

AJACS「シングルセルRNA-seqを知って・学んで・使う」

JST東京本部 2階 共創スペース

2024年12月23日(月) 13:40～14:20 (40分)

scRNA-seqデータを用いた細胞分類入門

飯田 溪太

大阪大学蛋白質研究所

講義内容の概要

今日、次世代シーケンサーを用いたゲノム解読技術が飛躍的に向上し、生命分子の網羅的情報(オミクス)が一細胞レベルで取得可能になっている。オミクスデータをどのように分析し、生命を包括的に理解するかという問いは、現代医学・生物学における共通の課題の一つである。

今回は、シングルセルRNAシーケンス(scRNA-seq)データの情報解析をテーマに、「細胞分類」と呼ばれる基本的な問題を考える。ふつう、細胞分類といえば、細胞を適当にクラスタリングしたのち、各クラスターを既知の細胞種に当てはめる(=分類する)ことを指す。しかし、がんのように分類自体が困難なデータの場合はどうすればよいだろうか？

想定する対象者

- ・シングルセル解析に興味のある方
- ・scRNA-seqデータをもとに遺伝子発現解析を行いたい人
- ・標準的な解析手法の限界を乗りこえたい人

アジェンダ

- RNA-seq とは？
- 標準的な解析手法
- ソフトウェアASURAT(阿修羅)の開発
- まとめと今後の展望

アジェンダ

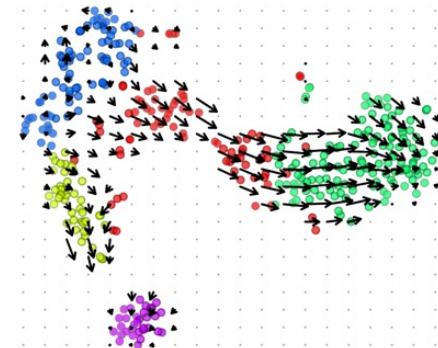
- RNA-seq とは？
- 標準的な解析手法
- ソフトウェアASURAT(阿修羅)の開発
- まとめと今後の展望

RNA-seq とは

少しだけ、生データの説明をする。

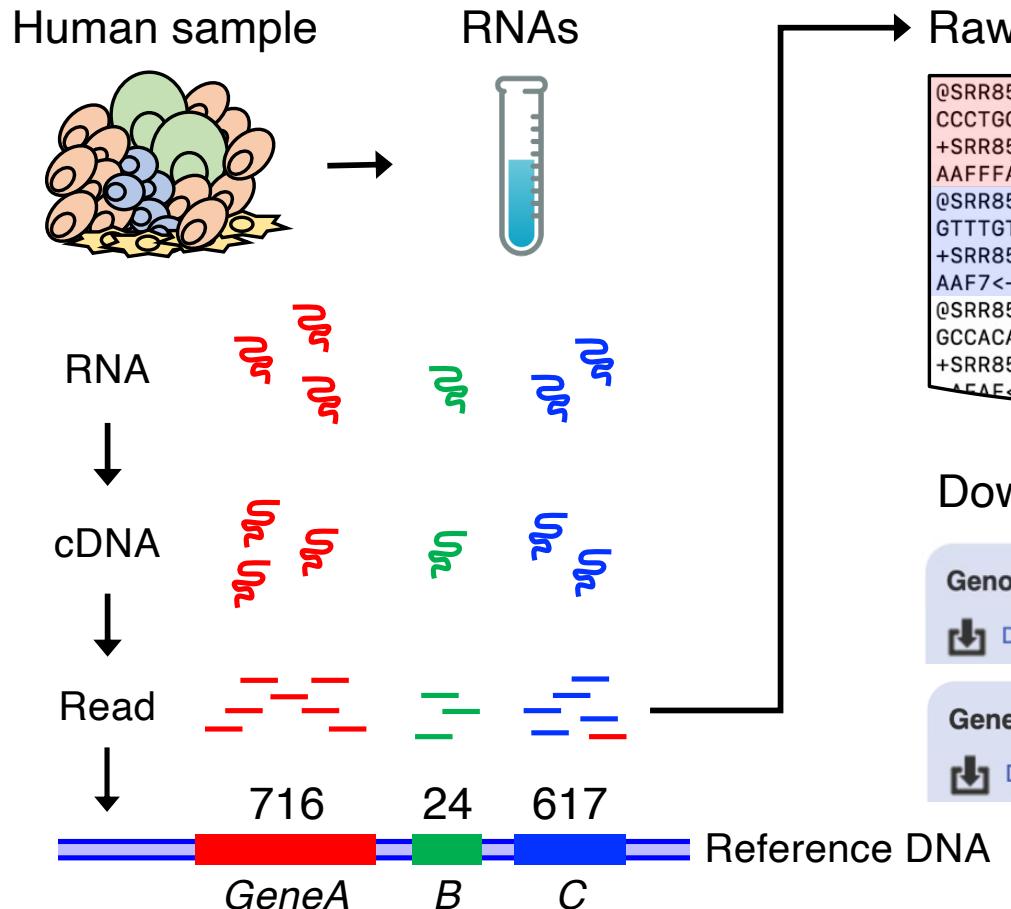
興味がない場合は「ふーん」で構わないが、バイオインフォマティクスの分野では重要。

実際、RNA velocity という考え方はここから生まれた。



RNA velocity: La Manno et al., 2018

バルク RNA-seq



Raw data (FASTQ files)

```
@SRR8518122.1 1 length=150  
CCCTGGACCTGTGGCAGGCAGCCGTGCAGGGTCTCCCCAGAACCC  
+SRR8518122.1 1 length=150  
AAFFFA7FFFJJFJJJJJAFF7AJJFJJ<JA7FJJF77A-<AA-<P  
@SRR8518122.2 2 length=150  
GTTTGTAGAATGAACGCTACAACAACTAAAGTACACTTGTAGA  
+SRR8518122.2 2 length=150  
AAF7<--AFFFJ7FJFJJJJFJJ-JJJJJFJJJ-7<FJJJFFFJ  
@SRR8518122.3 3 length=150  
GCCACATTCTTCACATTGTAGGGTTATCTTCAGTATGAATTTC  
+SRR8518122.3 3 length=150  
AAAE<-F-FJJ<FFJJJJJJ
```

- 1st row Read name
- 2nd row DNA sequence
- 3rd row (+symbol)
- 4th row Quality

※ Potentially this is personal information

Download ref genome from public database

Genome assembly: GRCh38.p14 (GCA_000001405.29)

Download DNA sequence (FASTA)

