トランスクリプトーム解析の今昔
なぜマイクロアレイ？
なぜRNA-Seq？

東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
門田 幸二（かどた こうじ）
http://www.iu.a.u-tokyo.ac.jp/~kadota/
kadota@iu.a.u-tokyo.ac.jp

Contents

- トランスクリプトーム解析の概要
- 各手法の長所・短所
  - マイクロアレイ、RNA-Seq、RT-PCRやSAGE
- 実データの比較（RNA-Seq vs. マイクロアレイ）
- RNA-Seqデータの正規化（の基礎）
  - マイクロアレイと異なる点（遺伝子の配列長による結果の偏り）
  - 基本的な考え（RPKM）

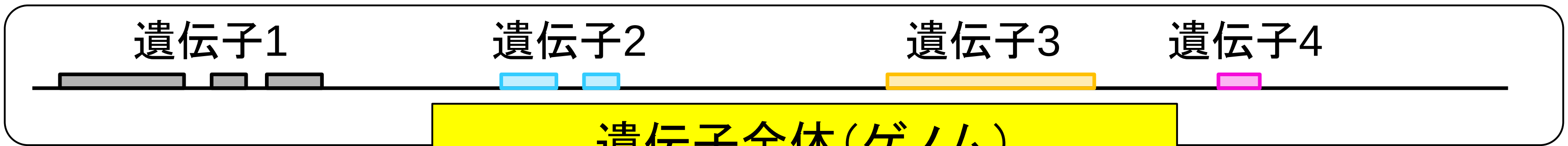
ねらい

各種トランスクリプトーム解析手法の長所、短所を理解し、その上でなぜ次世代シーケンサーによるトランスクリプトーム解析（RNA-Seq）が有用かを理解する

トランスクリプトームとは

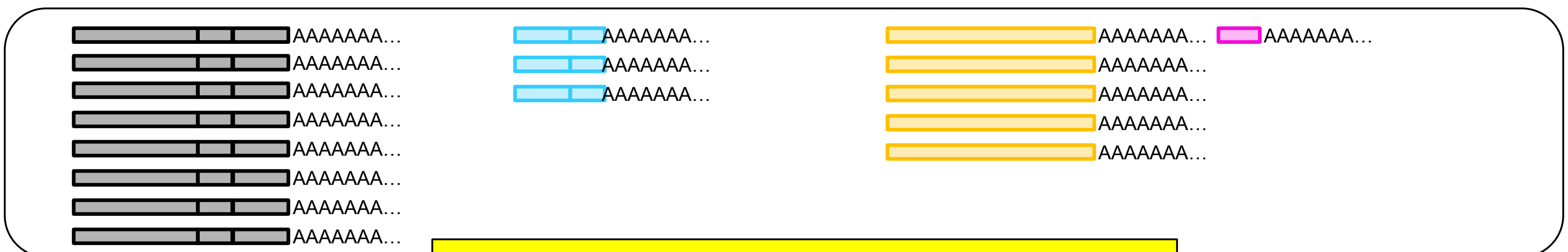
- ある状態のあるサンプル（例：目）のあるゲノムの領域

ヒト



遺伝子全体（ゲノム）

・どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）



転写物全体（トランスクリプトーム）

・遺伝子1は沢山転写されている（発現している）
・遺伝子4はごくわずかしか転写されてない
・…

トランスクリプトームとは

- ある状態のあるサンプル（例：目）のあるゲノムの領域

遺伝子全体（ゲノム）

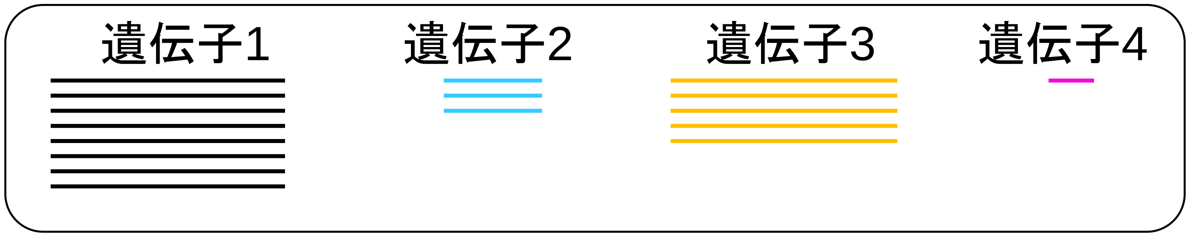
・どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）

