From Data Chaos to Data Harmony: Managing NGS Data in a Wet Lab

Jose Alejandro Romero Herrera

DeIC conference 2023-11-08

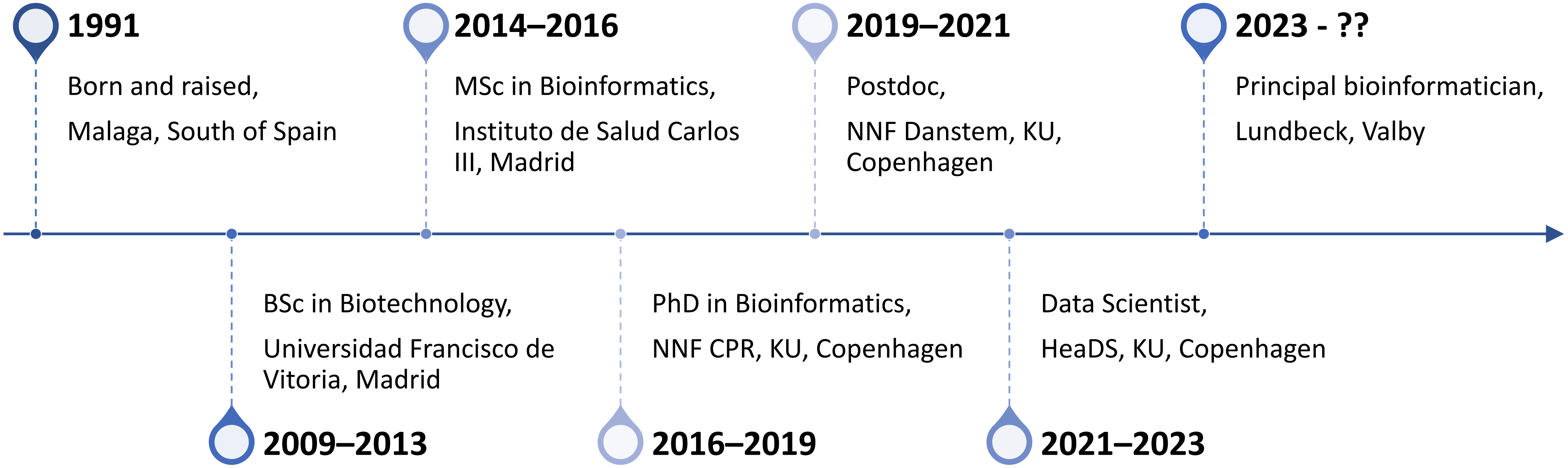
UNIVERSITY OF COPENHAGEN

HeaDS

Agenda

- About
- Intro
- The problem
- Simple solutions
- Future work
- Acknowledgments

About me



1991

Born and raised,

Malaga, South of Spain

2009–2013

BSc in Biotechnology,

Universidad Francisco de Vitoria, Madrid

2014–2016

MSc in Bioinformatics,

Instituto de Salud Carlos III, Madrid

2016–2019

PhD in Bioinformatics,

NNF CPR, KU, Copenhagen

2019–2021

Postdoc,

NNF Danstem, KU, Copenhagen

2021–2023

Data Scientist,

HeaDS, KU, Copenhagen

2023 - ??

Principal bioinformatician,

Lundbeck, Valby

About HeaDS

- Center for Health Data Science, KU
  - Strengthen health data science at KU
  - Serve as a hub for researchers
  - Provides consulting services
  - Teaching in various topics

Anders Krogh

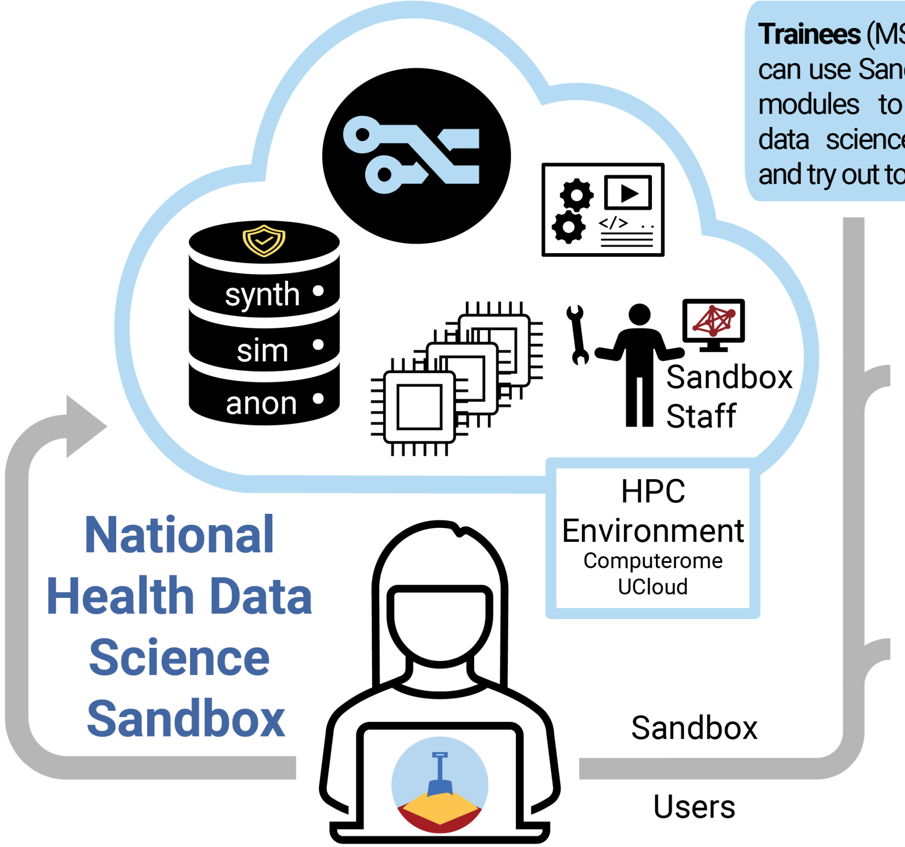
UCPH HeaDS

- Working for the Sandbox project
  - Multi-disciplinary
  - Multi-institutional
  - Virtual Environment for Health Data Science

Jennifer Bartell

National Health Data Science Sandbox

About Sandbox

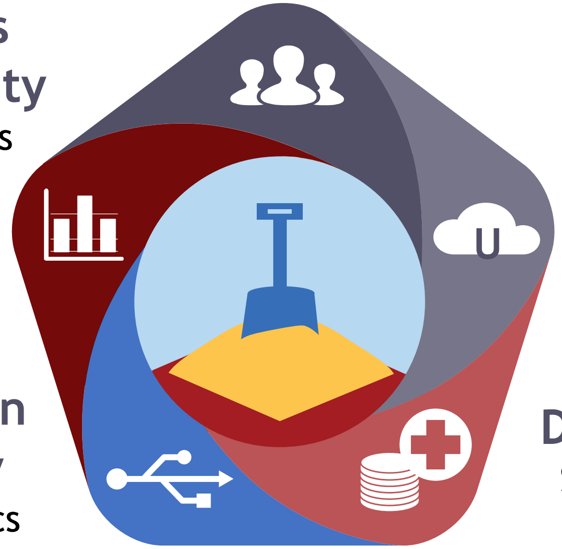


National Health Data Science Sandbox

Trainees (MS to professor) can use Sandbox teaching modules to learn health data science approaches and try out tools & pipelines

Researchers can access non-sensitive / synthetic health datasets for prototyping and exploratory research

Teachers can host data science courses and workshops with setup and support from Sandbox data scientists



