

+

# 次世代シーケンシングを 用いた遺伝子発現解析の ためのガイド

統合  
データベース  
講習会

AJACS  
尾張

2017/02/01

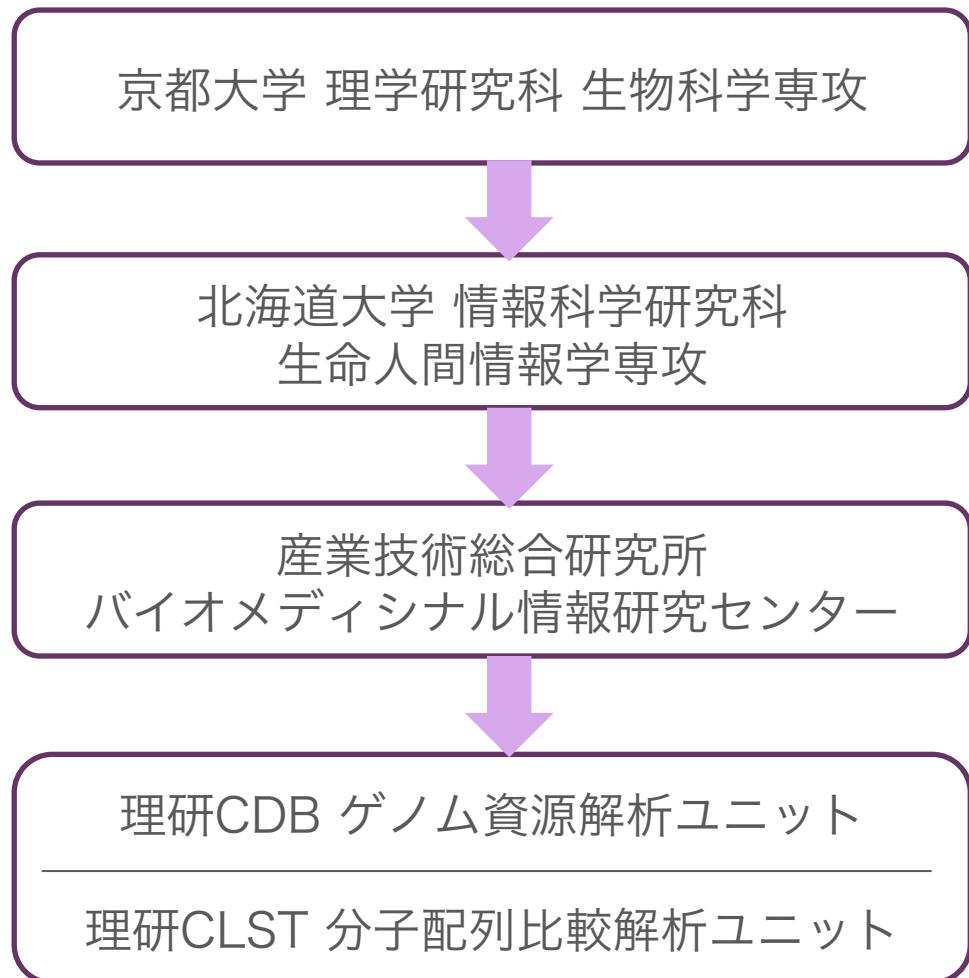
藤田保健  
衛生大学



理研CLST 原 雄一郎

+

# 自己紹介



## 分子進化学・比較ゲノム学

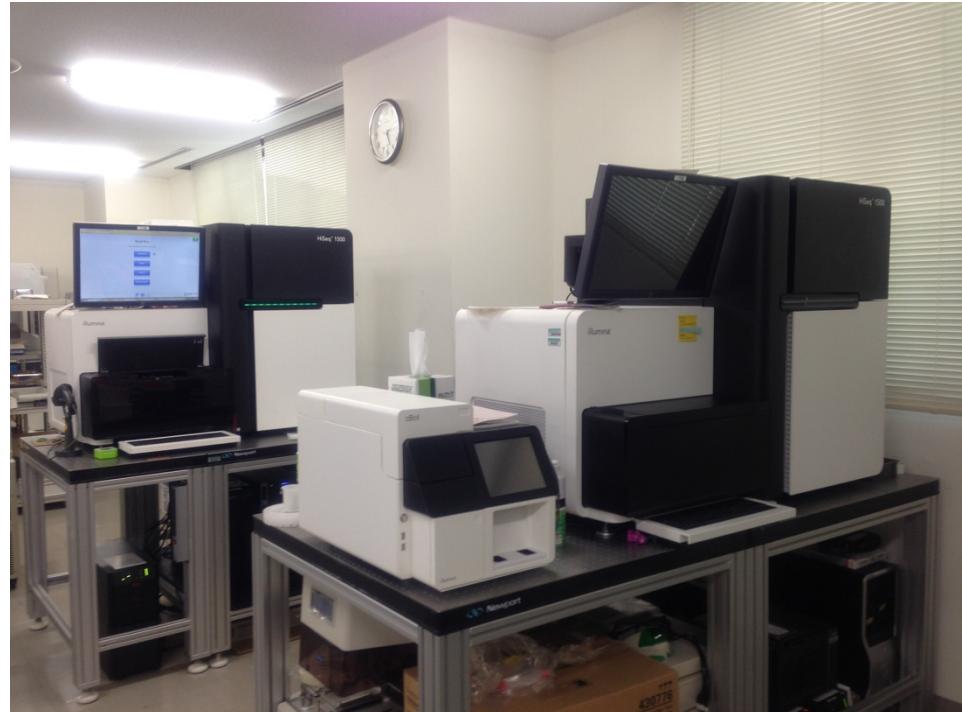
- 生物種の進化史を明らかにする
- 遺伝子の進化から表現型の進化を探る:
  - 遺伝子レパートリーの進化
  - ゲノム構造変異による進化

ヒト遺伝子データベース、分子進化  
データベースの開発運営

シーケンスコアラボにおける  
主に動物の形態形成を対象とした  
トランスクriプトーム解析

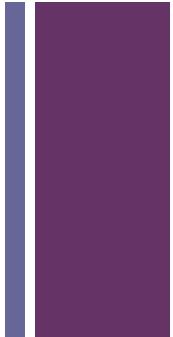
# + シーケンシングコア@神戸理研

- Illumina HiSeq1500 x2
- Illumina MiSeq
- Applied Biosystems 3730
- Applied Biosystems 3130

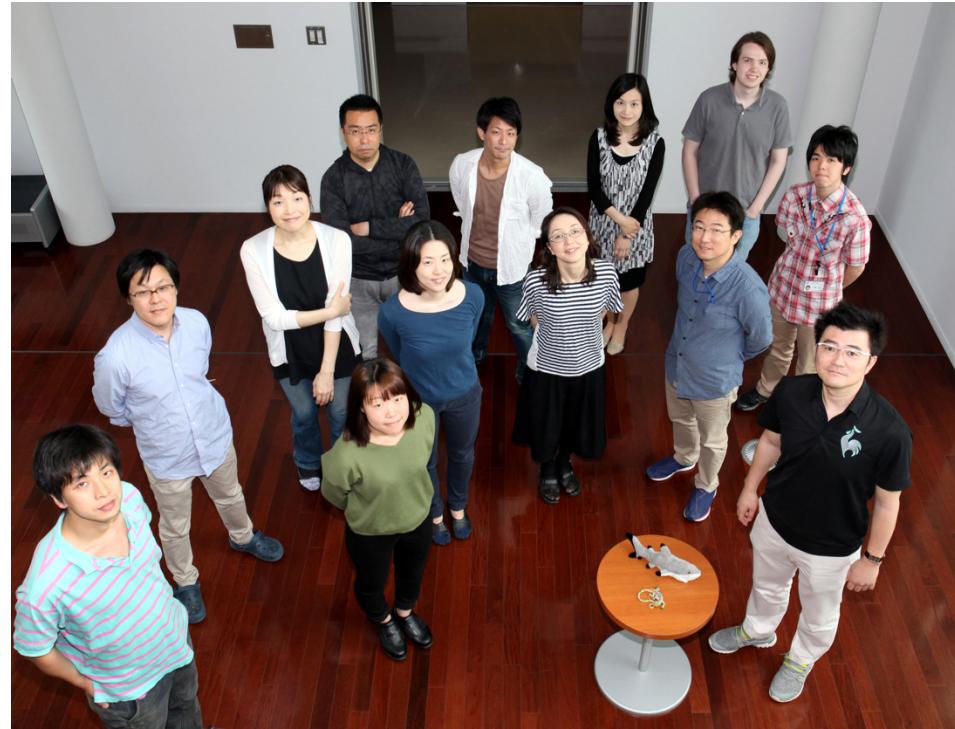


+

# シーケンシングコア@神戸理研



- WetとDryのスタッフが密に連携
- RNA-Seqを中心に、Chip-seq、ATAC-seq、新規ゲノム配列決定、エキソームシーケンシングを行っている



+

# 今日はこのような方を想定して 講習・実習を行います

- RNA-seqを用いて2群間の発現比較解析を行いたい
- なるべく自分でデータをハンドリングしたいけど、プログラムの実行だけでなく計算機のセットアップするスキルも無い
- まずはRNA-seqの実験を問題なく行い、データが持つ生物学的情報を正しく解釈するためのリテラシーをお伝えします。
- 実際のデータのハンドリング(計算機を用いてシーケンスリードから発現定量値を算出)については、概要のみ話します

+

# 今日はこのような方を想定して 講習・実習を行います

- RNA-seqを用いて2群間の発現比較解析を行いたい
- なるべく自分でデータをハンドリングしたいけど、プログラムの実行だけでなく計算機のセットアップするスキルも無い

## 伝えたいこと

- RNA-seqの入口: 実験計画の立て方
- RNA-seqの出口: データを見る目を養う
  - シーケンスリードのクオリティ評価
  - 発現定量値から導き出される生物学的特徴の抽出

# ローマは一日にしてならず



**NBDC** バイオサイエンスデータベースセンター

National Bioscience Database Center

English サイトマップ サイト内検索 検索...

ホーム NBDCについて 研究開発 公募情報 採用情報 イベント 人材支援 アクセス リンク

Home > 人材支援 > 支援 > 講習会 > 平成28年度NGSハンズオン講習会

平成28年度NGSハンズオン講習会

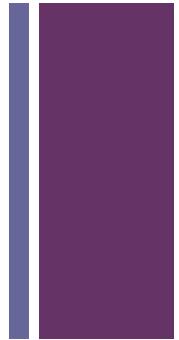
バイオインフォマティクス人材育成カリキュラム  
次世代シーケンサ(NGS)ハンズオン講習会を開催します。

H28.7.4

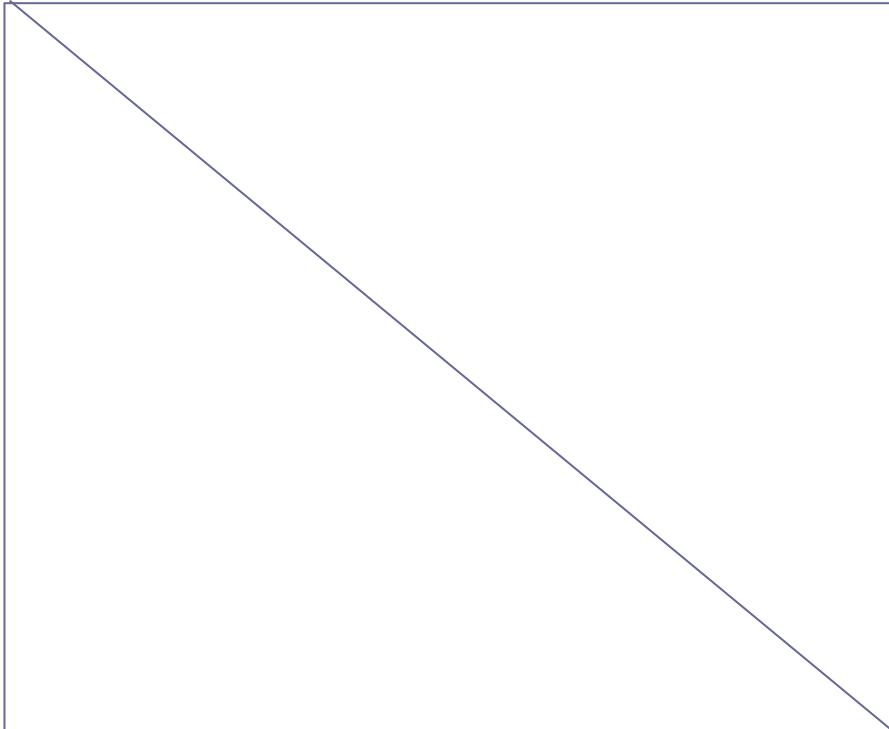
実施日	実施時間	大項目	タイトル	内容(予定)	担当講師(敬称体)
7月19日(水)	10:30~18:15	第1部 統計解析 (農学生命情報科学特論I)	PC環境の構築 Bio-LinuxとRのインストール状況確認	<ul style="list-style-type: none"> <li>Bio-Linux(第2回および一部で利用するovaファイル)の導入確認</li> <li>共有フォルダ設定完了確認</li> <li>RとBioconductorのインストール状況確認</li> <li>FASTAによるバイオデータのインストール確認</li> <li>講師指定の事前予習内容の再確認</li> <li>講習会期間中に貸与されるノートPCを用いた各種動作確認</li> </ul>	主催・共催機関
7月20日(木)	10:30~18:15		ゲノム解析、塩基配列解析	<ul style="list-style-type: none"> <li>NGS解析手順(ワークフロー)(例: Pipeline)との連携</li> <li>NGSデータの出力結果(FASTQ)の確認とその意味の理解</li> <li>NGSデータの出力結果解析(例: FASTQC)と連携</li> <li>塩基配列解析を行ったための基本スキルの復習や作成</li> <li>de novoアセンブリ時のエラーデータやゲノムサイズ推定の基本的な考え方</li> </ul>	
7月21日(金)	10:30~18:15		トランскriプトーム解析1	<ul style="list-style-type: none"> <li>カウンターナンバー等による塩基配列解析(Reads seq)</li> <li>スカラル度クロスランクイング、直交の解釈</li> <li>発現量比較(段階あり絶対比較)</li> <li>分子やゲノム、変異ゲノム</li> <li>次回実験(例: RNA-seq, ChIP-seq), および組換えの解釈</li> </ul>	門田幸二 (東京大学)
7月22日(土)	10:30~18:15		トランスクriプトーム解析2	<ul style="list-style-type: none"> <li>段階あり3回間比較(TOGLによるANOVA的な解釈)</li> <li>ゲノムアノテーション、post-hoc test</li> <li>遺伝子クロスオーバークラスター検出(PCA)</li> <li>段階あり3回間比較(NGSとBaySeq); より発現パターン分類</li> <li>段階なし3回間比較(TOGL)、および結果の説明</li> <li>PCAによる正規化、TOGL正規化を組み合わせた各種解析</li> </ul>	
7月25日(火)	10:30~18:15	第2部 NGS解析(初～中級)	NGS解析基礎	<ul style="list-style-type: none"> <li>ファイル形式 可視化 (IGV) quality check</li> <li>マッピング アセンブル</li> </ul>	山口昌樹 (アメリカフ) 鷹川美男 (アメリカフ)
7月26日(水)	10:30~18:15		ゲノムReseq、変異解析	代表的なバイオラインについての実習：ゲノムReseq、変異解析	山口昌樹 (アメリカフ) 三澤拓真 (アメリカフ)
7月27日(木)	10:30~18:15		RNA-seq	代表的なバイオラインについての実習	山口昌樹 (アメリカフ) 尾上広祐 (アメリカフ)
7月28日(金)	10:30~18:15		ChIP-seq	代表的なバイオラインについての実習	山口昌樹 (アメリカフ) 久保亮一 (アメリカフ)
8月1日(月)	10:30~18:15	第3部 NGS解析(中～上級) (農学生命情報科学特論II)	Linux環境でのデータ解析： JavaやRの利用法	<ul style="list-style-type: none"> <li>日本乳酸菌学会誌のNGS連載第4回の復習(特にFastQCとFastQC)</li> <li>乳酸菌連載第5回(FASTQCによるデータの確認)</li> <li>FASTQCによるデータの確認(FASTQC)</li> <li>Javaプログラミングの設定(JavaKeeper)</li> <li>Linux環境でのデータ利用法(対話モードとバッチモード)</li> </ul>	
8月2日(火)	10:30~18:15		Linux環境でのデータ解析： マッピング、トリミング、アンソブリ	<ul style="list-style-type: none"> <li>NGS連載第5回(残り)、第6回(7月まで) レジスターFastQCと用いたRNA-seqデータのマッピング 実験結果によるデータトリミング(Biostringsとfastx Toolkitによる)</li> <li>NGS連載第6回(残り)、第7回(8月まで) NGS連載第7回(残り)、NGS連載第8回(8月まで) 11luminous MSeriesデータの転換と前処理(FastQCとFastQC) de novo データマッピング(Velvet)</li> </ul>	
8月3日(水)	10:30~18:15		クラウド環境での連携、 コンダグリードデータの解析	<ul style="list-style-type: none"> <li>NGS連載第8回(残り)、ゲノムサイズ推定(GenomeSize) 配列長によるデータトリミング(Pythonプログラム実行と収容) NGS連載第9回(残り)、Velvetによるde novo データの公共化</li> <li>コンダグリード(Condigrid)データと公共化</li> <li>フリーソフト(era, FastQC, baxkit, SBR Toolkit, FastQC BBH Pipeline (BBAP)、エクスポート機能)</li> </ul>	門田幸二 (東京大学)
8月4日(木)	10:30~18:15		トランスクriプトームアセンブリ、 発現量推定	<ul style="list-style-type: none"> <li>de novoトランスクriプトームアセンブリ(Trinity k-Bridge) de novoトランスクriプトームアセンブリ(Trinity k-Bridge) 発現量推定(TIGAR)</li> </ul>	

+

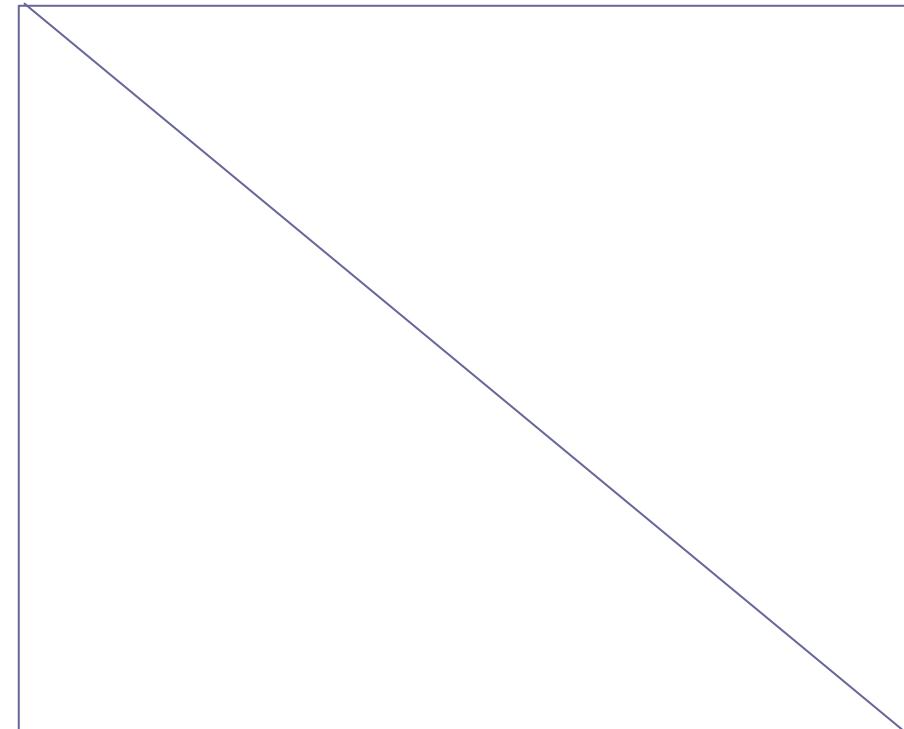
企業のシーケンシングサービスに付随する情報解析を活用するのも一案



A社



B社



ただし、実験のクオリティや生物学的解釈はクライアントに責任がある

# + 今日お話しすること

## 1. イントロダクション

- RNA-seq: 次世代シーケンシングデータから遺伝子発現を知る

## 2. RNA-seqの入口

- 実験計画: RNA-seqを行う前に
- 発現解析の流れ

## 3. RNA-seqの出口

- クオリティーチェック(←実習)
- データの可視化(←実習)

