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

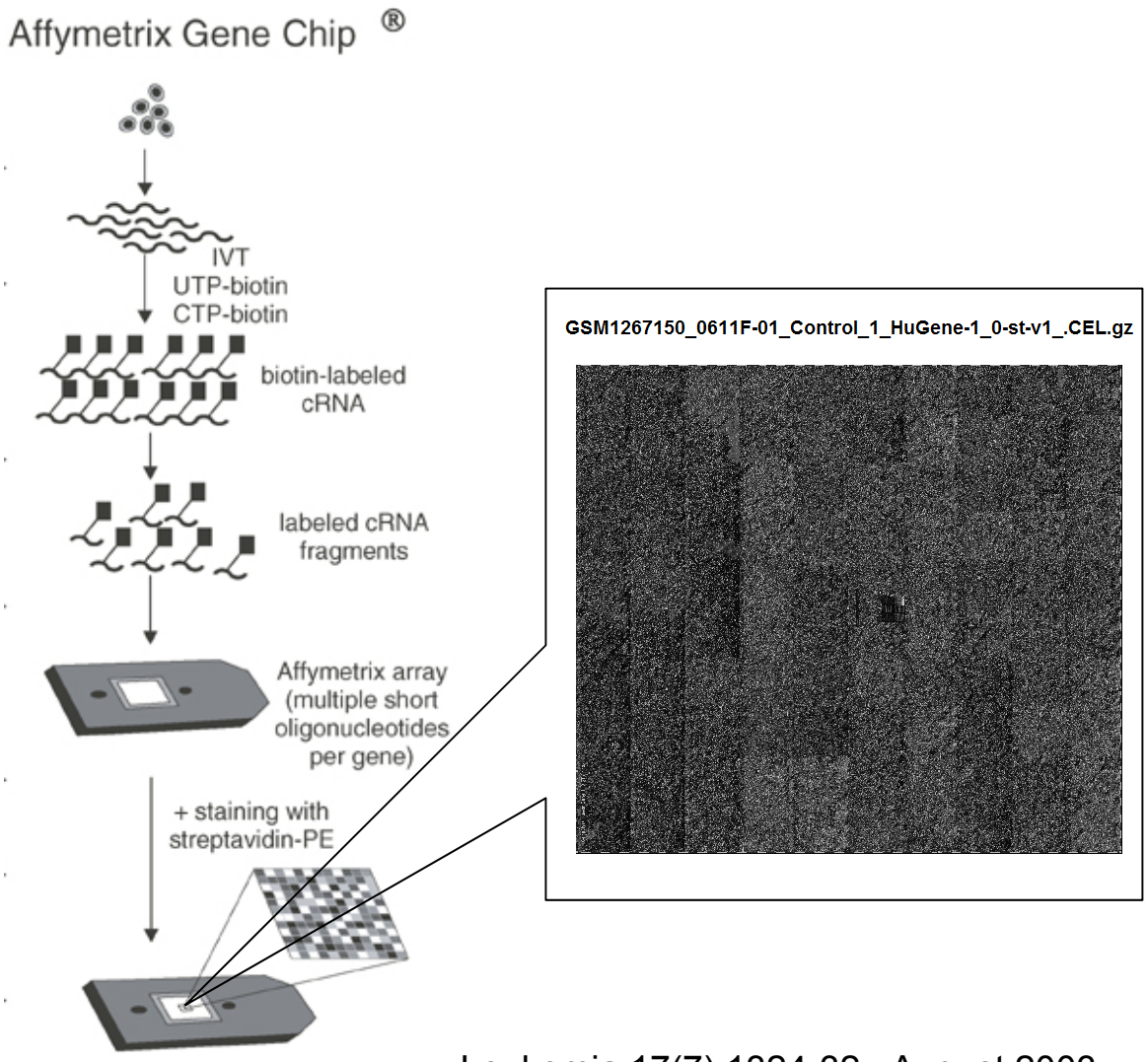
mRNA抽出

逆転写

ビオチンラベルcDNA

ハイブリダイズ

蛍光強度を測定

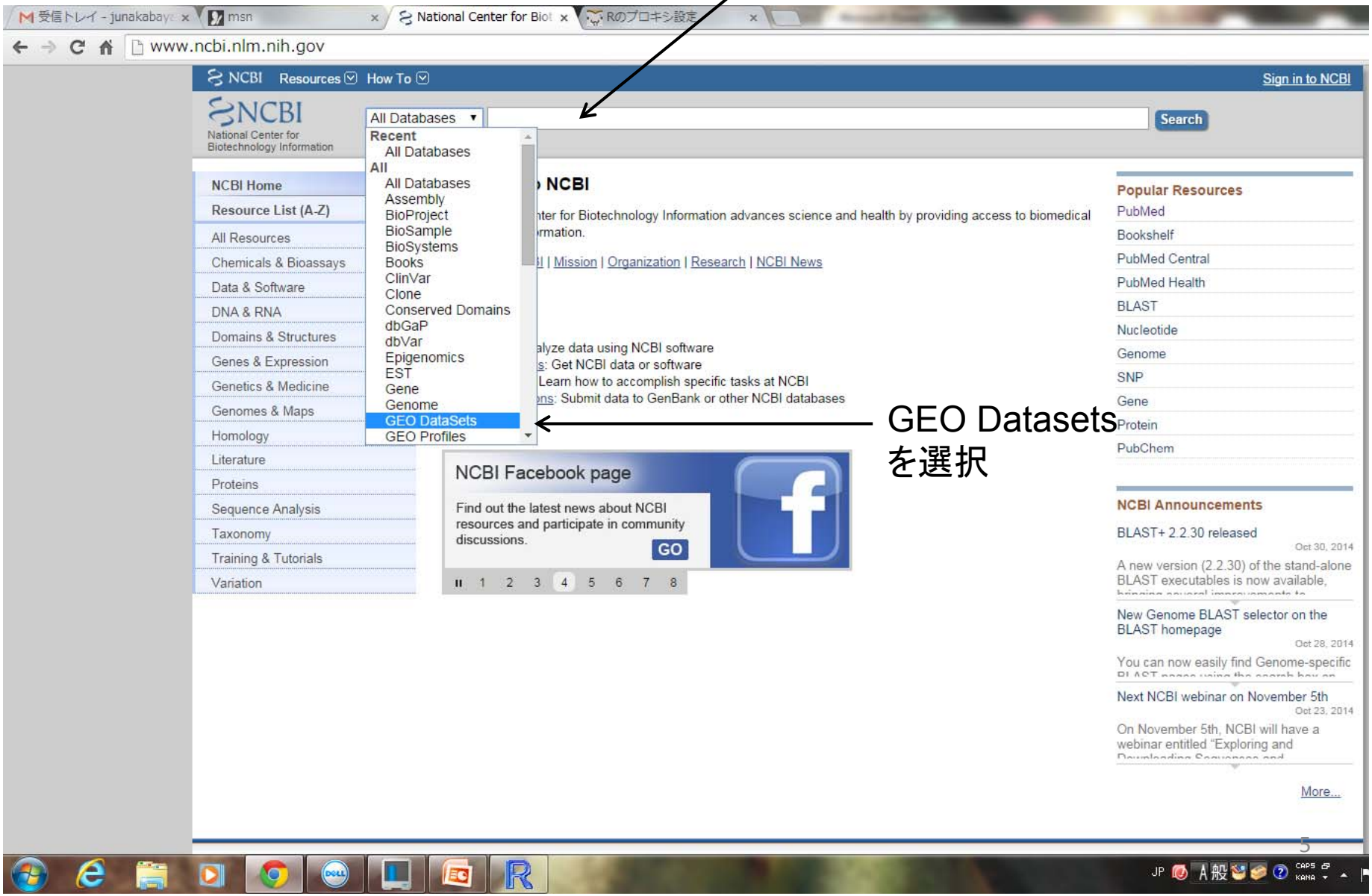


Leukemia 17(7):1324-32 · August 2003

GEOデータベース検索

http://ncbi.nlm.nih.gov

