

Cloud pour la Bioinformatique



Christophe Blanchet

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

Sequencing data



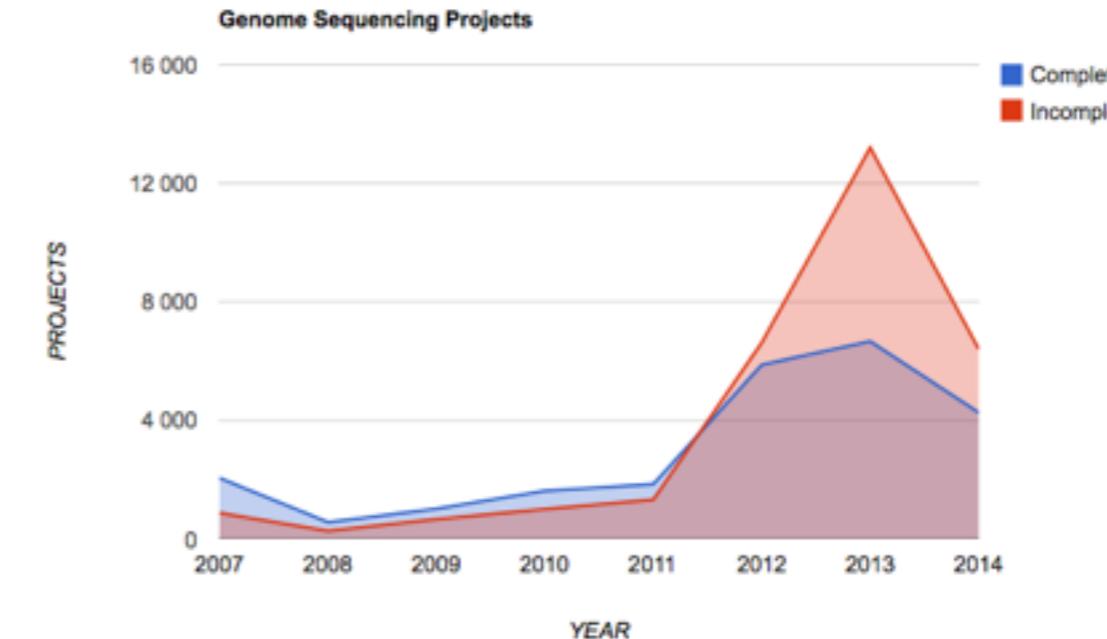
Next-Generation Sequencing Statistics

Vendor:	Roche			Illumina			ABI		
Technology:	454			Solexa GA			SOLiD		
Platform:	GS20	FLX	Ti	I	II	IIx	1	2	3
Reads: (M)	0.5	0.5	1.25	28	100	150	40	115	320
Fragment									
Read length:	100	200	400	35	50	100	25	35	50
Run time: (d)	0.25	0.3	0.4	3	3	5	6	5	8
Yield: (Gb)	0.05	0.1	0.5	1	5	15	1	4	16
Rate: (Gb/d)	0.2	0.33	1.25	0.33	1.67	3	0.34	1.6	2
Images: (TB)	0.01	0.01	0.03	0.5	1.1	2.8	1.8	2.5	1.9
PA Disk: (GB)	3	3	15	175	300	300	300	750	1200
PA CPU: (hr)	10	140	220	100	70	NA	NA	NA	NA
SRA: (GB)	0.5	1	4	30	50	2.5	100	140	600

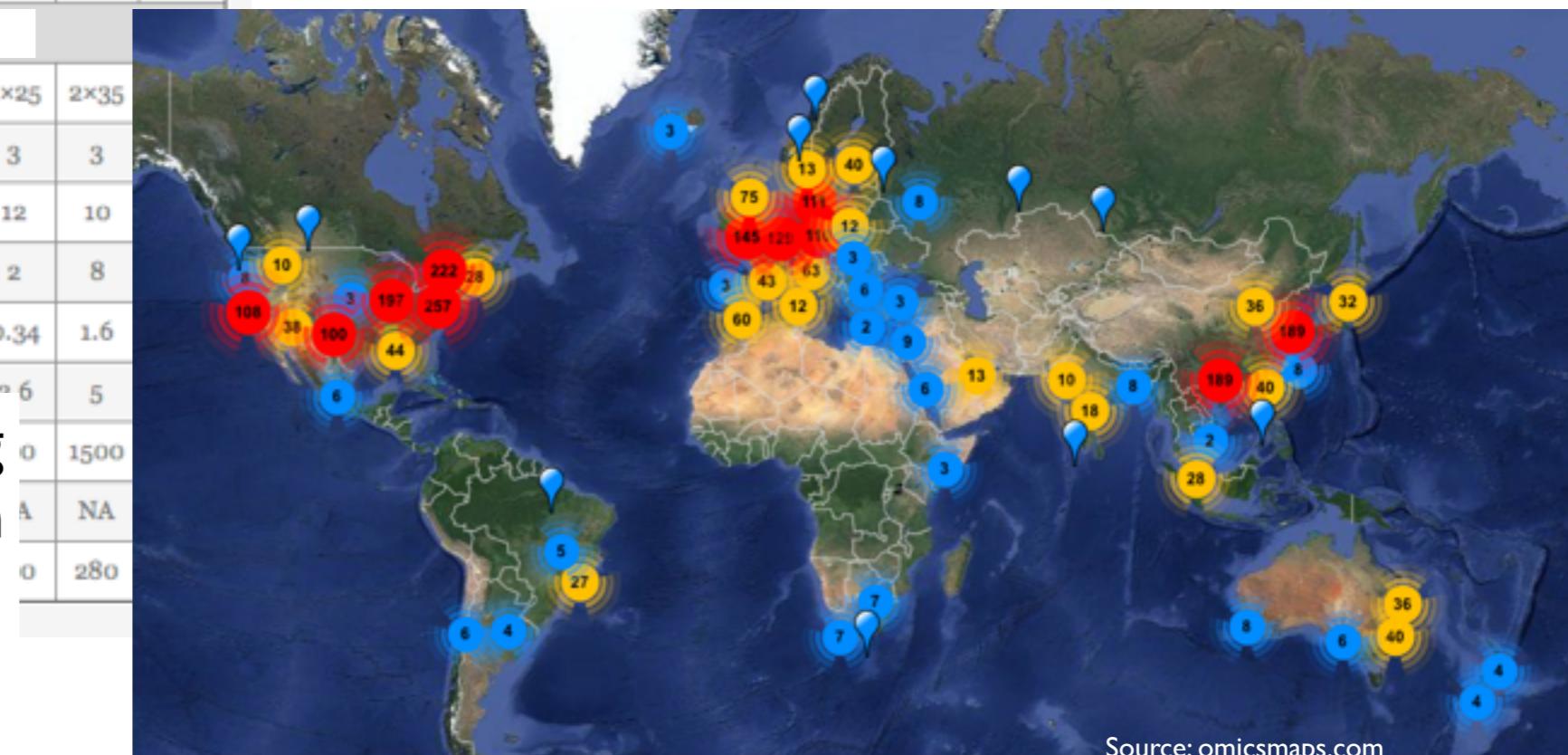
source: www.dolitigenomics.com/next-generation-

Read length:	200	400	2x35	2x50	2x100	2x25	2x35
Insert: (kb)	3.5	3.5	0.2	0.2	0.2	3	3
Run time: (d)	0.3	0.4	6	10	10	12	10
Yield: (Gb)	0.1	0.5	2	9	30	2	8
Rate: (Gb/d)	0.33	1.25	0.33	1.67	3	0.34	1.6
Transcripts (TR)	0.01	0.02	1	0.0	0.6	0.6	5

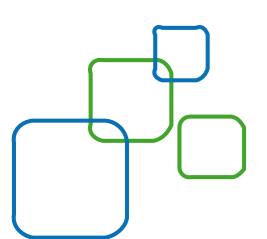
Complete genome sequencing
become a lab commodity with
NGS (cheap and efficient)



source: www.genomesonline.org

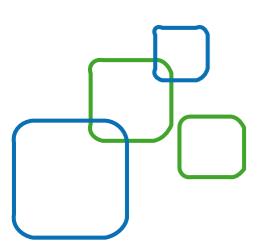


Source: omicsmaps.com

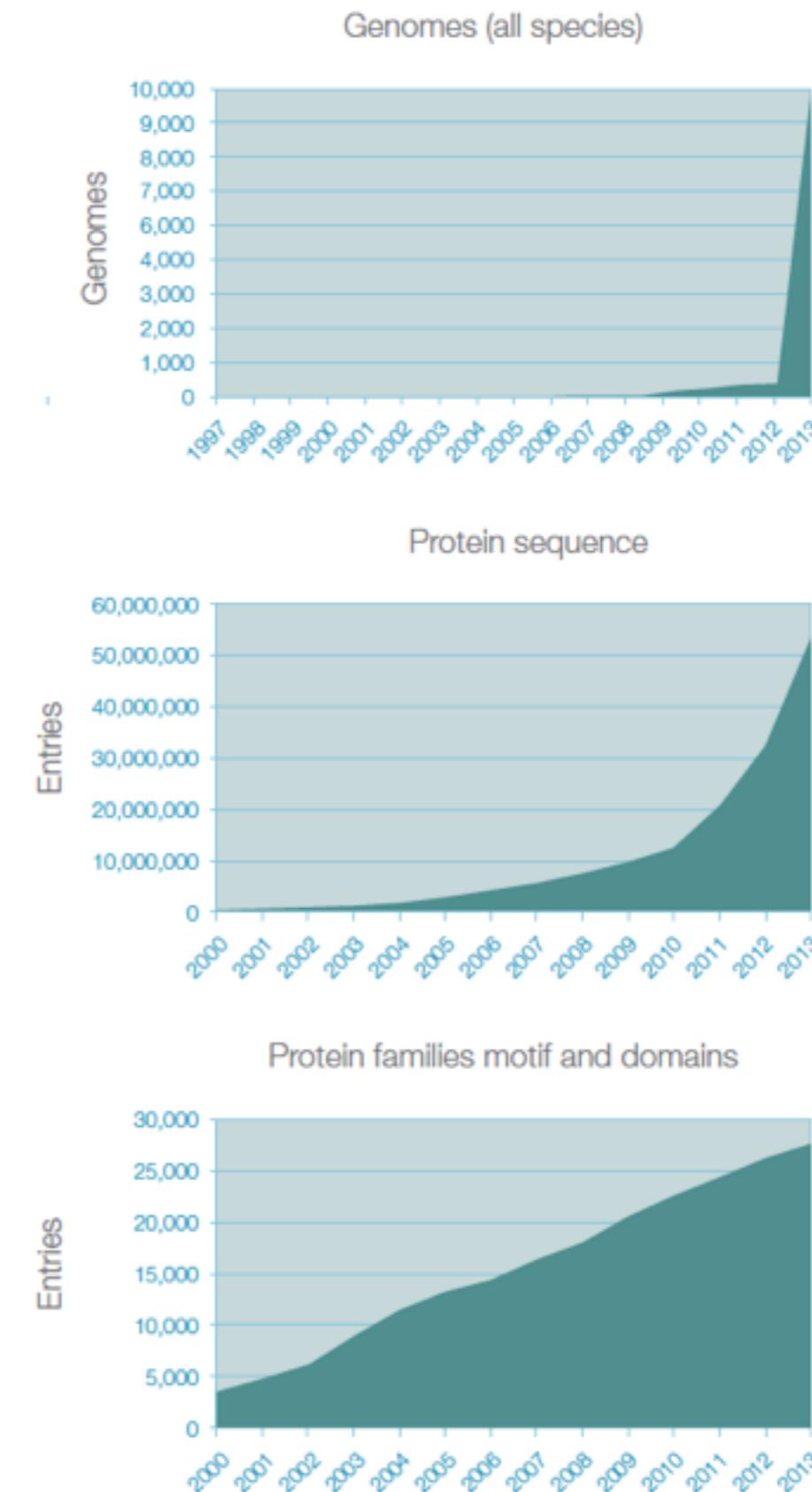
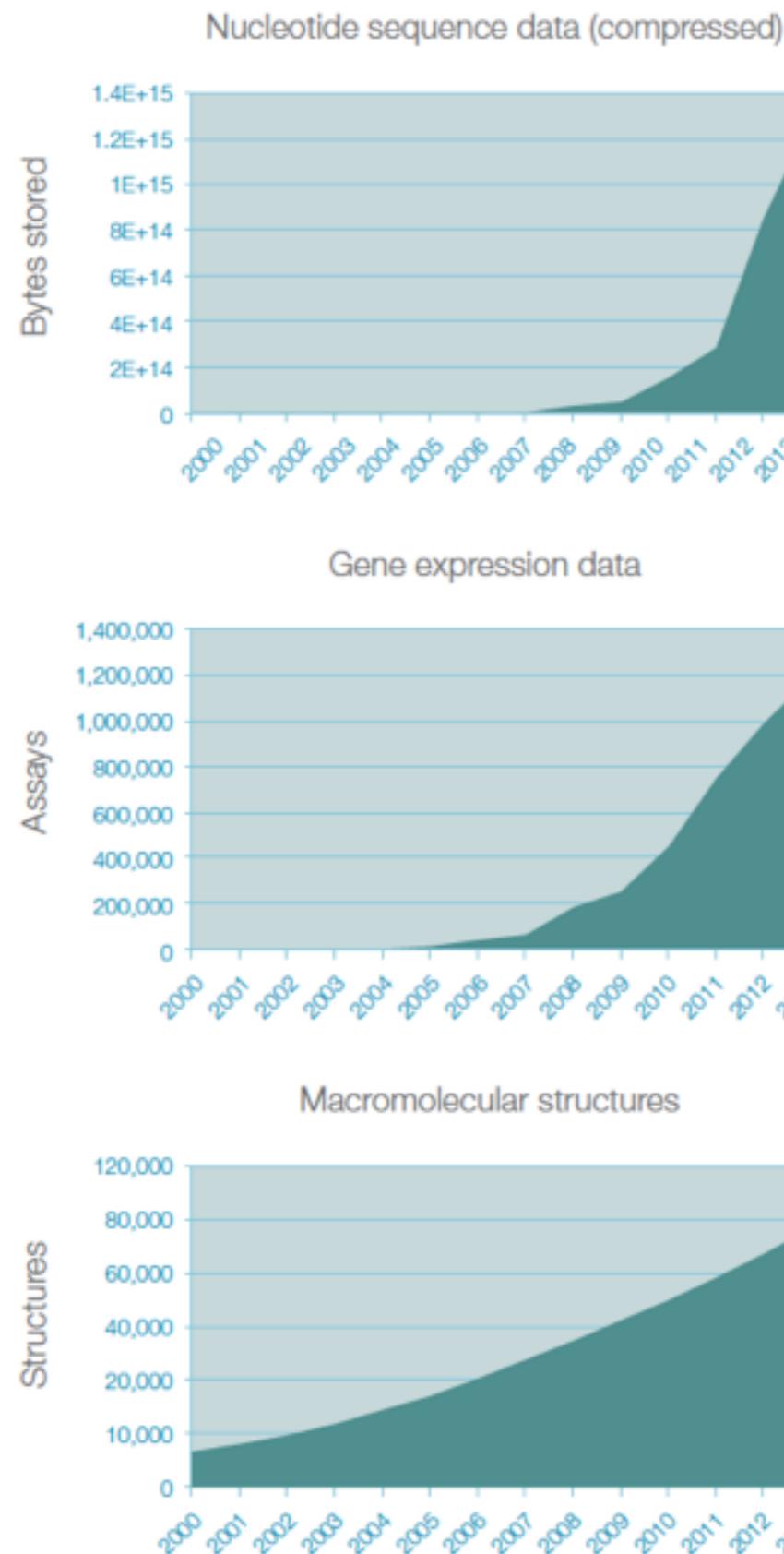


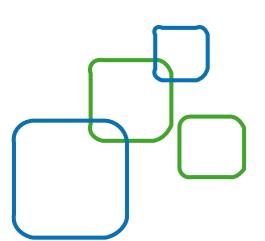
And other experimental data...





