2019年01月31日

やってみよう！
single cell RNA-seq解析ことはじめ

東京大学
定量生命科学研究所

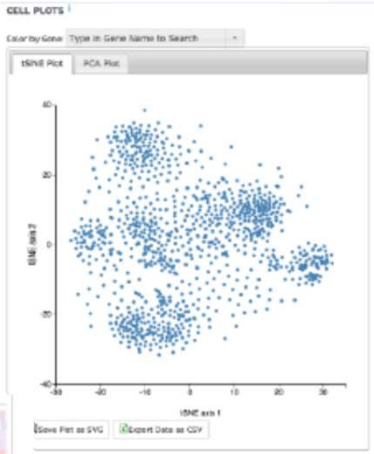
林　寛敦

scRNA-seq解析のワークフロー

ddSEQ

ddSEQ Single-Cell Isolator
(BIO-Rad)

SuewCell WTA 3′ Library Prep Kit
(illumine)



Experimental Design → **Sequence** → **Processing Reads** → **Preparing Expression Matrix** → **Biological Interpretation**

Cell Isolation
Library Preparaion
etc.

Cell Ranger
(Chromium: 10xgenomics)
etc.

その他
3rd. パーティー
(FlowJo, SeqGeq etc.)

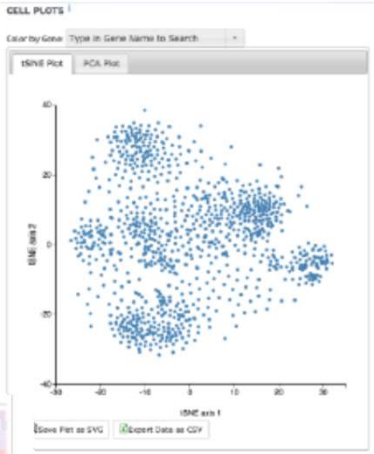
(illumine webinar RNA-seqをはじめよう　改変)

scRNA-seq解析のワークフロー

ddSEQの例

ddSEQ Single-Cell Isolator
(BIO-Rad)

SuewCell WTA 3′ Library Prep Kit
(illumine)



本日のメインテーマ

Experimental Design → **Sequence** → **Processing Reads** → **Preparing Expression Matrix** → **Biological Interpretation**

Cell Isolation
Library Preparaion
etc.

CellRanger
(Chromium: 10xgenomics)
etc.

その他
3rd. パーティー
ソフトウェア

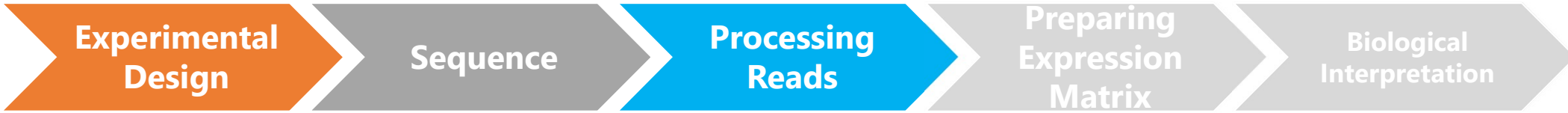
(illumine webinar RNA-seqをはじめよう　改変)

本日の流れ

- **1細胞RNA-seq (scRNA-seq)のワークフロー**
- **ddSEQのデータを用いた解析事例**
  - 解析の流れ・解析環境
    - BaseSpase → R (Seurat2)
  - R package Seurat2による解析
    - データのクオリティチェック
    - 細胞のクラスタリング
    - クラスター間で発現量に差がある遺伝子の抽出
    - クラスターのアノテーション
  - chromiumのデータとの比較
- **腫瘍検体のscRNA-seq解析事例**
  - monocle, velocytoによる解析例

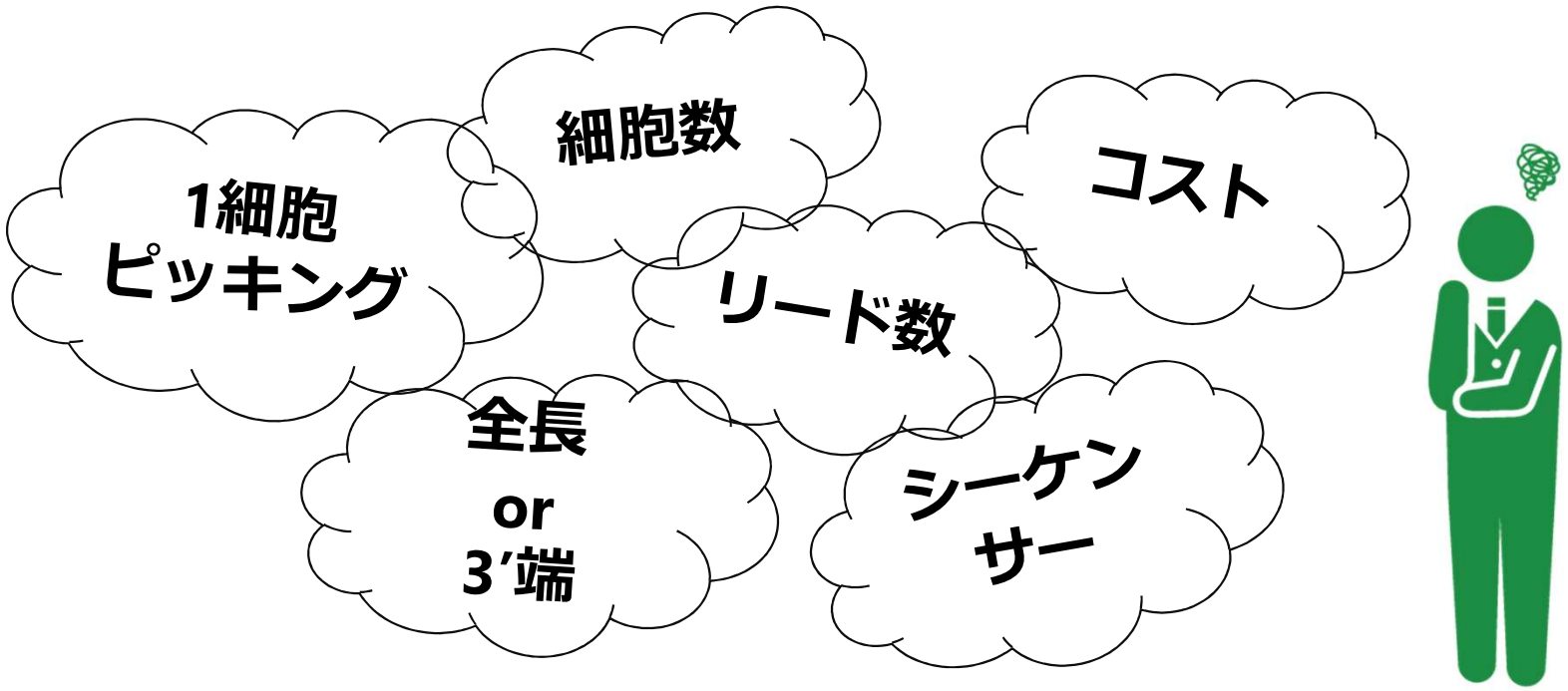
本日の流れ

- **1細胞RNA-seq (scRNA-seq)のワークフロー**
- **ddSEQのデータを用いた解析事例**
  - 解析の流れ・解析環境
    - BaseSpase → R (Seurat2)
  - R package Seurat2による解析
    - データのクオリティチェック
    - 細胞のクラスタリング
    - クラスター間で発現量に差がある遺伝子の抽出
    - クラスターのアノテーション
  - chromiumのデータとの比較
- **腫瘍検体のscRNA-seq解析事例**
  - monocle, velocytoによる解析例



Experimental Design
Sequence
Processing Reads

Single cell RNA-seqの解析プラットフォームは多数



目的に応じた解析プラットフォームの選択が重要

Experimental Design | Sequence | Processing Reads | Preparing Expression Matrix | Biological Interpretation

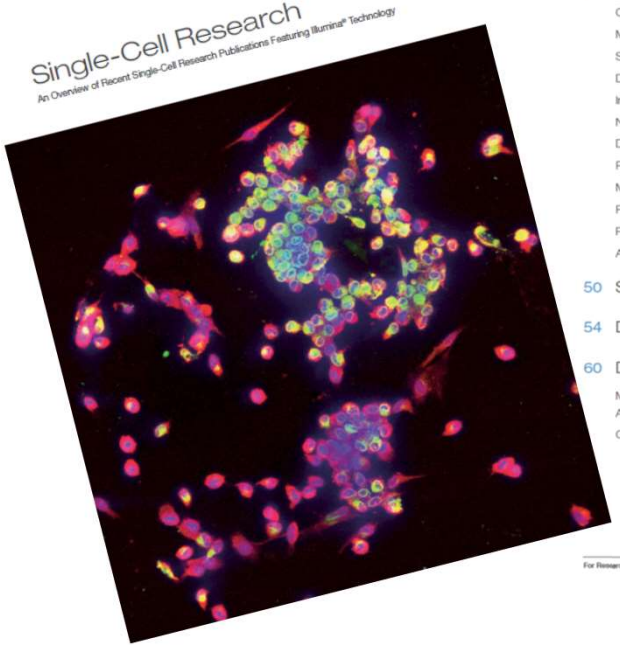


TABLE OF CONTENTS

- 様々な学術分野の動向を網羅 (Application)
- 1細胞を取得する技術の紹介 (Sample Preparation)
- データ解析ツールの紹介 (Data Analysis)
- **DNA**
- **Epigenome**
- **RNA**

に関する1細胞解析手法の紹介

手法ごとにメリット・デメリットが記載

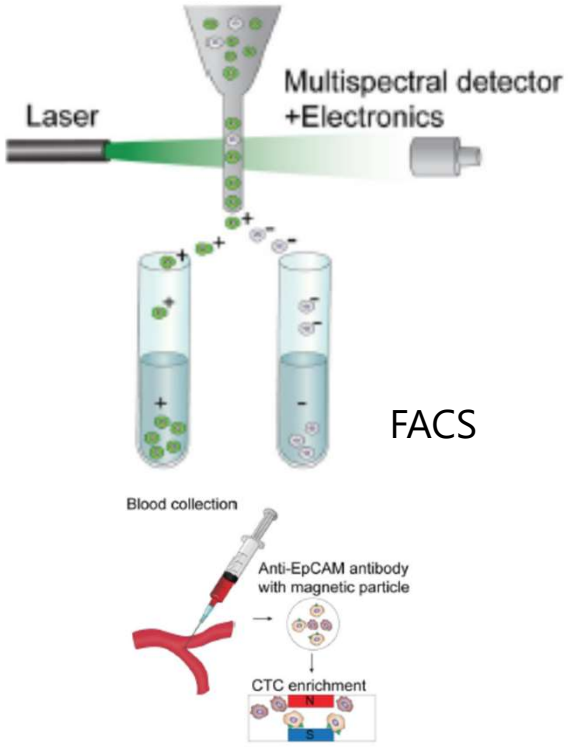
充実した参考文献

主にNGSによる1細胞解析の情報が網羅的にわかりやすくまとめられている

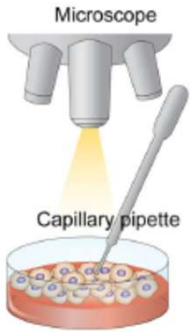
Experimental Design | Sequence | Processing Reads | Preparing Expression Matrix | Biological Interpretation

Isolation

Auto Sampling

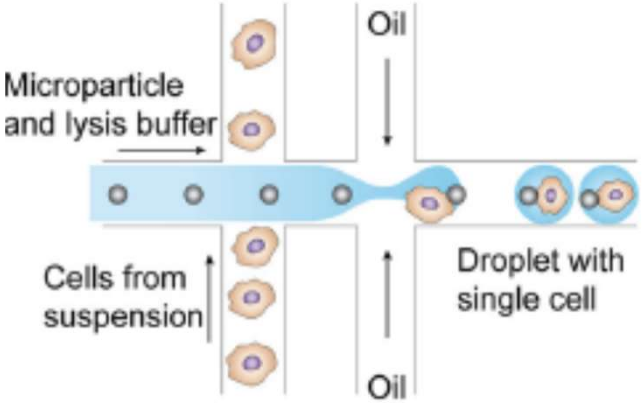


Manual Picking



MicroFluidics

Fluidigm C1



ddSEQ (Bio-Rad)
Chromium (10x genomics)

その他、様々な手法が存在

Hwang et al., Exp. Mol. Med. 50: 96 (2018)

Library Preparation

代表的な手法の例

Platform	Smart-seq	MARS-seq	CEL-seq	Drop-seq
Region	Full-length	3′ end	3′ end	3′ end
Target read depth (per cell)	(10^6)	(10^4)–(10^5)	(10^4)–(10^5)	(10^4)–(10^5)
UMI	None	Yes	Yes	Yes
Amplification	PCR	IVT	IVT	PCR
Feature	Isoform analysis	FACS sorting Multiplex barcoding	Linear amplification (pool cDNAs for IVT)	Emulsion Low cost

scRNA single-cell RNA sequencing, *Smart-seq* novel full-transcriptome mRNA-sequencing protocol, CEL-seq cell expression by linear amplification and sequencing, *Drop-seq* droplet sequencing, *IVT*in vitro transcription, *UMI*unique molecular identifier, *FACS*flow-activated cell sorting, *MARS-seq*massively parallel RNA single-cell sequencing framework

UMI = 分子バーコード = Unique mRNA

Full-length or 3’end?　　Number of cells?

Read depth?　　Cost?

Hwang et al., Exp. Mol. Med. 50: 96 (2018)

Smart-seq v2

Full-length
1 cell – 1 reaction

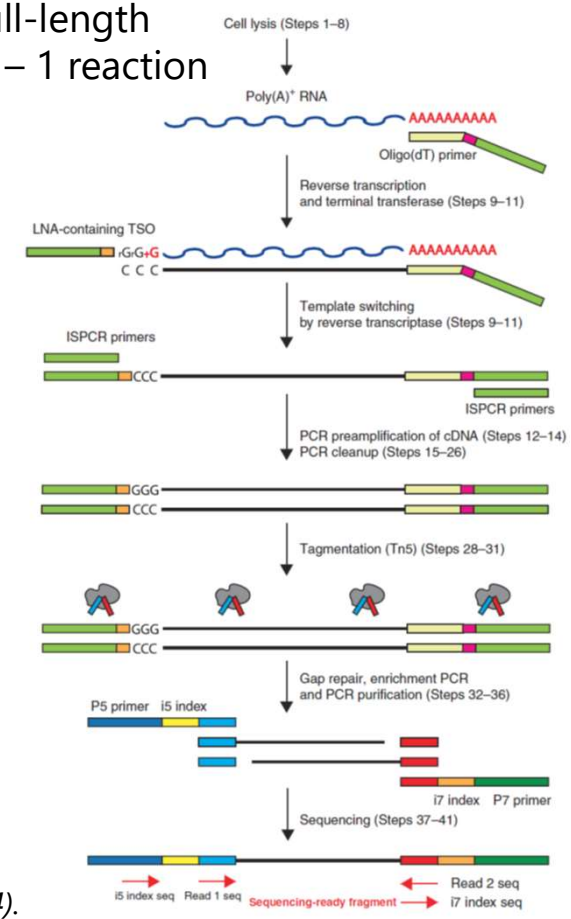
Capture Poly (A)⁺ RNA

Reverse Transcription
&
Template switching

PCR preamplification

Tagmentation (Tn5)

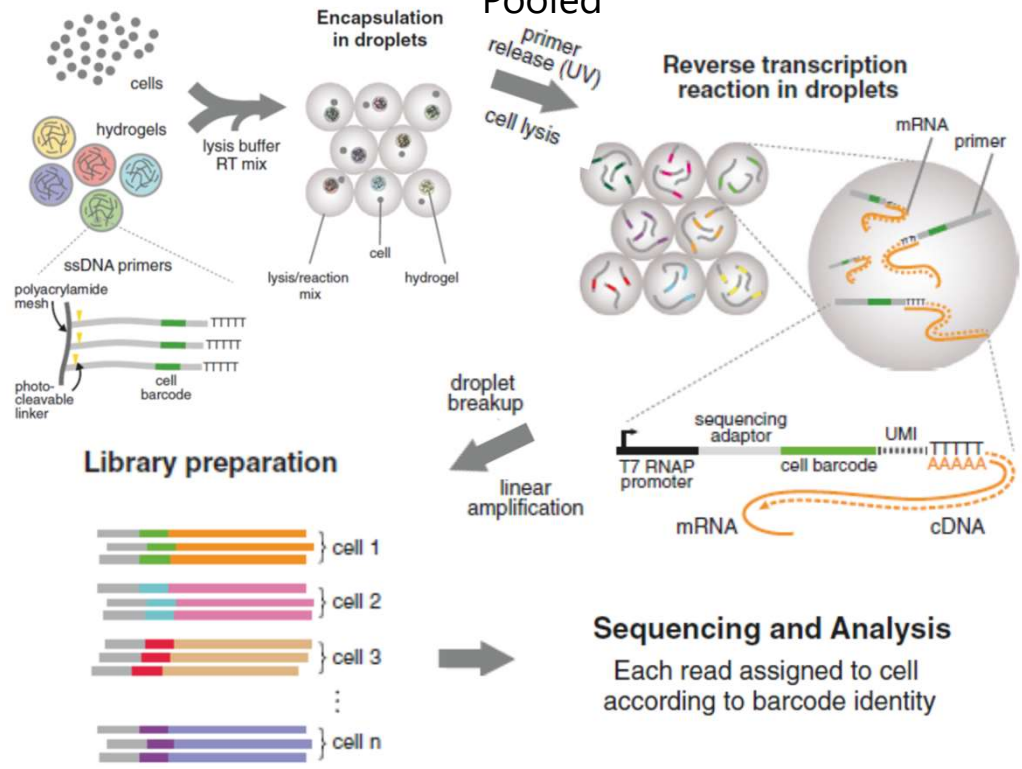
Index PCR
&
Sequencing



Nat Protoc. 9, 171-181 (2014).

