Bioconductor On The Cloud

EuroBioc2020
December 14th, 2020

Sehyun Oh, PhD

Sehyun.Oh@sph.cuny.edu
https://github.com/shbrief

Why Genomics in the Cloud?

COMPUTATIONAL TOOLS

Cloud computing for genomic data analysis and collaboration

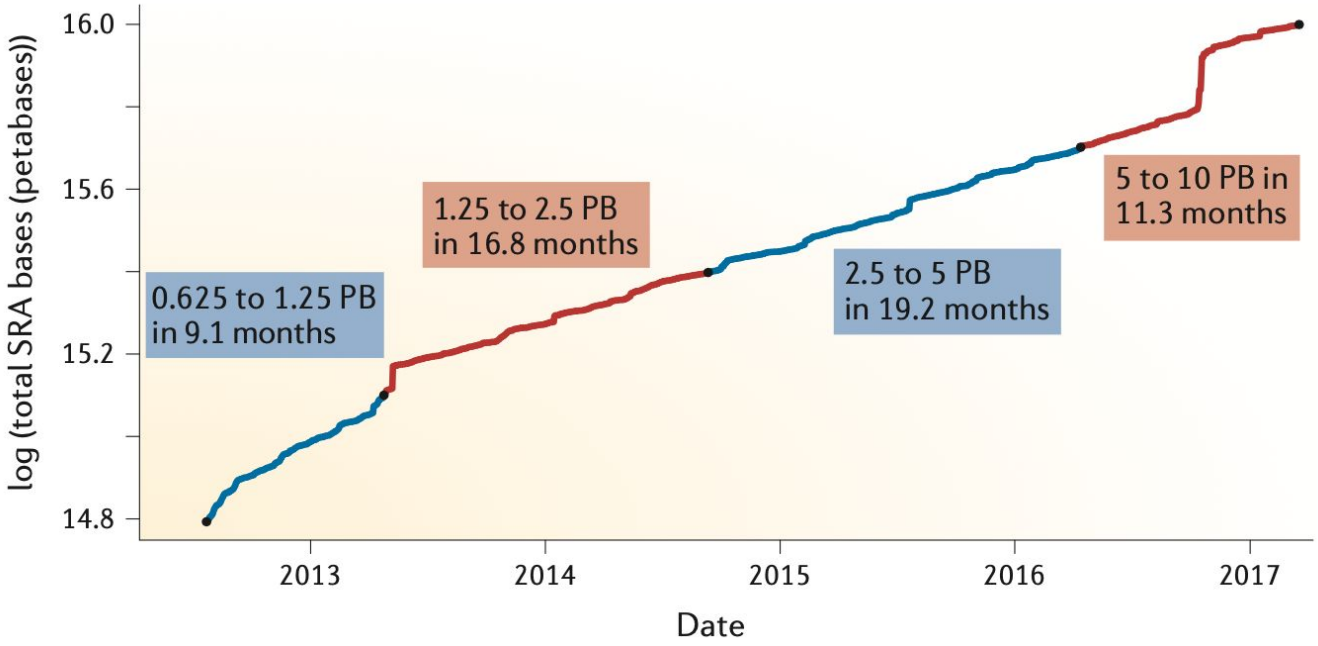
Ben Langmead[1] and Abhinav Nellore[2]

Abstract | Ne tudies based
on large sequ uencing
data have be earchers to
use large-sca ent
computers a n genomics
research. Her nd
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.

… its elasticity, reproducibility, and privacy features make it ideally suited …

https://doi.org/10.1038/nrg.2017.113 (Slide credit to Sean Davis)

Doubling every 18 months...



https://doi.org/10.1038/nrg.2017.113

30TB

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

(Slide credit to Sean Davis)

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.

(Slide credit to Sean Davis)

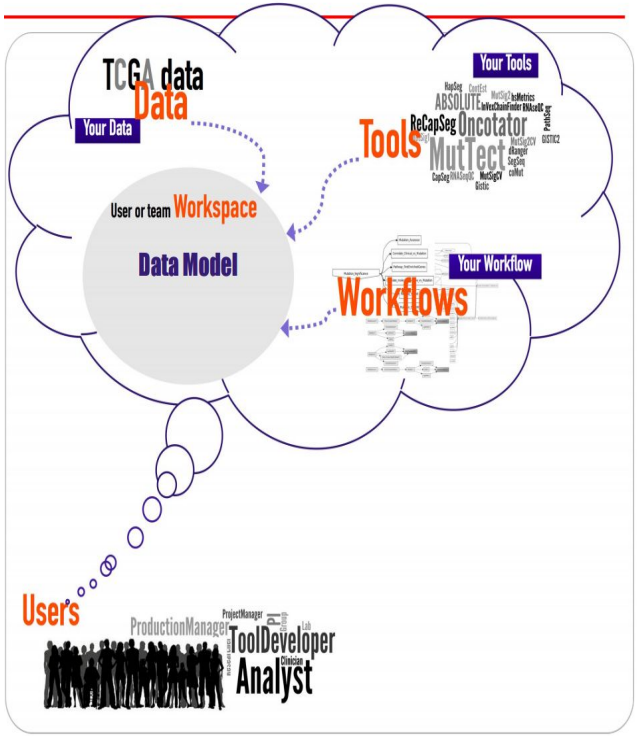
Table 1 | **Large genomics projects and resources**

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

https://doi.org/10.1038/nrg.2017.113

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



(Credit to Sean Davis)

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)

