Bioinformatics Cloud Services for Life Sciences

Christophe BLANCHET

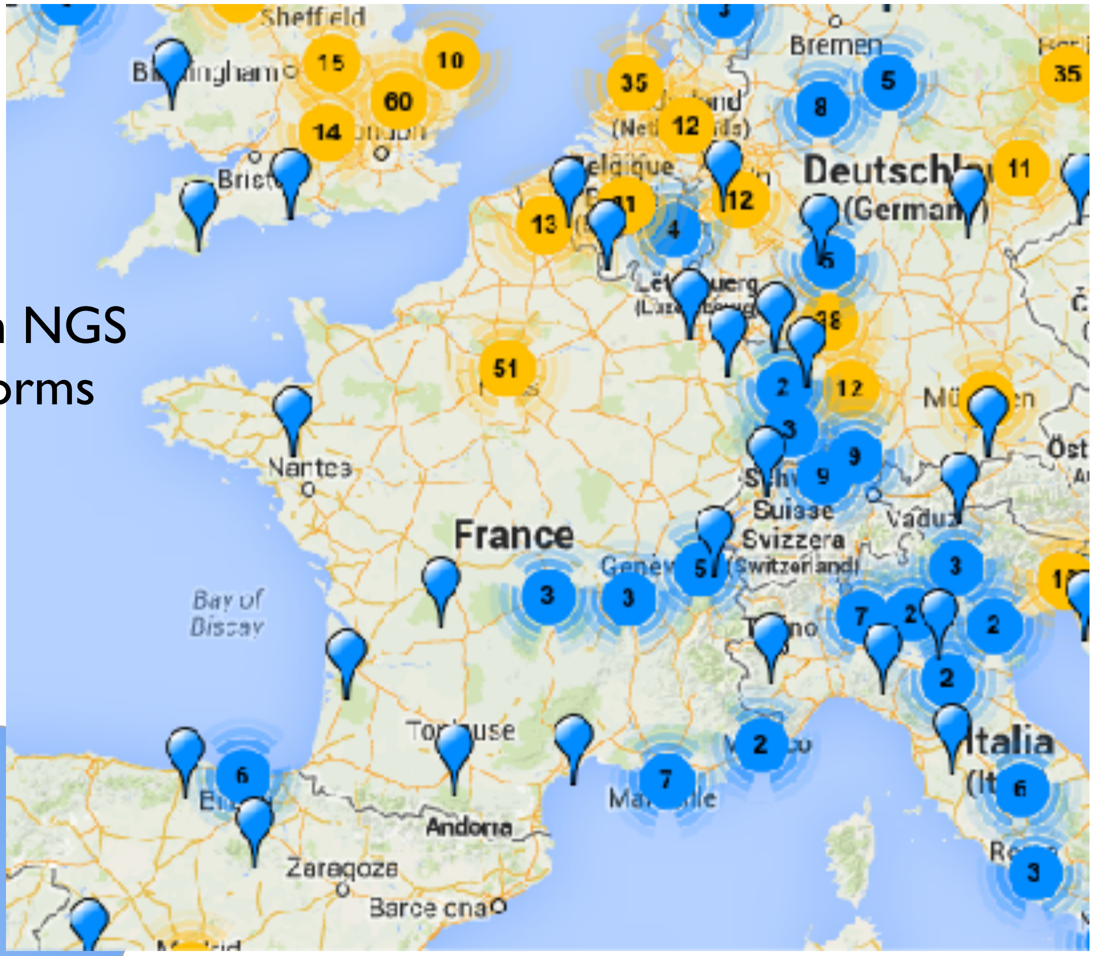
Institut Français de Bioinformatique - IFB
French Institute of Bioinformatics - ELIXIR-FR
CNRS UMS3601 - Gif-sur-Yvette - FRANCE

Nettab Conference
15 October 2015, Bari

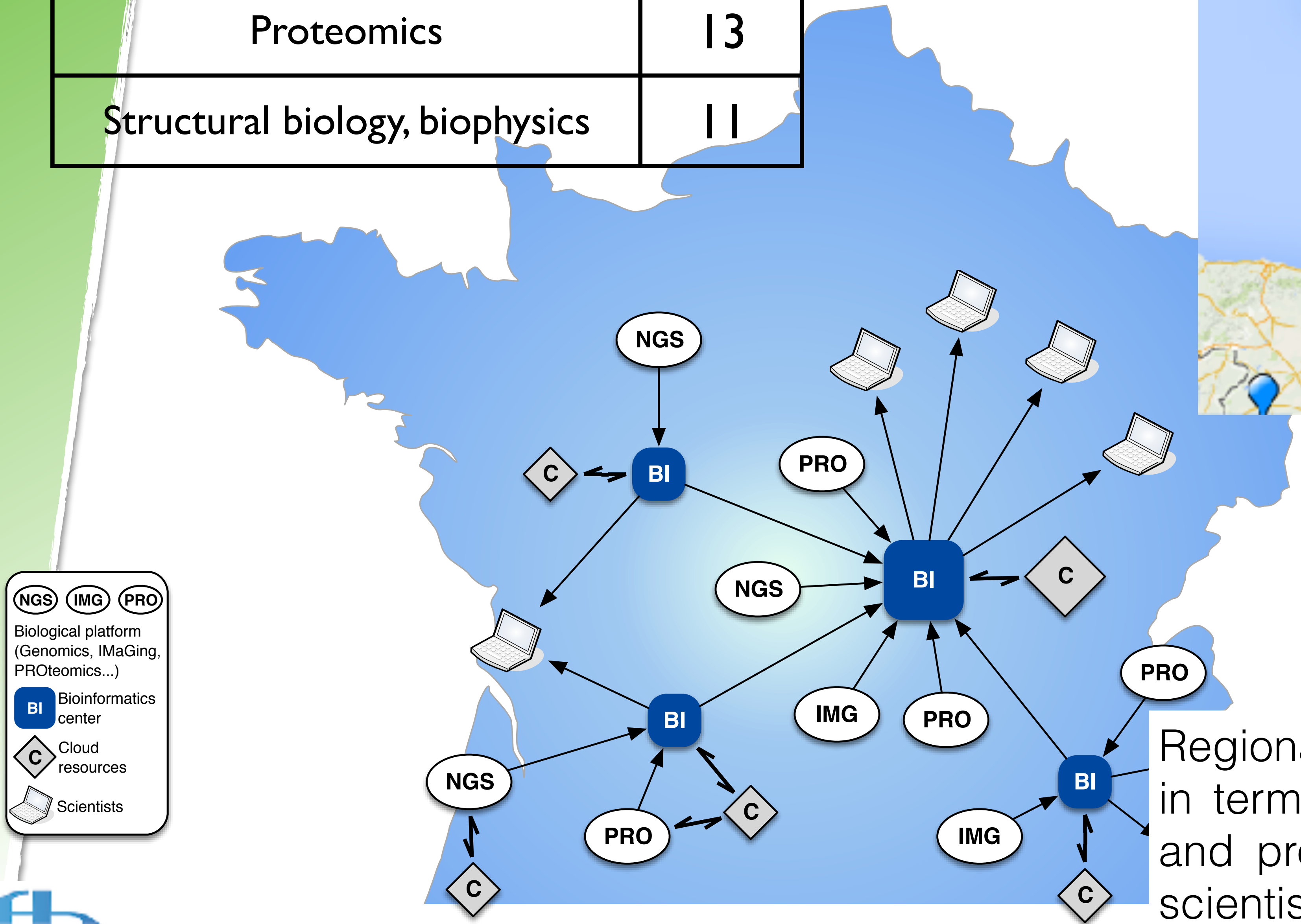
Experimental data producers in life sciences (FR)

French national platforms (GIS IBISA)	Nb
Cellular imaging	18
Genomics, transcriptomics	16
Proteomics	13
Structural biology, biophysics	11

French NGS platforms



Source: omicsmaps.com



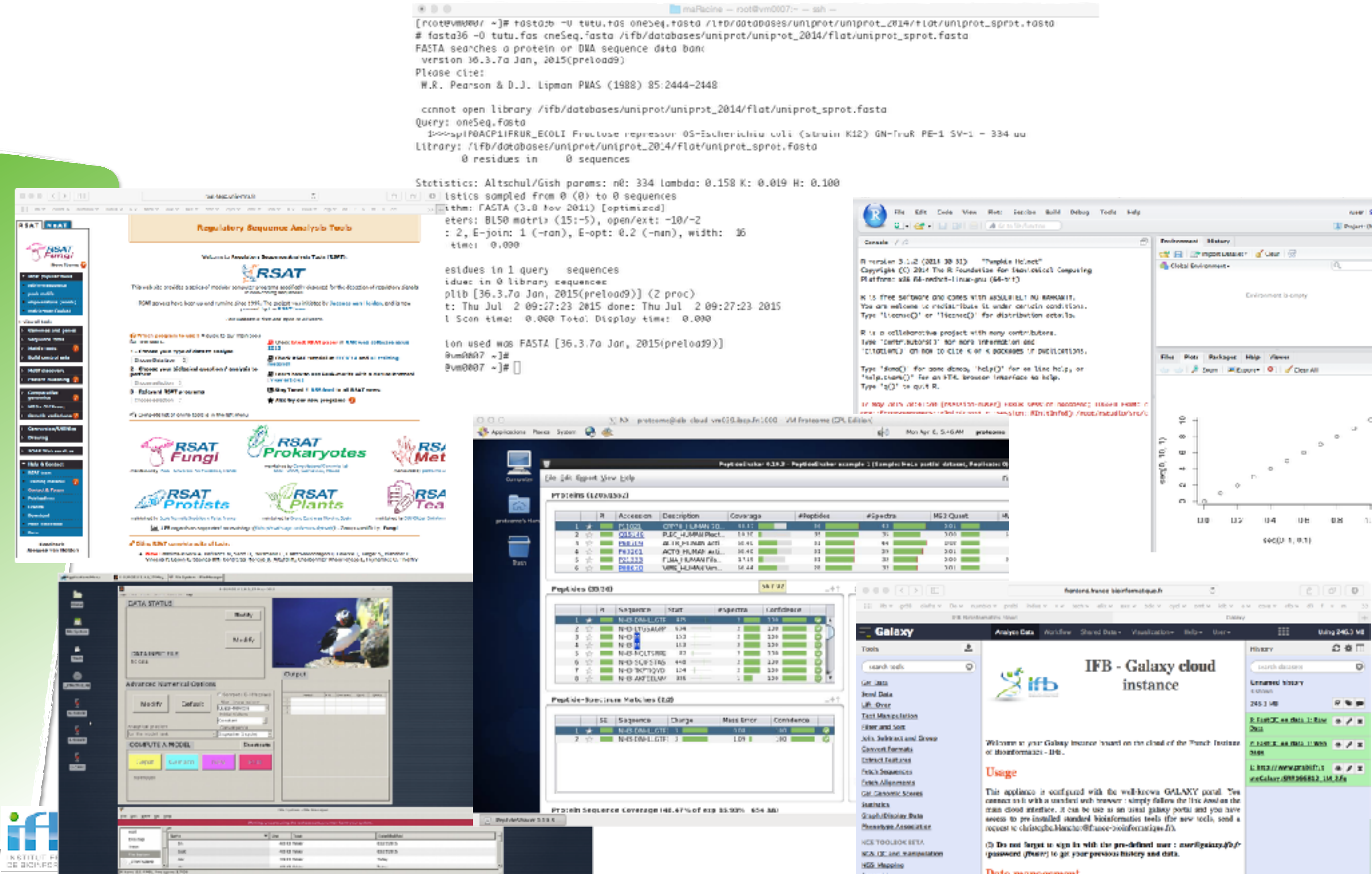
Regional centers distribute the load in terms of computing and storage, and provide better interactions with scientists

Un déluge de donnée. Blanchet C. et Collin O., 2011, Biofutur, 323: 64-67

A lot of bioinformatics tools

- ABYSS 1.3.4
- ARIA 2.3
- Bioconductor 2.11
- biomaj
- BLAST+ 2.2.27
- Blat 35
- Bowtie 0.12.8
- Bowtie2 2.0.0-beta7
- BWA 0.6.2
- BWA 0.7.10
- CAP3
- CD-HIT 4.6.1
- Clustal Omega 1.0.3
- CLUSTALW 2.1
- Cufflinks 2.0.2
- Cutadapt 1.2.1
- E-SURGE 1.9.0
- Exonerate 2.2.0
- eXpress 1.5.1
- FastA 3.6
- FastQC 0.10.1
- Galaxy portal
- GATK 2.3.4
- HMMer 3.0
- ImageJ 1.48
- khmer 1.1
- M-SURGE 1.8.5
- MEME 4.7
- MMSEQ 0.11.2a
- Mobyle
- MODAL
- MultAlin 5.4.1
- MUSCLE 3.8.31
- neo4j
- Oases 0.2.08
- OMSSA 2.1.9
- PeptideShaker 0.18.3
- phyml 3.1
- PREDATOR 2.1.2
- proline
- python 2.7
- R 2.13
- R 3.1.1
- R 3.1.2
- R-studio
- Ray 1.3
- RSAT
- samtools 0.1.18
- Samtools 1.1
- SearchGUI 1.10.4
- SeqClean
- Shiny
- Stacks
- STAR 2.4.0f1
- SuMo v1
- TGICL
- TopHat 2.0.6
- trim_galore 0.3.7
- Trinity 2.0.4
- U-CARE 2.3.2
- VCFtools 0.1.11
- Velvet 1.2.10
- X!tandem 12-10-01-1
- XPLOR-NIH 2.30
- …

Many interfaces



```
[root@vm0007 ~]# fasta36 -O tutu.fas oneSeq.fasta /ifb/databases/uniprot/uniprot_2014/flat/uniprot_sprot.fasta
# fasta36 -O tutu.fas oneSeq.fasta /ifb/databases/uniprot/uniprot_2014/flat/uniprot_sprot.fasta
FASTA searches a protein or DNA sequence data bank
 version 36.3.7a Jan, 2015(preload9)
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 cannot open library /ifb/databases/uniprot/uniprot_2014/flat/uniprot_sprot.fasta
Query: oneSeq.fasta
  1>>>sp|P0ACP1|FRUR_ECOLI Fructose repressor OS=Escherichia coli (strain K12) GN=fruR PE=1 SV=1 - 334 aa
Library: /ifb/databases/uniprot/uniprot_2014/flat/uniprot_sprot.fasta
        0 residues in     0 sequences

Statistics: Altschul/Gish params: n0: 334 Lambda: 0.158 K: 0.019 H: 0.100
Statistics sampled from 0 (0) to 0 sequences
Algorithm: FASTA (3.8 Nov 2011) [optimized]
Parameters: BL50 matrix (15:-5), open/ext: -10/-2
: 2, E-join: 1 (-ran), E-opt: 0.2 (-nan), width:  16
 time:  0.000

Residues in 1 query   sequences
idues in 0 library sequences
plib [36.3.7a Jan, 2015(preload9)] (2 proc)
t: Thu Jul  2 09:27:23 2015 done: Thu Jul  2 09:27:23 2015
l Scan time:  0.000 Total Display time:  0.000

ion used was FASTA [36.3.7a Jan, 2015(preload9)]
@vm0007 ~]#
@vm0007 ~]#
```

IFB - Galaxy cloud instance

Welcome to your Galaxy instance hosted on the cloud of the French Institute of Bioinformatics - IFB.