DNA sequence

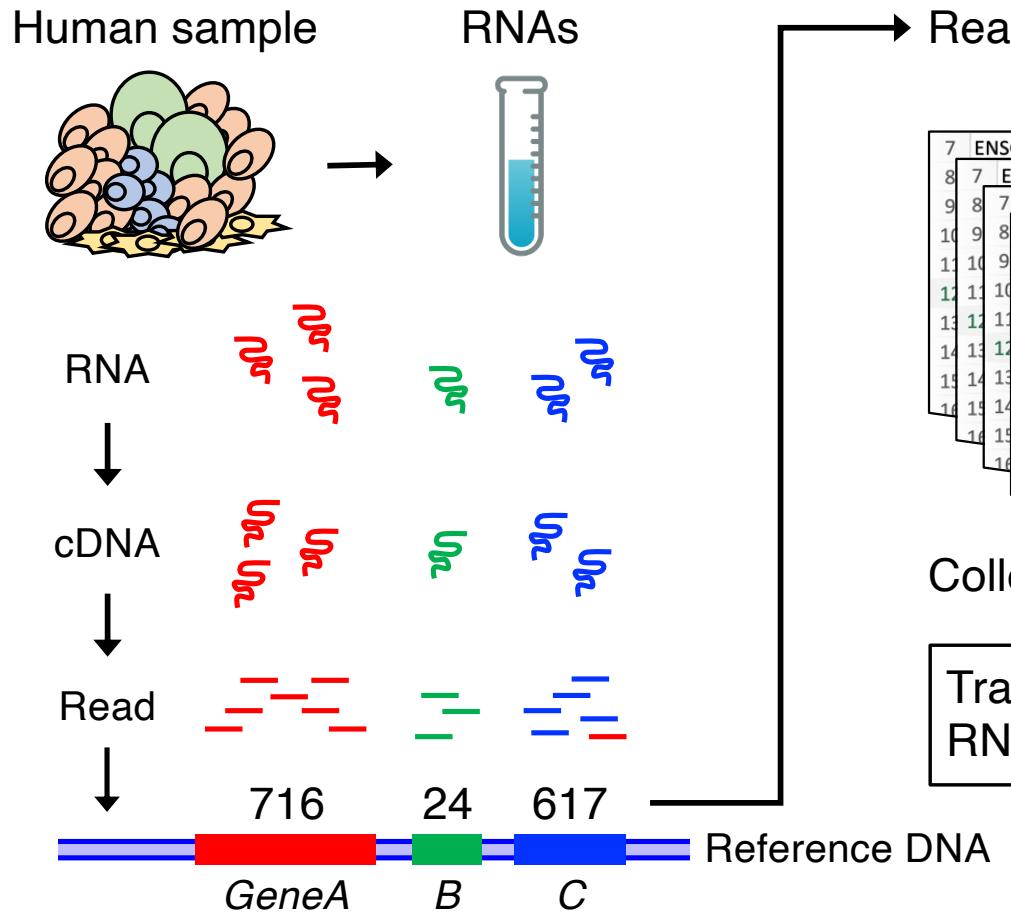
Gene annotation

Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins

Annotation info

<https://asia.ensembl.org>

バルク RNA-seq

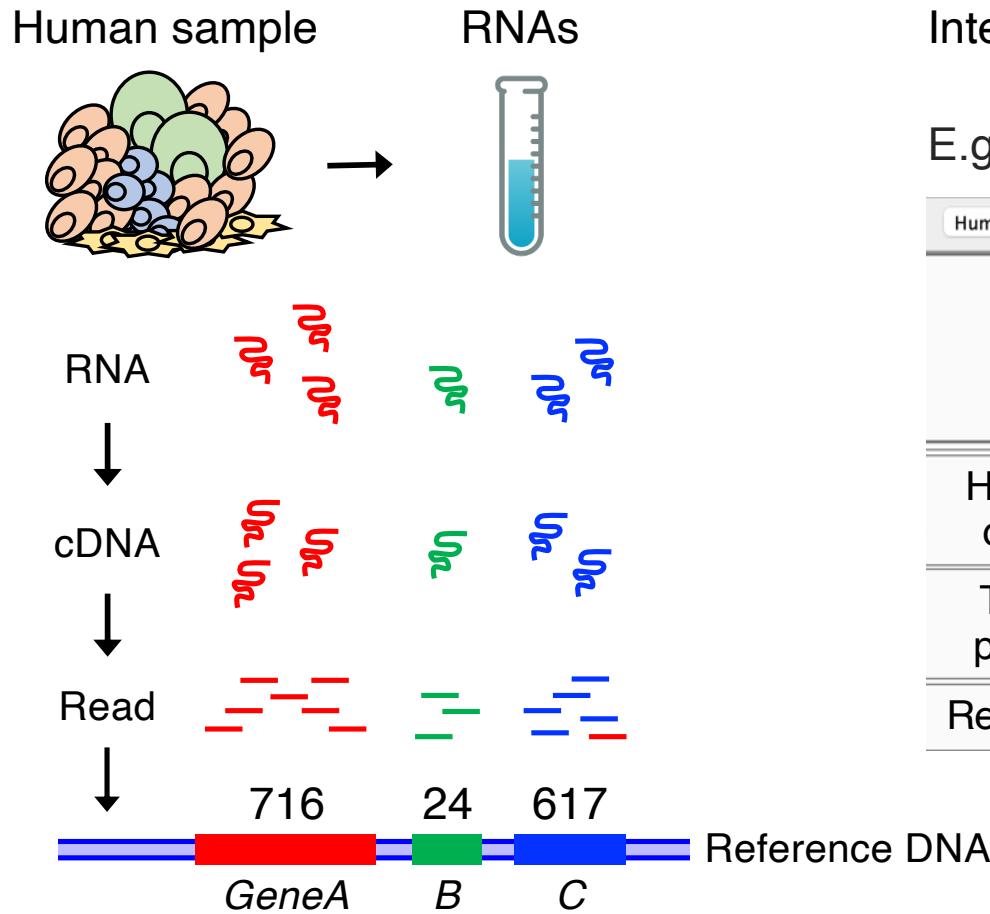


	Gene name	Expression level
7	ENSG00000000003.15	TSPAN6 protein_coding 716
8	ENSG00000000003.15	TSPAN6 protein_coding 892
9	ENSG00000000003.15	TSPAN6 protein_coding 551
10	ENSG00000000003.15	TSPAN6 protein_coding 340
11	ENSG00000000005.6	TNMD protein_coding 24
12	ENSG00000000419.13	DPM1 protein_coding 617
13	ENSG00000000457.14	SCYL3 protein_coding 769
14	ENSG00000000460.17	C1orf112 protein_coding 122
15	ENSG00000000938.13	FGR protein_coding 432
16	ENSG00000000971.16	CFH protein_coding 687
17	ENSG00000001036.14	FUCA2 protein_coding 1596
18	ENSG00000001084.13	GCLC protein_coding
19	ENSG00000001167.14	NEVA protein_coding

Collecting samples produces multivariate data

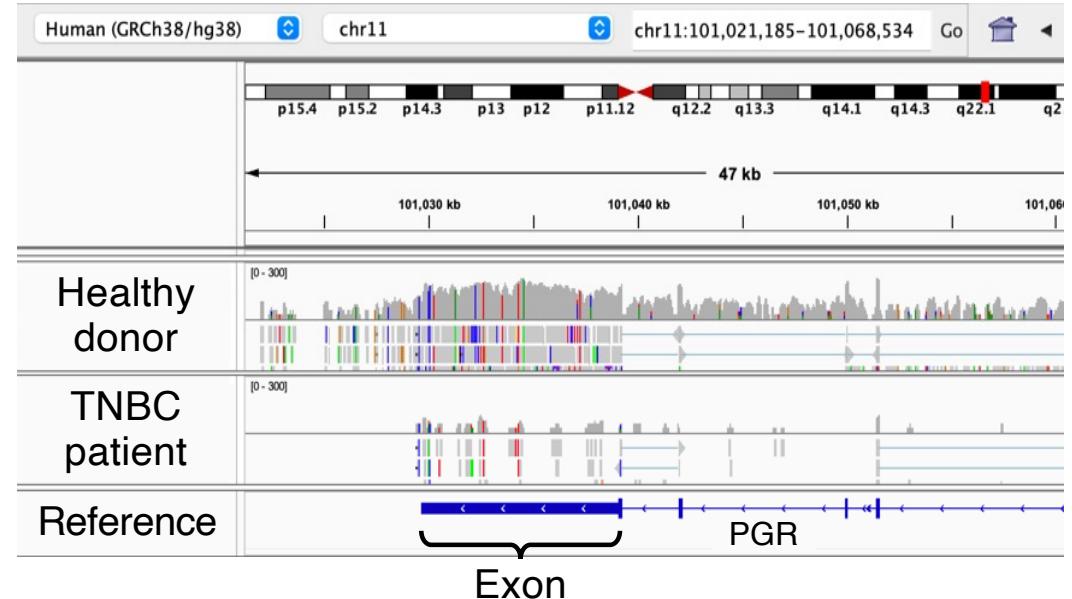
Transcriptome ... Genome-wide RNA information
RNA-sequence ... Experimental technique

バルク RNA-seq



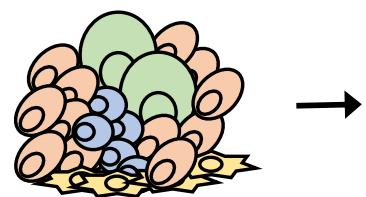
Integrated Genome Viewer

E.g., Triple negative breast cancer (TNBC) data



バルク RNA-seq

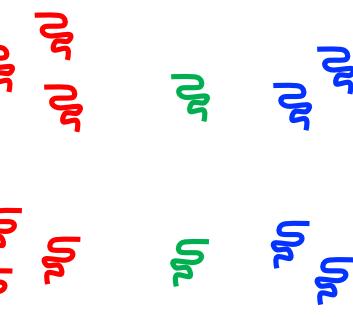
Human sample



RNAs



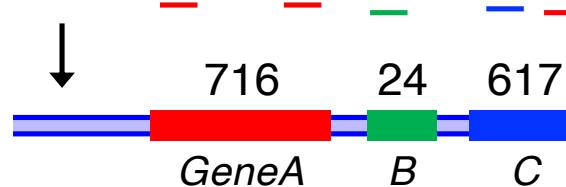
RNA



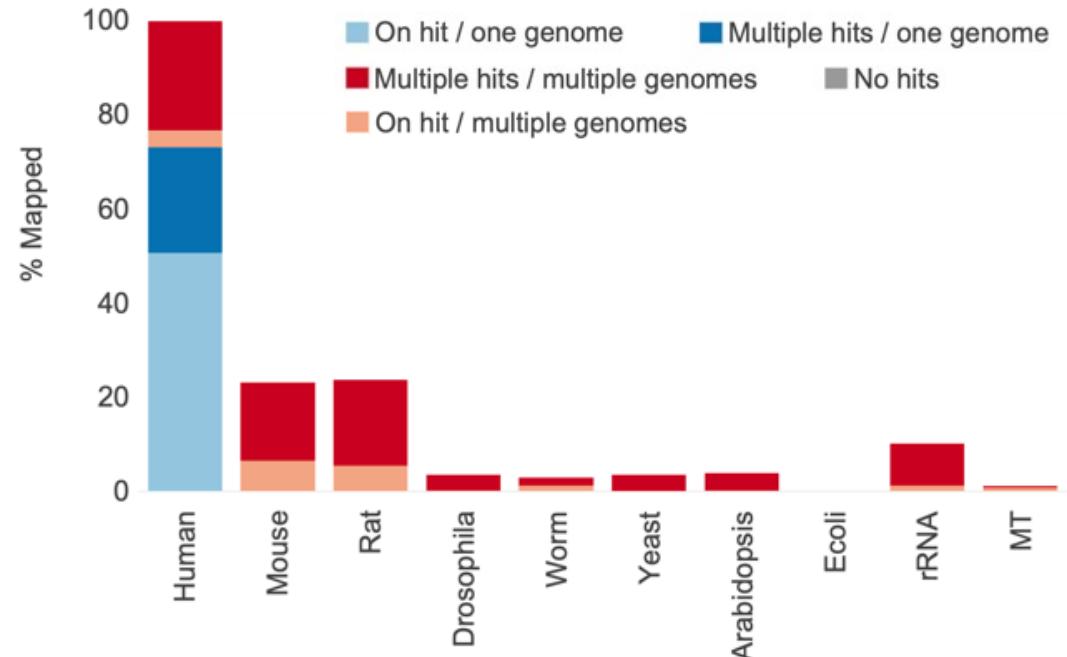
cDNA



Read

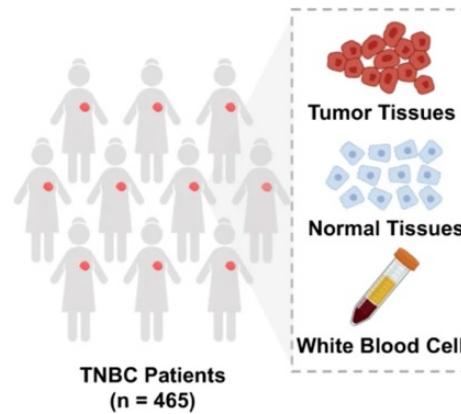
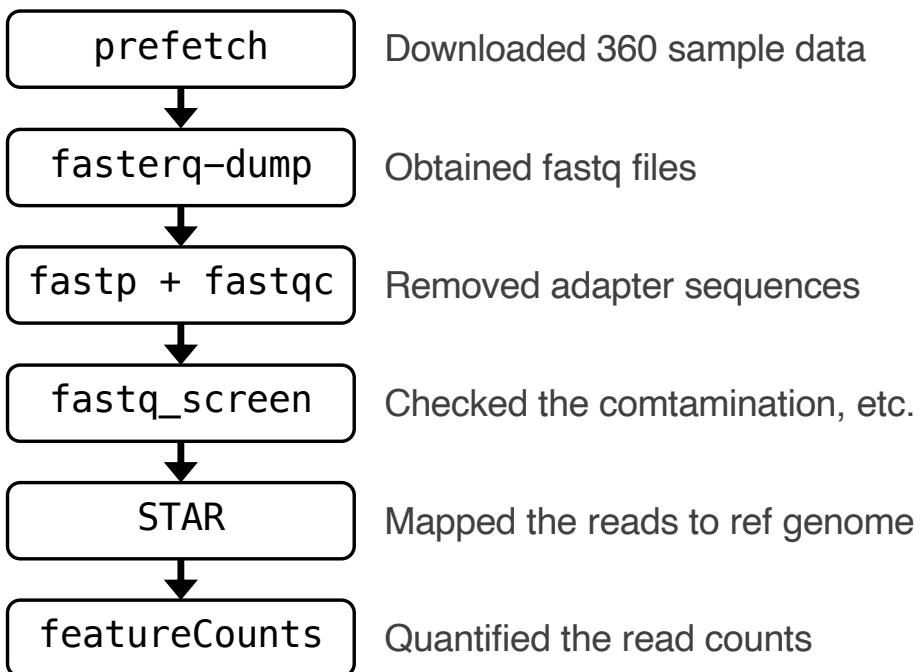


