RNAseqによる定量的解析とqPCR、マイクロアレイなどとの比較

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

自己紹介

- 1995年3月
  - 高知工業高等専門学校・工業化学科　卒業
- 1997年3月
  - 東京農工大学・工学部・物質生物工学科　卒業
- 1999年3月
  - 東京農工大学・大学院工学研究科・物質生物工学専攻　修士課程修了
- 2002年3月
  - 東京大学・大学院農学生命科学研究科・応用生命工学専攻　博士課程修了
  - 学位論文：「cDNAマイクロアレイを用いた遺伝子発現解析手法の開発」（指導教官：清水謙多郎教授）
- 2002/4/1~
  - 産総研・生命情報科学研究センター　産総研特別研究員
- 2003/11/1~
  - 放医研・先端遺伝子発現研究センター　研究員
- 2005/2/16~
  - 東京大学・大学院農学生命科学研究科　特任助手
- 2007/4/1~現在
  - 東京大学・大学院農学生命科学研究科　特任助教

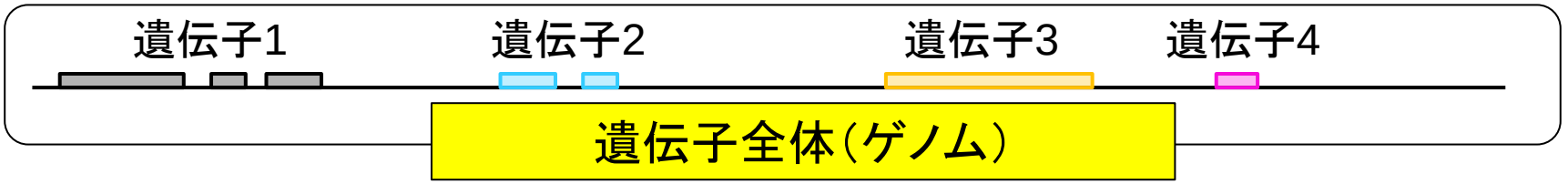
（2005/2/16~ 2007/4/1~現在）：アグリバイオインフォマティクスプログラム

Contents

- イントロダクション（発現レベルの数値化（定量化））
  - マイクロアレイ
  - RNA-seq（ゲノム配列既知のモデル生物の場合）
- 前処理（定量化や正規化）
  - RPKM、NAC、FVKM など
- 他のプラットフォーム（qPCRやマイクロアレイ）との比較
  - 発現量レベル（intra-sample）
  - サンプル間比較レベル（inter-sample）
- 非モデル生物のRNA-seq解析戦略
  - *de novo* transcriptome assembly → 発現変動コンティグ同定

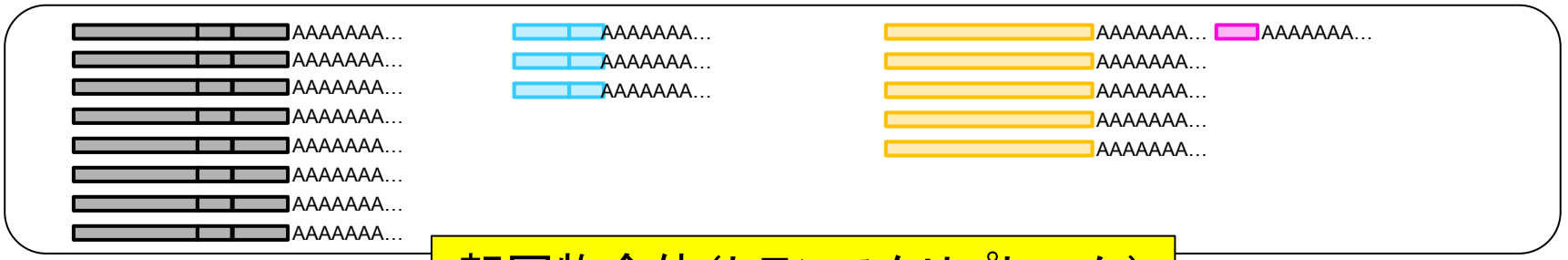
トランスクリプトームとは

- 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）

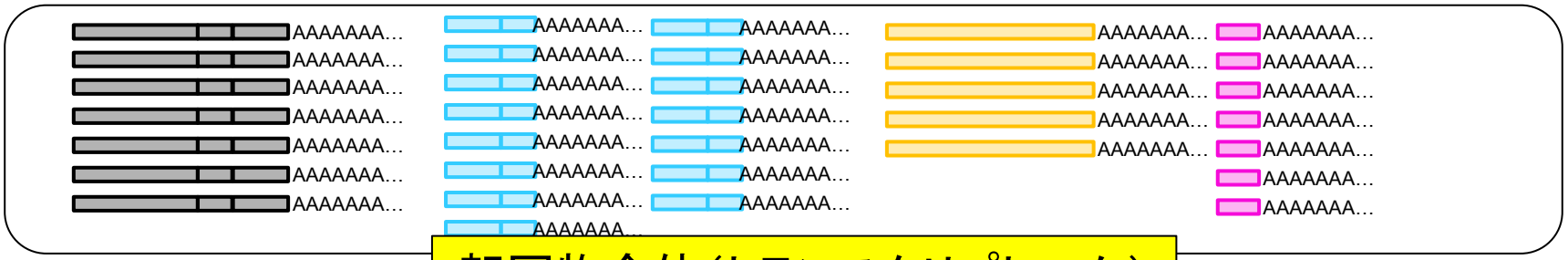
・遺伝子1は沢山転写されている（発現している）
・遺伝子4はごくわずかしか転写されてない
・…