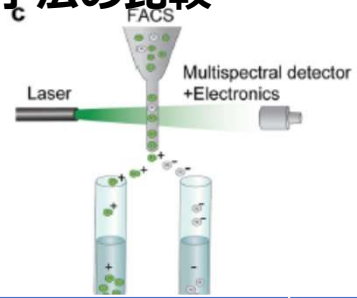
Droplet-base

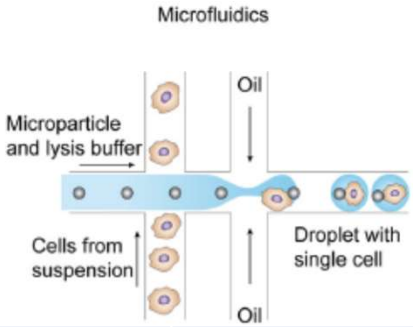
3'end UMI
Pooled



(inDrop-seq) *Cell 161:1187-1201 (2015).*

代表的な手法の比較





	Smart-seq	Droplet-base
配列・定量	**全長配列** （リード数）	**3’ 部分配列** (UMIによるカウント)
細胞数	100~384 cells	1,000-10,000 cells/well
1細胞当たりの同定遺伝子数	3,000遺伝子以上 （比較的安定）	1,500～6,000遺伝子 （細胞種による）
再解析	任意の細胞の再解析が可能 （個々の細胞が別ライブラリー） qRT-PCRも可能	個別の再解析不可能
特徴	集団構成推定が難しい （狭く深く） アイソフォームの解析可能	1細胞当たりの情報量が少ない （広く浅く） 多数の細胞を解析可能
シーケンスコスト (/1細胞)	そこそこ	非常に安価

研究目的に合わせて使い分け

Experimental Design | **Sequence** | Processing Reads | Preparing Expression Matrix | Biological Interpretation

シーケンス例

Hiseq2500 Rapid

Hiseq Rapid SBSキット V2	
リード長	100 bp x 2
シングルフローセル	50~60G
ランタイム	27時間
リード数	3 億ペアリード
クオリティ	100bp x 2で、80%以上の塩基がQ30を超える

Chromium (10x genomics)で調製した2,000細胞のサンプルを3サンプル分、合計6,000細胞をシーケンス

1細胞当たりのリード数の推奨リード数である50,000ペアリード/cell
80　～　100　円/cell (Total 6,000 cells)

NOVA-seq 6000

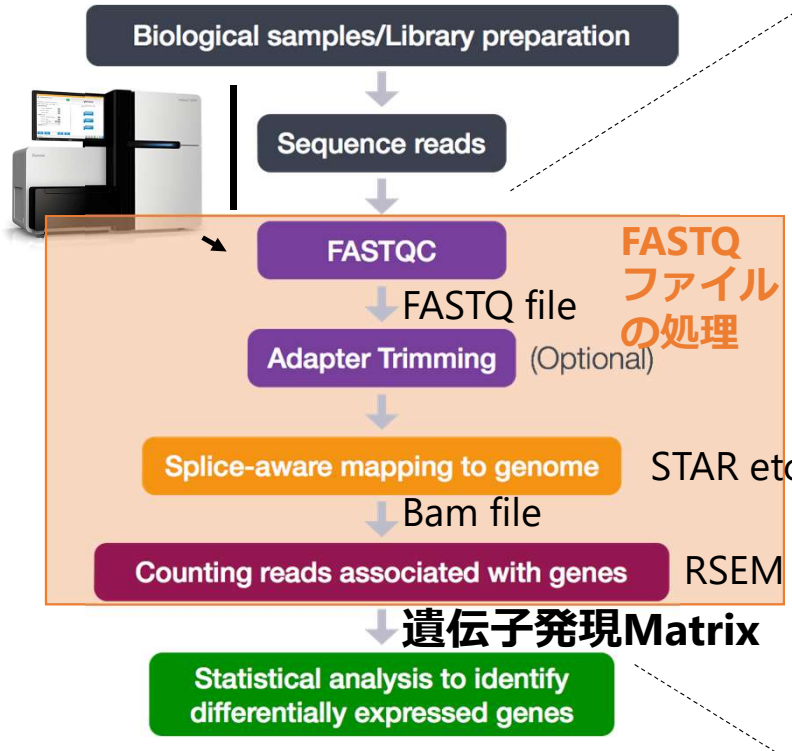
<u>より多くのサンプル、細胞数、リード数で安価にシーケンスしたい!!</u>

従来と同等のデータ量を約1/2のコストで実現可能

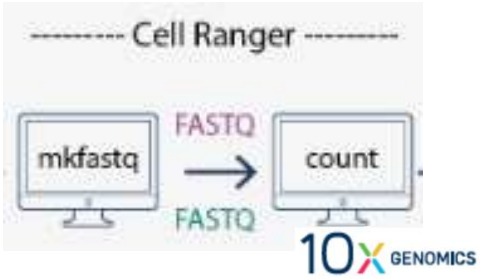
多くのシーケンス受託会社で稼働

Experimental Design | Sequence | **Processing Reads** | Preparing Expression Matrix | Biological Interpretation

scRNA-seq解析パイプライン例



Introduction to RNA-Seq using high-performance computing
https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/02_assessing_quality.html



(wikipedia)

BaseSpace

- 直感的に操作可能
- 主要アプリケーションを搭載
  - 全ゲノム解析
  - エクソーム解析
  - RNA-seq解析
  - scRNA-seq解析 (ddSEQ)
  - 腫瘍/正常細胞解析　など

Cell Ranger

- 10x genomicsが提供する NGS solutionのデータを解析
- CUI(コマンド) での操作だが簡便
- 充実したチュートリアル

自前の解析サーバー

- 構築の難度が高いが、その分汎用性が高い
- メンテナンスが面倒

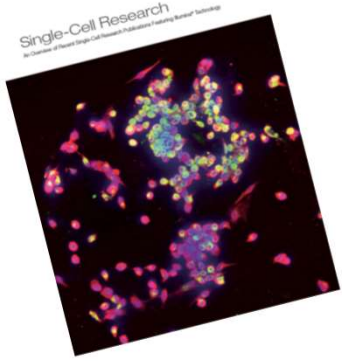
3rd Party

Experimental Design | Sequence | Processing Reads | Preparing Expression Matrix | Biological Interpretation

ここまでのまとめ

研究目的に適した解析プラットフォームを選択する

- 予算
- 細胞数・リード数
- 全長 or 3'end
- 細胞の取得方法・ライブラリー調製方法



シーケンスのコストをさげることはできる

- 多数のサンプルでNova-seq

サーバーやLinuxのスキルがなくとも解析できる方法はある

- BaseSpaceの活用
  ddSEQ (Droplet-base)やNexteraによって
  ライブラリー調製する方法 (Smart-seq V2など)には確実に対応
- プラットフォーマーの解析パイプライン
- 3rd Party解析ソフト、共同研究、外注

BaseSpace

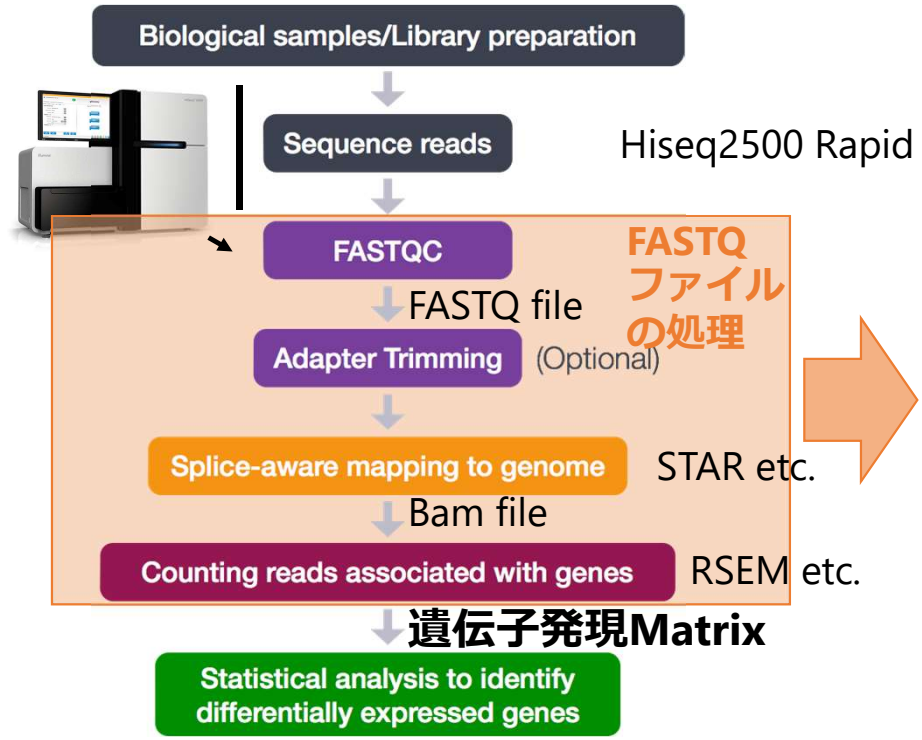
本日の流れ

- 1細胞RNA-seq (scRNA-seq)のワークフロー
- **ddSEQのデータを用いた解析事例**
  - 解析の流れ・解析環境
    - BaseSpase → R (Seurat2)
  - R package Seurat2による解析
    - データのクオリティチェック
    - 細胞のクラスタリング
    - クラスター間で発現量に差がある遺伝子の抽出
    - クラスターのアノテーション
  - chromiumのデータとの比較
- 腫瘍検体のscRNA-seq解析事例
  - monocle, velocytoによる解析例

Prepare Expression Matrix

Experimental Design | Sequence | **Processing Reads** | Preparing Expression Matrix | Biological Interpretation

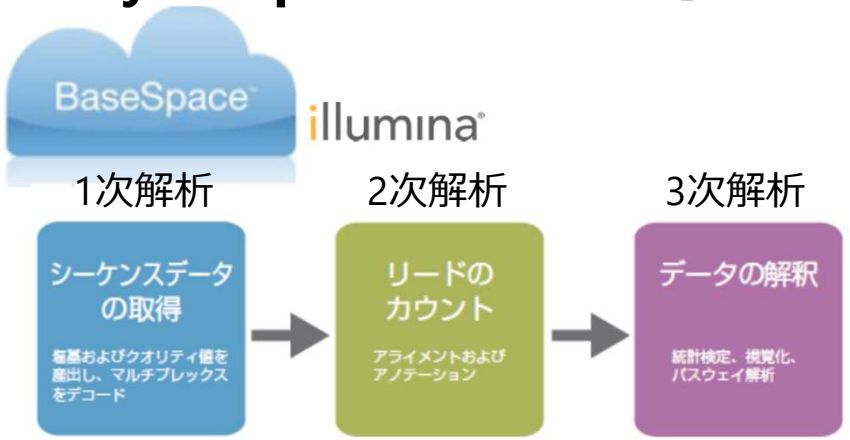
Library Preparation



ddSEQ Single-Cell Isolator (Bio-Rad)

SureCell WTA 3′ Library Prep Kit (illumine)

Data Analysis Pipeline for ddSEQ

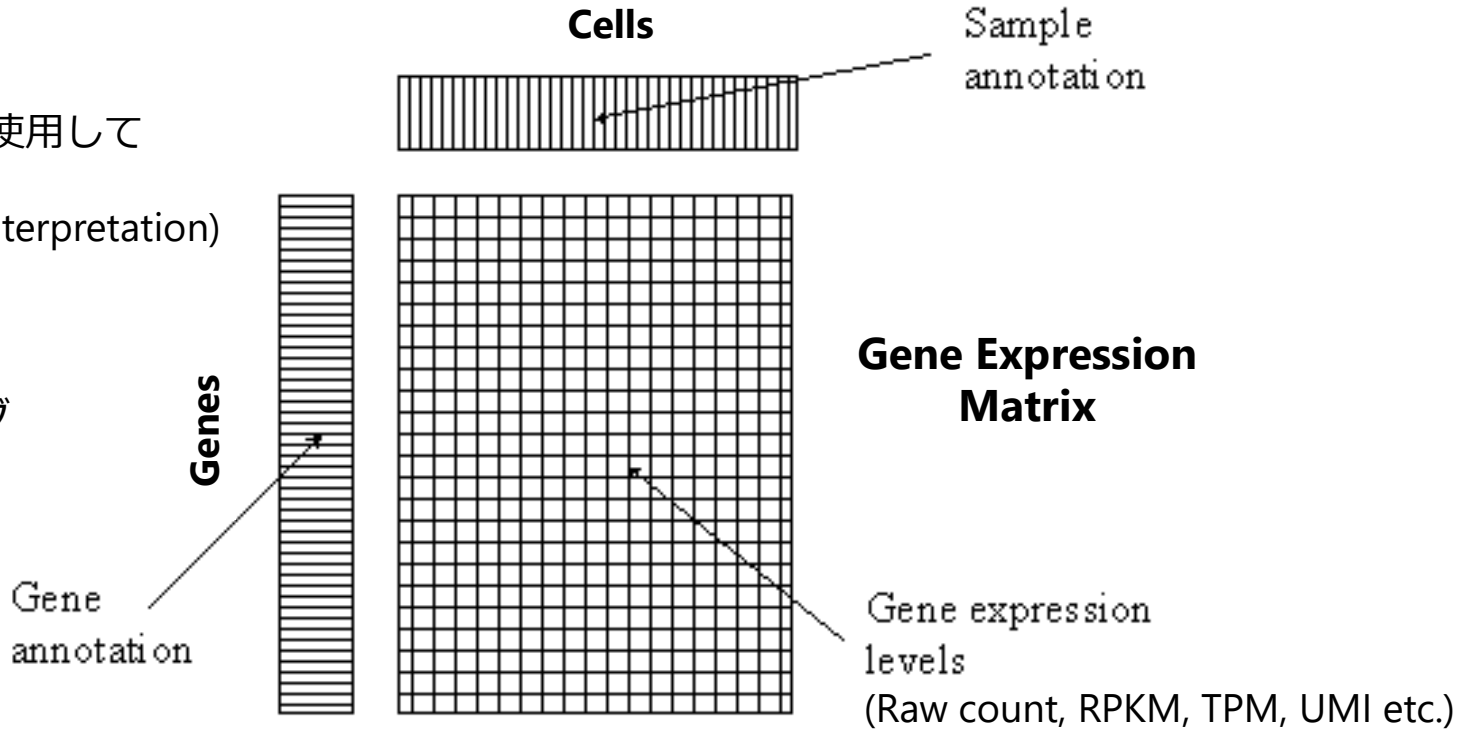


Introduction to RNA-Seq using high-performance computing
https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/02_assessing_quality.html

遺伝子発現マトリックス

大体の場合は、
遺伝子発現マトリックスを使用して
3次解析
(データの解釈, Biological Interpretation)

Ex.
発現量解析
クラスタリング



http://www.people.vcu.edu/~mreimers/OGMDA/gene_expression_matrix.gif

ddSEQのデータを用いた解析事例

- **Sample**

  Mixture of 4 cultured cell lines

  3 Ovarian Clear Cell Carcinoma (OCCC) Cell Lines

  OVISE　JHOC5　ES2

  1 Ovarian Surface Epithelial Cell Line

  OSE3

- **Target number of cells**

  1,200 cells (4 well)

Sequence Summary (ddSEQ)

解析結果 (BaseSpace)

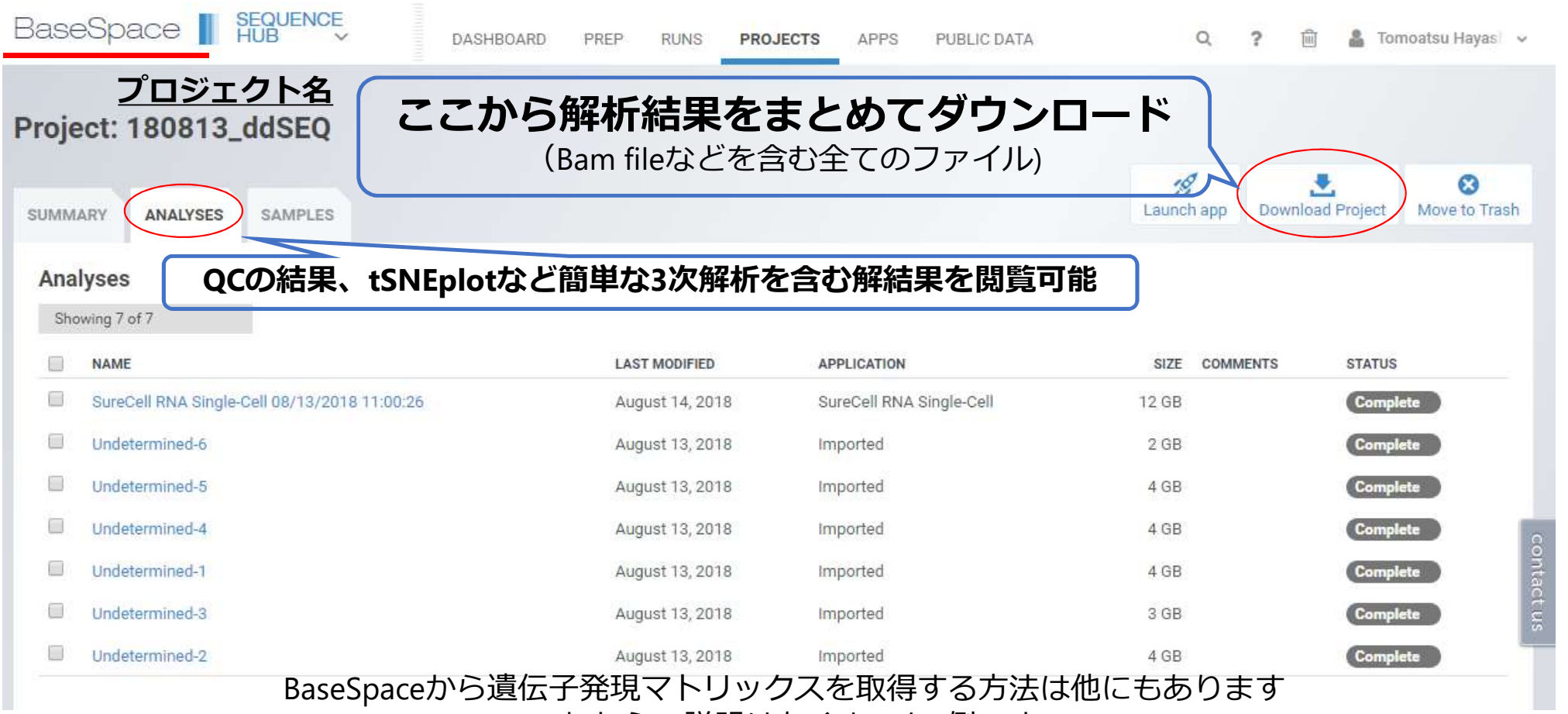
Cell Information

Total observed bead barcodes	169,199
Cells above background	1,312
Cells passing knee filter	1,212
UMI threshold for cells above background	652
UMI threshold for cells passing knee filter	1,139
Genic reads assigned to cells passing filter	29,786,321 (92.70%)
Median genic reads per cell passing filter	20,849
Median genic UMIs per cell passing filter	7,696
Median genes detected in cells passing filter	2,720

