

# Dockstore and FAIR

Denis Yuen, Ontario Institute for Cancer Research  
Beth Sheets, UC Santa Cruz Genomics Institute

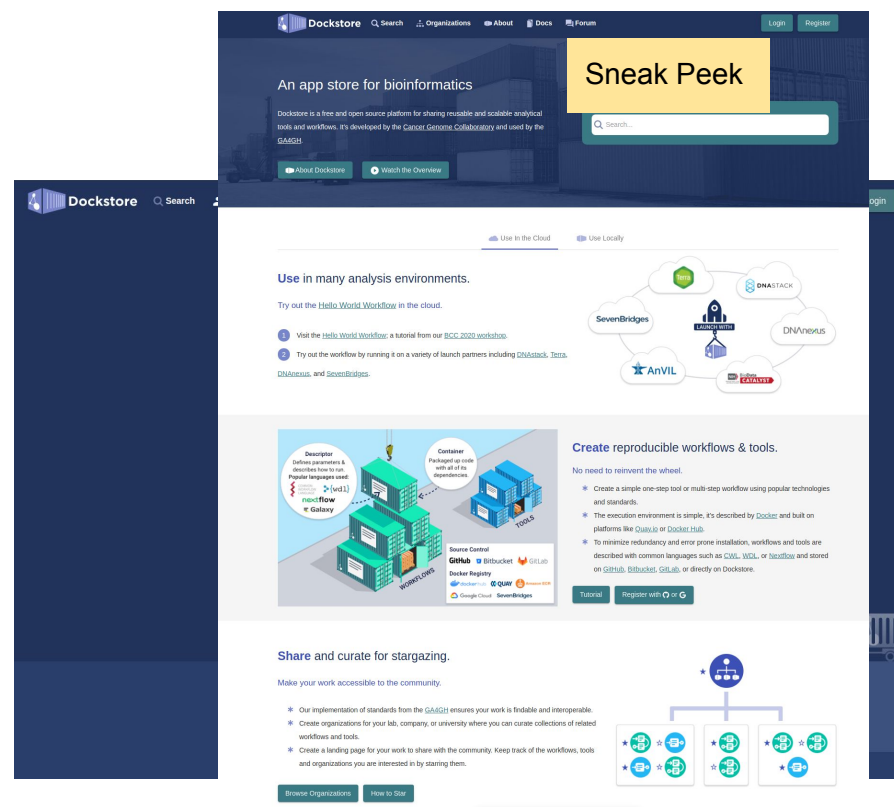
# Intro

- What is Dockstore and how did it come about?
- Dockstore and FAIR
- Best Practices Document
- Prominent FAIR Work on Dockstore
- Future Work

# What is Dockstore?

Dockstore is a free and open source platform for sharing scientific tools and workflows. It is a registry of Docker-based resources described using popular workflow languages in bioinformatics CWL, WDL, Nextflow, and Galaxy

- **Portability**
  - Run workflows in any environment that supports Docker
- **Interoperability**
  - Standardize computational analysis through GA4GH APIs
- **Reproducibility**
  - Create, Share, Use
  - Containers + Popular descriptor languages
- These aren't exactly FAIR



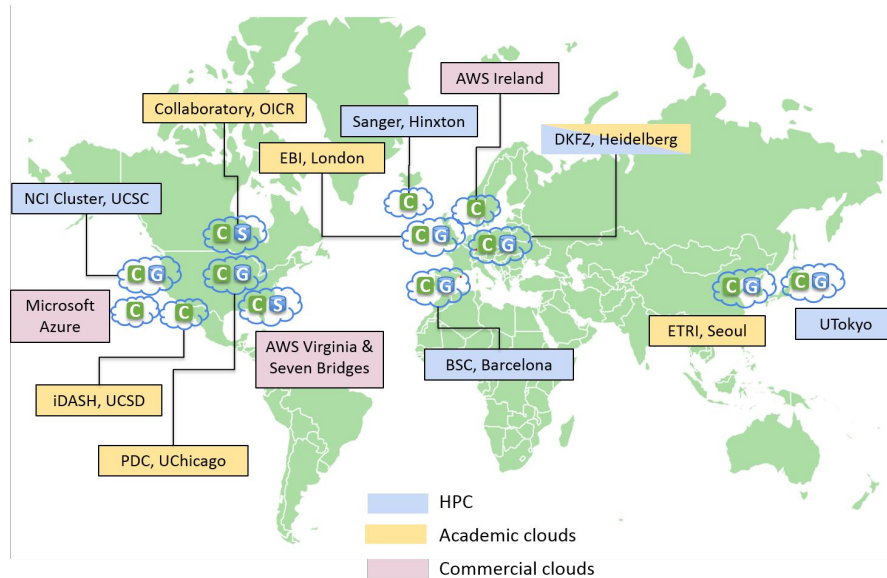
Now on version 1.10.2, first presented version 1.2.5 at BOSC2017  
\*\*note: all slides will be shared

