# NHLBI BioData Catalyst Community Hours Notes

## A Tour of the Analysis Workspaces with Seven Bridges and Terra

Wednesday, August 25, 2021 at 1 PM EDT

BioData Catalyst Community Hours is a monthly, hour-long event where users can learn about features of the ecosystem. The hour is split into time for presentation by a platform team and time for questions. Teams will showcase tools, new features, or tips that meet user needs. After the presentations, time is available for discussion and questions for platform reps from users.

During the community showcase we follow the NHLBI BioData Catalyst Statement of Conduct.

## NHLBI BioData Catalyst Representatives Present

| Name | Platform/Team | Role | Email |
|---|---|---|---|
| Amber Voght | BDC3 | Host | alvoght@renci.org |
| Paul Kerr | BDC3 | Co-Host | paulkerr@renci.org |
| Tony Patelunas | Seven Bridges | Co-Host | tony.patelunas@sbgenomics.com |
| Dave Roberson | Seven Bridges | Co-Host | dave.roberson@sbgenomics.com |
| Tiffany Miller | Terra | Co-Host | tiffanym@broadinstitute.org |
| Emily Hughes | PIC-SURE | Rep | emily_hughes@hms.harvard.edu |
| Simran Mikwana | PIC-SURE | Rep | simran_makwana@hms.harvard.edu |
| Sean Burke | Gen3 | Rep | svburke@uchicago.edu |
| Beth Sheets | Dockstore | Rep | esheets@ucsc.edu |

# Welcome and Agenda

**Slide 3: Agenda**

Session is focused on the analysis workspaces and requesting Pilot Credits.

- Topic presentations
    - Introduction to workspaces (10 minutes)
    - Tour of Terra (10 minutes)
    - Tour of Seven Bridges (10 minutes)
    - Requesting cloud credits and getting started on each platform (5 minutes)
- Discussion and questions
    - We have reps from all platforms to address questions.

**Slide 4: Welcome**

- Hosts and co-hosts
    - **BDC3**: Amber Voght, Paul Kerr
    - **Seven Bridges**: Dave Roberson, Tony Patelunas
    - **Terra**: Tiffany Miller
- We encourage you to submit unanswered questions, issues, and challenges - no matter how big or small - to our help desk
- Join the ecosystem: https://biodatacatalyst.nhlbi.nih.gov/contact/ecosystem

# Introduction to Workspaces

Rep: Tiffany Miller

**Slide 6: Harnessing the computing power of the cloud**

- Stop waiting in line for your university cluster: Set up workspaces on BioData Catalyst and process 1000s of samples at once using instances on AWS or Google Cloud

Within the NHLBI BioData Catalyst ecosystem you can take advantage of analysis platforms for all your data science needs without having to wait for your organization's computational resources. This means you can run 1000s of workflows on demand. The analysis platforms use cloud vendors like Google Cloud platform (GCP) and Amazon Web services (AWS) allowing you to utilize their vast data centers for running computational jobs. These two screenshots show the two platforms in the ecosystem where you can create and manage analyses in BioData Catalyst: Terra and Seven Bridges.

**Slide 7: Organize files, methods, and results in Workspaces/Projects**

- Workspaces serve as a place to organize files and tools.
- Set up and kick off analyses.
- Functionality to organize files & metadata.

You can organize your analyses in these platforms in a workspace or project. You can think of a workspace or project as a place to organize your data, configure and run analyses, and collaborate with others. Note that workspace is the term used in Terra and project is the term used in Seven Bridges, but fundamentally these computational sandboxes offer the same capabilities. It is a place where you can do all your data analysis end-to-end. By virtue of it being a platform on the cloud, you can easily collaborate with others during this process.

**Slide 8: Work privately or alongside collaborators**

- If you are the only member of a project, you are the only person who will have access to your data, tools, and analyses.
- To collaborate, add members to project. Administrative capabilities allow for granular permissions to limit what project members can see and do.

A workspace or project has granular permissions so that you can control who has access to it. It can be kept private or you can control who can view it, run analyses within it, or share it. This

means you don't have to worry about how collaborators will get access to your on-premise resources or data to collaborate on your project. You can securely give collaborators access to your workspace or project easily.