想定通りの細胞数がFilterを通過している

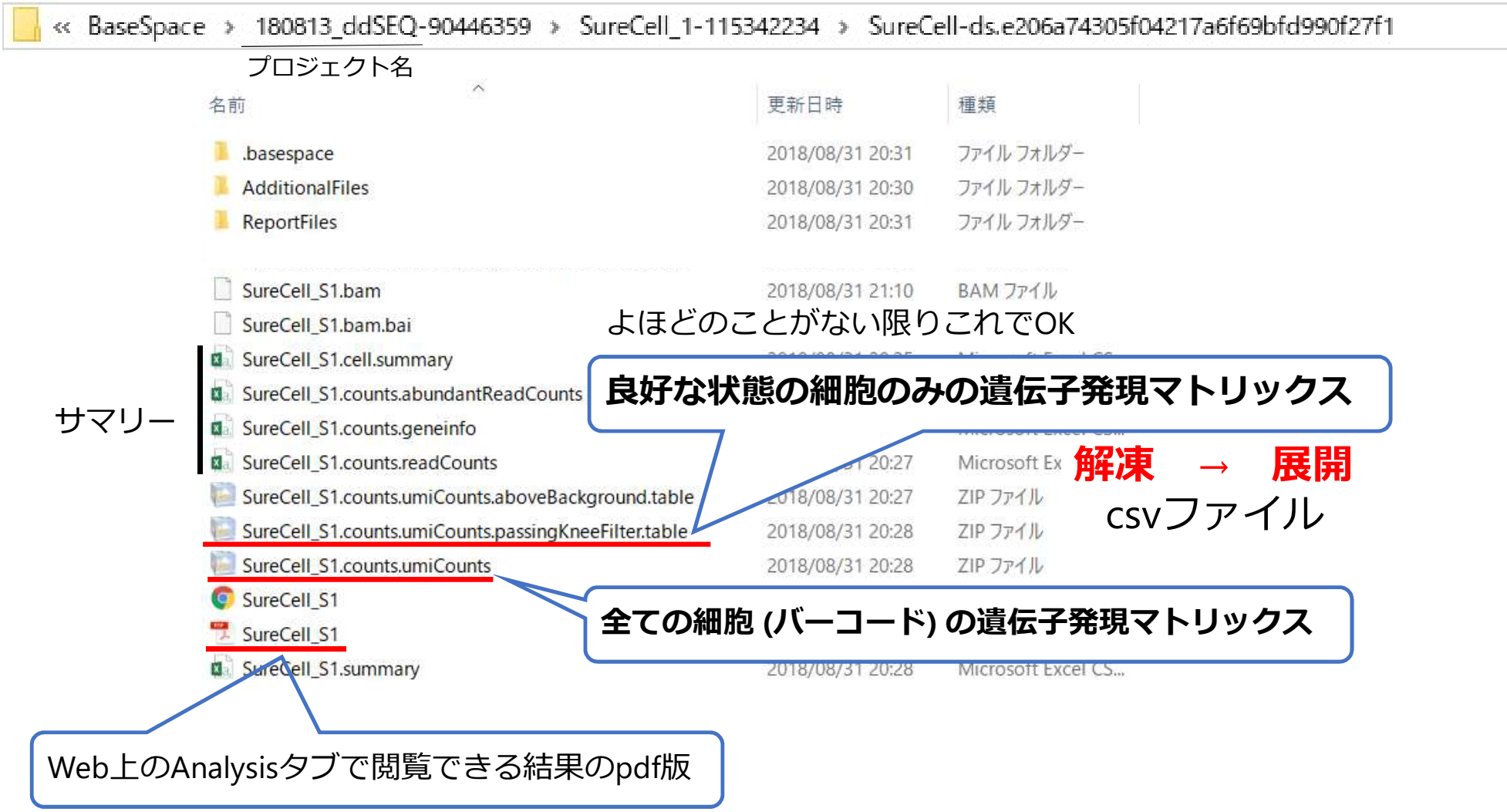
Preparing Expression Matrix from BaseSpace

プロジェクトの解析結果の画面



BaseSpaceから遺伝子発現マトリックスを取得する方法は他にもあります
これからの説明はあくまでも1例です

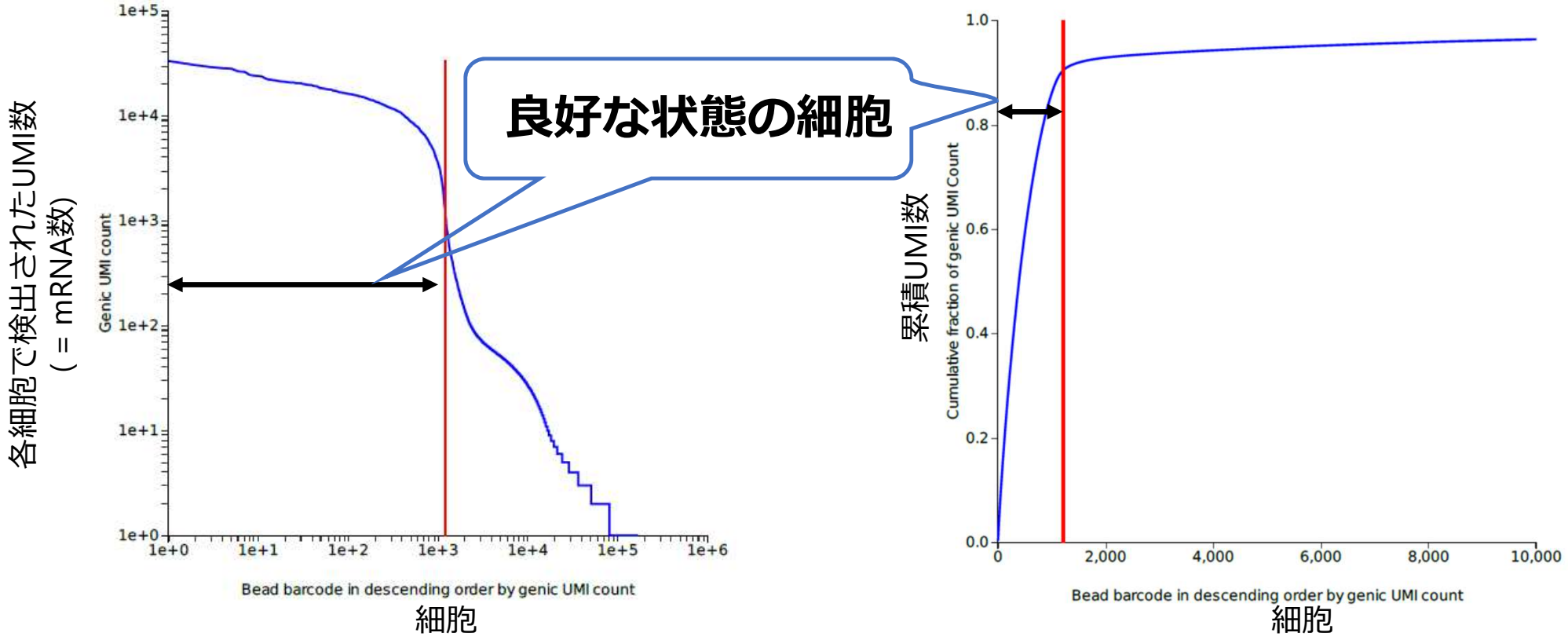
ダウンロード先をBaseSpaseフォルダに指定した場合



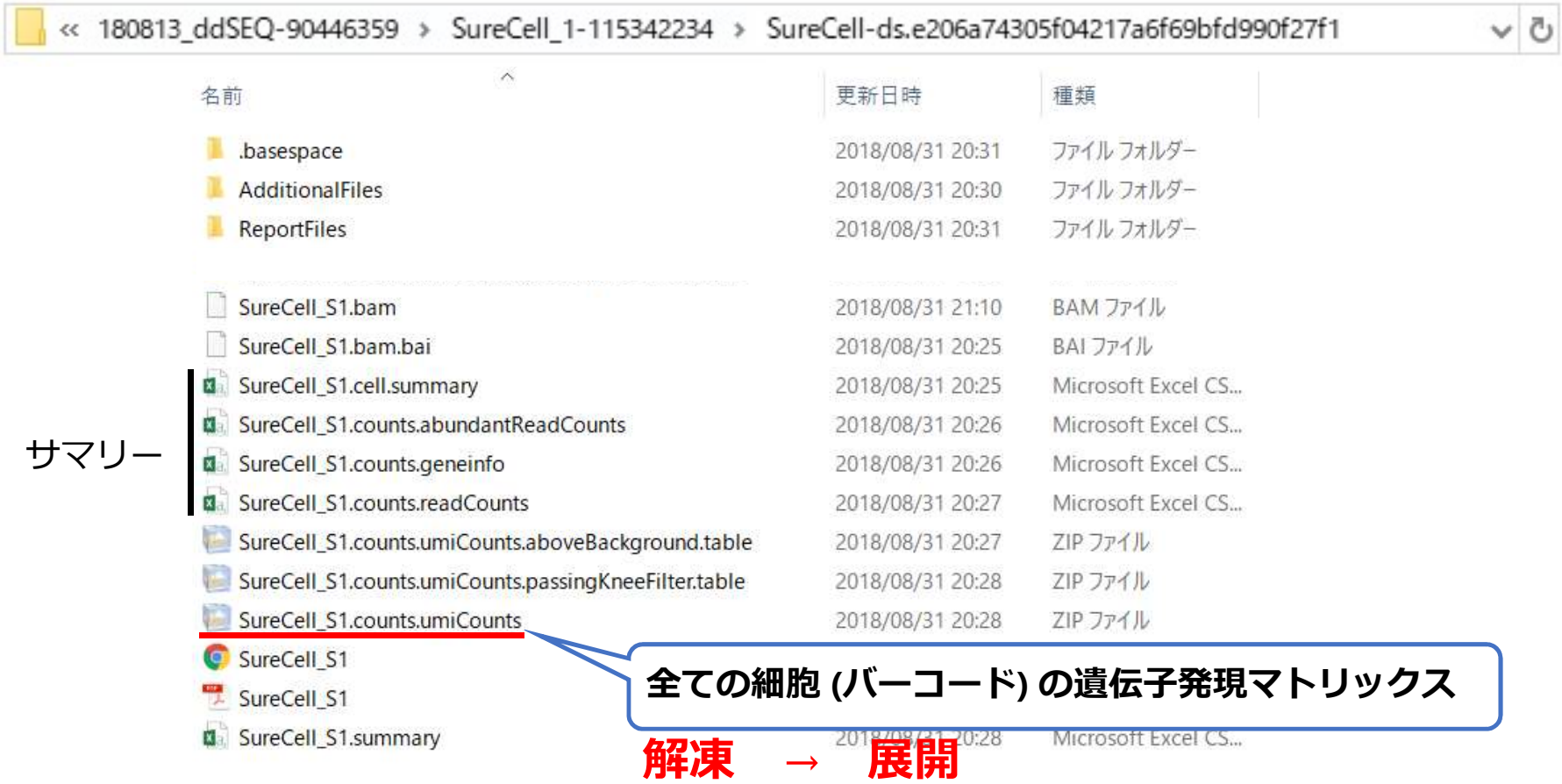
良好な状態の細胞？？

＝

細胞のダメージが少ない
適切な情報（遺伝子発現データ）を有している細胞



解析に使用する細胞を自分で選びたい



解析に使用する細胞を自分で選びたい

SureCell_S1.counts.umiCounts

Long Formatの形式

細胞バーコード　遺伝子　UMI Count

	A	B	C
1	CellId	GeneId	Count
2	atccggggtaggaattgg	A1BG	1
3	atccggggtaggaattgg	A1BG-AS1	1
4	atccggggtaggaattgg	A4GALT	1
5	atccggggtaggaattgg	AACS	3
6	atccggggtaggaattgg	AAED1	1
7	atccggggtaggaattgg	AAGAB	3
8	atccggggtaggaattgg	AAMP	2
9	atccggggtaggaattgg	AAR2	1
10	atccggggtaggaattgg	AARS	9
11	atccggggtaggaattgg	AARS2	3
12	atccggggtaggaattgg	AASDH	2
13	atccggggtaggaattgg	AASDHPPT	2
14	atccggggtaggaattgg	AATF	7
15	atccggggtaggaattgg	ABCA12	2
16	atccggggtaggaattgg	ABCB1	1
17	atccggggtaggaattgg	ABCC3	9
18	atccggggtaggaattgg	ABCC4	2

どれくらい長い？

最低でも、
フィルターをパスした細胞数 1,212 x 平均同定遺伝子数　2,720
= 3,296,640　行以上

Wide Format

	Runs ==>		
<== Genes GeneID	ERR1404137	ERR1404140	ERR1404150
b0001	532	60	655
b0002	317	46	132
b0003	92	20	75
b0004	150	33	85
b0005	507	31	82
b0006	62	11	31

Long Format

Runs	Genes	Values
ERR1399576	b0001	819
ERR1399576	b0002	4224
ERR1399576	b0003	1459
ERR1399576	b0004	2864
ERR1399576	b0005	92
ERR1399576	b0006	219
ERR1399576	b0007	116
ERR1399576	b0008	9327
ERR1399576	b0009	382
ERR1399576	b0010	90

一般的な遺伝子発現マトリックス
(http://genomespot.blogspot.com/2018/11/using-dee2-bulk-data-dumps.html)

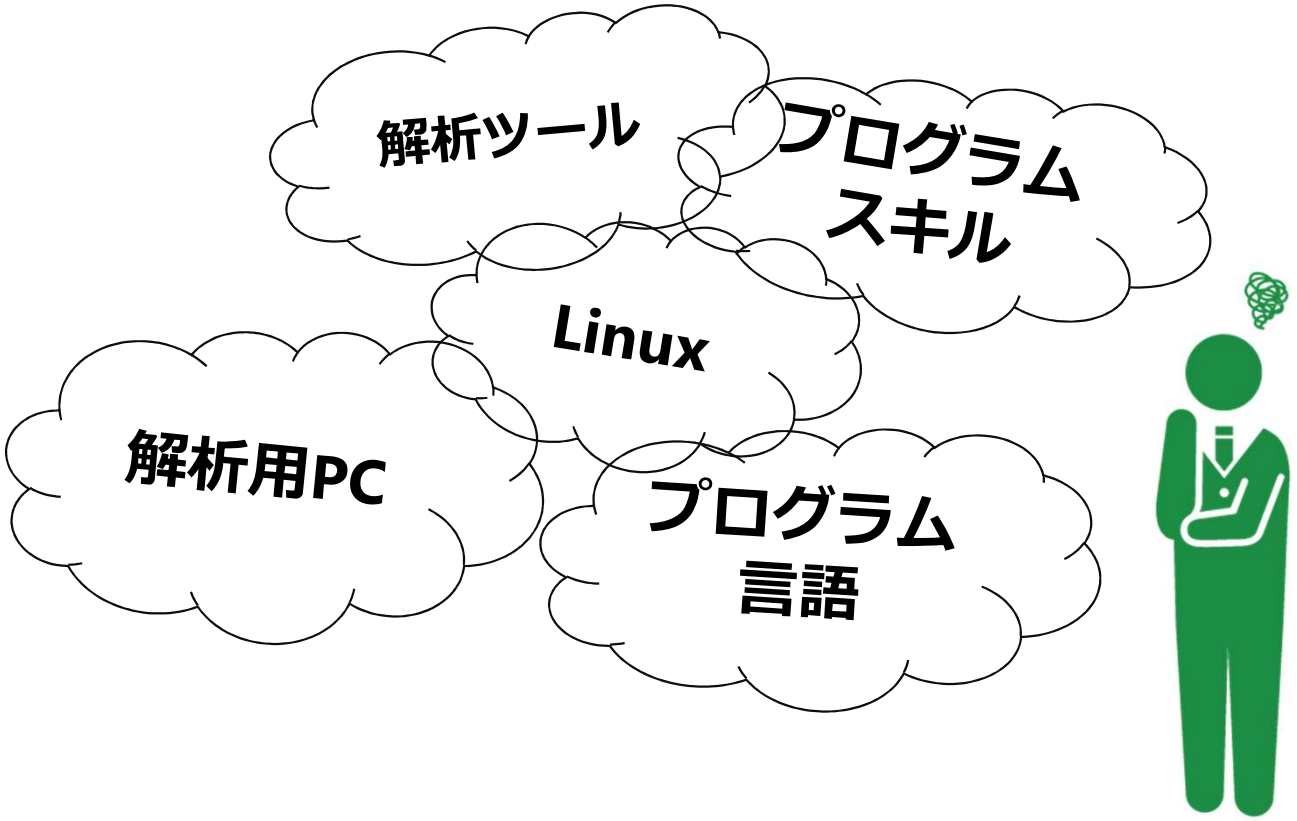
- Long Formatは、整然データ (Tidy Data)の形式
- Long Formatはデータ分析に扱いやすい形式
- WideとLongは相互変換可能 (R: tidyr package etc.)