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…）

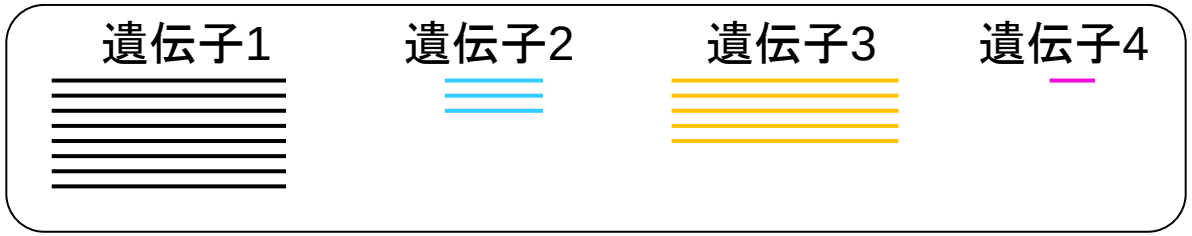


転写物全体（トランスクリプトーム）

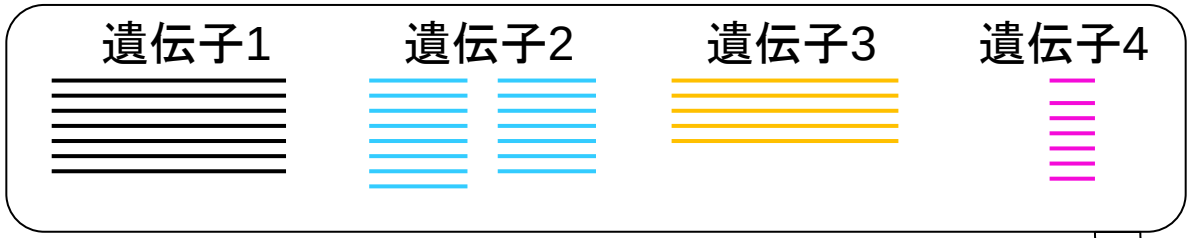
・遺伝子2は光刺激に応答して発現亢進
・遺伝子4も光刺激に応答して発現亢進

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



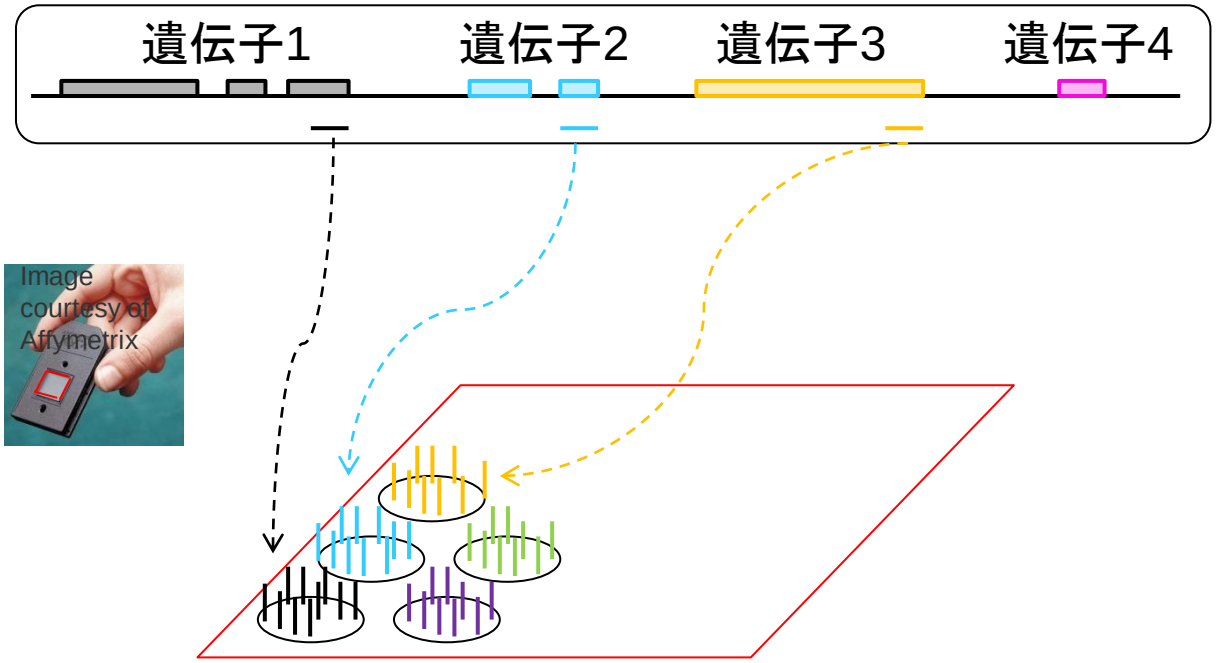
これがいわゆる
「遺伝子発現行列」

	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
- （電気泳動に基づく方法）
- 配列決定に基づく方法

トランスクリプトーム取得（マイクロアレイ）

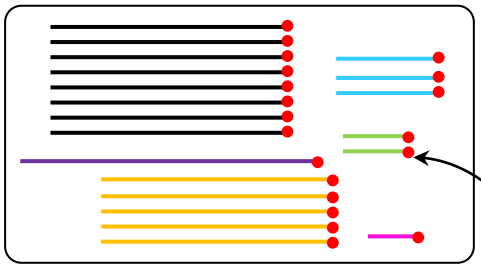
- よく研究されている生き物は多数の遺伝子（の配列情報）がわかっている



わかっている遺伝子（の配列の相補鎖）を搭載した”チップ”

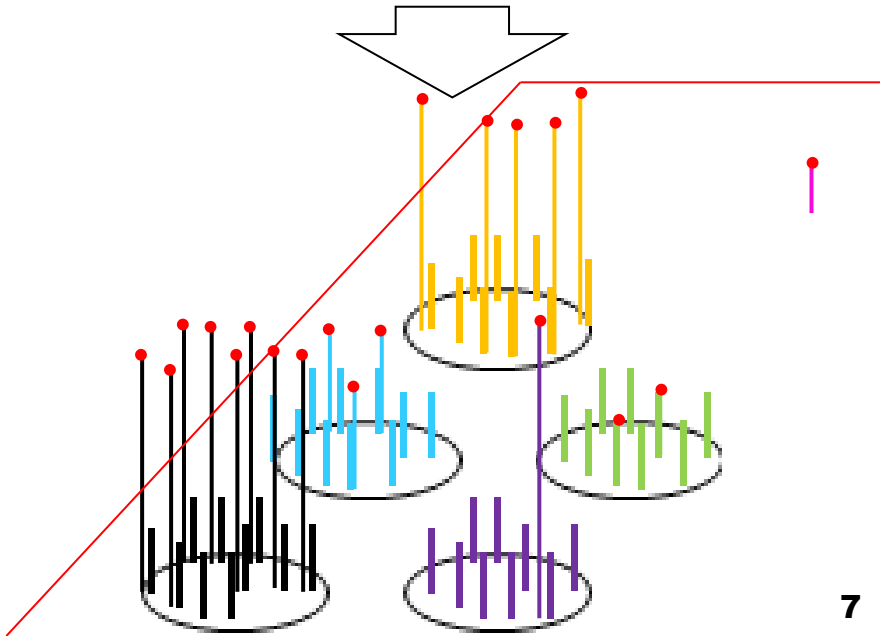
・メーカーによって搭載されている遺伝子の種類が異なる
→搭載されていない遺伝子（未知遺伝子含む、例：遺伝子4）の発現情報は測定不可…

光刺激前（T1）の目のトランスクリプトーム



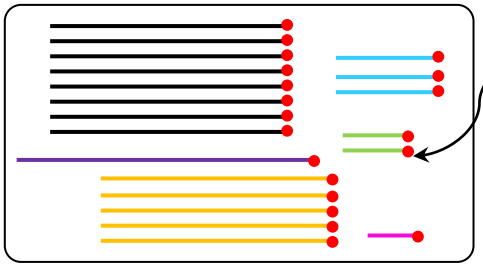
蛍光標識

ハイブリダイゼーション（二本鎖形成）

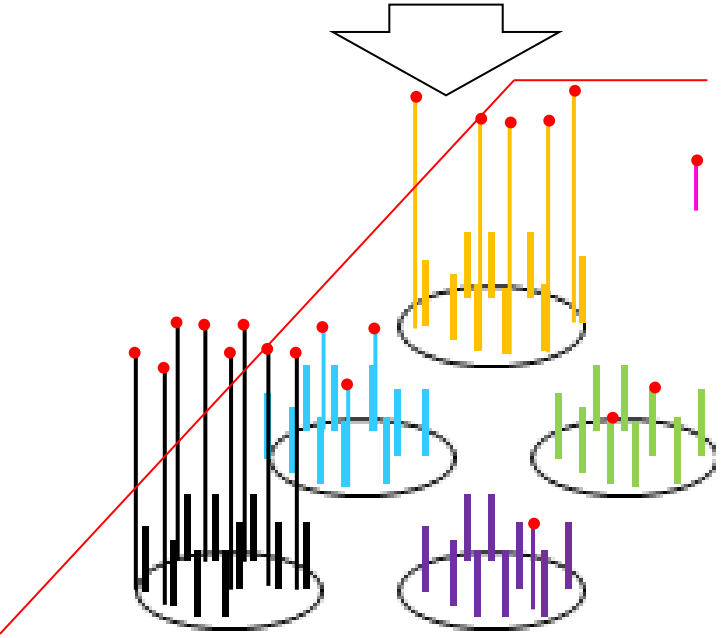


マイクロアレイデータ → 遺伝子発現行列

- 光刺激前（T1）の目のトランスクリプトーム

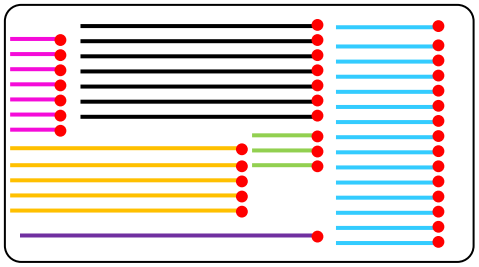


ハイブリダイゼーション
（二本鎖形成）



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

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
と
シグナル検出

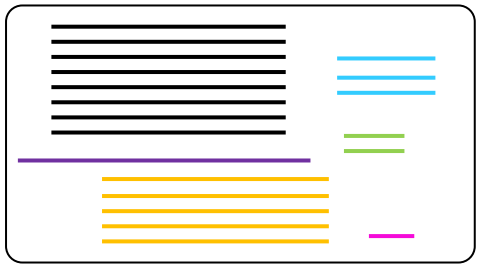
	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	?	?
遺伝子5	…	…
…	…	…