**Slide 9: Access and analyze 3.4 PB of hosted data on BioData Catalyst**

- Access data based on dbGaP permissions across the BioData Catalyst ecosystem.
- Select files and export to workspace -OR- form a cohort and work with associated files in workspace.
- Perform analysis on NHLBI's hosted data and avoid moving around large datasets.

In the ecosystem you can access and analyze 3.4 petabytes of hosted data. With your eRA Commons ID linked, particular datasets you are authorized to use in dbgap are accessible and exportable to the analysis platforms. This screenshot shows how you can export pointers to data from Gen3 to a workspace or project in Terra or Seven Bridges. By pointers to data I mean the pointers to the physical location of the files on the cloud. By pointing to the data, you won't have to pay for storage of these raw files for extended periods of time. You can make use of the files or portions of the files when you want to run your analysis. Sometimes the dataset you want to analyze isn't hosted yet. That isn't a problem though because you can bring it into the cloud yourself.

**Slide 10: Bring your own data**

- For smaller uploads, drag and drop from computer.
- For larger uploads, Command Line Upload.
- Access private cloud storage bucket from workspace and use as external file repository.

Users can upload data for which they have the appropriate approval, provided that they do not violate the terms of their Data Use Agreements, Limitations, or Institutional Review Board policies and guidelines. There are options for uploading your data to the cloud for accessibility in Terra and Seven Bridges. You can drag and drop small files via your browser or use the command line to upload large files or a set of files. If you already have data in GCP or AWS you can grant access to those buckets to your Terra or Seven Bridges account so you can use them for your analyses. Once the data is in the cloud you can start to analyze it. There are a few components required to be able to run software or tools on the cloud.

**Slide 11: Reproducible analyses using dockerized workflows**

- Find and export publicly available workflows from Dockstore.
- Docker containers allow you to package your required software/tools and easily run your workflow on any computing platform.
- With your workflow and docker, launch large jobs without waiting in a queue.

You need a dockerized workflow, but what does that mean?

First, you need to get your software or tool dependencies packaged in a way that is modular and can be reproduced on any unique virtual machine or cluster. Docker containers solve this problem. Docker containers package applications in "containers," allowing them to be portable among many systems.

Imagine when you are running a workflow on a machine in the cloud, that your Docker container is an environment running on these machines with all the software, language, and tool dependencies available. Terra and Seven Bridges will orchestrate the use of these machines from GCP or AWS, the docker container provided ensures all the technology is there, but the user must also provide the requirements and instructions for how their analysis should run.

That is the second component. You need to write your workflow in a language that the analysis platform can understand. You can use the Workflow Description Language (aka WDL) or the Common Workflow Language (CWL). Now all of that sounds like a lot of work, but luckily there is Dockstore. Dockstore, another product in the BioData Catalyst ecosystem, is sort of like a repository housing popular workflows and pointers to their corresponding docker images. You can make use of  common workflows so you don't have to recreate the wheel or view them to get ideas for writing your own pipeline. Dockstore is pictured here. Dockstore defines itself as an "open platform used by the GA4GH for sharing Docker-based tools described with either the Common Workflow Language (CWL), the Workflow Description Language (WDL), or Nextflow (NFL). Dockerized workflows come packaged with all of their requirements."

You can export workflows & tools to your workspace or project and run them in batches on several samples at once or whatever data type you are working with. Maybe you don't need to run workflows and just want to visualize your data quickly, compute statistics, or use another interactive app. This is known as an interactive analysis and is another key feature of the analysis platforms workspaces.

**Slide 12: Interactive Analysis**

- Spin up an interactive environment to analyze outputs of your workflows or any other data in the cloud.
- Built-in applications include Jupyter Notebooks and RStudio.

Within your workspace or project you can spin up a Jupyter Notebooks or Rstudio environment to further analyze your data. There are other apps available depending on the analysis platform that you can use as well.

If you haven't used them before, Jupyter notebooks are an open-source app that runs in a browser. They contain rich text commentary as well as code cells. They can execute any Python or R-based commands on data in real-time.

When you send a copy of a notebook with code cells that have been run to collaborators, they can view your results embedded right in the .jpynb file. Your workspace or project can also store these notebooks so that you can easily share or collaborate on them with all the data and relevant workflows stored in the same place.

