SBG-CGC: global architecture

Tiffany Delhomme

IARC course - analysing TCGA data in the cloud

28 feb 2018

International Agency for Research on Cancer
World Health Organization

Summary

International Agency for Research on Cancer

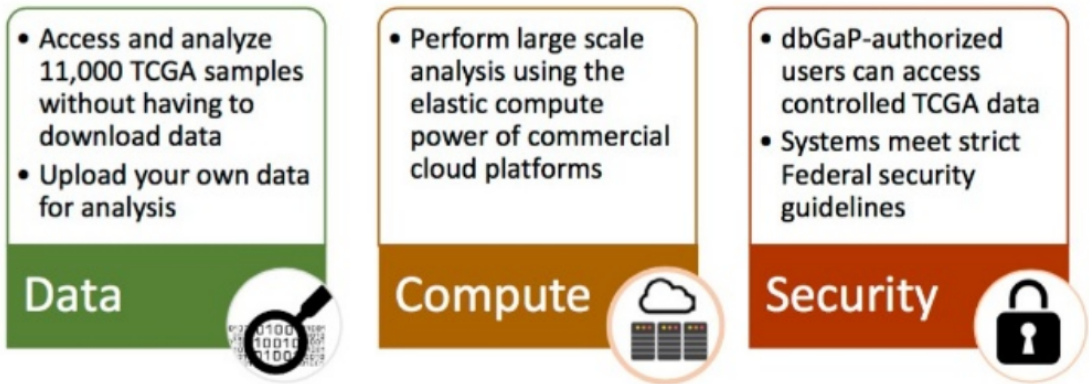
World Health Organization

NCI cloud pilots

Data, Computing and Security

Aim: providing big data ressources 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 ressources



International Agency for Research on Cancer

World Health Organization

Global overview

The Seven Bridges Plateform

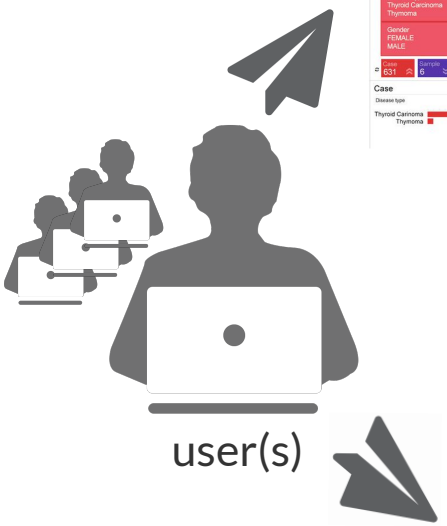
The CancerGenomicsCloud provides a simple and powerful way to analyze large quantities of genomic data:

user(s)

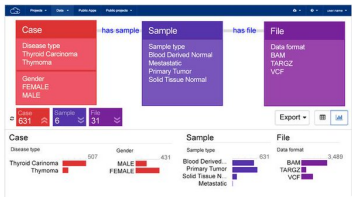
Global overview

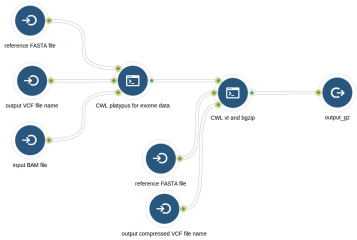
The Seven Bridges Plateform

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



web interface





API (R, Python)

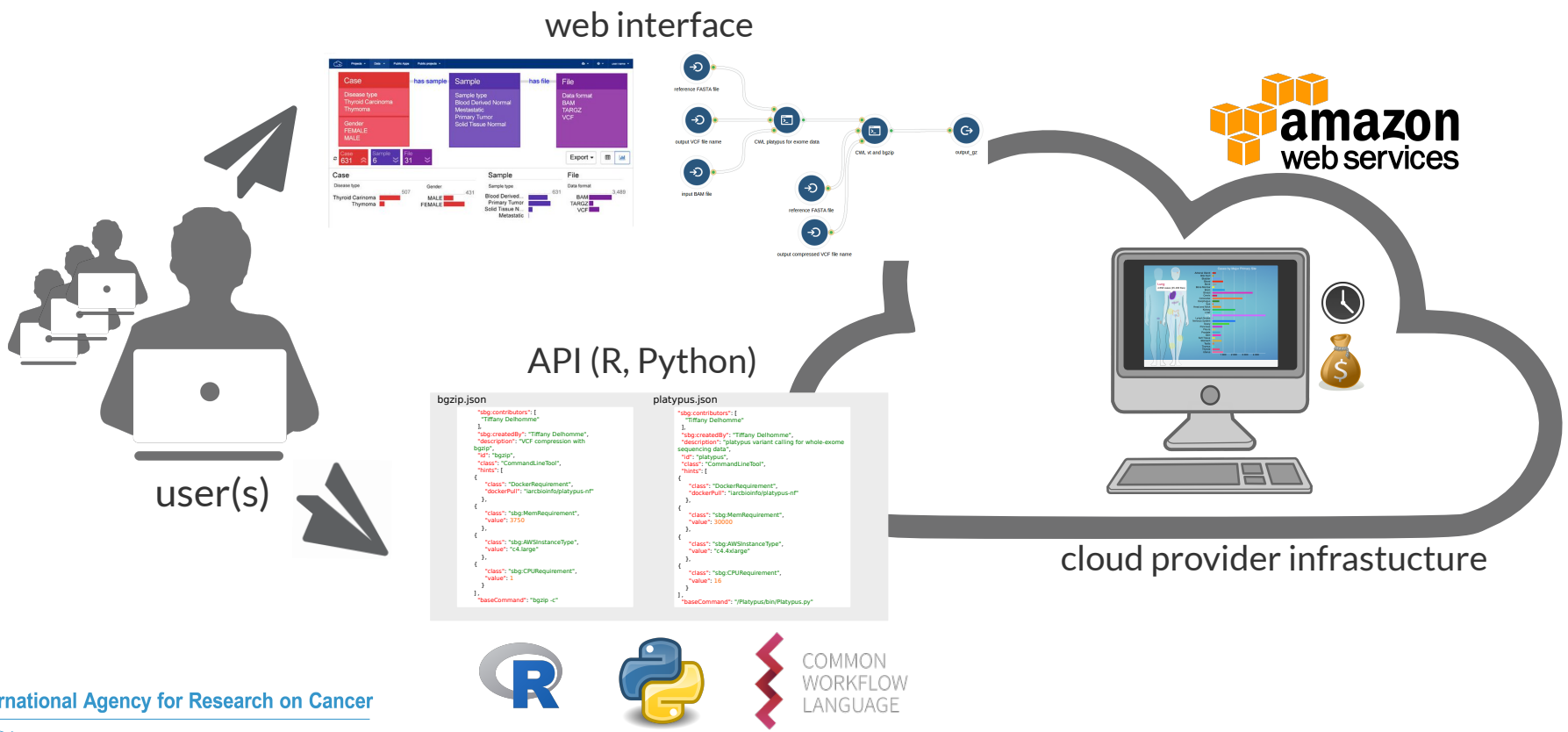
International Agency for Research on Cancer

World Health Organization

Global overview

The Seven Bridges Plateform

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



International Agency for Research on Cancer

World Health Organization

Global overview

The Seven Bridges Plateform

Seven Bridges **User flow chart**:

ACCESS	Access via the **visual interface**. and/or Access programmatically via our **API**.
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.
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**.
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
STORAGE and COMPUTE	Use customizable settings to **optimize and parallelize analyses**.
PRICING	We pass all charges from your cloud provider **without markup**.

