

# Optimization of a Galaxy pipeline in the Magerit supercomputer for analysis of the sugarcane microbiome

Loïc BOURGEOIS

April 2015



# Previously...

## What has been done

- Enabling the Galaxy toolshed for Magerit (chroot)
- Enabling heavy files upload through ftp
- Implementing Nuria's mothur workflow on galaxy

## What I wanted to do

- Populate the Galaxy instance with more tools
- Allow users to access data libraries in a simple way
- Improve Mothur's Galaxy wrappers

# Things don't always go as expected



## Sample of errors

- List index out of range error on some tools installation
- Mothur was giving segmentation faults while trying to rerun some analysis
- Galaxy would not accept Amalia's email, for some reason

# Things don't always go as expected

**martenson:** loic\_bourg: how old is your Galaxy?

**loic\_bourg:** martenson : its nearly one year old (revision is c458a0fe1ba8+)

**martenson:** the file that throws the exception does not even exist in the distribution now

:(

# Galaxy reinstallation

## Pros

- Compatibility with newer tools
- Maintenance
- New fonctionnalits
- Debugging (narrow down errors)

## Cons

- Possible incompatibility with Magerit installation
- Could take time

# Main steps of the Galaxy reinstallation

- Back up the old version
- Install and configure the newest version of Galaxy
- The newer version of Galaxy implements a CLI job runner for SLURM which needed to be tweaked
- All the software installed along Galaxy did not need to be reinstalled

# In the end

- Reinstalling Galaxy was actually easier than what I expected
- It solved errors I was getting on some tools installation
- It will be easier to keep the Galaxy instance up to date
- Mothur does not give anymore the segmentation faults that I could not explain
- A mothur's pipeline has been successfully ran for Amalia's data

# New tools

- Preprocessing tools : Trimmomatic, FastX-Toolkit, etc.
- Tools for RNAseq analysis from the Tuxedo protocol[4]: TopHat, Cufflinks, etc.
- Plotting tools...



# Data libraries

The screenshot shows the Galaxy / GEBIOPAD web interface in a Mozilla Firefox browser. The page title is "Data Library 'Global Data Library'". There are buttons for "Add datasets", "Add folder", and "Library Actions". Below the title, there is a section "Import your data here" with a folder icon and a dropdown menu showing "sff". A table lists the datasets in the library.

<input type="checkbox"/>	Name	Message	Data type	Date uploaded	File size
<input type="checkbox"/>	97_Silva_111_rep_set_MOD.fasta		fasta	Thu Apr 16 09:03:42 2015 (UTC)	188.6 MB
<input type="checkbox"/>	B11.fastq		fastq	Thu Apr 16 09:03:43 2015 (UTC)	2.4 MB
<input type="checkbox"/>	B21.fastq		fastq	Thu Apr 16 09:03:43 2015 (UTC)	900.4 KB
<input type="checkbox"/>	FS11.fastq		fastq	Thu Apr 16 09:03:43 2015 (UTC)	3.1 MB
<input type="checkbox"/>	FS21.fastq		fastq	Thu Apr 16 09:03:43 2015 (UTC)	1.1 MB
<input type="checkbox"/>	LS11.fastq		fastq	Thu Apr 16 09:03:42 2015 (UTC)	2.4 MB
<input type="checkbox"/>	LS21.fastq		fastq	Thu Apr 16 09:03:42 2015 (UTC)	992.5 KB
<input type="checkbox"/>	PS11.fastq		fastq	Thu Apr 16 09:03:43 2015 (UTC)	2.7 MB

Figure : An example of a data library

# Why is it better?

## Benefits

- Users can store and organize their data easily in them [2]
- It is possible to reimport easily the data for future analysis
- Data can be linked into galaxy history, instead of copied
- It stores the files "physicly" in a single folder

## But...

Only Galaxy administrators can create them

# Workaround

After discussing with the Galaxy developers, it appears that the only way to avoid requiring an admin is to use the galaxy API.

