

ウェビナー

# Chromium Controller による シングルセル情報解析の実際



KOTAI バイオテクノロジーズ  
Nicolas Sax

# 会社概要と共同研究先

## ■ 会社概要

会社名	KOTAI バイオテクノロジーズ株式会社
設立	2016年5月
代表取締役	山下和男
事業内容	免疫 AI プラットフォーム提供、 及び、それに関わるデータ解析等
所在地	大阪府吹田市山田丘 2-1 大阪大学産学共創 B 棟 2 階
資本金	252,320 千円
従業員数	12名（うち、研究開発 8名。2020年1月現在）
投資家	大阪大学ベンチャーキャピタル ニッセイキャピタル 信金キャピタル みなどキャピタル 池田泉州キャピタル
助成金	NEDO 「平成 29 年度 AI システム共同開発支援事業」 中小企業庁 「平成 30 年度 戦略的基盤技術高度化支援事業」 AMED 「エイズ対策実用化研究事業」「感染症実用化研究事業」「次世代がん医療創生研究事業」「革新的がん医療実用化研究事業」

## ■ 共同研究先



国内トップ研究機関と研究開発ネットワークを構築

# 研究開発チーム

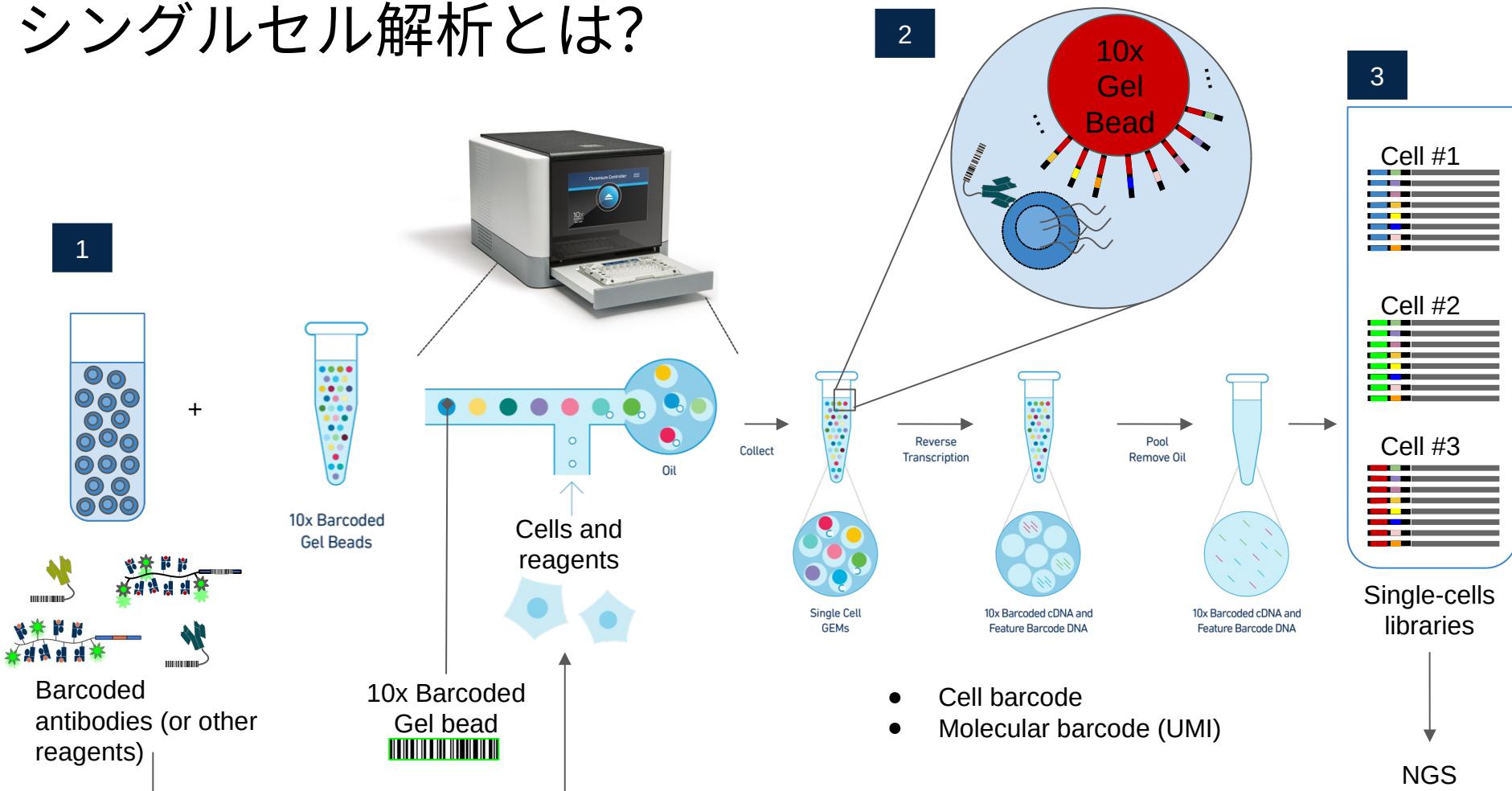


## 生物情報科学の専門家集団と最先端のビッグデータ実験基盤

研究開発部門外国人割合 62.5%



# シングルセル解析とは？



# シングルセル遺伝子発現： mRNA 量 vs タンパク質量

4

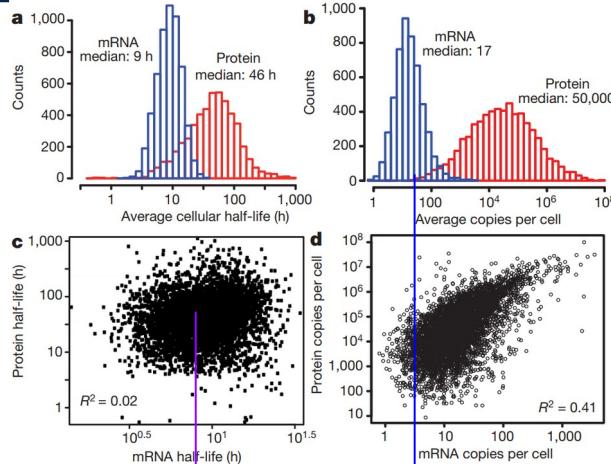
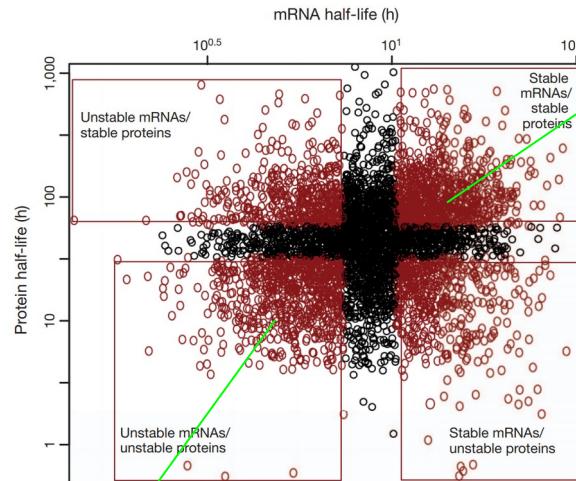


Figure 2 | mRNA and protein levels and half-lives. a, b, Histograms of mRNA (blue) and protein (red) half-lives (a) and levels (b). Proteins were on average 5 times more stable and 2,800 times more abundant than mRNAs and spanned a higher dynamic range. c, d, Although mRNA and protein levels correlated significantly, correlation of half-lives was virtually absent.

mRNA の安定性とタンパク質の安定性が全く相関しない ⇒  
FACS でこれまで使用されたマーカーが最適でない確率が高い：新たなマーカー遺伝子を使わないと行けない。

