Bringing the power of synthetic data generation to the masses

NCBI FAIR hackathon – April 15-16

Terra

BROAD INSTITUTE

DATA INPUTS

Exome Sequencing

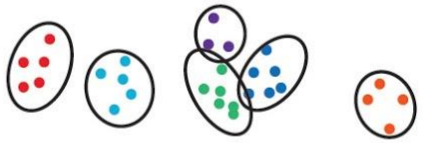
- 829 ToF patients (excl. carriers of known deletion)
- 1252 healthy controls
- Agilent SureSelectXT v4
- Illumina HiSeq2000

PROCESSING

Mapping and Variant Discovery

- MUGQIC GenPipes DNAseq including:
- Trimmomatic
- BWA 0.6.2 (b37/hg19)
- GATK 3.2 HaplotypeCaller
- QS (QUAL) > 100

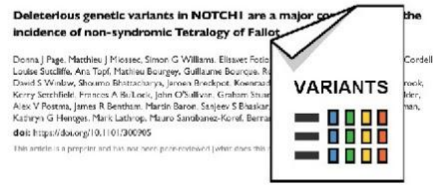
ANALYSIS



Effect Prediction and Clustering Analysis

- SnpEff + Gemini
- OMIM, GERP, IOOOG, ExAC
- MAF <= 0.001 in ExAC
- CADD >= 20
- 20 W_d statistic and test

SHARING



Preprint in biorXiv

- Summary of methods
- Table of 49 NOTCH1 variants
- Pers. communication with author to translate bash and Perl scripts

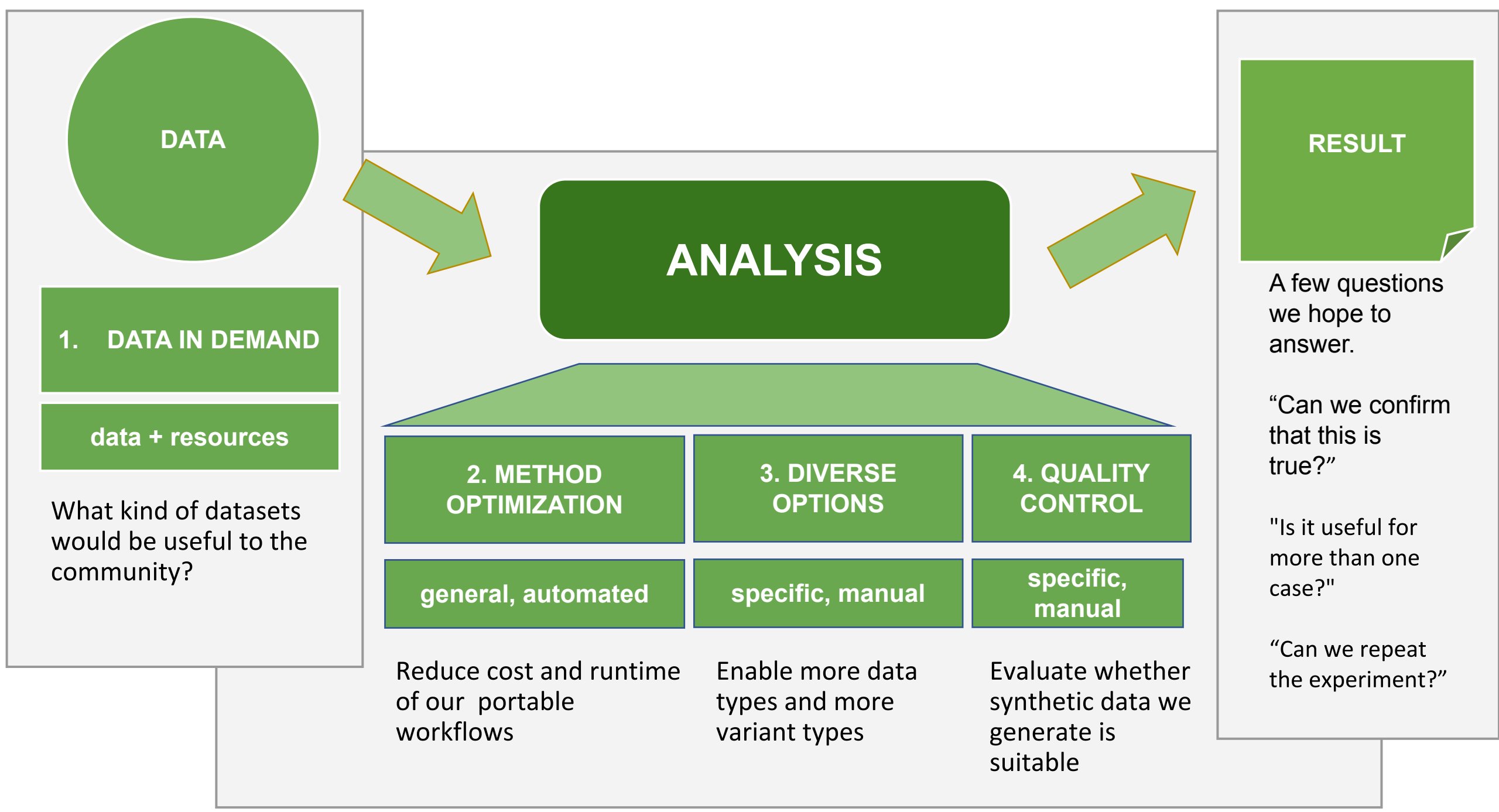
Generated synthetic data based on public 1000G Project data
Created case samples by spiking in the NOTCH1 variants
Joint variant discovery on cases + controls with GATK4

