



MICROBIOTA IN HEALTH AND DISEASE

The meeting of microbiota experts

CNRS Orléans - FRANCE - November 7th 2023

Current challenges in the analysis of microbiome data - Roles of national bioinformatics infrastructures

Claudine Médigue

Genoscope/LABGeM, UMR8030 Génomique Métabolique
Institut Français de Bioinformatique, UAR3601 IFB core



Biology is concern by the 3 + 1 dimensions of the « Big Data »

VARIETY

Heterogeneous data, raw or structured data

=> complex analysis

Data standardisation ?

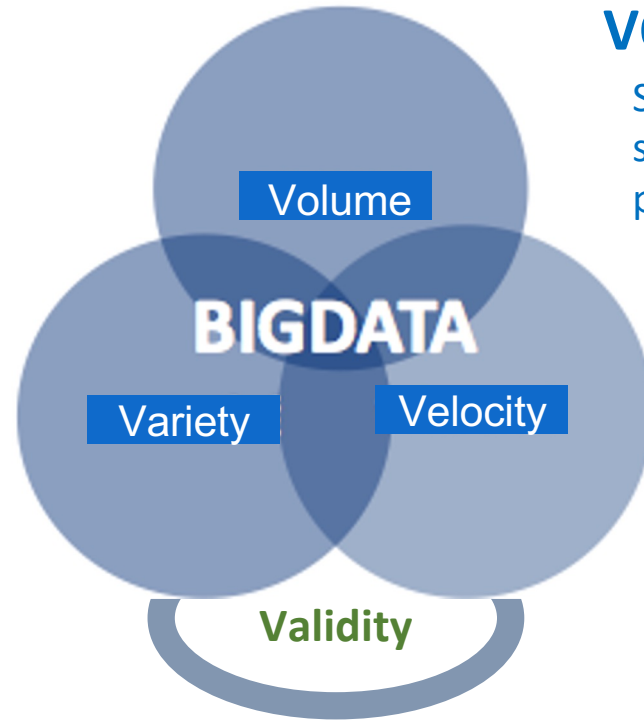
Data integration ?

VALIDITY

- Data quality
- Traceability : origin of data and metadata
- **FAIR principles**: how to make data Findable, Accessible, Interoperable and Reusable ?

Data standardisation ?

Open data and confidential/sensitive data ?



VOLUME/SIZE

Storage architectures for short and long term data preservation

Should we keep everything ?
How long ?

Nom	Symbole	Valeur
kilooctet	ko	10^3
mégaoctet	Mo	10^6
gigaoctet	Go	10^9
téraoctet	To	10^{12}
pétaoctet	Po	10^{15}
exaoctet	Eo	10^{18}
zettaoctet	Zo	10^{21}
yottaoctet	Yo	10^{24}

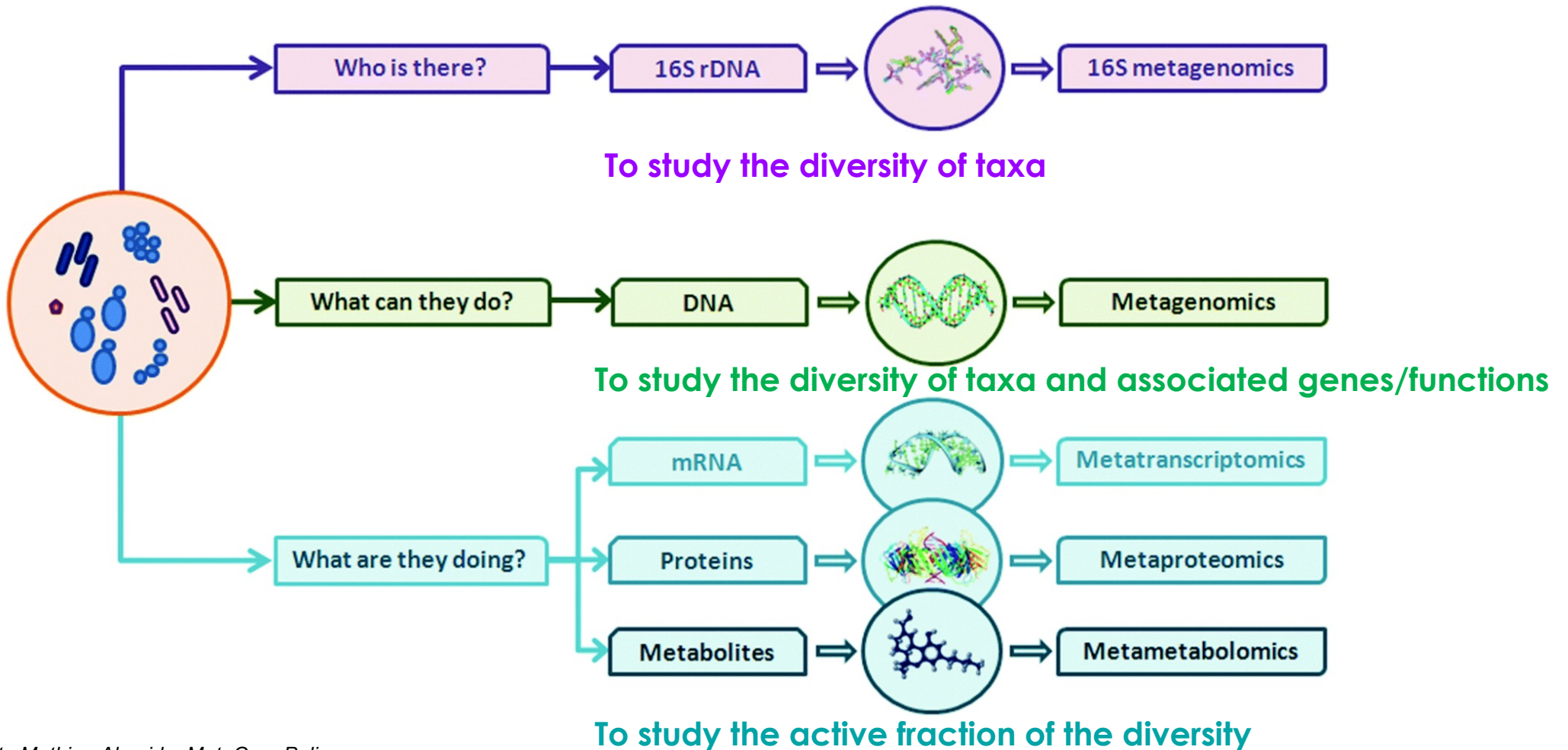
VELOCITY

Frequency at which data is generated, captured, analysed and shared
=> increasingly shorter data cycle / real time data analysis

Bioinformatics analysis of metagenomics data: pitfalls and challenges

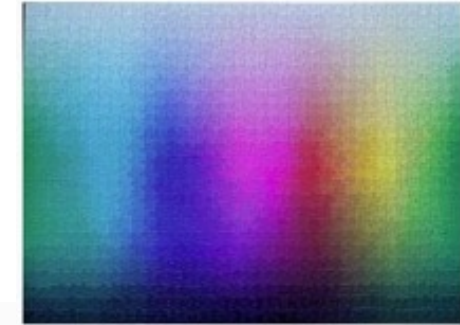


How to study the microbiome ?



