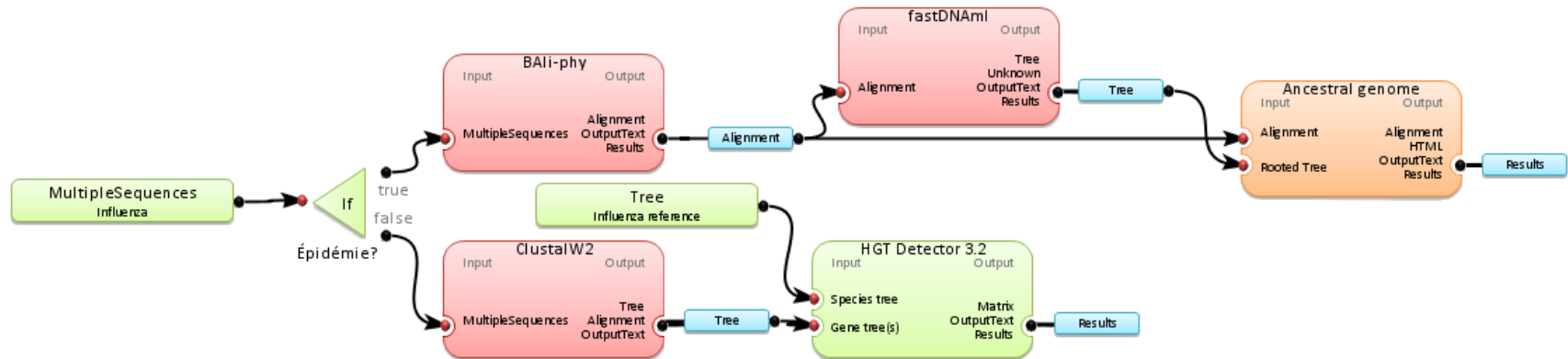


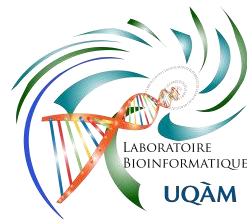
# Application des flux de travaux en bioinformatique



Etienne Lord, *Ph.D.*

Chercheur et développeur chez Agriculture Canada

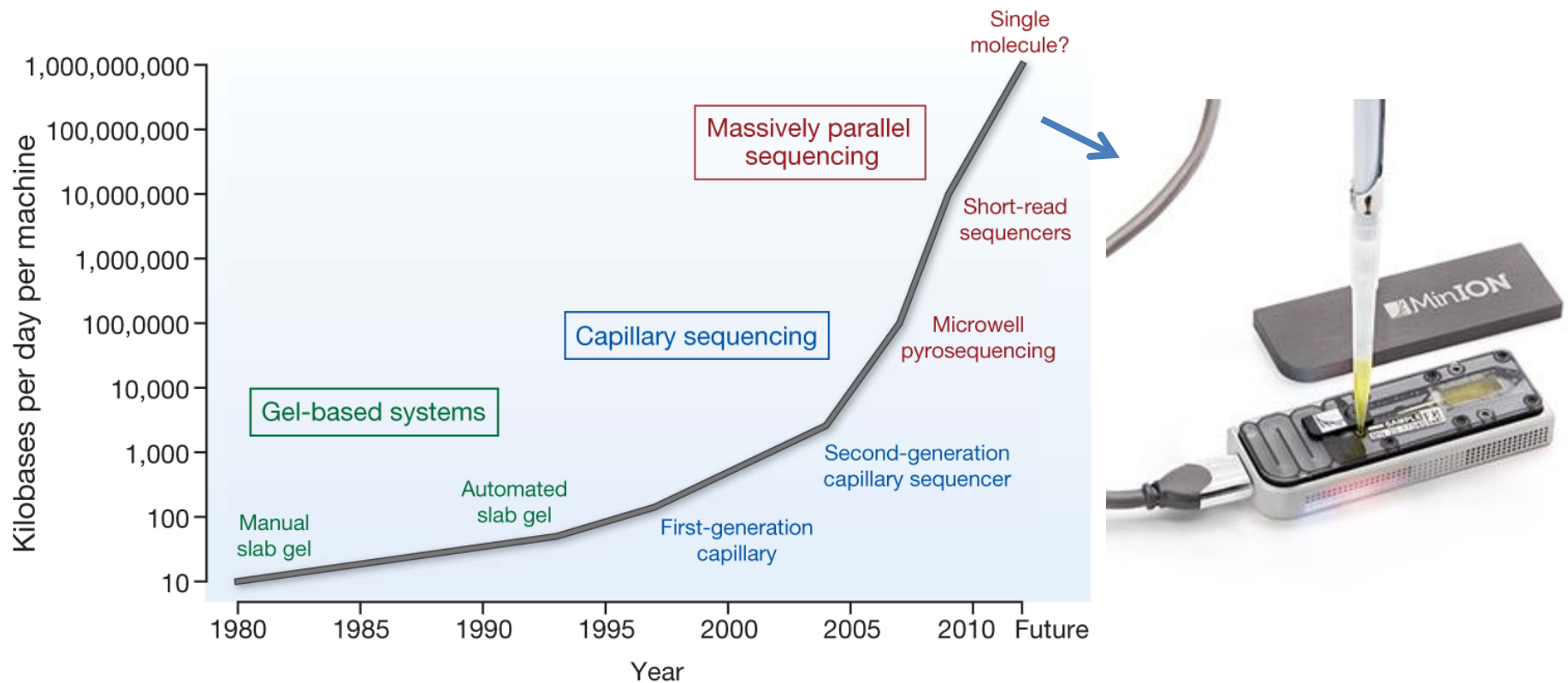
20 février 2018



- ❑ **L'évolution de la bioinformatique**
- ❑ **Les flux de travaux / flux de données**
  - ❑ Nomenclature des flux de données
  - ❑ Systèmes de gestions de flux de travaux
- ❑ **Les systèmes de gestions de flux de travaux en bioinformatique**
  - ❑ Galaxy
  - ❑ Taverna
  - ❑ Armadillo

} Et quelques autres...
- ❑ **La plate-forme *Armadillo* orientée vers la phylogénétique**
  - ❑ Cas d'utilisations
  - ❑ Orientation dans le développement et défis
- ❑ **Références**

# Évolution des techniques en biologie (problématique 1)

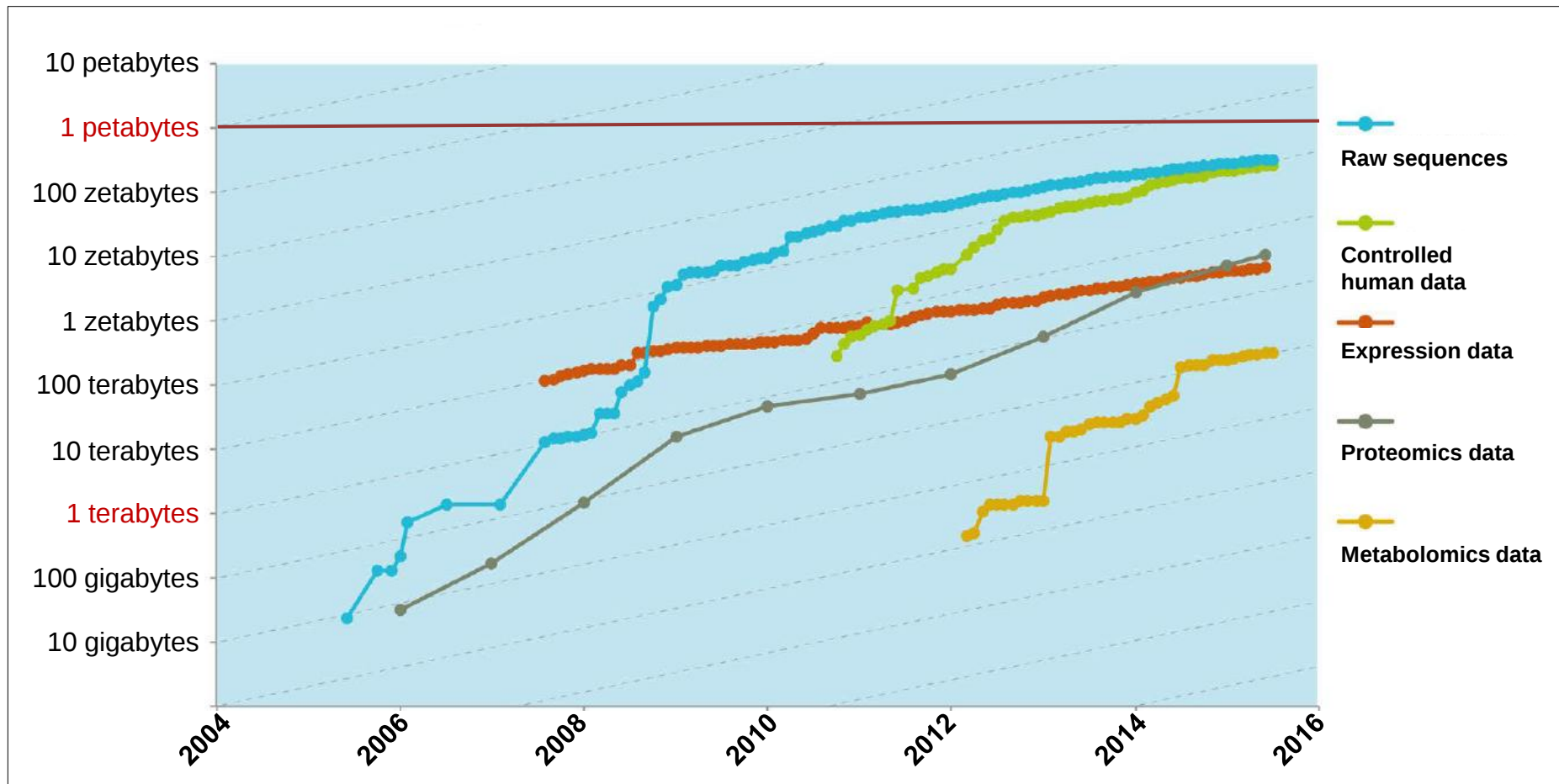


**Figure 3 | Improvements in the rate of DNA sequencing over the past 30 years and into the future.** From slab gels to capillary sequencing and second-generation sequencing technologies, there has been a more than a million-fold improvement in the rate of sequence generation over this time scale.

Séquenceur *MinION* de Oxford Nanopore  
[publications.nanoporetech.com/](https://publications.nanoporetech.com/)

Stratton MR, Campbell PJ, Futreal PA. (2009) The cancer genome. *Nature*. 458(7239):719-24.

# Évolution de la quantité des données (problématique 2)

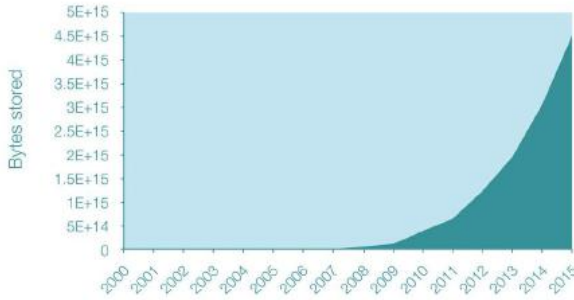


EMBL-European Bioinformatics Institute (2016) EMBL-EBI annual scientific report 2015

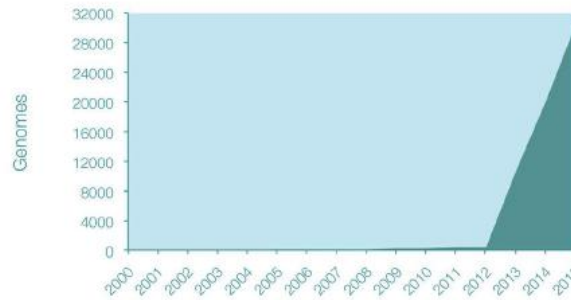
[http://www.ebi.ac.uk/sites/ebi.ac.uk/files/groups/external\\_relations/Documents/EMBL\\_EBI\\_ASR\\_2015\\_DigitalEdition23Jul2016.pdf](http://www.ebi.ac.uk/sites/ebi.ac.uk/files/groups/external_relations/Documents/EMBL_EBI_ASR_2015_DigitalEdition23Jul2016.pdf)

# Évolution des types de données (problématique 3)

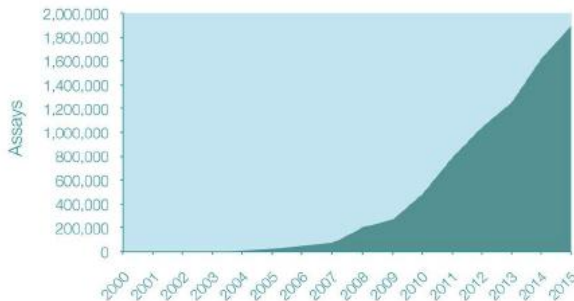
Nucleotide sequence data (compressed)



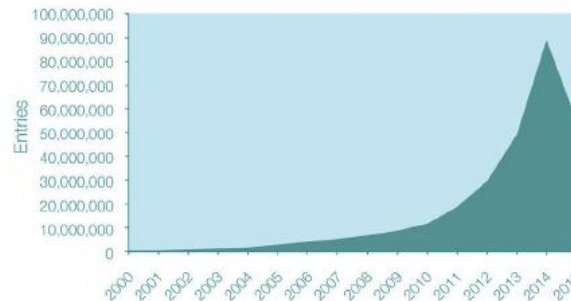
Genomes (all species)



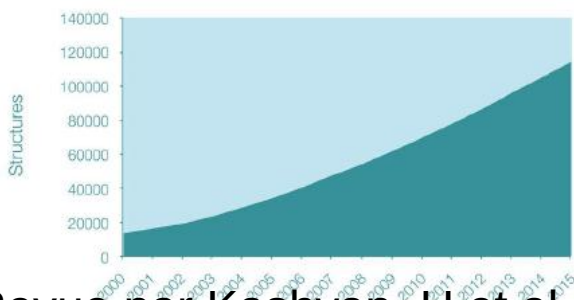
Gene expression data



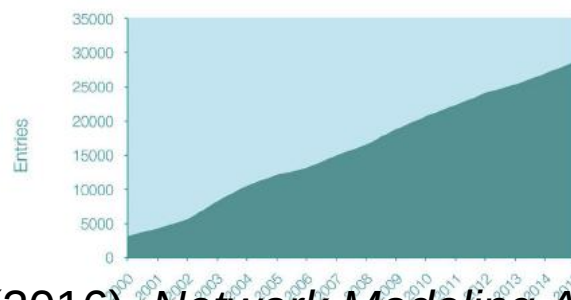
Protein sequence data



Macromolecular structures



Protein families motifs and domains

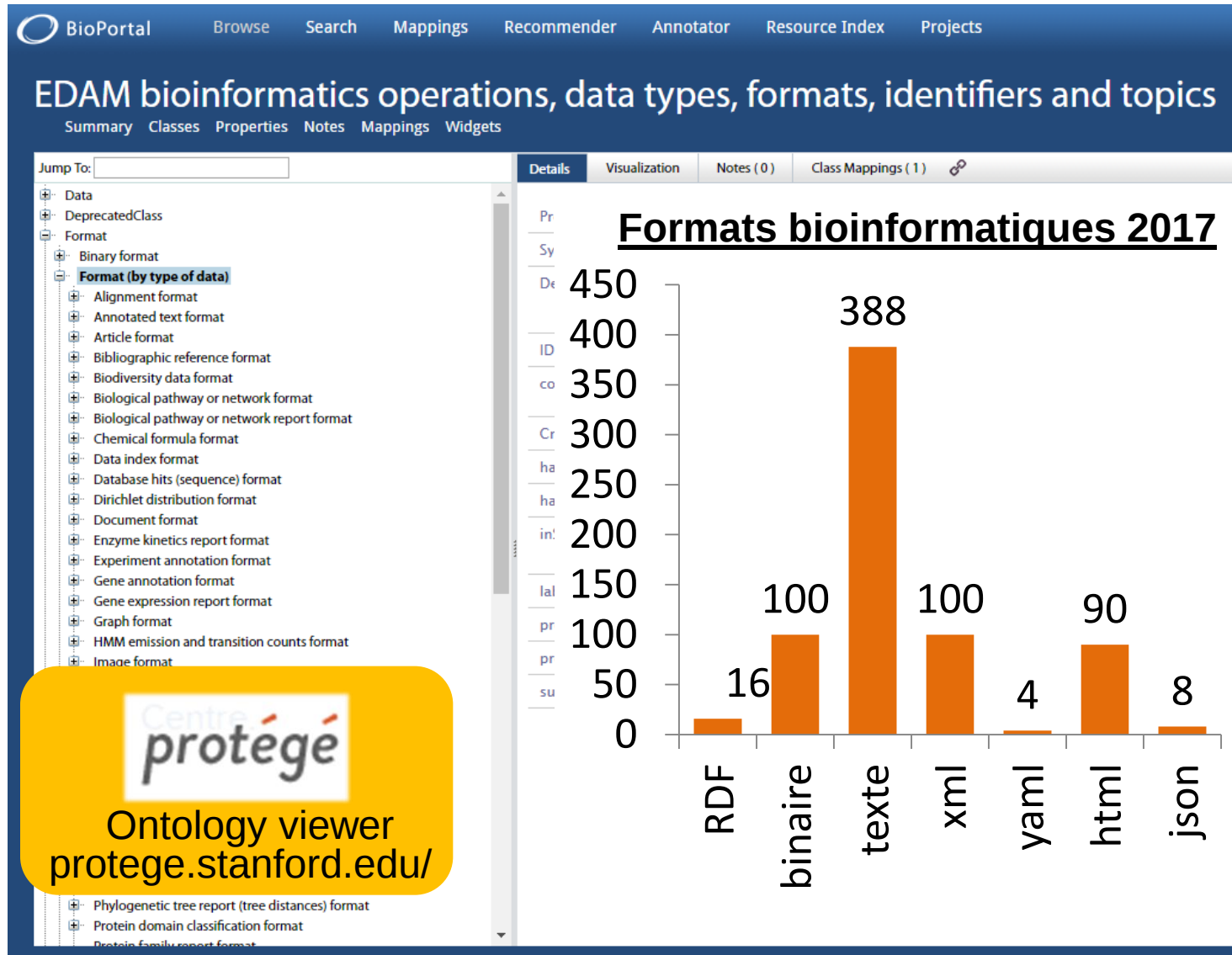


## Types de données

1. **Expression** (gènes, protéines, microarray)
2. **Séquences** (ADN, ARN, protéines).
3. **Structures** (2D, 3D)
4. **Interactions** (protéines-protein interactions)
5. **Réseaux** (pathways) et **ontologies**
6. **Raw data?** (Geo, etc.)

Revue par Kashyap, H et al. (2016). *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5(1), 28.

# Une myriade de formats (problématique 4)



<http://biportal.bioontology.org/ontologies/EDAM?p=classes>

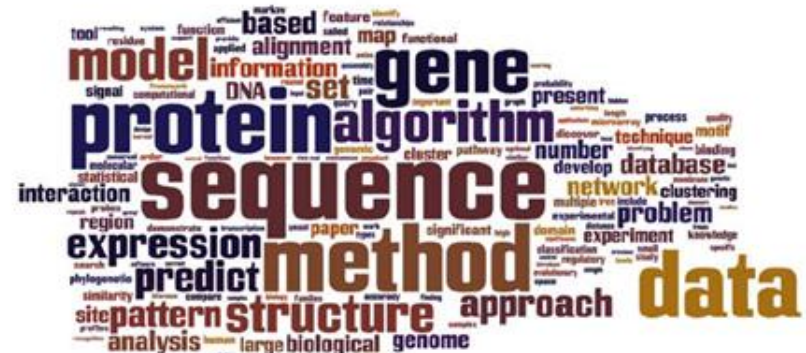
<http://edamontology.org/page>

# L'évolution de la bioinformatique

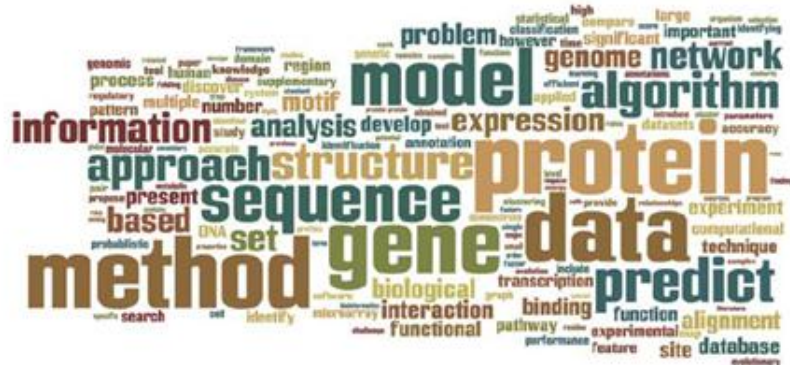
Les **méthodes**, les **réseaux** et les **données** deviennent de plus en plus importantes dans la recherche.



