



Genomics in the Cloud

Book Club - Week 14

March 1, 2021

Agenda

- Chapter 13: Assembling Your Own Workspace in Terra
- Additional resources
- Open discussion



Chapter 13: Assembling Your Own Workspace 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.

Our guest speaker

Dr. Adelaide
Rhodes



Getting started

Proxy Group ⓘ

PROXY_@firecloud.org

<https://app.terra.bio/#profile>



Bucket details



EDIT BUCKET



REFRESH BUCKET

genomics-book-test

Objects

Overview

Permissions

Bucket Lock



This bucket uses **fine-grained** access control, allowing you to specify access to individual objects. To control access uniformly at the bucket level, switch to uniform access control. [Learn more](#)

Edit

Add members

Remove



Filter by name or role

View by:

Members ▾



Type

Members ^

Role(s)



Granting access

Add members and roles for "genomics-book-test" 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

PROXY_@firecloud.org



Select a role

Condition

Type to filter

Dataproc

Firebase

Firebase Products

IAM

Other

Service Management

Storage

Storage Legacy

Storage Admin

Storage Object Admin

Storage Object Creator

Storage Object Viewer

MANAGE ROLES

3

Storage Object Admin

Full control of GCS objects.



6

Creating a new workspace

Create a New Workspace

Workspace name *

My first workspace

Billing project *

fccredits-cerium-white-3390

Description

Recreating the workspace from the genomics book

Authorization domain 

Select groups

CANCEL

CREATE WORKSPACE



Creating a new method

Create New Method ✕

Namespace

geraldine-and-brian

Only letters, numbers, underscores, dashes, and periods allowed

Name

scatter-hc

Only letters, numbers, underscores, dashes, and periods allowed

WDL

[Load from file...](#)

Selected: scatter-haplotypecaller.wdl

[Reset to file](#)

Undo

Redo

```
1 ## This workflow runs the HaplotypeCaller tool from GATK4 in GVCF mode
2 ## on a single sample in BAM format. The execution of the HaplotypeCaller
3 ## tool is parallelized using an intervals list file. The per-interval
4 ## output GVCF files are then merged to produce a single GVCF file for
5 ## the sample, which can then be used by the joint-discovery workflow
6 ## according to the GATK Best Practices for germline short variant
7 ## discovery.
8
9 version 1.0
10
11 workflow ScatterHaplotypeCallerGVCF {
12
13   input {
14     File input_bam
15     File input_bam_index
16     File intervals_list
17   }
18
19   String output_basename = basename(input_bam, ".bam")
20
21   Array[String] calling_intervals = read_lines(intervals_list)
22 }
```

Documentation (optional)

Edit

Preview

Side-by-side

[Populate from WDL comment](#)

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.

Synopsis (optional, 80 characters max)

Run scattered HaplotypeCaller (GATK4) in GVCF mode on a single sample BAM

Snapshot Comment (optional)

Cancel

Upload



Summary workflow page

METHOD

geraldine-and-brian/scatter-hc

SNAPSHOT


1 ▾


Export to Workspace...


Summary


WDL

Configurations

 | [Permissions...](#)

 | [Edit...](#)

 | [Clone...](#)

 | [Redact](#)

Synopsis

Run scattered HaplotypeCaller (GATK4) in GVCF mode on a single sample BAM

Method Owner

genomics.book@gmail.com

▼ Documentation

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.