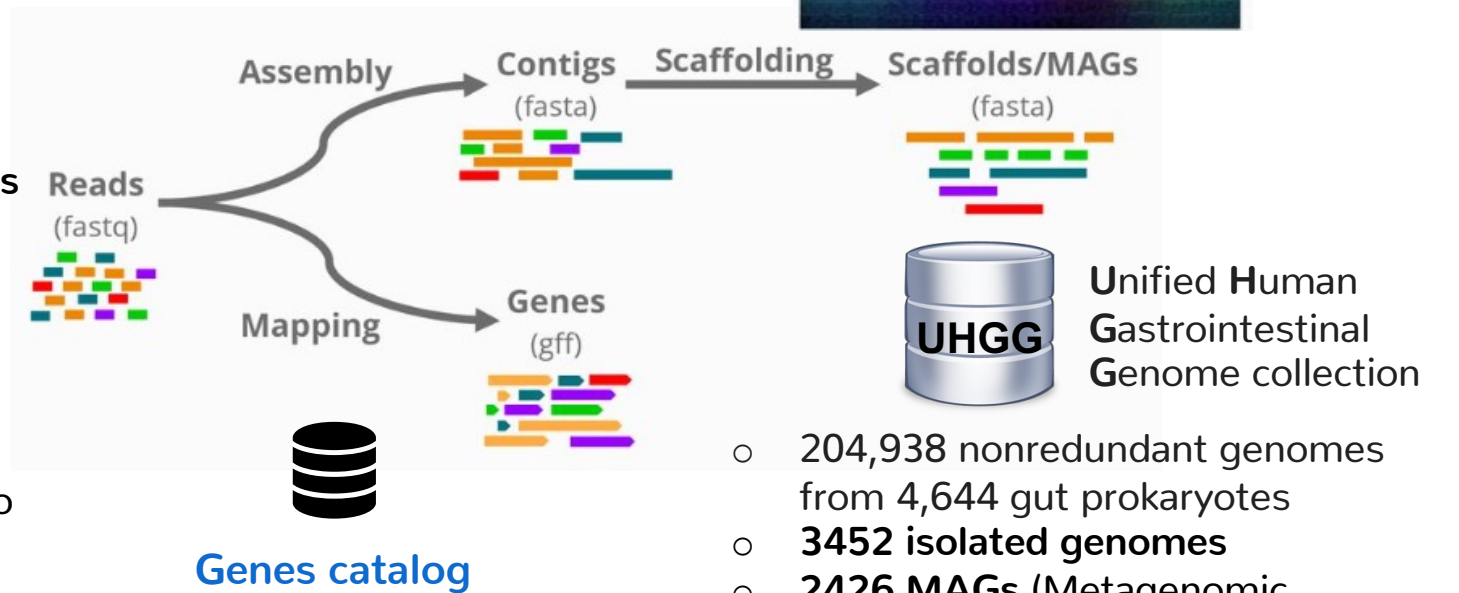
Credit : Mathieu Almeida, MetaGenoPolis

A first big challenge : assembly of metagenomics data !




Millions of reads

- thousands of different genomes mixed together
- most of the reads are missing due to **sequencing depth**
- some reads have **sequencing errors**
- **no reference genome** to refer to



From a lecture « Microbiome: metagenomics » [Beverly McDonald](#)

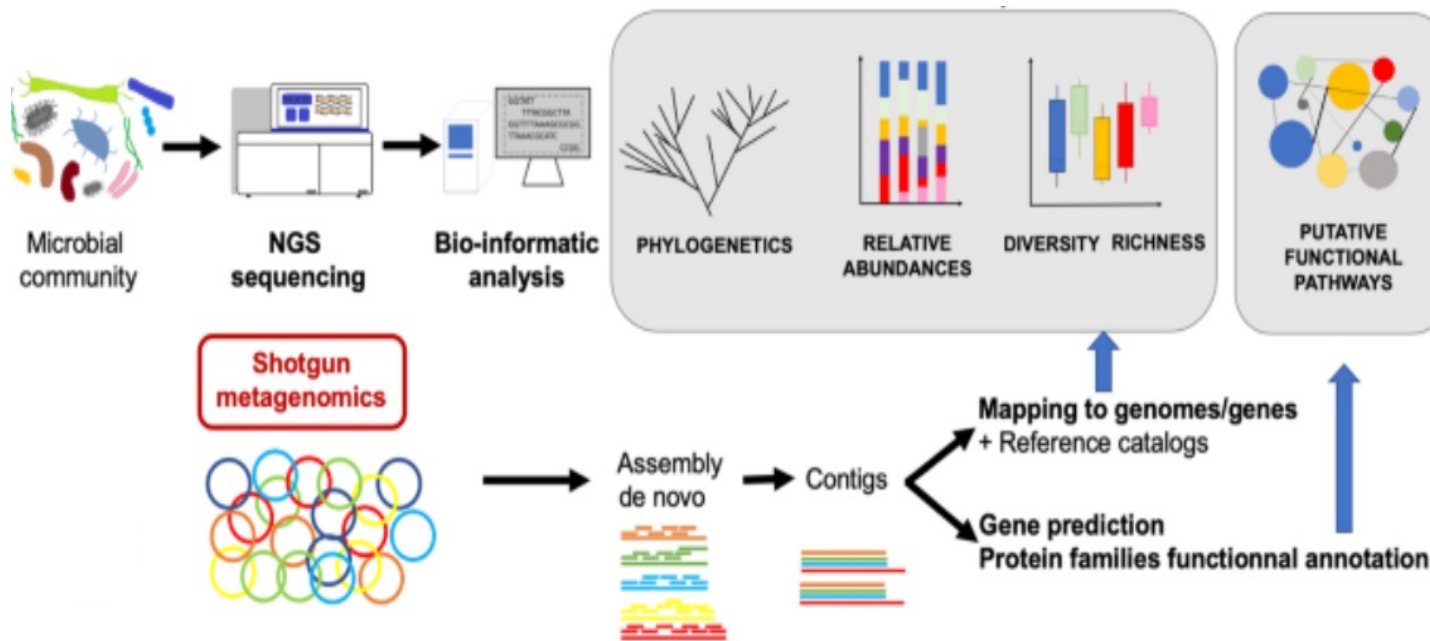
Tools & DB developed in P. Bork's Group (EMBL Heidelberg, DE)

 **Global Microbial Gene Catalog**



Structural and computational biology (EMBL Heidelberg)

