Bwb tutorial: Agenda

1. Presentation
2. Demo
3. Q&A

Code: https://github.com/BioDepot/BioDepot-workflow-builder

Container: https://hub.docker.com/r/biodepot/bwb

YouTube of tutorial: https://www.youtube.com/watch?v=r_03_UG1mBg

Building containerized workflows for RNA-seq data using the BioDepot-workflow-Builder (Bwb). Hung et al. bioRxiv 099010

Using BioDepot-workflow-Builder (Bwb) to create and execute reproducible bioinformatics workflows

Ling-Hong Hung, lhhung@uw.edu
Wes Lloyd, wlloyd@uw.edu
Ka Yee Yeung, kayee@uw.edu

University of Washington Tacoma

Motivation

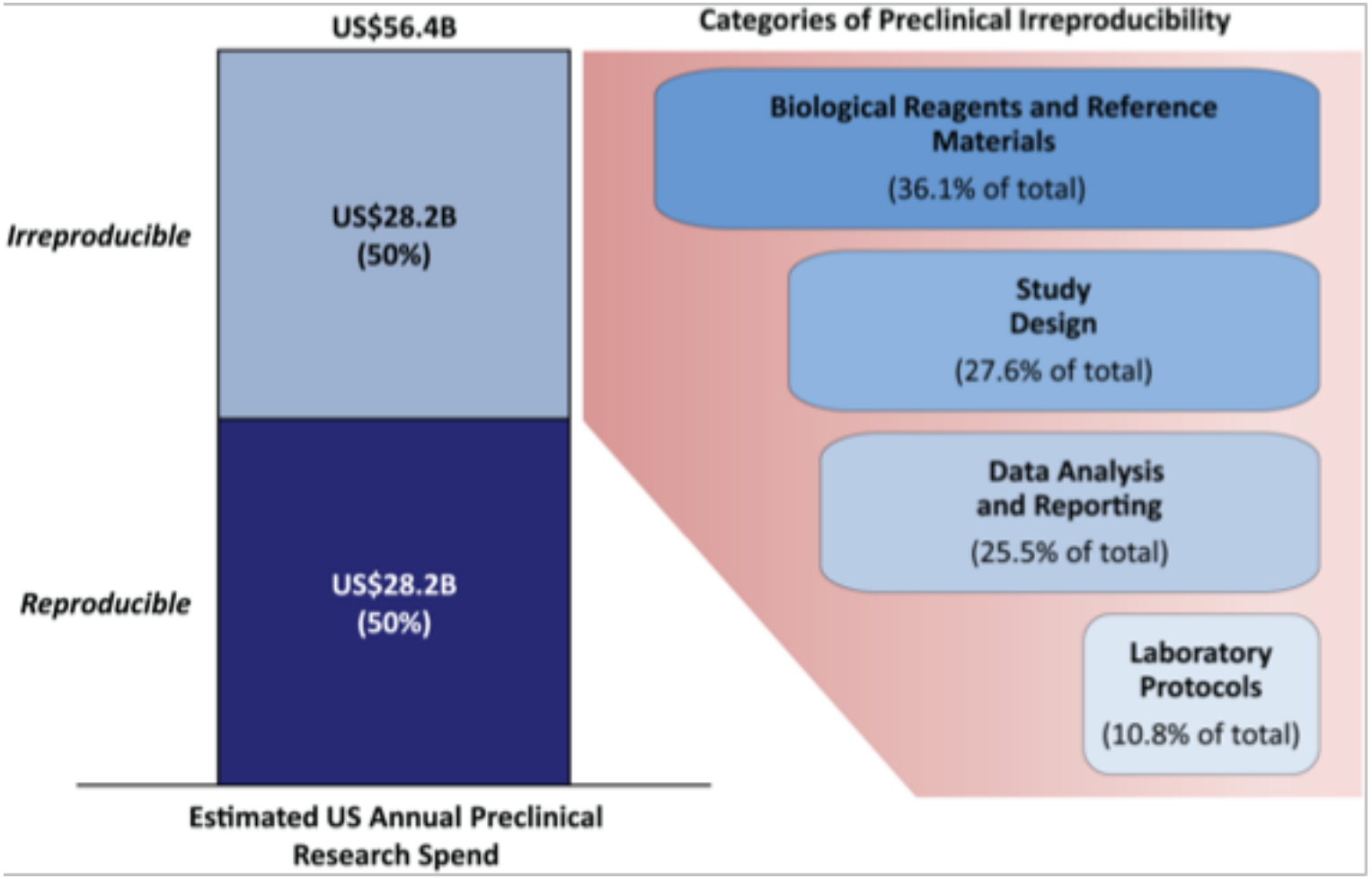
Transitioning from laptop/desktop/cluster based analyses to cloud based analyses

