

AJACS「シングルセルRNA-seqを知って・学んで・使う」

2024年12月23日 13:30～15:50

**遺伝子発現解析だけではない！
シングルセルデータを活かしまろう**

**理化学研究所 生命医科学研究センター
京都大学 ヒト生物学高等研究拠点
小口 綾貴子**



本日本話させて頂く内容

遺伝子の発現スイッチである**エンハンサー**

RNAの**5'末端**を捉える

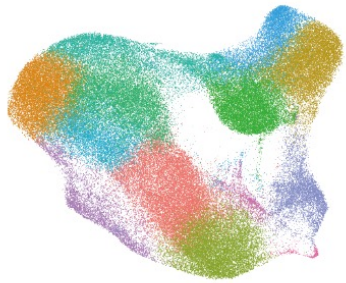
時代は**シングルセル**解析へ

独自のトランスクリプトーム解析法(ReapTEC法)の確立

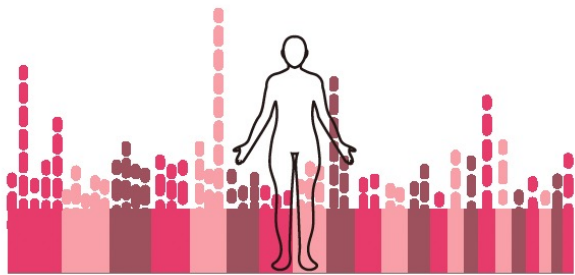
ReapTEC法の使い方

研究モチベーション：ヒト疾患の分子メカニズムを理解したい

1細胞**エンハンサー**解析技術を確認した



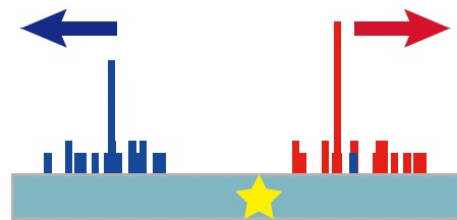
約100万個のヘルパーT細胞のシングルセルデータに適用



大規模な疾患ゲノム解析

ヘルパーT細胞の**エンハンサー**活性地図を構築

エンハンサーRNAの転写開始点

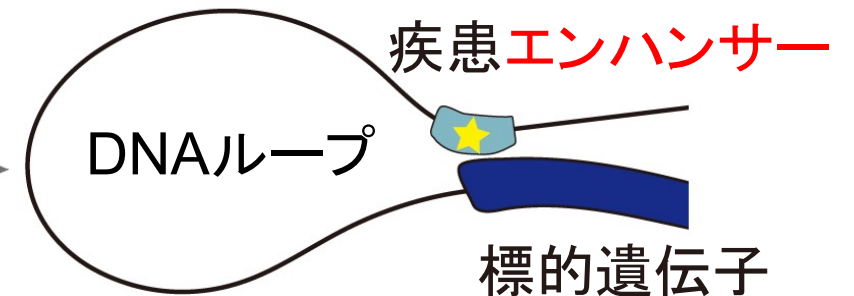


参照 TTCACAAGG
疾患 TTCATAAGG

↑
遺伝的変異

疾患**エンハンサー**を606個同定した

ヘルパーT細胞でゲノムの3次元構造を読み解いた

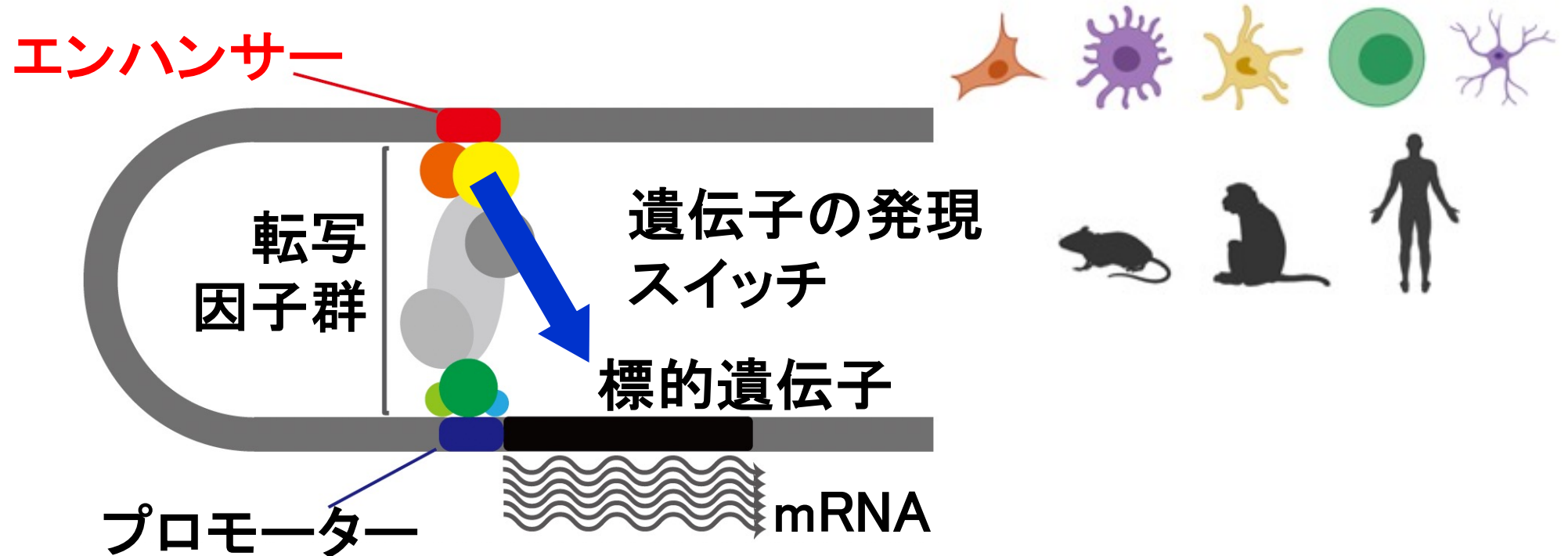


疾患**エンハンサー**の標的遺伝子を推定した

新しい治療標的分子を見出す土台を構築した

エンハンサーは細胞種特異的に標的遺伝子を活性化するが 依然エンハンサーの全貌は明らかではない

- ・配列上はプロモーターの遠位に存在する。
- ・空間的に標的遺伝子のプロモーターに近接する。



エンハンサーの理解のためには、
細胞種ごとのエンハンサーマップが必要

エンハンサーの領域に疾患の**遺伝子的変異**が濃縮している

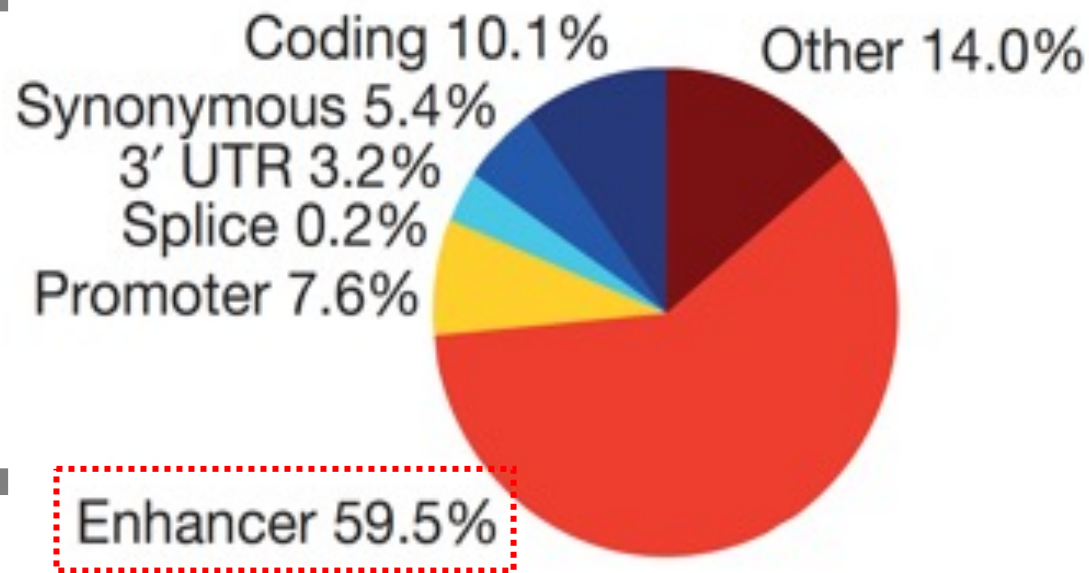
エンハンサー領域

健康な人 ...AGATG...

病気の人 ...AG**T**TG...

病気と関わる配列の違い
(遺伝的変異)

