

# The French bioinformatic cloud to analyse life science data



Christophe Blanchet

Institut Français de Bioinformatique - IFB  
French Institute of Bioinformatics - ELIXIR-FR  
CNRS UMS360I - Gif-sur-Yvette - FRANCE

Using clouds and VMs in bioinformatics training  
CSC, Helsinki, 24 May 2016

# IFB - Institut Français de Bioinformatique

## French distributed infrastructure for life-science information



<http://www.france-bioinformatique.fr>

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

- A national hub : IFB-core
- A network of 36 platforms

**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-FR Node**



**IFB-core**



# Relation with External Projects

## IFB is the ELIXIR French Node

- To promote consistency and complementarities between the components offered by the ELIXIR French node and those of other European nodes

## EXCELERATE (EU-H2020 676559)

- Services registry (WP-1 Tools Interoperability and Service Registry)
- Compute Platform (WP-4 Technical Services)
- Collaboration in connection with the e-Infrastructure
  - ★ WP5: The ELIXIR Interoperability Backbone
  - ★ WP6-9 : Use Cases
  - ★ WP10: ELIXIR Node Capacity Building and Communities of Practice

## CYCLONE (EU-H2020 644925)

- Scientific use cases: Securing human biomedical data, Cloud virtual pipeline for microbial genomes analysis, Live remote cloud processing of sequencing data...
- Multi-cloud deployment, AAI, Network-as-a-Service...

## EGI-Engage (EU-H2020 654142)

## BioDataCloud (French PIA INBS 2012)

- Industrial use cases about plant genomics assembling and visualization.

# IFB's e-Infrastructure

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

- A national IT resource + 35 platforms
- Hardware, data collections and bioinformatics tools

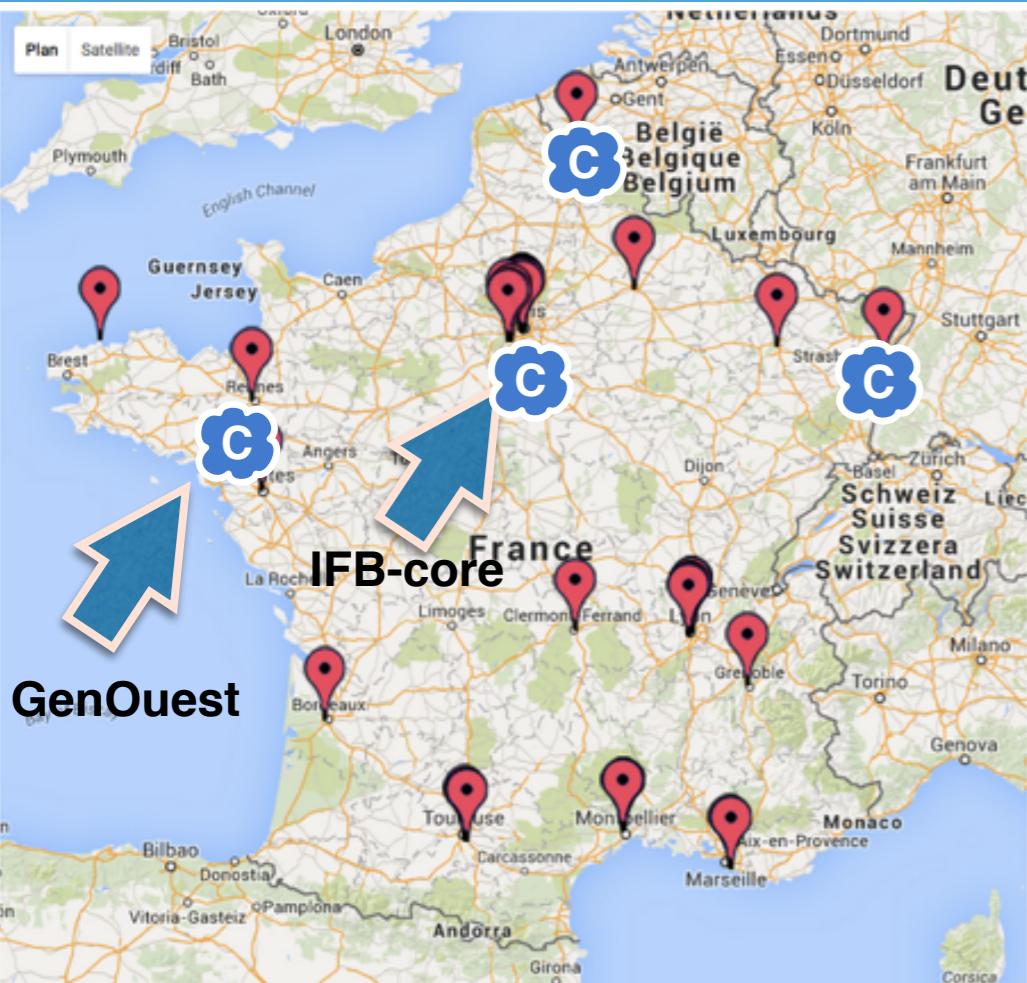
## Current resources

- 15,000 cores - 6 PB
- 4 running clouds



➡ Create a federation of clouds for life sciences

# Towards a multi-cloud infrastructure



## Requirements

- common identities and authorizations management based on community standards (EduGAIN)
- interoperability of virtual images (VM/container)
- tools for multi-cloud deployment (e.g. the SlipStream/ NuvLa broker)
- network management, security and propagation over several datacenters

## Solutions expected from

- ELIXIR/EXCELERATE and CYCLONE projects

SITE	Compute #cores	Storage #TB	RAM #GB	Largest VM	Technology	Location
IFB-core 2014-	200 (+160)	50 (+96)	2,000 (+1)	20c 256GB	StratusLab	CNRS-IDRIS, Paris
	5,000	1,000	27,500	64c 2TB	OpenStack	CNRS-IDRIS, Paris
2017	10,000	2,000+			OpenStack	CNRS-IDRIS, Paris
GenOuest 2014-	220 (+96)	8 (+20)	685	8c 32GB	OpenNebula	IRISA, Rennes

# Many tools



ABYSS 1.3.4	BWA 0.7.10	Mobyle	SearchGUI 1.10.4
ARIA 2.3	CAP3	MODAL	SeqClean
Bioconductor 2.11	CD-HIT 4.6.1	MultAlin 5.4.1	Shiny
biomaj	Clustal Omega 1.0.3	MUSCLE 3.8.3I	Stacks
BLAST+ 2.2.27	CLUSTALW 2.1	neo4j	STAR 2.4.0fI
Blat 35	Cufflinks 2.0.2	Oases 0.2.08	SuMo vI
Bowtie 0.12.8	Cutadapt 1.2.1	OMSSA 2.1.9	TGICL
Bowtie2 2.0.0-	E-SURGE 1.9.0	PeptideShaker 0.18.3	TopHat 2.0.6
beta7	Exonerate 2.2.0	phym 3.1	trim_galore 0.3.7
BWA 0.6.2	eXpress 1.5.1	PREDATOR 2.1.2	Trinity 2.0.4
	FastA 3.6	proline	U-CARE 2.3.2
	FastQC 0.10.1	python 2.7	VCFtools 0.1.11
	Galaxy portal	R 2.13	Velvet 1.2.10
	GATK 2.3.4	R 3.1.1	X!tandem
	HMMer 3.0	R 3.1.2	I2-10-01-I
	ImageJ 1.48	R-studio	XPLOR-NIH 2.30
	khmer 1.1	Ray 1.3	...
	M-SURGE 1.8.5	RSAT	
	MEME 4.7	samtools 0.1.18	
	MMSEQ 0.11.2a	Samtools 1.1	