GSE52452を入力して検索



GSE52452

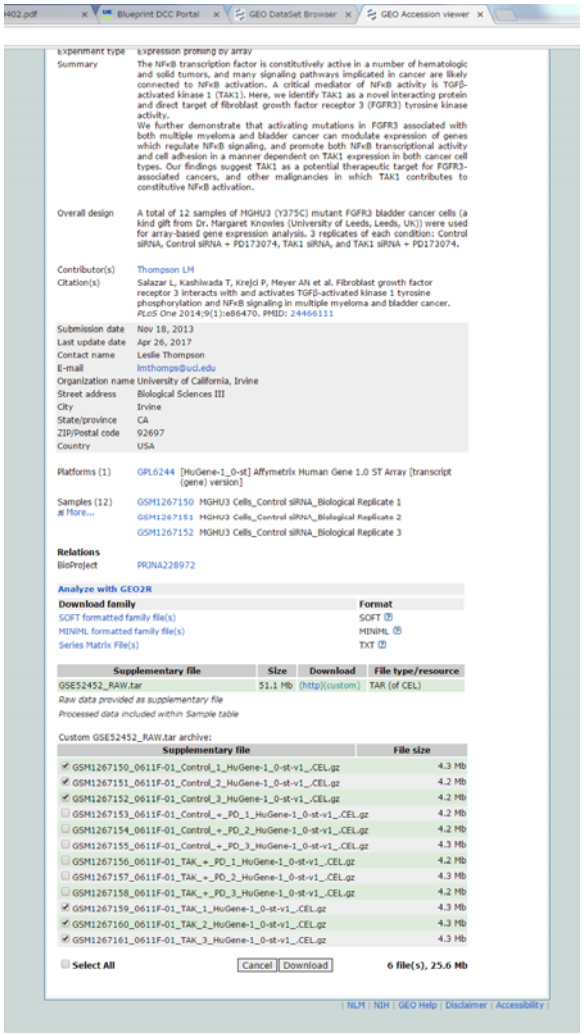
- Fibroblast growth factor receptor 3 interacts with and activates TGFβ-activated kinase 1 tyrosine phosphorylation and NFκB signaling in multiple myeloma and bladder cancer.

