バイオインフォマティクス
次世代シーケンサー（NGS）編

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

バイオインフォマティクス人材育成講座「スタンダードコース」

生命科学概論	生物学、遺伝学、分子生物学、生化学等の発展にバイオインフォマティクスが寄与するところについて概説する
情報科学概論	社会で活かされている身近な情報科学を導入として、生物分野においても情報科学の知識や技術が様々に活かされていることを概説する
バイオインフォマティクス総論	バイオインフォマティクスとは何か、その概要について歴史のふり返りと今後の展望から学び、基礎から応用、産業利用まで全体を俯瞰する
生命科学基礎 ○細胞構造 バイオインフォマティクスにお ○セントラルドグマ（DNA、RNA、タンパク質） ○PCR・電気泳動・シーケンシング	**情報科学基礎** ○確率・統計・検定 バイオインフォマティクスにお ○クラスタリング ○データベース
バイオインフォマティクス・基礎編	公共の研究機関等が提供しているデータベースサイトを利用しながら、配列解析、相同性検索、分子系統解析など、搭載データの活用を学ぶ
バイオインフォマティクス・応用編	WEBに公開されているツールを用いて、遺伝子の機能推定、タンパク質立体構造予測、ドッキング解析など、創薬分野への応用を事例に学ぶ
バイオインフォマティクス・NGS編	次世代、次々世代のシーケンサーの特徴について、その取り扱いや、コストなども踏まえた解説と、研究や産業における利活用について学ぶ

次世代シーケンサーを活用した実験解析について、トランスクリプトーム解析など最新の研究技術について学ぶ。

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　特任助教

遺伝子発現量、システムバイオロジー、トランスクリプトーム解析、マイクロアレイ、次世代シーケンサー、アラインメント

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

アグリバイオインフォマティクスプログラム

高専時代の成績もたいしたことない門田が、かれこれ10年以上バイオインフォマティクスの分野で楽しくやってます。

次世代シーケンサー

- Next-Generation Sequencer (NGS)
- 塩基配列を決定する実験機器のこと
- 特徴
  - 旧世代シーケンサーに比べ、一度に多数の塩基配列を決定することができる
  - **ゲノム**配列決定（ゲノム解読）や**トランスクリプトーム**解析手段としての応用が広がっている

オーム（Ome）研究

ome ：総体

DNA　遺伝子 (Gene) + ome　→ ゲノム（命の設計図）

転写（遺伝子DNAの情報をRNAに写すこと）

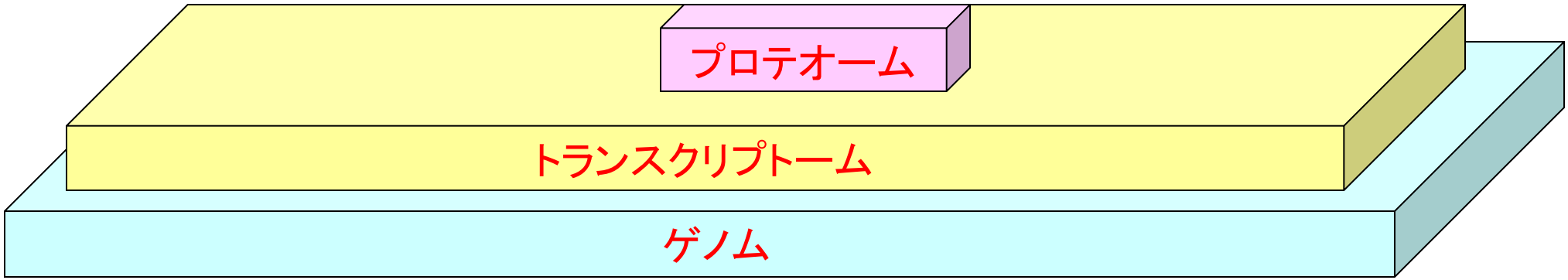
RNA　転写 (Transcription) + ome　→ トランスクリプトーム

翻訳（転写されたRNA情報をもとにタンパク質を作ること）

タンパク質　タンパク質 (Protein) + ome　→ プロテオーム

これまでの実験技術

全ゲノムショットガン法　　マイクロアレイ　　二次元電気泳動法

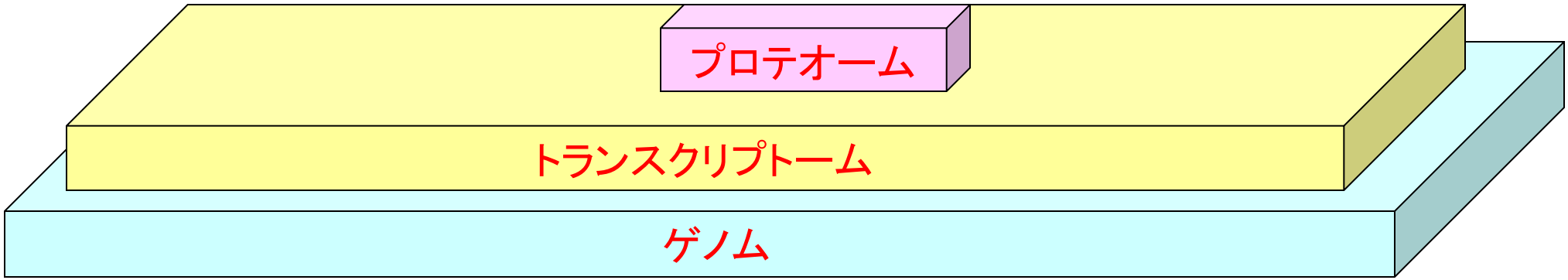


今後の実験技術

次世代シーケンサー（NGS）

GS FLX+　　5500xl SOLiD　　HiSeq 2000

二次元電気泳動法



NGSでゲノム解読の成果は続々と…

- パンダ（大熊猫）ゲノム解読（2008年）
  - ヒトゲノム解読に10年 → 半年
  - 猫よりも犬・熊に近い動物
- アジア人（中国人）一個体の全ゲノム配列決定（2008年）
- 国際プロジェクト
  - 1000人ゲノム計画（1人1人の遺伝情報の違いを詳細に調査）
  - 国際癌ゲノムプロジェクト
  - 感染症の同定
- 日本人の全ゲノム配列決定（2010年）
- **世界で初めてサンゴの全ゲノム解読に成功**（2011年7月）
  - サンゴと褐虫藻との共生メカニズム解明のための基盤情報取得
  - サンゴの白化現象（褐虫藻を失うこと）解明のための～
  - サンゴ礁の観光産業などの経済効果は2,500億円以上！

http://www.oist.jp/ja/press-room/press-releases/150/1047-pr-first-coral-genome.html

NGSの利活用(妄想?!)

- ○○のゲノム解読
  - 絶滅危惧種関連(ゲノム情報は沖縄にあり！)
    - 西表山猫とか…
  - バイオマスエタノール関連(エネルギー生産関連)
    - サトウキビとか…
- ○○と□□の比較ゲノム解析
  - ある有用な機能をもつ微生物(○○)ともたないもの(□□)
    - ○○のみがもつその機能と関連する遺伝子の同定
  - 長寿(沖縄) vs. 短命の県
- ○○と□□の比較トランスクリプトーム解析
  - ある有用な機能をもつ微生物(○○)ともたないもの(□□)
    - 発現に違いのある遺伝子同定

人材育成…

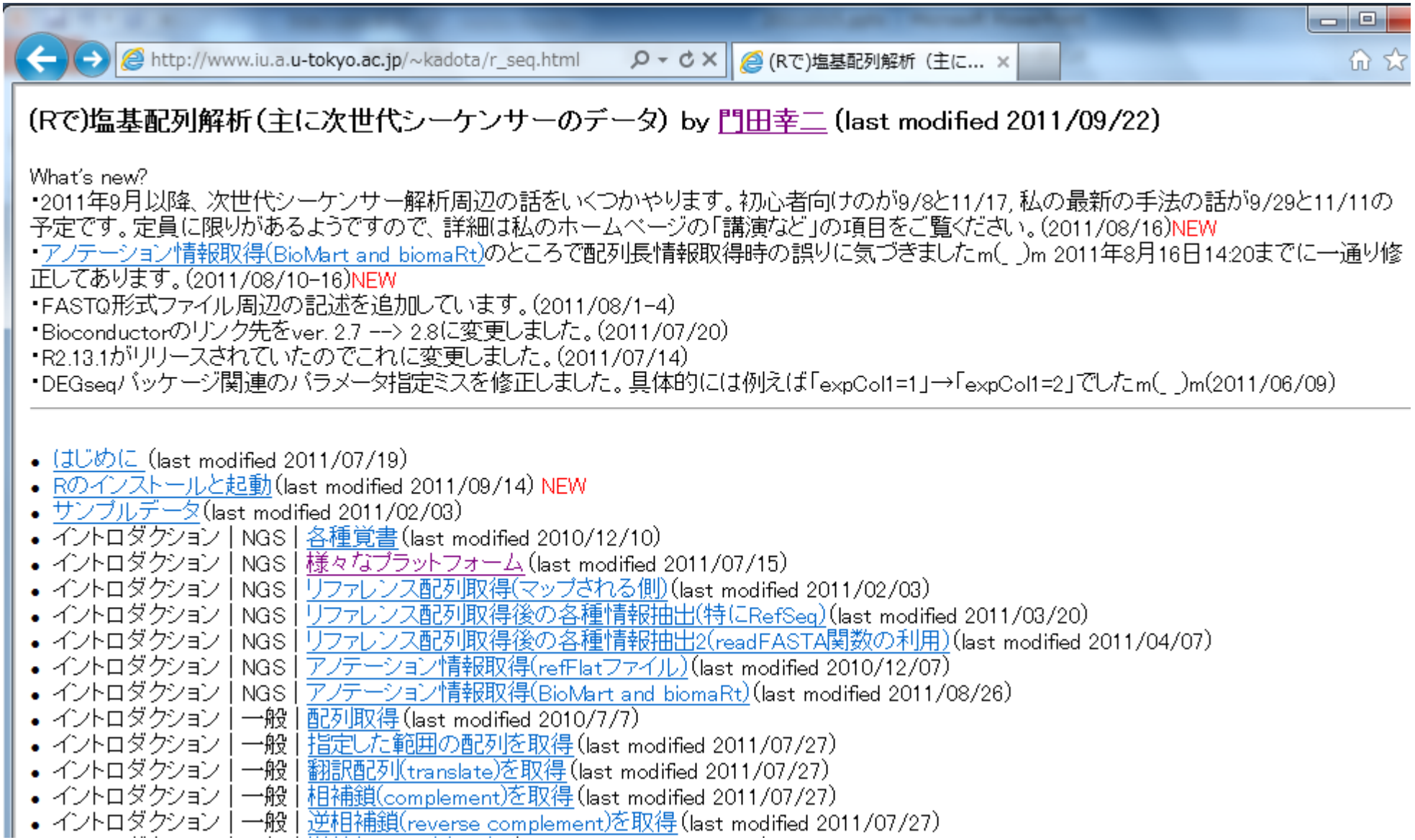
- 現状
  - NGSデータなどの大量実験データを自在に解析できるバイオインフォマティクス人材が不足
  - スキルのある人は引く手あまた
- 私の状況
  - 東大生のバイオインフォマティクス講義（90分×11回）
  - あちこちでセミナーや講習会の講師
  - 自分の研究（と共同研究の解析）を進める
  - メールでの質問対応（これも頻繁にくるので大変）
  - 初心者でもコピペでデータ解析可能なウェブページの更新
    - （Rで）マイクロアレイデータ解析
    - （Rで）塩基配列解析

ここでの講義も結局は自分のため

ねらい

- 次世代シーケンサー（NGS）を活用した実験解析について、トランスクリプトーム解析など最新の研究技術について学ぶ
- Rを利用することで、NGSから得られる塩基配列データの様々な解析が可能
  - プログラミング能力がなくても使いこなし術があれば…
- NGS解析を全部自力でやるにはLinuxのノウハウがある程度必要であることを実感してもらう
- バイオインフォマティクスの基本的なスキルを身につけることが重要
  - バイオインフォマティクス技術者認定試験合格を目指せ
  - 相関係数やエントロピーなどの要素技術を駆使すれば様々なデータ解析が可能であることを紹介

参考1



(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2011/09/22)

What's new?
・2011年9月以降、次世代シーケンサー解析周辺の話をいくつかやります。初心者向けのが9/8と11/17, 私の最新の手法の話が9/29と11/11の予定です。定員に限りがあるようですので、詳細は私のホームページの「講演など」の項目をご覧ください。(2011/08/16)NEW
・アノテーション情報取得(BioMart and biomaRt)のところで配列長情報取得時の誤りに気づきましたm(_ _)m 2011年8月16日14:20までに一通り修正してあります。(2011/08/10-16)NEW
・FASTQ形式ファイル周辺の記述を追加しています。(2011/08/1-4)
・Bioconductorのリンク先をver. 2.7 --> 2.8に変更しました。(2011/07/20)
・R2.13.1がリリースされていたのでこれに変更しました。(2011/07/14)
・DEGseqパッケージ関連のパラメータ指定ミスを修正しました。具体的には例えば「expCol1=1」→「expCol1=2」でしたm(_ _)m(2011/06/09)

- はじめに (last modified 2011/07/19)
- Rのインストールと起動 (last modified 2011/09/14) NEW
- サンプルデータ (last modified 2011/02/03)
- イントロダクション | NGS | 各種覚書 (last modified 2010/12/10)
- イントロダクション | NGS | 様々なプラットフォーム (last modified 2011/07/15)
- イントロダクション | NGS | リファレンス配列取得(マップされる側) (last modified 2011/02/03)
- イントロダクション | NGS | リファレンス配列取得後の各種情報抽出(特にRefSeq) (last modified 2011/03/20)
- イントロダクション | NGS | リファレンス配列取得後の各種情報抽出2(readFASTA関数の利用) (last modified 2011/04/07)
- イントロダクション | NGS | アノテーション情報取得(refFlatファイル) (last modified 2010/12/07)
- イントロダクション | NGS | アノテーション情報取得(BioMart and biomaRt) (last modified 2011/08/26)
- イントロダクション | 一般 | 配列取得 (last modified 2010/7/7)
- イントロダクション | 一般 | 指定した範囲の配列を取得 (last modified 2011/07/27)
- イントロダクション | 一般 | 翻訳配列(translate)を取得 (last modified 2011/07/27)
- イントロダクション | 一般 | 相補鎖(complement)を取得 (last modified 2011/07/27)
- イントロダクション | 一般 | 逆相補鎖(reverse complement)を取得 (last modified 2011/07/27)

参考2

(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2011/09/15)

What's new?
・最新の論文(Kadota and Shimizu, BMC Bioinformatics, 2011)の結果と絡めて、よくWADに対して寄せられる質問に対する回答を追加しました。(2011/08/02)NEW
・R2.13.1がリリースされていたのでこれに変更しました。(2011/07/14)NEW
・GSA (Efron_2007)の中身をちゃんと埋め始めましたが、まだ最後までは辿りつけてません(2010/8/30)
・Hook (Binder_2008)を追加しました(2010/8/10)
・Agilent two-color processing用のRパッケージを偶然発見したので(項目のみですが...)追加しました(2010/7/14)
・作図 | ROC曲線(ROC curve)を追加しました(2010/4/20)
・このページとは直接関係ありませんが、(Rで)塩基配列解析というページで主に次世代シーケンサーデータ解析を意識したページを作成しつつありますので、そっち方面の解析をRでやりたい方はそちらをご覧ください(2010/5/27)
・作図 | ROC曲線(ROC curve)を追加しました(2010/4/20)
・Linksのところにこのページの解析結果?!を可視化させるためのプラットフォーム情報などを追加しました(2010/4/20)
・ヒートマップのところのリンク切れなどを修正しました(2010/4/9)

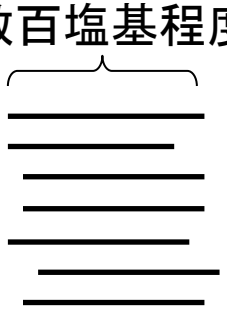
- はじめに (last modified 2009/8/7)
- Rのインストールと起動 (last modified 2011/07/14) NEW
- Rの昔のバージョンのインストール (last modified 20010/6/11)
- 使用例(初心者向け) (last modified 2011/09/15) NEW
- サンプルマイクロアレイデータ (last modified 2009/8/4)
- 発現データ取得 | Affymetrix data全体 | Celsius (Day_2007) (last modified 2007/11/13)
- 発現データ取得 | Gene Expression Omnibus (GEO)から | GEOquery (Davis_2007) (last modified 2009/8/5)
- 発現データ取得 | ArrayExpressから | ArrayExpress (last modified 2009/5/28)
- アノテーション情報取得 | Rのパッケージから (last modified 2009/8/5)
- アノテーション情報取得 | GEOから (last modified 2009/8/5)
- 正規化(cDNA or two-color or 二色法) について(last modified 2008/3/31)
- 正規化 | Stanford型 (or cDNA)マイクロアレイ (package: limma)
- 正規化 | Stanford型 (or cDNA)マイクロアレイ (package: marray)

シーケンサー新旧比較

- 旧世代シーケンサー（ABI3730など）
  - 800塩基程度の長さを読める
  - 数は少ない
  - 質は高い
- 次世代シーケンサー
  - 長さは短い（～数百塩基程度）
  - 数は多い
  - 質は低い

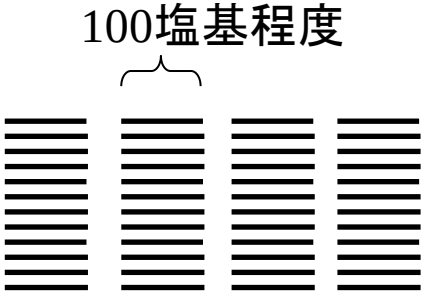
GS FLX+

数百塩基程度



5500xl SOLiD　HiSeq 2000

100塩基程度



ゲノム配列の決定はどうやって？

- 素朴な疑問（何が大変なのかわからない…）
  - Q：ゲノムが長い一本の塩基配列で、旧世代シーケンサーが一度に800塩基程度しか読めないのはわかった。だけど読めるところまで読んだら、最後の塩基のところからまた順番に読んでいけばいいじゃん！
  - A：それができないのでゲノムを物理的に切断した断片配列の配列決定（シークエンシング；sequencing）を行います。800塩基程度の配列の集合が手元にあるだけです。

どうやって、元のゲノム配列を再構築するのか？

de novo genome assembly

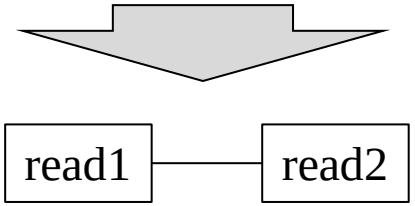
- *de novo*:「初めから、新規に」の意味
- 配列決定されたリードのみから、目的生物種のゲノム配列を決めること(組み立てること)
- 方法による分類(Miller et al., *Genomics*, **95**: 315-327, 2010)
  - Overlap-Layout-Consensus (OLC)アプローチ
    - 各リードを頂点(ノード)として、k個の共通連続塩基がある頂点同士を辺(エッジ)で結んだグラフを作成し、全ての頂点を通るパスを探索(ハミルトンパス問題)
    - 配列一致部分がある程度の長さ分必要なため、Roche 454など比較的長いリードのアセンブルに用いられる
  - Euler (or Eulerian path)アプローチ
    - リードを一塩基づつずらしたk個の連続塩基からなるk-merグラフを各リードごとに作成し、全リードの完全一致ノードをマージすることで「de Bruijnグラフ」を作成し、全ての辺を通るパスを探索(オイラーパス問題)
    - Illuminaなどの比較的短いリードのアセンブルに用いられる

