



SBG-CGC: global architecture

Tiffany Delhomme

IARC course - analysing TCGA data in the cloud

28 feb 2018

Summary

1 introduction

2 overview

3 material

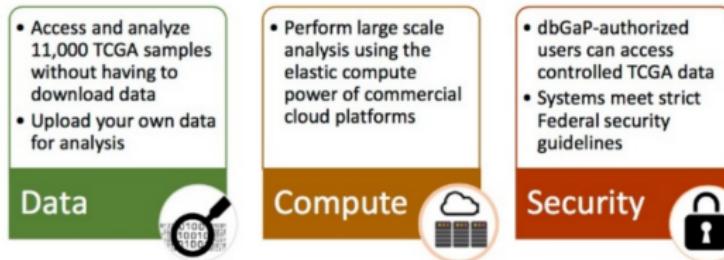
4 analyses

NCI cloud pilots

Data, Computing and Security

Aim: providing big data resources to accelerate cancer research

- in 2016, NCI has launched 3 Cancer Genomics Cloud Pilots
 - FireCloud from the Broad Institute
 - Cancer Genomics Cloud from the Institute for Systems Biology
 - **SevenBridges Cancer Genomics Cloud (SBG-CGC)**
- concept: co-localized data and computing resources



Global overview

The Seven Bridges Platform

The CancerGenomicsCloud provides a simple and powerful way to analyze large quantities of genomic data:

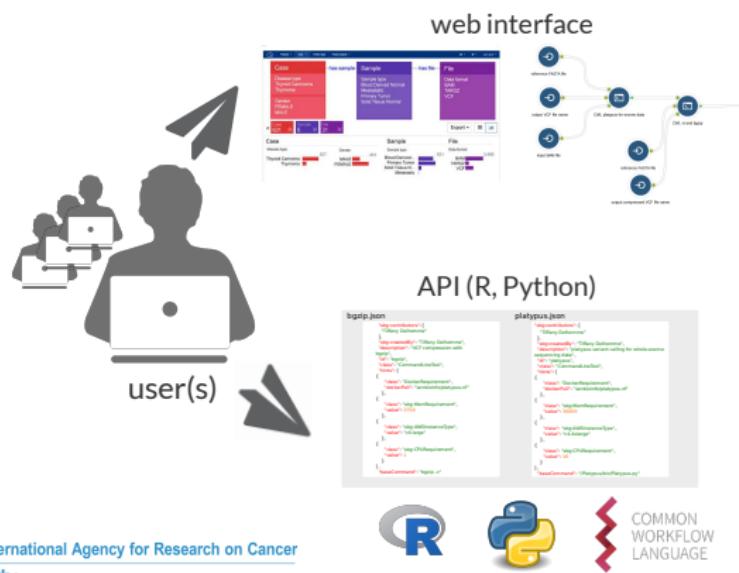


user(s)

Global overview

The Seven Bridges Platform

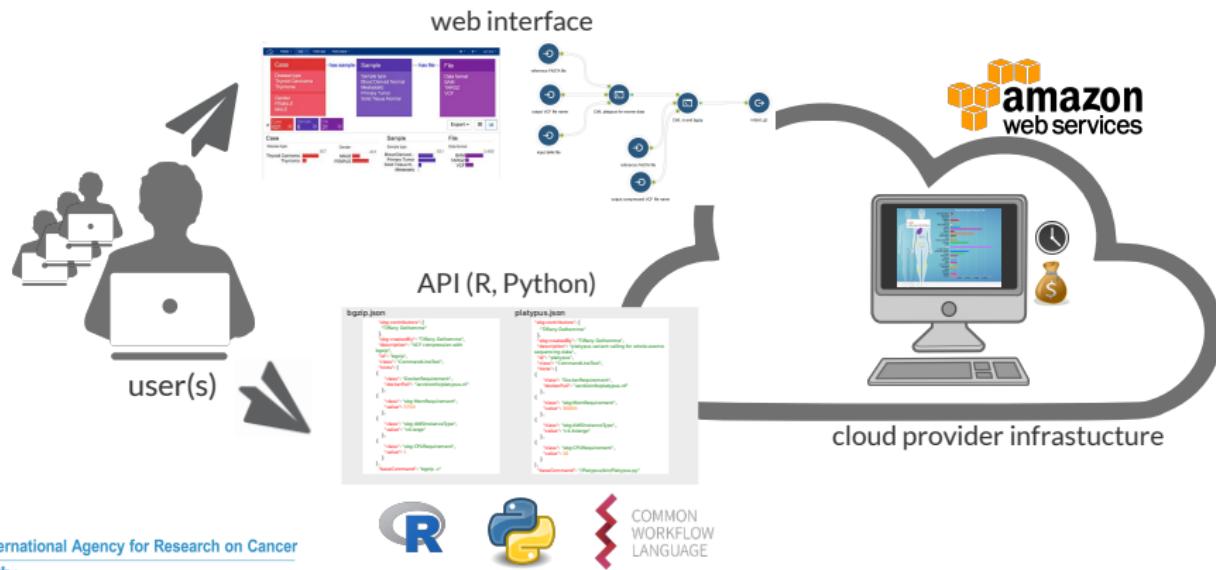
The CancerGenomicsCloud provides a simple and powerful way to analyze large quantities of genomic data:



Global overview

The Seven Bridges Platform

The CancerGenomicsCloud provides a simple and powerful way to analyze large quantities of genomic data:



Global overview

The Seven Bridges Platform

Seven Bridges User flow chart:

ACCESS	Access via the visual interface , and/or Access programmatically via our API .	ANALYSIS	Bring and store your own tools and workflows to analyze your data. and/or Use our ready-to-run bioinformatic analysis tools and workflows
PROJECT	Create a workspace for you and your team to store your data, perform analyses, and view results. Choose who has access to your project resources.	STORAGE and COMPUTE	Use customizable settings to optimize and parallelize analyses .
DATA	Bring and store your own files , and/or Use data from our public genomics reference files , and/or Easily query and use data from our large, publicly-hosted genomics datasets .	PRICING	We pass all charges from your cloud provider without markup .

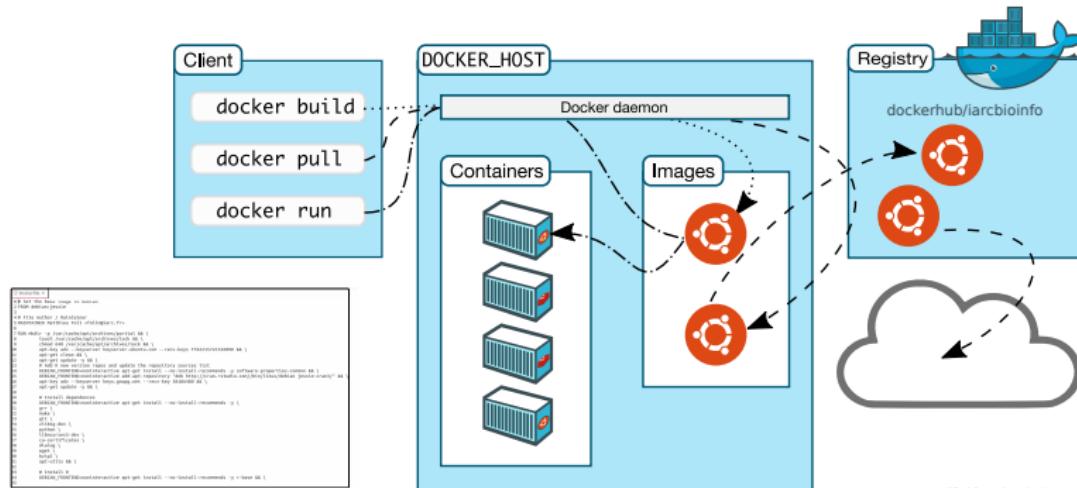
© 2014 Seven Bridges Genomics, Inc. All rights reserved.

Global overview

Docker for the deployment

How efficiently deploy your tool in the cloud?

