Diagnostyka molekularna (Galaxy Server) - platforma analizy danych z NGS

Table of Contents

Krótki opis usługi / Introduction

English follows Polish.

Usługa *"Diagnostyka molekularna (Galaxy Server) - platforma analizy danych z NGS"* jest instalacją znanego środowiska **Galaxy Server** **na potrzeby użytkowników PL-Grid**. Została wprowadzona na wniosek członków zespołów dziedzinowych LifeScience oraz Medycyna Spersonalizowana, ale może mieć zastosowanie także w innych dziedzinach pokrewnych. W największym skrócie jest to usługa pozwalająca przeprowadzać **zaawansowane, wielostopniowe analizy danych** pochodzących z tzw. sekwencjonowania następnej generacji (ang. **NGS - Next-Generation Sequencing**). W obrębie tej usługi można *zarządzać własnymi danymi* pochodzącymi z takich eksperymentów badawczych, *wykonywać złożone przetwarzanie* tych danych oraz *współdzielić uzyskane wyniki* z konkretnymi współużytkownikami platformy lub z wszystkimi członkami społeczności. Usługa posiada zaawansowane mechanizmy zapamiętywania obecnego stanu oraz tzw. *historii analiz*, dzięki czemu można łatwo odtworzyć ciąg zleconych analiz w celu np. ponowienia go dla innych danych bądź parametrów wejściowych. Wiele dalszych informacji odnośnie tej platformy można znaleźć na stronach projektu Galaxy.

Chociaż platforma Galaxy Server pozwala na szerokie zastosowanie w kilku dziedzinach bioinformatyki, opisana poniżej instalacja, przygotowana w ACK Cyfronet AGH, jest dostosowana **szczególnie do zastosowań typu NGS** (lista narzędzi znajduje się dalszej części podręcznika. Udostępniona wersja Galaxy Server wyposażona jest w szereg udogodnień związanych z osadzeniem usługi wewnątrz Infrastruktury PL-Grid.

Po krótkiej notce odnośnie procedury uzyskiwania konta PL-Grid, proponujemy wprowadzenie w postaci stosunkowo nieskomplikowanych ćwiczeń. Następie opisujemy pokrótce zainstalowane na platformie narzędzia i ich przeznaczenie, by przejść do zastosowań zaawansowanych. W tym rozdziale pokażemy, jak przeprowadzać pełne, istotne z punktu widzenia badań nad genetyką, ścieżki analiz (ang. *workflows*) z wykorzystaniem Galaxy Server. Pozostała część dokumentacji przygotowana została w języku angielskim.

The *"Molecular Diagnostic (Galaxy Server) - a platform for NGS analyses"* service is a deployment of popular environment called Galaxy Server for PL-Grid users. It was introduced on request from members of the PL-Grid LifeScience and Personalized Medicine community, but it may also be applied in other, similar domains of science. In short, this service allows to *perform advanced, multi-step analyses* of NGS (Next-Generation Sequencing) data. Using the platform, you may manage your own data sets of NGS experiments, run complex processing of that data and share acquired results with specific collaborators within the platform, or with every registered user. The service provides advanced mechanisms of current state and "history" *recording*, thanks to which you may easily trace back all the analyses you performed and, for instance, *reenact* these steps for different input data or parameters. For further information regarding the platform, please refer to the Galaxy project wiki pages.

Although the Galaxy Server platform is designed to support a wide range of applications from various areas of bioinformatics, the deployment of this service at ACC Cyfronet AGH (this documentation pertains specifically to that specific deployment) is especially suited to NGS-type of applications. We have decided that, to narrow the list of integrated tools down to a set hand-picked by our NGS experts (who are the co-authors of this documentation), will result in both better quality of service and easier learning curve. The list of these supported tools is provided in another section. Moreover, we have outfitted our own Galaxy Server deployment with a set of extensions helpful for any user of the PL-Grid Infrastructure - these are also described later, within following sections of this document.

After a short note regarding the PL-Grid registration procedure, we advise you to follow the introduction in a form of simple tutorial. Next, we provide a description of installed tools and their applications, which is followed by an advanced usage section. In this part we will show how you may perform full-fledged, scientifically-relevant genetic analyses (*workflows*) using the Galaxy Server.