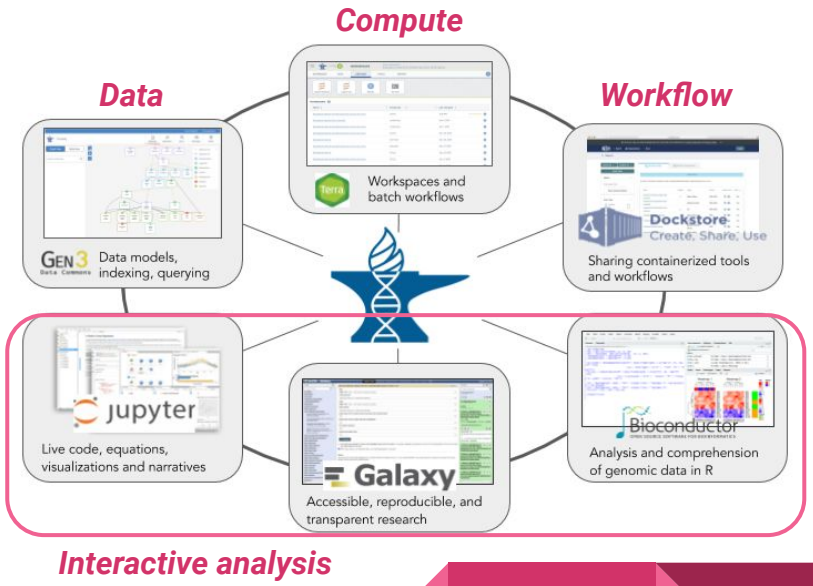
Contents

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

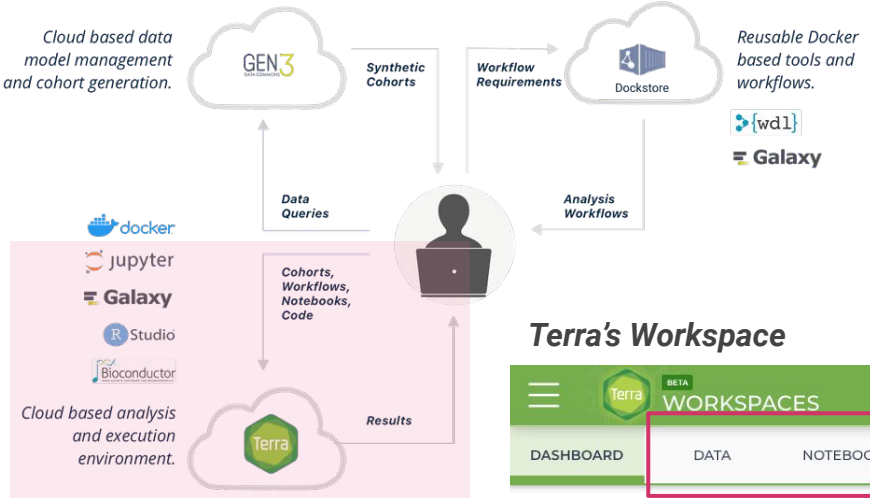
- NHGRI's AnVIL provides a unified environment for data management and compute.
  - no need for data movement
  - better security handling
  - provide elastic, shared computing resources



https://anvilproject.org/overview

Terra

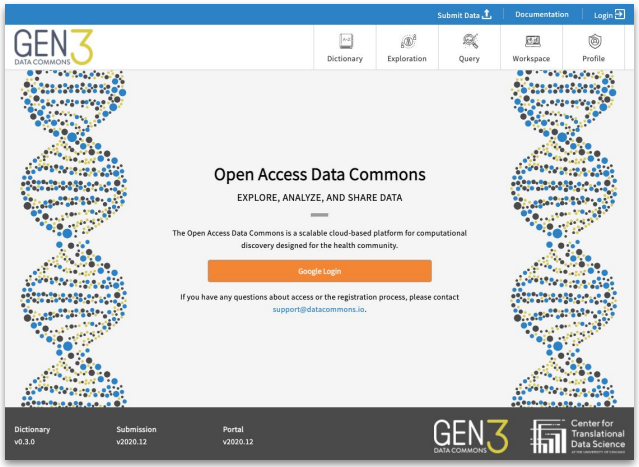
- Provides a compute environment with secure data and analysis sharing capabilities
- Provide interactive analysis interfaces such as Jupyter, RStudio, and Galaxy
- **On-demand computational capacity** sourced from Google Cloud Platform
- **Workspace** is the main building block of Terra

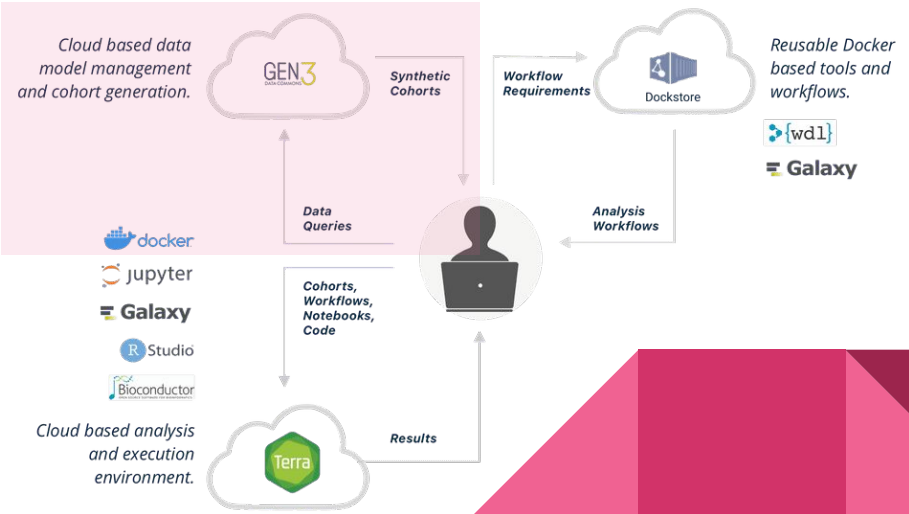


Terra's Workspace

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

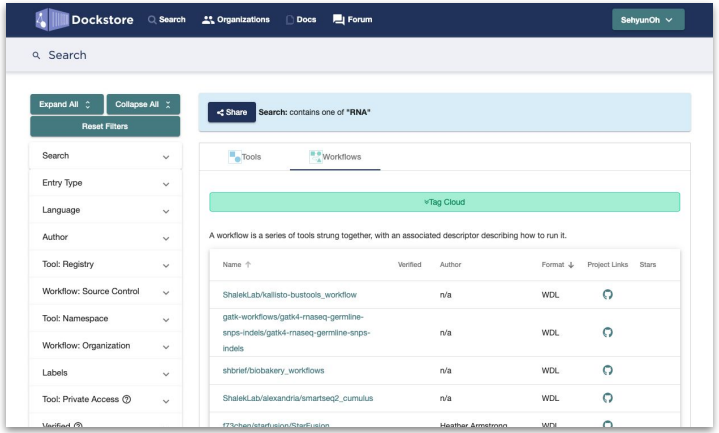
The Cancer Genome Atlas Presented by the National Cancer Institute