Translated scripts to WDL & R
Same analysis commands with same data resources

Public Terra Featured Workspace
Bundles all data, workflows, Jupyter notebook and results

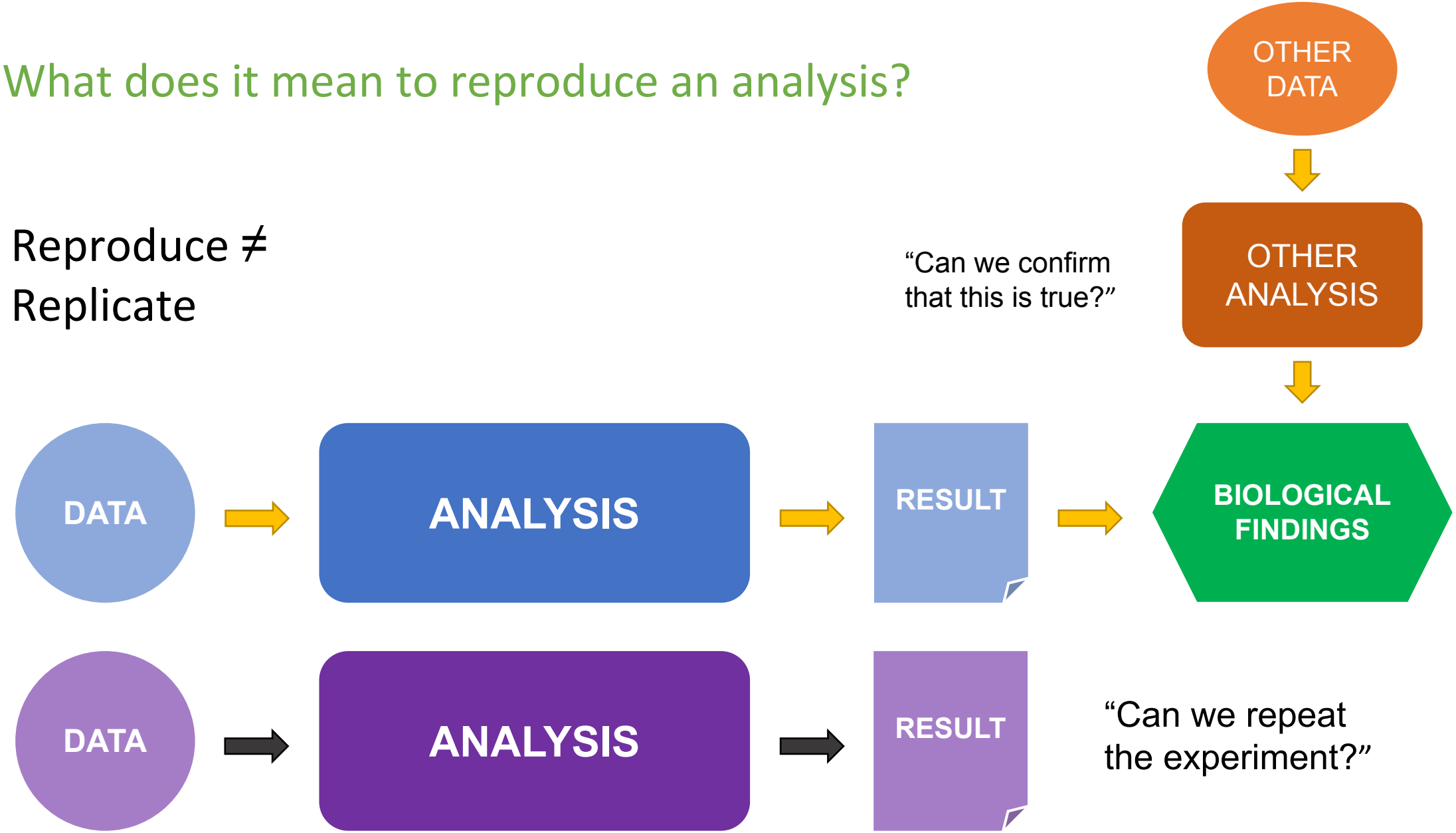
jupyter

Terra


DATA
1. DATA IN DEMAND
data + resources
What kind of datasets would be useful to the community?
ANALYSIS
2. METHOD OPTIMIZATION
general, automated
Reduce cost and runtime of our portable workflows
3. DIVERSE OPTIONS
specific, manual
Enable more data types and more variant types
4. QUALITY CONTROL
specific, manual
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

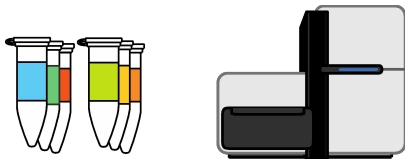
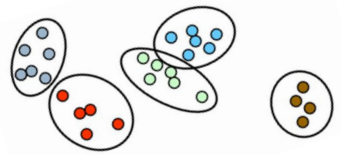
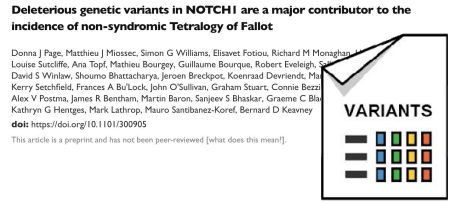
Deleterious genetic variants in NOTCH1 are a major contributor to the incidence of non-syndromic Tetralogy of Fallot

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

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

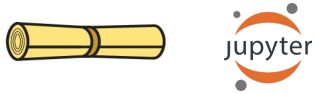