Becoming a user (Aktywowanie usługi)

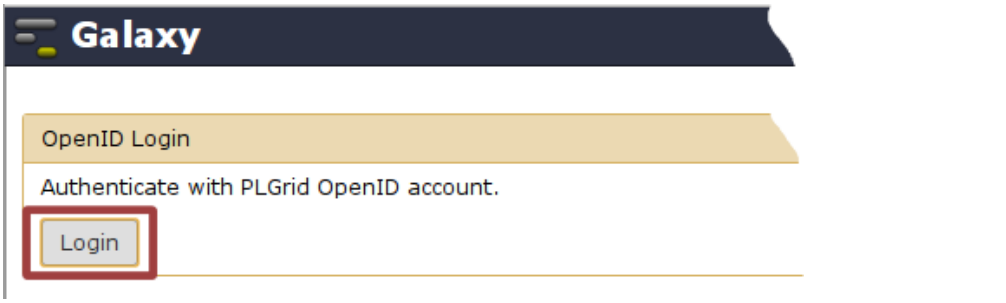
First of all, you have to know that the Galaxy Server at ACC Cyfronet AGH, as a part of the PL-Grid project, is *free of charge for any Polish scientist* . It is also free for *any foreign collaborator of Polish scientists* - so, if you don't work in Poland, but you have an ongoing research collaboration with one of Polish science institutes, you can freely use our service.

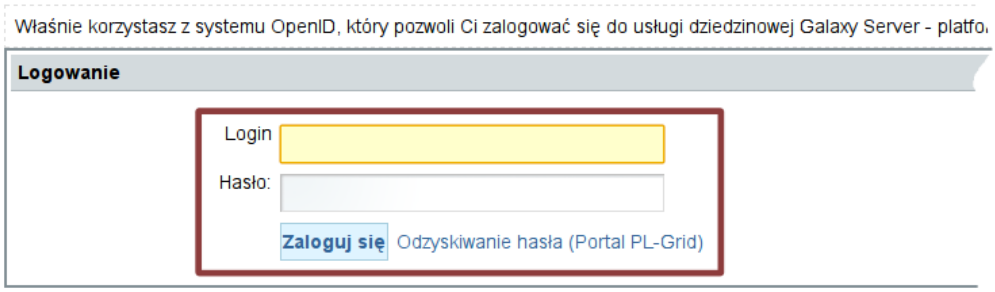
In order to do so, you have to register (and thus obtain your PL-Grid account) and setup your user settings, inside the **PL-Grid Portal** . We will go through that procedure step by step:

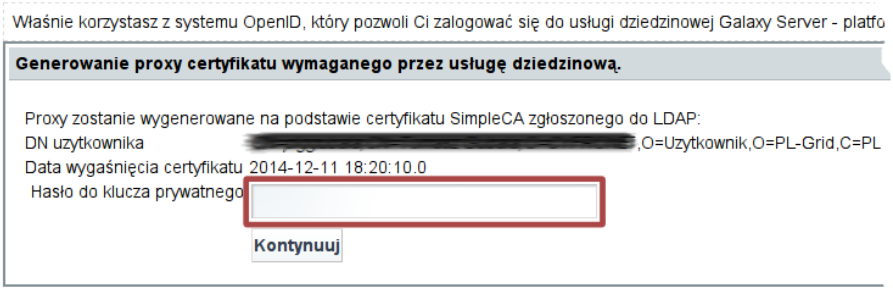
1. Registering as a PL-Grid user: please consult the Registration **steps section** of the **Podręcznik użytkownika** (PL) / **User Manual** (EN)
    a. if you are a foreigner, be prepared that you will need your Polish collaborator to confirm your identity - for the PL-Grid operatives to contact the correct person for the confirmation.

2. Requesting a certificate - again, please follow the **Aplikowanie, rejestracja i użycie certyfikatu** section of the manual for detailed instruction on how to automatically obtain your own PL-Grid certificate. Here, you will be asked to supply a certificate guard *passphrase* - please remember what you type as you are going to need it later on.
3. Requesting access to services; you will need to activate access to two services (you will find it inside Katalog aplikacji):
    - "*Dostęp do klastra ZEUS*"
    - "*Diagnostyka molekularna*"
4. Both services should be activated automatically for you, in a couple of minutes, and you will be informed of that fact by an e-mail. Then you are ready to navigate to **https://galaxy.plgrid.pl/** in order to start your first Galaxy session.
5. The service will use the infrastructure resources grant you have set as your default (it might be your personal resources grant).

