NE BRR Training Session June 12th, 2023

Welcome to a NE BRR Training Session hosted by Theiagen Genomics & the Massachusetts Department of Public Health

We appreciate your punctuality! Please give others a few minutes to arrive (and adjust their audio equipment). We will get started at **2:33 PM Eastern Time**. Thanks!

Docker for Public Health Bioinformatics

Week 4 - StaPH-B/docker-builds project & review

Monday June 12th, 2023

Curtis Kapsak, MS & Frank Ambrosio, MS | Theiagen Genomics

Course Introduction

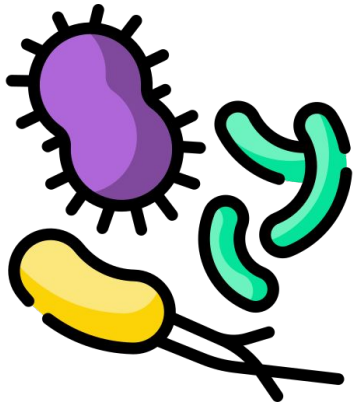
Training Workshop Overview

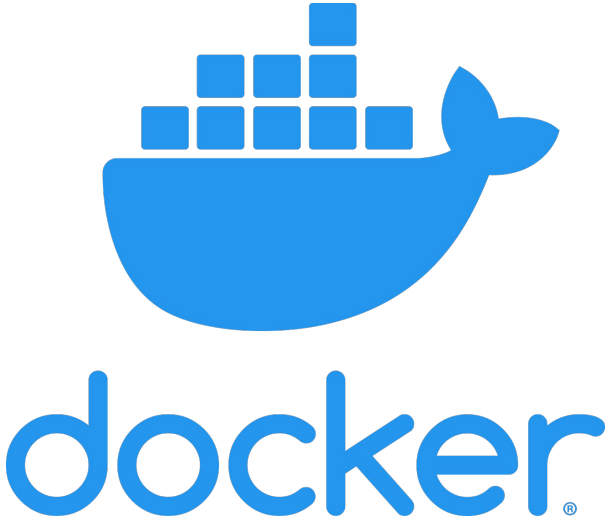
Training Information, Communication, and Support

- **Training Notion Page** created to host training resources and information
- **Support Contacts**:
  - support@terrapublichealth.zendesk.com

Main Course Objective

Learn about the concepts of Docker & containerization and their applications in public health bioinformatics





Training Workshop Overview

This workshop is an Intermediate/Advanced course

Great resources for more information regarding containers and pathogen genomics

For more **technical content**, get connected with various **pathogen genomics communities** such as **PHA4GE, StaPH-B, & micro-binfie**

- **StaPH-B Docker User Guide**
- **Ten Recommendations for supporting open pathogen genomic analysis in public health**
  - Highlights containers and workflow management systems in context of public health
- **A Primer on Infectious Disease Bacterial Genomics**
  - Introduction to analyzing pathogen genomics data

Course Structure

4-Week Virtual Training Workshop

- **All training sessions** will begin at 2:30pm Eastern Time
  - Live Lectures (90m) on Mondays
  - Office Hours (60m) on Wednesdays
  - Exceptions:
    - No sessions the week of APHL Annual conf. (May 22-25)
    - Week 2 lecture will occur Tue May 30th 2:30-4pm EST due to Memorial Day
- **Live lectures** will include **hands-on exercises**
  - To participate, please ensure that you have registered for a GitHub account

Course Content

Week One - Intro to Docker and Containerization

- **Lecture Content**: Introduction to Docker containers
- **Hands-on Exercises**: Utilize a docker container to download a *Klebsiella pneumoniae* genome and to run Kleborate

Week Two - Container Repositories and Writing Dockerfiles

- **Lecture Content**: Intro to various repositories for Docker containers e.g. StaPH-B docker-builds & biocontainers
- **Hands-On Exercise**: Build docker images using pre-existing dockerfiles

Course Content

Week Three - Developing custom Docker Images

- **Lecture Content**: Intro to development and testing practices for writing dockerfiles
- **Hands-on Exercise**: Create a new dockerfile NCBI datasets; assign homework: contribute dockerfile to StaPH-B docker-builds

