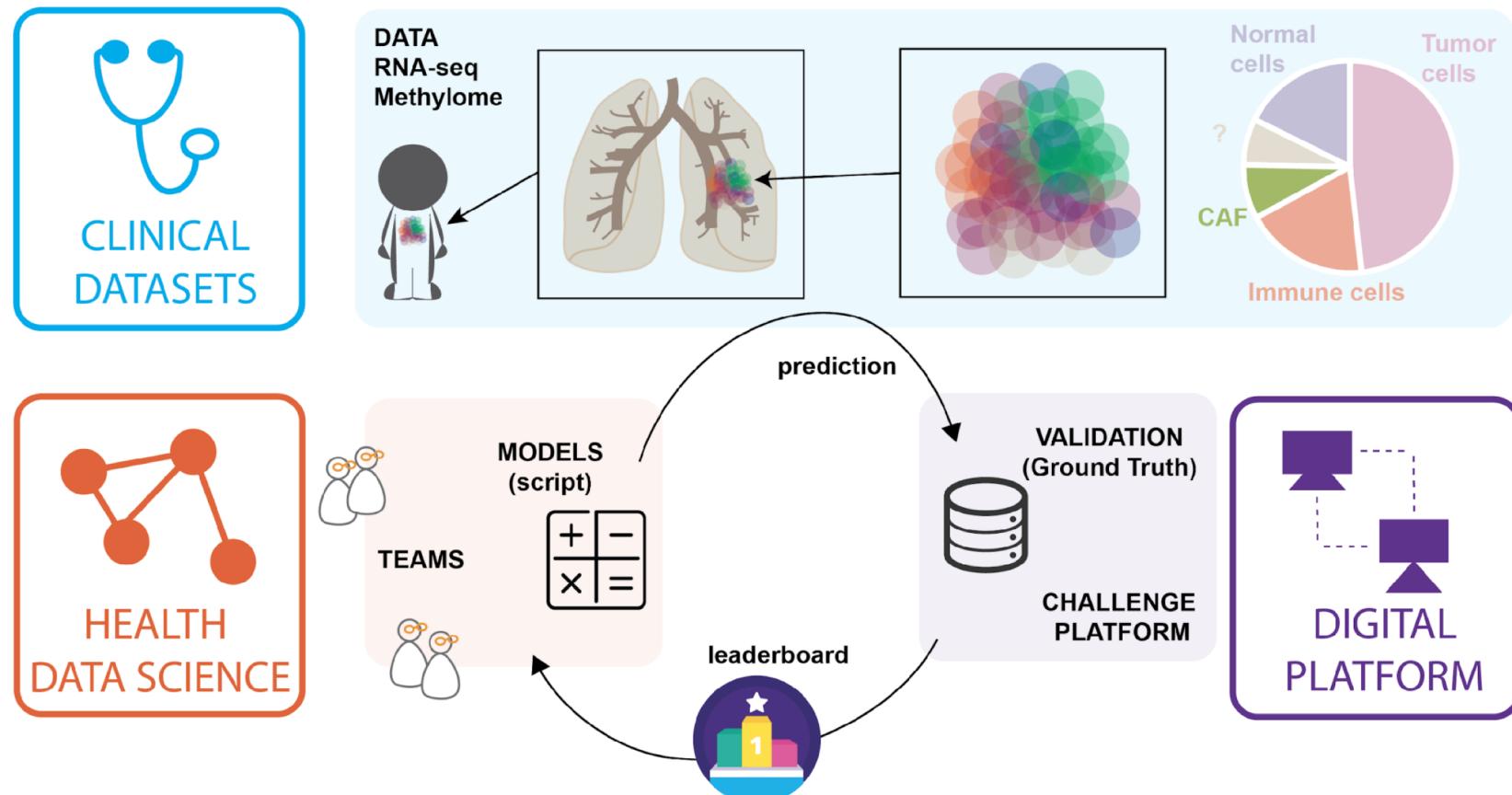


HADACA – Health Data Challenge

Deconvolution methods to quantify tumor heterogeneity

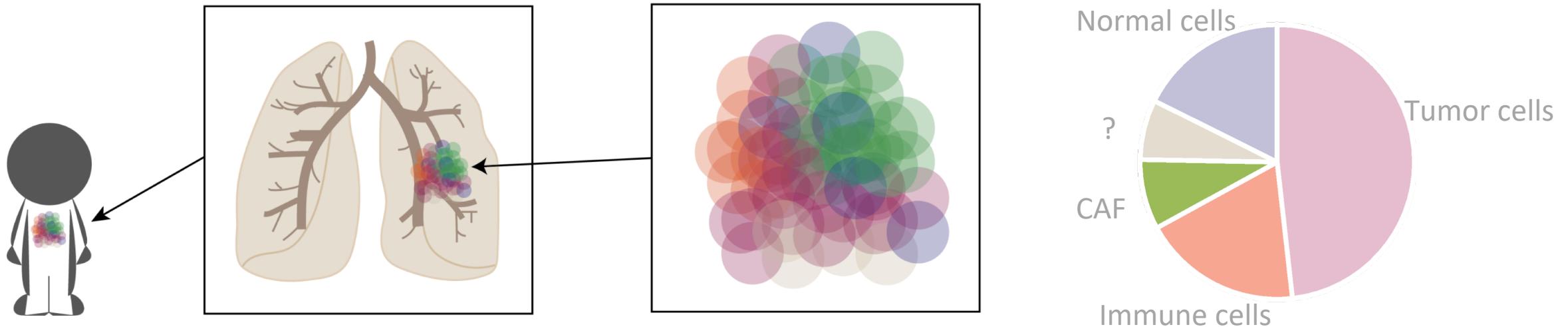


Introduction to the data challenge



Goal of our data challenge

Quantification of tumor heterogeneity



2nd edition of the data challenge

1st edition (2018)



- Methylation Data
- One cancer type
- Cell lines

2nd edition (2019)



- Methylation and transcriptomic Data
- Several cancer types
- Primary tumors / cell lines

Organizing team



Magali Richard
(TIMC-IMAG team)



