

# Bioconductor On The Cloud

EuroBioc2020

December 14<sup>th</sup>, 2020

Sehyun Oh, PhD

[Sehyun.Oh@sph.cuny.edu](mailto:Sehyun.Oh@sph.cuny.edu)  
<https://github.com/shbrief>

# Why Genomics in the Cloud?



COMPUTATIONAL TOOLS

## Cloud computing for genomic data analysis and collaboration

Ben Langmead<sup>1</sup> and Abhinav Nellore<sup>2</sup>

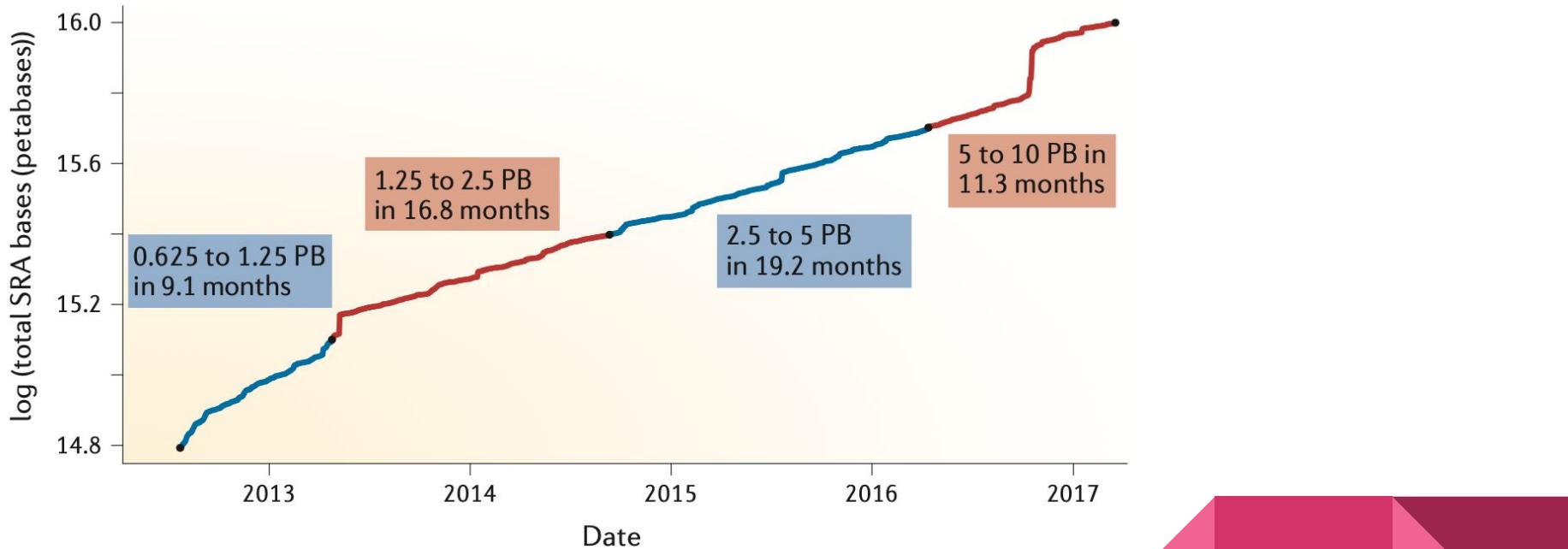
Abstract | Next-generation sequencing (NGS) data have been analyzed using large-scale computers and cloud research. Here,

***... its elasticity, reproducibility, and privacy features make it ideally suited ...***

large-scale collaborations, and argue that its elasticity, reproducibility and privacy features make it ideally suited for the large-scale reanalysis of publicly available archived data, including privacy-protected data.

studies based on large sequencing data have been used to enable researchers to perform analyses in genomics and

# *Doubling every 18 months...*



# 30TB

Approximate amount of public sequence data received and processed  
*daily* by the NCBI Sequence Read Archive (SRA).

# 87GB

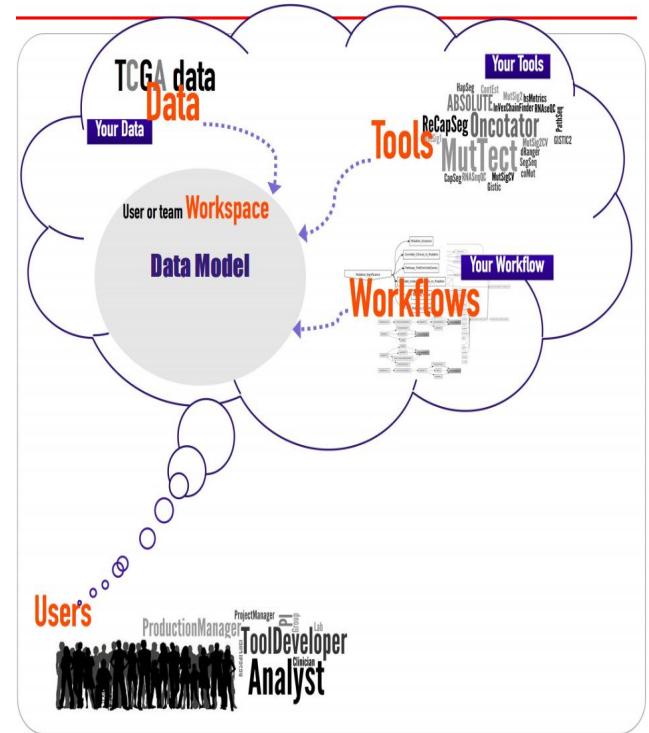
Amount of storage necessary for a ***single*** whole genome (40x coverage),  
requiring about 11 minutes of data transfer on a dedicated 1Gb/second  
network.

Table 1 | Large genomics projects and resources

Name	Website	Description
1000 Genomes Project (1KGP) <sup>102</sup>	<a href="http://www.internationalgenome.org">www.internationalgenome.org</a>	This project includes whole-genome and exome sequencing data from 2,504 individuals across 26 populations
Cancer Cell Line Encyclopedia (CCLE) <sup>115</sup>	<a href="http://portals.broadinstitute.org/ccle">portals.broadinstitute.org/ccle</a>	This resource includes data spanning 1,457 cancer cell lines
Encyclopedia of DNA Elements (ENCODE) <sup>33</sup>	<a href="http://www.encodeproject.org">www.encodeproject.org</a>	The goal of this project is to identify functional elements of the human genome using a gamut of sequencing assays across cell lines and tissues
Genome Aggregation Database (gnomAD) <sup>13</sup>	<a href="http://gnomad.broadinstitute.org">gnomad.broadinstitute.org</a>	This resource entails coverage and allele frequency information from over 120,000 exomes and 15,000 whole genomes
Genotype–Tissue Expression (GTEx) Portal <sup>15,16</sup>	<a href="http://gtexportal.org">gtexportal.org</a>	This effort has to date performed RNA sequencing or genotyping of 714 individuals across 53 tissues
Global Alliance for Genomics and Health (GA4GH) <sup>92</sup>	<a href="http://genomicsandhealth.org">genomicsandhealth.org</a>	This consortium of over 400 institutions aims to standardize secure sharing of genomic and clinical data
International Cancer Genome Consortium (ICGC) <sup>14</sup>	<a href="http://icgc.org">icgc.org</a>	This consortium spans 76 projects, including TCGA
Million Veterans Program (MVP) <sup>19</sup>	<a href="http://www.research.va.gov/mvp">www.research.va.gov/mvp</a>	This US programme aims to collect blood samples and health information from 1 million military veterans
Model Organism Encyclopedia of DNA Elements (modENCODE) <sup>25,85</sup>	<a href="http://www.modencode.org">www.modencode.org</a>	The goal of this effort is to identify functional elements of the <i>Drosophila melanogaster</i> and <i>Caenorhabditis elegans</i> genomes using a gamut of sequencing assays
Precision Medicine Initiative (PMI) <sup>18</sup>	<a href="http://allofus.nih.gov">allofus.nih.gov</a>	This US programme aims to collect genetic data from over 1 million individuals
The Cancer Genome Atlas (TCGA) <sup>116</sup>	<a href="http://cancergenome.nih.gov">cancergenome.nih.gov</a>	This resource includes data from 11,350 individuals spanning 33 cancer types
Trans-Omics for Precision Medicine (TOPMed) <sup>17</sup>	<a href="https://www.ncbiwg.org">https://www.ncbiwg.org</a>	The goal of this programme is to build a commons with omics data and associated clinical outcomes data across populations for research on heart, lung, blood and sleep disorders

