NCI Cloud Resources Hands-on

2018 ACM-BCB Workshop, Washington, DC

[**Workshop Material**](#_5cngat6lho0o) **2**

[Presentation Slides](#_pplmfsxxxtuz) 2

[GitHub Repo](#_p44cmg8gh2qb) 2

[Data provided by Kallisto -](#_pplmfsxxxtuz) 2

[Cite NCI Cancer Research Data Commons and NCI Cloud Resources](#_ckwo7qdd632o) 2

[**Contact us**](#_dc3i9zbubewa) **2**

[**Broad Institute - FireCloud**](#_nph4tahrza7g) **3**

[For the Workshop](#_3gkpd3afikb) 3

[What you need before registering](#_461rh6awpp94) 3

[How to get credits](#_956fmm4ncofw) 3

[Registering for FireCloud](#_h9sz0pd3tdz8) 3

[Create a Workspace](#_ammniyfgw7jn) 4

[Import Metadata](#_edzxs91iiqsn) 5

[Create Workflow (Methods/Config)](#_gu0pxtb7cjyt) 6

[Launch Analysis](#_3x1pdp6y3s33) 8

[Cite the FireCloud](#_6nlv96y9bwyd) 9

[**Seven Bridges Genomics - Cancer Genomics Cloud**](#_zq4qrnjvq4o) **10**

[For the Workshop](#_1358l6lnofks) 10

[What you need before registering](#_vhezuj1tqg) 10

[Registering for the Seven Bridges CGC](#_c2qpu9582hbx) 10

[How to get credits](#_p28pvn58ug9) 11

[Create a project](#_aujbi2ajwcyv) 12

[Add members, data and tools to a project](#_9904wwb77eyf) 12

[Cite the Seven Bridges CGC](#_gnpcpvbllfd4) 14

[**Institute of Systems Biology - Cancer Genomics Cloud**](#_eclfe58achjz) **15**

[For This Workshop](#_7hnmaft3yd7x) 15

[Login to ISB-CGC Google Cloud Console - https://console.cloud.google.com/](#_y0dbnji1stz5) 15

[Run Kallisto WDL in a Docker Container](#_8rxjk1yeiojz) 15

[Connect to VM by clicking “SSH”](#_bstvg1o8x9r5) 15

[Run Kallisto and View Output](#_8utnoses8t67) 16

[Stop the VM (Don’t forget)](#_tyh60uya3dmj) 17

[Other useful Docker commands](#_90k684msmz4b) 17

[Run Kallisto in Unix - bash commands](#_d9jtmgobpc2) 17

[Cite ISB-CGC](#_4xk6zlmu8dkc) 17

# Workshop Material

## [Presentation Slides](https://docs.google.com/presentation/d/1dZosV7waROzohx18rlSUC3PNVlW6xoUKAB8XmqpJJ7A/edit?usp=sharing)

## [GitHub Repo](https://github.com/stevetsa/2018-ACMBCB-workshop)

## Data provided by Kallisto -

<https://github.com/stevetsa/2018-ACMBCB-workshop/raw/master/reads_1.fastq.gz>

<https://github.com/stevetsa/2018-ACMBCB-workshop/raw/master/reads_2.fastq.gz>

<https://github.com/stevetsa/2018-ACMBCB-workshop/raw/master/transcripts.fasta.gz>

## Cite NCI Cancer Research Data Commons and NCI Cloud Resources

[Hinkson, I. V., Davidsen, T. M., Klemm, J. D., Kerlavage, A. R. & Kibbe, W. A. A Comprehensive Infrastructure for Big Data in Cancer Research: Accelerating Cancer Research and Precision Medicine. Front Cell Dev Biol 5, 83 (2017).](https://www.frontiersin.org/articles/10.3389/fcell.2017.00083/full)

# Contact us

[Steve Tsang](mailto:hsinyi.tsang@nih.gov)

[Durga Addepalli](mailto:kanakadurga.addepalli@nih.gov)

# 

# Broad Institute - FireCloud

<http://firecloud.org>

## For the Workshop

* Set up an account
* Create a workspace
* Clone a workspace
* Configure the workspace and methods
* Run analysis!

## What you need before registering

* **A Google account** This can be a Gmail account, an institutional Google Apps account, or any Google Apps account that you create. Note that you can create a Google Apps account associated with your existing email address; as explained [here](https://software.broadinstitute.org/firecloud/documentation/article?id=9846), using a Gmail address is not required.
* The name of your Institute and your institutional affiliation
* The name and email of your Principal Investigator (PI) or Program Lead

## How to get credits

To take advantage of this unique opportunity, all you need to do is [register](https://software.broadinstitute.org/firecloud/documentation/article?id=6816) for an account on the FireCloud portal (which is itself always free and open to all). When you log into [the FireCloud portal](https://portal.firecloud.org/) you will see a banner at the top of the screen welcoming you to the trial. Click **Start trial**, accept the Terms and Conditions, and you will be issued a FireCloud billing project with the free credit. It is as simple as that!

Note on the FireCloud Billing Project: The billing project will have the following naming convention: fccredits-xxx-xxx-xxxx. Be sure to use this billing project whenever prompted in FireCloud, such as when creating workspaces, launching analyses, or downloading data. Starting the trial will also activate the sixty-day countdown.

