

Research outputs beyond the paper

Code, research software & computational workflows

NBIS Data Management Team

data-management@scilifelab.se

Presented by Wolmar Nyberg Åkerström
Open Science Uppsala, Uppsala, Sweden
10 March 2023



Priem, J. (2013). Beyond the paper. Comment in Nature, 495(7442), 437–440. <https://doi.org/10.1038/495437a>

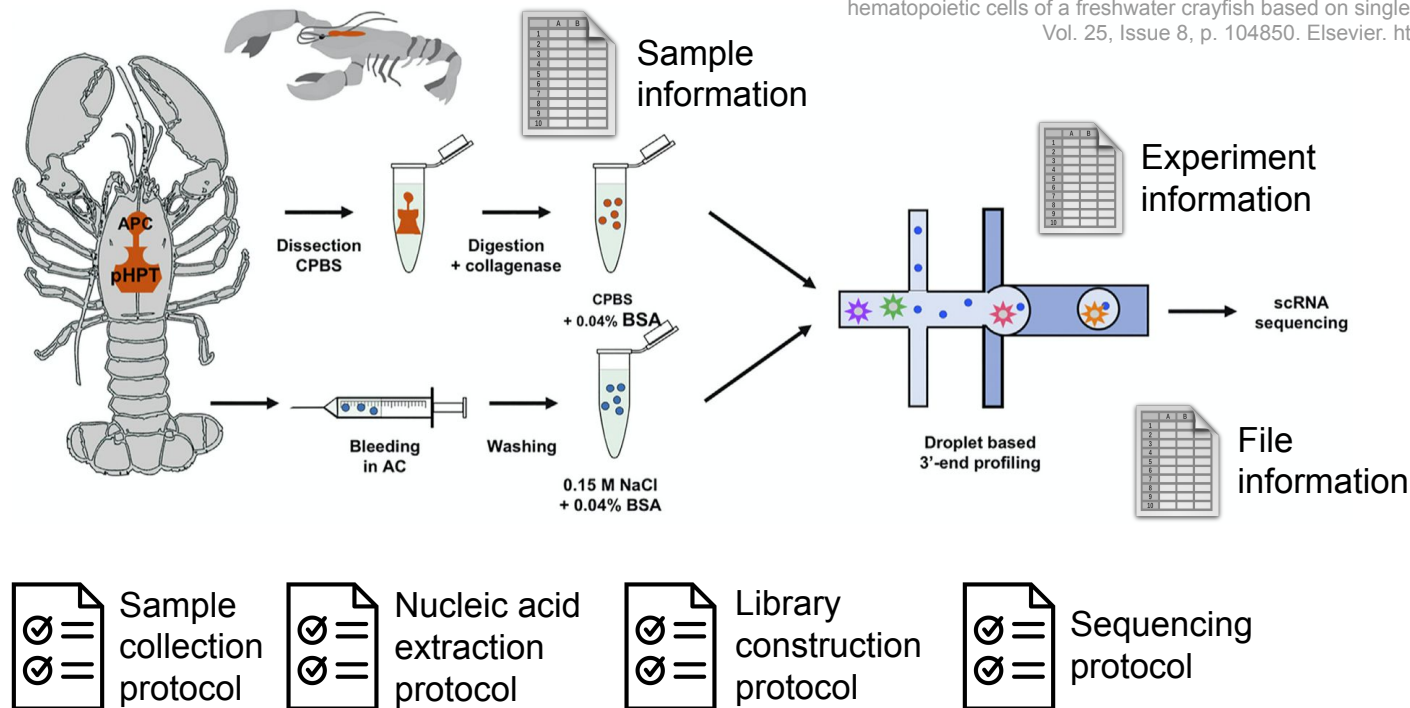
“We now have a unique opportunity as scholars to guide the evolution of our tools in directions that honour our values and benefit our communities. Here's what to do. First, try new things: publish new kinds of products, share them in new places and brag about them using new metrics.”

– Jason Priem, 2013

“Tools are emerging to facilitate this ‘share early, share often’ approach.”