# + 今日お話しすること

## 1. イントロダクション

- RNA-seq: 次世代シーケンシングデータから遺伝子発現を知る

## 2. RNA-seqの入口

- 実験計画: RNA-seqを行う前に
- 発現解析の流れ

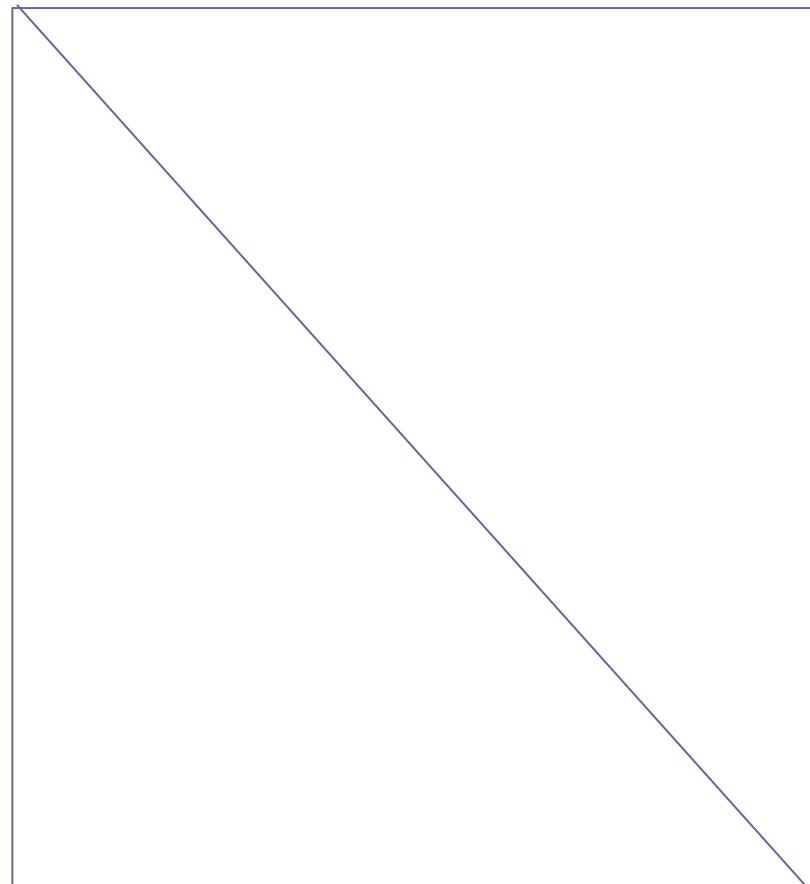
## 3. RNA-seqの出口

- クオリティーチェック(←実習)
- データの可視化(←実習)

+

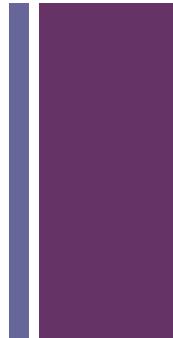
# 転写産物の配列情報を読み取る

- mRNA→cDNAに逆転写してDNA配列情報とし、シーケンサーで読み取る

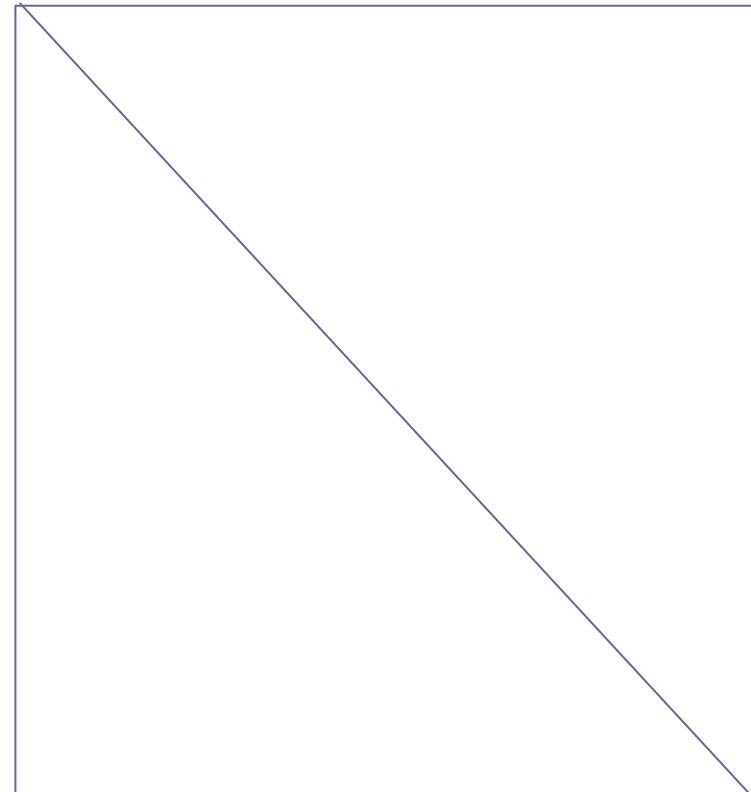


+

# RNA-seq



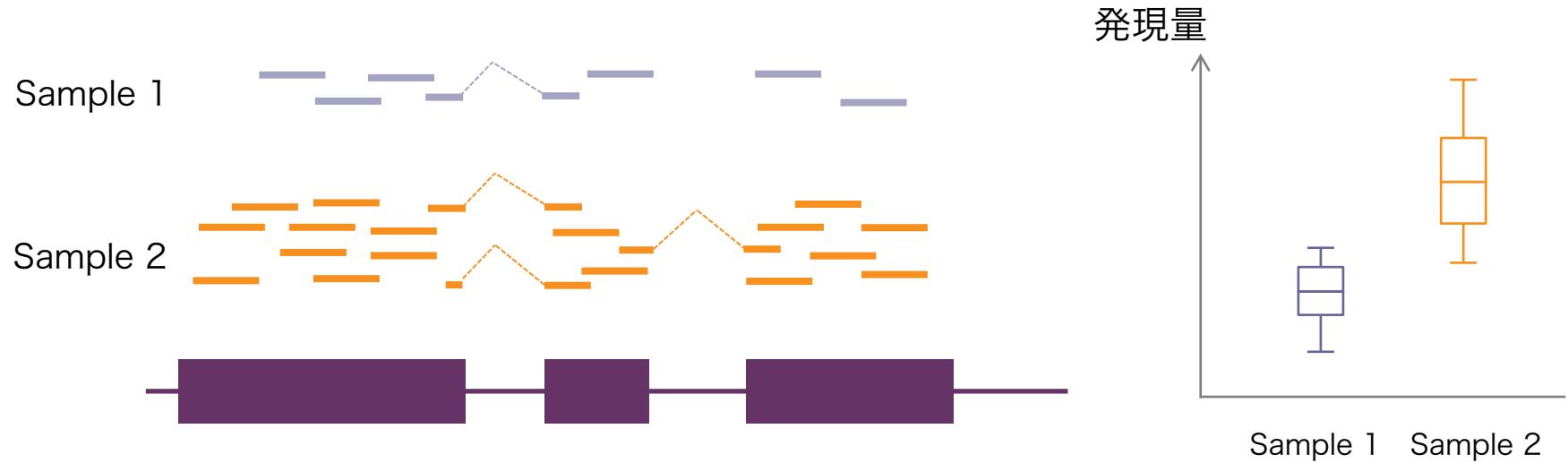
- 次世代シーケンシンサーを用いて、ある組織/細胞に発現する遺伝子配列を網羅的に読み取る(→トランスクriプトーム)
- 短い配列
  - 1本のシーケンスリードは100 bp前後
- 膨大な分子数
  - 動物サンプルの場合、1サンプルにつき1000万リード前後



+

# RNA-seqによる発現解析

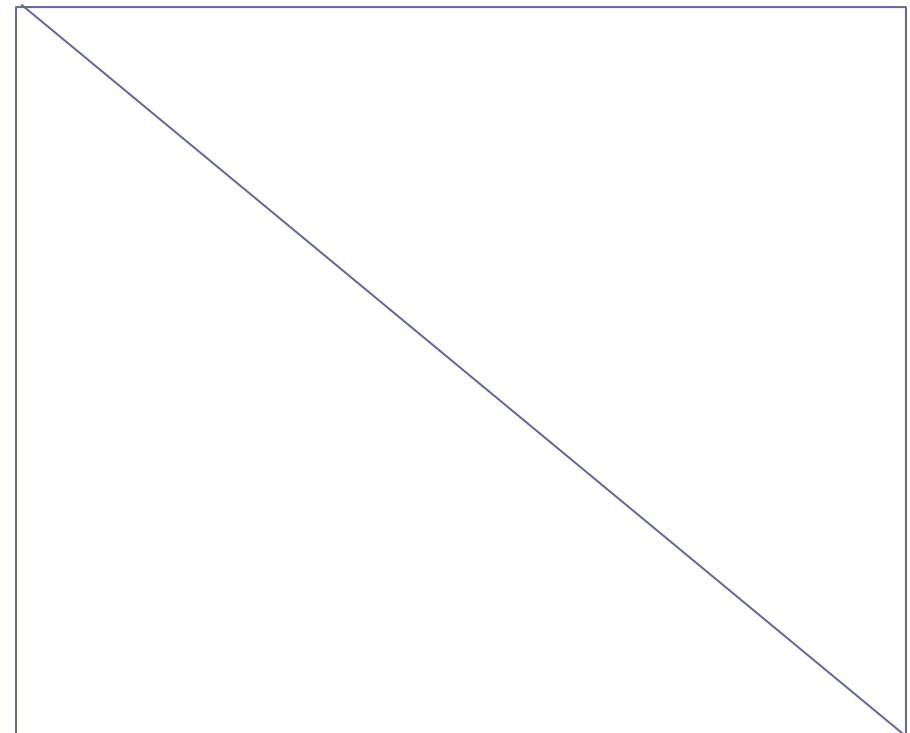
- 転写産物に貼り付けられるシーケンスリードの数から発現を定量化する
- 複数のサンプルで発現量を比較し、特異的発現を示す遺伝子を同定する
- 既知の遺伝子情報と照らし合わせ、新規の遺伝子やスプライシングバリエントを同定する



+

発現解析から例えばこんなことが  
調べられてきた

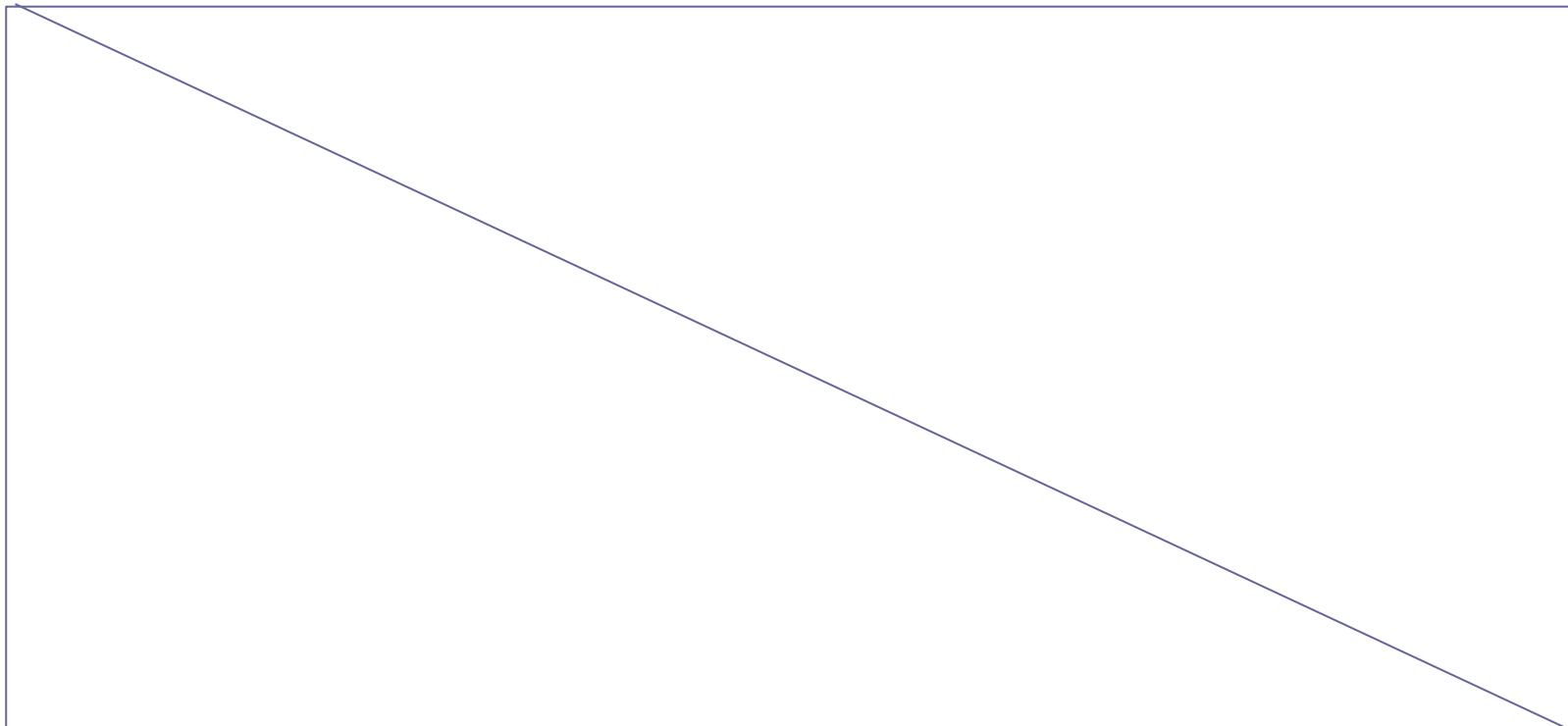
- ツメガエルの発生初期における遺伝子発現の変化



+

## 発現解析から例えばこんなことが 調べられてきた

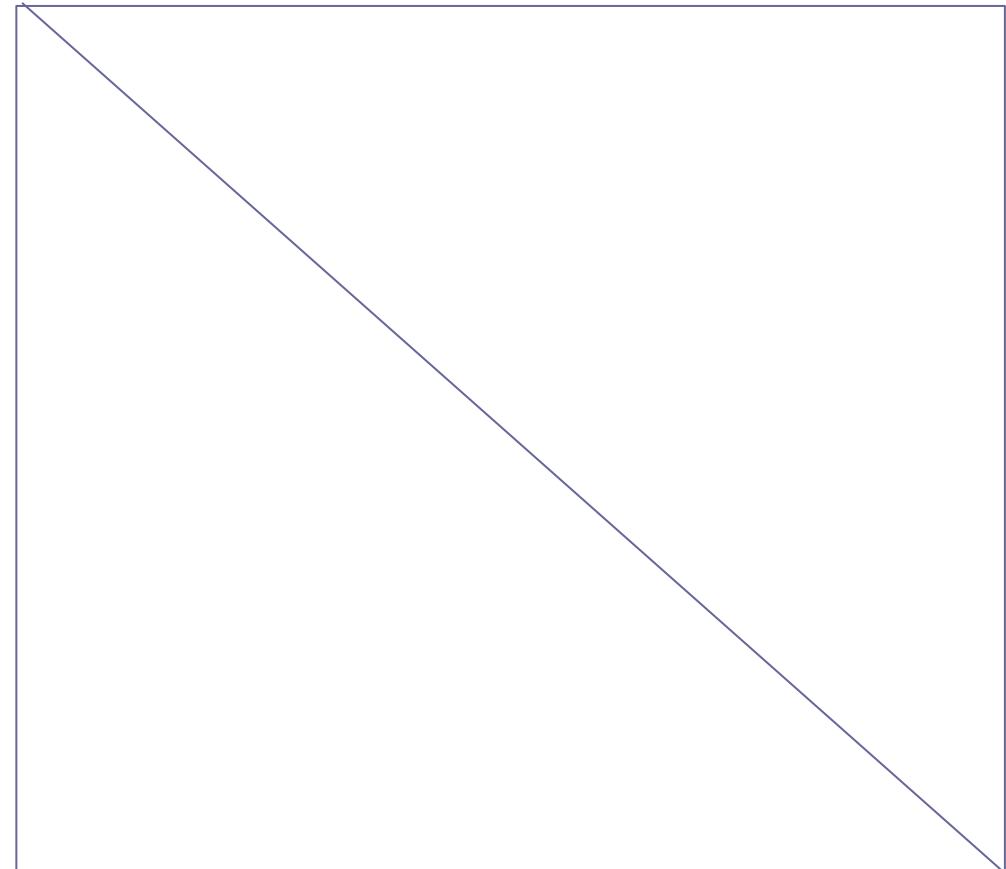
- ヒト成人の褐色脂肪細胞を特徴付ける遺伝子発現プロファイルの同定
  - ヒト成人の褐色脂肪細胞はマウスのベージュ脂肪細胞に似た発現プロファイルを示す
  - 発現解析をもとに新規マーカー遺伝子を同定



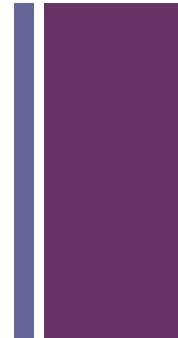
+

発現解析から例えばこんなことが  
調べられてきた

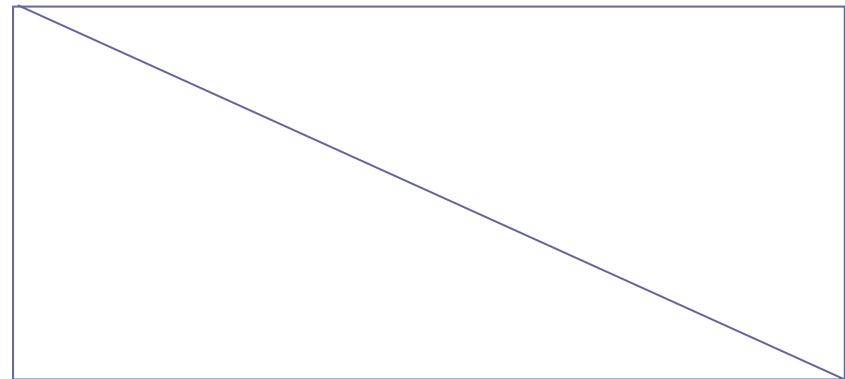
- がん細胞のトランск립トーム解析: fusion geneの同定



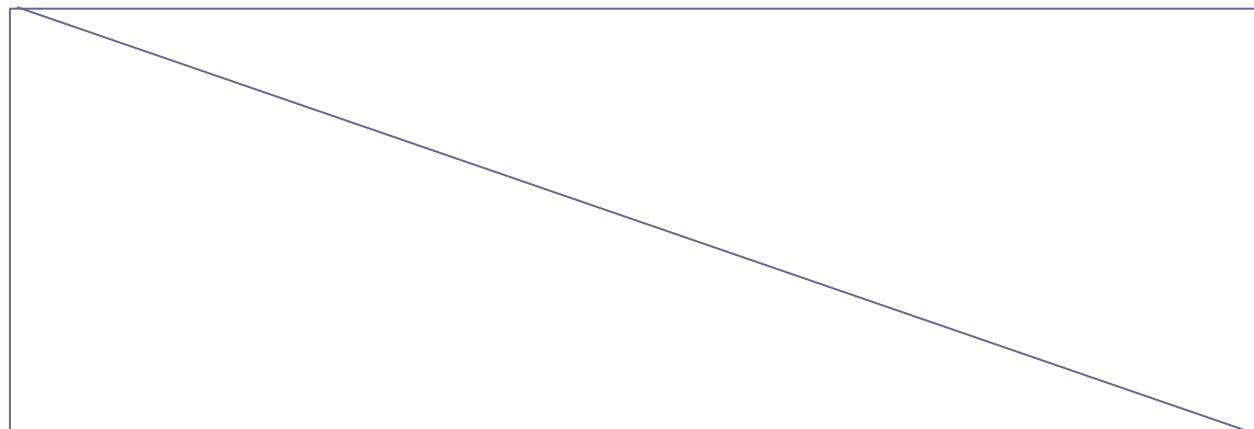
# + 1細胞RNA-seq



- 1細胞の転写産物を増幅してシーケンシング
- 組織のRNA-seqレベルではわからなかった細胞ごとの特徴を追跡できる
  - がん細胞
  - 中枢神経系
  - 幹細胞の分化



Sandberg Nature methods (2014).