整然データの詳細な解説

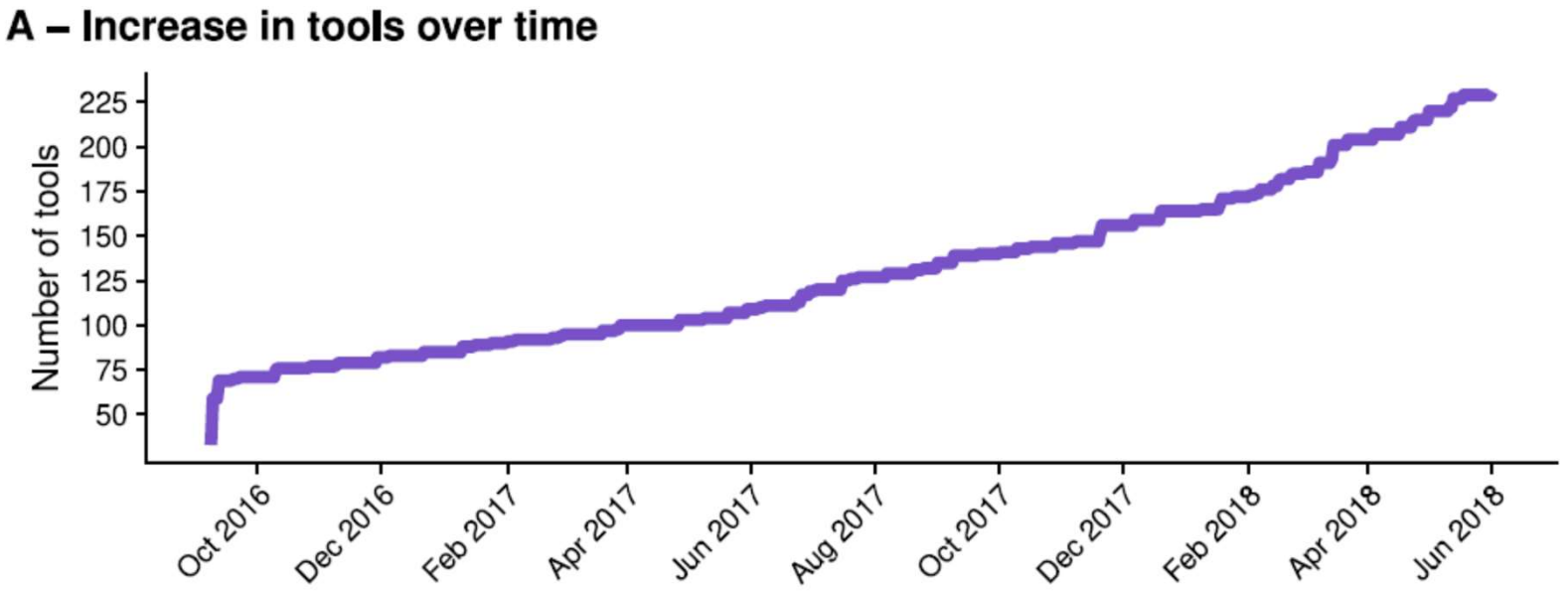
整然データとは何か — @f_nisihara
https://speakerdeck.com/fnshr/zheng-ran-detatutenani
http://id.fnshr.info/2017/01/09/tidy-data-intro/

Biological Interpretation

遺伝子発現マトリックスを使用してどのように解析したらよい？

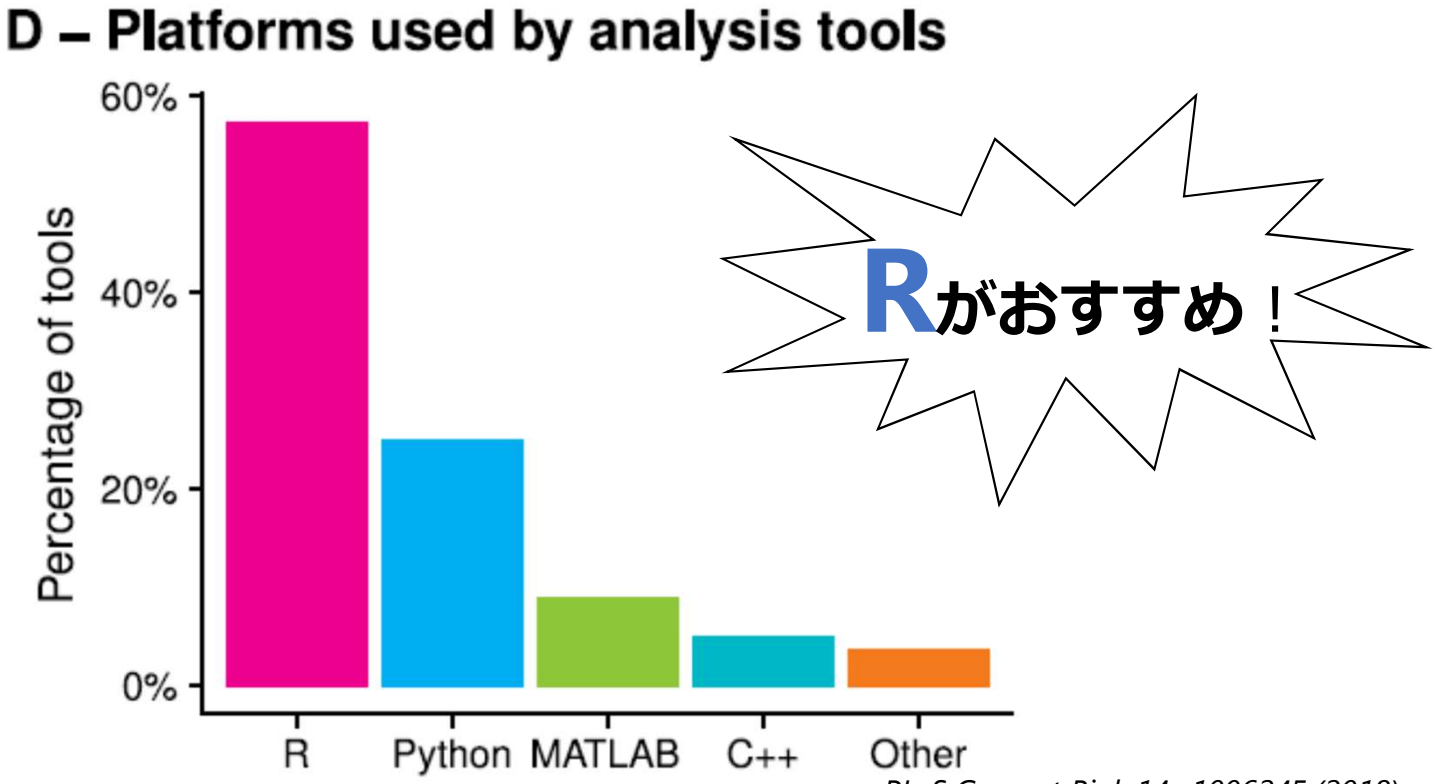


scRNA-seq用の解析ツールは毎年かなりの数が開発されている



PLoS Comput Biol. 14:e1006245 (2018).

Rに対応したツールが非常に多い！



PLoS Comput Biol. 14:e1006245 (2018).

解析環境 (R)

- OSを選ばない
  - MacでもLinuxでも、windowsでも！
- 基本的な使用方法について解説したページが多数
  - Rjp Wiki など
  - (Rで)塩基配列解析　門田先生

RjpWiki
http://www.okadajp.org/RWiki/?RjpWiki

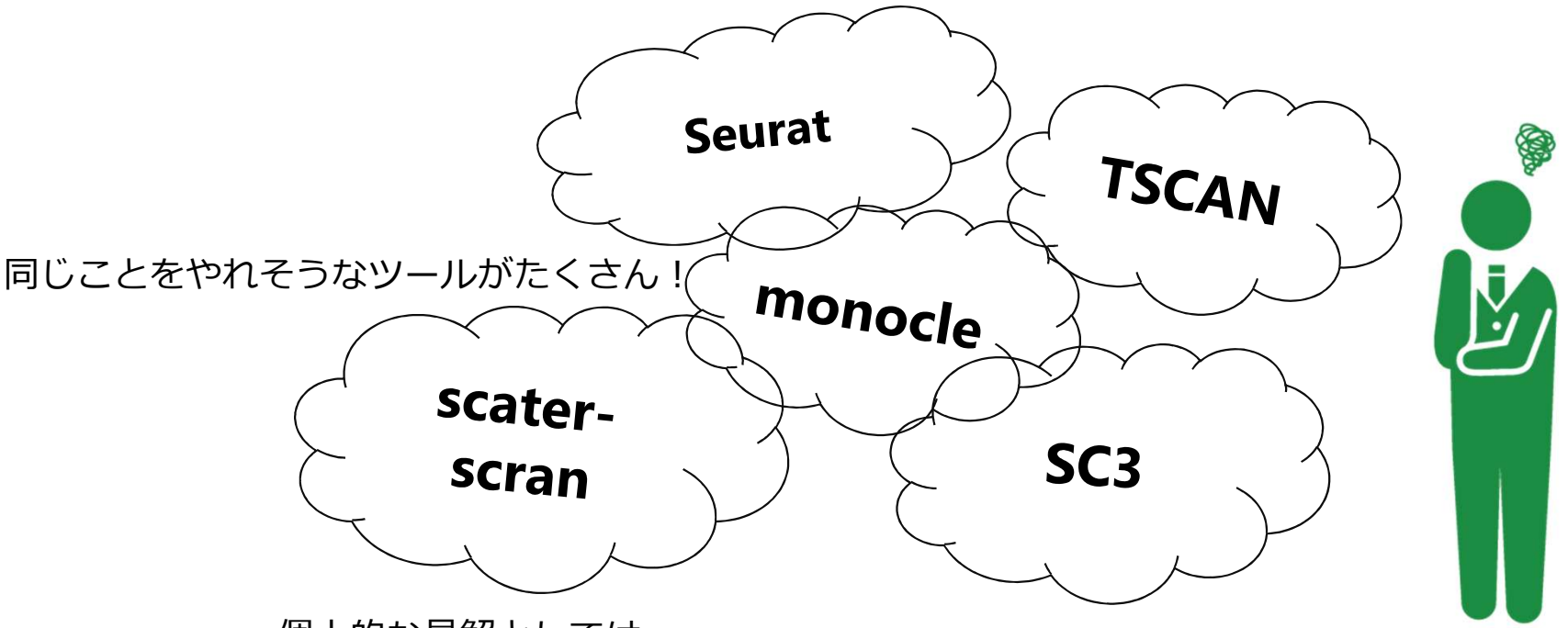
(Rで)塩基配列解析
(last modified 2018/08/30, since 2010)

- 書籍数も多い
- 様々な解析ツールがpackageという形で提供されている
- scRNA-seqを含むNGS解析に特化したpackageが多数提供されている
  - Bionconductor

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

- 対話型
  - エラーが起きても、エラー内容がわかりやすい

RのどのscRNA-seq解析用ツールを使用すればよいの？



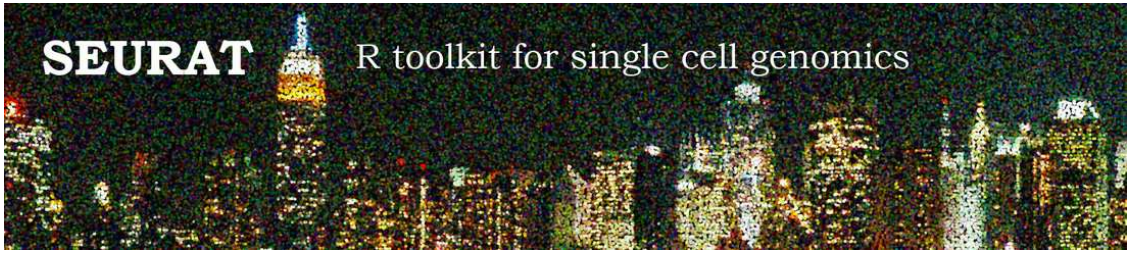
同じことをやれそうなツールがたくさん！

個人的な見解としては、

実際に使用してみて研究目的と相性が良いものを選ぶしかない

scRNA-seq解析のおすすめRパッケージ

Seurat (R package)



SATIJA Lab: http://satijalab.org/seurat/

- チュートリアル、FAQが充実
- QC、filtering、clusteringなどscRNA-seqの一連の解析をAll-in-oneで可能
- Gene Expression Matrixがあれば、解析可能
- メモリにやさしく動作が非常に軽快
- 異なる手法のデータセットを結合し、Normalizationやbatch effectの除去が可能
- DEGの手法の選択が多様
- 日進月歩のscRNA-seq解析手法を素早く導入（アップデート頻度が高い）

Table of Contents

Analysis of single cell RNA-seq data

GitHubにあるサンガー研究所 Martin Hembergチームによる
scRNA-seq解析トレーニングコース

https://hemberg-lab.github.io/scRNA.seq.course/index.html

Seuratだけが独立した項目として解説されている

Seuratは現時点 (2019年1月)において、scRNA-seq解析のスタンダード
と言えるかも？

少なくとも最初に使用するツールとしてはおすすめ

非常に参考になりますので、興味ある方は是非
（**「Hemberg scRNA」などで検索！！**）

本日の流れ

- 1細胞RNA-seq (scRNA-seq)のワークフロー
- **ddSEQのデータを用いた解析事例**
  - 解析の流れ・解析環境
    - BaseSpase → R (Seurat2)
  - R package Seuratによる解析
    - データのクオリティチェック
    - 細胞のクラスタリング
    - クラスター間で発現量に差がある遺伝子の抽出
    - クラスターのアノテーション
  - chromiumのデータとの比較
- 腫瘍検体のscRNA-seq解析事例
  - monocle, velocytoによる解析例

解析の準備-1

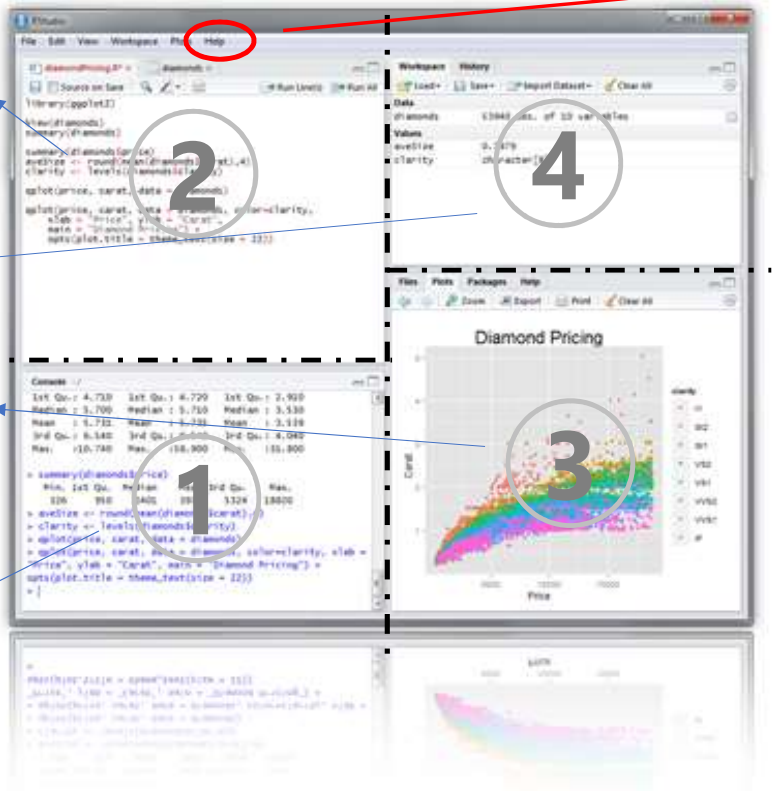
1. Rのための統合開発環境Rstudioをインストール

② Source Editor
Rのスクリプトを編集
ここからコードの実行も可能

④ Environment
作成したデータや変数の確認
実行履歴

③ File, Plots, Packages など
Fileへのアクセス
Plotデータの出力
R packageの管理、ヘルプ
など

① Console
対話的に操作するところ
コマンドの実行・結果の出力



(http://memorandum2015.sakura.ne.jp/index_rstudio.html)

Help
困ったらココ!!