正規化

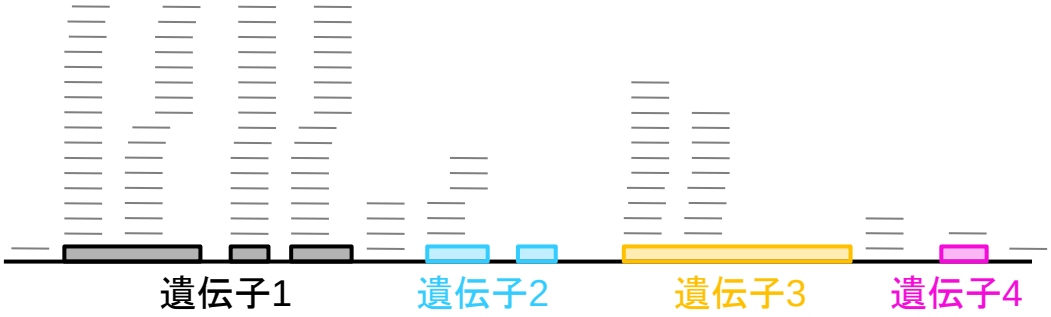
RNA-seqデータ → 遺伝子発現行列

- RNA-seq

光刺激前（T1）の目のトランスクリプトーム



ゲノム配列にマッピング



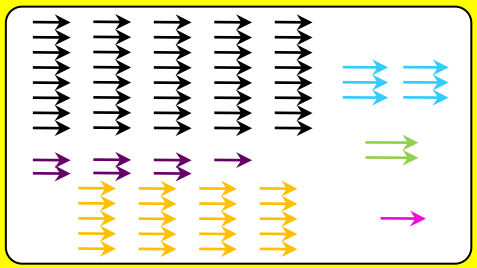
定量化(例：生のリード数をカウント)

	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	…
…	…

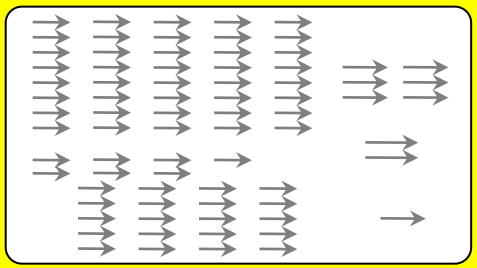
正規化

T1
8
3
5
1
…
…

－イメージ－
50-125塩基程度からなる配列が沢山ある



－実際－
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

前処理(定量化や正規化)

- 基本的な考え
  - サンプル間の総リード数の違いをいかに補正するか
  - 配列長由来の偏り(長いほど沢山sequenceされる)をいかに補正するか

    (長さの異なる複数のisoformsが存在する場合にその遺伝子の配列長をいかに定義するか)

- RPKM (Mortazavi et al., *Nat Methods*, 2008; **ERANGEの論文**)
  - Reads per kilobase of exon per million mapped reads
- NAC (Griffith et al., *Nat Methods*, 2010; **ALEXA-seqの論文**)
  - Normalized average coverage

本質的に同じ

- FPKM (Trapnell et al., *Nat Biotechnol*., 2010; **Cufflinksの論文**)
  - Fragments per kilobase of transcript per million mapped fragments
- FVKM (Lee et al., *Nucleic Acids Res*., 2010; **NEUMAの論文**)
  - Fragments per virtual kilobase per million mapped reads

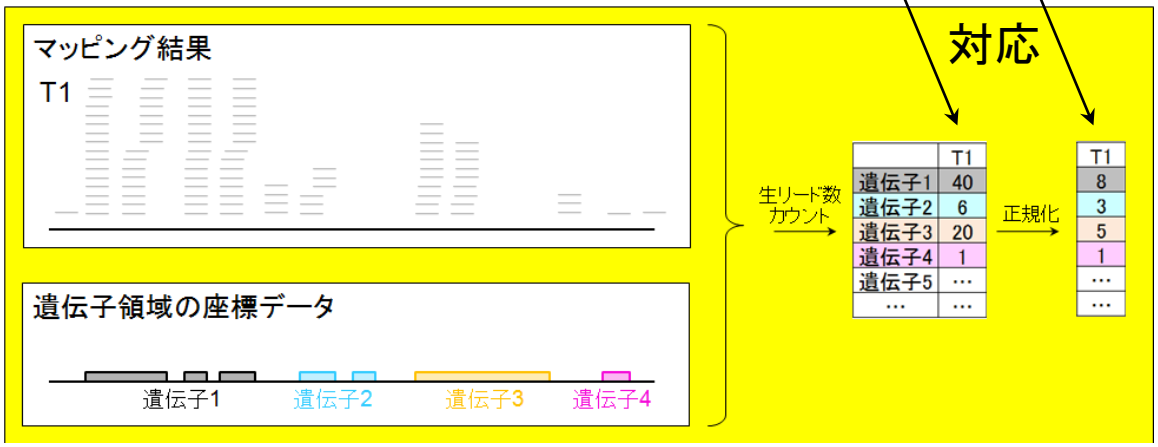
Multiple isoforms

…

RPKM (Mortazavi *et al*., 2008)

- Reads per kilobase of exon per million mapped reads

geneName	raw counts	RPKM	all reads	gene length
A1BG	744	82.9	5087097	1764
A1CF	159	13.7	5087097	2278
A2BP1	1	0.0	5087097	5415
A2LD1	4	0.6	5087097	1226
A2M	2373	100.3	5087097	4653



5087097 reads (重複を含む塩基配列数)がマップされており、そのうち **744** readsが**A1BG**という遺伝子のエクソン上にマップされていて、この遺伝子をRPKMという単位で定量化すると**82.9**となる。
どうやって計算してる？

RPKM (Mortazavi *et al*., 2008)

- 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」

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

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

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

$$\mathrm{A1BG} = 744 \times \frac{1{,}000{,}000{,}000}{1{,}764 \times 5{,}087{,}097} = 82.9$$

NAC (Griffith *et al*., 2010)

- Normalized average coverage
  - 1リードがx塩基の長さとして考える
  - 長さ補正

    ある遺伝子のaverage coverage (AC)は「その遺伝子上にマップされた総塩基数」を「その遺伝子の長さ」で割ったものなので、

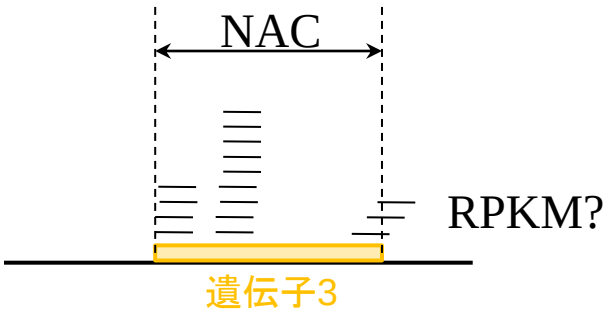
$$\mathrm{AC} = \frac{\text{raw counts} \times x}{\text{gene length}} = \frac{744 \times x}{1{,}764}$$

  - 総リード数補正

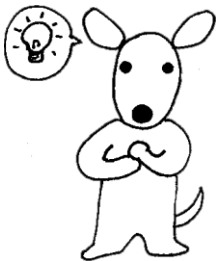
    サンプルごとにマップされたリードの総塩基数が異なるので、マップされたリードの総塩基数が10,000,000,000塩基だった場合に補正

$$\mathrm{NAC} = \mathrm{AC} \times \frac{10{,}000{,}000{,}000}{\text{all reads} \times x} = \text{raw counts} \times \frac{10{,}000{,}000{,}000}{\text{gene length} \times \text{all reads}} = 10 \times \mathrm{RPKM}$$

geneName	raw counts	RPKM	all reads	gene length
A1BG	744	82.9	5087097	1764
A1CF	159	13.7	5087097	2278
A2BP1	1	0.0	5087097	5415
A2LD1	4	0.6	5087097	1226
A2M	2373	100.3	5087097	4653

