

# Sustainable tool benchmarking and workflow development in Computational Biology

**Kim Philipp Jablonski**

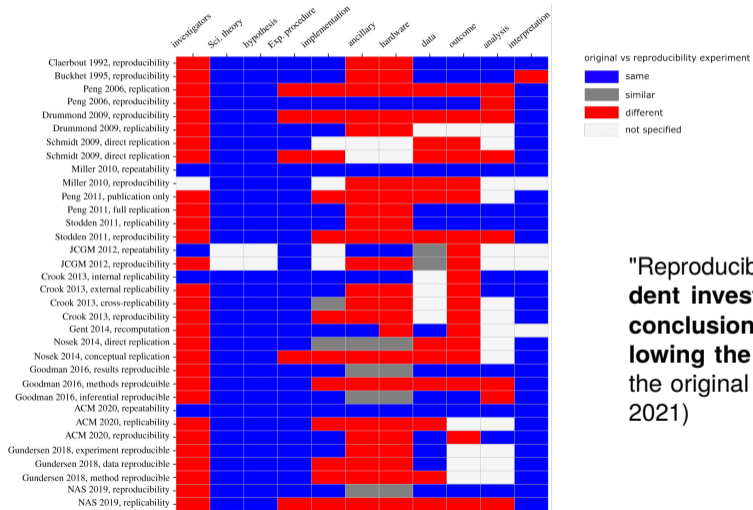
Computational Reproducibility Seminar – 2023.06.21

# Today's scaffolding

1. A biased and short review of the current state of sustainable data science
2. Automated workflows to save us from the **technical debt** of science ("scientific debt")
  - 2.1 Identifying cancer pathway dysregulations using differential causal effects
  - 2.2 Coherent pathway enrichment estimation by modeling inter-pathway dependencies using regularized regression
3. Automated workflows to enable **robust** and **large-scale** analyses
  - 3.1 The next generation of V-pipe: towards sustainable data processing workflows
4. What to do going forward?

# Current state of sustainable data science

# What is reproducibility?



"Reproducibility is the ability of **independent investigators** to draw the **same conclusions** from an experiment by **following the documentation** shared by the original investigators." (Gundersen, 2021)

# The reproducibility crisis

Reproducibility Project: Cancer Biology (Errington et al., 2021)

193 experiments from 53 papers

**2%**

experiments with open data

**70%**

of experiments required asking for key reagents

**69%**

of experiments needing a key reagent original authors were willing to share

**0%**

of protocols completely described

**32%**

of experiments the original authors were not helpful (or unresponsive)

**41%**

of experiments the original authors were very helpful

# Beyond reproducibility

# Identifying cancer pathway dysregulations using differential causal effects

Jablonski, K. P., Pirkl, M., Čevič, D., Bühlmann, P., & Beerenwinkel, N. (2022). *Bioinformatics*

---

# Coherent pathway enrichment estimation by modeling inter-pathway dependencies using regularized regression

Jablonski, K. P., & Beerenwinkel, N. (2022). *Submitted*

# Goal: develop novel computational methods

## *dce*

- Detect pathway dysregulations at edge-specific level
- Account for (latent) confounding factors using causal framework (intra-pathway dependencies)
- **Produce robust, well documented software package**
- **Make reproduction of all presented results as trivial as possible**

## *pareg*

- Make pathway enrichment robust for large, redundant pathway databases
- Implement generalizable benchmarking workflow
- **Produce robust, well documented software package**
- **Make reproduction of all presented results as trivial as possible**

## A detailed look at dce

# A detailed look at pareg

Linear model:

$$\begin{matrix}
 0 & 1 & 0 \\
 p_1 & & \\
 \vdots & & \\
 p_N & &
 \end{matrix}
 \begin{matrix}
 \mathbb{B} \\
 @ \\
 \vdots \\
 @
 \end{matrix}
 =
 \begin{matrix}
 0 & & & & \\
 X_{11} & X_{12} & & & \\
 X_{21} & X_{22} & & & \\
 \vdots & \vdots & \ddots & & \\
 X_{N1} & X_{N2} & & &
 \end{matrix}
 \begin{matrix}
 1 & 0 & 1 \\
 X_{1K} & & 1 \\
 X_{2K} & \mathbb{C} & \mathbb{B} \\
 \vdots & @ & \vdots \\
 X_{NK} & & \mathbb{C}
 \end{matrix}
 \begin{matrix}
 \mathbb{C} \\
 @ \\
 \vdots \\
 \mathbb{C}
 \end{matrix}
 ; \quad x_{ij} = \begin{cases} 1 & \text{if gene } i \text{ in pathway } j \\ 0 & \text{otherwise} \end{cases}$$