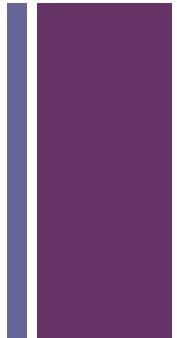
Patel, et al. Science (2014)

+

## シーケンシングしよう！でもその前に

- RNA-seqにはそれなりのコストがかかる(十数万円～)
- 後戻りできる箇所と出来ない箇所を認識する
  - サンプル調製→シーケンシング: 後戻りできない！
  - シーケンシングデータ解析: やり直しできる
- クオリティの低いデータをバイオインフォマティクス解析で挽回するのは困難
- バイオインフォマティクス解析では、クオリティの低いデータでからも何らかの結果を出してしまう
- たとえデータ解析を全て他の方に依頼したとしても、データを見る目を養うことは必須

## + Take-home message



Bioinformatics is not a magic!



# + 今日お話しすること

## 1. イントロダクション

- RNA-seq: 次世代シーケンシングデータから遺伝子発現を知る

## 2. RNA-seqの入口

- 実験計画: RNA-seqを行う前に
- 発現解析の流れ

## 3. RNA-seqの出口

- クオリティーチェック(←実習)
- データの可視化(←実習)

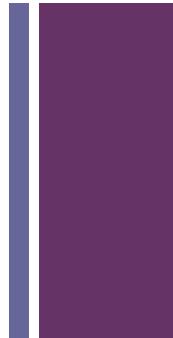
+

# 「よくある」 RNA-seq

- 特定の2群で発現に差異がある遺伝子を同定
  - 同じ幹細胞由来の異なる細胞種間→細胞の運命を決定する遺伝子の同定
  - Wild type vs. Mutants→ノックアウト/ノックイン/etc した遺伝子が及ぼす効果を測定
  - 同一の組織を異なる時間でサンプリング→時間軸に基づく遺伝子発現のトラッキング
- ある組織/細胞に発現する遺伝子のサンプリング
  - 発現遺伝子のカタログ化
  - lncRNAなどの新規転写産物の同定
  - ゲノムが解読されていない生物種の遺伝子カタログ作成(de novo トランскriプトームアセンブリ)

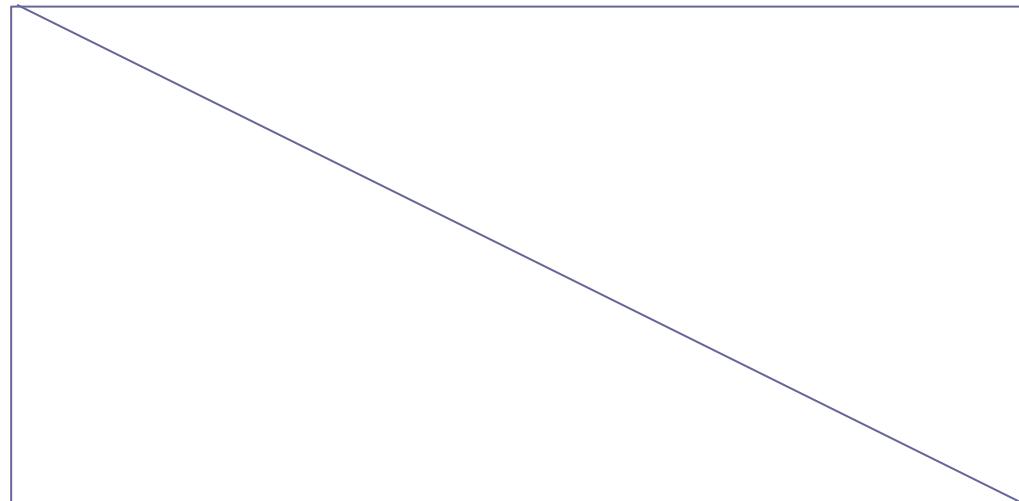
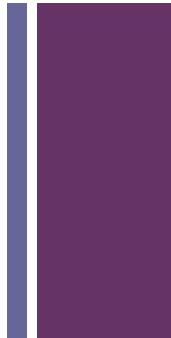
+

# 実験計画



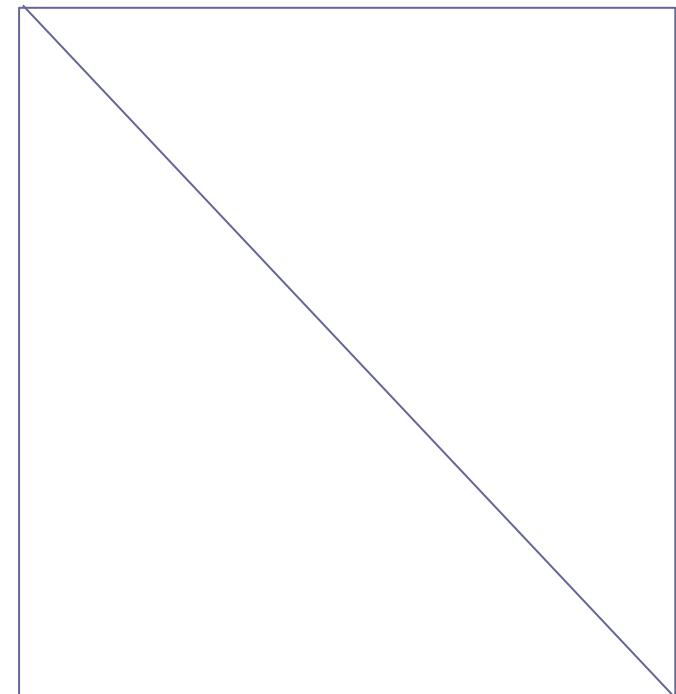
- 「よいデータ解析」を行うための計画を立てる
- データ解析担当も交えて立案を。「とりあえずシーケンスしたけど誰か解析して」はダメ
- Differential expression analysisに対するチェックポイント
  - Biological replicates
  - 細胞数(RNA量)を十分確保できるか
  - ライブライリ作成方法、どのライブライリ作成キットを用いるか
  - シーケンスするリード量(=フローセルサイズ)は？
  - リファレンスゲノムデータの入手可否
  - 実験系統とリファレンスゲノムの系統との遺伝的距離

# + Biological replicates



Klaus. The EMBO Journal (2015).

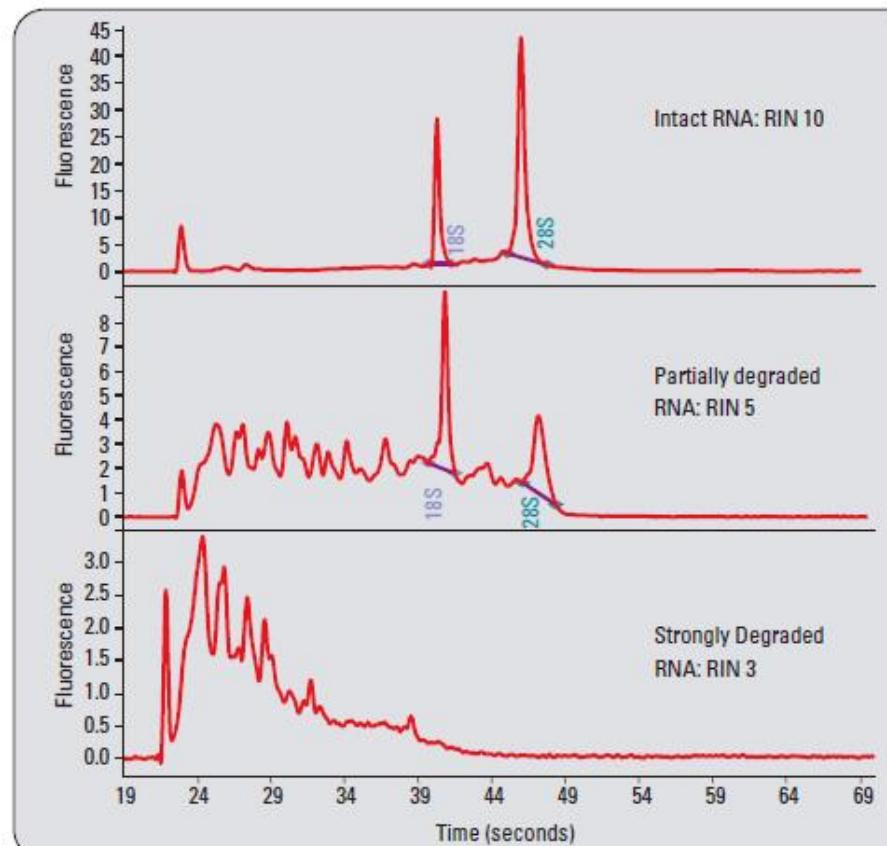
- Replicatesが多いほどロバストな結果を得やすい
- 我々は $\geq 3$  replicatesを推奨している



Regassa, A., et al. BMC genomics (2011).

# + ライブラリ調製

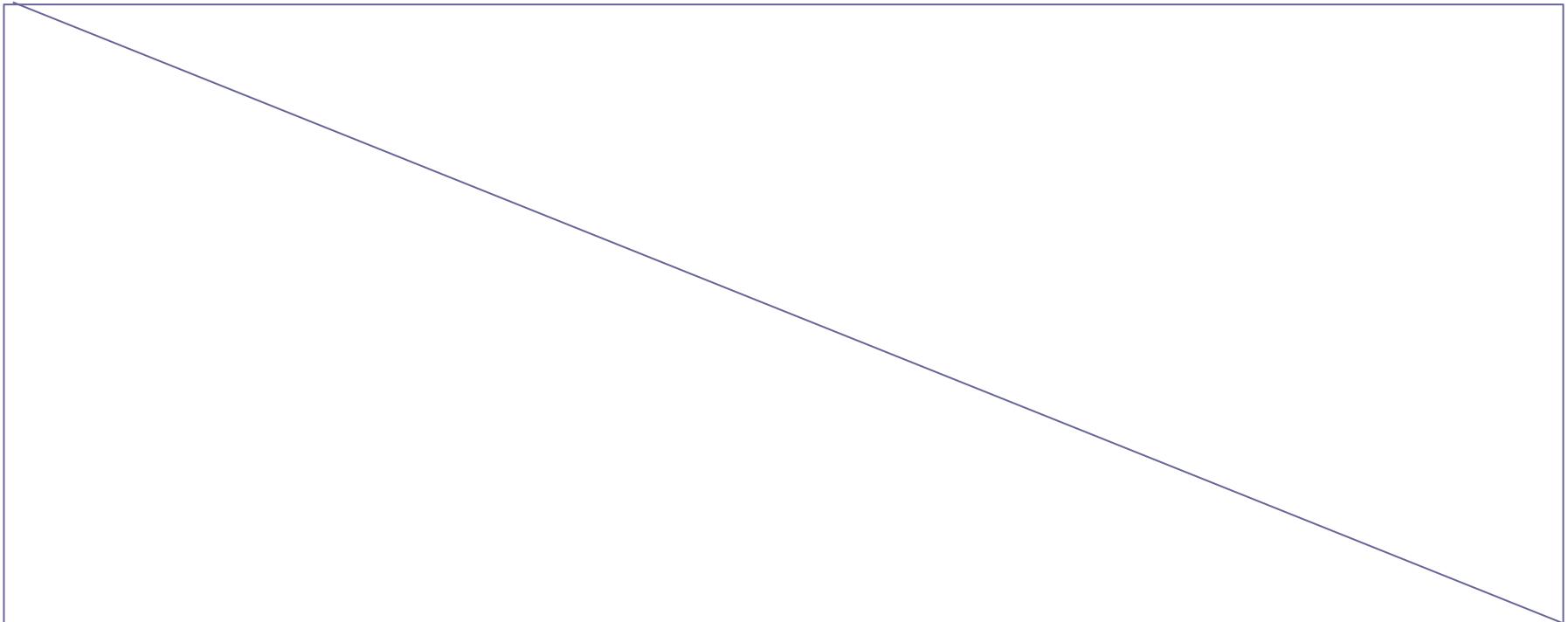
- RNA amount/quality
  - 1 μg total RNA
  - 分解していない(RNA Integrity Number (RIN)  $\geq 8$ を目安に)
  - 最小量のサンプルにあわせる
- mRNA isolation from total RNA
  - Poly-A selection or rRNA removal
- Fragmentation
  - Read length, SR/PEを考慮
- PCR cycle
  - 少ないほどPCRによるバイアスの影響が小さい
  - RNA amountが大きいほどサイクル数を小さくできる



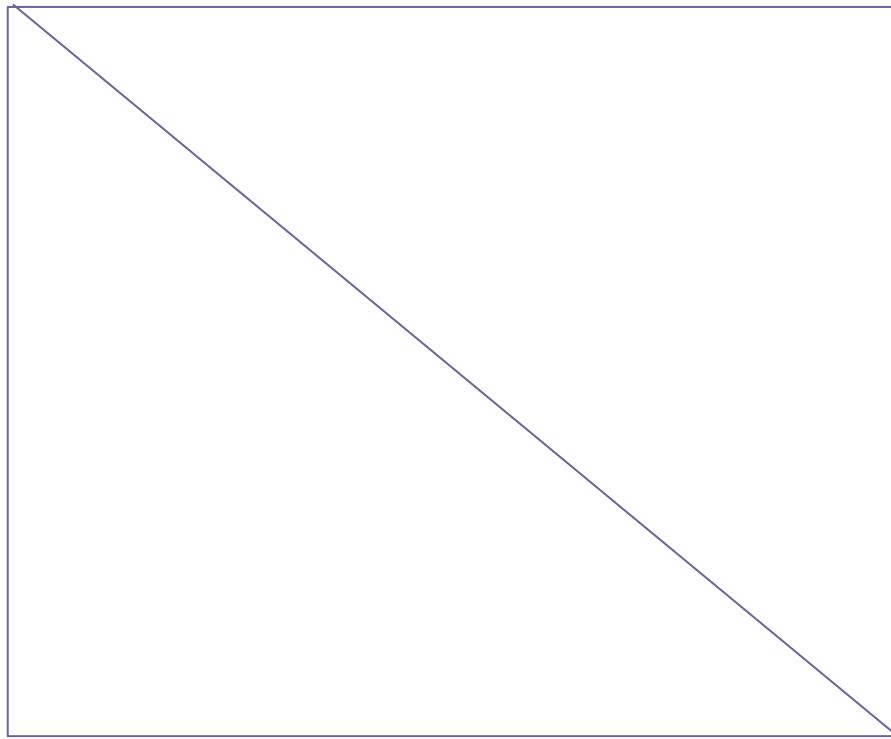
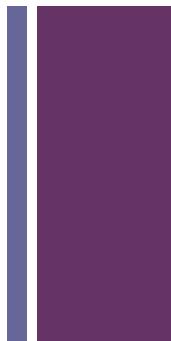
+

# Batch effects

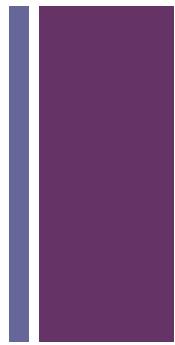
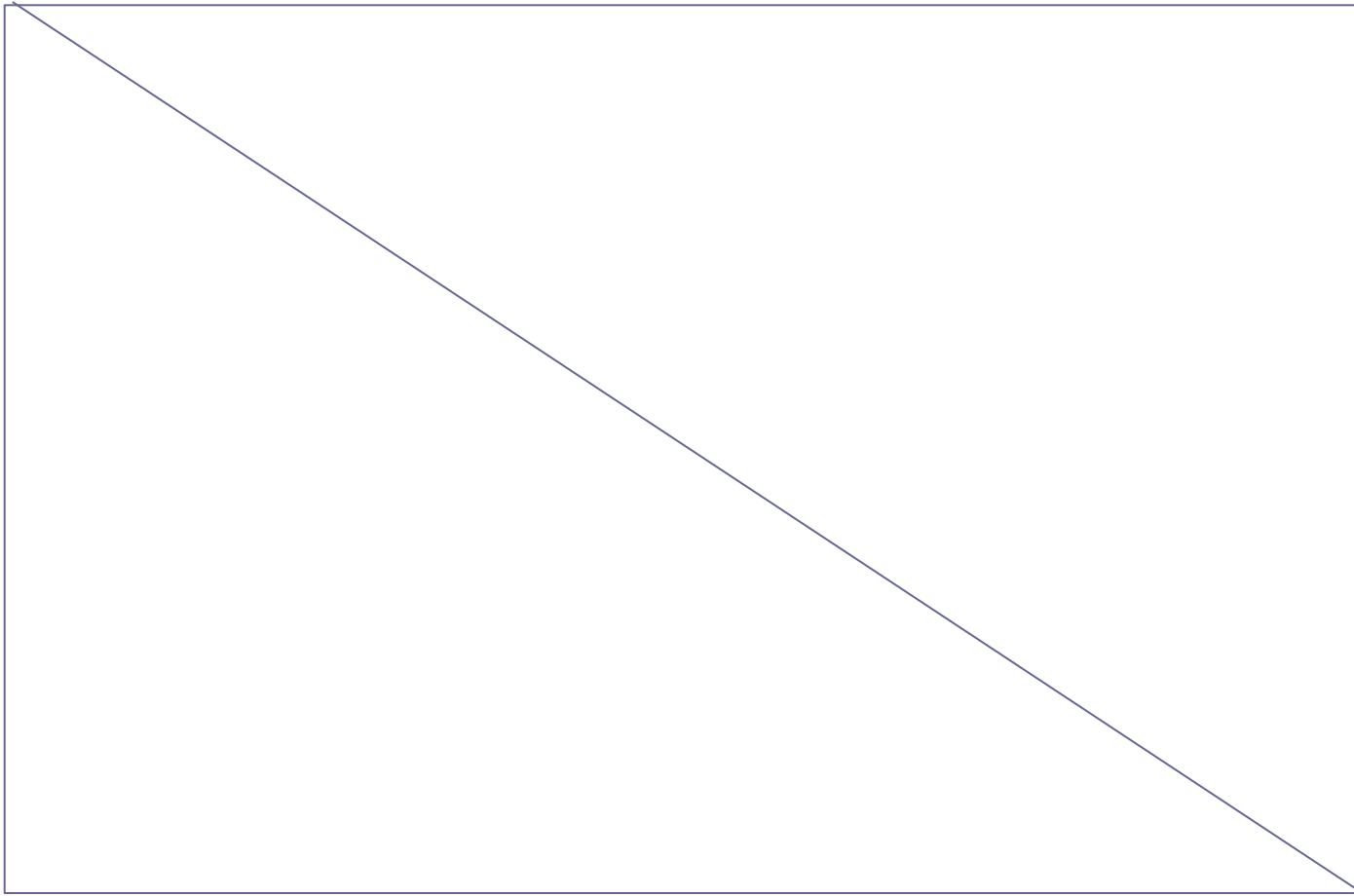
- 実験環境により生じる結果の変動



