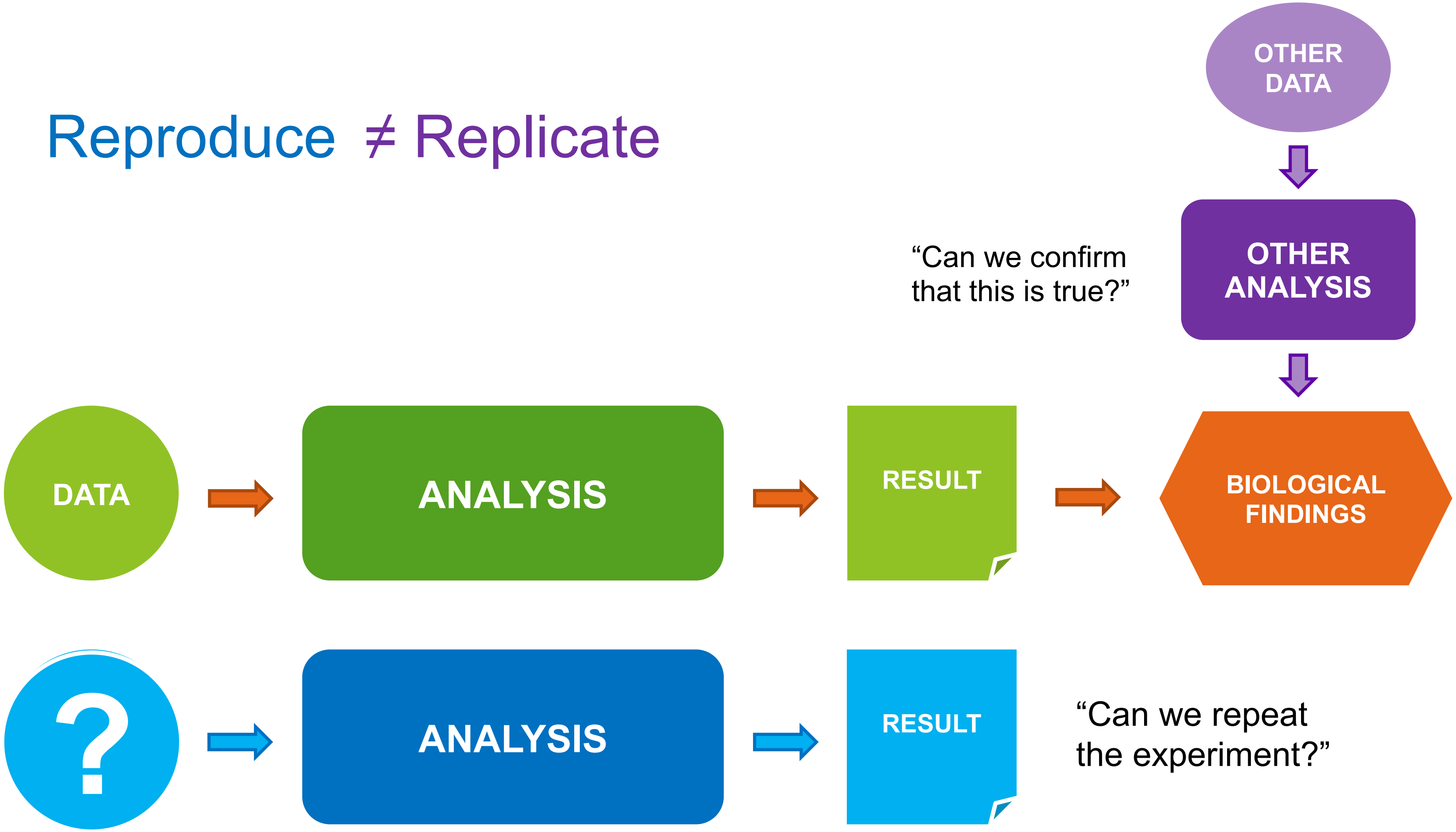
Bringing the Power of Synthetic (Sequence) Data to the Masses

FAIR Hackathon, BioIT World Apr 15-16, 2019

R for REUSE / REPRODUCIBILITY

What does reproducibility mean to you?