# Cloud computing advantages

- Scalability for both storage/data and compute resources
- Share and collaborate securely
- Reproducibility comes with shared infrastructure and code
- Reusability
- Democratize data access and, potentially, analysis



# Cloud-based genomics platforms

Cloud-based genomics platform is one of the promising solutions for rapidly growing size of sequencing data and many platforms already exist hosting different dataset and analysis tools. Below is the brief example of a few:

Platform	Hosted Data	Analysis Tools
Terra	CCDG, eMERGE, TCGA, TARGET, TOPMed, etc.	WDL, Notebook, RStudio, Galaxy
Seven Bridges	TOPMed	CWL, Notebook, RStudio
Seven Bridges	TCGA, TARGET, ICGC, etc.	CWL, Notebook, RStudio
ISB-CGC	TCGA, TARGET, etc.	GCP tools (e.g. Google BigQuery)

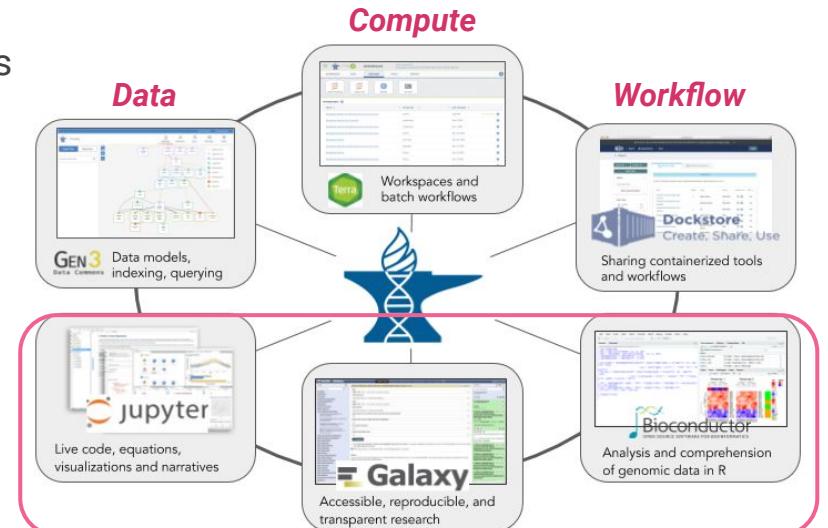
# Contents

1. Overview of Terra platform
2. Setup 'classroom' in Terra (*for teaching*)
3. Bioinformatics analysis on Terra UI (*for wet-lab scientists*)
4. Share your published work through Terra (*for bioinformaticians*)
5. Workflow package powered by Terra (*for developers*)

# What is Terra?

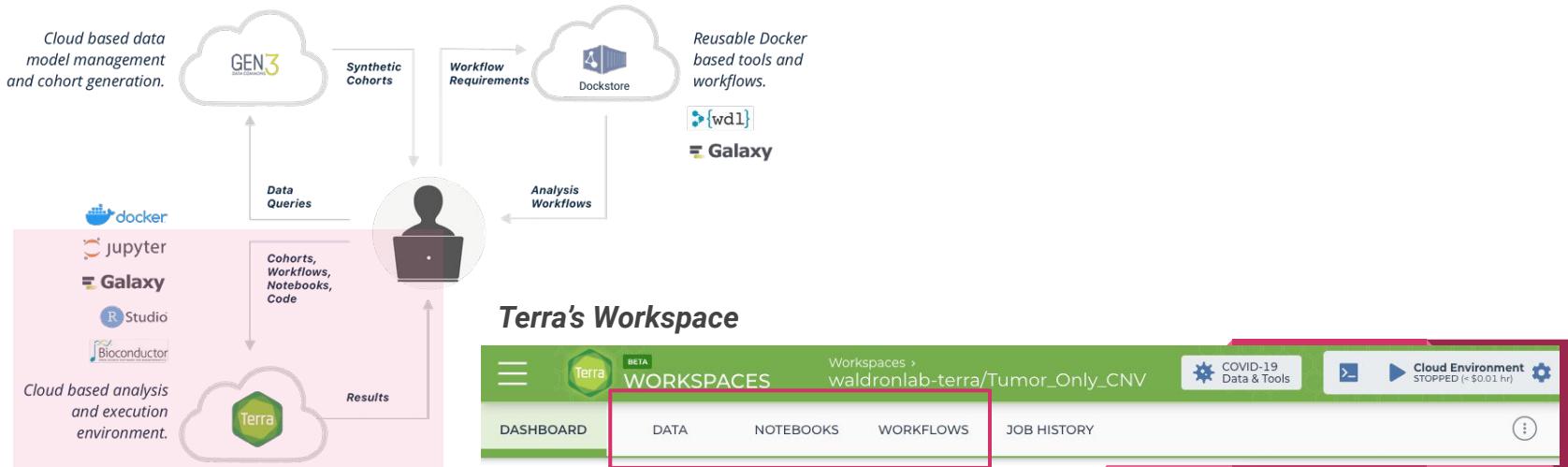
# AnVIL (Analysis, Visualization, and Informatics Lab-Space)

- Problems of the traditional model of genomic data sharing, which is centralized data warehouse such as dbGap from which researchers download data to analyze locally :
  - transfer/download cost
  - long transfer time
  - redundant compute infrastructure
  - security of protected data
- NHGRI's AnVIL provides a unified environment for data management and compute.
  - no need for data movement
  - better security handling
  - provide elastic, shared computing resources