Human disease-SNPs



Farh et al. *Nature* 2015

その遺伝的変異は、遺伝子の“質”ではなく“**量**”を**変化**させ、
疾患の発症に関わりうる。

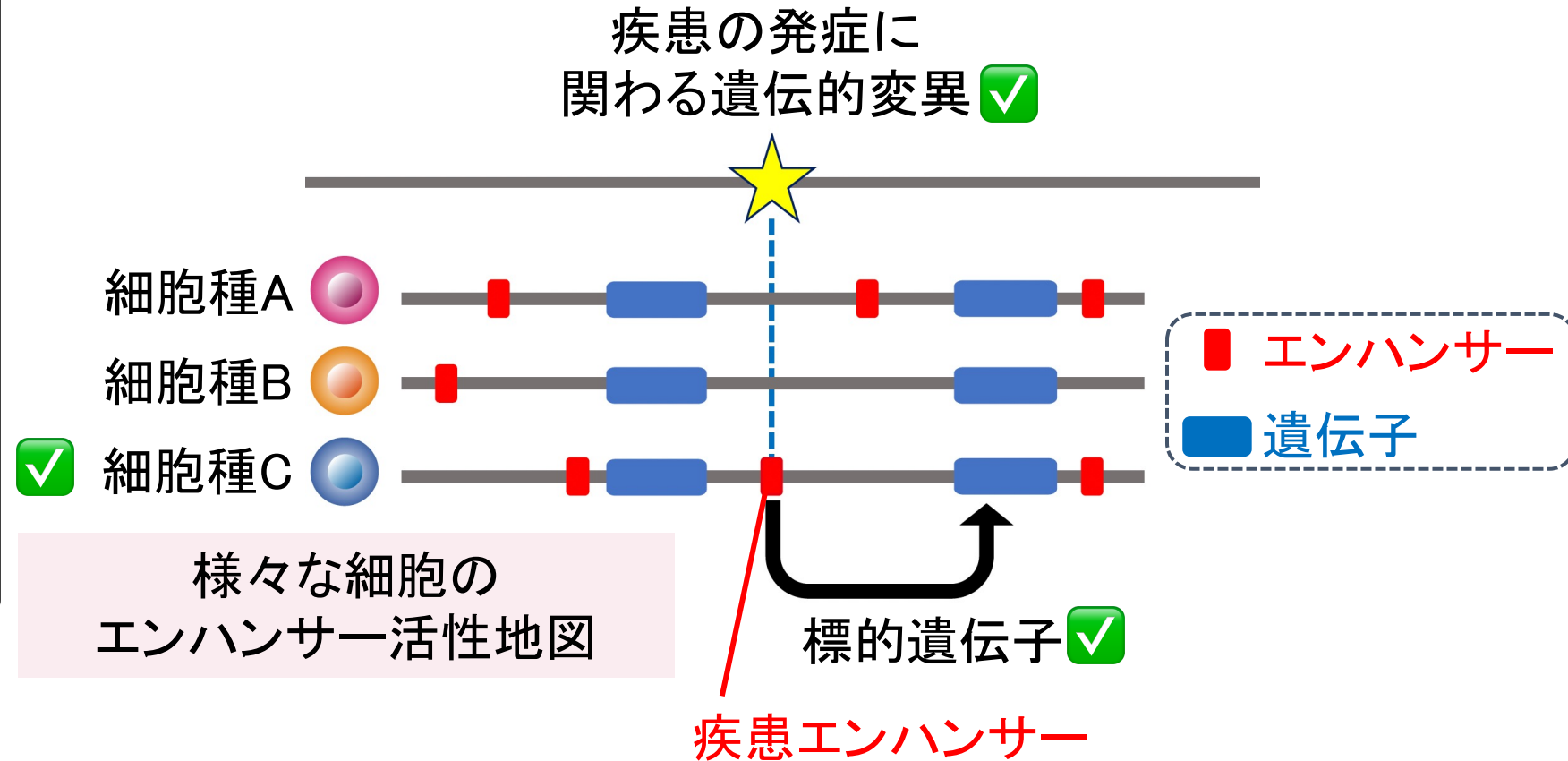
ヒトの疾患の理解のためヒトの各々の細胞で 活性化する**エンハンサーマップ**を作成することにした