## 1993-1997



## 1998-2002



## 2003-2007

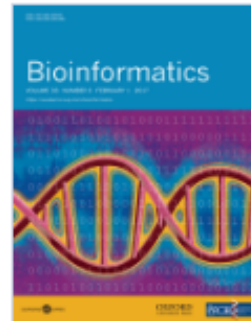


## 2008-2012

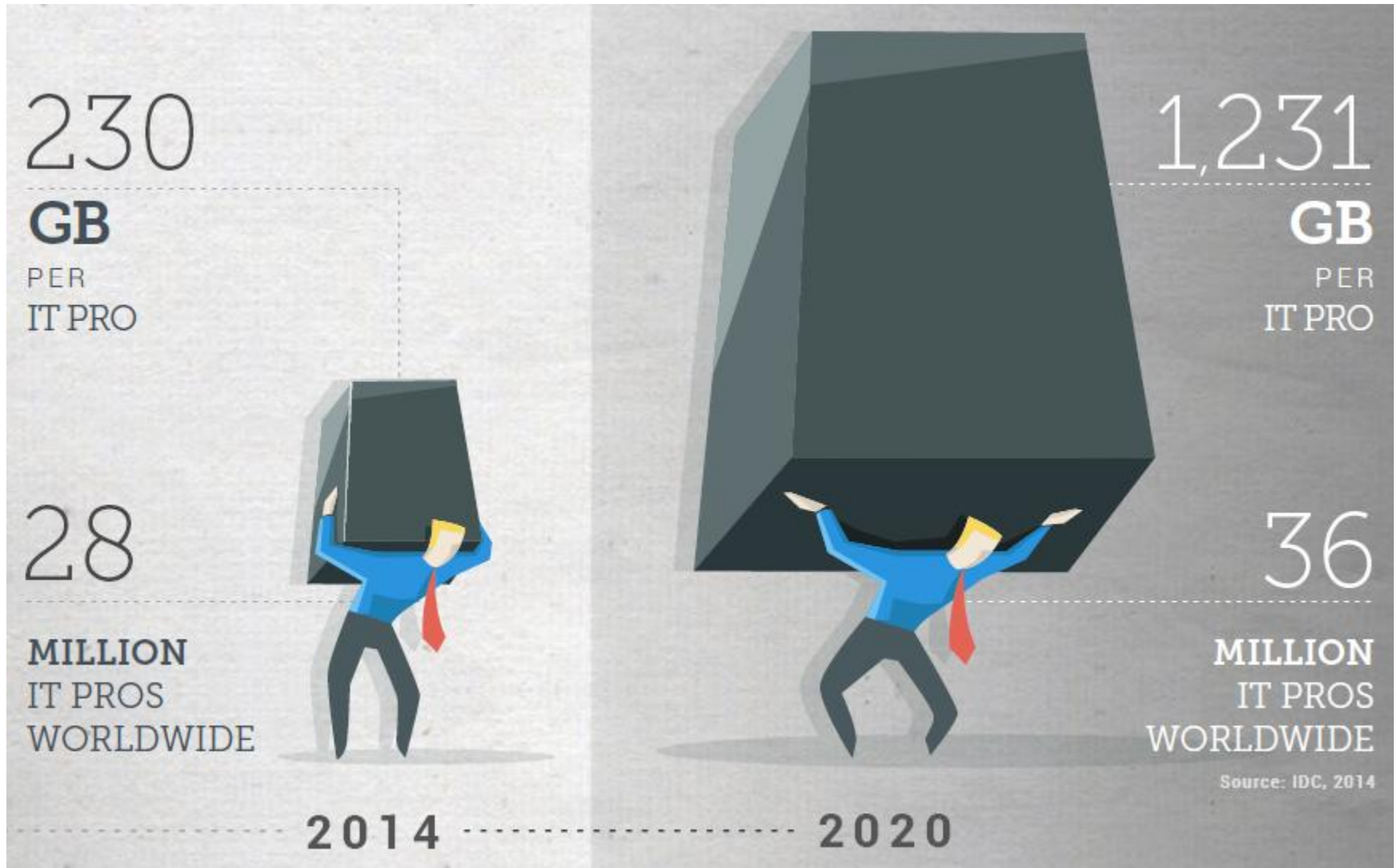
**Gibson TA. (2012) The roots of bioinformatics in ISMB. PLoS Comput Biol. 8(8):e1002679.**



## 2013-2018

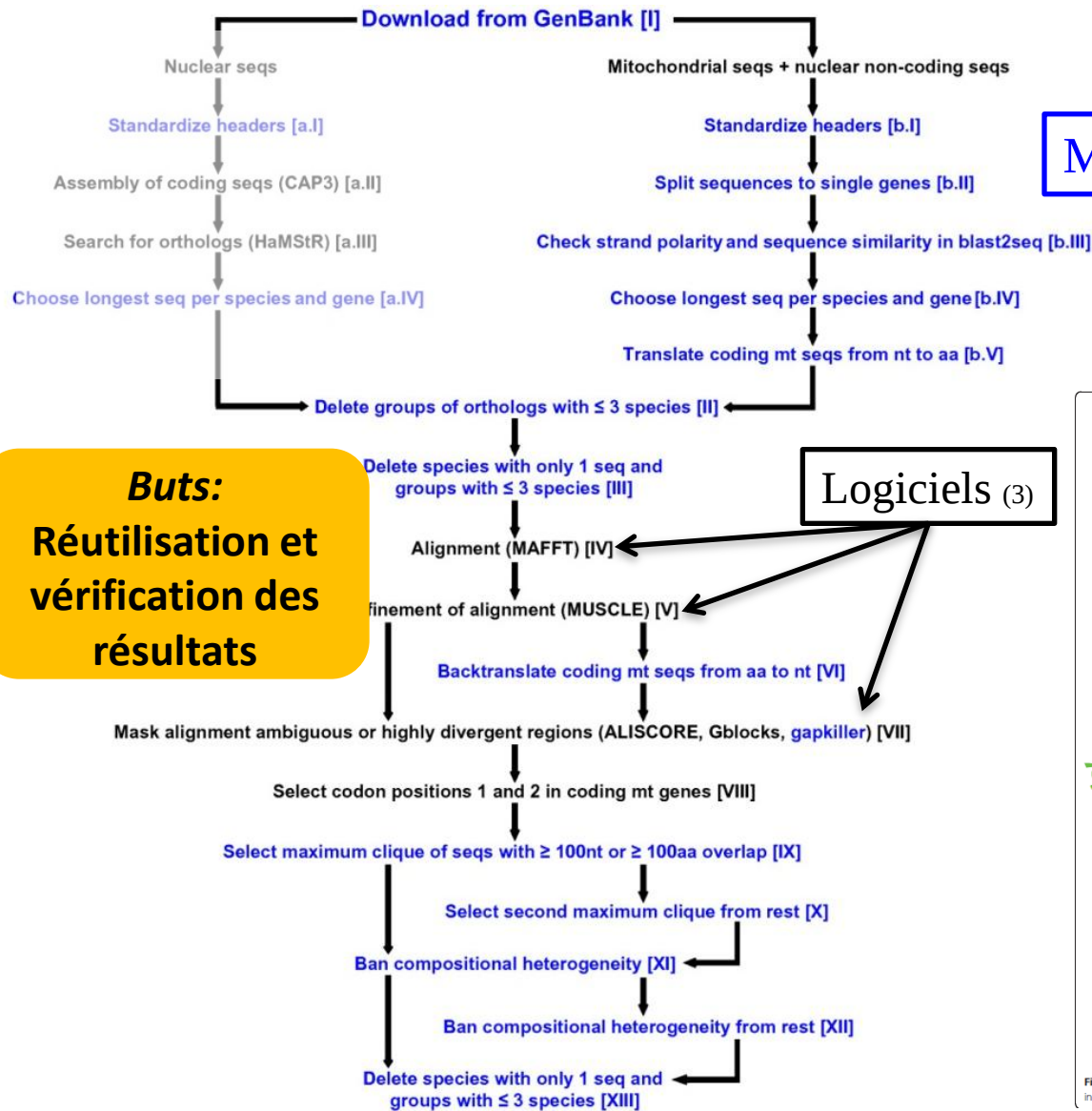


# Comment s'en sortir?



Turner V et al. (2014) The digital universe of opportunities: rich data and the increasing value of the internet of things. International Data Corporation, White Paper, IDC\_1672

# Exemple de flux de travaux en analyse phylogénétique

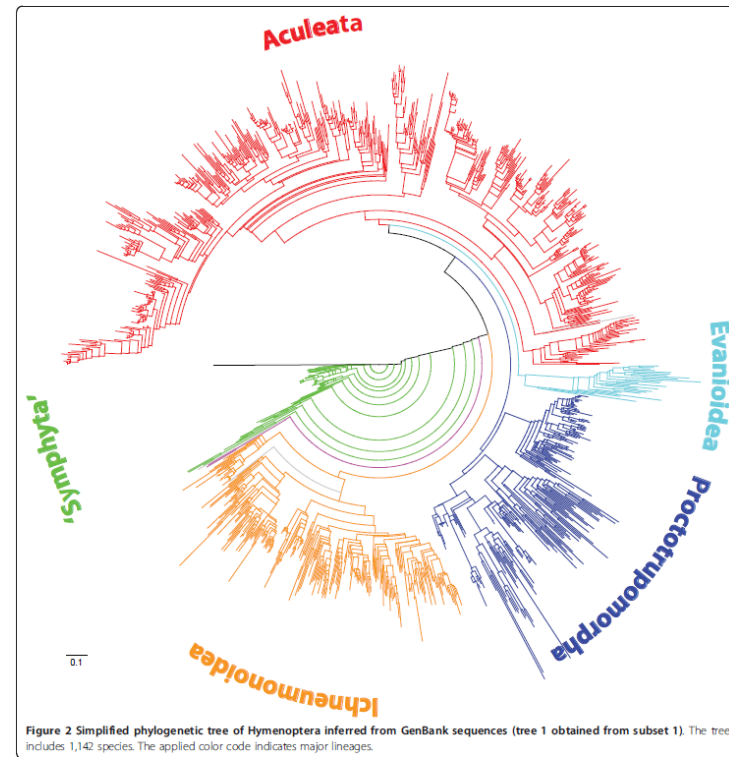


**Buts:**  
Réutilisation et  
vérification des  
résultats

Méthodes (13)

Logiciels (3)

120 000 séquences



Une reconstruction phylogénétique par Peters *et al.* (2011) *BMC Biology* 9, 55.

# Exemple d'application en 2015

## What's in my pot? Real-time species identification on the MinION

BioArxiv doi: <http://dx.doi.org/10.1101/030742>

Sissel Juul, Fernando Izquierdo, Adam Hurst, Xiaoguang Dai, Amber Wright, Eugene Kulesha, Roger Pettett & Daniel J Turner

### Abstract

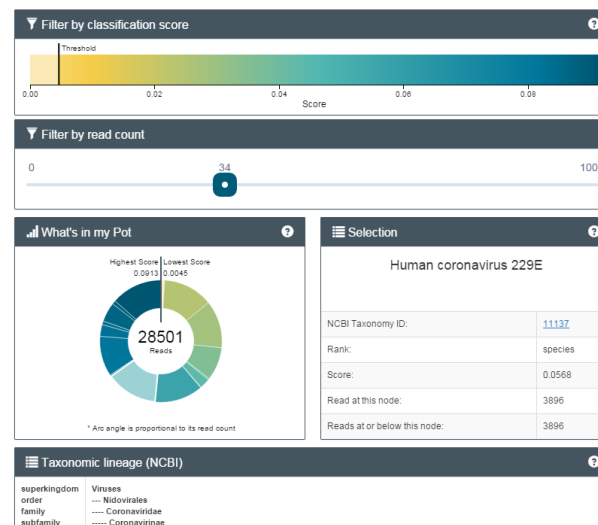
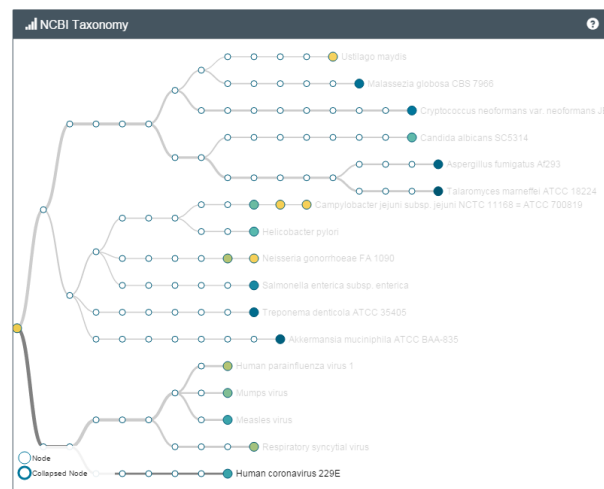
Whole genome sequencing on next-generation instruments provides an unbiased way to identify the organisms present in complex metagenomic samples. However, the time-to-result can be protracted because of fixed-time sequencing runs and cumbersome bioinformatics workflows. This limits the utility of the approach in settings where rapid species identification is crucial, such as in the quality control of food-chain components, or in during an outbreak of an infectious disease. Here we present What's in my Pot? (WIMP), a laboratory and analysis workflow in which, starting with an unprocessed sample, sequence data is generated and bacteria, viruses and fungi present in the sample are classified to subspecies and strain level in a quantitative manner, without prior knowledge of the sample composition, in approximately 3.5 hours. This workflow relies on the combination of Oxford Nanopore Technologies' MinION™ sensing device with a real-time species identification bioinformatics application.



WIMP Bacteria Virus Fungi k24 for SQK-MAP006 [Instance ID: 89822]

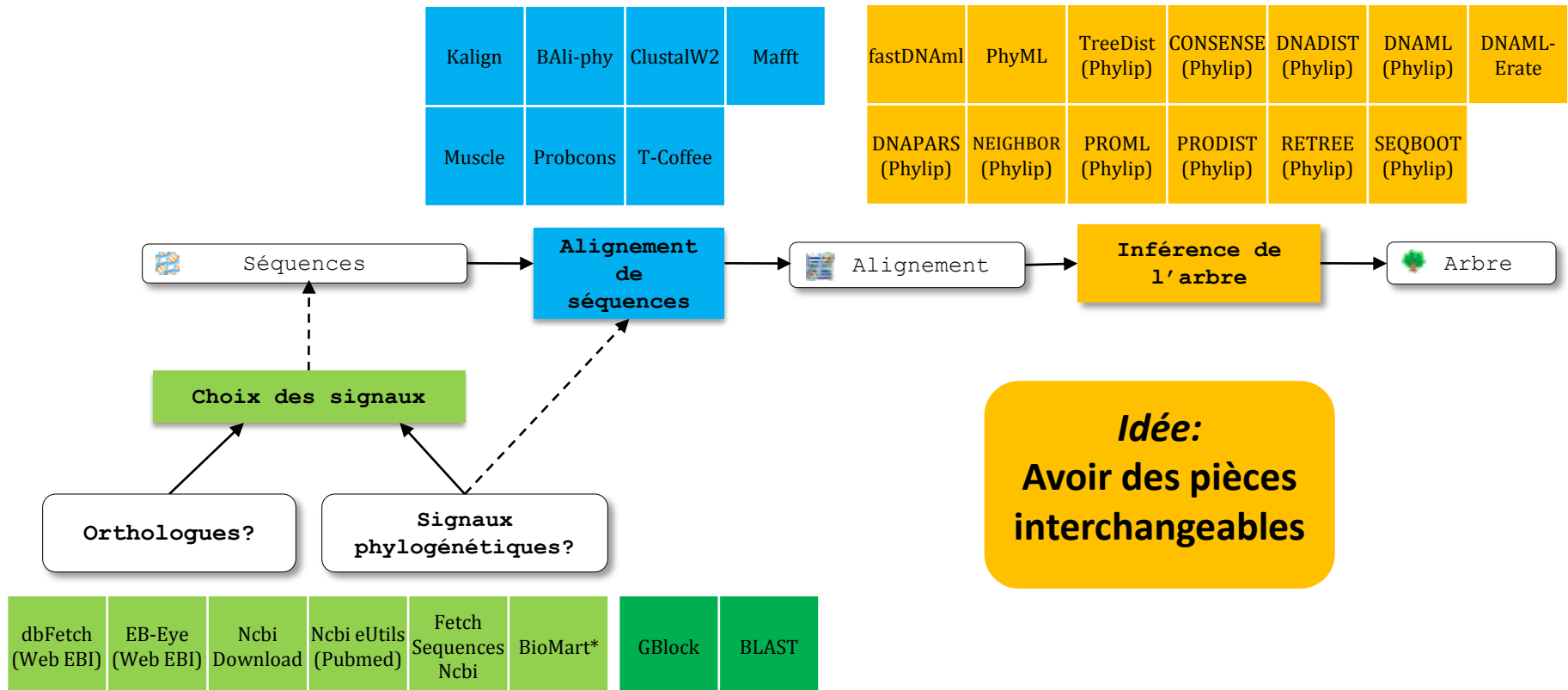
State: stopped  
Duration: 1d 20h 38m 48s  
Started (UTC): 2015-10-26 19:55:34  
Stopped (UTC): 2015-10-28 16:34:22

Application: WIMP Bacteria Virus Fungi k24 for SQK-MAP006  
Data Region: eu-west-1  
Export Metadata: CSV, XLS



# Flux de travaux: définition générale d'un workflow

Un flux de travaux est un **patron de tâches ordonnées** pouvant être **exécutées de manière répétitive**.



Schématisation de l'inférence d'un arbre phylogénétique.

Voir pour revue Philippe *et al.* (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602

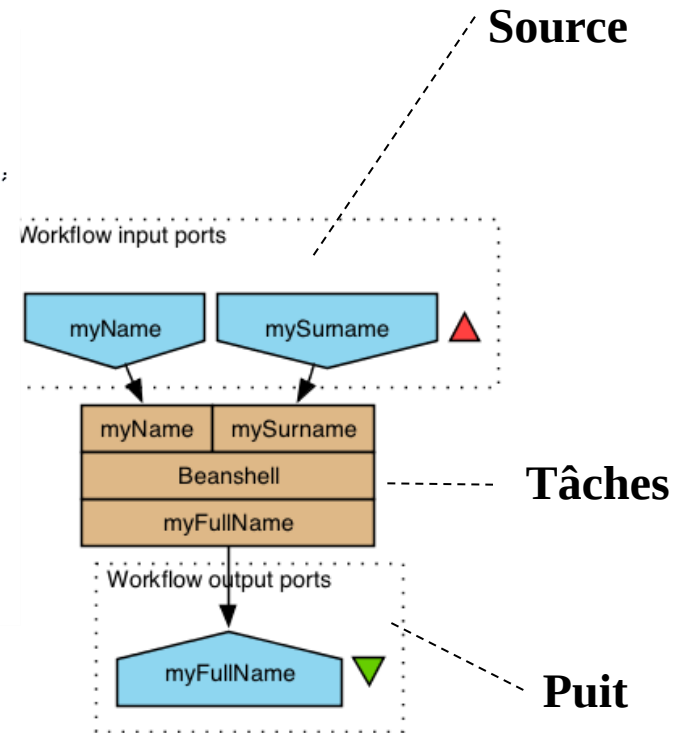
# Composantes des flux de travaux et avantages

```
public MultipleSequences Convert_to_Protein(MultipleSequences multi, boolean remove_stop, int code_number) {
    Config.log("Convert_to_Protein()");
    Translator translator=new Translator();
    GeneticCode code=translator.code.get(code_number);
    MultipleSequences newmulti=new MultipleSequences();
    newmulti.setName("Protein from "+multi.getName());
    newmulti.setNote("Conversion of MultipleSequences to Protein at "+Util.returnCurrentDateAndTime());
    for (biologic.Sequence s:multi.getSequences()) {
        String aa="";
        if (remove_stop) {
            aa=translator.translate_without_stop(s, code);
        } else {
            aa=translator.translate(s, code);
        }
        biologic.Sequence s2=s.clone();
        s2.setId(0);
        s2.setSequence(aa);
        s2.setSequence_type("aa");
        if (s2.getLen()==0) {
            Config.log("Unable to find start codon for "+s.getName());
            return multi;
        }
        Config.log(s2.getName()+"\t"+s2.getLen()+" aa ");
        newmulti.add(s2);
    }
    Config.log("Finish...");
    return newmulti;
}
```

*Code source*



- + **Portabilité (pas de recompilation)**
- + **Automatisation**
- + **Réutilisation des composants**
- + **Scale-up facilité**
- + **Simplification de la cognition**



Taverna Workflow System

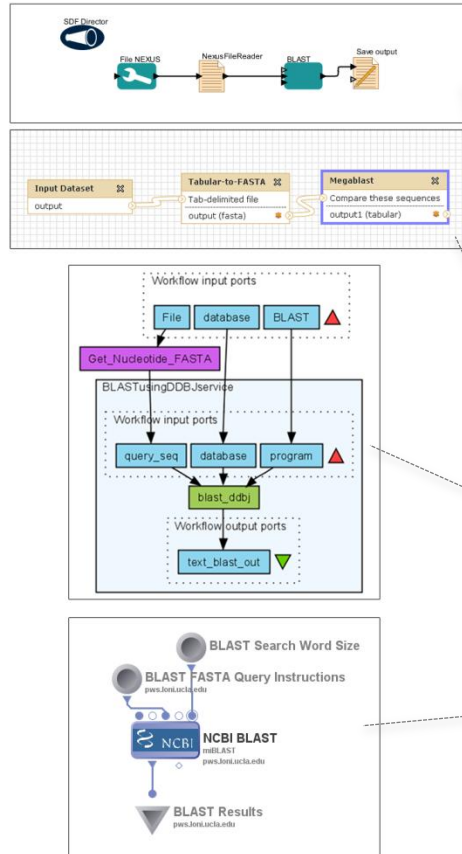
# Caractéristiques des systèmes de gestion de workflows (WfMS)

- Permet *l'interaction avec l'utilisateur* (suggestion avec ontologie).
- Supporte la *reproductibilité* dans les résultats.
- Intègre des services (e.g. services Web) et de *logiciels hétérogènes*.
- Supporte la *distribution et la gestion de données* hétérogènes.
- Support pour les *plate-formes hautes performances*.
- Permet la *surveillance* pendant le fonctionnement et permet la *reprise* même dans le cas d'erreurs.
- Permet une *interconnexion avec d'autres systèmes* (e.g. de flux de travaux, d'ordinateurs en réseau)

Lin, C., Lu, S., Fei, X., Chebotko, A., Pai, D., Lai, Z., Fotouhi, F., et Hua, J. (2009). A reference architecture for scientific workflow management systems and the VIEW SOA solution. *IEEE Transactions on Services Computing*, 2(1), 79-92.

Woollard, D., Medvidovic, N., Gil, Y., et Mattmann, C. A. (2008). Scientific Software as Workflows: From Discovery to Distribution. *IEEE Software*, 25(4), 37-43.

