CBI学会
October 24, 2023@東京

ゲノム科学におけるクラウド活用：独占型スパコン構築からデータ共有まで

Cloud Utilization in Genome Science: From Building Own Supercomputers to Sharing Data

重信　秀治
Shuji Shigenobu

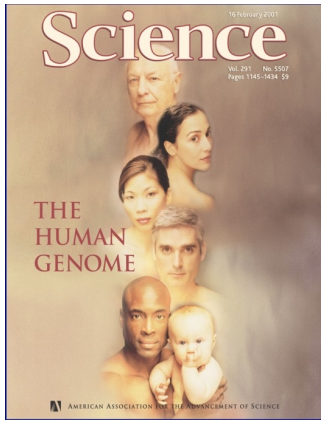
基礎生物学研究所 (NIBB)
National Institute for Basic Biology, NINS

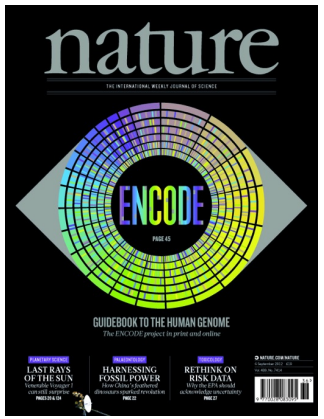
shige@nibb.ac.jp

本日のトピック

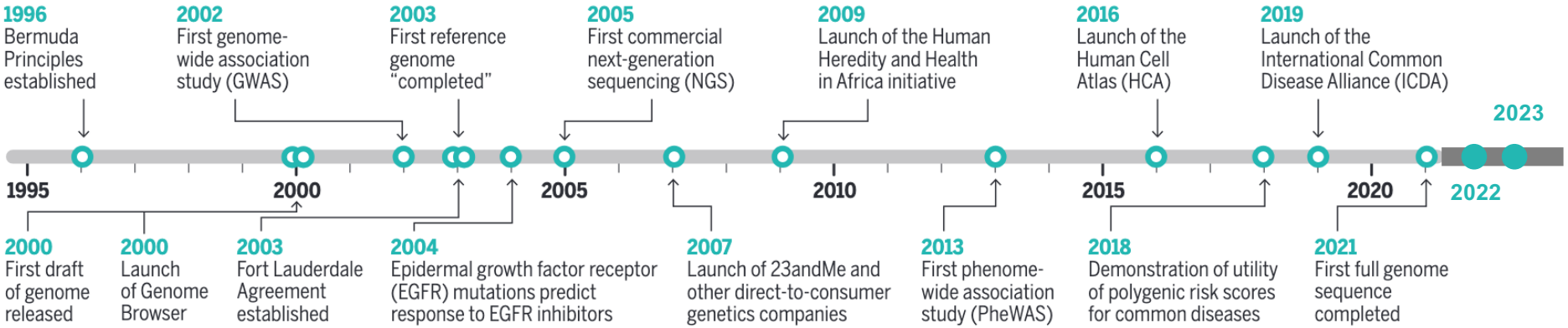
ゲノム科学分野、特に基礎研究、におけるクラウド活用

- 計算資源（HPC）の確保
- データ共有による共同研究促進





2000



(Rood et al. 2021 w/ modification)

急速に発展するゲノム科学

VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM

i5k

i5k: 5k insect genomes

2022

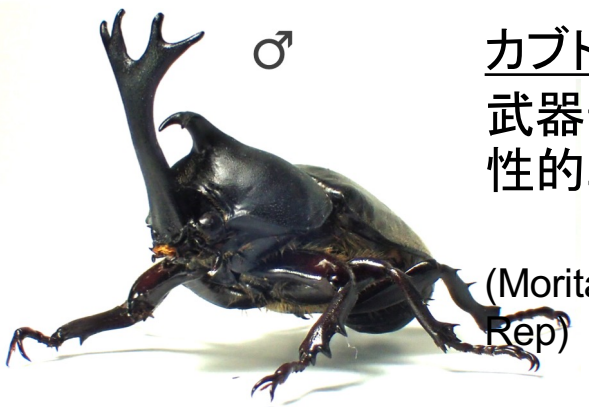
Y

2023

私が基礎生物学研究所で解読した昆虫ゲノム（抜粋）

エンドウヒゲナガアブラムシ
細胞内共生・環境適応・植物適応

（IAGC 2010 PLOS Biol, Shigenobu et al., 2000 Nature)



カブトムシ
武器形質の進化と性的二型

(Morita et al. 2023 Sci Rep)

ヤマトシロアリ
社会性の進化

(Shigenobu et al., 2022 PNAS)

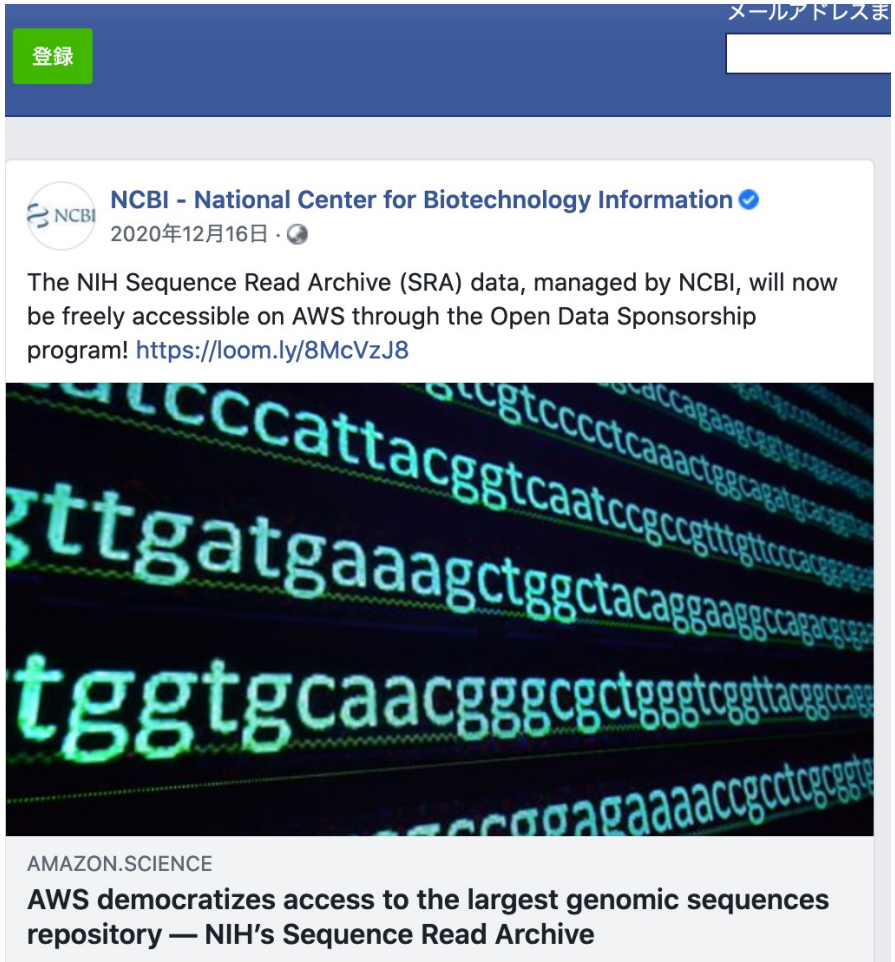
ヘイケボタル
発光機能の獲得と進化

(Fallon et al., 2018 eLife)

ますますビッグデータ化するゲノム生物学
どうする　コンピュータリソース？

ゲノム研究分野におけるクラウド化の動向

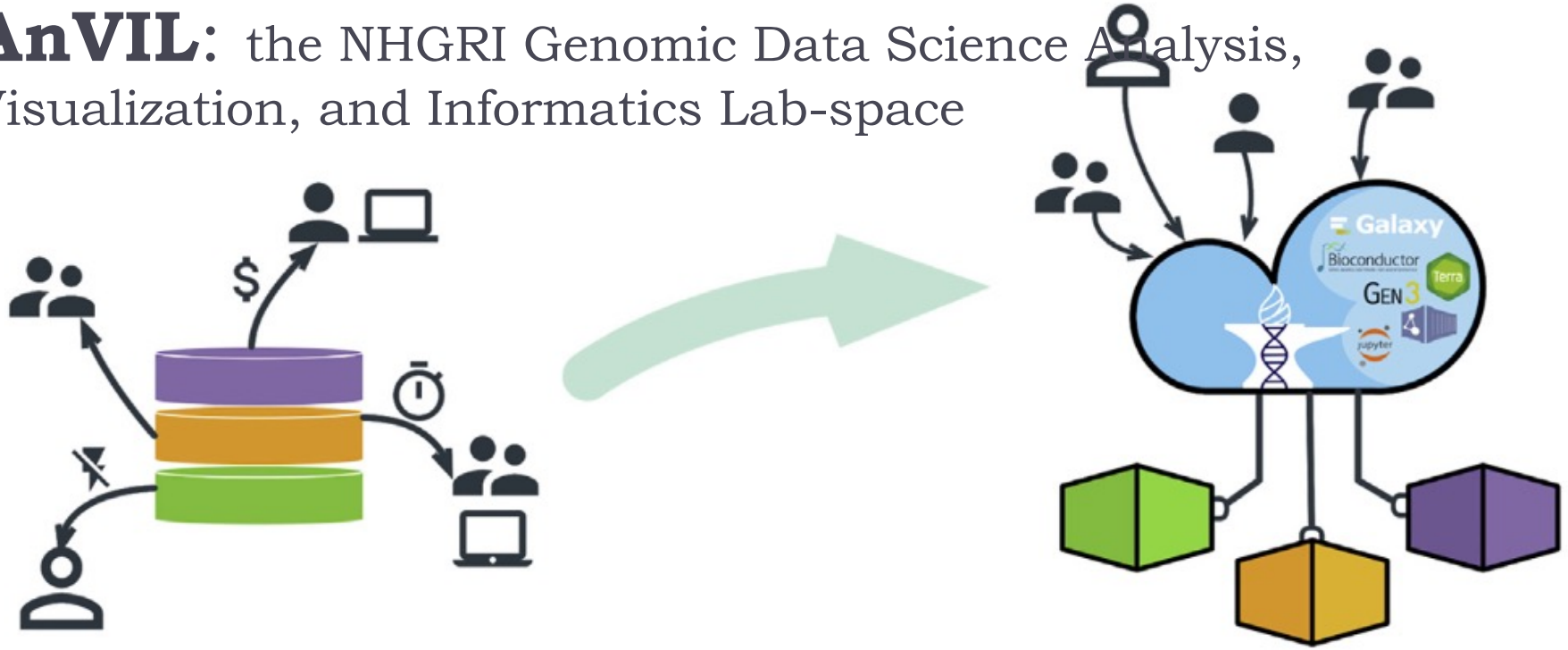
米国ではクラウド化の流れ



Facebook内ニュース 2020/12/16

- NCBIの次世代シーケンスデータ「SRA」はAmazon AWS上で提供される。
- SRA data:
  - 40 petabytes now
  - Will double every 1-1.5 year
- Saved in S3 storage service on AWS. Open access.
- NIHはSRA以外もクラウド化推進の姿勢