Usage

This appliance is configured with the well-known GALAXY portal. You connect to it with a standard web browser : simply follow the link Area on the main cloud interface. It can be use as an usual galaxy portal and you have access to pre-installed standard bioinformatics tools (for new tools, send a request to christophe.blanchet@france-bioinformatique.fr).

(!) Do not forget to sign in with the pre-defined user : user@galaxy.ifb.fr (password (ifbuser)) to get your previous history and data.

Data management

The French Institute of Bioinformatics and its e-infrastructure

History

Since 2004, ReNaBi is the National Network of Bioinformatics platforms with an IBiSA label (Infrastructures in Biology, Health and Agronomy)

In 2010, call of proposals “Infrastructures in Biology and Health” from the “Investments for the Future” initiative.

- ★ Project ReNaBi-IFB accepted in 2012 and endowed with 20m €

Other national infrastructures (NIs)

- ★ France Génomique : sequencing and genotyping NI
- ★ Profi : proteomics NI
- ★ Frisbi : structural biology NI
- ★ etc. (17 NIs all together) + 5 IHUs (Instituts Hospitaliers Universitaires) + 1 IRT (Institut de Recherche Technologique)

IFB - Institut Français de Bioinformatique

French distributed infrastructure for life-science information

INSTITUT FRANÇAIS DE BIOINFORMATIQUE

http://www.france-bioinformatique.fr

CNRS UMS3601. Avenue de la Terrasse, Bât 21. 91190 Gif-sur-Yvette

INRA SCIENCE & IMPACT · cnrs · Inserm · inria · cea

Mission : to make available core bioinformatics resources to the life science research community.

- To provide **support for national biology programs**
- To provide an **IT infrastructure** devoted to management and analysis of biological data
- To act as a middleman between the life science community and the bioinformatics/computer science research community

elixir

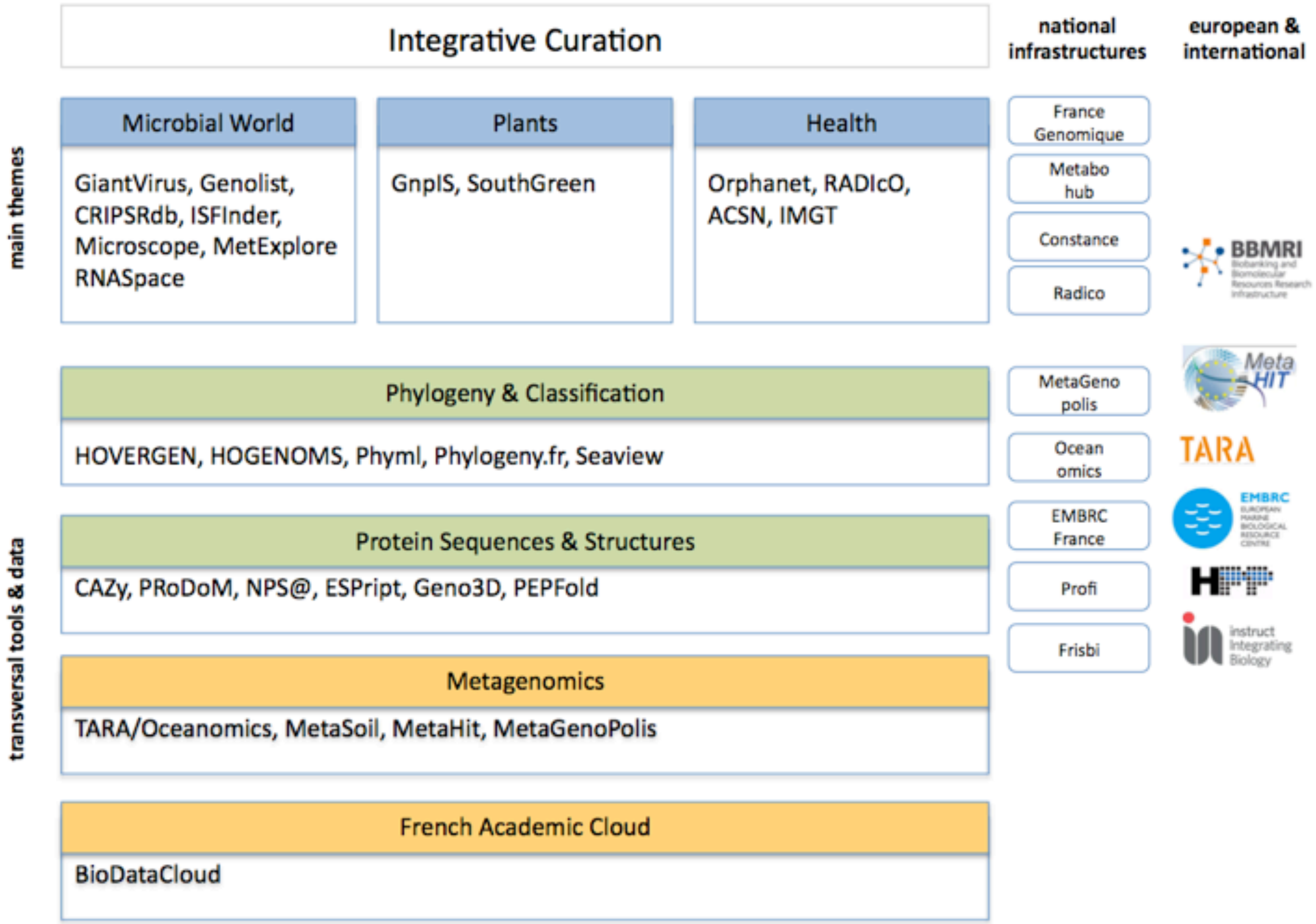
ELIXIR French Node

- The European distributed infrastructure for life-science information
- To optimize the **interactions and coordination** between the national level and ELIXIR and other ESFRI infrastructures in biomedical and environmental field,
- To promote **consistency and complementarities** between the components offered by the ELIXIR French node and those of other European nodes

Support to projects

Support to biological, biomedical or technological projects

- Large scale institutional projects and projects with other infrastructures
- Technological projects for developing services and tools
- Biology and biomedical research projects
- Services offered to industry



main themes

transversal tools & data

Integrative Curation

Microbial World: GiantVirus, Genolist, CRIPSRdb, ISFInder, Microscope, MetExplore RNASpace

Plants: GnpIS, SouthGreen

Health: Orphanet, RADIcO, ACSN, IMGT

Phylogeny & Classification: HOVERGEN, HOGENOMS, Phyml, Phylogeny.fr, Seaview

Protein Sequences & Structures: CAZy, PRoDoM, NPS@, ESPript, Geno3D, PEPFold

Metagenomics: TARA/Oceanomics, MetaSoil, MetaHit, MetaGenoPolis

French Academic Cloud: BioDataCloud

national infrastructures: France Genomique, Metabo hub, Constance, Radico, MetaGeno polis, Ocean omics, EMBRC France, Profi, Frisbi

european & international

BBMRI
Biobanking and Biomolecular Resources Research Infrastructure
MetaHIT
TARA
EMBRC
European Marine Biological Resource Centre
HPP
instruct
Integrating Biology

Call for new proposals in progress

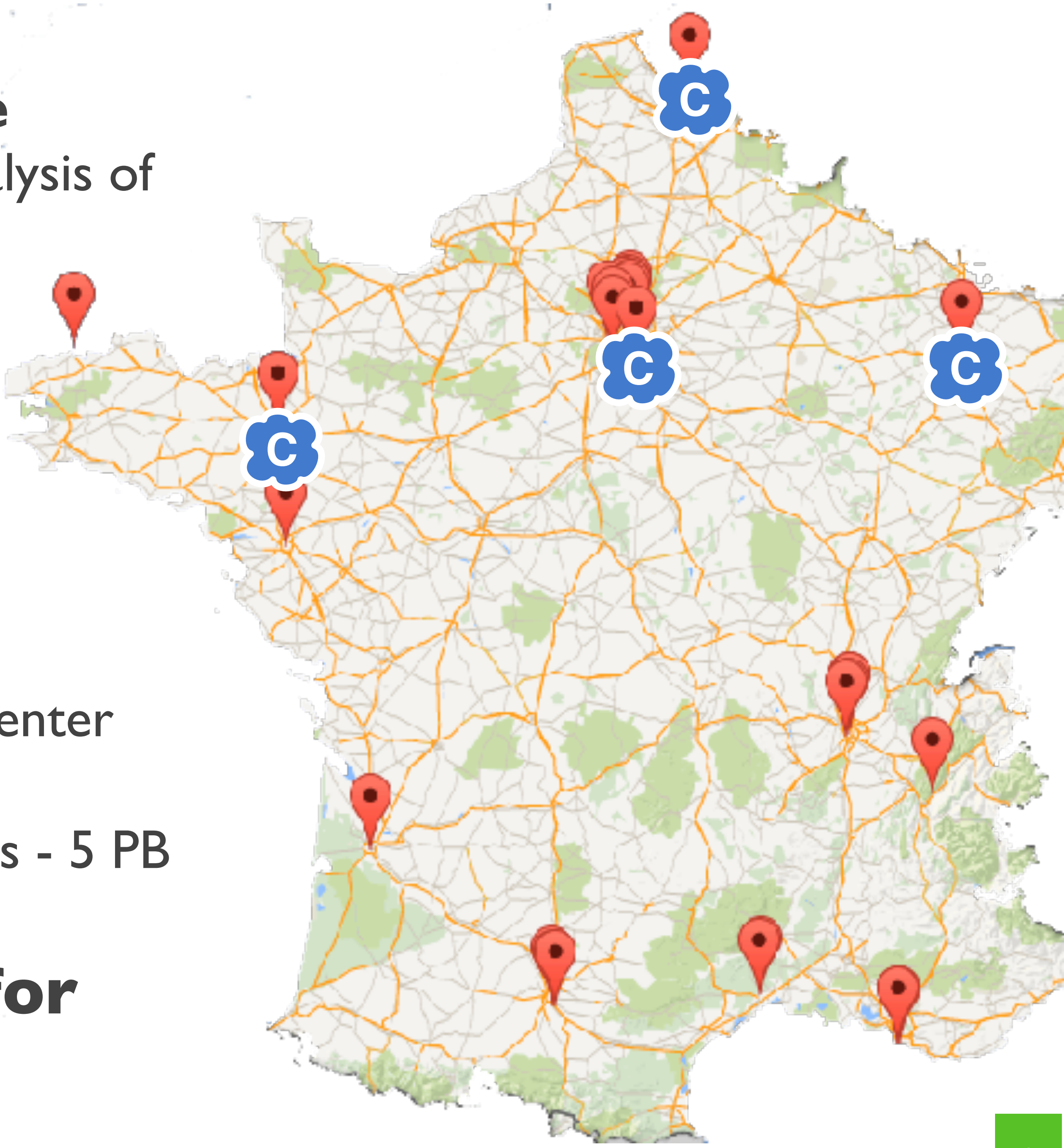
IFB e-Infrastructure

Mission : to provide core bioinformatics resources to the life science research community.

- To set up a **French IT infrastructure (cloud)** devoted to management and analysis of biological data
- To provide hardware, data collections and bioinformatics tools
- To collaborate with international infrastructure (ELIXIR)

Current resources

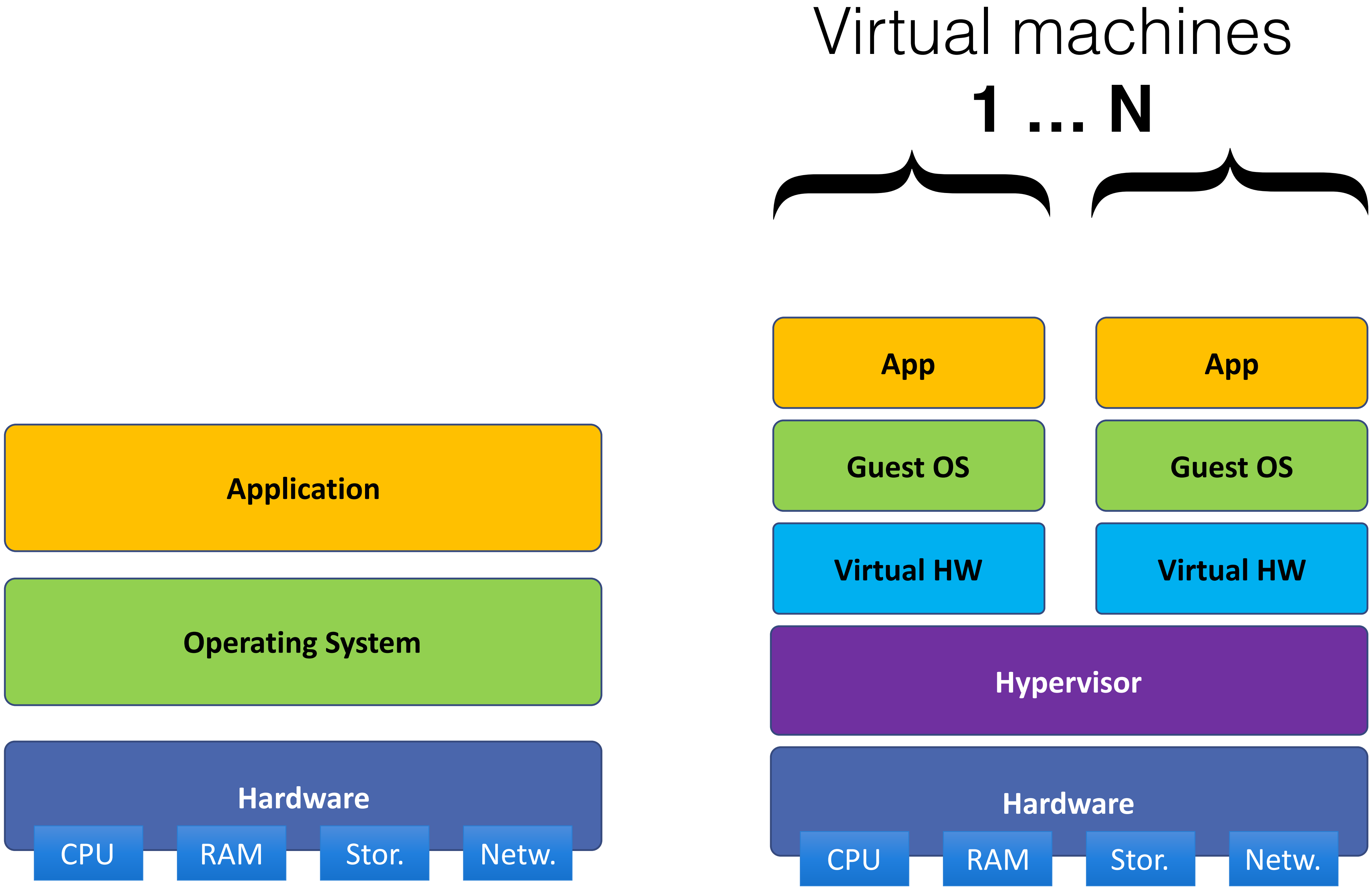
- A **national hub : IFB-core**
  IT resources hosted at CNRS IDRIS SC center
- A network of **regional centers**
  32 bioinformatics platforms - 15,000 cores - 5 PB
- 4 running clouds



