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?

Cloud computing for genomic data analysis and collaboration

Ben Langmead¹ and Abhinav Nellore²

Abstract | Next-generation sequencing studies based on large sequence data have become commonplace in genomics research. Here we review computational tools and 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.
Doubling every 18 months...

- 0.625 to 1.25 PB in 9.1 months
- 1.25 to 2.5 PB in 16.8 months
- 2.5 to 5 PB in 19.2 months
- 5 to 10 PB in 11.3 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).
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>
<thead>
<tr>
<th>Name</th>
<th>Website</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Genomes Project (1KGP)(^{102})</td>
<td><a href="http://www.internationalgenome.org">www.internationalgenome.org</a></td>
<td>This project includes whole-genome and exome sequencing data from 2,504 individuals across 26 populations</td>
</tr>
<tr>
<td>Cancer Cell Line Encyclopedia (CCLE)(^ {115})</td>
<td>portals.broadinstitute.org/ccle</td>
<td>This resource includes data spanning 1,457 cancer cell lines</td>
</tr>
<tr>
<td>Encyclopedia of DNA Elements (ENCODE)(^ {93})</td>
<td><a href="http://www.encodeproject.org">www.encodeproject.org</a></td>
<td>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</td>
</tr>
<tr>
<td>Genome Aggregation Database (gnomAD)(^ {13})</td>
<td>gnomad.broadinstitute.org</td>
<td>This resource entails coverage and allele frequency information from over 120,000 exomes and 15,000 whole genomes</td>
</tr>
<tr>
<td>Genotype–Tissue Expression (GTEx) Portal(^ {15,16})</td>
<td>gtexportal.org</td>
<td>This effort has to date performed RNA sequencing or genotyping of 714 individuals across 53 tissues</td>
</tr>
<tr>
<td>Global Alliance for Genomics and Health (GA4GH)(^ {98})</td>
<td>genomicsandhealth.org</td>
<td>This consortium of over 400 institutions aims to standardize secure sharing of genomic and clinical data</td>
</tr>
<tr>
<td>International Cancer Genome Consortium (ICGC)(^ {14})</td>
<td>icgc.org</td>
<td>This consortium spans 76 projects, including TCGA</td>
</tr>
<tr>
<td>Million Veterans Program (MVP)(^ {19})</td>
<td><a href="http://www.research.va.gov/mvp">www.research.va.gov/mvp</a></td>
<td>This US programme aims to collect blood samples and health information from 1 million military veterans</td>
</tr>
<tr>
<td>Model Organism Encyclopedia of DNA Elements (modENCODE)(^ {15,55})</td>
<td><a href="http://www.modencode.org">www.modencode.org</a></td>
<td>The goal of this effort is to identify functional elements of the <em>Drosophila melanogaster</em> and <em>Caenorhabditis elegans</em> genomes using a gamut of sequencing assays</td>
</tr>
<tr>
<td>Precision Medicine Initiative (PMI)(^ {19})</td>
<td>allofus.nih.gov</td>
<td>This US programme aims to collect genetic data from over 1 million individuals</td>
</tr>
<tr>
<td>The Cancer Genome Atlas (TCGA)(^ {136})</td>
<td>cancergenome.nih.gov</td>
<td>This resource includes data from 11,350 individuals spanning 33 cancer types</td>
</tr>
<tr>
<td>Trans-Omics for Precision Medicine (TOPMed)(^ {17})</td>
<td><a href="https://www.nhlbiwgs.org">https://www.nhlbiwgs.org</a></td>
<td>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</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>Platform</th>
<th>Hosted Data</th>
<th>Analysis Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terra</td>
<td>CCDG, eMERGE, TCGA, TARGET, TOPMed, etc.</td>
<td>WDL, Notebook, RStudio, Galaxy</td>
</tr>
<tr>
<td>Seven Bridges</td>
<td>TOPMed</td>
<td>CWL, Notebook, RStudio</td>
</tr>
<tr>
<td>Seven Bridges</td>
<td>TCGA, TARGET, ICGC, etc.</td>
<td>CWL, Notebook, RStudio</td>
</tr>
<tr>
<td>ISB-CGC</td>
<td>TCGA, TARGET, etc.</td>
<td>GCP tools (e.g. Google BigQuery)</td>
</tr>
</tbody>
</table>
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?
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

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Genomes High Coverage</td>
<td><img src="image" alt="1000 Genomes High Coverage" /></td>
</tr>
<tr>
<td>1000 Genomes Low Coverage</td>
<td><img src="image" alt="1000 Genomes Low Coverage" /></td>
</tr>
<tr>
<td>AMP Parkinson's Disease</td>
<td><img src="image" alt="AMP Parkinson's Disease" /></td>
</tr>
<tr>
<td>Baseline Health Study</td>
<td><img src="image" alt="Baseline Health Study" /></td>
</tr>
<tr>
<td>CCDG presented by NHGRI AnVIL</td>
<td><img src="image" alt="CCDG presented by NHGRI AnVIL" /></td>
</tr>
<tr>
<td>CMG presented by NHGRI AnVIL</td>
<td><img src="image" alt="CMG presented by NHGRI AnVIL" /></td>
</tr>
<tr>
<td>ENCODE Project</td>
<td><img src="image" alt="ENCODE Project" /></td>
</tr>
<tr>
<td>Broad Dataset Workspace Library</td>
<td><img src="image" alt="Broad Dataset Workspace Library" /></td>
</tr>
<tr>
<td>Framingham Heart Study Teaching Dataset</td>
<td><img src="image" alt="Framingham Heart Study Teaching Dataset" /></td>
</tr>
<tr>
<td>Human Cell Atlas</td>
<td><img src="image" alt="Human Cell Atlas" /></td>
</tr>
<tr>
<td>Neuroscience Multi-Omic Archive</td>
<td><img src="image" alt="Neuroscience Multi-Omic Archive" /></td>
</tr>
<tr>
<td>Therapeutically Applicable Research to Generate Effective Treatments (TARGET) presented by the National Cancer Institute</td>
<td><img src="image" alt="TARGET" /></td>
</tr>
<tr>
<td>The Cancer Genome Atlas Presented by the National Cancer Institute</td>
<td><img src="image" alt="The Cancer Genome Atlas" /></td>
</tr>
<tr>
<td>TopMed presented by NHLBI BioData Catalyst</td>
<td><img src="image" alt="TopMed presented by NHLBI BioData Catalyst" /></td>
</tr>
<tr>
<td>UK Biobank</td>
<td><img src="image" alt="UK Biobank" /></td>
</tr>
</tbody>
</table>
Workflows

