補遺

https://github.com/yuifu/2019seminar

scRNA-seq実験: 細胞の単離

マイクロピペッティング

レーザーキャプチャーマイクロダイセクション

FACS

微細流路 (Microfluidics) ベース

- Array microwell-based: C1, iCell8
- Droplet-based: Chromium, ddSeq, Dolomite Bio

scRNA-seq実験: RT, cDNA合成, 増幅

Poly(A) tailing + PCR

Template switching + PCR

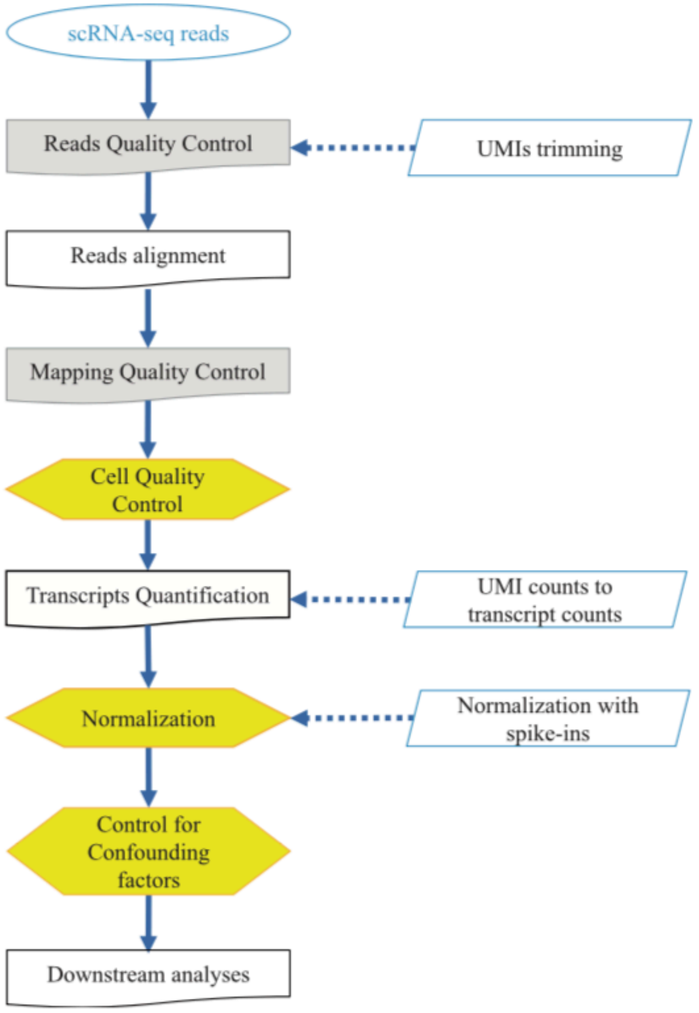
IVT

RamDA

Table 2. Main features of most popular single-cell protocols for mRNA capture, RT and amplification (latest update: December 2017)

Method	Reference	RT primers	cDNA synthesis	Amplification method	UMIs	Transcript coverage	Sample pooling
Tang *et al.*	Tang *et al.* [11]	Poly(T) + poly(A)	Poly(A) tailing	PCR	No	Nearly full-length	No
Quartz-seq2	Sasagawa *et al.* [12]	Poly(T) + poly(A)	Poly(A) tailing	PCR	Yes	3′-end	Yes
STRT-seq	Islam *et al.* [18]	Poly(T)	Template switching	PCR	Yes	5′-end	Yes
SMART-seq2	Picelli *et al.* [13]	Poly(T)	Template switching	PCR	No	Full-length	No
SCRB-seq	Soumillon *et al.* [14]	Poly(T)	Template switching	PCR	Yes	3′-end	No
Drop-seq	Macosko *et al.* [9]	Oligo-dT	Template switching	PCR	Yes	3′-end	Yes
Seq-Well	Gierahn *et al.* [15]	Poly(T)	Template switching	PCR	Yes	3′-end	Yes
SPLiT-seq	Rosenberg, Roco *et al.* [16]	Poly(T) (*in situ*)	Template switching	PCR	Yes	3′-end	Yes
CEL-seq2	Hashimshony *et al.* [17]	Poly(T)	IVT	IVT	Yes	3′-end	Yes

