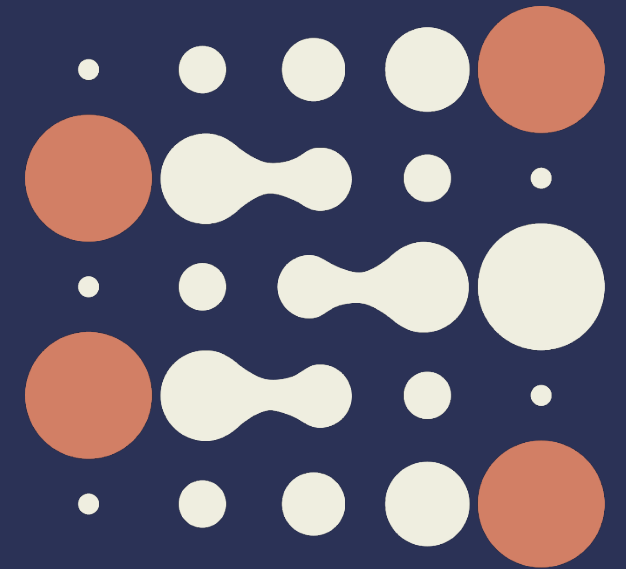


# Update on data science activities at IARC

Matthieu Foll, Vivian Viallon, Nicolas Tardy

Scientific Council 60<sup>th</sup> session – SC/60/6  
*Lyon, 7–9 February 2024*

International Agency  
for Research on Cancer



# The Data Science Steering Committee (DSSC)

- Oversees data science activities:
  - Bioinformatics and computational biology
  - Biostatistics
  - Scientific information technology (SIT)
- Composed of three working groups
- Promotes open-science:
  - FAIR principles for open data
  - open-source code sharing
  - reproducible research
- Stimulates best practices for data management, security and safety



# Activities of the IARC Bioinformatics WG

- ~ 20 active members spread across IARC scientific branches
- **Overarching objective:** facilitate interaction and knowledge sharing:
  - Organize regular internal seminar series
  - Organize training courses
  - Informal community discussions:
    - new methodologies,
    - promote the use of best-practice tools,
    - common technical aspects,
    - IT needs

# Activities of the IARC Statistical WG

- ~ 20 active members spread across IARC scientific branches
- **Overarching objective:**
  - To familiarize IARC colleagues with standard and advanced statistical tools
    - Ad-hoc technical support
    - Training courses
    - Research seminars (~ every 6 weeks)
      - causal inference
      - analysis of metabolomics or proteomics data
      - federated learning

THE UNIVERSITY OF MELBOURNE VICBioStat murdoch children's research institute

**Mediation analysis with multiple mediators:  
a target trial approach**

Margarita Moreno-Betancur

**Novel statistical analysis approaches for the analysis of  
metabolomics data**

IARC Statistics Working Group seminar

Jan Krumsiek

**Next generation pan-cancer blood  
proteome profiling using proximity  
extension assay**

20<sup>th</sup> December 2023

María Bueno Álvarez  
PhD student - Mathias Uhlen's Lab

SciLifeLab

**PDA:  
Privacy-preserving Distributed Algorithms and Statistical  
Inference in the Era of Networked Real-World Data**

Yong Chen, Ph.D., Professor of Biostatistics  
University of Pennsylvania

Seminar at International Agency for Research on Cancer  
Feb. 28, 2023

**PennCIL**  
A Computing • Inference • Learning  
lab at University of Pennsylvania  
<https://penncil.med.upenn.edu/>

# Activities of the IARC IT Working Group

3 active members:

ITS infrastructure manager, Scientist and IARC Data Protection Officer

Overarching objective: Provision IT resources to support IARC scientific activities

- In collaboration with other DSSC WGs and scientific teams identify evolving IT needs
  - Define, propose, and provision IT solutions enabling scientific projects.
  - Ensure IT resources (hardware and software) are fit to purpose and benefit the maximum of IARC scientific activities.
- Oversee IT projects progress and Monitor IT resources activities.
  - Weekly meeting for the Scientific IT platform projects.
  - Oversee activities and provision IT resources accordingly.
- Provide guidance and support to use existing resources

# Collaboration across data science WG

- **Knowledge sharing**

- Increasingly complex data, e.g., (cross-)omics data
- Synergism to be sought in the IARC statistics and bioinformatics communities
- Organization of joint seminars and training covering bioinformatics and statistics aspects:
  - **Statistics and programming with R**
  - **Introduction to multiple imputation for missing data**
  - **FAIR data principles in practice**
  - **Data visualization with R Shiny**
  - With support from the Human Resources Office/IARC learning team

- **SIT platform**

- Bottom-up approach to identify needs
- Development of resources (e.g., training, tutorials)

