

# GenomeDataAnalysis7

大阪大学大学院医学系研究科 遺伝統計学  
東京大学大学院医学系研究科 遺伝情報学  
理化学研究所生命医科学研究センター システム遺伝学チーム

<http://www.sg.med.osaka-u.ac.jp/index.html>

## GenomeDataAnalysis7

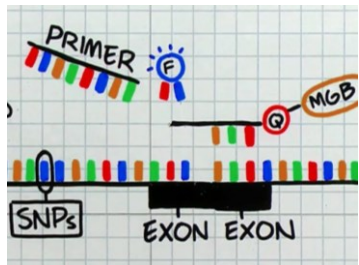
① シングルセル解析技術と情報解析

② Seuratを使ったシングルセル解析実習

本講義資料は、Windows PC上で  
C:¥SummerSchoolにフォルダを配置すること  
を想定しています。

# ① シングルセル解析技術と情報解析

## TaqManアッセイ



## SNPマイクロアレイ



## ショートリードNGS



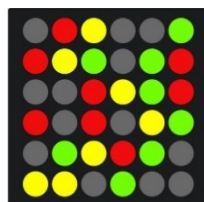
## ロングリードNGS



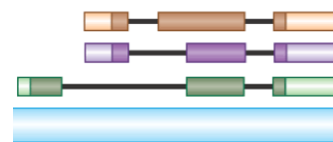
## RT-PCR



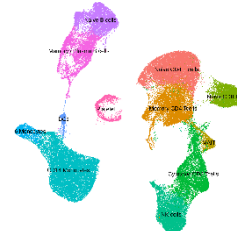
## 遺伝子マイクロアレイ



## RNA-seq



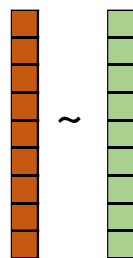
## シングルセル解析



## 分割表検定

	AA	AG	GG
Case	360	480	160
Control	250	500	250

## 回帰分析



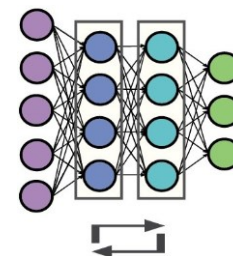
## 機械学習

Variants (90 K)

ACTGAACGCAAAC
CCAGTATTCTACCT
CATGACTGCAAAC
ACAGTATGCTACAT
CCTAAATGCTACCT
AAAGTCTTCTACCT
ACTATATGCTACCT

Individuals (170 K)

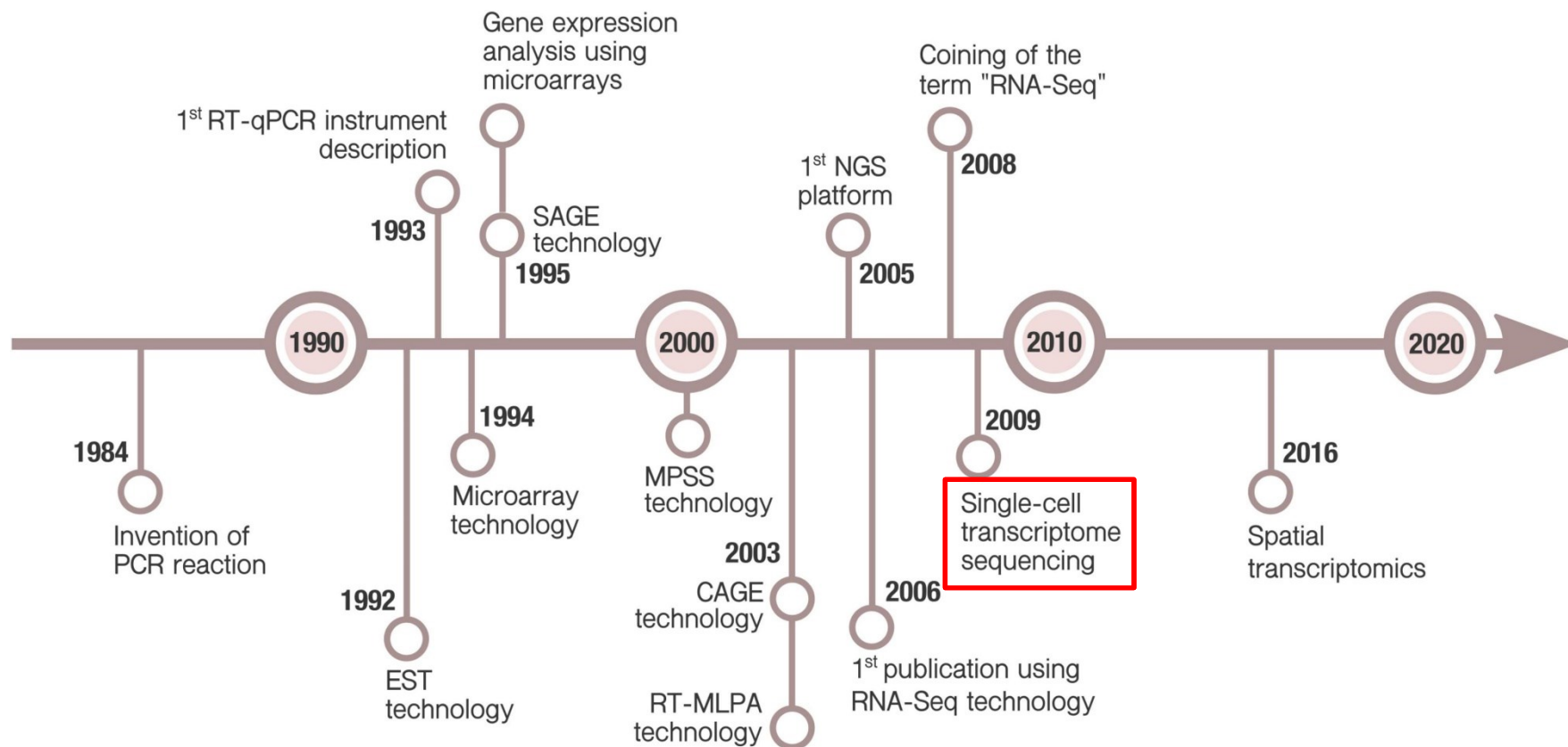
## 深層学習



- 観測・解析技術の進歩は、常に予想を上回る速度で進む。
- Wet・Dry双方の最新解析技術の先進的導入が生命科学に不可欠。
- 実験・解析原理の正確な理解が、革新的な研究を可能にする。

# ① シングルセル解析技術と情報解析

## 複数遺伝子発現量の同時定量解析手段の発展

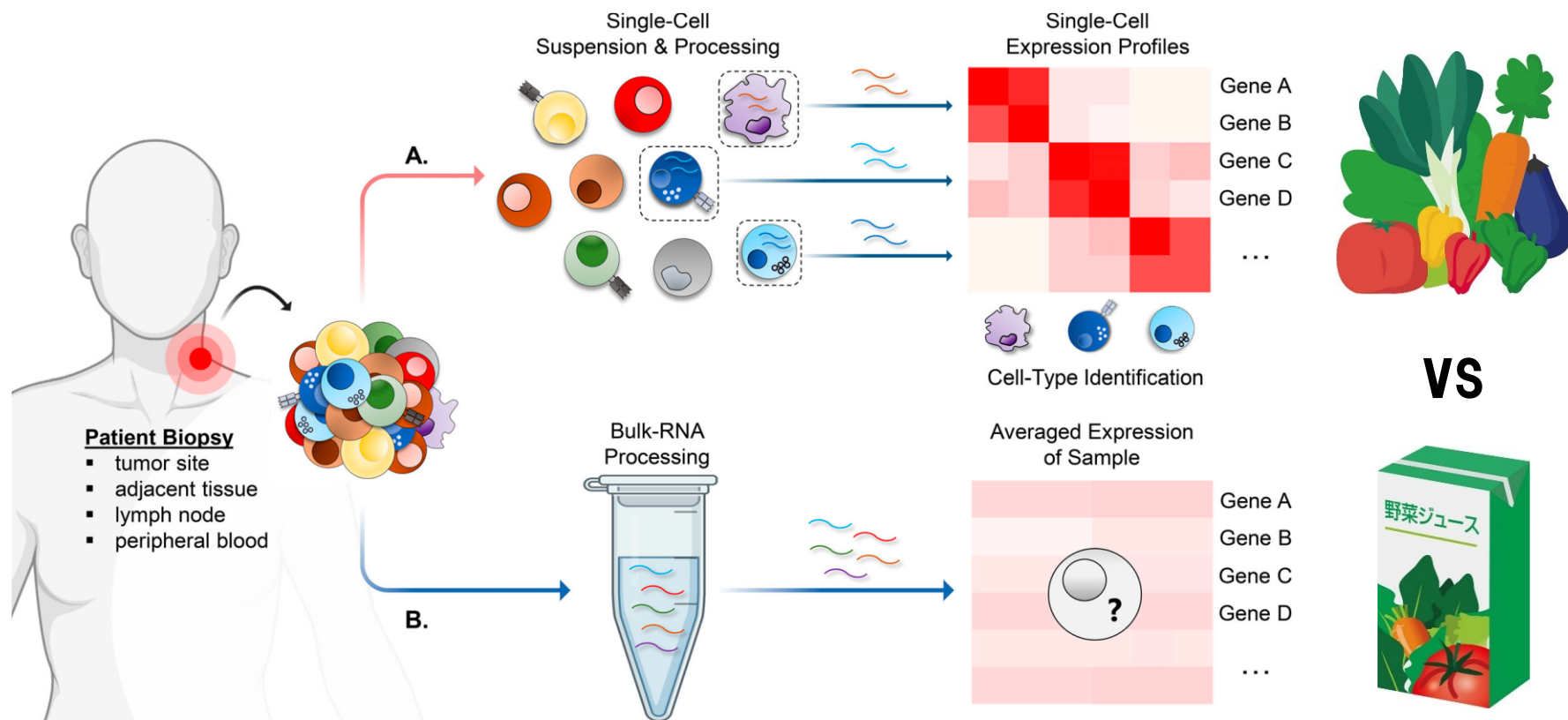


- 対象サンプルの**複数遺伝子の発現量を同時に定量化**する解析は、生命現象に伴う遺伝子動態の知見を得る上で、重要な手段です。
- 解析手段の一つとして、**シングルセル解析**が注目を集めています。



# ① シングルセル解析技術と情報解析

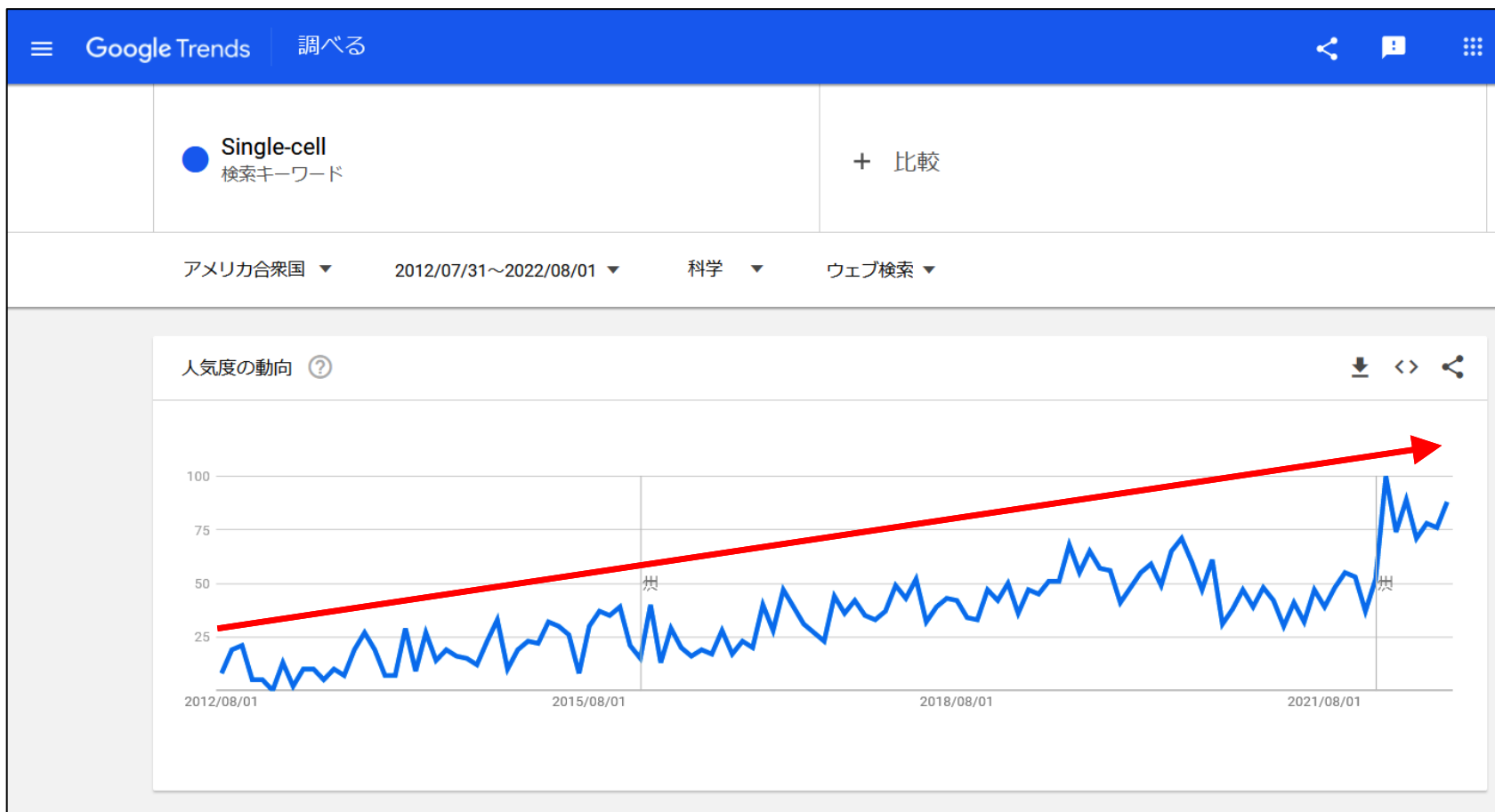
## Bulk RNA-seqとsingle cell RNA-seqの違い



・組織全体で平均化された遺伝子発現量を計測する従来のバルク(bulk)解析と異なり、シングルセル解析では、個別の一細胞における遺伝子発現量が観測可能になります。

# ① シングルセル解析技術と情報解析

## 過去10年間の”single cell”の使用人気度(by Google Trends)



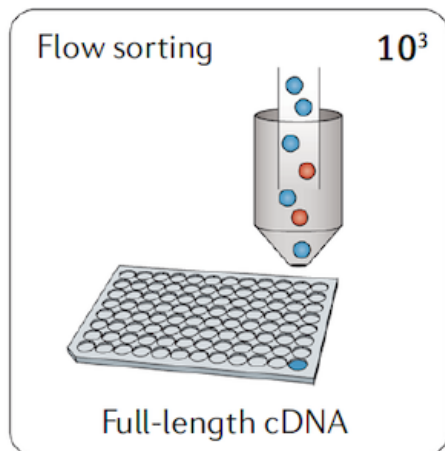
- 革新的な解析技術として、シングルセル解析の重要性は年々高まってきています。
- 生命科学のいずれの分野においても、向き合う必要性が生じています。

# ① シングルセル解析技術と情報解析

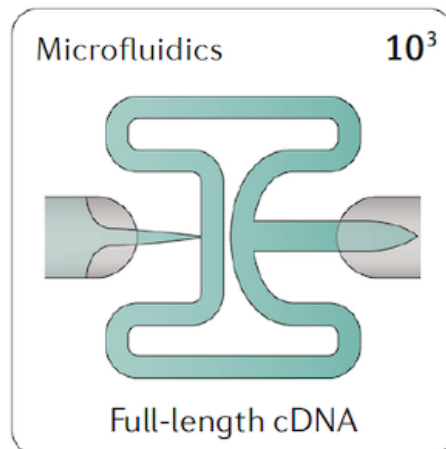
## シングルセル解析技術の進化

細胞分離法

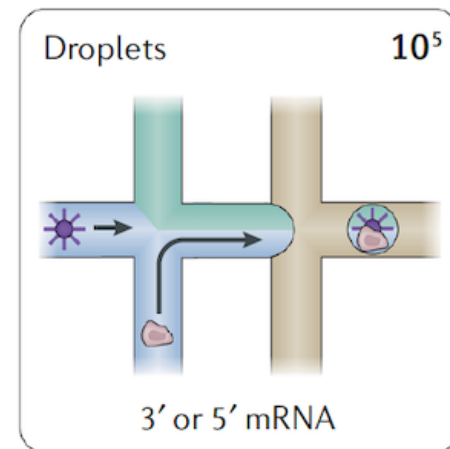
チューブ・  
ウェルプレート



微細流路トラップ  
(C1)



ドロップレット封入  
(Chromium)



1細胞の選択

○

○

×

検出遺伝子数

多い

多い (C1)  
少ない (C1 HT)

少ない

解析可能細胞数

少ない

最大 96/plate (C1)  
最大 800/plate (C1 HT)

1,000 -  
10,000

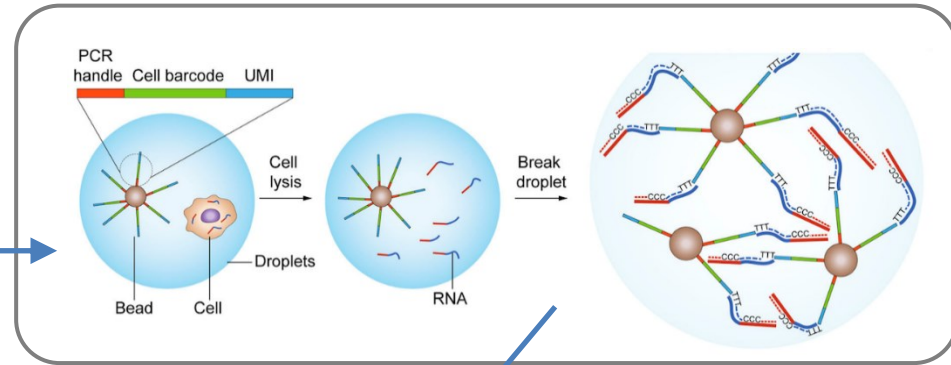
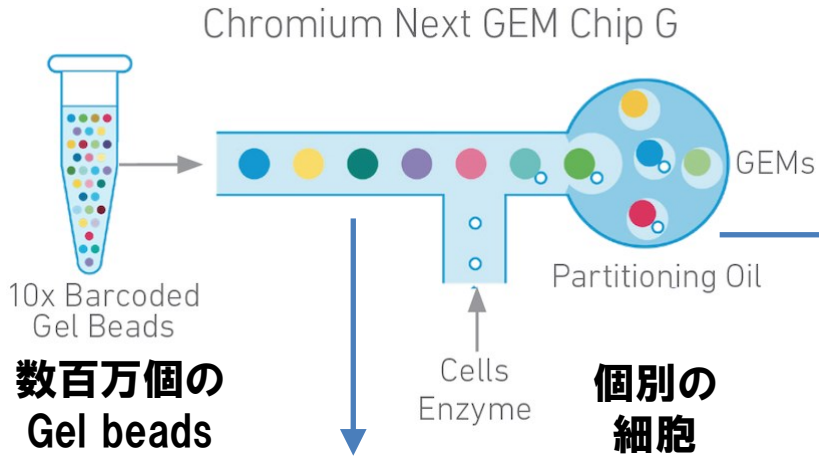
- シングルセル解析技術は、日々進化を遂げています。
- 最近、**ドロップレット(泡)**の中に一細胞を取り込む方法が主流です。

# ① シングルセル解析技術と情報解析

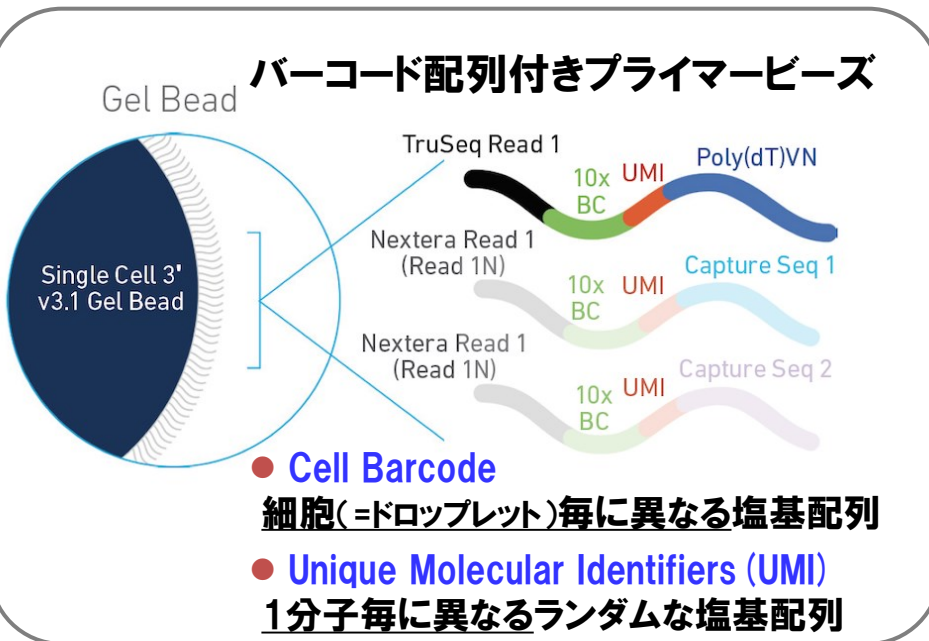
## ドロップレット封入型シングルセル解析(10X Genomics社Chromium)

~90-99%のBeadsは細胞封入なし

1 Beadsに1細胞を封入



### バーコード配列付きプライマービーズ



Cell barcode UMI cDNA (50-bp sequenced)

```

AAATTATGACGATGTGCTTG ..... GACTGCAC
COTTAGATGCGCAGGCCCGG ..... CTCATAGT
GACTACGAGTTAGTTTGT ..... GCTCATAA
GTTAAAGCTACCTAGCTGT ..... GATTTTCT
ACGTCACCTTTGTGGGGT ..... ATAAAGCTC
TTCCCTGTGTTATGGAGG ..... CCAGCAC
AGTCCATGTGCGGCAGGTTT ..... GTTGGCOT
AAATTATGACGAGTTTGT ..... AGATGGGG
CCAAAGATGTCCTTAGGCT ..... GGGGACGA
GTTAAAGCTACCAAGGCTT ..... CAAAGTTC
TTTTGACCACTCGTAGGG ..... TTCCAAGG
ACTGTCCATGCCCTGTGTA ..... TTGTCCT
CGTAAACATATACCGGTG ..... TTAACCG
    
```