Clémentine Decamps
(TIMC-IMAG team)



Alexis Arnaud
(TIMC-IMAG team)



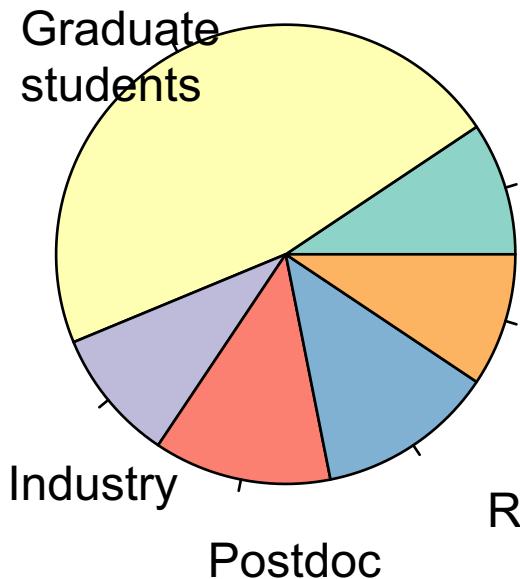
Yuna Blum
(CIT team)



Florent Petitprez
(CIT team)



About the participants (n=31)



10 teams of 3-4 people

Engineers

Bioinformatics

Undergrad
students

Researchers

Computer
science

Data
science

Medical science

Statistics

Institut Curie
UGA Grenoble
CRC Cordeliers
INSERM
CEA
Innate Pharma
Verteego
...

France



Byelorussia



Sweden



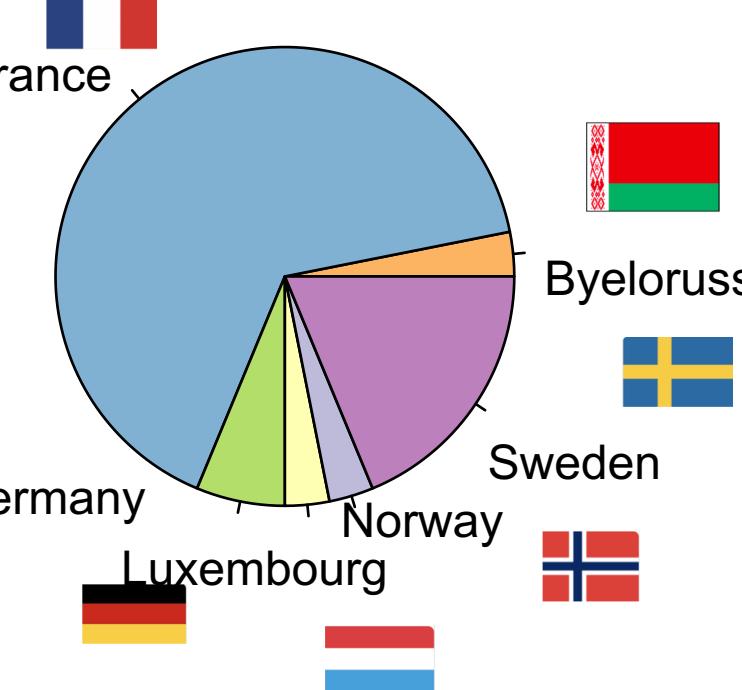
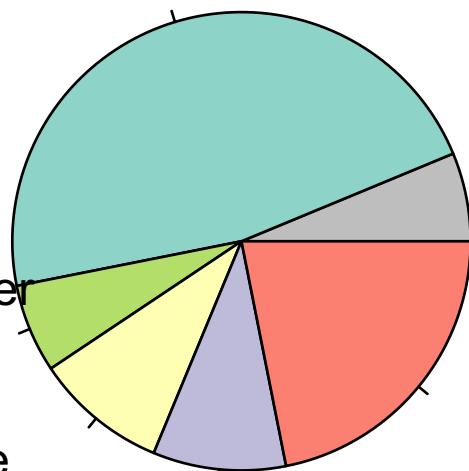
Norway