Aarhus University
Genomics

Southern Denmark University
Proteomics

Copenhagen University
Transcriptomics

Denmark DTU
Supercomputing

Aalborg University
Health records
& predictive modeling

UNIVERSITY OF COPENHAGEN

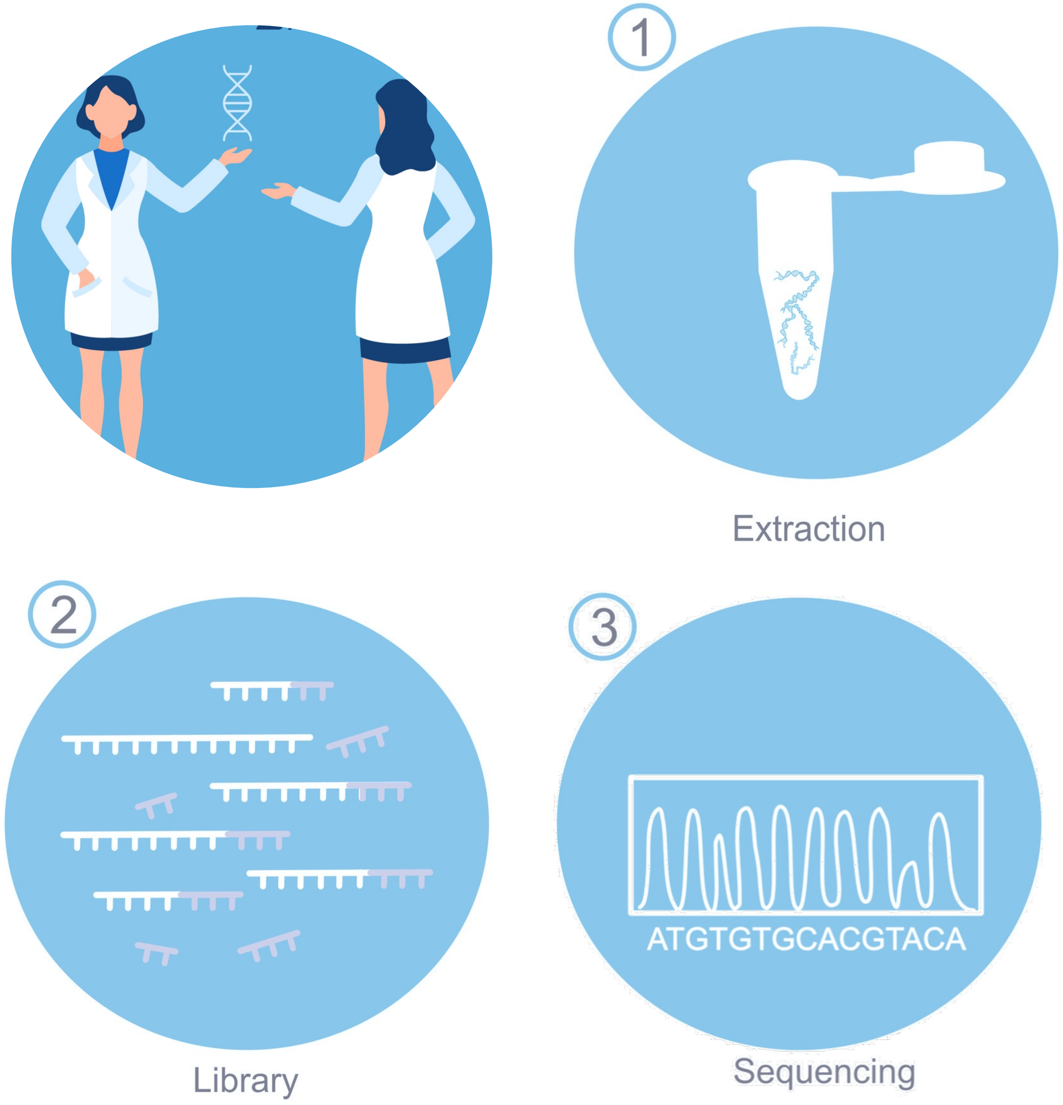
HeaDS

NGS data

- Next Generation Sequencing Data
  - Determine genetic information

- Applications:
  - Genomic research
  - personalized medicine
  - forensic science, and more

- Millions of DNA sequences generated
  - Really big datasets (Gb per sample)