# Motivations for Dockstore:

## The Pan Cancer Analysis of Whole Genomes (PCAWG)

### Cloud-Based, Distributed Collaboration

- <https://dcc.icgc.org/pcawg/>
- **~5,800 Whole Genomes**
  - **~2,800 Cancer Donors**
- **8 sites storing and sharing data via GNOS**
  - **300TB -> 900TB**
- **14 Cloud (and HPC) environments**
  - **3 Commercial, 7 OpenStack, 4 HPC**
  - **~630 VMs, ~15K cores, ~60TB of RAM**
- **2016 Dockstore** founded to standardize workflows across these environments



PanCancer Analysis  
OF WHOLE GENOMES



International  
Cancer Genome  
Consortium



Ontario Institute  
for Cancer Research

# What is Dockstore?

[dockstore / dockstore](#)

Code Issues 438 Pull requests 4 Actions Wiki Security Insights

Go to file Create

develop 27 branches 251 tags

**kathy-t** Record how Docker images used in workflows are specified (#4267) 5 days ago 2,813 commits

.circled	Upload image digest to S3 on quay build (#4274)	5 days ago
github	deterministic sorting of openapi.yaml? (#4136)	3 months ago
.maven.wrapper	Adding wrapper (#4046)	4 months ago
THIRD-PARTY-NOTICES	Standardize model boilerplate (#1356)	3 years ago
bom-internal	Upgrade to ES 7.10 (#4219)	last month
dockstore-common	Record how Docker images used in workflows are specified (#4267)	5 days ago
dockstore-event-consumer	Bump generated versions (#4270)	6 days ago
dockstore-integration-testing	Record how Docker images used in workflows are specified (#4267)	5 days ago
dockstore-language-plugin-parent	Bump generated versions (#4270)	6 days ago
dockstore-webservice	Record how Docker images used in workflows are specified (#4267)	5 days ago
git-hooks	Set up git-secrets/maven integration (#3788)	9 months ago
openapi-java-client	Bump generated versions (#4270)	6 days ago
reports	Bump generated versions (#4270)	6 days ago
scripts	don't append to dockstoreTest.yml multiple times (#4240)	28 days ago
swagger-java-bitbucket-client	Bump generated versions (#4270)	6 days ago
swagger-java-client	Bump generated versions (#4270)	6 days ago
swagger-java-discourse-client	Bump generated versions (#4270)	6 days ago
swagger-java-quay-client	Bump generated versions (#4270)	6 days ago
swagger-java-sam-client	Bump generated versions (#4270)	6 days ago
swagger-java-zenodo-client	Bump generated versions (#4270)	6 days ago
.codecov.yml	Stop failing builds for minor decreases in coverage (#1522)	3 years ago
.dockerignore	Push Docker Images to Quay.io	9 months ago
gitatowed	Upload image digest to S3 on quay build (#4274)	5 days ago
.gitignore	don't append to dockstoreTest.yml multiple times (#4240)	28 days ago

**About**

Our VM/Docker sharing infrastructure and management component

[dockstore.org/](#)

[docker](#) [workflow](#) [bioinformatics](#)

[nextflow](#) [containers](#) [vdl](#) [cwl](#)

[dockstore](#)

[Readme](#)

[View license](#)