 **PROGENOMES v3**



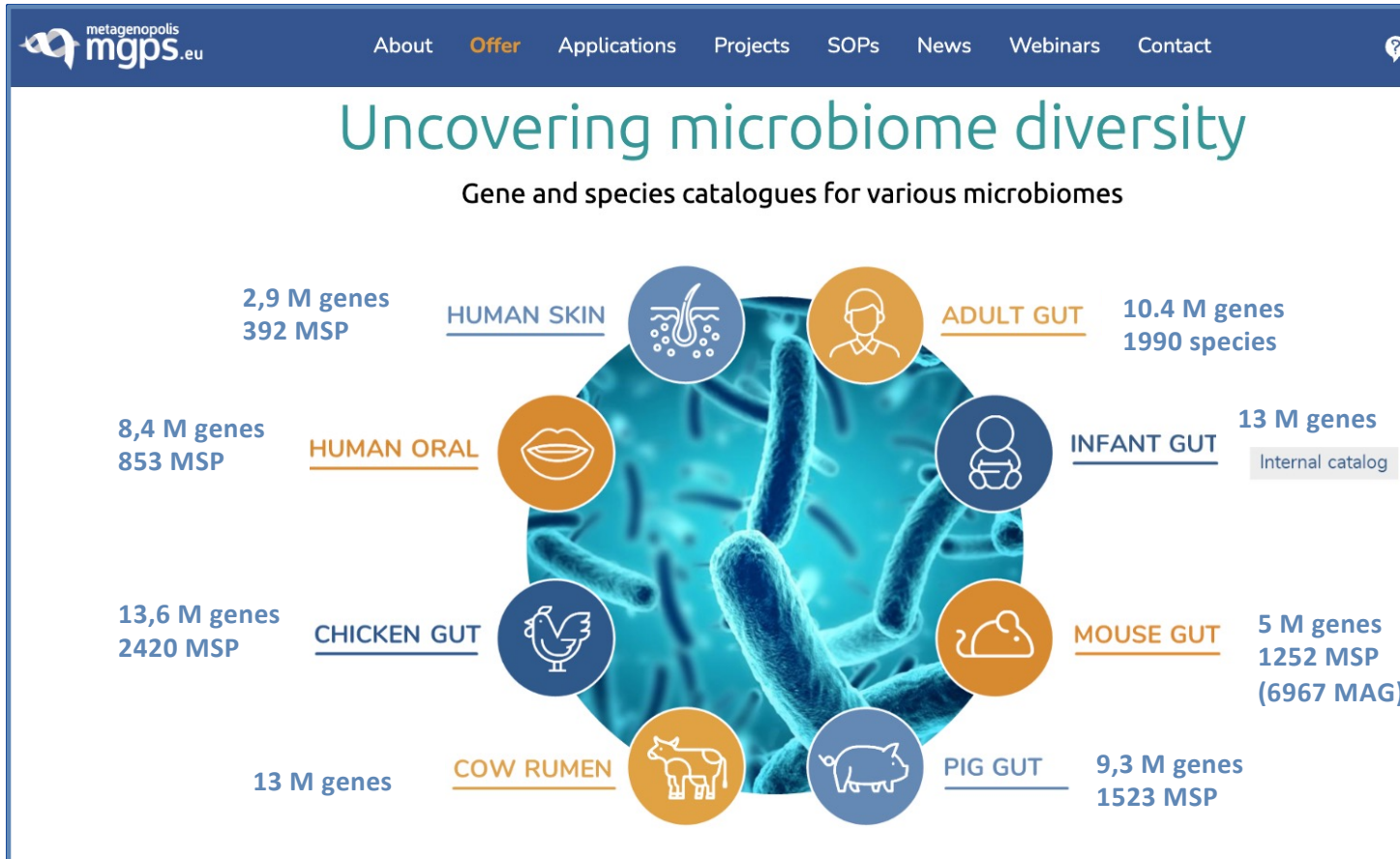
iPATH3

SMART

NOG EggNOG 6.0.0

Picture from « *Environmental Chemicals, the Human Microbiome, and Health Risk: A Research Strategy* » Washington (DC): [National Academies Press \(US\)](https://www.nationalacademies.org); 2017

Bioinformatics expertises at MetaGenoPolis [Inrae, Jouy, France]

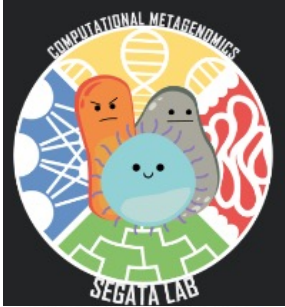


- De-novo construction of gene catalogues
- Identification, annotation and profiling of Metagenomics Pangenome Species (MPS)
- Meteor (Metagenomic Explorer): a software for profiling metagenomic data at gene level

<https://forgemia.inra.fr/metagenopolis/meteor>



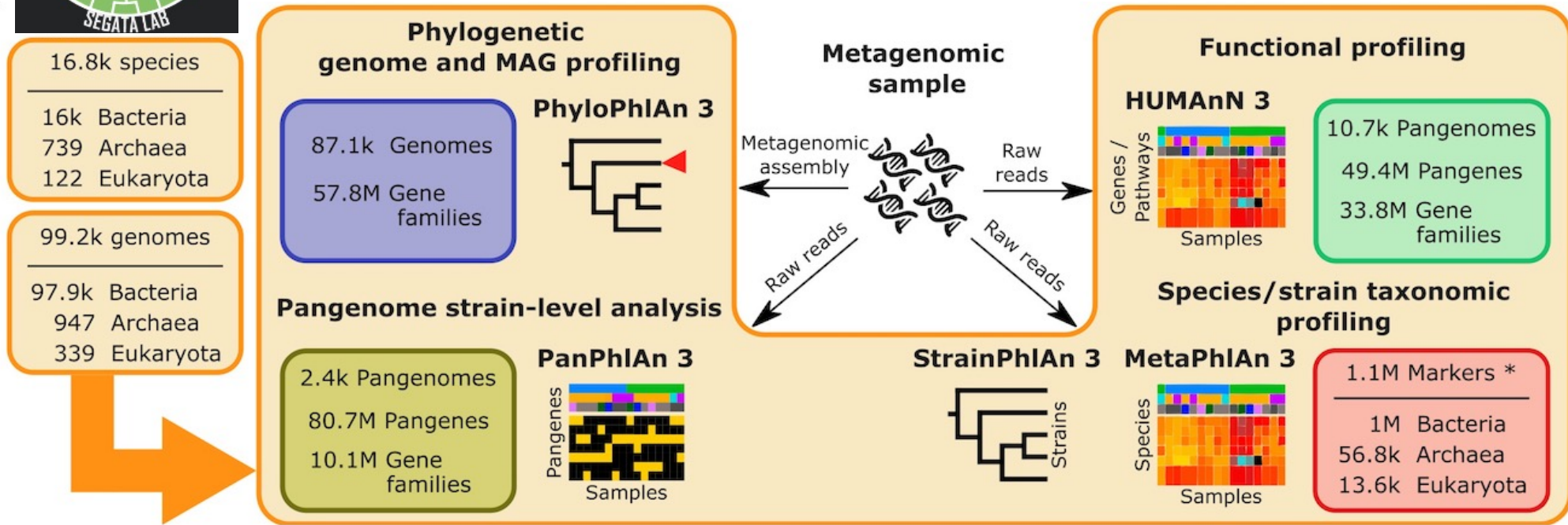
Catalogues are available at : <https://entrepot.recherche.data.gouv.fr/dataverse/mgp>



Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with **bioBakery 3**

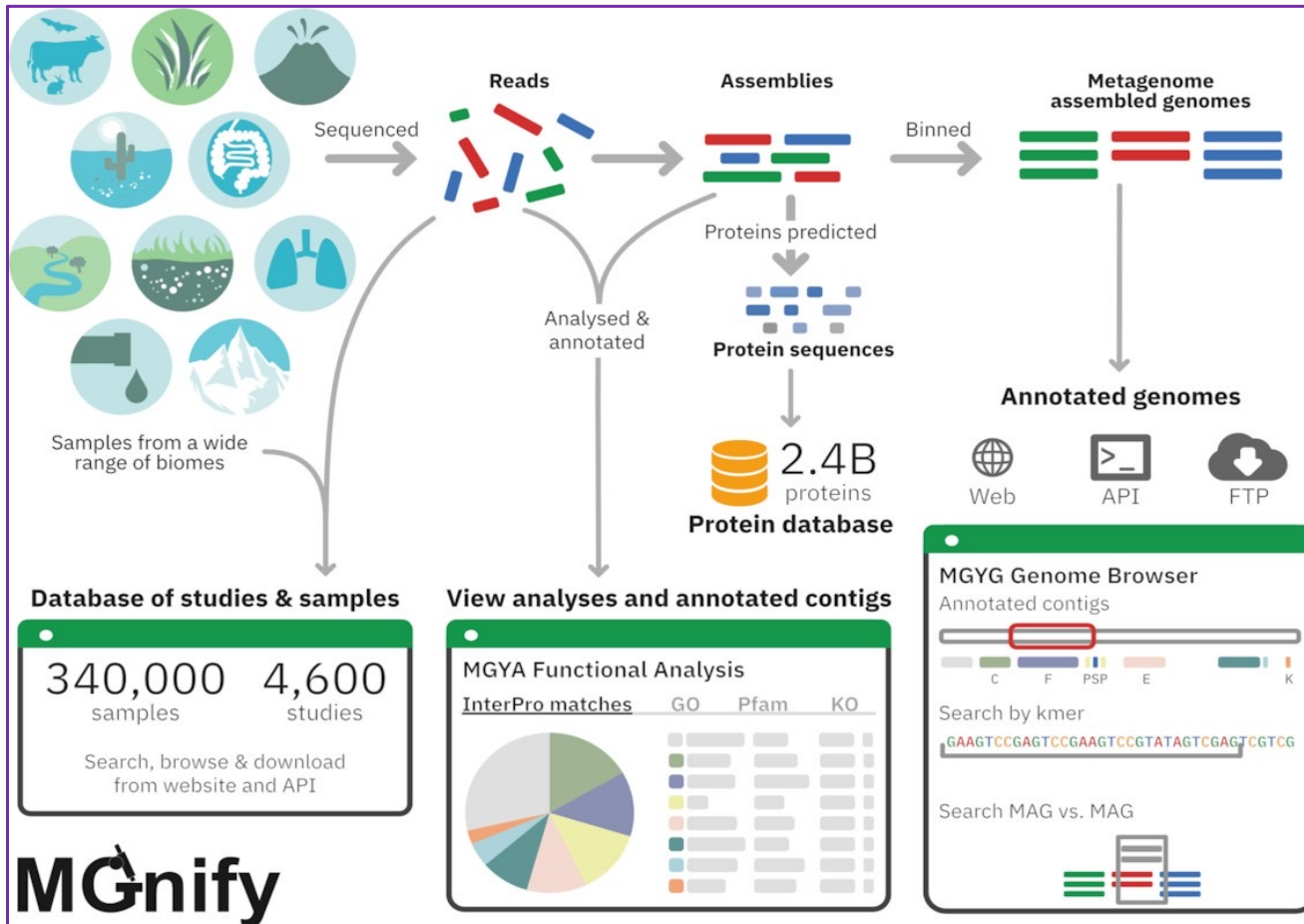
eLife, May 2021

<https://doi.org/10.7554/eLife.65088>



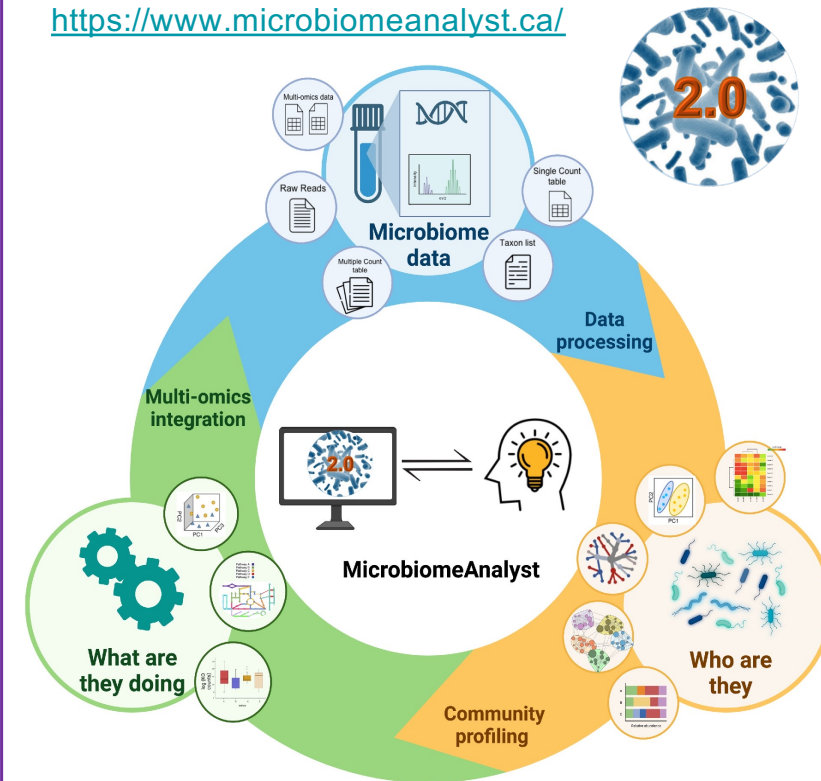
<http://segatalab.cibio.unitn.it/tools/biobakery/index.html> (tools freely downloadable)

Some famous web sites / platforms to study microbiomes



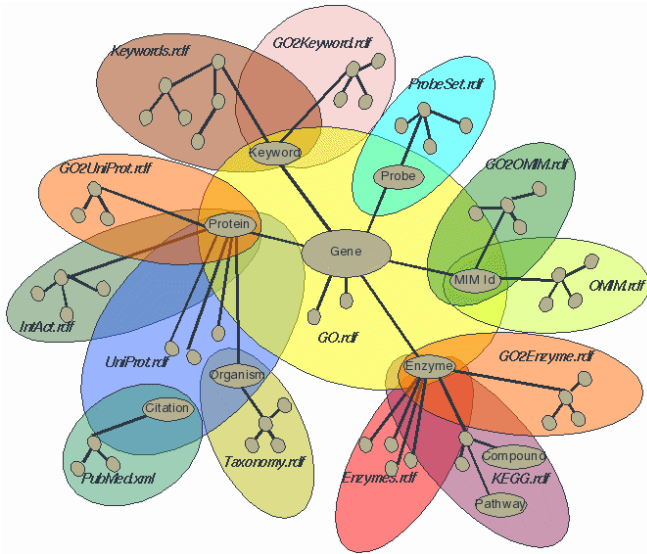
Nucleic Acids Res, Issue D1, 6 January 2023
<https://doi.org/10.1093/nar/gkac1080>

MicrobiomeAnalyst – comprehensive statistical, functional and integrative analysis of microbiome data
<https://www.microbiomeanalyst.ca/>



Nucleic Acids Res, Issue W1, 5 July 2023
<https://doi.org/10.1093/nar/gkad407>

To understand the challenge of data integration



Multiple databases



Syntactic

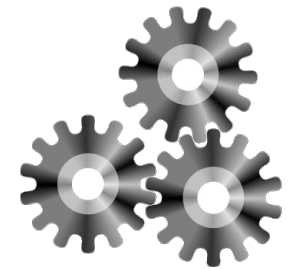
Various representation in different databases
=> **Conflicts of name, of data type, of attributes, ...**



Semantics

Different meaning, interpretation or use of the same data

- Definition of **standards** (syntactic interoperability)
 - Definition of **formal ontologies** (semantics interoperability)
- => Syntactic and semantics integration of biological data**



Standardisation of data and metadata

- **data standard** are used to
 - analyse, compare and exchange data
 - publish datasets in international resources
- **metadata standard** are used to
 - describe data richly and accurately, with the same vocabulary as the rest of your scientific community
 - make metadata interoperable and to allow other systems to exploit them
- **ontologies** are used to describe data and metadata
 - ontologies are hierarchies of well-defined and standardized vocabulary
 - their main purpose is **to make metadata searchable, comparable and machine-readable.**

The human microbiome field needs to reinforce their standards for accelerating its science !

- most datasets are used only by the researchers that have generated the data for their particular study without consideration of standardization :
 - submission in public resources is often a complex task and procedures are heterogeneous
 - metadata are often incomplete, inconsistent, redundant or not enough informative
- => transfer of products or applications from microbiome research to the medical field is still limited.

Need to educate researchers to make more accessible and easily shareable their data : towards FAIRification

What are the FAIR (Findable, Accessible, Interoperable, Reusable) principles ?



The **FAIR** principles were developed to:

- support **knowledge discovery and innovation** both by humans and machines
- support data and **knowledge integration**
- support new discoveries through the **harvest and analysis of multiple datasets** and outputs
- **promote sharing and reuse** of data
- improve the **reproducibility and reliability** of research

How to make project and data FAIR ?