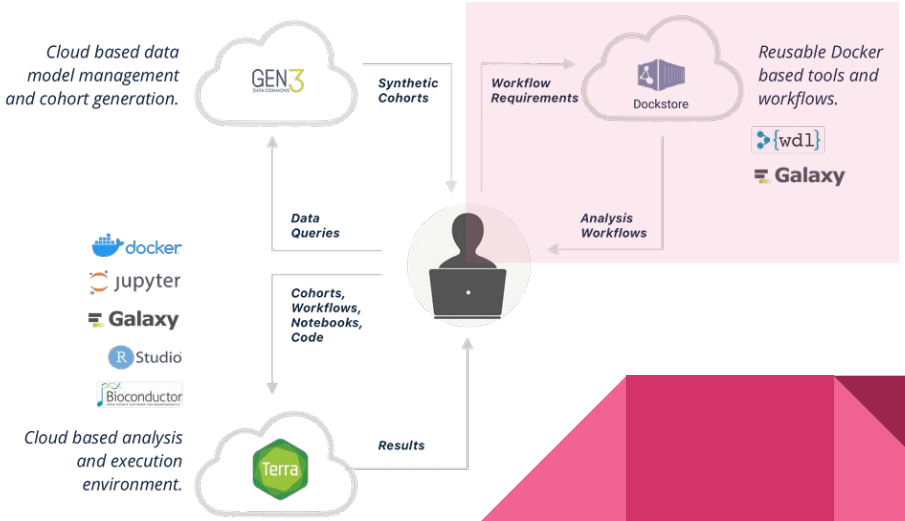
TopMed presented by NHLBI BioData Catalyst

UK Biobank

Workflows

- Currently, WDL (***W***orkflow ***D***escription ***L***anguage) is the only workflow language supported by Terra. (Cromwell, the execution engine, can takes WDL and CWL (***C***ommon ***W***orkflow ***L***anguage))
- Dockstore : a large collection of pre-built WDL workflows