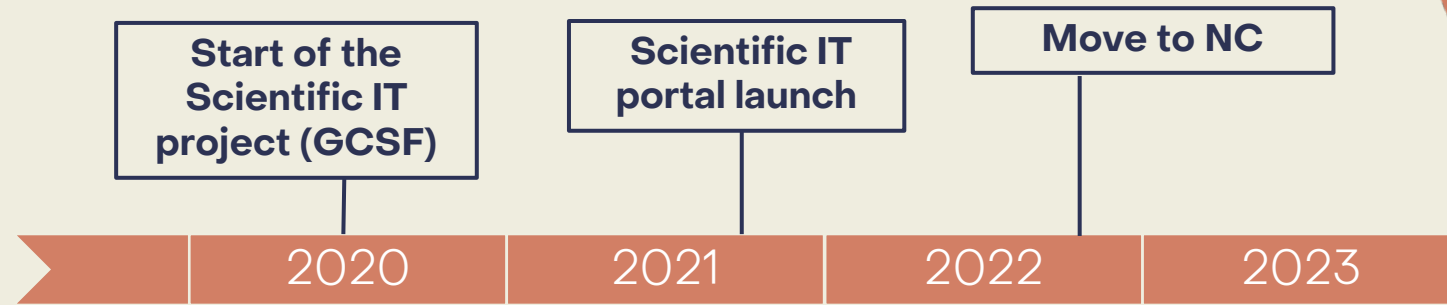
# Scientific IT Platform

Provide access to shared **centralized IT resources** for data storage and analysis and enhances:

- **Collaboration:** projects folders can be shared with multiple users;
- **Remote work:** using a web browser
- **Performance:** access to powerful machines
- **Cost effective:** avoids buying powerful personal computers
- **Security:** data stored in a secured environment and doesn't need to leave IARC premises
- **Compliance:** required by data owners to store sensitive/personal data



Scientific IT



## Phase 1

Provide **all IARC personnel** with a best-in-class **Scientific IT platform**


- + User-friendly **web portal**
- + Foundations to allow access to external collaborators (legal administrative, technical)
- + Data Protection Policy, Data Use Agreement



# Scientific IT Portal


OnDemand provides an integrated, single access point for all of your HPC resources.

## Pinned Apps A featured subset of all available apps




JupyterLab

System Installed App




RStudio

System Installed App




QuPath

System Installed App




IGV

System Installed App




Fiji

System Installed App




NextFlow Tower

System Installed App



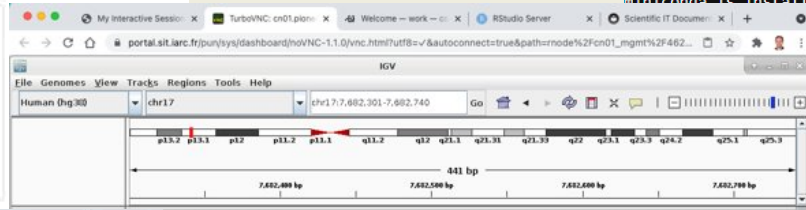
Visual Studio Code

System Installed App



Mega

System Installed App



```

My Interactive Sessions - SIT Portal x TurboVNC: cn05.pioneerx:1 (tardyn) x Dashboard - SIT Portal x
portals.it.iarc.fr/pun/sys/shell/ssh/osiris.iarc.iarc.lan

Host: osiris.iarc.lan
Last login: Mon Sep  6 16:41:56 2021 from 10.99.1.26

Welcome tardyn on OSIRIS Cluster, Iarc's HPC Infrastructure

Checking Users Environment

CHECK OUT Osiris documentation on the Scientific Computing Wiki
http://collab.iarc.fr/communities/ScientificComputing/

TIPS: About SCHEDULER: SLURM 20.02.3
visit http://slurm.schedmd.com
Specify the resources you need when submitting a job

TIPS: About DATA
/home : store your personal working file. Space limited. Incremental Backup
/data : store project structure. Check Guideline for more details about backup
/data/references : common references centrally managed
/data/databases : common databases centrally managed
/scratch : temporary file. No backup

TIPS: About APPLICATIONS
Use miniconda when appropriate to install package
miniconda is installed /home/tardyn/miniconda3
conda with conda might be in /app or ask it-services@iarc.fr
  
