Implementing Principles of Reproducible Research at Scale in a Bionformatics Core Facility: Challenges and Solutions

Hubert Rehrauer

University of Zurich UZH

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

FGCZ Genome Informatics

Vision

- Enable design to result workflows using Next Generation Sequencing
- Generate exciting results from the analysis of omics data for and with ETHZ/UZH researchers
- fill the gap (if there is any) between user skills/knowledge and existing bioinformatic tools

Notes

- Reproducibility is never asked for but considered as granted
- Nobody ever asks us about reproducibility

Mission of FGCZ Bioinformatics

Data Processing

- support **data generation** by the wet lab units of the FGCZ
- operate **data processing** infrastructure and implement **data quality control**

Data Analysis

- **operate data analysis infrastructure**
- **perform data analysis services** for ETH/UZH researchers
- collaborative data analysis within research projects
- training and education in omics areas

Reproducibility in the context of research projects

- Funding bodies set the scene for research projects

- SNF encourages reproducibility of analysis results and long-term data re-use
- Responsibility is with the PI
  - have proper strategies for data analysis and data management (DMP)
  - make sure project members apply appropriate practices
  - make sure service providers (core facilities) are compliant
- FGCZ Genome Informatics mainly acts as provider for part of the analysis trail

Research Cycle

- Bioinformatics service usually covers part of the research cycle
- FGCZ Bioinformatics
  - supports and consults on all steps
  - has full responsibility the *Data Processing* and *Data Study & Analytics*
  - supports *Planning* and *Data Preservation* according to DMPs

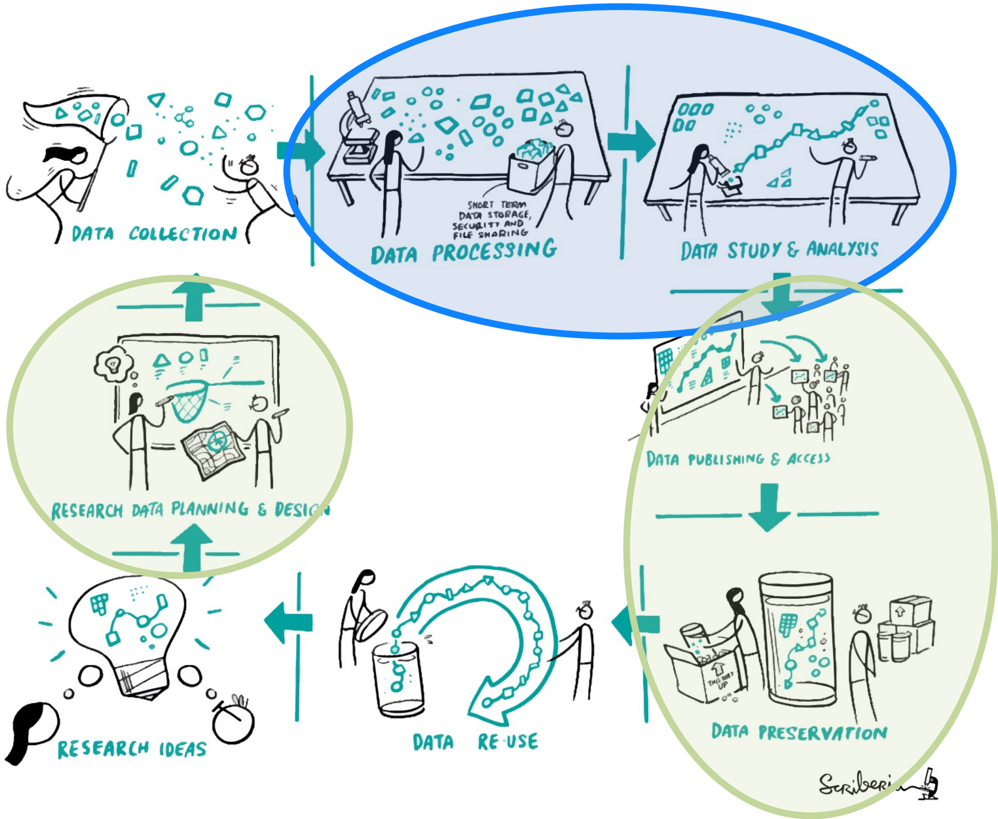
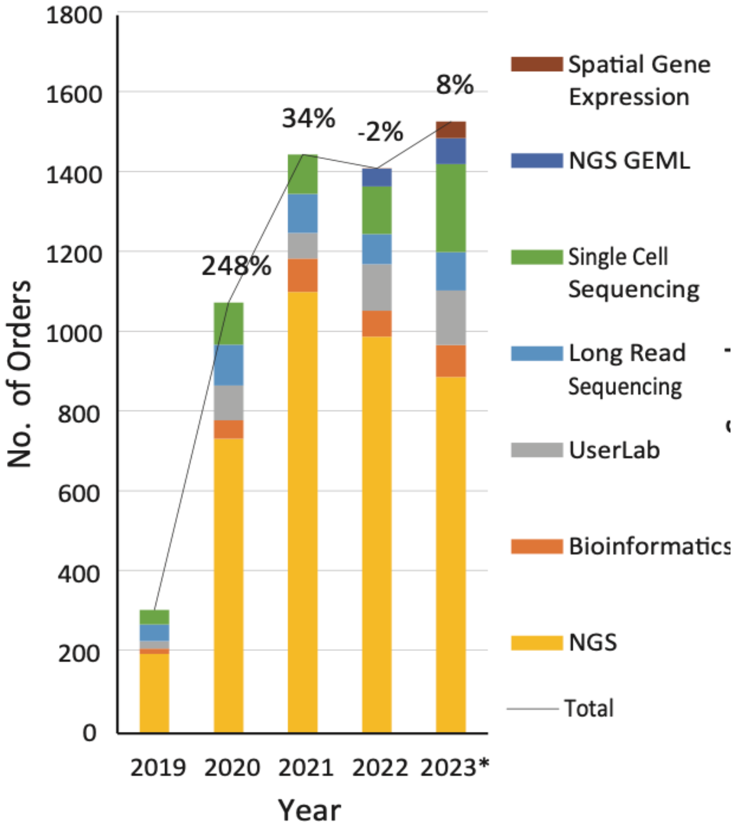
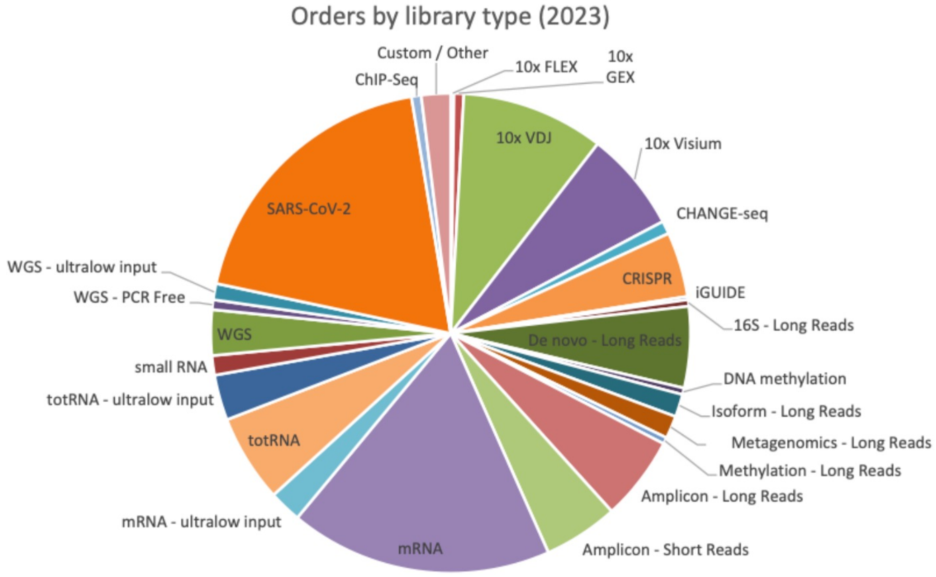


