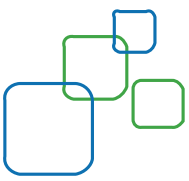


Cloud pour la Bioinformatique



Christophe Blanchet

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



Sequencing data

Next-Generation Sequencing Statistics

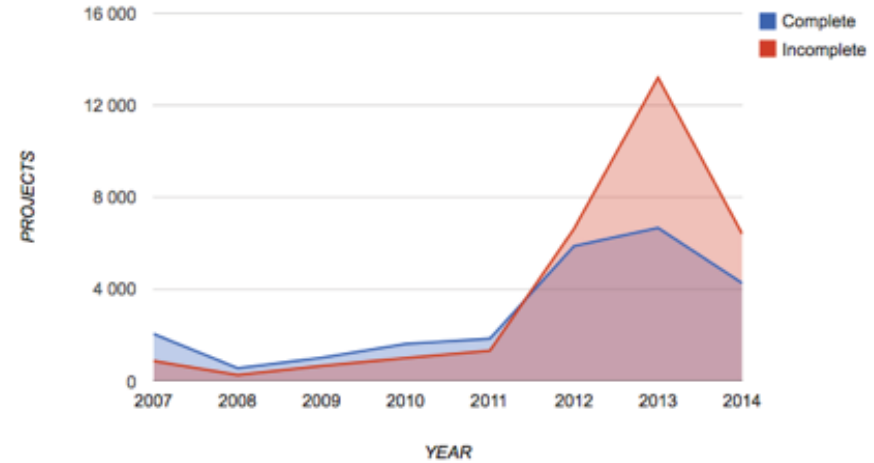
Vendor:	Roche			Illumina			ABI		
Technology:	454			Solexa GA			SOLID		
Platform:	GS20	FLX	T1	I	II	IIx	1	2	3
Reads: (M)	0.5	0.5	1.25	28	100	150	40	115	320
Fragment									
Read length:	100	200	400	35	50	100	25	35	50
Run time: (d)	0.25	0.3	0.4	3	3	5	6	5	8
Yield: (Gb)	0.05	0.1	0.5	1	5	15	1	4	16
Rate: (Gb/d)	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	25	100	140	600

source: www.dolitiogenomics.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
Images: (TB)		0.01	0.01	0.03	0.5	1.1	1.8	2.5

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

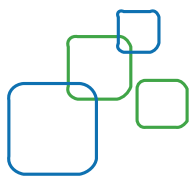
Genome Sequencing Projects



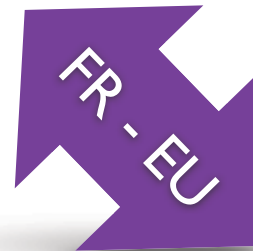
source: www.genomesonline.org

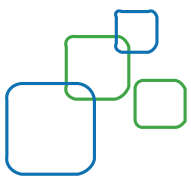


Source: omicsmaps.com



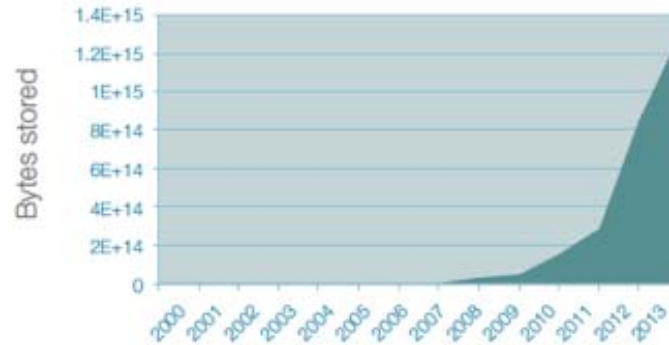
And other experimental data...



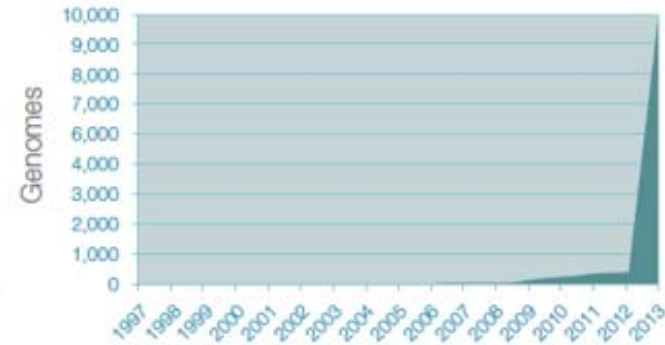


EMBL-EBI data resources growth

Nucleotide sequence data (compressed)



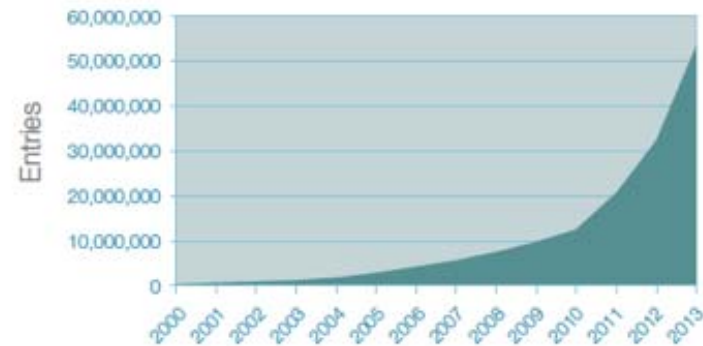
Genomes (all species)