# Terra Tour

Rep: Tiffany Miller

**Slide 15: Navigation in Terra**
Three minute video on the high-level navigation in Terra: https://youtu.be/3rH86vcAqK8

**Slide 16**: **Notes on Navigation**
The homepage for BioData Catalyst Powered by Terra looks a bit different than what was displayed in the video. See the image on the left. BioData Catalyst Powered by Terra has the ecosystem branding and different links on the home page. If you sign-in through either url though, you will be able to access your workspaces created via either portal. That means if you sign-in to app.terra.bio and create a workspace, you can access it via BioData Catalyst Powered by Terra and vice versa.

Another thing to note is you don't need an eRA commons ID to register. If you do not have one, you should get that process started, but you can also start exploring BioData Catalyst Powered by Terra without it. You can create and authenticate with any Google identity you have. When you register, on your profile page you can link your NIH account and link NHLBI BioData Catalyst Framework Services to get access to restricted datasets authorized by dbGap.

**Slide 17: Notes on Navigation**

- New Featured workspaces Organization
- https://terra.biodatacatalyst.nhlbi.nih.gov/#library/showcase
- Check out the BioData Catalyst Collection workspace

We recently updated our showcase page interface and I've shared a screenshot of it here. As the video went over, this page demos featured workspaces. Featured workspaces are packaged demonstrations of an analysis or toolkit on example data. They typically include an example dataset, configured workflows w/ dockers that can run on the example data set, or configured notebooks. They always have good documentation describing what the analysis does. Some provide cost and time estimates for running the workflows. It is worth a visit to this page and so I put a link here for you. You can clone these workspaces to run an analysis you find interesting on your own data.

They are both community curated and the Data Sciences Platform at Broad curates many as well. For example there are a lot of GATK Best practice workspaces.
For community curated workspaces, we have a process for validating them to ensure they have appropriate documentation and are runnable. One way we've seen researchers use this page is to package their publication's methods/dockers in a workspace so that others can more easily reproduce their work or build off of their work.

**Slide 18: Terra Differentiators**

| | |
|---|---|
| **Workflow Language** | Workflow Description Language (WDL) |
| **Cloud Provider** | Google Cloud Platform, Azure (*coming*) |
| **Applications** | Preloaded applications and options to bring-your-own through a user-friendly interface.<br><br>Galaxy (including full community toolshed), IGV, Seqr |
| **Interactive Analysis Features** | Highly customizable machines with persistent disks set up to save your work<br><br>Bioconductor, Hail, GATK and other popular bioinformatics tools preloaded. Available in "best practices" workspaces maintained/approved by the tool developers |

Now I am going to switch gears and talk about some items that are different between Seven Bridges and Terra. This sometimes helps researchers figure out which platform to start with when they join the ecosystem. If you are hearing this information and aren't quite sure if these differentiators matter to you, we recommend you try both platforms. By both analysis platforms being a part of the ecosystem, researchers have a lot of flexibility.

To start our differentiators,  the workflow language supported by Terra is the Workflow description language. The Cloud provider Terra sits on top of is the Google Cloud Platform and we are working on building capabilities to work on Azure by Microsoft as well.

In terms of applications, there are pre-loaded apps and the option to bring your own. You can launch Galaxy, IGV, or Seqr as well. More to come on what those apps are.
In terms of Interactive analysis features, you can customize your machine's specifications like CPUs, GPUs, memory and make use of persistent disks to back up any files or packages you imported to use again next time you spin up a quick environment. We also provide Terra maintained Hail, Bioconductor, and GATK application configurations to support popular use of these toolkits.

**Slide 19: openwdl.org**
WDL is a community driven language. This is a site where you can learn more about it with some helpful pointers to the GitHub, stack overflow, slack workspace, etc.  OpenWDL is led by a small core group who help govern the language specification and you can find out more about this group from this site as well.

**Slide 20: Cloud Providers**

- Google Cloud Platform
- Azure (coming)

Wanted to highlight that when you create a Terra billing project it creates a Google project with particular permissions, so you can actually go into the GCP cloud console and see your billing data, what resources you are actively using, and a variety of other info.

**Slide 21: Applications**

- Bring your own Docker image (built off the Terra base image) or initialization script
- Galaxy: https://galaxyproject.org/
- IGV: https://software.broadinstitute.org/software/igv/home
- Seqr: https://seqr.broadinstitute.org/

So you can bring your own Docker image extended from the Terra base image to install the software application, programming languages, and packages you need. Or provide an initialization script. Here are links to more info on Galaxy, IGV, and Seqr.

