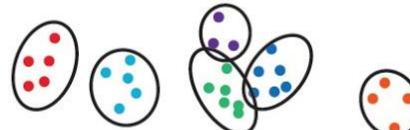
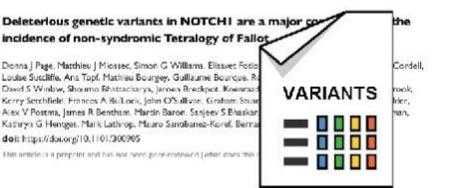
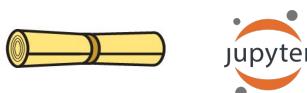


# **Bringing the power of synthetic data generation to the masses**

NCBI FAIR hackathon – April 15-16



DATA INPUTS	PROCESSING	ANALYSIS	SHARING
 <b>Exome Sequencing</b> <ul style="list-style-type: none"><li>• 829 ToF patients (excl. carriers of known deletion)</li><li>• 1252 healthy controls</li><li>• Agilent SureSelectXT v4</li><li>• Illumina HiSeq2000</li></ul>	 <b>Mapping and Variant Discovery</b> <ul style="list-style-type: none"><li>• MUGQIC GenPipes DNaseq including:<ul style="list-style-type: none"><li>• Trimmomatic</li><li>• BWA 0.6.2 (b37/hg19)</li><li>• GATK 3.2 HaplotypeCaller</li><li>• QS (QUAL) &gt; 100</li></ul></li></ul>	 <b>Effect Prediction and Clustering Analysis</b> <ul style="list-style-type: none"><li>• SnpEff + Gemini</li><li>• OMIM, GERP, ICGV, ExAC</li><li>• MAF &lt;= 0.001 in ExAC</li><li>• CADD &gt;= 20</li><li>• 20 W<sub>d</sub> statistic and test</li></ul>	 <b>Preprint in biorXiv</b> <ul style="list-style-type: none"><li>• Summary of methods</li><li>• Table of 49 NOTCH1 variants</li><li>• Pers. communication with author to translate bash and Perl scripts</li></ul>
<b>Generated synthetic data based on public 1000G Project data</b> <b>Created case samples by spiking in the NOTCH1 variants</b> <b>Joint variant discovery on cases + controls with GATK4</b>	<b>Translated scripts to WDL &amp; R</b> <b>Same analysis commands with same data resources</b>	<b>Public Terra Featured Workspace</b> <b>Bundles all data, workflows, Jupyter notebook and results</b>	





### 1. DATA IN DEMAND

data + resources

What kind of datasets would be useful to the community?

## ANALYSIS

### 2. METHOD OPTIMIZATION

general, automated

### 3. DIVERSE OPTIONS

specific, manual

### 4. QUALITY CONTROL

specific, manual

Reduce cost and runtime of our portable workflows

Enable more data types and more variant types

Evaluate whether synthetic data we generate is suitable

## RESULT

A few questions we hope to answer.