生産的

RStudioはRを使って何かを生み出す時に必要なものすべてを単一のカスタマイズ可能な環境にまとめたものです。その直感的なインタフェイスと強力なコーディングツールは作業をより早く終える助けとなるでしょう。

どこでも走る

RStudioはWindows, Mac OS X, Linuxといったすべての主要なプラットホームで利用できます。RStudioはRとともにサーバ上で走らせることもできるので，複数のユーザがウェブブラウザを用いてRStudio IDEを利用することもできます。

フリー&オープン

R同様，RStudioもソフトウェアの共有と改変，すべてのユーザにとってフリーソフトウェアであることを保証するopen source licenseの下に利用できます。

Download RStudio Beta

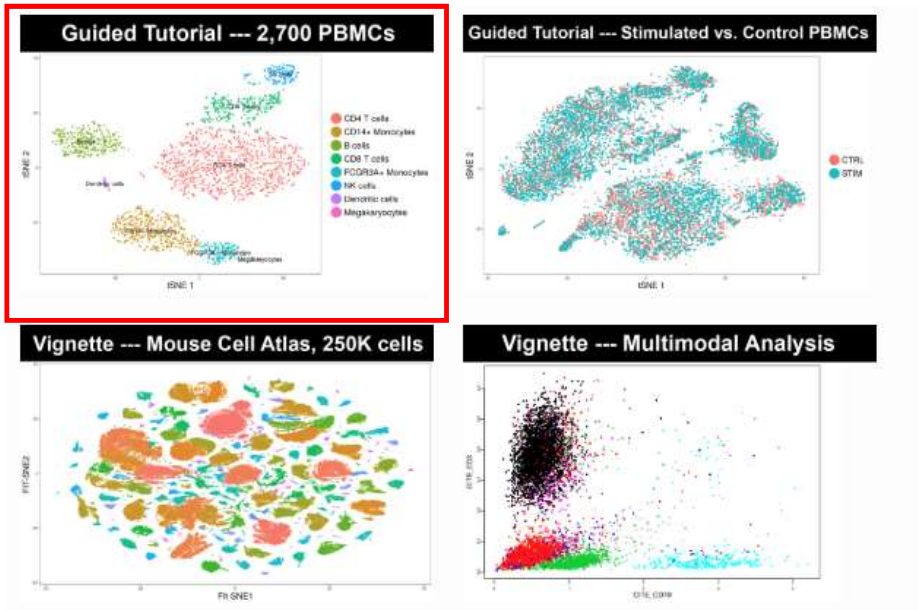
解析の準備-2

2. RstudioにSeurat Packageをインストール

更新頻度が高いので
定期的にアップデートを
確認すると良い

Experimental Design | Sequence | Processing Reads | Preparing Expression Matrix | **Biological Interpretation**

Seurat web Site



Get Started

→

Guided Tutorial –2,700 PBMCs

Seurat

1. **Seurat object class**
   遺伝子発現マトリックスをSeurat object classへ変換
2. **Expression QC**
   解析に使用する良好な細胞、遺伝子をフィルタリング
3. **Normalization**
4. **Highly variable genes**
5. **Dealing with confounders**
6. **Linear dimensionality reduction**
   次元削減（PCA→tSNE）
7. **Significant PCs**
   次元削減（PCA→tSNE）
8. **Clustering cells**
   細胞を分類
9. **Marker genes**
   クラスターに特徴的な遺伝子の抽出

解析環境

```
> sessionInfo()
R version 3.5.0 Patched (2018-05-30 r74806)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)

Matrix products: default

locale:
[1] LC_COLLATE=Japanese_Japan.932  LC_CTYPE=Japanese_Japan.932
LC_MONETARY=Japanese_Japan.932
[4] LC_NUMERIC=C                   LC_TIME=Japanese_Japan.932

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] Seurat_2.3.3          cowplot_0.9.2         Matrix_1.2-14         forcats_0.3.0         stringr_1.3.1
 [6] dplyr_0.7.8           purrr_0.2.5           readr_1.3.0           tidyr_0.8.2           tibble_1.4.2
[11] ggplot2_3.1.0         tidyverse_1.2.1.9000
```

#Windows10 Pro

プロセッサ	Intel(R) Core(TM) i5-4570 CPU @ 3.20GHz 3.20 GHz
実装 RAM	16.0 GB (15.9 GB 使用可能)

- **Method**

  ddSEQ (Droplet-base)

- **Sample**

  Mixture of 4 cultured cell lines

  3 Ovarian Clear Cell Carcinoma (OCCC) Cell Lines

  OVISE　JHOC5　ES2

  1 Ovarian Surface Epithelial Cell Line

  OSE3

- **Target number of cells**

  1,200 cells (4 well)

上記のデータを

RのSeurat packageを使用して解析してみる

Gene Expression Matrixの取り込み

BaseSpaceによって生成されたカンマ区切り (CSV形式) の遺伝子発現マトリックスを
データフレームとして読みこむ

例

```
library(Seurat)
library(dplyr)

df <-      read.table(file ="PATH/TO/YOUR/GENE/EXPRESSION/MATRIX/FILE",        #ファイルの場所を指定
                      sep = ",",                                                #csvファイルなので、カンマ区切りを指定
                                                                                #タブ区切りの場合は、"¥t"
                      header = TRUE,                                            #先頭の行が列名の時は、TRUEに
                      row.names = 1,                                            #1列目が行名であることを指定
                      stringsAsFactors = FALSE)                                 #文字列が誤変換されないように
dim(df)
df[1:5, 1:5]
```

```
> dim(df)
[1]  1212 26364

> df[1:5,1:5]
```

遺伝子

細胞	DDX11L1	WASH7P	MIR6859.3	MIR6859.2	MIR6859.4
atccggggtaggaattgg	0	0	0	0	0
atacttagatgtgaaggg	0	0	0	0	0
gagcttcggtcccttacg	0	0	0	0	0
caccacgaggccgtcggc	0	0	0	0	0
gcgcggcagactcttgaa	0	0	0	0	0

あれ！？
行（細胞）と列（遺伝子）
が逆！！

Gene Expression Matrixの取り込み

データフレーム形式からマトリックス形式への変換

例

```
mat <- as.matrix(df)

##行と列の入れ替え
mat <- t(mat)

##以下は SparseMatrixへの変換 (現在のSeuratのバージョンでは必要ない)
s.mat <- as(mat, "sparseMatrix")
s.mat[1:4,1:5]
```

```
> mat[1:5, 1:4]
          atccggggtaggaattgg atacttagatgtgaaggg gagcttcggtcccttacg caccacgaggccgtcggc
DDX11L1                   0                  0                  0                  0
WASH7P                    0                  0                  0                  0
MIR6859.3                 0                  0                  0                  0
MIR6859.2                 0                  0                  0                  0
MIR6859.4                 0                  0                  0                  0
```

行（遺伝子）-列（細胞）のマトリックスに変換された

参考: SaprseMatrix

```
5 x 4 sparse Matrix of class "dgCMatrix"
          atccggggtaggaattgg atacttagatgtgaaggg gagcttcggtcccttacg caccacgaggccgtcggc
DDX11L1                   .                  .                  .                  .
WASH7P                    .                  .                  .                  .
MIR6859.3                 .                  .                  .                  .
MIR6859.2                 .                  .                  .                  .
MIR6859.4                 .                  .                  .                  .
```

遺伝子発現マトリックスをSeurat Objectへ

例

```
ddseq <-                                          #Seurat Objectの名前（任意）
        CreateSeuratObject(                       #Seurat Objectを作成するためのコマンド
                raw.data = mat,                   #用意した遺伝子発現マトリックスを指定
                min.cells = 3,                    # 遺伝子フィルター
                min.genes = 200,                  # 細胞フィルター
                project = "ddSEQ"                 #プロジェクト名（任意）
                )
```

min.cells = 3

最低限、3つの細胞以上で発現が検出されている遺伝子のみを解析に使用

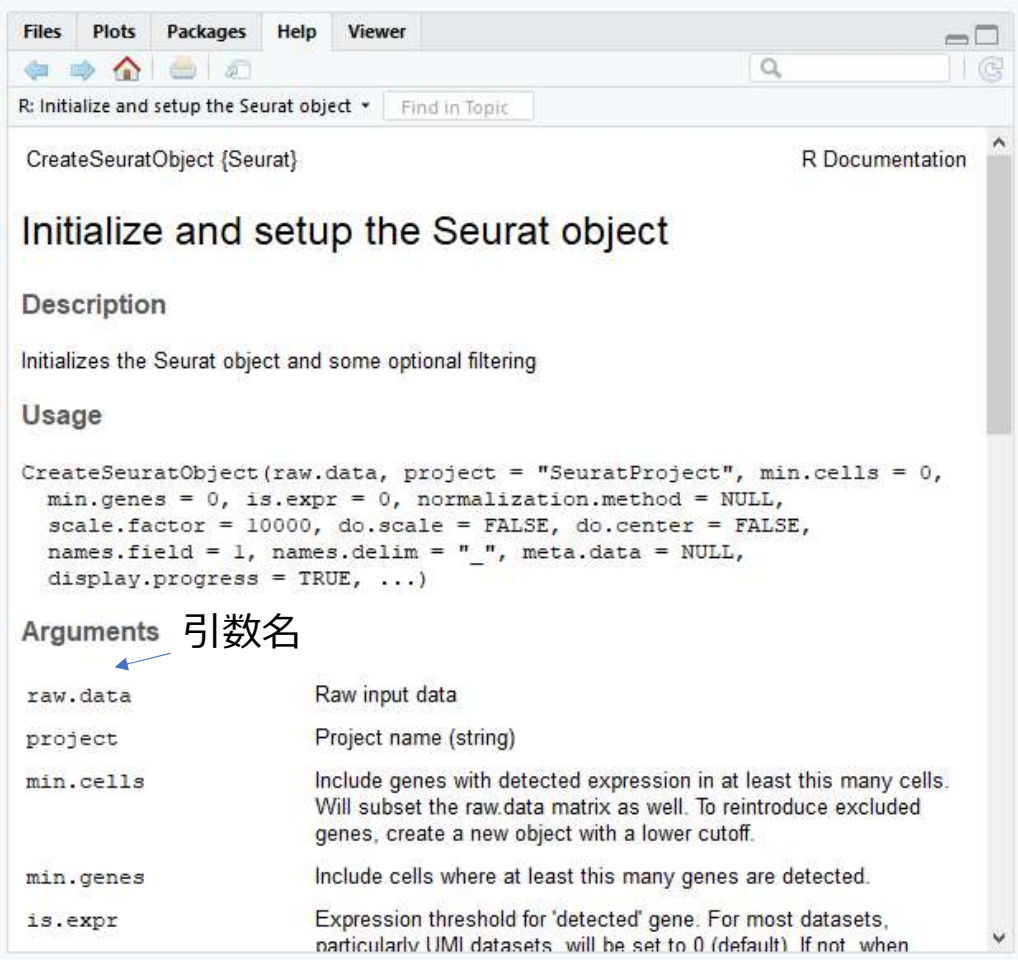
min.genes = 200

最低限、200種類の遺伝子が検出されている細胞のみを解析に使用

重要

研究目的、データの質に合わせてフィルタリングを調整する必要がある

コマンドの内容を詳しく知りたい



Files Plots Packages Help Viewer

R: Initialize and setup the Seurat object Find in Topic

CreateSeuratObject {Seurat} R Documentation

Initialize and setup the Seurat object

Description

Initializes the Seurat object and some optional filtering

Usage

```
CreateSeuratObject(raw.data, project = "SeuratProject", min.cells = 0,
  min.genes = 0, is.expr = 0, normalization.method = NULL,
  scale.factor = 10000, do.scale = FALSE, do.center = FALSE,
  names.field = 1, names.delim = "_", meta.data = NULL,
  display.progress = TRUE, ...)
```

Arguments 引数名

raw.data	Raw input data
project	Project name (string)
min.cells	Include genes with detected expression in at least this many cells. Will subset the raw.data matrix as well. To reintroduce excluded genes, create a new object with a lower cutoff.
min.genes	Include cells where at least this many genes are detected.
is.expr	Expression threshold for 'detected' gene. For most datasets, particularly UMI datasets, will be set to 0 (default). If not, when

CreateSeuratObject ()の例

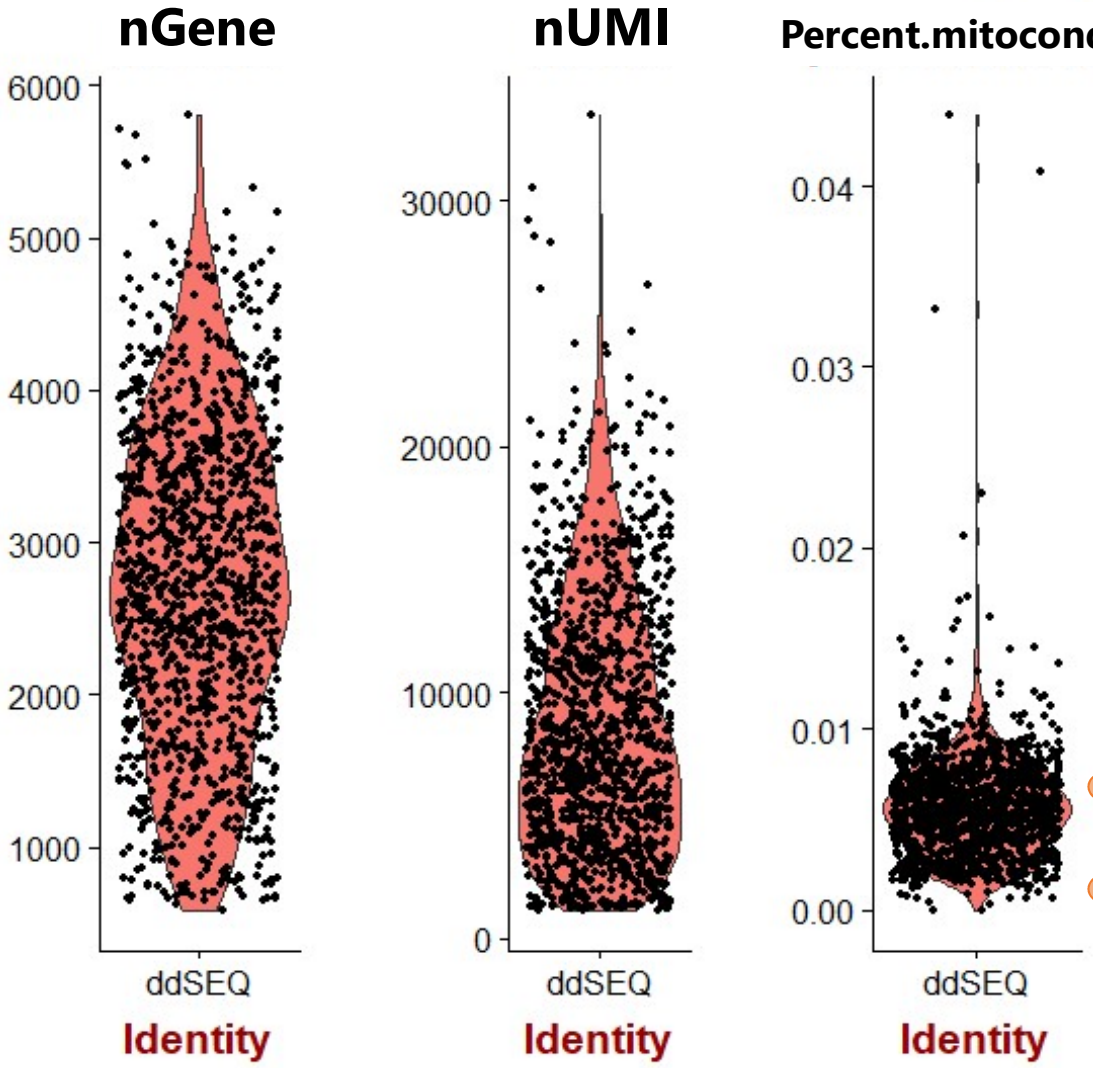
Rsutdioの右下の分割画面にあるpackagesタブからSeuratを選択
↓
Helpのタブに切り替わり、Seuratの命令文一覧が表示
↓
CreateSeuratObjectを選択 (左図)

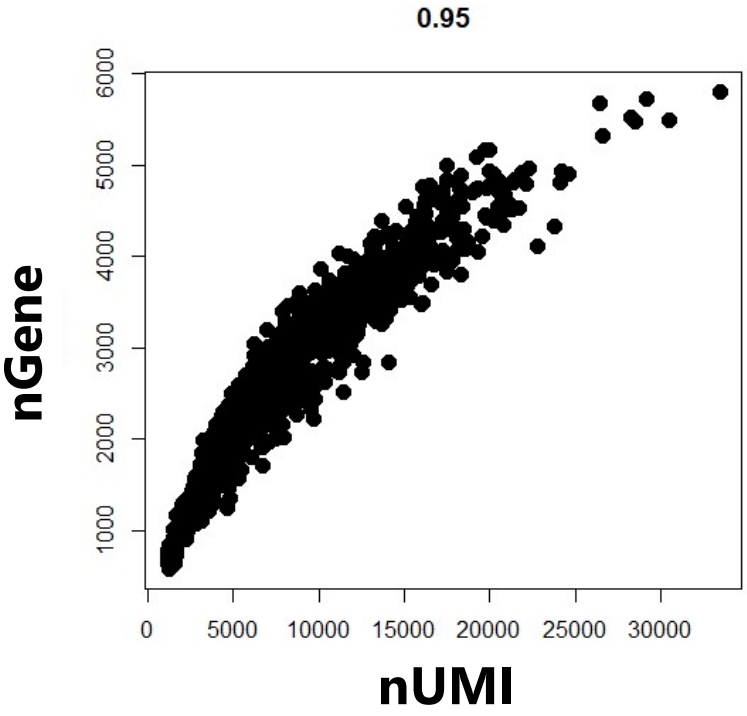
min.cellsやmin.genesが何を指定するものか記載されている

min.cells	Include genes with detected expression in at least this many cells. Will subset the raw.data matrix as well. To reintroduce excluded genes, create a new object with a lower cutoff.
min.genes	Include cells where at least this many genes are detected.