Logging in to the service is fairly straightforward. See the following screenshots for instructions:







The first step is to click the "**Login**" button on the Galaxy Server welcome screen (see the first image). Afterwards, you are being redirected to the central PL-Grid authentication server (so-called OpenID server). Here, you supply your login and password credentials - the same ones you have used to access the PL-Grid Portal (the second image). One more step is needed, however - due to the extra rights delegation procedure, you need to verify as a Grid certificate owner - you will be asked to retype the certificate *passphare* you have used before to generate the certificate in the PL-Grid Portal (the last image). After a successful login, you will be redirected back to the Galaxy web platform and you should be ready to start.

> **Tired of the two-step login?**
>
> You can speed up the login procedure a little bit. During the certificate generation step of setting up your PL-Grid account, you should have received your certificate in the form of a file (you can recognize it by its ".p12" extension). You may wish to import that certificate file to tour own browser . That way your browser will deal with the second step of the procedure for you.

Short Introduction | Basics Tutorial (Pierwsze kroki)

As a form of hands-on introduction to our service, we propose to perform a set of exercises that will get you acquainted with the Galaxy Server platform and the mode it operates. Before we get going with the NGS analysis workflows, however, we need to get through some basic information regarding the variety of different file types you are going to encounter shortly. Getting to know them by heart is one of the elements that are rather required in the NGS data processing. Afterwards, we have prepared three basic, introductory workflows for you to follow. Apart from teaching you the fundamentals of the Galaxy Server, you will surely find them quite useful as crucial steps in your future, more advanced NGS workflows.

1. Most common file formats for Next Generation Analysis
2. Quality control pipeline: Trimming and filtering features using Flexbar
3. Short reads alignment to human reference genome using BWA.
4. Short reads alignment to mouse transcriptome using Tophat2.

5. Analysis of Bisulfite-Seq (BS-Seq) data using Bismark

After having completed the above tutorials on your own, you should be acquainted with some of the constituting elements of the Galaxy platform. You know how to use your history panel to track your progress (or failures, for that matter), you are able to import shared items. What is more important, you know your way through the multitude of NGS-related data formats and you are able to assess the quality of your sequencer output. The alignment process will help you in the coming tutorials of more advanced nature.

> **GSI Proxy problem**
>
> During prolonged sessions with Galaxy, you may experience the GSI proxy outage issue. When a job in your history finishes with error (it is presented against the red background) and the error message (provided after you click on the job name) reads: *GSI proxy is invalid - please relogin,* you need to logout and login again to renew your GSI proxy session. This is a security mechanism, imposed on us by the PL-Grid Infrastructure security policy, in order to prevent malicious users to "steal" your GSI proxy and be able to impersonate you. Hopefully, this will not be a serious nuisance to you.

List of Integrated Tools

Bowtie - short read aligner. Installed version: **Bowtie2 2.2.0**. Location in Galaxy tool menu: **NGS: Mapping**.
Its purpose is to align sequencing reads to long reference sequences. It should be user particularly to align reads of about 50 up to 100s or 1,000s of characters to relatively long (e.g. mammalian) genomes. Bowtie 2 supports gapped, local, and paired-end alignment modes. Provided by the John Hopkins University. Main page: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml.

SAMtools - very popular set of utilities. Installed version: **SAMtools 0.1.18**, implementing SAM specification version 1.4. Location in Galaxy tool menu: **NGS: SAM Tools**.
SAMtools is a set of utilities which main purpose is to manage genomic data stored in SAM, BAM and pileup format files. Available tools include: *sam-to-bam, sort, merge, filter, mpileup, rmdup, flagstat*. Maintained by the project contributors. Main page: http://samtools.sourceforge.net/.

Picard - a set of tools to manipulate SAM and BAM datasets. Installed version: **Picard 1.104**. Location in Galaxy tool menu: **NGS: Picard**.
Picard provides a lot of utilities, which are useful for sequencing data analysis. In our installation of Galaxy we wrap most of them - if you require other utilities or running modes, please contact us. Maintained by the project contributors. Main page: http://picard.sourceforge.net/.