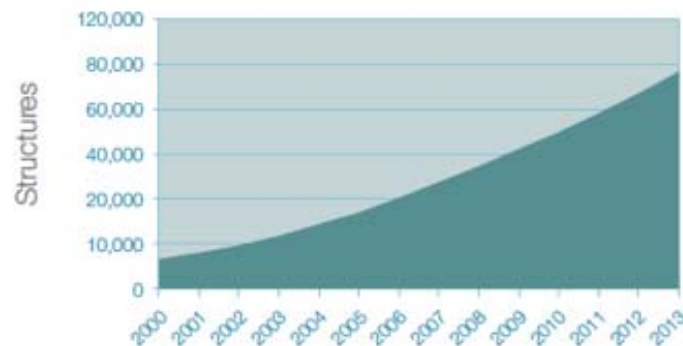
Gene expression data



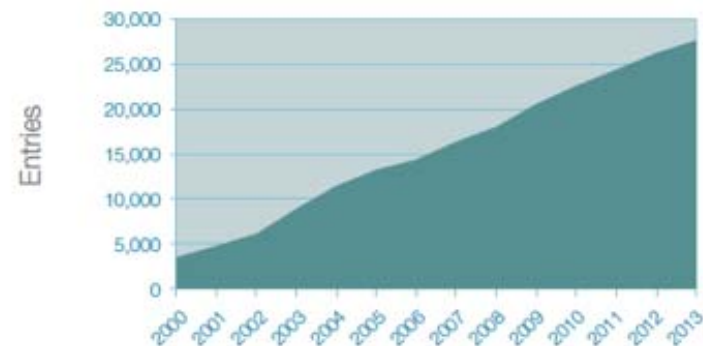
Protein sequence

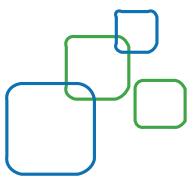


Macromolecular structures



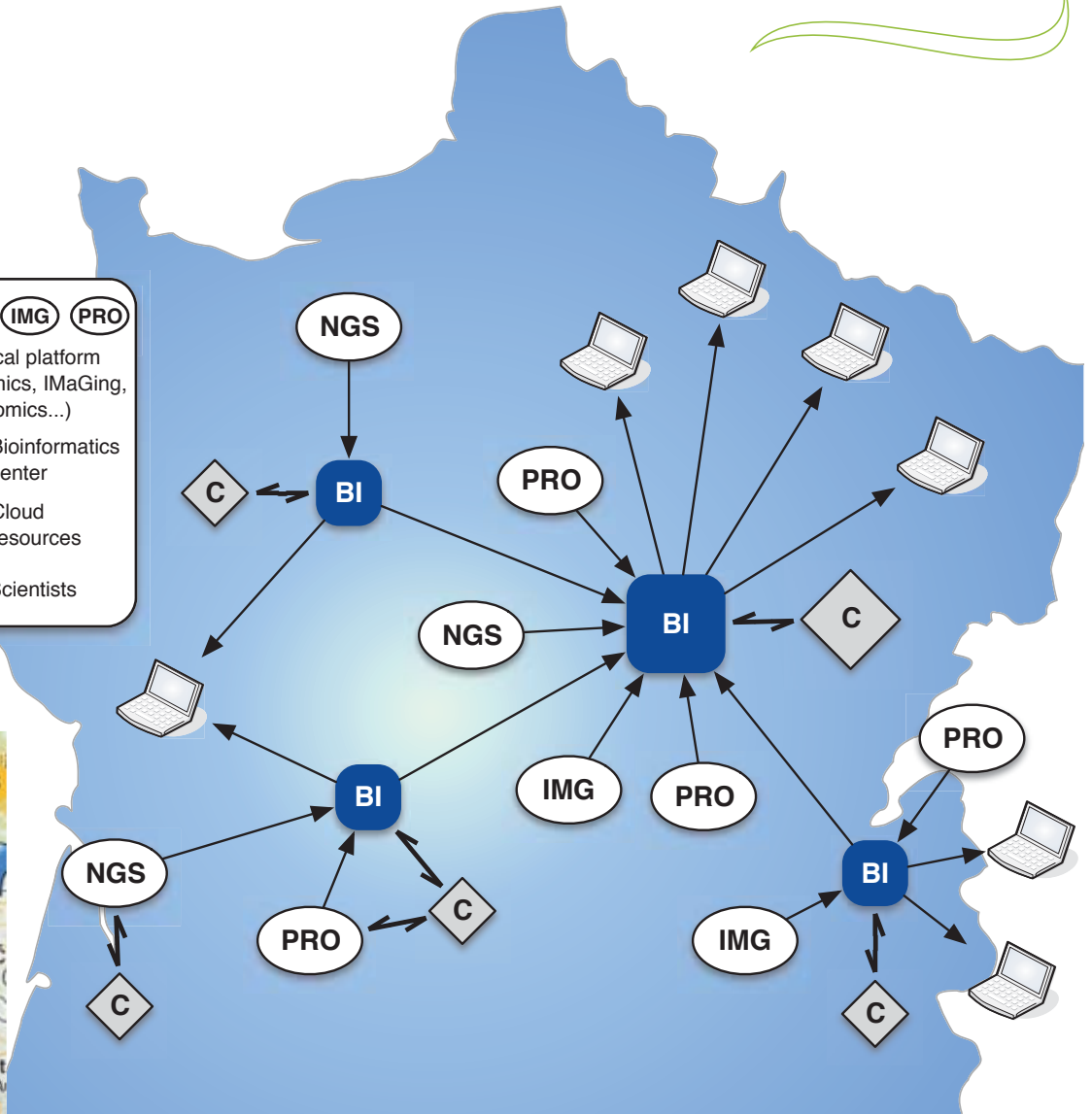
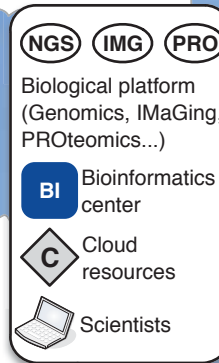
Protein families motif and domains





Plateformes Expérimentales en Biologie

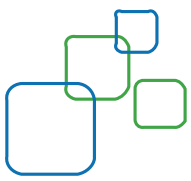
Plateformes nationales (GIS IBISA)	Nb
Imagerie cellulaire	19
Génomique, Transcriptomique	16
Protéomique	13
Biologie structurale, biophysique	11



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

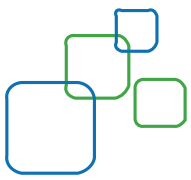


Infrastructures in Biology

The image displays several overlapping browser windows showcasing various bioinformatics infrastructures:

- GenQuest Bioinformatics Platform:** A window showing the URL `http://tools.genouest.org/tools/genouestsf/blast.php/en/` and the text "Run a new blast GenQuest".
- NCBI Resources:** A window showing the NCBI logo and a navigation menu with categories like "Data & Software", "DNA & RNA", "Domains & Structures", "Genes & Expression", "Genetics & Medicine", "Genomes & Maps", "Homology", and "Literature".
- EMBL-EBI:** A window showing the EMBL-EBI logo and a "Data Resource" list including ENA, UniProt, ArrayExpress, Ensembl, InterPro, and PDB.
- NPS@:** A window titled "Welcome to Network Protein Sequence @analysis at IBCP, FRANCE" with the URL `npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html`. It features the logo for "Pôle BioInformatique Lyonnais Network Protein Sequence Analysis" and the text "NPS@ is the IBCP contribution to PBIL in Lyon, France".
- MicroScope Home - MaGe:** A window showing the URL `www.genoscope.cns.fr/agc/microscope/home/index.php`. It includes a login section with "username" and "password" fields, and buttons for "LOGIN" and "SIGN UP". Below this are navigation buttons for "MaGe", "Genomic Tools", "Comparative Genomics", "Metabolism", "Search/Export", "Transcriptomics", and "Variant Discovery".
- Bioinformatics : Home:** A window showing the URL `bioinfo.genotoul.fr` and the logo for "genotoul bioinfo".

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



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

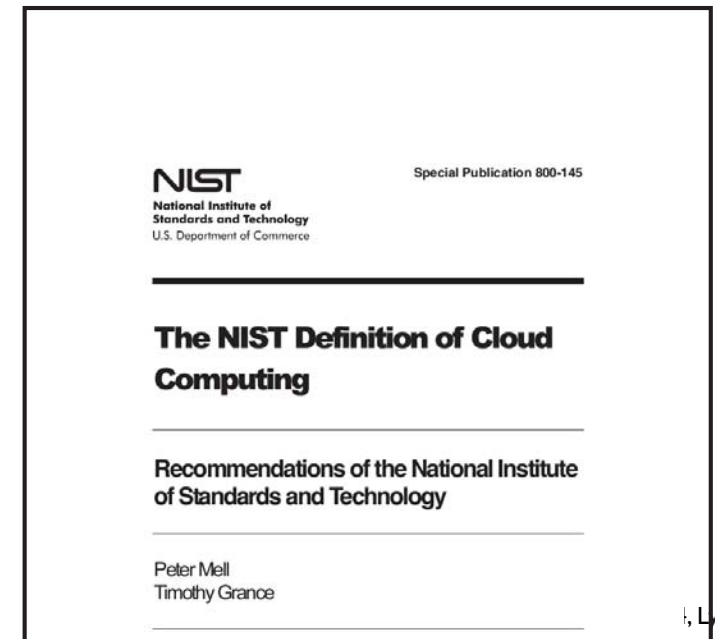
- Private
 - Single administrative domain, limited number of users
- Community
 - Different administrative domains with common interests & proc.
- Public
 - People outside of institute's administrative domain