We reserve the right to accept or reject applications for credits at our discretion. This is intended to protect against any attempts to abuse the credits program since our goal here is to give as many people as possible the opportunity to try out the pipelines. The credit funding is not endless, so we have the right to stop the program when we hit our budget.

## Registering for FireCloud

<https://software.broadinstitute.org/firecloud/documentation/article?id=12216>

1. Using Google Chrome, go to [the FireCloud portal](http://firecloud.dsde-prod.broadinstitute.org/) or click on **Use FireCloud** in the top menu bar of the documentation website.
2. Click **Register**, then **Sign In** with your Gmail or Google Apps account.
3. When asked to allow FireCloud to view your email address and basic profile info, please click **Allow**. You can find an explanation of what this entails in the [**FAQs**](https://software.broadinstitute.org/firecloud/documentation/article?id=6892).
4. You will be prompted to enter **New User Registration** information. Please add your first and last name, along with the other required (\*) fields.
5. After entering your registration information, click **Register**.

Note that you can optionally provide a **Contact Email** to receive information to help you get started, as well as periodic notifications. We recommend providing a **Contact Email** if the Gmail or Google Apps account you use to register for FireCloud is not your primary email. If you do not provide a **Contact Email**, we will send notifications to your Gmail or Google Apps account.

Initial Compute Credit available for 6 months

## Create a Workspace

<https://software.broadinstitute.org/firecloud/documentation/article?id=10746>

There are two ways to create a new [workspace](https://software.broadinstitute.org/firecloud/documentation/quickstart):

* From the Workspace List, by clicking **Create New Workspace...**
* If you want to copy an existing workspace, click **Clone...** Cloning will make a copy of all the content in the workspace except the submissions. This means analyses submission history will not be copied into the Monitor tab. Cloning will not copy files over to the new bucket.

Upon creation, you will be prompted to enter information about your new workspace (if you already obtained free credits [as detailed above](#_956fmm4ncofw), then a billing project has already been created for you, and you can ignore the block below):

* Pick the [Billing Project](https://software.broadinstitute.org/firecloud/documentation/article?id=9763) that will cover the compute, storage, and download costs for this workspace.
* Give the workspace a unique name (only letters, numbers, underscores, and dashes allowed). The name of the Billing Project + Name must be unique in all of the Google Cloud Platform.
* Enter a Description of what work will be done here.
* Choose a [Group](https://software.broadinstitute.org/firecloud/documentation/article?id=9553) or Groups to be in the [Authorization Domain](https://software.broadinstitute.org/firecloud/documentation/article?id=9524). These are used to secure your workspace access to a group of specific users.

Note: If the "Create New Workspace" button is greyed out, it means you do not have access to a FireCloud Billing Project. To set up billing, [follow this process flow](https://software.broadinstitute.org/firecloud/documentation/article?id=10444) or obtain free credits for a limited period of time by referring to the section ["How to get credits"](#_956fmm4ncofw). If someone you are working with already has billing set up and will allow you to use their billing project, they can follow these [instructions](https://software.broadinstitute.org/firecloud/documentation/article?id=9764) to set you up.

## Import Metadata

<https://software.broadinstitute.org/firecloud/documentation/article?id=10738>

You can import metadata into your workspace's data model by either *copying from an existing workspace* or *importing a file*. The following steps outline how to import metadata for your own projects, but in this workshop, we have already imported some sample metadata so you can run an analysis. Navigate to the “**Data**” tab in the **2018-ACMBCB-Workshop** workspace to view metadata for the patients and the samples.

Copying from an existing workspace

1. Choose the [workspace](https://software.broadinstitute.org/firecloud/documentation/article?id=9777) you want to import metadata from. You will notice that you can only import data from workspaces that are compatible with the [Authorization domain](https://software.broadinstitute.org/firecloud/documentation/article?id=9524) you have set.
2. Pick the participants, samples, pairs, or sets you want. Importing sets will bring over all the data required for the set. For example, if you import a sample set, the sample and participant data linked to the set will also be copied over.
3. Notes:
   1. Import conflicts can occur if you already have an identical participant, sample or pair in your workspace that matches what you are importing. FireCloud will notify you that the entity already exists in the workspace.
   2. Copying metadata from another workspace will not import any linked files into your workspace bucket. Rather, it will refer to file paths in the bucket of the workspace you copied. Thus, if that workspace bucket is deleted, your workspace data model will no longer refer to an existing bucket path.

Importing a file

You import metadata corresponding to entity type -- [Participant](https://software.broadinstitute.org/firecloud/documentation/article?id=9770), [Sample](https://software.broadinstitute.org/firecloud/documentation/article?id=9771), or [Pair](https://software.broadinstitute.org/firecloud/documentation/article?id=9772) -- by uploading load files in tab-separated-value format, a type of text file (.tsv or .txt). Separate files must be used for uploading each entity type. The first line of each file must contain the appropriate field names in their respective column headers. See the individual entity entries for examples of load files.

Note that for each of the basic entities, the data model also supports set entities, which are essentially lists of the basic entity type:

* Participant Set
* Sample Set
* Pair Set

In **set** load files, each line lists the membership of a non-set entity (e.g., participant) in a set (e.g., participant set). The first column contains the identifier of the set entity and the second column contains a key referencing a member of that set. For example, a load file for a participant set looks like this:

|  |  |
| --- | --- |
| membership:participant\_set\_id | participant\_id |
| TCGA\_COAD | TCGA-5M-AAT4 |
| TCGA\_COAD | TCGA-NH-A8F8 |
| TCGA\_BRCA | TCGA-A8-A07L |

Note that multiple rows in a set load file may have the same set entity id (e.g. TCGA\_COAD).