## What is an API?

GUIs are nice and user friendly, but it is impossible to automatize tasks with them. An API (Application Programming Interface) is an interface allowing to access programmatically to a given software.

# The galaxy API and Bioblend

Galaxy offers two ways to do so :

## A classic REST API

- Classic HTTP requests
- Complete
- Heavier in term of scripting

## Bioblend[1][3]

- Python wrapper on the REST API
- Not all the REST API functions are implemented
- Easier to use

# Workaround

First solution that worked, but with some problems :

## One Data Library for everyone with the right permissions

- Each new user in Galaxy would need to have permissions added to the data library automatically which would require a script running in background
- If a user do mistakes while importing a file, and admin would be required to fix it

Second solution, which I adopted :

## Write a Galaxy Tool allowing the user to create a data library using the API

- The user has full control on his data library
- Lighter in term of ressources, as the script is launched only when needed by the user
- Was not working until yesterday because of a "bug"

# Data libraries for everyone

The screenshot shows the Galaxy / GEBIOPAD web interface. The browser title is "Galaxy / GEBIOPAD - Mozilla Firefox". The address bar shows "138.4.110.239/galaxy/". The navigation bar includes "Galaxy / GEBIOPAD", "Analyze Data", "Workflow", "Shared Data", "Visualization", "Admin", "Help", "User", and "Using 1.4 GB".

In the "Tools" sidebar, the "Get Data" section is expanded, and the tool "Create User Data Library Allows any user to create a data library" is highlighted with a red box. Other tools listed include "Upload File from your computer", "UCSC Main table browser", "UCSC Test table browser", "UCSC Archaea table browser", "EBI SRA ENA SRA", "Get Microbial Data", "BioMart Central server", "BioMart Test server", "CBI Rice Mart rice mart", "GrameneMart Central server", "modENCODE fly server", and "Flymine server".

The main content area features a green welcome message: "Welcome to the Galaxy instance of GEBIOPAD! If you have any questions or meet any problems, please contact: Loïc BOURGEOIS : [loic.bourg@gmail.com](mailto:loic.bourg@gmail.com), Nuria LOZANO GARCIA : [nuria.lozano@upm.es](mailto:nuria.lozano@upm.es)". Below this, a paragraph describes Galaxy as an open, web-based platform for data intensive biomedical research, supported by various institutions.

The "History" panel on the right shows a search for datasets and a list of recent datasets, including "Test\_1", "62: 16S\_ecoli.fasta", "49: FastQC on data 1: RawData", "48: FastQC on data 1: Webpage", "45: stability.files", "44: stability.batch", "43: mouse.time.design", "42: mouse.dpw.metadata", and "41: Mock\_S280\_1001\_R2\_001.fastq".

Figure : New tool allowing data library creation

# What's next...

- Compare QIIME and Mothur's pipelines [3]
- Setup maintenance scripts and cron jobs
- Enable the alignment visualisation Tool on Galaxy
- Test the Tuxedo protocol in the galaxy instance for RNA seq analysis

Thank you for your attention!



# References

- [1] Simone Leo, Luca Pireddu, Gianmauro Cuccuru, Luca Lianas, Nicola Soranzo, Enis Afgan, and Gianluigi Zanetti. BioBlend.objects: metacomputing with Galaxy. Bioinformatics (Oxford, England), 30(19):2816–2817, October 2014.
- [2] Luisa Maria Seoane, Omar Al-Massadi, Mary Lage, Carlos Dieguez, and Felipe F. Casanueva. Ghrelin: from a GH-secretagogue to the regulation of food intake, sleep and anxiety. Pediatric endocrinology reviews: PER, 1 Suppl 3:432–437, August 2004.
- [3] Clare Sloggett, Nuwan Goonasekera, and Enis Afgan. BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics (Oxford, England), 29(13):1685–1686, July 2013.
- [4] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols, 7(3):562–578, March 2012.