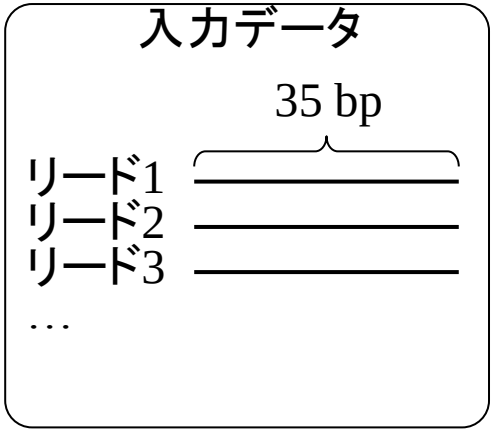
Overlap-Layout-Consensus (OLC)アプローチ

- k個（例：k=25）の共通連続塩基があるリード（頂点）同士を辺でむすぶ

read1：TGCCGACATGCATCCAAGTAGGAATCCTTAGCTTA

read2：CATCCAAGTAGGAATCCTTAGCTTAGCCAATGCGT

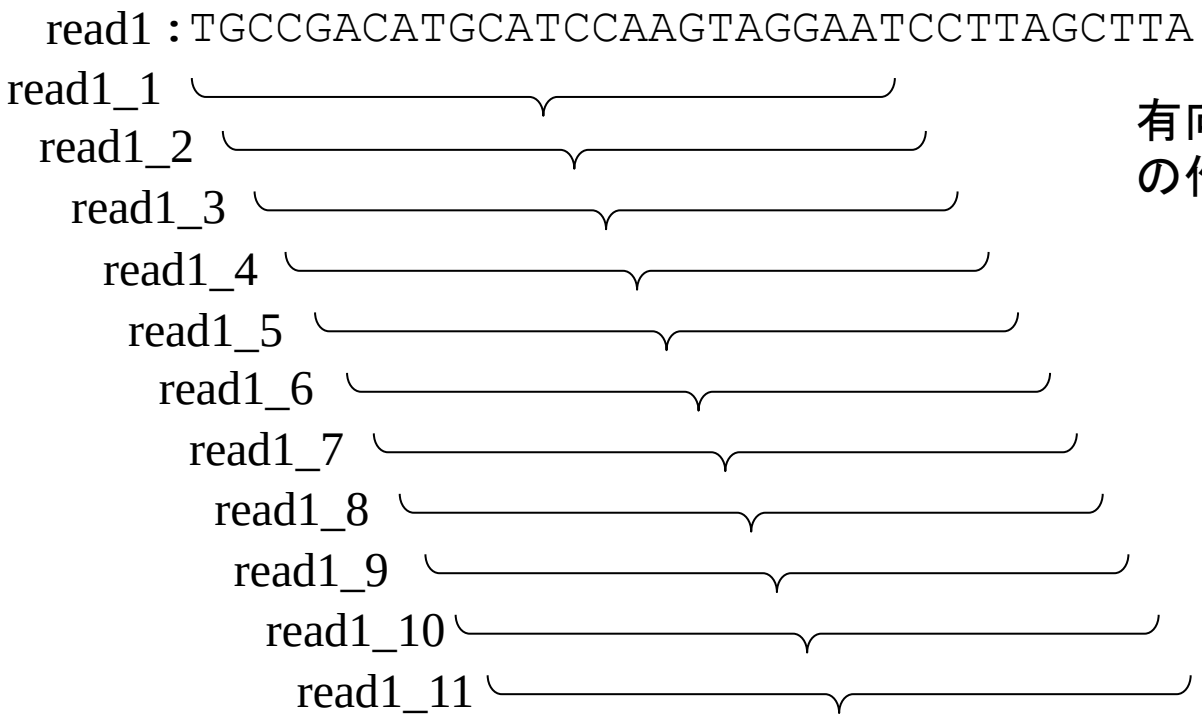


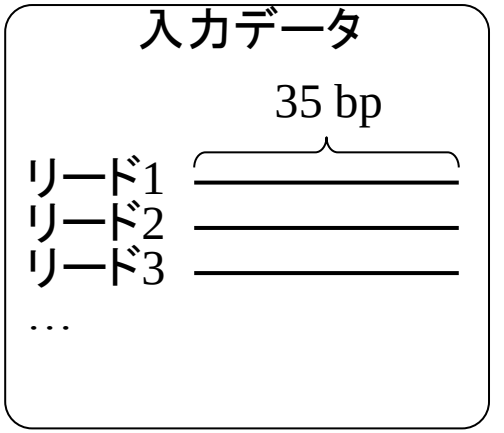


アセンブル＝全ての**頂点**を通るパス（経路）を探索すること

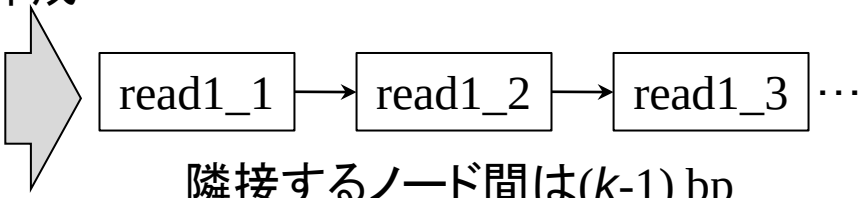
Euler (or Eulerian path)アプローチ

- 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





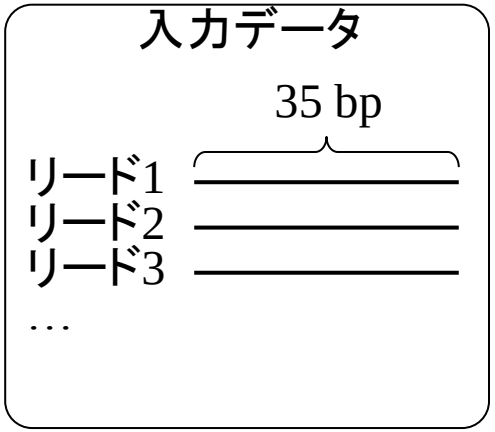
有向グラフの作成



隣接するノード間は(*k*-1) bpのオーバーラップ

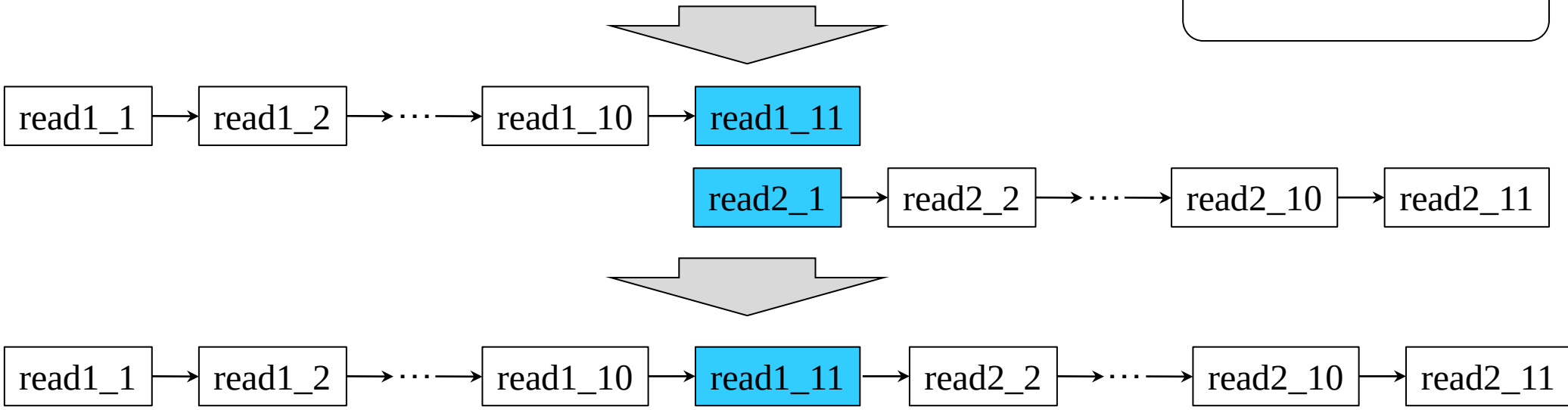
Euler (or Eulerian path)アプローチ

- 同一ノードをマージ



read1：TGCCGACATGCATCCAAGTAGGAATCCTTAGCTTA

read2：CATCCAAGTAGGAATCCTTAGCTTAGCCAATGCGT



全リードの情報をもとに同一ノードをマージしたグラフ（de Bruijnグラフ）
アセンブル＝全ての辺を通るパスを探索すること

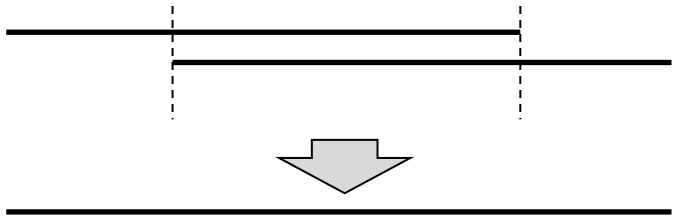
de novo genome assembly

- Overlap-Layout-Consensus (OLC)アプローチ
  - 生物種：Drosophila (Myers et al., *Science*, **287**: 2196-2204, 2000)
  - 全ゲノムショットガン
- Euler (or Eulerian path)アプローチ
  - 生物種：Giant panda (Li et al., *Nature*, **463**: 311-317, 2010)
  - Illumina Genome Analyzer (37paired-end)

パンダゲノムはたまたまうまくいった？！配列さえ読めばあとはボタン一つ押せばアセンブルされたゲノムが得られる…ほど簡単ではない

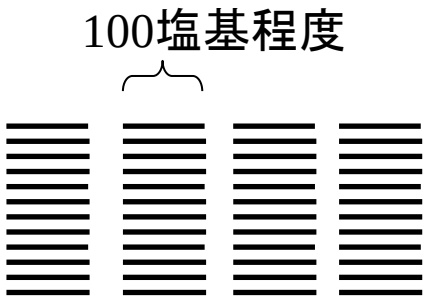
新規の（de novo）ゲノム配列決定は大変

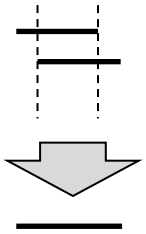
- 旧世代シーケンサー（ABI3730など）



一致（のりしろ）部分の領域大
→ 信頼性高い

- 次世代シーケンサー





一致（のりしろ）部分の領域小
→ 信頼性低い

NGSでゲノム解読の成果は続々と…?

- パンダ(大熊猫)ゲノム解読(2008年)
  - ヒトゲノム解読に10年 → 半年
  - 猫よりも犬・熊に近い動物
- アジア人(中国人)一個体の全ゲノム配列決定 (2008年)
- 国際プロジェクト
  - 1000
  - 国際
  - 感染症の同定
- 日本人の全ゲノム配列決定(2010年)

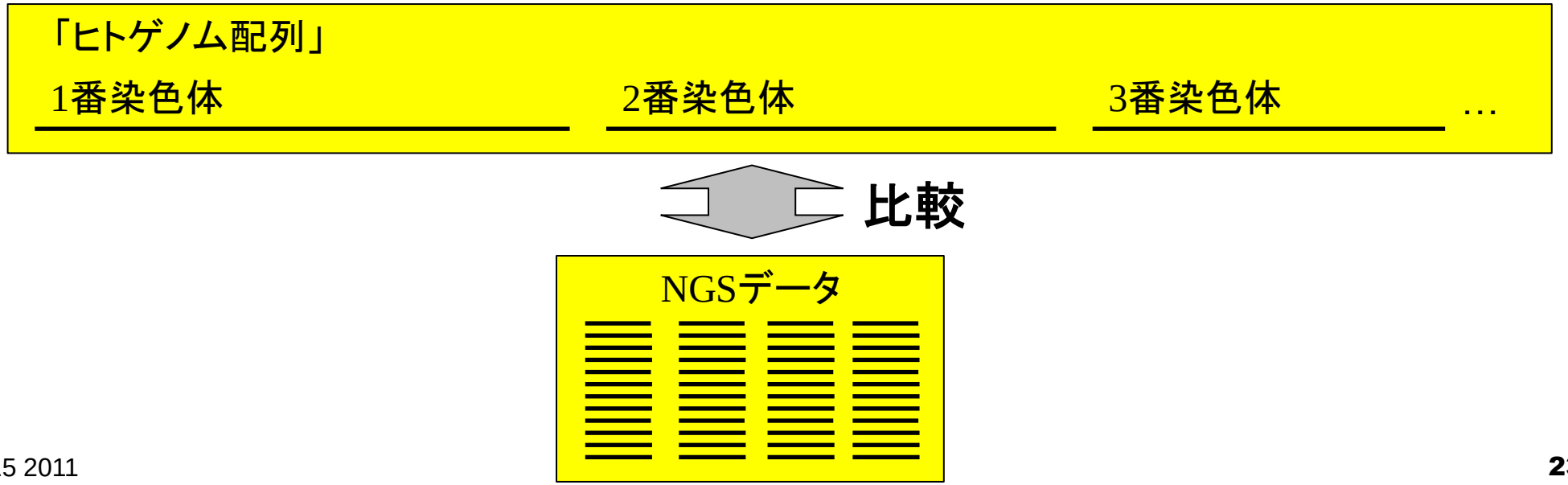
Resequencing(再配列決定)

- **世界で初めてサンゴの全ゲノム解読に成功**(2011年7月)
  - サンゴと褐虫藻との共生メカニズム解明のための基盤情報取得
  - サンゴの白化現象(褐虫藻を失うこと)解明のための～
  - サンゴ礁の観光産業などの経済効果は2,500億円以上！

http://www.oist.jp/ja/press-room/press-releases/150/1047-pr-first-coral-genome.html

Resequencing

- 既知の塩基配列と次世代シーケンサー（NGS）から得られた短い塩基配列（short read）を**比較**すること
  - ヒトゲノム配列は旧世代シーケンサーを用いて解読済み
  - 例：「日本人ゲノム解読」は、次世代シーケンサーを用いて日本人のNGS塩基配列データを取得し、「ヒトゲノム配列」と**比較**して、日本人特有の領域や配列の違いなどを発見しました、ということ。

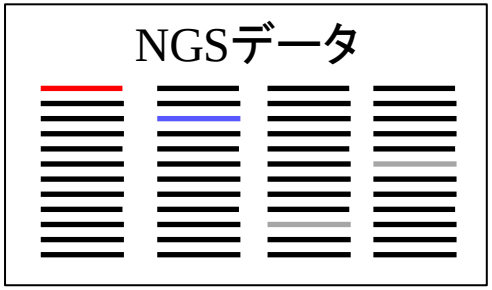


比較？

- NGSデータ中の数千万リード（一が数千万個あるということ）の各々がゲノム中のどこにマップされるか、マップされないのはどれか、などを調べるイメージ

「ヒトゲノム配列」

1番染色体　2番染色体　3番染色体 …



Linux上で動くNGSデータ専用のマッピングプログラムを用いて実行できます

NGS解析はLinux上で行うのが基本

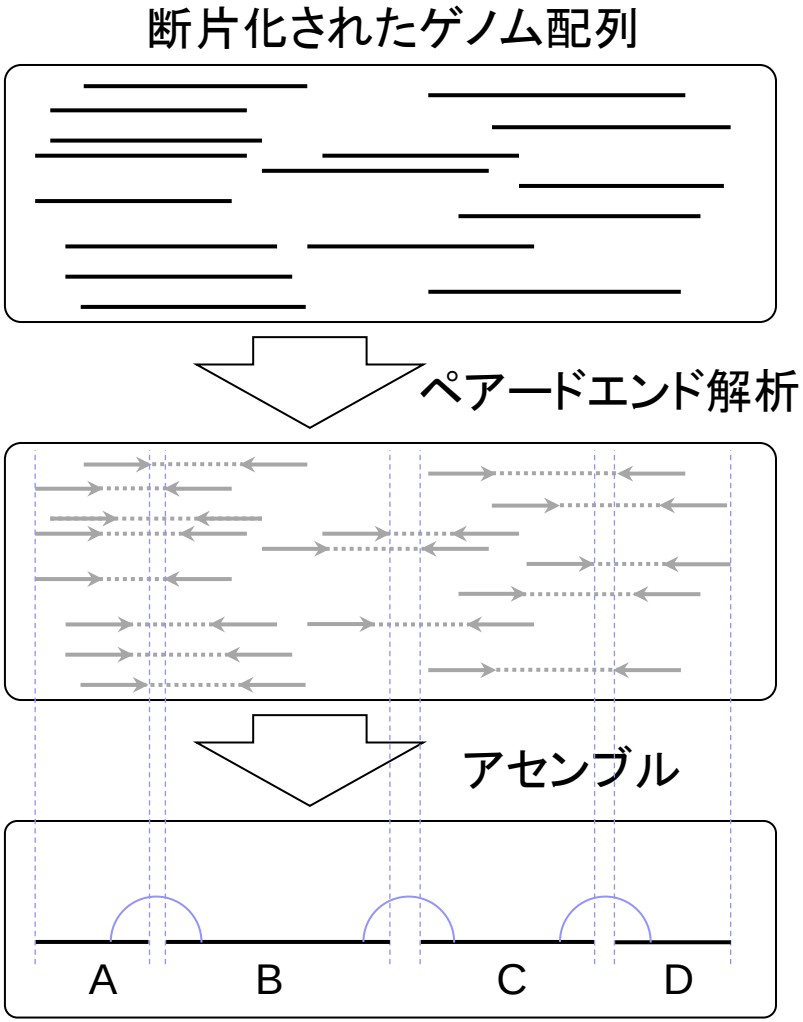
- 理由1：de novo assemblyやマッピングなどの基本的な解析部分を行うプログラムはLinux (UNIX)用が大多数
- 理由2：その後の解析はWindows版のRでもできるが、Linux版のRでもできる（しかも速い！）

→Linuxに慣れてる人は、Rを使って行う解析もLinux上でやる

Linuxを使いこなせるのがベストであることは間違いない

用語解説

- リード
  - Sequencerで読んだ塩基配列のこと
- コンティグ
  - 異なる複数のリードがACGTの切れ目なく連結されたもの
  - 右図ではA-Dの四つのコンティグ
- Scaffold
  - コンティグ間の位置関係を表したもの
  - 「A-D-B-C」ではなく「A-B-C-D」という関係
- N50
  - 得られた複数のコンティグを最も長いコンティグから順番に連結していったときにcombined total lengthの50%になったときのコンティグの長さ



用語解説など

- Coverage（カバレッジ）
  - ゲノム解読したいときなどに、解読するために必要とされる指標となる数値。ゲノムサイズ(X)に対する、sequencerで読んだ塩基配列長の和のこと。一般に、この数値が高ければ高いほどよい。
- kの数はいくつがいいの？
  - わかりません。。。複数のkの値を試すみたいです。
- アセンブル結果の評価基準は？
  - よくわかりません。平均コンティグ長やN50が論文の表でよく記述されます。このあたりの数値を大きくするだけなら、kの値を小さめにすればいいのですが、同時にそれはキメラコンティグを形成してしまう確率が上昇することを意味するからです。
- アセンブルプログラムを実行して得られる出力ファイルはどんな感じ？
  - （基本的に）multi-fasta形式のファイルです。

```
>contig1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
…
>contig2
ACGATGCAGCCTTAACGA…
>contig3
…
```

FASTQ形式（とFASTA形式）

- FASTA形式
  - 「“>”ではじまる一行のdescription行」と「配列情報」からなる形式
  - NGSのread長は短いので、実質的に一つのリードを二行で表現

```
>SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
```

- FASTQ形式
  - 一行目：「“@”ではじまる一行のdescription行」
  - 二行目：「配列情報」
  - 三行目：「”+”からはじまる一行（のdescription行）」
  - 四行目：「クオリティ情報」

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

http://en.wikipedia.org/wiki/FASTQ_format

塩基配列のクオリティ情報といえば…

- Phredスコア
  - Phredというベースコールプログラムから得られるQuality Value（QV値）のこと

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

http://en.wikipedia.org/wiki/Phred_quality_score

なぜFASTQ形式では、Phredスコアそのものでクオリティ情報を表現しないの？

理由:(容量)節約のため

Phred スコア	ASCII印字可能文字		Phred スコア	ASCII印字可能文字	
0	!	ASCII 33	21	6	ASCII 54
1	"	ASCII 34	22	7	ASCII 55
2	#	ASCII 35	23	8	ASCII 56
3	$	ASCII 36	24	9	ASCII 57
4	%	ASCII 37	25	:	ASCII 58
5	&	ASCII 38	26	;	ASCII 59
6	'	ASCII 39	27	<	ASCII 60
7	(	ASCII 40	28	=	ASCII 61
8	)	ASCII 41	29	>	ASCII 62
9	*	ASCII 42	30	?	ASCII 63
10	+	ASCII 43	31	@	ASCII 64
11	,	ASCII 44	32	A	ASCII 65
12	-	ASCII 45	33	B	ASCII 66
13	.	ASCII 46	34	C	ASCII 67
14	/	ASCII 47	35	D	ASCII 68
15	0	ASCII 48	36	E	ASCII 69
16	1	ASCII 49	37	F	ASCII 70
17	2	ASCII 50	38	G	ASCII 71
18	3	ASCII 51	39	H	ASCII 72
19	4	ASCII 52	40	I	ASCII 73
20	5	ASCII 53	…	…	…

- FASTQ形式中のクオリティ情報部分

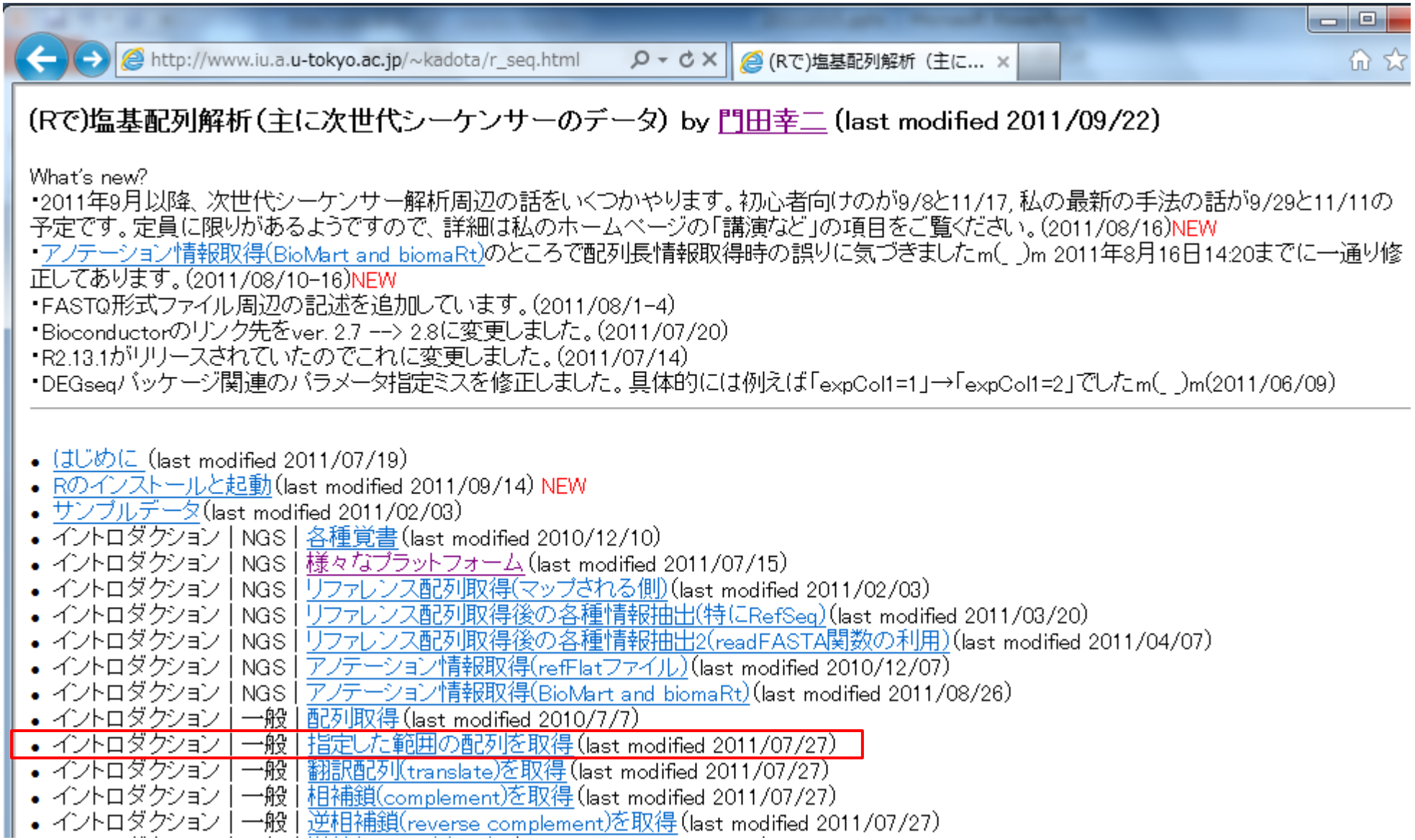
```
@SRR037439.375
GCGGTGTGTTTGTGGTATAGTGGTGCCCCGCCCCG
+SRR037439.375
IIII&IIIII?223<(<I2B*4@#/I"#"#''""+
```

- Phredスコア(QUAL形式)

```
40 40 40 40  5 40 40 40 40 40 30 17 17 18 27  7 27 40 17 33
 9 19 31  2 14 40  1  2  1  2  6  6  1  1 10
```

Phredスコアが*X*の場合「ASCII (*X*+33)」に対応する文字コードを割り当てる

実習



(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2011/09/22)

What's new?
・2011年9月以降、次世代シーケンサー解析周辺の話をいくつかやります。初心者向けのが9/8と11/17, 私の最新の手法の話が9/29と11/11の予定です。定員に限りがあるようですので、詳細は私のホームページの「講演など」の項目をご覧ください。(2011/08/16)NEW
・アノテーション情報取得(BioMart and biomaRt)のところで配列長情報取得時の誤りに気づきましたm(_ _)m 2011年8月16日14:20までに一通り修正してあります。(2011/08/10-16)NEW
・FASTQ形式ファイル周辺の記述を追加しています。(2011/08/1-4)
・Bioconductorのリンク先をver. 2.7 --> 2.8に変更しました。(2011/07/20)
・R2.13.1がリリースされていたのでこれに変更しました。(2011/07/14)
・DEGseqパッケージ関連のパラメータ指定ミスを修正しました。具体的には例えば「expCol1=1」→「expCol1=2」でしたm(_ _)m(2011/06/09)

- はじめに (last modified 2011/07/19)
- Rのインストールと起動 (last modified 2011/09/14) NEW
- サンプルデータ (last modified 2011/02/03)
- イントロダクション | NGS | 各種覚書 (last modified 2010/12/10)
- イントロダクション | NGS | 様々なプラットフォーム (last modified 2011/07/15)
- イントロダクション | NGS | リファレンス配列取得(マップされる側) (last modified 2011/02/03)
- イントロダクション | NGS | リファレンス配列取得後の各種情報抽出(特にRefSeq) (last modified 2011/03/20)
- イントロダクション | NGS | リファレンス配列取得後の各種情報抽出2(readFASTA関数の利用) (last modified 2011/04/07)
- イントロダクション | NGS | アノテーション情報取得(refFlatファイル) (last modified 2010/12/07)
- イントロダクション | NGS | アノテーション情報取得(BioMart and biomaRt) (last modified 2011/08/26)
- イントロダクション | 一般 | 配列取得 (last modified 2010/7/7)
- イントロダクション | 一般 | 指定した範囲の配列を取得 (last modified 2011/07/27)
- イントロダクション | 一般 | 翻訳配列(translate)を取得 (last modified 2011/07/27)
- イントロダクション | 一般 | 相補鎖(complement)を取得 (last modified 2011/07/27)
- イントロダクション | 一般 | 逆相補鎖(reverse complement)を取得 (last modified 2011/07/27)

http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#intr (Rで)塩基配列解析（主に...

