MetaGenSense : A Web application for analysis and visualization of High throughput Sequencing metagenomic data

Damien Correia[1], Olivia Doppelt-Azeroual[2], Jean-Baptiste Denis[3], Mathias Vandenbogaert[1], Valérie Caro[1]

[1]Pôle Génotypage des Pathogènes (PGP), Unité Environnement et Risques Infectieux (ERI), Institut Pasteur, Paris, France.

[2]Centre d'Informatique pour la Biologie (CIB), Institut Pasteur, Paris, France.

[3]Groupe Exploitation et Infrastructure (E&I), Institut Pasteur , Paris, France.

Abstract

The detection and characterization of emerging infectious agents has been a continuing public health concern. High throughput sequencing (or Next-Generation Sequencing; NGS) technologies have proven promising approaches for unbiased detections of pathogens in complex biological samples. They are efficient and provide comprehensive analyses. However, NGS yields millions of putatively representative reads per sample, such an amount, that efficient data management and visualization resource have become mandatory requirement, through a dedicated ***Laboratory Information Management System*** (LIMS), solely to provide perspective regarding the information contained in this huge amount of data.

We developed a managing and **analytical bioinformatics framework** that is engineered **to run associated and dedicated Galaxy[1] workflows** for **the detection and eventually classification of pathogens**. In essence, our primary purpose is **to assist the biologist in the process of deciding on the most relevant sample-specific sequences** in the supplied samples, and determine their relative abundance. To this end, a user-friendly interface is essential. A complete set of specific Galaxy pipelines, producing high quality reads and/or assemblies meaningful for biological interpretation, have been engineered, and serve as the driving engine for a graphical web-interface associating the sample's meta-data and its analysis results. This user-interface has been tailored to associate a bio-IT provider resources **(a Galaxy instance, sufficient storage and grid computing power)**, with the input data and its metadata. Hence, the web application allows scientists to easily interact with existing Galaxy metagenomic workflows, facilitates the organization, visualization and aggregation of the most significant and most meaningful bits of information from millions of genomic sequences. In more detail, communication between our Django-based interface [2] and Galaxy uses the Bioblend library[3]. It gives access to a Galaxy instance's main features, through scripted and automated commands. **Metadata** about samples, **runs** as well as **the workflow results** are stored in the LIMS.

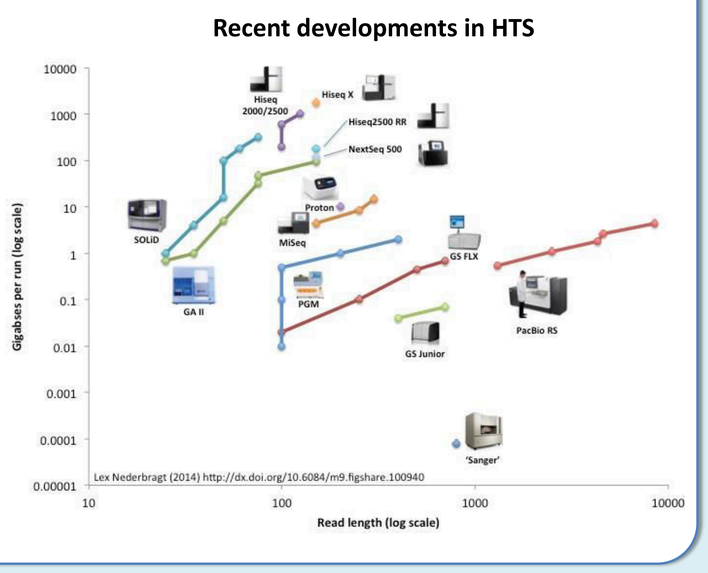
Visualization tools associating the sequencing raw data with the analysis results are as important as the analysis itself. The interface already integrates existing tools such as **KRONA** [4], and also enables sharing of scientific results with several project members. In the end, it will also allow the integration of other visualization tools (in development).

Background

Meta-genomics by HTS

High Throughput Sequencing (HTS) approaches yield millions of potentially interesting reads per sample. **Reads** are small fragment nucleotide stretches sequenced in millions of parallel reactions. They are described in FastQ format gathering sequence and the corresponding quality score:

```
@read1
ATTAAACCGGCAGGTC
+read1
efffcfdfea]^_^]B
```



Bioinformatics for HTS data

Systematic bioinformatic approaches are in use to analyse HTS data. Today, hundreds of tools are available, each with **specific parameters** and **diverse command lines**.

Galaxy[1], is a **scientific workflow management system**, which provides means to build multi-step computational analyses akin to a recipe.