```

powered by OPEN OnDemand

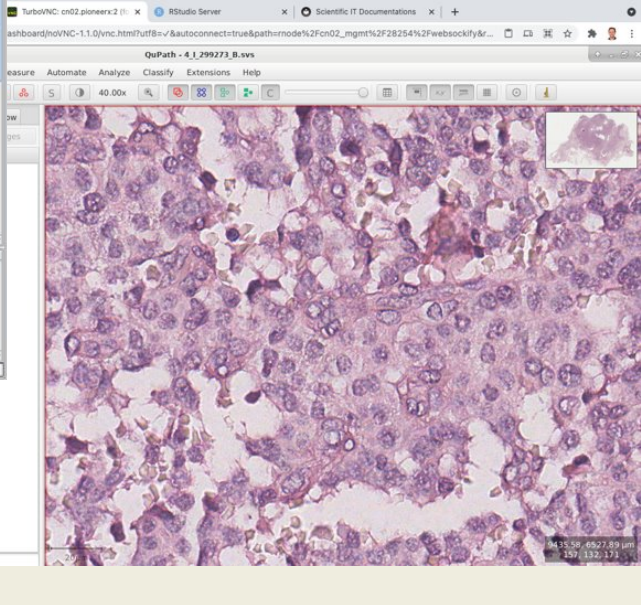
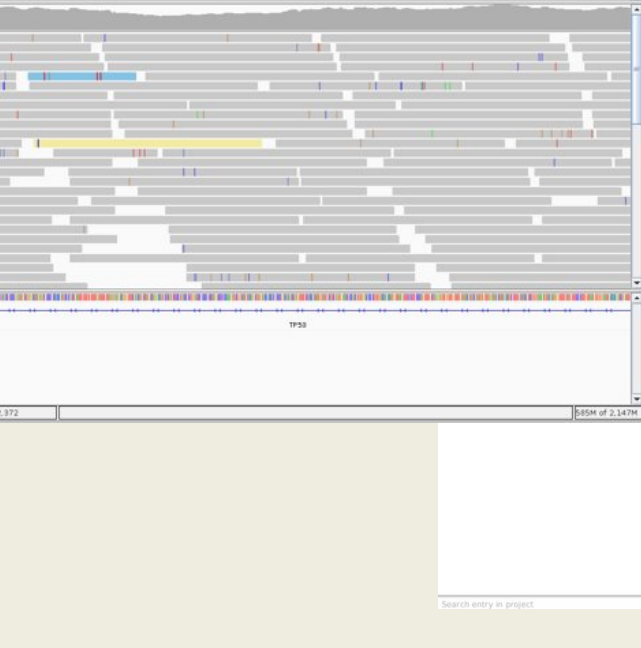
```

# WARNING: this is not the true CCF, when it's missing I just take the VAF
CCF = coalesce(CCF,adj,VAF)
Clonal = case_when(
  (is.no(Clonal) & between(CCF, 0.5, 1)) ~ TRUE,
  is.no(Clonal) & between(CCF, 0.0, 0.5) ~ FALSE,
  TRUE ~ Clonal) %>%
group_by(varID) %>%
fill(Driver.mutation) %>%
fill(Driver.mutation, direction = "up")

bla_all <- bla_all %>%
summarize(Change = case_when(
  sum(Clonal)==2 ~ "Clonal" ~> "clonal",
  diff(Clonal)==1 ~ "Subclonal" ~> "clonal",
  diff(Clonal)==1 ~ "Clonal" ~> "subclonal"),
  inner_join(bla_all, by = "varID") %>%
mutate(Clonality = fct_recodes(factor(Clonal), Clonal = "TRUE", Subclonal = "FALSE")) %>%
select(varID ~ "chr7_S5374773_GAA_G")

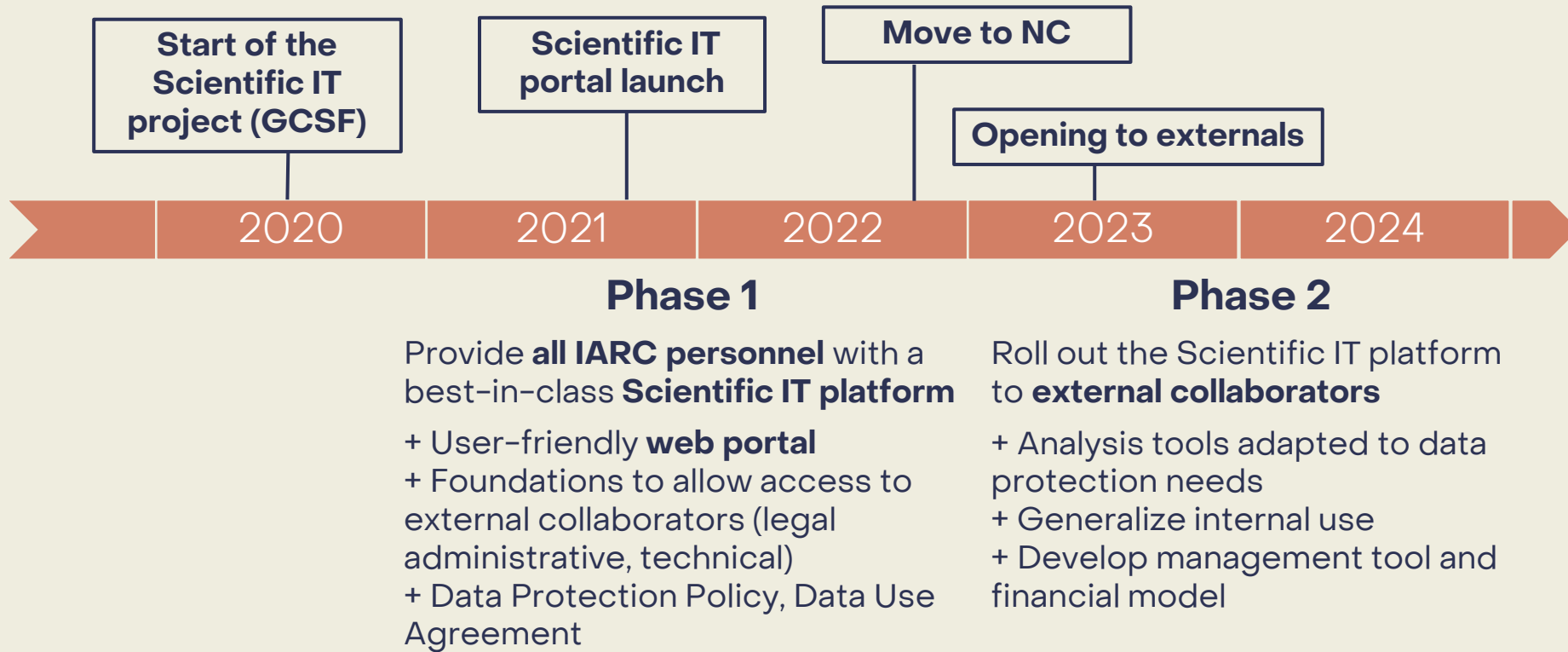
set.seed(17)

ggplot(UMAP_TM, aes(x = UMAP1, y = UMAP2, color = Molecular_clusters)) +
  geom_point(size=3, alpha=0.8) +
  scale_colour_manual(values=distinctive_cols) +
  labs(x = "UMAP dimension 1", y = "UMAP dimension 2", color = "") + theme_bw() + theme(legend.position = "bottom") +
  guides(col = guide_legend(nrow = 2))
ggplot(UMAP_TM, aes(x = UMAP1, y = UMAP2, color = SCL_mirna_profile)) +
  geom_point(size=3, alpha=0.8) +
  scale_colour_gradient2(colours = rev(brewer.pal(11, "spectral"))) +
  labs(x = "UMAP dimension 1", y = "UMAP dimension 2", color = "Neuronal score") + theme_bw() + theme(legend.position = "bottom")
ggplot(UMAP_TM, aes(x = UMAP1, y = UMAP2, color = proliferation_score)) +
  geom_point(size=3, alpha=0.8) +
  scale_colour_gradient2(low="yellow", high="red") +
  scale_colour_gradient2(colours = rev(brewer.pal(11, "spectral"))) +
  labs(x = "UMAP dimension 1", y = "UMAP dimension 2", color = "Proliferation score") + theme_bw() + theme(legend.position = "bottom")
  