1 細胞当たりの mRNA の分子数はタンパク質の 1 万分の一  
⇒ 全ての遺伝子を全ての細胞で検出できることが難しい

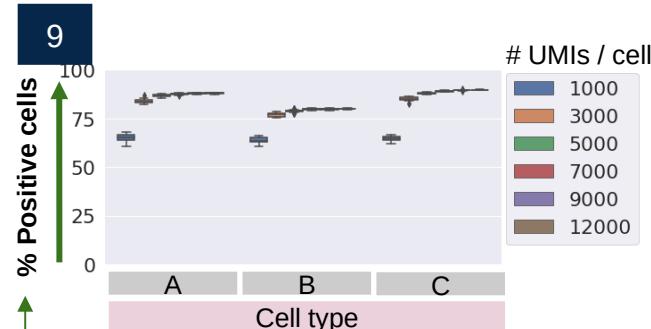
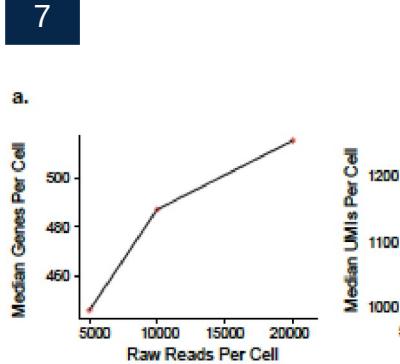
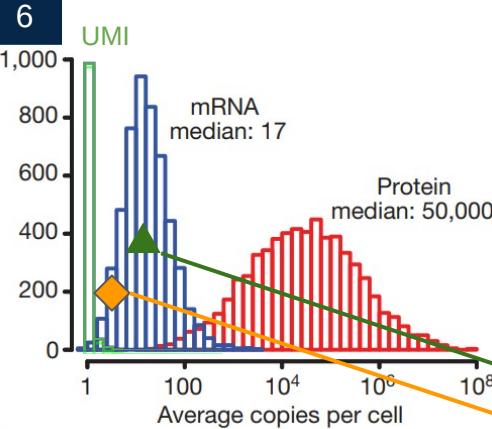
5



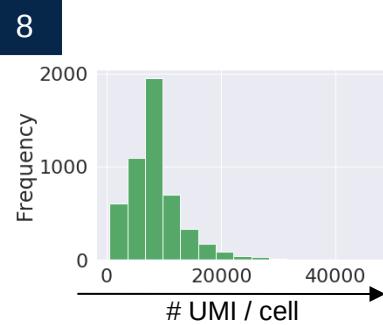
Source:  
Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., ... Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347), 337–342.  
<https://doi.org/10.1038/nature10098>

scRNA-seq should be considered a sampling. The underlying distribution can be mathematically modelled.

# シングルセル遺伝子発現： mRNA 量 vs UMI 数

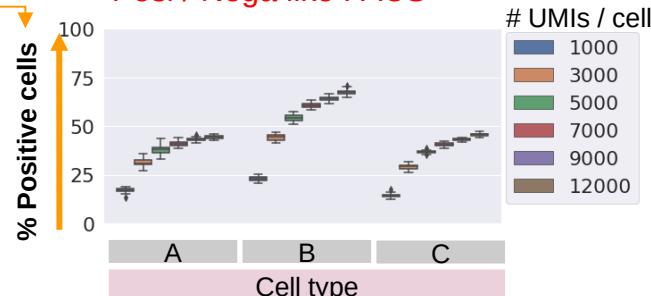


Detection can be UMI dependent → Should not think Posi / Nega like FACS



High heterogeneity and noise but also High throughput and high number of genes (10~30k genes per cell, vs 5~50 parameters for FACS/CyTOF)

Using 10x 5k PBMC (v3 chemistry)

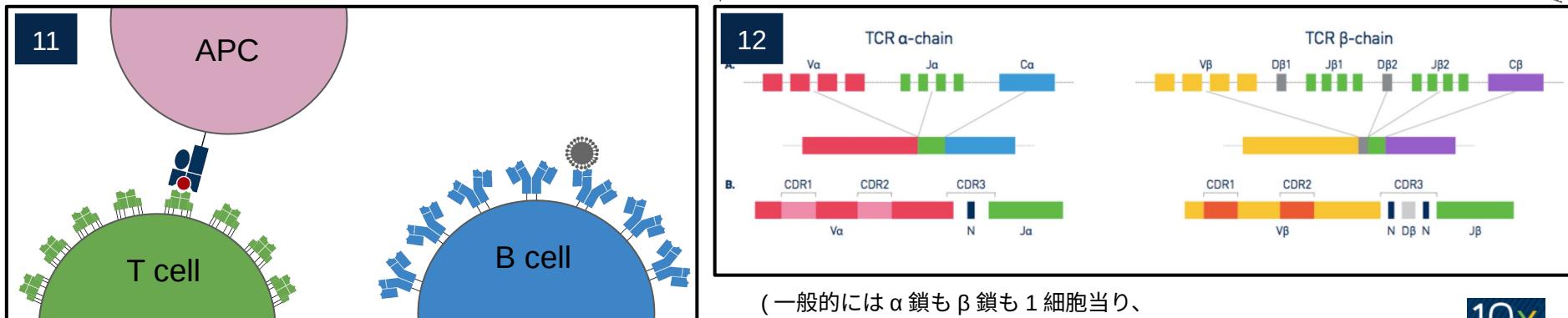
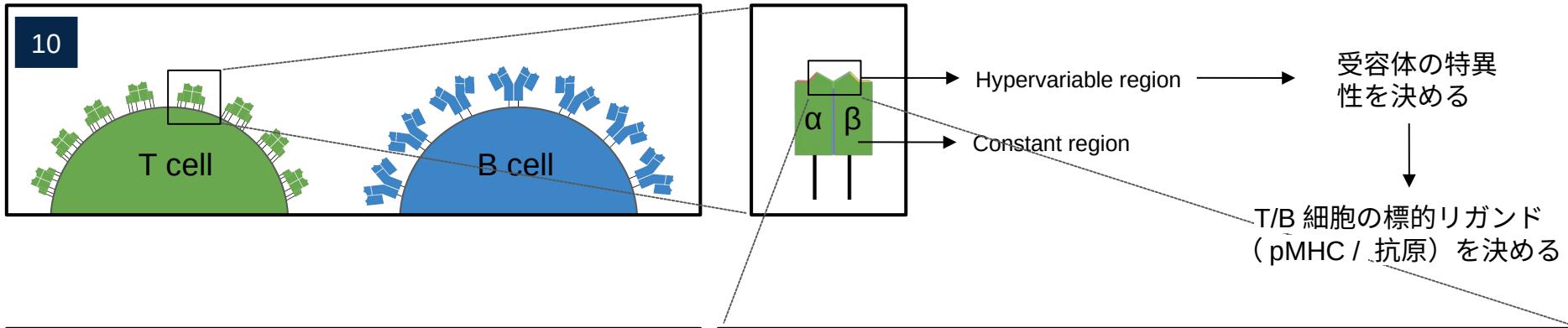


Sources: \* Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., ... Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1), 14049. <https://doi.org/10.1038/ncomms14049> Supp. Fig. 5

\* 5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor (v3 chemistry) [https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.2/5k\\_pbmc\\_v3](https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.2/5k_pbmc_v3)

# T・B 細胞受容体シークエンシング: V(D)J 遺伝子再構成

B and T cells express B/T-cell receptors, generated by an almost random recombination process resulting in a very high sequence diversity.



