

Galaxy (ゲノムビッグデータ解析) におけるオンデマンド機能の活用

株式会社アスケイド
那須野 淳

目次

生命科学系の研究分野において主にゲノムデータ解析用のプラットフォームとして使われている **Galaxy** について、オンデマンドクラウド機能を活用して解析環境を提供する事例をご紹介します。

1. データ解析プラットフォーム **Galaxy** とは
2. **Galaxy** によるNGSデータ解析について
3. クラスタ構成の **Galaxy** 環境
4. オンデマンドクラウド機能の活用

自己紹介

株式会社アスケイド

✦ 設立	2000年4月3日
✦ 本社	東京都新宿区
✦ 事業内容	ソフトウェアの受託開発 ソフトウェアに関するコンサルティング
✦ 主な開発実績	- 並列スーパーコンピュータOS開発 - 仮想化データセンター運用基盤開発 - 商用IaaS型クラウドサービス基盤開発 - 研究用アプリのインタークラウド環境適用開発

国立情報学研究所様におきまして、アカデミックインタークラウド基盤技術の研究開発プロジェクトに携わり、H27年度よりクラウド群連成基盤システムの開発／構築、関連研究の技術支援等を行っています。

1. データ解析プラットフォーム Galaxy とは

データ解析プラットフォーム Galaxy とは

- 生命科学系で利用されるWebベースのデータ解析システム
- 新型シーケンサーから得られる塩基配列データを扱うのが得意

The screenshot displays the Pitagora-Galaxy web interface. The browser address bar shows try.pitagora-galaxy.org/galaxy/. The main navigation bar includes links for **Analyze Data**, **Workflow**, **Shared Data**, **Visualization**, **Admin**, **Help**, and **User**. A status indicator shows **Using 0%**.

The left sidebar contains a **ツール** (Tools) section with a search bar containing **tophat**. Below the search bar, it lists **NGS: Mapping** with the entry **Tophat2 Gapped-read mapper for RNA-seq data**. A **Workflows** section below it lists **All workflows**.

The main content area is titled **Tophat2 (version 0.6)**. It contains the following configuration options:

- Is this library mate-paired?:** **Single-end**
- RNA-Seq FASTQ file:** **4: brain_2.fastq**
- Nucleotide-space:** Must have Sanger-scaled quality values with ASCII offset 33
- Use a built in reference genome or own from your history:** **Use a built-in genome**
Built-ins genomes were created using default options
- Select a reference genome:** **hg19**
If your genome of interest is not listed, contact the Galaxy team
- Tophat settings to use:** **Use Defaults**
You can use the default settings or set custom values for any of Tophat's parameters.
- Specify read group?:** **No**

An **Execute** button is located at the bottom of the configuration section.

Below the configuration section is the **Tophat Overview** section, which provides a description of the tool and its usage, including a citation: *Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36, 2013.*

The right sidebar contains a **ヒストリー** (History) section listing recent jobs. The jobs are numbered and include details about the tool used and the data files. The jobs are listed in descending order of age, with the most recent job at the top.

Jobs listed in the history:

- 25: Cufflinks on data 18: gene expression
- 22: Cufflinks on data 11: assembled transcript
- 21: Cufflinks on data 11: transcript expression
- 20: Cufflinks on data 11: gene expression
- 19: FastQC brain_2.fastq.html
- 18: Tophat2 on data 4 and data 3: accepted hits
- 17: Tophat2 on data 4 and data 3: splice junctions
- 16: Tophat2 on data 4 and data 3: deletions
- 15: Tophat2 on data 4 and data 3: insertions
- 14: Tophat2 on data 4 and data 3: align summary
- 13: FastQC brain_1.fastq.html
- 12: FastQC adrenal_2.fastq.html