- イントロダクション | 一般 | 指定した範囲の配列を取得

例1では、12塩基(AGTGACGGTCTT)からなる一つの塩基配列(description行が">kadota")からなるFASTA形式ファイル(sample1.fasta)を入力として、この塩基配列の任意の範囲 (始点が3, 終点が9)の配列を抽出し、得られた部分配列をFASTA形式ファイル(tmp.fasta)に出力するやり方を示します。
例2では、RefSeqのhuman mRNAのmulti-fasta形式のファイル (h_rna.fasta)が手元にあったとして、任意のRefSeq ID (例: NM_203348.1)の任意の範囲 (例:始点が2, 終点が5)の配列の抽出を行います。
ここでは、得られた部分配列をFASTA形式ファイル (ファイル名: tmp.fasta) で保存するやり方を示します。

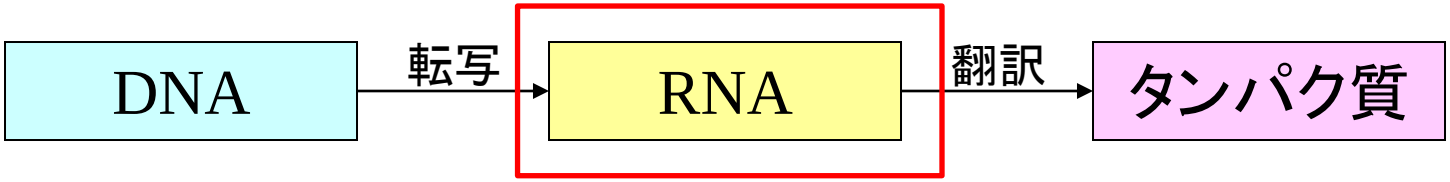
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

```
#例1：sample1.fastaファイルの場合
------    ここから    ------
in_f <- "sample1.fasta"                                  #multi-fasta形式のファイルを指定
out_f <- "tmp.fasta"                                     #出力ファイル名を指定
param <- c(3, 9)                                         #抽出したい範囲の始点と終点を指定

library(Biostrings)                                      #パッケージの読み込み
reads <- read.DNAStringSet(in_f, format="fasta")         #in_fで指定したファイルをFASTA形式で読み込み
out <- subseq(reads, param[1], param[2])                 #paramで指定した始点と終点の範囲の配列を抽出してoutに格納
write.XStringSet(out, file=out_f, format="fasta", width=80)      #outの中身をout_fで指定したファイル名で保存

------    ここまで    ------
#例2：h_rna.fastaファイルの場合
------    ここから    ------
in_f <- "h_rna.fasta"                                    #multi-fasta形式のファイルを指定
out_f <- "tmp.fasta"                                     #出力ファイル名を指定
param1 <- "NM_203348.1"                                  #取得したい配列のアクセッション番号を指定
param2 <- c(2, 5)                                        #抽出したい範囲の始点と終点を指定

library(Biostrings)                                      #パッケージの読み込み
reads <- read.DNAStringSet(in_f, format="fasta")         #in_fで指定したファイルをFASTA形式で読み込み
head(reads)                                              #readsオブジェクトの最初の一部を表示(ちなみに最後の一部を表示させたい場合
names(reads)                                             #readsオブジェクトのID (description部分)を表示させたい場合
obj <- reads[names(reads) == param1]                     #param1で指定したIDの配列のみ抽出してobjに格納
out <- subseq(obj, param2[1], param2[2])                 #param2で指定した始点と終点の範囲の配列を抽出してoutに格納
write.XStringSet(out, file=out_f, format="fasta", width=80)      #outの中身をout_fで指定したファイル名で保存

------    ここまで    --
```

Rはただの統計解析フリーソフトではありません

BioconductorのBiostringsのwebページ

NGSを用いたトランスクリプトーム解析

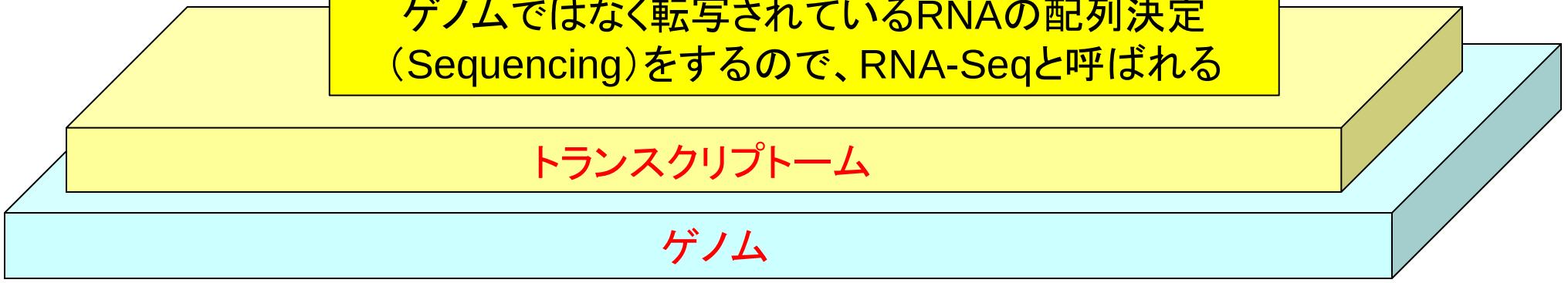


二次元電気泳動法

次世代シーケンサー（NGS）

GS FLX+　　5500xl SOLiD　　HiSeq 2000

ゲノムではなく転写されているRNAの配列決定（Sequencing）をするので、RNA-Seqと呼ばれる



トランスクリプトームとは

- ある特定の状態の組織や細胞中に存在する全RNA（転写物、transcripts）の総体
- 様々なトランスクリプトーム解析技術
  - マイクロアレイ
    - cDNAマイクロアレイ、Affymetrix GeneChip、タイリングアレイなど
  - 配列決定に基づく方法
    - EST、SAGEなど、次世代シーケンサー (NGS)
  - 電気泳動に基づく方法
    - Differential Display、AFLPなど

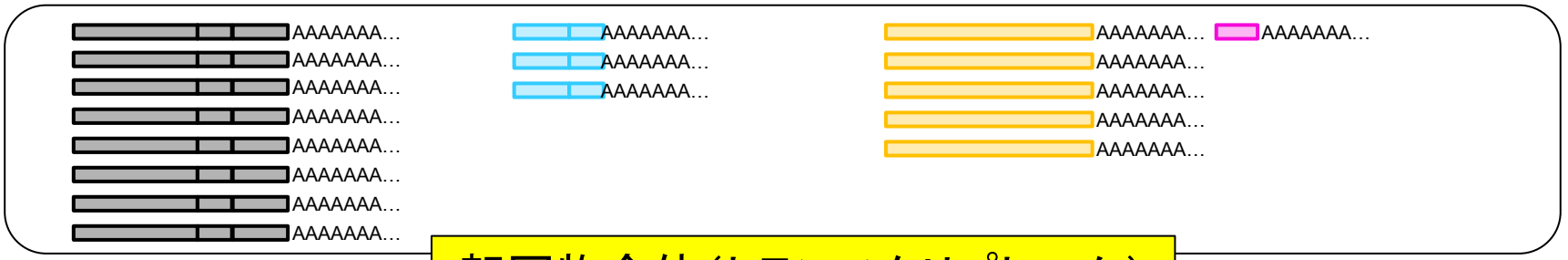
調べたい組織でどの遺伝子がどの程度発現しているのかを一度に観察

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）

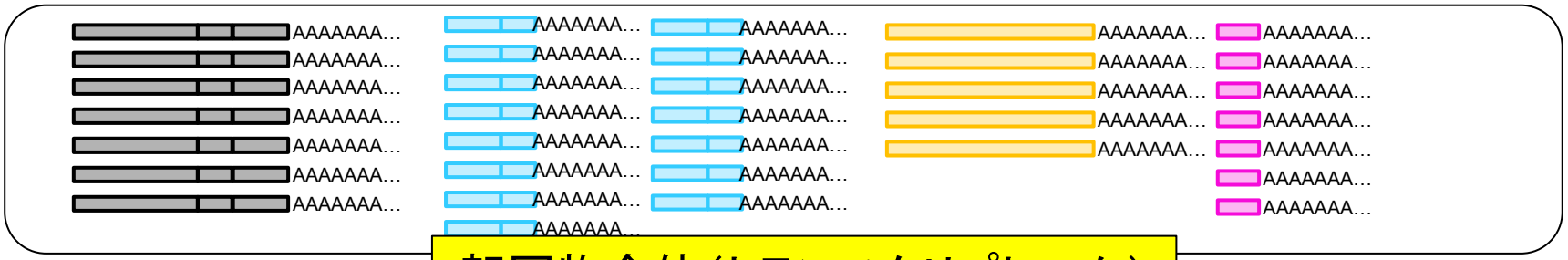
・遺伝子1は沢山転写されている（発現している）
・遺伝子4はごくわずかしか転写されてない
・…

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…）

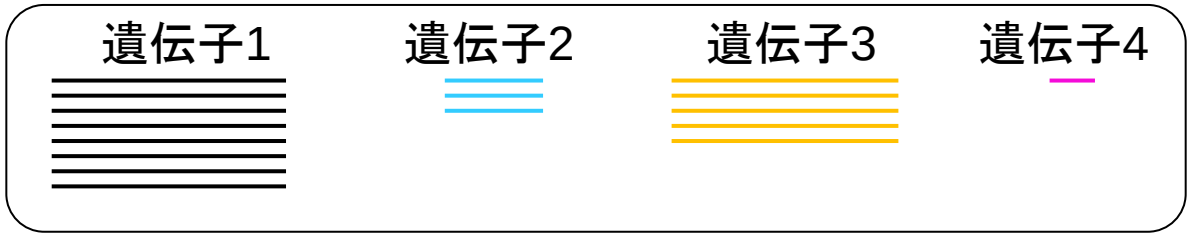


転写物全体（トランスクリプトーム）

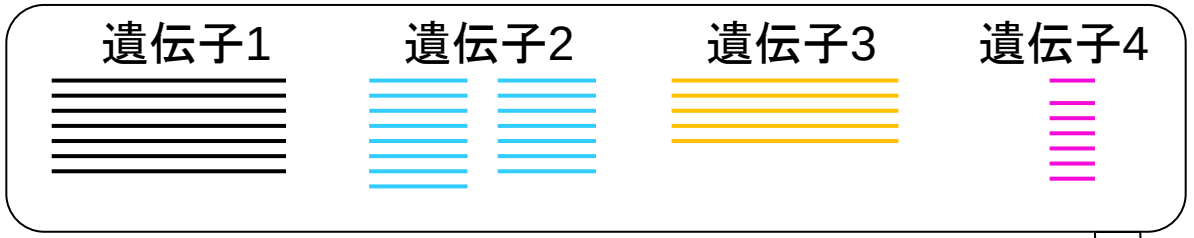
・遺伝子2は光刺激に応答して発現亢進
・遺伝子4も光刺激に応答して発現亢進

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



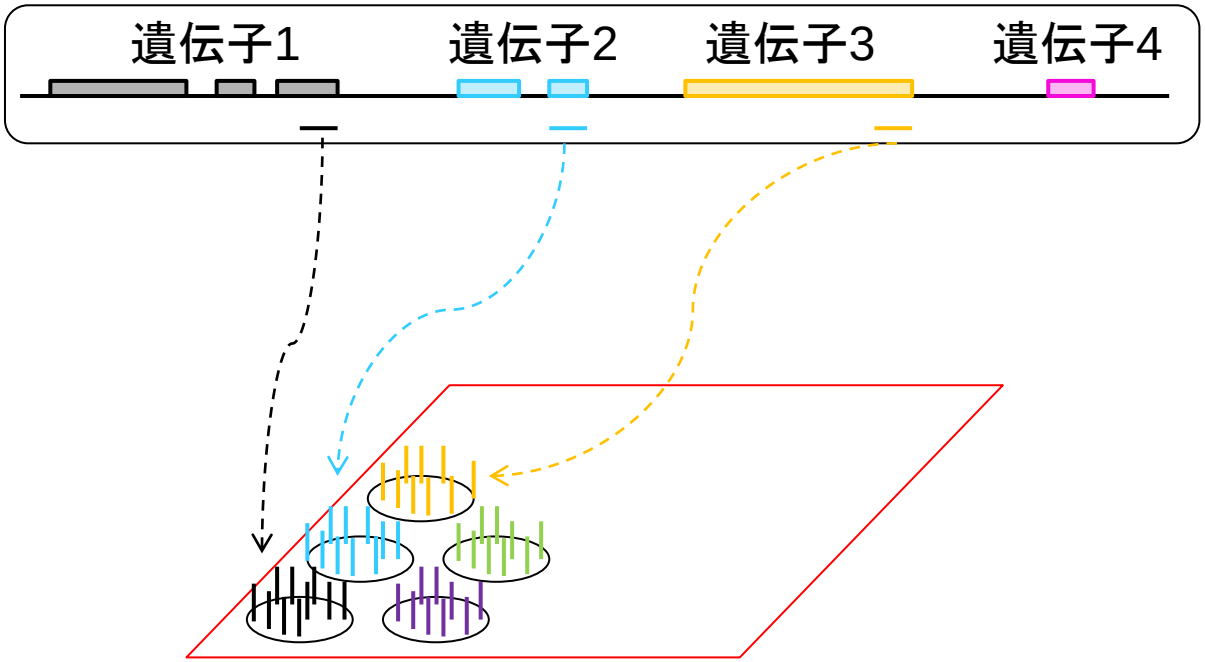
これがいわゆる
「遺伝子発現行列」

	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
- RNA-Seq（NGS）
- SAGE
- …

トランスクリプトーム取得（マイクロアレイ）

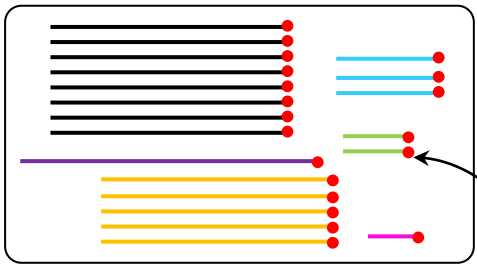
- よく研究されている生き物は多数の遺伝子（の配列情報）がわかっている



わかっている遺伝子（の配列の相補鎖）を搭載した”チップ”

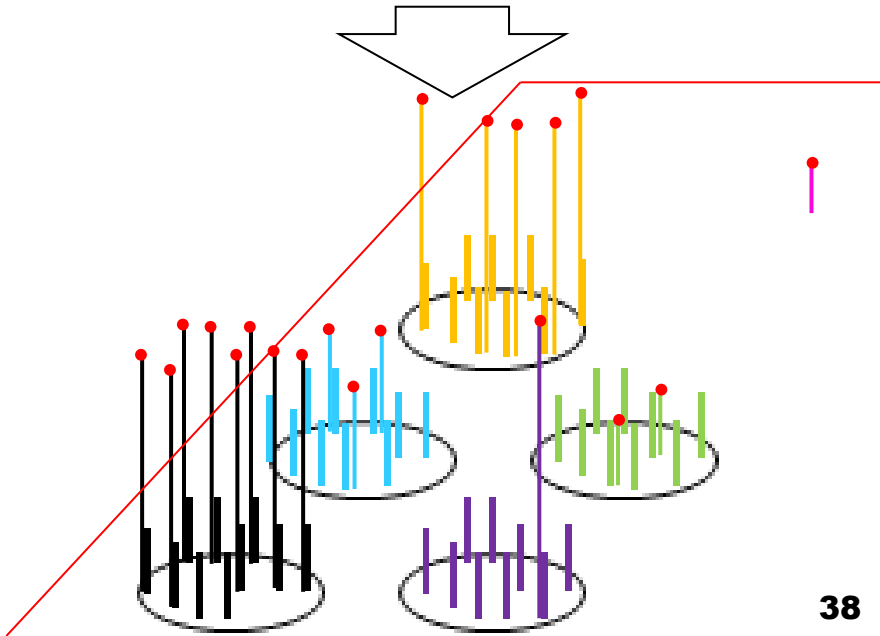
・メーカーによって搭載されている遺伝子の種類が異なる
→搭載されていない遺伝子（未知遺伝子含む、例：遺伝子4）の発現情報は測定不可…

光刺激前（T1）の目のトランスクリプトーム



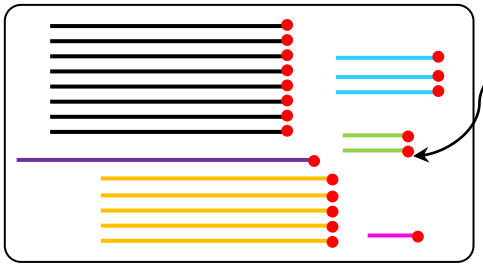
蛍光標識

ハイブリダイゼーション（二本鎖形成）

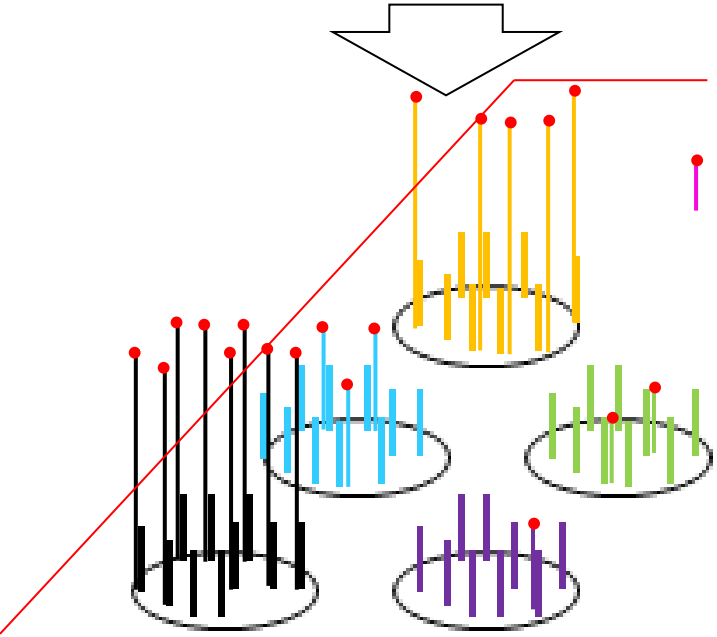


マイクロアレイデータ → 遺伝子発現行列

- 光刺激前（T1）の目のトランスクリプトーム

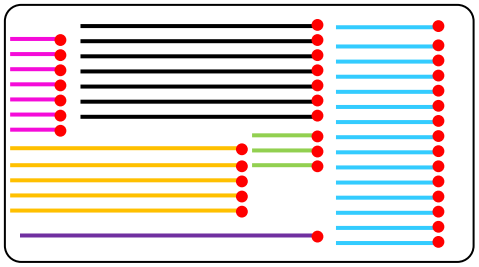


ハイブリダイゼーション
（二本鎖形成）



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



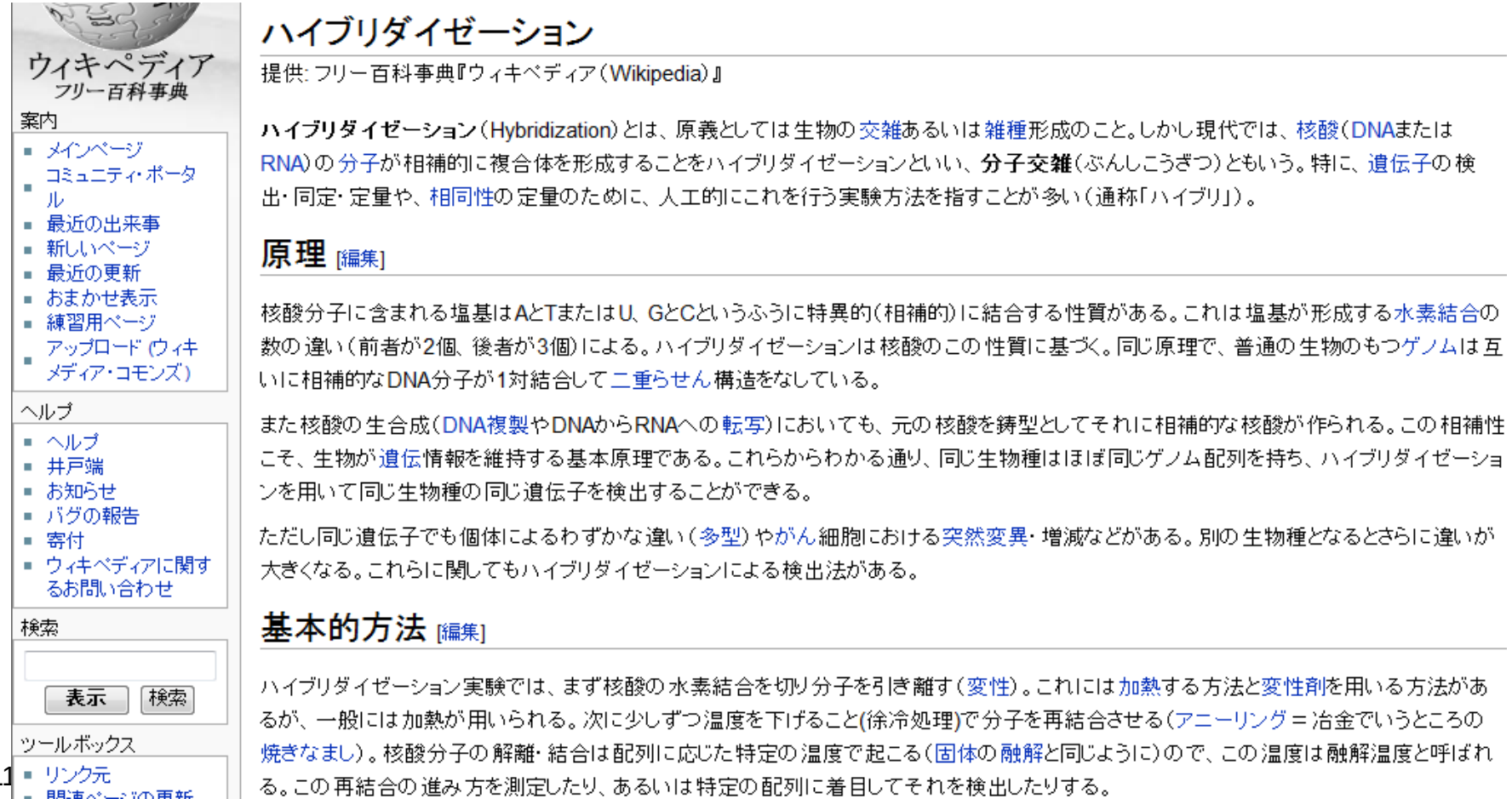
ハイブリダイゼーション
と
シグナル検出

	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

ハイブリダイゼーション

- 核酸（DNA or RNA）分子が相補的に複合体を形成すること
  - 核酸分子に含まれる塩基はAとT（or U）またはGとCというふうに相補的に結合する性質があるので、この性質を利用