Snapshot Comment

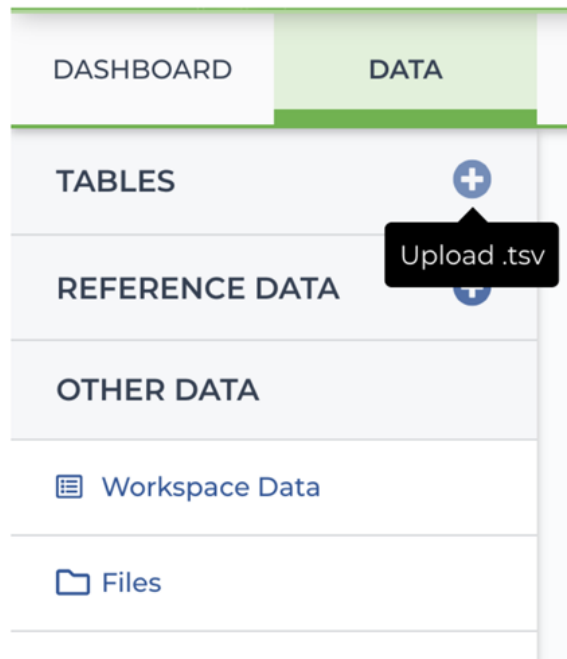
Created

January 5, 2020, 10:24 AM

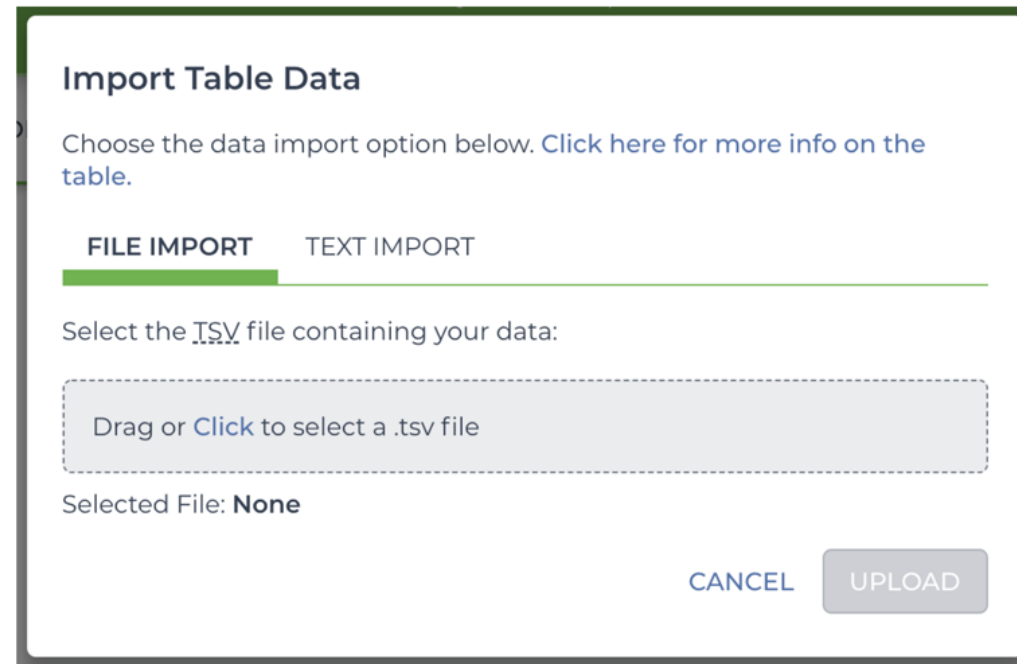


Adding the data table

	A	B	C	D	E	F
1	entity:book_sample_id	input_bam	input_bam_index			
2	mother	gs://genomics-in-t	gs://genomics-in-the-cloud/v1/data/germline/bams/mother.bai			
3	father	gs://genomics-in-t	gs://genomics-in-the-cloud/v1/data/germline/bams/father.bai			
4	son	gs://genomics-in-t	gs://genomics-in-the-cloud/v1/data/germline/bams/son.bai			
5						



A.



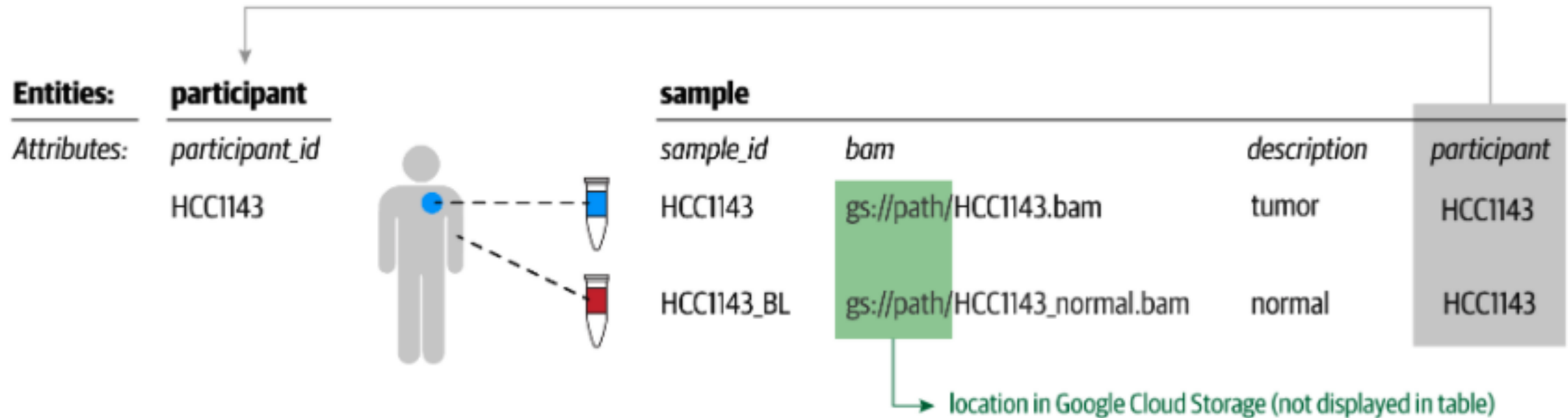
B.