EMBL-EBI data resources growth

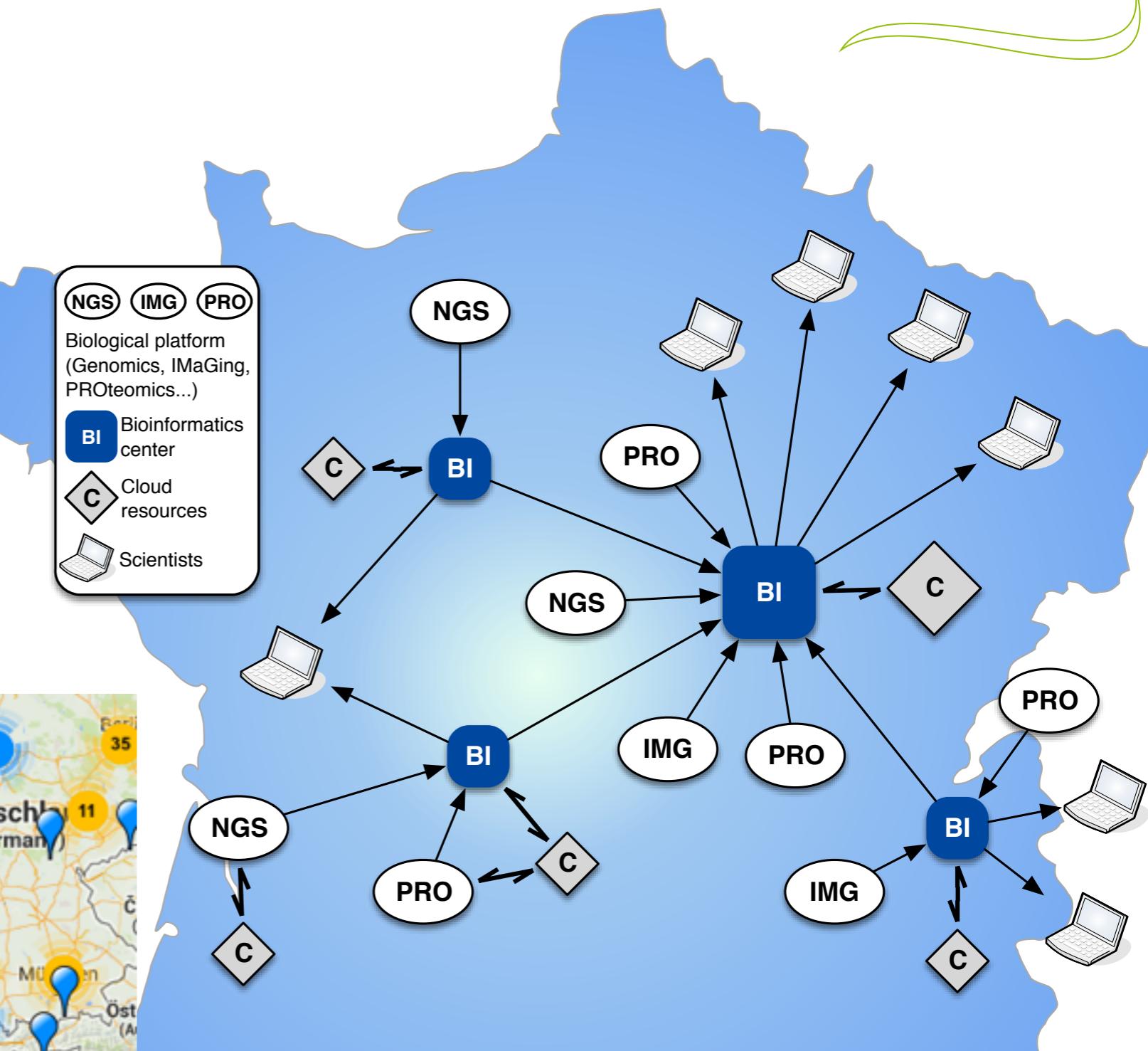




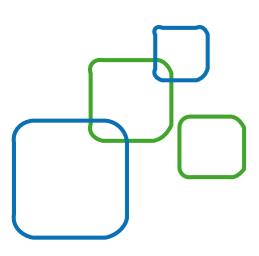
Plateformes Expérimentales en Biologie

Plateformes nationales (GIS IBISA)	Nb
Imagerie cellulaire	19
Génomique, Transcriptomique	16
Protéomique	13
Biologie structurale, biophysique	11

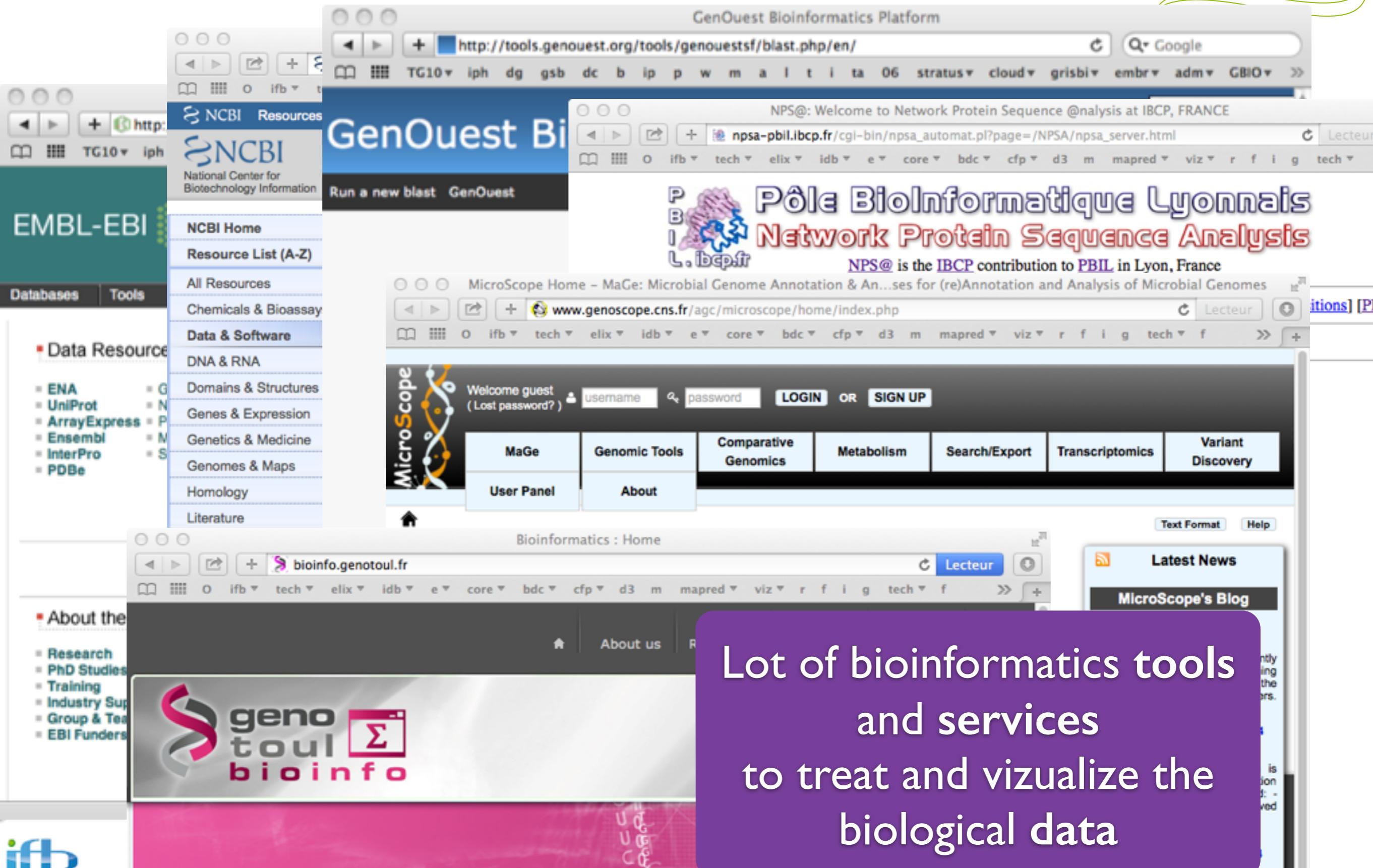
Localisation des plateformes NGS



Des sites intermédiaires permettent de répartir la charge en terme de stockage et de puissance de calcul tout en assurant une meilleure proximité avec les scientifiques



Infrastructures in Biology



NCBI Resources

NCBI National Center for Biotechnology Information

EMBL-EBI

Databases Tools

Data Resources

- ENA
- UniProt
- ArrayExpress
- Ensembl
- InterPro
- PDB
- Genomes & Maps
- Homology
- Literature

GenOuest Bioinformatics Platform

Run a new blast GenOuest

NPS@: Welcome to Network Protein Sequence @nalysis at IBCP, FRANCE

Pôle BioInformatique Lyonnais Network Protein Sequence Analysis

NPS@ is the IBCP contribution to PBIL in Lyon, France

MicroScope Home – MaGe: Microbial Genome Annotation & An...ses for (re)Annotation and Analysis of Microbial Genomes

Welcome guest (Lost password?)

username password LOGIN OR SIGN UP

MaGe Genomic Tools Comparative Genomics Metabolism Search/Export Transcriptomics Variant Discovery

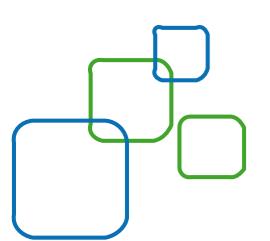
User Panel About

Bioinformatics : Home

Latest News

geno toul bioinfo

Lot of bioinformatics tools and services to treat and visualize the biological data



Bioinformatics Today



- Biological data are *big data*
 - 1552 online databases (NAR Database Issue 2014)
 - Institut Sanger, UK, 5 PB - Beijing Genome Institute, China, 7 sites, 20.6 PB

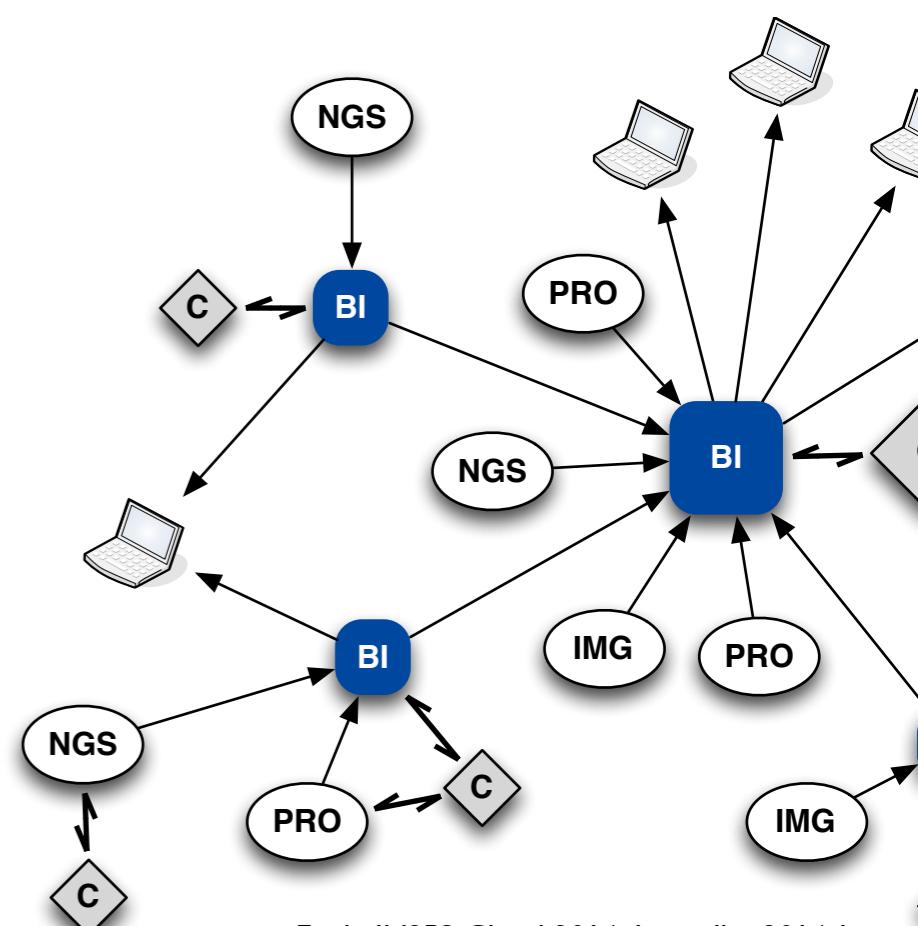
→ **Big data in many places**

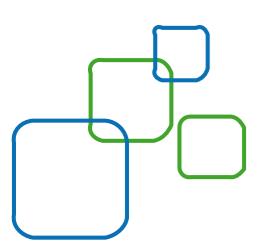
- Analysing such data became difficult
 - Scale-up of the analyses : gene/protein to complete genome/proteome, ...
 - Lot of different daily-used tools that need to be combined in workflows
 - Usual interfaces: portals, Web services,...