How do we facilitate sharing new workflows reproducibly?

How do we faciltate executing workflows on the cloud?

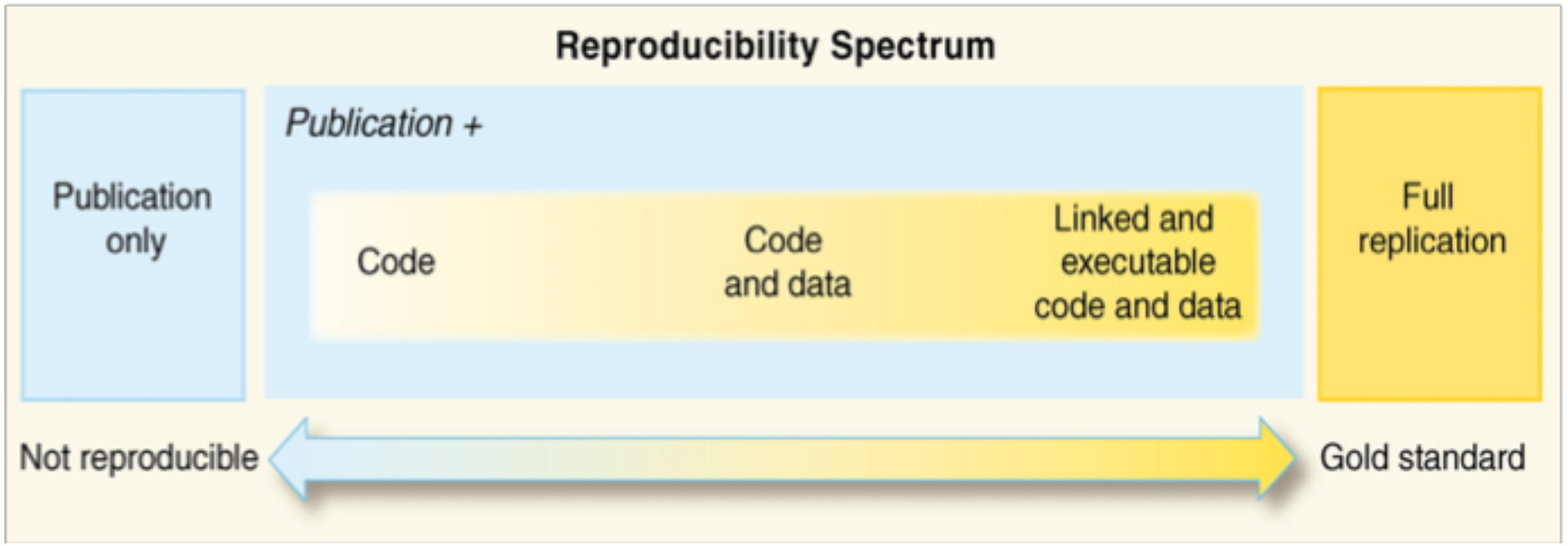
How do we get adoption and input from biologists?

Reproducibility in biology



Freedman et al. “The Economics of Reproducibility in Preclinical Research.” PLOS Biology 2015

The spectrum of reproducibility



Peng (2011). Reproducible research in computational science. *Science,334*(6060), 1226-1227.