➡**Create a federation of clouds for life sciences**

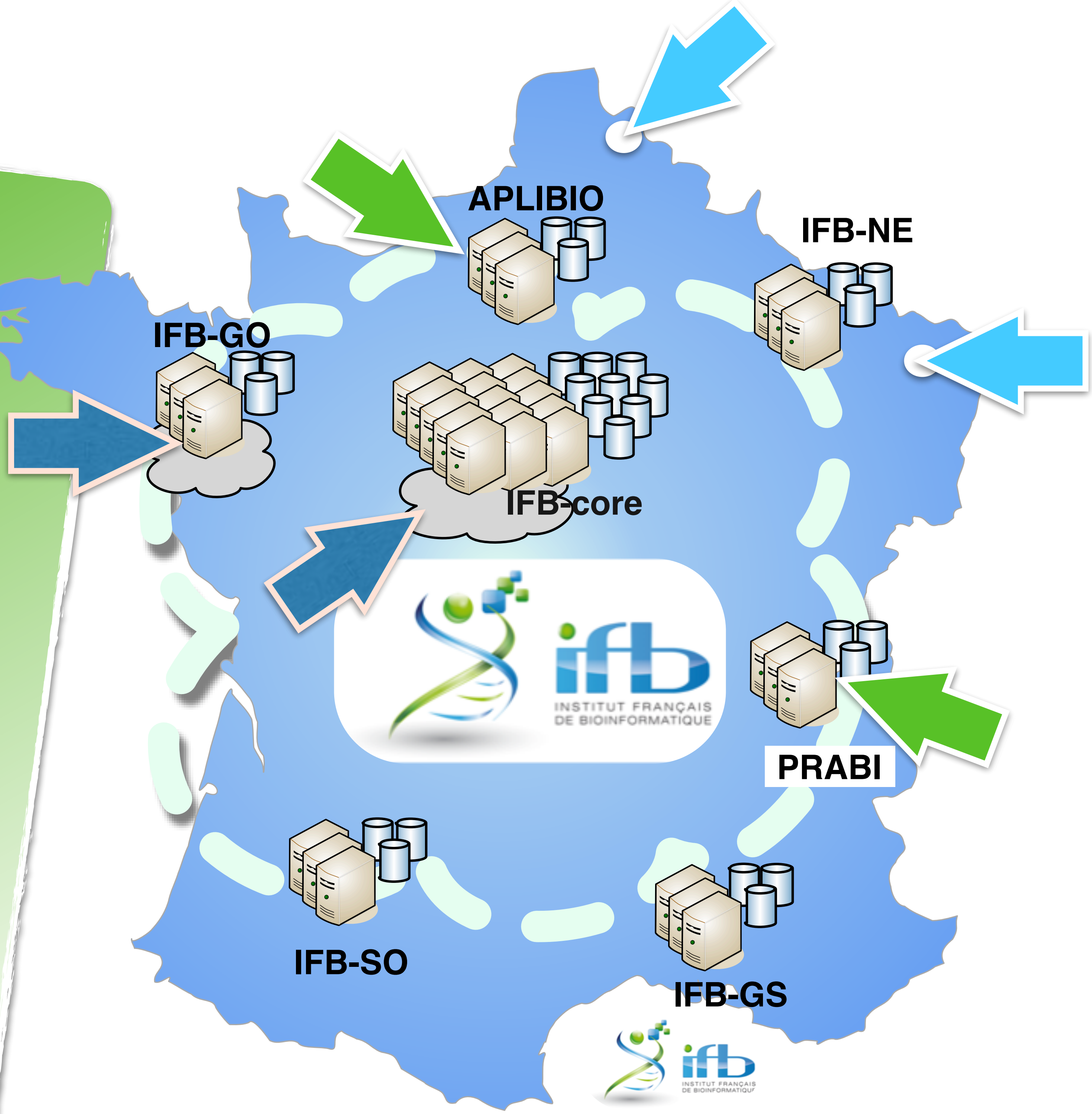
Virtualisation

With some limits…



Physical server

IFB's Cloud-s-



In IFB's premises

- IFB-core (Gif)
- GenOuest (Rennes)

In collaboration

- BiLille/Univ.Lille (Lille)
- BISTRO/IPHC-EGI fedcloud (Strasbourg)

PoC & experiments

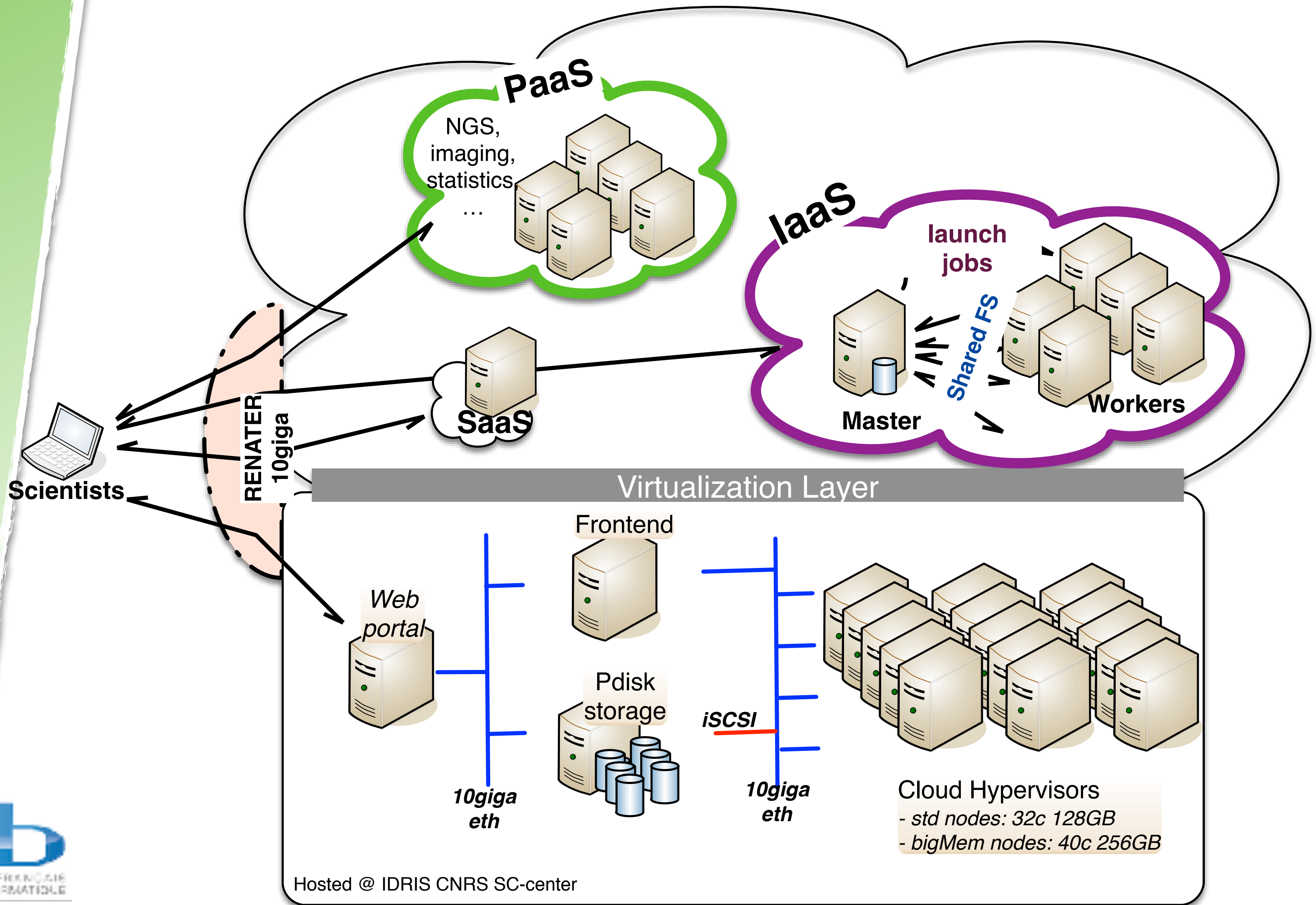
- URGI (Versailles)
- PRABI-LBBE (Lyon)

=> Towards a federation

- common identity and access management
- interoperability of VMs
- …

IFB-core's cloud

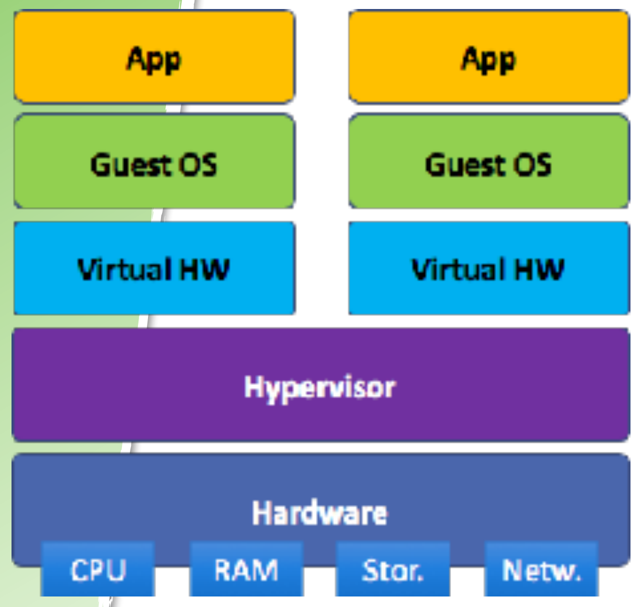
IFB-core	# Compute Cores	# TB Storage	# TB RAM	Max VM size	Technology	Location
Pilot	**200**	**50**	**2**	**40c 256GB**	**StratusLab**	**CNRS-IDRIS, Paris**
2016-S1	*3,000*	*500*	-	*?144c 3TB?*	*StratusLab*	*CNRS-IDRIS, Paris*
2017	*10,000*	*2,000*	-	*??*	*StratusLab*	*CNRS-IDRIS, Paris*

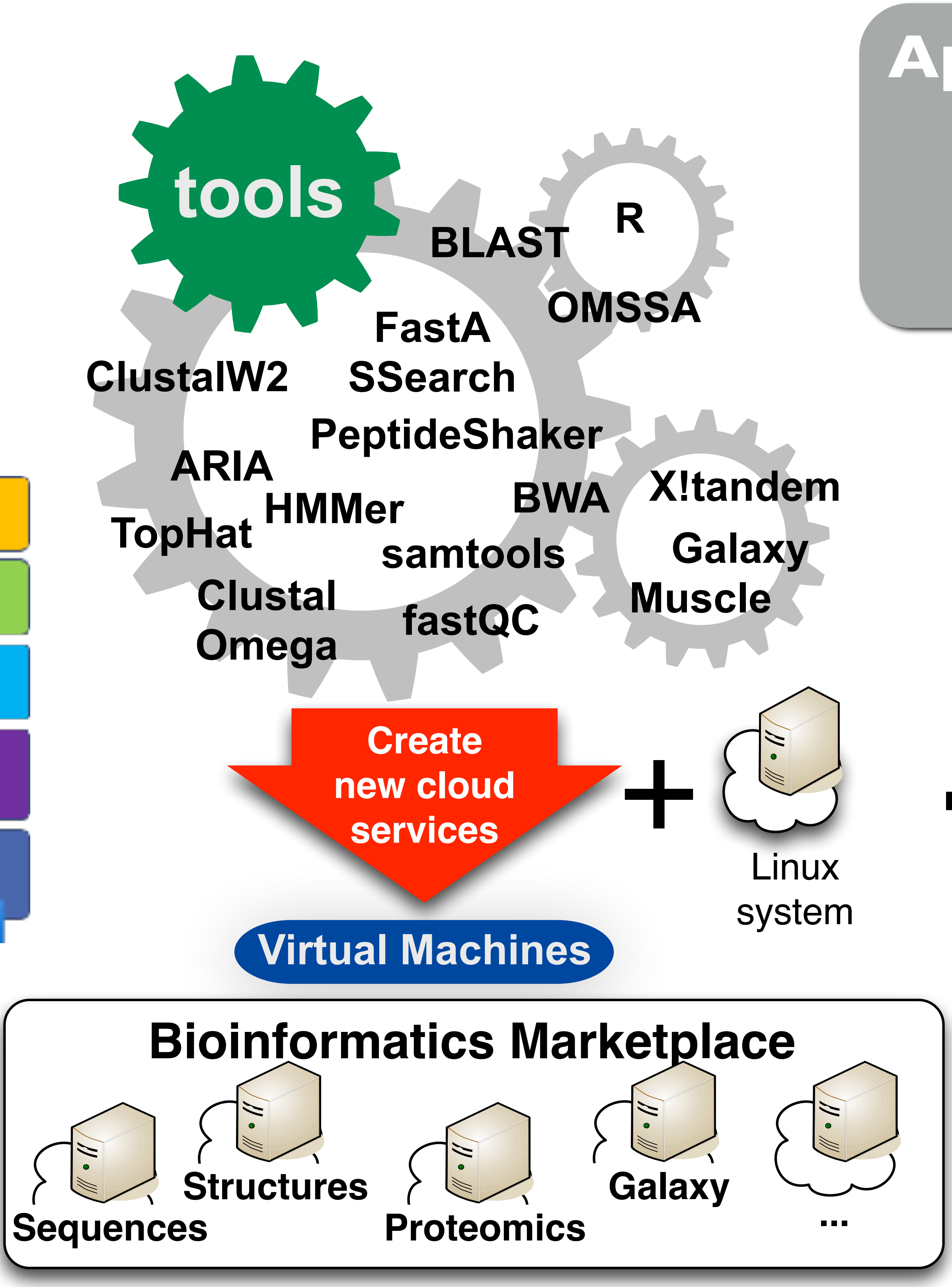


stratuslab

Provide scientists with
bioinformatics resources
- data and tools -
as cloud appliances

Create bioinformatics "appliances"





Appliance ?

- predefined virtual machine
- including tools, pipeline,recipes…
- Ready to run

+

Appliance annotation

- Title
- Description (w. controlled voc.)
  - ★ Topics
  - ★ Tools
- Contact
- Developer(s) and **maintainer(s)** !