E.g., FastQ Screen

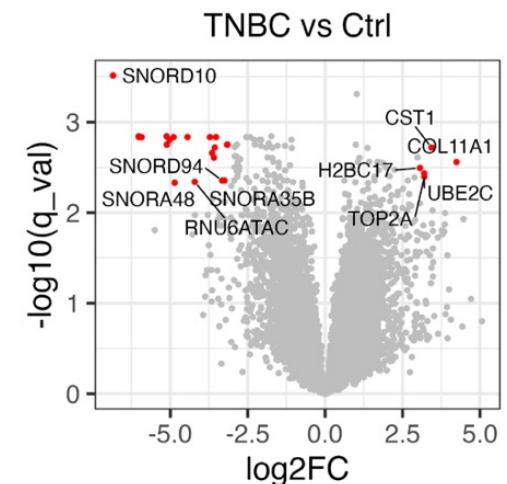


Data quality control (QC) is important

バルク RNA-seq

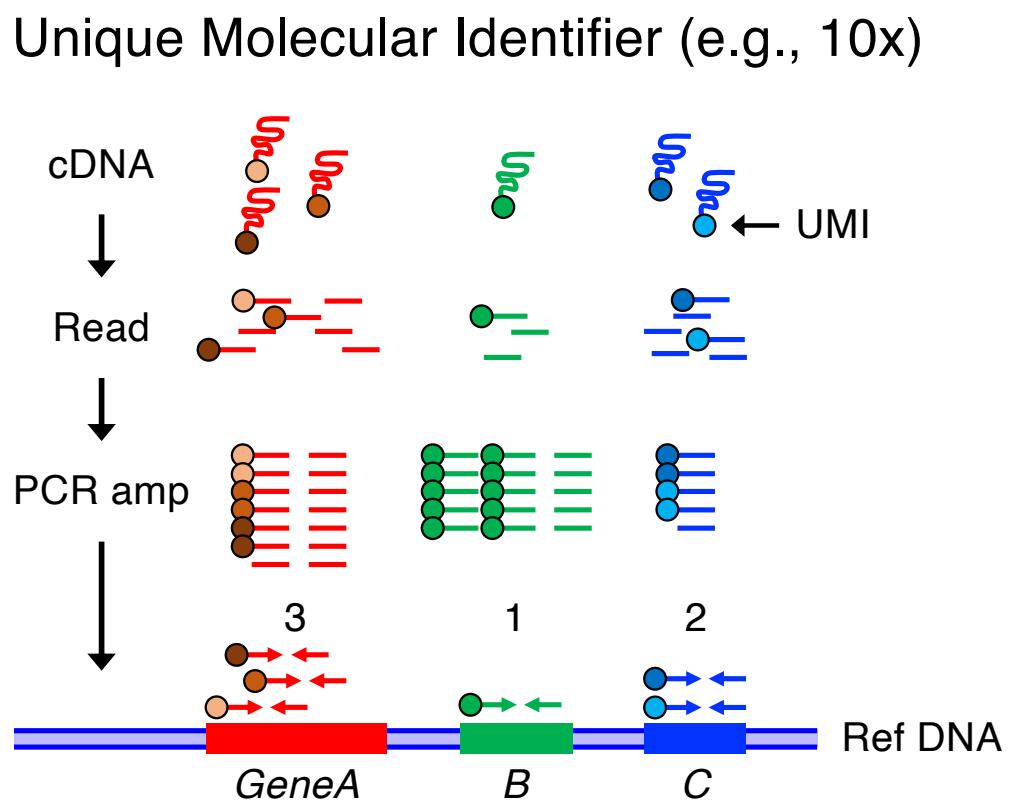
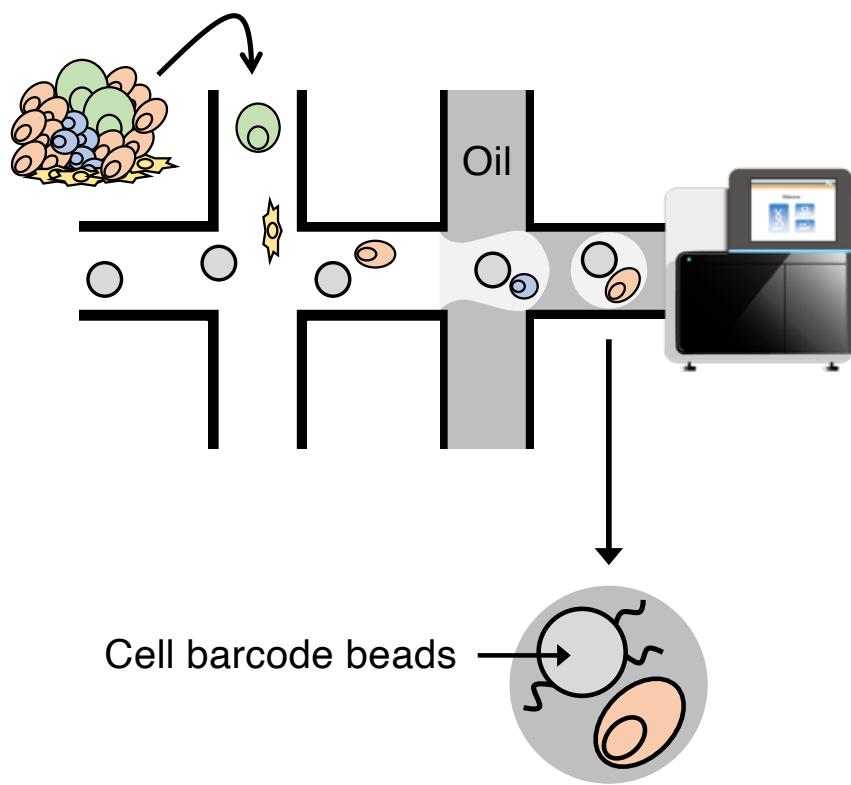


Chen et al., Sci Data, 2020

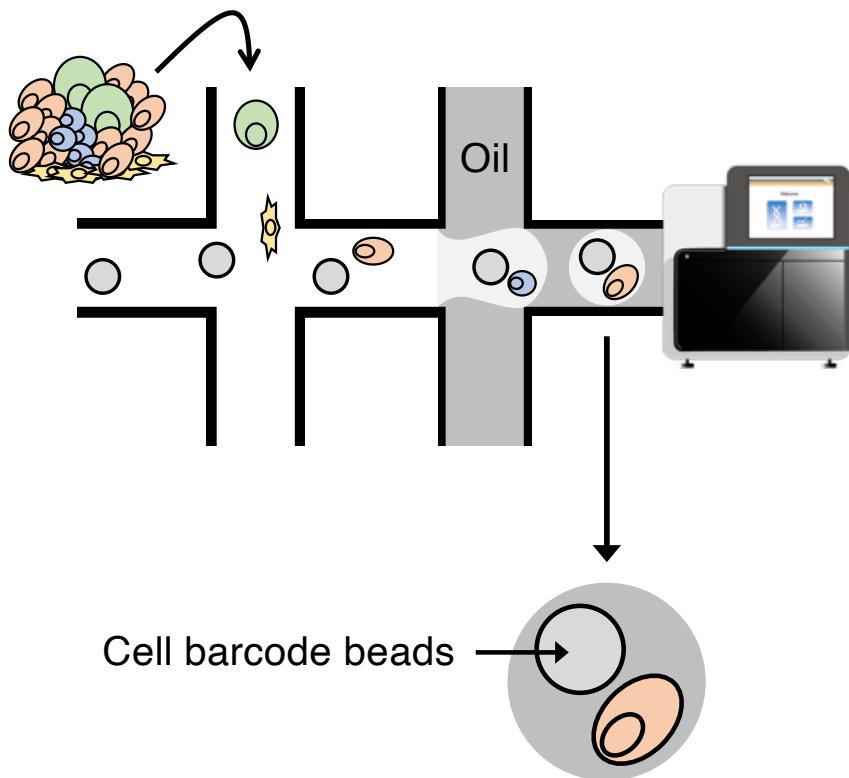


Bulk RNA-seq は安価なので予備検討にも用いられる

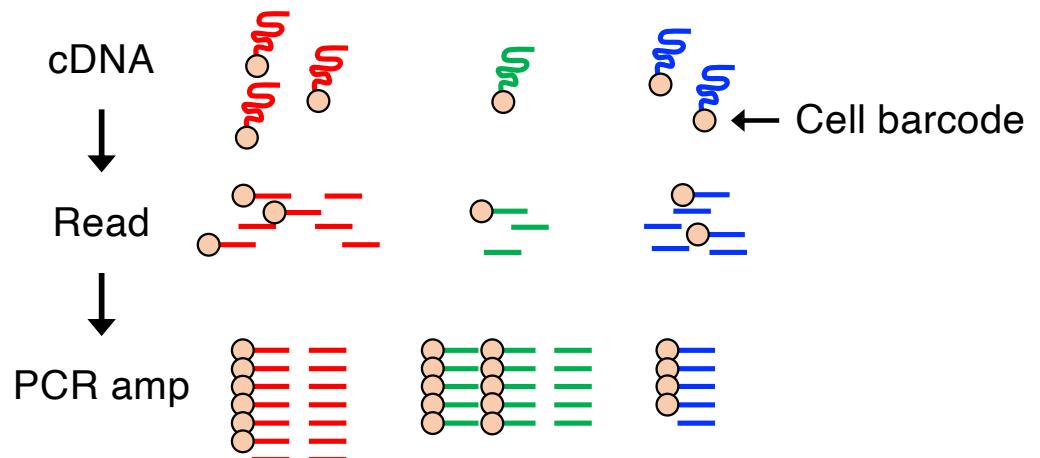
シングルセルRNA-seq (scRNA-seq)



シングルセルRNA-seq (scRNA-seq)

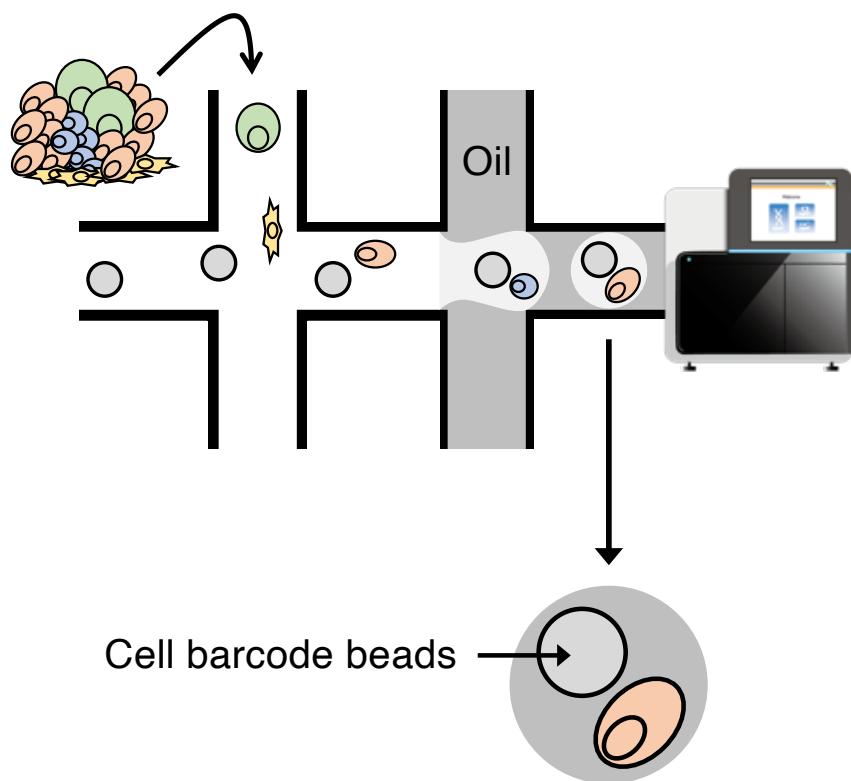


Non-UMI (e.g., SMART-seq2)



Non-UMIでは、PCRによる増幅バイアスが生じうる

シングルセルRNA-seq (scRNA-seq)



Read count table (single-cell version)