+



+



Kit 1

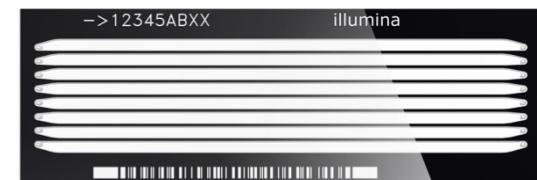
Kit 2

Kit 3

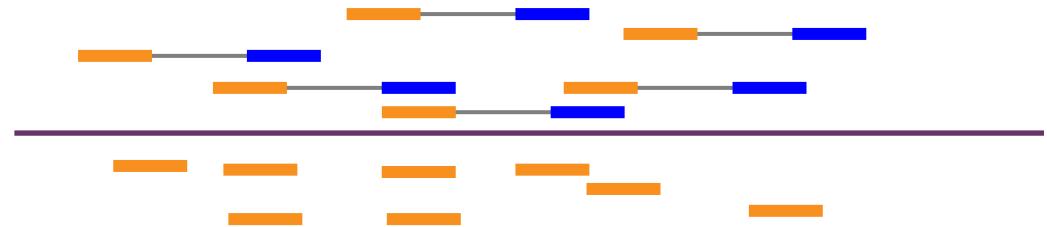
Kit 4

# + シーケンシング

- プラットフォーム
- サイクル数(リード長)
- リード数(レーン数)
- シングルエンドorペアエンド

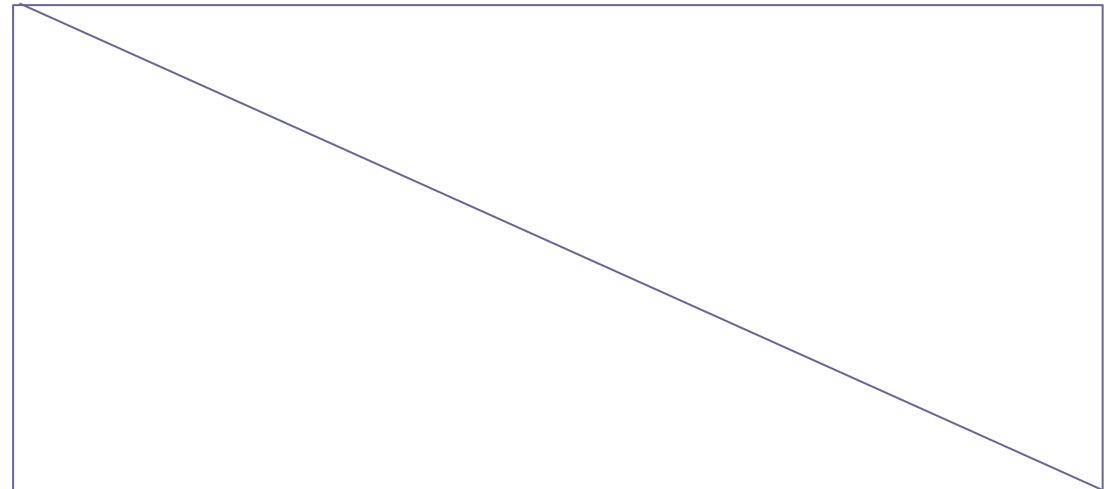
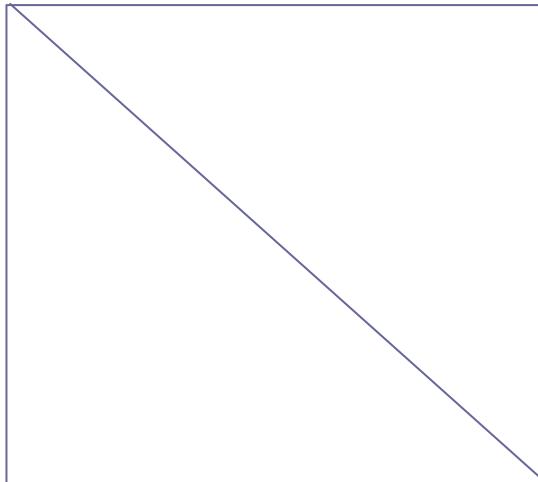


得たい情報量、予算にあわせて選択



# + リード数とreplicateのどちらをとるか？

- 10M reads/sampleを超えるとreplicatesを増やす方が感度が上がる
- Replicatesを増やすことは、発現比が小さいが有意に発現レベル異なる遺伝子の同定に有効
- Replicatesの数はfalse positiveにはほぼ影響ない
- 予算(= $\#$ reads)が限られているなら、サンプル数を増やす方に努めるのがよい



+

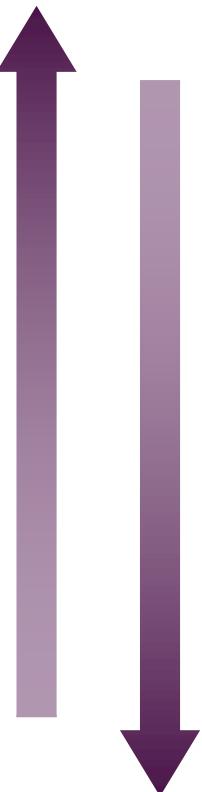
# どこでシークエンシングする？

- 所属する研究室
- 所属大学や機関のシークエンシングファシリティ
- 企業へのアウトソーシング
- 共同研究先
- シーケンスを行わず、データベースからシークエンスデータを取得する

+

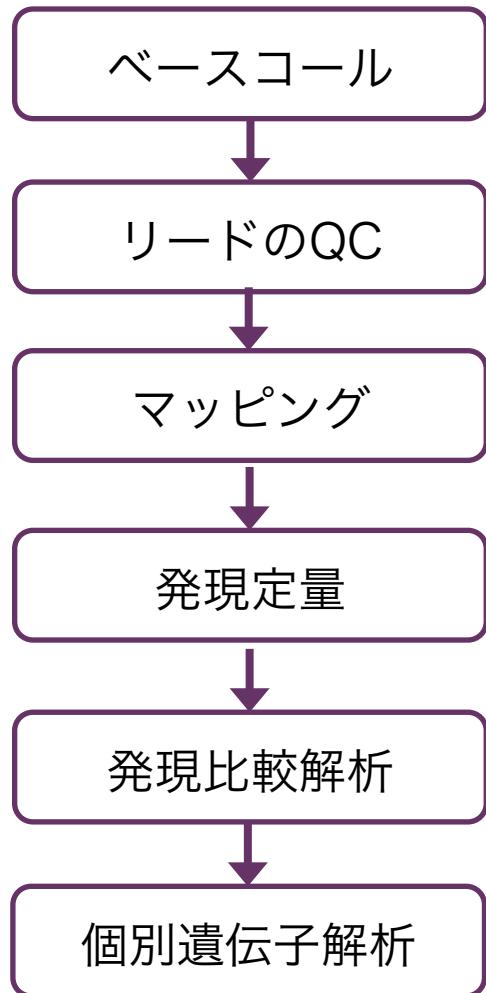
# どのかたちでデータを受け取る？

高いバイオインフォマティクスの  
スキルを要する

- 
- FASTQファイル: 自分でシーケンスしている。シーケンスのみアウトソーシングしている。バイオインフォマティクスのスキルがある
  - マッピングデータ(BAMファイル)、発現定量、発現比較解析の結果: アウトソーシングや共同研究で一次解析まで行ってもらっている
  - エンリッチメント解析の結果、各解析の図示: 全ての解析を依頼している

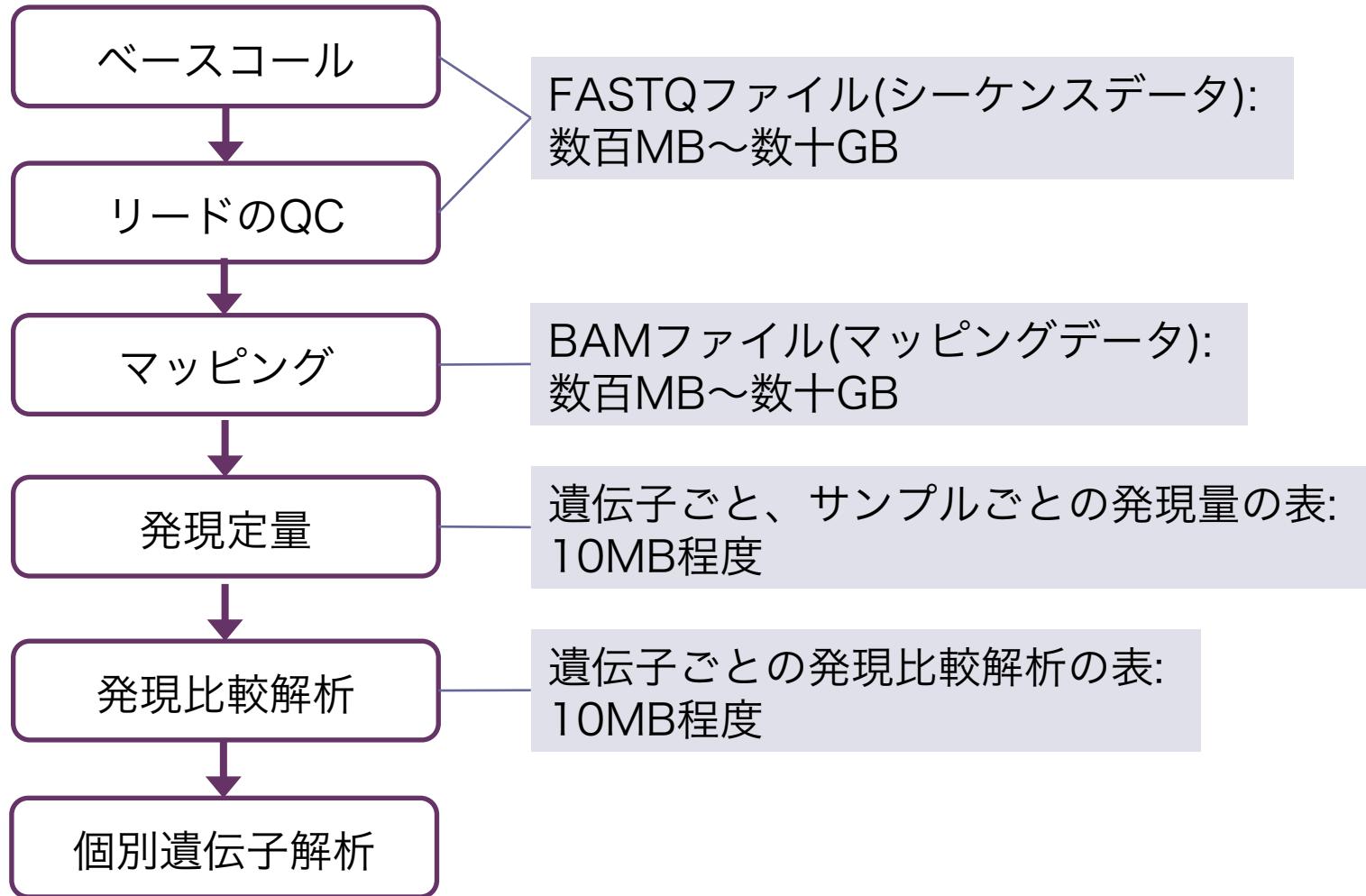
解析がより進んだ状態

# + データ解析の流れ



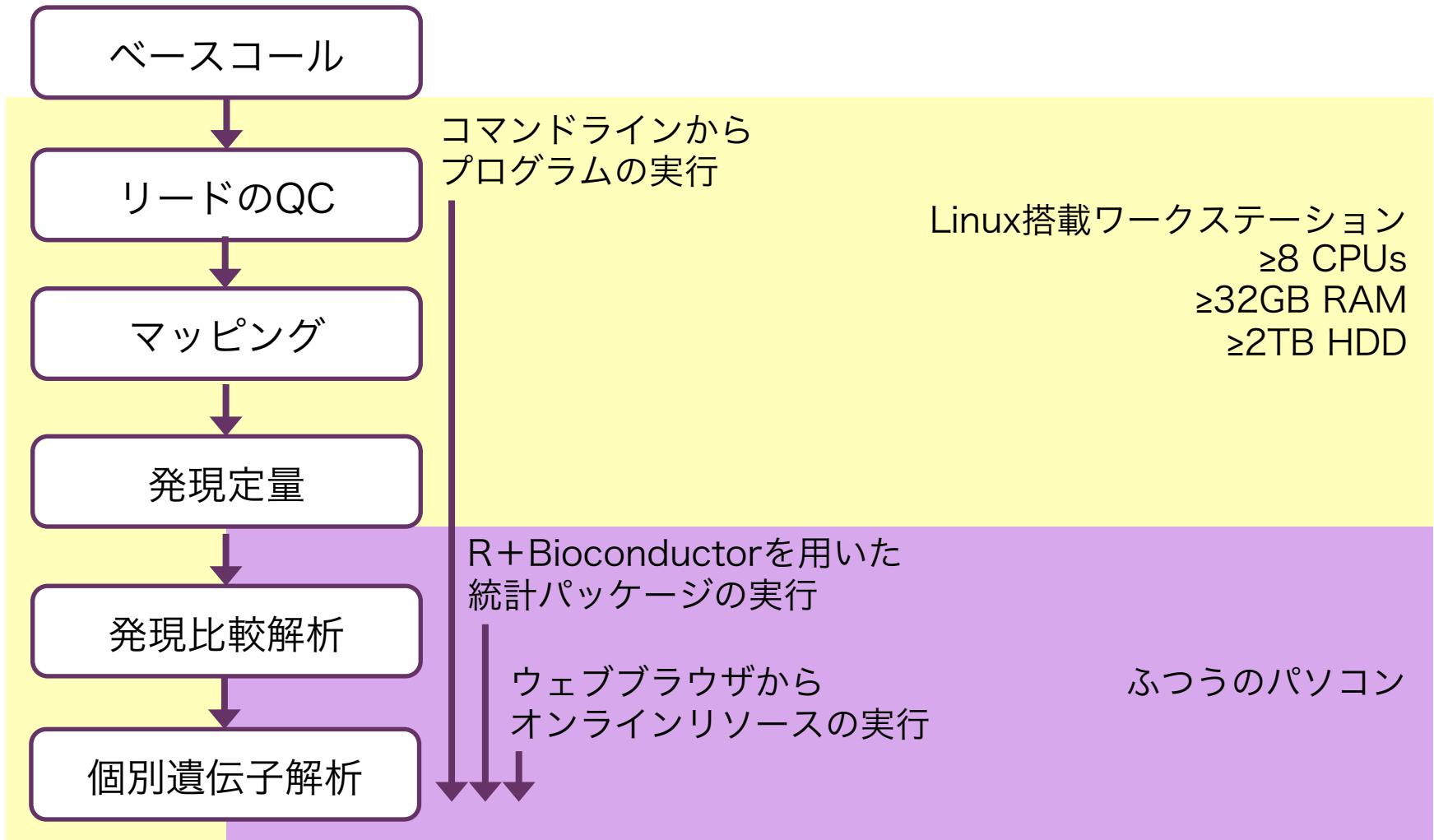
- 発現比較解析の一般的なフロー
- 全てのRNA-seqで同一の解析を行うわけではない
- 実験計画や産出されるデータによって解析を最適化する

# + データの容量(動物のRNA-seqの場合)



+

# 必要なハードウェア、スキル



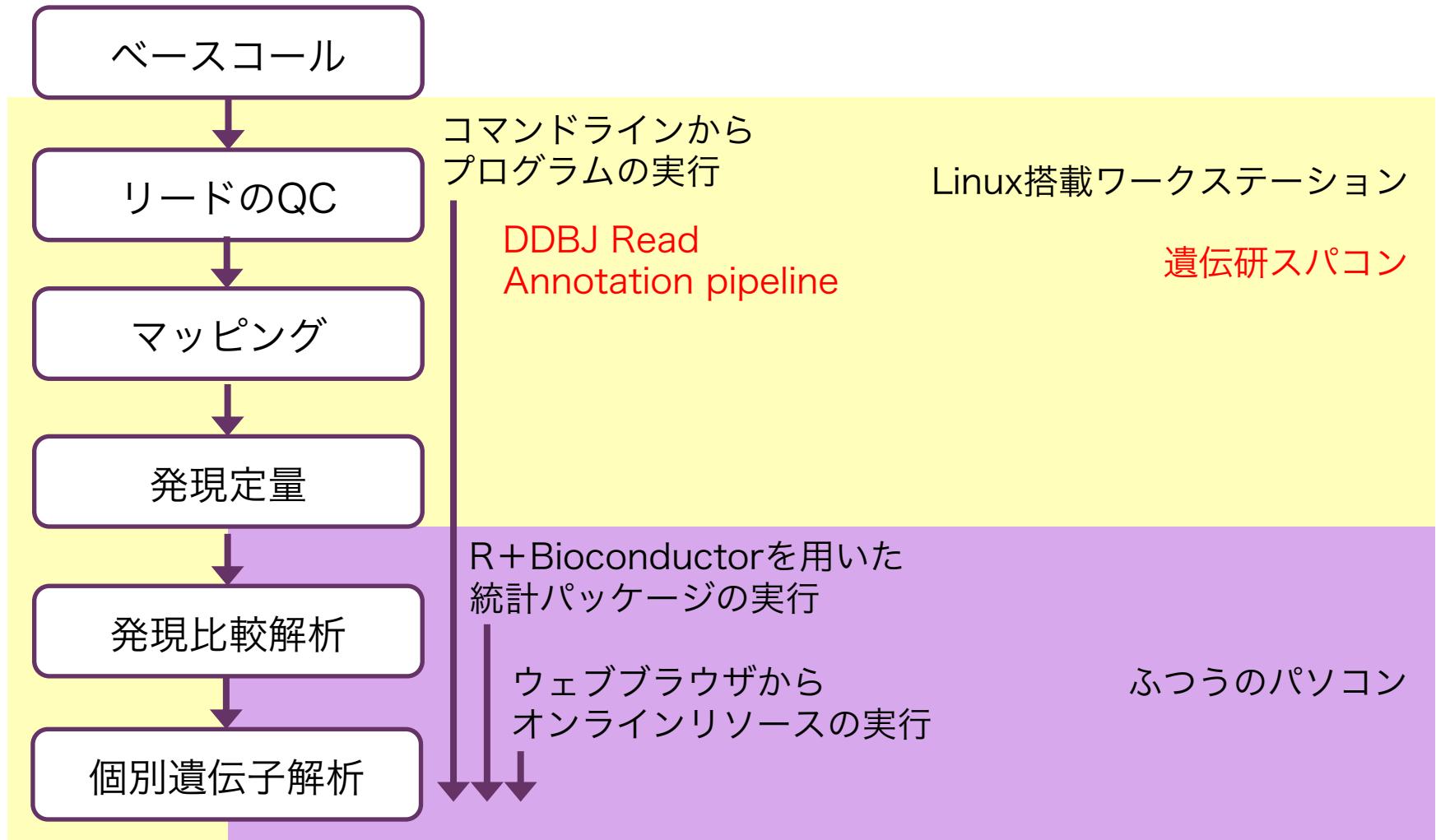
+

# 何が配列解析を困難にさせるのか？

- ワークステーションの導入、セットアップ、管理
  - ワークステーションの購入にはそれなりに費用がかかる
  - ハードウェアの故障やセキュリティへの対応にスキル要
- 解析プログラムのインストール、バージョン管理
  - 新規の解析プログラムが続々と発表されている
  - ソフトウェアのバージョンが頻繁にアップデートされる
- Linuxコマンドライン、R言語のスキル
  - R-studioの登場で若干親しみやすくなった