転写物全体（トランスクリプトーム）

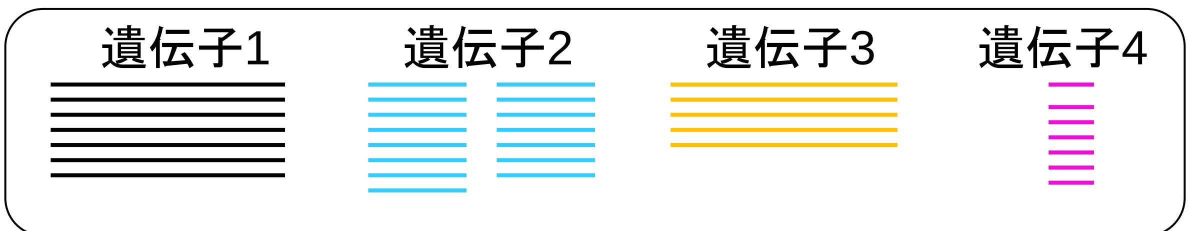
・遺伝子2は光刺激に応答して発現亢進
・遺伝子4も光刺激に応答して発現亢進

トランスクリプトーム情報を得る手段

- 光刺激前(T1)の目のトランスクリプトーム



- 光刺激後(T2)の目のトランスクリプトーム



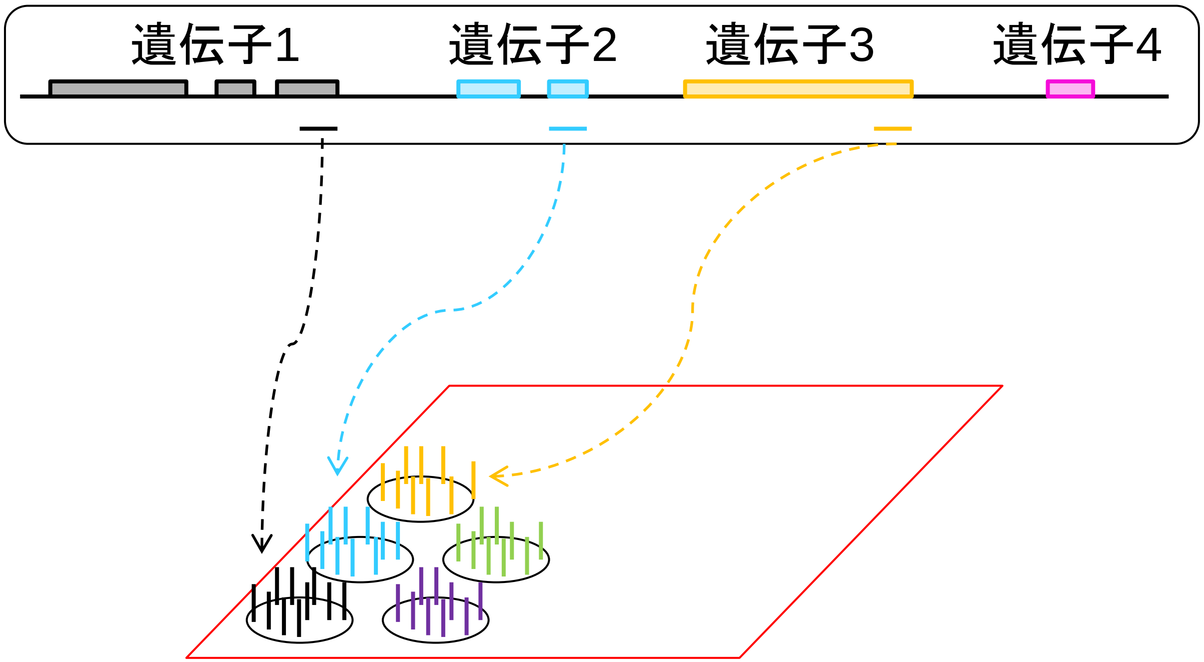
これがいわゆる
「遺伝子発現行列」

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

- マイクロアレイ
- RNA-Seq
- SAGE
- …

トランスクリプトーム取得（マイクロアレイ）

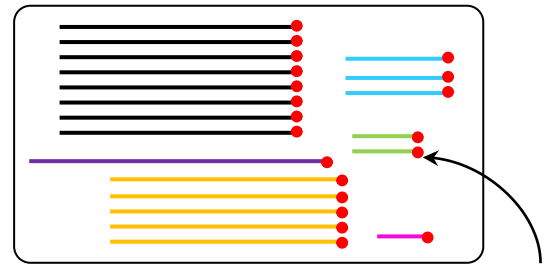
- よく研究されている生き物は多数の遺伝子（の配列情報）がわかっている



わかっている遺伝子（の配列の相補鎖）を搭載した”チップ”

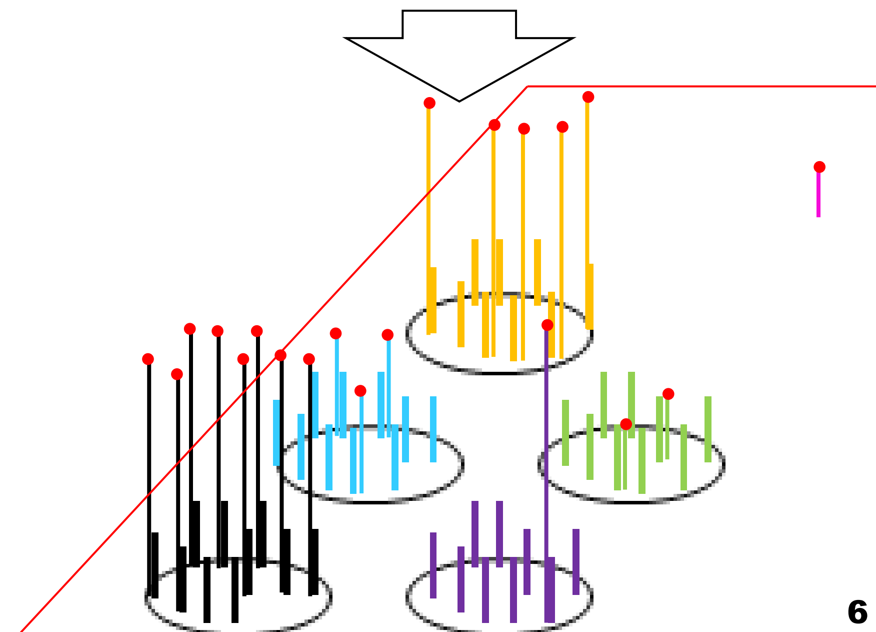
・メーカーによって搭載されている遺伝子の種類が異なる
→搭載されていない遺伝子（未知遺伝子含む、例：遺伝子4）の発現情報は測定不可…

光刺激前（T1）の目の
トランスクリプトーム



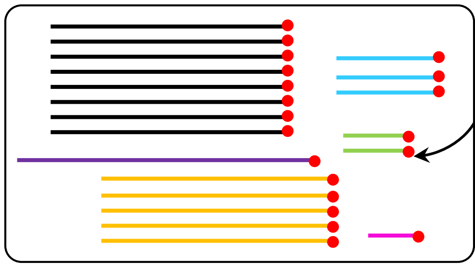
蛍光標識

ハイブリダイゼーション
（二本鎖形成）



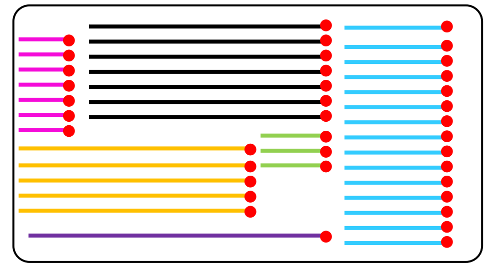
マイクロアレイデータ → 遺伝子発現行列

- 光刺激前(T1)の目のトランスクリプトーム



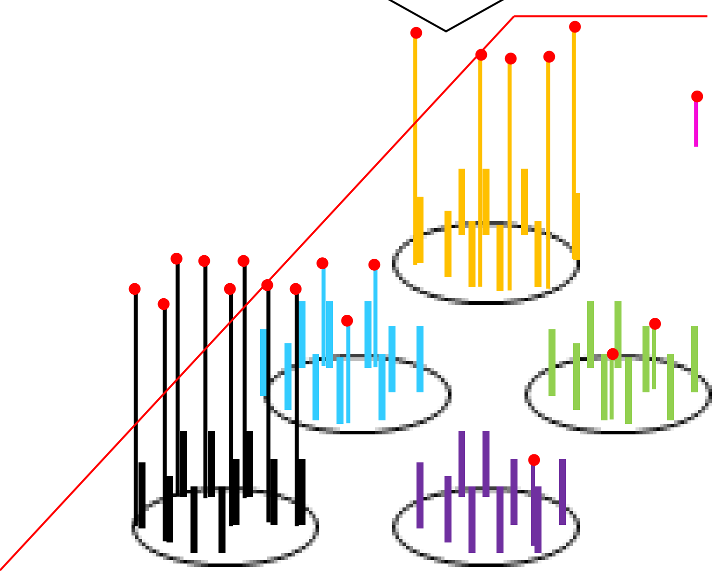
蛍光標識

光刺激後(T2)の目の
トランスクリプトーム



ハイブリダイゼーション
(二本鎖形成)