International Agency for Research on Cancer

World Health Organization

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

International Agency for Research on Cancer

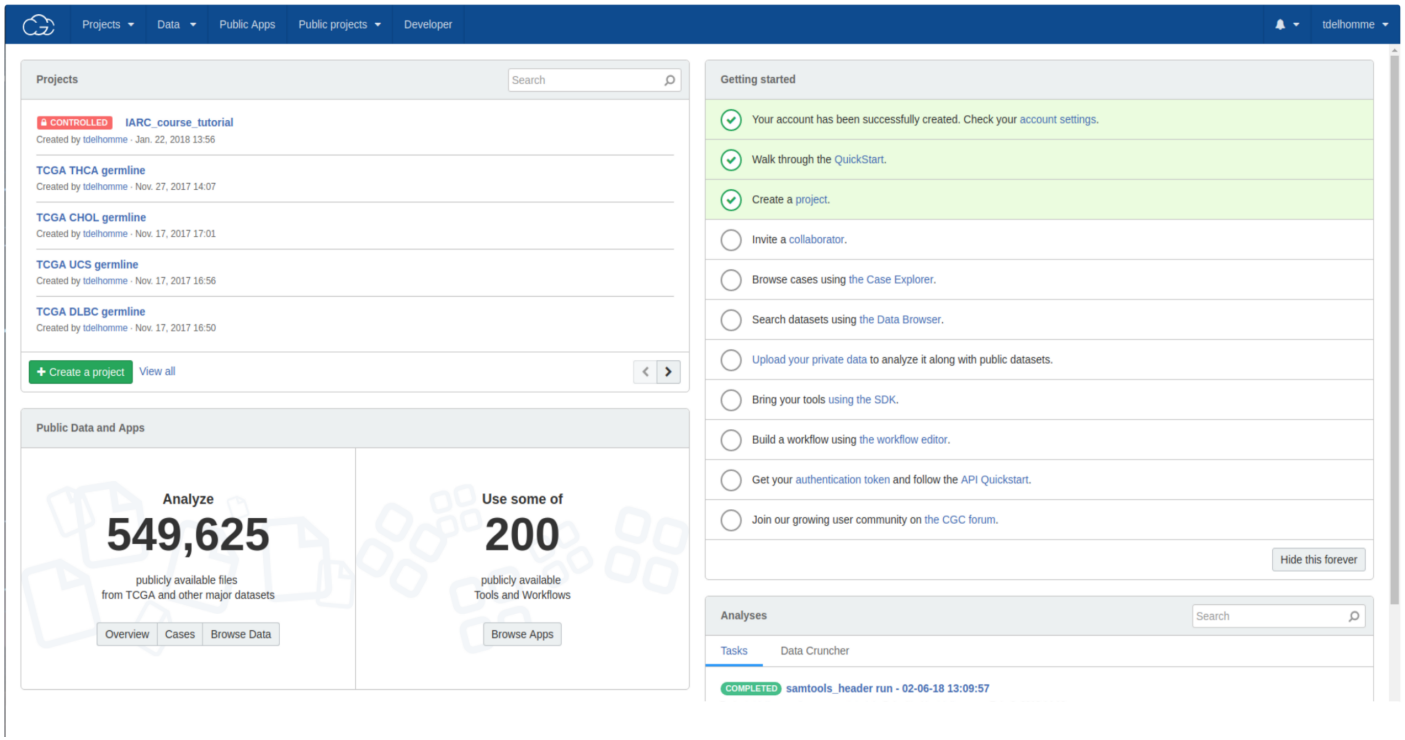
World Health Organization

MATERIAL

Web interface to facilitate integration with the plateform

Use to get data in the course

Web interface features



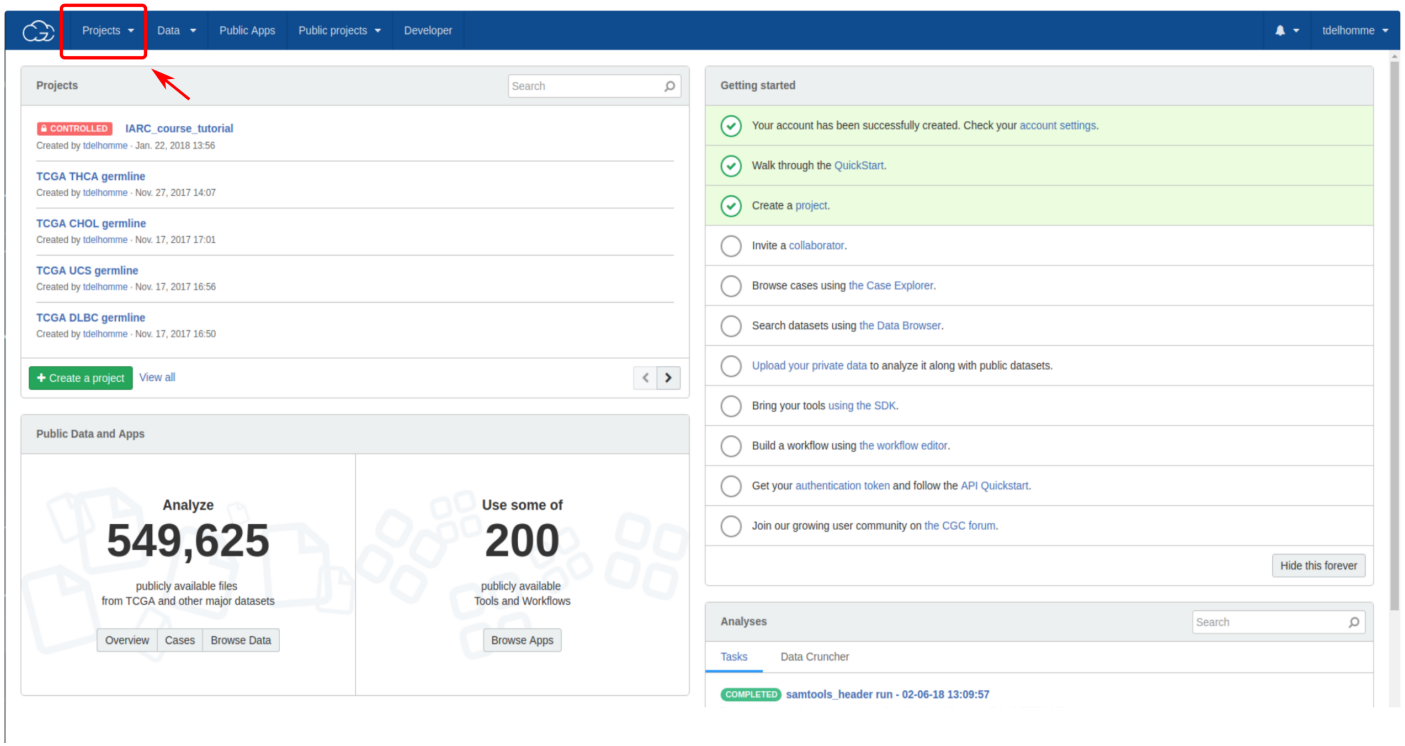
International Agency for Research on Cancer

