

Galaxy入門とワークフロー構築

ライフサイエンス統合データベースセンター
山口敦子

Galaxy とは

Galaxy はゲノムなどの生物学データを
対象としたウェブベースの解析環境

The screenshot shows the Galaxy web interface. The left sidebar contains a search bar and a list of analysis tools categorized under 'Search tools'. The main content area displays a search result for 'Galaxy', showing a thumbnail image of a robot and a brief description. On the right, there is a 'History' panel titled 'Unnamed history' containing several entries: '1: SGD POST on data_0', '2: Pasted Entry', '3: Pasted Entry', '4: BioM Custom track on data_1', '5: BL2-TC-GTF on data_1', and '6: UCSC Main on Human; KnownGene (genome)'. A red text overlay on the right side reads: 'ペンシルベニア州立大学を中心とした Galaxy team が開発' and provides the URL 'http://galaxyproject.org/'.

Galaxy を使ってできること

解析

用意されたツールを組み合わせて
データを解析する

ログイン
不要

保存

解析結果・履歴を保存
ツールの組み合わせをワークフローとして保存

共有

保存した解析結果・履歴・ワークフローを
ユーザが指定する共同研究者と共有

ログイン
必要

公開

保存した解析結果・履歴・ワークフローを
WWW上で公開

Galaxy を使うと…

解析

自分の実験データをいろんなツールを
いろんなパラメタを使って解析して、

保存

うまくいった解析結果はツールとパラメタの
組み合わせをワークフローに保存して、

共有

保存した解析結果やワークフローを
パラメタや解析履歴ごと共同研究者と
共有して、コメントもらって、

公開

論文にするときには、
サプライメンタルデータとして公開

ができます

Galaxy はどこで使える?

1. 公開サーバ:

<http://usegalaxy.org/>
<http://galaxy.dbcls.jp/>

など

2013年10月時点で、
49の公開サーバが世界中で
サービス中

The Galaxy Project's public server (usegalaxy.org, Main) can meet many needs, but it is not suitable for everything (see Choices for why) and cannot possibly scale to meet the entire world's needs.

Fortunately the Galaxy Community is helping out by installing Galaxy at their institutions and then making those installations either publicly available or open to their organizations or community.

This page lists such public or semi-public Galaxy servers.

To add your public Galaxy server to this list, please either just add it (hey, it's a wiki), or contact Galaxy Outreach <outreach AT galaxyproject DOT org>

Andromeda

- Links:
 - Andromeda
 - carver
 - Andromeda's implementation was the featured topic at [the Galaxy Dev Day 2011](#).

良い点: 無料, 自分でシステムをセットアップする必要がない
悪い点: 計算機資源が限られている(データ領域の制限など)

ちょっと試してみたい人,
大きくないデータしか扱わない人向け

Galaxy はどこで使える?

2. Amazon AWS:

Amazonが提供するクラウド上の仮想サーバサービス
(料金を支払って, 計算機資源が必要なときに必要なだけ
必要な性能の仮想サーバを借りることができるサービス)

仮想サーバを借りる際,
Galaxyが入ったものを選ぶことで,
自分だけのGalaxyを好きなだけ
好きなように使うことができる。

良い点: 自分でシステムを
セットアップする必要がない
悪い点: データを転送するたび,
計算するたびお金がかかる

お金がある人, 大きくないデータしか扱わない人向け

1 はじめの一歩: Galaxy のセットアップ手順
2 お持ちの Galaxy のセットアップ手順
3 AWS のアカウントを作成する
4 キーパスワードを作成する
5 Security Groups を作成する
6 EC2 のインスタンスを作成する
7 Galaxy のアカウントを作成する
8 初期化手順を確認しない
9 AMI を作成する
10 生成された AMI を Public にする

Galaxy はどこで使える?

3. お手持ちの計算機:

3-1. ソースプログラムからインストール

良い点: 自分の好きなように
カスタマイズ(ツールやDBの追加)が可能

悪い点: 計算機を用意する必要がある
計算機(Unix, Python, ネットワーク等)に関する知識が必要
計算機の知識がある人向け

3-2. お持ち帰り Galaxy の利用

VMWare player だけインストールすれば,
その上で DBCLS Galaxy と同じものを動かすことが可能.