Data model



Data model



The formal description of the various entities involved in your experimental design and how they relate to each other is called the *data model*.



1000 Genomes Project in Terra

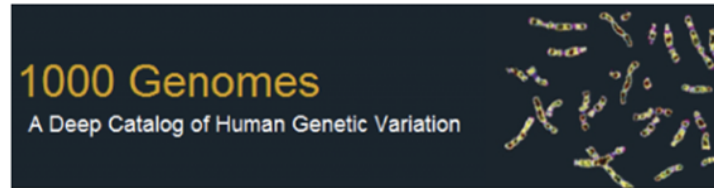


1000 Genomes High Coverage presented by NHGRI AnVIL

1000 Genomes project phase 3 samples sequenced to 30x coverage. This dataset is delivered as a workspace. You may clone this workspace to run analyses or copy specific samples to a workspace of your choice.

Participants: 2,504

[BROWSE DATA](#)



1000 Genomes Low Coverage

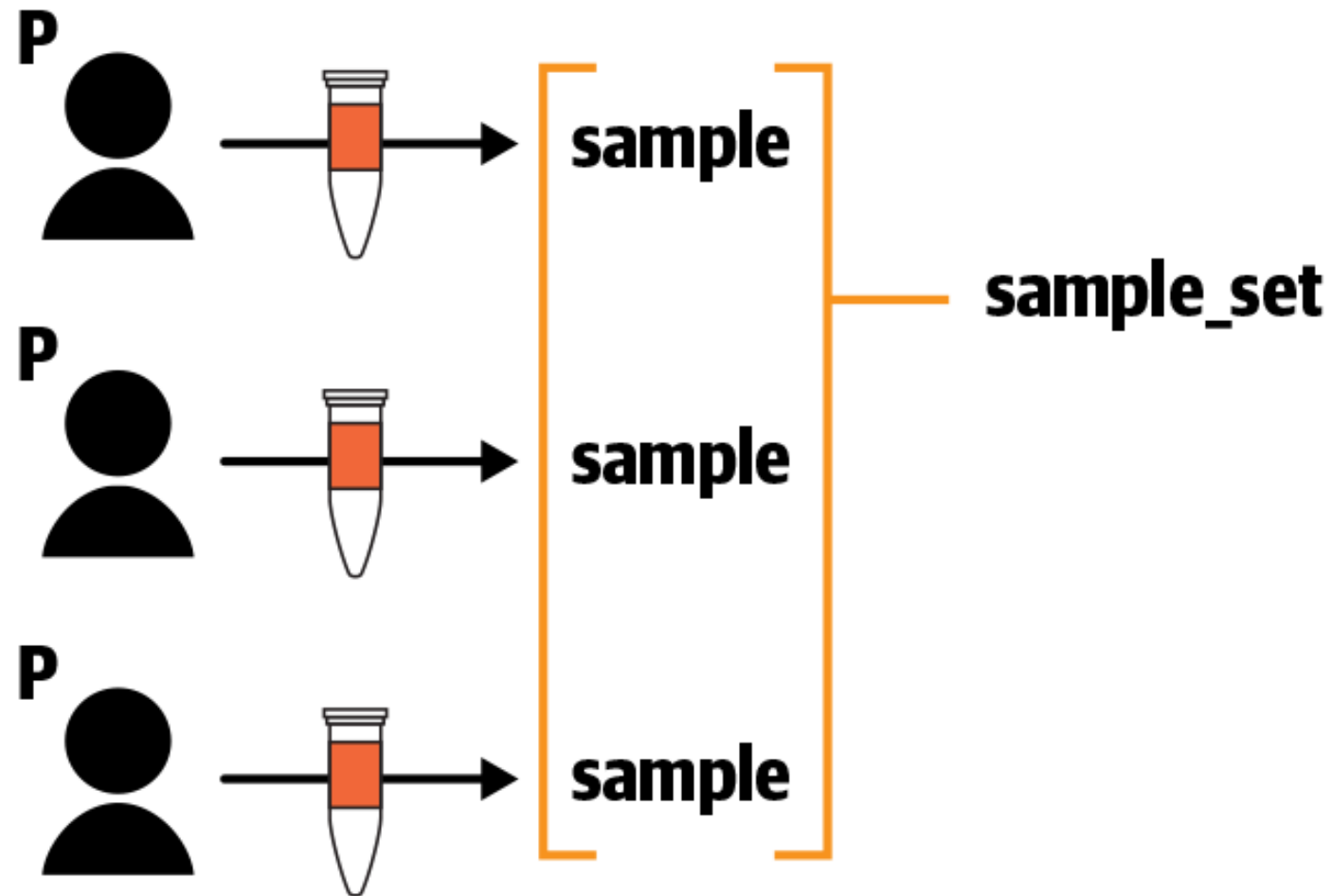
The [1000 Genomes Project](#) ran between 2008 and 2015, creating the largest public catalogue of human variation and genotype data. The goal of the 1000 Genomes Project was to find most genetic variants with frequencies of at least 1% in the populations studied.