- Currently, WDL (Workflow Description Language) is the only workflow language supported by Terra. (Cromwell, the execution engine, can take WDL and CWL (Common Workflow Language))

- Dockstore: a large collection of pre-built WDL workflows
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**

Cloud environment 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 $0 GB persistent disk to keep your data even after you delete your compute
  - Learn more about Persistent disks and where your disk is mounted

Create custom environment

**Custom Environment**

- e.g. select RStudio

- Application configuration
  - Default: (GATK 4.1.4.1, Python 3.7.9, R 4.0.3)
  - What's installed on this environment?
  - Updated Nov 9, 2020
  - Version: 1.0.11

- Cloud compute profile
  - CPUs: 4
  - Memory (GB): 15

- Startup script
  - URI

- Compute type
  - Standard VM

- Persistent disk size (GB)
  - Store your analysis data. Learn more about Persistent disks and where your disk is mounted
  - $0

Create
Costs for GCP resources

Hourly cost for custom environments

<table>
<thead>
<tr>
<th>Virtual CPUs</th>
<th>Memory</th>
<th>Price (USD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.75GB</td>
<td>$0.04749975</td>
</tr>
<tr>
<td>2</td>
<td>7.5GB</td>
<td>$0.0949995</td>
</tr>
<tr>
<td>4</td>
<td>15GB</td>
<td>$0.189999</td>
</tr>
<tr>
<td>8</td>
<td>30GB</td>
<td>$0.379998</td>
</tr>
<tr>
<td>16</td>
<td>60GB</td>
<td>$0.759996</td>
</tr>
<tr>
<td>32</td>
<td>120GB</td>
<td>$1.519992</td>
</tr>
<tr>
<td>64</td>
<td>240GB</td>
<td>$3.039984</td>
</tr>
</tbody>
</table>

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

**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 streamed in the previous vignette Manage 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.

```r
library(AnVILBulkRNASeq)
```

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

```r
install.packages("Bioconductor/DESeq2")
```

**Creating the DESeq2 dataset**
The files that are needed for the `DESeq2` analysis are the `qual.txt` files for each sample. We create the path to these files (for each sample) and save them to a file.

```r
file_path <- paste0("genew", ",/RNASeqRNASeq\1\qual.txt")
```

A simple object is needed for the analysis so we download the `GTFF` file associated with Arabidopsis thaliana and run `makeGFF3FileGTF()` on the downloaded file.

```r
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

ABOUT THE WORKSPACE

Reliable analysis of tumor CNV/SNV without matching normal

This workspace provides a fully reproducible example of copy number variation (CNV) and single nucleotide variants (SNV) analysis of tumor samples without matching normal profile, described in the recent publication [link].

Reliable analysis of clinical tumor-only whole exome sequencing data

Synthetic dataset from BioIT-Hackathon-2019-Synthetic-Data-Team workspace is loaded in this workspace. Data model for current workspace is based on synthetic dataset, so please modify input/output attributes based on the data model of your own datasets.

Overview

Allele-specific copy number alteration (CNA) analysis is essential to study the functional impact of single nucleotide variants (SNV) and the process of tumorigenesis. Most commonly used tools in the field rely on high quality genome-wide data with matched normal profiles, limiting their applicability in clinical settings.

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’
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)
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)
> 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/HSM74NY_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.

```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

https://huttenhower.sph.harvard.edu/biobakery_workflows/
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

Launch workflow

List outputs

Download outputs
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/](https://gen3.org/)
- Dockstore: [https://dockstore.org/](https://dockstore.org/)
- 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)
- BioC-AnVIL Slack Channel: [https://join.slack.com/share/zt-k04vu3kl-mtu6MlitdX8VB7Bx1k~FLg](https://join.slack.com/share/zt-k04vu3kl-mtu6MlitdX8VB7Bx1k~FLg)
- BioC-AnVIL project website: [https://bioconductor.github.io/AnVIL_Admin/](https://bioconductor.github.io/AnVIL_Admin/)
- biobakeR: [https://github.com/shbrief/biobakeR](https://github.com/shbrief/biobakeR)
- Get $300 Google credits: [https://support.terra.bio/hc/en-us/articles/360046295092](https://support.terra.bio/hc/en-us/articles/360046295092)
- Contact for an inquiry on BioC-AnVIL credit: Sehyun.Oh@sph.cuny.edu