

# Research outputs beyond the paper

*Code, research software & computational workflows*

NBIS Data Management Team

[data-management@scilifelab.se](mailto:data-management@scilifelab.se)

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



<https://doi.org/10.17044/scilifelab.22249429>

# Beyond the paper



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.”

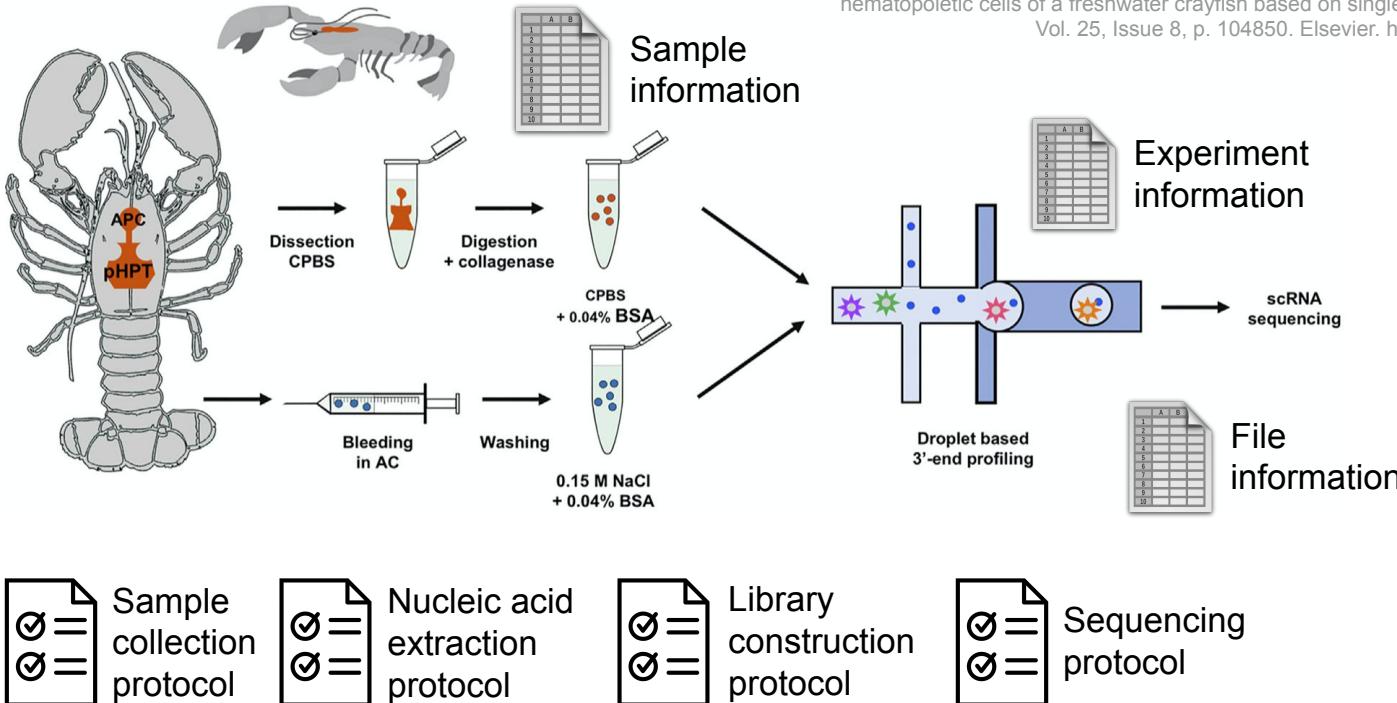
– Jason Priem, 2013

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

# Shared models, methods & protocols



"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



**Sequence File Formats | FASTQ**

illumina Search Illumina.com

INFORMATICS Overview Infrastructure Setup Sequencing Data Analysis Biological Interpretation

## File Formats for Illumina Sequencing

Numerous options are provided 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

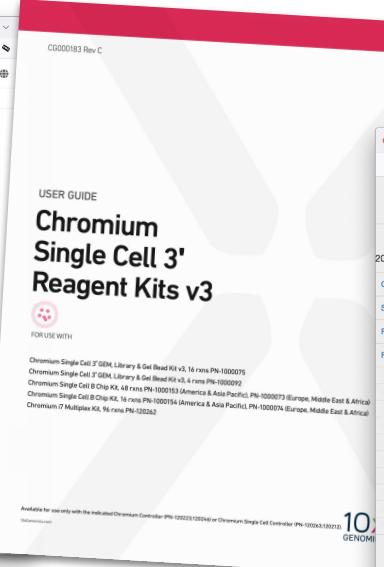
FASTQ is a text-based sequencing data file format that stores both raw sequence data and quality scores. FASTQ files have become the standard format for storing NGS data from Illumina sequencing systems, and can be used as input for a wide variety of secondary data analysis solutions.