Appliances - Topics

Bioimaging

Ecology of population

Genomics tools

Mass Spectrometry

Molecular structural analysis

Multiple Sequence Alignment

Nucleotide and Protein sequence searching

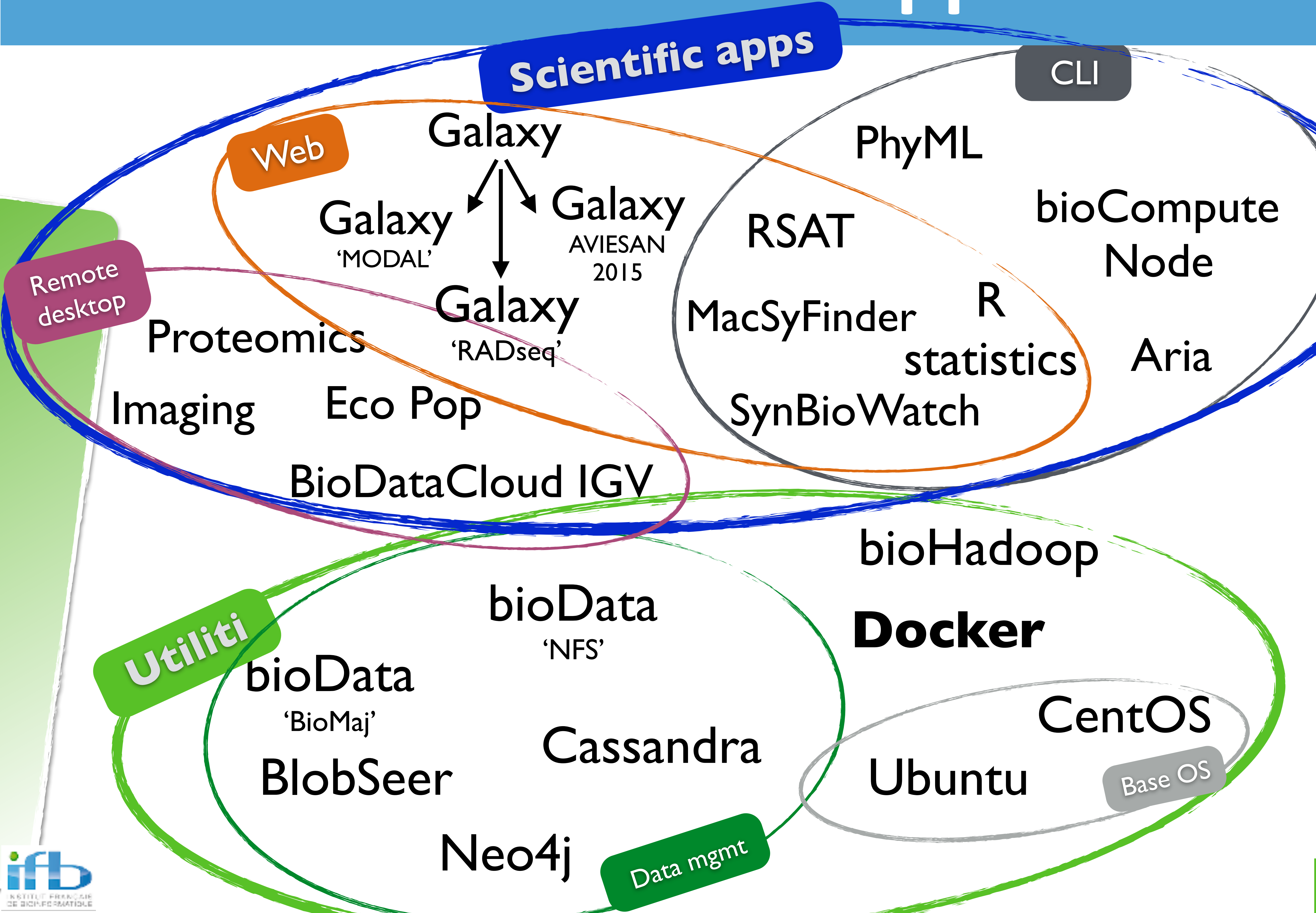
Proteomics

Public databases

Sequence analysis

…

IFB's bioinformatics appliances



Scientific apps

CLI

Web

Galaxy

Galaxy
'MODAL'

Galaxy
AVIESAN
2015

Galaxy
'RADseq'

Remote
desktop

Proteomics

Imaging

Eco Pop

BioDataCloud IGV

PhyML

RSAT

bioCompute
Node

MacSyFinder

R
statistics

Aria

SynBioWatch

bioHadoop

Docker

Utiliti

bioData
'NFS'

bioData
'BioMaj'

Cassandra

BlobSeer

Neo4j

Data mgmt

CentOS

Ubuntu

Base OS

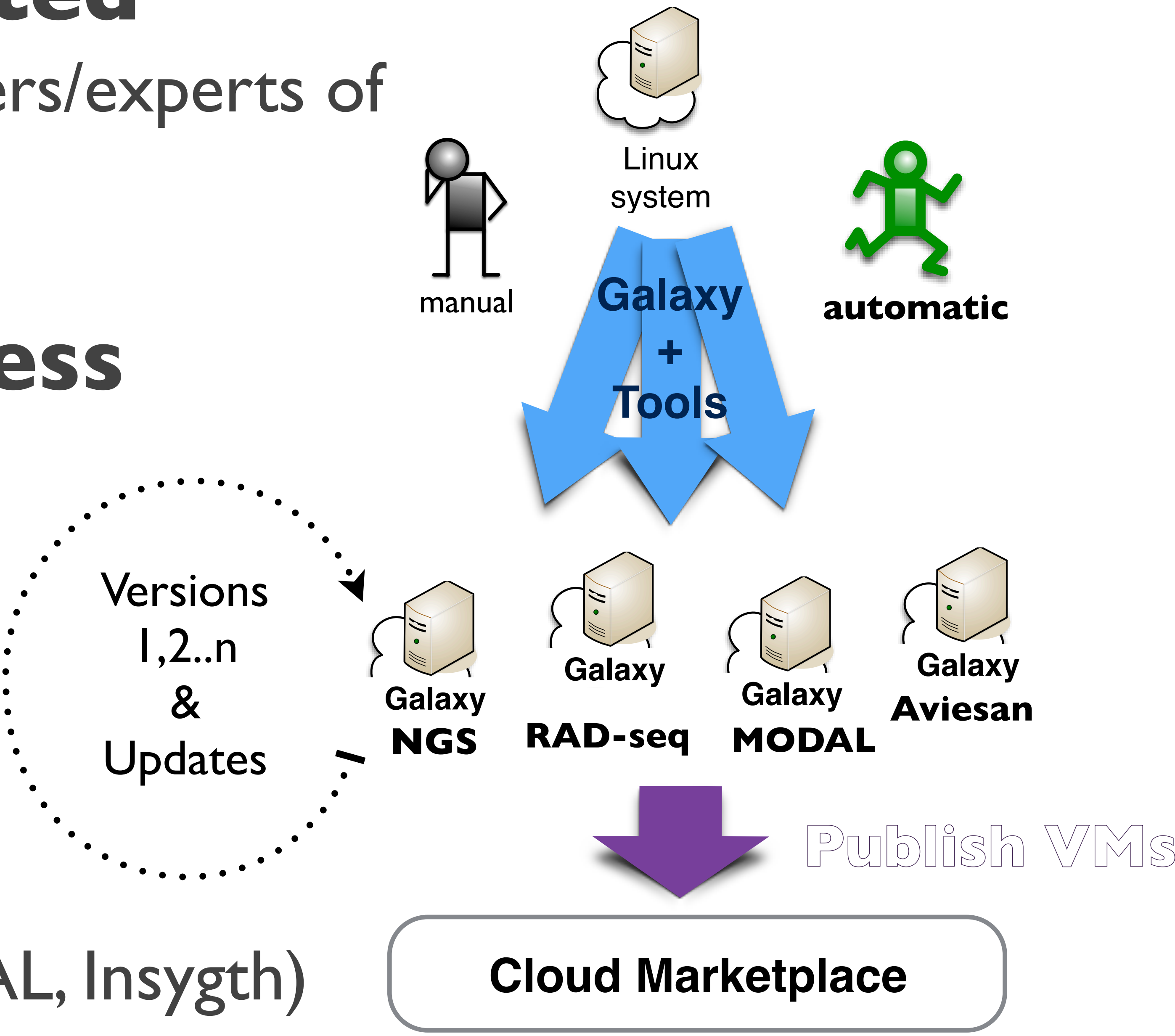
Help developers to create appliance

Appliances are created

- by the life science developers/experts of different domains

Appliances in progress

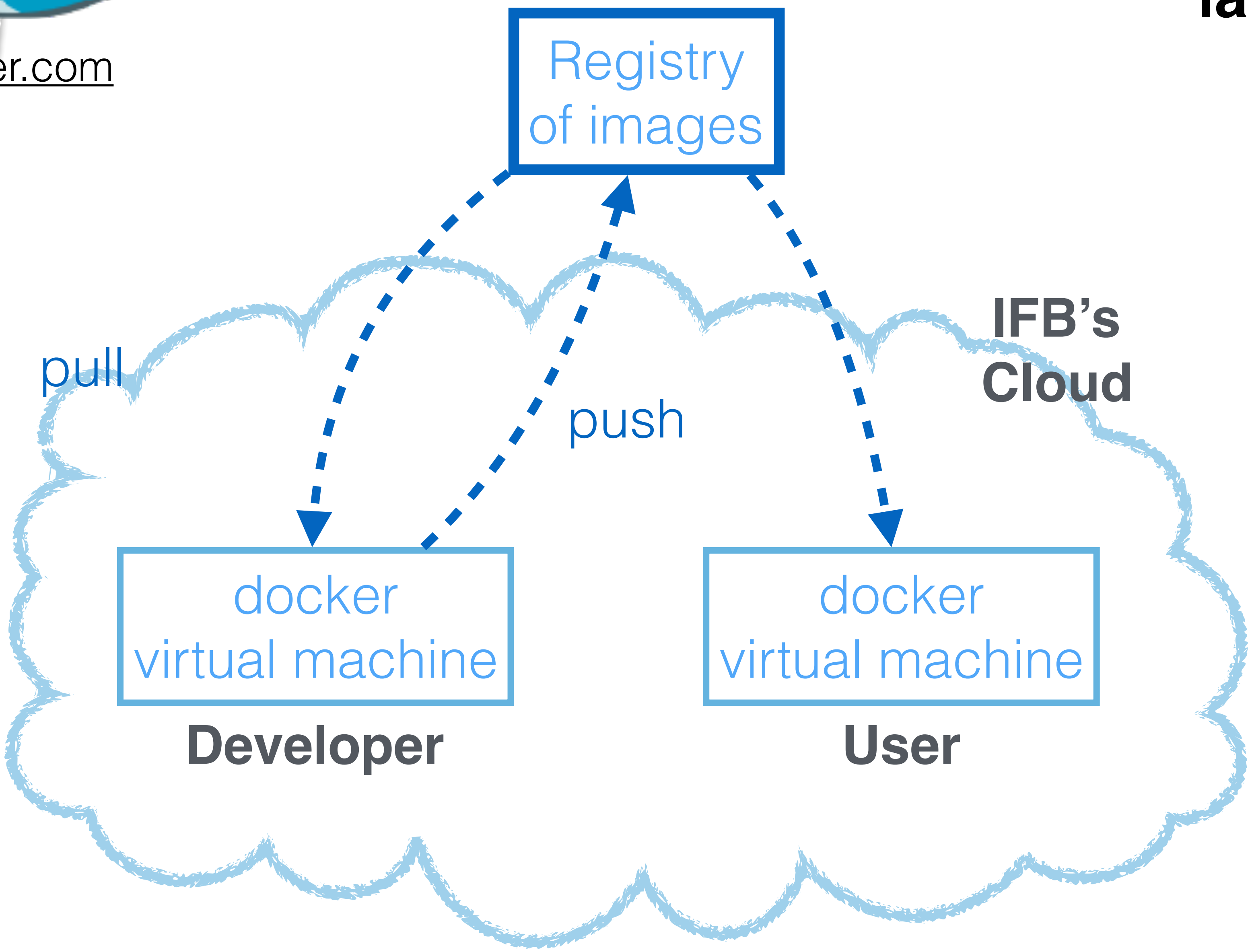
- BioDataCloud-RNAseq
- ProFi
- REPET
- TriAnnot
- Clinical NGS for cancerology (CLB & CFB)
- Bacterial genomics (AGMIAL, Insygth)
- Metagenomics (iMetAMOS)
- …



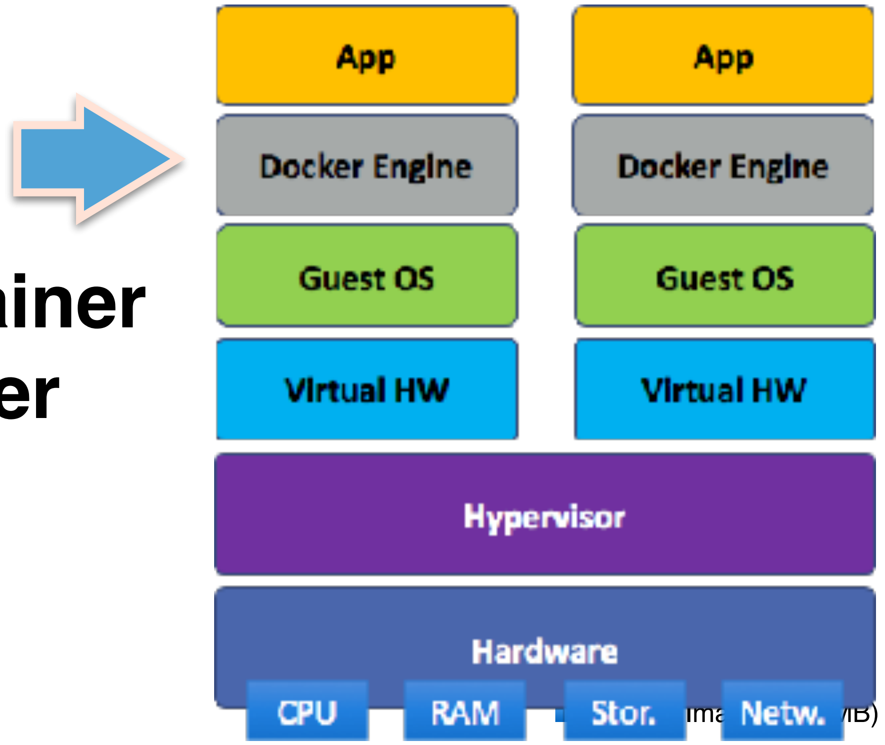
Docking bioinformatics tools

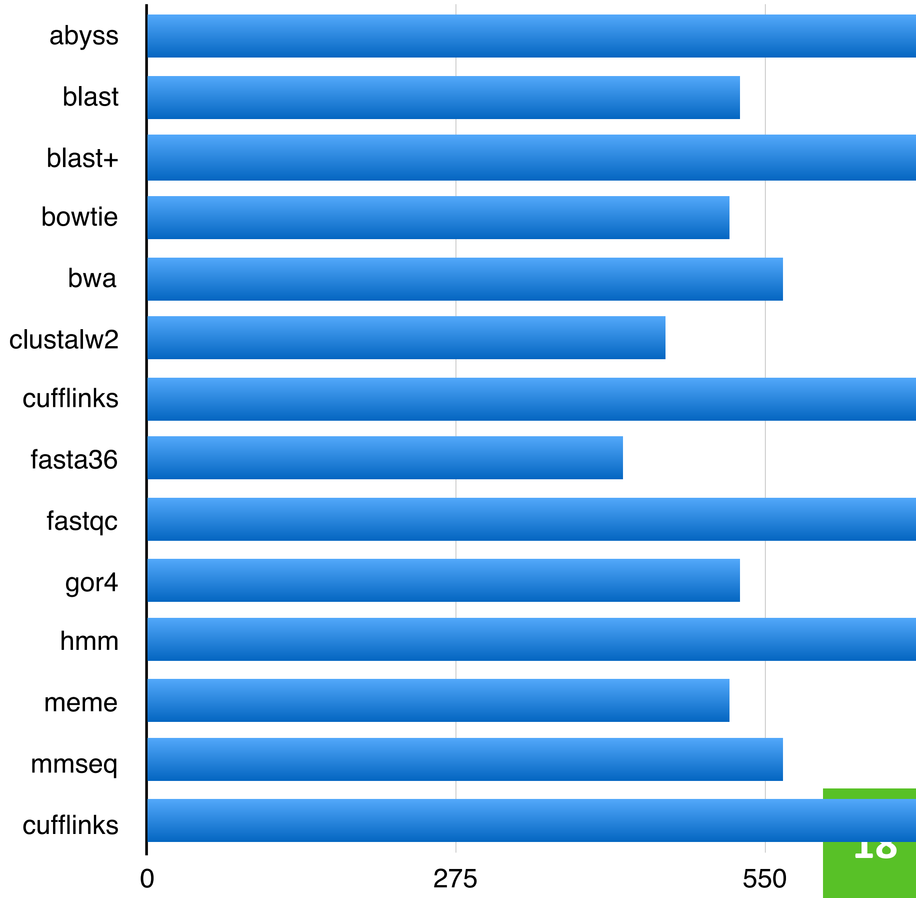
docker.com

IFB’s docker hub
(@GenOuest)