WDL

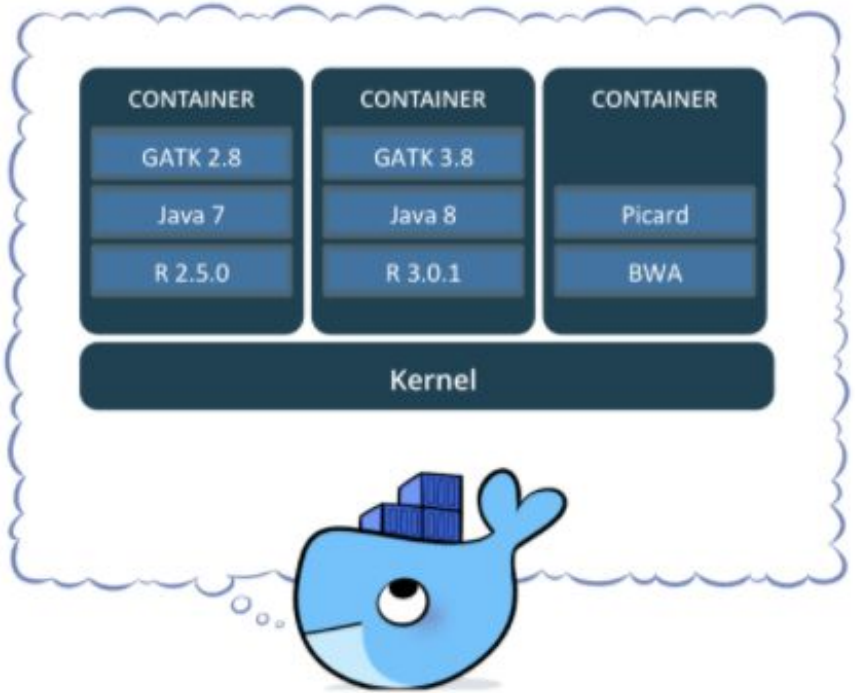
```
version 1.0

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)

https://support.terra.bio/hc/en-us/articles/360037340472-Docker-container-overview

Setup ‘classroom’ in Terra

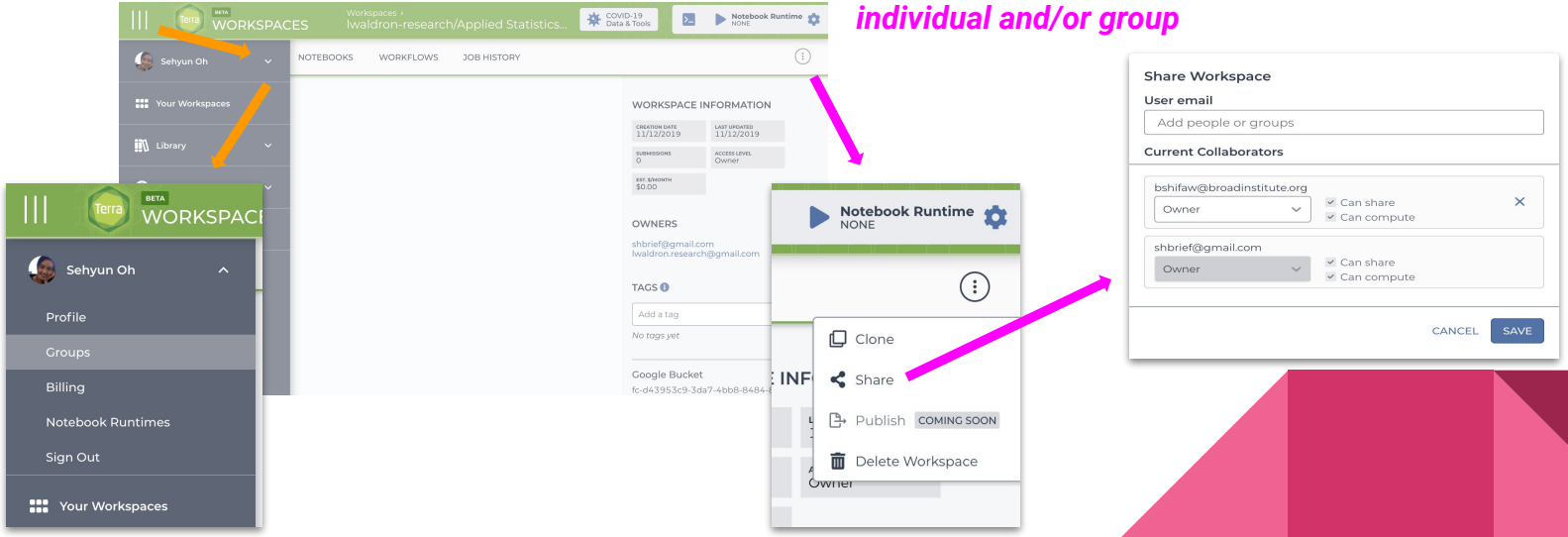
(For teaching)

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.

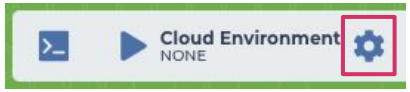
Create a group to share workspace / billing

Share a workspace with individual and/or group

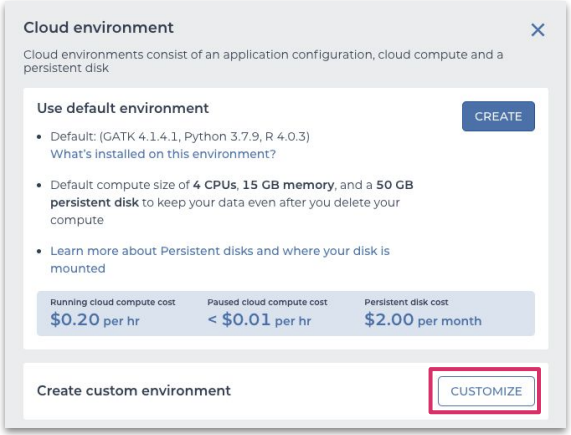


Flexible runtime environment

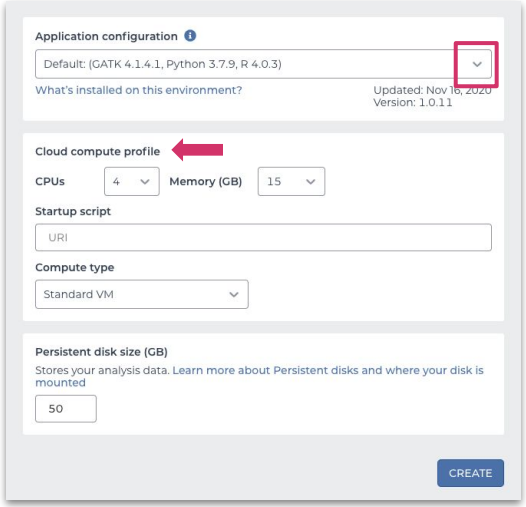
- Creating runtime is very intuitive with the cost/hr information.



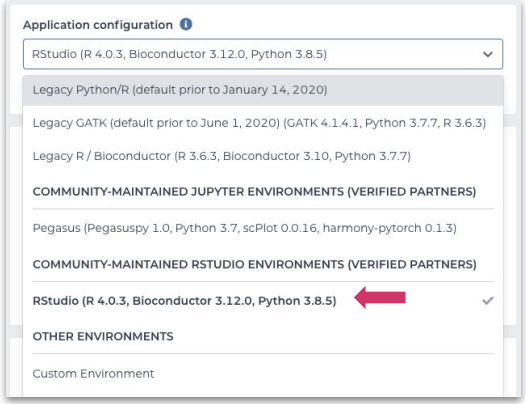
Default



Custom Environment



e.g. select RStudio



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)

