



Check for updates

SOFTWARE TOOL ARTICLE

AnVILWorkflow: A runnable workflow package for Cloud-implemented bioinformatics analysis pipelines

[version 1; peer review: 2 approved]

Sehyun Oh^{1,2}, Kai Gravel-Pucillo^{1,2}, Marcel Ramos^{1,2}, Michael C. Schatz^{3,4}, Sean Davis⁵, Vincent Carey^{id}⁶, Martin Morgan⁷, Levi Waldron^{id}^{1,2}

¹Institute for Implementation Science in Population Health, City University of New York School of Public Health, New York, New York, USA

²Department of Epidemiology and Biostatistics, City University of New York School of Public Health, New York, New York, USA

³Department of Biology, Johns Hopkins University, Baltimore, Maryland, USA

⁴Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA

⁵Departments of Biomedical Informatics and Medicine,, University of Colorado Anschutz School of Medicine, Denver, Colorado, USA

⁶Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA

⁷Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, New York, USA

V1 First published: 21 Oct 2024, 13:1257
<https://doi.org/10.12688/f1000research.155449.1>
Latest published: 21 Oct 2024, 13:1257
<https://doi.org/10.12688/f1000research.155449.1>

Abstract

Advancements in sequencing technologies and the development of new data collection methods produce large volumes of biological data. The Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) provides a cloud-based platform for democratizing access to large-scale genomics data and analysis tools. However, utilizing the full capabilities of AnVIL can be challenging for researchers without extensive bioinformatics expertise, especially for executing complex workflows. We present the *AnVILWorkflow* R package, which enables the convenient execution of bioinformatics workflows hosted on AnVIL directly from an R environment. *AnVILWorkflow* simplifies the setup of the cloud computing environment, input data formatting, workflow submission, and retrieval of results through intuitive functions. We demonstrate the utility of *AnVILWorkflow* for three use cases: bulk RNA-seq analysis with *Salmon*, metagenomics analysis with *bioBakery*, and digital pathology image processing with *PathML*. The key features of *AnVILWorkflow* include user-friendly browsing of available data and workflows, seamless integration of R and non-R tools within a reproducible analysis pipeline, and accessibility to scalable computing resources without direct management overhead. *AnVILWorkflow* lowers the barrier to utilizing AnVIL's resources, especially for exploratory analyses or bulk processing with established workflows. This

Open Peer Review

Approval Status



	1	2
version 1 21 Oct 2024	 view	 view
1. Fabricio Almeida-Silva ^{id} , Ghent University, Ghent, Belgium		
2. Zachary S.L Foster ^{id} , USDA Agricultural Research Service, Corvallis, USA		

Any reports and responses or comments on the article can be found at the end of the article.

empowers a broader community of researchers to leverage the latest genomics tools and datasets using familiar R syntax. This package is distributed through the Bioconductor project (<https://bioconductor.org/packages/AnVILWorkflow>), and the source code is available through GitHub (<https://github.com/shbrief/AnVILWorkflow>).

Keywords

Cloud computing, Genomics, Workflows, R/Bioconductor, AnVIL



This article is included in the **Bioconductor** gateway.

Corresponding author: Sehyun Oh (Sehyun.Oh@sph.cuny.edu)

Author roles: **Oh S:** Conceptualization, Formal Analysis, Investigation, Methodology, Software, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Gravel-Pucillo K:** Data Curation, Software; **Ramos M:** Methodology, Software, Validation; **Schatz MC:** Conceptualization, Writing – Review & Editing; **Davis S:** Conceptualization, Writing – Review & Editing; **Carey V:** Conceptualization, Methodology, Supervision; **Morgan M:** Conceptualization, Methodology, Resources, Software; **Waldron L:** Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: SO, KG, MR, MS, VC, MM, and LW were supported by the National Human Genome Research Institute (U24HG010263).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 Oh S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Oh S, Gravel-Pucillo K, Ramos M *et al.* **AnVILWorkflow: A runnable workflow package for Cloud-implemented bioinformatics analysis pipelines [version 1; peer review: 2 approved]** F1000Research 2024, 13:1257 <https://doi.org/10.12688/f1000research.155449.1>