ウィキペディア
フリー百科事典

案内
- メインページ
- コミュニティ・ポータル
- 最近の出来事
- 新しいページ
- 最近の更新
- おまかせ表示
- 練習用ページ
- アップロード (ウィキメディア・コモンズ)

ヘルプ
- ヘルプ
- 井戸端
- お知らせ
- バグの報告
- 寄付
- ウィキペディアに関するお問い合わせ

検索
表示 検索

ツールボックス
- リンク元

ハイブリダイゼーション

提供: フリー百科事典『ウィキペディア（Wikipedia）』

ハイブリダイゼーション（Hybridization）とは、原義としては生物の交雑あるいは雑種形成のこと。しかし現代では、核酸（DNAまたはRNA）の分子が相補的に複合体を形成することをハイブリダイゼーションといい、**分子交雑**（ぶんしこうざつ）ともいう。特に、遺伝子の検出・同定・定量や、相同性の定量のために、人工的にこれを行う実験方法を指すことが多い（通称「ハイブリ」）。

原理 [編集]

核酸分子に含まれる塩基はAとTまたはU、GとCというふうに特異的（相補的）に結合する性質がある。これは塩基が形成する水素結合の数の違い（前者が2個、後者が3個）による。ハイブリダイゼーションは核酸のこの性質に基づく。同じ原理で、普通の生物のもつゲノムは互いに相補的なDNA分子が1対結合して二重らせん構造をなしている。

また核酸の生合成（DNA複製やDNAからRNAへの転写）においても、元の核酸を鋳型としてそれに相補的な核酸が作られる。この相補性こそ、生物が遺伝情報を維持する基本原理である。これらからわかる通り、同じ生物種はほぼ同じゲノム配列を持ち、ハイブリダイゼーションを用いて同じ生物種の同じ遺伝子を検出することができる。

ただし同じ遺伝子でも個体によるわずかな違い（多型）やがん細胞における突然変異・増減などがある。別の生物種となるとさらに違いが大きくなる。これらに関してもハイブリダイゼーションによる検出法がある。

基本的方法 [編集]

ハイブリダイゼーション実験では、まず核酸の水素結合を切り分子を引き離す（変性）。これには加熱する方法と変性剤を用いる方法があるが、一般には加熱が用いられる。次に少しずつ温度を下げること(徐冷処理)で分子を再結合させる（アニーリング＝冶金でいうところの焼きなまし）。核酸分子の解離・結合は配列に応じた特定の温度で起こる（固体の融解と同じように）ので、この温度は融解温度と呼ばれる。この再結合の進み方を測定したり、あるいは特定の配列に着目してそれを検出したりする。

マイクロアレイは実績がある

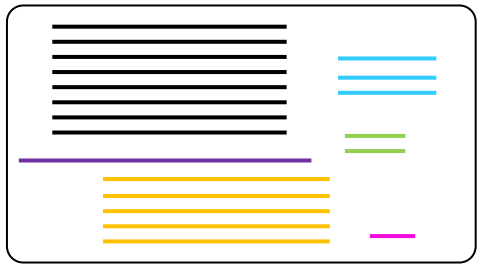
- 「MammaPrint」: 乳癌予後予測検査サービス
  - 2008年3月
  - 乳癌手術を受けた患者の転移・再発の可能性に関する情報提供
  - 70遺伝子の活性を測定
  - 不必要な補助化学療法などを避けることが可能（ローリスク群）
- 「oncotype DX」: 早期浸潤性乳癌の術後再発予測サービス
  - 2007年2月
  - 再発リスクの数値化および化学療法の効果予測
  - 21遺伝子を解析
  - 必要以上の化学療法を回避
- 「GeneSearch」: 乳癌の術中リンパ節転移迅速診断
  - 2007年7月

既に臨床診断に利用されている

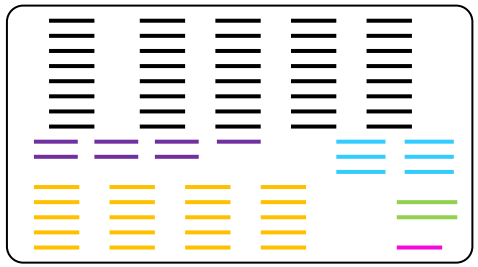
RNA-Seqデータ → 遺伝子発現行列

- 次世代シーケンサー（Illumina社の場合）

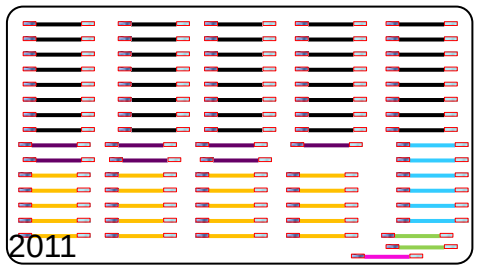
光刺激前（T1）の目のトランスクリプトーム



数百塩基程度
に断片化



二種類のアダプター
配列を両末端に付加

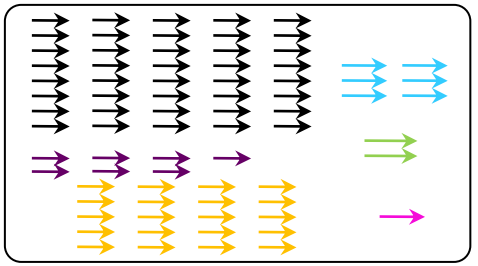


配列決定

・ペアードエンド法
断片配列の両末端が数百塩基以内の対の二種類の配列が得られる

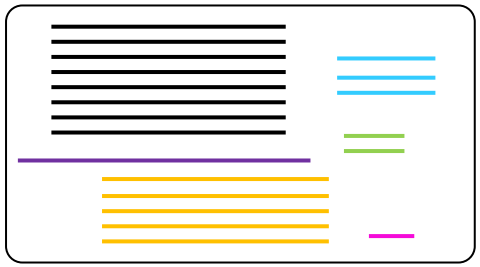
・シングルエンド法

シングルエンド法
の場合

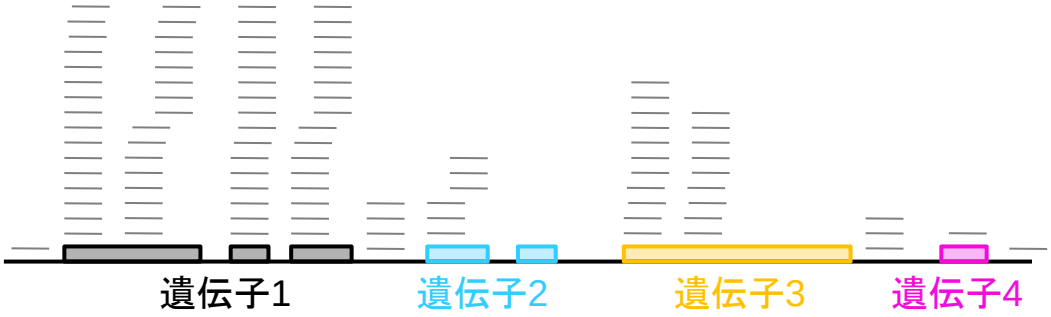


RNA-Seqデータ → 遺伝子発現行列

光刺激前（T1）の目のトランスクリプトーム



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)

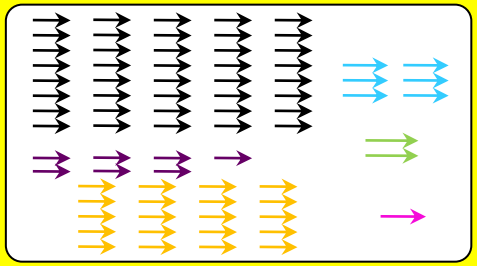
	T1
遺伝子1	40
遺伝子2	6
遺伝子3	20
遺伝子4	1
遺伝子5	…
…	…

正規化

T1
8
3
5
1
…
…

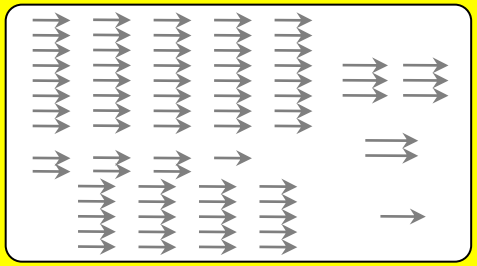
－イメージ－

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

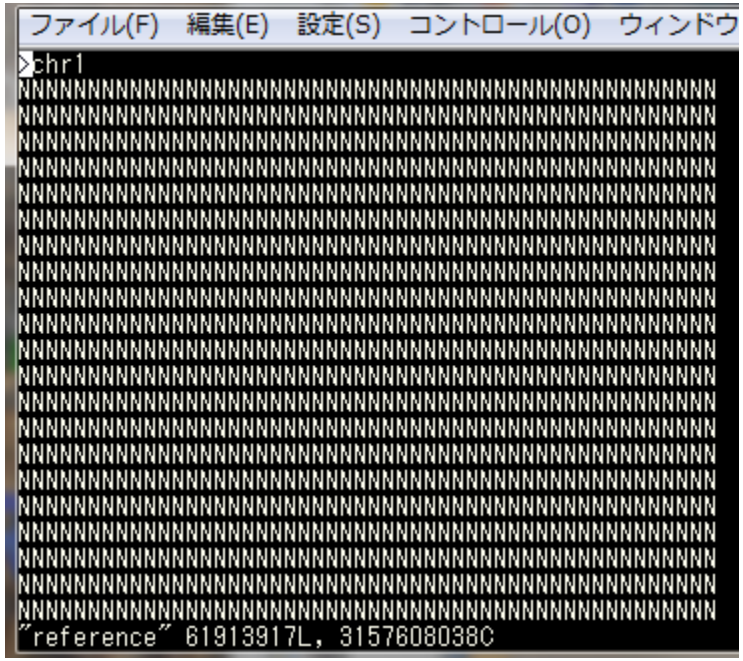


（短い）配列を読んだものという意味で（ショート）リードなどと呼ばれる

ゲノムにマップ

- 実データ(ヒトの場合)

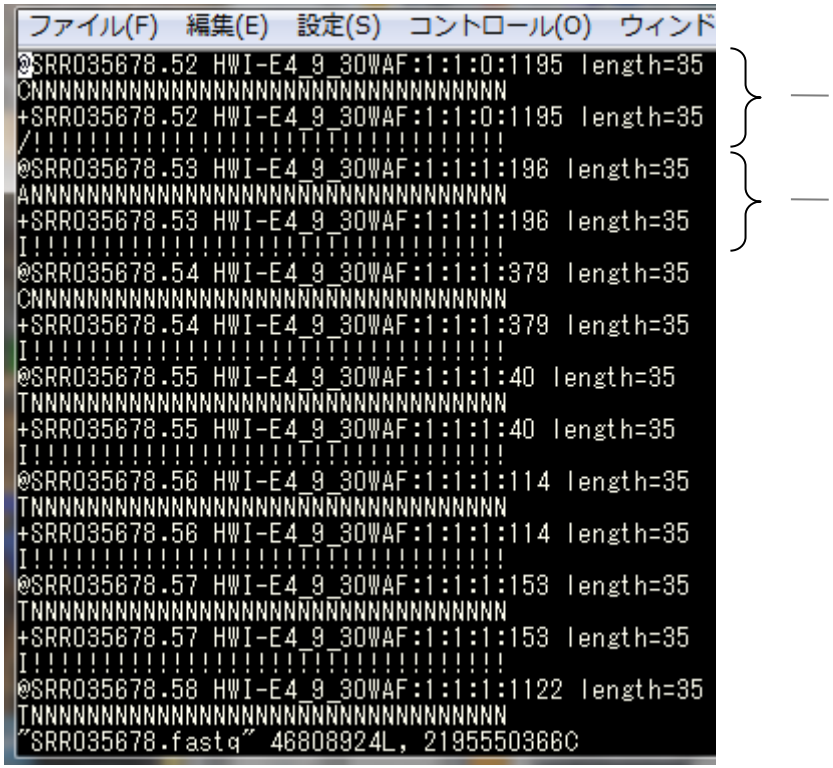
マップされる側のリファレンスゲノム配列



・1-22番染色体＋X＋Y
・約6200万行のファイル
・約3GBのサイズ

chr1
chr2
…

マップする側の塩基配列(FASTQ形式)

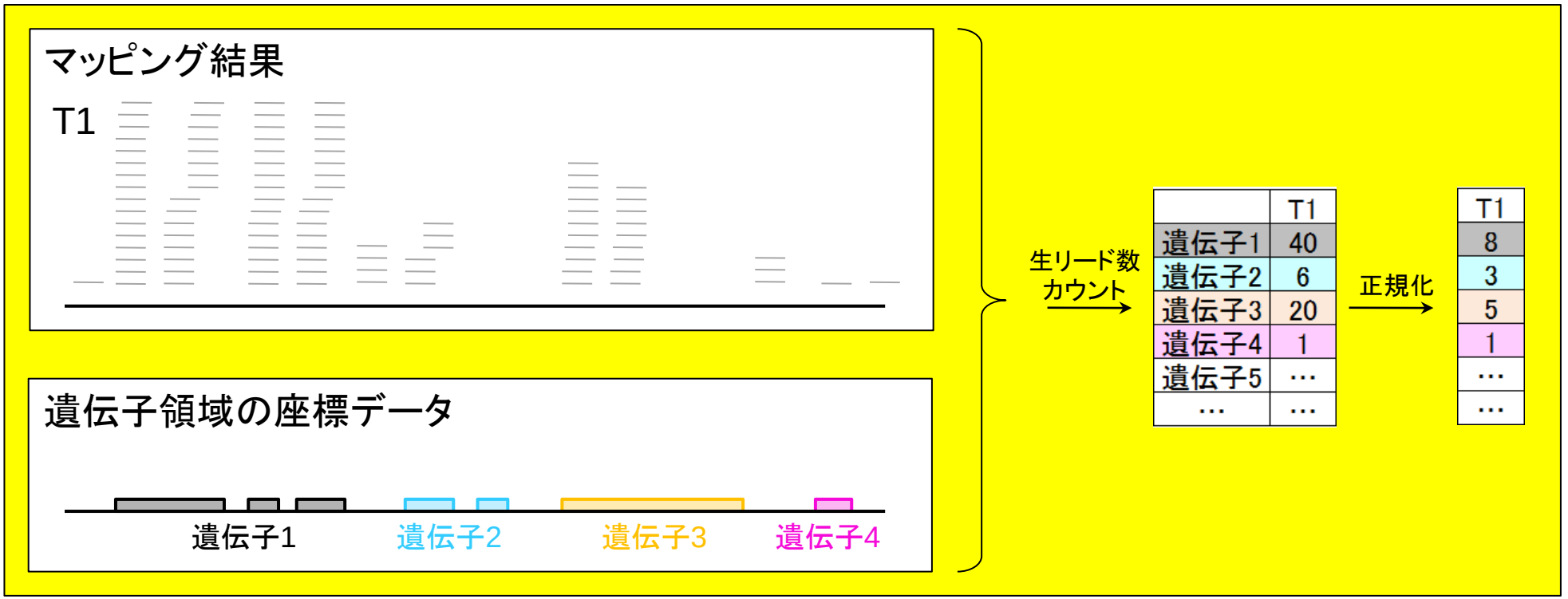


・約47万行
・1 配列(1 read)を4行で表現
・1配列35塩基長(この場合)

各readが染色体上のどこに一致するか
座標情報を出力するのがマッピングプログラム

塩基配列データ → 遺伝子発現行列

- 遺伝子領域の座標データがないと遺伝子発現行列は作れない



	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	…
…	…

T1
8
3
5
1
…
…

(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2010/7/20)

What's new?

- イントロダクション | NGS | 各種覚書 (last modified 2010/7/14) NEW
- イントロダクション | NGS | 様々なプラットフォーム (last modified 2010/7/7)
- イントロダクション | NGS | リファレンス配列取得(マップされる側) (last modified 2010/7/9)
- イントロダクション | NGS | アノテーション情報取得(refFlatファイル) (last modified 2010/7/9)
- イントロダクション | 一般 | 配列取得 (last modified 2010/7/7)

塩基配列データ → 遺伝子発現行列

- 遺伝子領域の座標データファイル（例：refFlat形式）

	A	B	C	D	E	F	G	H	I	J	K
1	SNAR-G2	NR_024244	chr19	-	49534925	49535044	49535044	49535044	1	49534925,	49535044,
2	SNAR-D	NR_024243	chr19	-	50643458	50643577	50643577	50643577	1	50643458,	50643577,
3	SNORD113-5	NR_003233	chr14	+	101404523	101404600	101404600	101404600	1	101404523,	101404600,
4	UBL5	NM_024292	chr19	+	9938567	9940797	9939013	9940684	5	9938567,9939002,9939267,9939512,9940640,	9938740,9939069,

A: 遺伝子シンボル
B: 遺伝子名
C: 染色体番号
D: 鎖の向き(+鎖 or –鎖)
E：転写開始位置
F：転写終結位置
G：コーディング領域の開始位置
H：コーディング領域の終結位置
I：エクソンの数
J：エクソンの開始位置
K：エクソンの終結位置

座標データファイルも無料で公開されている

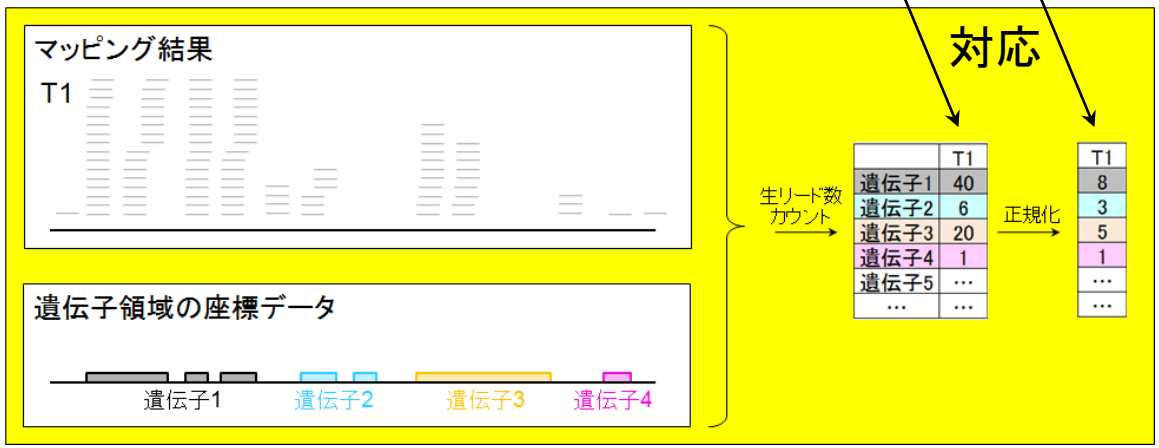
塩基配列データ → 遺伝子発現行列

(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2010/7/20)

What
重複なし) (last modified 2010/6/8)
- 前処理 | ゲノムへのマッピング結果から既知遺伝子の発現レベル(RPKM)への変換 (last modified 2010/9/14) NEW
- 前処理 | サンプル間比較を行うための正規化について(RPM, RPKM, ...) (last modified 2010/7/20)

の結果ファイル

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



このサンプルを次世代シーケンサーにかけると**5087097** reads (重複を含む塩基配列数)からなるデータが得られており、そのうち**744** readsが**A1BG**という遺伝子上にマップされていて、この遺伝子の正規化後の発現レベルは**82.9** RPKMですよ。

データの正規化

- RPM正規化（マイクロアレイなどと同じところ）
  - Reads per million mapped readsの略
  - サンプルごとに読まれた総リード（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

「生read数 : 総read数 = x : 1,000,000 」
A1BGの場合は「744 : 5,087,097 = x : 1,000,000」

$$x = \text{生read数} \times \frac{1000000}{\text{総read数}} = 744 \times \frac{1000000}{5087097} = 146.3$$

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（次世代シーケンサ特有）
  - Reads per kilobase of exon per million mapped readsの略
  - 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）だった場合」に補正

$$\text{生read数} \times \frac{1000000}{\text{総read数}} \times \frac{1000}{\text{配列長}} = 744 \times \frac{1000000}{5087097} \times \frac{1000}{1764} = 82.9$$