Reproduce ≠ Replicate



OTHER DATA

OTHER ANALYSIS

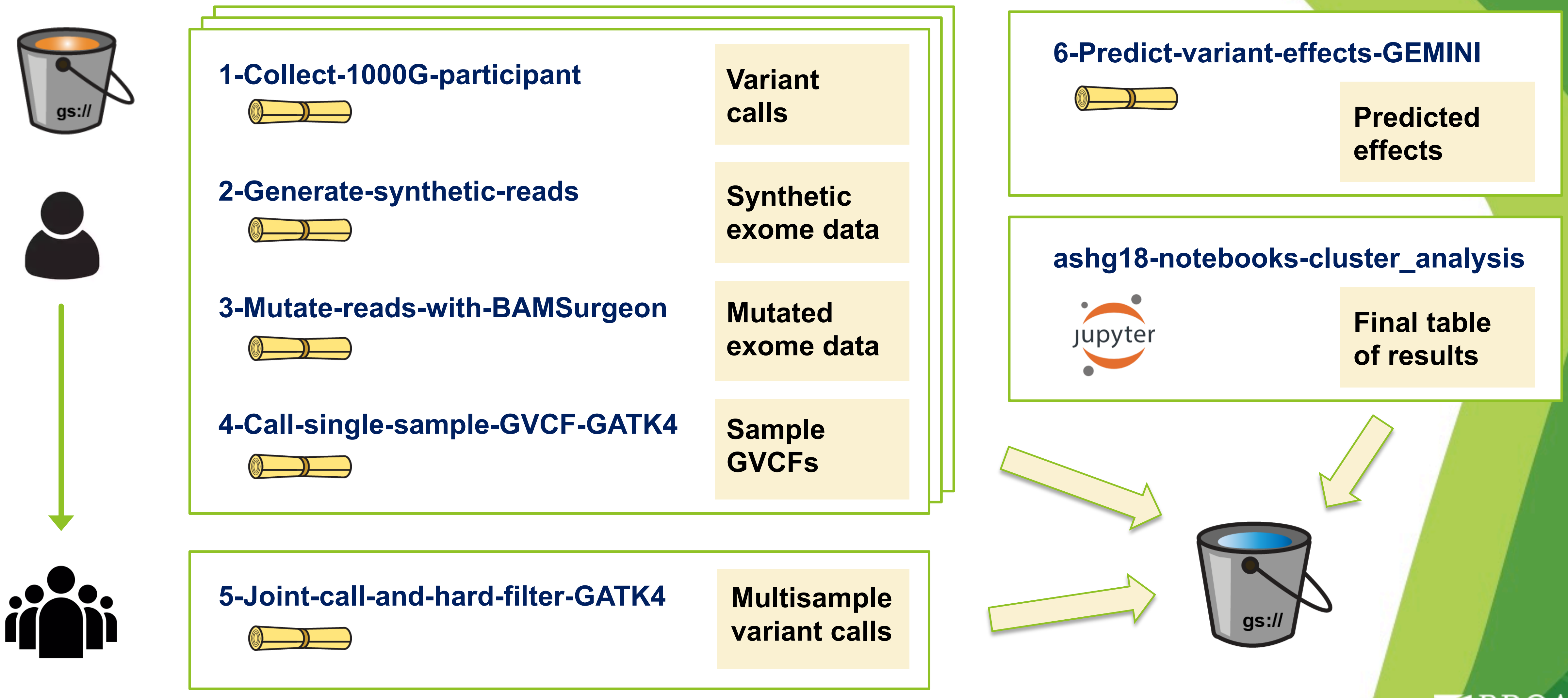
“Can we confirm that this is true?”

DATA → ANALYSIS → RESULT → BIOLOGICAL FINDINGS

? → ANALYSIS → RESULT

“Can we repeat the experiment?”