Objective function (with similarity matrix  $(g_{ij})$ ):

$$\mathbf{b} = \arg \min_{\mathbf{z}} \log(\text{likelihood}(\mathbf{y}; \mathbf{X}; \mathbf{z})) + \underbrace{\sum_{j=1}^K \|\mathbf{z}_j\|_{k_1}}_{\text{LASSO}} + \underbrace{\sum_{i=1}^K \sum_{j=1}^K \frac{z_i z_j}{2} g_{ij}}_{\text{network fusion}}$$

# Sustainable work ows as your trusty companion in the endless journey of science

Organizing your projects as work ows has steeper initial learning curve but pays o in long-term sustainability

Enables rapid iterations and safe prototyping and thus con dence in your results

Gives us a handle on scienti c debt

What this could look like in practice (1/2)

What this could look like in practice (2/2)

# The next generation of V-pipe: towards sustainable data processing workflows

Jablonski, K. P., Topolsky, I., Fuhrmann, L., Langer, B. & Beerenwinkel, N. In preparation

# Introduction

SARS-CoV-2 emerged in late 2019  
and caused COVID-19 pandemic  
(Hu et al., 2021)

575,887,049 confirmed cases  
including 6,398,412 deaths  
worldwide (World Health  
Organization, 2022)

# Introduction

SARS-CoV-2 emerged in late 2019  
and caused COVID-19 pandemic  
(Hu et al., 2021)

575,887,049 confirmed cases  
including 6,398,412 deaths  
worldwide (World Health  
Organization, 2022)

Swiss SARS-CoV-2 Sequencing  
Consortium leads largest  
sequencing effort in Switzerland

Enables genomic surveillance via  
NextStrain and CoV-Spectrum

# Introduction

Large-scale analyses on HPC  
clusters

Quickly adaptable to new  
questions

Robust performance

# Aims

Enable large-scale genomic surveillance programs

Make work ow reproducible, adaptable, and transparent, i.e., sustainable

Benchmark viral diversity estimation, a core V-pipe feature

# Sustainable workflow design

Automated testing

Docker containers

Virus-specific  
configuration files

Scripts for sample import  
and submission

Dynamic documentation

Project website/ mailing  
list

# Summary

- *V-pipe* is an integral part of Swiss SARS-CoV-2 monitoring efforts
- Has been applied in many projects
  - Alm et al., 2020; S. Nadeau, Beckmann, et al., 2020; Kuipers et al., 2020; Chen et al., 2021; S. Nadeau, T. G. Vaughan, et al., 2021; Jahn et al., 2022
- Quickly usable by independent researchers for novel viruses
- Core features are benchmarked in future-proof way



Pipeline overview Usage **SARS-CoV-2** Literature About Contact

V-pipe: A bioinformatics pipeline for viral sequencing data

New version v2.99.2 of V-pipe [has been released](#)

## Introduction

Virus populations exist as heterogeneous ensembles of genomes within their hosts. This genetic diversity is associated with viral pathogenesis, virulence, and disease progression, and it can be probed using high-throughput sequencing technologies.



Install from GitHub!

Get the Docker image!

Snakemake the workflow!

bio tools expasy resource sib resource License Apache 2.0

# What now?

## What to do going forward?

- Don't expect to do everything in one night – **incremental** actions lead to **long-term** success and prevent early burnout
- Sustainable research is not a binary decision – many small steps will nudge you in the right direction
- Focus on slowly shifting your **culture of reproducibility**
- Choose your favorite workflow management system (shout-out to Snakemake)
- These efforts go hand in hand with following **software engineering best practices**

Combining exciting research, sustainable workflow development, and proper software engineering is worth the effort!

# Acknowledgements

## Doctoral Examination Committee

Niko Beerenwinkel

Peter Bühlmann

Caroline Uhler

Petra Dittrich



## Computational Biology Group

Fritz Bayer

Nico Borgsmüller

Pawel Czyż

Arthur Dondi

Monica Drăgan

David Dreifuss

Falé Ferreira

Lara Furhmann

Johannes Gawron

Pelin Burcak Icer

Jack Kuipers

Xiang Ge Luo

Marco Roncador

Kevin Rupp

## Former members & more

Aashil Batavia

Anil Tuncel

Christos  
Dimitrakopoulos

Simon Dirmeier

Francesco Marass

Lisa Lamberti

Martin Pirkl

Unispital waiting  
rooms

Mathias Cardner

Jochen Singer

Susana Céspedes

Domagoj Čevič

Anne Bertolini

Franziska Singer

And You!

