



# Genomics in the Cloud

Book Club - Week 12

February 15, 2021

# Agenda

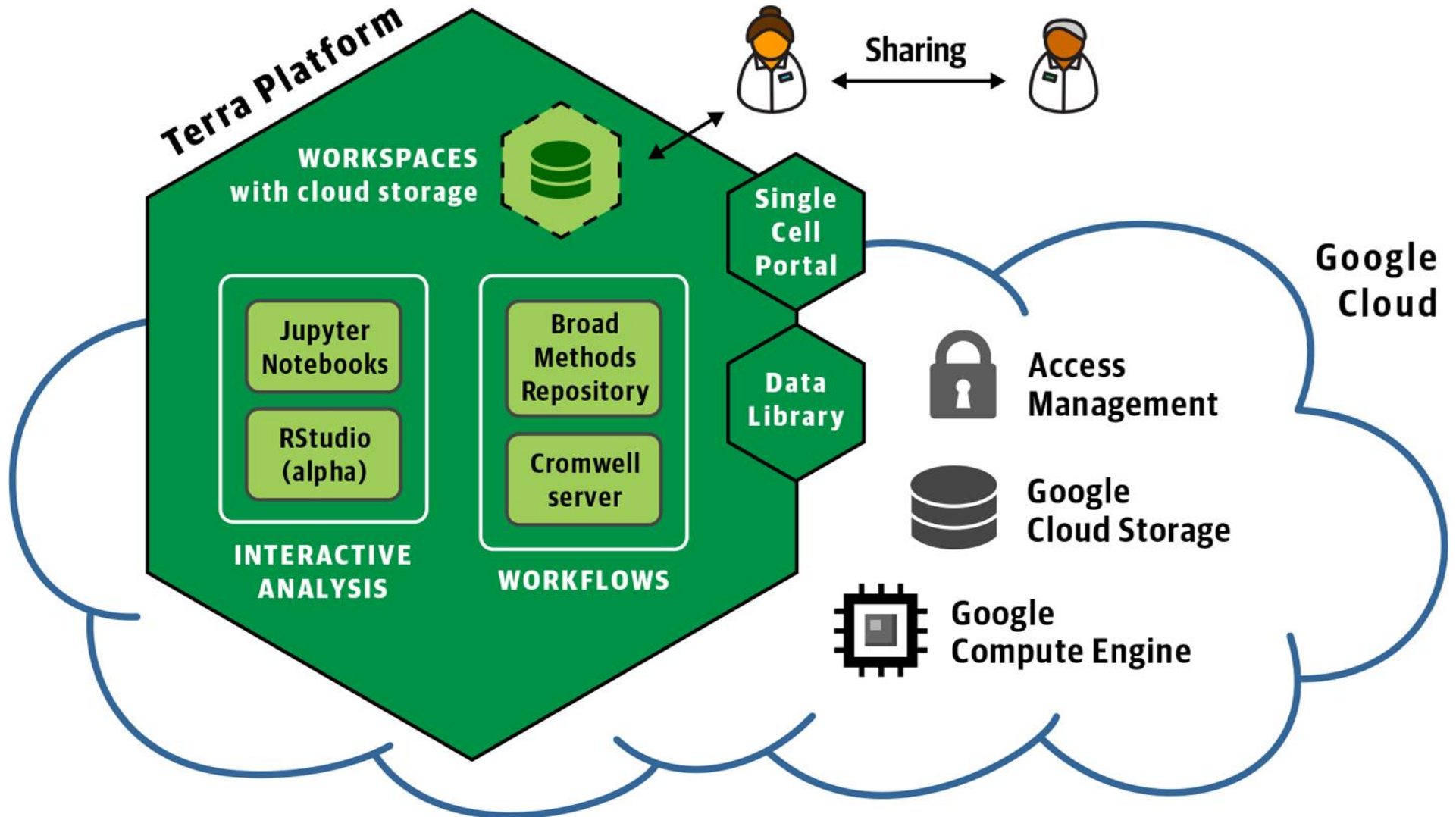
- Chapter 11: Running Many Workflows Conveniently in Terra
- Additional resources
- Open discussion



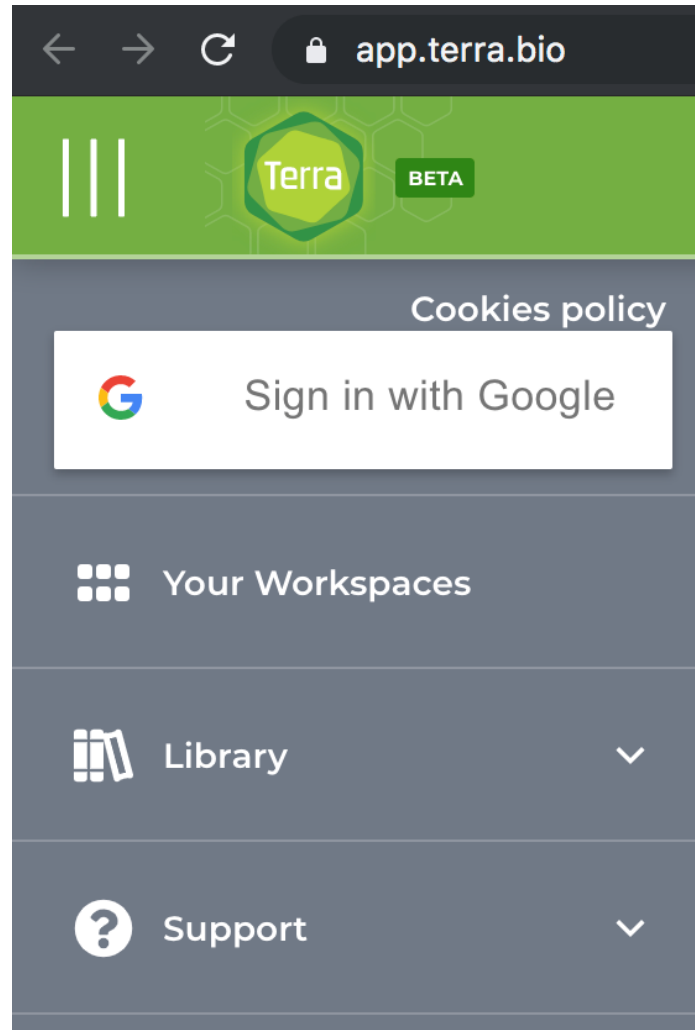
# Chapter 11: Running Many Workflows Conveniently in Terra

*Genomics in the Cloud* by Geraldine A. Van der Auwera and Brian D. O'Connor (O'Reilly). Copyright 2020 The Broad Institute, Inc. and Brian O'Connor, 978-1-491-97519-0.