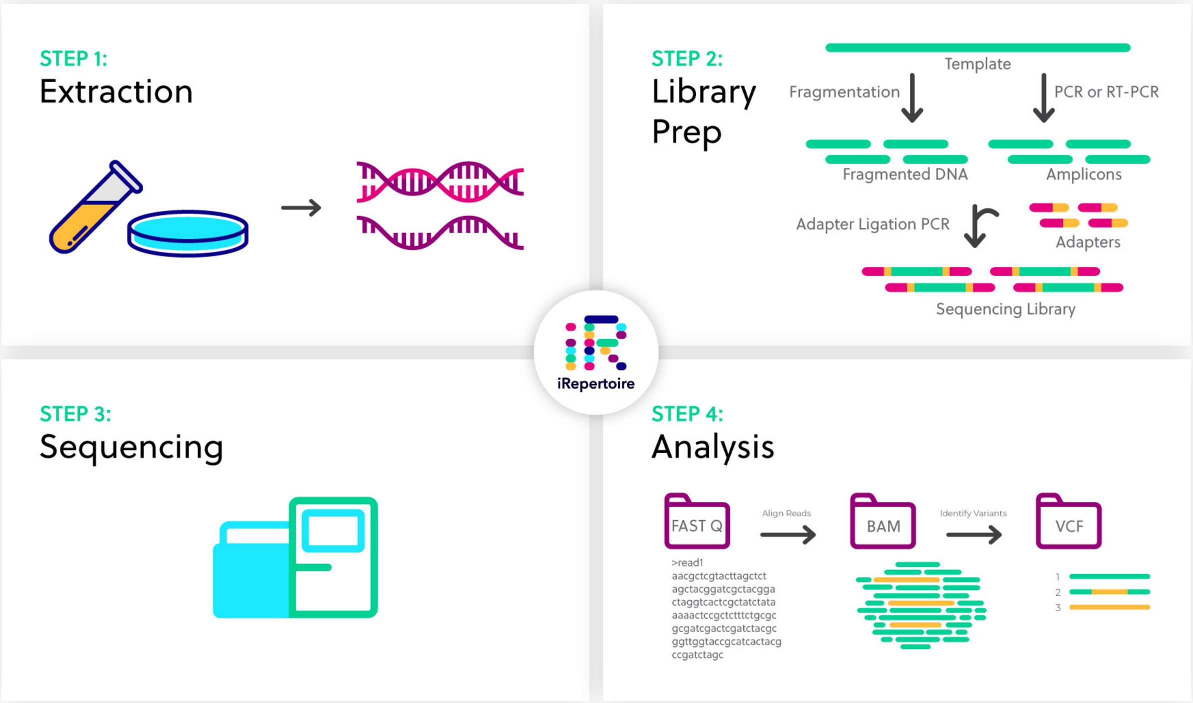


RDM for wet lab

Labguru

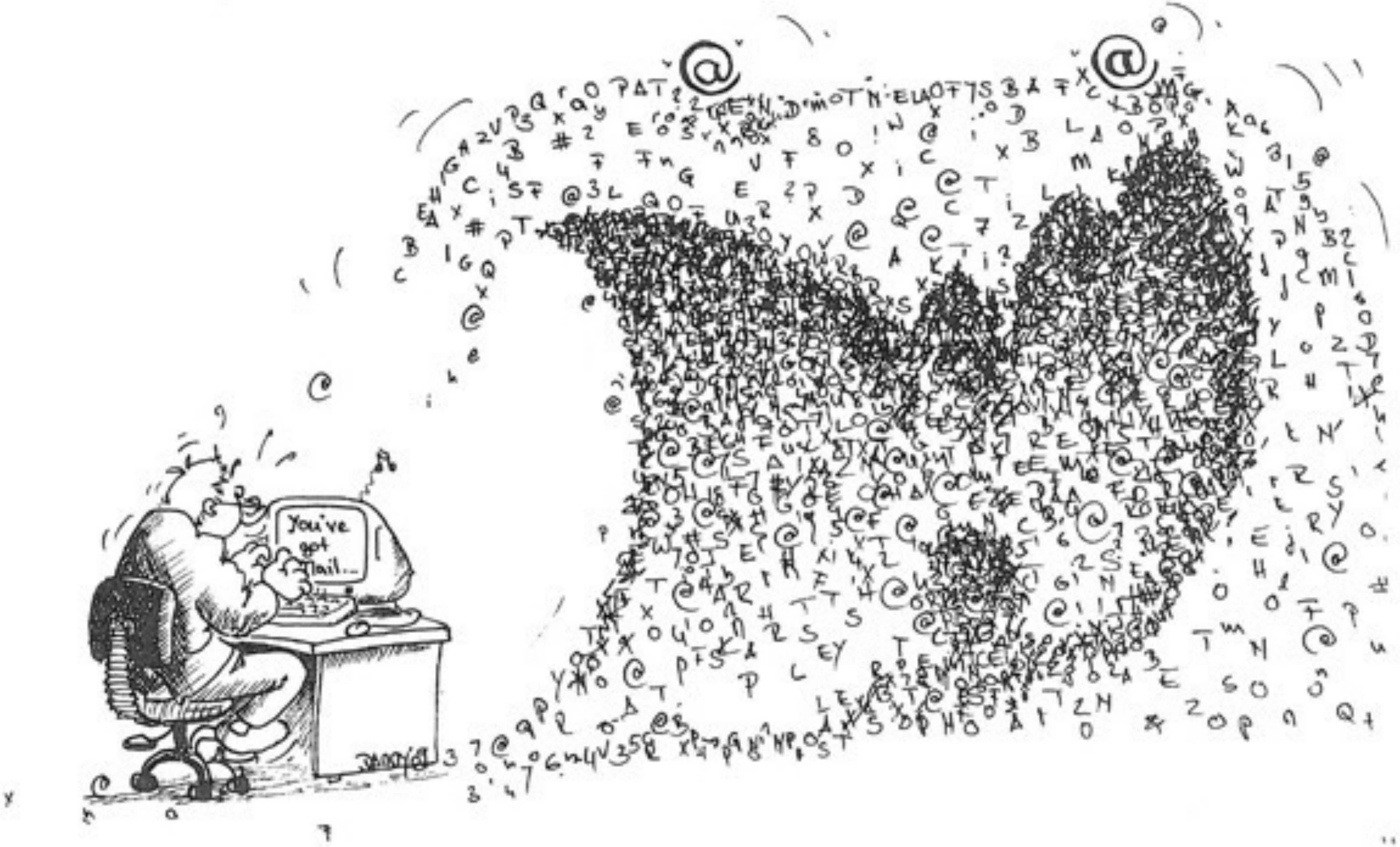
Benchling

- RDM supported options
  - ELNs and LIMS
- Tracks:
  - Experiments, protocols, samples
  - Order of chemicals
  - Instruments and physical storage



- **Bioinformatics data and its analysis is not tracked at all!**

Bioinformatics in a wet lab



Data chaos

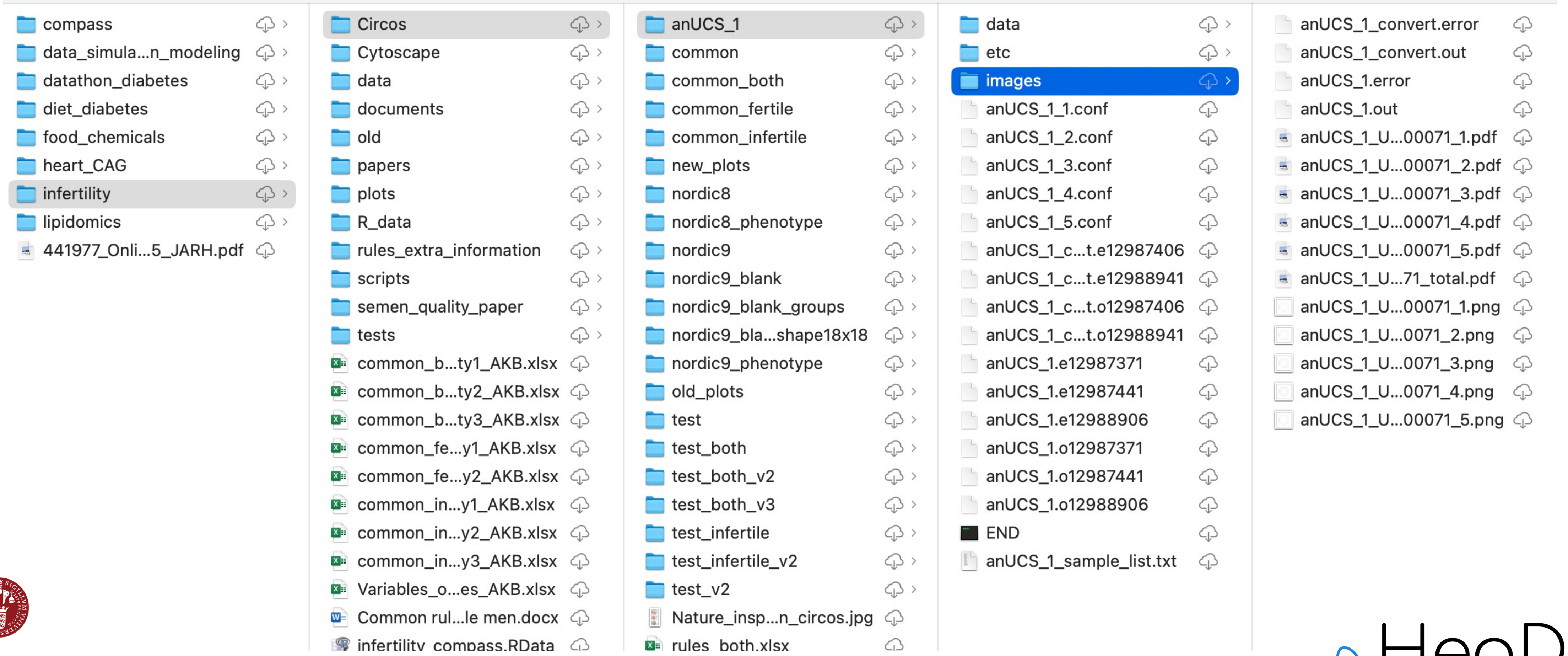
- Data accumulated for several years
- No common structure
- No common file naming conventions
- No metadata or documentation
- No provenance
- Massive problems when staff leaves

Digital version of a chaotic desktop

Wonderlane on Unsplash

Data chaos

Me, myself and only I to blame



compass
data_simula...n_modeling
datathon_diabetes
diet_diabetes
food_chemicals
heart_CAG
infertility
lipidomics
441977_Onli...5_JARH.pdf

Circos
Cytoscape
data
documents
old
papers
plots
R_data
rules_extra_information
scripts
semen_quality_paper
tests
common_b...ty1_AKB.xlsx
common_b...ty2_AKB.xlsx
common_b...ty3_AKB.xlsx
common_fe...y1_AKB.xlsx
common_fe...y2_AKB.xlsx
common_in...y1_AKB.xlsx
common_in...y2_AKB.xlsx
common_in...y3_AKB.xlsx
Variables_o...es_AKB.xlsx
Common rul...le men.docx
infertility_compass.RData

