

Swiss Institute of
Bioinformatics

Kubernetes for Biomedical Analysis

Personalized Health Informatics Group (PHI)
SIB Swiss Institute of Bioinformatics

Kevin Sayers, Workflow Developer @KevinTSayers

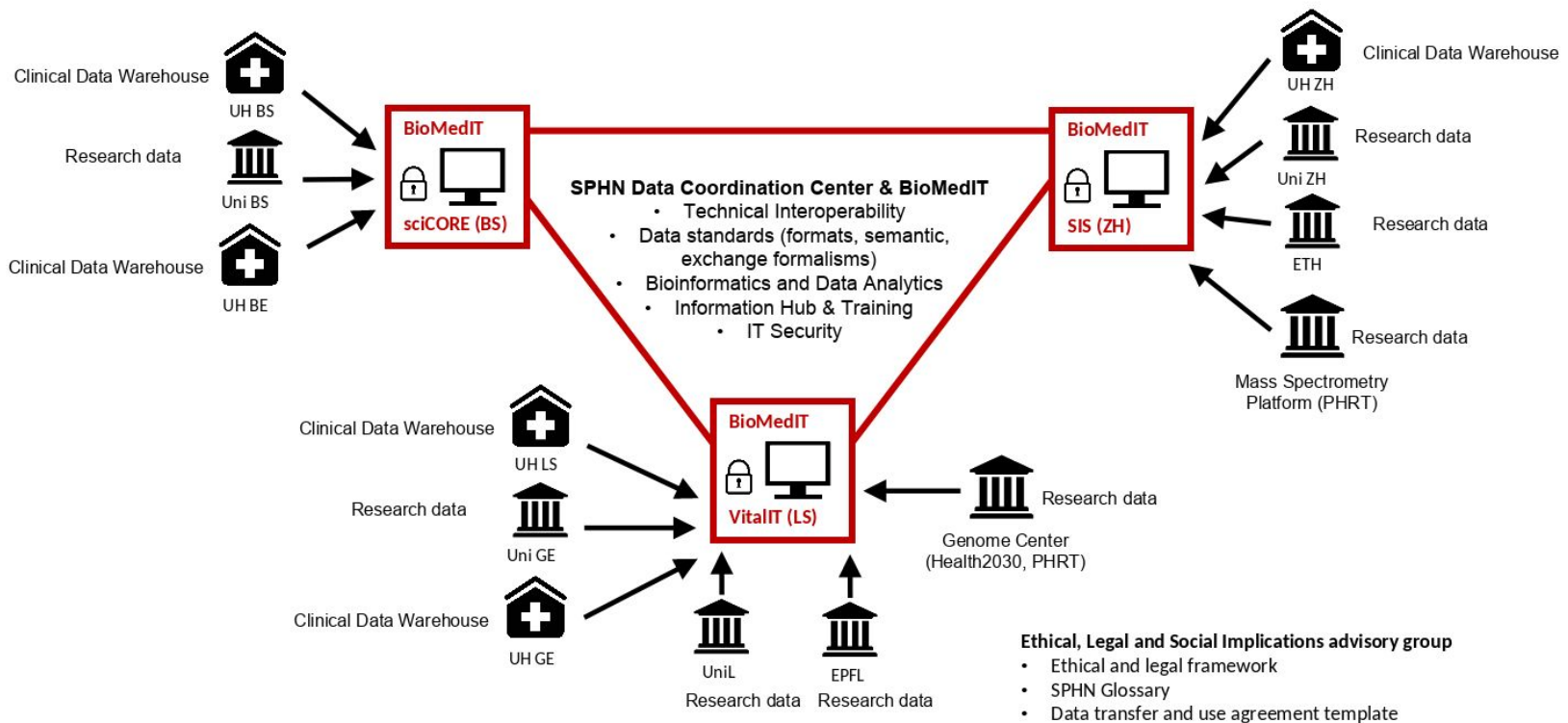


www.sib.swiss

SPHN Context in a nutshell

- **Building an interoperable IT infrastructure for biomedical researchers**
- **BioMedIT nodes provide secure computing environments**
- **Diverse analytical needs from interactive notebooks to HPC and machine learning**
- **Take the compute to the data**
- **Sensitive data!**
- **Early days of BioMedIT analysis**

