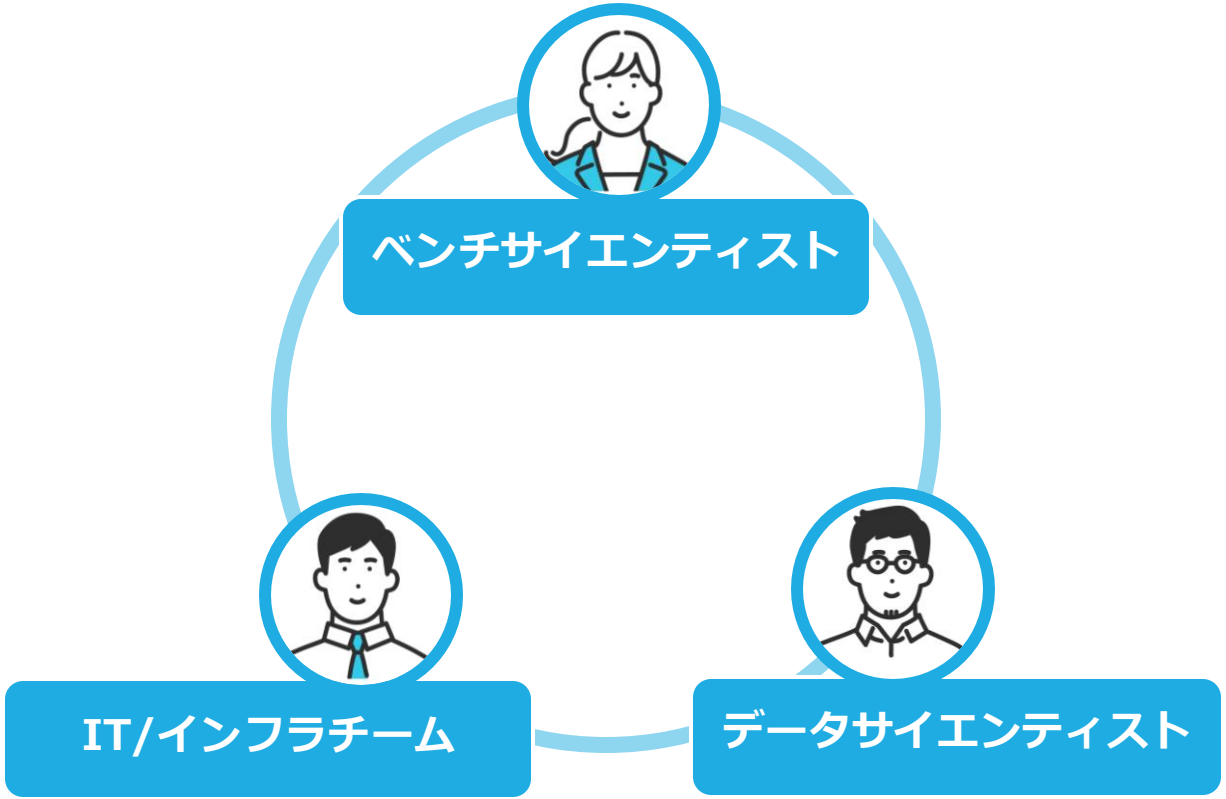
ゲノミクスDev/Opsを実現するアーキテクチャー

薬師寺 秀樹


ベンチサイエンティスト
IT/インフラチーム
データサイエンティスト

basepair

ベンチサイエンティスト

データサイエンティスト

IT/インフラチーム

ASTRIDE

basepair

ベンチサイエンティスト

データサイエンティスト

IT/インフラチーム

- インフォマティクスは自分では難しい。
- 会社支給のPCではスペックが足りない。
- インフォマティシャンに頼むのも申し訳ない。

ASTRIDE

basepair

ベンチサイエンティスト

データサイエンティスト

IT/インフラチーム

- 個別リクエストの対応が追いつかない。
- パイプラインを随時アップデートする必要がある。
- AI/MLなど新しい技術を導入し、データドリブンを勧めたい。
- オミックスデータはバラバラに格納されていてアクセスが煩雑

ASTRIDE

basepair

ベンチサイエンティスト

データサイエンティスト

IT/インフラチーム

- クラウドは利用しているが、コストを抑える必要がある。
- オミックス解析用の環境を用意したいが、ユーザー・データ管理を構築する必要がある。

ASTRIDE

basepair

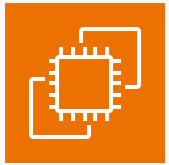
環境構築には様々な側面がある

UI/UX	GUI	Visualization
	Data sharing	API
Bioinformatics	Pipelines	Ref Genome
Infrastructure	Computing	Storage
	User/Data Mgmt	Security
	CI/CD	Cost

EC2 Instances (784)

Based on your inputs, this is the lowest-cost EC2 instance: **t2.nano**

Chosen instance: **t2.nano** | **Family: t2** | **1vCPU** | **0.5 GiB Memory**



Search instance type

Search by instance name or filter by keyword

Instance family Info: Any Instance family

vCPUs: Any vCPUs

Memory (GiB): Any Memory (GiB)

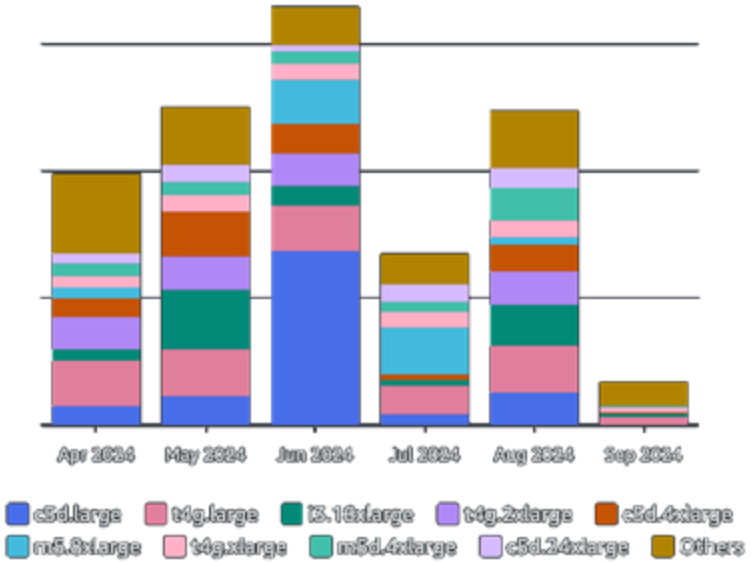
Network performance: Any Network Performance

Show only current generation instances.

