

CINECA

Common Infrastructure for National Cohorts in Europe, Canada, and Africa

Federated analysis for polygenic risk score calculations

Presenter: Will Rayner and Anshika Chowdhary (Helmholtz Munich)

Host: Marta Lloret Llinares (EMBL-EBI)



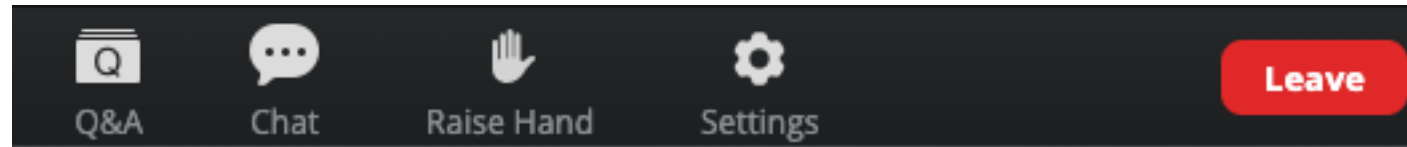
This project has received funding from the European Union's Horizon 2020 research and Innovation programme under grant agreement No. 825775



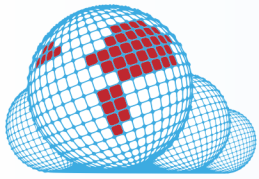
This webinar is being recorded
It will be available on the CINECA website and YouTube channel



Audience Q&A Session



Please write your questions
on the Zoom Q&A



CINECA

Common Infrastructure for National Cohorts in Europe, Canada and Africa

The vision:

Accelerating disease research and
improving health by facilitating
transcontinental human data exchange

Stay
informed

@CinecaProject



www.cineca-project.eu



The challenges:





Today's presenters

Will Rayner is the head of the Data and Analytics group at the Institute of Translational Genomics in the Computational Health Department at Helmholtz Munich. He is interested in all aspects of data management and data privacy and has been leading the CINECA polygenic risk score use case.

Anshika Chowdhary is a Data informatician, at the Institute of Translational Genomics at Helmholtz Munich. She has been working on the development of the workflows for the polygenic risk score use case, and analysis of the eQTL catalog on the datasets at HMGU.

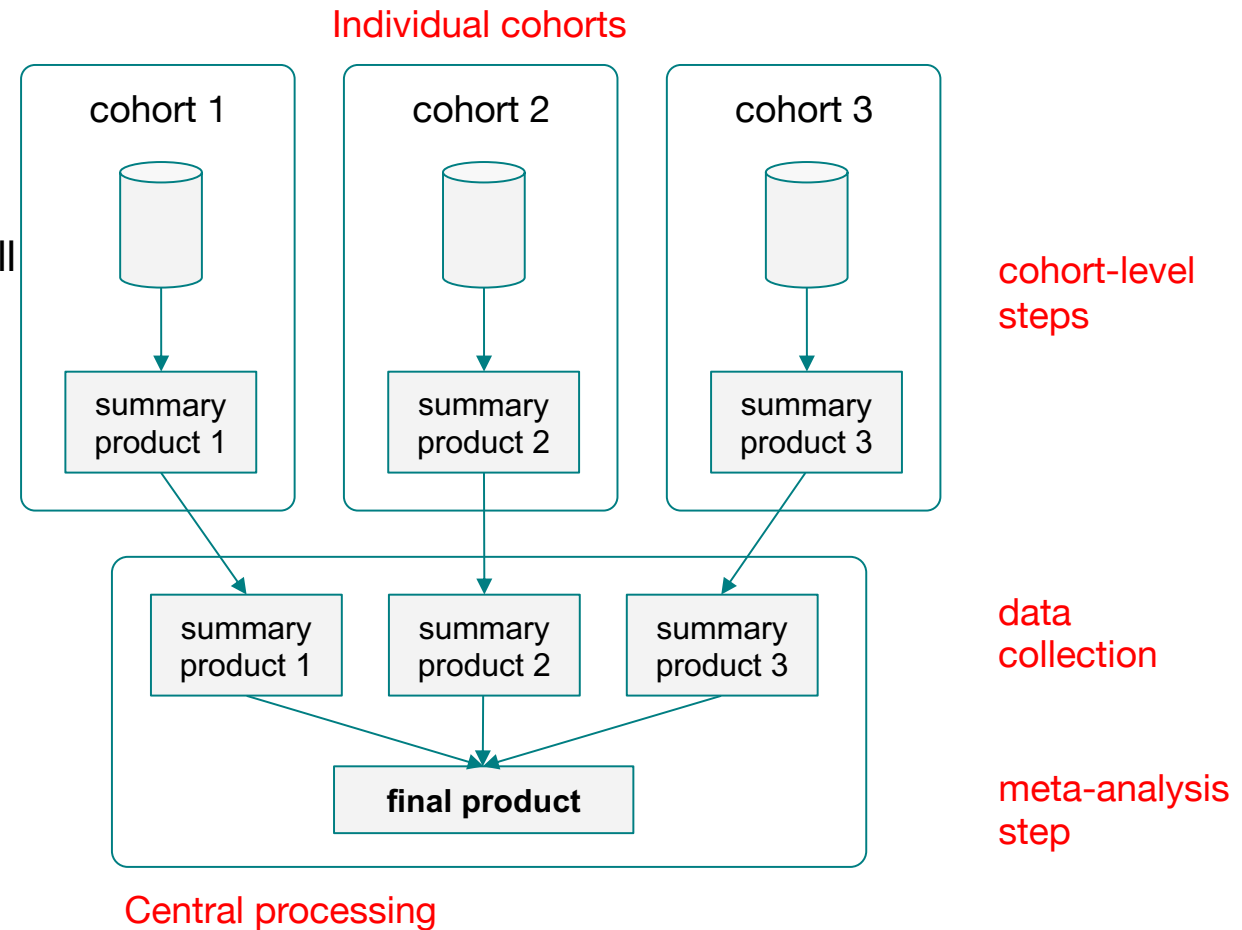
Federated Analyses for Polygenic Risk Score Calculations

