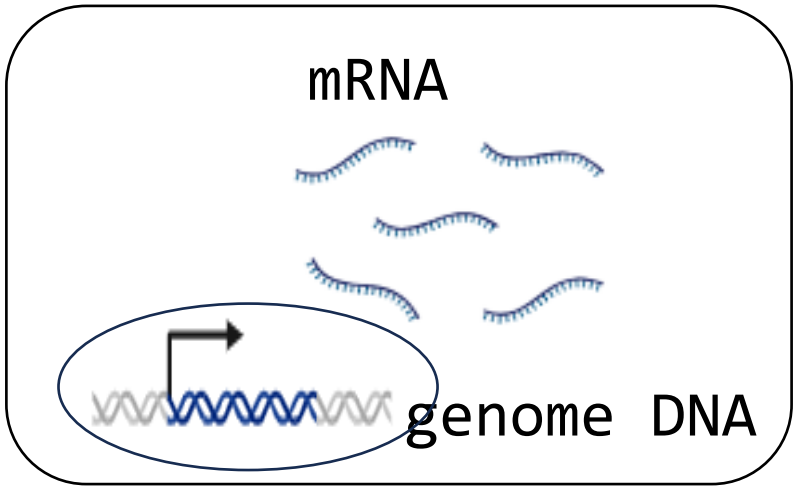
RNA-seq

東京科学大学　ILA国府台
中林潤

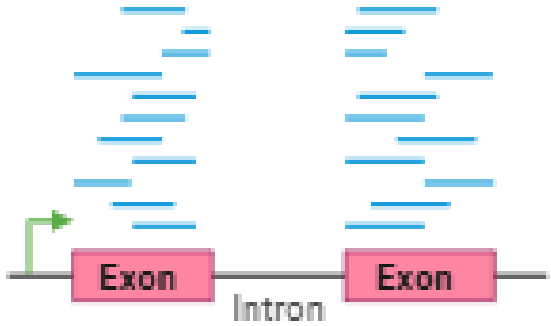
JH人材育成課　バイオインフォマティシャン育成講座　⑨

RNA-seq

- 細胞内に発現している全ての遺伝子の発現量を測定する方法



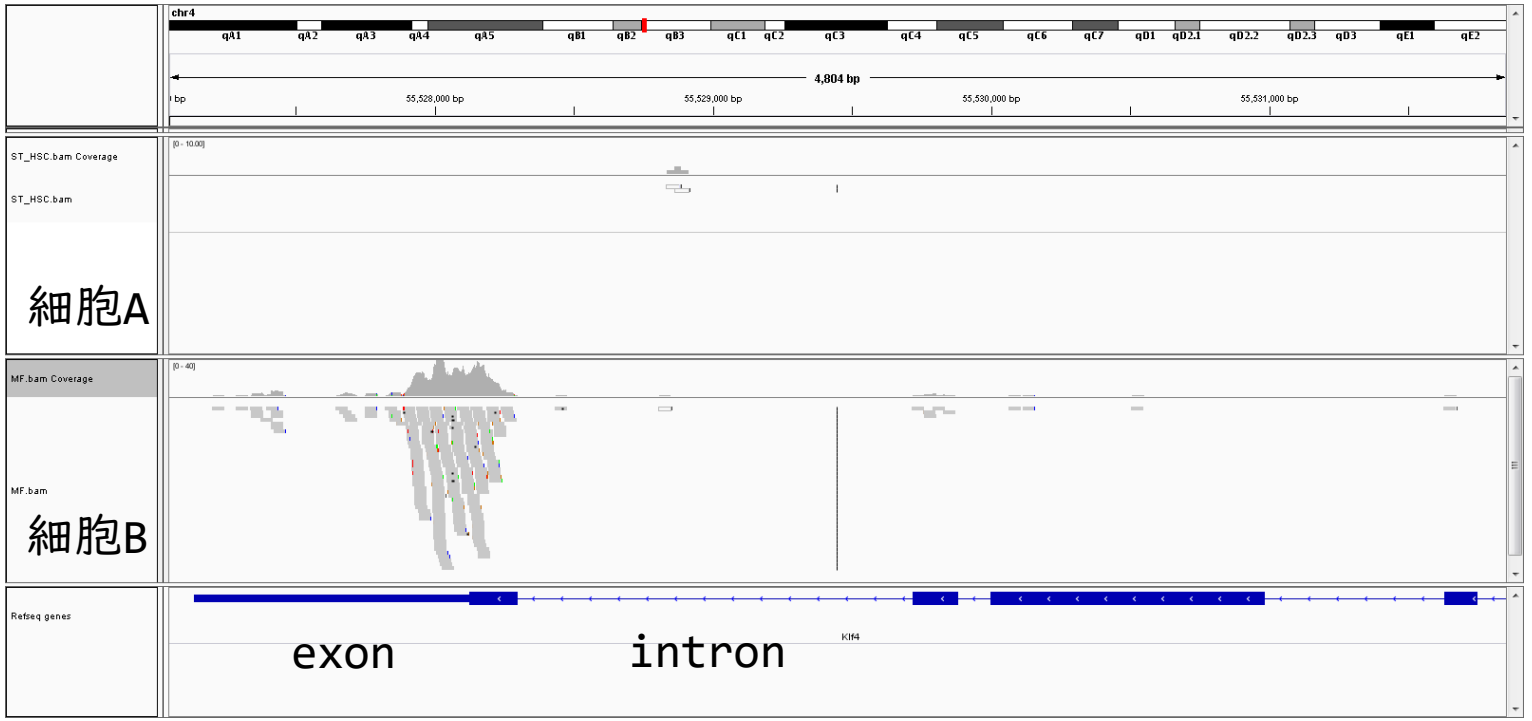
mRNAを抽出
　oligo dT
　rRNA除去
断片化
ライブラリを作成
次世代シーケンサー