  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



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

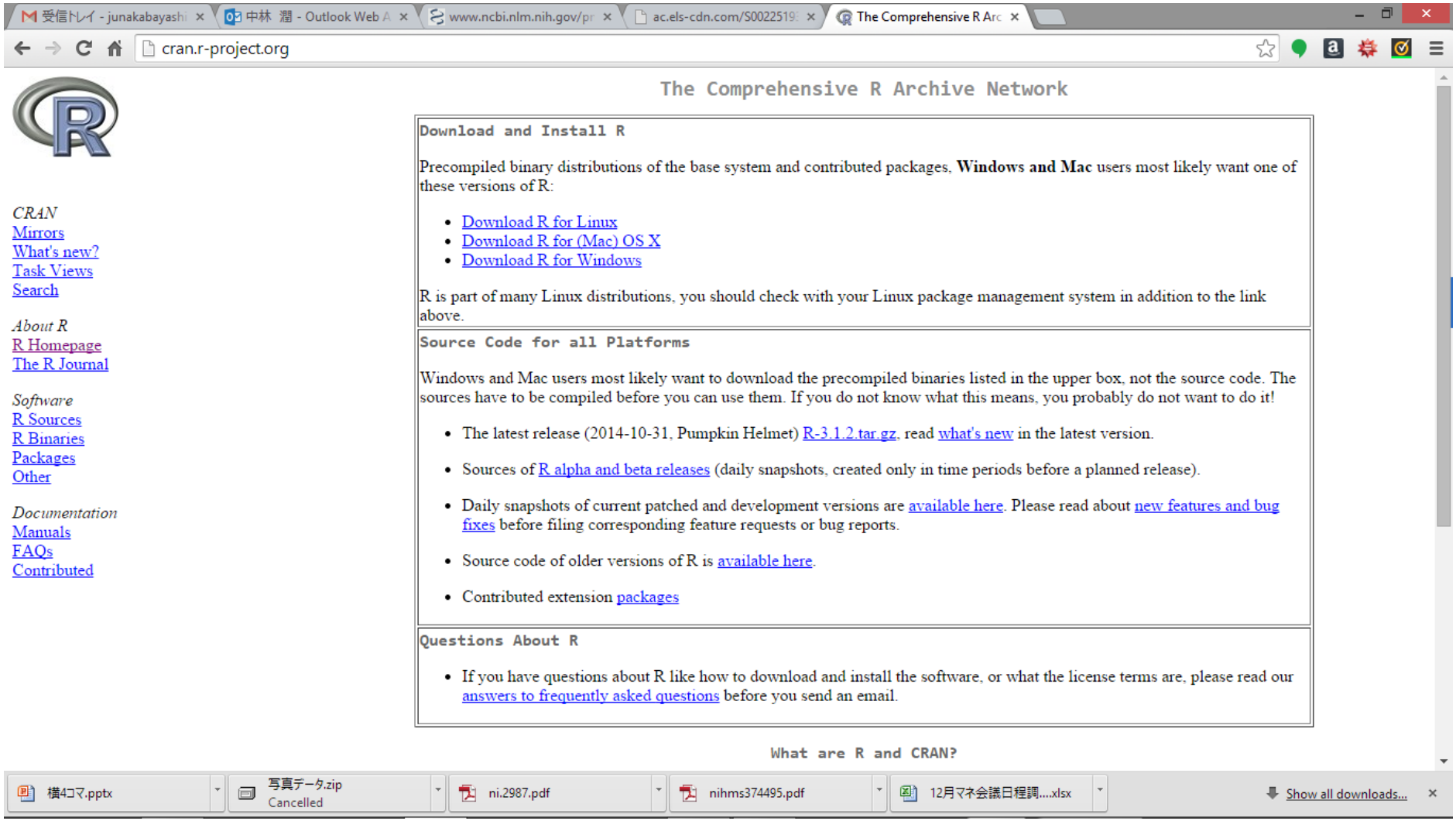
- Windows Mac Linuxで使用可能

- 様々な研究分野で広く使われている

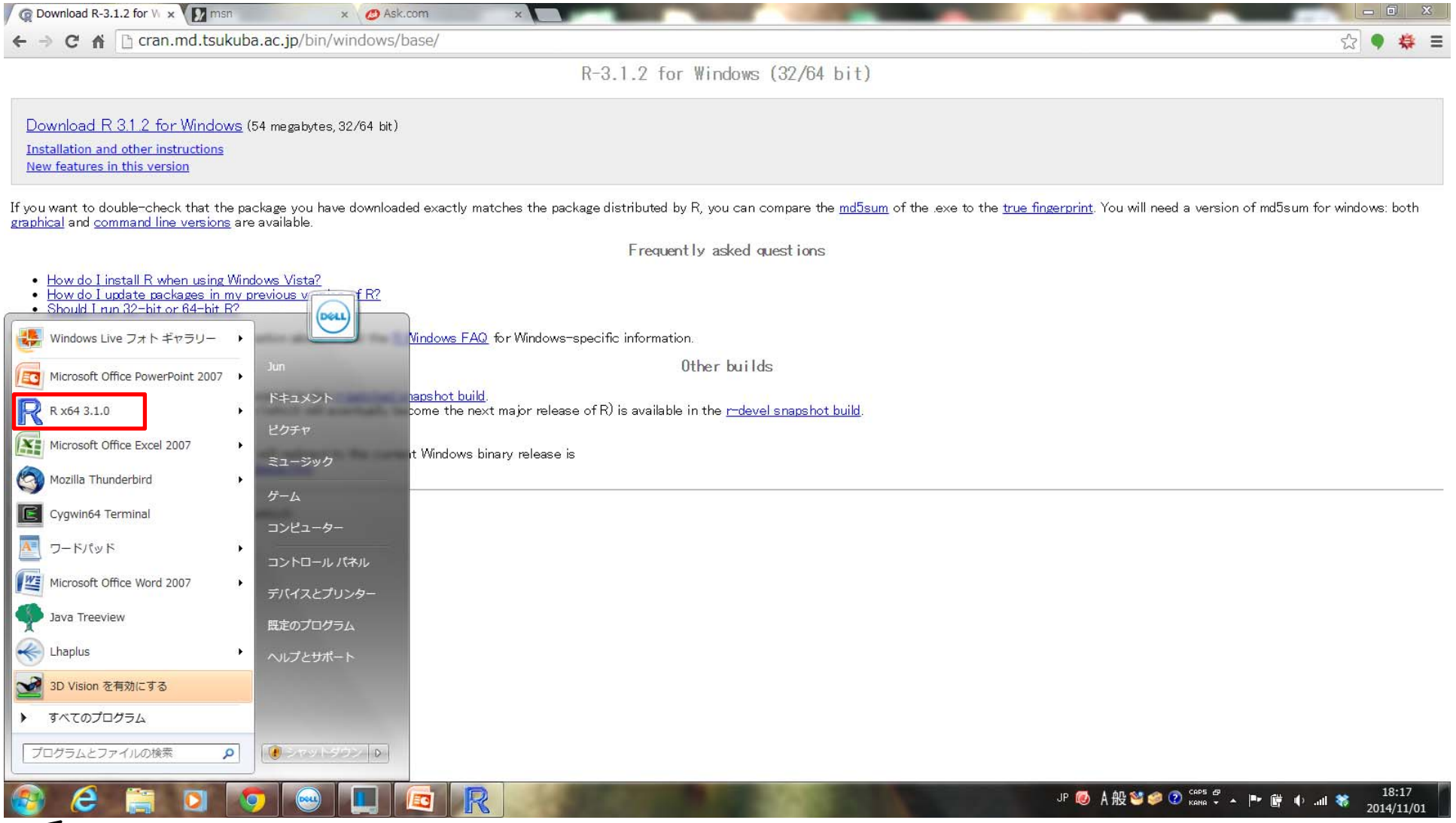
- 参考