Usageにはdefautでの設定も記載されている
例えば、
min.cellsやmin.genesの値を特に指定しなかった場合、
min.cells = 0, min.genes = 0として処理される
（つまりは、フィルター無し）

QC and selecting cells for further analysis





- Good correlation between nUMI and nGene
- Low percent of Mitocondria genes

QC and selecting cells for further analysis

例

```
ddseq <-
        FilterCells(                                          #細胞をフィルターするためのコマンド
                object = ddseq,                               #解析するSeuratObject
                subset.names = c("nGene", "percent.mito"),    #フィルタリングに使用する情報
                low.thresholds = c(-Inf, -Inf),               #下限
                high.thresholds = c(5000, 0.1)                #上限
                )
```

low.thresholds = c(-Inf, -Inf)

細胞ごとの遺伝子数 (nGene)、ミトコンドリア遺伝子の割合 (percent.mito) の下限は足切りしない

注意）
例えば、免疫細胞と上皮細胞ではそもそも発現しているmRNA数が異なる（免疫細胞: 少, 上皮細胞: 大) ので、nGeneの下限を高く設定すると、免疫細胞が除外されてしまう可能性がある
解析目的に応じて、下限の足切り遺伝子数を設定する

high.thresholds = c(5000, 0.1)

細胞ごとの遺伝子数 (nGene)、ミトコンドリア遺伝子の割合 (percent.mito) がそれぞれ5,000遺伝子以上または10%以上の細胞は解析に使用しない

=2個以上の細胞のデータとうたがわれるもの、調子が悪い細胞など今後の解析のノイズとなる細胞を除外

Normalizing the data

例

```
ddseq <-
        NormalizeData(
                object = ddseq,
                normalization.method = “LogNormalize”,      #Normalizationの方法
                scale.factor = 10000)                       #RNA-seqのRPKMやTPMのM (10^6)に相当
```

scale.factor = 10000

細胞ごとのカウント数 (リード数またはnUMIなど）を10,000に揃える

Normalization後のデータはどこ？

Normalized Gene Expression Matrix

ddseq@data
ddseq@data[200:204, 100:104]　　#200~204番目の５つの遺伝子と100～104番目の5つの細胞を抽出

SparseMatrixの形式で格納されている

```
5 x 5 sparse Matrix of class "dgCMatrix"
           gcttgtttggatctaggt caagtcggtgcttaatag aagccatggcaggcttgt gccgttaaagaaacggac gaaataaattggtgagac
MTOR.AS1                   .                  .                  .                  .                  .
ANGPTL7                    .                  .                  .                  .                  .
UBIAD1                     .          0.4851475                  .          0.4862542                  .
PTCHD2                     .                  .                  .                  .                  .
LOC101929181               .                  .                  .                  .                  .
```

Raw Gene Expression Matrix

ddseq@raw.data

BaseSpaceで生成された遺伝子発現マトリックスと同じ（ただし、形式はSparse Matrix)

```
5 x 5 sparse Matrix of class "dgCMatrix"
           gcttgtttggatctaggt caagtcggtgcttaatag aagccatggcaggcttgt gccgttaaagaaacggac gaaataaattggtgagac
MTOR.AS1                   .                  .                  .                  .                  .
ANGPTL7                    .                  .                  .                  .                  .
UBIAD1                     .                  1                  .                  1                  .
PTCHD2                     .                  .                  .                  .                  .
LOC101929181               .                  .                  .                  .                  .
```

Z-score

ddseq@scale.data

Heatmapを作成するときなどに使用する
通常のMatrix形式

SparseMatrixから通常のMatrixへの変換は、as.matrix()で可能
例: mat <- as.matrix(ddseq@raw.data) #先ほどBaseSpaceのcsvから取り込んだmatと同じ

Perform linear dimensional reduction

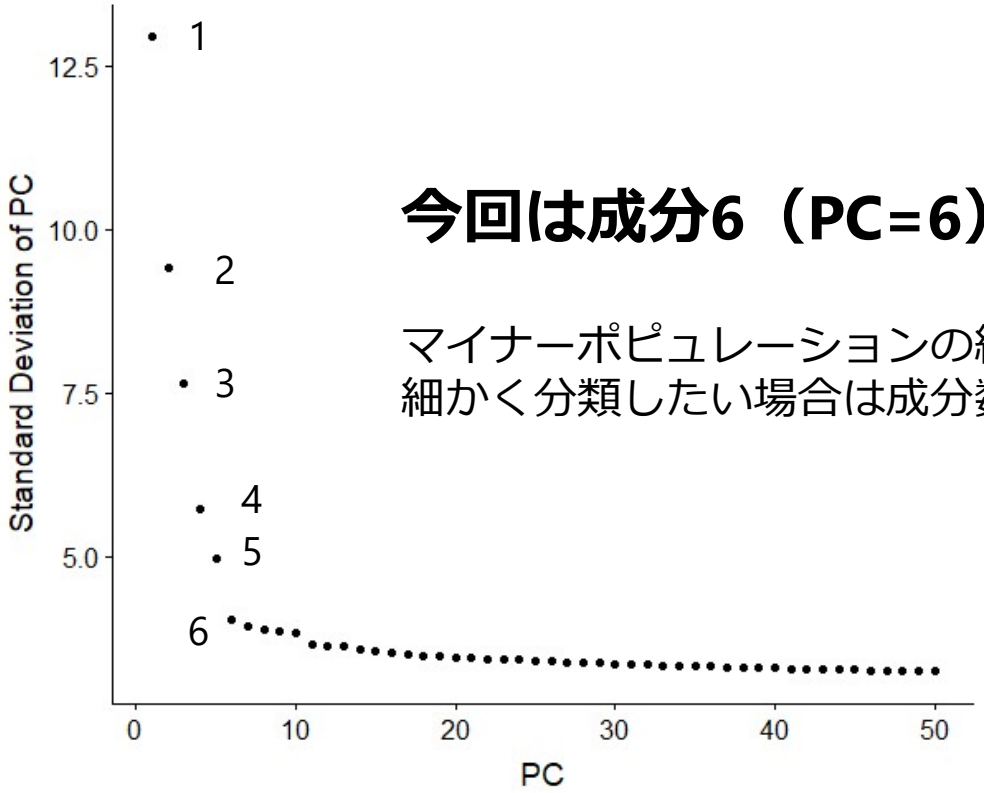
解析に使用する次元削減手法とその数を決定する

例

```
ddseq <- RunPCA(                                    #主成分解析
                object = ddeseq,
                pc.genes = ddseq@var.genes,         #予め算出した細胞間で分散が大きい
                                                    （発現変動が大きい）遺伝子リスト
                do.print = TRUE,                    #結果を出力するか否か
                pcs.print = 1:10,                   #結果を出力するか主成分の数
                genes.print = 10,                   #出力する遺伝子の数
                pcs.compute = 50                    #計算する成分の数
                )

#今後の解析に使用する成分の検討
PCElbowPlot(object = ddseq, num.pc = 50)
#または
ddseq <- JackStraw(object = ddseq, num.replicate = 100, display.progress = FALSE)  #少し時間がかかる
JackStrawPlot(object = ddseq, PCs = 1:12)
```

Perform linear dimensional reduction

PCElbowPlot(object = ddseq, num.pc = 50)



今回は成分6（PC=6）までを使用

マイナーポピュレーションの細胞の同定など
細かく分類したい場合は成分数を増やす

Clustering

例

```
ddseq <-
        FindClusters(                          #細胞を分類するためのコマンド
                object = ddseq,
                reduction.type = “pca”,        #次元削減の手法名
                dims.use = 1:6,                #解析に使用する次元数
                resolution = 0.2,              #クラスタリングの解像度
                                               #値が大きいほど細かくクラスタリング
                                               #複数の解像度を指定可能 (ex. resolution = c(0.2, 0.8, 1.2))
                print.output = 0,
                save.SNN = TRUE
                )
```

パラメーターが多数あるので、最適なクラスタリング結果になるように調整する

主には、解析に使用するPC（成分）の数 (dims.use)と解像度 (resolution)

Seuratのクラスタリング手法の詳細

A smart local moving algorithm for large-scale modularity-based community detection
Waltman and van Eck *The European Physical Journal B* 86:471 (2013)

クラスタリングの結果はどこ？-1

```
ddseq@meta.data
head(ddseq@meta.data)
```

FindClusters()で複数の解像度を指定した場合、
解像度ごとの結果が格納されている

行名=rownames

細胞のバーコード

	nGene	nUMI	orig.ident	percent.mito	res.0.2	res.0.8	res.1.2
aacgtgggattggttaac	4902	24652	ddSEQ	0.006612040	0	0	8
catagagagcttgctgag	4937	24177	ddSEQ	0.003143483	1	2	2
aaagaagaggcccttacg	4807	24090	ddSEQ	0.005853051	0	0	8
aattggccacgctcagtg	4333	23769	ddSEQ	0.001977365	1	4	4
acaaggattagtcaaccg	4118	22768	ddSEQ	0.003074491	1	4	4
aagccacgaaagggtgct	4962	22271	ddSEQ	0.009923219	2	5	6
・							
・							
・							

@meta.dataには、細胞の情報がまとまっている

クラスタリングの結果はどこ？-2

例

```
ddseq@ident
head(ddseq@ident)
```

FindClusters()の結果は、ddseq@identにも格納される
ddseq@identが、tSNE-plotや遺伝子発現量解析おけるクラスター情報として使用される
複数の解像度を指定して解析した場合は、一番最後の解析結果が格納される（ようだ？）

10 cluster (res.1.2)

```
aacgtgggattggttaac catagagagcttgctgag aaagaagaggcccttacg aattggccacgctcagtg acaaggattagtcaaccg aagccacgaaagggtgct
                8                   2                   8                   4                   4                   6
Levels: 0 1 2 3 4 5 6 7 8 9
```

@identは、名前付きのファクター型

解像度0.2の結果で今後の解析を行いたい

例

```
res02<- as.factor(ddseq@meta.data$res.0.2)      #@meta.dataのres.0.2列をファクター型として抽出
names(res02) <- rownames(ddseq@meta.data)       #name属性として対応する細胞（バーコード）を付加
ddseq@ident <- res02                            #ddseq@identに解像度0.2の結果を代入
head(ddseq@ident)
```

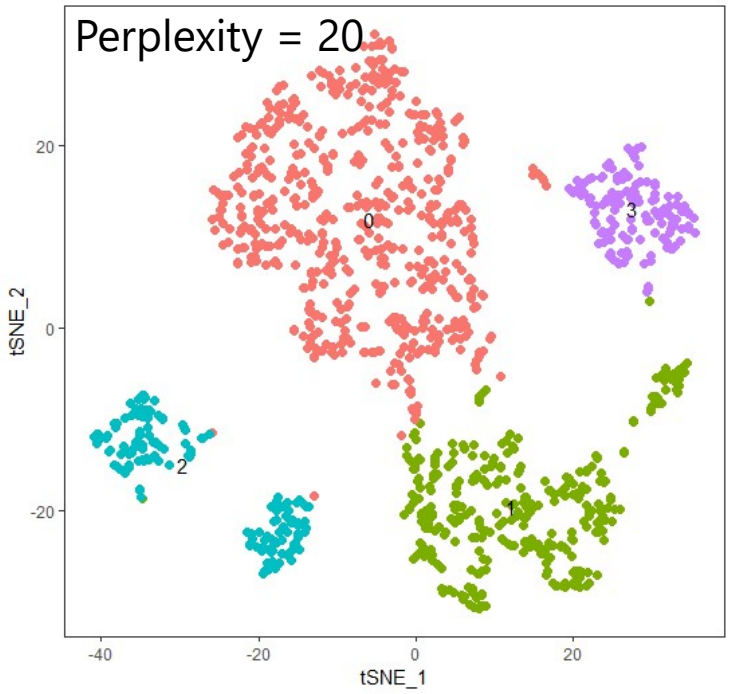
4 cluster (res.0.2)

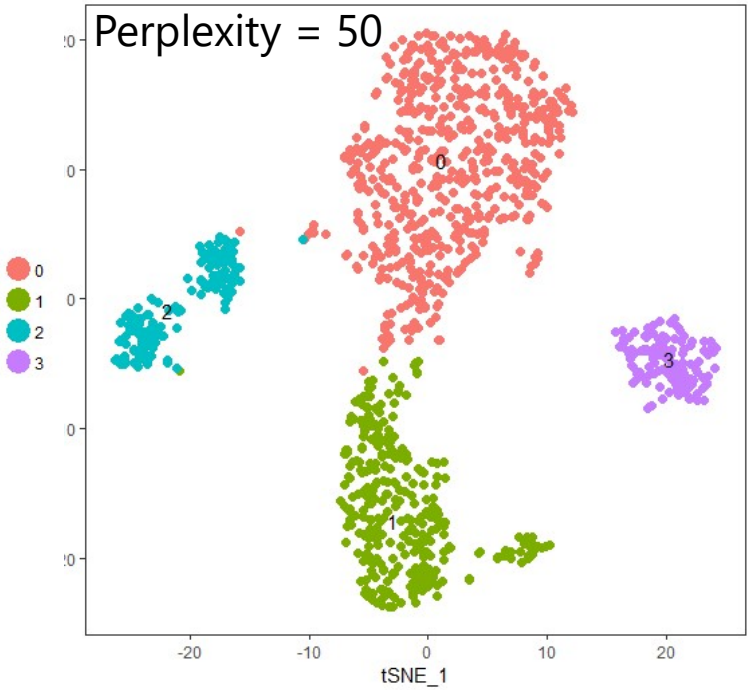
```
aacgtgggattggttaac catagagagcttgctgag aaagaagaggcccttacg aattggccacgctcagtg acaaggattagtcaaccg aagccacgaaagggtgct
                0                   1                   0                   1                   1                   2
Levels: 0 1 2 3
```

tSNE plot

例

```
ddseq <-
        RunTSNE(
        object = ddseq,
        dims.use = 1:6,                          #クラスターを探した時と同じ次元数が好ましい
        do.fast = TRUE,
        perplexity = 50)                         #tSNEのパラメーター複雑度。大分形が変わる。
TSNEPlot(object = ddseq, pt.size = 2, do.label = TRUE)#tSNE-plotの出力
```