ゲノム配列にマッピング
リード数を発現量として定量

RNA発現量

細胞からmRNAを抽出→ライブラリ作成→次世代シーケンサー



リード数をカウント→発現量

発現量の算出

遺伝子A　　遺伝子B

遺伝子長の長い遺伝子には多くのフラグメントがマップされる

サンプル①

遺伝子A

サンプル②

遺伝子A

総リード数が多ければマップされるフラグメント数も多くなる

FPKM（Fragments per kilobase of exon per million reads mapped）

- マップされた総リード1000000当たり、1kb当たりのフラグメント数

- 総リード数と遺伝子長で補正したカウント数
  総リード数を100万、遺伝子長を1Kbに補正する
  $FPKM_i = \frac{q_i}{\frac{\sum_j q_j}{10^6} \times \frac{l_i}{10^3}}$
  q_i：カウント数
  l_i：遺伝子長

- 正確に発現量を反映しないケースもあるので注意が必要

TPM（Transcripts Per Million）

- FPKMの代替として提案された補正法
- 総リード数1000000、遺伝子長1Kb当たりのリード数
- 計算方法がFPKMと異なる

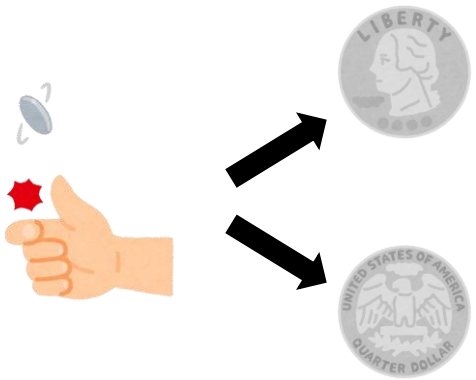
$$A_i = \frac{q_i}{l_i} \times 10^3$$

$$TPM_i = \frac{A_i}{\sum_j A_j} \times 10^6$$

先に遺伝子長で補正して、その後総リード数で補正する。

ベルヌーイ試行

ベルヌーイ試行：二つの結果だけが可能な独立な試行を繰り返すとき、
結果の起こる確率が試行を通じて一定であるような試行



表：$\frac{1}{2}$

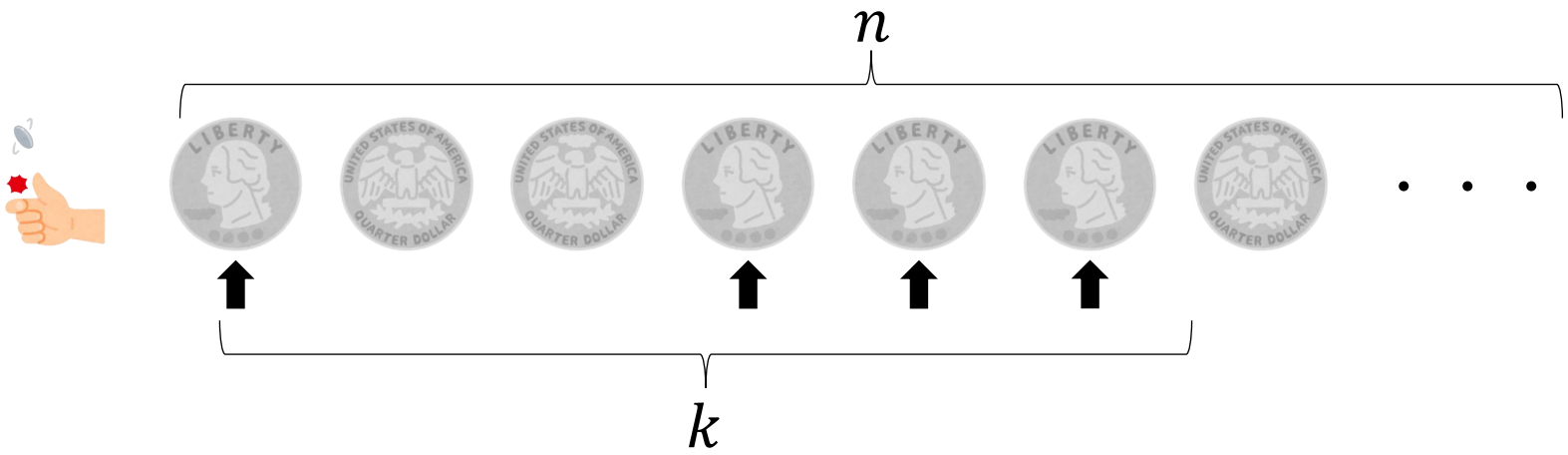
裏：$\frac{1}{2}$

成功：p
失敗：$1-p$

二項分布

二項分布：ベルヌーイ試行をn回独立に行ったとき、k回成功する確率

成功：p
失敗：$1-p$



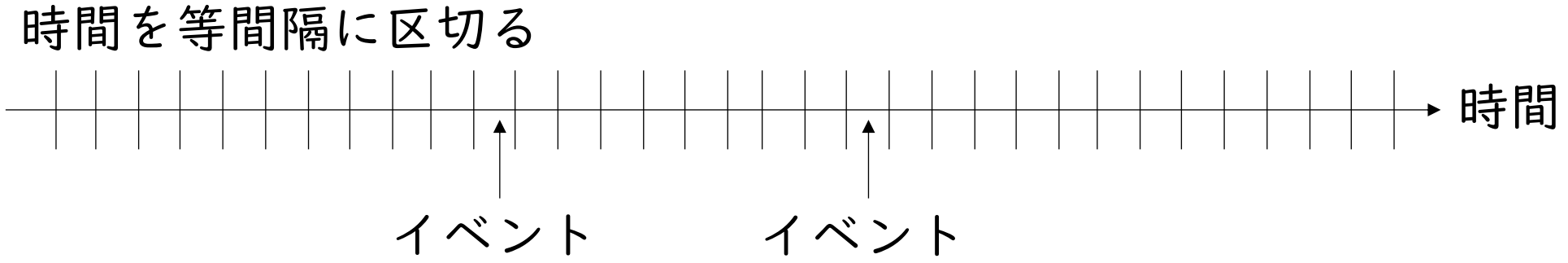
$$B_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \binom{n}{k} = {}_nC_k = \frac{n!}{k!(n-k)!} \text{のこと}$$

ポアソン分布

- 平均でλ回起こる稀にしか起こらない事象が、十分長い時間経過する間にk回起こる確率

- 二項分布の生起確率pが非常に小さく、試行回数nが非常に大きい場合に相当する

- イベントが起きる（成功）、起きない（失敗）のベルヌーイ試行を繰り返す



偶然起こる事象が従う確率分布