→ **Datacenters with ease of access/use**

- Distributed resources
 - Experimental platforms: NGS, imaging, ...
 - Bioinformatics platforms

→ **Federation of datacenters**





Cloud ?



● Essential characteristics

- On-demand self-service
 - No human intervention
- Broad network access
 - Fast, reliable remote access
- Rapid elasticity
 - Scale based on app. needs
- Resource pooling
 - Multi-tenant sharing
- Measured service
 - Direct or indirect economic model with measured use

● Deployment models

- Private
 - Single administrative domain, limited number of users
- Community
 - Different administrative domains with common interests & proc.
- Public
 - People outside of institute's administrative domain

● Hybrid

- Federation via combination of other deployment models

● Service models

- Software as a Service (SaaS)
 - Direct (scalable) hosting of end user applications
- Platform as a Service (PaaS)
 - Framework and infrastructure for creating web applications
- Infrastructure as a Service (IaaS)
 - Access to remote virtual machines
 - Machines with root access

NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

Special Publication 800-145

The NIST Definition of Cloud Computing

Recommendations of the National Institute of Standards and Technology

Peter Mell
Timothy Grance

I, Lyon

<http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>

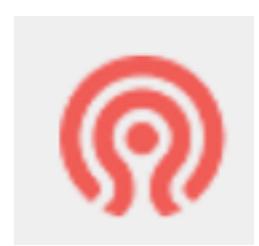
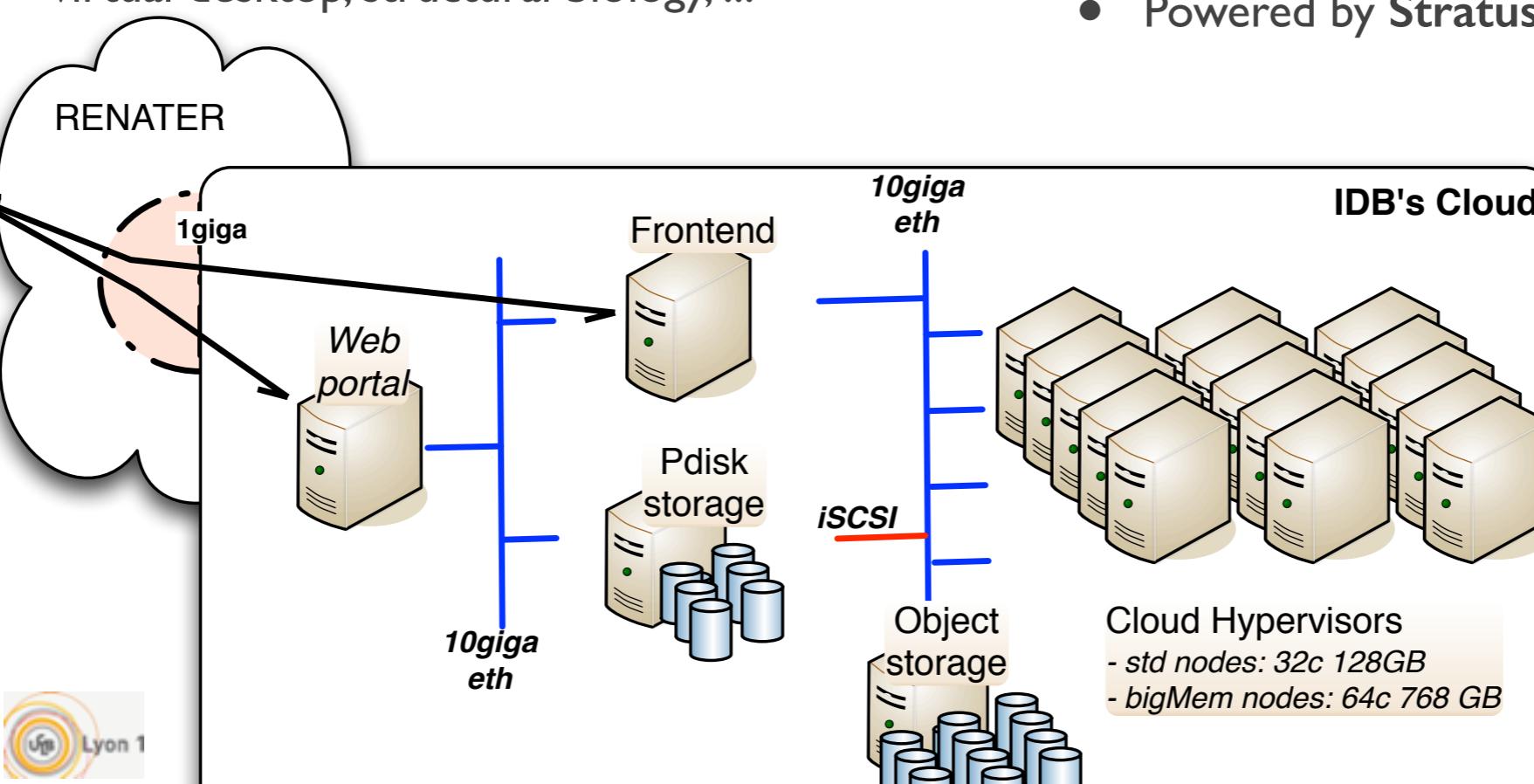
Cloud IDB

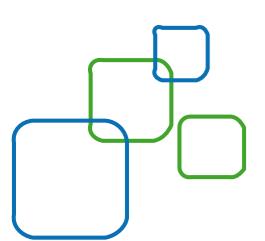
- **Cloud workbench for Biology**
 - Infrastructure Distributed for Biology
<https://idee-b.ibcp.fr/cloud.html>
 - Running since Sept. 2011
IBCP FR3302 CNRS-Univ. Lyon 1, Lyon, France
 - opened to Biology community
 - 14 bioinformatics appliances: Galaxy portal, standard compute nodes, proteomics, virtual desktop, structural biology, ...

- +70 users from all IFB regional centers PRABI 16, APLIBIO 28, RENABI-NE 13, -GO 7, -SO 2, -GS 5
- VMs up to 32cores-768GB RAM

- **Infrastructure**

- Compute +900cores +4TB ram
 - Standard nodes (32c-128GB)
 - Bigmen nodes (64c 768GB)
- Storage +250TB
 - Virtual disks, large-scale object storage (S3)
- Powered by StratusLab and CEPH



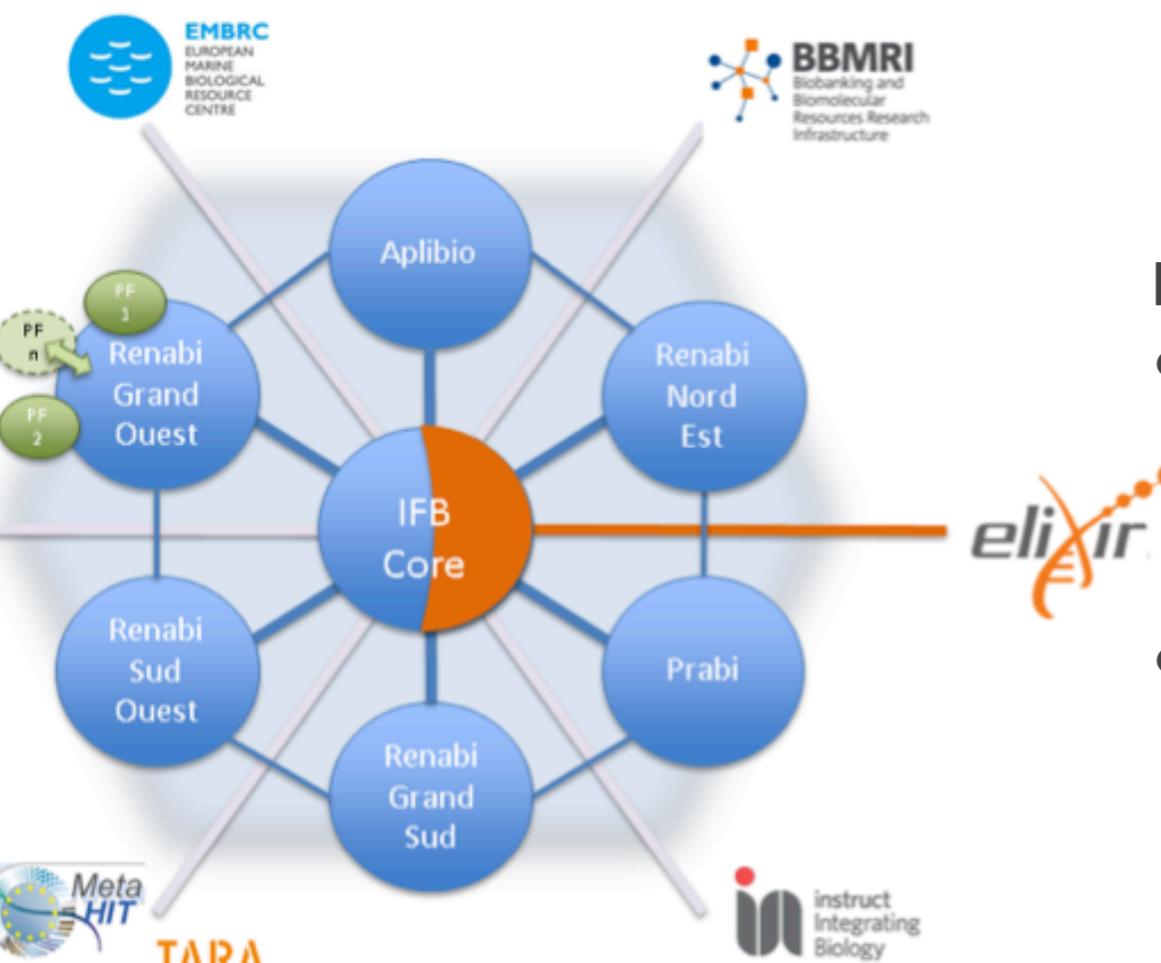


French Institute of Bioinformatics - IFB



Mission : to make available core bioinformatics resources to the national/international 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



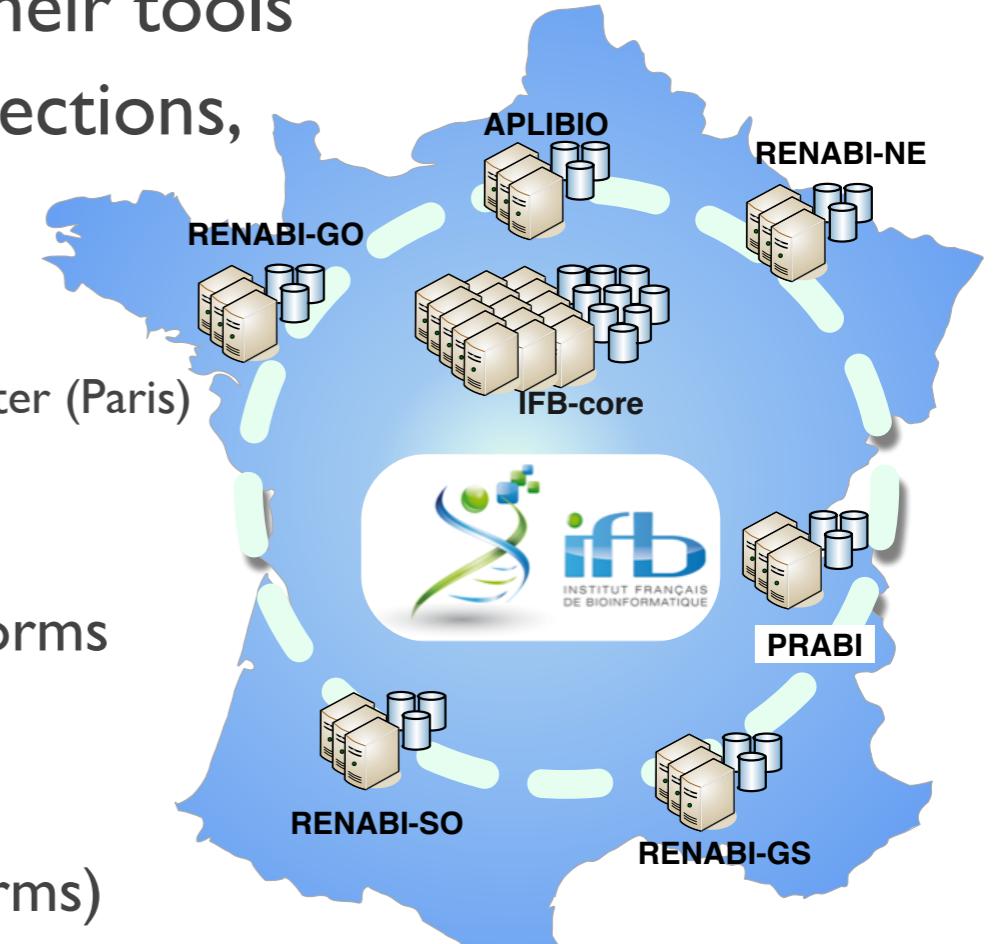
ELIXIR French Node

- optimizing the interactions and coordination between the national level and ELIXIR and other ESFRI infrastructures in biomedical and environmental field,
- promoting consistency and complementarities between the components offered by the ELIXIR French node and those of other European nodes

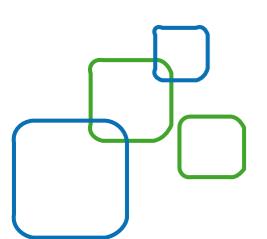


IFB e-Infrastructure

- **Support** : help members to deploy and use their tools
- **e-infrastructure**: hardware, biology data collections, bioinformatics tools
- **Academic cloud for life science**
 - a **core** ressource 'IFB-core' hosted at CNRS IDRIS SC center (Paris)
 - + **regional** resources
 - 6 regional bioinformatics centers with 2 clouds
 - 11,000 cores - 6 PB but +20 bioinformatics platforms
 - Create a **federation of clouds** for life science
- **Technical organization**
 - **GRISBI**: a national technical group (all national platforms)
 - Participation to **ELIXIR** task forces



Cloud Ressources	Location	# Compute Cores	# TB Storage	# TB RAM	Max VM size	Technology
IFB-core	CNRS-IDRIS, Paris	100	50	1	40c 256GB	StratusLab
IFB-core 2014	CNRS-IDRIS, Paris	4,000	500	-	96c 1TB	StratusLab
IFB-core 2015	CNRS-IDRIS, Paris	10,000	2,000	-	96c 2TB	StratusLab
idee-B	PRABI-IBCP, Lyon	1,000	380	4	64c 768GB	StratusLab
Genocloud	IFB-GO, Rennes	240	8	1	-	ONE



Extended cloud functionalities for bioinformatics

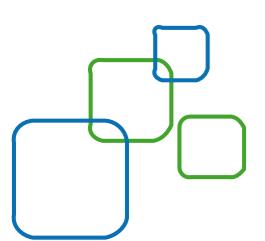


Native cloud services

- Authentication
- Virtual machine management
- Persistent disk service
- Client CLI
- etc.