What's in the box? Reconstructing the analysis based on the preprint

INPUT	PROCESSING	ANALYSIS	SHARING
data + resources	general, automated	specific, manual	materials & methods
 **Exome sequence data**	 **Mapping & Variant Discovery**	 **Effect Prediction & Clustering Analysis**	 **Preprint in biorXiv***
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	**Supplemental materials** Table of 49 *NOTCH1* variants + Summary of methods

* 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 *NOTCH1* variants in supplemental materials
Generated synthetic data based on 1000G Cases created by spiking in the *NOTCH1* variants Joint variant discovery on cases + controls with GATK4		**Translated scripts Same commands with same data resources**	**Terra workspace Contains all data, workflows and notebook**



* 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

1-Collect-1000G-participant → Variant calls

2-Generate-synthetic-reads → Synthetic exome data

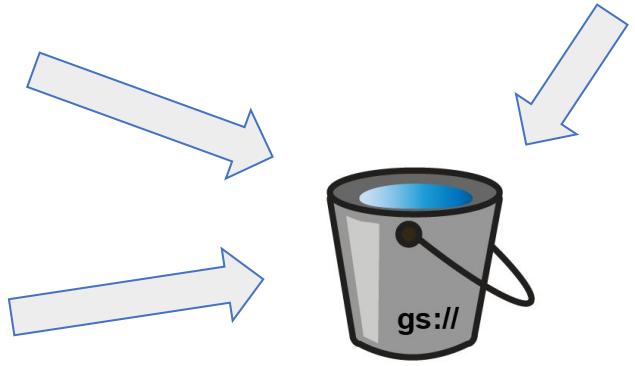
3-Mutate-reads-with-BAMSurgeon → Mutated exome data

