ETH zürich

D BSSE

Computational Biology Group

Sustainable tool benchmarking and workflow development in Computational Biology

Kim Philipp Jablonski

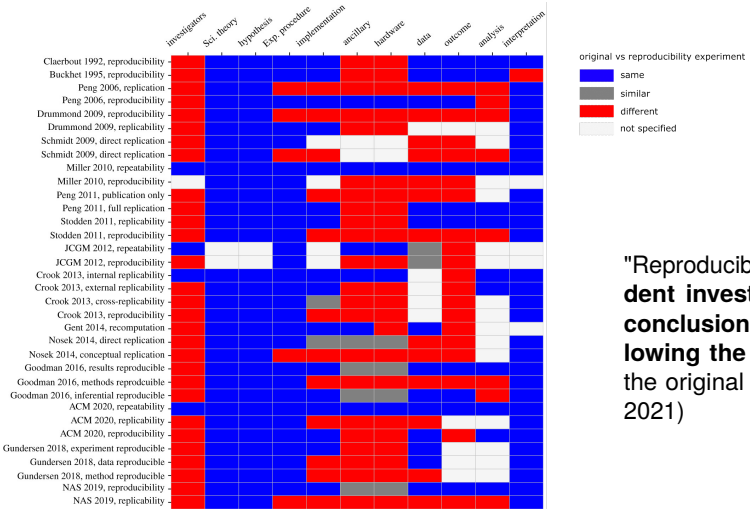
Computational Reproducibility Seminar – 2023.06.21

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?



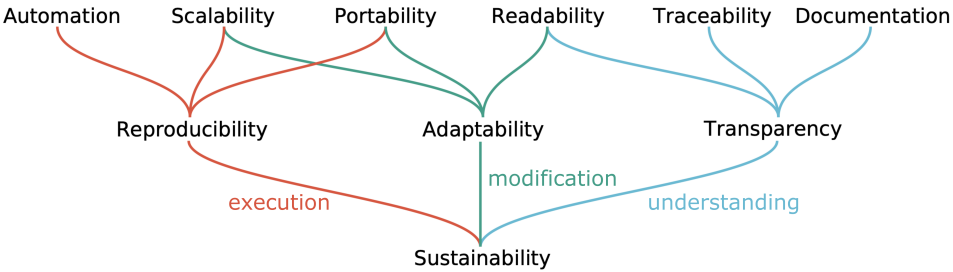
"Reproducibility is the ability of **independent investigators** to draw the **same conclusions** from an experiment by **following the documentation** shared by the original investigators." (Gundersen, 2021)

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



Adapted from figure 1 in Mölder et al., 2021

Identifying cancer pathway dysregulations using differential causal effects

Jablonski, K. P., Pirkl, M., Ćevid, 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

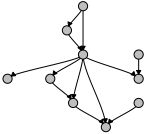
- Detect pathway dysregulations at edge-specific level
- Account for (latent) confounding factors using causal framework (intra-pathway dependencies)
- **Produce robust, well documented software package**
- **Make reproduction of all presented results as trivial as possible**

pareg

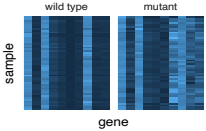
- Make pathway enrichment robust for large, redundant pathway databases
- Implement generalizable benchmarking workflow
- **Produce robust, well documented software package**
- **Make reproduction of all presented results as trivial as possible**

A detailed look at *dce*

A: Underlying Causal Structure

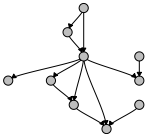


B: Expression Matrices

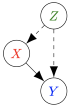


C: Pathway Databases

D: Prior Causal Knowledge

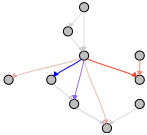


E: Causal Inference



$$\vdots$$
$$Y \sim \beta + \beta_X X + \beta_Z Z$$
$$\vdots$$

F: Causal Pathway Perturbations



A detailed look at *pareg*

Linear model:

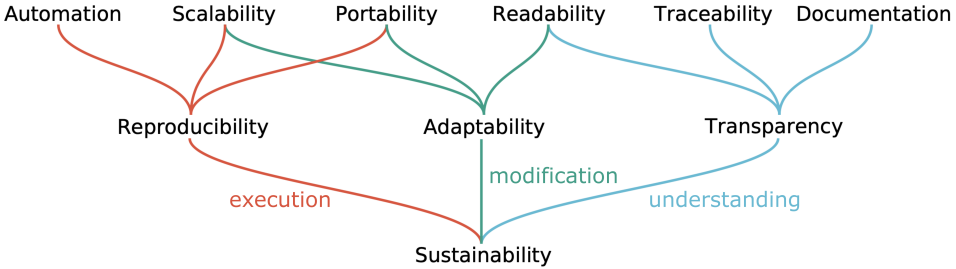
$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}, \quad x_{ij} = \begin{cases} 1 & \text{if gene } i \text{ in pathway } j \\ 0 & \text{otherwise} \end{cases}$$

