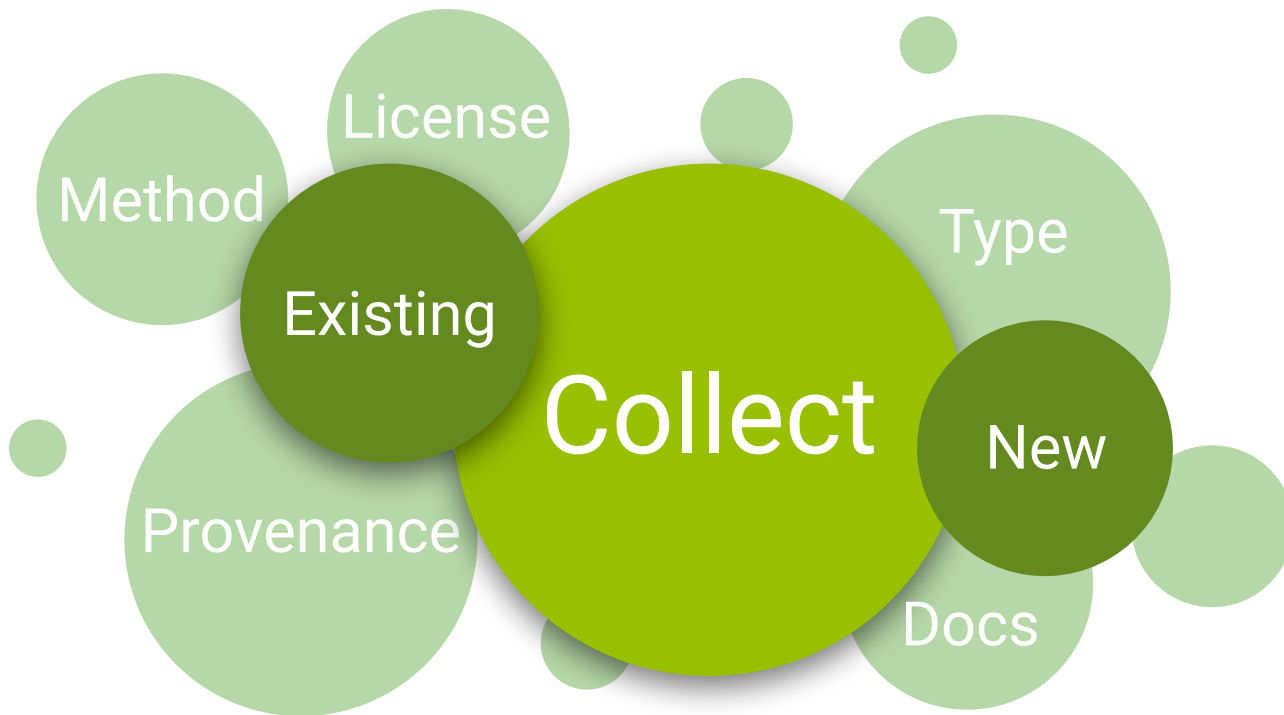


Data collection

NBIS Data Management Team
data-management@scilifelab.se

<https://doi.org/10.17044/scilifelab.c.6820587>





The RDMkit data lifecycle: Collecting



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



- **What is data collection?**
- **Why is data collection important?**
the collection phase lays the foundation for the quality of both the data and its documentation

- **What should be considered for data collection?**

- Capture the provenance e.g. of samples, researchers and instruments.
- Define the experimental design including a collection plan (e.g. repetitions, controls, randomisation) in advance.
- Calibrate the instruments.
- Find suitable repository to store the data.
- Identify suitable metadata standards.

Your tasks

Data organisation

Best practices to name and organise research data.

Data quality

How to ensure high quality of research data.

Existing data

How to find and reuse existing data.

Identifiers

How to use identifiers for research data.

Documentation and metadata

How to document and describe your data.