Galaxy Workflow / Tools

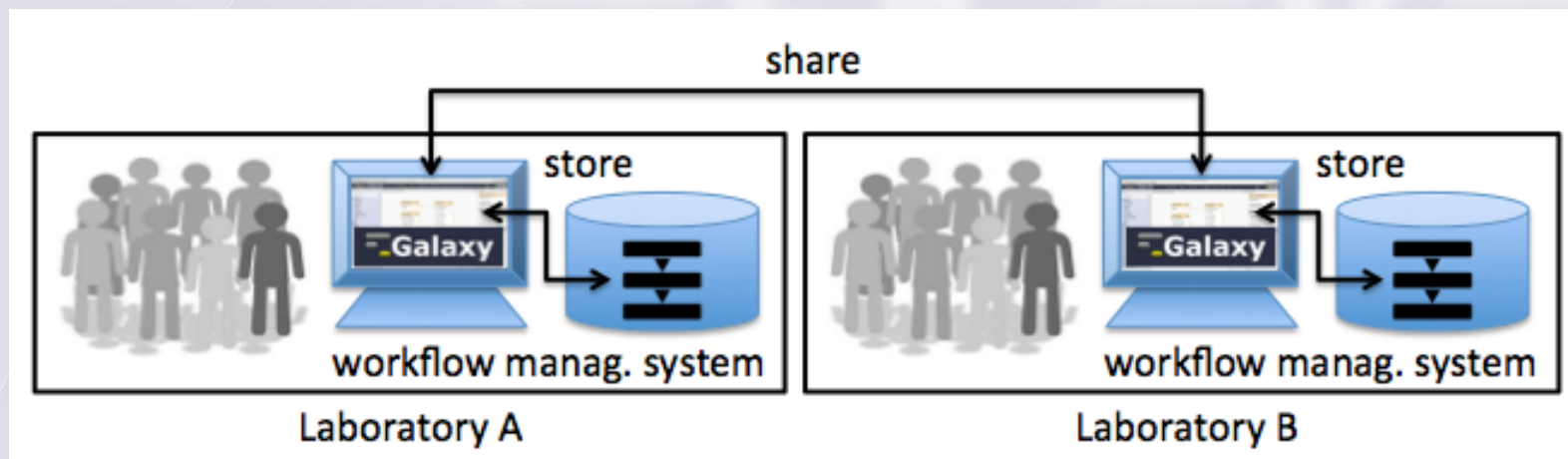
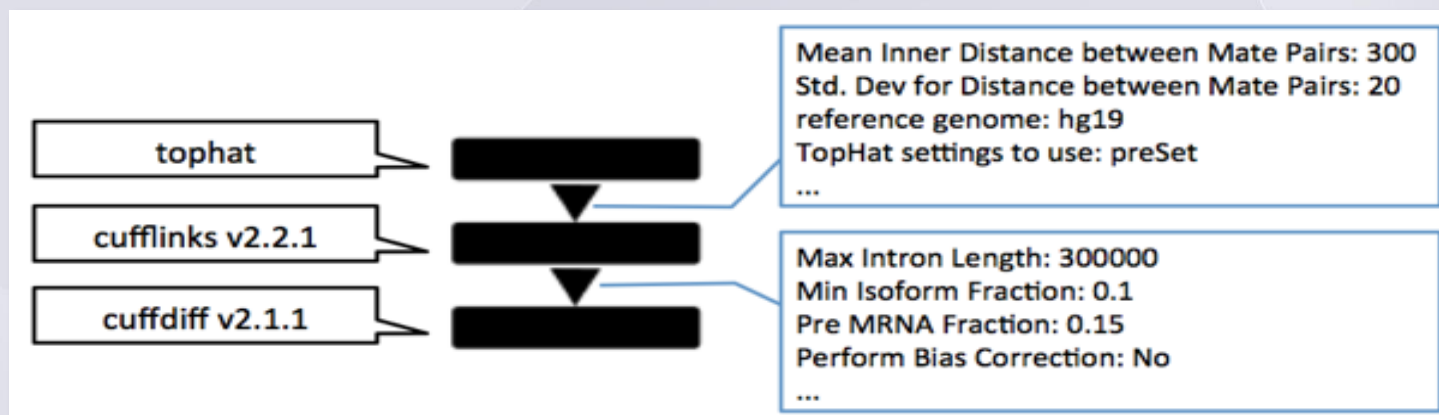
- 複数のツールを組み合わせた解析パイプライン = **ワークフロー**を構築し、Webブラウザ上で実行することが可能。

The screenshot displays the Pitagora-Galaxy web interface, specifically the Workflow Editor. The browser address bar shows the URL: `try.pitagora-galaxy.org/galaxy/workflow/editor?id=f2db41e1fa331b3e`. The interface is divided into several sections:

- Left Panel (Tools):** A sidebar with a search bar and a list of tool categories including Get Data, Send Data, Lift-Over, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Statistics, Graph/Display Data, Evolution, Motif Tools, NGS: QC and manipulation, NGS: Mapping, NGS: RNA Analysis, NGS: Simulation, Phenotype Association, NGS: Peak Calling, Operate on Genomic Intervals, Workflow control, and Inputs.
- Workflow Canvas:** The central workspace titled "Workflow Canvas | RNA-seq 01". It contains a grid of tool nodes connected by lines representing data flow. The tools visible include:
 - FastQC: Read QC (receiving input from an "Input dataset" node).
 - Tophat2 (receiving input from an "Input dataset" node and FastQC output).
 - Cufflinks (receiving input from a Tophat2 node).
 - Another Tophat2 node (receiving input from an "Input dataset" node and Cufflinks output).
- Right Panel (Details):** A panel titled "Details" showing the configuration for the selected "Tool: Tophat2". It includes:
 - Version: 0.6
 - Is this library mate-paired?: Paired-end
 - RNA-Seq FASTQ file, forward reads: Data input 'input1' (fastqsanger)
 - RNA-Seq FASTQ file, reverse reads: Data input 'input2' (fastqsanger)
 - Mean Inner Distance between Mate Pairs: 300
 - Std. Dev for Distance between Mate Pairs: 20
 - Report discordant pair alignments?: Yes
 - Use a built in reference genome or own from your history: Use a built-in genome
 - Select a reference genome: hg19
 - TopHat settings to use: Use Defaults
 - Specify read group?: No