TopHat - A spliced read mapper for RNA-Seq. Installed version: **TopHat2 2.0.10**. Location in Galaxy tool menu: **NGS: RNA Analysis**.
This tool is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using **Bowtie**, and then analyzes the mapping results to identify splice junctions between exons. Provided by Johns Hopkins University, University of California and Harvard University. Main page: http://tophat.cbcb.umd.edu/ .

Genome Analysis Toolkit (GATK) - a software package to analyse next-generation resequencing data. Installed version: **GATK 2.8-1**. Location in Galaxy tool menu: **NGS: GATK Tools (beta)** .
GATK offers a wide variety of tools, with a primary focus on variant discovery and genotyping as well as strong emphasis on data quality assurance. GATK is provided and maintained by the Broad Institute. Main page: http://www.broadinstitute.org/gatk/.

MACS - Model-based Analysis for ChIP-Seq for short reads sequencers such as Illumina. Installed version: **MACS 1.4.1**. Location in Galaxy tool menu: **NGS: Peak Calling** .
MACS empirically models the length of the sequenced ChIP fragments, which tends to be shorter than sonication or library construction size estimates, and uses it to improve the spatial resolution of predicted binding sites. MACS is provided by Xiaole Shirley Liu Lab at the Harvard School of Public Health. Main page: http://liulab.dfci.harvard.edu/MACS/.

Bismark - a software package to map bisulfite treated sequencing reads to a genome of interest and perform methylation calls. Installed version: 0.14.5. Location in Galaxy tool menu: NGS: Bi-seq analysis. Bismark is provided by Babraham Institute. Main page: http://www.bioinformatics.babraham.ac.uk/projects/bismark/.

The above are just the most notable toolkits and services integrated into Galaxy, but there are much more (e.g., Cufflinks, (p) BWA, FastQC, Bismark, bedtools, snpEff). In case you need another tool and wrapper available in our Galaxy, just give us a hint at PL-Grid HelpDesk (please, pick the Galaxy queue) and we'll be happy to assist (provided, the license of the tool allows us install it on our resources). What is also important - all these packages are installed on the Zeus computational cluster (the Galaxy server simply shares these packages with Zeus). Hence, in case you would rather use them with traditional command line, you may always log into the Zeus machine (zeus.cyfronet.pl - SSH protocol) and these packages should be available as modules to be loaded. Use the `module load` command to get access to them.

How to Acknowledge PL-Grid Support

In the case you have published some work that used the results obtained by using this service, please include the following sentence in the acknowledgements section of the publication:

PL-Grid Acknowledgments Form

This research was supported in part by PL-Grid Infrastructure.

Praca została wykonana z wykorzystaniem Infrastruktury PL-Grid.

Contact

In case of any trouble with our Galaxy installation or if you 'd like to ask us something, please contact us using **the HelpDesk system** . There, you may login using your usual PL-Grid credentials - then, please write us a message mentioning **"Galaxy"** somewhere inside (so it will be quickly redirected to our inboxes). You may use either Polish or English language.

In case you don't have a PL-Grid account yet (so it is not possible to log into the HelpDesk system), or you just want to make it simpler, you may send us an e-mail directly to: **helpdesk@plgrid.pl** .

Authors

This Galaxy installation is provided and maintained by ACC Cyfronet AGH and Klaster LifeScience Kraków. The experts from the following institutions provided workflows and tool wrappers for our Galaxy installation:

- Institute of Pharmacology, Polish Academy Sciences
- Omicron Laboratory - Enabling OMICs high-thROughput techNologies, Faculty of Medicine, Jagiellonian University Medical College
- National Research Institute of Animal Production

The Galaxy Server software is provided by Center for Comparative Genomics and Bioinformatics at Penn State , and the Biology and Mathematics and Computer Science departments at Emory University:

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. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. "Galaxy: a web-based genome analysis tool for experimentalists". *Current Protocols in Molecular Biology*. 2010 Jan; Chapter 19:Unit 19.10.1-21.
3. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. "Galaxy: a platform for interactive large-scale genome analysis." *Genome Research*. 2005 Oct; 15(10):1451-5.

List of contributors:

- Kacper Żukowski
- Maciej Filocha
- Marcin Piechota
- Piotr Radkowski
- Tadeusz Szymocha
- Tomasz Gubala
- Tomasz Waller

Adres usługi / Service URL address:

https://galaxy.plgrid.pl/