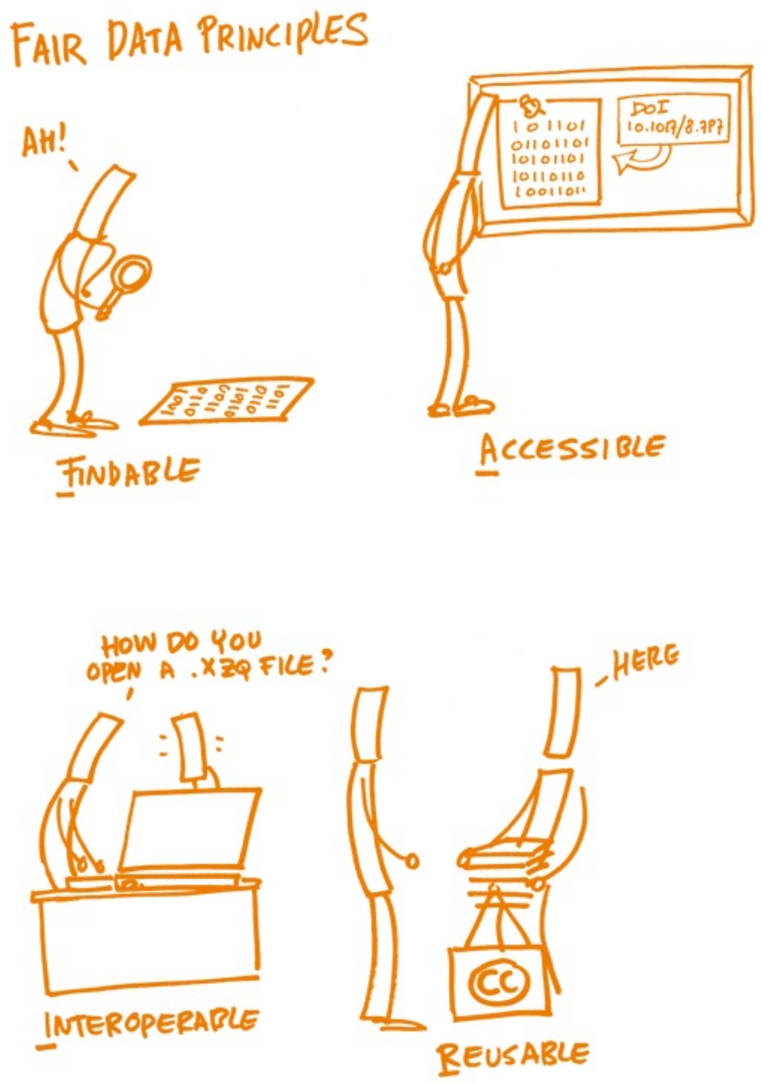
anUCS_1
common
common_both
common_fertile
common_infertile
new_plots
nordic8
nordic8_phenotype
nordic9
nordic9_blank
nordic9_blank_groups
nordic9_bla...shape18x18
nordic9_phenotype
old_plots
test
test_both
test_both_v2
test_both_v3
test_infertile
test_infertile_v2
test_v2
Nature_insp...n_circos.jpg
rules_both.xlsx

data
etc
images
anUCS_1_1.conf
anUCS_1_2.conf
anUCS_1_3.conf
anUCS_1_4.conf
anUCS_1_5.conf
anUCS_1_c...t.e12987406
anUCS_1_c...t.e12988941
anUCS_1_c...t.o12987406
anUCS_1_c...t.o12988941
anUCS_1.e12987371
anUCS_1.e12987441
anUCS_1.e12988906
anUCS_1.o12987371
anUCS_1.o12987441
anUCS_1.o12988906
END
anUCS_1_sample_list.txt

anUCS_1_convert.error
anUCS_1_convert.out
anUCS_1.error
anUCS_1.out
anUCS_1_U...00071_1.pdf
anUCS_1_U...00071_2.pdf
anUCS_1_U...00071_3.pdf
anUCS_1_U...00071_4.pdf
anUCS_1_U...00071_5.pdf
anUCS_1_U...71_total.pdf
anUCS_1_U...00071_1.png
anUCS_1_U...0071_2.png
anUCS_1_U...0071_3.png
anUCS_1_U...0071_4.png
anUCS_1_U...00071_5.png

Solution: FAIR principles

- Many guidelines on FAIR principles
- Not much applied to NGS data
  - Some few examples, but not applications
- Simple guidelines for any bioinformatician
  - Basic command line experience
  - Version control with git/GitHub



Simple rules for NGS RDM

1. Adhere to **folder structures** and **naming conventions** using **templates**
2. Fill a **DMP template** that it is **prefilled** with common information
3. Create and fill **metadata file** and **README file** in each folder
4. Make **a browsable database** from all metadata files
5. Use **community-curated workflows** for data preprocessing
6. **Version control** data analysis with git/Github
7. Display **data analysis reports** with GitHub Pages
8. **Archive** data and data analysis in **repositories** (Zenodo-GEO/Annotare)

1. Folder templates and naming conventions

- Create custom templates using COOKIECUTTER

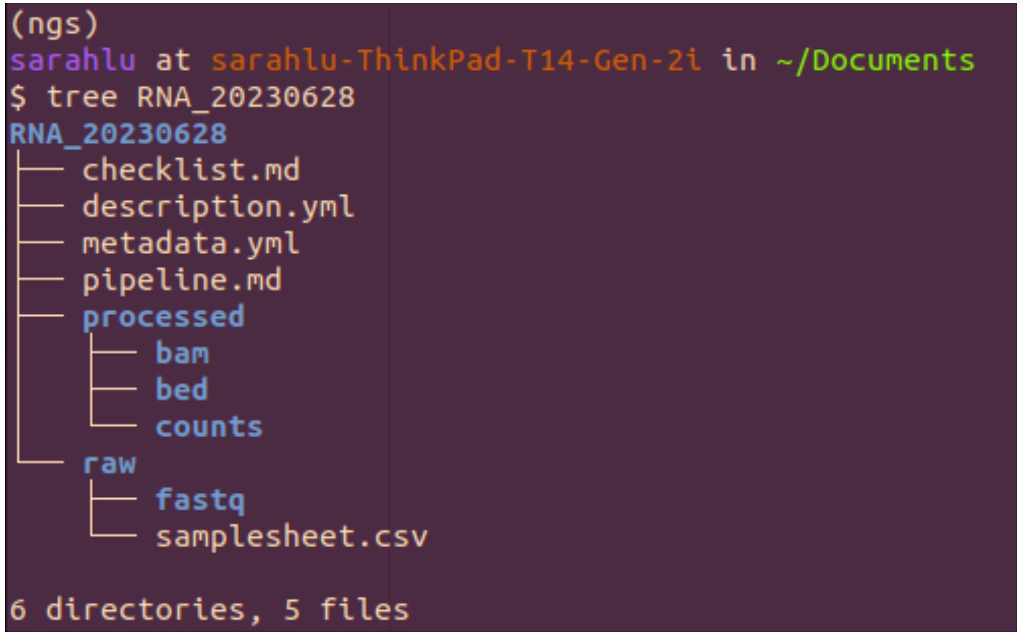
- Command line utility
  - Very flexible
  - Simple, but can do complex things

- Two templates
  - *Assay* folders: NGS data
  - *Project* folders: Data analyses

1. Folder templates and naming conventions

Data folder ("assays")

- Subfolders → NGS experiment data
- Read only, no duplicates
- Unique ID, human readable
- Metadata file
  - ID, keywords, tech, author, date, etc.
  - Controlled vocabularies
- README.md: additional details
- Data organization:
  - raw, processed, pipeline
  - pipeline: community curated workflows