```





# Scientific IT Platform



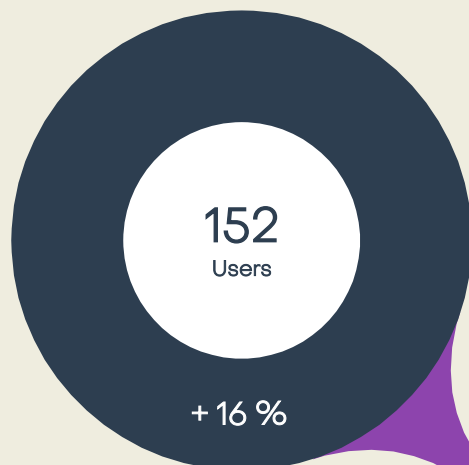
Continuous capacity & performance increase following demand; equipment renewal

# Scientific IT Platform

## Key Activities Indicators



Scientific IT

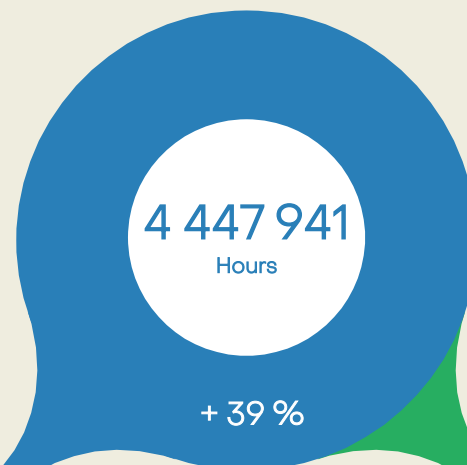
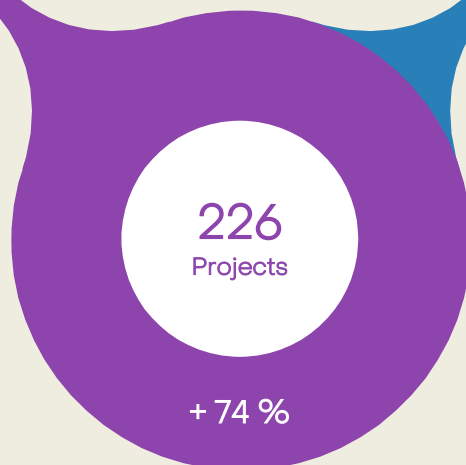


### Users

Users of the Scientific platform resources with different profiles and needs (storage, HPC, RStudio ... )

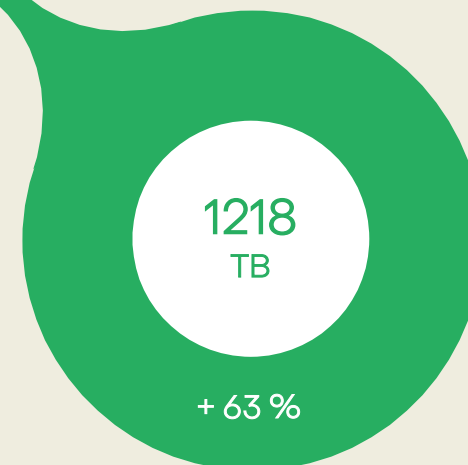
### Projects

Projects from all Scientific branches (active or archived)



### Compute Usage

Hours of CPU running analysis.