Container layer





Managing biological data

Collections of reference data

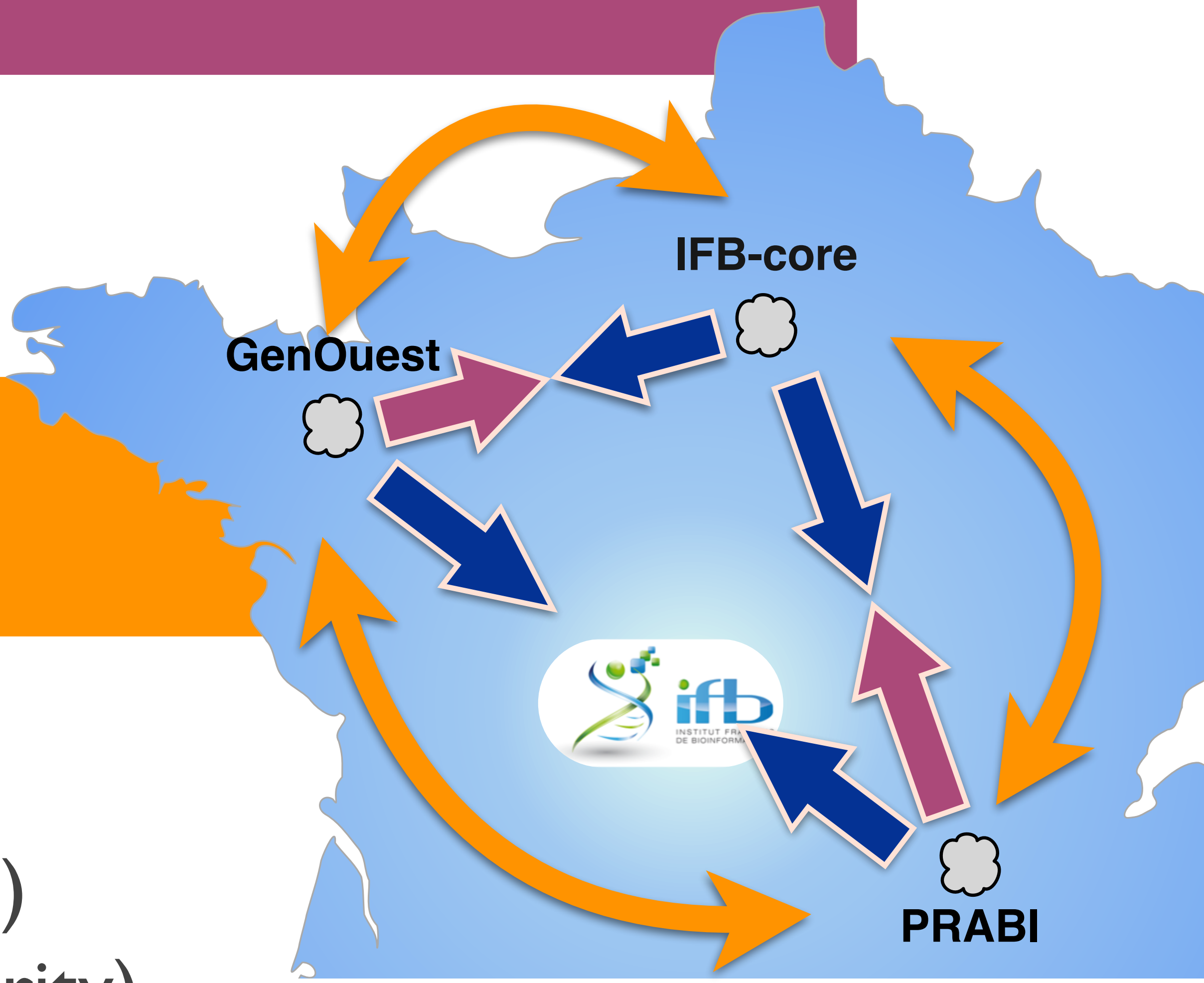
- Databases updates and index built in IFB-core (BioMAJ)
- Transfers from IFB-core to regional PFs

Experimental data: archiving (and treatment)

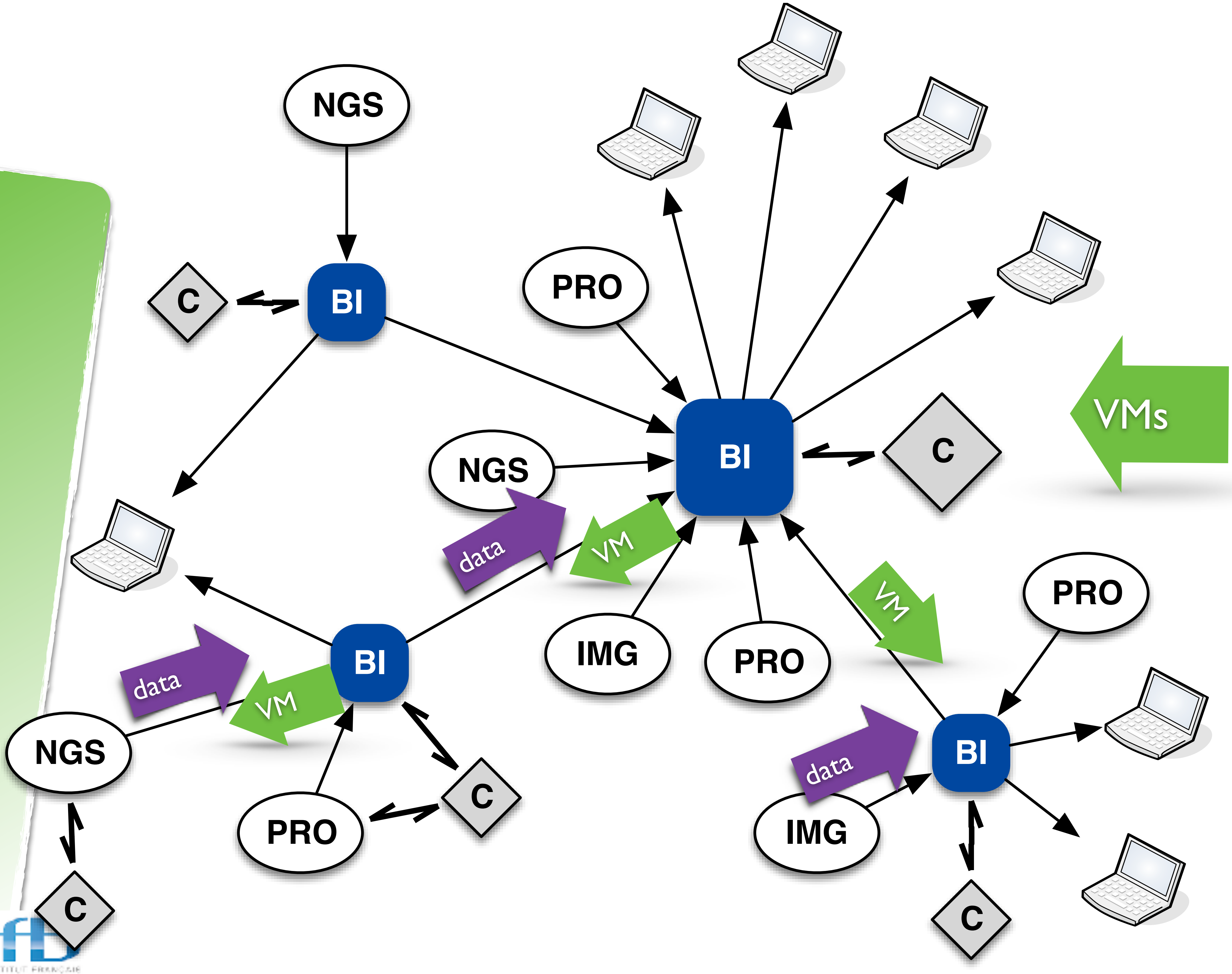
- Regional desks: deposit
- Replicate to IFB-core (iRODS?)

User data: distribution, optimisation, security

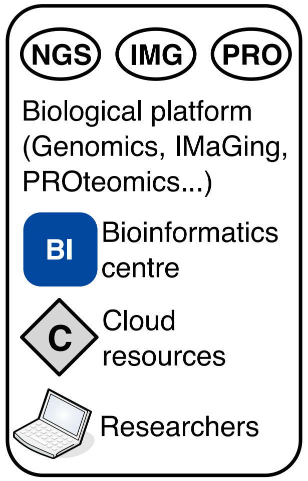
- Object storage (replication)
- Multi-site noSQL (distribution)
- Multi-site workflow (optimisation)
- Biomedical data (end-to-end security)



Move VMs rather than data



IFB's marketplace & VMs repository for life sciences



Biological platform (Genomics, IMaGing, PROteomics...)