(Hundreds of millions of reads)

cDNA alignment to genome and group results by cell

```

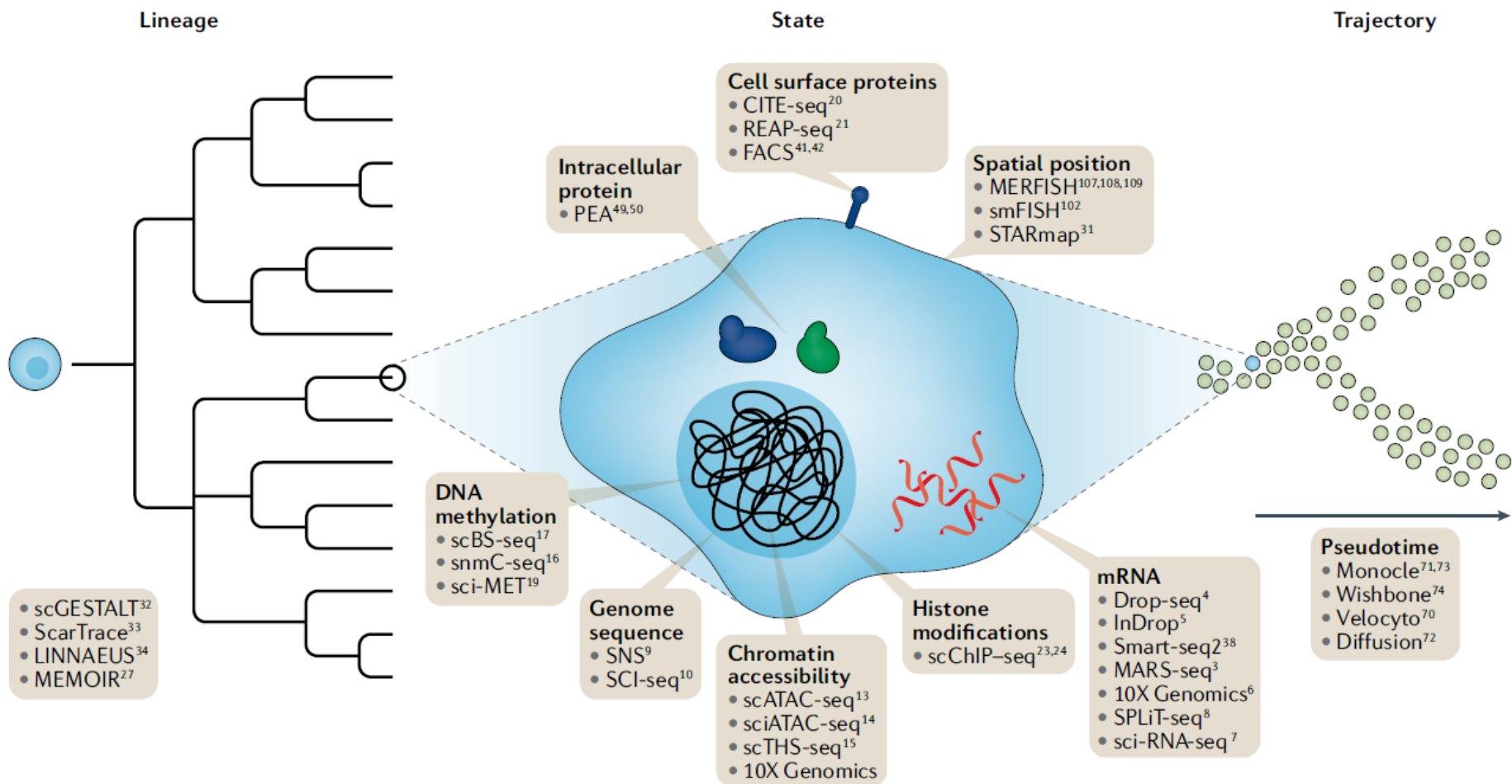
Cell 1 { TTCCCTGTGTTATGGAGG ..... CGGTGTA } DDX51
        { TTCCCTGTGTTATGGAGG ..... CCAGCAC } NOP2
        { TTCCCTGTGTTATGGAGG ..... AAATGTC } ACTB
Cell 2 { COTTAGATGCGCAGGCCCGG ..... CTCATAGT } LBR
        { COTTAGATGCGCAGTTATA ..... ACGGTAC } ODF2
        { COTTAGATGCGCAGGATT ..... AGCCTTT } HIF1A
Cell 3 { AAATTATGACGAGTTTGT ..... GGGGATTA } ACTB
        { AAATTATGACGAGTTTGT ..... GACTGCAC } RPS15
Cell 4 { GTTAAAGCTACCTAGCTGT ..... GATTTTCT } GTPBP4
        { GTTAAAGCTACCGAGAAGT ..... GTTGGCT } GAPDH
        { GTTAAAGCTACCAAGGCTT ..... CAAAGTTC } ARL1
        { GTTAAAGCTACCTCCGCTC ..... TCCAGTCG }
    
```

(Thousands of cells)

Cell Barcodeでグループ化し、cDNAアライメント

# ① シングルセル解析技術と情報解析

## シングルセル解析技術で観測可能なヒトオミクス情報

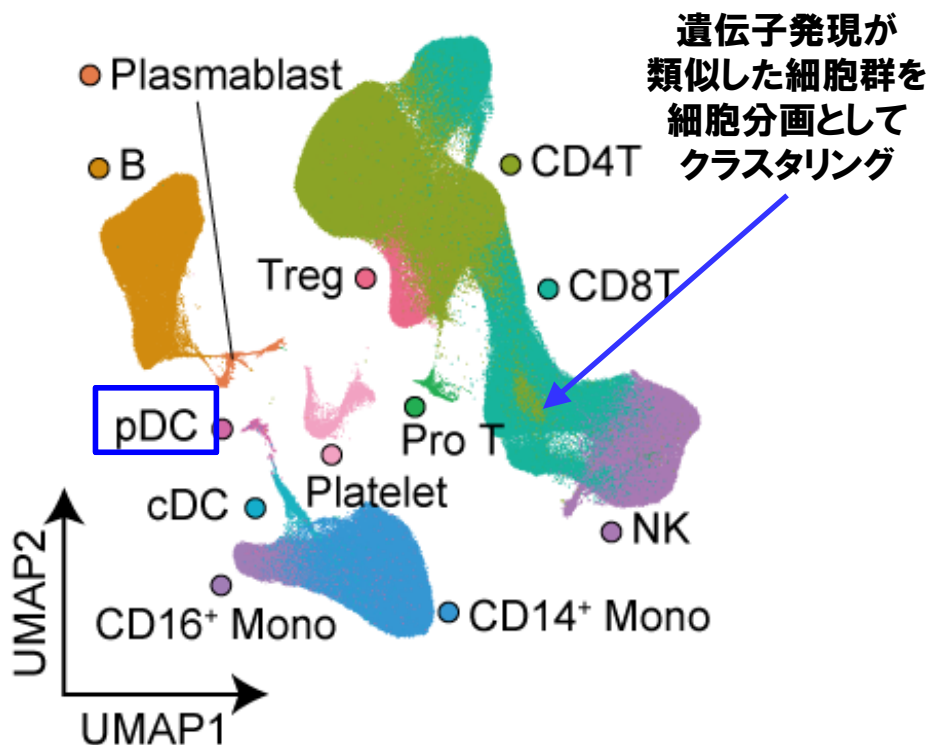


• 遺伝子発現だけでなく、エピゲノム修飾やタンパク質など**多層的なオミクス情報**も、一細胞解像度で観測可能になっています。

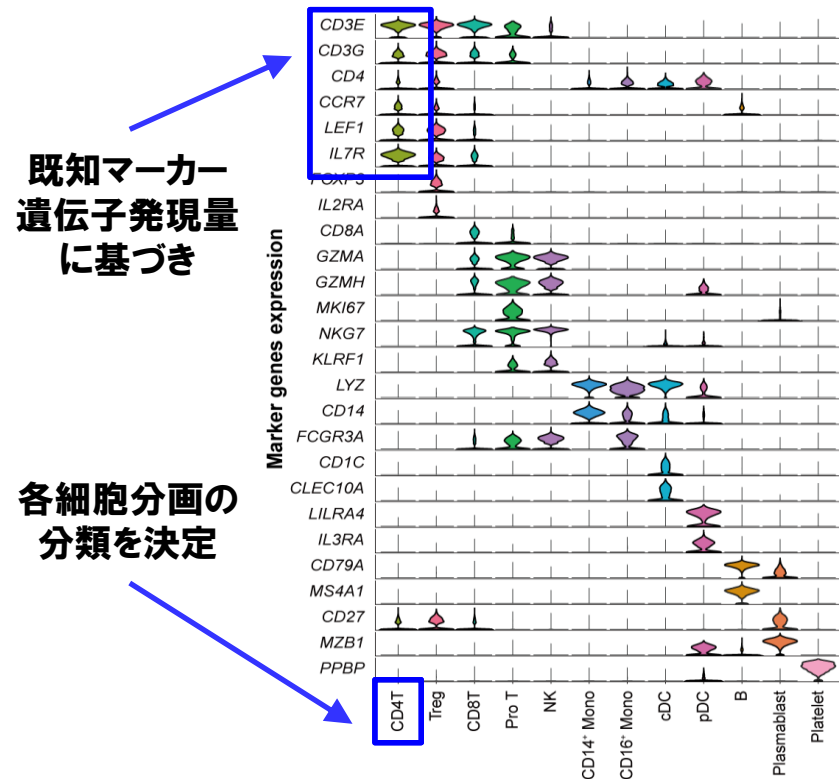
# ① シングルセル解析技術と情報解析

## シングルセル解析による末梢血液中の細胞分画の同定

### 末梢血PBMCの細胞分画



### 細胞分画の遺伝子発現プロファイル



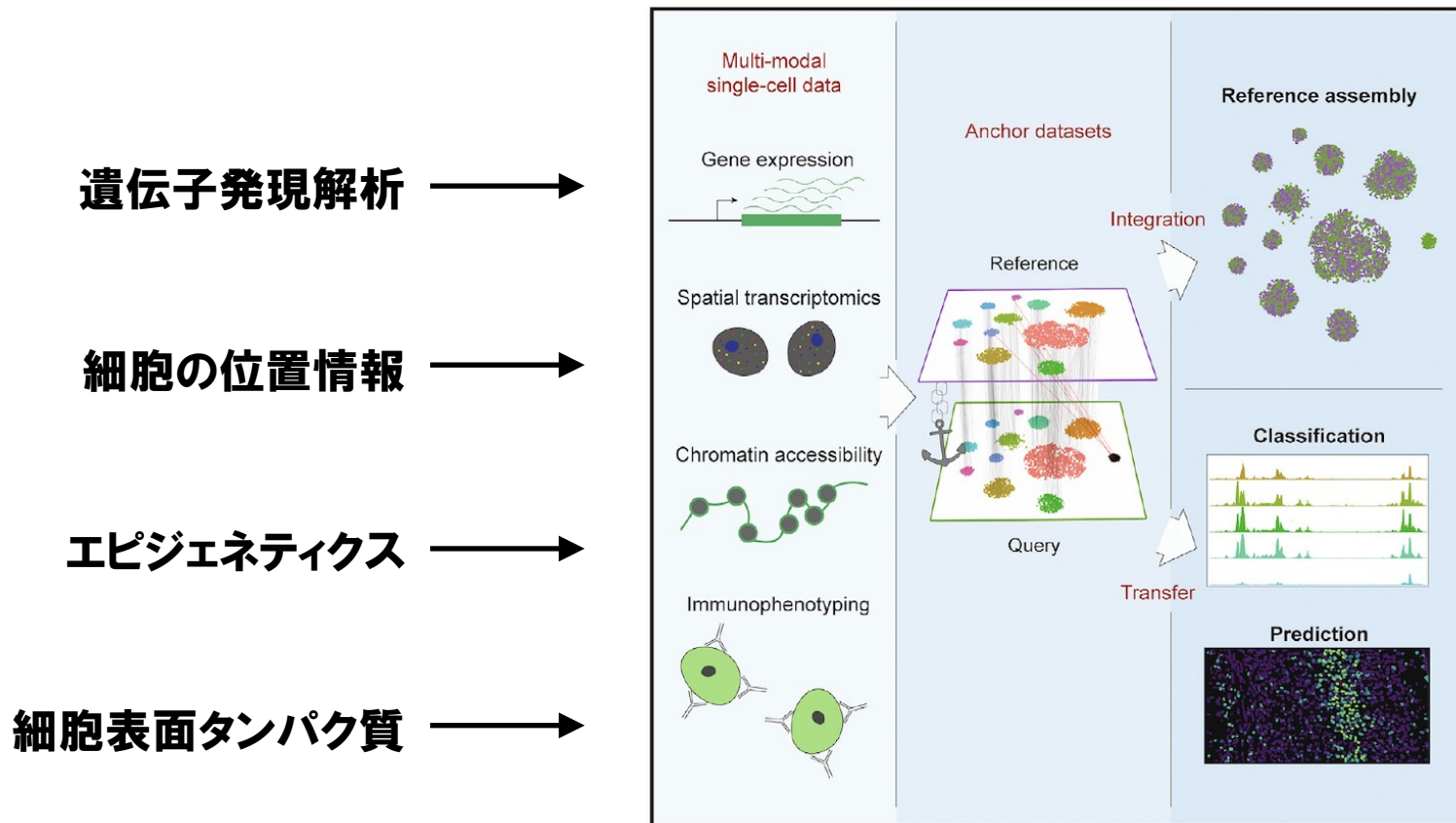
• シングルセル解析で得られた、各細胞の遺伝子発現プロファイルを参照することで、細胞分画の詳細な分類が可能になります。

• 数が極めて少ない細胞分画(例: plasmacytoid dendritic cell; pDC)や、新規細胞分画の同定も可能になります。(Namkoong H and Edahiro R et al. *Nature* 2022)



# ① シングルセル解析技術と情報解析

## シングルセル・オミクス統合解析による細胞動態解明

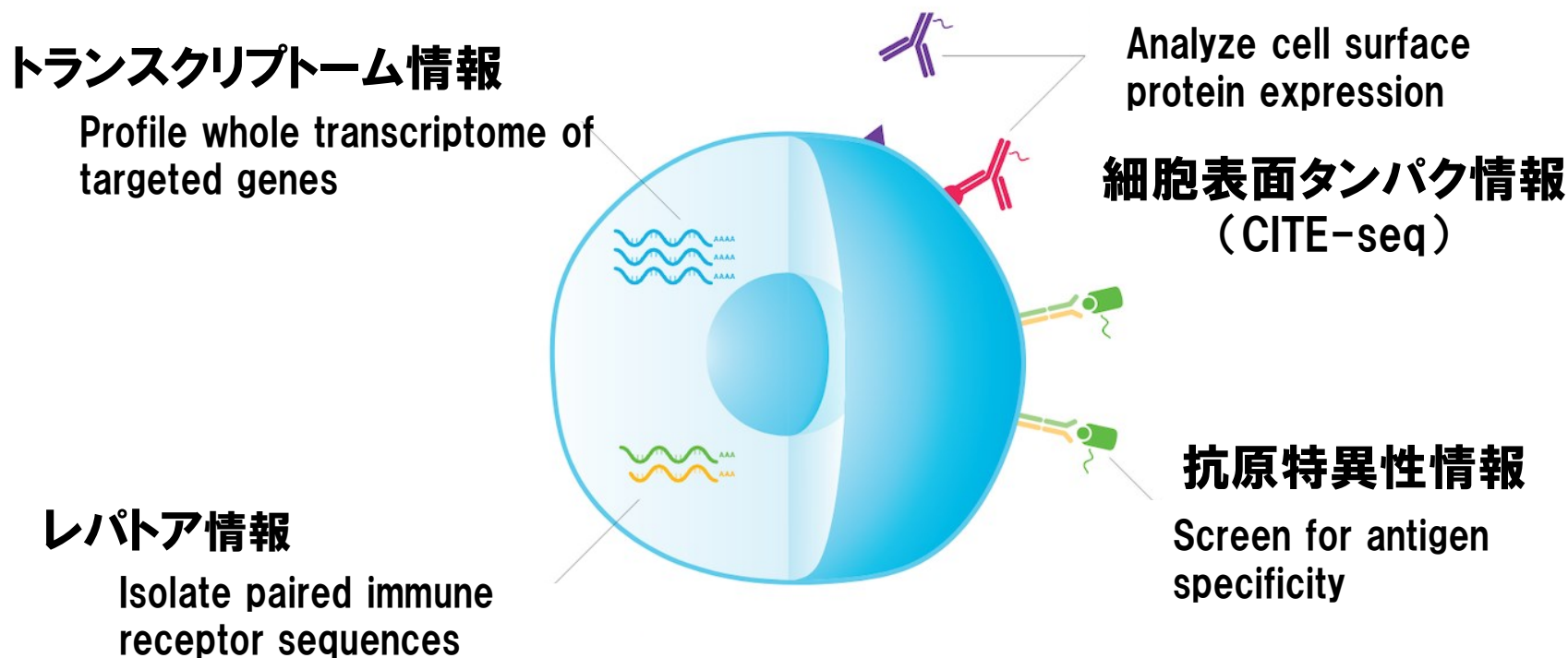


- シングルセル遺伝子発現解析を、異なるシングルセル解析データと統合することで、詳細な細胞動態解明が可能になります。
- 一細胞解像度オミクス情報の統合解析ツールの開発も進んでいます。

(<https://satijalab.org/seurat/>, Stuart T et al. *Cell* 2019)

# ① シングルセル解析技術と情報解析

## シングルセル・オミクス統合解析による細胞動態解明

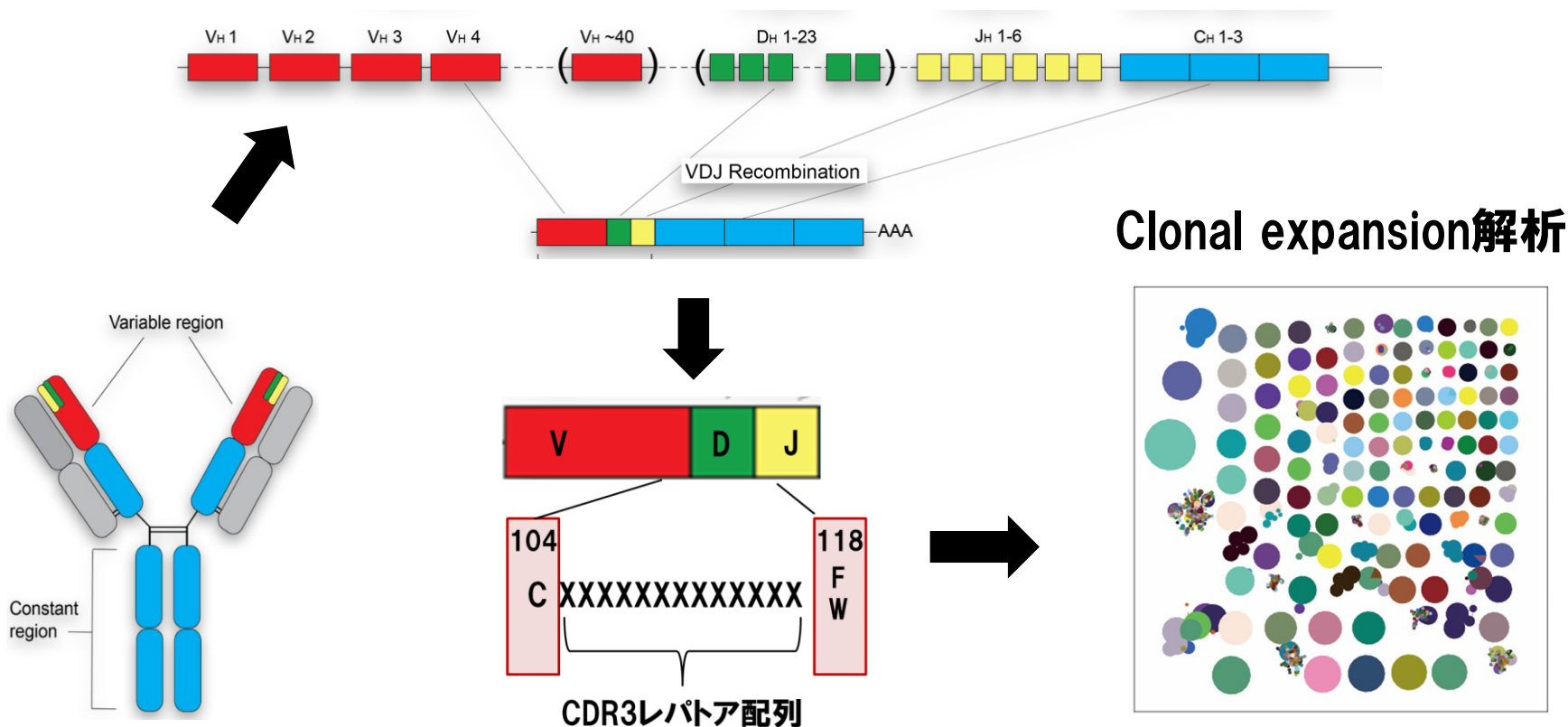


- 同一の細胞サンプル群から、複数のオミクス情報を同時に取得可能なシングルセル解析技術が実現化しつつあります。
- 同一のシングル細胞に対する複数のオミクス情報を統合することで、より詳細な細胞分画分類や細胞間ネットワークの解明が可能になります。



# ① シングルセル解析技術と情報解析

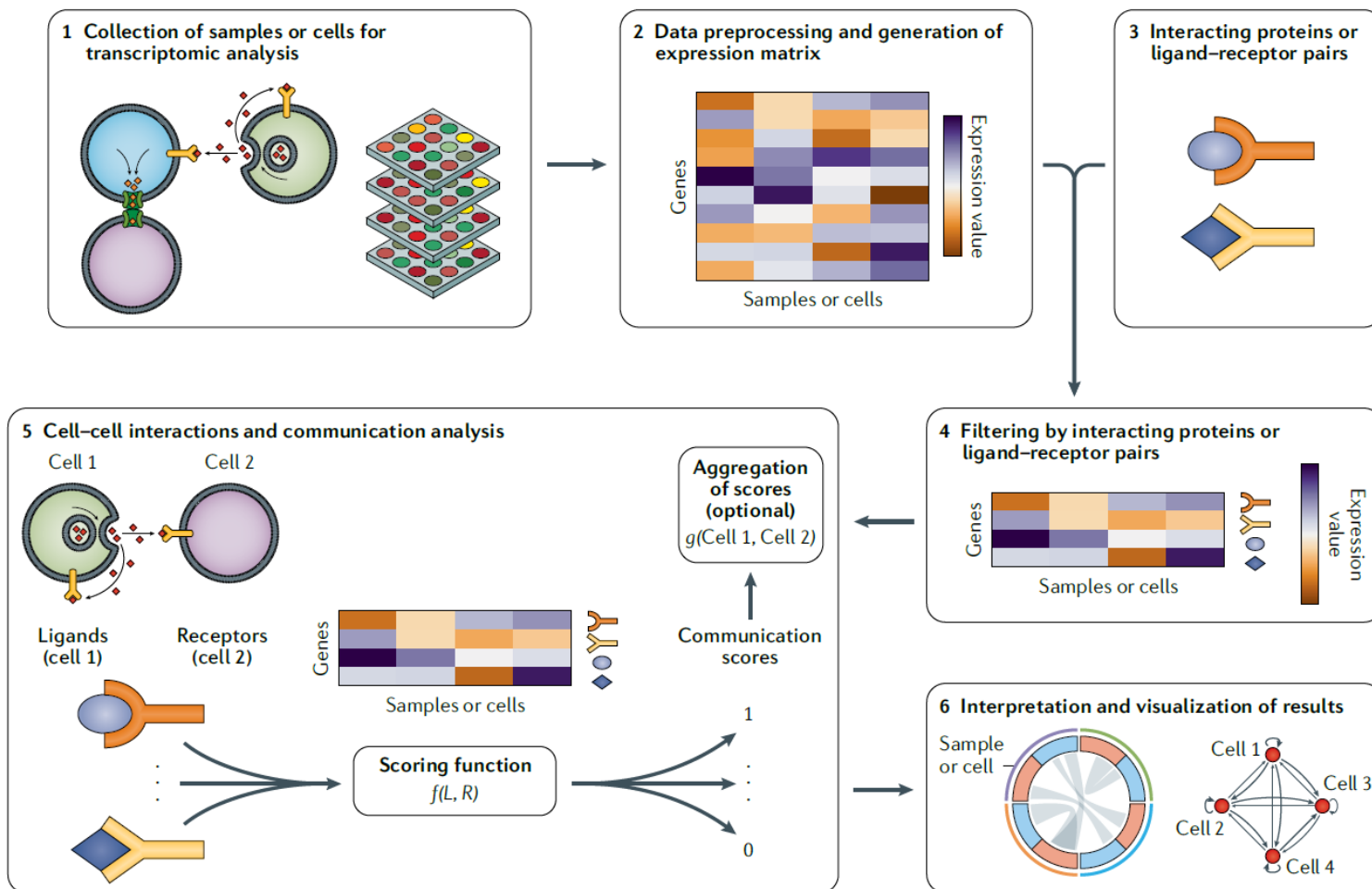
## T細胞・B細胞受容体CDR3配列の解読



- T細胞・B細胞受容体の抗原特異性を決定するVDJ領域のCDR3配列をシングルセル解析(5' 端対応版)による網羅的解読が可能。
- 「どの疾患のどのT/B細胞分画でどのCDR3配列がclonal expansionしているのか」を容易に同定することが可能になりました。