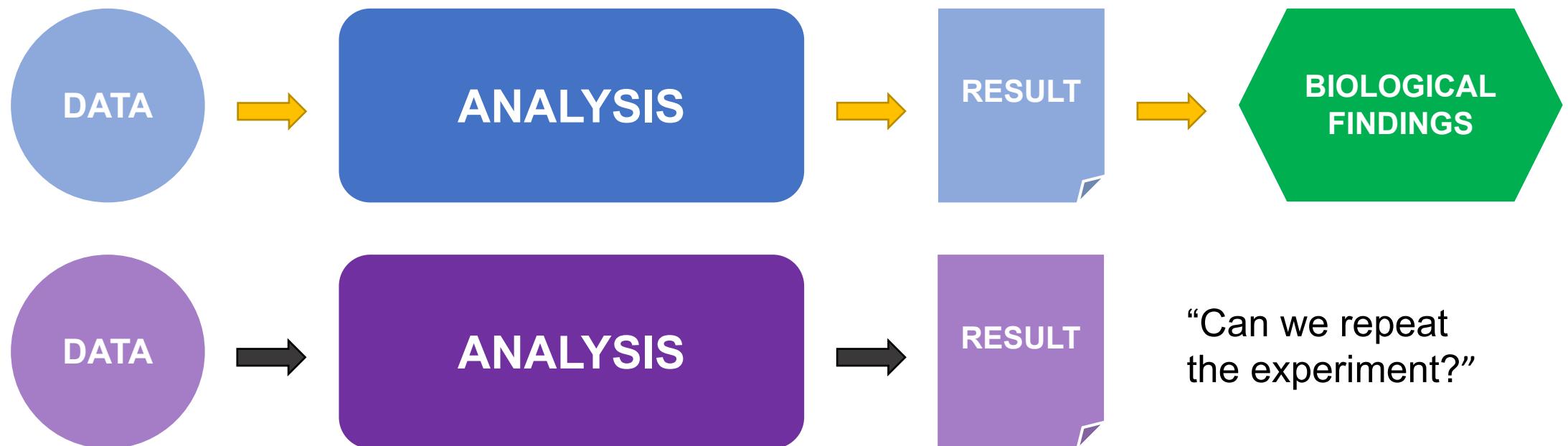
"Can we confirm that this is true?"

"Is it useful for more than one case?"

"Can we repeat the experiment?"

# What does it mean to reproduce an analysis?

Reproduce ≠  
Replicate



# Case study

## **Deleterious genetic variants in NOTCH1 are a major contributor to the incidence of non-syndromic Tetralogy of Fallot**

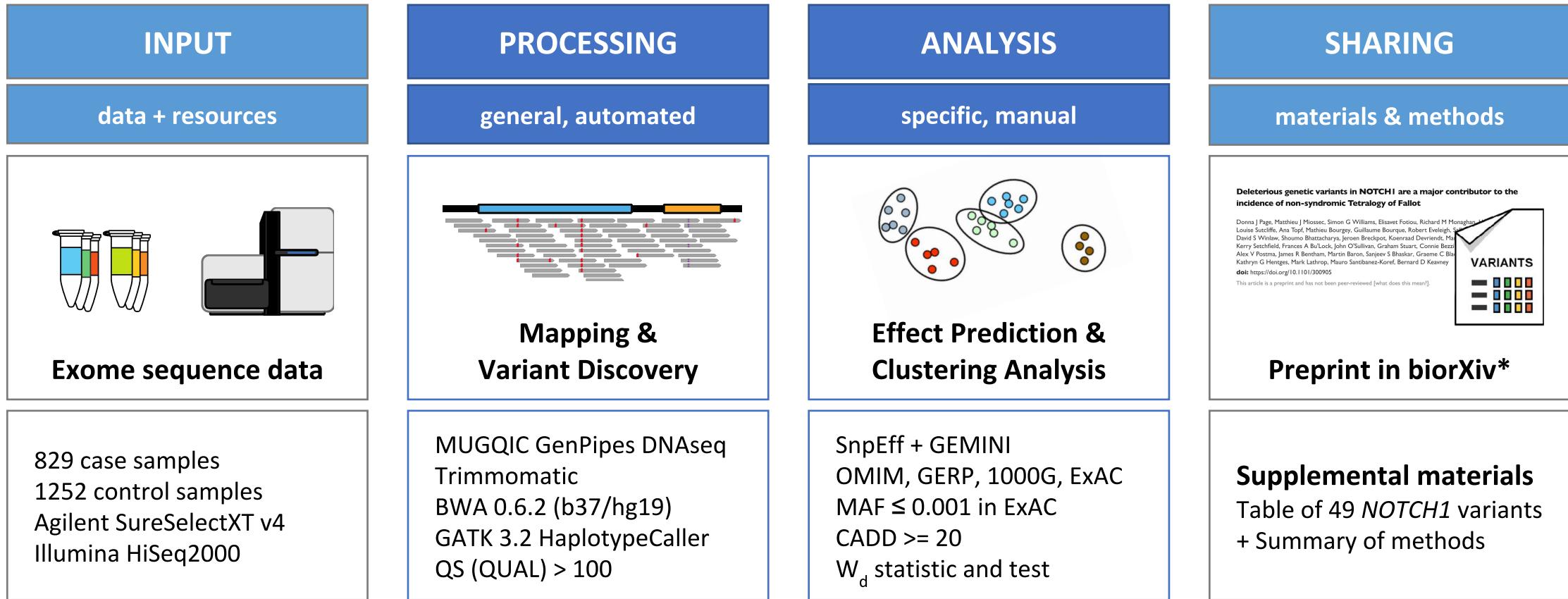
Donna J Page, Matthieu J Miossec, Simon G Williams, Elisavet Fotiou, Richard M Monaghan, Heather J Cordell, Louise Sutcliffe, Ana Topf, Mathieu Bourgey, Guillaume Bourque, Robert Eveleigh, Sally L Dunwoodie, David S Winlaw, Shoumo Bhattacharya, Jeroen Breckpot, Koenraad Devriendt, Marc Gewillig, David Brook, Kerry Setchfield, Frances A Bu'Lock, John O'Sullivan, Graham Stuart, Connie Bezzina, Barbara J.M. Mulder, Alex V Postma, James R Bentham, Martin Baron, Sanjeev S Bhaskar, Graeme C Black, William G Newman, Kathryn G Hentges, Mark Lathrop, Mauro Santibanez-Koref, Bernard D Keavney

