



2017年 第一回バイオインフォマティクス実習

発現変動遺伝子の抽出
二群間の発現プロファイル比較

先端医科学研究センター
バイオインフォマティクス解析室 中林潤

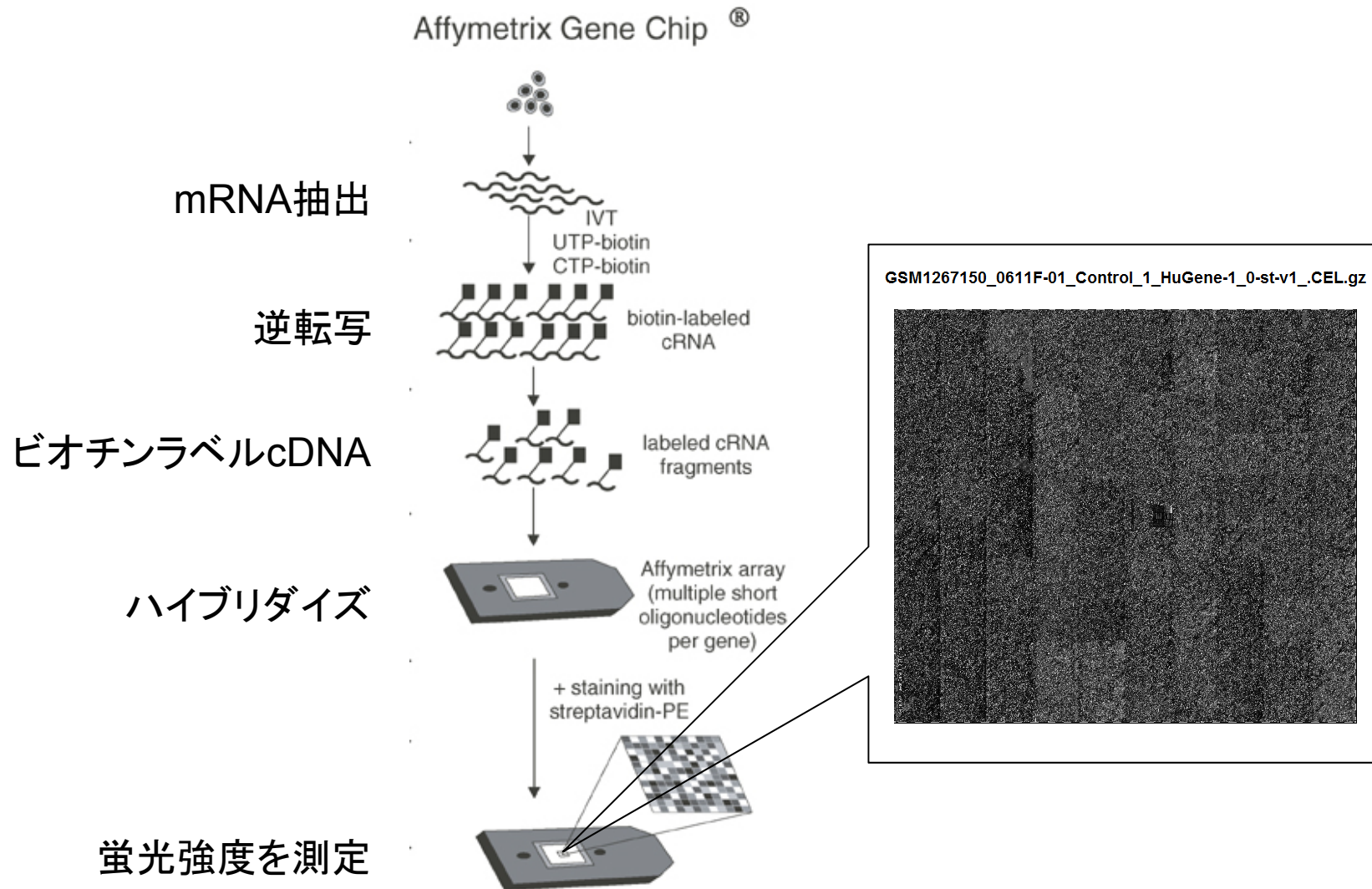
実習の進め方

- サブモニターで講師の作業を見る
- 自分の手元のPCで同じ手順を実行する
サブモニターと配布資料を参考にしてください

本日の実習内容

- GEOデータベースから発現プロファイルを取得
- 統計解析ソフトRを使ってファイルの読み込み
- 二群間で比較し、発現変動遺伝子を抽出

microarray



GEOデータベース検索

http://ncbi.nlm.nih.gov

GSE52452を入力して検索

The screenshot shows the NCBI website interface. At the top, there is a search bar with the text "GSE52452" entered. Below the search bar, a dropdown menu is open, showing a list of databases. The "GEO DataSets" option is highlighted in blue. An arrow points from the text "GEO Datasets を選択" to this highlighted option. Another arrow points from the text "GSE52452を入力して検索" to the search bar. The website header includes "NCBI Resources" and "How To" menus. The left sidebar contains a "Resource List (A-Z)" with various categories like "Chemicals & Bioassays", "Data & Software", etc. The main content area features a "NCBI Facebook page" widget. The right sidebar includes "Popular Resources" (PubMed, Bookshelf, etc.) and "NCBI Announcements". The bottom of the page shows a Windows taskbar with various application icons.