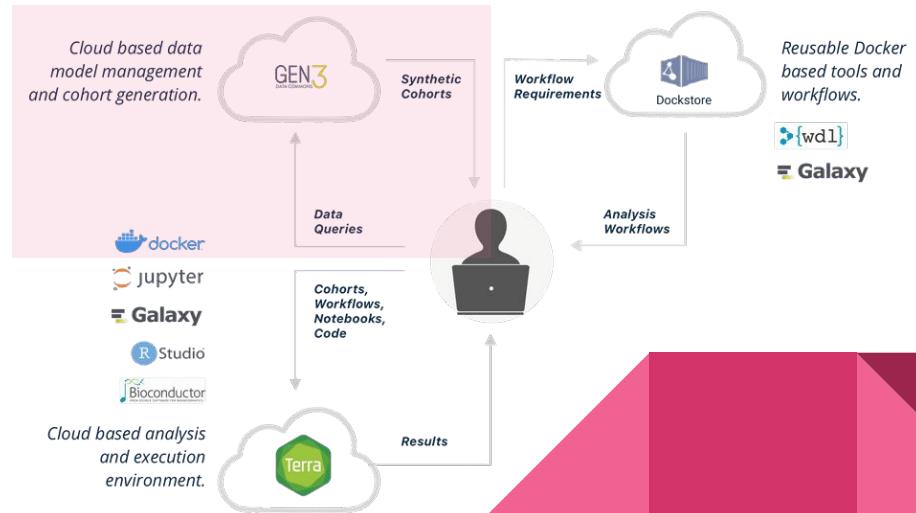
# Terra

- Provides a compute environment with secure data and analysis sharing capabilities
- Provide interactive analysis interfaces such as Jupyter, RStudio, and Galaxy
- **On-demand computational capacity** sourced from Google Cloud Platform
- **Workspace** is the main building block of Terra



# Data

- Data model management and cohort generation by Gen3
- Features :
  - Easy authentication
  - No storage and transfer costs for the data hosted by Terra



## Currently available datasets in AnVIL/Terra



1000 Genomes High Coverage



1000 Genomes Low Coverage



AMP Parkinson's Disease



Baseline Health Study



CCDG presented by NHGRI AnVIL



CMG presented by NHGRI AnVIL



ENCODE Project



Broad Dataset Workspace Library



Framingham Heart Study  
Teaching Dataset



Human Cell Atlas



Neuroscience Multi-Omic Archive



Therapeutically Applicable  
Research to Generate Effective  
Treatments (TARGET) presented  
by the National Cancer Institute



The Cancer Genome Atlas  
Presented by the National Cancer  
Institute



TopMed presented by NHLBI  
BioData Catalyst



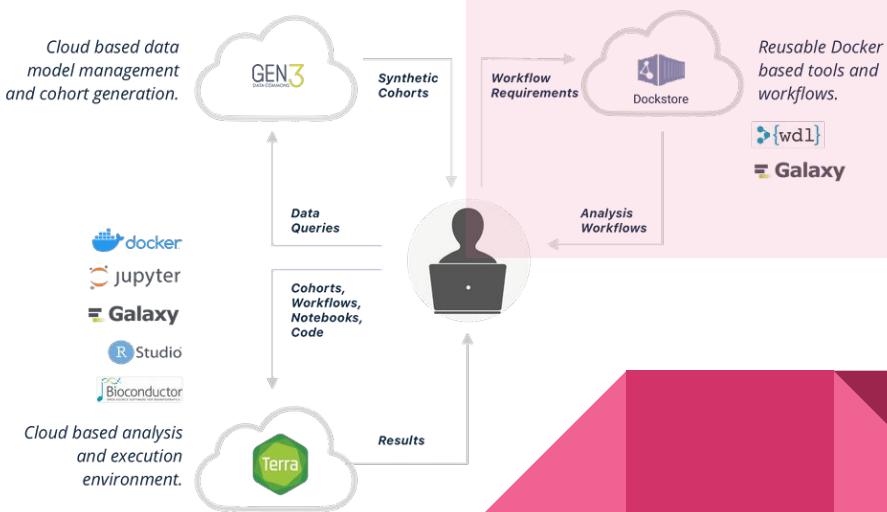
UK Biobank

# Workflows

- Currently, WDL (**Workflow Description Language**) is the only workflow language supported by Terra. (Cromwell, the execution engine, can takes WDL and CWL (**Common Workflow Language**))
- Dockstore : a large collection of pre-built WDL workflows

The screenshot shows the Dockstore web interface. At the top, there's a navigation bar with links for Search, Organizations, Docs, Forum, and a user profile for SehyunOh. Below the navigation is a search bar with the placeholder "Search: contains one of 'RNA'". To the left is a sidebar with filters for Expand All, Collapse All, Reset Filters, Search, Entry Type, Language, Author, Tool Registry, Workflow Source Control, Tool Namespace, Workflow Organization, Labels, and Tool Private Access. The main content area displays a table of search results for WDL workflows. The columns include Name, Verified, Author, Format, Project Links, and Stars. The results listed are:

Name	Verified	Author	Format	Project Links	Stars
ShalekLab/kallisto-bustools_workflow	n/a		WDL	<a href="#">View</a>	0
gatk-workflows/gatk4-maseq-germline-snps-indels	n/a		WDL	<a href="#">View</a>	0
shbrief/biobakery_workflows	n/a		WDL	<a href="#">View</a>	0
ShalekLab/alexandria/smartsq2_cumulus	n/a	Heather Armstrong	WDL	<a href="#">View</a>	0



# WDL

```
version 1.0

task hello {
    input {
        String name
    }

    command {
        echo 'hello ${name}!'
    }
    output {
        File response = stdout()
    }
    runtime {
        docker: 'ubuntu:latest'
    }
}

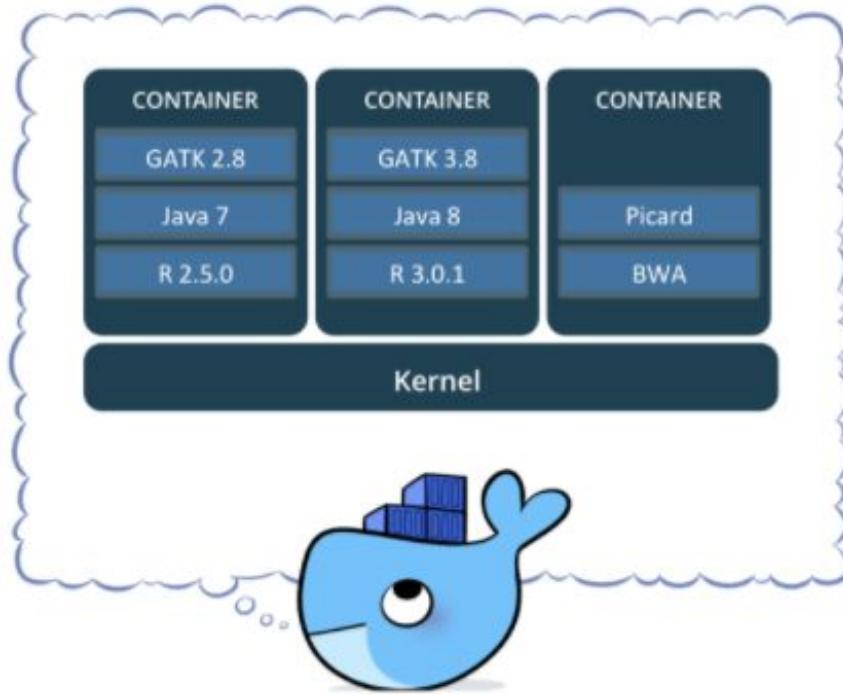
workflow test {
    call hello
}
```

- Top-level components: workflow, task, and call
- Core task-level components: command and output
- Default runtime attributes ([here](#)):

```
runtime {
    "docker": "ubuntu:latest",
    "cpu": 1,
    "memory": "2G",
    "preemptible": 0
}
```

- Additional runtime attributes can be found [here](#).