Luxembourg



Germany



Invited speakers



**Jérôme
Cros**
AP-HP, Paris
France

- Tumor heterogeneity:
the clinician's point of view



Michael Scherer
Max-Planck-Institut für
Informatik,
Saarbrücken,
Germany

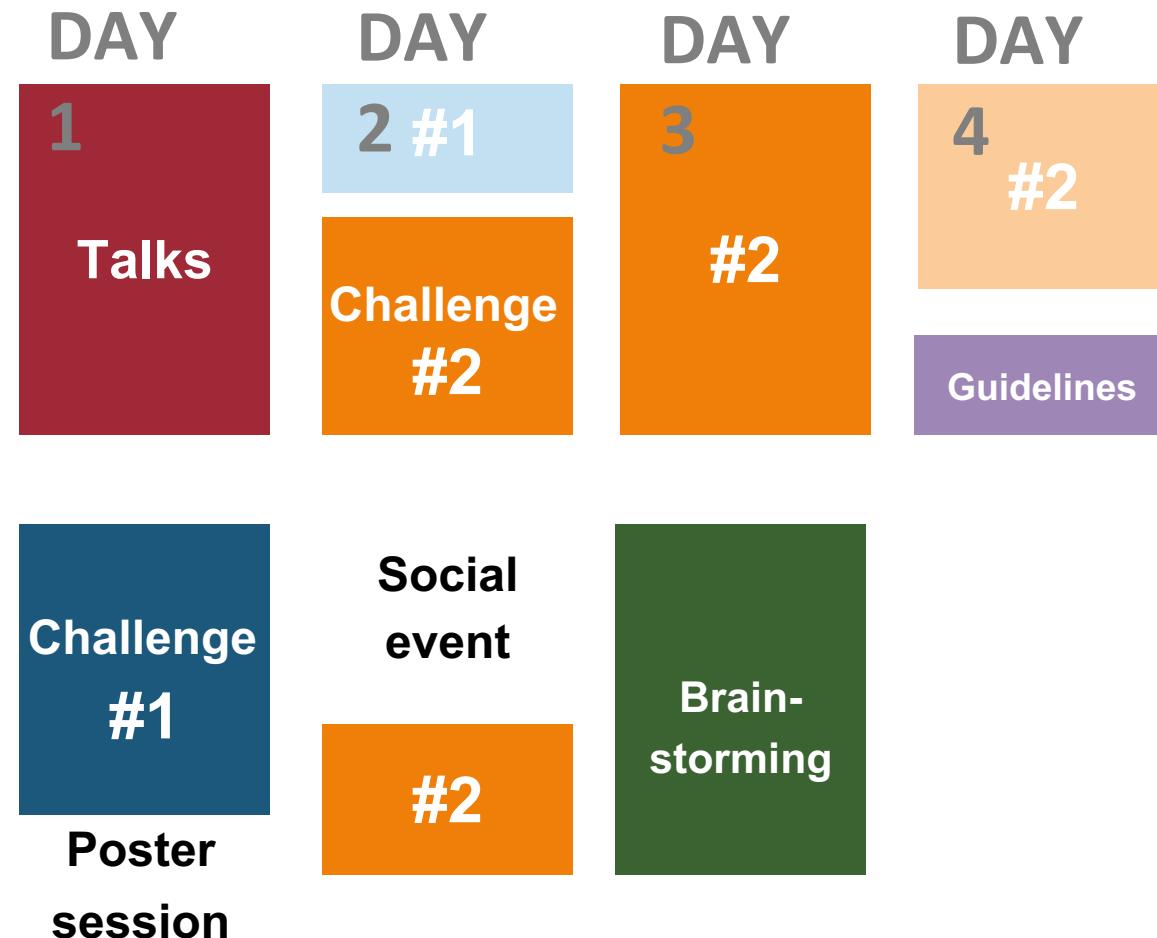
- Methylation and
deconvolution methods



Francisco Avila Cobos
Ghent University, Gand
Belgium

- Transcriptome and
deconvolution
methods

Agenda



“The basics” during poster session

“Deconvolution methods”

DECONVOLUTION METHODS

INTRODUCTION: WHAT IS A DECONVOLUTION/ UNMIXING/ SOURCE SEPARATION PROBLEM?

The figure shows a scenario where multiple people are speaking into microphones, and the goal is to estimate the proportion of each speaker's voice in the recorded audio. This is a classic deconvolution problem.

NATURE OF OMICS DATA (DNA/RNA) USED FOR ESTIMATION OF CELL TYPE PROPORTIONS

This section discusses the nature of omics data used for estimating cell type proportions. It includes a diagram showing the estimation of the relative amount of RNA of all genes expressed in a tissue of a given individual observed at a time. It also covers the estimation of the relative amount of DNA methylation (Dnase, chromatin & TAD) at all resolved CpG sites using the genome-wide approach.

MATHEMATICAL FOUNDATIONS OF DECONVOLUTION

This section provides a detailed explanation of the mathematical foundations of deconvolution. It states that deconvolution is an algorithmic process used to reverse the effects of convolution or recorded data. The concept of deconvolution is to find the solution of a convolution equation of the form: $y = h * x + n$. In this equation, y is the observed data, x is the true signal, h is the convolution kernel, and n is the noise. The objective of deconvolution is to find the solution of the convolution equation of the form: $y = h * x + n$.

DATA PRE-PROCESSING

This section covers various steps of data pre-processing, including filtering, normalizing, and removing batch effects.

ESTIMATING THE NUMBER OF CELL TYPES