"Protocol" icon by Justin Blake from thenounproject.com

Söderhäll, I., Fasterius, E., Ekblom, C., & Söderhäll, K. (2022). Characterization of hemocytes and hematopoietic cells of a freshwater crayfish based on single-cell transcriptome analysis. In *iScience* Vol. 25, Issue 8, p. 104850. Elsevier. <https://doi.org/10.1016/j.isci.2022.104850>



- checksums.md5
- SampleSheet.csv
- ▼ SI-GA-F2_1
 - TJ-2700-1_S1_L001_R1_001.fastq.gz
 - TJ-2700-1_S1_L001_R2_001.fastq.gz
 - TJ-2700-1_S1_L002_R1_001.fastq.gz
 - TJ-2700-1_S1_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_2
 - TJ-2700-1_S2_L001_R1_001.fastq.gz
 - TJ-2700-1_S2_L001_R2_001.fastq.gz
 - TJ-2700-1_S2_L002_R1_001.fastq.gz
 - TJ-2700-1_S2_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_3
 - TJ-2700-1_S3_L001_R1_001.fastq.gz
 - TJ-2700-1_S3_L001_R2_001.fastq.gz
 - TJ-2700-1_S3_L002_R1_001.fastq.gz
 - TJ-2700-1_S3_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_4
 - TJ-2700-1_S4_L001_R1_001.fastq.gz
 - TJ-2700-1_S4_L001_R2_001.fastq.gz
 - TJ-2700-1_S4_L002_R1_001.fastq.gz
 - TJ-2700-1_S4_L002_R2_001.fastq.gz

File formats and data specifications



File Formats for Illumina Sequencing

Numerous options are available for converting data to compatible sequence file formats such as FASTQ files, and for downstream analysis of sequencing data. Illumina sequencers are designed so data can be easily streamed into Illumina Connected Analytics and BaseSpace Sequence Hub for cloud-based data management, analysis, and collaboration.

Raw data files are provided in sequence file formats that are compatible, or easily converted, to standardized data formats for streamlined aggregation and mining of large cohorts. With the DRAGEN BioIT platform, the newest file format, FASTQ.ORA is available. FASTQ.ORA is a lossless compression file reducing the size, time to transfer, and storage cost.

FASTQ Sequence File Format

FASTQ is a text-based sequencing data file format that stores both raw sequence data and quality scores. FASTQ files have become the standard format for storing NGS data from Illumina sequencing systems, and can be used as input for a wide variety of secondary data analysis solutions.

The MiniSeq and MiSeq Sequencing Systems provide the option to automatically convert data from BCL to FASTQ format, so separate conversion software is not required.

[Learn More About FASTQ Files](#)

FASTQ.ORA Sequence File Format

FASTQ.ORA is a binary compressed file format of the text-based FASTQ sequencing data file format. fastq.ora files are up to 5x smaller than their corresponding fastq.gz files without compromising data integrity. All fastq.ora files can be read using the free decompression software available [here](#). Once installed, a simple command can be used to directly pipe the output of decompression on the fly into a wide range of popular mapping tools such as BWA, STAR, and Bowtie. DRAGEN.ORA compression is available with the DRAGEN server and on-board the NextSeq1000/2000.

Chromium Single Cell 3' Reagent Kits v3

FOR USE WITH

Chromium Single Cell 3' GEN Library & Gel Bead Kit v3, 16 runs PN-1000075
Chromium Single Cell 3' GEN Library & Gel Bead Kit v3, 4 runs PN-1000092
Chromium Single Cell 8 Chip Kit, 16 runs PN-1000153 (America & Asia Pacific), PN-1000172 (Europe, Middle East & Africa)
Chromium Single Cell 8 Chip Kit, 16 runs PN-1000154 (America & Asia Pacific), PN-1000172 (Europe, Middle East & Africa)
Chromium 17 Multiplex Kit, 16 runs PN-100262