First published: 21 Oct 2024, 13:1257 <https://doi.org/10.12688/f1000research.155449.1>

Introduction

The NHGRI's Genomic Data Science *Analysis, Visualization, and Informatics Lab-space* (AnVIL) consortium was launched in 2018, aiming to democratize genomics data.¹ AnVIL enables easy sharing of genomics data by organizing databases, bioinformatics pipelines for large-scale data processing, and interactive downstream analysis in one Cloud-based platform. AnVIL,² also the name of the platform from the AnVIL project, implements the FAIR data-sharing philosophy and provides a graphical user interface (GUI, supported by Terra³), making it more accessible for researchers without programming backgrounds. However, a GUI tends to be less efficient and slower than a command line interface (CLI), especially for bulk analyses, still requires learning a new platform, and does not support version control and text-based workflows, often included as best practices for reproducible computational research.⁴

Bioconductor's *AnVIL* package is an AnVIL API wrapper that provides R-friendly, programming-based functionalities to leverage flexible and scalable cloud-based resources implemented in the AnVIL platform. With the *AnVIL* package, users can easily access workflows, data, and Cloud-based computing resources managed by AnVIL. However, the AnVIL package is not customized for workflow execution tasks. Instead, *AnVIL* covers all the resources related to the AnVIL platform, such as interaction with the repository for Docker-based genomic analysis tools and workflows (Dockstore⁵), leveraging cloud resources (Leonardo⁶), and data search and digestion (Gen3⁷). Many *AnVIL* functions also expose API commands directly, requiring a deep understanding of the underlying AnVIL workspace structures and data models to use for workflow execution. Also, it is a general package without individual support on any workspace and provides no metadata curation. Because most Bioconductor users focus on data analysis, a convenient R-friendly way of accessing and utilizing AnVIL resources is needed. Here, we present the *AnVILWorkflow* package to meet this need. *AnVILWorkflow* package is a convenient, fit-for-purpose wrapper around the *AnVIL* package with the following features optimized for workflow execution:

- Intuitive function names and returned values
- Support workflow-specific documentations
- Enable to set up a Cloud environment with a single function call
- Return error messages that are easy to interpret and actionable
- Essential metadata curation for more efficient data browsing

Users can apply *AnVILWorkflow* on any workspace they can access, including 347 public workspaces (snapshot on 8.28.23) available to anyone with an AnVIL account. We present three use cases where we ran non-R-based bioinformatics analysis tools using conventional R syntax: *Salmon*,⁸ *bioBakery*,⁹ and *PathML*.¹⁰ *Salmon* is a widely used RNA sequencing analysis tool for quantifying the expression of transcripts and is based on the command-line interface. Its downstream analysis involves many R/Bioconductor packages, such as *DESeq2*, *edgeR*, and *limma*. *bioBakery* is a widely used whole metagenomic shotgun (WMS) sequencing data analysis environment, mainly relying on Python. *PathML* is a general-purpose research toolkit for computational pathology, including many functionalities in digital pathology data analysis, such as strain normalization, nucleus segmentation, and tissue detection. *PathML* takes raw image files and returns the processed image data in an hdf5 format for further downstream analysis, including machine learning methods.

AnVIL provides comprehensive resources for biomedical data analysis, including data (e.g., genomics), workflows for bulk analysis, and interactive analysis apps (i.e., Galaxy, Jupyter Notebooks, and RStudio) under the workspace. Workflows are often a limiting factor in bioinformatics analysis due to computing demands and the bioinformatics expertise required. Thus, the *AnVILWorkflow* package makes the workflow-related resources from AnVIL more accessible and easier to use, especially for R users (Figure 1).

While AnVIL manages workflow orchestration and workspace metadata and provides default setups that simplify decision-making for users, users still need to manage their data storage and cloud costs. Genomics data, especially their raw and intermediate forms, are very large, so data storage can be costly if the sample size increases. Storage costs incur and can be managed in two ways - storage itself and transfer. For example, using regional storage instead of multi-region, cleaning up intermediate results, and storing infrequently accessed data in low-cost storage (e.g., nearline or coldline storage from Google Cloud) can reduce per-sample costs. Analyzing data stored in one region using Virtual Machine (VM) compute resources in a different region incurs data transfer charges, so centralizing all storage and computing in a single region can be more cost-efficient by not only reducing the storage cost but also avoiding data transfer charges.