# Terra platform overview



# Create an account



<https://app.terra.bio>



# Terra registration form



# TERRA

## New User Registration

First Name \*

Last Name \*

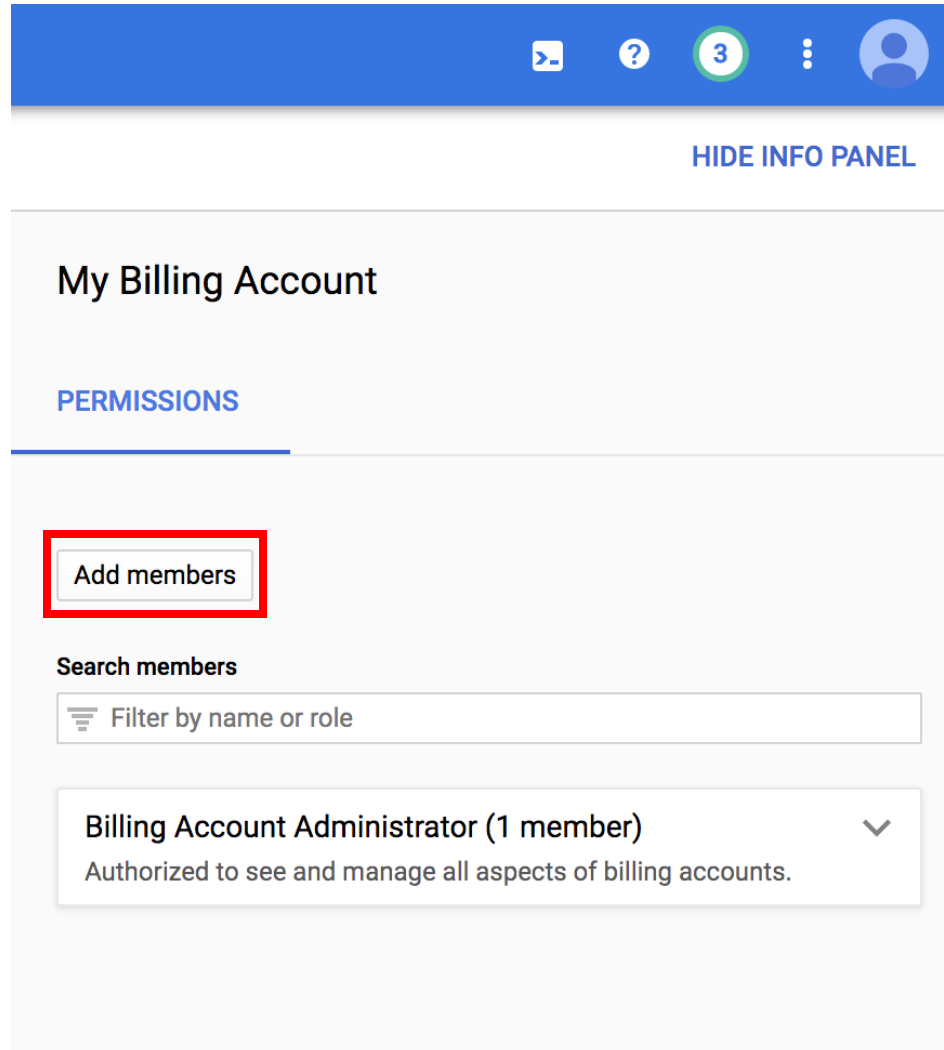
Contact Email for Notifications \*

REGISTER

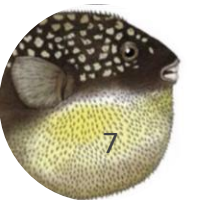
CANCEL



# Connect an existing billing account



The screenshot displays a user interface for managing a billing account. At the top, a blue navigation bar contains icons for a chat bubble, a question mark, a green circle with the number '3', a vertical ellipsis, and a user profile icon. Below this bar, a link labeled 'HIDE INFO PANEL' is visible. The main content area is titled 'My Billing Account' and features a 'PERMISSIONS' tab. A red rectangular box highlights the 'Add members' button. Below this button is a 'Search members' section with a search bar containing the placeholder text 'Filter by name or role'. At the bottom, a list item shows 'Billing Account Administrator (1 member)' with a dropdown arrow, and a description: 'Authorized to see and manage all aspects of billing accounts.'



# Add the Terra billing user

Add members to "My Billing Account"

## Add members and roles for "My Billing Account" resource

Enter one or more members below. Then select a role for these members to grant them access to your resources. Multiple roles allowed. [Learn more](#)

New members

terra-billing@terra.bio ✕



Add terra-billing@terra.bio

Role

Type to filter

Billing	Billing Account Administrator
Cloud Composer	Billing Account User
Dataflow	Billing Account Viewer
Dataproc	
Error Reporting	
Firebase	
IAM	
Logging	

MANAGE ROLES





# Add existing billing account to Terra

## Create Billing Project

Enter name \*

Name must be unique and cannot be changed.

