NGSデータから遺伝子発現を見るためのホップ＆ステップ

統合データベース講習会

AJACS 伊予

2015/09/25

愛媛大学

理研CLST　原 雄一郎

自己紹介

京都大学 理学研究科 生物科学専攻

北海道大学 情報科学研究科
生命人間情報学専攻

産業技術総合研究所
バイオメディシナル情報研究センター

理研CDB ゲノム資源解析ユニット

理研CLST 分子配列比較解析ユニット

分子進化学・比較ゲノム学

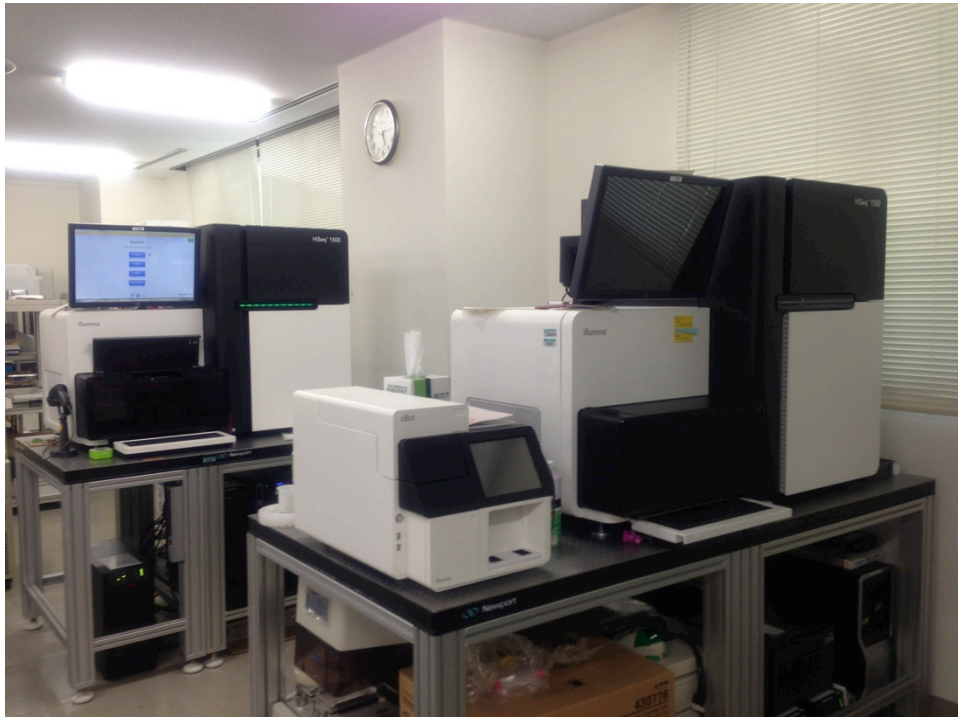
- 生物種の進化史を明らかにする
- 遺伝子の進化から表現型の進化を探る:
  - 遺伝子レパートリーの進化
  - ゲノム構造変異による進化

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
トランスクリプトーム解析

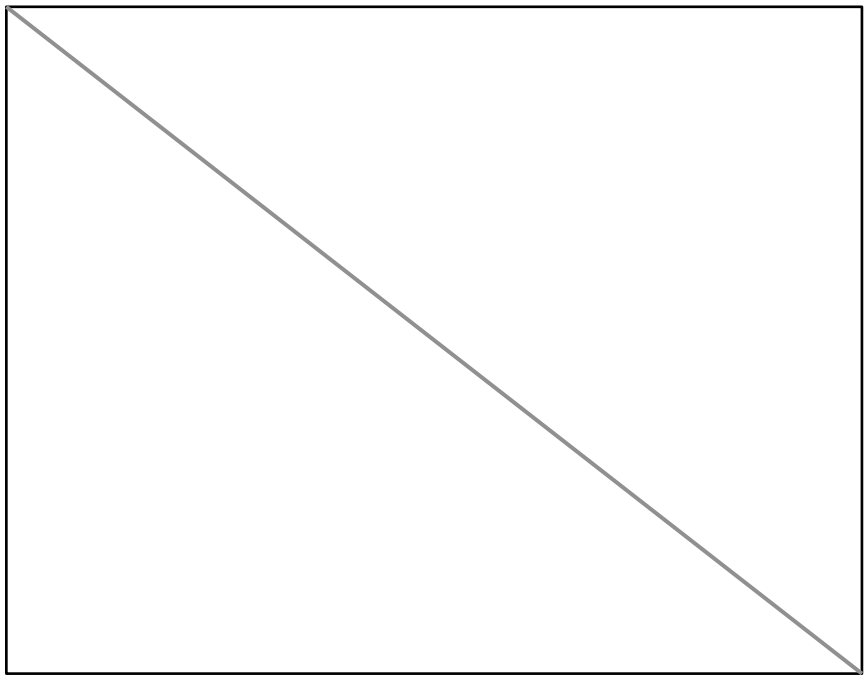
シーケンシングコア@神戸理研

- 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、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



今日お話しすること

ホップ

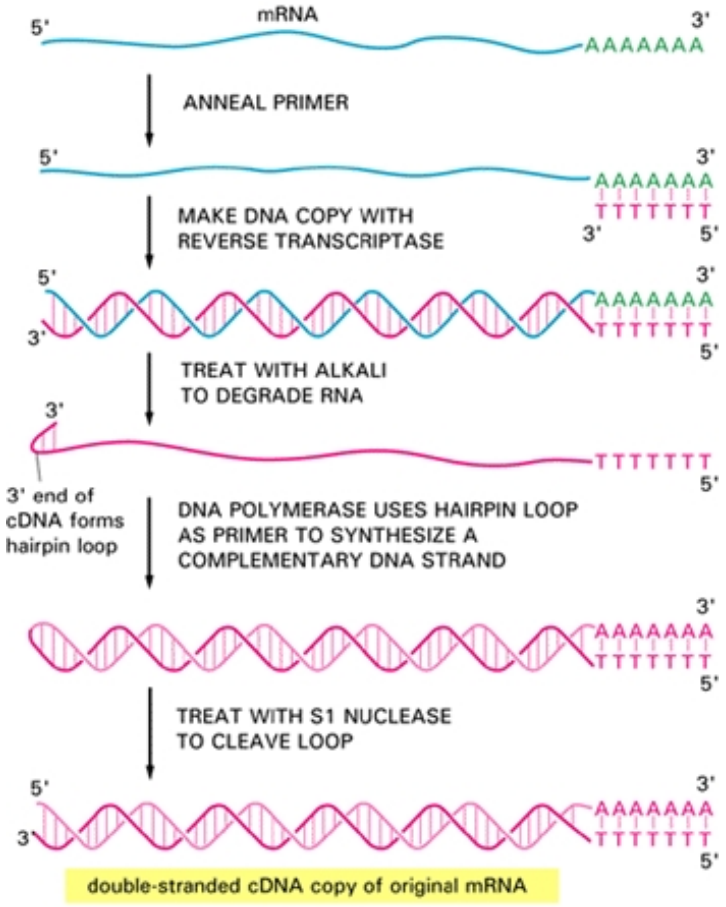
- RNA-seq: 次世代シーケンシングデータから遺伝子発現を知る
- 実験計画: RNA-seqを行う前に
- 発現解析の流れ

ステップ

- シーケンスデータ、発現データに触れてみる
- 非モデル生物のRNA-seq

転写産物の配列情報を読み取る

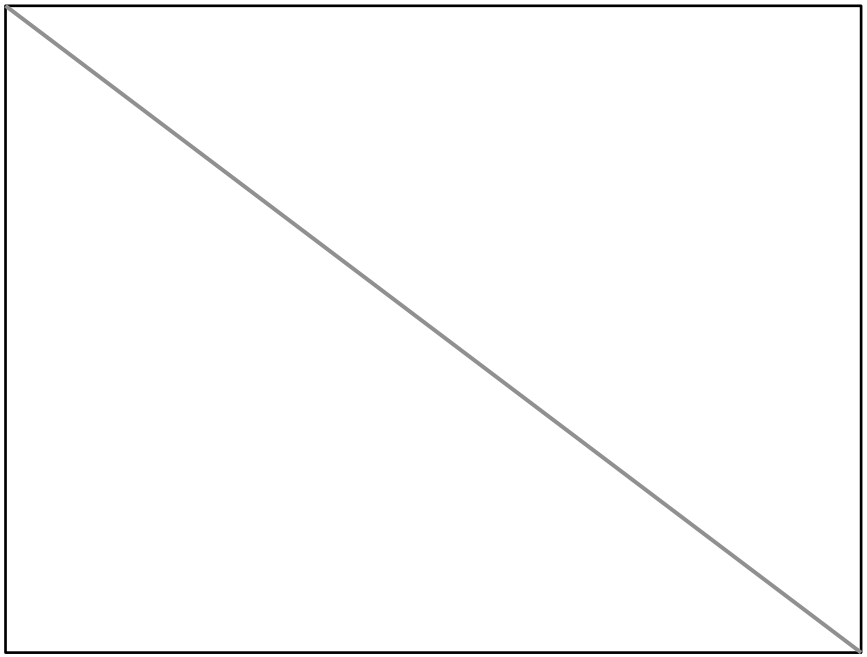
- mRNA→cDNAに逆転写してDNA配列情報とし、シーケンサーで読み取る



Molecular Biology of the Cell. 3rd edition.
Alberts B. et al.
New York: Garland Science; 1994.

サンガーシーケンス時代…

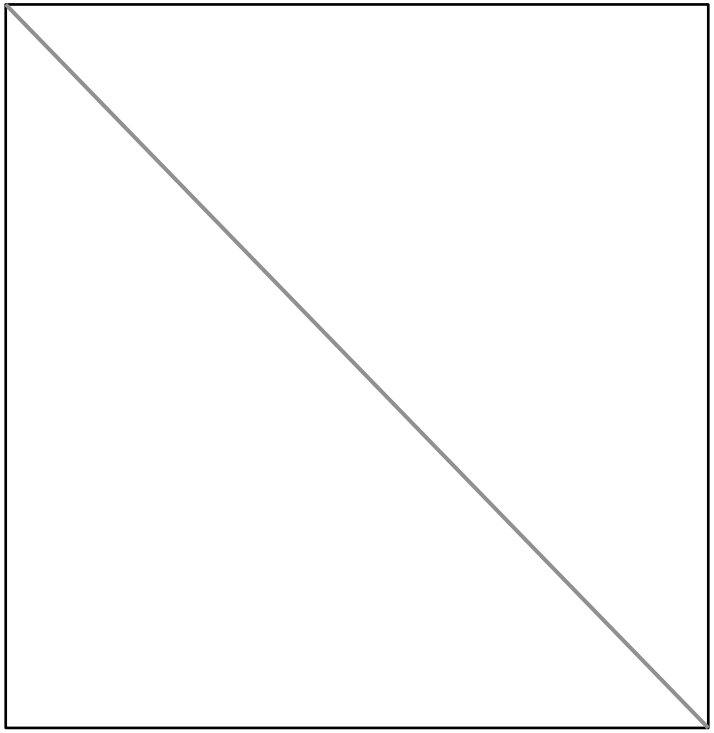
- 個別の転写産物をシーケンシング
- ハイスループット化
  - EST (Expressed Sequence Tag)
  - Full-length cDNA project



Pandey et al., *Trends in biochemical sciences* (1999).