Participants: 3,500

[BROWSE DATA](#)



Data model for 1000 Genomes data set



Copy data to Workspace

TABLES +

participant (2504)

sample (2504)

sample_set (1)

REFERENCE DATA +

OTHER DATA

Workspace Data

DOWNLOAD ALL ROWS

COPY PAGE TO CLIPBOARD

5 rows selected ⓘ

Search

<input type="checkbox"/>	sample_id ↓	cram	gVCF	gVCF
<input checked="" type="checkbox"/>	SRS000030	NA06985.final.cram	NA06985.haplotypeCalls.er.raw.g.vc...	NA06985.haplotypeCalls.er.raw.g.vc...
<input checked="" type="checkbox"/>	SRS000031	NA06986.final.cram	NA06986.haplotypeCalls.er.raw.g.vc...	NA06986.haplotypeCalls.er.raw.g.vc...
<input checked="" type="checkbox"/>	SRS000032	NA06994.final.cram	NA06994.haplotypeCalls.er.raw.g.vc...	NA06994.haplotypeCalls.er.raw.g.vc...
<input checked="" type="checkbox"/>	SRS000033	NA07000.final.cram	NA07000.haplotypeCalls.er.raw.g.vc...	NA07000.haplotypeCalls.er.raw.g.vc...
<input checked="" type="checkbox"/>	SRS000034	NA07037.final.cram	NA07037.haplotypeCalls.er.raw.g.vc...	NA07037.haplotypeCalls.er.raw.g.vc...

Download as TSV

Open with...

Export to Workspace

Send the selected data to another workspace



Direct import of TSV-formatted data

Import Table Data

Choose the data import option below. [Click here for more info on the table.](#)

FILE IMPORT

TEXT IMPORT

Copy and paste tab separated data here:

Clear

```
entity:sample_set_id  
federated-dataset
```



Data with the type 'sample_set' already exists in this workspace.
Uploading more data for the same type may overwrite some entries.

CANCEL

UPLOAD



Start / end rows of sample_set_membership.tsv

	A	B
1	membership:sample_set_id	sample
2	1000G-high-coverage-2019-all	SRS000030
3	1000G-high-coverage-2019-all	SRS000031

...

2505	1000G-high-coverage-2019-all	SRS000631
2506	one_sample	NA12878



Updated membership file with 25 samples

	A	B
1	membership:sample_set_id	sample
2	federated-dataset	SRS000030
3	federated-dataset	SRS000031

...

25	federated-dataset	SRS000055
26	federated-dataset	NA12878



sample_set table showing three sample sets

<input type="checkbox"/> ▼	sample_set_id ↓	samples
<input type="checkbox"/>	1000G-high-coverage-2019-all	2504 entities
<input type="checkbox"/>	federated-dataset	25 entities
<input type="checkbox"/>	one_sample	1 entity



Input configuration details

JointGenotyping	input_gvcfs	Array[File]	<div>this.samples.gvcf</div> {...}
JointGenotyping	input_gvcfs_indices	Array[File]	<div>this.samples.gvcf_index</div> {...}



Data tables – more info

<https://support.terra.bio/hc/en-us/articles/360051043031-Making-modifying-and-deleting-data-tables>

How to make tables of tumor-normal sample pairs



Uploading tables in a workspace



Tip title

Tip text.



Upload (nested) tables in a particular order!