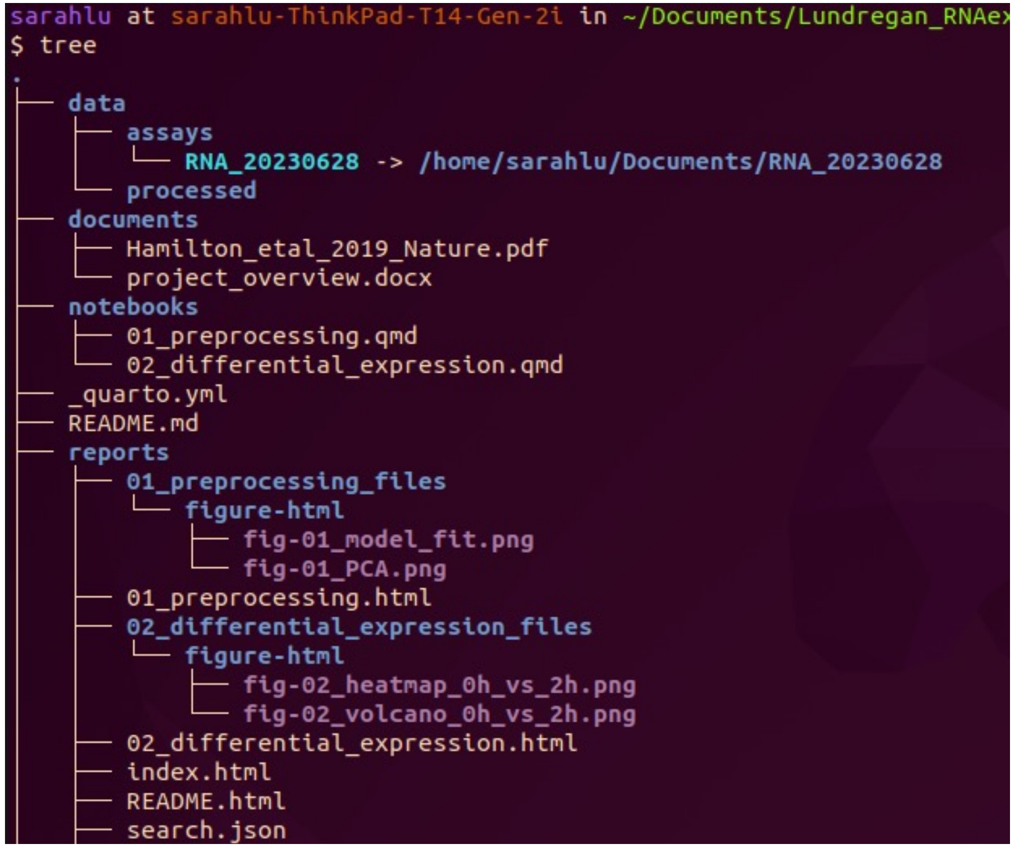
```
(ngs)
sarahlu at sarahlu-ThinkPad-T14-Gen-2i in ~/Documents
$ tree RNA_20230628
RNA_20230628
├── checklist.md
├── description.yml
├── metadata.yml
├── pipeline.md
├── processed
│   ├── bam
│   ├── bed
│   └── counts
└── raw
    ├── fastq
    └── samplesheet.csv

6 directories, 5 files
```

1. Folder templates and naming conventions

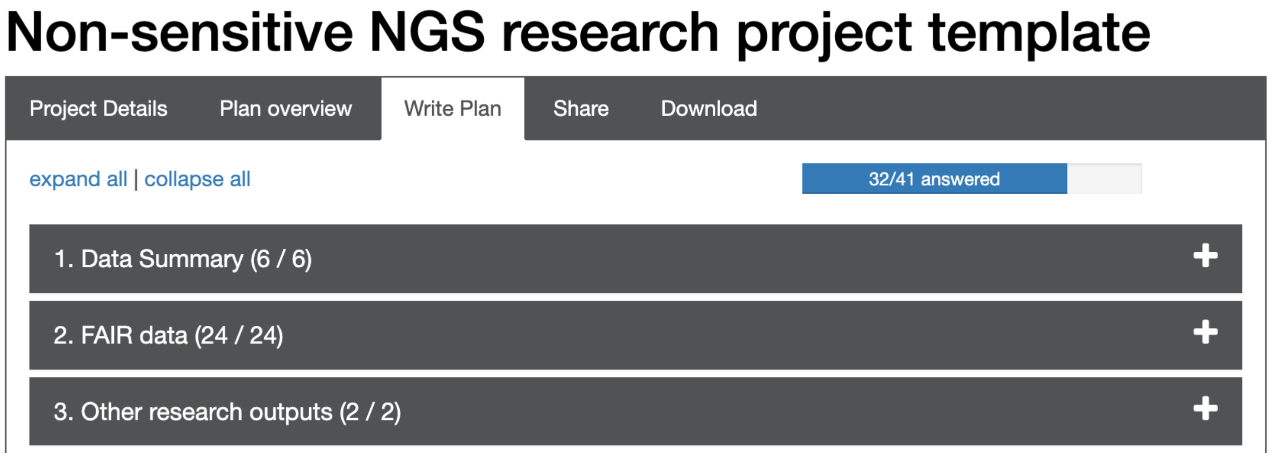
Project folder (“projects”)

- Subfolders → research project
- Simlink multiple *Assays*, no copies!
- Version controlled
- ID: AUTHOR_DESC_YYYYMMDD
- Metadata and README file
- Folder organization:
  - data, notebooks, reports, scripts, results, docs
- Naming conventions for results and figures
  - Heatmap_DEA_TreatVsControl_YYYYMMDD.tiff



```
sarahlu at sarahlu-ThinkPad-T14-Gen-2i in ~/Documents/Lundregan_RNAex
$ tree
.
├── data
│   ├── assays
│   │   └── RNA_20230628 -> /home/sarahlu/Documents/RNA_20230628
│   └── processed
├── documents
│   ├── Hamilton_etal_2019_Nature.pdf
│   └── project_overview.docx
├── notebooks
│   ├── 01_preprocessing.qmd
│   └── 02_differential_expression.qmd
├── _quarto.yml
├── README.md
├── reports
│   ├── 01_preprocessing_files
│   │   └── figure-html
│   │       ├── fig-01_model_fit.png
│   │       └── fig-01_PCA.png
│   ├── 01_preprocessing.html
│   ├── 02_differential_expression_files
│   │   └── figure-html
│   │       ├── fig-02_heatmap_0h_vs_2h.png
│   │       └── fig-02_volcano_0h_vs_2h.png
│   ├── 02_differential_expression.html
│   ├── index.html
│   ├── README.html
│   ├── search.json
```

2. Prefilled DMP template



Non-sensitive NGS research project template

1. Data Summary (6 / 6)

2. FAIR data (24 / 24)

3. Other research outputs (2 / 2)

- Prefilled with repetitive info
  - GEO, Github, Labguru…
  - Metadata and standards used
  - How it adheres to FAIR
- Streamline the process of writing your DMP
- Shared publicly
- More templates can be created depending on data
  - Sensitive datasets
  - Other omics data
  - Imaging data

3. Metadata and README

- Cookiecutter template will require you to fill metadata fields
- Collected metadata will be saved in a metadata.yml file
- Short descriptions in README.txt

Metadata field	Convention	Example
assay_type	-	ChIP-seq
owner	<Initials>	JARH
creation_date	<YYYYMMDD>	20231108
platform	-	Illumina
organism	<Genus species>	Homo sapiens
nsamples	<integer>	9