Select billing account \*



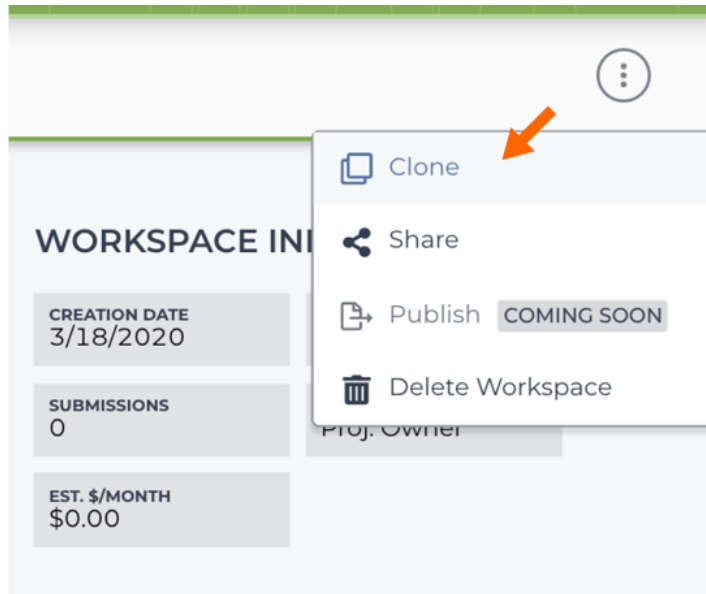
CANCEL

CREATE BILLING PROJECT



# Clone the preconfigured workspace

A.



B.

Clone a workspace

Workspace name \*

Genomics in the Cloud v1 copy

Billing project \*

Select a billing project

fccredits-cerium-white-3390

Companion workspace for Genomics in the Cloud, an O'Reilly book by Geraldine A. Van der Auwera and Brian D. O'Connor.

Read it [online in the O'Reilly Safari library]

Authorization domain ⓘ

Select groups

CANCEL CLONE WORKSPACE



# Available workflow configurations

Navigation bar: DASHBOARD DATA NOTEBOOKS **WORKFLOWS** JOB HISTORY

WORKFLOWS

SEARCH WORKFLOWS Sort By: Alphabetical

Find a Workflow

scatter-hc.data-table

V. 1  
Source: Terra

scatter-hc.filepaths

V. 1  
Source: Terra



# Workflow information summary

## ⓘ scatter-hc.filepaths

Snapshot:  ▼

Source: [genomics-in-the-cloud/scatter-hc/1](#)

Synopsis: Run GATK4 HaplotypeCaller parallelized by interval



- ✓ This workflow runs the HaplotypeCaller tool from GATK4 in GVCF mode on a single sample in BAM format. The execution of the HaplotypeCaller tool is parallelized using an intervals list file. The per-interval output GVCF files are then merged to produce a single GVCF file for the sample, which can then be used by the joint-discovery workflow according to the GATK Best Practices for germline short variant discovery.
- Run workflow with inputs defined by file paths
- Run workflow(s) with inputs defined by data table



# Workflow inputs

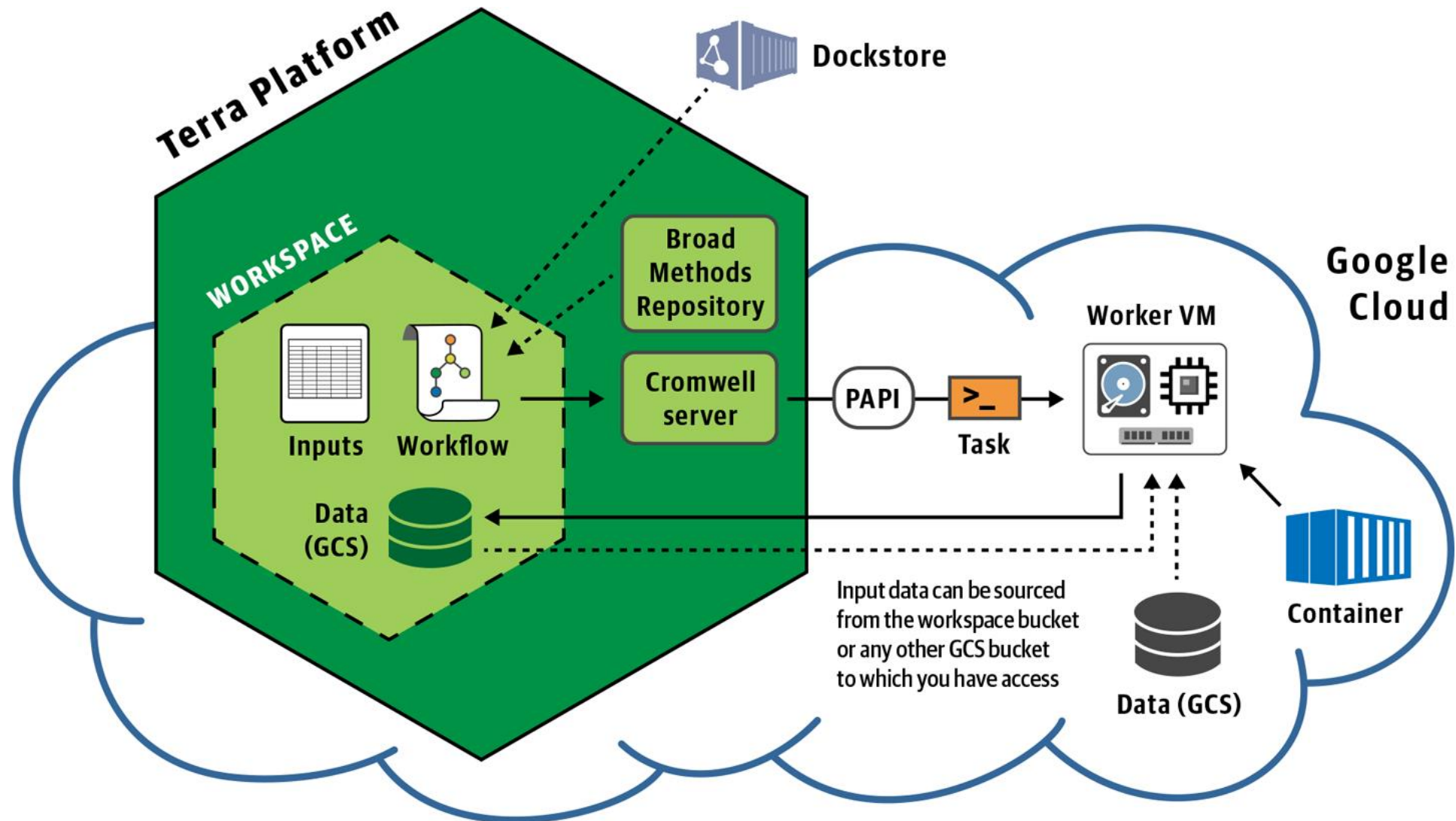
SCRIPT .. INPUTS .. OUTPUTS .. RUN ANALYSIS