The MiSeq 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 parse the fastq.ora files using standard FASTQ parsing tools or a range of popular mapping tools such as BWA,<sup>1</sup> STAR<sup>2</sup> and Bowtie.<sup>3</sup> DRAGEN.ORA compression is available with the DRAGEN server and on-board the NextSeq1000/2000.



M Report for project TJ-2700 on runfolder 201126\_A00605\_0172\_BHVTNDRXX

General Stats  
Sequencing Metadata  
FastQC  
Sequence Quality Histograms  
Per Sequence Quality Scores  
Per Base Sequence Content  
Per Sequence GC Content  
Per Base N Content  
Sequence Length Distribution  
Overrepresented sequences  
Adapter Content  
Status Checks

## MultiQC

Report for project TJ-2700 on runfolder 201126\_A00605\_0172\_BHVTNDRXX

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](mailto:seq@medsci.uu.se)

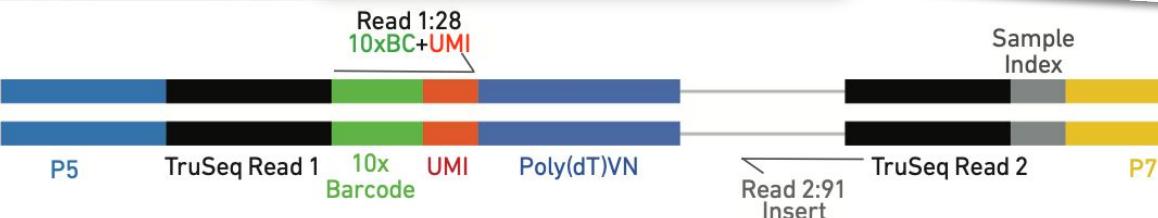
Report generated on 2020-11-27, 01:07 based on data in: /sequreports-data/nxf\_work/7b/d1a47e223927cc74172e0713e8f7a8

Welcome! Not sure where to start? Watch a tutorial video (0:06) don't show again

### 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

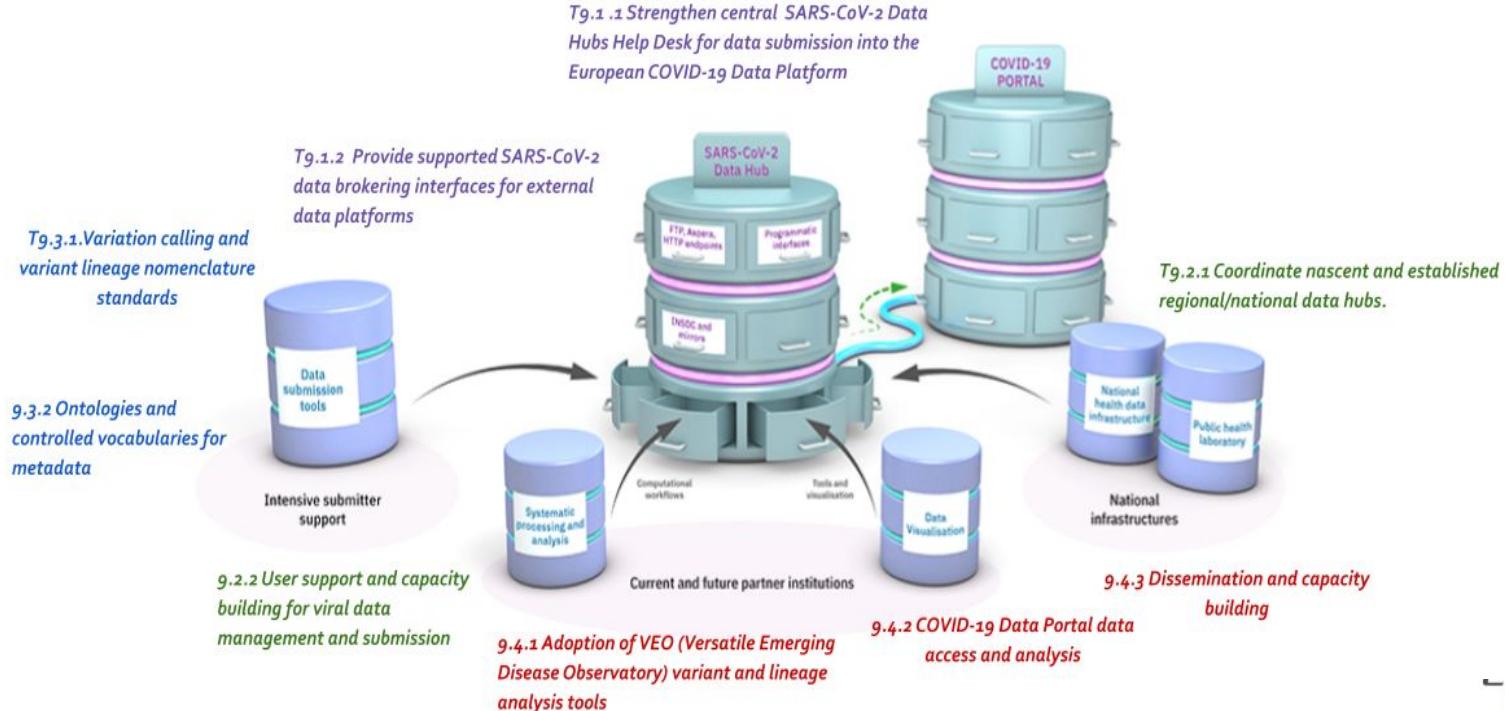
- [201126\\_A00605\\_0172\\_BHVTNDRXX\\_TJ-2700\\_multicqc\\_report.html](#)
- [checksums.md5](#)
- [SampleSheet.csv](#)