### Total Data Stored

Projects	613TB	+47%
Home	8TB	+57%
Archives	530TB	+82%

%: 2022-2023

# Scientific IT Platform Highlights



Scientific IT

## 1<sup>st</sup> SEM 2022

### Centralized Analysis

- SIT Portal New App (Mega, Fiji)

### Storage

- Merge IARC Core IT and SIT Infrastructure for Backup.

### Governance

- Creation of a working group composed of Legal, Scientific and IT people for the External Collaboration Pilot.
- External Collaborator Pilot Start

### Move

- Data Center Move to Nouveau Centre planification

## 1<sup>st</sup> SEM 2023

### Storage

- Storage Renewal Project kick Off

### Governance

- User Requirement Specifications delivered by "Do IT Now" IT Consultants.

### Move

- Data Center fully operational in Nouveau Centre

### Centralized Analysis

- Computational Resources Extension

### Governance

- Availability of Data structure for Consortium
- Kickoff project to start User Requirement Specifications by "Do IT Now" Consultant.

### Move

- Data Center effective Move to Nouveau Centre

## 2<sup>nd</sup> SEM 2022

### Centralized Analysis

- Software Portfolio Extension Project Kick Off

### Storage

- New Storage Architecture provisioning (Architecture design & RFP & Purchase)

### Governance

- Start of External Collaborator Pilot Phase 2
- Financial Model for SIT sustainability

## 2<sup>nd</sup> SEM 2023

# Scientific IT Platform

## Access to external collaborators pilot



Scientific IT

A 1st phase allowing access to external collaborators started in April 2022,

- led to the development of,
  - Data Use Agreement Template,
  - Administratives processes,
  - Technical documentation
- allowed the evaluation and the documentation of the SIT requirement such as,
  - Back-office management Tools for contracts, external collaborators access, licences, projects, ...
  - A sustainable financial model,

Those needs were **also identified in the User Specification Requirement** delivered by “Do IT Now” in 2022/2023.

A 2<sup>nd</sup> phase allowing access to a wider set of external collaborators started in June 2023 and will last while requirements are being addressed.

# Science made possible

- **Large omics analyses:** e.g. Mutographs, Rare Cancers Genomics projects etc.
- **Data-hub:** e.g. EPIC, LC3, InterLymph...
- **Emerging area:** deep-learning and AI

nature genetics



Article

<https://doi.org/10.1038/s41588-023-01321-1>

## **Multiomic analysis of malignant pleural mesothelioma identifies molecular axes and specialized tumor profiles driving intertumor heterogeneity**