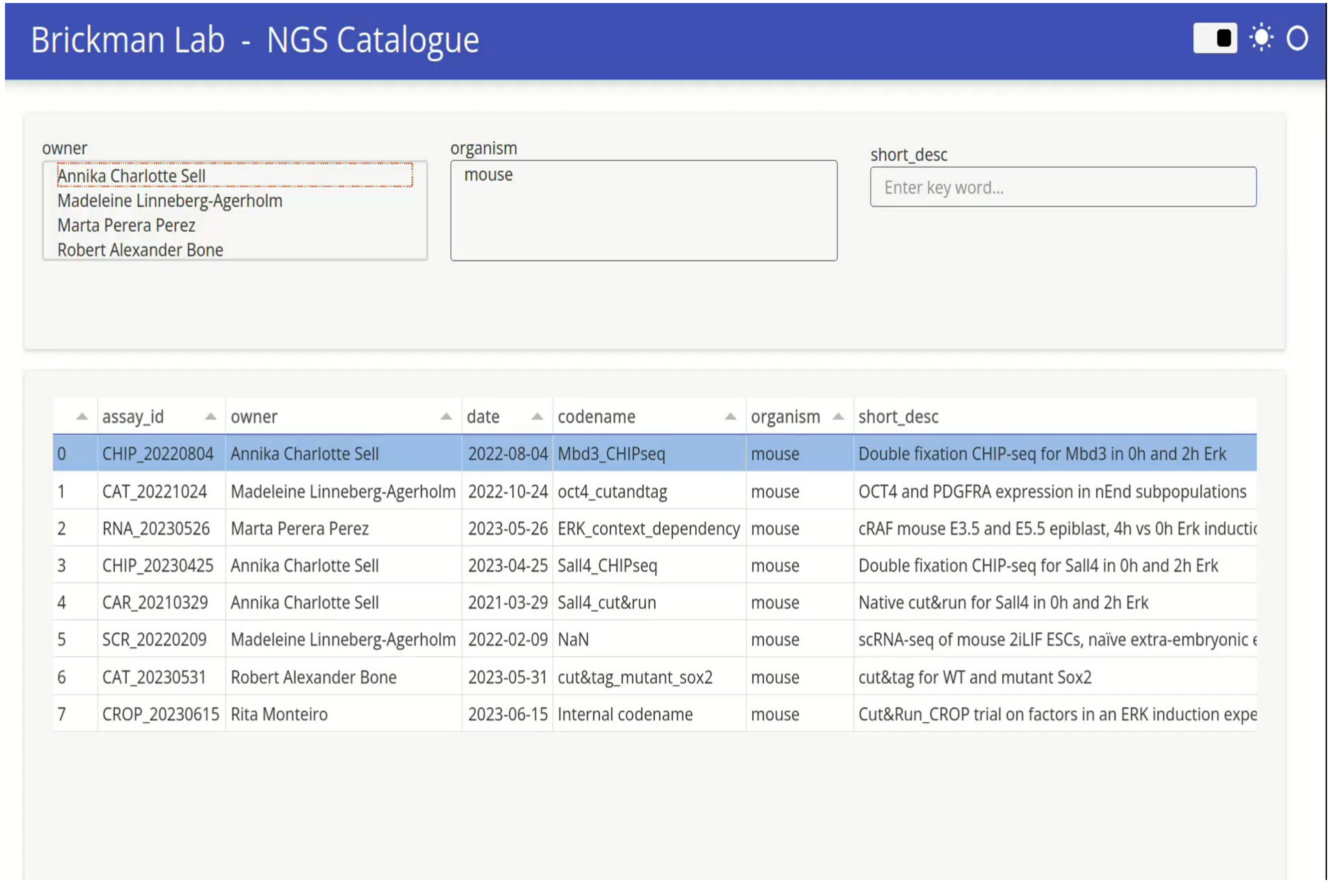
```
# NGS Analysis Project: Exploring Gene Expression in
Human Tissues

## Aims

This project aims to investigate gene expression
patterns across various human tissues using Next
Generation Sequencing (NGS) data. By analyzing the
transcriptomes of different tissues, we seek to uncover
tissue-specific gene expression profiles and identify
potential markers associated with specific biological
functions or diseases.
```

4. Browsable database

- Collect metadata files
- Create a tsv or SQL database
- Browse it with Shiny R app
- Very useful for all lab members



owner: Annika Charlotte Sell, Madeleine Linneberg-Agerholm, Marta Perera Perez, Robert Alexander Bone

organism: mouse

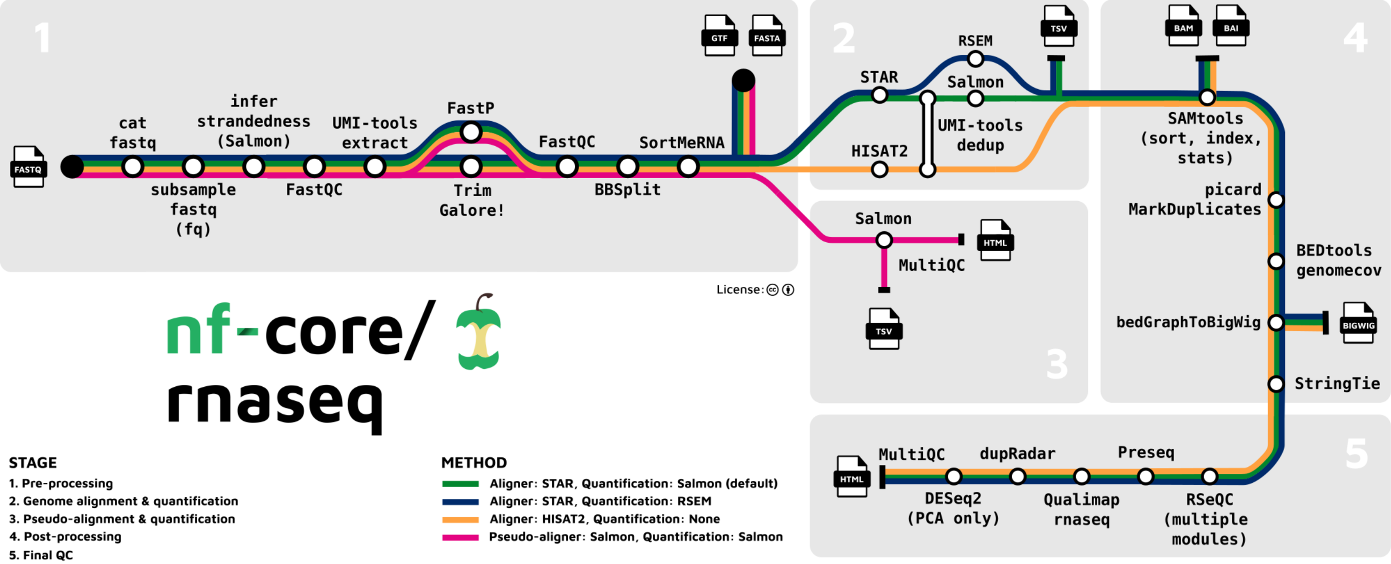
short_desc: Enter key word...

	assay_id	owner	date	codename	organism	short_desc
0	CHIP_20220804	Annika Charlotte Sell	2022-08-04	Mbd3_CHIPseq	mouse	Double fixation CHIP-seq for Mbd3 in 0h and 2h Erk
1	CAT_20221024	Madeleine Linneberg-Agerholm	2022-10-24	oct4_cutandtag	mouse	OCT4 and PDGFRA expression in nEnd subpopulations
2	RNA_20230526	Marta Perera Perez	2023-05-26	ERK_context_dependency	mouse	cRAF mouse E3.5 and E5.5 epiblast, 4h vs 0h Erk inductic
3	CHIP_20230425	Annika Charlotte Sell	2023-04-25	Sall4_CHIPseq	mouse	Double fixation CHIP-seq for Sall4 in 0h and 2h Erk
4	CAR_20210329	Annika Charlotte Sell	2021-03-29	Sall4_cut&run	mouse	Native cut&run for Sall4 in 0h and 2h Erk
5	SCR_20220209	Madeleine Linneberg-Agerholm	2022-02-09	NaN	mouse	scRNA-seq of mouse 2iLIF ESCs, naïve extra-embryonic e
6	CAT_20230531	Robert Alexander Bone	2023-05-31	cut&tag_mutant_sox2	mouse	cut&tag for WT and mutant Sox2
7	CROP_20230615	Rita Monteiro	2023-06-15	Internal codename	mouse	Cut&Run_CROP trial on factors in an ERK induction expe

5. Community-curated workflows

nf-core

- A community effort to collect a curated set of analysis pipelines
- Uses nextflow language, specialized for reproducible workflows
- Many gold-standard bioinformatics and NGS pipelines