# Les plates-formes de flux de travaux en bioinformatique



Plates-formes de flux de travaux bioinformatiques

| Plate-forme            | Classe                      | Flux de travaux               | Accès au « nuage » | Application     |
|------------------------|-----------------------------|-------------------------------|--------------------|-----------------|
| <b>Kepler (2004)</b>   | <b>Locale</b>               | <i>Data-flow</i>              | Oui (Ecogrid)      | Général         |
| <b>Galaxy (2005)</b>   | <b>Web/Locale</b>           | <i>Data-flow</i>              | Oui (Amazon EC2)   | Bioinformatique |
| <b>Taverna (2003)</b>  | <b>Locale, services Web</b> | <i>Data-flow</i>              | Oui (myGRID)       | Bioinformatique |
| <b>LONI (2003)</b>     | <b>Locale, services Web</b> | <i>Data-flow</i>              | Oui                | Bioinformatique |
| <b>Bio-Jeti (2009)</b> | <b>Locale, services Web</b> | <i>Control-flow</i>           | Oui                | Bioinformatique |
| <b>Triana (1997)</b>   | <b>Client-Serveur</b>       | <i>Data-flow/Control-flow</i> | Oui                | Général         |

**Revue récente:** Ocaña K, de Oliveira D. Parallel computing in genomic research: advances and applications. *Advances and Applications in Bioinformatics and Chemistry* : AABC. 2015;8:23-35.

# Répertoire de workflows (myexperiment.org)



myExperiment makes it easy to **find**, **use** and **share** scientific workflows and other **Research Objects**, and to build **communities**.

myExperiment interface showing navigation options and workflow exploration tools.

**First time visitor? Try these videos:**

- Project Introduction
- Bioinformatics Case Study

**Use myExperiment to...**

- Find **Workflows**
- Share Your **Workflows** and **Files**
- Create and Find **Packs** of Items
- Find **People** and Make Friends
- Create and Join **Groups**
- Build your Profile and Reputation
- Tag and Rate things
- Write Reviews and Comments

**Explore**

Workflow Inputs: queries, program, database

Workflow Outputs: component\_output, final\_output

**Find Workflows**

**About myExperiment**

- Join the Mailing List
- myExperiment Publications
- For Developers
- Give us Feedback
- The BioCatalogue Project

**Register**

**or Login:**

Username or Email:

Password:

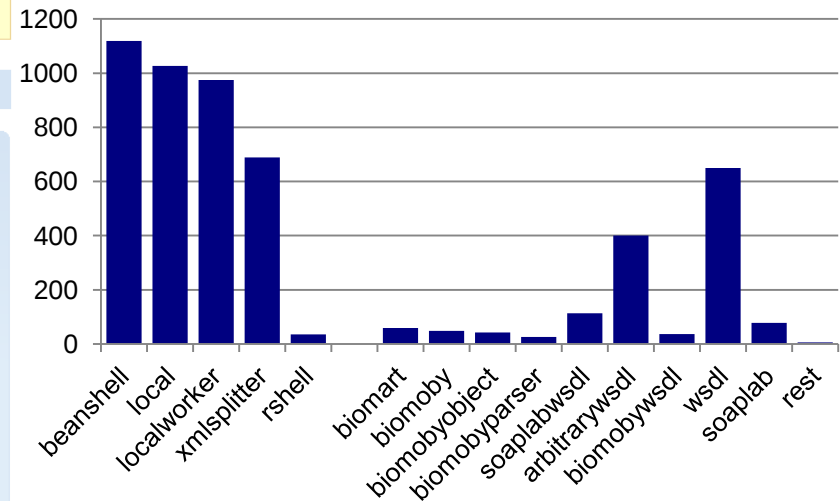
Remember me: ☐

Or use OpenID:

(eg. name.myopenid.com)

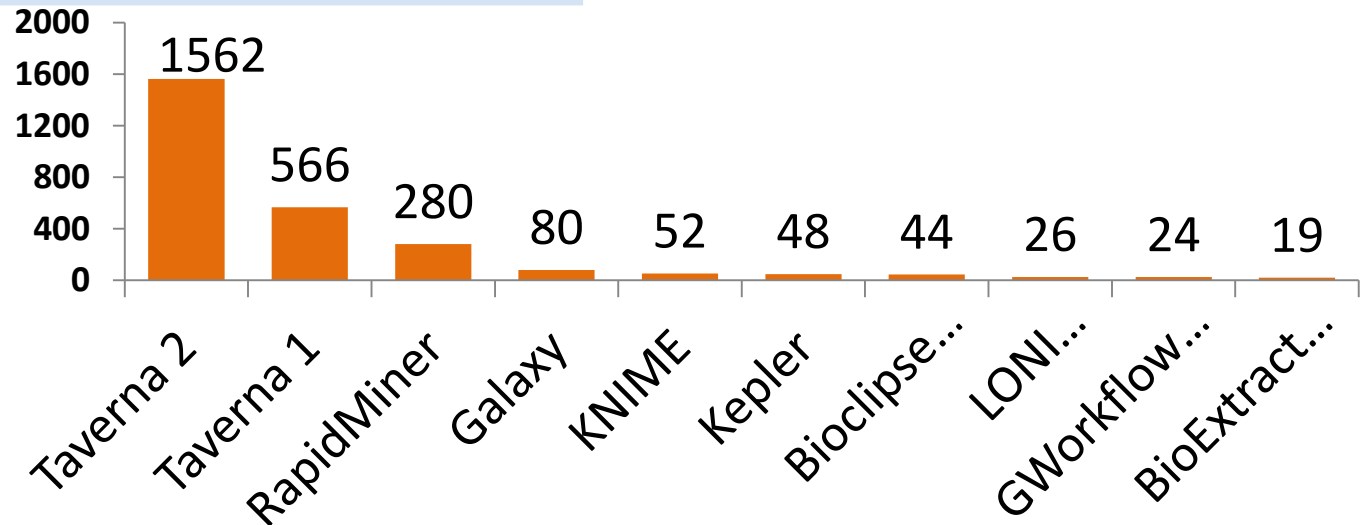
**Login**

[Forgot Password?](#)



myExperiment has over 5000 members, 250 groups, 2000 workflows, 450 files and 150 packs

**En 2018:**  
**3892 workflows**  
**~ 10 600 membres**



Serveur Web ou Serveur Local | Langage **Python et PHP** | **Data-flow**  
Données externes | Bioinformatique | Définition des outils en format XML

The screenshot displays the Galaxy web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main area is the 'Workflow Canvas | imported: metagenomic analysis'. It shows a workflow with three tools: 'Input dataset' (output), 'Tabular-to-FASTA' (input: Tab-delimited file, output: output (fasta)), and 'Megablast' (input: Compare these sequences, output: output1 (tabular)). The left sidebar lists various tool categories like 'Get Data', 'Text Manipulation', 'FASTA manipulation', etc. The right sidebar shows 'Details' for the workflow, including 'Edit Workflow Attributes' with fields for Name, Tags, and Annotation / Notes.

- Les données sont des fichiers
- Base de données **postgresql**
- Préférence pour les formats de types csv et tsv (*table*)

Le système de gestion de flux de travaux Galaxy (<http://galaxy.psu.edu>)  
Giardine, B. et al. (2005). Galaxy: a platform for interactive large-scale genome analysis,  
Genome Res.15:1451-1455.

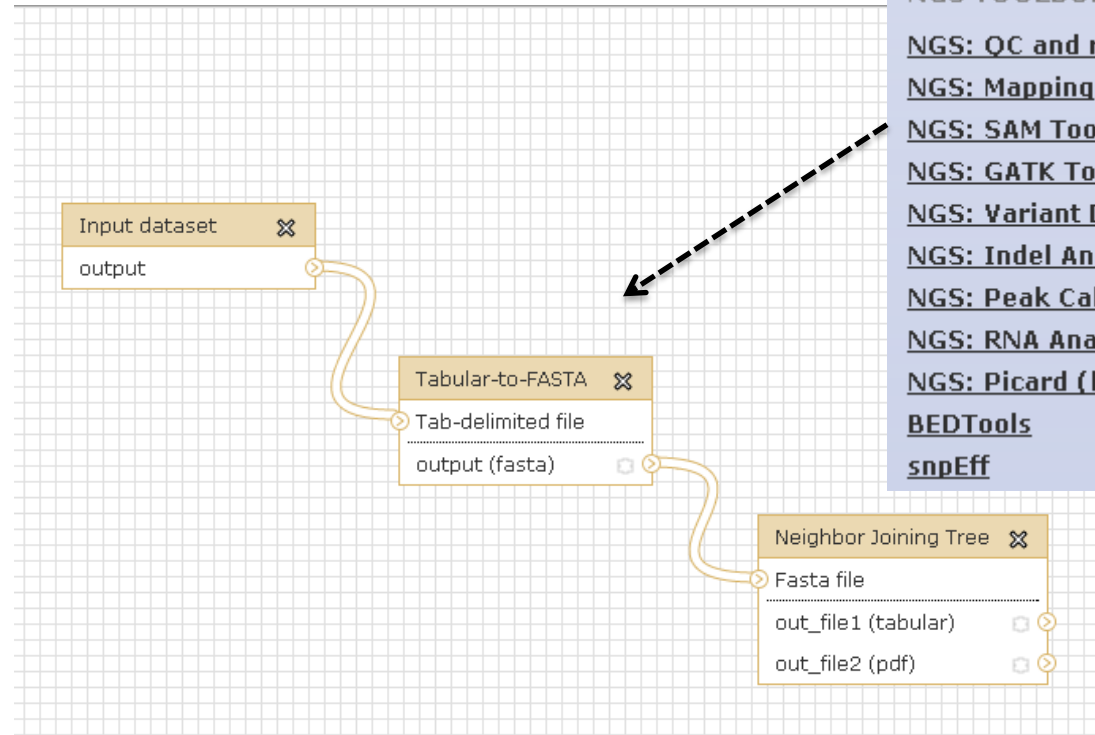
**Galaxy**

Tools

search tools

**Get Data**

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- BX table browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- GenomeSpace import from file browser



## Next generation sequencing

**NGS TOOLBOX BETA**

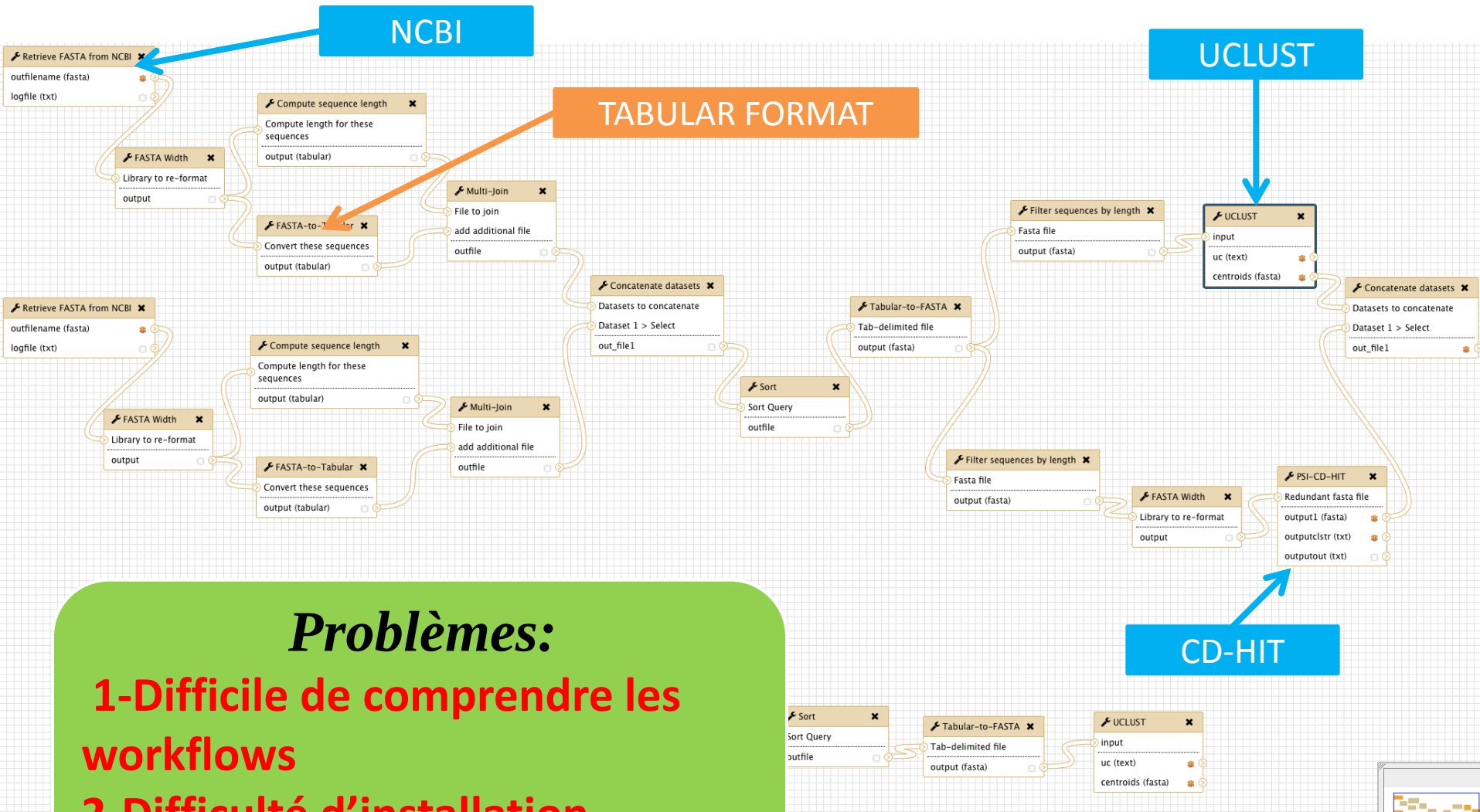
- [NGS: QC and manipulation](#)
- [NGS: Mapping](#)
- [NGS: SAM Tools](#)
- [NGS: GATK Tools \(beta\)](#)
- [NGS: Variant Detection](#)
- [NGS: Indel Analysis](#)
- [NGS: Peak Calling](#)
- [NGS: RNA Analysis](#)
- [NGS: Picard \(beta\)](#)
- [BEDTools](#)
- [snpEff](#)

## Exemples de flux de travaux

[http://main.g2.bx.psu.edu/workflow/list\\_published](http://main.g2.bx.psu.edu/workflow/list_published)

## Serveurs publics

<https://wiki.galaxyproject.org/PublicGalaxyServers>

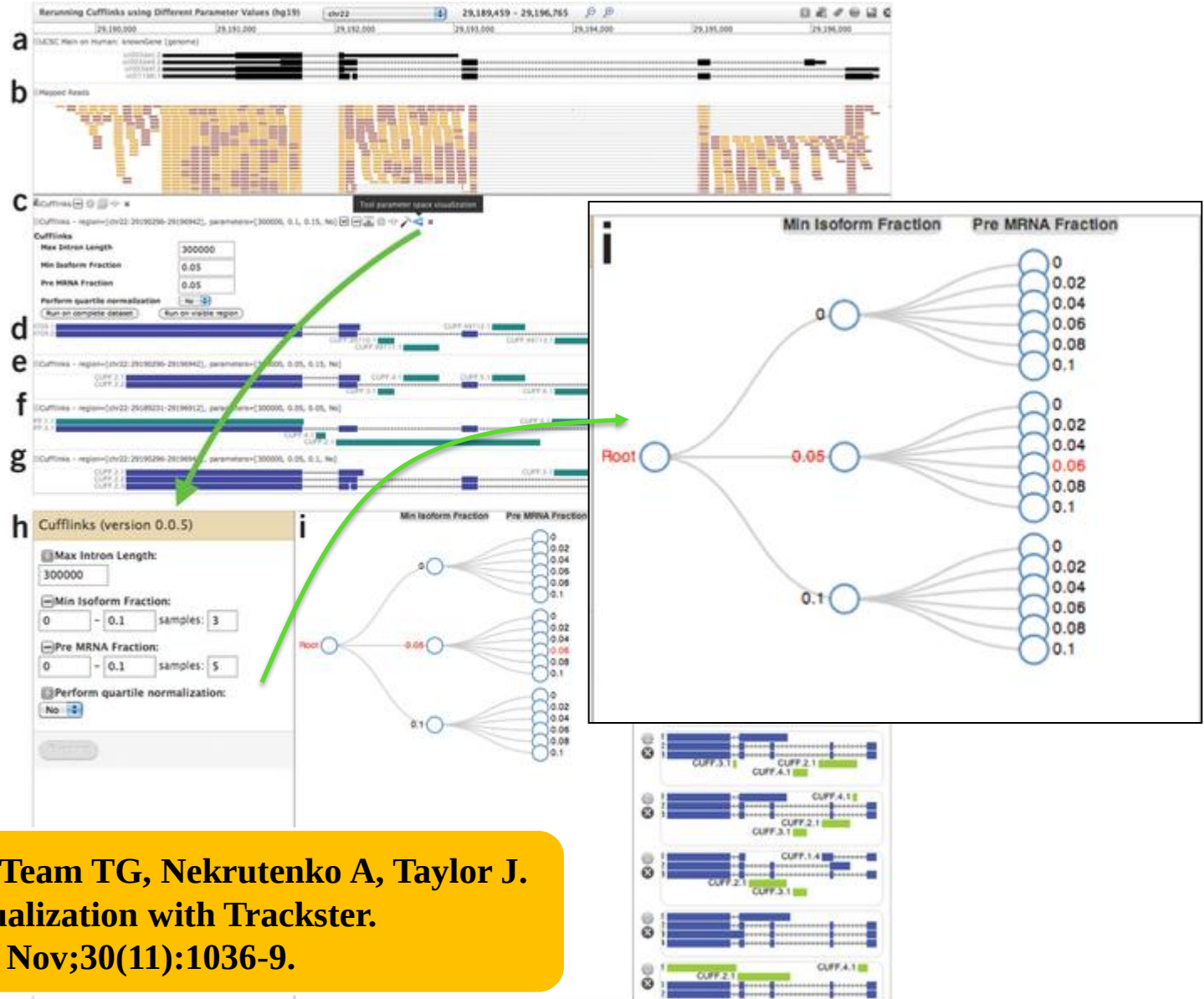


# 5455 publications

<http://www.citeulike.org/group/16008/>

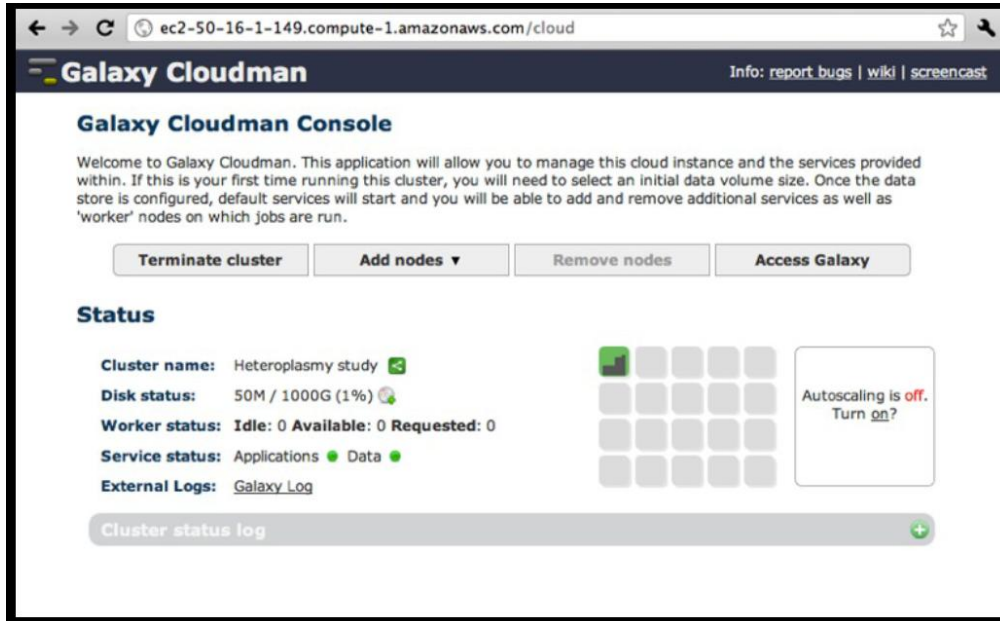
<https://www.zotero.org/groups/1732893/galaxy?>

# Galaxy (workflows imbriqués)



**Goecks J, Coraor N, Team TG, Nekrutenko A, Taylor J.**  
**NGS analyses by visualization with Trackster.**  
**Nat Biotechnol. 2012 Nov;30(11):1036-9.**

# Galaxy (Amazon Cloud) et Tool Shed



The screenshot shows the Galaxy Cloudman console in a web browser. The URL is [ec2-50-16-1-149.compute-1.amazonaws.com/cloud](https://ec2-50-16-1-149.compute-1.amazonaws.com/cloud). The page title is "Galaxy Cloudman". Below the title, there's a welcome message and a grid of buttons: "Terminate cluster", "Add nodes", "Remove nodes", and "Access Galaxy". The "Status" section shows the cluster name "Heteroplasmy study", disk status "50M / 1000G (1%)", worker status "Idle: 0 Available: 0 Requested: 0", and service status "Applications" and "Data". There's also a note about autoscaling being off.