https://cloud.google.com/compute/vm-instance-pricing#n1_predefined

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

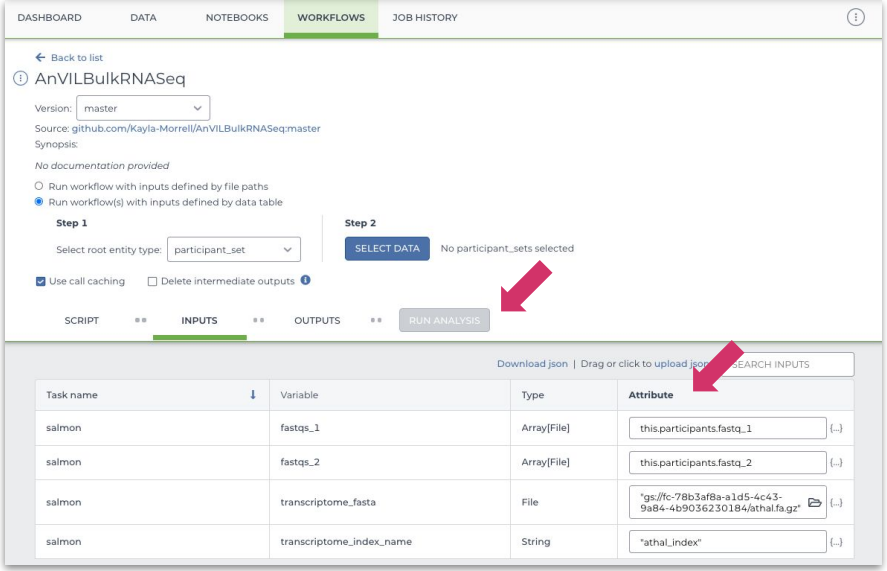
(For wet-lab scientists)

RNA sequencing analysis

- Terra workspace : [Bioconductor-Workflow-DESeq2]

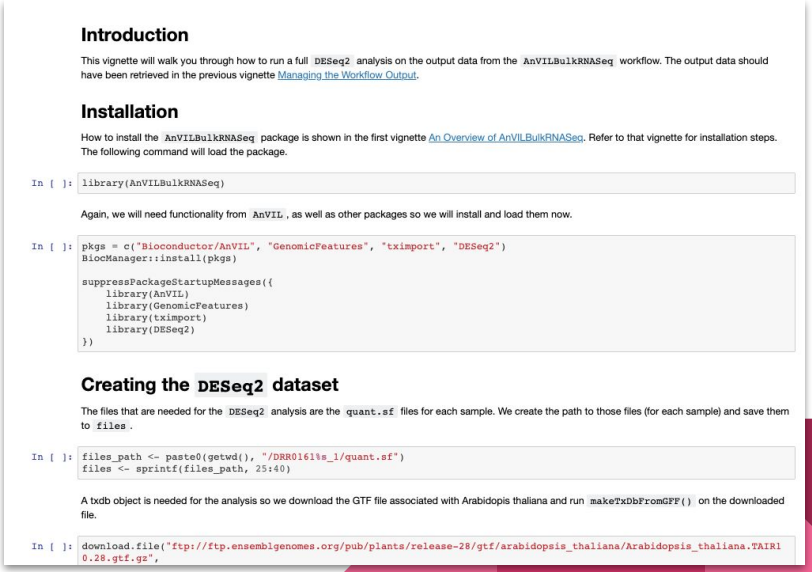
Workflow

for `fastq` → count matrix using salmon



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 AnVILBulkRNASeq. 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`.

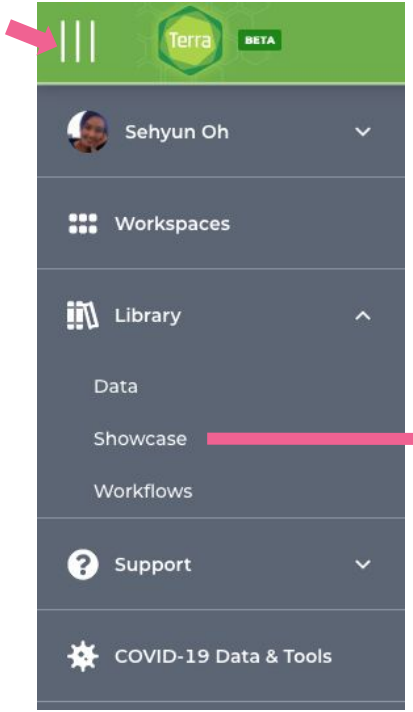
```
In [ ]: files_path <- paste0(getwd(), "/DRR0161%s_1/quant.sf")
        files <- sprintf(files_path, 25:40)
```

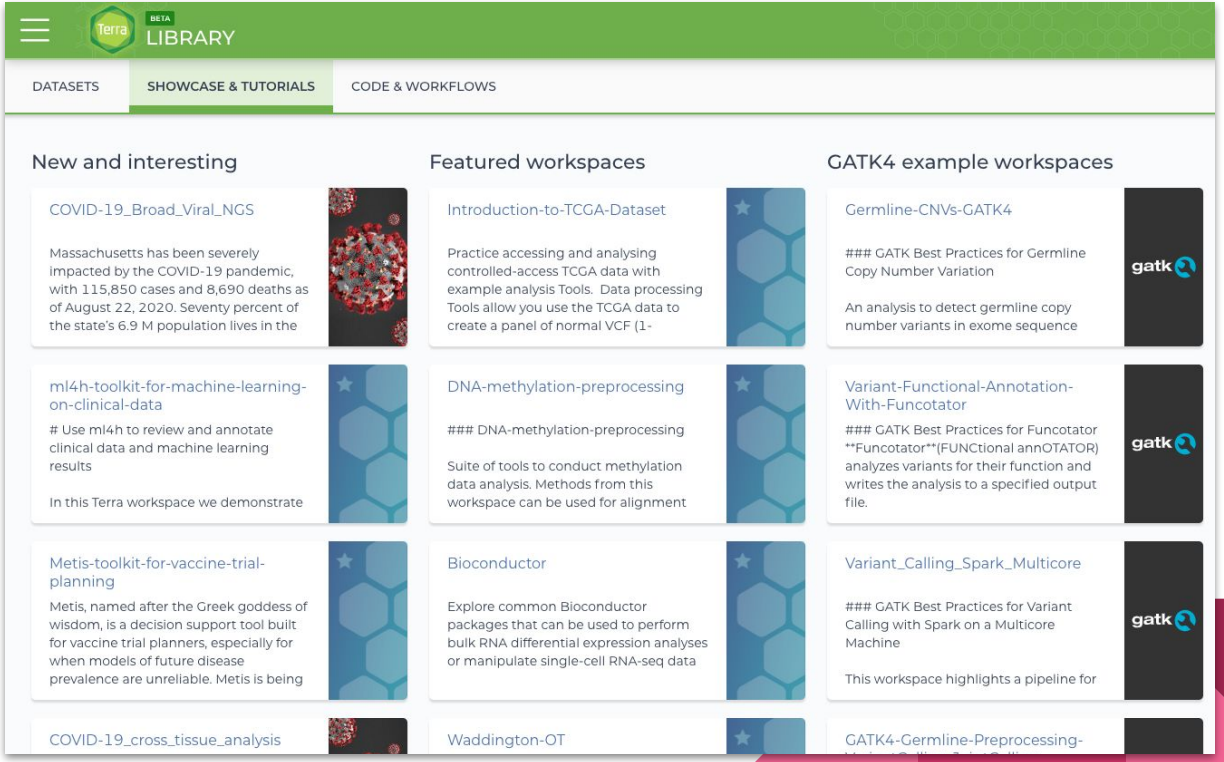
A txdb object is needed for the analysis so we download the GTF file associated with Arabidopsis thaliana and run `makeTxDbFromGFF()` on the downloaded file.