	Instance name	vCPUs	Memory	Network Performance	Storage	On-Demand Hourly Cost	CurrentGeneration	Potential Effective Hourly Cost (Savings %)
○	x8g.medium	1	16 GiB	Up to 12.5 Gigabit	EBS only	0.09945	Yes	0.0000 (100%)
○	m6i.large	2	8 GiB	Up to 12500 Megabit	EBS only	0.0995	Yes	0.0000 (100%)
○	m5.large	2	8 GiB	Up to 10 Gigabit	EBS only	0.1	Yes	0.0000 (100%)
○	m4.large	2	8 GiB	Moderate	EBS only	0.1035	Yes	0.0000 (100%)
○	c4.large	2	3.75 GiB	Moderate	EBS only	0.104	Yes	0.0000 (100%)
○	m7i.large	2	8 GiB	Up to 12500 Megabit	EBS only	0.1043	Yes	0.0000 (100%)
○	r6g.large	2	16 GiB	Up to 10 Gigabit	EBS only	0.1043	Yes	0.0000 (100%)
○	c7a.large	2	4 GiB	Up to 12500 Megabit	EBS only	0.10614	Yes	0.0000 (100%)
○	r7g.large	2	16 GiB	Up to 12500 Megabit	EBS only	0.1106	Yes	0.0000 (100%)
○	c5n.large	2	5.25 GiB	Up to 25 Gigabit	EBS only	0.112	Yes	0.0000 (100%)

Basepairでよく使うインスタンス

必要な時に、必要なだけインスタンスを起動。解析後は自動的にシャットダウン。



C5 : Compute Optimized

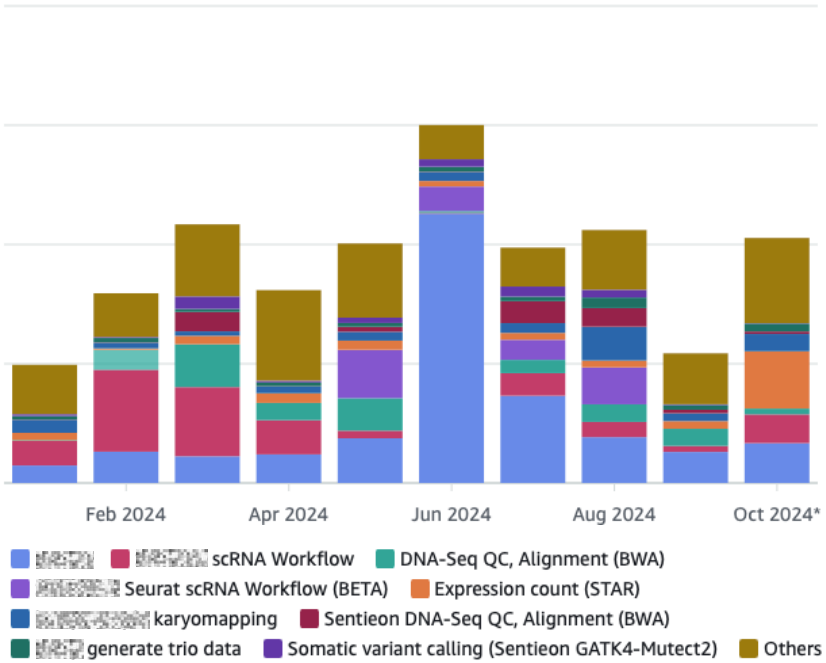
I3 : Storage Optimized

M5 : General Purpose

T4g : General Purpose

Model	vCPU	Memory (GiB)	Instance Storage (NVMe SSD, GB)
c5d.large	2	4	1 x 50
c5d.4xlarge	16	32	1 x 400
c5d.24xlarge	96	192	4 x 900

パイプラインごと



- scRNA-Seq、WGSが高負荷
- RNA-Seqは数は多いが低負荷

- パイプラインごとにインスタンスを選択
- 常時モニターし適宜調整

ワークフローエンジン

- **Amazon SWF（Simple Workflow Service）**

> “連携ロジックを完全に制御することが可能ですが、アプリケーションの開発はより複雑になります”

- **AWS Step Functions（AWS推奨）**

> 新規アプリケーションには Step Functions を利用することを検討してください

Basepairは Temporal

Spot instance

- AWSクラウドの未使用のEC2容量を活用
- オンデマンド価格と比較して最大90%のコスト削減

Spot Instance interruption – Amazon EC2 terminates, stops, or hibernates your Spot Instance when Amazon EC2 needs the capacity back. Amazon EC2 provides a Spot Instance interruption notice, which gives the instance a two-minute warning before it is interrupted.
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-instances.html

あなたのスポットインスタンスが
停止されるまであと

２：００

Basepairの解

- **積極的に活用**
  - 特に実行時間の短いパイプライン（RNA-Seq関係など）には有効
- ***Spot Instance interruption* に対応**
  - 2分前通知を受け取ると自動的にシャットダウン、開放後自動再開する仕組みを構築

