

Anatomy of a Modular Toolchain for Reproducible Bioinformatics Workflows

Kelly Sovacool^{1,2}, Samarth Mathur^{1,2}, Vishal Koparde^{1,2}, Parthav Jailwala^{1,2}

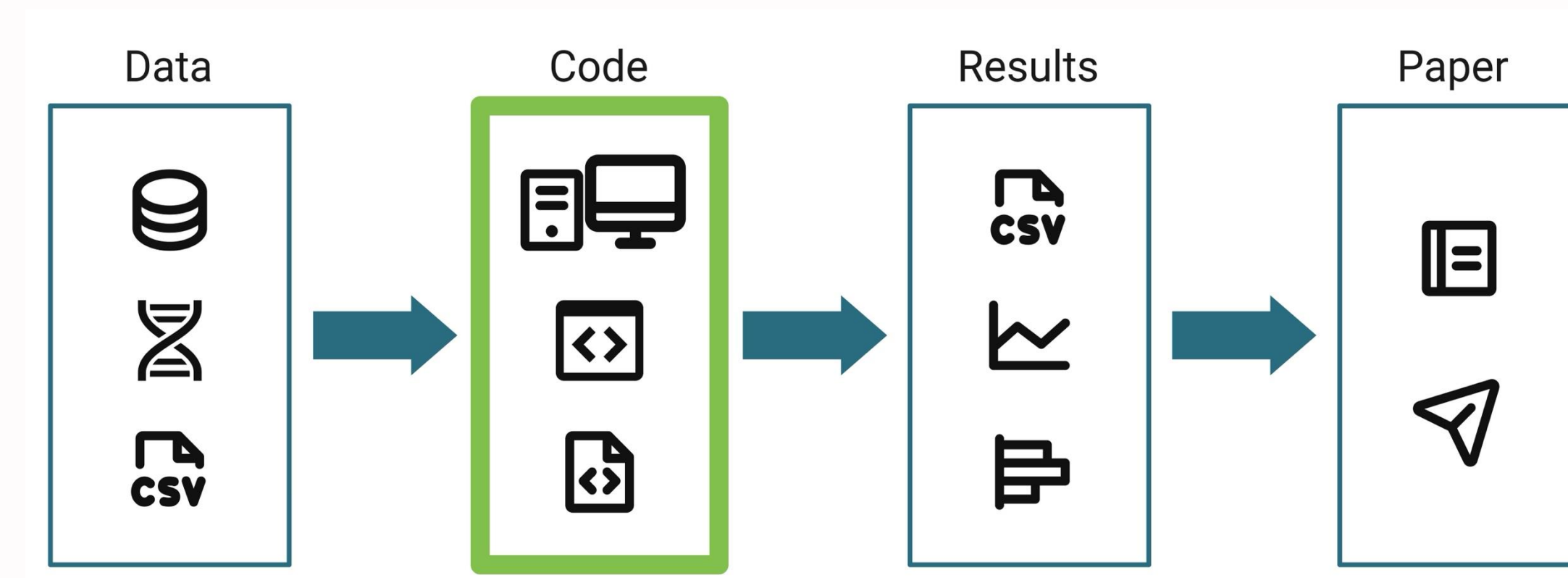


¹ Advanced Biomedical Computational Science, Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA

² CCR Collaborative Bioinformatics Resource, Office of Science and Technology Resources, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

INTRODUCTION

Reproducibility is the foundation of trustworthy data-intensive biological research. Reproducible research relies on the quality of the software that generated it.