● 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 with root access



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

Cloud IDB

- **Cloud workbench for Biology**

- Infrastructure Distributed for Biology

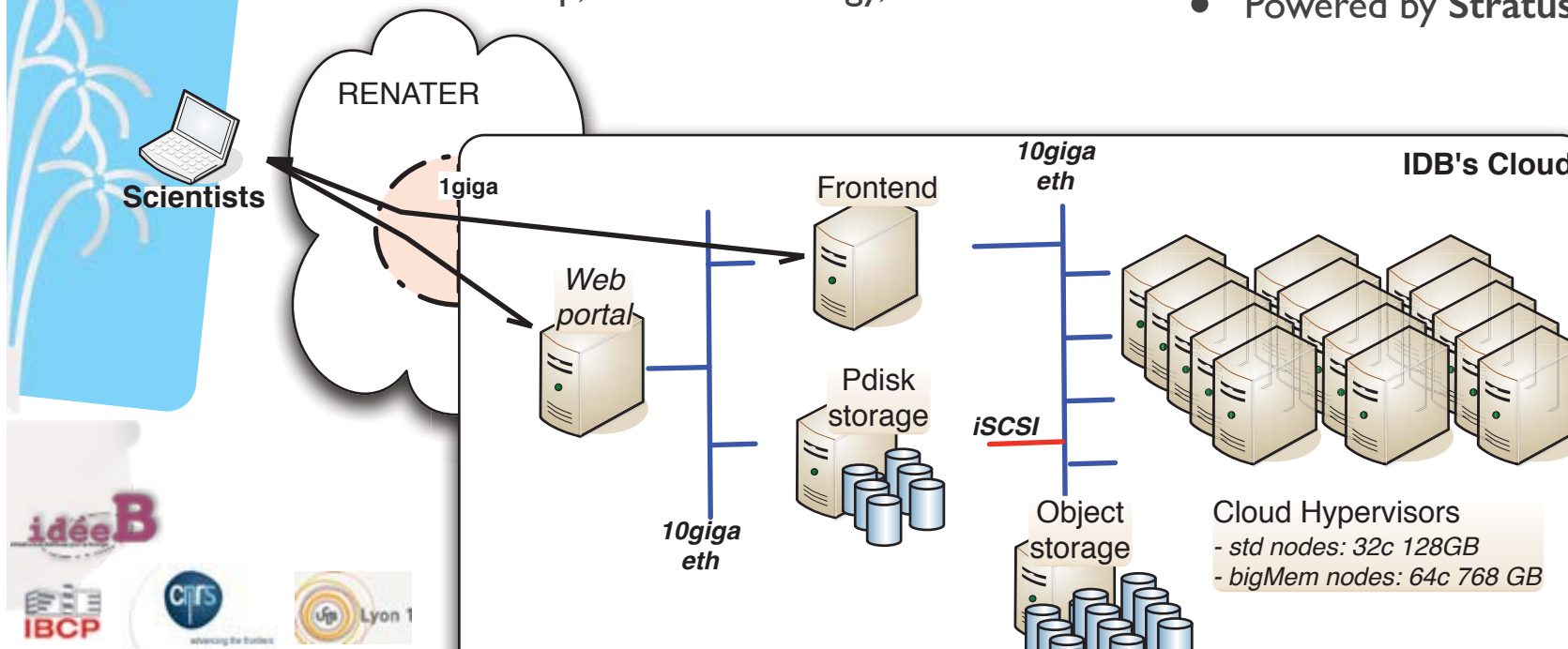
<https://idee-b.ibcp.fr/cloud.html>

- Running since Sept. 2011
IBCP FR3302 CNRS-Univ. lyon I, Lyon, France
- opened to Biology community
- 14 bioinformatics appliances: Galaxy portal, standard compute nodes, proteomics, virtual desktop, structural biology, ...

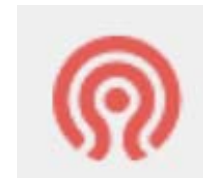
- +70 users from all IFB regional centers
PRABI 16, APLIBIO 28, RENABI-NE 13, -GO 7, -SO 2, -GS 5
- VMs up to 32cores-768GB RAM

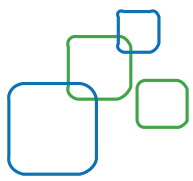
- **Infrastructure**

- 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



stratuslab

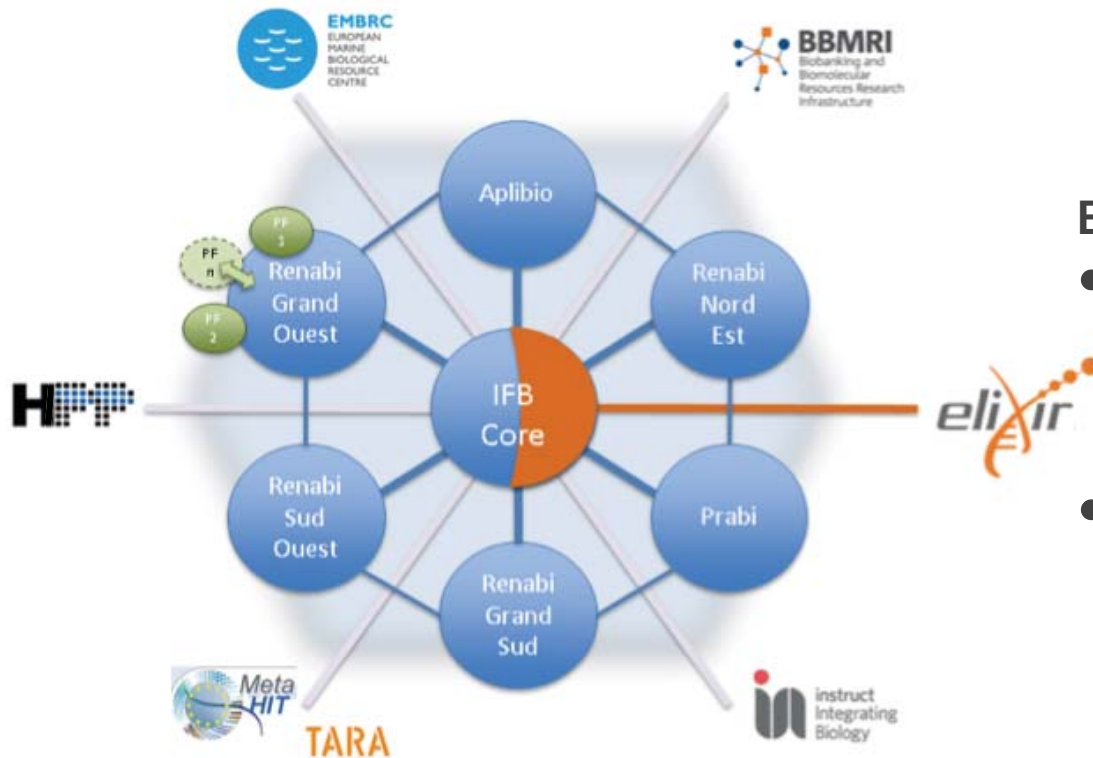




French Institute of Bioinformatics - IFB

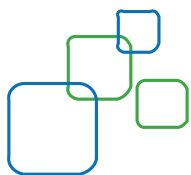
Mission : to make available core bioinformatics resources to the national/international life science research community.

- 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



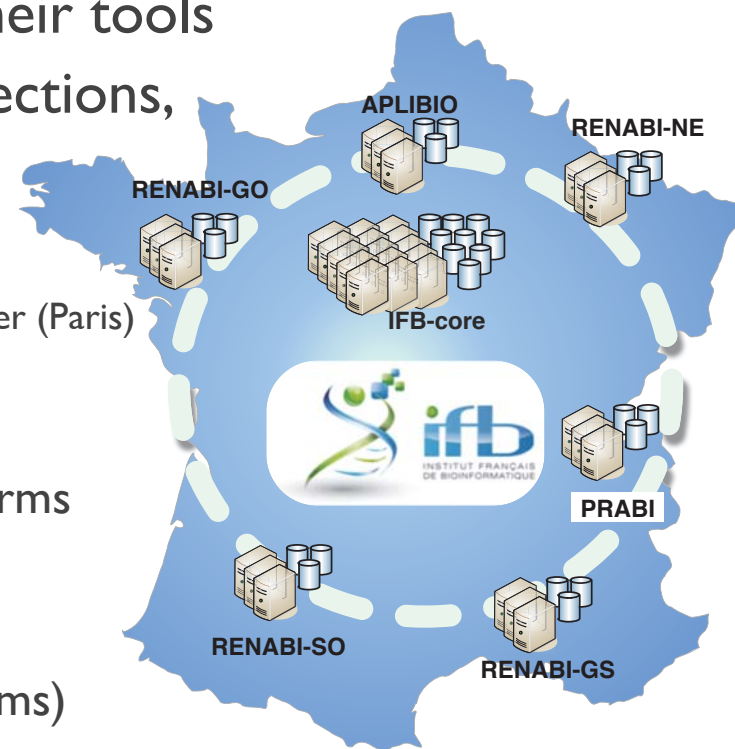
ELIXIR French Node

- 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

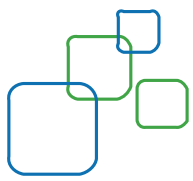


IFB e-Infrastructure

- **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
<i>IFB-core 2014</i>	<i>CNRS-IDRIS, Paris</i>	<i>4,000</i>	<i>500</i>	<i>-</i>	<i>96c 1TB</i>	<i>StratusLab</i>
<i>IFB-core 2015</i>	<i>CNRS-IDRIS, Paris</i>	<i>10,000</i>	<i>2,000</i>	<i>-</i>	<i>96c 2TB</i>	<i>StratusLab</i>
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