良い点: 自分でシステムをセットアップする必要がない
悪い点: 計算機を用意する必要がある
計算機を用意できる人向け

Galaxy を使ってできること

解析

用意されたツールを組み合わせて
データを解析する

保存

解析結果・履歴を保存
ツールの組み合わせをワークフローとして保存

共有

保存した解析結果・履歴・ワークフローを
ユーザが指定する共同研究者と共有

公開

保存した解析結果・履歴・ワークフローを
WWW上で公開

解析

1. データをGalaxyに取り込む

1-1. データをGalaxyにアップロードする

1-2. データを外部のデータベースから取り込む

2. 取り込まれたデータを、 用意されているツールで解析する

ヒトの22番染色体の BED 形式のデータについて、
各区間がどの exon に対応するかを調べる

BED

ゲノム上の区間を表現するデータフォーマット
(UCSC によって定義されたフォーマット。
Galaxy 上ではツールの入出力形式として頻繁に使われる)

例: chr22 1000 5000 cloneA 960 + 1000 5000 255,0,0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 255,0,0 2 433,399, 0,3601

染色体番号
名前
ストランド
RGB色指定
ブロック開始位置リスト

開始座標
終了座標
スコア
特徴開始座標
特徴終了座標
ブロック数
ブロックサイズリスト

必須項目 オプション項目

データの取り込み

1.1 データをアップロードする

例: 手持ちの BED 形式のデータをアップロード

GetData → Upload File from your computer を選択

1.2 データを外部DBからGalaxyに送る

例: UCSC からヒトの22番染色体の exon データを

BED形式でGalaxyに送る

GetData → UCSC Main table browser を選択

<http://galaxy.dbcls.jp/> で実際にやってみます

ヒストリー

Galaxy の特徴の一つである、データ置き場システム

ユーザの取った
操作履歴が
データとして
ここに残る



ヒストリーの読み方

ユーザのアクション情報
(ツール、パラメタなど)

データをダウンロード

8: UCSC Main on Human: knownGene (chr22:1-51304566)
12,410 regions
format: bed, database: hg19
display at UCSC main
display in IGB Local Web
display at Ensembl Current
display at RVViewer main

1.Chrom 2.Start 3.End 4.Name

1.Chrom	2.Start	3.End	4.Name
chr22	16258185	16258303	uc002slh.1_c
chr22	16266928	16267095	uc002slh.1_c
chr22	16268136	16268181	uc002slh.1_c
chr22	16269872	16269943	uc002slh.1_c
chr22	16275206	16275277	uc002slh.1_c
chr22	16277747	16277885	uc002slh.1_c

データの詳細を見る
ヒストリーを編集する
ヒストリーを隠す
メタデータ
ビューワで見る
UCSC Genome Browser
Integrated Genome Browser
Ensembl のビューワ
VISTA Region Viewer

データの一部

データをツールで解析

おおまかな手順

- 1) ツールを選ぶ
- 2) ツールの入力として、ヒストリーにあるデータを選ぶ
必要なパラメタを入力する

例:

先程アップロードしたデータ

ヒトの22番染色体の BED 形式の手持ちのデータについて,

各区間がどの exon に対応するかを調べる

先程UCSCから取得したデータ

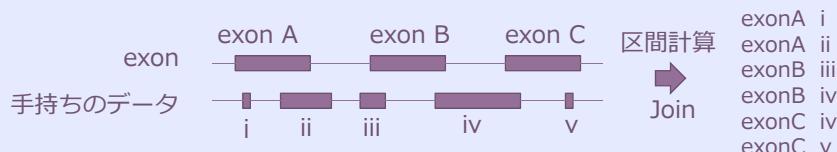
二つのデータの区間の重なりの対応関係を計算するツールを選べばよい

区間計算

区間を対象にした計算

Galaxy では Operate on Genomic Intervals に多数用意

区間の重なりの対応関係を計算するのは Join



<http://galaxy.dbcls.jp/> で実際にやってみます

Galaxy を使ってできること

解析

用意されたツールを組み合わせて
データを解析する

保存

解析結果・履歴を保存
ツールの組み合わせをワークフローとして保存

共有

保存した解析結果・履歴・ワークフローを
ユーザが指定する共同研究者と共有

公開

保存した解析結果・履歴・ワークフローを
WWW上で公開

保存

1. ヒストリーの保存
2. ワークフローの保存

いずれもログインが必要

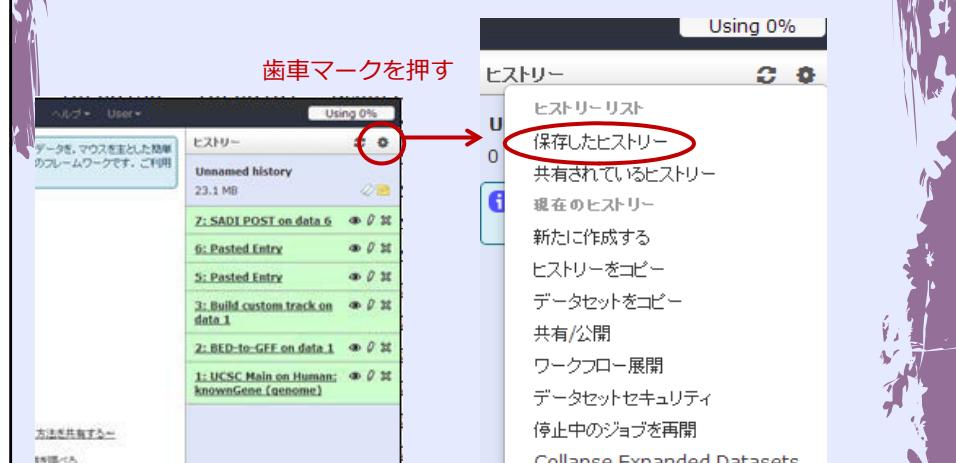
ログインするためには…

- Galaxy アカウントを作成してログイン
メールアドレス, Public nameを登録してアカウント作成
- OpenID アカウントでログイン
Google, Yahoo!, DBCLS OpenID などでログイン可能

ヒストリーの保存

ヒストリーはログインしていれば自動的に保存される

保存されたヒストリーを見る：



ワークフロー作成

ツール、パラメタの組み合わせを
ワークフローとして作成可能

二通りの作成方法:

- ・ヒストリーからワークフロー作成
- ・Workflow Canvas を使って
ワークフロー作成

作成されたワークフローは自動的に保存

ヒストリーからワークフロー作成

The screenshot shows the Galaxy web interface. On the left, the 'History' panel lists several history items, including 'Unnamed history' (23.1 MB) and various pasted entries and build custom tracks. A red arrow points from the text '歯車マークを押す' (Press the gear icon) to the gear icon in the top right corner of the 'Unnamed history' item. To the right of the history panel is a context menu with the following options:

- Using 0%
- ヒストリー
- ヒストリーリスト
- 保存したヒストリー
- 共有されているヒストリー
- 現在のヒストリー
- 新たに作成する
- ヒストリーをコピー
- データセットをコピー
- 共有/公開
- ワークフロー展開 (highlighted with a red oval)
- データセットセキュリティ
- 停止中のジョブを再開
- Collapse Expanded Datasets

<http://galaxy.dbcls.jp/> で実際にやってみます

ワークフローの管理

画面上部の「ワークフロー」をクリック

空のワークフローを作成

これまで作成したワークフローの一覧

Name	# of Steps
Test Workflow 2	3
Test Workflow	3
SNPdbの発がんlexonをchr22から探し	6
Unnamed workflow	0
Unnamed workflow	0
TogoWS	3
Workflow constructed from history 'Unnamed history'	5
Workflow constructed from history 'Unnamed history'	9
Unnamed workflow	0
Workflow constructed from history 'Unnamed history'	7

Workflows shared with you by others
No workflows have been shared with you.