MultiQC

Report for project TJ-2700 on runfolder 201126_A00605_0172_BHVVTNDRXX

NGI Uppsala - SNP&SEQ Technology Platform

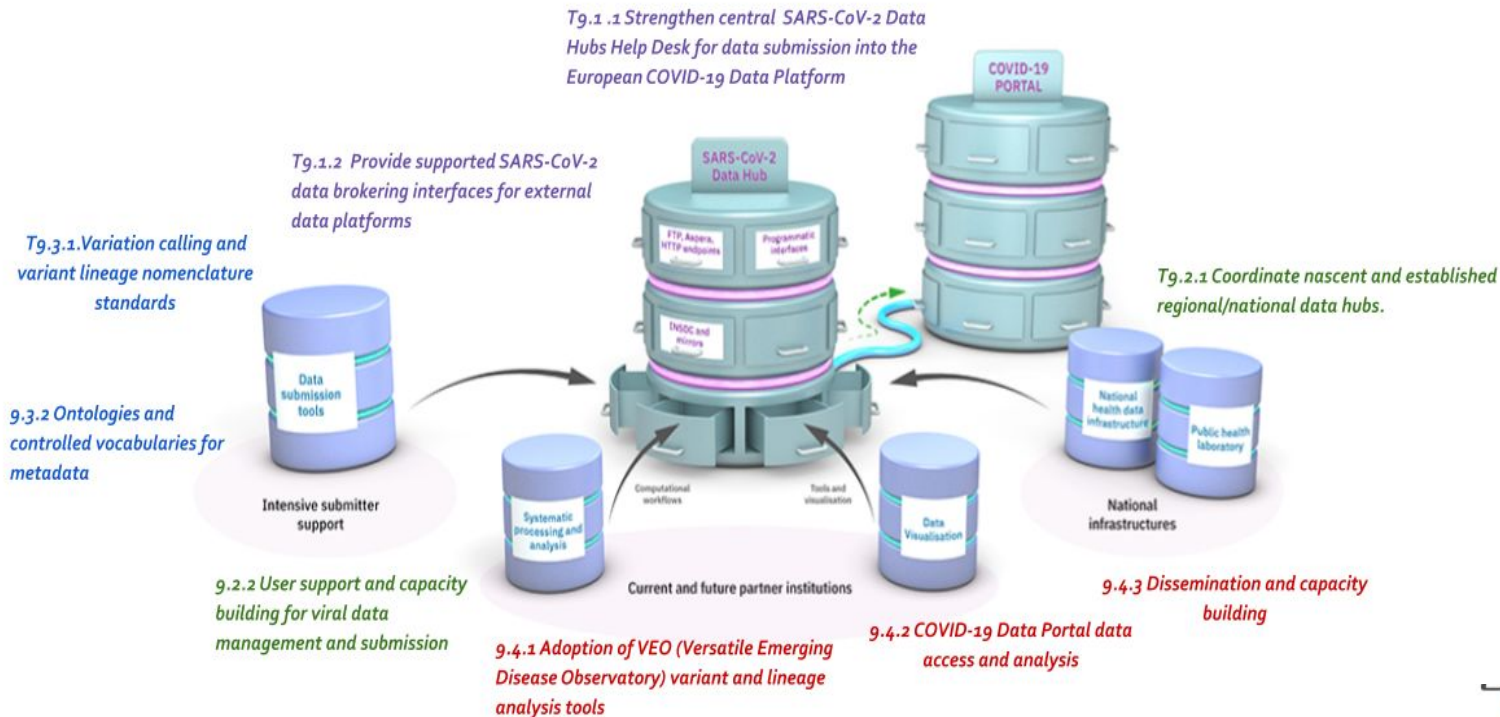
This is a report containing quality control information about your project run at the SNP&SEQ Technology Platform. If you have any questions, please do not hesitate to contact us at seq@medsci.uu.se

Report generated on 2020-11-27, 01:07 based on data in: /seqreports-data/nef_work/2b/d14a7e2d23927c741f2a6713e8fa8

General Statistics

Sample Name	% GC	Length	M Seqs
TJ-2700-1_S1_L001_R1_001	46%	28 bp	53.5
TJ-2700-1_S1_L001_R2_001	48%	91 bp	53.5
TJ-2700-1_S1_L002_R1_001	46%	28 bp	53.5
TJ-2700-1_S1_L002_R2_001	48%	91 bp	53.5