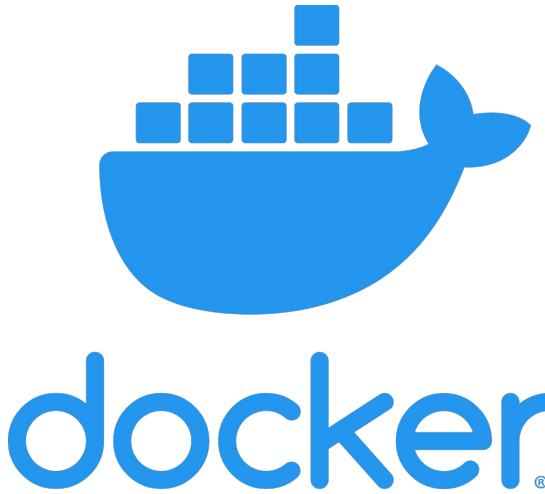
Week Four - StaPH-B docker-builds project

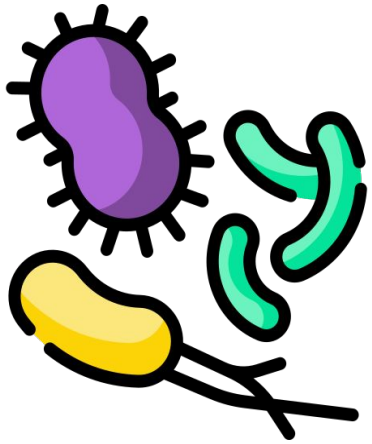
- **Lecture Content**: Review of the StaPH-B docker-builds project and code repository
- **Hands-On Exercise**: Develop a dockerfile and create a pull request

Container Repositories and Writing Dockerfiles

Goals by End of Week Four

- 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





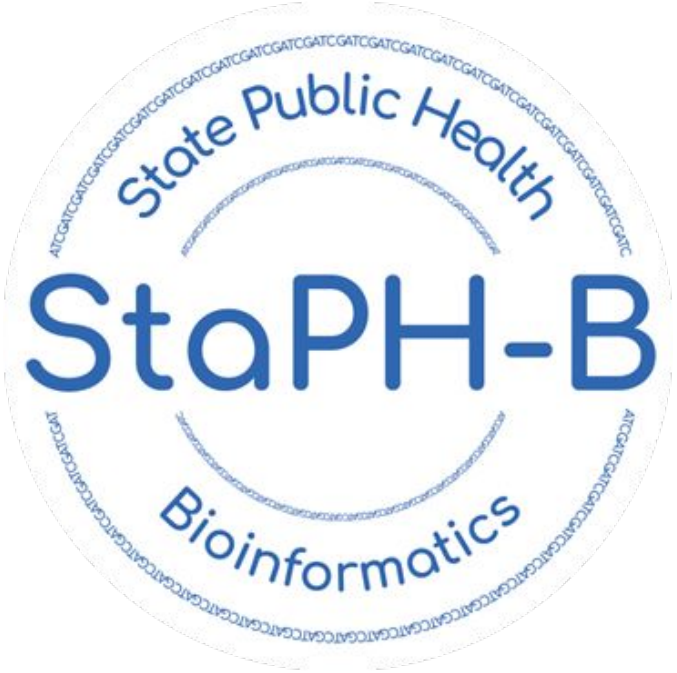
Outline

- StaPH-B & docker-builds project
- Barriers to bioinformatics in PHLs
- Post-training survey (15 min)
- Review Weeks 1,2, & 3
  - Dockerfiles & **`docker build`**
  - Best practices for writing Dockerfiles
  - Strategies for creating and testing dockerfiles
- Homework - update or create your own dockerfile

Week 4
StaPH-B/docker-builds project
&
Review Weeks 1-3

StaPH-B
State Public Health Bioinformatics

What is StaPH-B?



State **P**ublic **H**ealth **B**ioinformatics 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
  - Partner with CDC and APHL for coordination and support
- https://staphb.org/
- Join us! Slack invite link (will expire 2023-07-08): https://join.slack.com/t/staph-b-dev/shared_invite/zt-1wowycedz-fzjCz~XGZpM8HGNuJxJK4w

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
    - myself (Curtis), 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