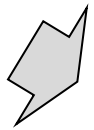


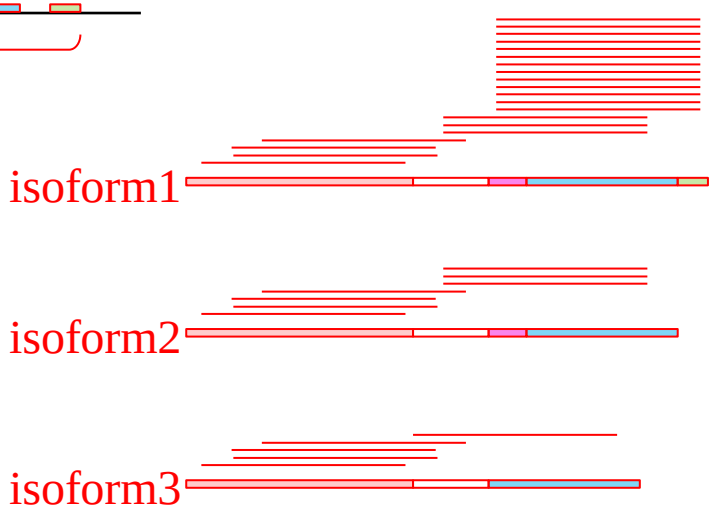
NACとRPKMは本質的に同じだが、NACのほうがより厳密



複数アイソフォーム対策

- 「元の遺伝子（補正後）のgene length」値をいかに見積もるか？
  - FPKM (Trapnell et al., *Nat Biotechnol*., 2010; **Cufflinksの論文**)
    - 複数のisoformsの長さと発現量をもとに、「発現量で重みをつけた平均値」を採用





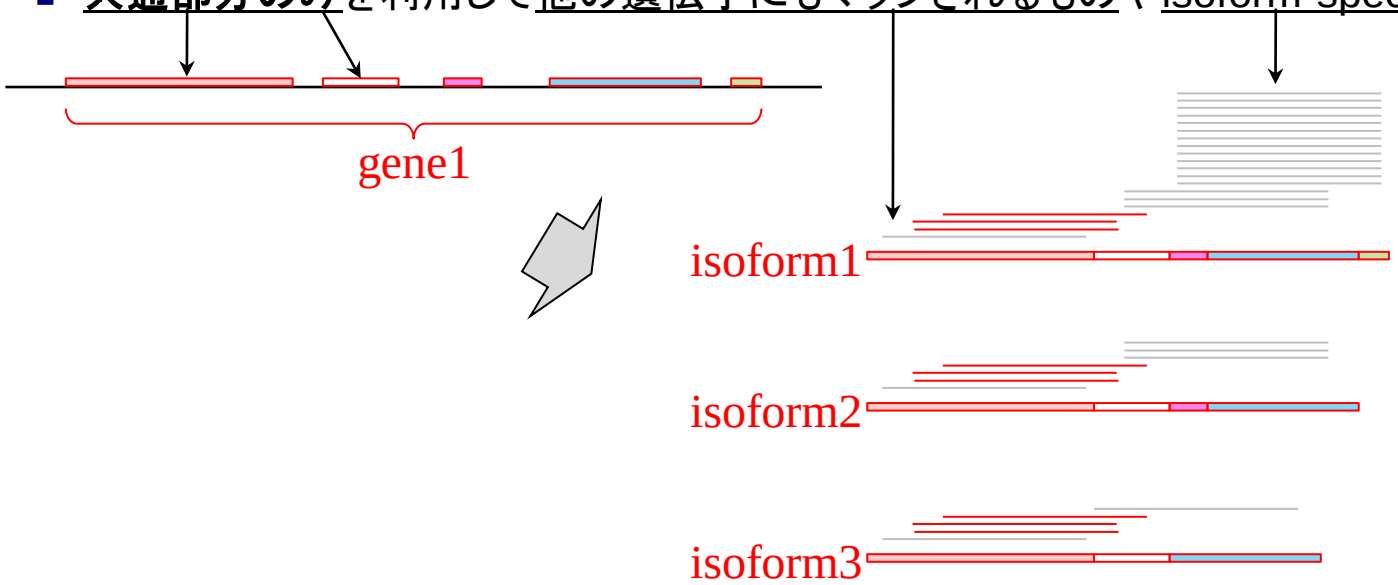
	長さ	発現量
isoform1	69bp	20
isoform2	65bp	7
isoform3	60bp	5

$$\text{補正後のgene length} = \frac{20\times 69 + 7*65 + 5*60}{20+7+5} = 66.72\ \text{bp}$$

$$\text{raw counts} \times \frac{\text{定数}}{\text{gene length} \times \text{all reads}}$$

複数アイソフォーム対策

- 「元の遺伝子（補正後）のgene length」値をいかに見積もるか？
  - FVKM (Lee et al., *Nucleic Acids Res*., 2010; **NEUMA**の論文)
    - **共通部分のみ**を利用して他の遺伝子にもマップされるものやisoform-specificなものは使わない

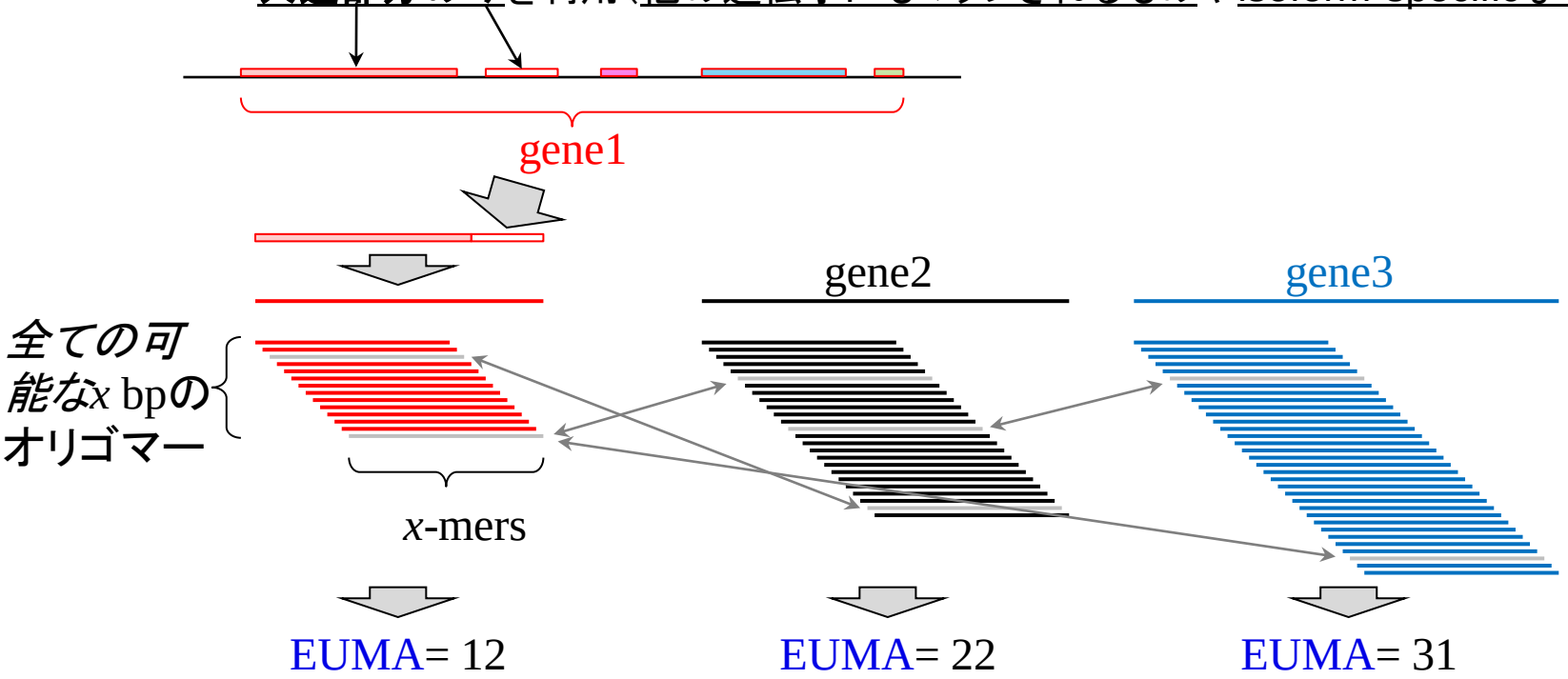


raw count (原著論文ではgNIR) = 3

$$\text{raw counts} \times \frac{\text{定数}}{\text{gene length} \times \text{all reads}}$$

