Cloud
Ready for Bioinformatics ?

C. Blanchet and C. Gauthey

Plateforme ‘Infrastructure Distribuée pour la Biologie’

Journée ‘Clouds pour le Calcul Scientifique, Paris Sud’ (Orsay, France)

27 novembre 2012

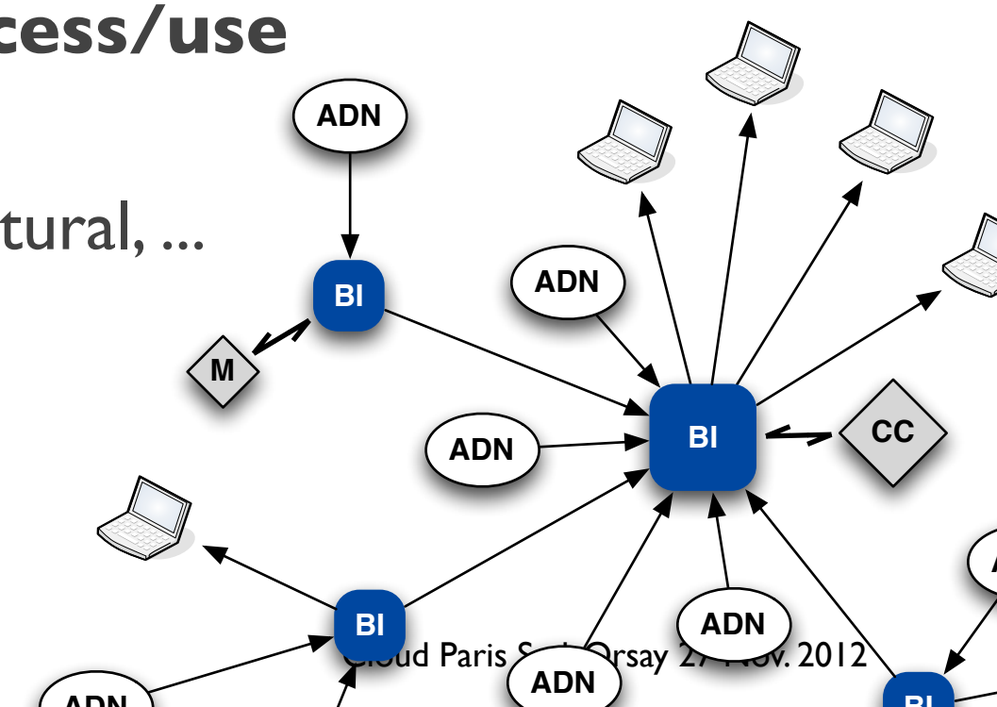
Bioinformatics Today

- Size of biological data are tremendous
  - Institut Sanger, UK, 5 PB
  - Beijing Genome Institute, China, 4 sites, 10 PB

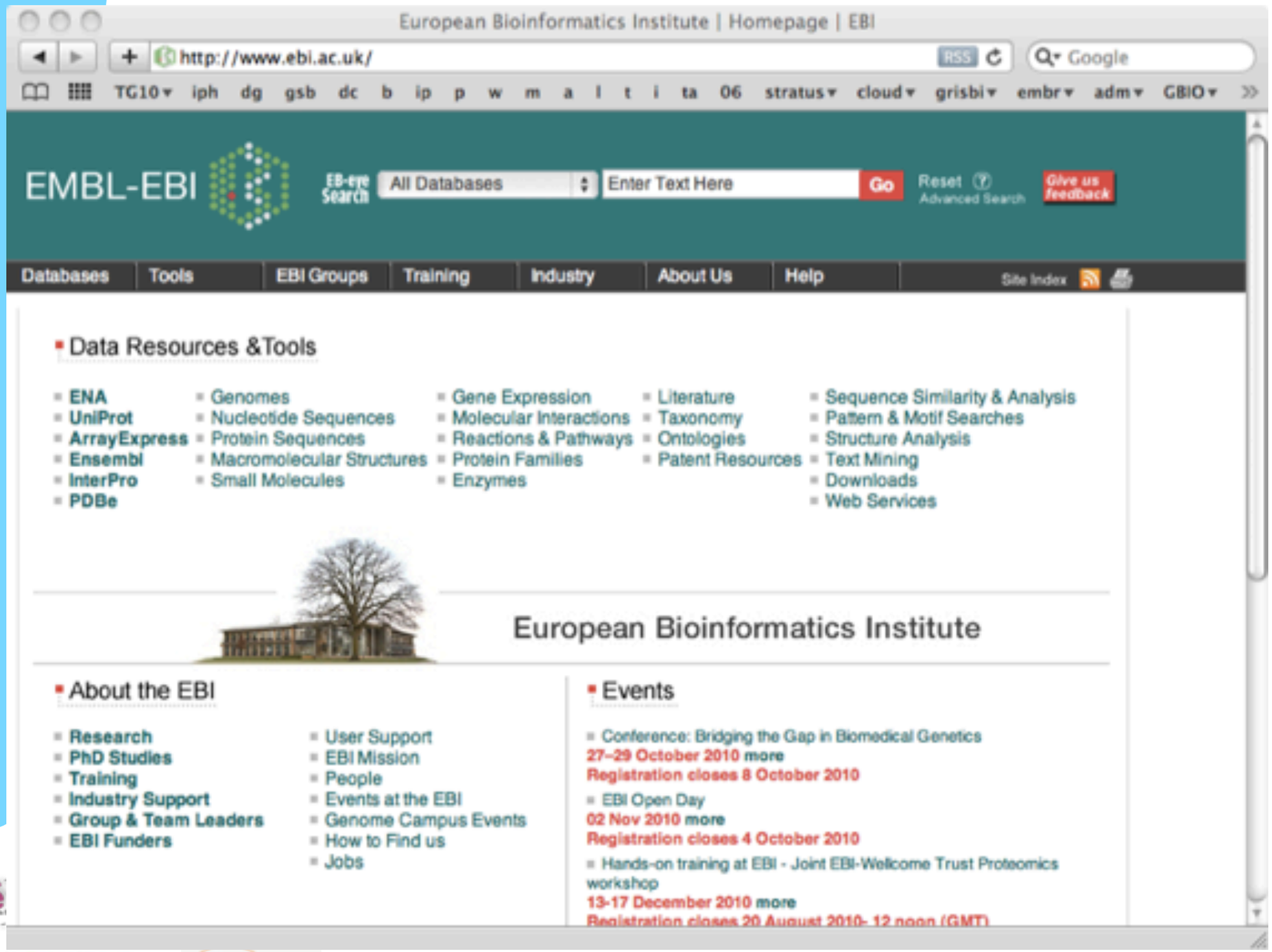
  ➡ **Huge data in lot of places**
- Analysing such data became difficult
  - Scale-up of the analyses : gene/protein to complete genome/proteome, ...
  - Lot of different daily-used tools
  - That need to be combined in workflows
  - Usual interfaces: portals, Web services, federation,...

  ➡ **Datacenters with ease of access/use**
- Distributed resources
  - Experimental platforms: NGS, structural, ...
  - Bioinformatics platfomr