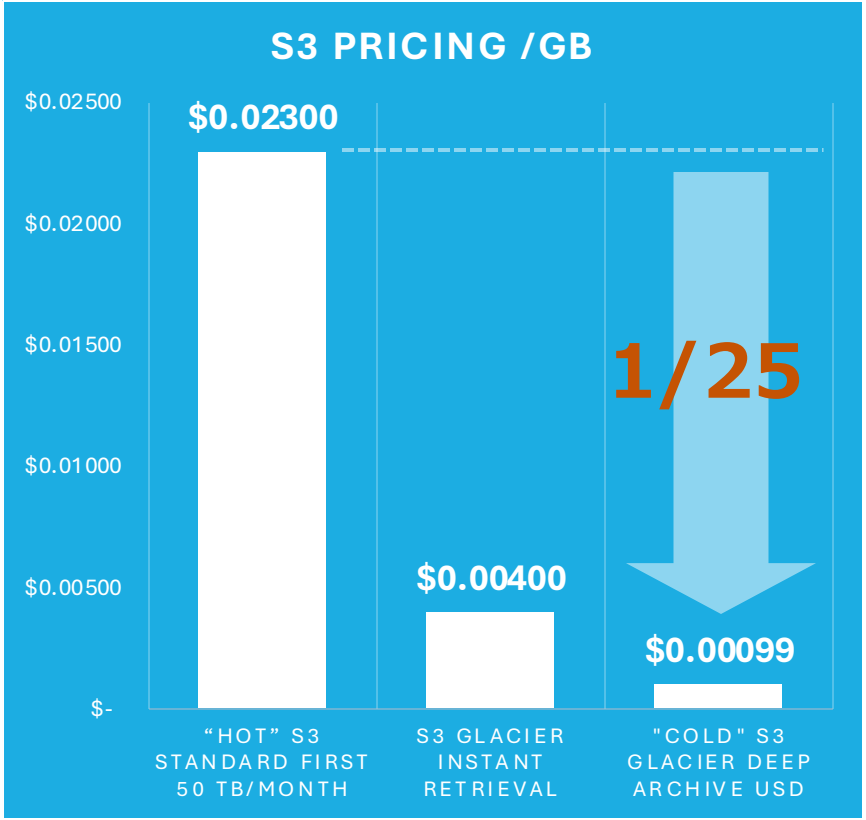
Save ~50%

basepair

Storage

- S3 Intelligent-Tiering
- **S3 Standard**
- S3 Standard-Infrequent Access
- S3 One Zone-Infrequent Access
- **S3 Glacier Instant Retrieval**
- S3 Glacier Flexible Retrieval
- **S3 Glacier Deep Archive**
- S3 Outposts