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(→トランスクリプトーム)
- 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

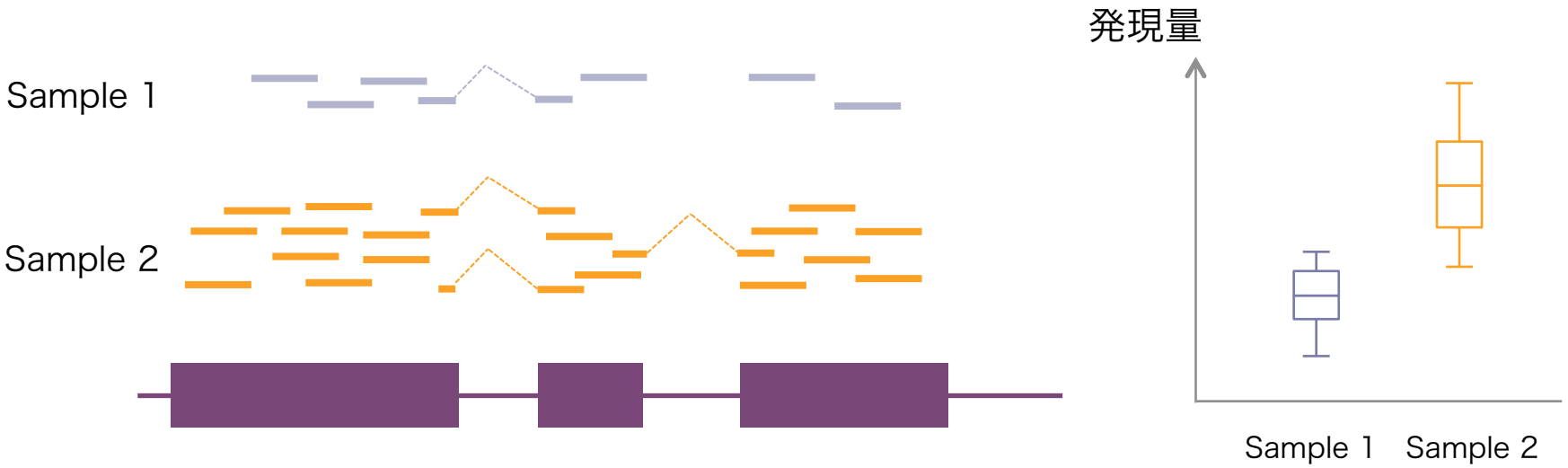


膨大なシーケンスデータから何がわかるか

- 遺伝子発現の定量化、サンプル間における発現差異の同定
- 未知転写産物の同定（スプライシングバリアント、新規遺伝子、ncRNA)
- 非モデル生物の転写産物カタログの作成
- アリル特異的な発現の同定

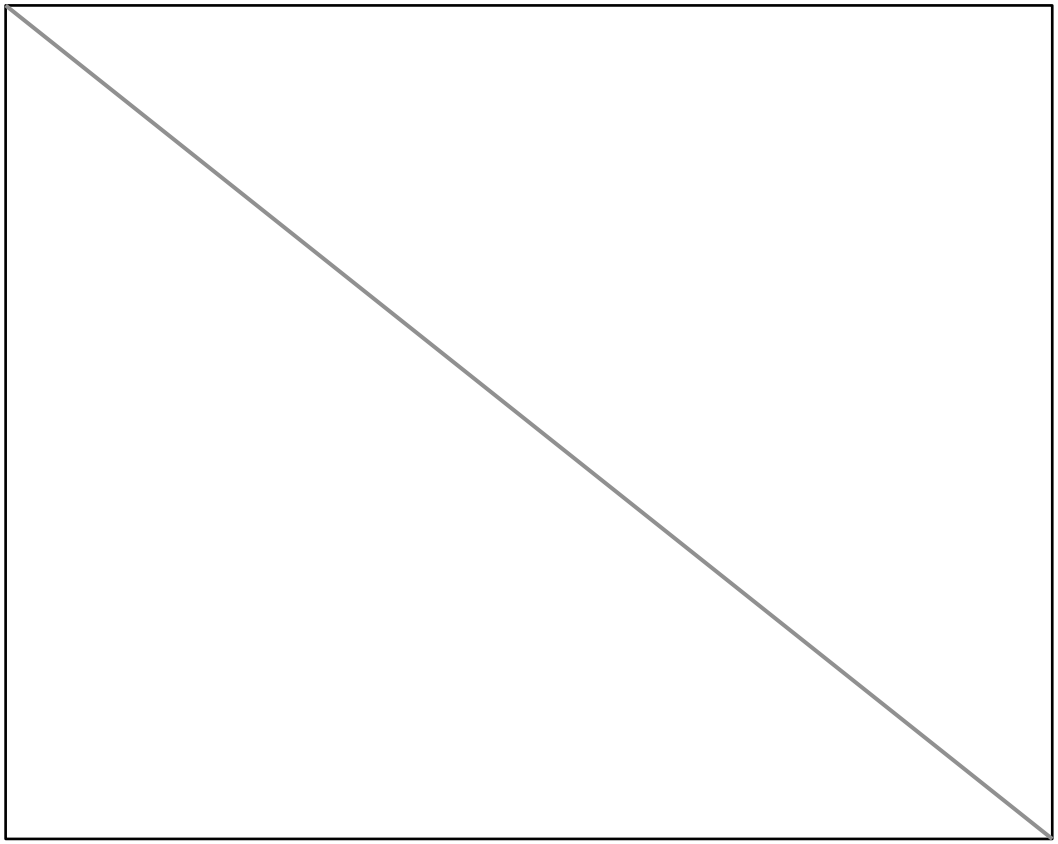
RNA-seqによる発現解析

- 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比較し、新規の遺伝子やスプライシングバリアントを同定する



発現解析から例えばこんなことが調べられてきた

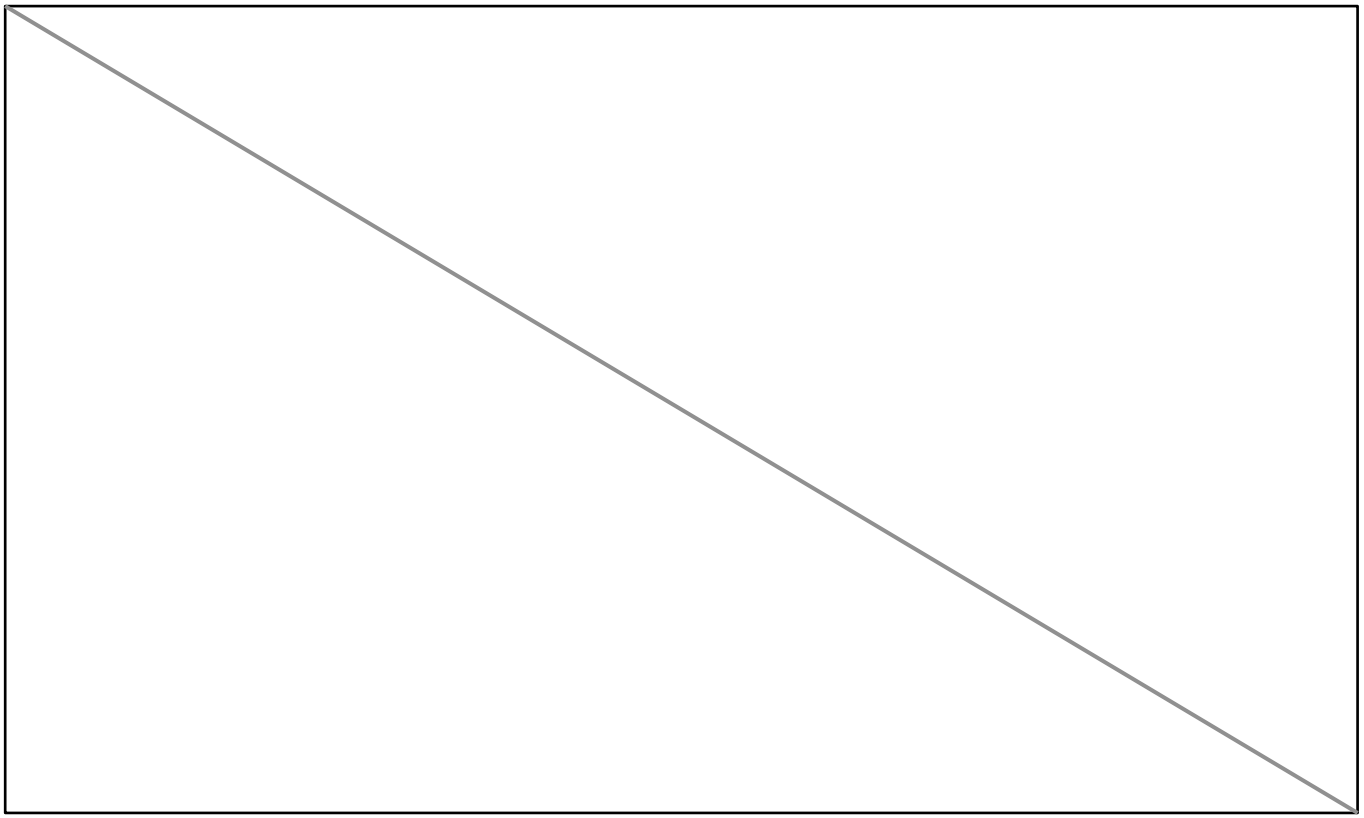
- 乾燥ストレス下における
  トウモロコシ子房での
  応答
- 正常に育つ個体と比べて
  サイクリン遺伝子の発現
  量が低下→G1, G2期の
  細胞が蓄積する



Kakumanu, et al. *Plant physiology* (2012).

発現解析から例えばこんなことが調べられてきた

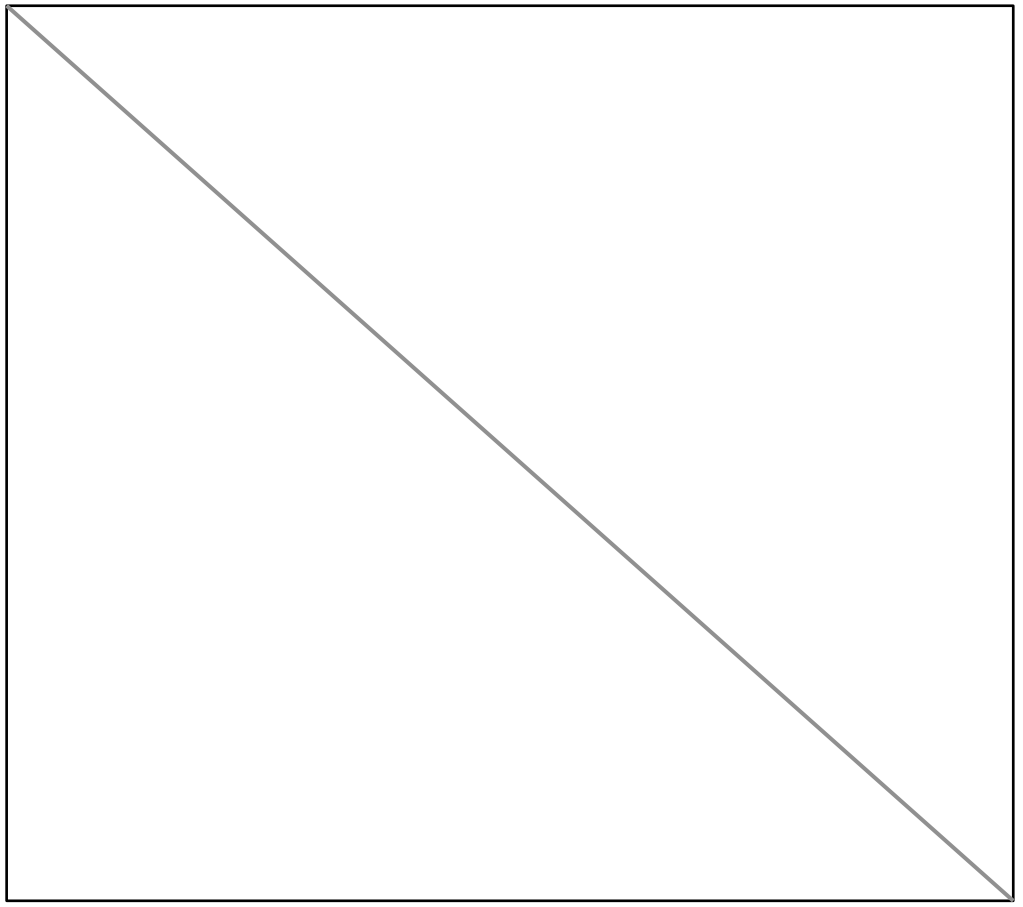
- ツメガエルの発生初期における遺伝子発現の変化



Tan, et al. *Genome research* (2013).