実行履歴 / 入力データ / ワークフローの共有

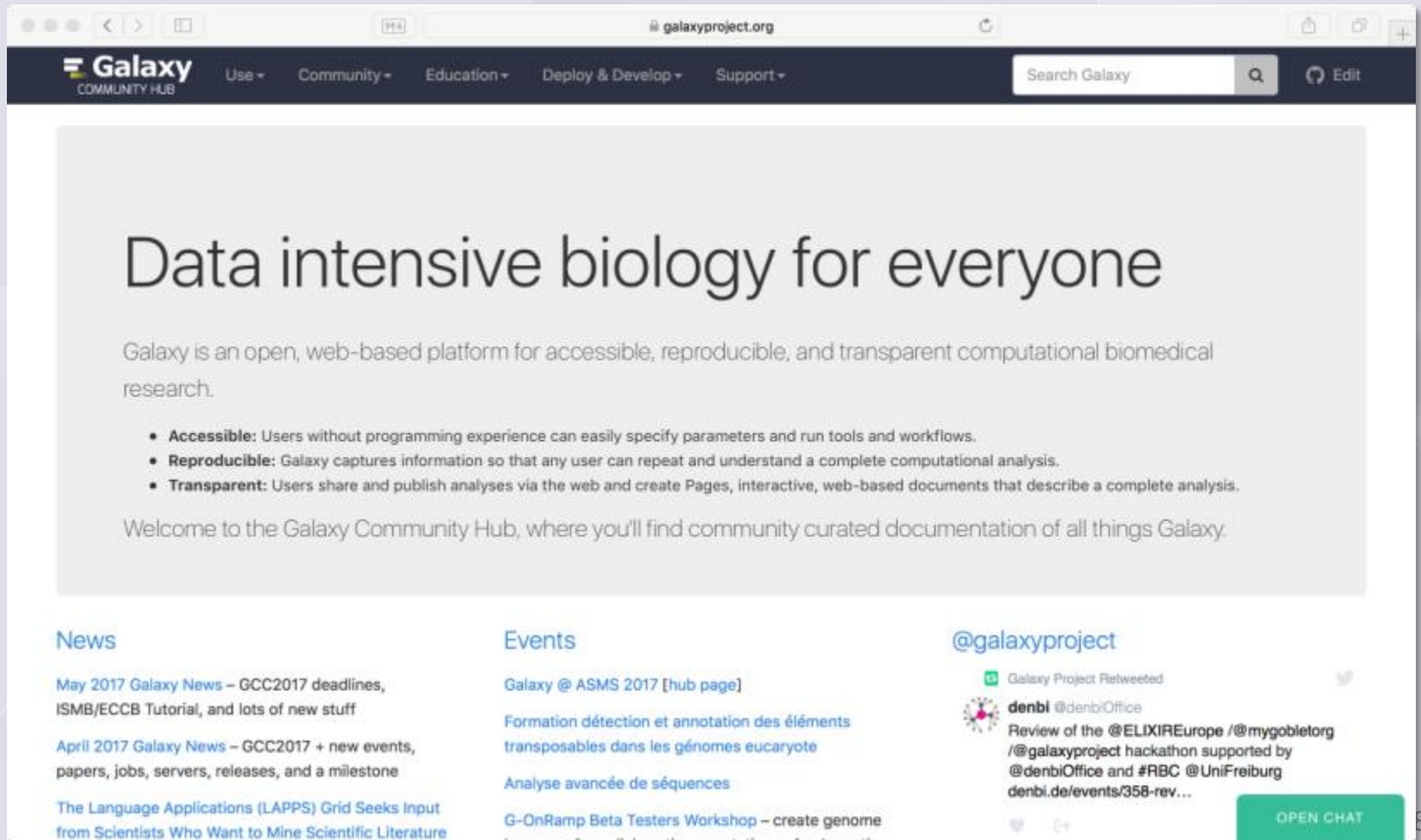
- 入力データや解析結果の関連を履歴管理することで同一処理の繰り返し、過去のデータ解析の再実行が容易
- ワークフロー、各データ、実行履歴を研究者間で受け渡し可能



開発体制・コミュニティ

■ The Galaxy Project

Penn Stateのメンバーを中心としたGalaxyチームにより開発されている。
Official web > galaxyproject.org



The screenshot shows the Galaxy Project website. The header features the Galaxy logo and navigation links: Use, Community, Education, Deploy & Develop, and Support. A search bar and an 'Edit' button are also present. The main content area has a large heading 'Data intensive biology for everyone' followed by a description of Galaxy as an open, web-based platform for accessible, reproducible, and transparent computational biomedical research. Below this, three bullet points highlight key features: Accessible, Reproducible, and Transparent. A welcome message follows. The footer is divided into three columns: News, Events, and a social media section for @galaxyproject. The News section lists recent updates from May and April 2017. The Events section lists upcoming workshops and seminars. The social media section shows a tweet from denbi (@denbiOffice) about a review of the ELIXIREurope hackathon.

Galaxy
COMMUNITY HUB

Use • Community • Education • Deploy & Develop • Support •

Search Galaxy

Edit

Data intensive biology for everyone

Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.

- **Accessible:** Users without programming experience can easily specify parameters and run tools and workflows.
- **Reproducible:** Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- **Transparent:** Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

Welcome to the Galaxy Community Hub, where you'll find community curated documentation of all things Galaxy.

News

May 2017 Galaxy News – GCC2017 deadlines, ISMB/ECCB Tutorial, and lots of new stuff

April 2017 Galaxy News – GCC2017 + new events, papers, jobs, servers, releases, and a milestone

The Language Applications (LAPPS) Grid Seeks Input from Scientists Who Want to Mine Scientific Literature

Events

Galaxy @ ASMS 2017 [hub page]

Formation détection et annotation des éléments transposables dans les génomes eucaryote

Analyse avancée de séquences

G-OnRamp Beta Testers Workshop – create genome

@galaxyproject

Galaxy Project Retweeted

denbi @denbiOffice

Review of the @ELIXIREurope /@mygobletorg /@galaxyproject hackathon supported by @denbiOffice and #RBC @UniFreiburg denbi.de/events/358-rev...

OPEN CHAT

Galaxyの利用形態

1. Public Serverを利用

世界中の様々な機関によって、GalaxyをインストールしたWebサーバが公開されている。アカウント申請することでサーバ上で解析を行う。

- galaxyproject.org/public-galaxy-servers
- usegalaxy.org
- galaxy.dbcls.jp
- try.pitagora-galaxy.org/galaxy