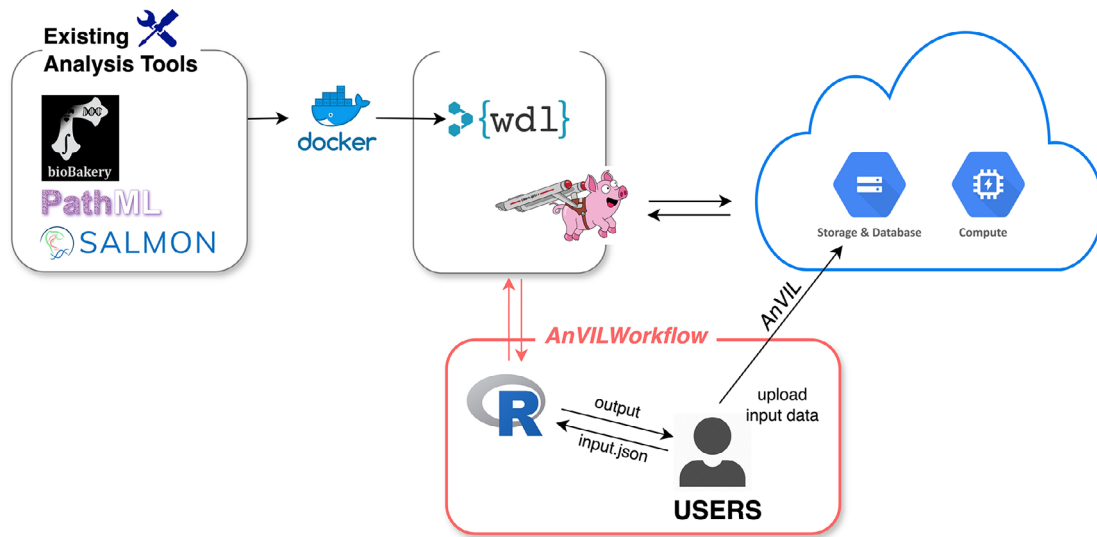


Figure 1. Overview of AnVILWorkflow package. AnVIL's workflow description language (WDL) specified the runtime environment, which includes proper docker containers for existing analysis tools and computing resources. Cromwell, a scientific workflow execution and management system, runs WDL workflows on the cloud. *AnVILWorkflow* users can easily run established workflows developed by experts and utilize the cloud resources without configuring or taking maintenance responsibilities.

Currently, the AnVIL workspaces use the us-central1 as a default region, and any artifacts generated from the workflow execution, unless specified, are saved in the same-region bucket linked to the workspace. If users use the default region configured by AnVIL, bringing their data stored in the default region, us-central1, will save the data transfer charge. Additionally, open and controlled access genomic datasets hosted in AnVIL are stored in the us-multi-region, so there are no storage and transfer charges for users using the default workspace configuration. Downloading data to the user's workstation or laptop is subject to charges, currently \$0.08 to \$0.12 per GB, depending on the amount of data¹¹ and geography of the transfer, and transfer from the US to another continent is more expensive than within the US transfer.

While browsing existing resources through *AnVILWorkflow* is free, running workflows charge computing costs. *AnVILWorkflow* is designed to use existing workflows, which usually predefine computing resources optimized for the types of analyses, simplifying computing-related cost management. You can further reduce the run cost using call caching and preemptive instances. For example, if your workflow runs in fewer than 24 hours since a preemptible VM lasts 24 hours at most, you can save up to 80% by using preemptible VMs.

The cost management for a group of users can be efficiently managed through the AnVIL billing project. One billing account can be shared by adding email addresses under the billing project. The billing project offers details on each workspace, including workspace owner and spent reports, so we can easily identify 'who' uses 'how much' for 'what'. In addition to the workspace-level expense reports, users can further enhance cost monitoring by configuring spending reporting.¹² This allows users to closely monitor the expenditure associated with each workflow execution.