ポアソン分布の確率密度

二項分布

$$B_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\lambda = np$$

$$p = \frac{\lambda}{n}$$

$$n \to \infty$$

$$B_{n,p}(k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{n(n-1)\cdots}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{n(n-1)\cdots(n-k-1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= 1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}$$

$n \to \infty$のとき

1

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$$

1

$$P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

ポアソン分布の期待値と分散

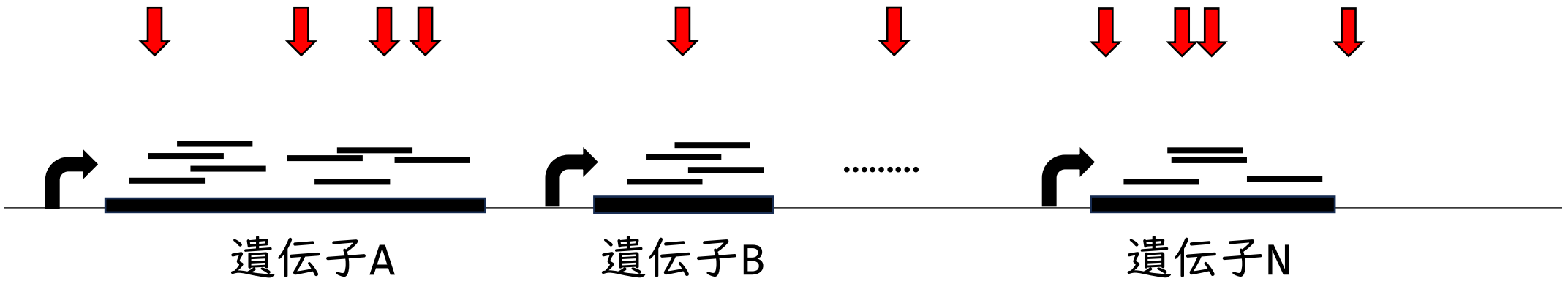
$$E[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda}$$
$$= \lambda \sum_{k=0}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}$$
$k' = k - 1$とおく
$$= \lambda \sum_{k=0}^{\infty} \frac{\lambda^{k'}}{k'!} e^{-\lambda}$$
$$= \lambda$$

$$V[X] = E\left[X^2\right] - (E[X])^2$$
$$= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2$$
$$= \sum_{k=0}^{\infty} \{k(k-1) + k\} \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2$$
$$= \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} e^{-\lambda} + \lambda \sum_{k=0}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} - \lambda^2$$
$$= \lambda^2 + \lambda - \lambda^2$$
$$= \lambda$$