でも、そんな単純じゃない！



ASTRIDE


ストレージのジレンマ
古いデータも
解析したい！
vs
コストを
抑えねば！

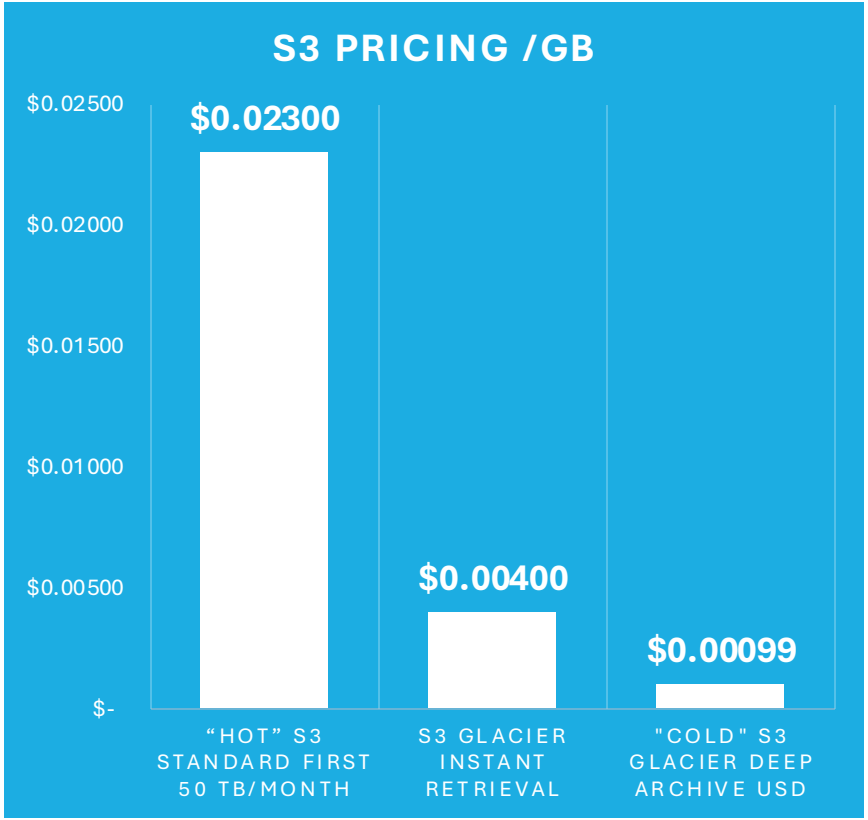
basepair

Basepairの解
オリジナルロジックを構築

- データタイプによる傾向
- Hot ⇄ COLD 直移行
- 積極的なアーカイブ

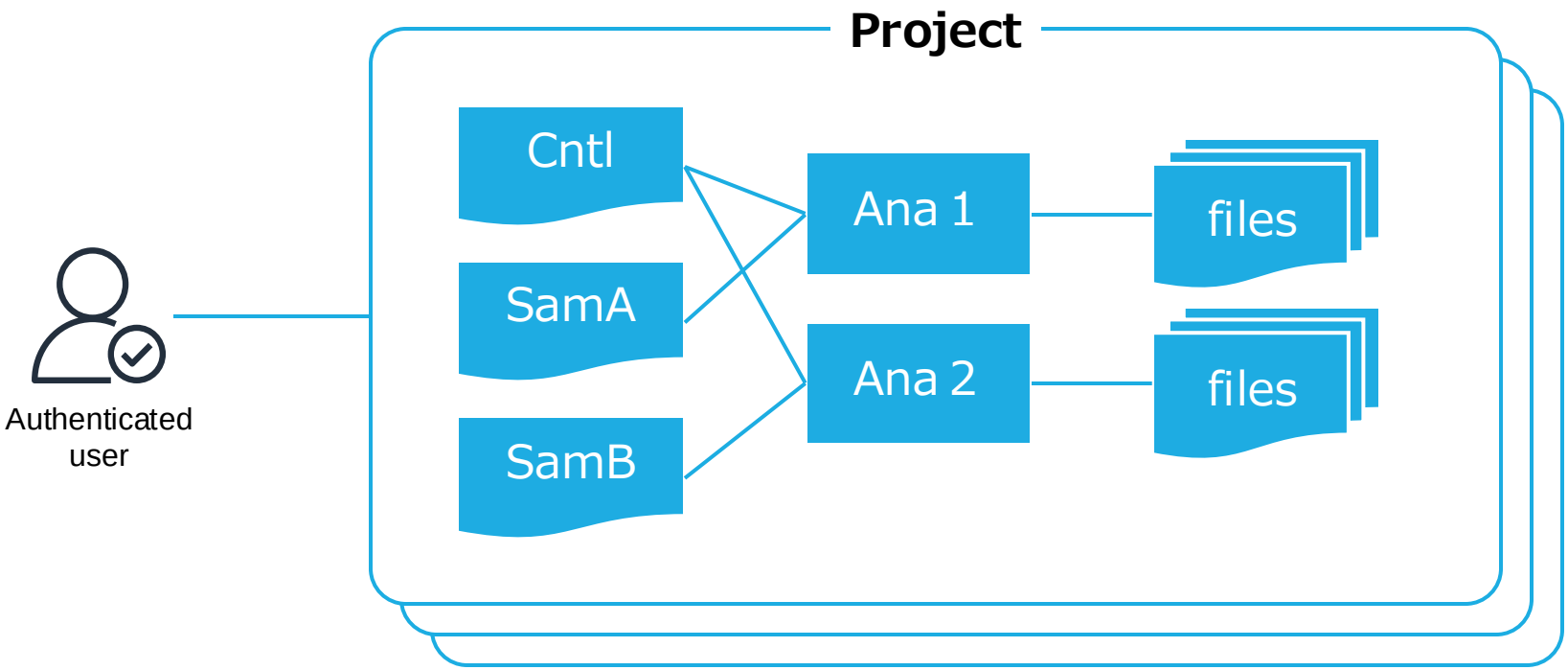
（ユーザーのAWSアカウントのS3の場合は、ユーザーのポリシーに準ずる）

Save ~80%

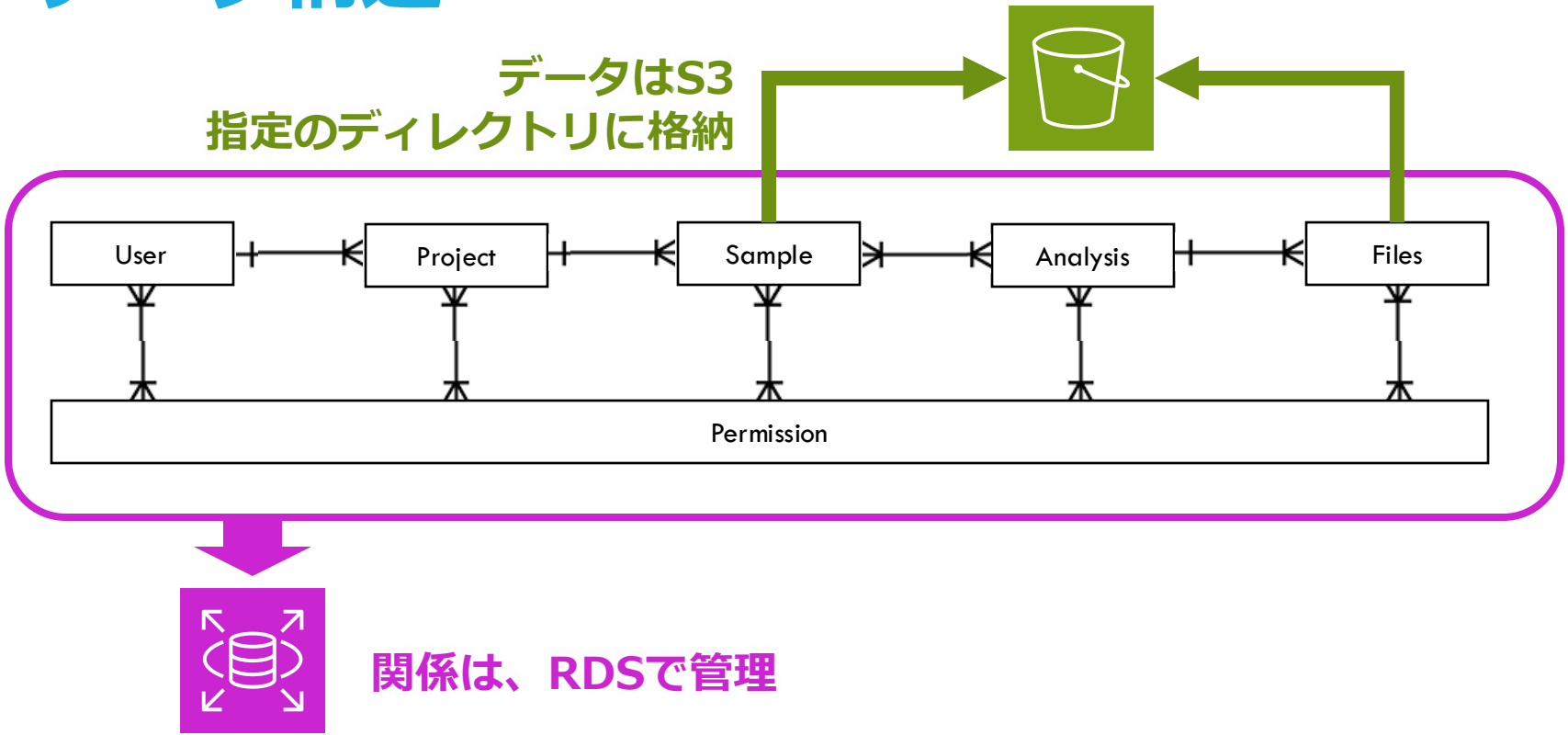


ASTRIDE

プロジェクトの構造

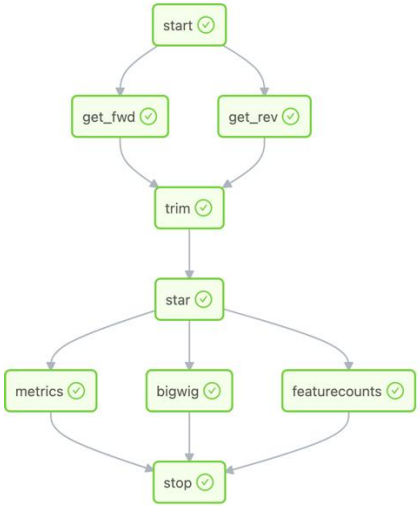