**Galaxy Cloudman Console**

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud instance and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

**Buttons:** Terminate cluster, Add nodes, Remove nodes, Access Galaxy

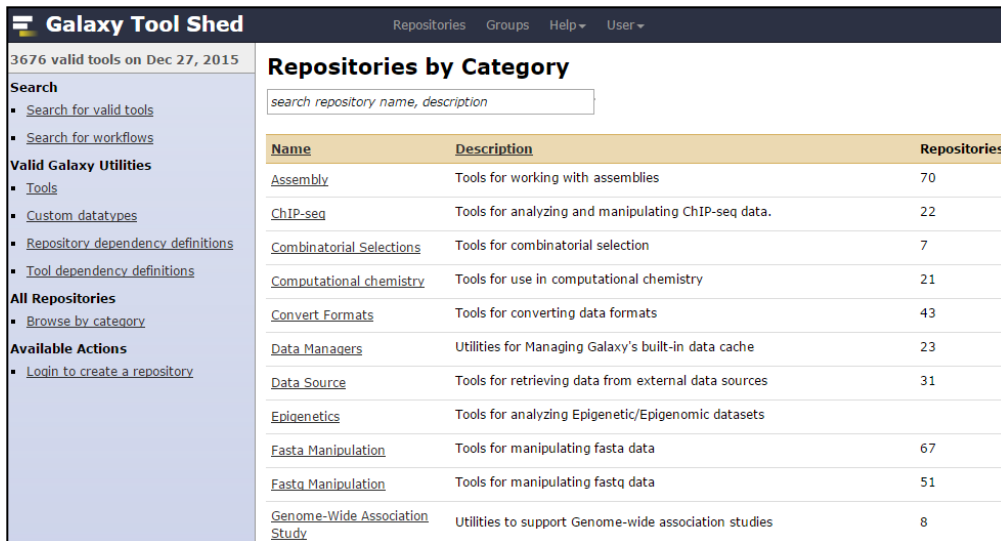
**Status**

Cluster name: Heteroplasmy study  
Disk status: 50M / 1000G (1%)  
Worker status: Idle: 0 Available: 0 Requested: 0  
Service status: Applications Data  
External Logs: [Galaxy Log](#)

Autoscaling is off. Turn on?



Afgan, E *et al.* (2012) Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics. 11 Suppl 12:S4.



The screenshot shows the Galaxy Tool Shed interface. The top bar says "Galaxy Tool Shed" and "3676 valid tools on Dec 27, 2015". There's a search bar and a list of repositories categorized by tool type. The categories include Assembly, ChIP-seq, Combinatorial Selections, Computational chemistry, Convert Formats, Data Managers, Data Source, Epigenetics, Fasta Manipulation, and Genome-Wide Association Study.

**Galaxy Tool Shed**

3676 valid tools on Dec 27, 2015

**Search**

- Search for valid tools
- Search for workflows

**Valid Galaxy Utilities**

- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

**All Repositories**

- Browse by category

**Available Actions**

- Login to create a repository

**Repositories by Category**

| Name                          | Description                                          | Repositories |
|-------------------------------|------------------------------------------------------|--------------|
| Assembly                      | Tools for working with assemblies                    | 70           |
| ChIP-seq                      | Tools for analyzing and manipulating ChIP-seq data.  | 22           |
| Combinatorial Selections      | Tools for combinatorial selection                    | 7            |
| Computational chemistry       | Tools for use in computational chemistry             | 21           |
| Convert Formats               | Tools for converting data formats                    | 43           |
| Data Managers                 | Utilities for Managing Galaxy's built-in data cache  | 23           |
| Data Source                   | Tools for retrieving data from external data sources | 31           |
| Epigenetics                   | Tools for analyzing Epigenetic/Epigenomic datasets   | 67           |
| Fasta Manipulation            | Tools for manipulating fasta data                    | 51           |
| Genome-Wide Association Study | Utilities to support Genome-wide association studies | 8            |



<https://toolshed.g2.bx.psu.edu/>

46 outils seulement en phylogénétique!

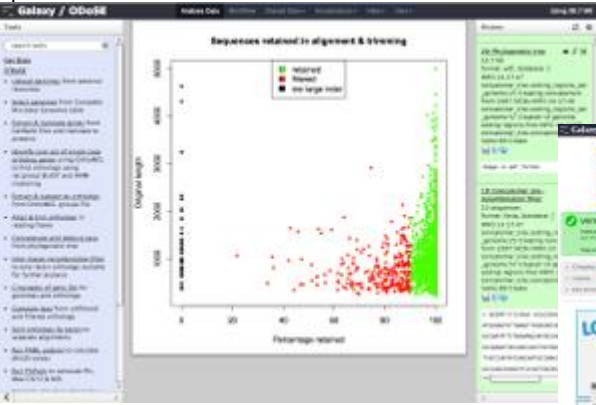

**Galaxy Wiki**

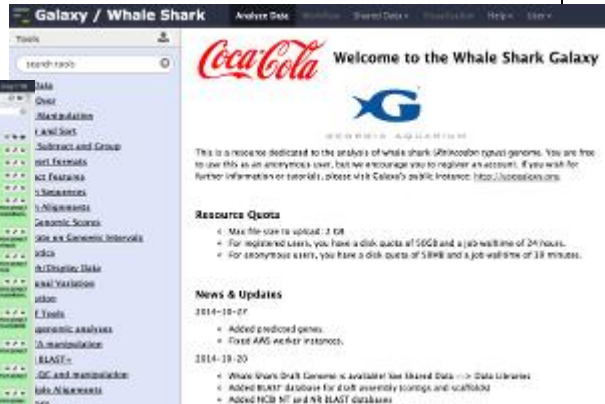
Connexion | Rechercher :

PublicGalaxyServers




Public Galaxy Servers  
and *still* counting



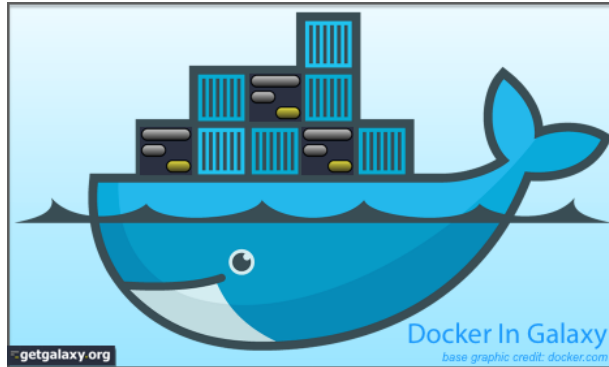


<https://wiki.galaxyproject.org/PublicGalaxyServers>

# Machines Virtuelles pour Galaxy



VirtualBox



## Galaxy Virtual Appliance Directory

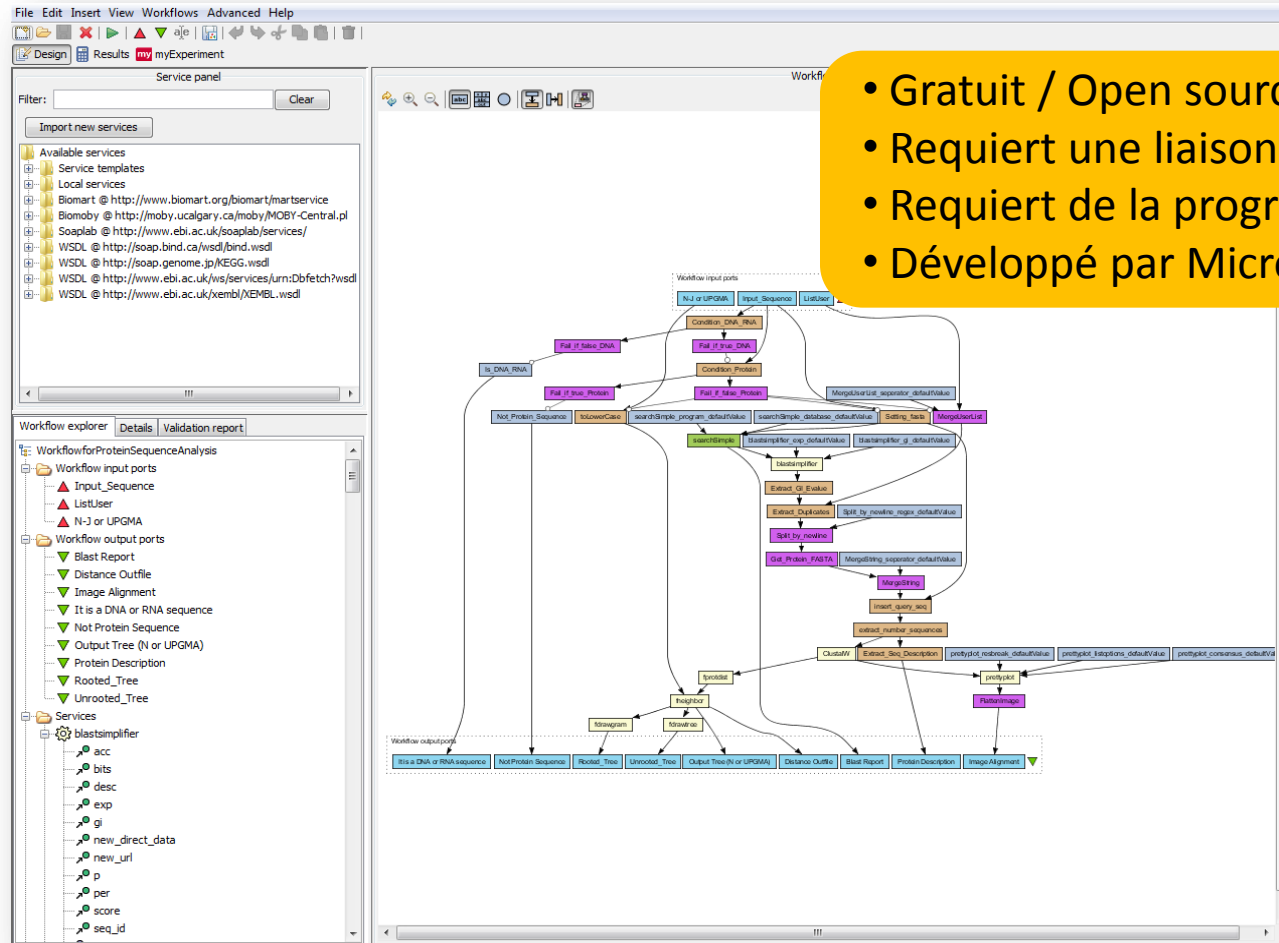
This lists all known available Galaxy virtual appliances. The goal is to make it easy for the Galaxy community to find and deploy them, thus benefiting from the (generous!) work of others.

| Appliance                                                    | Technology | Domains                                                                                      | Description                                                                                                                                                                               | Owners                                                                                                                                  | Date Created/Updated |
|--------------------------------------------------------------|------------|----------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| <a href="#">DDBJ Read Annotation Pipeline</a>                | VM         | Reference genome mapping, <i>de novo</i> assembly, and structural and functional annotations | VirtualBox VM created in 2015 so that "Researchers may utilize the virtual machine image for users' sensitive datasets such as human personal genomic sequences."                         | DDBJ Read Annotation Pipeline, Development team                                                                                         | 2015                 |
| <a href="#">RNAmapper</a>                                    | VM, AMI    | Identify genome region linked to a mutation, and causal candidate mutations                  | <a href="#">RNAmapper</a> uses RNA-seq data to identify both a region of the genome linked to a mutation as well as candidate mutations that may be causal for the phenotype of interest. | Moen's Lab, Fred Hutchinson Cancer Research Center; Megason Lab, Harvard Medical School                                                 | 2012/08/03           |
| <a href="#">MegaMapper</a>                                   | VM, AMI    | positional cloning of mutations                                                              | <a href="#">MegaMapper</a> is a computational pipeline for positional cloning of mutations by whole genome sequencing.                                                                    | Megason Lab, Harvard Medical School                                                                                                     | 2013/05/02           |
| <a href="#">Automated Selection of Hotspots (ASH)</a>        | VM         | hotspot detection                                                                            | automated detection quantitative ranking of hotspots to support histopathologists in selecting the 'hottest' hotspot areas in adrenocortical carcinoma.                                   | David van Zessen <d.vanzessen@erasmusmc.nl>, Erasmus Medical Center                                                                     | 2014/05/16           |
| <a href="#">UC Davis Bioinformatics Core Training Galaxy</a> | AMI        | Training                                                                                     | Includes all needed tools and datasets for a weeklong workshop.                                                                                                                           | UC Davis Bioinformatics Core                                                                                                            | 2014/06/20           |
| <a href="#">Immunoglobulin Galaxy</a>                        | VM         | immunoglobulin                                                                               | a set of tools for easy detection and quantitation of immunoglobulin heavy chain alternative transcripts.                                                                                 | David van Zessen <d.vanzessen@erasmusmc.nl>, Erasmus Medical Center                                                                     | 2014/08/19           |
| <a href="#">ballaxy</a>                                      | Docker     | computer aided drug design, molecular modelling                                              | Contains the BALL (Biochemical Algorithms Library) Project tools, i.e. computer aided drug design and molecular modelling based on protein and ligand structure data.                     | Hans-Peter Lenhof Lab, Saarland University; Oliver Kohlbacher Lab, University of Tübingen; Andreas Hildebrandt Lab, University of Mainz | 2014/09/02           |
| <a href="#">Pitagora-Galaxy</a>                              | VM, AMI    | General Purpose                                                                              | Includes the tools and workflows that are described at <a href="#">our wiki</a> .                                                                                                         | Genome Science Division, RCAST, University of Tokyo                                                                                     | 2014/11/20           |
| <a href="#">SADI-Galaxy Container</a>                        | Docker     | Web services, RDF                                                                            | Make SADI services available in a Galaxy platform.                                                                                                                                        | Mikel Egaña Aranguren                                                                                                                   | 2015/01/07           |
| <a href="#">RNA Galaxy Workbench</a>                         | Docker     | RNA-Seq                                                                                      | A Docker based RNA-Seq workbench                                                                                                                                                          | RNA Bioinformatics Center, part of the German Network for Bioinformatics Infrastructure (de.NBI).                                       | 2015/02/17           |

<https://wiki.galaxyproject.org/VirtualAppliances>, <https://hub.docker.com/r/anhi/ballaxy/>

# Taverna (v2.5 et 3.0)

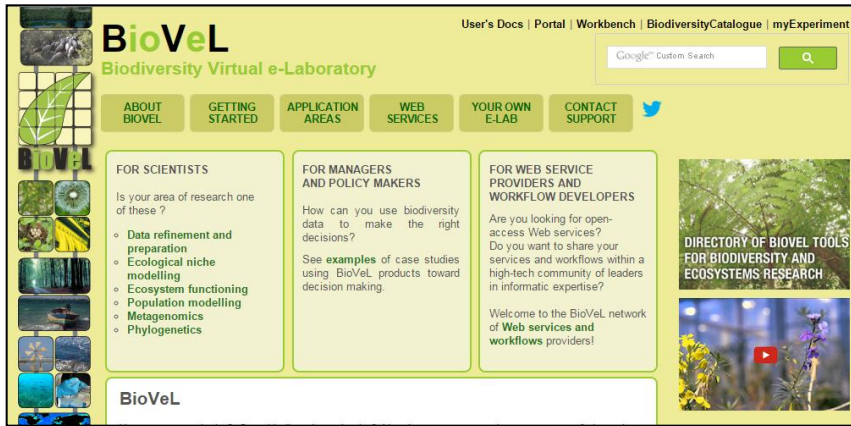
Serveur Web ou Serveur Local | Langage **Java** | Données externes  
**Data-flow** et **control-flow** avec programmation | Bioinformatique



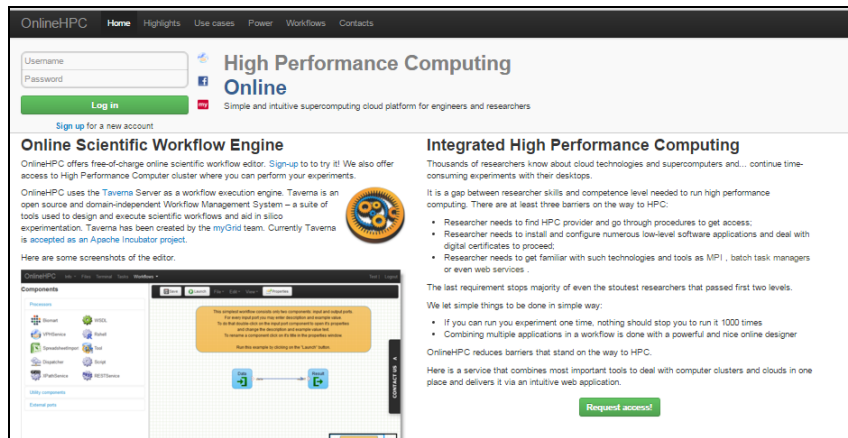
- Gratuit / Open source
- Requiert une liaison Internet
- Requiert de la programmation (java)
- Développé par Microsoft, Apache

<https://taverna.incubator.apache.org/documentation/taverna-galaxy/>  
Stevens, et al. (2003). myGrid: personalised bioinformatics on the information grid.  
Bioinformatics, 19:i302-i304.

# Répertoire de Services Webs / Intégration



<https://www.biodiversitycatalogue.org/>  
<https://www.biovel.eu/>



<http://onlinehpc.com/site/main>

## Integrated with other myGrid tools

- **my experiment** – workflow sharing environment for scientists
- **BioCatalogue** – curated catalogue of Web services

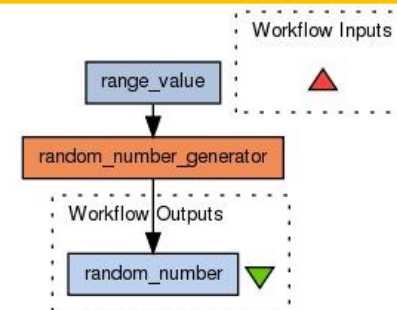
## Works with ...

- **bio::smart** – data integration system for large scale data querying
- **SoapLab** – Web services wrapping existing command-line analysis programs (such as EMBOSS sequence analysis)
- **SADI** – semantic Web services described using SADI
- **R** – software for statistical computing and graphics
- **Bioconductor** – tools for the analysis and comprehension of high-throughput genomic data
- A number of third party plugins

Intégration des langages de programmation:  
Python, R, Perl, Java.

Exemple de python:

[www.myexperiment.org/workflows/2475](http://www.myexperiment.org/workflows/2475)



# Biodiversity Catalogue

**BiodiversityCatalogue**  
"The Biodiversity Sciences Web Services Registry"

Getting Started | About Us | Contact Us | API Docs

Sign up | Sign in

Search:  **Go!** | [Home](#) | [Services](#) | [Register a Service](#) | [Service Providers](#) | [Search by Data](#) | [Latest](#)

[Home](#) + [Services](#)

**Services** | **Register a Service** | **Service Providers** | **Search by Data**

synonym | accepted name | class | classification | ecology | family | genus | kingdom | order | phylum

**Filtering**

Current Filters Applied  
none

Select filters from below...

**Enable tag filters**

**Service Types (2)**

[REST](#) (63)  
[SOAP](#) (10)

**Service Categories (59)**

- ☒ [Taxonomy](#) (34)
- ☒ [Analysis](#) (9)
- ☒ [Modelling](#) (10)
- ☐ [Geospatial](#) (9)
- ☒ [Data](#) (26)
- ☒ [Visualization](#) (5)
- ☐ [Publishing](#) (5)
- ☐ [Shim](#) (2)
- ☐ [Infrastructure](#) (5)
- ☒ [Context](#) (25)
- ☒ [Created By](#) (0)

Search within results:  **Go!**

Displaying services **1 - 21** of **71** in total

« Previous **1** **2** **3** **4** Next »

**Include archived?** ☐  
**Sort by:** Newest  
**Per page:** 21  
**View:** Grid

**Lifewatch-WB ... ope database** REST

Categories: none [Help categorise this...](#)

The ecotopes are delineated based on 2015 orthophotos and 2012/2013 LIDAR dataset. Each polygon is then characterised...

Provider: [maps-elle-ucl-ac-be](#)

**Prokaryotic ... e Up-to-Date** REST

Categories: [GFBio](#) [Taxonomy](#)

[Taxonomic Name Resolution](#)  
[Taxonomic Synonym Resolution](#)  
[Checklist and Classification](#)

PNU is a compilation of all names of Bacteria and Archaea which have been validly published according to the Bacterio...

Provider: [Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures](#)

**DTN Taxon Lists Services** REST

Categories: [Taxonomy](#) [Taxonomic Name Resolution](#)  
[Taxonomic Synonym Resolution](#)  
[Checklist and Classification](#)

The DWB REST Webservice for Taxon Lists is part of a Diversity Workbench (DWB) services network. It is delivering bas...

Provider: [SNSB](#)

**BioCASE Moni ... Webservices** REST

Categories: [Taxonomic Diversity](#) [Data Retrieval](#) [GFBio](#)

The BioCASE Monitor Service BMS 2.0 is a tool for coordinators of networks of biodiversity databases

**GFBio Terminology Server** REST

Categories: [GFBio](#) [Data Retrieval](#)  
[Data Refinement](#) [Knowledge Organisation](#)

The GFBio Terminology Server is a core component of the GFBio infrastructure and allows users to share various kinds...

**BioNet Web Services** REST

Categories: none [Help categorise this...](#)

BioNet Web Services will progressively make all of the data held in BioNet open and accessible via a standardised API...

