Theiagen®

GENOMICS

We will be starting soon

Docker for Public Health Bioinformatics

Week 04 - StaPH-B/docker-builds project & Review

PRESENTED BY:

Inês Mendes, PhD

A Mid-Atlantic Workforce Development Offering Provided by the Division of Consolidated Laboratory Services in Collaboration with Theiagen Genomics

Course Introduction

Training Workshop Resources

Training Information, Communication, and Support

- **GitHub Repository** created to host training resources and information:
  - https://github.com/theiagen/Mid-Atlantic-Docker4PH-2025
- **Support contact:**
  - support@theiagen.com

Theiagen® GENOMICS

Course Agenda

Docker for Public Health Bioinformatics

Week 4 - April 15/17, 2025

- StaPH-B/docker-builds project & Review
- Hands-on Exercise: Employing StaPH-B Docker Images

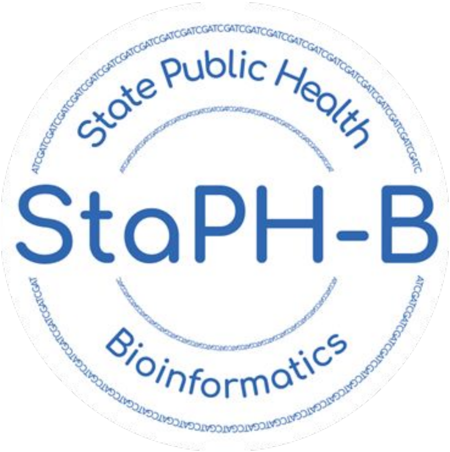
OBJECTIVE

Goals by End of Week 4

- Learn the history & goals of the StaPH-B/docker-builds project
- Learn strategies for contributing to the StaPH-B/docker-builds project
- Review course content from weeks 1-3

StaPH-B/ docker-builds project

What is StaPH-B?



State Public Health Bioinformatics working group

- Started in 2017
- Public health scientists interested in addressing common barriers
- **Mission:**
    - Support construction and maintenance of bioinformatics infrastructure in state & local Public Health laboratories
    - Provide training and resources for fundamentals and practice of bioinformatics
    - Development of bioinformatics resources including tools, pipelines, and documentation

What is StaPH-B?

State Public Health Bioinformatics working group

- Partner with CDC and APHL for coordination and support
- https://staphb.org/
  - Join us! Slack invite link: https://join.slack.com/t/staph-b-dev/shared_invite/zt-w4ivhtq9-2XypNGWXY9vmyeWf0LABng

Barriers to bioinformatics in PHLs

- Vast landscape of compute infrastructure
  - On-premise servers/workstations
  - High performance compute cluster
  - Public cloud
  - None
- Limited experience working with open source software (OSS)
- Limited IT support beyond typical desktop/network support

StaPH-B/docker-builds

- Project started in Sep 2018, shortly after StaPH-B started
  - Led by former APHL/CDC bioinformatics & AR fellows
    - Curtis Kapsak, Kelsey Florek, Erin Young, Kevin Libuit, others
- CO state PHL - containerized their bioinformatics workflows for bacterial WGS & phylogenetic analysis
  - Similar containerization efforts began at other state & local PHLs
  - PulseNet made the switch to WGS in 2019
- Many labs starting to adopt cloud resources
  - **We needed a way to easily install and run bioinformatics software in a reproducible manner**

StaPH-B/docker-builds

Goals

- **Improve distribution of OSS used for PH bioinformatics analyses**
  - Provide access to freely available software that can run on any compute infrastructure
- **Maximize reproducibility of analyses**
  - CLIA/CAP validation
- **Simplify bioinformatics workflow development**
  - Spend more time on the science, less time installing software
- **Provide thorough documentation**
  - The community thanks you for this!