Mangiante *et al.* 2023

nature communications



Article

<https://doi.org/10.1038/s41467-023-37979-8>

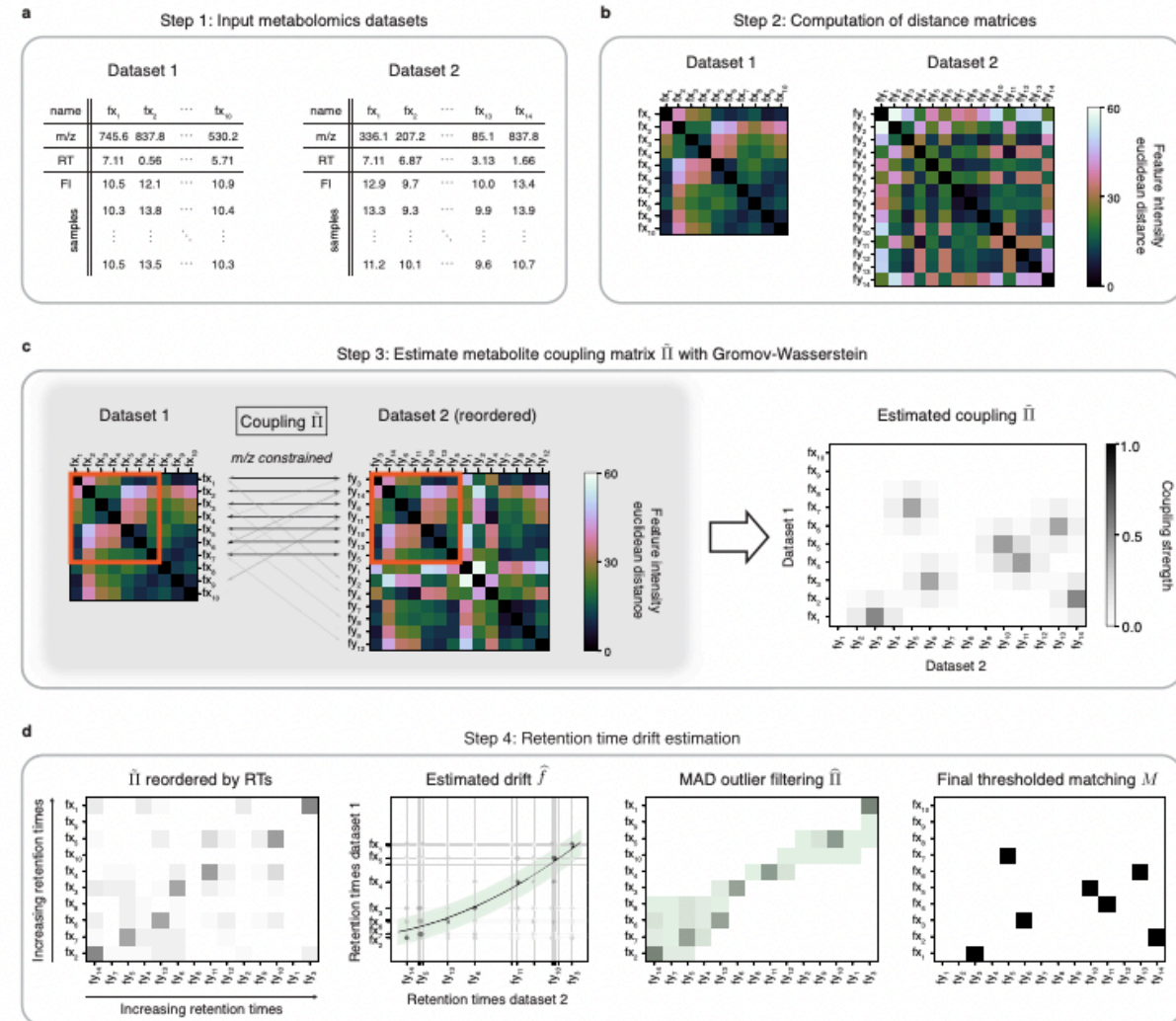
## **The blood proteome of imminent lung cancer diagnosis**

The Lung Cancer Cohort Consortium (LC3) 2023

# Optimal transport for automatic alignment of untargeted metabolomic data

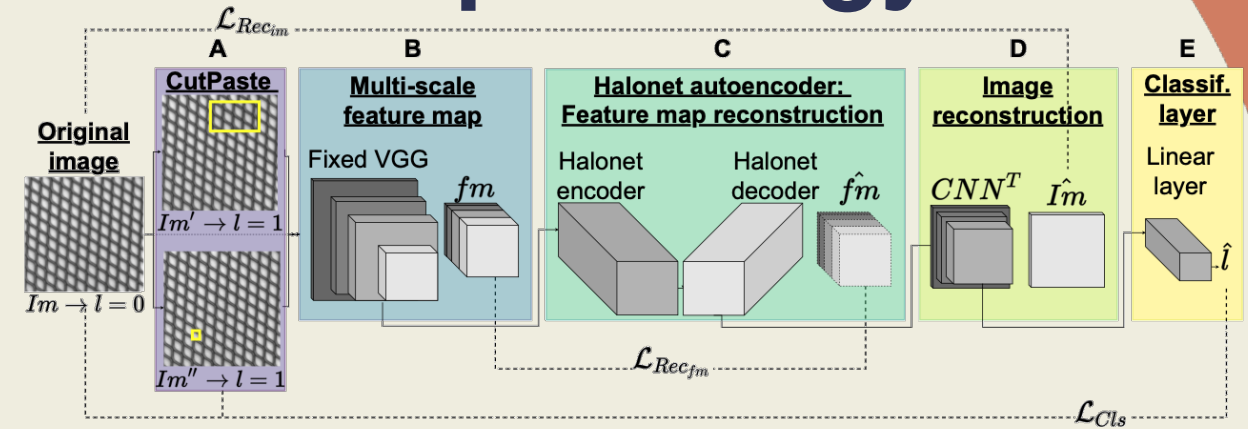
Breeur *et al.* eLife, (To appear)

- Manually pooling untargeted metabolomics data acquired in different studies is challenging and cumbersome
- We introduce **GromovMatcher**, a flexible algorithm that automatically combines untargeted metabolomics datasets using **optimal transport**
- GromovMatcher delivers superior alignment accuracy and robustness compared to existing approaches
- Application in **EPIC**, to identify **possible biomarkers of alcohol intake and study their link with risk of several cancers**