# Many data

## Collections of reference data (TB)

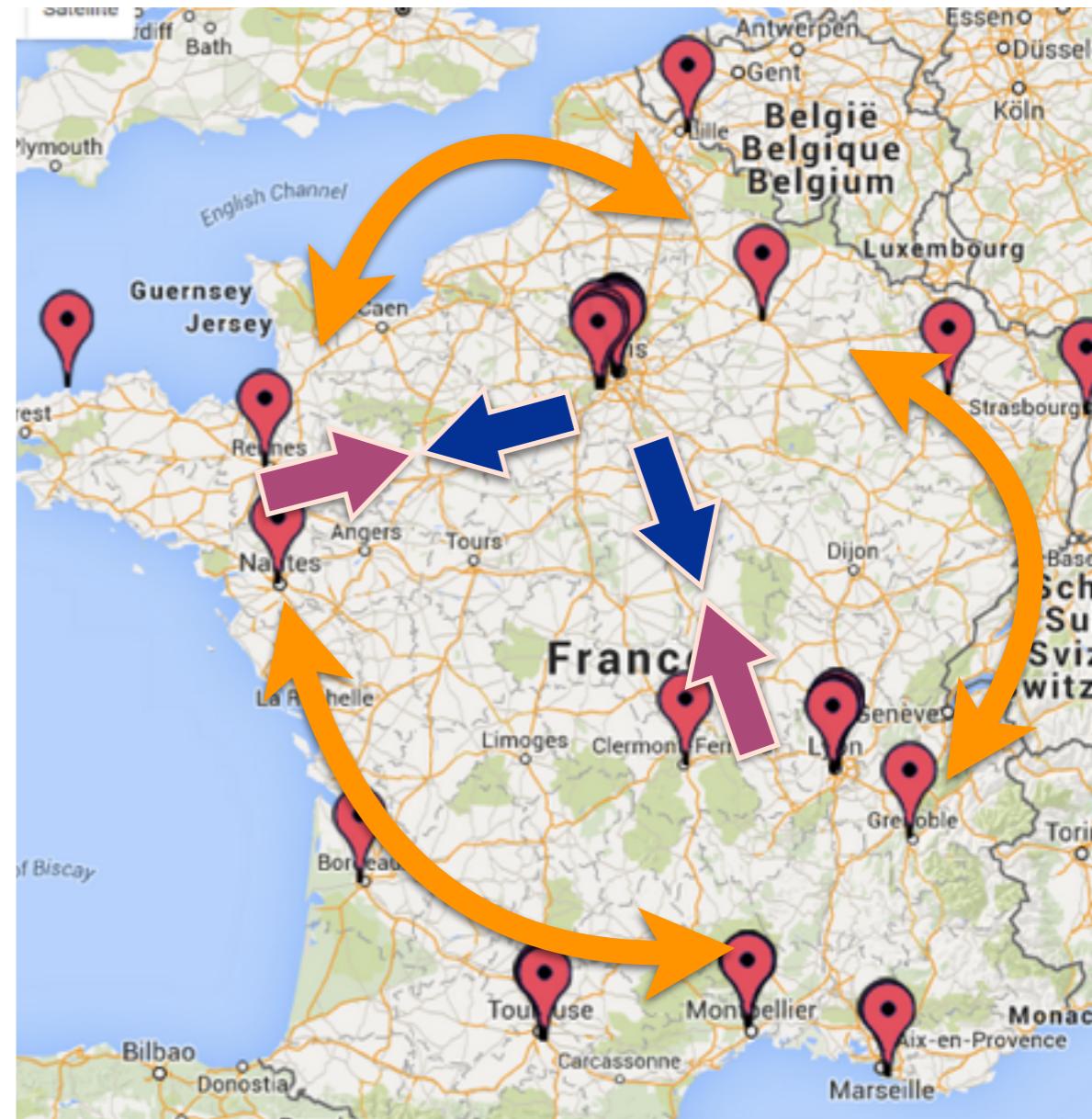
- IFB-Core mirrors and computes indexes
- PFs can subscribe to databases update
- Transfer from IFB-core to regional PFs
  - ★ BioMAJ, iRODS, gridftp

## Experimental data (TB)

- Regional desks for initial deposit
- Local archiving and analysis
- Transfer from regional PFs to IFB-core
  - ★ iRODS, gridftp

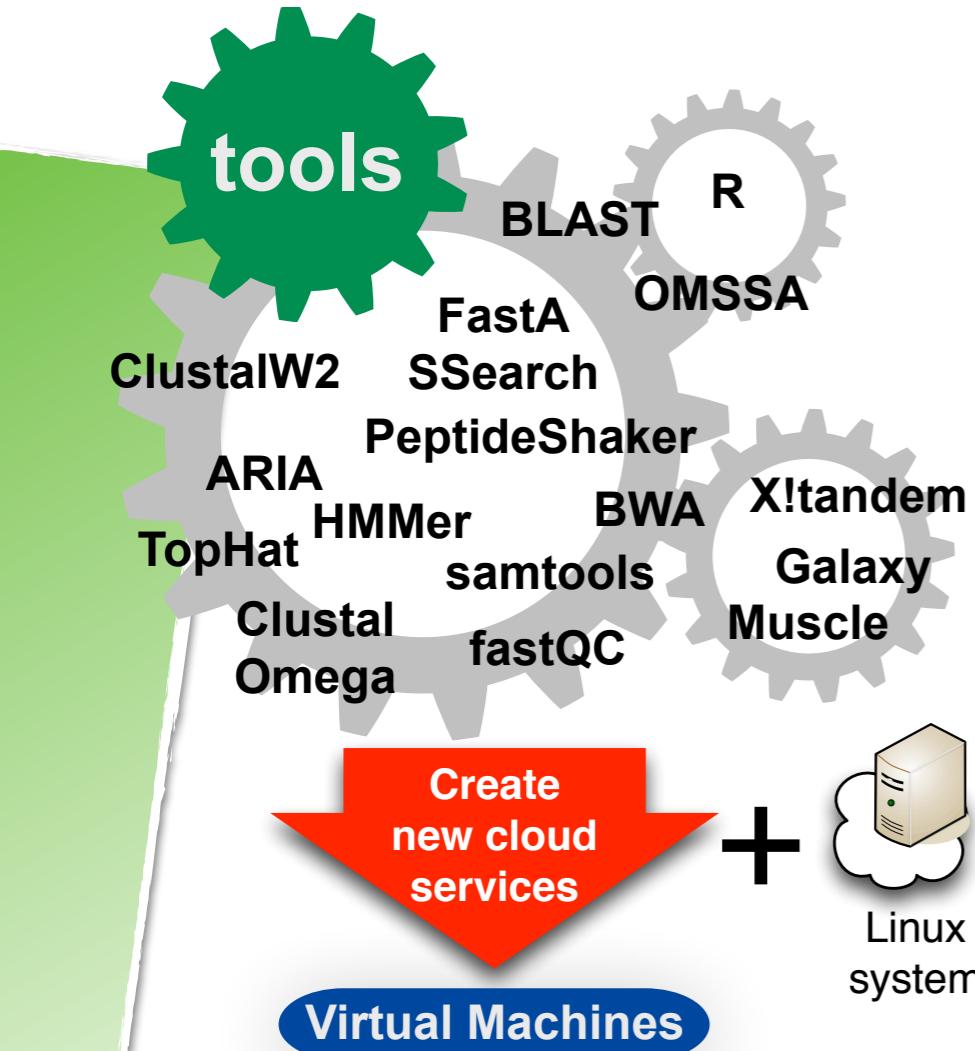
## User data (MB-TB)

- Many (can be large),
- Heterogeneous
- Unpredictable distribution
  - ★ iRODS, object storage, noSQL