発現解析から例えばこんなことが調べられてきた

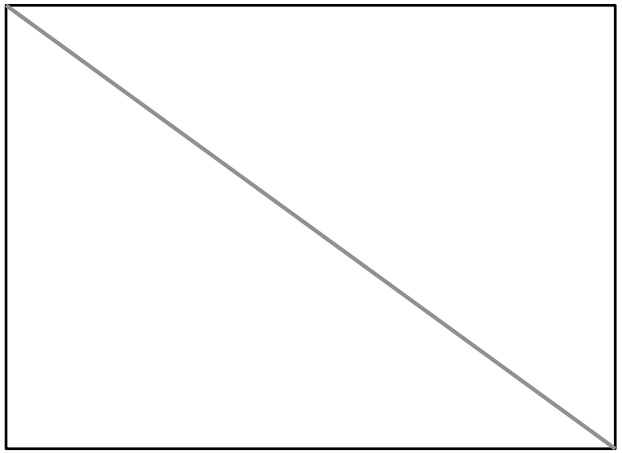
- がん細胞のトランスクリプトーム解析: fusion geneの同定



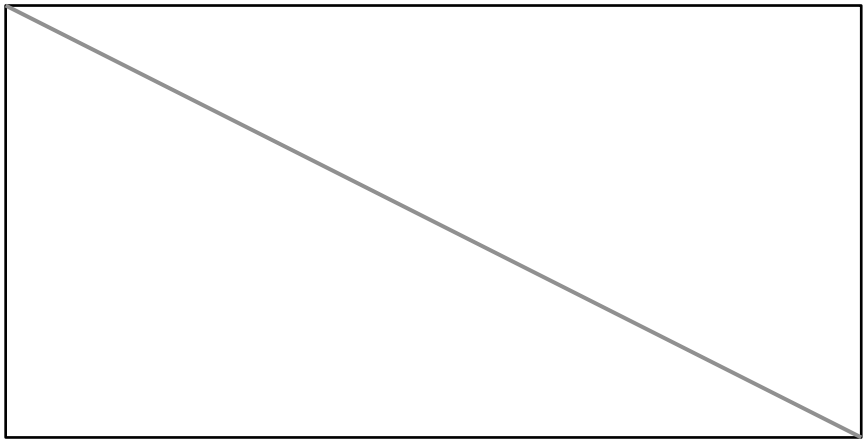
Northcott et al. *Nature* (2012).

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).

RNA-seqを行うとすれば…

たとえば…

異なる表現型をもつ2グループにおいて、表現型の差異を決める遺伝子を知りたい！

- 2グループで発現に差異がある遺伝子を同定
- 発現差がある遺伝子が表現型を決めている・・・？

どこでシーケンシングする？

- 所属する研究室
- 所属大学や機関のシーケ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
- もしくは、公共データベースからシーケンスデータを取得する

シーケンシングしよう！データ解析しよう！でもその前に

- RNA-seqにはそれなりのコストがかかる(数十万円〜)
- 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
- シーケンシングまでの行程でミスってクオリティの低いデータを出してしまったら、データ解析で挽回するのは困難

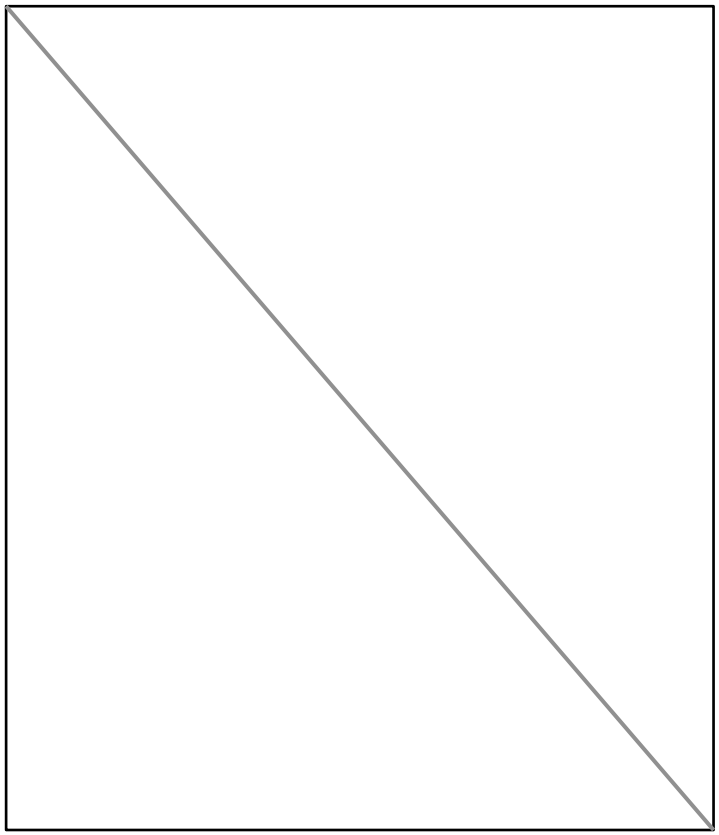
Take-home message

Bioinformatics is not a magic!

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
  - Biological replicates

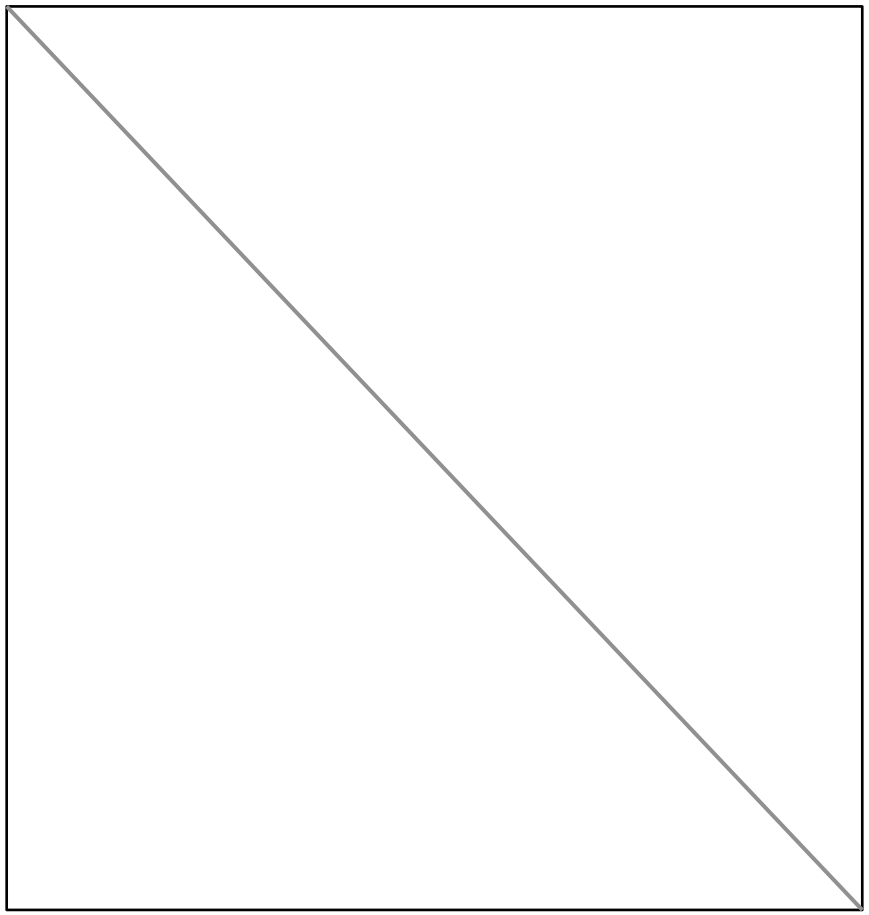
    ≥3 replicates
  - リファレンスゲノムデータの入手可否
  - 実験系統とリファレンスゲノムの系統との遺伝的距離
  - 細胞数(RNA量)を十分確保できるか



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

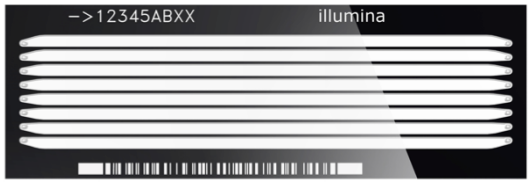
ライブラリ調製

- RNA amount/quality
  - 1μg total RNA
  - 分解していたらダメ
- mRNA isolation
- Fragmentation
  - シーケンスリード長、読み方(シングル/ペアエンド)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

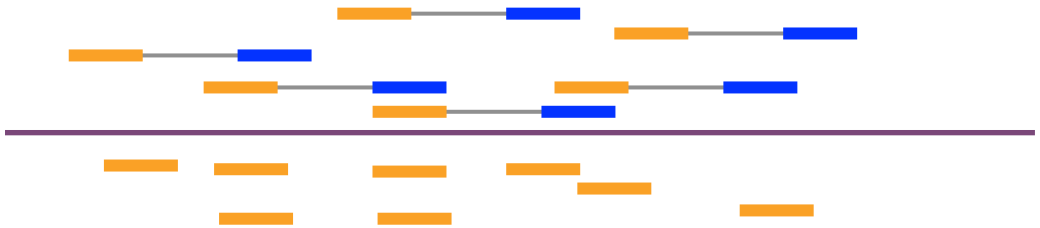


シーケンシング

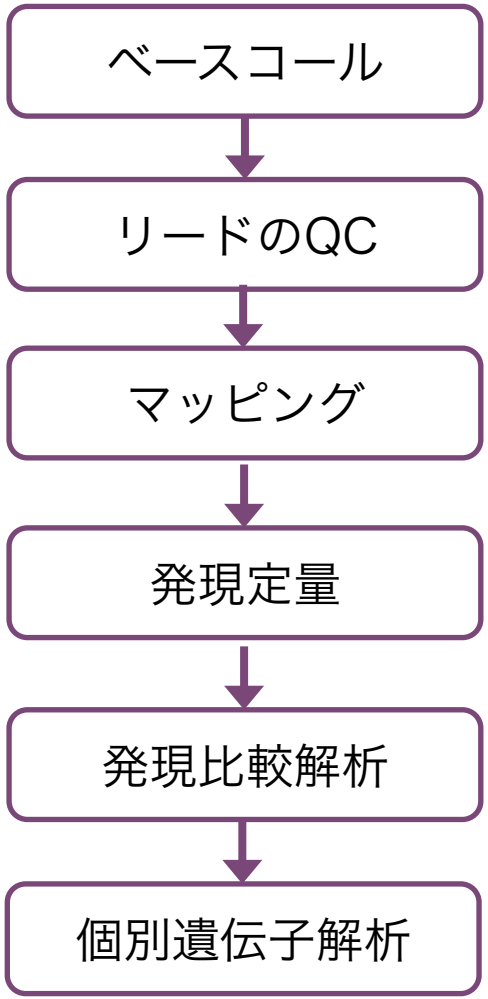
- プラットフォーム
- サイクル数(リード長)
- リード数(レーン数)
- シングルエンドorペアエンド



得たい情報量にあわせて選択



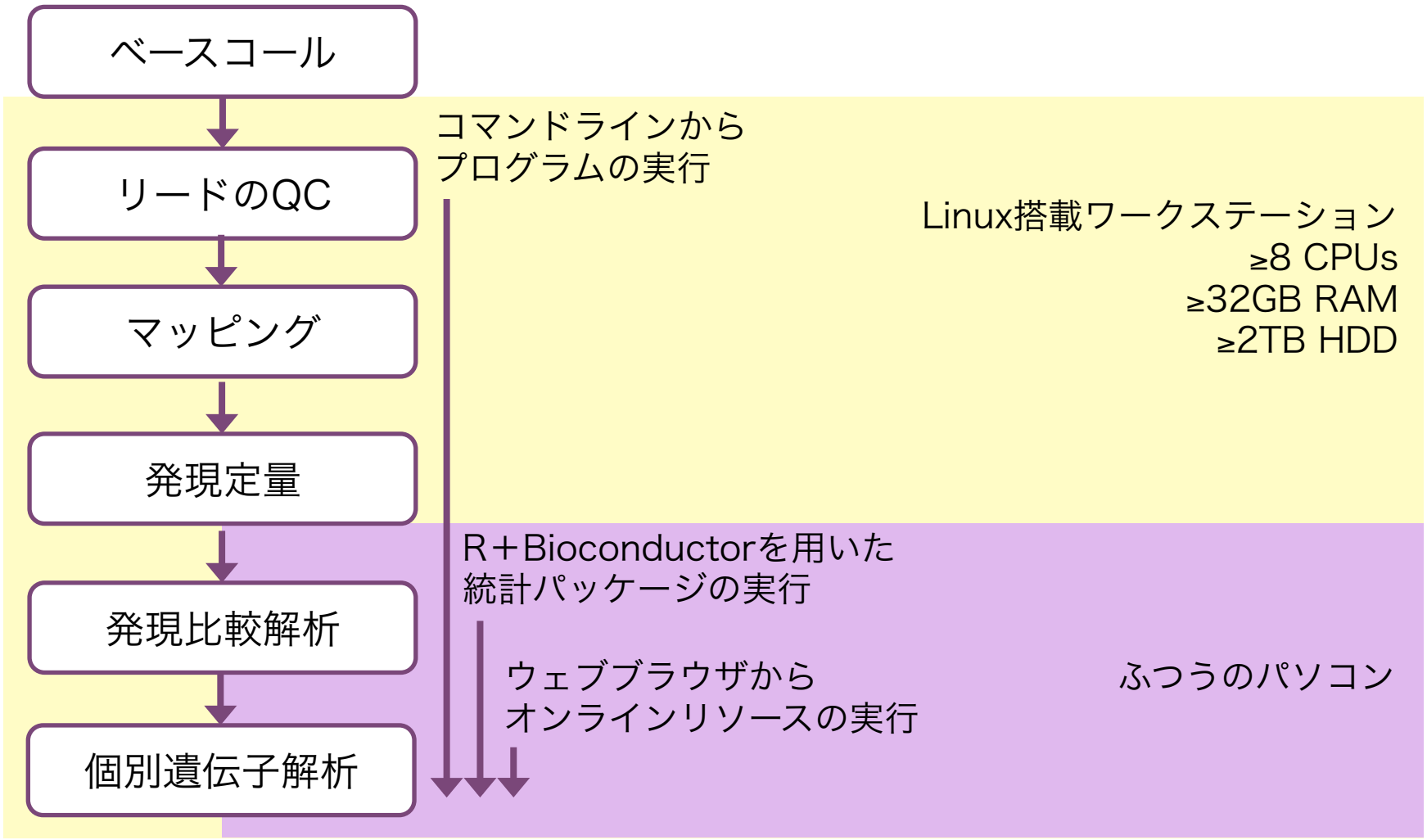
データ解析の流れ



- 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)