# Docker Container



A container encapsulates  
**all the software dependencies**  
associated with running a program

**Benefits:**  
Portability, Reusability, and Reproducibility

**Repositories:**  
Docker Hub,  
Dockstore,  
GCR (Google Container Registry)

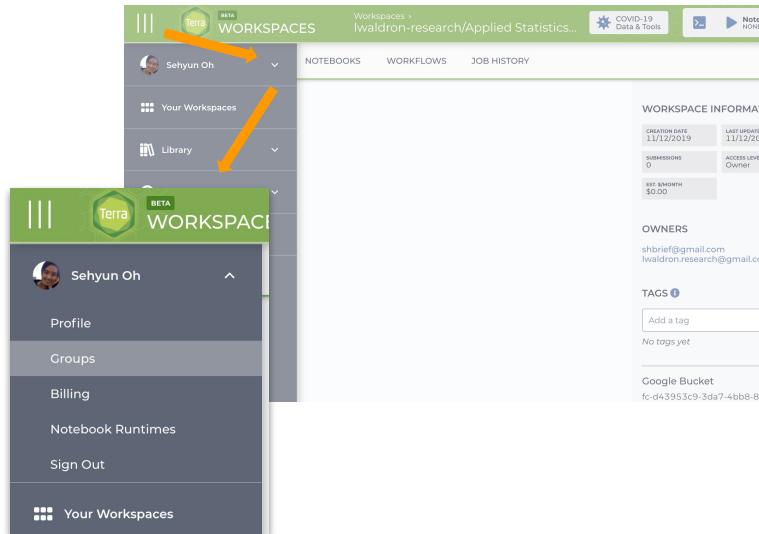
# **Setup 'classroom' in Terra**

*(For teaching)*

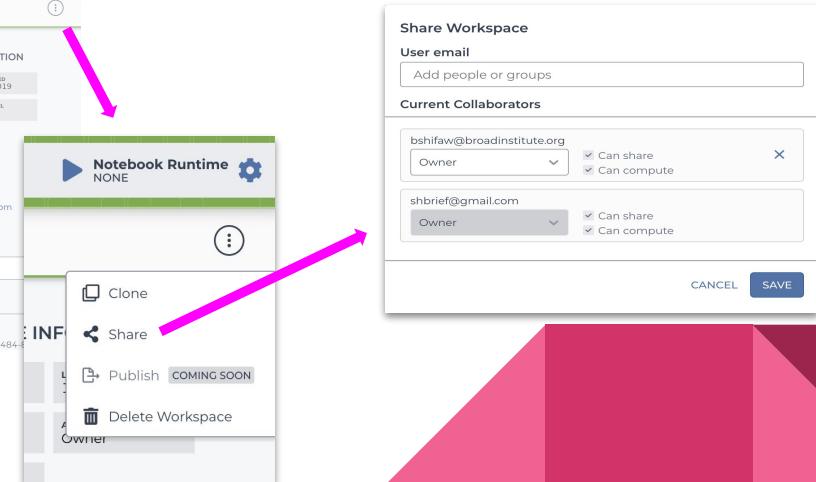
# Hassle-free setup

- Terra workspace allows a setup of a lecture or lab in advance of sharing it, where everyone uses the same runtime environment - no setup or compatibility issues.

*Create a group to share workspace / billing*



*Share a workspace with individual and/or group*



# Flexible runtime environment

- Creating runtime is very intuitive with the cost/hr information.

The diagram illustrates the process of creating a flexible runtime environment. It starts with a 'Default' Cloud Environment, which is then used to create a 'Custom Environment'. The 'Custom Environment' configuration page shows fields for Application configuration, Cloud compute profile (with highlighted CPU and Memory inputs), Startup script, Compute type (Standard VM), and Persistent disk size (50 GB). A red arrow points to the 'CREATE' button. To the right, a dropdown menu for selecting environments shows 'RStudio (R 4.0.3, Bioconductor 3.12.0, Python 3.8.5)' selected, with another red arrow pointing to it. The entire process is labeled 'e.g. select RStudio'.

**Default**

**Cloud environment**

- Cloud environments consist of an application configuration, cloud compute and a persistent disk
- Use default environment
  - Default: (GATK 4.1.4.1, Python 3.7.9, R 4.0.3)  
What's installed on this environment?
  - Default compute size of **4 CPUs, 15 GB memory**, and a **50 GB persistent disk** to keep your data even after you delete your compute
  - Learn more about Persistent disks and where your disk is mounted
- Running cloud compute cost: \$0.20 per hr  
Paused cloud compute cost: < \$0.01 per hr  
Persistent disk cost: \$2.00 per month

**Custom Environment**

**e.g. select RStudio**

Application configuration: RStudio (R 4.0.3, Bioconductor 3.12.0, Python 3.8.5)

Cloud compute profile: 4 CPUs, 15 GB Memory

Startup script: URI

Compute type: Standard VM

Persistent disk size (GB): 50

CREATE

RStudio (R 4.0.3, Bioconductor 3.12.0, Python 3.8.5)

Legacy Python/R (default prior to January 14, 2020)

Legacy GATK (default prior to June 1, 2020) (GATK 4.1.4.1, Python 3.7.7, R 3.6.3)

Legacy R / Bioconductor (R 3.6.3, Bioconductor 3.10, Python 3.7.7)

COMMUNITY-MAINTAINED JUPYTER ENVIRONMENTS (VERIFIED PARTNERS)

Pegasus (Pegasuspy 1.0, Python 3.7, scPlot 0.0.16, harmony-pytorch 0.1.3)

COMMUNITY-MAINTAINED RSTUDIO ENVIRONMENTS (VERIFIED PARTNERS)

RStudio (R 4.0.3, Bioconductor 3.12.0, Python 3.8.5)

OTHER ENVIRONMENTS

Custom Environment

# Costs for GCP resources

## Hourly cost for custom environments

Virtual CPUs	Memory	Price (USD)
1	3.75GB	\$0.04749975
2	7.5GB	\$0.0949995
4	15GB	\$0.189999
8	30GB	\$0.379998
16	60GB	\$0.759996
32	120GB	\$1.519992
64	240GB	\$3.039984

## Persistent disk pricing

\$0.040 per GB / month in USD

(e.g. 50GB persistent disk costs \$2.00 per month.)

## Cost-saving strategies

- Auto-shutdown (for notebooks)
- Use call caching (for workflows)
- Delete intermediate outputs (for workflows)

# Summary

- **Positives:**
  - No setup or compatibility issues
  - Each student selects a compute environment with known cost per hour : each student can select what they need
  - Terra's auto-suspension of notebook runtimes helped keep costs low
  - Students only have to login, don't have to set up billing
- **Negatives:**
  - Each student selects a compute environment with known cost per hour: No way to identify an over-spending student or to limit what runtime they must use
  - No cost breakdown per student
- **Notes:**
  - Billing is post-pay (you can find out how much has been spent with ~24h delay)

# **Bioinformatics analysis on Terra UI**

*(For wet-lab scientists)*

# RNA sequencing analysis

- Terra workspace : [Bioconductor-Workflow-DESeq2](#)

## Workflow for fastq → count matrix using salmon

The screenshot shows the Terra workspace interface with the 'WORKFLOWS' tab selected. The workflow titled 'AnVILBulkRNASEq' is displayed. Step 1 shows 'SELECT DATA' with a dropdown for 'Root entity type' set to 'participant\_set'. Step 2 shows 'SELECT DATA' with 'No participant\_sets selected'. Below these steps is a table of inputs:

Task name	Variable	Type	Attribute
salmon	fastqs_1	Array[File]	this.participants.fastq_1
salmon	fastqs_2	Array[File]	this.participants.fastq_2
salmon	transcriptome.fasta	File	'gs://fc-78b3afba-a1d5-4c43-9a84-4b9036230184/athal.fasta'
salmon	transcriptome_index_name	String	"athal.index"

## Notebook for interactive analysis

### Introduction

This vignette will walk you through how to run a full DESeq2 analysis on the output data from the AnVILBulkRNASEq workflow. The output data should have been retrieved in the previous vignette [Managing the Workflow Output](#).