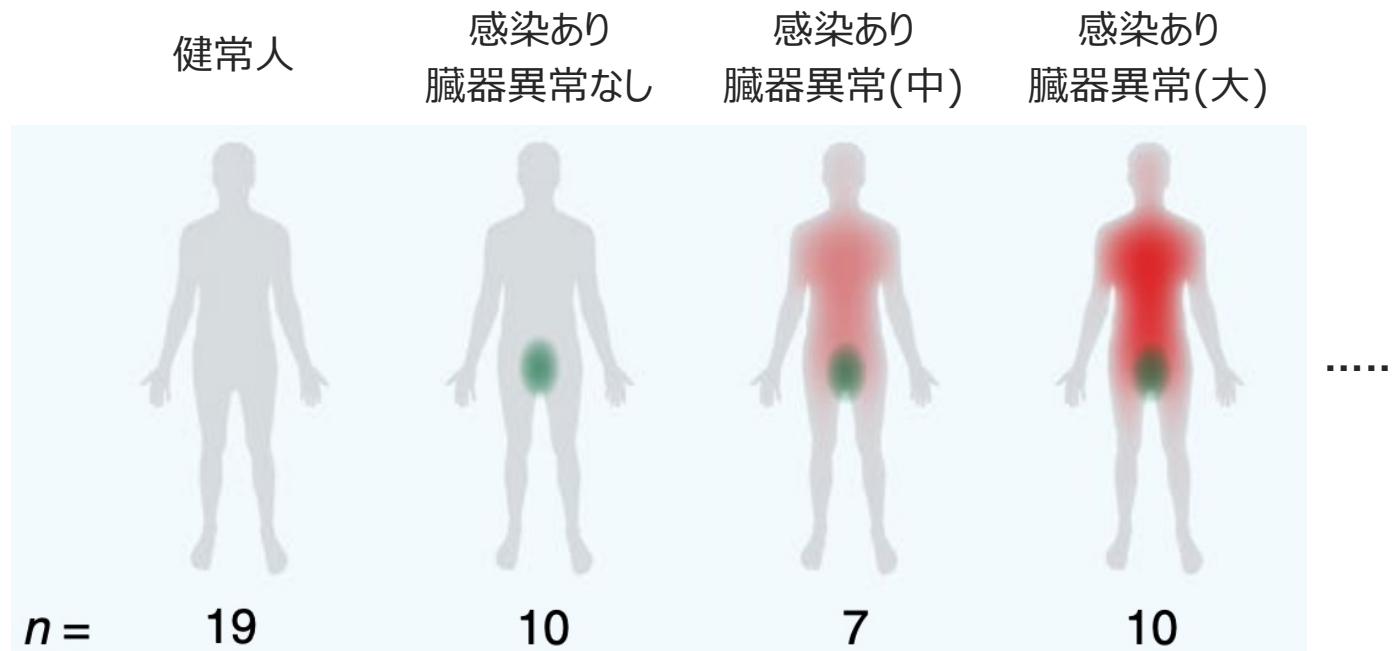
Genes	Cells
GENE, AACCTGAGACGCTT-1	AAACCTGAGCAAT
RP11-34P13.7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	
RP11-34P13.8, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	
RP11-34P13.9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	
F0538757.3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	
F0538757.2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	
AP006222.2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	
RP4-669L17.10, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	
RP5-857K21.4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	
RP11-206L10.9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	
RP11-206L10.9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	

アジェンダ

- RNA-seq とは？
- 標準的な解析手法
- ソフトウェアASURAT(阿修羅)の開発
- まとめと今後の展望

敗血症のシングルセル解析

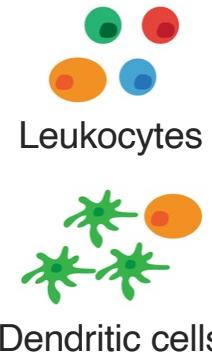
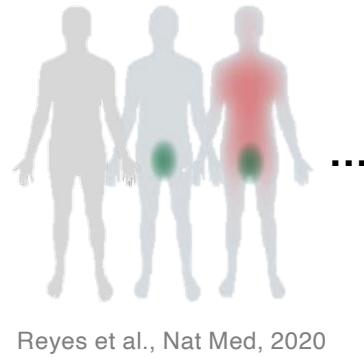
細菌感染性の敗血症（炎症 → 多臓器不全、非常に高い死亡率）



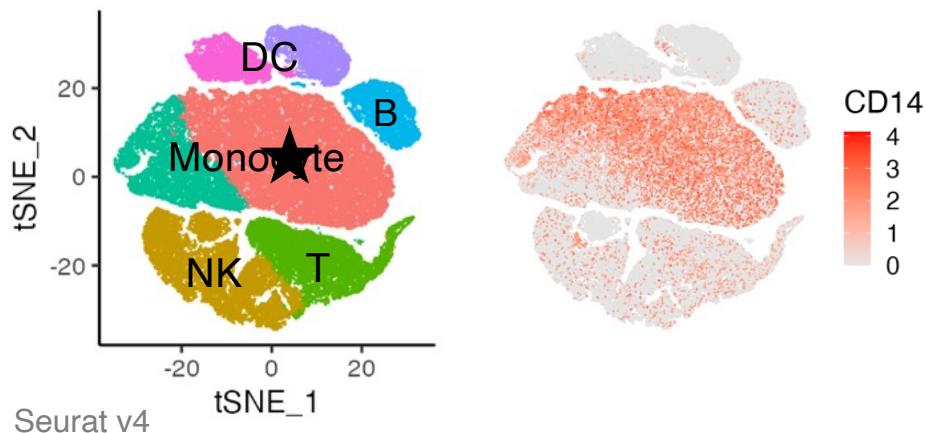
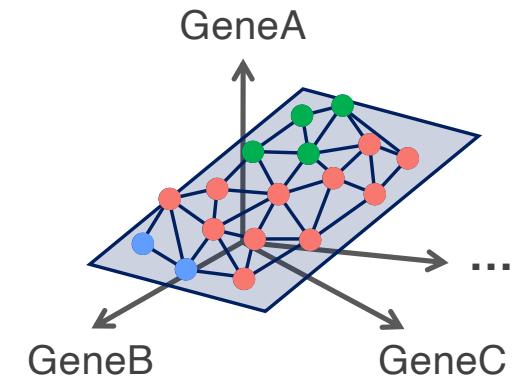
Reyes et al., Nat Med, 2020

敗血症のシングルセル解析

被験者($n = 65$)



GENE, AACCTGAGACGCTT-1, AACCTGAGCAAT
RP11-34P13.7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
RP11-34P13.8, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
RP11-34P13.9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
F0538757.3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
F0538757.2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
AP006222.2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
RP4-669L17.10, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
RP5-857K21.4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
RP11-206L10.9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0



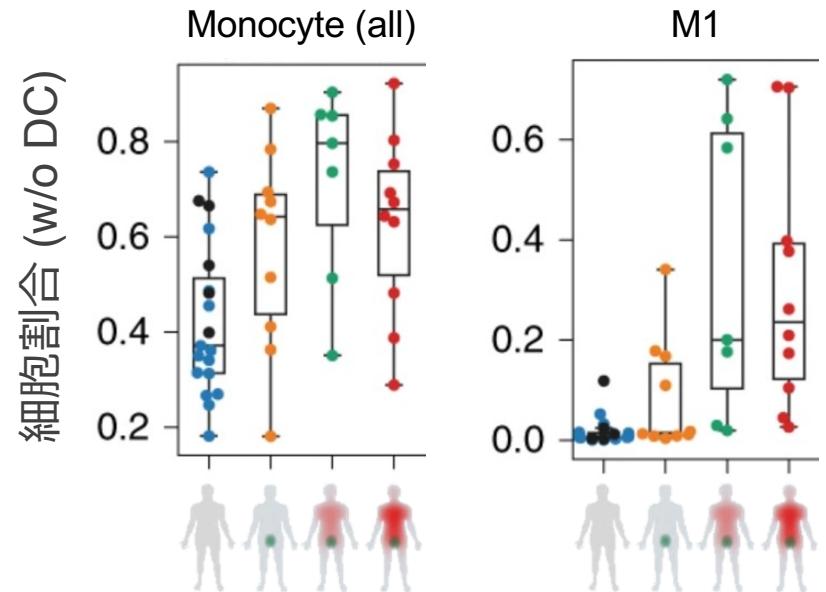
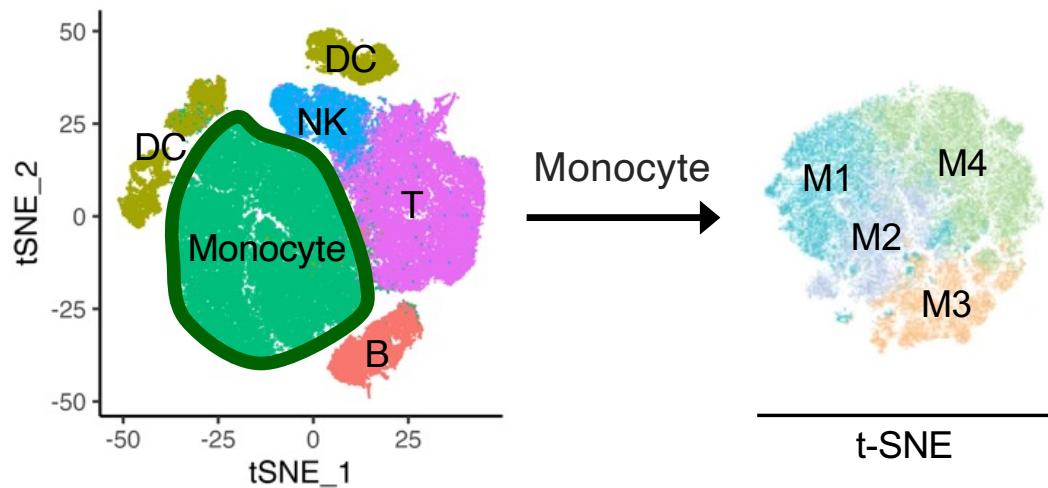
クラスター★で発現量が有意に高い遺伝子は361個もある ($p_{adj} < 10^{-100}$)

→ 文献調査

クラスタリング結果の生物学的解釈が常に問題となる
最適なクラスタリングとは？ (情報科学 ≠ 生物学)