Methods

Implementation

AnVILWorkflow package provides all the functionalities required to run workflows available in AnVIL from the local R session - from the environment setup to the output download. One prerequisite is to create an AnVIL account from the AnVIL web portal. AnVIL account provides two required inputs to run workflows remotely: 1) the email address associated with the user's account and 2) the billing project name to cover the computing cost.

AnVIL-hosted workflows can be run using four main functions: `setCloudEnv`, `cloneWorkspace`, `runWorkflow`, and `getOutput`. The `setCloudEnv` function accepts the AnVIL account email and billing project name and sets up your local R environment so you can access AnVIL and Cloud-computing resources. The `cloneWorkspace` function creates the user's copy of a 'template' workspace, and the `runWorkflow` executes the workflow. The `getOutput` function can check the outputs from successfully executed workflows and download user-specified files to a local computer.

User input can be provided through the `updateInput` function, which accepts two different forms of tables depending on the workflows - AnVIL's data model or URLs pointing to data files stored in Google Cloud buckets. The input data formats are already specified in the workflow scripts (Workflow Description Language, WDL¹³). Other accessory functions are available to monitor submission progress (`monitorWorkflow`), stop submitted workflow (`stopWorkflow`), and get Dashboard content (`getDashboard`).

The `AnVILBrowse` function allows users to browse AnVIL resources using keywords. This function runs instantaneously because the *AnVILWorkflow* package includes the metadata snapshot on all the publicly accessible AnVIL workspaces and their workflows and data. It performs basic metadata harmonization, allowing more efficient browsing and filtering, such as selecting workspaces based on the study size or participants' ages. Users can also browse non-public workspaces they have access to using the `getMetaTables` function; however, this process can take a while, depending on the number of workspaces a user has access to.

Operation

The use cases demonstrated below include demo input data in the template workspaces, so the R scripts below can run the listed use cases from the local computer. Ready-to-run examples that can be used to test the process on the user's own AnVIL account are available in the *AnVILWorkflow* package vignette. Genome Analysis Toolkit Variant Discovery (GATK) best-practice pipelines¹⁴ are not demonstrated here but are also available as AnVIL workspaces.

```
## Setup the account
setCloudEnv(accountEmail = {AnVIL account email},
             billingProjectName = {AnVIL billing project name})

## Clone the workspace of your interest
newName <- {Unique name for your copy of workspace}
cloneWorkspace(workspaceName = newName, templateName = templateName)

## Run workflow
runWorkflow(workspaceName = newName,
             workflowName = {name of the workflow if there is more than one
                             in the workspace of your interest})

## Get workflow outputs
getOutput(workspaceName = newName)
```

The main features of the demo workspaces and their workflow-specific input data preparation process are described below.

Results

Use cases

Bulk RNA sequencing data analysis

Salmon workflow uses AnVIL's data model and requires four essential inputs - `fastq1`, `fastq2`, `fasta`, and `transcriptome index name`. This workflow can be easily applied to the consortium data hosted in AnVIL, which follows AnVIL's data model. With the default runtime environment configured for this workflow (1 CPU, 2GB memory, and 10GB SSD disk), processing 16 demo samples (32 fastq files, ~1 GB per file) took about 30 minutes and cost \$0.12.

Whole metagenomic shotgun data analysis

bioBakery is a metagenome analysis environment composed of Python-based tools, reference databases, and command-line-based workflows. It processes raw shotgun sequencing data into microbial community feature profiles, summary reports, and figures.⁹ *bioBakery*'s whole metagenome shotgun (`wmgx`) and visualization (`wmgx_vis`) workflows are implemented as an AnVIL workspace. The current version of the *AnVILWorkflow* supports *bioBakery* version 3.¹⁵ While users can customize this workflow to a great degree, only six inputs are sufficient to run a standard, optimized version of this workflow. Those six inputs are:

- Name of the Trimmomatic adaptor type (for demo data, *NexteraPE*)
- Your project name
- Extension of input files (for demo data, `.fastq.gz`)
- A table of your sequencing file (fastq) names stored in the Google Cloud Storage bucket
- Input file identifier for paired-end sequencing (for demo data, `_R1` and `_R2`)

