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 渓太

大阪大学蛋白質研究所

講義内容の概要

今日、次世代シークエ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。

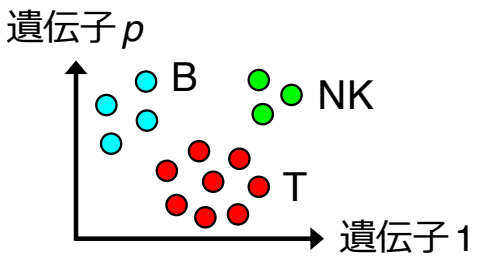
今回は、シングルセルRNAシークエンス(scRNA-seq)データの情報解析をテーマに、「細胞分類」と呼ばれる基本的な問題を考える。ふつう、細胞分類といえば、細胞集団をクラスタリングしたのち、各クラスターを既知の細胞種に当てはめる(＝分類する)ことを指す。しかし、がんのように分類自体が困難なデータの場合はどうすればよいだろうか？

ふつうの細胞分類とは？

- 細胞１の発現データ
- 細胞２の発現データ

⋮

- 細胞 *n* の発現データ



事前知識

B細胞のマーカー遺伝子
T細胞のマーカー遺伝子

⋮

NK細胞のマーカー遺伝子

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

アジェンダ

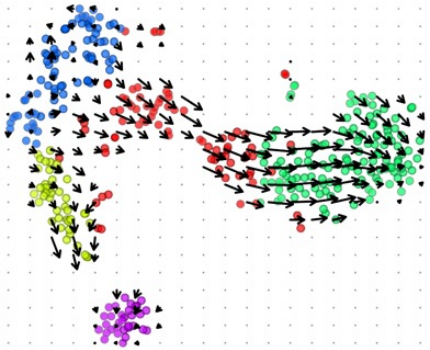
- RNA-seq とは？
- 標準的な解析手法
- ソフトウェアASURAT(阿修羅)の開発
- まとめ

RNA-seq とは

少しだけ、生データの説明をする。

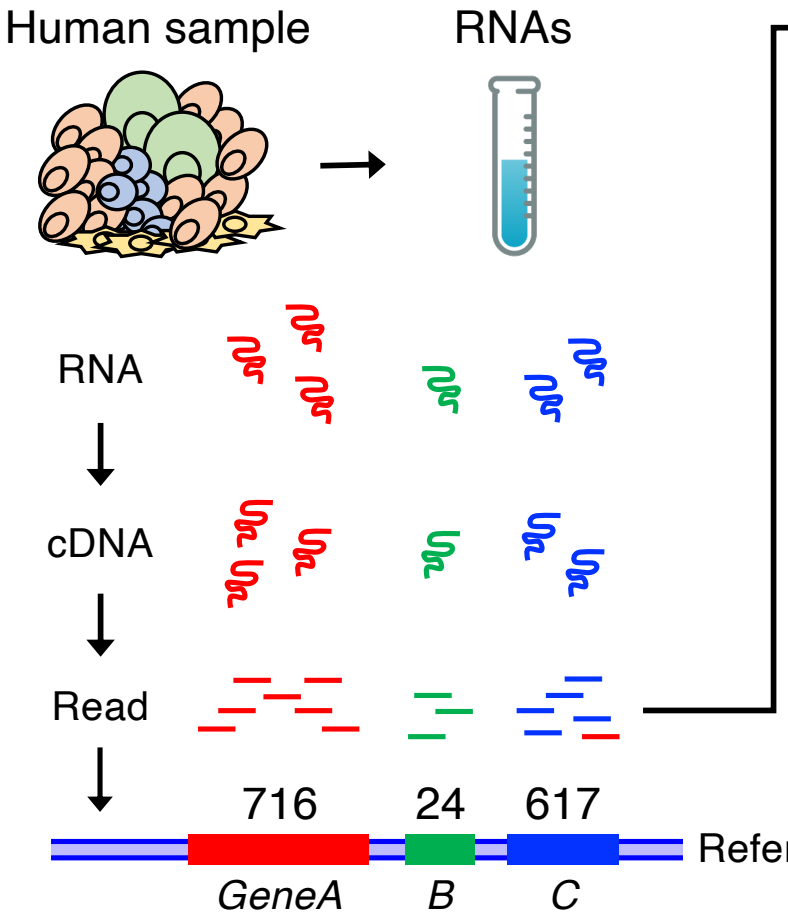
興味がない場合は「ふーん」で構わないが、バイオインフォマティクスの分野では重要。