[Molin+, 2018, Briefings in Bioinformatics]



[Molin+, 2018, Briefings in Bioinformatics]

scRNA-seqデータ解析: 前処理 (1)

Read QC

FastQC, fastp, sinQC, Scater

Triming

Trimmomatic, cutadapt, fastp

UMI trimming

Demultiplexing

scPipe

demultiplexer_quartz-seq2 [1]

Mapping (Read alignment)

STAR, HISAT2, Kallisto (pseudo-alignment)

Low-quality cell detection

RSeQC, SCell, Cellity

[1] https://github.com/rikenbit/demultiplexer_quartz-seq2

scRNA-seqデータ解析: 前処理 (2)

Expression level quantification

- Alignment-free: Kallisto
- Alignment-required: FeatureCounts, RSEM, StringTie, HTSeq

UMI-counting

- UMI-tools, Je
- 補正: TRUmiCount

collapsing (類似したUMI, cell barcodeを一つのまとめる)

collapsing

scRNA-seqデータ解析: 前処理 (3)

正規化

Linnorm, scran

バッチエフェクト補正

バッチ補正済み発現量行列:

MNNs, Scanorama, scMerge

低次元表現行列: Seurat, MINT

共変量補正 (細胞周期とか)

BASiCS

variance decomposition-based

regression-based

scLVM (knowledge-based)

ccRemover

Oscope (cyclic genes)

次元圧縮

PCA

t-SNE

Diffusion maps (Destiny)

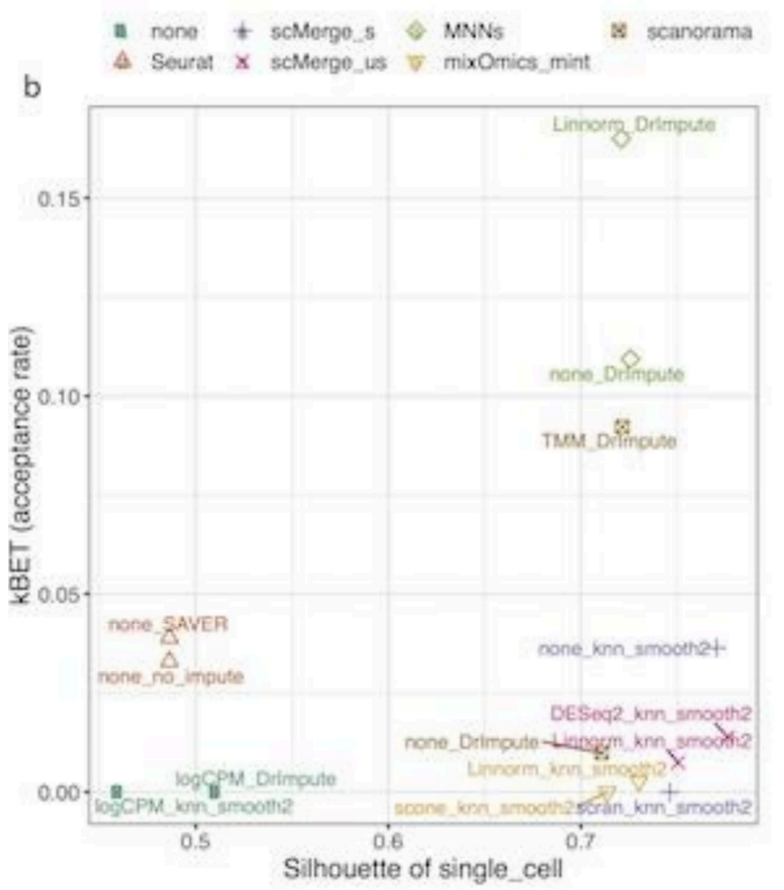
バッチ効果

MNN

特に良い

Seurat

あまりよくない (t-SNEが距離を保存しないから？)



Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments [Tian+, 2019, Nature Methods]

t-SNEについて

手法:

最近傍ネットワーク上でのランダムウォークによって、元データ空間での局所的な距離を保持しながら低次元に写像

特徴:

非線形（遺伝子発現量と低次元空間での座標が線形でない）

確率的（何回もやると結果が変わる）

perplexity (近傍点の数を反映) によっても結果が変わる

実践上の注意:

何度もシード（確率的挙動を制御）とperplexityを変えて結果をみてみるのがよい

実験デザイン (1)

[Molin+, 2018, Briefings in Bioinformatics]

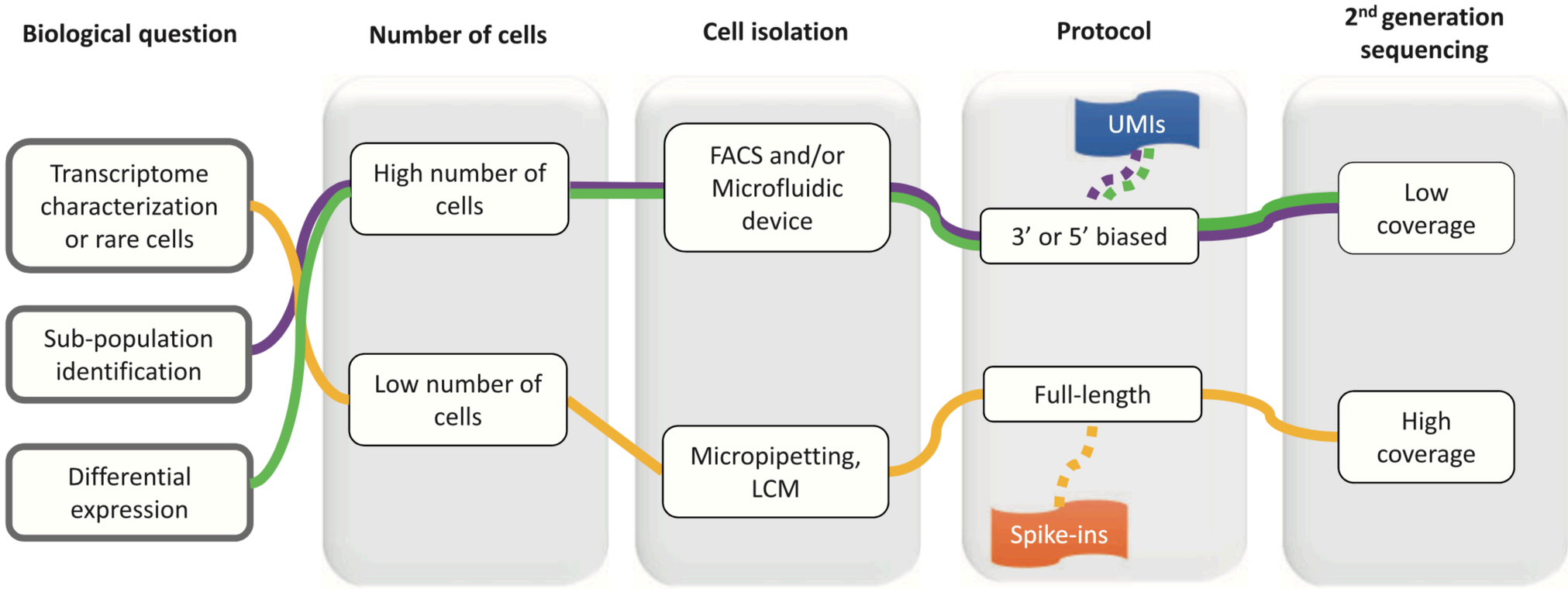


Figure 3. Workflow diagram of possible experimental designs of a scRNA-seq experiment. Dashed lines represent optional choices. Figure inspired by Cannoodt *et al.* [82].