敗血症のシングルセル解析

Scanpy を使用したクラスタリング



Reyes et al., Nat Med, 2020

結論：単球の亜集団 M1 は敗血症になると急増する

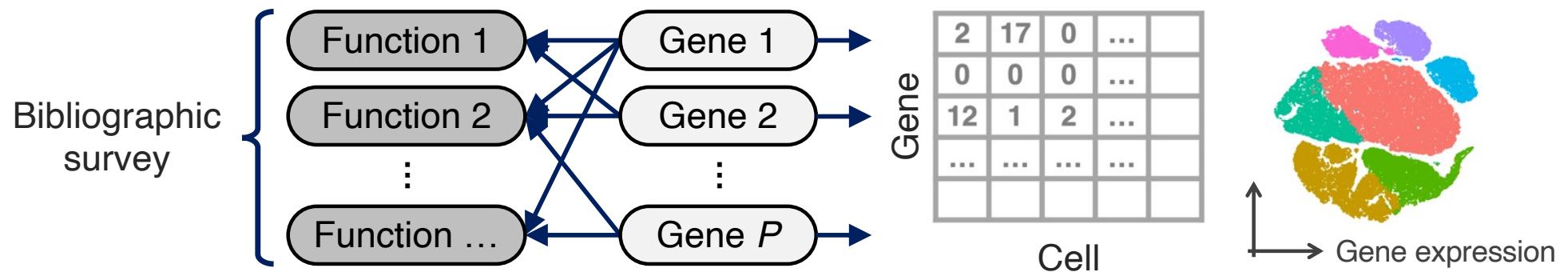
問題点：クラスタリングの信頼性について疑問が残る

アジェンダ

- RNA-seq とは？
- 標準的な解析手法
- ソフトウェアASURAT(阿修羅)の開発
- まとめと今後の展望

「生物学的解釈」を数学的に定義する

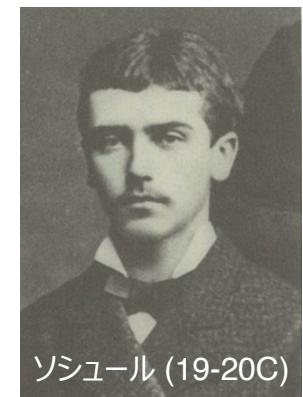
従来のアプローチ



「生物学的解釈」を数学的に定義したい

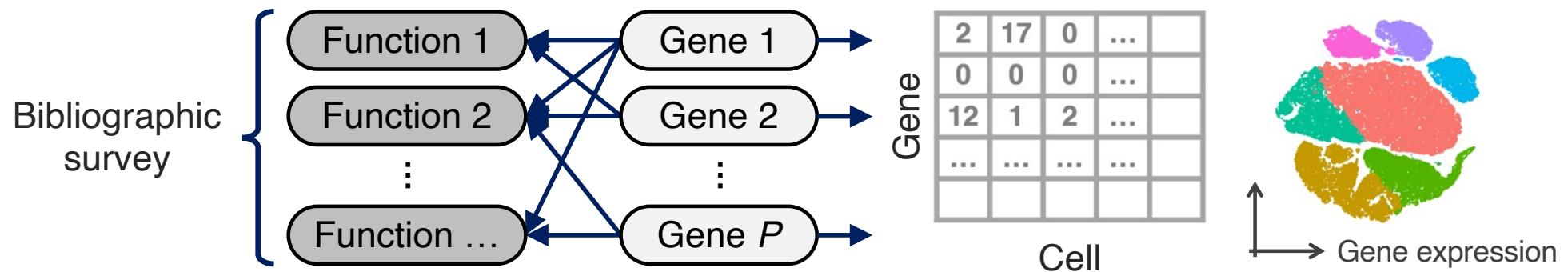
「解釈」=「意味づけ」

「意味」とは？ → 記号学

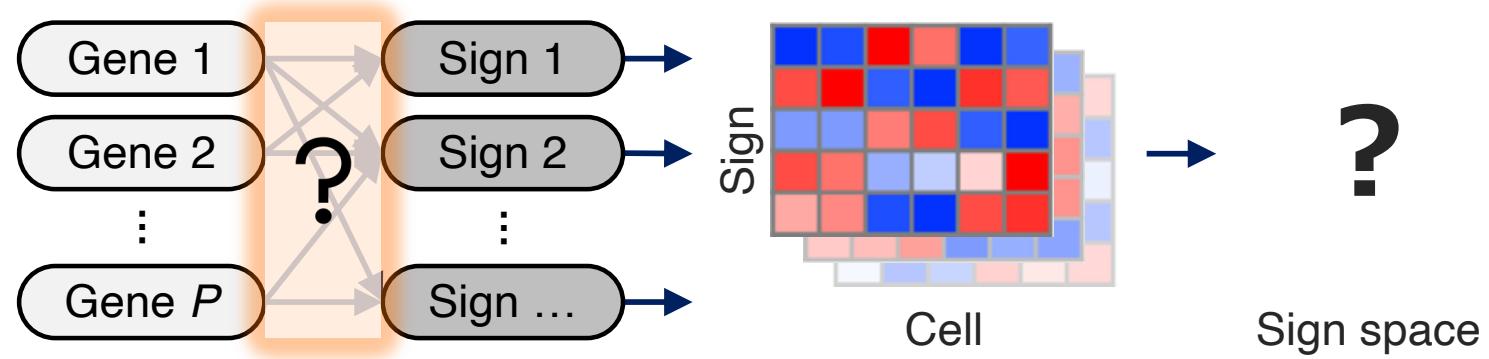


「生物学的解釈」を数学的に定義する

従来のアプローチ



記号学的アプローチ



知識データベースの活用

例：細胞オントロジーのデータベース

ID	Description	IC	Count	Gene
CL:0000000	cell	0	12072	DRB1/HMGN2/HSPB1/HYAL1/IDE/IGHM/IL1RN/IL4R/IL5/ARVCF/CAPN3/CLIC2/DHX8/DHX9/DHX15/STMN1/M1/MIR138-2/MIR9-1/MIR9-2/MIR9-3/MIR675/CD3G/MAP1/EFHC1/LHX6/FUT10/DIXDC1/SRGAP2C/SYNE2/SUN1
CL:0000003	native cell	0.280344255995428	7	ERCC1/WRN/TP63/CHEK2/ROMO1/SALL1/NEK4
CL:0000014	germ line stem cell	8.09923031047809	7	PIWIL2/PRDM14/NANOS2/ETV5/ING2/ZBTB16/STRA8
CL:0000015	male germ cell	3.90654984753513	21	H2AX/HSPA2/MLH1/SYCP1/TNP1/TRIP13/REC8/TCFL5/

2020年12月時点、383の細胞型に関するIDが登録されている

知識データベースの活用

例：遺伝子オントロジーのデータベース

ID	Description	IC	Count	Gene
GO:0008150	biological_process	0	18866	AS1/PITPNM1/PDIA4/NFE2L3/RNF14/VPS9D1/CARTPT/RA AS1/NEU3/FAXDC2/FAM114A2/ZNF275/ALDH1L1/FTCD/F 8/BLOC1S6/FBXW8/FBXO25/FBXO24/TSPAN17/FBXO22/S AS/MBTPS2/HOOK1/CDC40/CHST15/ZMYND10/PLA1A/U 4-1/IGHV3-30-3/KAT14/PBXIP1/CBX8/RCN3/ZNF286A/ZNF 6/TSPYL2/RBM26/PRSS22/OXCT2/PERP/MMP27/NPAS3/CJ
GO:0000003	reproduction	4.03787500459501	1503	ABAT/ACR/ACOX1/ACVR1/ACVR1B/ACVR2A/ADA/ADA 2/MIR21/ASTL/LHX8/LRRC52/RIMBP3B/PIWIL3/SPATA31C
GO:0002376	immune system process	2.78567002936747	3287	A1BG/A2M/SERPINA3/ABL1/AOC1/ABR/ACAA1/ACLY/A H/MR1/HLX/HMGB1/HMGB2/HMGB3/HMGN2/HMOX1/HI 5/IGHV7-81/IGHV6-1/IGHV5-10-1/IGHV5-51/IGHV4-38-2/IC

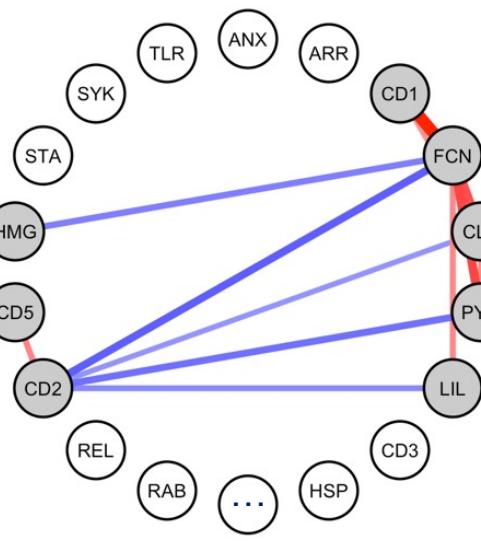
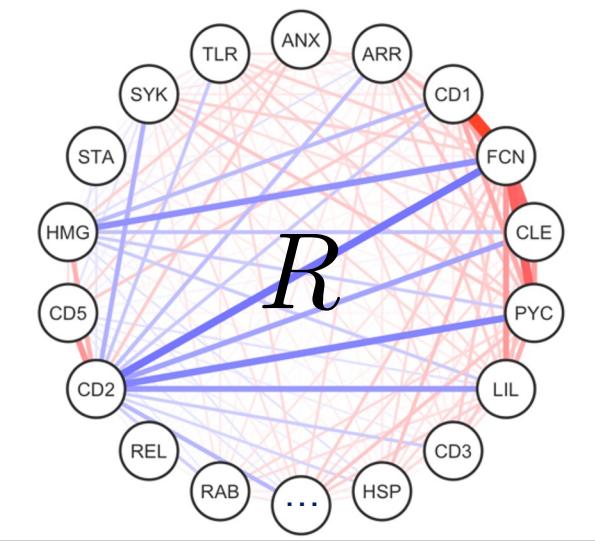
2020年12月時点、44272個の生物機能に関するIDが登録されている

「記号」の定義

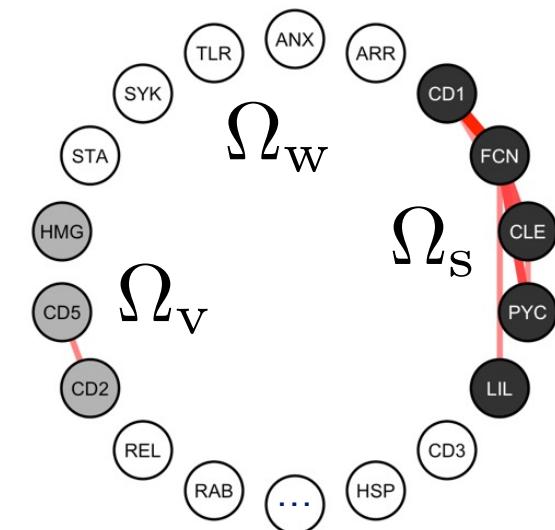
T = “Interleukin-8 production” (GO:0032637)

Ω = Gene set

R = “Relation” among genes



K-means clustering



三つ組 (T, Ω_s, R) and (T, Ω_v, R) を記号と名付ける

記号(sign)の生成

Public databases

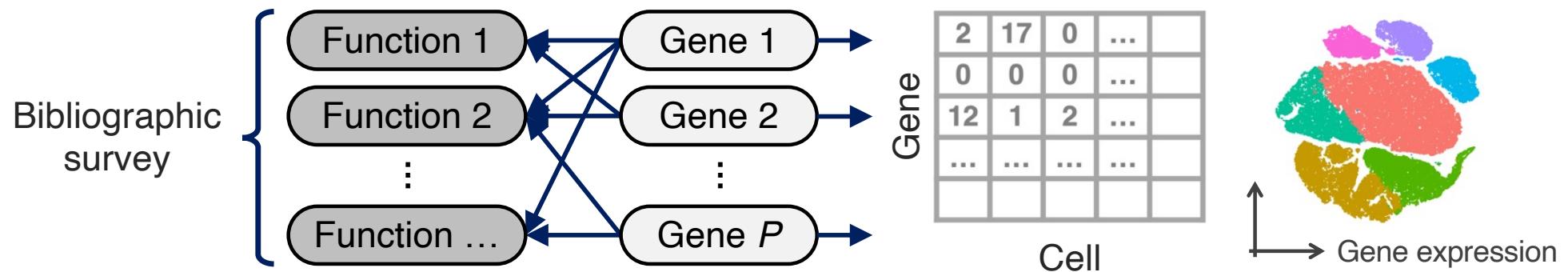


...