遺伝子発現行列 → 様々な解析が可能

- RPKM正規化後の遺伝子発現行列（ファイル名：data.txt）

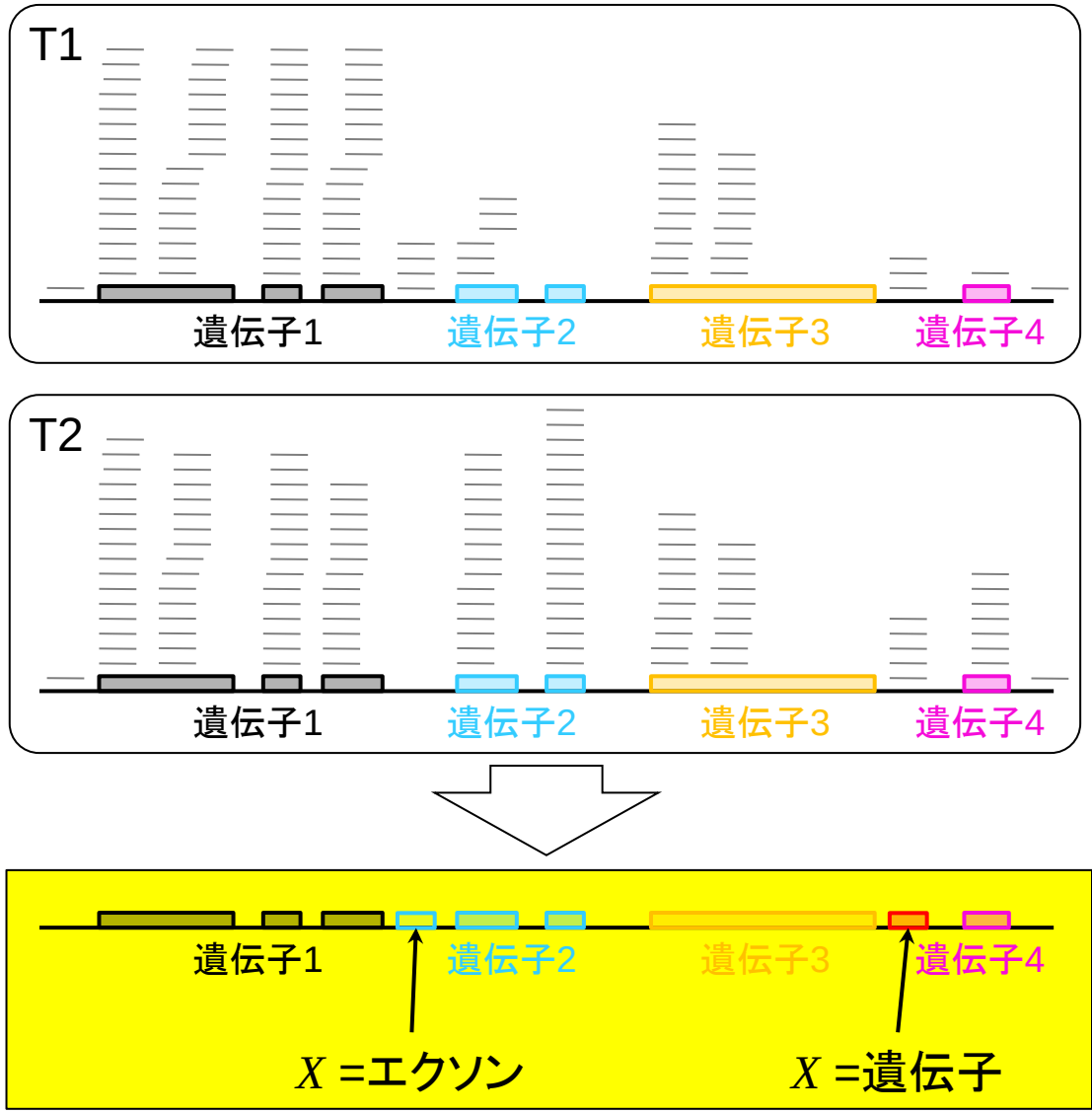
14サンプル
（A: 7サンプル、B：7サンプル）

21,717遺伝子

symbol	A1	A2	A3	A4	A5	A6	A7	B1	B2	B3	B4	B5	B6	B7
A1BG	7.2	7.7	7.6	8.0	7.6	7.1	7.1	3.5	4.4	3.6	3.9	4.2	4.2	4.0
A1CF	3.1	3.4	2.9	3.1	3.8	2.5	3.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0
A2BP1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	26.3	23.7	23.1	23.7	25.4	24.3	23.5
A2LD1	1.6	2.1	2.7	2.5	1.8	2.4	1.5	1.2	1.4	1.4	1.2	0.9	1.9	1.0
A2M	93.8	94.9	91.3	91.5	91.8	94.3	93.7	23.1	23.5	22.6	23.6	22.4	23.3	24.6
A2ML1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.3	0.4	0.5	0.2	0.2	0.4	0.3
A4GALT	7.2	7.6	7.2	5.1	7.1	7.9	6.1	3.6	3.5	4.5	4.0	2.9	3.8	3.6
A4GNT	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0
AAA1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AAAS	33.0	33.5	33.8	34.6	33.0	34.6	33.7	12.4	11.2	11.8	12.0	12.3	13.3	13.4
AACS	12.4	13.1	13.4	12.0	12.9	12.6	12.5	14.9	13.6	11.9	13.4	14.0	14.1	13.2
AACSL	0.3	0.4	0.9	0.8	0.8	0.8	0.6	0.0	0.0	0.1	0.0	0.1	0.2	0.1
AADAC	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AADACL2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AADACL3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AADACL4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AADAT	2.2	1.8	1.9	2.0	1.2	2.4	2.0	1.4	1.3	1.7	0.9	2.2	1.2	1.7
AAGAB	14.0	15.8	13.7	13.9	15.6	15.6	15.3	6.6	6.4	6.3	6.1	7.1	6.7	6.3
...														

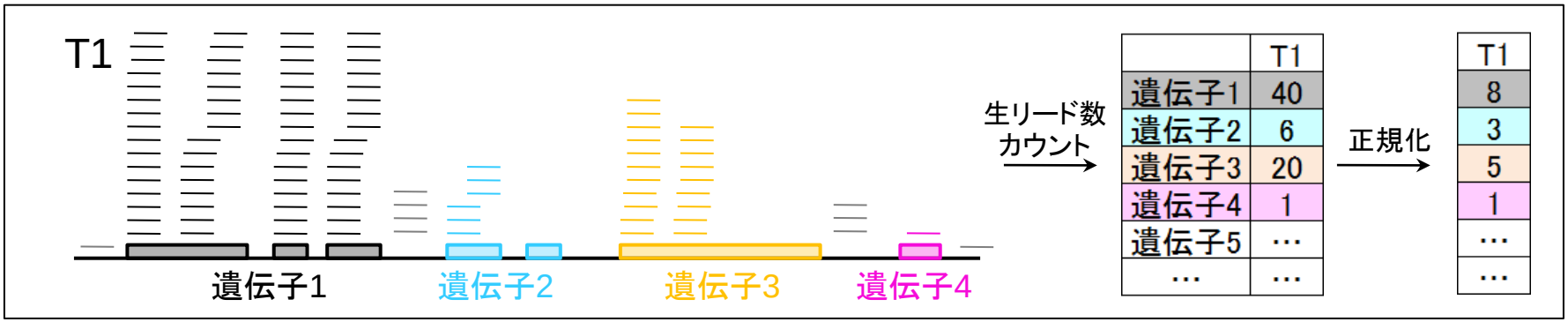
次世代シーケンサーの無限の可能性

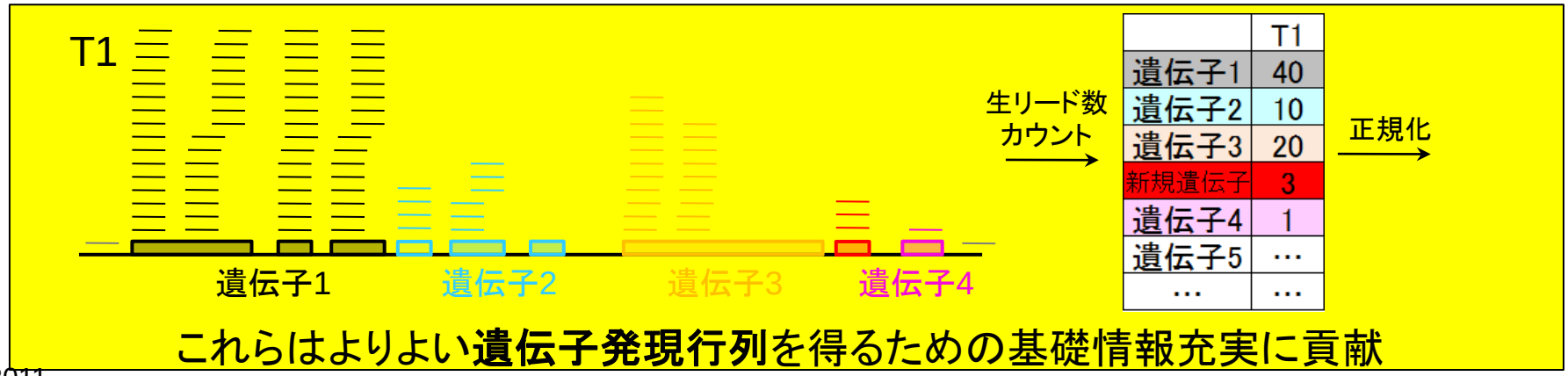
- 新規Xの同定
  - X =エクソン, 遺伝子, …



次世代シーケンサーの無限の可能性

- 「新規ゲノム配列決定」
- 「新規Xの同定」





トランスクリプトームとは

- ある特定の状態の組織や細胞中に存在する全RNA（転写物、transcripts）の総体
- 様々なトランスクリプトーム解析技術
  - マイクロアレイ
    - cDNAマイクロアレイ、Affymetrix GeneChip、タイリングアレイなど
  - 配列決定に基づく方法
    - EST、SAGEなど、次世代シーケンサー (NGS)
  - 電気泳動に基づく方法
    - Differential Display、AFLPなど

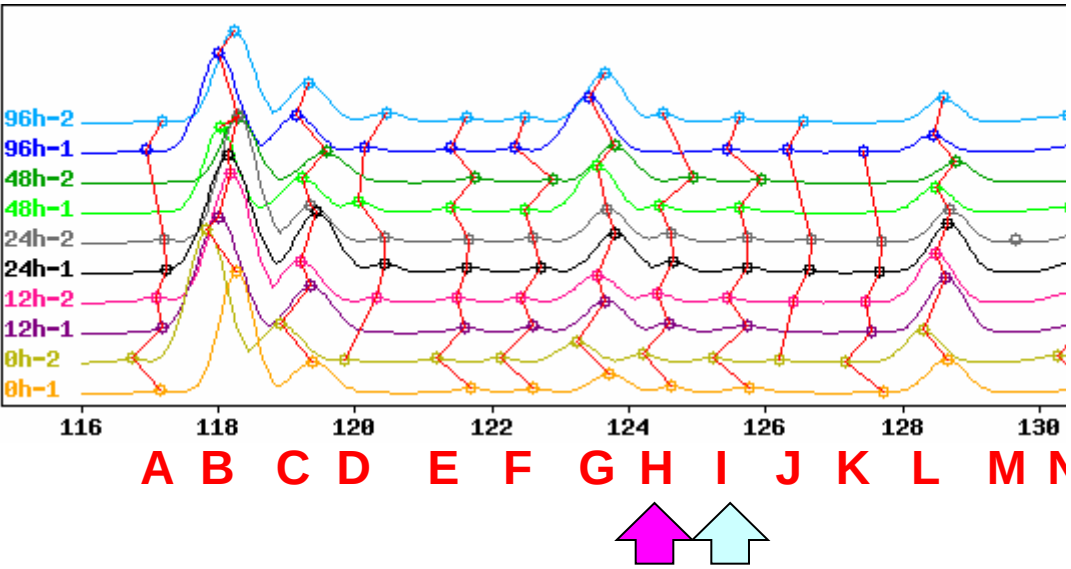
調べたい組織でどの遺伝子がどの程度発現しているのかを一度に観察

電気泳動データ → 遺伝子発現行列

- マイクロアレイ（や塩基配列データ）では**遺伝子発現行列**が出発点
- 電気泳動データは**遺伝子発現行列**の作成が簡単ではない

比較する実験数が増えるほど、同一遺伝子の認識（アラインメント）精度が下がるから

理想的なアラインメント



遺伝子発現行列

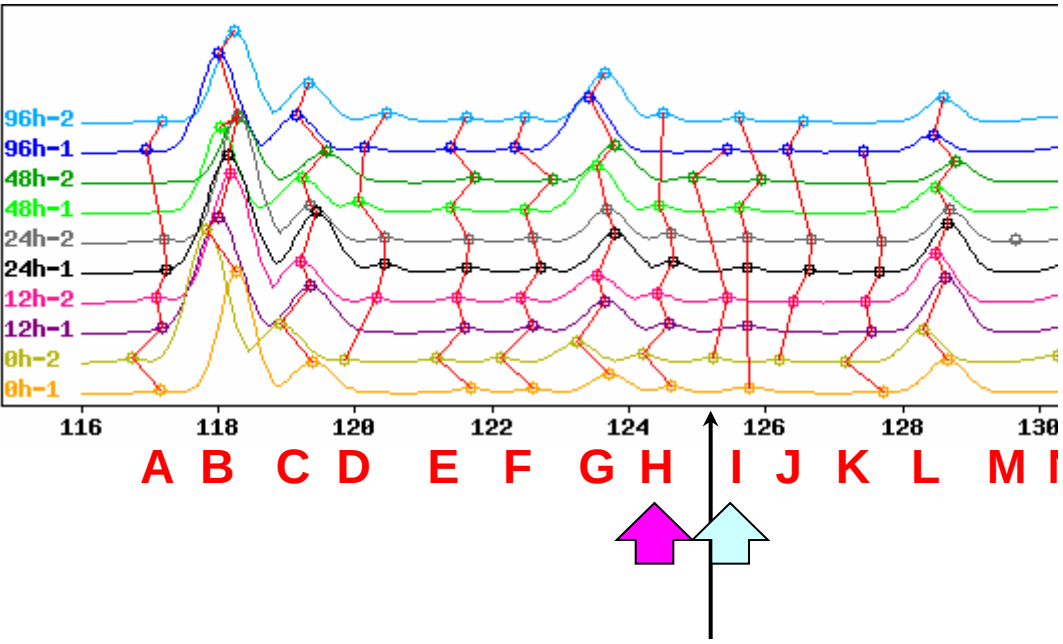
	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	**33**	**41**	**47**	**49**	**55**	**48**	**34**	**27**		**43**
I	**28**	**24**	**34**	**25**	**30**	**27**	**25**	**14**	**18**	**23**
J		16		8	13	17			14	7
K	7	9	8	9	8	9			4	
L	168	163	275	242	246	165	126	102	85	123
M						10				
N	30	34	54	49	68	52	25	11	33	31

電気泳動データ → 遺伝子発現行列

- マイクロアレイ（や塩基配列データ）では**遺伝子発現行列**が出発点
- 電気泳動データは**遺伝子発現行列**の作成が簡単ではない

比較する実験数が増えるほど、同一遺伝子の認識（アラインメント）精度が下がるから

現実…



遺伝子発現行列

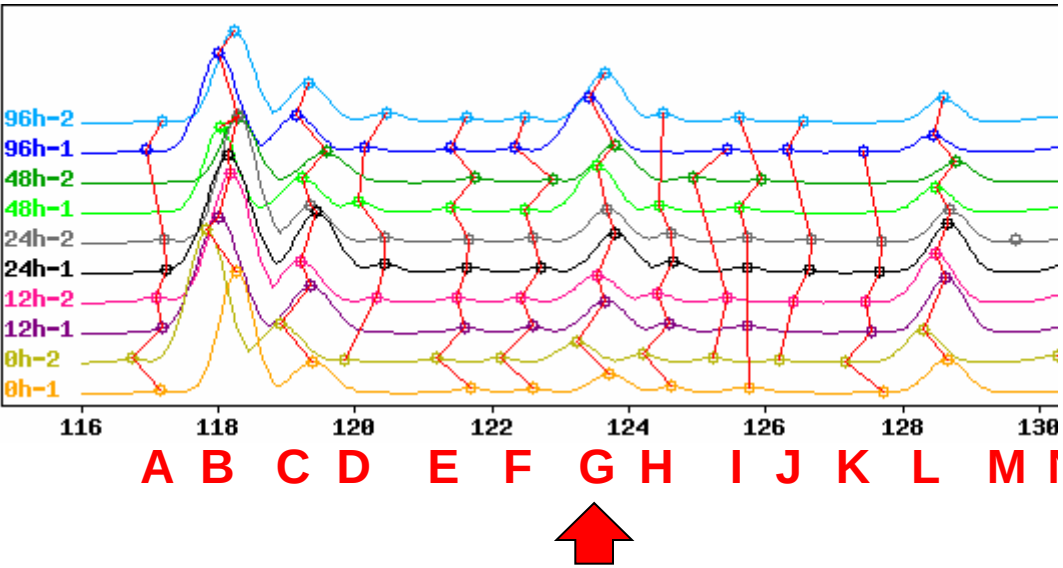
	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	33	41	47	49	55	48	34			43
H'I'		24		25				27	18	
I	28		34		30	27	25	14		23
J		16		8	13	17			14	7
K	7	9	8	9	8	9			4	
L	168	163	275	242	246	165	126	102	85	123
M						10				
N	30	34	54	49	68	52	25	11	33	31

電気泳動データ → 遺伝子発現行列

- マイクロアレイ（や塩基配列データ）では**遺伝子発現行列**が出発点
- 電気泳動データは**遺伝子発現行列**の作成が簡単ではない

比較する実験数が増えるほど、同一遺伝子の認識（アラインメント）精度が下がるから

現実…



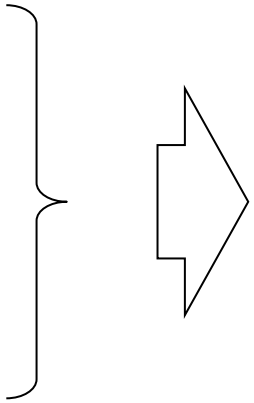
遺伝子発現行列

	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	33	41	47	49	55	48	34			43
H'I'		24		25				27	18	
I	28		34		30	27	25	14		23
J		16		8	13	17			14	7
K	7	9	8	9	8	9			4	
L	168	163	275	242	246	165	126	102	85	123
M						10				
N	30	34	54	49	68	52	25	11	33	31

Gの発現パターンは本当に全部G由来？！

ここまでのまとめ

- 様々なトランスクリプトーム解析技術を紹介
  - マイクロアレイ
  - 配列決定に基づく方法
    - 次世代シーケンサー (NGS)
  - 電気泳動に基づく方法



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
…	…	…

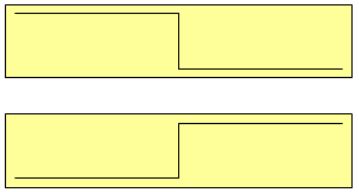
どの実験技術由来データも「遺伝子発現行列」の形式に変換可能

バイオインフォマティクス要素技術

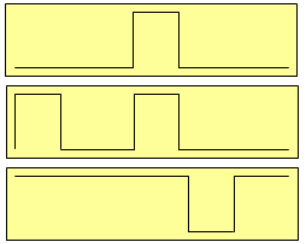
- 「相関係数」や「エントロピー」などの応用例を紹介

↓

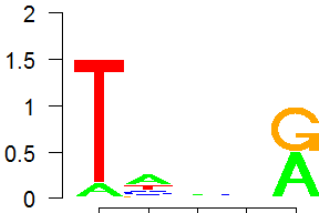
二群間比較



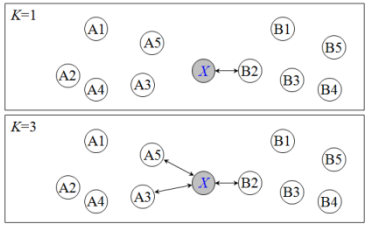
組織特異的遺伝子



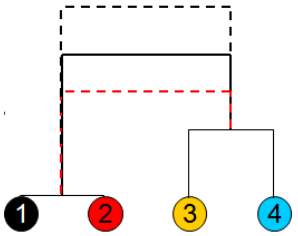
Sequence logo



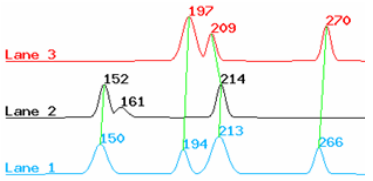
分類（診断）



クラスタリング



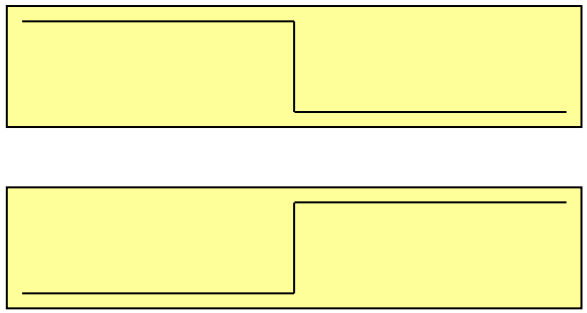
同一ピーク同定



様々な遺伝子発現行列

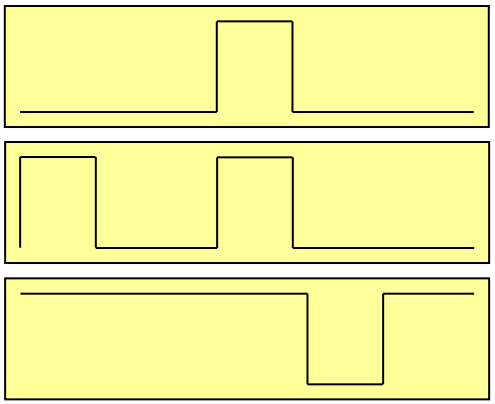
1. 二群間比較

	A群 A1	A2	…	B群 B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	



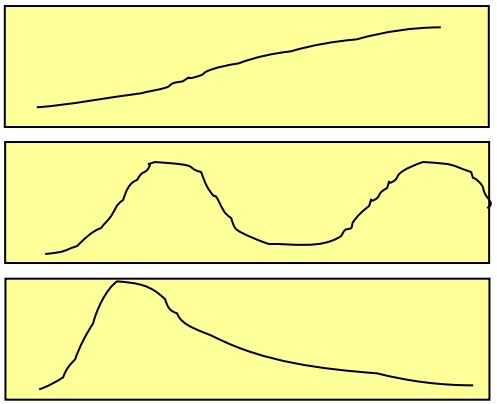
2. 様々な組織（条件）

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



3. 時系列データ

	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



理想的なパターンと似たパターンを示す遺伝子を検出

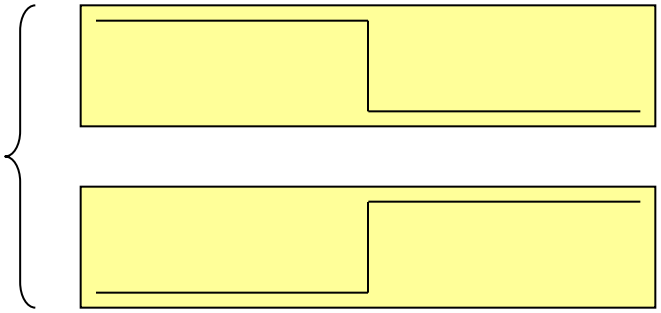
解析例(二群間比較)