⇒ Docker to embed it in a virtual container that can run on **any machine**



modified from docs.docker.com

MATERIAL

Web interface to facilitate integration with the platform

Use to get data in the course

Web interface features

The screenshot shows a web-based application interface with the following sections:

- Projects:** A list of projects:
 - IARC_course_tutorial (CONTROLED)
 - TCGA THCA genome
 - TCGA CHOL genome
 - TCGA UCEC genome
 - TCGA DLBC genome
- Create a project:** A button to start a new project.
- Public Data and Apps:** Statistics and links:
 - Analyze: 549,625 publicly available files from TCGA and other major datasets. Buttons: Overview, Cases, Browse Data.
 - Use some of 200 publicly available Tools and Workflows. Button: Browse Apps.
- Getting started:** A list of steps to get started:
 - Your account has been successfully created. Check your account settings.
 - Walk through the QuickStart.
 - Create a project.
 - Invite a collaborator.
 - Browse cases using the Case Explorer.
 - Search datasets using the Data Browser.
 - Upload your private data to analyze it along with public datasets.
 - Bring your tools using the SDK.
 - Build a workflow using the workflow editor.
 - Get your authentication token and follow the API Quickstart.
 - Join our growing user community on the CGC forum.
- Analyses:** A list of tasks:
 - Completed: sanctools_header run - 02-06-18 13:09:57

Web interface to facilitate integration with the platform

Use to get data in the course

1. Manage your project

The screenshot shows the cBioPortal web interface. At the top, there is a navigation bar with links for 'Projects', 'Data', 'Public Apps', 'Public projects', and 'Developer'. The 'Projects' link is highlighted with a red box and a cursor arrow pointing to it. Below the navigation bar, there is a section titled 'Projects' with a search bar. It lists several projects:

- UARC_course_tutorial (Created by tdelhomme - Jan. 22, 2018 13:58)
- TCGA THCA germline (Created by tdelhomme - Nov. 27, 2017 14:07)
- TCGA CHOL germline (Created by tdelhomme - Nov. 17, 2017 17:02)
- TCGA UCB germline (Created by tdelhomme - Nov. 17, 2017 18:16)
- TCGA DLBC germline (Created by tdelhomme - Nov. 17, 2017 18:36)

Below the project list, there are two buttons: 'Create a project' (highlighted with a green box) and 'View all'. To the right of the project list, there is a 'Getting started' section with a list of steps, many of which are checked off:

- Your account has been successfully created. Check your account settings.
- Walk through the Quickstart.
- Create a project. (This step is checked off.)
- Invite a collaborator.
- Browse cases using the Case Explorer.
- Search datasets using the Data browser.
- Upload your private data to analyze it along with public datasets.
- Bring your tools using the SDK.
- Build a workflow using the workflow editor.
- Get your authentication token and follow the API Quickstart.
- Join our growing user community on the CGC forums.

At the bottom of the interface, there is a 'Hide this forever' button. On the right side, there is a 'Analyses' section with a 'Tasks' list:

- Completed: samtools_header run - 02-06-18 13:09:57

Web interface to facilitate integration with the platform

Use to get data in the course

2. Explore available data

The screenshot shows the cBioPortal web interface. At the top, there is a navigation bar with links for 'Projects', 'Data +', 'Public Apps', 'Public projects', and 'Developer'. The 'Data +' link is highlighted with a red box and a cursor arrow pointing to it. Below the navigation bar, there are three main sections: 'Projects', 'Public Data and Apps', and 'Analyses'. The 'Projects' section lists several projects, including 'IARC_cancer_tutorial' (selected), 'TCGA THCA germline', 'TCGA CHOL germline', 'TCGA UCEC germline', and 'TCGA DLBC germline'. The 'Public Data and Apps' section features two large callout boxes: one for 'Analyze 549,625 publicly available files from TCGA and other major datasets' and another for 'Use some of 200 publicly available Tools and Workflows'. The 'Analyses' section shows a single task named 'saantools_header run - 02-06-18 13:09:57' with a status of 'COMPLETED'. On the right side of the interface, there is a 'Getting started' sidebar with a list of steps, many of which are checked off.

- ✓ Your account has been successfully created. Check your account settings.
- ✓ Walk through the Quickstart.
- ✓ Create a project.
- Invite a collaborator.
- Browse cases using the Case Explorer.
- Search datasets using the Data browser.
- Upload your private data to analyze it along with public datasets.
- Bring your tools using the SDK.
- Build a workflow using the workflow editor.
- Get your authentication token and follow the API Quickstart.
- Join our growing user community on the CGC forums.

