

# 実験から論文までのNGS解析サービス

---

株式会社セルイノベーター

土井 淳

〒812-8582 福岡市東区馬出3-1-1

九州大学コラボ・ステーションI

共同実験室4-1

<http://www.cell-innovator.com/jp/>

cell innovator

# 株式会社セルイノベーター

---

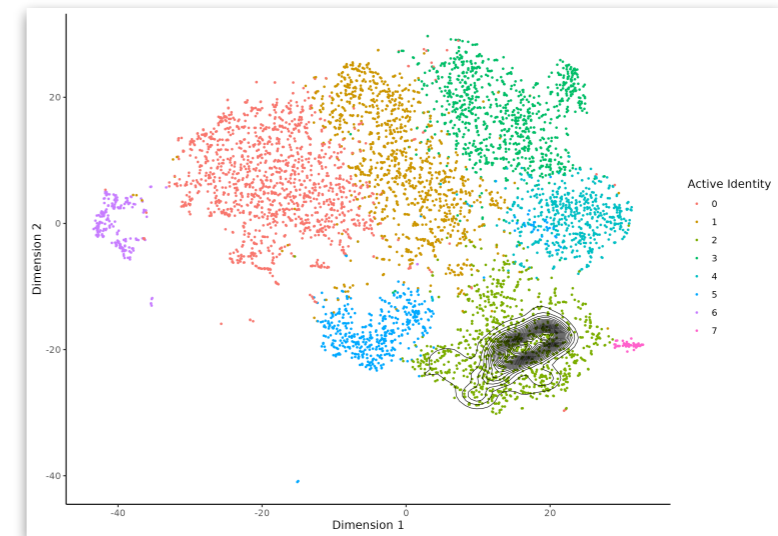
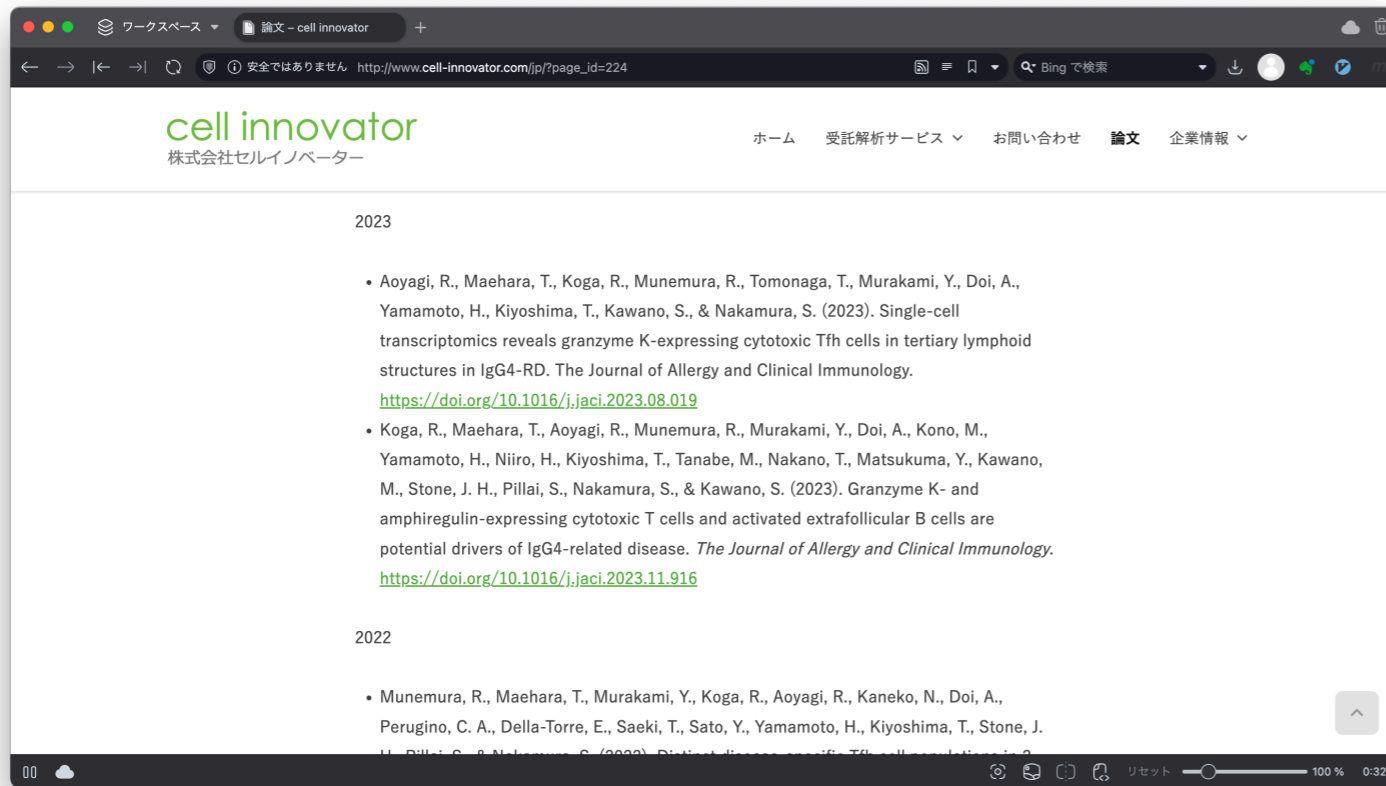
- 九州大学発ベンチャー
  - 九州大学コラボ・ステーションI 共同実験室4-1
- 実験から論文までをサポート
  - 実験：サンプル回収時のご相談から、RNA抽出などに対応。
  - 論文まで：ボルケーノプロットやヒートマップなど、論文のフィギュアの実験、マテメソの記述、GEOへのデータ登録もサポート。
- 自社設備として、DNBSEQ を所有。
  - サンプル数が多い場合、お得なレーン買切りプラン。