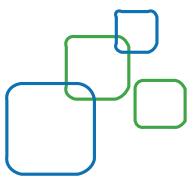


Native cloud services

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

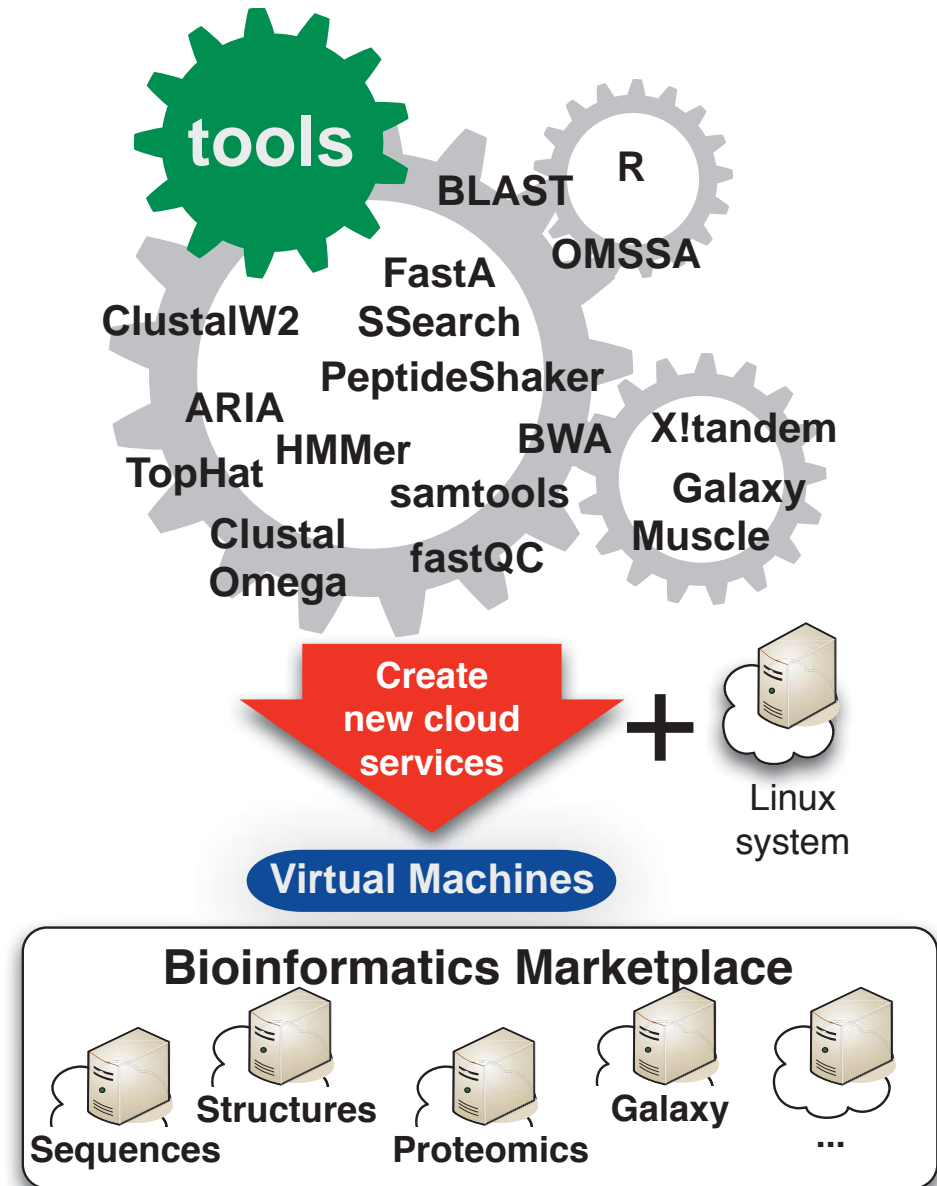
IDB

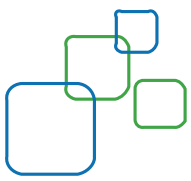
- **Bioinformatics appliances**
 - integrate bioinformatics tools and workflows
- **Bioinformatics marketplace**
 - focus on bioinformatics appliances
 - satisfy visibility constraints for some bioinformatics appliances (confidentiality)
- **Bioinformatics metadata “bio:tool”**
 - annotate appliances 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

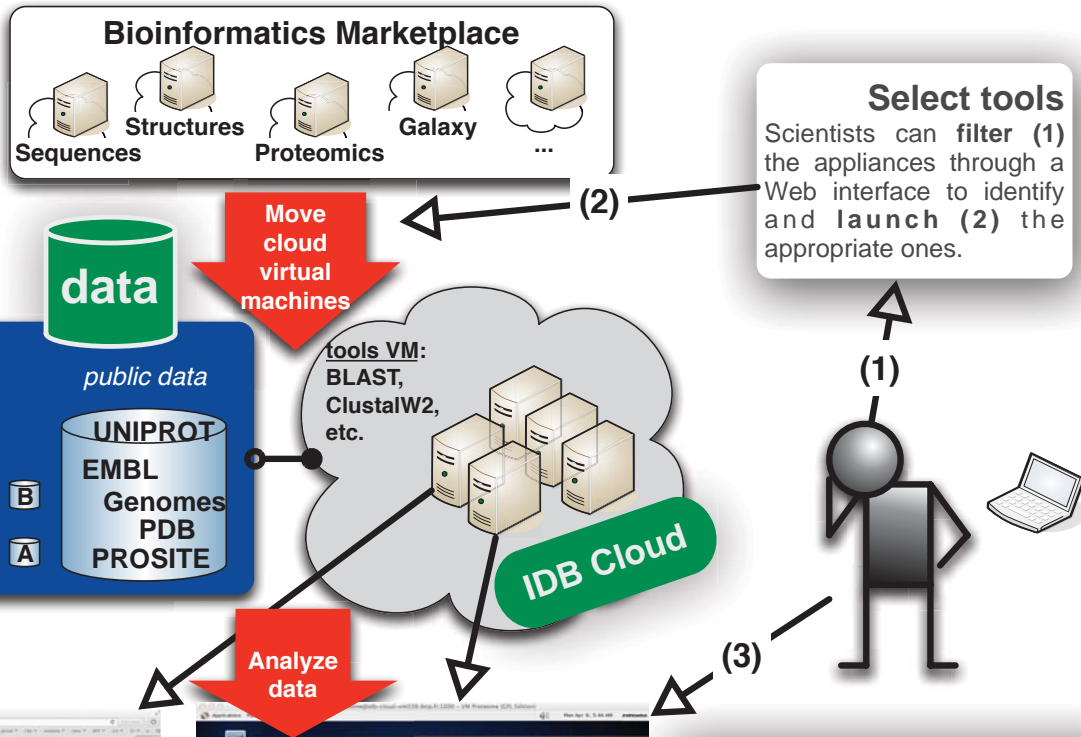
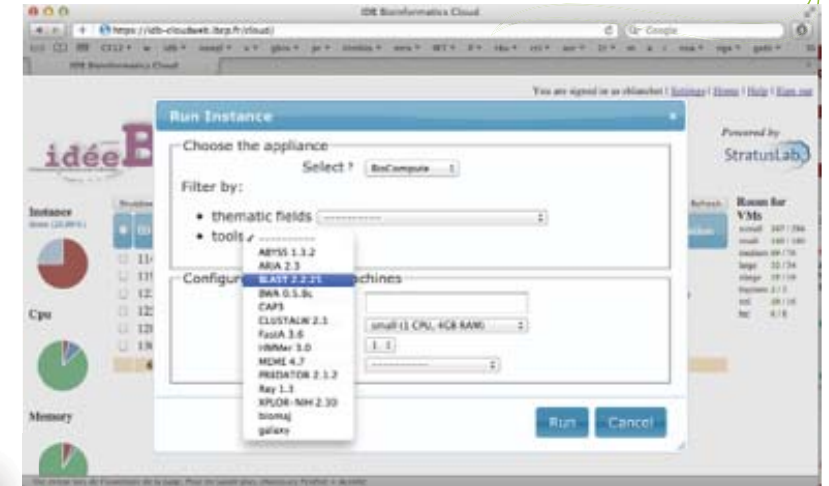
- Bioinformatics appliances are usual virtual machines
 - small : few GB, easy to convert in most virtualization formats
- Installed and pre-configured with bioinformatics tools
 - e.g. BLAST, Clustalw, ARIA, MEME, HMMer, TopHat, BWA, Samtools, etc.
- Recorded in a marketplace
 - devoted to bioinformatics