### Installation

How to install the `AnVILBulkRNASEq` package is shown in the first vignette [An Overview of AnVIL BulkRNASEq](#). Refer to that vignette for installation steps. The following command will load the package.

```
In [ ]: library(AnVILBulkRNASEq)
```

Again, we will need functionality from `AnVIL`, as well as other packages so we will install and load them now.

```
In [ ]: pkgs = c("Bioconductor>AnVIL", "GenomicFeatures", "tximport", "DESeq2")
BiocManager::install(pkgs)

suppressPackageStartupMessages({
  library(AnVIL)
  library(GenomicFeatures)
  library(tximport)
  library(DESeq2)
})
```

### Creating the DESeq2 dataset

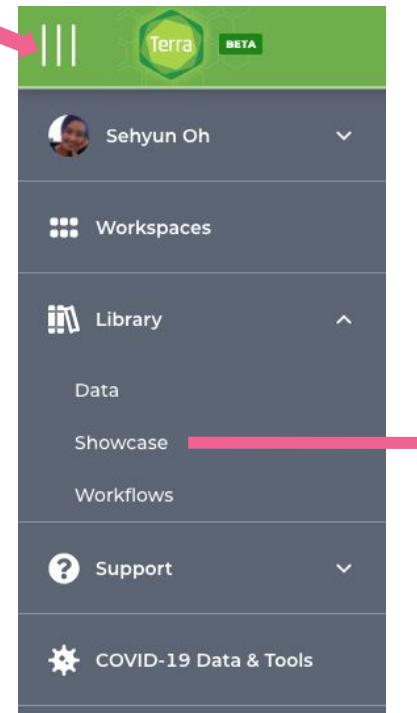
The files that are needed for the DESeq2 analysis are the `.quant.sf` files for each sample. We create the path to those files (for each sample) and save them to `files`.

```
In [ ]: files_path <- paste0(getwd(), "/DRR0161%<_1/quant.sf")
files <- sprintf(files_path, 25:40)
```

A txdb object is needed for the analysis so we download the GTF file associated with Arabidopsis thaliana and run `makeTxDbFromGFF()` on the downloaded file.

```
In [ ]: download.file("ftp://ftp.ensemblgenomes.org/pub/plants/release-28/gtf/arabidopsis_thaliana/Arabidopsis_thaliana.TAIR1
0.28.gtf.gz",
```

# Available workspaces under Showcase



The screenshot shows the 'Showcase & Tutorials' section of the Terra library. The header includes the Terra logo and a 'BETA' badge. The main content is organized into three columns:

- New and interesting:**
  - COVID-19\_Broad\_Viral\_NGS: A workspace for analyzing COVID-19 viral NGS data from the Broad Institute.
  - ml4h-toolkit-for-machine-learning-on-clinical-data: A workspace for machine learning on clinical data using the ml4h toolkit.
  - Metis-toolkit-for-vaccine-trial-planning: A workspace for decision support tool Metis, used for vaccine trial planning.
  - COVID-19\_cross\_tissue\_analysis: A workspace for cross-tissue analysis of COVID-19 data.
- Featured workspaces:**
  - Introduction-to-TCGA-Dataset: A workspace for introducing the TCGA dataset.
  - DNA-methylation-preprocessing: A workspace for DNA methylation data preprocessing.
  - Bioconductor: A workspace for exploring common Bioconductor packages.
  - Waddington-OT: A workspace for Waddington-OT analysis.
- GATK4 example workspaces:**
  - Germline-CNVs-GATK4: A workspace for germline CNV analysis using GATK4.
  - Variant-Functional-Annotation-With-Funcotator: A workspace for variant functional annotation using Funcotator.
  - Variant\_Calling\_Spark\_Multicore: A workspace for variant calling using Spark on a multicore machine.
  - GATK4-Germline-Preprocessing: A workspace for germline preprocessing using GATK4.

# Available workspaces under Workspaces

The image shows the Terra Workspaces interface. On the left, there is a sidebar with a user profile for Sehyun Oh, a 'Workspaces' section highlighted with a red arrow, and links for 'Library', 'Support', and 'COVID-19 Data & Tools'. The main area is titled 'WORKSPACES' and shows a search bar with 'scRNA-seq' and a filter for 'PUBLIC (3)'. A red arrow points to the 'PUBLIC (3)' button. Below this, a table lists three workspaces:

Name	Last Modified	Created By	Access Level
InferCNV SCP scRNA-seq ## inferCNV	Jan 6, 2020	bshifaw@broadinstitute.org	Reader
PrimateRetinalCellAtlas-RegevSanes-Retina-BroadInstit... Human retinal cells from 1 subject (approximately 4000 cells) pro...	Feb 20, 2019	nsharif@broadinstitute.org	Reader
scRNA-seq-cloud ## Single-Cell / Single-Nucleus RNA-seq (Sc/SnRNA-seq) toolkit wi...	May 12, 2020	bshifaw@broadinstitute.org	Reader

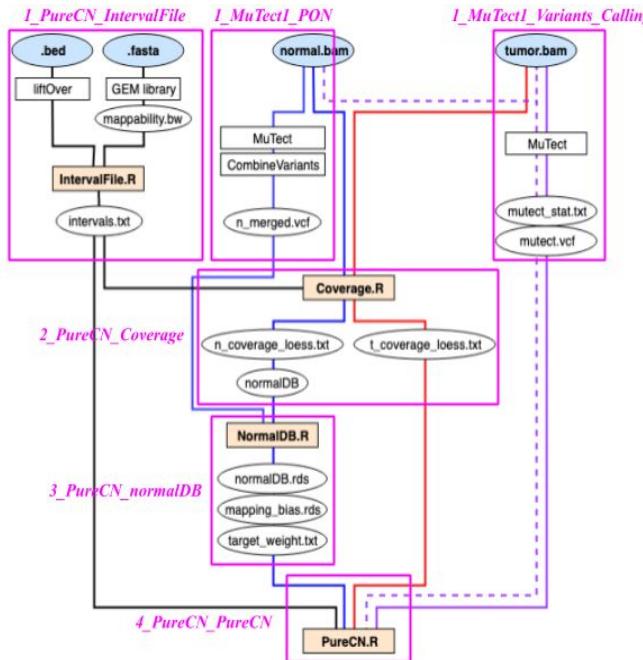
# Summary

- Minimum coding thanks to the GUI and pre-implemented workflows and notebooks
- Available workspaces:
  - RNA sequencing analysis
  - Single cell analysis
  - genesis-GWAS
  - GATK best practices from Broad Institute
- Access Google Cloud resources
- [Note] Your data should be stored in Google cloud storage. And if you want to use Terra's data model, you need to create and upload the table of your data stored in Google cloud storage.