現状では、細胞数を取るか、フルレングスを取るかの二択

実験デザイン (2): バイアスと対抗策

Table 3. Main sources of bias in scRNA-seq experiments and solutions for limiting their impact

Source of bias	Type	Effect	Current solutions
RNA capture and RT efficiency	Technical	Stochastic zeroes	Spike-ins, statistical modelling
cDNA amplification	Technical	Loss of quantification accuracy	UMIs, statistical modelling
Batch effects	Technical	Introduce a signal different from the true biological signal	Statistical modelling
HVGs, transcriptional burst	Biological	Increase variance in the data	Statistical modelling
Cell-cycle stage, differentiation state, etc.	Biological	Confuse the true biological signal	Cell visualization, statistical modelling

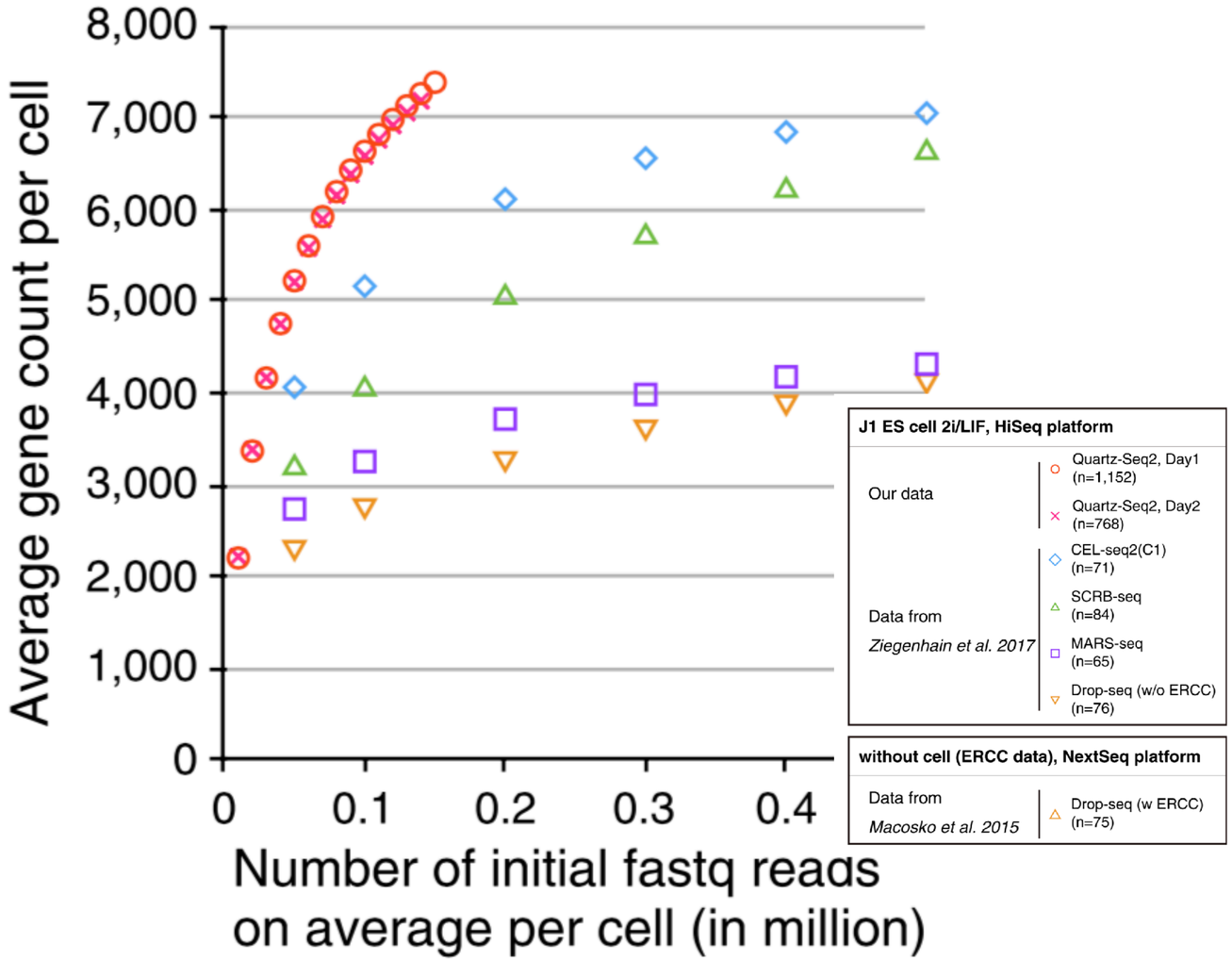
[Molin+, 2018, Briefings in Bioinformatics]