STAGE

1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

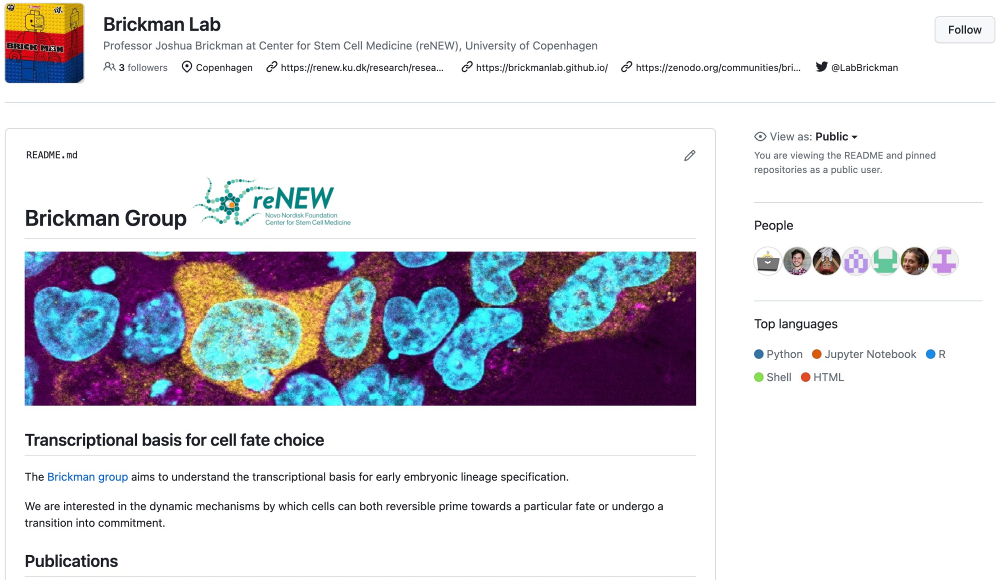
METHOD

🟩 Aligner: STAR, Quantification: Salmon (default)
🟦 Aligner: STAR, Quantification: RSEM
🟧 Aligner: HISAT2, Quantification: None
🟪 Pseudo-aligner: Salmon, Quantification: Salmon

6. Version control data analysis

GitHub

- Use version control framework for your data analysis: Git
  - NGS data is too big though
- Online repositories: GitHub
  - Create your lab organization
- Benefits
  - Enhance collaboration
  - Public sharing of the analysis of your data
  - Easier tracking of changes and results



Brickman Lab

Professor Joshua Brickman at Center for Stem Cell Medicine (reNEW), University of Copenhagen

3 followers · Copenhagen · https://renew.ku.dk/research/resea... · https://brickmanlab.github.io/ · https://zenodo.org/communities/bri... · @LabBrickman

Follow

README.md

Brickman Group

reNEW Novo Nordisk Foundation Center for Stem Cell Medicine

Transcriptional basis for cell fate choice

The Brickman group aims to understand the transcriptional basis for early embryonic lineage specification.

We are interested in the dynamic mechanisms by which cells can both reversible prime towards a particular fate or undergo a transition into commitment.

Publications

View as: Public

You are viewing the README and pinned repositories as a public user.

People

Top languages

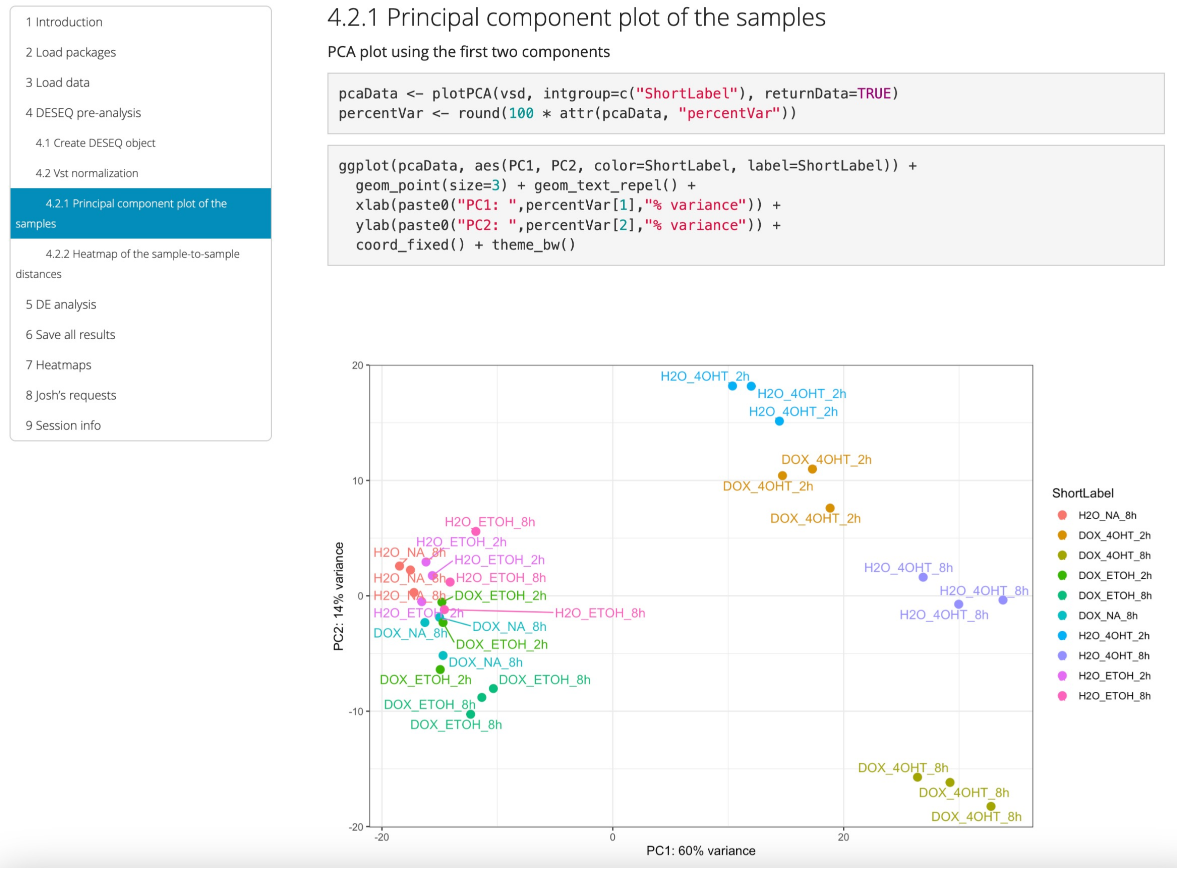
● Python ● Jupyter Notebook ● R ● Shell ● HTML

UNIVERSITY OF COPENHAGEN

HeaDS

7. Data analysis reports

- Lab webpage using GitHub Pages
  - Intro to lab
  - Feature papers or data analyses
- Display your data analysis
  - Public after publication
  - Transparency of results and analysis
- Could be used as tool documentation



1 Introduction
2 Load packages
3 Load data
4 DESEQ pre-analysis
4.1 Create DESEQ object
4.2 Vst normalization
4.2.1 Principal component plot of the samples
4.2.2 Heatmap of the sample-to-sample distances
5 DE analysis
6 Save all results
7 Heatmaps
8 Josh's requests
9 Session info