# 相談の例

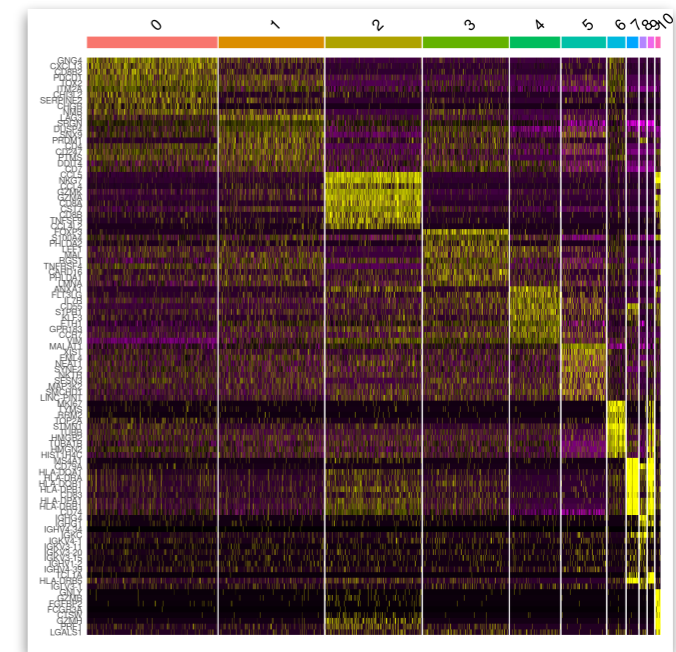
---

- 実験について
  - サンプル回収がうまくいかない
  - RNA が分解している
  - 試薬に何を使ったらいいか
- データ解析
  - 発現変動遺伝子の結果から、何をすべきか？
    - エンリッチメント解析（GO 解析、GSEA 解析）
    - パスウェイ解析
  - ゲノム配列の決定（糸状菌など）