ハイブリダイゼーション
と
シグナル検出



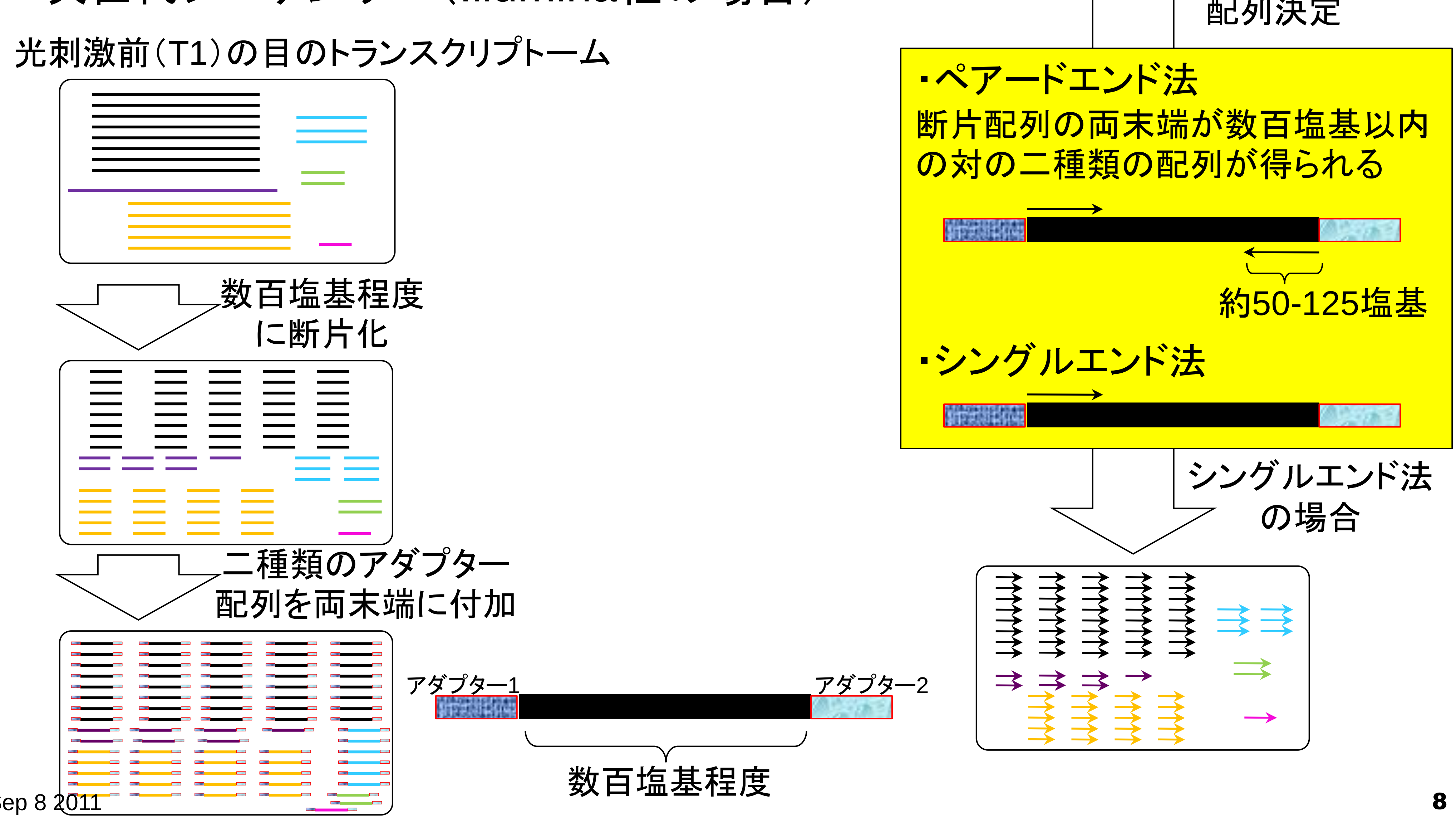
専用の検出器で各
遺伝子に対応する
領域の蛍光シグナ
ル強度を測定

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	?	?
遺伝子5	…	…
…	…	…

正規化

トランスクリプトーム取得(RNA-Seq)

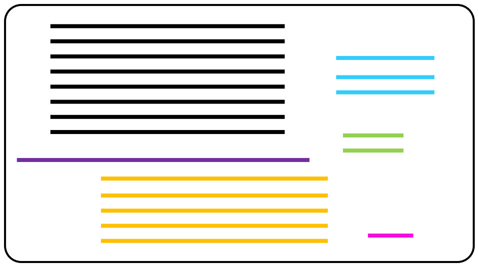
- 次世代シーケンサー(Illumina社の場合)



RNA-Seqデータ → 遺伝子発現行列

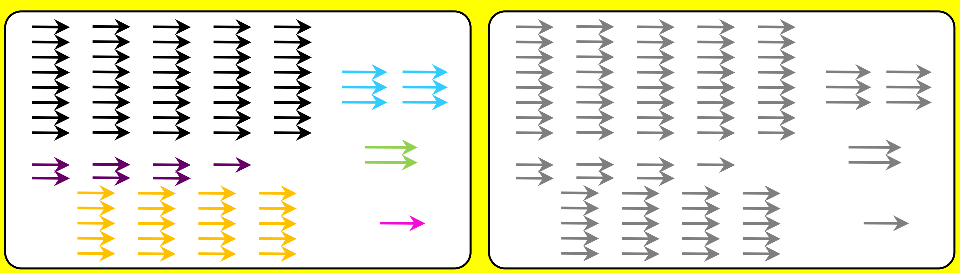
- RNA-seq

光刺激前(T1)の目のトランスクリプトーム



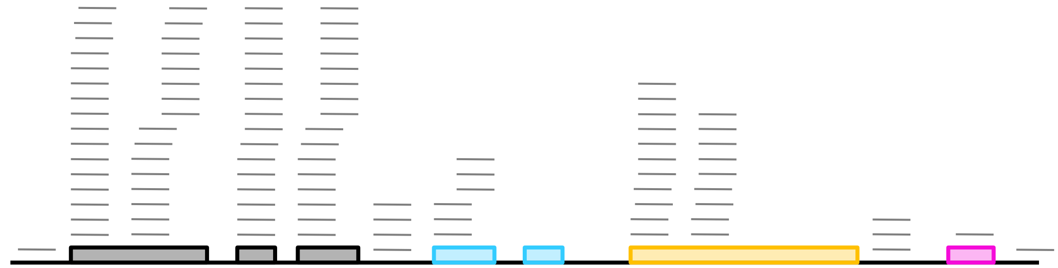
ーイメージー
50-125塩基程度からなる配列が沢山ある

ー実際ー
数百万個の配列があり、どの遺伝子に対応するか不明



(短い)配列を読んだものという意味で(ショート)リードなどと呼ばれる

ゲノム配列にマッピング



定量化(例：生のリード数をカウント)

