

References data

Maxime U Garcia
SciLifeLab | National Genomics Infrastructure
nf-core

ANF - workflow et reproductibilité

2025-11-25..27



The Challenge of Reproducibility

Hidden reproducibility issues are like an iceberg



First, we tried to re-run the analysis with the code and data provided by the authors.

Second, we reimplemented the whole method in a Python package...



Experimenting with reproducibility:
a case study of robustness in bioinformatics
<https://doi.org/10.1093/gigascience/giy077>

The five pillars of computational reproducibility:
bioinformatics and beyond
<https://doi.org/10.1093/bib/bbad375>



Vocabulary of Reproducibility

Permalink: <http://nojhan.net/tfd/vocabulary-of-reproducibility.html>

2025-03-31 © nojhan — CC-BY-SA v4

Reference: <http://arxiv.org/abs/2102.03380v2>

Random Factors

e.g. seed, batch



Repeatable

Exactly repeat the original experiment, generating precisely the same results.



Reproducible

Test whether the original results were a statistical anomaly.



Replicable

Test whether one can independently reach the same conclusion.



Generalisable

Test whether the conclusion extends beyond the original experiments.



Original



New



New



New

Artifacts

e.g. input data



Original



Original



New



New

Fixed Factors

e.g. source code



Original



Original



Original



New

So let's make crêpes, but France 1954 style



STAGE

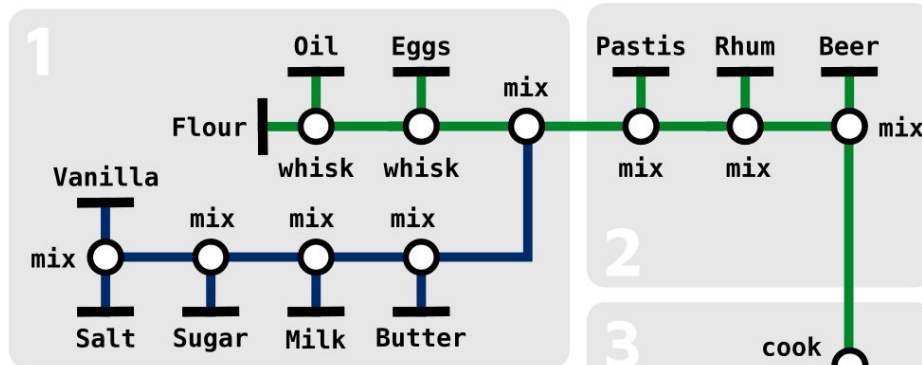
1. Prepare batter
2. Perfume and relax batter
3. Cook and enjoy



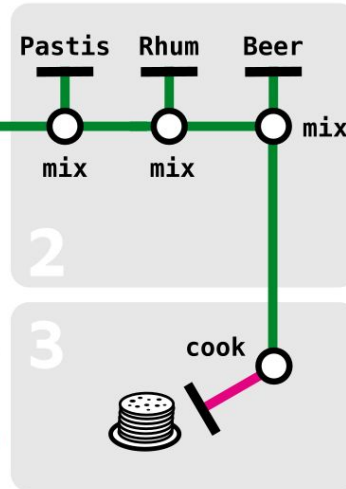
APPLIANCE

- Bowl
- Saucepan
- Crêpière

License:



nf-core/
crêpes



cook

INSPIRED BY

Raymond Oliver - 1954



Input data for pipelines

- Data to be analyzed
- Reference data
- Metadata



nf-core



nf-core - what do we do?