This section discusses the estimation of the number of cell types from single-cell RNA-seq data. It highlights that the number of cell types can be estimated by performing PCA on the raw data.

ACKNOWLEDGEMENTS

Universitat de Girona, I3-Health, Universitat de Barcelona, Institut d'Estudis Clínics de Barcelona, Institut d'Investigacions Biomèdiques de Bellvitge (IDIBELL), Institut de Recerca i Tecnologia Industrial (IRI).

“Transcriptomic data”

TRANSCRIPTOMIC DATA

INTRODUCTION: A REMINDER OF CELL BIOLOGY

This section provides a reminder of cell biology, showing the human body as a collection of cells, each containing a nucleus with chromosomes. It also shows the flow of information from DNA to RNA to protein.

TRANSCRIPTOME DEFINITION

The transcriptome is defined as the estimation of the relative amount of RNA of all genes expressed in a tissue of a given individual observed at a time. It includes a diagram of a gene being transcribed into mRNA.

HOW TO ACCESS THE TRANSCRIPTOME

This section describes two types of technologies: Microarray and RNA-seq. Microarray involves hybridization of cRNA to a grid of probes, while RNA-seq involves sequencing of cDNA libraries. Both technologies measure gene expression levels.

THE TECHNOLOGY

This section provides a detailed overview of the transcriptomic technology, including the use of microarrays and RNA-seq, and the analysis of gene expression data.

DATA PROCESSING AND ANALYSIS

This section covers the processing and analysis of transcriptomic data, including differential analysis, clustering, and PCA.

FAQ AND Q&A

Frequently asked questions and answers related to transcriptomic data analysis.

ACKNOWLEDGEMENTS

Universitat de Girona, I3-Health, Universitat de Barcelona, Institut d'Estudis Clínics de Barcelona, Institut d'Investigacions Biomèdiques de Bellvitge (IDIBELL), Institut de Recerca i Tecnologia Industrial (IRI).

“Methylome data”

DNA METHYLATION DATA

Introduction: DNA structure and epigenetics

This section introduces DNA structure and epigenetics. It shows the DNA double helix with methyl groups and discusses how DNA sequence (P-C-G) is methylated by enzymes to form the chromatin.