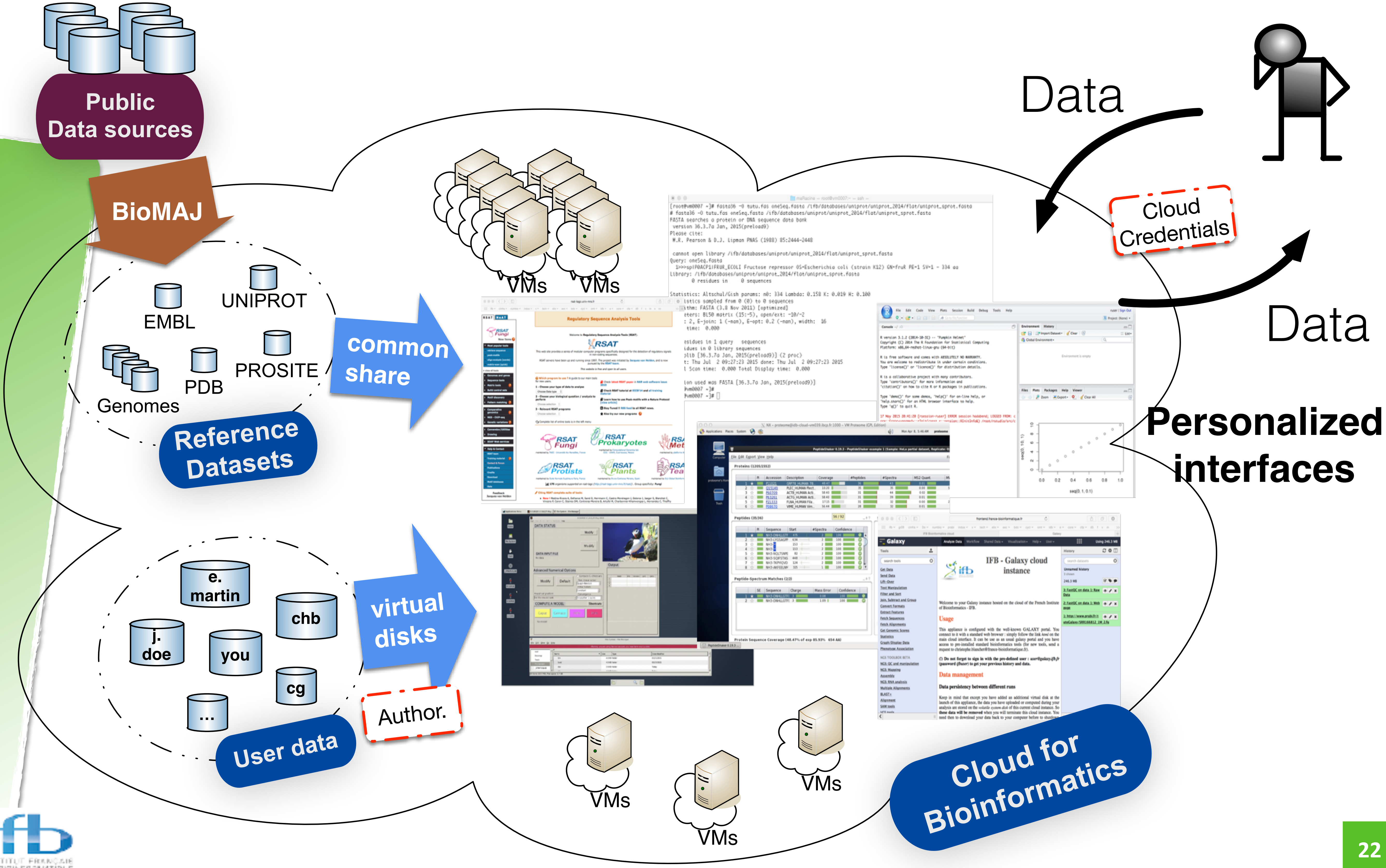
Bioinformatics centre

Cloud resources

Researchers

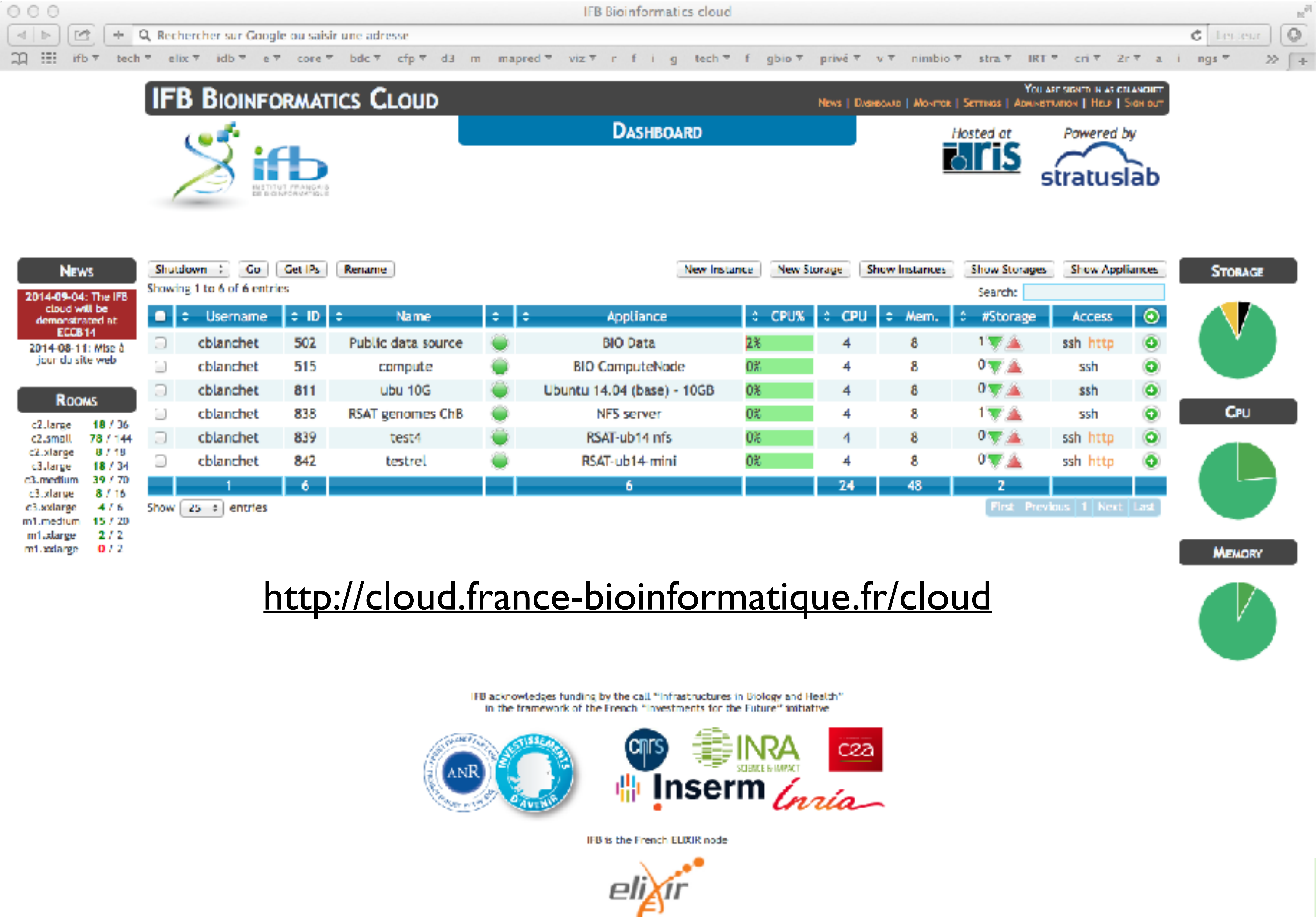
IFB's bioinformatics cloud services

A cloud for Bioinformatics



```
[root@vm0007 ~]# fasta36 -O tutu.fas oneSeq.fasta /ifb/databases/uniprot/uniprot_2014/flat/uniprot_sprot.fasta
# fasta36 -O tutu.fas oneSeq.fasta /ifb/databases/uniprot/uniprot_2014/flat/uniprot_sprot.fasta
FASTA searches a protein or DNA sequence data bank
 version 36.3.7a Jan, 2015(preload9)
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 cannot open library /ifb/databases/uniprot/uniprot_2014/flat/uniprot_sprot.fasta
Query: oneSeq.fasta
  1>>>sp|P0ACP1|FRUR_ECOLI Fructose repressor OS=Escherichia coli (strain K12) GN=fruR PE=1 SV=1 - 334 aa
Library: /ifb/databases/uniprot/uniprot_2014/flat/uniprot_sprot.fasta
        0 residues in     0 sequences

Statistics: Altschul/Gish params: n0: 334 Lambda: 0.158 K: 0.019 H: 0.100
 istics sampled from 0 (0) to 0 sequences
 ithm: FASTA (3.8 Nov 2011) [optimized]
 eters: BL50 matrix (15:-5), open/ext: -10/-2
 : 2, E-join: 1 (-nan), E-opt: 0.2 (-nan), width:  16
  time:  0.000

 esidues in 1 query   sequences
 idues in 0 library sequences
 plib [36.3.7a Jan, 2015(preload9)] (2 proc)
 t: Thu Jul  2 09:27:23 2015 done: Thu Jul  2 09:27:23 2015
 l Scan time:  0.000 Total Display time:  0.000

 ion used was FASTA [36.3.7a Jan, 2015(preload9)]
 @vm0007 ~]#
 @vm0007 ~]#
```

A cloud driven through a web dashboard


IFB Bioinformatics cloud
Rechercher sur Google ou saisir une adresse
ifb tech elix idb e core bdc cfp d3 m mapred viz r f i g tech f gbio privé v nimbio stra IRT cri 2r a i ngs
IFB Bioinformatics Cloud
You are signed in as cblanchet
News | Dashboard | Monitor | Settings | Administration | Help | Sign out
Dashboard
Hosted at
Powered by
stratuslab
Shutdown
Go
Get IPs
Rename
New Instance
New Storage
Show Instances
Show Storages
Show Appliances
Search:
Storage
Cpu
Memory
First
Previous
1
Next
Last

News

2014-09-04: The IFB cloud will be demonstrated at ECCB14

2014-08-11: Mise à jour du site web

Rooms

c2.large	18 / 36
c2.small	78 / 144
c2.xlarge	8 / 18
c3.large	18 / 34
c3.medium	39 / 70
c3.xlarge	8 / 16
c3.xxlarge	4 / 6
m1.medium	15 / 20
m1.xlarge	2 / 2
m1.xxlarge	0 / 2

Showing 1 to 6 of 6 entries

□	Username	ID	Name		Appliance	CPU%	CPU	Mem.	#Storage	Access	
□	cblanchet	502	Public data source	●	BIO Data	2%	4	8	1	ssh http	
□	cblanchet	515	compute	●	BIO ComputeNode	0%	4	8	0	ssh	
□	cblanchet	811	ubu 10G	●	Ubuntu 14.04 (base) - 10GB	0%	4	8	0	ssh	
□	cblanchet	838	RSAT genomes ChB	●	NFS server	0%	4	8	1	ssh	
□	cblanchet	839	test4	●	RSAT-ub14 nfs	0%	4	8	0	ssh http	
□	cblanchet	842	testrel	●	RSAT-ub14-mini	0%	4	8	0	ssh http	
	1	6			6		24	48	2		

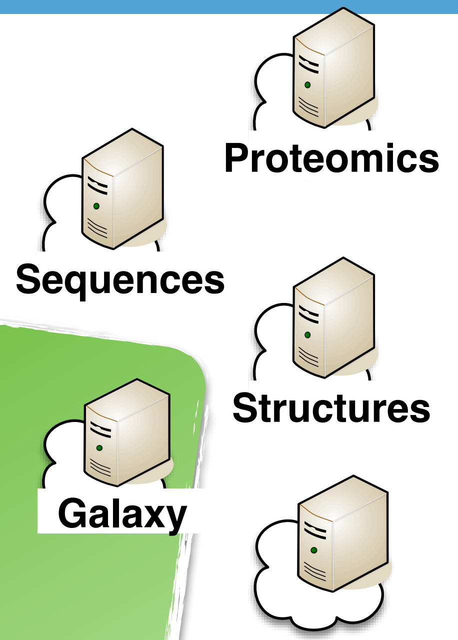
Show 25 entries

http://cloud.france-bioinformatique.fr/cloud

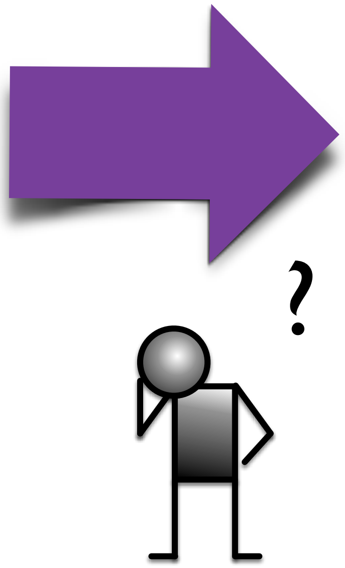
IFB acknowledges funding by the call "Infrastructures in Biology and Health"
in the framework of the French "Investments for the Future" initiative

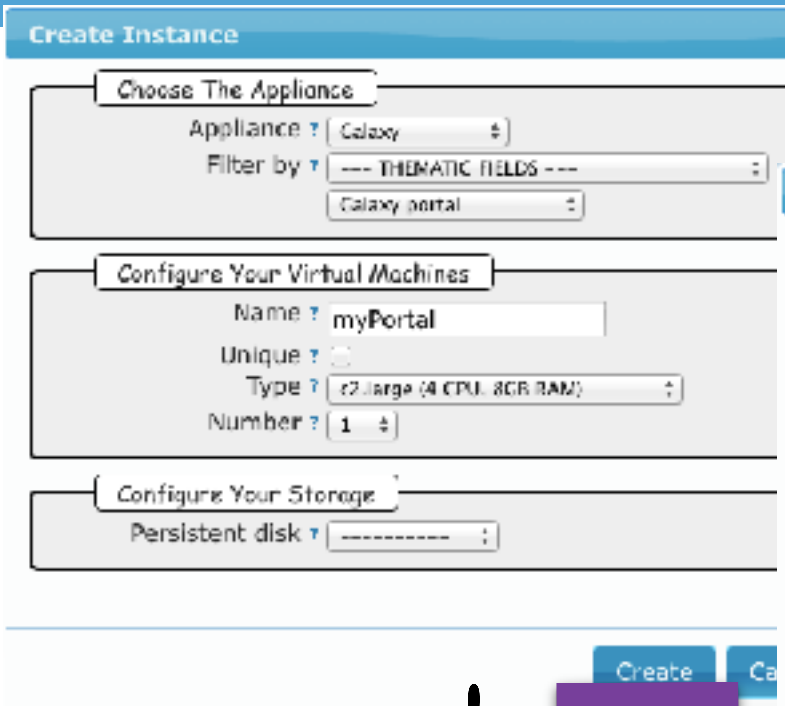
IFB is the French ELIXIR node

Browse the marketplace and run an App !



IFB's… bioinformatics marketplace




Create Instance

Choose The Appliance

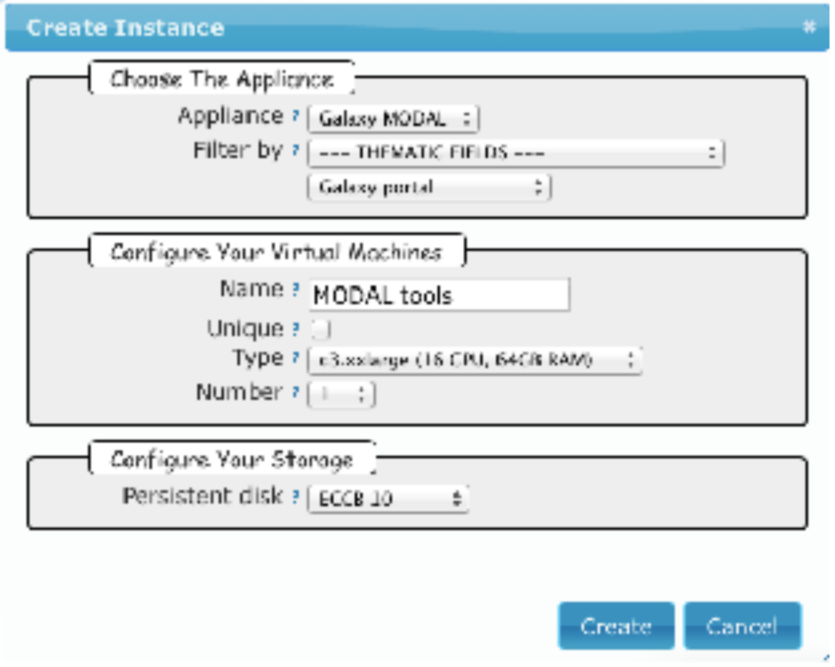
- Appliance ? Galaxy
- Filter by ? --- THEMATIC FIELDS ---
- Galaxy portal

Configure Your Virtual Machines

- Name ? myPortal
- Unique ? □
- Type ? c2.large (4 CPU, 8GB RAM)
- Number ? 1

Configure Your Storage

- Persistent disk ? ----------


Create Instance

Choose The Appliance

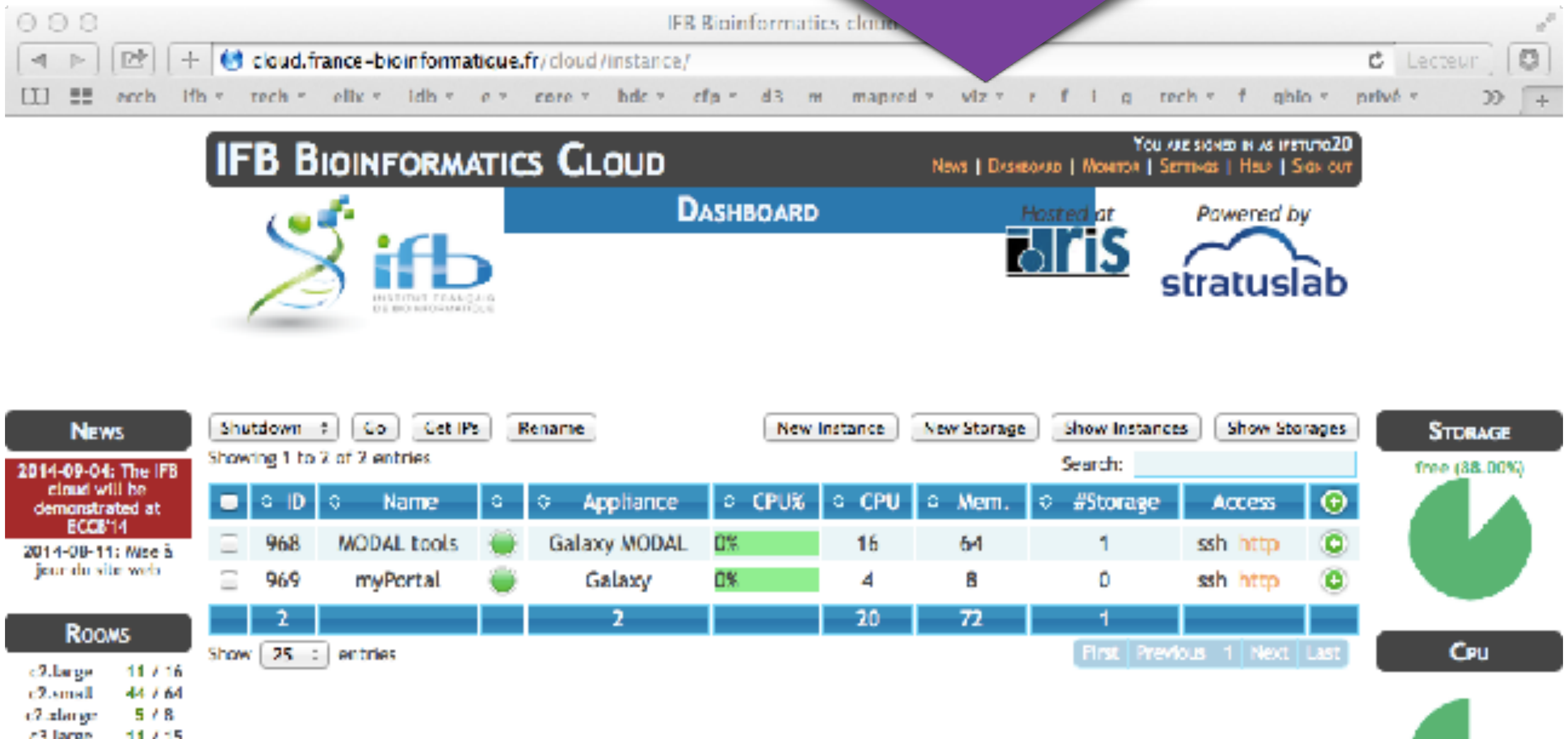
- Appliance ? Galaxy MODAL
- Filter by ? --- THEMATIC FIELDS ---
- Galaxy portal

Configure Your Virtual Machines

- Name ? MODAL tools
- Unique ? □
- Type ? c3.xxlarge (16 CPU, 64GB RAM)
- Number ? 1

Configure Your Storage

- Persistent disk ? ECCB 10


IFB Bioinformatics cloud

cloud.france-bioinformatique.fr/cloud/instance/

Lecteur
arch ifb tech ellx ldh e core hdc cfa d3 m mapred viz r f l g tech f gbio privé

IFB Bioinformatics Cloud

You are signed in as ifbtuto20

News | Dashboard | Monitor | Settings | Help | Sign out
Dashboard
Hosted at idris
Powered by stratuslab
News

2014-09-04: The IFB cloud will be demonstrated at ECCB'14

2014-08-11: Mise à jour du site web

Shutdown
Go
Get IPs
Rename
New Instance
New Storage
Show Instances
Show Storages

Showing 1 to 2 of 2 entries

Search:

□	ID	Name		Appliance	CPU%	CPU	Mem.	#Storage	Access	
□	968	MODAL tools	●	Galaxy MODAL	0%	16	64	1	ssh http	⊕
□	969	myPortal	●	Galaxy	0%	4	8	0	ssh http	⊕
	2			2		20	72	1		

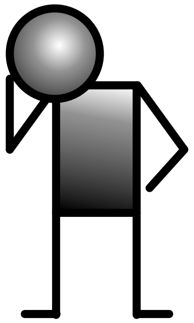
Show 25 entries

c2.large	11 / 16
c2.small	44 / 64
c2.xlarge	5 / 8
c3.large	11 / 15

RAINBio : Registry of bioinformatics tools and VMs

Prototype

Query :
topic ? tool ?
VM ?