VDJ Figures are from

10X  
GENOMICS

# T・B 細胞受容体シークエンシング：レパトア解析

13

Bulk

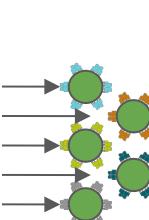
# Reads	Vgene	aaSeqCDR3	Jgene
77	TRBV29-1	CSEVGLPQQGTGEQYF	TRBJ2-7
72	TRBV4-1	CASSQDPFQTGETQYF	TRBJ2-5
60	TRBV15	CATSRGPQGTGSRSTEAFF	TRBJ1-1
52	TRBV4-1	CASSSLPGQQGPPGANVLTF	TRBJ2-6
37	TRBV4-1	CASSGDGTYEQYF	TRBJ2-7
30	TRBV6-2	CASSYGSNEQFF	TRBJ2-1
28	TRBV11-3	CASSLGTGETDTQYF	TRBJ2-3
26	TRBV5-1	CASSPGAPVGGARYNEQFF	TRBJ2-1
26	TRBV11-2	CASSLGGQGSFFTEAFF	TRBJ1-1
26	TRBV5-5	CASSRDSGRRLRSPLHF	TRBJ1-6
25	TRBV7-2	CASSLDSGTRYYEQYF	TRBJ2-7
23	TRBV27	CASRTGGVNEQYF	TRBJ2-7

Higher throughput and lower cost,  
but usually only beta / heavy chain.  
If both chains, no pairing  
information

14

Single-cell

Cell barcode	Vgene.alpha	aaSeqCDR3.alpha	Jgene.alpha	Vgene.beta	aaSeqCDR3.beta	Jgene.beta
TGGGAAGAGCGTTCGG	TRA9-2	CALSYQAGALTIF	TRAJ15	TRBV4-1	CASSGGVVEAFF	TRBJ1-1
CAGCATACAGGACGTA	TRA1-2	CASYNTDKLIF	TRAJ34	TRBV7-9	CASSLLAGGPYNEQFF	TRBJ2-1
CTGCCTAGTCGCTGGT	TRA17	CATAYSGNTPLVF	TRAJ29	TRBV20-1	CSARDPQGLSYEQYF	TRBJ2-7
ACTATCTCAGCATGAG	TRA1-2	CASMYSNYQLIW	TRAJ33	TRBV6-5	CASSPASNQPQHF	TRBJ1-5
CTCGTACCATACGCTA	TRA4	CLVGDGFSASKIIF	TRAJ3	TRBV7-2	CASSPGQAINEKLF	TRBJ1-4
GTCAAGTGTCCAGTGC	TRA1-2	CAVTLSGNTPLVF	TRAJ29	TRBV24-1	CATSDSRRAGDYEQYF	TRBJ2-7
TTTACTGAGCGTTAC	TRA23/DV6	CATLTYQGGSEKLVF	TRAJ57	TRBV7-2	CASSLATAMNTEAFF	TRBJ1-1
GTGCATAAGAGACGAA	TRA26-2	CILRGSGGSNYKLTF	TRAJ53	TRBV28	CASSHVRVLRSPLHF	TRBJ1-6
GGCTGGTCAGCGTTCG	TRA26-1	CIAPILNTGTASKLTF	TRAJ44	TRBV7-2	CASSLAVLGAYEQYF	TRBJ2-7
CTTAGGACACGACATCG	TRA13-1	CAASFSGGYNKLIF	TRAJ4	TRBV5-1	CASSLAPRGDYEQYF	TRBJ2-7
TGCCAAACAGACGCAA	TRA21	CAVTGGGNKLTF	TRAJ10	TRBV20-1	CSATKDGYSYNEQFF	TRBJ2-1
TCGGTAACAGTTCATG	TRA29/DV5	CAAIYSGGGADGLTF	TRAJ45	TRBV4-3	CASSQDRLARDEQYF	TRBJ2-7



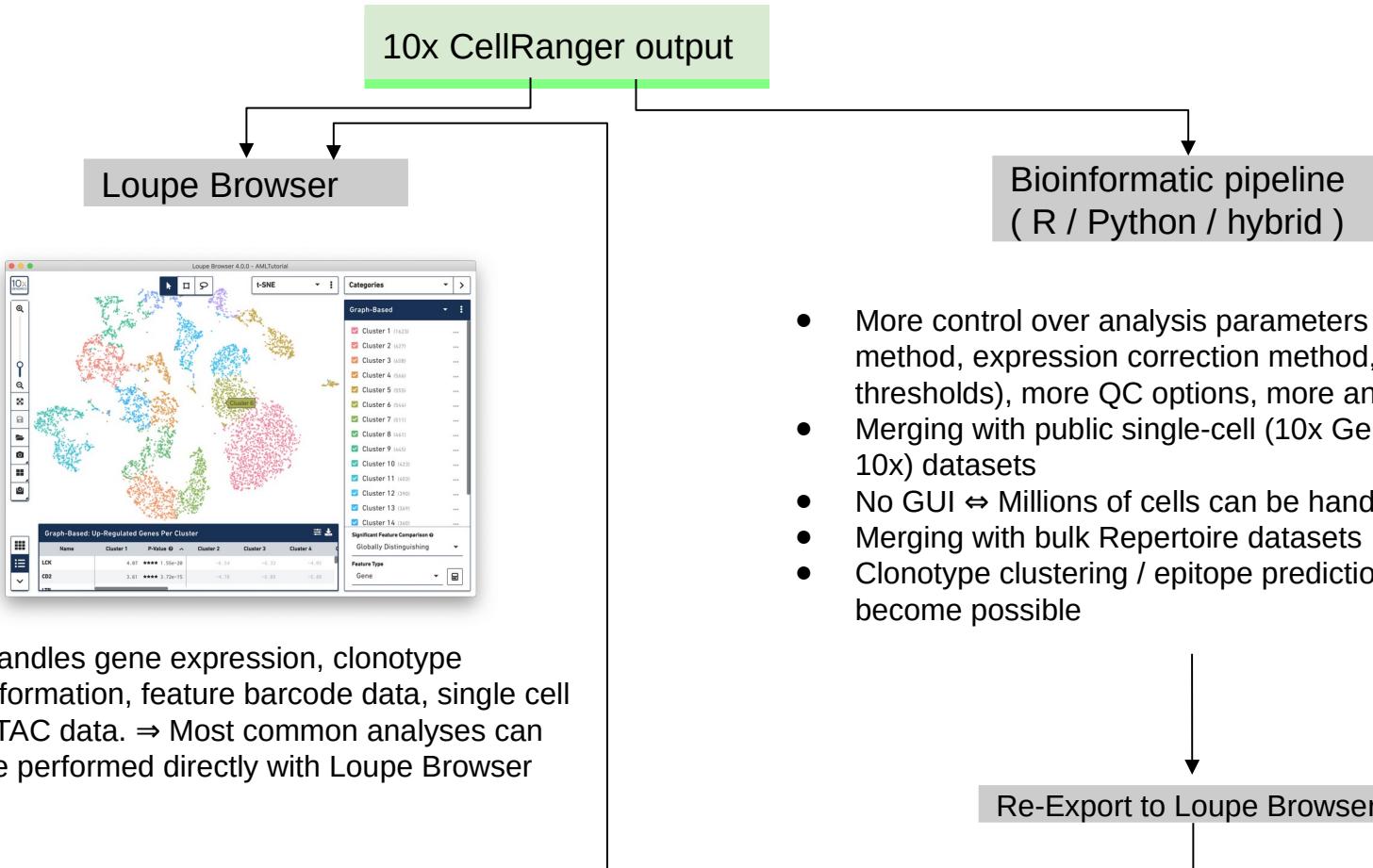
\* Cell barcode information ⇒ 遺伝子発現情報  
とリンクできる。

\* Paired sequence information (alpha & beta,  
heavy & light) ⇒ 高感度解析・タンパク質発現  
・立体構造予測が可能に。

⇒ 特異性 & 細胞表現型 が同時に分かる。

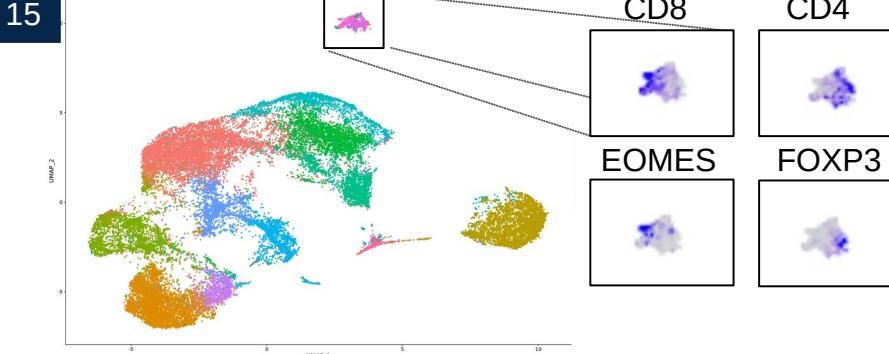
\* Unlike Gene Expression, transcript sequence  
is more important than transcript abundance