The seven required databases are already linked to this workflow, and nine additional optional inputs are available for further customization. Optional inputs are for workflow customization, such as bypassing functional profiling (default is false) and maximum memory usage for different tasks (default is 32GB for functional profiling by *HUMAnN*, 8GB for quality control by *Kneaddata*, and 24GB for taxonomic profiling by *MetaPhlAn*). This workflow uses call caching and preemptive instances by default for cost efficiency. Processing six paired-end demo samples (mean file size ~380MB) with the optimized default setting without using preemptive instances took about 5 hours and cost around \$6.50. With the preemptive instances, it can take longer but cost less. Compared to the existing options, such as *Nephele*,¹⁶ *AnVILWorkflow* allows a programmatic approach and more flexible customization options.

Histopathology image processing using PathML

We implemented the hematoxylin-eosin (HE) stain normalization process of *PathML* as an *AnVIL* workspace. This workflow accepts an SVS file as input and returns original and normalized images as PNG files. There are two required inputs - Google Cloud Storage URI, where the input SVS image file is stored, and the sample name. Processing one publicly available image (CMU-1_Small_Region.svs, 1.8MB)¹⁷ with the default runtime (4 CPU, 16GB memory) took about 8 minutes and cost \$0.01. This simple but robust analysis setup can support clinical use cases, such as pathologists who process a large number of images in a short time, by offering guidance and cross-validation options.

Discussion

The *AnVILWorkflow* package enables users to conduct complex and computationally intense analyses with minimal bioinformatics expertise through well-established workflows within *AnVIL* and versatile cloud resources directly from standard laptops using the familiar R syntax. The major advantages *AnVILWorkflow* provides over the existing approaches include 1) a minimal entry barrier, negating the need for software installations, preparation of properly versioned reference data, or construction and oversight of workflows, 2) leveraging flexible cloud computing resources without the need to learn or handle them directly, 3) user-friendly functions that provide enhanced information, and 4) improved reproducibility and interoperability by seamlessly linking multiple analysis steps, conducted in both R and non-R based tools, within a single R vignette. However, there are still some limitations. For instance, certain customizations of the workflows are limited or require a more profound understanding of the workflows. Despite not being inherently more costly than an in-house server, the pay-per-use structure requires careful planning and management. The absence of an integrated versioning system in *AnVIL* workspaces requires users to manually monitor new versions. In conclusion, *AnVILWorkflow* proves most advantages for analyzing a bulk of samples on relatively simple workflows (i.e., single-stage workflow procedure) or for exploratory data analysis for non-technical users, particularly when employing well-established analysis workflows.

Ethics and consent

Ethical approval and consent were not required.

Authors' contributions

SO, LW, and VC conceived the research idea. SO, KG, MR, and MM developed the software. SO and KG performed the benchmarking analyses. SO, LW, and KG wrote the manuscript. LW, SD, MS, and MR reviewed the manuscript.

Data availability

- Figshare: Test datasets for *bioBakery* and *PathML* workflows; <https://doi.org/10.6084/m9.figshare.27018421.v3>¹⁸

The project contains the following underlying data:

- IBDMDB: six pairs of WMS sequencing files and their sample-level metadata
- PathML_data: one input and two output files
- The test datasets for bulk RNAseq analysis workflow have been deposited in the European Nucleotide Archive (ENA); accession numbers are DRR016125-DRR016140; <https://www.ebi.ac.uk/ena/browser/view/PRJDB2508>¹⁹
- License: [Creative Commons Attribution 4.0 International](#) (CC BY 4.0)

Software availability

- Source code available from: <https://github.com/shbrief/AnVILWorkflow>
- Software available from: <https://www.bioconductor.org/packages/release/bioc/html/AnVILWorkflow.html>
- Archived software available from DOI: [10.5281/zenodo.13868810](https://doi.org/10.5281/zenodo.13868810)²⁰
- Software license: Artistic-2.0

Acknowledgments

Not applicable.