Data sensitivity

How to identify the sensitivity of different research data types

Data storage

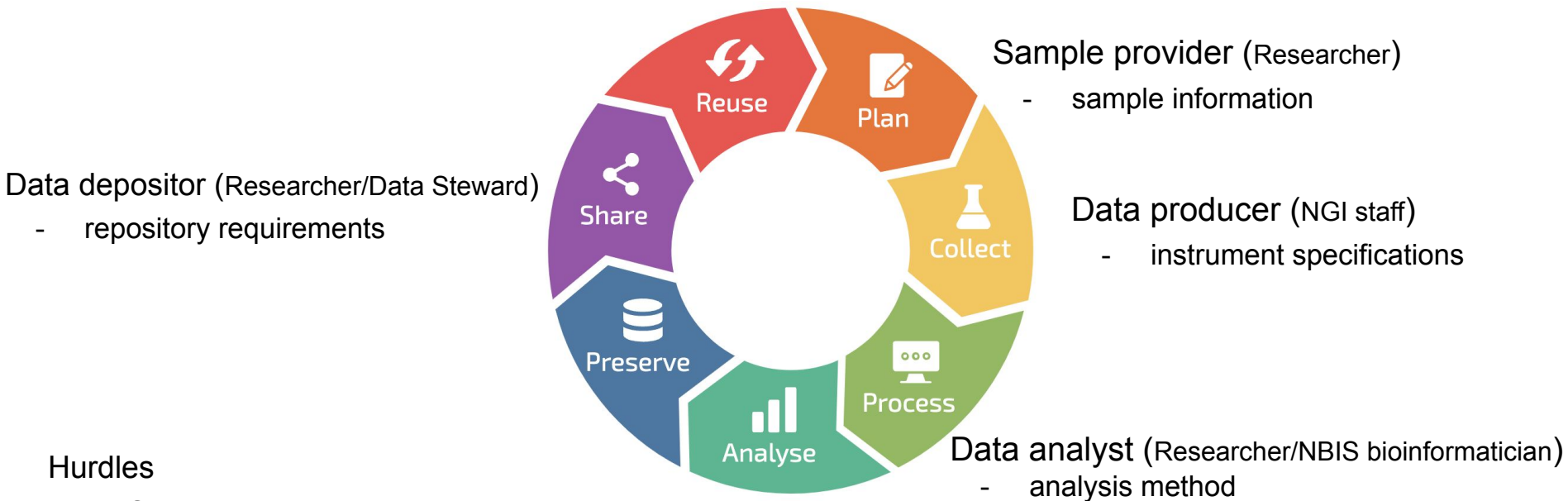
How to find appropriate storage solutions.

Data provenance

How to record information about data provenance.

Data collection

different perspectives and responsibilities



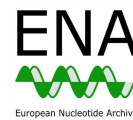
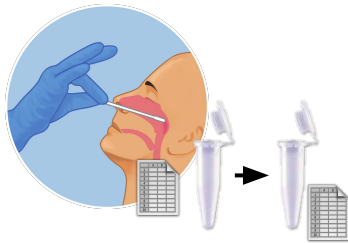
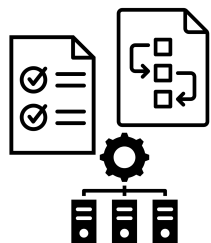
Hurdles

- Shared terminology (ontology)
- File formats for data and documentation (metadata)
- Data delivery mechanisms

Moving towards FAIR by design



“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, ...