Project
Cntl
SamA
SamB
Ana 1
Ana 2
files
files
Authenticated user

データ構造

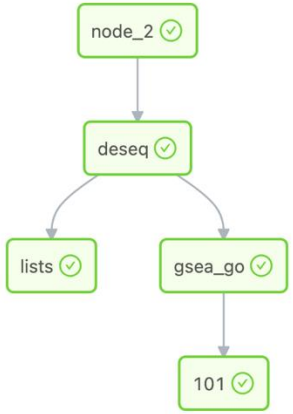

データはS3
指定のディレクトリに格納
User
Project
Sample
Analysis
Files
Permission
関係は、RDSで管理

Basepairのパイプライン
（Differential Expressionの場合）

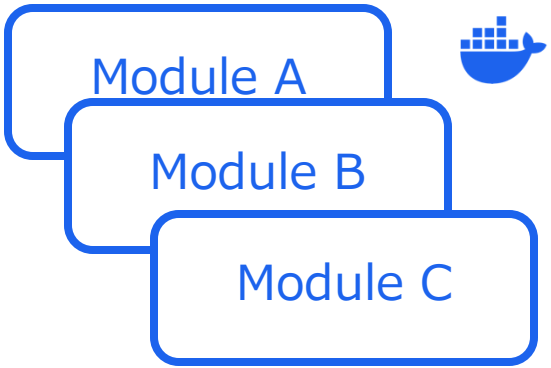
Expressio Count (STAR)



Differential Expression (DESeq2)