# 解析パイプライン



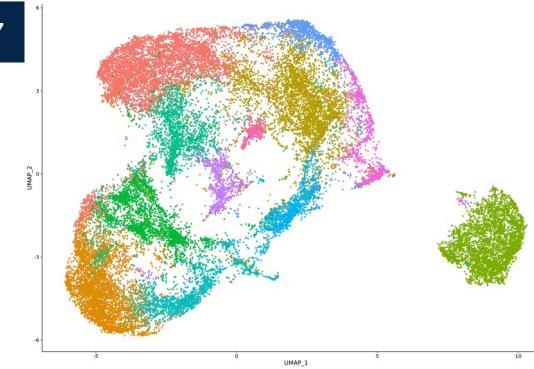
# 解析パイプライン：バイアス除去 (1)

Example: CD3 陽性で FACS ソートされた T 細胞

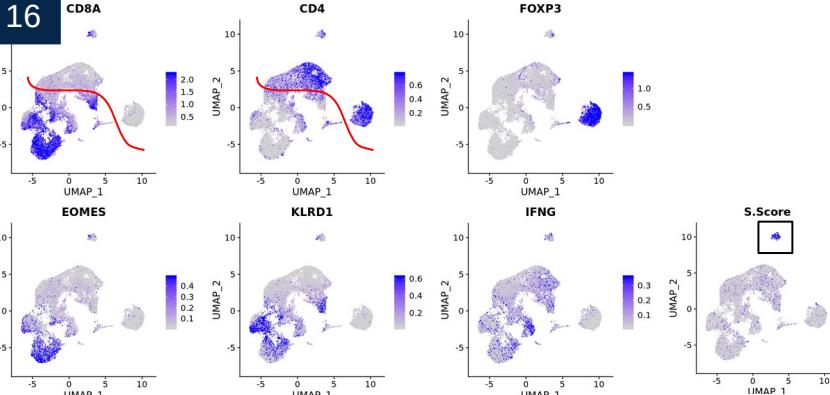


The “outlier cluster” can also be split into a CD8 side, CD4 side, memory CD8 and Tregs, just like the full-size dataset.

17



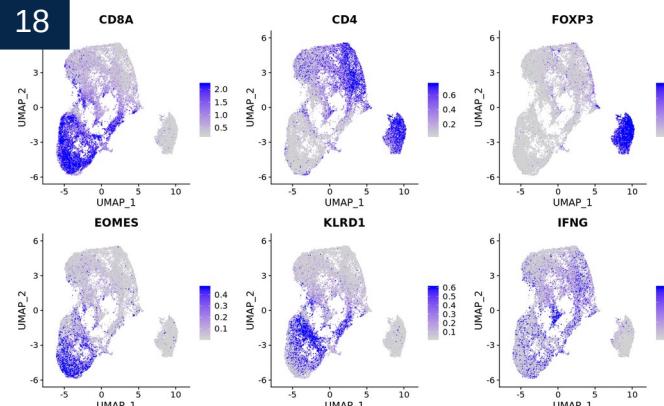
Regressing out: Mitochondrial gene fraction (MT%)



細胞周期補正が結果に大きく影響するデータセットが多い。

(補正しない方が良いデータセットもある。)

18

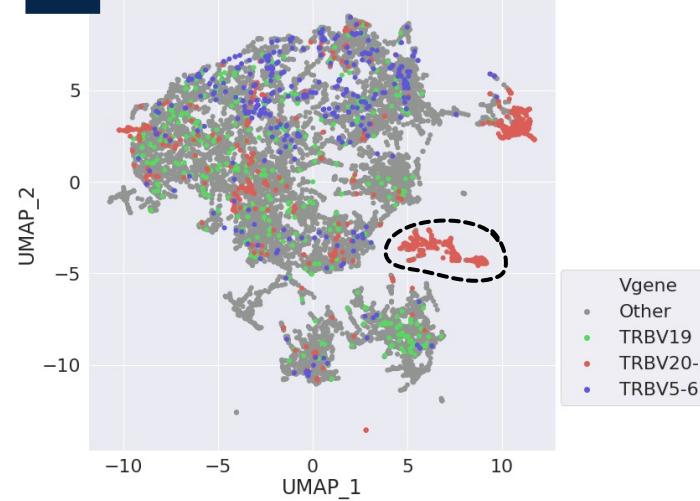


# 解析パイプライン：バイアス除去 (2)

Example: 腫瘍浸潤 T 細胞 (TIL) データセット：受容体遺伝子の統合

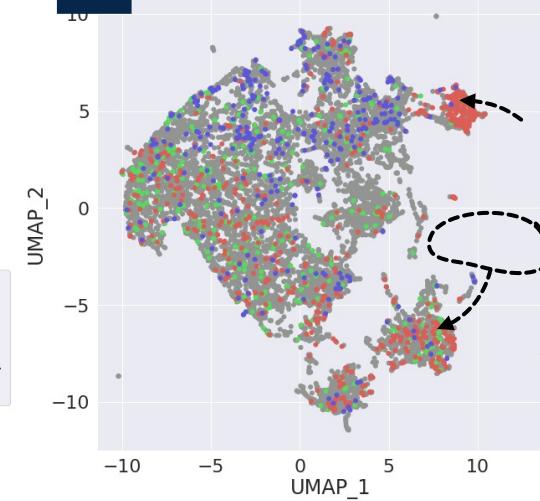
19

Keeping V/D/J genes



20

Without V/D/J genes



- Cluster cells based on their phenotype (gene expression), not their specificity.
- Analyze specificity separately, using 10x VDJ.
- V/J/C gene removal also helps merging 10x 5' and 3' datasets (more V counts for 5', more C counts for 3' technologies)

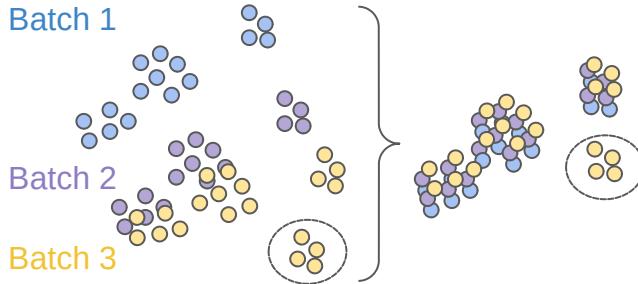
T・B 細胞に着目する際に、V 遺伝子がクラスタリングのフィーチャーとして使用される時もある。  
(特に expanded したクローンが存在する時に要注意)

**Left side:** keeping all TRxV/D/J genes.

**Right side:** replacing all TRBV (eg. TRBV19, TRBV20-1, TRBV5-6, TRBV27 → TRBV, same for TRA, TRD, and D / J region-coding genes) by the V, D, J gene with the highest UMI counts.  
(Apart from that, the exact same pipeline, filtering and clustering parameters were used.)