# Research information resources

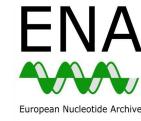
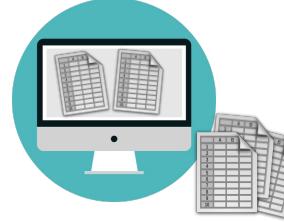
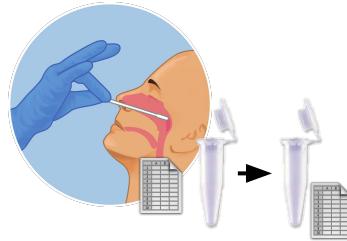
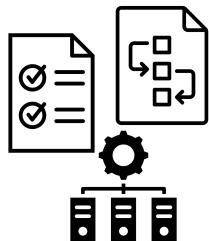


[https://rdmkit.elixir-europe.org/data\\_brokering](https://rdmkit.elixir-europe.org/data_brokering)



# 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, ...

# Managing software and data



- 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**



# Procedures for quality control



Slides 17–22, Data collection at NBIS, <https://doi.org/10.17044/scilifelab.21557151>

The data do you have and their current characteristics

**NBIS** What data do you rely on?

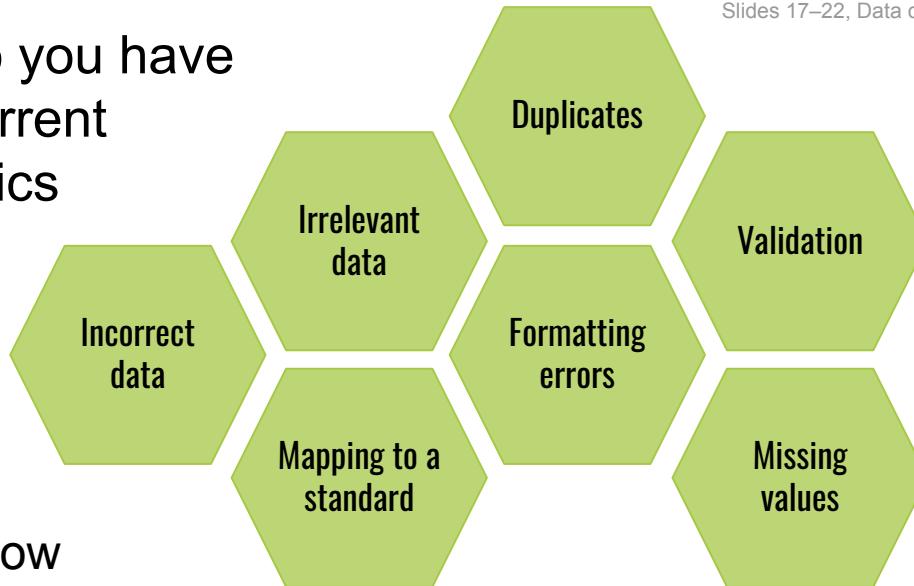
- Existing data to access Where and under what conditions are the data available? How and when will you get access to them?
- New data to be created What will be measured? Where? By whom? And using what instrumentation/methods?