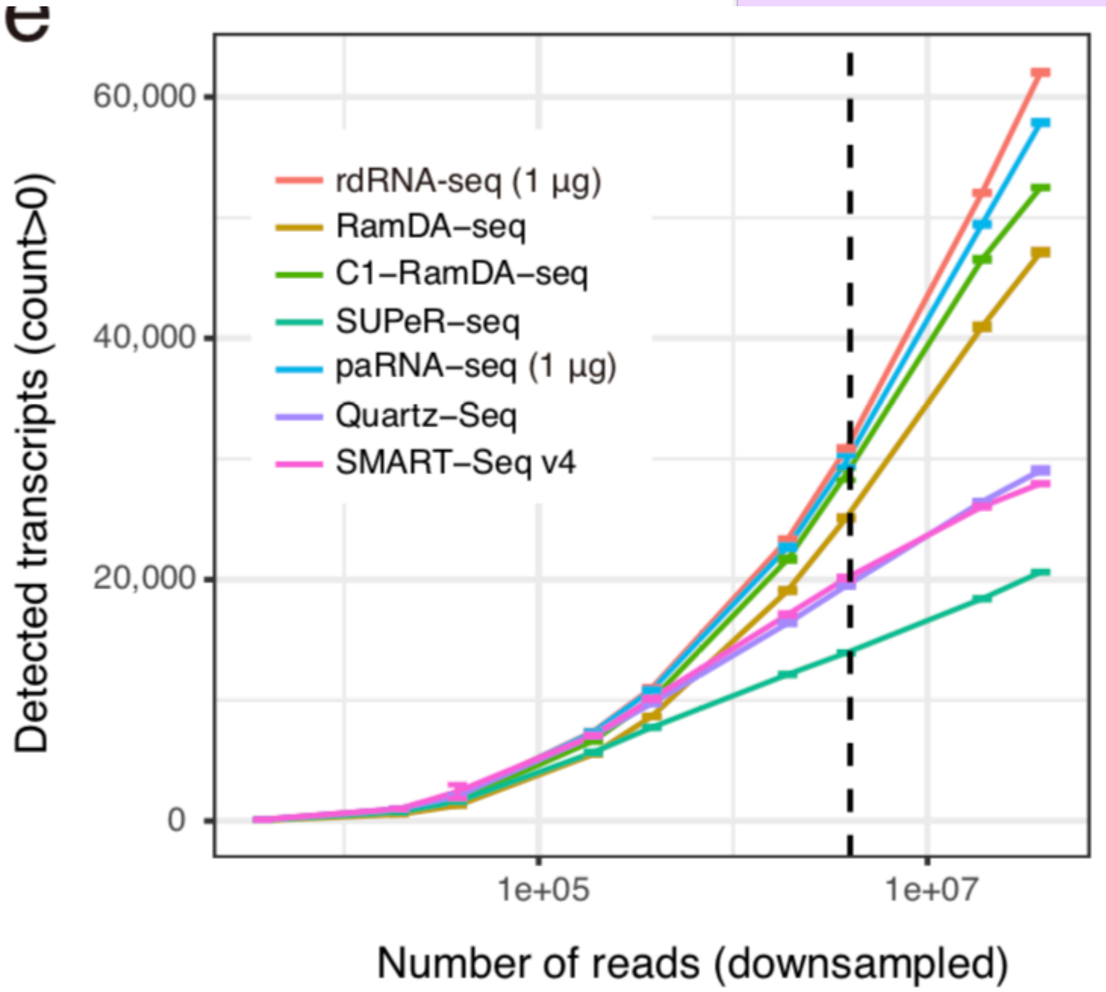
実験: Spilke-in, UMI

解析: 種々のバイアスの統計モデリング、可視化

この他にも、細胞サイズや細胞が含むRNA量の違いも影響する

シークエンスするリード数が多いほど検出遺伝子数は多くなる





[Sasagawa+, 2018, Genome Biology]

[Hayashi+, 2018, Nat. Commun.]