## Biodiversity Virtual e-Laboratory

**Phylogenetics** Bayesian phylogenetic inference workflows are for performing phylogenetic inference for systematics and diversity research. Bayesian methods guide selection of the evolutionary model and a post hoc validation of the inference is also made. Phylogenetic partitioning of the diversity across samples allows study of mutual information between phylogeny and environmental variables

purl: <http://www.purl.ox.ac.uk/researchobj/myexp-370>

Portal: <https://www.portal.biovel.eu/workflows/466>

<https://www.portal.biovel.eu/workflows/549>

<https://www.portal.biovel.eu/workflows/550>

<https://www.portal.biovel.eu/workflows/525>

PDP workflow, using PhyloH for partitioning environmental sequencing data using both categorical and phylogenetic information

purl: <http://www.purl.ox.ac.uk/workflow/myexp-3570.5>

Portal: <https://www.portal.biovel.eu/workflows/434>

<https://www.portal.biovel.eu/workflows/71>

MSA-PAD workflow performs a multiple DNA sequence alignment coding for multiple/single protein domains invoking two alignment modes: gene and genome

Gene mode purl: <http://www.purl.ox.ac.uk/workflow/myexp-4549.1>

Portal: <https://www.portal.biovel.eu/workflows/712> (access on request)

Genome mode purl: <http://www.purl.ox.ac.uk/workflow/myexp-4551.1>

Portal: <https://www.portal.biovel.eu/workflows/713> (access on request)

SUPERSMART (self-updating platform for estimating rates of speciation and migration, ages and relationships of taxa) is a pipeline analytical environment for large-scale phylogenetic data mining, taxonomic name resolution, tree inference and fossil-based tree calibration

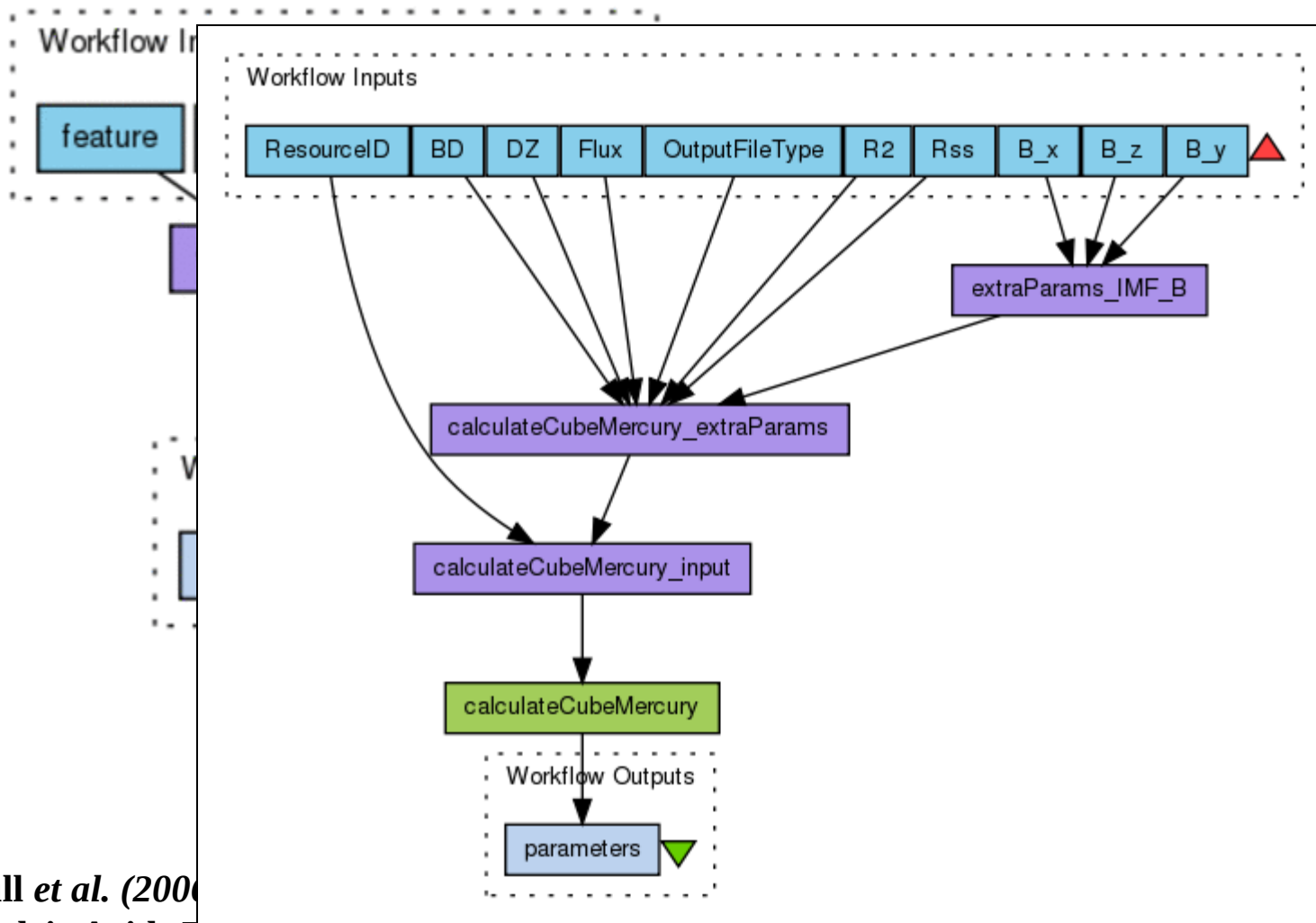
url: <https://www.biodiversitycatalogue.org/services/78>

Scientific studies: [44–47]

web service

Hardisty et al. (2016) BioVeL: a virtual laboratory for data analysis and modelling in biodiversity science and ecology. BMC Ecol. 2016 Oct 20;16(1):49. PubMed PMID: 27765035

# Approche de Taverna



ples  
non une  
nom des

Hull *et al.* (2006) *Nucleic Acids Research*, vol. 34, 729-732.

Oinn *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*. 20(17):3045-54.



The screenshot shows the GitHub repository for Taverna, specifically the 'taverna / taverna-prov' branch. The repository has 37 watches, 1 star, and 1 fork. The main content area displays the repository name and a link to the W3 Provenance WG ontology export for Taverna. A yellow callout box is overlaid on the left side of the repository view, containing text about adding information on data provenance and ontology. Below the repository name, there are buttons for 'Create new file', 'Upload files', 'Find file', and 'Clone or download'. The latest commit is 485cde9 on 21 Oct 2014, 3 years ago. The repository is licensed under LGPL-3.0 and has 1 contributor.

taverna / taverna-prov

Watch 37 Star 1 Fork 1

Code Issues 2 Pull requests 0 Projects 0 Wiki Pulse Graphs

W3 Provenance WG ontology export for Taverna <http://www.w3.org/2011/prov/wiki/Main...>

1 contributor LGPL-3.0

Create new file Upload files Find file Clone or download

Latest commit 485cde9 on 21 Oct 2014  
github.com/taverna/taverna-prov  
3 years ago

example simplified sparql

## Command Line Tool 2.5

Taverna Command Line Tool 2.5 is basically the Taverna Workbench stripped of its GUI so you can run workflows from a command prompt headlessly (i.e. without the graphical environment).

It is available in several editions. The [Core edition](#) contains services for running [scientific workflows](#) in any domain, accessing general services such as REST or SOAP Web services and external command line tools.

Domain-specific editions (see below) add support for specific service types.

[Enterprise Taverna](#) is an 'all inclusive' edition which contains all available service types, equivalent to the Core edition with all official plugins included.

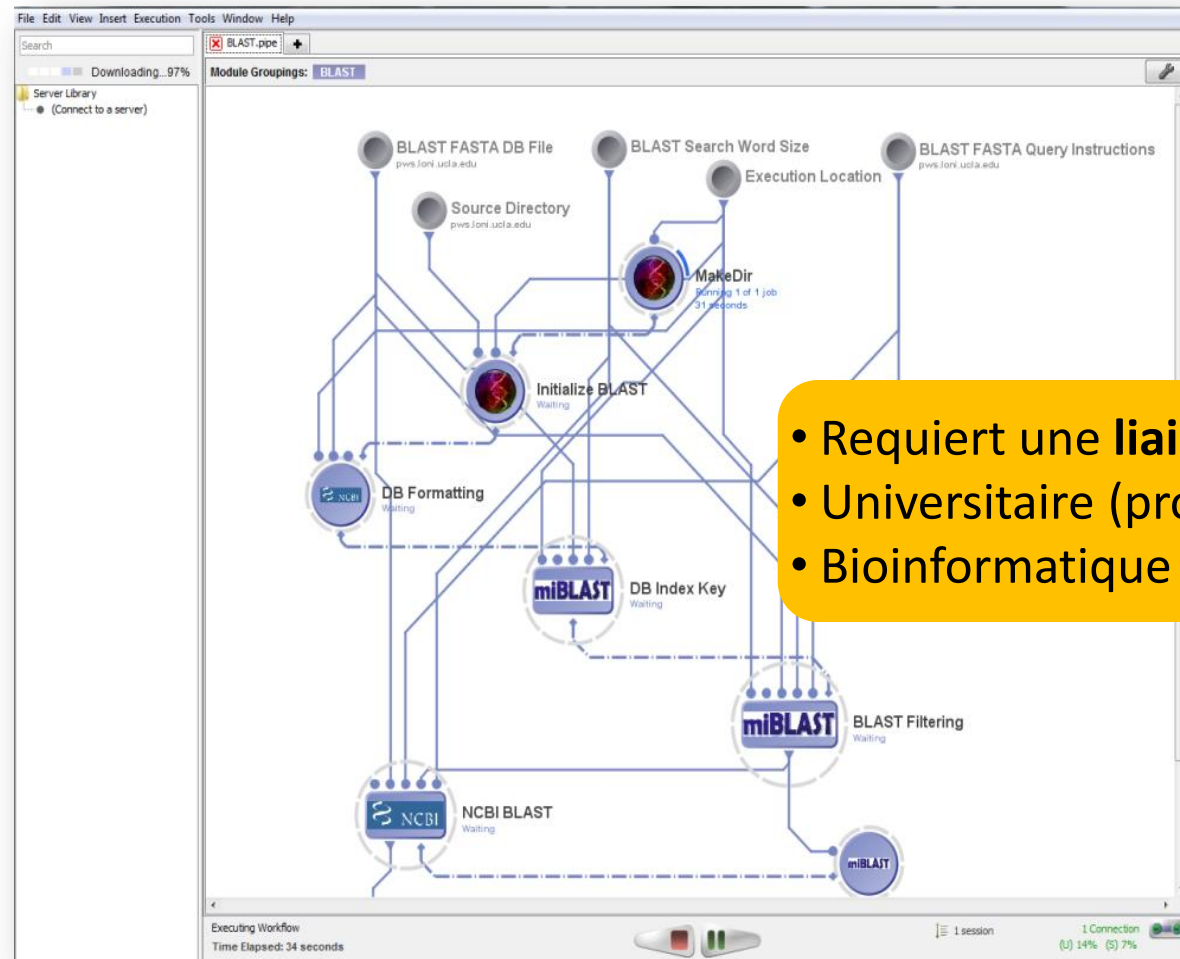
### Available editions

Check [system requirements](#) before you download Taverna.

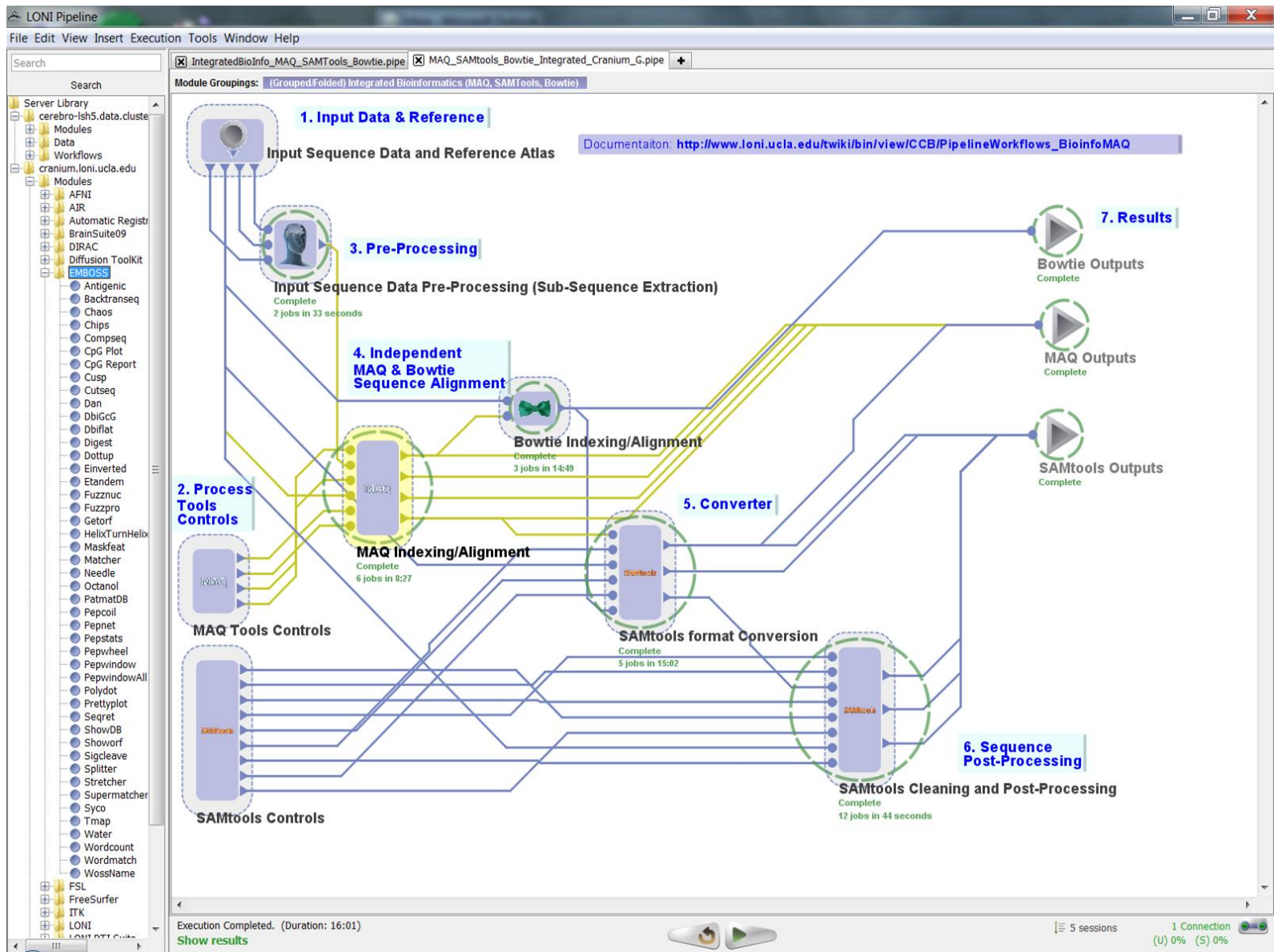
- Taverna Command Line Tool Core 2.5
- Taverna Command Line Tool Astronomy 2.5
- Taverna Command Line Tool Bioinformatics 2.5
- Taverna Command Line Tool Biodiversity 2.5
- Taverna Command Line Tool Digital Preservation 2.5
- Taverna Command Line Tool Enterprise 2.5

Quelques outils bioinformatiques disponibles

## Service webs et Local | Langage **Java** | Données externes Data-flow **sans** programmation | Bioinformatique



[pipeline.loni.ucla.edu](http://pipeline.loni.ucla.edu)



[http://www.ccb.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows\\_BioinfoMAQ](http://www.ccb.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoMAQ)

The screenshot displays the Bio-Jeti software interface. The title bar reads "Pfam-A family list - C:\Documents and Settings\more26107809\mes documents\Bio-Jeti\pfam-sibs-demo\project\models\Pfam\_A\_family\_li...". The menu bar includes File, Edit, Project, SIB, Edge, Graph, View, Mode, Extras, Plugins, and Help. The toolbar contains various icons for file operations and workflow management. The left sidebar shows a project tree under "JABC Projects" with a "Demo Project" containing "models", "mini", "web-shop", and "Welcome.xml". The main workspace is titled "Pfam-A annotations" and "Pfam-A family list". It contains a text description of the Pfam-A family list service and a workflow diagram. The workflow diagram shows a sequence of steps: "PfamAFamilyList (text)" (red star icon) connected by a "default" arrow to "show result" (blue document icon), which is connected by an "ok" arrow to "PfamAFamilyList (XML)" (red star icon), which is then connected by a "default" arrow to another "show result" (blue document icon). The status bar at the bottom shows "SN 5293868311", "0 Cells selected", "41 MB - 85%", and "Developer" mode.

**Projects** | **SIBs**

JABC Projects

- Demo Project [E:\JABC\projects\Demo]
  - models
    - Countdown.xml
    - TextFile.xml
  - web-shop
    - Welcome.xml
  - <classpath>

**APs** | **Reference** | **SIB Icon** | **Module Viewer**

**SIB** | **Graph** | **Draw** | **LocalChecker** | **GEAR [Basic]**

**Service Descriptions**

**Pfam-A annotations** | **Pfam-A family list**

The Pfam-A family list service can be used to retrieve the list of all Pfam-A families in the latest Pfam release. This example demonstrates the use of the service.

```

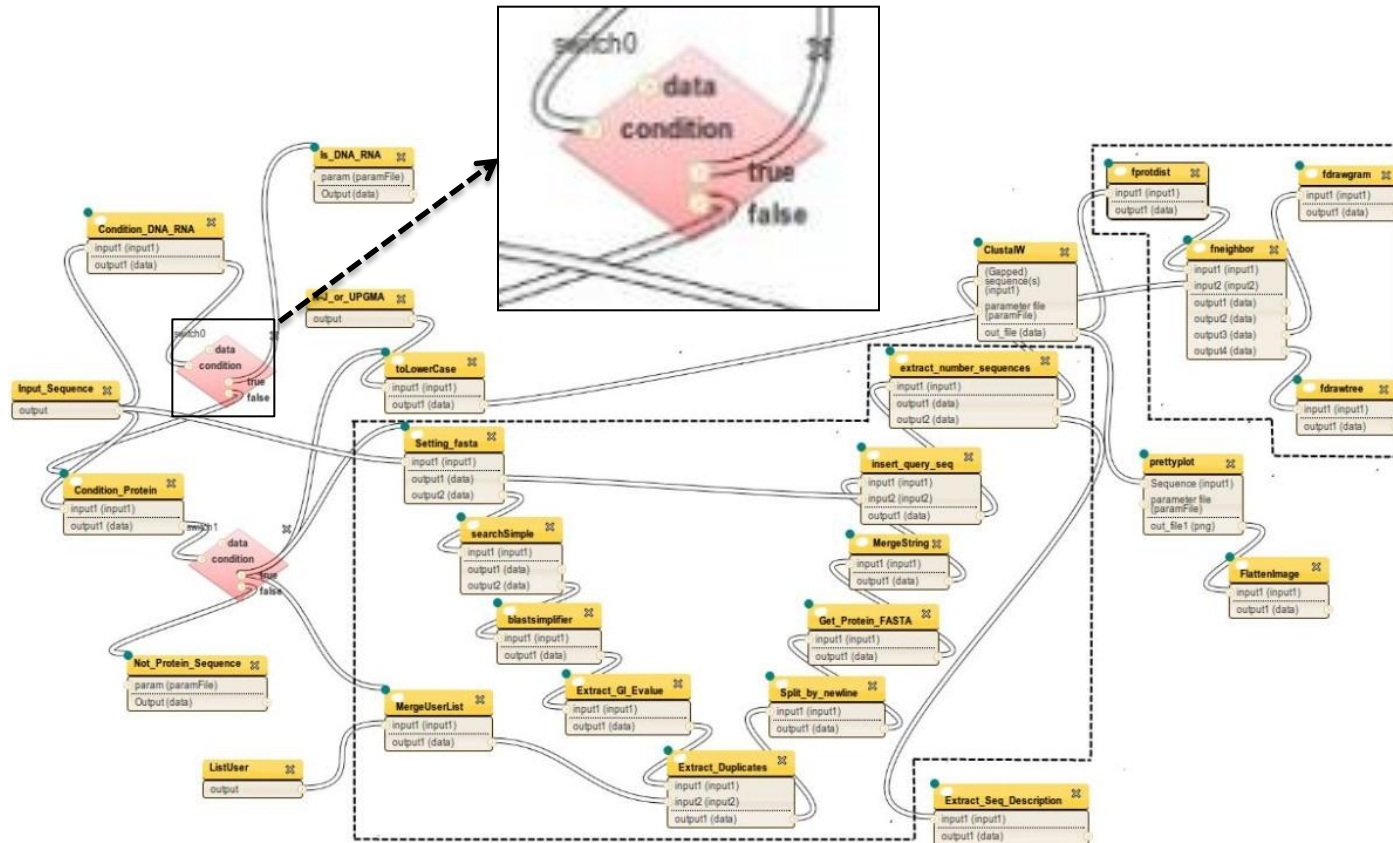
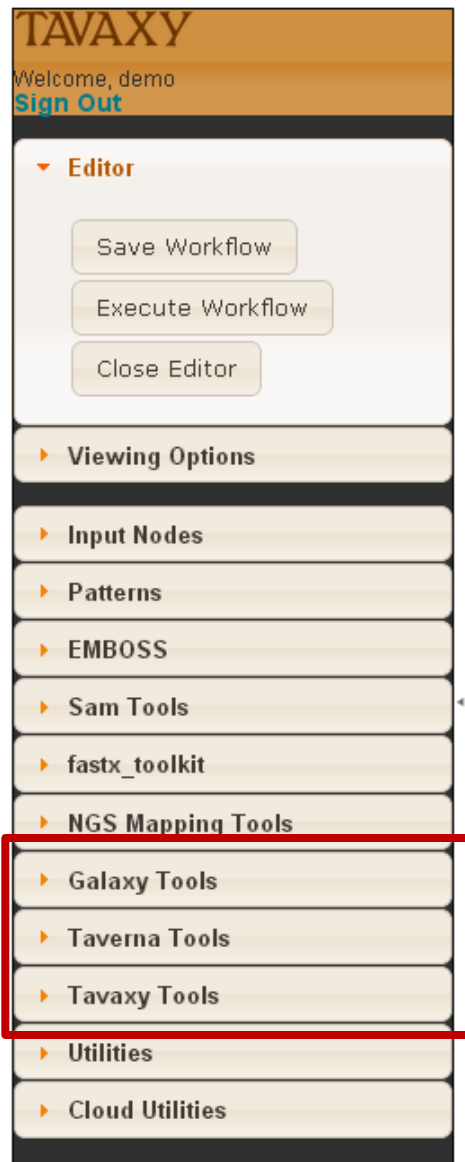
graph LR
    A[PfamAFamilyList (text)] -- default --> B[show result]
    B -- ok --> C[PfamAFamilyList (XML)]
    C -- default --> D[show result]
    
```

SN 5293868311 | 0 Cells selected | 41 MB - 85% | ALT Developer 100%

- Exploration de modèles et « model checking »
- Transformation de workflows en exécutables (Java et C++)
- Contient des structures de contrôle (boucles for)
- Bioinformatique