Web interface to facilitate integration with the platform

Use to get data in the course

3. Explore public apps

The screenshot shows the cBioPortal web interface. At the top, there is a navigation bar with tabs: 'Projects', 'Data', 'Public Apps' (which is highlighted with a red box and a red arrow pointing to it), 'Public projects', and 'Developer'. Below the navigation bar, there is a 'Projects' section listing several datasets:

- TCGA UARC breast cancer (Created by Delhomme - Jan. 22, 2018 13:58)
- TCGA THCA gliomma (Created by Delhomme - Nov. 27, 2017 14:07)
- TCGA CHOL gliomma (Created by Delhomme - Nov. 17, 2017 17:02)
- TCGA UCB gliomma (Created by Delhomme - Nov. 17, 2017 18:06)
- TCGA DLBC gliomma (Created by Delhomme - Nov. 17, 2017 18:06)

Below the projects section, there is a 'Public Data and Apps' section with two main sections: 'Analyze' and 'Use some of'. The 'Analyze' section displays the number '549,625' and the text 'publicly available files from TCGA and other major datasets'. The 'Use some of' section displays the number '200' and the text 'publicly available Tools and Workflows'. At the bottom of this section are buttons for 'Overview', 'Cases', and 'Browser Data'. To the right of these sections is a 'Getting started' sidebar with a list of steps, many of which are checked off:

- Your account has been successfully created. Check your account settings.
- Walk through the Quickstart.
- Create a project.
- Invite a collaborator.
- Browse cases using the Case Explorer.
- Search datasets using the Data browser.
- Upload your private data to analyze it along with public datasets.
- Bring your tools using the SDK.
- Build a workflow using the workflow editor.
- Get your authentication token and follow the API Quickstart.
- Join our growing user community on the CGC forums.

At the very bottom of the page, there is a footer with the text 'International Agency for Research on Cancer' and the World Health Organization logo. On the right side of the footer, there are several small icons for navigating between slides or sections.

Web interface to facilitate integration with the platform

Use to get data in the course

4. Get developer requests

The screenshot shows the cBioPortal web interface. At the top, there is a navigation bar with links for 'Projects', 'Data', 'Public Apps', 'Public projects', and 'Developer'. The 'Developer' link is highlighted with a red box and a cursor arrow pointing to it. Below the navigation bar, there is a 'Projects' section listing several projects: 'IARC_cancer_tutorial' (created by Delhomme on Jan. 22, 2018 13:58), 'TCGA THCA germline' (created by Delhomme on Nov. 27, 2017 14:07), 'TCGA CHOL germline' (created by Delhomme on Nov. 17, 2017 17:01), 'TCGA UCEC germline' (created by Delhomme on Nov. 17, 2017 18:16), and 'TCGA DLBC germline' (created by Delhomme on Nov. 17, 2017 18:16). There are buttons for 'Create a project' and 'View all'. Below the projects is a 'Public Data and Apps' section. It features two large boxes: one for 'Analyze' showing '549,625' publicly available files from TCGA and other major datasets, and another for 'Use some of 200' publicly available Tools and Workflows. Buttons for 'Overview', 'Cases', and 'Browse Data' are under the analyze box, and 'Browse Apps' is under the tools box. To the right of these sections is a 'Getting started' sidebar with a list of items, many of which are checked off with green checkmarks. The items include: 'Your account has been successfully created. Check your account settings.', 'Walk through the Quickstart.', 'Create a project.', 'Invite a collaborator.', 'Browse cases using the Case Explorer.', 'Search datasets using the Data browser.', 'Upload your private data to analyze it along with public datasets.', 'Bring your tools using the SDK.', 'Build a workflow using the workflow editor.', 'Get your authentication token and follow the API Quickstart.', and 'Join our growing user community on the CGC forums.' A 'Hide this forever' button is at the bottom of this sidebar. At the very bottom of the page, there is a footer with the International Agency for Research on Cancer logo, the World Health Organization logo, and a navigation bar with icons for back, forward, search, and other functions.