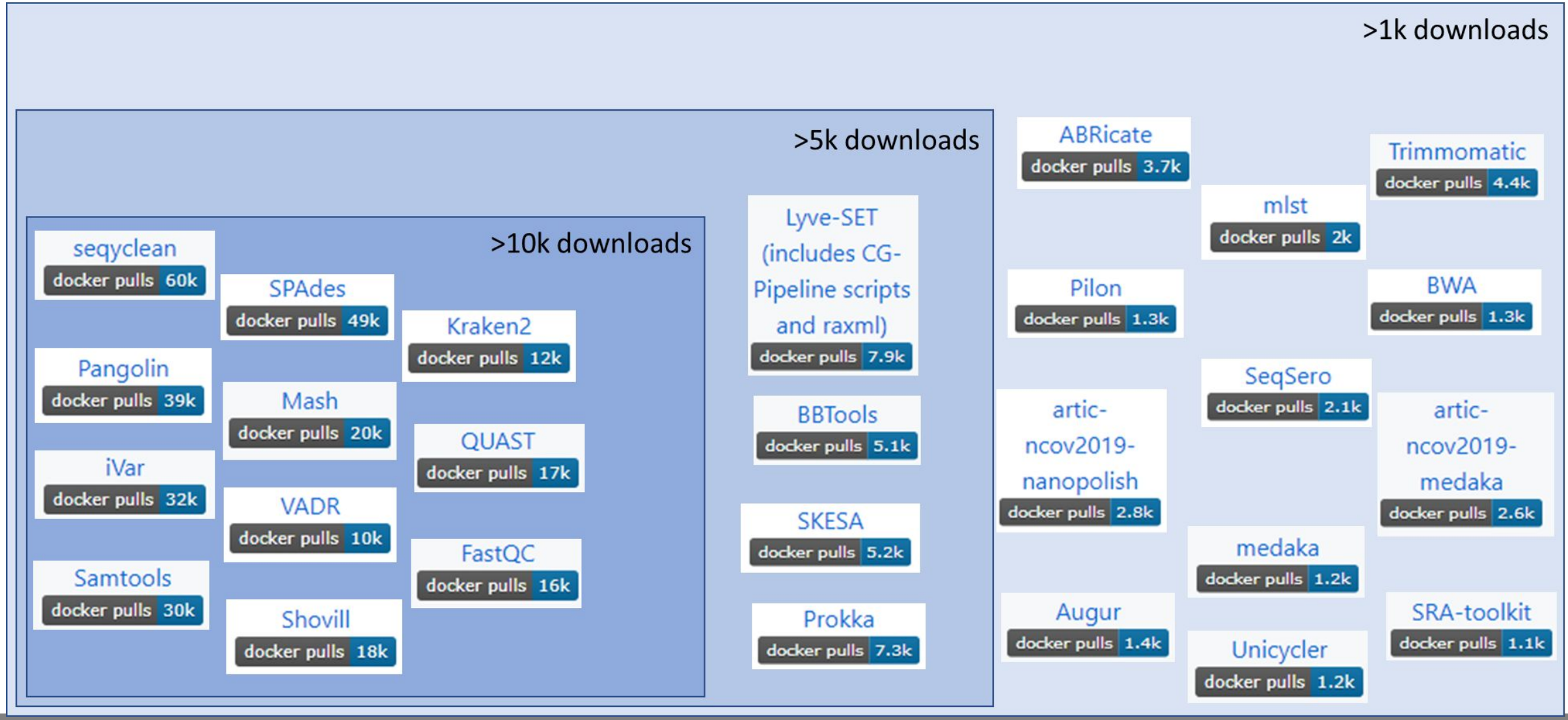
StaPH-B/docker-builds

- 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

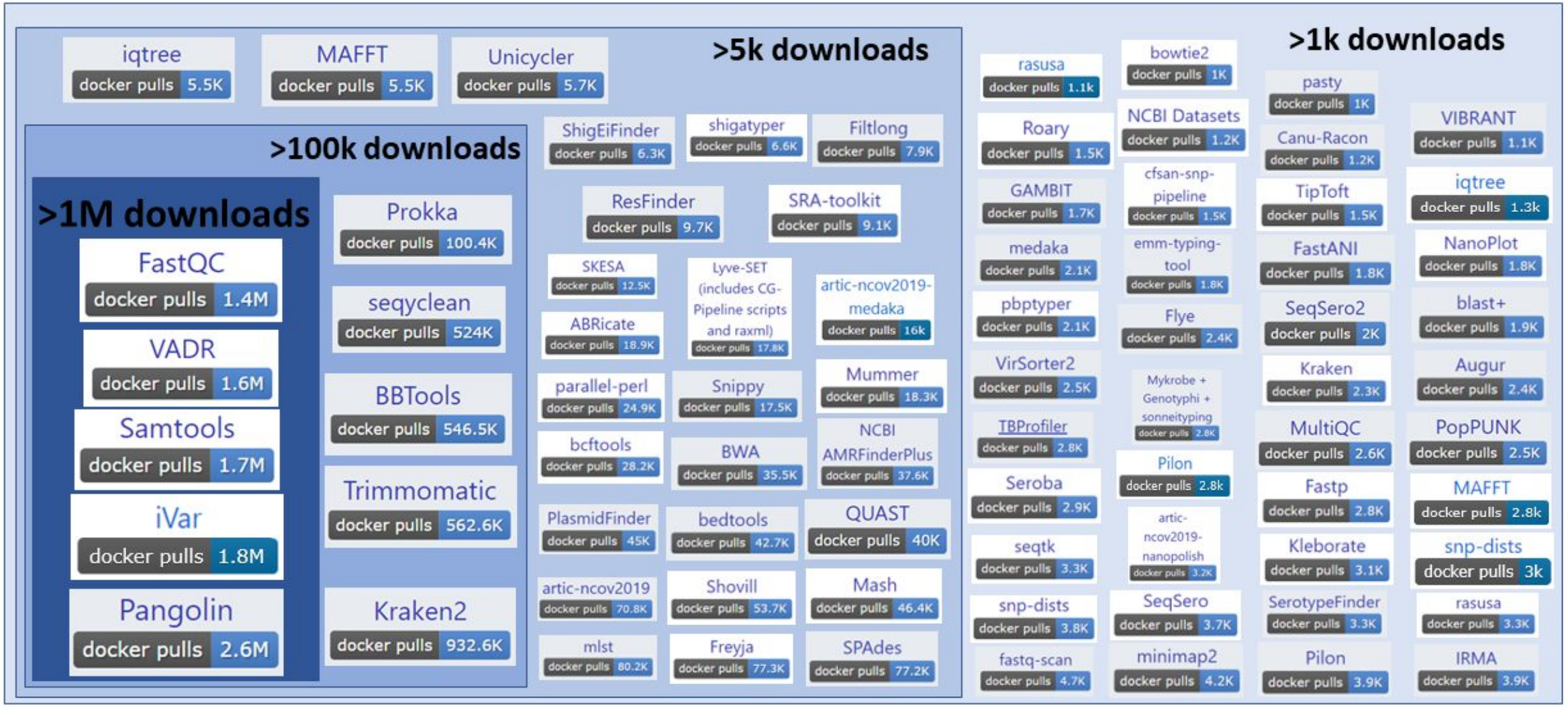
Most downloaded Docker images
of pulls as of 2021-03-08



StaPH-B/docker-builds

Most downloaded Docker images

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



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!

Questions?

Please fill out post-training survey.
10 min. Resume 3:06pm

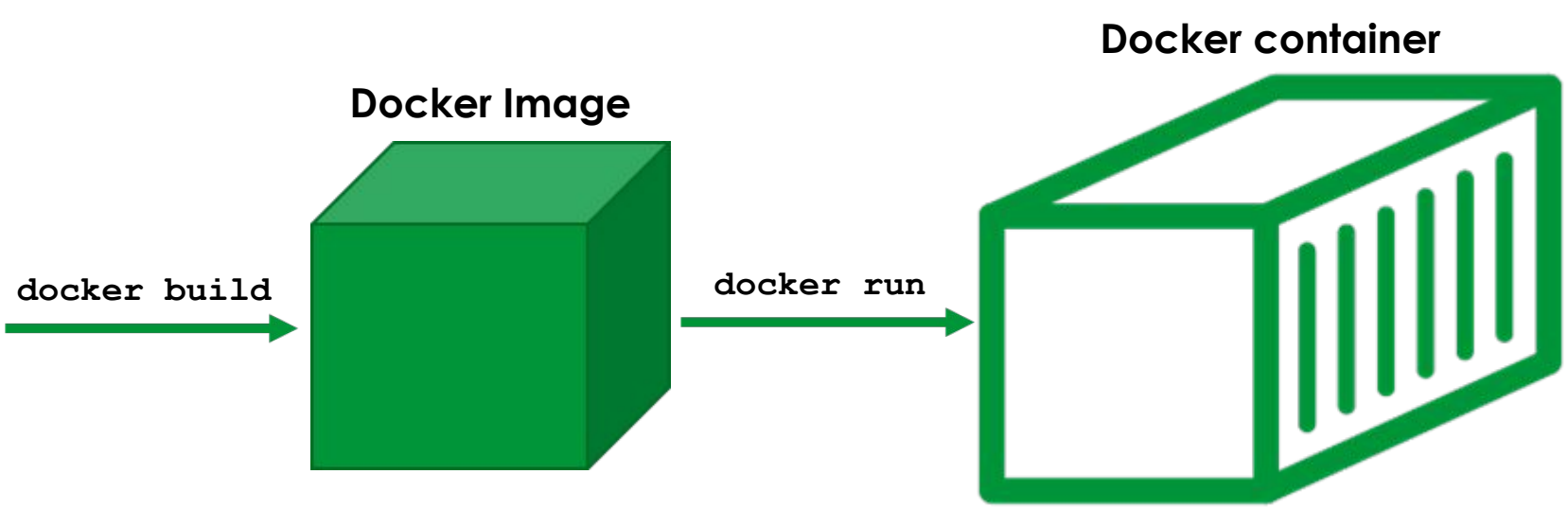
https://forms.gle/kbmL9oCHjPGm2Y639

Week 1 Review

- **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 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
- *There are a few other instructions, but these are the main ones

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/
- General command structure:

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

- example using SPAdes 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

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

Homework!