SPHN IT infrastructure - a decentralized approach

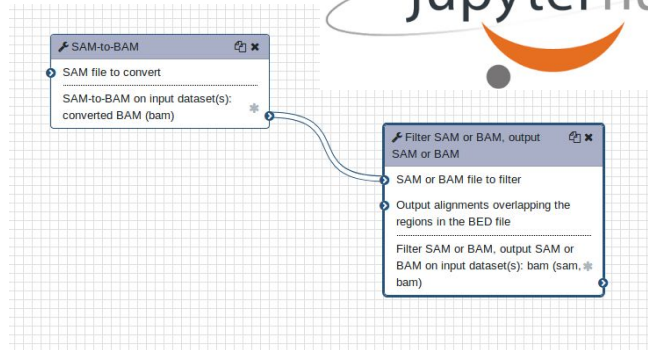
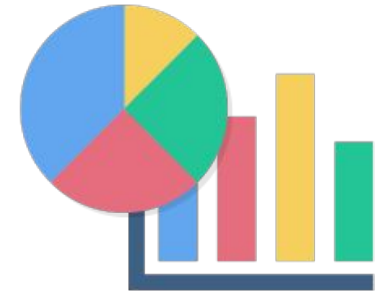


Bioinformatics analysis flow

GGCCCGGCAGCA
GGATGATGCTCTC
CCGGGCCAAGCC
GGCTGTGCGGAG
CACCCCGCCGCA
GGGGGACAGGCG
GAGGAGAAAGG
GAAGAAGGTGCC
ACAGATCG

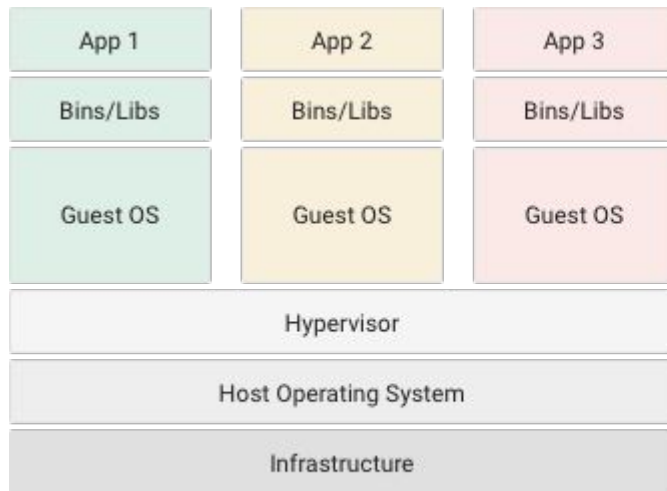
```
Terminal
kevin@kevin-XPS-15-9550:~$ nextflow run nextflow-io/rnaseq-nf -with-docker
NEXTFLOW - version 18.10.1
Launching 'nextflow-io/rnaseq-nf' [jovial_gautier] - revision: 9740761258 [master]
NOTE: Your local project version looks outdated - a different revision is available in the remote repository [a81115ecae]
RNASEQ - NF PIPELINE
=====
transcriptome: /home/kevin/.nextflow/assets/nextflow-io/rnaseq-nf/data/ggal/ggal_liver_1_48850000_49020000.Ggal71.500bpflank.fa
reads         : /home/kevin/.nextflow/assets/nextflow-io/rnaseq-nf/data/ggal/*_{1,2}.fq
outdir        : results

[warm up] executor > local
[38/561ca3] Submitted process > fastqc (FASTQC on ggal_out)
[43/9921ec] Submitted process > index (ggal_liver_1_48850000_49020000)
[c2/2c46a1] Submitted process > fastqc (FASTQC on ggal_liver)
```

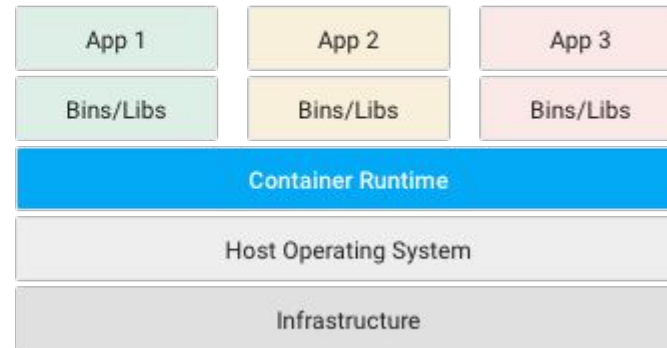


What are containers

- Provide a method for packaging application and dependencies
- Container runtimes: Docker and Singularity
- Portable across platforms