多くはタンパクをコーディングしないゲノム領域
に存在し、意義不明であった。



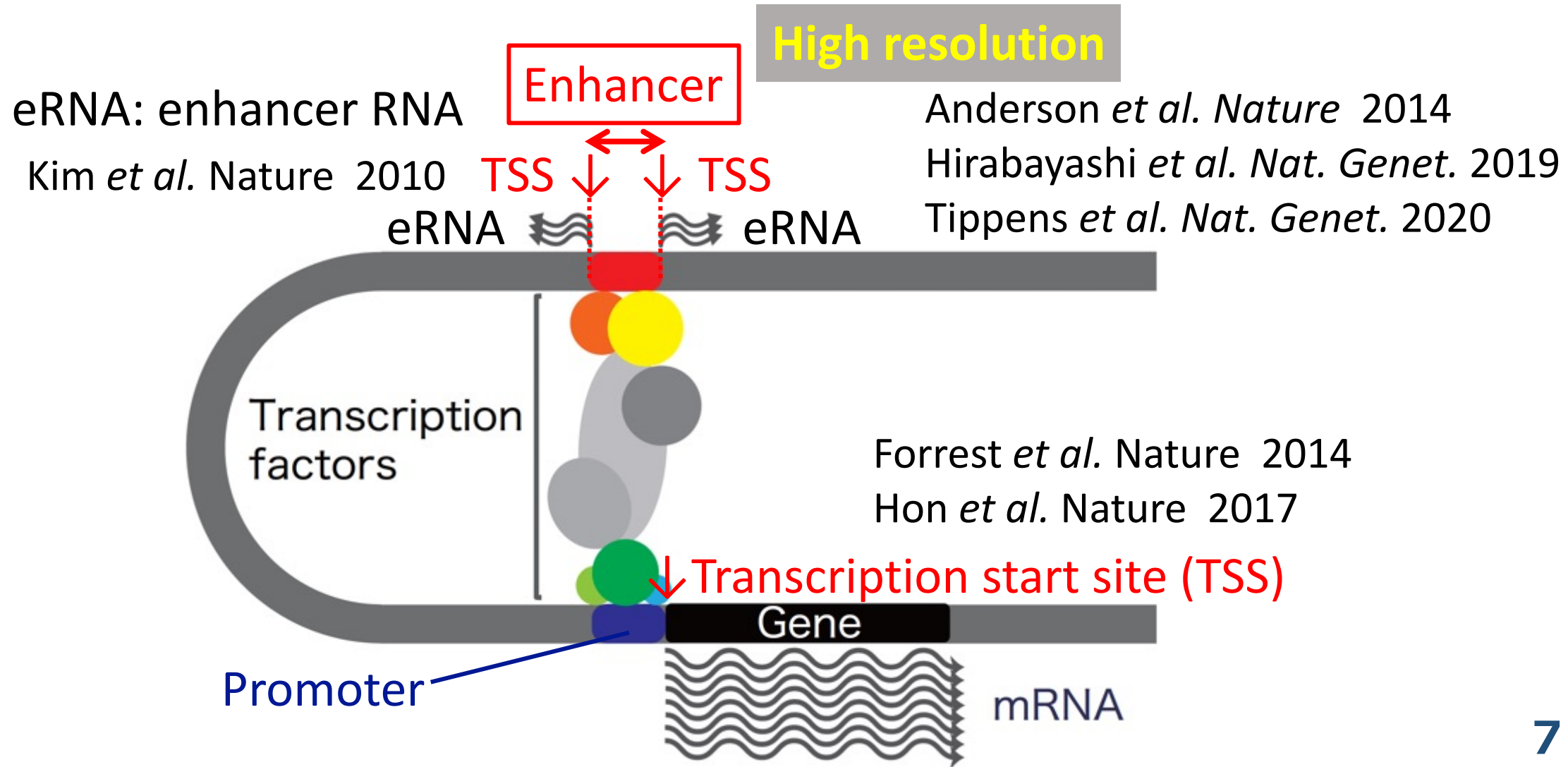
Ozaki et al.
2002

理研で世界初のゲノムワイド関連
解析 (GWAS) が実施され、疾患
感受性領域が次々に報告されて
いる。

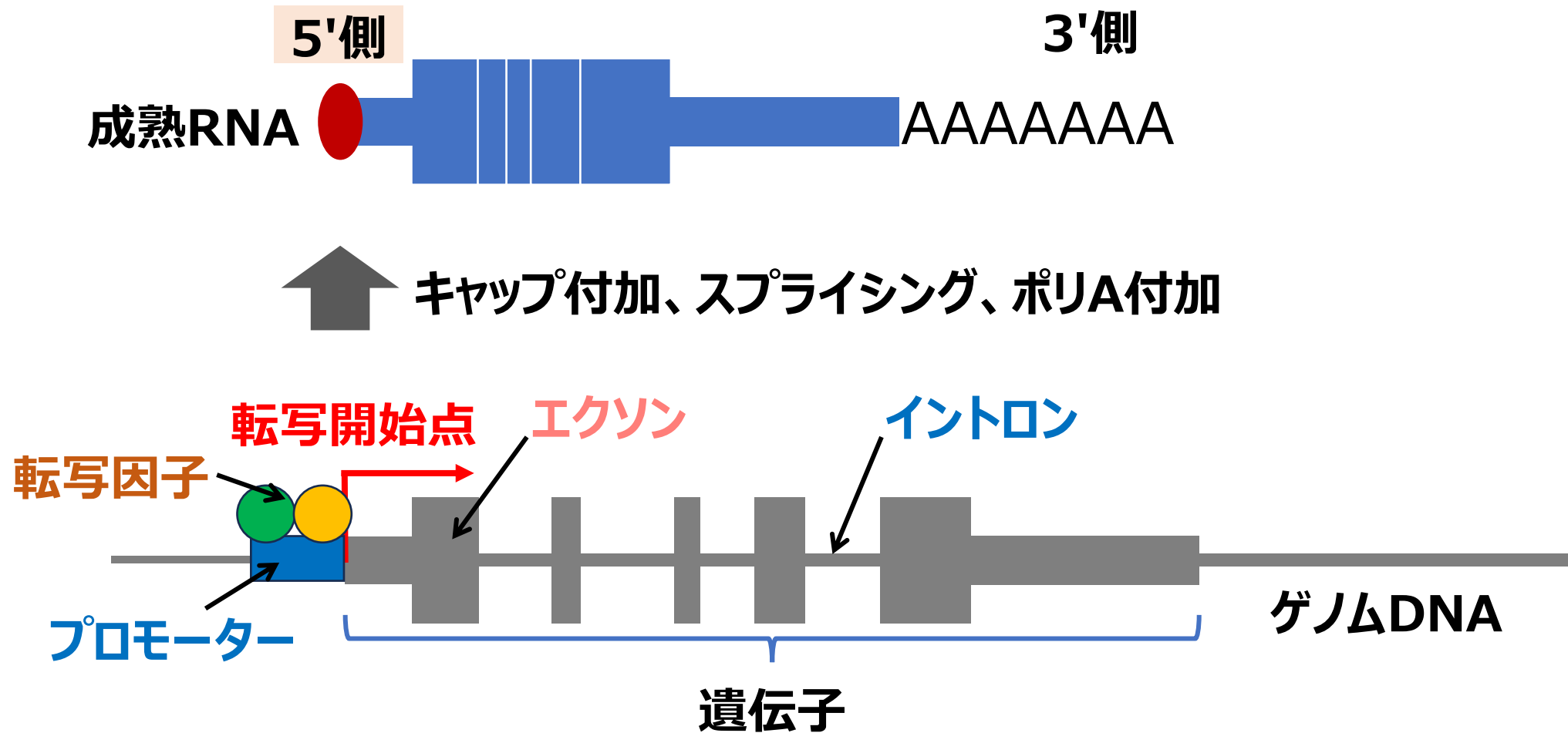


RNAの転写開始点を捉えることでプロモーターだけでなく 活性化した状態のエンハンサー領域も同定できる

“Active” enhancers produce bidirectional eRNAs

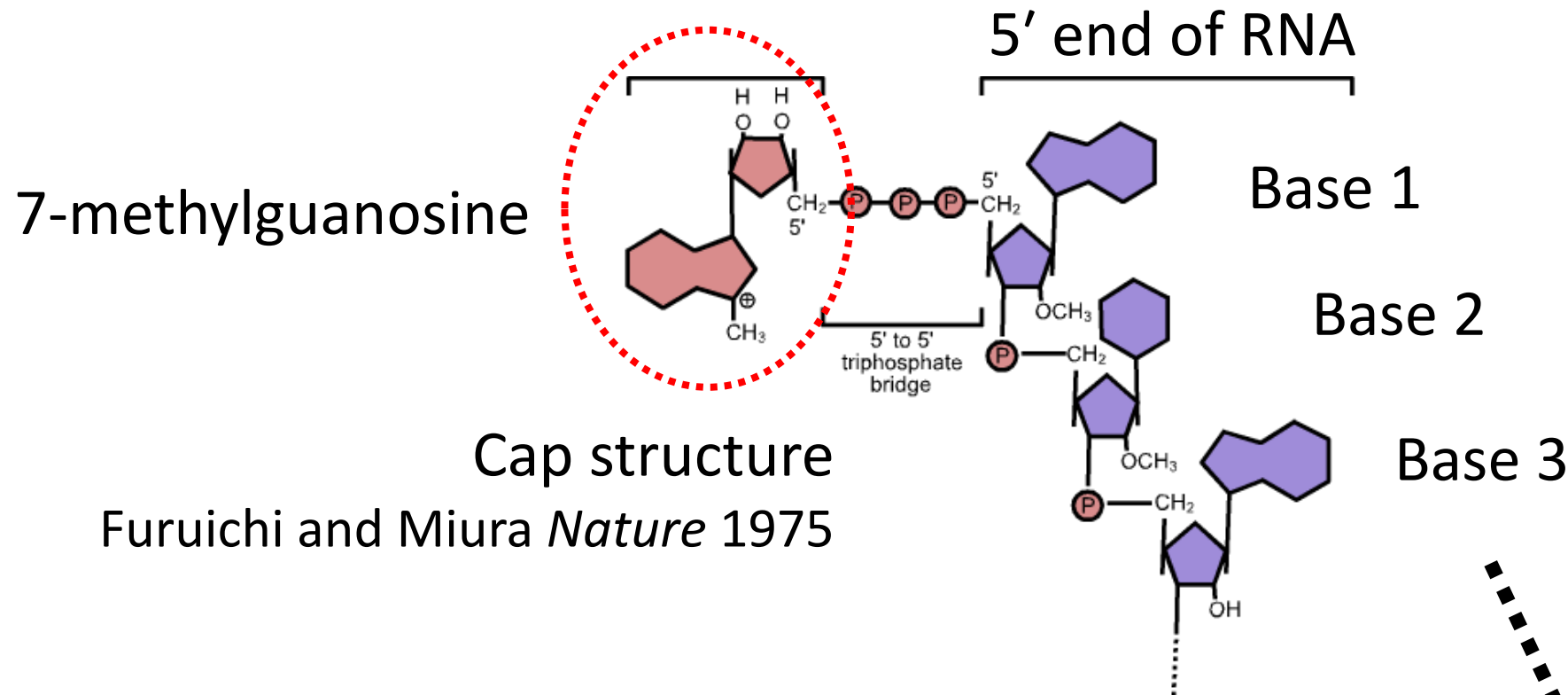


ゲノム情報の流れ : DNA → RNA → Protein



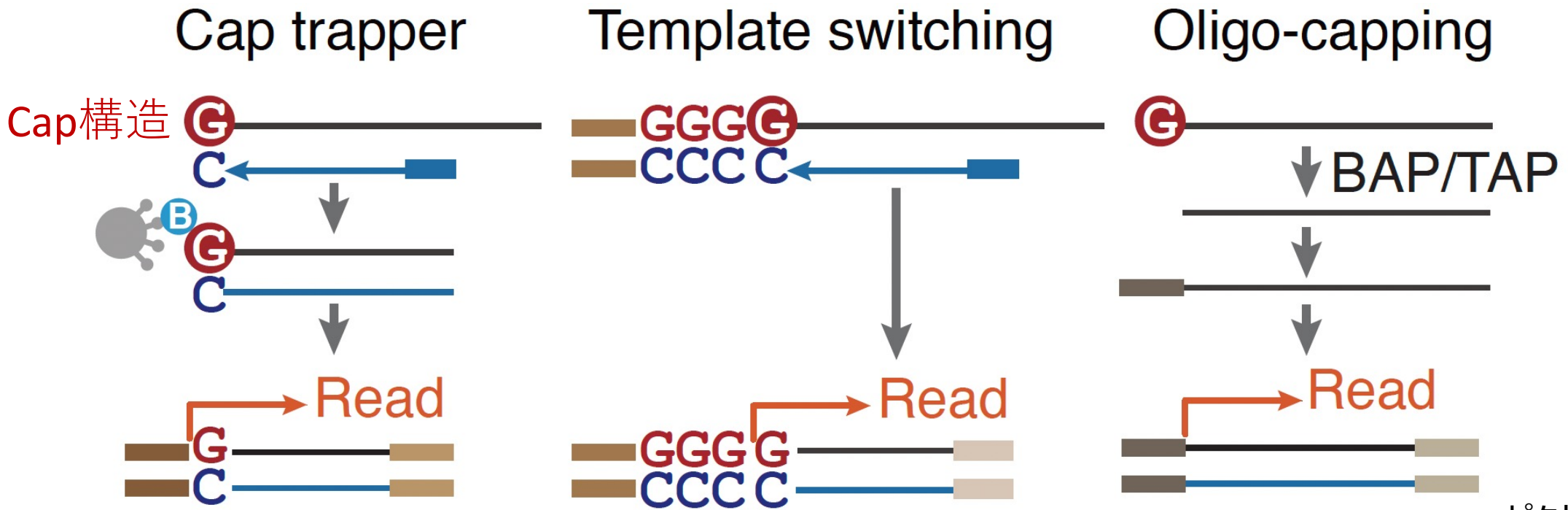
RNA先頭の5' 末端はCap化されている

- RNAポリメラーゼIIにより転写されるRNAの5'末端に付加される。
- 転写された直後に付加される。
- RNAの安定性や翻訳効率に関与する。
- mRNAワクチンにも利用されている。



シングルセルではTemplate switching法で5'末端を捉えることが多い

5'末端を捉える主な3つの技術



CAGE法

(Cap Analysis of Gene Expression)

高い正確性

Carninci *et al.*
Genomics 1996

10x 5'キット

少量から可

Zhu *et al.*
Biotechniques 2001

GRO-cap

Kazuo *et al.*
Gene 1994

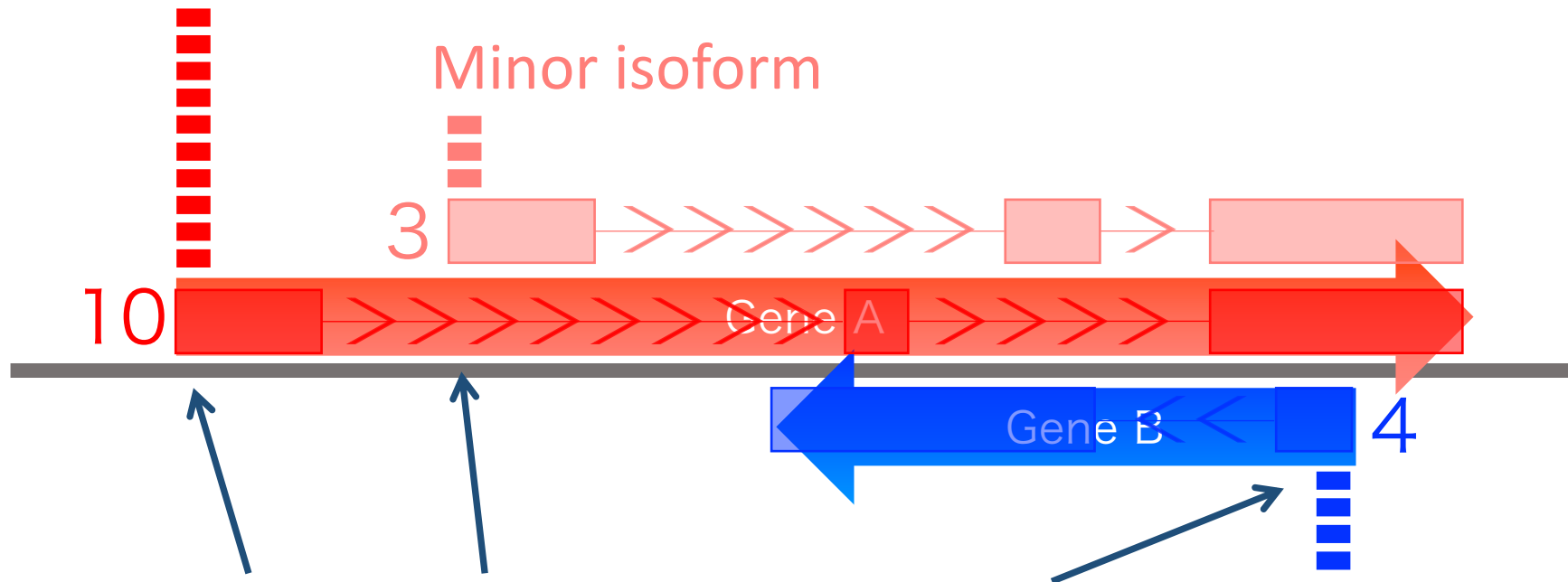


ピクトグラム
車いすテニス
東京パラリンピック2021

Adiconis *et al.*
Nat Methods 2018

RNAの5'末端 (転写開始点)を解析することは 通常のRNA-seqにはないメリットがある

Major isoform ① アイソフォームごとの発現量の解析ができる。



② 転写因子が付くプロモーター領域の解析ができる。
それにより遺伝子の転写制御の理解に繋がる。



時代は、Bulk解析からシングルセルへ

3'キット vs 5'キット

ポイント

Single Cell Gene Expression 3'キット

3' gene expression profiling at scale with single cell resolution.