# ① シングルセル解析技術と情報解析

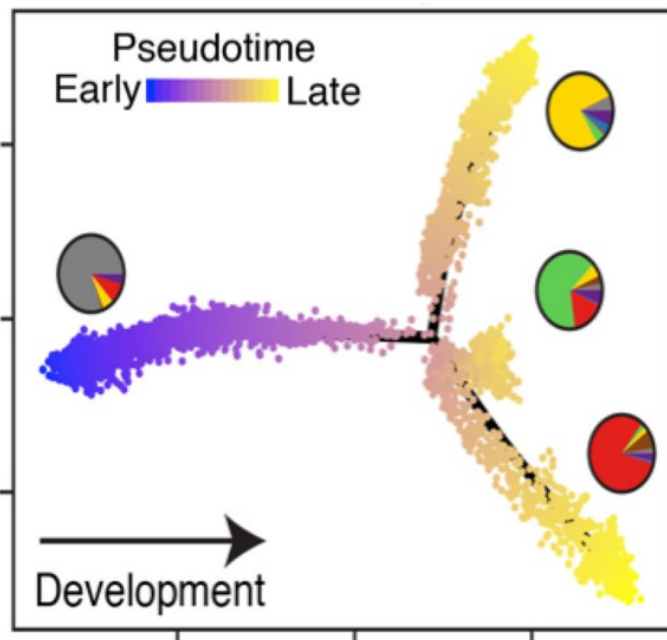
## シングルセル情報を活用したcell-cell interaction network解析



• 受容体-リガンドや下流パスウェイ遺伝子発現が連動する細胞分画ペアに基づき、**cell-cell interaction network**の間接的な観測が可能。

# ① シングルセル解析技術と情報解析

Pseudotimeによる細胞分化予測



RNA velocityによる細胞分化予測

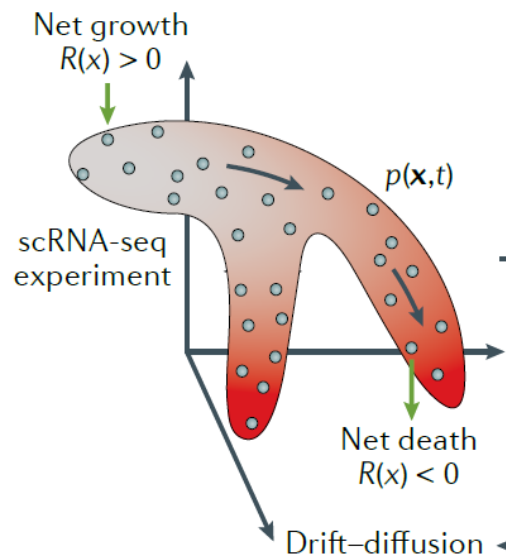


- シングルセル解析を利用して、**細胞分化や推移状態を推測するための Trajectory inference (軌道推定)**の解析手法が多く開発されている。
- 細胞間の不均一性を、疑似的に分化時間軸に投影、と解釈される。
- 遺伝子発現量変化の類似性に基づく **pseudotime** や、unspliced/spliced mRNA発現量比に基づく **RNA velocity** が有名です。

# ① シングルセル解析技術と情報解析

Waddington's landscape model

High-dimensional phase space,  $x$



Potential energy  $F(x = \{x_1, x_2, \dots\})$

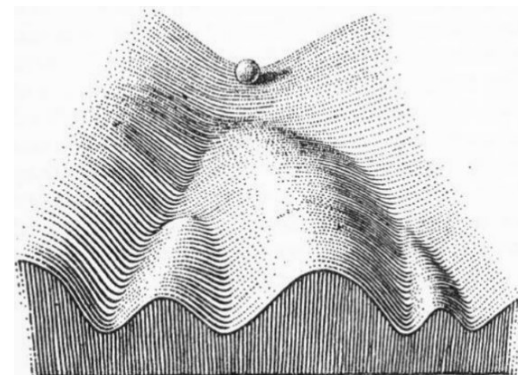
$$\frac{\partial p}{\partial t} = 0$$
$$D, R(x)$$

State variable,  $x_1$

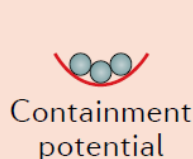
Markov chain equivalence

State variable,  $x_2$

Cell-cell graph



$$F(x) = V(x) + U(x)$$



## Advantages

- Works for a high-dimensional phase space
- Can easily capture more general dynamics (e.g. multifurcations)
- Stochasticity and cell-fate probabilities

## Limitations

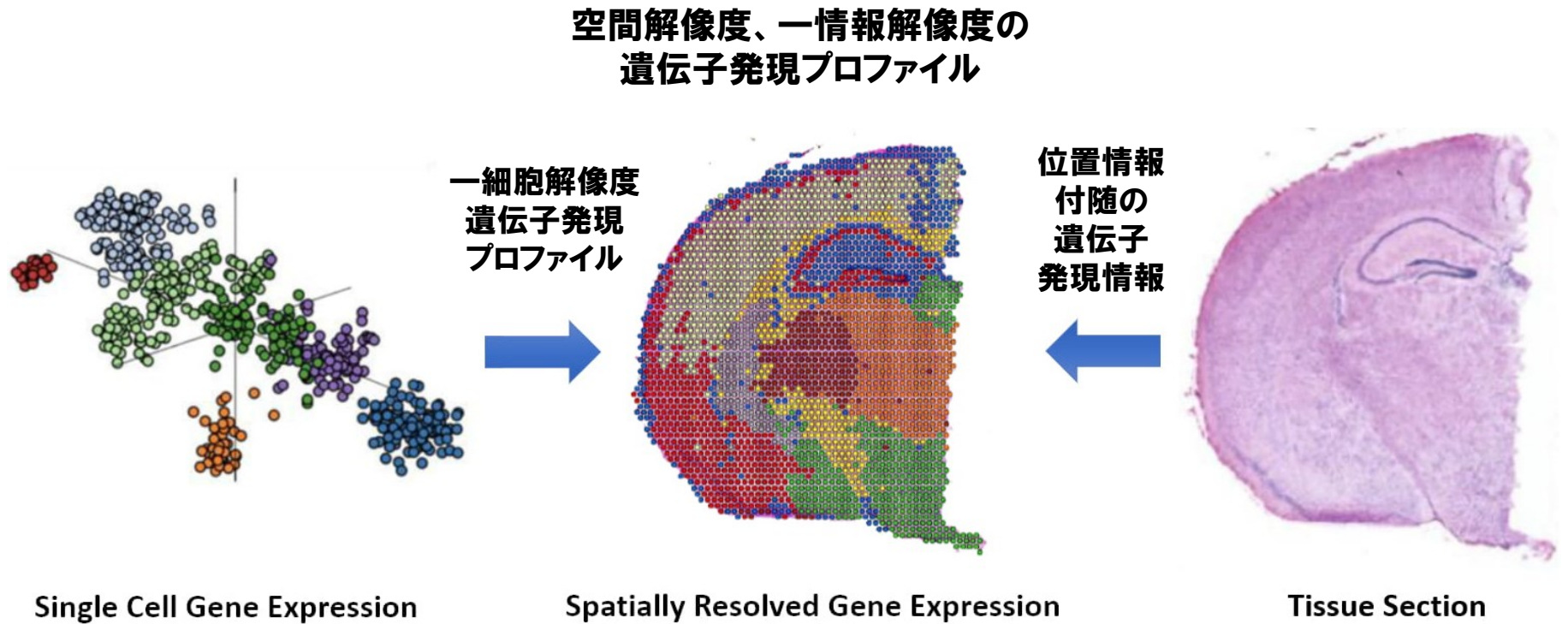
- Still requires some parameter input
- Sensitive to choice of phase-space coordinates
- Steady-state assumption
- $V$  is not modelled from the bottom up

• シングルセル解析情報と数理モデルを統合することで、未病状態から疾患発症へと至る動的な過程をモデル化する試みも始まっています。



# ① シングルセル解析技術と情報解析

## 空間トランスクリプトームとシングルセル解析の統合



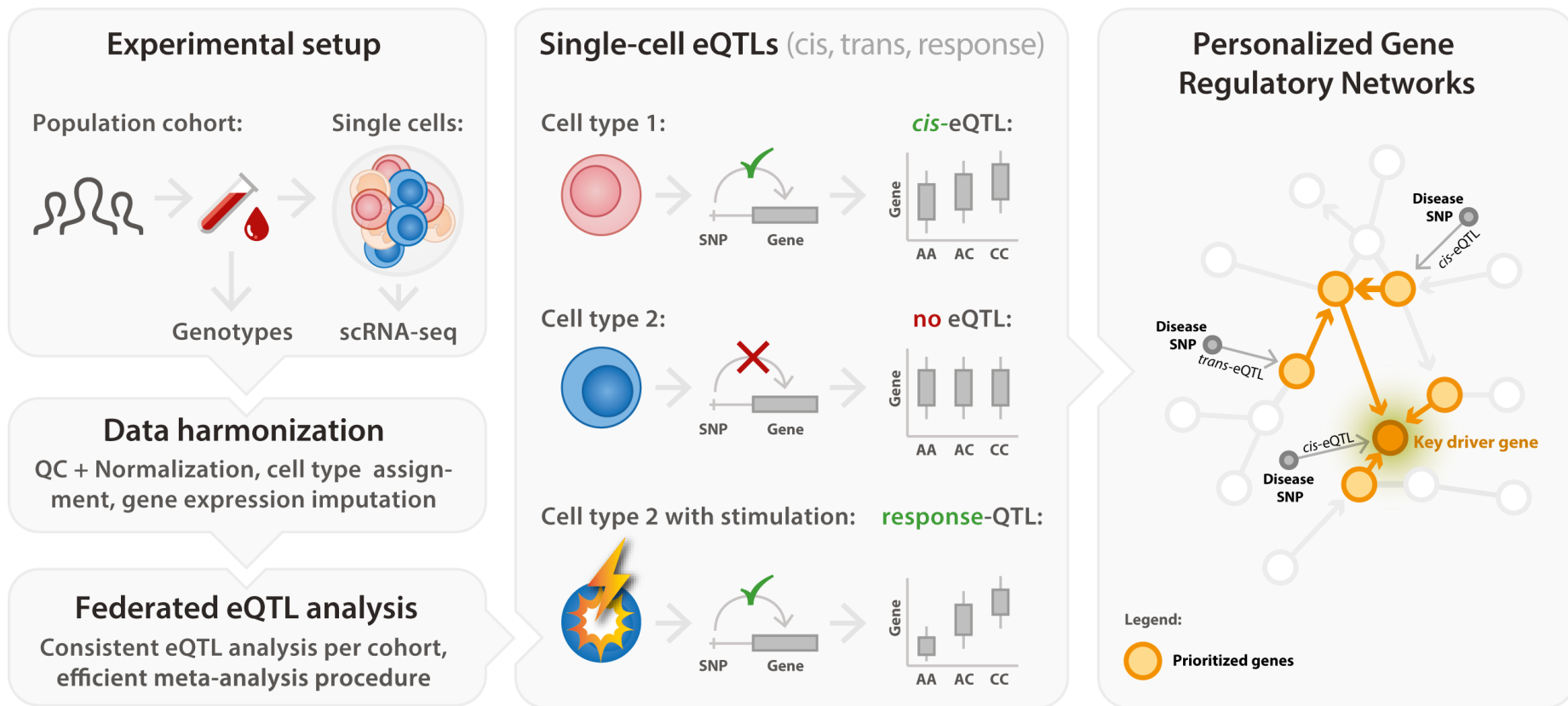
- 空間トランスクリプトーム (spatial transcriptome) とシングルセル解析の統合により、**遺伝子発現の時空間分布**の定量化が可能になります。

(<https://sites.dartmouth.edu/cqb/>, Longo SK et al. *Nat Rev Genet* 2021)<sup>17</sup>

# ① シングルセル解析技術と情報解析

## 集団ゲノム情報とシングルセル解析の統合

### Single cell eQTLGen Consortium

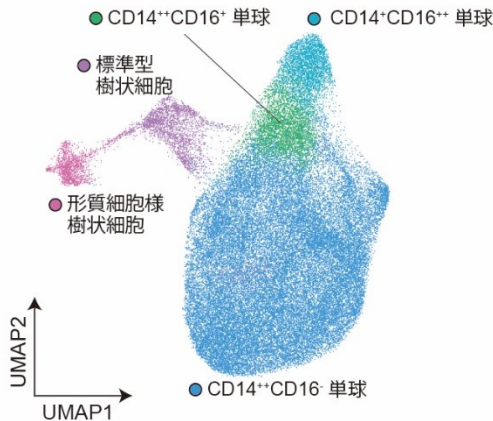


• 遺伝子変異が集団内で細胞組織的な遺伝子発現量に与える量的影響(=eQTL効果)も、**シングルセルeQTL解析**を通じたデータベース構築へとシフトしています。

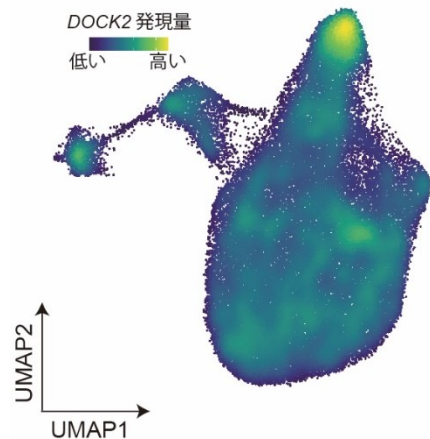
# ① シングルセル解析技術と情報解析

## COVID-19重症化遺伝子DOCK2の細胞組織・疾患特異的eQTL効果

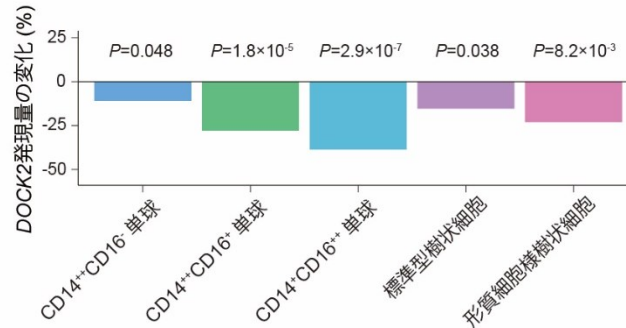
単球・樹状細胞のクラスタリングの結果



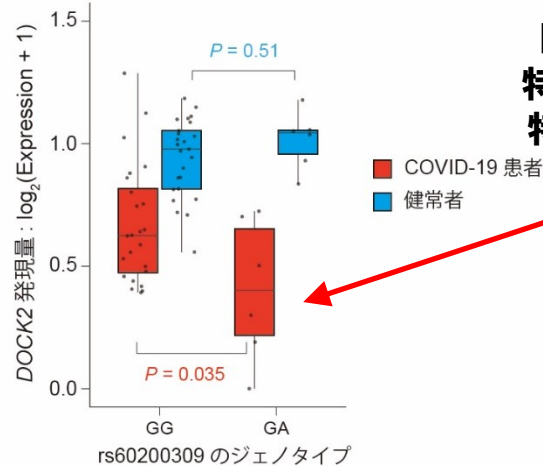
シングルセルレベルでの DOCK2 発現分布



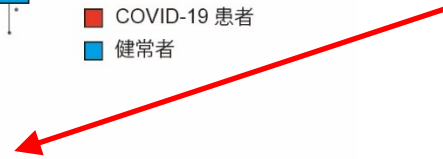
各クラスターでの DOCK2 発現量の比較  
(重症 COVID-19 患者 30 名 vs 健常者 31 名)



CD14<sup>+</sup>CD16<sup>++</sup> 単球における  
リスクバリエーションの DOCK2 発現量への影響



Non-classical monocyte  
特異的かつCOVID-19感染  
特異的な一細胞eQTL効果



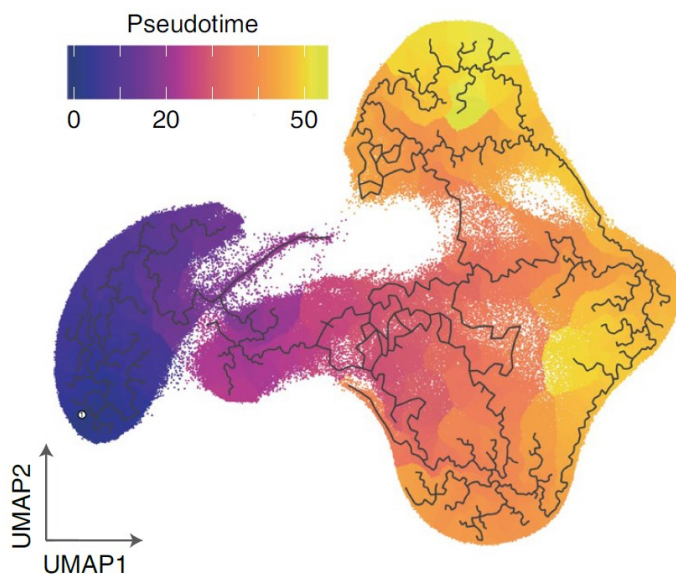
・疾患感受性遺伝子変異が細胞組織特異的かつ疾患特異的に一細胞 eQTL効果を持つ例が報告されています(= context-specific eQTL効果)。

(Namkoong H and Edahiro R et al. *Nature* 2022)

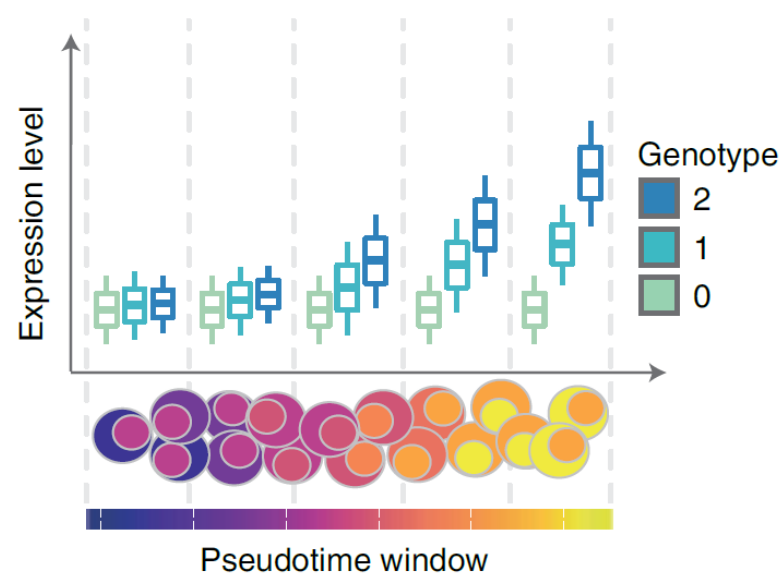
# ① シングルセル解析技術と情報解析

## T細胞の分化ダイナミクスに対するeQTL効果

刺激下におけるT細胞分化過程  
の軌道推定(by pseudotime)



Pseudotime推定軸に応じた  
eQTL効果の経時的・経分化的変化

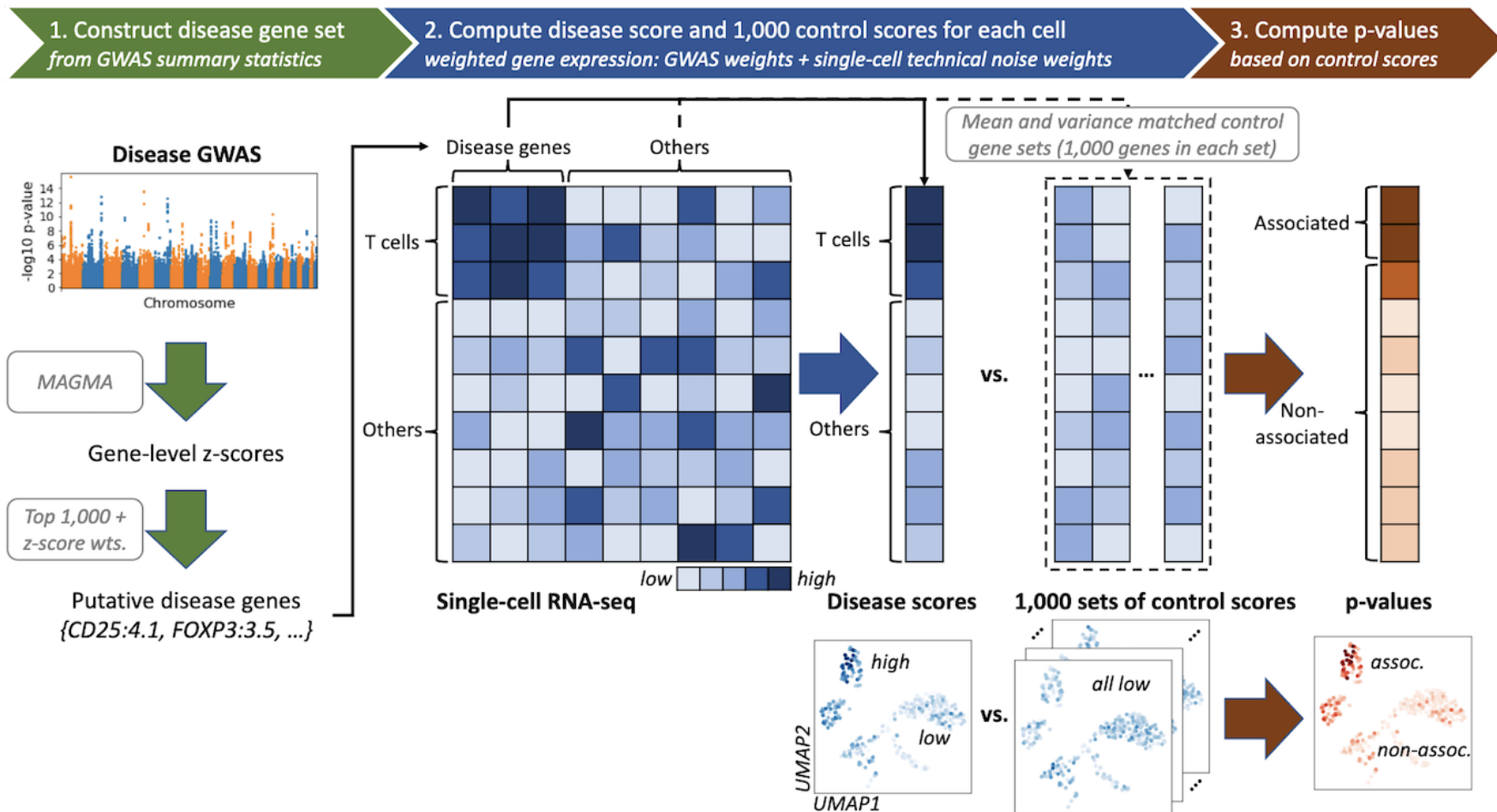


- Pseudotimeなどの軌道推定手法に基づき得られた細胞分化ダイナミクスに応じて、eQTL効果の経時的・経分化的な変化も報告されています(=ジェノタイプ毎に細胞分化が異なる現象、dynamic eQTL効果)。