ポアソン分布の期待値と分散はともにλ

発現量の分布の推定

ランダムにフラグメントがゲノム上に分布すると考える



ある遺伝子にフラグメントがマッピングされるのは偶然による事象

ポアソン分布の問題点

- 期待値と分散が等しい
- 遺伝子の発現量は分散が大きい
- 遺伝子発現量の分布がポアソン分布に従うと仮定するのは不適切

負の二項分布

- 成功確率pのベルヌーイ試行で、r回成功するまでにk回失敗する確率

$$P(k) = \binom{r+k-1}{k} p^r (1-p)^k$$

- 期待値

$$E[k] = \frac{r(1-p)}{p} = \mu$$

- 分散

$$V[k] = \frac{r(1-p)}{p^2} = \mu + \phi\mu^2$$

期待値<分散となるような分布を仮定することができる。

Benjamini-Hochberg法（BH法）

- False Discovery Rate（FDR）：棄却した帰無仮説の中で、正しい帰無仮説の割合

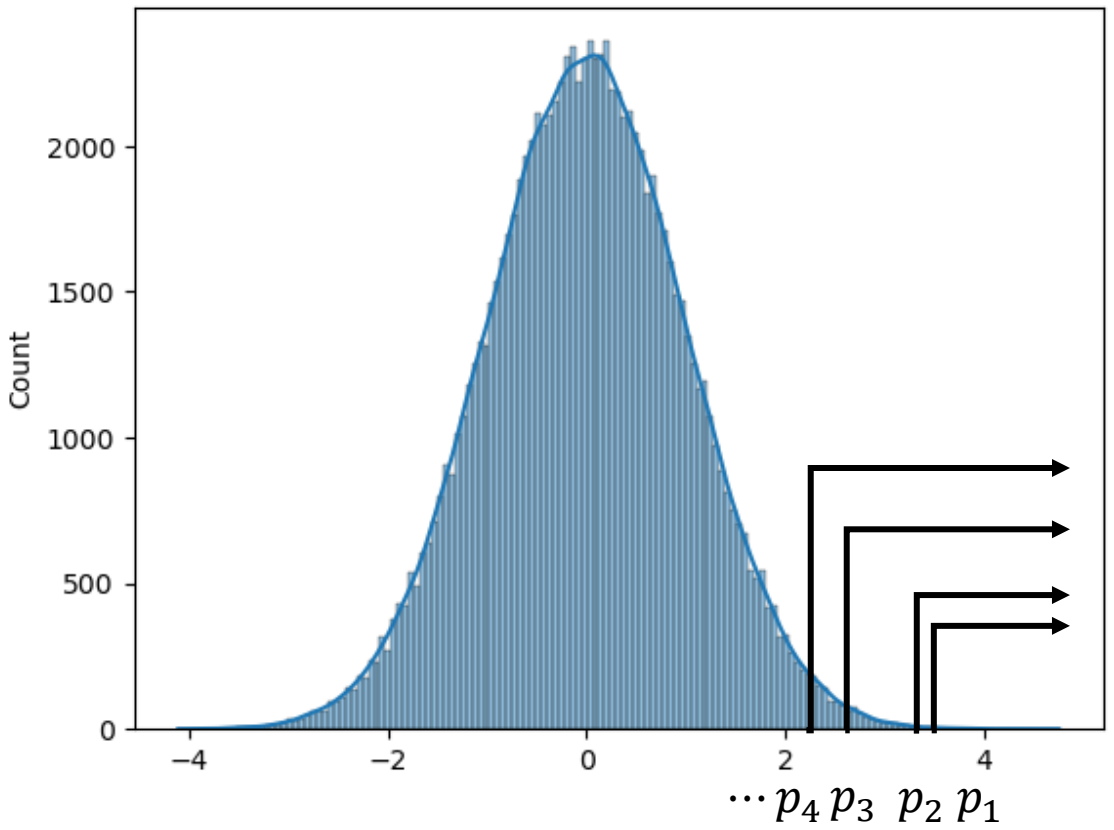
		真実	
		帰無仮説は正しい	帰無仮説は誤り
検定	帰無仮説を棄却	False Positive（FP）	True Positive（TP）
	帰無仮説を採用	True Negative（TN）	False Negative（FN）