  ➡ **Federation**



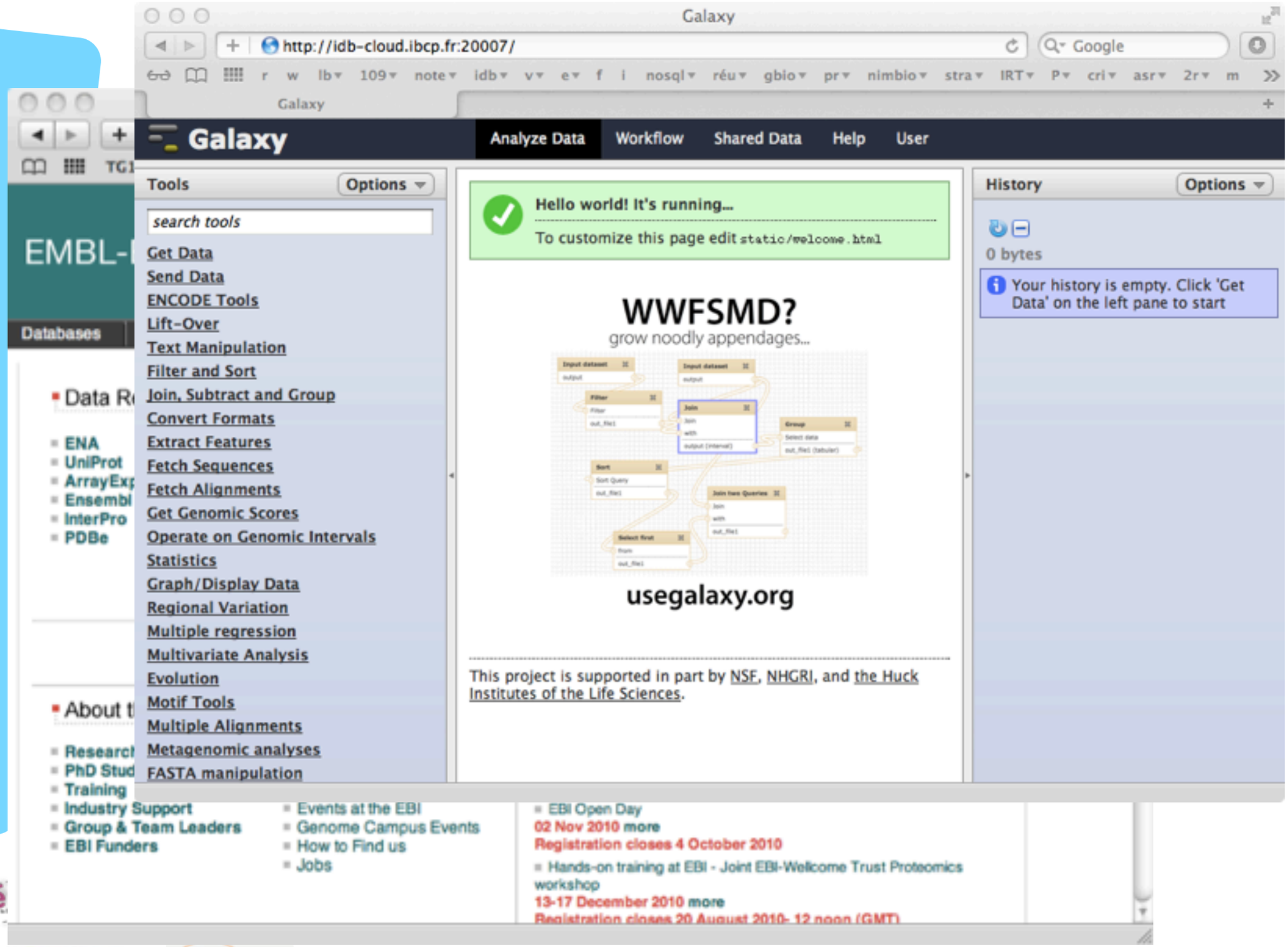
Infrastructure in Biology



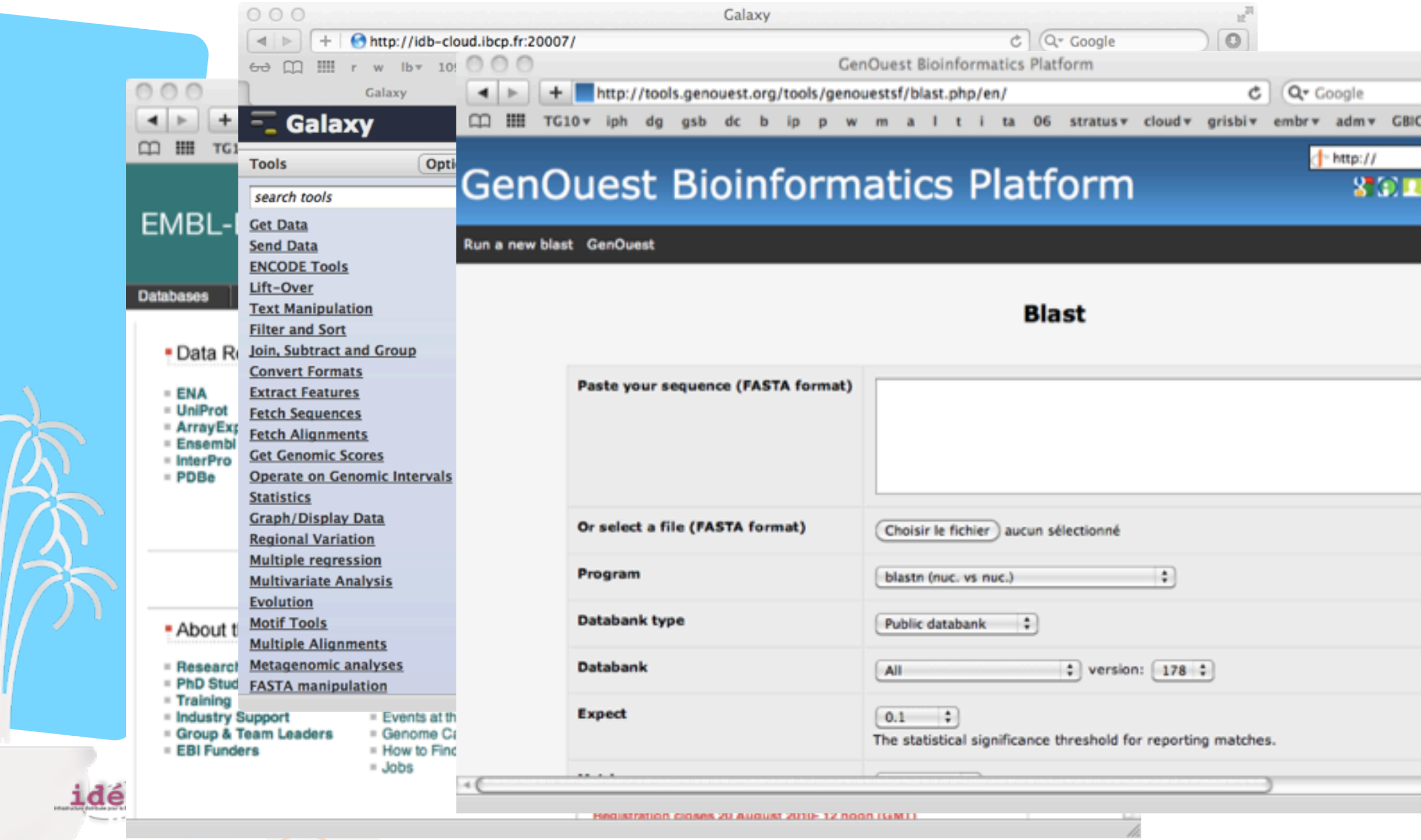
IBCP

Lyon 1

Infrastructure in Biology



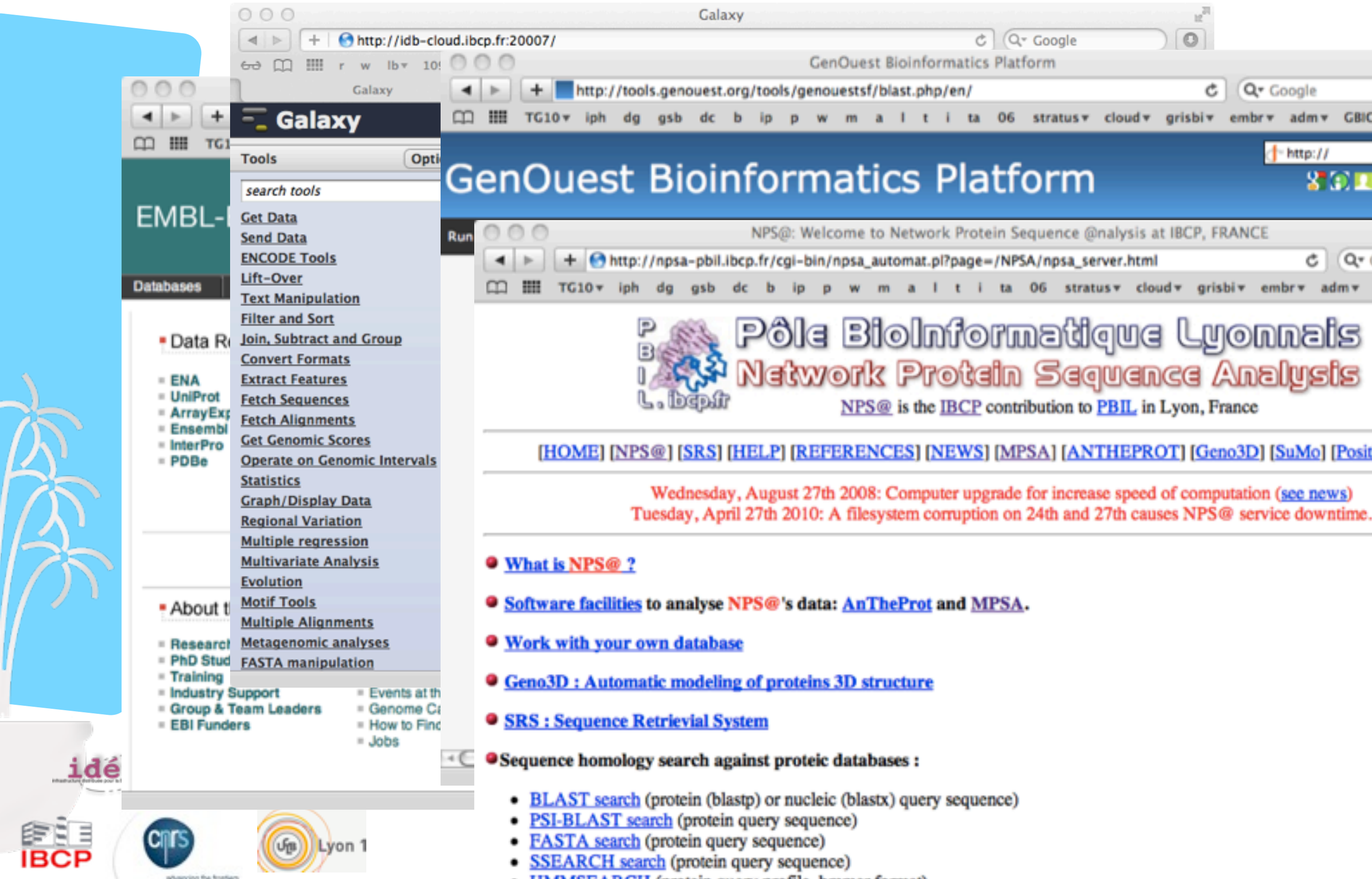
Infrastructure in Biology



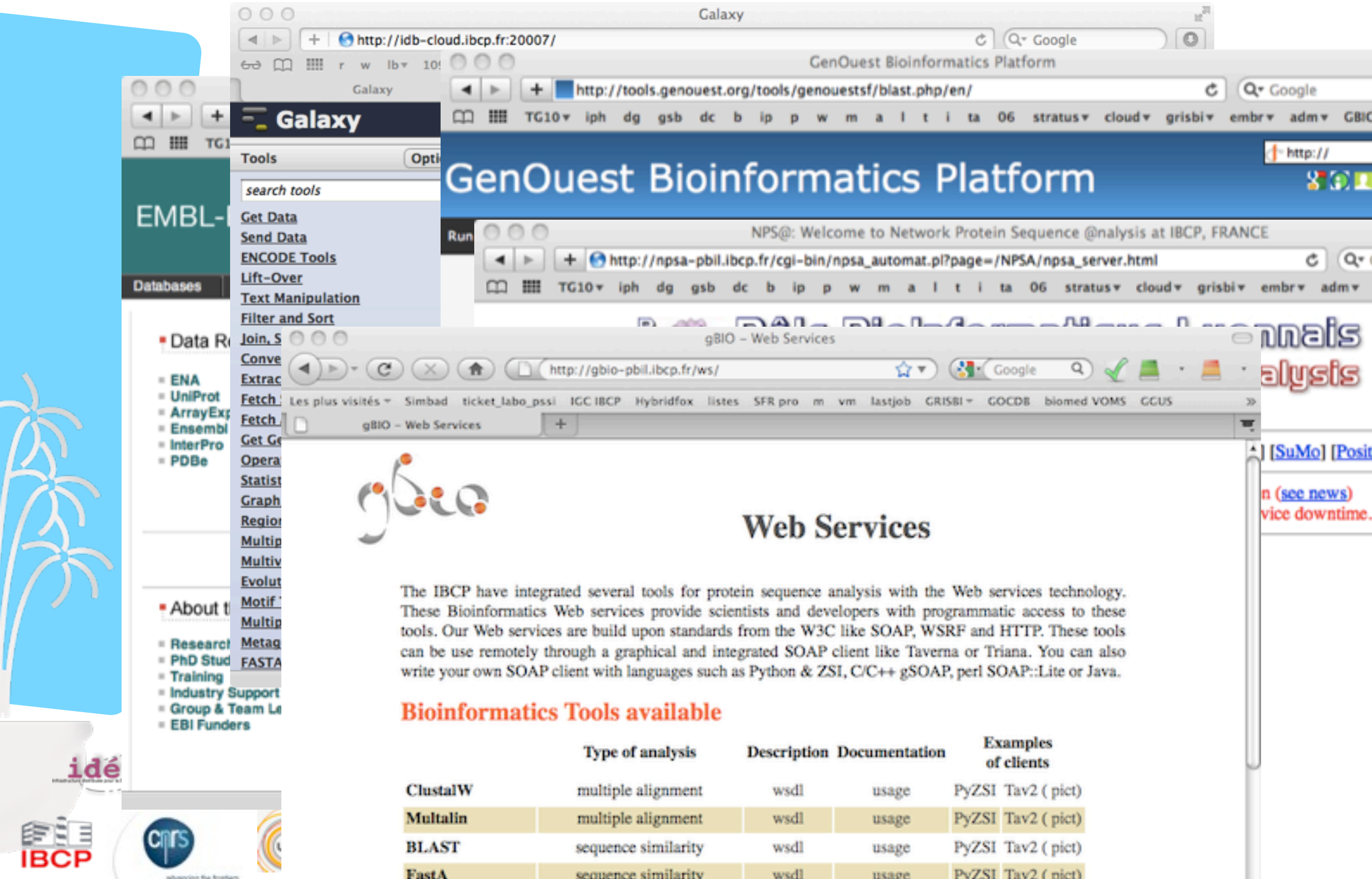
IBCP

Cloud Paris Sud, Orsay 27 Nov. 2012

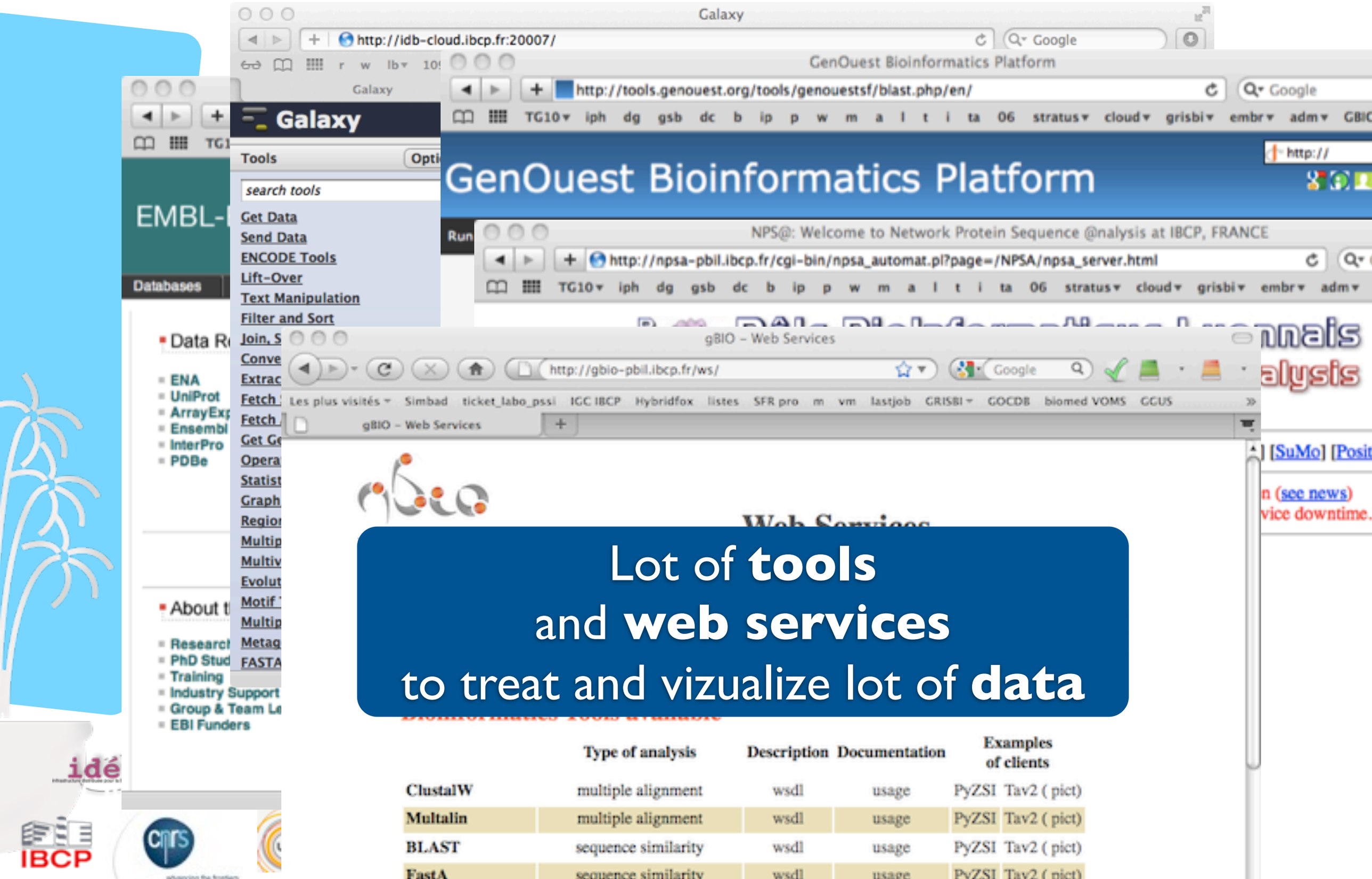
Infrastructure in Biology



Infrastructure in Biology



Infrastructure in Biology

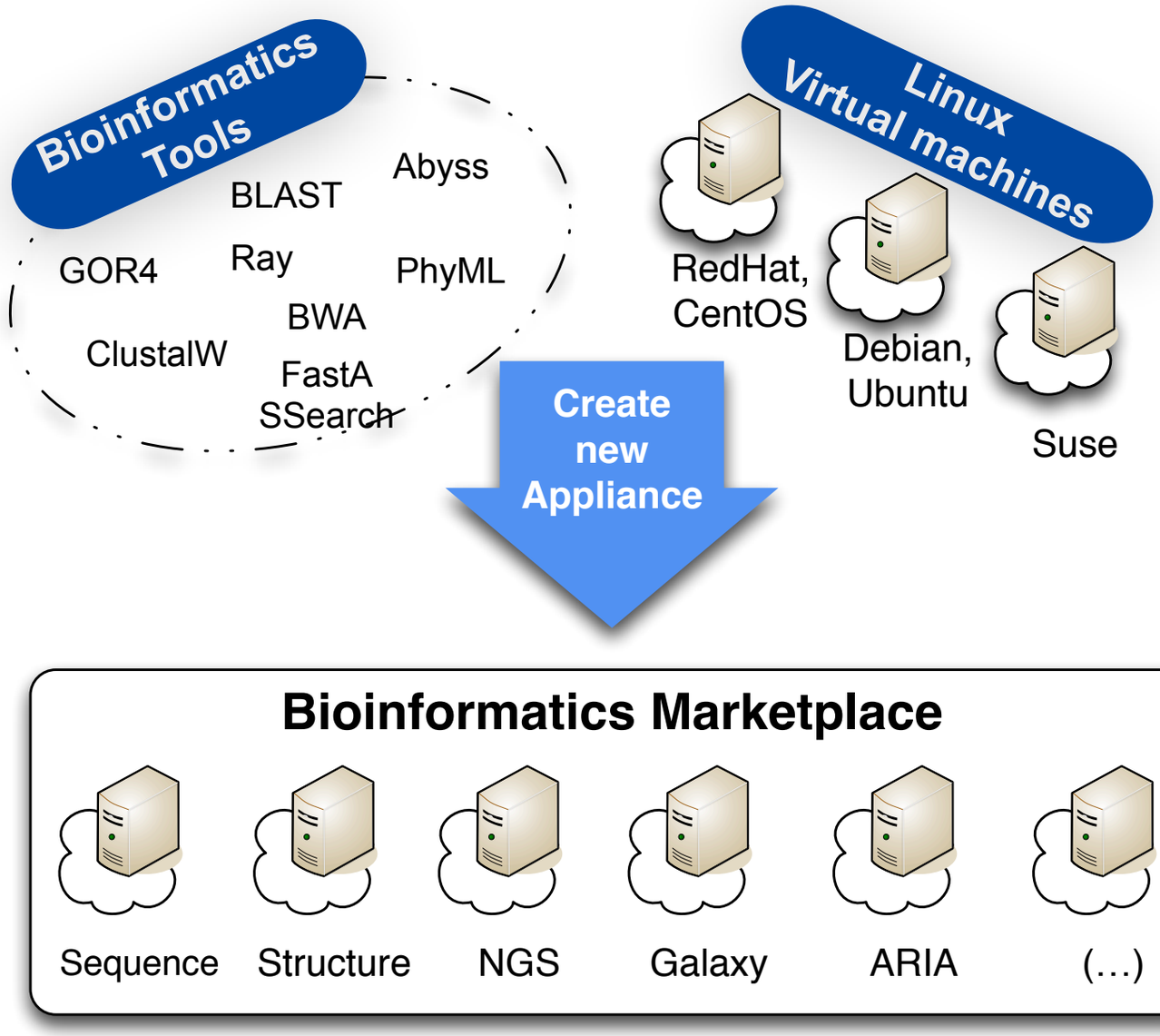


The scene

- Core facility providers
  - Is it easy to deploy lot of (incompatible) tools ?
  - To make them connected to public databases ?
  - To limit transfer of huge data ?
  - To provide users with their own computing resources ?
  - With their own isolated storage ?

- Scientific users
  - Is it easy to access/use these tools ?
  - To adapt to your usage ?
  - To get your/other tools deployed on a datacenter ?
  - To combine them ?

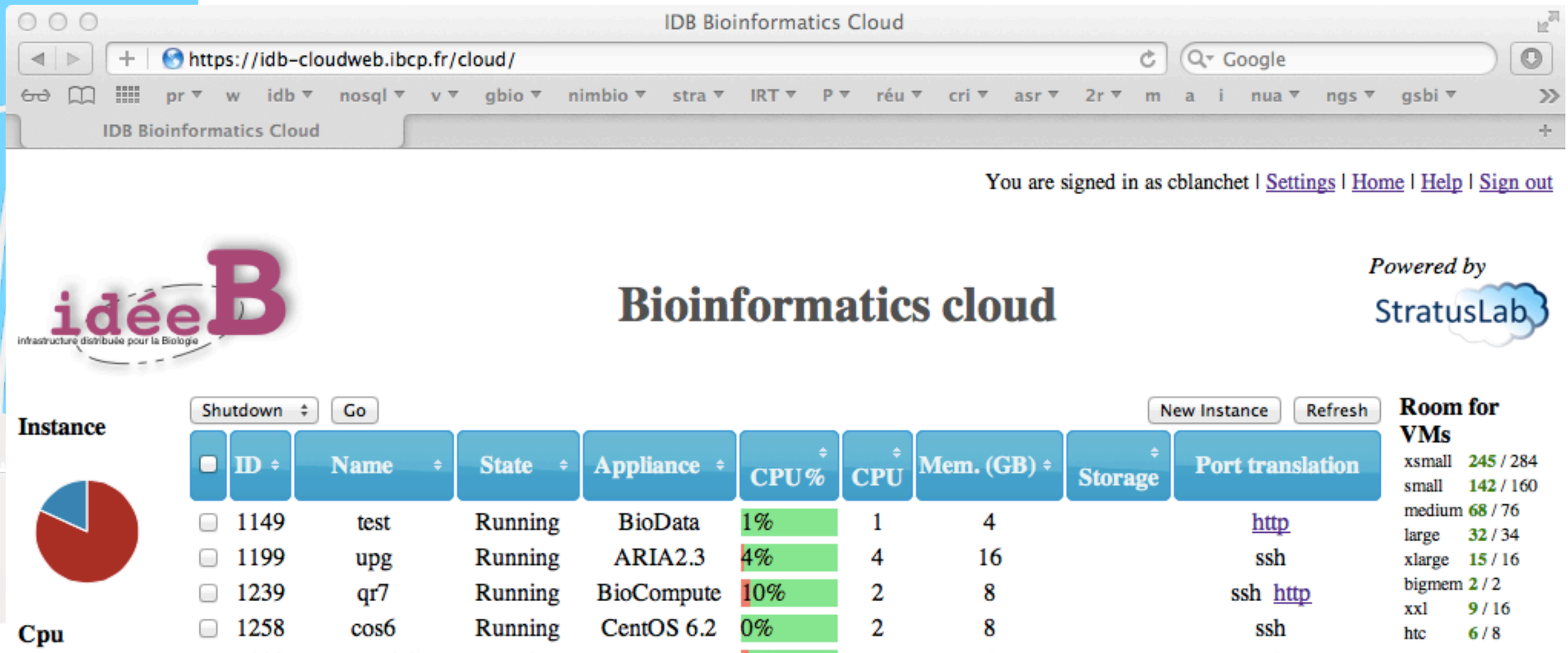
Integrate Bioinformatics Tools in Cloud



- Predefined virtual machines
  - small : few GB, easy to convert in most virtualization formats
- Installed and pre-configured with common bioinformatics tools
  - e.g. BLAST, Clustalw, ARIA, MEME, HMMer, TopHat, BWA, Samtools, etc.

IDB's Cloud Workbench

- Demonstrate usefulness of cloud for Biology
- 13 turnkey bioinformatics appliances (as of Nov. 2012)
- Running since Sept. 2011, opened to Biology community
  - 184 cores, 536 GB RAM, 36 TB de stockage
  - Powered by Toolkit StratusLab (EU FP7 INFSO-RI-261552) Rel. 1.4
- Specific developments of a Web interface ready for biologists



Bioinformatics Appliances

ARIA 2.3

Endorser: *christophe.blanchet@ibcp.fr*
Identifier: *N_zDsconV86gvkjZtt7D-ePv4M6*
Created: *2012-10-11T14:01:38Z*

This appliance is part of the StratusLab bioinformatics usecase TOSCANI (TOwards StruCtural AssignmeNt Improvement). The goal is to improve the determination of protein...

More...