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
CCCTGGACCTGTGGCAGGCGGCCGTGCAGGGTCCTCCCCAGAACCT
+SRR8518122.1 1 length=150
AAFFFA7FFFJJFJJJJJAFF7AJJFJJ<JA7FJJF77A-<AA-<F
@SRR8518122.2 2 length=150
GTTTGTCAGAATGAACGCTACAACAATCTAAAAGTACACTTGTAGA
+SRR8518122.2 2 length=150
AAF7<--AFFFJ7FJFJFJJJJFJJ-JJJJJFJJJ-7<FJJJFFFJ
@SRR8518122.3 3 length=150
GCCACATTCTTCACATTTGTAGGGTTTATCTTCAGTATGAATTTCC
+SRR8518122.3 3 length=150
```

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)

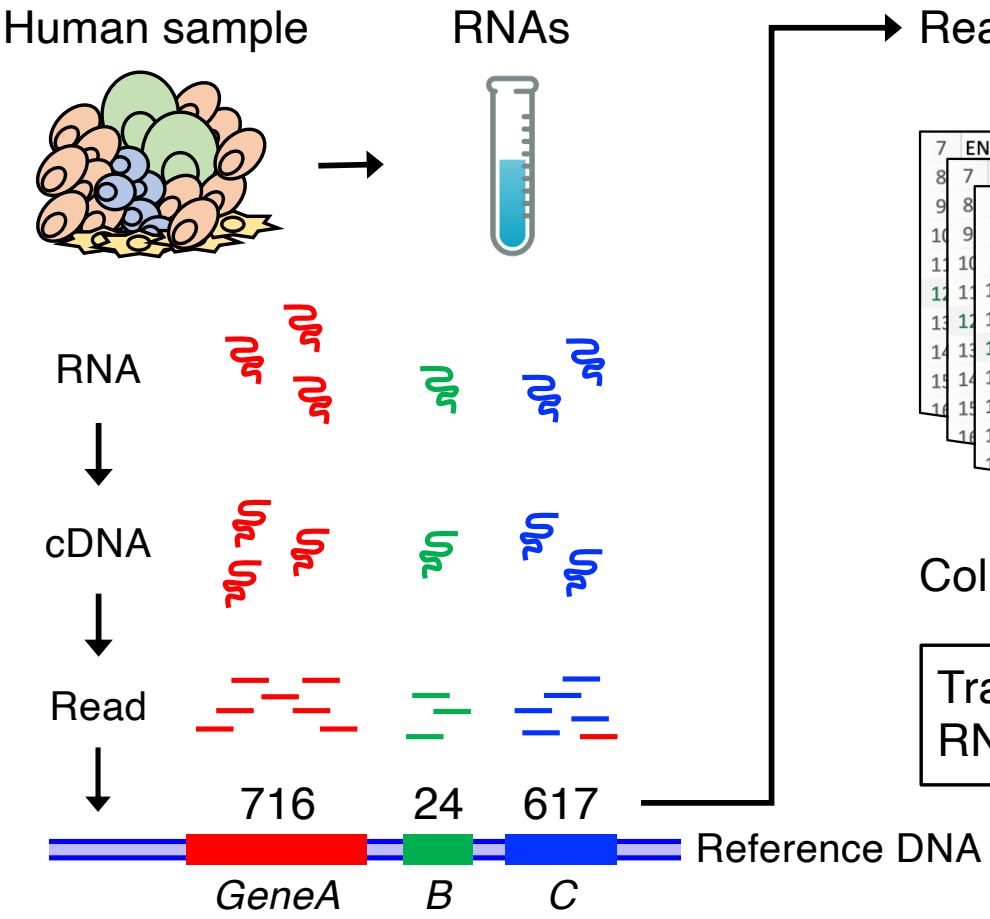
DNA sequence

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

Annotation info

https://asia.ensembl.org

バルクRNA-seq



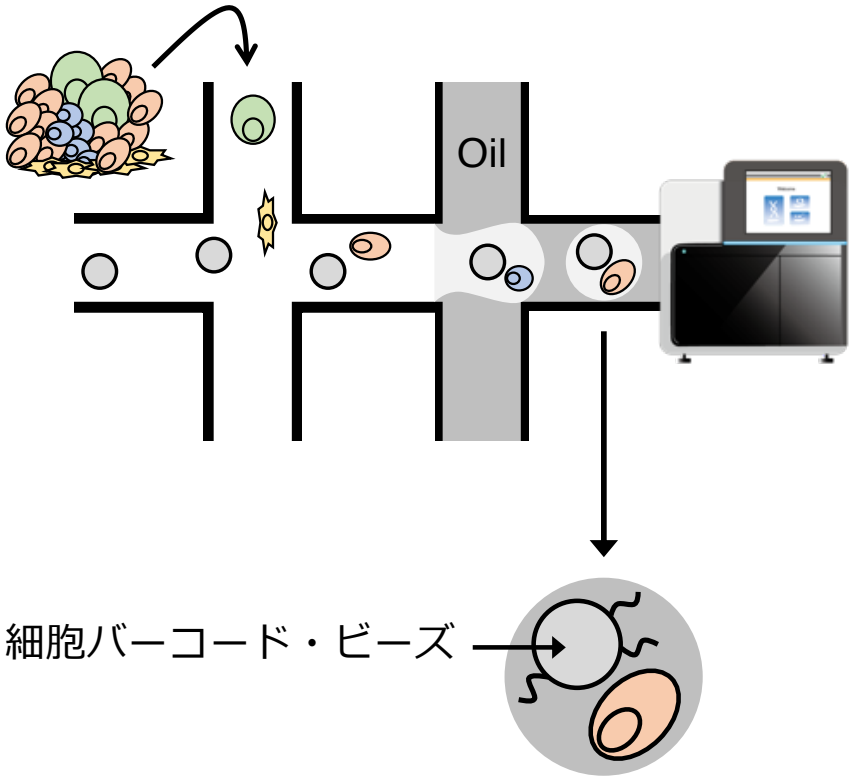
Read count table

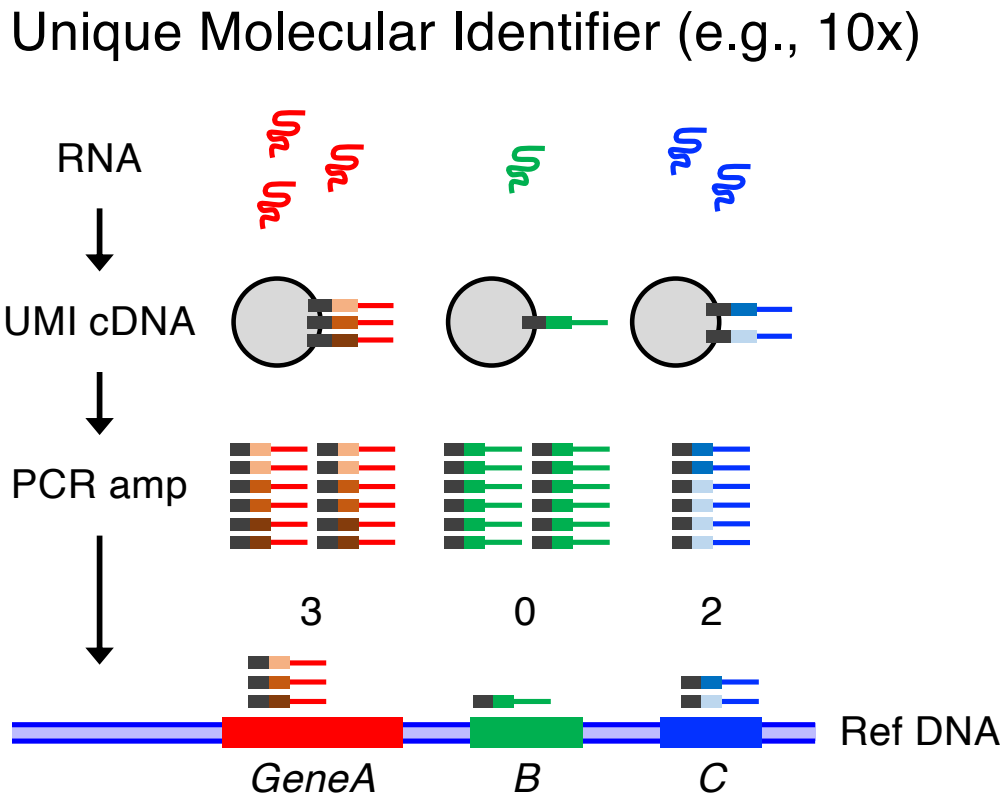
		Gene name		Expression level
7	ENSG00000000003.15	TSPAN6	protein_coding	716
7	ENSG00000000003.15	TSPAN6	protein_coding	892
7	ENSG00000000003.15	TSPAN6	protein_coding	551
7	ENSG00000000003.15	TSPAN6	protein_coding	340
8	ENSG00000000005.6	TNMD	protein_coding	24
9	ENSG00000000419.13	DPM1	protein_coding	617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