If data tables reference entities in another table, the dependent table needs to be uploaded first. The order is as follows ("A > B" means entity type A must be uploaded before entity type B):

- participants > samples
 - samples > pairs
 - participants > participant sets
 - samples > sample sets
 - pairs > pair sets
 - set membership > set entity
- (e.g. *participants* before *samples* before *sample set membership* before *sample set entity*)



"Joint discovery" workflows in Dockstore

Expand All ^

Collapse All x

Search

Enter search term
joint discovery

Open Advanced Search

Entry Type

☒ workflow (7)

Language

☒ WDL (7)

Author

☐ n/a (4)

Share

Search: contains one of "joint, discovery"AND the Entry Type is workflow AND the Language is WDL

Browse Tools

Browse Workflows

Tag Cloud

A workflow is a series of tools strung together, with an associated descriptor describing how to run it.

Name	Verified	Author	Format	Project Links	Stars ↓
gatk-workflows/gatk4-germline-snp-indels/gatk4-germline-snp-indels-haplotypecaller-gvcf-calling		n/a	WDL	GitHub	
gatk-workflows/gatk4-germline-snp-indels/haplotypecaller-gvcf-gatk4-nio		n/a	WDL	GitHub	
gatk-workflows/gatk4-germline-snp-indels		n/a	WDL	GitHub	
gatk-workflows/gatk4-germline-snp-indels/joint-discovery-gatk4		n/a	WDL	GitHub	



Changing an Existing Workflow

A few months ago, Terra.bio announced a public workspace containing best-practices workflows for viral genome analysis developed by Dr. Danny Park's Viral Genomics group and used to process COVID-19 research data. Here is a blog post describing the workflows in the new workspace.

<https://terra.bio/workflow-updates-to-the-covid-19-workspace-better-viral-assembly-and-phylogenetics-with-nextstrain/>

This workspace contains the Broad's viral genomics team's reference based viral assembly tool (assemble_refbased.wdl), which they've updated to reflect these best practices and is appropriate for use on any Illumina data generated from SARS-CoV-2.

The Starting Point of the workspace can be either fastq files downloaded from the Sequence Read Archive (SRA) or a list of SRA accession numbers that will generate unaligned BAM files (uBAM) from SRA files.



Changing an Existing Workflow – Why?

During this same time period, the NIH National Center for Biotechnology Information was making the leap to cloud

<https://www.ncbi.nlm.nih.gov/sars-cov-2/>

SARS-CoV-2 datasets are now loaded into publicly available buckets on GCP

- 1.) <gs://nih-sequence-read-archive/run/ERR4308581/> - bucket location for this accession number
- 2.) <gs://nih-sequence-read-archive/sra-src/> - bucket locations of submissions (before conversion to SRA)

SRA Aligned Read Formats (SARFs) provide a few more options other than fastq-dump, for example,

- 1.) contigs created from the raw reads in the run (fasta format) and
- 2.) reads are aligned back to the contigs (sam format).

<https://www.ncbi.nlm.nih.gov/sra/docs/sra-aligned-read-format/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6771016/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7764237/>



Changing an Existing Workflow

```
gsutil ls gs://nih-sequence-read-archive/sra-src/*/fastq*
```

```
gs://nih-sequence-read-archive/sra-src/ERR4080576/DK_ALAB-SSI-147_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080578/DK_ALAB-SSI-149_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080581/DK_ALAB-SSI-154_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080582/DK_ALAB-SSI-155_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080583/DK_ALAB-SSI-156_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080586/DK_ALAB-SSI-159_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080587/DK_ALAB-SSI-160_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080588/DK_ALAB-SSI-162_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080593/DK_ALAB-SSI-168_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080594/DK_ALAB-SSI-169_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080595/DK_ALAB-SSI-171_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080596/DK_ALAB-SSI-172_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080597/DK_ALAB-SSI-173_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080598/DK_ALAB-SSI-174_2020.fastq.gz
gs://nih-sequence-read-archive/sra-src/ERR4080599/DK_ALAB-SSI-183_2020.fastq.gz
```



Changing an Existing Workflow

The Challenge:

1. Test direct ingress from publicly available data in GCP instead of using efetch to load from on-prem resources
2. Test the replacement of the assembly portion of the workflow with SARF objects and compare quality
3. Optimize the cost/speed of running repeated runs on the bucket, which is updated regularly
4. Build notebooks that take advantage of BigQuery Metadata sets (note not currently linked from here due to recent GCP updates)