**Slide 22: Interactive Analysis Tours**

- Best practice workspaces made available by tool developers
- Check out the BioData Catalyst Collection workspace

And to conclude, take a look at the BioData Catalyst collection workspace for a bunch of handy notebooks curated by the amazing team at UCSC. Thanks for your attention and now I will pass it on to the Seven Bridges team.

## Seven Bridges

**Slide 24: Live Demo**

Live demo by Dave Roberson of Seven Bridges platform (recording available on community forum)

**Slide 25: Seven Bridges Differentiators**

| | |
|---|---|
| **Workflow Language** | Common Workflow Language (CWL) |
| **Cloud Provider** | Amazon Web Services, Google Cloud Platform; capability to connect private cloud bucket directly to platform |
| **Applications** | Annotation Explorer: Query annotations for all SNVs and dbSNP INDELS (+8 billion variants and ~700 annotations)<br><br>Genomic Data Overview: Query TOPMed variants and see aggregated results. Compliant with Genomic Data Sharing Agreements. |
| **Interactive Analysis Features** | SAS<br><br>Commonly used libraries like Bioconductor are preloaded on the Docker image for spinning up the RStudio or Jupyterlab environments. |

# Requesting Pilot Credits

**Rep: Paul Kerr, User Services at BDC3**

**Slide 27: What are Cloud Credits?**
Cloud credits are required to perform certain tasks, such as running analyses or storing results. Note that users are not charged for the storage of the datasets that are hosted on BioData Catalyst. However, when that data is used in your analyses, then you will need to cover the costs of the computation and the storage of the derived results. Also, users who upload or import their own data into BioData Catalyst will need to cover the storage costs for that data.

Many factors are involved in estimating the amount of cloud credits needed. For instance, computation costs are influenced by the size of the analysis and the length of time it takes to run. Likewise, storage costs are influenced by the file size and the length of time that the data is stored. Our next Community Hours session in September will include a deeper dive into cloud credits, with representatives from the workspaces discussing tips for estimating and controlling costs, and writing BioData Catalyst funding into your grant proposals.

**Web resource**: Cloud Costs and Credits

**Slide 28: Try out both platforms with Pilot Credits**
New users of BioData Catalyst may apply for an initial $500 in cloud credits, also known as pilot credits, and many analyses can actually be completed for that amount or less. For larger tasks, you can use the credits to test and evaluate the ecosystem for things such as piloting pipelines on smaller samples and estimating how much a full analysis will cost.

Cloud credits can be applied to either, or both, workspaces. So your $500 in credits can go entirely to Terra or Seven Bridges, or you can split it into $250 on both workspaces. If you don't already have CWL tools or WDL tools, and you're flexible about which workspace to use, we recommend trying both so you can test and evaluate them with your chosen data to make an informed decision about which workspace is the best fit for your research. And, if you like, you can later transfer any unused credits from one workspace onto the workspace of your choice.

**Slide 29: Cloud Credits Workflow**
Here is the workflow for requesting pilot credits. First, if you haven't already, sign up for the BioData Catalyst community, which provides access to the help desk, forums, newsletters and more. Then, sign up for either Terra or Seven Bridges or both. After you've signed up for a

workspace, you can then complete the Cloud Credits Request Form. Again, you can use all your credits on a single workspace or split the credits between the two workspaces. After you've spent your pilot credits, you can cover future costs through grant funding, a credit card, or a purchase order.

**Slide 30: Web Form**

Here is our Cloud Costs and Credits webpage, which includes links to helpful articles on estimating and controlling cloud costs along with the form to request credits. Remember to complete the form after you've signed up for a workspace. The form asks for some basic information, including PI and collaborator names if available, and your role and organization.

At the bottom it says you "may request one of the following", and here you would choose the $500 in initial pilot cloud credits. I would note that the other option here is "Additional cloud credits." As you know, NHLBI's mission focuses on heart, lung, blood and sleep disorders, and researchers in those areas may apply for additional cloud credits beyond the initial $500 for specific projects. Please note that these requests are closely reviewed by NHLBI leadership and this should not be seen as a sustainable way to pay for large projects.