Fig. 8 The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807. #

https://the-turing-way.netlify.app/reproducible-research/overview

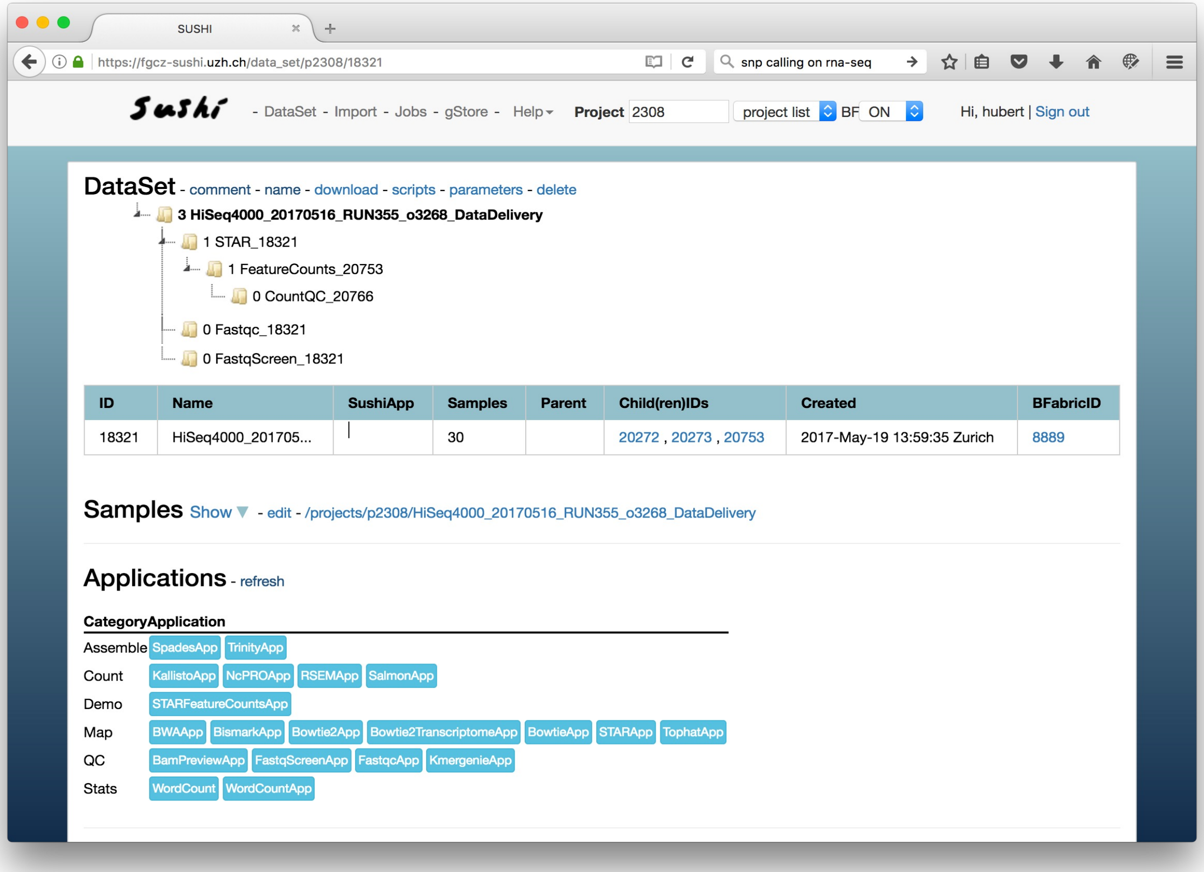
Core Facility: Throughput and Diversity





- Large Number of experiments
- Many different types of analyses
- *Aspiration to be reproducible*