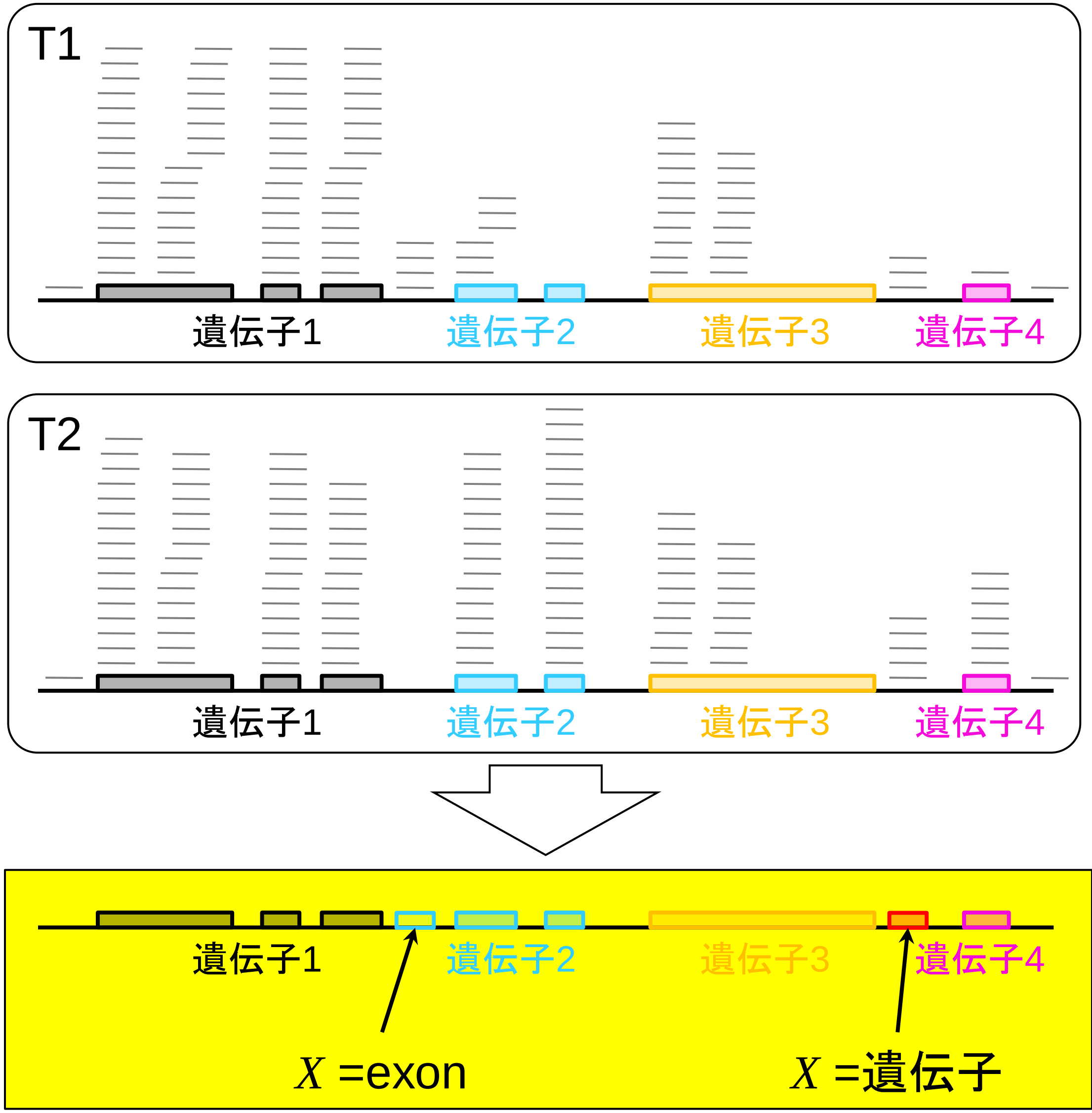
	T1
遺伝子1	40
遺伝子2	6
遺伝子3	20
遺伝子4	1
遺伝子5	…
…	…

正規化

T1
8
3
5
1
…
…

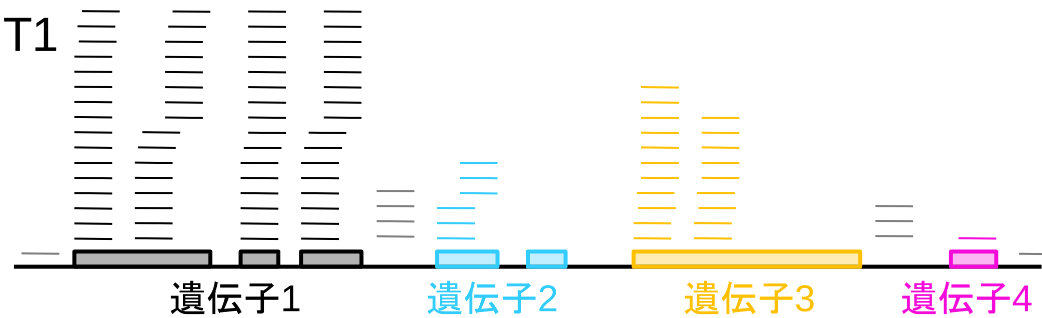
RNA-Seqの長所

- 新規Xの同定
  - X =exon, 遺伝子, …



RNA-Seqの長所

- 新規Xの同定

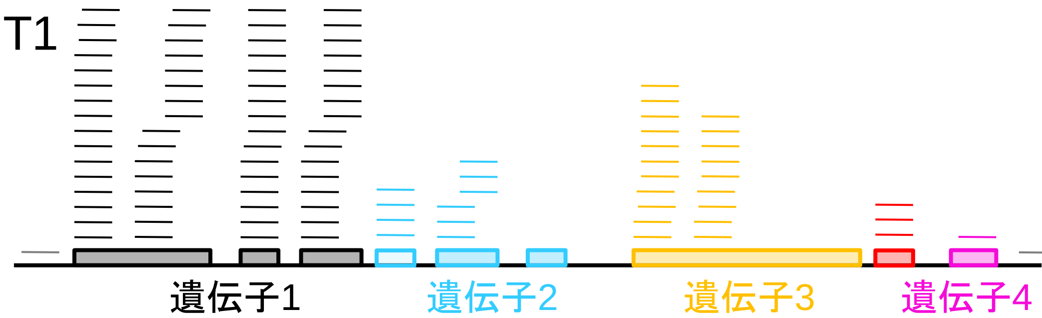


生リード数カウント →

	T1
遺伝子1	40
遺伝子2	6
遺伝子3	20
遺伝子4	1
遺伝子5	…
…	…

正規化 →

T1
8
3
5
1
…
…



生リード数カウント →

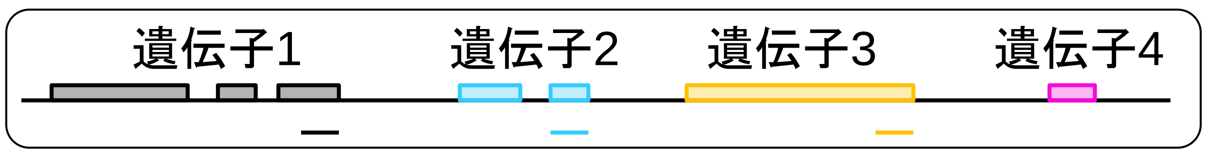
	T1
遺伝子1	40
遺伝子2	10
遺伝子3	20
新規遺伝子	3
遺伝子4	1
遺伝子5	…
…	…