Reproducibility exercise on a DREAM Challenge

Students in a 500-level bioinformatics class were asked to reproduce results from a crowd sourcing challenge in 2016-17.

Given: time series gene expression data (time 0, 24 hours) across 4 different viruses (H1N1, H3N2, RSV, Rhinovirus) in 7 studies.

Goal: build predictors to distinguish people who become contagious after exposure to flu and other respiratory viruses.

Question: How many do you think we can reproduce?

Challenges:

- Missing files: e.g. missing pre-processed data, need access to Google drive
- Environment setup: e.g. version conflict, memory, disk space
- Code issues: e.g. don't have software, code simply doesn't run, missing library

Challenges in reproducing bioinformatics analysis results

- Parameters not given or incorrectly given
- Environment assumed i.e. directory structure
- Configuration assumed
- Different OS/software versions
- Bioinformatics analyses usually involve a pipeline consisting of many tools.
- These tools often have graphical user interfaces.

Our solution: We share the computing environment, not just code with Bwb.

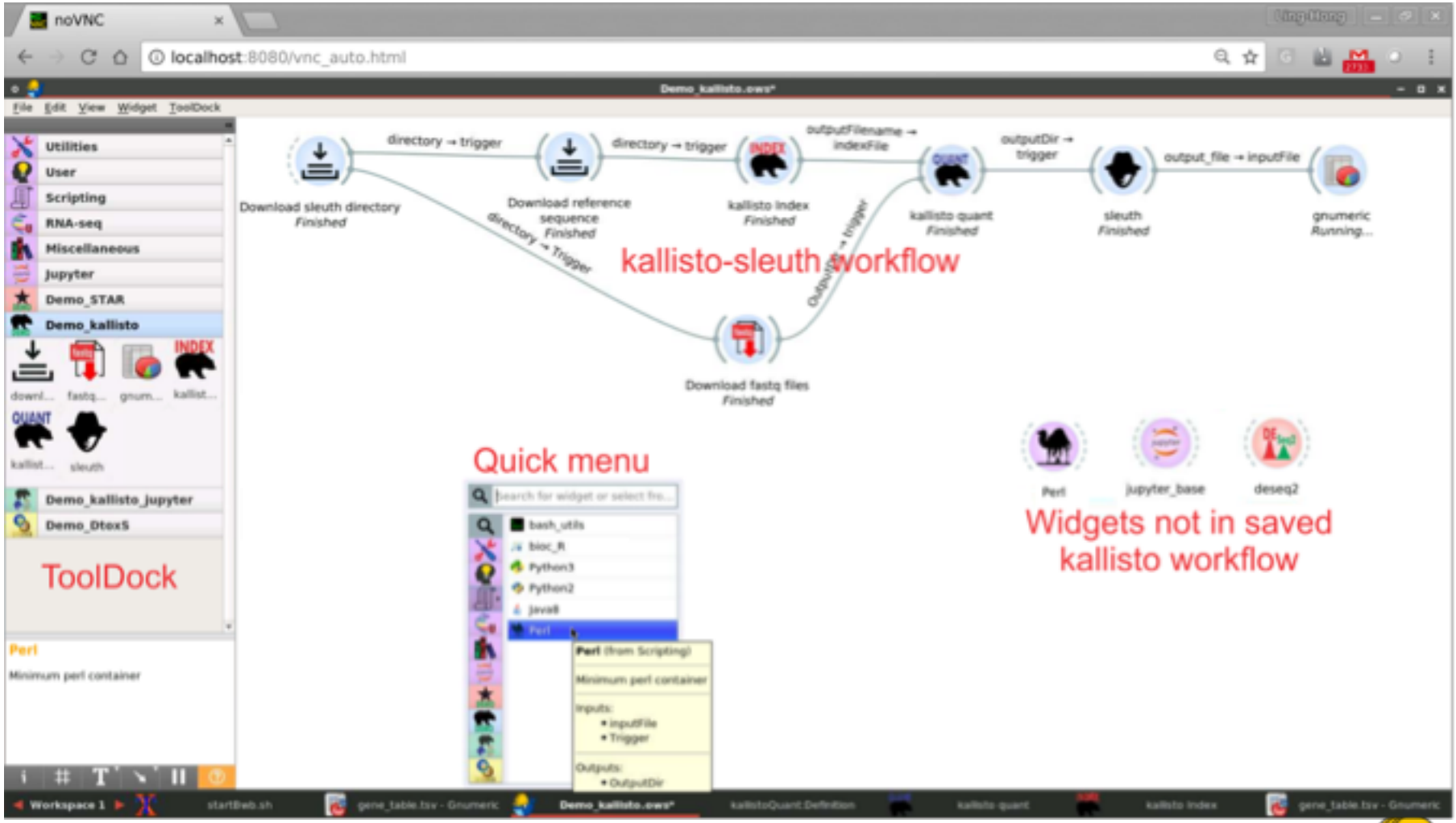
- Docker containers wrap up a piece of software in a complete file system that contains everything it needs to run: code, runtime, system tools, system libraries – anything you can install on a server.
- It will give the same results regardless of the host environment
- Docker Hub: repository of Docker containers