**Releases** 251

1.10.2 Latest on 8 Mar

+ 250 releases

**Packages**

No packages published

**Contributors** 34

+ 23 contributors

**Languages**

Java 75.5% Web 20.1%

Nextflow 2.3%

Common Workflow Language 0.7%

Mustache 0.7% Scale 0.4%

Other 0.3%

## Dockstore API 1.6.0

[ Base URL: [dockstore.org/api/](#) ]  
[https://dockstore.org/swagger-json](#)

This describes the dockstore API, a webservice that manages pairs of Docker images and associated metadata such as CWL documents and Dockerfiles used to build those images

Terms of service  
Dockstore@ga4gh - Website  
Send email to Dockstore@ga4gh  
Apache License Version 2.0  
Dockstore documentation

### Schemes

HTTPS

Authorize

Filter by tag

**entries** Interact with entries in Dockstore regardless of whether they are containers or workflows

**PUT** `/entries/{id}/aliases` Update the aliases linked to a entry in Dockstore.

**GET** `/entries/{id}/collections` Get the collections and organizations that contain the published entry

**containers** List and register entries in the dockstore (pairs of images + metadata (CWL and Dockerfile))

**GET** `/containers/{containerId}` Retrieve a tool.

**PUT** `/containers/{containerId}` Update the tool with the given tool.

**DELETE** `/containers/{containerId}` Delete a tool.

**GET** `/containers/{containerId}/cwl` Get the primary CWL descriptor file on Github.

**GET** `/containers/{containerId}/cwl/{relative-path}` Get the corresponding CWL descriptor file on Github.

**Github:** <https://github.com/dockstore/dockstore>

**Swagger:** <https://dockstore.org/api/static/swagger-ui/index.html#>

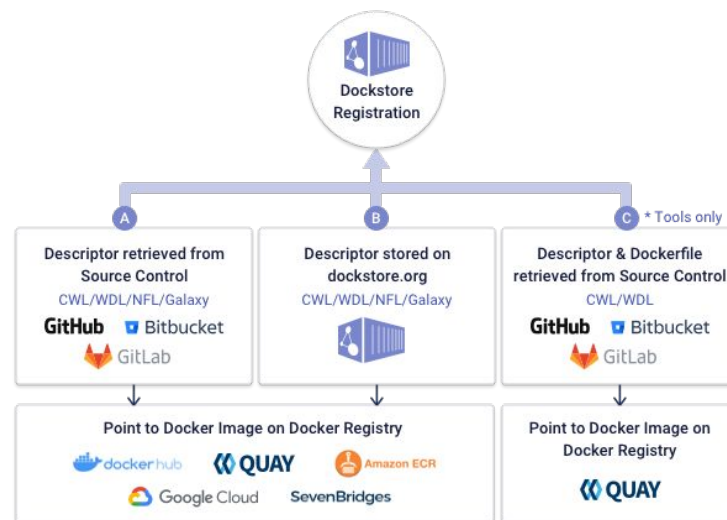
# Containers and Descriptor Languages

- Containers
  - Docker
- Descriptor [Language Support](#)
  - CWL, WDL, Nextflow, and now Galaxy!
    - Syntax Highlighting, Validation
    - Visualization
      - Including additional visualization options from [view.commonwl.org](http://view.commonwl.org) and EPAM WDL Viewer
    - Running
    - Metadata parsing
      - Authorship, contact info, description
      - I/O file types (CWL)
      - References to Docker images



# How to get workflows onto Dockstore

- Getting Started [guides](#)
  - Describe Docker, CWL, WDL, Nextflow, Galaxy
- Current recommendation
  - Sync workflows with GitHub apps
  - [guide](#)
- Workshop tutorials
  - [posters and talks](#)
  - [youtube](#)



# Running workflows from Dockstore

- Use in the Cloud
  - Launch with [partners](#)
  - Click-through to launch partners like how you click through to hotels on Google Maps, TripAdvisor
    - [Demo link](#)
- Use Locally
  - CLI helps walkthrough how to run CWL and WDL workflows locally for development purposes





# Features that help with FAIR



## Findable

When publishing on Dockstore, we strongly suggest including robust metadata and human readable instructions. Dockstore parses metadata and allows it to be searchable, which helps others find workflows and tools easily. DOIs can be generated for citations in publications.