The original project broad.io/ASHG2018



1-Collect-1000G-participant

Variant calls

2-Generate-synthetic-reads

Synthetic exome data

3-Mutate-reads-with-BAMSurgeon

Mutated exome data

4-Call-single-sample-GVCF-GATK4

Sample GVCFs

5-Joint-call-and-hard-filter-GATK4

Multisample variant calls

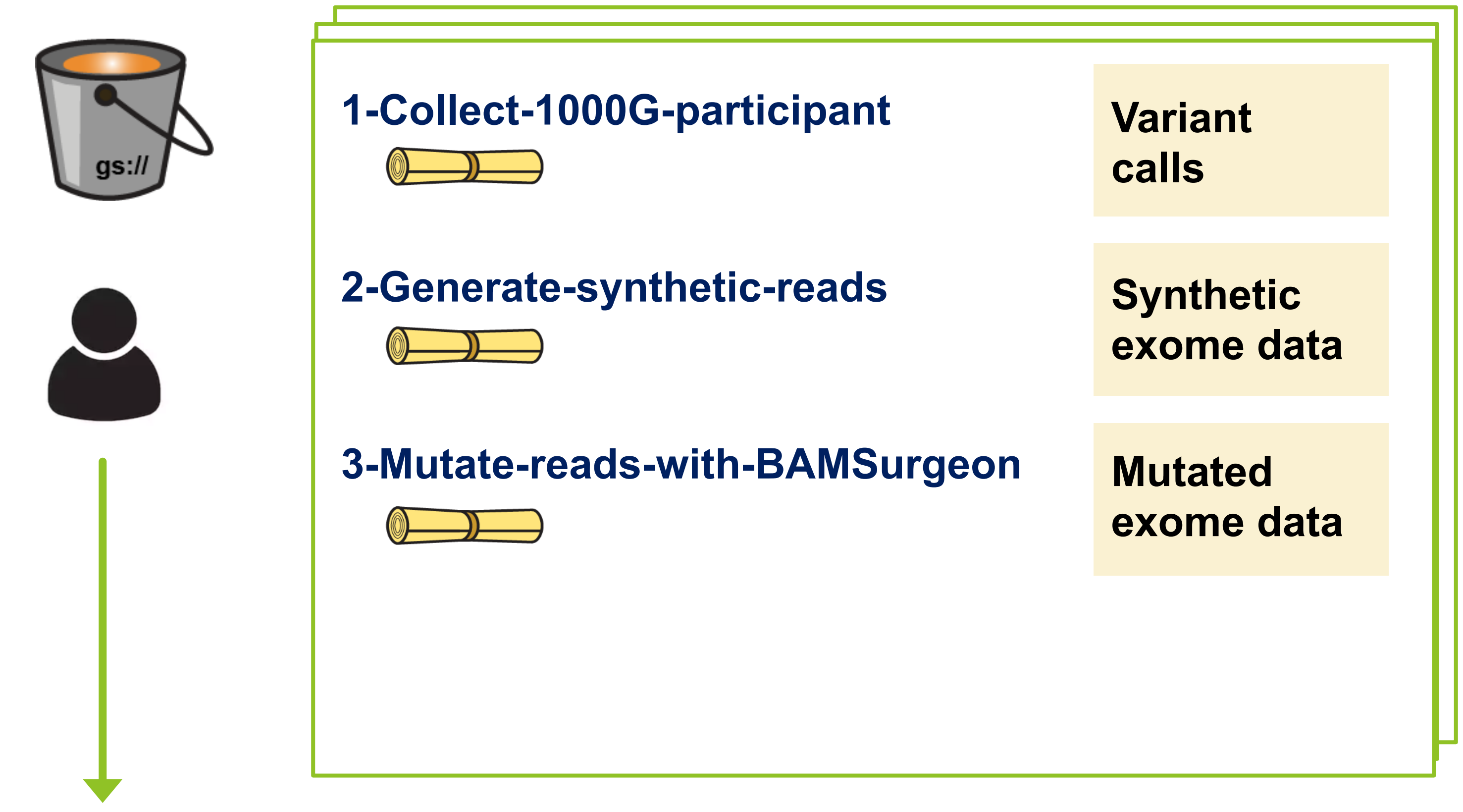
6-Predict-variant-effects-GEMINI

Predicted effects

ashg18-notebooks-cluster_analysis

Final table of results

This hackathon project broad.io/FAIRdatahack2019



1-Collect-1000G-participant

2-Generate-synthetic-reads