# ① シングルセル解析技術と情報解析

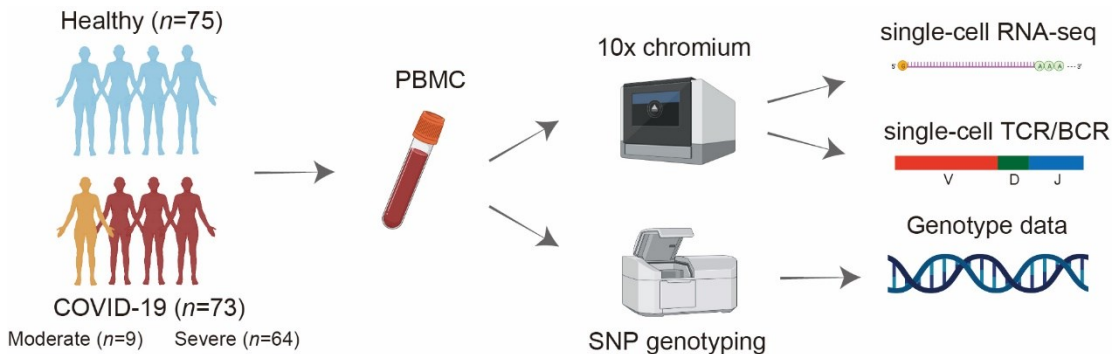
## 疾患GWAS解析結果とシングルセル解析の統合



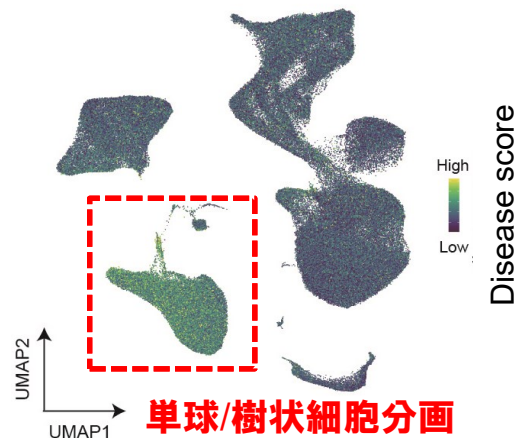
• **疾患ゲノム解析**と細胞組織特異的エピゲノム情報を統合する横断的オミクス解析も、シングルセル解析データの統合へとシフトしています。

# ① シングルセル解析技術と情報解析

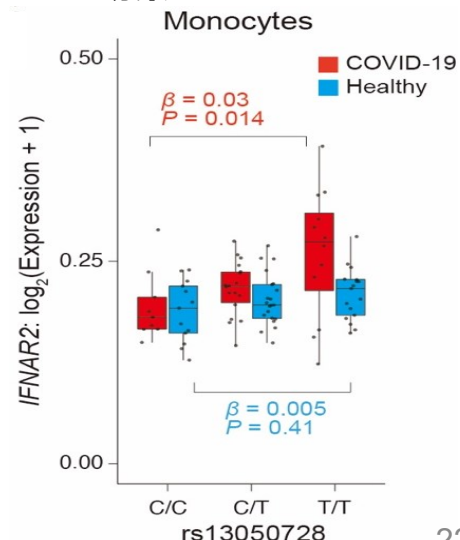
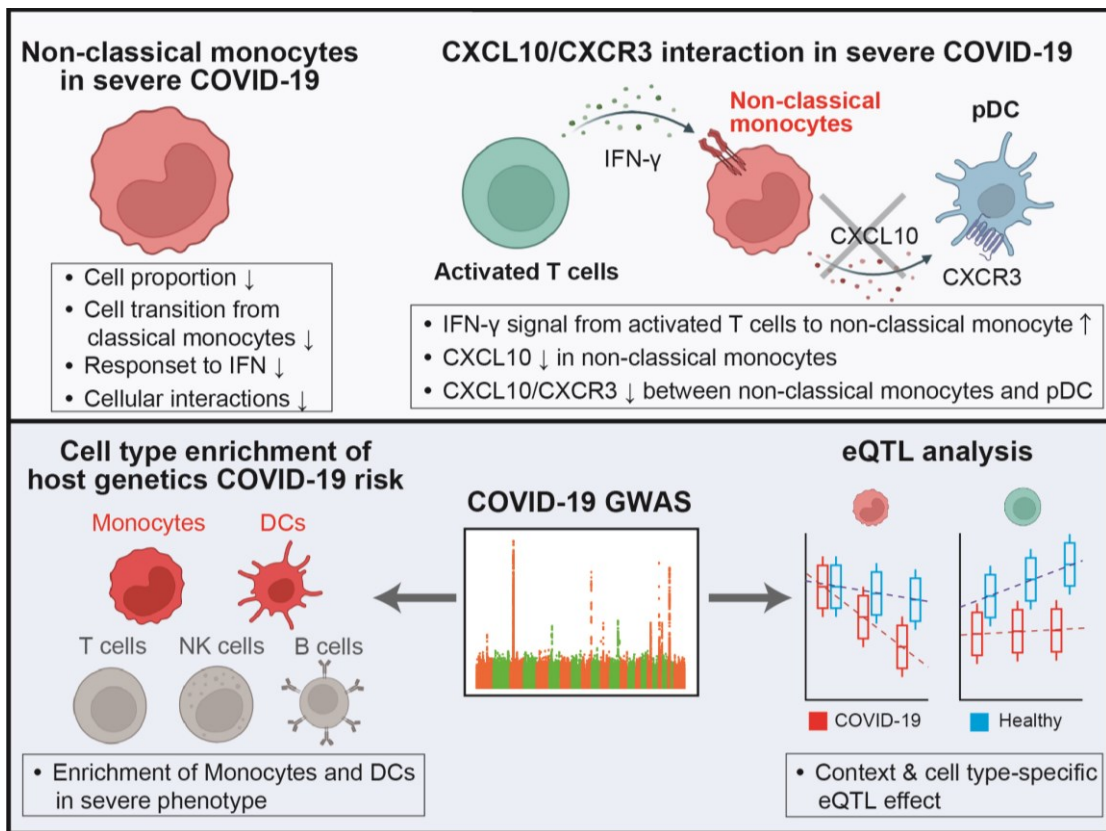
## 日本人集団COVID-19患者血液シングルセル解析



### ● 重症COVID-19のPolygenic risk投影



### ● 単球特異的かつCOVID-19特異的な一細胞eQTL効果

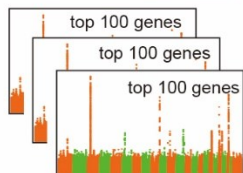


# ① シングルセル解析技術と情報解析

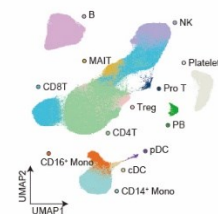
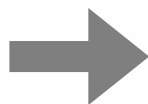
## Polygenic riskをシングルセルへ投影：遺伝的背景の細胞組織特異性

COVID-19 GWAS  
(3 phenotypes)

COVID-19 HGI. *Nature* 2022



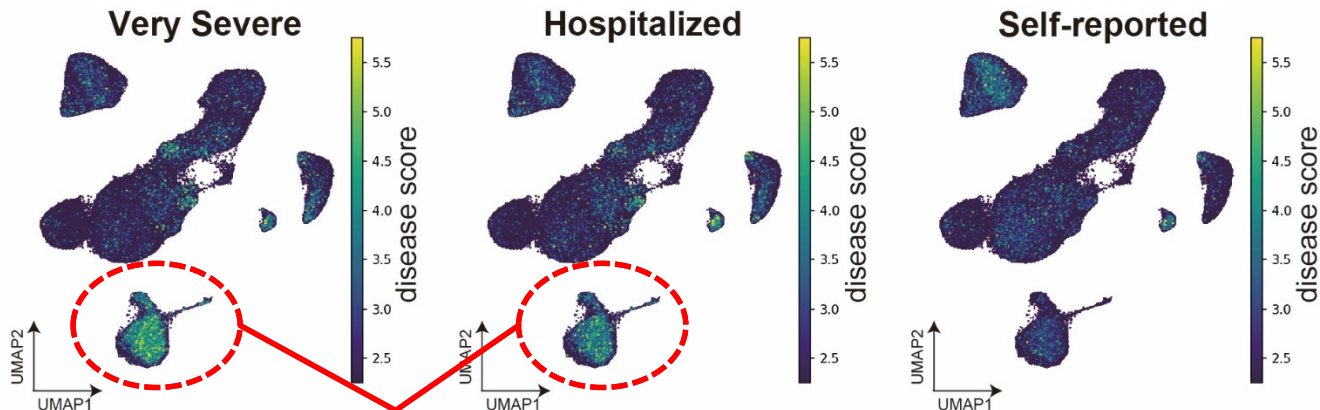
+



scRNA-seq data

PRS & single cell integration by scDRS

(Zhang M.J. et al. *Nat Genet* 2022)



単球分画にpolygenic risk集積

Very Severe  
Hospitalized  
Self-reported

Very Severe	×		×	■	■	
Hospitalized	×		×	■	■	
Self-reported				■	■	■
	CD4T	CD8T	NK	Mono	DC	B
				cMono	intMono	ncMono
				cDC	pDC	

□ Sig. cell type-disease association

× Sig. within-cell type heterogeneity

cell type-disease association



•ゲノムワイド関連解析の結果をシングルセル解析に投影することで、疾患の遺伝的背景と密接に関連した細胞分画の同定が可能に。

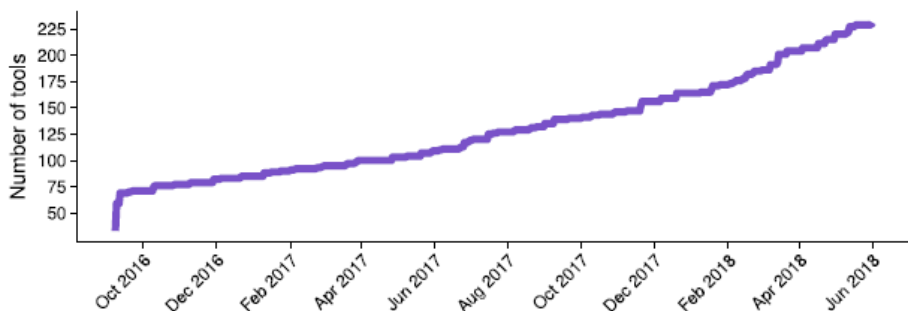
•COVID-19の遺伝的背景は、重症患者例を中心に単球分画にenrich。

(Edahiro R et al. *Nat Genet* 2023)

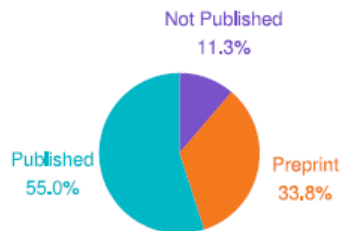
# ① シングルセル解析技術と情報解析

## シングルセル情報解析アルゴリズム: 戦国時代

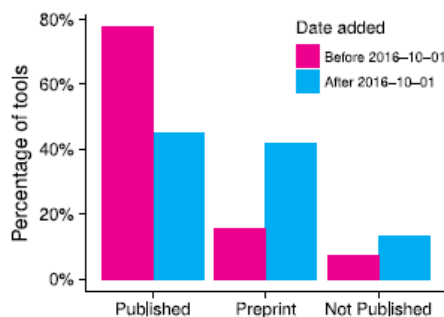
A – Increase in tools over time



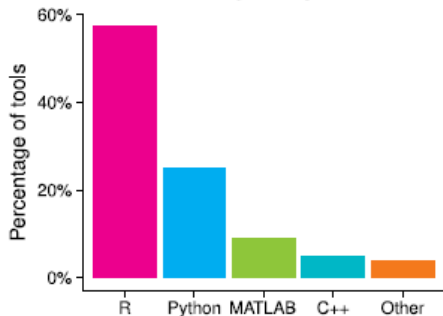
B – Publication status



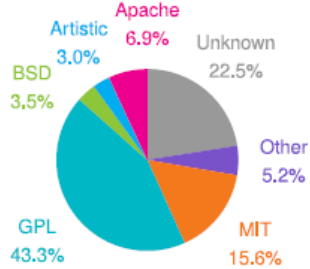
C – Publication status over time



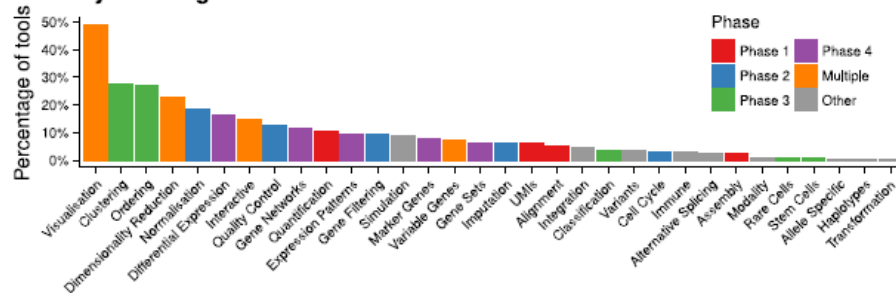
D – Platforms used by analysis tools



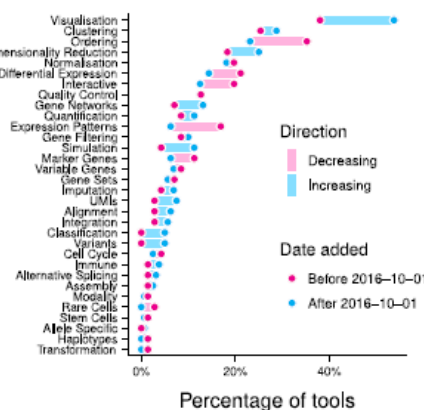
E – Associated software licenses



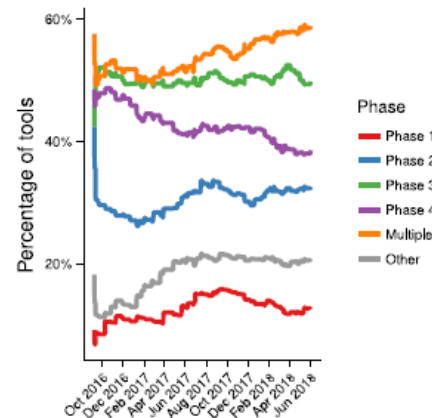
A – Analysis categories



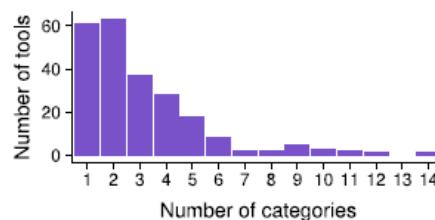
B – Change in analysis categories



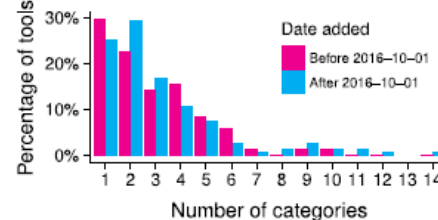
C – Analysis phases over time



D – Number of categories per tool



E – Categories per tool by date added



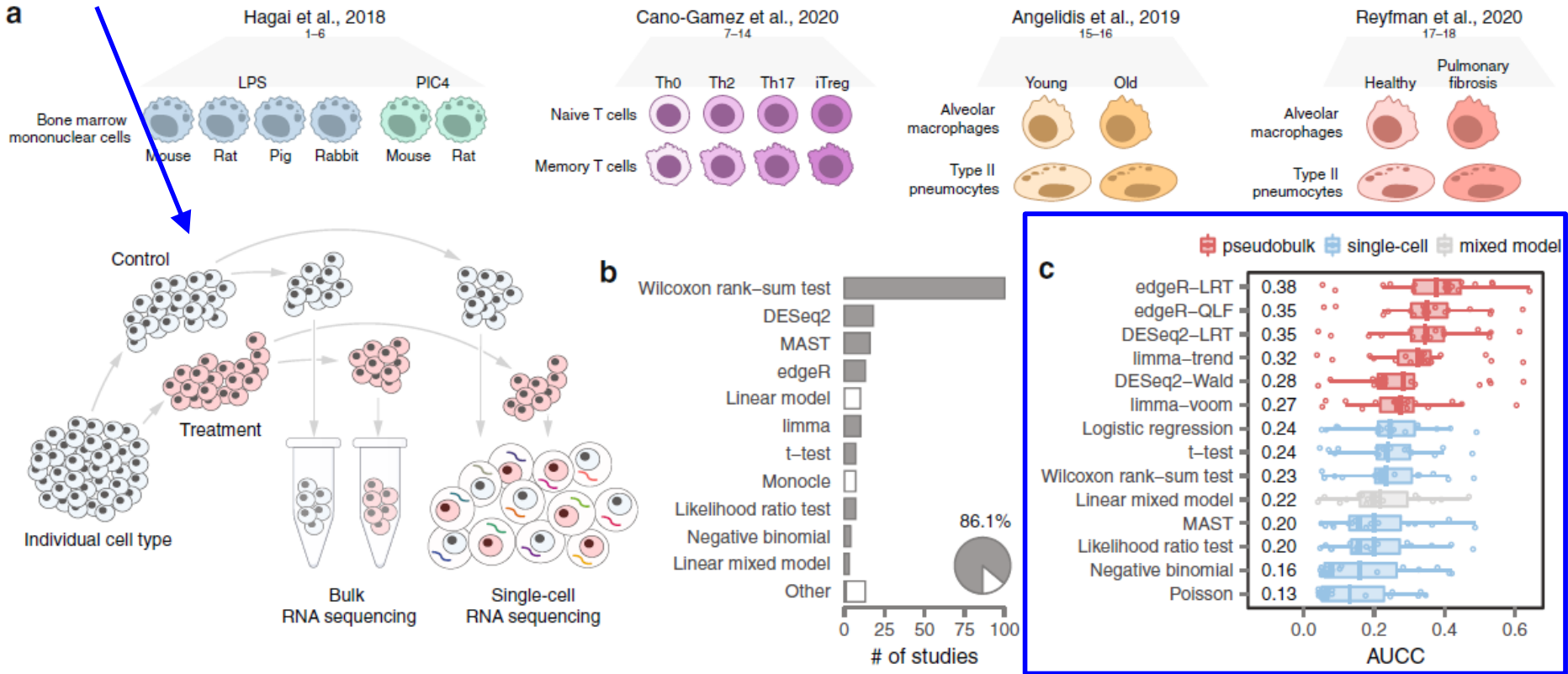
• シングルセル情報解析ソフトウェアが乱立し、まさに戦国時代!!



# ① シングルセル解析技術と情報解析

## シングルセル情報解析アルゴリズム: 戦国時代

疑似的にDEG解析データを作成

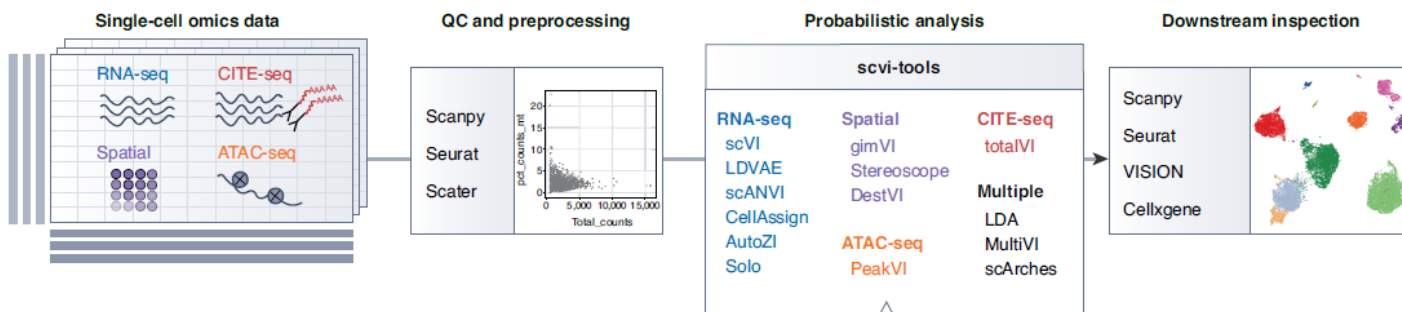


解析手法によって、DEG結果が大きく異なる!!

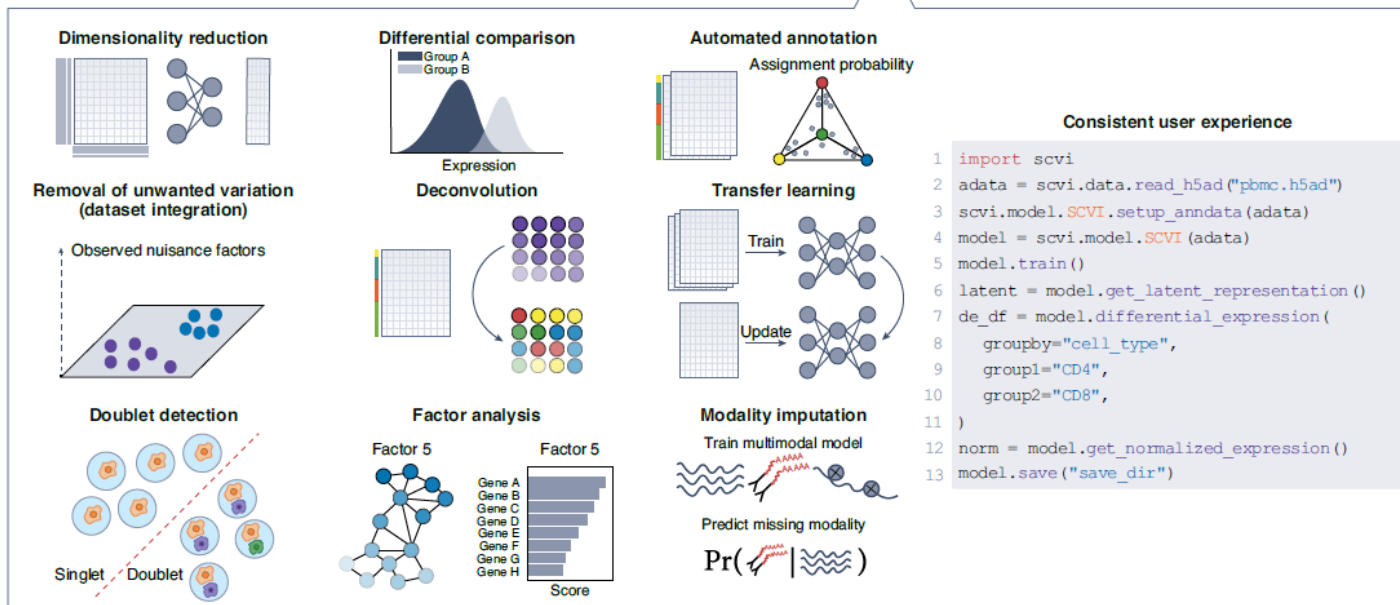
・頻用される遺伝子発現量比較解析(differential expression gene: DEG)においても、スタンダードな解析手法が定まっていない状況。