Order for uploading Load Files

Load files **must be imported in a strict order** due to references to other entities.

The order is as follows ("A > B" means entity type A must precede B in order of upload):

* participants > samples
* samples > pairs
* participants > participant sets
* samples > sample sets
* pairs > pair sets
* set membership > set entity, e.g., participants > samples > sample set membership > sample set entity.

## Create Workflow (Methods/Config)

<https://software.broadinstitute.org/firecloud/documentation/article?id=11123>

This tutorial covers the basic formatting for configuring a method's inputs and outputs in a workspace. There is also a [video tutorial](https://www.youtube.com/watch?v=bVU0Rdu0m78), that explains this and other concepts in more detail if you prefer.

Navigate to the “**Method Configurations**” tab in the **2018-ACMBCB-Workshop** workspace, and you should see a configuration with the name “**kallisto**.” Click on the name (kallisto) and have a look at the configuration.

Under the header **WDL**, you’ll see the option “Expand,” which you can click on to see the contents of the [WDL](https://software.broadinstitute.org/wdl/documentation/quickstart), where we have defined inputs and outputs for the workflow. FireCloud interprets the WDL and provides you an input "form" to fill out. The outputs part of the form is optional.

This example shows just the workflow portion of a WDL and what the inputs and outputs look like filled out in FireCloud.

**Look at the WDL code. How many workflows and tasks do you see?**

Answer: The WDL includes one workflow with two tasks. There are inputs listed for the workflow and separately for the tasks, followed by outputs for the workflow.

In the Inputs diagram, you can tell the difference between **workflow** inputs and **task** inputs by looking at the name. The first four inputs named CramtoBamFlow, are the workflow inputs according to the WDL.

After the name of the task or workflow, the other columns are:

* Variable = the name of the input given in the WDL
* Type = Integer, String, Boolean, File, or Array of these
* Attribute = the actual value (this is the part you configure)

Running a method without the workspace data model

Sometimes users come to FireCloud with a preexisting WDL that was successfully running on their local infrastructure and would now like to test it quickly in a cloud environment. These users can take advantage of the ability to run their analysis without setting up the data model using the following steps:

1. Upload the method to the Method Repository and export it to a workspace.
2. Uncheck the checkbox “Configure inputs/outputs using the Workspace Data Model” within the method configuration.
3. Upload the input json file by clicking “Populate with a .json file.” FireCloud then populates the configuration with the attributes listed in the json file. Review and edit if necessary.
4. Click "Launch Analysis". If you want to use call caching, leave this checkbox checked.

The output files will be placed in the Google bucket after completion and will not be registered in the data model.

Attribute format per type

When you fill out the Attribute section per input, you have to follow the formatting requirements based on the type listed. See the Inputs diagram above for examples.

1. Integer - No formatting required.
2. String - Quotes required. *e.g.* "my string"
3. Boolean- Quotes required. Case insensitive so "true" or "TRue" or "TrUE" are the same.
4. File - can be referenced from the Google bucket, data model, or workspace attribute section. See the **Referencing files** section below for details.
5. Array[X] - Lists of these attributes can be entered with a comma between each item. *e.g.* "a","b","c" or 1,2,3 or "true","True","TruE","TRUE"

Referencing files

* **Google bucket** - Use *"gs://url-to-file-in-bucket"* to reference the [Google bucket](https://software.broadinstitute.org/firecloud/documentation/article?id=9778) file. Please note the quotes are necessary if you are directly referencing a file URL, but the quotes are *not* necessary if you reference a file using the data model or workspace attributes below.
* **Data model** - Suitable for referencing several files listed in your [data model](https://software.broadinstitute.org/firecloud/documentation/article?id=9769) without hard coding values or having to adjust your method configuration when you add more data to the table. You can call the files listed under the name of a column by typing this. plus the column title. Make sure to leave the checkbox checked “Configure inputs/outputs using the Workspace Data Model” so that FireCloud registers that you are using the data model. The keyword this. tells FireCloud to look at your data model in the table you set as your root entity. So if you set your root entity to "sample" when you imported the method to your workspace, then FireCloud will look in the "sample" table for an [attribute](https://software.broadinstitute.org/firecloud/documentation/article?id=10503) (a column) with the name you specify. e.g. this.sample\_id would look in the "sample" table for the "sample\_id" attribute. This expression also gives you the flexibility to dive into attributes that exist on any entity that the method config is running on. For example, say your method is to be run on a [pair](https://software.broadinstitute.org/firecloud/documentation/article?id=9772). The pair table contains a control\_sample\_id, a case\_sample\_id, and their corresponding bam files. Say your WDL task requires the case\_sample\_bam input. You’d type this.case\_sample\_id.case\_sample\_bam
* **Workspace attribute** - Storing an input as a workspace attribute is convenient if you are using a file over and over again in multiple methods. If the file path changes, you only have one place to update, similar to global variables in scripting. You can call this by typing workspace. plus the attribute key. For example, workspace.ref\_fasta or workspace.ref\_dict If you type workspace. into the method configuration, all the workspace attributes available will auto-populate below.