World Health Organization

Web interface to facilitate integration with the plateform

Use to get data in the course

1. Manage your project



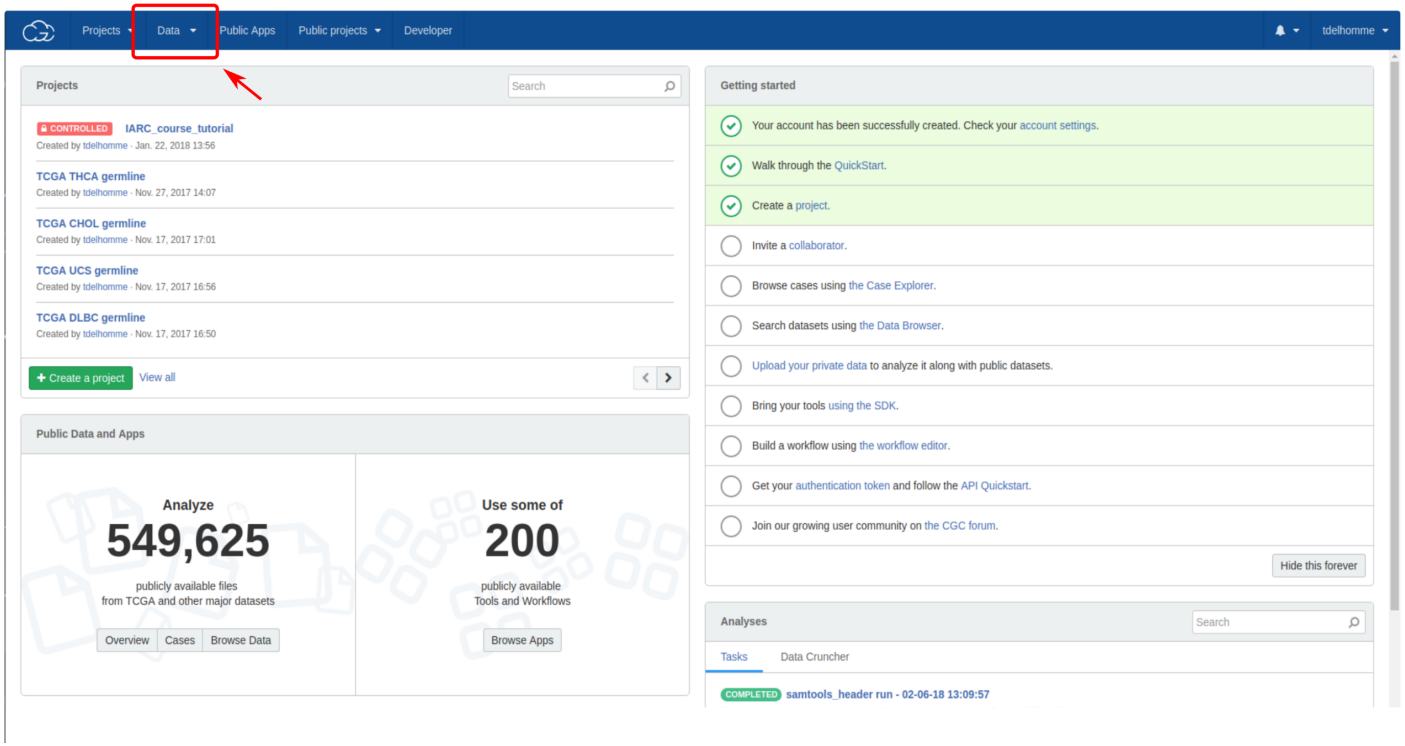
International Agency for Research on Cancer

World Health Organization

Web interface to facilitate integration with the plateform

Use to get data in the course

2. Explore available data



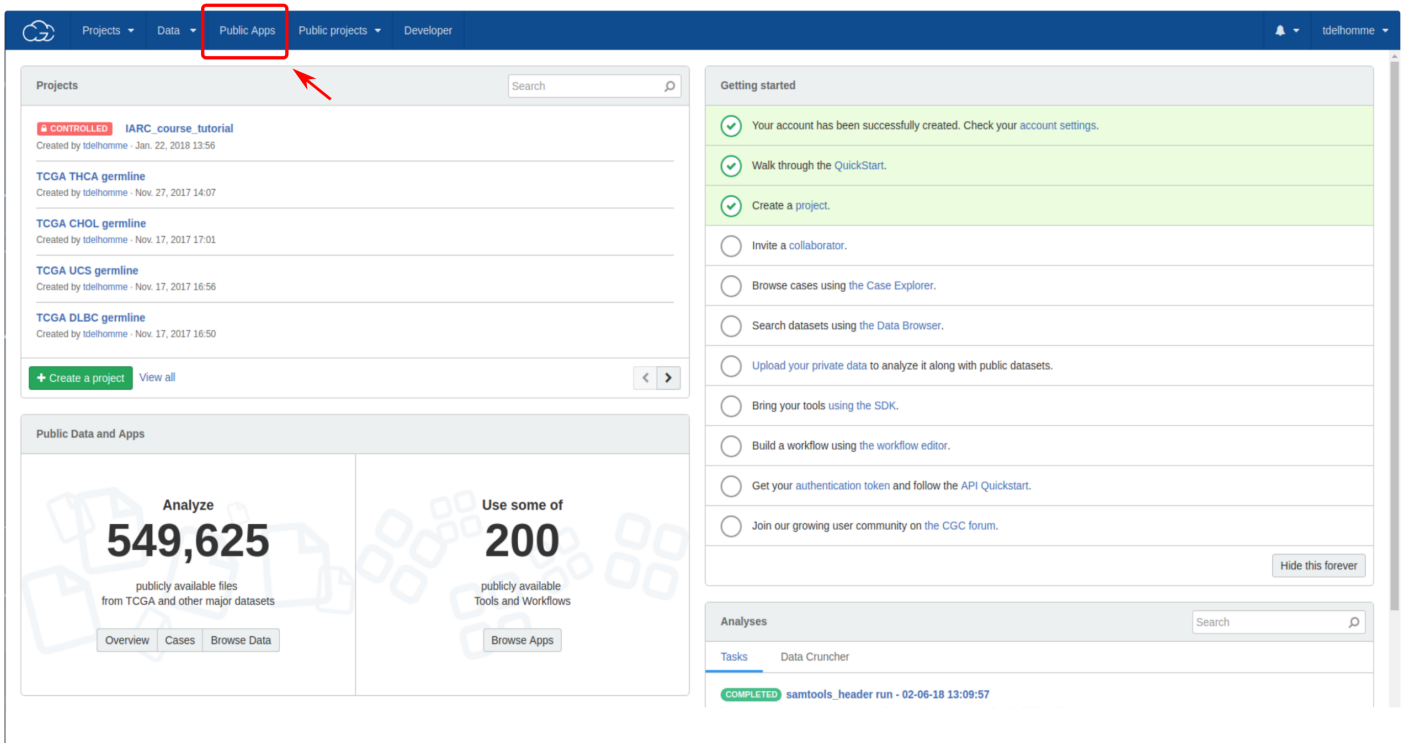
International Agency for Research on Cancer