複数アイソフォーム対策

- 「元の遺伝子（補正後）のgene length」値をいかに見積もるか？
  - FVKM (Lee et al., *Nucleic Acids Res*., 2010; **NEUMA**の論文)
    - **共通部分のみ**を利用（他の遺伝子にもマップされるものやisoform-specificなものは使わない）

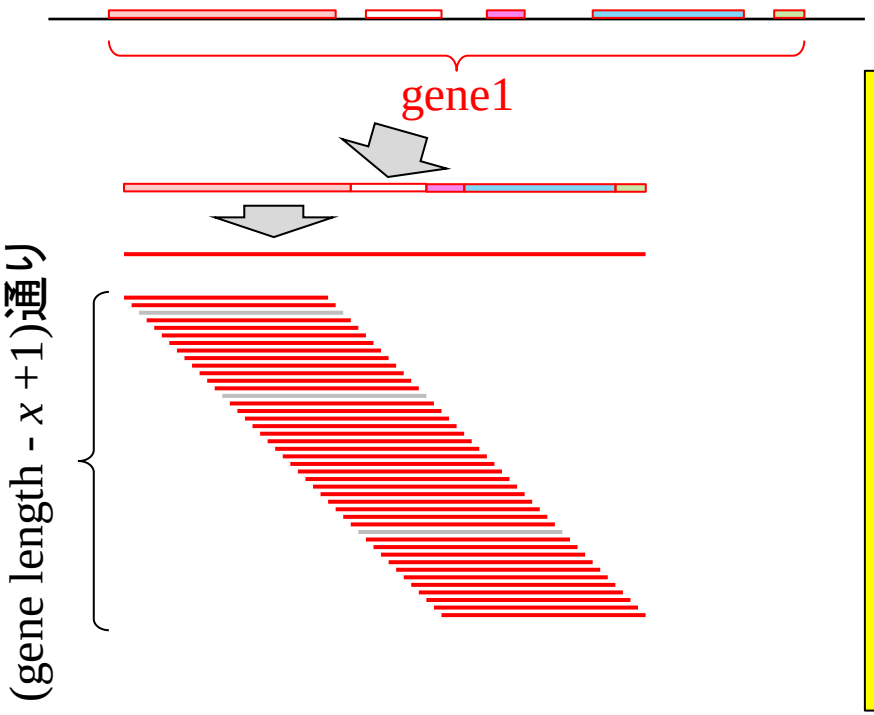


$$\mathrm{FVKM} = 3 \times \frac{1{,}000{,}000{,}000}{12 \times \text{all reads}}$$

$$\text{raw counts} \times \frac{\text{定数}}{\text{gene length} \times \text{all reads}}$$

複数アイソフォーム対策

- 「元の遺伝子（補正後）のgene length」値をいかに見積もるか？
  - virtual length (Sultan et al., *Science*, 2008)
    - **全エクソンの領域**を利用



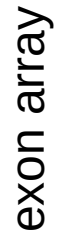
他の遺伝子上にはなくユニークにヒットする*x*-merの数の期待値（theoretical total number of unique *x*-mers）を”virtual length”と定義

raw countsのほうも100%マッチでユニークにマップされるリード数のみをカウント

$$\text{raw counts} \times \frac{\text{定数}}{\text{gene length} \times \text{all reads}}$$

他のプラットフォームとの比較(vs. microarray)

- 発現量レベル（intra-sample）



2,434 genes

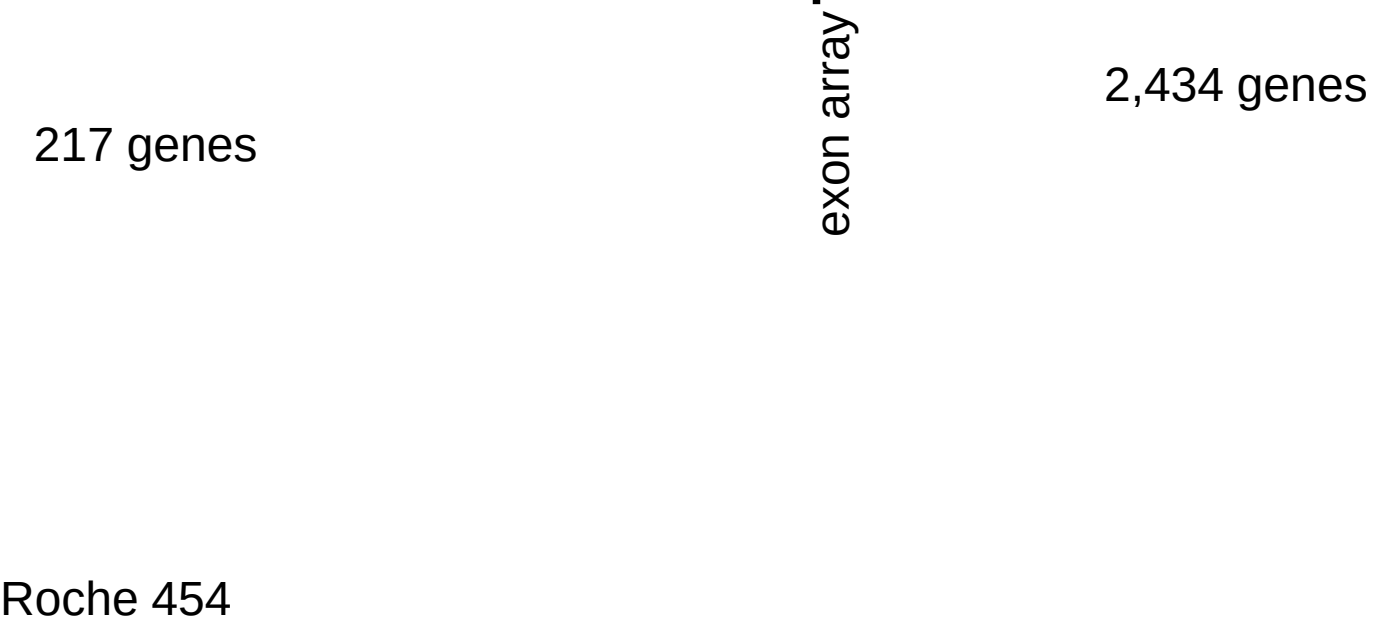
$\log_2$(NAC)

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

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

他のプラットフォームとの比較(vs. microarray)

- サンプル間比較レベル（inter-sample）



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

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

他のプラットフォームとの比較(vs. qPCR)

- 発現量レベル（intra-sample）

FVKM

FPKM

27 genes

RPKM

RPKM

Lee et al., *Nucleic Acids Res.*, 2010のFig. 2

他のプラットフォームとの比較(vs. qPCR)