# ① シングルセル解析技術と情報解析

## シングルセル情報解析アルゴリズム: 戦国時代



b



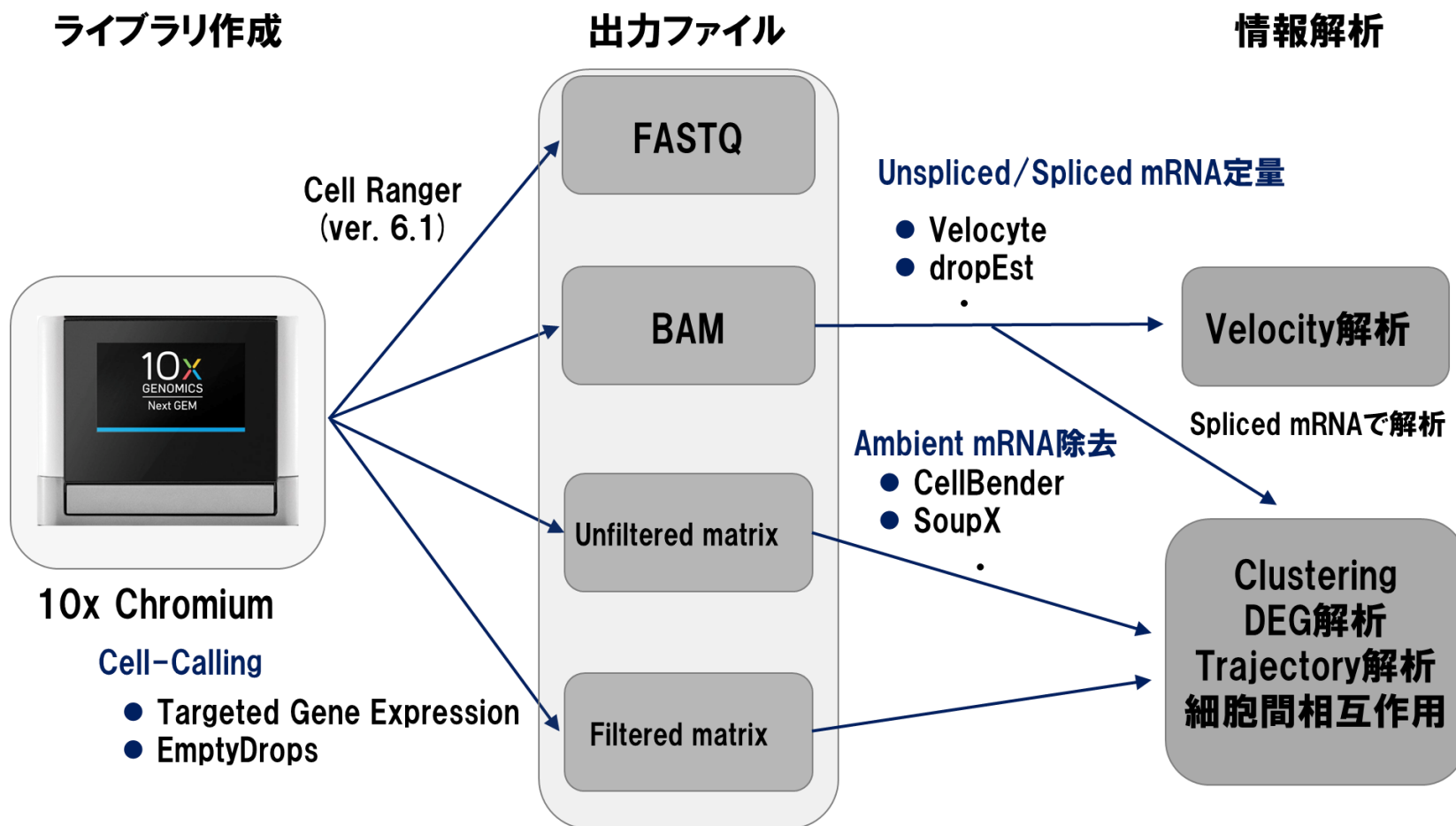
• 細胞分画、遺伝子発現、レパトア解析、cell-cell interaction、軌道推定、一通りの解析アルゴリズムを試すのも大変！

• まさに「解析アルゴリズムの洪水」状態。

(Gayoso A et al. *Nat Med* 2022)

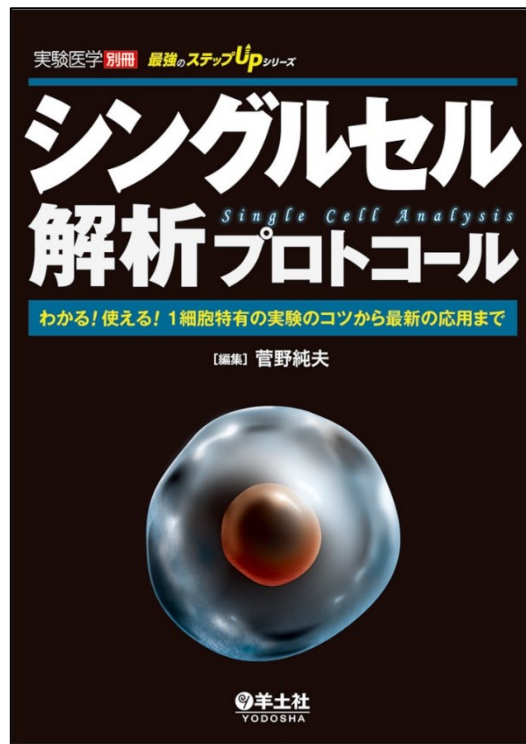
# ① シングルセル解析技術と情報解析

## シングルセル情報解析アルゴリズム: 戦国時代



- 10X Chromium謹製の解析ツールで遺伝子発現マトリックスを出力し、出来合いのツールで解析すれば一通りのことは誰でもできるように。
- しかし…それだけでいいのでしょうか？

# ① シングルセル解析技術と情報解析



- シングルセル解析の研究分野は、実験技術・情報解析技術共に発展のスピードが速く、最新情報の効率的な把握が重要となります。
- 英文原著論文はもちろん大事ですが、国内総説誌やweb媒体の活用も有用な手段と考えられます。



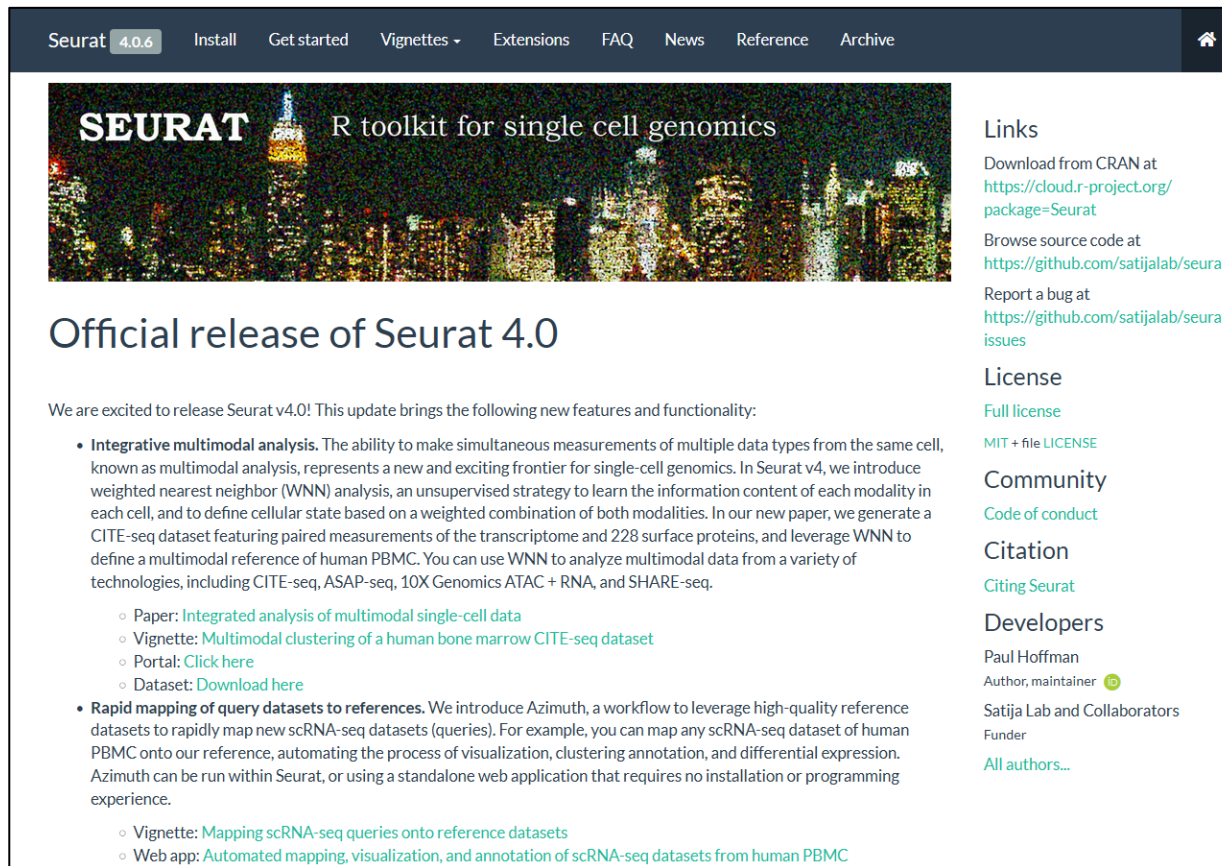
## GenomeDataAnalysis7

① シングルセル解析技術と情報解析

② Seuratを使ったシングルセル解析実習

本講義資料は、Windows PC上で  
C:¥SummerSchoolにフォルダを配置すること  
を想定しています。

## ② Seuratを使ったシングルセル解析実習



Seurat 4.0.6 Install Get started Vignettes Extensions FAQ News Reference Archive

# SEURAT

R toolkit for single cell genomics

## Official release of Seurat 4.0

We are excited to release Seurat v4.0! This update brings the following new features and functionality:


- **Integrative multimodal analysis.** The ability to make simultaneous measurements of multiple data types from the same cell, known as multimodal analysis, represents a new and exciting frontier for single-cell genomics. In Seurat v4, we introduce weighted nearest neighbor (WNN) analysis, an unsupervised strategy to learn the information content of each modality in each cell, and to define cellular state based on a weighted combination of both modalities. In our new paper, we generate a CITE-seq dataset featuring paired measurements of the transcriptome and 228 surface proteins, and leverage WNN to define a multimodal reference of human PBMC. You can use WNN to analyze multimodal data from a variety of technologies, including CITE-seq, ASAP-seq, 10X Genomics ATAC + RNA, and SHARE-seq.
  - Paper: [Integrated analysis of multimodal single-cell data](#)
  - Vignette: [Multimodal clustering of a human bone marrow CITE-seq dataset](#)
  - Portal: [Click here](#)
  - Dataset: [Download here](#)
- **Rapid mapping of query datasets to references.** We introduce Azimuth, a workflow to leverage high-quality reference datasets to rapidly map new scRNA-seq datasets (queries). For example, you can map any scRNA-seq dataset of human PBMC onto our reference, automating the process of visualization, clustering annotation, and differential expression. Azimuth can be run within Seurat, or using a standalone web application that requires no installation or programming experience.
  - Vignette: [Mapping scRNA-seq queries onto reference datasets](#)
  - Web app: [Automated mapping, visualization, and annotation of scRNA-seq datasets from human PBMC](#)

**Links**  
Download from CRAN at <https://cloud.r-project.org/package=Seurat>  
Browse source code at <https://github.com/satijalab/seurat/>  
Report a bug at <https://github.com/satijalab/seurat/issues>

**License**  
[Full license](#)  
MIT + file [LICENSE](#)

**Community**  
[Code of conduct](#)

**Citation**  
[Citing Seurat](#)

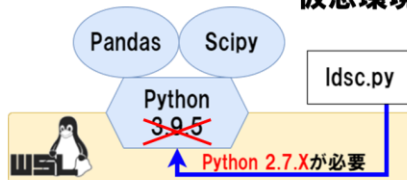
**Developers**  
Paul Hoffman  
Author, maintainer   
Satija Lab and Collaborators  
Funder  
[All authors...](#)

<https://satijalab.org/seurat/>

- **Seurat**は、シングルセル解析の代表的なツールとして知られています。
- 統計ソフトR上で動くプログラムとして、インストール方法や使用方法チュートリアル、FAQと共に、web上で公開されています。

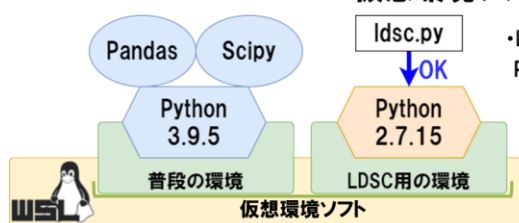
## ② Seuratを使ったシングルセル解析実習

### 仮想環境ソフトなし



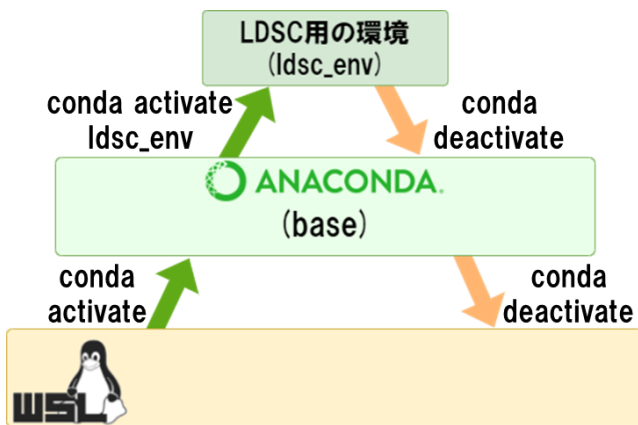
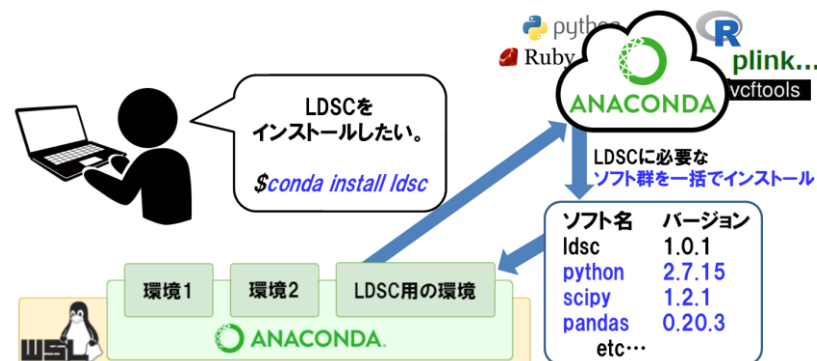
- Pythonを一度すべて消して入れなおす。  
(→今までに構築したモジュールも消える)
- Python2.7を追加で入れる。  
(→コマンドが重複するなど、混乱が生じる)

### 仮想環境ソフトあり



- LDSC用の環境を新規に作成してそこにPython2.7をインストールする。

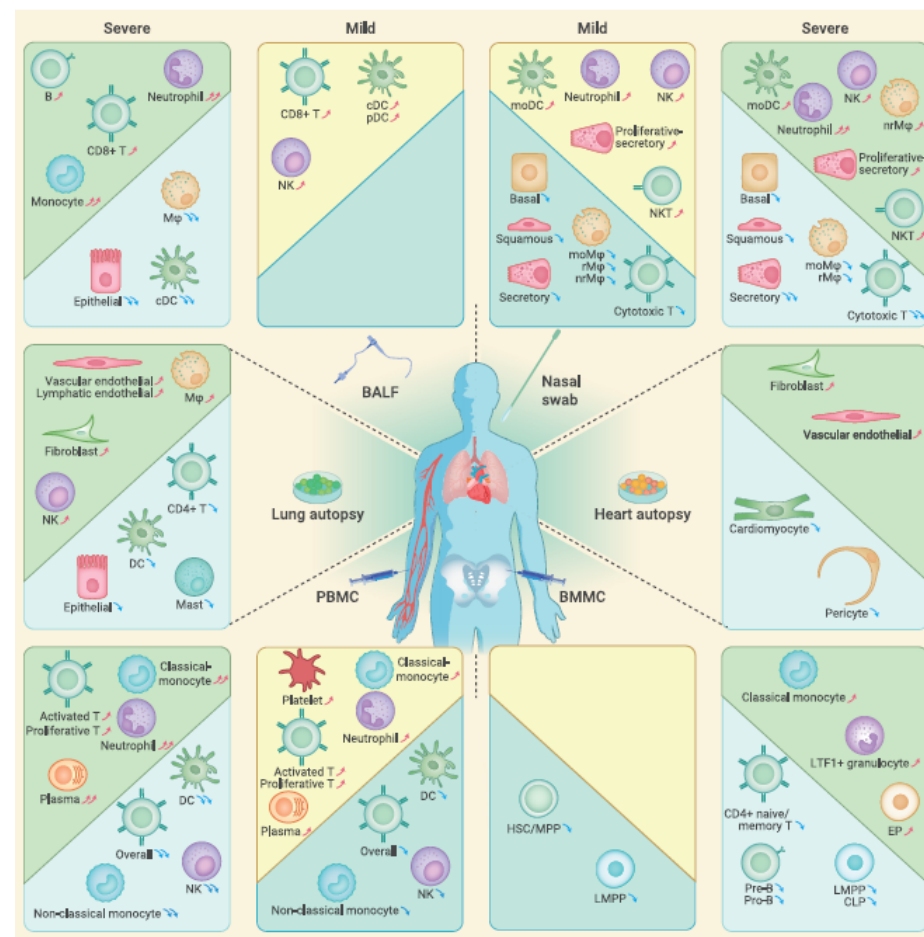
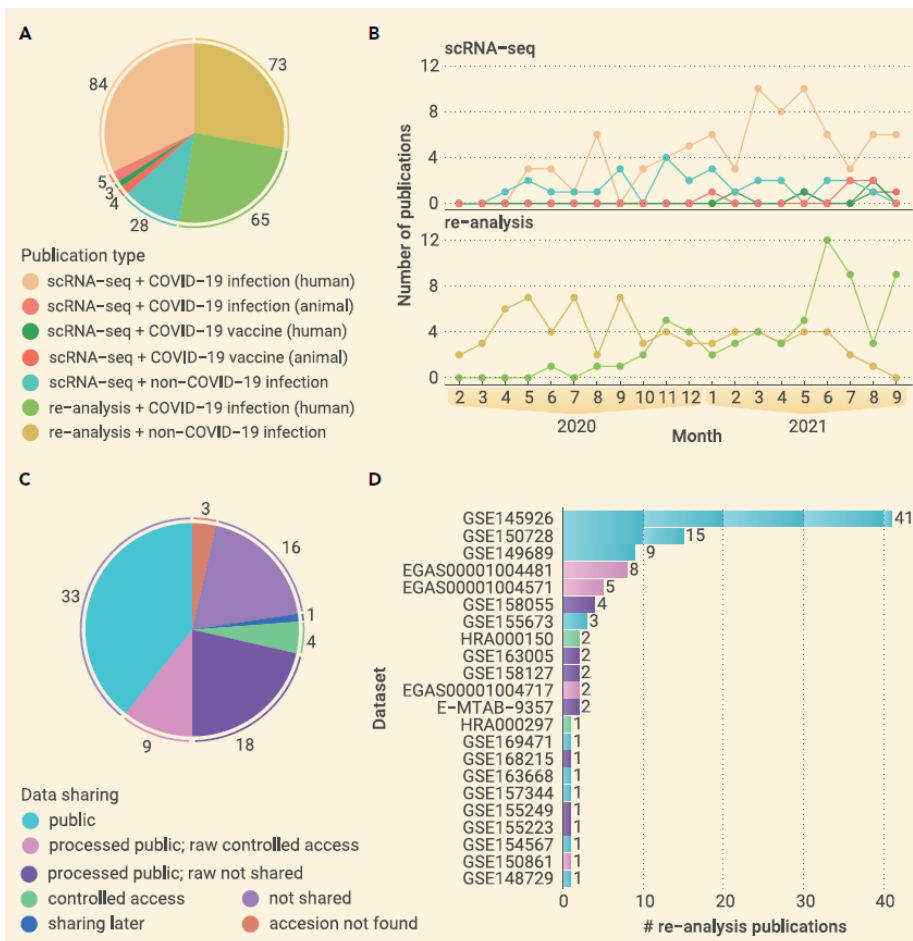
環境の切り替えも簡単!



```
yokada@LAPTOP-U1NCURTH: ~  
(ldsc_env) yokada@LAPTOP-U1NCURTH: ~$  
  
yokada@LAPTOP-U1NCURTH: ~  
(base) yokada@LAPTOP-U1NCURTH: ~$  
  
yokada@LAPTOP-U1NCURTH: ~  
yokada@LAPTOP-U1NCURTH: ~$
```

- Seuratのインストールには、仮想環境ソフトAnacondaを使用します。
- Anacondaの説明、導入方法や使用方法については、[手順書](#)および[GenomeDataAnalysis6](#)の演習内容を参照してください。

## ② Seuratを使ったシングルセル解析実習



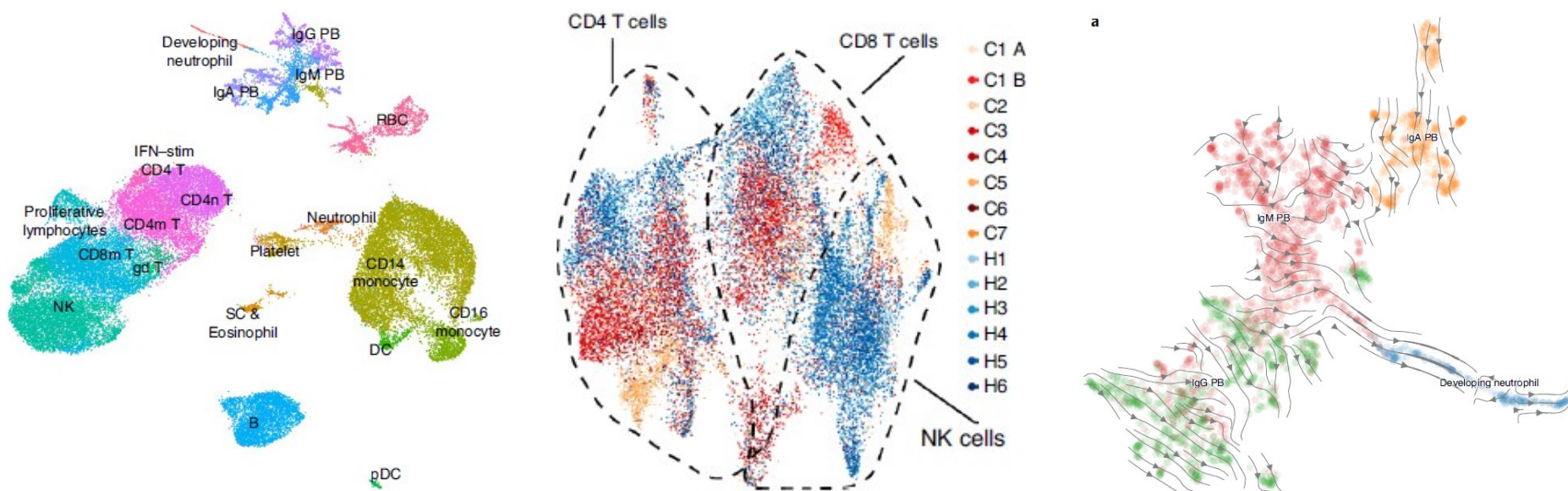
- COVID-19研究では、多彩なオミクス解析とデータ蓄積が進みました。
- シングルセル解析でも、多数の論文報告と公共データ登録が存在。

## ② Seuratを使ったシングルセル解析実習

# A single-cell atlas of the peripheral immune response in patients with severe COVID-19

nature  
medicine

Aaron J. Wilk<sup>1,2,5</sup>, Arjun Rustagi<sup>3,5</sup>, Nancy Q. Zhao<sup>2,5</sup>, Jonasel Roque<sup>3</sup>, Giovanni J. Martínez-Colón<sup>3</sup>, Julia L. McKechnie<sup>2</sup>, Geoffrey T. Ivison<sup>2</sup>, Thanmayi Ranganath<sup>3</sup>, Rosemary Vergara<sup>3</sup>,