Other options
Configure your workflow menu

Workflow Canvasで作成

ワークフローネームを押すとリストが表示

Edit を押すと Workflow Canvas が起動

空のワークフローから簡単なワークフローを作成してみます

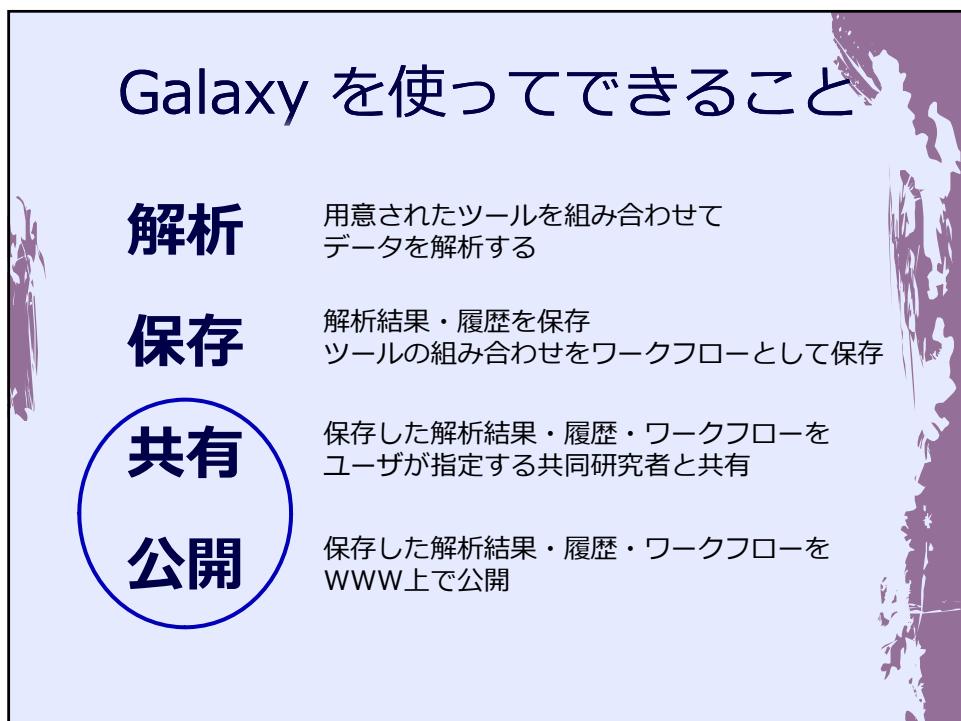
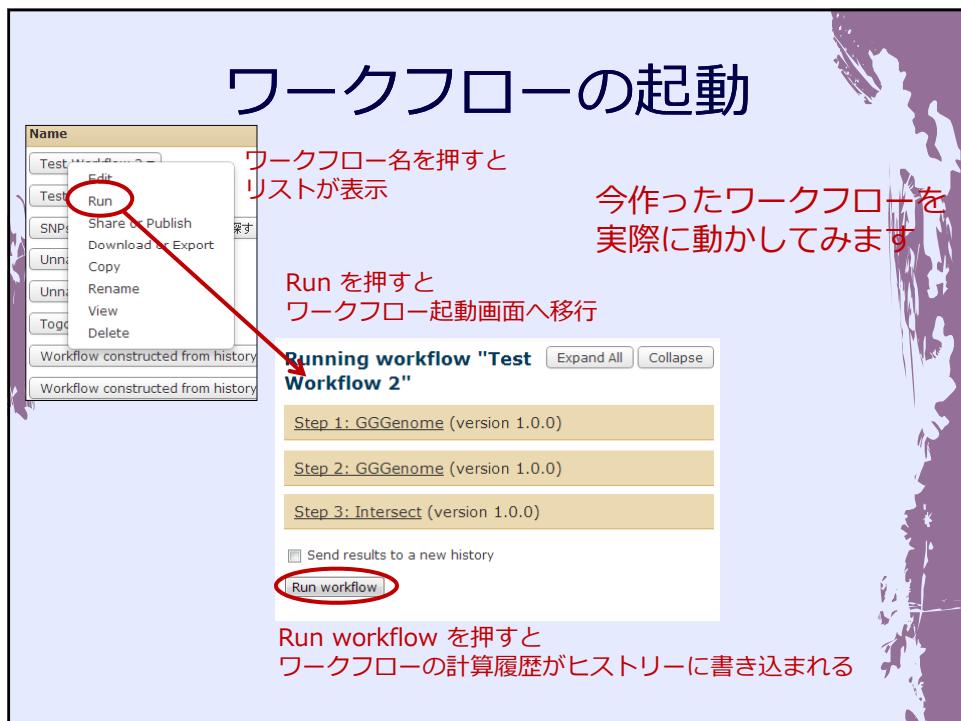
Workflow Canvas | Test Workflow 2

Workflow Attributes

- Name: Test Workflow 2
- Tags:
- Annotation / Notes:

```

graph TD
    A[GGGenome output (bed)] --> B[Intersect]
    B --> C[GGGenome output (bed)]
    
```