GSE52452

- Fibroblast growth factor receptor 3 interacts with and activates TGF β -activated kinase 1 tyrosine phosphorylation and NF κ B signaling in multiple myeloma and bladder cancer.

Salazar L, et al. 2014 PLoS One 9(1):e86470

- MGH-U3 (膀胱がん細胞株)
- Control vs TAK1 siRNA

データダウンロード

Experiment type: Expression profiling by array

Summary: The NF- κ B transcription factor is constitutively active in a number of hematologic and solid tumors, and many signaling pathways implicated in cancer are likely connected to NF- κ B activation. A critical mediator of NF- κ B activity is TGF β -activated kinase 1 (TAK1). Here, we identify TAK1 as a novel interacting protein and direct target of fibroblast growth factor receptor 3 (FGFR3) tyrosine kinase activity. We further demonstrate that activating mutations in FGFR3 associated with both multiple myeloma and bladder cancer can modulate expression of genes which regulate NF- κ B signaling, and promote both NF- κ B transcriptional activity and cell adhesion in a manner dependent on TAK1 expression in both cancer cell types. Our findings suggest TAK1 as a potential therapeutic target for FGFR3-associated cancers, and other malignancies in which TAK1 contributes to constitutive NF- κ B activation.

Overall design: A total of 12 samples of MGHU3 (Y275C) mutant FGFR3 bladder cancer cells (a kind gift from Dr. Margaret Knowles (University of Leeds, Leeds, UK)) were used for array-based gene expression analysis. 3 replicates of each condition: Control siRNA, Control siRNA + PD173074, TAK1 siRNA, and TAK1 siRNA + PD173074.

Contributor(s): Thompson LM

Citation(s): Salazar L, Kashiwada T, Krejci P, Meyer AN et al. Fibroblast growth factor receptor 3 interacts with and activates TGF β -activated kinase 1 tyrosine phosphorylation and NF- κ B signaling in multiple myeloma and bladder cancer. *PLoS One* 2014;9(1):e86470. PMID: 24466111

Submission date: Nov 18, 2013
Last update date: Apr 26, 2017
Contact name: Leslie Thompson
E-mail: lmthompson@uci.edu
Organization name: University of California, Irvine
Street address: Biological Sciences III
City: Irvine
State/province: CA
ZIP/Postal code: 92697
Country: USA

Platforms (1): GPL6244 [HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]

Samples (12): GSM1267150 MGHU3 Cells_Control siRNA_Biological Replicate 1
GSM1267151 MGHU3 Cells_Control siRNA_Biological Replicate 2
GSM1267152 MGHU3 Cells_Control siRNA_Biological Replicate 3

Relations: BioProject: PRJNA228972

Analyze with GEO2R

Download family: Format: SOFT, MINML, TXT

Supplementary file	Size	Download	File type/resource
GSE52452_RAW.tar	51.1 Mb	(http/custom)	TAR (of CEL)

Raw data provided as supplementary file
Processed data included within Sample table

Custom GSE52452_RAW.tar archive:

Supplementary file	File size
<input checked="" type="checkbox"/> GSM1267150_0611F-01_Control_1_HuGene-1_0-st-v1_CEL.gz	4.3 Mb
<input checked="" type="checkbox"/> GSM1267151_0611F-01_Control_2_HuGene-1_0-st-v1_CEL.gz	4.2 Mb
<input checked="" type="checkbox"/> GSM1267152_0611F-01_Control_3_HuGene-1_0-st-v1_CEL.gz	4.2 Mb
<input type="checkbox"/> GSM1267153_0611F-01_Control_PD_1_HuGene-1_0-st-v1_CEL.gz	4.2 Mb
<input type="checkbox"/> GSM1267154_0611F-01_Control_PD_2_HuGene-1_0-st-v1_CEL.gz	4.2 Mb
<input type="checkbox"/> GSM1267155_0611F-01_Control_PD_3_HuGene-1_0-st-v1_CEL.gz	4.3 Mb
<input type="checkbox"/> GSM1267156_0611F-01_TAK_PD_1_HuGene-1_0-st-v1_CEL.gz	4.2 Mb
<input type="checkbox"/> GSM1267157_0611F-01_TAK_PD_2_HuGene-1_0-st-v1_CEL.gz	4.3 Mb
<input type="checkbox"/> GSM1267158_0611F-01_TAK_PD_3_HuGene-1_0-st-v1_CEL.gz	4.2 Mb
<input checked="" type="checkbox"/> GSM1267159_0611F-01_TAK_1_HuGene-1_0-st-v1_CEL.gz	4.3 Mb
<input checked="" type="checkbox"/> GSM1267160_0611F-01_TAK_2_HuGene-1_0-st-v1_CEL.gz	4.3 Mb
<input checked="" type="checkbox"/> GSM1267161_0611F-01_TAK_3_HuGene-1_0-st-v1_CEL.gz	4.3 Mb