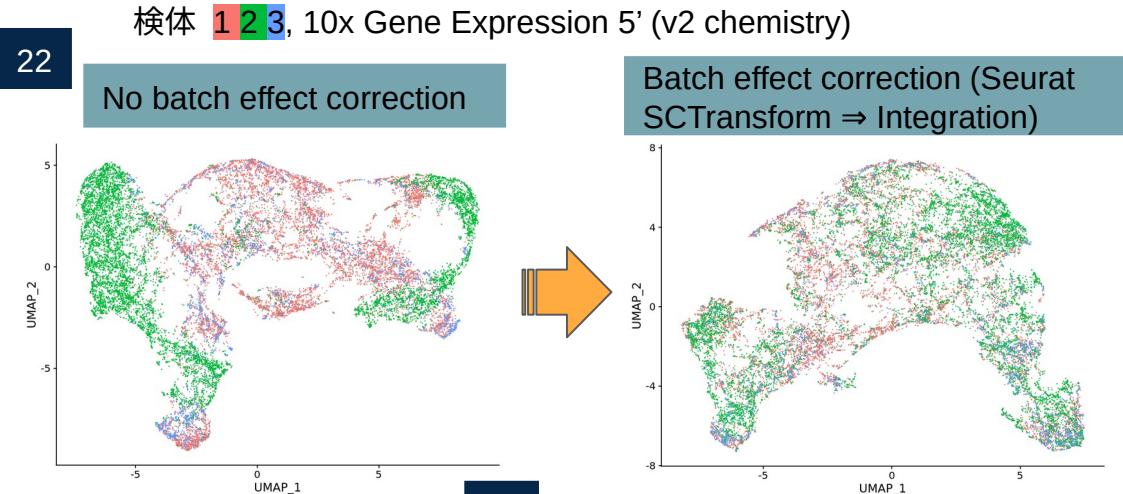
# 解析パイプライン: Batch effect 補正

21



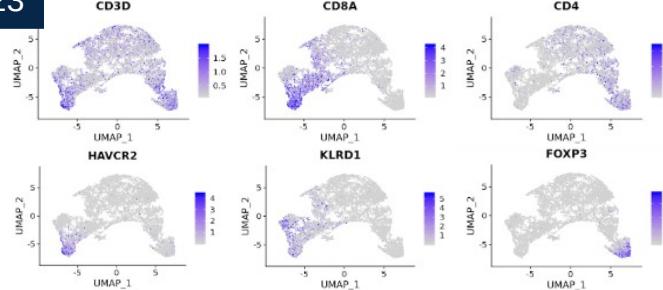
A number of experimental factors such as cell viability, time on ice, cell count accuracy, experimentalist, kit chemistry will affect the final library and metrics such as UMIs / cell, number of genes detected. ⇒ Direct concatenation of normalized data will often result in noticeable differences between datasets, even in the case of technical replicates.

22



23

Batch correction algorithms usually assume that datasets with similar cell contents will be merged ⇒ Must be careful to not overcorrect

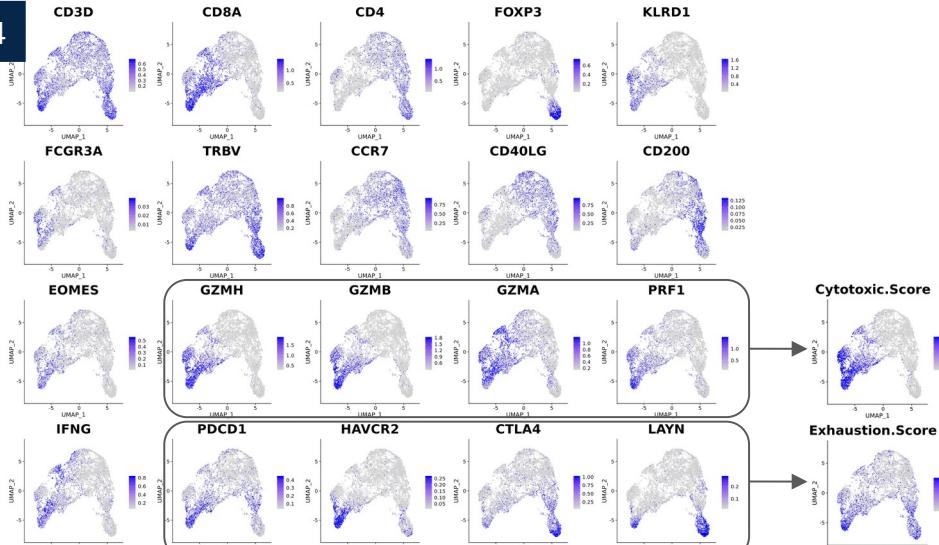


Method: Removing TR\*V/D/J genes and replacing them by 1 gene with their maximum UMI counts, calculate mitochondrial gene content (MT%), keep only cells with nFeature\_RNA between 200 and 2000, MT% < 10, followed by 'standard' preprocessing with Seurat (SCTransform) regressing out MT%, then (**left side**): Concatenation ⇒ PCA ⇒ UMAP plotting | (**Right side**) SelectIntegrationFeatures ⇒ PrepSCTIntegration ⇒ FindIntegrationAnchors ⇒ IntegrateData ⇒ PCA ⇒ UMAP plotting

# 解析パイプライン: Cell Typing

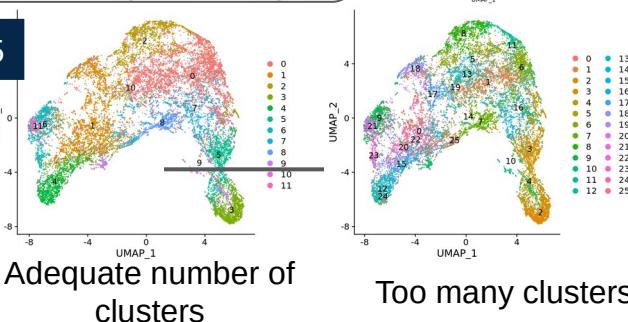
By individual marker genes, marker gene lists:

24



Get a general idea of the number of subtypes by plotting common marker genes  
⇒ Manually assign a cell type to each cluster

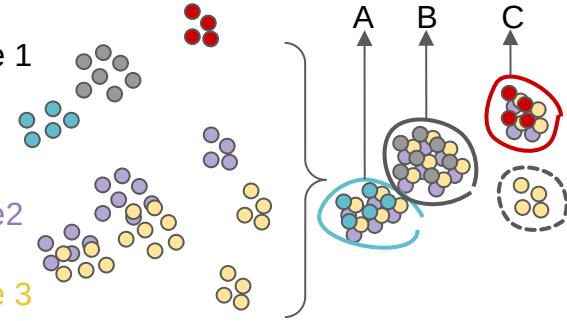
25



26

Merging with labelled datasets:

Sample 1



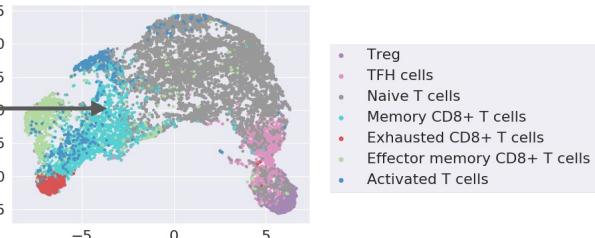
Sample2

Sample 3

Can be automated (and part of the merging process) and works well if expected cell types are similar to the reference dataset

Automatic cell typing functions / packages

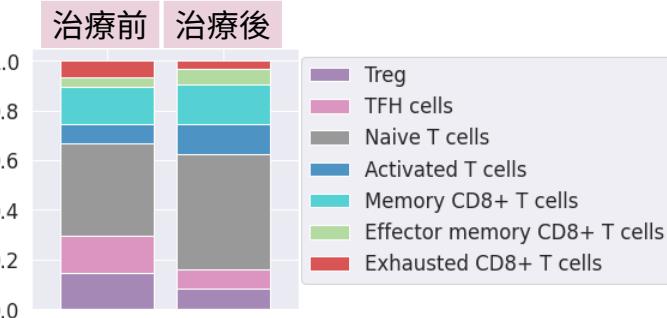
A lot of papers published in 2019~2020. No 'clear winner' yet but the field is moving fast, and the number of reference datasets is increasing quickly !