AnVIL: the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space



- AnVIL platform is NHGRI-supported data commons running on the Google Cloud Platform (GCP).
- Terra: analysis platform, Gen3: data search and artificial cohort creation; Dockstore : Docker-based genomic analysis tools and workflows.
- Support interactive analysis tools: Jupyter notebooks, Bioconductor, RStudio, and Galaxy.

https://anvilproject.org/

Schatz *et al. Cell Genom* (2022)

AWS Offers Genomics Analysis Platform

Amazon Genomics CLI

AWS HealthOmics (Amazon Omics)

日本のアカデミアにおけるクラウド動向

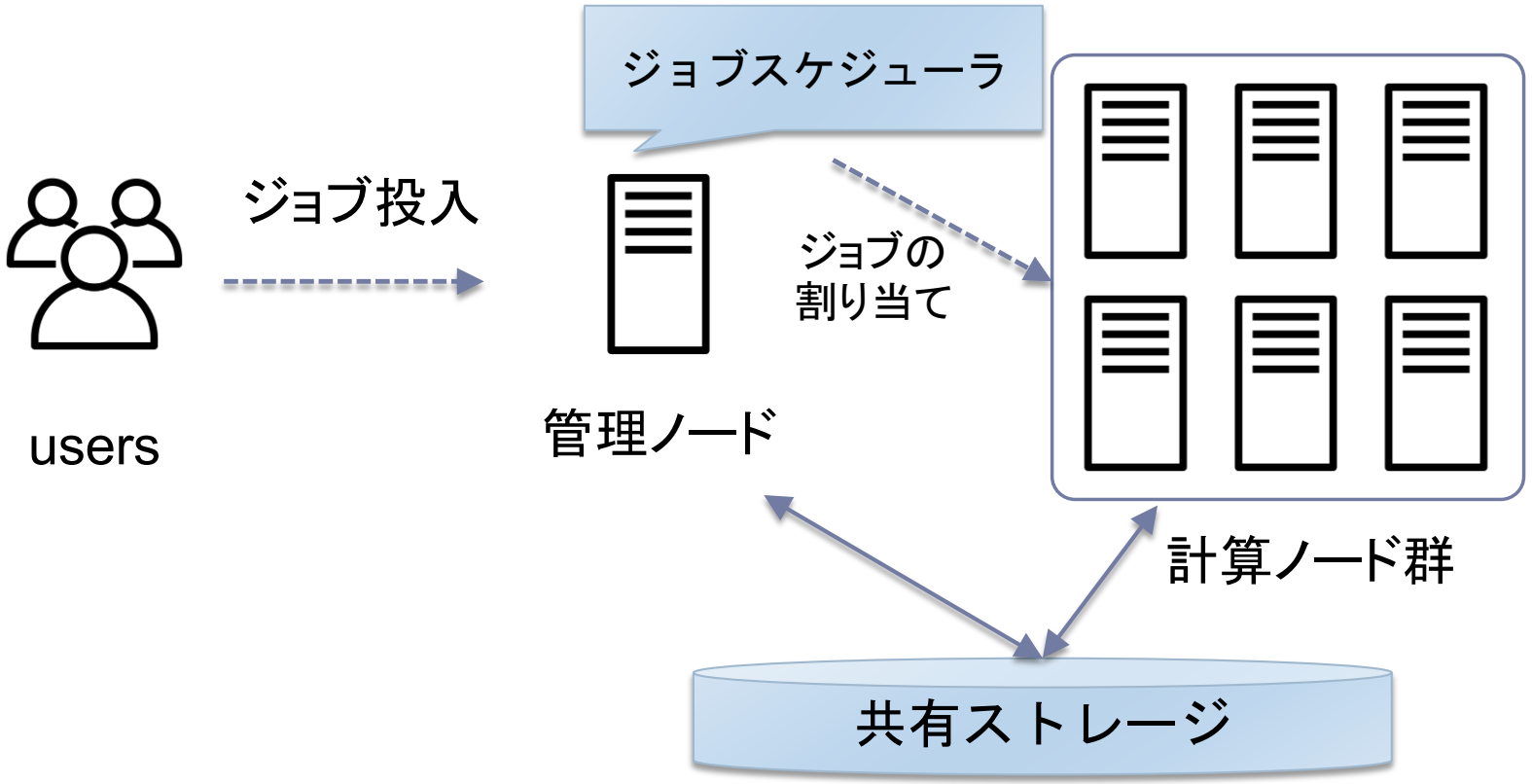
- 2018年、政府は「クラウドサービスの利用推進」を宣言
- 2023年5月、競争的研究費の直接経費からクラウド利用料の支出が可能であることが明確化
- 日本のゲノム科学界隈では、まだそれほどクラウド利用は浸透してない様子だが、先端的な利用例が報告されつつある。
  - (e.g.) Nagasaki, M. et al. Design and implementation of a hybrid cloud system for large-scale human genomic research. Hum Genome Var 10, 6 (2023).

ますますビッグデータ化するゲノム生物学
どうする　コンピュータリソース？

1. HPC（High-Performance Computing）
2. ゲノム関連の情報とツールの共有

HPC（High-Performance Computing）on AWS

HPC環境の典型的な構成要素



多くの大学や研究機関では共有のHPC（クラスターコンピュータ・スパコン）を整備。

共用HPCの問題点

- 共用のオンプレミスHPCではジョブ待ちが頻繁。
- OSやライブラリのバージョンが固定。ソフトウェアが動かない場合、対処が困難。
- オンプレミスサーバは調達に時間がかかり、手続きも煩雑。メンテナンスなどの運用負担。

ゲノム科学の基礎研究に特有の事情

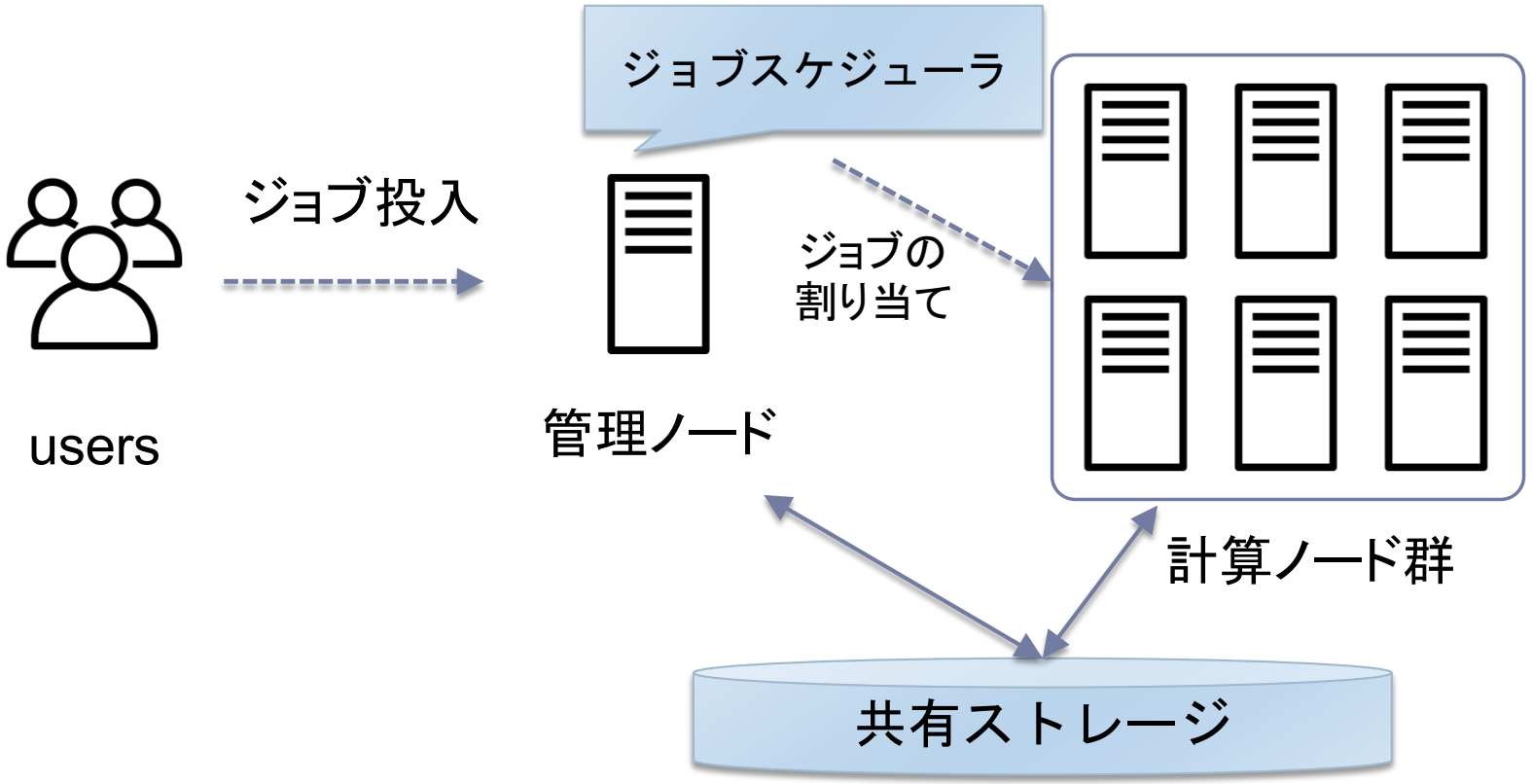
- 次世代シーケンスデータなどファイルサイズが大きくなりがち
- 基礎研究においては頻繁なトライ＆エラーが重要
- ゲノム解析のワークフローでは多様なソフトウェアを組み合わせて利用する。それぞれのソフトウェアごとに必要なコンピュータリソースの特性が異なる。