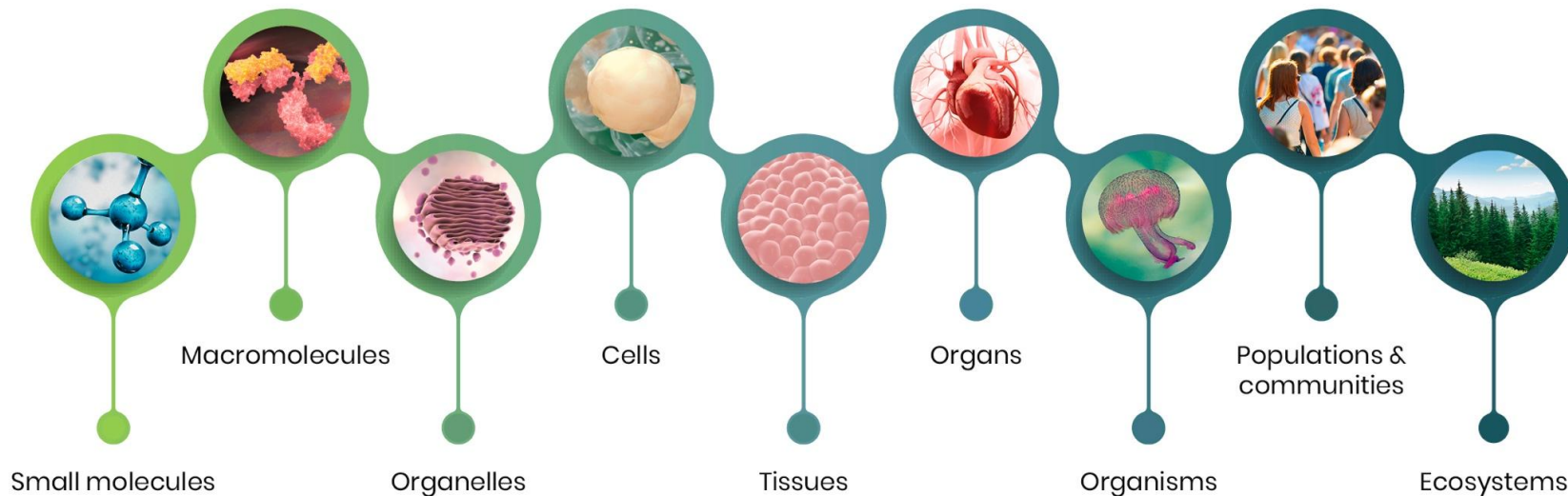
Data and computational workflows

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, ...

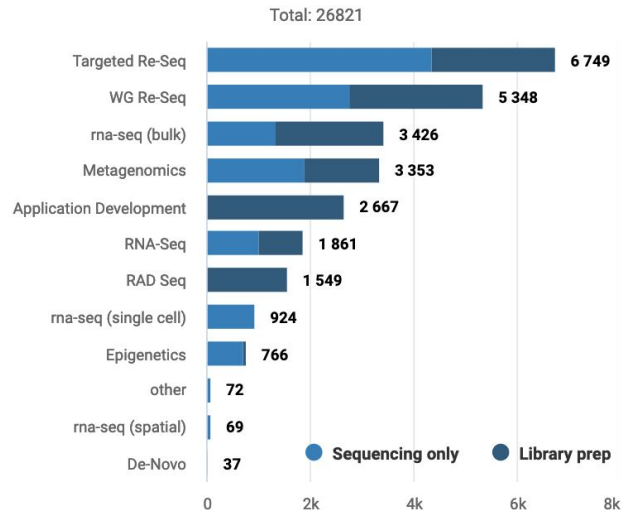
SciLifeLab generates a variety of data





- Instruments include *Ion Torrent*, *Illumina*, *Pacific Biosciences*, and *Oxford Nanopore*
- Automated quality checks and analyses in a number of applications
- Samples must be prepared to fulfill the requirements for the method or kit