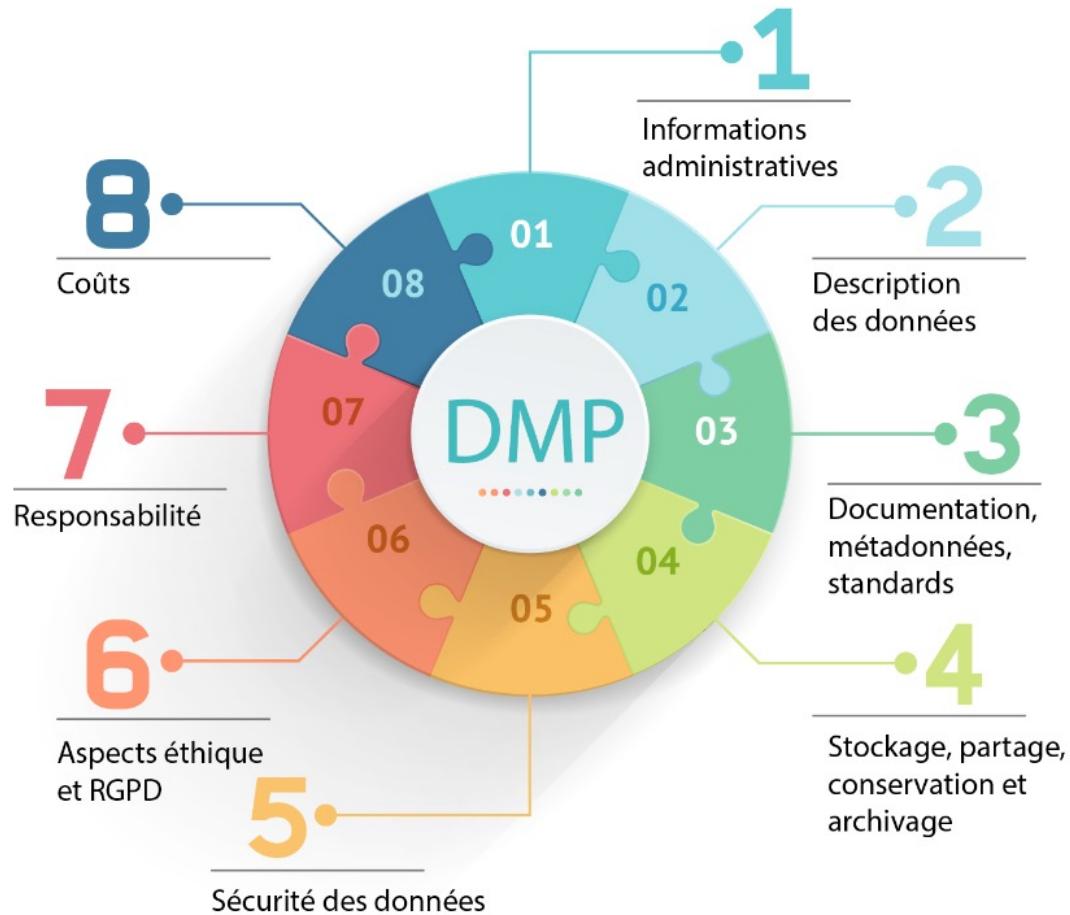
<https://faircookbook.elixir-europe.org/content/recipes/introduction/fairification-process.html>

1) Use common and adapted vocabularies, ontologies, standards, formats and check-lists

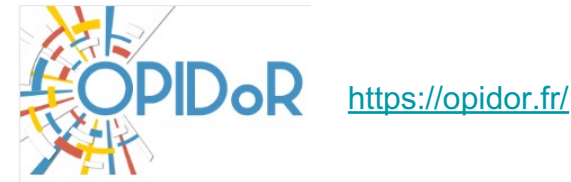
e.g. STORMS = Strengthening The Organization and Reporting of Microbiome Studies (<https://www.stormsmicrobiome.org/figures/>)

2) Data Management Plan (DMP)...

Data Management Plan (DMP)



1. Plan the management of the project
2. Describe how the data is obtained
3. Ensure the data is understandable
4. Providing appropriate data storage
5. Describe how data are secured
6. Clarify the legal and ethical framework
7. Define everyone's responsibilities
8. Estimate the costs !



How to associate specific phenotypic metadata (clinical, diet, lifestyle...) ?

- Standards as GSC-MIMS (Genomic Standard Consortium about the Minimum Information about Metagenomic Sequence) are not complete enough to take into account these phenotypic data.
- Some standards for dealing with clinical data:
 - SNOMED CT (<https://www.snomed.org>)
 - LOINC (<https://loinc.org/>)
 - FHIR (<https://www.hl7.org/fhir/>) - FAIRness for FHIR: Towards Making Health Datasets FAIR Using HL7 FHIR (Martinez-Garcia et al, 2022)
- Phenotypic metadata should be **considered sensitive personal (health) data according to GDPR** (General Data Protection Regulation) : **take care about authorization of data sharing !**
- More precisely, datasets that include health information and other sensitive personal data **should be sufficiently structured or anonymized for lowering identification risk.**

So, what about metagenomic sequences from human microbiomes ?...

Is metagenomic data considered as sensitive personal data ?

nature microbiology



Article

<https://doi.org/10.1038/s41564-023-01381-3>

Reconstruction of the personal information from human genome reads in gut metagenome sequencing data

Received: 6 April 2022

Accepted: 12 April 2023

Published online: 15 May 2023

Check for updates

Yoshihiko Tomofuji^{1,2,3}, Kyoto Sonehara^{1,2,4}, Toshihiro Kishikawa^{1,5,6}, Yuichi Maeda^{2,7,8}, Kotaro Ogawa⁹, Shuhei Kawabata¹⁰, Takuro Nii^{7,8}, Tatsusada Okuno⁹, Eri Oguro-Igashira^{7,8}, Makoto Kinoshita⁹, Masatoshi Takagaki¹⁰, Kenichi Yamamoto^{11,12}, Takashi Kurakawa⁸, Mayu Yagita-Sakamaki^{7,8}, Akiko Hosokawa^{9,13}, Daisuke Motooka^{2,14}, Yuki Matsumoto¹⁴, Hidetoshi Matsuoka¹⁵, Maiko Yoshimura¹⁵, Shiro Ohshima¹⁵, Shota Nakamura^{2,14,16}, Hidenori Inohara⁵, Haruhiko Kishima¹⁰, Hideki Mochizuki⁹, Kiyoshi Takeda^{8,16,17}, Atsushi Kumanogoh^{2,7,8} & Yukinori Okada^{1,2,3,4,12,16} ✉

=> Importance to **eliminate human DNA trace in metagenomic samples** in order to reduce risk of identification



Identifying personal microbiomes using metagenomic codes

Eric A. Franzosa^{a,b}, Katherine Huang^b, James F. Meadow^c, Dirk Gevers^b, Katherine P. Lemon^{d,e}, Brendan J. M. Bohannan^c, and Curtis Huttenhower^{a,b,1}

^aBiostatistics Department, Harvard School of Public Health, Boston, MA 02115; ^bMicrobial Systems and Communities, Genome Sequencing and Analysis Program, The Broad Institute, Cambridge, MA 02142; ^cInstitute of Ecology and Evolution, University of Oregon, Eugene, OR 97403; ^dDepartment of Microbiology, The Forsyth Institute, Cambridge, MA 02142; and ^eDivision of Infectious Diseases, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115

Edited by Ralph R. Isberg, Howard Hughes Medical Institute, Tufts University School of Medicine, Boston, MA, and approved April 6, 2015 (received for review December 15, 2014)

- Microbiome of a same individual is quiet stable (at gene or strain level) during time.
- Microbiome-based identifiability does not match the exceptionally high specificity of genomic identifiability
- However, **microbiome-based identifiability is possible** for a nontrivial fraction of individuals in a typical cohort: a **potential genetic information privacy issue** not typically considered in microbiome study design.