# 解析パイプライン：特異性・レパートア解析のメリット

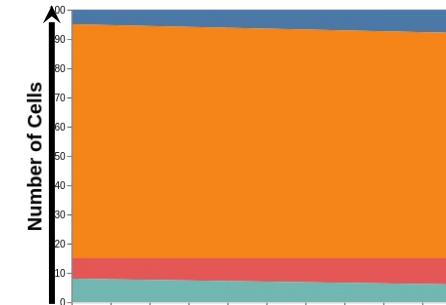
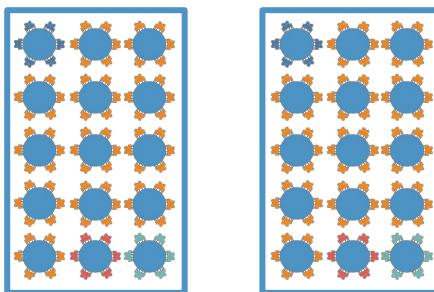
27

## 表現型レベルで比較

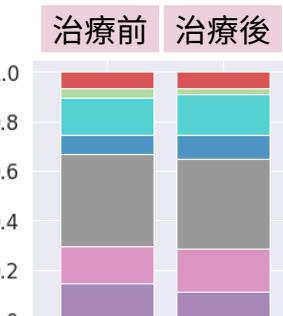


28

## 特異性レベルで比較

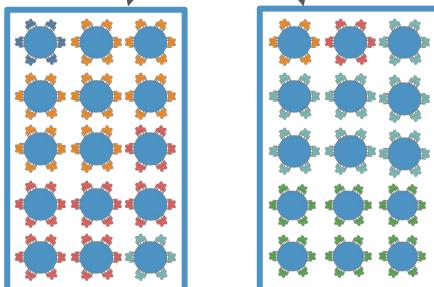


特異性に変動がない



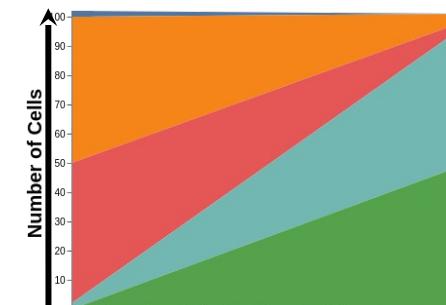
29

各集団の細胞数に変動がない $\Leftrightarrow$ 免疫状態は安定している?



治療前

治療後



新たなクローンが拡大し、標的が変わった?

# 解析パイプライン：フローチャート

scRNA-Seq

CellRanger counts matrix

- \* V/D/J gene removal or aggregation
- \* MT%, # features, # UMIs → Remove outliers
- \* Normalization and scaling
- \* Add useful composite metrics to help cell typing (“Cytotoxic score”, “Exhaustion score”)

Merging with Public data

- \* (Multiple datasets:) Integration
- \* Dimension reduction and plotting (PCA, UMAP)
- \* Add V/D/J genes back for plotting if necessary

Merging with VDJ data

- \* Cell typing (manually with markers, automated with references)
- \* Add clonotype expansion (# cells, % of dataset)
- \* Calculate clonotype diversity per dataset
- \* Calculate clonotype enrichment / depletion / stability vs other available samples

scVDJ-Seq

CellRanger clonotypes csv

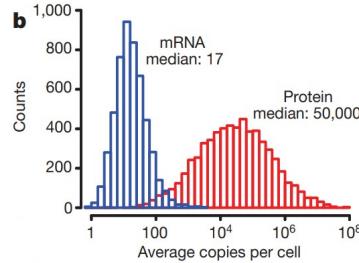
- \* Calculate confidence score from UMIs (multiple clonotypes with similar # UMIs in one cell can be true “polyspecific cells”, or cell multiplets, but hard to analyze in any case)
- \* Clonotype generation probability (pGen) calculation
- \* (Cancer datasets:) Map known reactive clonotypes from in vitro experiment (cell killing assays, IFNy secretion assays...)

## Fully integrated analysis:

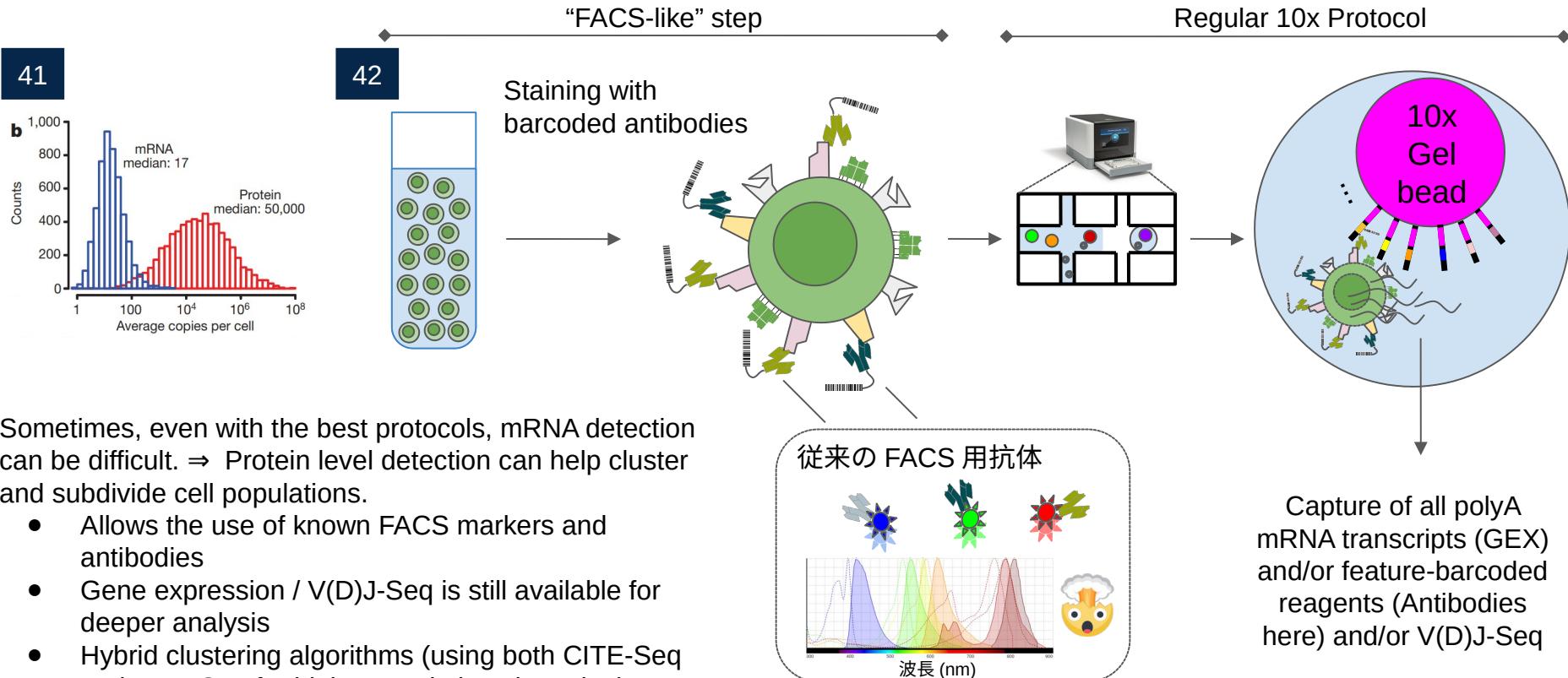
- \* Diversity by cell type
- \* Differential gene expression between clonotype categories, between cell types
- \* Top clonotypes mapping (by cell type, on 2D UMAP)
- \* Clonotype tracking between samples (CIRCOS plots, Streamgraph)

# Feature barcoding: CITE-Seq

41



42

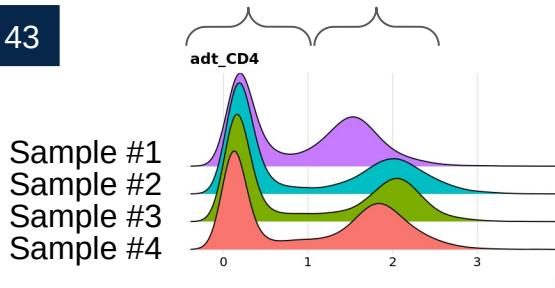


Sometimes, even with the best protocols, mRNA detection can be difficult. ⇒ Protein level detection can help cluster and subdivide cell populations.