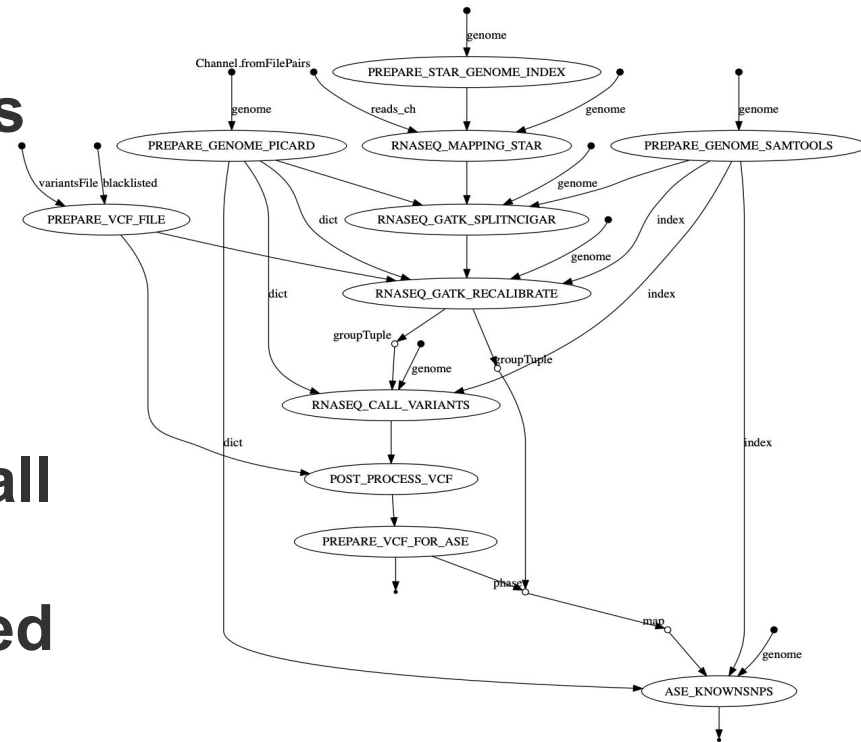
Virtual Machines



Containers

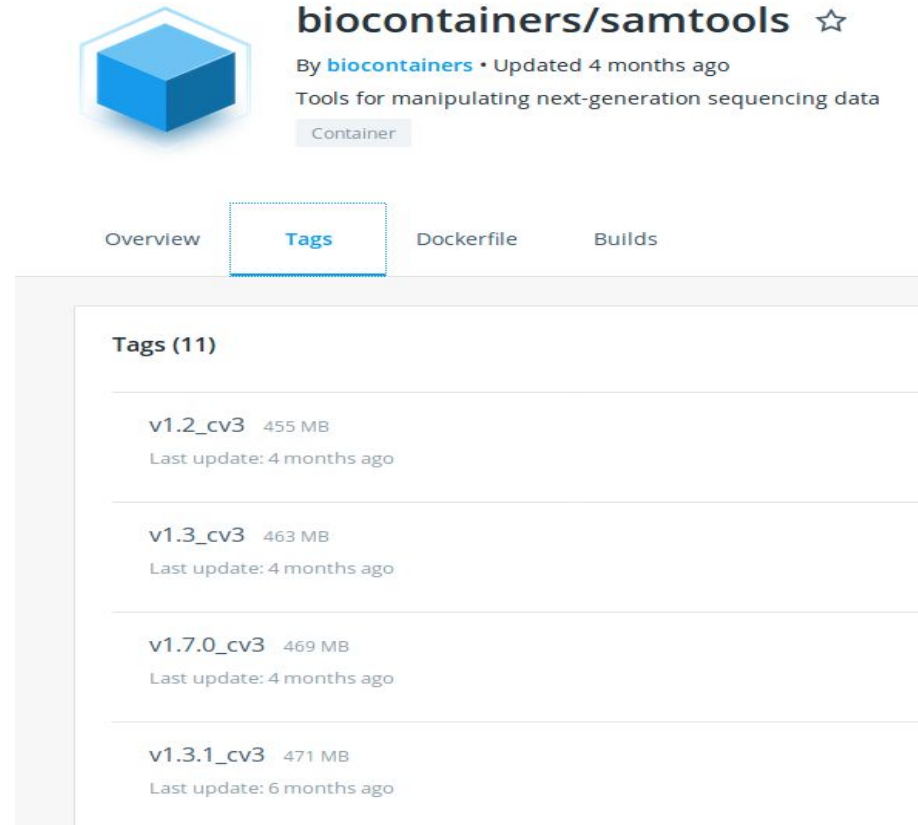
Containers address a need in bioinformatics