パイプライン実装



docker

Module Definition

- パス、コマンド構造
- インプット
- アウトプット

GitHub

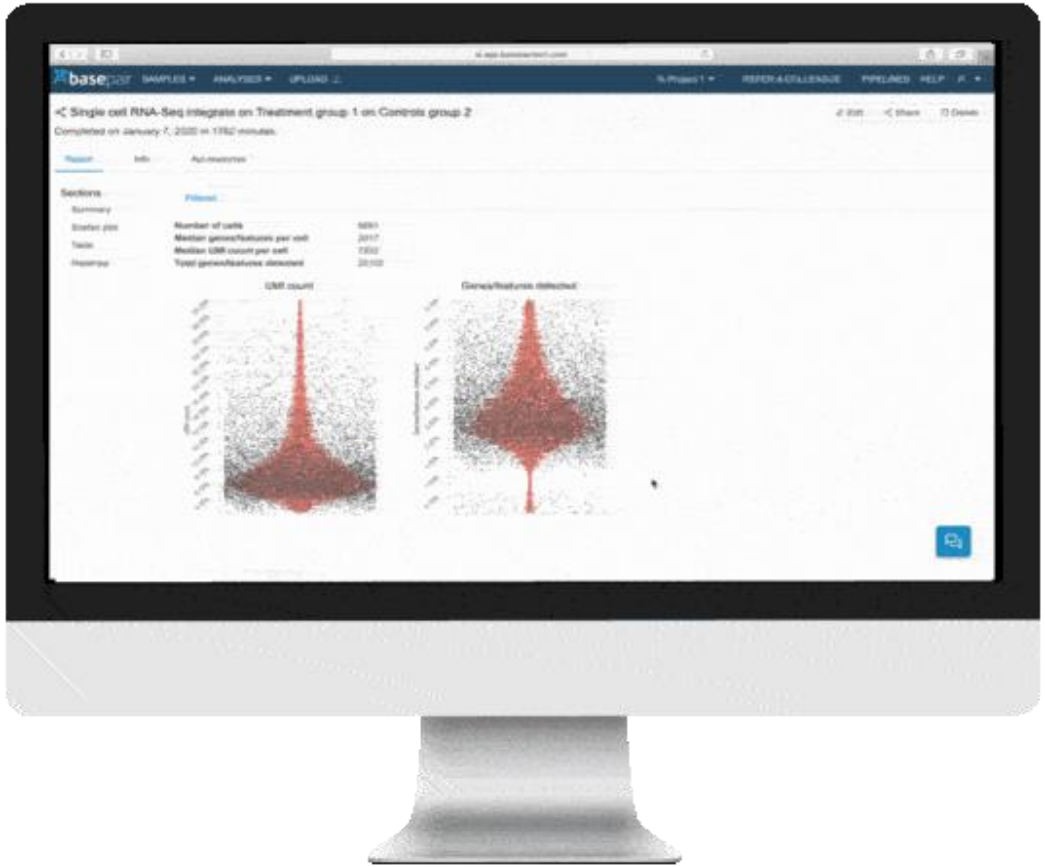
Workflow definition

- ノードのコレクション
- エッジのコレクション
- マッピング

nextflow CWL {wdl} でもOK

ASTRIDE

basepair



ASTRIDE

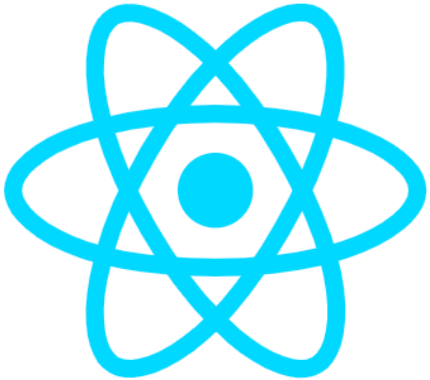
basepair

フロントエンド

Ant Design

+

React

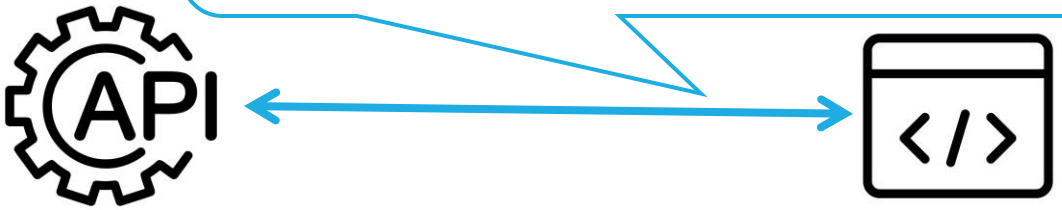


+

API

basepair

他アプリケーションから操作
- サンプル、解析、結果
- アップロード、ダウンロード、実行、削除



REST API

```
import basepair

# CREATE SAMPLE
data = {
    "name": "Sample 10",
    "genome": "hg19",
    "datatype": "dna-seq",
    "file1": "/path/to/file1.fastq.gz",
    "file2": "/path/to/file2.fastq.gz",
}
sample_id = basepair.create_sample(data)


# RUN ANALYSIS
basepair.create_analysis(workflow_id=5, sample_id)
```

Python API

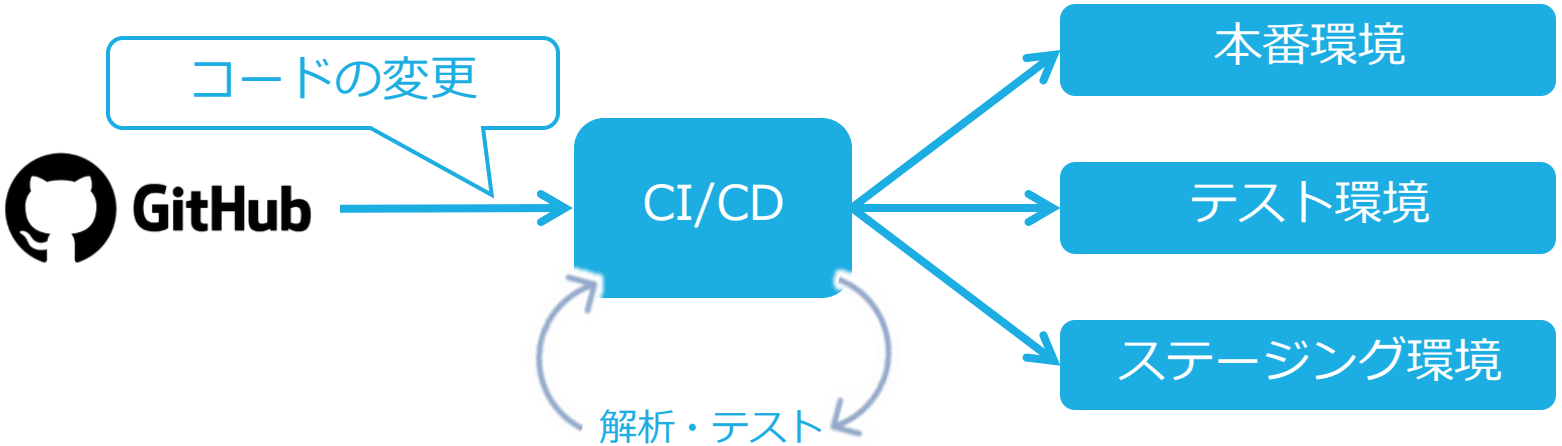
```
basepair --action create-sample --name "Sample 5"
\
--datatype chip-seq --genome mm9 --file1
sample_5.1.fastq.gz

basepair --action create-analysis -w 14 -s 4111 -
s 4112 -c 5112 -c 5113 -c 5114
```