Epigenetics

Epigenetics is the study of all the modifications to the DNA that do not involve a change in the DNA sequence. These modifications can influence gene expression, some examples are histone modifications or the DNA methylation (methyl).

Details of DNA methylation

This section provides a detailed explanation of DNA methylation, including its definition, how it is measured, and its reversibility.

Technology using Array

This section describes the Illumina methylation assay, which uses biotinylated cytosines to detect methylated DNA.

Analyse of methylation data

This section covers the analysis of methylation data, including the calculation of average methylation and the generation of heatmaps.

ACKNOWLEDGEMENTS

Universitat de Girona, I3-Health, Universitat de Barcelona, Institut d'Estudis Clínics de Barcelona, Institut d'Investigacions Biomèdiques de Bellvitge (IDIBELL), Institut de Recerca i Tecnologia Industrial (IRI).

“Codalab platform & docker”

Codalab platform & Docker containers

Challenge on Codalab platform

Codalab is an open-source web-based platform that hosts challenges for students, conferences, and researchers with the aim to enforce collaboration and reproducible research.

Participant submissions

Participants can submit their source code or results. The source code is executed on Codalab servers using Docker containers in order to produce a result. The submitted or computed results are compared to the solutions in order to give a score.

Competitions

Codalab competitions: <https://competitions.codalab.org>

Codalab documentation

<https://github.com/codalab/codalab-competitions/wiki>

Containerized Applications with Docker

This section provides instructions for running a container, importing and saving containers, and using Docker volumes.

Import and save container

- Download a docker image from Docker Hub
- Start local instance based on the certain image
- Interactive mode (with docker, dockerfile, dockerfile.Dockerfile)
- Save local image (with docker, dockerfile, dockerfile.Dockerfile)

Docker run

- Print the list of active containers
- Or else clean stopped containers

Docker build

- Builds a Docker image
- Creates a Dockerfile
- Builds a Docker image from the Dockerfile

Docker volume

- Create a Docker volume between host and container
- Mount a directory from the host to the container