- サンプル間比較レベル（inter-sample）

Griffith et al., *Nat Methods*, 2010のFig. 2

前処理は重要（遺伝子発現行列作成時）

- 発現量補正の基本形

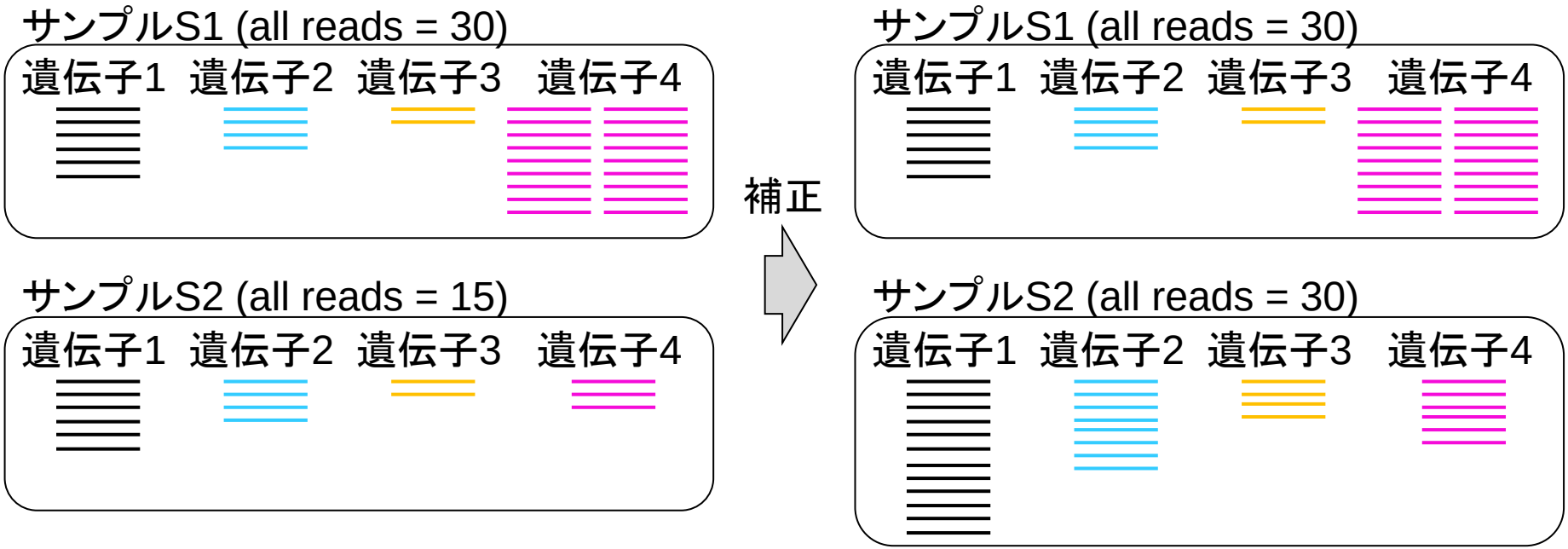
$$\text{raw counts} \times \frac{\text{定数}}{\text{gene length} \times \text{all reads}}$$

- **発現量レベル（intra-sample）の（プラットフォーム間）比較**
  - all readsの項はなくてもよい
- **サンプル間比較（inter-sample）の場合、「基本形」ではまだ不十分**
  - Bullard et al., *BMC Bioinformatics*, 2010
    - RPKM補正でもまだ、発現変動遺伝子が配列長の長いものに偏る
    - t統計量$/\sqrt{\text{gene length}}$ で若干緩和される
  - Robinson and Oshlack, *Genome Biol*., 2010
    - サンプル中の「**RNA組成の違い**」による影響は甚大
    - 付加的な正規化係数（TMM）を掛けることで影響が緩和される

「RNA組成の違い」のイメージ

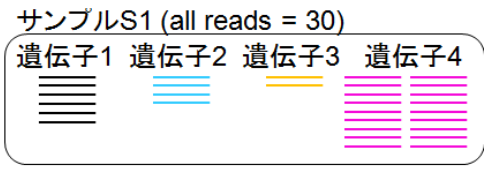
- 仮定
  - 全4遺伝子
  - 長さが同じ（gene lengthの項を無視できるので）
  - 遺伝子4だけが発現変動遺伝子

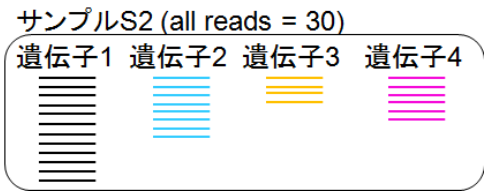
$$\text{raw counts} \times \frac{\text{定数}}{\cancel{\text{gene length}} \times \text{all reads}}$$



補正結果：S1で高発現が1個, S2で高発現が3個

M-A plot (R-I plot)

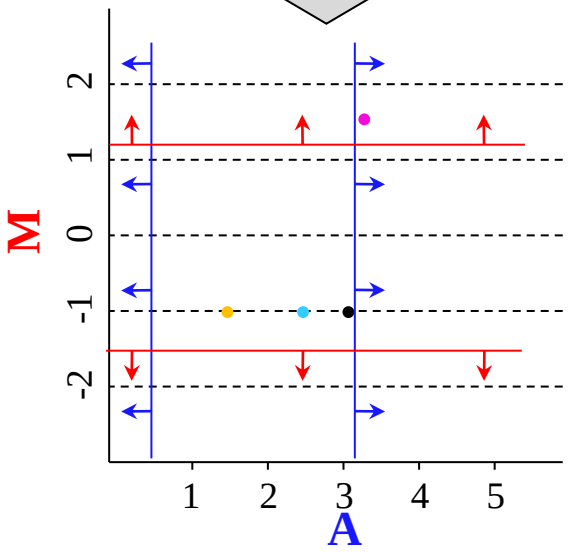


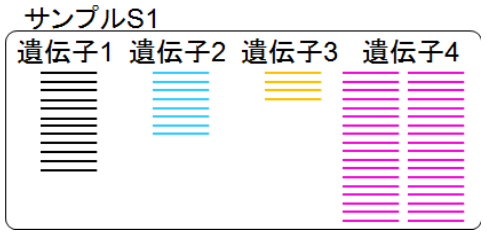


	S1	S2
遺伝子1	6	12
遺伝子2	4	8
遺伝子3	2	4
遺伝子4	18	6

	$\log_2$(S1)	$\log_2$(S2)
遺伝子1	2.58	3.58
遺伝子2	2.00	3.00
遺伝子3	2.00	2.00
遺伝子4	4.17	2.58

縦軸（発現比） R: $\log_2$(S1/S2)	横軸（全体的な発現レベル） I: log2(sqrt(S1×S2))
M: $\log_2$(S1)-$\log_2$(S2)	A: ($\log_2$(S1)+$\log_2$(S2))/2
-1.00	3.08
-1.00	2.50
-1.00	1.50
1.58	3.38





	$\log_2$(S1) - TMM	$\log_2$(S2)
遺伝子1	3.58	3.58
遺伝子2	3.00	3.00
遺伝子3	2.00	2.00
遺伝子4	5.17	2.58

TMM = -1

横（and 縦）軸で上位下位のx（and y）%をTrim
→ 残りのデータでMのMean（TMM）を計算

TMM補正するしないで…

- 得られたDEGセット中の割合
  - TMM補正なし（Marioni et al., *Genome Res.*, 2008）
    - サンプルS1（Liver）：22%
    - サンプルS2（Kidney）：78%
  - TMM補正あり（Robinson and Oshlack, *Genome Biol.*, 2010）
    - サンプルS1（Liver）：47%
    - サンプルS2（Kidney）：53%

- **基本形で発現量補正 → 追加補正 → その後の解析**

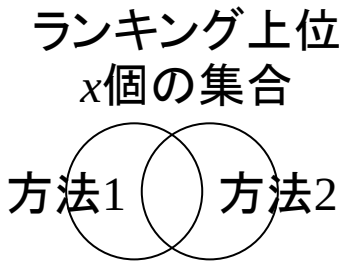
$$\text{raw counts} \times \frac{\text{定数}}{\text{gene length} \times \text{all reads}}$$