```
In [ ]: download.file("ftp://ftp.ensemblgenomes.org/pub/plants/release-28/gtf/arabidopsis_thaliana/Arabidopsis_thaliana.TAIR10.28.gtf.gz",
```

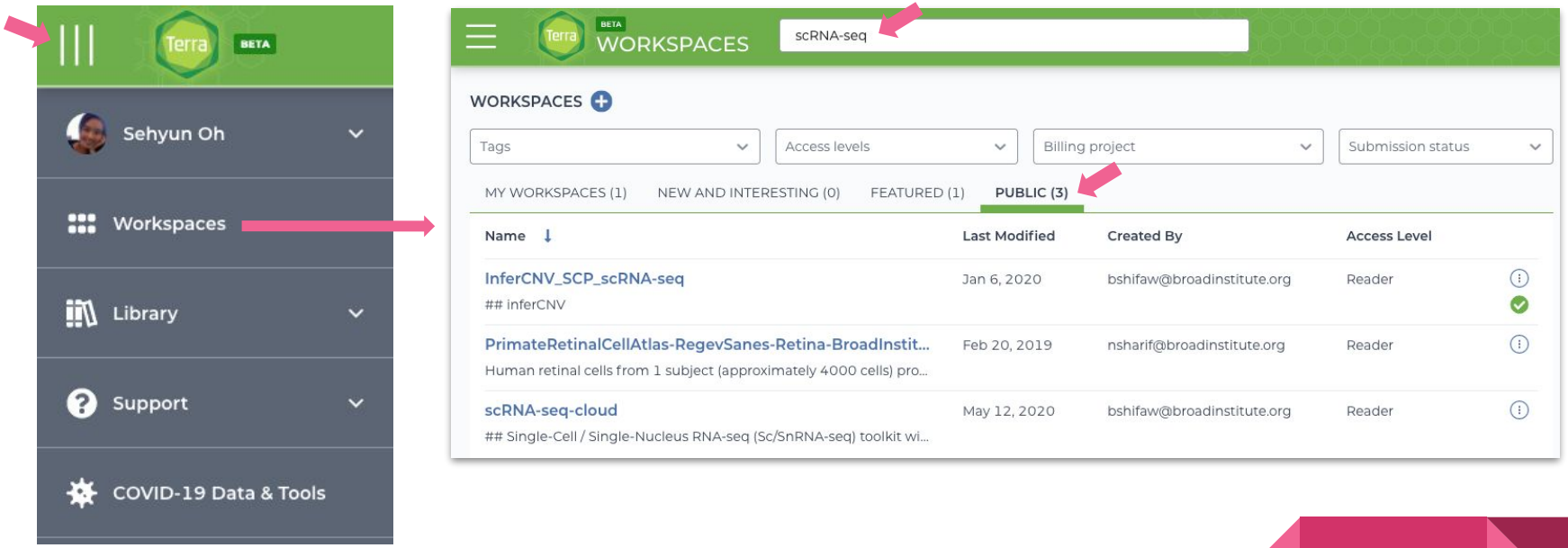

For wet-lab scientists

Available workspaces under *Showcase*





Available workspaces under *Workspaces*



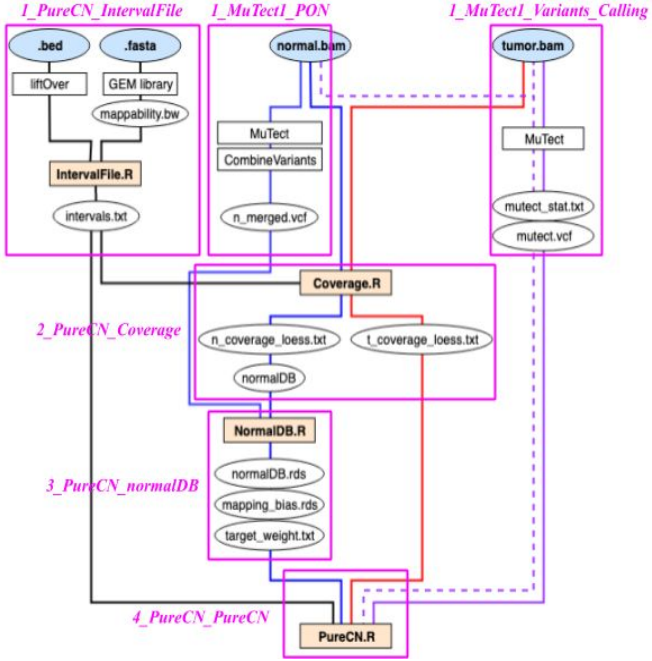
Summary

- 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



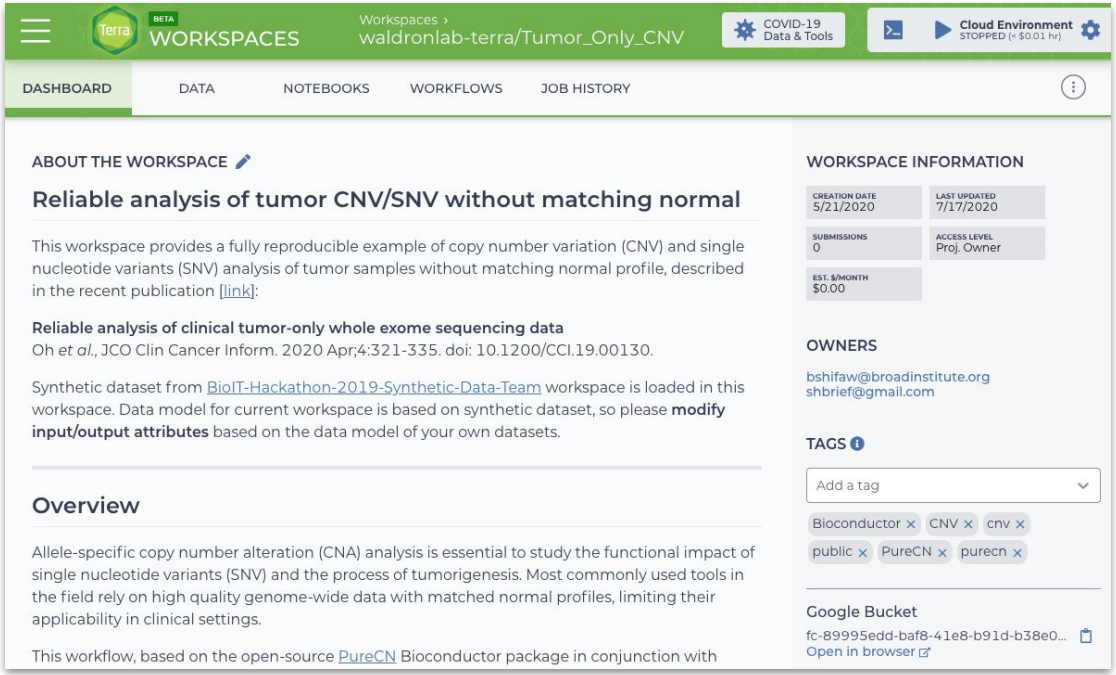
- Reliable analysis of clinical tumor-only whole-exome sequencing data

  (Oh *et al.*, JCO Clin Cancer Inform, 2020)

- Terra workspace : Tumor_Only_CNV

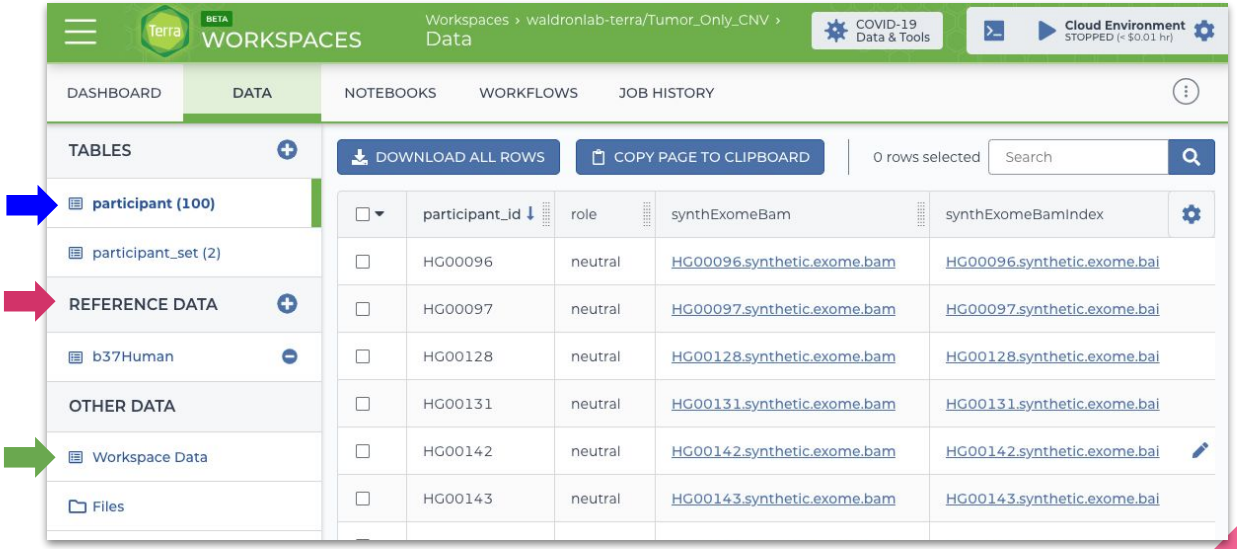
Dashboard

- Contains information on the workspace



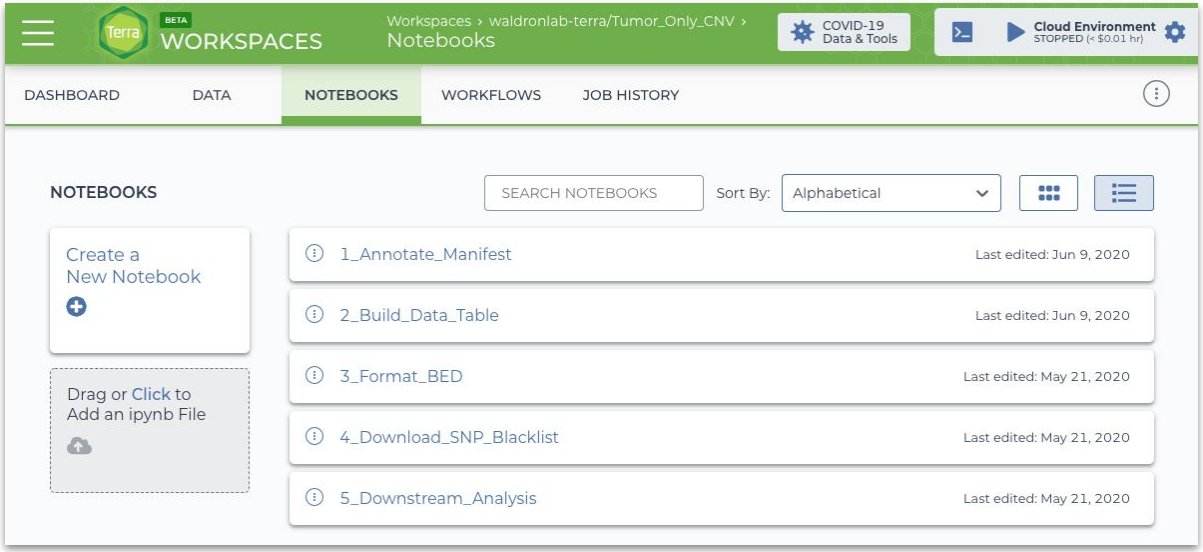
Data

- 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'



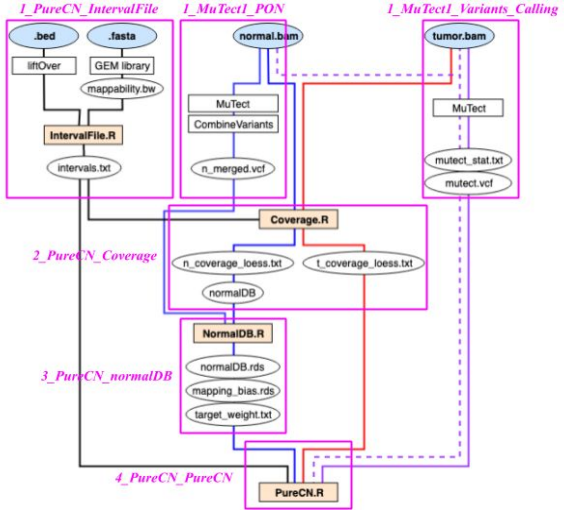
Notebooks

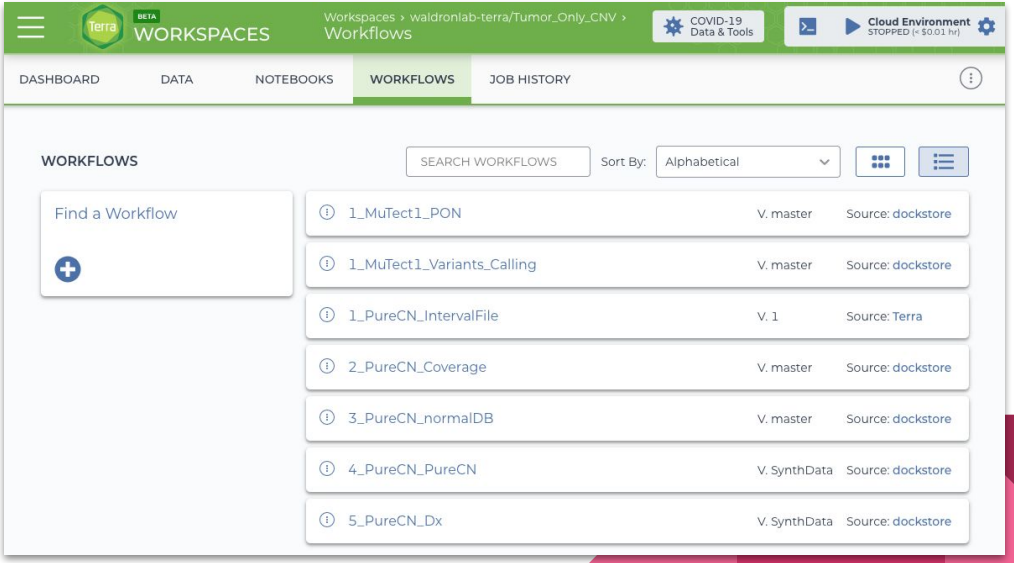