- **Bioinformatic workflows combine many analysis tools**
- **Common to have conflicting dependencies**
- **Tools can be difficult to install**
- **Bioinformaticians are focused first and foremost on the analysis**



Reproducibility

- **Containers provided a way to pull specific versions and monitor hash of images**
- **Clinically very important**
- **Example: Differential gene expression, minor differences when using non-containerized versions (<https://doi.org/10.1038/nbt.3820>)**



 **biocontainers/samtools** ☆
By [biocontainers](#) • Updated 4 months ago
Tools for manipulating next-generation sequencing data
Container

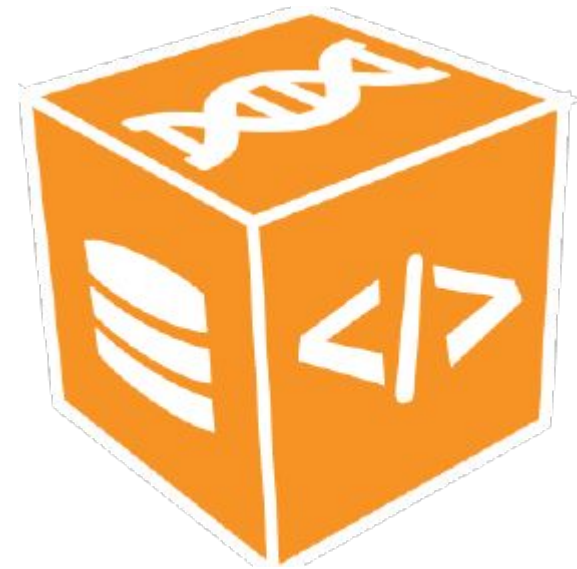
Overview **Tags** Dockerfile Builds

Tags (11)

v1.2_cv3	455 MB	Last update: 4 months ago
v1.3_cv3	463 MB	Last update: 4 months ago
v1.7.0_cv3	469 MB	Last update: 4 months ago
v1.3.1_cv3	471 MB	Last update: 6 months ago

BioContainers for bioinformatic tools

- **Community based container registry for bioinformatics tools**
- **Built using Bioconda recipes**
- **Over 4000 containers**
- **17.8M pulls for Samtools on quay.io**
- **Explicit version tagging to promote reproducibility (no latest tag)**



Where do we go now that we have a container

- **Containerizing bioinformatics tools seems valuable now what?**
- **Workflows could be run in containers manually**
- **HPC schedulers can be used to run a container**
- **“Cloud native” solutions**