CINECA Webinar 31 Jan 2023

W. Rayner
Anshika Chowdhary

What is Federated Genetic Data Analysis?

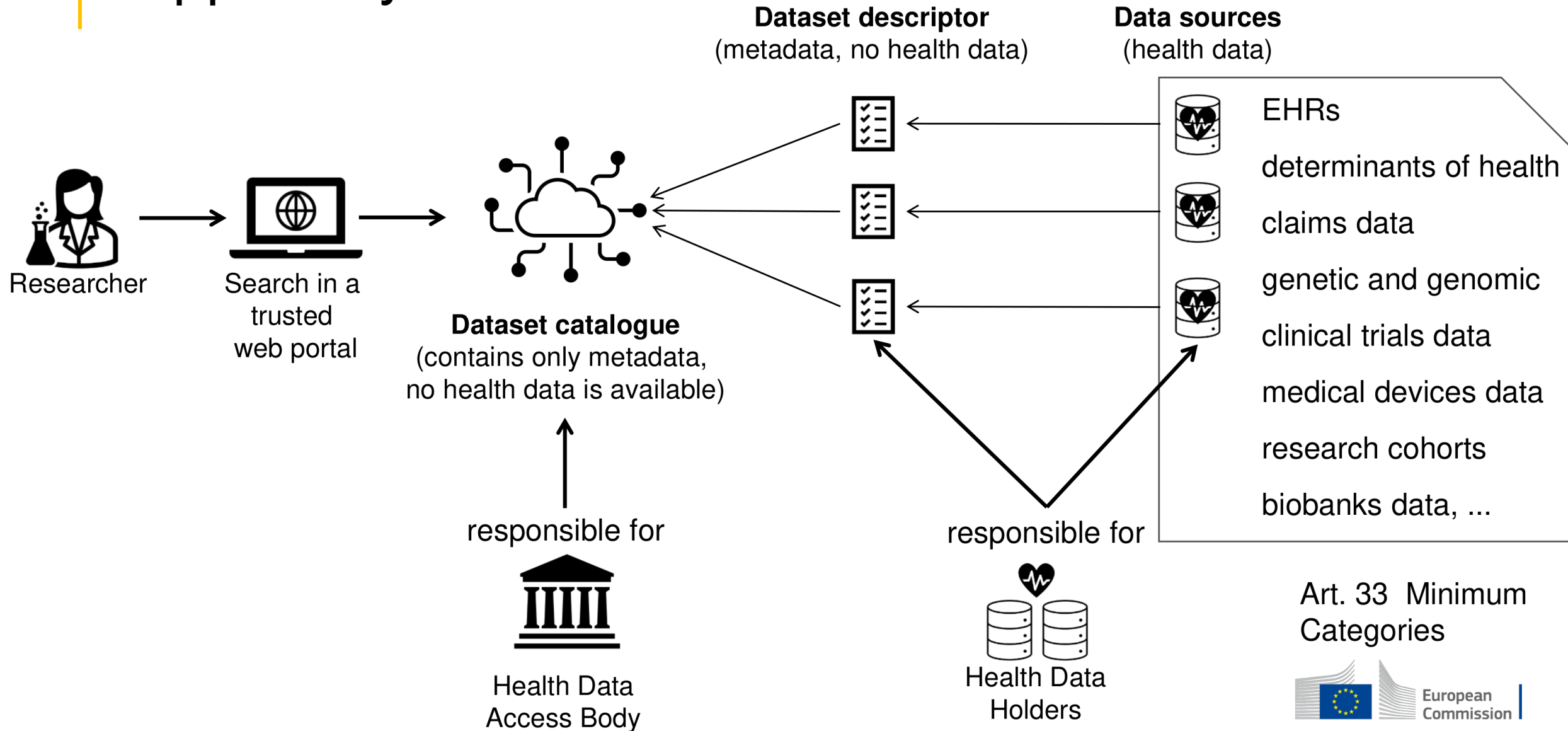
- Conducting analyses on multiple genomics data sets residing on different computer systems as if they were all residing locally
- Large consortia are doing meta-analyses across summary data from many cohorts e.g. GIANT, GLGC, DIAMANTE in what is effectively a manual federated analysis
 - Utilises the power of large data sets, ideally from different ancestries
 - Has yielded a large number of associations with common complex diseases
 - Slow, analyses can take years