Online repository of Docker images

<https://hub.docker.com/>

DISCUSSION

Discuss <https://codalab.org/discuss>



The challenge

Your goal: estimate the matrix A

The matrix A represents the proportion of each cell type in each patient

You have this

↓

D	Patients
Sites	Expression/ Methylation
=	

T

Types

Sites

Expression/
Methylation

=

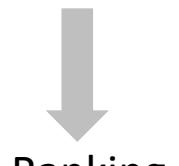
T	Types
Sites	Expression/ Methylation
x	

You search this

↓

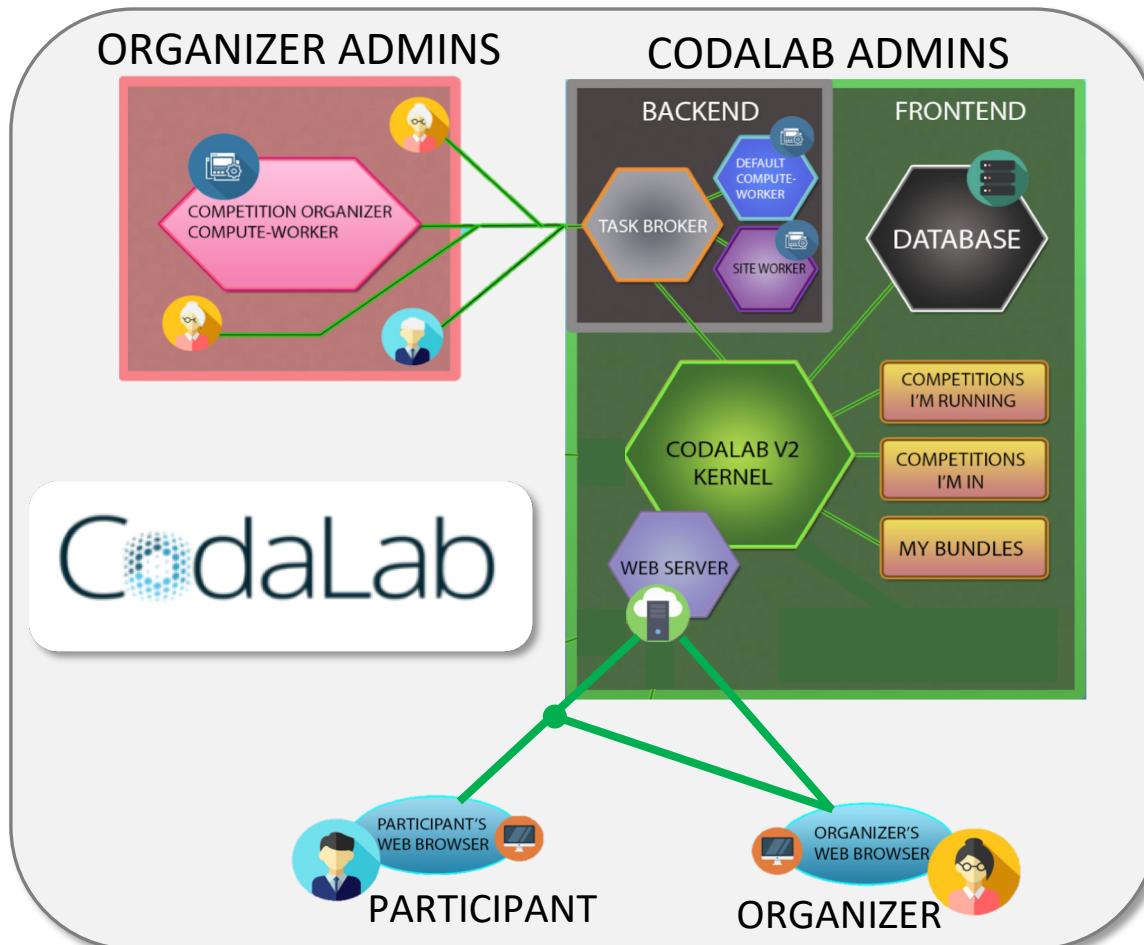
A	Patients
Types	Proportions

Mean absolute error
 $mean(abs(A - \hat{A}))$



Code that estimates A

The challenge platform



Open-source platform

- ? Enables participants to submit their codes
- ? Automatically rank the participants



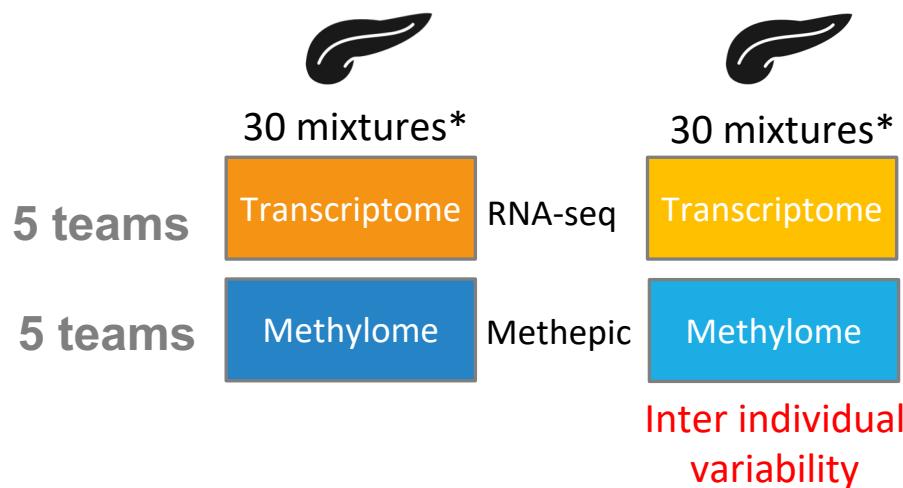
Alexis (IT) dedicated to the computing resource management

Challenges of increasing complexity



Challenge #1

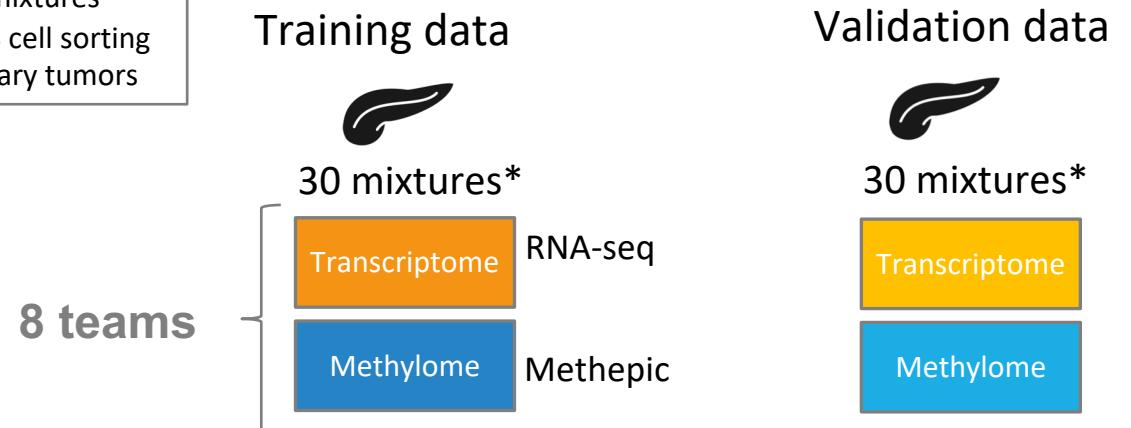
Aim: learn how to use collab, discover the dataset, manage to submit a deconvolution script on either RNA-seq or methylome