References

- Schatz MC, *et al.*: Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genom.* 2022; **2**.
- Terra. [Reference Source](#)
- Terra. [Reference Source](#)
- Sandve GK, Nekrutenko A, Taylor J, *et al.*: simple rules for reproducible computational research. *PLoS Comput. Biol.* 2013; **9**: e1003285. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yuen D, *et al.*: The Dockstore: enhancing a community platform for sharing reproducible and accessible computational protocols. *Nucleic Acids Res.* 2021; **49**: W624–W632. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Leonardo: Notebook Service: (Github).
- Hughes L, *et al.*: Harmonization of clinical data across Gen3 data commons. *J. Clin. Orthod.* 2019; **37**: e18094–e18094.
- Patro R, Duggal G, Love MI, *et al.*: Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods.* 2017; **14**: 417–419. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- McIver LJ, *et al.*: bioBakery: a meta'omic analysis environment. *Bioinformatics.* 2018; **34**: 1235–1237. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rosenthal J, *et al.*: Building Tools for Machine Learning and Artificial Intelligence in Cancer Research: Best Practices and a Case Study with the PathML Toolkit for Computational Pathology. *Mol. Cancer Res.* 2022; **20**: 202–206. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Pricing: Google Cloud. [Reference Source](#)
- How much did my workflow cost? Terra Support. [Reference Source](#)
- Voss K, Gentry J, Van der Auwera G: Full-stack genomics pipelining with GATK4 + WDL + Cromwell. 2017. Preprint at [Publisher Full Text](#)
- Van Der Auwera GO, Connor BD: *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media; 2020.
- Beghini F, *et al.*: Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *elife.* 2021; **10**. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Weber N, *et al.*: Nephele: a cloud platform for simplified, standardized and reproducible microbiome data analysis. *Bioinformatics.* 2018; **34**: 1411–1413. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Aperio SVS. [Reference Source](#)
- Oh S: Test datasets for the AnVILWorkflow package. *figshare.* 2024. [Publisher Full Text](#)
- EMBL-EBI: ENA Browser. [Reference Source](#)
- Oh S: AnVILWorkflow. *Zenodo.* 2024. [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 11 December 2024

<https://doi.org/10.5256/f1000research.170635.r342030>

© 2024 Foster Z. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Zachary S.L Foster 

USDA Agricultural Research Service, Corvallis, OR, USA

The authors describe an R package for submitting and interacting with pipelines hosted on the Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL). The goal of this package is to make it easier for people primarily familiar with R to use AnVIL resources without learning other systems. This is primarily beneficial for those inexperienced in bioinformatics and that need to run large analyses using linux-based tools. The paper is succinct and written well. I appreciate the factual and unbiased tone of the writing. In particular, the lack of exaggeration of the significance of the work is refreshing.

The tools itself seems fairly simple and is primarily consists of “quality of life” wrappers for the AnVIL R packages, which seem to more closely follow the implementation of the underlying API. AnVILWorkflow has about 10 commands, each 50-100 lines of R code. The code is well documented and seems tidy, but there are a few minor improvements that could be made, which I mention below. I installed the package from Bioconductor and the few function I tried seemed to work, but I did not run any analyses, since this would require an account on Terra and would cost money.

Overall, this article is a nice and succinct description of a tool meant to make running analyses on Terra/AnVIL more user-friendly.

Below are specific comments for each section.

Introduction

- Define acronym NHGRI
- “Enable to set up a Cloud environment with a single function call” sounds off. Maybe “Cloud environment initialization with a single function call.”
- Should “spent reports” be “spending reports” or “expense reports”?
- “However, the AnVIL package is not customized for workflow execution tasks.”: The source code for AnVILWorkflow::runWorkflow calls AnVIL::Terra()\$createSubmission, so it seems that the AnVIL package does allow for workflow execution. Are you trying to say that it is not as convenient for workflow execution? Might want to rephrase this to make it more clear/accurate.

Results

- “Name of the Trimmomatic adaptor type”: Is this the right phrasing? Trimmomatic is the program used to remove adapters, but is not a type of adaptor. Perhaps something like “Sequencing adaptor type”?