European Health Data Space (EHDS)





- First of several data spaces planned for Europe
- Individuals will have control over who has access to their health data
- EU wide storage of Electronic Health Records (EHRs)
- Data will be stored in one, or more, data centres in each country
- Possibility to use the data for research (secondary use)
- Given the increase in health sequencing this will be key for future genomics research

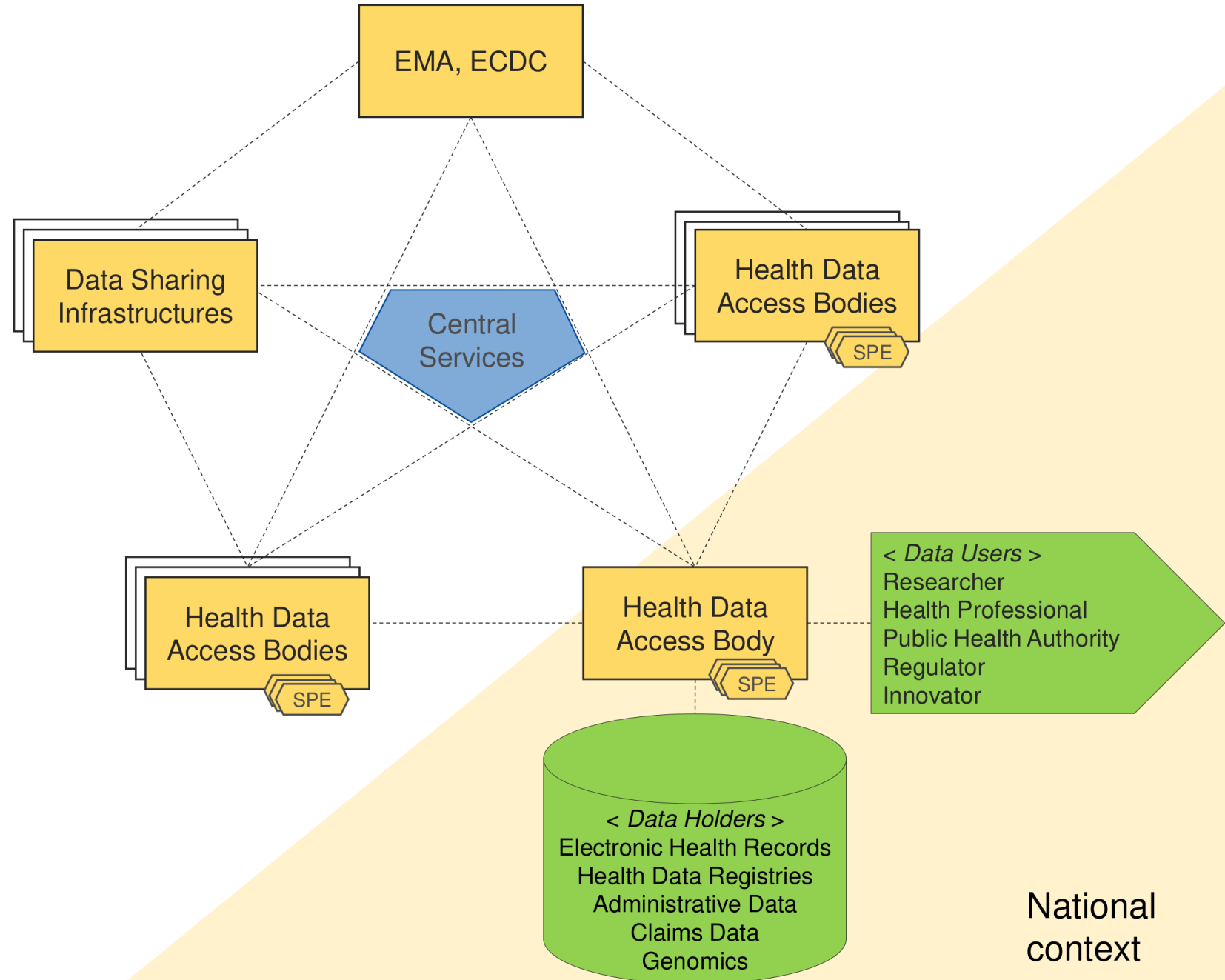
What health data exists to support my research?