SUSHI


SUSHI
https://fgcz-sushi.uzh.ch/data_set/p2308/18321
snp calling on rna-seq
Sushi
- DataSet - Import - Jobs - gStore - Help
Project 2308
project list
BF ON
Hi, hubert | Sign out

DataSet - comment - name - download - scripts - parameters - delete

- 3 HiSeq4000_20170516_RUN355_o3268_DataDelivery
  - 1 STAR_18321
    - 1 FeatureCounts_20753
      - 0 CountQC_20766
  - 0 Fastqc_18321
  - 0 FastqScreen_18321

ID	Name	SushiApp	Samples	Parent	Child(ren)IDs	Created	BFabricID
18321	HiSeq4000_201705...		30		20272 , 20273 , 20753	2017-May-19 13:59:35 Zurich	8889

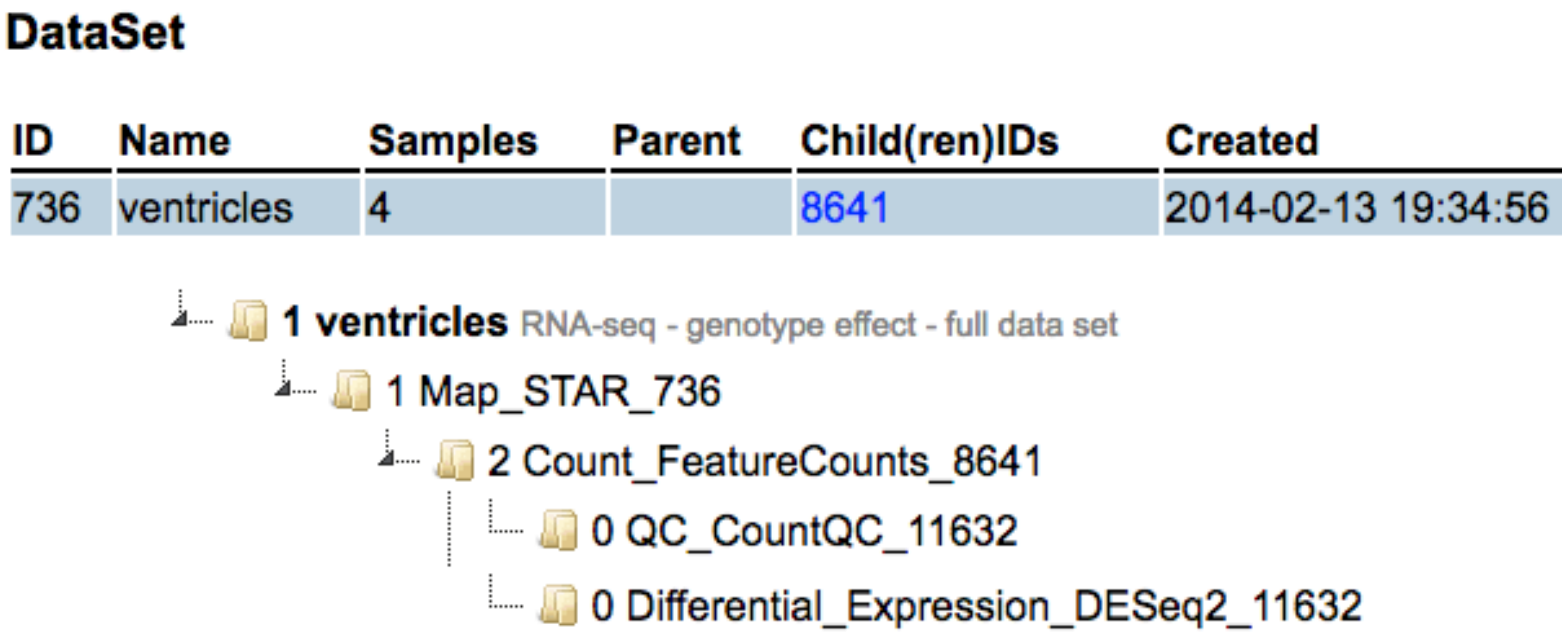
Samples Show - edit - /projects/p2308/HiSeq4000_20170516_RUN355_o3268_DataDelivery

Applications - refresh

Category	Application
Assemble	SpadesApp TrinityApp
Count	KallistoApp NcPROApp RSEMApp SalmonApp
Demo	STARFeatureCountsApp
Map	BWAApp BismarkApp Bowtie2App Bowtie2TranscriptomeApp BowtieApp STARApp TophatApp
QC	BamPreviewApp FastqScreenApp FastqcApp KmergenieApp
Stats	WordCount WordCountApp