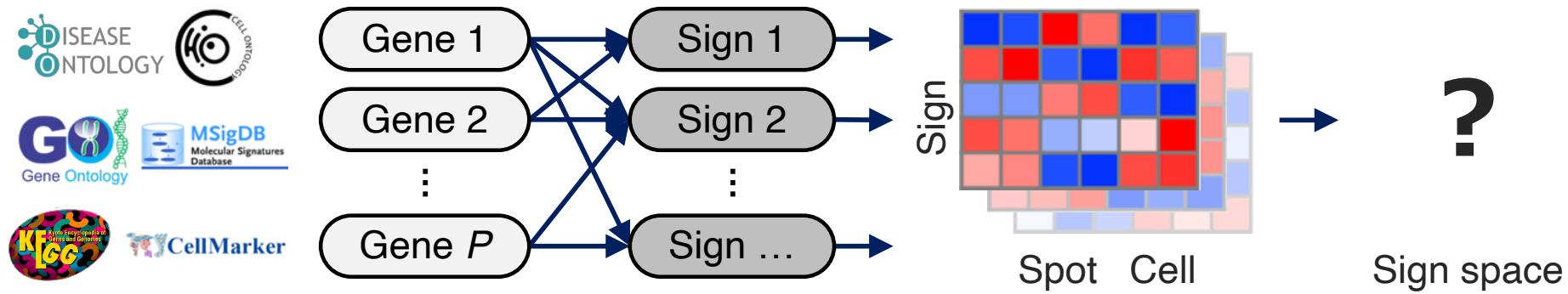
SignID	Description	IC	StrgCorrGene	VariCorrGene	WeakCorrGene
1	GO:0000003	reproduction	4.037875	NA	NA
2	GO:0007610	behavior	5.017167	NA	NA
3	GO:0019740	nitrogen utilization	10.371865	NA	NA
4	GO:0022414	reproductive process	4.039296	NA	NA
5	GO:0022610	biological adhesion	4.062674	NA	NA
6	GO:0040007	growth	4.854412	NA	NA
7	GO:0040011	locomotion	3.686254	NA	NA
8	GO:0043473	pigmentation	6.893706	NA	NA
9	GO:0048511	rhythmic process	5.780794	NAMPT/PTGDS/MYCBP2/R...	HNRNPD/NONO/PRKDC/ET...
10	GO:0051703	intraspecies interac...	7.973970	NA	NA
11	GO:0051704	multi-organism proce...	4.575199	NA	NA
12	GO:0098754	detoxification	6.950865	S100A9/GPX1/SOD2/CD3...	ADH5/PIM1/TXN/ALOXE...
13	GO:0110148	biomineralization	6.610665	NA	ACOX1/ADA/ADAM10/ADM...
...
15	GO:0052	sexual reproduction	4.789873	NA	...

「生物学的解釈」を数学的に定義する

従来のアプローチ



記号学的アプローチ



ASURAT(阿修羅)の開発

YouTubeからも公開

 Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers

Search:

Home » Bioconductor 3.15 » Software Packages » ASURAT

ASURAT

platforms all rank 2112 / 2140 support 0 / 0 in Bioc < 6 months

build ok updated before release dependencies 48

DOI: [10.18129/B9.bioc.ASURAT](https://doi.org/10.18129/B9.bioc.ASURAT)  

Functional annotation-driven unsupervised clustering for single-cell data

Bioconductor version: Release (3.15)

ASURAT is a software for single-cell data analysis. Using ASURAT, one can simultaneously perform unsupervised clustering and biological interpretation in terms of cell type, disease, biological process, and signaling pathway activity. Inputting a single-cell RNA-seq data and knowledge-based databases, such as Cell Ontology, Gene Ontology, KEGG, etc., ASURAT transforms gene expression tables into original multivariate tables, termed sign-by-sample matrices (SSMs).

Author: Keita Iida [aut, cre] , Johannes Nicolaus Wibisana [ctb]

Maintainer: Keita Iida <kiida@protein.osaka-u.ac.jp>

TO GO TV

Bio"Pack"athon2022#11 @Online

細胞の多面的分類を行う
遺伝子発現解析ツールの開発と
Bioconductor奮闘記

飯田渓太
Bio"Pack"athon

DBCLS

Documentation

Bioconductor

- Package [vignettes](#)
- [Workflows](#) for learning
- Several [online books](#) for comprehensive coverage of a particular research field, biological question, or technology.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about

O'REILLY

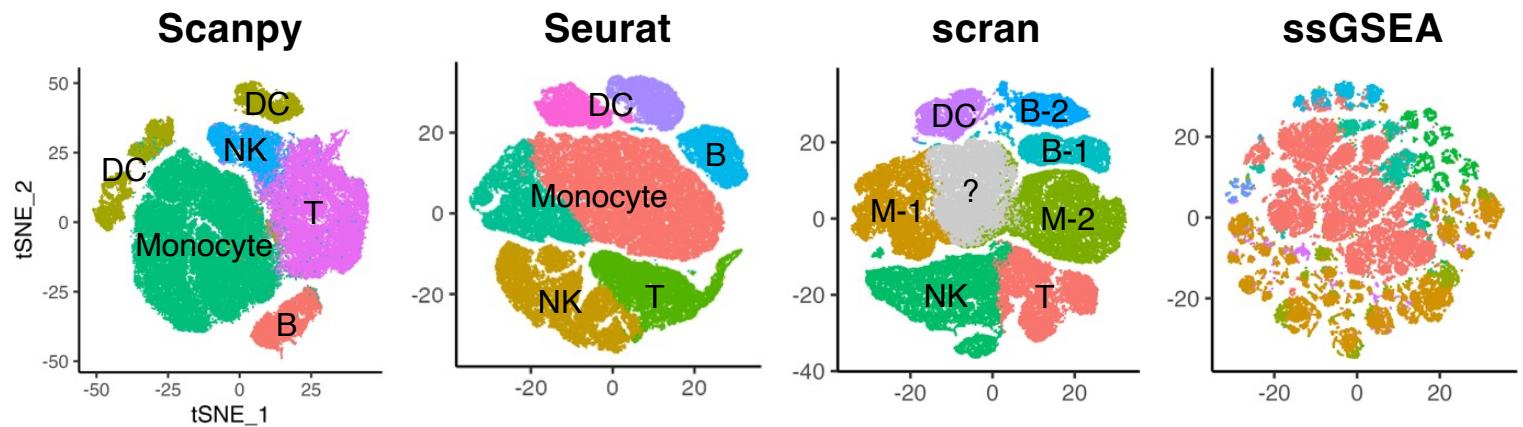
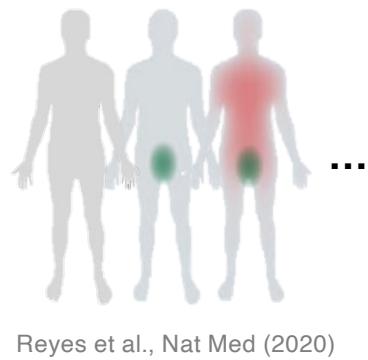
Rパッケージ
開発入門

テスト、文書化、コード共有の手法を学ぶ

Hadley Wickham 著
高川 伸人 翻訳
石井 由美子 翻訳
吉原 駿

ASURATによる敗血症のシングルセル解析

被験者($n = 65$)

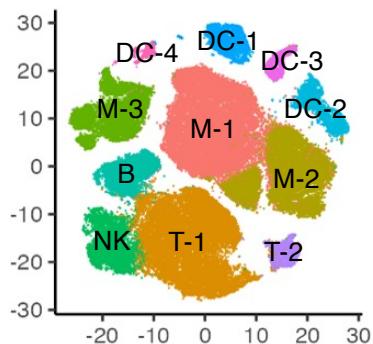


ASURAT

細胞型のデータベース



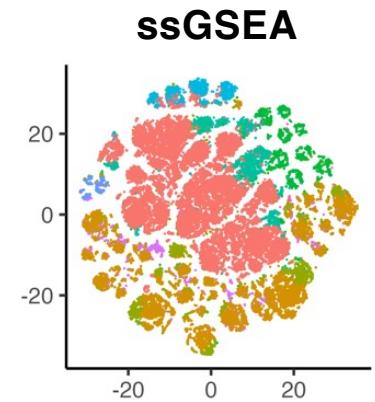
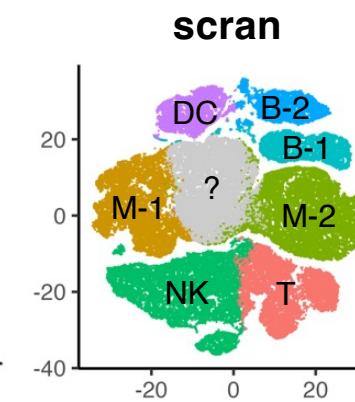
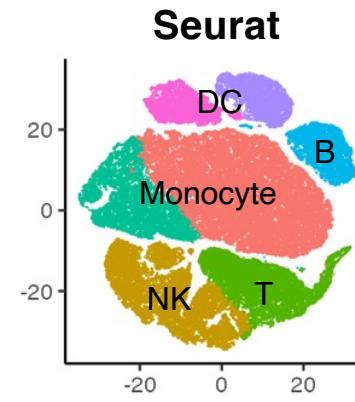
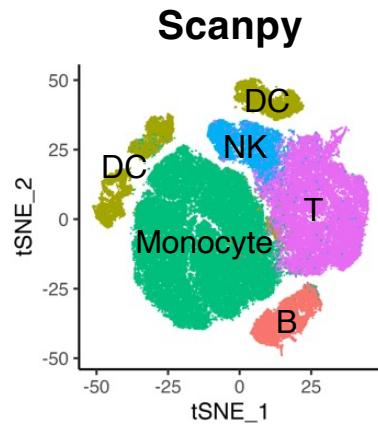
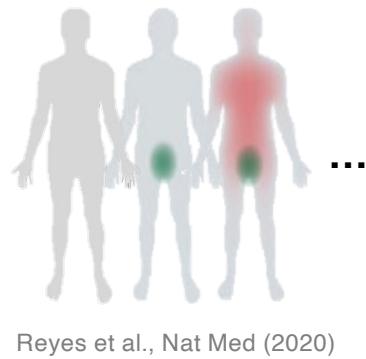
細胞型の記号空間



Iida et al., *Bioinformatics*, 2022

ASURATによる敗血症のシングルセル解析

被験者($n = 65$)

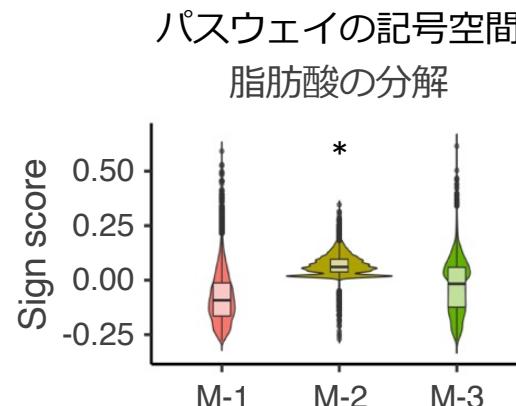
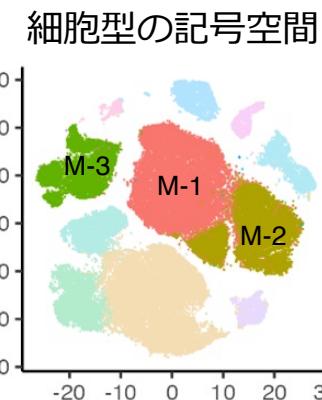