Samples in 2022

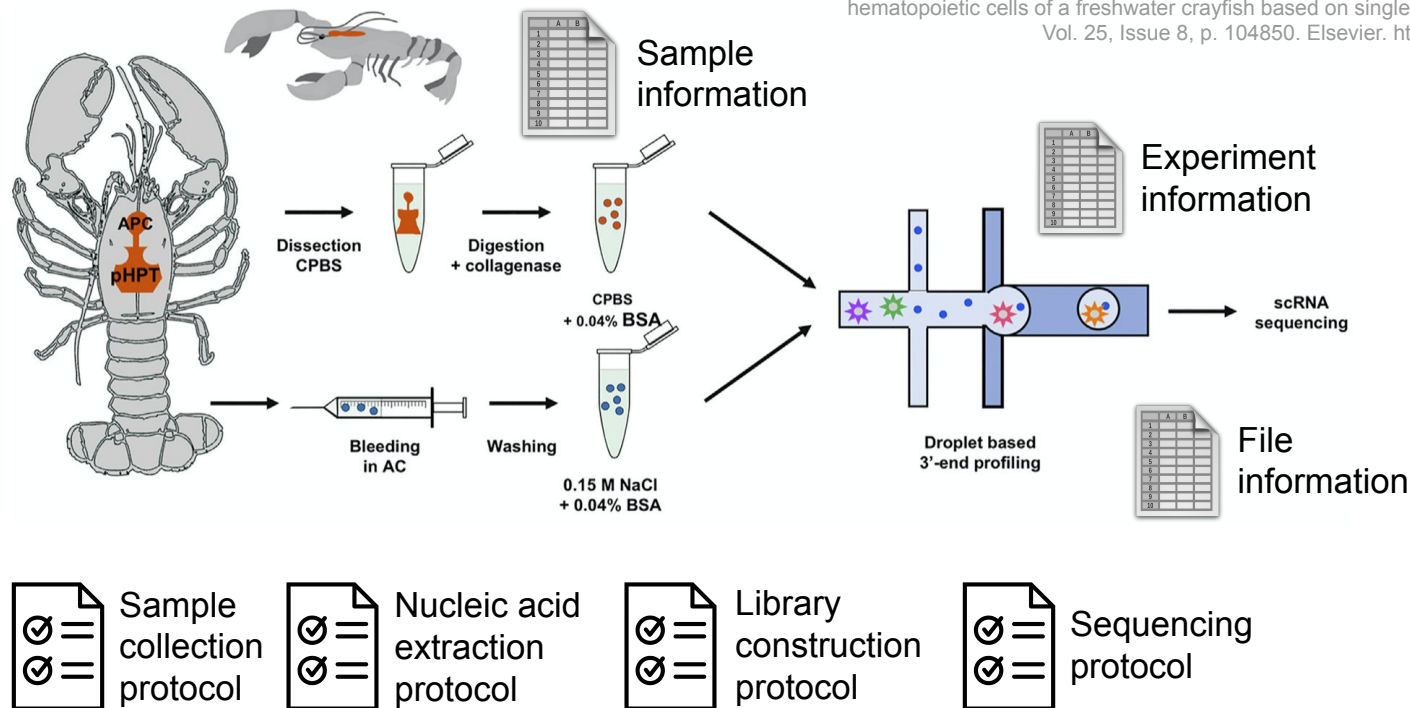


Single-cell RNA sequencing example



"Protocol" icon by Justin Blake from thenounproject.com

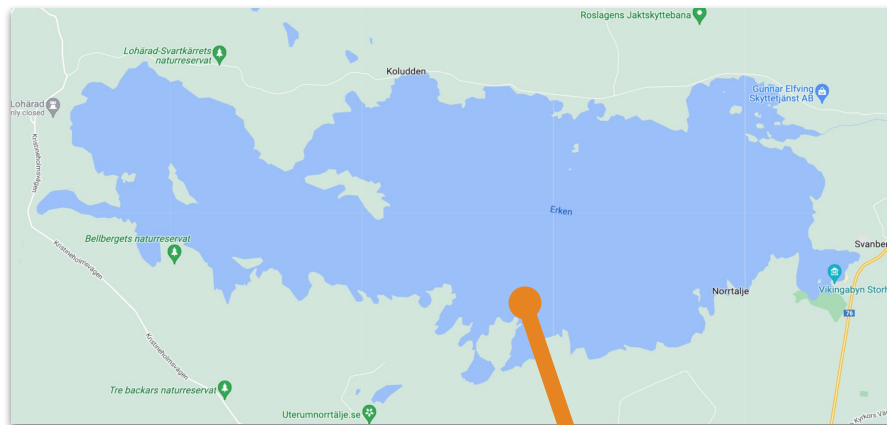
Söderhäll, I., Fasterius, E., Ekblom, C., & Söderhäll, K. (2022). Characterization of hemocytes and hematopoietic cells of a freshwater crayfish based on single-cell transcriptome analysis. In *iScience* Vol. 25, Issue 8, p. 104850. Elsevier. <https://doi.org/10.1016/j.isci.2022.104850>