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

http://cran.r-project.org

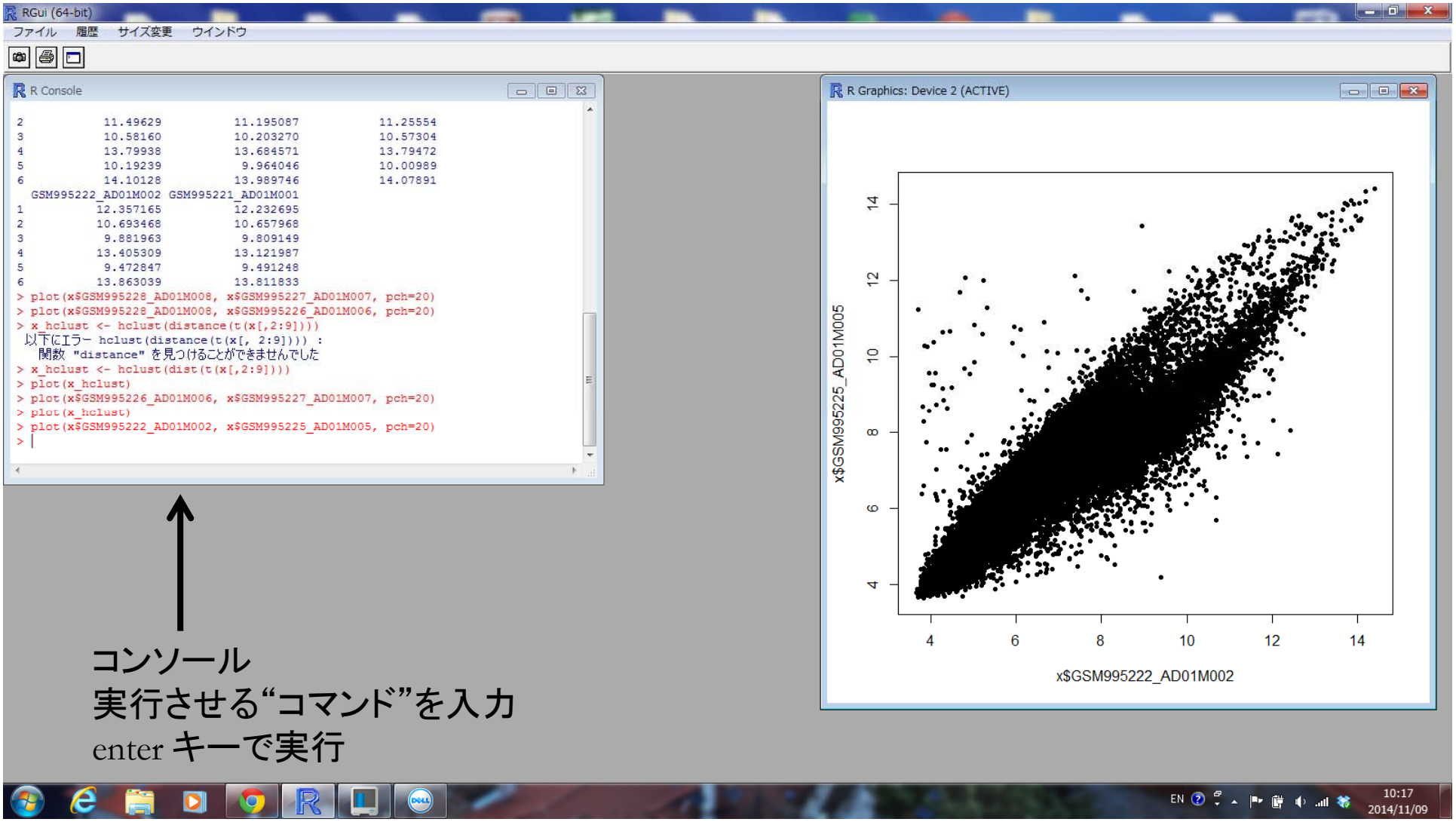


Rの起動



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

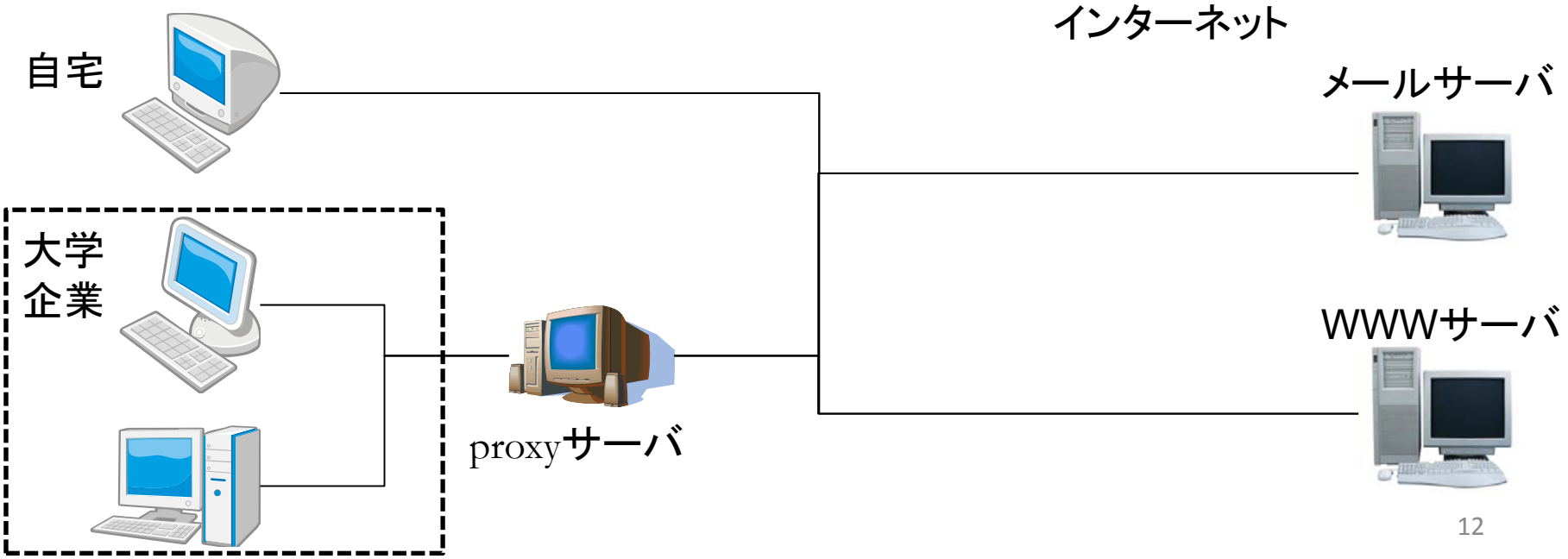

RGui (64-bit)
ファイル 履歴 サイズ変更 ウインドウ