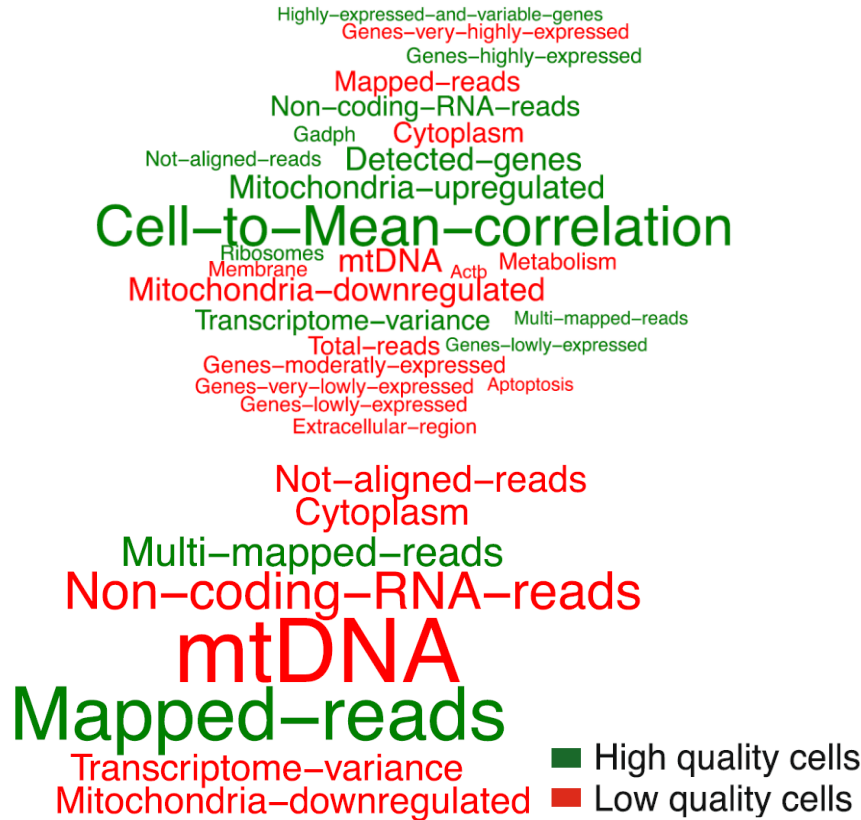


- 本演習では、COVID-19シングルセル解析公開データを使用します。
- 計算コスト軽減の観点から、QC実施済のオリジナルデータの一部を削り軽量化しています(COVID-19感染者7名 vs 対照群6名)。

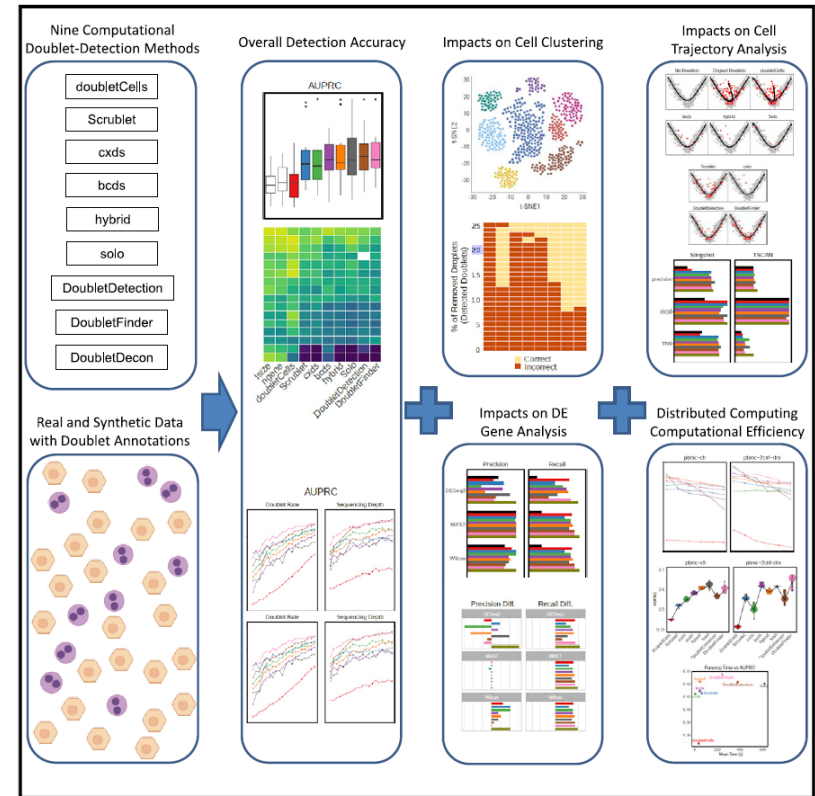


# ② Seuratを使ったシングルセル解析実習

## 低品質細胞の特徴可視化



## Doublet同定アルゴリズムのベンチマーク比較



- シングルセル解析はQCが重要で、**低品質細胞**(例:死細胞・ミトコンドリア遺伝子高発現細胞)や、**Doublet**(単一ドロップに複数細胞が封入)を除外します。
- データセット毎にどの手法・閾値を適用するか、個別検討が必要です。

## ② Seuratを使ったシングルセル解析実習

```
statgen@statgen-PC: ~
```

```
$ conda activate
```

※ファイル”SingleCell\_Command.txt”を開いて、内容をLinuxコマンドにコピー&ペーストして下さい。

```
(base) statgen@statgen-PC: ~
```

```
$ conda activate seurat_env
```

← Anaconda環境の起動

```
(seurat_env) statgen@statgen-PC: ~
```

```
$ cd /mnt/c/SummerSchool/GenomeDataAnalysis7/Analysis/
```

```
(seurat_env) statgen@statgen-PC: ~
```

```
$ R ← Linux上でのRの起動
```

- AnacondaでSeuratの実行に必要な**仮想環境**を起動します。  
(シングルセル解析演習は、Cygwin環境やM1 M2 macには対応していません。)
- Seuratは統計ソフトR上で動くpackageとして作られています。
- Linux環境上でRを起動し、その上でSeuratを実行します。  
(Seuratのインストールに必要なR packageは、Anacondaによりインストール済です。)

## ② Seuratを使ったシングルセル解析実習

```
> rm(list = ls(all = TRUE));
```

```
> setwd("/mnt/c/SummerSchool/GenomeDataAnalysis7/Analysis");
```

↑ 保存データを初期化 & R上でディレクトリの移動

```
> library(Seurat); ← Seuratライブラリを起動
```

```
> library(ggplot2); ← ggplot2ライブラリを起動
```

※ファイル”SingleCell\_Command.txt”を開いて、内容をRコマンドにコピー&ペーストして下さい。

```
> data <- readRDS("NatMed2020_COVID7HC6.rds"); ← シングルセルデータの読み込み
```

```
> data;
```

An object of class Seurat

26361 features across 44721 samples within 1 assay

Active assay: RNA (26361 features, 0 variable features)

2 dimensional reductions calculated: pca, umap

- **解析シングルセルデータ(”NatMed2020\_COVID7HC6.rds”)を読み込みます。**
- **Seuratオブジェクトという形式であり、44,721細胞に対して26,361遺伝子の発現量データが格納されています。**
- **細胞群2次元プロットに必要なPCA・UMAP座標情報※も計算済です。**  
(※:本講義では座標計算方法は説明しません。Seurat公式チュートリアルをご参照下さい<sup>36</sup>。)



## ② Seuratを使ったシングルセル解析実習

```
> str(data);
```

```
Formal class 'Seurat' [package "SeuratObject"] with 13 slots
```

```
..@ assays      :List of 1
```

```
...$ RNA:Formal class 'Assay' [package "SeuratObject"] with 8 slots
```

```
.....@ counts   :Formal class 'dgCMatrix' [package "Matrix"] with 6 slots
```

```
.....@ i        : int [1:51089274] 1062 1230 1535 1924 1965 2035 2358 2573 2671 2691 ...
```

```
.....@ p        : int [1:44722] 0 125 285 497 809 1145 1496 1870 2254 2646 ...
```

```
.....@ Dim      : int [1:2] 26361 44721
```

```
.....@ Dimnames:List of 2
```

```
.....$ : chr [1:26361] "5S-rRNA" "7SK" "A1BG" "A1BG-AS1" ...
```

```
.....$ : chr [1:44721] "covid_555_1.1" "covid_555_1.2" "covid_555_1.3" "covid_555_1.7" ...
```

```
.....@ x        : num [1:51089274] 1 1 4 1 1 1 1 1 1 5 ...
```

```
.....@ factors  : list()
```

```
.....@ data     :Formal class 'dgCMatrix' [package "Matrix"] with 6 slots
```

```
.....@ i        : int [1:51089274] 1062 1230 1535 1924 1965 2035 2358 2573 2671 2691 ...
```

```
.....@ p        : int [1:44722] 0 125 285 497 809 1145 1496 1870 2254 2646 ...
```

```
.....@ Dim      : int [1:2] 26361 44721
```

```
⋮
```

- 今回のデータは、ベクトルやテーブルが複数の階層で束ねられたオブジェクト形式のため、`str()` 関数を使ってデータの構造を把握します。<sup>37</sup>

## ② Seuratを使ったシングルセル解析実習

```
> str(data, max.level=2);  
Formal class 'Seurat' [package "SeuratObject"] with 13 slots  
..@ assays      :List of 1  
..@ meta.data   :'data.frame':   44721 obs. of  6 variables:  
..@ active.assay: chr "RNA"  
..@ active.ident: Factor w/ 13 levels "C1","C2","C3",...: 1 1 1 1 1 1 1 1 1 1 ...  
.. ..- attr(*, "names")= chr [1:44721] "covid_555_1.1" "covid_555_1.2" "covid_555_1.3"  
"covid_555_1.7" ...  
..@ graphs      : Named list()  
..@ neighbors   : list()  
..@ reductions  :List of 2  
..@ images      : list()  
..@ project.name: chr "SeuratProject"  
..@ misc        : list()  
..@ version     :Classes 'package_version', 'numeric_version' hidden list of 1  
..@ commands    : Named list()
```

- **引数”max.level”の指定で、各階層に応じたデータ構造が把握できます。**
- **@assays, @meta.data, @graphsなど、個別のデータの名称が付与されていることがわかります。**

## ② Seuratを使ったシングルセル解析実習

> data@assays\$RNA@data[20:35,1:5];

← 行列データへの直接アクセス

> GetAssayData(data, slot = "data")[20:35,1:5];

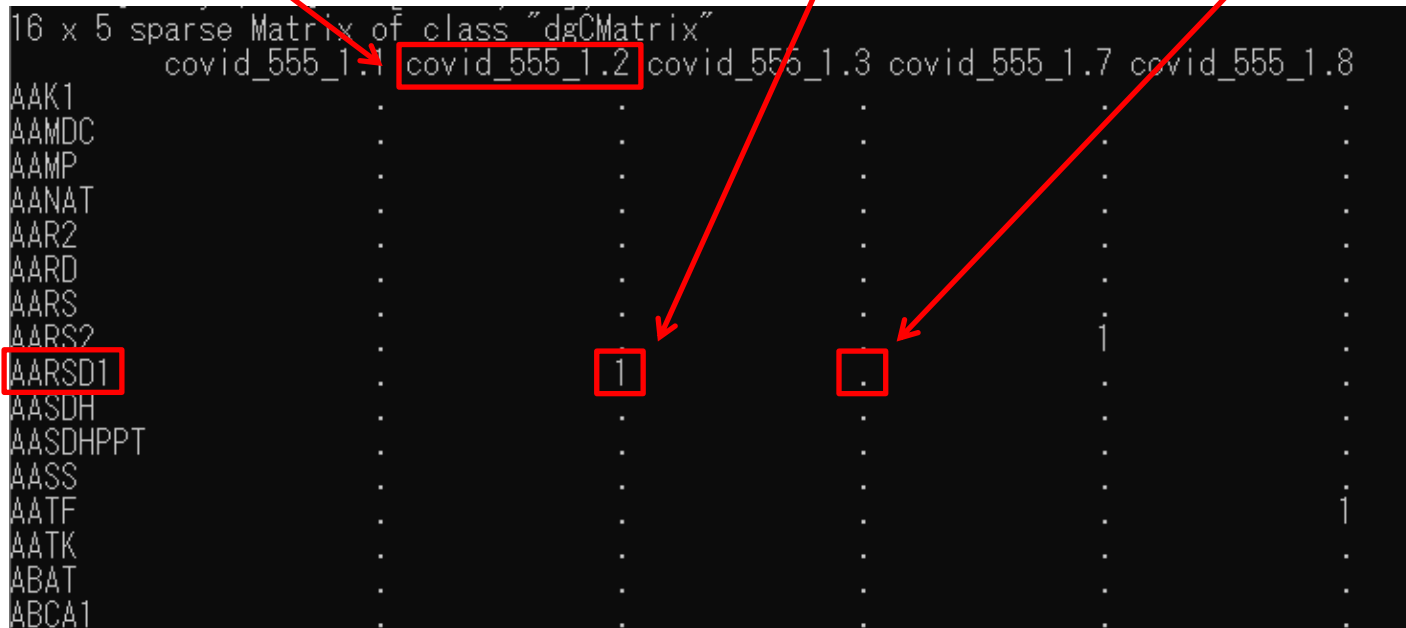
← GetAssayData()関数を用いたアクセス

個別の細胞のID

遺伝子発現を1リード観測

遺伝子発現は観測されず

遺伝子



- @assaysデータには細胞×遺伝子の発現量行列が格納されています。直接アクセスやSeurat内臓のGetAssayData()関数で確認可能です。
- "covid\_555\_1.2"という細胞で、AARSD1遺伝子の発現が1リード観測されていることがわかります。

## ② Seuratを使ったシングルセル解析実習

```
> dataN <- NormalizeData(data, normalization.method = "LogNormalize", scale.factor = 10000); ← 遺伝子発現量行列データを正規化
```

```
> GetAssayData(dataN, slot = "data")[20:35,1:5]; ← 正規化後のデータを確認
```

正規化後は遺伝子発現量が整数から小数へ

```
16 x 5 sparse Matrix of class "dgCMatix"
      covid_555_1.1 covid_555_1.2 covid_555_1.3 covid_555_1.7 covid_555_1.8
AAK1      .          .          .          .          .
AAMDC      .          .          .          .          .
AAMP      .          .          .          .          .
AANAT      .          .          .          .          .
AAR2      .          .          .          .          .
AARD      .          .          .          .          .
AARS      .          .          .          .          .
AARS2      .          .          .          1.638542      .
AARSD1      .          .          .          .          .
AASDH      .          .          .          .          .
AASDHPPT      .          .          .          .          .
AASS      .          .          .          .          .
AATF      .          .          .          .          1.685227
AATK      .          .          .          .          .
ABAT      .          .          .          .          .
ABCA1      .          .          .          .          .
```

2.312454

- 細胞特異的な実験バイアスの除外目的で、データを**正規化**します。
- 正規化には様々な方法が存在し、解析ツールによっても異なります。
- 今回は解析コストの観点から、**細胞毎に10,000リードあたりの発現量に補正**してから対数変換を行う、基本的な正規化を実施しています。

## ② Seuratを使ったシングルセル解析実習

> head(dataN@meta.data); ← @meta.dataのヘッダー情報を確認

個別の細胞のID	発現遺伝子のリード数	発現遺伝子の個数	Annotateされた細胞種類	サンプルID	形質情報
covid_555_1.1	1222	125	RBC	C1	COVID
covid_555_1.2	1099	160	Class-switched B	C1	COVID
covid_555_1.3	1055	212	IgG PB	C1	COVID
covid_555_1.7	2411	312	Class-switched B	C1	COVID
covid_555_1.8	2276	336	IgA PB	C1	COVID
covid_555_1.11	1166	351	IgA PB	C1	COVID

> table(dataN@meta.data\$cell.type); ← @meta.data\$cell.typeのヘッダー情報を確認

"RBC" "Class-switched B" "IgG PB" "Class-switched B" "IgA PB" "IgA PB"

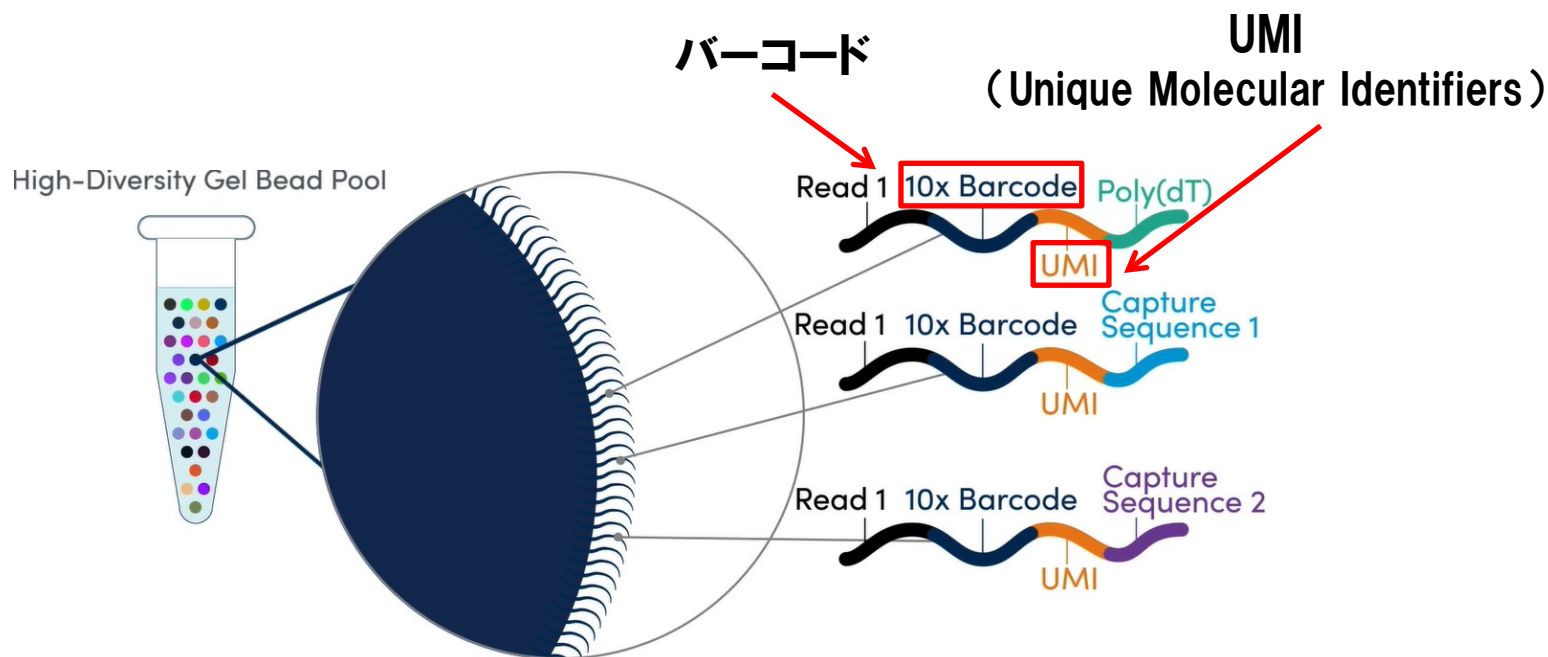
- Seuratオブジェクトの@meta.dataには、各細胞に付与されたメタデータの行列ファイルが格納されています。内容を確認してみましょう。
- 各行が各細胞に相当し、1行目の細胞の付与情報は以下の通りです。  
細胞ID:covid\_555\_1.1、発現遺伝子のリード数:1,222、発現遺伝子の個数:125、細胞種類:RBC、サンプルID報:C1、形質情報:COVID群

(※列の構成や意味はデータに依存して異なるため、その都度、確認する必要があります。)<sup>41</sup>



## ② Seuratを使ったシングルセル解析実習

### シングルセル解析リードに含まれるバーコード・UMI(10X Chromium)

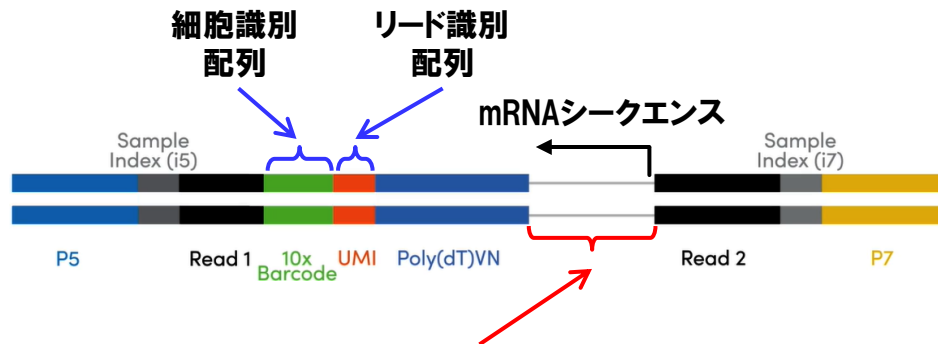


- ドロップレット封入型では、どの細胞由来のリードか識別する目的で、各細胞(=各ドロップレット)に固有の**バーコード**や、各リードに固有の**ランダム塩基配列(Unique Molecular Identifiers: UMI)**が割り振られています。
- 各細胞で観測された独立なUMIの個数に基づき、PCR増幅による重複を差し引いて、一細胞内における発現遺伝子数が把握できます。

## ② Seuratを使ったシングルセル解析実習

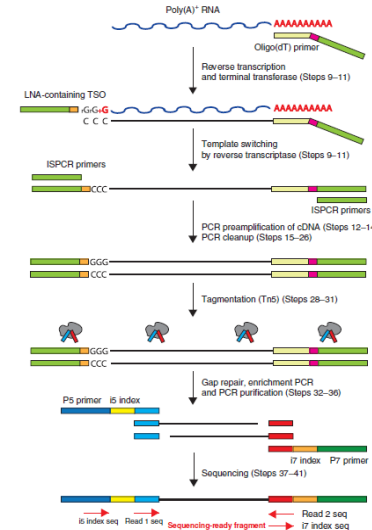
### シングルセル解析でシーケンスされるmRNAの長さ

#### ドロップレット封入型(10X Chromium)



- ・シーケンスされるmRNA長は100-150塩基
- ・mRNAの5' 末端か3' 末端のどちらか一方をライブラリー作成時に選択

#### Smart-seq

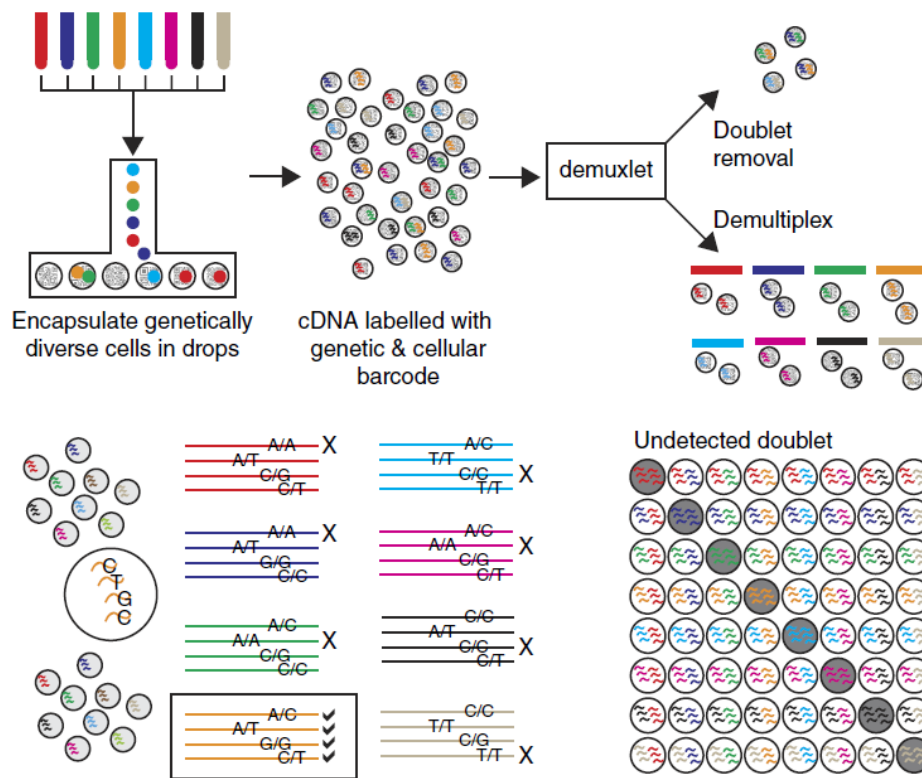


原理的には  
mRNA全長の  
シーケンス  
が可能

- ・シングルセルシーケンス対象となるmRNA長は手法により異なります。
- ・ドロップレット封入型(10X Chromium)では、100-150塩基と短めであり、mRNAの5' 末端/3' 末端のどちらか一方がシーケンスされます。
- ・mRNA全長シーケンス可能な解析手法として、smart-seqがあります。
- ・シーケンス手法、ライブラリー作成手法、入力細胞数、リード長、リード数など、シングルセル解析の実験条件はその都度確認が必要です。