2. 自分の環境にGalaxyをインストールして利用

専用サーバを用意する。一般的なLinuxやmacOSが搭載されたマシンであればインストールはそれほど難しくない。

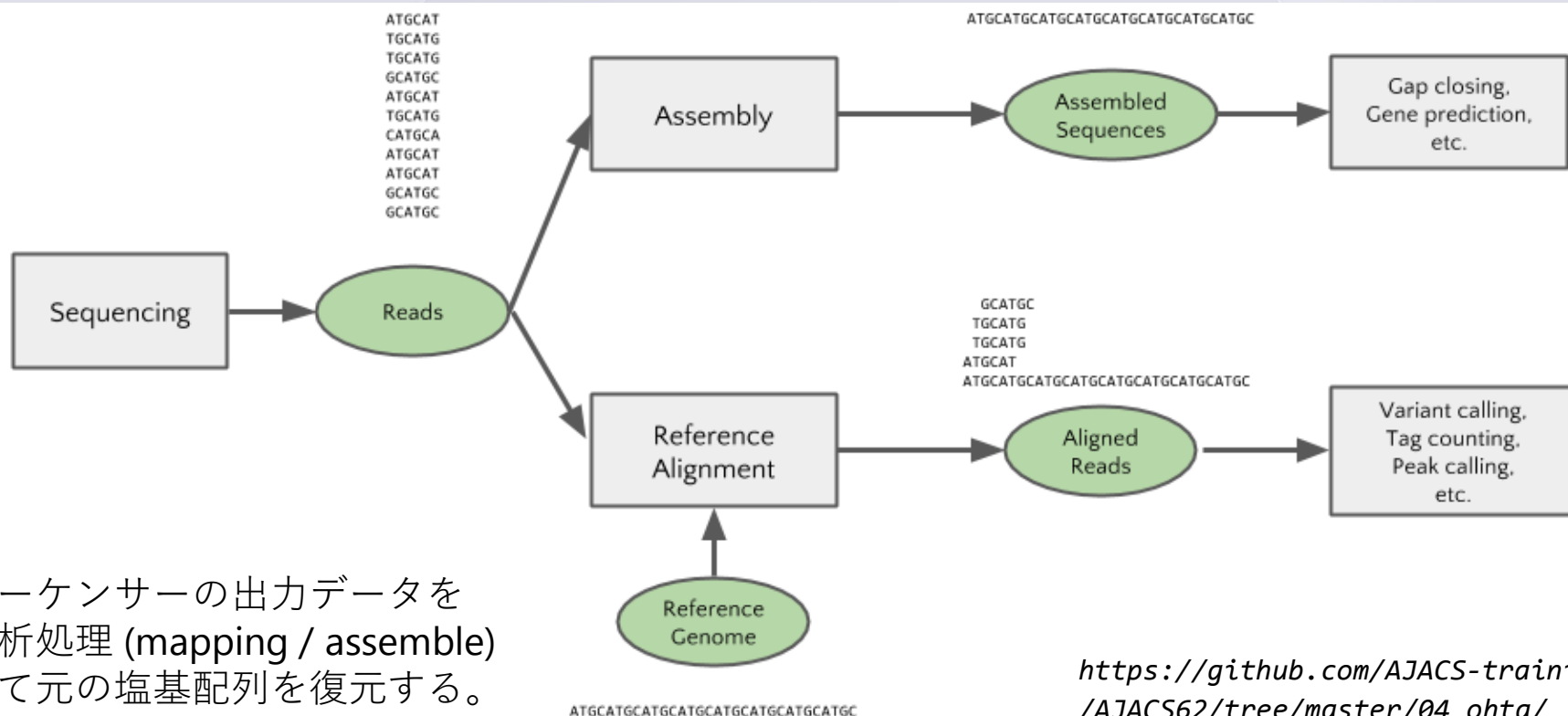
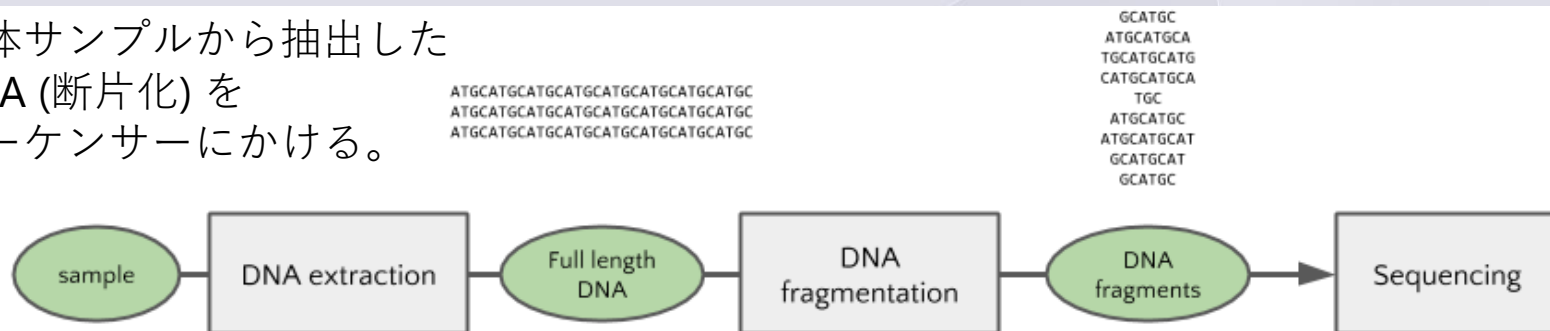
- インターネットにアップロードできないデータを解析したい
または自身で必要なツールをインストールしたい場合
- Galaxy Project のGithubからソースコードを取得してセットアップ
⇒コンパイル不要、run.sh のみで初期設定 & サーバ起動
galaxyproject.org/admin/get-galaxy

2. GalaxyによるNGSデータ解析

一般的なNGSデータ解析フロー

NGS: Next Generation Sequencer
(次世代シーケンサー)