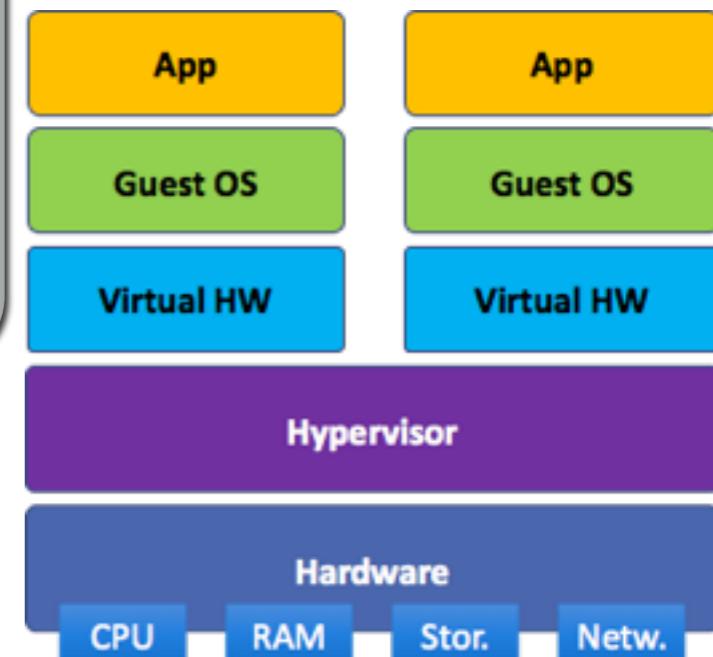
Two main issues : Transfer / Biomedical data

# Create bioinformatics “appliances”



## What is an appliance ?

- predefined image of a virtual machine
- with tools, pipeline, recipes...
- ready to run
- small size image (GB)

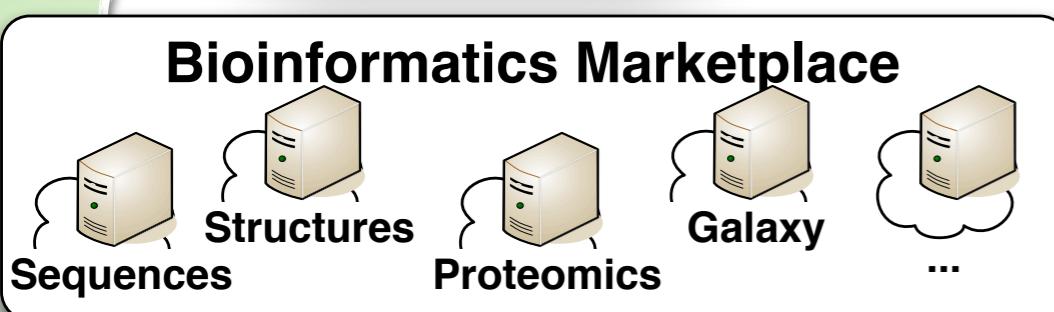


## Created by

- developers
- experts of different life science domains

## How to help the developers?

- training sessions for developers
- cloud support team of IFB-core
- individual hackathon at IFB-core



# RAINBio : registry of bioinformatics VM images

<http://cloud.france-bioinformatique.fr/rainbio/>

The screenshot shows a web-based application for managing bioinformatics VM images. At the top, there's a navigation bar with the IFB logo, 'Rainbio' button, 'Browse EDAM' button, 'Sign in' button, and 'Help' button. Below the navigation, a message states: 'Appliances come with tools that are labeled by EDAM topics. Here we propose a view of these Appliances, tools and topics.' There are three tabs: 'Appliances' (selected), 'Tools', and 'EDAM topics'. The main area displays a grid of appliance entries. Each entry includes a thumbnail, the appliance name, a list of tools it contains, and a list of EDAM topics. For example, the 'BIO compute node (3.3)' entry lists tools like fastqc, Bioconductor 2.11, TopHat, MMSEQ 0.11.2a, and many others, along with topics such as Molecular biology, Statistics and probability, Bioinformatics, Sequence analysis, RNA splicing, Mapping, Data architecture, analysis and design, Mathematics, Sequence comparison, Genomics, Biology, Protein structure analysis, Protein structure prediction, Sequence composition, complexity and repeats, Protein folds and structural domains, Sequence sites, features and motifs, and Sequence assembly.

Appliance	Tools	Topics
BIO compute node (3.3)	fastqc, Bioconductor 2.11, TopHat, MMSEQ 0.11.2a, samtools 0.1.18, R, muscle, MultAlin 5.4.1, Clustal Omega, Bowtie2 2.0.0-beta7, bowtie, Ray, BWA, ABYSS, predator, MEME 4.7, HMMER, fasta, CLUSTALW 2.1, cap3, BLAST+ 2.2.27	Molecular biology, Statistics and probability, Bioinformatics, Sequence analysis, RNA splicing, Mapping, Data architecture, analysis and design, Mathematics, Sequence comparison, Genomics, Biology, Protein structure analysis, Protein structure prediction, Sequence composition, complexity and repeats, Protein folds and structural domains, Sequence sites, features and motifs, Sequence assembly
Bacterial genomics (Insyght) (1.2)	Web interface, BLAST+ 2.2.30, python 2.7, HMMER	Sequence composition, complexity and repeats, Protein folds and structural domains, Sequence sites, features and motifs, Sequence comparison
Bio Imaging (1.1)	Bureau virtuel, ImageJ	Data architecture, analysis and design, Imaging
BioDataCloud Assemblage (1.0)		
BioDataCloud IGV (1.0)	Bureau virtuel, IGV - Integrative Genome Viewer	
BioDataCloud RNA-seq (1.2)	CD-HIT 4.6.1, khmer, SeqClean, TGICL, Trinity 2.0.4, pyth, eXpress, Samtools 1.1, trim_galore 0.3.7, Exonerate 2.2.0, STAR 2.4.0f1, BWA, oases, velvet	
BioPerl (2015-12)	BioPerl, python 2.7	

## Metadata from

- ELIXIR bio.tools
- cloud marketplace
- docker hub
- BioShaddock

## Scientific apps

- NGS
- Biolmaging
- Proteomics
- ...