Web interface to facilitate integration with the platform

Use to get data in the course

5. Get account details

The screenshot shows the cBioPortal web interface. At the top, there is a navigation bar with links for Projects, Data, Public Apps, Public projects, and Developer. On the far right of the header, there is a dropdown menu labeled "tiffomome" with a red arrow pointing to it. Below the header, the main content area is divided into several sections:

- Projects:** A list of projects:
 - IARC course tutorial** (CONTROLLER, created by tiffomome on Jan 22, 2018 13:58)
 - TCGA THCA genome** (created by tiffomome on Nov 27, 2017 14:07)
 - TCGA CHOL genome** (created by tiffomome on Nov 17, 2017 17:01)
 - TCGA UCEC genome** (created by tiffomome on Nov 17, 2017 18:16)
 - TCGA DLBC genome** (created by tiffomome on Nov 17, 2017 18:50)
- Create a project:** A button to start a new project.
- Public Data and Apps:** Two main sections:
 - Analyze:** Shows a large number **549,625** and a link to "publicly available files from TCGA and other major datasets". Buttons for Overview, Cases, and Browse Data are present.
 - Use some of 200:** Shows a large number **200** and a link to "publicly available Tools and Workflows". A button for Browse Apps is present.
- Getting started:** A list of steps:
 - Your account has been successfully created. Check your account settings.
 - Walk through the QuickStart.
 - Create a project.
 - Invite a collaborator.
 - Browse cases using the Case Explorer.
 - Search datasets using the Data Browser.
 - Upload your private data to analyze it along with public datasets.
 - Bring your tools using the SDK.
 - Build a workflow using the workflow editor.
 - Get your authentication token and follow the API Quickstart.
 - Join our growing user community on the CGC forum.
- Analyses:** A section showing a completed task: **santools_header run - 02-06-18 13:09:57**.

Multi-users projects to provide collaboration

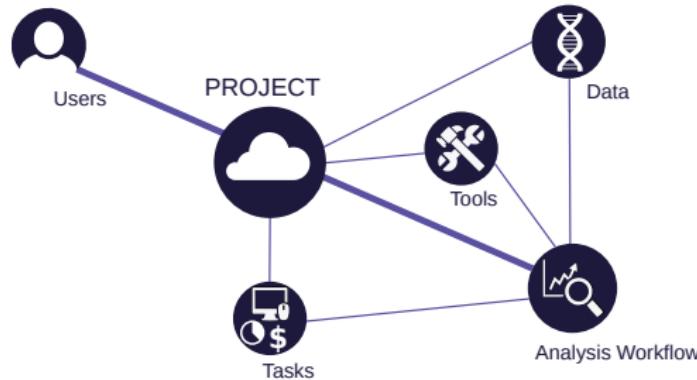
- projects are the core building **blocks** of the CGC

Multi-users projects to provide collaboration

- projects are the core building **blocks** of the CGC
- one project corresponds to a particular scientific **investigation**

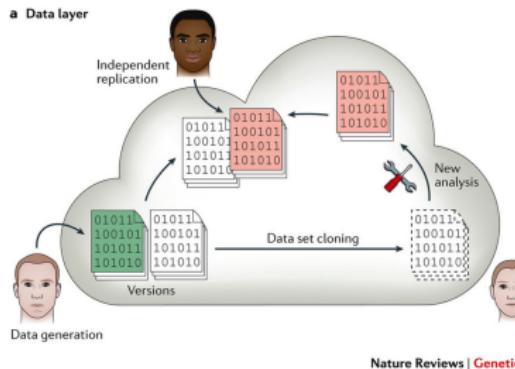
Multi-users projects to provide collaboration

- projects are the core building **blocks** of the CGC
- one project corresponds to a particular scientific **investigation**
- it serves to package **data, analysis workflow** and **results**



Multi-users projects to provide collaboration

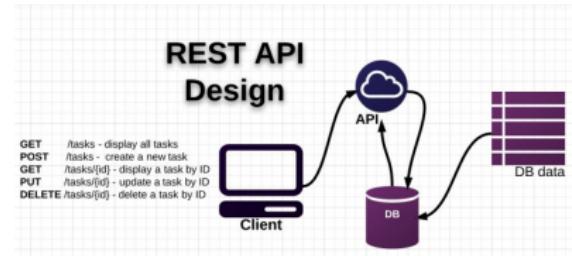
- projects are the core building **blocks** of the CGC
- one project corresponds to a particular scientific **investigation**
- it serves to package **data, analysis workflow and results**
- project **members** are the collaborators (individual defined rights)



API to provide reproducibility

API: application programming interface

- API allows you to interact with the SBG-CGC as you would via the visual interface.
- SBG-CGC API is a set of RESTful web services intended for integration and automation



- the aim of this API is to read/write information on the platform, to automate analyses executions

API to provide reproducibility

Focus on R interface for the API in the course

R/Bioconductor package: [sevenbridges-r](#)