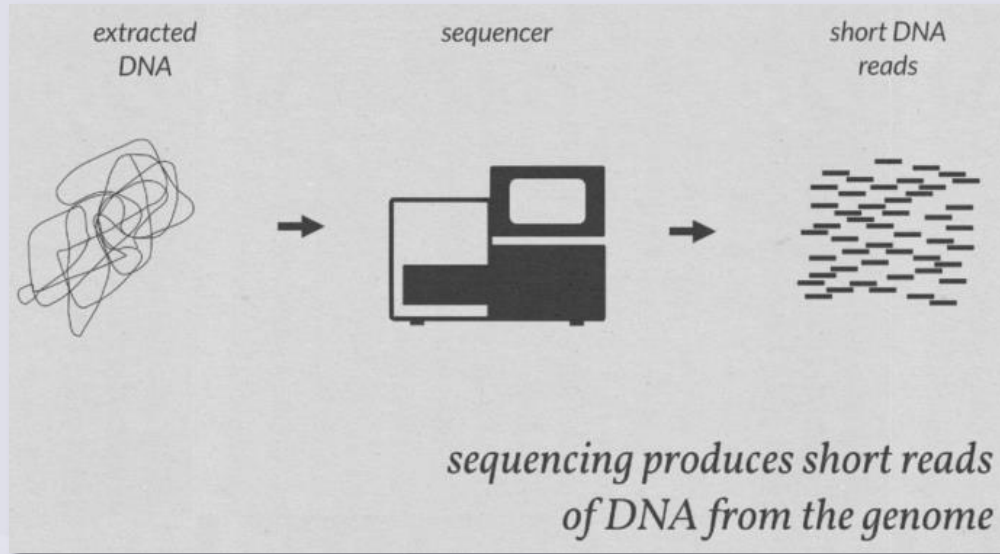
生体サンプルから抽出した
DNA (断片化) を
シーケンサーにかける。



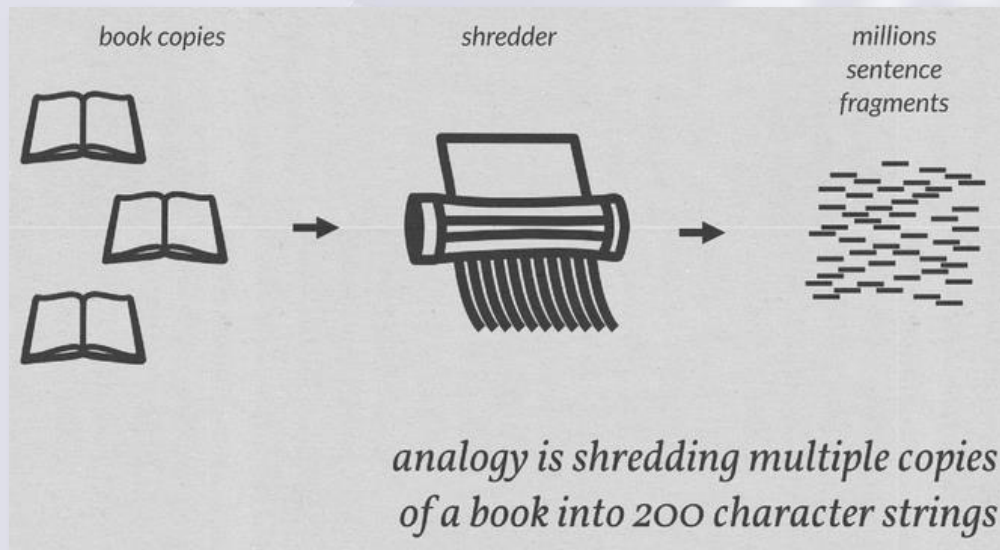
シーケンサーの出力データを
解析処理 (mapping / assemble)
して元の塩基配列を復元する。

https://github.com/AJACS-training/AJACS62/tree/master/04_ohta/

NGSの原理を簡単に説明すると…



生体サンプルをシーケンサーに入れると、短く断片化された塩基配列のリスト
＝フラットファイルのテキストデータで出力



DNAシーケンサーをシュレッダーに例えると…

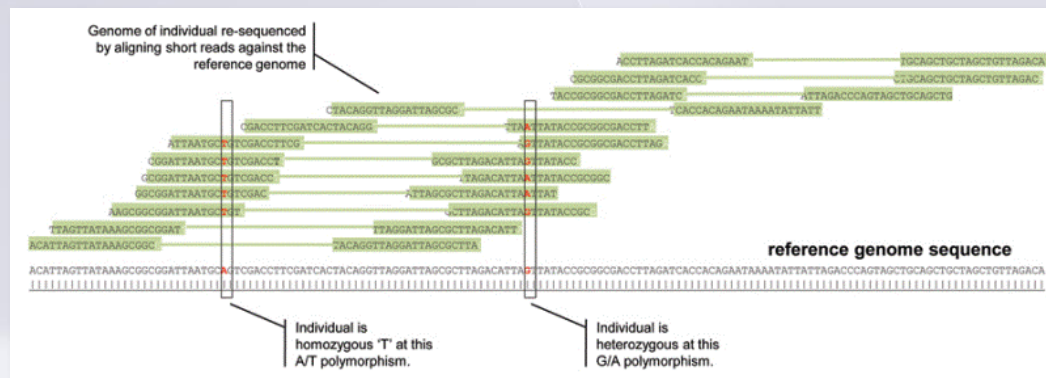
GalaxyによるNGSデータ解析

- データ処理の種類

リファレンスゲノムの有無によりシーケンス後のデータ処理が異なる。

- リファレンスアラインメント
(マッピング)

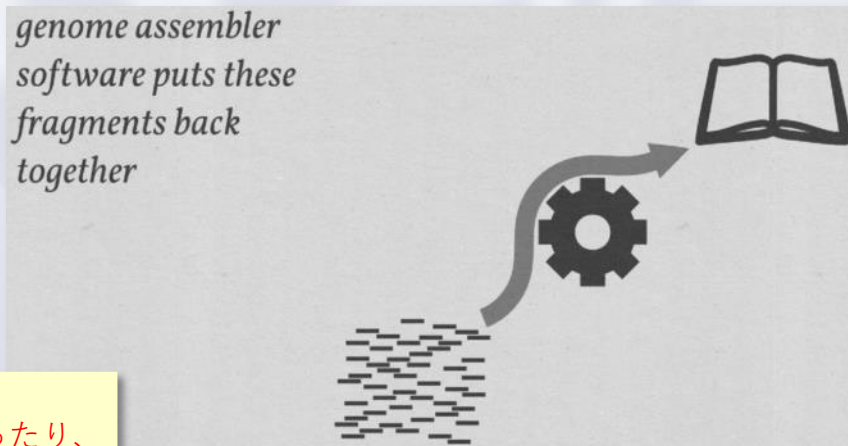
シーケンサから得られた
短い塩基配列（リード）を
リファレンスゲノムに沿って
並べて復元。