- 5 Jupyter notebooks written in R → 4 for data pre-processing and 1 for downstream analysis
- AnVIL package enables a direct connection between ‘Data’ and ‘Notebooks’



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.)





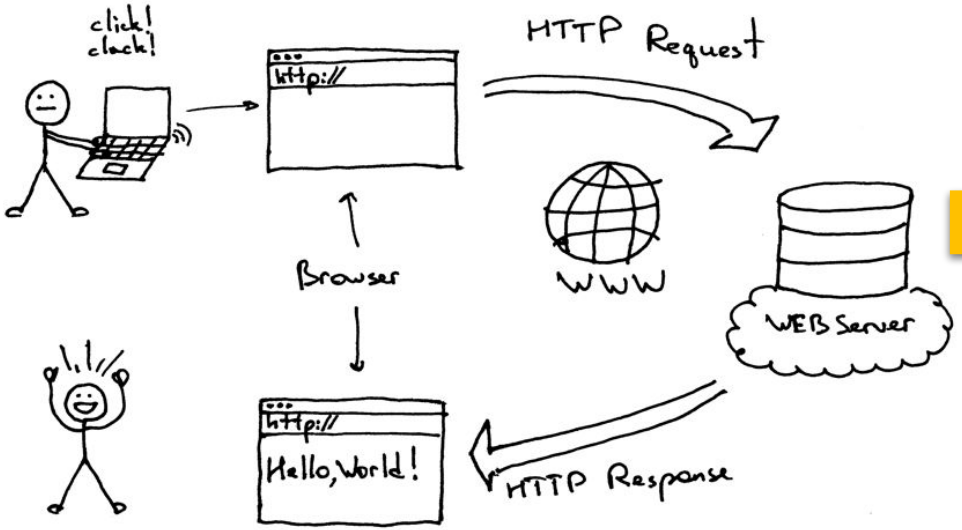
Summary

- 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)



(Slide credit to Sean Davis)

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

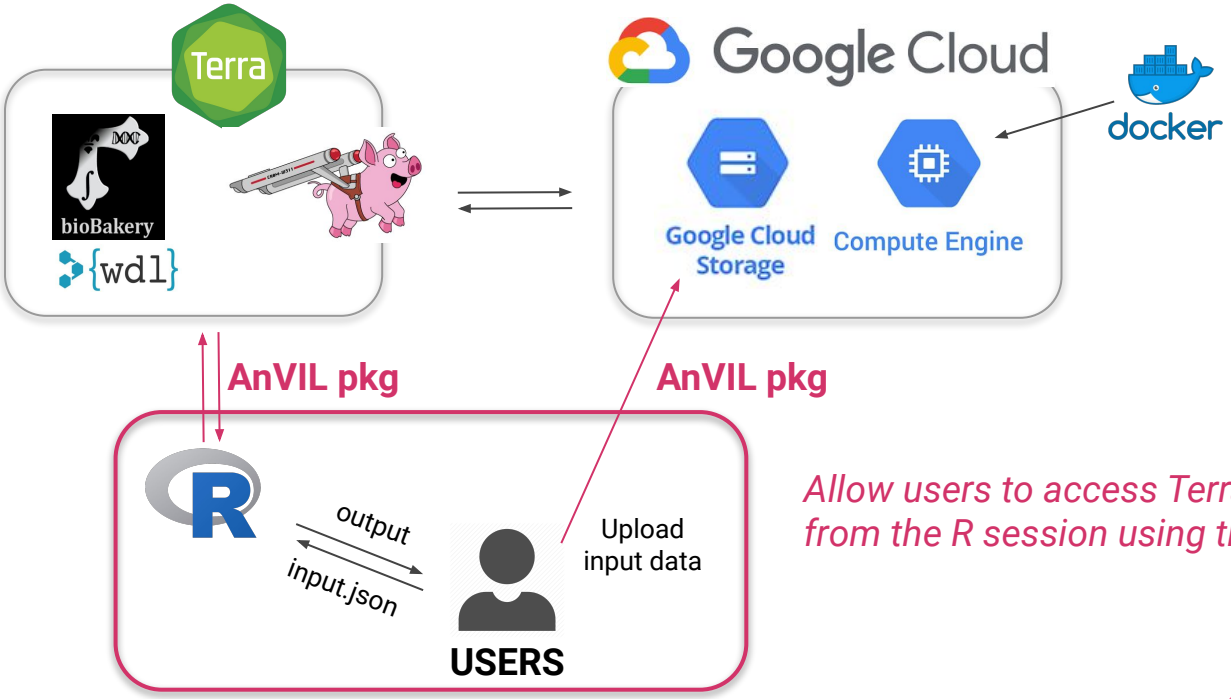
```
> src <- "gs://biobaker/"
> gsutil_ls(src)
[1] "gs://biobaker/ibdmdb_demo_metadata_test.txt" "gs://biobaker/ibdmdb_file_list_test.txt"
> pathToFastq <- "gs://biobaker/ibdmdb_file_list_test.txt"
> read.table(gsutil_pipe(pathToFastq), sep = "\t")
                                                                    V1
1 gs://fc-7130738a-5cde-4238-b00a-e07eba6047f2/IBDMDB/HSM7J4NY_R1.fastq.gz
2 gs://fc-7130738a-5cde-4238-b00a-e07eba6047f2/IBDMDB/HSMA33OT_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