ゲノム科学の基礎研究に適した「HPC」はクラウドで実現可能か？

Yes

AWS Parallel Cluster

HPC環境の典型的な構成要素



AWSが提供するHPCフレームワーク

AWS Parallel Cluster

AWS ParallelClusterとは

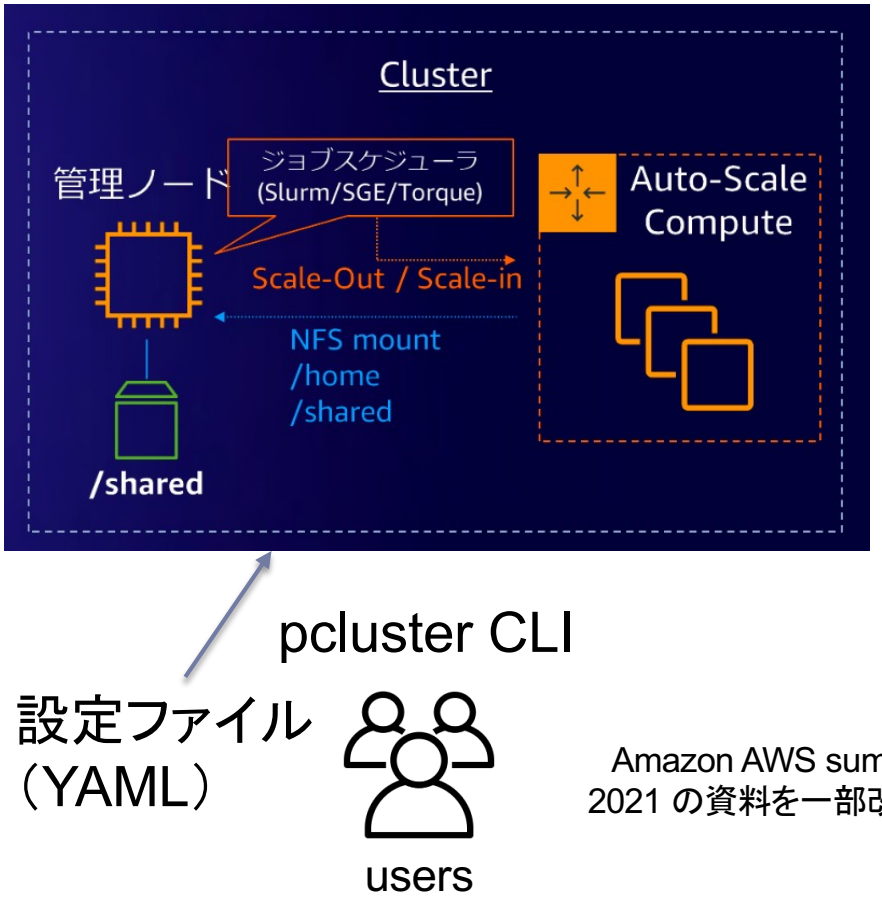
数コマンド操作で、ジョブ投入に応じて自動でスケールするクラスタをAWS上に構築可能なAWS公式のオープンソースソフトウェア

主な特徴

- HPC向けジョブスケジューラ(Slurm)とauto-scaling を連携した環境を作成
- 使用するOSや、インスタンスタイプ、ネットワーク環境、ストレージ構成などを柔軟にカスタマイズ可能
- OSS (https://github.com/aws/aws-parallelcluster)

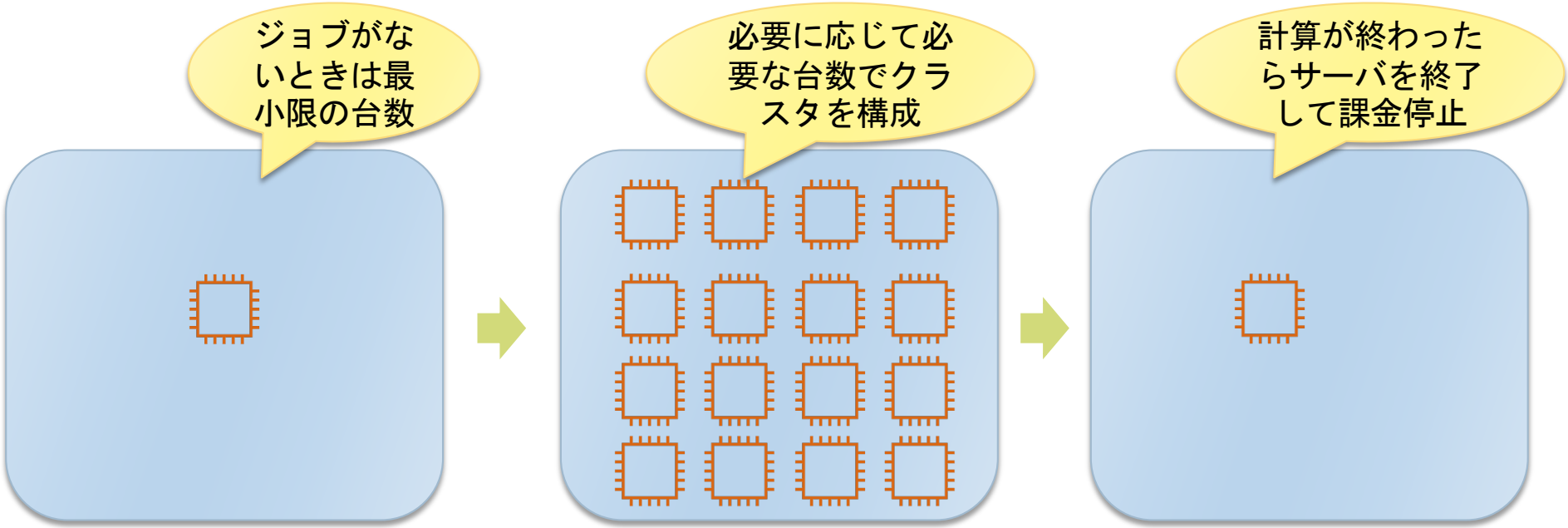
利用方法

- Configファイルを YAMLフォーマットで記述。`pcluster create` コマンドを実行する。
- 管理ノードにログイン。実行したいジョブのシェルスクリプトを作成し、ジョブ投入。



Amazon AWS summit 2021 の資料を一部改変

柔軟に自動スケールするクラスタ



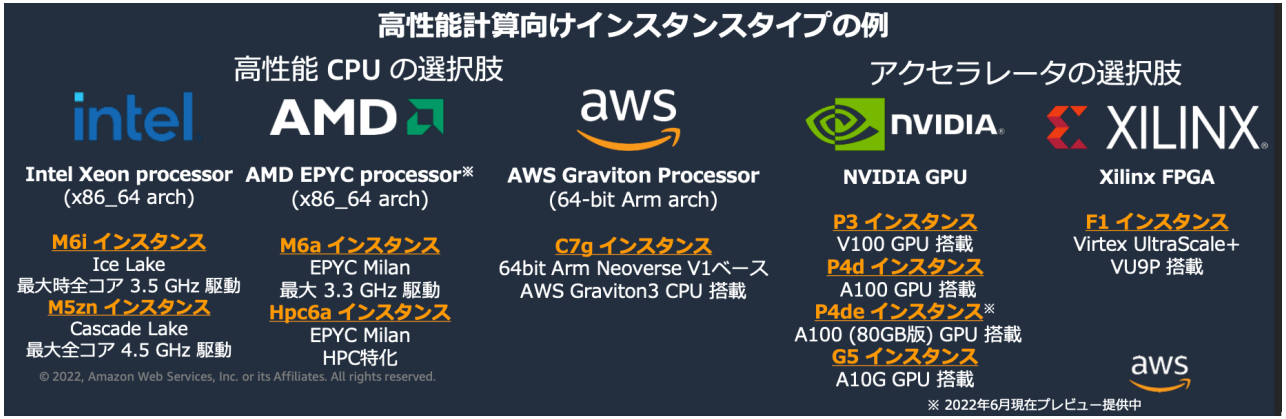
技術的メリット

- ジョブ待ちの無いHPC環境を実現。
- 必要な時に必要な分の計算リソースを一気に確保＝＞解析時間短縮。

コスト的メリット

- 費用は「時間x台数」。無駄がない。
- スポットインスタンス（最大90%OFF）を使えば大幅なコスト節約。

多様なスペックのEC2インスタンスを選択可



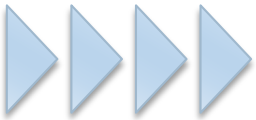
Amazon AWS summit
2021 の資料を一部改変

ゲノム解析のワークフローでは多様なソフトウェアを組み合わせて利用する。それぞれのソフトウェアごとに必要なコンピュータリソースの特性が異なる。インスタンスタイプの選択肢が広いことはアドバンテージ。

- C, Mシリーズ（汎用・計算）はコア数とメモリのバランスが取れており、大小インスタンスを組み合わせることで幅広いアプリケーションに対応可能。
- I/Oの激しいジョブはローカルストレージを備えるdタイプのインスタンスが有用（例: M5d.8xlarge: 32 CPUs, 128 GB RAM, 2x600 SSD）
- ゲノム解析には高メモリを要求するソフトウェアも多い。R、X、ハイメモリシリーズが有用。
- 機械学習、AIにはGPUシリーズが有用もしくは必須（例：AlphaFold2）
- スポットインスタンスでコスト削減

