

Biologia: NGS Galaxy

Krótki opis usługi

Usługa jest przeznaczona dla biologów oraz bioinformatyków.

NGS Galaxy jest implementacją popularnego środowiska zarządzania zadaniami pozwalającego na uproszczone uruchamianie analiz bioinformatycznych na zasobach obliczeniowych PLGrid. W ramach usługi szczególny nacisk położony został na udostępnienie możliwie dużej liczby narzędzi związanych z analizą danych pochodzących z eksperymentów opartych o metody wysokoprzepustowego sekwencjonowania. Wykonywanie analiz z użyciem **Galaxy** oparte jest o wygodny interfejs dostępny w formie serwisu internetowego, pozwalającego na intuicyjne zarządzanie danymi, narzędziami oraz wynikami. Wbudowane moduły wizualizacji pozwalają na przejrzystą i efektywną analizę wyników.

Aktywowanie usługi

Aby skorzystać z usługi **NGS Galaxy**, należy mieć aktywne [konto w Infrastrukturze PLGrid](#) oraz aktywną [afiliację](#).

Do usługi **NGS Galaxy** może uzyskać dostęp każdy użytkownik PLGrid, który jest użytkownikiem usługi **Molecular Biology Data Analysis Toolkit**. W celu uzyskania dostępu do tej usługi należy po zalogowaniu się w [Portal PLGrid](#) w lewym menu wybrać zakładkę **Usługi**. Po przejściu do tej zakładki należy kliknąć zielony przycisk **Zarządzaj usługami**, znajdujący się w prawym górnym rogu. Spowoduje to przejście do [Katalogu Aplikacji i Usług](#) (KAiU), gdzie należy wyszukać usługę **Molecular Biology Data Analysis Toolkit**, a następnie o nią aplikować.

Przyznanie usługi następuje automatycznie. Gdy usługa zostanie przyznana, pojawi się na liście usług w portalu PLGrid ze statusem „active”. Status usługi będzie również widoczny w KAiU jako *"Status użytkownika: Usługa aktywna — usługa dostępna dla użytkownika"*.