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

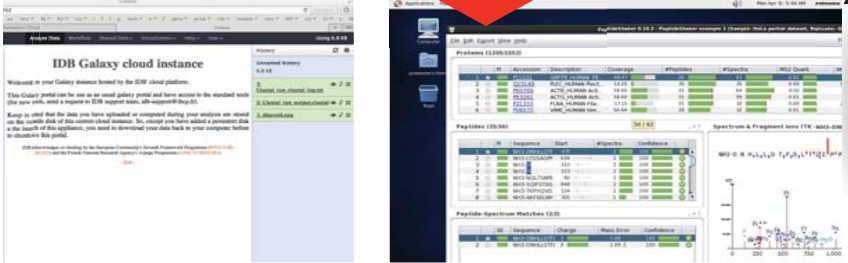


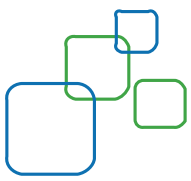
Select tools
Scientists can filter (1) the appliances through a Web interface to identify and launch (2) the appropriate ones.

Use tools (3)
Scientists have access to their own cloud resources through web portal, remote virtual desktop or SSH.

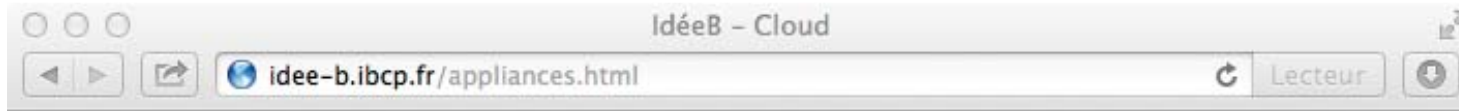
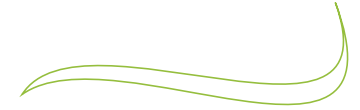
- Filter images with metadata related to bioinformatics
 - attribute <bio:tool> in VM manifests
 - scientists can select the appropriate appliance according to the tools required for their analyses
 - e.g. the BLAST tool

Deploy on several clouds





Appliances page



Bioinformatics Cloud Appliances

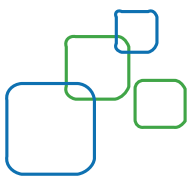
[Databases](#) | [Tools](#) | [Cloud](#) | [Grid](#) | [Documentation](#) | [Sign in](#)
[Appliances](#) | [Cloud interface](#)

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.

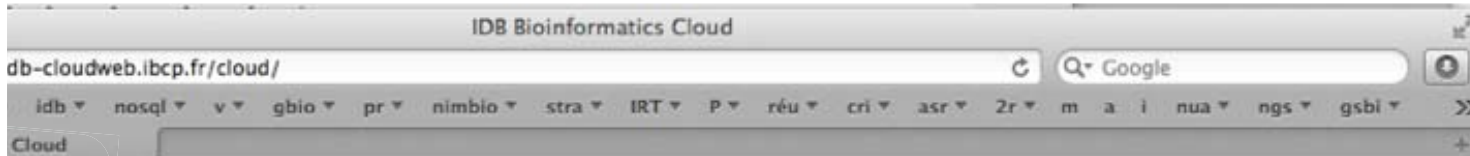
You can get a description of each appliance by *clicking on their name* in the list below. *To run your own instances*, click on the corresponding power button. Then, you will be redirected to a pre-filled form to create your instances.

▶ Bioinformatics compute node	⏻
▶ Galaxy portal	⏻
▼ Proteomics	⏻
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)	⏻

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



Filter appliances with tools description



Run Instance

Choose the appliance

Select ?

Filter by:

- thematic fields
- tools

Configure your virtual machines

Name ?

Type ?

Number ?

Storage ?

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

Run Instance

Choose the appliance

Select ?

Filter by:

- thematic fields
- tools

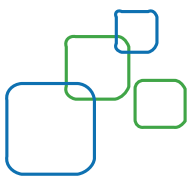
Configure your virtual machines

-
- ABYSS 1.3.2
- ARIA 2.3
- BLAST 2.2.25
- 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

la page. Pour en savoir plus, choisissez Fenêtre > Activité.

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

https://cloud.ifb.idris.fr/cloud/


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



Bioinformatics cloud



Powered by stratuslab



Hosted at iris


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

Showing 1 to 6 of 6 entries Search:


ID	Name	Appliance	CPU%	CPU	Mem.	#Storage	Access
94	Public data source	BIO Data	3%	4	16	0	ssh http
357	test2	RSAT 0.1	0%	4	8	0	ssh http
365	proxy	Galaxy 4.1	0%	4	8	1	ssh http
369	hotplug	BIO ComputeNode	0%	4	8	1	ssh
385	testrel	Galaxy 4.2	0%	4	8	1	ssh http
390	test-cleaner	Ubuntu 14.04	0%	2	8	0	ssh
6		6		22	56	3	

Show 25 entries First Previous 1 Next Last


Instance



Storage

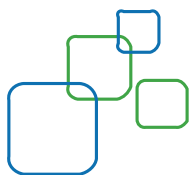


Cpu



Room for VMs

c2.large	25 / 36
c2.small	105 / 144
c2.xlarge	12 / 18
c3.large	24 / 34
c3.medium	50 / 70
c3.xlarge	11 / 16
c3.xxlarge	5 / 6
m1.medium	14 / 20
m1.xlarge	1 / 2
m1.xxlarge	1 / 2