Outputs

If you are using the data model for your analysis, you can optionally fill out the outputs with the same nomenclature (workspace., this., etc.). It is optional because your outputs will go directly into your bucket without defining anything here. If you want links to the output file destination in your data model, you need to define it here. Determine the name of the column or use a pre-existing column and type this. in front of it. For example, this.analysis\_ready\_bam will output the BAM to the column called analysis\_ready\_bam in the sample tab (if you chose to run this method on a sample). If the column header doesn’t exist now, the method will create it after execution.

A note on versioning

When you create a method or a method configuration, FireCloud will give it a number (starting at 1) to identify the version. This is called a snapshot ID and can be found in the method and method config header along with the name, owner, documentation, etc. Every time a method or method config is edited, FireCloud automatically adds 1 to the ID. You can keep track of what method snapshots you have launched in the Monitor tab.

## Launch Analysis

<https://software.broadinstitute.org/firecloud/documentation/article?id=10738>

Launching an analysis is pretty simple once you have the [method inputs configured](https://software.broadinstitute.org/firecloud/documentation/article?id=11123). This [video demo](https://www.youtube.com/watch?v=G_HTuXYZayU&list=PLbhQIHW7hn4jt-_gw8FN-ax6W2sVPYdcQ) shows how to launch an analysis with a Featured Workspace, where the methods are configured for you. The steps are detailed below.

1. Click the **Launch Analysis…** button in the [Method configuration tab](https://software.broadinstitute.org/firecloud/documentation/quickstart?page=config).
2. If you are **not** using the data model (i.e. you have unchecked the checkbox *Configure inputs/outputs using the Workspace Data Model*) then you can skip to Step 3. If you are using the data model, now you can select what sample, participant, pair, or set you want the method to run on. For example, if the Root entity type is equal to "sample," choose a sample. If you want the method to run on multiple samples, you will have to create a sample set and **Define an Expression**, this.samples, that will expand the workflow from running on a single sample to multiple samples in the set. See the **Defining Expressions** section for more detail. Note: You can change the Root entity type of a method configuration, by selecting **Edit Configuration...** and choosing from the drop-down options.
3. Decide whether to turn [call caching](https://software.broadinstitute.org/firecloud/documentation/article?id=9313) on or off.
4. Click Launch. This starts a conversation between FireCloud, [Cromwell](https://cromwell.readthedocs.io/en/develop/), [PAPI](https://software.broadinstitute.org/firecloud/documentation/article?id=11116), and Dockerhub or other image registry. FireCloud will then take you to the [Monitor tab](https://software.broadinstitute.org/firecloud/documentation/quickstart?page=monitor), to monitor your submission and workflows.

Defining Expressions

Defining Expressions is most commonly required when you want to run on sets of [participants](https://software.broadinstitute.org/firecloud/documentation/article?id=9770), [samples](https://software.broadinstitute.org/firecloud/documentation/article?id=9771), or [pairs](https://software.broadinstitute.org/firecloud/documentation/article?id=9772) instead of one item.

For example, see the screenshot below of the Test\_data sample set with two samples. To get this method to run on both samples, you add the expression this.samples

See how **samples** is the name of the third column in the sample\_set table? FireCloud reads this as, “run the method on whatever is in the samples column for the Test\_data sample set.” One workflow will be created for each sample in the set and submitted all at once.

## Cite the FireCloud

<https://software.broadinstitute.org/firecloud/documentation/article?id=12216>

[Birger, C. et al. FireCloud, a scalable cloud-based platform for collaborative genome analysis: Strategies for reducing and controlling costs. bioRxiv 209494 (2017). doi:10.1101/209494](https://www.biorxiv.org/content/early/2017/11/03/209494.full.pdf+html)

# 

# Seven Bridges Genomics - Cancer Genomics Cloud

<http://www.cancergenomicscloud.org>

Overview of Topics:

What you need before registering

Registering for the Seven Bridges CGC

How to get credits

User settings

Create a project

Add members, data and tools to a project

Launch an analysis

Cite the SBG CGC

## For the Workshop

* Set up an account
* Create a project
* Copy the data (or upload from links below) and confirm metadata setting

<https://github.com/stevetsa/2018-ACMBCB-workshop/raw/master/reads_1.fastq.gz>

<https://github.com/stevetsa/2018-ACMBCB-workshop/raw/master/reads_2.fastq.gz>

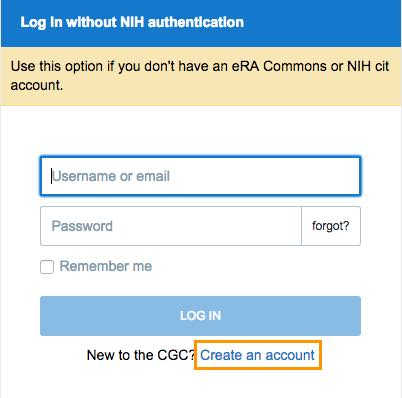
<https://github.com/stevetsa/2018-ACMBCB-workshop/raw/master/transcripts.fasta.gz>

* Set metadata for paired files
* Create a workflow -with Kallisto Index and Kallisto Quant
* Run analysis!