StaPH-B/docker-builds

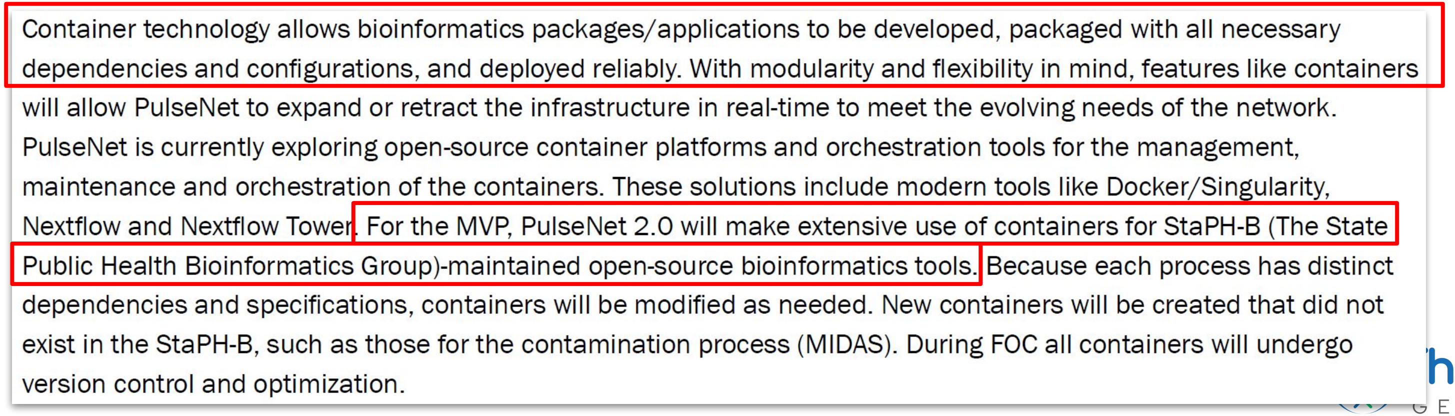
A few bioinformatics workflows that utilize StaPH-B docker containers:

- All Theiagen WDL workflows (TheiaCov, TheiaProk, TheiaEuk, all utility workflows, etc.)
- C-BIRD bacterial WGS workflow from Kutluhan & the CT PHL
  - https://github.com/Kincekara/C-BIRD
- UT PHL
  - Grandeur - https://github.com/UPHL-BioNGS/Grandeur
  - Cecret - https://github.com/UPHL-BioNGS/Cecret
- StaPH-B toolkit
  - https://github.com/StaPH-B/staphb_toolkit
- WI PHL
  - spriggan - https://github.com/wslh-bio/spriggan
  - dryad - https://github.com/wslh-bio/dryad

StaPH-B/docker-builds

A few bioinformatics workflows that utilize StaPH-B docker containers:

- PulseNet 2.0 workflows
  - Excerpt from PulseNet 2.0 White Paper:
  - https://www.aphl.org/aboutAPHL/publications/Documents/PulseNet-2.0-White-Paper.pdf