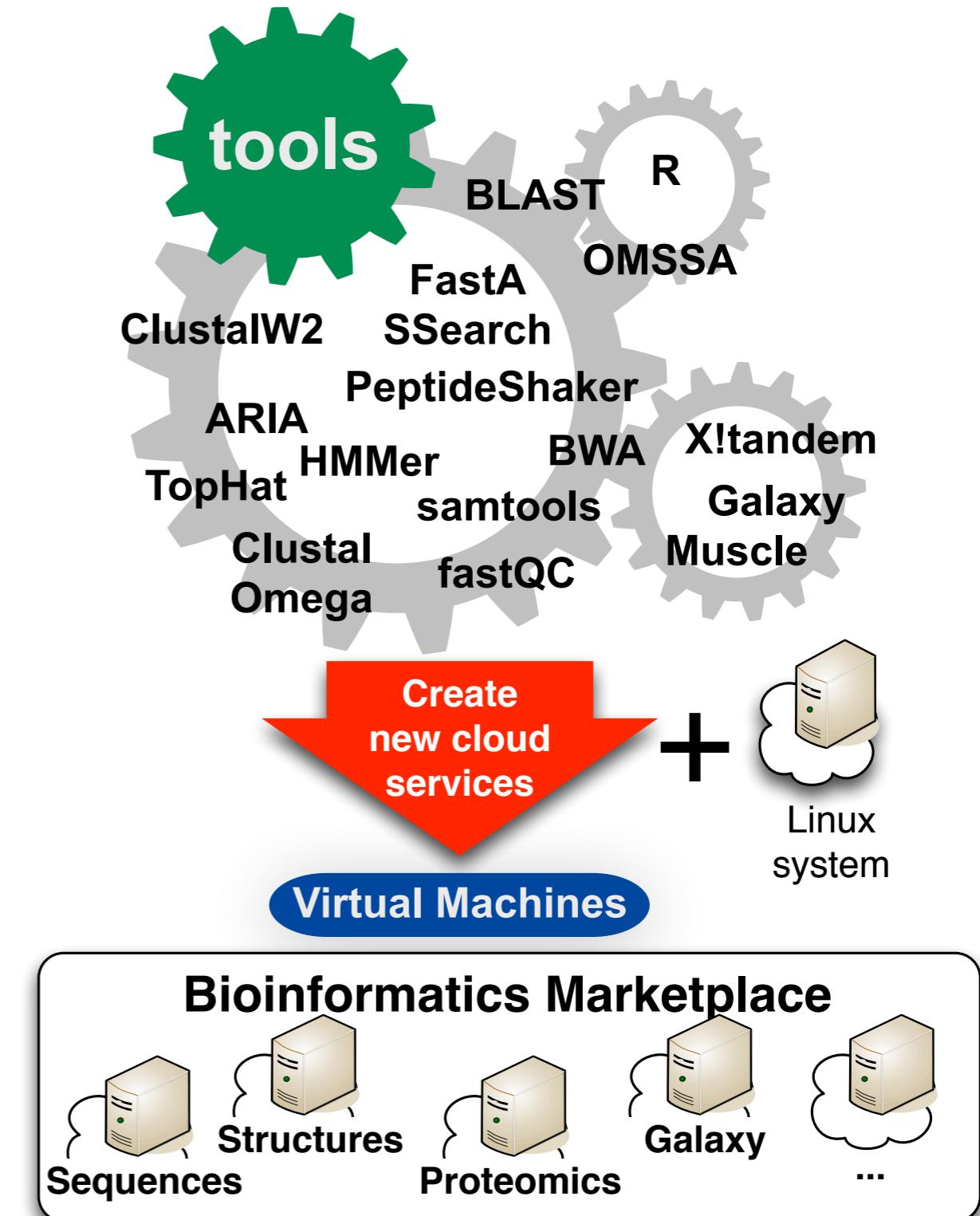


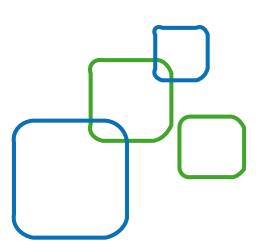
- **Bioinformatics appliances**
 - integrate bioinformatics tools and workflows
- **Bioinformatics marketplace**
 - focus on bioinformatics appliances
 - satisfy visibility constraints for some bioinformatics appliances (confidentiality)
- **Bioinformatics metadata “bio:tool”**
 - annotate appliances with attributes related to bioinformatics tools
 - help to select suitable bioinformatics appliances containing the required tools
- **Integrated Web interface**
 - VM & virtual disks management
 - filter bioinformatics appliances with “bio:tool”
- **CEPH storage backend**
 - large scale and distributed storage
 - reliable by replication
 - high-throughput IO
 - single unified storage cluster for all interfaces: block, object and file system



Bioinformatics cloud appliances

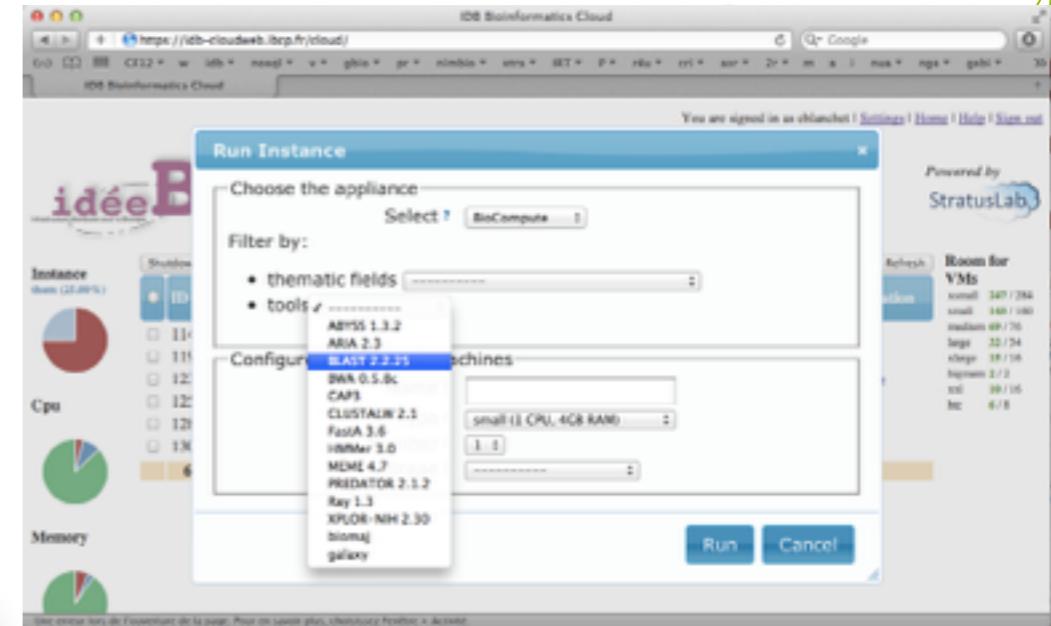
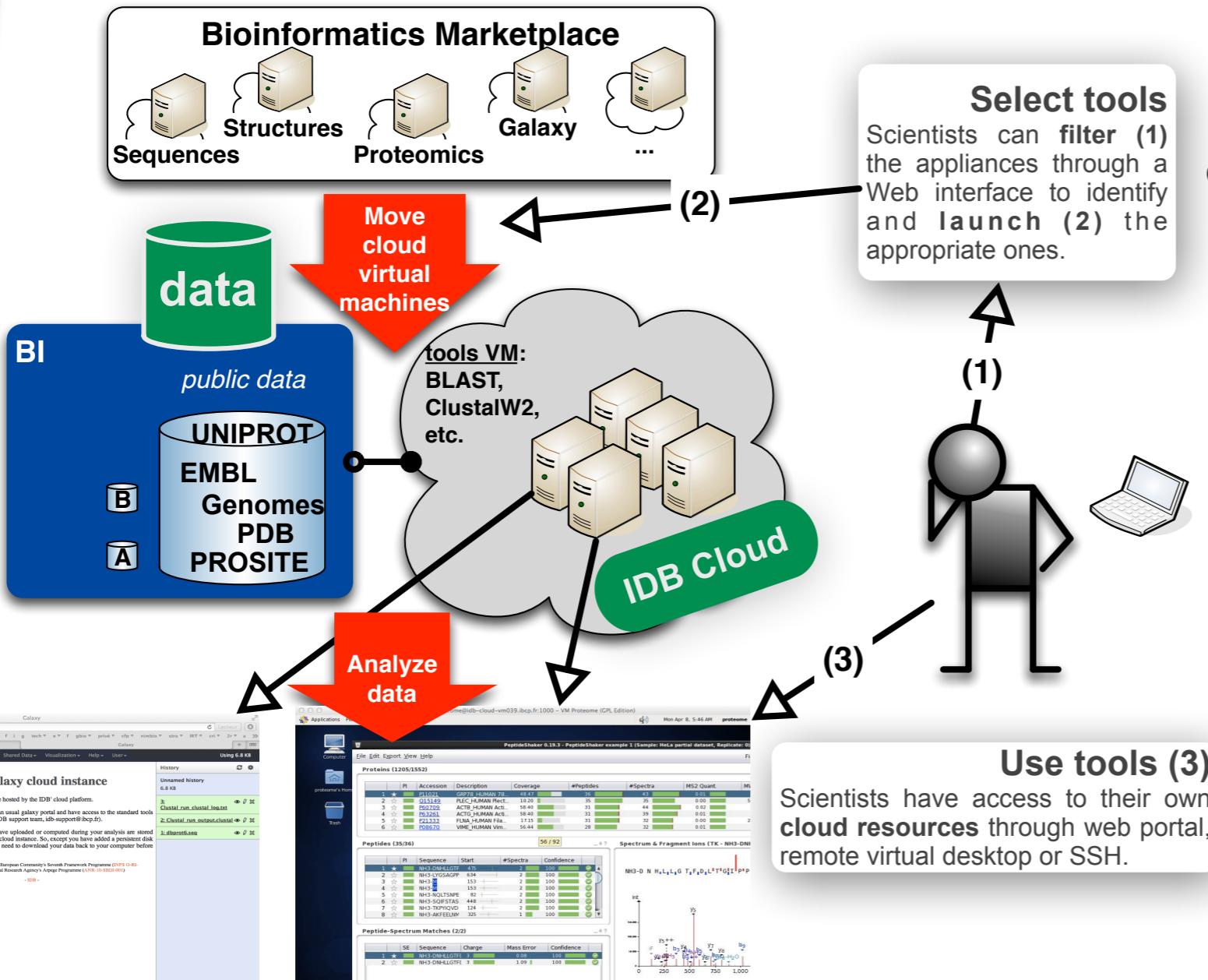
- Bioinformatics appliances are usual virtual machines
 - small : few GB, easy to convert in most virtualization formats
- Installed and pre-configured with bioinformatics tools
 - e.g. BLAST, Clustalw, ARIA, MEME, HMMer, TopHat, BWA, Samtools, etc.
- Recorded in a marketplace
 - devoted to bioinformatics



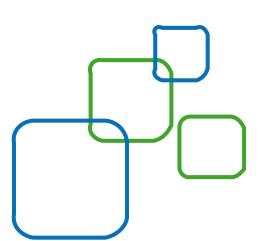


Run bioinformatics appliances

- Bioinformatics marketplace
 - both a virtual machines repository
 - Store life science VMs
 - and a catalogue
 - Help users to select the appropriate VM for their analysis



- Filter images with metadata related to bioinformatics
 - attribute <bio:tool> in VM manifests
 - scientists can select the appropriate appliance according to the tools required for their analyses
 - e.g. the BLAST tool
- Deploy on several clouds



Appliances page



The screenshot shows a web browser window with the following details:

- Title bar: IdéeB - Cloud
- Address bar: idee-b.ibcp.fr/appliances.html
- Toolbar: Back, Forward, Stop, Refresh, Home, and a download icon.
- Content area:
 - Header: Bioinformatics Cloud Appliances
 - Navigation menu: Databases | Tools | Cloud | Grid | Documentation | Sign in | Appliances | Cloud interface
 - Text: "We provide different bioinformatics cloud appliances ready-to-run. A cloud appliance is a predefined virtual machine with pre-installed tools and workflows. Most of these appliances can be associated with one of your virtual disk."
 - Text: "You can get a description of each appliance by *clicking on their name* in the list below. *To run your own instances*, click on the corresponding power button. Then, you will be redirected to a pre-filled form to create your instances."
 - List of appliances:
 - Bioinformatics compute node
 - Galaxy portal
 - Proteomics
 - Description: Bioinformatics virtual appliance for protein identification from mass spectrometry data. Contain OMSSA and X!Tandem tools, PeptideShaker and SearchGUI graphic interfaces.
 - ARIA (Ambiguous Restraints for Iterative Assignment)
 - RIO (Raw Ion Output)



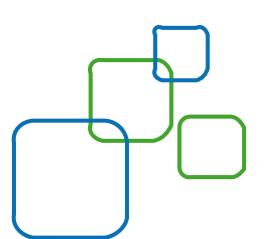
[Databases](#) | [Tools](#) | [Cloud](#) | [Grid](#) | [Documentation](#) | [Sign in](#)
[Appliances](#) | [Cloud interface](#)

We provide different bioinformatics cloud appliances ready-to-run. A cloud appliance is a predefined virtual machine with pre-installed tools and workflows. Most of these appliances can be associated with one of your virtual disk.