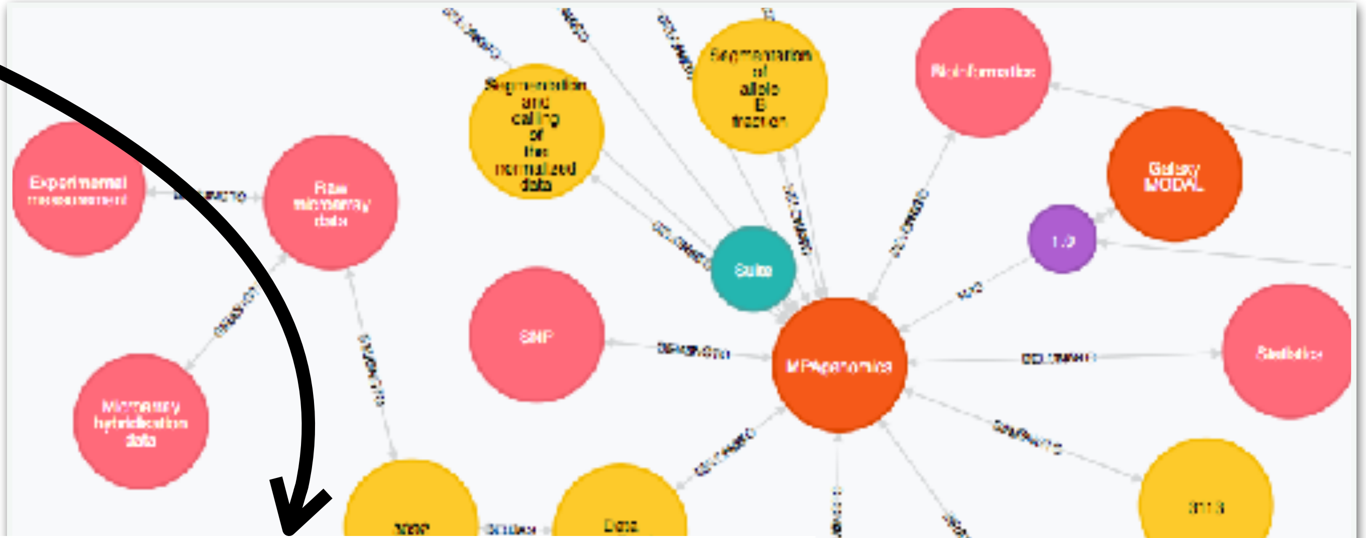
Life science researcher

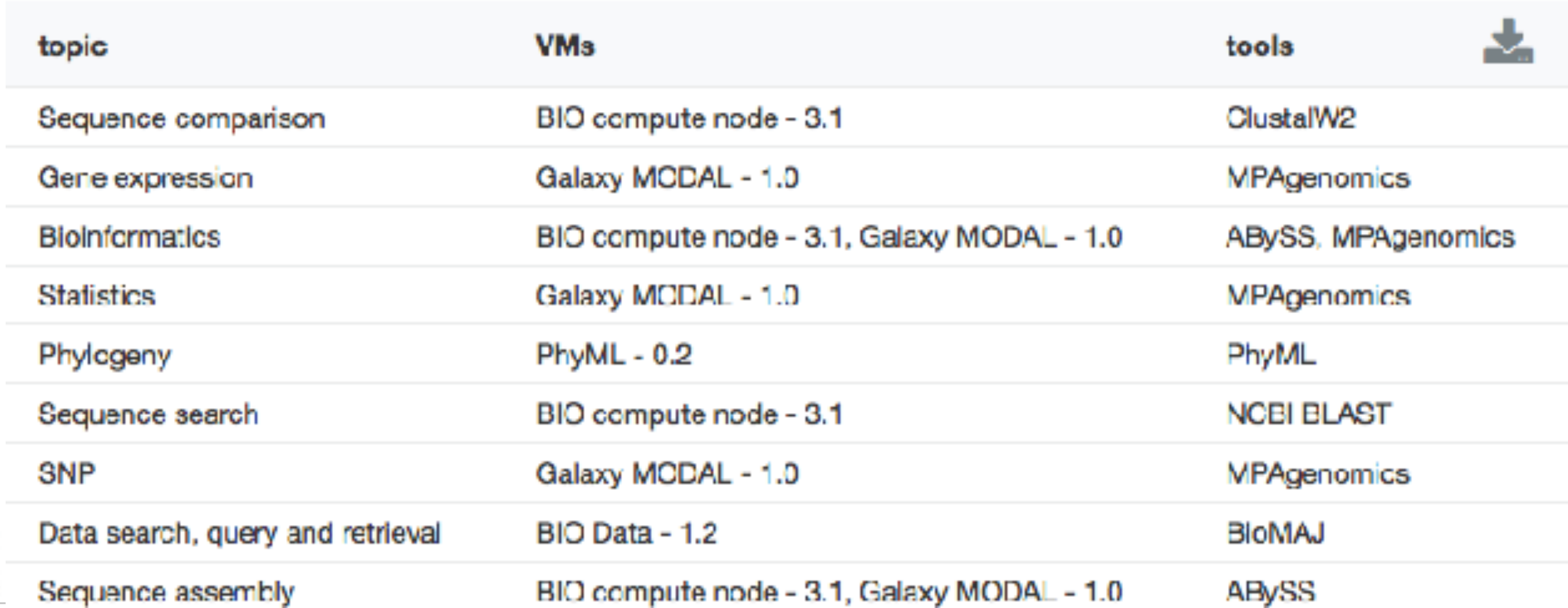
IFB's Cloud Marketplace

ELIXIR's Services Registry

VMs

Tools



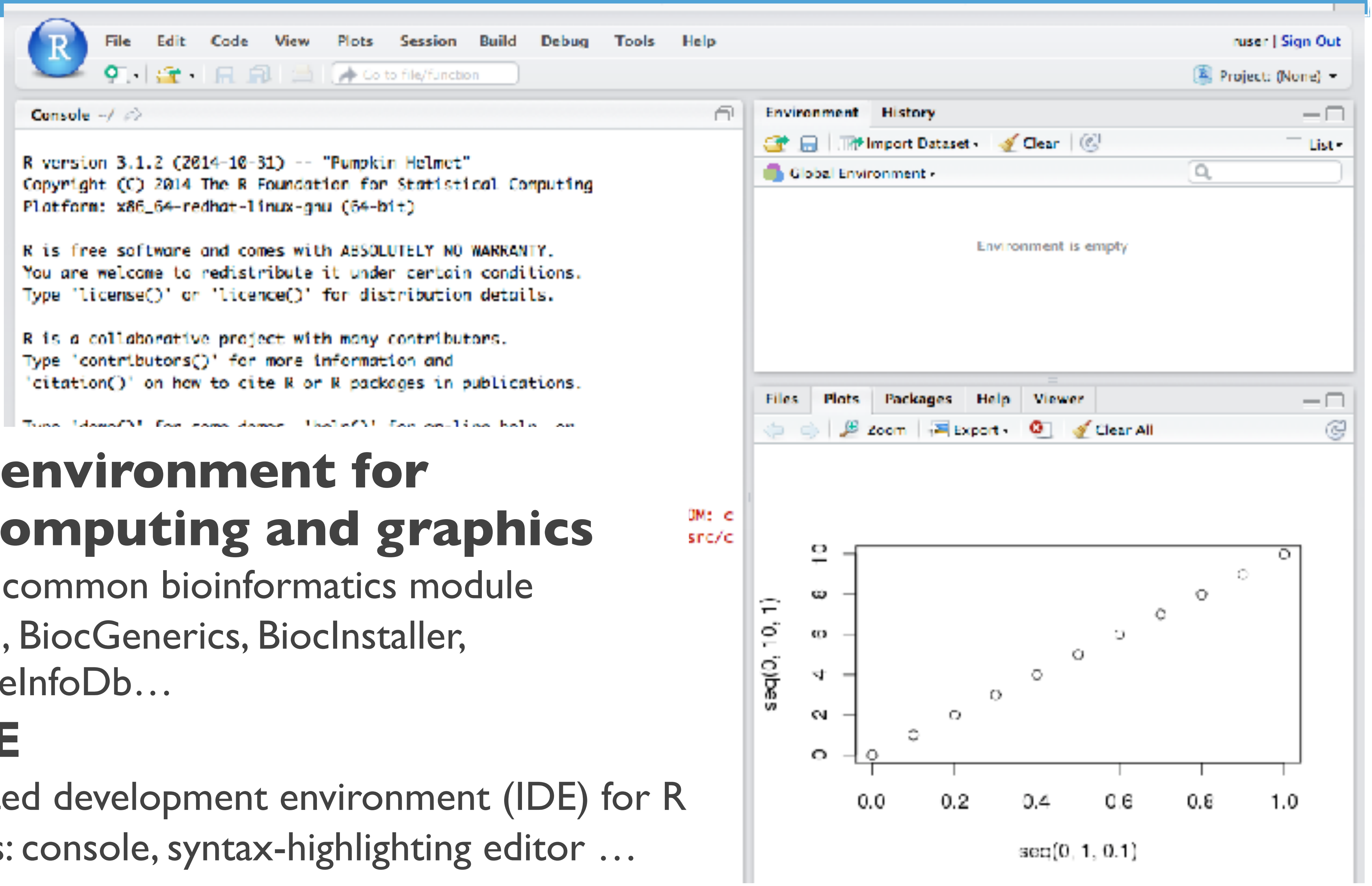


topic	VMs	tools
Sequence comparison	BIO compute node - 3.1	ClustalW2
Gene expression	Galaxy MODAL - 1.0	MPAgenomics
Bioinformatics	BIO compute node - 3.1, Galaxy MODAL - 1.0	ABySS, MPAgenomics
Statistics	Galaxy MODAL - 1.0	MPAgenomics
Phylogeny	PhyML - 0.2	PhyML
Sequence search	BIO compute node - 3.1	NCBI BLAST
SNP	Galaxy MODAL - 1.0	MPAgenomics
Data search, query and retrieval	BIO Data - 1.2	BioMAJ
Sequence assembly	BIO compute node - 3.1, Galaxy MODAL - 1.0	ABySS

RAINBio

Graph DB (Neo4J)

App R Statistical Computing



```
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
```

R software environment for statistical computing and graphics

- include common bioinformatics module
- Biobase, BiocGenerics, BiocInstaller, GenomeInfoDb…

RStudio IDE

- integrated development environment (IDE) for R
- features: console, syntax-highlighting editor …

Shiny web framework

- powerful web framework for building web applications using R.
- without requiring HTML, CSS, or JavaScript knowledge.

Contact: Stéphane Delmotte (IFB PRABI-LBBE)

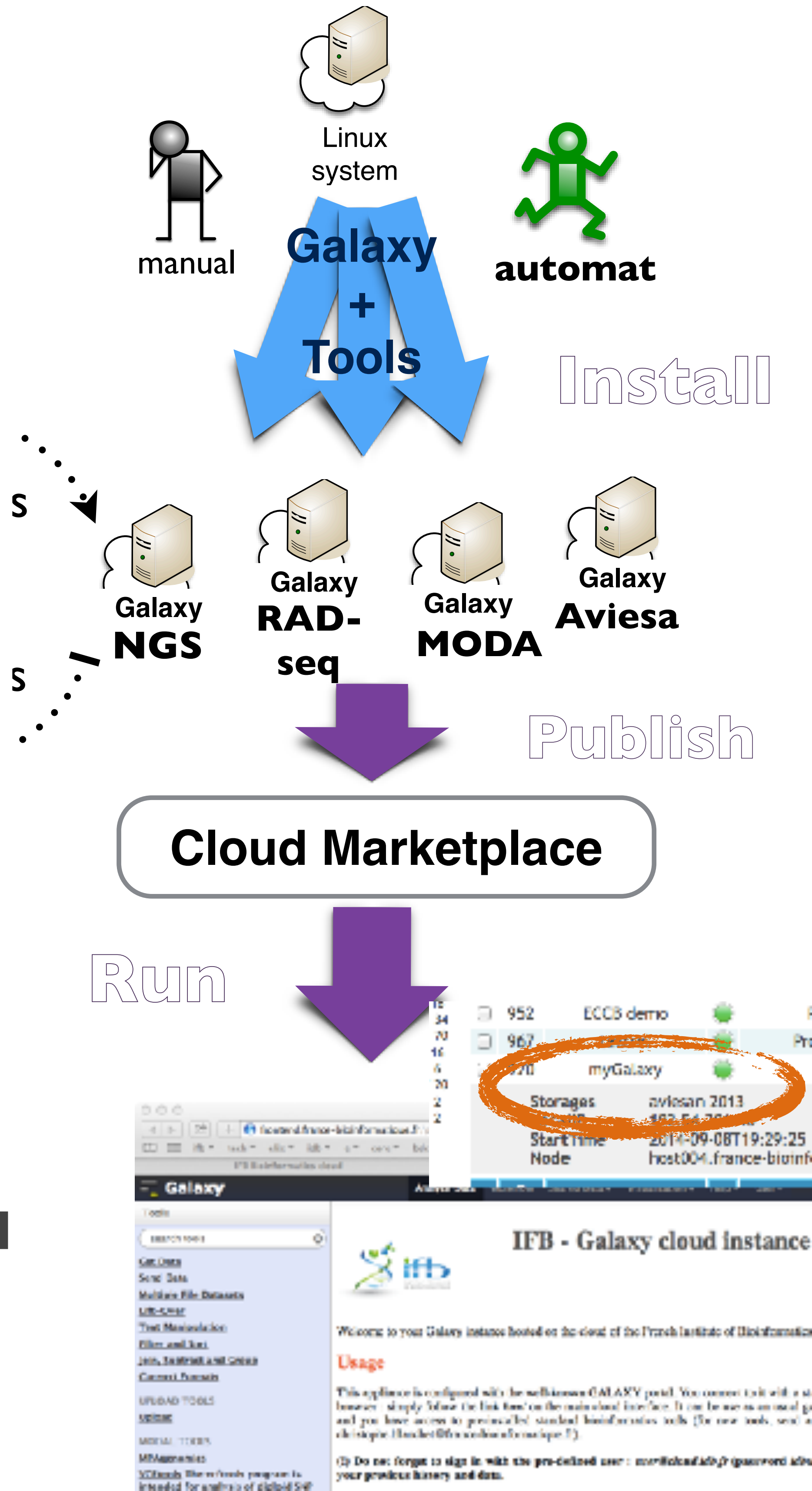
AppS Cloud Galaxy Portal

Galaxy portal is widely used in the community

- analyse NGS data (mainly but not only)
- connected to community knowledge: data and indexes, tools, workflows