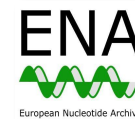
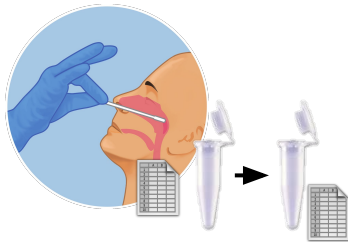
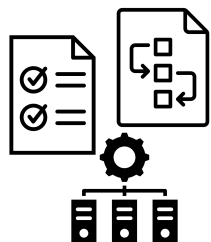
201126_A00605_0172_BHVVTNDRXX_TJ-2700_multiqc_report.html
checksums.md5
SampleSheet.csv



Design for transparency and reuse



“Protocol” & “project plan” icons by Justin Blake, and “infrastructure” icon by Eko Purnomo, from thenounproject.com



Study & data
design

Sampling
& specimen
collection

Sample
preparation

Sample analysis
& data generation

Data processing
to prepare inputs
for analysis

Data
analysis

Communicating
results

Procedures

data protection,
ethics permit,
infrastructure,
standards,
protocols,
data dictionaries,
data access, ...

Biosamples and instruments

populations (statistical) and inclusion criteria,
physical processing steps,
working storage conditions,
long-term storage location,
sample quality assessment,
sample annotations,
reagents, instruments, kits, ...

Data and computational workflows

digital processing steps,
working storage conditions,
long-term storage location,
data quality assessment,
sample/data annotations,
reference data,
analysis method...

Outputs

publications,
data,
tools,
workflows,
reports,
dashboards, ...



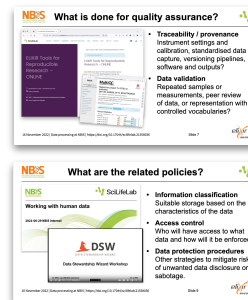
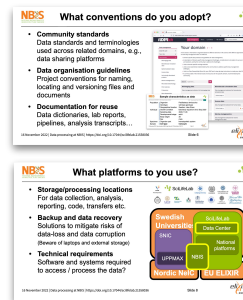
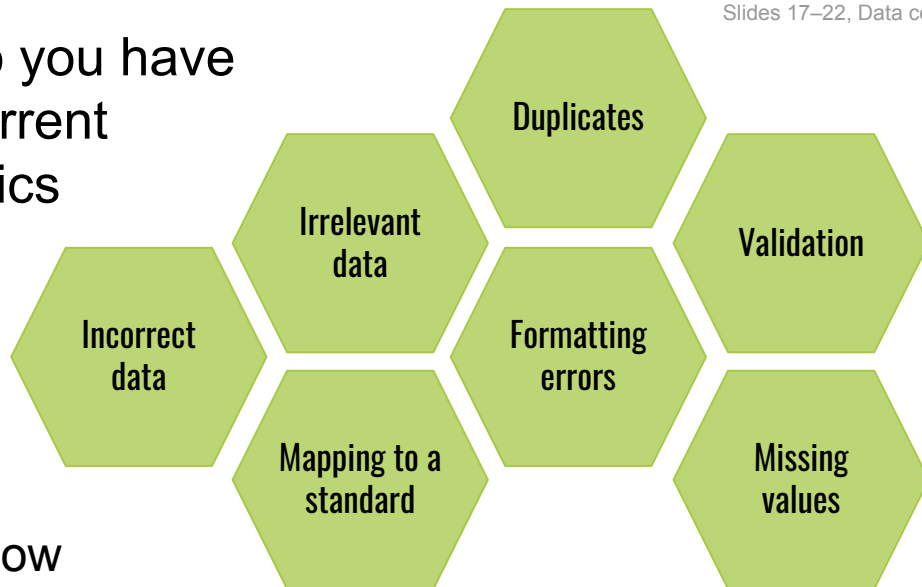
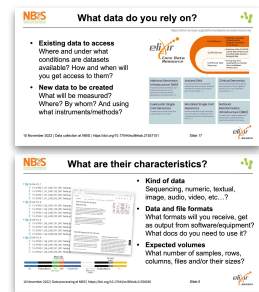
- Make project more efficient by implementing **good practices for handling research data & software**
- Establish procedures to **address all aspects of data management** throughout the project life cycle
- Adopt best-practice guidelines that encourage **Reproducible Research, Open Science & FAIR principles**





The data do you have
and their current
characteristics

The characteristics that
you want, ready for the
platforms you will use



Typical workflow

Inspecting

Harmonising

Verifying

Reporting/Documenting

Detect unexpected, incorrect,
and inconsistent data, etc.