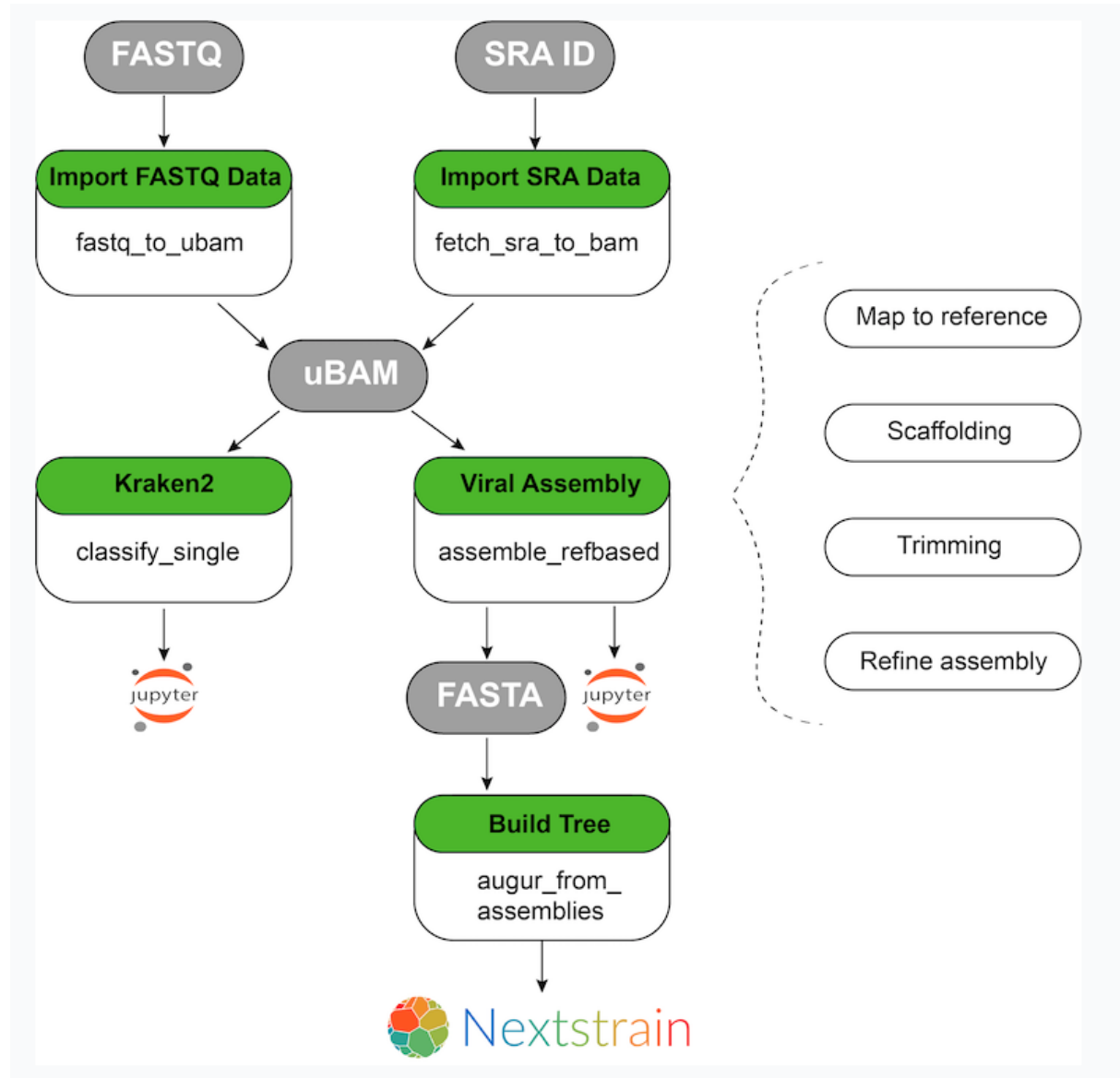
<https://console.cloud.google.com/marketplace/product/national-library-of-medicine/ncbi-covid-data>

Addendum: instructions to access SRA bq from the command line are here:

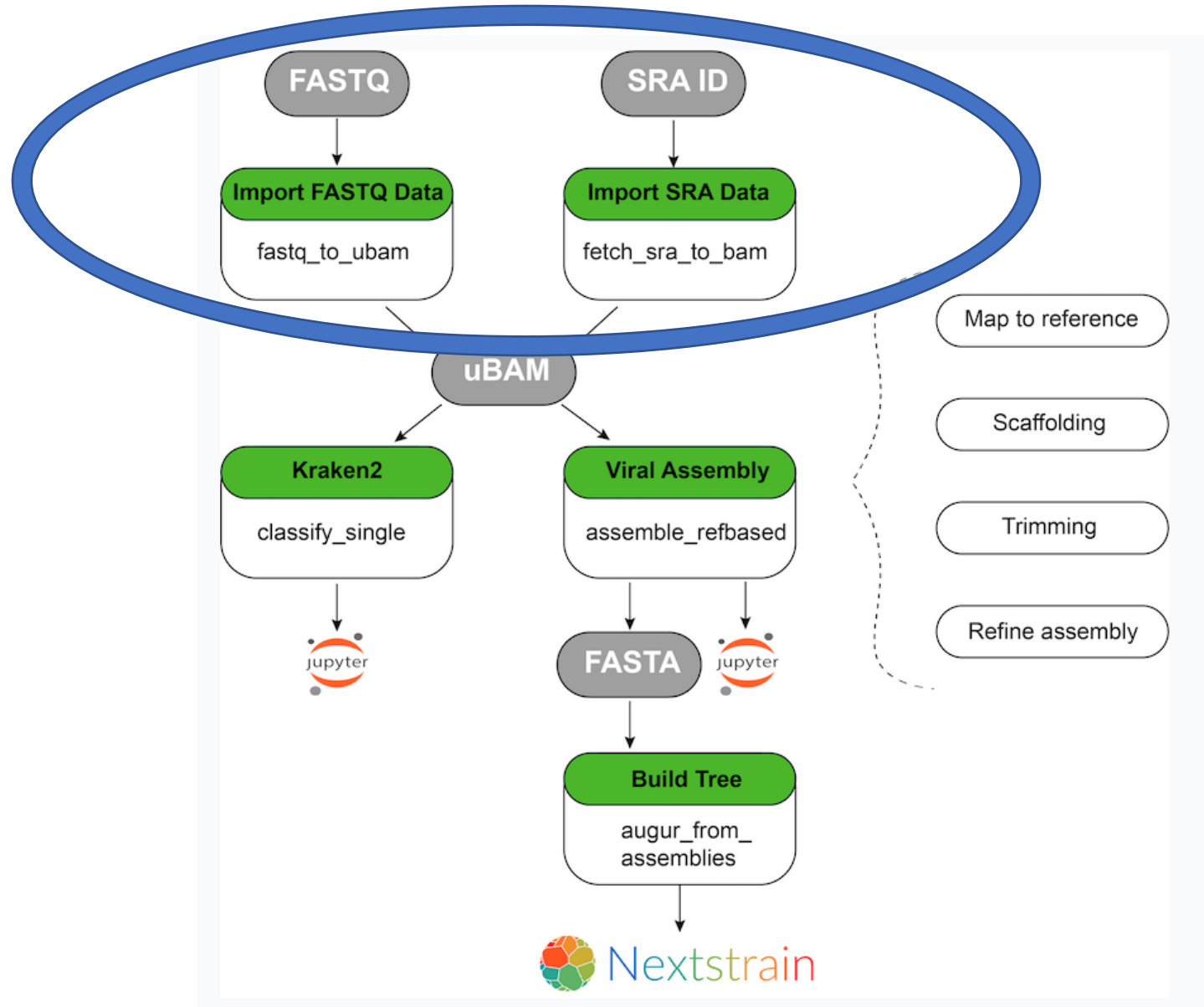
<https://www.ncbi.nlm.nih.gov/sra/docs/sra-bigquery/>