3-Mutate-reads-with-BAMSurgeon

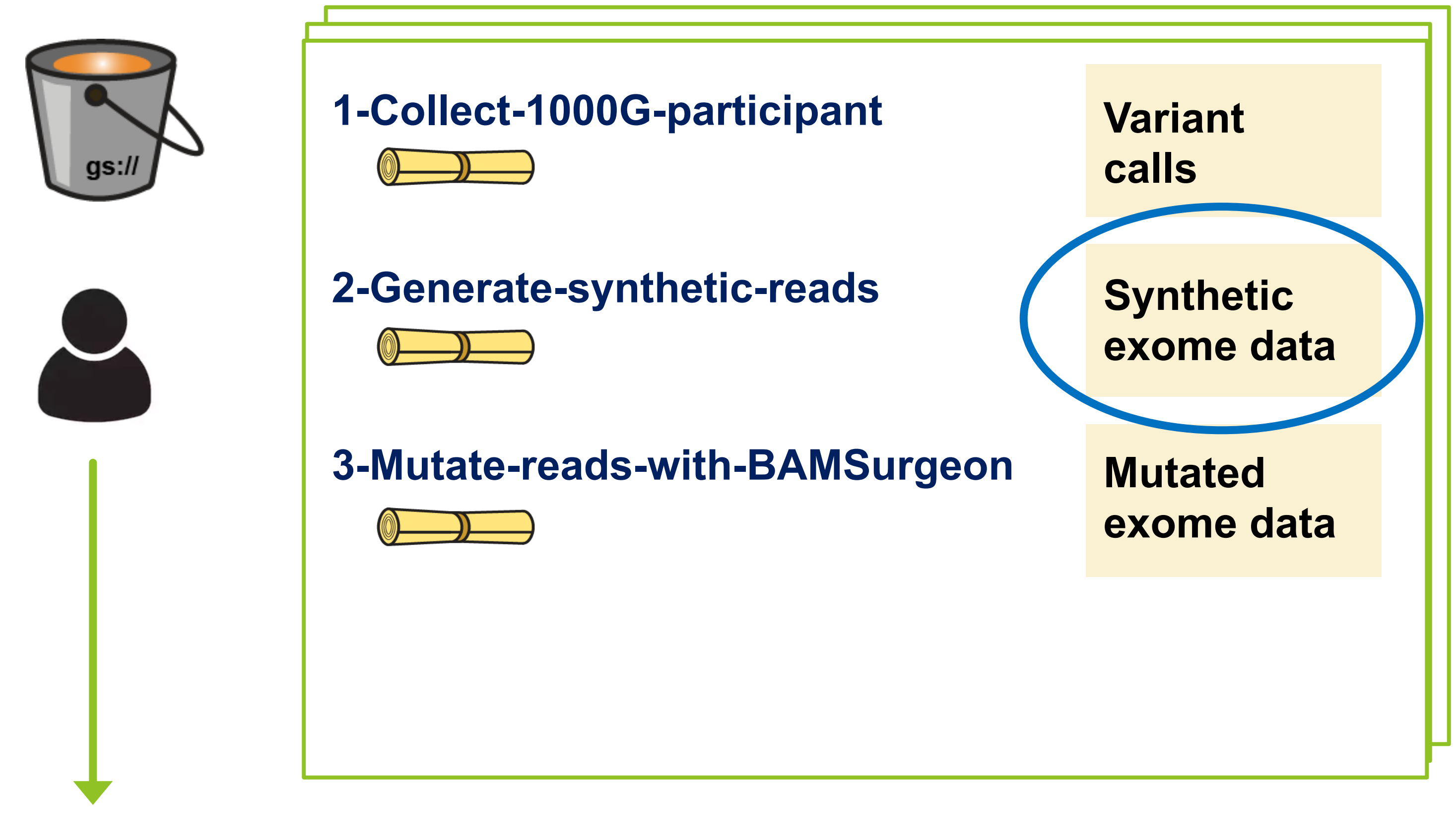
Variant calls

Synthetic exome data

Mutated exome data

How do we turn this into a FAIR community resource to empower biomedical researchers to leverage the underlying tools more easily?

This hackathon project broad.io/FAIRdatahack2019

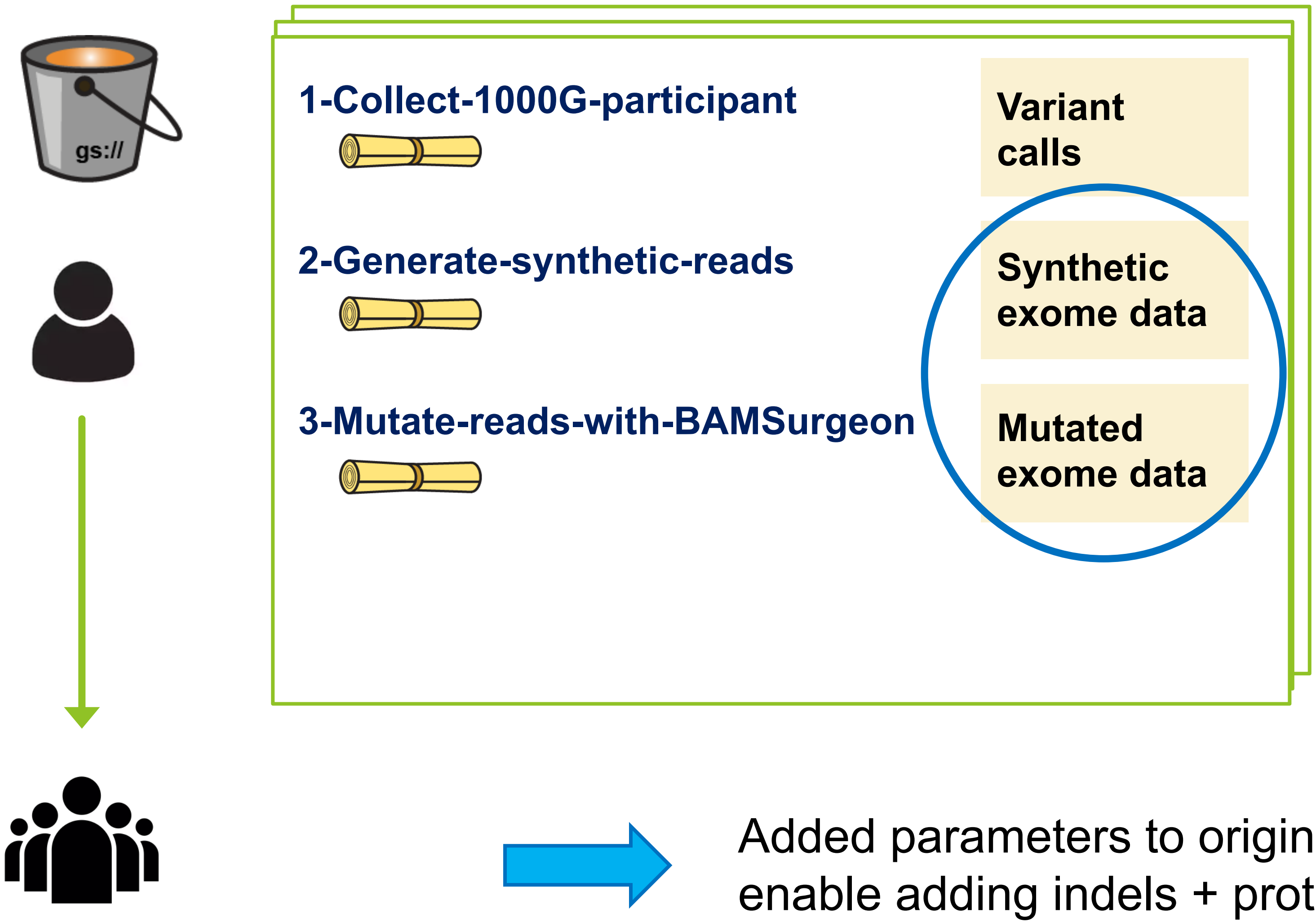

gs://
1-Collect-1000G-participant
2-Generate-synthetic-reads
3-Mutate-reads-with-BAMSurgeon
Variant calls
Synthetic exome data
Mutated exome data

#1 - Data in demand

What kind of datasets would be useful to the community?