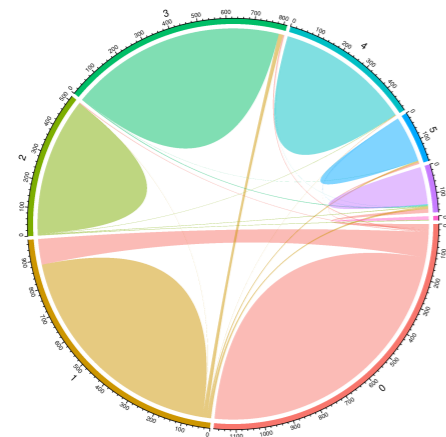
# 実験から論文までのトータルサポート



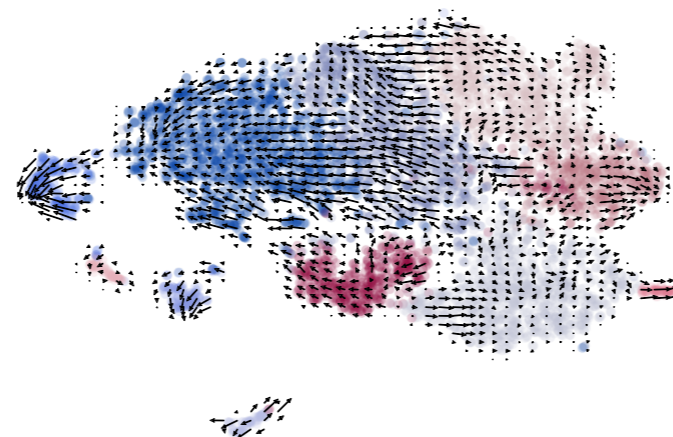
scRNA-seq



cell innovator



TCR レパトア解析



velocity 解析

# 受託解析

---

- 対応している解析
  - RNA-seq, small RNA-seq, BS-seq, ChIP-seq
  - WGS, WES
  - de novo assembly
  - Amplicon sequence, meta genome sequence, etc.
- scRNA-seq
- 空間トランスクリプトーム
- マイクロアレイ (Agilent, Thermofisher)
- SNPアレイ (ジャポニカアレイ)

受託企業としては、すべてに対応できる計算機環境が必要

# 受託解析におけるクラウド活用

---

「論文までをサポート」をかかげるセルイノベーターにおいて、  
下記の3つがクラウドを導入している理由

1. データの保管
2. 結果の再現性
3. 新規手法への挑戦

# 1. データの保管

---

- 「論文投稿までサポート」
  - 博士論文の場合、3年から5年かかることも。
  - 論文投稿時に GEO など公共データベースへの生データ登録が必要。
    - シーケンス結果の fastq は、1サンプルあたり数ギガバイト。
    - トリム後、マッピング後 (bam) => **3倍のデータ量**に。
- バックアップ
  - ファイルのバックアップは、2コピーで十分か？ => **6倍のデータ量**。
  - RAID でもハードディスクの故障とは無縁ではない。
- 委託されたデータのダウンロード納品
  - 期限内にダウンロードできているか？



# クラウドストレージで解決



- たくさんの NAS に囲まれたデスクから解放。
- ハードウェア故障の不安から解放。
- 凍結保存 (Glacier Deep) なら、コストを抑えて、長期間保存。



## 2. 結果の再現性

---

- 同じ結果を再現 => 同じ計算機環境が必要。
  - OS は? Windows, Linux, macOS
  - OS のバージョンは? (18.04 LTS, 22.04 LTS)
  - Samtools のバージョン?
  - scRNA-seq の解析に用いられる R と Python
    - R のバージョンは?
      - bioconductor
      - Seurat v4? v5?
    - Python 2系? 3系?
      - バーチャル環境は、 venv? conda?