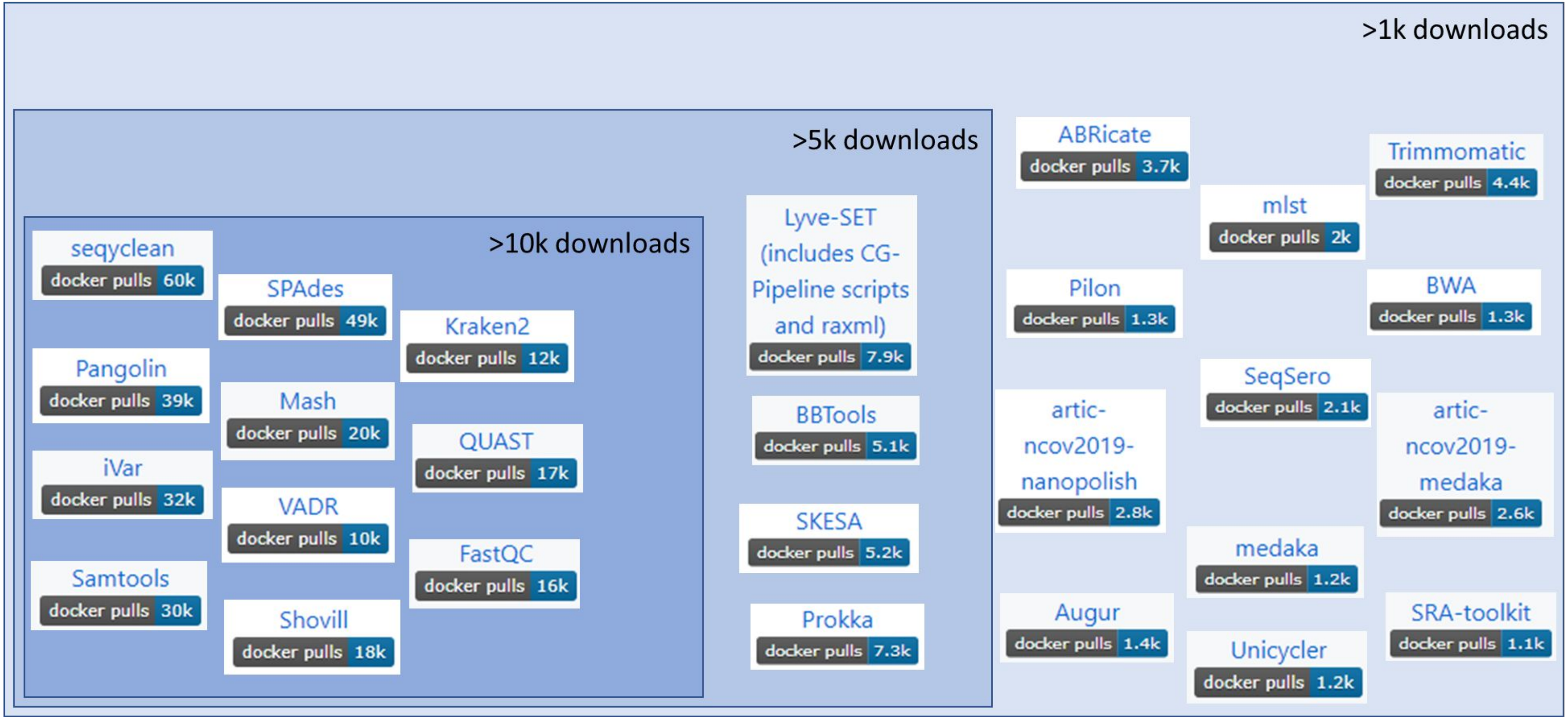


Container technology allows bioinformatics packages/applications to be developed, packaged with all necessary dependencies and configurations, and deployed reliably. With modularity and flexibility in mind, features like containers will allow PulseNet to expand or retract the infrastructure in real-time to meet the evolving needs of the network. PulseNet is currently exploring open-source container platforms and orchestration tools for the management, maintenance and orchestration of the containers. These solutions include modern tools like Docker/Singularity, Nextflow and Nextflow Tower. For the MVP, PulseNet 2.0 will make extensive use of containers for StaPH-B (The State Public Health Bioinformatics Group)-maintained open-source bioinformatics tools. Because each process has distinct dependencies and specifications, containers will be modified as needed. New containers will be created that did not exist in the StaPH-B, such as those for the contamination process (MIDAS). During FOC all containers will undergo version control and optimization.