Collecting samples produces multivariate data

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

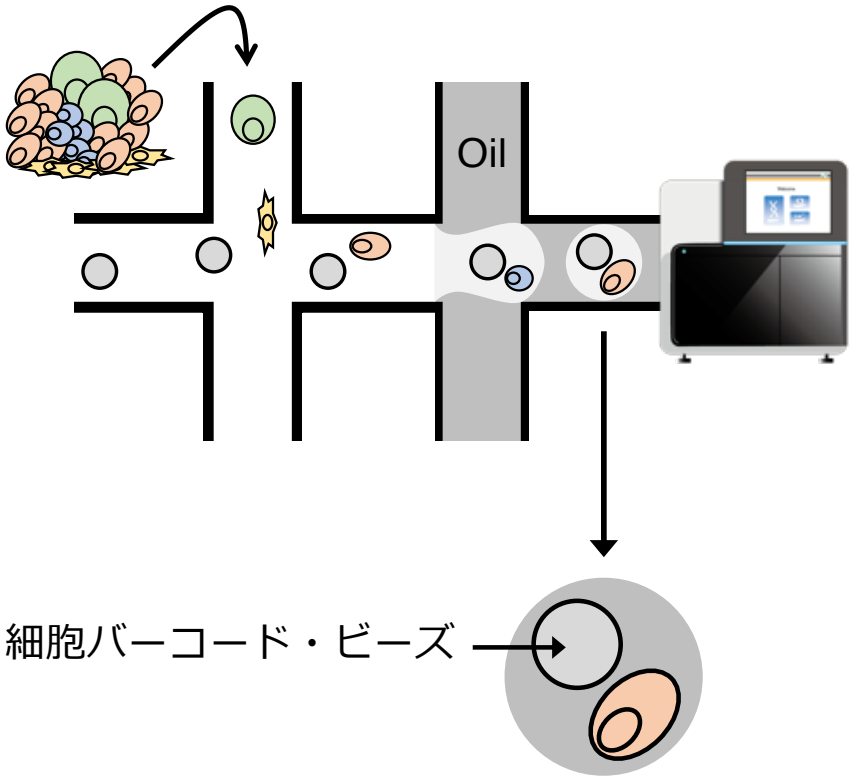
シングルセル RNA-seq (scRNA-seq)

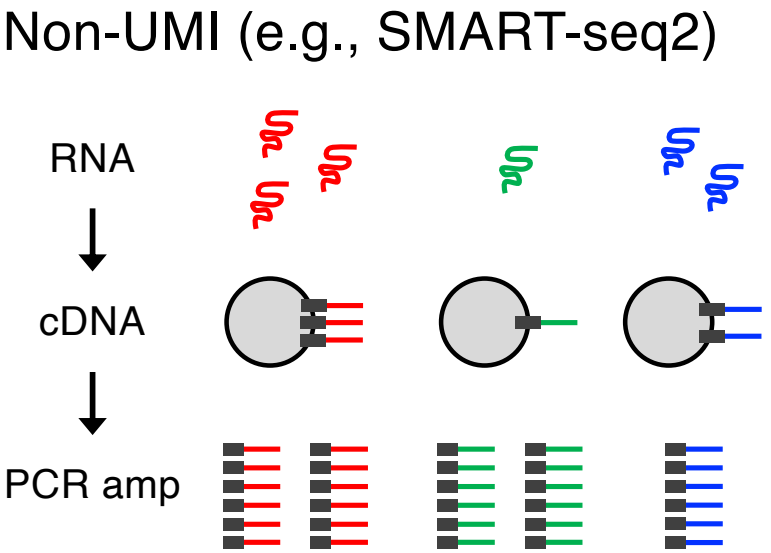




画像の出展: https://togotv.dbcls.jp

シングルセル RNA-seq (scRNA-seq)

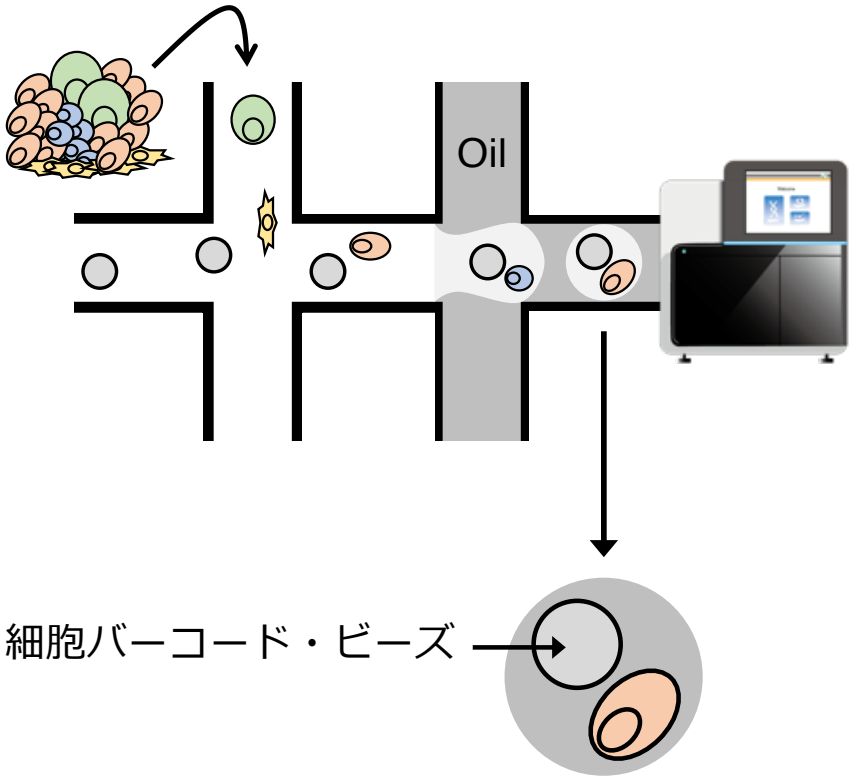




Non-UMI では、PCRによる増幅バイアスが生じうる

画像の出展: https://togotv.dbcls.jp

シングルセル RNA-seq (scRNA-seq)



Read count table (single-cell version)

Cells

Genes

GENE	AAACCTGAGACGCTTT-1	AAACCTGAGCAAT
RP11-34P13.7	0	0
RP11-34P13.8	0	0
RP11-34P13.9	0	0
FO538757.3	0	0
FO538757.2	0	0
AP006222.2	0	0
RP4-669L17.10	0	0
RP5-857K21.4	0	0
RP11-206L10.9	0	0

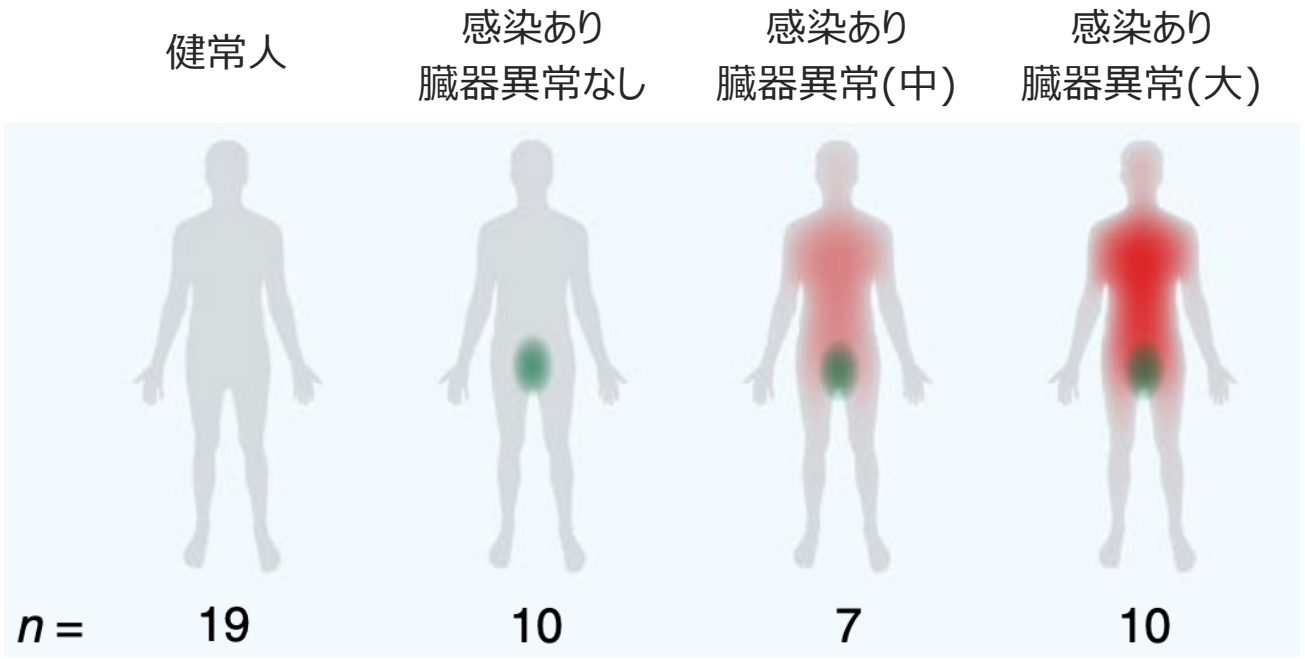
画像の出展: https://togotv.dbcls.jp

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

敗血症のシングルセル解析

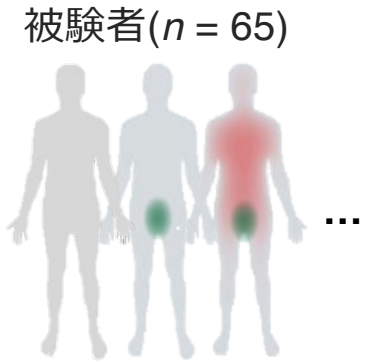
細菌感染性の敗血症（炎症 → 多臓器不全、非常に高い死亡率）



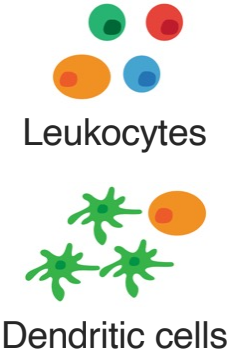
……

Reyes et al., Nat Med, 2020

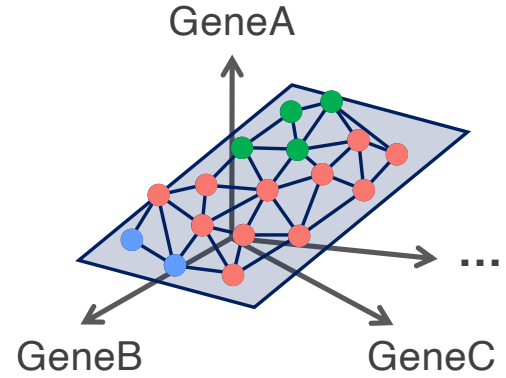
標準的なアプローチ

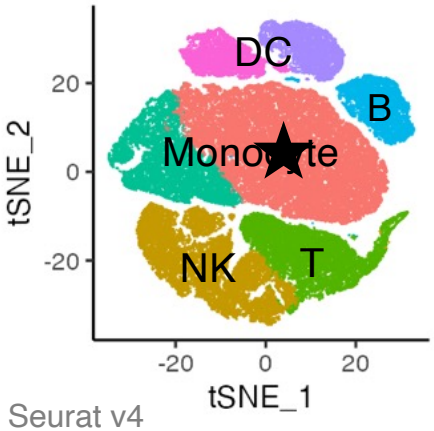


Reyes et al., Nat Med, 2020

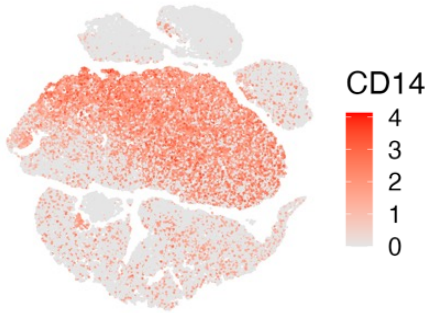