https://www.docker.com/what-docker

BioDepot-workflow-Builder (Bwb)

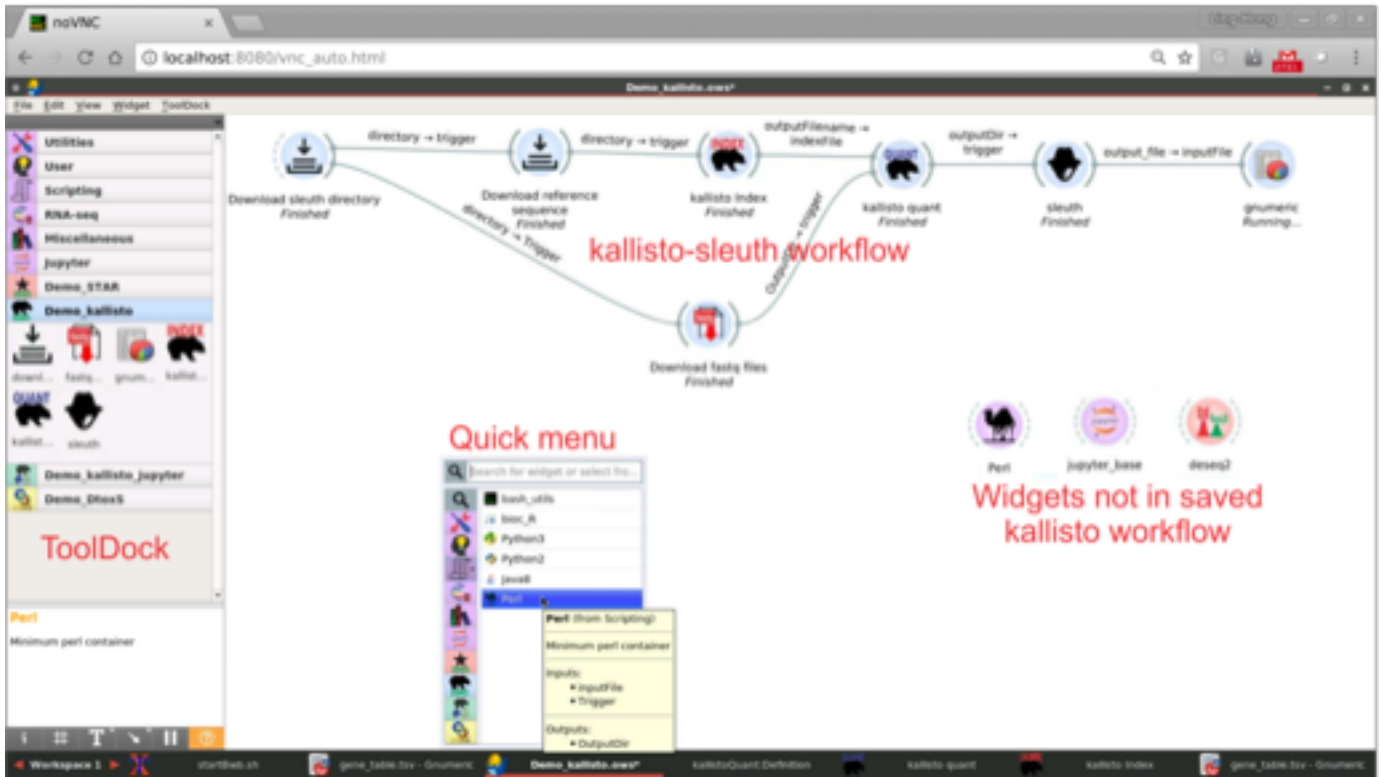
- **Drag-and-drop** form based interface
- **Modular**: Build modular bioinformatics workflows using widgets. Mix-and-match.
- **Reproducible results:** each module is encapsulated by a software container, computing environment and parameters are maintained
- **Extensible:** can add easily new widgets using a form-based builder.
- Support graphical output in modules
- Not affected by third party software upgrades.
- Easily deployed across different platforms.