正規化 →

・“トランスクリプトーム（転写物の全体像）”の理解への一番の近道
・よりよい**遺伝子発現行列**を得るための基礎情報充実に貢献

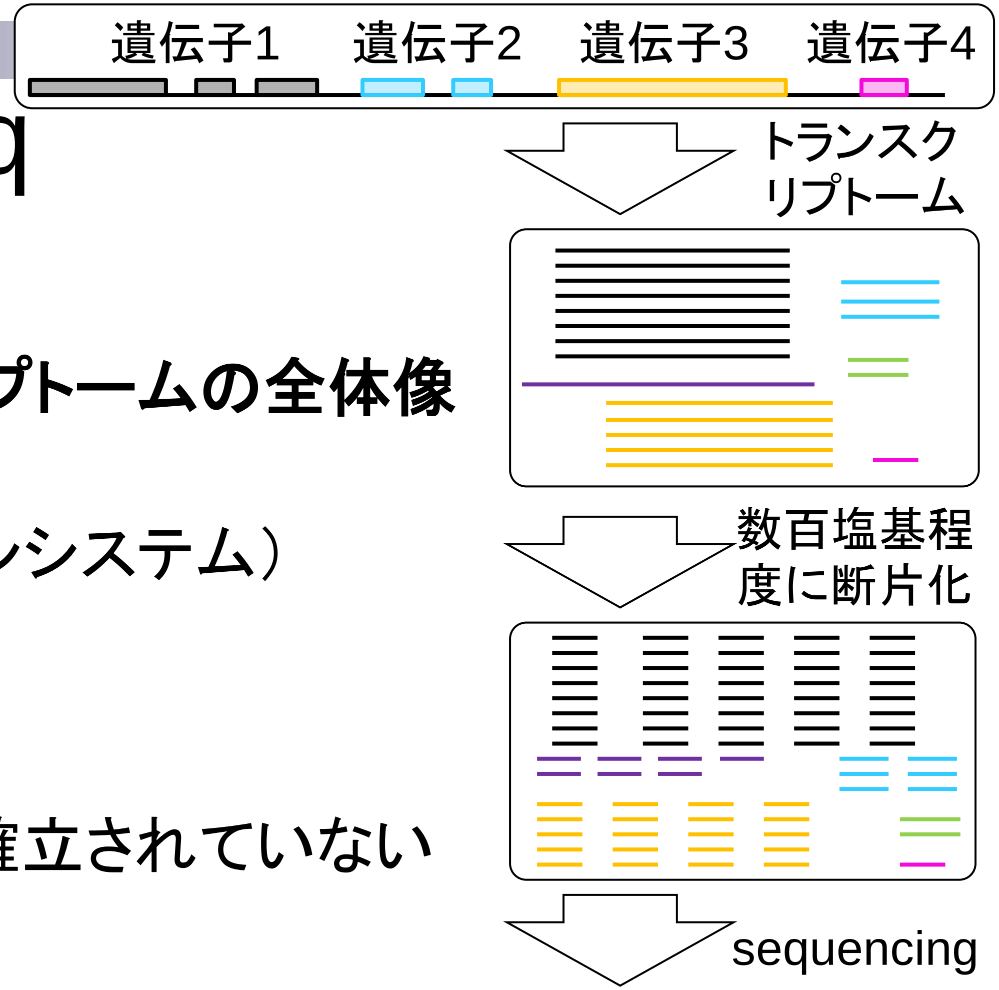
長所・短所：（発現解析用）マイクロアレイ

- 長所
  - すでに診断用マイクロアレイが市販されているなど**長年の実績**
  - お手軽、各種データ解析ツールが豊富
- 短所
  - （プローブ搭載のために）解析対象の塩基配列情報を予め知っておく必要がある。（クローズドシステム）
  - プローブが搭載されていない遺伝子の発現レベルは測定不可能（未知遺伝子も当然対象外）



- 主なユーザー
  - 主な解析対象が（アノテーション情報が豊富な）モデル生物で、既知遺伝子のみでいい、という研究者

長所・短所：RNA-Seq



- 長所
  - （未知遺伝子を含む）**トランスクリプトームの全体像を理解することが原理的に可能**
  - 事前情報を必要としない（オープンシステム）
  - ダイナミックレンジが広い
- 短所
  - データ解析が大変、解析手法が確立されていない
- 主なユーザー
  - 無制限（モデル生物・非モデル生物を問わない）
  - （お金持ち…）

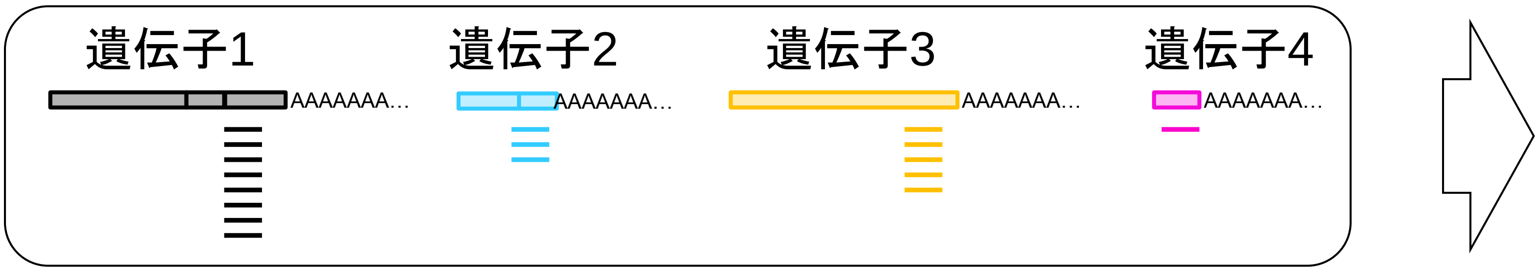
長所・短所：RT-PCR

- 長所?!
  - このテクノロジーで得られた測定結果が「最も信頼性が高いはず!（ゴールドスタンダード）」と多くのbiologistが思っている…
- 短所
  - 用いたプライマー次第で結果が変わる
  - Low-throughput（RT-PCRでのトランスクリプトーム解析は事実上不可能）
- 主なユーザー
  - （論文を通すために）マイクロアレイ（やRNA-Seq）解析を行った結果得られた候補遺伝子群のうちのいくつかの発現を確認しておこうと思った研究者