- offers SUSHI apps to perform individual analysis steps
- ultimately generates an entire analysis trail
- relies on a web server and an associated database to run analysis steps

http://fgcz-sushi.uzh.ch



DataSet

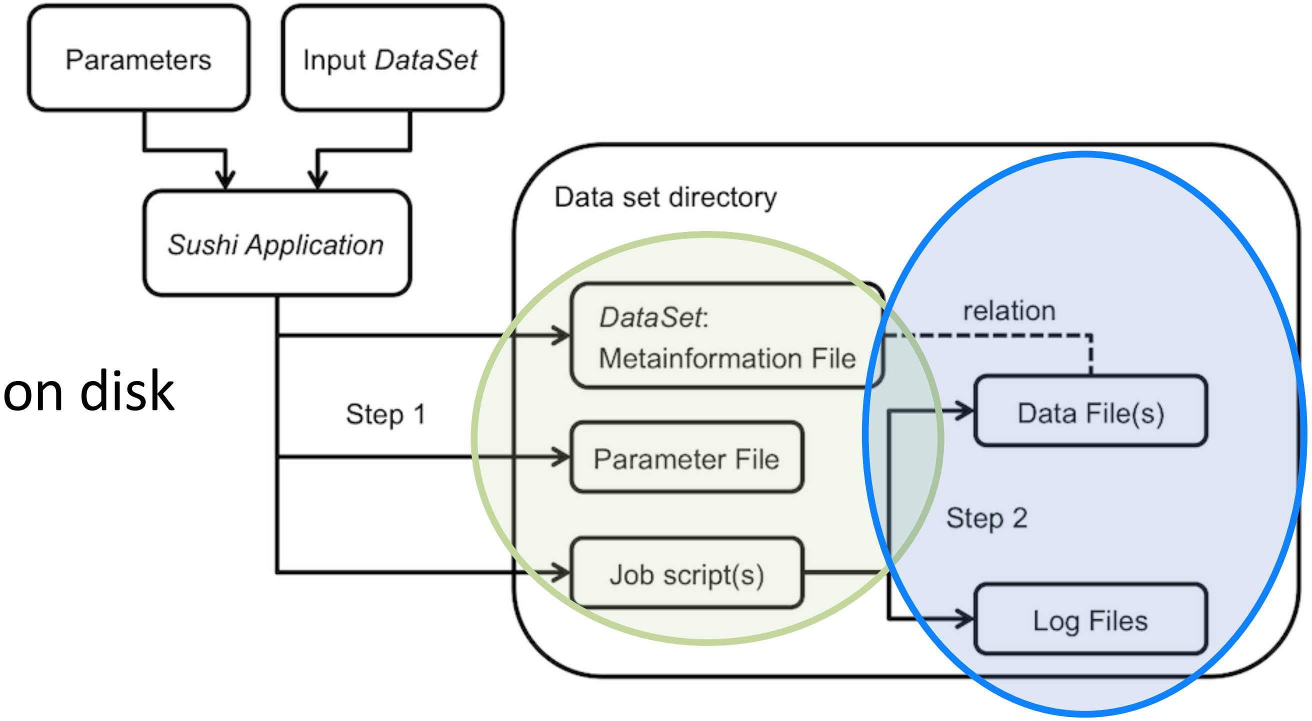
ID	Name	Samples	Parent	Child(ren)IDs	Created
736	ventricles	4		8641	2014-02-13 19:34:56

- **1 ventricles** RNA-seq - genotype effect - full data set
  - 1 Map_STAR_736
    - 2 Count_FeatureCounts_8641
      - 0 QC_CountQC_11632
      - 0 Differential_Expression_DESeq2_11632

- web-based
- fully reproducible
- self-contained data sets

Hatakeyama et al. 2016

SUSHI Informatics



Step 1:

- Analysis Generation
- Done by the SUSHI App
- Writes the intended analysis on disk

Step 2:

- Analysis Execution
- Performs the analysis
- Writes results on disk

SUSHI Applications are not stored persistently

Analysis Generation is not long-term repeatable

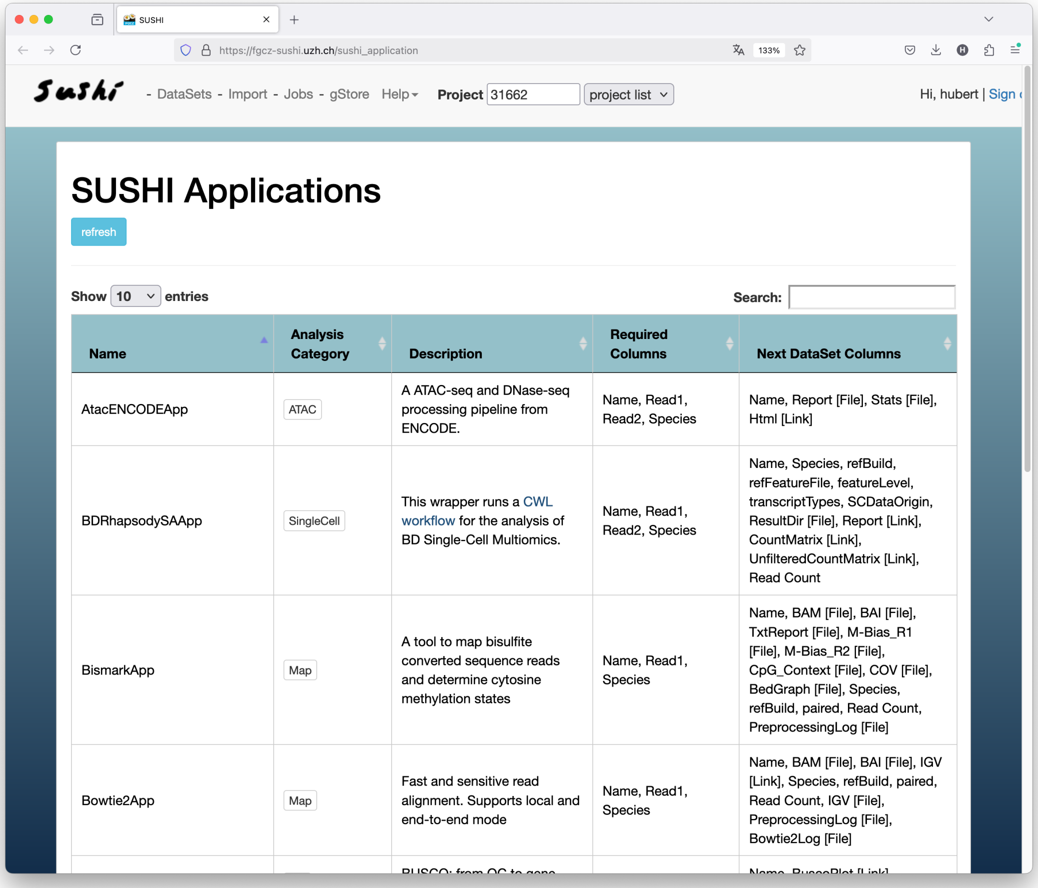
Actual Data Analysis is repeatable

SUSHI Apps and Workflows

- SUSHI models only individual steps
- Workflows are built as a series of steps; workflows are not explicitly modelled and managed
- Nevertheless: An individual step may actually perform an entire workflow in one step (e.g if the step executes an nf-core workflow)

- With a focus on steps there are fewer things to maintain

Available SUSHI Apps



SUSHI
https://fgcz-sushi.uzh.ch/sushi_application
133%
Sushi
- DataSets - Import - Jobs - gStore Help
Project 31662
project list
Hi, hubert | Sign

SUSHI Applications

Name	Analysis Category	Description	Required Columns	Next DataSet Columns
AtacENCODEApp	ATAC	A ATAC-seq and DNase-seq processing pipeline from ENCODE.	Name, Read1, Read2, Species	Name, Report [File], Stats [File], Html [Link]
BDRhapsodySAApp	SingleCell	This wrapper runs a CWL workflow for the analysis of BD Single-Cell Multiomics.	Name, Read1, Read2, Species	Name, Species, refBuild, refFeatureFile, featureLevel, transcriptTypes, SCDataOrigin, ResultDir [File], Report [Link], CountMatrix [Link], UnfilteredCountMatrix [Link], Read Count
BismarkApp	Map	A tool to map bisulfite converted sequence reads and determine cytosine methylation states	Name, Read1, Species	Name, BAM [File], BAI [File], TxtReport [File], M-Bias_R1 [File], M-Bias_R2 [File], CpG_Context [File], COV [File], BedGraph [File], Species, refBuild, paired, Read Count, PreprocessingLog [File]
Bowtie2App	Map	Fast and sensitive read alignment. Supports local and end-to-end mode	Name, Read1, Species	Name, BAM [File], BAI [File], IGV [Link], Species, refBuild, paired, Read Count, IGV [File], PreprocessingLog [File], Bowtie2Log [File]
		BUSCO: from QC to gene		Name, BuscoPlot [Link]

- Currently ~80 Apps are available
- Apps are plugins that can be removed or updated based on needs

Going stepwise

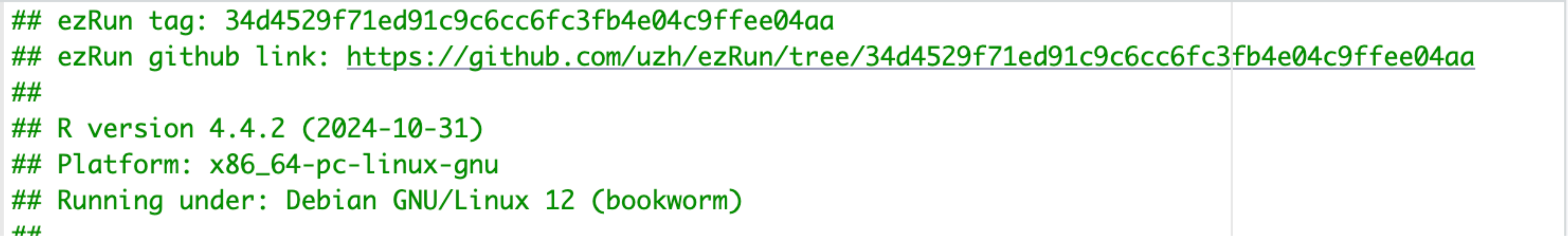
- Advantages:
  - modification of analysis trail can happen ad hoc
  - encourages revisiting of results of individual steps and adapt subsequent steps
  - number of steps much smaller than the number of sequences of steps, a.k.a. workflows
  - analysis steps re-used in different workflows

FGCZ Genome Informatics – data analysis portfolio

- What do those job scripts do?
- Software environment:
  - R with > 1000 packages installed
  - module environment with ~60 tools where we support different versions
  - ~100 tools installed with one version
  - ~ 30 conda environments for group-wise use
  - numerous personal conda environments, for individual use

Software Versions: Updates vs Consistency

- Conflicting interests:
  - run latest version vs consistent analysis within a project
- Update strategy:
  - Keep different version of tools within the modules; SUSHI can call specific versions of tools → support for different versions in different projects
  - Update R and all R-packages twice a year
  - All own code is git version-controlled and continuously updated; we keep track of the git tag
    → let's us continuously update our code



```
## ezRun tag: 34d4529f71ed91c9c6cc6fc3fb4e04c9ffee04aa
## ezRun github link: https://github.com/uzh/ezRun/tree/34d4529f71ed91c9c6cc6fc3fb4e04c9ffee04aa
##
## R version 4.4.2 (2024-10-31)
## Platform: x86_64-pc-linux-gnu
## Running under: Debian GNU/Linux 12 (bookworm)
##
```

Software life cycle

- SUSHI offers push-button repeatability but only for a short term
  - applications in SUSHI are retired if unused or if superseded
    - retired applications are effectively removed
    - retired applications can no longer be run on new data
    - previous analysis can only be repeated on command line
- SUSHI doesn’t keep any legacy!