Single Cell Gene Expression Flex

Fixed RNA Profiling assay for comprehensive probe-based gene expression profiling with single cell resolution.

どっちでもない

ポイント

Single Cell Immune Profiling 5'キット

5' gene expression alongside V(D)J repertoire profiling and antigen specificity of T and B cells.

Single Cell Multiome ATAC + Gene Expression

3'キット

ポイント

Combined profiling of 3' gene expression and chromatin accessibility from the same cell.

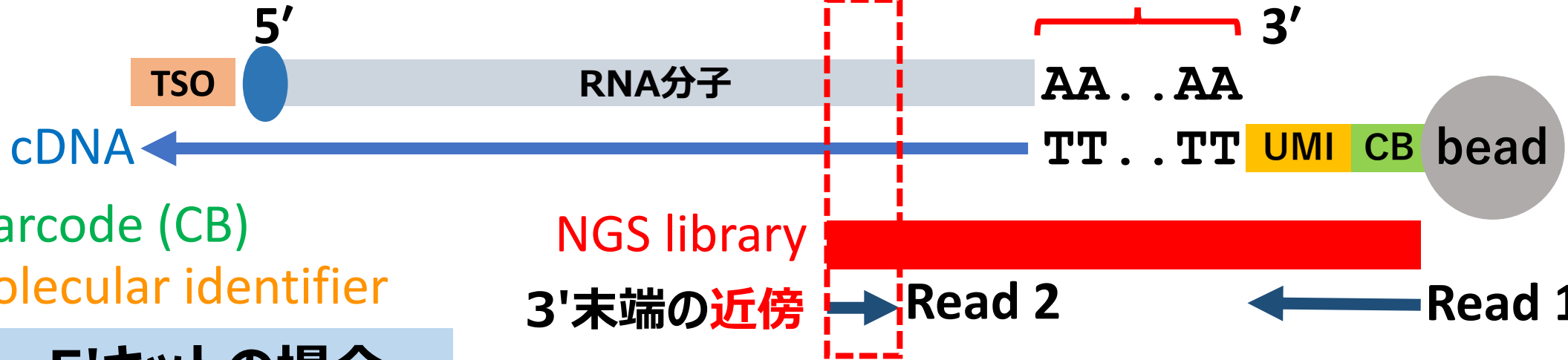
10x GENOMICS

10x GENOMICS社様のウェブサイトより



今回は一般的な3'キットでなく、5'キットを用いるのがミソ

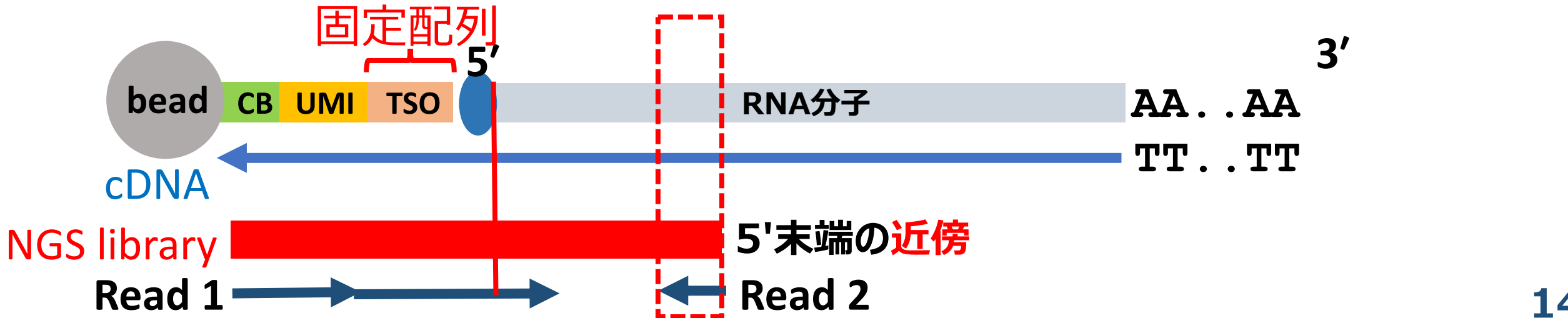
3'キットの場合



■ 10x Cell Barcode (CB)

■ Unique molecular identifier (UMI)

5'キットの場合

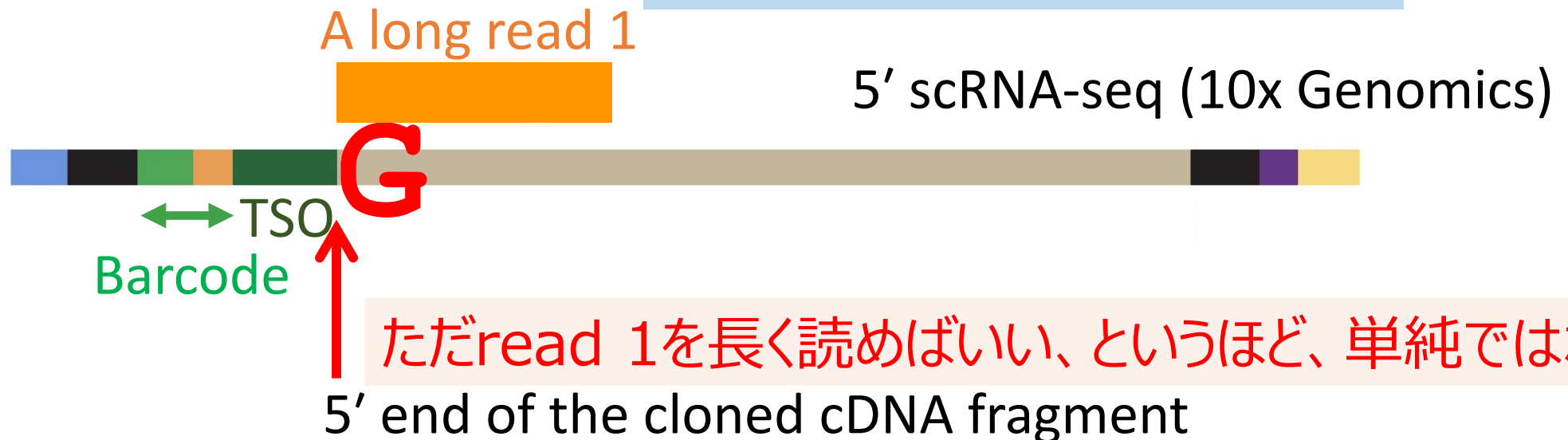


Read 1は意外に長く読める (開発秘話)

- 共同研究先のライブラリのシーケンスを外注。
- 通常はRead 2は長めに読み、Read 1は細胞バーコードや分子バーコード(UMI)の20数塩基だけ読む。
- しかし、**たまたま150 bp PE**でシーケンスされたものが納品。
- Read 1のシーケンスクオリティが予想よりは良かった！

- 10x Barcode (16 nt)
- UMI (10 or 12 nt)

DO NOT follow the protocol
2 × 150 bp paired-end run

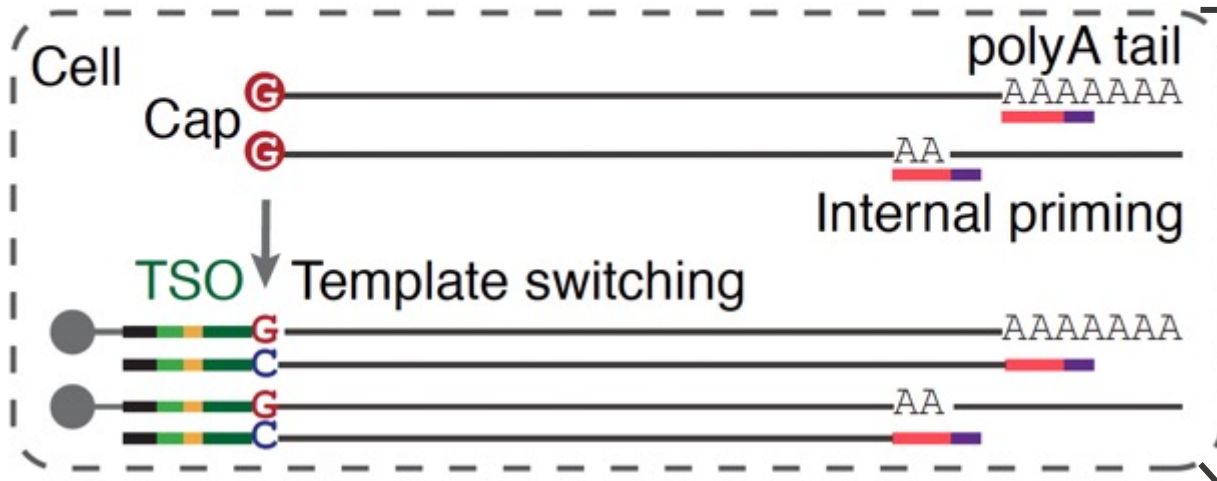


5' scRNA-seqを使って転写開始点(TSSs)を捉える

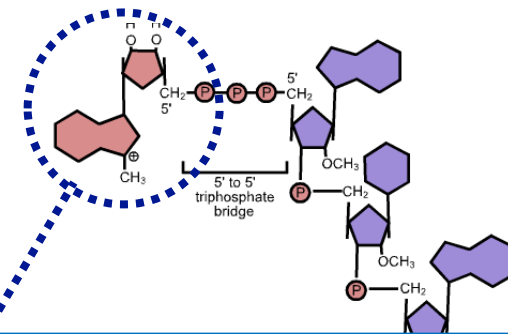
Point ① 5' Kitを使う

5' scRNA-seq
(10x genomics)

Chromium Next GEM Single Cell 5' Kit



m7G 5' end of RNA



Point ③ Cap構造も逆転写が起きる



Reverse transcription

Ohtake, H. *et al.* *DNA Res.* 2004

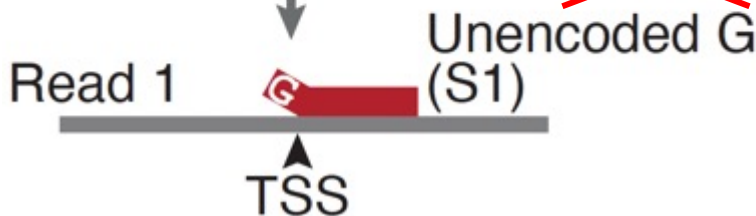
Library preparation

Read 1 (150 bp)