You can get a description of each appliance by *clicking on their name* in the list below. *To run your own instances*, click on the corresponding power button. Then, you will be redirected to a pre-filled form to create your instances.

- Bioinformatics compute node
- Galaxy portal
- Proteomics
 - Description: Bioinformatics virtual appliance for protein identification from mass spectrometry data. Contain OMSSA and X!Tandem tools, PeptideShaker and SearchGUI graphic interfaces.
- ARIA (Ambiguous Restraints for Iterative Assignment)
- RIO (Raw Ion Output)

- List of existing appliances
- Appliance description and doc
- Direct launch
- ‘Power’ button



Filter appliances with tools description

IDB Bioinformatics Cloud

db-cloudweb.ibcp.fr/cloud/

Cloud

Run Instance

Choose the appliance

Select ? ARIA2.3

Filter by:

- thematic fields -----
- tools -----

Genomics tools

- ✓ Molecular structural analysis
- Multiple Sequence Alignment
- Nucleotide and Protein sequence searching
- Public databases
- Sequence analysis

Configure your virtual machine

Name ?

Type ? small (1 CPU, 4GB RAM)

Number ? 1

Storage ? -----

Instance them (25.00%)

Cpu

Memory

idéeB infrastructure distribuée pour la Bioinformatique

Shutdown

ID

114 115 123 125 128 130

6

Une erreur lors de l'ouverture de la page. Pour en savoir plus, choisissez Fenêtre > Activité.

IDB Bioinformatics Cloud

https://idb-cloudweb.ibcp.fr/cloud/

You are signed in

Run Instance

Choose the appliance

Select ? BioCompute

Filter by:

- thematic fields -----
- tools -----

Configurable machines

Name ?

Type ? small (1 CPU, 4GB RAM)

Number ? 1

Storage ? -----

BLAST 2.2.25

ABYSS 1.3.2

ARIA 2.3

BWA 0.5.8c

CAP3

CLUSTALW 2.1

FastA 3.6

HMMer 3.0

MEME 4.7

PREDATOR 2.1.2

Ray 1.3

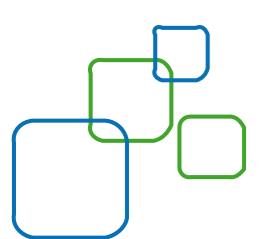
XPLOR-NIH 2.30

biomaj

galaxy

Run

Une erreur lors de l'ouverture de la page. Pour en savoir plus, choisissez Fenêtre > Activité.



A cloud driven through a simple web interface



Bioinformatics cloud

You are signed in as cblanchet | [Settings](#) | [Instances](#) | [Monitor](#) | [Help](#) | [Sign out](#)

 INSTITUT FRANÇAIS DE BIOINFORMATIQUE

 Hosted at 

Instance

Showing 1 to 6 of 6 entries

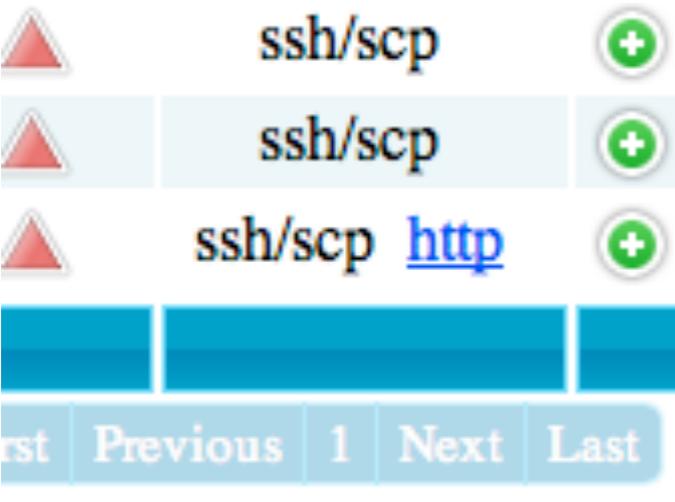
<input type="checkbox"/>	ID	Name	Appliance	CPU%	CPU	Mem.	#Storage	Access	
<input type="checkbox"/>	94	Public data source	BIO Data	3%	4	16	0	ssh http	
<input type="checkbox"/>	357	test2	RSAT 0.1	0%	4	8	0	ssh http	
<input type="checkbox"/>	365	proxy	Galaxy 4.1	0%	4	8	1	ssh http	
<input type="checkbox"/>	369	hotplug	BIO ComputeNode	0%	4	8	1	ssh	
<input type="checkbox"/>	385	testrel	Galaxy 4.2	0%	4	8	1	ssh http	
<input type="checkbox"/>	390	test-cleaner	Ubuntu 14.04	0%	2	8	0	ssh	

Show 25 entries

Storage

Cpu

Connection to VMs



Connection Information

You can connect to the **ssh/scp** port with:

```
ssh -A -p 20062 root@idb-cloud.ibcp.fr  
scp -P 20062 <file> root@idb-cloud.ibcp.fr:
```

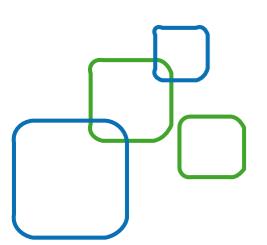
Close

A screenshot of a terminal window titled "maRacine — root@idb-cloud-vm050:~ — ssh — 27". The window shows a successful SSH login from the user "cblanchet" to the host "idb1:maRacine". The terminal displays the following command and output:

```
idb1:maRacine cblanchet$ ssh -A -p 20062 root@idb-cloud.ibcp.fr  
Last login: Mon May 20 15:05:28 2013 from mtl01-1-88-161-187-9.fbx.pr  
oad.net  
[root@idb-cloud-vm050 ~]# ls  
anaconda-ks.cfg  install  install.log.syslog  
cleaner.sh        install.log  mydisk  
[root@idb-cloud-vm050 ~]# 
```

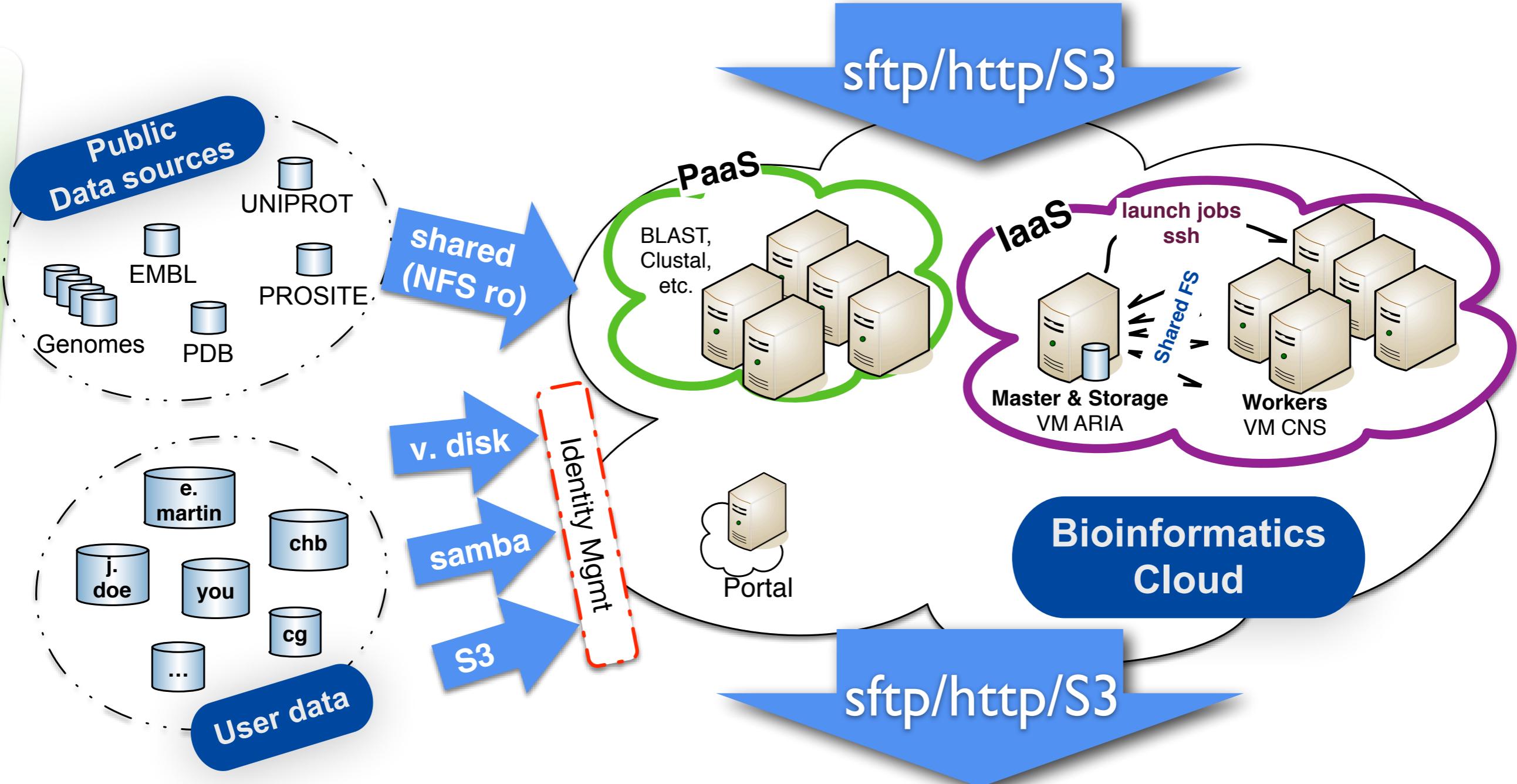
	Cloud IDB
	proteome@idb-cloud.ibcp
	GNOME
	1280x1024
	Enabled