```
GENE,AAACCTGAGACGCTTT-1,AAACCTGAGCAAT
RP11-34P13.7,0,0,0,0,0,0,0,0,0,0,0,0,
RP11-34P13.8,0,0,0,0,0,0,0,0,0,0,0,0,
RP11-34P13.9,0,0,0,0,0,0,0,0,0,0,0,0,
FO538757.3,0,0,0,0,0,0,0,0,0,0,0,0,
FO538757.2,0,0,0,0,0,0,0,0,0,0,0,0,
AP006222.2,0,0,0,0,0,0,0,0,0,0,0,0,
RP4-669L17.10,0,0,0,0,0,0,0,0,0,0,0,0
RP5-857K21.4,0,0,0,0,0,0,0,0,0,0,0,0
RP11-206L10.9,0,0,0,0,
```





Seurat v4



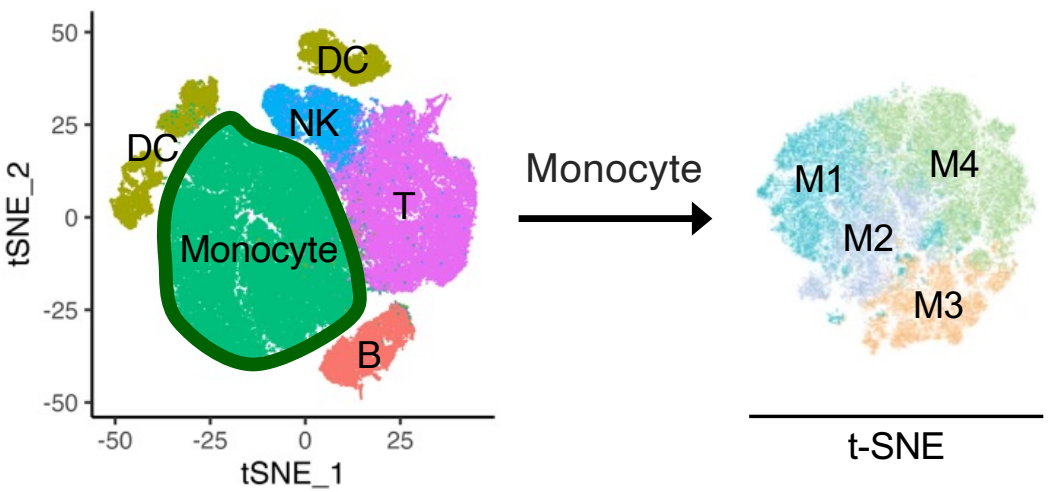
クラスター★で発現量が有意に高い遺伝子は361個もある ($p_{adj} < 10^{-100}$)

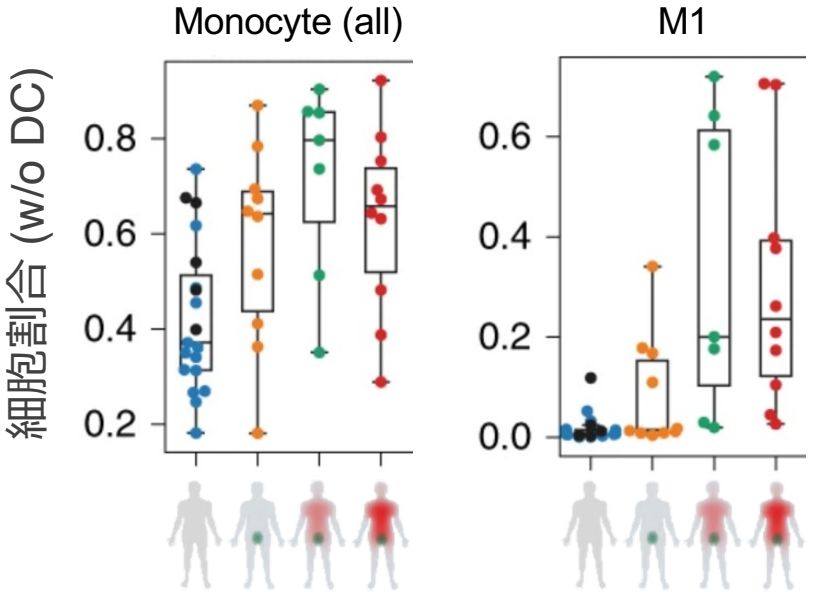
⟶ 文献調査

クラスタリング結果の生物学的解釈は常に問題となる

標準的なアプローチ

Scanpy を使用したクラスタリング





Reyes et al., Nat Med, 2020

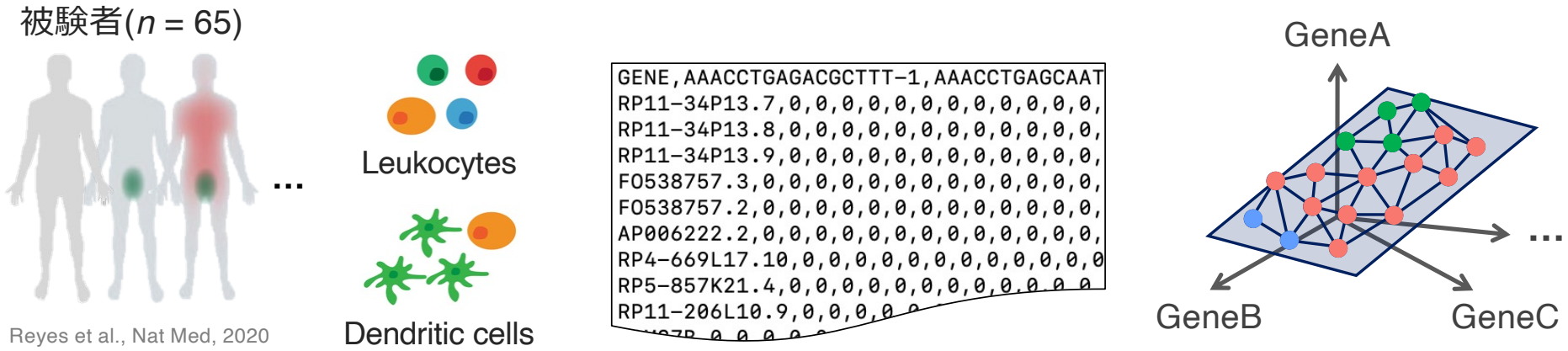
結論：単球の亜集団 M1 は敗血症になると急増する

問題点：クラスタリングの信頼性について疑問が残る

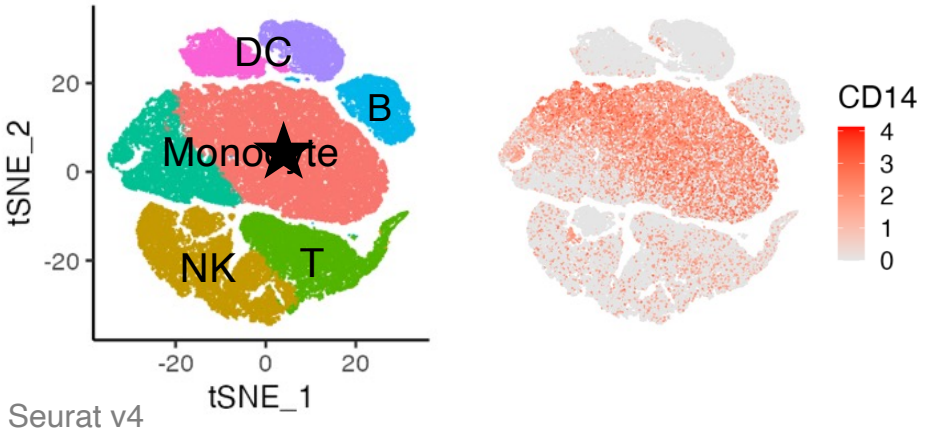
アジェンダ

- 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

再掲：標準的なアプローチ



Reyes et al., Nat Med, 2020



Seurat v4

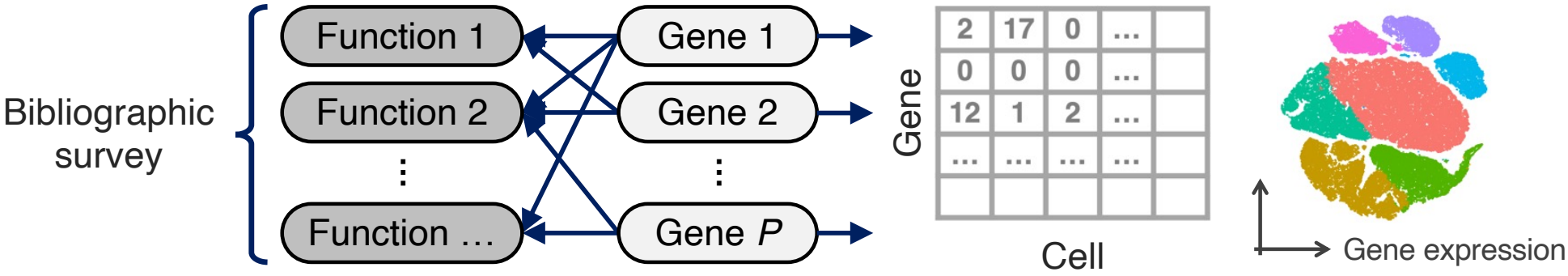
クラスター★で発現量が有意に高い遺伝子は361個もある ($p_{adj} < 10^{-100}$)

⟶ 文献調査

クラスタリング結果の生物学的解釈は常に問題となる

「生物学的解釈」の問題

従来のアプローチ

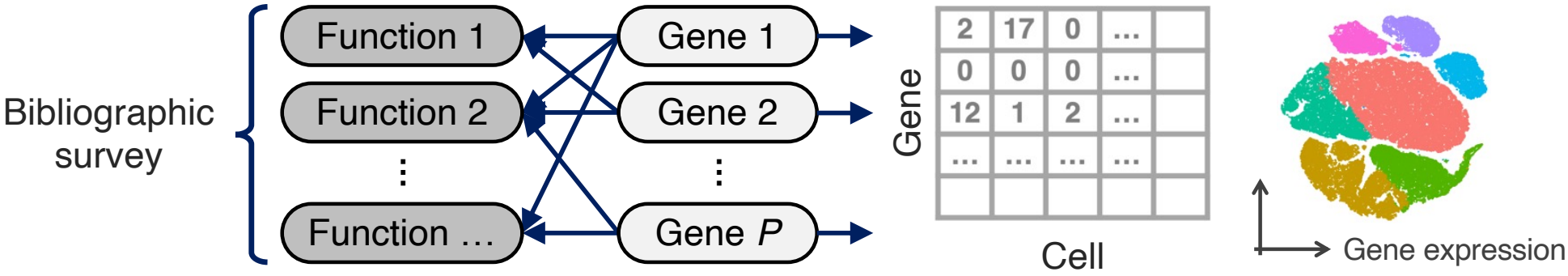


超簡単だが、クラスタリング結果の **生物学的解釈** が常に問題となる

ジレンマ：最適なクラスタリングとは何か？ (情報科学 ≠ 生物学)

「生物学的解釈」を数学的に定義する

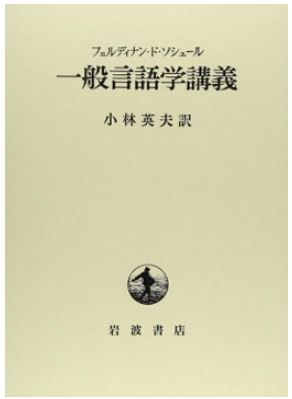
従来のアプローチ

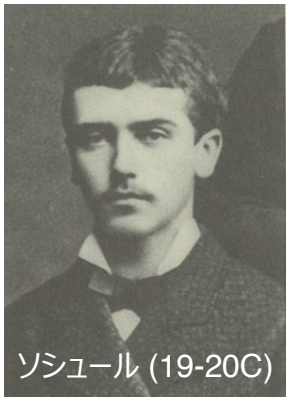


「生物学的解釈」を数学的に定義したい

「解釈」≒「意味づけ」

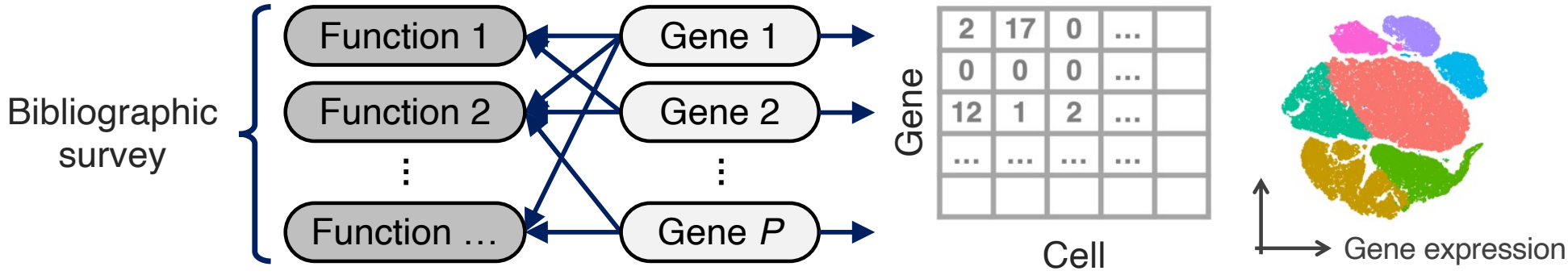
「意味」とは？ → **記号学**



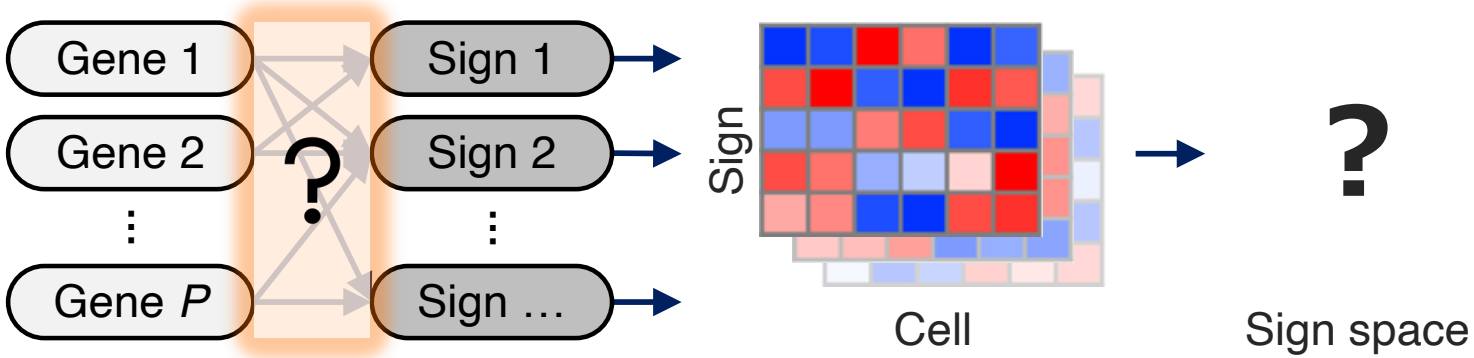


記号学的アプローチ

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/CI
GO:0000003	reproduction	4.03787500459501	1503	ABAT/ACR/ACOX1/ACVR1/ACVR1B/ACVR2A/ADA/ADAI 2/MIR21/ASTL/LHX8/LRRC52/RIMBP3B/PIWIL3/SPATA31(