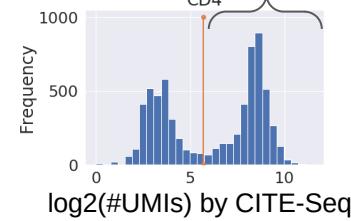
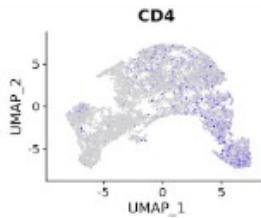
- Allows the use of known FACS markers and antibodies
- Gene expression / V(D)J-Seq is still available for deeper analysis
- Hybrid clustering algorithms (using both CITE-Seq and RNA-Seq for higher resolution clustering) are being developed / released.

# Feature barcoding: CITE-Seq

43



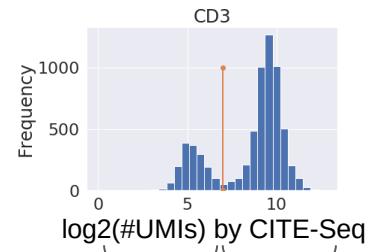
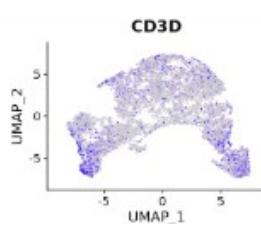
44



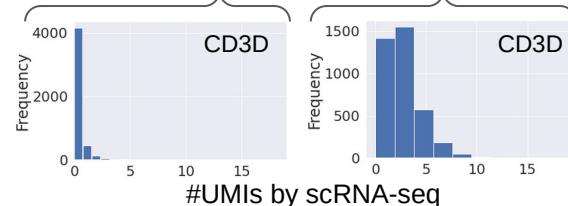
Similarly to FACS, bimodal distribution:  
some background signal in negative  
cells, high signal in positive cells.

- \* No need to model / estimate a distribution, values can be directly normalized
- \* Batch effect is mostly staining-related and easier to correct.

45

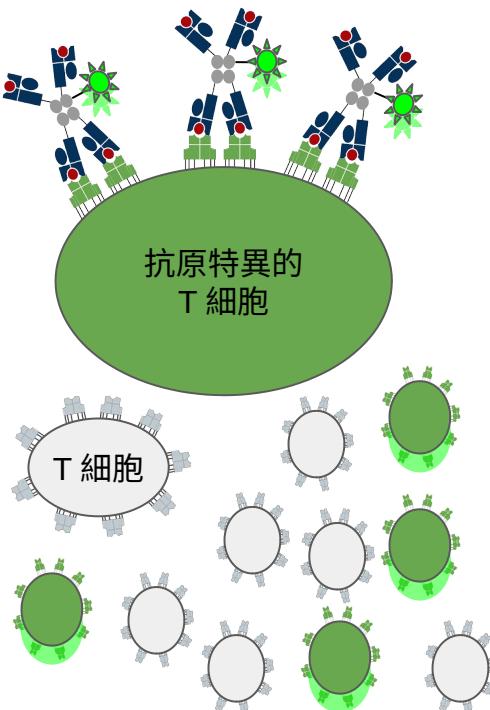


“Positive” populations get ~5000x more UMIs by CITE-Seq than scRNA-Seq  $\Rightarrow$  FACS-like approach and “gating” become possible.



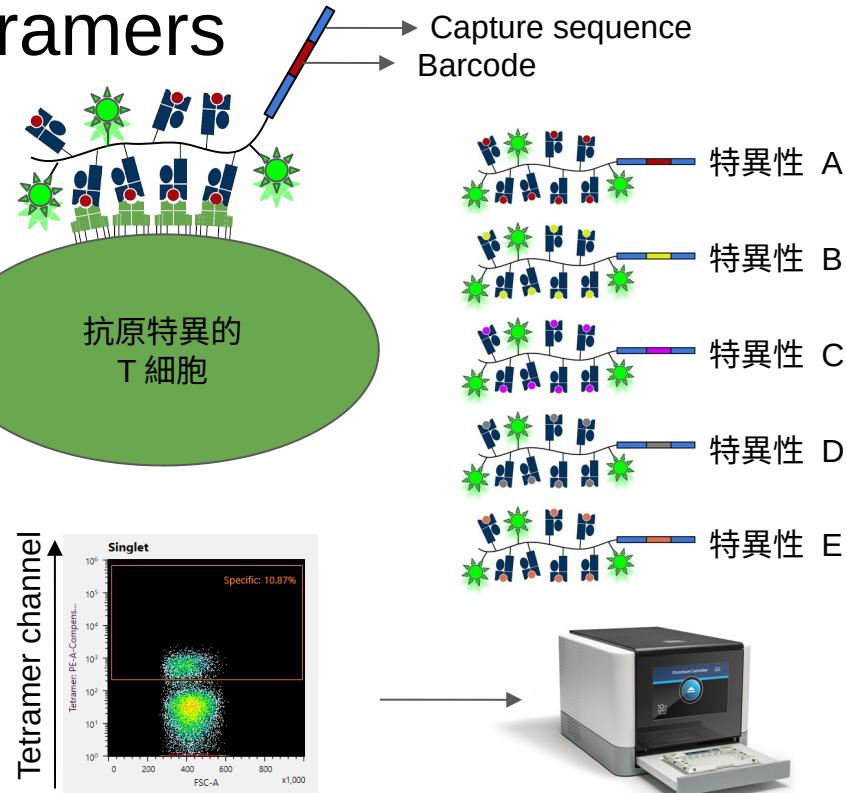
# これからの展開 : dCODE dextramers

## 46 従来の特異的な T 細胞の同定



- Single-well sorting  
⇒ Cloning of alpha/beta chains
- Bulk sorting ⇒ Bulk sequencing

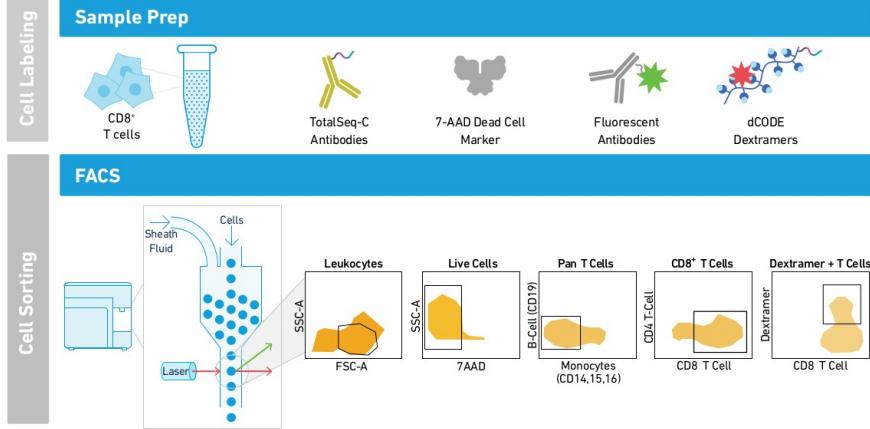
## 47



複数の特異性を同時に同定できる。  
抗原 (pMHC) の種類と数は予算でしか限られていない

# これからの展開 : dCODE dextramers

48



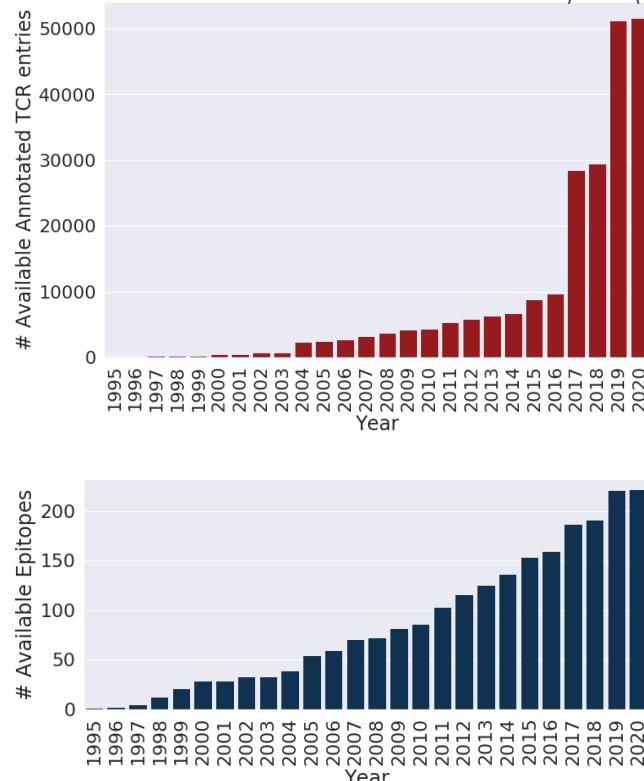
10x 社が去年公開した”dCODE データセット”: フィルタリング後でも特異性 40 種類以上。クロノタイプ数 2 万以上  
⇒ Why is annotated data important ?

