

統計学入門

北海道大学大学院農学院
(兼) 数理・データサイエンス
教育研究センター
佐藤昌直

そのためには

- 測定、実験計画を見直せるように
- 仕組みを知る
- 試す - R
- 統計用語・表記に慣れる

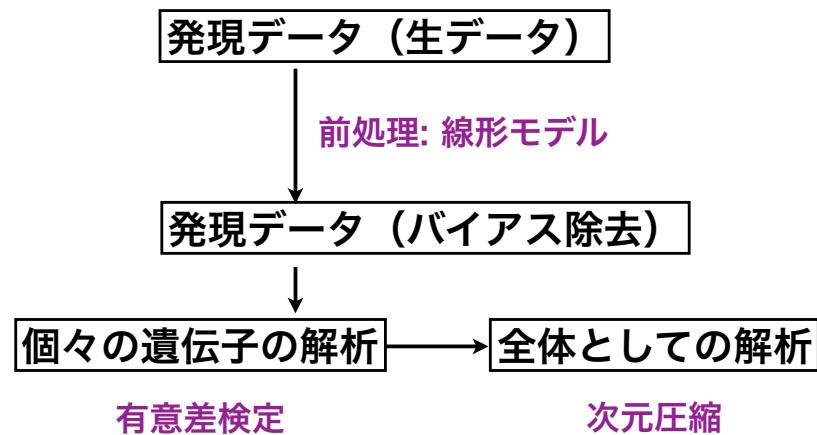
私が重視しているポイント

- 研究全体における統計の役割、
実験と統計との連携を意識する
- 遺伝子発現解析に必要な統計の
基礎概念を解説する
- “*statistical mind*”を養う

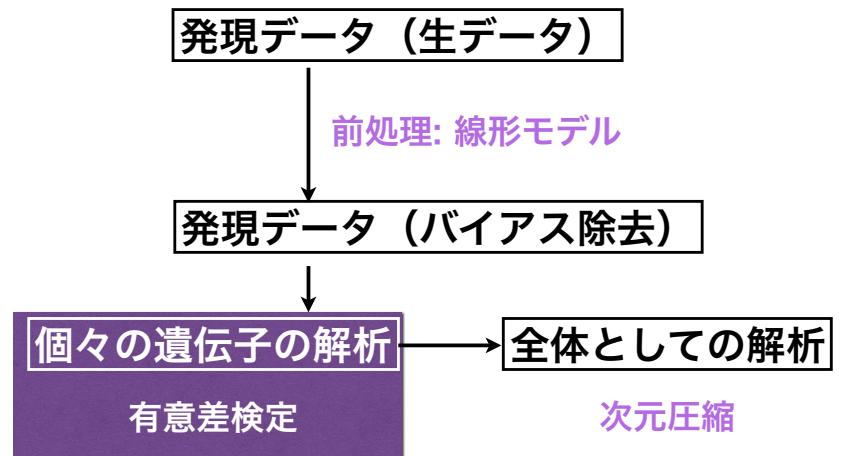
基本的な統計の用途

- 仮説検定
- 予測 (モデル構築)

遺伝子発現解析における統計の役割



遺伝子発現解析における統計の役割



仮説検定 - t 検定を例に

ねらい

t 検定から検定の背景知識を得る:

- 検定の流れを知る
- 勉強のとっかかりを作る

用語の意味の整理

- 統計量、確率分布、自由度、 p 値

統計における検定の手続き

1. 仮説を立てる
2. 統計量を求める
3. 求めた統計量を確率分布に照らし合わせる
4. 判定: 求めた確率と棄却限界値との比較

2. 統計量を求める:

ポイント

統計量: データから導いた
具体的な数値

↔ **母数:** 未知の数値

我々ができること: 少数の測定値（標本）から
「母集団」を推定すること

1. 仮説を立てる:

帰無仮説

statistical mind

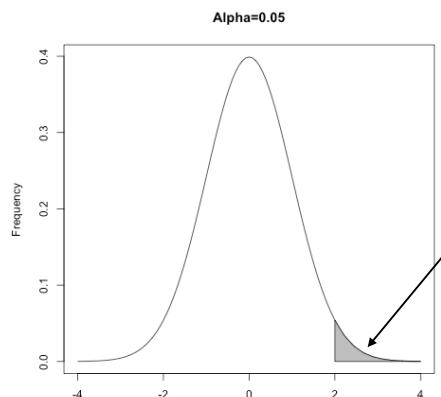
最終的に棄却される仮定:

「AとBに差がある」かを検定する場合は
「AとBには差がない」と仮定する

例1. 野生型と変異体Aの遺伝子xの発現量に違いがあるか？

例2. 野生型と変異体Aの遺伝子発現プロファイル間の相関
係数は0.35だった。これらは有意に相関していると
考えられるか？

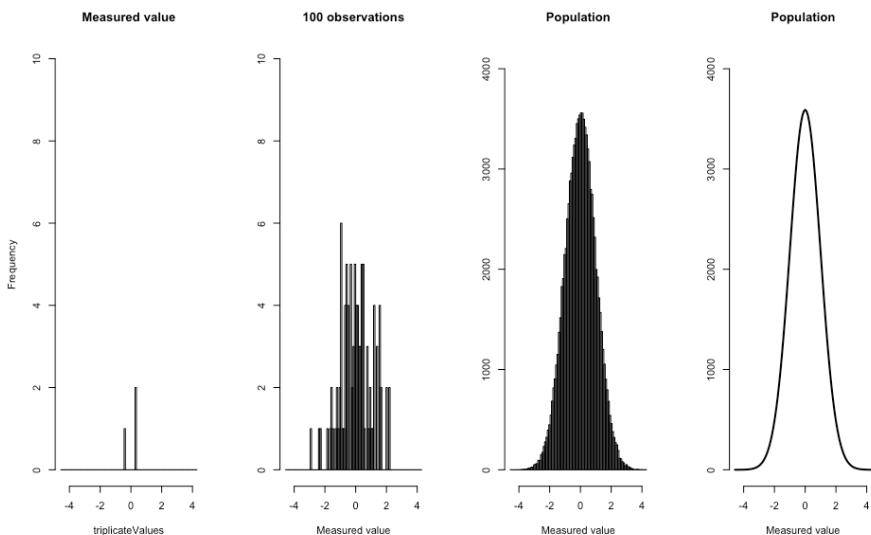
3. 確率分布と照らし合わせる



統計量

棄却限界値によって
規定される面積
(通例: 全体の5%)

確率分布? 面積?