Download json | Drag or click to upload json

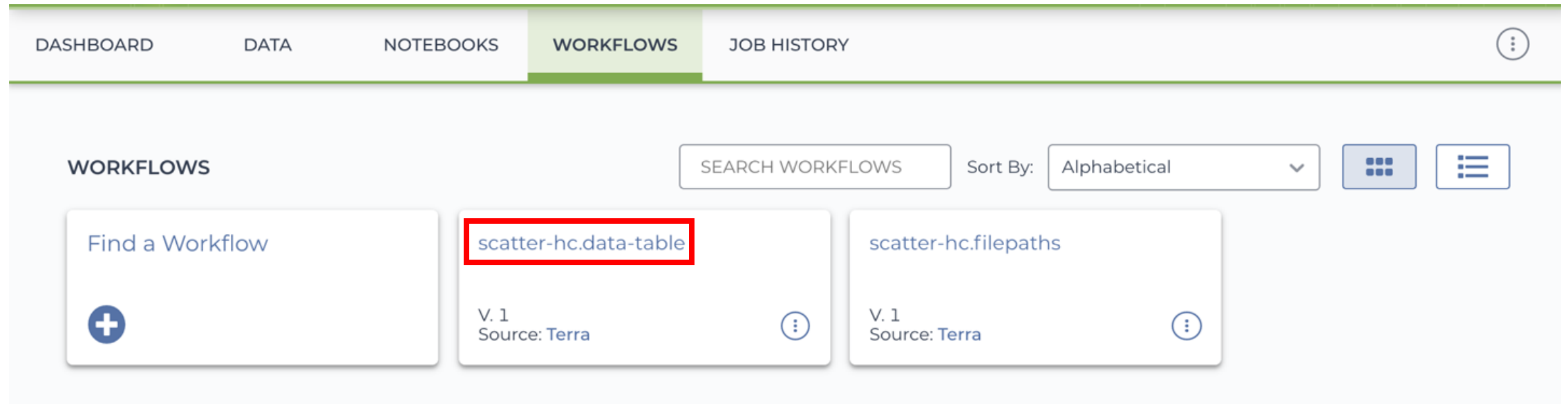
Task name	Variable	Type	Attribute
HaplotypeCallerGVCF	docker_image	String	<input type="text" value="us.gcr.io/broad-gatk/gatk:4.1.3.0"/> {...}
HaplotypeCallerGVCF	java_opt	String	<input type="text" value="-Xmx8G"/> {...}
HaplotypeCallerGVCF	ref_dict	File	<input type="text" value="gs://genomics-in-the-cloud/v1/data/germline/ref/ref.dict"/>  {...}
HaplotypeCallerGVCF	ref_fasta	File	<input type="text" value="gs://genomics-in-the-cloud/v1/data/germline/ref/ref.fasta"/>  {...}



# Workflow submission in Terra



# Using the data table



The screenshot displays the Terra Workflows interface. At the top, a navigation bar includes links for DASHBOARD, DATA, NOTEBOOKS, WORKFLOWS (which is the active tab), and JOB HISTORY. Below the navigation bar, the 'WORKFLOWS' section is visible. It features a search bar labeled 'SEARCH WORKFLOWS', a 'Sort By' dropdown menu set to 'Alphabetical', and two view toggle buttons (grid and list). The workflow list contains three items: a 'Find a Workflow' button with a plus icon, a workflow named 'scatter-hc.data-table' (highlighted with a red box), and a workflow named 'scatter-hc.filepaths'. Each workflow entry shows 'V. 1' and 'Source: Terra', along with an information icon.

DASHBOARD DATA NOTEBOOKS **WORKFLOWS** JOB HISTORY

WORKFLOWS

SEARCH WORKFLOWS Sort By: Alphabetical

Find a Workflow

**scatter-hc.data-table**

V. 1  
Source: Terra

scatter-hc.filepaths

V. 1  
Source: Terra



# Using the data table

- Run workflow with inputs defined by file paths
- Run workflow(s) with inputs defined by data table

## Step 1

Select root entity type:

## Step 2

**SELECT DATA**

all 3 book\_samples (will create a new set named "scatter-hc-data-table\_2020-03-19T05-02-55")

SCRIPT .. INPUTS .. OUTPUTS .. RUN ANALYSIS			
Download json   Drag or click to upload json SEARCH INPUTS			
Task name	Variable	Type	Attribute
HaplotypeCallerGVCF	docker_image	String	<input type="text" value="workspace.gatk_docker"/> {...}
HaplotypeCallerGVCF	java_opt	String	<input type="text" value="-Xmx8G"/> {...}
HaplotypeCallerGVCF	ref_dict	File	<input type="text" value="workspace.ref_dict"/> {...}
...			
MergeVCFs	java_opt	String	<input type="text" value="-Xmx8G"/> {...}
ScatterHaplotypeCallerGVCF	input_bam	File	<input type="text" value="this.input_bam"/> {...}
ScatterHaplotypeCallerGVCF	input_bam_index	File	<input type="text" value="this.input_bam_index"/> {...}
ScatterHaplotypeCallerGVCF	intervals_list	File	<input type="text" value="workspace.intervals_list_min"/> {...}

**workspace.gatk\_docker**

**this.input.bam**





# The book\_sample table

DOWNLOAD ALL ROWS

COPY PAGE TO CLIPBOARD

2 rows selected

<input type="checkbox"/>	book_sample_id	input_bam	input_bam_in
<input checked="" type="checkbox"/>	father	<a href="#">father.bam</a>	<a href="#">father.bai</a>
<input checked="" type="checkbox"/>	mother	<a href="#">mother.bam</a>	<a href="#">mother.bai</a>
<input type="checkbox"/>	son	<a href="#">son.bam</a>	<a href="#">son.bai</a>

Download as TSV

Open with...