Challenge #2

More populations (some similar)

Aim: test and develop method to quantify tumor heterogeneity using both RNA-seq and methylome data



*In silico mixtures
from FACS cell sorting
from primary tumors

The awards



**Winner of the main
Challenge (#2)**



**Winner of the training
Challenge (#1)**



Best poster award

How to participate?

- (1) **Register** to the challenge on Codalab
- (2) **Find** your teammates
- (3) **Download** the starting kit and the public dataset

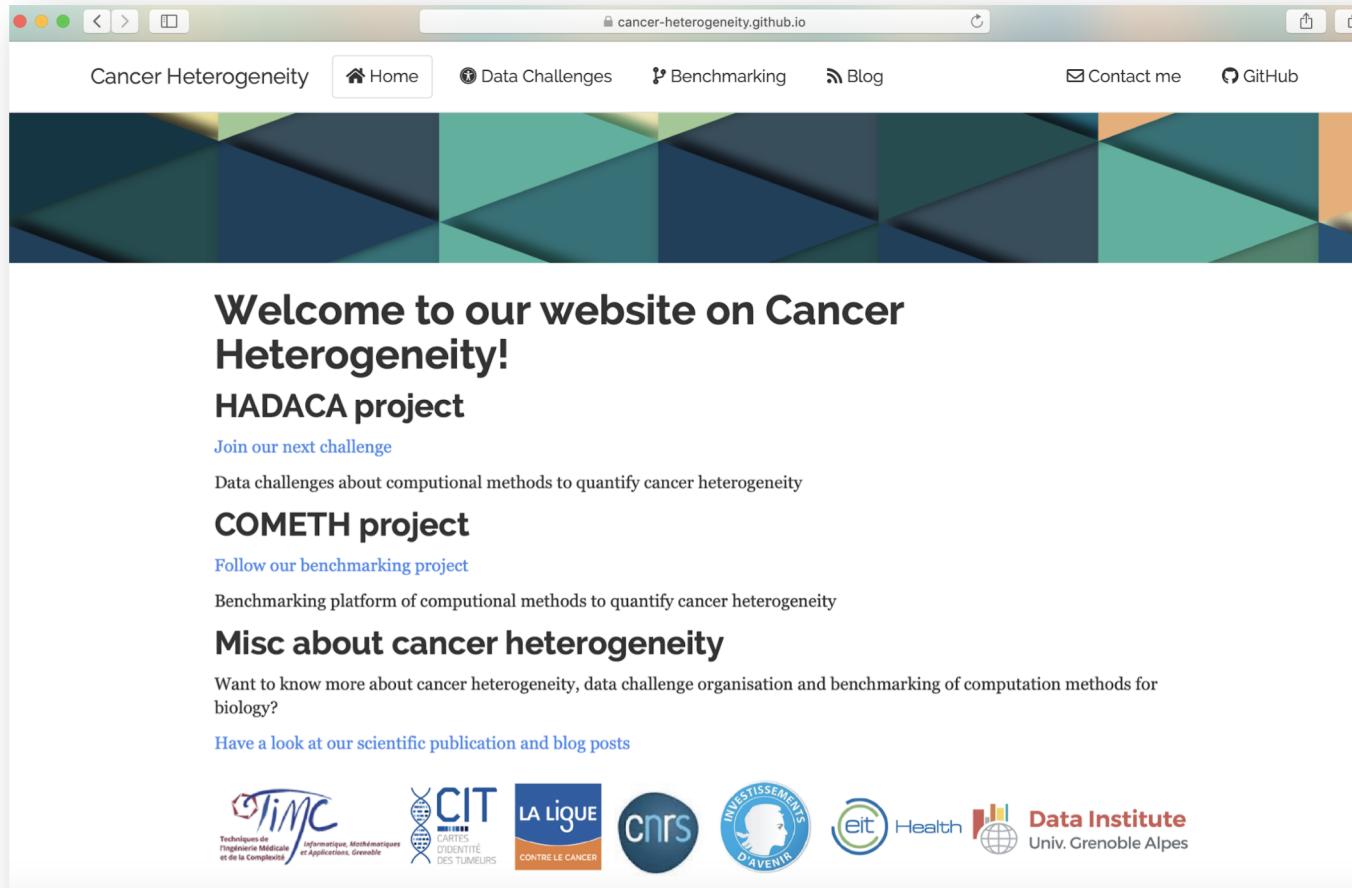
CHALLENGE BEGINS

- (1) **Work** in group to find deconvolution methods
- (2) **Submit** your results on the Codalab platform
- (3) **Improve** your score

CHALLENGE ENDS

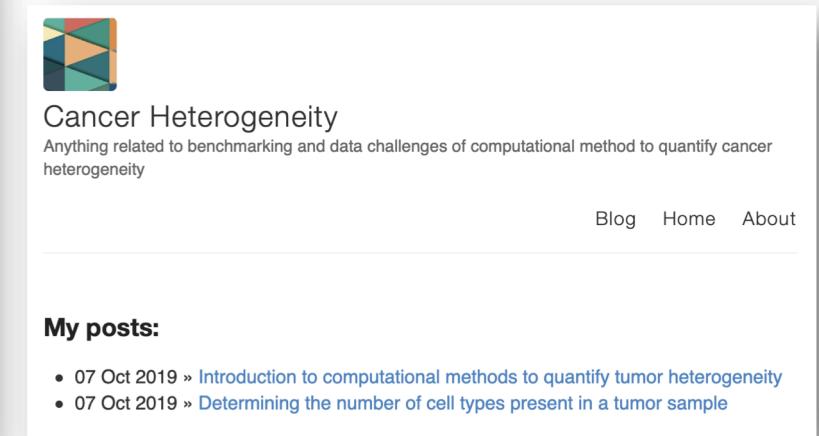
Restitution of your work (3 slides per team // PDF format)
at the end of each challenge (send it on time to Alexis!)

Website



The screenshot shows the homepage of the website cancer-heterogeneity.github.io. The header includes a navigation bar with links for Home, Data Challenges, Benchmarking, Blog, Contact me, and GitHub. Below the header is a large, abstract geometric background image composed of triangles in shades of teal, blue, and orange. The main content area features a large heading "Welcome to our website on Cancer Heterogeneity!" followed by "HADACA project". Below this, there are sections for "Join our next challenge" (with a link to "Data challenges about computational methods to quantify cancer heterogeneity"), "COMETH project" (with a link to "Benchmarking platform of computational methods to quantify cancer heterogeneity"), and "Misc about cancer heterogeneity" (with a link to "Want to know more about cancer heterogeneity, data challenge