統計的検定の手続き

1. 仮説を立てる

2つのサンプル間で遺伝子発現量
(平均値) の違いがある?

2. 統計量を求める

平均、標準誤差、自由度から
t統計量を求める

3. 求めた統計量を確率
分布に照らし合わせる

t分布からp値を求める

4. 判定: 求めた確率と
棄却限界値との比較

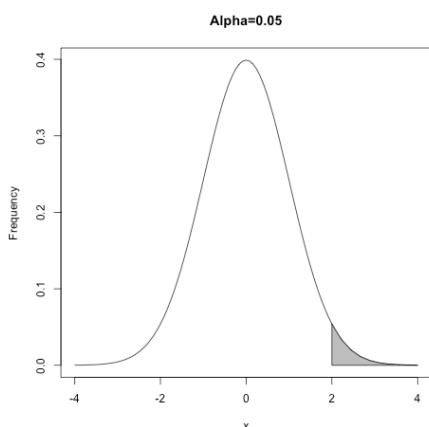
有意差の判定

4. 判定: 帰無仮説が棄却されるか?

帰無仮説

最終的に棄却
される仮説:

「AとBに差が
ある」かを検
定する場合は
「AとBには差
がない」と仮
定する



ポイント

2. 統計量を求める:

統計量: データから導いた
具体的な数値

母数 : 未知の数値

我々ができること: 少数の測定値（標本）から
「母集団」を推定すること

代表値

平均値: 相加平均。すべてのデータを足して、データ数で割って得られる値

- (バー) は 平均を表す
- ^ (ハット) は 推定を表す

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

中央値: データを小さいものから順に並べたときに中央にくる値。データの分布に依存しない。

$n-1$?

なぜ、平均を求める時と分散を求める時では分母が変わるので？

自由度: 統計量を求めるのに使うことができる「独立」な標本数

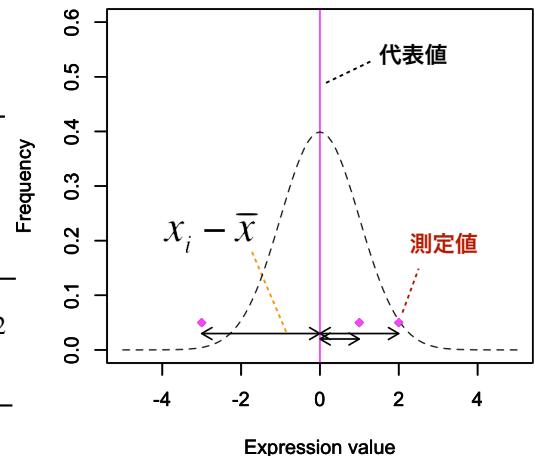
ばらつき: 分散／偏差

分散:

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

標準偏差:

$$\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

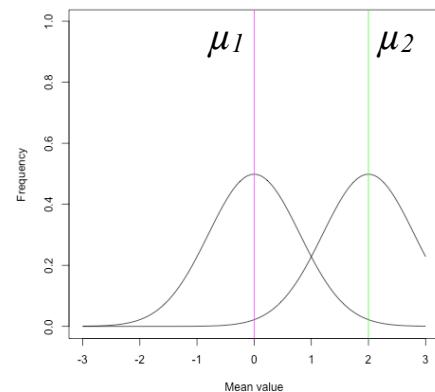


t 検定:

2サンプルの平均の検定

- 平均値 = μ_1, μ_2
- データは正規分布

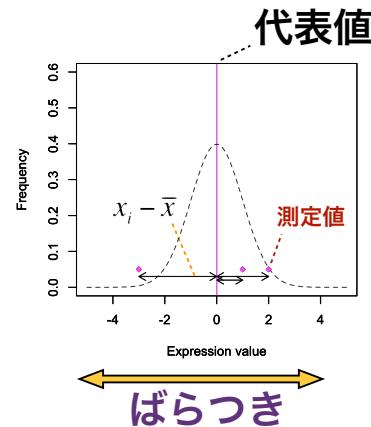
ほぼ全ての検定方法に前提がある



母集団を推定する統計量

1. 代表値

2. ばらつきの範囲



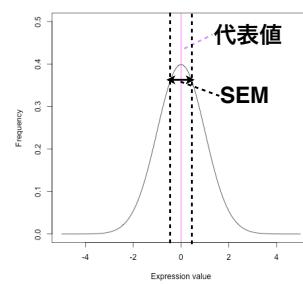
統計量その2: *statistical mind*

平均値もあくまで推定値

(平均) 標準誤差:

「統計量」の偏差

$$SEM = \frac{s}{\sqrt{n}}$$



統計量その1

平均値: 相加平均。すべてのデータを足して、データ数で割って得られる値

$$\bar{x} = \hat{x} = \frac{\sum_{i=1}^n x_i}{n}$$

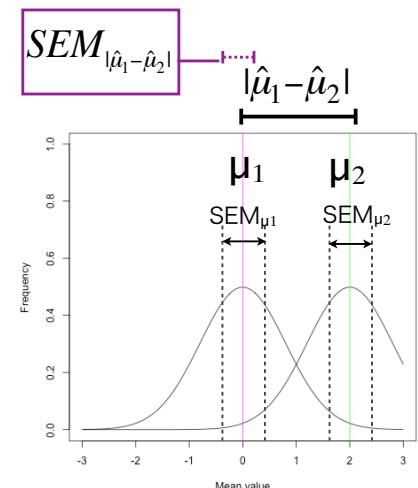
statistical mind

統計量その3:

平均の差とその誤差

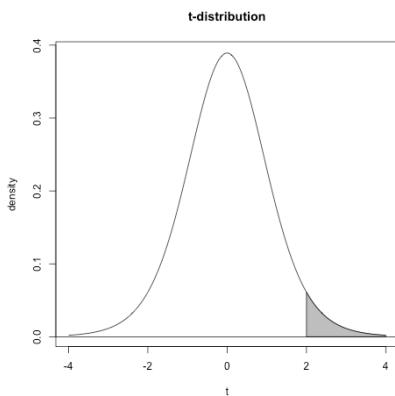
t統計量

$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{SEM_{|\hat{\mu}_1 - \hat{\mu}_2|}}$$