Select All 6 file(s), 25.6 Mb

GSM1267150_0611F-01_Control_1_HuGene-1_0-st-v1_CEL.gz
GSM1267151_0611F-01_Control_2_HuGene-1_0-st-v1_CEL.gz
GSM1267152_0611F-01_Control_3_HuGene-1_0-st-v1_CEL.gz
GSM1267159_0611F-01_TAK_1_HuGene-1_0-st-v1_CEL.gz
GSM1267160_0611F-01_TAK_2_HuGene-1_0-st-v1_CEL.gz
GSM1267161_0611F-01_TAK_3_HuGene-1_0-st-v1_CEL.gz
をチェックしてダウンロードをクリックする。

ダウンロードされたGSE52452_RAW.tarをダブルクリックしてデスクトップのGSE52452_RAWフォルダに解凍する。

統計解析ソフトR

- オープンソースの統計解析ソフト

<http://cran.r-project.org>

で配布

- Windows Mac Linuxで使用可能
- 様々な研究分野で広く使われている
- 参考

<http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html>

http://cran.r-project.org

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2014-10-31, Pumpkin Helmet) [R-3.1.2.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

What are R and CRAN?

横4コマ.pptx | 写真データ.zip Cancelled | ni.2987.pdf | nihms374495.pdf | 12月マネ会議日程調...xlsx | Show all downloads...

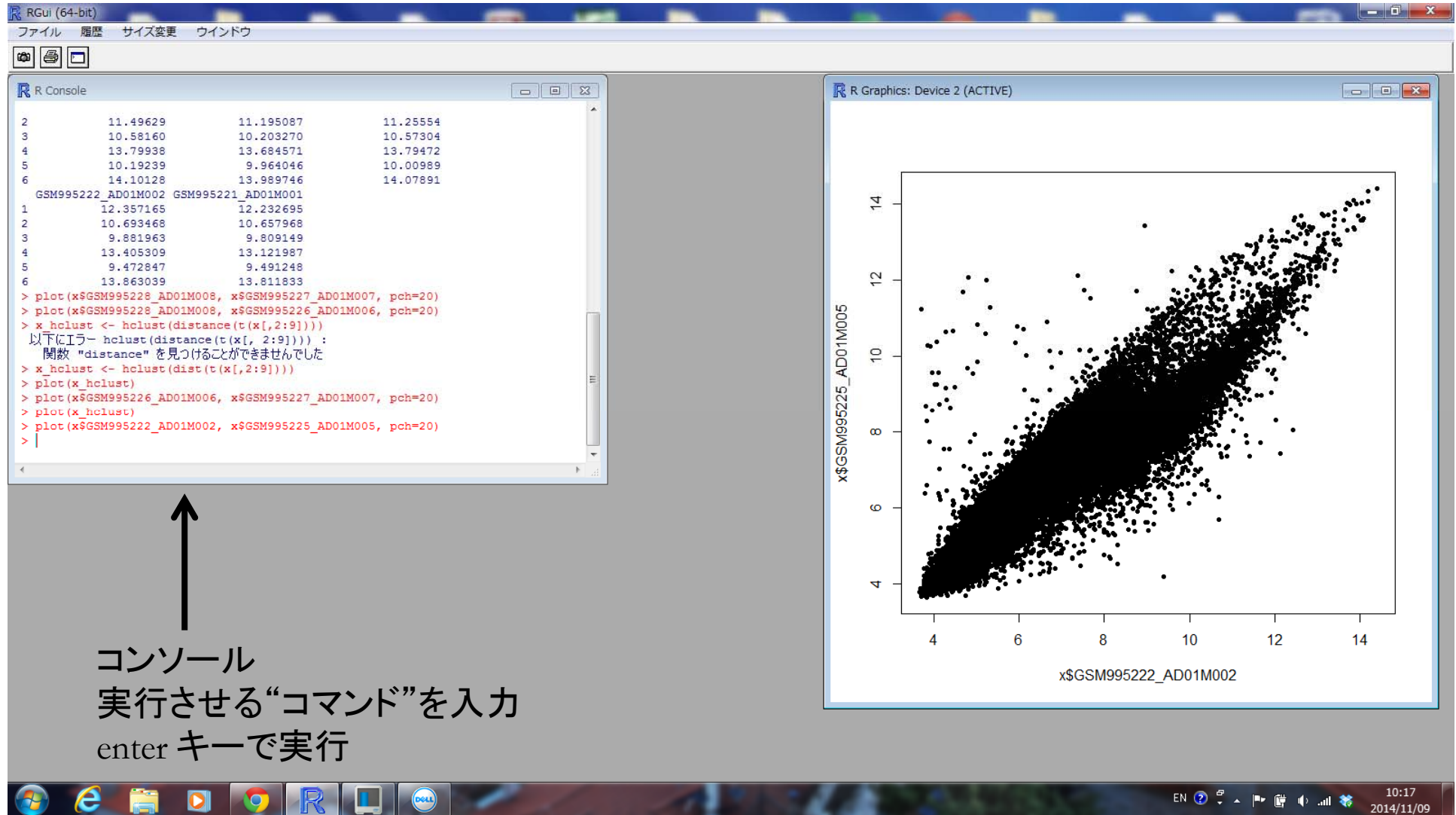
Rの起動

The screenshot shows a web browser window with the URL `cran.md.tsukuba.ac.jp/bin/windows/base/`. The page title is "R-3.1.2 for Windows (32/64 bit)". Below the title, there are links for "Download R 3.1.2 for Windows (54 megabytes, 32/64 bit)", "Installation and other instructions", and "New features in this version". A paragraph of text explains how to verify the downloaded package using `md5sum` and provides links for "graphical" and "command line versions". Below this, there is a section for "Frequently asked questions" with several links, and a section for "Other builds" with links to "snapshot build" and "r-devel snapshot build".

