

オープンソフトウェア  
**Bioconductor**  
の紹介

三共株式会社  
田崎 康一

# Bioconductor what?

- Implemented on R
  - R is poorman's S (要するに GPL な S)
    - だから R に付随する膨大な統計パッケージが使える
- Specialized for (statistical) Analysis in Biology
  - Microarray, SAGE, CGH
  - MS 系も入れたけど, まだ道半ば
- プログラマ向け
  - マクロが強力な反面, GUI は弱い
  - GeneSpring や SpotFire と併用がお勧め

# Bioconductor why?

- Reproducibility of Analysis
  - ソースコードが公開されてるので traceable
    - Excel のマクロは曲者. マニュアルと齟齬がある
      - FTEST は片側 P値とあるが両側 (だから有意なのを見落とすかも)
    - 旧バージョンが入らないので reproducibility の保証にはならないかも
- 最新の解析手法
  - 誰かが書いてくれるはず (or 自分で書く)
- 綺麗なグラフ
  - Ref. <http://www.bioconductor.org/Screenshots/index.html>

# NOT Bioconductor why?

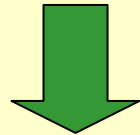
- スプレッドシートっぽく使いたい
  - 可能だが, そういう時は エクセルが一番!
- Word にグラフ・解析結果を張り込みたい
  - 可能だが Excel や SpotFire には敵わない
    - ・ LaTeX 使いなら問題ないけど.....
- グラフとデータを行ったり来たり
  - SpotFire などが良いと思う
- GUI でぐりぐり
  - イマイチ. そもそも GUI は目的じゃない

# Bioconductor is **NOT** 統計ソフト

- データ読み込みツール & 基礎データ集
  - Affymetrics の .cel ファイル読み込み
  - GO アノテーション
  - とは言え良く使う統計・表示モジュールは便利
    - GO term との association
    - FDR (false discovery rate)
    - plot.mat
    - Normalize ???
  - 要するに, R の統計パッケージを使うためのソフト

# Bioconductor summary

- Bio 系基礎データ・アプリ for R
- 一括処理・最先端技術に強い
- GUI は弱いし, 他アプリとの連携も弱い



やることが決まった後の一括処理が最適

あと凝りまくったグラフも.....

# Bioconductor の利用例

- CHP ファイルから MAS5 データを生成
  - データ読み込み
- ALL/AML を分類するマーカ遺伝子探索
  - R による統計解析との融合
  - 基礎的データ(GO) の利用

# .CHP から MAS5

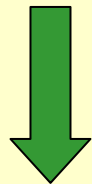
大昔(2001年頃)に取ったデータは Mas4  
今は Mas5 の時代だし, こっちのほうが良いから  
昔のデータを Mas5 に直したい

```
esets <- ReadAffy()  
exprs <- mas5(esets, sc=100)  
exprs2excel(exprs,  
"output.csv")
```

これだけで, フォルダ内にある全ての .chp ファイルを  
Mas5 (平均 100 にスケールして) 値を output.csv に出力する

# ALL/AML を分類する遺伝子探索

ALL/AML は鑑別が難しい & 療法が異なるので  
簡単に鑑別する方法が欲しい



Golub et.al. のデータを使って

R による統計解析

Random permutation を使って false positives を省く例



GOHyperG を使った生物学的解釈

DNA modification (GO:0006304 ALL>AML) や  
Immune cell chemotaxis (GO:0030595 ALL<AML) な  
遺伝子が多い

# Random Permutation

```
nLoop <- 1000
countData <- data.frame(list(count=rep(0, 7129)))
for(i in 1:nLoop){
  t1 <- sample(allIndex, length(allIndex)/2)
  t2 <- sample(amlIndex, length(amlIndex)/2)
  tmp <- apply(exprs(golubTrain), 1, function(x){
    w <- wilcox.test(x[t1], x[t2], alternaitve="two.sided")
    c(w$p.value, median(x[t1]), median(x[t2]))
  })
  tmp <- t(tmp)
  tNames <- rownames(tmp[tmp[,1]<0.05,])
  countData[tNames,1] <- countData[tNames,1] + 1
}
markers <- names(sort(names(countData[,1]),
                      decreasing=TRUE)[1:50])
```

こんな感じで ALL/AML を分類するマーカを求める

# GOHyperG の利用

どんな機能の遺伝子が多くマーカに含まれているのかを計算

```
tmpLL <- as.numeric(mget(markers, env=hu6800LOCUSID))  
myLL <- unique(tmpLL[is.na(tmpLL)==FALSE])  
xx <- GOHyperG(myLL, "hu6800", "BP")  
for(x in names(xx$pvalues[xx$pvalues<1e-2])){  
  cat(x, xx$pvalues[x], "¥n")  
}
```

太字部分が Bioconductor 提供部分



DNA modification (GO:0006304 ALL>AML) や  
Immune cell chemotaxis (GO:0030595 ALL<AML) な  
遺伝子が多い

# 計算結果の表示



Bioconductor 提供の plot.mat と R 付随の hclust を利用

# その他

- 解析(手法)に凝るより, 測定・サンプル調整に手をかけた方が良い
  - 良いデータに勝るものなし
- Bioconductor (R) はバッドノウハウの山
  - 「奥が深い」症候群に注意
- 提供データにおかしなコトも.....
  - YEASTGO2PROBE と  
YEASTGO2ALLPROBES
    - オープンソースの強みを生かすべき

# まとめ

- Bioconductor は統計ソフトじゃない
  - 統計処理をするために必要な機能と
  - 基礎的なデータと
  - 処理後の表示機能
- バッチ処理に威力を発揮
- 統計処理の前段階では別のソフトが良い
- 大抵の解析は bioconductor (+R)で充分