World Health Organization

Web interface to facilitate integration with the plateform

Use to get data in the course

3. Explore public apps



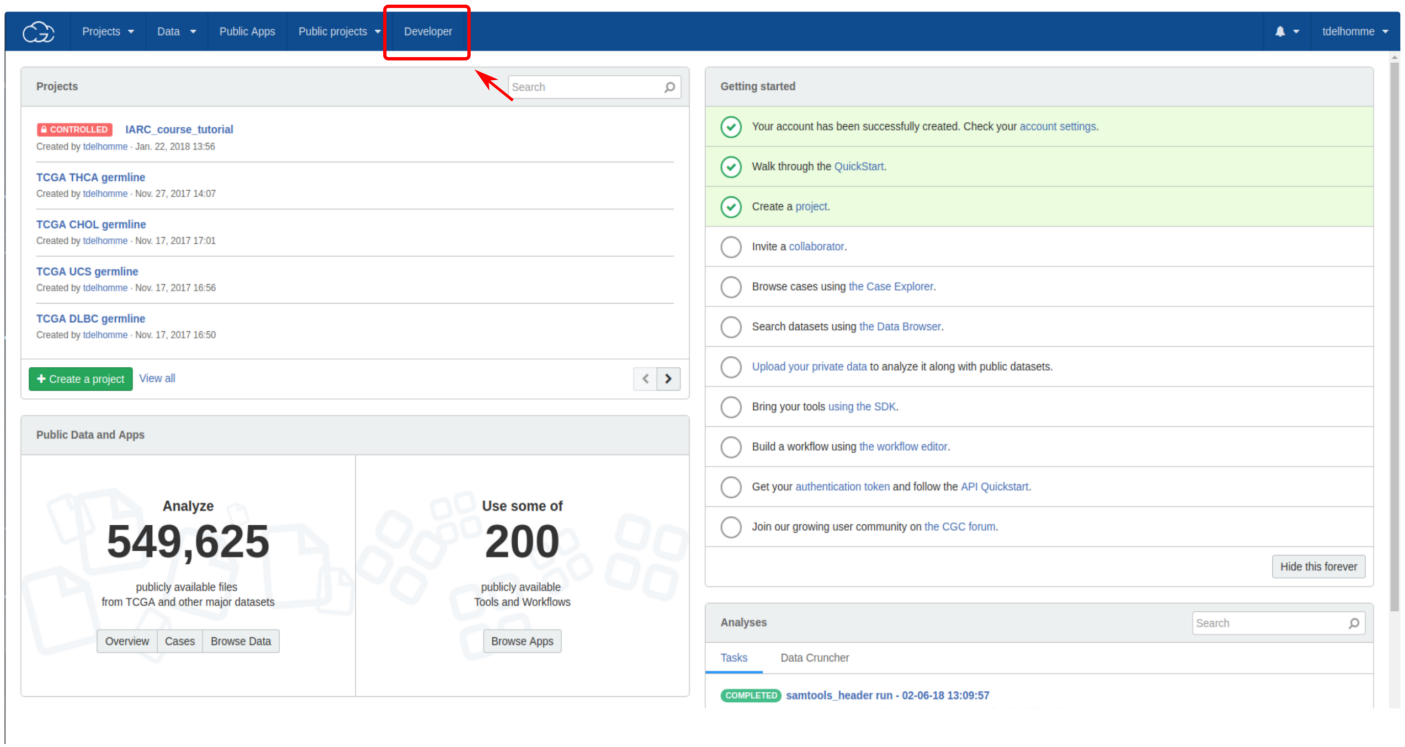
International Agency for Research on Cancer

World Health Organization

Web interface to facilitate integration with the plateform

Use to get data in the course

4. Get developer requests



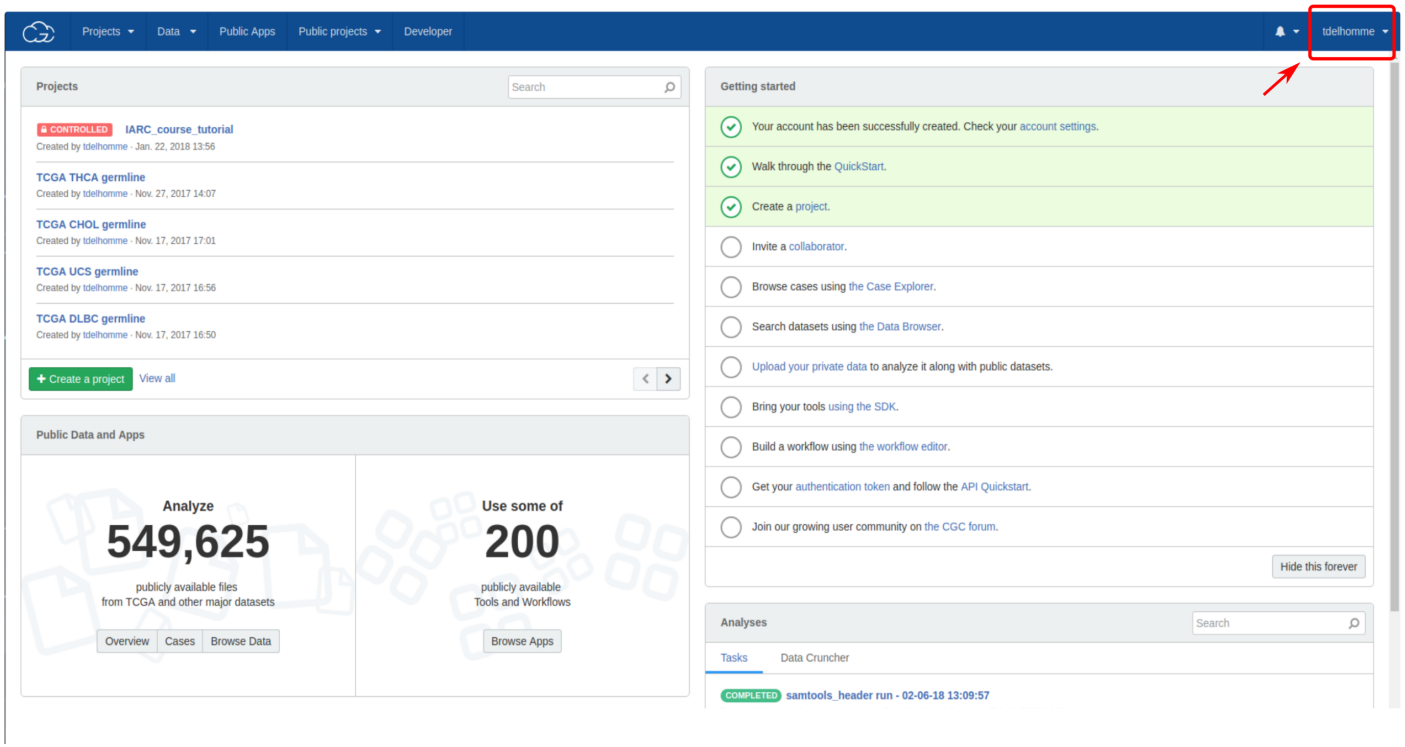
International Agency for Research on Cancer

World Health Organization

Web interface to facilitate integration with the plateform

Use to get data in the course

5. Get account details



International Agency for Research on Cancer

World Health Organization

Multi-users projects to provide collaboration

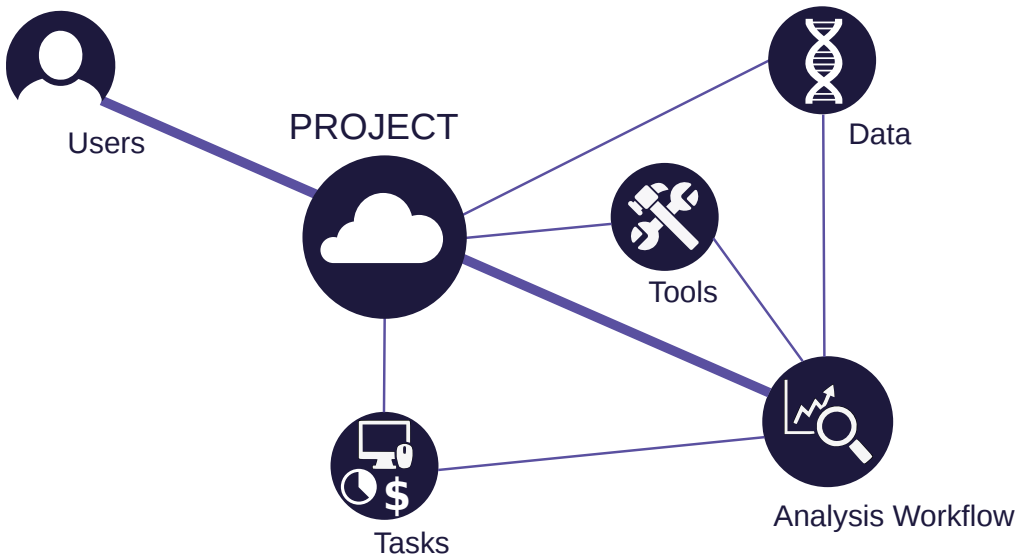
- projects are the core building **blocks** of the CGC