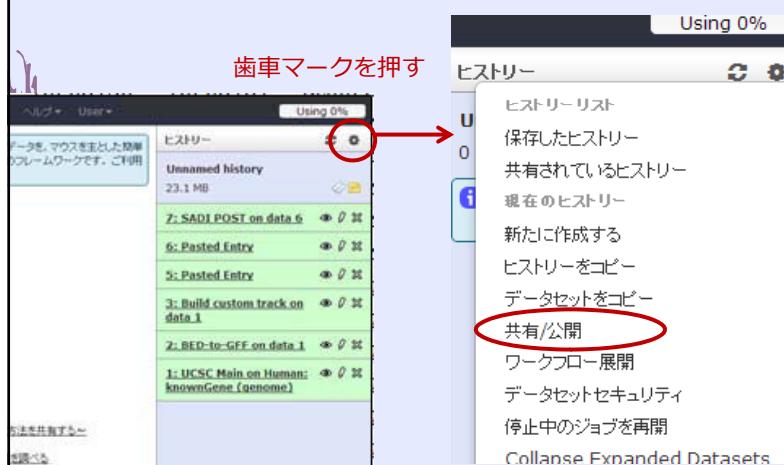
共有・公開

今回は

- 1.ヒストリーを共有・公開
- 2.ワークフローを共有・公開

の二つに絞ります

ヒストリーを共有・公開



ヒストリーを共有・公開

Share or Publish History 'Unnamed history'

Make History Accessible via Link and Publish It

This history is currently restricted so that only you and the users listed below can access it. You can:

- Make History Accessible via Link**
Generates a web link that you can share with other people so that they can view and import the history.
- Make History Accessible and Publish**
Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, where it is publicly listed and searchable.

Share History with Individual Users

You have not shared this history with any users.

Back to Histories List

さきほどのヒストリーを共有・公開してみます

公開

- リンクを知ってる人は誰でも見ることが可能に
- 公開ヒストリリストに載せる(=Galaxyユーザ全員から見られる)

共有

特定ユーザと共有

公開ヒストリーリスト

画面上部の「共有データ」をクリック

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
Unnamed history		t-hatakeyama	★★★★★		Feb 28, 2013
Unnamed history		atsuko	★★★★★		Jan 26, 2012
Unnamed history		taendo	★★★★★		Nov 30, 2011
LOAD Queries		nakao	★★★★★		Nov 28, 2011
Exon SNPs		atsuko	★★★★★		Jun 03, 2011
history for db migration test		thata	★★★★★		Feb 16, 2011
Unnamed history		fujiwara	★★★★★		Jan 11, 2011
upstream analysis		masakazu	★★★★★		Nov 16, 2010
TogoWS Test		atsuko	★★★★★		Aug 02, 2010

ワークフローを共有・公開

画面上部の「ワークフロー」をクリック

Share or Publish

Name	# of Steps
Test Workflow 2	3
Test Workflow	3
SNP	6
Unnamed workflow	0
Unnamed workflow	0
TogoWS	3
Workflow constructed from history 'Unnamed history'	5
Workflow constructed from history 'Unnamed history'	9
Unnamed workflow	0
Workflow constructed from history 'Unnamed history'	7

Workflows shared with you by others

No workflows have been shared with you.

Other options

Configure your workflow menu

ワークフローを共有・公開

Share or Publish Workflow 'Test Workflow 2'

Make Workflow Accessible via Link and Publish It

This workflow is currently restricted so that only you and the users listed below can access it. You can:

- Make Workflow Accessible via Link
- Make Workflow Accessible and Publish

Share Workflow with Individual Users

You have not shared this workflow with any users.

Share with a user

Back to Workflows List

公開

1. リンクを知ってる人は誰でも使うことが可能に
2. 公開ワークフローリストに載せる
(=Galaxyユーザ全員から使える)

共有

特定ユーザと共有

さきほどのワークフローを共有・公開してみます

公開ワークフローリスト

画面上部の
「共有データ」をクリック

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
test workflow		thata	★★★★★		Feb 17, 2011
SNPsの数が多いexonをchr22から探す		atsuko	★★★★★		Oct 20, 2010
TogoWS		atsuko	★★★★★		Jun 15, 2010

まとめ

Galaxy を使うと、共同研究に必要な解析・保存・共有・公開がヒストリーやワークフローを通じてスムーズにできる

利用目的や利用規模によって、Galaxy 公開サーバ、クラウド上の Galaxy、自力でインストール、お持ち帰り Galaxy の使い分けが望ましい

最後にちょっとだけ宣伝

DBCLS Galaxy

<http://galaxy.dbcls.jp/>

今日デモで利用した公開サーバ

メニューの日本語化, DBCLSで開発した便利なツールの追加, セマンティックウェブ対応をしたものを作成中

DBCLS Galaxy

An Integrated Environment

for Databases in Life Science

ライフサイエンスにおけるデータベース統合環境として

Galaxy の日本での普及を目指して



お持ち帰り Galaxy

<ftp://ftp.dbcls.jp/galaxy/>

上記 DBCLS Galaxy を VMWare 上で使えるようにしたもの。