Discussion

- “Despite not being inherently more costly than an in-house server, the pay-per-use structure requires careful planning and management.”: I like that this comparison is brought up. Cloud based analysis sounds very tempting at first, but I have used it for genomics data analysis and I am not convinced its a way to save money. In fact, I expect that in-house computing is actually cheaper in the long run. A powerful desktop can run the vast majority of bioinformatic workflows. Scaling and massive one-off analyses are the main benefits of cloud computing. Surprise costs are a big problem, especially to those that have little experience in commercial cloud computing, which is the target audience of this article. It might be a good idea to add a stronger warning about surprise costs. I also like that data egress fees were brought up earlier.
- “In conclusion, AnVILWorkflow proves most advantages for analyzing a bulk of samples on relatively simple workflows”: Is the “most advantages” phrased right? “most advantageous”? Also, the lack of exaggeration when describing the significance of the work is refreshing.

Github.io documentation

- The image on the home page does not load. Looks like a broken link.
- There are some errors that look like they came from chunks with `echo = FALSE` (e.g. “Error : ‘avworkspace_clone’ failed:” and “Error in mget(arg_names, environment): object ‘billingProjectName’ not found”). Perhaps the Rmd/Qmd needs to be fixed and rerun or make errors not printed?
- Typically there are installation instruction on the README, especially in combination with a quick start guide. I know there are instructions on Bioconductor, but the first place many people look is the Github repository README, so they should be there too. Consider adding the command to install the development version too using `devtools::install_github()`.

The source code

- The code is documented well using roxygen2 and seems tidy Overall
- The jsonlite package is used for JSON parser, but a much faster package is available called RcppSimdJson. This should be a simple drop in replacement for the most part and will make json parsing 100s of times faster and more RAM efficient. However, this might not matter if only small JSON files are being parsed.
- There are some dependencies that could be replaced with base R code, which would make maintaining the code easier: dplyr, plyr, stringr, and tidyr. Some of these might not be worth the effort to replace, but stringr for example is only used in one place to convert something to title case. The only non-trivial thing dplyr and plyr are doing are table joins and the merge function can do this in base R. `dplyr::filter` could easily be replaced with standard R subsetting. I think the tidyverse packages are great for interactive use and data analysis scripts, but not ideal as dependencies since they are large, complex, and frequently break backwards compatibility. However, there is always a balance between dependencies and development time/effort, so this should be considered on a case-by-case basis.

- Consider adding a LICENSE file to the repository. I know its in the DESCRIPTION file, but that is specific to R, and LICENSE is the standard for most software and then Github will be able to read it and display the license, which is important information for many users.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am primary a bioinformatic software developer with experience in pathogen diagnostics, metabarcoding, and developing pipelines. Of relevance to this article, I have developed R packages and pipelines that run on commercial cloud services.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 04 December 2024

<https://doi.org/10.5256/f1000research.170635.r342032>

© 2024 Almeida-Silva F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Fabricio Almeida-Silva 

VIB Center for Plant Systems Biology, Ghent University, Ghent, Belgium

In this manuscript, Oh et al present AnVILWorkflow, an R package that allows running and managing AnVIL workflows directly from an R session. This package is a valuable resource for the bioinformatics community, especially to novice bioinformaticians or researchers who are not familiar with the details of AnVIL. AnVILWorkflow has the potential of making AnVIL more widely

used and known, simplifying the analysis of biological data and helping decelerate the unnecessary growth of lab-specific bioinformatics pipelines that do (roughly) the same thing. The manuscript is well written, and it nicely describes what AnVIL is, positive and negative aspects, and how to use AnVIL workflows in R. Overall, I believe the manuscript is in great shape in its current format. I have only a few minor points that I believe could help readers:

(1) Authors mention some approaches to reduce costs, such as using call caching and pre-emptive instances. I believe it would be useful if authors could expand on that, giving clear examples (maybe using one of the three example workflows in the manuscript) on how such cost reductions can be achieved.

(2) For people who have never used AnVIL before and want to consider using it after reading this paper, I think authors could add a few sentences in the Discussion describing how to set up AnVIL in a lab. For instance, since it works on a pay-per-use basis, should a lab have a single account or should each member of the lab have its own? How should accounts be managed to reduce redundancy and costs?

(3) The whole idea of AnVIL seems very similar to Galaxy. I know the purpose of this manuscript is to describe AnVILWorkflow, not AnVIL itself, but maybe mentioning Galaxy in the Introduction and how AnVIL differs from Galaxy (highlighting advantages and disadvantages) would be useful to readers.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research