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

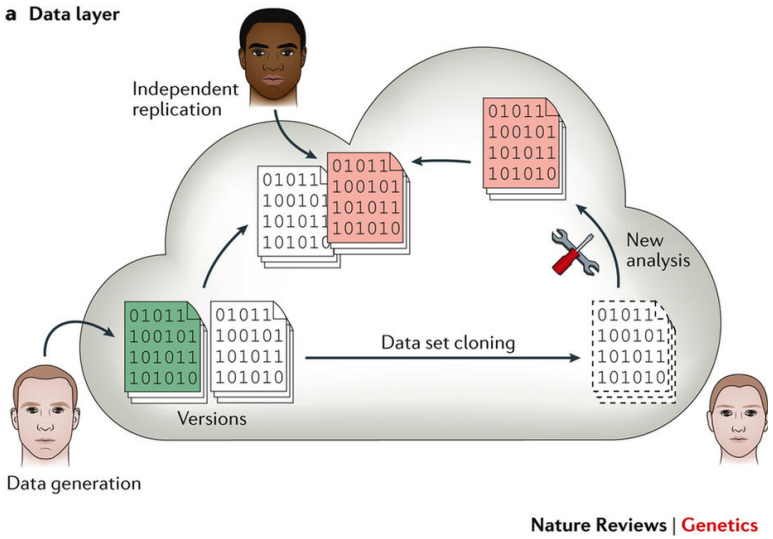
- projects are the core building **blocks** of the CGC
- one project corresponds to a particular scientific **investigation**
- it serves to package **data**, **analysis workflow** and **results**



International Agency for Research on Cancer
World Health Organization

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)

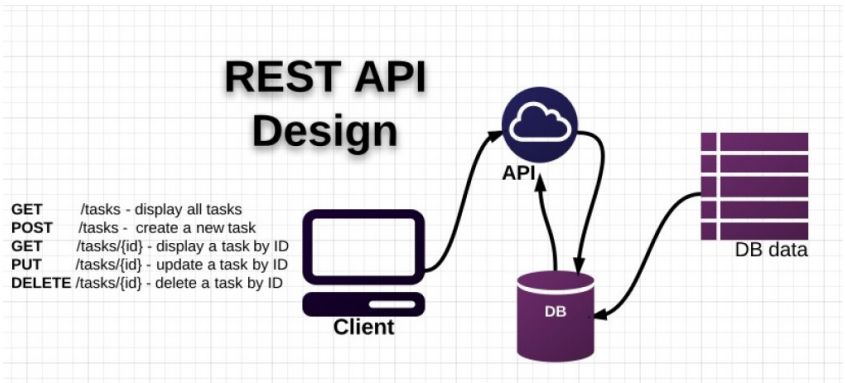


International Agency for Research on Cancer
World Health Organization

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

- Authentification 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

- Authentification 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

- Authentification 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

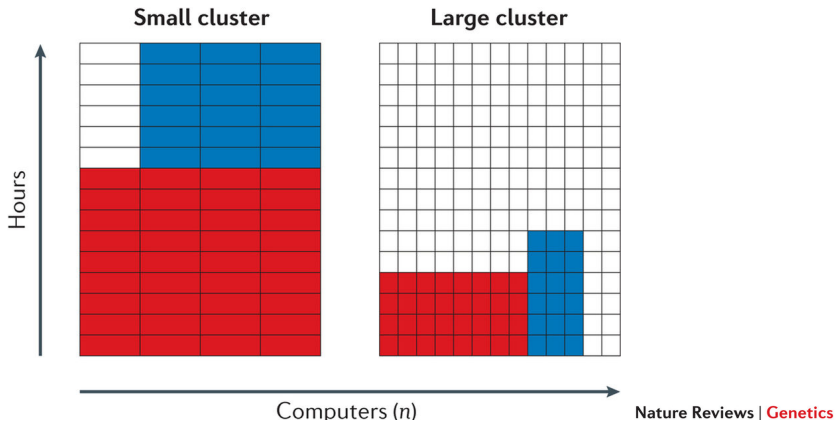
Focus on R interface for the API in the course

R/Bioconductor package: sevenbridges-r

- Authentification 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



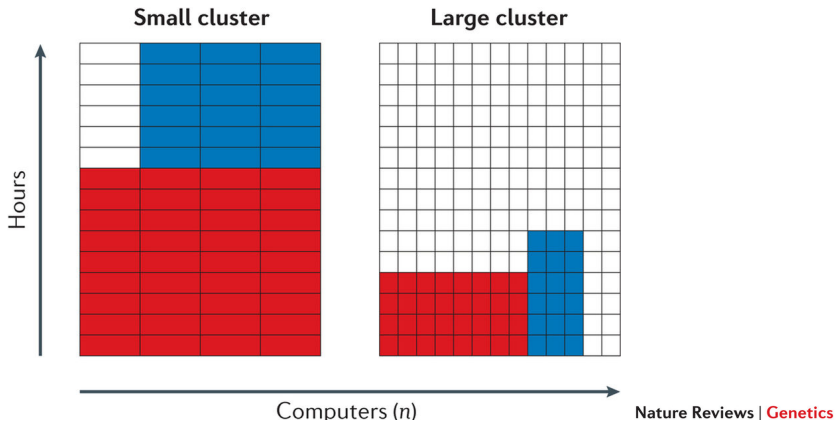
Each tool is executed on a computation instance in the cloud

- instances are virtual computers

International Agency for Research on Cancer
World Health Organization

CGC provides elasticity

Amazon Web Services Instances

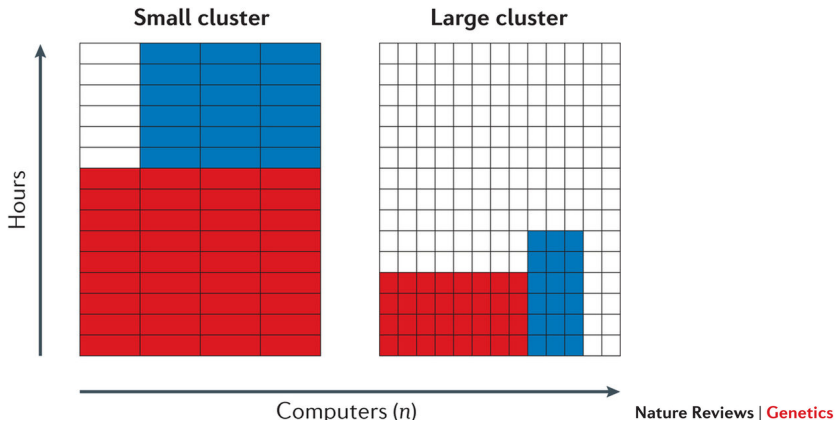


Each tool is executed on a computation instance in the cloud

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

CGC provides elasticity

Amazon Web Services Instances



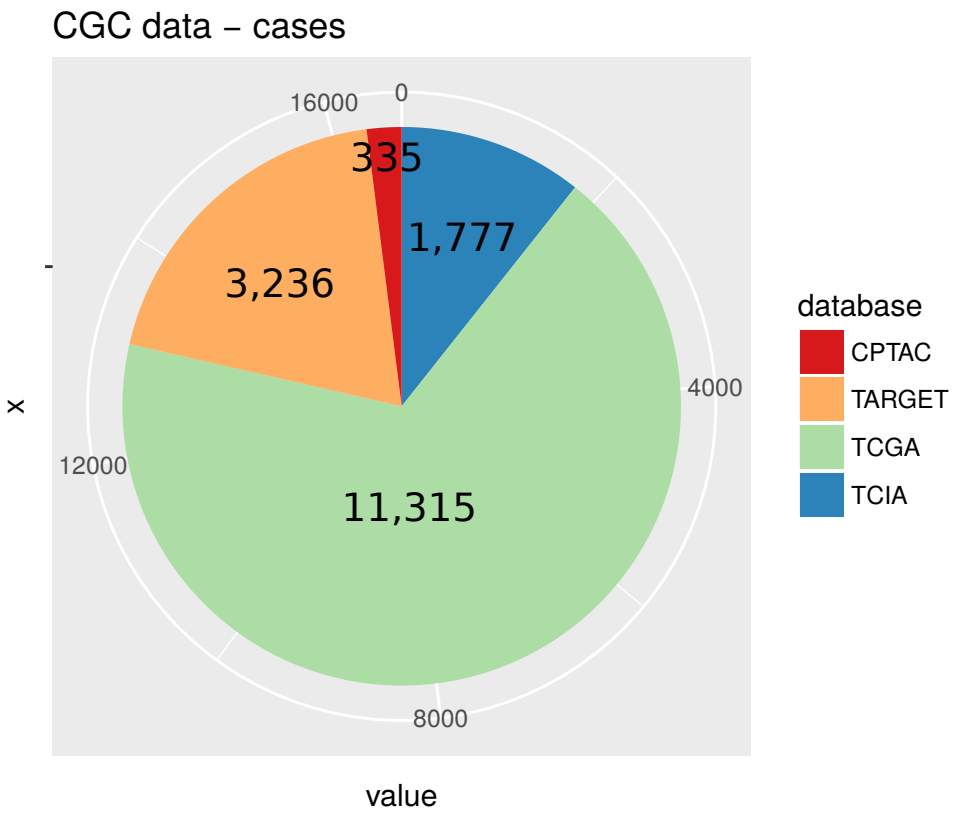
Each tool is executed on a computation instance in the cloud