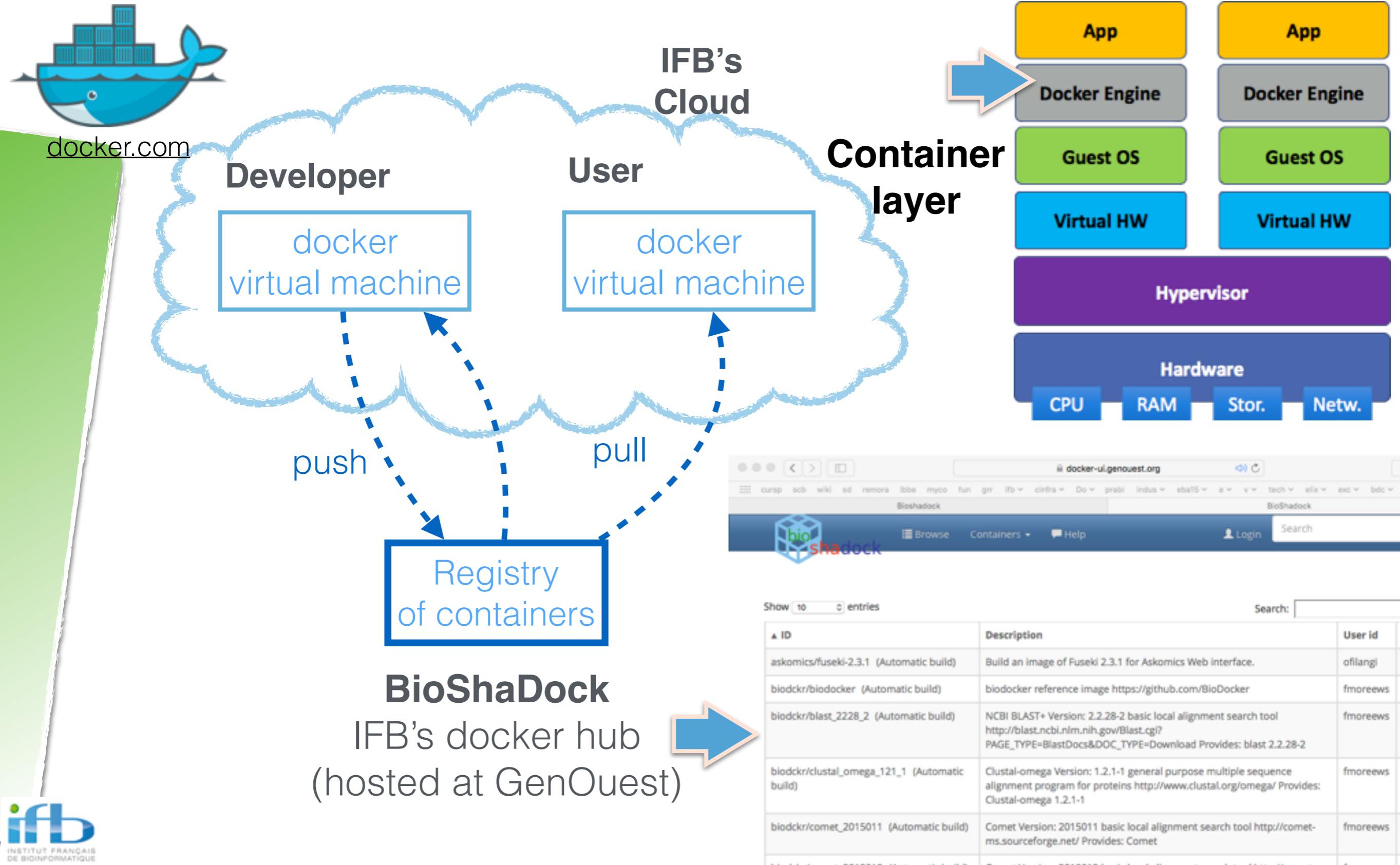
## Utilities

- base OS
- data mgmt
- batch sched.

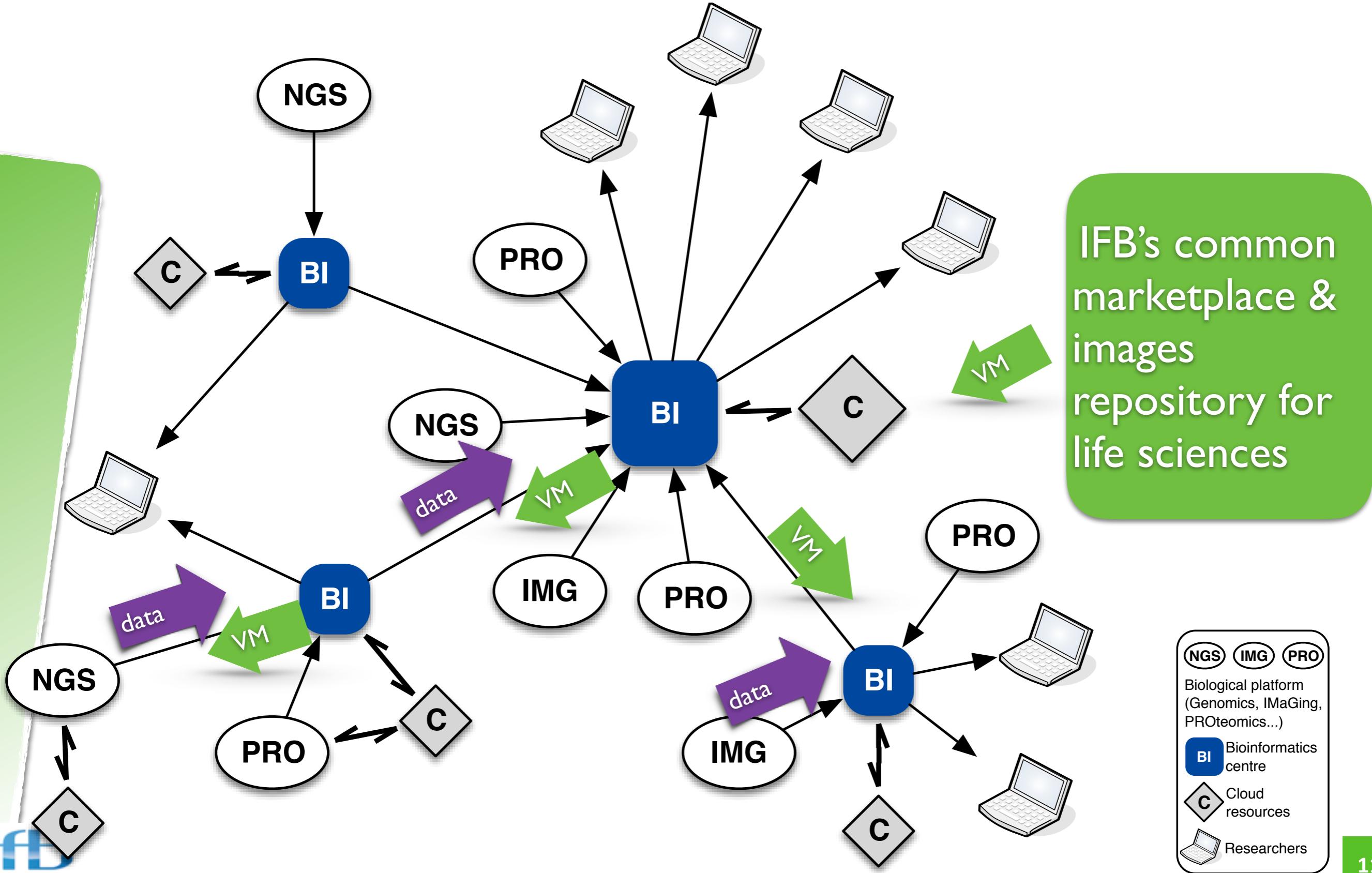
## Interfaces

- CLI
- Web portal
- Remote Desktop

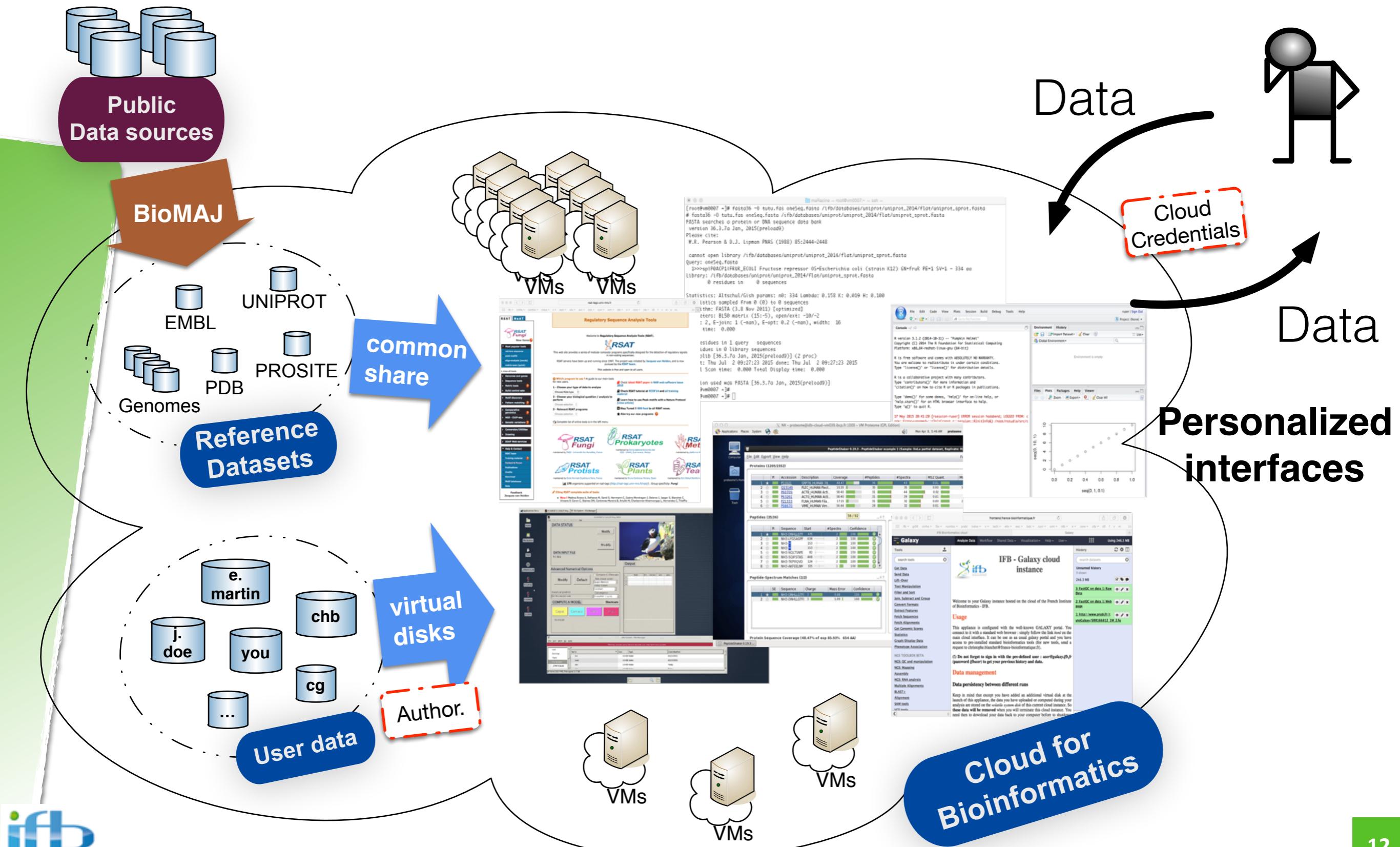
# Simplify tools integration with docker