- Authentication via CGC token
 - preserve access rights

API to provide reproducibility

Focus on R interface for the API in the course

R/Bioconductor package: [sevenbridges-r](#)

- Authentication via CGC token
 - preserve access rights
- Access to and manipulate data
 - access to your project and work on it

API to provide reproducibility

Focus on R interface for the API in the course

R/Bioconductor package: [sevenbridges-r](#)

- Authentication via CGC token
 - preserve access rights
- Access to and manipulate data
 - access to your project and work on it
- Build tools and workflows
 - use CWL Tool interface or import JSON

API to provide reproducibility

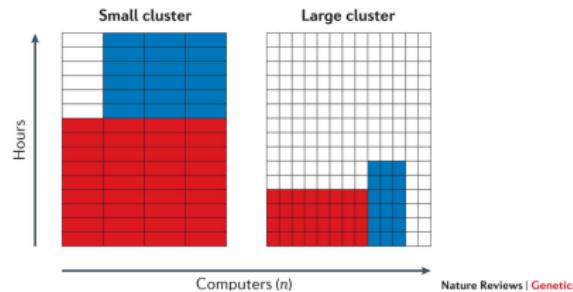
Focus on R interface for the API in the course

R/Bioconductor package: [sevenbridges-r](#)

- Authentication via CGC token
 - preserve access rights
- Access to and manipulate data
 - access to your project and work on it
- Build tools and workflows
 - use CWL Tool interface or import JSON
- Run reproducible analyses
 - loop over files or run on batches

CGC provides elasticity

Amazon Web Services Instances

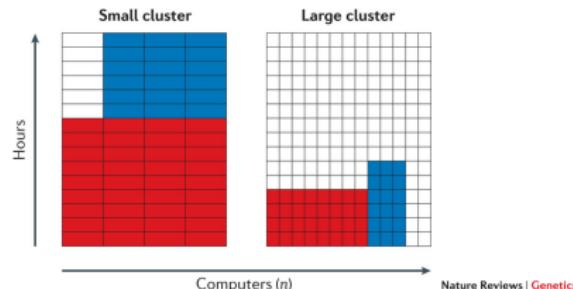


Each tool is executed on a computation instance in the cloud

- instances are virtual computers

CGC provides elasticity

Amazon Web Services Instances

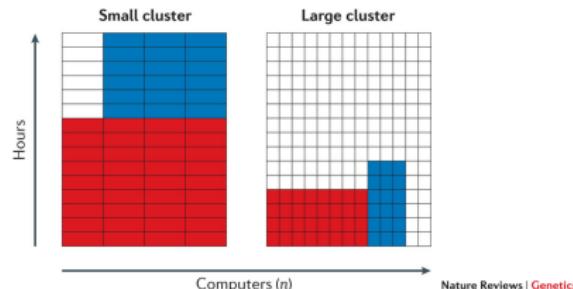


Each tool is executed on a computation instance in the cloud

- instances are virtual computers
- each instance type has a particular allocations of CPU and memory

CGC provides elasticity

Amazon Web Services Instances



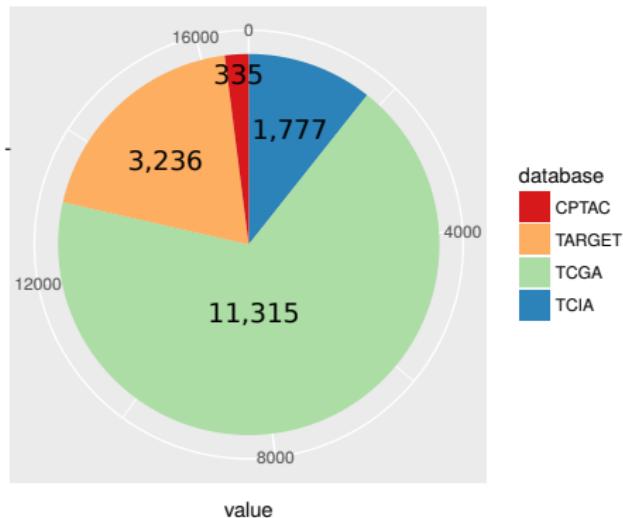
Each tool is executed on a computation instance in the cloud