CentOS 6

Endorser: *christophe.blanchet@ibcp.fr*
Identifier: *Bl8HibMjBB6uu231adQUkqyGtnI*
Created: *2012-09-27T12:04:15Z*

A minimal installation for CentOS 6.x. Only root account configured. Firewall enabled with SSH and HTTP port open. SELinux disabled. Enhanced StratusLab contextualization used...

More...

BIO data

Endorser: *christophe.blanchet@ibcp.fr*
Identifier: *FtCJFZ7xO5uxKyzThGRX9Ex5cqR*
Created: *2012-09-24T08:56:26Z*

Biological databases repository appliance built by IDB-IBCP (CNRS, Lyon, France. http://idee-b.ibcp.fr). The following databases are installed and available: SwissProt,...

More...

Mobyle

Endorser: *christophe.blanchet@ibcp.fr*
Identifier: *NaCGZfy9NxClc3ISU158RHrG0ik*
Created: *2012-09-07T14:20:38Z*

This appliance provides cloud users with a fully functional Mobyle portal. Mobyle is a framework and web portal

Bioinformatics Appliances (2)

Protein identification

Endorser: *christophe.blanchet@ibcp.fr*
Identifier: *H6KPqxYlZRdlhPhs2ZKIENiiVyx*
Created: *2012-10-23T14:05:07Z*

Bioinformatics virtual appliance for protein identification from mass spectrometry data. Contains OMSSA, X!Tandem, PeptideShaker and SearchGUI tools. Constructed by IDB...

More...

Galaxy portal

Endorser: *christophe.blanchet@ibcp.fr*
Identifier: *OqucGN3bQD9FdIenGRlqZ4ZNNHW*
Created: *2012-10-11T15:11:59Z*

Bioinformatics gateway appliance configured with the GALAXY portal, built by CNRS IBCP-IDB. You have also access to the pre-installed bioionformatics tools through the web...

More...

Hadoop MapReduce

Endorser: *clement.gauthey@ibcp.fr*
Identifier: *PElfkAp5mOwULVh1KLsprFcji0s*
Created: *2012-10-11T14:42:36Z*

This appliance provides an easy way to deploy an Hadoop MapReduce cluster. You just need to run the bash script hadoop-create-cluster with a nodes list and an username in parameters and wait few...

More...