# **Share your published work through Terra**

*(For bioinformaticians)*

# CNV analysis



- Reliable analysis of clinical tumor-only whole-exome sequencing data  
(Oh et al., JCO Clin Cancer Inform, 2020)
- Terra workspace : [Tumor\\_Only\\_CNV](#)

# Dashboard

- Contains information on the workspace

The screenshot shows the Terra Workspaces interface. At the top, there's a green header bar with the Terra logo, a 'WORKSPACES' section labeled 'BETA', and navigation links for 'Workspaces > waldronlab-terra/Tumor\_Only\_CNV'. On the right of the header are icons for 'COVID-19 Data & Tools' (with a progress bar), 'Cloud Environment' (status: STOPPED (~\$0.01 hr)), and a gear icon.

The main content area has tabs for 'DASHBOARD' (selected), 'DATA', 'NOTEBOOKS', 'WORKFLOWS', and 'JOB HISTORY'. Below these tabs is a 'ABOUT THE WORKSPACE' section containing a title 'Reliable analysis of tumor CNV/SNV without matching normal' and a detailed description of the workspace's purpose and a link to a recent publication. It also includes a note about a synthetic dataset from another workspace.

On the right side, there are three main sections: 'WORKSPACE INFORMATION' (Creation Date: 5/21/2020, Last Updated: 7/17/2020, Submissions: 0, Access Level: Proj. Owner, Est. \$/Month: \$0.00), 'OWNERS' (email addresses: bshifaw@broadinstitute.org, shbrief@gmail.com), and 'TAGS' (a dropdown menu with options like Bioconductor, CNV, cnv, public, PureCN, purecn). Below the tags is a 'Google Bucket' section with a URL and an 'Open in browser' button.

At the bottom left, there's an 'Overview' section with a detailed paragraph about allele-specific copy number alteration (CNA) analysis and its limitations. At the very bottom, a note states: 'This workflow, based on the open-source [PureCN](#) Bioconductor package in conjunction with'.

# Data

- Paper used TCGA controlled data (BAM files) → **Synthetic dataset** for public workspace
- Pre-populated public reference files (provided by Terra) → available under '**Reference Data**'
- Researcher's own data (e.g. BED file in Google bucket) → linked under '**Workspace Data**'

The screenshot shows the Terra Workspaces interface. The top navigation bar includes 'WORKSPACES' (BETA), 'Data', 'COVID-19 Data & Tools', and 'Cloud Environment STOPPED (< \$0.01 hr)'. The main area has tabs for 'DASHBOARD', 'DATA', 'NOTEBOOKS', 'WORKFLOWS', and 'JOB HISTORY'. On the left, a sidebar lists 'TABLES' (with 'participant (100)' highlighted by a blue arrow), 'REFERENCE DATA' (highlighted by a red arrow), 'OTHER DATA' (with 'b37Human' listed), and 'Workspace Data' (highlighted by a green arrow). The central table displays rows for participant\_id (HG00096, HG00097, HG00128, HG00131, HG00142, HG00143) with roles 'neutral' and paths to synthetic.bam and synthetic.bai files.

	participant_id	role	synthExomeBam	synthExomeBamIndex
	HG00096	neutral	<a href="#">HG00096.synthetic.exome.bam</a>	<a href="#">HG00096.synthetic.exome.bai</a>
	HG00097	neutral	<a href="#">HG00097.synthetic.exome.bam</a>	<a href="#">HG00097.synthetic.exome.bai</a>
	HG00128	neutral	<a href="#">HG00128.synthetic.exome.bam</a>	<a href="#">HG00128.synthetic.exome.bai</a>
	HG00131	neutral	<a href="#">HG00131.synthetic.exome.bam</a>	<a href="#">HG00131.synthetic.exome.bai</a>
	HG00142	neutral	<a href="#">HG00142.synthetic.exome.bam</a>	<a href="#">HG00142.synthetic.exome.bai</a>
	HG00143	neutral	<a href="#">HG00143.synthetic.exome.bam</a>	<a href="#">HG00143.synthetic.exome.bai</a>

# Notebooks

- 5 Jupyter notebooks written in R → 4 for data pre-processing and 1 for downstream analysis
- AnVIL package enables a direct connection between ‘Data’ and ‘Notebooks’

The screenshot shows the Terra Workspaces interface. At the top, there's a green header bar with the Terra logo, the word "WORKSPACES" (with a "BETA" badge), and navigation links for "Workspaces", "Notebooks", "Workflows", and "Job History". On the right of the header are icons for "COVID-19 Data & Tools", "Cloud Environment" (status: STOPPED (~ \$0.01 hr)), and a gear icon.

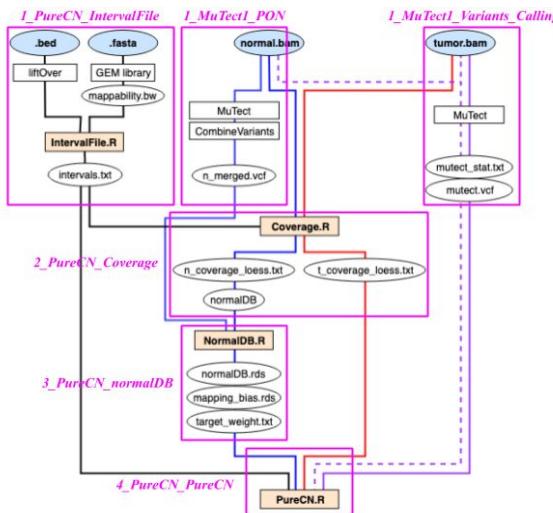
The main content area is titled "NOTEBOOKS". It features a search bar labeled "SEARCH NOTEBOOKS" and a dropdown menu "Sort By: Alphabetical". Below these are five cards, each representing a notebook:

- 1\_Annotate\_Manifest - Last edited: Jun 9, 2020
- 2\_Build\_Data\_Table - Last edited: Jun 9, 2020
- 3\_Format\_BED - Last edited: May 21, 2020
- 4\_Download\_SNP\_Blacklist - Last edited: May 21, 2020
- 5\_Downstream\_Analysis - Last edited: May 21, 2020

To the left of the notebook list, there are two boxes: one for creating a new notebook and another for adding an ipynb file.

# Workflows

- Pipeline was implemented into 7 WDL workflows in Terra, based on their modularity and input/output requirements.
- These workflows incorporate many different runtime environments (e.g. GATK, MuTect, Bioconductor, etc.)