organisation and benchmarking of computation methods for biology?"). At the bottom, there is a section for "Have a look at our scientific publication and blog posts" and a row of logos for various partners: TiMC, CIT, LA LIGUE CONTRE LE CANCER, CNRS, eit Health, and Data Institute Univ. Grenoble Alpes.

- General information
- FAQ
- Blog posts
- List of methods
- ...



The screenshot shows a blog post page titled "Cancer Heterogeneity". The page includes a small geometric graphic icon, a brief description ("Anything related to benchmarking and data challenges of computational method to quantify cancer heterogeneity"), and navigation links for "Blog", "Home", and "About". Below this, there is a section titled "My posts:" with two recent entries: "07 Oct 2019 » Introduction to computational methods to quantify tumor heterogeneity" and "07 Oct 2019 » Determining the number of cell types present in a tumor sample".

Practical organization

- **Breakfast** : 7.30-9.15am (Level 4)
- **Dinner** starts at 7.45pm
- **Bar** open after lunch (Level 3) and after Dinner
 Coffee/tea offered (specify that you are from the Data Challenge)
- **Breaks** will take place in the bar
- **The poster session** will be in the mezzanine (Level 6)
 Poster numbers are in the program
- **For working sessions**, use the mezzanine or the 'La Scolette' room
- **Hike** will be on Wednesday afternoon (2 groups)

Objectives of the week

Share interdisciplinary knowledge

Learn good coding practices

Discover methylome and transcriptomes specificities

Assess state of the art of deconvolution methods

Have fun

See you tomorrow at 9am, salle ‘La Scolette’, (or tonight at the bar)

Thank you for your attention !