$$FDR = \frac{FP}{FP + TP}$$

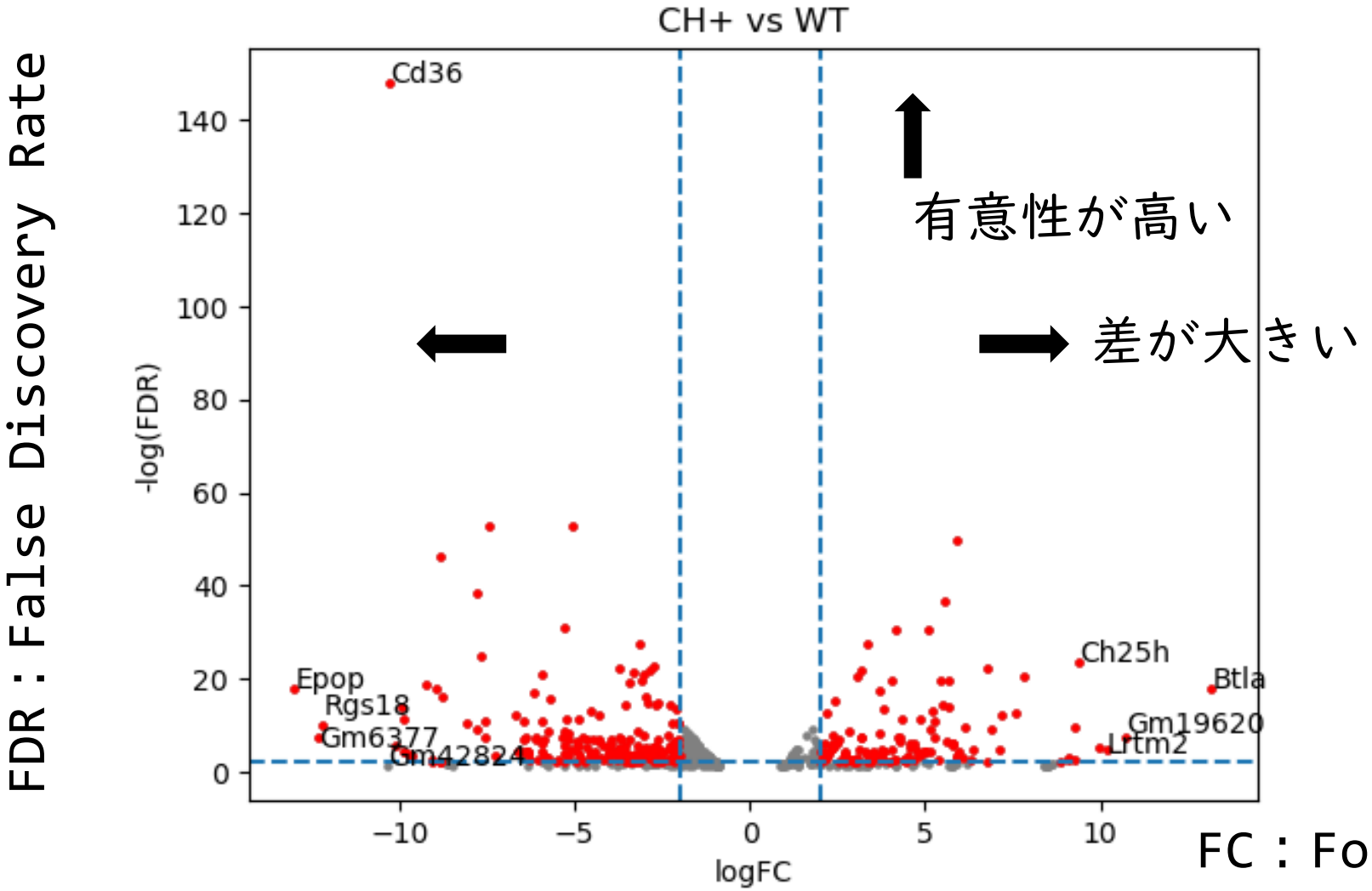
- Benjamini-Hochberg法：FDRをある値以下に抑える方法
- False Positiveをある程度許容するという考え方

Benjamini-Hochberg法

- N回検定を行った場合
- p_iを小さい順に並べる
- $q_i = \frac{Np_i}{i}$を計算
- q_i <FDR閾値であれば有意とする
- q_iが最大となるものまでを有意とする