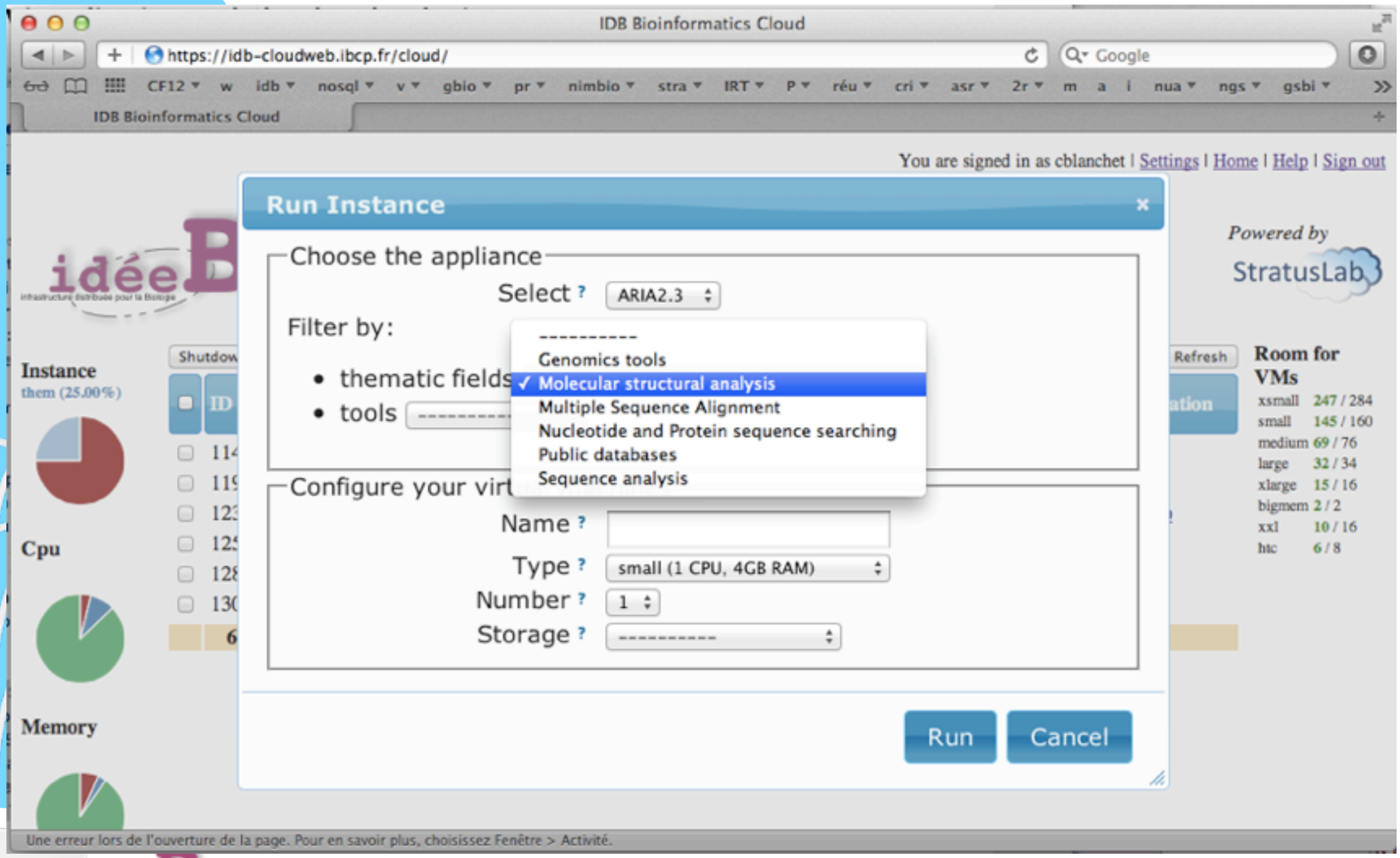
BIO compute node

Endorser: *christophe.blanchet@ibcp.fr*

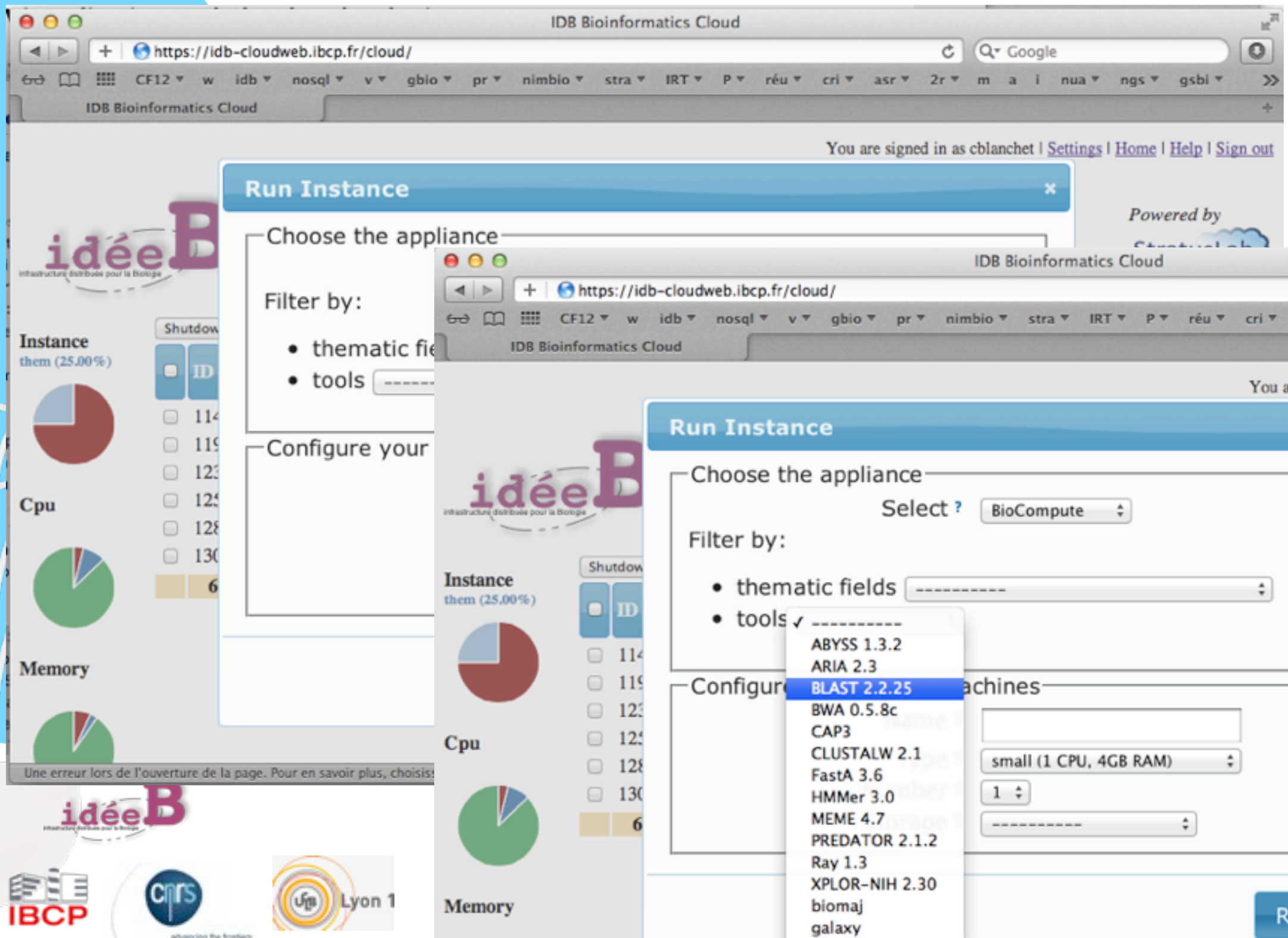
idéeB

IBCP cnrs Lyon 1

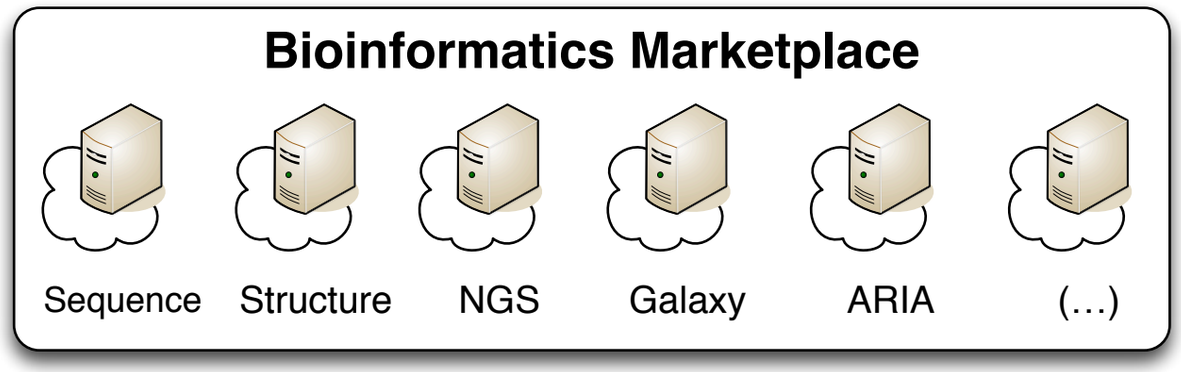
Select your bioinformatics tools

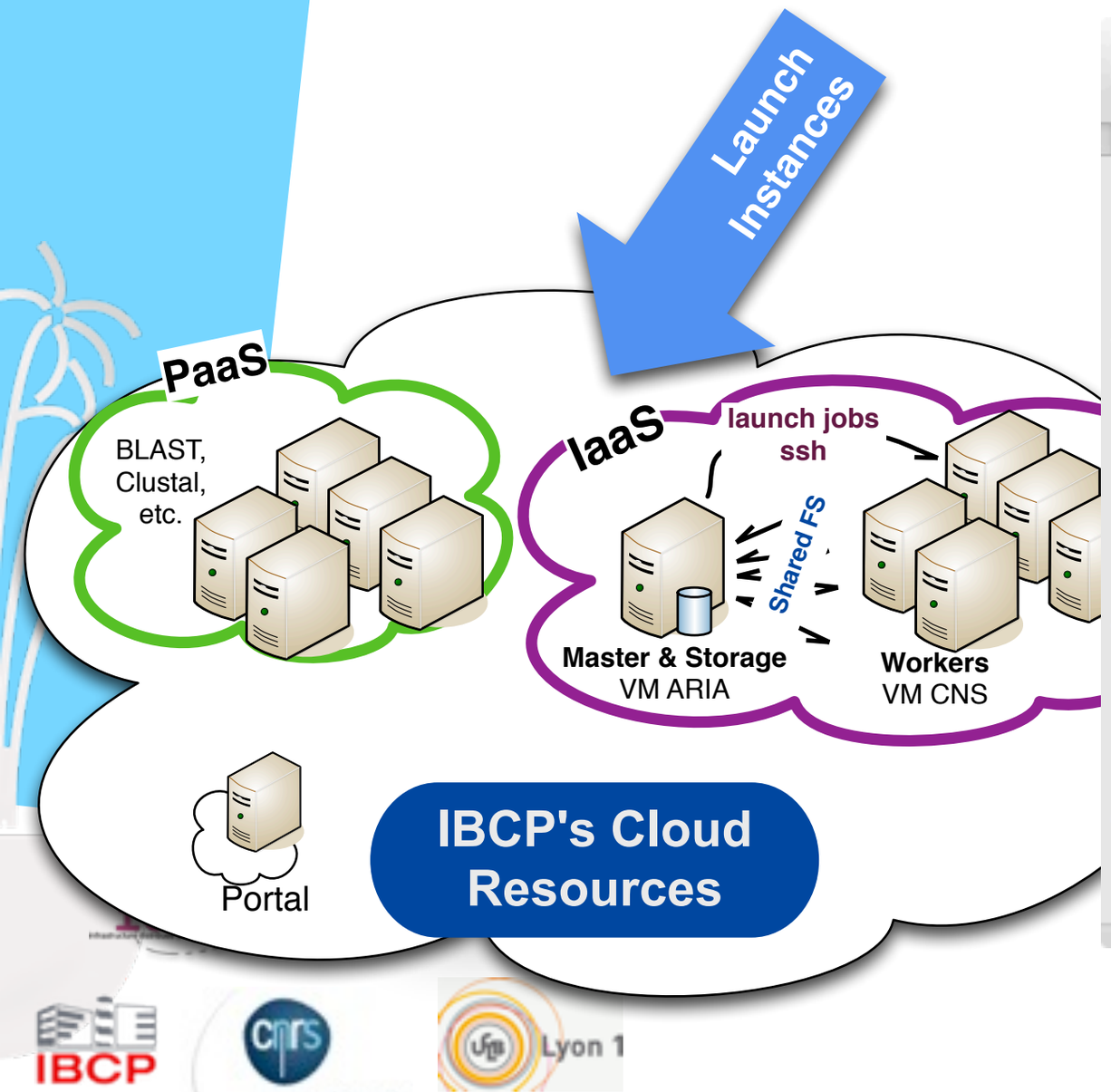


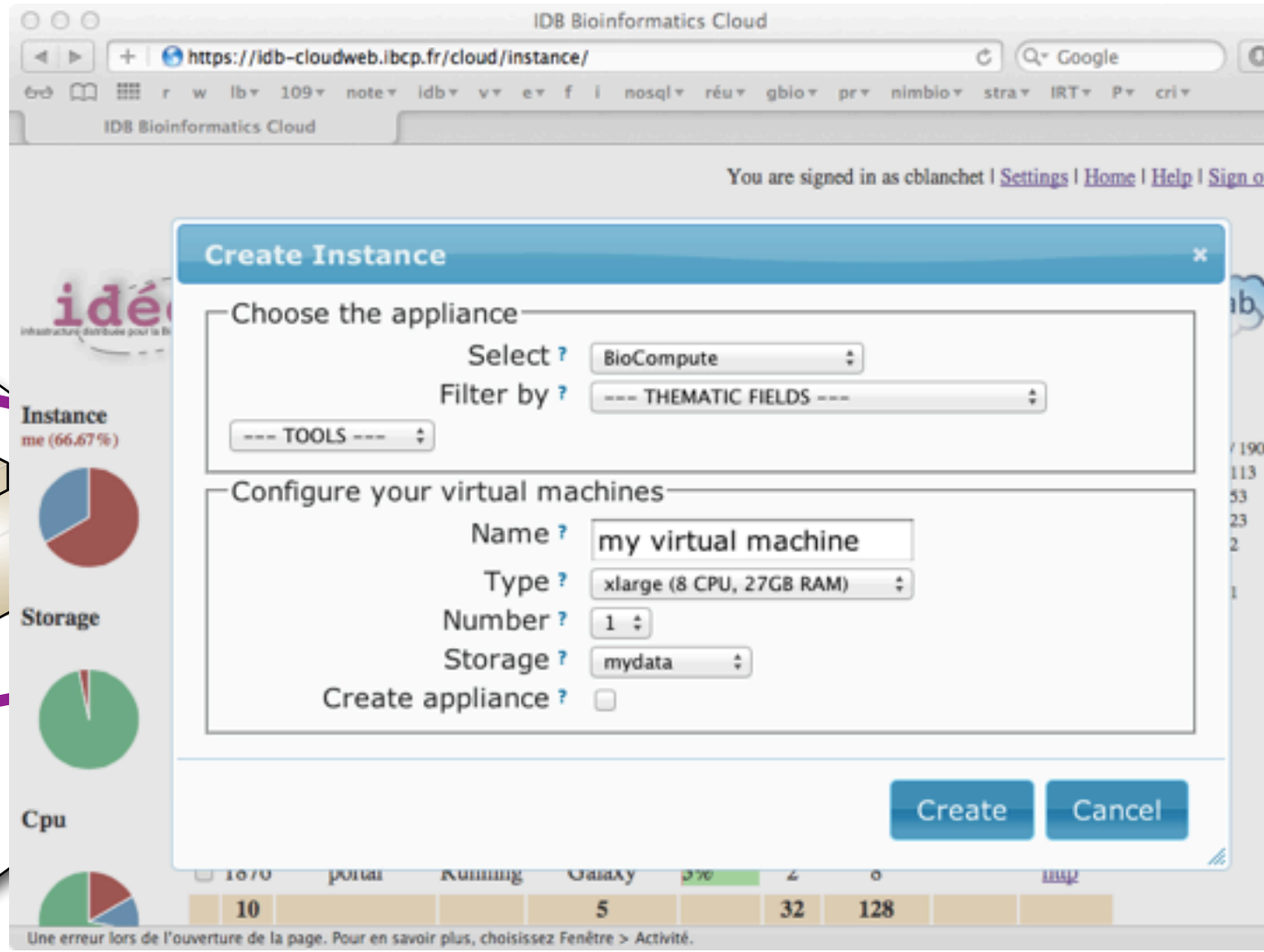
Select your bioinformatics tools



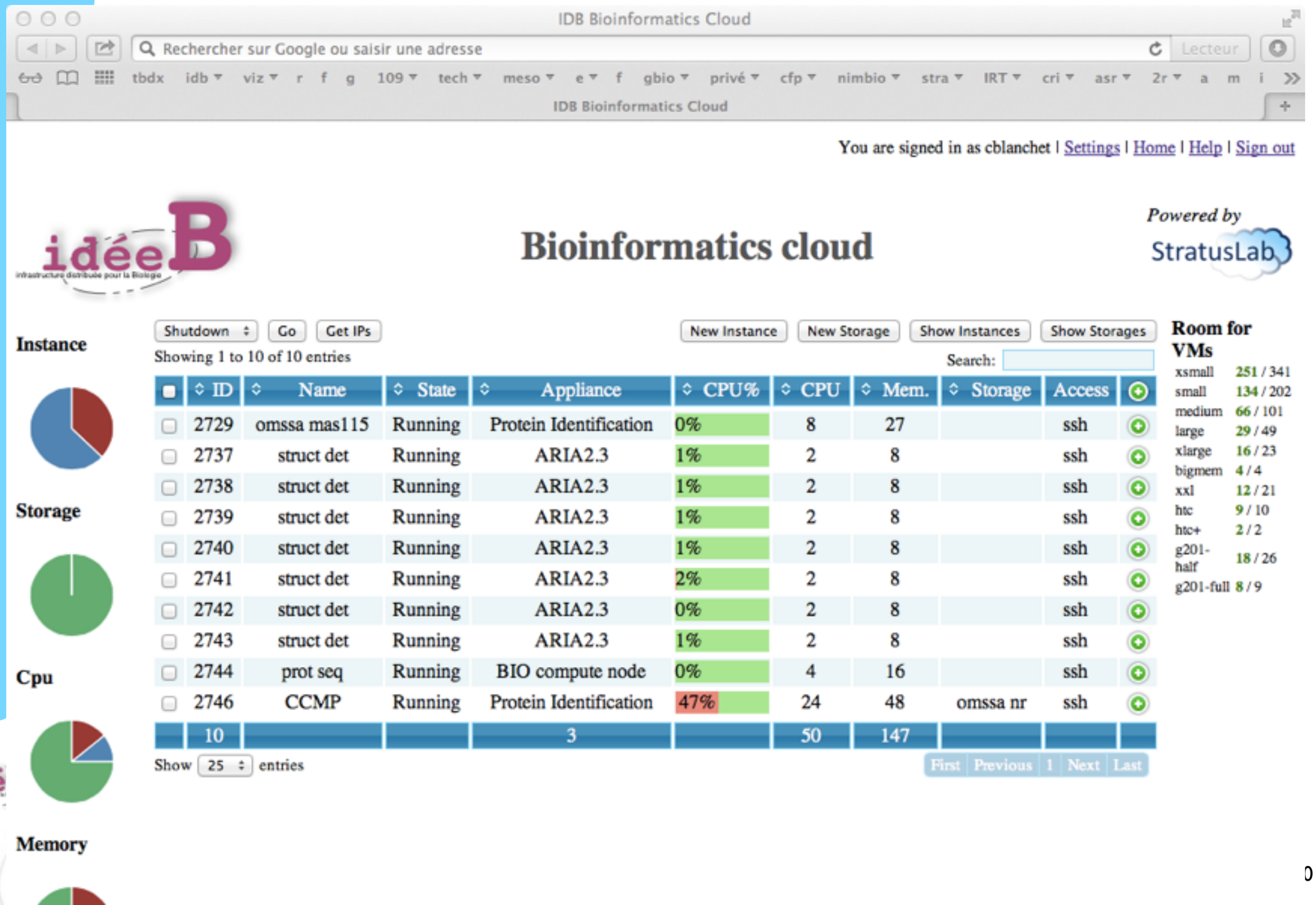
Run Bioinformatics Cloud Instances



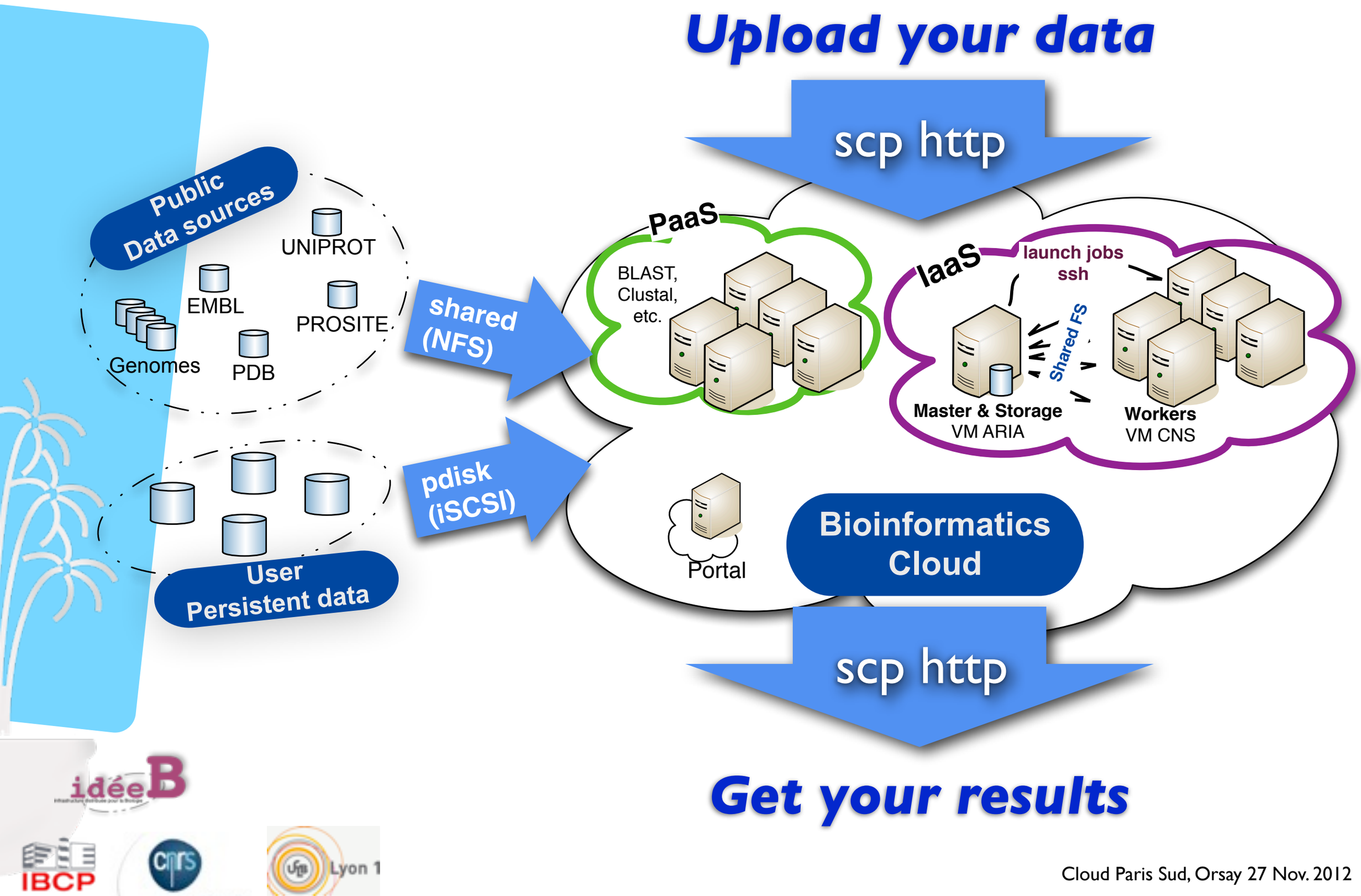




Manage your Cloud Instances



Data in cloud



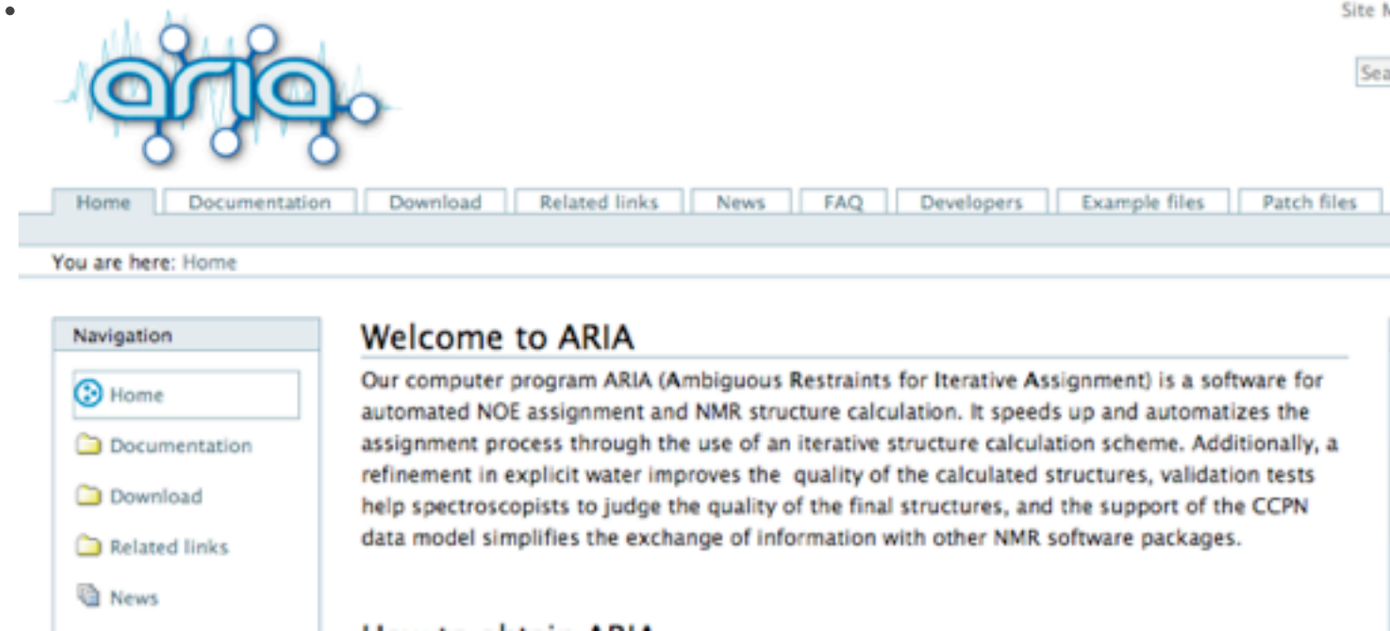
Examples

- Structural Biology: the TOSCANI usecase
- Galaxy portal for NGS analyses