4.2.1 Principal component plot of the samples

PCA plot using the first two components

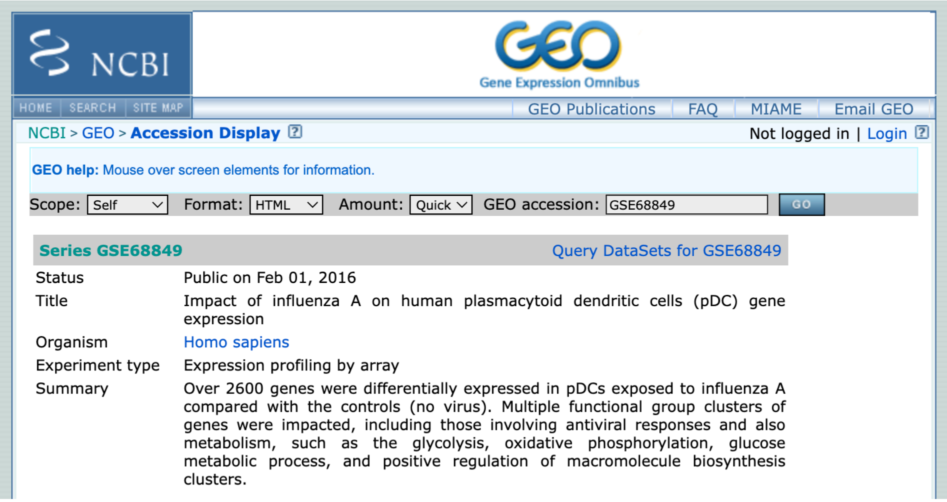
```
pcaData <- plotPCA(vsd, intgroup=c("ShortLabel"), returnData=TRUE)
percentVar <- round(100 * attr(pcaData, "percentVar"))
```

```
ggplot(pcaData, aes(PC1, PC2, color=ShortLabel, label=ShortLabel)) +
  geom_point(size=3) + geom_text_repel() +
  xlab(paste0("PC1: ",percentVar[1],"% variance")) +
  ylab(paste0("PC2: ",percentVar[2],"% variance")) +
  coord_fixed() + theme_bw()
```

8. Archiving

NGS data - Domain repository

- GEO or Annotare
- NGS metadata standards



GEO help: Mouse over screen elements for information.

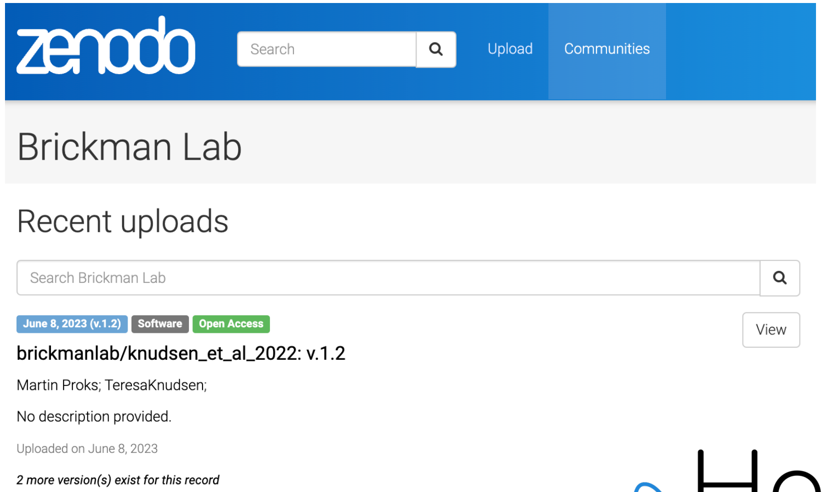
Scope: Self Format: HTML Amount: Quick GEO accession: GSE68849

Series GSE68849 Query DataSets for GSE68849

Status	Public on Feb 01, 2016
Title	Impact of influenza A on human plasmacytoid dendritic cells (pDC) gene expression
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Over 2600 genes were differentially expressed in pDCs exposed to influenza A compared with the controls (no virus). Multiple functional group clusters of genes were impacted, including those involving antiviral responses and also metabolism, such as the glycolysis, oxidative phosphorylation, glucose metabolic process, and positive regulation of macromolecule biosynthesis clusters.

Data analysis - Generic repository

- Zenodo
- DOI minting
- Synchronize with GitHub



Brickman Lab

Recent uploads

June 8, 2023 (v.1.2) Software Open Access

brickmanlab/knudsen_et_al_2022: v.1.2

Martin Proks; TeresaKnudsen;

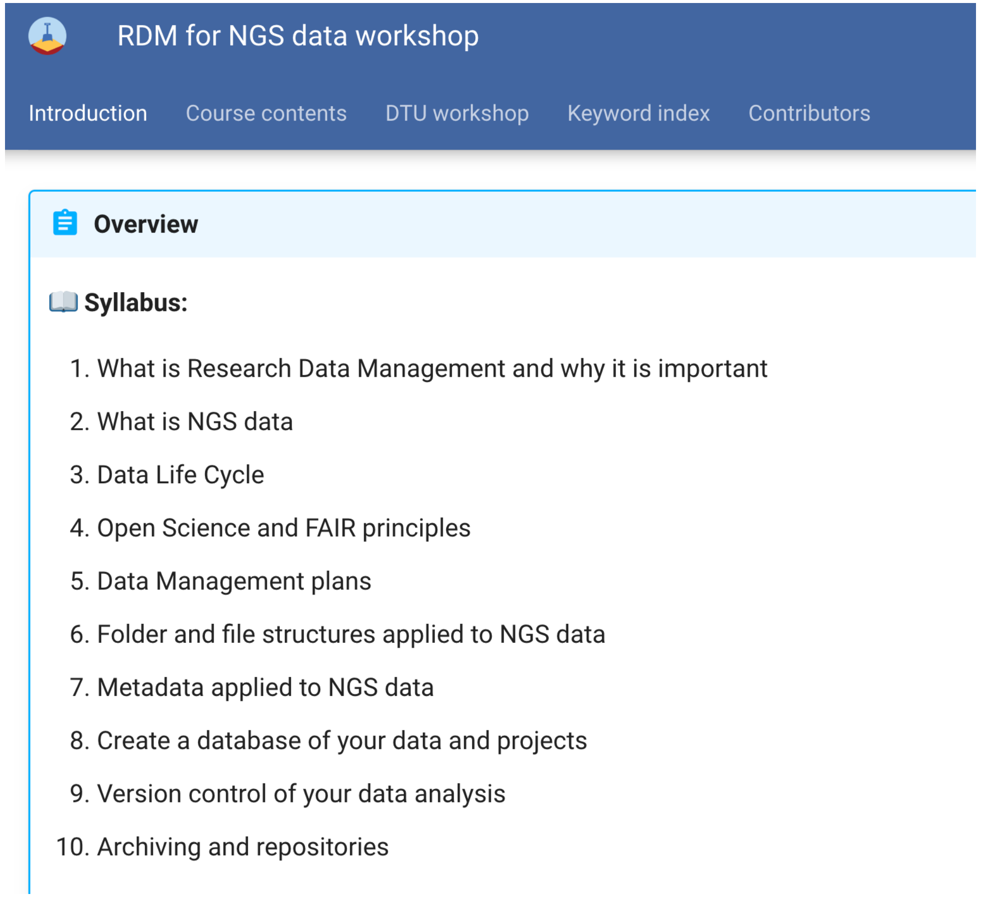
No description provided.

Uploaded on June 8, 2023

2 more version(s) exist for this record

Online course

https://hds-sandbox.github.io/RDM_NGS_course/



RDM for NGS data workshop

Introduction Course contents DTU workshop Keyword index Contributors

Overview

📖 **Syllabus:**

1. What is Research Data Management and why it is important
2. What is NGS data
3. Data Life Cycle
4. Open Science and FAIR principles
5. Data Management plans
6. Folder and file structures applied to NGS data
7. Metadata applied to NGS data
8. Create a database of your data and projects
9. Version control of your data analysis
10. Archiving and repositories

Future work

- Metadata ontologies and controlled vocabularies
  - Not easy to implement and enforce
  - Some examples are provided for different metadata fields

- Version control of the actual NGS data

- Need of command line experience and work for cookiecutter
  - Difficult for experimentalists

- Interaction between LIMS + ELN systems
  - Cross-linking data analysis with wet lab RDM
  - One place to go for everything?

Acknowledgements

Brickman lab

HeaDS center

Thank you!