Dockstore:
Enhancing a community platform for sharing cloud-agnostic research tools

Louise Cabansay, UC Santa Cruz Genomics Institute
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 CWL, WDL, and Nextflow.

- **Portability**
  - Run workflows in any environment that supports Docker

- **Interoperability**
  - Standardize computational analysis through GA4GH APIs

- **Reproducibility**
  - Create, Share, Use
  - Containers + Popular descriptor languages

Now on version 1.6.0, first presented version 1.25 at BOSC2017
What is Dockstore?

Github: https://github.com/dockstore/dockstore

Swagger: https://dockstore.org/api/static/swagger-ui/index.html#
Motivations for Dockstore:
The Pan Cancer Analysis of Whole Genomes (PCAWG)
Cloud-Based, Distributed Collaboration

- International Cancer Genome Consortium (ICGC)
- ~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
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)

**Sharing API**

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

**GitHub page:** [https://github.com/ga4gh/tool-registry-service-schemas](https://github.com/ga4gh/tool-registry-service-schemas)
**Latest draft release:** 1.0.0, working on 2.0.0 now
GA4GH Cloud Work Stream APIs

- Tool Registry Service (TRS)
- Workflow Execution Service (WES)
- Task Execution Service (TES)
- Data Object Service (DOS) (now the Data Repository Service, DRS)

Sharing Tools and Workflows
Executing Workflows
Executing Tasks
Accessing Data
Dockstore & GA4GH Cloud Workstream
Dockstore, going beyond TRS

- GA4GH - TRS API
  - Standardized subset of Dockstore and Biocontainers
  - We provide a [validator](#) to help those implementing TRS

- Dockstore API
  - Extends TRS with functionality specific to Dockstore
    - Catalog that brings together multiple external repositories
    - Authenticated access
    - Advanced language support
    - Launch analysis
    - Community and collaboration oriented features
      - Organizations, starring, platform verification
Containers, Descriptors, and Parameter Files

**Tool:** Container + Descriptor

**Workflow:** Tool(s) + Descriptor

**Descriptor:**

The workflow language used to describe how to run your tool or workflow/pipeline.

- Which tools/containers
- What steps and when
- Where I/O parameters
- How to allocate compute requirements
- Can also include metadata information

**Parameter File (wdl, cwl):**

- Specifies the actual input/output files and where to find them (local, ftp, http, or cloud)
- JSON, YAML

```json
{
    "bam_input": {
        "class": "File",
        "format": "http://edamontology.org/format_2572",
        "path": "/tmp/NA12878.chrom20.ILLUMINA.bam"
    },
    "bamstats_report": {
        "class": "File",
        "path": "/tmp/bamstats_report.zip"
    }
}
```
Dockstore as a Registry

- Searchable and Centralized Catalogue
- Many ways to get your tool or workflow into Dockstore!
  - External Hosting: Github, Bitbucket, Quay.io, Docker Hub
  - Direct Hosting (currently just workflows)
Ways to Register to Dockstore

Through linked account via supported public APIs

Register Workflow

1. Workflow storage type
2. Create a workflow

You can write your CWL/WDL and store it on Dockstore.org, or you can register a CWL/WDL/NextFlow from an existing third party repository, such as GitHub, Bitbucket, etc.

- Use CWL, WDL or Nextflow from GitHub, Bitbucket, etc.
- Create and save CWL, WDL, or Nextflow on Dockstore.org

OR through manual registration pointing to external source repo or directly on Dockstore.org
Ways to Register to Dockstore

(A) Automated builds using quay.io and the source control repos triggering them

(B) Retrieve workflow and tool descriptors from external source control directly

(B/C) point to docker image(s) on quay or dockerhub

(C) Store workflow descriptors on dockstore.org
Containers and Descriptor Languages

- Containers
  - Docker

- Descriptor Language Support
  - CWL, WDL, and now Nextflow!
    - Syntax Highlighting, Validation
    - Visualization
      - Including additional visualization options from view.commonwl.org and EPAM WDL Viewer
    - Metadata parsing
      - Authorship, contact info, description
      - Markdown syntax (CWL, WDL)
      - I/O Filetypes (CWL)
Visualizations

Built-in for CWL, WDL, and Nextflow (pictured)

[view.commonwl.org](view.commonwl.org) for CWL

EPAM Pipeline Builder for WDL
Launching Analysis

- GA4GH TRS-WES Orchestrator
  - FireCloud
  - Terra
- Partner Platforms
  - DNAstack
  - DNAnexus
  - Terra
  - Local Execution via Dockstore CLI
Dockstore CLI