- Proteomics

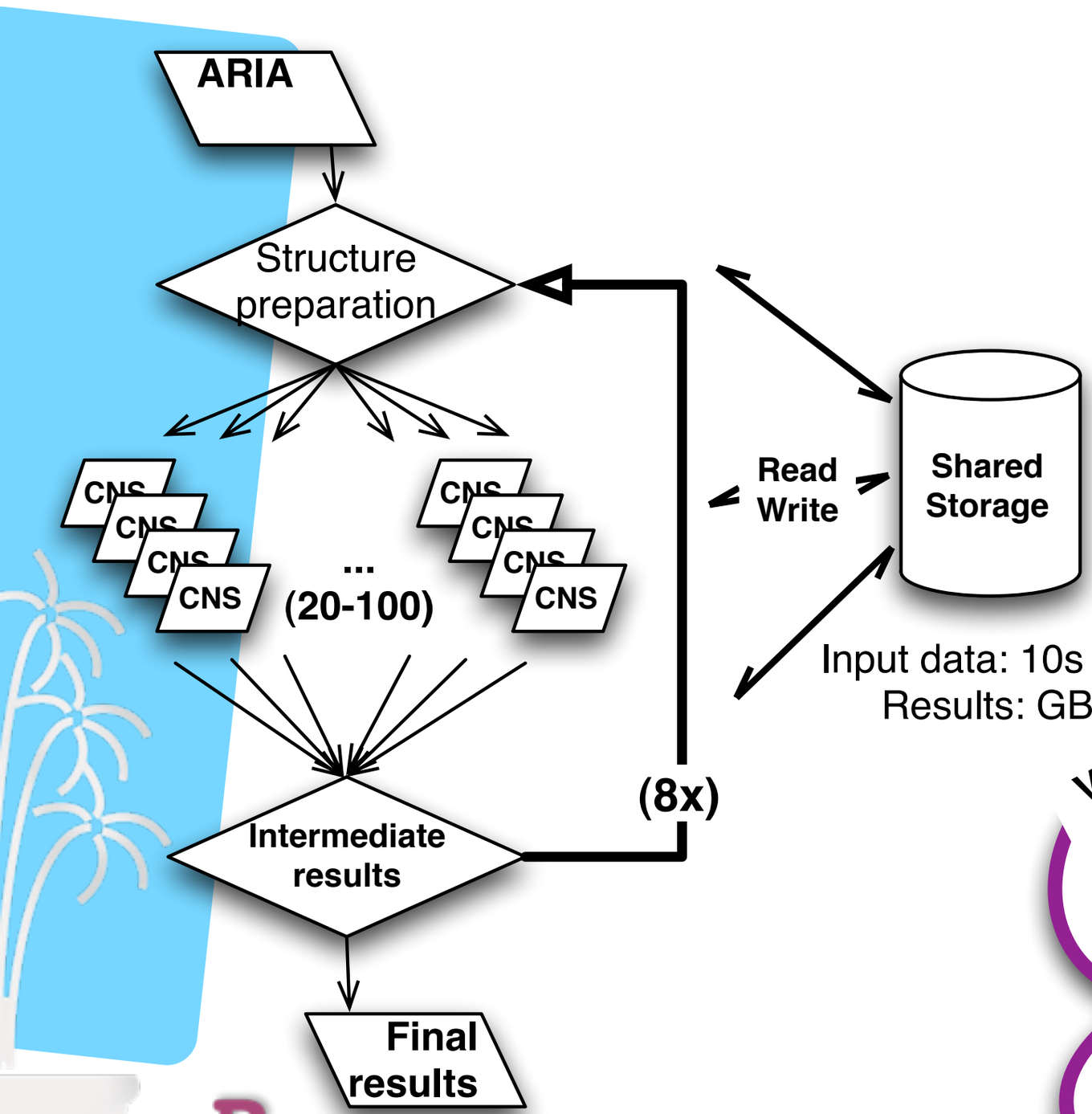
Structural Biology

- TOwards StruCtural AssignmeNt Improvement
  - To improve the determination of protein structures based on Nuclear Magnetic Resonance (NMR) information with ARIA software
  - Large computational needs.
  - A NMR laboratory will not specially invest in building a cluster of about 100 nodes to be able to run such NMR structure calculations.
  - Flexibility of the cloud to deploy the different required bioinformatics tools can accelerate such a procedure.
  - Commercial interest in providing such tools to structural biologists on a “pay as you go” basis.