- instances are virtual computers
- each instance type has a particular allocations of CPU and memory
- pay-as-you-go (PAYG) pricing principles
- available instances are named and catalogued

Available data in the CGC cloud

Public databases + private data

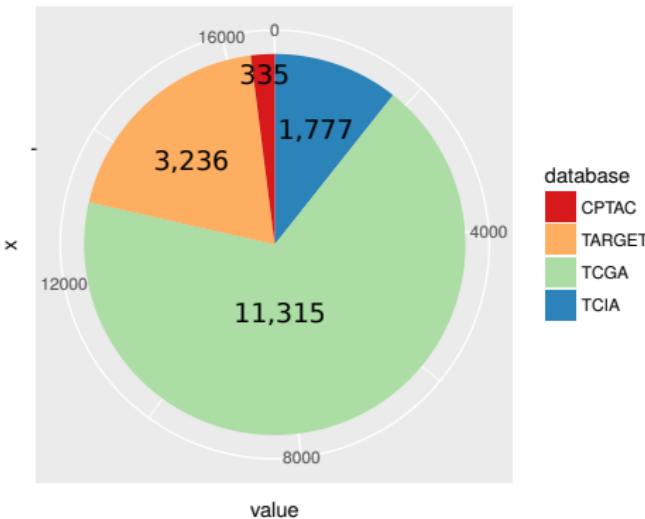
CGC data – cases



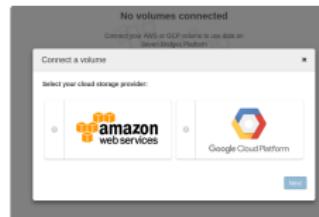
Available data in the CGC cloud

Public databases + private data

CGC data – cases



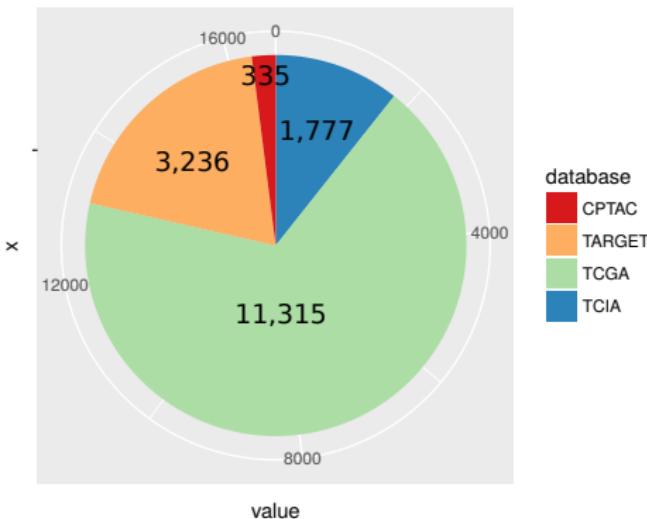
+ AWS or GCP cloud volumes



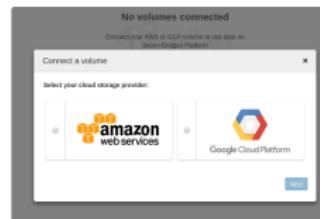
Available data in the CGC cloud

Public databases + private data

CGC data – cases



+ AWS or GCP cloud volumes



+ local data

- CGC web interface Uploader
- CGC command line uploader
- CGC APIs

ANALYSES

Public and private tools

- a **public app** can be used inside a project once copied, and then can be edited
- a **private app** should be defined in term of I/O
- a **workflow** is a chain of connected tools

PUBLIC TOOLS

Screenshot of a web interface showing a list of public bioinformatics tools:

- RNA-seq Alignment - STAR**: STAR 2.4.0b. Alignment to a reference genome and transcriptome presents the first step of RNA-Seq analysis. This pipeline uses STAR.
- Whole Exome Sequencing GATK 2.3.9.-HLI**: HLD pipeline applies an open-coding HLD pipeline to a genome (known as Exome). The exome is extracted to comprise ~1%
- Fusion Transcript Detection - ChimericScan**: Fusion Transcript Detector - ChimericScan detects and sterilizes fusion transcripts from paired-end RNA-seq.
- Veriscan2 Workflow from BAM**: Veriscan2 workflow is an extended methodology supported by the authors, where user can start analysis from BAM.
- FastQC Analysis**: The FastQC tool, developed by the Broad Institute, analyzes sequence data from FASTQ, BAM, or SAM files. It provides quality control (QC) processing.
- Alignment Metrics QC**: BGQ tool 1. Running the pipeline will process your alignment files to generate the quality of your alignment. Provides