A handy command line resource for developers to play with tools/workflows locally before running them on a cloud

- DRS file provisioning
- WES beta support
- Locally
  - Cromwell
  - Cwltool

Execution with the Dockstore Command Line Interface (CLI)

```
provision input files → stage CWL/WDL → execute with Cromwell /cwltool → provision output files
```

Simple Dockstore Command Line
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

- **Collections**
  - Playlist of workflows or tools highlighted by an Organization
  - Markdown descriptions
Usability Improvements

- **Improved Published Tool/Workflow Interface**
  - Labels - improve searchability
  - Metadata and markdown display
    - Authorship, contact info, descriptions
    - Readme-like descriptions (CWL, WDL)
  - Versioning: display all tags/versions, set defaults, check validity, verify compatible platforms
  - Link to checker workflow

---

Workflows: Check the workflow for any updates in the specific tool and see if there are any problems. The workflow is part of the PanCancer Analysis of Whole Genomes (PCAWG) project. It is a combination of the IGCC PCAWG Sequre DeNovo/Workflow (EMBL) workflow for structural variation and the IGCC PCAWG DKFZ Workflow workflow for SNVs, indels, and copy number variants.

For more information, see the PCAWG project page and the GitHub page for the code.

A list of all inputs and parameters with corresponding descriptions can be found at the bottom of the page.

**Common Use Cases**

This tool is used to call variants from Tumor/Normal pair of BAM files, previously generated by using the IGCC-PCAWG-Sequre-BWA-Workflow tool available in Public Apps. The tool is used in the same manner as the VCF files in the IGCC PCAWG dataset. To do this, set:

- Tumor/Normal pair in BAM format, provided via Input BAM Files port.
Usability Improvements

- **Search Interface**
  - Search name, author, organization, descriptor, etc
  - Filter results using facets
    - Descriptor Language
    - Labels
Usability Improvements

- Lots of new documentation and tutorials!
- Example Topics:
  - Getting Started With…(Docker, Nextflow, CWL, WDL)
  - Launching Tools and Workflows
  - Writing checker workflows
  - Developing File Provisioning Plugins
  - Creating Organizations
  - And many more!

https://docs.dockstore.org/
Highlighted Community Content

GA4GH-DREAM Workflow Testbed (2018)

via James Eddy
Dockstore will serve as an official workflow provider for the Commons Alliance, a common infrastructure collaboration to facilitate combining *Data + Cloud Compute*
Next Release - 1.7.0 (Late August/September)

- Prototype for services/apps
  - Dockerized genome browsers, jupyter notebooks, and reference data providers
  - Playground for Github Apps and automatic update syncing
- Immutable workflow versions
  - DOI issuing via linked Zenodo credentials
- Language support updates
  - WDL draft-3 to 1.0
  - cwltool update (possibly CWL 1.1 support)
- Display of logs from when we test and verify workflows
- Improved security of your data and logging, monitoring of the site
- Summary of changes at https://docs.dockstore.org/changelog/
Future Work

- Signing of Entries on Dockstore
  - Verify ownership and integrity of Docker images
- Alternative Containerization Support
  - Singularity and/or uDocker support
- Additional Workflow Languages
- Services/“Apps”
- Testbed Expansion
  - Automated testing for workflows across cloud platforms
- Open to collaborations with partners for any of these kind of projects!
The Dockstore Team

Lincoln Stein
Denis Yuen
Andrew Duncan
Gary Luu

Brian O’Connor
Benedict Paten
Cricket Sloan
Charles Overbeck
Walt Shands
David Steinberg

Louise Cabansay
Natalie Alvarez
Ryan Bautista
Melaina Galaspie
Shreya Radesh
Emily Soth
Acknowledgements

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

Handy stuff in case of questions
## Language Support

More info: [https://docs.dockstore.org/docs/user-tutorials/language-support/](https://docs.dockstore.org/docs/user-tutorials/language-support/)