+

# 必要なハードウェア、スキル



# + 今日お話しすること

## 1. イントロダクション

- RNA-seq: 次世代シーケンシングデータから遺伝子発現を知る

## 2. RNA-seqの入口

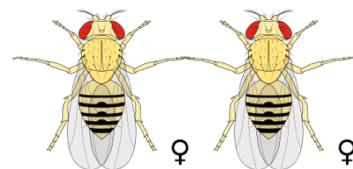
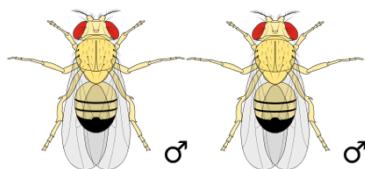
- 実験計画: RNA-seqを行う前に
- 発現解析の流れ

## 3. RNA-seqの出口

- クオリティーチェック(←実習)
- データの可視化(←実習)

# + テストケース:

- ショウジョウバエの脳において遺伝子発現に雌雄差はあるか？を知るためのRNA-seq
- 雌雄ごとに2つのreplicates
- 1ライブラリに成虫60個体分の脳から抽出したRNA
- Illumina HiSeq 2000を用いて51 nt single-endでシーケンス
- 配列データ(fastq), マッピングデータ(BAM), 発現定量と発現比較の表(タブ区切りテキスト)が納品された



Catalán et al. BMC Genomics 2012, 13:654  
<http://www.biomedcentral.com/1471-2164/13/654>

RESEARCH ARTICLE

Open Access

Population and sex differences in *Drosophila melanogaster* brain gene expression

Ana Catalán, Stephan Hutter and John Parsch\*



# + シーケンスデータのQC (クオリティチェック)

シーケンスリードのクオリティが低下する要因

- サンプルとは無関係な配列
  - アダプタ配列
  - PhiX
- クオリティーの低い塩基が含まれる配列
  - Quality value → 推定されるエラー率
    - $Q = -10\log(\text{エラー率})$
    - Q:10→20→30, エラー率:10%→1%→0.1%
- PCRによる配列の重複が多い
- ごく少ない種類の配列がデータの大半を占める

クオリティをプログラムでチェックする **FASTQC**



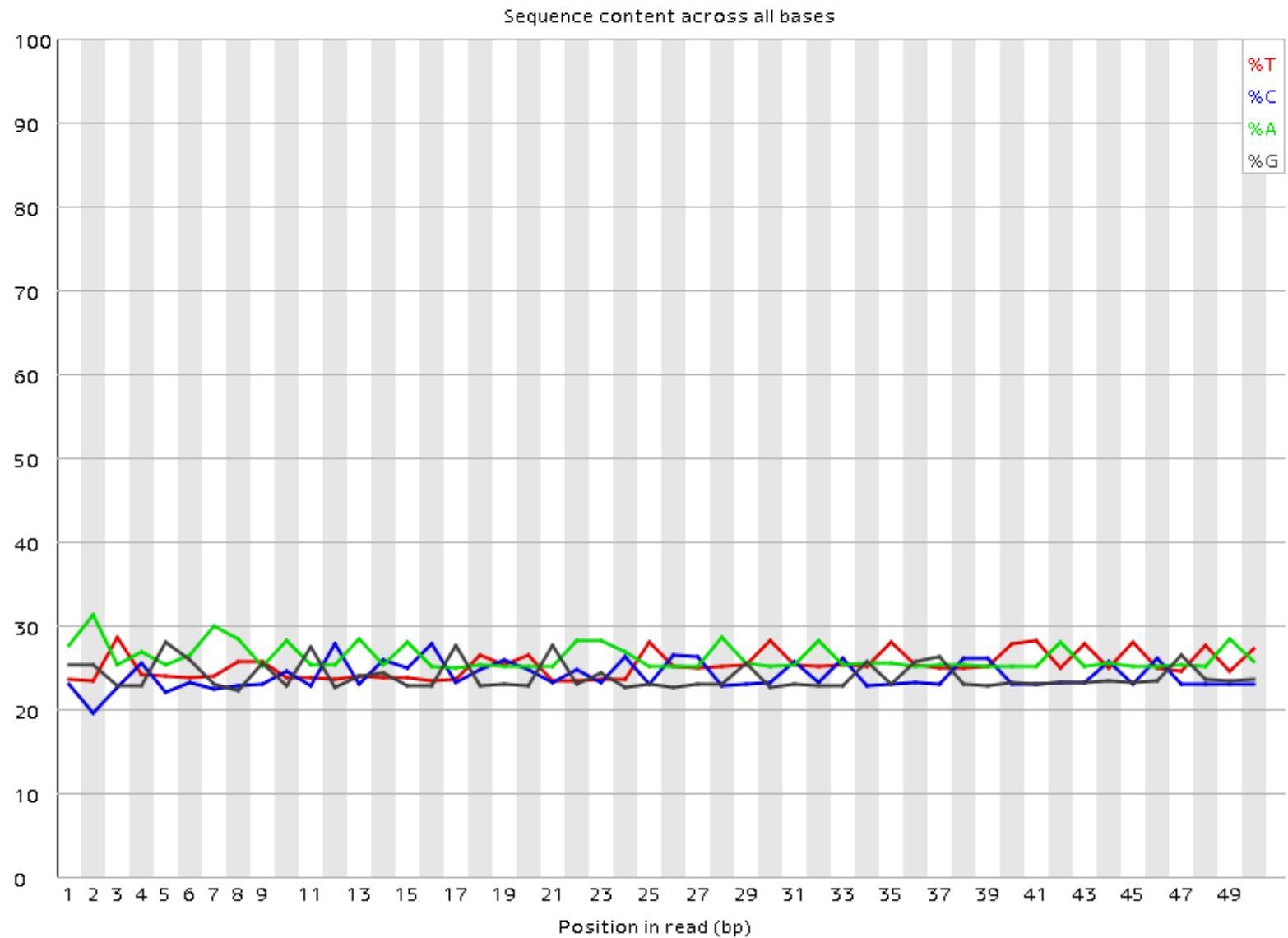
+

## 演習 1-1:

- 取得した配列のクオリティを見てみましょう
  - FASTQCプログラムを実行します
- 以下の点について観察しデータの質について考えてみましょう
  - 塩基ごとのクオリティの傾向は？
  - 塩基組成のばらつきは？
  - 重複した配列の頻度は？
  - アダプタ配列の混入は？

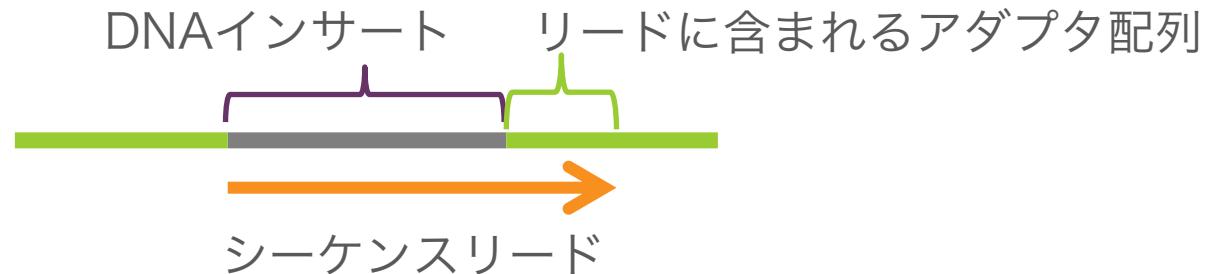


## Per base sequence content

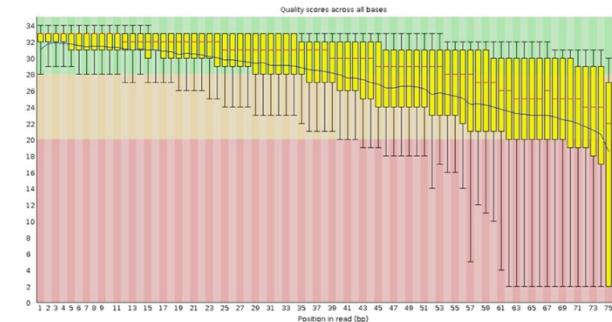


# + “使える”データを抽出する

## ■ アダプタ配列の除去



- クオリティスコアが低い塩基の除去
  - 3'端からクオリティの低い( $Q < 30$ )配列を削る
  - クオリティの低い( $Q < 30$ )塩基を一定の割合(20%)以上含む配列を除去する



## ■ クオリティーコントロールプログラム

Trim\_galore!, cutadapt, TagDust, FASTX\_Toolkit

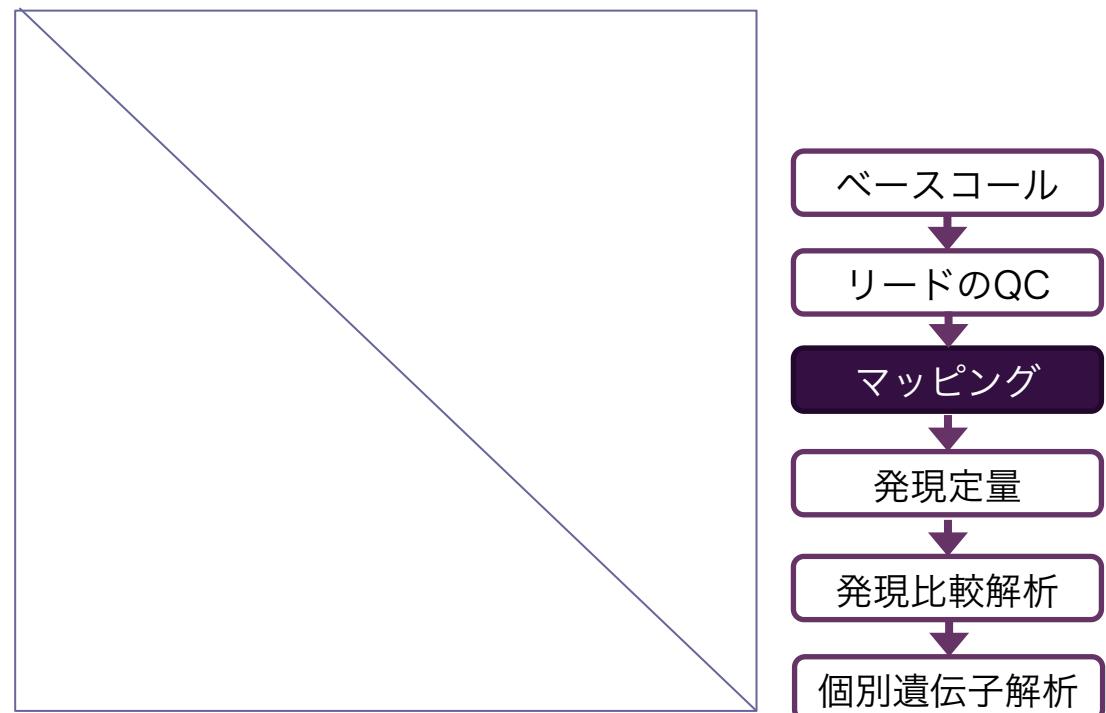
PRINSEQ

## <sup>+</sup> 演習1-2:

- アダプタやクオリティの低い配列を取り除いた配列のクオリティをチェックしてみましょう
- 配列のクオリティは向上していましたか？

# + マッピング

- シーケンスリードをリファレンス配列に貼り付ける
  - スプライシングを考慮してゲノム配列にマップする  
Tophat2, HISAT
  - 転写産物にマップする  
BWA, Bowtie2



# + Post-mapping QC

マッピングしてみないとわからないデータの質もある

- リファレンス配列にマップされるリードの割合は高いか？
  - コンタミネーションの可能性
- マップされるリードに偏りはないか？
  - インサートサイズ
  - Gene body(転写産物のどの領域にマップされるか)
- 使用するソフトウェア

RSeQC, RNA-SeQC,Picard tools

Qualimap

# マッピングプログラムもマップ率を出力してくれる

Left reads:

```
Input      : 30085625
Mapped     : 22261750 (74.0% of input)
of these:   311824 ( 1.4%) have multiple alignments (2449 have >20)
```

Right reads:

```
Input      : 30085625
Mapped     : 21357840 (71.0% of input)
of these:   297038 ( 1.4%) have multiple alignments (2357 have >20)
```

72.5% overall read mapping rate.

Aligned pairs: 20497588

```
of these:   282204 ( 1.4%) have multiple alignments
            89333 ( 0.4%) are discordant alignments
```

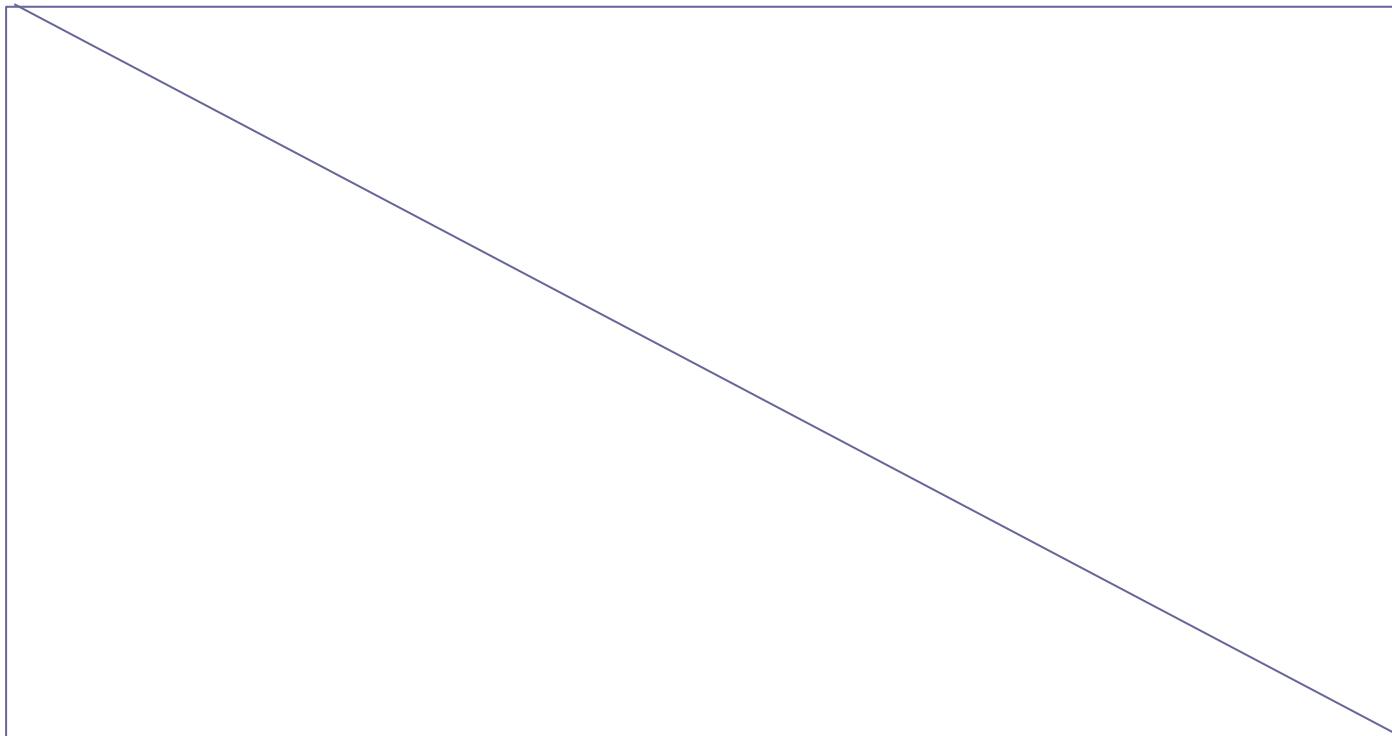
67.8% concordant pair alignment rate.

# + 演習2

- マッピングデータのクオリティを見てみましょう
- Qualimapを用います

### **RNA extraction and high-throughput sequencing**

Total RNA extraction and DNase I digestion were performed using the MasterPure RNA Purification Kit (Epicentre). cDNA library construction and high-throughput sequencing were performed by GATC Biotech (Konstanz, Germany). Briefly, poly-A mRNA was purified and fragmented by sonication. First-strand, single-end cDNA was synthesized using random primers. Eight tagged libraries were generated, pooled and run on two lanes of a HiSeq 2000 sequencer (Illumina) to generate single reads of 50 bases. All sequences have been submitted to the GEO database under the series GSE40907.



Sigurgeirsson et al., 2014. PLoS One.