<http://www.historyofnimg.org.uk/mill-hill-essays/essays-yearly-volumes/2010-2/bringing-it-all-back-home-next-generation-sequencing-technology-and-you/>

- ゲノムアセンブル

リード同士の配列相同性を元に
短いリードを繋いで、
長いリードを組み立てる。
(アセンブル=本の復元)



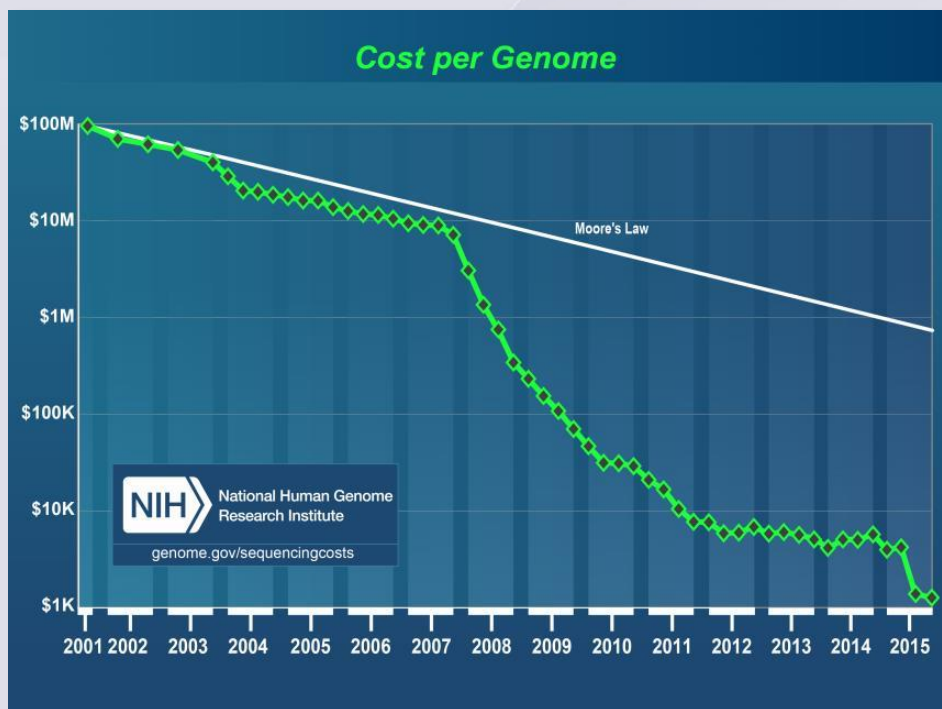
どちらの処理も膨大な計算資源を要する。
数GBのゲノムの新規アセンブルに20TBものメモリを使ったり、
データ処理に数日～数週間かかることもある。

<https://speakerdeck.com/michaelbarton/ranking-genome-assemblers-with-docker-containers-dockercon-eu-2014>

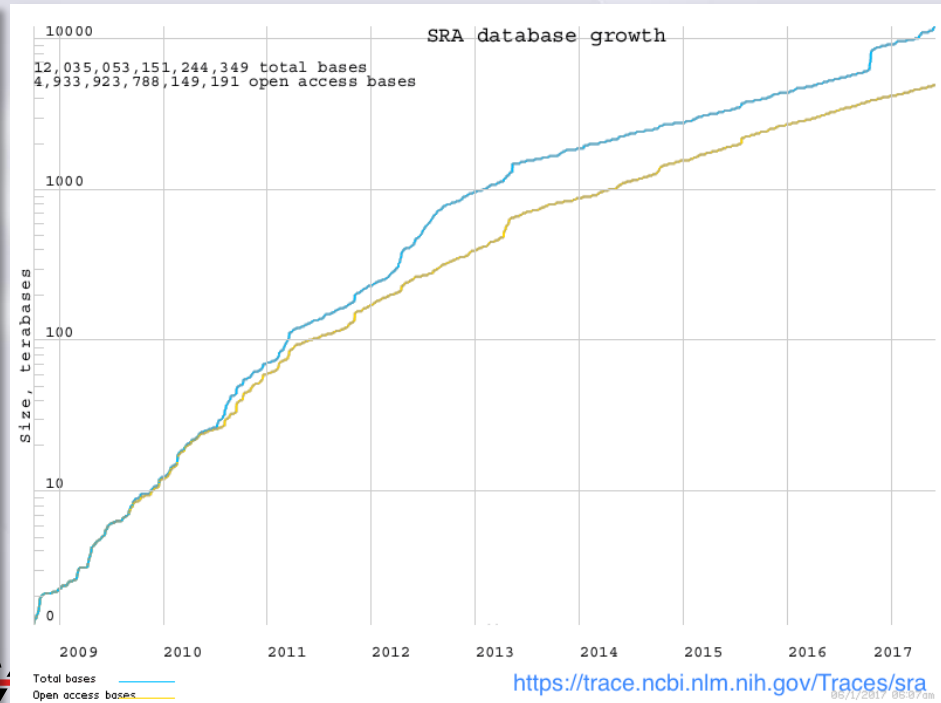
ゲノム情報解析の現状、現場の問題意識

- 実験機械の技術革新によりデータのサイズ・量が増大
- 解析目的によって異なるツール・アルゴリズムが使われる
- データ解析結果に対する再現性要求
- データ解析環境・インフラ構築に対する再現性要求