## What you need before registering

## Registering for the Seven Bridges CGC

1. There are two ways to sign up for the CGC:
   1. External account: If you already have an account with eRA Commons or NIH CIT, you can register using the associated credentials.
   2. Dedicated CGC account: If you do not have or do not wish to register using an external account, you can create a new account using a standard email and password. You can always [upgrade your CGC account](https://docs.cancergenomicscloud.org/v1.0/docs/account-settings-and-billing#section-link-your-era-commons-account-to-your-cgc-account) to be associated with your eRA Commons/NIH CIT account at a later time.
2. To create a CGC account using your email address:
   1. Go to <http://www.cancergenomicscloud.org/> and click “Create an account”.



* 1. Click “Register with a good old email/password combo”.
  2. Complete the registration form, including agreeing to the [CGC Terms](http://www.cancergenomicscloud.org/terms) and [Data Use policy](http://cancergenome.nih.gov/pdfs/Data_Use_Certv082014).
  3. Click “Register”. The CGC will send you an email containing a verification link.
  4. Open the email, and click “Confirm your email.”

1. To instead create an account using your eRA Commons or NIH CIT account:
   1. Click “Create an account”.
   2. Click “Register via an external account”.
   3. You will be redirected to the external iTrust login page, where you should enter your eRA Commons or NIH CIT credentials and click “Log in”.
   4. Next, you will be asked to authorize Seven Bridges to manage your access to Controlled Data. If you agree, click “Yes, I authorize.”
   5. You will be redirected back to the CGC registration form, which you should complete by entering the additional information required (your name, email, and organization) and agreeing to the [CGC Terms](http://www.cancergenomicscloud.org/terms) and [Data Use policy](http://cancergenome.nih.gov/pdfs/Data_Use_Certv082014).
   6. Click “Proceed to the CGC” to begin using the platform.
   7. If your eRA Commons account is associated with an approved dbGaP data access request for TCGA or TARGET Controlled Data, you should be able to access those data on the CGC within 24 hours of registering your account.

## How to get credits

1. NCI provides credits to assist researchers with the computational and storage costs associated with using the CGC. You will start with $300 of compute and storage credits upon account registration.
2. Once you have used the initial $300 of credits, you will be asked to provide a credit card or purchase order number to cover any additional data storage and/or computation costs associated with your research on the CGC.
3. There are no costs associated with using the CGC beyond the cost of the Amazon Web Services (AWS) users access as part of their research, such as:
   1. data storage costs for files they upload or generate as part of their analyses
   2. compute costs they incur while running their analyses on the CGC.
4. All costs are directly based on [Amazon Web Services (AWS) on-demand instance pricing and S3 data storage](https://aws.amazon.com/ec2/pricing/).
5. The CGC team also accepts proposals on a rolling basis for large cancer research and tool development that will be conducted on the CGC.
   1. Users can work with the CGC Team to prepare a cost estimate and timeline for the proposed project, and, as warranted by the project objectives, funded projects will be granted up to $10,000 in credits to cover data storage and computation costs on the CGC.
   2. Requests from graduate students and postdocs are particularly encouraged.
   3. Contact the CGC Team at cgc@sbgenomics.com with the subject ‘CGC Collaborative Project Support Request’.

## Create a project

1. There are two types of projects you can create on the CGC: standard projects and controlled projects (those that can contain Controlled Data).
2. Click the *Projects* tab on the top navigation bar and select *Create a project* from the drop-down menu.
3. Name your project.
4. Define the project type if you have Controlled Data access. If you don't have Controlled Data access, all your projects are standard projects by default.
5. Set your billing group.
   1. Each new CGC account comes with [Amazon Web Services (AWS)](https://aws.amazon.com/) computation and storage credits in a *Pilot Fund* billing group, which should be selected by default. You should be able to run your first analysis without adding additional funding.
6. Click *Create* to finish.

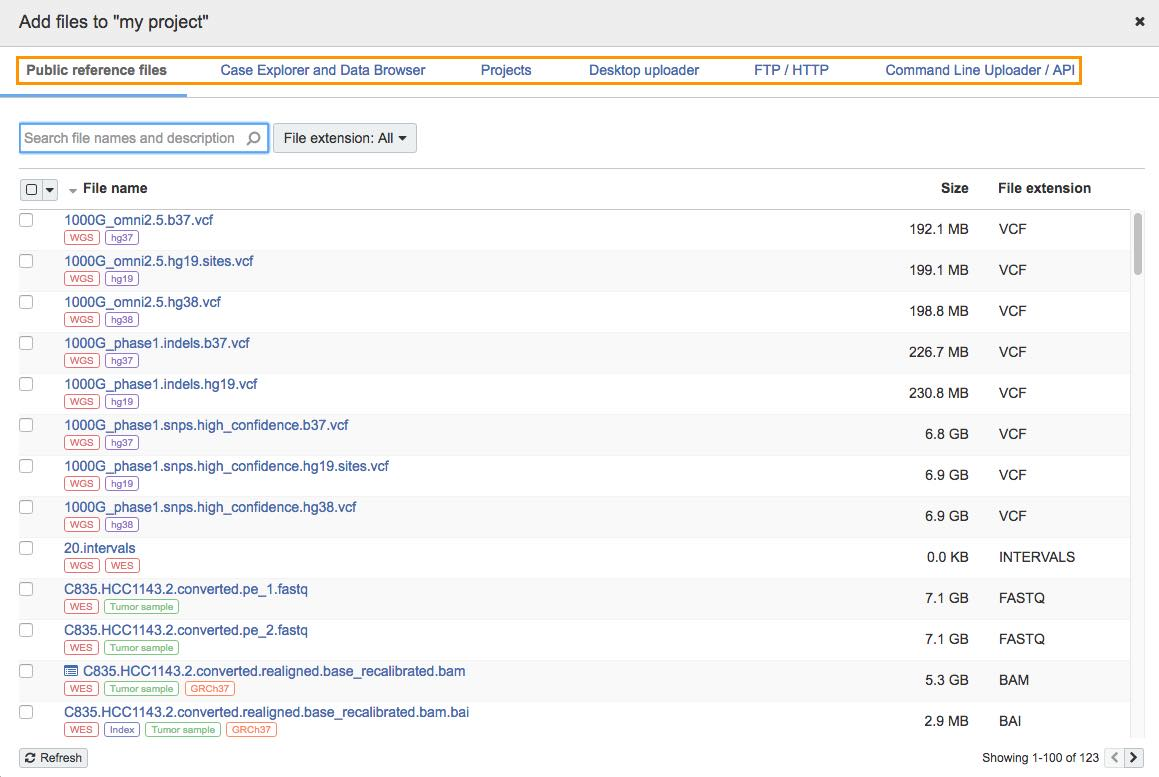
## Add members, data and tools to a project

