Building Galaxy Japan Community

Ryota Yamanaka[1], Koichi Ashizaki[2], Tazro Ohta[3], Hiroyuki Aburatani[1]

1. Genome Science Division, RCAST, The University of Tokyo, Tokyo
2. Laboratory for Disease Systems Modeling, IMS, RIKEN, Yokohama
3. Database Center for Life Science, ROIS, Kashiwa

BACKGROUND

What is Galaxy?

Galaxy is an **open source, web-based platform for data intensive biomedical researches**, and has been widely used for analysis of NGS data. On this platform, We can easily build analysis pipelines with combinations of public or custom tools, and those workflows can be shared and ran again later. (**Fig 1,2**)

Fig 1: Data processing with Galaxy

Regional Galaxy Communities

Recently, local Galaxy communities are growing . For example, **Galaxy UK community** has opened a web portal to share information about local events, servers and even job offers. Another example is that, **Galaxy France community** runs its website and mailing list in French language, and it makes easier for local engineers to join bioinformatics projects. Moreover, these local communities may be essential to manage community cloud environments, which is getting larger and replacing on-premises platforms in each institute. For instance, **the Genome Virtual Laboratory in Australia** manages nation-wide cloud computational resources, and it enables users to launch a pre-configured Galaxy on the cloud. To provide such common platforms, building a consensus of the community is important.

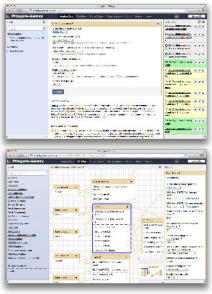

Fig 2: Galaxy's web interfaces

No Galaxy Japan Community Yet

However, in Japan, we have had no Galaxy community yet. Even though many Galaxy events are listed on Galaxy Wiki, none of them have taken place in Japan. As a consequence, we can hardly find contribution of Japanese institutes to Galaxy Community Conference, to Galaxy mailing lists, or to training network (**Fig 3**). Therefore, we have launched a website in Japanese language and start holding a monthly public workshop in Tokyo, so that we will find out potential demand of Galaxy-related knowledge in Japan. Also, we are engaged on a project called Pitagora-Galaxy, which **simplifies the installation process** of Galaxy and **ensures its reproducibility** on analyses workflows.

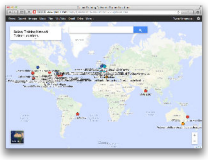

Fig 3: Galaxy Training Network

PROJECT OVERVIEW

Pitagora-Galaxy Project

Pitagora-Galaxy (http://www.pitagora-galaxy.org/) proposes an efficient method to share data analysis workflows for biomedical researchers and improve reproducibility of data analyses using Galaxy, **virtualization** and a **community website**. To make effective use of wide-spreading NGS data, we aim to promote inter-laboratory knowledge sharing based on this platform (**Fig 4**).

Fig 4: Pitagora-Galaxy Portal

Virtualization & Community Website

We think there are still some challenges to fully achieve the goals of Galaxy, reusability and reproducibility of workflows.

- Galaxy **workflows and tools are not shared** among research institutes well.
- It is **not easy to keep maintaining or to reconstruct** the Galaxy environments where we run our workflows before.

To solve the problems described above, we are running a website for sharing users' know-how, and distributing a virtual environment where we configured Galaxy with selected workflows and tools. Now, you can immediately use our analysis workflows on the following three environments (**Fig 5**).

- Access to our **public server** for testing.
- Download the **virtual machine** to your own PC or server.
- Launch **AMI** (Amazon machine image) on AWS cloud.

Since Pitagora-Galaxy enables us to run the same workflows on any infrastructure and rebuild the environments in any time, we can quickly use Galaxy, and at the same time, ensure the reproducibility of the analyses.