16 November 2020 | Data collection at NBIS | <https://doi.org/10.17044/scilifelab.21557151> | Slides 17

**NBIS** What are their characteristics?

- Kind of data Sequencing, numeric, textual, images, video, etc.
- Data and file formats What formats will you receive, get the data in, or use for downstream analysis? What does it need to use it?
- Expected volumes What are the expected sizes, rows, columns, files and/or their sizes?

16 November 2020 | Data collection at NBIS | <https://doi.org/10.17044/scilifelab.21557151> | Slides 18



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

**NBIS** What conventions do you adopt?

- Community standards Data standards and terminologies used across related domains, e.g., data models
- Data organisation guidelines Project conventions for naming, location and versioning files and documents
- Documentation for reuse Data distribution, protocols, pipelines, analysis transcripts...

16 November 2020 | Data collection at NBIS | <https://doi.org/10.17044/scilifelab.21557151> | Slides 19

**NBIS** What platforms to you use?

- Storage/processing locations For data collection, analysis, reporting, code, transfers etc.
- Backup and data recovery Data protection (against data-loss and data corruption) (levels of backup and external storage)
- Tools and software Software and systems required to access / process the data?

16 November 2020 | Data collection at NBIS | <https://doi.org/10.17044/scilifelab.21557151> | Slides 20

**NBIS** What is done for quality assurance?

- Traceability / provenance Instrument settings and calibration, standardized data output, data processing pipelines, software and outputs?
- Data validation Data validation methods, peer review of data, or representation with controlled vocabularies?

16 November 2020 | Data collection at NBIS | <https://doi.org/10.17044/scilifelab.21557151> | Slides 21

**NBIS** What are the related policies?

- Information classification Suitable storage based on the characteristics of the data
- Access controls Who will have access to what data and how can it be enforced?
- Data protection procedures Other strategies to mitigate risks of unwanted data disclosure or leakage

16 November 2020 | Data collection at NBIS | <https://doi.org/10.17044/scilifelab.21557151> | Slides 22

Typical workflow

Inspecting

Harmonising

Verifying

Reporting/Documenting

Detect unexpected, incorrect, and inconsistent data, etc.

Fix or remove anomalies, transform, convert etc.

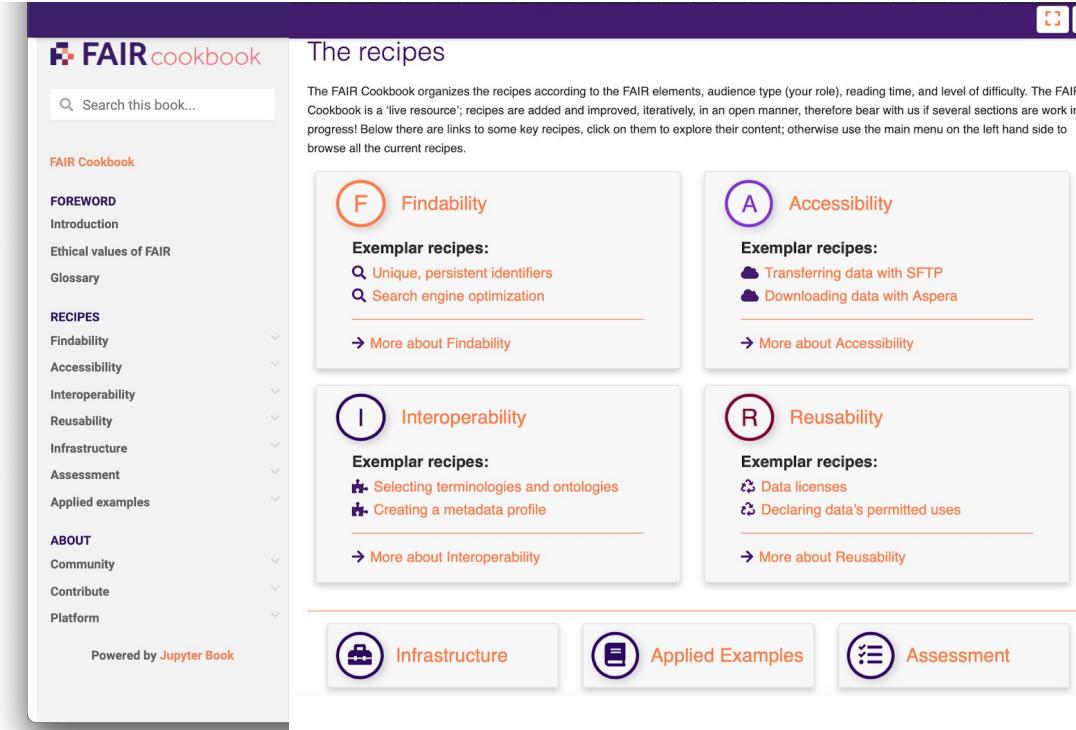
Test that the results are complete and correct.