File Tools Windows Help

BAM QC: saliva.sorted.bam

BAM QC: ERR089819.bam

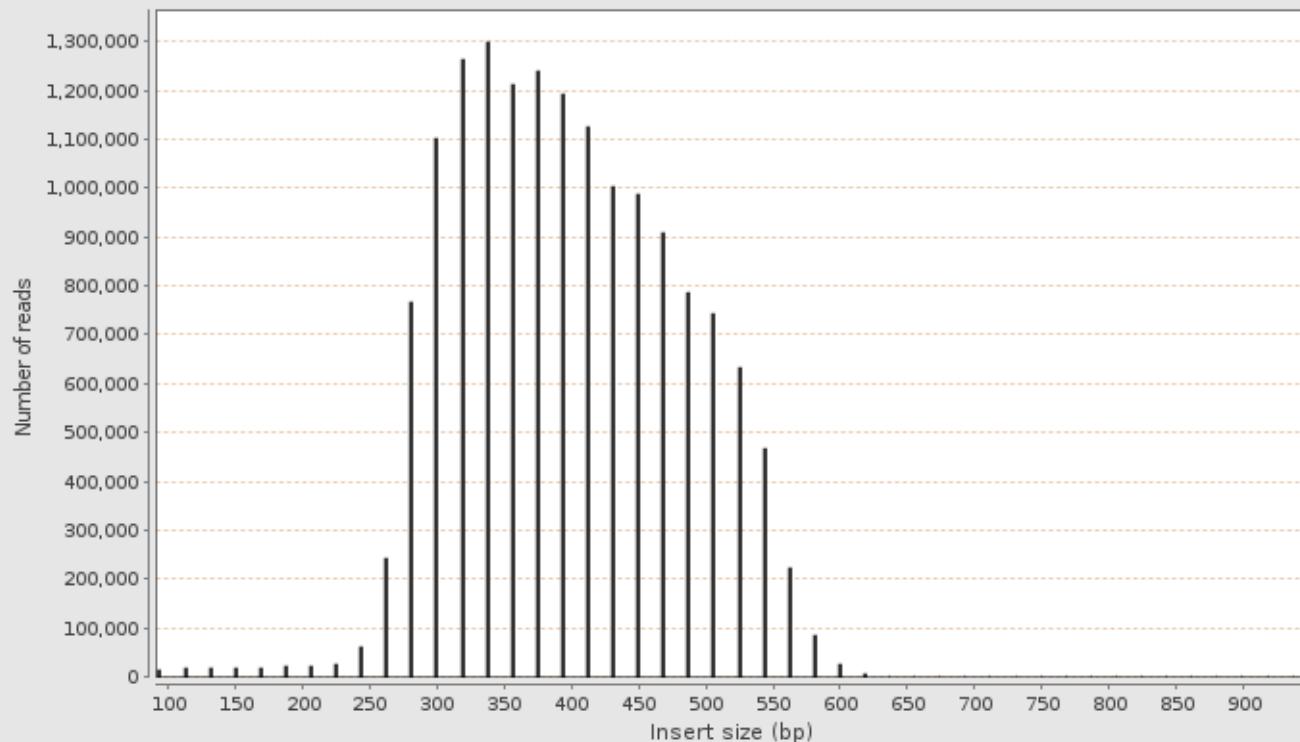
Counts QC: Data Analysis

Results

- Summary
- Input
- Coverage across reference
- Coverage Histogram
- Coverage Histogram (0-50X)
- Duplication Rate Histogram
- Genome Fraction Coverage
- Mapped Reads Nucleotide Content
- Mapped Reads GC-content Distribution
- Mapping Quality Across Reference
- Mapping Quality Histogram
- Insert Size Across Reference
- Insert Size Histogram

## Insert Size Histogram

ERR089819.bam



+

# 発現定量

- エキソン/転写産物マップされたリードの数→発現量
  - マップされるリードが多いほど発現量は高い
- 遺伝子(転写産物)の情報
  - 既知の遺伝子モデル
  - マッピングデータから遺伝子構造を推定する
- 遺伝子レベルか、isoform(転写産物)単位か
- リード数や遺伝子長による正規化
- 発現定量プログラム: マッピングプログラムとの相性

Tophat2, HISAT

Bowtie2

BWA

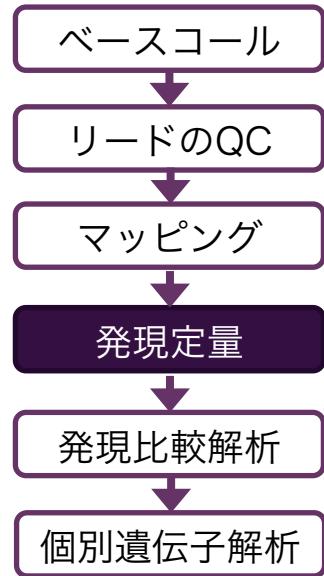
マッピングなし

Cufflinks, Stringtie

RSEM

Express

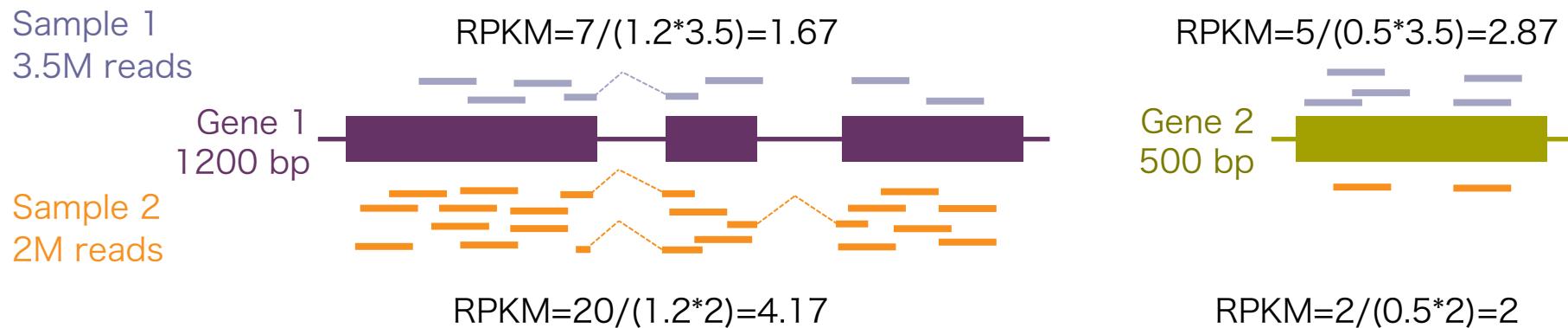
Sailfish, RNA-skim



+

# 発現定量値

- マップされたリードの数(raw count)は、リードの総数、転写産物の長さに応じて正規化される
  - CPM: (Counts per million)
  - RPKM/FPKM: Reads/Fragments per kilobase of exon per million mapped sequence reads)
  - TPM: (Transcripts per million)



+

# 発現定量値の表(の例)

- Raw countと正規化された定量値の両方が出力される
- 遺伝子単位と転写産物単位の定量が行われる

Gene\_count.txt

	A	B	C	D	E	F	G	H
1	track_id	Female_1	Female_2	Male_1	Male_2	SeqAcc	Gene_Sym	Gene_ID
2	FBgn0000008	517.998	531	1043	1105	FBtr0071764 FBtr0100521 FBtr0071763	a	FBgn0000008
3	FBgn0000014	4	1	1	3	FBtr0083388 FBtr0083387 FBtr0300485	abd-A	FBgn0000014
4	FBgn0000015	0	1	1	2	FBtr0083381 FBtr0083382 FBtr0083383	Abd-B	FBgn0000015
5	FBgn0000017	2872	3625	6971	8276	FBtr0075357 FBtr0112790 FBtr0330130	Abl	FBgn0000017
6	FBgn0000018	120	70	156	193	FBtr0080168	abo	FBgn0000018

Gene\_rpkm.txt

	A	B	C	D	E	F	G	H
1	track_id	Female_1	Female_2	Male_1	Male_2	SeqAcc	Gene_Sym	Gene_ID
2	FBgn0000008	7.62773	8.60866	8.23751	8.19216	FBtr0071764 FBtr0100521 FBtr0071763	a	FBgn0000008
3	FBgn0000014	0.0705568	0.0194496	0.00950885	0.0266906	FBtr0083388 FBtr0083387 FBtr0300485	abd-A	FBgn0000014
4	FBgn0000015	0	0.0159892	0.00781747	0.0284491	FBtr0083381 FBtr0083382 FBtr0083383	Abd-B	FBgn0000015
5	FBgn0000017	15.42	21.4842	20.3823	22.8015	FBtr0075357 FBtr0112790 FBtr0330130	Abl	FBgn0000017
6	FBgn0000018	4.96409	3.19301	3.4791	4.027	FBtr0080168	abo	FBgn0000018

# + R(F)PKM or TPM?

リード数の正規化

$$\text{RPKM}_g = \frac{r_g \times 10^9}{\text{fl}_g \times R}$$

$$\text{TPM} = \frac{r_g \times \text{rl} \times 10^6}{\text{fl}_g \times T}$$

転写産物数の正規化

$$T = \sum_{g \in G} \frac{r_g \times \text{rl}}{\text{fl}_g}$$

$r_g$ : 遺伝子gのリード数

$\text{fl}_g$ : 遺伝子gの長さ

R: 総リード数

rl: シーケンスリード長

T: シーケンスランの全転写産物数

総リード数Rに依存するRPKMの総和とは異なり、TPMの総和は同じ生物のあらゆるサンプルで不变

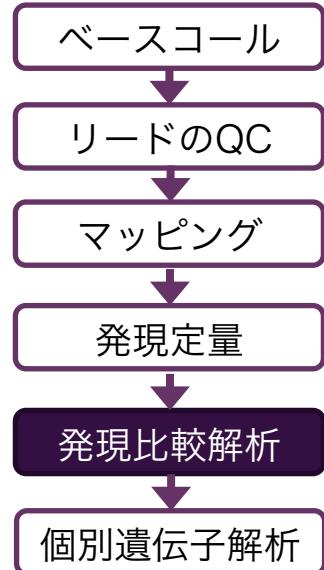
RNA-seq解析プログラムのアルゴリズム開発者達はR(F)PKMの使用をやめてTPMを使おうと主張している

+

# 発現比較解析

- 雄と雌の2群間で発現量に差が無いか統計検定を行う
- 遺伝子ごとに検定を行う(→多重検定の補正)
- 発現量に有意差がある遺伝子を「雌雄で発現差がある遺伝子」とする
- 発現比較解析のプログラム
  - よく使われているプログラムはRに実装されている
  - Raw readsが用いるデータ

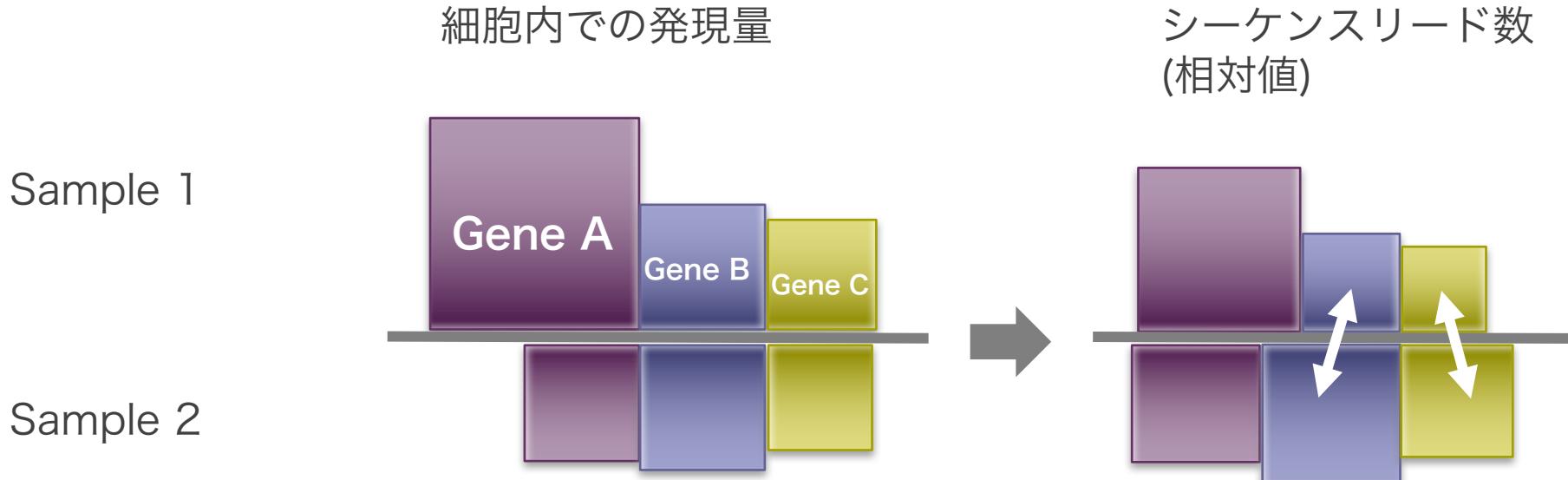
DESeq2, edgeR in R+Bioconductor



+

# 発現比較解析

- 発現遺伝子の構成により正規化が必要



- 発現比較解析パッケージに正規化プログラムも含まれている

+

# 発現比較解析の結果(の例)

- diffexpr\_edgeR\_Europe.txt: edgeRによる発現比較解析の結果

遺伝子(もしくは  
転写産物)のID

P値

多重検定の補正を  
受けた統計値

二群の発現量  
の比( $\log_2$ )

二群の発現量  
の平均( $\log_2$ )

各群の発現量  
の平均

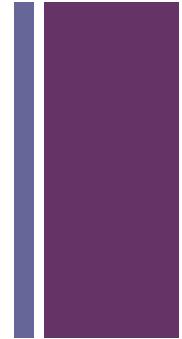
	A	B	C	D	E	F	G	H
1	track_id	P-value	Q-value	logFC	logCPM	Maximum fold-change	Mean CPM: Female	Mean CPM: Male
2	FBgn0019660	1.67E-192	2.75E-188	7.8935	6.7241	241.5522	0.8613	208.0573
3	FBgn0019661	1.14E-140	9.39E-137	8.0124	10.1075	258.3368	8.5029	2196.6075
4	FBgn0022702	2.25E-54	1.23E-50	-3.5618	4.6573	11.7981	47.9201	4.0617
5	FBgn0037236	1.53E-34	6.29E-31	-2.6104	4.5491	6.1053	41.3838	6.7783
6	FBgn0085249	1.25E-33	4.10E-30	-2.4733	5.3653	5.5554	71.0853	12.7958
7	FBgn0085353	1.24E-32	3.39E-29	-2.7307	4.4593	6.6446	39.3279	5.9187

+

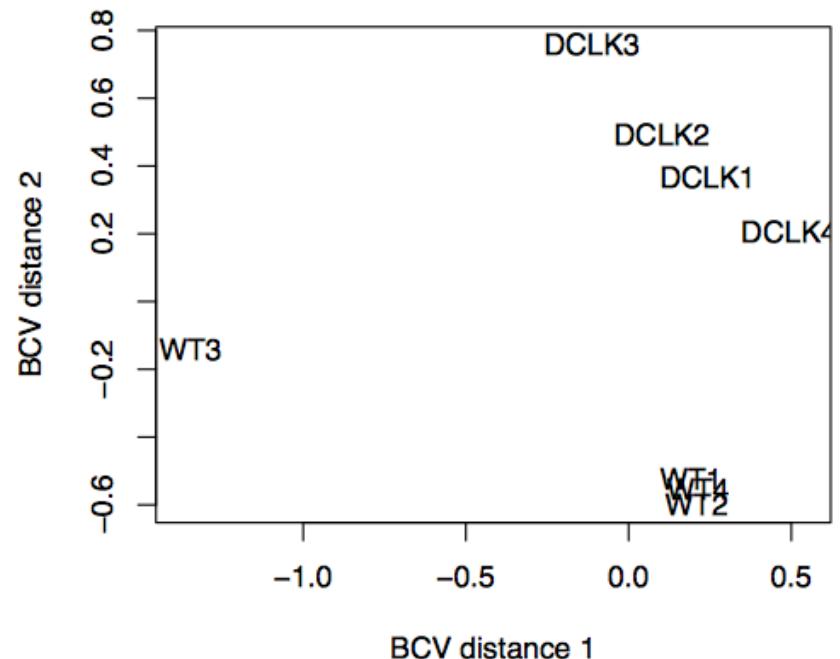
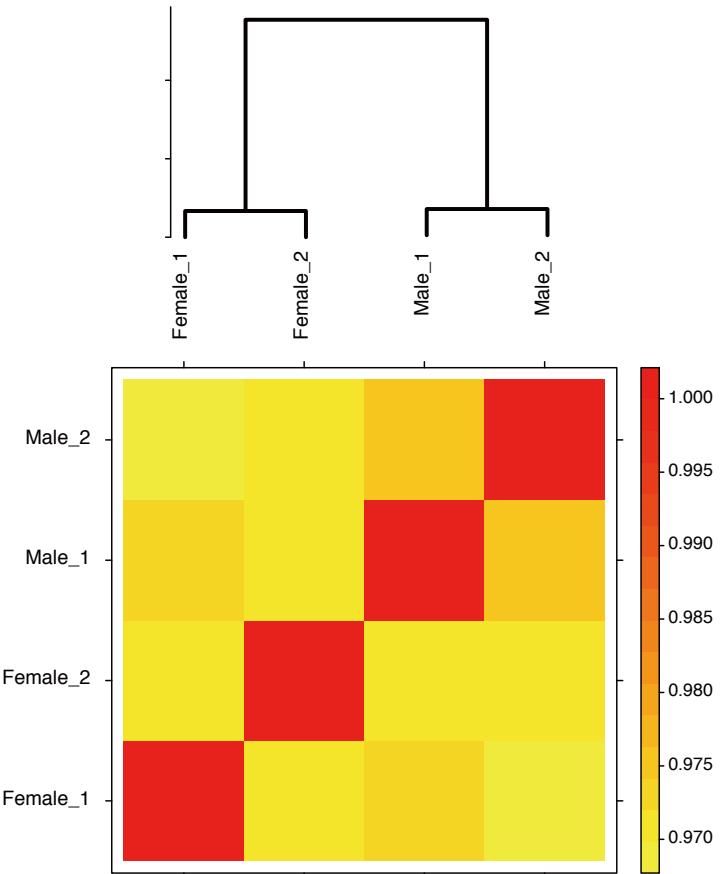
# 多重検定の補正

- 帰無仮説: 「群間で発現量に差が無い」
- 対立仮説(帰無仮説を棄却): 「群間で発現量に差がある」
- 有意水準→帰無仮説が正しいにもかかわらず棄却してしまう確率
- 同種の検定を繰り返すとこの確率は有意水準を大きく上回ってしまう
  - 有意水準5%の検定を10000回繰り返すとき、帰無仮説が実際には正しいのに棄却してしまう検定が500回起きうる
  - 検定の回数を増やしていくと、検定で棄却されるうち本当は帰無仮説が正しいケースが大部分になってしまう
- False Discovery Rate (FDR): 繰り返される検定で棄却されるケースのうち本当は帰無仮説が正しい確率