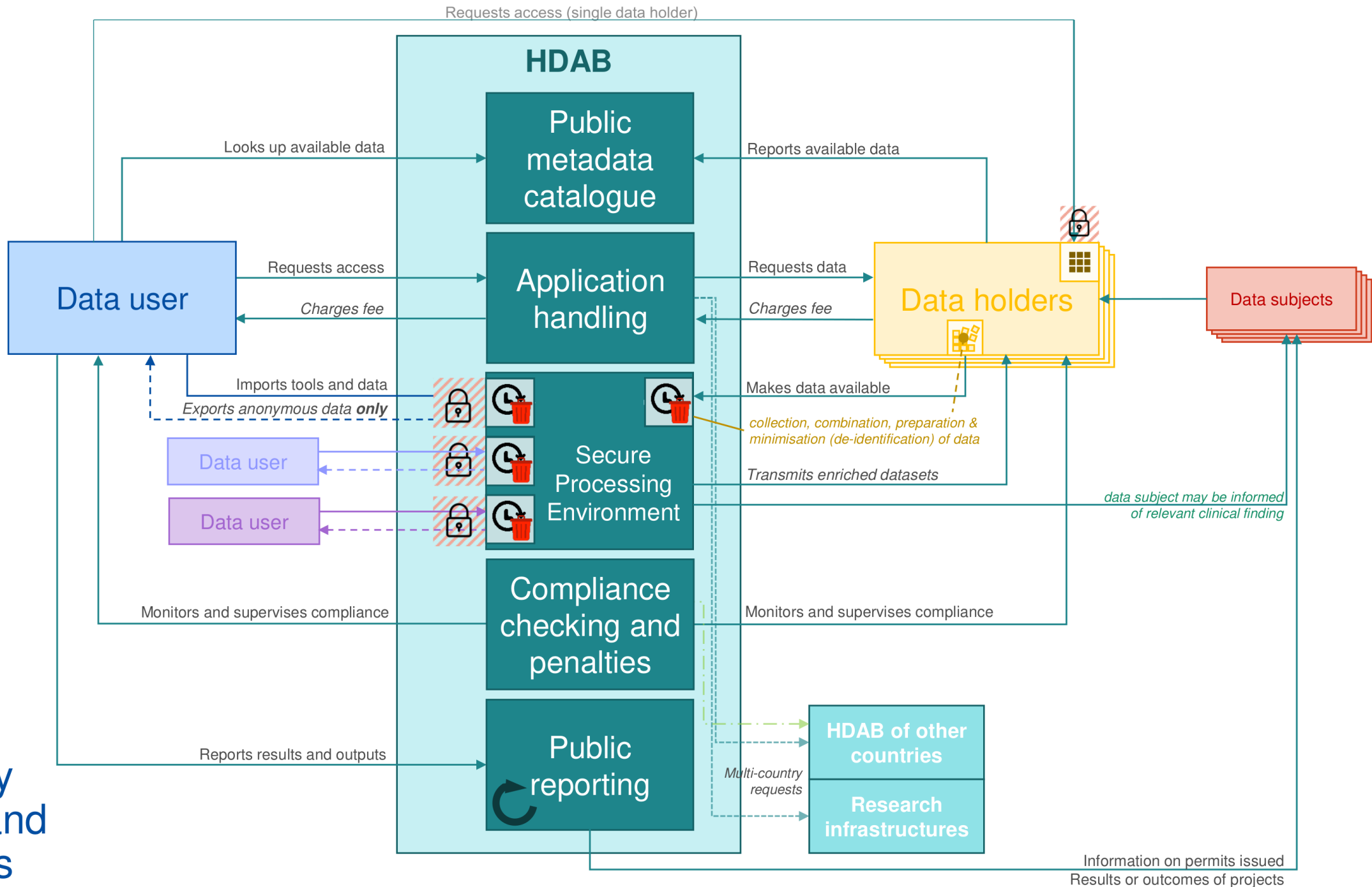


Cross-border secondary use infrastructure

HealthData@EU

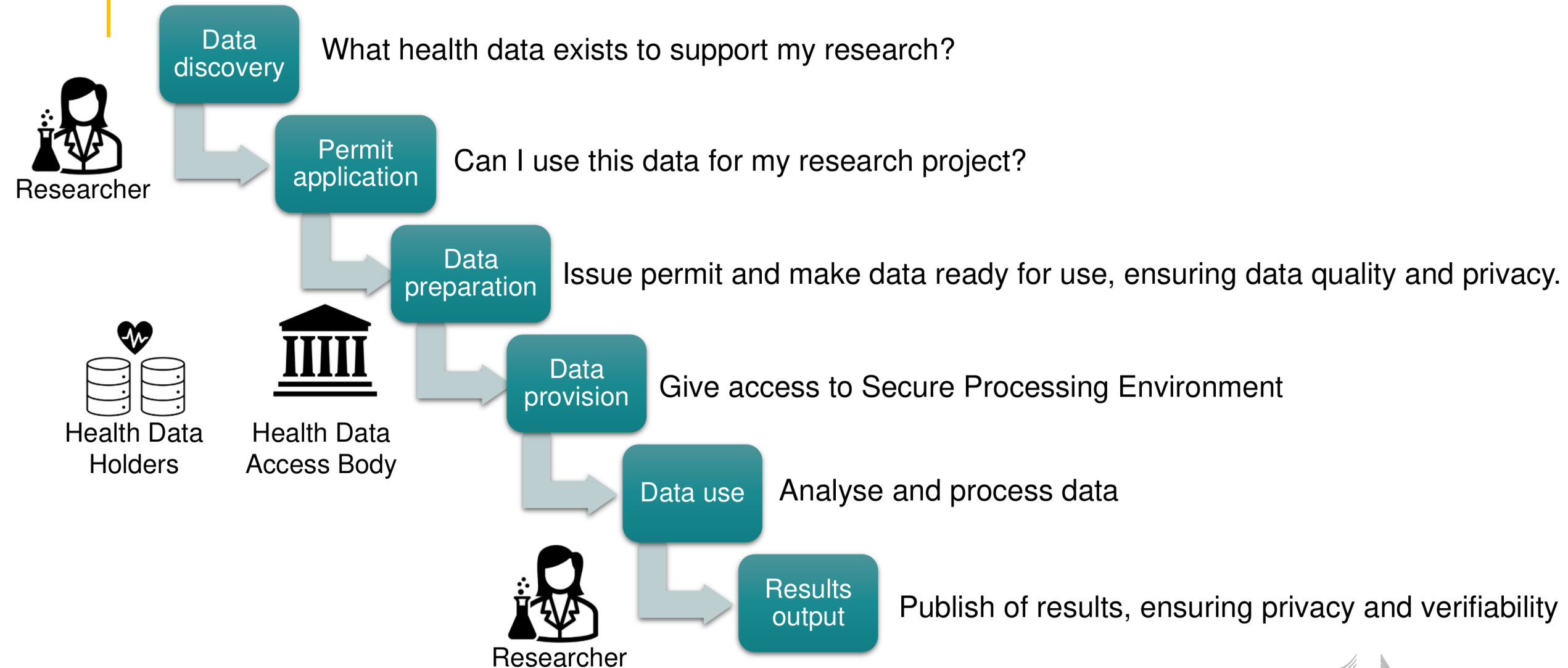
-  Central support services provided by EC
-  National data management services provided by authorised participants
-  Secure Processing Environments
-  Local services provided by/to local partners