Connection to VMs



- ssh/scp
 - ssh/scp
 - ssh/scp [http](#)
- First Previous 1 Next Last

X2Go Client

Cloud IDB

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

Cloud IFB

- proteome@VM IP
- GNOME
- 800x600
- Enabled

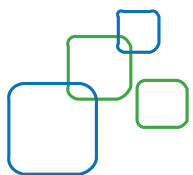
Connection Information

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

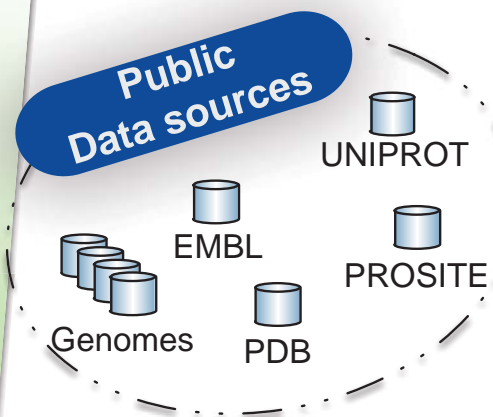
```
maRacine — root@idb-cloud-vm050:~ — ssh — Ⓜ7  
idb1:maRacine cblanchet$ ssh -A -p 20062 root@idb-cloud.ibcp.fr  
Last login: Mon May 20 15:05:28 2013 from mtl01-1-88-161-187-9.fbx.pr  
oxad.net  
[root@idb-cloud-vm050 ~]# ls  
anaconda-ks.cfg  install      install.log.syslog  
cleaner.sh      install.log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



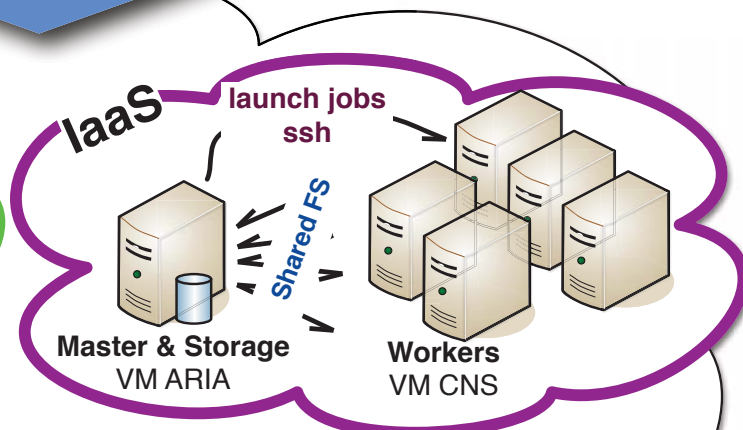
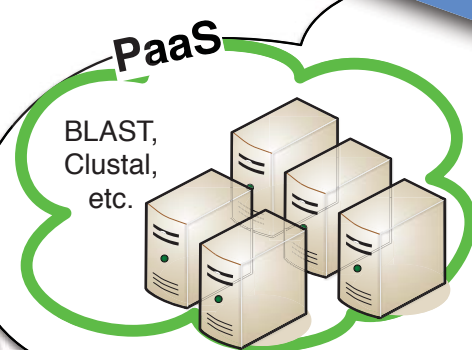
Cloud Storage for Biological Data

Upload your data

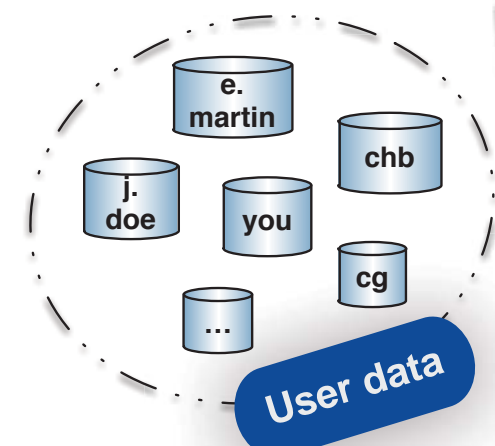
sftp/http/S3



shared (NFS ro)



Bioinformatics Cloud

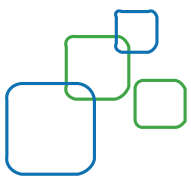


v. disk
samba
S3

Identity Mgmt

sftp/http/S3

Get your results



Exchanging data with VMs

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

Galaxy

idb-cloud.ibcp.fr:20133

Galaxy

Analyze Data Workflow Shared Data Visualization

Tools

search tools

Upload File (version 1.1.3)

File Format: Auto-detect

Which format? See help below

File: Choisir le fichier aucun fichier sél.

TIP: Due to browser limitations, uploading files large guaranteed to fail. To upload large files, use the URL FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or file.

Files uploaded via FTP:

File	Size	Date
<input type="checkbox"/> SampleC.1.fastq	269.0 KB	05/22/2013 1
<input type="checkbox"/> SampleC.2.fastq	269.0 KB	05/22/2013 1
<input type="checkbox"/> SampleB.2.fastq	262.7 KB	05/22/2013 1
<input type="checkbox"/> vcfutils.pl	15.3 KB	05/22/2013 1
<input type="checkbox"/> SampleA.2.fastq	244.2 KB	05/22/2013 1

```
dataset — title — bash — %9
idb1:dataset cblanchet$ scp -P 20132 * root@idb-cloud.ibcp.fr:upload_dir/
SampleA.1.fastq      100% 244KB 244.2KB/s  00:00
SampleA.2.fastq      100% 244KB 244.2KB/s  00:00
SampleB.1.fastq      100% 263KB 262.7KB/s  00:00
SampleB.2.fastq      100% 263KB 262.7KB/s  00:00
SampleC.1.fastq      100% 269KB 269.0KB/s  00:00
SampleC.2.fastq      100% 269KB 269.0KB/s  00:00
ref1.fasta           100%  10KB   9.9KB/s   00:00
run_analyses.sh      100% 1321    1.3KB/s   00:00
vcfutils.pl          100%  15KB  15.3KB/s   00:00
idb1:dataset cblanchet$
```

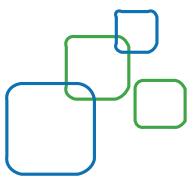
idb-cloud.ibcp.fr - SFTP

Nouvelle connexion Connexion rapide Action Actualiser Edition Se déconnecter

/root

nom du fichier	Taille	Date de modification
anaconda-ks.cfg	2.5 KB	02/03/12 09:09
cleaner.sh	338 B	27/09/12 13:33
install	--	27/09/12 13:51
install.log	10.5 KB	02/03/12 09:09
install.log.syslog	3.8 KB	02/03/12 09:08
mydisk	10 B	16/11/12 16:03

6 Fichiers



Moving VMs vs Data

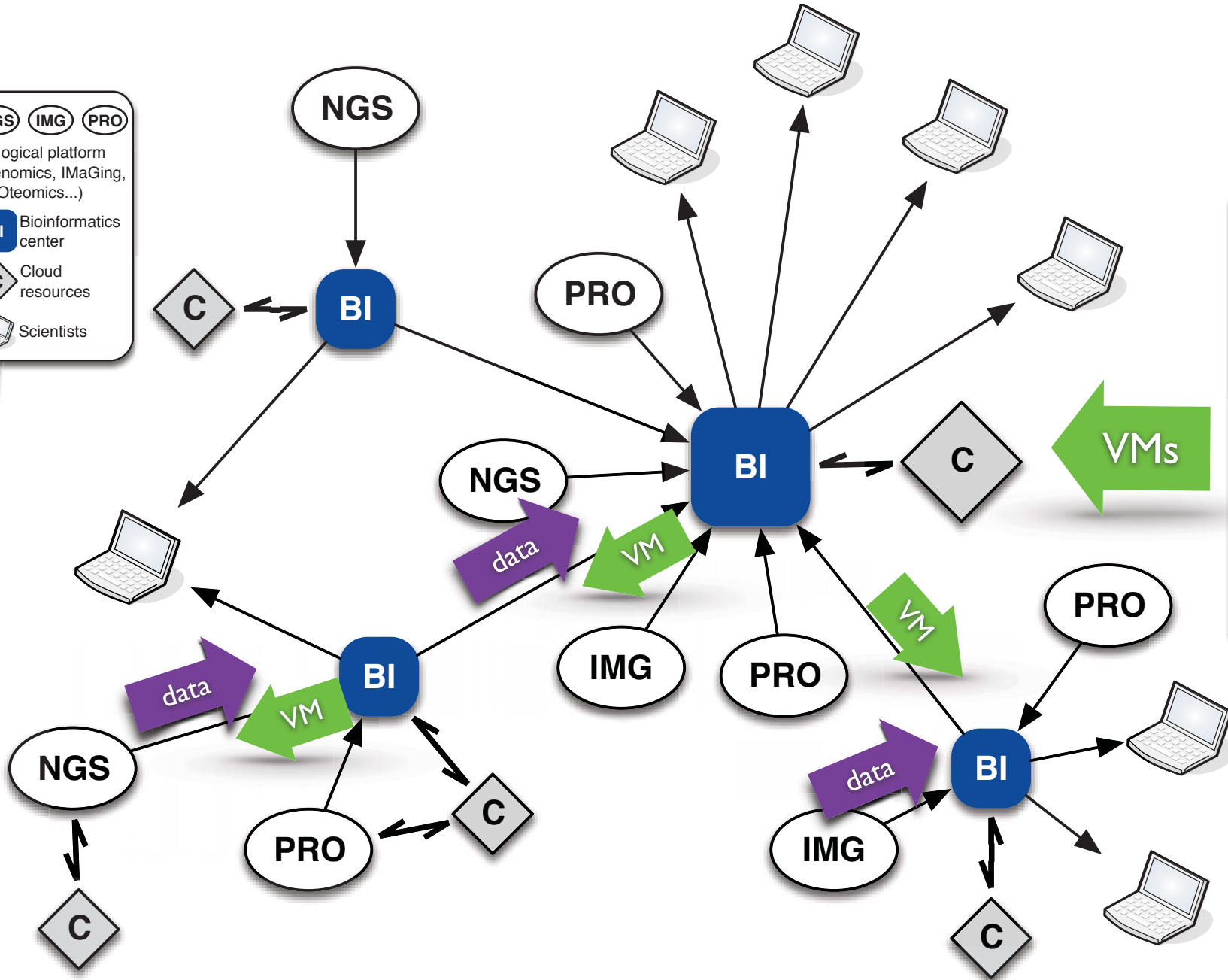


NGS **IMG** **PRO**
Biological platform
(Genomics, IMAging, PROteomics...)

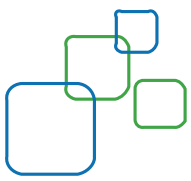
BI Bioinformatics center

C Cloud resources

Scientists



IFB
Bioinformatics
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.
 - Automatically shared to the VMs
- Cluster mode
 - turn several instances in a single virtual cluster
 - shared file system
 - batch scheduling

Metadata

Home | Endorsers | Query | Upload | About

Metadata

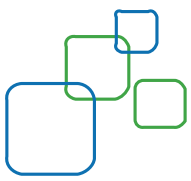
Show 10 entries

BIO compute node

Endorser: christophe.blanchet@ibcp.fr
Identifier: O2fHwIZlxLDoxcuCmqwoWVGbBM
Created: 2014-04-04T15:34:44Z
Kind: machine

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, simp96, 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: stratus-run-instance --contextualization "BIO_DB_SERVER=idb-...". This appliance can also be used as a hot mount. See the documentation on the Idee-B'