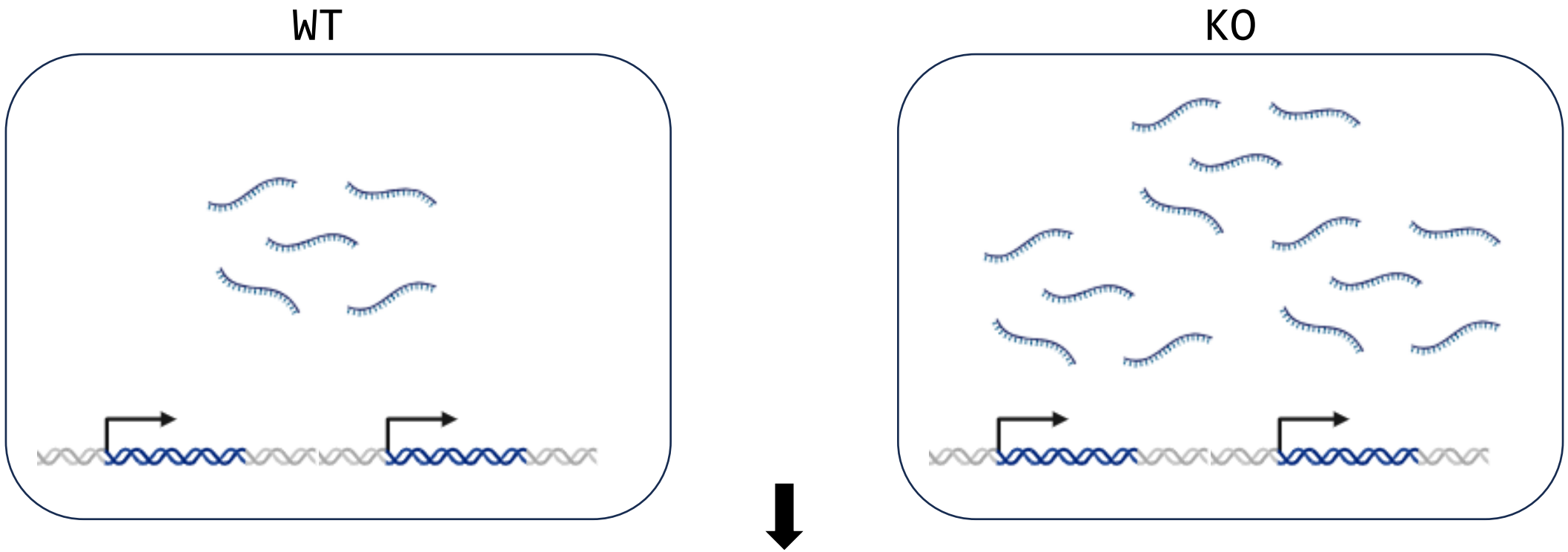
発現変動遺伝子の抽出



volcano plot

FC：Fold Change
sample①/sample②

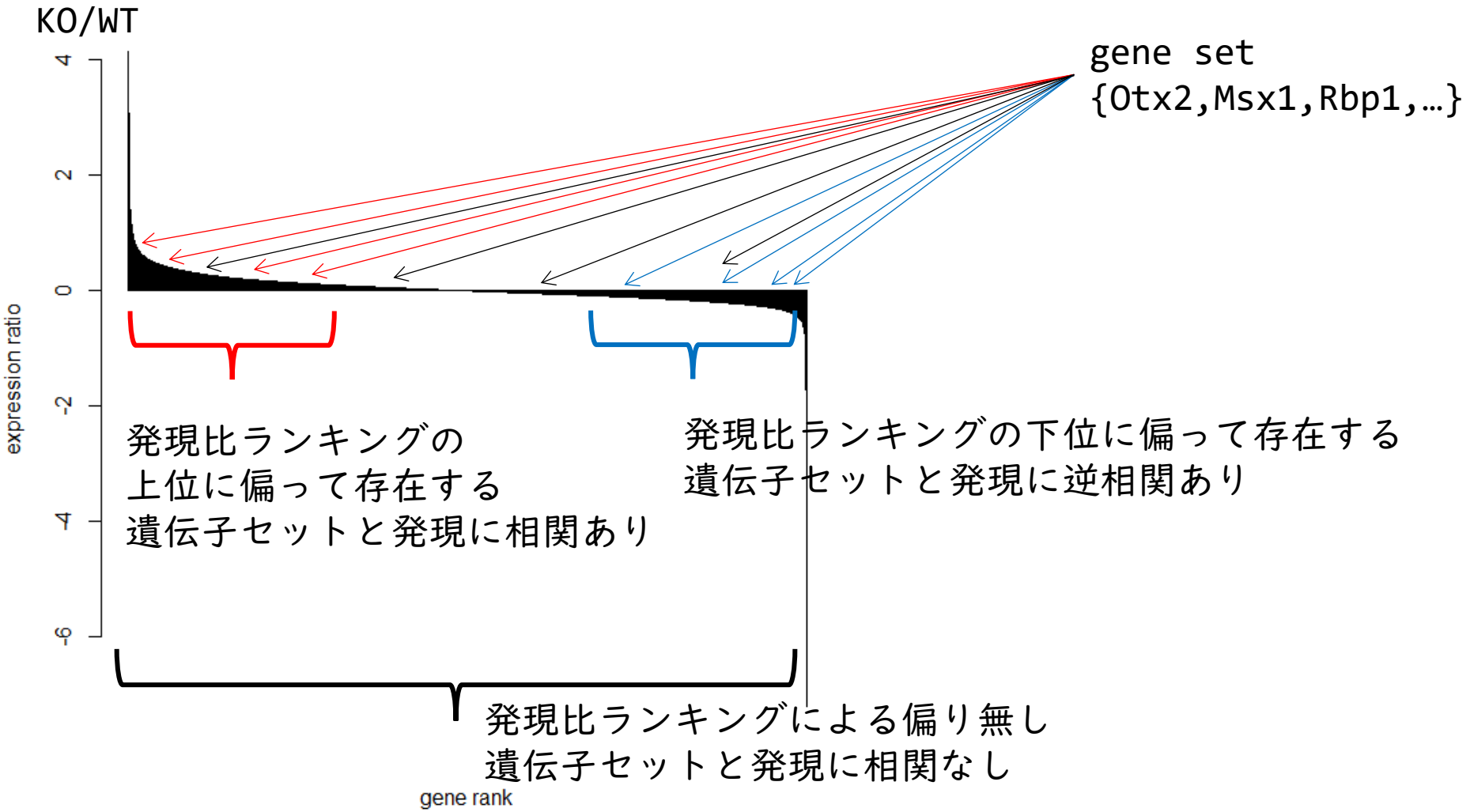
発現変動遺伝子の機能解析



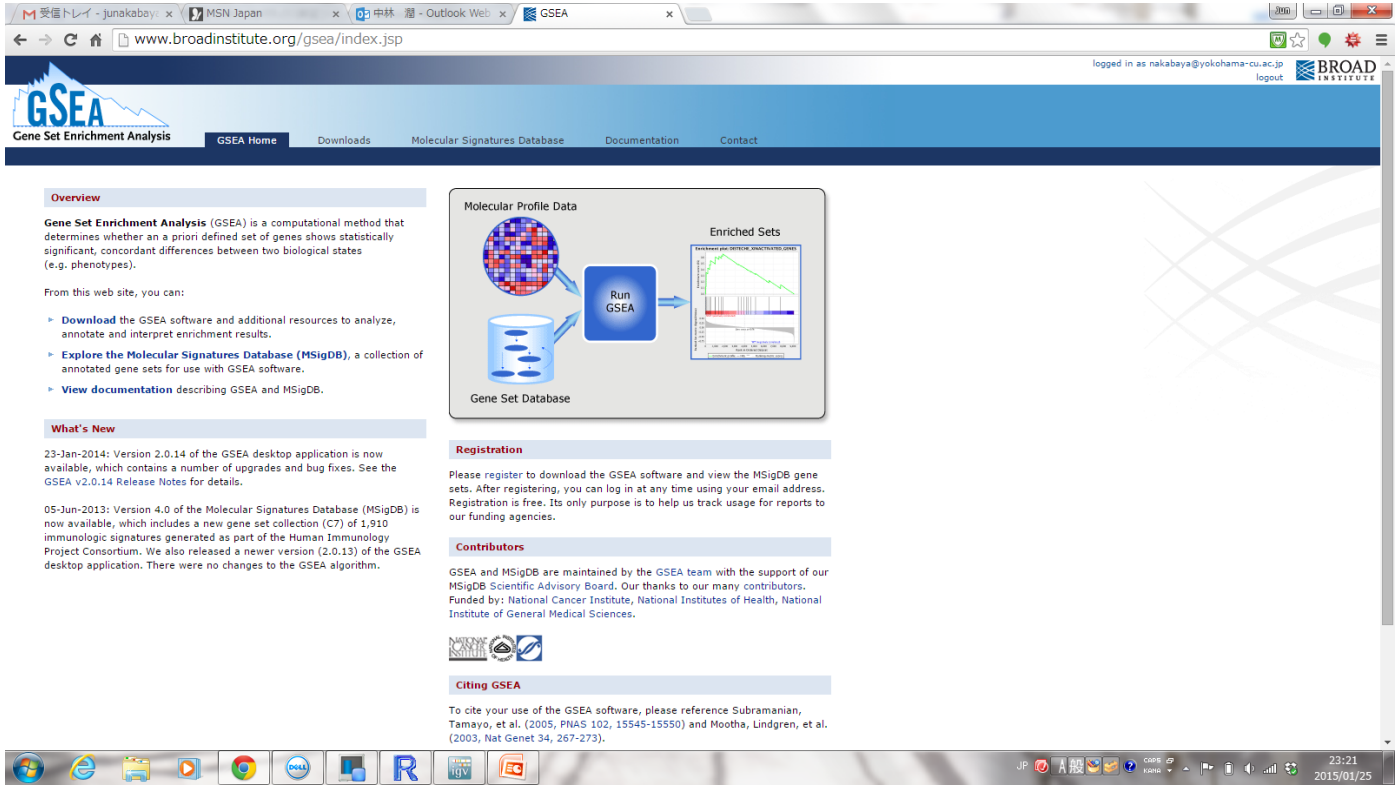
発現変動遺伝子を抽出
発現量が変動した遺伝子の機能を推定したい