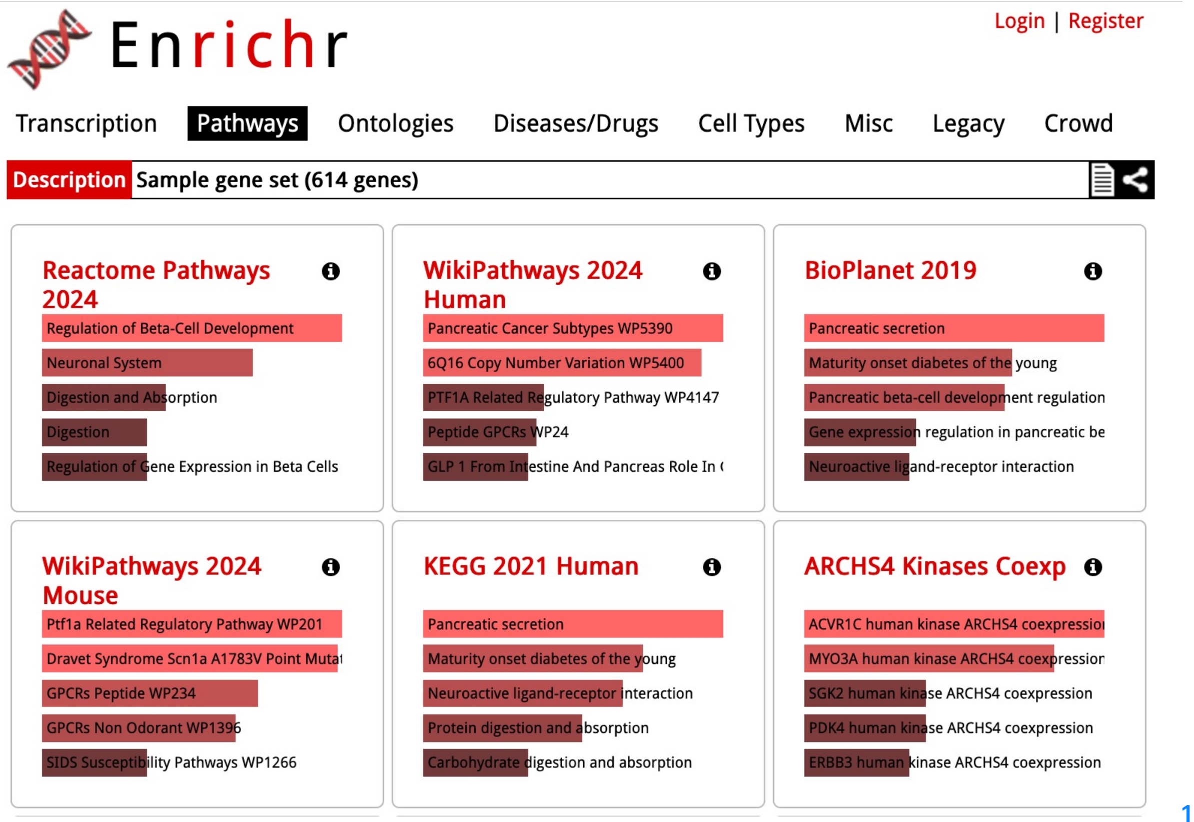
External Data: Genomes and Annotations

- Data analysis often uses additional data/knowledge from external variable sources, e.g. genome assemblies and annotation
- We keep
  - local versions: https://fgcz-gstore.uzh.ch/reference/Homo_sapiens/GENCODE/GRCh38.p13/Annotation/Release_34-2024-10-17/Genes/
  - script that generated the local version

- Generation of local genome copy is not repeatable if source files disappear from provider (NCBI, GENCODE, ….)

External Data: Enrichr

- API calls to external systems, e.g. Enrichr that host pathway knowledge bases → reproducibility not guaranteed



Enrichr

Login | Register

Transcription | Pathways | Ontologies | Diseases/Drugs | Cell Types | Misc | Legacy | Crowd

Description Sample gene set (614 genes)

Reactome Pathways 2024
- Regulation of Beta-Cell Development
- Neuronal System
- Digestion and Absorption
- Digestion
- Regulation of Gene Expression in Beta Cells

WikiPathways 2024 Human
- Pancreatic Cancer Subtypes WP5390
- 6Q16 Copy Number Variation WP5400
- PTF1A Related Regulatory Pathway WP4147
- Peptide GPCRs WP24
- GLP 1 From Intestine And Pancreas Role In (

BioPlanet 2019
- Pancreatic secretion
- Maturity onset diabetes of the young
- Pancreatic beta-cell development regulation
- Gene expression regulation in pancreatic be
- Neuroactive ligand-receptor interaction