	Cloud IFB
	proteome@VM IP
	GNOME
	800x600
	Enabled

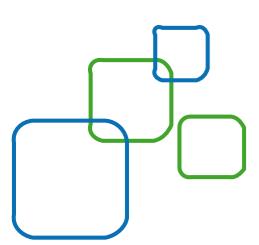


Cloud Storage for Biological Data

Upload your data



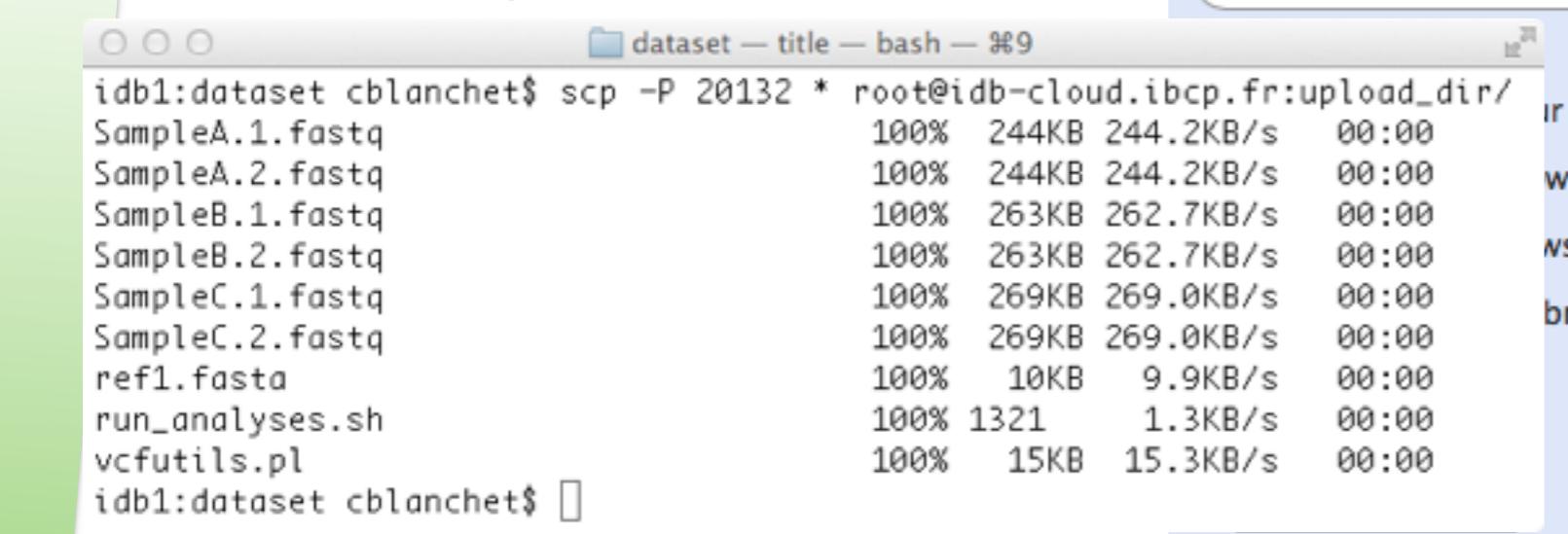
Get your results



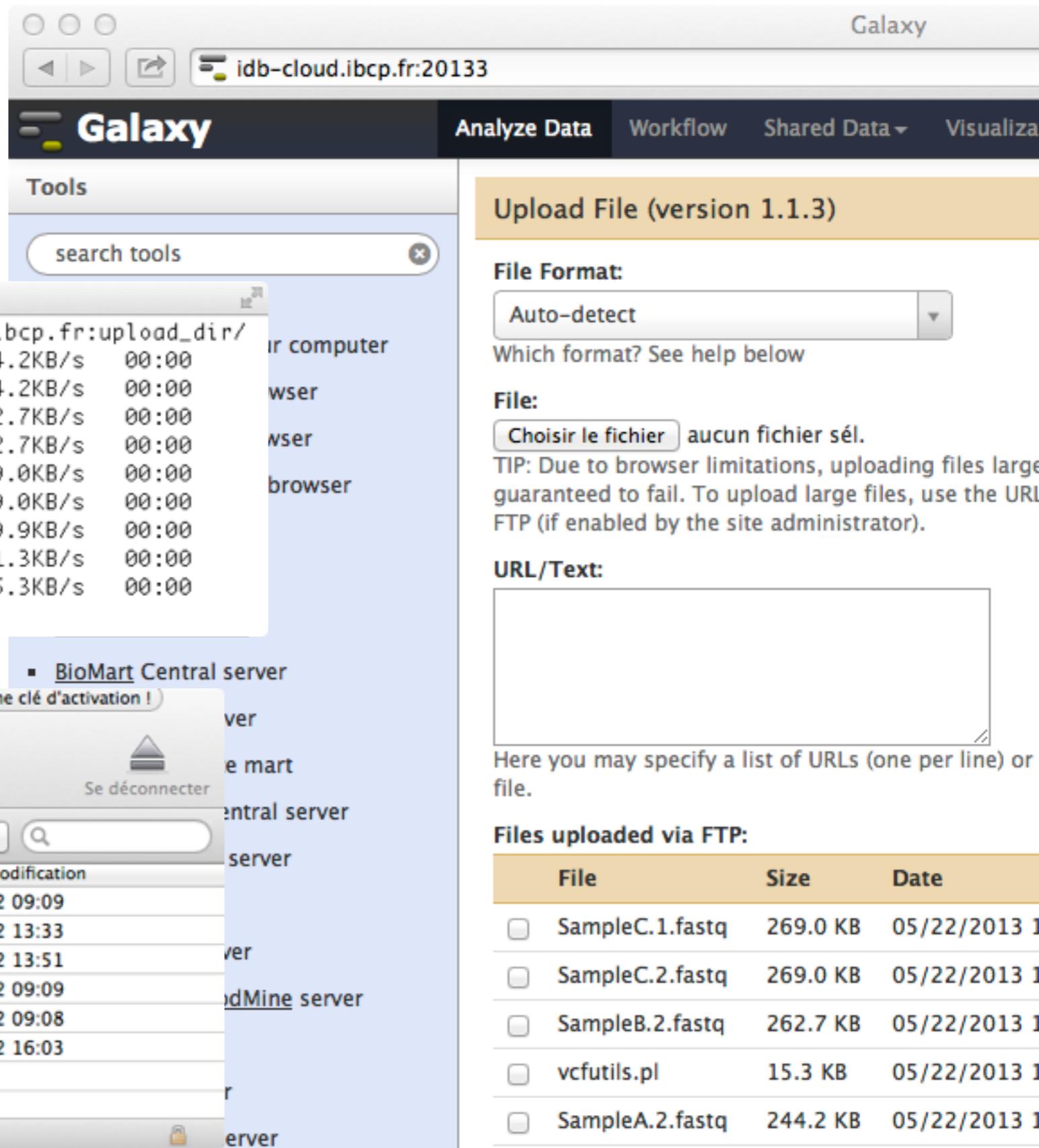
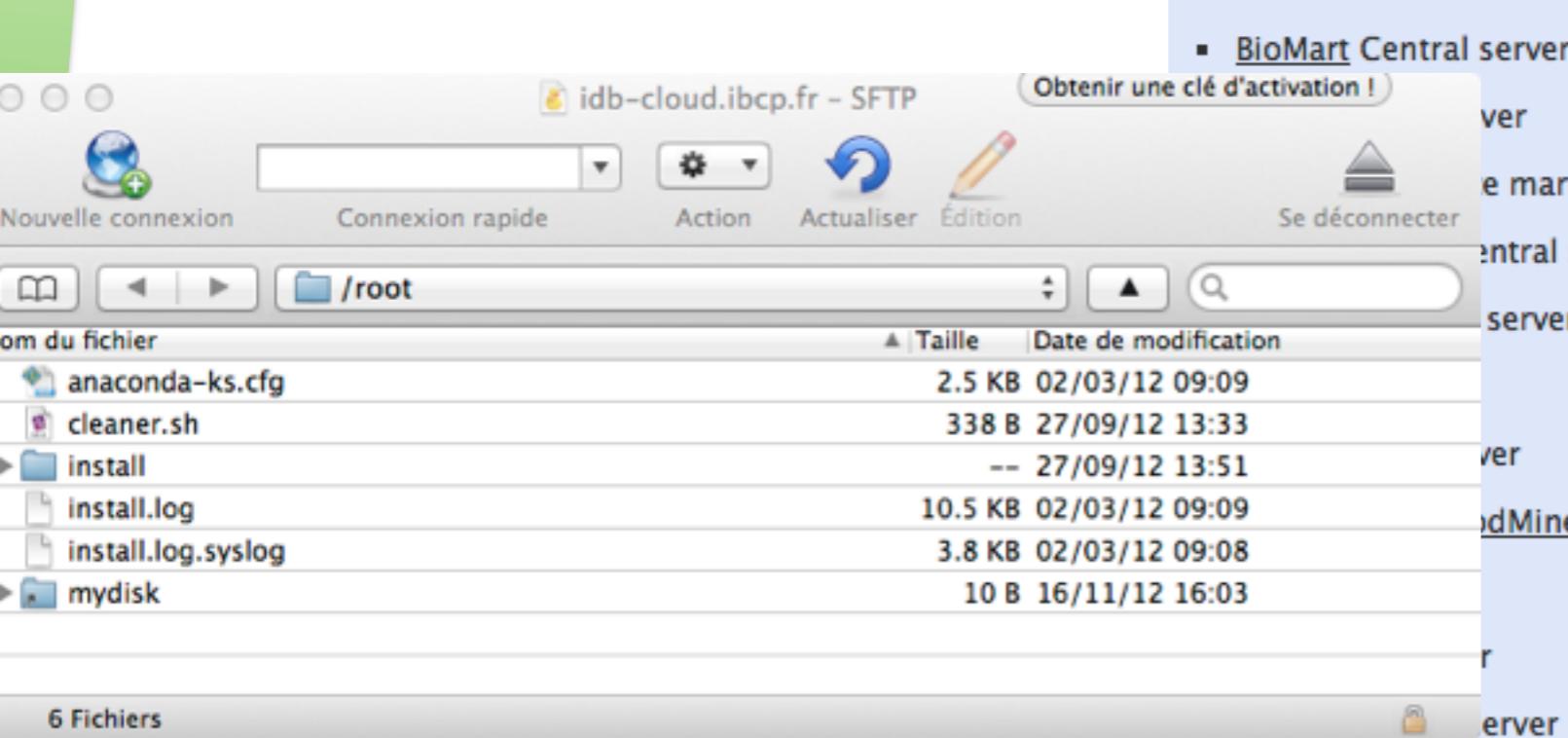
Exchanging data with VMs



- CLI ‘scp/sftp’
- GUI: Cyberduck, Transmit
- Integrated: Galaxy



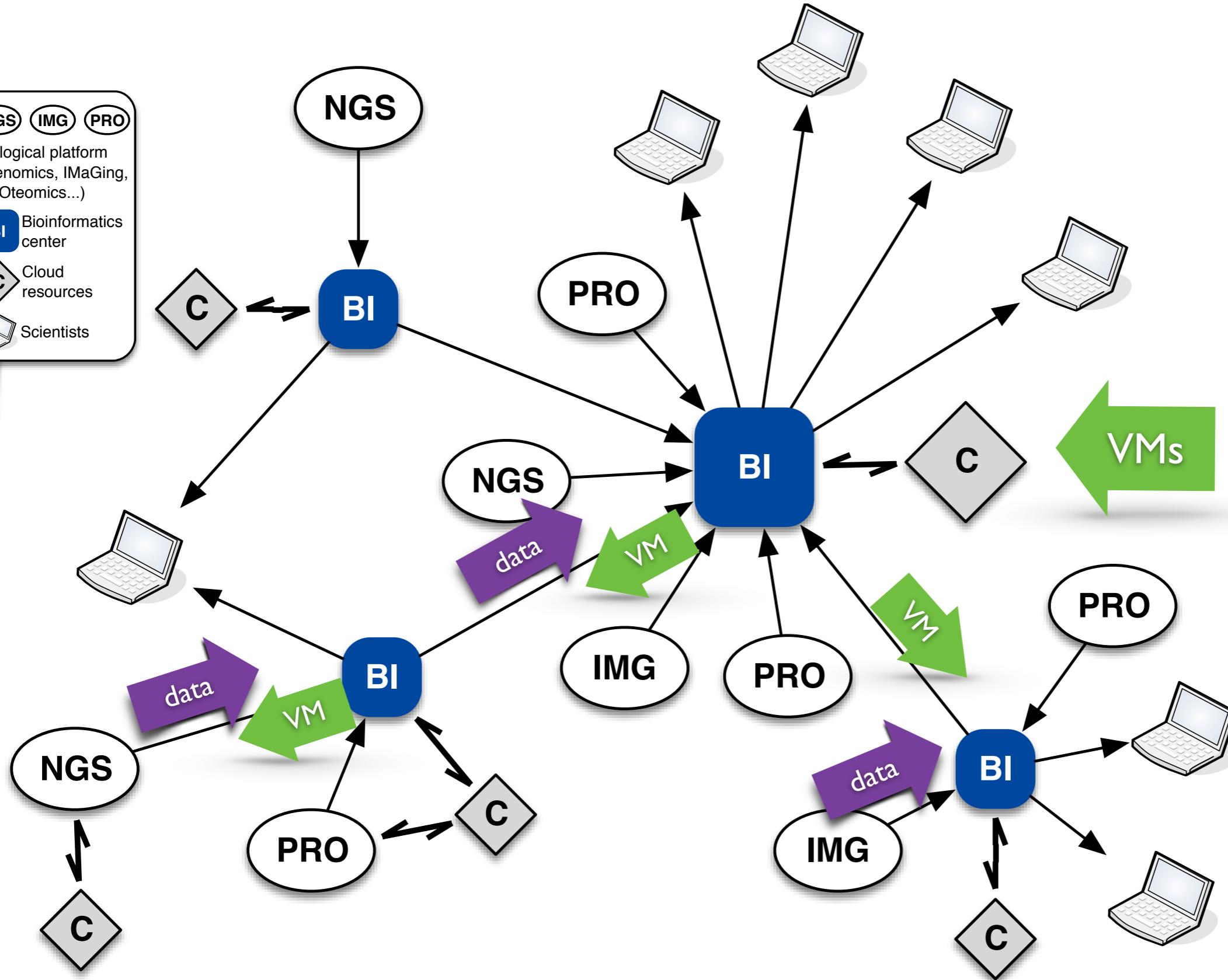
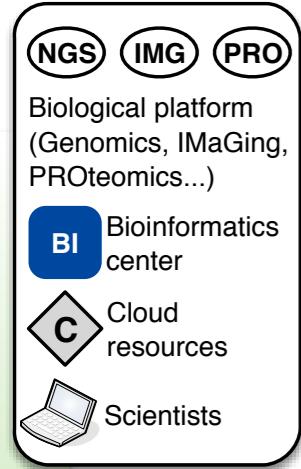
```
idb1:dataset cblanchet$ scp -P 20132 * root@idb-cloud.ibcp.fr:upload_dir/
SampleA.1.fastq
SampleA.2.fastq
SampleB.1.fastq
SampleB.2.fastq
SampleC.1.fastq
SampleC.2.fastq
ref1.fasta
run_analyses.sh
vcfutils.pl
idb1:dataset cblanchet$
```



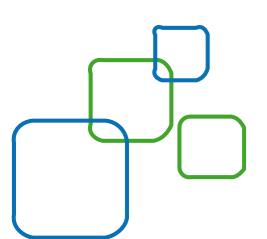
The screenshot shows the Galaxy web interface. At the top, the URL "idb-cloud.ibcp.fr:20133" is entered in the address bar. The main content area has a yellow header "Upload File (version 1.1.3)". It includes sections for "File Format" (set to "Auto-detect"), "File" (with a note about browser limitations), and "URL/Text". Below these, a large text input field contains the uploaded file names: SampleC.1.fastq, SampleC.2.fastq, SampleB.2.fastq, vcfutils.pl, and SampleA.2.fastq. A section titled "Files uploaded via FTP:" lists these same files along with their sizes and dates.

File	Size	Date
SampleC.1.fastq	269.0 KB	05/22/2013
SampleC.2.fastq	269.0 KB	05/22/2013
SampleB.2.fastq	262.7 KB	05/22/2013
vcfutils.pl	15.3 KB	05/22/2013
SampleA.2.fastq	244.2 KB	05/22/2013

Moving VMs vs Data



IFB
Bioinfor-
matics
marketplace
& VMs
repository



Case I: Standard Bioinformatics node

- appliance ‘Biocompute’
- Use your own instance(s)
- 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 data
 - Uniprot, EMBL, genomes, PDB, etc.
 - Automaticaly shared to the VMs
- Cluster mode
 - turn several instances in a single virtual cluster
 - shared file system
 - batch scheduling

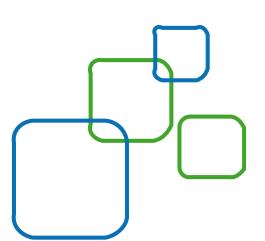
The screenshot displays the stratuslab web interface. At the top, there's a header with navigation links like Home, Endorsers, Query, Upload, and About. Below the header, a large blue cloud icon with the word "stratuslab" is visible. The main content area is titled "Metadata". It shows a table with one entry:

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

A detailed description follows the table:
Bioinformatics compute appliance built by CNRS IBCP-IDB. The following bioinformatics tools are installed and available from the command line: abyss, blast+, bioconductor, bowtie, bowtie2, bwa, cap3, clustal-omega, clustalw2, fasta36, gor4, hmm, meme, mmseq, multalin, muscle, predator, ray, R, samtools, simpa96, tophat, tophat2. To log in, use ssh with your key and the 'root' account. You have also access to the tools through a web portal, simply connect to your virtual machine with a standard web browser. The appliance can mount the cloud biological database repository (if available) by giving the corresponding contextualization parameters with the stratus-run-instance command. For example to run this appliance on the IBCP cloud, the command looks like:

```
maRacine — root@idb-cloud-vm050:~ — ssh — #7
idb1:maRacine cblanchet$ ssh -A -p 20062 root@idb-cloud.ibcp.fr
Last login: Mon May 20 15:05:28 2013 from mtl01-1-88-161-187-9.fbx.pr
oxid.net
[root@idb-cloud-vm050 ~]# ls
anaconda-ks.cfg  install  install.log.syslog
cleaner.sh        install.log  mydisk
[root@idb-cloud-vm050 ~]#
```

On the right side of the terminal window, there is some explanatory text about the command and its options.