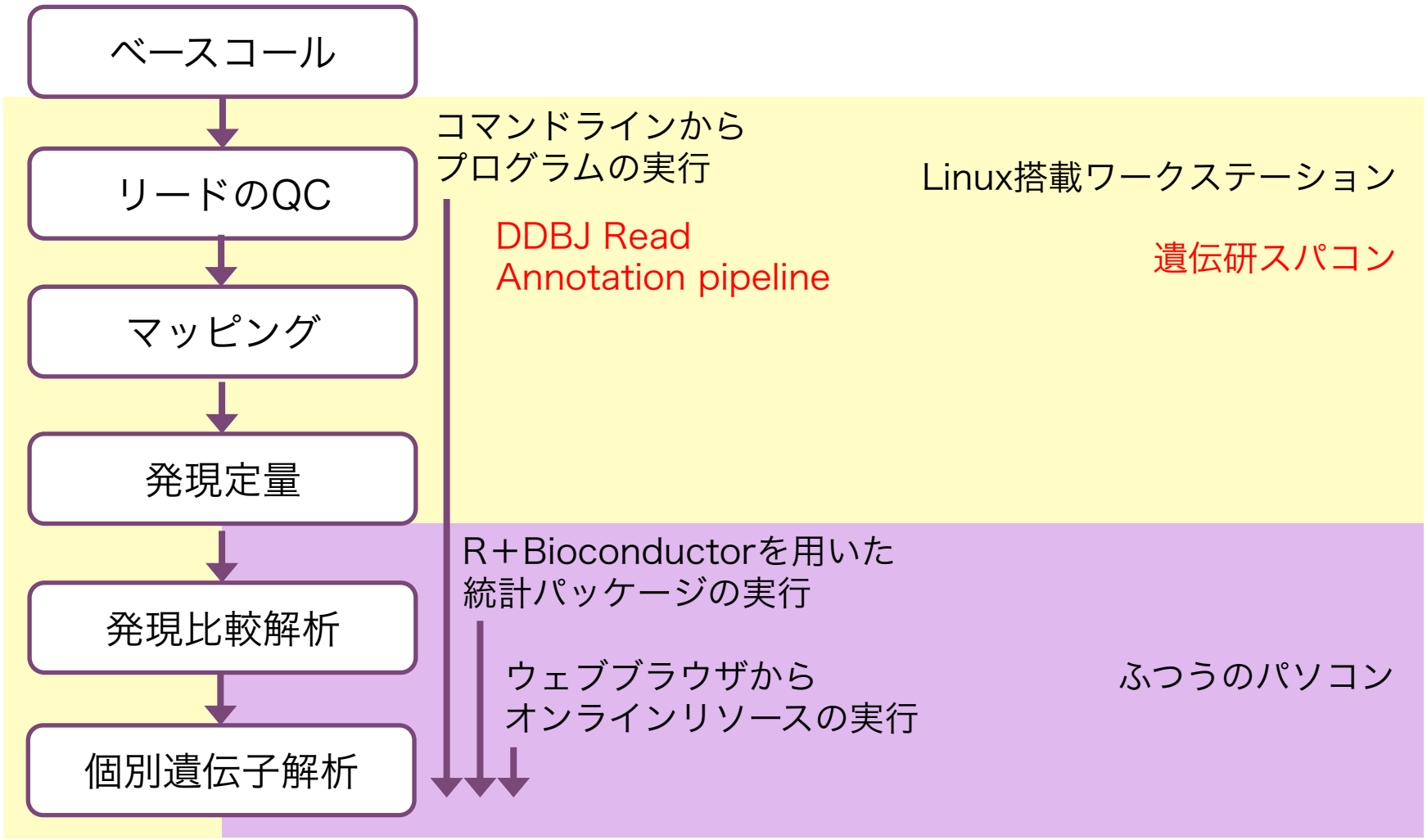
必要なハードウェア、スキル



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

ホップ

- 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
- 実験計画: RNA-seqを行う前に
- 発現解析の流れ

ステップ

- シーケンスデータ、発現データに触れてみる
- 非モデル生物のRNA-seq

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？


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

BMC Genomics

RESEARCH ARTICLE Open Access

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

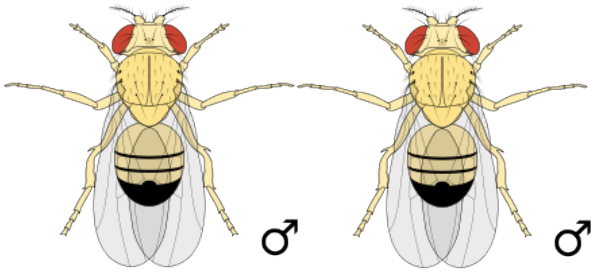
Ana Catalán, Stephan Hutter and John Parsch*

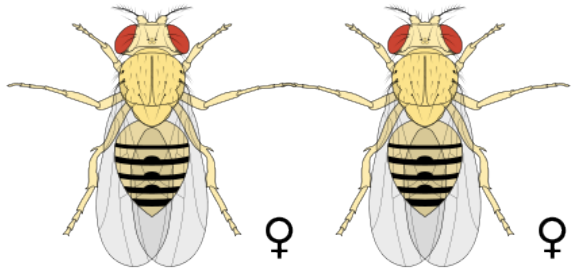
- 成虫雌雄それぞれ60個体分の脳から抽出したRNAを1ライブラリとする
- 各ライブラリは、Illumina HiSeq 2000にて51 bp single-endでシーケンス
- 2 Biological replicates per sex
- 使用するプログラムを環境によって色分けして表示

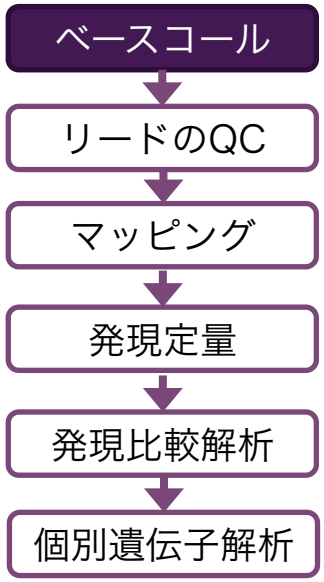
コマンドライン	Linuxワークステーション
コマンドライン・GUI 兼用	Windows, Mac, Linux

シーケンスデータの入手

- 4サンプル分(雄x2, 雌x2)をの配列データを取得する
- シーケンサーが出力したデータからコールした塩基配列データ (FASTQ形式)を取得する
  - シーケンシングプラットフォームによって方法が異なる
- もしくは公共データベースから取得する







FASTQフォーマット

塩基配列とそのクオリティスコアを同時に格納するテキストフォーマット

```
① ー @HWI-ST143:445:C044NACXX:6:1101:1440:1978 1:N:0:ATCACG
② ー CAAAATTATATCTTAATCCAACATCGAGGTCGCAATCTTTTTTATCGAT
③ ー +
④ ー 1=DDFFDFHFHFIJJJJJJJJJJIJJJGIGIJJJJIJIJJJJIJIJIBH
```

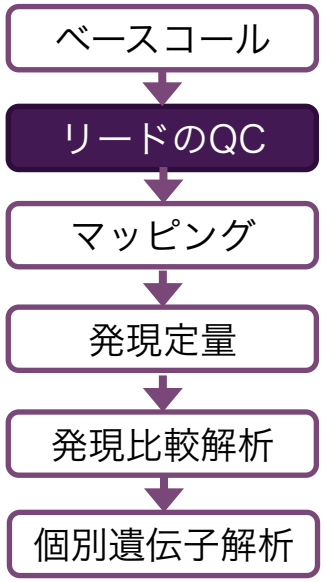
① “@”で始まり、配列のID等を記す
② 塩基配列を記す
③ “+” を記す
④ 塩基配列のクオリティスコア(ASCII文字)を記す

SRA

```
@SRR571454.1 HWI-ST143:445:C044NACXX:6:1101:1440:1978 length=51
CAAAATTATATCTTAATCCAACATCGAGGTCGCAATCTTTTTTATCGAT
+SRR571454.1 HWI-ST143:445:C044NACXX:6:1101:1440:1978 length=51
1=DDFFDFHFHFIJJJJJJJJJJIJJJGIGIJJJJIJIJJJJIJIJIBH
```

シーケンスデータのQC (クオリティチェック)

- 4サンプルそれぞれについてシーケンスリードのクオリティをチェックする
- クオリティーの低い領域をトリムする、あるいはリードごと除去する
- クオリティが保証されたリードより構成されるFASTQファイルが出力される



シーケンスデータのQC (クォリティチェック)

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 = -10log(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

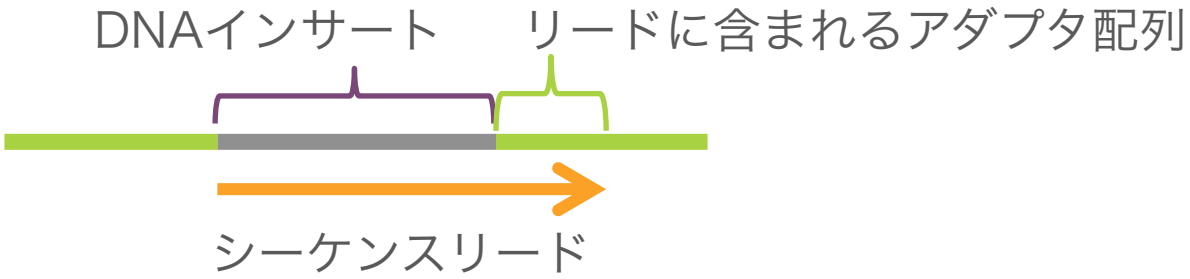
クオリティをプログラムでチェックする　FASTQC

演習 1 -1:

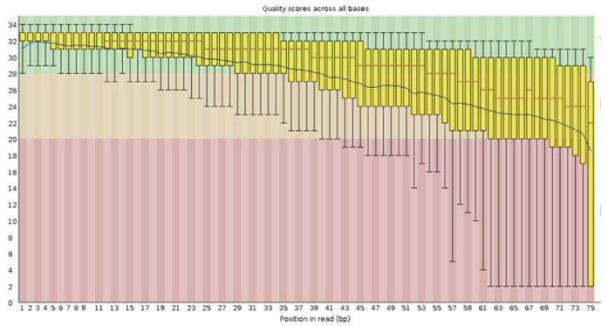
- 取得した配列のクオリティを見てみましょう
  - FASTQCプログラムを使います
- 以下の点について観察しデータの質について考えてみましょう
  - 塩基ごとのクオリティの傾向は？
  - 塩基組成のばらつきは？
  - 重複した配列の頻度は？
  - アダプタ配列の混入は？

