Analyzing TCGA data in the cloud

Tiffany Delhomme
Matthieu Foll

International Agency for Research on Cancer
Lyon, France

Learning objectives

- After completing this workshop, participants will be able to run their own computational tools on the cloud using TCGA data using:
  - the SevenBridges web interface to select and retrieve TCGA data,
  - Docker and DockerHub to build and store containers to deploy their own computational tools,
  - the Common Workflow Language (CWL) to describe the pipelines to run,
  - the SevenBridges R api to run automatically reproducible analyses

Agenda

Wednesday 28 February

09:00-10:00	Introduction to cloud computing and the SevenBridges architecture
10:00-10:30	Introduction to TCGA data
10:30-11:00	Break
11:00-11:30	Introduction to the SevenBridges web interface to run analyses
11:30-12:30	Practical application: run your first basic analysis in the cloud

Thursday 1 March

09:00-09:30	Introduction to Docker and DockerHub
09:30-11:00	Practical application: building your own Docker container and run it in the cloud
11:00-11:30	Break
11:30-12:30	Introduction to the R api and the CWL language

2 March **Friday 2 March**

09:00-12:30	Practical application: running your own practical project in the cloud using the R api, CWL and Docker.
12:30-14:00	Lunch Break
14:00-17:00	Practical application: running your own practical project in the cloud using the R api, CWL and Docker.

Github page

IARCbioinfo / SBG-CGC_course2018

Unwatch 8 | Unstar 2 | Fork 2

Code | Issues 3 | Pull requests 1 | Projects 0 | Wiki | Insights | Settings

IARC course on analyzing TCGA data in the SevenBridges Genomics CancerGenomicsCloud (SBG-CGC) Edit

Add topics

15 commits | 1 branch | 0 releases | 1 contributor | GPL-3.0

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

tdelhomme add guidelines after testing on IARC laptop — Latest commit 5eeb907 16 hours ago

demo_code	add required for output_file_name in json	8 days ago
project1-needlestack	add the 3 project folders	2 days ago
project2-neutrality	add the 3 project folders	2 days ago
project3-RNAseq-cellpop	add the 3 project folders	2 days ago
LICENSE	Initial commit	8 days ago
README.md	add guidelines after testing on IARC laptop	16 hours ago

README.md

SBG-CGC_course2018

IARC course on analyzing TCGA data in the SevenBridges Genomics CancerGenomicsCloud (SBG-CGC).

Description of the course

Learning objectives
After completing this workshop, participants will be able to run their own computational tools on the cloud using TCGA data using:

International Ag

World He
Organizat

3 Open ✓ 0 Closed Author Labels Projects Milestones Assignee Sort

- **Project 3: RNA-seq cell population** collaborative project protected TCGA data
  #3 opened 2 days ago by tdelhomme
- **Project 2: Neutral tumor evolution testing** collaborative project public TCGA data 5
  #2 opened 2 days ago by tdelhomme
- **Project 1: needlestack variant calling** collaborative project protected TCGA data 2
  #1 opened 2 days ago by tdelhomme

Project 2: Neutral tumor evolution testing #2

Open tdelhomme opened this issue 2 days ago · 5 comments

tdelhomme commented 2 days ago • edited Member

Test neutral tumor evolution model described in Williams et al. on TCGA data.

Data: public MAF files

Script: R script, which output regression coefficient and slope of the model, by sample.

Guidelines: Loop on files

Project source code and documentation is hosted here.

tdelhomme added the collaborative project label 2 days ago

tdelhomme assigned NoemieL 2 days ago

tdelhomme commented 2 days ago • edited Member

@NoemieL, you should have an R script computing this on a VCF file, you just should adapt this to MAF file. Be careful, you have one MAF per cohort.
@aurelieGabriel would give a presentation on TCGA data and so also on MAF format.

mfoll commented 23 hours ago Owner

Note that Williams et al. selected only samples with a tumor purity >70%. Estimates of tumor purity for most TCGA cohorts are available in COSMIC under "ASCAT Ploidy and Purity Estimates".

mfoll commented 23 hours ago Owner

One difficulty is that you have to filter on allelic fraction (>0.12 & <0.25 for example), and MAF files don't always contain this information. VCF do, but are protected and are not filtered with the same QC. Maybe check in which cohort the allelic fraction is available in the MAF file with @aurelieGabriel?

aurelieGabriel commented 21 hours ago • edited Member

In every MAF file the following columns are now reported: t_depth (Read depth across the locus in tumor BAM) and t_alt_count (Read depth supporting the variant allele in tumor BAM).