PRIVATE TOOLS



Analyses

Description of an analysis workflow

1. CWL tool description

logzip.json

```
["logCommand": 1, "Tiffany Delhomme", "logCommand": "Tiffany Delhomme", "logCommand": "logzip compression with", "logCommand": "args", "logCommand": "logzip", "logCommand": "CompressLogFile", "logCommand": "1", "logCommand": "Viken", "logCommand": "DockerRequirement", "logCommand": "dockerPull", "logCommand": "tinklebox/platypus:al", "logCommand": "Viken", "logCommand": "logMemRequirement", "logCommand": "value", "logCommand": "3750", "logCommand": "Viken", "logCommand": "logDiskRequirementType", "logCommand": "value", "logCommand": "10000", "logCommand": "Viken", "logCommand": "logCPURequirement", "logCommand": "value", "logCommand": "1", "logCommand": "Viken", "logCommand": "logCommand", "logCommand": "logzip -z"}
```

platypus.json

```
["logCommand": 1, "Tiffany Delhomme", "logCommand": "Tiffany Delhomme", "logCommand": "platypus", "logCommand": "platypus variant calling for whole-exome sequencing data", "logCommand": "args", "logCommand": "Viken", "logCommand": "CompressLogFile", "logCommand": "1", "logCommand": "Viken", "logCommand": "DockerRequirement", "logCommand": "dockerPull", "logCommand": "tinklebox/platypus:al", "logCommand": "Viken", "logCommand": "logMemRequirement", "logCommand": "value", "logCommand": "35000", "logCommand": "Viken", "logCommand": "logDiskRequirementType", "logCommand": "value", "logCommand": "100000", "logCommand": "Viken", "logCommand": "logCPURequirement", "logCommand": "value", "logCommand": "33", "logCommand": "Viken", "logCommand": "logCommand", "logCommand": "PlatypusOnPlatypus"}
```

Analyses

Description of an analysis workflow

1. CWL tool description

```
begin.json
{
  "id": "Workflow_1",
  "cwlVersion": "v1.0",
  "inputs": [
    {
      "name": "input_file",
      "type": "File"
    }
  ],
  "outputs": [
    {
      "name": "output_file",
      "type": "File"
    }
  ],
  "steps": [
    {
      "id": "Step_1",
      "label": "Step 1: Command Line Tool",
      "command": "ls"
    },
    {
      "id": "Step_2",
      "label": "Step 2: DockerRequirement",
      "image": "tiddlyanalysis/tiddlyanalysis:v0.1"
    },
    {
      "id": "Step_3",
      "label": "Step 3: SingularityRequirement",
      "image": "singularityhub/tiddlyanalysis:v0.1"
    }
  ]
}

platypus.json
{
  "id": "Workflow_1",
  "cwlVersion": "v1.0",
  "inputs": [
    {
      "name": "input_file",
      "type": "File"
    }
  ],
  "outputs": [
    {
      "name": "output_file",
      "type": "File"
    }
  ],
  "steps": [
    {
      "id": "Step_1",
      "label": "Step 1: Command Line Tool",
      "command": "ls"
    },
    {
      "id": "Step_2",
      "label": "Step 2: DockerRequirement",
      "image": "tiddlyanalysis/tiddlyanalysis:v0.1"
    },
    {
      "id": "Step_3",
      "label": "Step 3: SingularityRequirement",
      "image": "singularityhub/tiddlyanalysis:v0.1"
    }
  ]
}
```

2. CGC workflow integration



Analyses

Description of an analysis workflow

1. CWL tool description



2. CGC workflow integration



3. Querying data



Analyses

Description of an analysis workflow

1. CWL tool description



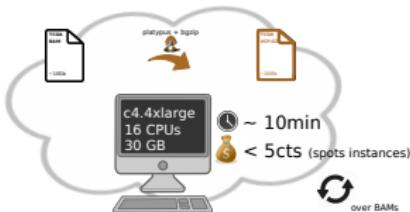
2. CGC workflow integration



3. Querying data



4. Running analyses



Analyses

example of task monitoring

tcga-kirc-germline (n=718)