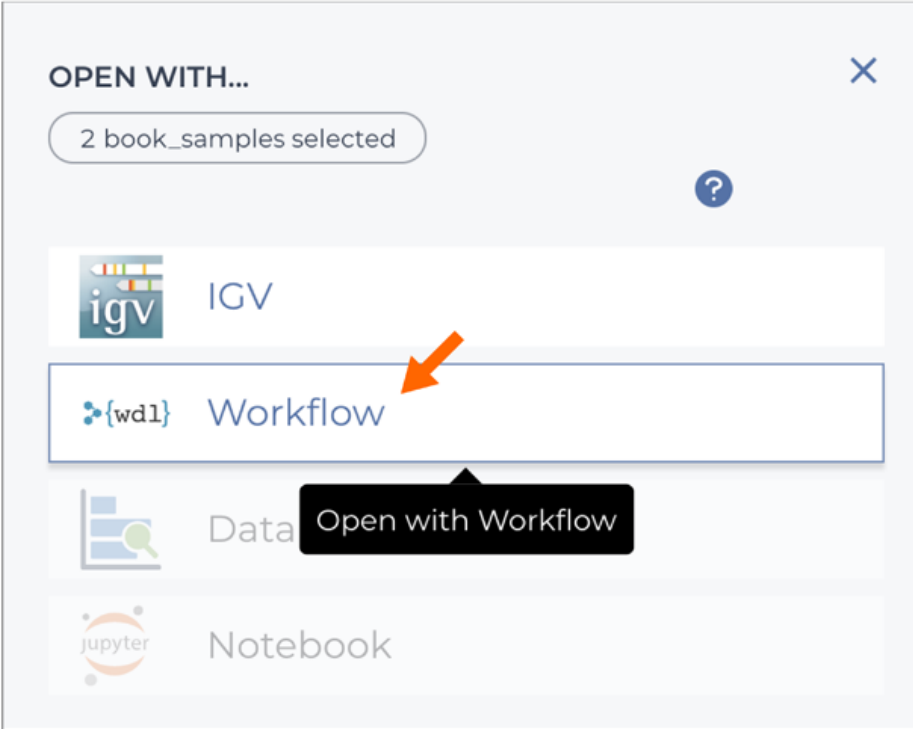
Export to Workspace

Delete Data



# Specifying the workflow

A.



OPEN WITH...

2 book\_samples selected

IGV

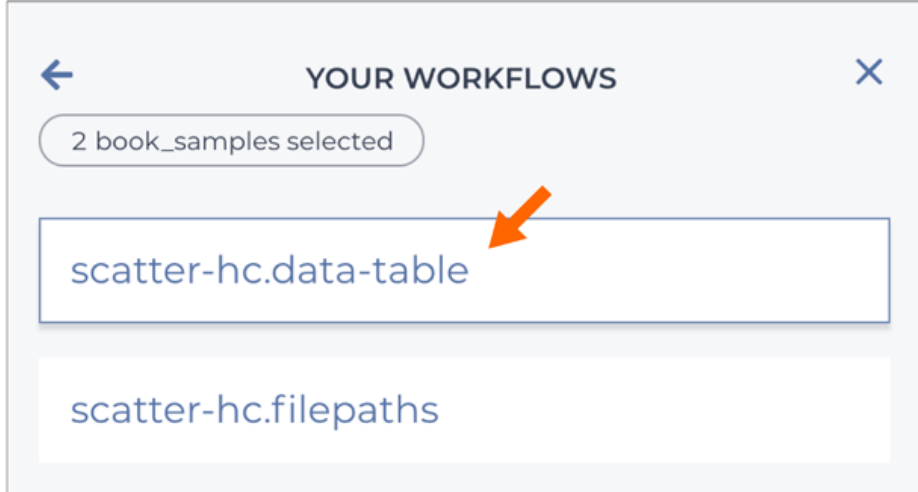
{wd1} Workflow

Data

Open with Workflow

jupyter Notebook

B.



← YOUR WORKFLOWS

2 book\_samples selected

scatter-hc.data-table

scatter-hc.filepaths



# Monitoring the workflow

Submission (click for details)	Data entity	No. of Workflo...	Status	Actions	Submitted	Submission...
scatter-hc.data-table Submitted by genomics.book@gmail.com	scatter-hc-data-table_2...	2	Submitted	<button>ABORT WORKFLOWS</button>	Today	21dccf11-c...
scatter-hc.filepaths Submitted by genomics.book@gmail.com		1	Done		Today	d48d9fb5-...

LIST VIEW

INPUTS

OUTPUTS

LABELS

TIMING DIAGRAM

Task Name	Status	Start	Duration	Inputs	Outputs	Links	Attempts
<a href="#">HaplotypeCallerGVCF</a>							
MergeVCFs							1

Task Name

[HaplotypeCallerGVCF](#)