Cloud advantages :

- User is **administrator of his/her own Galaxy** instance: he/she can install data and tools
- Preserve **workflows and results in cloud storage**
- Help the integration of monthly updates and new tools
- Different appliances can be available at the same time:
  - ★ a basic one with common tools for NGS
  - ★ specific ones for a domain or a set of tools
    e.g. Galaxy-MODAL, Galaxy-RADseq, **EBA-ChIP-Seq**
  - ★ or for training: create a special appliance with dedicated datasets, tools or workflows
    e.g. AVIESAN school 2015



952 ECCB demo
967
myGalaxy
Storages aviesan 2013
Start Time 2014-09-08T19:29:25
Node host004.france-bioinf

IFB - Galaxy cloud instance

Usage

App Multi-genomes browser

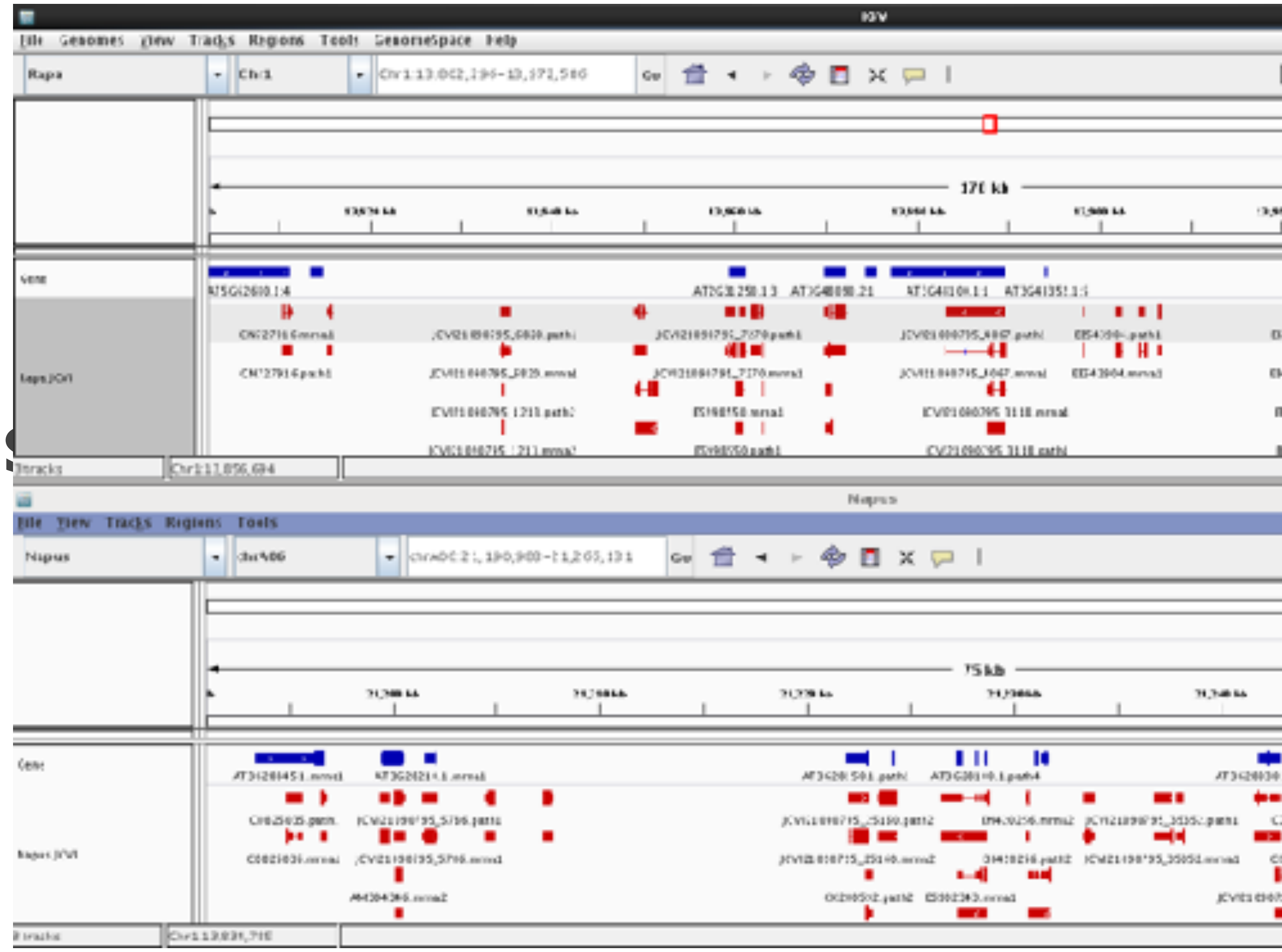
Integrative Genomics Viewer

Based on IGV

Ready to deploy in the cloud close to the datasets

Remote virtual desktop

- transfer only graphical visualization
- based on NX protocol

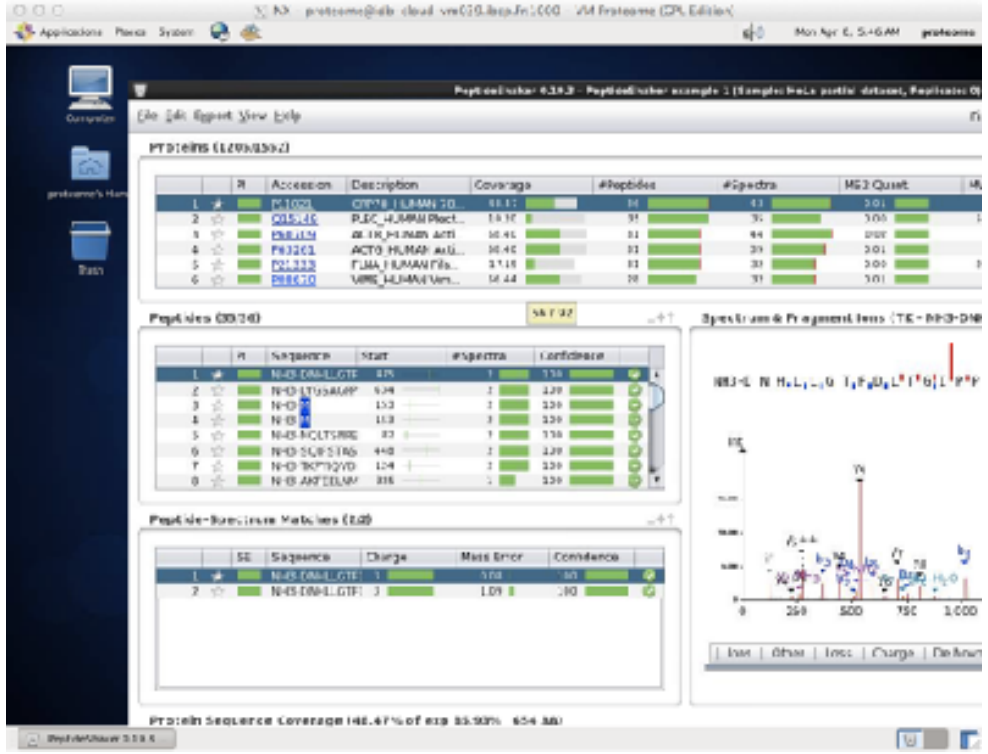

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Rapa
Chr1
Go
170 kb
Tracks
Napus
File View Tracks Regions Tools
Napus
Go
75 kb

Contact: Marie-Laure Franchinard (IFB MIGALE)

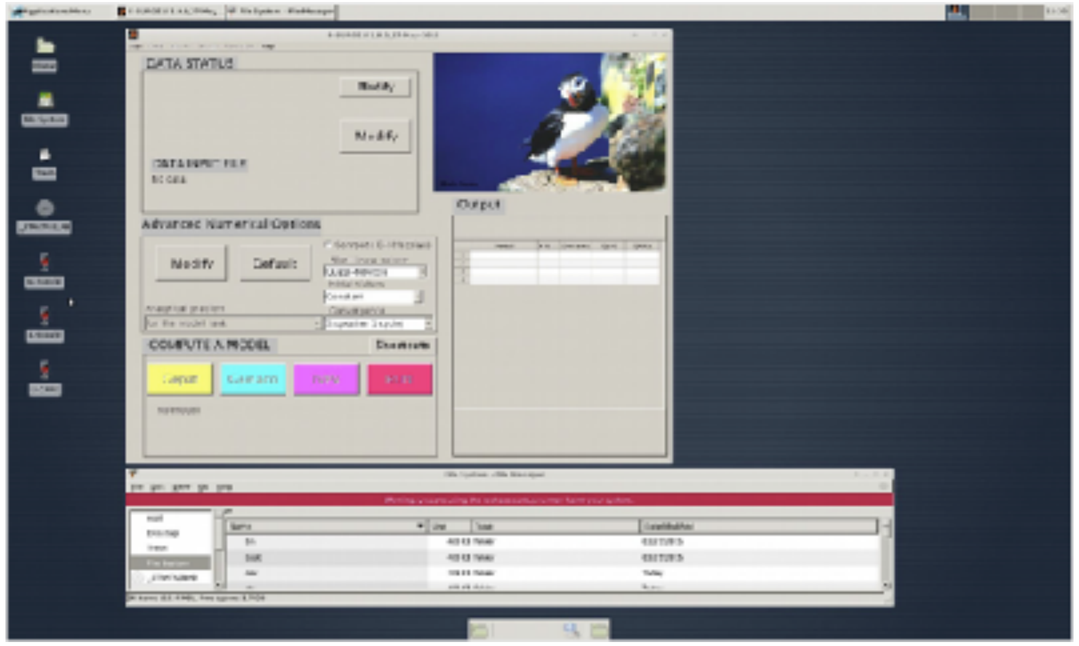
Funded by the French BIODATACLOUD project.

And other apps …

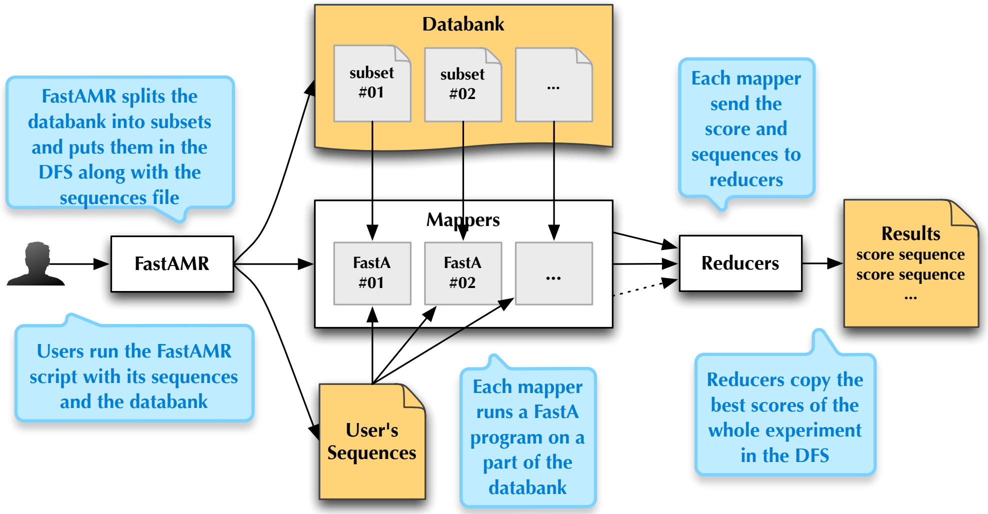
RSAT



Proteomics



Ecology of populations



FastAMR splits the databank into subsets and puts them in the DFS along with the sequences file

Databank

subset #01

subset #02

...

Each mapper send the score and sequences to reducers

FastAMR

Mappers

FastA #01

FastA #02

...

Reducers

Results
score sequence
score sequence
...

Users run the FastAMR script with its sequences and the databank

User's Sequences

Each mapper runs a FastA program on a part of the databank

Reducers copy the best scores of the whole experiment in the DFS

Hadoop

etc.

Conclusion: IFB's cloud today

22 bioinformatics appliances already available

- **+ 10 in progress** by the experts of the different life sciences domains
  - ★ BioDataCloud-**RNAseq**, **ProFi**, **Clinical NGS for cancerology** (2x), REPET, TriAnnot, Galaxy **RAD-seq**, **Bacterial genomics**, **iMetAMOS**…
- IFB supports different domain-specific developments
  - ★ Microbial Bioinformatics, Evolutionary bioinformatics, Plant bioinformatics, Structural Biology, NGS data processing, biomedical data analysis…
  - ★ Call for new proposals in progress

Scientific production - 239 users (October 2015)

- opened to members of IFB (standard allocated resources)
- opened to partners, academic and industry, infrastructures and projects: e.g. BioDataCloud, ProFi, MetaboHub, …
- extra resources allocation according to scientific and financial criteria

Training

- Scientific school "Cumulo NumBio - Cloud Computing for Life Sciences" (Aussois, June 2015)
- IFB's tutorials for cloud end-users and appliance developers
- tutorial at ECCB'14 about 'Analysis of Cis-Regulatory Motifs from High-Throughput Sequence Sets'
- Bioinformatics Masters in Marseille (2014) and Rouen (2015)
- Scientific school about Genomics with Galaxy (2015)

Questions ?

http://www.france-bioinformatique.fr

Acknowledgments

- **IFB** members
  - IFB hub: **Patricia**, **Awa, Jean-François, Mohamed, Jonathan, Maxime, Dominique**
    *Alumni : **Marie, Quentin***
    ➡ ***we are hiring !***
  - Working group IFB-GRISBI (co-chair with Olivier Collin)

- **Appliances developers**
  Samuel Blanck (Inria Lille), **Jacques** van Helden (TAGC), **Stéphane** Delmotte (PRABI-LBBE), **Bruno** Spataro (PRABI-LBBE), **Marie-Laure** Franchinard (MIGALE), **Anis** Djari (BioinfoGenoToul), **Bertrand** Néron (Institut Pasteur), **Adrien** Josso (MicroScope), **Thomas** Lacroix (MIGALE), **Christian** Baudet (CLB), **Germain** Paimparay & **Baptiste** Brault (CFB)**…**

- **CNRS IDRIS**: R. Medeiros, C. Gauthey and staff
- **StratusLab** members

- IFB is funded by **French programs PIA INBS 2012, BioDataCloud**
- **EU H2020 projects, CYCLONE (644925) and EGI-Engage (6541**