## ② Seuratを使ったシングルセル解析実習

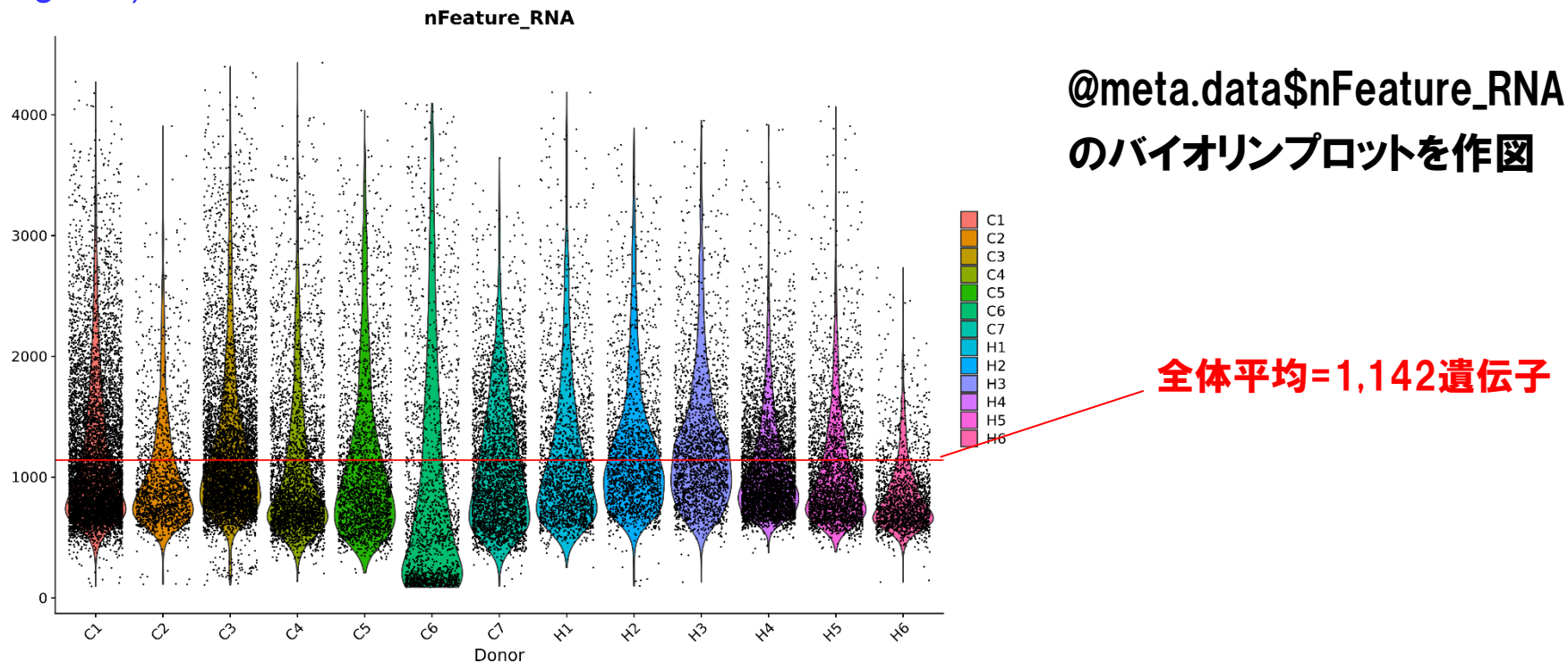
### Demultiplexによる各細胞が由来するサンプルの同定



- シングルセルシーケンスされたmRNA配列上の遺伝子変異に着目したり、(別途取得された)個人のゲノム配列と比較することで、各細胞がどのサンプル由来かを判定可能になり、“demultiplex”と呼ばれています。
- 複数サンプル由来の細胞を混ぜてシングルセル解析する際に有用です。

## ② Seuratを使ったシングルセル解析実習

```
> gfile1 <- VlnPlot(data, features="nFeature_RNA") +  
  geom_hline(yintercept=mean(data@meta.data$nFeature_RNA), color="red") + xlab("Donor");  
> ggsave(gfile1, filename="NatMed2020_COVID7HC6.Seurat.nFeature.Violinplot.png", width=12,  
  height=8);
```



@meta.data\$nFeature\_RNA  
のバイオリンプロットを作図

全体平均=1,142遺伝子

- 各細胞毎の発現遺伝子数の分布は、平均1,142遺伝子と少なめです。
- 大半の遺伝子(約20,000個)の発現が検出可能なbulk RNA-seqと比べ、  
検出できる遺伝子数が少ない点が、現在のシングルセル解析の課題。

## ② Seuratを使ったシングルセル解析実習

> table(dataN@meta.data\$cell.type); ← @meta.data\$cell.typeの構成要素を確認

```
Activated Granulocyte      B      CD14 Monocyte
                        206      3681      10339
CD16 Monocyte              CD4 T      CD4m T
                        1348      459      4098
CD4n T                     CD8eff T      CD8m T
                        3840      697      6065
Class-switched B           DC      gd T
                        1341      456      448
IgA PB                    IgG PB      Neutrophil
                        956      1107      301
NK                        pDC      Platelet
                        6857      234      527
RBC                      SC & Eosinophil
                        1437      324
```

> table(dataN@meta.data\$Status);

```
Healthy  COVID
16627    28094
```

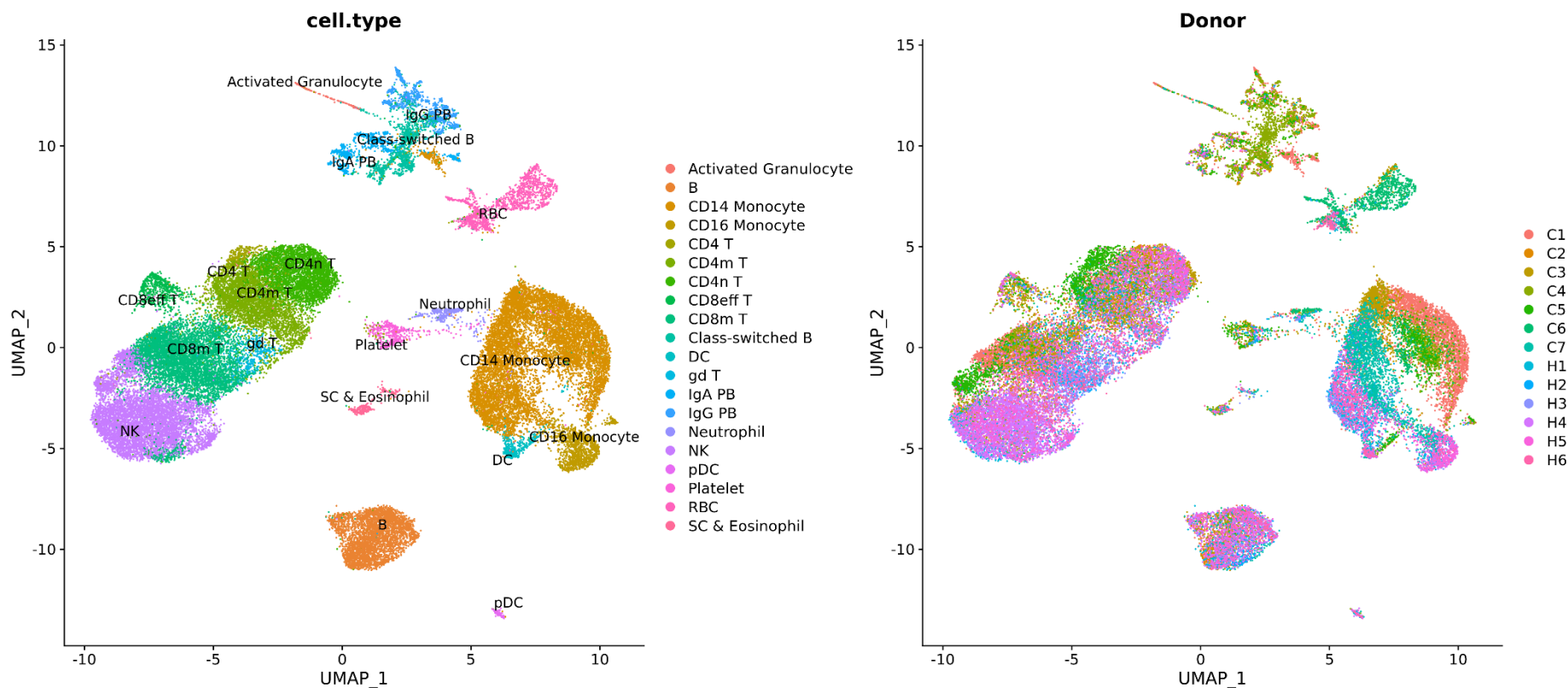
- @meta.dataの各列の構成要素を確認してみましょう。
- table () 関数で\$cell.typeを集計すると、各細胞種類毎の細胞数が表示されます。CD14+ Monocyteが10,339細胞と一番多いようです。
- COVID-19患者群が28,094細胞、対照群が16,627細胞になります。



## ② Seuratを使ったシングルセル解析実習

```
> gfile2 <- DimPlot(dataN, group.by = "cell.type", label=TRUE, repel=TRUE);  
> gfile3 <- DimPlot(dataN, group.by = "Donor");  
> ggsave(gfile2+gfile3, filename = "NatMed2020_COVID7HC6.Seurat.UMAP.png", width=18,  
height=8);
```

← 2次元プロットを作成し、\$cell.typeや\$Donorの情報に基づき分類



- **Dimplot () 関数**を用いて、細胞群の2次元プロット(UMAP)を作図します。  
(※:今回は各細胞の細胞分画のannotationは実施せず、予め付与されたものを使用します。)

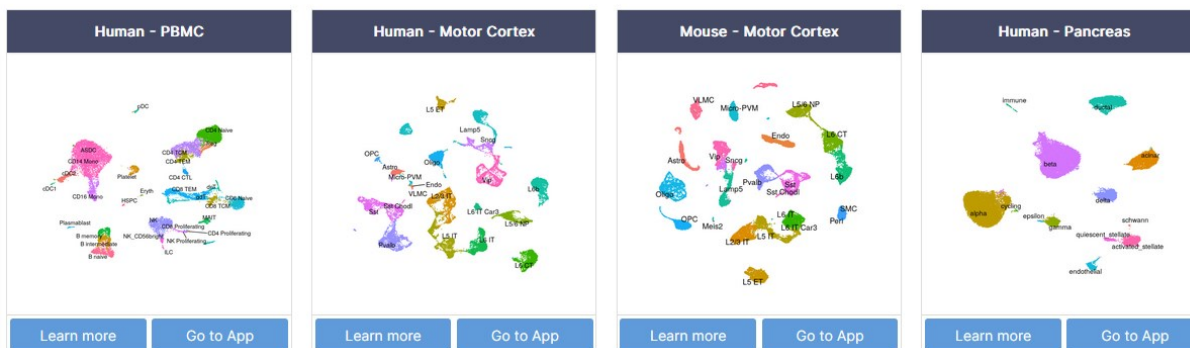
## ② Seuratを使ったシングルセル解析実習



Azimuth is a web application that uses an annotated reference dataset to **automate the processing, analysis, and interpretation of a new single-cell RNA-seq experiment**. Azimuth leverages a 'reference-based mapping' pipeline that inputs a counts matrix of gene expression in single cells, and performs normalization, visualization, cell annotation, and differential expression (biomarker discovery). All results can be explored within the app, and easily downloaded for additional downstream analysis.

The development of Azimuth is led by the New York Genome Center Mapping Component as part of the NIH Human Biomolecular Atlas Project (HuBMAP). Eight molecular reference maps are currently available, with more coming soon.

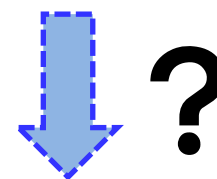
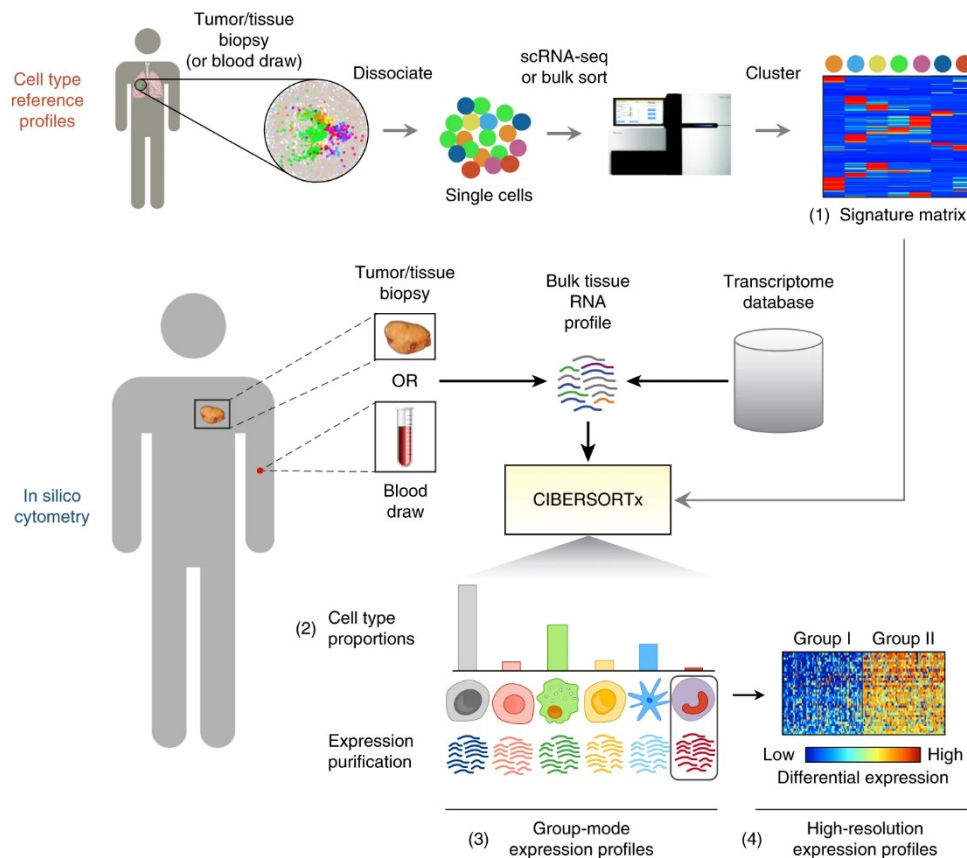
### References



- 既知の細胞組織特異的なマーカー遺伝子群の発現量を参照することで、個別の細胞が、**どの細胞分画に属するか推定**することができます。
- 様々な細胞分画推定ツールやマーカー遺伝子情報が公開されています。

## ② Seuratを使ったシングルセル解析実習

### CibersortXによるbulk RNA-seqデータのdeconvolution



- Bulk RNA-seqを対象に、細胞特異的遺伝子発現情報を参照し、各サンプルの細胞分画割合を定量推定する解析が、**Deconvolution**です。
- 有用ですが、シングルセル解析の直接観測には精度が劣る印象です。

## ② Seuratを使ったシングルセル解析実習

```
> dataN_cd16mono <- subset(dataN, cell.type=="CD16 Monocyte");  
> all_genes <- rownames(dataN_cd16mono);  
> filter_genes <- all_genes[grep("^MT-|^RP[SL]^RNA", all_genes, invert = TRUE)]; ←  
> de.markers <- FindMarkers(dataN_cd16mono, ident.1 = "COVID", ident.2 = "Healthy",  
  group.by="Status", features = filter_genes);  
> head(de.markers, n=15);
```

多重検定補正後P値

正規表現を用いて、

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
PIM1	4.462346e-74	1.7318084	0.538	0.102	1.176319e-69
IFI27	3.563478e-57	3.0811696	0.266	0.003	9.393683e-53
PABPC1	5.231588e-55	-0.6884141	0.970	0.995	1.379099e-50
AHNAK	2.090046e-51	-0.9148276	0.855	0.978	5.509569e-47
PLAC8	7.787258e-51	1.4276754	0.640	0.271	2.052799e-46
NAP1L1	1.855325e-49	-0.8927568	0.804	0.949	4.890823e-45
SOCS3	2.000120e-49	1.2803191	0.298	0.027	5.272516e-45
HLA-E	7.007038e-46	-0.5579238	0.982	0.999	1.847125e-41
IFITM3	1.182850e-45	1.0546346	0.871	0.670	3.118111e-41
GBP1	1.913828e-45	1.3142940	0.667	0.302	5.045041e-41
GBP5	1.056399e-43	1.4988166	0.564	0.224	2.784773e-39
FCGR1A	2.844721e-42	1.0000918	0.284	0.033	7.498968e-38
STAT1	9.117915e-42	1.1110547	0.721	0.390	2.403574e-37
TMSB10	1.707603e-37	0.5987025	0.988	0.958	4.501412e-33
WARS	2.878312e-36	1.0223132	0.885	0.797	7.587517e-32

遺伝子

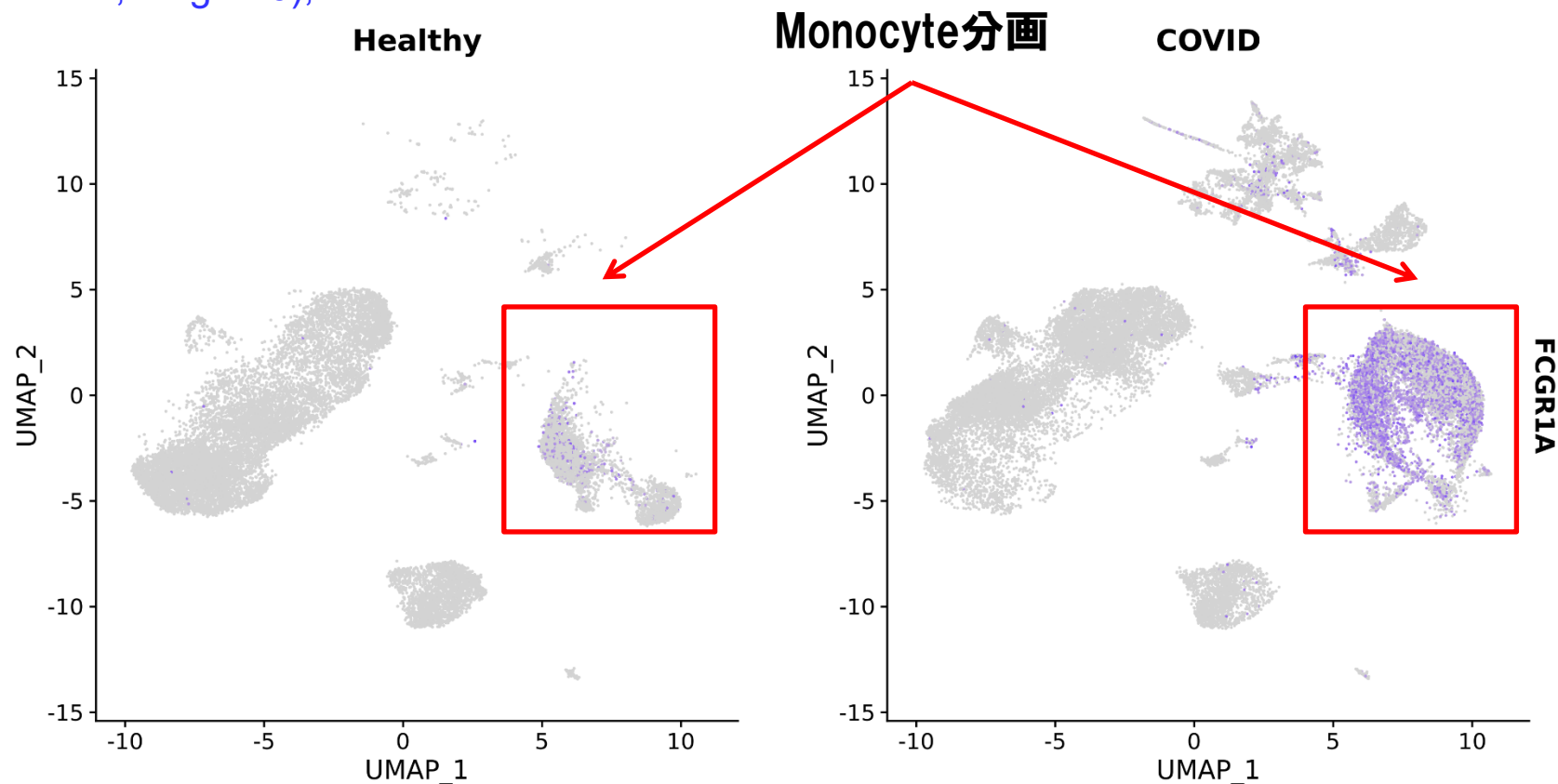
- ミトコンドリア遺伝子("MT-...")
  - リボソーム遺伝子("RPS/RPL...")
  - RNA遺伝子("RNA...")
- を解析対象から除外。

• CD16+ Monocyte分画(単球)においてケースコントロール間の発現量が異なる遺伝子群を同定してみます。FindMarkers()関数に実装されたWilcoxon rank sum test\*を使用してみます。

(※:様々な比較手法があり、全細胞へのWilcoxon検定はバイアスが入りやすい点に注意。)

## ② Seuratを使ったシングルセル解析実習

```
> gfile4 <- FeaturePlot(dataN, features = "FCGR1A" , split.by= "Status");  
> ggsave(gfile4, filename = "NatMed2020_COVID7HC6.Seurat.Featureplot.FCGR1A.png",  
  width=12, height=6);
```

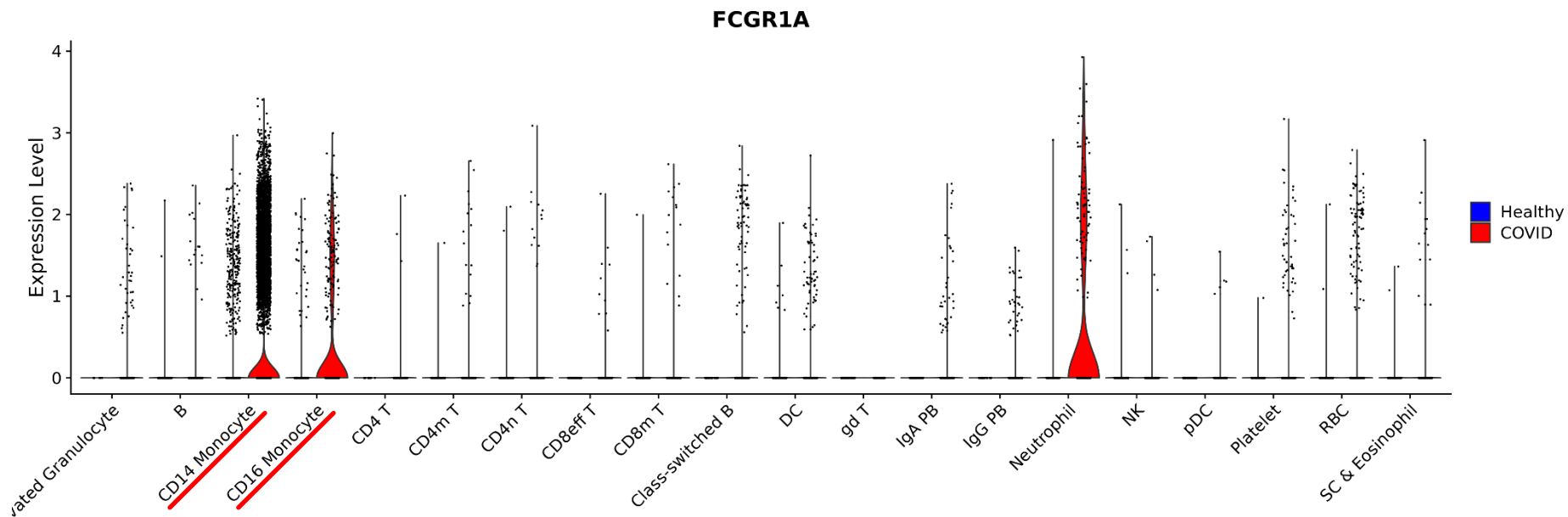


- 発現量差が認められたFCGR1A遺伝子に着目してみましょう。
- **Featureplot () 関数**を用いて、特定feature(=FCGR1A)をハイライトします。
- CD14+/CD16+ Monocyte分画で、ケース群での高発現を認めます。