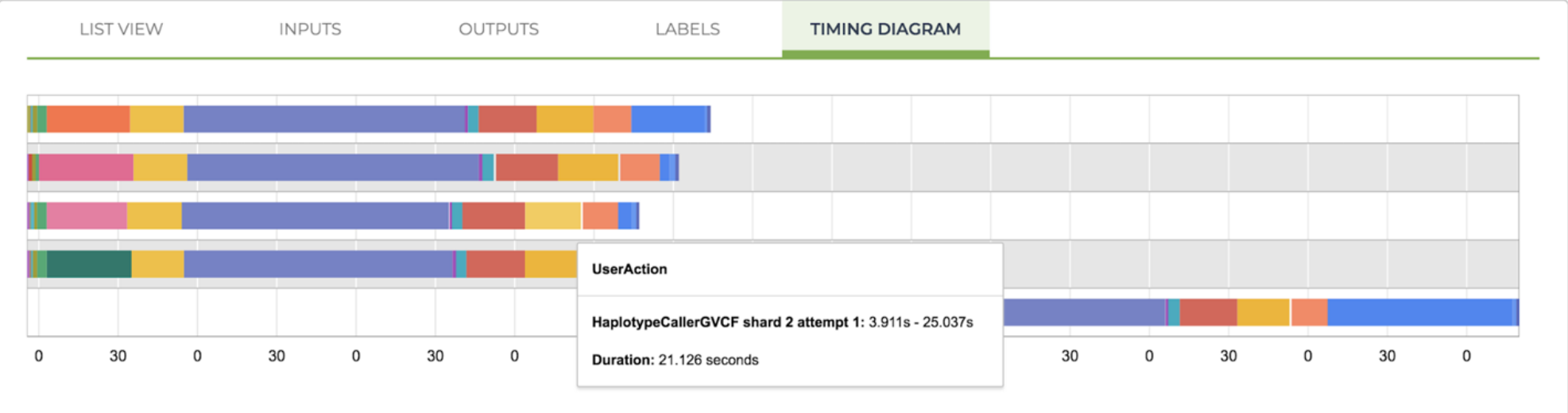
MergeVCFs

Scattered: 4 shards

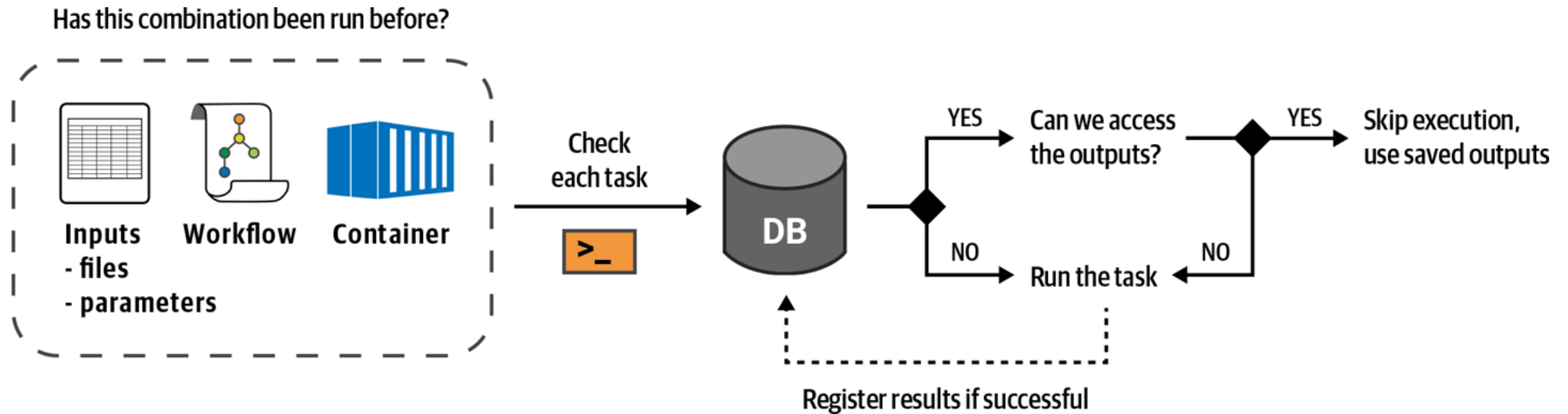
4 0 0 0



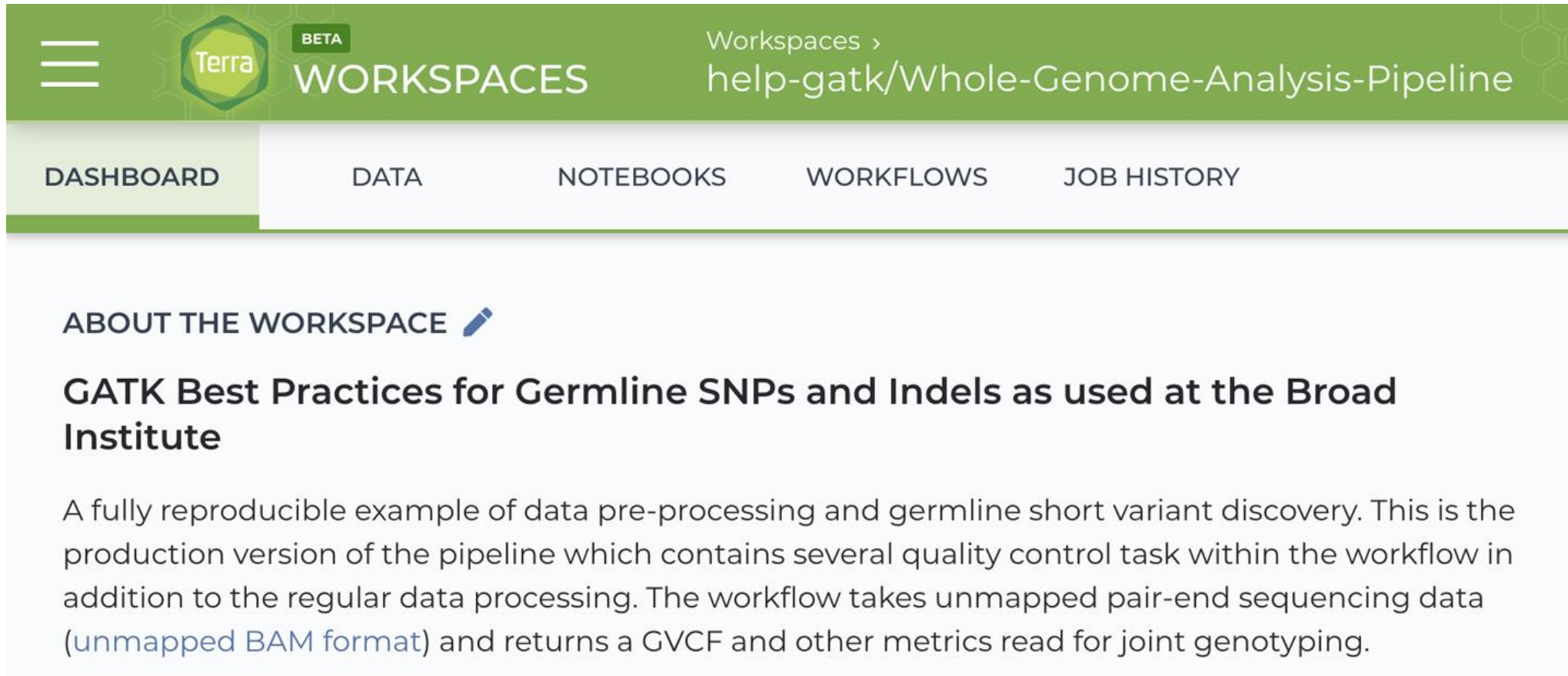
# Timing diagram



# Cromwell caching mechanism



# GATK Best Practices at full scale



The screenshot shows the Terra Workspaces interface. At the top, there is a green header bar with the Terra logo (a green hexagon with 'Terra' inside) and a 'BETA' badge. To the right of the logo, the text 'WORKSPACES' is displayed. Further right, the text 'Workspaces > help-gatk/Whole-Genome-Analysis-Pipeline' is visible. Below the header bar is a navigation bar with five tabs: 'DASHBOARD' (highlighted in green), 'DATA', 'NOTEBOOKS', 'WORKFLOWS', and 'JOB HISTORY'. Below the navigation bar, the main content area has a heading 'ABOUT THE WORKSPACE' followed by a blue pencil icon. The title of the workspace is 'GATK Best Practices for Germline SNPs and Indels as used at the Broad Institute'. Below the title is a paragraph of text: 'A fully reproducible example of data pre-processing and germline short variant