GO:0002376	immune system process	2.78567002936747	3287	A1BG/A2M/SERPINA3/ABL1/AOC1/ABR/ACAA1/ACLY/AC H/MR1/HLX/HMGB1/HMGB2/HMGB3/HMGN2/HMOX1/HI 5/IGHV7-81/IGHV6-1/IGHV5-10-1/IGHV5-51/IGHV4-38-2/IG

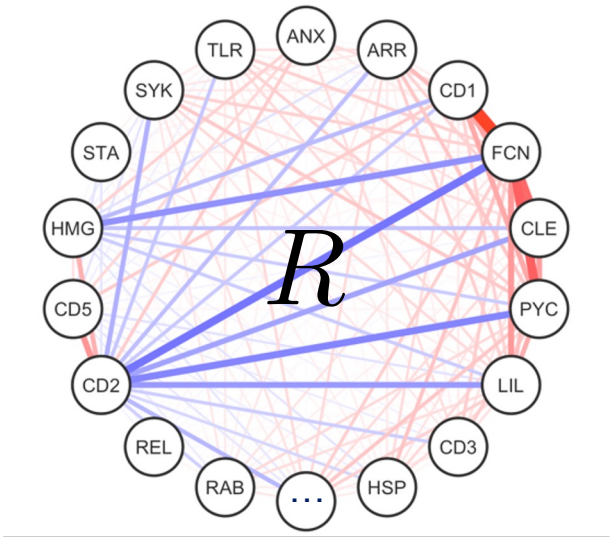
2020年12月時点、44272個の生物機能に関するIDが登録されている

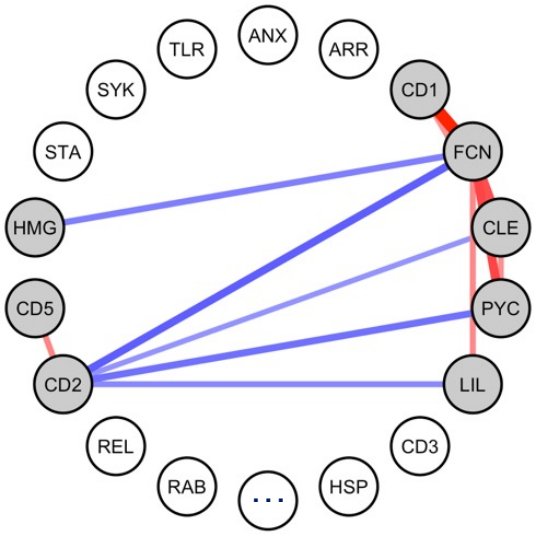
「記号」の定義

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

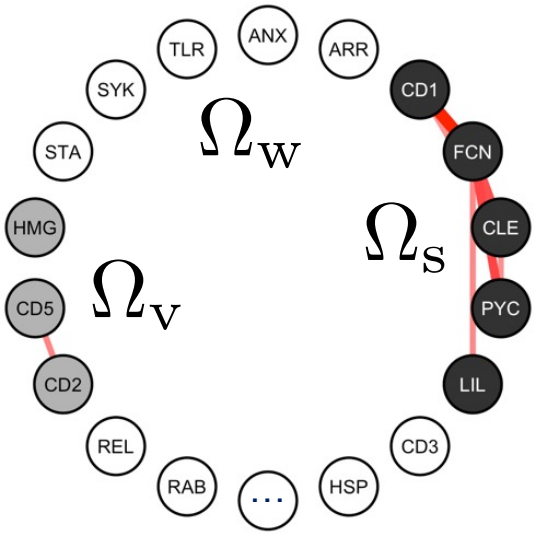
Ω = Gene set

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

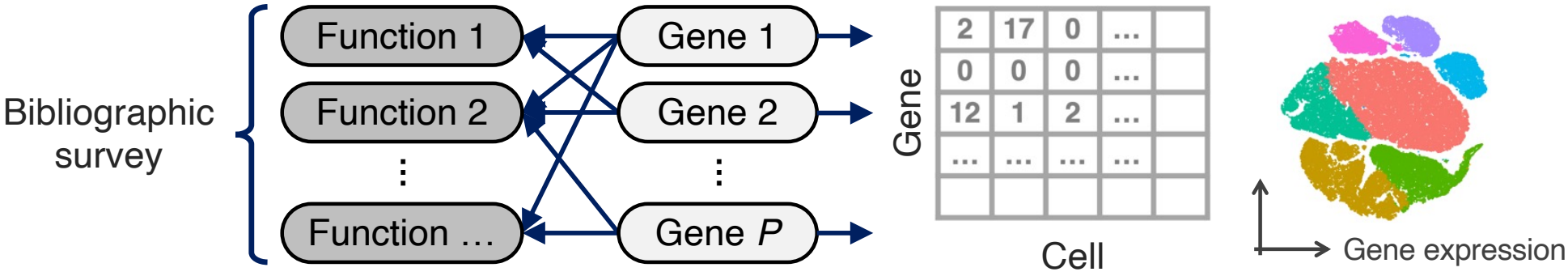
...

	SignID	Description	IC	StrgCorrGene	VariCorrGene	WeakCorrGene
1	GO:0000003	reproduction	4.037875	NA	NA	ACOX1/ADA/ADAM10/ADM...
2	GO:0007610	behavior	5.017167	NA	NA	ADA/AHCY/AMFR/APP/AP...
3	GO:0019740	nitrogen utilization	10.371865	NA	NA	BAX/BCL2/GLUL
4	GO:0022414	reproductive process	4.039296	NA	NA	ACOX1/ADA/ADAM10/ADM...
5	GO:0022610	biological adhesion	4.062674	NA	NA	ACTB/ACTG1/ACTN4/ACT...
6	GO:0040007	growth	4.854412	NA	NA	ACTL6A/ D1/A...
7	GO:0040011	locomotion	3.686254	NA	NA	AAMP/A CTN...
8	GO:0043473	pigmentation	6.893706	NA	NA	AP1G1/A CL2...
9	GO:0048511	rhythmic process	5.780794	NAMPT/PTGDS/MYCBP2/R...	HNRNPD/NONO/PRKDC/ET...	ADA/AHCY/AHR/ARRB1/A...
10	GO:0051703	intraspecies interac...	7.973970	NA	NA	CX3CR1/EIF4EBP2/MECP...
11	GO:0051704	multi-organism proce...	4.575199	NA	NA	ACOX1/ADA/ADAM10/ADM...
12	GO:0098754	detoxification	6.950865	S100A9/GPX1/SOD2/CD3...	ADH5/PIM1/TXN/ALOX5A	
13	GO:0110148	biomineralization	6.610665	NA		
		sexual reproduction	4.789873			

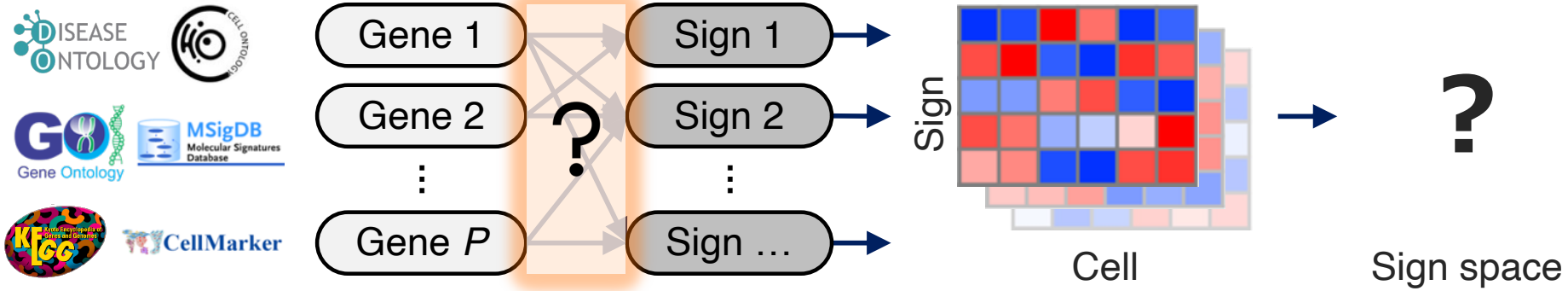
Ω_S Ω_V Ω_W

「生物学的解釈」を数学的に定義する

従来のアプローチ