**doi:** <https://doi.org/10.1101/300905>

This article is a preprint and has not been peer-reviewed [what does this mean?].

<https://www.biorxiv.org/content/early/2018/04/13/300905>

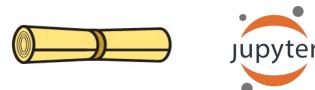
# What's in the box? Reconstructing the analysis based on the preprint



\* Page, Miossec *et al.*, 2018. **Deleterious genetic variants in NOTCH1 are a major contributor to the incidence of non-syndromic Tetralogy of Fallot**  
<https://www.biorxiv.org/content/early/2018/04/13/300905>

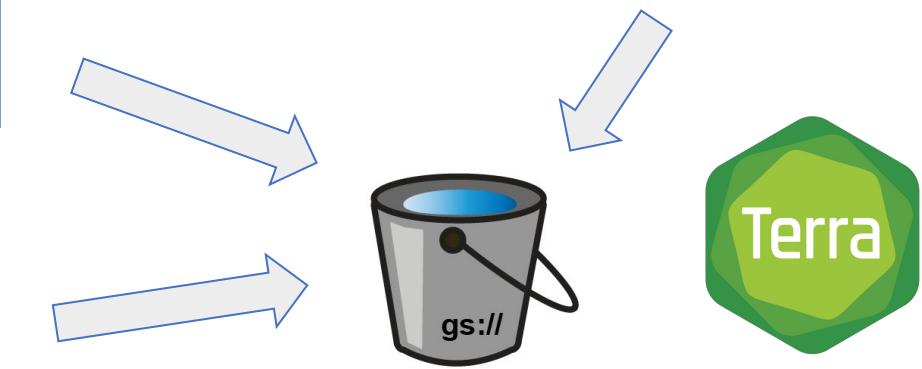
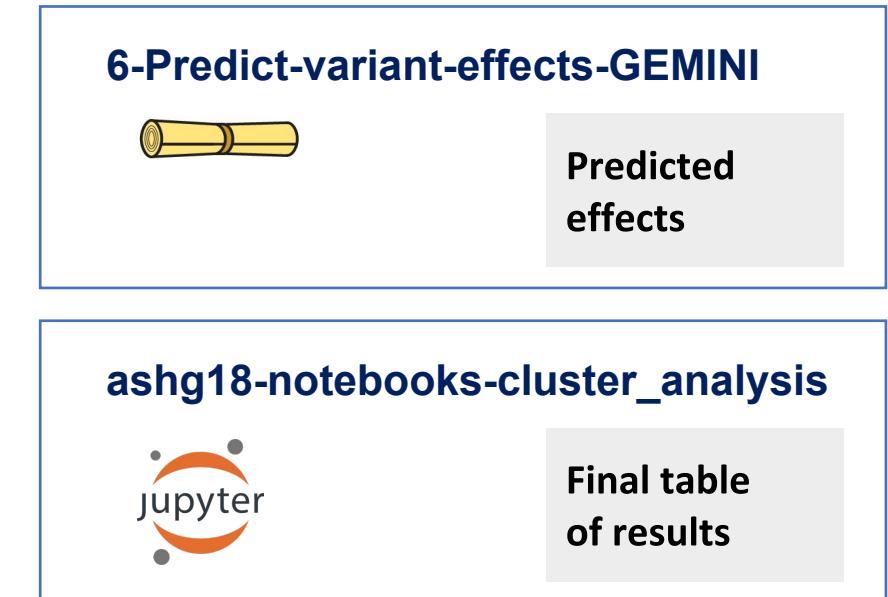
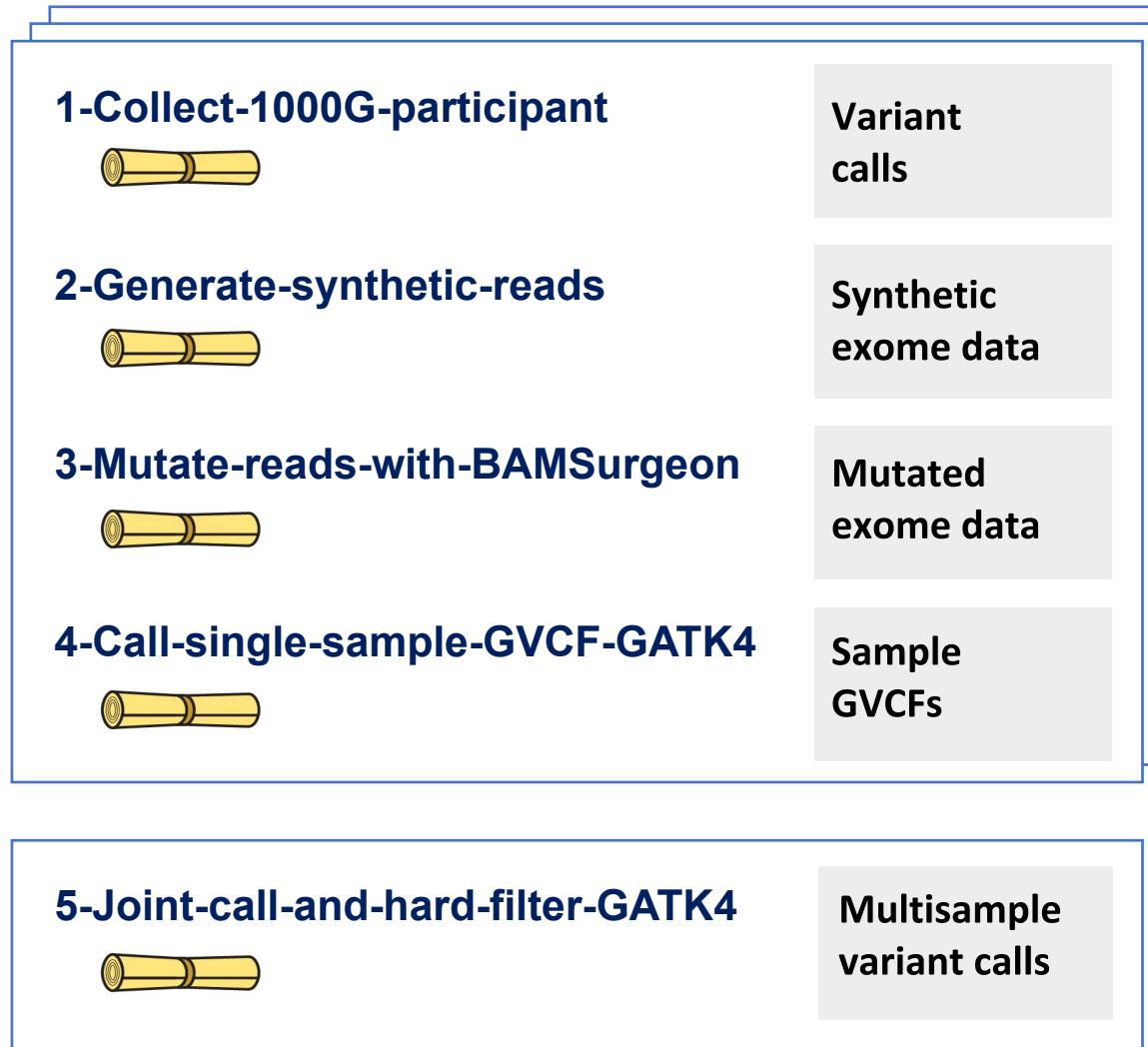
# Our approach to reproducing the work