```
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
oxad.net
[root@idb-cloud-vm050 ~]# ls
anaconda-ks.cfg  install      install.log.syslog
cleaner.sh      install.log  mydisk
[root@idb-cloud-vm050 ~]#
```



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

marketplace.ibc.idris.fr/metadata

Home | Endorsers | Query | Upload | About

Metadata

Show 10 entries

Galaxy portal

Endorser: christophe.blanchet@ibcp.fr
Identifier: GDqP1arAKmWzR2PB-tCEDsHbu7n
Created: 2013-11-21T15:14:39Z
Kind: machine

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>).

[More...](#)

<input type="checkbox"/>	1875	blast	Running	BioCompute	2%	2	8	ssh http
<input type="checkbox"/>	1876	portal	Running	Galaxy	2%	2	8	http
<input type="checkbox"/>	1877	blast machine	Running	BioCompute	2%	4	16	ssh http

idb-cloud.ibcp.fr:20023

IDB Bioinformatics Cloud

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Using 4.9 GB

IDB Galaxy cloud instance

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

Usage

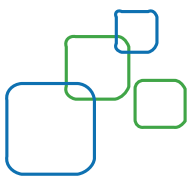
This appliance is configured with the well-known GALAXY portal. You connect to it with a standard web browser : simply follow the link on the main IDB cloud interface. It can be used as an usual galaxy portal and you have access to pre-installed standard bioinformatics tools (for new tools, send a request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

Tools: search tools

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats

Unamed history: 4.9 GB

- 23: Clustal run clustal log.txt
- 22: Clustal run output.clustal
- 21: dbprot6.seq

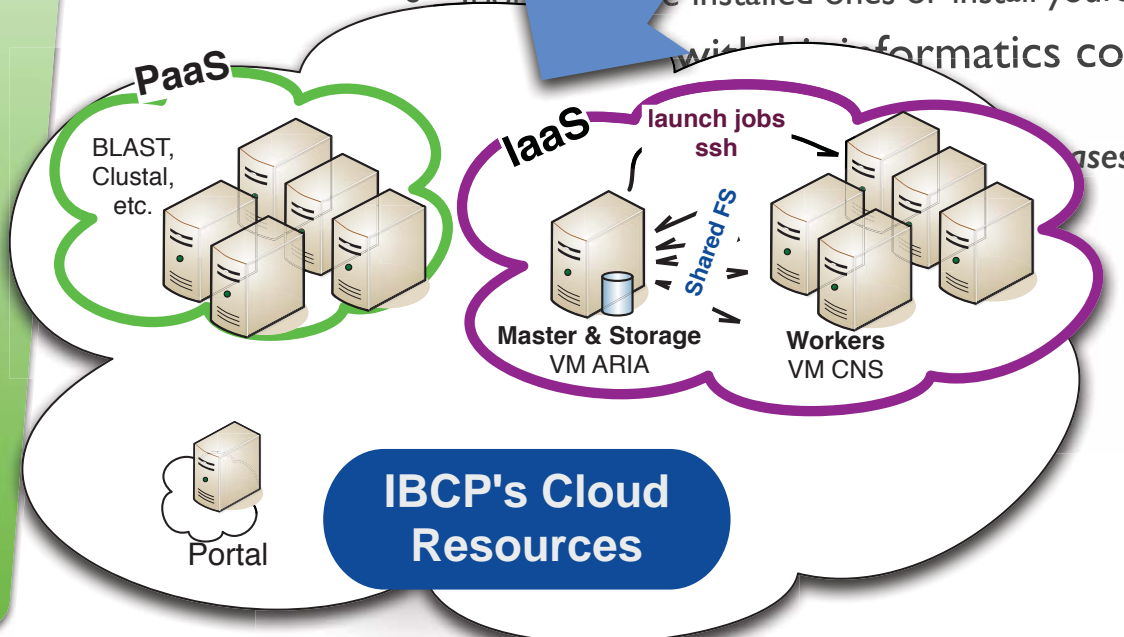


Run your Galaxy Portal on Cloud

Bioinformatics Marketplace

Sequence Structure NGS Galaxy ARIA (...)

- Stay Connected to Standard Data &
 - User data: attach datafiles or attach pdisk
 - Reference databases: mount biodata server s
 - Tools: pre-installed ones or install yours



Create Instance

Choose The Appliance

Appliance ?

Filter by ?

Configure Your Virtual Machines

Name ?

Unique ?

Type ?

Number ?

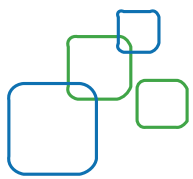
Create appliance ?

Configure Your Storage

Persistent disk ?

OR Volatile disk size ?

Create



Connect to your Galaxy Portal

<input type="checkbox"/>	5012	mapred#2		BIO MapReduce		4	16	0			ssh/scp	
<input type="checkbox"/>	5013	mapred#3		BIO MapReduce		4	16	0			ssh/scp	
<input type="checkbox"/>	5140	ma VM galaxy		Galaxy		4	16	1			ssh/scp http	

10	7	52	269	2
----	---	----	-----	---

Show entries

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*

Tools

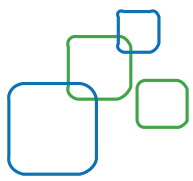
search tools

- 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

History

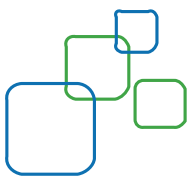
Using 4.9 GB

- 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.fastq
- 16: SampleC.1.fastq
- 15: SampleC.2.fastq
- 14: SampleB.1.fastq
- 13: SampleA.1.fastq
- 12: SampleB.2.fastq