Velculescu et al., *Science*, **270**:484-487, 1995

SAGE

- Serial Analysis of Gene Expressionの略
- mRNAの**3’末端に近い数十塩基**をSAGEタグとして配列決定



	T1
遺伝子1	8
遺伝子2	3
遺伝子3	5
遺伝子4	1
…	…

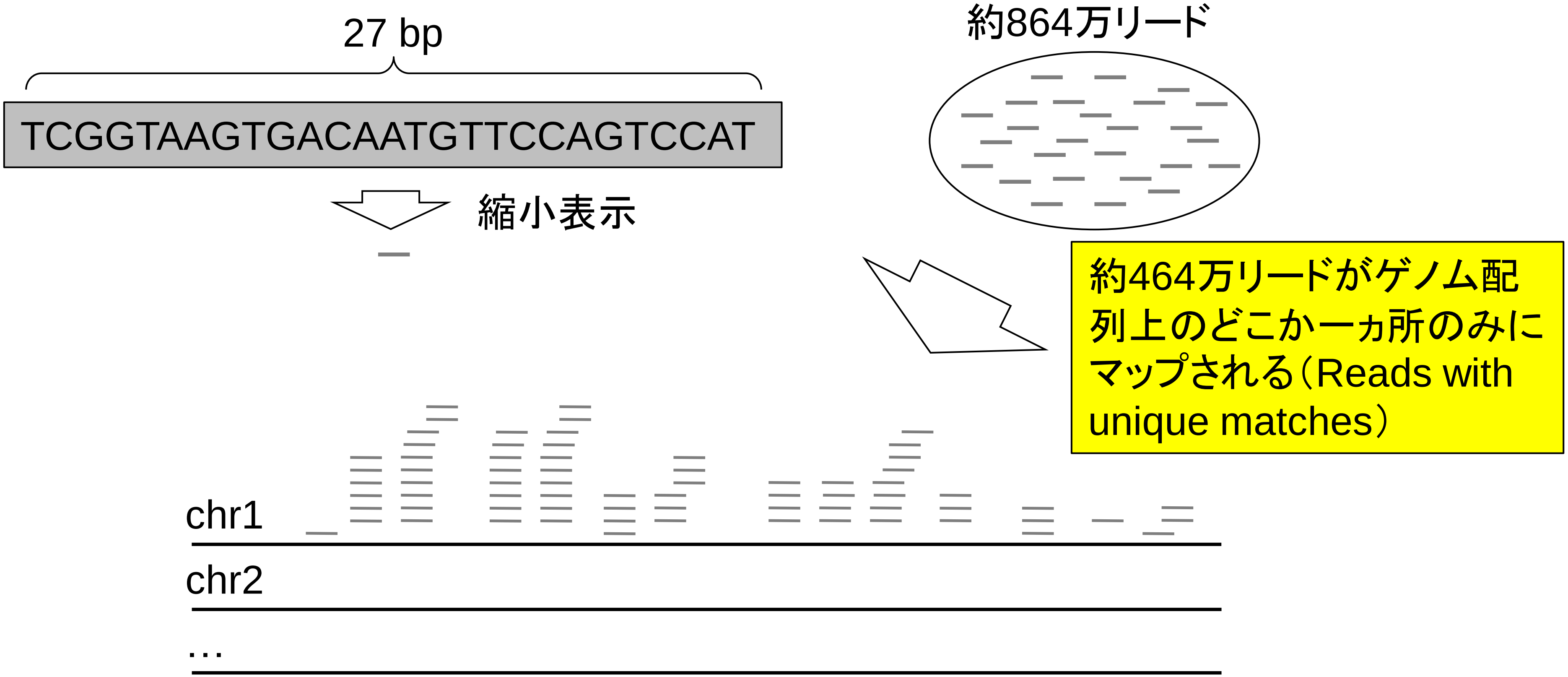
- 様々な改良版
  - 21bp読めるLongSAGE (Saha et al., *Nature Biotechnol*., 2002)
  - 26bp読めるSuperSAGE (Matsumura et al., *Proc. Natl. Acad. Sci. USA*, 2003)
  - 5’-end SAGE (Hashimoto et al., *Nature Biotechnol*., 2004)
  - 約37bp (ditagとして)読めるDeepSAGE (Nielsen et al., *Nucleic Acids Res*., 2006)
  - NGS用SuperSAGE (HT-SuperSAGE; Matsumura et al., *PLoS One*, 2010)

長所・短所：SAGE（RNA-Seqとの対比）

- 長所
  - （転写物の一部に特化しているので原理的に）ダイナミックレンジが広い
  - リード長がほぼ一定のため、RNA-Seqで問題となる「解析結果の配列長依存性（gene length-related bias）」とは無縁（後述）
- 短所
  - （転写物の一部に特化しているが故に原理的には）トランスクリプトームの全体像の理解は不可能（例：選択的スプライシング）
  - （制限酵素で切断しているので）制限酵素認識部位（NlaIIIなど）を持たない転写物の測定は困難
- 主なユーザー
  - 上記の長所を重要視する研究者

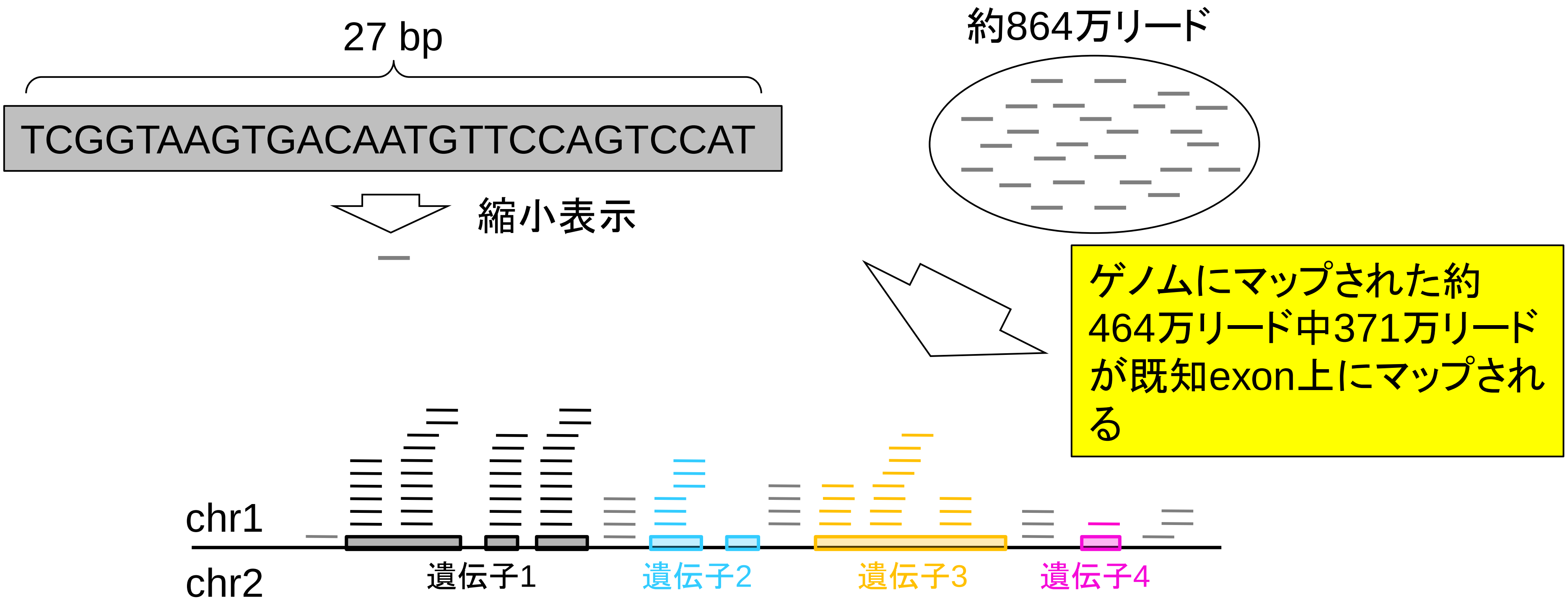
実データの比較（RNA-Seq vs. マイクロアレイ）

- Human embryonic kidney (HEK) 293T cells（とB cells）
  - **マイクロアレイ**: Illumina HumanRef8 V2.0 BeadChips
  - **RNA-Seq**: Illumina 1G Genome Analyzer



実データの比較（RNA-Seq vs. マイクロアレイ）

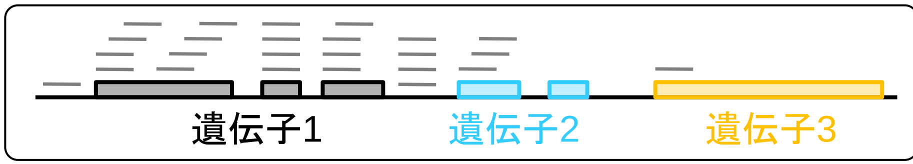
- Human embryonic kidney (HEK) 293T cells（とB cells）
  - **マイクロアレイ**：Illumina HumanRef8 V2.0 BeadChips
  - **RNA-Seq**: Illumina 1G Genome Analyzer



・既知エクソン領域以外にマップされたものは新規exonの可能性！
・大抵のマイクロアレイとの比較はアレイ上に搭載されている既知遺伝子についてのみ!