Source:

<https://www.immudex.com/media/118671/tf119302-sitc-2018-immudex-poster-in-collaboration-with-10x-genomics-dcode-dextramer-technology.pdf>

49

Chromium を用いた pMHC 特異的 T 細胞のデータセット 2つでこれまで蓄積されたデータの 75%!



VDJ-db statistics as of 2020.03

# これからの展開：特異性予測

51

50

受容体情報 + pMHC 情報

	cd3.alpha	v.alpha	L.alpha	cd3.beta	v.beta	d.beta	L.beta	mhc.a	antigen.epitope
CADLGQNLIF	TRAV5	TRAJ42	CASEGWGNGSYEQYF	TRBV6-5	TRB12-7	HLA-A*03:01	KLGQALQAK		
CALSDPGYSSASKIF	TRAV19	TRAJ32	CASQNLNGDDEQFF	TRBV12-5	TRB01	HLA-A*03:01	KLGQALQAK		
CANNSVQVLLTFT	TRAV12	TRAJ45	CASLNGVQVLLTFT	TRBV14-3	TRB02	HLA-A*03:01	KLGQALQAK		
CAVRSSGGSYIPITF	TRAV13	TRAJ36	CATSGDGNQPIPF	TRBV14-5	TRB02	HLA-A*03:01	KLGQALQAK		
CAVGNRDKKIF	TRAV22	TRAJ30	CASVYMLDPIHF	TRBV9	TRB14-6	HLA-A*03:01	KLGQALQAK		
CAPHSQVQVQIF	TRAV15	TRAJ35	CATGQVQVQVQIF	TRBV9-9	TRB02	HLA-A*03:01	KLGQALQAK		
CAVRSPVQVTAAGKSTF	TRAV38-20V	TRAJ27	CASPTTYDPIQYF	TRBV28	TRB01	HLA-A*03:01	KLGQALQAK		
CAVQVQVQVQVQVQVQVQF	TRAV20	TRAJ36	CATSPSPGGSVNEQFF	TRBV24-1	TRB01	HLA-A*11:01	IVTDTSVVK		
CAYTKDQKIF	TRAV17	TRAJ42	CASPTTYDPIQYF	TRBV6-5	TRB01	HLA-A*03:01	KLGQALQAK		
CAPGGSQQNLIF	TRAV5	TRAJ42	CASPTGLEAFK	TRBV6-6	TRB01	HLA-A*03:01	KLGQALQAK		
CAPVQVQVQVQVQVQVQF	TRAV18	TRAJ36	CASPTTYDPIQYF	TRBV12-5	TRB02	HLA-A*03:01	KLGQALQAK		
CALSEGNAGGTSYKLTIF	TRAV19	TRAJ52	CASCVQVYFSGT	TRBV14-5	TRB02	HLA-A*03:01	KLGQALQAK		
CAVMNDRDKIF	TRAV13	TRAJ30	CASSLMGQTYNEQFF	TRBV14-5	TRB02	HLA-A*11:03	IVTDTSVVK		
CAVNNRDKKIF	TRAV14	TRAJ30	CASVYMLDPIHF	TRBV12-2	TRB01	HLA-A*11:03	KLGQALQAK		
CAVQAVGSGDGKLKF	TRAV20	TRAJ36	CASVYMLDPIHF	TRBV11-2	TRB01	HLA-A*11:03	KLGQALQAK		
CAEPGSQQNLIF	TRAV27	TRAJ42	CASVYMLDPIHF	TRBV19	TRB01	HLA-A*03:01	KLGQALQAK		
CAVSVQVQVQVQVQVQF	TRAV19	TRAJ36	CASVYMLDPIHF	TRBV28	TRB01	HLA-A*11:03	KLGQALQAK		
CAESKNNARLMF	TRAV3	TRAJ31	CASLSGVQEYF	TRBV6-5	TRB01	HLA-A*11:03	AVPRKPSDAK		
CAVSVQVQVQVQVQVQF	TRAV14-2	TRAJ36	CASVYMLDPIHF	TRBV12-5	TRB01	HLA-A*11:03	KLGQALQAK		
CAVSPHDYSLF	TRAV12	TRAJ20	CASVYMLDPIHF	TRBV11-2	TRB01	HLA-A*11:03	IVTDTSVVK		

Annotated Data

機械学習



立体構造予測

配列マッチングで  
特異性同定



配列類似性で  
特異性予測



立体構造類似性で  
特異性予測

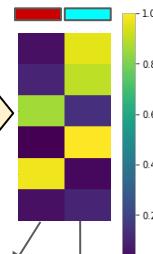


- 健常人のレパートアに比べて、統計的に“濃縮”された特異性の計算
- 既知がんペプチド特異的クローンの予測

52

(例)

各抗原の  
認識予測



Influenza A\*02:GILGFVFTL

COVID-19 Spike A\*24:VYSTGSNVF

末梢血からの免疫状態のスナップショット

CD20

Influenza

CD3

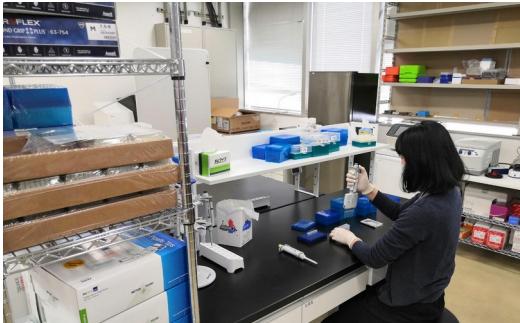
COVID-19

Simultaneous  
phenotype &  
Specificity profiling

B 細胞にも適用可能  
なアプローチ

# KOTAI のシングルセル受託サービス

<https://www.kotai-bio.com/service/>



KOTAIバイオ受託サービス  
by KOTAI Biotechnologies, Inc.

ホーム サービス サービスの流れ お問い合わせ

**10X GENOMICS**  
Certified Service Provider  
Chromium Single Cell Gene Expression

Campaign

シングルセルRNA-seq解析  
シングルセルレバトア解析  
オープニングキャンペーン実施中！

受託サービスの開始を記念して、2020年6月30日までディスカウント致します！

シングルセルRNA-seq解析の詳細はこちら  
シングルセルレバトア解析の詳細はこちら

シングルセルRNA-seq解析  
細胞にそれぞれ異なるバーコードを付与することで、細胞個々のmRNA遺伝子発現を調べることができます。

シングルセルレバトア解析  
1細胞単位のTCR / BCR 発現を調べることができます。TCR α鎖 / β鎖あるいはBCR H鎖 / L鎖のペアを検出することができます。

バイオインフォマティクス情報解析  
情報科学や統計学などのアルゴリズムを用いて、Fastqデータを分析します。

お気軽にお問い合わせください！

# Acknowledgements

国立がん研究センター  
千葉県がんセンター  
富樫庸介先生

国立がん研究センター  
西川博嘉先生

大阪大学  
微生物病研究所 遺伝子情報実験センター

Biken Biomics

10x Genomics

KOTAI バイオ  
実験部・研究開発部の皆さん!

ご清聴ありがとうございました

Questions ? Comments ? (positive or negative!)

[sax@kotai-bio.com](mailto:sax@kotai-bio.com)