- 11: ref1.fasta



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

Metadata

Home | Endorsers | Query | Upload | About

Metadata

Show 10 entries

Proteomics

Endorser: christophe.blanchet@ibcp.fr
Identifier: POCtUXnTejwxUbam6U1sj7uuah3
Created: 2014-04-04T13:39:36Z
Kind: machine

Bioinformatics virtual appliance for protein identification from mass spectrometry data. Contains OMSSA, X!Tander PeptideShaker and SearchGUI tools. Details on IDB web sit <http://idee-b.ibcp.fr>.

[More...](#)

PeptideShaker 0.19.3 - PeptideShaker example 1 (Sample: HeLa partial dataset, Replicate: 0)

	PI	Accession	Description	Coverage	#Peptides	#Spectra	MS2 Quant.	MI
1	★	P11021	GRP78_HUMAN 78...	48.47	36	43	0.01	
2	★	Q15149	PLEC_HUMAN Plect...	10.20	35	35	0.00	5
3	★	P60709	ACTB_HUMAN Acti...	58.40	31	44	0.02	
4	★	P63261	ACTG_HUMAN Acti...	58.40	31	39	0.01	
5	★	P21333	FLNA_HUMAN Fila...	17.15	31	32	0.00	2
6	★	P08670	VIME_HUMAN Vim...	56.44	28	32	0.01	

	PI	Sequence	Start	#Spectra	Confidence
1	★	NH3-DNHLLGTF	475	2	100
2	★	NH3-LYGSAGPP	634	2	100
3	★	NH3-...	153	2	100
4	★	NH3-...	153	2	100
5	★	NH3-NQLTSNPE	82	2	100
6	★	NH3-SQIFSTAS	448	2	100
7	★	NH3-TKPYIQVD	124	2	100
8	★	NH3-AKFEELNM	325	1	100

Peptide-Spectrum Matches (2/2)

	SE	Sequence	Charge	Mass Error	Confidence
1	★	NH3-DNHLLGTFI	3	0.08	100

Spectrum & Fragment Ions (TK - NH3-DNI)

NH3-D N H₁ L₁ L₁ G T₁ F₁ D₁ L¹ T¹ G¹ I¹ P¹ P

Int

Y5

Y5++

Y4

Y3

Y2

Y1

b3

b2

b1

b0

Y7

Y8

Y9

Y10

Y11

Y12

Y13

Y14

Y15

Y16

Y17

Y18

Y19

Y20

Y21

Y22

Y23

Y24

Y25

Y26

Y27

Y28

Y29

Y30

Y31

Y32

Y33

Y34

Y35

Y36

Y37

Y38

Y39

Y40

Y41

Y42

Y43

Y44

Y45

Y46

Y47

Y48

Y49

Y50

Y51

Y52

Y53

Y54

Y55

Y56

Y57

Y58

Y59

Y60

Y61

Y62

Y63

Y64

Y65

Y66

Y67

Y68

Y69

Y70

Y71

Y72

Y73

Y74

Y75

Y76

Y77

Y78

Y79

Y80

Y81

Y82

Y83

Y84

Y85

Y86

Y87

Y88

Y89

Y90

Y91

Y92

Y93

Y94

Y95

Y96

Y97

Y98

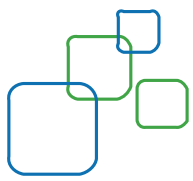
Y99

Y100

yon

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 *n* 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)

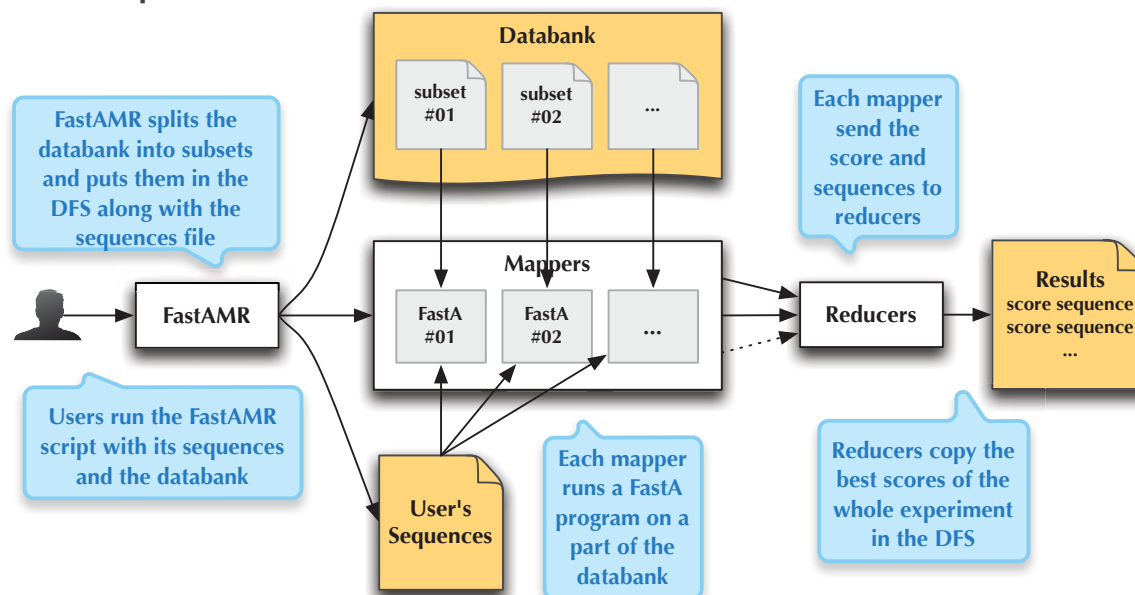
[More...](#)

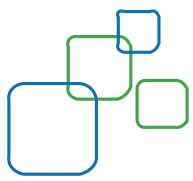
Hadoop MapReduce

Endorser: *clement.gauthey@ibcp.fr*
Identifier: *BttU7uNM5UT1haUigV57xySi2rr*
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 (Created for the French project MapReduce, ANR ARPEGE, 2010-2013, mapreduce.inria.fr)

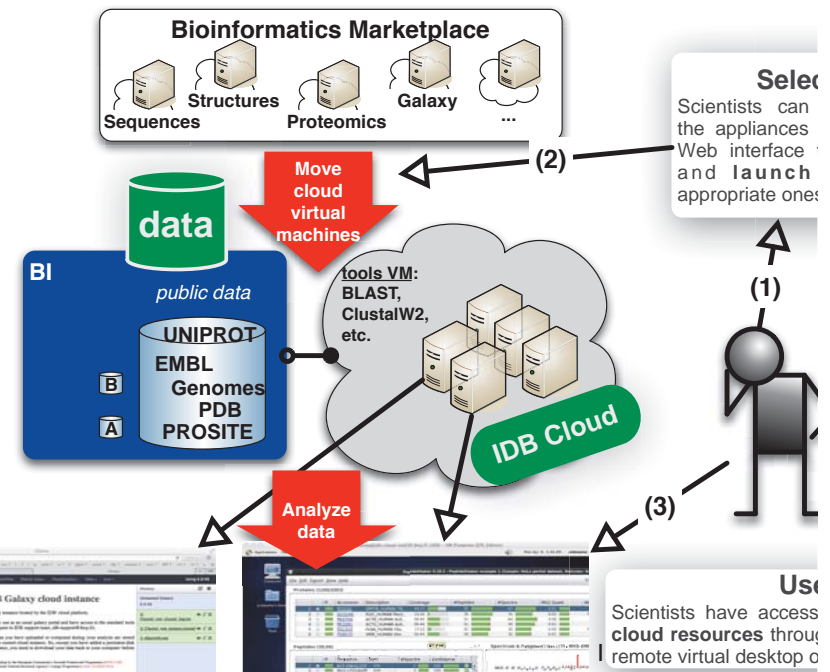
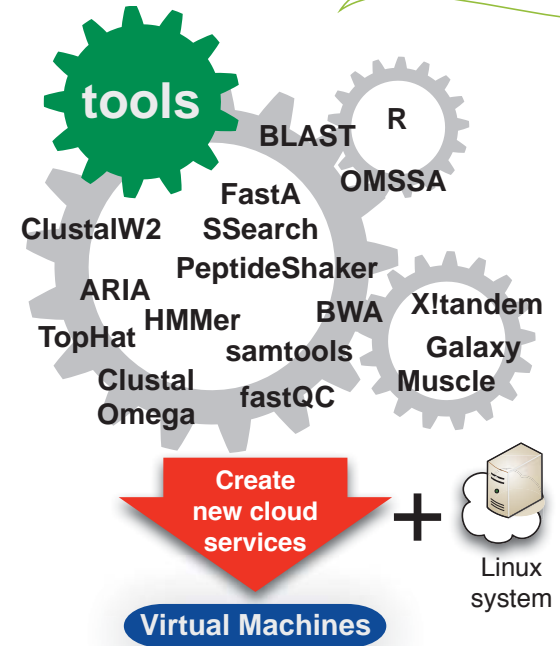
[More...](#)

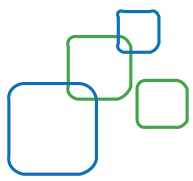




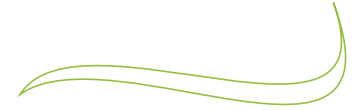
Cloud it be done ?

- 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

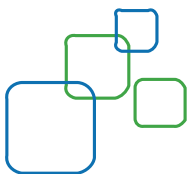




Perspectives



- 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
(INFOS-RI-261552)
French National Research Agency's Arpege Programme
(ANR-10-SEGI-001).
- IFB's funding by French program PIA INBS 2012

