartment \rightarrow main	Attribute <i>id</i> has changed: default_compartment → main				
Species					
Sucrose	Attribute compartment has changed: default_compartment → main				
ADPGam	Attribute compartment has changed: default_compartment → main				
PPam	Attribute compartment has changed: default_compartment → main				
Poyt	Attribute compartment has changed: default_compartment main				
F6Pcyt	Attribute compartment has changed: default_compartment → main				
ADPam	Attribute compartment has changed: default_compartment → main				
UDPcyt	Attribute compartment has changed: default_compartment → main				
Glucoseam	Attribute <i>compartment</i> has changed: default_compartment → main				
G6Pam	Attribute compartment has changed: default_compartment → main				



Reproducing simulations

SED-ML Simulation Result: bachmann2011









Effects of interest Physical conditions Technical variations Biological variations

Samples

UID 11	Name 11	DateOfBirth 11	Sex	Treatment1	Treatment1Route
MUS-200901ENG-1	19-1892	29 August 2019	F	saline	intraperitoneal
MUS-200901ENG-2	19-2135	19 September 2019	М	saline	intraperitoneal
MUS-200901ENG-3	19-2136	19 September 2019	м	saline	intraperitoneal
MUS-200901ENG-4	19-2137	19 September 2019	М	saline	intraperitoneal
MUS-200901ENG-5	19-2140	27 September 2019	F	saline	intraperitoneal
MUS-200901ENG-6	19-2141	27 September 2019	М	saline	intraperitoneal
MUS-200901ENG-7	19-2142	27 September 2019	м	saline	intraperitoneal
MUS-200901ENG-8	19-2143	27 September 2019	м	saline	intraperitoneal
MUS-200901ENG-9	19-2148	28 September 2019	F	saline	intraperitoneal
MUS-200901ENG-10	19-2149	28 September 2019	F	saline	intraperitoneal
Search UID	Search Name	Search Scientist	Search Protocol	Search Treatment1	Search Treatment1Route

Biosamples



More examples here https://fairdomhub.org/sample_types



Create snapshot and get DOI for your reaseach asset

Snapshot 8 (25th Apr 2016)



DOI: 10.15490/seek.1.investigation.74



Publishing...snapshot and assign DOIs Credits and Citations



Fenner et al, A Data Citation Roadmap for Scholarly Data Repositories doi: https://doi.org/10.1101/097196

Project Commons Organise > Share > Disseminate











Quyen Nguyen



Standards

Martin Golebiewski









Norman Morrison

Dagmar Waltemath



Alexey Kolodkin **Luxembourg Centre for Systems Biomedicine (LCSB)**













Jacky Snoep





Olga Krebs





Martin Peters, Martin Scharm Systems Biology Bioinformatics University of Rostock, Germany

PALs, users developers







HITS



Wolfgang Mueller







Jakub Straszewski



Let's try it - Hands On





https://sandbox7.fairdomhub.org/

https://docs.google.com/document/d/1QtM-vXkGzqxGQguu0h4kMEpItPX0Qzr/edit