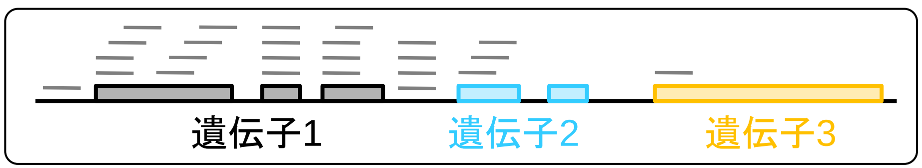
実データの比較（RNA-Seq vs. マイクロアレイ）

- マイクロアレイ上に搭載されている13,118遺伝子について、「発現している」とされた遺伝子数の比較
  - 閾値緩め（≧1 read）の場合



発現してる　　してる　　してる

  - 閾値厳しめ（≧5 read）の場合



（発現してる）　（してない）　（してない）

発現遺伝子数（HEK細胞の結果）

	RNA-Seqのみ	共通	マイクロアレイのみ
閾値緩め	2,918	9,238	110
閾値厳しめ	1,419	8,574	774

RNA-seqでのみ発現している遺伝子数 >> マイクロアレイでのみ

実データの比較(RNA-Seq vs. マイクロアレイ)

- 「HEK cells versus B cells」のlog ratio分布の比較

7,043 genes

Sultan et al., *Science*, **321**:956-960, 2008のFig. 2C

発現している:○, 発現してない:×

	マイクロアレイ			RNA-Seq			
	B cells	HEK cells	log比計算	B cells	HEK cells	log比計算	Plot
gene1	○	○	○	○	○	○	○
gene2	○	○	○	○	×	×	×
gene3	×	○	×	○	○	○	×
gene4	○	○	○	○	○	○	○
gene5	×	○	×	×	○	×	×
gene6	○	○	○	○	○	○	○
gene7	○	○	○	○	○	○	○
gene8	×	×	×	×	×	×	×
gene9	○	○	○	○	○	○	○
gene10	×	×	×	○	○	○	×
gene11	○	○	○	○	○	○	○
...							

全体として高発現側の遺伝子群の発現レベルは似ている

他の比較結果（RNA-Seq vs. マイクロアレイ）

- log ratio分布の比較（横軸：RNA-Seq, 縦軸：マイクロアレイ）

Mane et al., *BMC Genomics*, 2009のSuppl. Fig.の下半分

Griffith et al., *Nat Methods*, 2010のSuppl. Fig. 9b(A)

どの報告結果もだいたいこんな感じです

他の比較結果（RNA-Seq vs. マイクロアレイ）

- 発現量レベルの比較
  - LiverサンプルのRNA-Seqデータ vs. マイクロアレイデータ

Mortazavi et al., *Nat Methods*, 2008のFig. 3c

RPKM?

マイクロアレイデータの正規化

- 「各サンプルから測定されたシグナル強度の和は一定」と仮定
  - チップ上の遺伝子数が少ない場合は非現実的だが、数千～数万種類の遺伝子が搭載されているので妥当(だろう)

	sample1	sample2
gene1	10.5	12.4
gene2	6.4	7.1
gene3	8.0	8.5
gene4	10.8	11.4
gene5	5.6	6.7
gene6	8.4	8.9
gene7	6.2	7.0
gene8	6.1	6.8
gene9	6.6	6.5
gene10	5.1	5.8
総和	73.7	81.1

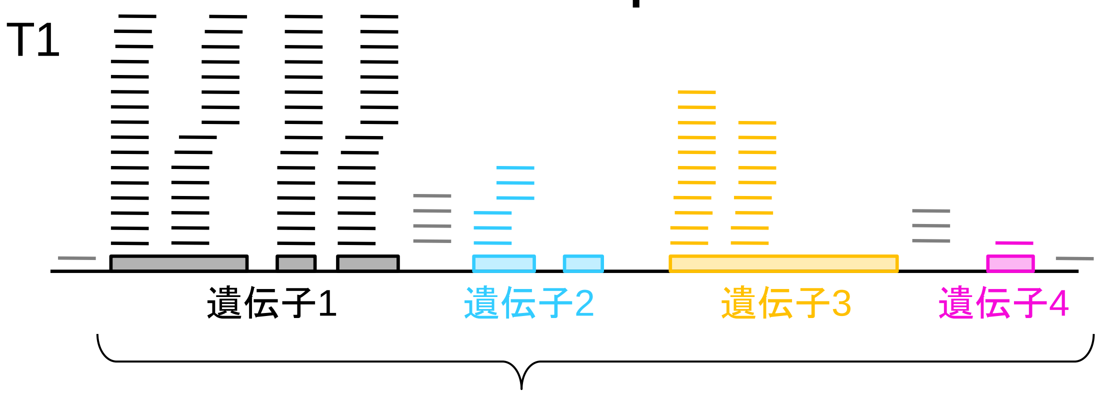
グローバル正規化

	sample1	sample2
gene1	14.2	15.3
gene2	8.7	8.8
gene3	10.9	10.5
gene4	14.7	14.1
gene5	7.6	8.3
gene6	11.4	11.0
gene7	8.4	8.6
gene8	8.3	8.4
gene9	9.0	8.0
gene10	6.9	7.2
総和	100.0	100.0

背景:サンプル(or chip)ごとにシグナル強度の総和は異なる
対策:総和が任意の値(例では100)になるような正規化係数を掛ける
例:sample1の正規化係数= 100 / **73.7**

RNA-Seqデータの正規化(の一部)

- 「各サンプルからsequenceされた総リード数は一定」と仮定



	T1	T2
遺伝子1	40	7
遺伝子2	6	15
遺伝子3	20	5
遺伝子4	1	1
総リード数	**67**	**28**

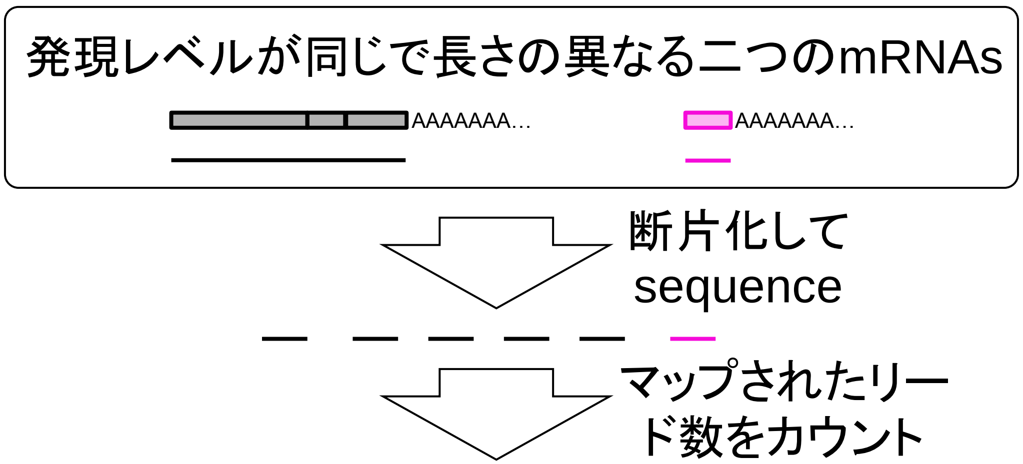
RPM正規化

	T1	T2
遺伝子1	597014.9	250000.0
遺伝子2	89552.2	535714.3
遺伝子3	298507.5	178571.4
遺伝子4	14925.4	35714.3
総リード数	1000000	1000000