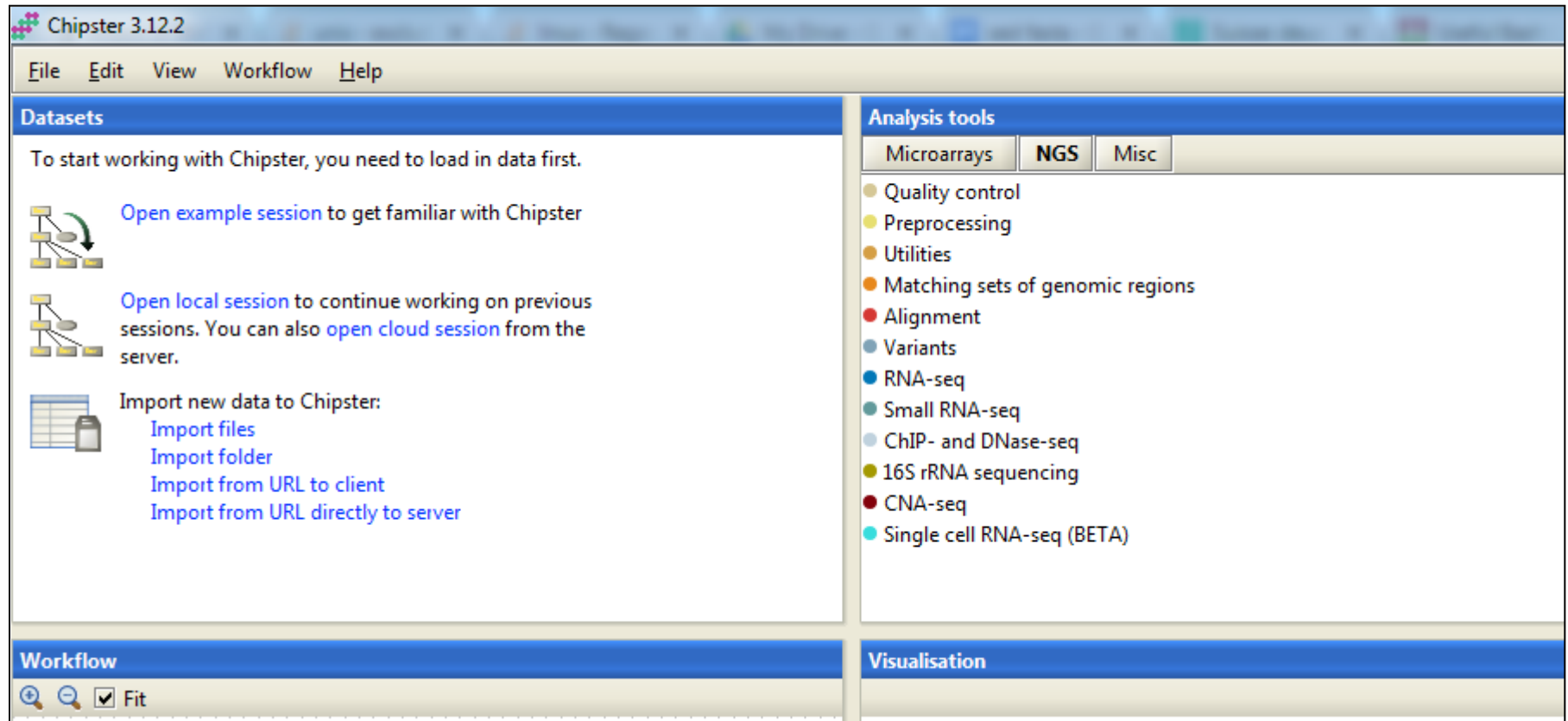
<http://ls5-www.cs.tu-dortmund.de/projects/biojeti/index.php>

# Tavaxy : lien entre Taverna et Galaxy



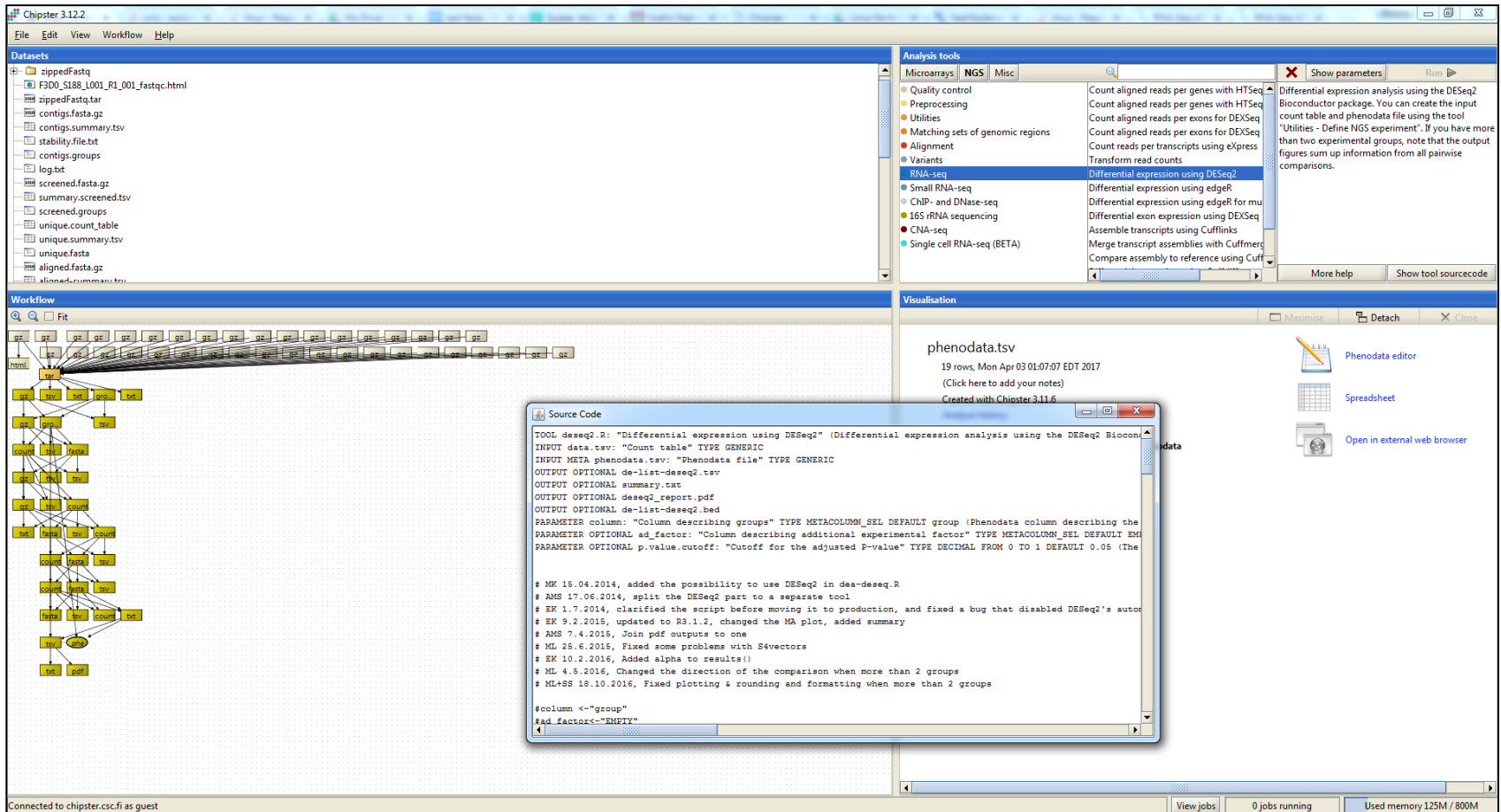
Abouelhoda M, Issa SA, Ghanem M. Tavaxy: integrating Taverna and Galaxy workflows with cloud computing support. BMC Bioinformatics. 2012 May 4;13:77.

# Chipster



<https://chipster.github.io/chipster/>

<https://chipster.csc.fi>



Kallio, M. Aleks, Jarno T. Tuimala, Taavi Hupponen, Petri Klemelä, Massimiliano Gentile, Ilari Scheinin, Mikko Koski, Janne Käki, and Eija I. Korpelainen. "Chipster: user-friendly analysis software for microarray and other high-throughput data." *BMC genomics* 12, no. 1 (2011): 507.

# Exemple Chipster

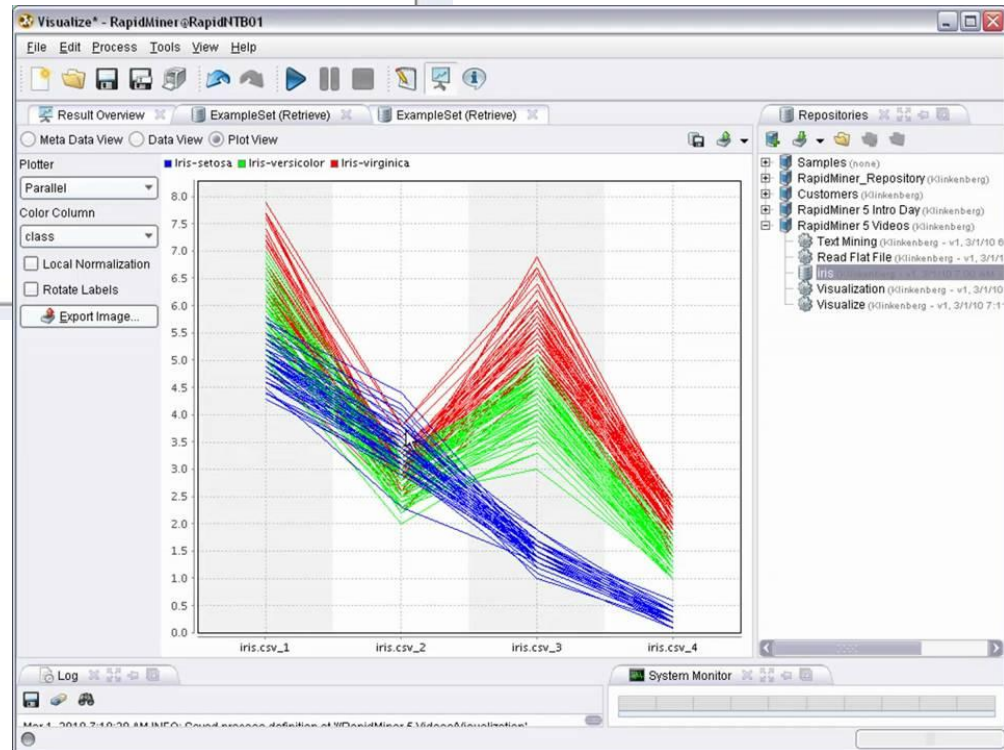
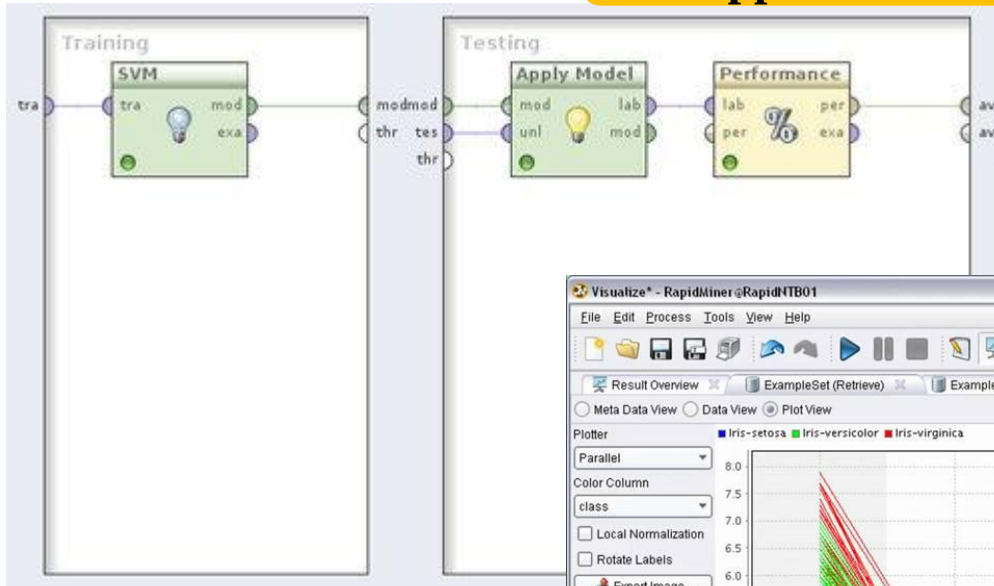
```
TOOL fastx-quality-filter.R: "Filter reads for quality with FastX" (Filters out reads
which contain quality scores below the user-specified criteria. This tool is based on
the FASTQ Quality Filter tool of the FASTX package.)
INPUT reads.fastq TYPE GENERIC
OUTPUT quality-filtered.fastq.gz
OUTPUT quality-filtered.log
PARAMETER quality: "Quality cut-off value" TYPE INTEGER FROM 1 TO 100 DEFAULT 20 (What
is the minimum quality score to keep.)
PARAMETER percentage: "Minimum percent of bases that must have that quality" TYPE
INTEGER FROM 1 TO 100 DEFAULT 90 (Percent of bases in sequence that must have quality
equal to or higher than the cut-off value.)
PARAMETER quality.format: "Quality value format used" TYPE [sanger: Sanger, illuminaold:
"Illumina GA v1.3-1.5"] DEFAULT sanger (What quality encoding is used in your FASTQ
file. Select Sanger if your data comes from Illumina 1.8 or later, SOLiD or 454.)

# EK 28.6.2011
# AMS 11.3.2014, gzip fastq outputs
# check out if the file is compressed and if so unzip it
source(file.path(chipster.common.path, "zip-utils.R"))
unzipIfGZipFile("reads.fastq")
# binary
binary <- c(file.path(chipster.tools.path, "fastx", "bin", "fastq_quality_filter"))
# command
quality.scale <- ifelse(quality.format == "sanger", "-Q 33", "")
command <- paste(binary, "-v", "-q", quality, "-p", percentage, quality.scale, "-i
reads.fastq -o quality-filtered.fastq > quality-filtered.log")

# run
system(command)
system("gzip *.fastq")
```



Locale | Langage **Java** | *Data-flow*  
Application en AI et Machine Learning



Armadillo v1.1

[adn.bioinfo.uqam.ca/armadillo](http://adn.bioinfo.uqam.ca/armadillo)

<https://github.com/armadilloUQAM/armadillo2>



HOMEPAGE

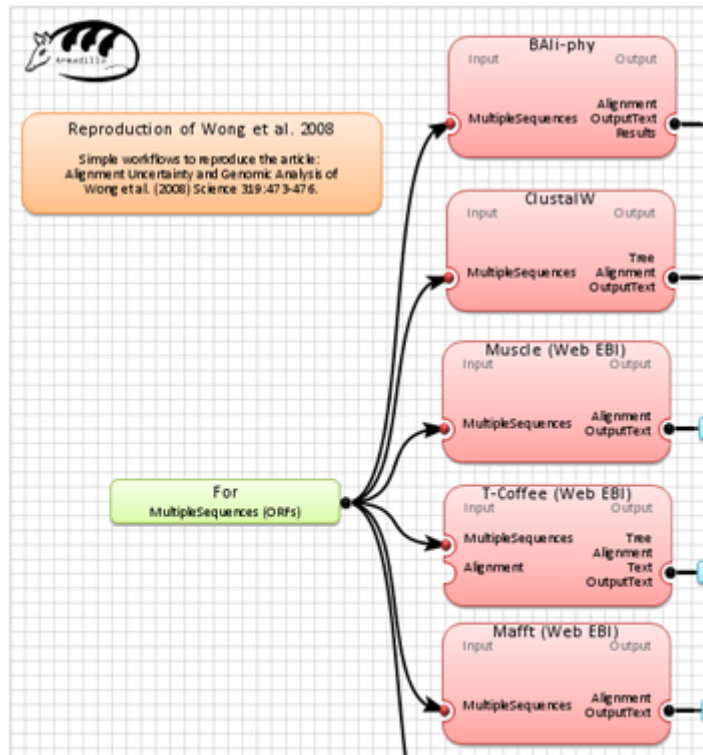
APPLICATIONS

TUTORIALS

WIKI

DOWNLOAD

A sample phylogenetic workflow reproduction



OPEN ACCESS Freely available online

PLOS one

## Armadillo 1.1: An Original Workflow Platform for Designing and Conducting Phylogenetic Analysis and Simulations

Etienne Lord<sup>1</sup>, Mickael Leclercq<sup>1</sup>, Alix Boc<sup>2</sup>, Abdoulaye Baniré Diallo<sup>1</sup>, Vladimir Makarenkov<sup>1\*</sup>

<sup>1</sup> Département d'informatique, Université du Québec à Montréal, Montréal, Canada, <sup>2</sup> Département de sciences biologiques, Université de Montréal, Montréal, Canada

### Abstract

In this paper we introduce Armadillo v1.1, a novel workflow platform dedicated to designing and conducting phylogenetic studies, including comprehensive simulations. A number of important phylogenetic and general bioinformatics tools have been included in the first software release. As Armadillo is an open-source project, it allows scientists to develop their own modules as well as to integrate existing computer applications. Using our workflow platform, different complex phylogenetic tasks can be modeled and presented in a single workflow without any prior knowledge of programming techniques. The first version of Armadillo was successfully used by professors of bioinformatics at Université du Québec à Montréal during graduate computational biology courses taught in 2010–11. The program and its source code are freely available at: <http://www.bioinfo.uqam.ca/armadillo>.

**Citation:** Lord E, Leclercq M, Boc A, Diallo AB, Makarenkov V (2012) Armadillo 1.1: An Original Workflow Platform for Designing and Conducting Phylogenetic Analysis and Simulations. PLOS ONE 7(1): e29903. doi:10.1371/journal.pone.0029903

**Editor:** Simon Joly, Montreal Botanical Garden, Canada

**Received:** June 16, 2011; **Accepted:** December 8, 2011; **Published:** January 11, 2012

**Copyright:** © 2012 Lord et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** EL and AB are Natural Science and Engineering Research Council of Canada fellows (NSERC). VM and ABD hold NSERC discovery grants. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: [vladimir.makarenkov@uqam.ca](mailto:vladimir.makarenkov@uqam.ca)

### Introduction

Bioinformatics is a fast-evolving field that encompasses

Meanwhile, the increasing use of genomic and phylogenetic data fuels the need for pipeline managing software. For instance, Cicarelli et al. [9] developed an automatable procedure for

# Armadillo v1.0 (adn.bioinfo.uqam.ca/armadillo)

Armadillo 1.0 - C:\armadillo2\trunk\armadillo\projects\ldoris3.db

File Edit Manager Tools Help

Toolbox Workflow Output Generate Report

Workflow Tools

Tools Database Current Workflow

Armadillo  
(+) (Beta testing)  
(+) Alignment  
(+) Ancestral Reconstruction  
(+) Blast  
(+) Comments  
(+) Conversion  
(+) EBI  
(+) Horizontal Genes Transfer  
(+) HGT Detector 3.2  
(+) Laktans  
(+) Phylotree v2.1 - R  
(+) Input  
(+) Logical conditions  
(+) Model Testing  
(+) Ncbi eUtils

Expand / Collapse toolbox

Enter search query and press Enter

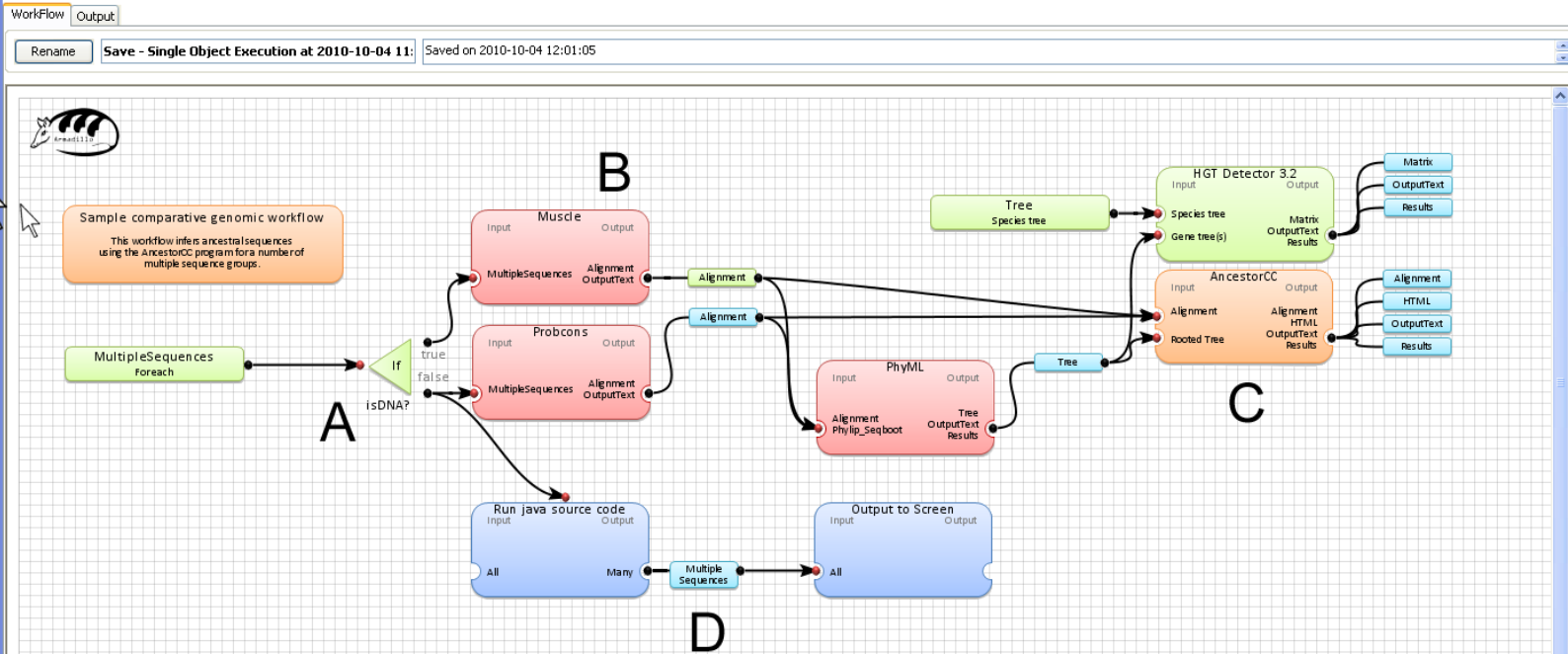
Database Content

(+) Sequence [16320]  
(+) MultipleSequences [12]  
(+) Alignment [12]  
(+) Ancestor  
(+) Tree [35]  
(+) MultipleTrees [27]  
(+) Matrix [8]  
(+) Genome  
(+) Workflows [59]  
(+) RunWorkflow [26]  
(+) Text [61]