Galaxy community provides, a tool sharing system called **toolshed** which facilitate a lot, tool installation in Galaxy Instances.

In Paris's Institut Pasteur, a galaxy team, manages support and tools installation, for more than 100 users.

One last very interesting functionality, is reproducibility, essential for the kind of analysis realized in the **MetaGenSense** project.

Bioblend is a python library built to remotely interact with Galaxy. To work, a galaxy instance *Object* is created using two parameters, a running Galaxy instance URL and an API key, generated by the user for authentication.

LIMS

Laboratory Information Management System

Biological samples

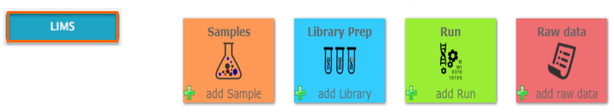
- Human
- Animal
- Environmental

A **LIMS** is an application to manage laboratory data. It is mainly used to track biological samples and associated metadata by recording them in a specific refined database. Data can be recorded using a client interface

MetaGenSense

A dedicated **LIMS**

1. **Record information about HTS projects**. For each project, the following items can be stored:

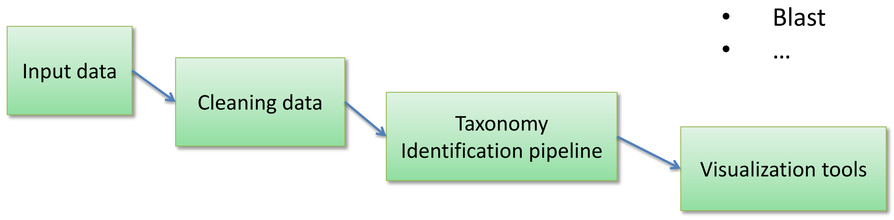


Metadata such as, sample localization, origin, extraction method, as well as comments can be stored to ensure sample traceability. The raw data is what's bioinformatically analyzed.

2. **Record information about bioinformatics analysis.** It is one of the main features. Which galaxy workflow was ran on which raw data. The user also gets to choose which results he wants to store in the LIMS, there is no automatic storage.

Pre-designed Galaxy workflows

Two metagenomics workflows are available in this application, one for paired-end reads, and one for single read analysis. Administrators are able to add new worflows as long as they have the workflow galaxy ID. The basic steps of the workflows are:



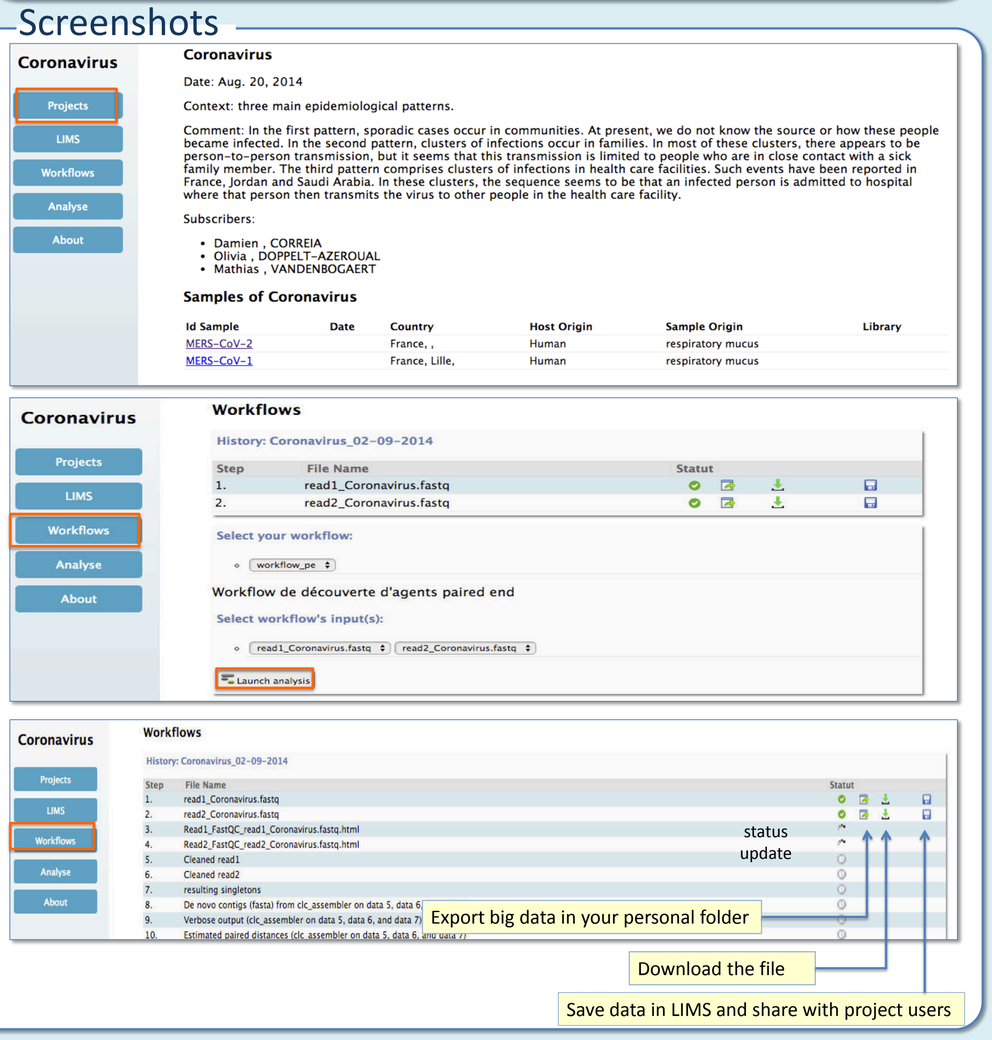
HTS data analyses

- Read control quality (FqCleaner)
- De novo Assembly
- Mapping
- Blast
- …

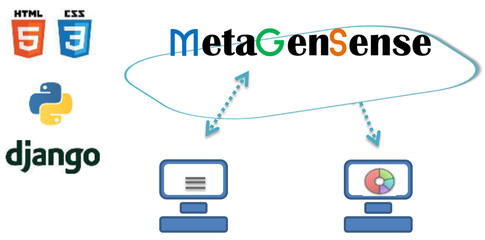
Grid computing

- Parallel operation
- Cluster resource
- Multi-JOBs

Screenshots



Architecture



Start. User creates a project in the LIMS

End. User consults results

bioblend

1. Folder is created in Galaxy user's library (project name)
2. Raw data selected are imported in the project folder
3. Project data can be displayed
4. Data are uploaded in the user's history
5. Selection of input(s) for the analysis
6. Workflow launching
7. Analysis status is available at each step of the execution
8. Results are available to be retrieved (downloaded, stored in LIMS, opened in the application)

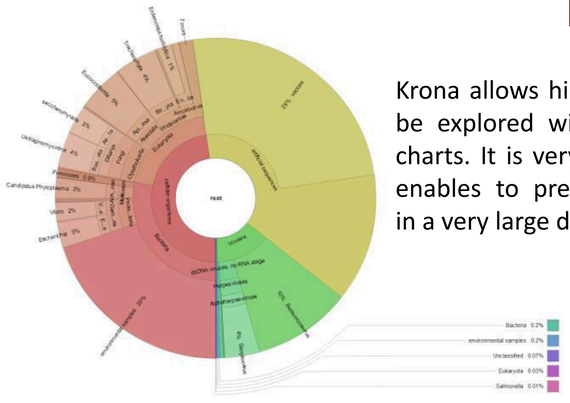
Jobs are launched by galaxy on a computing cluster

Visualization tools

Interactive visualization systems allow scientists to intuitively navigate and focus on the most pertinent results that might correspond to a family of microorganisms.

Through a link in the analyse part of MetaGenSense, the Krona pie chart is available. The scientific will be able to identify specific agents detected through the workflow.

If the analysis in terminated, the user can choose to download the results, or share them with other users logged on MetaGenSense



KRONA

Krona allows hierarchical data to be explored with zoomable pie charts. It is very interesting as it enables to preview information in a very large data set of taxa.

Conclusion

MetaGenSense helps to automate workflows from Galaxy, make biologists unfamiliar with designing workflows to use the Galaxy interface, and quickly obtain analysis results from HTS sequencing projects. It uses Galaxy as a workflow management software and bioblend API to remotely manage the data upload, the workflow launching as well as the results analysis. Visualization of end-results is an important component, and is subject to further developments.

References

[1] Goecks J, Nekrutenko A, Taylor J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010 Aug 25;11(8):R86.

[2] Django core team. *URL* http://www.djangoproject.com, 2011.

[3] C. Sloggett, N. Goonasekera, and E. Afgan; BioBlend: automating pipeline analyses within Galaxy and CloudMan, *Bioinformatics* (2013) 29 (13): 1685-1686.

[4] B.D. Ondov, N.H. Bergman and A.M. Phillippy, Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* 2011, 12:385.

Acknowledgments

- French Society of BioInformatics (SFBI) for fellowship grant.
- eBio platform (Institut Français de Bioinformatique) for bioinformatics support.
- CEA (Commissariat d'Energie Atomique) for financing Damien Correia.