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 (jupyter) → Final table of results



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

Biomedical Researchers

All of Us App

Terra App

Custom Apps & Portals

Data Generators

Data Wrangling, Storage & Access

Terra API

Data

Workspaces

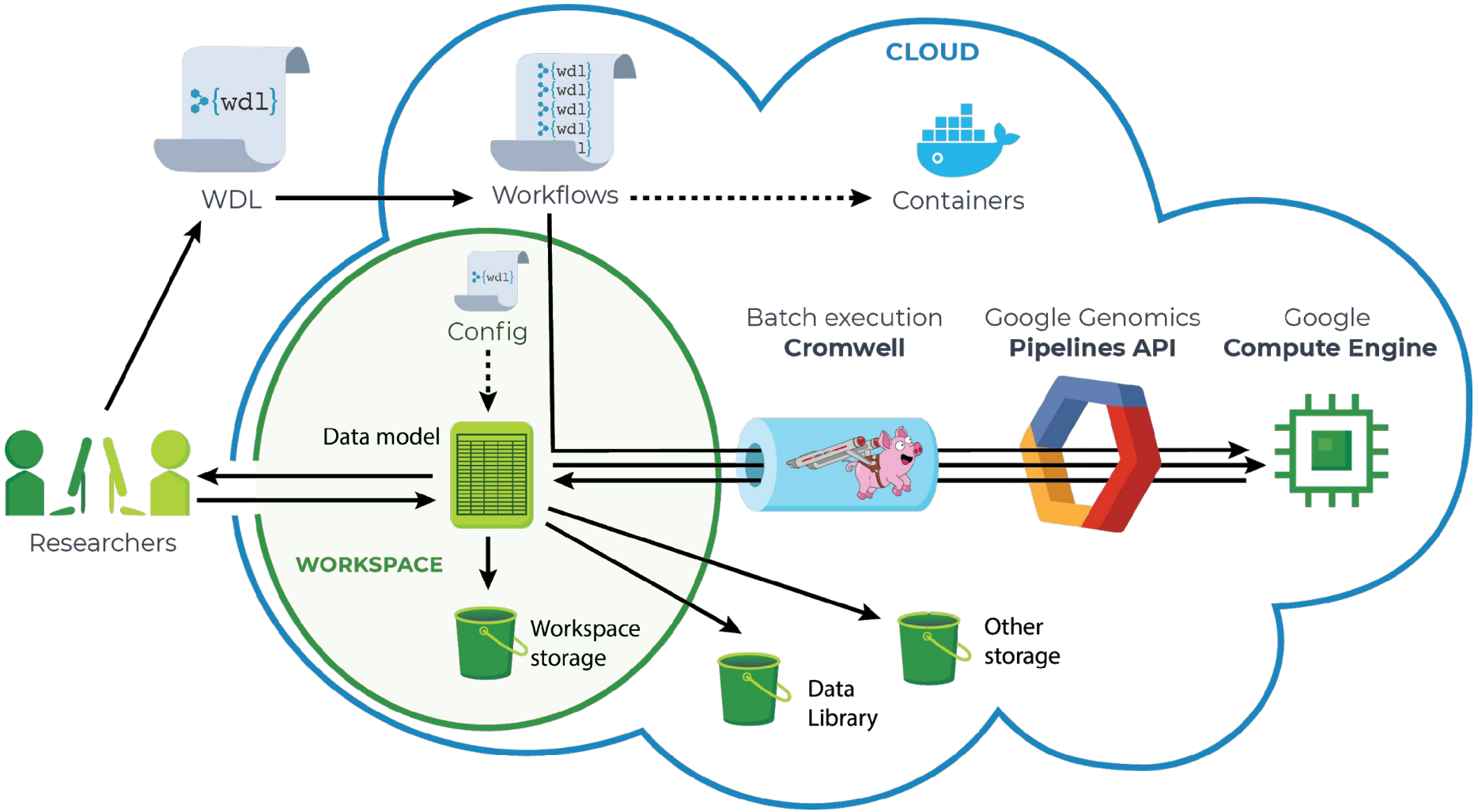
Tools

Services

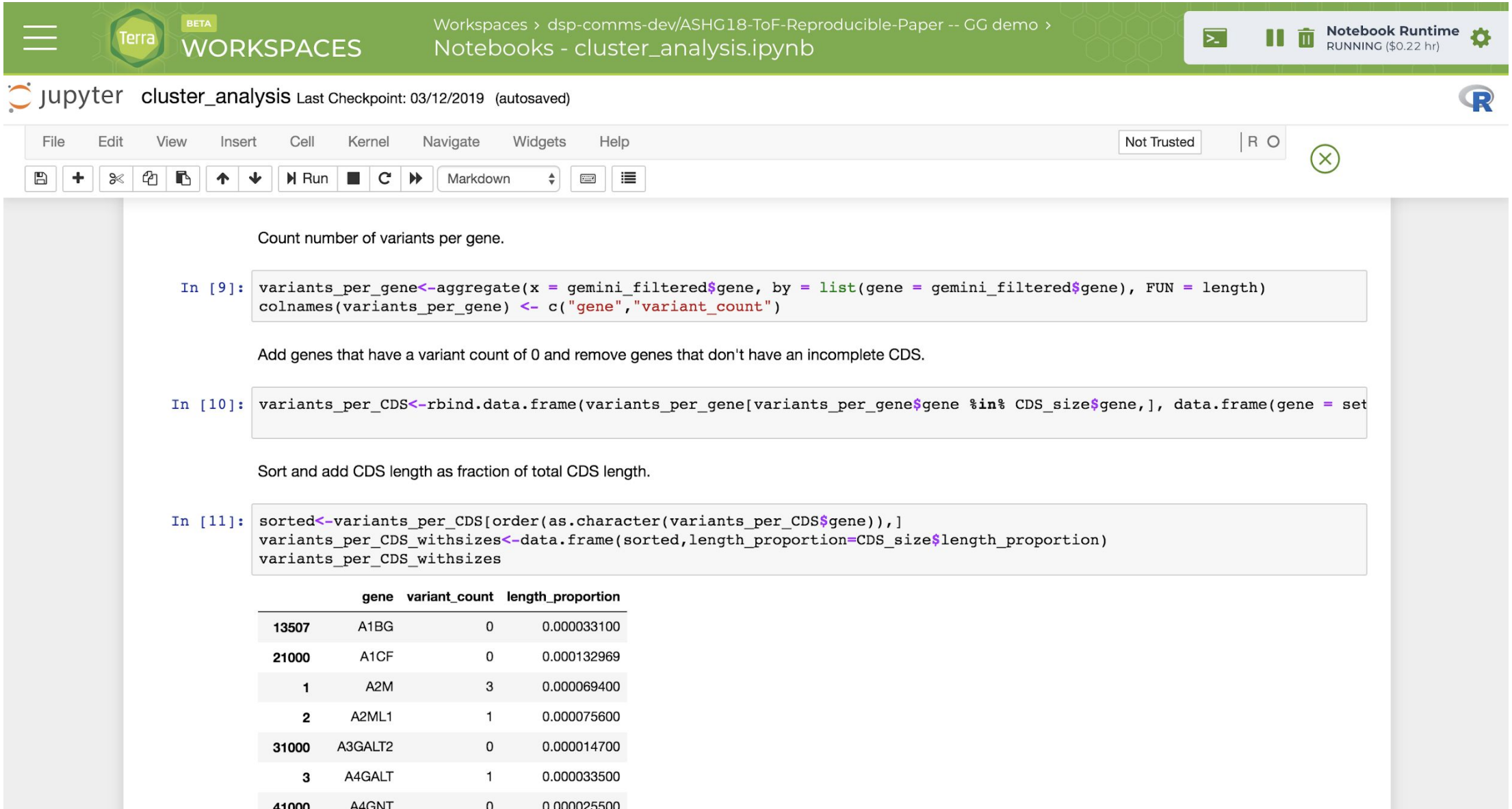
Tool Developers

Tool Generation & Maintenance

Workflow execution in Terra



Interactive analysis with Notebooks in Terra



	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