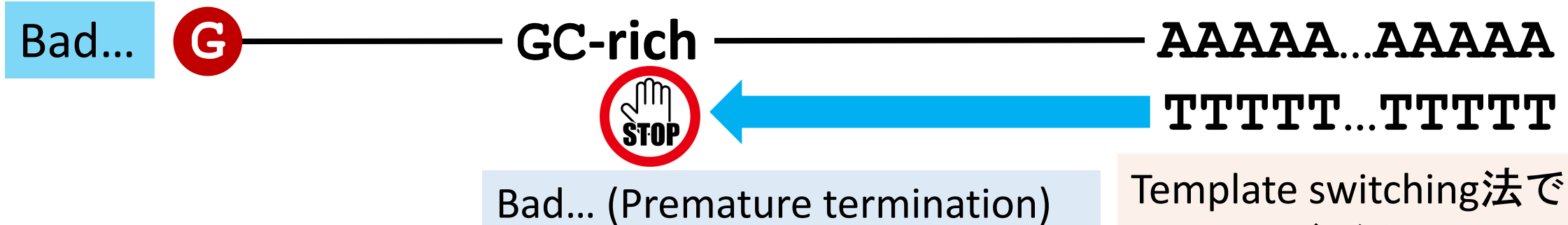
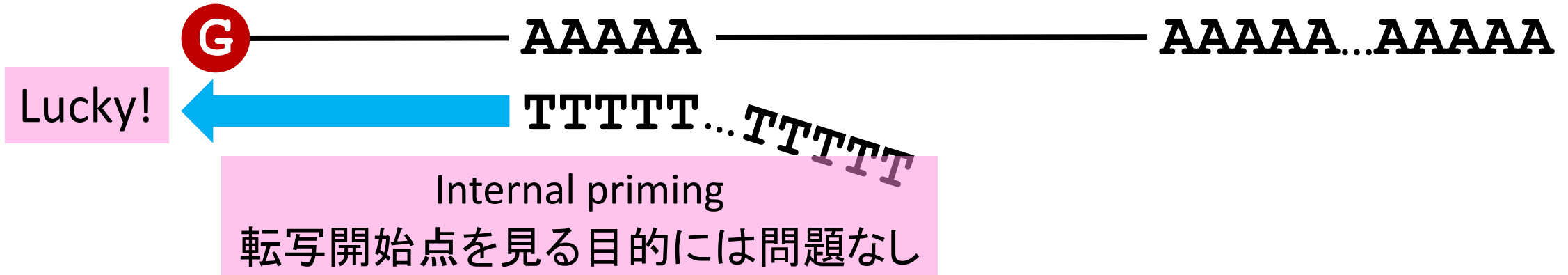
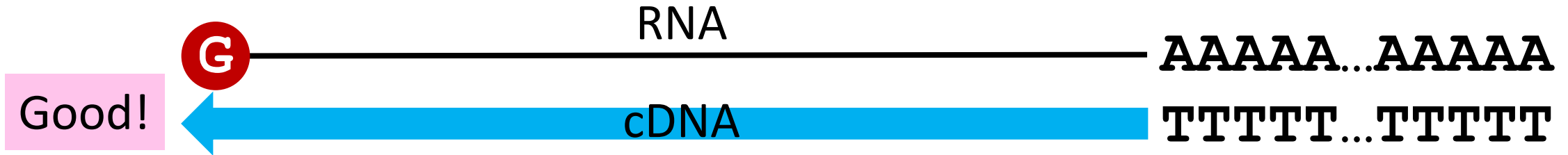
Point ② Read 1を長く読む (Don't follow the protocol)

~~Read 2~~

Point ④ ノイズリードを除去する (ReapTEC法)



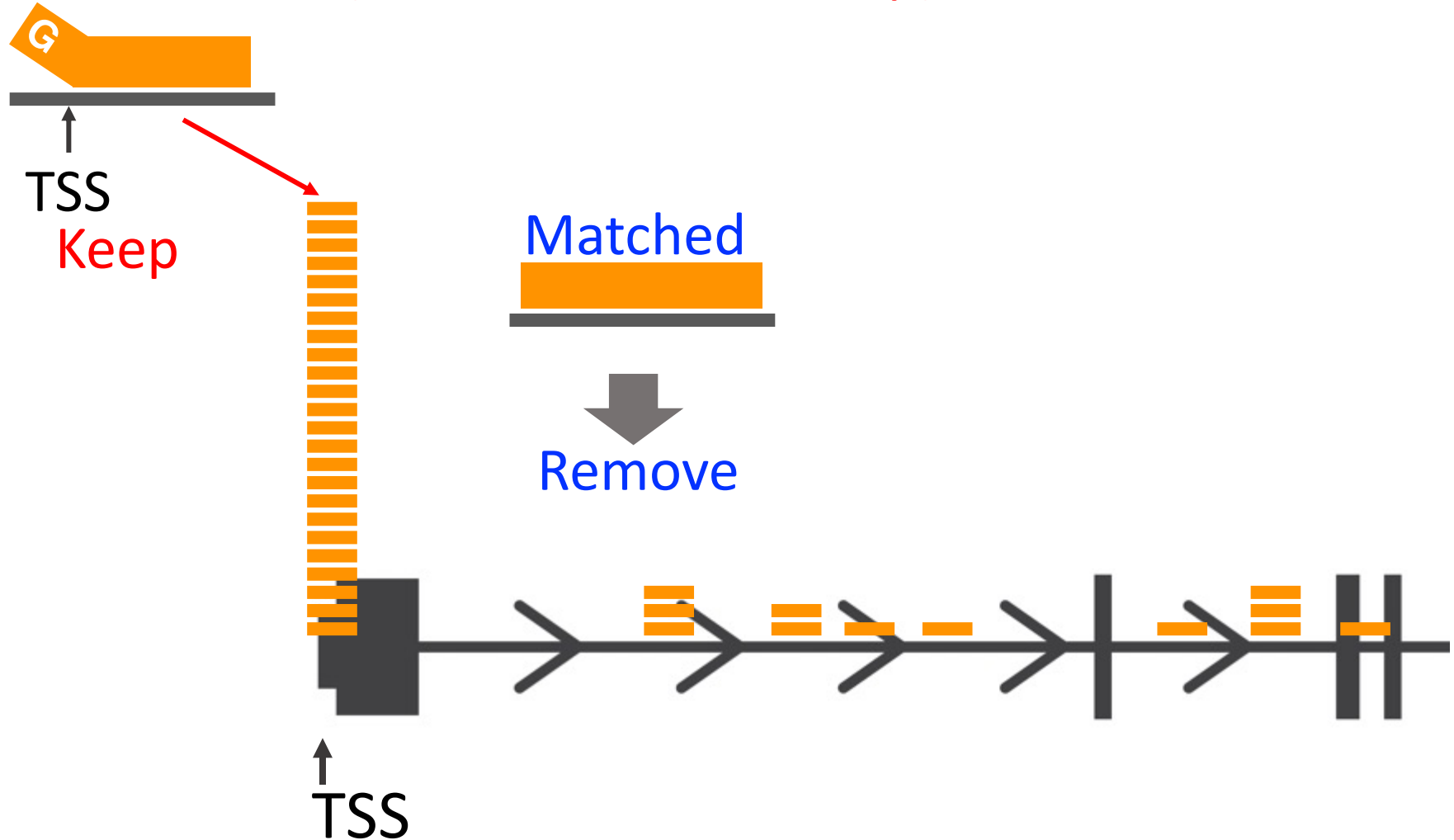
生リードには転写開始点を捉えていないノイズリードが生じる



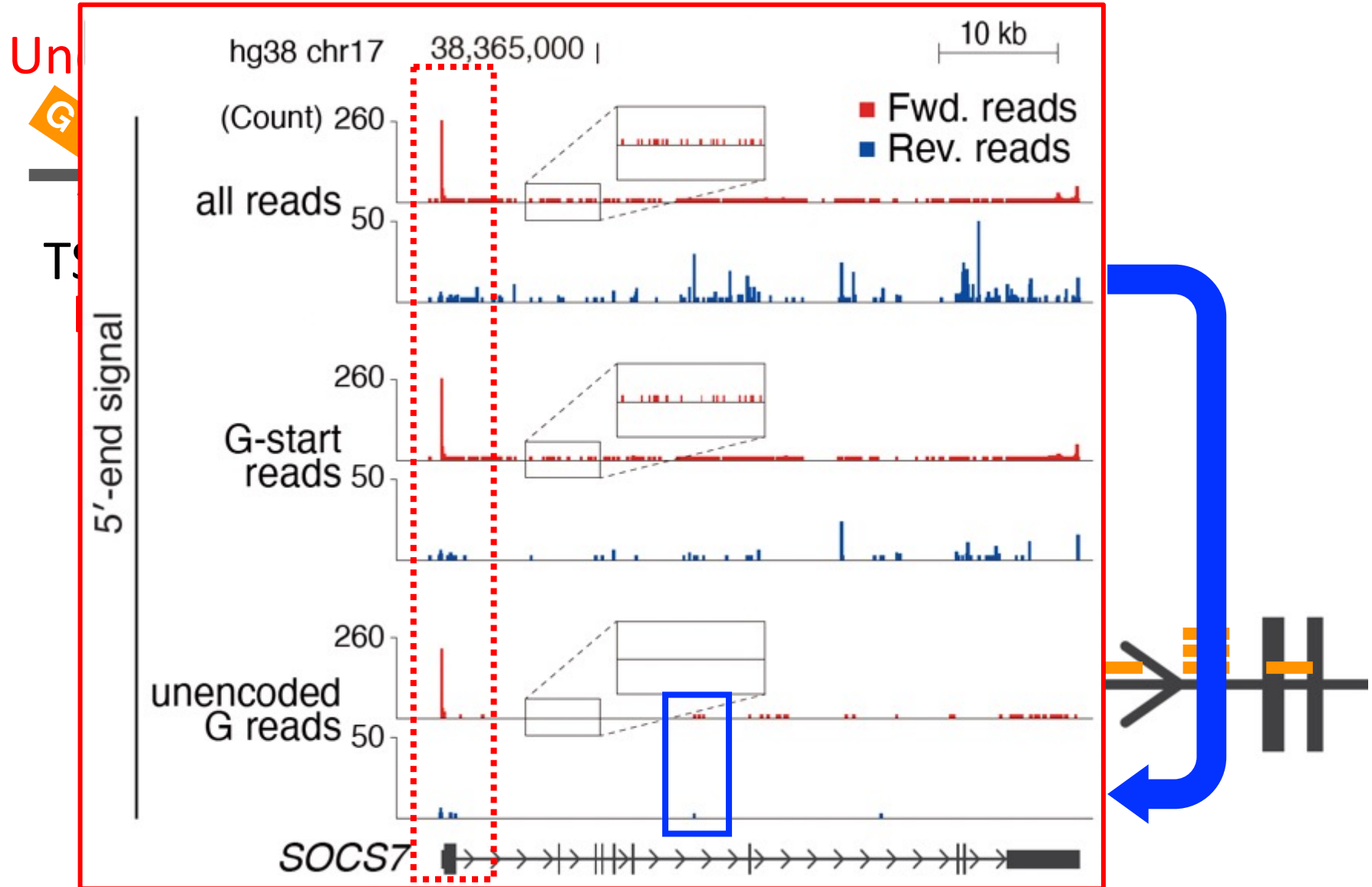
Template switching法では
このノイズが多い。

Cap構造由来のGを持つ (unencoded G)リードのみを残すことでリードをフィルタリングした

Unencoded G (derived from m7G cap)