# + 発現データを可視化する (R+Bioconductor)

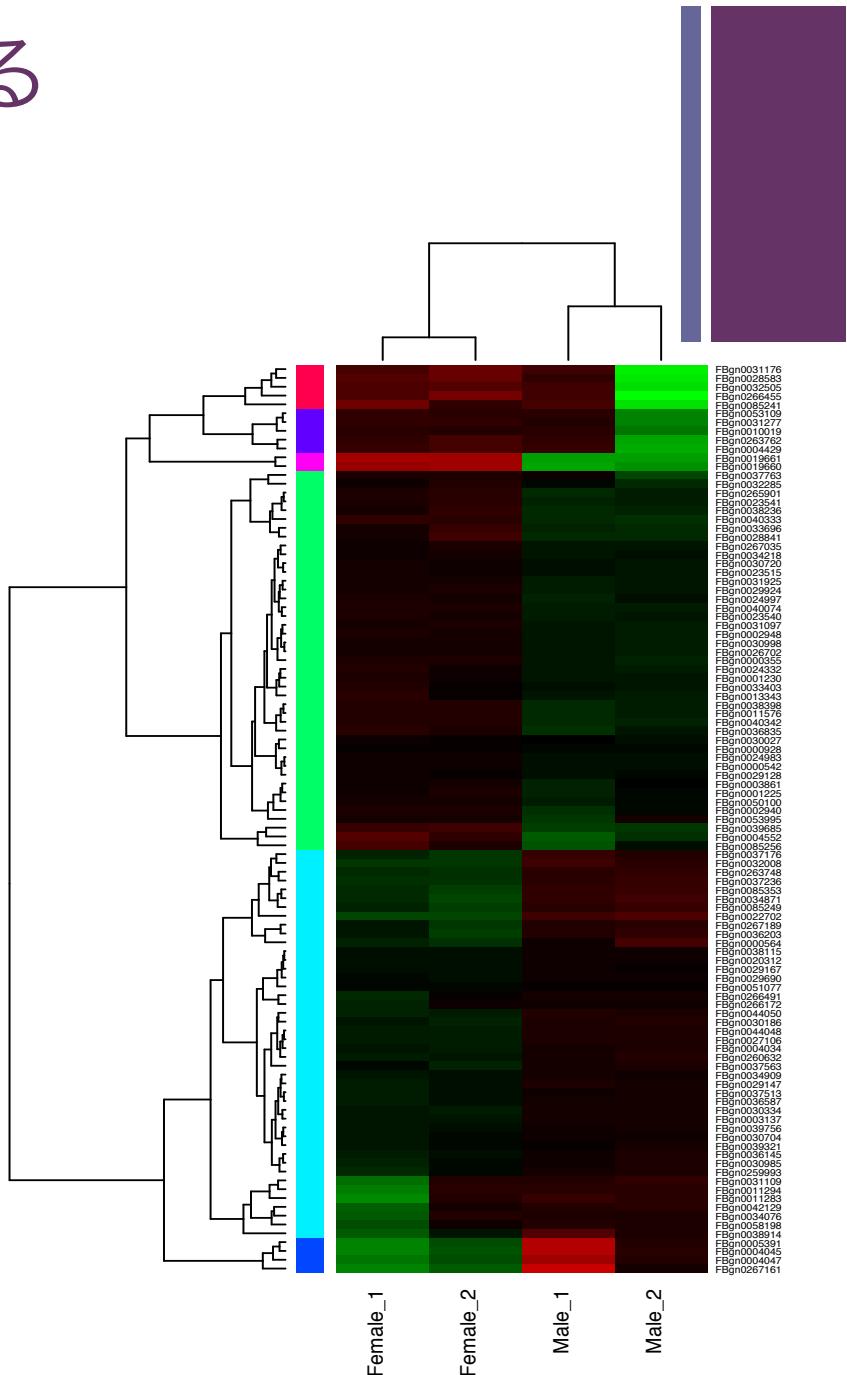
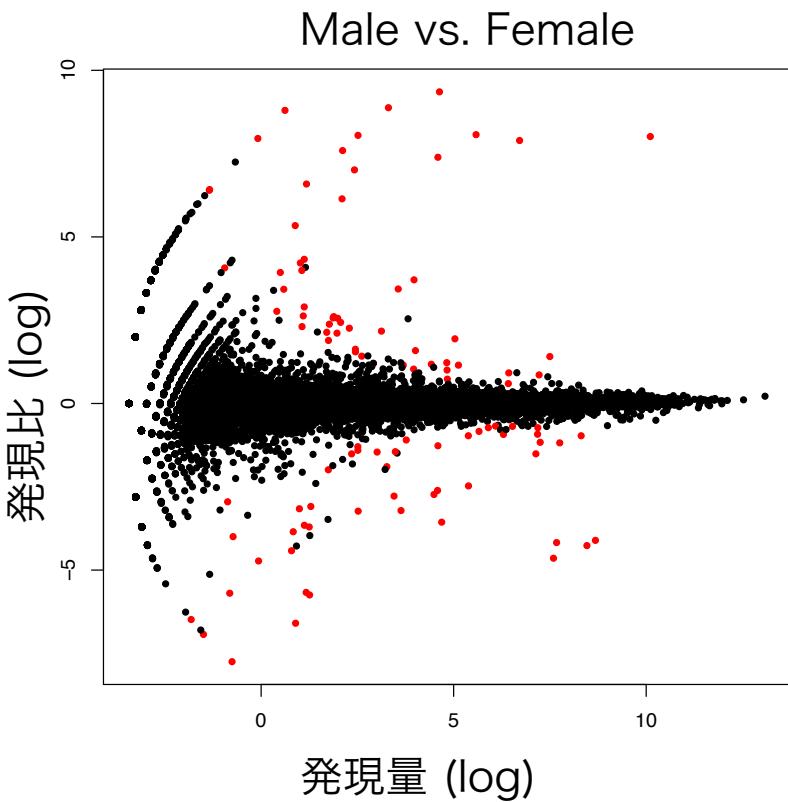


- 遺伝子プロファイルに基づくサンプルのクラスタリング



# + 発現データを可視化する (R+Bioconductor)

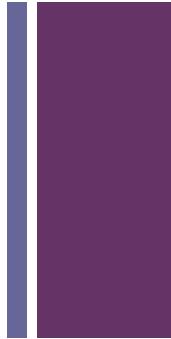
- 群間で発現に差がある遺伝子



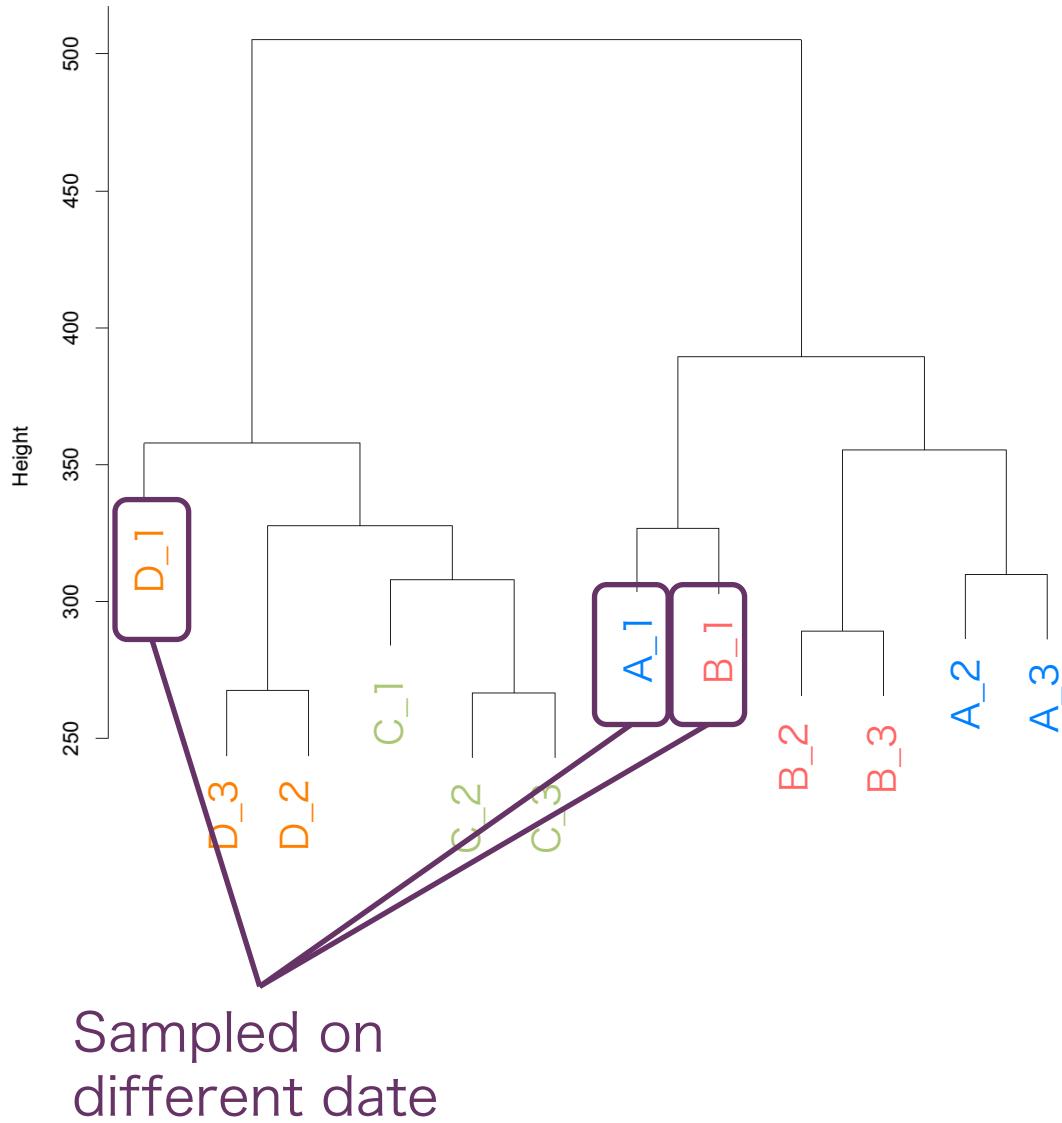
# + 演習3

- 発現プロファイルが雌雄のBiological replicatesごとに明確に分かれるか確認しましょう
- Rstudioを用いてRPKM値の類似度からデンドログラムを作ります

+



Sample clustering

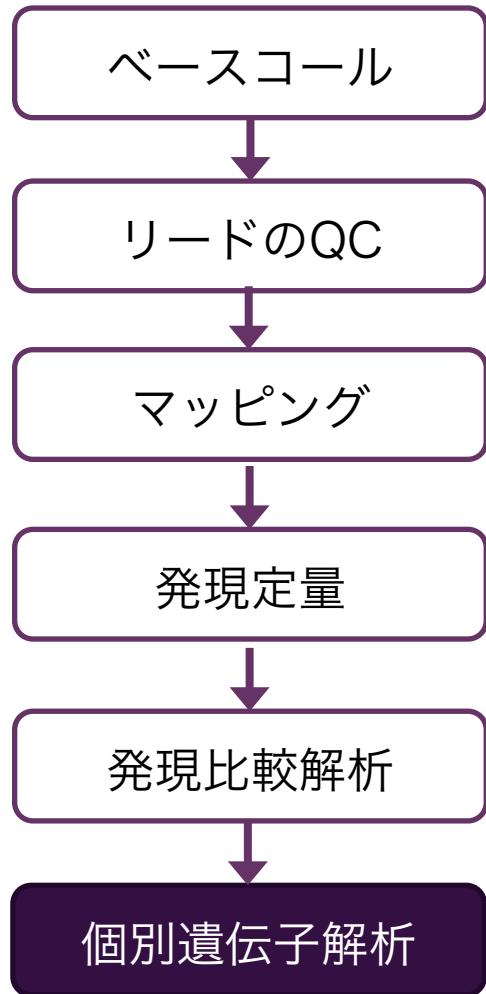


# + 演習4

- 発現比較解析の結果を可視化しましょう
  - Rstudioを用いてM-A plot, Volcano plotを作成します。
  - 異なる有意差の閾値を設定してプロットを比較してみましょう

+

# 発現差がある遺伝子から何を得るか



- 個別の遺伝子の機能解析
  - 発現同定
  - ノックアウト・ノックダウン
- 既存の知識をもとに、発現差がある遺伝子「群」としての特徴を発見
  - Gene Ontologyエンリッチメント解析
  - パスウェイ解析

# + 演習5

- IGV (Interactive Genomics Viewer): 遺伝子1つ1つの発現を見てみましょう
  - ハエのRNA-seqのBAMファイルをロードしましょう
  - オス、メスともに発現量が高い遺伝子を見てみましょう
  - オス、メスで発現差がある遺伝子を見てみましょう

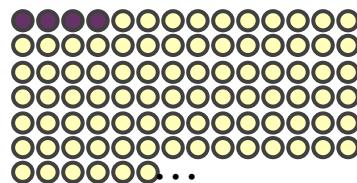
# + エンリッチメント解析

- ある遺伝子セット(e.g. 発現変動遺伝子)に統計学的に有意に特徴付けられる機能関連情報を同定

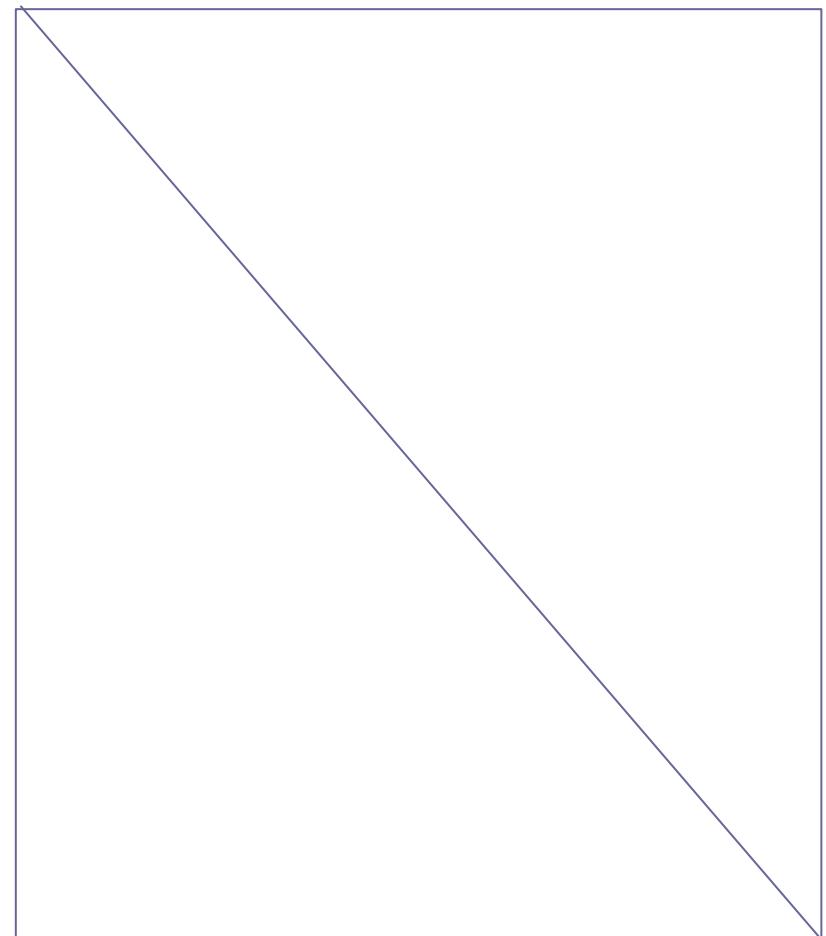
遺伝子セット



残りの遺伝子



- 機能関連情報(アノテーション)
  - Gene Ontology (GO)
  - KEGG
  - Reactome



# David

(<https://david.ncifcrf.gov>)

- 最も使用されている  
エンリッチメント解析  
ウェブツール

- 現在のバージョンは  
6.8 (2016. Oct)。  
1つ前(version 6.7)は  
7年前リリース

DAVID Bioinformatics Resources 6.8  
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About Us

\*\*\* Welcome to DAVID 6.8 with updated Knowledgebase ([more info](#)). \*\*\*  
\*\*\* If you are looking for DAVID 6.7, please visit our [development site](#). \*\*\*

Recommending: A paper published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.8

2003 - 2016

What's Important in DAVID?

- New requirement to cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID

DAVID Bioinformatic Resources Citations

in	Sublist	Category	Term	RT	Genes	Count	%	Value
ssoc	GOTERM_BP_ALL	response to chemical stimulus	RT	14	8.2%	6.1E-5		
nctc	GOTERM_BP_ALL	response to abiotic stimulus	RT	15	8.8%	6.5E-5		
itcra	GOTERM_MP_ALL	protein binding	RT	55	32.2%	8.0E-5		
ifiers	GOTERM_BP_ALL	response to bacteria	RT	7	4.1%	1.7E-4		
	GOTERM_MP_ALL	iron ion binding	RT	10	5.8%	2.6E-4		
	GOTERM_BP_ALL	cell-cell signaling	RT	15	8.8%	4.0E-4		
	GOTERM_BP_ALL	defense response to bacteria	RT	6	3.5%	5.4E-4		
	GOTERM_BP_ALL	regulation of hydrolase activity	RT	6	3.5%	6.1E-4		
	GOTERM_BP_ALL	regulation of GTPase activity	RT	5	2.9%	9.8E-4		
	GOTERM_BP_ALL	response to stress	RT	22	12.9%	9.9E-4		
	GOTERM_BP_ALL	response to other organism	RT	15	8.8%	1.2E-3		
	GOTERM_MP_ALL	heme binding	RT	6	3.5%	1.3E-3		
	GOTERM_BP_ALL	tetrapyrrole binding	RT	6	3.5%	1.3E-3		
	GOTERM_MP_ALL	response to stimulus	RT	40	23.4%	1.4E-3		
	GOTERM_MF_ALL	receptor binding	RT	14	8.2%	1.6E-3		
	GOTERM_BP_ALL	response to pest, pathogen or parasite	RT	14	8.2%	2.0E-3		
	GOTERM_BP_ALL	behavior	RT	8	4.7%	2.2E-3		
	GOTERM_BP_ALL	defense response	RT	23	13.5%	2.2E-3		
	GOTERM_MP_ALL	oxygen binding	RT	4	2.3%	2.7E-3		
	GOTERM_BP_ALL	inflammatory response	RT	8	4.7%	3.3E-3		
	GOTERM_MF_ALL	sodium ion binding	RT	5	2.9%	3.6E-3		
	GOTERM_BP_ALL	response to biotic stimulus	RT	23	13.5%	3.8E-3		
	GOTERM_MP_ALL	carbohydrate binding	RT	8	4.7%	3.9E-3		
	GOTERM_BP_ALL	sodium ion transport	RT	6	3.5%	4.0E-3		

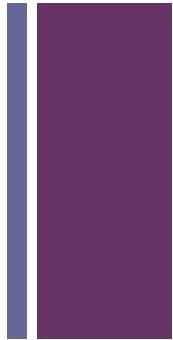
Main Annotation Sources

Data Sources	Release / Download Date	DAVID Update Date
ENSEMBL	Mar 2016	May 2016
ENTREZ	May 2016	May 2016
UNIPROT	May 2016	May 2016

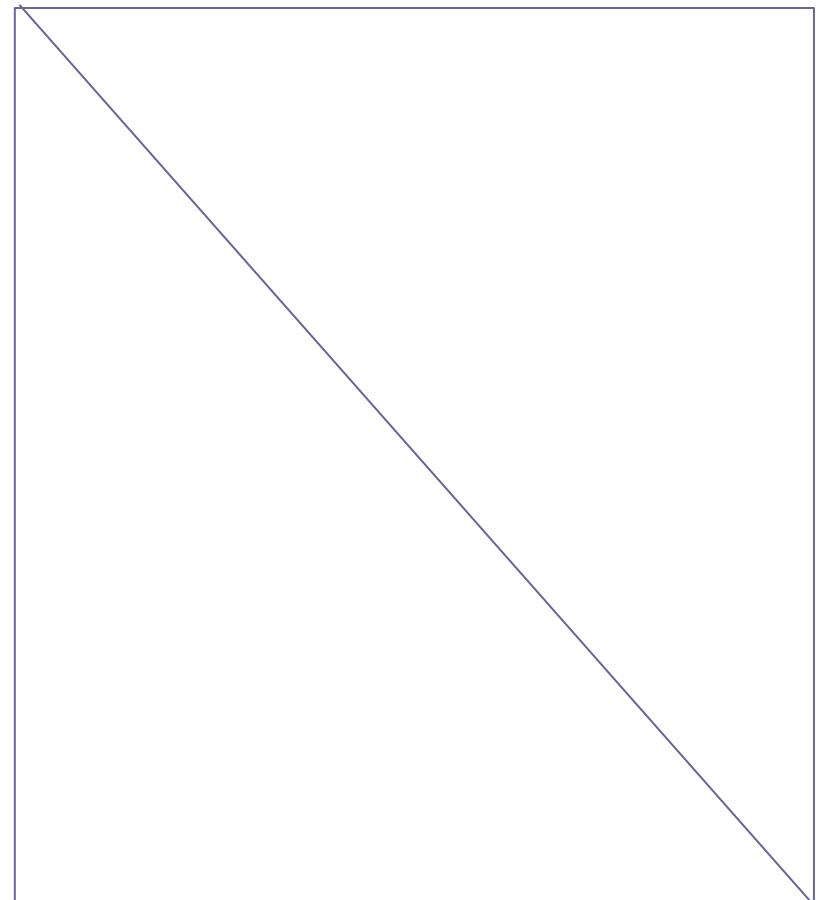
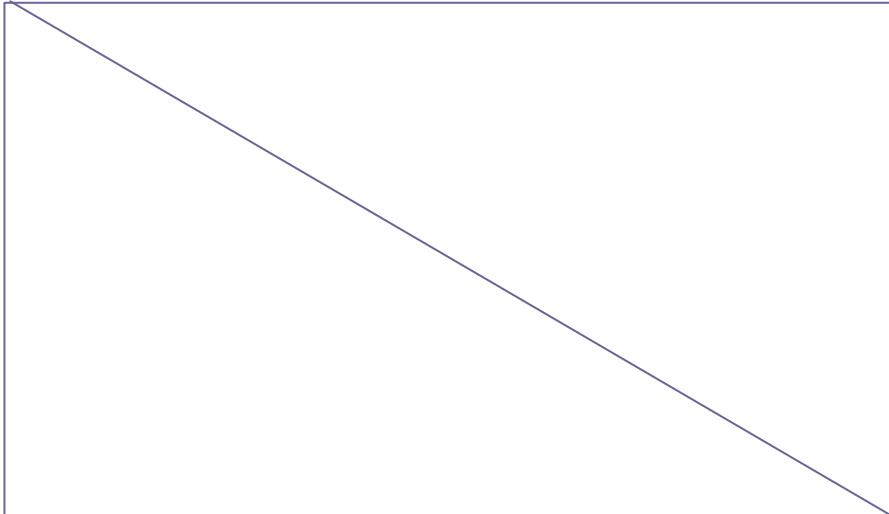
Secondary Sources