discovery. This is the production version of the pipeline which contains several quality control task within the workflow in addition to the regular data processing. The workflow takes unmapped pair-end sequencing data (unmapped BAM format) and returns a GVCF and other metrics read for joint genotyping.'



# Workflows and subworkflows

LIST VIEW

Task Name

[UnmappedBamToAligned...](#)

[AggregatedBamQC](#)

CollectRawWgsMetrics

CollectWgsMetrics

[BamToGvcf](#)

[BamToCram](#)

BamToGvcf

ID: 503d2bfa-8444-4f86-aaf0-caf65753bc16

Status: Succeeded

Tasks: 5 succeeded, 0 failed, 0 currently being processed

Submitted: Aug 13, 2019

Started: Aug 13, 2019

Ended: Aug 14, 2019 (19h 8m)

LIST VIEW

INPUTS

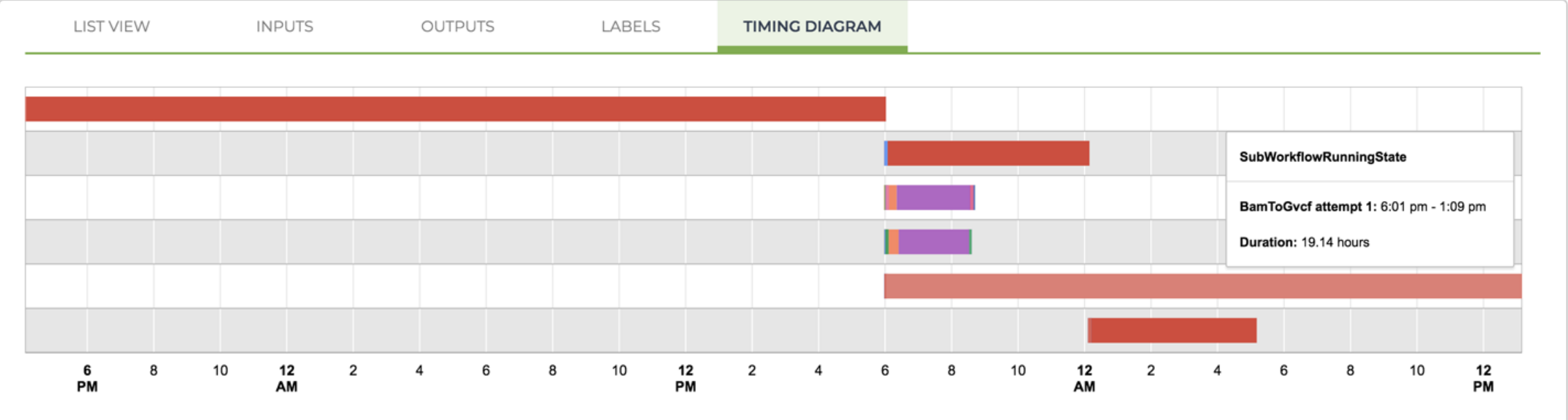
OUTPUTS

TIMING DIAGRAM

Task Name	Status	Start	Duration	Inputs	Outputs	Links	Attempts
ScatterIntervalList	✓	Aug 13, 2019	0h 4m				1
<a href="#">HaplotypeCallerGATK3</a>	✓	Aug 13, 2019	16h 30m				
MergeVCFs	✓	Aug 14, 2019	1h 8m				1
CollectVariantCallingMetrics	✓	Aug 14, 2019	1h 24m				1
ValidateVCF	✓	Aug 14, 2019	1h 24m				1