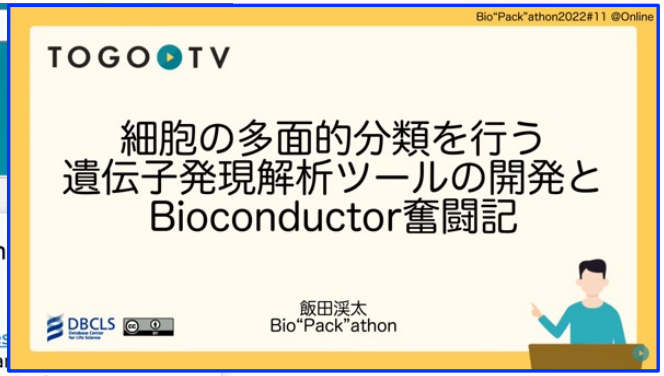


記号学的アプローチ



ASURAT(阿修羅)の開発

YouTubeからも公開



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers

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

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 at protein.osaka-u.ac.jp>

Documentation

Bioconductor

- Package vignettes
- Workflows for lea
- 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

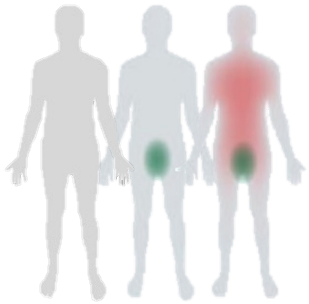
Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about

ASURATによる敗血症のシングルセル解析

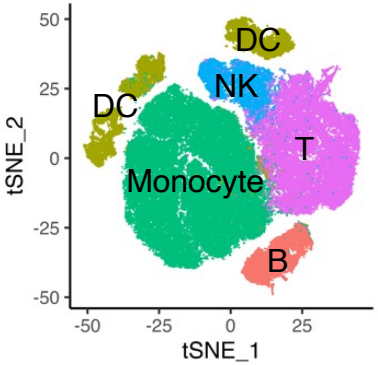
被験者(n = 65)



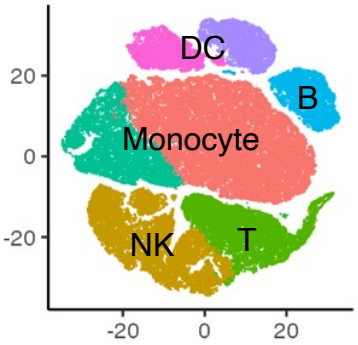
...

Reyes et al., Nat Med (2020)

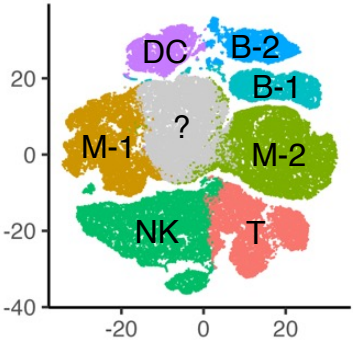
Scanpy



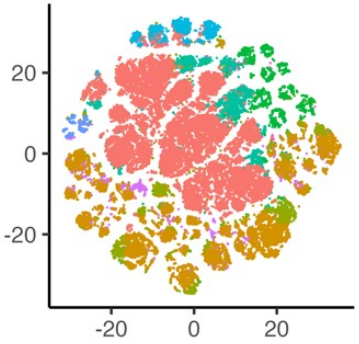
Seurat



scran



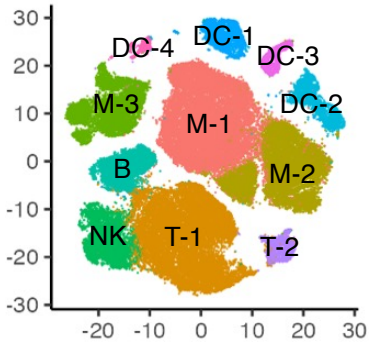
ssGSEA



ASURAT

細胞型のデータベース

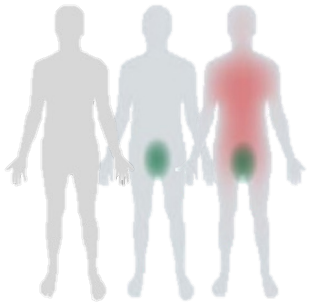
細胞型の記号空間



Iida et al., ***Bioinformatics***, 2022

ASURATによる敗血症のシングルセル解析

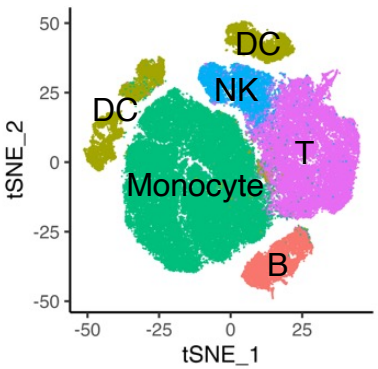
被験者(n = 65)



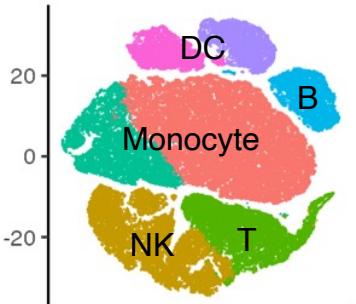
...

Reyes et al., Nat Med (2020)

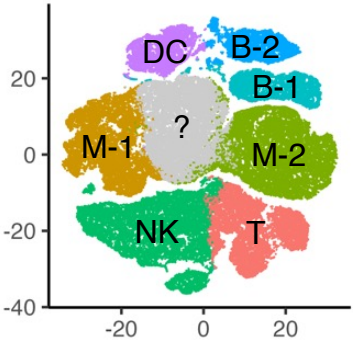
Scanpy



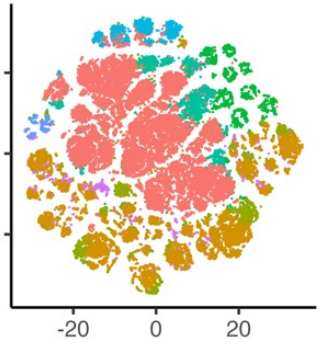
Seurat



scran



ssGSEA

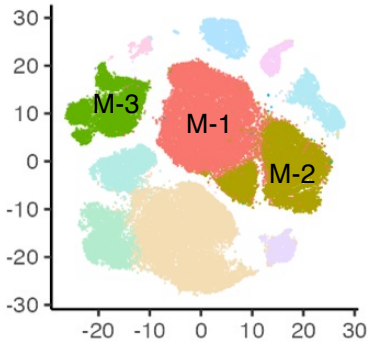


ASURAT

細胞型のデータベース

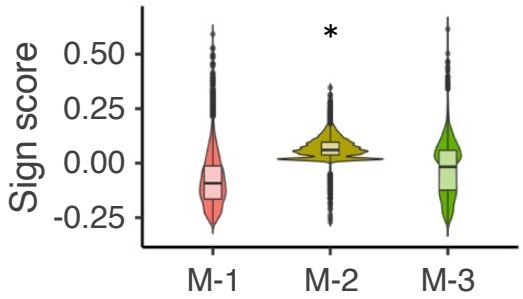
パスウェイのデータベース

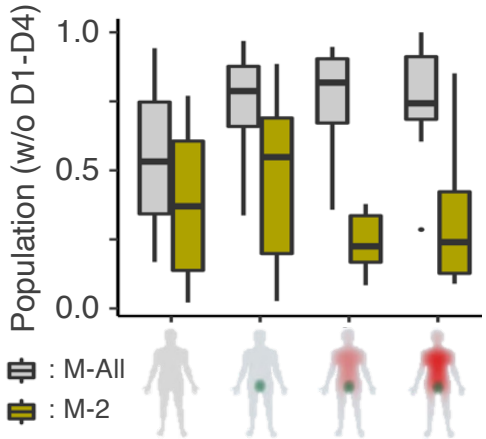
細胞型の記号空間



パスウェイの記号空間

脂肪酸の分解



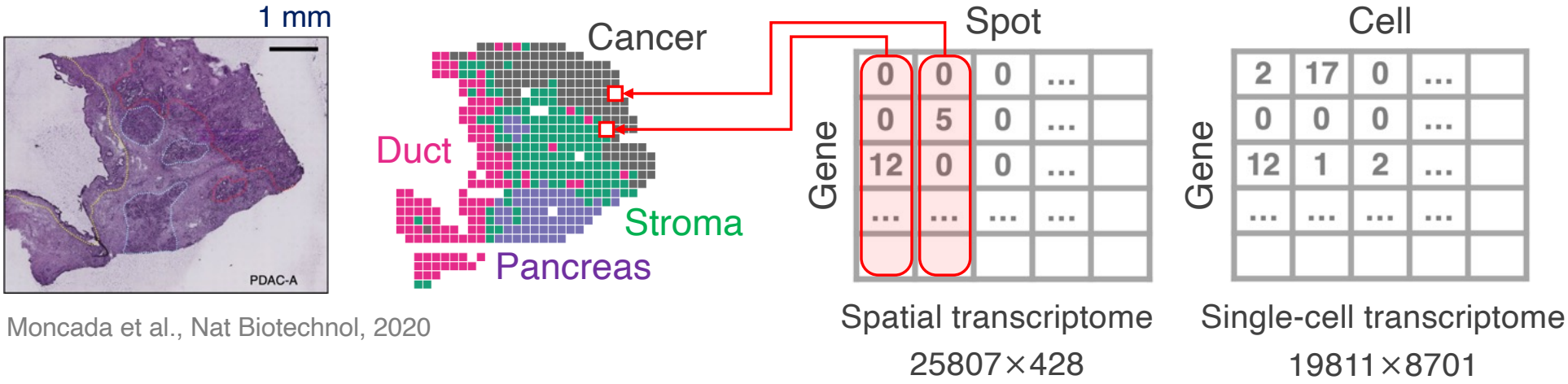


lida et al., ***Bioinformatics***, 2022

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

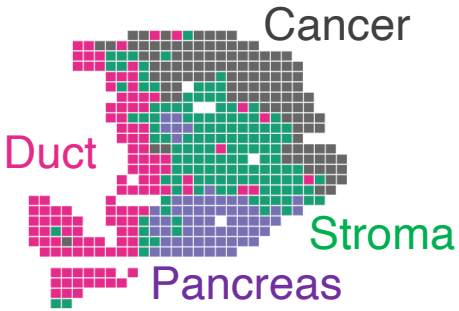
目的　１細胞・空間オミクスデータを用いて悪性細胞をみつける(細胞分類)