Data Sources	Release / Download Date	DAVID Update Date
AFFYMETRIX	Jun 2015	May 2016
AGILENT	Dec 2013	May 2016
BBID	Sep 2009	May 2016
BIOCARTA	Nov 2014	May 2016
CGAP_EST_QUARTILE	Oct 2006	May 2016
CGAP_SAGE_QUARTILE	Oct 2006	May 2016
COG_ONTOLOGY	Sep 2009	May 2016
GENE_ONTOLOGY	Apr 2016	May 2016
GNF_U133A_QUARTILE	Oct 2006	May 2016
KEGG	Dec 2015	May 2016
UCSC_TFBBS	Sep 2009	May 2016
UP_SEQ_FEATURE	Sep 2009	May 2016
UP_TISSUE	Sep 2009	May 2016
ZFIN_ANATOMY	Sep 2009	May 2016

+



- 古いアノテーション情報を用いたエンリッチメント解析では、結果の感度が低くなってしまう
- ウェブツールのリリース日に注意すること



# + 演習6

- Enrichment analysisを行ってみましょう
  - オスで発現量が高い遺伝子にエンリッチされるGO, KEGGパスウェイを探しましょう
  - 同じくメスで発現量が高い遺伝子にエンリッチされるGO, KEGGパスウェイを探しましょう
  - Davidを使います

# + g:profiler (<http://biit.cs.ut.ee/gprofiler/>)

### ■ 更新が頻繁

- Ensemblリリースごとにアーカイブを作成



**Welcome!** [Contact](#) [FAQ](#) [R / APIs](#) [Beta](#) [Archive](#)

**g:GOST Gene Group Functional Profiling**

**g:Cocoa Compact Compare of Annotations**

**g:Convert Gene ID Converter**

**g:Sorter Expression Similarity Search**

**g:Orth Orthology search**

**g:SNPense Convert rsID**

J. Reimand, T. Arak, P. Adler, L. Kolberg, S. Reisberg, H. Peterson, J. Viljo: g:Profiler --- a web server for functional interpretation of gene lists (2016 update) Nucleic Acids Research 2016; doi: 10.1093/nar/gkw199

**[?]** **Organism**  
Homo sapiens

**[?]** **Query (genes, proteins, probes)**

**Options**

**[?]** Significant only  
**[?]** Ordered query  
**[?]** No electronic GO annotations  
**[?]** Chromosomal regions  
**[?]** Hierarchical sorting  
**[?]** Hierarchical filtering  
[Show all terms \(no filtering\)](#)  
**[?]** Output type  
**Graphical (PNG)**  
[Show advanced options](#)

**[?]** **Gene Ontology** **Biological process** **Cellular component** **Molecular function**  
Inferred from experiment [IDA, IPI, IMP, IGI, IEP]  
Direct assay [IDA] / Mutant phenotype [IMP]  
Genetic interaction [IGI] / Physical interaction [IPI]  
Traceable author [TAS] / Non-traceable author [NAS] / Inferred by curator [IC]  
Expression pattern [IEP] / Sequence or structural similarity [ISS] / Genomic context [IGC]  
Biological aspect of ancestor [IBA] / Rapid divergence [IRD]  
Reviewed computational analysis [RCA] / Electronic annotation [IEA]  
No biological data [ND] / Not annotated or not in background [NA]  
Biological pathways   
Regulatory motifs in DNA   
Protein databases   
Human Phenotype Ontology (sequence homologs in other species)   
Online Mendelian Inheritance in Man

**[?]** **or Term ID:**

g:Profiler! Clear

Example or random query

g:Profiler version r1709\_e87\_eg34. Version info

**Archives**

**g:Profiler Archives** stores all the past stable versions of g:Profiler, including Ensembl and Ensembl Genomes versions. This allows for the reproducibility of analyses that have been retired since running an analysis. The following archived g:Profiler instances are available:

- Ensembl **62**, Ensembl Genomes **9** (rev 0998, build date 2011-05-07)
  - Ensembl **64**, Ensembl Genomes **11** (rev 1070, build date 2012-02-14)
  - Ensembl **65**, Ensembl Genomes **12** (rev 1075, build date 2012-02-25)
  - Ensembl **68**, Ensembl Genomes **15** (rev 1137, build date 2012-08-09)
  - Ensembl **68**, Ensembl Genomes **15** (rev 1177, build date 2012-08-09)
  - Ensembl **69**, Ensembl Genomes **16** (rev 1185, build date 2012-11-09)
  - Ensembl **72**, Ensembl Genomes **19** (rev 1227, build date 2013-07-18)
  - Ensembl **75**, Ensembl Genomes **22** (rev 1270, build date 2014-04-11)
  - Ensembl **78**, Ensembl Genomes **25** (rev 1353, build date 2015-02-06)
  - Ensembl **79**, Ensembl Genomes **26** (rev 1395, build date 2015-04-09)
  - Ensembl **80**, Ensembl Genomes **27** (rev 1435, build date 2015-07-07)
  - Ensembl **81**, Ensembl Genomes **28** (rev 1440, build date 2015-09-08)
  - Ensembl **82**, Ensembl Genomes **29** (rev 1477, build date 2015-10-20)
  - Ensembl **83**, Ensembl Genomes **30** (rev 1488, build date 2015-12-21)
  - Ensembl **83**, Ensembl Genomes **30** (rev 1507, build date 2015-12-21)
  - Ensembl **83**, Ensembl Genomes **30** (rev 1536, build date 2016-02-02)
  - Ensembl **84**, Ensembl Genomes **31** (rev 1615, build date 2016-05-06)
  - Ensembl **84**, Ensembl Genomes **31** (rev 1622, build date 2016-05-06)
  - Ensembl **85**, Ensembl Genomes **32** (rev 1665, build date 2016-09-05)
  - Ensembl **86**, Ensembl Genomes **33** (rev 1705, build date 2016-11-02)
  - Ensembl **87**, Ensembl Genomes **34** (rev 1709, build date 2016-12-13)

+

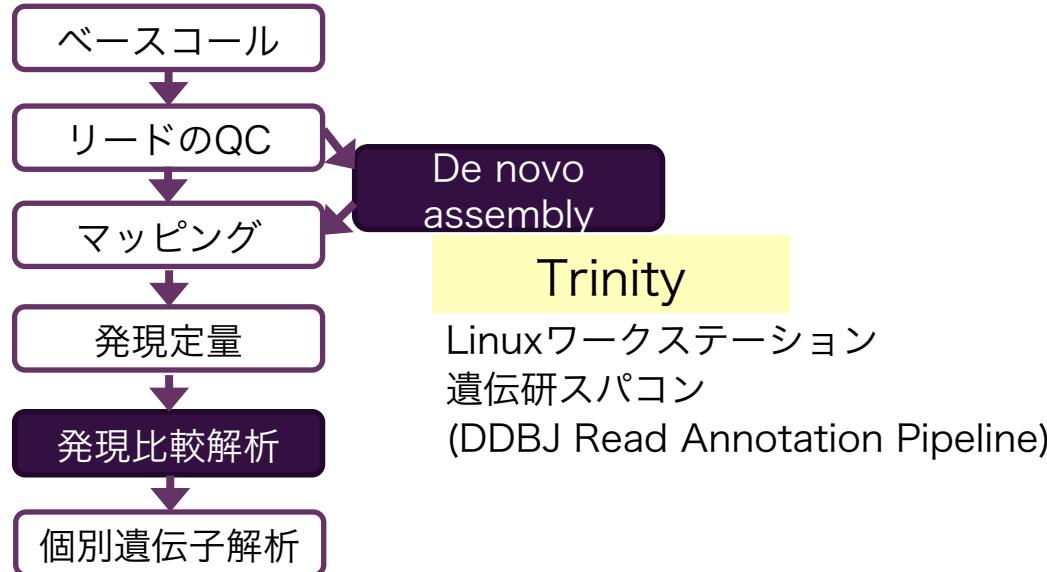
## リファレンスゲノムが無い生物の 発現比較解析

- ゲノムを読む
  - ゲノムアセンブリ
  - 遺伝子モデルの構築
  - 全ての遺伝子を予測できるが、費用対効果が悪い
- トランスクリプトームアセンブリ
  - 発現比較解析のリードを用いてアセンブリできる
  - 費用対効果が高い
  - 発現する遺伝子のみ得られる

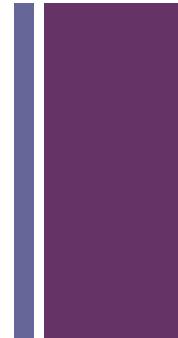
+

# *De novo transcriptome assembly*

- リファレンスゲノムがシーケンスされていない種では、RNA-seqのシーケンスリードをそのままアセンブルしてトランскルiptオーム配列を再構築する
  - 発現比較解析のためのリファレンス
  - 非モデル生物のシーケンスリソース



+ Genome/transcriptome sequencing  
in Kobe



+

## *De novo*トランスクリプトームアセンブリには どれだけシーケンスすれば十分？

- シーケンス量が少なすぎる→低発現遺伝子配列を復元できない
- シーケンス量が多すぎる→コスト増、よくわからない配列 (cryptic unstable transcripts)が増える
- ゲノム・トランスクリプトームアセンブリの網羅性を測る指標
  - 配列の長さ (N50)
  - アセンブリの中に復元される遺伝子数  
→CEGMAパイプラインを用いて、復元されたリファレンス遺伝子の数から評価する

# + gVolante: アセンブリの完全度を評価するウェブツール

- 真核生物に広く保存する遺伝子群(CEG, Core Eukaryote Genes)をリファレンスとすることが多い
- 系統特異的なゲノム進化に対応させるには特定の系統に着目したリファレンス遺伝子が効果的
- CVG (Core Vertebrate Genes)  
脊椎動物に1コピーしか存在しない遺伝子群のセット  
ヤモリトランскриプトームを用いた網羅性評価では、CEGより高い正確性と解像度を示した



gVolante - Completeness / <https://gvolante.riken.jp/index.html>

これは 英語 のページです。翻訳しますか？ いいえ 翻訳

HOME ANALYSIS TUTORIAL YOUR RESULTS DATABASE FAQ ABOUT LINKS

### Why gVolante?

More accurate assessment for genome assembly!

gVolante provides an online interface for completeness assessment of user's original or publicly available sequence datasets as well as for browsing results of completeness assessment performed on publicly available genome and transcriptome assemblies.

Preparation of high-quality genome or transcriptome sequence datasets for a study system of one's interest is a crucial step for modern biology, and can bring about various effects on downstream analyses. Commonly used metrics for assessing the quality of genome and transcriptome assemblies are based on sequence lengths, such as 'N50 length'. In fact, those length-based metrics are superficial, and cannot take into account their composition, namely the coverage of genes and the accuracy of reconstructed sequences in there, which matter in various biological analyses. In contrast, assessment referring to a set of pre-selected conserved genes can provide a complementary metric of completeness taking the composition of given sequences into account.

In this web site gVolante, you can run completeness assessment on the set of sequences of your interest, by means of computing the coverage of pre-selected conserved genes, in addition to the sequence length-based metrics.

[READ TUTORIAL](#)

### CVG: Core Vertebrate Genes

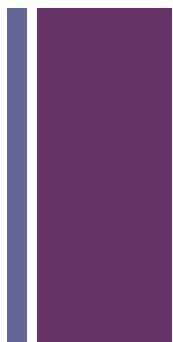
**A**

853,193 ChorNGOs  
One-to-one ortholog groups conserved in the 26 core vertebrates: zebrafish, and sea lamprey  
463  
292  
270  
223  
CVG

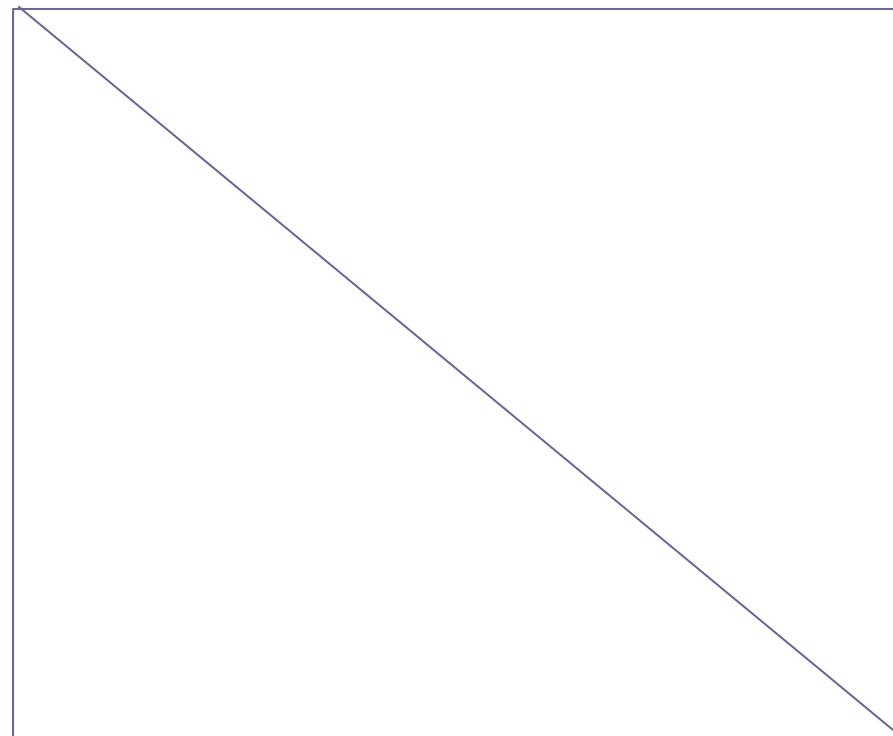
One-to-one ortholog groups with at least one ortholog in lungfishes (*Catlocarla intestinalis* and *C. aavignyi*)  
One-to-one ortholog groups conserved in the elephant shark genome  
Validation of the one-to-one orthology based on Ensembl gene trees

gVolante allows you to choose Core Vertebrate Genes (CVG), a new reference gene set of 233 ortholog groups that is compatible with completeness assessment particularly of vertebrate genomes and transcriptomes (Hara et al., 2015). Every group in CVG contains one-to-one orthologs as a single gene (without any paralogs generated in the vertebrate lineage nor gene loss) of all the vertebrate genomes including Chondrichthyes and Cyclostomata that were selected for screening. Our pilot assessments on genome assemblies of diverse vertebrates and embryonic transcriptome assemblies of the Madagascar ground gecko (*Paroedura picta*)

**B**



# + ショートリードシーケンサーの限界



# + ロングリードシーケンサー

- 完全長cDNAを読みきれるほど  
のリード長をもつ
- 1分子シーケンシングを用いて  
いるのでGC含量による読まれ  
やすさのバイアスが無い
- スループットとエラー率に課題



# 参考書籍



トランスクリプトーム解析 (シリーズ Useful R 7)

単行本 - 2014/4/9

門田 幸二 (著), 金 明哲 (編集)

★★★★☆ 2件のカスタマーレビュー



次世代シークエンス解析スタンダード～NGSのポテンシャルを活かしきるWET&DRY 単行本 -

2014/8/23

二階堂 愛 (編集)



シリーズ：細胞工学 > 細胞工学別冊

細胞工学別冊

次世代シークエンサーDRY解析教本

監修：清水厚志(岩手医科大学／いわて東北メディカル・メガバンク機構教授)  
井上秀雅(情報・システム研究機構 ライフサイエンス統合データベースセンター 特任准教授)

サイズ：B5変型判

頁数：408ページ

定価：本体5,400円（税別）

発行年月：2015年10月15日発行

ISBN\_10 : 4-7809-0920-1

ISBN\_13 : 978-4-7809-0920-3

# + 名古屋近郊ならば…

- 基生研(岡崎市)でRNA-seqチュートリアルが開催されています



基礎生物学研究所  
ゲノムインフォマティクス・トレーニングコース

[「RNA-seq入門 - NGSの基礎からde novo解析まで」\(準備編\)/\(実践編\)](#)  
2017年2月23日(木)～24日(金)/2017年3月9日(木)～10日(金)

[「BLAST自由自在～配列解析の極意をマスターする」](#)  
2016年12月1日(木)

◆ 過去のコース

2016年8月25-26日 / 9月8-9日	<a href="#">「RNA-seq入門 - NGSの基礎からde novo解析まで」(準備編) / (実践編)</a>
2016年3月10日～11日	<a href="#">「RNA-seq入門 - NGSの基礎からde novo解析まで」</a>
2016年2月25日～26日	<a href="#">「準備編 - UNIXとRの基礎」</a>

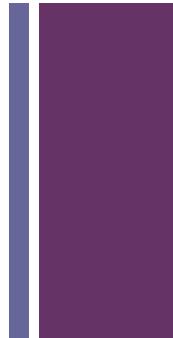
+

# コマンドラインを用いた解析

- NGS速習・NGSハンズオン講習会 (NBDC)  
<http://biosciencedbc.jp/human/human-resources/workshop>
- お家でできるMac Bookでやる次世代シーケンスデータ解析  
(緒方さん、日本バイオデータ)  
<http://www.ipad-zine.com/b/1520/>  
(閉鎖されたのでグーグル検索してみてください)
- BioPapyrus (孫さん、東大)  
<http://biopapyrus.net>

+

# R+Bioconductor



- Rstudioの導入(統合TV)

<http://tgotv.dbcls.jp/ja/20140221.html>

- 坊農さん(DBCLS)の講義のまとめ

<http://motdb.dbcls.jp/?AJACS51%2Fbono>

- (Rで)塩基配列解析(門田さん、東大)

[http://www.iu.a.u-tokyo.ac.jp/~kadota/r\\_seq.html](http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html)

# <sup>+</sup>Gene ontology + エンリッチメント解析

- Gene Ontologyを使って特定遺伝子の機能情報を検索する  
(統合TV)

<http://tогotv.dbcls.jp/ja/20111028.html>

- DAVIDの使い方 実践編 (統合TV)

<http://tогotv.dbcls.jp/ja/20130528.html>