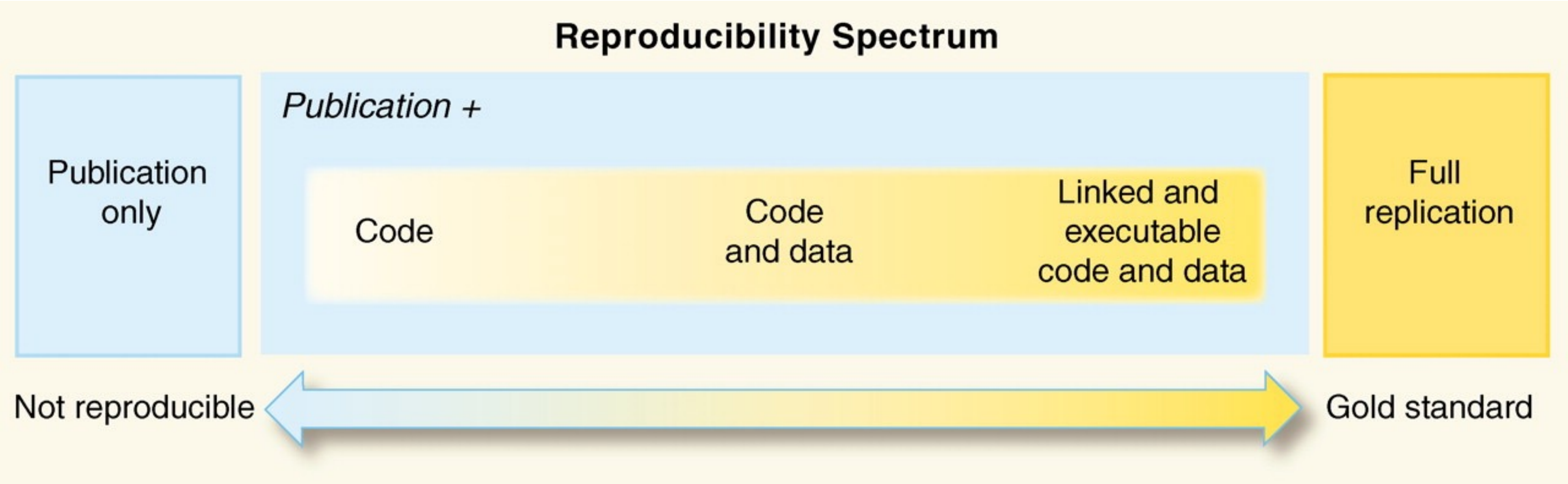
WikiPathways 2024 Mouse
- Ptf1a Related Regulatory Pathway WP201
- Dravet Syndrome Scn1a A1783V Point Mutat
- GPCRs Peptide WP234
- GPCRs Non Odorant WP1396
- SIDS Susceptibility Pathways WP1266

KEGG 2021 Human
- Pancreatic secretion
- Maturity onset diabetes of the young
- Neuroactive ligand-receptor interaction
- Protein digestion and absorption
- Carbohydrate digestion and absorption

ARCHS4 Kinases Coexp
- ACVR1C human kinase ARCHS4 coexpressio
- MYO3A human kinase ARCHS4 coexpression
- SGK2 human kinase ARCHS4 coexpression
- PDK4 human kinase ARCHS4 coexpression
- ERBB3 human kinase ARCHS4 coexpression

Where are we on the Reproducibility Spectrum?

- We commit to provide analysis code and, to limited extend, data



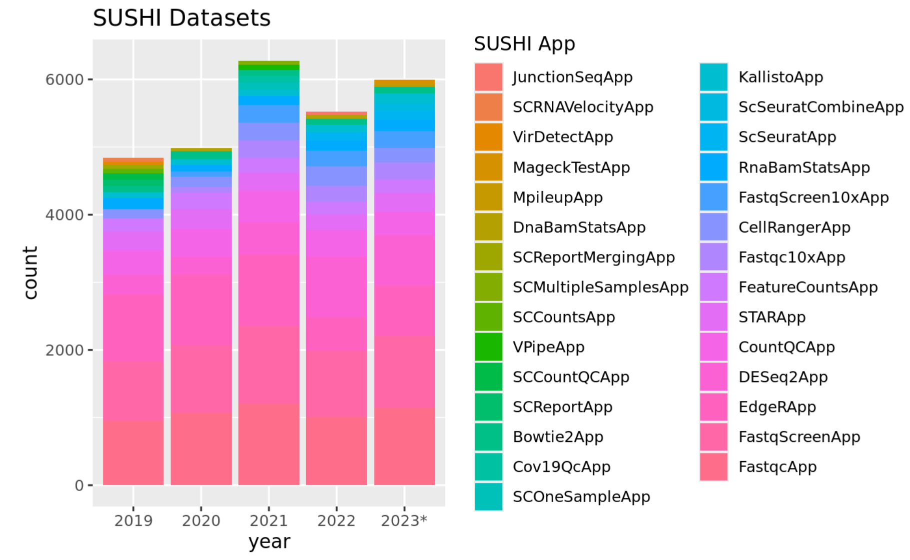
https://www.science.org/doi/10.1126/science.1213847

The issue with Data Rights

- Data rights are managed by the Principal Investigator
- PI is responsible for consent
- Mandate of core facility is usually to process and analyse the data not to long-term store the data

- Long-term storage implies long-term costs and requires funding which is usually not provided by PI
- PIs usually provide only the minimal context information on the samples that is necessary to achieve the analysis goals for the NGS data
  BUT: without full context information, the data has only limited value

Analysis Throughput & Requests for Reproduction



- Requests for repetition:
  - very few
- Requests for reanalysis
  - dozens per year (updated tools; different thresholds for sensitivity/specificity, …)
- Failed reanalysis/repetition
  - none