# Move virtual images rather than data



# Use IFB's Bioinformatics Cloud

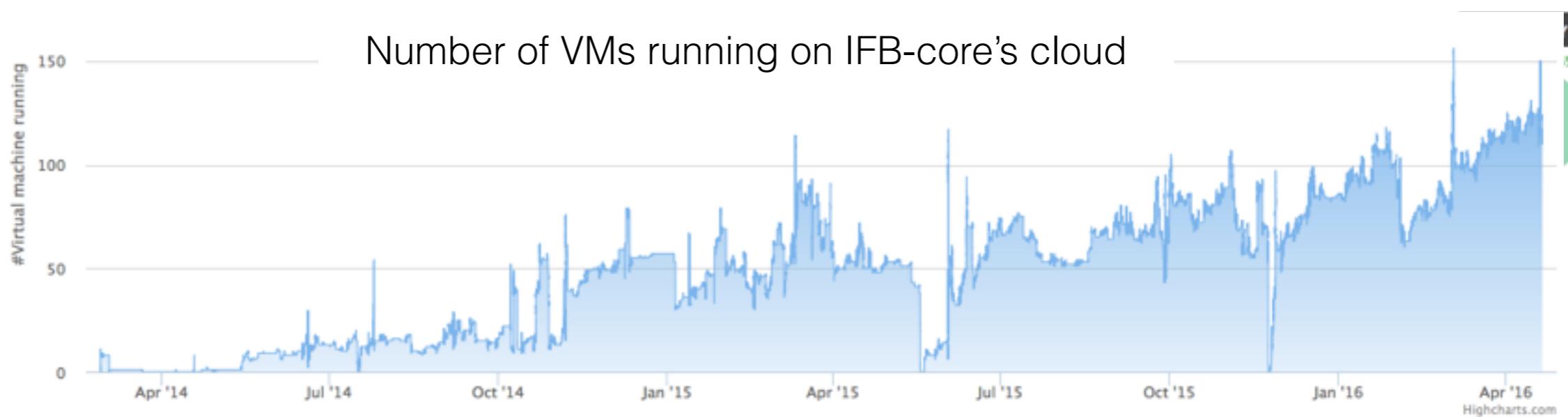


# IFB's Cloud Usage

The screenshot shows the IFB Bioinformatics Cloud dashboard. At the top, there is a navigation bar with links to various projects like cursp, wiki, sd, myco, fun, grr, ifb, cinfra, Do, prabi, indus, e, v, tech, elix, elex, bdc, cycl, egi, madics, masto, omt, and a search bar. Below the navigation is the IFB logo and the title "IFB BIOINFORMATICS CLOUD". The dashboard header includes "DASHBOARD", "HOSTED AT idris", and "POWERED BY stratuslab". A news section shows 4 entries, and a table lists "ROOM FOR VMs" with columns for Name, Appliance, CPU%, CPU, Mem., #Storage, Access, and a status column. To the right, there are three circular performance indicators: one green circle at the top, one labeled "CPU free (97.28%)", and another green circle at the bottom. The main area features a line chart titled "Number of VMs running on IFB-core's cloud" showing the count over time from April 2014 to April 2016.

Room	VM Count
c2.large	37 / 92
c2.small	167 / 368
c2.xlarge	17 / 46
c3.large	34 / 83
c3.medium	76 / 175
c3.xlarge	15 / 37
c3.xxlarge	4 / 13
m1.large	7 / 15
m1.medium	22 / 41
m1.xlarge	1 / 2
m1.xxlarge	1 / 2

Number of VMs running on IFB-core's cloud



# App Biocompute

## Standard bioinformatics node

### With pre-installed standard bioinformatics tools

- BLAST, FastA, SSearch, HMM,...
- ClustalW2, Clustal-Omega, Muscle,..
- Bowtie(2), BWA, samtools, ...
- MEME, R, etc.

### Connected to public reference datasets

- Uniprot, EMBL, genomes, PDB, etc.
- Automatically shared with the VMs

### Cluster mode

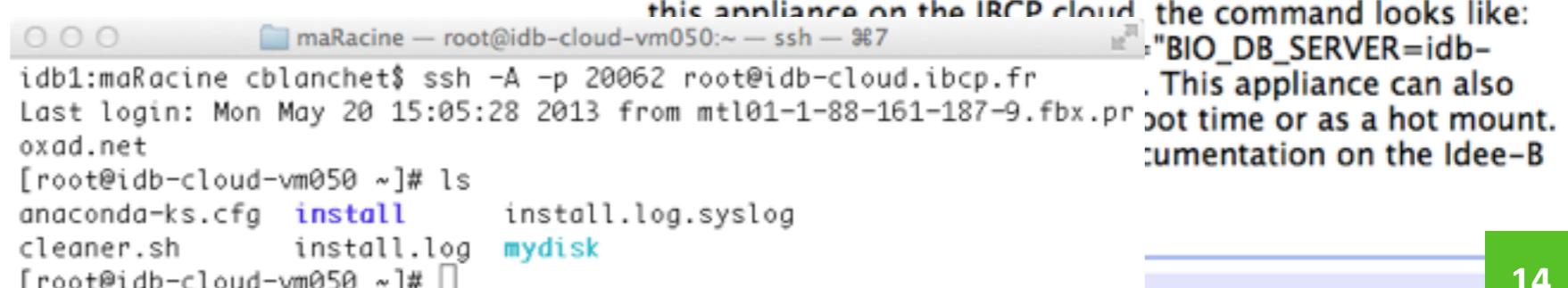
- turn several instances in a single virtual cluster
- shared file system
- batch scheduling



The screenshot shows a web-based interface for managing cloud resources. At the top, there's a toolbar with icons for navigation and a search bar containing 'marketplace.ifb.idris.fr/metadata'. Below the toolbar, a header menu includes 'Home', 'Endorsers', 'Query', 'Upload', and 'About'. The main content area is titled 'Metadata' and displays a table with one row. The table has columns for 'Show' (set to 10 entries) and a search bar labeled 'BIO compute node'. The single row in the table is titled 'BIO compute node' and contains the following details:

Endorser:	christophe.blanchet@ibcp.fr
Identifier:	O2fHwlZlxLDoxcuCmqwoWVGBpBM
Created:	2014-04-04T15:34:44Z
Kind:	machine

A descriptive text block follows, explaining the nature of the appliance and how it can be used. It mentions tools like abyss, blast+, bioconductor, bowtie, bowtie2, bwa, cap3, clustal-omega, clustalw2, fasta36, gor4, hmm, meme, mmseq, multalin, muscle, predator, ray, R, samtools, simpa96, tophat, tophat2. It also notes the availability of a shared file system and batch scheduling.