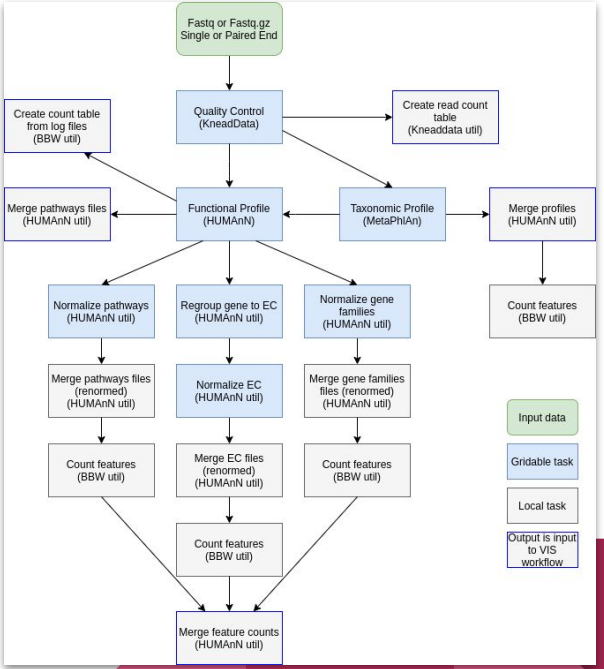


Allow users to access Terra and GCP resources from the R session using their own laptop !!

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 succesfully 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
```

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
```

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-d85538f85502/87adddce-5f43-40b0-a5a1-f7a
4-848f-bbbe6f46de64/call-FunctionalProfile/shard-0/cacheCopy/HSM7J4NY_genefamili
Copying gs://fc-071d1d53-e186-44ad-8951-d85538f85502/87adddce-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 familar 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-rpci-anvil/Bioconductor-Workflow-DESeq2
- Tumor_Only_CNV workspace : https://app.terra.bio/#workspaces/waldronlab-terra/Tumor_Only_CNV
- bioBakery : https://huttenhower.sph.harvard.edu/biobakery_workflows/
- bioBakery workspace (contact me to access): 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-mtu6MlitdX8VB7Bx1k~FLg
- BioC-AnVIL project website : 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
- Reference book : https://www.amazon.com/Genomics-Cloud-GATK-Spark-Docker/dp/1491975199