Expand / Collapse datasets

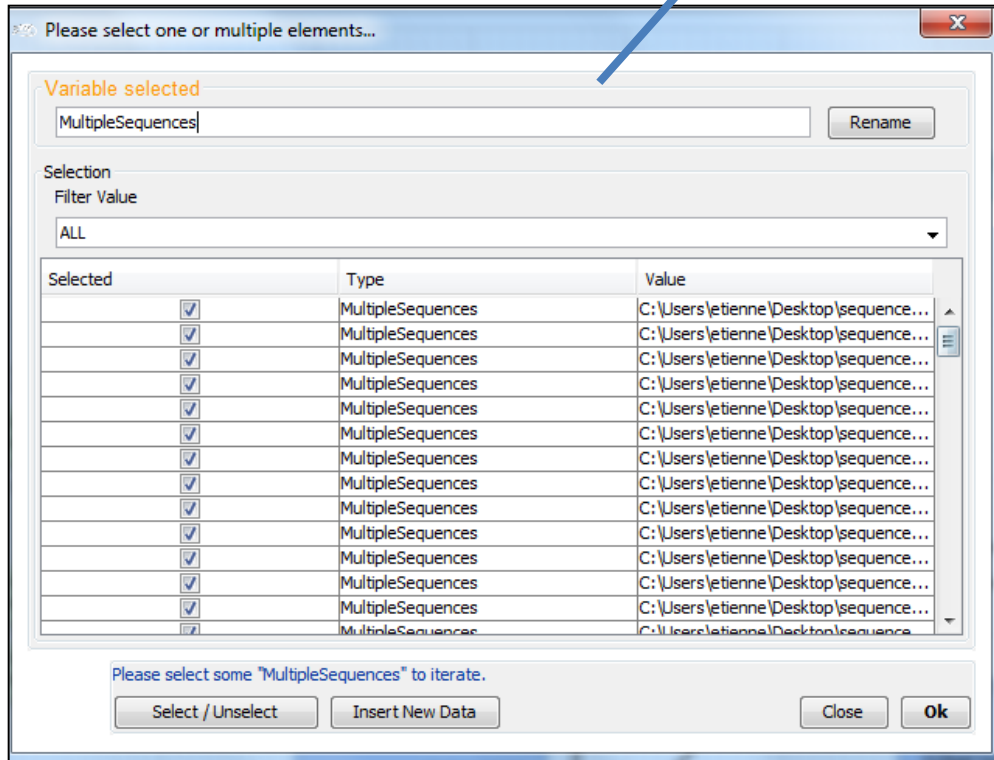
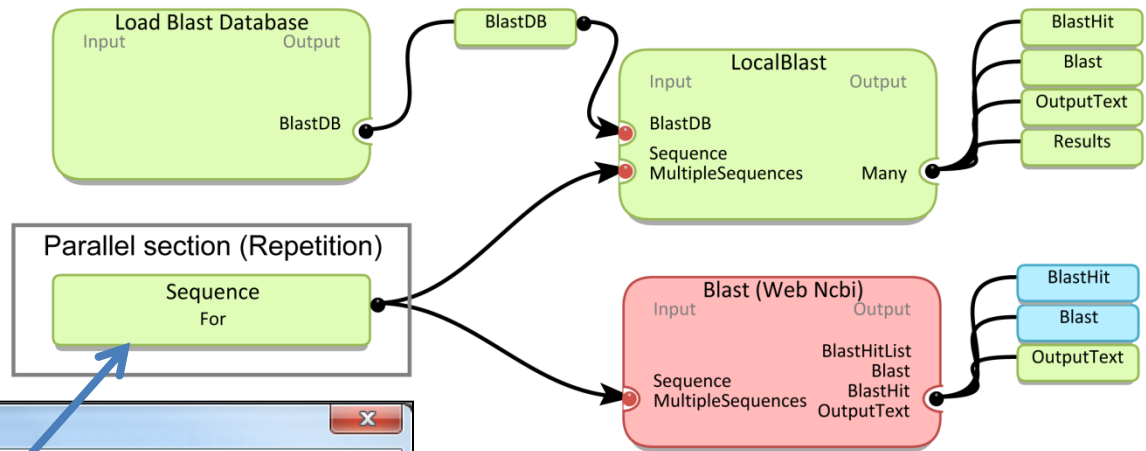
Drag and Drop to the Workflow

Locale | Langage **Java** | Données internes (SQLite)  
**Data-flow et Control-flow sans programmation | Phylogénétique**



- (A) Une structure de *control-flow* (if).
- (B) Différents **types** d'alignements de séquences.
- (C) Différentes **couleurs** pour ajouter au processus cognitif.
- (D) Exécution conditionnelle de code source *Java*.

## Armadillo : structures de répétitions



## Types de données

Sequence,  
MultipleSequences, Tree,  
MultipleTrees, Blast,  
Alignment, Genome, Text

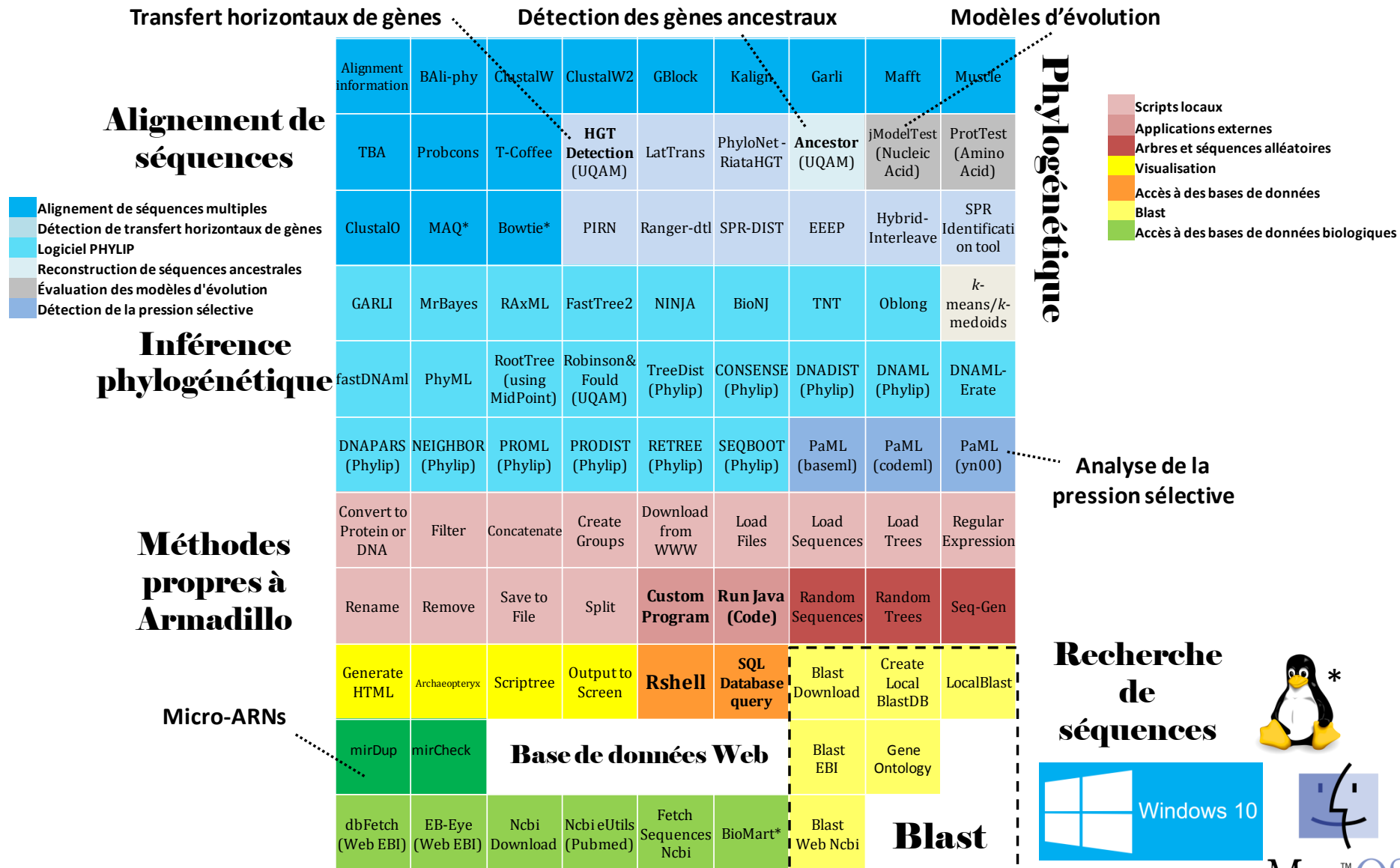
## Dernière version

[github.com/armadilloUQAM/armadillo2](https://github.com/armadilloUQAM/armadillo2)



En développement,  
exécutables sous Docker.io

# Armadillo : plusieurs applications incluses ...



\* Les applications sous Linux sont en cours de développement



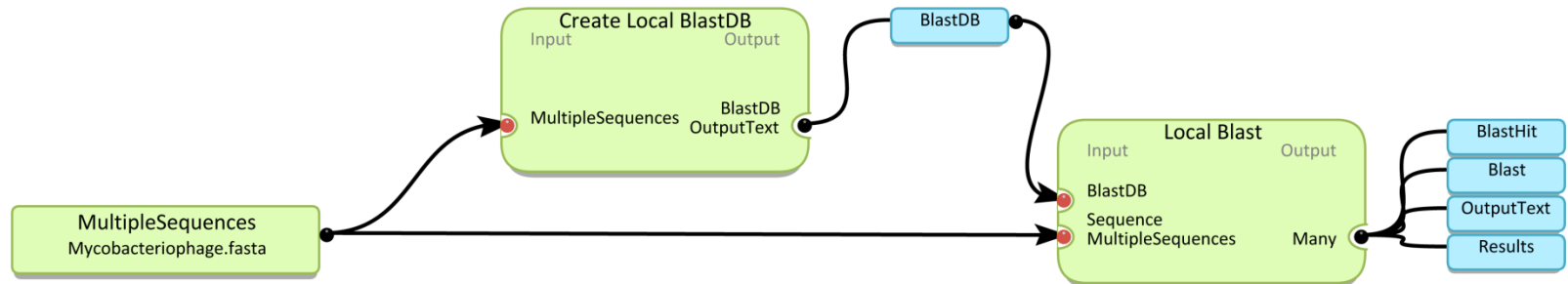
Mac OS

# Démo Armadillo



## Sample local BLAST

Using NSBI BLAST. This workflow create a local blast database from the MultipleSequences and then, as a verification BLAST the MultipleSequences against it.



# ... et une plate-forme qui pourrait être adaptée à d'autres domaines.

|                           |                  |                           |                         |                   |                     |                  |                          |                       |
|---------------------------|------------------|---------------------------|-------------------------|-------------------|---------------------|------------------|--------------------------|-----------------------|
| Alignment information     | BALI-phy         | ClustalW                  | ClustalW2               | GBlock            | Kalign              | Garli            | Mafft                    | Muscle                |
| TBA                       | Probcons         | T-Coffee                  | HGT Detector (UQAM)     | LatTrans          | PhyloNet - RiataHGT | Ancestor (UQAM)  | ModelTest (Nucleic Acid) | ProtTest (Amino Acid) |
| fastDNAmI                 | PhyML            | RootTree (using MidPoint) | Robinsons Fould (PhyML) | TreeDist (PhyML)  | CONSENSE (PhyML)    | DNADIST (PhyML)  | DNAML (PhyML)            | DNAML-Erate           |
| DNAPARS (PhyML)           | NEIGHBOR (PhyML) | PROML (PhyML)             | PRODIST (PhyML)         | RETRIE (PhyML)    | SEQBOOT (PhyML)     | PaML (baseml)    | PaML (codeml)            | PaML (yn00)           |
| Convert to Protein or DNA | Filter           | Concatenate               | Create Groups           | Download from WWW | Load Files          | Load Sequences   | Load Trees               | Regular Expression    |
| Rename                    | Remove           | Save to File              | Split                   | Custom Program    | Run Java (Code)     | Random Sequences | Random Trees             | Seq-Gen               |
| Generate HTML             | Archaeopteryx    | Scriptree                 | Output to Screen        | Rshell*           | SQL Database query  | Blast Download   | Create Local BlastDB     | LocalBlast            |

|                   |                  |               |                      |                      |          |                |
|-------------------|------------------|---------------|----------------------|----------------------|----------|----------------|
| dbFetch (Web EBI) | EB-Eye (Web EBI) | Ncbi Download | Ncbi eUtils (Pubmed) | Fetch Sequences Ncbi | BioMart* | Blast Web Ncbi |
|-------------------|------------------|---------------|----------------------|----------------------|----------|----------------|

## Phylogénétique

En développement  
Version *BeanShell*

## Une sélection de commandes Linux

|         |          |       |          |          |          |           |            |          |
|---------|----------|-------|----------|----------|----------|-----------|------------|----------|
| cd      | ls       | rm    | cp       |          | mv       | chown     | chmod      | chgrp    |
| pwd     | mkdir    | rmdir | scp      | rcp      | mmv      | su / sudo | date       | times    |
| sleep   | uniq     | nice  | slocate  | locate   | df       | bzip2     | tar        | gzip     |
| man     | help     | read  | paste    | find     | du       | sort      | tsort      | tree     |
| top     | ps       | bg    | cron     | crontab  | time     | env       | export     | kill     |
| let     | declare  | set   | unset    | uuencode | uudecode | alias     | adduser    | addgroup |
| printf  | clear    | echo  | exec     | true     | false    | wget      | awk/gawk   | sed      |
| command | function | if    | until    | eval     | more     | wc        | grep/egrep | head     |
| case    | for      | while | continue | break    | less     | tail      | cat        | cut      |

# Problématiques actuelles

| Défis actuels<br>( <i>Romano, 2008</i> )          | 2017<br>(solutions et recherches)                                                                               |
|---------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| Abstraction limité des types de données           | Ontologies bioinformatiques  |
| Accès programmatiques aux données difficiles      | Services REST/SOAP                                                                                              |
| Mauvaises performances des WfMS                   | Apache Hadoop, etc.                                                                                             |
| Ressources (applications) utilisables limités     | Ontologies? Docker?          |
| Mauvaise sémantique limite la réutilisation       | Un langage de workflow commun: CWL                                                                              |
| Difficultés à conserver les traces des exécutions | Ontologies?                 |

Romano P. (2008) Automation of in-silico data analysis processes through workflow management systems. Brief Bioinform. 9(1):57-68

# Common workflow language : CWL



<https://github.com/common-workflow-language/common-workflow-language>

- Arvados Project (Curoverse)
- Broad Institute
- Cincinnati Children's Hospital
- Galaxy Project
- Harvard Chan School of Public Health
- Institut Pasteur
- Oregon Health & Science University
- Sanger Institute
- Seven Bridges
- Taverna
- UC Santa Cruz
- UC Davis

cwltool (reference implementation)

Rabix

Arvados

Galaxy

Parallel Recipes

Toil

CancerCollaboratory

Airflow (SciDAP)



**Broad Institute CWL meetup 2015-11-13.pdf**

- Modeled as user defined function (UDF)
- Typed input / output signature
- Inputs & outputs are fully specified
- Tool executions are isolated from one another & from parent process
- Well defined execution process
  1. Collect & validate inputs
  2. Map input file paths to locations inside container
  3. Build tool command line
  4. Build Docker invocation
  5. Execute
  6. Collect & validate outputs

**Broad Institute CWL meetup 2015-11-13.pdf**

# CWL : SamTools (<http://samtools.sourceforge.net/>)

```
class: CommandLineTool
cwlVersion: draft-3
description: Sort by chromosomal coordinates
```

```
requirements:
- class: DockerRequirement
  dockerPull: scidap/samtools:v1.2-216-gdffc67f
```

```
inputs:
- id: input
  type: File
  inputBinding:
    position: 1

- id: output_name
  type: string
  inputBinding:
    position: 2
```

```
outputs:
- id: output
  type: File
  outputBinding:
    glob: $(inputs.output_name)
```

```
baseCommand: [samtools, sort]
```

← File type & metadata

← Runtime environment



← Input parameters

← Output parameters

← Executable

# Existing workflow system

- <https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems>

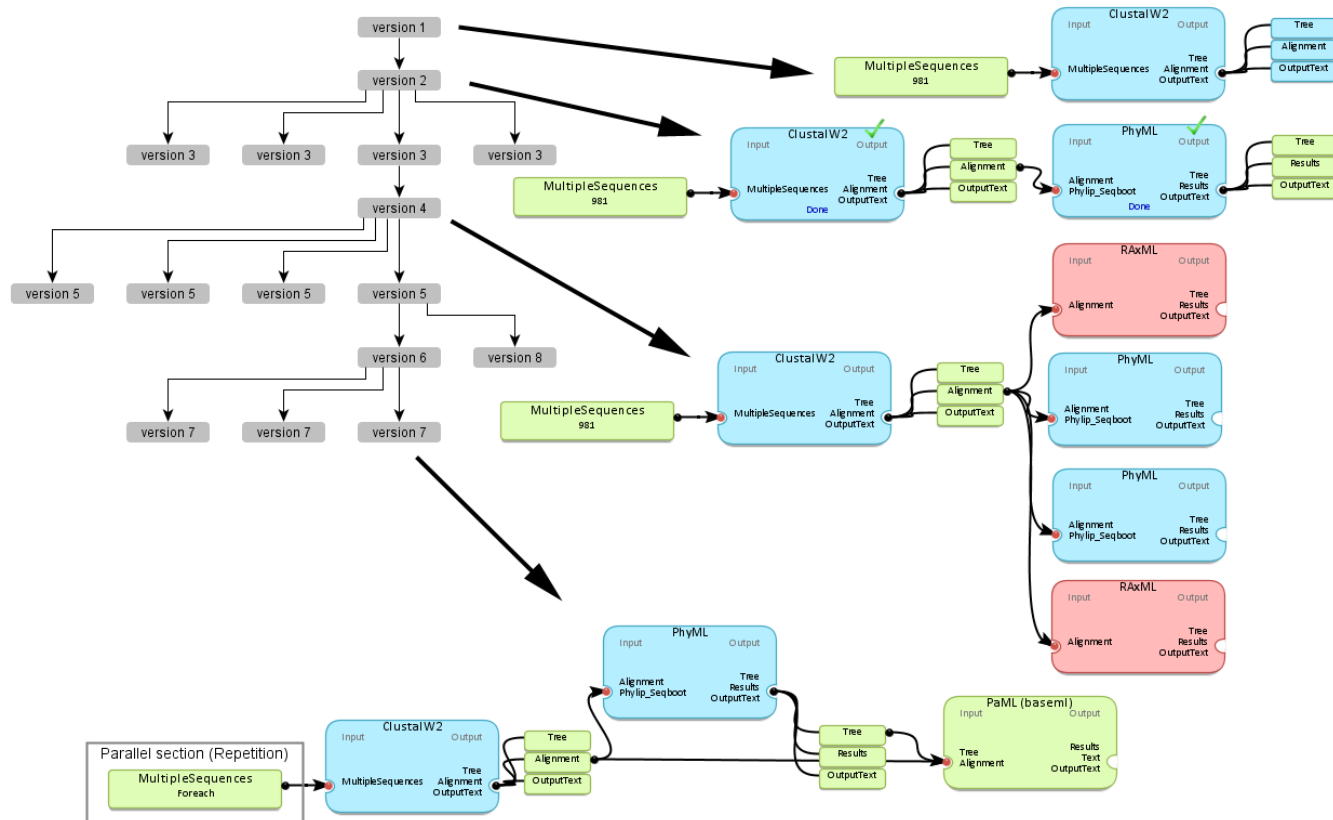


## Example: Grep and Count Workflow



<https://github.com/common-workflow-language/workflows/blob/master/workflows/presentation-demo/grep-and-count.cwl>

**Problématique:** une expérimentation *in silico* évolue et résulte en différentes versions d'un flux de travaux.  
**Comment les comparer?**



Lord, E., Diallo, A. B., Makarenkov, V. (2015). Classification of bioinformatics workflows using weighted versions of partitioning and hierarchical clustering algorithms. *BMC Bioinformatics*, 16(1), 68.