The screenshot shows a terminal window with a blue header bar. The terminal prompt is 'maRacine — root@idb-cloud-vm050:~ — ssh — %7'. The user is running the command 'ssh -A -p 20062 root@idb-cloud.ibcp.fr'. The output shows the user logging in from 'mon May 20 15:05:28 2013 from mtl01-1-88-161-187-9.fbx.pr'. The user then runs 'ls' to list files in the current directory, which include 'anaconda-ks.cfg', 'install', 'install.log', 'install.log.syslog', 'cleaner.sh', 'mydisk', and '[root@idb-cloud-vm050 ~]#'. The terminal window has a light gray background and white text.

# App R Statistical Computing



A screenshot of the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Tools, and Help. The top right corner shows 'ruser | Sign Out' and 'Project: (None)'. The main window has several panes: 'Console' showing R version 3.1.2 startup messages; 'Environment' pane showing 'Global Environment' with 'Environment is empty'; 'Plots' pane showing a scatter plot of points from seq(0, 1, 0.1); 'Packages' pane showing 'DM: c' and 'src/c'; and 'Files' pane showing file navigation buttons and a search bar.

## 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)

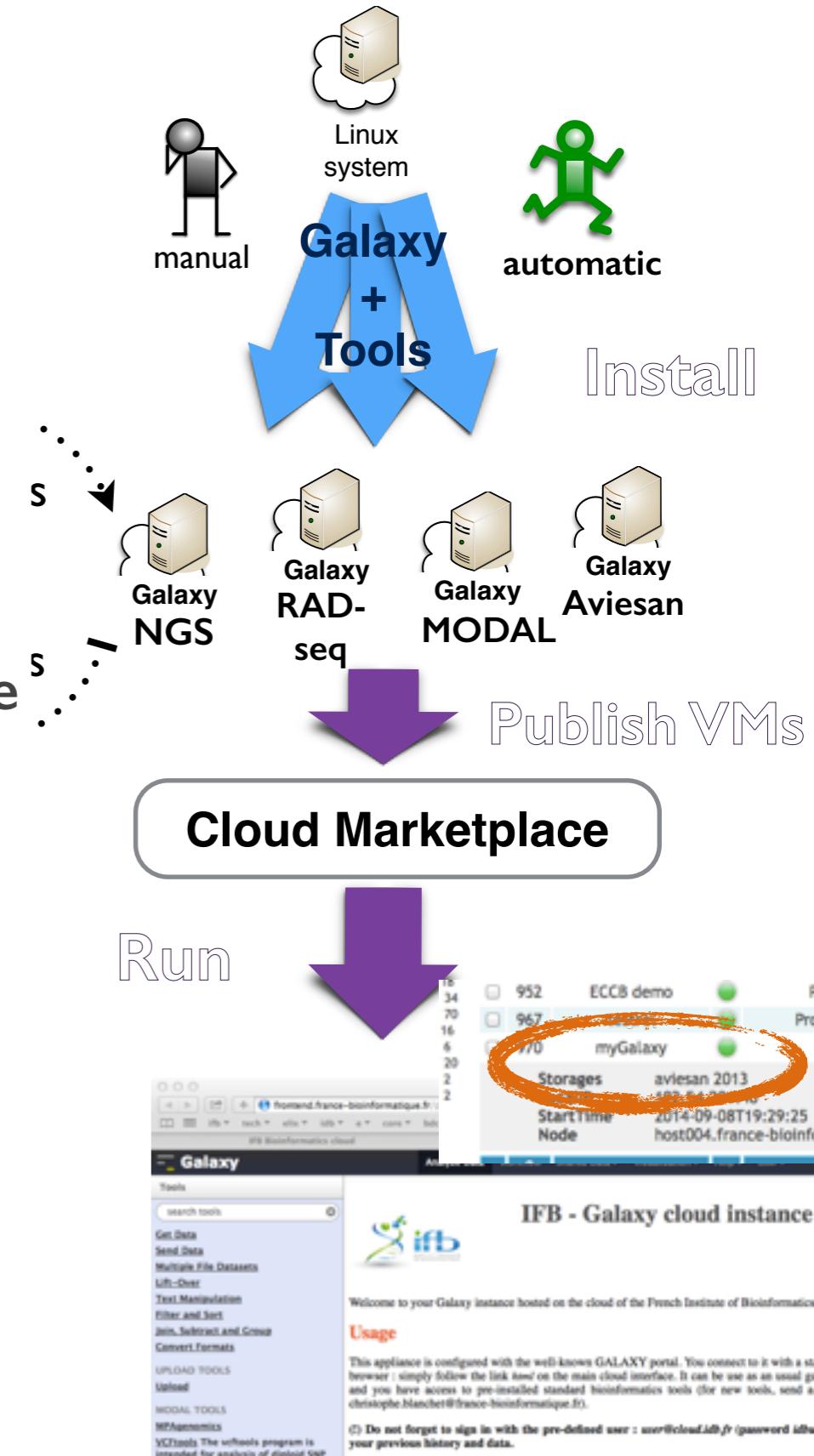