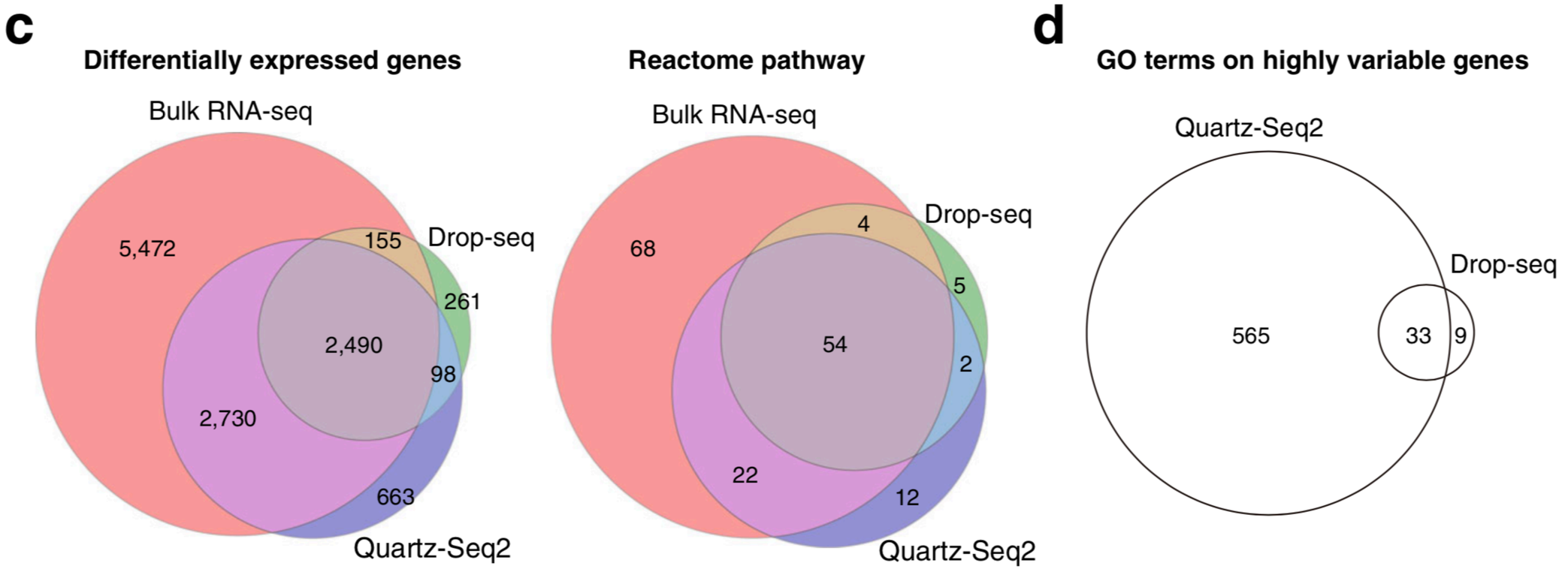
Pollen et al. [85] state that 50 000 reads per cell are sufficient to detect the majority of genes that contribute to the overall popu- lation variance and that a sequencing depth between 5000 and 50 000 reads per cell is sufficient to detect cell subpopulations. However, at low depth, it is not possible to detect most of the lowly expressed genes [85].

Shalek et al. [86] state that 1 million reads per cell are sufficient to accurately estimate the mean and variance of gene expression [30].

Consistently, Tung et al. [87] suggest that using a sequencing depth of 1.5 million reads per cell is sufficient to be able to monitor also the lowly expressed genes.