Gene Set Enrichment Analysis（GSEA）

- 特定の遺伝子セットと発現比の間に相関があるか調べる

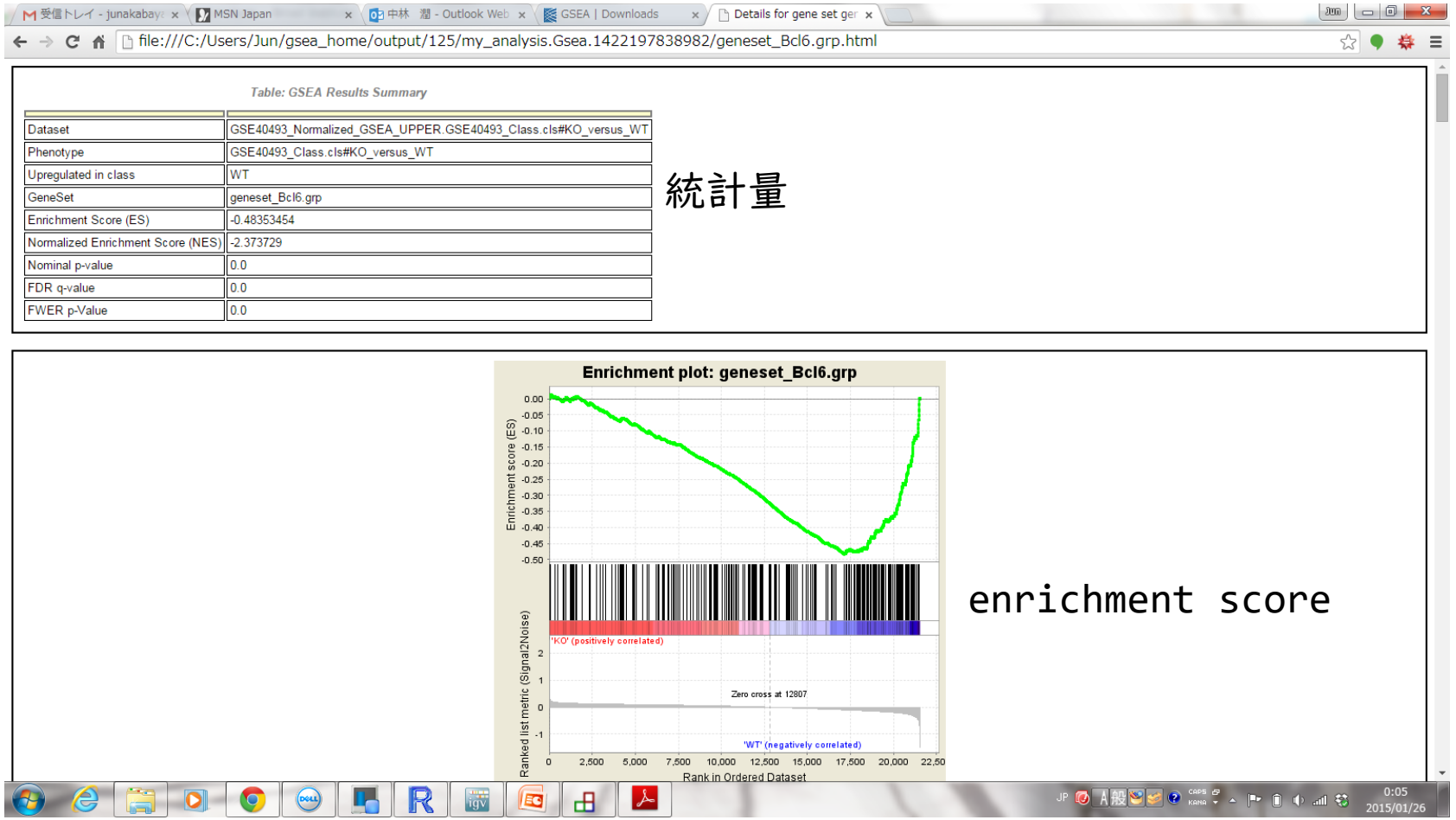


Broad Institute



http://www.broadinstitute.org/gsea/index.jsp

結果の表示



まとめ

- マッピングされたフラグメントのカウント数を発現量に換算する。

- TPM

$$A_i = \frac{q_i}{l_i} \times 10^3$$

$$TPM_i = \frac{A_i}{\sum_j A_j} \times 10^6$$

- 遺伝子発現量は負の二項分布に従うと仮定する。

- 有意性の検定にはFDRを用いる。

まとめ

- GSEA（Gene Set Enrichment Analysis）
- 特定の機能に関わる遺伝子セットと、二群間の発現比の関係を調べる。
- 遺伝子発現の変化が特定の機能と関係しているかどうか推定する。