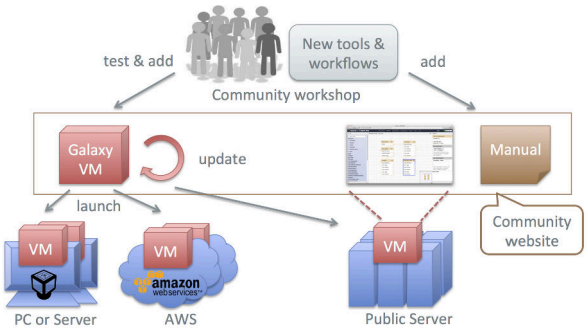

Fig 5: Project overview of Pitagora-Galaxy

PROGRESS

Beta Release

Pitagora-Galaxy has been already **released as beta versions** in virtual machine image and AMI. Although currently only two testing workflows (for RNA-seq analysis and ChIP-seq analysis) and related tools are installed, we are going to increase the collection of workflows and tools and **update the virtual machine image monthly**. We have already setup Pitagora-Galaxy and started using it **in three different laboratories** and have been collecting their feedbacks. The workflows designed by those institutes will be packaged in the future releases and their manuals will be uploaded on our community website (**Fig 6**).

Fig 6: Manuals for workflows

Developing User Assistance Tool

Pitagora-galaxy project focuses on efficient and full use of Galaxy and it does not aim to develop new functions on Galaxy or modify Galaxy. At the same time, some functions of Galaxy are too extensible but complicated for Galaxy beginners and some are still under development. In those cases, we develop **tools for providing workarounds** for Galaxy beginners. The following tools are the examples.

- Reference download
- Local file browser

The **reference download** tool enables users to easily download and setup reference data, such as FASTA genome files or BWA index files. Although Galaxy project provides Data Manager function, it is still a beta version, and also, the procedure to get the reference datasets is (extensible but) a little complicated for Galaxy beginners. (**Fig 7**)

The **local file browser** tool helps Galaxy administrators to build a web interface to list the datafiles on their server, and users can search datafiles on local servers and import them to Galaxy immediately. While Galaxy has Sample Tracking function, for those laboratories which have already been managing sequence data files, using our browser tool may be a simpler workaround. (**Fig 8**)

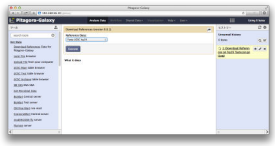

Fig 7: Reference download

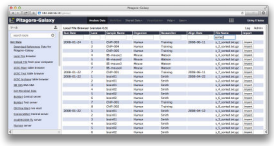

Fig 8: Local file browser

FUTURE PLAN

Container Version

We are also making Docker container image of Pitagora-Galaxy (but this is not yet released). Using Docker, we can also use the same Galaxy environment on Linux container, which has **performance merit** against virtual machine. Also, we can manage the recipe for creating Pitagora-Galaxy as dockerfile (= concept of Infrastructure as Code) and strengthen Pitagora-Galaxy's reproducibility. Since it is possible to create virtual machine and AMI from Docker image, we will be able to distribute the same Galaxy environment in all three different methods (**Fig 9**).

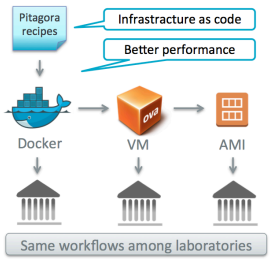

Fig 9: Creating images with recipes

Building Community

We are holding a **monthly workshop** for developers of Pitagora-Galaxy. Although we had only 7 researchers from 6 different institutes in the workshop of this month, we plan to publically announce about this workshop since the next month. This workshop is particularly important for designing Pitagora-Galaxy, because this will allow us to collect the actual analysis workflows share the exactly same workflows among those institutes.

Also, we are going to introduce Pitagora-Galaxy to potential users in domestic **conferences and training sessions**. Specially, Pitagora-Galaxy is useful to make hands-on session about the usage of Galaxy. Currently, one conference presentation and one training session are scheduled by the end of this year, and they are shown at our website.

SUMMARY

- We are launching **Galaxy Japan community** with a website and workshops.
- We are distributing a virtual machine for **easier installation** and **reproducibility** of Galaxy environment. This image is currently available in VM or AMI (or Docker in the future.)
- Our **user supporting tools are also included** in the image. Those tools are useful as temporal workarounds for easier use of Galaxy and they will promote introducing Galaxy.
- Another key advantage of using Pitagora-Galaxy is that **different institutes can use the same workflows** on the exactly same environment. Those workflows are added through community workshops.

Univ.Tokyo RCAST IMS RCAI-CGM DBCLS Database Center for Life Science