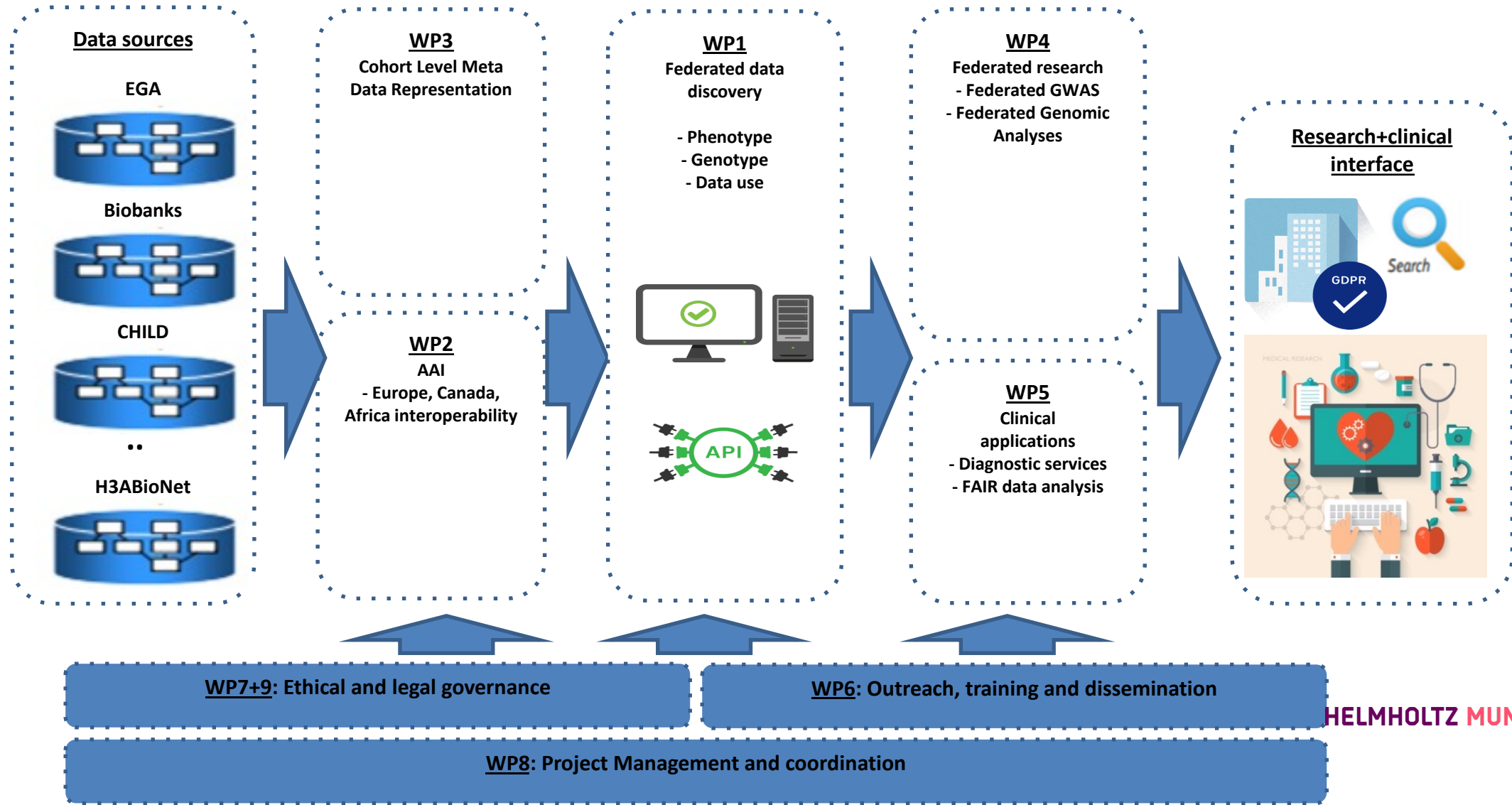


HDABs key functions and interactions

Overview of a generic data access approval process

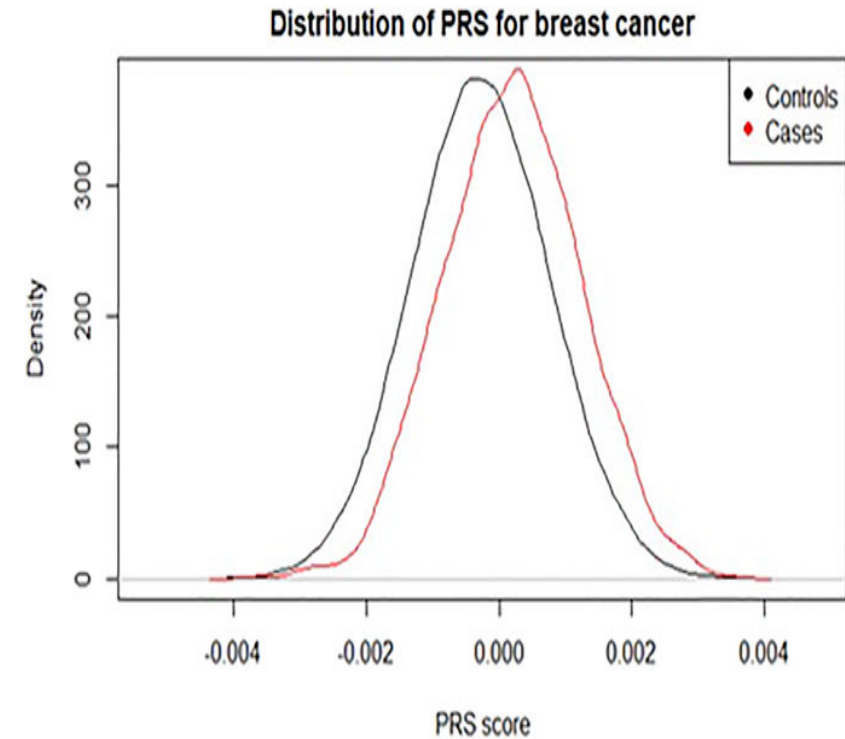


CINECA Structure



Polygenic scores

- A polygenic risk score (PRS) is a single-value estimate of an individual's genetic disposition to a trait or disease
- It is calculated by the sum of an individual's risk alleles, weighted by risk allele effect sizes derived from genome-wide associated study (GWAS) data.
- Applied in the prediction of individual risks for certain disease traits and phenotypes



Work package 4: PRS computation pipeline

Workflow language: **nextflow**

- Reproducible work flows
- Allows portablility and parallel execution
- Inclusion of other scripts for resusabilty
- Scalable to interoperable pipelines

```
nextflow.enable.dsl=2

process sayHello {
  input:
  val cheers
  output:
  stdout

  """
  echo $cheers
  """
}

workflow {
  channel.of('Ciao','Hello','Hola') | sayHello | view
}
```


Nextflow concepts

Main.nf

```
nextflow.enable.dsl=2

process foo {
  output:
    path 'foo.txt'

  script:
    """
    your_command > foo.txt
    """
}

process bar {
  input:
    path x

  output:
    path 'bar.txt'

  script:
    """
    another_command $x > bar.txt
    """
}

workflow {
  data = channel.fromPath('/some/path/*.txt')
  foo()
  bar(data)
}
```

Nextflow.config

```
process {
  withName:foo {
    container = 'image_name_1'
  }
  withName:bar {
    container = 'image_name_2'
  }
}
charliecloud {
  enabled = true
}
```

Containers: Docker,
Singularity, Podman, etc.

Executors: Slurm, AWS
batch, Kubernetes, etc.

Input/Output params file

Environment configurations
Conda , dependency loading

Nextflow features

- Metrics with regards to computation resource usage
- Pipeline sharing
- Step by step tracing and logging of each process in the workflow
- Emails and report generation

Steps implemented in workflow (1):

Module1: VCF to bed files conversion

Module2 : Target data QC steps:

- removing SNPs low genotyping rate,
 - out of Hardy-weinberg equilibrium,
 - low minor allele frequency duplicate SNPs,
 - ambiguous SNPs; strand matching.
-
- Imputation check

Steps implemented in workflow(2):

Module 3: Base QC steps

- Downloading GWAS file if provided
- Effect allele
- Mismatching SNPs removal
- Duplicate and Ambiguous SNPs removal
- To check: genome build

Module 4: PRS computation:

- Clumping, PRS computation (best fit PRS) using PRSice and/or Ldpred

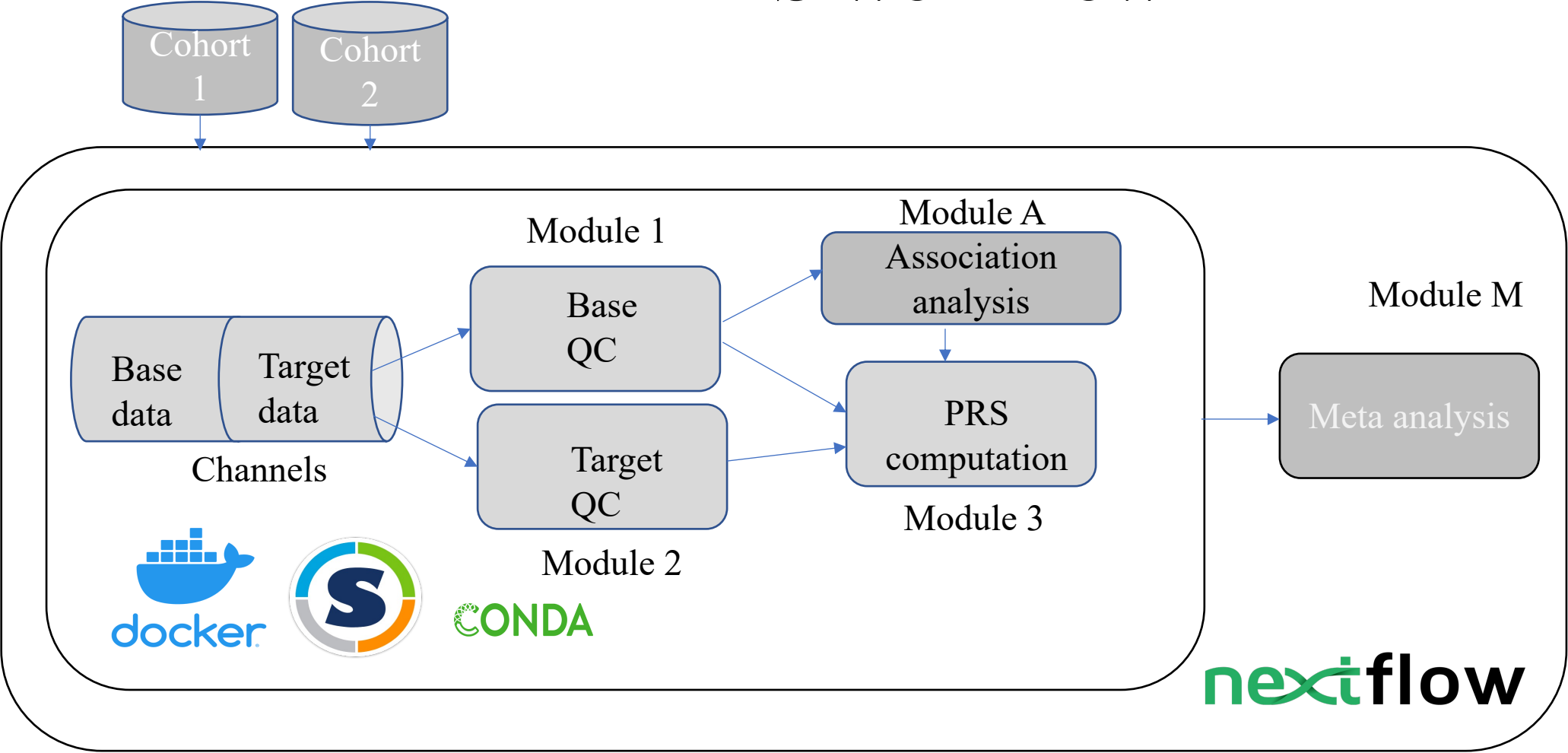
Running the Workflow

- Project migrated/checked out from desired computing environment
- Config file loaded with required runtime parameters
- Main.nf run from the desired computing environment
- Processes loaded individually in the workflow in the main script (main.nf) when dsl2 version enabled
- Imputation check and GWAS pipelines can be called separately for the target and base data separately
- Intermediate results from each process are stored in the “*work*” directory
- The status of the pipeline can be checked in either the slurm/nextflow log files

Use Case

- Genotype data: CINECA UK1 synthetic dataset, derived from the 1000 genomes project , 2508 samples, 84739838 SNPs
- Phenotype data: Height data
- Base data: Height GWAS
- Currently loaded in the project, user can provide other phenotypes and covariates corresponding to their dataset