Personal data protection and Open Science in the field of human microbiota science



Environment International
Volume 165, July 2022, 107334



Position paper on management of personal data in environment and health research in Europe

Govarts *et al.* : <https://doi.org/10.1016/j.envint.2022.107334>

« Balancing the aim of open science driven **FAIR** data management with **GDPR** compliant personal data protection safeguards is now a common challenge for many research projects dealing with (sensitive) personal data »

FAIR is not equal to Open: The 'A' in FAIR stands for 'Accessible under well defined conditions'
Open as possible, Closed as necessary

Technical solutions to anticipate future regulations

- Access-controlled repository (e.g. EGA)
- **Federated infrastructure** (bring algorithms to the data ; e.g, FeMAI, Microb-AI-ome projects)
- **Virtual Research Environments** (data remotely accessible and combinable with other data sources without the need for transferring the data ; e.g. DataSHIELD [Gaye et al. 2014])
- **Sensitive (health) data repository** : security bubble with strong access control

This requires several considerations such as secure data storage and data exchange (authentication, authorization)



INSTITUT FRANÇAIS DE BIOINFORMATIQUE



The French Bioinformatics Infrastructure & ongoing or ready to start National projects



IFB, a federation of bioinformatics facilities and research teams

A distributed infrastructure

- 36 platforms and research teams
- Wide geographic coverage
- Expertise in all the domains of bioinformatics

Mutualised resources

- Funding for the national infrastructure PIA1 RENABI-IFB (22.8M€), PIA3 MUDIS4LS (16.5M€)
- Missions organised around task forces mutualised between platforms

Reaching out all the communities

- Fundamental biology, health, agriculture, environment

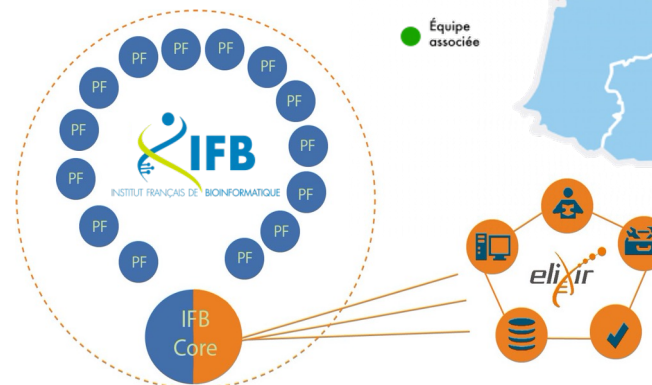
Coordination: IFB-core (UAR 3601)

- multi-organisms (CNRS, Inserm, INRAE, CEA)
- mission : coordination and management of the mutualised resources
- ELIXIR-FR : French node of [ELIXIR](#), the European bioinformatics infrastructure (ESFRI)

22 member platforms
7 associated platforms
8 associated research teams
>400 experts (~200 FTE)



Paris & Île-de-France	
IFB Core	PlantBioinfoPF
ARTbio	RPBS
Curie Bioinfo	I2BC Bioinfo
DAC	ORPHANET
MICROSCOPE	PB-IBENS
MIGALE	Biomics
Pasteur HUB	EvryRNA



Missions and activities

Digital infrastructure



- National Network of Computing Resources (**NNCR**)
- Two computing modalities: **cluster + cloud**
- Data protection (**GDPR**)

11 Po storage
21K CPUs

Software tools



- **Development** of specialized software
- **Packaging** (conda)
- **Virtualization** (appliances, containers)
- **Best practices** of software engineering
- **Collaborative development, open code**

> 900
available tools

User support



- A single entry point for all the IFB services and platforms
- Support from the conception to the publication
- Data management plans (**DMP**)
- Conception and implementation of **workflows**
- **Data science**

400 experts
~200 FTE

9115 users
310 collaborative projects



Knowledge bases



- Development of databases and knowledge bases
- Technical support to database developers
- Deployment of databases developed by French teams

Training

- **Thematic schools** (NGS, multi-omics, phylogeny, biostat.)
- **Bring Your Own Data (BYOD)** format
- Webinars, MOOC
- Open science: **FAIR-data** and **FAIR-bioinfo** training
- Adaptation to the evolution of the needs

2022 KPIs
136 training sessions
2326 trainees
341 days of training

Prospective and innovation



- Identification of emerging needs in bioinformatics
- Pilot-projects to address the needs:
 - Integrative bioinformatics
 - Integration of health data
 - Artificial intelligence for biology and health

Open science & interoperability



- Data management
- Interoperability, standards, ontologies...
- Data brokering (user support to deposit data in repositories)
- Machine-actionable Data Management Plans (maDMPs)
- Collaboration with data-producing national infrastructures

Support to researchers in life science

National Network of Computing Resource (NNCR)

- Two complementary **operating technologies** (6 batch clusters ; 8 clouds)
- 9 French sites
- 27,000+ CPUs-HT, 135 Tb RAM, **14 Pb storage**

Managed by mutualised teams (**32 persons, 7.5 FTE**)

- Engineers from IFB platforms who share expertise
- Mutual aid to solve problems
- Sharing solutions (developments, deployment recipes)

Collaboration with regional and National computing centers (CCIN2P3, IDRIS, CINES) G

Compute & Storage infrastructure	2021	2022
Total number of active accounts	10 900	10 329
New accounts opened per year	2 884	4 272

<https://www.france-bioinformatique.fr/en/ifb-clusters/>
<https://www.france-bioinformatique.fr/en/ifb-cloud/>

Support to researchers in Life Sciences

=> support biologists for the analysis of their data and the exploitation of their results.

- **Anticipation:** ideally (but unfortunately not frequently), help for the **design of the projects** (approaches, sampling plan, feasibility, required person•months)
- **A la carte:** each project requires customized workflows, individual analyses, specific interpretation.
- **Long lasting:** data analysis and result exploitation typically take 1-2 years between raw data production and publication
- **Exploratory:** questions and approaches are generally evolving / refining during the project
- **Subcontracting vs collaboration:** long-term projects can hardly rely on invoiced services.
Alternative: collaboration with co-publication

<https://www.france-bioinformatique.fr/en/help-desk/>

Training

- **Thematic schools**
 - Next-generation sequencing [EBAII](#)
 - Integrative bioinformatics ([ETBii 2023](#))
- **DUBii: Diplôme universitaire en bioinformatique intégrative [IFB + UnivParis Diderot (2019-2021)]**
- **IFB training courses about FAIR principles**
 - **FAIR data** (with WG Open Science & Interoperability)
 - **FAIR bioinfo** (with I2BC): reproducibility of the analyses
- **Mutualized resources and trainer toolboxes**
 - [IFB training catalog](#)
 - [Moodle](#) platform
 - Forge logicielle gitlab IFB
 - Galaxy Training Network
- **Innovation and good practices**
 - FAIRization of training material
 - Training paths (successions of training courses)
 - IFB trainer network
 - [e-learning workgroup](#)



Training	2021	2022
Training sessions organised	132	160
Total number of trainees	4 458	4 936
Percent trainees satisfied or very satisfied	95%	92%

<https://www.france-bioinformatique.fr/formations/>

Open science & Interoperability

- **OpenLink**: data management dashboard (Funding: ANR-19-DATA-0011-01)
- **maDMP4LS**: machine-actionable Data Management Plans, linking the allocation of storage resources with the DMP (Funding: ANR-19-DATA-0017-01)
- **Modular DMPs** for data integration (collaboration with data-producing national infrastructures)
- **Structure DMP** for platforms of the national infrastructures in life sciences
- **Metark**: software tool facilitating data submission to international repositories (ongoing development)
- **FAIRchecker**: tool to quantify the compliance of Web resources to FAIR criteria

<https://fair-checker.france-bioinformatique.fr/>



FAIR-Checker

Improve the FAIRness of your web resources

Welcome

FAIR-Checker is a tool aimed at assessing FAIR principles and empowering data provider to enhance the quality of their digital resources.