- 二群間比較
  - A:癌サンプル
  - B:正常組織サンプル

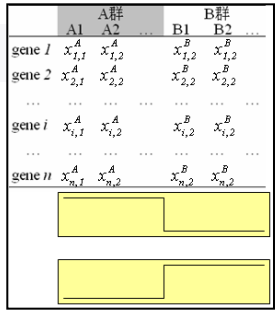
→腫瘍マーカー候補の探索

	A群			B群		
	A1	A2	…	B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	

癌と正常で発現の異なる遺伝子（発現変動遺伝子）を同定



解析例（二群間比較）



- パターンマッチング法
  - 理想的なパターンyとの類似度が高い順にランキング

$$\text{相関係数 } r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (-1 \le r \le 1)$$

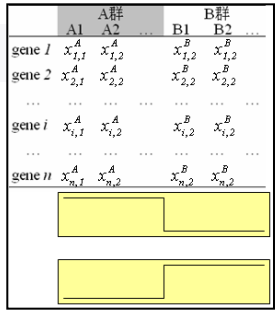
	A群						B群				
	A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	B5
y	1	1	1	1	1	1	0	0	0	0	0
x_{gene1}	87	79	91	82	90	84	12	21	19	13	17
x_{gene2}	56	122	106	47	84	98	7	44	2	11	18
x_{gene3}	15	28	33	9	27	41	48	46	52	50	49

$$r_{gene1} = \frac{18.85}{36.32 \times 0.52} = 0.994$$

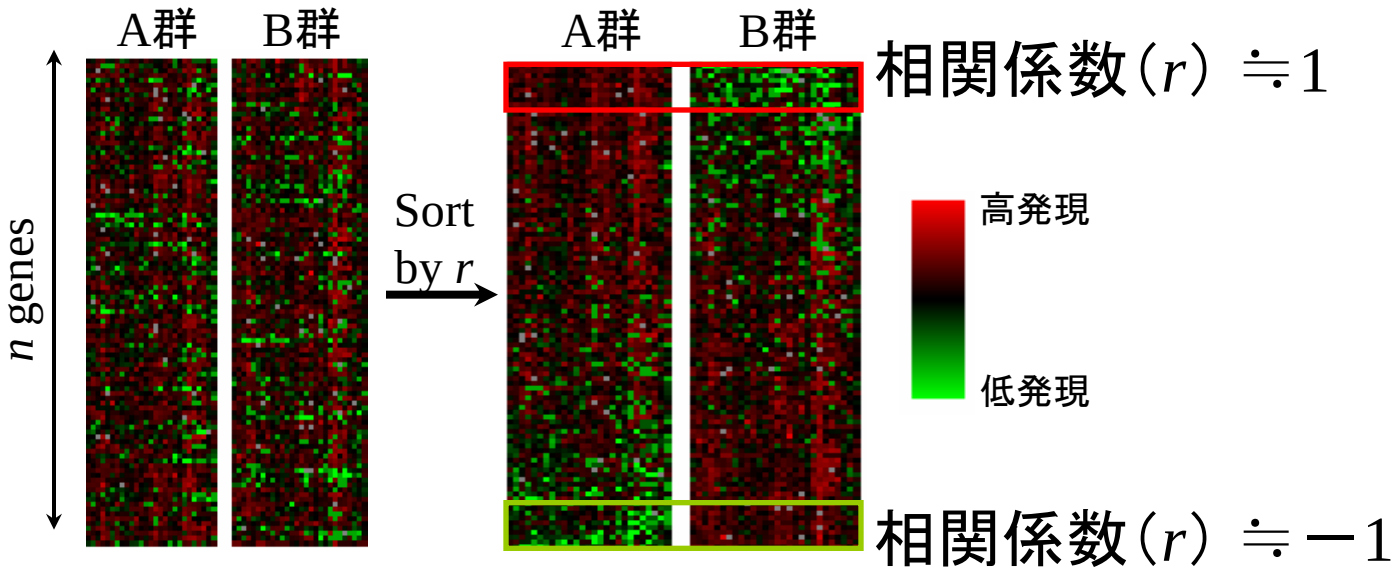
$$r_{gene2} = \frac{18.85}{42.87 \times 0.52} = 0.842$$

$$r_{gene3} = \frac{-6.41}{14.88 \times 0.52} = -0.825$$

解析例（二群間比較）



- パターンマッチング法
  - 理想的なパターンyとの類似度が高い順にランキング



解析例(二群間比較)

- Golub *et al.*, *Science*, 1999.
  - A:ALL(27サンプル)
  - B:AML(11サンプル)

急性リンパ性白血病　　急性骨髄性白血病

発現の異なる遺伝子群を同定するとともに、分類(診断)に適用

実習(二群間比較)

(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2010/09/01)

What's new?
- GSA (Efron_2007)の中身をちゃんと埋め始めましたが、まだ最後までは辿りつけてません(2010/8/30)

- 解析 | 発現変動遺伝子 | 二群間 | 対応なし | PPLR (Liu_2006) (last modified 2009/7/25)
- 解析 | 発現変動遺伝子 | 二群間 | 対応なし | Rank products (Breitling_2004) (last modified 2009/11/2)
- 解析 | 発現変動遺伝子 | 二群間 | 対応なし | Empirical bayes statistic (Smyth_2004) (last modified 2009/7/25)
- 解析 | 発現変動遺伝子 | 二群間 | 対応なし | samroc (Broberg_2003) (last modified 2009/7/25)
- 解析 | 発現変動遺伝子 | 二群間 | 対応なし | SAM (Tusher_2001) (last modified 2009/7/25)
- 解析 | 発現変動遺伝子 | 二群間 | 対応なし | Student's t-test (last modified 2009/7/28)
- 解析 | 発現変動遺伝子 | 二群間 | 対応なし | Welch t-test (last modified 2009/7/28)
- 解析 | 発現変動遺伝子 | 二群間 | 対応なし | Mann-Whitney U-test (last modified 2009/7/28)
- 解析 | 発現変動遺伝子 | 二群間 | 対応なし | パターンマッチング法 (last modified 2011/10/13) NEW

• 解析 | 発現変動遺伝子 | 二群間 | 対応なし | パターンマッチング法

パターンマッチング法を用いて、二群間での発現変動遺伝子の同定を行うやり方を紹介します。

「ファイル」-「ディレクトリの変更」で解析したいサンプルマイクロアレイデータ15中のsample16.txtファイル(遺伝子発現データ)とsample16_cl.txtファイル(クラスラベルデータ)を置いてあるディレクトリに移動し、以下をコピペ

1. クラスラベル情報ファイル (sample16_cl.txt) を読み込んでテンプレートパターン情報を得る場合：

```
------    ここから    ------
in_f1 <- "sample16.txt"                                          #入力ファイル名1(発現データ)を指定
in_f2 <- "sample16_cl.txt"                                       #入力ファイル名2 (テンプレート情報) を指定
out_f <- "hoge.txt"                                              #出力ファイル名を指定

data <- read.table(in_f1, header=TRUE, row.names=1, sep="¥t", quote="")#入力ファイル1を読み込んでdataに格納
hoge <- read.table(in_f2, sep="¥t", quote="")                    #入力ファイル2を読み込んでhogeに格納
data.cl <- hoge[,2]                                              #テンプレートパターンベクトルdata.clを作成
r <- apply(data, 1, cor, y=data.cl)                              #各 (行) 遺伝子についてテンプレートパターンdata.clとの相関係数を計算
tmp <- cbind(rownames(data), data, r)                            #入力データの右側に相関係数rのベクトルを結合した結果をtmpに格納。
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定したファイル名で保存。

------    ここまで    ------
```

実習(二群間比較)

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ

R Console

```
>
>
>
> in_f1 <- "sample16.txt"                                          #入力ファ$
> in_f2 <- "sample16_cl.txt"                                       #入力ファ$
> out_f <- "hoge.txt"                                              #出力ファ$
>
> data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t", quote="")#入$
> hoge <- read.table(in_f2, sep="\t", quote="")                    #入力ファ$
> data.cl <- hoge[,2]
> r <- apply(data, 1, cor, y=c
> tmp <- cbind(rownames(data),
> write.table(tmp, out_f, sep=
>
>
```

hoge.txt - Microsoft Excel

ファイル ホーム 挿入 ページ レイアウト 数式 データ 校閲 表示 活用しよう！エクセル

Q10

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	rownan	A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	B5	r
2	gene1	87	79	91	82	90	84	12	21	19	13	17	0.994
3	gene2	56	122	106	47	84	98	7	44	2	11	18	0.842
4	gene3	15	28	33	9	27	41	48	46	52	50	49	−0.825

hoge

コマンド 100%

様々な遺伝子発現行列

1. 二群間比較

	A群 A1	A2	…	B群 B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	

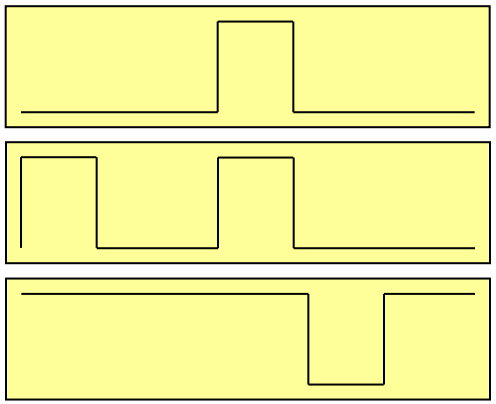
2. 様々な組織（条件）

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

脳特異的高発現

心臓と脳特異的高発現

肺特異的低発現



3. 時系列データ

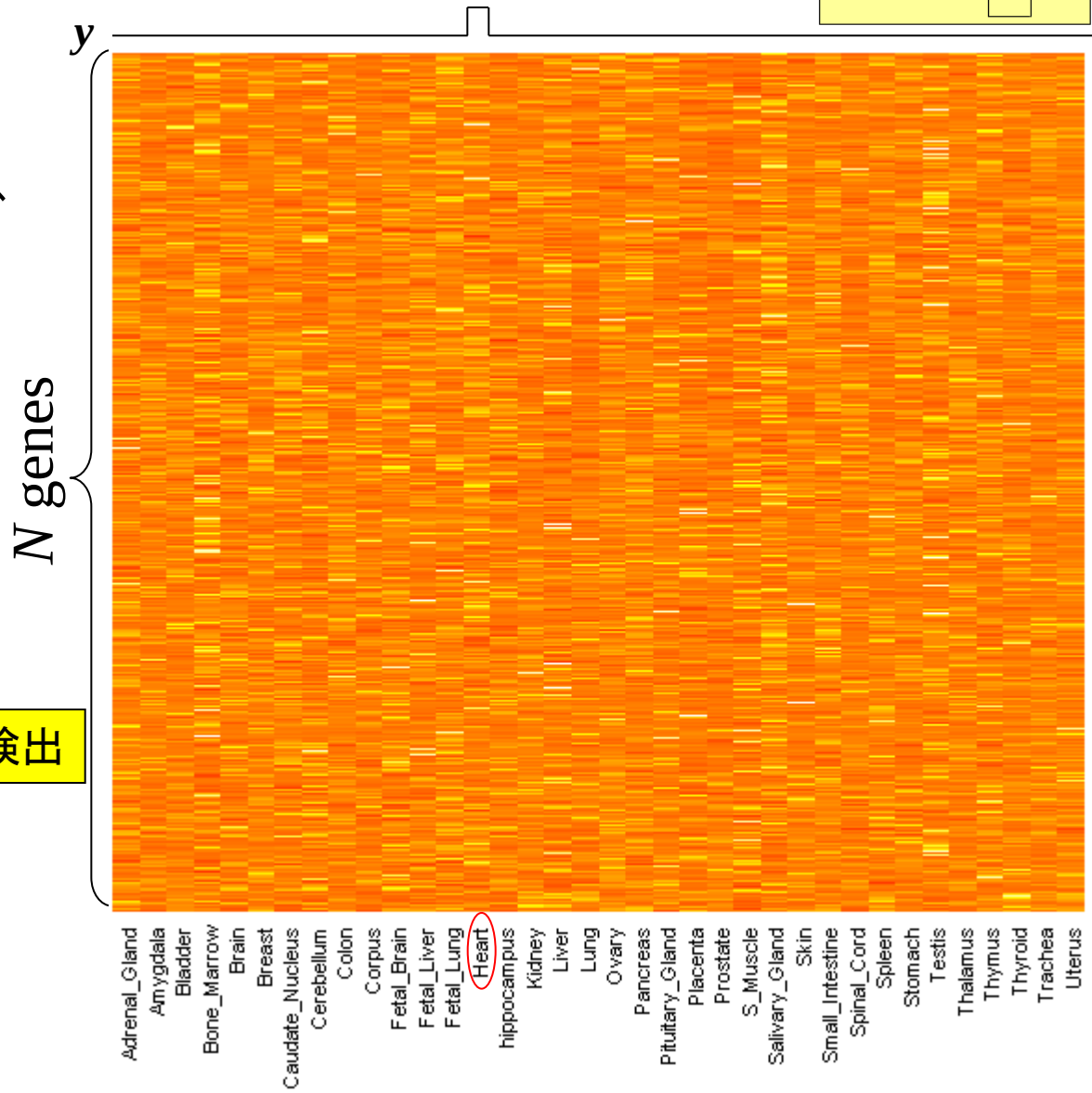
	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

理想的なパターンと似たパターンを示す遺伝子を検出

解析例(多サンプル間比較)

- パターンマッチング法
  - 理想的なパターン*y*との類似度が高い順にランキング

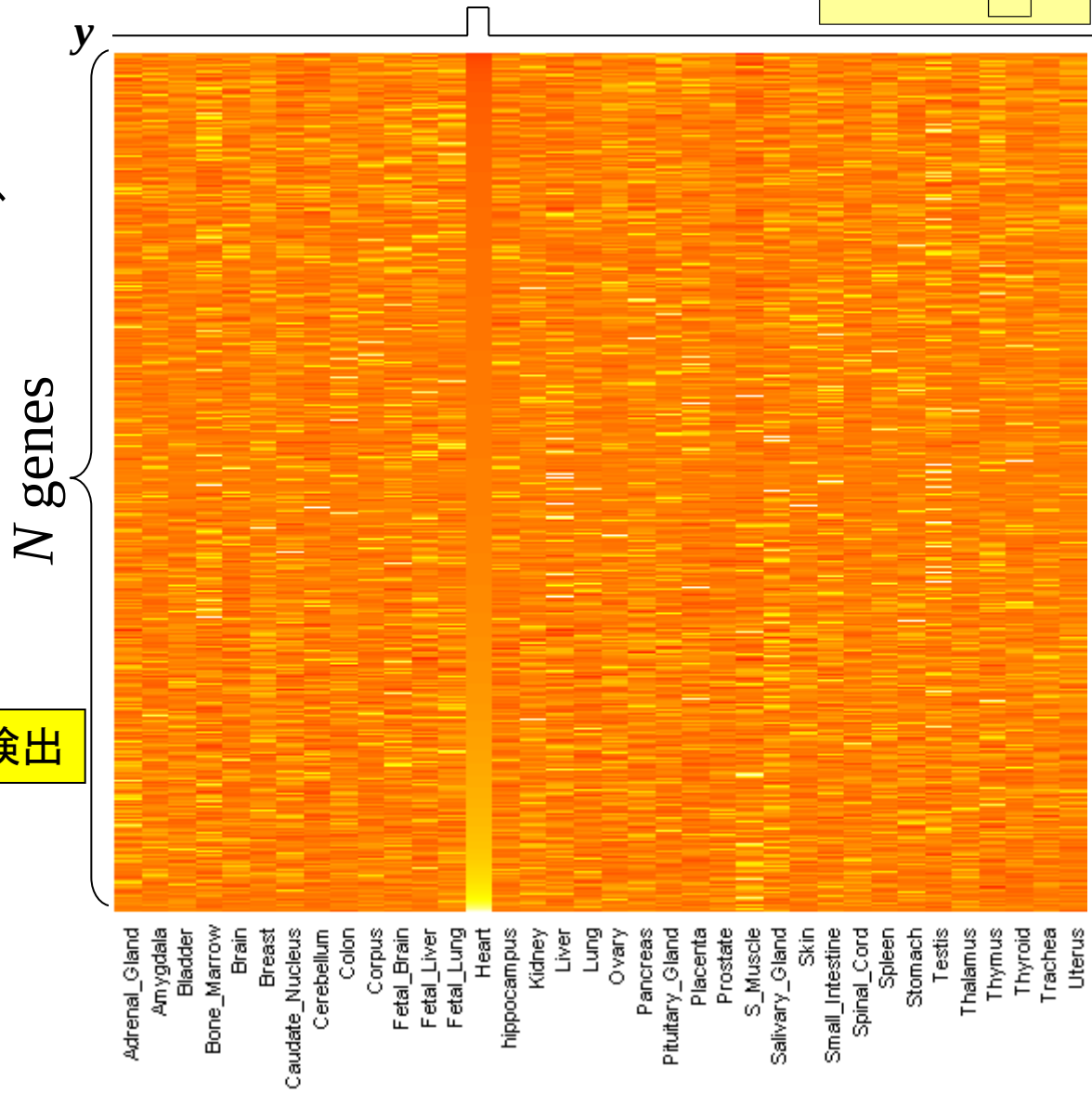
例:心臓特異的パターンを示す遺伝子群の検出



解析例(多サンプル間比較)

- パターンマッチング法
  - 理想的なパターンyとの類似度が高い順にランキング

例:心臓特異的パターンを示す遺伝子群の検出



実習(組織特異的遺伝子検出)

(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2011/10/13)

What's new?

- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜について(last modified 2011/6/6) NEW
- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜ROKU(Kadota_2006) (last modified 2009/07/30)
- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜Sprent's non-parametric method (Ge_2005) (last modified 2009/07/31)
- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜Schug's H(x) statistic (Schug_2005) (last modified 2011/10/13) NEW
- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜Schug's Q statistic (Schug_2005) (last modified 2009/07/31)
- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜Ueda's AIC-based method (Kadota_2003) (last modified 2009/07/31)
- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜パターンマッチング法(テンプレートマッチング法) (last modified 2011/10/13) NEW
- 解析｜発現変動遺伝子｜時系列データ｜Periodic genes｜Lomb-Scargle periodogram (Glynn_2006) (last modified 2006/7/11)

解析｜発現変動遺伝子｜組織特異的(選択的)発現遺伝子｜パターンマッチング法(テンプレートマッチング法)

(基本的には、解析｜似た発現パターンを持つ遺伝子の同定をご覧ください。)
パターンマッチング法を用いて、指定した理想的なパターンとの類似度が高い遺伝子の同定を行うやり方を紹介します。

「ファイル」－「ディレクトリの変更」で解析したいサンプルマイクロアレイデータ14中のsample15.txtファイル(遺伝子発現データ)とsample15_cl.txtファイル(sample4で特異的高発現パターンを検出するためのテンプレートパターンのデータ)を置いてあるディレクトリに移動し、以下をコピペ

```
------    ここから    ------
in_f1 <- "sample15.txt"                                          #入力ファイル名1(発現データ)を指定
in_f2 <- "sample15_cl.txt"                                       #入力ファイル名2（テンプレート情報）を指定
out_f <- "hoge.txt"                                              #出力ファイル名を指定

data <- read.table(in_f1, header=TRUE, row.names=1, sep="¥t", quote="")#入力ファイル1を読み込んでdataに格納
hoge <- read.table(in_f2, sep="¥t", quote="")                    #入力ファイル2を読み込んでhogeに格納
data.cl <- hoge[,2]                                              #テンプレートパターンベクトルdata.clを作成
r <- apply(data, 1, cor, y=data.cl)                              #各（行）遺伝子についてテンプレートパターンdata.clとの相関係数を計算した
tmp <- cbind(rownames(data), data, r)                            #入力データの右側に相関係数rのベクトルを結合した結果をtmpに格納。
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定したファイル名で保存。

------    ここまで    ------
```

実習（組織特異的遺伝子検出）

- 入力データ1（遺伝子発現データファイル：sample15.txt）

	A	B	C	D	E	F	G	H	I
1	id	tissue1	tissue2	tissue3	tissue4	tissue5	tissue6	tissue7	tissue8
2	gene1	1	0	0	9	0	0	0	0
3	gene2	5	2	1	2	4	6	3	5
4	gene3	6	6	6	6	6	6	6	6
5	gene4	4	4	4	4	10	4	4	4
6	gene5	10	10	10	10	4	10	10	10

- 入力データ2（テンプレートパターンファイル：sample15_cl.txt）

	A	B
1	tissue1	0
2	tissue2	0
3	tissue3	0
4	tissue4	1
5	tissue5	0
6	tissue6	0
7	tissue7	0
8	tissue8	0

実習(組織特異的遺伝子検出)