おおよそ、
tSNEplot上でおおよそ
４つの島、
4つのクラスターに分類

Mixture of 4 cultured cell lines

3 Ovarian Clear Cell Carcinoma (OCCC) Cell Lines

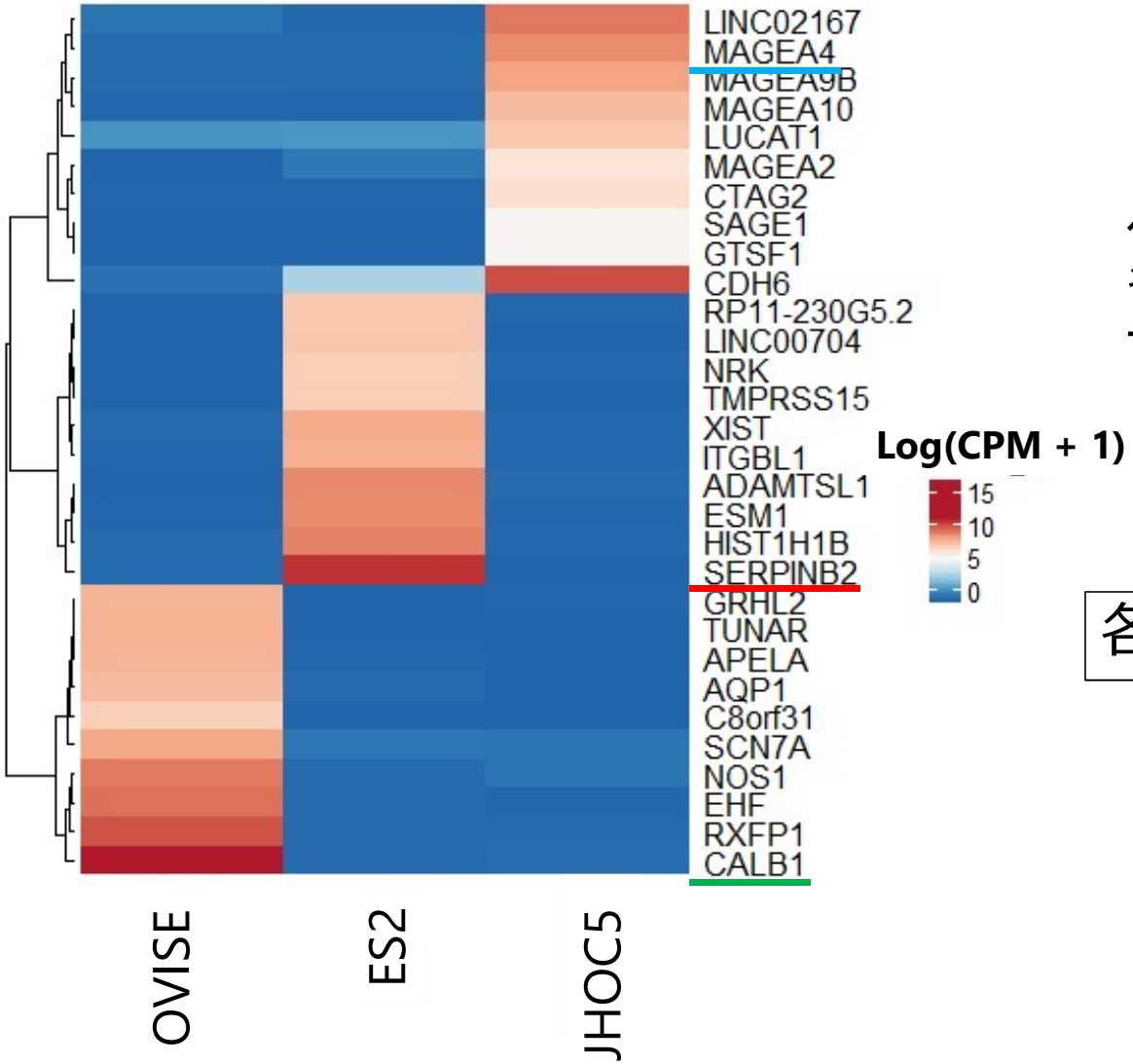
OVISE　JHOC5　ES2

1 Ovarian Surface Epithelial Cell Line

OSE3

どのクラスターがどの細胞に対応しているのか？

Extracting the DEG from bulkRNA-seq data



DEG: Differentially Expressed Genes

バルクRNA-seqのデータから
各細胞で特異的に発現が高い
Top10遺伝子を抽出

↓

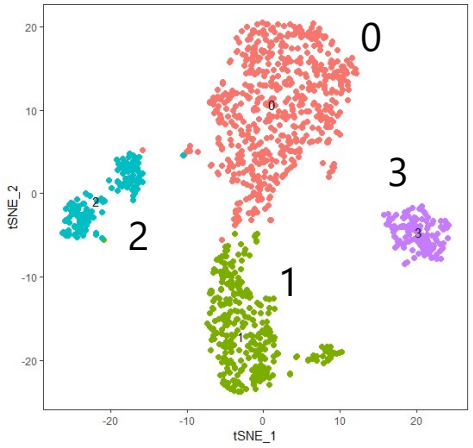
各細胞のマーカー遺伝子を決定

OVISE → **CALB1**

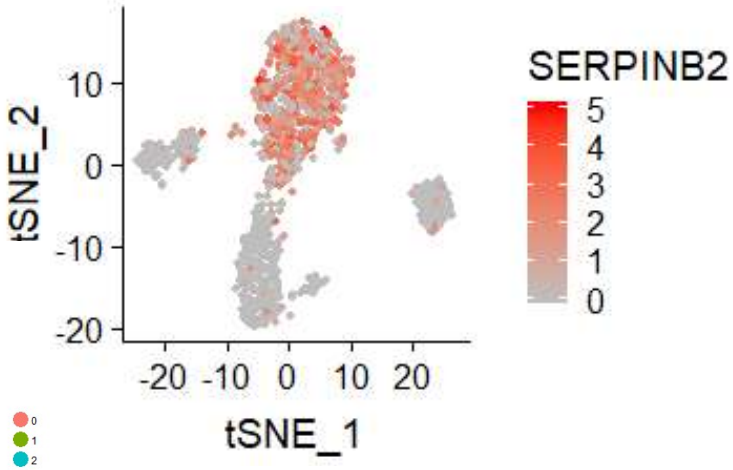
ES2 → **SERPINB2**

JHOC5 → **MEGEA4**

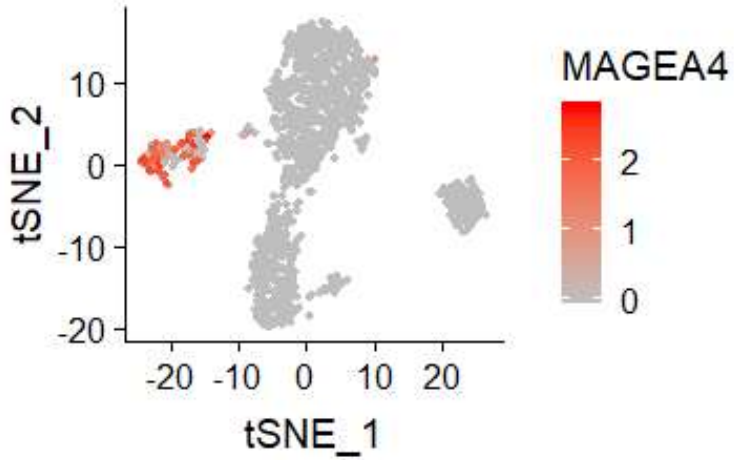
Feature plot (marker gene)



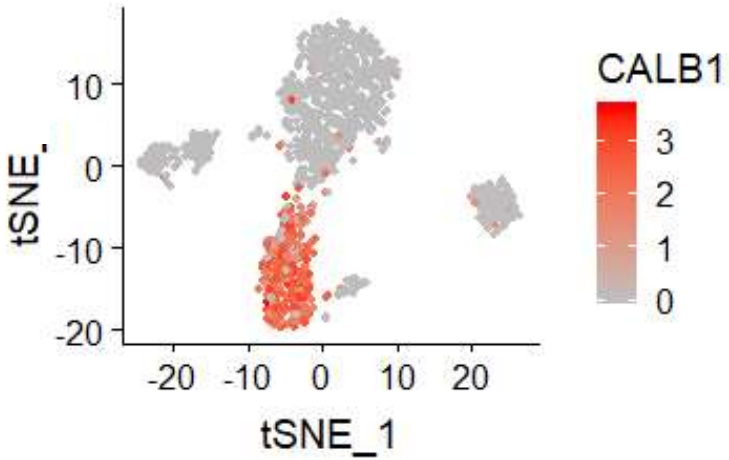
SERPINB2



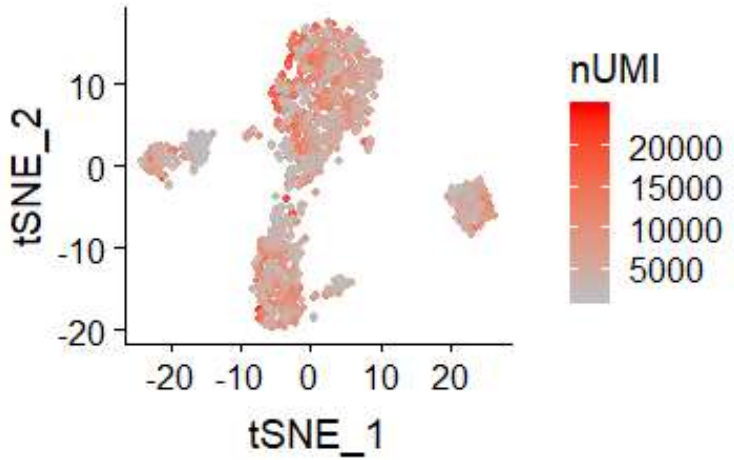
MEGEA4



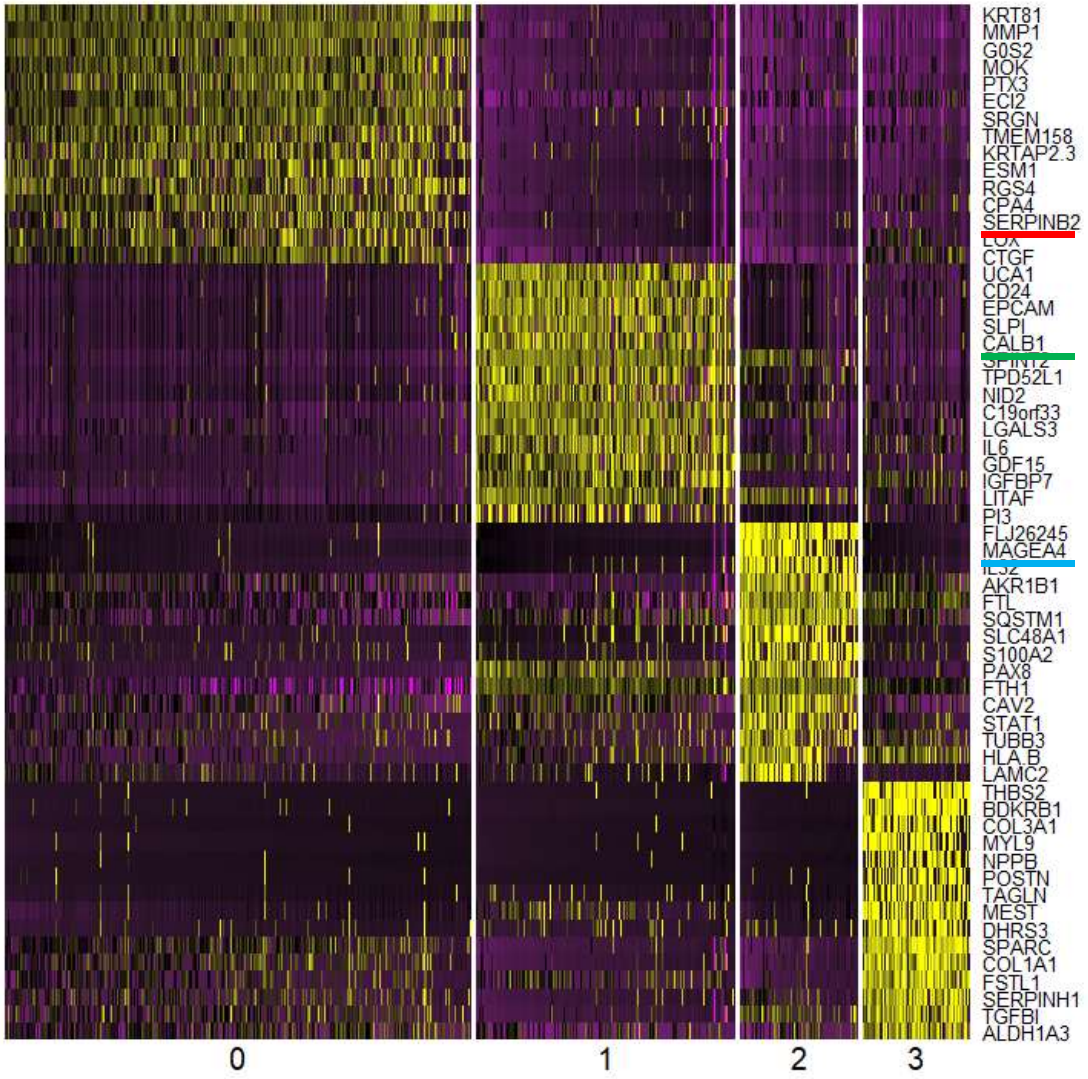
CALB1



nUMI



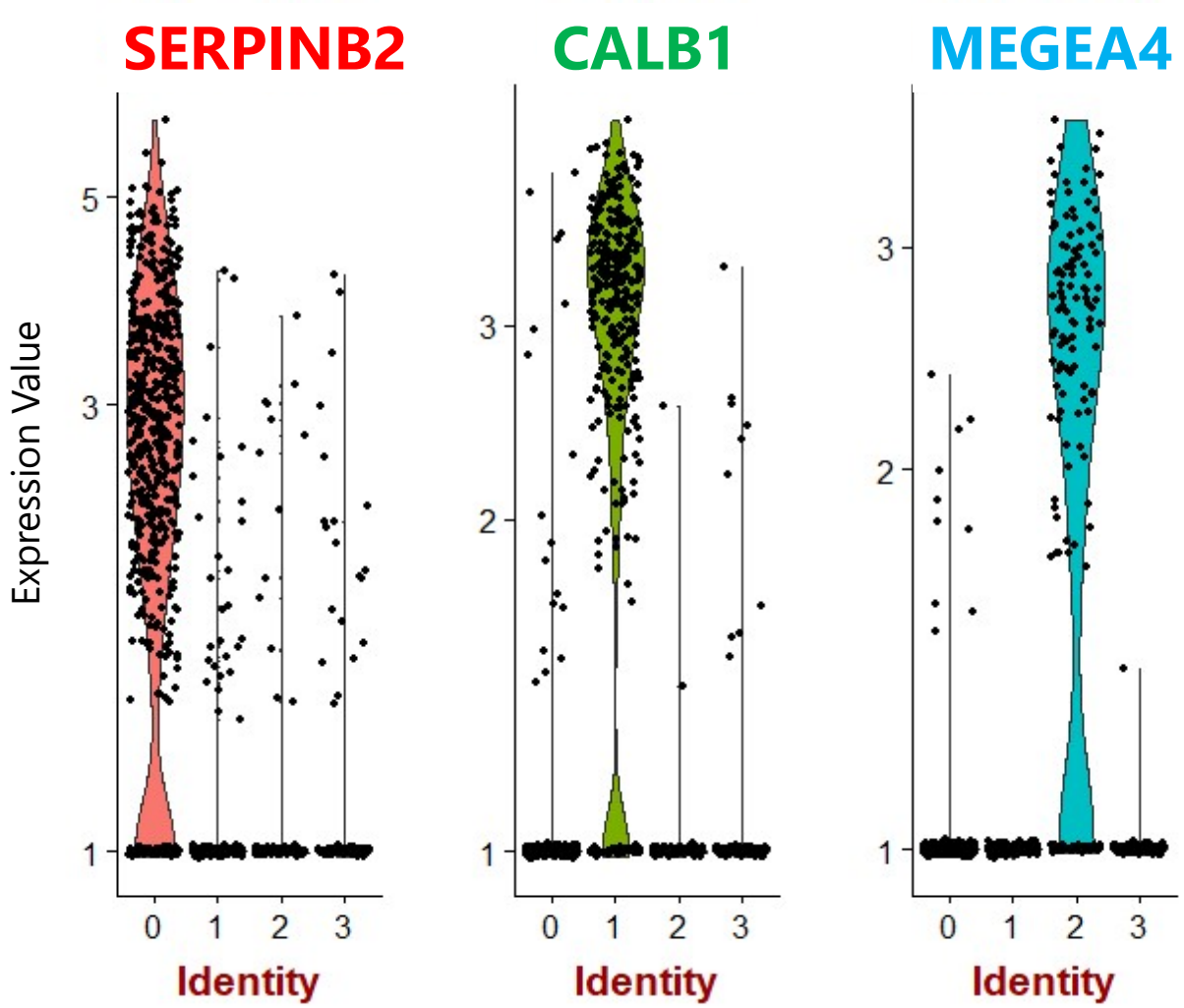
DEG in each cluster at single cell resolution



例

```
all.markers <-
            FindAllMarkers(
            object = ddseq,
             min.pct = 0.1,
            thresh.use = 0.1)

top15 <- all.markers %>%
            group_by(cluster) %>%
             top_n(15, avg_logFC)
#%>%: パイプライン演算子
#dplyr package参照

DoHeatmap(
            object = ddseq,
            genes.use = top15$gene,
            use.scaled = TRUE,
            slim.col.label = TRUE,
            remove.key = TRUE)
#use.scaled = TRUEの場合、ddseq@scale.dataの値が使用される
```

Violin plot (marker gene)



Cluster-0 **(C0)**
ES2?

Cluster-1 **(C1)**
OVISE?

Cluster-2 **(C2)**
JHOC5?

Cluster-3 **(C3)**
OSE3?