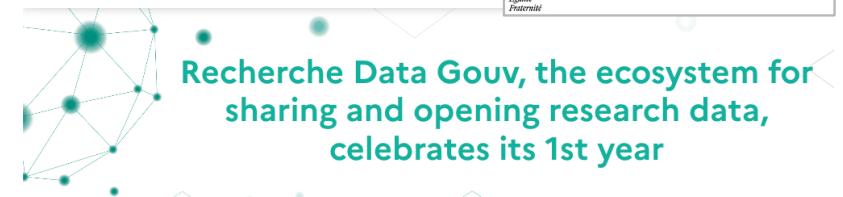
Data providers and consumers can **check** how FAIR are web resources. Developers can explore and **inspect** metadata exposed in web resources.



Check ✓

Inspect 🔍

Journal of Biomedical Semantics, 14: 7 (2023)



In 2022, IFB was designated as a thematic reference center for Life and Health data.

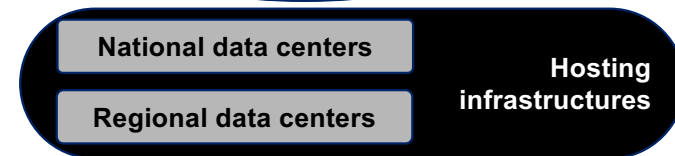
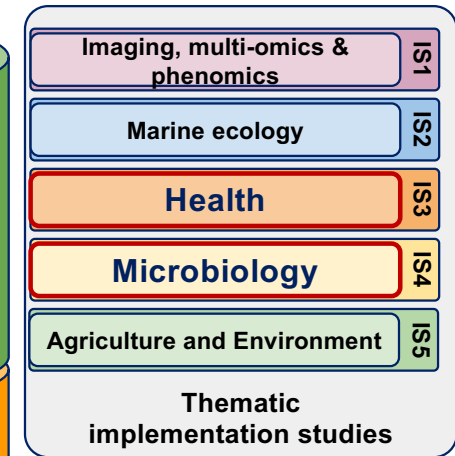
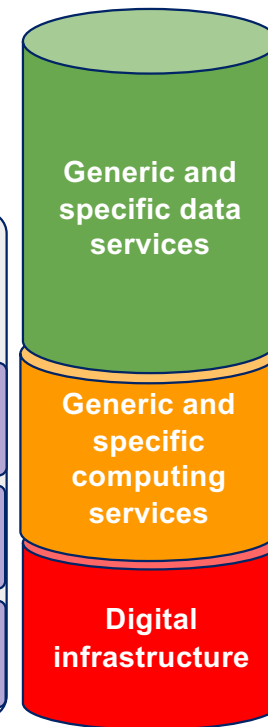
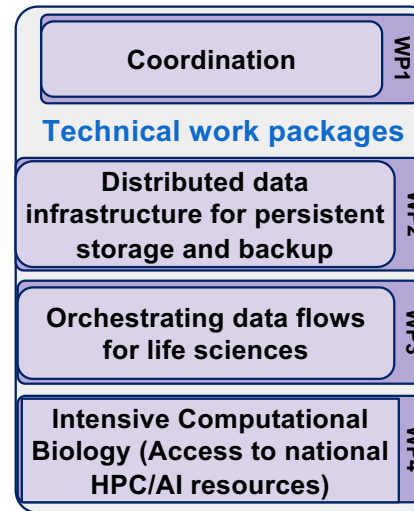
Mutualised Digital Spaces for Life Sciences and Health [PIA3 ongoing]

Main axes of the MUDIS4LS project

- Enlarging the national coverage of the NNCR
- Anchoring in national and regional data centers
- Subcontracting with some regional mesocenters
- **Data securing** during the projects (mid-term storage)
- **Orchestrating data flows** along the whole life cycle
 - maDMP-based management
 - data brokering
 - linking to electronic lab book
- Data **FAIR**ification
- **5 thematic implementation studies** ensuring relevance to address user needs

Organisation

- 39 partner teams
- 170 persons, 3000 PM (2.5 person•centuries)
- 14 partner organisms
- 4 national + 7 regional data centers, 4 partner mesocenters
- Allocated funding: **16,5 M€**



J. van Helden



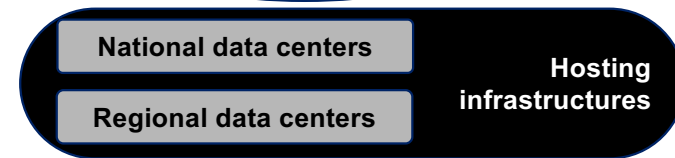
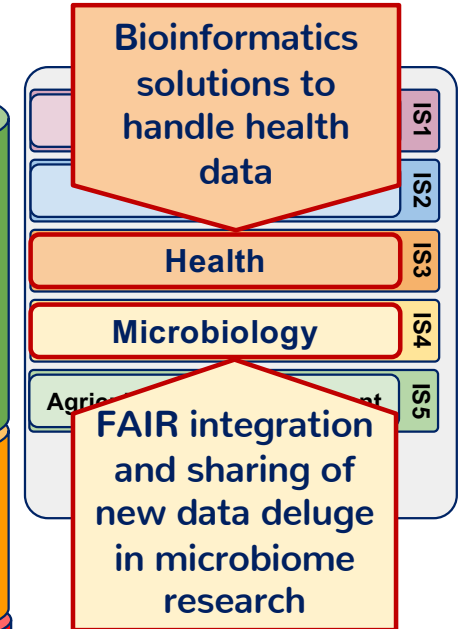
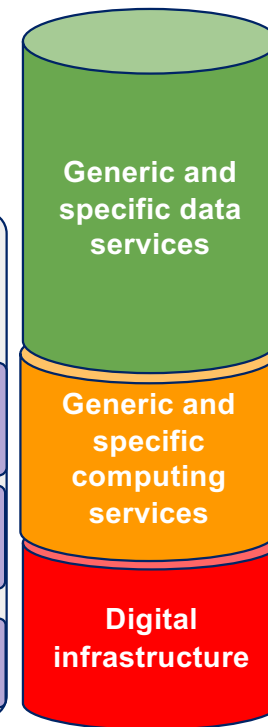
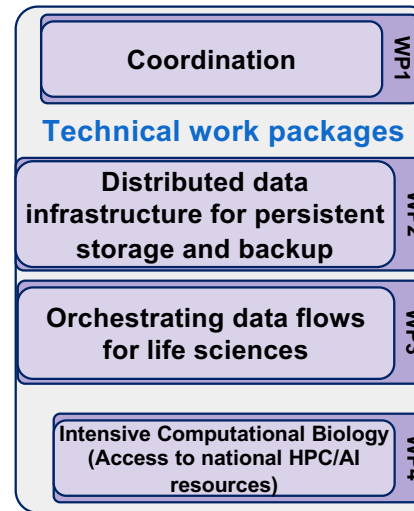
Mutualised Digital Spaces for Life Sciences and Health [PIA3 ongoing]

Main axes of the MUDIS4LS project

- Enlarging the national coverage of the NNCR
- Anchoring in national and regional data centers
- Subcontracting with some regional mesocenters
- **Data securing** during the projects (mid-term storage)
- **Orchestrating data flows** along the whole life cycle
 - maDMP-based management
 - data brokering
 - linking to electronic lab book
- Data **FAIR**ification
- **5 thematic implementation studies** ensuring relevance to address user needs

Organisation

- 39 partner teams
- 170 persons, 3000 PM (2.5 person•centuries)
- 14 partner organisms
- 4 national + 7 regional data centers, 4 partner mesocenters
- Allocated funding: **16,5 M€**