Overlaid on the bottom left of the browser window is a Windows Start menu. The "R x64 3.1.0" icon is highlighted with a red box. The Start menu also shows other applications like "Windows Live フォトギャラリー", "Microsoft Office PowerPoint 2007", "Microsoft Office Excel 2007", "Mozilla Thunderbird", "Cygwin64 Terminal", "ワードパッド", "Microsoft Office Word 2007", "Java Treeview", "Lhaplus", and "3D Vision を有効にする". The taskbar at the bottom shows icons for Internet Explorer, File Explorer, Chrome, Dell, and R. The system tray on the right shows the time as 18:17 and the date as 2014/11/01.

スタートメニューからRを選択して起動

Rのコンソール



The screenshot displays the R GUI interface. The R Console window on the left shows the execution of R code. The R Graphics window on the right displays a scatter plot of two variables, x\$GSM995222_AD01M002 and x\$GSM995225_AD01M005, showing a strong positive correlation.

```
R Console  
2      11.49629      11.195087      11.25554  
3      10.58160      10.203270      10.57304  
4      13.79938      13.684571      13.79472  
5      10.19239      9.964046      10.00989  
6      14.10128      13.989746      14.07891  
GSM995222_AD01M002 GSM995221_AD01M001  
1      12.357165      12.232695  
2      10.693468      10.657968  
3      9.881963      9.809149  
4      13.405309      13.121987  
5      9.472847      9.491248  
6      13.863039      13.811833  
> plot(x$GSM995228_AD01M008, x$GSM995227_AD01M007, pch=20)  
> plot(x$GSM995228_AD01M008, x$GSM995226_AD01M006, pch=20)  
> x_hclust <- hclust(distance(t(x[,2:9])))  
> x_hclust <- hclust(dist(t(x[,2:9])))  
以下にエラー hclust(distance(t(x[,2:9]))) :  
関数 "distance" を見つけることができませんでした  
> x_hclust <- hclust(dist(t(x[,2:9])))  
> plot(x_hclust)  
> plot(x$GSM995226_AD01M006, x$GSM995227_AD01M007, pch=20)  
> plot(x_hclust)  
> plot(x$GSM995222_AD01M002, x$GSM995225_AD01M005, pch=20)  
> |
```

R Graphics: Device 2 (ACTIVE)
Scatter plot showing the relationship between x\$GSM995222_AD01M002 (X-axis) and x\$GSM995225_AD01M005 (Y-axis). Both axes range from 4 to 14. The plot shows a dense cluster of points with a strong positive linear correlation.

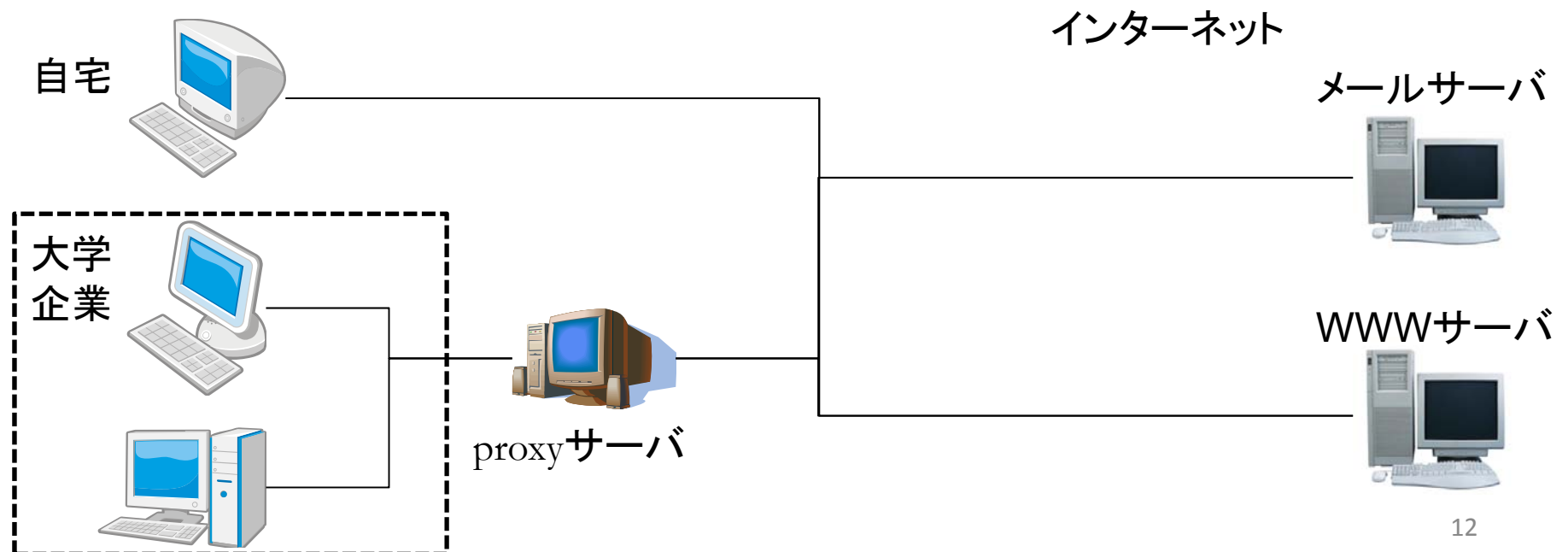


コンソール
実行させる“コマンド”を入力
enter キーで実行

proxyの設定(横浜市大の場合)

```
R console [-] [ ] [X]  
>Sys.setenv(http_proxy="http://proxy.med.yokohama-cu.ac.jp:8080")  
>Sys.getenv("http_proxy")
```