確率分布-t分布

- 得られたt統計量がどのくらいの確率で起きるか
- t分布（確率分布）を標本のt統計量と自由度を使って参照



【おさらい】自由度: 統計量を求めるのに使うことができる独立な標本数

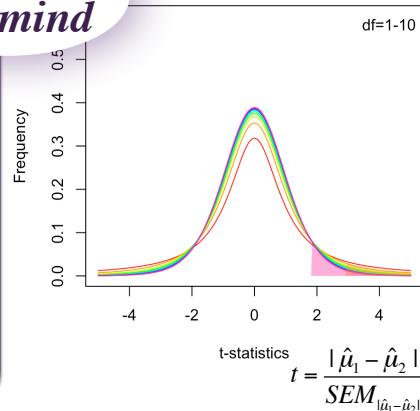
データの分布、仮説検定に即した確率分布を使う

我々の測定では

- 母分散が未知
- したがって確率密度は自由度によって変化

→正規分布ではなく、t分布

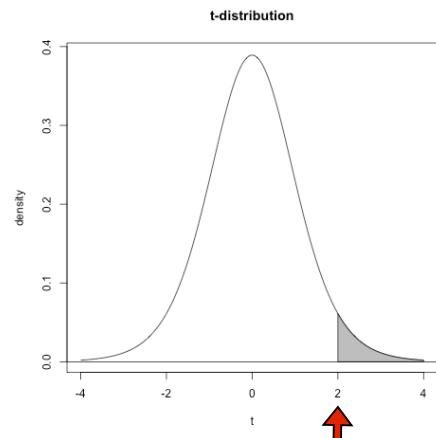
statistical mind



$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{SEM_{|\hat{\mu}_1 - \hat{\mu}_2|}}$$

p値とは：

- 標本に基づいた統計量が帰無仮説の下、起きうる確率
- 汎用される閾値（危険率）：0.05



研究における手続き

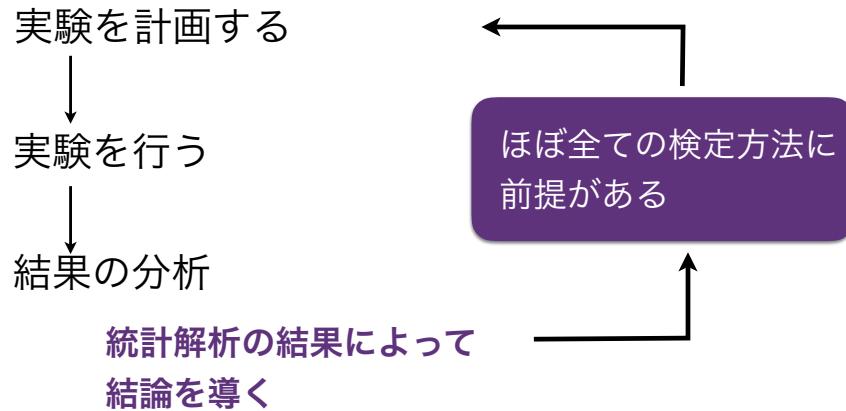
実験を計画する

↓
実験を行う

↓
結果の分析

統計解析の結果によって
結論を導く

現実には: 実験デザインはデータを取得する「前」に練ってある必要がある



ポイント

ほぼ全ての検定方法に前提がある

ポイント

ex. *t* 検定: 正規分布、等分散

どの確率分布を想定する？

連続値: **正規分布**、ガンマ分布 (非負)

離散値 (カウントデータ) :

ポアソン分布 (平均=分散= λ)

負の二項分布 (λ がガンマ分布)

多重検定の補正

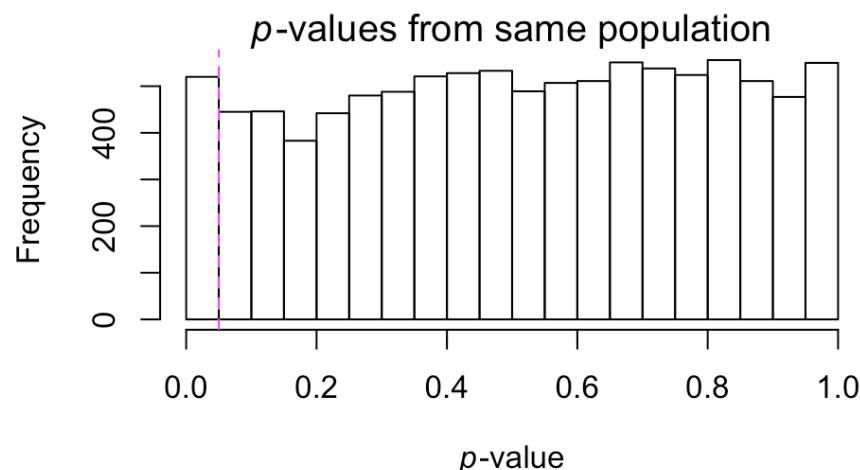
+ 統計検定における重要な概念

*p*値とは:

- 標本に基づいた統計量が帰無仮説の下、起きうる確率
- 汎用される危険率 (閾値) :
0.05 = 100回に5回起きる

同一平均値集団間の*t*検定の繰り返しをシミュレートしてみましょう
source("\$_HOME/data/MS/generate_null_p-values.R")

同一平均値集団間の t 検定でも $p < 0.05$ が得られる



多重検定の補正の必要性

- $p = 0.05$ の検定を 100 回繰り返すと
5 回はランダムに間違い
- NGS 解析では数万回以上繰り返す

多重検定の補正

1. Bonferroni タイプ

2. False discovery rate (FDR):

- Benjamini-Hochberg [R:p.adjust]
- Storey [R:qvalue]