ASTRIDE

CI（継続的インテグレーション）/CD（継続的デリバリー）

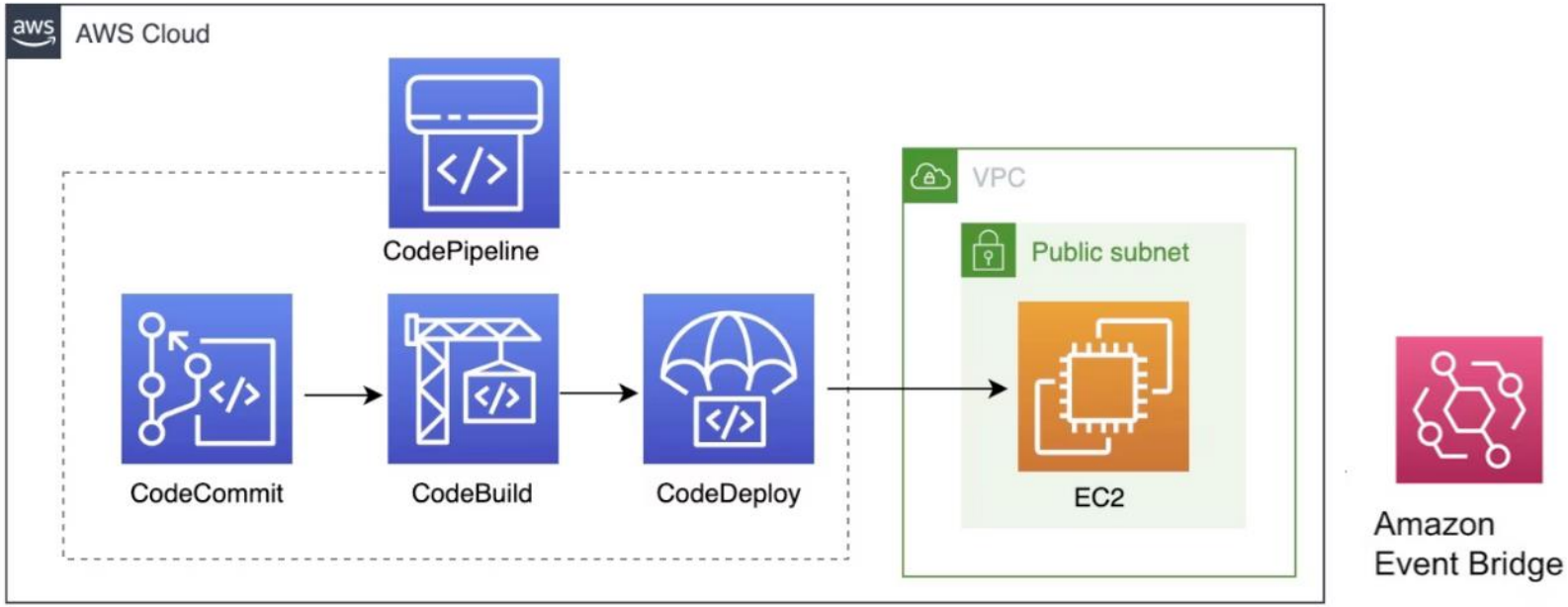
basepair

- ソフトウェアの変更を常にテストし、自動で本番環境に適用できるような状態にしておく、DevOpsを実現するツール



ASTRIDE

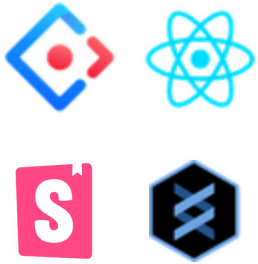
AWSで組むと…


AWS Cloud
CodePipeline
CodeCommit
CodeBuild
CodeDeploy
VPC
Public subnet
EC2
Amazon
Event Bridge

Architecture of basepair

basepair

Frontend



Pipeline

YAML

{wdl} CWL

nextflow

Backend

Database

Computing

Storage

Temporal

Code manage

GitHub

docker

CI/CD

Amazon CloudWatch

ASTRIDE

basepair

Architecture of basepair

Frontend

Pipeline

YAML

{wdl} CWL

nextflow

Backend

Database Computing Storage

Temporal

Code manage

GitHub

docker

CI/CD

Amazon CloudWatch

ASTRIDE

Platform approach

ABOUT US

We are dedicated to realizing the potential of natural killer (NK) cells for the treatment of cancer. Our proprietary technology is designed to harness the power of these important pathogen-fighting immune cells and is uniquely capable of enhancing their ability to search and destroy tumor cells.

NKarta

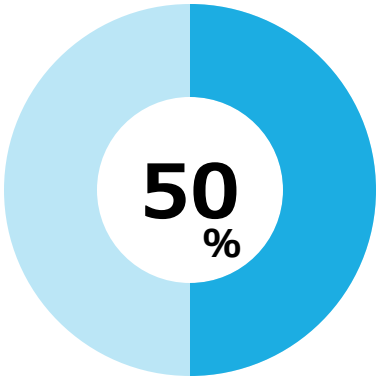
“私たちの小規模チームは時間の約 25% を解放し、より価値のあるデータ マイニング タスクに集中できるようになりました。

さらに、ベンチサイエンティストは今では、小さな反復的なタスクに取り組むよう依頼するのではなく、十分な情報に基づいた質問を持って私のチームにやって来ます。”

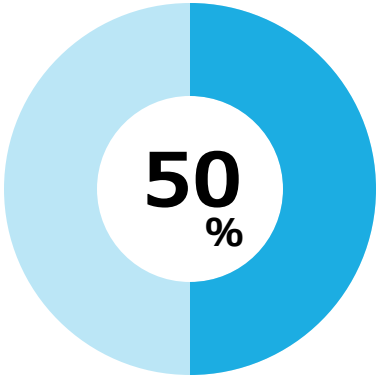
- Sombeet Sahu, Nkarta, associate director of bioinformatics