“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

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？

解析するためのリファレンスデータ

- ゲノム配列データとアノテーションされた遺伝子データを取得
  - iGenomes (Illuminaが提供する配列データとアノテーション)

    https://support.illumina.com/sequencing/sequencing_software/igenome.html
  - USCS Genome Browser (基本的に動物のデータ、様々なアノテーション情報をダウンロードできる)

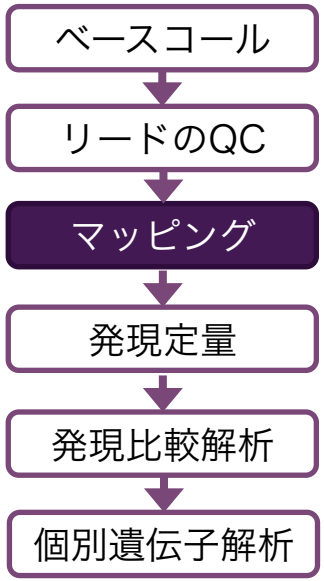
    https://genome.ucsc.edu
  - Ensembl/Ensembl Genomes (脊椎動物の他にも様々な分類群のゲノムデータを格納)

    http://ensembl.org

    http://ensemblgenomes.org

マッピング

- 最新リリースのショウジョウバエゲノム配列データおよび遺伝子アノテーションデータを取得する
- アノテーションデータの遺伝子構造をガイドにして、QCを行った各サンプルのリードを、スプライシング構造を考慮してゲノム配列にマッピングする
- FASTQファイルを入力とし、BAM/SAMファイルを出力とする

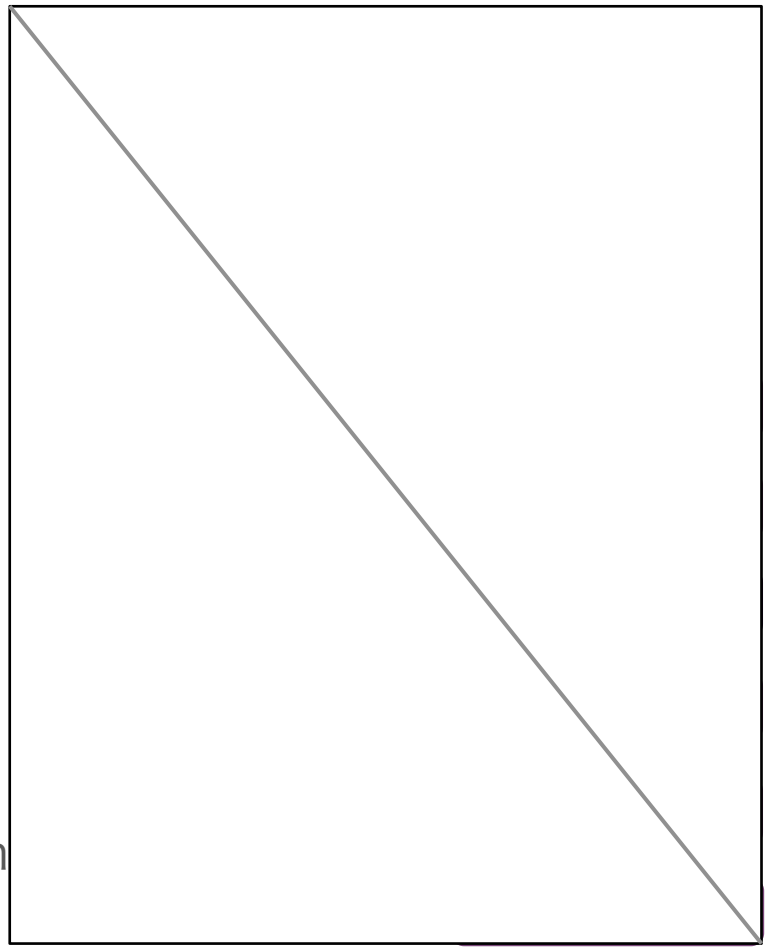


マッピング

- 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
- Input: FASTQ配列、リファレンス配列のインデックスデータ
- Output: SAM/BAMフォーマット
  - sam: sequence alignment map
  - bam: binary compressed version of sam



Kim, et al. " *Genome Biol* (2013).

SAMファイル

- タブ区切りファイル
- リードがマップされた位置、一致度、マッピングの状態を示す

```
SRR571454.17627901 0   2L  202320  50  50M *   0   0
TTATACTACGATTATTTATCCAGACGCGTATTTAATTATAATTAATATGT
@@DDDDDH<CFHIFHIIGII4CGEIIII@DHGIIIIGIIEGEHCFHBGHI
AS:i:-3 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:20T29
YT:Z:UU XS:A:+  NH:i:1

SRR571454.8545335  16  2L  1013813 50  33M1I15M    *   0   0
TCCAAGTGCTCACGCCAAATCAAAGTCGAATGCAAAAAACTCAATTCAT
CIJJJJIGHFG@HGGCGCIIJJIIJJIJIJJJJJIHGFADFFFFED?=<
AS:i:-8 XN:i:0  XM:i:0  XO:i:1  XG:i:1  NM:i:1  MD:Z:48
YT:Z:UU XS:A:-  NH:i:1
```

これで1行

詳しくは、http://cell-innovation.nig.ac.jp/wiki/tiki-index.php?page=SAM など

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

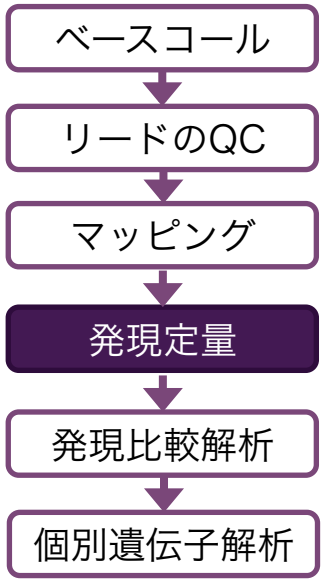
Qualimap

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

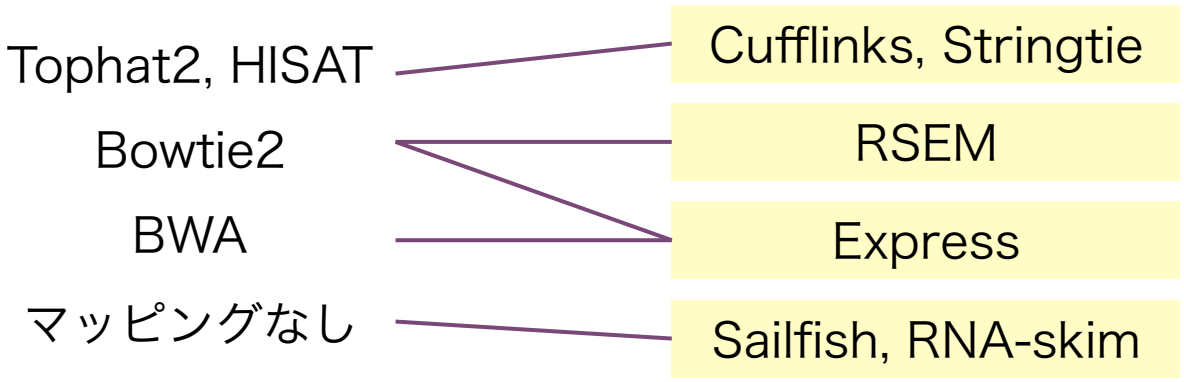
発現定量

- 各ライブラリのマッピングファイルおよび遺伝子のアノテーション情報に基づき、遺伝子ごとにマップされるリード数をカウントする
- 各ライブラリの各遺伝子において、リード数や遺伝子長で正規化した値を発現量とし、表に出力する

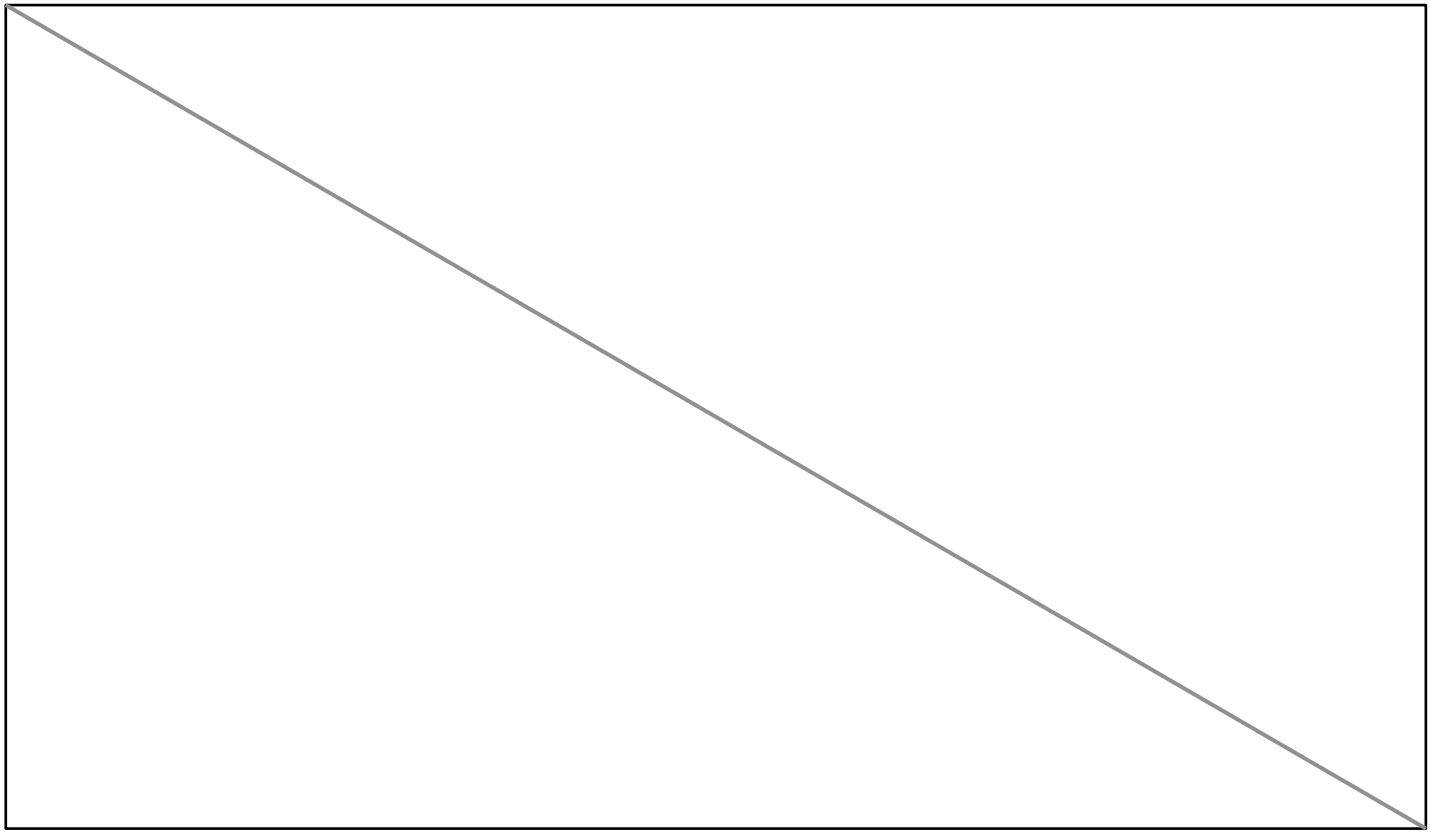


発現定量

- 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
  - 既知のアノテーション情報を事前に与える
  - 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
- 発現定量プログラム: マッピングプログラムとの相性



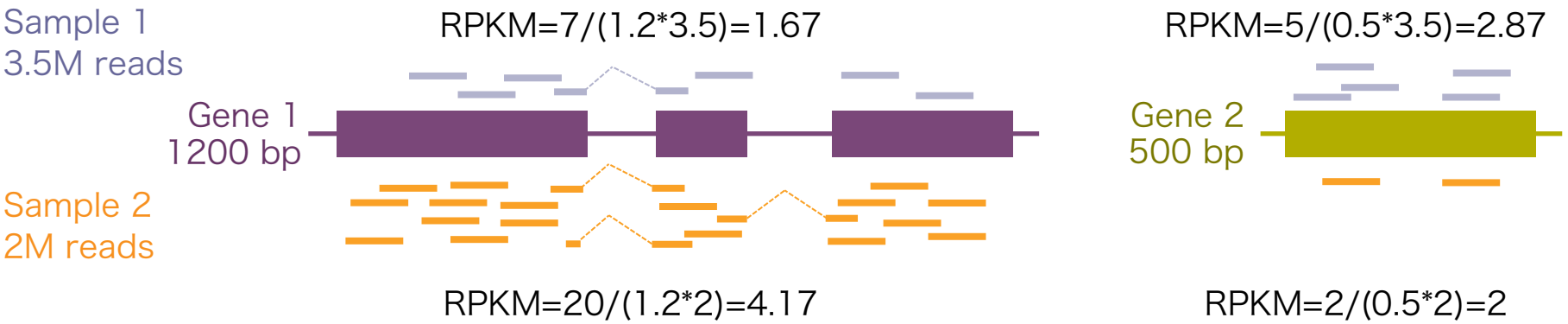
Tuxedo pipeline



Trapnell, et al. *Nature biotechnology* (2010).