Case 2: Cloud Galaxy portal for NGS analyses

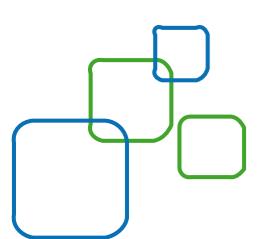
- Analyse NGS data
 - portal Galaxy is widely used in the community
 - connected to large public data: sequences and indexes
 - large user data (GBs)
- Preserve workflows and results (cloud virtual disk)
- Different domain-specific instances (RNAseq, ChIPseq, etc.)
- For training: create a special instance derived from the main instance but with dedicated datasets
- Help the integration of monthly updates

The screenshot shows the Galaxy web interface for an IDB Galaxy cloud instance. At the top, there's a navigation bar with links like 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. Below the navigation is a search bar labeled 'search tools'. On the left, there's a sidebar with links for 'Get Data', 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', and 'Convert Formats'. The main content area has a title 'IDB Galaxy cloud instance' and a welcome message: 'Welcome to your Galaxy instance hosted by the IDB's cloud platform.' Under the 'Usage' section, it says: 'This appliance is configured with the well-known GALAXY portal. You connect to it with a standard web browser : simply follow the link on the main IDB cloud interface. It can be used as an usual galaxy portal and you have access to pre-installed standard bioinformatics tools (for new tools, send a request to IDB support team: idb.support@ibcp.fr)'. To the right, there's a 'History' panel titled 'unnamed history' showing four items: 'Clustal run clustal log.txt', 'Clustal run output.clustal', and 'dbprot6.seq', each with a download icon. Above the Galaxy interface, there's another window titled 'Metadata' from 'stratuslab' showing details about a 'Galaxy portal' instance. The 'http' link in both the Galaxy interface and the Metadata window is circled in red.

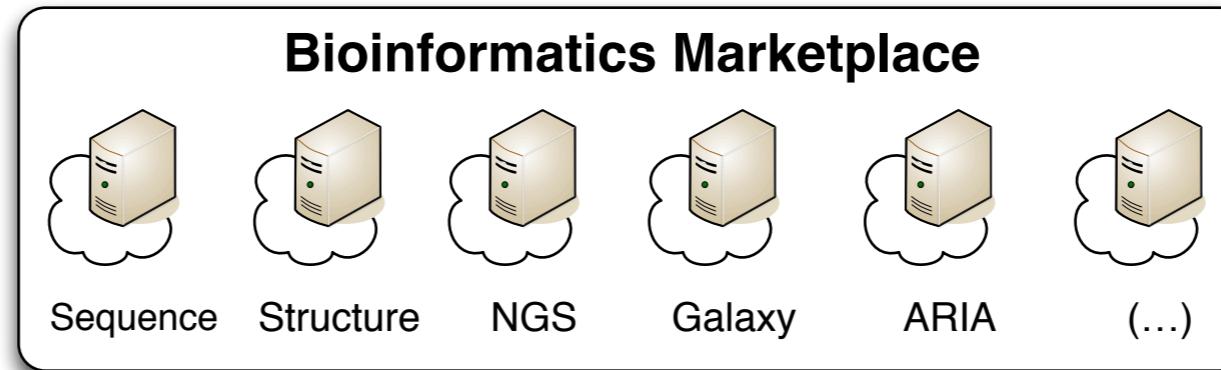
The screenshot shows the 'Metadata' page from the stratuslab portal. At the top, there's a header with links for 'Home', 'Endorsers', 'Query', 'Upload', and 'About'. Below the header, there's a logo for 'stratuslab'. The main content area has a title 'Galaxy portal' and a table with the following data:

Endorser	christophe.blanchet@ibcp.fr
Identifier	GOqP1arAKmWzR2PB-tCEdShbu7n
Created	2013-11-21T15:14:39Z
Kind	machine

Below the table, there's a detailed description: 'Bioinformatics gateway appliance configured with the GALAXY portal, built by CNRS IBCP-IDB. You will have access to the pre-installed bioinformatics tools through the web portal. Connect to your own Galaxy portal with a standard web browser, simply follow the link on the main IDB cloud interface. For more details, see relative documentation on the Idee-B site (<http://idee-b.ibcp.fr>).'. There's also a 'More...' link at the bottom.



Run your Galaxy Portal on Cloud



- Stay Connected Standard Data &
 - User data: upload datafiles or attach pdisk
 - Reference databases: mount biodata servers
 - Tools: use pre-installed ones or install yours

Launch Instances

Create Instance

Choose The Appliance

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

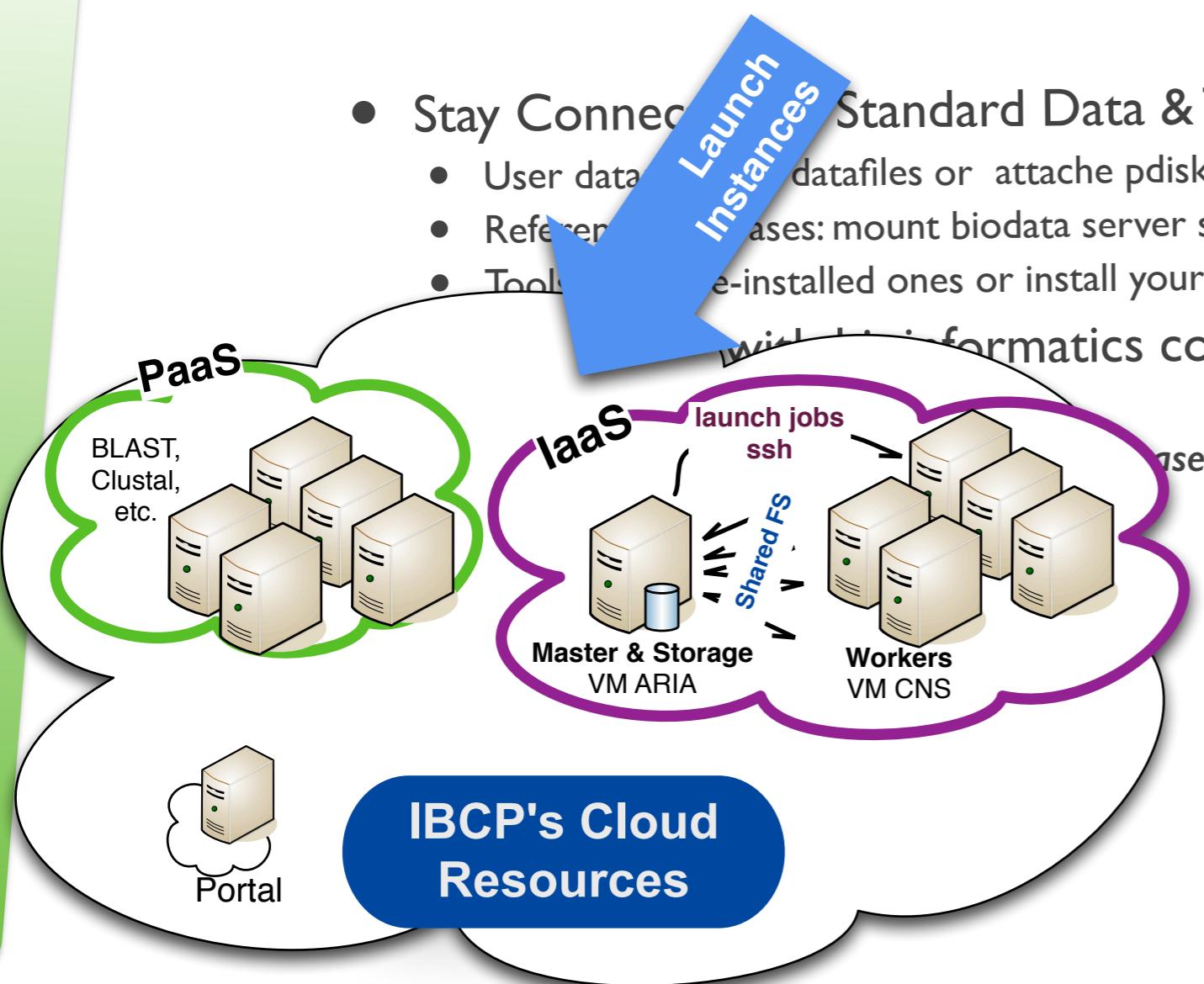
Configure Your Virtual Machines

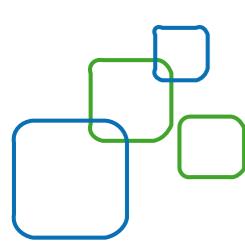
Name ? ma VM galaxy
Unique ?
Type ? large (4 CPU, 16GB RAM)
Number ? 1
Create appliance ?

Configure Your Storage

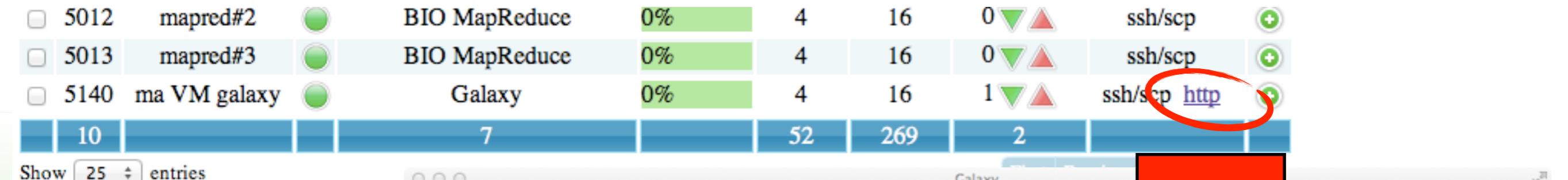
Persistent disk ? stockage galaxy
OR Volatile disk size ?

Create





Connect to your Galaxy Portal



5012 mapred#2 BIO MapReduce 0% 4 16 0 ssh/scp +
5013 mapred#3 BIO MapReduce 0% 4 16 0 ssh/scp +
5140 ma VM galaxy Galaxy 0% 4 16 1 ssh/scp [http](#) +

Show 25 entries

Galaxy

idb-cloud.ibcp.fr:20023

IDB Bioinformatics Cloud

Galaxy Analyze Data Workflow Shared Data Visualization Help Using 4.9 GB

Tools search tools

Get Data Send Data ENCODE Tools Lift-Over Text Manipulation Filter and Sort Join, Subtract and Group Convert Formats Extract Features Fetch Sequences Fetch Alignments Get Genomic Scores Operate on Genomic Intervals Statistics Wavelet Analysis Graph/Display Data Regional Variation Multiple regression Multivariate Analysis Evolution Motif Tools Multiple Alignments Metagenomic analyses FASTA manipulation

IDB Galaxy cloud instance

Welcome to your Galaxy instance hosted by the IDB's cloud platform.

Usage

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

Data management

Data persistency between different runs

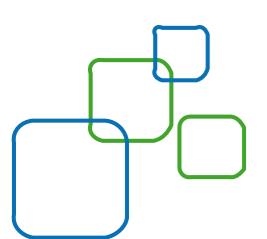
Keep in mind that except you have added a persistent disk at the launch of this appliance, the data you have uploaded or computed during your analysis are stored on the *volatile disk* of this current cloud instance. So **these data will be removed** when you will terminate this cloud instance. You need then to download your data back to your computer before to shutdown this portal. When this appliance is run in association with one of your virtual disks, the history and the data of your Galaxy portal is stored for a further execution. Don't forget to attach your favorite virtual disk in the 'Create instance' form.

Large files

(!) Don't forget to sign in with the pre-defined user : `user@cloud.idb.fr` (password `idbuser`).

Galaxy provides users with the 'FTP upload method' to upload large files. On the IDB's cloud

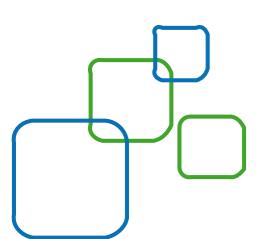
History
Unnamed history 4.9 GB
23: Clustal run clustal.log.txt
22: Clustal run output.clustal
21: dbprot6.seq
20: A.bam
19: A2.fq
18: A1.fq
17: SampleA.2.fasta
16: SampleC.1.fasta
15: SampleC.2.fasta
14: SampleB.1.fasta
13: SampleA.1.fasta
12: SampleB.2.fasta
11: ref1.fasta
10:



Advantages of Cloud for Galaxy



- Added value of cloud for Galaxy,
 - for scientific analyses: user-specific resources, isolated, different domain-specific instances (RNAseq, ChIPseq, Variants, ...)
 - for training: create a special instance derived from the main but with dedicated datasets
 - Examples of training with Galaxy: Mai 2013 Galaxy Lille, Nov 2013 Aviesan Bioinformatics School
 - For integration of monthly updates
 - for development & operations (DevOps): different versions at the same time
- Bioinformatics cloud (e.g. IDB)
 - Tightly connected to existing bioinformatics resources
 - Linked to public biological databases
 - In collaboration with the French Institute of Bioinformatics



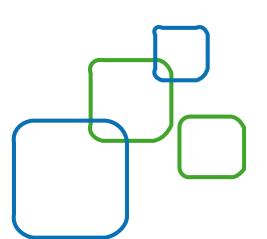
Case 3: Proteomics virtual desktop

- Motivation
 - Collaboration with a mass spectroscopy platform
 - Running out of space on their local resources
- Protein identification tools
 - Mass experimental data
 - Reference databases : nr, Swiss-Prot
 - Reference screening tools: OMSSA, X!Tandem
- User interface
 - Remote Virtual Desktop (NX)
 - Reference GUIs
 - SearchGUI
 - PeptidShaker

The screenshot shows the stratuslab Metadata interface. At the top, there's a navigation bar with links for Home, Endorsers, Query, Upload, and About. Below it is a search bar with the term "proteomics". A large blue cloud icon with the word "stratuslab" is on the left. The main content area displays a single entry for a proteomics application. The entry includes fields for Endorser (christophe.blanchet@ibcp.fr), Identifier (P0CtUXnTejwxUbam6U1s7uuah3), Created (2014-04-04T13:39:36Z), Kind (machine), and a detailed description of the Bioinformatics virtual appliance for protein identification from mass spectrometry data, mentioning OMSSA, X!Tandem, PeptideShaker, and SearchGUI tools, with a link to the IDB website.