Identified top needs based on literature
+ feedback from hackathon participants

This hackathon project broad.io/FAIRdatahack2019

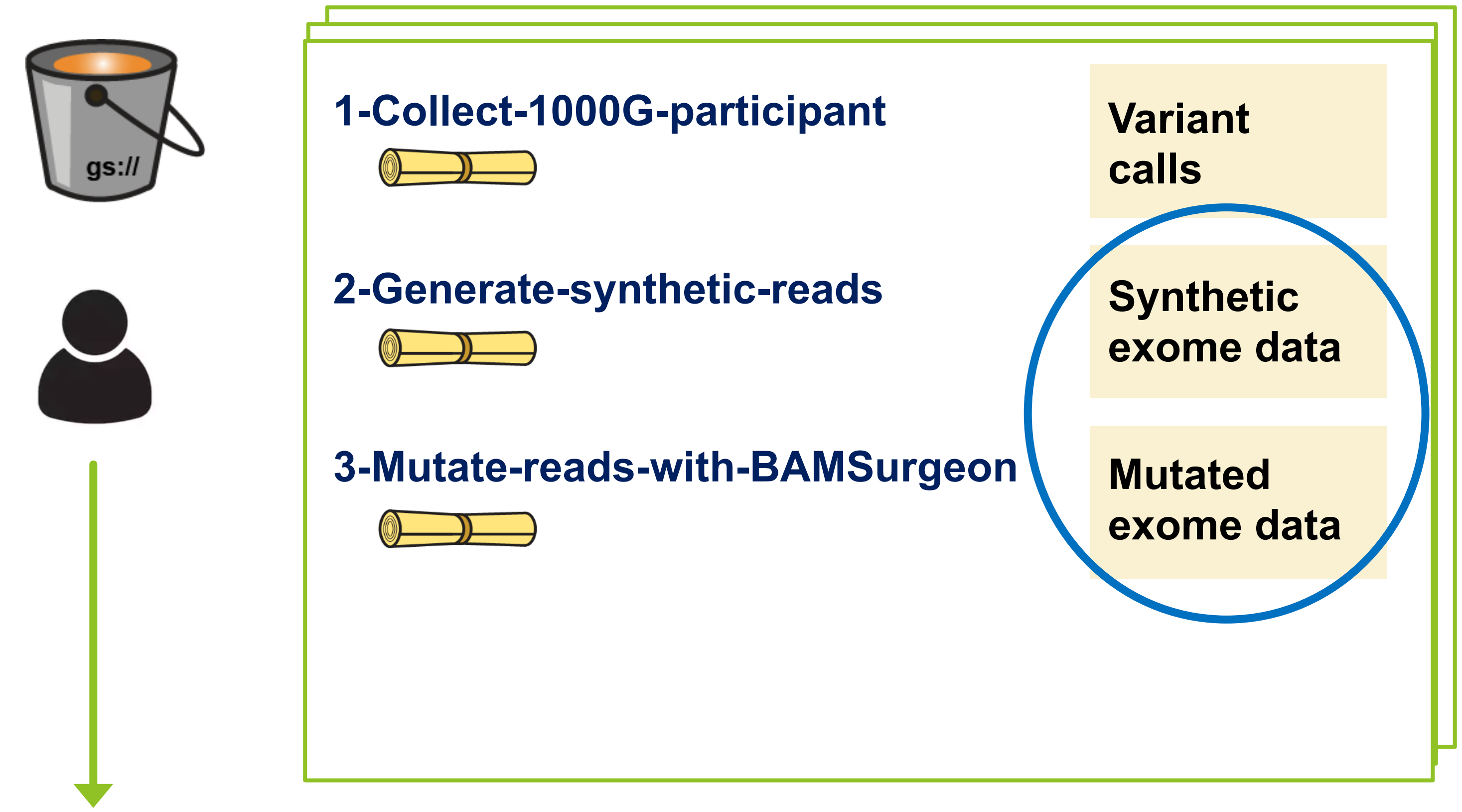

gs://
1-Collect-1000G-participant
2-Generate-synthetic-reads
3-Mutate-reads-with-BAMSurgeon
Variant calls
Synthetic exome data
Mutated exome data