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ

R Console

```
>
> in_f1 <- "sample15.txt"                    #入力ﾌｧｲ
> in_f2 <- "sample15_cl.txt"                 #入力ﾌｧｲ
> out_f <- "hoge.txt"                        #出力ﾌｧｲ
>
> data <- read.table
> hoge <- read.table
> data.cl <- hoge[,2]
> r <- apply(data, 1,
 警告ﾒｯｾｰｼﾞ:
In FUN(newX[, i], ...
> tmp <- cbind(rownar
> write.table(tmp, ou
>
>
```

hoge.txt - Microsoft Excel

ファイル ホーム 挿入 ページ レイアウト 数式 データ 校閲 表示 活用しよう！エクセル

N10

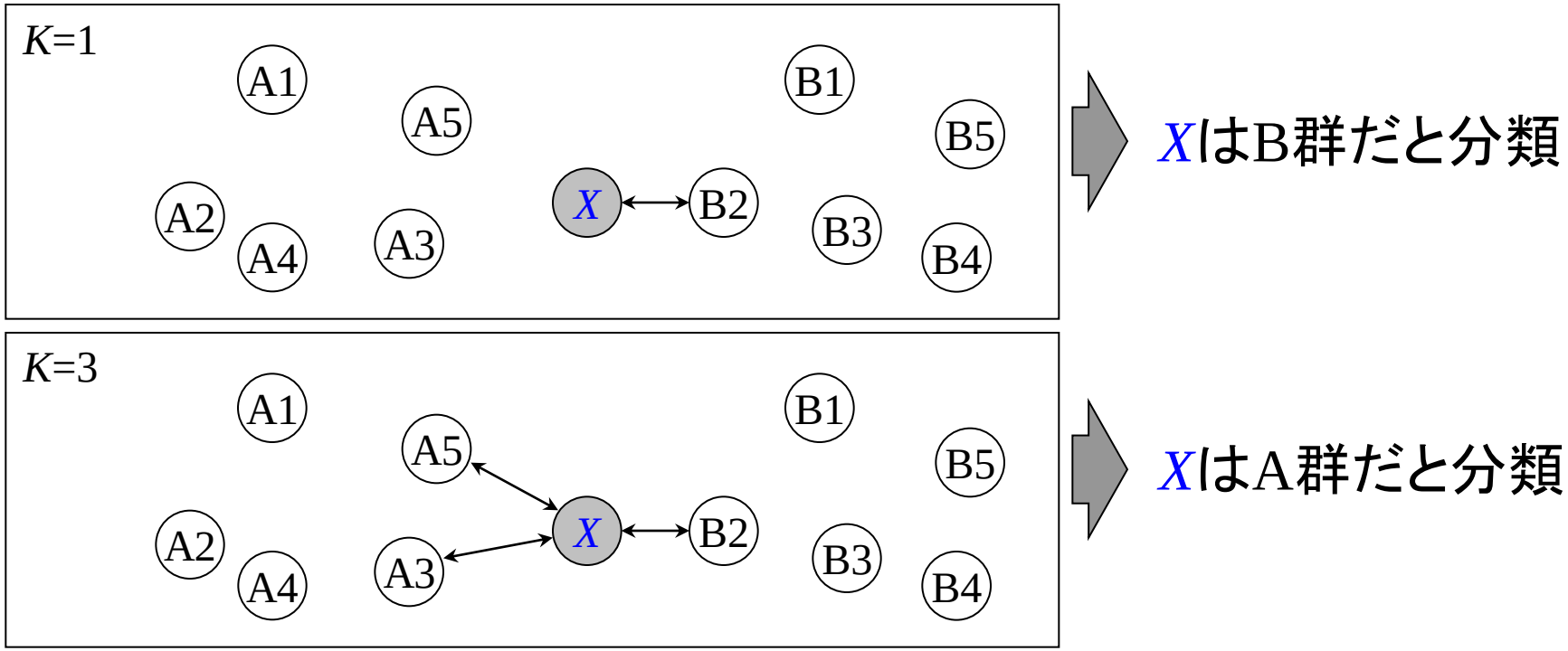
	A	B	C	D	E	F	G	H	I	J
1	rownan	tissue1	tissue2	tissue3	tissue4	tissue5	tissue6	tissue7	tissue8	r
2	gene1	1	0	0	9	0	0	0	0	0.994
3	gene2	5	2	1	2	4	6	3	5	−0.342
4	gene3	6	6	6	6	6	6	6	6	NA
5	gene4	4	4	4	4	10	4	4	4	−0.143
6	gene5	10	10	10	10	4	10	10	10	0.143

hoge

コマンド 100%

解析例(分類)

- *K*-Nearest Neighbor (*K*-最近傍法)
  - 目的:未知サンプル*X*をAまたはBに分類
    - 未知サンプル*X*からの**距離**がもっとも近い*K*個のサンプルのうち、所属するクラスが最も多いクラスに分類



距離(非類似度)の定義

- 目的:xとyの発現パターンの距離Dを定義したい
  - 似ていればDが0になるようにしたい

相関係数 $\boldsymbol{r_{xy}} = \dfrac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$ $\quad (-1 \le r_{xy} \le 1)$

	X	B2
i	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
5	x_5	y_5
…	…	…
n	x_n	y_n

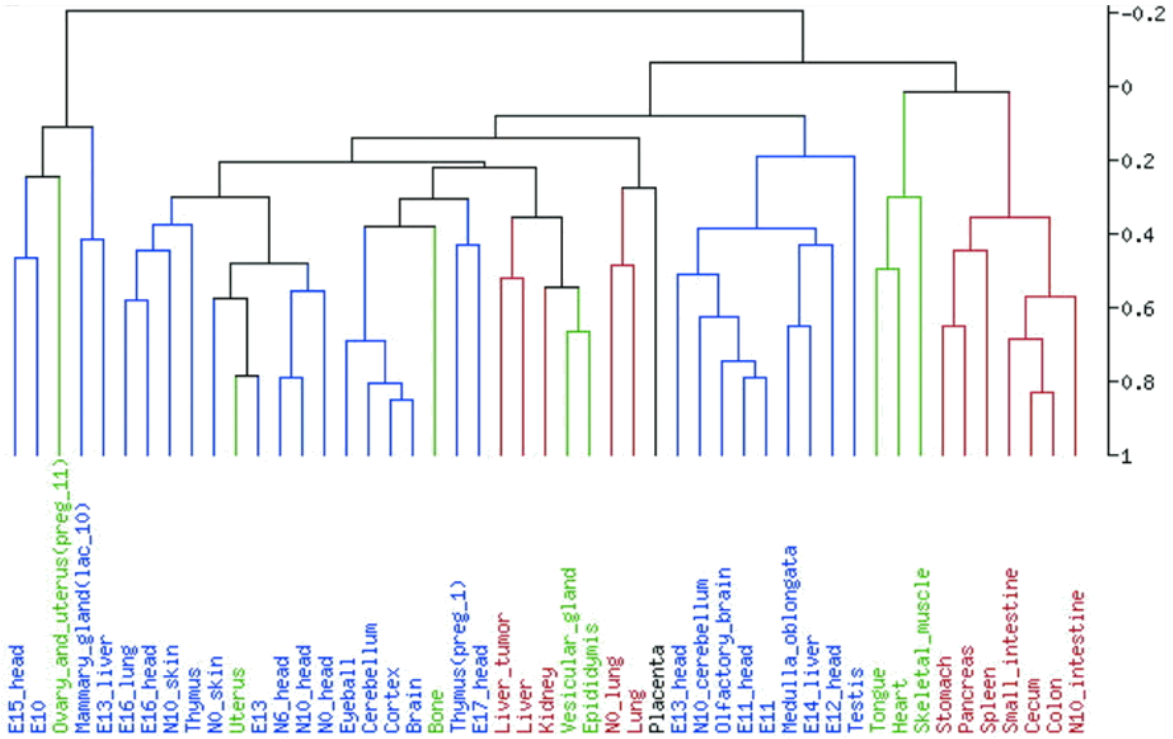
- xとyの発現パターンが酷似 → $r \approx 1$
- xとyの発現パターンがばらばら → $r \approx 0$
- xとyの発現パターンがほぼ正反対 → $r \approx -1$

距離 $\boldsymbol{D} = 1 - r \quad (0 \le D \le 2)$

- $\boldsymbol{r} = 1 \rightarrow \boldsymbol{D} = 1 - 1 = 0$
- $\boldsymbol{r} = 0 \rightarrow \boldsymbol{D} = 1 - 0 = 1$
- $\boldsymbol{r} = -1 \rightarrow \boldsymbol{D} = 1 - (-1) = 2$

解析例（クラスタリング）

- **階層的クラスタリング**
  - 発現パターンの類似した遺伝子（サンプル）を集めて系統樹を作成



解析例（クラスタリング）

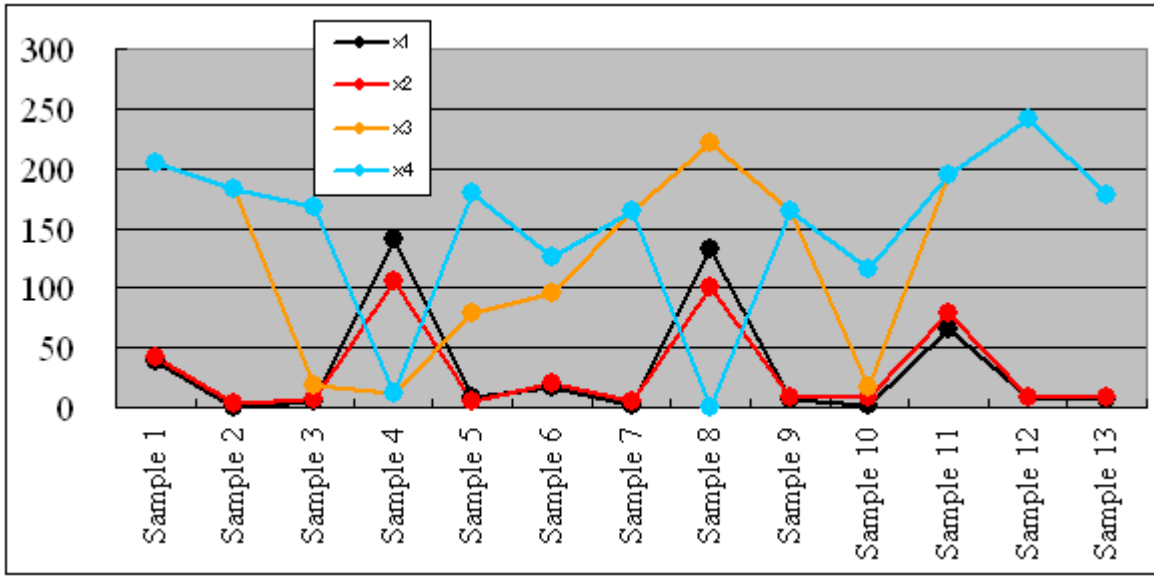
- サンプル間クラスタリング
  - Bittner *et al.*, *Nature*, 2000

悪性度の高い癌の**サブタイプ**を発見

解析例(クラスタリング)

- 階層的クラスタリング

①遺伝子間距離を計算



例:4遺伝子の場合

	x^1	x^2	x^3	x^4
Sample 1	38	42	204	204
Sample 2	0	3	182	182
Sample 3	6	6	19	168
Sample 4	141	106	11	11
Sample 5	8	5	79	179
Sample 6	16	20	96	126
Sample 7	2	5	164	164
Sample 8	132	101	222	0
Sample 9	7	9	164	164
Sample 10	2	8	16	116
Sample 11	66	79	195	195
Sample 12	8	9	241	241
Sample 13	6	8	177	177

距離 $D = 1 - r \quad (0 \le D \le 2)$ ⇒ 距離 $D = \frac{1-r}{2} \quad (0 \le D \le 1)$

相関係数 $r_{1,2} = 0.98 \rightarrow$ 距離 $D_{1,2} = \frac{1-0.98}{2} = 0.01$

相関係数 $r_{1,3} = -0.01 \rightarrow$ 距離 $D_{1,3} = \frac{1-(-0.01)}{2} = 0.50$

相関係数 $r_{1,4} = -0.78 \rightarrow$ 距離 $D_{1,4} = \frac{1-(-0.78)}{2} = 0.89$

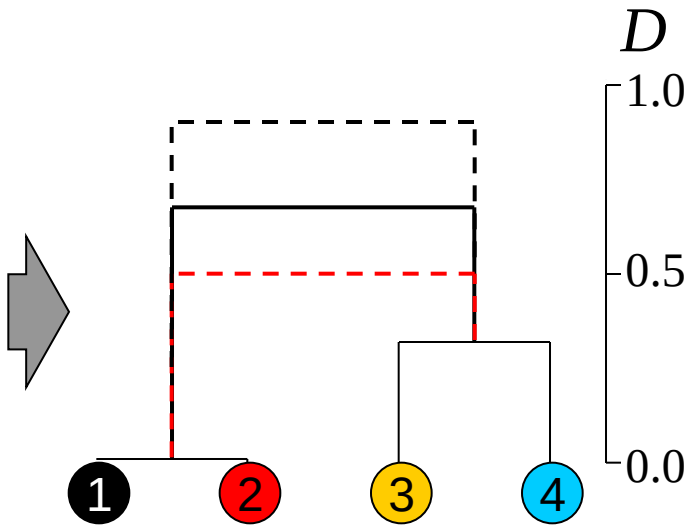
…

解析例（クラスタリング）

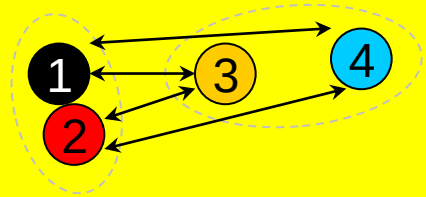
- 階層的クラスタリング

  ②樹形図を作成

	x^2	x^3	x^4
x^1	0.01	0.50	0.89
x^2		0.47	0.84
x^3			0.32



平均連結法の場合



$$(D_{1,3}+D_{1,4}+D_{2,3}+D_{2,4})/4$$
$$=(0.50+0.89+0.47+0.84)/4$$
$$=0.68$$

単連結法の場合

$$\min(D_{1,3},D_{1,4},D_{2,3},D_{2,4})$$
$$=0.47$$

完全連結法の場合

$$\max(D_{1,3},D_{1,4},D_{2,3},D_{2,4})$$
$$=0.89$$

実習（サンプル間クラスタリング）

(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2010/09/01)

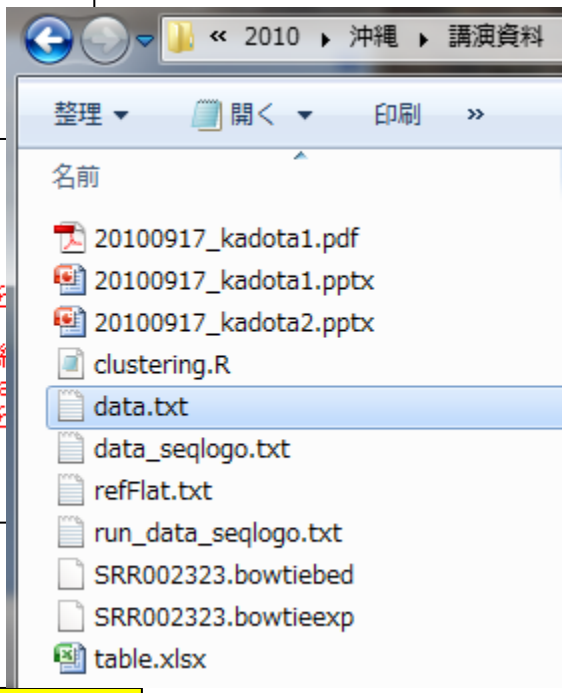
What's new?
- GSA (Efron_2007)の中身をちゃんと埋め始めましたが、まだ最後までは辿りつけてません(2010/8/30)
- Hook (Binder_2008)を追加しました(2010/8/10)

- 解析 | クラスタリング | 階層的 | について(last modified 2009/8/12)
- 解析 | クラスタリング | 階層的 | pvclust (Suzuki_2006) (last modified 2010/8/5) NEW
- 解析 | クラスタリング | 階層的 | hclust (last modified 2010/1/29)
- 解析 | クラスタリング | 階層的 | hclust後の詳細な解析 (last modified 2009/8/7)
- 解析 | クラスタリング | 階層的 | 最適なクラスター数を見積る (last modified 2009/9/9)
- 解析 | クラスタリング | 非階層的 | K-means
- 解析 | クラスタリング | 非階層的 | 自己組織化マップ(SOM)

2. サンプル間クラスタリングの場合(類似度:「1-相関係数」、方法:平均連結法(average)):

・R Graphics画面上に表示したい場合：

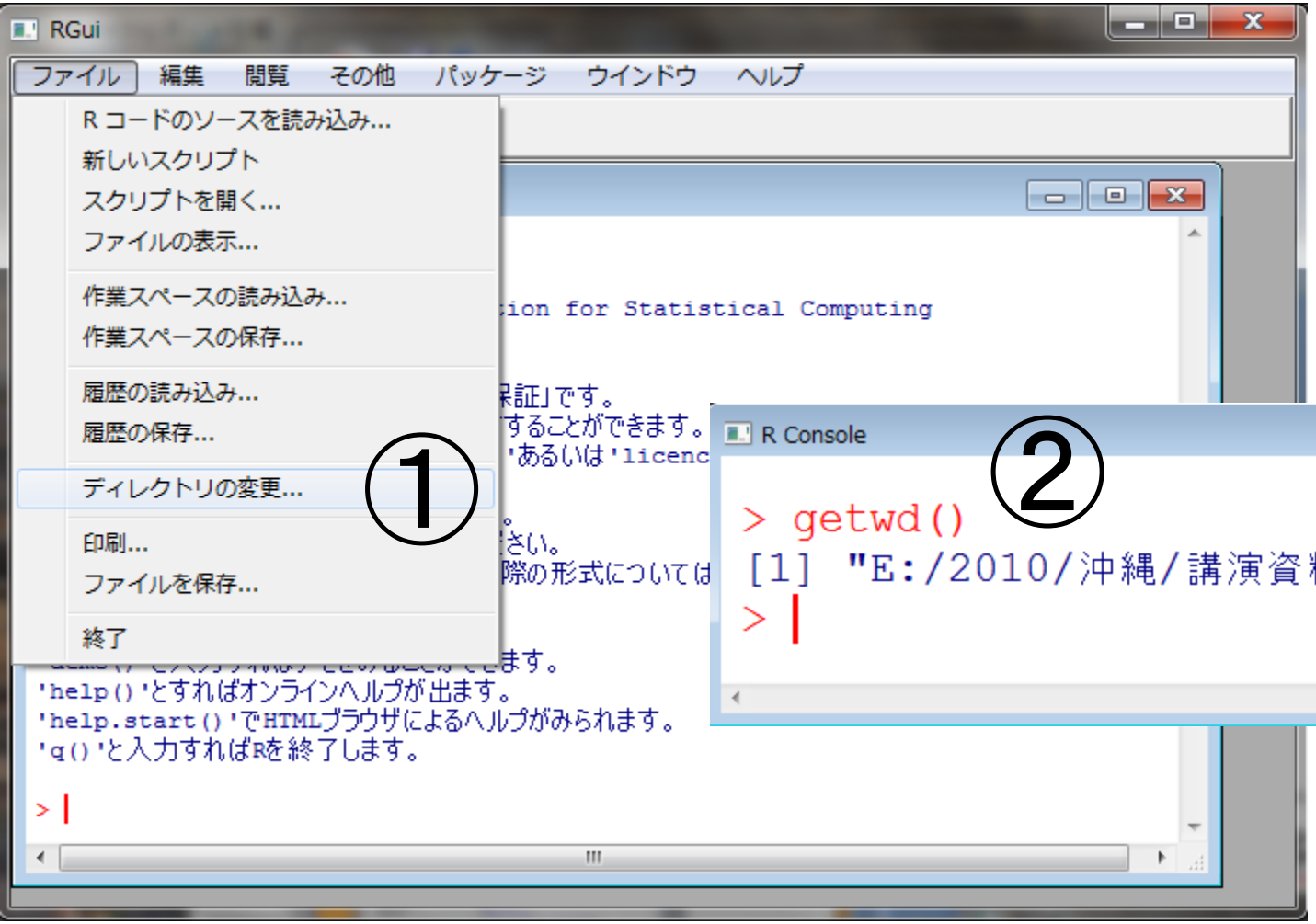
```
------    ここから    ------
in_f <- "sample3.txt"                                          #入力ファイル名(発現データファイル)を
param2 <- "average"                                            #方法(method)を指定
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#発現データを読み込んでdataに格
data.dist <- as.dist(1 - cor(data))                            #サンプル間の距離を計算し、結果をdata
out <- hclust(data.dist, method=param2)                        #階層的クラスタリングを実行し、結果を
plot(out)                                                      #樹形図（デンドログラム）の表示

------    ここまで    ------
```



解析したいのは「… – 2010 – 沖縄 – 講演資料」フォルダ中の「data.txt」ファイル

実習（サンプル間クラスタリング）

①Rを起動し、「ファイル」ー「ディレクトリの変更」で解析したいファイル(data.txt)を置いてあるディレクトリに移動。②念のため確認



実習（サンプル間クラスタリング）

③入力ファイル名の部分を変更したものを用意し、④R Console上でコピペ

③

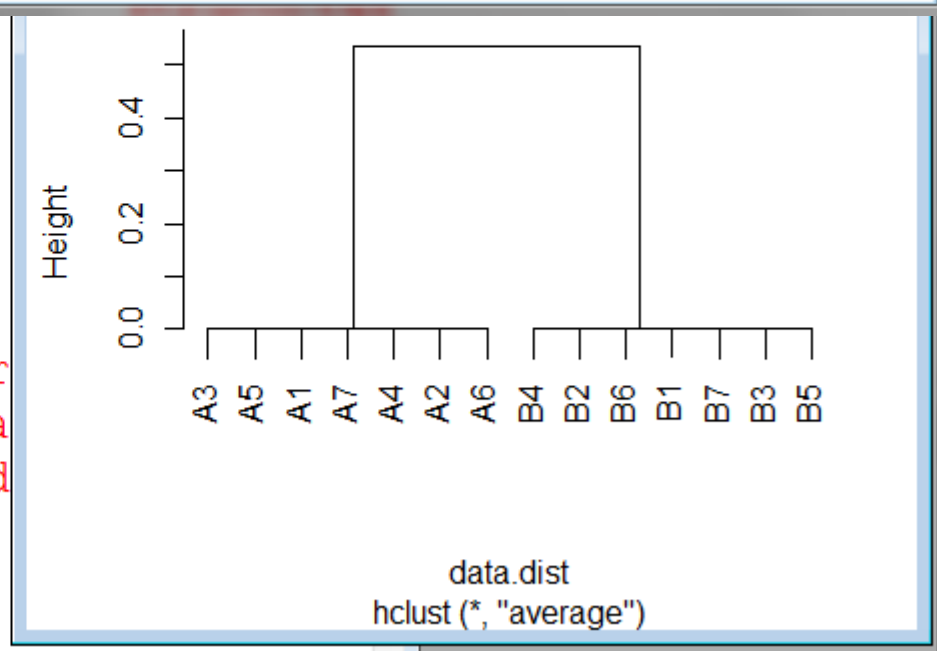
clustering.R - メモ帳

ファイル(F)　編集(E)　書式(O)　表示(V)　ヘルプ(H)

```
in_f <- "data.txt"                                          #入力ファイル名(発現データファイル)を指定
param2 <- "average"                                         #方法(method)を指定
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#発現データを読み込んでdataに格納。
data.dist <- as.dist(1 - cor(data))                         #サンプル間の距離を計算し、結果をdata.distに格納
out <- hclust(data.dist, method=param2)                     #階層的クラスタリングを実行し、結果をoutに格納
plot(out)                                                   #樹形図（デンドログラム）の表示
```