StaPH-B/docker-builds

Most downloaded Docker images

# of pulls as of 2021-03-08



>1k downloads

>5k downloads

>10k downloads

seqyclean
docker pulls 60k

SPAdes
docker pulls 49k

Kraken2
docker pulls 12k

Pangolin
docker pulls 39k

Mash
docker pulls 20k

QUAST
docker pulls 17k

iVar
docker pulls 32k

VADR
docker pulls 10k

FastQC
docker pulls 16k

Samtools
docker pulls 30k

Shovill
docker pulls 18k

Lyve-SET (includes CG-Pipeline scripts and raxml)
docker pulls 7.9k

BBTools
docker pulls 5.1k

SKESA
docker pulls 5.2k

Prokka
docker pulls 7.3k

ABRicate
docker pulls 3.7k

Trimmomatic
docker pulls 4.4k

mlst
docker pulls 2k

Pilon
docker pulls 1.3k

BWA
docker pulls 1.3k

SeqSero
docker pulls 2.1k

artic-ncov2019-nanopolish
docker pulls 2.8k

artic-ncov2019-medaka
docker pulls 2.6k

medaka
docker pulls 1.2k

Augur
docker pulls 1.4k

SRA-toolkit
docker pulls 1.1k

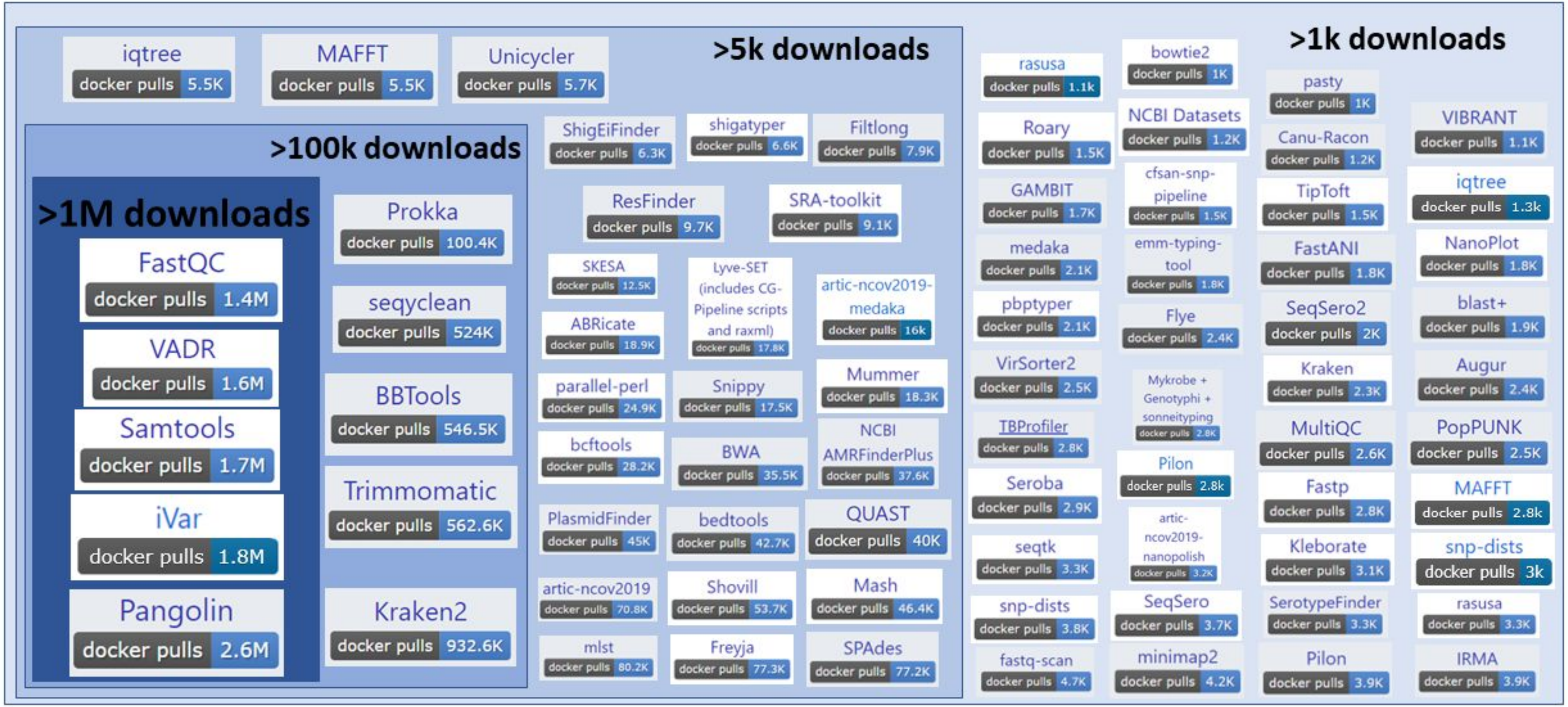
Unicycler
docker pulls 1.2k

StaPH-B/docker-builds

Most downloaded Docker images

of pulls reported by DockerHub as of 2023-06-10

State Public Health StaPH-B Bioinformatics



>1M downloads

FastQC	docker pulls	1.4M
VADR	docker pulls	1.6M
Samtools	docker pulls	1.7M
iVar	docker pulls	1.8M
Pangolin	docker pulls	2.6M

>100k downloads

Prokka	docker pulls	100.4K
seqyclean	docker pulls	524K
BBTools	docker pulls	546.5K
Trimmomatic	docker pulls	562.6K
Kraken2	docker pulls	932.6K

>5k downloads

iqtree	docker pulls	5.5K
MAFFT	docker pulls	5.5K
Unicycler	docker pulls	5.7K
ShigEiFinder	docker pulls	6.3K
shigatyper	docker pulls	6.6K
Filtlong	docker pulls	7.9K
ResFinder	docker pulls	9.7K
SRA-toolkit	docker pulls	9.1K
SKESA	docker pulls	12.5K
ABRicate	docker pulls	18.9K
parallel-perl	docker pulls	24.9K
bcftools	docker pulls	28.2K
PlasmidFinder	docker pulls	45K
artic-ncov2019	docker pulls	70.8K
mlst	docker pulls	80.2K
Lyve-SET (includes CG-Pipeline scripts and raxml)	docker pulls	17.8K
Snippy	docker pulls	17.5K
BWA	docker pulls	35.5K
bedtools	docker pulls	42.7K
Shovill	docker pulls	53.7K
Freyja	docker pulls	77.3K
artic-ncov2019-medaka	docker pulls	16k
Mummer	docker pulls	18.3K
NCBI AMRFinderPlus	docker pulls	37.6K
QUAST	docker pulls	40K
Mash	docker pulls	46.4K
SPAdes	docker pulls	77.2K