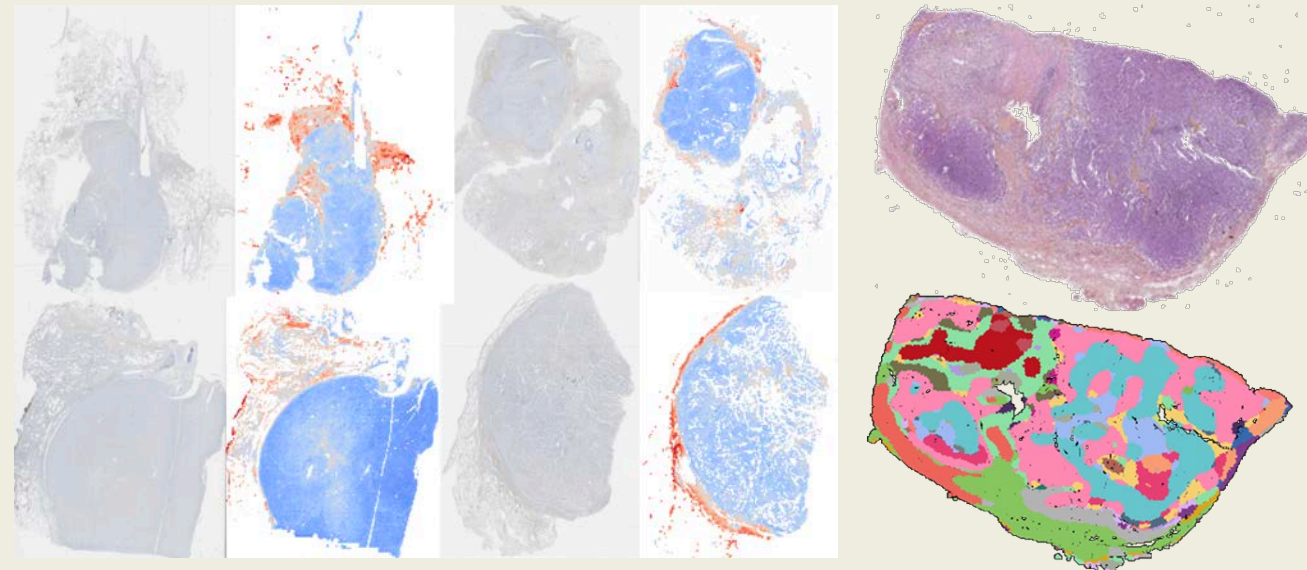


# Deep-learning for computational pathology

- We developed HaloAE: a local transformer auto-encoder for **anomaly detection**
- Drastically decreases memory and computation complexity
- Allows for the first time the application of the transformer architecture to histopathological whole slide images