## Accessible

Dockstore never requires a user to log in to search and inspect contents for workflows and tools. Links to source repositories are provided. Once found, analyses can be moved from environment-to-environment (e.g. [Terra](#), [DNAnystack](#), and [DNAnexus](#)) and yet be guaranteed to run on anything that supports Docker.



## Interoperable

Docker repositories (e.g. [Docker Hub](#)) and source control repositories (e.g. [GitHub](#)) provide much of the needed infrastructure. Contributors are encouraged to provide clear instructions, test data and checker workflows to ensure their software is usable in any environment. Standardized APIs enable the simple launching of workflows to a variety of compute platforms.



## Reusable

One of the ultimate goals of Dockstore is to provide workflows and tools that others can find, reuse and build upon. To ensure this, contributors are encouraged to provide clear documentation and the exact version of the container in their descriptor files.

- Caveat: Dockstore was developed by software developers to help software developers run workflows in a reproducible and portable way, nonetheless we've accumulated a few features over the years that we believe help with FAIR and welcome suggestions for new ones in our [GitHub issues](#) or [forum](#) or next weeks panel discussion
- Also examining some of this from the perspective of FAIR4S <https://www.force11.org/group/fair-4-research-software-fair4rs-working-group>

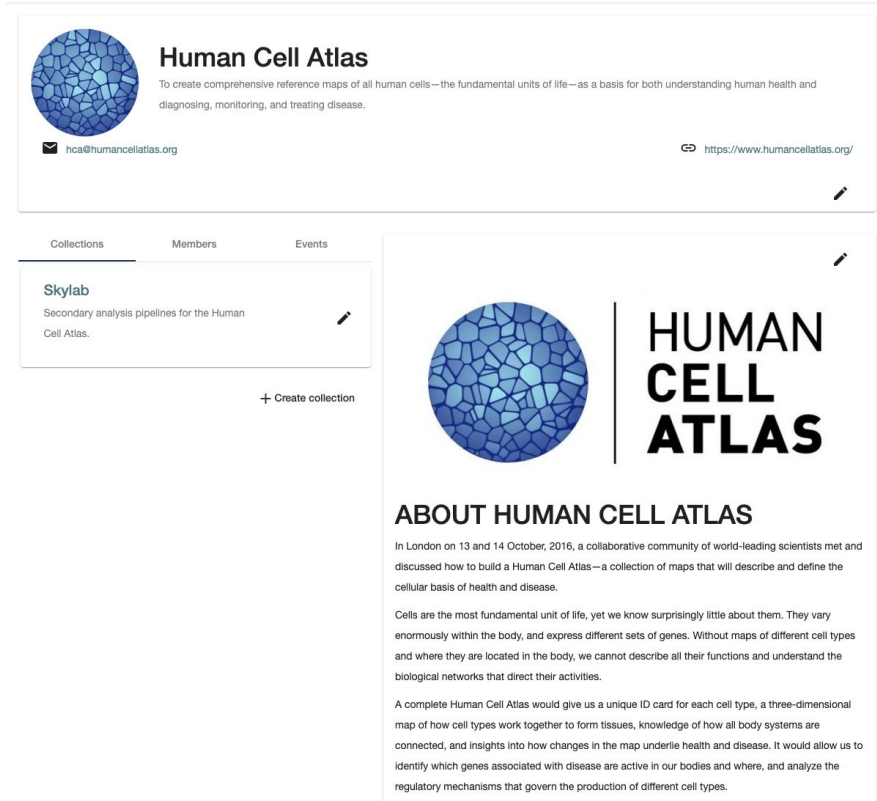
# Features that help with FAIR: Findable



- ID structure (also different versions) based on source control workspace
  - For example:  
<https://dockstore.org/workflows/github.com/gatk-workflows/gatk4-rnaseq-germline-snp-indels/gatk4-rnaseq-germline-snp-indels:1.1.1>
- Faceted Search
  - Metadata harvested from descriptors
  - Free text search
  - UI configurable labels
- Organizations and Collections
  - Similar to channels and playlists
- DOI export

# Organizations and Collections

- Organizations
  - A place for groups, labs, consortiums, etc to showcase their projects, collaborate, and group sets of tools/workflows into 'collections'
  - Markdown descriptions
  - Membership roles
  - [ORCID links](#)
- Collections
  - Playlist of workflows or tools highlighted by an Organization
  - Markdown descriptions



The screenshot displays the Human Cell Atlas website. The top header features the Human Cell Atlas logo (a blue mosaic circle) and the text "Human Cell Atlas" with a subtitle: "To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease." Below this is the email "hca@humancellatlas.org" and the URL "https://www.humancellatlas.org/".