・TMM
・median
etc…

・発現変動遺伝子検出
・分類
・クラスタリング
etc…

マイクロアレイからの知見（発現変動遺伝子;DEG）

- Jeffery et al., *BMC Bioinformatics*, 2006
  - 別のランキング法を用いると違った結果に



一致は8-21%！（再現性低い…）

- Kadota et al., *Algorithms Mol. Biol.*, 2008,2009
  - 既知のDEGは全体的に発現レベルが高い
  - **ランキング法**は「*t*-test系とFold Change系」に大別でき、この間の比較で再現性低下
  - 遺伝子発現行列作成時に用いる**前処理法**（Affymetrixの場合）の違いの影響もある
    →ランキング法と前処理法の組合せが大事
  - **感度・特異度**が高いランキング法：Rank products or WAD
  - **再現性**：WAD（前処理法によらず）
- Hu and Xu, *BMC Genomics*, 2010
  - **感度・特異度**：WAD > *t*-test > Fold change > Rank products
    - 上位1,000遺伝子までで評価
    - 前処理法として何を使ったか不明（公共DBはMAS-preprocessed dataが大半で、Rank productsとの相性悪い）

非モデル生物のトランスクリプトーム解析

- *de novo* **genome** assembly用プログラム
  - Velvet (Zerbino and Birney, *Genome Res*., 2008)
  - ABySS (Simpson et al., *Genome Res*., 2009)
  - EULER-SR (Chaisson et al., *Genome Res*., 2009)
  - etc…
- *de novo* **transcriptome** assembly用プログラム（特にIllumina）
  - Multiple-k (Surget-Groba and Montoya-Burgos, *Genome Res*., 2010)
  - Trans-ABySS (Robertson et al., *Nat Methods*, 2010)
  - Rnnotator (Martin et al., *BMC Genomics*, 2010)
  - Oases (Schulz and Zerbino, unpublished)

2010年の夏以降からtranscriptome用のものが続々登場

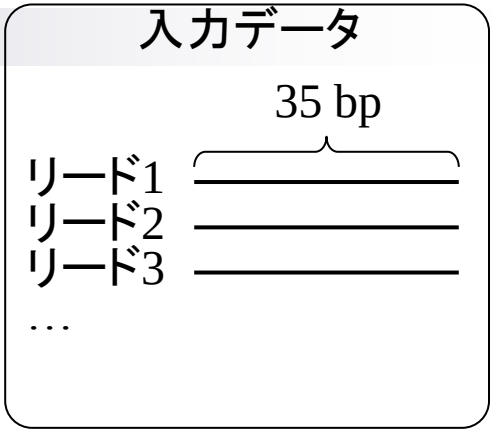
de novo **transcriptome** assembly

- 目的：（short）readsのデータから転写物ごとのコンティグを得る
- アセンブリの基本戦略
  1. （計算を軽くするため、ユニークなリード配列の集合にしておく）
  2. *de novo* **genome** assembly用プログラムを複数の*k*値で実行
     - 転写物の場合はcoverageが多様である
     - 転写物が高（or 低）発現のときはhigh (or low) coverageであることを意味する
     - *k*を大きくすると高発現転写物がアセンブルされる確率が上がる（低感度高特異度）
     - *k*を小さくすると低発現転写物がアセンブルされる確率が上がる(がキメラも増える；高感度低特異度)
     - Rnnotator：*k*=19, 21, …, 33
     - Multiple-*k*: *k*=19, 21, …, 29
     - Trans-ABySS: *k*=26, 27, …, 49

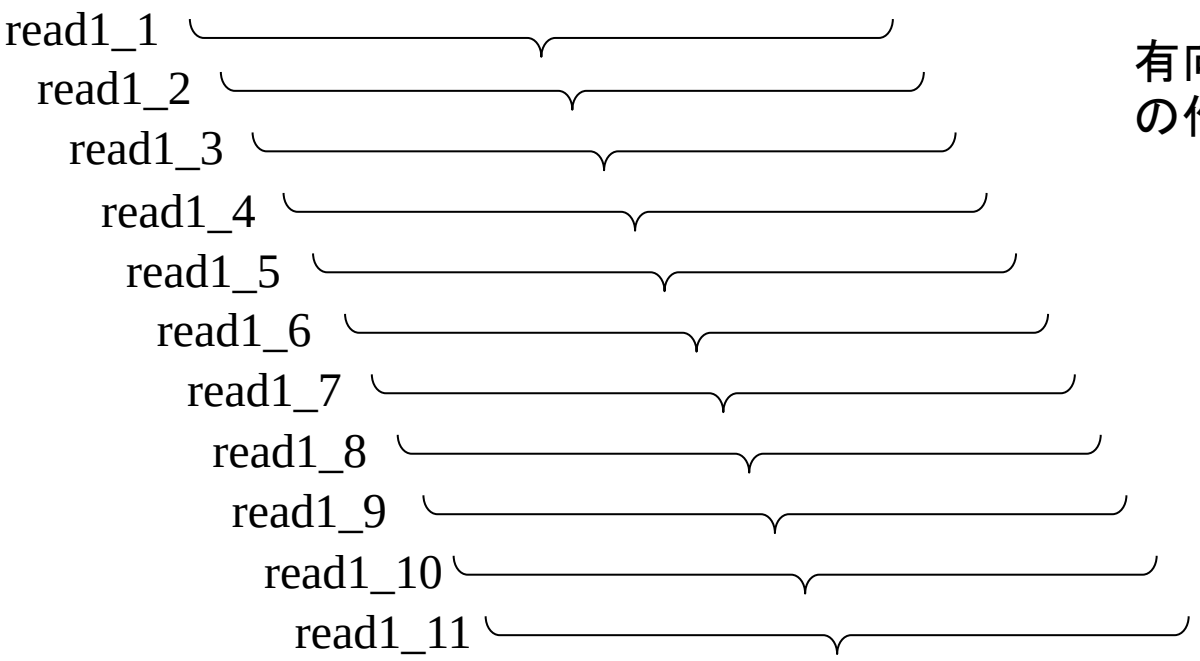
いろいろ試して、できるだけ転写物のcoverageを上げる
（読んだリードの長さLによってkの探索範囲を変更）

35 bpのsingle-endで*k*を考える

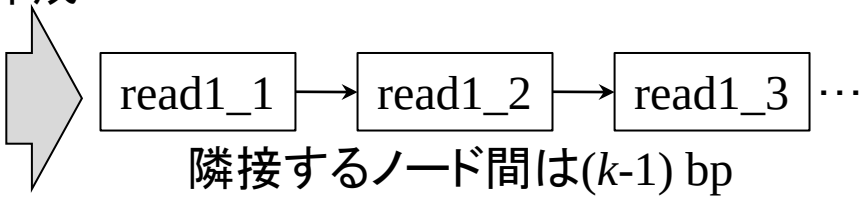
- 各リードを全ての可能な*k*-mer (*k* < 35の任意の値; 例えば*k*=25)に分割して有向グラフを作成