発現定量

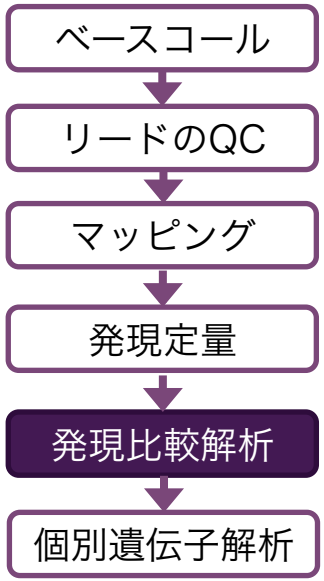
- マップされたリードの数は、リードの総数、転写産物の長さに応じて正規化される
  - CPM: (Counts per million)
  - RPKM/FPKM: Reads/Fragments per kilobase of exon per million mapped sequence reads)



  - TPM: (Transcripts per million)

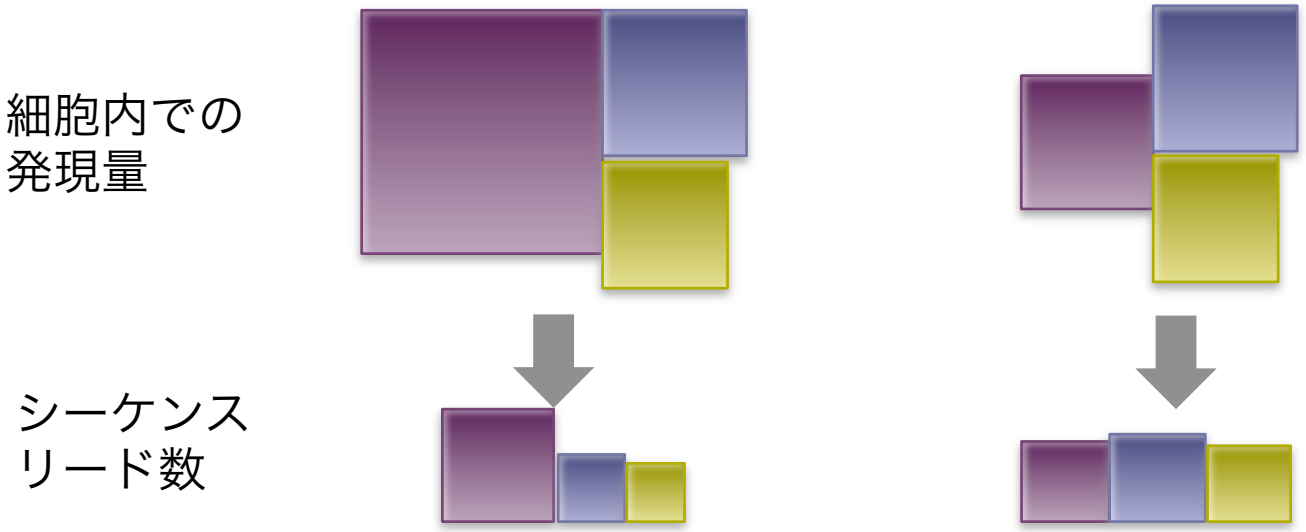
発現比較解析

- 雄と雌に分け、2群間で発現量(CPM)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

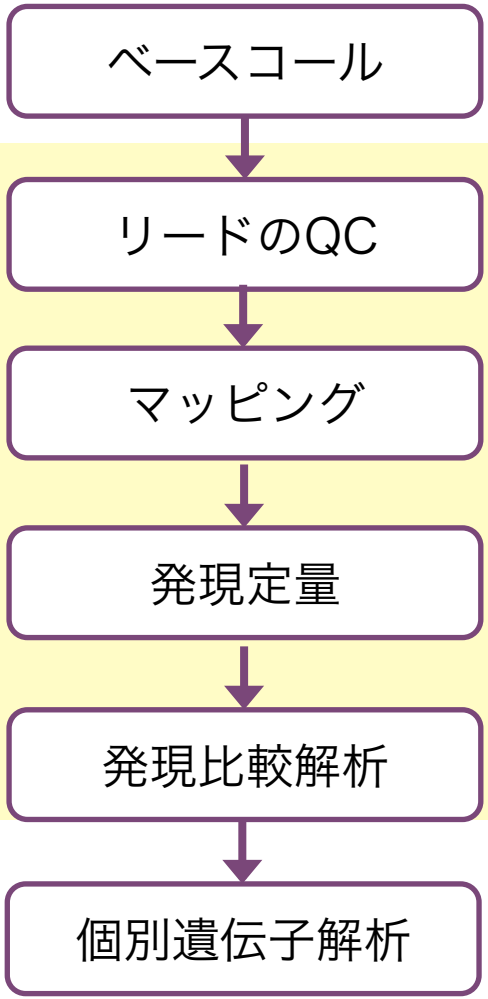
- 発現遺伝子の構成により正規化が必要



- 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

DESeq2, edgeR *in R+Bioconductor*

発現比較解析の実行



FASTQC
Trim_galore

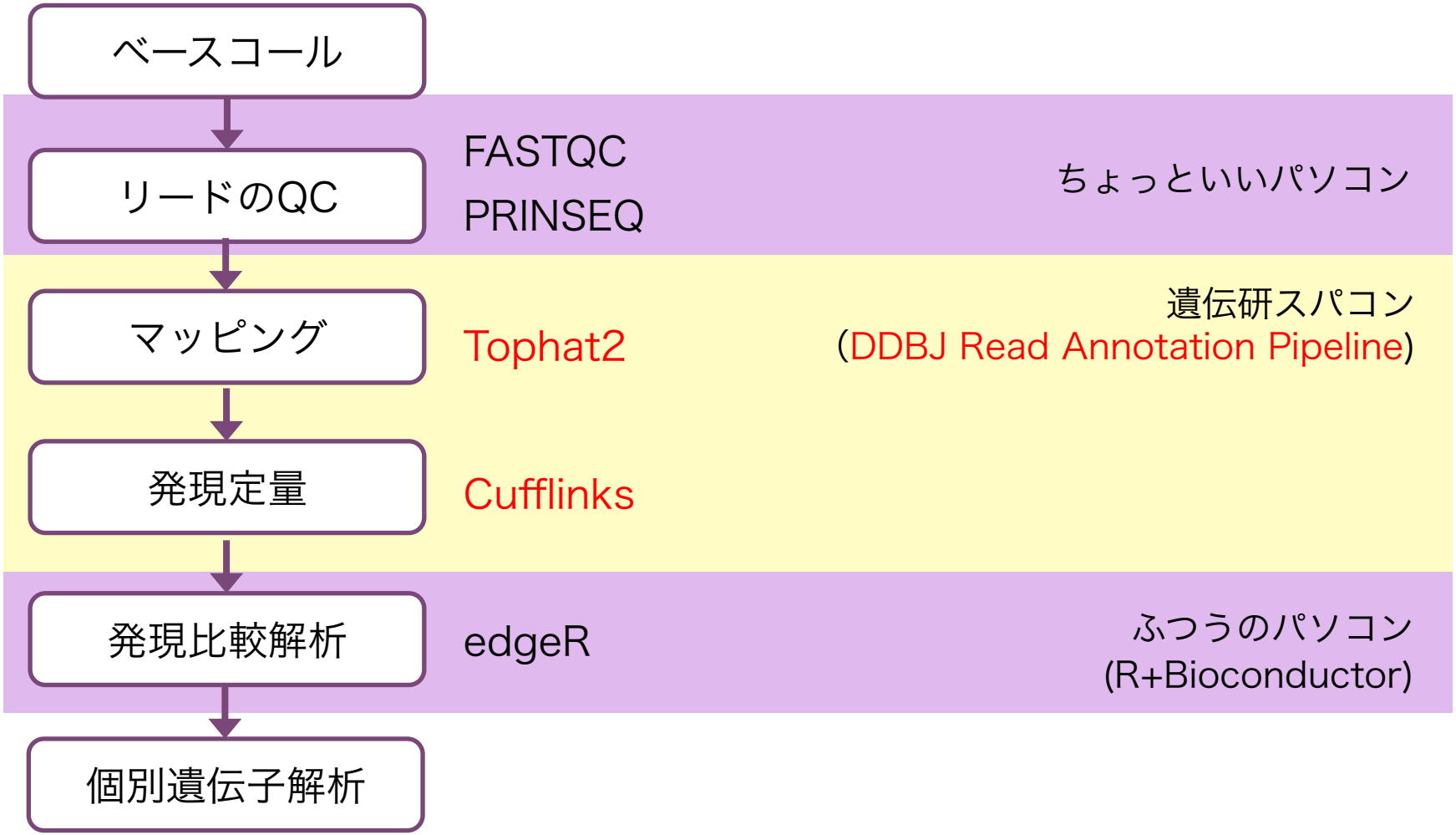
Tophat2
RSeQC

Cufflinks

edgeR

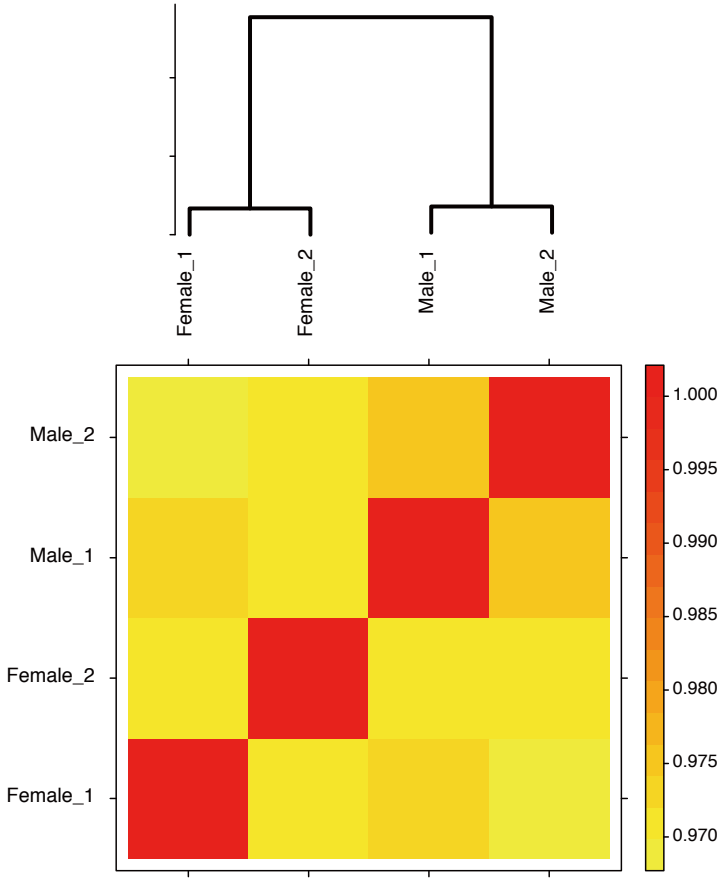
Linux搭載ワークステーション

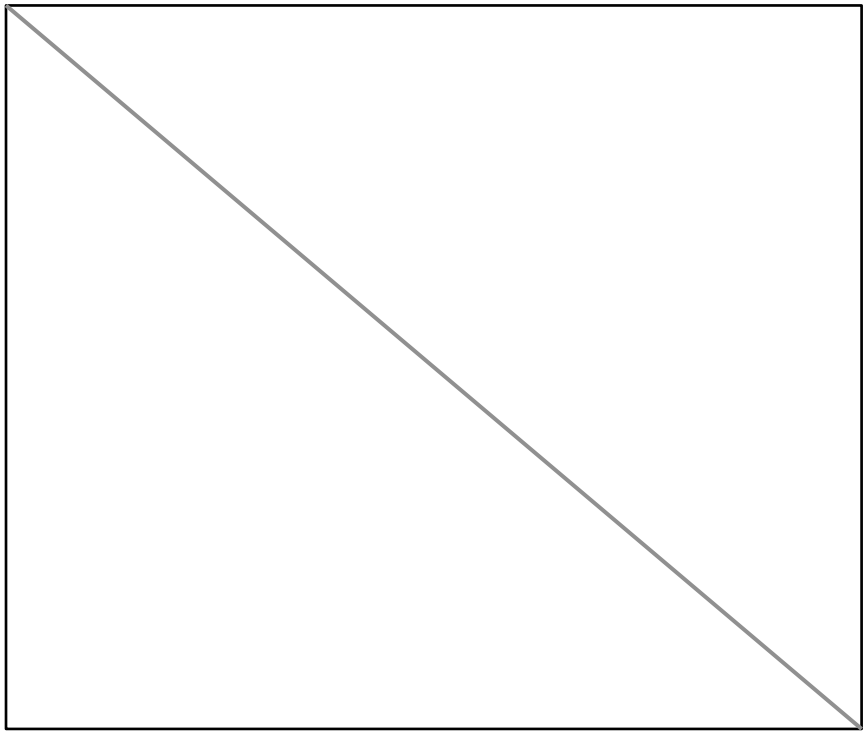
発現比較解析の実行
(コマンド不使用、無料ソフトウェア)