Moncada et al., Nat Biotechnol, 2020

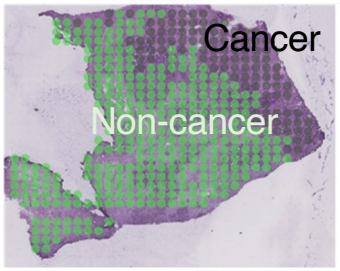
膵管腺癌の空間オミクス解析

Previous results



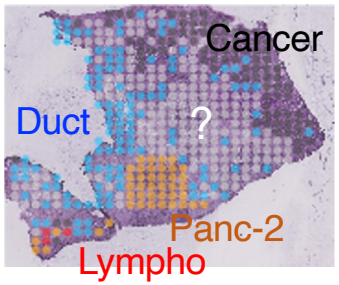
Moncada et al., 2020

Method: ssGSEA



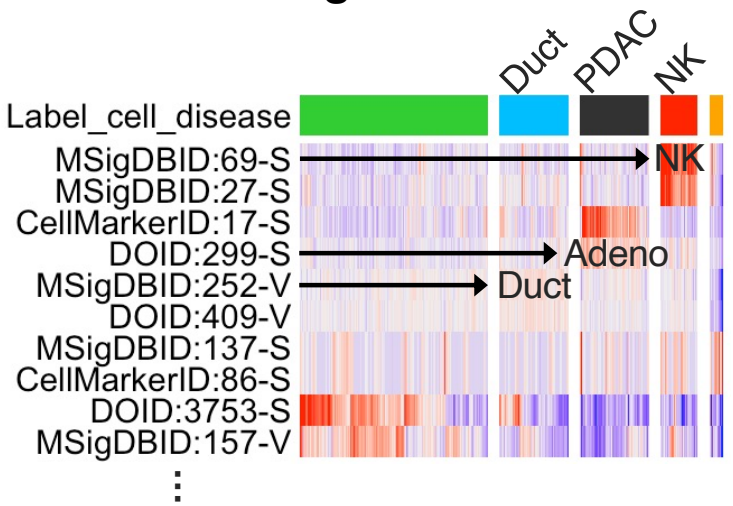
Subramanian et al., 2020

Method: PAGODA



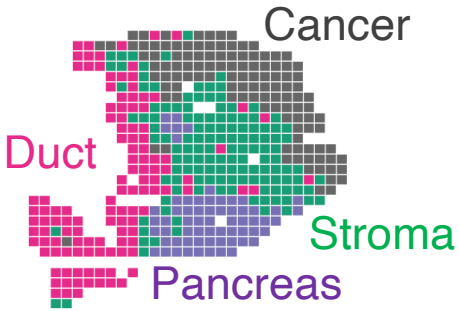
Fan et al., 2016

Our result using ASURAT



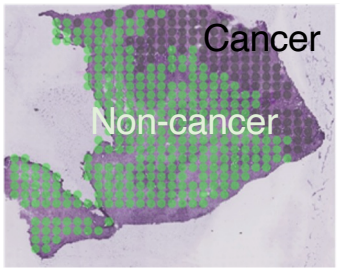
膵管腺癌の空間オミクス解析

Previous results



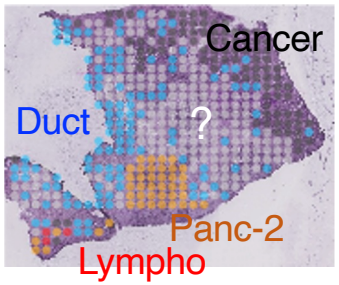
Moncada et al., 2020

Method: ssGSEA



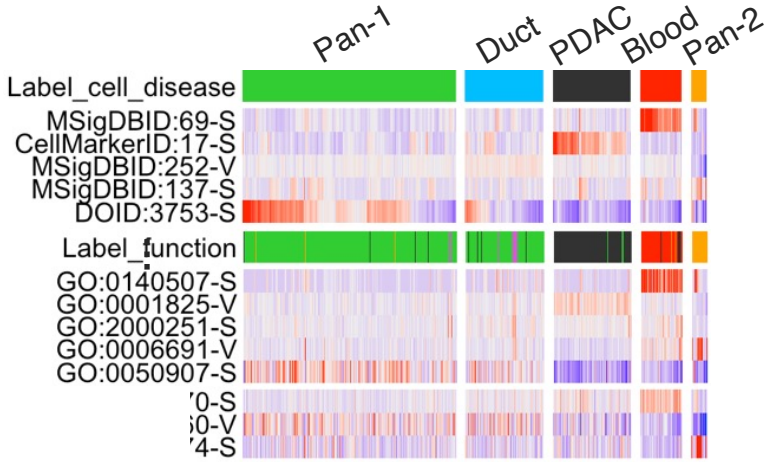
Subramanian et al., 2020

Method: PAGODA



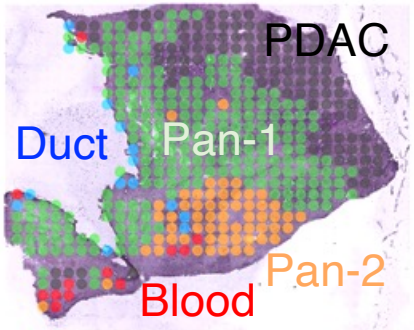
Fan et al., 2016

Our result using ASURAT

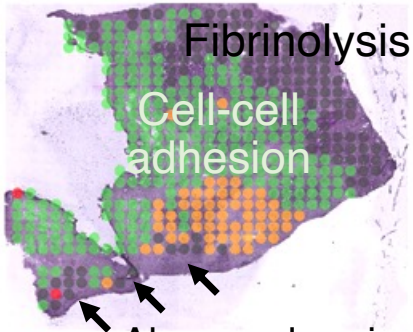


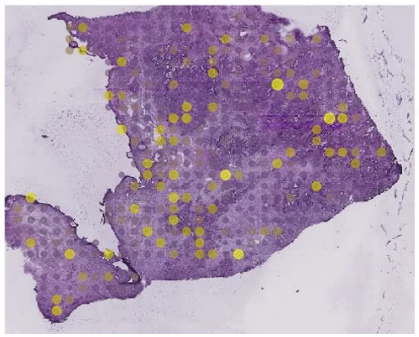
(KEGG:pathway) path:hsa00010-S
Glycolysis / Gluconeogenesis

Disease and cell type



Biological process

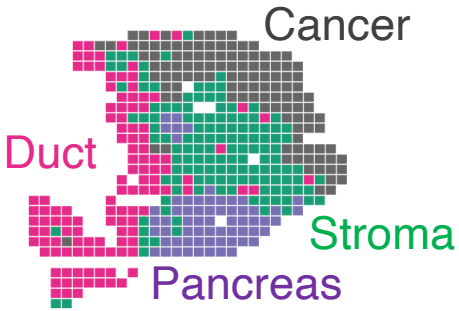




Iida* et al., ***Bioinformatics***, 2022

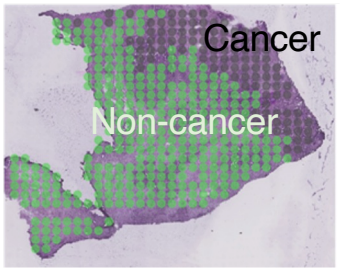
膵管腺癌の空間オミクス解析

Previous results



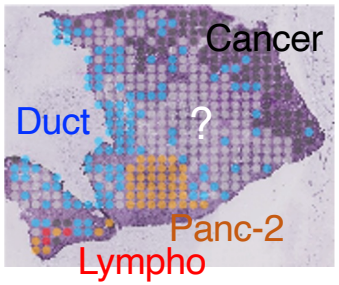
Moncada et al., 2020

Method: ssGSEA



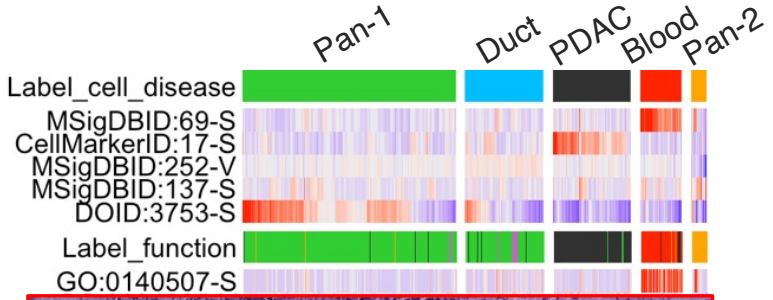
Subramanian et al., 2020

Method: PAGODA

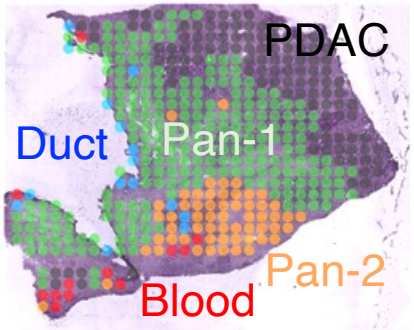


Fan et al., 2016

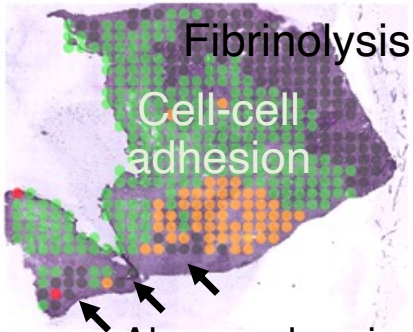
Our result using ASURAT



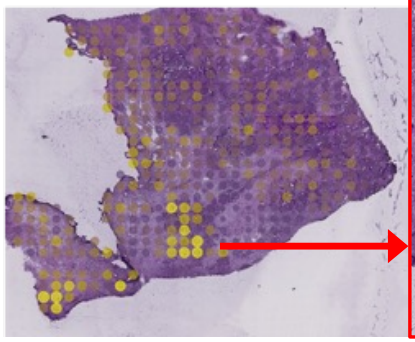
Disease and cell type

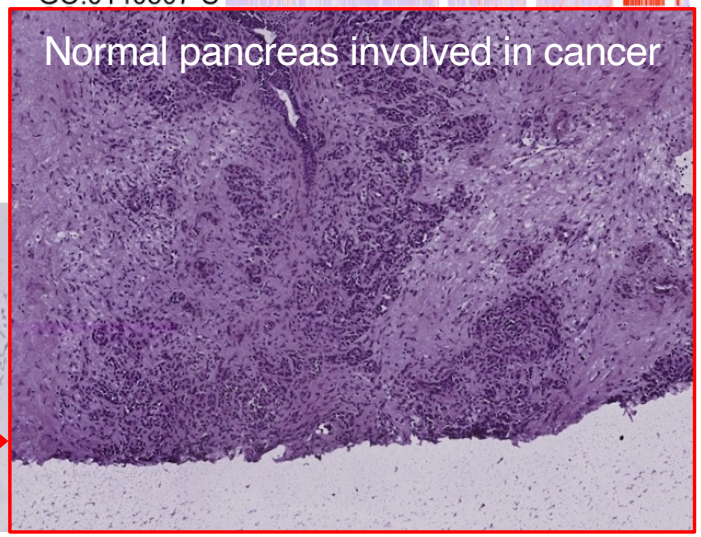


Biological process



Pathway: Th17 diff

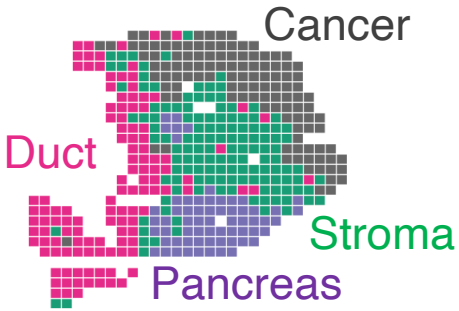




lida* et al., ***Bioinformatics***, 2022

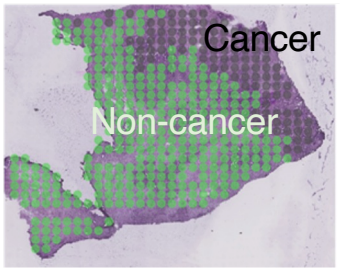