Cap構造由来のGを持つ (unencoded G)リードのみを残すことでリードをフィルタリングした

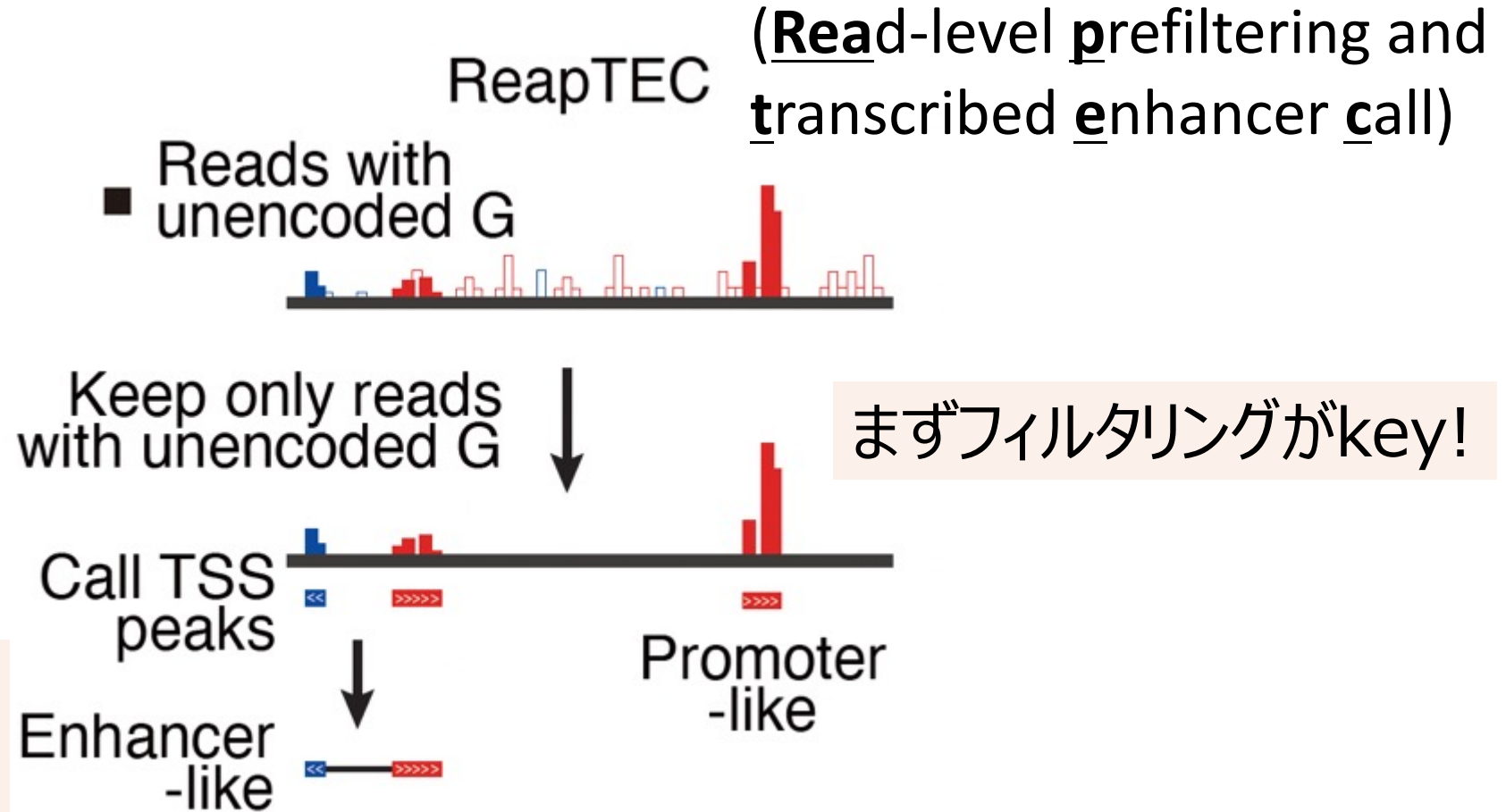


このフィルタリング方法をエンハンサーコールに活用する ReapTEC法を開発した

<https://github.com/MurakawaLab/ReapTEC>



両方向にeRNAを合成する
活性化エンハンサー領域



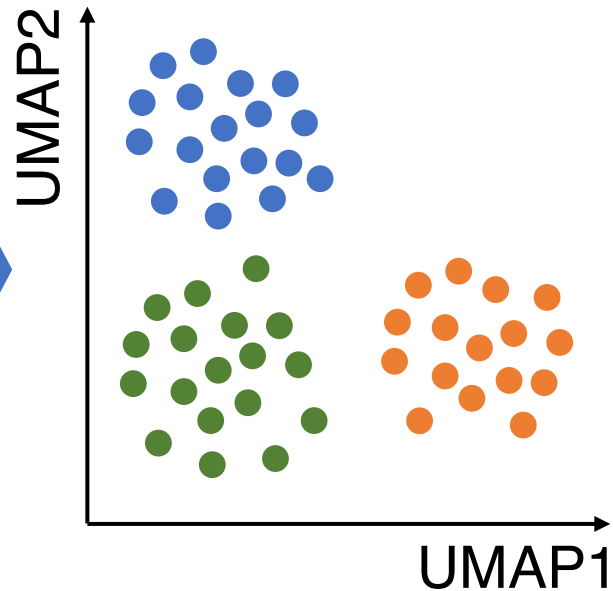
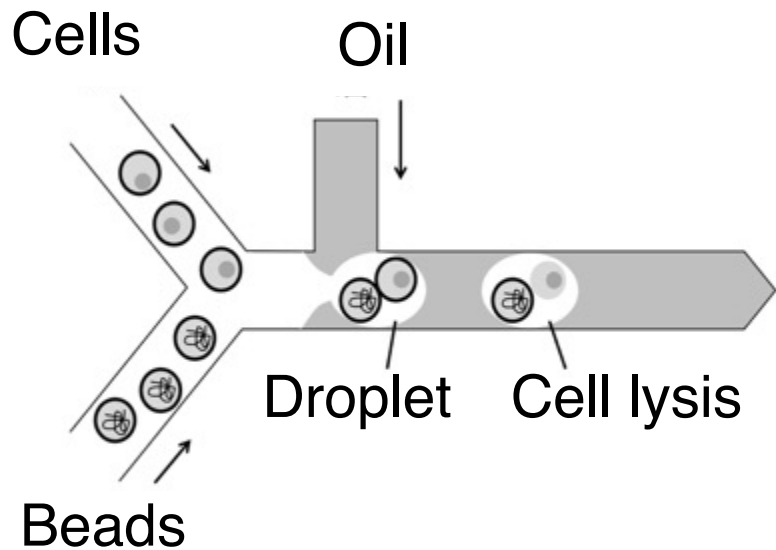
bidirectionally transcribed candidate enhancer (btcEnhs)