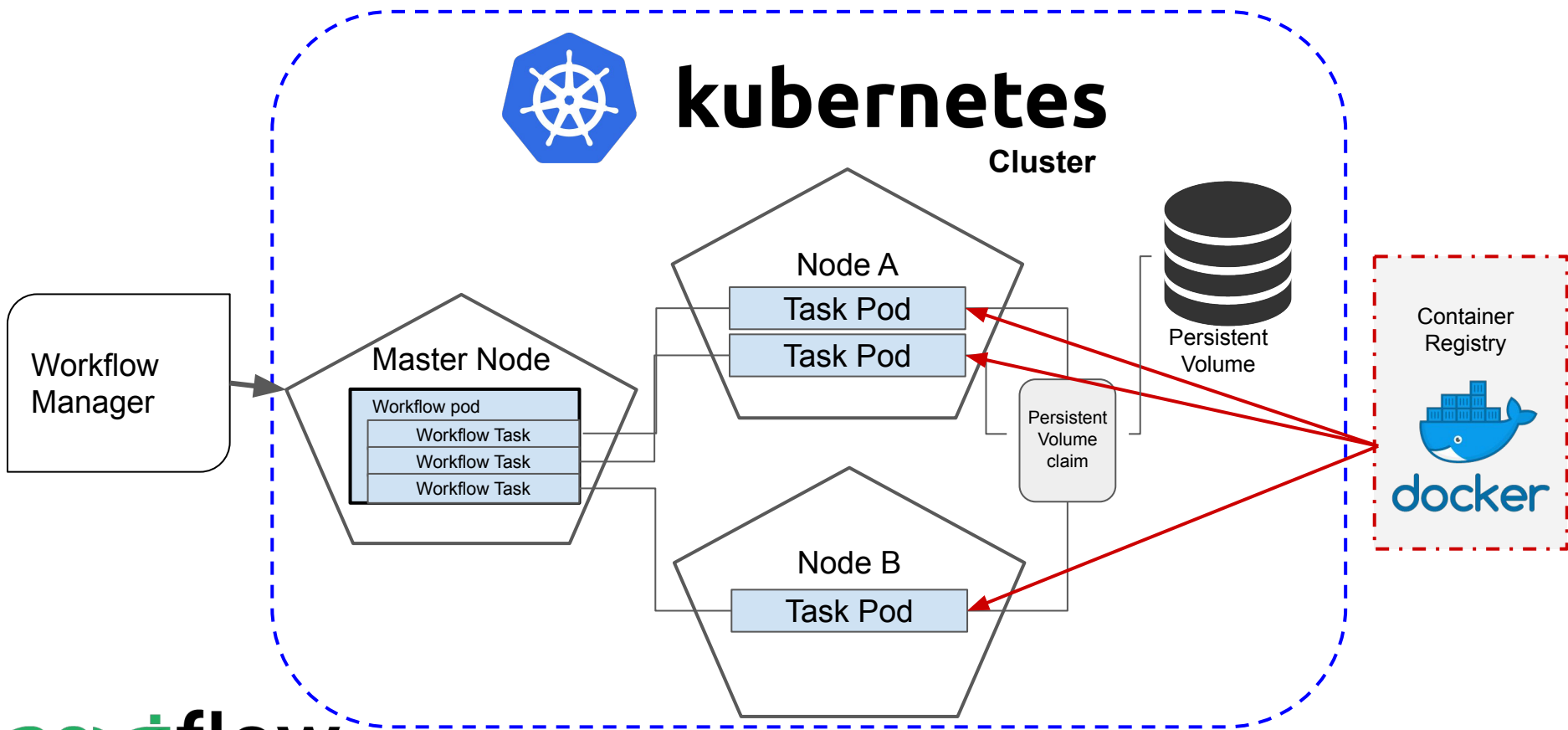
Kubernetes

- **Container orchestration system**
- **Launch containers across multiple hosts, a pod is one or more containers**
- **Abstraction of networking and storage**
- **Provides methods for controlling where pods are scheduled**
- **On-premises or Cloud based**

Kubernetes for bioinformatics analysis

- **Helm charts can be used to deploy JupyterHub and Galaxy**
- **Libraries for parallelizing code on Kubernetes such as Python Dask**
- **Deploy apps such as R Shiny**
- **Workflow managers can run batch analysis jobs on Kubernetes**

Batch processing on Kubernetes



nextflow

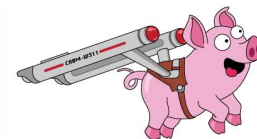


reana

Galaxy
PROJECT



TESK



COMMON
WORKFLOW
LANGUAGE

A concrete batch example

nextflow

```
...
process index {
  container 'combinelab/salmon'
  pod = [nodeSelector: 'memory=high']

  input:
  file transcriptome from transcriptome_file

  output:
  file 'index' into index_ch

  script:
  """
  salmon index $transcriptome -i index
  """
}
```

Pod YAML
File



On prem Kubernetes
cluster




Google Kubernetes Engine

```
> nextflow kuberun KevinSayers/rnaseq-nf -v nfs:/pvdata
```

Elsewhere in the Community



Developing specifications for deploying containerized workflows for genomics. Including APIs for batch processes (TES/WES) and registries with containerized tools (TRS/Biocontainers) 



ELIXIR Compute Platform focusing on developing community wide platform for running containerized workflows. Current work on determining requirements for analysis on sensitive data. Kubernetes implementation of GA4GH TES API



Swiss Data Science Center developing RENKU, a data science platform with provenance tracking and auto workflow generation, which is deployed on Kubernetes

Challenges

- **Security of containers. Pulling containers from public registries, escaping the container**
- **Security of Kubernetes, multiple projects on a single cluster vs. project specific clusters**
- **Federation of Kubernetes clusters**
- **Singularity as a container runtime in Kubernetes**

In short

- **Containers are seeing recent adoption in the field of bioinformatics**
- **Kubernetes is flexible enough to run batch workflows and interactive tools**
- **Further work on security and federation needed**