R起動直後に実行しないと設定されないことがあります。



Packageのインストール

Package

複数の関数をまとめたものがパッケージとして提供されている。


Bioconductor.org

- バイオインフォマティクス関連のパッケージを配布しているサイト

<http://bioconductor.org>

The screenshot shows the Bioconductor.org homepage. At the top left is the Bioconductor logo with the tagline "OPEN SOURCE SOFTWARE FOR BIOINFORMATICS". To the right of the logo is a search bar. Below the logo is a navigation menu with links for "Home", "Install", "Help", "Developers", and "About". The main content area is divided into several sections: "About Bioconductor" with a paragraph describing the project, "News" with a list of recent updates, "Install" with a list of links for getting started, "Learn" with a list of resources for mastering tools, "Use" with a list of links for creating solutions, and "Develop" with a list of links for contributing to the project.

org

 **Bioconductor**
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

[Home](#) [Install](#) [Help](#) [Developers](#) [About](#)

About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, [934 software packages](#), and an active user community. Bioconductor is also available as an [Amazon Machine Image \(AMI\)](#).

News

- Recent literature citations are now collated on the updated [publications](#) page.
- [Bioconductor 3.0 is released!](#)
- Use the [support site](#) to get help installing, learning and using Bioconductor.
- Learning R / Bioconductor for Sequence Analysis [course material](#) and [videos](#) now

Install »

Get started with *Bioconductor*

- [Install Bioconductor](#)
- [Explore packages](#)
- [Support](#)
- [Latest newsletter](#)
- [Follow us on Twitter](#)
- [Using R](#)

Learn »

Master *Bioconductor* tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

Use »

Create bioinformatic solutions with *Bioconductor*

- [Software](#), [Annotation](#), and [Experiment](#) packages
- [Amazon Machine Image](#)
- [Latest release announcement](#)

Develop »

Contribute to *Bioconductor*

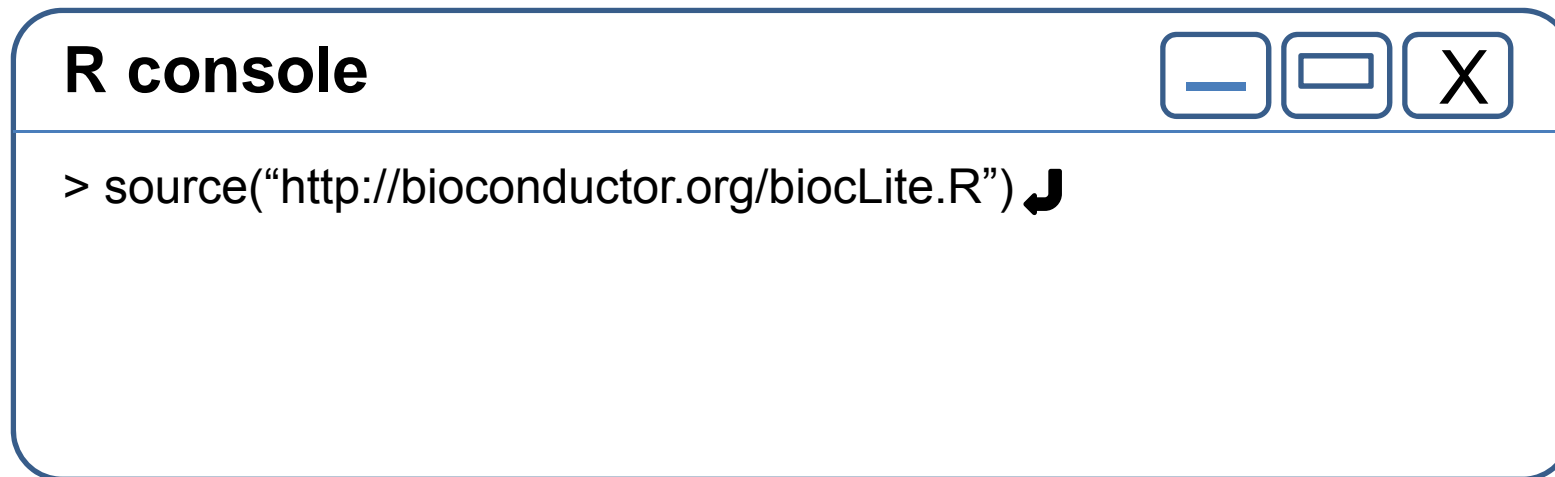
- [Use Bioc 'devel'](#)
- ['Devel' Software](#), [Annotation](#) and [Experiment](#) packages
- [Package guidelines](#)
- [New package submission](#)