J. van Helden



French Research Priority Plan on AntiMicrobial Resistance (AMR)



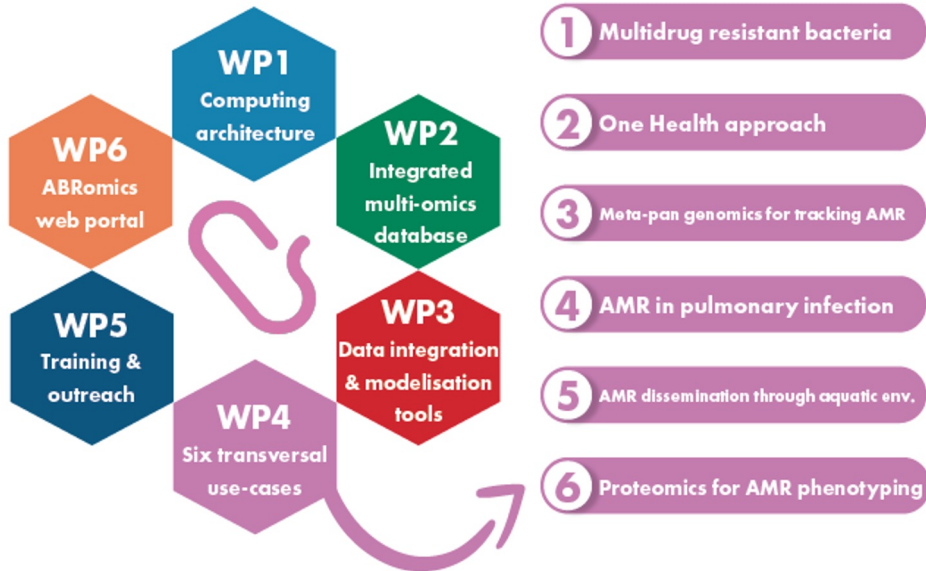
ILS SONT PRÉCIEUX, UTILISONS-LES MIEUX.

A multi-omics data platform, 2 M€ for 4 years (2021-2025)

Coordination : C. Médigue (IFB) and P. Glaser (IP)



Work packages of ABRomics



UC 3, 4 & 5 : metagenomics data analysis pipelines, visualisation and exploration

<https://www.abromics.fr>

The screenshot shows the ABRomics website homepage. The header includes the ABRomics logo and navigation links: About, ABRomics Platform, Use-Cases, Activities, Publications, and a search icon. The main heading reads "The French national multiomics platform for Antibiotic Resistance research and surveillance". Below this, a sub-heading states: "ABRomics is an online community-driven platform to scale up and improve surveillance and research on antibiotic resistance from a One Health perspective." A purple banner indicates "Available at the beginning of 2024". Three main action buttons are displayed: "Explore ABRomics-DB" (with a magnifying glass icon), "Upload Data & Analysis" (with a download icon), and "Data Integration & Modelling" (with a gear icon). The background features a microscopic image of bacteria.

PEPR SAMS (Food Systems, Microbiomes and Health) : Cloud4SAMS



Secured computing spaces for the data access and analysis of the PEPR SAMS, 1,86 M€ for 4 years (2024-2028)

Coordination : N. Pons (MGP) and C. Médigue (IFB)



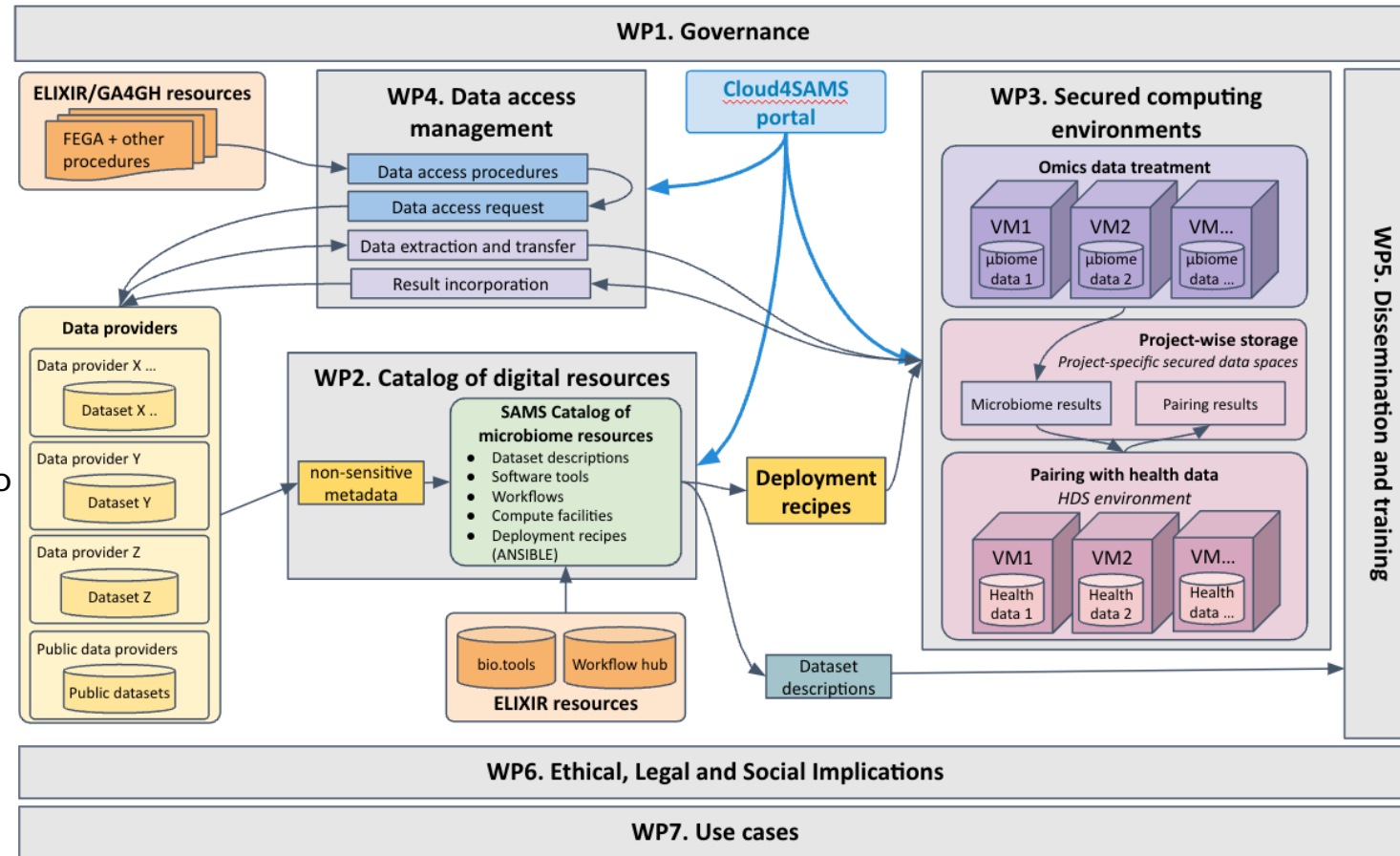
Federation of computing resources indexed in the Cloud4SAMS catalog (WP2) :

- datasets produced by microbiome projects
- software tools and WFs
- computing and storage platforms suitable for processing microbiome data and matching them with health data

=> used to define **deployment recipes** describing the procedures to instantiate a **virtual machine in a secure cloud (WP3)**

WP4 to manage the requests to the data access of each project

WP6 to define the legal framework for data sharing and the appropriate level of cybersecurity





MICROBIOTA IN HEALTH AND DISEASE

The meeting of microbiota experts

CNRS Orléans - FRANCE - November 7th 2023

Thanks for your attention !

Any questions ?....