Workflow	Version	Source
1_MuTect1_PON	V. master	Source: dockstore
1_MuTect1_Variants_Calling	V. master	Source: dockstore
1_PureCN_IntervalFile	V. 1	Source: Terra
2_PureCN_Coverage	V. master	Source: dockstore
3_PureCN_normalDB	V. master	Source: dockstore
4_PureCN_PureCN	V. SynthData	Source: dockstore
5_PureCN_Dx	V. SynthData	Source: dockstore

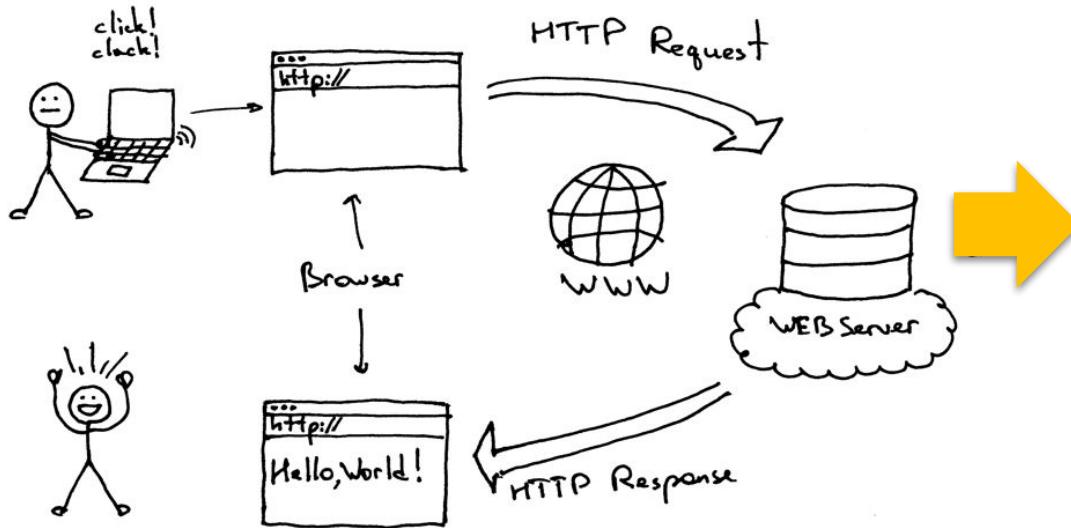
# Summary

- The major benefits of having Terra workspace for research papers are :
  1. Data storage, compute-intensive pipeline, and downstream analyses are all available in one place
  2. Improved the reproducibility
  3. Sharing code and providing additional information not included in the paper are available through the workspace
- One potential downside is that for a complicating pipeline, like CNV analysis, writing WDL and managing data model can be non-trivial.

# **Workflow package powered by Terra**

*(For Developers)*

# API (Application Programming Interface)



# AnVIL package

**For the end-users**, AnVIL provides fast binary package installation, utilities for working with Terra / AnVIL table and data resources, and convenient functions for file movement to and from Google cloud storage.

Using `gcloud_*`() for account management

```
> gcloud_account() # authentication account  
[1] "shbrief@gmail.com"  
> gcloud_project() # billing project information  
[1] "bioinfo"
```

Using `gsutil_*`() for file and bucket management

```
> src <- "gs://biobaker/"  
> gsutil_ls(src)  
[1] "gs://biobaker/ibdmdb_demo_metadata_test.txt" "gs://biobaker/ibdmdb_file_list_test.txt"  
> pathToFastq <- "gs://biobaker/ibdmdb_file_list_test.txt"  
> read.table(gsutil_pipe(pathToFastq), sep = "\t")  
V1  
1 gs://fc-7130738a-5cde-4238-b00a-e07eba6047f2/IBDMDB/HSM7J4NY_R1.fastq.gz  
2 gs://fc-7130738a-5cde-4238-b00a-e07eba6047f2/IBDMDB/HSMA330T_R1.fastq.gz
```

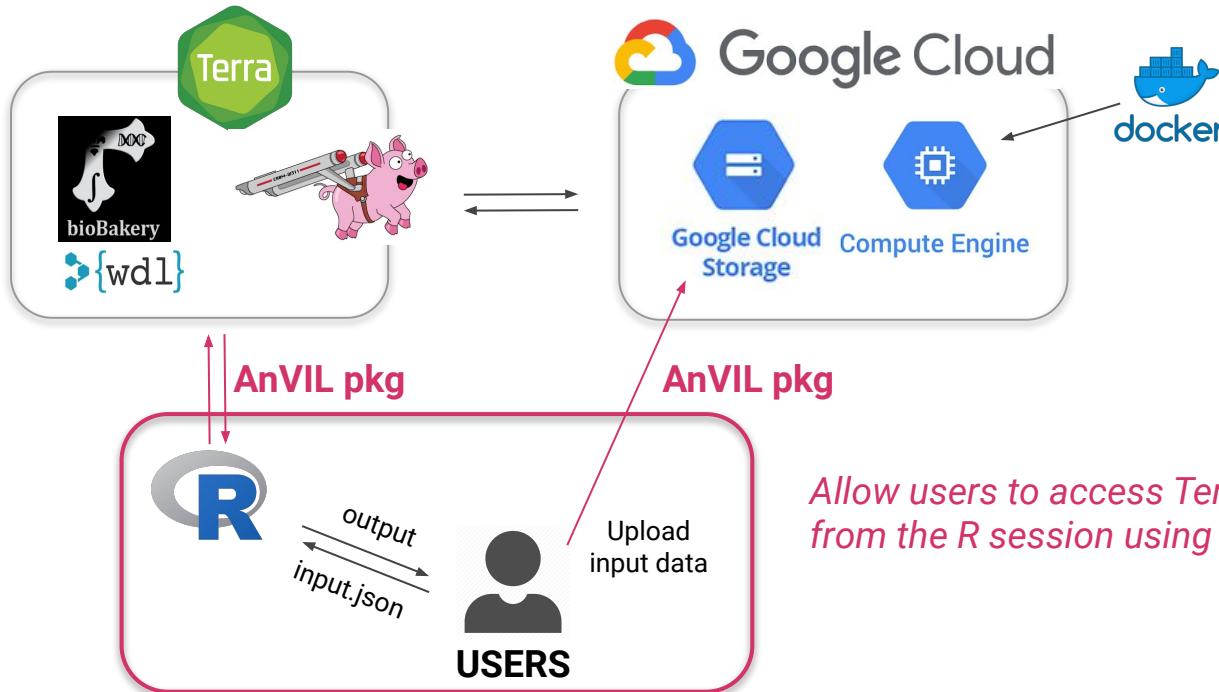
# AnVIL package

**For the developers**, AnVIL provides programmatic access to the Terra, Leonardo, Dockstore, and Gen3 RESTful programming interface, including helper functions to transform JSON responses to the formats more amenable to manipulation in R.

```
> ## Create an instance of service
> terra <- Terra()
> ## Invoke endpoints
> terra$status()
Response [https://api.firecloud.org/status]
  Date: 2020-12-13 00:12
  Status: 200
  Content-Type: application/json
  Size: 245 B

> ## Process responses
> status <- terra$status()
> class(status) # defined in the httr package
[1] "response"
```