Drag-and-drop GUI



Widgets

- Represent modular units in a workflow, such as an executable, script or a basic operation.
- Each widget executes a Docker container
- Can be dragged to the canvas and connected to form workflows.



Workflows

When the pipeline is to be run, the Bwb/Orange engine follows the graph, executing each widget and then propagating the output to connected widgets which are then executed, until the entire graph is traversed.

Running Bwb

Bwb can be deployed on your local host or any cloud platform with Docker installed.

1. Download the Bwb Docker image

```
docker pull biodepot/bwb:latest
```

2. Start the Bwb container

```
docker run --rm -it -p 6080:6080 -v ${PWD}:/data \
 -v /var/run:/var/run -v /tmp/.X11-unix:/tmp/.X11-unix biodepot/bwb
```

To access the container open up a browser window and type in localhost (or the IP of the container) followed by the port number 6080 i.e. localhost:6080

Bwb: Implementation details

- Drag-and-drop interface based on OrangeML library (http://orange.biolab.si)
- OrangeML: widget is a manually written Python file
- Bwb widgets
  - stores the widget definition in 3 JSON (JavaScript Object Notation) files, 1 icon file and 1 auto-generated Python file in a directory.
  - parameters specified in widgets stored in JSON and xml files.
- Wrapped the Bwb interface in a GUIdock-VNC [Mittal et al. 2017] container.

RNA-seq case study

Differential expression of genes:

(count the number of transcripts made by genes under different conditions)

-align short sequence reads to genome sequenc to identify transcript (intensive C/C++ executables)

-statistical analyses to calculate which transcripts/genes are differentially expressed (R or Python libraries)

-visualization (R/Python/Jupyter

Example

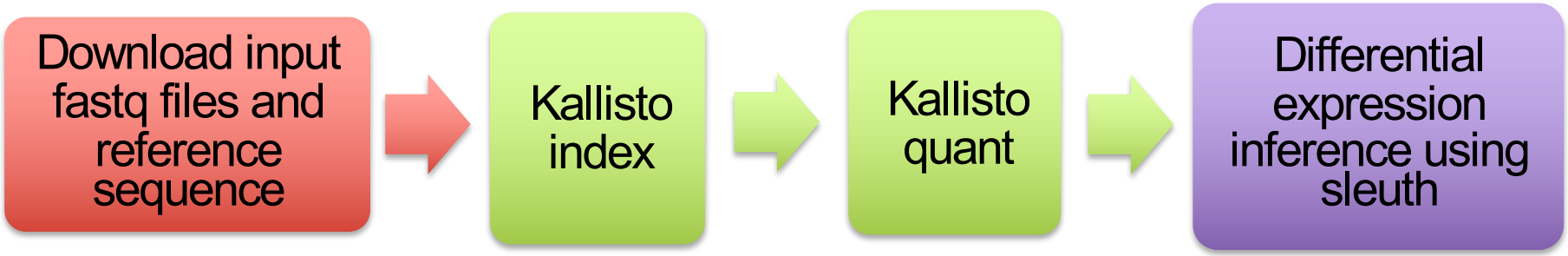
Kallisto-sleuth RNA-seq workflow

Ex: RNA-seq workflow using Kallisto and Sleuth [Pachter Lab @ Caltech]