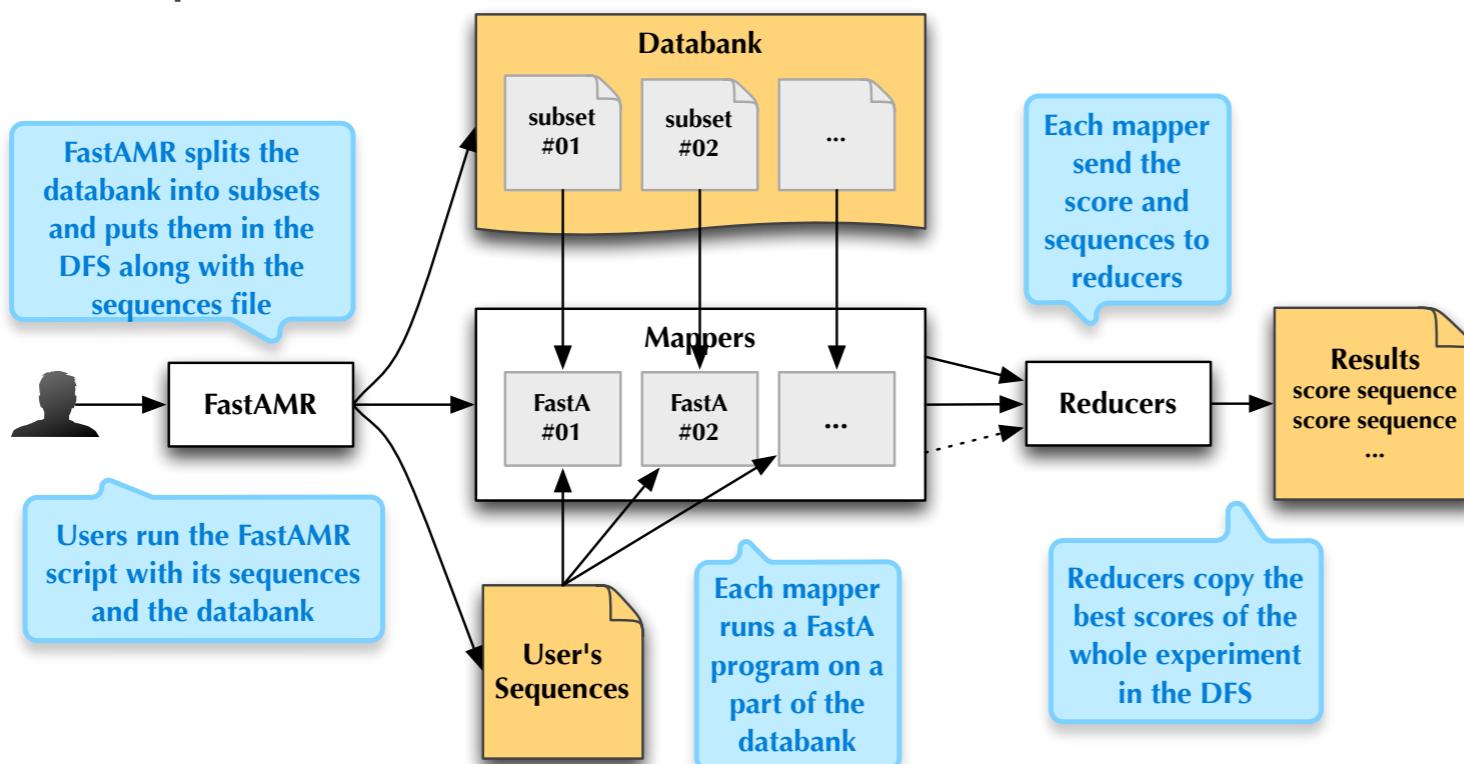
The screenshot shows the NX desktop environment. On the left is a dark blue desktop panel with icons for Computer, proteome's Home, and Trash. The main window is titled "PeptideShaker 0.19.3 - PeptideShaker example 1 (Sample: HeLa partial dataset, Replicate: 0)". It contains several tabs: "Proteins (1205/1552)", "Peptides (35/36)", "Spectrum & Fragment Ions (TK - NH3-DNF)", and "Peptide-Spectrum Matches (2/2)". The "Proteins" tab shows a table with columns: #, RI, Accession, Description, Coverage, #Peptides, #Spectra, MS2 Quant., and Mw. The "Peptides" tab shows a table with columns: #, RI, Sequence, Start, #Spectra, and Confidence. The "Spectrum & Fragment Ions" tab displays a mass spectrum plot with peaks labeled with peptide sequences like NH3-DNHLGLTF, NH3-LYGSAGPP, etc. The "Peptide-Spectrum Matches" tab shows a table with columns: #, SE, Sequence, Charge, Mass Error, and Confidence.

OMSSA
x!



Case 4: Hadoop for Life Science

- Provide turnkey virtual machine with pre-configured mapreduce framework
 - Accelerate biological bigdata analysis
 - Hadoop MapReduce 1.0.4
- Appliances (2)
 - provide standard hadoop: including mapreduce and HDFS
 - with integrated bioinformatics tools
- Example of sequence similarity searching
 - FastA & SSearch
 - deploy database of sequences in HDFS
 - compare each structure to others



Developed in the context of the French project
MapReduce, ANR ARPEGE

BIO MapReduce

Endorser: clement.gauthey@ibcp.fr
Identifier: J46wxrwGLdnoSskmb0JlfGv8UpY
Created: 2013-05-17T11:13:08Z
Kind: machine

This appliance provides an easy way to deploy a Hadoop MapReduce cluster (v1.0.4) with pre-installed bioinformatics tools such as FastA. You just need to run the bash script `hadoop-create-cluster` with a nodes list and an username parameters and wait few minutes until the process is completed. Then you can login to the user account and submit your Hadoop jobs or interact with Hadoop filesystem. You can extend a current cluster by submitting a list of new nodes to the script. A FastA MapReduce example is also provided under the directory `/usr/local/share/fasta`. (Created for the French project MapReduce, ANR ARPEGE, 2010-2013, mapreduce.inria.fr)

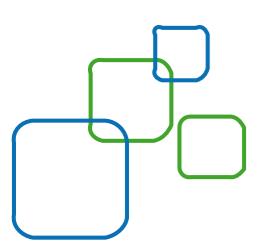
[More...](#)

Hadoop MapReduce

Endorser: clement.gauthey@ibcp.fr
Identifier: BtU7uNM5UT1haUigVS7xySI2rr
Created: 2013-05-17T09:33:52Z
Kind: machine

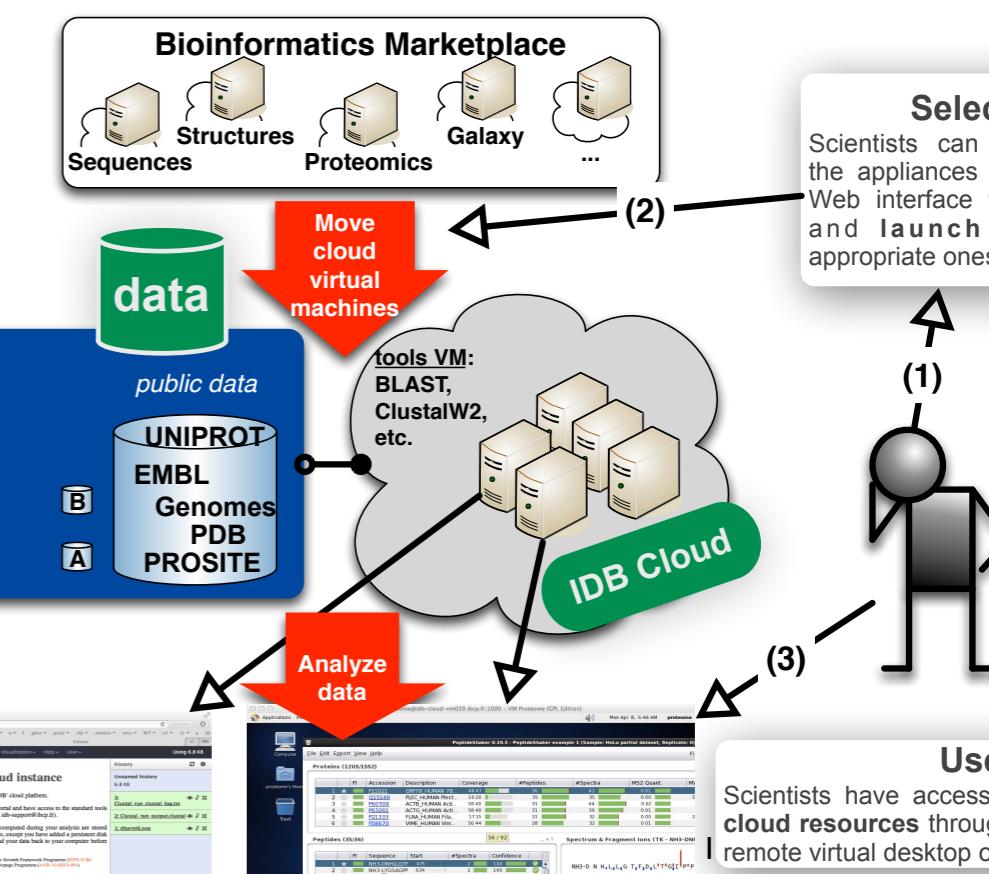
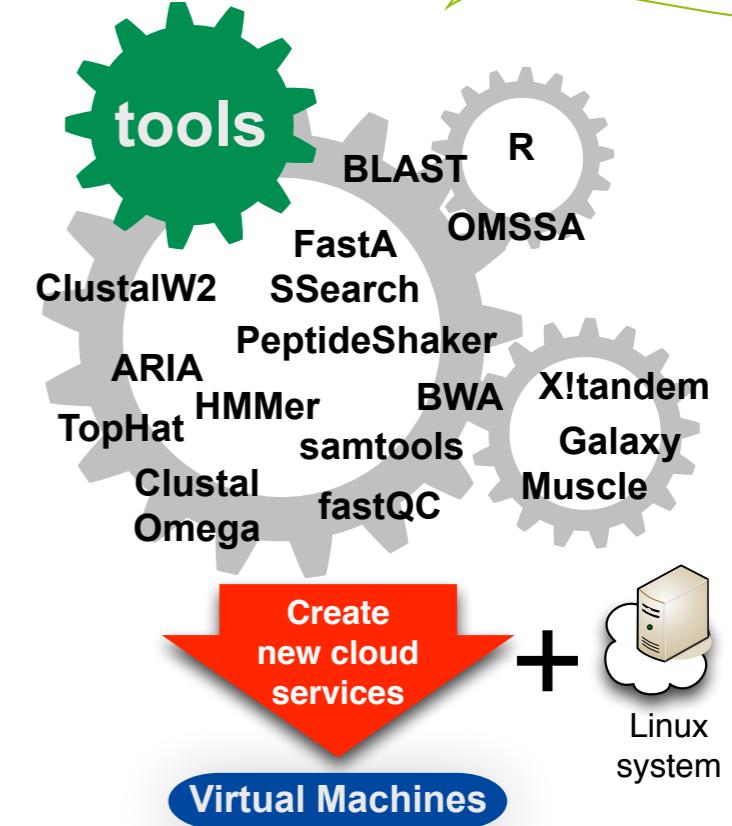
This appliance provides an easy way to deploy an Hadoop MapReduce cluster (v1.0.4). You just need to run the bash script `hadoop-create-cluster` with a nodes list and an username in parameters and wait few minutes until the process is completed. Then you can login to the user account and submit your Hadoop jobs or interact with Hadoop filesystem. Enjoy! In addition, you can extend a current cluster by submitting a list of new nodes to the command `hadoop-add-node`. (Created for the French project MapReduce, ANR ARPEGE, 2010-2013, mapreduce.inria.fr)

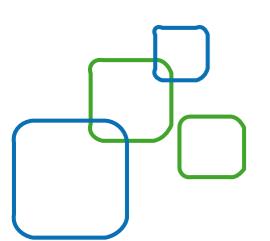
[More...](#)



Cloud it be done ?

- IFB's cloud for life science simplify access to biological data and tools
 - integrate tools and pipelines in turnkey cloud appliances
 - is tightly connected to existing bioinformatics resources, e.g. public reference data sources...
 - 14 bioinformatics appliances: standard compute nodes, proteomics virtual desktop, Galaxy portal, structural biology...
 - +70 users from all IFB regional centers
PRABI 16, APLIBIO 28, RENABI-NE 13, -GO 7, -SO 2, -GS 5
- Bioinformatics marketplace
 - store images related to life science
 - help users to select the appropriate VM for their analysis





Perspectives



- Create bioinformatics appliances
 - by the experts of the domains
 - make them available to the scientists
- IFB established priorities: 5 scientific domains
 - Microbial Bioinformatics
 - Evolutionary bioinformatics
 - Plant bioinformatics
 - Structural Biology
 - NGS data processing
- and 3 technical pilots
 - Appliances interoperability between different cloud infrastructures
 - Distributing biological data with distributed noSQL engine
 - Live remote cloud processing of sequencing data



Questions ?

Acknowledgments

- Clément Gauthey (IDB-IBCP)
- StratusLab members
- IDB's co-funding by
European Community's Seventh Framework Programme
(INFSO-RI-261552)
French National Research Agency's Arpege Programme
(ANR-10-SEGI-001).
- IFB's funding by French program PIA INBS 2012