# AppS Cloud Galaxy Portal

## 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



# App Multi-genomes browser



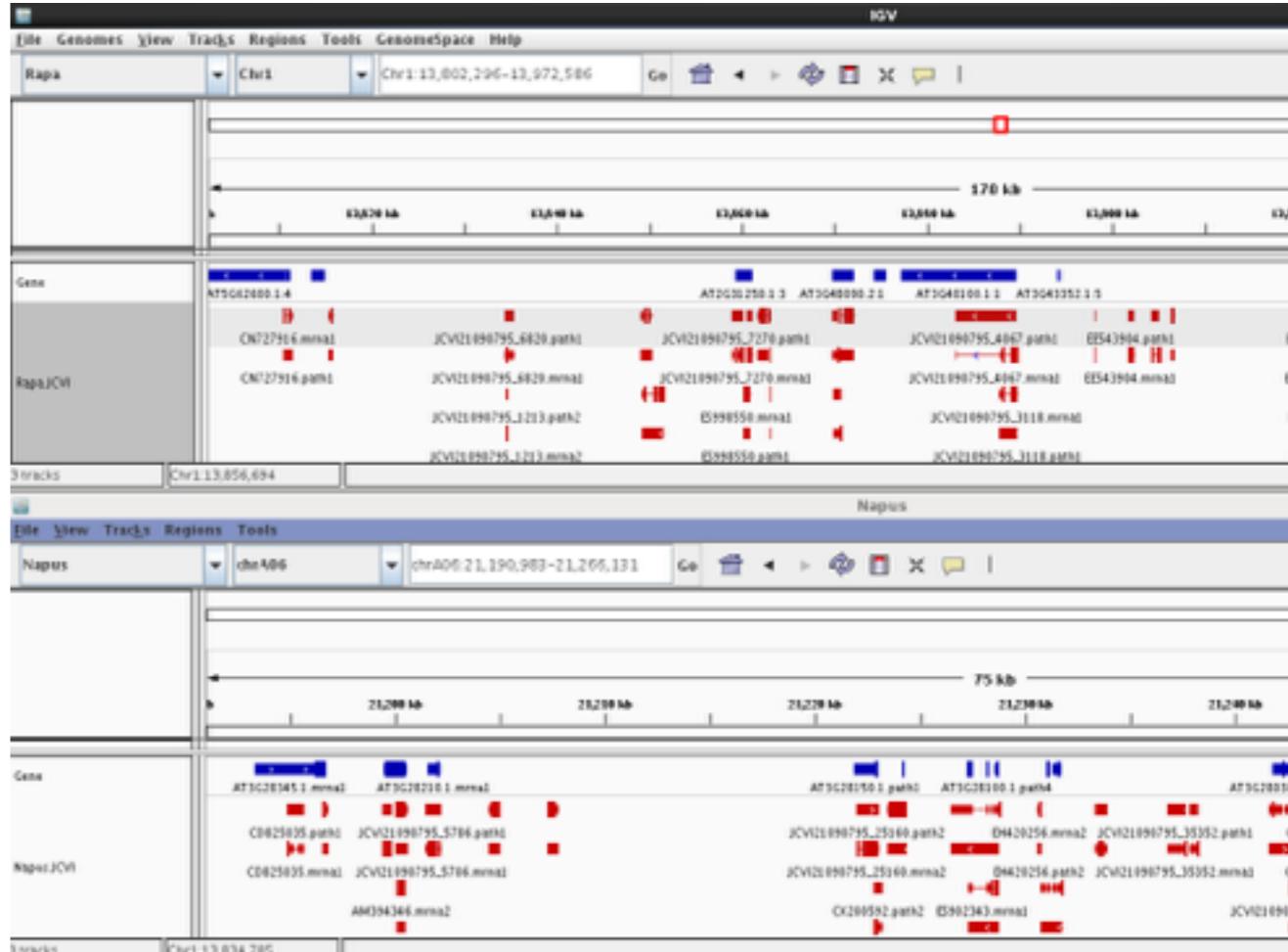
Integrative  
Genomics  
Viewer

Based on IGV

Ready to deploy in the cloud  
close to the datasets

Remote virtual desktop

- transfer only graphical visualization
- based on NX protocol



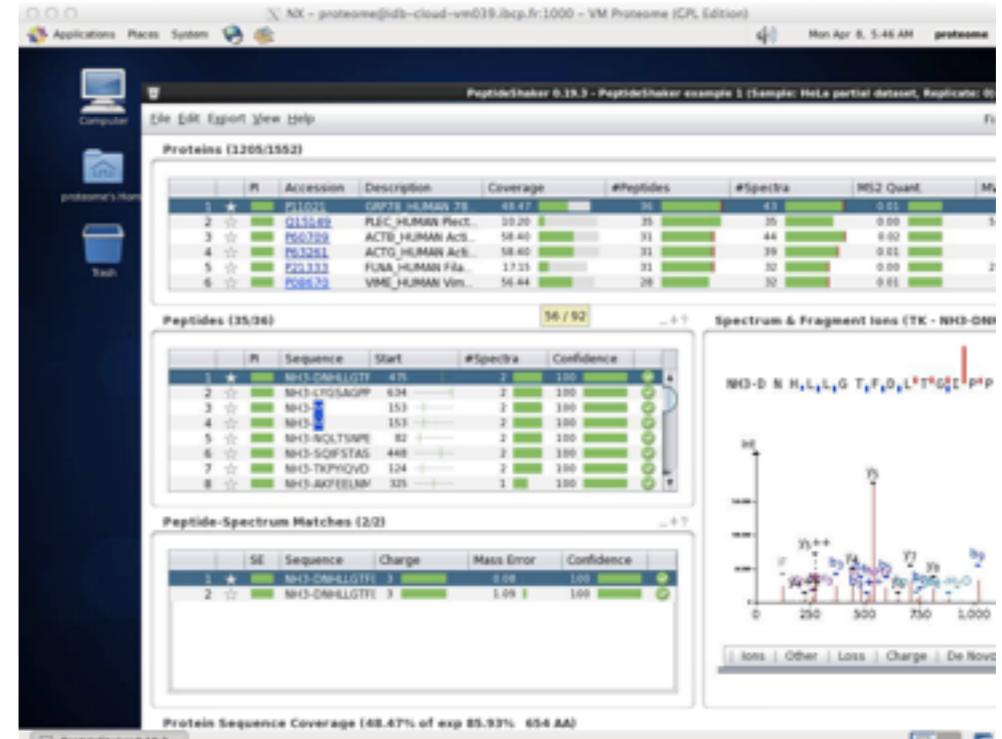
Contact: Marie-Laure Franchinard  
(IFB MIGALE)

Funded by the French  
BIODATACLOUD project.

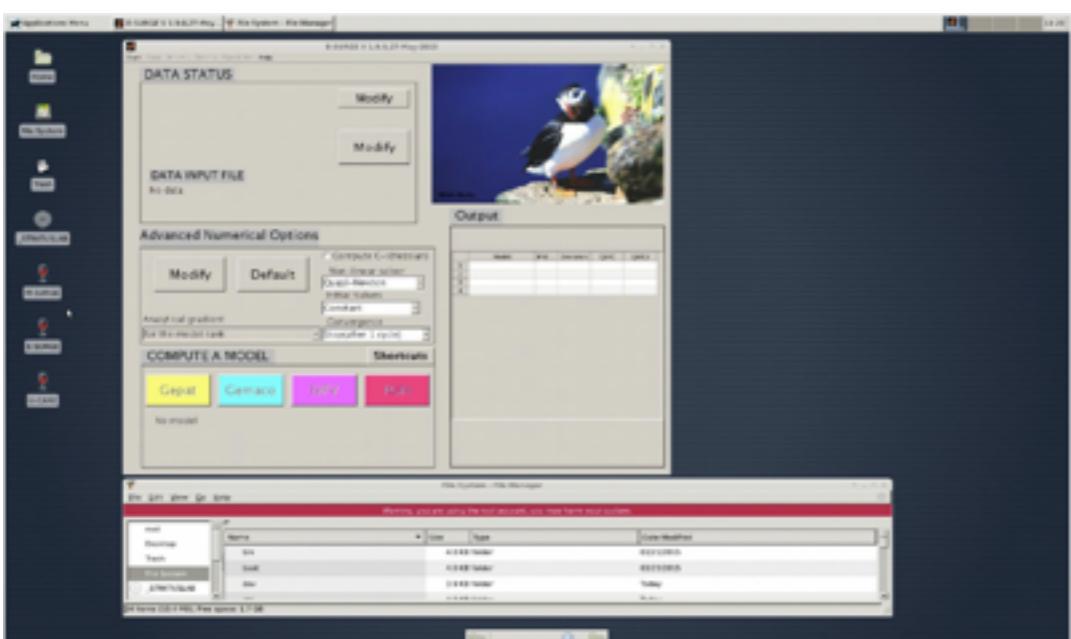
bpifrance



# And other apps ...

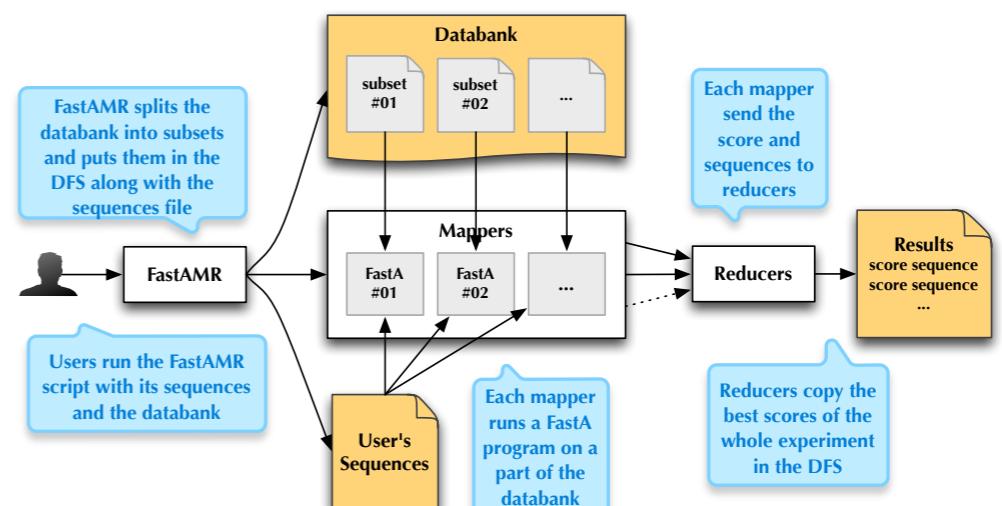


## RSAT



**Ecology of populations**

## Proteomics



**Hadoop**

**etc.**

# Cloud training sessions 2014-15

Title	Teachers	Dates	Part.	Level
Tutoriels Cloud pour la Biologie <i>Gif-sur-Yvette</i>	IFB-core, GenOuest	juin 2014	23	débutant
Tutoriels Cloud pour la Biologie Rennes	GenOuest, IFB-core	nov. 2014	20	débutant
GRISBi-27 <i>Gif-sur-Yvette</i>	IFB-core	mars 2015	27	admsys, développeur
Ecole Cumulo Numbio <i>Aussois</i>	IFB-core, LRI, Grid5000, FranceGrilles, KerData	juin 2015	53	mixte
Journées work packages bioinformatique IFB <i>Gif-sur-Yvette</i>	IFB-core	juin 2015	15	développeur bioinformatique
Journée IBC <i>Montpellier</i>	IFB-core	nov. 2015	20	débutant