>1k downloads

rasusa	docker pulls	1.1k
Roary	docker pulls	1.5K
GAMBIT	docker pulls	1.7K
medaka	docker pulls	2.1K
pbptyper	docker pulls	2.1K
VirSorter2	docker pulls	2.5K
TBProfiler	docker pulls	2.8K
Seroba	docker pulls	2.9K
seqtk	docker pulls	3.3K
snp-dists	docker pulls	3.8K
fastq-scan	docker pulls	4.7K
bowtie2	docker pulls	1K
NCBI Datasets	docker pulls	1.2K
cfsan-snp-pipeline	docker pulls	1.5K
emm-typing-tool	docker pulls	1.8K
Flye	docker pulls	2.4K
Mykrobe + Genotyphi + sonneityping	docker pulls	2.8K
Pilon	docker pulls	2.8k
artic-ncov2019-nanopolish	docker pulls	3.2K
SeqSero	docker pulls	3.7K
minimap2	docker pulls	4.2K
pasty	docker pulls	1K
Canu-Racon	docker pulls	1.2K
TipToft	docker pulls	1.5K
FastANI	docker pulls	1.8K
SeqSero2	docker pulls	2K
Kraken	docker pulls	2.3K
MultiQC	docker pulls	2.6K
Fastp	docker pulls	2.8K
Kleborate	docker pulls	3.1K
SerotypeFinder	docker pulls	3.3K
Pilon	docker pulls	3.9K
VIBRANT	docker pulls	1.1K
iqtree	docker pulls	1.3k
NanoPlot	docker pulls	1.8K
blast+	docker pulls	1.9K
Augur	docker pulls	2.4K
PopPUNK	docker pulls	2.5K
MAFFT	docker pulls	2.8k
snp-dists	docker pulls	3k
rasusa	docker pulls	3.3K
IRMA	docker pulls	3.9K

StaPH-B/docker-builds Summary

- The field of public health bioinformatics has adopted container technologies
- Use of docker containers addresses barriers that are common to public health labs
  - Increases portability
  - Increases reproducibility
  - Simplifies of bioinfo workflow development
- Community-led effort!

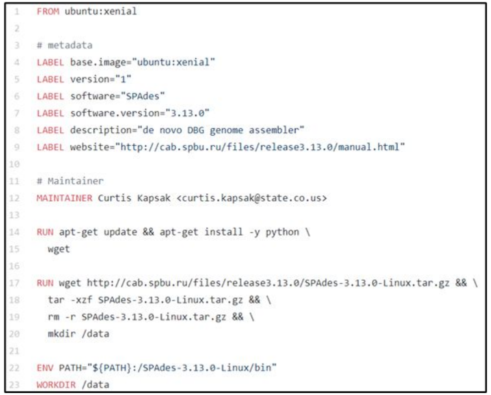
Week 1 - 3 Review

Week 1 Review

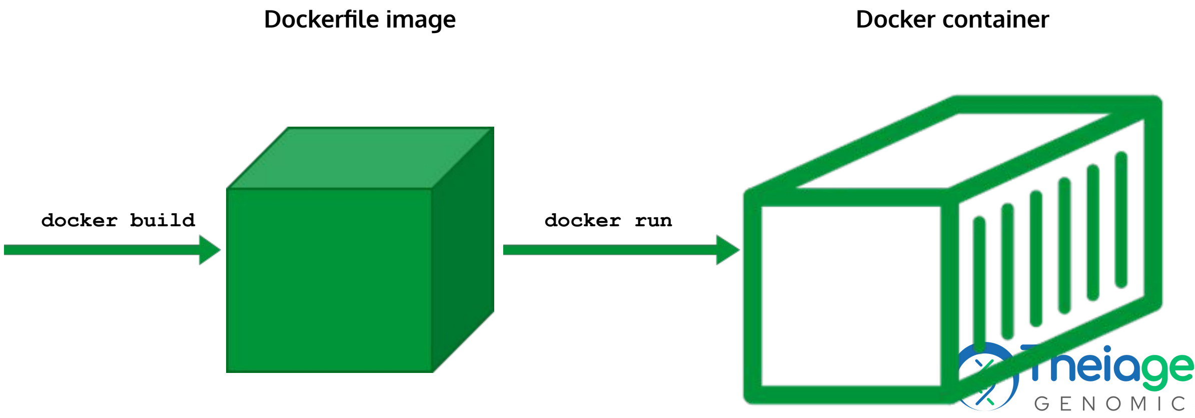
Summary:

- **Dockerfile** is used to create the docker **image**
- Docker **image** is used to create the docker **container**
    - Container **is the runnable instance of an image**

Dockerfile



```
FROM ubuntu:xenial

# metadata
LABEL base.image="ubuntu:xenial"
LABEL version="1"
LABEL software="SPAdes"
LABEL software.version="3.13.0"
LABEL description="de novo DBG genome assembler"
LABEL website="http://cab.spbu.ru/files/release3.13.0/manual.html"

# Maintainer
MAINTAINER Curtis Kapsak <curtis.kapsak@state.co.us>

RUN apt-get update && apt-get install -y python \
  wget

RUN wget http://cab.spbu.ru/files/release3.13.0/SPAdes-3.13.0-Linux.tar.gz && \
  tar -xzf SPAdes-3.13.0-Linux.tar.gz && \
  rm -r SPAdes-3.13.0-Linux.tar.gz && \
  mkdir /data

ENV PATH="${PATH}:/SPAdes-3.13.0-Linux/bin"
WORKDIR /data
```



Week 1 review

Docker Images can be built locally **or** pre-built images can be downloaded from public repositories like:

- **Docker hub:** https://hub.docker.com/
- **Quay.io:** https://quay.io/
- **GitHub container registry (GHCR):** https://ghcr.io/
- Cloud provider container registries:
  - GCP Artifact Registry
  - Amazon Elastic Container Registry
  - Microsoft Azure Container Registry
- Private registries are an (paid) option

Week 1 Review

- **Docker Hub:** https://hub.docker.com
- **Quay.io:** https://quay.io/

Week 2 Review

Dockerfile instructions

- **`FROM`** defines the base docker image
- **`ARG`** set environmental variables ONLY available during build time
- **`ENV`** set environmental variables that persist during and after build time
- **`RUN`** executes a command in a new layer
- **`WORKDIR`** sets the working directory for executing commands
- **`COPY`** (and **`ADD`**) copy files into the docker image
- **`LABEL`** adds metadata to your docker image

Week 2 Review

Docker build

- Builds an image from a Dockerfile
- At a minimum, requires a Dockerfile. Some dockerfiles require other files for building (scripts, databases, etc.)
- Official docs: https://docs.docker.com/engine/reference/commandline/build/

```
docker build --tag <name>:<tag> <directory-with-dockerfile>
```

```
docker build --tag spades:3.15.5 spades/3.15.5/
```

Week 3 Review

I want to create a dockerfile, where do I start?

- Easiest - Use & modify an existing dockerfile
- A bit more challenging - start from a template dockerfile
- Most challenging - writing a dockerfile from scratch

Week 3 Review

Best practices for writing dockerfiles

- One docker container should be used for one purpose - one bioinfo tool*
- Fewer layers = better. **`RUN`**, **`COPY`**, and **`ADD`** instructions add layers
- No "large" databases or files. Large means >1GB
- Only install what is necessary
- **`docker build`** often while writing Dockerfile. Trial and error as much as necessary!
- Use a Dockerfile linter (Docker VSCode extension) to catch errors before you **`docker build`**

Week 3 Review

More best practices for writing dockerfiles

- Read the tool's documentation. Familiarize yourself with the installation procedure.
- Use **`docker build --progress=plain`** so that all STDOUT/STDERR is printed to screen - can see every command being executed
- If looking for the location of files, launch interactive container to see where files are located: **`docker run -it <image>`**
- alternatively - add **`ls`**, **`find`**, or other commands in your dockerfile
- Make sure that required files (scripts, databases, etc. files) are **readable and executable to all users**. You may have to use **`chmod`** command to change permissions on files

Hands-On Exercise

Exercise 04: Employing StaPH-B Docker Images

Exercise Goal:

- Navigate to StaPH-B Docker repository
- [Optional] Contribute to StaPH-B's repository

Post-Training Feedback Form

Post-Training Feedback Form

- **Anonymous feedback form** to evaluate course delivery:
    - https://forms.gle/q55XabtYLhjpHCPf9
- **Support materials:**
    - **GitHub Repository** created to host training resources and information:
        - https://github.com/theiagen/Mid-Atlantic-Docker4PH-2025
    - **Support contact:**
        - support@theiagen.com

Theiagen®

GENOMICS

www.theiagen.com
support@theiagen.com