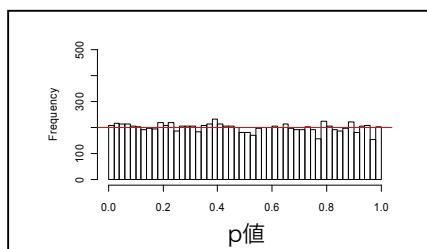
Bonferroni タイプの多重検定の補正

危険率を検定数で調整

$$\text{危険率} = \alpha / k$$

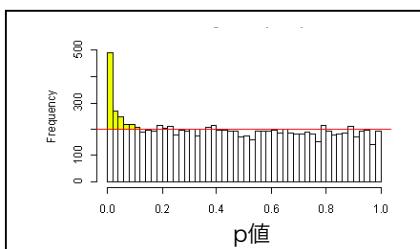
α : 元の危険率、
 k : 検定数

False Discovery Rate (FDR)



帰無仮説

全ての範囲のp値が
同等の頻度で観察される
←どのp値を選んでも
ランダムに選ぶのと同じ



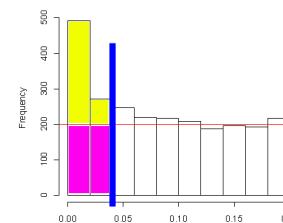
観察

あるp値（閾値）以下のp
値は有意な検定結果である
→では、ランダムに生じて
しまう各p値の頻度は？

False Discovery Rate (FDR)

q値:

補正されたp値。そのq値以
下の検定のうち、どのくら
いの割合でfalse positiveが
含まれているか。



ポイント

p 値、 q 値の違い

p 値の視点: $\text{FP}/(\text{TN}+\text{FP})$

q 値の視点: $\text{TP}/(\text{TP}+\text{FP})$

検定

		+	-
+	+	True positive	False negative
	-	False positive	True negative

復習／発展学習

- p 値とは？
- 統計解析の結果は確率
 - トランск립トーム解析では多数繰り返す
→ 多重検定の補正
 - 多重検定の補正における仮定
例) 時系列データの比較にFDRは使えない

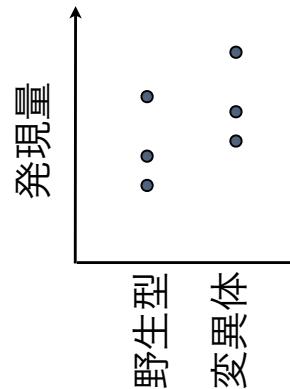
データのばらつきと 実験デザイン・統計学的観点

我々にできる事
少数の測定値（標本）から
「母集団」を推定すること

生体サンプルを繰り返し取る:
biological replicates

同一サンプルを繰り返し測る:
technical replicates

測定データはバラつく



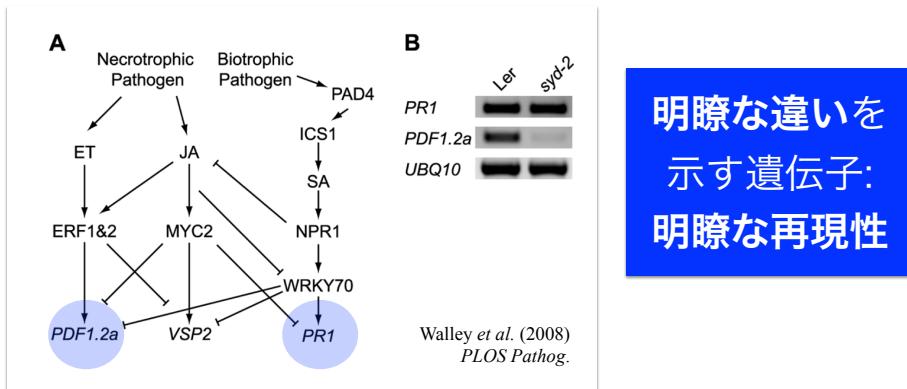
- 実験（測定）を反復する
- 何を「真」と考えるか
- 論文として発表できる
データには**再現性**が必要