INPUT	PROCESSING	ANALYSIS	SHARING
data + resources	general, automated	specific, manual	materials & methods
Exome sequence data	Mapping & Variant Discovery	Effect Prediction & Clustering Analysis	Preprint in BioarXiv*
829 case samples 1252 control samples Agilent SureSelectXT v4 Illumina HiSeq2000	MUGQIC GenPipes DNaseq Trimmomatic BWA 0.6.2 (b37/hg19) GATK 3.2 HaplotypeCaller QS (QUAL) > 100	SnpEff + Gemini OMIM, GERP, 1000G, ExAC MAF ≤ 0.001 in ExAC CADD ≥ 20 $W_d$ statistic and test	Methods summary + list of 49 <i>NOTCH1</i> variants in supplemental materials
<p><b>Generated synthetic data based on 1000G</b> <b>Cases created by spiking in the <i>NOTCH1</i> variants</b> <b>Joint variant discovery on cases + controls with GATK4</b></p>			<p><b>Terra workspace</b> <b>Contains all data, workflows and notebook</b></p>



\* Page, Miossec *et al.*, 2018. **Deleterious genetic variants in NOTCH1 are a major contributor to the incidence of non-syndromic Tetralogy of Fallot**  
<https://www.biorxiv.org/content/early/2018/04/13/300905>

# 6 workflows + 1 notebook in a Terra workspace



<gs://firecloud-workshops/181017-ashg18>

# Project objectives

- 1. Data in demand

What kind of datasets would be useful to the community?

- 2. Diversifying Options

Enable more data types and more variant types

- 3. Method optimization:

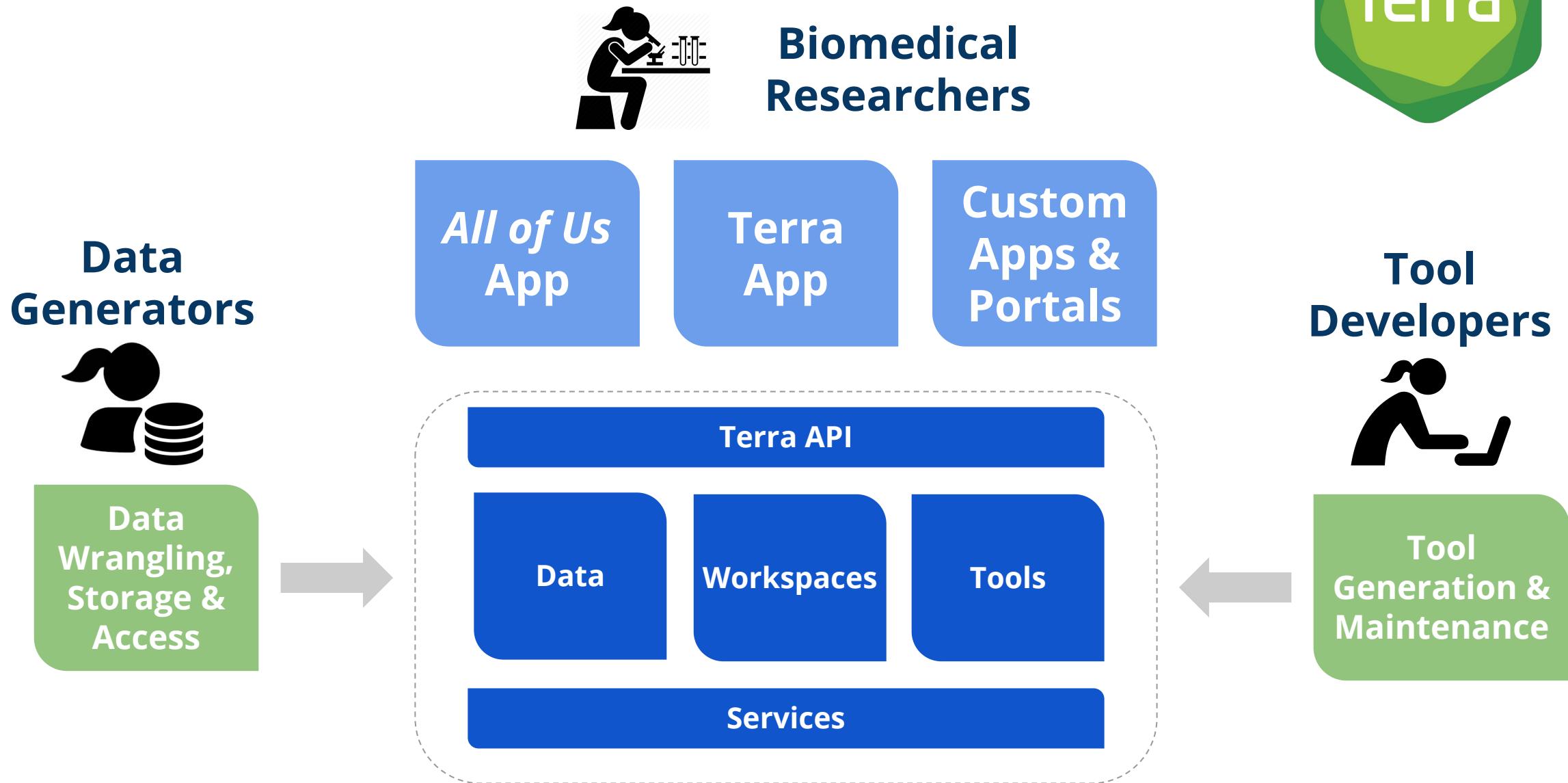
Reduce cost and runtime of our workflows