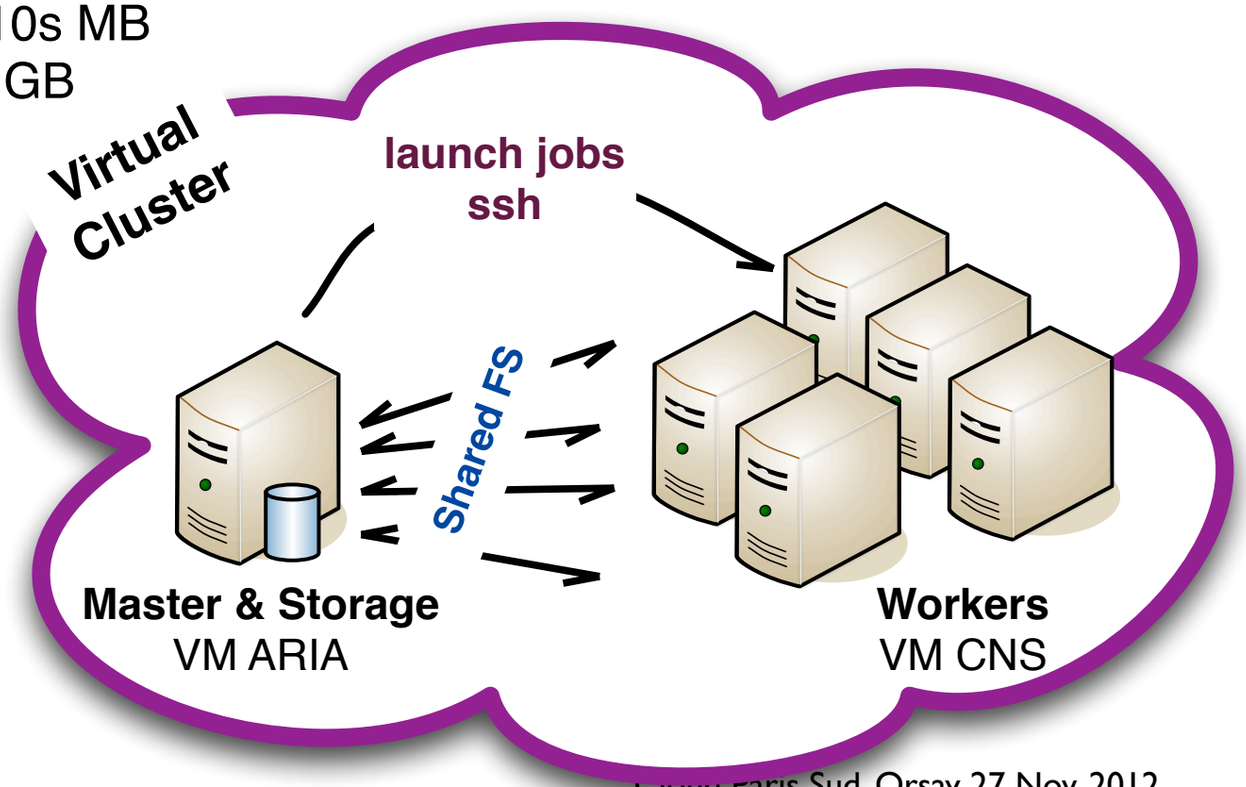
- *Endorsers:*
  Institut Pasteur Paris
  and CNRS IBCP



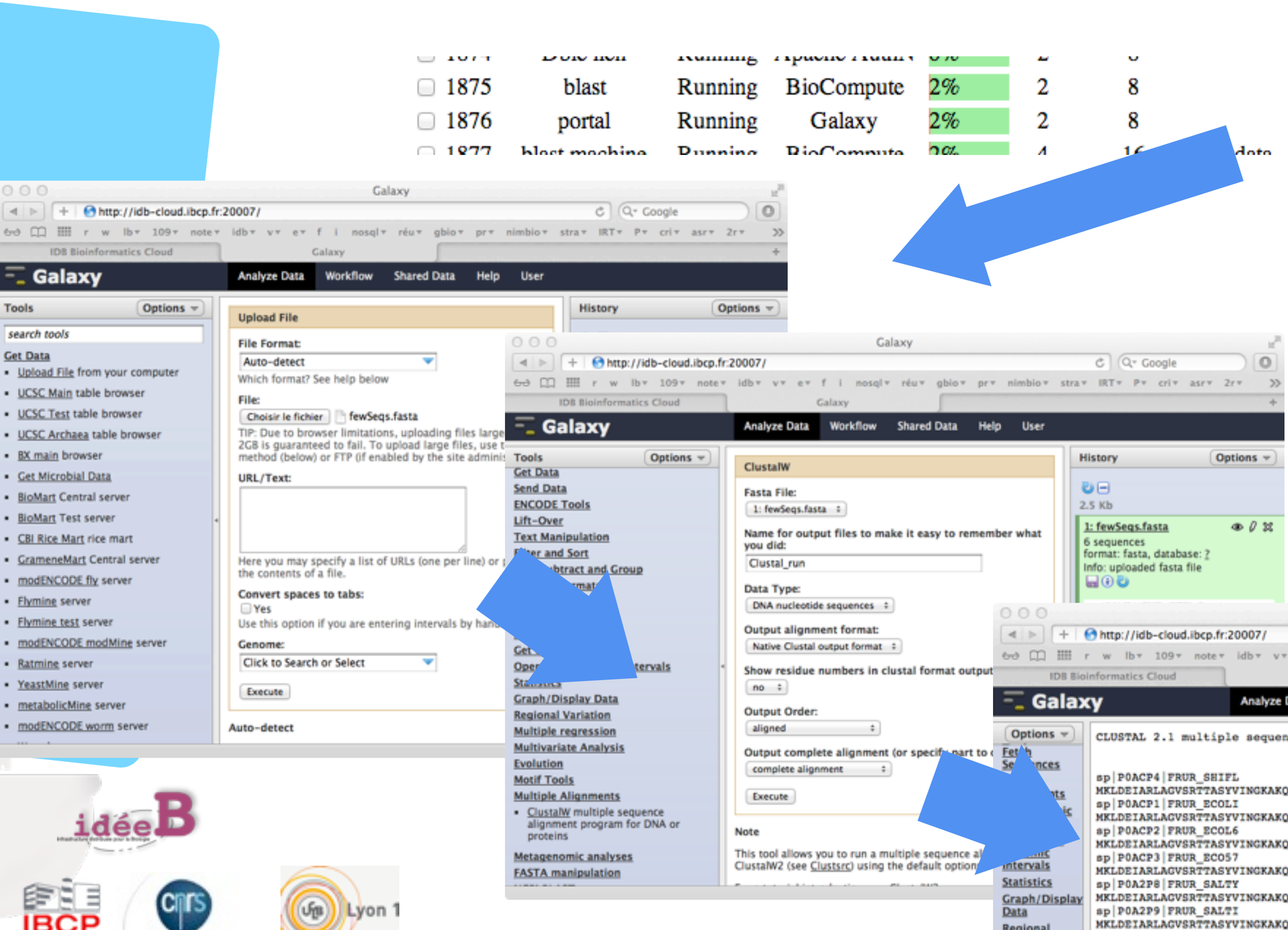
IaaS deployment of ARIA



Significant increase in the number of calculated protein conformations improves the statistics on the NMR conformations and can help to overcome the ambiguity bottleneck.