Fix or remove anomalies,
transform, convert etc.

Test that the results are
complete and correct.

Record changes made and
quality assessments.

Recipes for processing data



<https://faircookbook.elixir-europe.org>



- Over 70 **recipes** covering all operational aspects of FAIR data management
- Recipes are **citable** (PID) and **credited** to authors (ORCID)
- Covering **technical processes** with FAIRification **examples** in the life sciences, incl.
 - omics
 - pre-clinical
 - clinical areas
- **Use it, contribute to it, and recommend it!**

The screenshot shows the FAIR Cookbook website. The header is purple with the FAIR Cookbook logo and navigation icons. A left sidebar contains a search bar and a menu with categories like FOREWORD, RECIPES, and ABOUT. The main content area is titled 'The recipes' and explains that the cookbook organizes recipes by FAIR elements, audience type, reading time, and difficulty. It lists four main categories: Findability (F), Accessibility (A), Interoperability (I), and Reusability (R). Each category has a list of exemplar recipes and a link to 'More about' that category. At the bottom, there are three buttons for 'Infrastructure', 'Applied Examples', and 'Assessment'.



covid19.galaxyproject X

search name, annotation, owner, and tag:

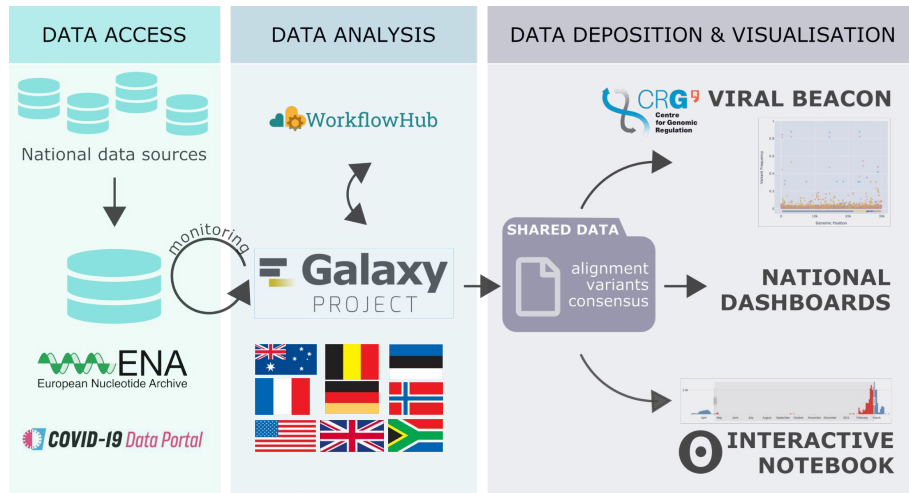


https://usegalaxy.org/workflows/list_published

Advanced Search

Name	Annotation	Owner	Community Rating	Community Tags
COVID-19: variation analysis on WGS SE data	Call variants from WGS (non-ampliconic) single-end reads.	wolfgang-maier	★★★★★	covid-19 covid19.galaxyproject.org
COVID-19: variation analysis on ARTIC PE data	Call variants from ampliconic paired-end reads.	wolfgang-maier	★★★★★	covid-19 artic covid19.galaxyproject.org
COVID-19: variation analysis on WGS PE data	Call variants from WGS (non-ampliconic) paired-end reads.	wolfgang-maier	★★★★★	covid-19 covid19.galaxyproject.org
COVID-19: consensus construction	Build a consensus sequence from a list of variants. Hard-mask regions with low coverage and sites with called, but filtered variants. Note: Sites with...	wolfgang-maier	★★★★★	covid-19 covid19.galaxyproject.org
COVID-19: variation analysis of ARTIC ONT data	A Galaxy workflow that replaces the ARTIC minion shell command	wolfgang-maier	★★★★★	ont covid-19 artic covid19.galaxyproject.org
COVID-19: variation analysis reporting	Generate variant reports for the output of SARS-CoV-2 variation analysis workflows	wolfgang-maier	★★★★★	covid-19 covid19.galaxyproject.org

- Galaxy is a platform connecting **data**, computational **workflows**, **visualisations** and other **services**
- Supports **reproducible**, and **transparent** computational analysis
- Community committed to **improving tools & workflows**



Galaxy

Galaxy
EUROPE

Galaxy
AUSTRALIA



Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.