➤ ゲノム解析のコストは近年急激に低下



➤ 公共データリポジトリのデータサイズは伸び続ける



再現性に対する要求

■ データ解析（結果）の再現性

- ツールが変わると解析結果が変わる
- データ解析者（利用者）とツール開発者（実装者）は必ずしも同じではない
 - ⇒ 利用者がツールの挙動を完全に把握しているとは限らない
- 昨今は、ツールの「コンテナ化」が解決の一助となっている

■ データ解析環境（構築）の再現性

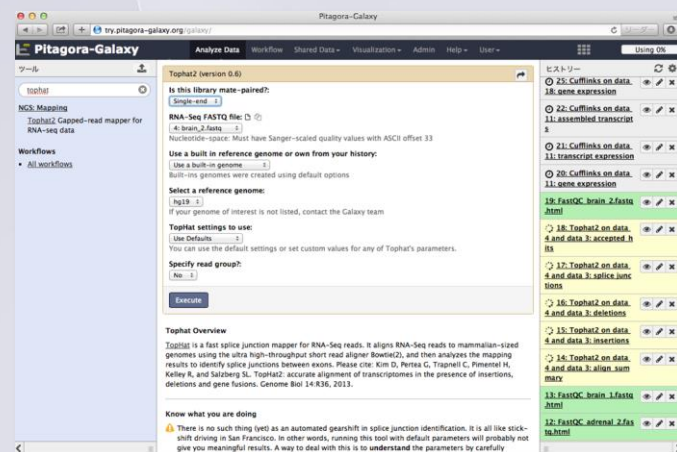
- 環境が変わるとツールが動かなくなる
 - 対象データが大きくなる / 増える
 - バッチ処理を大量のサンプルに対して実行する
- ツールの依存関係問題
- 実行環境を再構築（デプロイ）できるかどうかの問題
＝インフラの再現性 (Immutable Infrastructure)
 - 読み書きが高速で巨大な外部ストレージ接続
 - 大規模メモリ搭載システム
 - 分散実行ジョブスケジューリングシステム連携

3. クラスタ構成のGalaxy環境

Galaxy の基本アーキテクチャ

Frontend (Web UI)

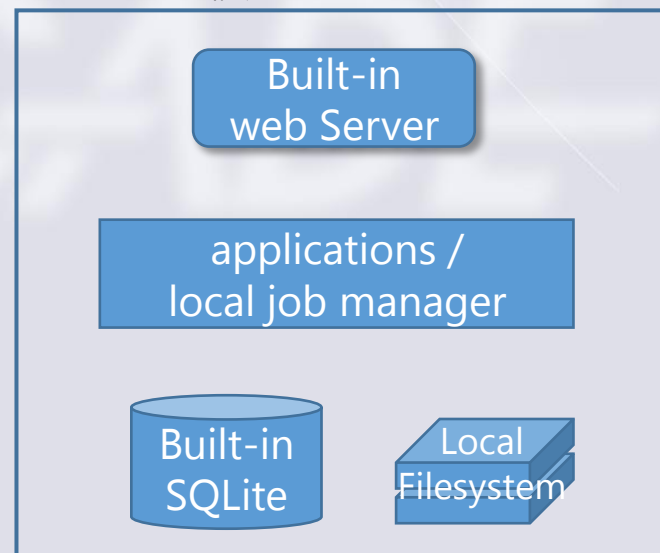
- シンプルで一貫性があるUI
 - JavaScript driven
 - Backbone for MVC
 - webpack & RequireJS for modules
- 主な対象ユーザは実験系研究者
⇒プログラミングやコマンドライン実行の知識・経験を前提としない



Backend (Python Server)

- プラグイン可能なインタフェースで構成
多様な技術に適応可能
 - Database Server (by SQLAlchemy)
 - Job Manager & Cluster
 - Storage
 - Web Proxy
- All-in-One構成での稼働も可
 - Built-in HTTP server (WSGI)
 - Built-in database (SQLite)
 - Local job manager

All-in-One 構成



クラスタ構成のGalaxy環境

基本構成ではローカルシステム内でジョブ実行されるが、Backend側のプラグインを設定することで、クラスタ上で実行するように構成することも可能。

分散リソーススケジューラ

- Univa Grid Engine
- Slurm
- HTCondor
- PBS Pro.
- その他、DRMAA I/F 対応製品
- ※ ジョブ実行プラグイン(Job Runner)
として独自実装してもよい

共有ファイルシステム

- NFS
 - GlusterFS
 - Lustre
- etc ...

DBサーバ

- PostgreSQL
- MySQL

Proxyサーバ (Load Balance)

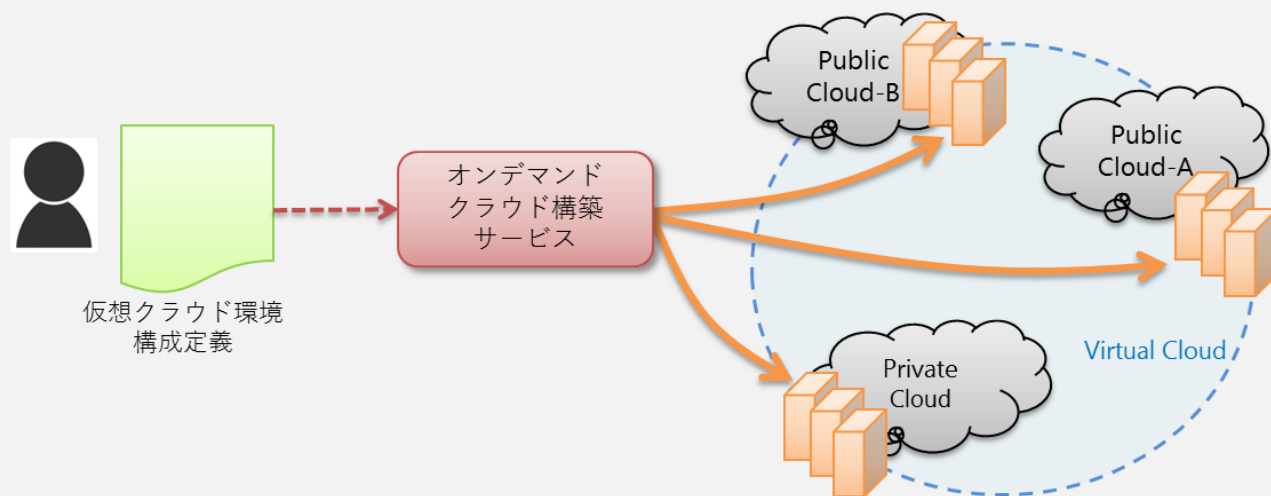
- Apache
 - Nginx
- etc ...

