pawsey

eResearch Australasia
Brisbane, 16 – 20 October 2023

Streamlining software deployment and access: two container stories

Mark Gray, on behalf of

M. De La Pierre, A. Stott, S. Beecroft, V. Sochat, P. Di Tommaso, T. Gamblin

Australian Government

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

CSIRO

Curtin University

Murdoch University

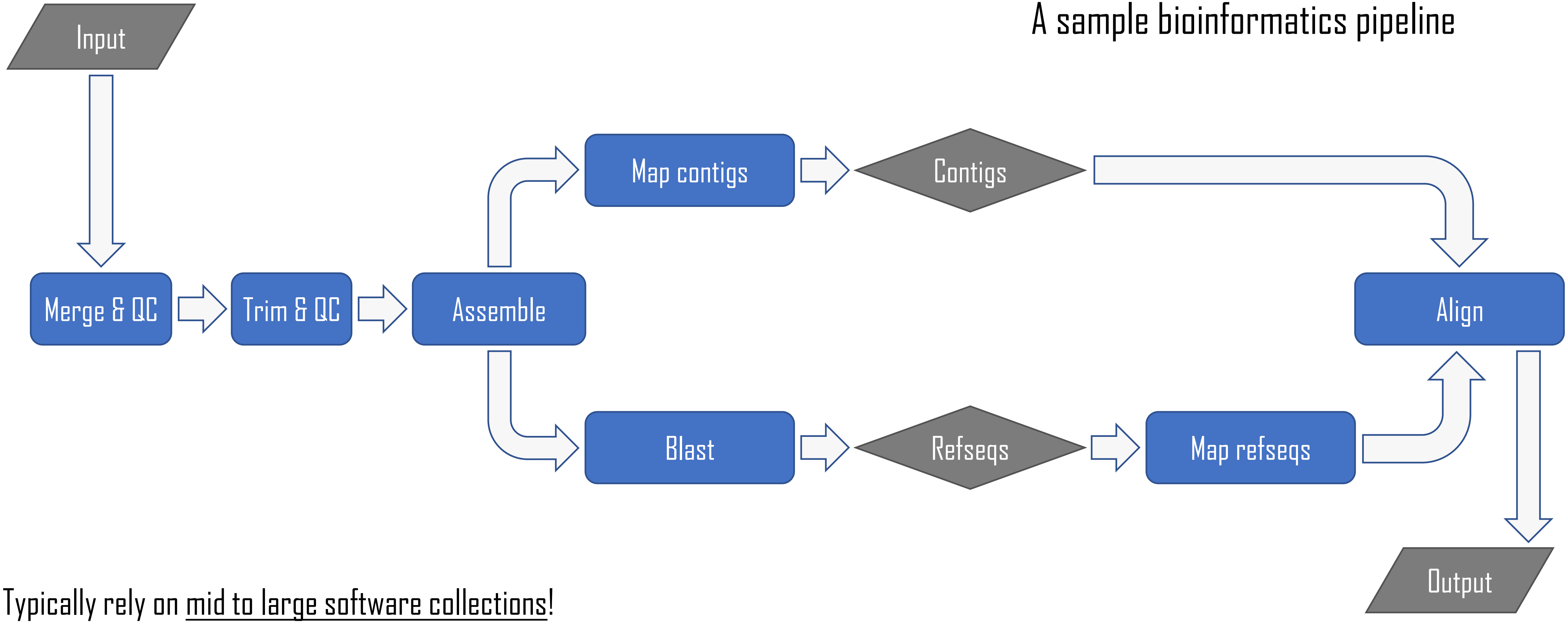
Government of Western Australia

ECU Edith Cowan University

The University of Western Australia

Scientific workflows

A sample bioinformatics pipeline



Typically rely on <u>mid to large software collections</u>!

A question back in 2018 – or was it earlier?

- Can we reduce the "software pain" incurred when attempting to setup and run life sciences (and other) data-driven pipelines?

- This would benefit researchers, support staff, infrastructure providers … and science, if reproducibility is also improved

Answer: YES! … Use containers, you folks!

Is handing containers off to users the ultimate answer?

Teaching large cohorts of researchers to design and implement containers in their workflow is still quite painful

(to be entirely honest, it is a bit less painful)

Containers aka ... shift the pain from installation to training

Container story #1

- What if we could hide containers to users?

- Use containers to ease software deployment
- Provide users with usual interface and experience

- … and what if we could scale the approach to a large, pre-existing container collection, such as the BioContainers? (9000+ tools)

Container story #1: collaboration around Singularity HPC

- Container Modules using Singularity HPC

- Collaboration between Pawsey and Vanessa Sochat (ex Singularity, now at LLNL)

- Integration of several new features
  - expanded support for modules
  - support for Docker and Podman in addition to Singularity
  - validated integration with workflow tools
  - support for MPI and GPUs
- Automated, scalable generation of container manifests, enabling inclusion of whole BioContainers collection

Container story #1: outcomes

- Publicly and globally available:
  - https://singularity-hpc.readthedocs.io
- Requires:
  - Python, module system, container runtime
- Minimal setup:

```
# setup
pip install singularity-hpc
module use $( shpc config get module_base )

# install a tool
shpc install quay.io/biocontainers/fastqc:0.11.9--0

# use the tool as usual
module load quay.io/biocontainers/fastqc/0.11.9–0
fastqc --version
```

- Peer-reviewed publication at:
  - https://doi.org/10.5334/jors.451

Container story #2

- Since inception, the Nextflow workflow engine has provided a transparent interface to:
  - container runtimes, the Conda package manager, HPC schedulers, Cloud providers
- In late 2022, **Wave Containers** were introduced in Nextflow
  - dynamically wrap single or multiple Conda packages in a container
  - add other required scripts or tools in a modular way
  - behind the scenes caching for faster deployment
  - completely transparent to users

- What can we add on top of this, from an HPC perspective?
- Answer: architecture optimised containers using the Spack package manager!

Container story #2: collaboration around Spack Wave containers

- Integration of Spack within Nextflow and Wave containers

- Collaboration between Pawsey, Paolo Di Tommaso at Seqera Labs (Nextflow) and Todd Gamblin at LLNL (Spack)

- Integration in standard Nextflow
  - transparently use existing Spack installation in your infrastructure to deploy software within Nextflow pipelines

- Integration in Wave containers
  - transparently deploy containerised, architecture optimised Spack packages through containers
  - still under development (binary caching)
  - benefits not only HPC, also Cloud (e.g. ARM instances in AWS)

Container story #2: outcomes

Spack in standard Nextflow

- Publicly available from Nextflow version 23.04
  - https://github.com/nextflow-io/nextflow
  - https://www.nextflow.io/docs/latest/spack.html
- Requires:
  - Java VM, Python, Spack

```
# nextflow configuration

process.spack = 'fastqc@0.11.9'
spack.enabled = true
```

Spack in Wave containers

- Service under development, available over next few months
- Requires:
  - Java VM, container runtime

```
# nextflow configuration

process.spack = 'fastqc@0.11.9’
singularity.enabled = true
wave {
  strategy = ['spack']
  enabled = true
}
```

Publication under preparation

The message from these stories

- Containers technology as a starting point to improve user experience of researchers running computational workflows
  - higher productivity through reduced software deployment time
  - lower barrier to access through simple, intuitive user interfaces

- Collaboration and open-source as enablers to increase impact and reach of outcomes

Thanks!

Check out our annual report