④

```
'q()'と入力すればRを終了します。

> getwd()
[1] "E:/2010/沖縄/講演資料"
> in_f <- "data.txt"
> param2 <- "average"
> data <- read.table(in_f, header
> data.dist <- as.dist(1 - cor(da
> out <- hclust(data.dist, method
> plot(out)
>
```

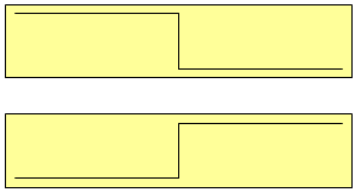


バイオインフォマティクス要素技術

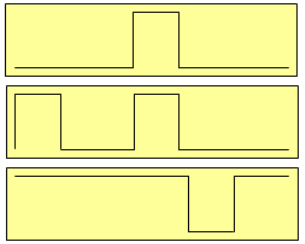
- 「相関係数」や「エントロピー」などの応用例を紹介

↓

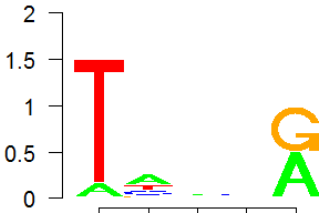
二群間比較



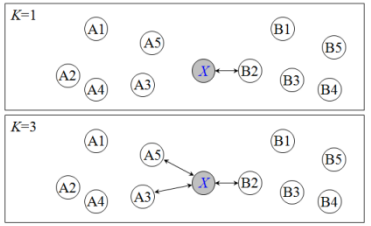
組織特異的遺伝子



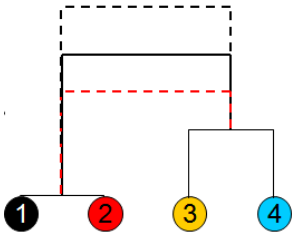
Sequence logo



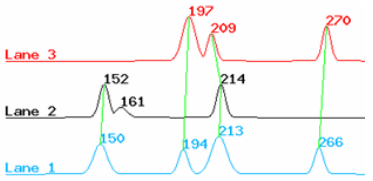
分類（診断）



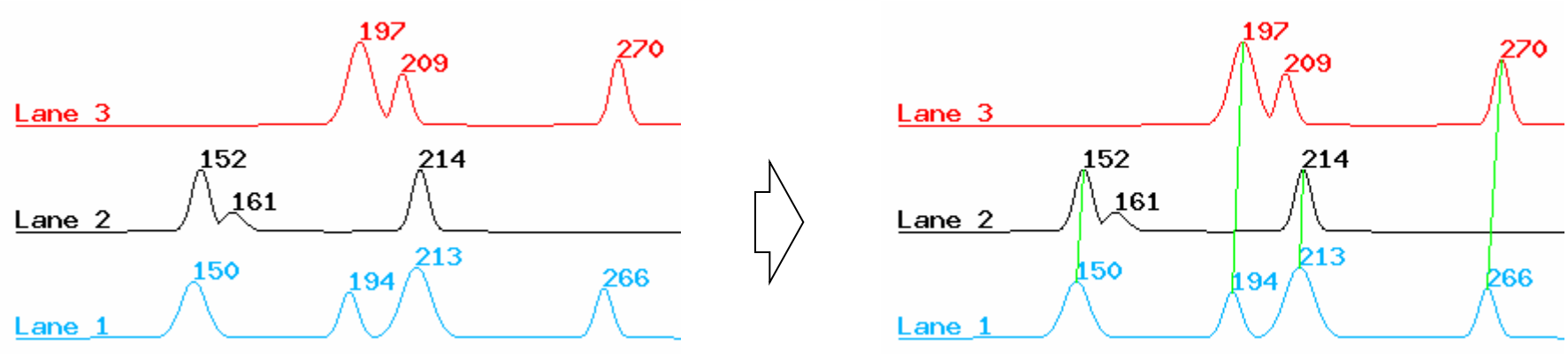
クラスタリング



同一ピーク同定



クラスタリングの考えを同一ピーク認識に応用



②ピーク間距離を計算

②'クラスター間距離が最短のものをマージ

Lane	M.W.
1	150
1	194
1	213
1	266
2	152
2	161
2	214
3	197
3	209
3	270

①分子量でソート

Lane	M.W.	
1	150	
2	152	2
2	161	9
1	194	33
3	197	3
3	209	12
1	213	4
2	214	1
1	266	52
3	270	4

c.	TDF
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0

組織特異的遺伝子検出にエントロピーを利用

- 遺伝子iのエントロピー $H(x_i) = -\sum_{j=1}^{N} p_{ij} \log_2(p_{ij})$, where $p_{ij} = x_{ij} / \sum_{j=1}^{N} x_{ij}$

N: 組織数（jの数）= 8

Hの取りうる範囲: $0 \leqq H \leqq \log_2 N \rightarrow 0 \leqq H \leqq 3$

x_{ij} (i: 遺伝子, j: 組織)

j	遺伝子1	遺伝子2	遺伝子3	遺伝子4	遺伝子5	...
組織1	1	5	6	4	10	...
組織2	0	2	6	4	10	...
組織3	0	1	6	4	10	...
組織4	9	2	6	4	10	...
組織5	0	4	6	10	4	
組織6	0	6	6	4	10	
組織7	0	3	6	4	10	
組織8	0	5	6	4	10	
$\Sigma_j x_{ij}$	10	28	48	38	74	

p_{ij} (i: 遺伝子, j: 組織)

j	遺伝子1	遺伝子2	遺伝子3	遺伝子4	遺伝子5	...
1	0.10	0.18	0.13	0.11	0.14	...
2	0.00	0.07	0.13	0.11	0.14	...
3	0.00	0.04	0.13	0.11	0.14	...
4	0.90	0.07	0.13	0.11	0.14	...
5	0.00	0.14	0.13	0.26	0.05	
6	0.00	0.21	0.13	0.11	0.14	
7	0.00	0.11	0.13	0.11	0.14	
8	0.00	0.18	0.13	0.11	0.14	
Σ_j	1.00	1.00	1.00	1.00	1.00	

$-p_{ij} \log_2(p_{ij})$ (i: 遺伝子, j: 組織)

j	遺伝子1	遺伝子2	遺伝子3	遺伝子4	遺伝子5	...
1	0.33	0.44	0.38	0.34	0.39	...
2	0.00	0.27	0.38	0.34	0.39	...
3	0.00	0.17	0.38	0.34	0.39	...
4	0.14	0.27	0.38	0.34	0.39	...
5	0.00	0.40	0.38	0.51	0.23	
6	0.00	0.48	0.38	0.34	0.39	
7	0.00	0.35	0.38	0.34	0.39	
8	0.00	0.44	0.38	0.34	0.39	
Σ_j	0.47	2.83	3.00	2.90	2.96	

組織特異的遺伝子は低いエントロピー

そうでないものは高い値

実習(組織特異的遺伝子検出)

(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2011/10/13)

What's new?
・最新の論文(Kadota and Shimizu, BMC Bioinformatics, 2011)の結果と絡めて、よくWADに対して寄せられる質問に対する回答を追加しました。(2011/08/02)NEW
・R2.13.1がリリースされていたのでこれに変更しました。(2011/07/14)NEW
・GSA (Efron_2007)の中身をちゃんと埋め始めましたが、まだ最後までは辿りつけてません(2010/8/30)
・Hook (Binder_2008)を追加しました(2010/8/10)
・Agilent two-color processing用のRパッケージを偶然発見したので(項目のみですが...)追加しました(2010/7/14)
・作図 | ROC曲線(ROC curve)を追加しました(2010/4/20)
・このページとは直接関係ありませんが、(Rで)塩基配列解析というページで主に次世代シーケンサーデータ解析を意識したページを作成しつつありますので、そっち方面の解析をRでやりたい方はそちらをご覧ください(2010/5/27)
・作図 | ROC曲線(ROC curve)を追加しました(2010/4/20)
・Linksのところに
・ヒートマップのと

- はじめに (last
- Rのインストール
- Rの昔のバージ
使用例(初心者

- 解析 | 発現変動遺伝子 | 二群間 | 対応あり | 時系列 | について(last modified 2008/3/14)
- 解析 | 発現変動遺伝子 | 二群間 | 対応あり | 時系列 | Di Camillo's method 2 (Di Camillo_2007) (last modified 2008/3/14)
- 解析 | 発現変動遺伝子 | 二群間 | 対応あり | 時系列 | maSigPro (Conesa_2006) (last modified 2009/8/6)
- 解析 | 発現変動遺伝子 | 多群間 | 対応なし | について(last modified 2008/3/17)
- 解析 | 発現変動遺伝子 | 多群間 | 対応なし | 一元配置分散分析(One-way ANOVA) (last modified 2009/07/29)
- 解析 | 発現変動遺伝子 | 多群間 | 対応なし | Kruskal-Wallis(クラスカル-ウォリス)検定 (last modified 2009/07/29)
- 解析 | 発現変動遺伝子 | 組織特異的(選択的)パターン | について(last modified 2011/6/6) NEW
- 解析 | 発現変動遺伝子 | 組織特異的(選択的)パターン | ROKU(Kadota_2006) (last modified 2009/07/30)
- 解析 | 発現変動遺伝子 | 組織特異的(選択的)パターン | Sprent's non-parametric method (Ge_2005) (last modified 2009/07/31)
- 解析 | 発現変動遺伝子 | 組織特異的(選択的)パターン | Schug's H(x) statistic (Schug_2005) (last modified 2011/10/13) NEW
- 解析 | 発現変動遺伝子 | 組織特異的(選択的)パターン | Schug's Q statistic (Schug_2005) (last modified 2009/07/31)
- 解析 | 発現変動遺伝子 | 組織特異的(選択的)パターン | Ueda's AIC-based method (Kadota_2003) (last modified 2009/07/31)
- 解析 | 発現変動遺伝子 | 組織特異的(選択的)パターン | パターンマッチング法(テンプレートマッチング法) (last modified 2009/7/14)
- 解析 | 発現変動遺伝子 | 時系列データ | Periodic genes | Lomb-Scargle periodogram (Glynn_2006) (last modified 2006/7/11)
- 解析 | 発現変動遺伝子 | 時系列データ | Periodic genes | GeneCycle (Ahdesmaki_2005) (last modified 2009/8/3)
- 解析 | 発現変動遺伝子 | 時系列データ | non-periodic genes | について(last modified 2008/3/17)
- 解析 | 発現変動遺伝子 | 時系列データ | non-periodic genes | maSigPro (Conesa_2006) (last modified 2009/8/3)
- 解析 | 発現変動遺伝子 | 時系列データ | non-periodic genes | SAM (Tusher_2001) (last modified 2009/8/3)

実習(組織特異的遺伝子検出)

「ファイル」-「ディレクトリの変更」で解析したい遺伝子発現行列のファイルを置いてあるディレクトリに移動し、以下をコピペ

1. 入力ファイルがGDS1096_rma.txtの場合

```
------    ここから    ------
in_f <- "GDS1096_rma.txt"
out_f <- "hoge.txt"
source("http://www.iu.a.u-tokyo.ac.jp/
data <- read.table(in_f, header=TRUE, r
entropy_score <- apply(data, 1, shannor
tmp <- cbind(rownames(data), data, entr
write.table(tmp, out_f, sep="¥t", apper

#以下は（こんなこともできますという）お
#全遺伝子をentropy_scoreの低い順にソー
out_f2 <- "hoge2.txt"
tmp2 <- tmp[order(entropy_score),]
write.table(tmp2, out_f2, sep="¥t", app

------    ここまで    ------
```

sample15.txt - Microsoft Excel

ファイル | ホーム | 挿入 | ページ レイアウト | 数式 | データ | 校閲 | 表示 | 活用しよう！エクセル

L8

	A	B	C	D	E	F	G	H	I
1	id	tissue1	tissue2	tissue3	tissue4	tissue5	tissue6	tissue7	tissue8
2	gene1	1	0	0	9	0	0	0	0
3	gene2	5	2	1	2	4	6	3	5
4	gene3	6	6	6	6	6	6	6	6
5	gene4	4	4	4	4	10	4	4	4
6	gene5	10	10	10	10	4	10	10	10

2. 入力ファイルがsample15.txtの場合：

```
------    ここから    ------
in_f <- "sample15.txt"                                          #入力ファイル名を指定
out_f <- "hoge.txt"                                             #出力ファイル名を指定
source("http://www.iu.a.u-tokyo.ac.jp/~kadota/R/R_functions.R")  #エントロピーを計算するshannon.entropy関数を含むフ
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#入力ファイルを読み込んでdataに格納
entropy_score <- apply(data, 1, shannon.entropy)                #エントロピーを計算した結果をentropy_scoreに格納
tmp <- cbind(rownames(data), data, entropy_score)               #入力データの右側に計算したentropy_scoreを結合した結
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定したファイル名で保存。

#以下は（こんなこともできますという）おまけ
#全遺伝子をentropy_scoreの低い順にソートした結果を得たい場合：
out_f2 <- "hoge2.txt"                                           #出力ファイル名を指定
tmp2 <- tmp[order(entropy_score),]                              #順位(entropy_score)でソートした結果をtmp2に格納
write.table(tmp2, out_f2, sep="¥t", append=F, quote=F, row.names=F)#tmp2の中身をout_f2で指定したファイル名で保存。

------    ここまで    ------
```

参考文献 (Schug et al., Genome Biol., 2005)

実習(組織特異的遺伝子検出)

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ

R Console

```
> in_f <- "sample
> out_f <- "hoge
> source("http:/
> data <- read.t
> entropy_score
> tmp <- cbind(r
> write.table(tmp
>
> #以下は(こんなこと
> #全遺伝子をentr
> out_f2 <- "hog
> tmp2 <- tmp[or
> write.table(tmp
>
>
```

hoge.txt

	A	B	C	D	E	F	G	H	I	J
1	rownan	tissue1	tissue2	tissue3	tissue4	tissue5	tissue6	tissue7	tissue8	entropy_score
2	gene1	1	0	0	9	0	0	0	0	0.469
3	gene2	5	2	1	2	4	6	3	5	2.826
4	gene3	6	6	6	6	6	6	6	6	3.000
5	gene4	4	4	4	4	10	4	4	4	2.900
6	gene5	10	10	10	10	4	10	10	10	2.959

hoge

hoge2.txt

	A	B	C	D	E	F	G	H	I	J
1	rownan	tissue1	tissue2	tissue3	tissue4	tissue5	tissue6	tissue7	tissue8	entropy_score
2	gene1	1	0	0	9	0	0	0	0	0.469
3	gene2	5	2	1	2	4	6	3	5	2.826
4	gene4	4	4	4	4	10	4	4	4	2.900
5	gene5	10	10	10	10	4	10	10	10	2.959
6	gene3	6	6	6	6	6	6	6	6	3.000

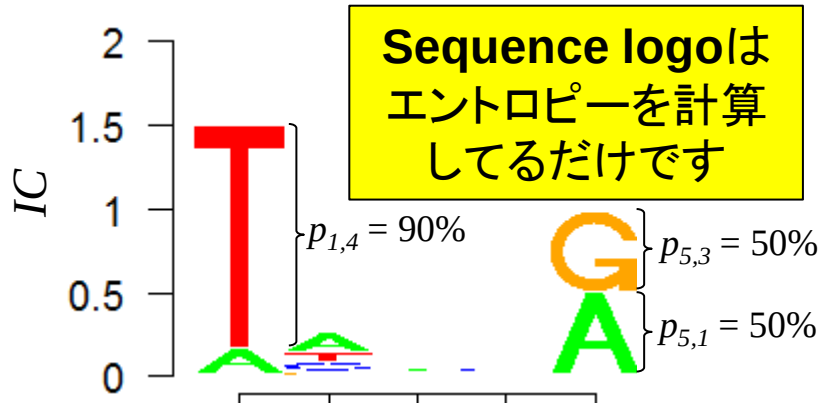
配列モチーフなどの表現にエントロピーを利用

- position iの情報量 $IC_i = \log_2(N) - H(x_i)$
  $= 2$

N: 塩基の種類数= 4
Hの取りうる範囲: $0 \leqq H \leqq \log_2 N$

position i

	1	2	3	4	5	…
配列 1	T	A	C	G	G	…
配列 2	T	A	A	C	G	…
配列 3	T	G	T	A	G	…
配列 4	A	C	T	T	A	…
配列 5	T	T	G	G	A	…
配列 6	T	C	A	A	G	…
配列 7	T	A	C	T	A	…
配列 8	T	T	G	C	A	…
配列 9	T	A	A	C	A	…
配列10	T	A	C	T	G	…



Sequence logoはエントロピーを計算してるだけです

IC	1.53	0.24	0.03	0.03	1.00	…

x_{ij}	1	2	3	4	5	…
Aの数 (j=1)	1	5	3	2	5	…
Cの数 (j=2)	0	2	3	3	0	…
Gの数 (j=3)	0	1	2	2	5	…
Tの数 (j=4)	9	2	2	3	0	…
$\Sigma_j\ x_{ij}$	10	10	10	10	10	

j

p_{ij}	1	2	3	4	5	…
1	0.1	0.5	0.3	0.2	0.5	…
2	0.0	0.2	0.3	0.3	0.0	…
3	0.0	0.1	0.2	0.2	0.5	…
4	0.9	0.2	0.2	0.3	0.0	…
Σ_j	1.0	1.0	1.0	1.0	1.0	

$-p_{ij}\log_2(p_{ij})$	1	2	3	4	5	…
1	0.33	0.50	0.52	0.46	0.50	…
2	0.00	0.46	0.52	0.52	0.00	…
3	0.00	0.33	0.46	0.46	0.50	…
4	0.14	0.46	0.46	0.52	0.00	…
$H = \Sigma_j$	0.47	1.76	1.97	1.97	1.00	

実習（Sequence logo）

(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2010/9/15)

What's new?
・(Rで)マイクロアレイデータ解析中の解析法もここで使う予定(あるいはそちらにリンクさせる予定)です(2010/6/3)
・思いつくままにつらつらと書いています。ある程度たまってきたら、項目名を含め大幅にリニューアルする予定です(2010/5/27)

- 解析 | 一般 | アラインメント(ペアワイズ;基本編2) (last modified 2010/6/8)
- 解析 | 一般 | アラインメント(ペアワイズ;応用編) (last modified 2010/6/8)
- 解析 | 一般 | パターンマッチング (last modified 2010/6/30)
- 解析 | 一般 | GC含量 (last modified 2010/7/1)
- 解析 | 一般 | Sequence logos (Schneider_1990) (last modified 2010/9/15) NEW
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | MARS (last modified 2010/7/9)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | MARS (last modified 2010/7/9)

2. 入力ファイルが塩基組成のファイルの場合：

```
------    ここから    ------
in_f <- "data_seqlogo.txt"                    #読み込みたいファイル名を指定してin_fに格納
library(Biostrings)                           #パッケージの読み込み
library(seqLogo)                              #パッケージの読み込み
hoge <- hoge <- read.table(in_f)              #in_fで指定したファイルの読み込み
out <- makePWM(hoge)                          #情報量(information content; ic)を計算している
seqLogo(out)                                  #塩基組成やicの情報を含むoutを入力としてsequence logoを描画。単

------    ここまで    ------
```

実習（Sequence logo）

data_seqlogo.txt

0.1	0.5	0.3	0.2	0.5
0	0.2	0.3	0.3	0
0	0.1	0.2	0.2	0.5
0.9	0.2	0.2	0.3	0

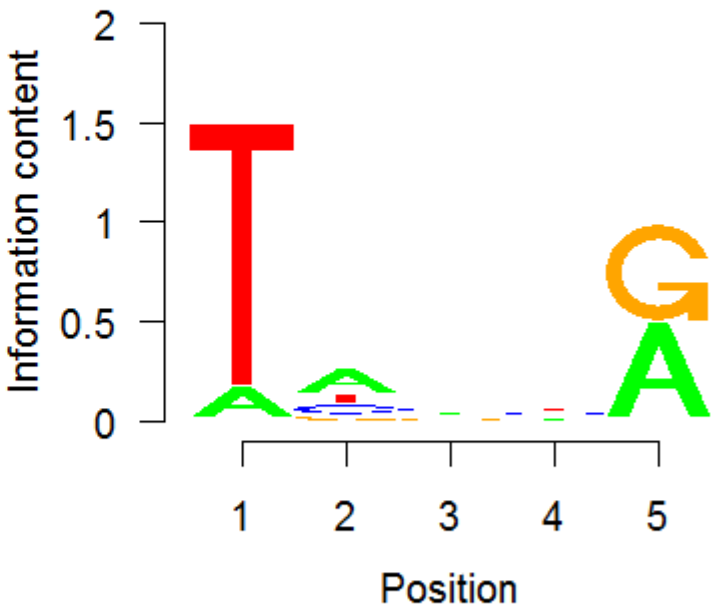
RGui

ファイル　履歴　サイズ変更　ウインドウ

R Console

```
>
> in_f <- "data_seqlogo.txt"                    $
> library(Biostrings)                           $
 要求されたパッケージ IRanges をロード中です

 次のパッケージを付け加えます: 'IRanges'

The following object(s) are masked from 'pack$

    cbind, Map, mapply, order, paste,
    pmax, pmax.int, pmin, pmin.int,
    rbind, rep.int, table

> library(seqLogo)                              $
 要求されたパッケージ grid をロード中です
> hoge <- hoge <- read.table(in_f)              $
> out <- makePWM(hoge)                          $
> seqLogo(out)                                  $
>
>
```

R Graphics: Device 2 (ACTIVE)



まとめ

- 次世代シーケンサー（NGS）を活用した実験解析について、トランスクリプトーム解析など最新の研究技術について学ぶ
- Rを利用することで、NGSから得られる塩基配列データの様々な解析が可能
  - プログラミング能力がなくても使いこなし術があれば…
- NGS解析を全部自力でやるにはLinuxのノウハウがある程度必要であることを実感してもらう
- バイオインフォマティクスの基本的なスキルを身につけることが重要
  - バイオインフォマティクス技術者認定試験合格を目指せ
  - 相関係数やエントロピーなどの要素技術を駆使すれば様々なデータ解析が可能であることを紹介

次世代シーケンサデータもRの**コピペ**で解析可能
→ 頭脳労働

バイオインフォ要素技術の習得は大事だが、それだけでも様々な種類の実験データに対応可能

10:00-19:00（完全週休二日）の研究生活です