- 4. Quality control

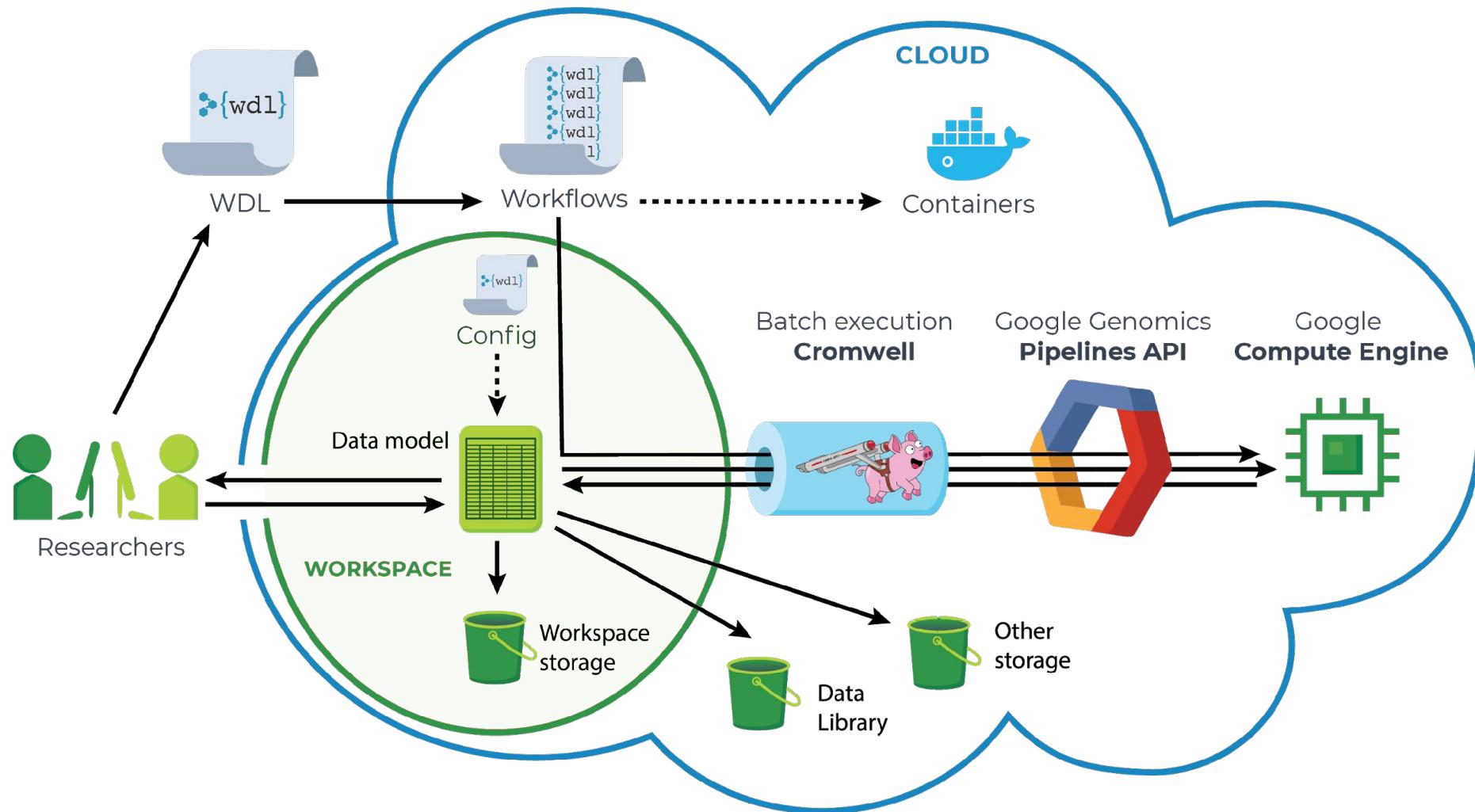
Evaluate that the synthetic data we generate is suitable

**All computational work can be done on Terra**

# Terra is a platform for Data Access & Analysis



# Workflow execution in Terra



# Interactive analysis with Notebooks in Terra

The screenshot shows a Jupyter Notebook interface within a Terra workspace. The top navigation bar includes 'WORKSPACES' (BETA), 'Notebooks - cluster\_analysis.ipynb', and 'Notebook Runtime RUNNING (\$0.22 hr)'. The notebook title is 'cluster\_analysis' with a last checkpoint on '03/12/2019' (autosaved). The toolbar below the title includes standard Jupyter controls like File, Edit, View, Insert, Cell, Kernel, Navigate, Widgets, Help, and a 'Not Trusted' status.