定量的測定が可能且つ要求される時代の 再現性のあるデータとは何か？

- 何が再現されうるか？再現されたとするか？
- いつ行っても再現できる？
- どこで行っても再現できる？
- 誰が行っても再現できる？

非NGS測定：“マーカー遺伝子”測定

- 何が再現されうるか？再現されたとするか？



明瞭な違いを
示す遺伝子:
明瞭な再現性

定量的測定が可能且つ要求される時代の 再現性のあるデータとは何か？

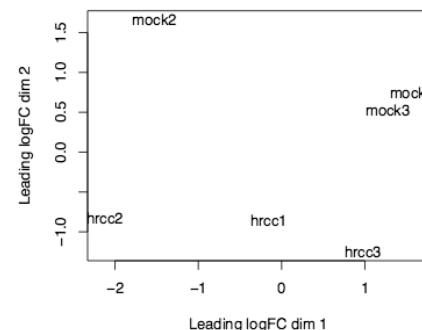
- 何が再現されうるか？再現されたとするか？

- いつ行っても再現できる？
- どこで行っても再現できる？
- 誰が行っても再現できる？

バラつきの
定量と
説明変数への
割当て

“トランск립トーム”測定

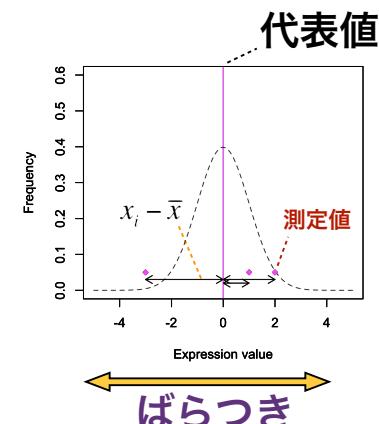
- 何が再現されうるか？再現されたとするか？



網羅的測定:
再現性の
再定義

Chen et al. (2015) edgeR User's Guide page 63

ここまで統計量はサンプルという
一要因のみを考慮



分散分析・線形モデル: 多変数データを系統立てて解析する - 実験デザインと統計の連携

解析の流れ

発現データ（生データ）

前処理: 線形モデル

発現データ（バイアス除去）

個々の遺伝子の解析 —— 全体としての解析

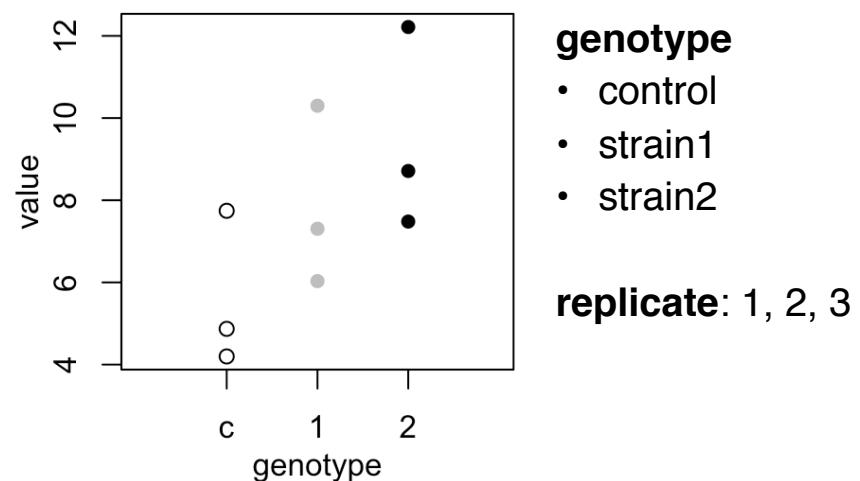
有意差検定

次元圧縮

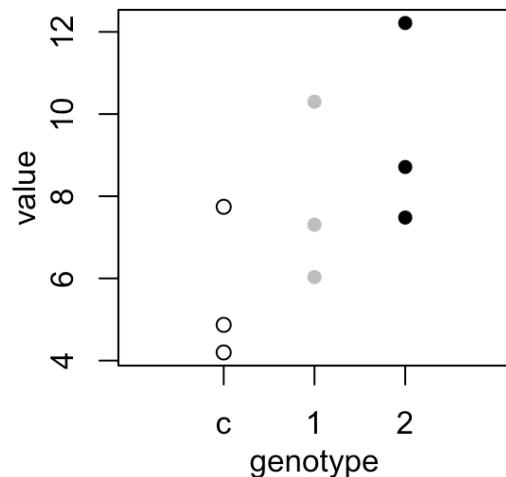
目標

- 線形モデルの概念を掴む
- 実験デザインがどう統計に影響するかを考えるきっかけをつかむ

あるRT-qPCR実験: 生データ



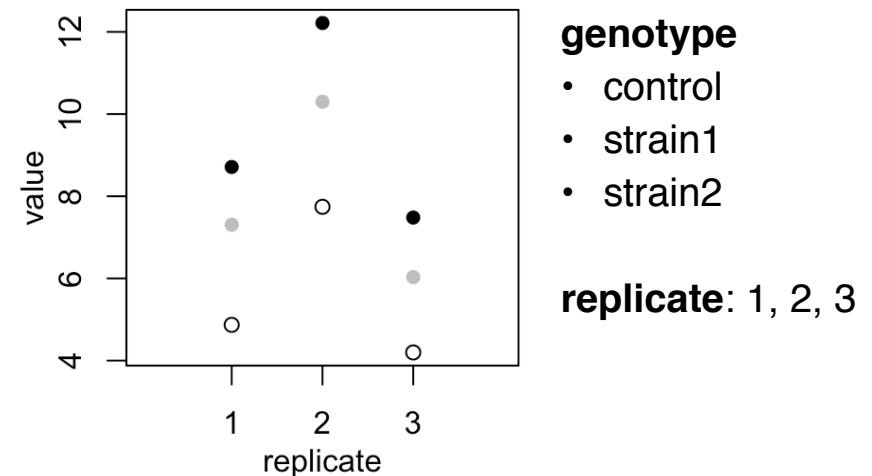
あるRT-qPCR実験: t検定結果



p-values

- control vs strain1
= 0.2456
- control vs strain2
= 0.1011

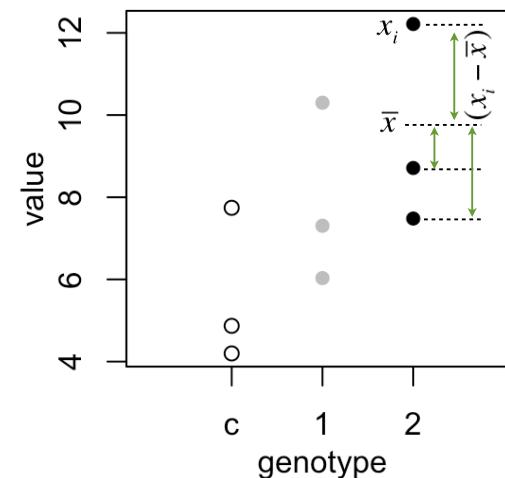
生データをreplicateについて可視化



ポイント

検定から推定（予測・モデル構築）へ：
線形モデルへの転換

線形モデルで考えてみる：モデル表記



$$x_i = \bar{x} + (x_i - \bar{x})$$

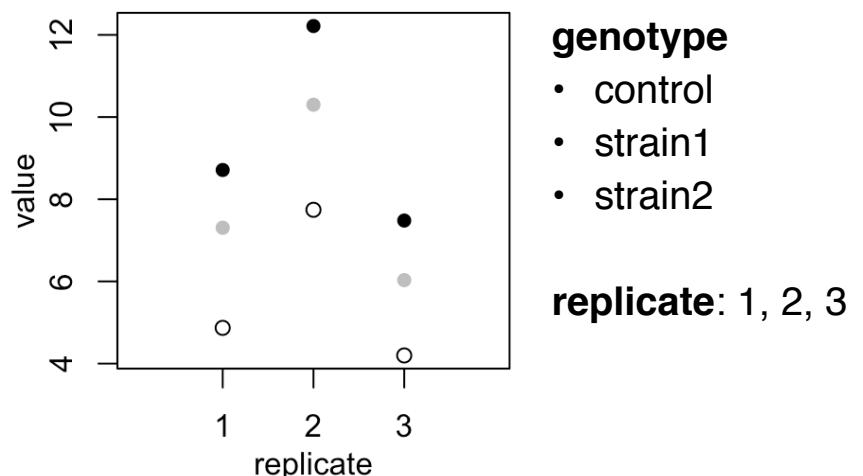
線形モデルで考えてみる：モデル表記

$$x_i = \bar{x} + (x_i - \bar{x})$$