REPRODUCIBILITY
Same data; same methods


REPLICABILITY
Different data; same methods


ROBUSTNESS
Same data; different methods


GENERALIZABILITY
Different data; different methods

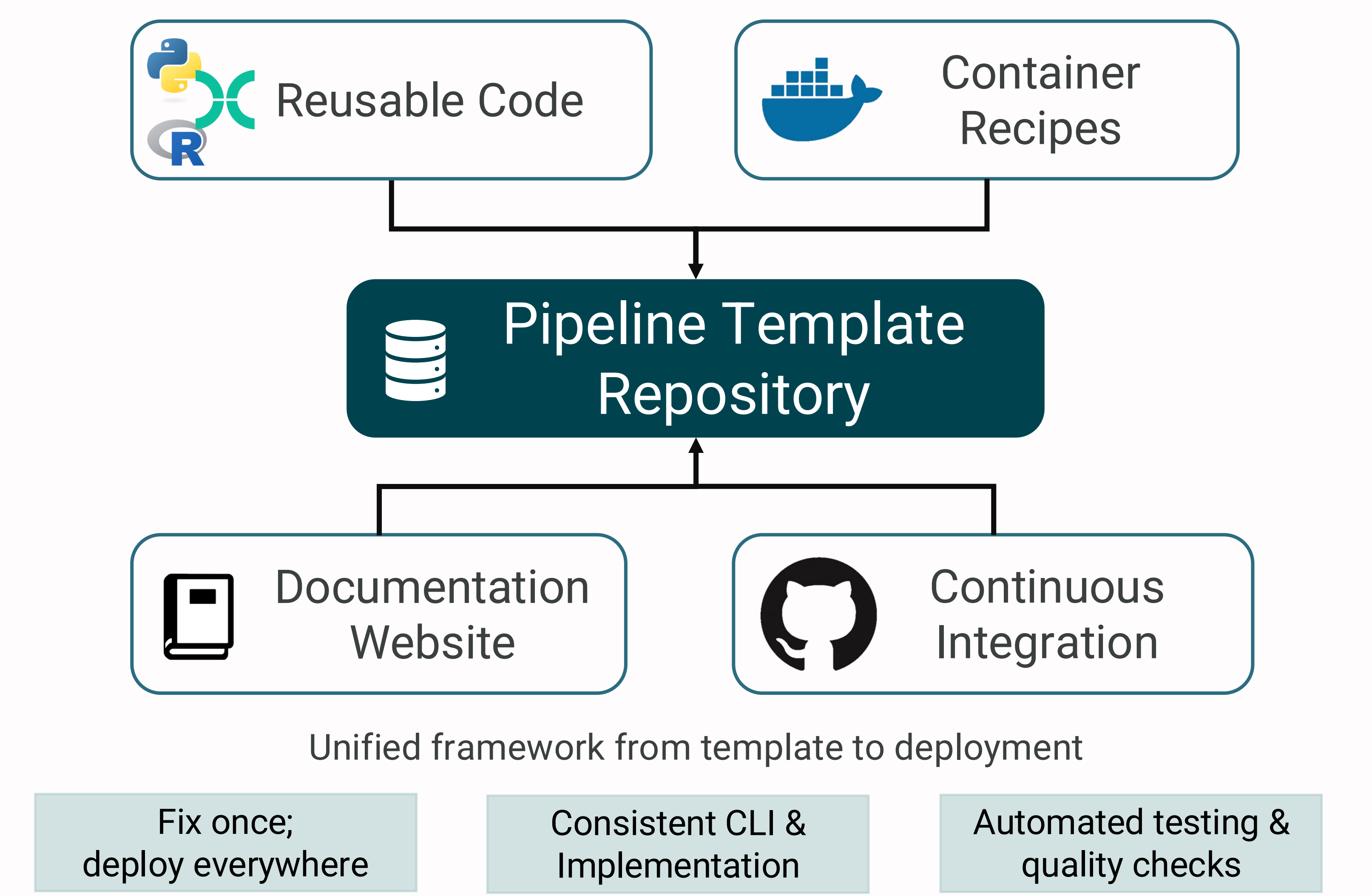
Schloss PD. 2018. [10.1128/mBio.00525-18](https://doi.org/10.1128/mBio.00525-18).