Add members to your project:

1. Only project administrators can invite collaborators to a project.
2. Click *Invite new members* (or *Manage project members* in case there are already members in the project) on the dashboard of your chosen project.
3. Begin typing your collaborator's CGC username or their email address. The CGC will autosuggest all users with whom you have previously collaborated.
4. If you try to add a project member who does not have access to Controlled Data to a controlled project, you will receive a warning message, and the member won't be added.
5. If your collaborator does not have an account on the CGC, he/she will receive an email inviting him/her to create an account and join the project. In order to begin working together, your collaborator needs to register for an account on the CGC.
6. As the project administrator, you can secure your project by setting the permissions for your collaborators.
7. The different permissions you can specify are: read, write, copy, execute, admin.

Add data to your project:

1. Click on the *Files* tab.
2. Select *+Add Files*.
3. Select the correct repository from the navigation panel, as shown below.



1. For this hands-on session - Select FTP/HTTP and enter the following urls

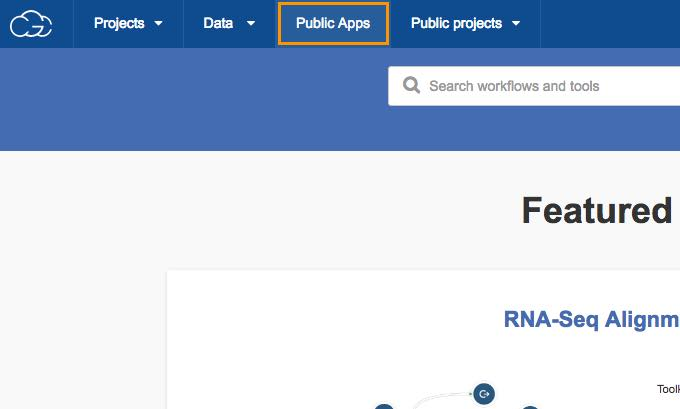
<https://github.com/stevetsa/2018-ACMBCB-workshop/raw/master/reads_1.fastq.gz>

<https://github.com/stevetsa/2018-ACMBCB-workshop/raw/master/reads_2.fastq.gz>

<https://github.com/stevetsa/2018-ACMBCB-workshop/raw/master/transcripts.fasta.gz>

Add tools to your project:

1. The CGC offers a repository of publicly available apps that are suitable for many different types of data analysis.
2. Apps include both tools (individual bioinformatics utilities) and workflows (chains or pipelines of connected tools).
3. Access the Public Apps gallery. Identify the type of app:  denotes a Workflow, and  denotes a Tool.



1. Enter the desired keyword “Kallisto” into the search field and press *ENTER*. You will be presented with apps matching the search criteria.
2. Click *Copy* to copy “Kallisto Index” and “Kallisto Quant” into your project. Use this option if you'd like to edit your app.
3. Alternatively, click *Run* to create a draft task. Use this option if you'd like to run the app without editing it.

Launch an analysis

1. Select the appropriate input files to run Kallisto, enter the parameters if any, and click *Run*.

Documentation:

<https://docs.cancergenomicscloud.org/>

Forum:

<https://docs.cancergenomicscloud.org/discuss>

Support:

[cgc@sbgenomics.com](mailto:cgc@sbgenomics.com)

## Cite the Seven Bridges CGC

[Lau, J. W. et al. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research. *Cancer Res*. 77, e3–e6 (2017).](http://cancerres.aacrjournals.org/content/77/21/e3)