https://aws.amazon.com/jp/solutions/case-studies/basepair-case-study/

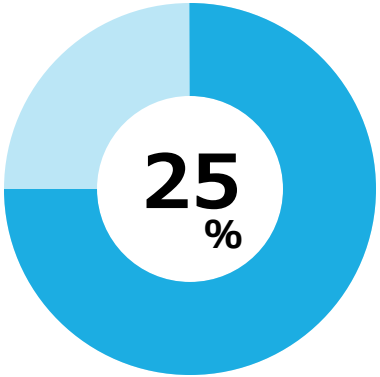
Basepair Impact at NKarta



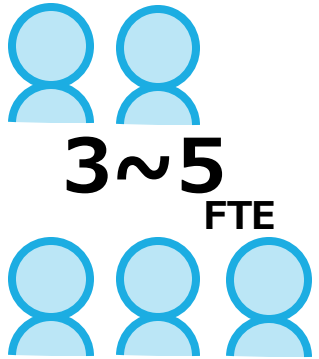
Computer costs saved



Response time improved



Time saved in routine analysis

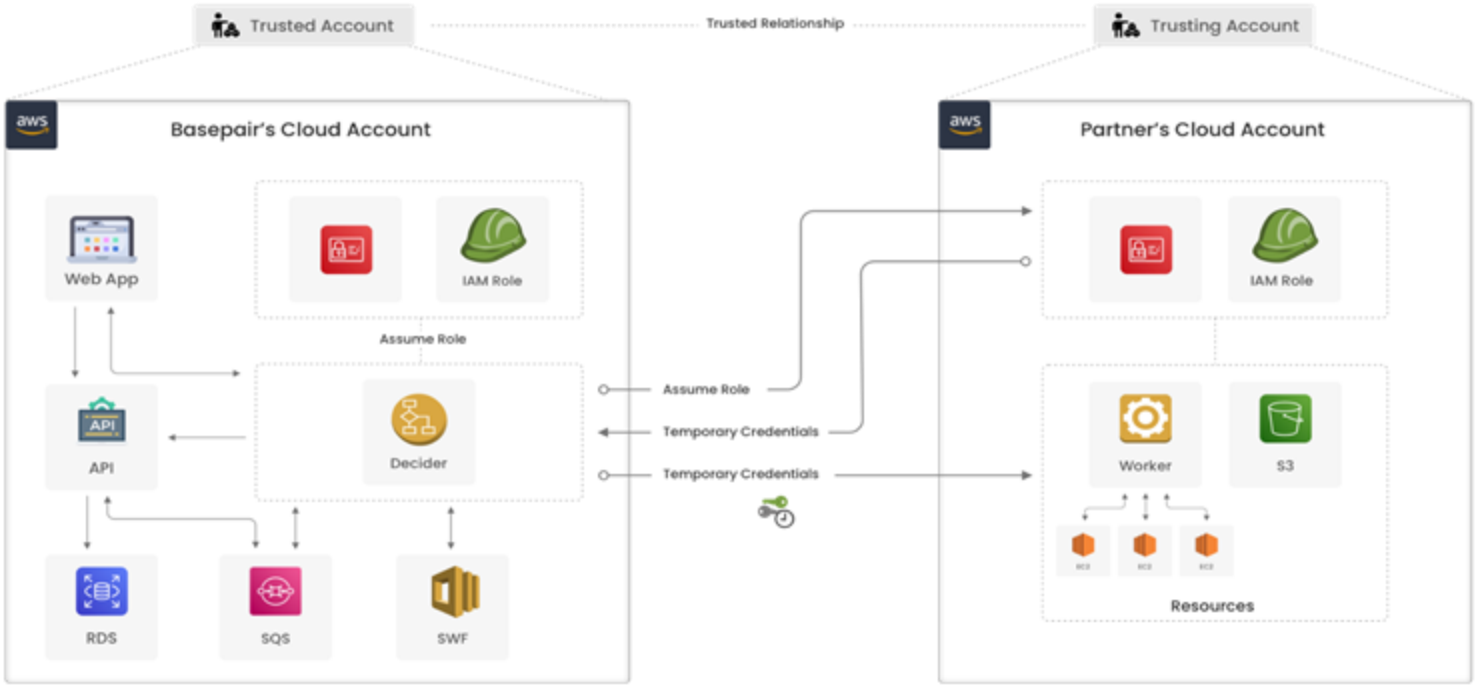


Engineer jobs saved

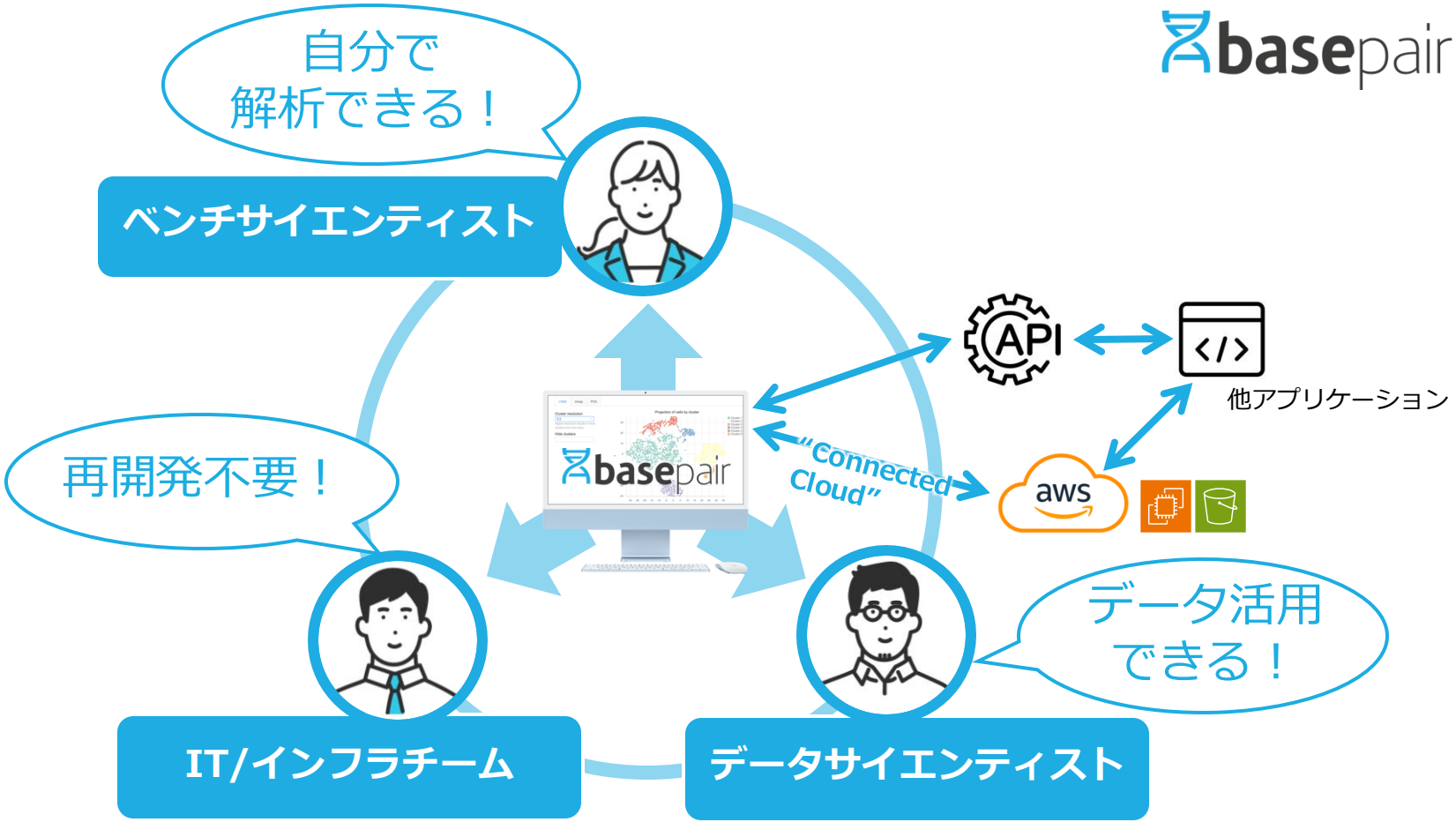
+ Facilitated: scientists' ability to analyze raw data

“Connected Cloud”

basepair



ASTRIDE


basepair
自分で
解析できる！
ベンチサイエンティスト
API
他アプリケーション
basepair
“Connected
Cloud”
aws
再開発不要！
データ活用
できる！
IT/インフラチーム
データサイエンティスト

Democratize Omics Data Analysis with Basepair


basepair


ASTRIDE