The main content area has three tabs: "Collections", "Members", and "Events". The "Collections" tab is active, showing a collection named "Skylab" with the description "Secondary analysis pipelines for the Human Cell Atlas." and a "+ Create collection" button.

To the right, there is a large graphic with the Human Cell Atlas logo and the text "HUMAN CELL ATLAS". Below this is a section titled "ABOUT HUMAN CELL ATLAS" with two paragraphs of text.

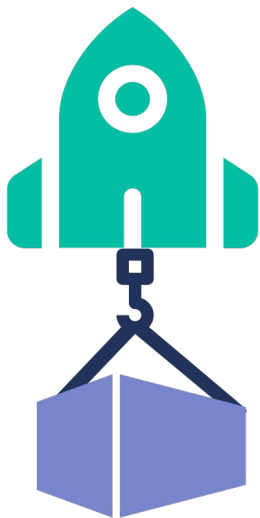
**ABOUT HUMAN CELL ATLAS**

In London on 13 and 14 October, 2016, a collaborative community of world-leading scientists met and discussed how to build a Human Cell Atlas—a collection of maps that will describe and define the cellular basis of health and disease.

Cells are the most fundamental unit of life, yet we know surprisingly little about them. They vary enormously within the body, and express different sets of genes. Without maps of different cell types and where they are located in the body, we cannot describe all their functions and understand the biological networks that direct their activities.

A complete Human Cell Atlas would give us a unique ID card for each cell type, a three-dimensional map of how cell types work together to form tissues, knowledge of how all body systems are connected, and insights into how changes in the map underlie health and disease. It would allow us to identify which genes associated with disease are active in our bodies and where, and analyze the regulatory mechanisms that govern the production of different cell types.

# Features that help with FAIR: Accessible

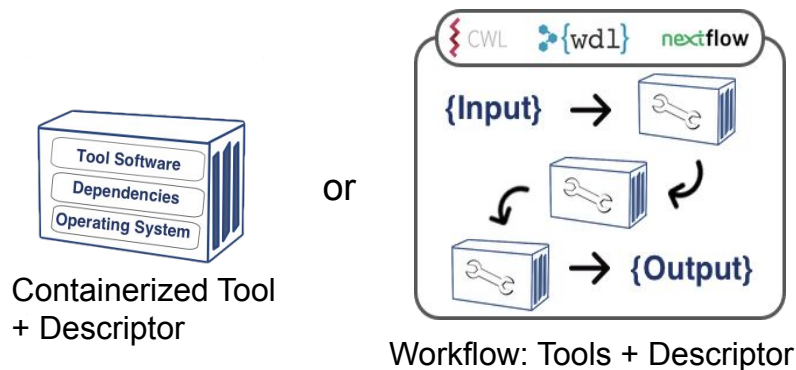


- Standardized protocol: TRS implementation
  - V1, v2 beta, and v2 final
- Published workflows accessible without login
- Links back to source control
  - Dockstore stores a version of the workflow that can be shared with others
  - Snapshot feature
    - Integration with Zenodo to mint DOIs

# GA4GH Tool Registry Service (TRS) API

Knowing we didn't have all the answers, Dockstore helped establish the Global Alliance for Genomics and Health (GA4GH) Tool Registry API standard for listing and describing available tools for exchange, indexing, and searching

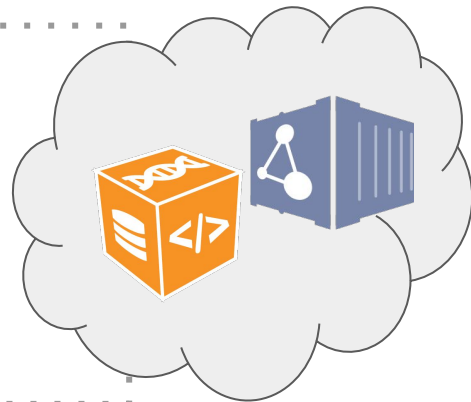
- Stand alone containerized tools
- Workflows with multiple tools wrapped in descriptor languages (Common Workflow Language, Workflow Descriptor Language, Nextflow, Galaxy)