#2 – Diversifying Options

Enable generation of more data types and more variant types

Added parameters to original workflows to enable adding indels + prototyped SVs

This hackathon project broad.io/FAIRdatahack2019



gs://

1-Collect-1000G-participant

2-Generate-synthetic-reads

3-Mutate-reads-with-BAMSurgeon

Variant calls

Synthetic exome data

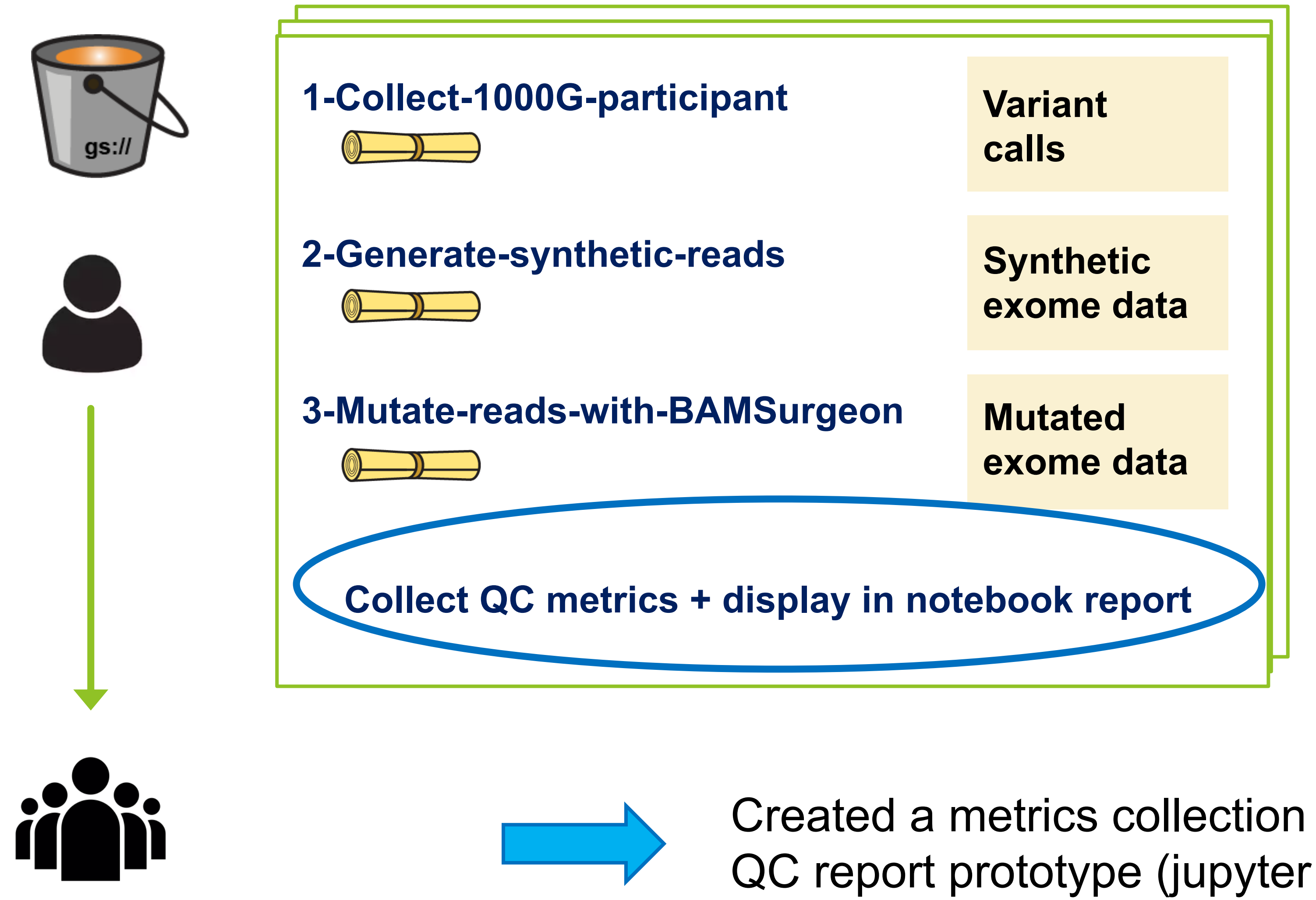
Mutated exome data

#3 – Method Optimization

Reduce cost and runtime of our workflows

Identified performance bottlenecks and started exploring use of multithreading options