Packageのインストール

今回使用するpackage

- “affy”
Affymetrixデータ処理
- “genefilter”
t-test実行
- “mygene”
ID変換

Bioconductor, biocLiteの設定



R console

```
> source("http://bioconductor.org/biocLite.R") ↵
```

Bioconductor

バイオインフォマティクス関連のパッケージを配布しているサイト

biocLite.R

バイオインフォマティクス関連のパッケージをインストールするインストーラ
パッケージ間の依存関係やバージョンの整合性を調整してくれる。

Package のインストール

R console



```
> biocLite("affy") ↵
```

```
> library(affy) ↵
```

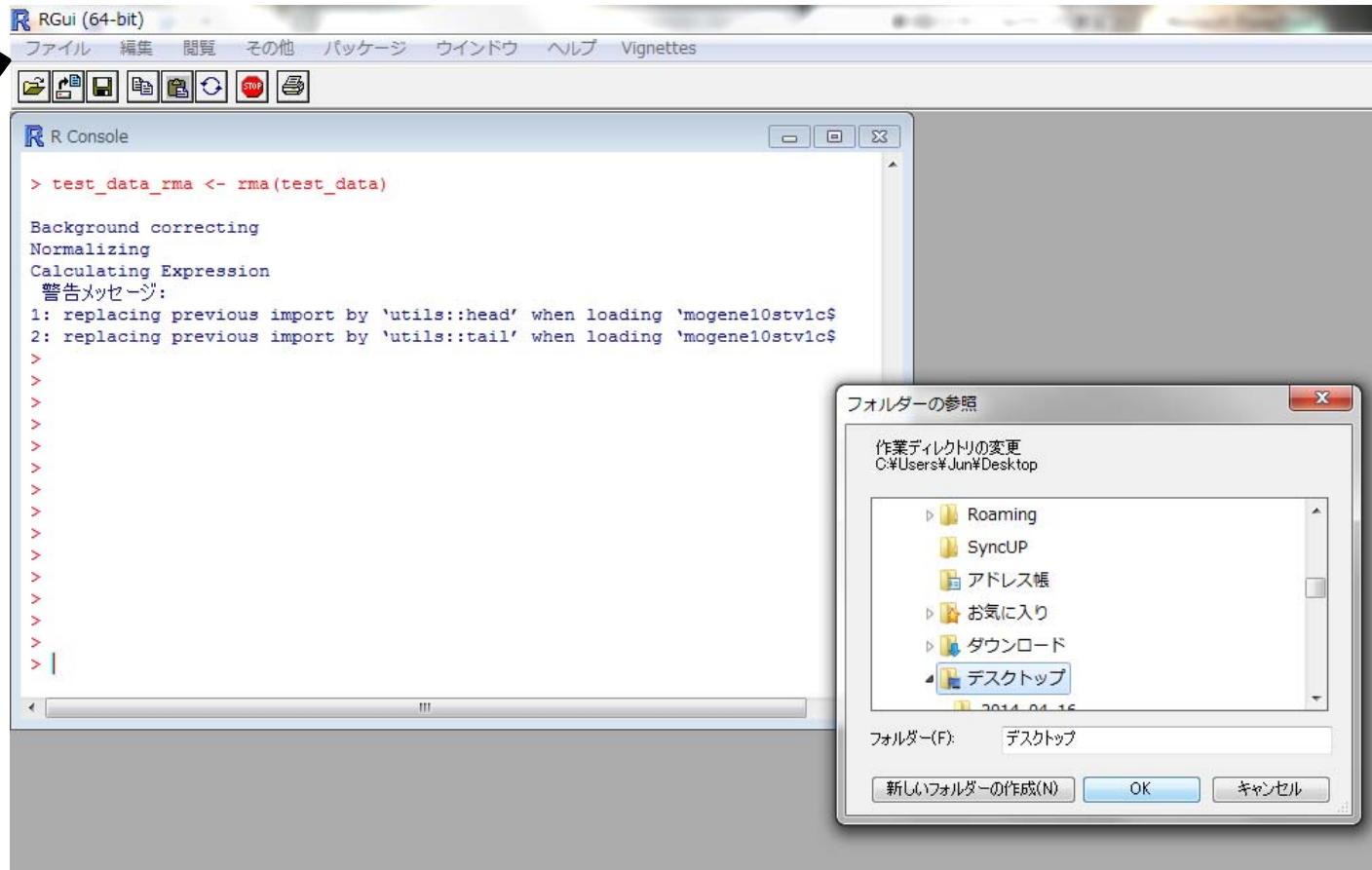
```
> biocLite("genefilter") ↵
```

```
> library(genefilter) ↵
```

```
> biocLite("mygene") ↵
```