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

$$\widehat{\beta} = \underset{\beta,\phi}{\arg\min} \left(-\log(\underbrace{\mathcal{L}(\beta,\phi|\mathbf{Y},\mathbf{X})}_{\text{likelihood}}) + \lambda \underbrace{\|\beta\|_1}_{\text{LASSO}} + \psi \underbrace{\sum_{i=1}^{K}\sum_{j=1}^{K} \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|_2^2 g_{ij}}_{\text{network fusion}} \right)$$

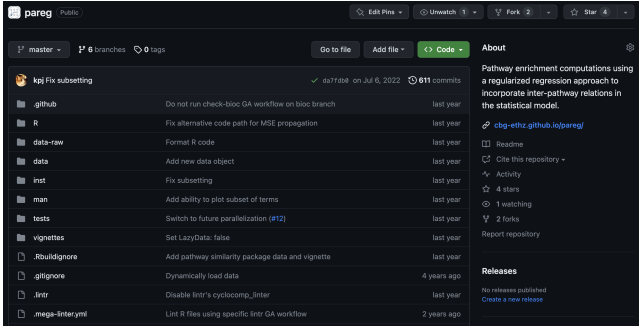
Sustainable workflows as your trusty companion in the endless journey of science

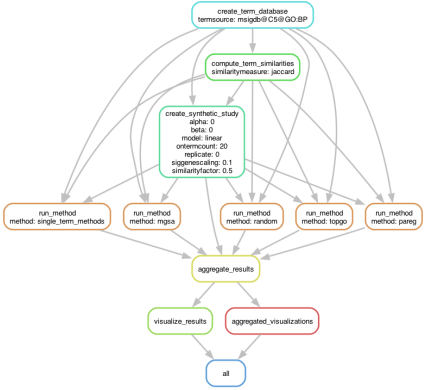
- Organizing your projects as workflows has steeper initial learning curve but pays off in **long-term sustainability**
- Enables **rapid** iterations and **safe** prototyping and thus **confidence** in your results
- Gives us a handle on **scientific debt**



Adapted from figure 1 in Mölder et al., 2021

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





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

dce

platforms all | rank 1024 / 2140 | support 0 / 0 | in Bioc 1 year
build ok | updated < 1 month | dependencies 236

DOI: 10.18129/B9.bioc.dce

Pathway Enrichment Based on Differential Causal Effects

Bioconductor version: Release (3.15)

Compute differential causal effects (dce) on (biological) networks. Given observational samples from a control experiment and non-control (e.g., cancer) for two genes A and B, we can compute differential causal effects with a (generalized) linear regression. If the causal effect of gene A on gene B in the control samples is different from the causal effect in the non-control samples the dce will differ from zero. We regularize the dce computation by the inclusion of prior network information from pathway databases such as KEGG.

Author: Kim Philipp Jablonski [aut, cre], Martin Pirkl [aut]

Maintainer: Kim Philipp Jablonski <kim.philipp.jablonski at gmail.com>

Citation (from within R, enter `citation("dce")`):

Jablonski, Philipp K, Pirkl, Martin, 'Cevid, Domagoj, B"uhlmann, Peter, Beerenwinkel, Niko (2021). "Identifying cancer pathway dysregulations using differential causal effects." *Bioinformatics*. doi: 10.1093/bioinformatics/btab847, https://doi.org/10.1093/bioinformatics/btab847.

dce 0.99.6 | Get started | Reference | Articles | Changelog

dce

Compute differential causal effects on (biological) networks. Check out our vignettes for more information.

> Publication: https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btab847/6470558

Installation

Install the latest stable version from Bioconductor:

```
BiocManager::install("dce")
```

Install the latest development version from GitHub:

```
remotes::install_github("cbg-ethz/dce")
```