Other approaches: Galaxy

SUSHI

- repeat: rerun the static job script from the SHELL
- workflows: not modelled

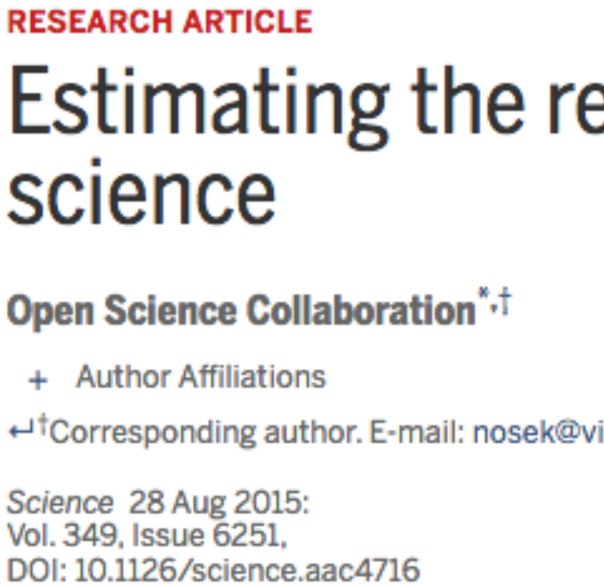
- requires compute environment available

GALAXY

- repeatability: rerun the workflow in the GALAXY interface
- workflows: explicitly modelled
- keeps track of history of workflows

- requires GALAXY instance

Repro Challenge: False Positives



RESEARCH ARTICLE

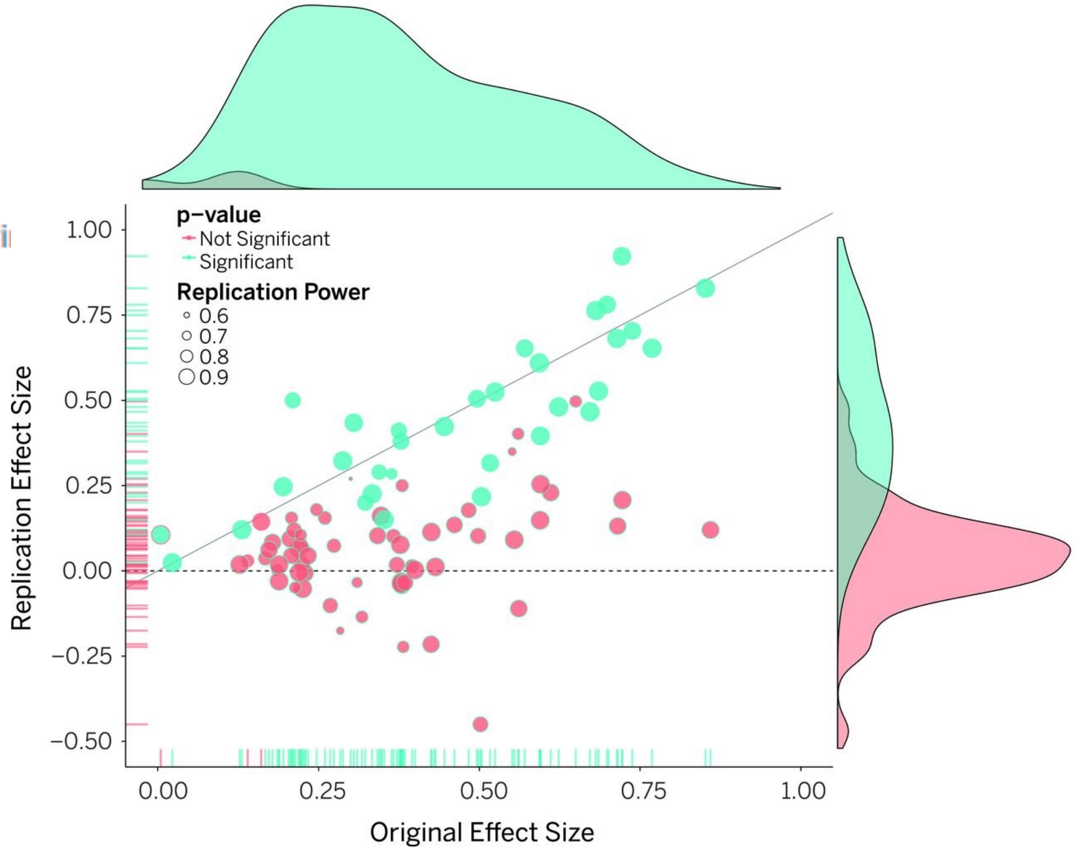
Estimating the reproducibility of psychological science

Open Science Collaboration[*,†]

+ Author Affiliations

↵[†]Corresponding author. E-mail: nosek@vi

Science 28 Aug 2015:
Vol. 349, Issue 6251,
DOI: 10.1126/science.aac4716



Summary

- Commitment to generate biologically interpretable and reproducible results
- Constraints:
  - dozens of workflows for hundreds of projects in a changing environment with data rights managed by users
  - analysis often needs to integrate in a larger workflow
- Minimalistic approach to Reproducibility
  - documented analysis workflows with scripts that enable long-term **repeatability** by a skilled bioinformatician (report the information necessary to understand and repeat the analysis)
  - push-button **repeatability** only available for ~1 year
- Results of analysis steps as self-contained and documented folders that enable reproducibility by other groups
- Long-term average of ~3 requests per year for repeated analysis