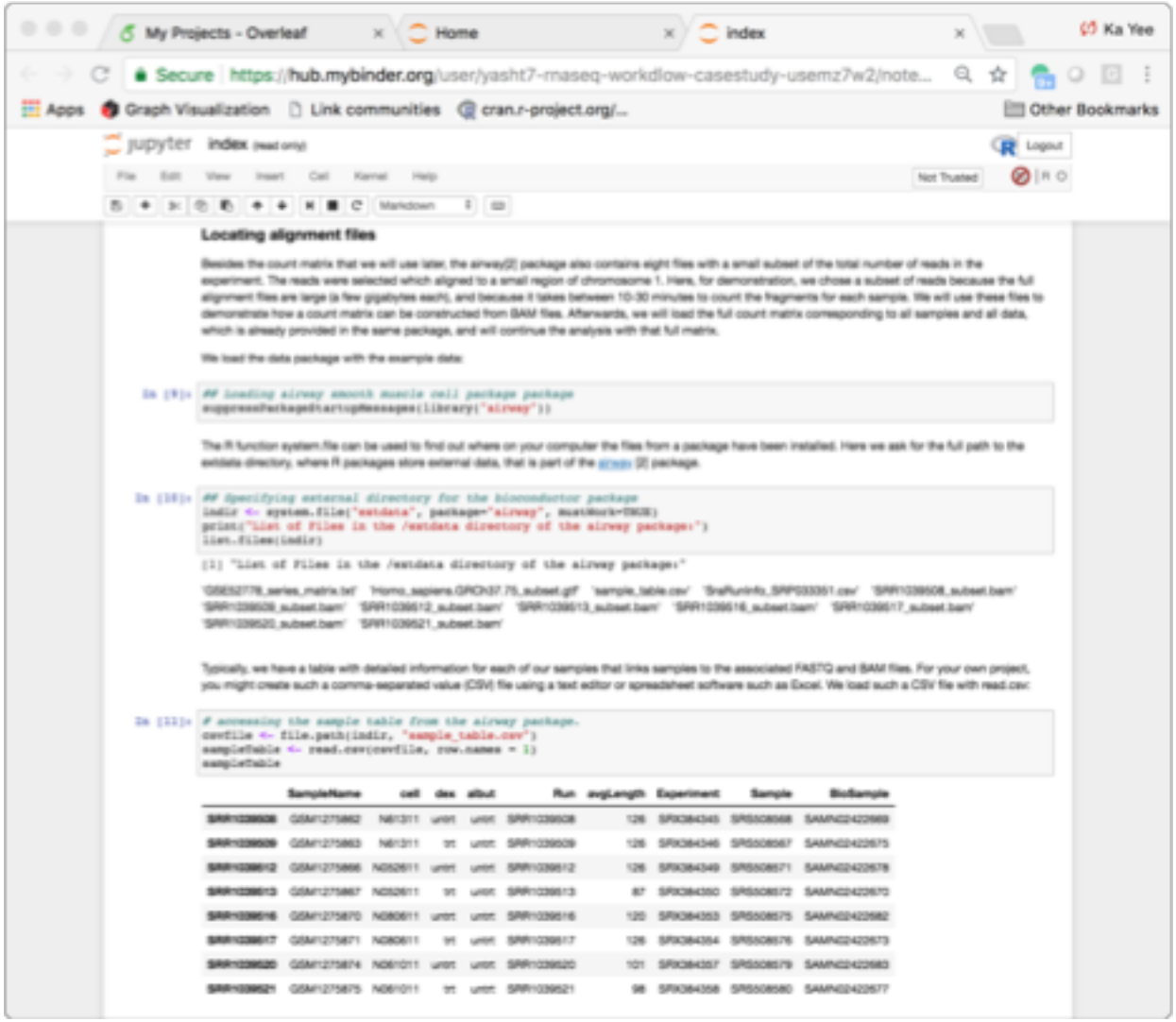
- Kallisto [Bray et al. 2016]
  - a program for fast RNA-Seq quantification based on pseudo-alignment.
  - Can quantify 30 million human reads in less than 3 minutes on a desktop computer using only the read sequences and a transcriptome index that itself takes less than 10 minutes to build.
  - Written in C++
- Sleuth [Pimentel et al. 2017]
  - Written in R. Shiny app with plots.
  - Dependences on other R packages
  - Recommended installation: biocLite or conda
  - Issues with the latest version of R