Before

AWS Parallel Cluster



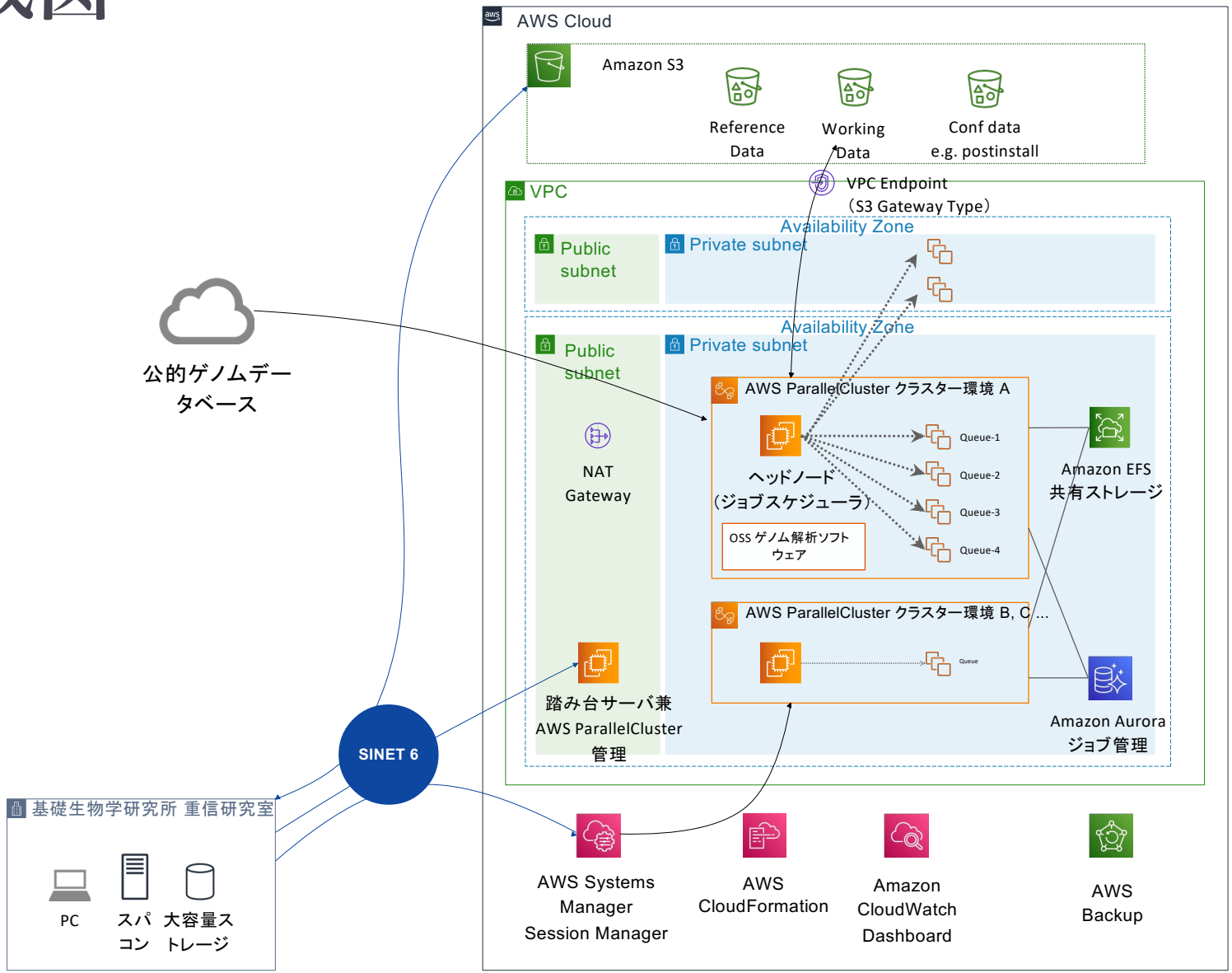
After

- 共用のオンプレミスHPCではジョブ待ちが頻繁。
- OSやライブラリのバージョンが固定。ソフトウェアが動かない場合、対処が困難。
- オンプレミスサーバは調達に時間がかかり、手続きも煩雑。メンテナンスなどの運用負担。

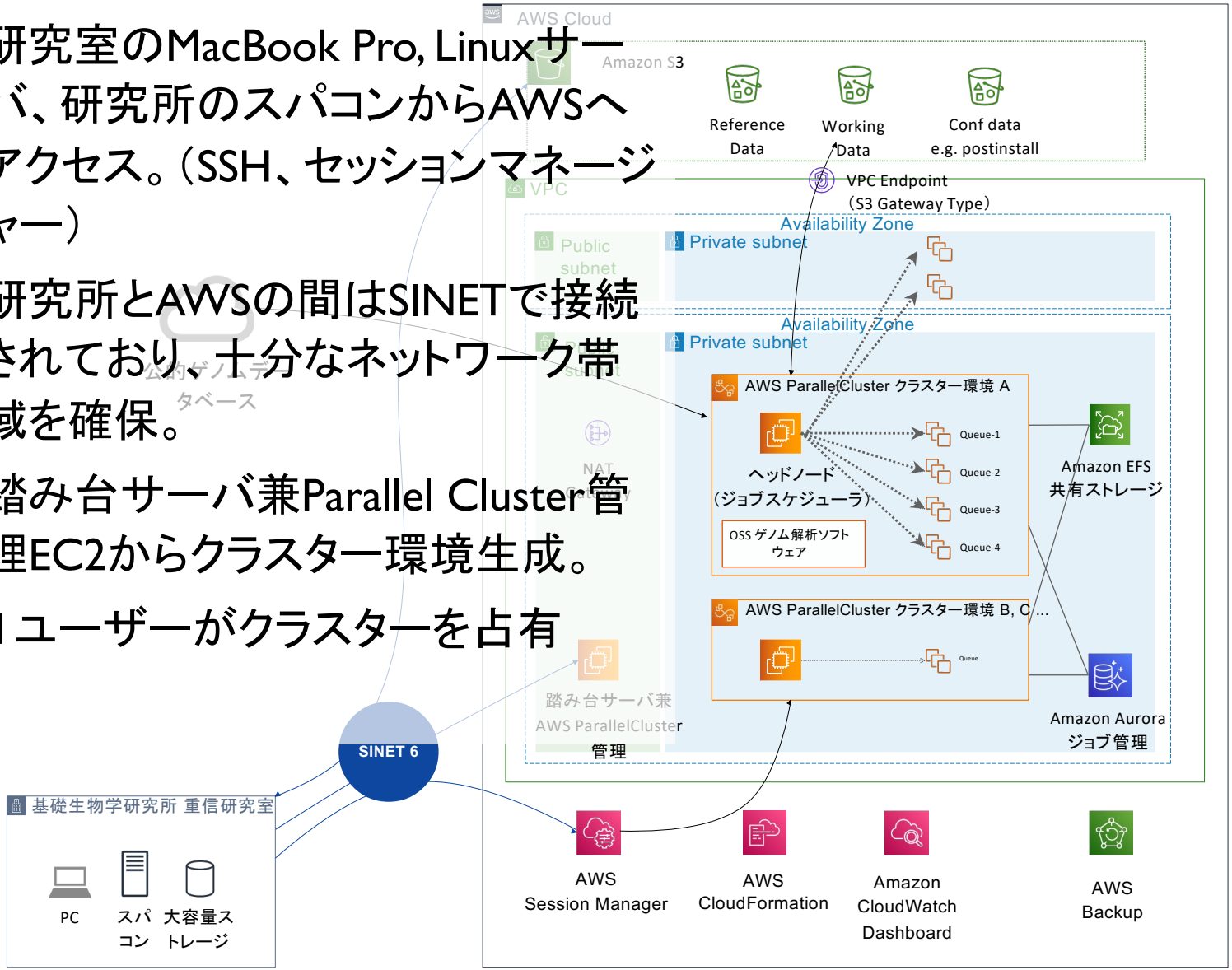
- HPCの計算機リソースを必要な時に必要な分を即時確保。
- 理想的な計算環境を、root権限でスクラップ＆ビルド。
- 事務手続きやコンピュータメンテナンスから解放されて、研究そのものに集中。