$$x_i = \bar{x} + \varepsilon_i$$

残差 (観察値-推定値):
想定要因では説明できない
データの変動

生データをreplicateについて可視化



観察値を複数要因の
影響に起因するものとして分解

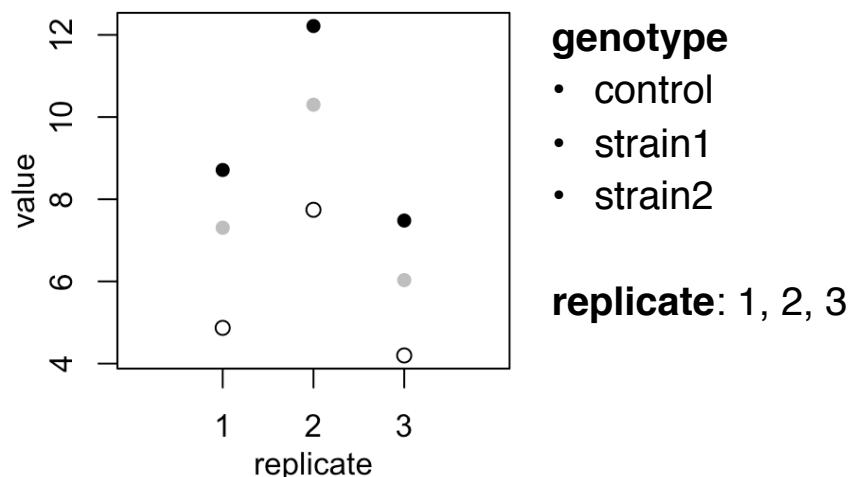
$$x_i = \bar{x} + (x_i - \bar{x})$$

$$x_i = \bar{x} + \varepsilon_i$$

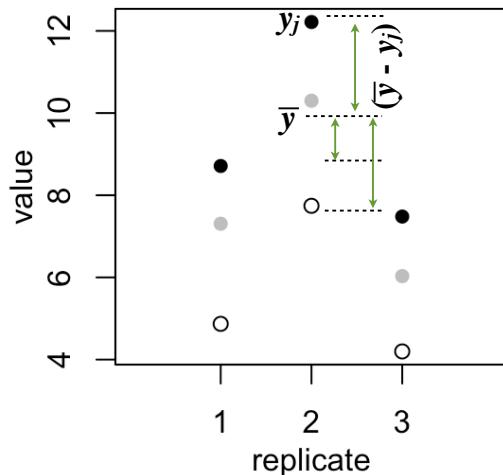
genotypeとreplicateの
影響を同時に
考えられないか？

$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

生データをreplicateについて可視化



replicateの影響も推定



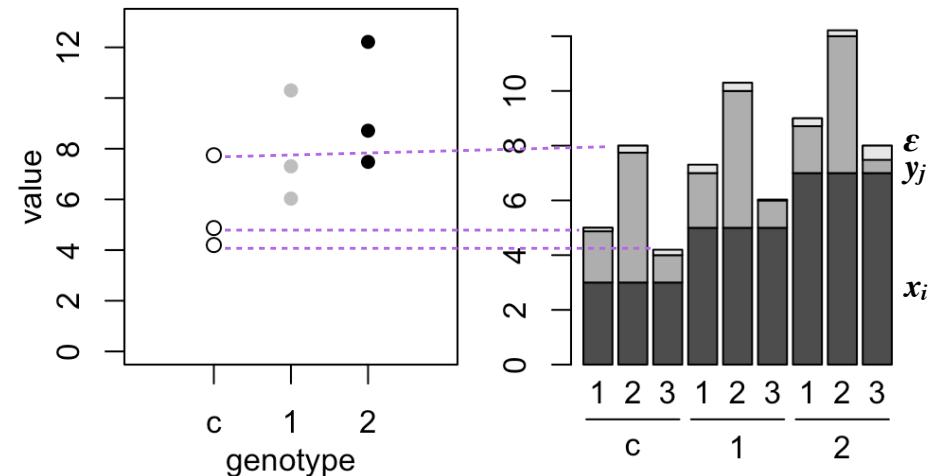
genotype

- control
- strain1
- strain2

replicate: 1, 2, 3

genotype, replicateの影響を

同時に推定する: $O_{ij} = x_i + y_j + \varepsilon_{ij}$



分散分析・線形モデルの枠組み

$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

教科書・論文での書き方

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

応答変数

説明変数

線形モデルとは

応答変数 \sim 説明変数1 + 説明変数2 + ... + 誤差

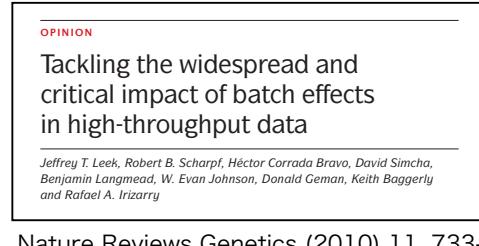
と観察値を説明する（かもしれない）
変数の足し算で応答変数への貢献度を
推定する

- R: lm, glm, glmFitなどの関数を使う

実験デザインの重要性

- omicsデータは“batch effect”と呼ばれる体系的なバイアスが混入する。

例: 実験時期、実験者、餌



Nature Reviews Genetics (2010) 11, 733-

- 線形モデルで推定・除去

実験デザインの重要性

ポイント

- 線形モデルで推定・除去

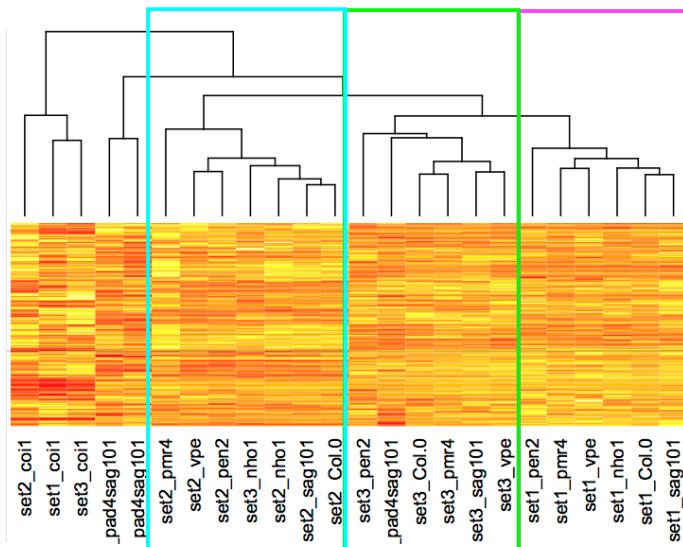
$$O_{ij} = \alpha_i + \beta_j + \varepsilon_{ij}$$