→ **別の方法でも確認**

Gene correlation between bulkRNA-seq and scRNA-seq

50%以上の細胞で検出されている遺伝子について、各クラスターにおける平均の発現量を算出

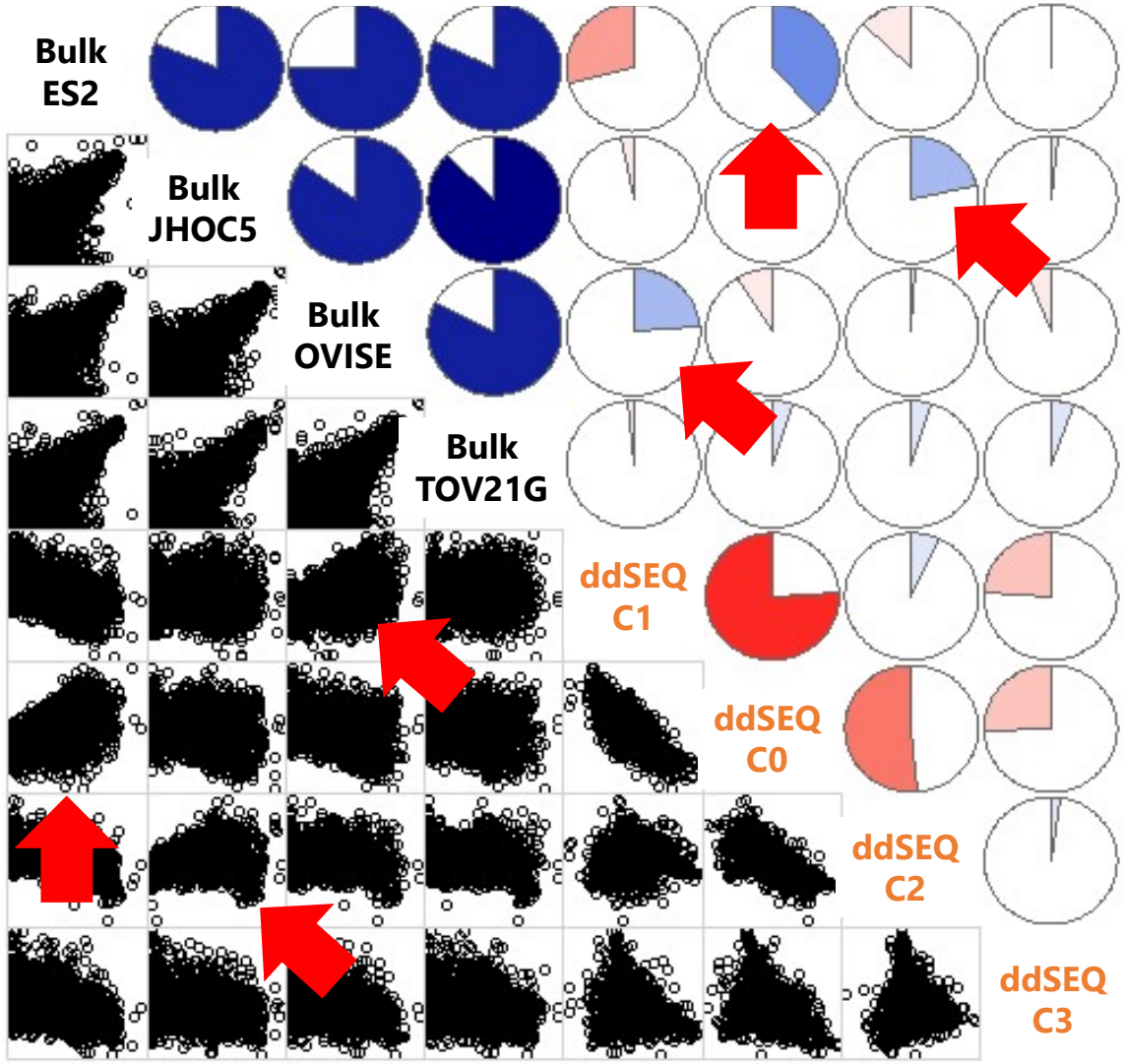
ddSEQ data

遺伝子の発現パターンを比較

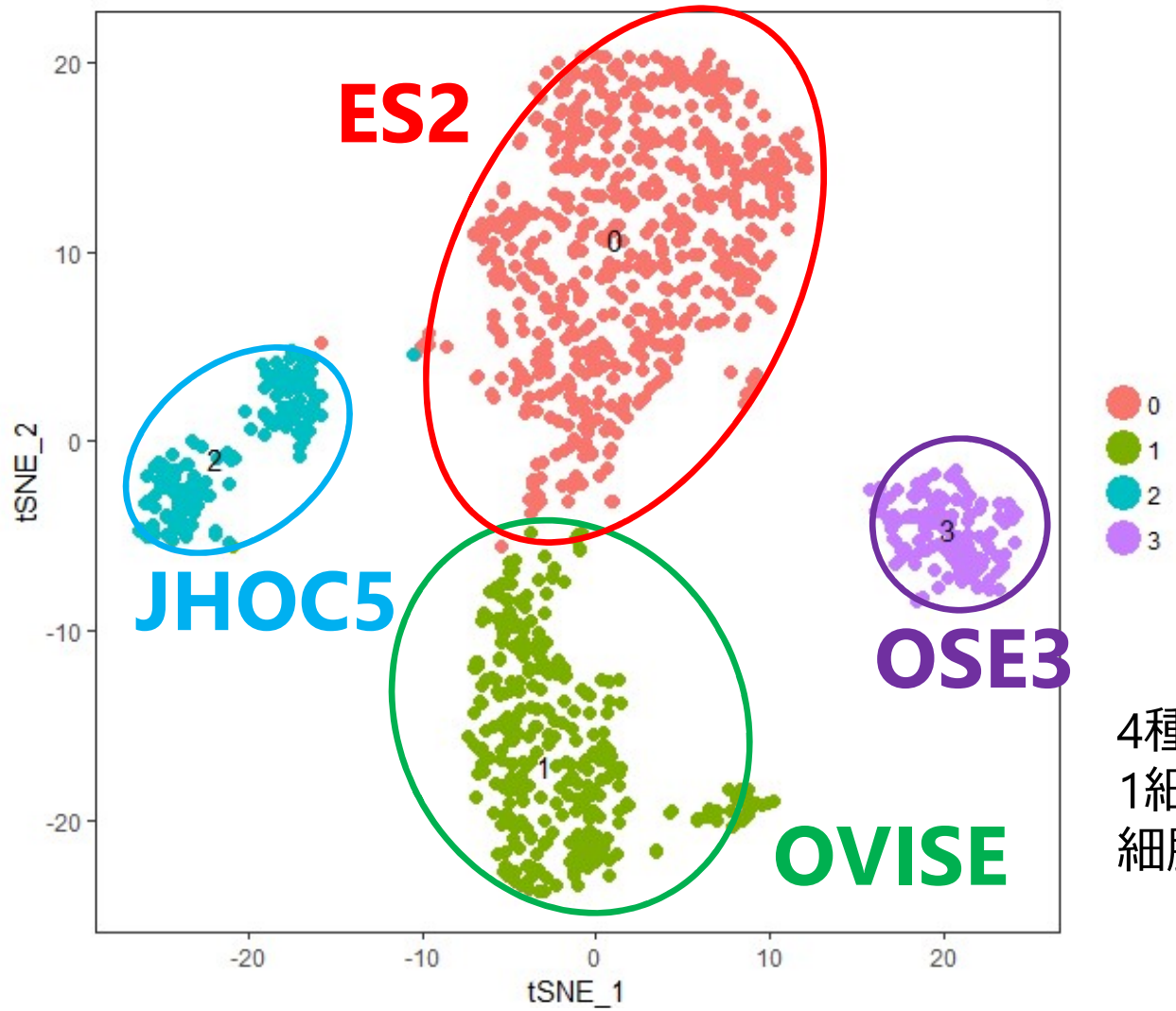
BulkRNA-seq data

上記遺伝子の発現プロファイルのみを使用
（edgeR: TMM → CPM)

青: 正の相関係数
赤: 負の相関係数



Final annotation of each cluster



Cluster-0 **(C0)**
ES2

Cluster-1 **(C1)**
OVISE

Cluster-2 **(C2)**
JHOC5

Cluster-3 **(C3)**
OSE3

4種類の培養細胞を混合した
1細胞RNA-seqのデータから、
細胞の種類ごとに分類できた

保存

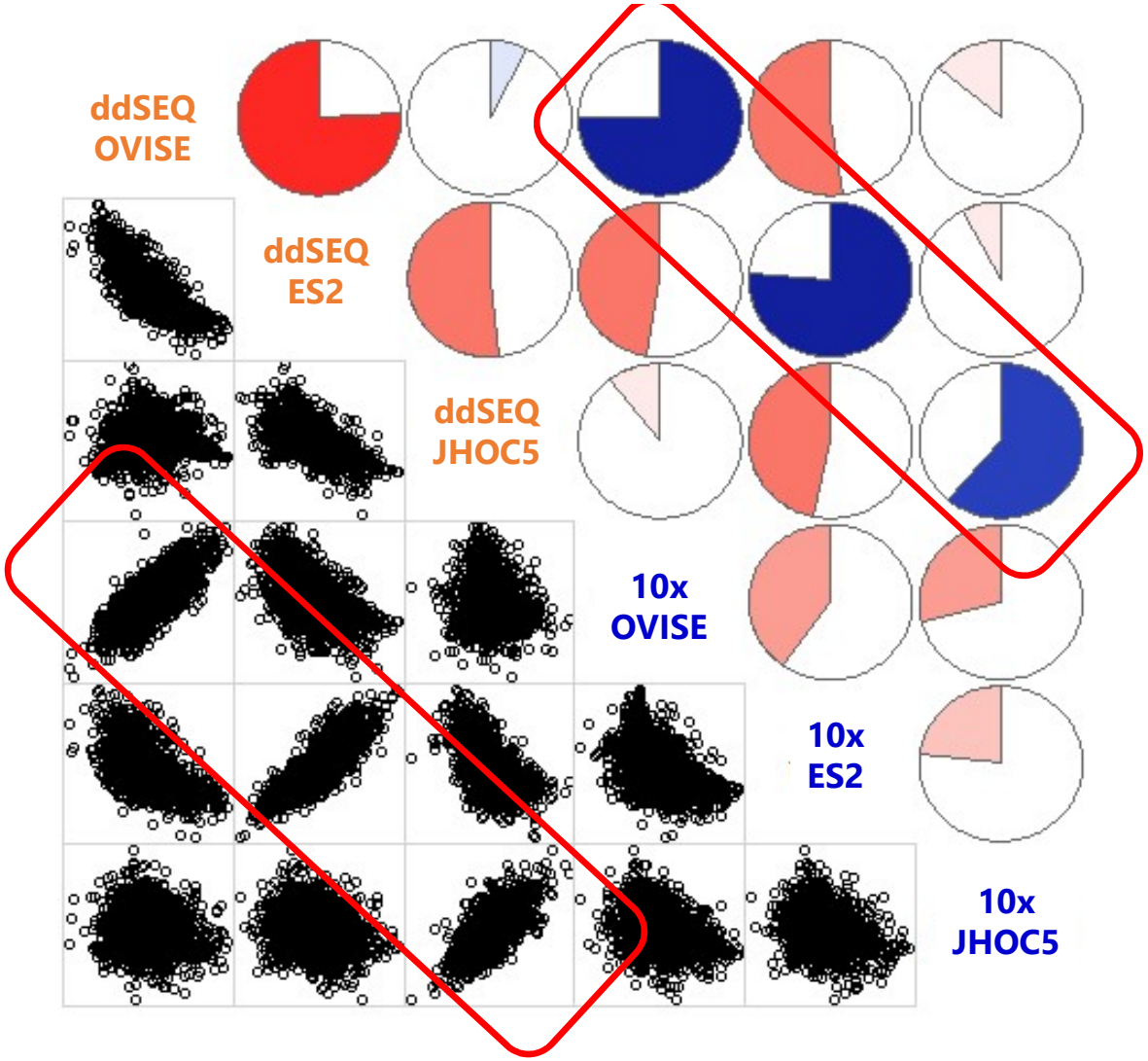
ここまで解析したddseq (Seurat Oblect) を.rds形式で保存

```
saveRDS(ddseq, “/PATH/TO/YOUR/SAVE/DIRECTORY/****.rds" )
```

読み込むときは、

```
ddseq <- loadRDS(“/PATH/TO/YOUR/SAVE/DIRECTORY/****.rds" )
```

Gene correlation between ddSEQ and chromium

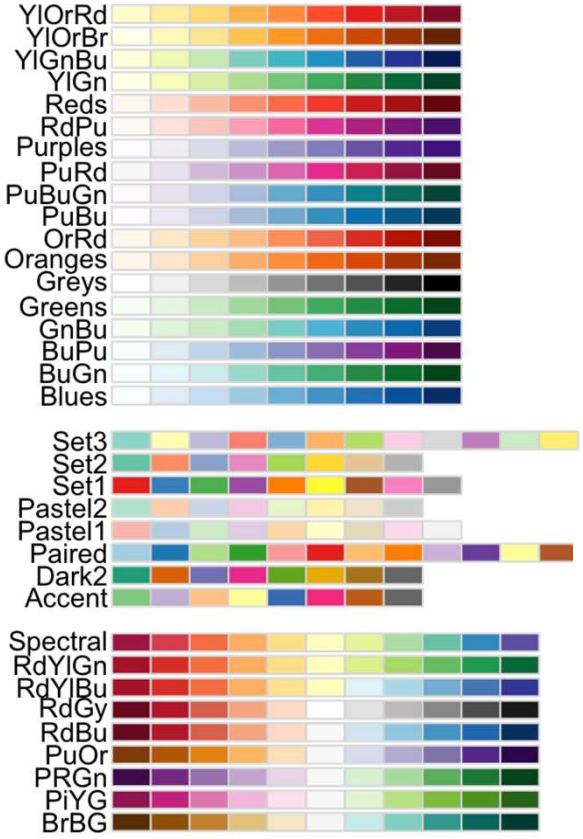


同じDroplet-baseの
プラットフォーム機器である
ddSEQ (Bio-Rad)
と
Chromium (10x genomics)
のデータの比較

ddSEQのデータと
Chromiumのデータには
高い相関性がある

色を変えたい！-1

よく使用される色のセット　例



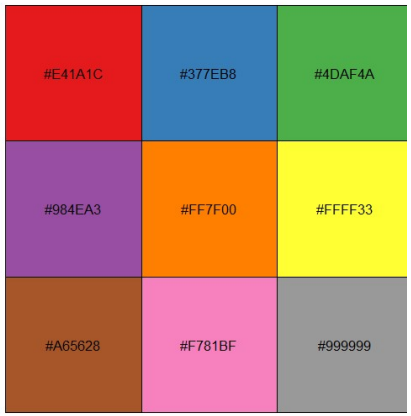
RColorBrewer
display.brewer.all()

```
library(RColorBrewer)

col <- colorRampPalette(brewer.pal(9, "Set1"))#カラーパレットを生成するコマンド
res02.col <- col(9)[1:4]                      #全9色のSet1から最初の4色を選択
scales::show_col(res02.col)                   #色見本の表示
cluster.02 <- levels(ddseq@ident)             #クラスター名を取得
names(res02.col) <- cluster.02                #色に対してクラスター名を与える
res02.col                                     #name属性付文字列ベクター
```

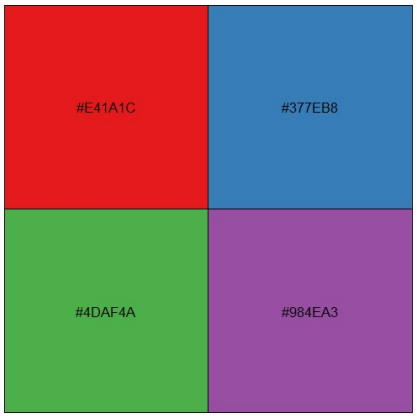
```
> res02.col
        0         1         2         3
"#E41A1C" "#377EB8" "#4DAF4A" "#984EA3"
```

クラスター名

Name Vector



Set1



res02.col

16進法で表記された
RGBカラーコード

クラスター名と
カラーの対応をつければ
どんな色にも変更可能

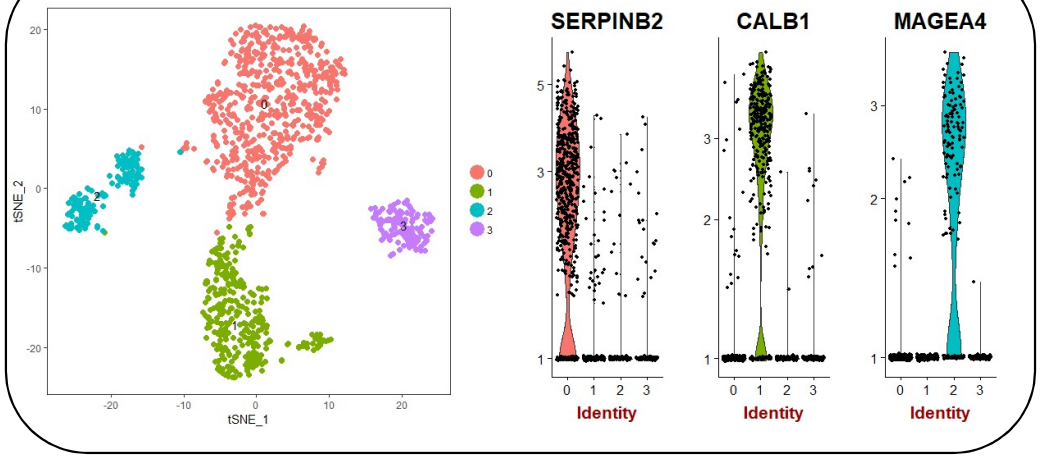
色を変えたい！-2

例

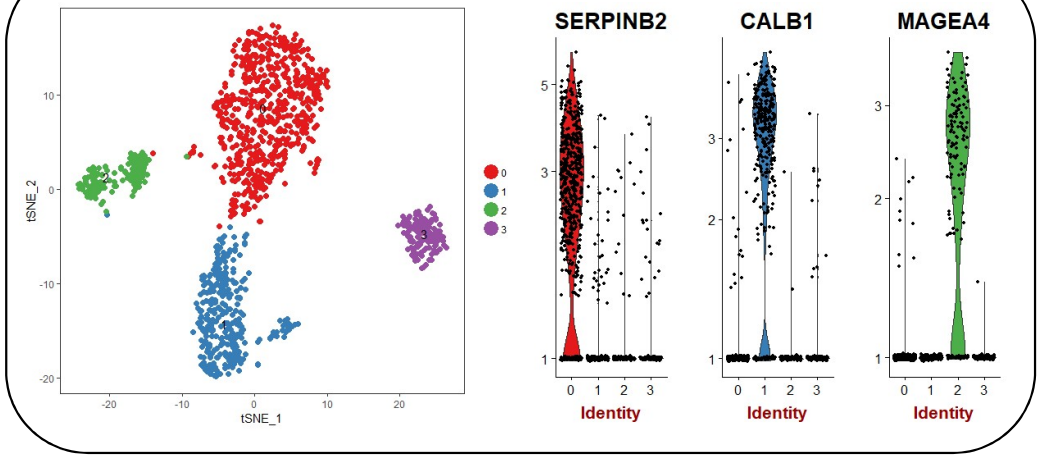
```
TSNEPlot(object = ddseq, pt.size = 2, do.label = TRUE, colors.use = res02.col)

VlnPlot(object = ddseq, features.plot = c("SERPINB2", "CALB1", "MAGEA4"),
        y.log = TRUE, cols.use = res02.col)
```

同じ内容でも引数の名前が異なることがあるので注意
TSNEPlot　→ color.use　VinPlot → cols.use

変更前



変更後



デフォルトの色はR package “ggplot2”のデフォルト色

まとめ

多数存在する解析ツールから研究目的に適したツールを選択する

実際に使用してみる、論文での使用頻度　など

scRNA-seqの解析ツールの多くはR

プログラムスキルがなくともチュートリアルを参照してどうにかなるので、チュートリアルが充実しているツールがおすすめ

今回使用したSeurat　など
チュートリアルに従って操作しているうちに、R言語への理解も深まる

コマンドでわからないことがあったらHelpを見る

設定可能な引数の情報、コマンドの処理内容などがわかる
それでもわからない場合は、身近のわかる人に聞く（そのような人が近くにいることが重要！）

まだ情報解析したことがない方も是非トライしてみてください

Experimental Design → Sequence → Processing Reads → Preparing Expression Matrix → Biological Interpretation

ご静聴ありがとうございました！

プログラムスキルの有無にかかわらず、
是非、情報解析にトライしてみてください

楽しく実りあるシングルセルRNA-seq解析ライフを送りましょう