当研究室での運用の実際

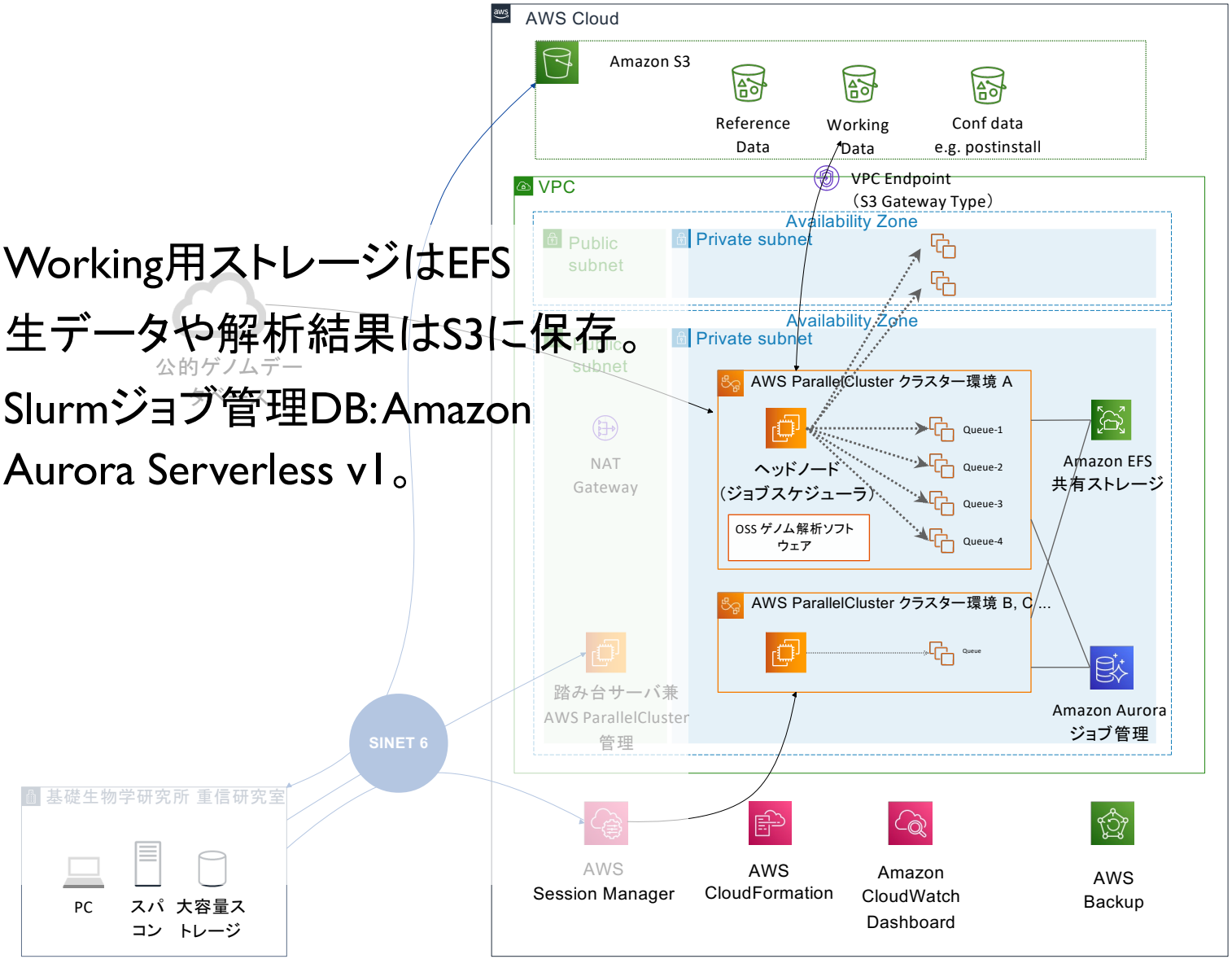
構成図


AWS Cloud
Amazon S3
Reference Data
Working Data
Conf data
e.g. postinstall
VPC Endpoint
（S3 Gateway Type）
VPC
Availability Zone
Public subnet
Private subnet
Availability Zone
Public subnet
Private subnet
公的ゲノムデータベース
AWS ParallelCluster クラスター環境 A
Queue-1
Queue-2
Queue-3
Queue-4
ヘッドノード
（ジョブスケジューラ）
OSS ゲノム解析ソフトウェア
Amazon EFS
共有ストレージ
NAT Gateway
AWS ParallelCluster クラスター環境 B, C…
Queue
踏み台サーバ兼
AWS ParallelCluster
管理
Amazon Aurora
ジョブ管理
SINET 6
基礎生物学研究所 重信研究室
PC
スパコン
大容量ストレージ
AWS Systems Manager
Session Manager
AWS CloudFormation
Amazon CloudWatch
Dashboard
AWS Backup

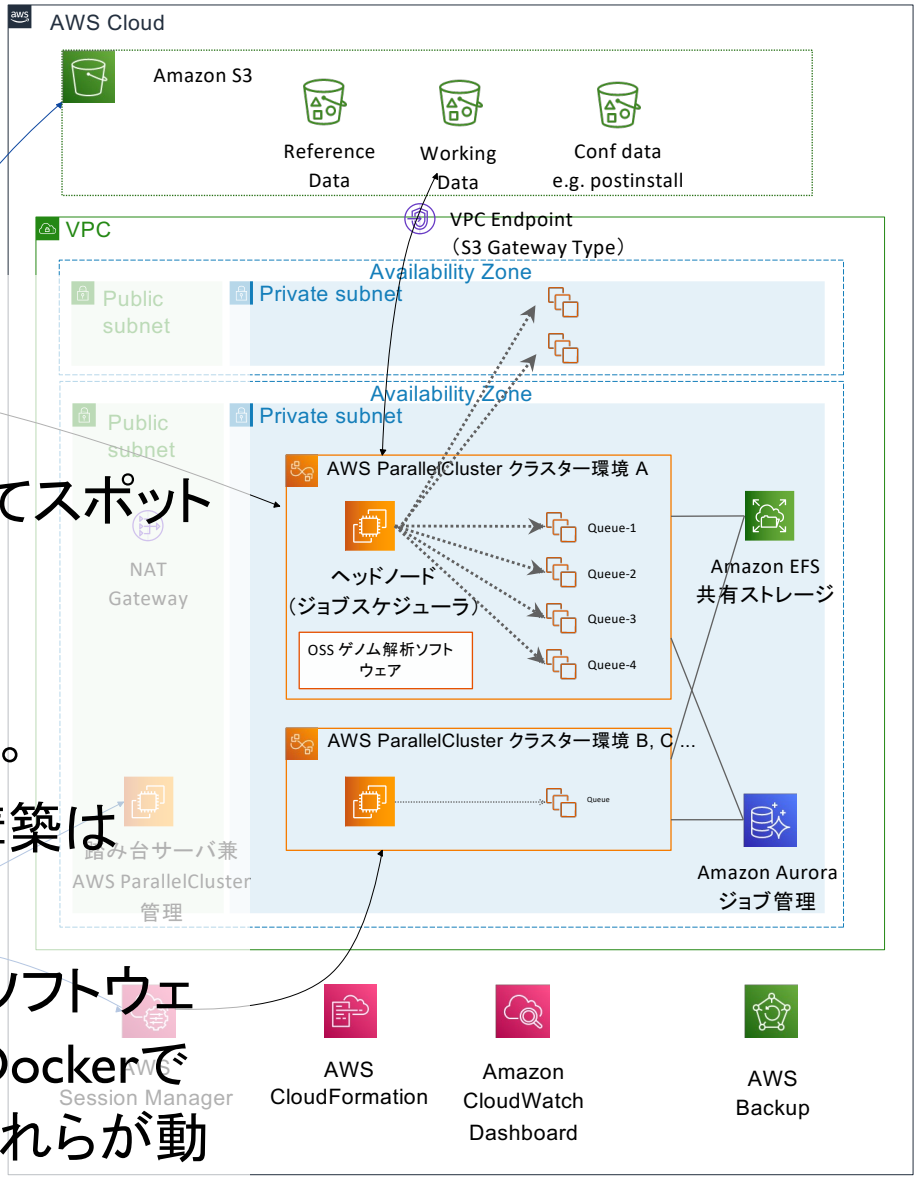
- 研究室のMacBook Pro, Linuxサーバ、研究所のスパコンからAWSへアクセス。（SSH、セッションマネージャー）
- 研究所とAWSの間はSINETで接続されており、十分なネットワーク帯域を確保。
- 踏み台サーバ兼Parallel Cluster管理EC2からクラスター環境生成。
- 1ユーザーがクラスターを占有



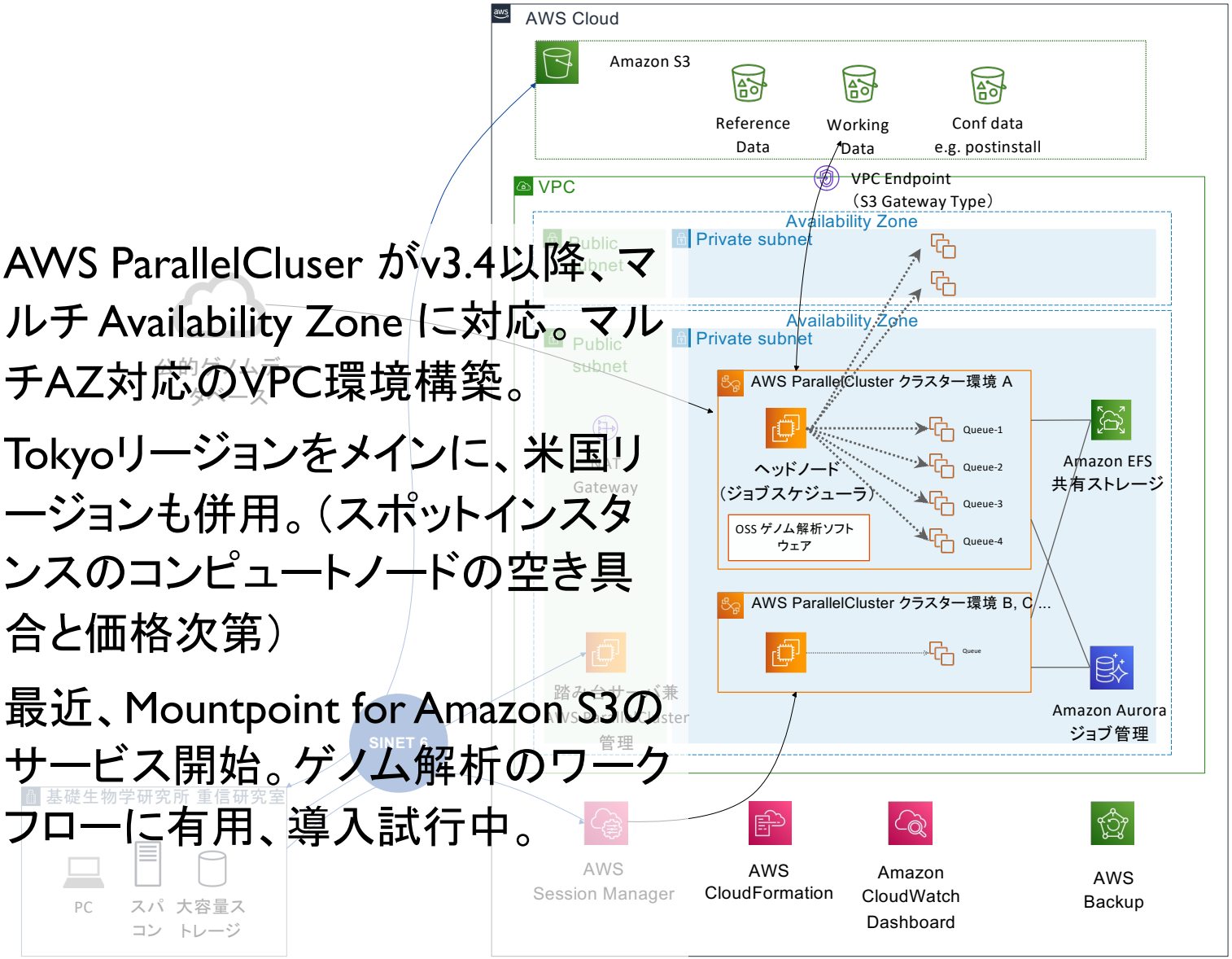
- Working用ストレージはEFS
- 生データや解析結果はS3に保存。
- Slurmジョブ管理DB: Amazon Aurora Serverless v1。



- OS: Ubuntu 20.04 LTS
- CPU: intel系
- 計算ノードは基本的に全てスポットインスタンス。
- AWSのリソース構築はCloudFormationで自動化。
- ヘッドノードの計算環境構築はAnsibleによって自動化。
- バイオインフォマティクスソフトウェアは、conda (Bioconda), Dockerで提供されることが多い。これらが動くように初期設定。