# Différents encodages des flux de travaux

Flux de  
travaux

**Encodage**

mesure de  
distance

méthode de  
regroupement

Nombre de  
groupes

Matrice binaire

| Encoding of Type I       | W1 | W2 | W3 | W4 | W5 | Weights for Encoding of Type I |
|--------------------------|----|----|----|----|----|--------------------------------|
| Blast (NCBI)             | 0  | 0  | 0  | 1  | 0  | 0.35                           |
| ClustalW2                | 0  | 1  | 0  | 0  | 1  | 0.49                           |
| HGT Detector 3.2         | 1  | 1  | 1  | 0  | 1  | 0.88                           |
| Muscle                   | 1  | 0  | 0  | 0  | 1  | 0.41                           |
| PROTML (Phylip)          | 1  | 0  | 0  | 0  | 0  | 0.68                           |
| PhyML (1)                | 0  | 1  | 1  | 0  | 1  | 1.13                           |
| PhyML (2)                | 0  | 0  | 0  | 0  | 1  | 1.13                           |
| Probcons                 | 0  | 0  | 1  | 0  | 0  | 0.55                           |
| Robinson&Foulds distance | 0  | 0  | 0  | 1  | 0  | 0.25                           |
| SEQBOOT                  | 1  | 0  | 0  | 0  | 0  | 0.14                           |
| Seq-Gen                  | 0  | 1  | 0  | 1  | 0  | 0.43                           |

Disperser les tâches  
similaires pour des  
**exécutions plus  
rapides**

-- Vecteur de temps moyens

| Encoding of Type II      | W1 | W2 | W3 | W4 | W5 | Weights for Encoding of Type II |
|--------------------------|----|----|----|----|----|---------------------------------|
| Blast (NCBI)             | 0  | 0  | 0  | 1  | 0  | 0.10                            |
| ClustalW2                | 0  | 1  | 0  | 0  | 1  | 0.10                            |
| HGT Detector 3.2         | 1  | 1  | 1  | 0  | 1  | 1.00                            |
| Muscle                   | 1  | 0  | 0  | 0  | 1  | 0.10                            |
| PROTML (Phylip)          | 1  | 0  | 0  | 0  | 0  | 0.10                            |
| PhyML                    | 0  | 1  | 1  | 0  | 2  | 0.10                            |
| Probcons                 | 0  | 0  | 1  | 0  | 0  | 0.10                            |
| Robinson&Foulds distance | 0  | 0  | 0  | 1  | 0  | 0.10                            |
| SEQBOOT                  | 1  | 0  | 0  | 0  | 0  | 0.10                            |
| Seq-Gen                  | 0  | 1  | 0  | 1  | 0  | 0.10                            |

Grouper les flux de  
travaux similaires en  
**fonction de mots-clés**

-- Vecteur de poids

Matrice d'occurrences

# Différents encodages des flux de travaux

Matrice d'occurrences

| Encoding of Type III             | W1 | W2 | W3 | W4 | W5 | Weights for Encoding of Type III |
|----------------------------------|----|----|----|----|----|----------------------------------|
| Blast (NCBI)                     | 0  | 0  | 0  | 1  | 0  | 0.35                             |
| HGT Detector 3.2                 | 1  | 1  | 1  | 0  | 1  | 0.88                             |
| Robinson&Foulds distance         | 0  | 0  | 0  | 1  | 0  | 0.25                             |
| ClustalW2→PhyML                  | 0  | 1  | 0  | 0  | 1  | 1.62                             |
| Muscle→PhyML                     | 0  | 0  | 0  | 0  | 1  | 1.54                             |
| Muscle→SEQBOOT (Phylip)          | 1  | 0  | 0  | 0  | 0  | 0.55                             |
| PROTML (Phylip)→HGT Detector 3.2 | 1  | 0  | 0  | 0  | 0  | 1.56                             |
| PhyML→HGT Detector 3.2           | 0  | 1  | 1  | 0  | 2  | 2.01                             |
| Probcons→PhyML                   | 0  | 0  | 1  | 0  | 0  | 1.68                             |
| SEQBOOT (Phylip)→PROTML (Phylip) | 1  | 0  | 0  | 0  | 0  | 0.82                             |
| Seq-Gen→Blast (NCBI)             | 0  | 0  | 0  | 1  | 0  | 0.78                             |
| Seq-Gen→ClustalW2                | 0  | 1  | 0  | 0  | 0  | 0.92                             |

Disperser les tâches similaires pour des **exécutions plus rapides**

-- Vecteur de temps moyens

| Encoding of Type IV              | W1 | W2 | W3 | W4 | W5 | Weights for Encoding of Type IV |
|----------------------------------|----|----|----|----|----|---------------------------------|
| Blast (NCBI)                     | 0  | 0  | 0  | 1  | 0  | 0.10                            |
| HGT Detector 3.2                 | 1  | 1  | 1  | 0  | 1  | 1.00                            |
| Robinson&Foulds distance         | 0  | 0  | 0  | 1  | 0  | 0.10                            |
| ClustalW2→PhyML                  | 0  | 1  | 0  | 0  | 1  | 0.10                            |
| Muscle→PhyML                     | 0  | 0  | 0  | 0  | 1  | 0.10                            |
| Muscle→SEQBOOT (Phylip)          | 1  | 0  | 0  | 0  | 0  | 0.10                            |
| PROTML (Phylip)→HGT Detector 3.2 | 1  | 0  | 0  | 0  | 0  | 1.00                            |
| PhyML→HGT Detector 3.2           | 0  | 1  | 1  | 0  | 2  | 1.00                            |
| Probcons→PhyML                   | 0  | 0  | 1  | 0  | 0  | 0.10                            |
| SEQBOOT (Phylip)→PROTML (Phylip) | 1  | 0  | 0  | 0  | 0  | 0.10                            |
| Seq-Gen→Blast (NCBI)             | 0  | 0  | 0  | 1  | 0  | 0.10                            |
| Seq-Gen→ClustalW2                | 0  | 1  | 0  | 0  | 0  | 0.10                            |
| INPUT_Sequences                  | 1  | 0  | 1  | 0  | 1  | 1.00                            |
| INPUT_Tree                       | 1  | 1  | 1  | 2  | 0  | 1.00                            |
| OUTPUT_Blast (NCBI)              | 0  | 0  | 0  | 1  | 0  | 1.00                            |
| OUTPUT_Matrix                    | 1  | 1  | 1  | 1  | 1  | 1.00                            |
| OUTPUT_MultipleTrees             | 0  | 0  | 0  | 1  | 0  | 1.00                            |
| OUTPUT_OutputText                | 1  | 1  | 1  | 2  | 1  | 1.00                            |
| OUTPUT_Results                   | 1  | 1  | 1  | 1  | 1  | 1.00                            |

Encourager la **réutilisation des données** (génomiques)

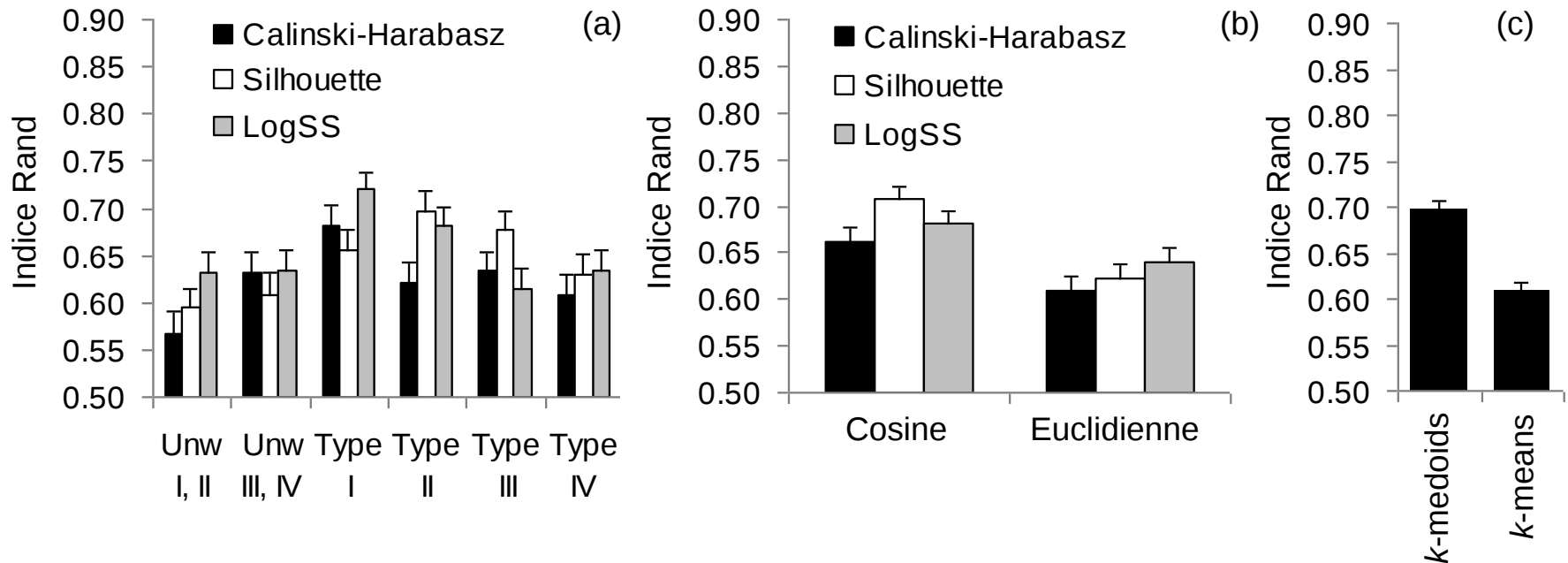
-- Vecteur de poids

Matrice d'occurrences

Paires de tâches

# Conclusions pour le regroupement par partitionnement

- (a) Effet des critères d'optimisation pour les encodages non pondérés (les deux premiers jeux de barres) et pondérés (quatre derniers jeux de barres);
- (b) Effet de la mesure de distances;
- (c) Effet de l'algorithme de partitionnement appliqué;



**La distance cosine pondérée, utilisée avec l'algorithme *k-medoids*, l'encodage de Type I et l'indice Silhouette, montre la meilleure performance.**

**N.B. Les plus grandes valeurs de l'indice Rand sont les meilleures.**

## Workflow spaghetti



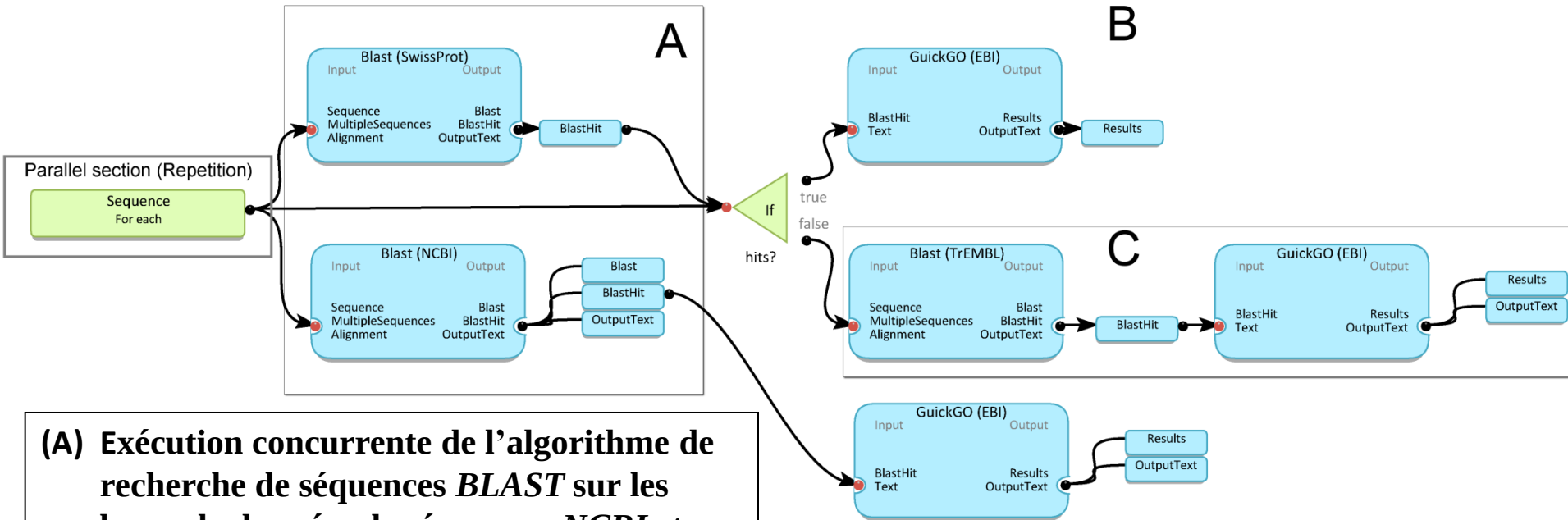
# Flux de travaux bioinformatiques

Agharbaoui,Z., Leclercq, M., Remita,M.A., Badawi, M., **Lord, E.**, Houde,M., Danyluk, J., Diallo, A.B., Sarhan, F. (2015) An integrative approach to identify hexaploid wheat miRNAome associated with development and tolerance to abiotic stress. *BMC Genomics* 16:339

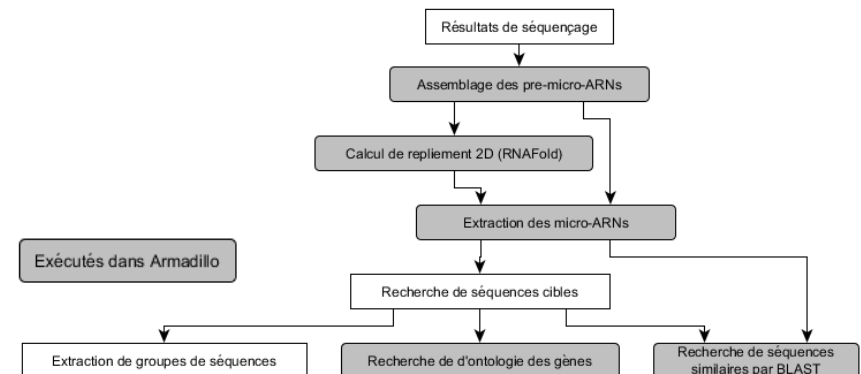
Remita MA, **Lord, E.**, et al. (2015) WMP: A novel comprehensive wheat miRNA database, including related bioinformatics software.  
doi: <http://dx.doi.org/10.1101/024893>

# Exemple : Flux de travaux utilisés pour annoter les séquences cibles

Création d'une banque de données de petits ARNs chez le blé.  
<http://wheat.bioinfo.uqam.ca/>



- (A) Exécution concurrente de l'algorithme de recherche de séquences *BLAST* sur les bases de données de séquences *NCBI* et *SwissProt*.
- (B) Exécution conditionnelle de la recherche ontologique si on a des résultats en (A).
- (C) Recherche par la méthode *BLAST*, mais sur la base de données *TReMBL* et recherche ontologique sur ces résultats, s'il n'y a pas de résultats en (A).

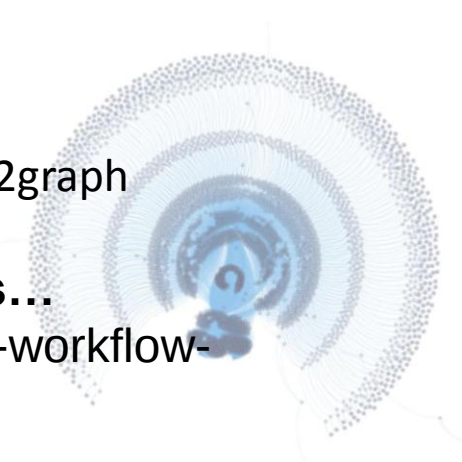


# Autres paradigmes : les langages de type *Pipe*

| Langages     | Exemple                                                                                                                                                                                                                    | Liens                                                                                                                                              |
|--------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| Swift        | <pre>foreach protein in proteinList { runBLAST(protein); }</pre>                                                                                                                                                           | <a href="http://swift-lang.org/main/">http://swift-lang.org/main/</a>                                                                              |
| GNU Make     | <pre>output.txt: bin/analyse input.txt bin/analyse input.txt &gt; output.txt</pre>                                                                                                                                         | <a href="http://bost.ocks.org/mike/make/">bost.ocks.org/mike/make/</a><br><a href="http://www.bioinformaticszen.com">www.bioinformaticszen.com</a> |
| MakeFlow*    | <pre>part1 part2 part3: input.data split.py ./split.py input.data</pre>                                                                                                                                                    | <a href="http://ccl.cse.nd.edu/software/makeflow/">ccl.cse.nd.edu/software/makeflow/</a>                                                           |
| Apache Spark | <pre>val text_file = spark.textFile("hdfs://...") val word_counts = text_file.flatMap(lambda line: line.split())             .map(lambda word: (word, .reduceByKey(lambda a, b: a+b) val count = word_counts.count()</pre> | <a href="http://spark.apache.org/">http://spark.apache.org/</a>                                                                                    |



Pierre Lindenbaum  
(<http://plindenbaum.blogspot.ca/>)  
<https://github.com/lindenb/makefile2graph>



**Encore: une liste des workflows en bioinformatiques...**

<https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems>

# Autres paradigmes : Big Data

[Home](#)[Documentation](#)[Powered By Apex](#)[Roadmap](#)[Community](#)[Github](#)[Download](#)

## Apache Apex™

Enterprise-grade unified stream and batch processing engine.

*Now with event-time windowing and high-level API.*

### Enterprise Grade

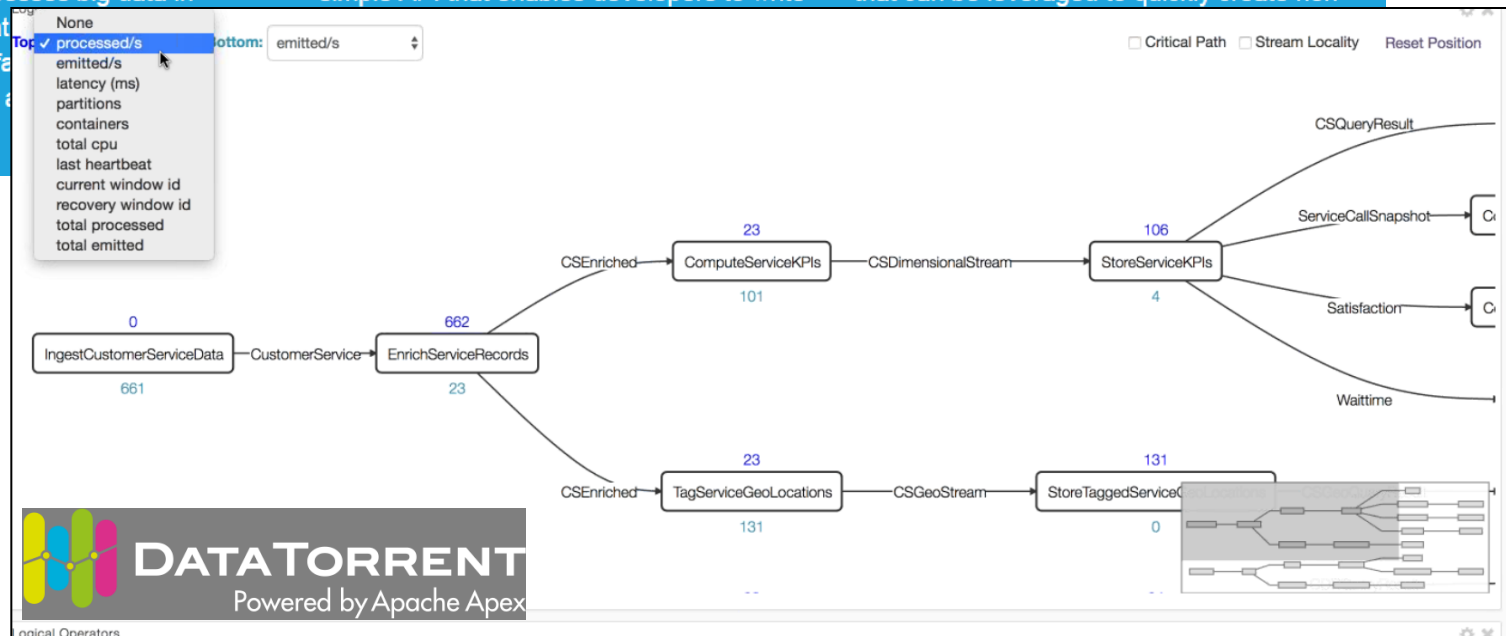
Apex is a Hadoop YARN native platform that unifies stream and batch processing. It processes big data information in a way that is highly performant, fault-tolerant, secure, distributed, and

### Low Barrier-to-Entry

Write your business logic and leave all operability to the platform. It provides a simple API that enables developers to write

### Modular

The Apex platform comes with Malhar, a library of operators (units of functionality) that can be leveraged to quickly create non-

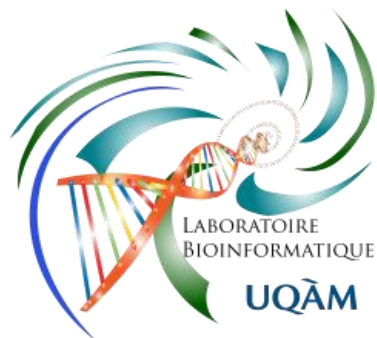


- La plate-forme *Armadillo* ne requiert pas de programmation.
- Elle ne nécessite pas d'applications externes et permet le partage de workflows.
- La plate-forme *Armadillo* pourrait être adaptée à d'autres domaines scientifiques.
- Il reste beaucoup de recherches à effectuer dans le domaine des flux de travaux: ontologie, dispersion des tâches, meilleure abstraction, etc.
- Une meilleure intégration des services webs serait aussi souhaitée ainsi que l'intégration du CWL.

**Vladimir Makarenkov, Université du Québec à Montréal**

**Abdoulaye Baniré Diallo, Université du Québec à Montréal**

**François-Joseph Lapointe, Université de Montréal**



**UQÀM**

Université   
de Montréal



*Fonds de recherche  
sur la nature  
et les technologies*

Québec 



**NSERC  
CRSNG**

- Lord, E., Leclercq, M., Boc, A., Diallo, A.B., et Makarenkov, V. (2012). Armadillo 1.1: an original workflow platform for designing and conducting phylogenetic analysis and simulations. *PloS One*, 7(1), e 29903.
- Oinn, T., Greenwood, M., Addis, M., Alpdemir, M. N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M. R., Senger, M., Stevens, R., Wipat, A., et Wroe, C. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10), 1067-1100.
- Oinn, T., Li, P., Kell, D. B., Goble, C., Goderis, A., Greenwood, M., Hull, D., Stevens, R., Turl, D., et Zhao, J. (2007). Taverna/myGrid: aligning a workflow system with the life sciences community. Dans *Workflows for e-Science* (p. 300-319). Springer London.
- Sharma, A., Rai, A., et Lal, S. (2013). Workflow management systems for gene sequence analysis and evolutionary studies - A Review. *Bioinformation*, 9(13), 663-672.
- Silva, V., Chirigati, F., Maia, K., Ogasawara, E., Oliveira, D., Braganholo, V., Murta, L., et Mattoso, M. (2011). Similarity-based workflow clustering. *Journal of Computational Interdisciplinary Sciences*, 2(1), 23-35.