Ex: RNA-seq workflow using Kallisto and Sleuth [Pachter Lab]

- Sample data: GSE37704
- RNA-seq data in which lung fibroblasts in response to loss of the developmental transcription factor HOXA1 [Trapnell et al. 2013]
- 6 samples: 3 knockdown, 3 control



The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text.



Demo

Take home message

- Reproducible results when containers are deployed across different hardware and software configurations.
- Dockerization makes it easy to swap out a component and not worry about dependencies. Each module can be run in a different environment.
- UI to create your own containerized widgets that run your customized code.

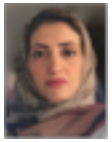
Future Work

- BwB
  - Scheduler
  - Support the Common Workflow Language and YAML
- Benchmarking bioinformatics workflows on the cloud and build predictive models
- Improve the scalability of workflows

Using Docker containers to address the reproducibility of bioinformatics analyses

GUIdock: Using Docker containers with a common graphics user interface to address the reproducibility of research. Ling-Hong Hung, Daniel Kristiyanto, Sung Bong Lee, Ka Yee Yeung. PLOS One 2016, 11(4):e0152686.

GUIdock-VNC: using a graphical desktop sharing system to provide a browser-based interface for containerized software. Varun Mittal, Ling-Hong Hung, Jayant Keswani, Daniel Kristiyanto, Sung Bong Lee, Ka Yee Yeung. Gigascience 2017, 6(4): 1-6.



Reproducible Bioconductor Workflows Using Browser-based Interactive Notebooks and Containers. Reem Almugbel, Ling-Hong Hung, Jiaming Hu, Abeer Almutairy, Nicole Ortogero, Yashaswi Tamta, Ka Yee Yeung. JAMIA 2018, 25(1):4-12.

Hot-starting software containers for bioinformatics analyses. Pai Zhang, Ling-Hong Hung, Wes Lloyd, Ka Yee Yeung. Gigascience 2018, 7(8): giy092.

Embedding containerized workflows inside data science notebooks enhances reproducibility. Jiaming Hu, Ling-Hong Hung, Ka Yee Yeung. bioRxiv 309567.

Serverless computing provides on-demand high performance computing for biomedical research. Dimitar Kumanov, Ling-Hong Hung, Wes Lloyd, Ka Yee Yeung. arXiv:1807.11659.

UW Team

Ka Yee Yeung

Wes Lloyd

Ling-Hong Hung

Reem Almugbel
Abeer Almutairy
Huazeng Deng
Saranya Ravishankar Devi
Jiaming Hu
Alyssa Ingersoll
Nicole Kauer
Jayant Keswani
Daniel Kristiyanto
Sung Bong Lee
Xiao Liang

Trevor Meiss
Varun Mittal
Radhika Agumbe Sridhar
Kuangdi Yu
Pai Zhang

Icahn School of Medicine at Mount Sinai

Eric Sobie
Yuguang Xiong

Thank you

For more information, visit
https://github.com/BioDepot/

NIH grant R01GM126019

NIH National Institute of General Medical Sciences

amazon web services

Google Cloud