- AWS ParallelCluser がv3.4以降、マルチ Availability Zone に対応。マルチAZ対応のVPC環境構築。
- Tokyoリージョンをメインに、米国リージョンも併用。（スポットインスタンスのコンピュートノードの空き具合と価格次第）
- 最近、Mountpoint for Amazon S3のサービス開始。ゲノム解析のワークフローに有用、導入試行中。



例：ヘビゲノムのBUSCO解析

シマヘビゲノムのBUSCO5の解析をParallelClusterで、５パターンで並列解析。
BUSCOとはゲノムアセンブリの完全性をユニバーサル遺伝子セットの網羅率で評価するソフトウェア。

- BUSCO解析をデータベースを5パターンで一気に実施。
- BUSCOはconda環境でインストール済。（Dockerの利用も可能）
- インスタンスはm5d.8xlarge, 12x, 16x を選択。
- バッチジョブスケジューラ slurmにジョブを投げると、自動的に仮想マシンが立ち上がる。
- 経費節約のためにスポットインスタンスを利用した。
- I/Oの激しいジョブなので、インラインSSDを持つインスタンスタイプを選択。インラインSSDをマウントする/scratchで作業し、計算終了後、rsyncでEFS下のホームディレクトリに計算結果をコピー。
- ５パターンの解析は、それぞれ1h46m, 1h35m, 3h57m, 4h04m, 5h10mで計算終了。

```
#!/bin/bash

#SBATCH -p medium-m5d
#SBATCH -N 1
#SBATCH -n 16

INPUT=shimahebi_genome.fasta
MODE=genome
LINEAGE=metazoan #ここをvertebrataなど異なるデータベースで合計５パターン
NCPU=16

busco -i $INPUT \
    -l $LINEAGE \
    --out_path /scratch \
    -o $OUTF \
    -m $MODE \
    -c $NCPU \

rsync -avh /scratch/$OUTF ./
```

AWS ParallelCluster 個人的使用感

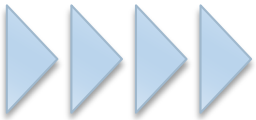
- オンプレミスのバッチジョブとほぼ同じ感覚で使用できる。
- オンプレミスの共用サーバと異なり、ユーザーが一人でクラスターを占有できる。ジョブの待ち時間がゼロ。スパコン級環境独占の圧倒的快適さ。
- 計算機資源（インスタンス、ストレージ）は事実上無限大。
- 仮想マシン上に理想的な計算環境を自らがroot権限でスクラップ＆ビルドすることができる、高い自由度により、フットワーク軽く、研究が進められる。
- 高性能なEC2インスタンスは結構高価。しかしスポットインスタンスを利用することでコストを数分の1に抑えられる。
- 最近、スポットインスタンスが以前より混んできた。対策として、マルチAZ対応や海外リージョン利用。
- 実行するソフトウェアの特性とAWSの料金体系を理解してインスタンスの選択やバッチジョブスクリプトを最適化する必要がある。
- AWSの知識が必須。ユーザーがAWS力を高めていく必要がある。

PI（研究室主宰者）目線でクラウドHPCの管理運営面でのメリット

- 研究時間捻出：マシンのメンテナンスなどの雑用から解放され、研究そのものに集中することができる。
- 柔軟なコスト管理：使用した分のみ支払い。＝＞研究費に制限のある小さな研究室でも気軽にHPCが利用できる。
- 調達手続き不要：高額なコンピュータをオンプレで購入する場合、入札等で導入まで数ヶ月以上かかる上、手続きが面倒。
- セキュリティ：自分で計算機を管理するより確実・安全。（もちろん十分な知識と対策は必要）。
- オンプレの共用スパコンを否定するものではない。効果的なハイブリッド活用。

Before

AWS Parallel Cluster



After

- 共用のオンプレミスHPCではジョブ待ちが頻繁。
- OSやライブラリのバージョンが固定。ソフトウェアが動かない場合、対処が困難。
- オンプレミスサーバは調達に時間がかかり、手続きも煩雑。メンテナンスなどの運用負担。

- HPCの計算機リソースを必要な時に必要な分を即時確保。
- 理想的な計算環境を、root権限でスクラップ＆ビルド。
- 事務手続きやコンピュータメンテナンスから解放されて、研究そのものに集中。

– HPCはクラウドで独占型スパコンを実現可能。快適。
– ゲノム科学や基礎科学に特有のニーズにも柔軟に対応。

ゲノム情報をクラウドを介して共有する

HPCで得られたゲノム解析結果をどのように共有するか？

- 「ゲノムブラウザ」を介して、共同研究者とデータを共有する。
- 「ゲノムブラウザ」の構築にはクラウドが最適。

ゲノムブラウザの例

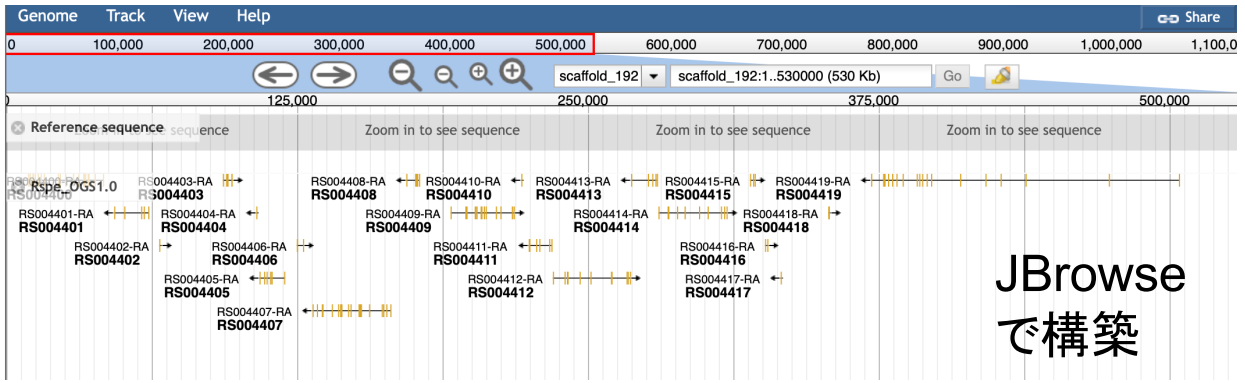
基礎生物学研究所 National Institute for Basic Biology
富山大学 UNIVERSITY OF TOYAMA
東京大学 大学院 理学系研究科・理学部 SCHOOL OF SCIENCE, THE UNIVERSITY OF TOKYO
慶應義塾大学
琉球大学 UNIVERSITY OF THE RYUKYUS

２０２２年１月１９日

高度な社会性を持つシロアリのゲノム情報を解読

～遺伝子重複が社会性進化の原動力であることを明らかに～

Shigenobu et al., 2022 PNAS



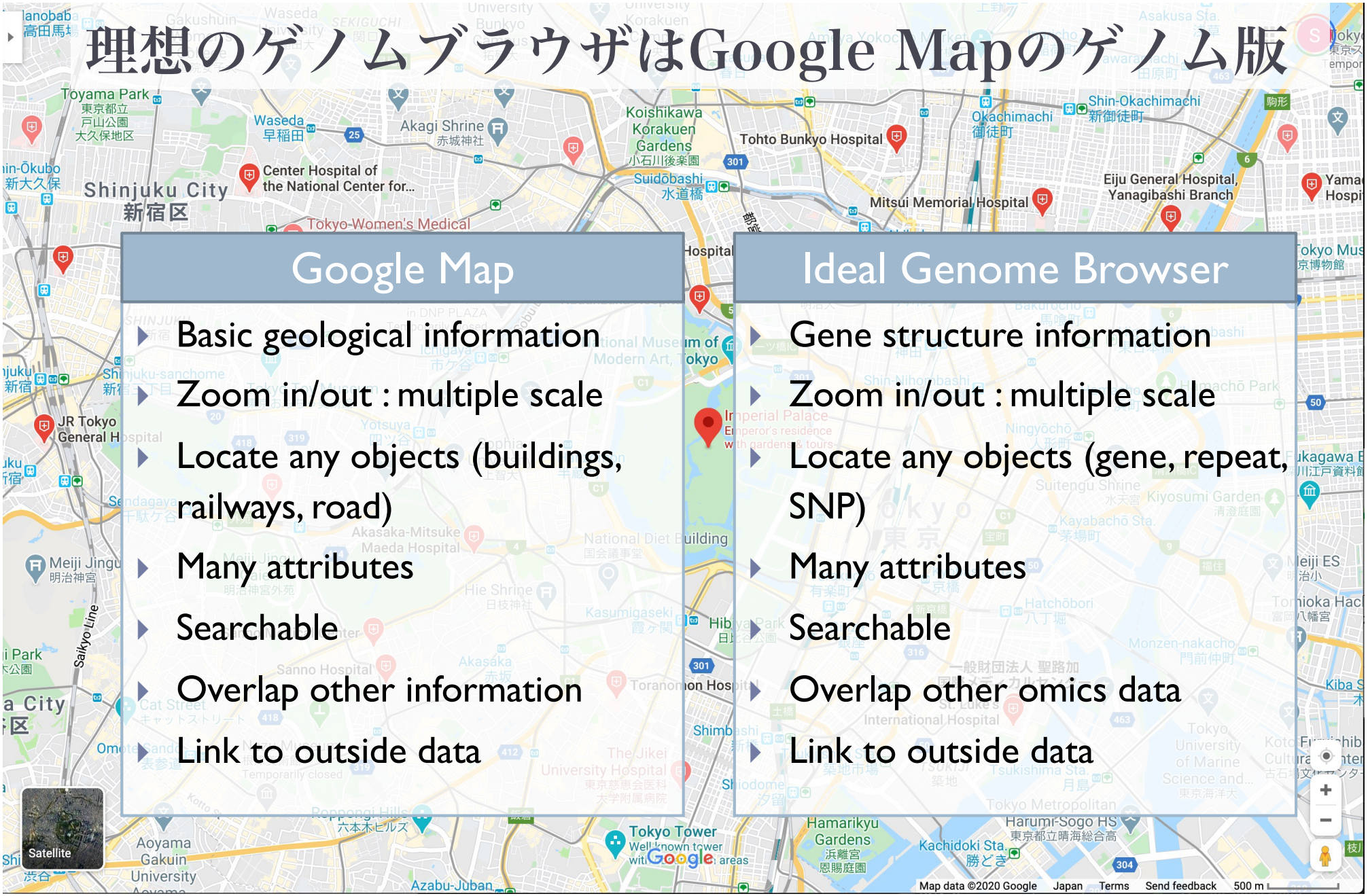
公開ゲノムブラウザ　http://www.termite.nibb.info/retsp/