発現データを可視化する (R+Bioconductor)

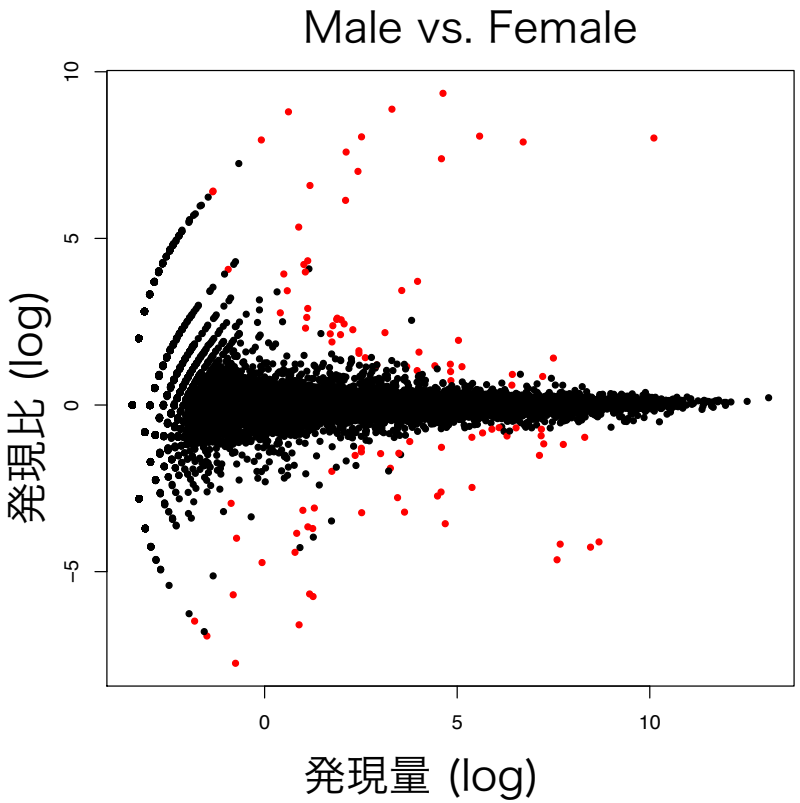
- 遺伝子プロファイルに基づくサンプルのクラスタリング

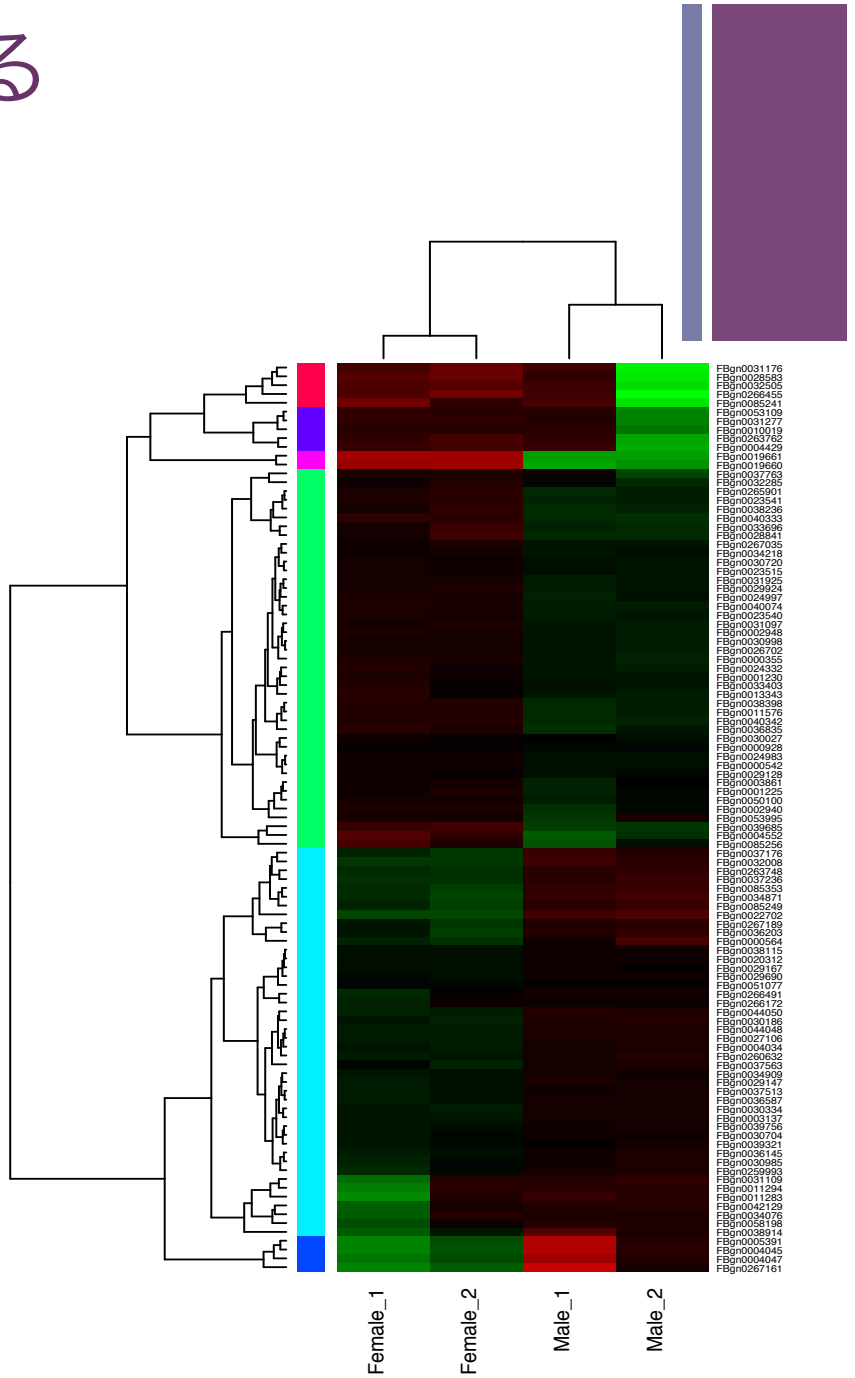




発現データを可視化する (R+Bioconductor)

- 群間で発現に差がある遺伝子

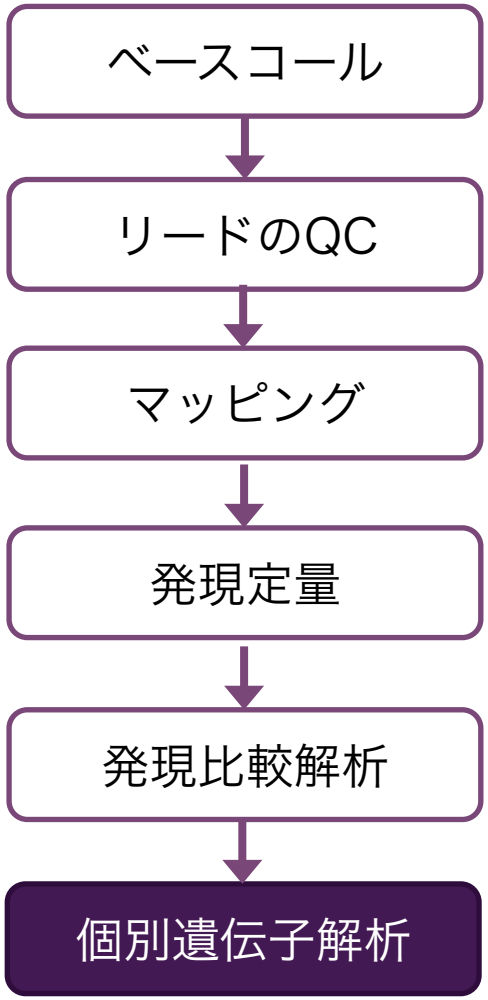




演習3

- 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

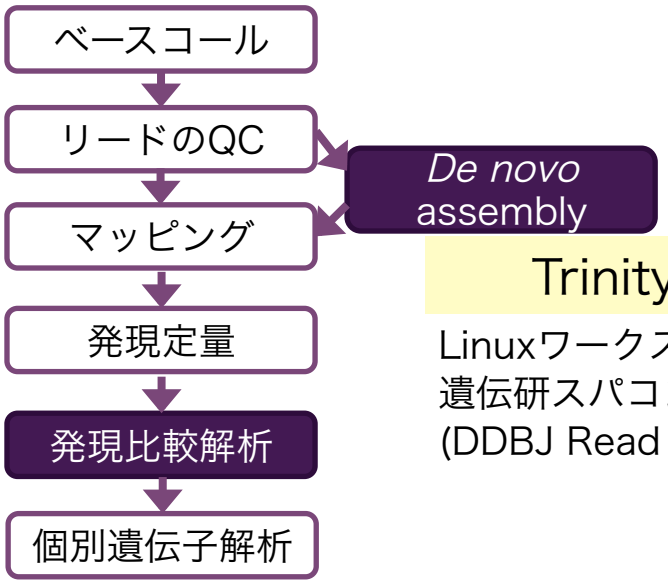
発現差がある遺伝子から何を得るか



- 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

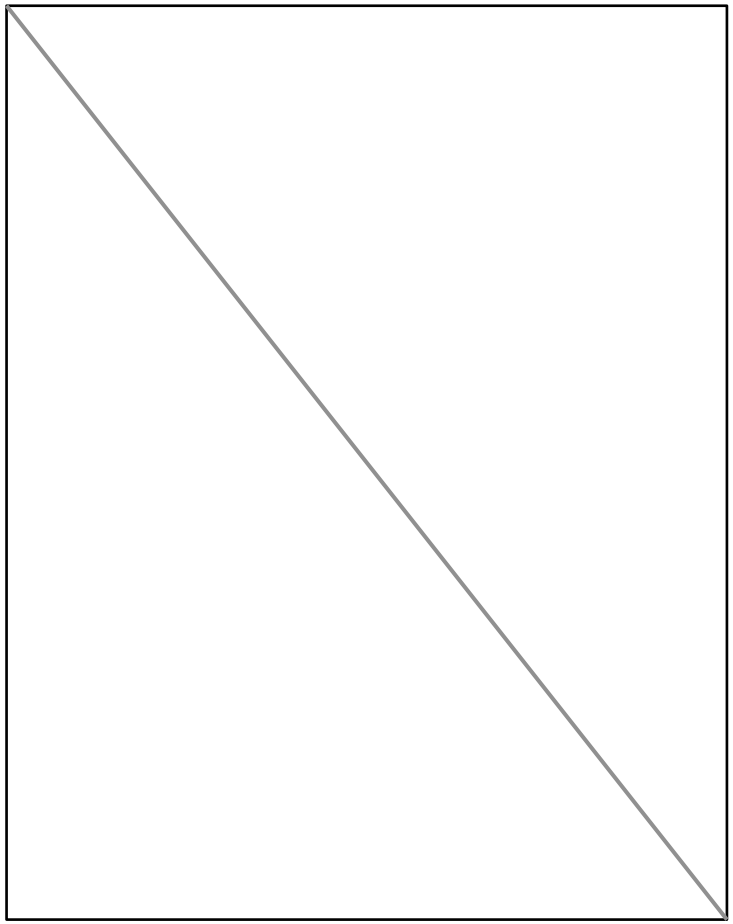
De novo transcriptome assembly

- 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スクリプト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