膵管腺癌の空間オミクス解析

Previous results



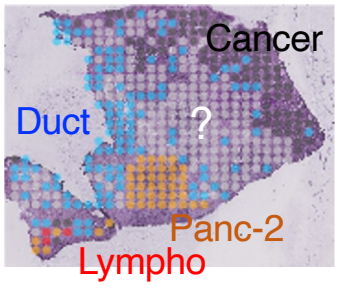
Moncada et al., 2020

Method: ssGSEA



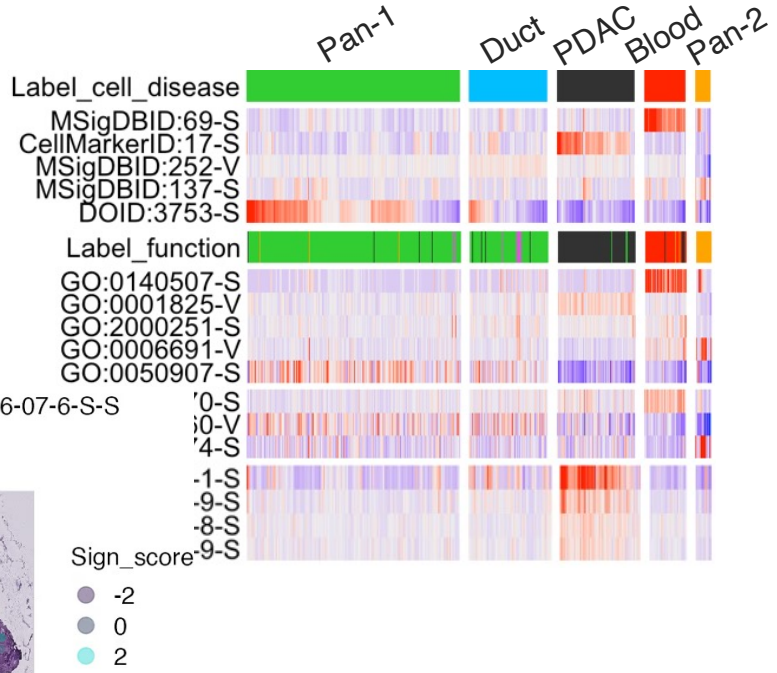
Subramanian et al., 2020

Method: PAGODA



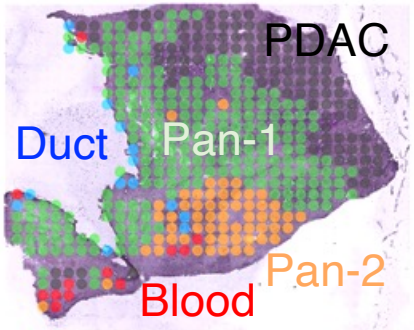
Fan et al., 2016

Our result using ASURAT

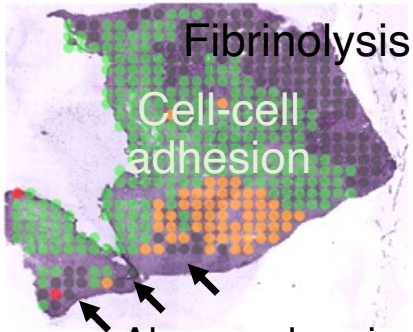


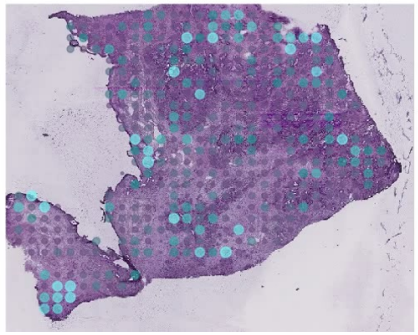
(Depmap:drug) BRD-A00077618-236-07-6-S-S
8-bromo-cGMP (sensitive)

Disease and cell type



Biological process

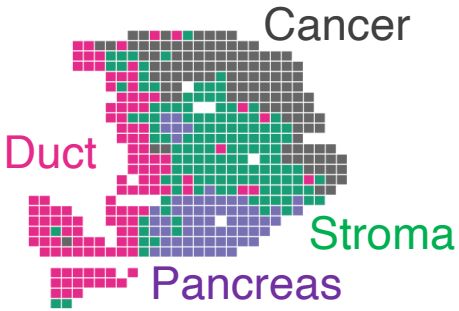




lida* et al., ***Bioinformatics***, 2022

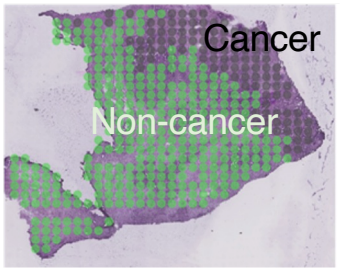
膵管腺癌の空間オミクス解析

Previous results



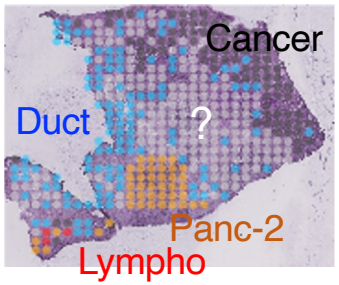
Moncada et al., 2020

Method: ssGSEA



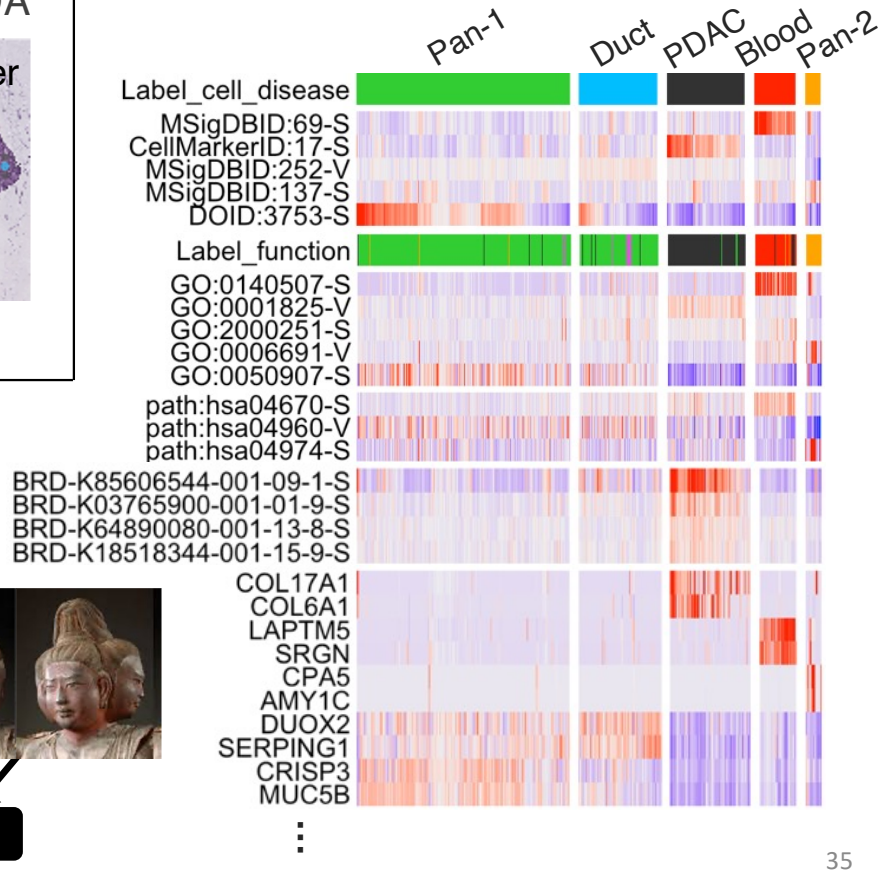
Subramanian et al., 2020

Method: PAGODA

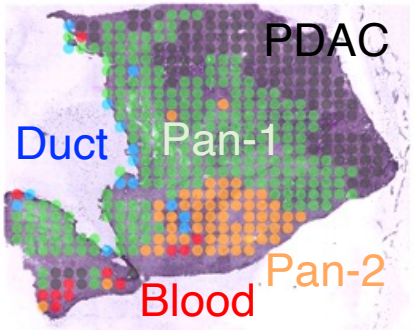


Fan et al., 2016

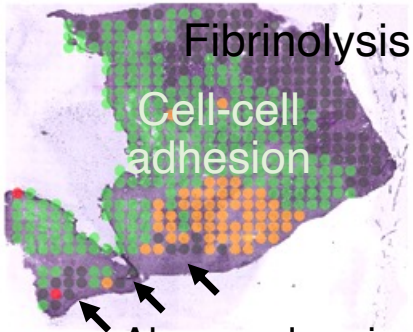
Our result using ASURAT



Disease and cell type

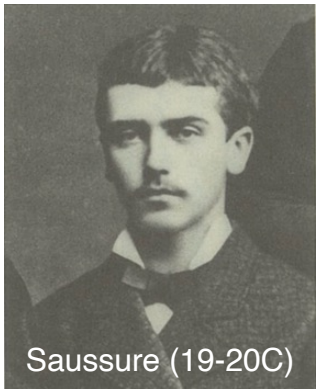


Biological process

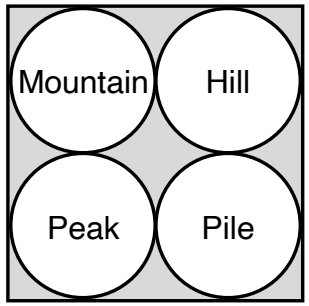


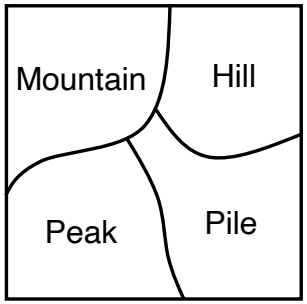
lida* et al., ***Bioinformatics***, 2022

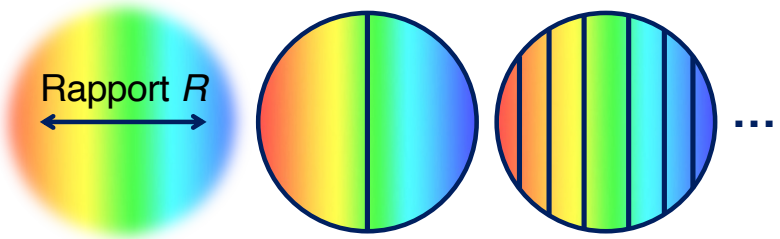
まとめ

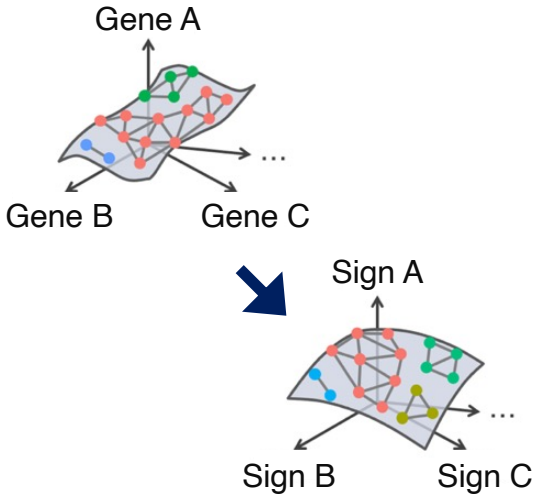


Nomenclaturism Semiology

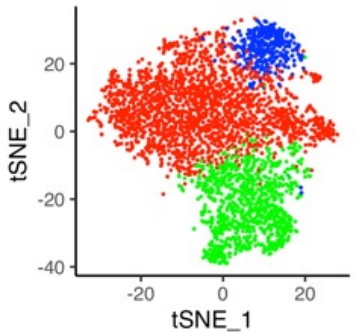




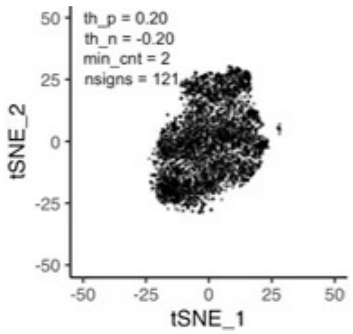




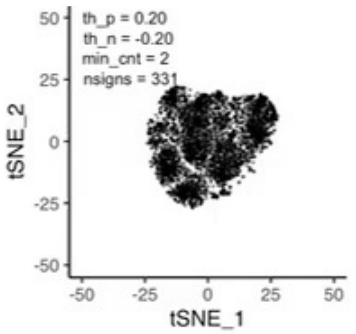
Cervical cancer data Masuda, Iida, et al., ***Am J Pathol***, 2023



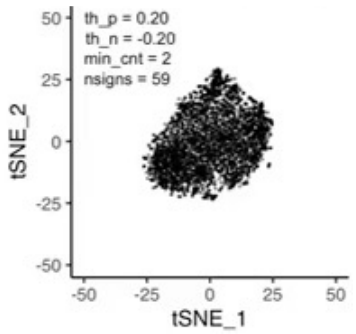
Expression space



Disease



Biological process



Signaling pathway

Any question is very welcomed