<table>
<thead>
<tr>
<th>Feature</th>
<th>CWL</th>
<th>WDL</th>
<th>Nextflow</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dockstore site</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tool registration</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Workflow registration</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Hosted Workflows</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DAG Display</td>
<td>Yes (cwl version&gt;=1.0) [1]</td>
<td>Yes (wdl version&lt;=draft-2) [3]</td>
<td>Limited support</td>
</tr>
<tr>
<td>Tool Tab Display</td>
<td>Yes (cwl version&gt;=1.0)</td>
<td>Yes (wdl version&lt;=draft-2)</td>
<td>Yes</td>
</tr>
<tr>
<td>Launch-with Platforms</td>
<td>Not yet!</td>
<td>FireCloud (workflows only) [2]</td>
<td>Not yet!</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DNAnexus (workflows only) [3]</td>
<td></td>
</tr>
<tr>
<td>Metadata Display</td>
<td>Yes</td>
<td>Yes (wdl version&lt;=draft-2)</td>
<td>Yes</td>
</tr>
</tbody>
</table>

| **Dockstore CLI**        |                      |                      |                       |
| Local workflow engines   | owitool, Cromwell    | Cromwell [4]         | Nextflow              |
| File Provisioning In     | Local File System    | Local File System    | Local File System     |
|                          | HTTP                 | HTTP                 | HTTP                  |
|                          | FTP                  | FTP                  | FTP                   |
|                          | S3 via plugins       | S3 via plugins       | S3                    |
|                          | Data Object Service  | Data Object Service  |                       |
| Plugins Support          | s3                   | s3                   | No                    |
|                          | s3cmd                | s3cmd                |                       |
|                          | icgc-get             | icgc-get             |                       |
|                          | Data Object Service  | Data Object Service  |                       |
| File Provisioning Out    | Local File System    | Local File System    | Local File System     |
|                          | HTTP                 | HTTP                 | S3                    |
|                          | FTP                  | FTP                  |                       |
|                          | S3 via plugins       | S3 via plugins       |                       |
| Notifications            | Yes                  | Yes                  | No                    |
Dockstore Ecosystem

Dockstore is thankful to its many contributors, users, and partners. This community has pulled together a library of over 450 tools and workflows. In the diagram to the right we’ve highlighted a few select contributors to give a sense of what has been occurring in this space.
Importing from Dockstore

github.com/DataBiosphere/topmed-workflows/UM_variant_caller_wdl
v 1.32.0

⚠️ Please note: Dockstore cannot guarantee that the WDL and Docker image referenced by this Workflow will not change. We advise you to review the WDL before future runs.

```plaintext
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22


## This is the U of Michigan variant caller workflow WDL for the workflow
## https://github.com/statgen/topmed_freeze3_calling
##
## It uses a Docker image built with software tools that can reproduce
## variant calls compatible to TopMed Freeze 3a
##
## NOTE: This workflow assumes that input CRAM files have been built with
## human reference genome. In particular for the TopMed CRAM files the
## reference genome files to use are located here:
## ftp://share.sph.umich.edu/gotcloud/ref/hg38DH-db142-v1.tgz
##
## workflow TopMedVariantCaller {

  Boolean? calculate_DNA_contamination
  Boolean calculate_contamination = select_first([calculate_DNA_contamin
  Boolean? dynamically_calculate_file_size
  Boolean dynamically_calculate_disk_requirement = select_first([dynamic
```
Importing github.com/DataBiosphere/topmed-workflows/Functional_Equivalence:1.30.0 from Dockstore

⚠ Please note: Dockstore cannot guarantee that the WDL and Docker image referenced by this Workflow will not change. We advise you to review the WDL before future runs.

Name
Functional_Equivalence

Destination Workspace
Select a workspace

Export

---

```plaintext
# Copyright Broad Institute, 2017
#
# This WDL pipeline implements CGM Pipeline standard to process
# high-throughput sequencing data for downstream processing.
#
# Requirements/Expectations:
# 1) Human whole-genome pair-end sequencing data in unfixed BAM (SAM) format
# 2) One or more read groups, one per BAM file, all belonging to a single sample (SAM)
# 3) Input SAM files must additionally comply with the following requirements:
# 4) Read groups must use the same suffix (.bam) or be untagged
# 5) Reads must pass validation by ValidateSamFile
# 6) Reads are provided in query-sorted order
# 7) Reads must have an RG tag
# 8) Reference genome must be HG38 with ALT contigs
#
# Runtime parameters are optimized for Broad's Google Cloud Platform implementation.
# For program versions, see docker containers.
#
# LICENSE:
# This script is released under the WDL source code license (BSD-3) (see LICENSE in
# https://github.com/broadinstitute/wdl). Note however that the programs it calls may
# be subject to different licenses. Users are responsible for checking that they are
# authorized to run all programs before running this script. Please see the docker
# page at https://hub.docker.com/broadinstitute/genomes-in-the-cloud/ for detailed
# licensing information pertaining to the included programs.
#
# WORKFLOW DEFINITION
Workflow PairEndSingleSampleWorkflow {
  File wgs_evaluation_interval_list
  File wgs_coverage_interval_list
  String sample_name
  String base_file_name
  Array[File] flankcall_unmapped_bams
  ...
```