Changing an Existing Workflow



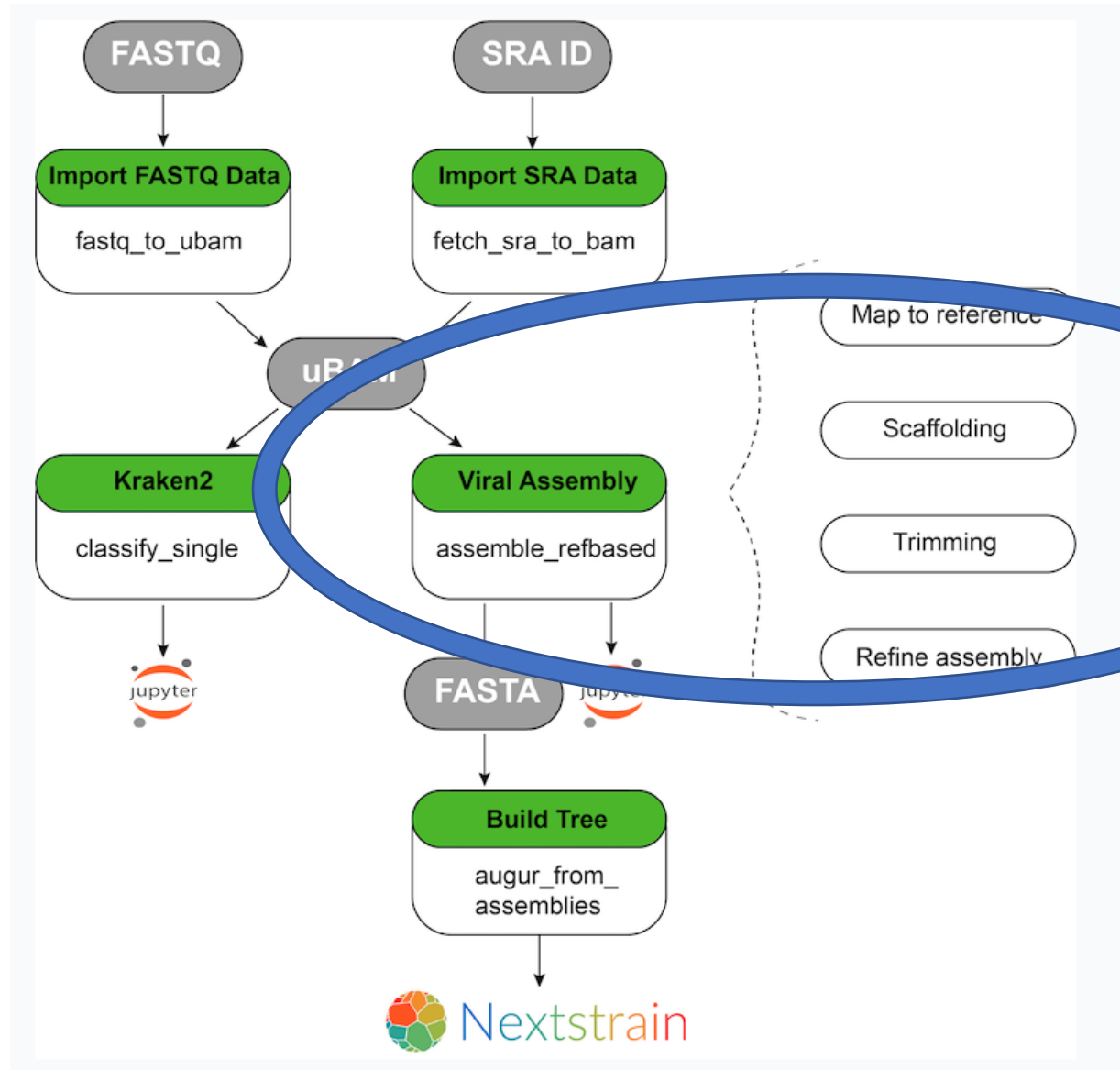
Changing an Existing Workflow



Idea 1:
Direct
ingress
from GCP
buckets



Changing an Existing Workflow



**Idea 2: Pull
in SARF's
objects
instead of
doing
assembly
manually**



Changing an Existing Workflow

jupyter ncbi-datasets-virus (unsaved changes)

[Visit repo](#)[Copy Binder link](#)

Not Trusted

Python 3

Memory: 124 / 2048 MB

Using `ncbi.datasets` library to download and parse virus datasets

The objective of this notebook is to use the `ncbi.datasets` python library to download and extract genome and annotation data for the coronavirus family of viruses, including SARS-CoV-2.

First, let's import the python modules we'll use. Be sure you have first installed the requirements in 'requirements.txt' into your virtual environment.

```
In [1]: import ncbi.datasets
import json
import jsonlines
import os
import csv
import zipfile
import pandas as pd
from pyfaidx import Fasta
from google.protobuf.json_format import ParseDict
import ncbi.datasets.vlalpha1.reports.virus_pb2 as virus_report_pb2
from collections import Counter
from datetime import datetime, timezone, timedelta
```

We will need an API object specific to retrieving viral data. To see all the possible API instances, [visit the documentation on GitHub](#). Let's go ahead and create the API object.

```
In [1]: virus_api = ncbi.datasets.VirusApi(ncbi.datasets.ApiClient())
```



Additional resources

- Terra resources
 - <https://app.terra.bio/#library/datasets>
 - <https://support.terra.bio/hc/en-us/articles/360026775691> (authorization domains)
 - <https://support.terra.bio/hc/en-us/articles/360025674392-Finding-the-workflow-method-you-need-and-its-JSON-in-the-Methods-Repository> (methods repository)
 - <https://support.terra.bio/hc/en-us/articles/360025758392> (workspace tables)
- Analysis, Visualization and Informatics Lab-space (AnVIL)
 - <https://www.genome.gov/Funded-Programs-Projects/Computational-Genomics-and-Data-Science-Program/Genomic-Analysis-Visualization-Informatics-Lab-space-AnVIL>
- Dockstore resources
 - <https://dockstore.org>
 - <https://docs.dockstore.org>





Thank you for joining us today!

Next week: Chapter 14

Next meeting: March 8, 2021