Rizzetto et al. [83] suggest a minimum read length of 100 bp paired-end with sequencing depth >250 000 reads per cell for accurate detection of gene expression or identification of cell subpopulations, and for minimizing the technical noise.

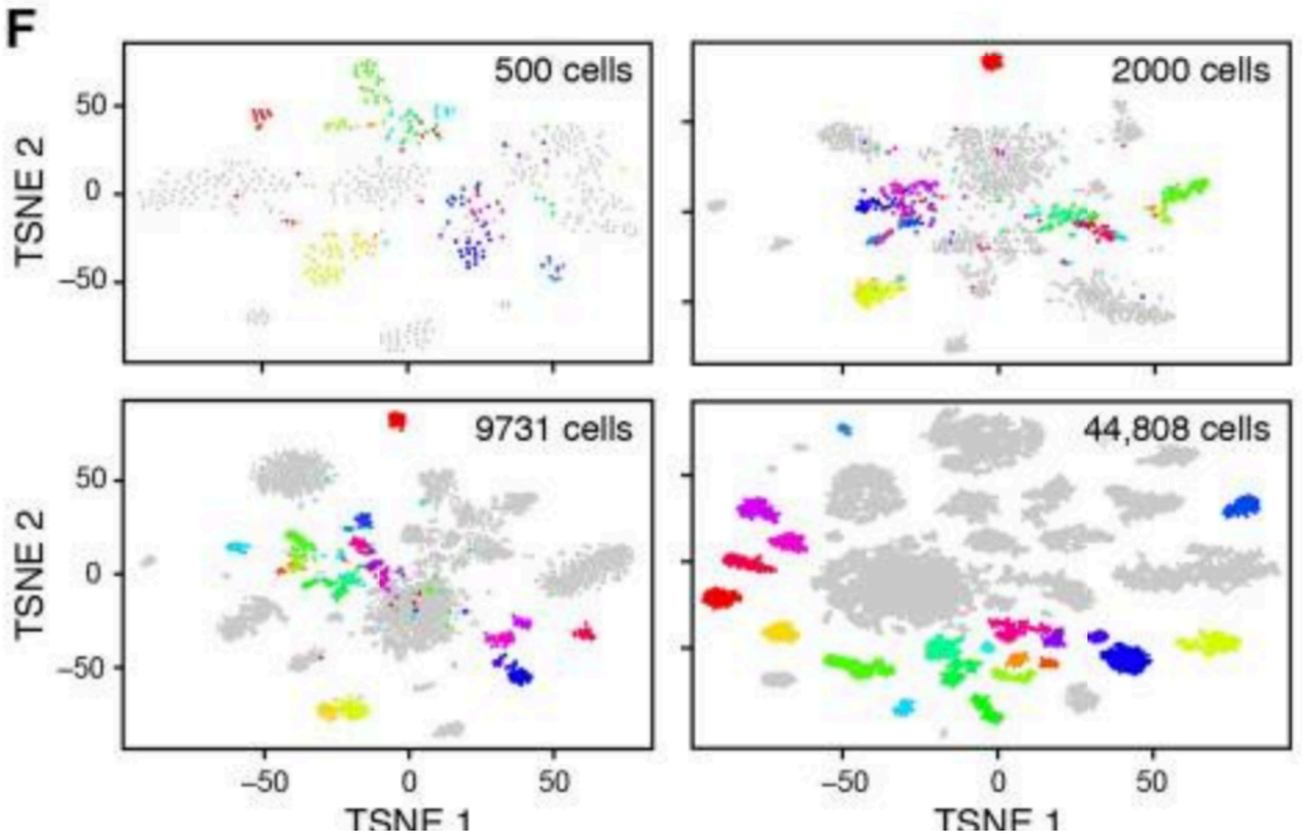
高い検出遺伝子数のメリット: DEG解析



DEG（発現変動遺伝子）や変動する遺伝子機能（GO解析など）が多く発見できる

[Sasagawa+, 2018, Genome Biology]

多い細胞数のメリット:



細胞型、細胞亜集団、新規サブタイプがより多く見つかる

Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets [Macosko+, 2016, Cell]