This hackathon project broad.io/FAIRdatahack2019



1-Collect-1000G-participant

Variant calls

2-Generate-synthetic-reads

Synthetic exome data

3-Mutate-reads-with-BAMSurgeon

Mutated exome data

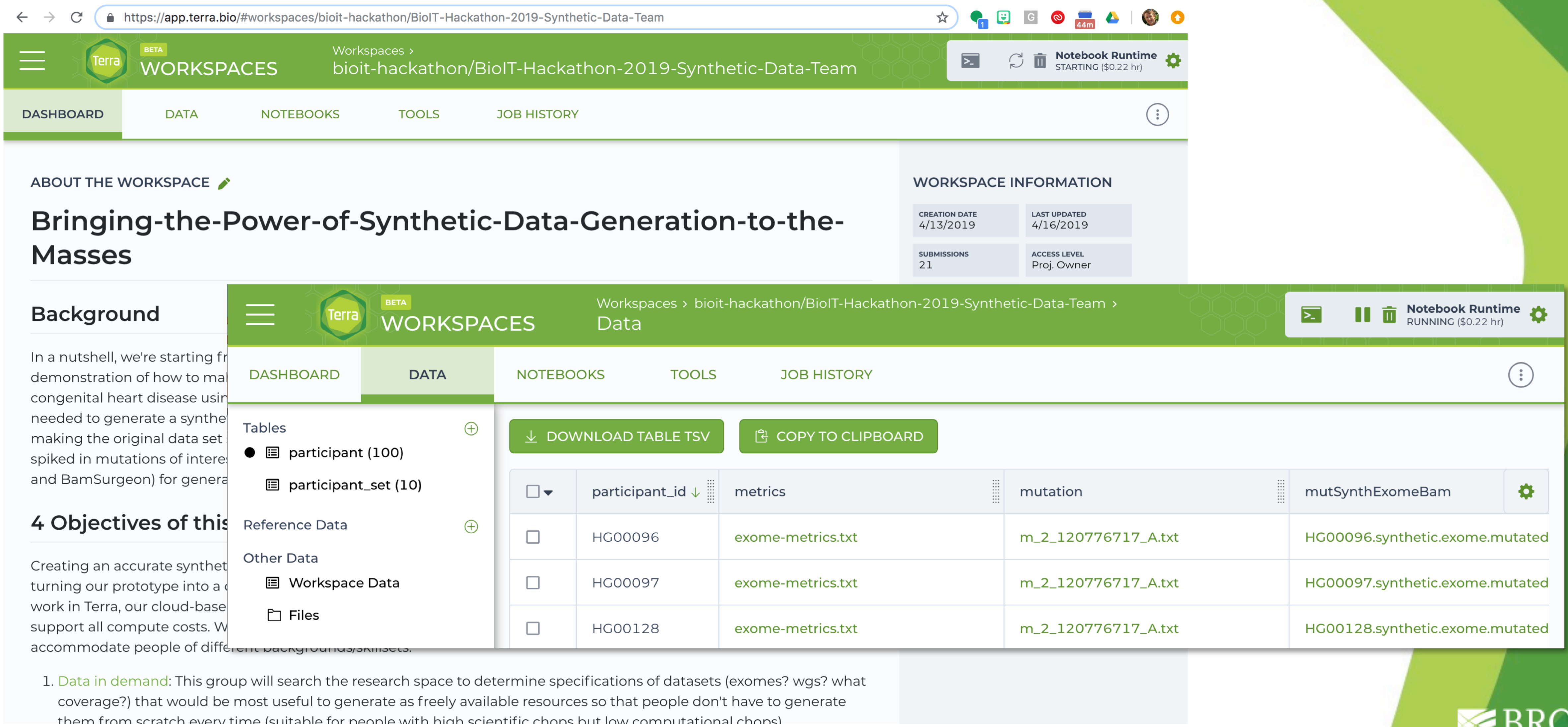
Collect QC metrics + display in notebook report

#4 – Quality Control

Evaluate whether the synthetic data we generate is suitable

Created a metrics collection workflow and a QC report prototype (jupyter notebook)

The Workspace



https://app.terra.bio/#workspaces/bioit-hackathon/BioIT-Hackathon-2019-Synthetic-Data-Team

ABOUT THE WORKSPACE

Bringing-the-Power-of-Synthetic-Data-Generation-to-the-Masses

WORKSPACE INFORMATION

CREATION DATE 4/13/2019	LAST UPDATED 4/16/2019
SUBMISSIONS 21	ACCESS LEVEL Proj. Owner

Background

In a nutshell, we're starting fr demonstration of how to ma congenital heart disease usir needed to generate a synthe making the original data set spiked in mutations of intere and BamSurgeon) for genera