- 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

International Agency for Research on Cancer
World Health Organization

Available data in the CGC cloud

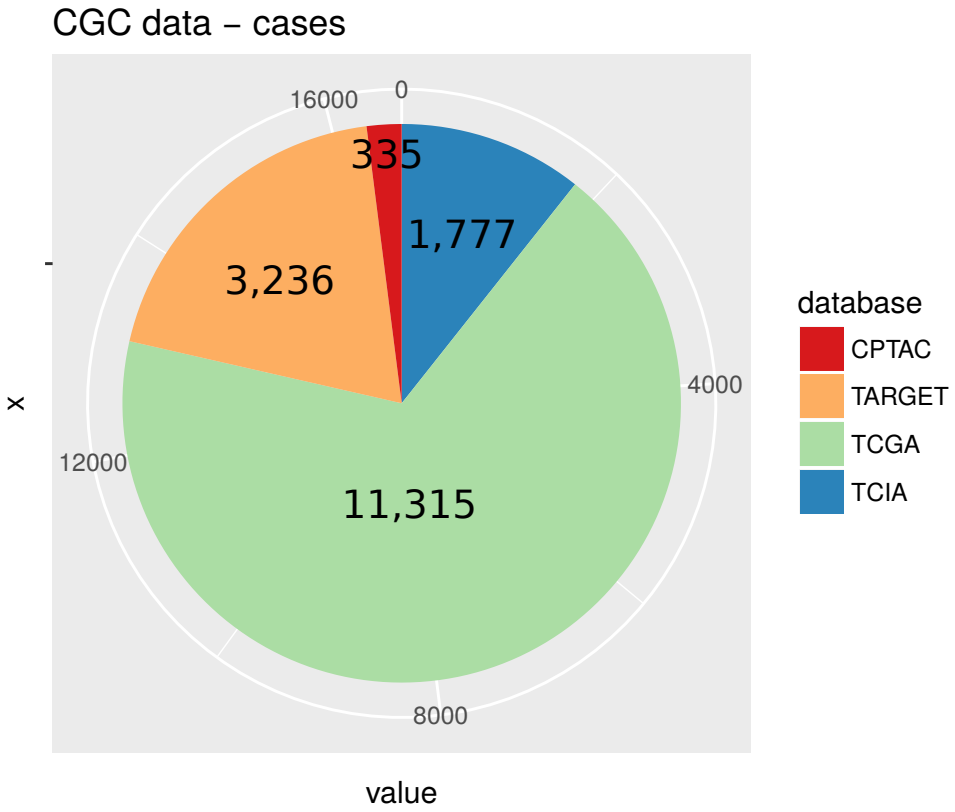
Public databases + private data



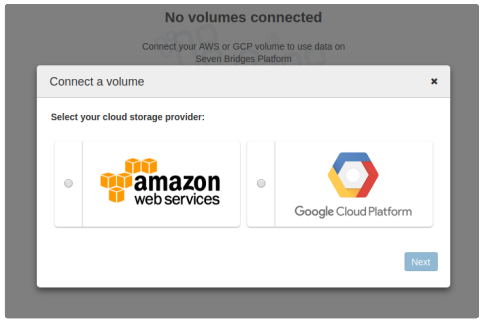
International Agency for Research on Cancer
World Health Organization