ASURAT

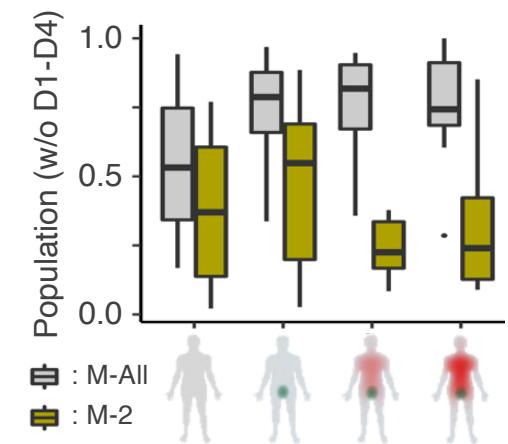
細胞型のデータベース



パスウェイのデータベース



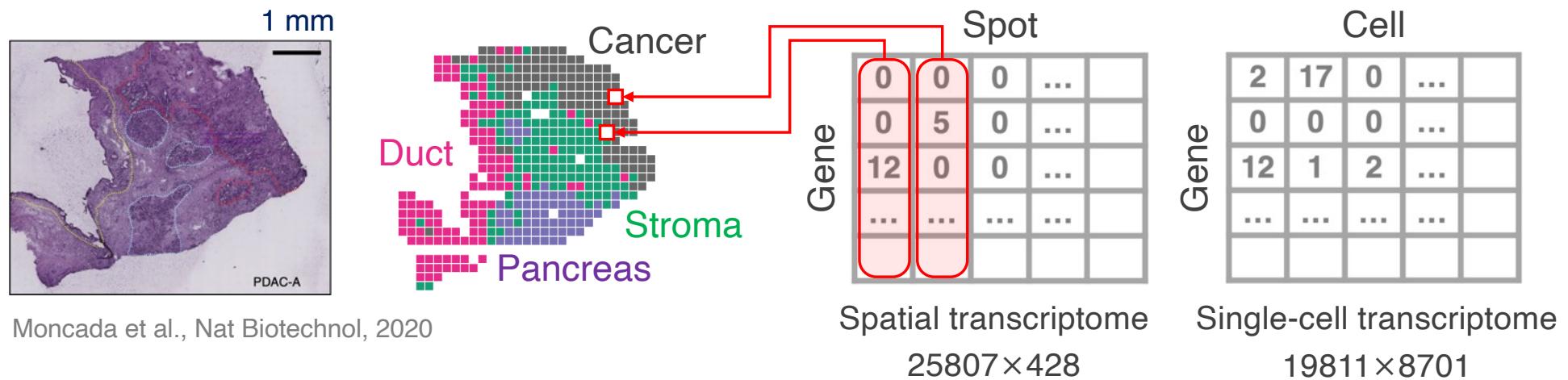
Iida et al., *Bioinformatics*, 2022



がんのデータ解析にも適用可能か？

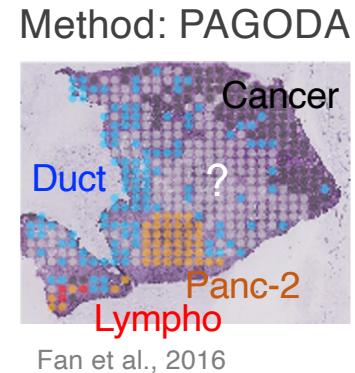
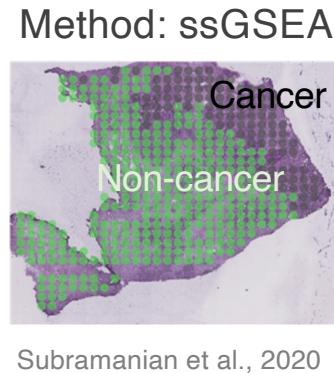
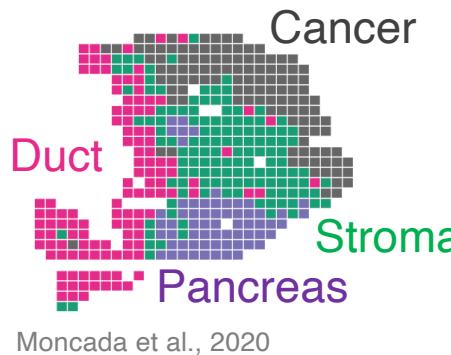
膵管腺癌の空間オミクス解析

目的 1 細胞・空間オミクスデータを用いて悪性細胞をみつける(細胞分類)

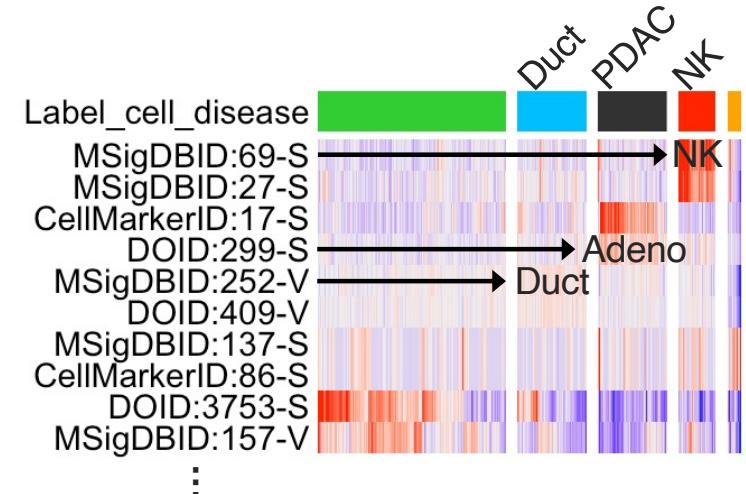


膵管腺癌の空間オミクス解析

Previous results

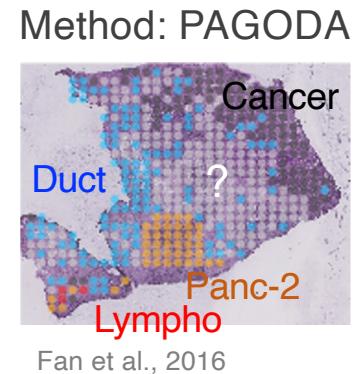
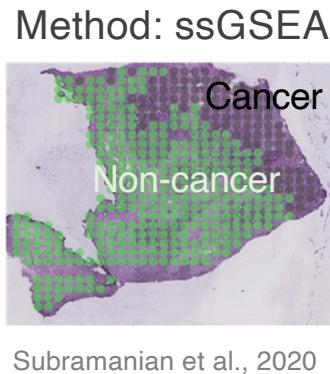
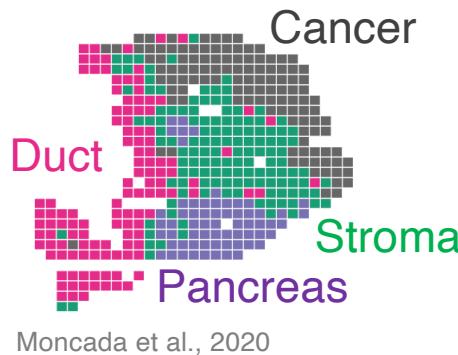


Our result using ASURAT

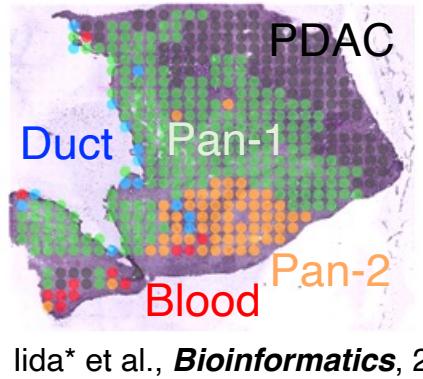


膵管腺癌の空間オミクス解析

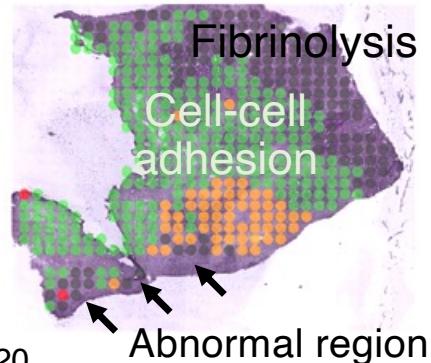
Previous results



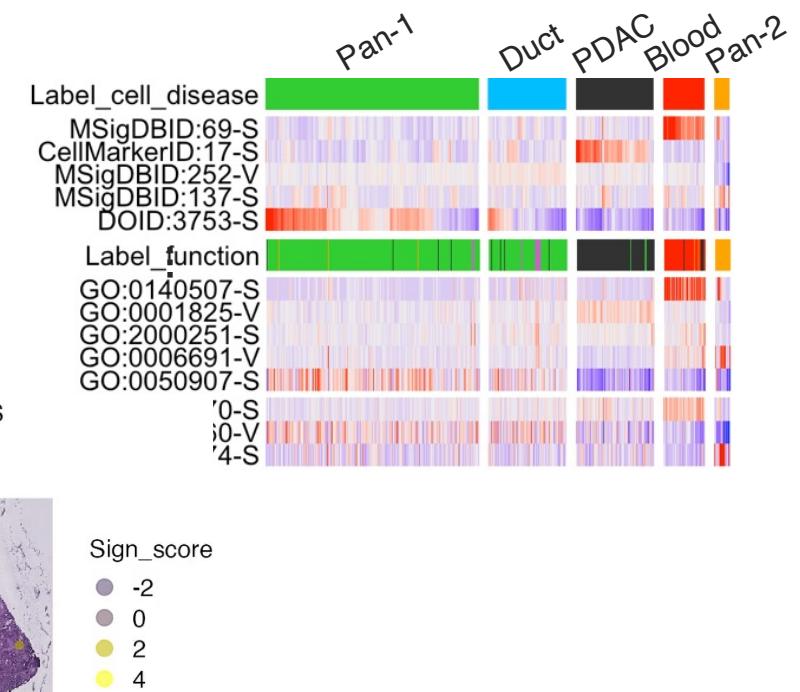
Disease and cell type



Biological process

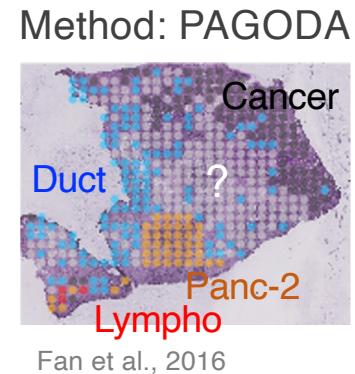
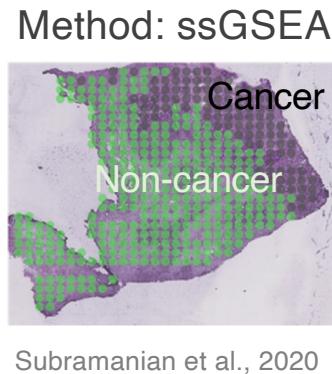
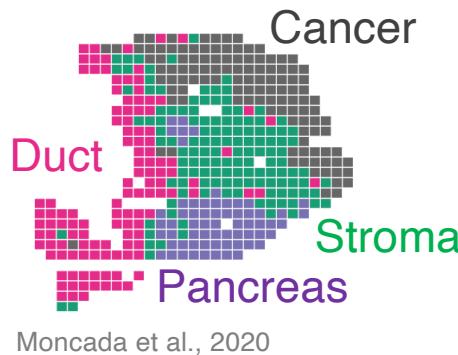