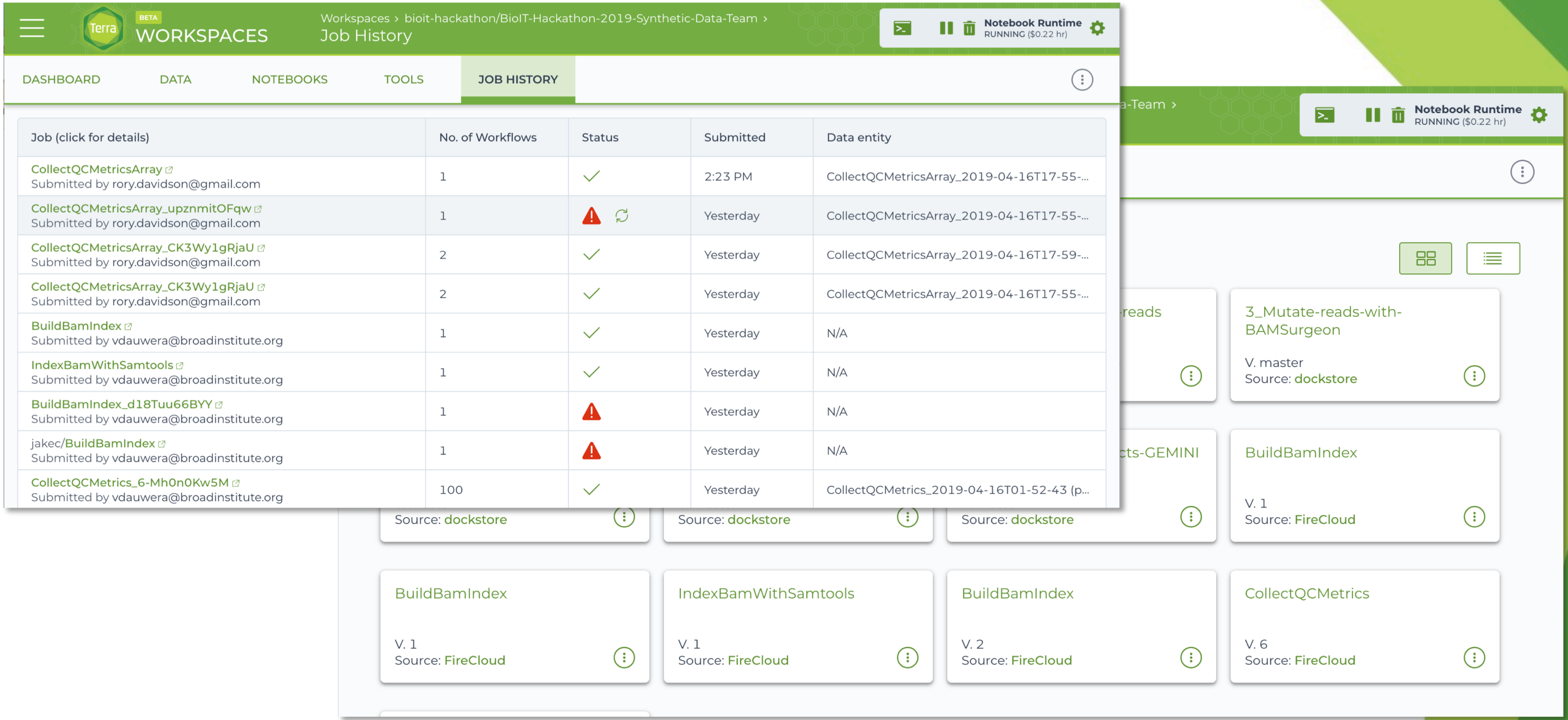
4 Objectives of this

Creating an accurate synthet turning our prototype into a work in Terra, our cloud-base support all compute costs. W accommodate people of different backgrounds/skillsets.

1. Data in demand: This group will search the research space to determine specifications of datasets (exomes? wgs? what coverage?) that would be most useful to generate as freely available resources so that people don't have to generate them from scratch every time (suitable for people with high scientific chops but low computational chops)

□	participant_id ↓	metrics	mutation	mutSynthExomeBam
□	HG00096	exome-metrics.txt	m_2_120776717_A.txt	HG00096.synthetic.exome.mutated
□	HG00097	exome-metrics.txt	m_2_120776717_A.txt	HG00097.synthetic.exome.mutated
□	HG00128	exome-metrics.txt	m_2_120776717_A.txt	HG00128.synthetic.exome.mutated

The Workspace



Job (click for details)	No. of Workflows	Status	Submitted	Data entity
CollectQCMetricsArray Submitted by rory.davidson@gmail.com	1	✓	2:23 PM	CollectQCMetricsArray_2019-04-16T17-55-...
CollectQCMetricsArray_upznmitOFqw Submitted by rory.davidson@gmail.com	1	⚠ ↻	Yesterday	CollectQCMetricsArray_2019-04-16T17-55-...
CollectQCMetricsArray_CK3Wy1gRjaU Submitted by rory.davidson@gmail.com	2	✓	Yesterday	CollectQCMetricsArray_2019-04-16T17-59-...
CollectQCMetricsArray_CK3Wy1gRjaU Submitted by rory.davidson@gmail.com	2	✓	Yesterday	CollectQCMetricsArray_2019-04-16T17-55-...
BuildBamIndex Submitted by vdauwera@broadinstitute.org	1	✓	Yesterday	N/A
IndexBamWithSamtools Submitted by vdauwera@broadinstitute.org	1	✓	Yesterday	N/A
BuildBamIndex_d18Tuu66BYY Submitted by vdauwera@broadinstitute.org	1	⚠	Yesterday	N/A
jakec/BuildBamIndex Submitted by vdauwera@broadinstitute.org	1	⚠	Yesterday	N/A
CollectQCMetrics_6-Mh0n0Kw5M Submitted by vdauwera@broadinstitute.org	100	✓	Yesterday	CollectQCMetrics_2019-04-16T01-52-43 (p...

reads

3_Mutate-reads-with-BAMSurgeon
V. master
Source: dockstore

cts-GEMINI

BuildBamIndex
V. 1
Source: FireCloud

Source: dockstore

Source: dockstore

Source: dockstore

BuildBamIndex
V. 1
Source: FireCloud

IndexBamWithSamtools
V. 1
Source: FireCloud

BuildBamIndex
V. 2
Source: FireCloud

CollectQCMetrics
V. 6
Source: FireCloud

The Next Steps

Off-the-shelf synthetic data catalog

User-friendly tools for generating custom synthetic (sequence) datasets

The Team

Broadies

Adelaide Rhodes
Allie Hajian
Anton Kovalsky
Ruchi Munshi
Tiffany Miller
Geraldine Van der Auwera

Guest stars

Ernesto Andrianantoandro
Dan Rozelle
Jay Moore
Rory Davidson
Roma Kurilov
Vrinda Pareek