BEST PRACTICES

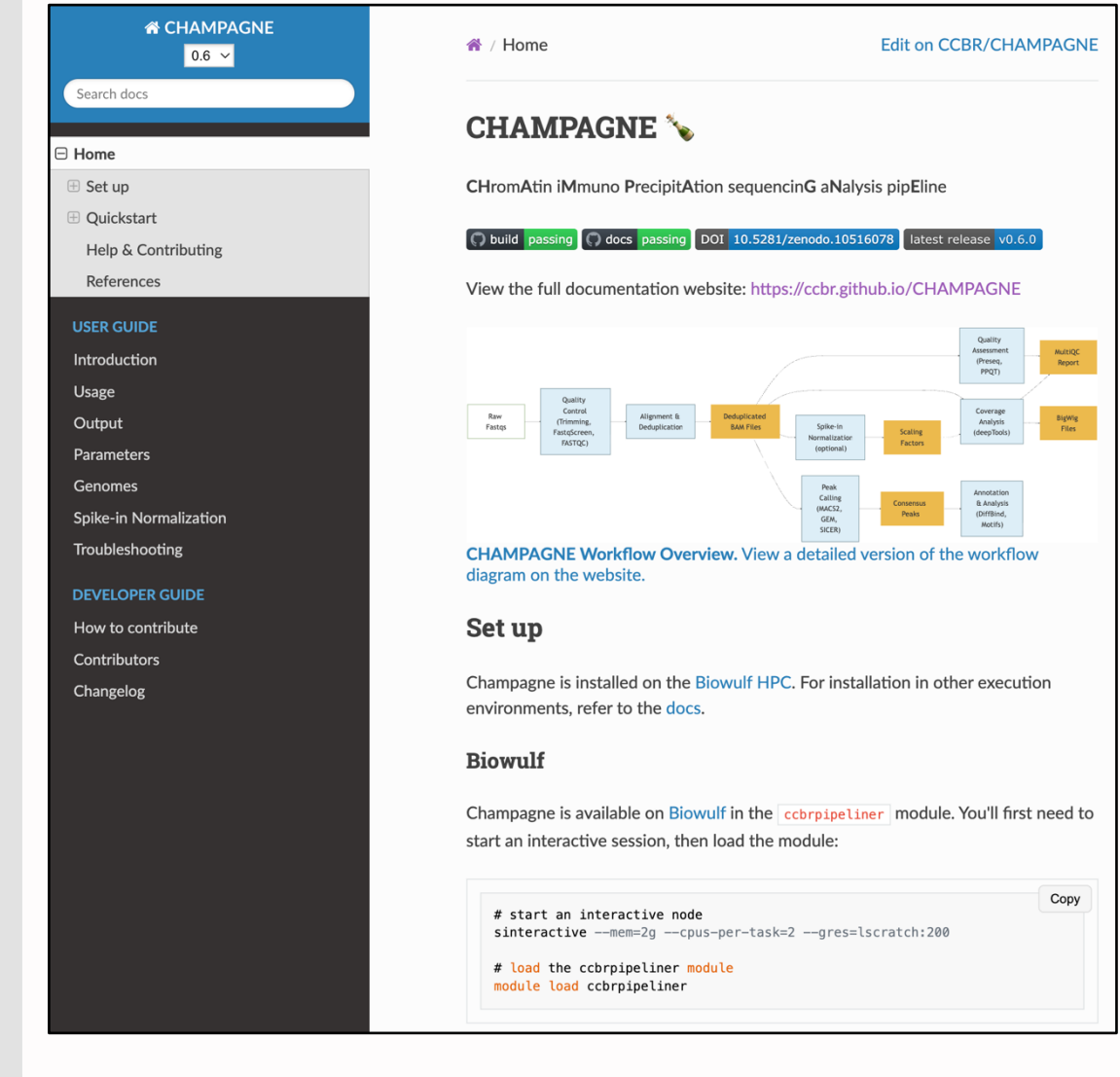
Modular Design	Code Reuse
Version Control	Documentation
Containerize Dependencies	Continuous Integration

Following established best practices in research software engineering helps to prevent accumulation of technical debt, streamline maintenance processes, and enable long-term sustainability of software projects.

TOOLCHAIN FOR REPRODUCIBLE BIOINFORMATICS



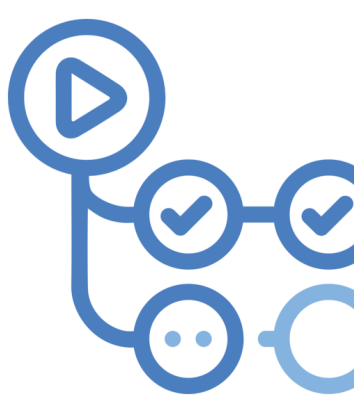
CLI, DOCUMENTATION, AUTOMATION



Consistent interface, usage guidelines, and documentation across all pipelines provides a cohesive user-experience and simplifies user training.

```
$ champagne run --input samplesheet.csv
```

```
$ sinclair run --input samplesheet.csv --params-file ...
```



Pull Request process ensures every change is reviewed thoroughly. GitHub Actions runs CI checks on every change.