エンハンサーやATACのピークはクラスターベースでコールした






1. Single-cell RNA-seq
2. Single-nuclei ATAC-seq






Cell-type clustering using gene expression

Pseudo-bulk analysis to identify transcribed enhancers and ATAC-seq peaks

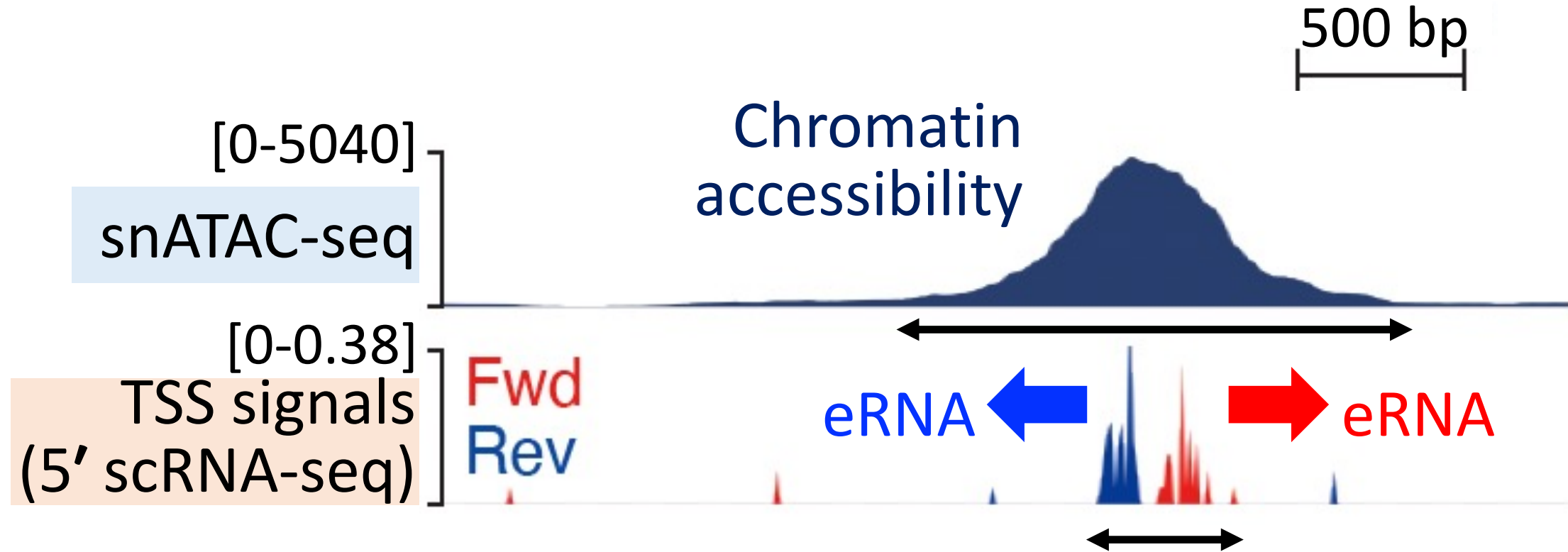


Cell-type A  =  +  +  + 

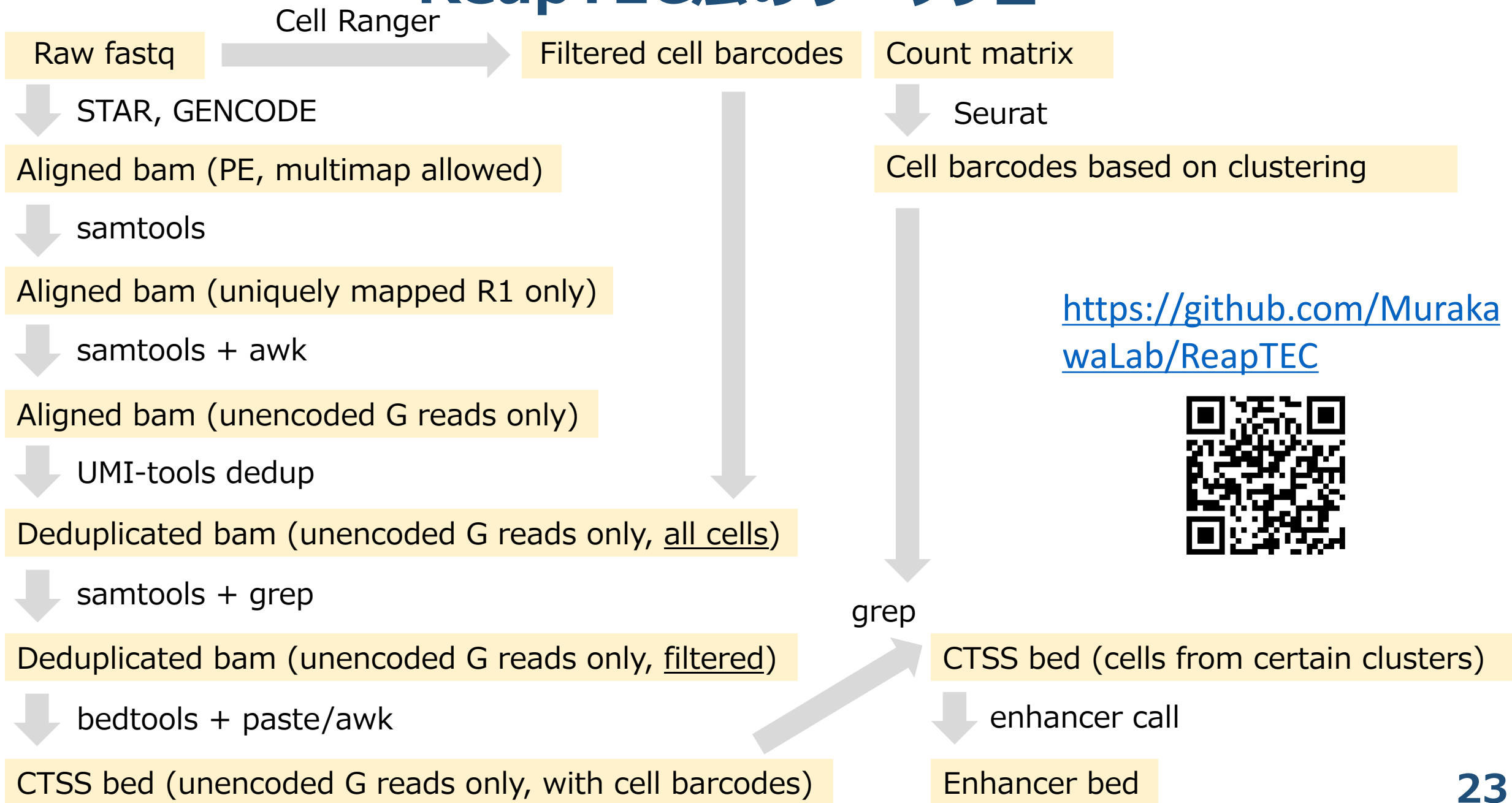
Cell-type B  =  +  +  + 

Cell-type C  =  +  +  + 

エンハンサー-RNAを捉えることで高い塩基解像度で エンハンサー領域を同定できる



ReapTEC法のワークフロー



使用するソフトウェア（バージョンは動作確認済みのもの）

- Cell Ranger v7.0.0 (<https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest>)
- STAR v2.7.10a (<https://github.com/alexdobin/STAR>)
- umi-tools v1.1.2 (<https://github.com/CGATOxford/UMI-tools/releases/tag/1.1.2>)
- samtools v1.15.1 (<http://www.htslib.org/>)
- bedtools v2.30.0 (<https://github.com/arq5x/bedtools2>)
- Seurat v4.3.0 (<https://satijalab.org/seurat/>)

1. count matrix を生成する

通常解析通りにCell Rangerのcellranger countを用いてcount matrixのデータを生成する。

Shell

```
$ cellranger count --id=Sample1_Cellranger \  
--fastqs=Sample1 \  
--transcriptome=reference_data
```

2.有効な細胞バーコードのリストを作成する

①で得られた“filtered_feature_bc_matrix”ディレクトリにはCell Rangerのフィルタリングを通過した細胞のデータが格納されている。同ディレクトリ内のbarcodes.tsv.gzから細胞バーコードのリストを作成する。

Shell

```
#各行の-1の文字を削除して出力する  
$ zcat barcodes.tsv.gz \  
| sed -e 's/-1//g' > Sample1_whitelist.txt
```

3. FASTQファイルを別途マッピングする

ステップ1ですでにBAMファイルを得ているが、Cell Ranger側の仕様により、直接転写開始点解析に使用できないため、改めて生のFASTQファイルをインプットとしてSTAR (STARsolo) でpaired-endマッピングを行う。

Shell

```
$ STAR --runThreadN 12 --genomeDir index/ \  
--readFilesIn Sample1_S1_R1_001.fastq.gz Sample1_S1_R2_001.fastq.gz \  
--soloCBwhitelist Sample1_whitelist.txt \  
--soloBarcodeMate 1 --clip5pNbases 39 0 \  
--readFilesCommand zcat --soloType CB_UMI_Simple \  
--soloCBstart 1 --soloCBlen 16 --soloUMIstart 17 --soloUMIlen 10 \  
--soloStrand Reverse --outFileNamePrefix Sample1_ \  
--outSAMtype BAM SortedByCoordinate \  
--soloCBmatchWLtype 1MM_multi_Nbase_pseudocounts \  
--soloUMIdedup 1MM_Directional_UMItools \  
--outSAMattributes NH HI nM AS CR UR CB UB GX GN sS sQ sM
```

3. FASTQファイルを別途マッピングする (5' GEM-X ver.)

5' GEM-Xのキットを使っている場合は、ステップ3で以下のように修正が必要です。
(UMIの長さが変更になったため。)

Shell

```
$ STAR --runThreadN 12 --genomeDir index/ \  
--readFilesIn Sample1_S1_R1_001.fastq.gz Sample1_S1_R2_001.fastq.gz \  
--soloCBwhitelist Sample1_whitelist.txt \  
--soloBarcodeMate 1 --clip5pNbases 41 0 \  
--readFilesCommand zcat --soloType CB_UMI_Simple \  
--soloCBstart 1 --soloCBlen 16 --soloUMIstart 17 --soloUMIlen 12 \  
--soloStrand Reverse --outFileNamePrefix Sample1_ \  
--outSAMtype BAM SortedByCoordinate \  
--soloCBmatchWLtype 1MM_multi_Nbase_pseudocounts \  
--soloUMIidedup 1MM_Directional_UMItools \  
--outSAMattributes NH HI nM AS CR UR CB UB GX GN sS sQ sM
```

4. uniquely mapped readのRead 1のみを抽出する

Samtoolsを用いてステップ3で得られたBAMファイルから一箇所でのみ（ユニークに）マッピングされたリード（uniquely mapped read）を取得する（-q 255）。それと同時に転写開始点の情報を含んでいるread 1のみを抽出する（-f 64）。

Shell

```
$ samtools view -@ 12 -hbf 64 -q 255 \  
Sample1_Aligned.sortedByCoord.out.bam \  
> Sample1_unique_R1.bam
```


5. 真の転写開始点情報を持つリードを抽出する (unencoded G filtering)

リファレンスゲノムに存在しない、キャップ由来の「G」(unencoded G) を持つリード (= 真の転写開始点の情報を持つ) を取得する。Read 1の先頭から39塩基は細胞バーコード(16塩基)、UMI (10塩基)、TSO配列 (13塩基) 由来であるため、unencoded Gを持つリードでは40塩基目に「G」が存在する (reverse strandではC) 。Read 1のこの先頭40塩基がリファレンスゲノムにマッピングされない場合、これらはSTAR (STARsolo)でマッピングすると「soft-clip (S) 」と認識される。BAMファイルの6列目を確認すると「40S***」(forward strand) あるいは「***M40S」(reverse strand) と表記されている。この情報を元に、unencoded Gを持つリードを抽出する。

注意：5' GEM-Xでは、UMIが12塩基のため、「42S***」(forward strand) あるいは「***M42S」(reverse strand) と表記される。
ステップ5も一部修正必要。

5. 真の転写開始点情報を持つリードを抽出する (unencoded G filtering)

Shell

```
$ samtools view -@ 12 -H Sample1_unique_R1.bam \  
> Sample1_header.sam  
  
$ samtools view -@ 12 -F 16 Sample1_unique_R1.bam \  
| awk -F '\t' 'BEGIN {OFS="\t"} {BASE = substr($10, 40, 1); \  
if ($6 ~ /^40S[0-9]/ && BASE == "G") {print $0}}' \  
> Sample1_SoftclipG_F.sam #40S***の場合  
  
$ samtools view -@ 12 -f 16 Sample1_unique_R1.bam \  
| awk -F '\t' 'BEGIN {OFS="\t"} { ALT = substr($10, length($10)-39, 1); \  
if ($6 ~ /[0-9]M40S$/ && ALT == "C") {print $0}}' \  
> Sample1_SoftclipG_R.sam #***M40Sの場合  
  
$ cat Sample1_header.sam \  
Sample1_SoftclipG_F.sam \  
Sample1_SoftclipG_R.sam \  
| samtools sort -@ 12 -O bam -o SoftclipG_Sample1.bam
```

5. 真の転写開始点情報を持つリードを抽出する (unencoded G filtering, 5' GEM-X ver.)