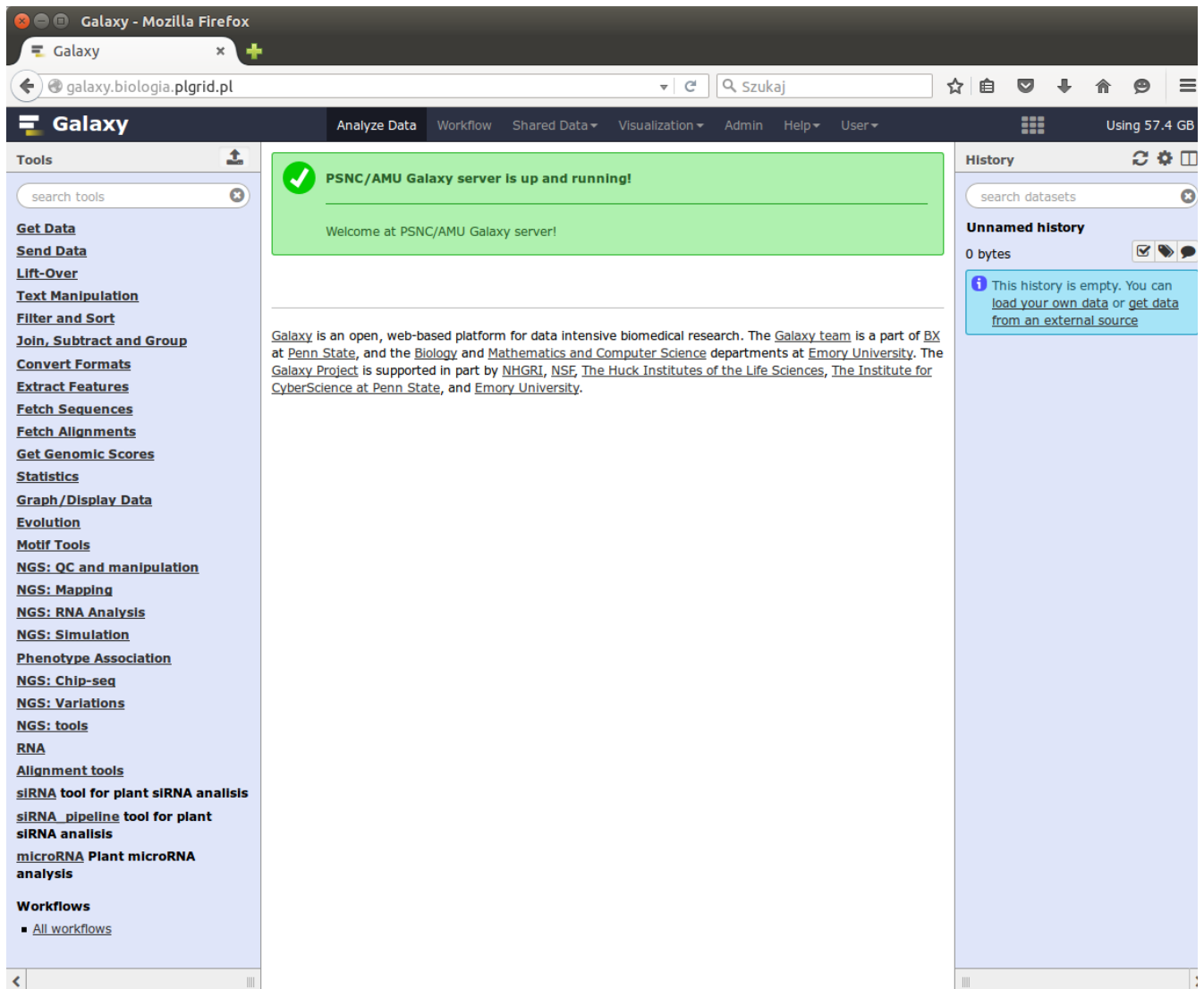
Pierwsze kroki

Uruchomienie usługi

Uruchomienie usługi **NGS Galaxy** możliwe jest na każdym komputerze wyposażonym w przeglądarkę internetową. Usługa dostępna jest poprzez portal **MBDAT** bądź pod adresem: <https://galaxy.biologia.plgrid.pl>. Podczas wczytywania strony wyświetlony zostanie monit o zalogowanie z użyciem konta PLGrid.

Organizacja interfejsu

Po zalogowaniu wyświetlona zostanie strona startowa usługi:



Interfejs składa się z trzech paneli: po lewej stronie zlokalizowana jest lista dostępnych narzędzi pogrupowanych według kategorii funkcjonalnych, środkowy panel służy do wyświetlania formularzy poszczególnych narzędzi oraz wizualizacji wyników, natomiast w prawym panelu znajduje się historia pracy w której wyświetlane są pliki wysłane do **Galaxy** oraz pliki wynikowe uzyskane w ramach prowadzonych analiz.

Przykładowy scenariusz użycia: analiza danych RNA-seq

1. Załaduj z dysku pliki z folderu **Sesja2**:

adrenal_1.fastq
adrenal_2.fastq
brain_1.fastq
brain_2.fastq

Aby to zrobić, wybierz narzędzie **Get Data -> Upload File**, następnie użyj przycisku **Choose local file**. Można zaznaczyć wszystkie pliki jednocześnie z wciśniętym klawiszem **Shift**. Następnie należy określić format plików. Dla plików **.fastq** w kolumnie **Type** wybierz **fastqsanger** (pliki fastq z kodowaniem jakości w skali Sanger). Naciśnij przycisk **Start**. Podczas transferu można zamknąć okno przyciskiem **Close** bez przerywania transferu. Aby do niego wrócić należy ponownie wybrać **Get Data -> Upload File**.

2. Pobieranie z UCSC Genome Browser plików z adnotacją genomową w formacie GTF. Narzędzie: **Get Data -> UCSC Main table browser**.

Aby pobrać adnotacje znanych genów o wysokiej wiarygodności dla genomu człowieka w wersji złożenia hg19 dla chromosomu 19 wybierz następujące opcje:

- clade: Mammal
- genome: Human
- assembly: Feb. 2009 (GRCh37/hg19)
- group: Genes and Gene Predictions
- track: UCSC Genes
- table: knownGene
- region: zaznacz opcję position i wpisz chr19

- output format: GTF – gene transfer format, zaznacz opcję Send output to Galaxy

Wciśnij przycisk **get output** i następnie potwierdź przyciskiem **Send query to Galaxy**. Plik z adnotacją pojawi się w panelu historii Galaxy.

3. Pobieranie plików z adnotacją genomową w formacie BED. Powtórz poprzedni punkt, tym razem w opcjach na stronie UCSC Genome Browser wybierając output format: **BED – browser extensible data**. Tym razem po wciśnięciu przycisku **get output**, pojawi się strona z dodatkowymi opcjami. Pozostaw ustawienia domyślne, i wciśnij przycisk **Send query to Galaxy**
4. Mapowanie odczytów do genomu referencyjnego. Narzędzie: **NGS: Mapping -> Bowtie2 – map reads against reference genome**. Wybierz następujące parametry:

- Is this single or paired library: Paired end
- Pierwszy plik FASTQ: adrenal_1.fastq (plik zawierający odczyty „forward” (zazwyczaj zawiera ‘_1’ albo ‘f’ w nazwie)
- Drugi plik FASTQ: adrenal_2.fastq (plik zawierający odczyty „reverse” (zazwyczaj zawiera ‘_2’ albo ‘r’ w nazwie)
- Will you select a reference genome from your history or use a built-in index?: Use a built-in genome index
- Select reference genome: Human (Homo sapiens) (b37): hg19 Canonical

Wciśnij przycisk **Execute**. Spróbuj powtórzyć procedurę dla plików **brain_1.fastq** i **brain_2.fastq**. Pamiętaj, że możesz wybrać tylko pliki o zgodnym formacie. Podczas importu plików, dla **brain** ustawiliśmy opcję **Auto-detect**. Klikając na nazwę pliku w historii rozwiną się informacje na temat pliku. Sprawdź, czy pliki **brain** mają format **fastqsanger**. Jeśli nie, wciśnij ikonę ołówka, a następnie w zakładce **Datatype** zmień typ pliku na **fastqsanger**.

5. Wizualizuj uzyskany plik BAM w przeglądarce genomowej. W panelu historii kliknij na nazwę pliku i poszukaj ikonki z wykresem, która po najechaniu na nią myszką powinna być adnotowana jako **Visualize in Trackster**. Po jej wciśnięciu zostaniesz przekierowany do okna wizualizacji. Aby utworzyć nową wizualizację, należy wpisać jej nazwę (dowolną) oraz wybrać genom **Human Feb. 2009 (GRCh37/hg19) (hg19)** oraz wcisnąć przycisk **Create**. Dane są dostępne dla rejonu: Chr19:3000000:3500000, dlatego z rozwijanego menu na górze strony należy wybrać chr19 i powiększyć fragment chromosomu na którym widoczne są zmapowane odczyty. Aby porównać rozłożenie odczytów z pozycjami genów, należy załadować plik GTF. W tym celu naciśnij ikonę **+** znajdującą się w prawym górnym rogu (**Add tracks**). Wybierz plik GTF pobrany wcześniej z UCSC Genome Browser. Zwróć uwagę, że po odpowiednim zawężeniu obserwowanego regionu zmienia się reprezentacja odczytów oraz sposób wyświetlania genów. Po zakończeniu, wciśnij ikonę **save** a następnie **close**. W ten sposób wizualizacja ta będzie dostępna później poprzez menu górnej belki **Visualization -> Saved visualizations**.
6. Analiza jakości mapowania. Narzędzie: pakiet RseQC. Wszystkie narzędzia z pakietu uruchomić należy na plikach BAM uzyskanych za pomocą mapowania programem Bowtie2 (Input file) oraz pliku BED zawierającym adnotację chromosomu 19 pobranym w punkcie 3 tej sesji (reference gene model). W ramach pakietu należy wybrać następujące narzędzia i opcje:

- **NGS: QC and manipulation -> Gene Body Coverage (BAM)**
- **NGS: QC and manipulation -> RPKM Saturation**. Opcje: Strand-specific?: Pair-End RNA-seq z pierwszym układem odczytów.
- **NGS: QC and manipulation -> Read Distribution**
- **NGS: QC and manipulation -> BAM/SAM Mapping Stats**

Uruchom narzędzie dla plików BAM uzyskanych z mapowania próbek adrenal oraz brain. Porównaj wyniki.

7. Testowanie statystyczne genów pod względem różnicowej ekspresji. Narzędzie: **NGS: RNA Analysis -> DESeq2**. Jako plik GFF podaj plik GTF z adnotacją chromosomu 19 pobrany w punkcie 2 tej sesji. Należy dodać po jednym replikacie dla każdej z grup. Jako pierwszy wskaż plik BAM będący wynikiem mapowania odczytów pochodzących z mózgu, a w ramach grupy drugiej wskaż plik BAM będący wynikiem mapowania odczytów z nadnerczy.
8. Przeanalizuj uzyskane wyniki zawierające listę genów ulegających różnicowej ekspresji.

Gdzie szukać dalszych informacji?

Szczegółowe informacje o użytkowaniu infrastruktury PLGrid znajdują się w [Podręczniku Użytkownika](#).

Szczegółowa dokumentacja platformy Galaxy znajduje się na stronie: <https://galaxyproject.org/>

Materiały szkoleniowe opracowane przez deweloperów platformy Galaxy dostępne są na stronie: <https://wiki.galaxyproject.org/Learn/Screencasts>

Informacje o usługach dziedzinowych *Biologia* dostępne są na stronie: <http://biologia.plgrid.pl/>

Uzyskanie informacji/helpdesk PLGrid: [dokumentacji o pomocy](#)