Reads Per Million mapped reads(RPM)
正規化後の総リード数が100万(one million)になるように補正
例:T1の正規化係数 = 1000000 / **67**

配列長の補正

- 配列長が長い遺伝子ほど沢山sequenceされる
  - それらの遺伝子上にマップされる生のリード数が増加傾向
  - 配列長が長い遺伝子ほど発現レベルが高い傾向になる



mRNA	リード数
AAAAAAA...	5
AAAAAAA...	1

一つのサンプル内での異なる遺伝子間の発現レベルの高低を(配列長を考慮せずに)比較することはできない

Mortazavi et al., *Nature Methods*, **5**: 621-628, 2008

配列長の補正

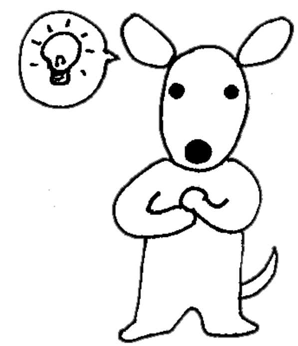
mRNA	リード数	配列長(in bp)
AAAAAAA...	5	1500
AAAAAAA...	1	300

- 前提条件:配列長が既知
- 補正の基本戦略:配列長で割る
  - 「1 / 配列長」を掛ける場合

    →「塩基あたりの平均のリード数」を計算しているのと等価
  - 「1000 / 配列長」を掛ける場合

    →「その遺伝子の配列長が1000bpだったときのリード数」と等価

    Reads Per Kilobase (of exon)



Mortazavi et al., *Nature Methods*, **5**: 621-628, 2008

RPKM

- RPM正規化（マイクロアレイなどと同じところ）
  - Reads per million mapped reads
  - サンプルごとにマップされた総リード（塩基配列）数が異なる。
    →各遺伝子のマップされたリード数を「総read数が100万（one million）だった場合」に補正

「raw counts : all reads= RPM : 1,000,000 」
A1BGの場合は「744 : 5,087,097 = RPM : 1,000,000」

$$\mathrm{RPM} = \text{raw counts} \times \boxed{\frac{1{,}000{,}000}{\text{all reads}}} = 744 \times \frac{1{,}000{,}000}{5{,}087{,}097} = 146.3$$

geneName	raw counts	RPKM	all reads	gene length	RPM
A1BG	744	82.9	5087097	1764	146.3
A1CF	159	13.7	5087097	2278	31.3
A2BP1	1	0.0	5087097	5415	0.2
A2LD1	4	0.6	5087097	1226	0.8

- RPKM正規化（RNA-Seq特有）
  - Reads per kilobase of exon per million mapped reads
  - 遺伝子の配列長が長いほど配列決定(sequence)される確率が上昇
    →各遺伝子の配列長を「1000塩基（one kilobase）の長さだった場合」に補正

$$\mathrm{RPKM} = \underbrace{\text{raw counts} \times \boxed{\frac{1{,}000{,}000}{\text{all reads}}}}_{\mathrm{RPM}} \times \boxed{\frac{1{,}000}{\text{gene length}}} = \text{raw counts} \times \frac{1{,}000{,}000{,}000}{\text{gene length} \times \text{all reads}}$$

解析結果が配列長依存という問題…

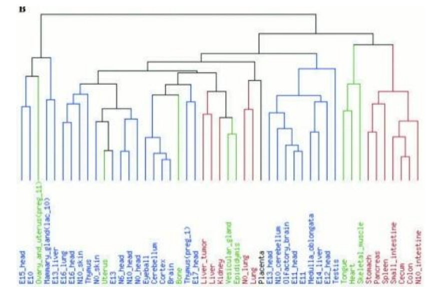
- 二群間比較など発現変動遺伝子(DEG)検出が目的の場合、(いわゆる発現比でランキングする方法以外の)統計的方法を用いると、配列長の長いものほどDEGと判定される確率が上昇してしまう

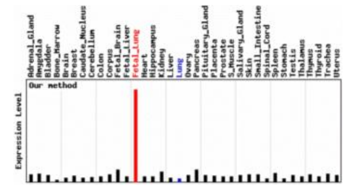
Oshlack and Wakefield, *Biology Direct*, **4**:14, 2009のFig 1

理由:長い遺伝子ほどバラツキが小さくなる傾向

正規化後のRNA-Seqデータ

- マイクロアレイデータと同様の解析が可能
  - クラスタリング
    - 似た発現パターンを持つ遺伝子やサンプルの同定
  - 発現変動遺伝子
    - 二群間比較、組織特異的遺伝子など
  - GSEA解析（どの遺伝子セットが動いているか）
    - Gene Ontology解析、パスウェイ解析など



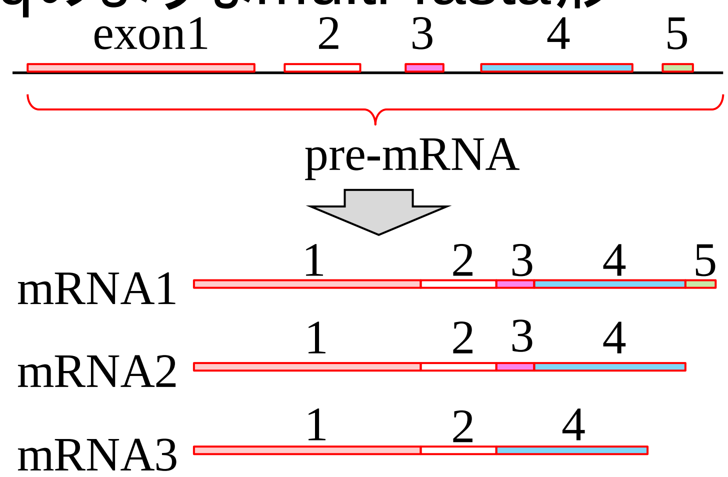


Ranking	Geneset_name
1	KEGG_TERPENOID_BACKBONE_BIOSYNTHESIS
2	KEGG_BIOSYNTHESIS_OF_UNSATURATED_FATTY_ACIDS
3	KEGG_FATTY_ACID_METABOLISM
4	KEGG_NITROGEN_METABOLISM
5	KEGG_LIMONENE_AND_PINENE_DEGRADATION

解析の基本的なイメージはマイクロアレイと同じです

なぜRNA-Seq？

- マイクロアレイに搭載されていない転写物も解析可能
  - 転写物全体の配列情報を取得可能（RefSeqのようなmulti-fasta形式のファイルをゲットできるイメージ）
  - 選択的スプライシングの全体像の理解
  - 発現変動exonの同定



モデル生物：より詳細なレベルでの理解
非モデル生物：（まずは）全体像の把握