- 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

- Obtain freshwater crayfish (adult males) from lake Erken, Sweden (59.8 N 18.6 E)
- Maintain in the crayfish facility in running tap water, 10-12°C, 12:12 light:dark cycle...
- Feed once a week

Map data © Google. Crayfish derived from Illustration in Söderhäll et al. (2022).



Map data © Google



Slide 10

Sample information as data



ENA Checklist: ERC00001 – ENA default sample checklist

OLS / Experimental Factor Ontology

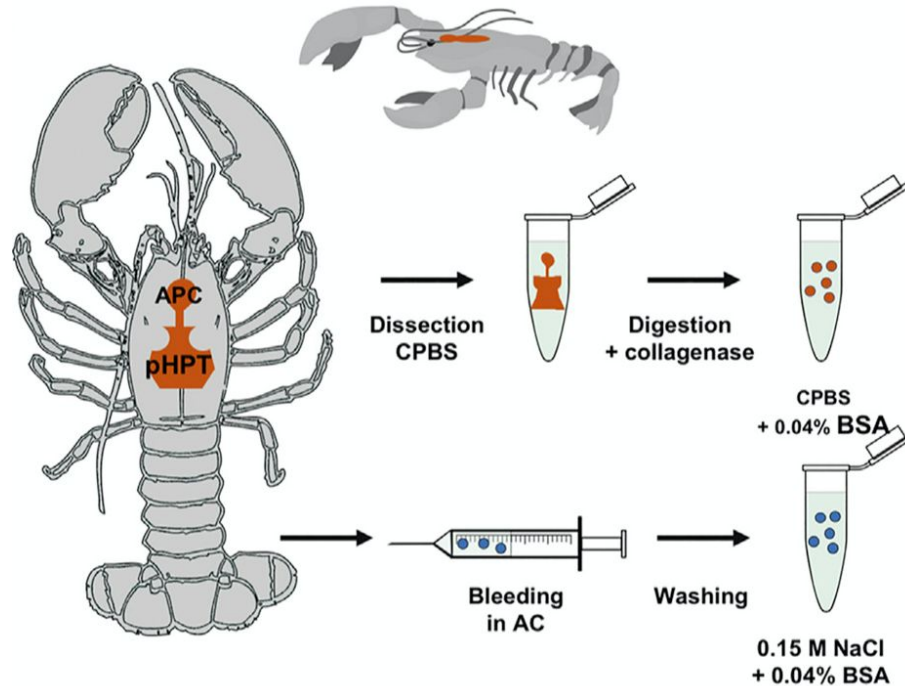
EFO

Population	Organism	Pacifastacus leniusculus
	Genotype	wild type genotype
	Geographical location	Sweden, Lake Erken
Individual	Growth condition	laboratory aquarium since Sep 2020
	Sampling date	2020-11-06
	Developmental stage	adult
	Body weight	35
	Sex	male
Specimen	Organism part	hematopoietic system
Sample	Cell type	hemocyte