（AWS EC2, Route 53 等を利用）

ゲノムブラウザ

- ゲノムブラウザ（Genome Browser）は、ゲノムの情報を視覚的に閲覧・検索するためのツールやウェブサイト。
- 遺伝子構造に関する情報や、その遺伝子に関連するアノテーション情報（機能や発現データなど）を閲覧。
- 公開ゲノムブラウザの例としては、UCSC Genome BrowserやEnsemblなど。これらのブラウザはウェブベースでアクセスできる。
- 自前で構築するためのツールとしては、JBrowse2とIGV (Integrative Genome Viewer)が有名。

理想のゲノムブラウザはGoogle Mapのゲノム版



Buels *et al. Genome Biology* (2016) 17:66
DOI 10.1186/s13059-016-0924-1

Genome Biology

SOFTWARE Open Access

JBrowse: a dynamic web platform for genome visualization and analysis

Robert Buels[1], Eric Yao[1], Colin M. Diesh[2], Richard D. Hayes[3,6], Monica Munoz-Torres[3], Gregg Helt[3,4], David M. Goodstein[3,6], Christine G. Elsik[2], Suzanna E. Lewis[3], Lincoln Stein[5,7] and Ian H. Holmes[1,3*]

Abstract

Background: JBrowse is a fast and full-featured genome browser built with JavaScript and HTML5. It is easily embedded into websites or apps but can also be served as a standalone web page.

Results: Overall improvements to speed and scalability are accompanied by specific enhancements that support complex interactive queries on large track sets. Analysis functions can readily be added using the plugin framework; most visual aspects of tracks can also be customized, along with clicks, mouseovers, menus, and popup boxes. JBrowse can also be used to browse local annotation files offline and to generate high-resolution figures for publication.

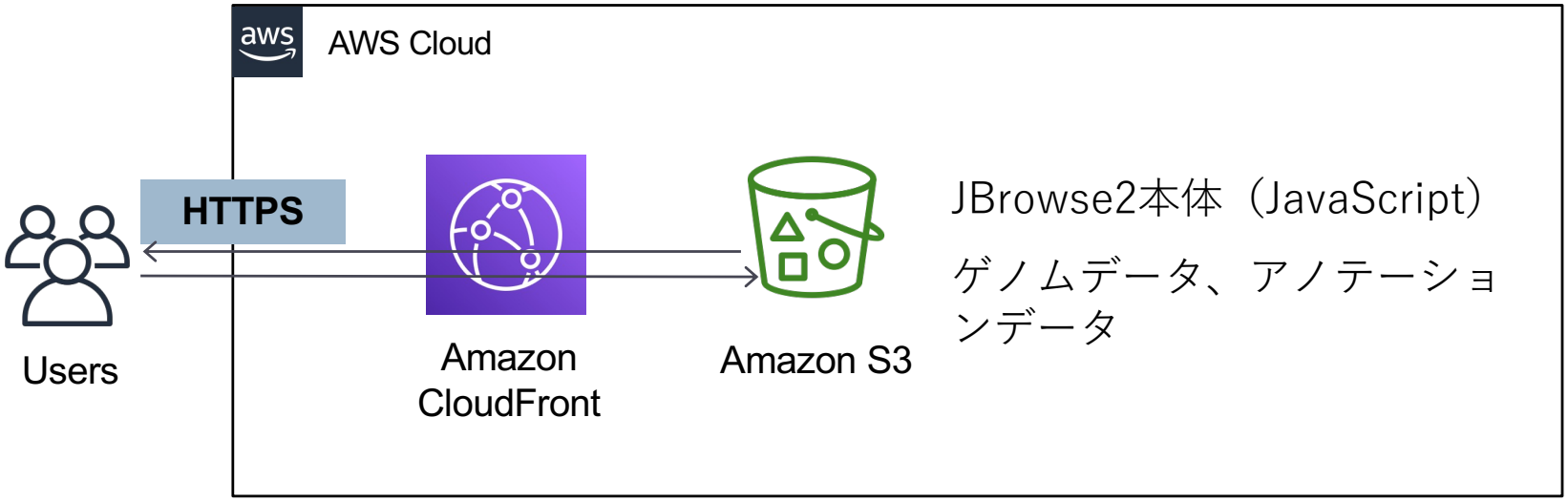
Conclusions: JBrowse is a mature web application suitable for genome visualization and analysis.

Keywords: Genome, Browser, Bioinformatics

http://jbrowse.org/

新規ゲノムブラウザをJBrowse2を用いてAWS上にサーバレスに構築する

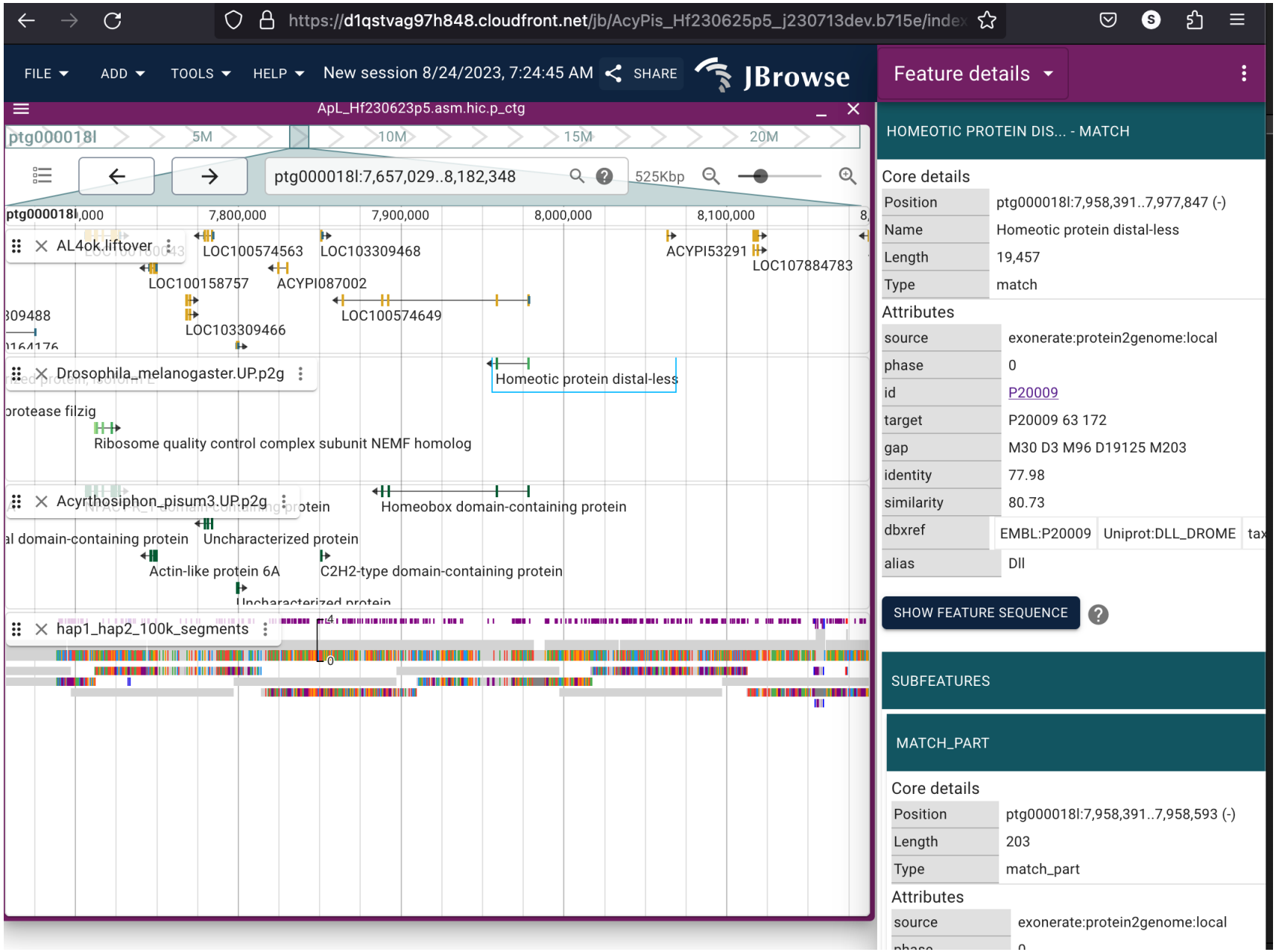
- JBrowse2はJavaScriptで書かれている。リレーショナルデータベースは不要。
- S3 + CloudFront のサーバレスアーキテクチャーで実装



ゲノムブラウザをAWS上にサーバレスに構築するメリット

- サーバレスなのでサーバの管理が不要。
- コストはデータ保存料とデータ転送料のみで安価。
- CloudFrontの機能により、高負荷耐性、世界中へ高速配信。
- 最近は大学や研究機関のネットワークセキュリティが強化され、内部ネットワークに公開サーバを設置することがNGであることがほとんど。

例）当研究室で解読したアブラムシゲノムのゲノムブラウザ



HPCで得られたゲノム解析結果をどのように共有するか？

- 「ゲノムブラウザ」を介して、共同研究者とデータを共有する。
- 「ゲノムブラウザ」の構築にはクラウドが最適。

- 配列検索ツール「BLAST」のニーズが高い。
- BLASTをサーバレス化した。

National Library of Medicine
National Center for Biotechnology Information

Log in

GCP users

NCBI公式クラウド型BLAST

Blast-help > Blast Searches at a Cloud Provider > ElasticBlast

- ElasticBLAST
- Overview
- Requirements
- Quickstart
- Tutorials
- IAM Policy
- Budget
- Commands
- Configuration variables
- Exit codes
- Tips for GCP
- Limiting search by taxonomy
- Support
- Troubleshooting
- Known issues on AWS
- Known issues on GCP

ElasticBLAST, version 1.1.0

ElasticBLAST is a cloud-based tool to perform your BLAST searches faster and make you more effective.