Scalable, Reproducible, Sustainable Bioinformatics

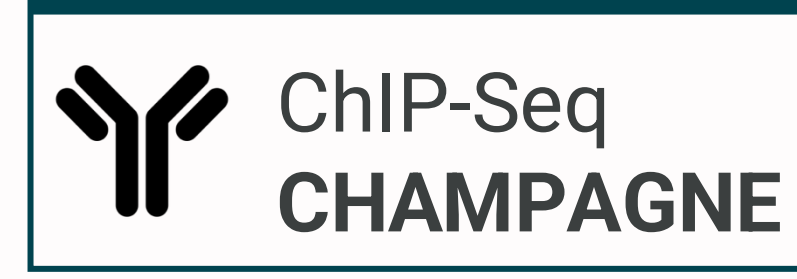
NIH Biowulf HPC
Secure, Scalable, and high performance computing platform

GitHub Open Source
Collaborate, Contribute, Reuse

Documentation
Reference Guides, Tutorials, Best Practices

10+ pipelines on NIH Biowulf HPC

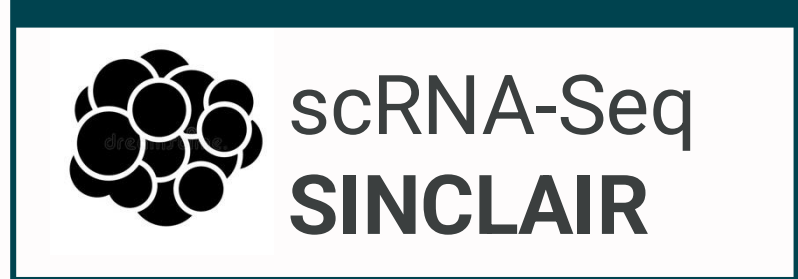
\$ module load ccbripipeliner



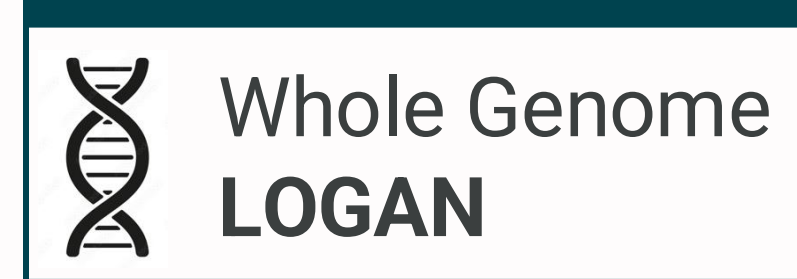
ChIP-Seq
CHAMPAGNE



CRISPR-Seq
CRISPIN



scRNA-Seq
SINCLAIR



Whole Genome
LOGAN

References

Schloss, P. D. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *mBio* **9**, e00525-18, /mbio/9/3/mBio.00525-18.atom (2018).