1台のサーバーで対応  
しようとするとは困難

# クラウドのバーチャルマシンで解決



- 解析環境構築には時間がかかる。（Docker, Singularity という手もあるが。）
- クラウドなら、マシンイメージ (AMI) として計算時点の環境をそのまま保存。
  - 古い環境でしか動作しないプログラムも解決。

### 3. 新規手法への挑戦

---

- 「この論文と同じ解析をやりたい」
  - 新しいプログラムに対応する必要がある。=> コンパイルから。
- 前述の「同じ解析環境を維持する」とは相反する。
  - 例えば、Python にパッケージを導入する場合。
  - Conda で環境を維持していたとしても、conda install できないことも。
  - パッケージによっては、pip でのインストール。
  - どうなるか？
    - 新たに導入したパッケージは動作するが、これまで動作していた別のパッケージが動かなくなる。=> numpy のバージョンが影響など。
    - 最悪、パッケージの依存関係がバラバラに。=> Conda 環境の再構築。



# クラウドで解決 (AMI イメージ)



- これも AMI イメージであれば、既存環境を残したまま、新しいパッケージにチャレンジできる。
- 全く別の環境なので、パッケージの依存関係も気にしなくていい。
- 一度、構築した環境は、そのまま保持しておける。

# クラウド環境の利用

---

- outer\_space: 自社開発のクラウド環境
- AWS 上に構築
  - RNA-seq や、WGS など必要に応じたインスタンスを起動。
  - 複数サンプルを20台のインスタンスを立ち上げて並列処理。
  - メモリを多くする計算にも、256GB メモリのインスタンスで対応。
  - 計算結果は、クラウドストレージ (S3) に保存。
  - 解析環境を AMI イメージとして保存。 => 後日再現するため。
  - スポットインスタンスを使用して、コスト削減。

# クラウドを利用した解析環境

Amazon マシンイメージ (AMI) (29) 情報

自己所有

属性またはタグ でAMI を検索

Name	AMI名	AMI ID	ソース
bs-seq	bismark_hisat2_gd	ami-04	
outer_space	bwa_hisat2_mapping20190417	ami-0e	
outer_space	VEP_offline for c5	ami-00	
	deg_report	ami-01	
outer_space	RSEM_AMI_for_c5_20220719	ami-05	
snpeff_202405	snpeff_AMI_mm10_hg19	ami-0f	
signac	R_Signac_ready_with_Vnc20240502	ami-06	
signac	R_Seurat_ready_with_Vnc20230722	ami-0d	
outer_space_b...	outer_space20220722	ami-0f	

The screenshot shows the OUTER SPACE web interface. The main content area displays a form for submitting a sequence job. The form includes a title 'ticket0000/WT', a description 'sequence job submission', and a 'Script parameters (url)' field. Below the form are several buttons: 'HardWork', 'Download to S3', 'SRA to S3', 'QC+Trimming', 'Mapping\_hisat2 OR Mapping\_bwa', 'Count OR Variant Call', and 'VEP'. There are also checkboxes for 'with --no-spliced-alignment (HISAT2)', 'with --ploidy=1 (bcftools)', and 'keep instance alive (for debugging)'. A dropdown menu for 'with reference =' is set to '--'. On the right side, there is a 'SequenceJobs (133)' section listing several jobs with their status, options, and submission times.

解析用マシンイメージ

ジョブ実行用インターフェイス

# 1からクラウド環境を作成する必要はありません

---

- そう、 **Basepair** ならね。
- とりあえず、発現変動遺伝子を確認したい。
  - GSEA の結果も自動で生成。
- できるだけコストを抑えたい。
  - 1サンプルから解析可能。
- その後の解析に不安。。。
  - セルイノベーターにて、データ解析のサポートプランあり。



# まとめ

---

- 論文までのサポートを実現するには
  - データの保管
    - => クラウドストレージの活用 (S3)。
  - 結果の再現性
    - => 仮想マシン (AMI) による解析環境の保存。
  - 新規手法への挑戦
    - => 仮想マシン (AMI) によるパッケージの依存関係に悩まされない環境構築。
- クラウド環境としての **Basepair**
  - セルイノベーターにて、サポートプランあり。