**We teach foundational coding
and data science skills to
researchers worldwide.**



FIVE RECOMMENDATIONS FOR FAIR SOFTWARE

LET'S GO! →

ENDORSE



In which journals should I publish my software?

By Neil Chue Hong.

Until there is a radical change in the way that academic credit is given, the principal record of scientific research is still the peer-reviewed publication. Given that software is a fundamental part of doing science in the digital age, the question we are often asked is: *where can I publish papers which are primarily focused on my scientific software?*

The following is a list of journals which accept submissions that are primarily about the software, and not necessarily on new algorithms or new science. There is an expectation that the use of the software will



image courtesy of shimray

Tags

- Neil Chue Hong
- Publications
- Journals



Upcoming event

nf-core Training - March 2023

A set of global online Nextflow and nf-core training events

15:00 CET, March 13, 2023  Training

[Event Details](#)

Event countdown:

2 days,
22h 50m 28s



A community effort to collect a curated set of analysis pipelines built using
Nextflow.

[VIEW
PIPELINES](#)



Curating research artifacts to support scientific integrity.

The CUrating for REproducibility (CuRe) Consortium supports curation of research data and review of code and associated digital scholarly objects for the purpose of facilitating the digital preservation of the evidence-base necessary for future understanding, evaluation, and reproducibility of scientific claims.

10 Things

News

Data Quality Review

CURE Training

Get Involved!

10 Things for Curating Reproducible and FAIR Research

Computational reproducibility requires a village. This document is primarily for data curators and information professionals who are charged with verifying that a computation can be executed and can reproduce prespecified results. Secondly, it is for researchers, publishers, editors, reviewers, and others who have a stake in creating, using, sharing, publishing, or preserving reproducible research.

The 10 Things for Curating Reproducible and FAIR Research is the result of the collaborative efforts of members of the Research Data Alliance (RDA) CURE-FAIR Working Group. The original 10 Things document was accepted by RDA as an endorsed recommendation cited below:

Arguillas, F., Christian, T., Gooch, M., Honeyman, T., & Peer, L. (2022). *10 Things for Curating Reproducible and FAIR Research* (Version 1.1). Research Data Alliance. <https://doi.org/10.15497/RDA00074>

Barker et al (2022). Introducing the FAIR Principles for research software. Scientific Data, 9(1), 622. <https://doi.org/10.1038/s41597-022-01710-x>

Article | [Open Access](#) | [Published: 14 October 2022](#)

Introducing the FAIR Principles for research software

[Michelle Barker](#) , [Neil P. Chue Hong](#), [Daniel S. Katz](#), [Anna-Lena Lamprecht](#), [Carlos Martinez-Ortiz](#), [Fotis Psomopoulos](#), [Jennifer Harrow](#), [Leyla Jael Castro](#), [Morane Gruenpeter](#), [Paula Andrea Martinez](#) & [Tom Honeyman](#)

[Scientific Data](#) **9**, Article number: 622 (2022) | [Cite this article](#)

9418 Accesses | **2** Citations | **243** Altmetric | [Metrics](#)

Abstract

Research software is a fundamental and vital part of research, yet significant challenges to discoverability, productivity, quality, reproducibility, and sustainability exist. Improving the practice of scholarship is a common goal of the open science, open source, and FAIR (Findable, Accessible, Interoperable and Reusable) communities and research software is now being understood as a type of digital object to which FAIR should be applied. This emergence reflects a maturation of the research community to better understand the crucial

Download PDF



Sections

References

[Abstract](#)

[Introduction](#)

[Results](#)

[Discussion](#)

[Methods](#)

[Data availability](#)

[Code availability](#)

[References](#)

[Acknowledgements](#)

[Author information](#)