All the resources we use

- Public data
- AWS iGenomes
- annotation-cache
- test-data
- Bundled with pipelines
- User provided



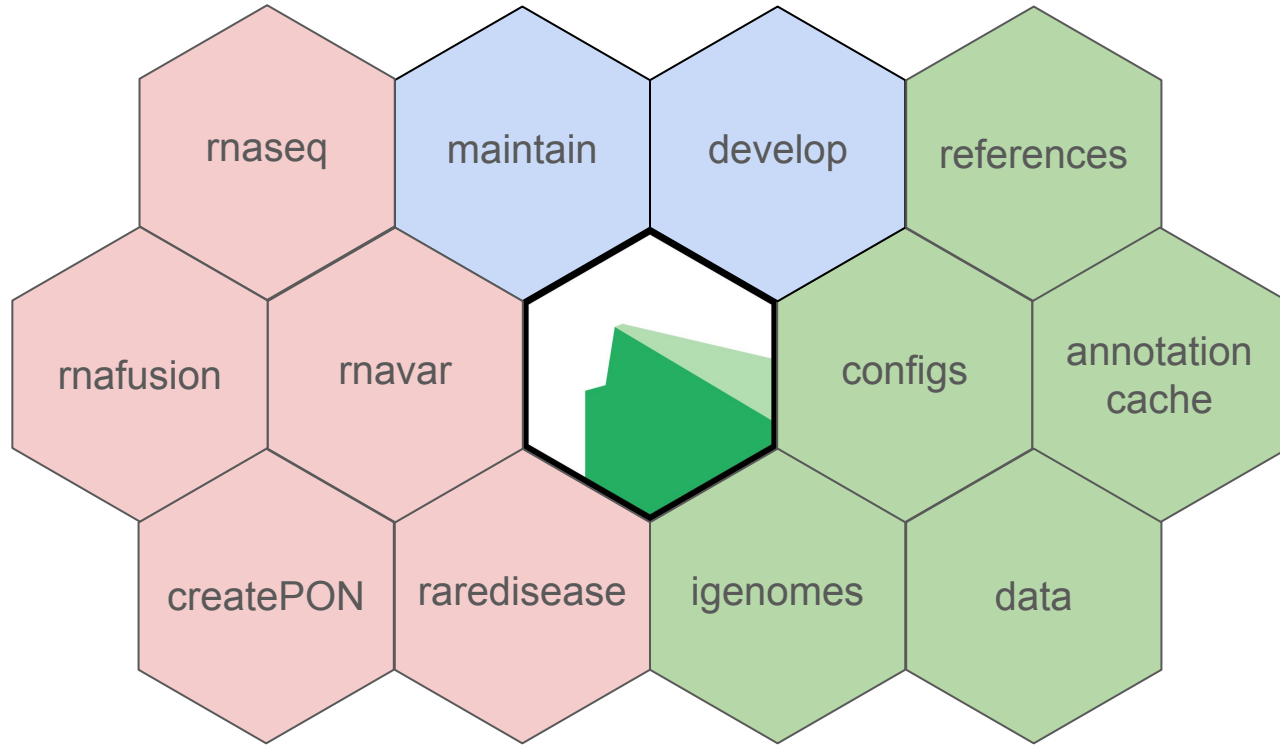
Sarek

The park

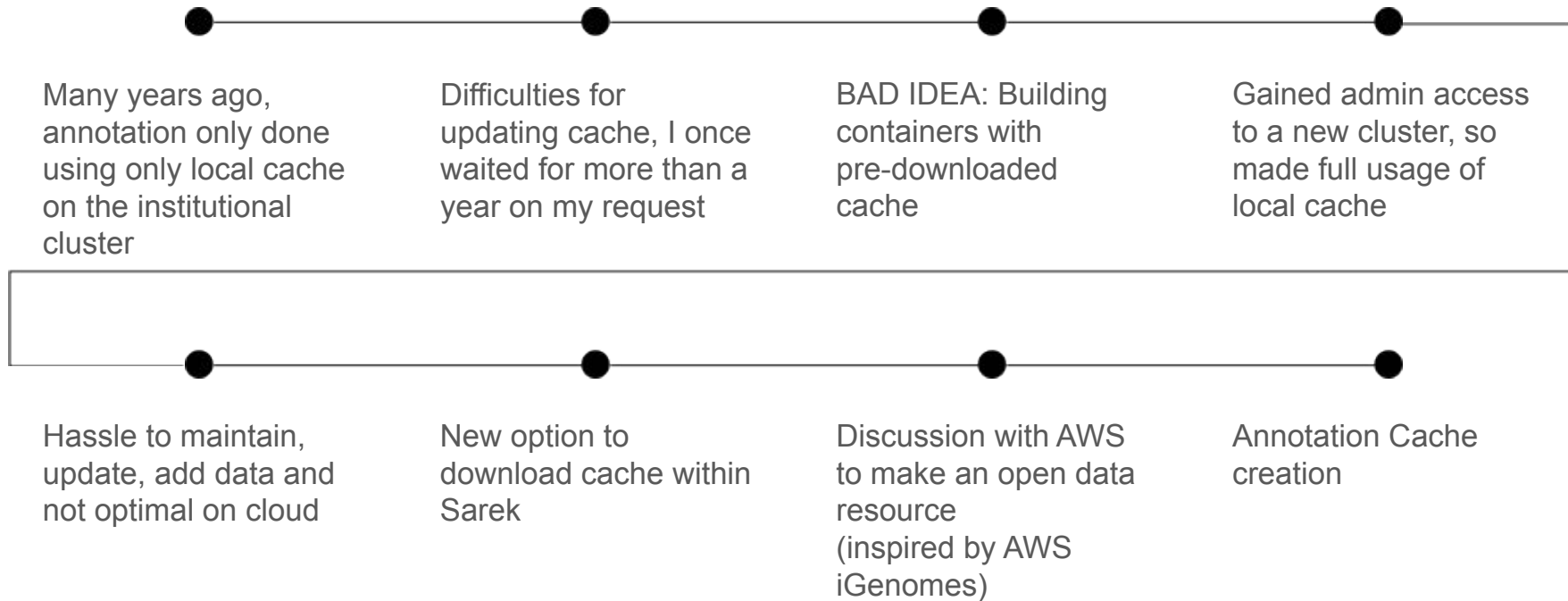


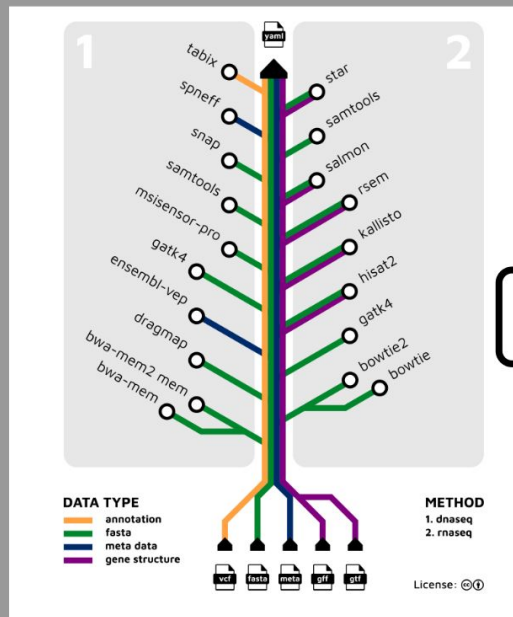
Sarek is a National Park in Northern Sweden.
Pretty inaccessible.
With beautiful landscape.
The name is an allegory for the genomic landscape to explore with the pipeline.

Sarek and its ecosystem

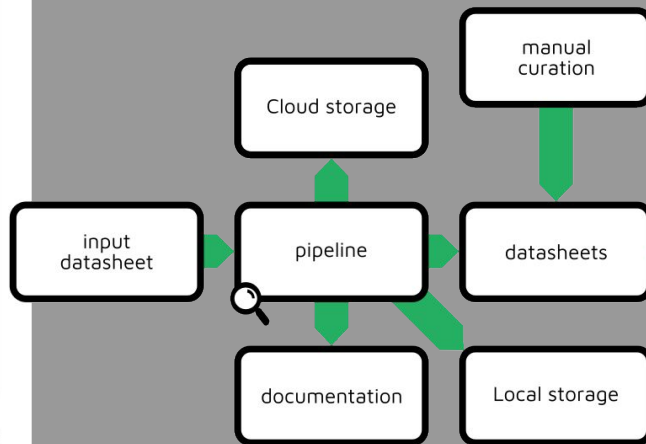


Annotation-cache





The pipeline



Reproducible creation, data provenance,
automation, flexible,
Any genome, local or cloud, up-to-date, easy to use