Galaxy portal for NGS analyses



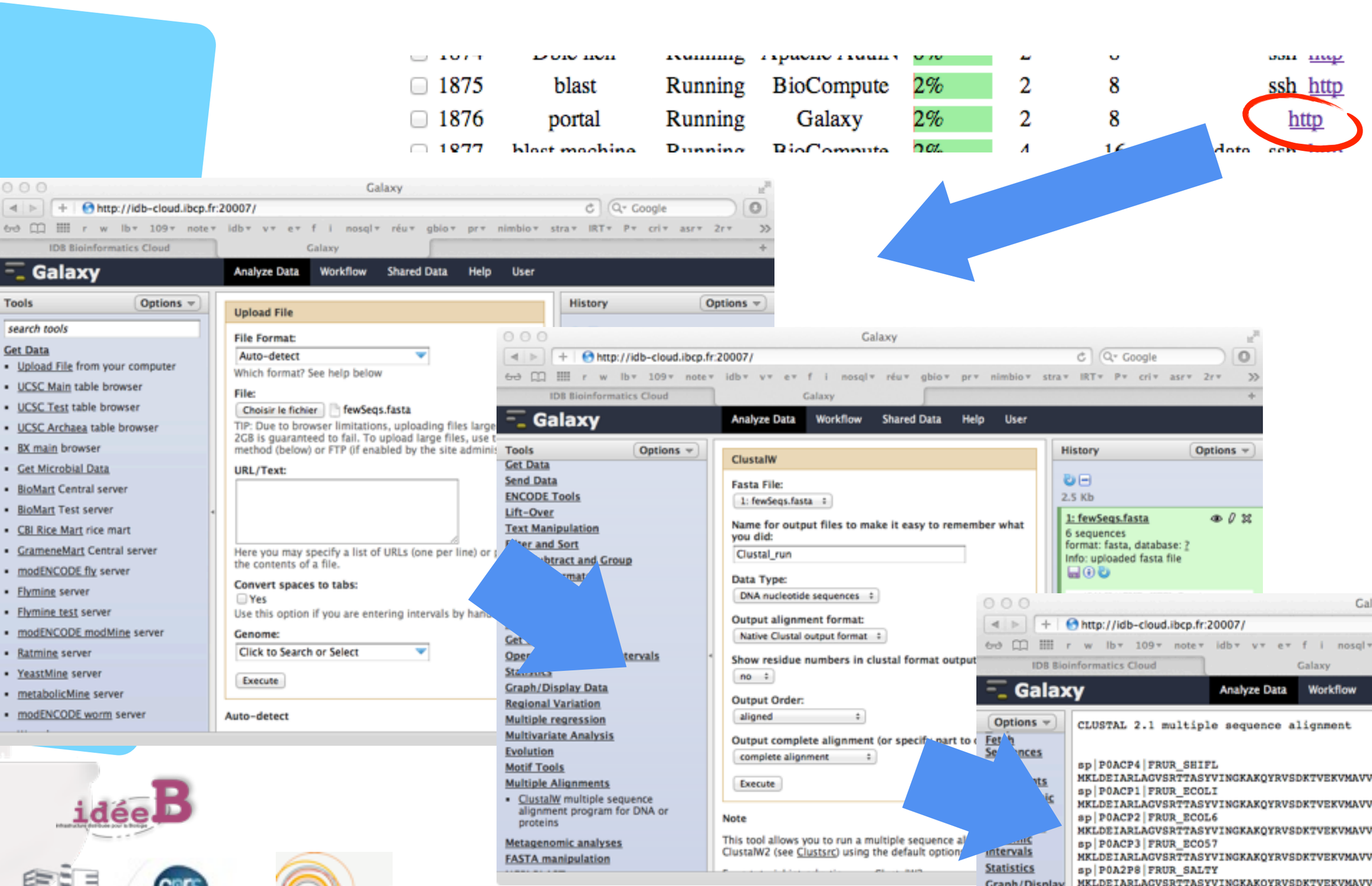
idéeB

IBCP

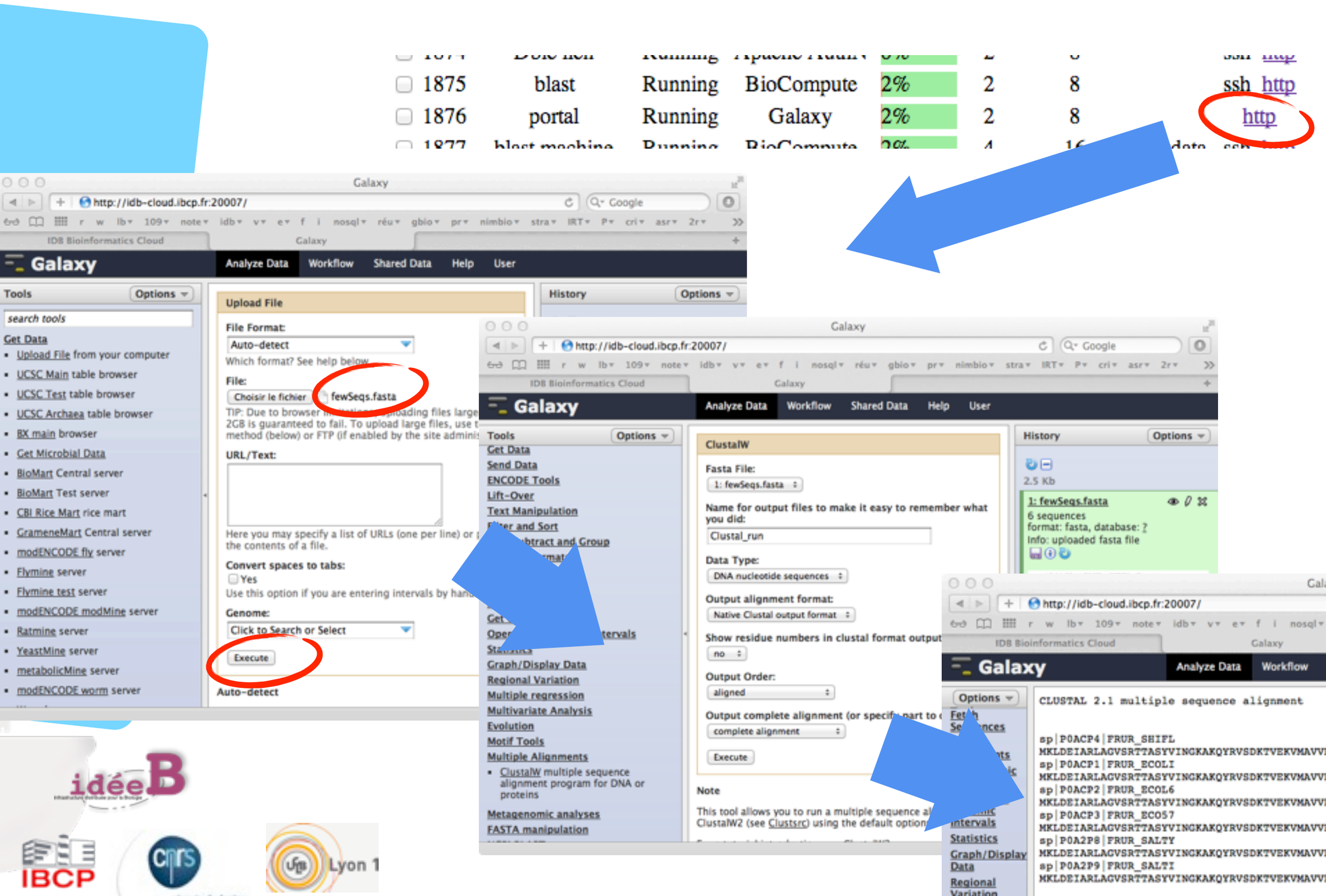
cnrs

Lyon 1

Galaxy portal for NGS analyses



Galaxy portal for NGS analyses



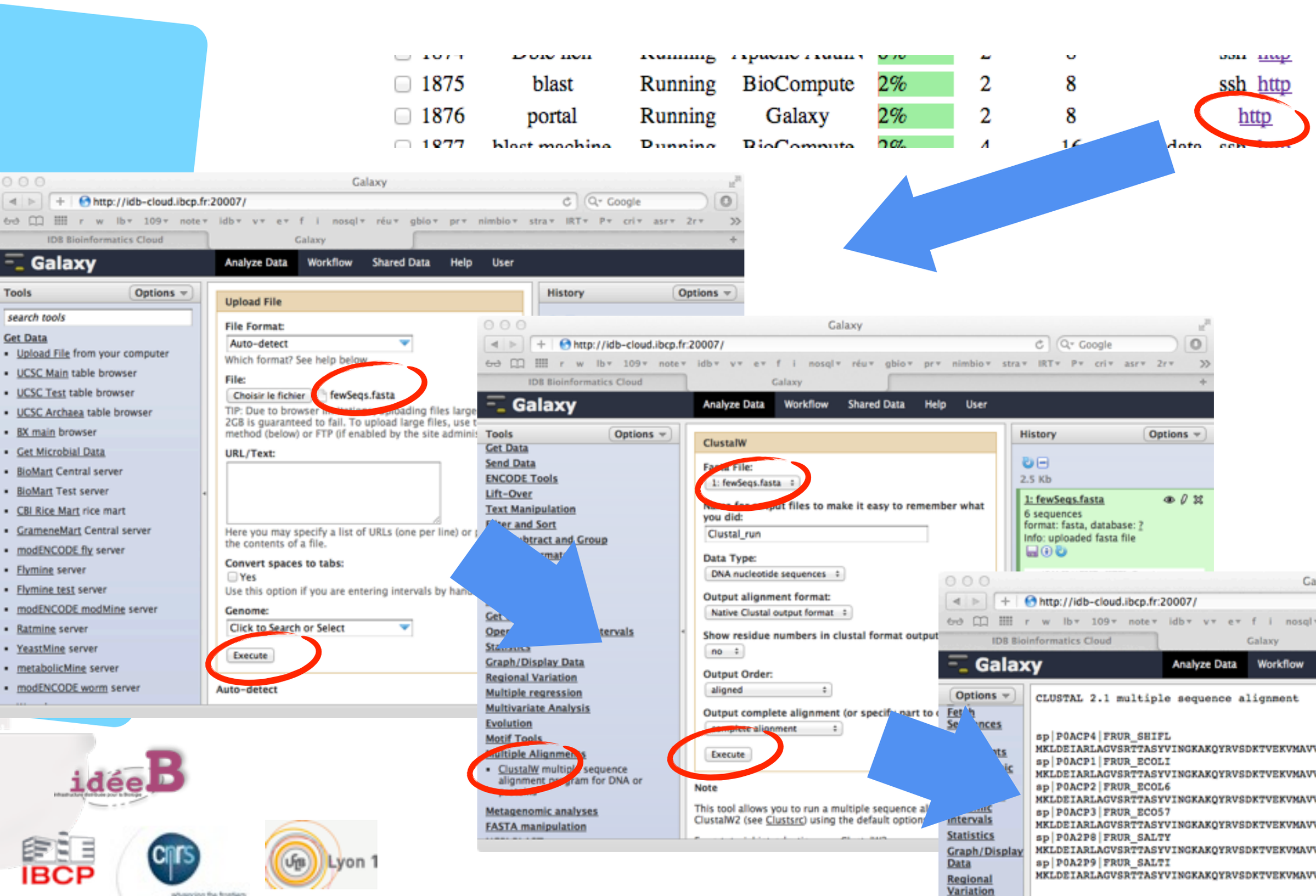
idéeB

IBCP

cnrs

Lyon 1

Galaxy portal for NGS analyses



idéeB

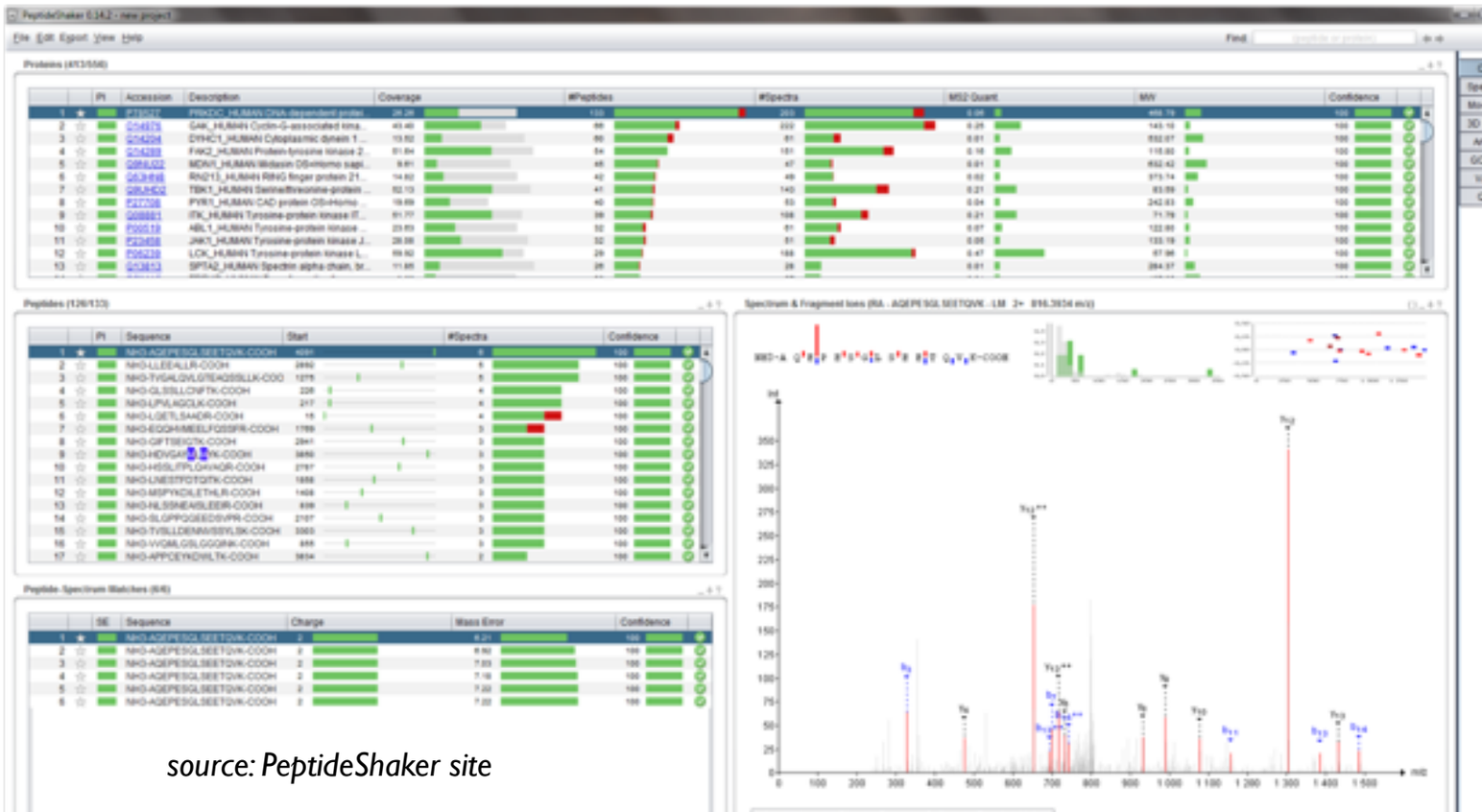
IBCP

cnrs

Lyon 1

Proteomics

- Motivation
  - Mass spectroscopy platform
  - Running out of space on their local resources
- Protein identification
  - Mass spectroscopy experimental data
  - reference databases : nr, Swiss-Prot
  - Screening tools: OMSSA, X!Tandem
- User interface
  - Remote display
  - Common GUIs
    - SearchGUI
    - PeptidShaker



source: PeptideShaker site

Conclusion

- Provide turnkey bioinformatics cloud services
  - Standard tools and pipelines
  - Ready to run on core facilities and commercial datacenters
  - Easier to transfer appliances than data (GB vs TB)
- Cloud infrastructure tightly connected to existing bioinformatics infrastructure
  - Public IDB's bioinformatics cloud
  - Linked to public biological databases
  - Collaboration with national Research Infrastructures like the French RENABI GRISBI
- Ease the access
  - Web interface for cloud management
  - Usual scientific gateways
  - Persistent and large ubiquitous storage

What's next ?

- Help bioinformatics centres to provide academic and commercial community with bioinformatics services!
  - Pre-defined bioinformatics appliances, webservices and portal (PaaS)
  - Multi-nodes applications, e.g. ARIA, or comprehensive pipelines (IaaS)
  - Referenced in a bioinformatics marketplace
  - Ready to deploy on the future cloud infrastructure of the French Bioinformatics Institute (IA ReNaBi-IFB)

Questions ?

Platform ‘ Infrastructure Distributed for Biology - IDB ’ acknowledges

Institut Pasteur: M Nilges, T Malliavin, F. Mareuil
StratusLab partners

co-funding by the European Community's Seventh Framework Programme (INFSO-RI-261552) and the French National Research Agency's Arpege Programme (ANR-10-SEGI-001)

http://idee-b.ibcp.fr