Available data in the CGC cloud

Public databases + private data



+ AWS or GCP cloud volumes

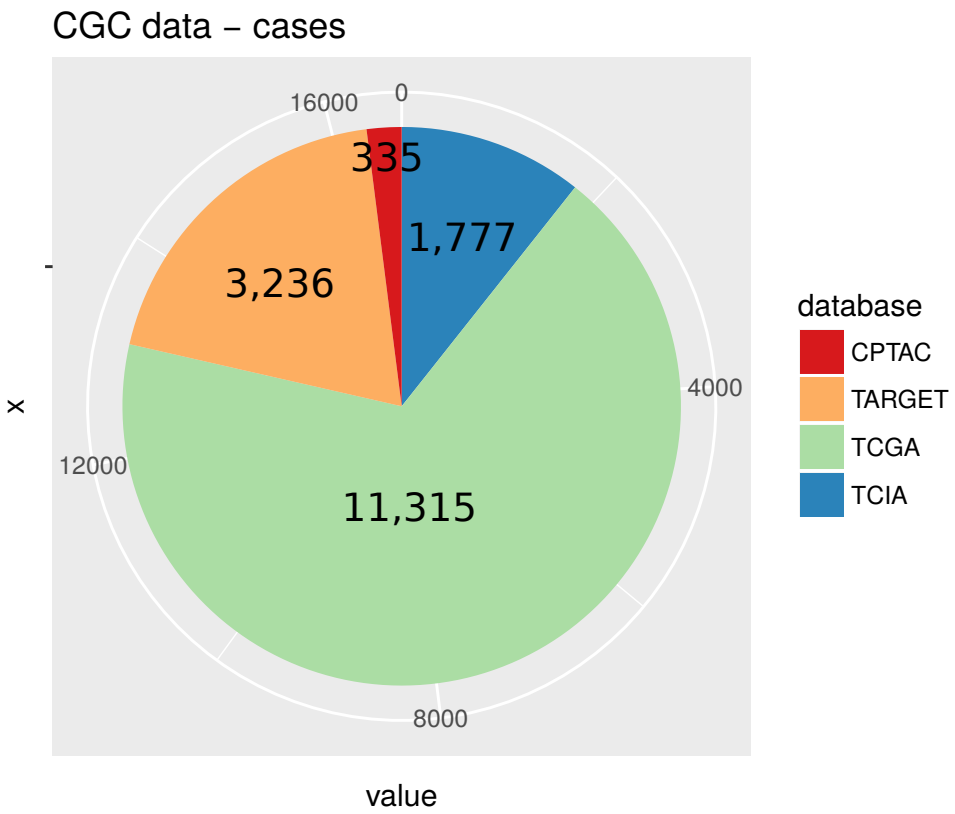


International Agency for Research on Cancer

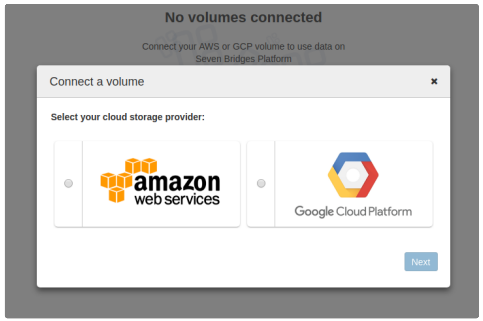
World Health Organization

Available data in the CGC cloud

Public databases + private data



+ AWS or GCP cloud volumes



+ local data

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

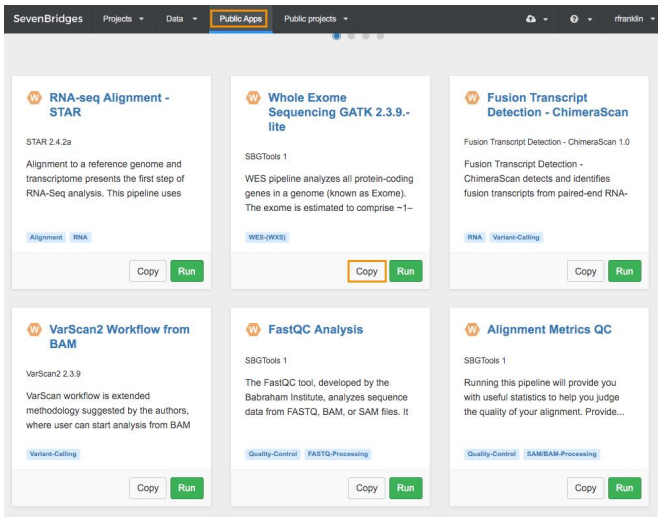
International Agency for Research on Cancer
World Health Organization

ANALYSES

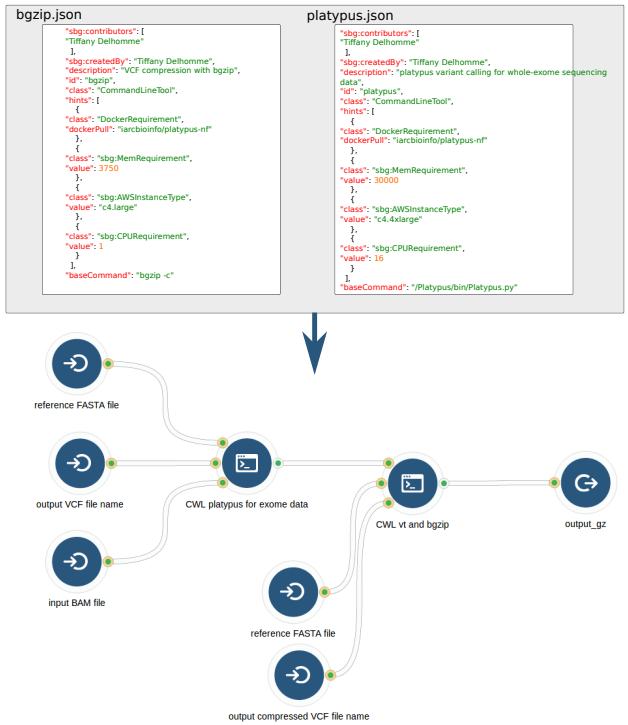
Public and private tools

- 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



PRIVATE TOOLS



International Agency for Research on Cancer

World Health Organization

Analyses

Description of an analysis workflow

1. CWL tool description

bgzip.json


```
"sbg:contributors": [
  "Tiffany Delhomme"
],
"sbg:createdBy": "Tiffany Delhomme",
"description": "VCF compression with
bgzip",
"id": "bgzip",
"class": "CommandLineTool",
"hints": [
  {
    "class": "DockerRequirement",
    "dockerPull": "iarcbioinfo/platypus-nf"
  },
  {
    "class": "sbg:MemRequirement",
    "value": 3750
  },
  {
    "class": "sbg:AWSInstanceType",
    "value": "c4.large"
  },
  {
    "class": "sbg:CPURequirement",
    "value": 1
  }
],
"baseCommand": "bgzip -c"
```


platypus.json


```
"sbg:contributors": [
  "Tiffany Delhomme"
],
"sbg:createdBy": "Tiffany Delhomme",
"description": "platypus variant calling for whole-exome
sequencing data",
"id": "platypus",
"class": "CommandLineTool",
"hints": [
  {
    "class": "DockerRequirement",
    "dockerPull": "iarcbioinfo/platypus-nf"
  },
  {
    "class": "sbg:MemRequirement",
    "value": 30000
  },
  {
    "class": "sbg:AWSInstanceType",
    "value": "c4.4xlarge"
  },
  {
    "class": "sbg:CPURequirement",
    "value": 16
  }
],
"baseCommand": "/Platypus/bin/Platypus.py"
```


International Agency for Research on Cancer

World Health Organization

Analyses

Description of an analysis workflow

1. CWL tool description

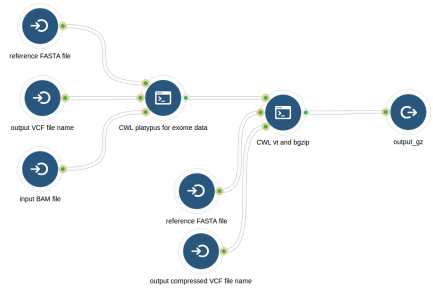
bgzip.json


```
"sbg:contributors": [
  "Tiffany Delhomme"
],
"sbg:createdBy": "Tiffany Delhomme",
"description": "VCF compression with
bgzip",
"id": "bgzip",
"class": "CommandLineTool",
"hints": [
  {
    "class": "DockerRequirement",
    "dockerPull": "iarcbioinfo/platypus-nf"
  },
  {
    "class": "sbg:MemRequirement",
    "value": 3750
  },
  {
    "class": "sbg:AWSInstanceType",
    "value": "c4.large"
  },
  {
    "class": "sbg:CPURequirement",
    "value": 1
  }
],
"baseCommand": "bgzip -c"
```


platypus.json


```
"sbg:contributors": [
  "Tiffany Delhomme"
],
"sbg:createdBy": "Tiffany Delhomme",
"description": "platypus variant calling for whole-exome
sequencing data",
"id": "platypus",
"class": "CommandLineTool",
"hints": [
  {
    "class": "DockerRequirement",
    "dockerPull": "iarcbioinfo/platypus-nf"
  },
  {
    "class": "sbg:MemRequirement",
    "value": 30000
  },
  {
    "class": "sbg:AWSInstanceType",
    "value": "c4.4xlarge"
  },
  {
    "class": "sbg:CPURequirement",
    "value": 16
  }
],
"baseCommand": "/Platypus/bin/Platypus.py"
```


2. CGC workflow integration



International Agency for Research on Cancer

World Health Organization

Analyses

Description of an analysis workflow

1. CWL tool description

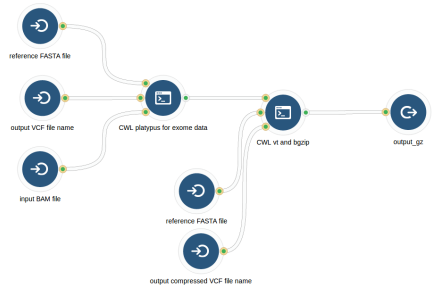
bgzip.json


```
"sbg:contributors": [
  "Tiffany Delhomme"
],
"sbg:createdBy": "Tiffany Delhomme",
"description": "VCF compression with
bgzip",
"id": "bgzip",
"class": "CommandLineTool",
"hints": [
  {
    "class": "DockerRequirement",
    "dockerPull": "iarcbioinfo/platypus-nf"
  },
  {
    "class": "sbg:MemRequirement",
    "value": 3750
  },
  {
    "class": "sbg:AWSInstanceType",
    "value": "c4.large"
  },
  {
    "class": "sbg:CPURequirement",
    "value": 1
  }
],
"baseCommand": "bgzip -c"
```