## ② Seuratを使ったシングルセル解析実習

```
> gfile5 <- VlnPlot(dataN, features = "FCGR1A", split.by = "Status", group.by="cell.type", cols=c  
  ("blue","red"));  
> ggsave(gfile5, filename = "NatMed2020_COVID7HC6.Seurat.Violinplot.FCGR1A.png", width=16,  
  height=6);
```



- **VlnPlot () 関数**を用いて、各細胞分画毎の発現量分布を確認します。
- **CD14+ / CD16+ Monocyte**分画に加え、**Neutrophil**分画(好中球)でもケース群で高発現していることがわかりました。

## ② Seuratを使ったシングルセル解析実習

```
> HLAgenes <- c("HLA-DMA", "HLA-DMB", "HLA-DOA", "HLA-DOB", "HLA-DPA1", "HLA-DPB1",  
"HLA-DQA1", "HLA-DQA2", "HLA-DQB1", "HLA-DQB1-AS1", "HLA-DQB2", "HLA-DRA", "HLA-  
DRB1", "HLA-DRB5");  
> dataNH <- AddModuleScore(dataN, features=list(HLAgenes), name="class2hla");  
> head(dataNH);
```

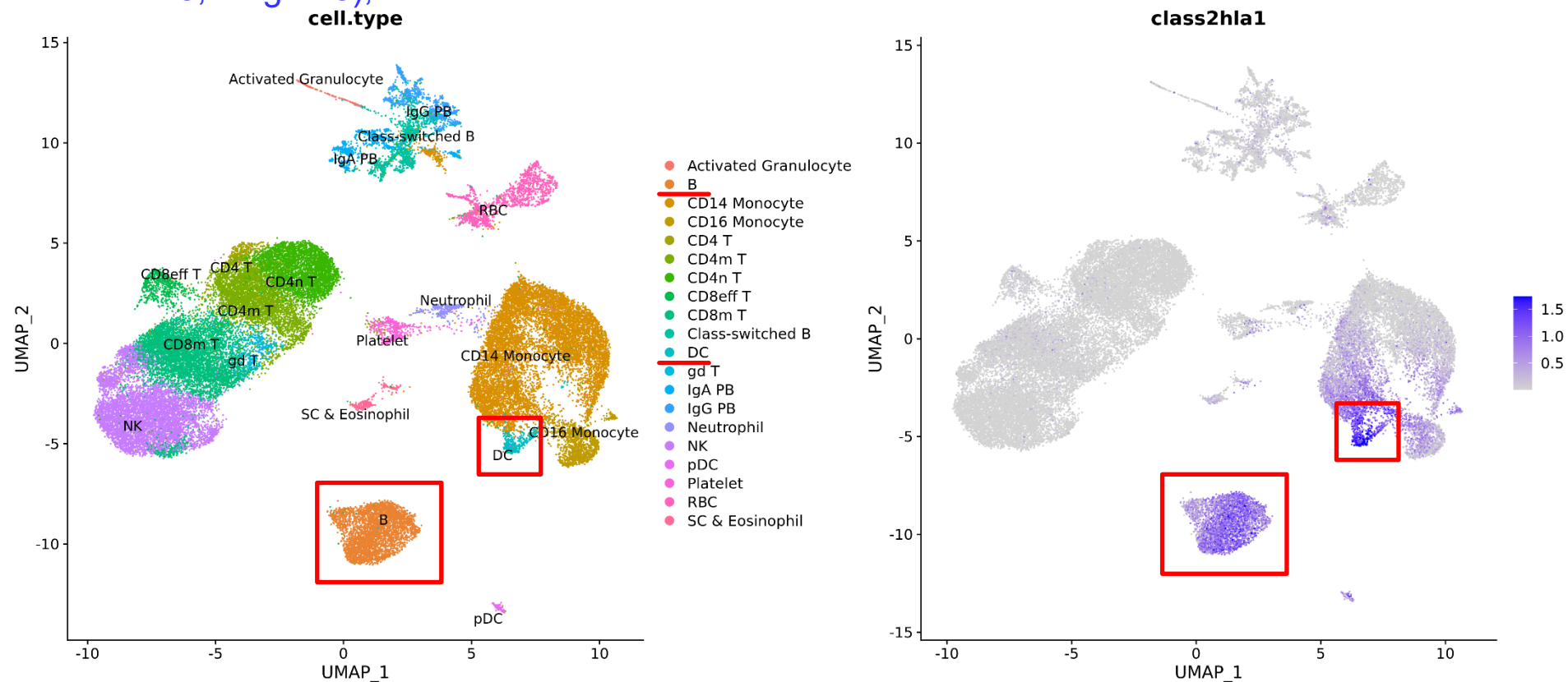
新たなメタデータとして付与

	orig.ident	nCount_RNA	nFeature_RNA	cell.type	Donor		Status	class2hla1
covid_555_1.1	covid_555_1	1222	125	RBC	C1	covid_555_1.1	COVID	-0.044251815
covid_555_1.2	covid_555_1	1099	160	Class-switched B	C1	covid_555_1.2	COVID	0.113498058
covid_555_1.3	covid_555_1	1055	212	IgG PB	C1	covid_555_1.3	COVID	-0.068152661
covid_555_1.7	covid_555_1	2411	312	Class-switched B	C1	covid_555_1.7	COVID	-0.078902627
covid_555_1.8	covid_555_1	2276	336	IgA PB	C1	covid_555_1.8	COVID	-0.096960292
covid_555_1.11	covid_555_1	1166	351	IgA PB	C1	covid_555_1.11	COVID	-0.144158706
covid_555_1.12	covid_555_1	1080	374	CD14 Monocyte	C1	covid_555_1.12	COVID	-0.169123354
covid_555_1.13	covid_555_1	1109	384	Class-switched B	C1	covid_555_1.13	COVID	0.006662368
covid_555_1.14	covid_555_1	1032	392	CD14 Monocyte	C1	covid_555_1.14	COVID	0.016202991
covid_555_1.15	covid_555_1	1048	401	CD14 Monocyte	C1	covid_555_1.15	COVID	0.013911387

- 複数遺伝子のセットを定義し、新たなメタデータとして付加が可能です。
- **AddModuleScore ()** 関数を用いて、**セット内に含まれる遺伝子の発現量の平均値 (=gene score)**を計算し※、付与してみます。
- 今回は、クラスII HLA遺伝子群を”\$class2hla”として定義してみます。  
(※:gene scoreの計算にあたっては、別途コントロール遺伝子群をランダムに定義し、細胞間のデータの違いを揃える目的で使用しています。詳細はSeuratチュートリアルをご参照ください。)

## ② Seuratを使ったシングルセル解析実習

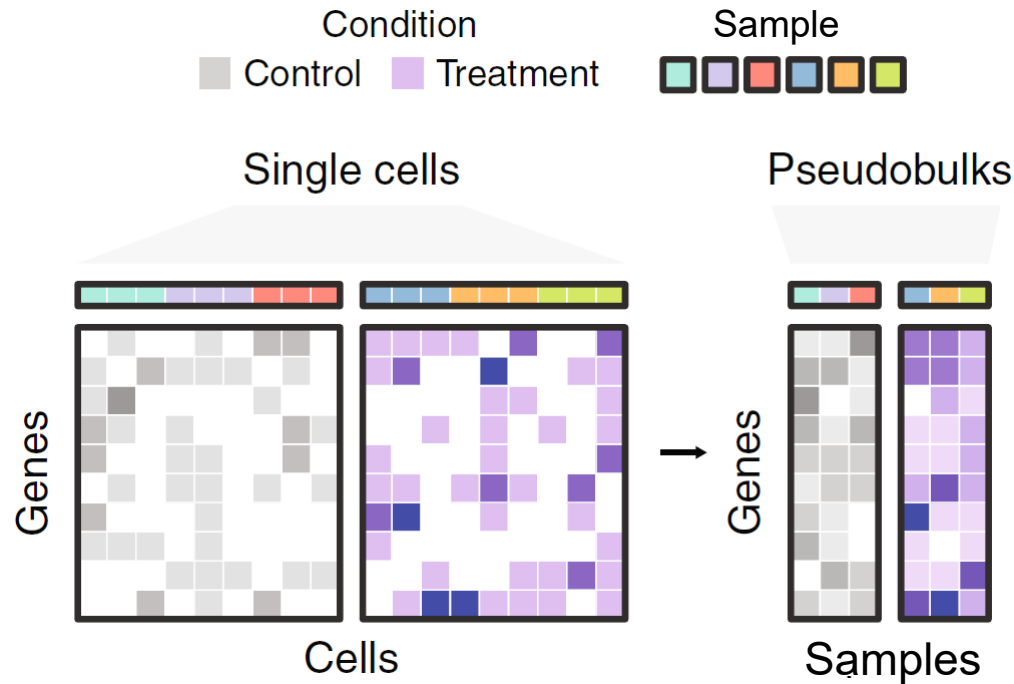
```
> gfile6 <- FeaturePlot(dataNH, features = "class2hla1", min.cutoff= "q1" , max.cutoff= "q99");  
> ggsave(gfile2+gfile6, filename = "NatMed2020_COVID7HC6.Seurat.Featureplot.HLAgenes.png",  
width=18, height=8);
```



- **Featureplot () 関数**で、HLA遺伝子群のgene score分布を確認します。
- B cell分画(B細胞)や、Dendritic Cell分画(DC; 樹状細胞)など、抗原提示細胞で高発現していることが確認できます。

## ② Seuratを使ったシングルセル解析実習

### シングルセルデータにおけるpseudo-bulk解析

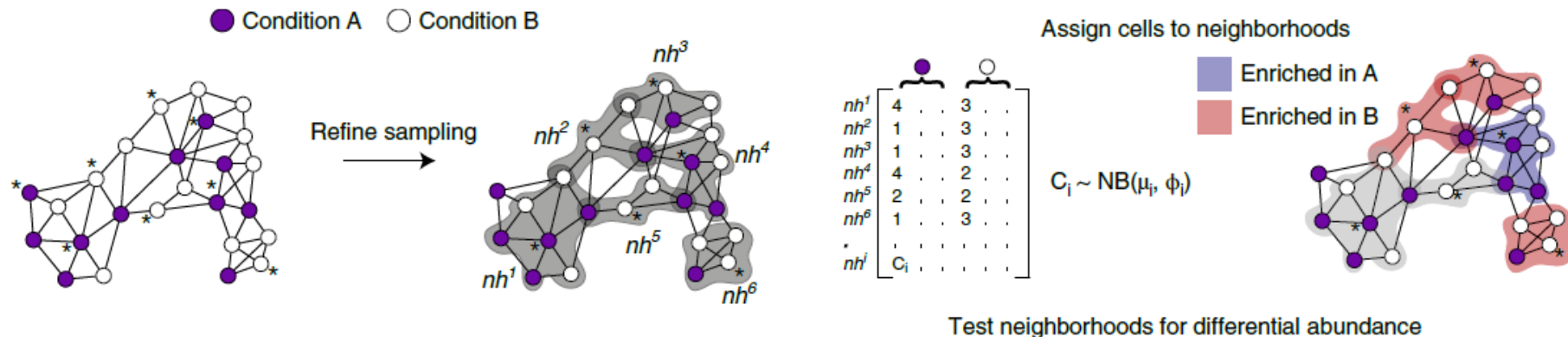


- 複数の細胞の遺伝子発現情報を、**サンプルや細胞分画単位に集約して扱う解析手法を、pseudo-bulk解析**と呼びます。
- Pseudo-bulk解析には、低発現遺伝子の扱い、疎なデータ、細胞間の不均一性といったシングルセル解析の技術的課題を軽減し、**既存の bulk RNA-seq解析手法を適用可能にする**など、**メリット**があります。

(Squair JW et al. *Nat Commun* 2021より改変)

## ② Seuratを使ったシングルセル解析実習

### グラフ構造に基づいた、似ている一細胞同士の紐づけ解析(Milo)



- Pseudo-bulk解析では、取得した一細胞解像度の情報が喪失します。
- 一方、全細胞を個別に扱う解析は計算コストの問題があります。
- 発現プロファイルが似ている細胞同士を互いに紐づけたり、まとめてグループ化して、中間的な性質のデータとして扱う方法も存在します。
- 本演習では、Miloパッケージを使用した解析を行います。
- Miloは、グラフ構造を利用して、各細胞の発現プロファイルから、似ている情報を持つ細胞を”neighborhoods”と定義します。

## ② Seuratを使ったシングルセル解析実習

```
> library(miloR);                ← Miloおよび関連ライブラリの起動
> library(SingleCellExperiment);
> library(patchwork);

> dataN_sub <- subset(dataN, downsample=1000); ← 計算負荷軽減のため最大1000細胞
> dataN_sce <- as.SingleCellExperiment(dataN_sub); ← となるようダウンサンプリング。その後オ
> dataN_milo <- Milo(dataN_sce);           ← ブジェクト形式をMilo解析用に変換

> dataN_milo <- buildGraph(dataN_milo, k = 20, d = 30);
> dataN_milo <- makeNhoods(dataN_milo, prop = 0.1, k = 20, d=30, refined = TRUE);
> dataN_milo <- countCells(dataN_milo, meta.data = data.frame(colData(dataN_milo)),
  sample="Donor");
> dataN_milo <- calcNhhoodDistance(dataN_milo, d=30);
  ↑ グラフ構造を基にNeighborhoodsを作成(実行時間:5分程度)
> dataN_design <- data.frame(colData(dataN_milo))[,c("Donor", "Status")];
> dataN_design <- unique(dataN_design);
> rownames(dataN_design) <- dataN_design$Donor; ← ケースコントロール情報を抽出

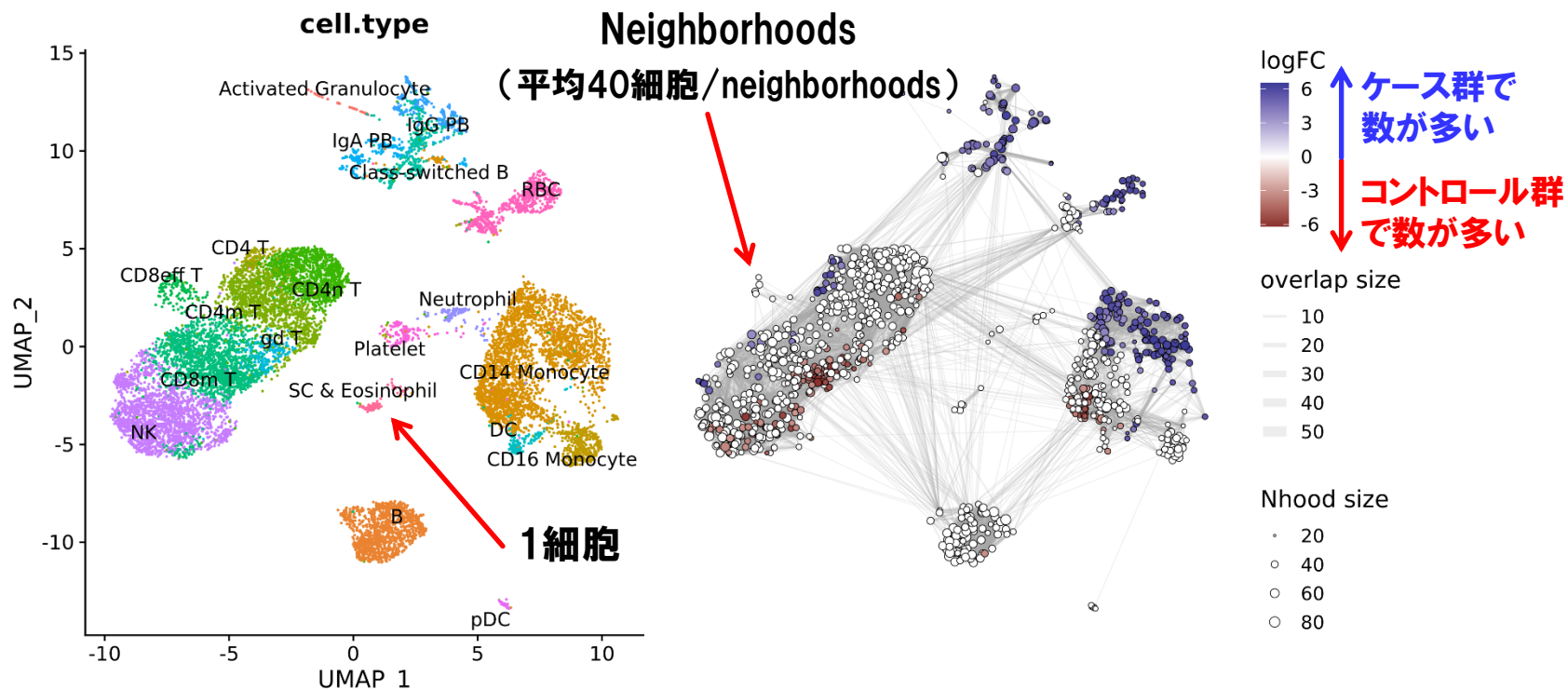
> da_results <- testNhoods(dataN_milo, design = ~ Status, design.df = dataN_design);
  ↑ ケースコントロール間でのNeighborhoodsのAbundance analysisを実施
```

• Miloパッケージを用いた解析を進めます(やや複雑な手順となります)。



## ② Seuratを使ったシングルセル解析実習

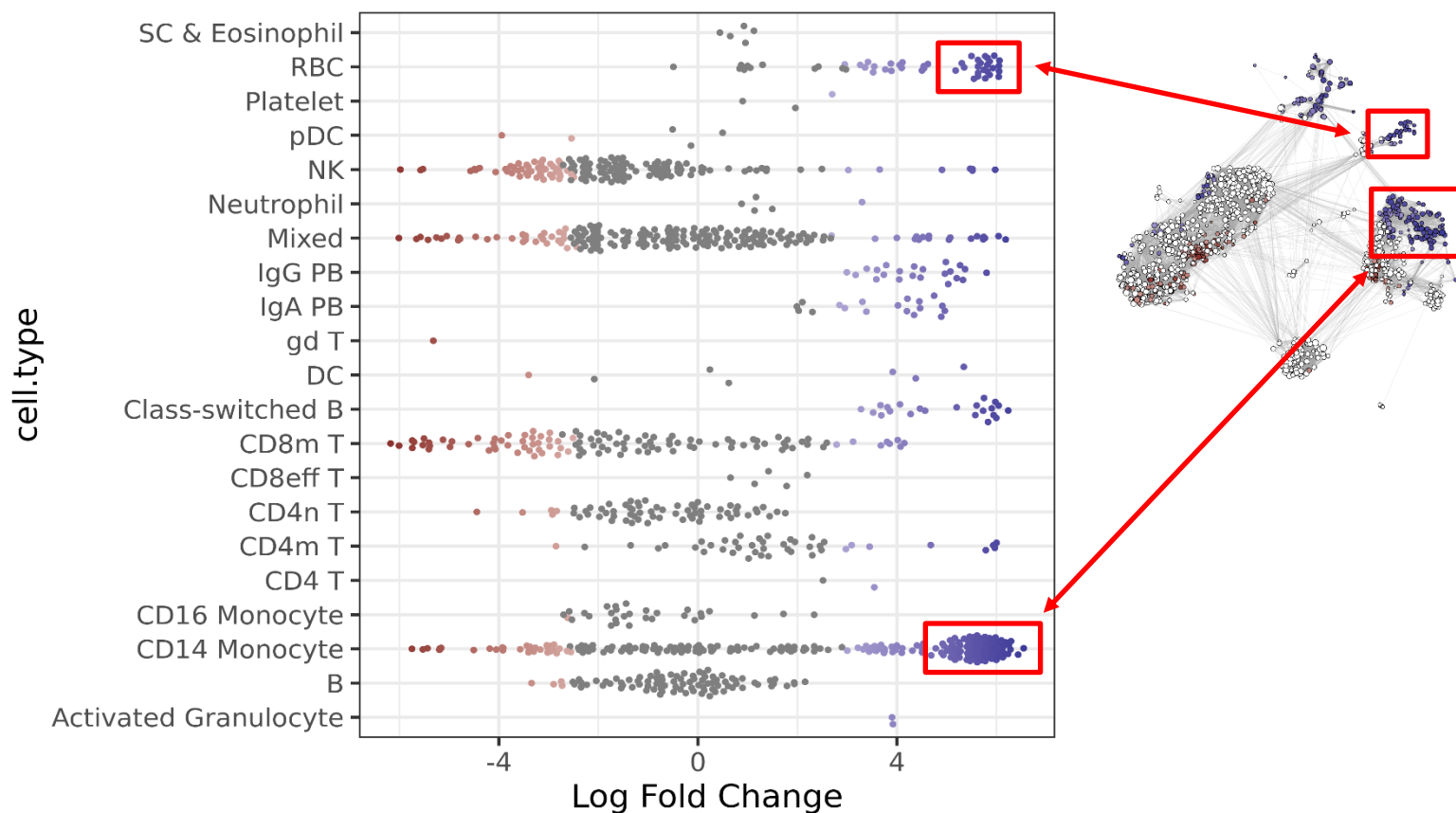
```
> dataN_milo <- buildNhoodGraph(dataN_milo);  
> gfile7 <- DimPlot(dataN_sub, group.by = "cell.type", label=TRUE, repel=TRUE) + NoLegend();  
> gfile8 <- plotNhoodGraphDA(dataN_milo, da_results, alpha=0.05) + plot_layout(guides="collect");  
> ggsave(gfile7+gfile8, filename = "NatMed2020_COVID7HC6.Milo.UMAP_Neighborhoods.png",  
width=12, height=6);
```



- 13,000細胞を対象に、1,074個のneighborhoodsが定義されました※
  - 各neighborhoodsで、ケースコントロール間の細胞数の大小を比較。
- (※:定義されるneighborhoods数は、グラフ構造推定時のパラメーターや乱数seedに依存。)

## ② Seuratを使ったシングルセル解析実習

```
> da_results <- annotateNhoods(dataN_milo, da_results, coldata_col="cell.type");  
> da_results$cell.type <- ifelse(da_results$cell.type_fraction < 0.7, "Mixed", da_results$cell.type);  
> gfile9 <- plotDAbeeswarm(da_results, group.by="cell.type");  
> ggsave(gfile9, filename = "NatMed2020_COVID7HC6.MIlo.beeswarmplot.png",width=8, height=8);
```



- Neighborhoods単位の細胞数ケースコントロール比較を、細胞分画に別に集計。CD14+ Monocyte・RBC(赤血球)のケース群での増加を確認

## 終わりに

- シングルセル解析を巡る現状について確認の上、有名な解析ツールであるSeuratを使ったシングルセルデータ解析演習を行いました。
- シングルセル解析技術の発展と普及は著しく、次々と**新しい実験技術や情報解析ツール**が開発されています。
- 今後は、**実験技術のハイスループット化とコスト低下**が進み、より**大規模サンプル由来の多彩な細胞組織**を対象に、**多数の一細胞オミクス情報**を、**多層的に取得**するようになると予測されます。
- 常に最先端の情報を取り入れ、実践することが重要になりそうです。
- 一方、高額な実験機器や専門的な情報解析技術が要求されるため、一つの研究室で完結するのではなく、**施設内コアファシリティ**の充実や**専門家との共同研究体制、情報共有の場**の確保が鍵となりそうです。