- Now that we've learned some of the tips and tricks for writing dockerfiles, let's put our knowledge to the test and write a new dockerfile.
- Let's share our dockerfiles & images with the community & contribute to the StaPH-B docker-builds project.
  - https://github.com/StaPH-B/docker-builds
- Please see the separate slide deck with instructions on how to contribute.

Homework!

- bioinfo tools & versions where dockerfiles are needed!
  - beginner/easy
    - update dragonflye v1.1.1 - current dockerfile needs version update. No GitHub issue yet - **Jessie**
    - update fastp v0.23.4 - current dockerfile needs version update. No GitHub issue yet
    - update minimap2 v2.26 - current dockerfile needs version update. No GitHub issue yet
    - update seqkit v2.4.0 - current dockerfile needs version update. No GitHub issue yet - **Luc**
    - update snp-sites v2.5.1 - current dockerfile missing app and test layers. GitHub issue - **Sean**
    - update kSNP3 v3.1 - current dockerfile missing app and test layers. GitHub issue - **Kari**
    - update colorid 0.1.4.3 - current dockerfile missing app and test layers. GitHub issue
    - update hmmer v3.3 - current dockerfile missing app and test layers. GitHub issue - **Neranjan**
    - update clustalo v1.2.4 - current dockerfile missing app and test layers. GitHub issue
  - intermediate
    - seqtk v1.4 - have dockerfile for v1.3, needs updating
    - sra-tools (AKA sra-toolkit) v3.0.5 - have dockerfile for 2.9.2, needs updating
    - krakenuniq 1.0.4 - have dockerfile in progress
  - advanced
    - Krocus v1.0.3 - start w/ dockerfile template
    - Meningotype v0.8.2-beta - start w dockerfile template
    - MIDAS v1.3.2 - start with other example dockerfiles (but tweak to StaPH-B requirements)
    - Samtools - start with existing dockerfile, update with new build stage - **Kutluhan**
  - Have ideas for tools not listed here?

Homework!

- Once you have been assigned a tool, it is your homework to follow the instructions for creating & testing a dockerfile, and submitting a Pull Request via GitHub to contribute your code to the StaPH-B docker-builds project
- Feel free to work on this task at your own pace in the GitPod environment
  - NOTE: you will need to create a new GitPod workspace for this, see slide deck for instructions
- Please use time during office hours this week and next week to ask questions, seek help & advice as you develop.
- Curtis (and potentially other StaPH-B maintainers) will review your code, make suggestions for improvements, help troubleshoot, etc. to guide you through the process

Further reading & resources

- StaPH-B Github repo and docker hub account
  - https://github.com/StaPH-B/docker-builds
  - https://hub.docker.com/u/staphb
- Docker Documentation - a wealth of info here. Note that we use Docker Community Edition, as you have to pay for the Enterprise Edition
  - https://docs.docker.com/
- An awesome tutorial/workshop on docker for bioinformatics
  - https://github.com/PawseySC/bio-workshop-18
- Template for your Dockerfile
  - https://github.com/StaPH-B/docker-builds/blob/master/dockerfile-template/Dockerfile
- Some best practices
  - https://staphb.org/docker-builds/make_containers/
- Search for docker images and (sometimes) Dockerfiles here:
  - http://hub.docker.com/
  - https://quay.io/
- "What is Docker?"(~11 min)
  - https://www.youtube.com/watch?time_continue=1&v=aLipr7tTuA4

Acknowledgements

- MA DPH
- Members of StaPH-B & the docker-builds contributors & maintainers
  - Erin Young, UT PHL
  - Kelsey Florek, WI PHL
  - Kevin Libuit, Theiagen Genomics
  - Frank Ambrosio, Theiagen Genomics
  - many more awesome people!
    - StaPH-B docker-builds contributors: https://github.com/StaPH-B/docker-builds#authorsmaintainers
- APHL
- CDC

StaPH-B
State Public Health Bioinformatics