4. オンデマンドクラウド機能の活用

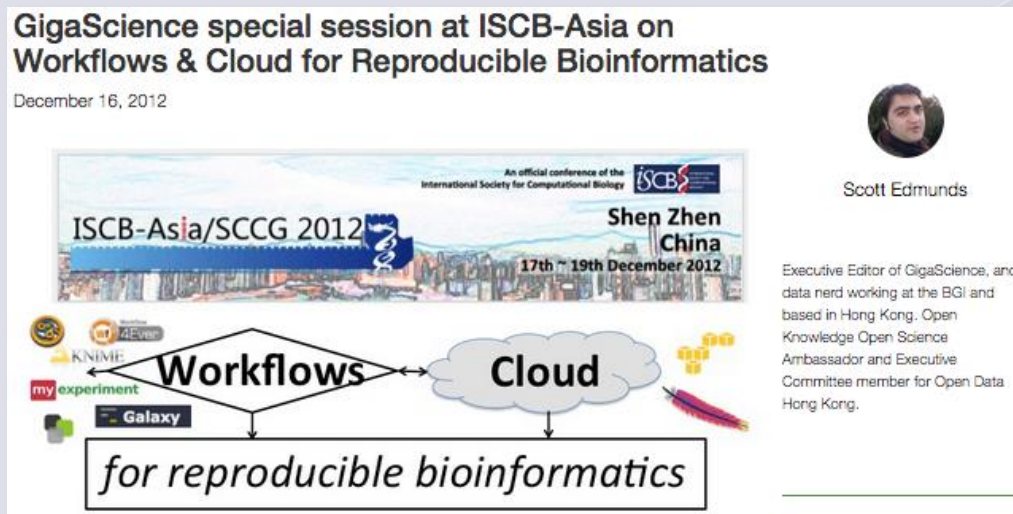
オンデマンドクラウド機能の活用

- 期待されるメリット：

- ① 論文掲載のデータ解析プロセス・結果の再現性確認（追試）
- ② 論文掲載の手法に基づく追加 / 発展 / 応用した新たな解析
- ③ 遠隔地に存在する大規模データを利用した解析
- ④ 高い機密性が要求されるデータ解析
- ⑤ 独立したデータ解析環境、プロジェクト別データ解析環境の作成



論文掲載のデータ解析プロセス・結果の再現性確認



<http://gigasciencejournal.com/blog/gigascience-special-session-at-iscb-asia-on-workflows-cloud-for-reproducible-bioinformatics/>

生命科学分野でのデータ解析の反復可能性と再現性確保は以前から重視されている。

論文の解析環境をパッケージングし、同一条件のGalaxyを仮想クラウド環境で作成する。

入力データ	データ取得先情報 = 公共データベースの登録IDなど
ツール	公開リポジトリで版数付けされたものでコンテナ化
ワークフロー	ツールの組み合わせ手順（パイプライン）をGalaxy Workflowとして定義
実行時パラメータ	各ツールの実行時スレッド数など、ワークフロー内の変更可能な設定値
計算資源	実行先クラウド、ノード数、マシンタイプ (CPU, Memory, Disk)

遠隔地にある大規模データ利用

- Galaxyで解析対象としたいデータは様々な組織が保持し、巨大なものもある
- DDBJ(日), EMBL(欧), GenBank(米) などの国際塩基配列データベースに蓄積されたデータを利用した研究
- ツールを実行する計算ノードと遠隔地の大規模データ間のデータアクセスが課題となる

実現方法：

- a. 大規模データの近くにGalaxyを立ち上げる
- b. ワークフローの一部を大規模データの近くで実行する

⇒ 計算資源やデータアクセス手段などに応じた仮想クラウド環境の構成を定義する。

高い機密性が要求されるデータの解析

- 「個人ゲノム情報」などプライバシーにかかわるデータ
or 独自性が高く公開したくないケース
 - ある範囲外に持ち出すことができないデータ
or アクセスが厳しく制限されているデータ
(通信路の秘匿性、厳密に設定されたアクセス権)
 - 暗号化して格納されているデータ
- ⇒ 「移動が許されないデータ」という特徴からは
「遠隔地の大規模データ利用」と同様の扱いとなる

独立した（プロジェクト別）データ解析環境作成

- 様々な構成に柔軟に適合するGalaxy環境を作成したい
- ゲノム情報解析に使われるソフトウェアは複雑化している
- セットアップ作業などにはインフラ構築に関する知識・経験が必要
- Wet(実験)系の研究者にとって敷居が高く、作業コストは無視できない

用途別の選択肢：

- a. 特定のワークフロー実行用のプレ・インストール環境
- b. 素の状態のGalaxy
- c. コマンドラインで実行するGalaxy

⇒ 仮想クラウド環境の構成定義を研究コミュニティ内で流通

仮想クラウド環境の構成定義（例）

Galaxyクラスタの構成定義におけるUnitおよびUnit間の依存関係