Our result using ASURAT

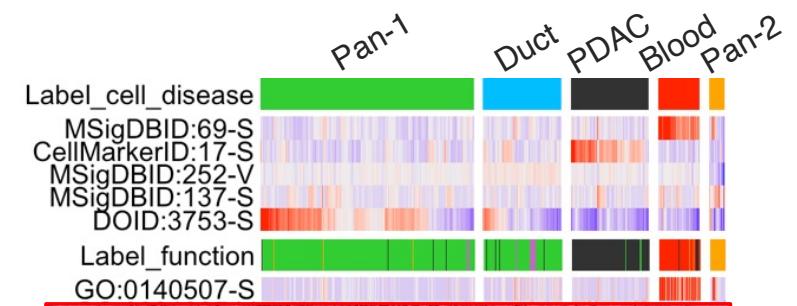


膵管腺癌の空間オミクス解析

Previous results

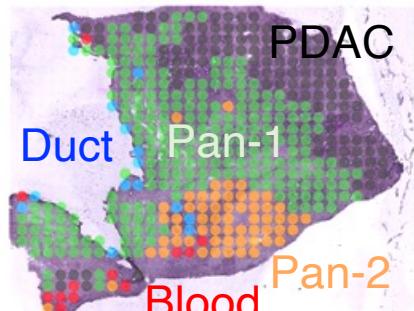


Our result using ASURAT

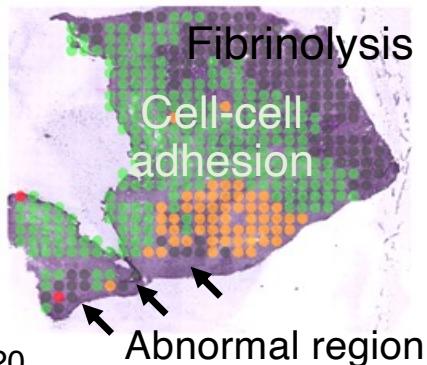


Normal pancreas involved in cancer

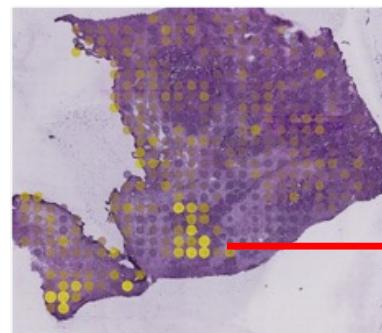
Disease and cell type



Biological process

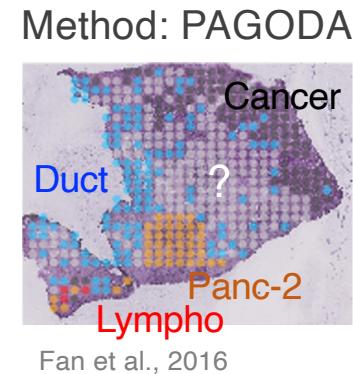
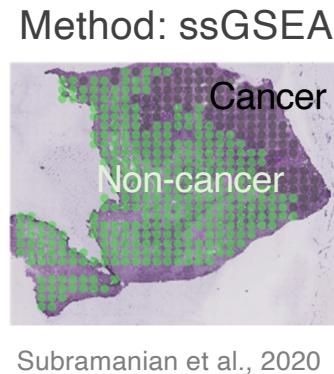
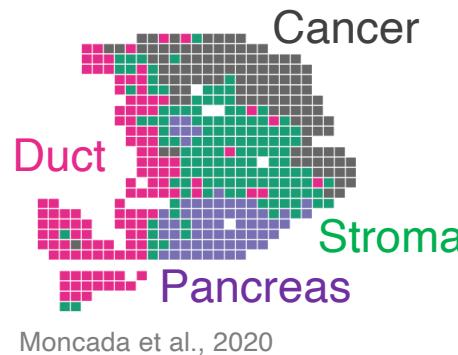


Pathway: Th17 diff

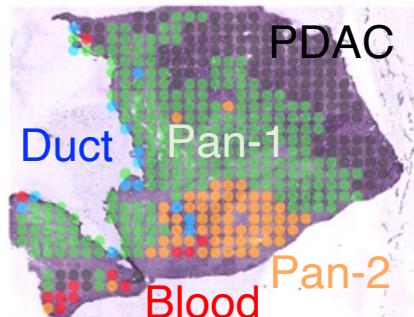


膵管腺癌の空間オミクス解析

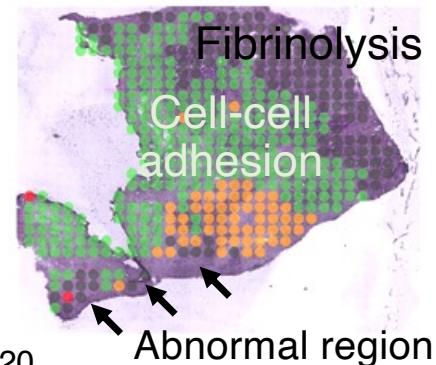
Previous results



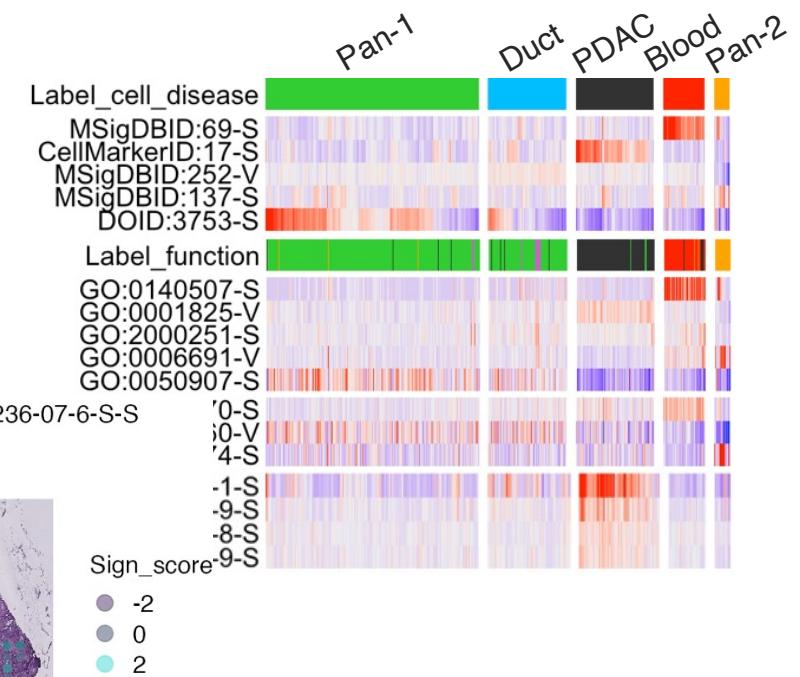
Disease and cell type



Biological process

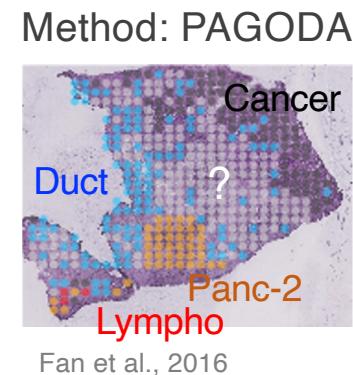
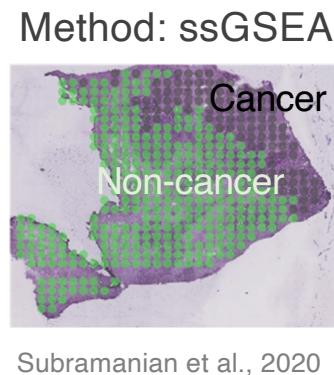
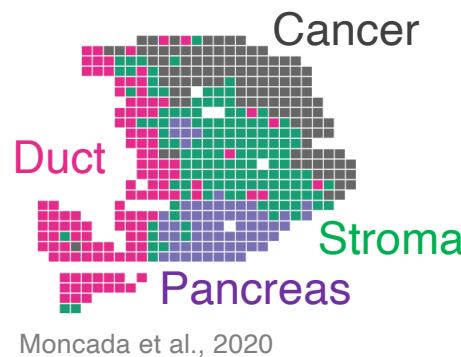


Our result using ASURAT

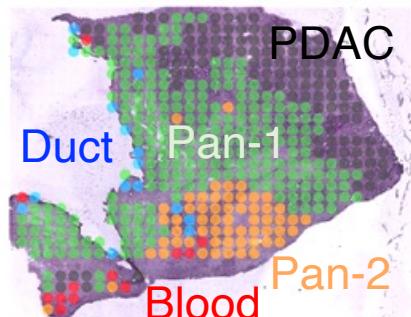


膵管腺癌の空間オミクス解析

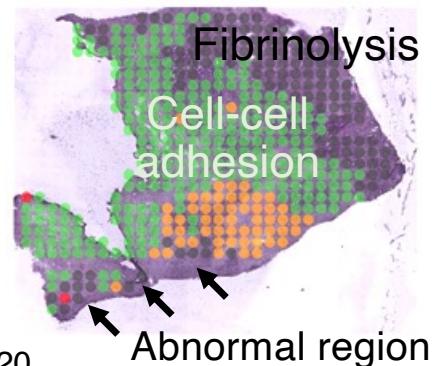
Previous results



Disease and cell type

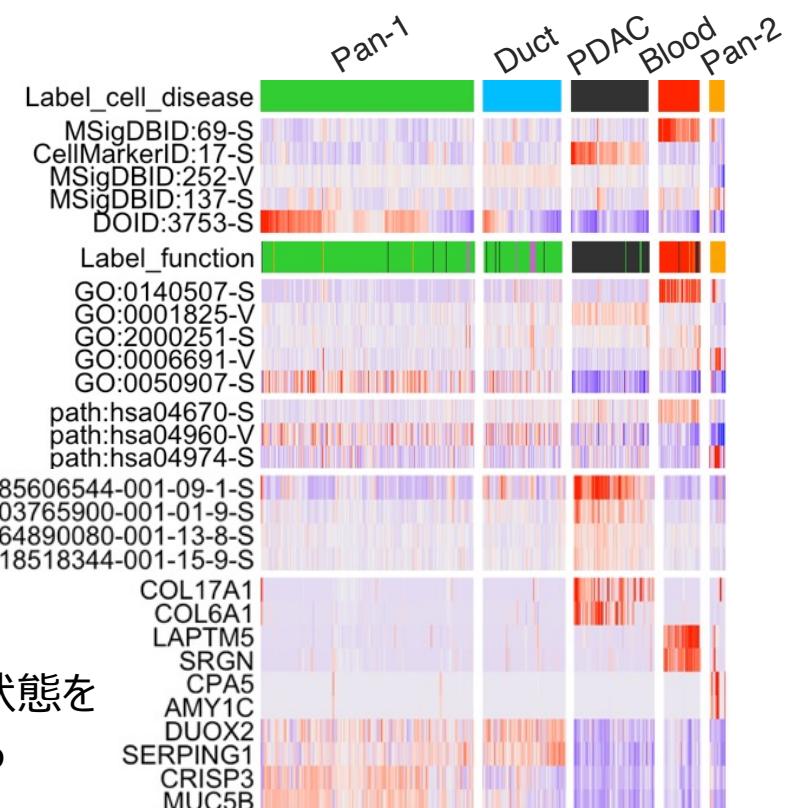


Biological process



各スポット、細胞の状態を
多面的に分析できる

Our result using ASURAT



まとめ