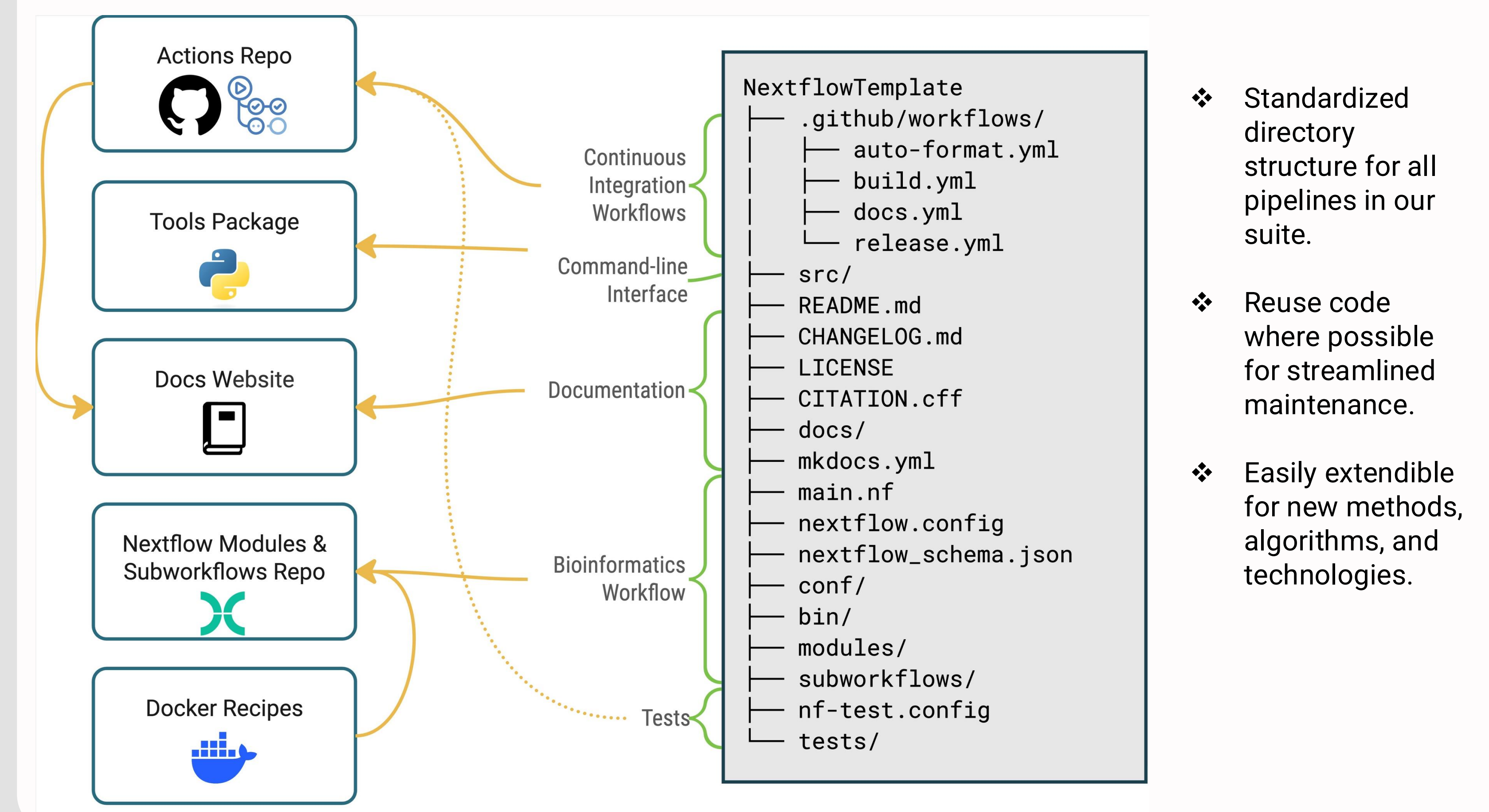
Ewels, P. A. et al. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* **38**, 276–278 (2020).

Wilson, G. et al. Best Practices for Scientific Computing. *PLoS Biol* **12**, e1001745 (2014).

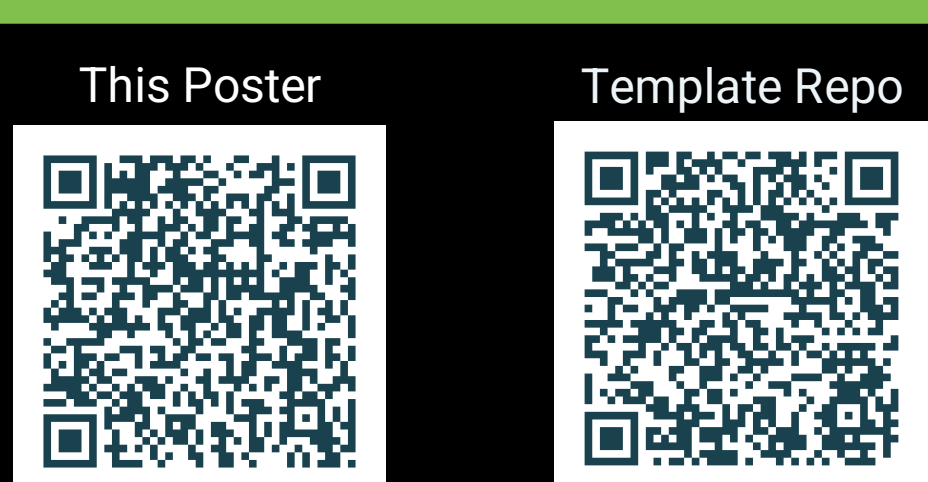
Wilson, G. et al. Good enough practices in scientific computing. *PLOS Computational Biology* **13**, e1005510 (2017).

Roach, M. J. et al. Ten simple rules and a template for creating workflows-as-applications. *PLoS Comput Biol* **18**, e1010705 (2022).

PIPELINE TEMPLATE



- Standardized directory structure for all pipelines in our suite.
- Reuse code where possible for streamlined maintenance.
- Easily extendible for new methods, algorithms, and technologies.



Advancing reproducible computational research through
high-quality bioinformatics software and data analysis pipelines