Links
View on Bioconductor
Browse source code
Report a bug

License
GPL-3

Citation
Citing dce

Developers
Kim Philipp Jablonski
Author, maintainer
Martin Pirkl
Author

Dev status

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

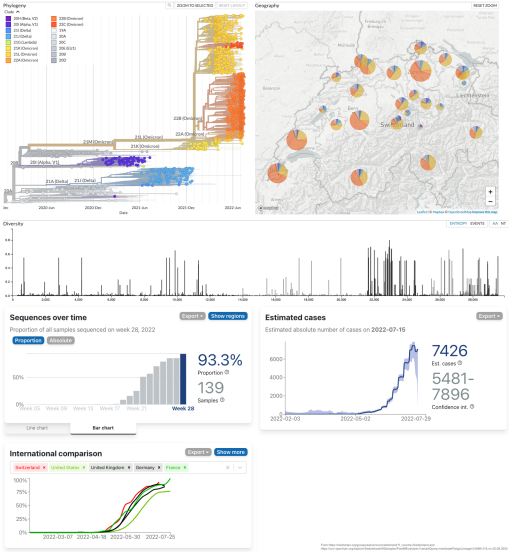
Jablonski, K. P., Topolsky, I., Fuhrmann, L., Langer, B. & Beerenwinkel, N. In preparation

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

VIOLLIER

labor team w ag

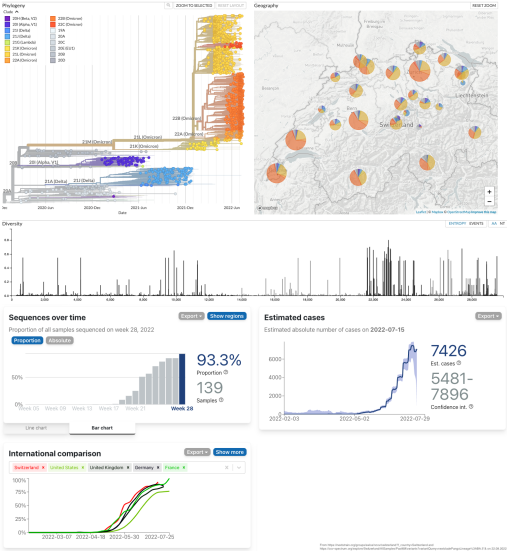
Universitätsspital Basel

health2030 genome center

- Large-scale analyses on HPC clusters
- Quickly adaptable to new questions
- Robust performance

ENA European Nucleotide Archive

GISAID



Aims

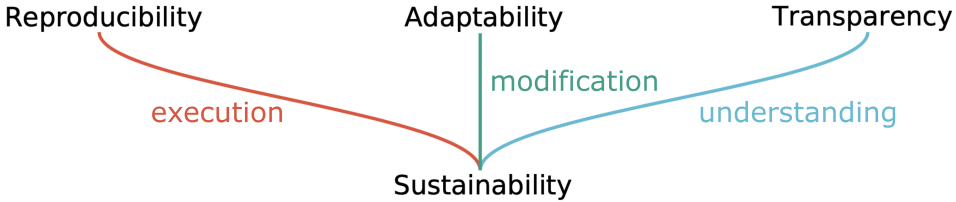
- Enable large-scale genomic surveillance programs
- Make workflow reproducible, adaptable, and transparent, i.e., sustainable
- Benchmark viral diversity estimation, a core V-pipe feature

Sustainable workflow design

- Automated testing
- Docker containers

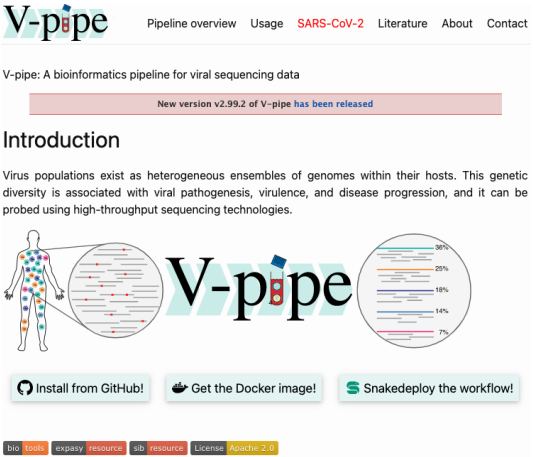
- Virus-specific configuration files
- Scripts for sample import and submission

- 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



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

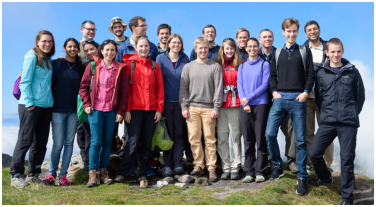
Doctoral Examination Committee

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

Anıl Tuncel

Christos Dimitrakopoulos

Simon Dirmeier

Francesco Marass

Lisa Lamberti

Martin Pirkl

Unispital waiting rooms

Mathias Cardner

Jochen Singer

Susana Céspedes

Domagoj Ćevid

Anne Bertolini

Franziska Singer

And You!

ETH zürich

D BSSE

Computational Biology Group

[thanks]

kpj
@kpj_py
kim.philipp.jablonski@gmail.com