# Timing diagram





# Downloading data...or not

**A.**

### File Details

Filename  
NA12878.ubams.list

Preview

```
gs://broad-public-datasets/NA12878/unmapped/HJYFJCCXX.4.  
gs://broad-public-datasets/NA12878/unmapped/HJYFJCCXX.5.  
gs://broad-public-datasets/NA12878/unmapped/HJYFJCCXX.6.  
gs://broad-public-datasets/NA12878/unmapped/HJYFJCCXX.7.  
gs://broad-public-datasets/NA12878/unmapped/HJYFJCCXX.8.  
gs://broad-public-datasets/NA12878/unmapped/HJYN2CCXX.1.  
gs://broad-public-datasets/NA12878/unmapped/HK35MCCXX.1.  
gs://broad-public-datasets/NA12878/unmapped/HK35MCCXX.2.  
gs://broad-public-datasets/NA12878/unmapped/HK35MCCXX.3.  
gs://broad-public-datasets/NA12878/unmapped/HK35MCCXX.4.  
gs://broad-public-datasets/NA12878/unmapped/HK35MCCXX.5.  
gs://broad-public-datasets/NA12878/unmapped/HK35MCCXX.6.  
gs://broad-public-datasets/NA12878/unmapped/HK35MCCXX.7.  
gs://broad-public-datasets/NA12878/unmapped/HK35MCCXX.8.  
gs://broad-public-datasets/NA12878/unmapped/HK35MCCXX.1.
```

File size  
1.9 KB

View this file in the Google Cloud Storage Browser

DOWNLOAD FOR < \$0.01\*

Terminal download command

```
gsutil cp gs://broad-public-datasets/NA12878/unma
```

> More Information

\* Estimated. Download cost may be higher in China or Australia.

DONE

**B.**

### File Details

Filename  
NA12878.filtered.vcf.gz

File can't be previewed.

File size  
185.67 MB

View this file in the Google Cloud Storage Browser

DOWNLOAD FOR \$0.02\*

Terminal download command

```
gsutil cp gs://fc-ef9abb30-2a10-4229-b574-895e3ac
```

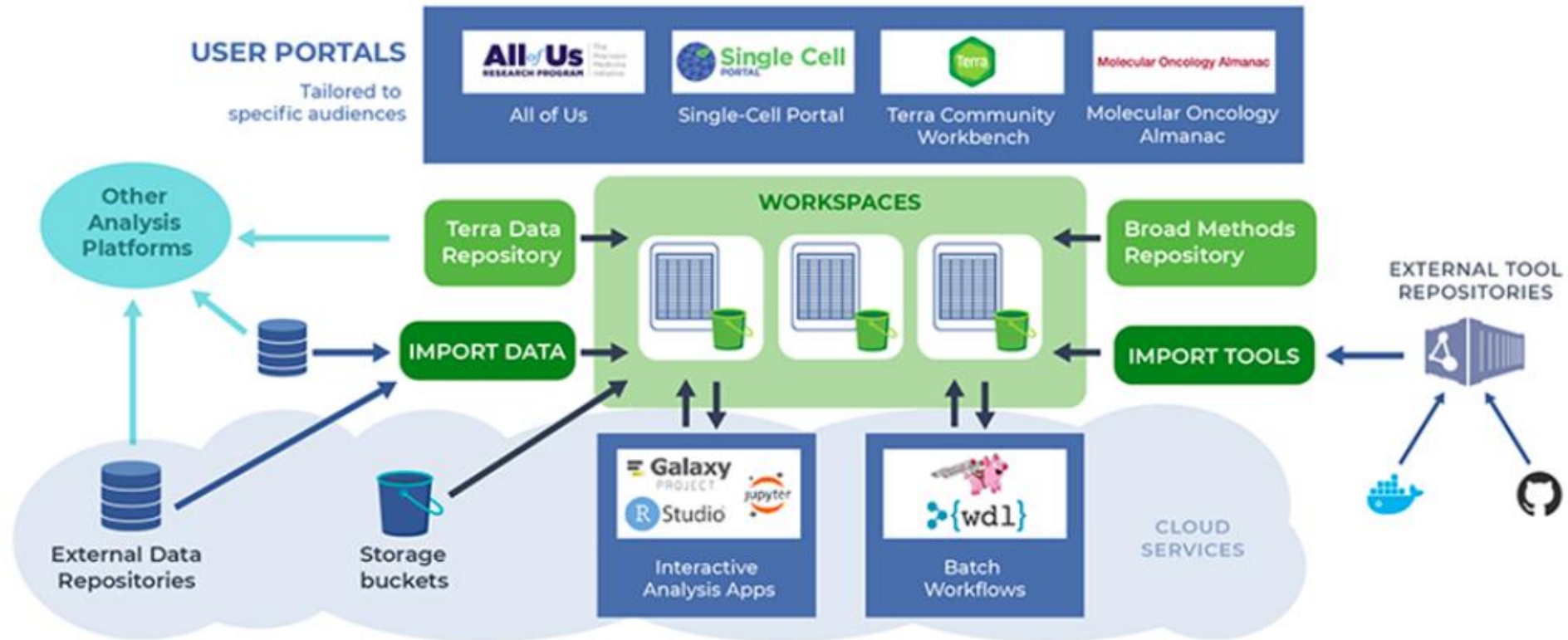
> More Information

\* Estimated. Download cost may be higher in China or Australia.

DONE



# Terra is part of an open data ecosystem



# Additional resources

- Terra documentation
  - [https://www.youtube.com/playlist?list=PLh\\_zJaZ9uQ7P0w6bMLWgL8oDul2EiNlv6](https://www.youtube.com/playlist?list=PLh_zJaZ9uQ7P0w6bMLWgL8oDul2EiNlv6)
  - <https://app.terra.bio/#workspaces/fc-product-demo/Terra-Data-Tables-Quickstart>
  - <https://app.terra.bio/#workspaces/fc-product-demo/Terra-Workflows-Quickstart>
- Terra Library Showcase
  - <https://app.terra.bio/#library/showcase>
  - <https://app.terra.bio/#workspaces/warp-pipelines/Whole-Genome-Analysis-Pipeline>
- Terra community forum
  - <https://support.terra.bio/hc/en-us/community/topics/360000500432>



A detailed illustration of a pufferfish, likely a species of porcupinefish, with its mouth open. The fish has a dark, mottled pattern on its upper body and a lighter, more uniform pattern on its lower body. Its mouth is wide open, revealing a row of small, sharp teeth. The background is a plain, light color.

# Thank you for joining us today!

Next week: Chapter 12

Next meeting: February 22, 2021