## Sharing API

GET `/api/ga4gh/v2/tools  
/{id}/versions/{version_id}`

GET `/api/ga4gh/v2/tools`



- **GitHub page:** <https://github.com/ga4gh/tool-registry-service-schemas>
- **Latest release:** 2.0.0, working on 2.0.1 now

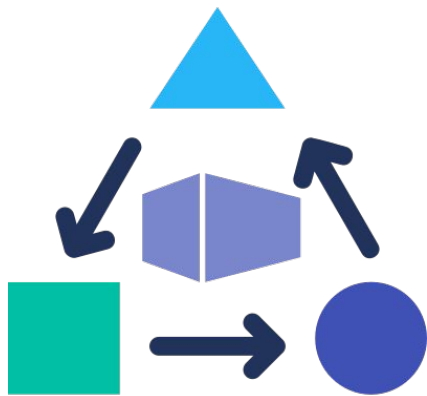


**Global Alliance**  
for Genomics & Health

# Dockstore - relationship with TRS

- GA4GH - TRS API
  - Standardized subset of Dockstore and Biocontainers
  - We provide a [validator](#) to help those implementing TRS
- Dockstore API
  - Alongside TRS with functionality specific to Dockstore
    - Authenticated access
    - Advanced language support
    - User operations
    - Community and collaboration oriented features
      - Organizations, starring, platform verification

# Features that help with FAIR: Interoperable



- JSON parameter files
  - Sample input parameters
  - Test data (preferably open access)
- Checker workflows
  - Test whether a workflow is portable
  - Used as part of a [GA4GH-DREAM](#) challenge
- Coordinate with efforts via Cloud Workstream and with FASP

# Features that help with FAIR: Reusable



New doc: [Best Practices for Secure and FAIR Workflows](#)

- Covers:
  - Referencing immutable containers
  - Suggestions to include in README
  - Licensing
  - How to set up your workflow for citation
- Developed in partnership with [NHLBI BioData Catalyst](#)



# Prominent FAIR Work on Dockstore

- Broad Institute's [Viral Genomics Collection](#)
  - Provide tutorials in Terra workspaces (execution environment) that outline in detail the steps to set up and execute their collection of workflows, including configured jsons.
  - Published [DOIs](#) of their workflows in the methods of their [Science paper](#)
- nf-core [organization](#)
  - Includes [guidelines for workflow developers](#) with minimum requirements and templates
- NHLBI BioData Catalyst [GWAS using GENESIS](#) collection
  - Paired Jupyter notebook for generating inputs for workflows to run in Terra
  - Tutorials in Terra
  - Cloud cost examples shared in README

# Further Reading

- NAR paper at <https://doi.org/10.1093/nar/gkab346>
- <https://docs.dockstore.org/> with tutorials
- And keep an eye out for 1.11!
  - Start of a more consistent and user-friendly UI
  - Galaxy launch-with buttons
  - Control publishing and control branches with github apps (.dockstore.yml)
  - Behind the scenes: if things like FedRAMP or CIS AWS standards mean something to you
  - And more!

# The Dockstore Team



Lincoln Stein

Denis Yuen

Gary Luu

Gregory Hogue

Kathy Tran



UNIVERSITY OF CALIFORNIA

SANTA CRUZ

Genomics  
Institute

Benedict Paten

Brian O'Connor

Charles Overbeck

Beth Sheets

Walt Shands

David Steinberg

Natalie Perez

Ben Vizzier

Richard Hansen

Charles Reid

Ash O'Farrell

Avani Khadilkar

# Acknowledgements



This work was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-168).



UNIVERSITY OF CALIFORNIA

SANTA CRUZ

Genomics  
Institute

Funded by:



National Institutes of Health



BISTI

Biomedical Information  
Science and Technology Initiative



National Heart, Lung,  
and Blood Institute



NHGRI

# Extra Slides

Handy stuff in case of questions