Shell

```
$ samtools view -@ 12 -H Sample1_unique_R1.bam \  
> Sample1_header.sam  
  
$ samtools view -@ 12 -F 16 Sample1_unique_R1.bam \  
| awk -F '\t' 'BEGIN {OFS="\t"} {BASE = substr($10, 42, 1); \  
if ($6 ~ /^42S[0-9]/ && BASE == "G") {print $0}}' \  
> Sample1_SoftclipG_F.sam #42S***の場合  
  
$ samtools view -@ 12 -f 16 Sample1_unique_R1.bam \  
| awk -F '\t' 'BEGIN {OFS="\t"} { ALT = substr($10, length($10)-41, 1); \  
if ($6 ~ /[0-9]M42S$/ && ALT == "C") {print $0}}' \  
> Sample1_SoftclipG_R.sam #***M42Sの場合  
  
$ cat Sample1_header.sam \  
Sample1_SoftclipG_F.sam \  
Sample1_SoftclipG_R.sam \  
| samtools sort -@ 12 -O bam -o SoftclipG_Sample1.bam
```

6.重複しているリードを除去する

ライブラリー調整中にPCRを行っているため、シーケンスされたデータには細胞バーコードやUMIが全く同一の重複したリード (duplicate reads) が含まれている。UMI-toolsのdedupでは、細胞バーコードとUMI配列に基づいて重複したリードを除去することができる。

Shell

```
$ umi_tools dedup --per-cell \  
-I SoftclipG_Sample1.bam \  
--extract-umi-method=tag \  
--umi-tag=UR \  
--cell-tag=CR \  
-S SoftclipG_Sample1_deduplicated.bam
```

7. 細胞バーコードが一致していないリードを除去する

ステップ1で得られた細胞バーコードを持つ細胞のリードのみを抽出する。

Shell

```
#STARの出力に合わせて、バーコードのリストにCB:Zを付加する。
$ awk '{print "CB:Z:"$1}' Sample1_whitelist.txt \
  > Sample1_cell_barcode.txt

#header部分だけ抜き出す。
$ samtools view -@ 12 -H SoftclipG_Sample1_deduplicated.bam \
  > SAM_header

#bamファイルからバーコードのリストと一致するリードを抽出する。
$ samtools view -@ 12 SoftclipG_Sample1_deduplicated.bam \
  | LC_ALL=C grep -F -f Sample1_cell_barcode.txt \
  > filtered_SAM_body

#headerとメインファイルを再度統合する。
$ cat SAM_header filtered_SAM_body > filtered.sam

#samファイルをbamファイルに変換する。
$ samtools view -@ 12 -b filtered.sam \
  > SoftclipG_Sample1_filtered.bam
```

8. 転写開始点ごとのカウントファイル (CTSS.bedファイル) を作成する

転写開始点の情報を含んだリードから、bedtoolsのbamToBedを用いて転写開始点ごとのカウントファイルを取得する。細胞バーコードごとに転写開始点のリードをカウントすること (CTSS.fwd.rev.cell.barcode.bed) で、どの細胞由来の情報かが分かるため、後に目的の細胞のみを取り出した解析も可能となる。

8. 転写開始点ごとのカウントファイル (CTSS.bedファイル) を作成する

Shell

#バーコードをbamファイルから取得する。

```
$ samtools view -@ 12 SoftclipG_Sample1_filtered.bam \  
  | awk 'BEGIN{OFS="\t"}{print $23}' \  
  > Sample1_cell_barcode_tmp.txt
```

#bamファイルをbedファイルに変換する。

```
$ bamToBed -i SoftclipG_Sample1_filtered.bam \  
  | paste Sample1_cell_barcode_tmp.txt \  
  | awk 'BEGIN {OFS="\t"} \  
  { if($6=="+") {print $1, $2, $2+1, ".", $7, $6} \  
  else {print $1, $3-1, $3, ".", $7, $6}}' \  
  | sort -k1,1 -k2,2n -k6,6 \  
  | bedtools groupby -g 1,2,3,4,5,6 -c 1 -o count \  
  | awk 'BEGIN{OFS="\t"}{if ($1 ~ /chr/) print $1, $2, $3, $5, $7, $6}' \  
  > Sample1.CTSS.fwd.rev.cell.barcode.bed
```

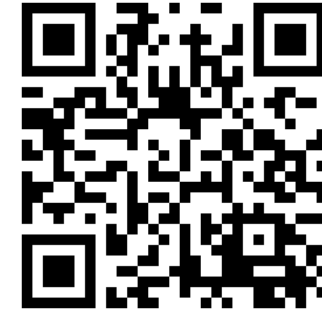
forwardストランドでは、転写開始点、転写開始点+1の位置が、reverseストランドでは、転写開始点、転写開始点-1の位置が記されたbedファイルを作成する。

バーコードごとに同一の転写開始点のリードをカウントする。

9.解析データの品質の確認を行う(オプション)

CTSS.fwd.rev.cell.barcode.bedを用いて、既知のプロモーター領域やエンハンサー領域と重なるリードをカウントし、発現解析を行う。各領域におけるリードのマッピング率やデータの相関性を見ることで、データの妥当性を確認する (sanity check) 。

10. 転写性エンハンサー領域を同定する



2014年に報告されたAndersson らのスクリプト

(<https://github.com/anderssonrobin/enhancers>) をベースに双方向性にeRNAが転写されるエンハンサー領域を同定する。

転写はある程度領域を持って生じるため、スクリプト内では、まず各リードの転写開始点 (tag) を集合させたクラスター (tag cluster) が作成される。

Anderssonらのオリジナルのスクリプトでは近接した20塩基以内のtagをtag clusterとしている)が、今回は10塩基以内のtagをtag clusterとしている。また、すでにunencoded G filteringを行いノイズリードは除去されているため、1 tagでもtag clusterと認めることで、エンハンサー領域の検出感度を高めている。

11. 目的の細胞集団の細胞バーコードを抽出する

Seuratなどを用いて細胞をクラスタリングし、各クラスターを構成する細胞の細胞バーコードを取得する。

R

```
Cluster0_object <- subset(Seurat_object, idents = 0)
Cluster0_cell_barcode <- colnames(Cluster0_object@assays$RNA@counts)
write.table(Cluster0_cell_barcode, "Cluster0_cell_barcode.txt", sep="\t", col.names=F, row.names=F, quote=F)
```

12. クラスター毎のCTSS.bedファイルを取得する

ステップ8で得られたファイルからステップ11で得たそれぞれのクラスターの細胞バーコードリストを用いて目的の細胞のデータを抽出する。

Shell

```
$ grep Sample1.CTSS.fwd.rev.cell.barcode.bed \  
    -f Cluster0_cell_barcode.txt \  
    > Sample1_cluster0.CTSS.fwd.rev.cell.barcode.bed
```

このデータにより、転写開始点解析やエンハンサー領域の同定(ステップ9)をシングルセル遺伝子発現解析で得られたクラスターごとに行うことができる。

ReapTEC法の留意点

たまたま転写開始点の一塩基上流のリファレンスゲノムが「G」であった場合、キャップ由来の「G」を持つリードもReapTEC法では除去される。

エンハンサーRNAは発現値が低いため、ノイズに紛れやすい。より確からしいリードだけ残すことで、より確からしいエンハンサー領域を同定するを目的としている。

基本的に転写はある一点から起こることはなく、ある程度の領域を持って起こるため、10塩基以内に存在するtagをtag clusterにまとめる過程（ステップ10）で、領域としては部分的にリカバリーされている。

Take home messages

5' scRNA-seqがオススメ

Read 1は長く読む

いつもの解析に加えてReapTEC法を使えば、

転写開始点解析、転写性エンハンサーの同定が可能

ReapTEC法を行うために、実験やシーケンスの追加料金は発生しません！



謝辞



RESEARCH ARTICLE

全ての著者と関係者の方々に御礼申し上げます。

IMMUNOLOGY

An atlas of transcribed enhancers across helper T cell diversity for decoding human diseases

Akiko Oguchi^{1,2,3†}, Akari Suzuki^{4†}, Shuichiro Komatsu^{1,5†}, Hiroyuki Yoshitomi^{2,6}, Shruti Bhagat², Raku Son^{1,2,3}, Raoul Jean Pierre Bonnal⁵, Shohei Kojima⁷, Masaru Koido^{8,9,10}, Kazuhiro Takeuchi^{1,2,11}, Keiko Myouzen⁴, Gyo Inoue⁴, Tomoya Hirai^{1,12}, Hiromi Sano¹, Yujiro Takegami¹³, Ai Kanemaru¹³, Itaru Yamaguchi¹³, Yuki Ishikawa¹⁰, Nao Tanaka¹⁰, Shigeki Hirabayashi^{1,14,15}, Riyo Konishi¹⁶, Sho Sekito^{2,17}, Takahiro Inoue¹⁷, Juha Kere^{18,19,20}, Shunichi Takeda^{21,22}, Akifumi Takaori-Kondo¹⁴, Itaru Endo¹², Shinpei Kawaoka^{16,23}, Hideya Kawaji^{24,25,26}, Kazuyoshi Ishigaki²⁷, Hideki Ueno^{2,6}, Yoshihide Hayashizaki^{13,26}, Massimiliano Pagani^{5,28}, Piero Carninci^{29,30}, Motoko Yanagita^{2,3}, ITEC Consortium[‡], Nicholas Parrish⁷, Chikashi Terao^{10,31,32*}, Kazuhiko Yamamoto^{4*}, Yasuhiro Murakawa^{1,2,5,11*}

番号は各々の所属機関、論文に掲載情報