Nucleic acid extraction protocol



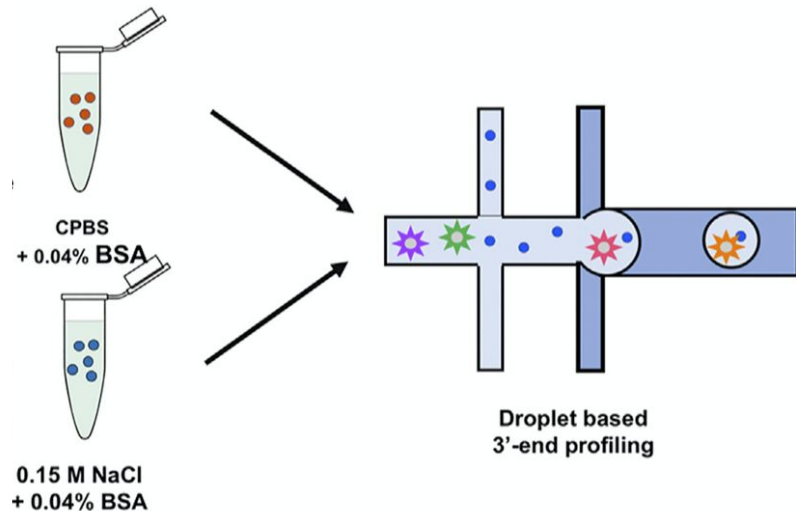
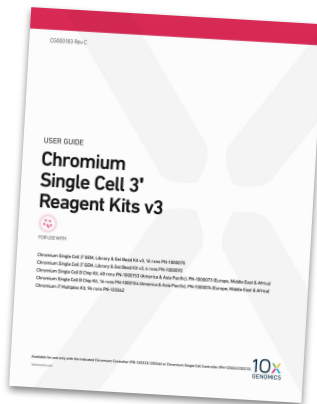
Protocol illustration derived from Illustration in Söderhäll et al. (2022).



- Dissect and digest into single cells by incubation in 300 µl of 0.1% collagenase ... at room temperature for 20 min on a rotating plate ... then filtered through a 40 mm cell strainer
- Pool isolated cells from four animals for scRNA-seq

- Prepare sequencing libraries using Chromium Single Cell 3' reagent kit v3 (cat# 1000075/1000073/120262, 10xGenomics)
- According to the manufacturer's protocol

CG000183
Single Cell 3' Reagent Kit
User Guide, v3 chemistry,
10xGenomics



Sequencing protocol



- 28+8+0+91 bp read length
- NovaSeq 6000 system
- SP flowcell
- v1 sequencing chemistry
- Include a sequencing library for the phage PhiX as 1% spike-in in the sequencing run

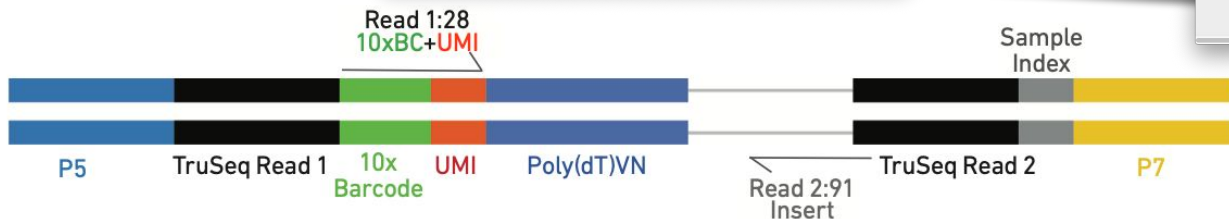
Files delivered by NGI



- ▼ 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

The image shows three overlapping screenshots. The leftmost is the Illumina website's 'File Formats for Illumina Sequencing' page, which lists various sequencing formats and provides links to download them. The middle screenshot is the 'USER GUIDE' for 'Chromium Single Cell 3' Reagent Kits v3', showing the product packaging and a list of compatible kits. The rightmost screenshot is a 'MultiQC' report for project TJ-2700 on the runfolder 201126_A00605_0172_BHVVTNDRXX. The report includes a table of general statistics for each sample.