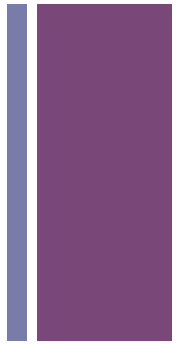
Trinity

Linuxワークステーション
遺伝研スパコン
(DDBJ Read Annotation Pipeline)



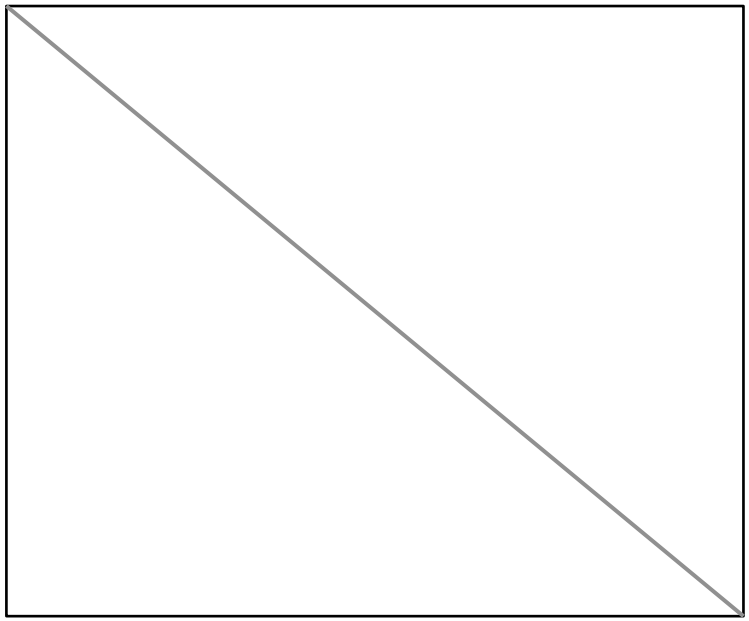
Iyer and Arul. *Nature biotechnology* (2011).

Genome/transcriptome sequencing in Kobe



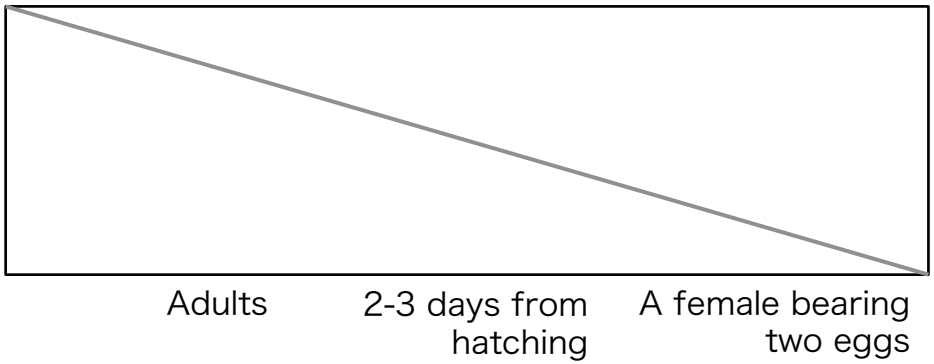
+ヤモリトランスクリプトーム

ニワトリとマウスだけでは
羊膜類の形態多様性を明かせない
→爬虫類のモデル



Nomura et al. (2013)

ソメワケササクレヤモリ
(*Paroedura picta*)



- 性成熟は6ヶ月程度。10日2個卵を産む
- 比較的殻が固い→胚操作を行いやすい
- 胚発生のステージが詳細に知られている
- ヤモリシーケンシングプロジェクト
  - トランスクリプトーム: 完了
  - ゲノム: シーケンシング完了、解析中

*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？

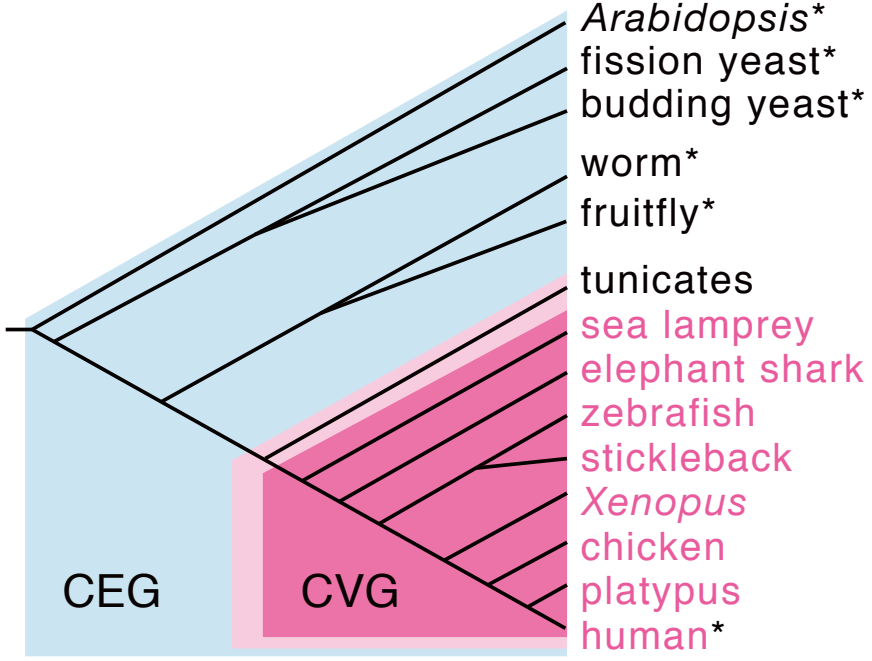
- シーケンス量が少なすぎる→低発現遺伝子配列を復元できない
- シーケンス量が多すぎる→コスト増、よくわからない配列が増える
- ゲノム・トランスクリプトームアセンブリの網羅性を測る指標
  - 配列の長さ (N50)
  - トランスクリプトームアセンブリの中に復元される遺伝子数

    →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

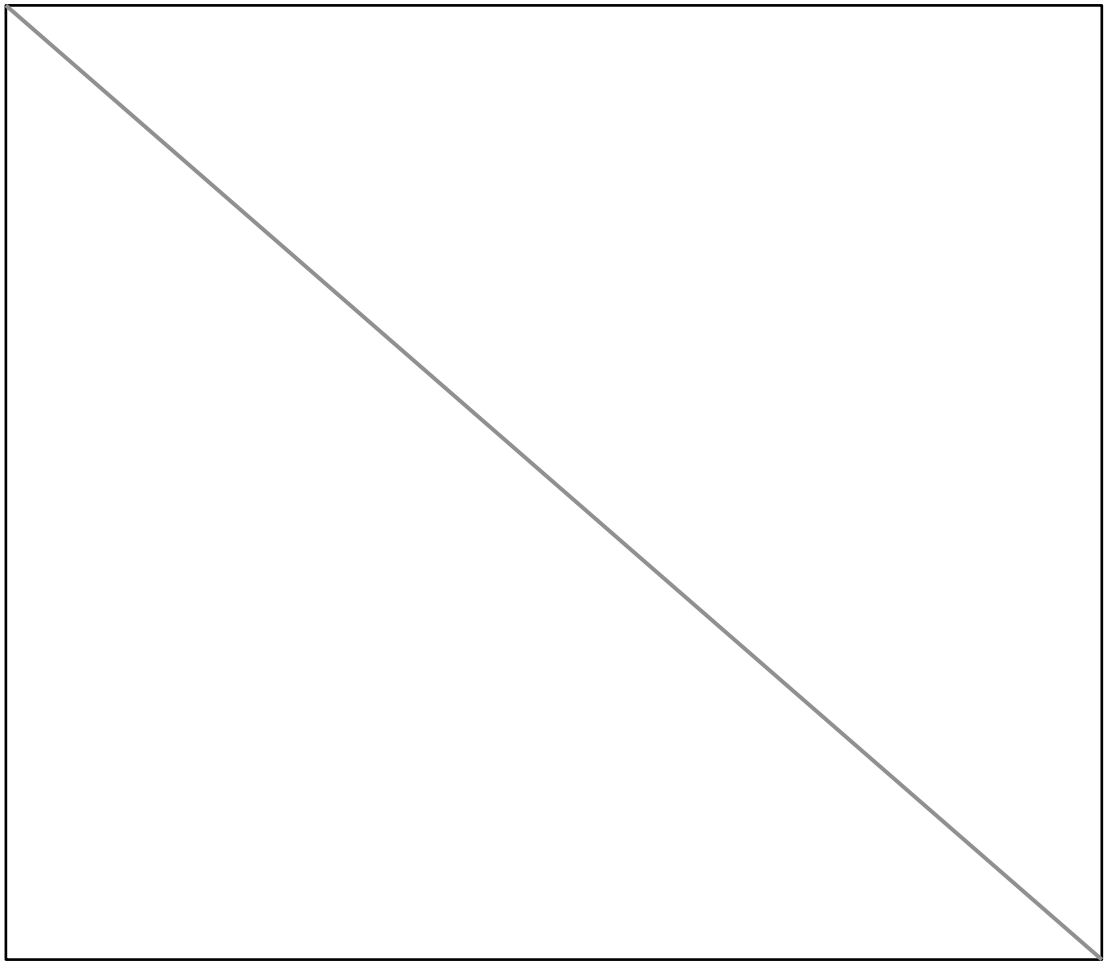
特定の系統にフォーカスしたリファレンス遺伝子セット

- 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

  ヤモリトランスクリ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



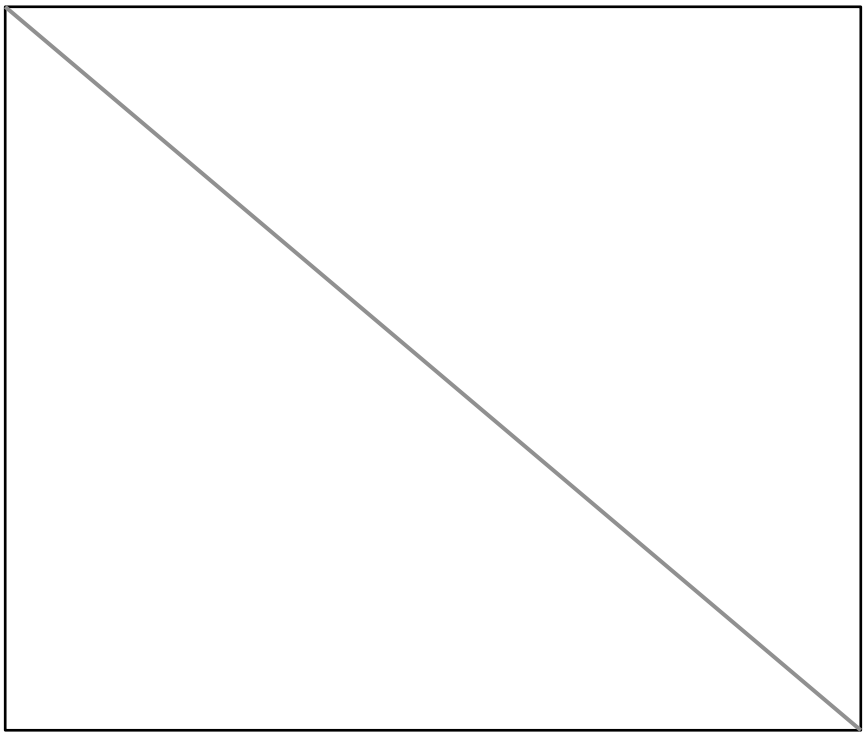
ショートリードシーケンサーの限界



Korf. *Nature methods* (2013).

Iso-seq

- PacBio RSIIシーケンサーを用いてcDNAをまるごとシーケンスする
- 平均10Kbのロングリードをシーケンス→大部分の転写配列をシーケンス可能



http://www.digital-biology.co.jp/allianced/products/pacbio/