And lastly you'll choose your preferred workspace - Seven Bridges, Terra or both. Please note that it can take a few business days for cloud credits requests to be approved and applied to the workspace you choose.

**Slide 31: Next Steps**

- Follow one or both of the platform **Getting Started Guides**
    - Terra Getting Started
    - Seven Bridges Getting Started
- Join our next **Community Hours** on Wednesday, September 15th
- Help Desk: https://biodatacatalyst.nhlbi.nih.gov/contact

Next steps include reading the Getting Started Guide for your preferred workspace. Also be sure to join our next Community Hours in September which will include a detailed discussion on estimating, controlling and optimizing cloud costs, and writing BioData Catalyst funding into your grant proposals. And as always, feel free to contact our Help Desk with any questions or feedback you might have. Thank you.

## Discussion and Questions

- Send any questions to our help desk.
- Another option is to have one-on-one conversations with us.

**Question**: Are pilot cloud credits limited per lab, or can multiple users in a lab apply?
**Answer (Paul, BDC3)**: There is no limitation on who can apply. The requests are reviewed closely by NHLBI leadership, so if requests were redundant, that might raise some eyebrows, but in general free to apply and there are no inherent limitations.

**Question**: In AWS and GCP, if you are working directly on those platforms, it is relatively easy to leave something running and burn through a huge amount of cloud credits and it's not always easy to remediate that in any way - not always a great appeals process. Can you say anything of what that looks like in the context of these cloud credits?
**Answer (Tony, Seven Bridges)**: On Seven Bridges when you make a project, there is a block function that is on by default that you can turn off or adjust the time and say this project or these analyses have a time cap of 60 minutes. So if you accidentally leave it running or close the tab without stopping the analysis, it automatically will cap out. Again, that is customizable and adjustable, you can change it, and that can help you control costs.
**Answer (Tiffany, Terra)**: In terms of interactive analysis, if you are not actively using it, you can auto pause your environment so that you are not racking up costs for it while you aren't using it, that is something you can set and customize. In terms of cloud credits, in my previous experience, in the back-end there are processes for whoever is giving out these credits, to disassociate the billing from the project if you have passed your threshold for these credits. I am not sure if that is happening in this scenario, but that is what a lot of them do.
**Answer (Paul, BDC3)**: Regarding remediations, there is not an official process that currently exists, but GCP does accept requests if things are run in error. We are hoping that the tools mentioned, as far as controlling things before it gets to that point, are likely to be sufficient. But if you end up in this situation, reach out to the help desk and we will see what we can do to help the situation.

**Question:** As a Ph.D student now, I do not have permission to access for dbGap data. In this case, how can I sync to access dbGap data if I'm going to use in each platform?
**Answer (Tiffany, Terra)**: There is information on how to request access to dbGaP data and get that process started. In this case, you want to get that process started so you can eventually access dbGaP data. There is also open access data that you can use and PIC-SURE and Gen3

are good places to search for that to use later in the analysis platforms.  That would be one way to use other data if you don't quite have what you need access to yet.
**Answer (Tony, Seven Bridges)**: You can test the platform as mentioned with open data. My understanding is you have to go through your PI to have them request the dbGaP data after you identify the datasets you need, using a literature search or wherever you are finding that.
**Answer (Simran, PIC-SURE)**: You would need to be added as a dbGaP downloader under a PI to gain access to dbGaP data, and then that downloader access syncs automatically with the Terra, Seven Bridges, and Gen3 platforms.
**Answer (Amber, BDC3)**: You can find information on our previous session from August 11 regarding data access and exploration - including information on data access requests - on our community forum and those materials may be helpful.

**Tony, Seven Bridges:** I would like to emphasize and recommend that you register for the community, if you haven't already. Try out both Seven Bridges and Terra platforms. Reach out if you have questions.

## Join the ecosystem

- If you have not joined the ecosystem, make sure to join to learn about future events and follow the GUIDE for getting started.
- Join here: https://biodatacatalyst.nhlbi.nih.gov/contact/ecosystem

## Thank you and next hours

- We will be having another Community Hours in three weeks, on **Wednesday, September 15th, at 1 PM EDT**.
- The topic is "Cloud Credits". We will discuss budgeting and writing cloud credits into your grant proposals.
- Reps from all platforms will be available to answer your questions.
- Submit questions in advance, if you have them.
- You can register now: https://bit.ly/3yeQddp