# IBI-1 - IFB cloud basics

**Public :** tout public

**Fréquence :** bimestrielle

**Niveau requis**

- débutant sur le cloud,
- utilisateur de services bioinformatiques

**Objectifs**

- Savoir utiliser le cloud IFB pour des analyses de données biologiques : exécuter ses propres machines virtuelles et transférer ses données entre son poste de travail et le cloud et récupérer les résultats.

**Contenu:**

- Présentation du cloud IFB ;
- Présentation du tableau de bord ;
- Déploiement des machines virtuelles ;
- Gestion des données avec des disques virtuels, leur gestion ;
- Les différents types de connexion aux VMs (SSH, Web et bureau à distance).
- Pratique : utilisation de l'appliance Galaxy et d'un bureau virtuel

**Documents:**

- Diapos : Présentation du cloud IFB
- Fascicule : Utilisation du cloud IFB

**Sessions**

- 3rd February 2016
- 16th April 2016

# IBI-2 - IFB cloud advanced

**Public :** Utilisateur du cloud

**Fréquence :** trimestrielle

**Niveau requis**

- Avoir suivi IBI-1 (ou équivalent) et maîtriser la ligne de commande

**Objectifs**

- Savoir déployer une application complexe pour l'analyse intensive de données biologiques de grande taille.
- Savoir adapter les machines virtuelles disponibles pour répondre à des besoins plus complexes.

**Contenu:**

- Déploiement d'une application bioinformatique complexe comprenant plusieurs machines virtuelles
- Installation de logiciels à partir d'archives (codes source ou binaires) ou à l'aide de scripts (approver)
- Utilisation de conteneurs docker pour l'installation de logiciels bioinformatiques
- Intégration des ressources de données (génome, annotation ...) grâce à l'appliance Biomaj ou avec d'autres solutions
- Gestion des données avec des disques virtuels en NFS

**Documents:**

- Diapos : présentation des fonctionnalités avancées du cloud IFB
- Fascicule : cas pratique des fonctionnalités les plus fréquemment utilisées

**Session**

- 3rd May 2016

# IBI-3 - IFB cloud for developers

## Public :

- Développeurs avec une pratique du cloud IFB

Fréquence : semestrielle

## Niveau requis

- avoir suivi IBI-2 (ou équivalent) et connaitre le système d'exploitation Linux

## Objectifs

- Savoir intégrer un logiciel ou un pipeline bioinformatique dans une machine virtuelle pour une diffusion et mise à disposition sur le cloud IFB.

## Contenu:

- Présentation des bonnes pratiques de création d'appliance.
- Présentation des fonctionnalités avancées disponibles dans le cloud IFB : le montage automatique des collections de données publiques de référence, la contextualisation d'un portail web, la configuration des disques virtuels pour la conservation des paramètres d'un logiciel ou portail ...
- Présentation des différents modèles d'intégration: archives (codes source ou binaires), scripts d'installation automatique (approver, puppet), conteneurs (docker).
- Création de conteneurs docker pour le déploiement de logiciels bioinformatiques
- Choix et configuration de l'interface pour les utilisateurs (CLI, portail web, bureau virtuel à distance)
- Rédaction d'une description pour le référencement dans le cloud IFB

## Documents:

- Diapos : Présentation des bonnes pratiques de création d'appliance sur le cloud IFB
- Fascicule : Exemples de cas pratique

## Sessions

- 16th mars 2016 (FG)
- July 2016

# Master courses

Title	Teachers	Dates	Students	Special apps (*)
Master cours « cis-régulation » (M2) <i>Univ. Marseille</i>	J. van Helden, A. Griffon	oct. 2014	45	RSAT
Master Bioinformatique (M1) <i>Univ. Rouen</i>	S. Gallina	jan.-fév. 2015	10	Non <i>BIO ComputeNode Galaxy</i>
Polytech Biotech <i>Univ. Marseille</i>	D. Puthier	déc. 2015	38	TAGC Cours-Unix
Master Bioinformatique ADC (M1) <i>Univ. Lyon</i>	P. Veber	déc. 2015	18	Jupyter Notebook
Master Bionformatique AMI2B (M2) <i>Univ. Paris-Saclay</i>	D. Gautheret, F. Lemoine, C. Billerey	déc. 2015 -jan. 2016	20	COURS M2 Paris-Saclay 2015

\* The special appliance was built by the teachers with the help of the IFB-core staff.

# Conclusion

## Pilote infrastructure is running since 2014

- 2 clouds in IFB's datacenters, 2 others in collaboration with academia

## 36 bioinformatics appliances already available

- for different life science domains, several developments in progress

## Scientific usage - 357 users

- 9500+ VMs created since 2014 on the IFB-core's cloud

## Training (2014-16)

- Users & developers : 8 cloud sessions, 4 scientific trainings
- Students : 5 master courses (130 st.), 3 already planned for 2016

## To whom it is available

- Members of IFB and French life science community
  - ★ academic and commercial
- Partners of academic/industry infrastructures and projects:
  - ★ national: BioDataCloud, ProFi, MetaboHub...
  - ★ European: ELIXIR-EXCELERATE, CYCLONE, EGI-ENGAGE...
- *Including trainers and trainees.*

# Prospects

**Extend the national cloud site**

**Help regional centers to deploy a cloud**

**Federate the national and regional clouds**

**Expand the catalogue of bioinformatics cloud services  
(with new VM/container images)**

**Make resources access simple with AAI and multi-cloud deployment**

**Provide users with a collaborative working space**

**Set up a continuous integration environment for the development and validation of bioinformatics images**

# Questions ?

<http://www.france-bioinformatique.fr>

## Acknowledgments

- **IFB members**
  - IFB-core : Awa, Bryan, Dominique, Jean-François, Jonathan, Frédéric, Mohamed, Patricia, Sandrine, Victoria (*Alumni : Marie, Maxime, Quentin*)
  - Working group **IFB-GRISBI** : Olivier Collin and members
- **Appliances developers**  
 Samuel Blanck (Inria Lille), Jacques van Helden (TAGC), Stéphane Delmotte (PRABI-LBBE), Bruno Spataro (PRABI-LBBE), Marie-Laure Frachinard (MIGALE), Anis Djari (BioinfoGenoToul), Bertrand Néron (Institut Pasteur), Adrien Josso (MicroScope), Thomas Lacroix (MIGALE), Christian Baudet (CLB), Germain Paimparay & Baptiste Brault (CFB), Olivier Inizan (URGI), Jocelyn Brayet (Institut Curie), Guillaume Brysbaert (Bilille), Guy Perrière (PRABI) ...
- **CNRS IDRIS, StratusLab developers**
- IFB is funded by French programs **PIA INBS 2012**
- EU H2020 projects: **CYCLONE (644925)**, **EXCELERATE (676559)** and **EGI-Engage (654142)**