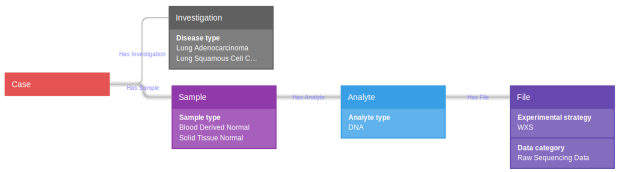

platypus.json


```
"sbg:contributors": [
  "Tiffany Delhomme"
],
"sbg:createdBy": "Tiffany Delhomme",
"description": "platypus variant calling for whole-exome
sequencing data",
"id": "platypus",
"class": "CommandLineTool",
"hints": [
  {
    "class": "DockerRequirement",
    "dockerPull": "iarcbioinfo/platypus-nf"
  },
  {
    "class": "sbg:MemRequirement",
    "value": 30000
  },
  {
    "class": "sbg:AWSInstanceType",
    "value": "c4.4xlarge"
  },
  {
    "class": "sbg:CPURequirement",
    "value": 16
  }
],
"baseCommand": "/Platypus/bin/Platypus.py"
```


2. CGC workflow integration



3. Querying data



International Agency for Research on Cancer

World Health Organization

Analyses

Description of an analysis workflow

1. CWL tool description

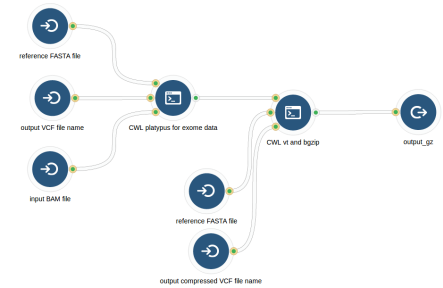
bgzip.json


```
"sbg:contributors": [
  "Tiffany Delhomme"
],
"sbg:createdBy": "Tiffany Delhomme",
"description": "VCF compression with
bgzip",
"id": "bgzip",
"class": "CommandLineTool",
"hints": [
  {
    "class": "DockerRequirement",
    "dockerPull": "iarcbioinfo/platypus-nf"
  },
  {
    "class": "sbg:MemRequirement",
    "value": 3750
  },
  {
    "class": "sbg:AWSInstanceType",
    "value": "c4.large"
  },
  {
    "class": "sbg:CPURequirement",
    "value": 1
  }
],
"baseCommand": "bgzip -c"
```

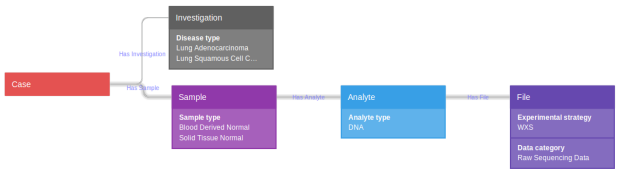

platypus.json


```
"sbg:contributors": [
  "Tiffany Delhomme"
],
"sbg:createdBy": "Tiffany Delhomme",
"description": "platypus variant calling for whole-exome
sequencing data",
"id": "platypus",
"class": "CommandLineTool",
"hints": [
  {
    "class": "DockerRequirement",
    "dockerPull": "iarcbioinfo/platypus-nf"
  },
  {
    "class": "sbg:MemRequirement",
    "value": 30000
  },
  {
    "class": "sbg:AWSInstanceType",
    "value": "c4.4xlarge"
  },
  {
    "class": "sbg:CPURequirement",
    "value": 16
  }
],
"baseCommand": "/Platypus/bin/Platypus.py"
```

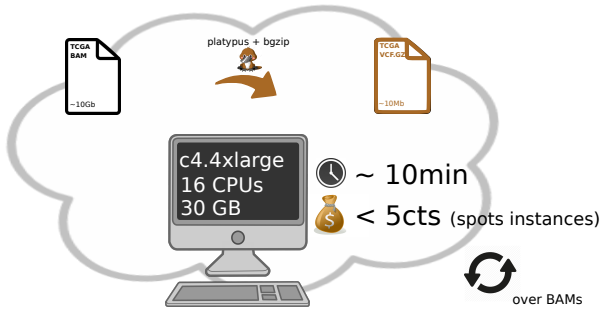

2. CGC workflow integration



3. Querying data



4. Running analyses

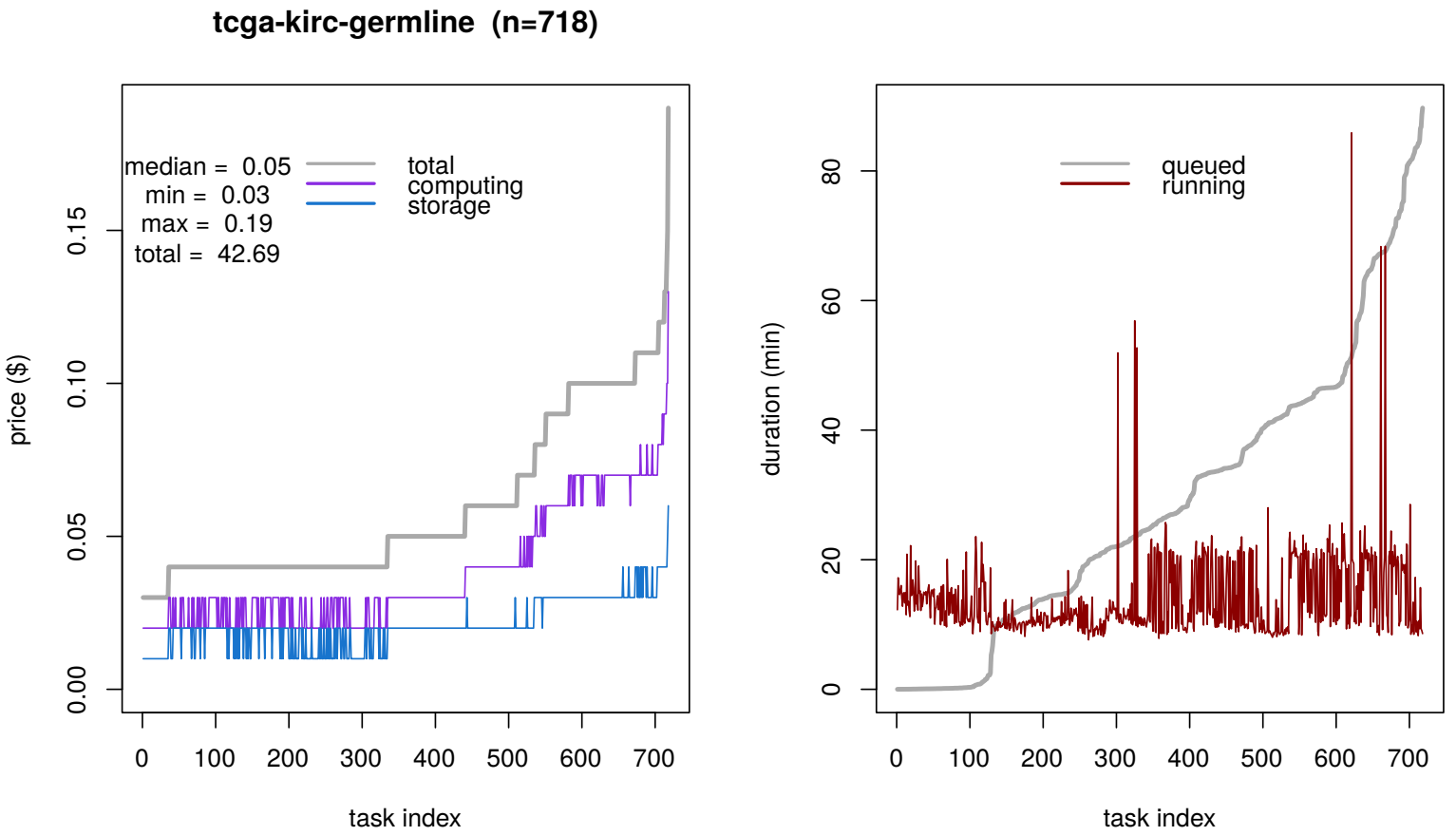


International Agency for Research on Cancer

World Health Organization

Analyses

example of task monitoring



International Agency for Research on Cancer

World Health Organization


Thank You