Record changes made and quality assessments.

# Recipes for processing data



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



The FAIR Cookbook organizes the recipes according to the FAIR elements, audience type (your role), reading time, and level of difficulty. The FAIR Cookbook is a 'live resource'; recipes are added and improved, iteratively, in an open manner, therefore bear with us if several sections are work in progress! Below there are links to some key recipes, click on them to explore their content; otherwise use the main menu on the left hand side to browse all the current recipes.

- F Findability**
  - Exemplar recipes:**
    - Q Unique, persistent identifiers
    - Q Search engine optimization
  - [More about Findability](#)
- A Accessibility**
  - Exemplar recipes:**
    - Cloud Transferring data with SFTP
    - Cloud Downloading data with Aspera
  - [More about Accessibility](#)
- I Interoperability**
  - Exemplar recipes:**
    - Hand Selecting terminologies and ontologies
    - Hand Creating a metadata profile
  - [More about Interoperability](#)
- R Reusability**
  - Exemplar recipes:**
    - Hand Data licenses
    - Hand Declaring data's permitted uses
  - [More about Reusability](#)
- Infrastructure**
- Applied Examples**
- Assessment**



- 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!**

# Computational tools and services



covid19.galaxyproject ✘

search name, annotation, owner, and tag:



[https://usegalaxy.org/workflows/list\\_published](https://usegalaxy.org/workflows/list_published)

Advanced Search

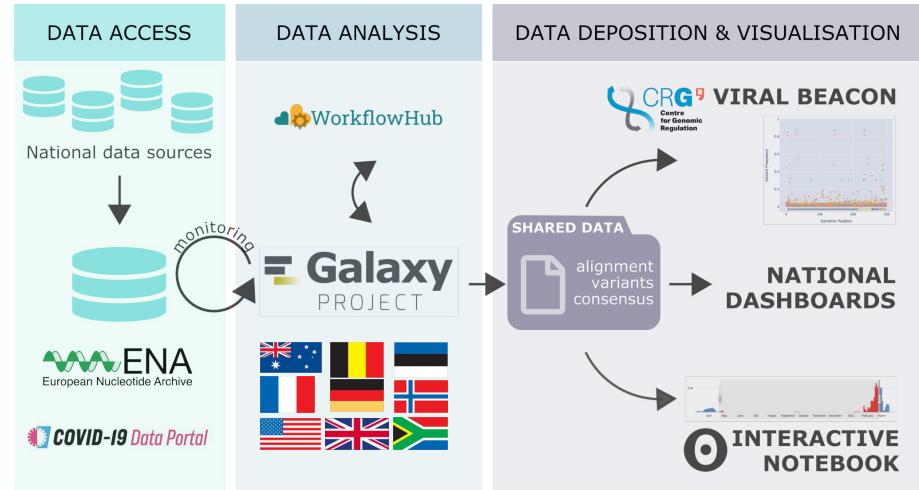
Name	Annotation	Owner	Community Rating	Community Tags
COVID-19: variation analysis on WGS SE data	Call variants from WGS (non-ampliconic) single-end reads.	wolfgang-maier		covid-19 covid19.galaxyproject.org
COVID-19: variation analysis on ARTIC PE data	Call variants from ampliconic paired-end reads.	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

# Platforms linking data and services



<https://galaxyproject.org/eu/>

- 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   = Galaxy



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



# Coding skills for researchers



<https://carpentries.org>



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



## FIVE RECOMMENDATIONS FOR FAIR SOFTWARE

LET'S GO! →

ENDORSE





<http://software.ac.uk/which-journals-should-i-publish-my-software>



Software  
Sustainability  
Institute

About

Programmes and Events

Resources

search



## 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



### Tags

- Neil Chue Hong
- Publications
- Journals

# Curated computational workflows



<https://nf-co.re>

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

# nf-core

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 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. Secondarily, 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>

10 Things

News

Data Quality Review

CURE Training

Get Involved!



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](#)