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

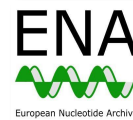
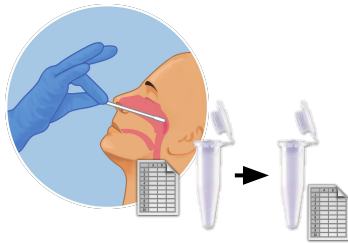
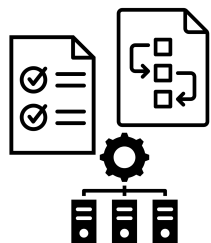


- 201126_A00605_0172_BHVVTNDRXX_TJ-2700_multiqc_report.html
- checksums.md5
- SampleSheet.csv

Moving towards FAIR by design



“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

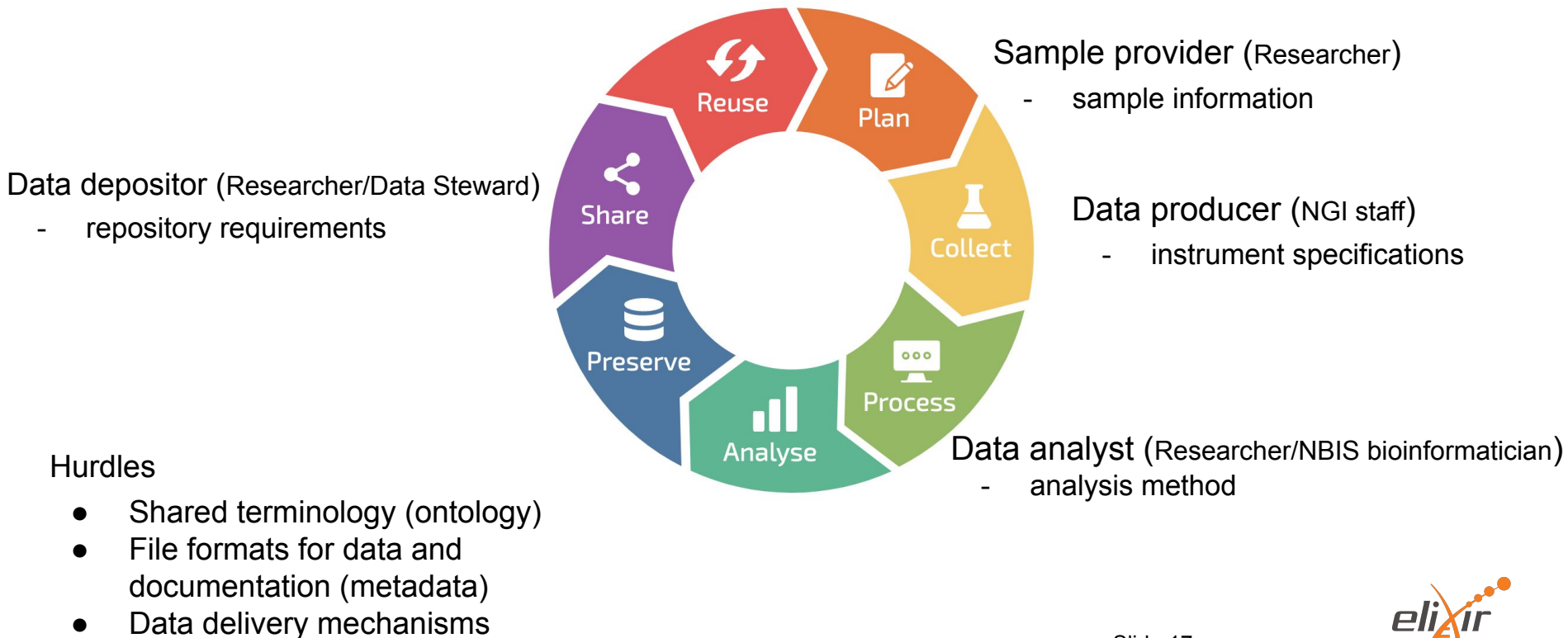
digital processing steps,
working storage conditions,
long-term storage location,
data quality assessment,
sample/data annotations,
reference data,
analysis method...

Outputs

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

Data collection

different perspectives and responsibilities





1. What data do you need to collect from your users?
2. What data do you need to collect from your providers?
(instrument manufacturers/pipeline maintainers etc.)
3. What data do your users need to collect from you or your providers?
 - Are you capturing/documenting all this data?
 - How can you make that data more FAIR and open?
 - General data vs project specific data
4. What data do your indirect users need to collect from you?
(NBIS bioinformaticians, data stewards etc.)