read1 : TGCCGACATGCATCCAAGTAGGAATCCTTAGCTTA



有向グラフの作成

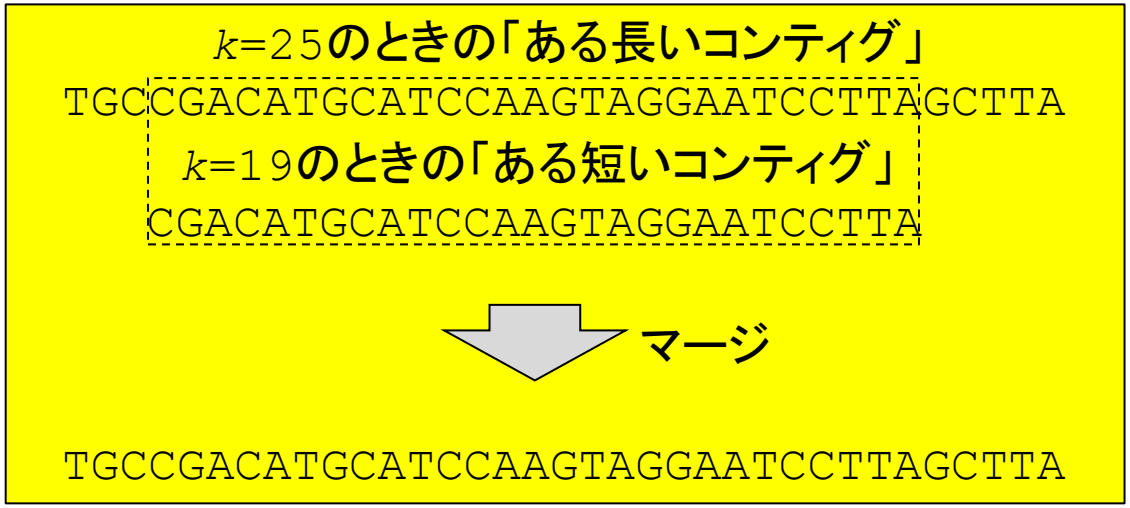


隣接するノード間は(*k*-1) bpのオーバーラップ

全リードのグラフ情報をもとに同一ノードをマージしたグラフ(de Bruijn graph)を作成し、オイラーパス問題として解く(=コンティグを得る)

de novo **transcriptome** assembly

- 目的：(short)readsのデータから転写物ごとのコンティグを得る
- アセンブリの基本戦略
  3. それぞれの*k*値を用いて独立してアセンブルを行った結果から、長いコンティグ中に短いコンティグが100%マッチになるものはマージしていくことでnon-redundant setにする



de novo **transcriptome** assembly

- 目的：（short）readsのデータから転写物ごとのコンティグを得る
- アセンブリの基本戦略
  4. キメラコンティグを分割

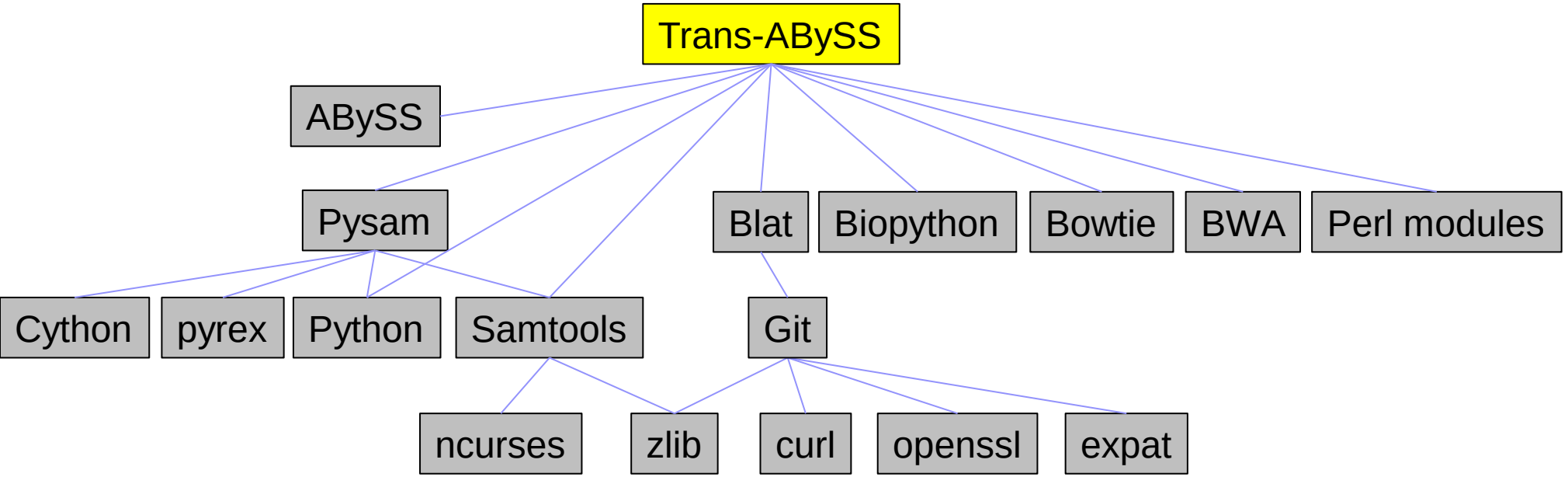
Martin et al., *BMC Genomics*, 2010のFig. 3b

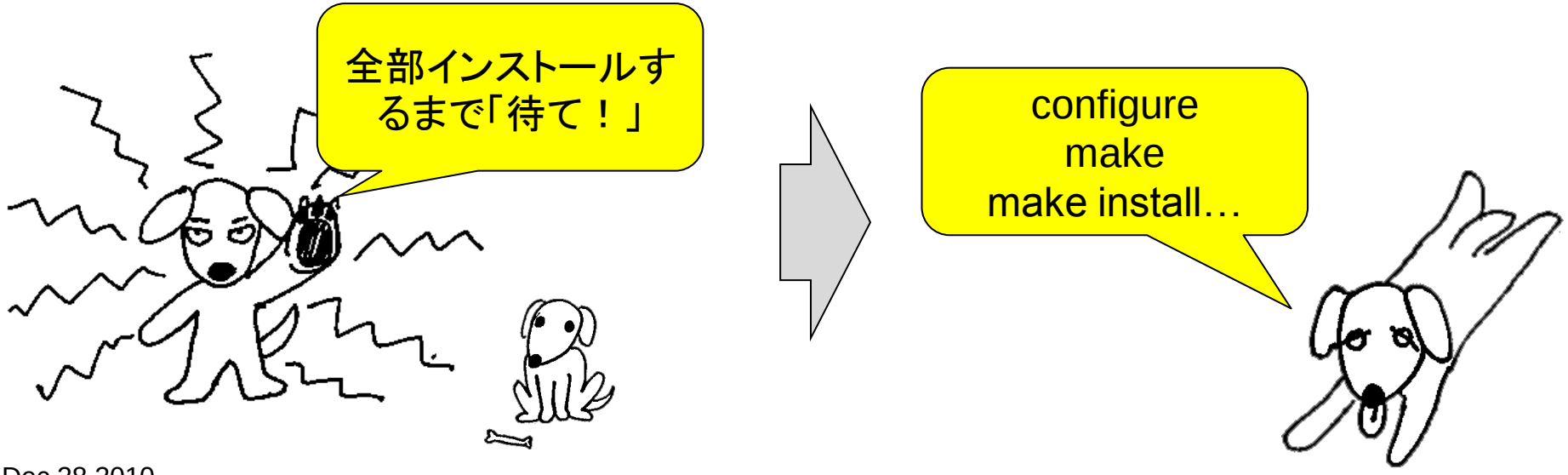
コンティグに再びリードをマップさせてforward側と
reverse側で明確にcoverageが異なるところで分離

非モデル生物の比較トランスクリプトーム解析戦略

1. 比較する複数サンプル（samples A and B）由来のリードを一つにまとめたセットを用意
2. *de novo* transcriptome assemblyプログラムを実行し、コンティグのセット(transcriptome sequence)を得る
3. Transcriptome sequenceに各サンプル由来リードを（Bowtieなどを用いて）マップ
   - 発現量の定量化はNEUMA的な考え方でunique readsの結果のみ採用
   - （正規化は二つのサンプル由来リードがマップされているコンティグの発現レベルのみを考慮し、TMM正規化のような考え方を採用）

要求されること（例：Trans-ABySS）





謝辞

東京大学　大学院農学生命科学研究科

清水 謙多郎　教授

嶋田 透　教授

グラント

- 若手研究(B)（H21年度- ）:「マイクロアレイ解析の再現性・感度・特異度を飛躍的に向上させるデータ解析手法の開発」(代表)
- 新学術領域研究(研究領域提案型)(H22年度-):「非モデル生物におけるゲノム解析法の確立」(分担)