# 

# Institute of Systems Biology - Cancer Genomics Cloud

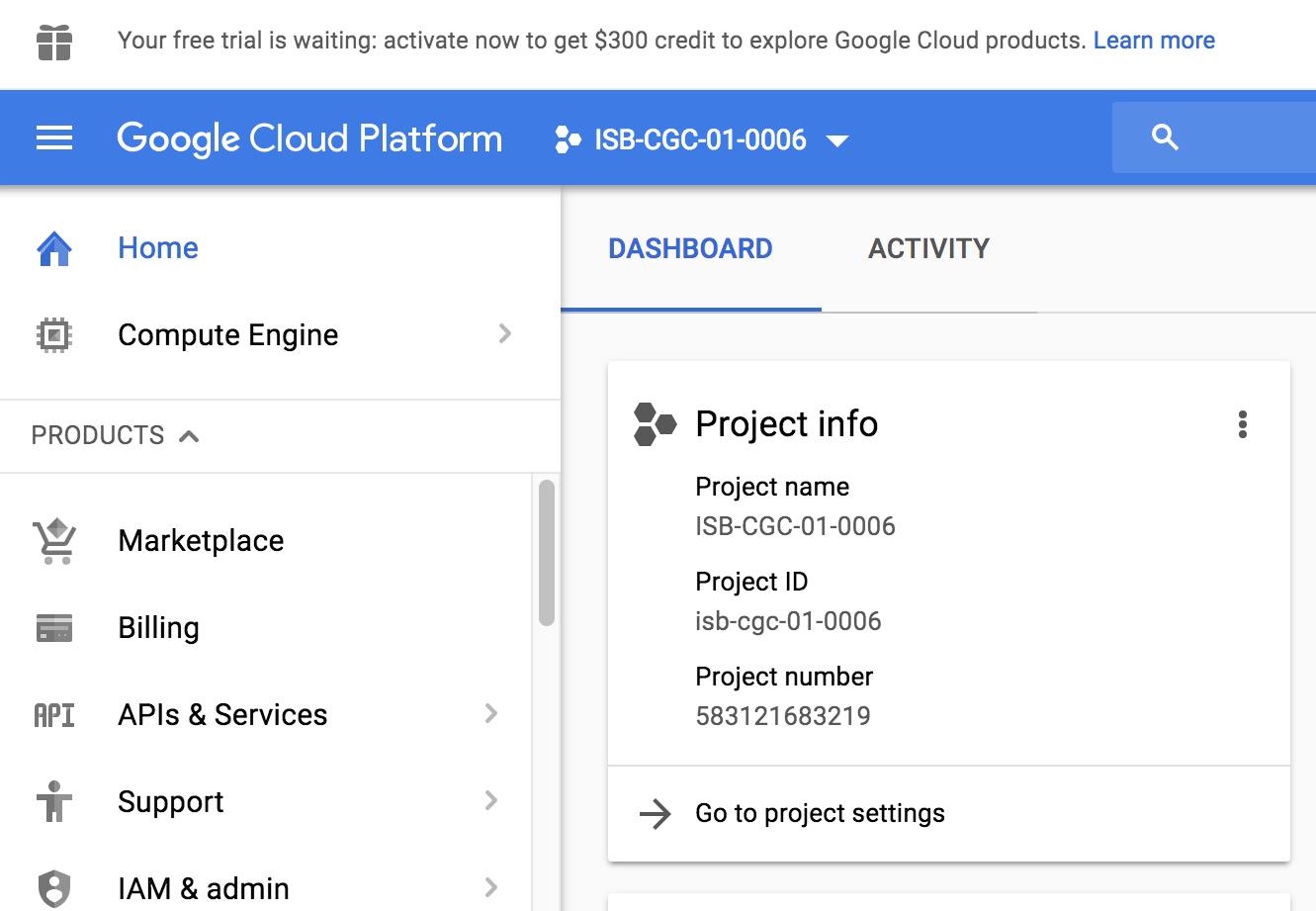
[http://cgc.systemsbiology.net](http://cgc.systemsbiology.net/)

## For This Workshop

### Login to ISB-CGC Google Cloud Console:

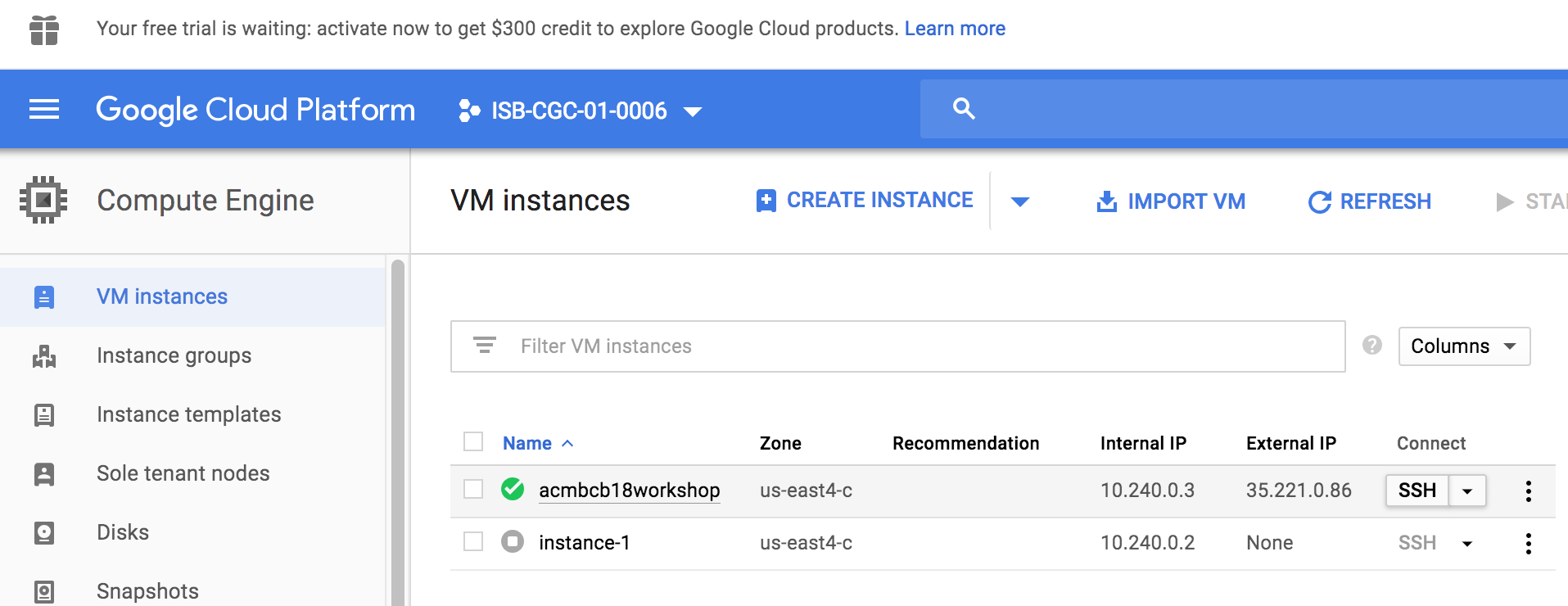
### Follow this link and sign in using the Google ID you input to be added to the billing group: <https://console.cloud.google.com/>