PRS workflow



kubernetes

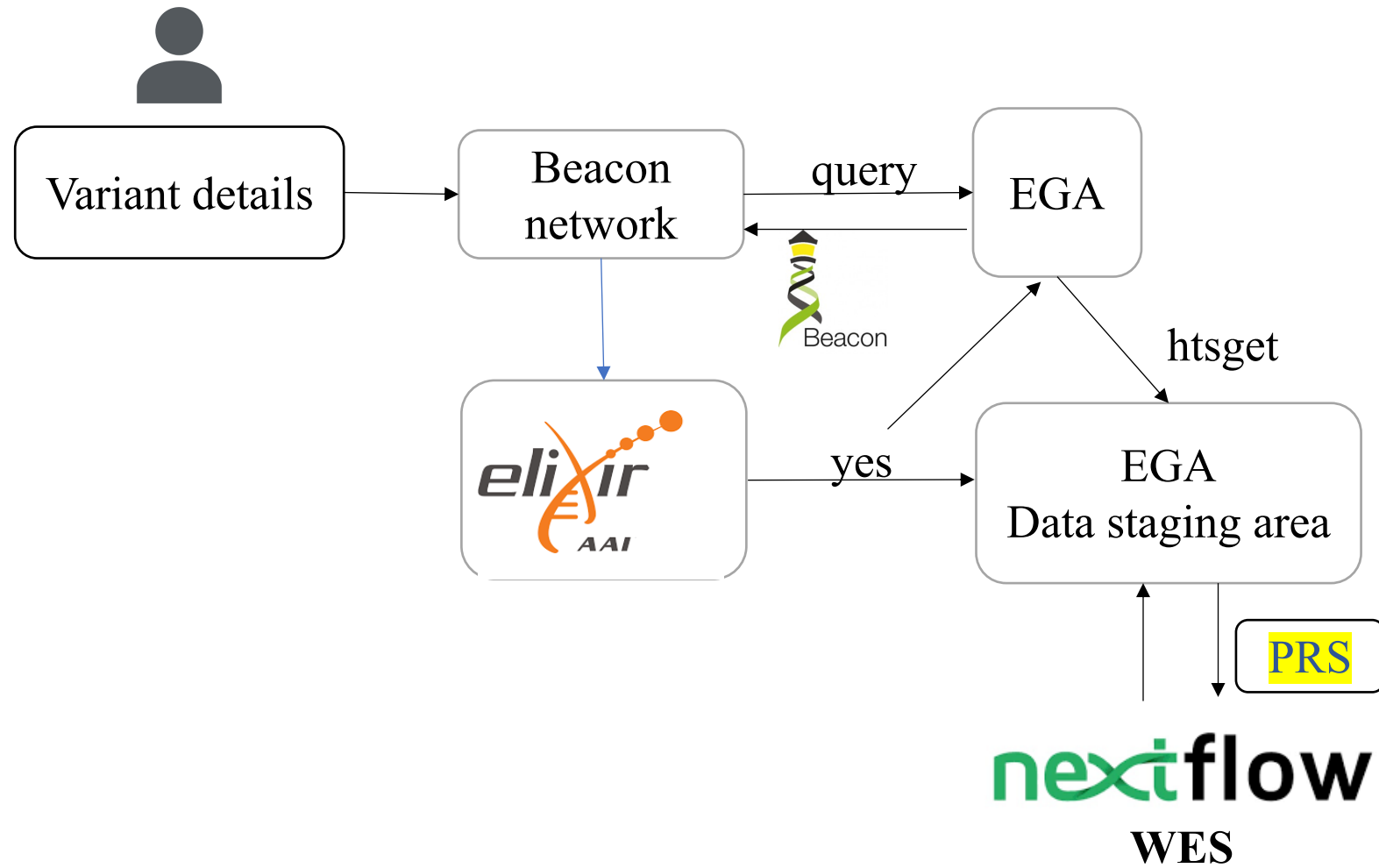


Demo

Federated Analysis

- Pipeline implemented in DSL2
- Tested on slurm (HMGU, UTartu) and Kubernetes(EKS) clusters
- eQTL catalogue pipelines also run on the HMGU cluster for monocytes dataset
- GA4GH guidelines

Federated Framework



Acknowledgements:

HELMHOLTZ MUNICH →

Will Rayner

Andrei Barysenka

YC Park

Leslie Glass

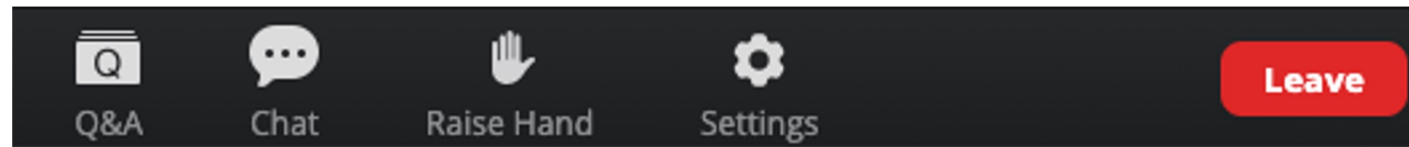


Kaur Alasso

Nurlan Kerimov



Questions?



Please write your questions on
the Zoom Q&A

Title: Federated analysis for polygenic risk score calculations

Presenter: Will Rayner and Anshika Chowdhary



Next CINECA webinar

Title: The case of data reuse: ethical, legal and societal issues in international genomic data access and sharing

Presenter: Melanie Goisauf and Emmanuelle Rial-Sebbag

Date: 9 February 2023

Time: 15:00 CET

Registration and details: <https://www.cineca-project.eu/news-events-all/ethical-legal-societal-issues-international-genomic-data>



CINECA

Common Infrastructure for
National Cohorts in Europe,
Canada, and Africa

Workshop series

Hosted by CINECA

Federated data analysis

Organisers: Nicola Mulder; Mamana Mbiyavanga - (UCT, South Africa)
Saskia Hiltemann - (EMC, The Netherlands)
Kim Gurwitz; Marta Lloret Llinares; Daniel Thomas Lopez; Carles Garcia Linares - (EMBL-EBI, UK)



Cape Town, South Africa



26 February 2023



14:00-17:30 SAST / 12:00-15:00 UTC

More Info: <https://bit.ly/federated-analysis-workshop>

info@cineca-project.eu - <https://cineca-project.eu>