1

mfoll commented 21 hours ago Owner

Good news! (It was not the case before).

International Agency for Research on Cancer

World Health Organization

What this course is NOT!

- A Unix/R/HPC/bioinformatics course
  - These are pre-requisites. There are other courses for this, including MOOCs or quick tutorials (Unix).
- A comprehensive course on:
  - TCGA data
  - Cloud computing or even the SevenBridges cloud
  - Docker
  - CWL
  - …

What this course is NOT!

- It will NOT change you overnight into an expert in those topics.
- It will NOT answer all your questions
- Not for the lazy, this is work

What this course is

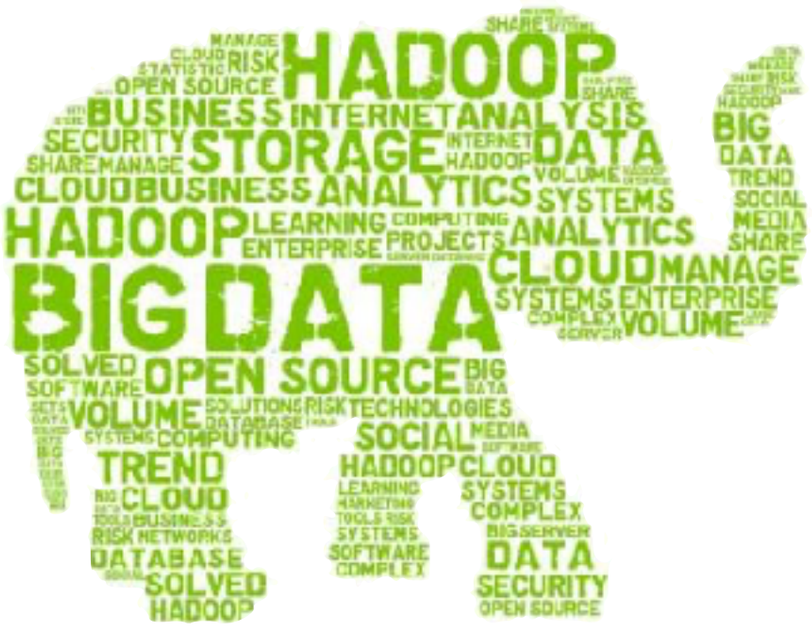
- A generous knowledge sharing exercise
  - To avoid you unnecessary reading of 100s of pages about cloud computing, Docker, CWL etc.
  - Assembling the puzzle of skills needed in each field
  - Giving you the minimal set of skills to get started
- A nice day working on interesting projects together!

What is cloud computing?

- A lot of things we don't care about in bioinformatics
- A lot of buzz words that only interest business IT-geeks
- What we want:
  - Machines to run analyses (HPC)
  - Data to analyses
  - Bioinformatics software

Why using cloud computing?

- The hardware
  - If you don't have access to a HPC cluster, you can create one in the cloud
  - You only pay when you need it and it is "elastic"
- The data
  - Analyse the data where it is (e.g. TCGA data)
- The services
  - Easy to run software



International Agency for Research on Cancer

World Health Organization

TCGA data and cloud computing

- Constrains due to the presence of protected data (managing access rights)
- NIH Genomic Data Commons (GDC):
  - collaborates with NCI Cloud Resources to democratize access to NCI-generated genomic
  - three groups developing cloud-based platforms
    - Broad Institute
    - Institute for Systems Biology (ISB)
    - Seven Bridges Genomics

Reference

Focus on Computer Resources

Cancer Research

The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research

Jessica W. Lau, Erik Lehnert, Anurag Sethi, Raunaq Malhotra, Gaurav Kaushik, Zeynep Onder, Nick Groves-Kirkby, Aleksandar Mihajlovic, Jack DiGiovanna, Mladen Srdic, Dragan Bajcic, Jelena Radenkovic, Vladimir Mladenovic, Damir Krstanovic, Vladan Arsenijevic, Djordje Klisic, Milan Mitrovic, Igor Bogicevic, Deniz Kural, and Brandi Davis-Dusenbery; for The Seven Bridges CGC Team

International Agency for Research on Cancer

World Health Organization