ElasticBLAST is ideal for users who have a large number (thousands or more) of queries to BLAST or who prefer to use cloud infrastructure for their searches. It can run BLAST searches that cannot be done on NCBI WebBLAST and runs them more quickly than stand-alone BLAST+.

ElasticBLAST speeds up your work by distributing your searches across multiple cloud instances. The ability to scale resources in this way allows large numbers of queries to be searched in a shorter time that you normally could with BLAST+.

The National Center for Biotechnology Information (NCBI), part of the National Library of Medicine at the NIH, develops and maintains ElasticBLAST.

ElasticBLAST status: beta

Platforms available: AWS, GCP (account required)

Camacho *et al. BMC Bioinformatics* (2023) 24:117
https://doi.org/10.1186/s12859-023-05245-9

BMC Bioinformatics

SOFTWARE Open Access

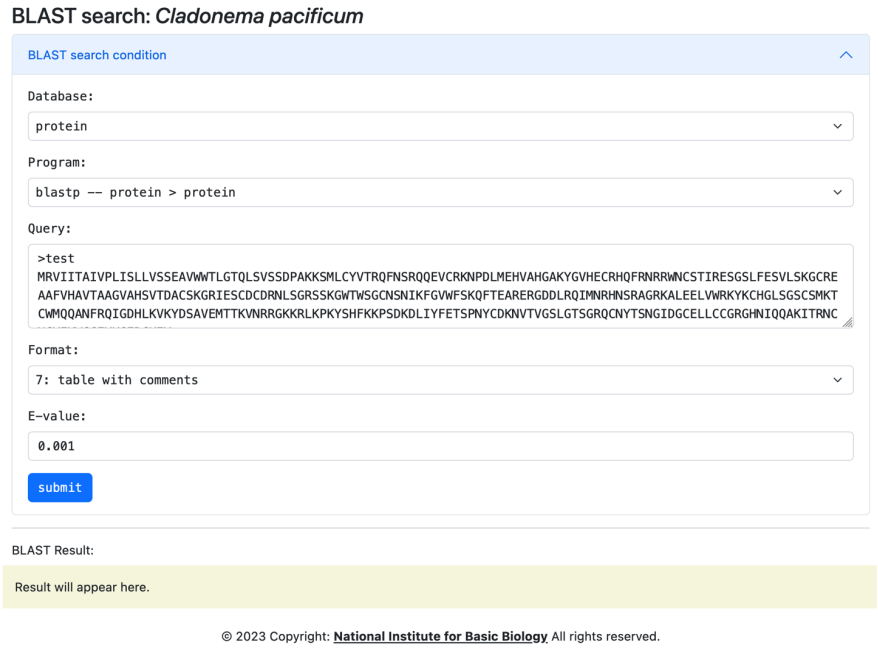
ElasticBLAST: accelerating sequence search via cloud computing

Christiam Camacho, Grzegorz M. Boratyn, Victor Joukov, Roberto Vera Alvarez and Thomas L. Madden*

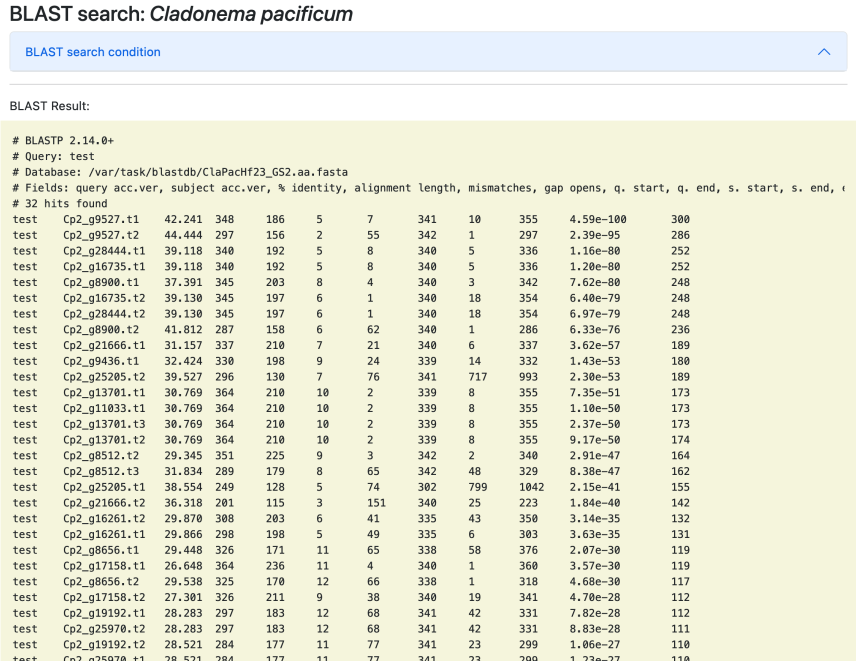
https://blast.ncbi.nlm.nih.gov/doc/elastic-blast/

Camacho et al. 2023

Our Light-weight Serverless BLAST

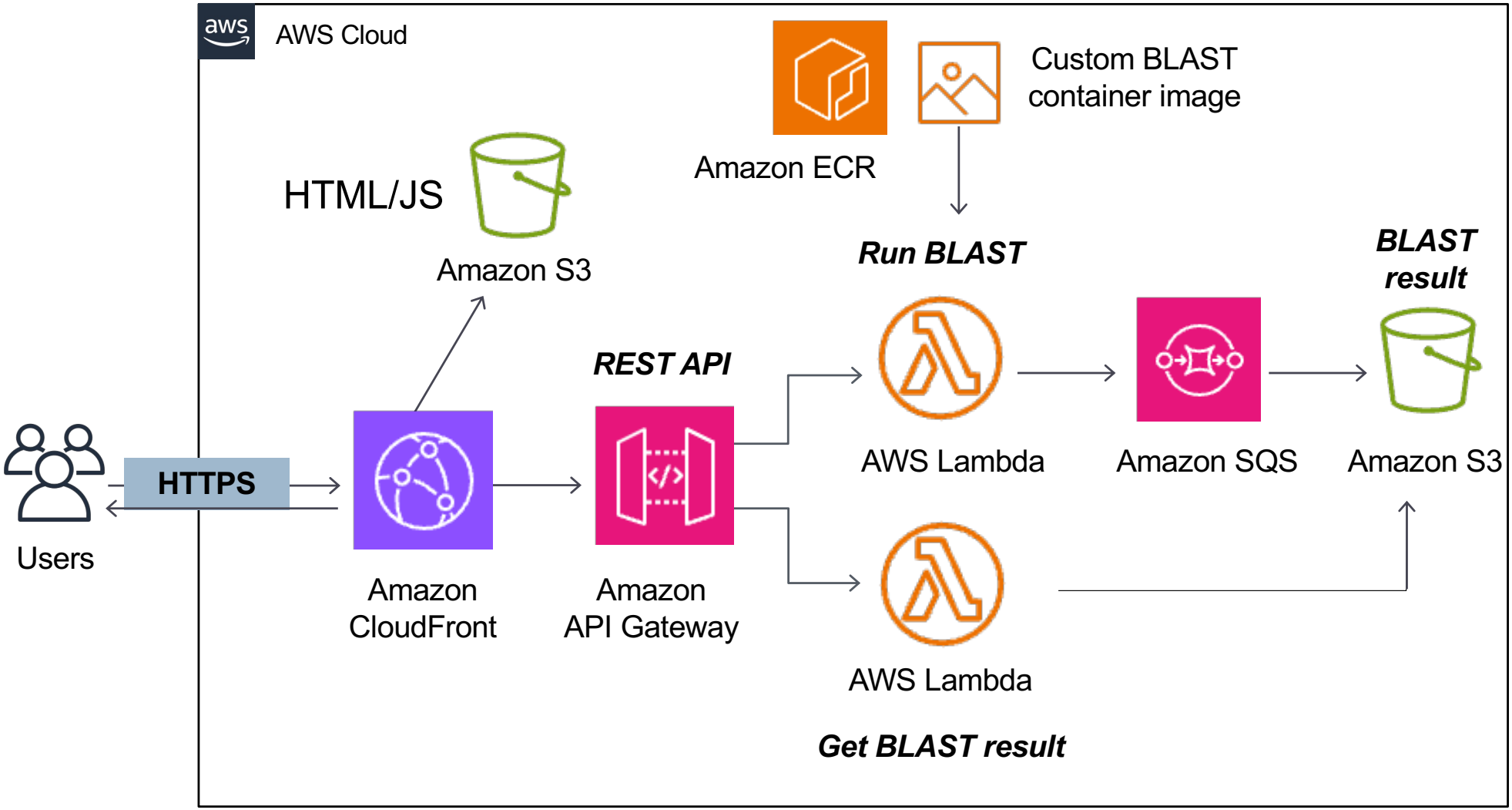


入力画面



結果画面

Our light-weght serverless BLAST



BLASTはNCBIよりUNIX CUI環境で動くバイナリが提供されている。これをDockerコンテナ化した上でAWS Lambda化。フロントエンドはJavaScriptで実装。ユーザーがウェブインターフェイスからquery配列を入力すると、非同期にBLAST実装Lambdaが起動し、計算が終わったら、結果が画面に表示される。

HPCで得られたゲノム解析結果をどのように共有するか？

- 「ゲノムブラウザ」を介して、共同研究者とデータを共有する。
- 「ゲノムブラウザ」の構築にはクラウドが最適。

- 配列検索ツール「BLAST」のニーズが高い。
- BLASTをサーバレス化した。

ゲノム解析後のゲノム情報の共有やツール提供にはクラウドが有効。開発後のメンテナンス面を考慮すると、サーバレスアーキテクチャを目指すべきである。

まとめ

ゲノム科学、特に基礎研究、におけるクラウド活用

- 計算資源（HPC）の確保
  - HPCはクラウドで独占的スパコンを実現可能。快適。
  - ゲノム科学や基礎科学に特有のニーズにも柔軟に対応。
- データ共有による共同研究促進
  - ゲノムブラウザ等ゲノム情報の共有やツール提供にはクラウドが有効。
  - サーバレスアーキテクチャを目指すべき。

謝辞：基礎生物学研究所・TSBセンター、進化ゲノミクス研究室
科研費・倉谷新学術ほか
クラスメソッド、アマゾンAWS