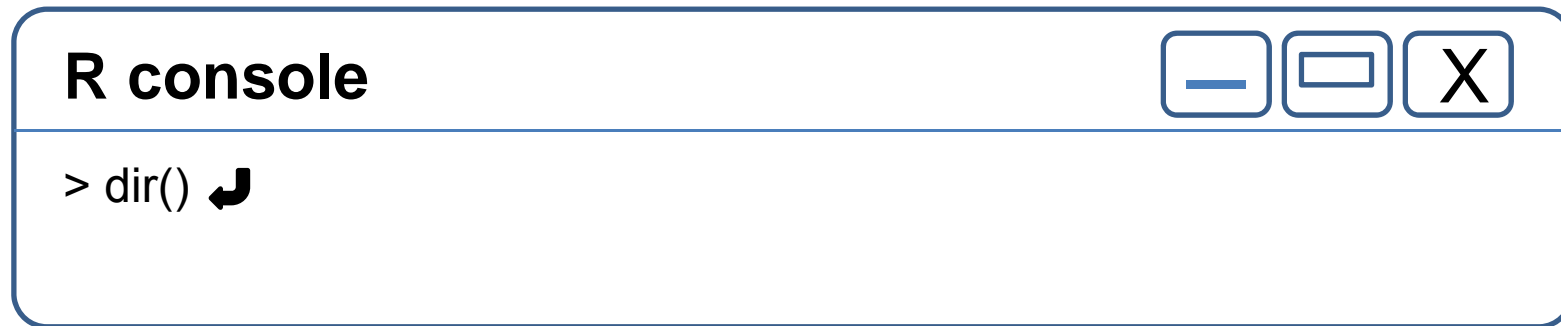
```
> library(mygene) ↵
```

作業ディレクトリに移動



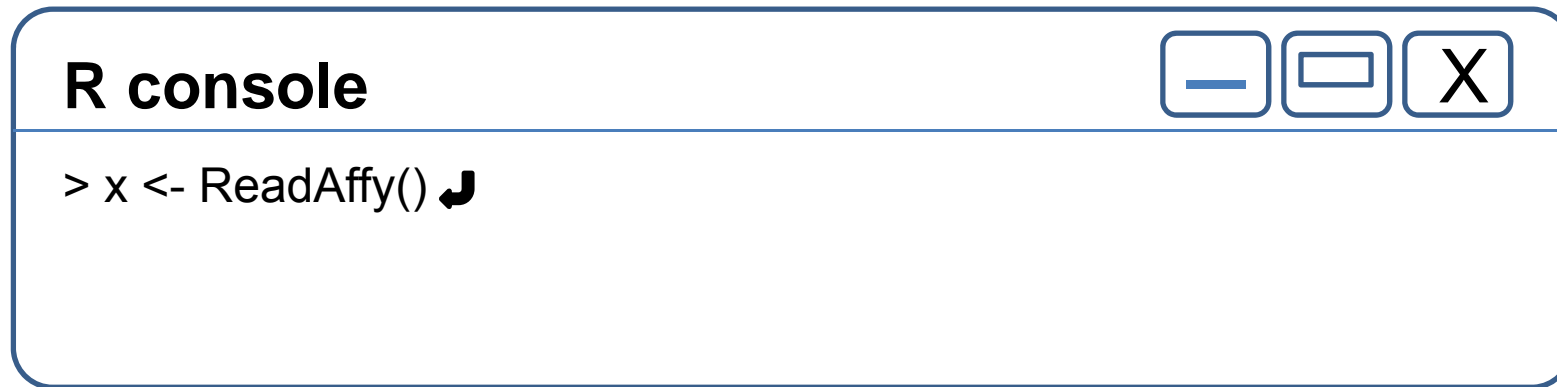
ファイルメニューから“ディレクトリの変更”を選択
作業ディレクトリを選択（読み書き可能な各自のアカウントフォルダを選択）
次回ログイン時に、今回書き込んだデータが保持されます。

確認



コンソールにCELファイル名が表示されたら、データの取得とディレクトリの変更が完了しています。

データの読み込み

A screenshot of an R console window. The window has a title bar with the text "R console" and three standard window control buttons (minimize, maximize, close) on the right. The main area of the window contains the R command "> x <- ReadAffy()" followed by a carriage return symbol (↵).

```
R console
> x <- ReadAffy() ↵
```

ReadAffy() : 作業フォルダ内のCELファイルの内容を読み込む
CELファイルはgz圧縮した状態でも読み込み可能。

rma法で正規化

R console

```
> y <- rma(x) ↵  
> z <- exprs(z)  
> write.exprs(y, "GSE52452_Exp.txt")
```

rma() : rma法で正規化

exprs() : 発現量の \log_2 値を抽出

write.exprs() : 発現量の \log_2 値をファイルに出力

RMA (Robust Multi-Array Average) 法

Exploration, normalization, and summaries of high density oligonucleotide array probe level data.

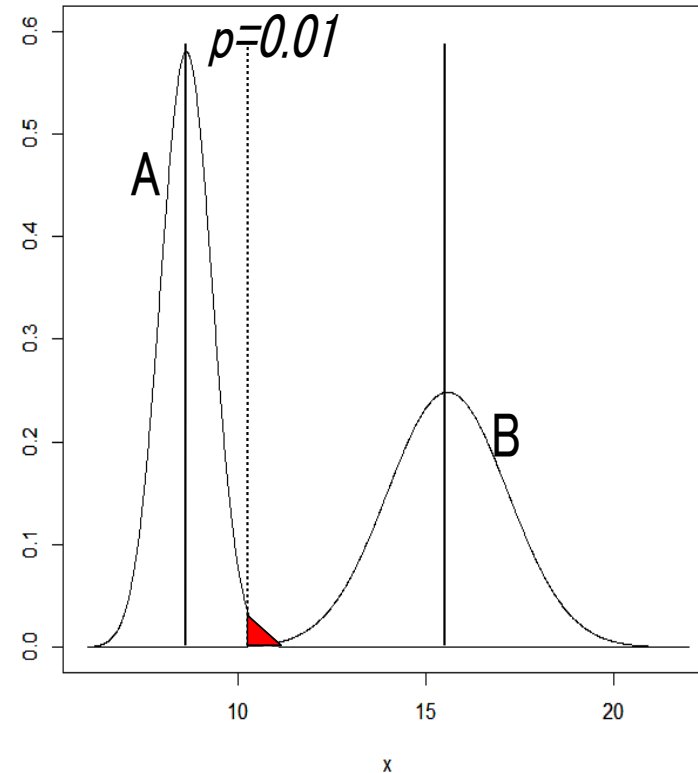
Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP

Biostatistics 2003 4(2):249-64

二群間の比較

A群	B群
7.89	16.28
9.60	16.75
9.07	13.21
8.31	17.01
8.30	14.69

帰無仮説
A群、B群は平均の等しい
母集団から得られた



B群のデータが得られる確率は1%以下



帰無仮説を棄却し、二群間には差があると判定する

検定による過誤と多重検定問題

実際/判定	陰性	陽性
陰性	True Negative (TN)	False Positive (FP)
陽性	False Negative (FN)	True Positive (TP)

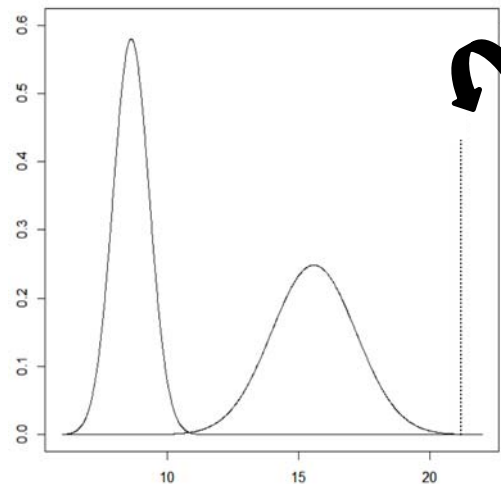
有意水準 $P < 0.01$: 実測データが得られる確率は1%未満
遺伝子発現データ 20000超 → 200個以上のFalse Positive

Bonferroni補正
有意水準を検定回数で割る

$$\hat{p} = p / N$$

$$\hat{p} = 0.01 / 20000$$

$$= 5 \times 10^{-7}$$



非現実的な有意
水準が要求され
る

False Discovery Rate (FDR)

BH法：1995年にBenjaminiとHochbergによって提唱された。
 $FDR=FP/(FP+TP)$ を指標にする手法。ある程度のFPを許容する考え方。

H : 帰無仮説
 N : 遺伝子の総数
 p : *p-value*

H_1	H_2	H_3	...	H_N
p_1	p_2	p_3	...	p_N

一定の割合でFPを含むと仮定する

$$q_{(i)} = \frac{p_{(i)} \times N}{i} \quad (i=1,2,\dots,N)$$

$p_{(i)} \times N$: 有意水準 $p_{(i)}$ で全ての帰無仮説を棄却したときのFPの推定数

実際の計算方法

p 値を低いものから並べる

q 値を計算

$i, j=1, 2, \dots, N$ かつ $i < j$ について、 $q_i > q_j$ なら $q_i = q_j$ とする

$q_i > q^*$ となる帰無仮説を全て棄却する

統計解析ソフトR genefilterパッケージ

R console



```
> w <- rowttests(as.matrix(z), factor(c("C", "C", "C", "S", "S", "S")))↵  
> head(w) ↵  
> q <- p.adjust(w$p.value, method = "BH") ↵  
> w <- cbind(z, w, q) ↵  
> head(w) ↵
```

rowttests : 列ごとにt-testを実行

p.adjust : BH法でFDRを計算

cbind : 列を結合

統計解析ソフトR mygeneパッケージ

R console



```
> result1 <- queryMany(rownames(subset(w, w$q.value < 0.05 & w$dm > 1)),  
+ scopes = "reporter", species = "human", fields = "symbol")  
> result2 <- queryMany(rownames(subset(w, w$q.value < 0.05 & w$dm < -1)),  
+ scopes = "reporter", species = "human", fields = "symbol"))  
> result1$symbol  
> result2$symbol  
> write.table(result1, "overExpressed.txt", quote=F, row.names=F, sep="¥t")  
> write.table(result2, "underExpressed.txt", quote=F, row.names=F, sep="¥t")
```

queryMany() : IDを変換

変換前のIDをscopesに記述する。Affymetrix probe ID は reporter

変換後のIDをfieldsに記述する。gene symbolはsymbol

生物種をspeciesに記述する。ヒトはhuman、マウスはmouse

まとめ

- CEL ファイルフォーマットの遺伝子発現プロファイルを GEO データベースから取得。
- 統計解析ソフト R の affy パッケージを使ってデータの読み込み、rma 補正を行う。
- genefilter パッケージを使って二群間の比較を行い、発現変動遺伝子を抽出。
- mygene パッケージを使って遺伝子の Affymetrix probe ID を遺伝子名に変換。