The notebook content consists of three code cells:

- In [9]:**

```
variants_per_gene<-aggregate(x = gemini_filtered$gene, by = list(gene = gemini_filtered$gene), FUN = length)
colnames(variants_per_gene) <- c("gene", "variant_count")
```
- In [10]:**

```
variants_per_CDS<-rbind.data.frame(variants_per_gene[variants_per_gene$gene %in% CDS_size$gene,], data.frame(gene = set
```
- In [11]:**

```
sorted<-variants_per_CDS[order(as.character(variants_per_CDS$gene)),]
variants_per_CDS_withsizes<-data.frame(sorted,length_proportion=CDS_size$length_proportion)
variants_per_CDS_withsizes
```

The output of the third cell is a table:

gene	variant_count	length_proportion	
13507	A1BG	0	0.000033100
21000	A1CF	0	0.000132969
1	A2M	3	0.000069400
2	A2ML1	1	0.000075600
31000	A3GALT2	0	0.000014700
3	A4GALT	1	0.000033500
41000	A4GNT	0	0.000025500

# Example analysis and tool workspaces: fully loaded out of the box

← → C https://app.terra.bio/#library/showcase

**Terra** **BETA** **LIBRARY**

DATASETS SHOWCASE & TUTORIALS CODE & TOOLS

### GATK4 example workspaces

**Germline-SNPs-Indels-GATK4-hg38**  
### GATK Best Practices for Germline SNPs & Indels  
  
The purpose of this workspace is to provide a fully reproducible example of

**Somatic-CNVs-GATK4**  
### GATK Best Practices for Somatic CNV Discovery  
  
The purpose of this workspace is to provide a fully reproducible example of

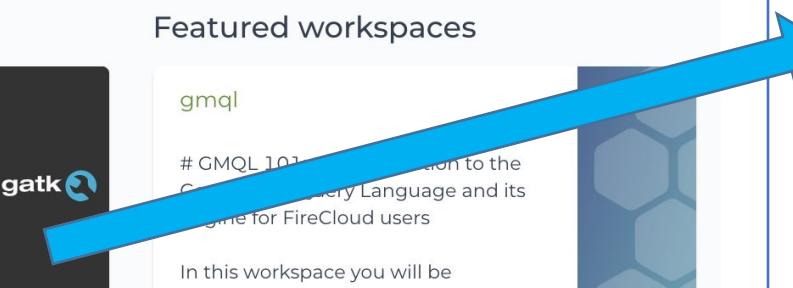
**Somatic-SNVs-Indels-GATK4**  
### GATK Best Practices for Single Tumor-Normal Pair or Single Tumor Sample  
  
The purpose of this workspace is to

### Featured workspaces

**gmql**  
# GMQL 101: An Introduction to the GMQL Query Language and its Use for FireCloud users  
  
In this workspace you will be

**Reproducibility\_Case\_Study\_Tetralogy\_of\_Fallot**  
## Reproducing the paper: Variant analysis of Tetralogy of Fallot  
  
### Overview

**Seq-Format-Conversion**  
### Methods for format conversion  
  
The purpose of this workspace is to provide users with example wdl's for converting their sequence data. The



← → C https://app.terra.bio/#workspaces/help-gatk/Germline-SNPs-Indels-GATK4-hg38/tools/gatk/3

**Terra** **BETA** **WORKSPACES**

Workspaces > help-gatk/Germline-SNPs-Indels-GATK4-hg38 > tools > 3-Joint-Discovery

### 3-Joint-Discovery

Snapshot 1.1.0  
Source: [github.com/gatk-workflows/gatk4-germline-snps-indels/joint-discovery-gatk4](https://github.com/gatk-workflows/gatk4-germline-snps-indels/joint-discovery-gatk4)  
Synopsis:  
*No documentation provided*

Process single workflow from files  
 Process multiple workflows from: **Sample Set** **Select Data**  
0 selected sample\_set

**SCRIPT** -> **INPUTS** -> **OUTPUTS** -> **RUN ANALYSIS**

Show optional inputs

Task name	Variable	Type	Attributed
JointGenotyping	input_gvcfs_indices	Array[File]	
JointGenotyping	indel_recalibration_annotation_values	Array[String]	
JointGenotyping	snp_recalibration_tranche_values	Array[String]	
JointGenotyping	omni_resource_vcf_index	File	
JointGenotyping	eval_interval_list	File	
JointGenotyping	one_thousand_genomes_resource_vcf_index	File	
JointGenotyping	one_thousand_genomes_resource_vcf	File	