Click on “Select a Project” at the top of the screen and select “ISB-CGC-01-006”

Click on the navigation menu in the top left corner of the screen (represented by three parallel lines) and select “Compute Engine,” then “VM instances.” You should now see a list of VM instances, or at least, one instance called “acmbcb18workshop.”

### Run Kallisto [WDL](https://github.com/stevetsa/2018-ACMBCB-workshop/blob/master/FireCloud/kallisto_cfg.wdl) in a Docker Container

#### Connect to VM by clicking “SSH”



#### Run Kallisto and View Output

gcloud auth login

## copy and paste URL to a new browser then copy and paste the authentication back to the terminal

git clone <https://github.com/stevetsa/2018-ACMBCB-workshop.git>

## -v flag mount `pwd` on host to `pwd` in container

## -w working directory in the container

cd 2018-ACMBCB-workshop

docker run --rm -v `pwd`:`pwd` -w `pwd` -i -t stevetsa/2018-acmbcb-workshop

### inside container <root prompt> ###

#java -jar /opt/womtool-34.jar inputs kallisto\_cfg.wdl > kallisto\_inputs.json

### modify kallisto\_inputs.json

cd ISB-CGC

java -jar /opt/cromwell-34.jar run kallisto\_cfg.wdl -i kallisto\_inputs.json > kallisto\_workflow.$USER.log

gsutil cp kallisto\_workflow.$USER.log gs://stevetsa/2018-ACMBCB

## To scroll in SSH window

## Use Ctrl-Shift-PageUp / Ctrl-Shift-PageDn (Windows and Linux)

## Fn-Shift-Up / Fn-Shift-Down (on Mac OS X).

### 

[2018-08-22 18:29:26,77] [info] SingleWorkflowRunnerActor workflow finished with status 'Succeeded'.

{

"outputs": {

"kallisto.quant.TAR": "/home/ubuntu/2018-ACMBCB-workshop/cromwell-executions/kallisto/cfa39df3-8d0f-4819-bd45-ef82ad9af602/call-quant/execution/quant.tar.gz",

"kallisto.index.indexFASTA": "/home/ubuntu/2018-ACMBCB-workshop/cromwell-executions/kallisto/cfa39df3-8d0f-4819-bd45-ef82ad9af602/call-index/execution/transcript\_index"

},

"id": "cfa39df3-8d0f-4819-bd45-ef82ad9af602"

}

ls -al /home/ubuntu/2018-ACMBCB-workshop/cromwell-executions/kallisto/cfa39df3-8d0f-4819-bd45-ef82ad9af602/call-quant/execution/outputDir

total 52

drwxr-xr-x 2 root root 4096 Aug 22 18:29 **.**

drwxr-xr-x 3 root root 4096 Aug 22 18:29 **..**

-rw-r--r-- 1 root root 35848 Aug 22 18:29 abundance.h5

-rw-r--r-- 1 root root 558 Aug 22 18:29 abundance.tsv

-rw-r--r-- 1 root root 736 Aug 22 18:29 run\_info.json

ls /home/ubuntu/2018-ACMBCB-workshop/cromwell-executions/kallisto/cfa39df3-8d0f-4819-bd45-ef82ad9af602/call-index/execution/ -al

-rw-r--r-- 2 root root 399346 Aug 22 18:29 transcript\_index

## Stop the VM (Don’t forget)

## Other useful Docker commands

docker rm $(docker ps -a -q)

## Run Kallisto in Unix - bash commands

kallisto index -i transcripts\_index transcripts.fasta.gz

kallisto quant -i transcripts\_index -o output\_dir reads\_1.fastq.gz reads\_2.fastq.gz

## Cite ISB-CGC

[Reynolds, S. M. et al. The ISB Cancer Genomics Cloud: A Flexible Cloud-Based Platform for Cancer Genomics Research. Cancer Res. 77, e7–e10 (2017).](http://cancerres.aacrjournals.org/content/77/21/e7.long)