α_i : 遺伝子型／処理など注目している効果の要因

β_j : 反復（実験日時）／実験者などバイアス要因

- α_i の推定値、標準誤差のみを使う

batch effect の トランскRIPTームへの影響



定量的測定が可能且つ要求される時代の 再現性のあるデータとは何か？

- 何が再現されうるか？再現されたとするか？

- いつ行っても再現できる？
- どこで行っても再現できる？
- 誰が行っても再現できる？

系統的な
バラつきの
定量と
説明変数への
割当て

R (edgeR) での実装

```
> x <- read.delim("TableOfCounts.txt", row.names="Symbol")
> group <- factor(c(1,1,2,2))
> y <- DGEList(counts=x, group=group)
> y <- calcNormFactors(y)
> design <- model.matrix(~group)
> y <- estimateDisp(y, design)
> fit <- glmFit(y, design)
> lrt <- glmLRT(fit, coef=2)
> topTags(lrt)
```

Chen, et al., edgeR User's Guide (December 26, 2017)

model.matrixで生成される出力

```
group      <- factor(c(rep("M", 3), rep("H", 3)))
replicates <- factor(c(1:3, 1:3))
model.matrix(~group+replicates)

(Intercept) groupM replicates2 replicates3
1           1       1         0         0
2           1       1         1         0
3           1       1         0         1
4           1       0         0         0
5           1       0         1         0
6           1       0         0         1

attr(", "assign")
[1] 0 1 2 2
attr(", "contrasts")
attr(", "contrasts")$group
[1] "contr.treatment"
```

0と1の行列
contrasts

Rを用いた線形モデルにおける

実験デザイン指定: factor, model.matrix

```
> x <- read.delim("TableOfCounts.txt", row.names="Symbol")
> group <- factor(c(1,1,2,2))
> y <- DGEList(counts=x, group=group)
> y <- calcNormFactors(y)
> design <- model.matrix(~group)
> y <- estimateDisp(y, design)
> fit <- glmFit(y, design)
> lrt <- glmLRT(fit, coef=2)
> topTags(lrt)
```

Chen, et al., edgeR User's Guide (December 26, 2017)

model.matrixで生成される出力

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

	(Intercept)	groupM	replicates2	replicates3
1	1	1	0	0
2	1	1	1	0
3	1	1	0	1
4	1	0	0	0
5	1	0	1	0
6	1	0	0	1

ポイント

線形モデルとmodel.matrixの関係

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

↓
i, jを書き下すと

$$O_{M,1} \sim \mu + \alpha_M + \beta_1 + \varepsilon_{M,1}$$

$$O_{M,2} \sim \mu + \alpha_M + \beta_2 + \varepsilon_{M,2}$$

$$O_{M,3} \sim \mu + \alpha_M + \beta_3 + \varepsilon_{M,3}$$

$$O_{H,1} \sim \mu + \alpha_H + \beta_1 + \varepsilon_{H,1}$$

$$O_{H,2} \sim \mu + \alpha_H + \beta_2 + \varepsilon_{H,2}$$

$$O_{H,3} \sim \mu + \alpha_H + \beta_3 + \varepsilon_{H,3}$$

ポイント

線形モデルとmodel.matrixの関係

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

$$O_{M,1} \sim \mu + \alpha_M + \beta_1 + \varepsilon_{M,1}$$

$$O_{M,1} \sim \mu + \alpha_M + \alpha_H + \beta_1 + \beta_2 + \beta_3 + \varepsilon_{M,1}$$

$$O_{M,1} \sim \mu \times 1 + \alpha_M \times 1 + \alpha_H + \beta_1 + \beta_2 \times 0 + \beta_3 \times 0 + \varepsilon_{M,1}$$

contrasts: 1番目の水準の係数を0として残りと比較

$$O_{M,1} \sim \mu \times 1 + \alpha_M \times 1 + \beta_2 \times 0 + \beta_3 \times 0 + \varepsilon_{M,1}$$