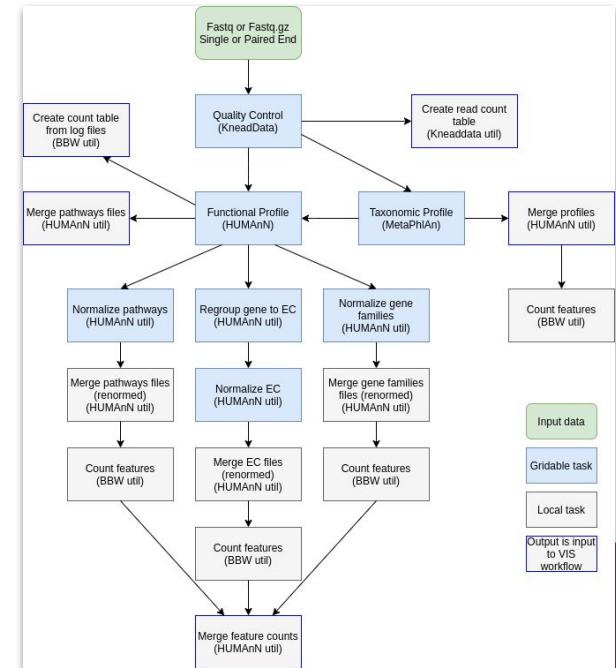
# 'Runnable' workflow package



# Microbiome analysis

- bioBakery workflows is a collection of workflows and tasks for executing common microbial community analyses using standardized, validated tools and parameters.
- Potential blockers for using bioBakery
  - Limited compute and storage resources
  - Unfamiliarity of Python
  - Non-trivial setup process
- Terra workspace : [mtx\\_workflow\\_biobakery\\_version3](#)
  - Whole metagenome shotgun
  - Requirements: Python v2.7+, AnADAMA, KneadData, MetaPhlAn, HUMAnN
  - Tasks: quality control, taxonomic and functional profiling

*DIY workflow*



# biobakeR package

## 0. Prerequisite

[Set up Terra account](#) : you get the required inputs, `accountEmail` and `billingProjectName`, for biobakeR.  
Place your data in Google Cloud Bucket

## 1. Input

`cloneWorkspace()`: copies the template workspace containing the bioBakery workflow.  
`updateInput()` : takes user's inputs.

## 2. Run workflow

`launchWorkflow()` : launches the bioBakery workflow in Terra.

## 3. Result

`monitorSubmission()` : allows you to monitor the status of your workflow run.  
`listOutput()` : displays the list of your workflow outputs.  
`getOutput()` : allows you to download your outputs.

## Clone workspace

```
> cloneWorkspace(accountEmail, billingProjectName, workspaceName = "test")
[1] "Workspace is successfully cloned"
```

## Launch workflow

```
> launchWorkflow(accountEmail, billingProjectName, workspaceName)
[1] "Workflow is successfully launched."
> submissions <- monitorSubmission(accountEmail, billingProjectName, workspaceName)
> submissions
# A tibble: 57 x 6
  submissionId          submitter    submissionDate   status
  <chr>                <chr>        <dttm>           <chr>
  1 0c915297-f8c2-4a29-b642-39a7c9e7974b shbrief@gmail.com 2020-12-13 02:17:56 Submit
  2 80b04b78-22f4-42af-842a-0e0f7e60cc9a shbrief@gmail.com 2020-12-13 02:17:12 Submit
```

## List outputs

```
> listOutput(accountEmail, billingProjectName, workspaceName, submission_id,
+             keyword = "HSM7J4NY.*.tsv")
# A tibble: 9 x 4
  file                  workflow  task      path
  <chr>                <chr>     <chr>    <chr>
  1 HSM7J4NY_genefamilie... workflowM... call-Functiona... gs://fc-071d1d53-e186-44ad-89
  2 HSM7J4NY_pathabundan... workflowM... call-Functiona... gs://fc-071d1d53-e186-44ad-89
  3 HSM7J4NY_pathcoverage... workflowM... call-Functiona... gs://fc-071d1d53-e186-44ad-89
```

## Download outputs

```
> HSM7J4NY_dir <- "~/data2/biobakeR/inst/extdata/outputs/HSM7J4NY"
> getOutput(accountEmail, billingProjectName, workspaceName, submission_id,
+            keyword = "HSM7J4NY.*.tsv", dest_dir = HSM7J4NY_dir)
Copying gs://fc-071d1d53-e186-44ad-8951-d85538f85502/87addfce-5f43-40b0-a5a1-f7a
4-848f-bbbe6f46de64/call-FunctionalProfile/shard-0/cacheCopy/HSM7J4NY_genefamili
Copying gs://fc-071d1d53-e186-44ad-8951-d85538f85502/87addfce-5f43-40b0-a5a1-f7a
```

# Summary

- With biobakeR package, users can run python tools using Google Cloud resources from R session on their own laptop
- Runnable workflow packages can minimize the overhead for R users  
→ Users don't need to setup computing environment nor need to learn WDL, Terra, and GCP to run Terra-implemented workflows

# Conclusions

1. Terra offers an easy way to share bioinformatics work with the identical runtime environment, facilitating collaboration and teaching.
2. Terra workspaces and workflows enable complicating bioinformatics analyses with minimum coding.
3. You can increase the reproducibility of your work by sharing it through Terra, where you can host data, workflow, and downstream analysis all together.
4. Workflow package powered by Terra allows users to utilize Google cloud resources and even non-R tools from R session on their own laptop in a familiar way

# Acknowledgements

## Waldron's lab

- Levi Waldron
- Marcel Ramos

## Bioconductor-AnVIL team

- Martin Morgan
- Vince Carey
- Nitesh Turaga
- Lori Shepherd
- BJ Stubbs
- Kayla Interdonato

## Funding

NHGRI supports AnVIL through [cooperative agreement awards](#) to the [Broad Institute](#) (#5U24HG010262) and [Johns Hopkins University](#) (#5U24HG010263).

# Links

- Gen3 : <https://gen3.org/>
- Dockstore : <https://dockstore.org/>
- Default WDL runtime attributes :  
<https://support.terra.bio/hc/en-us/articles/360046944671-Default-runtime-attributes-for-workflow-submissions>
- RNA Sequencing Analysis Workspace (contact me to access):  
<https://app.terra.bio/#workspaces/bioconductor-rpc1-anvil/Bioconductor-Workflow-DESeq2>
- Tumor\_Only\_CNV workspace :  
[https://app.terra.bio/#workspaces/waldronlab-terra/Tumor\\_Only\\_CNV](https://app.terra.bio/#workspaces/waldronlab-terra/Tumor_Only_CNV)
- bioBakery : [https://huttenhower.sph.harvard.edu/biobakery\\_workflows/](https://huttenhower.sph.harvard.edu/biobakery_workflows/)
- bioBakery workspace (contact me to access):  
[https://app.terra.bio/#workspaces/rjxmicrobiome/mtx\\_workflow\\_biobakery\\_version3](https://app.terra.bio/#workspaces/rjxmicrobiome/mtx_workflow_biobakery_version3)
- Set up Terra account :  
<https://support.terra.bio/hc/en-us/articles/360034677651-Account-setup-and-exploring-Terra>
- BioC2020 Workshop on AnVIL/Terra : <http://waldronlab.io/AnVILWorkshop/>
- BioC-AnVIL Slack Channel : <https://join.slack.com/share/zt-k04vu3kl-mtu6MltdX8VB7Bx1k~FLg>
- BioC-AnVIL project website : [https://bioconductor.github.io/AnVIL\\_Admin/](https://bioconductor.github.io/AnVIL_Admin/)
- biobakeR : <https://github.com/shbrief/biobakeR>
- Get \$300 Google credits : <https://support.terra.bio/hc/en-us/articles/360046295092>
- Contact for an inquiry on BioC-AnVIL credit : [Sehyun.Oh@sph.cuny.edu](mailto:Sehyun.Oh@sph.cuny.edu)
- Reference book : <https://www.amazon.com/Genomics-Cloud-GATK-Spark-Docker/dp/1491975199>