Mathian VISIGRAPP 2023

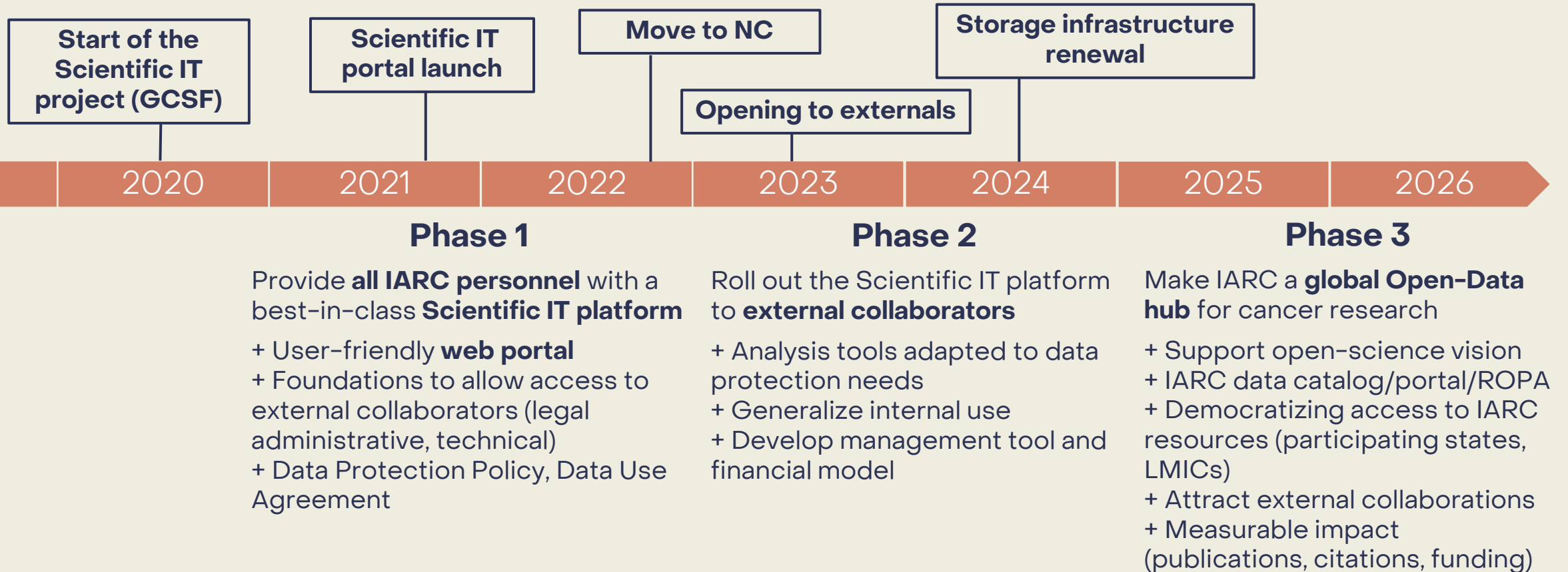


**Thank you for your attention!**

**Questions?**



# Future developments

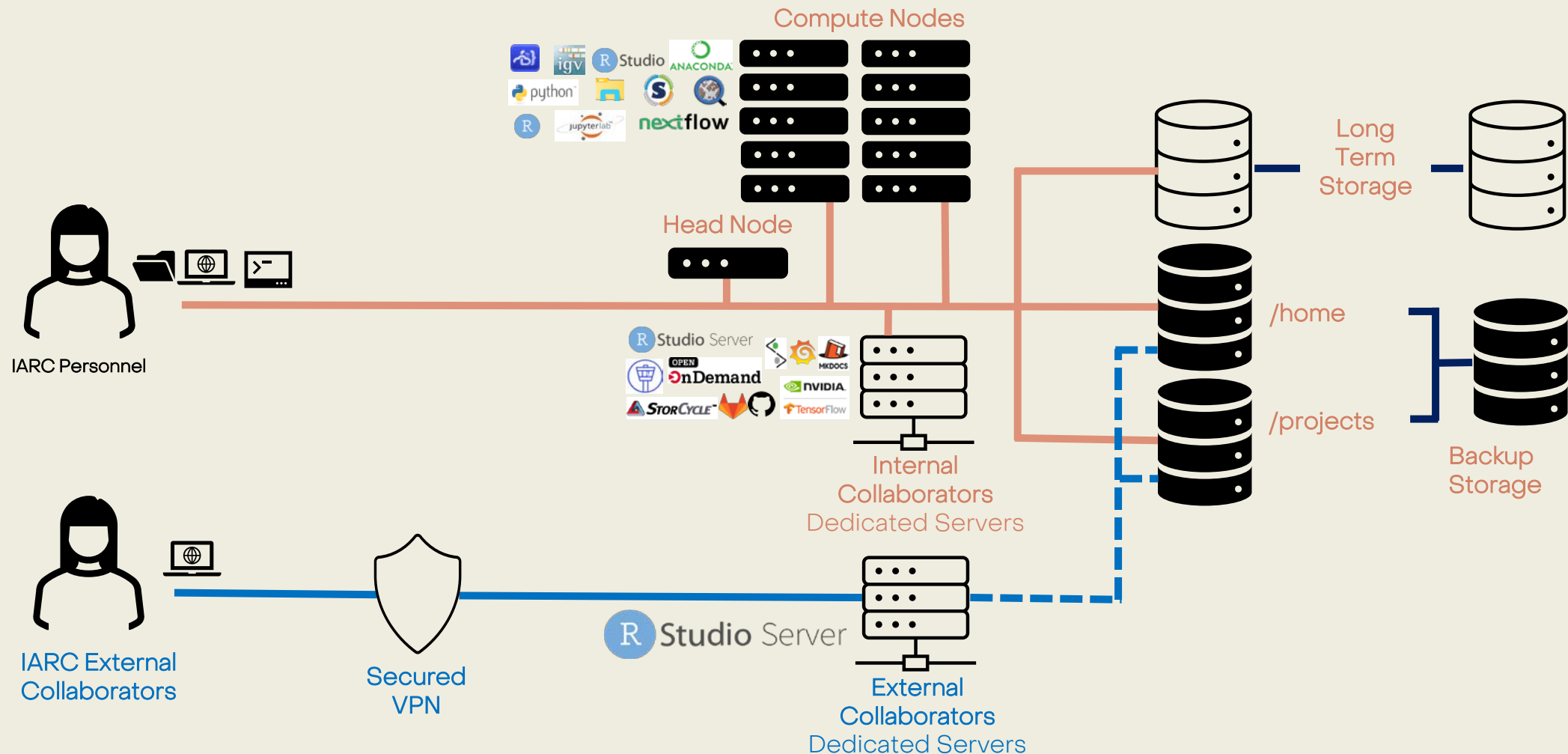


Continuous capacity & performance increase following demand; equipment renewal

# Scientific IT Platform



Scientific IT



# Scientific IT Platform:

## User Specifications Requirement



Scientific IT

Through workshop with scientific, IT and the administration, “Do IT Now” Consultant evaluated the SIT platform and documented the SIT User Specification Requirement from both a scientific and an administrative point of view,

The document describe requirements in 6 main areas :

- Data Storage,
- Data Analysis,
- Back-office Management tool,
- Financial Model,
- Scientific Data Management tool,
- Security

# Scientific IT Platform: Storage Renewal Project



Scientific IT

Following "Do IT Now" USR, the storage renewal project started in 2023 by the evaluation of systems architectures, technologies, and manufacturer. The selected scenario is the merge of storage for IARC core service and storage for scientific IT.

The publication of a “Request for Proposal” allowed the selection of the best value for money solution which was ordered in December 2023.

The implementation will take place during the first semester of 2024 providing high performance storage, enabling advanced, secure and efficient data storage,

- Full Flash performance,
- Hardware Encryption,
- Data replication for efficient and secure backup,
- Data Compression and Deduplication,
- ...