1	1	0	0
---	---	---	---

ポイント

model.matrix, contrasts, 実験デザインの関係

観察数: 6

$$O_{M,1} \sim \mu + \alpha_M + \beta_1 + \varepsilon_{M,1}$$

$$O_{M,2} \sim \mu + \alpha_M + \beta_2 + \varepsilon_{M,2}$$

$$O_{M,3} \sim \mu + \alpha_M + \beta_3 + \varepsilon_{M,3}$$

$$O_{H,1} \sim \mu + \alpha_H + \beta_1 + \varepsilon_{H,1}$$

$$O_{H,2} \sim \mu + \alpha_H + \beta_2 + \varepsilon_{H,2}$$

$$O_{H,3} \sim \mu + \alpha_H + \beta_3 + \varepsilon_{H,3}$$

推定する係数の数: 6

$$\mu, \alpha_M, \alpha_H, \beta_1, \beta_2, \beta_3$$

推定したい係数の数よりも
観察数が多くなくてはならない

contrasts: 1番目の水準の

係数を0として残りと比較
→係数の数を削減

model.matrix, contrasts, 実験デザインの関係

観察数: 6

$$O_{M,1} \sim \mu + \varepsilon_{M,1}$$

$$O_{M,2} \sim \mu + \beta_2 + \varepsilon_{M,2}$$

$$O_{M,3} \sim \mu + \beta_3 + \varepsilon_{M,3}$$

$$O_{H,1} \sim \mu + \alpha_H + \varepsilon_{H,1}$$

$$O_{H,2} \sim \mu + \alpha_H + \beta_2 + \varepsilon_{H,2}$$

$$O_{H,3} \sim \mu + \alpha_H + \beta_3 + \varepsilon_{H,3}$$

推定する係数の数: 4

$$\mu, \alpha_H, \beta_2, \beta_3$$

推定したい係数の数よりも
観察数が多くなくてはならない

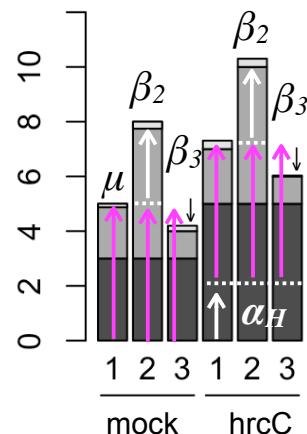
contrasts: 1番目の水準の

係数を0として残りと比較
→係数の数を削減

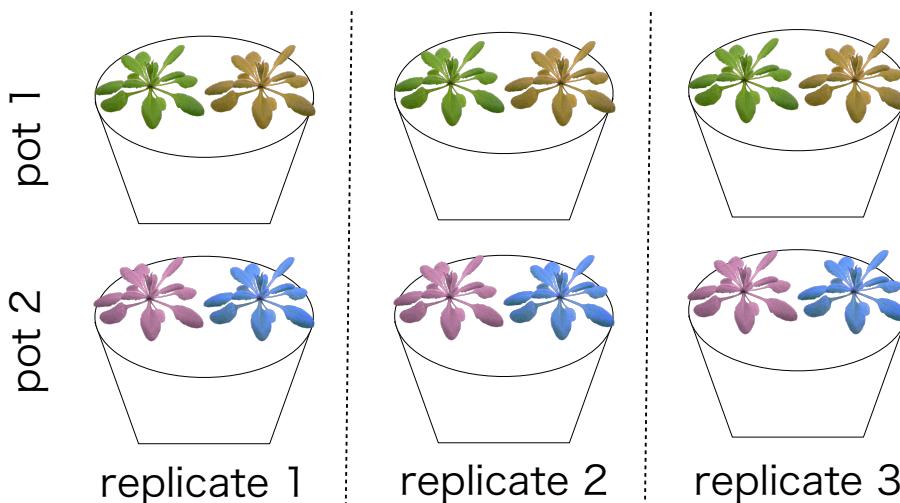
線形モデルにおける係数推定のイメージ：
最小二乗法を使いながら連立方程式を解く

$$\begin{aligned}
 O_{M,1} &\sim \mu + \varepsilon_{M,1} \\
 O_{M,2} &\sim \mu + \beta_2 + \varepsilon_{M,2} \\
 O_{M,3} &\sim \mu + \beta_3 + \varepsilon_{M,3} \\
 O_{H,1} &\sim \mu + \alpha_H + \varepsilon_{H,1} \\
 O_{H,2} &\sim \mu + \alpha_H + \beta_2 + \varepsilon_{H,2} \\
 O_{H,3} &\sim \mu + \alpha_H + \beta_3 + \varepsilon_{H,3}
 \end{aligned}$$

$\mu, \alpha_M, \alpha_H, \beta_1, \beta_2, \beta_3$



実験デザインの重要性:
genotype+replicate+potモデルを当てはめるには？



ポイント

model.matrixまとめ

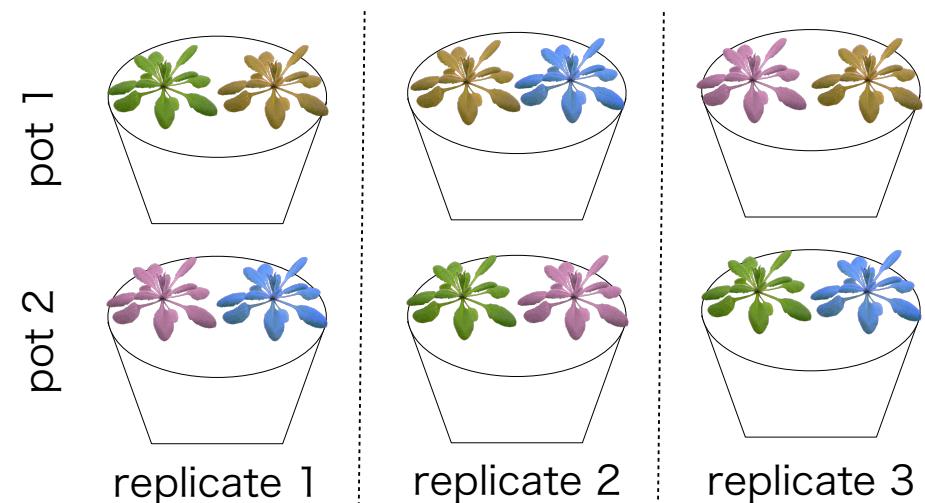
$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

	(Intercept)	groupM	replicates2	replicates3
1	1	1	0	0
2	1	1	1	0
3	1	1	0	1
4	1	0	0	0
5	1	0	1	0
6	1	0	0	1

ポイント

1. 0と1の意味
2. (この場合の) contrastsの概念: μ =replicate1の処理Hの係数
3. 観察数、実験デザインとの関連

実験デザインの重要性:
genotype+replicate+potモデルを当てはめるには？



ポイント

実験デザインの重要性

- 要因効果を推定するための実験デザイン

- 各実験要因を適切に反復させた実験デザイン
(発展学習: 無作為化)

- 実験デザインとモデル

- 要因: データ取得「前」に想定しておくもの
- データの変動を説明しない要因を解析時に減らすことは可能。一方、実験デザイン時に計画しなかった要因を増せない。

(少しだけ) 線形モデル→一般化線形モデル

[予測]

実現象に即し、データにあてはまるモデル

どの確率分布を想定する？

連続値：正規分布 [R: lm]

離散値（カウントデータ）：

負の二項分布 [R: glmFit, glm.nb]

まとめ

- 計測データセットに影響を与える要因が一つではない場合、分散分析・線形モデルの枠組みが有効
- 理論を理解するのは難しいかもしれないが、実行はRで簡単に行える。理解に努める努力と実験デザインと連動したモデルを立てることが重要

復習／発展学習

- 回帰（最小二乗法）: contrast、切片
- 実験計画法
- 交互作用
- Bioconductor: limma、edgeRパッケージ