R Console
2 11.49629 11.195087 11.25554
3 10.58160 10.203270 10.57304
4 13.79938 13.684571 13.79472
5 10.19239 9.964046 10.00989
6 14.10128 13.989746 14.07891
GSM995222_AD01M002 GSM995221_AD01M001
1 12.357165 12.232695
2 10.693468 10.657968
3 9.881963 9.809149
4 13.405309 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])))
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)
x$GSM995225_AD01M005
x$GSM995222_AD01M002
4
6
8
10
12
14
コンソール
実行させる“コマンド”を入力
enter キーで実行
EN
10:17
2014/11/09

proxyの設定（横浜市大の場合）

R console

```
>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

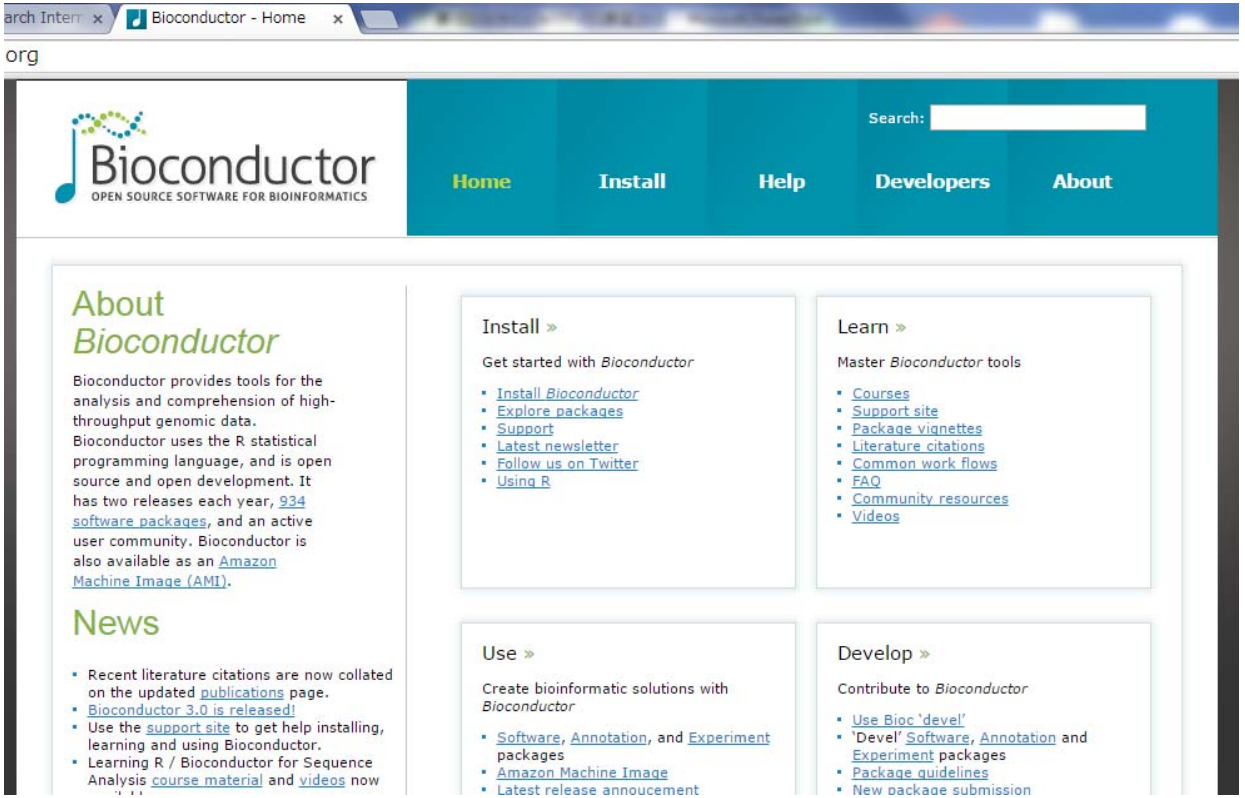
Package

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



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

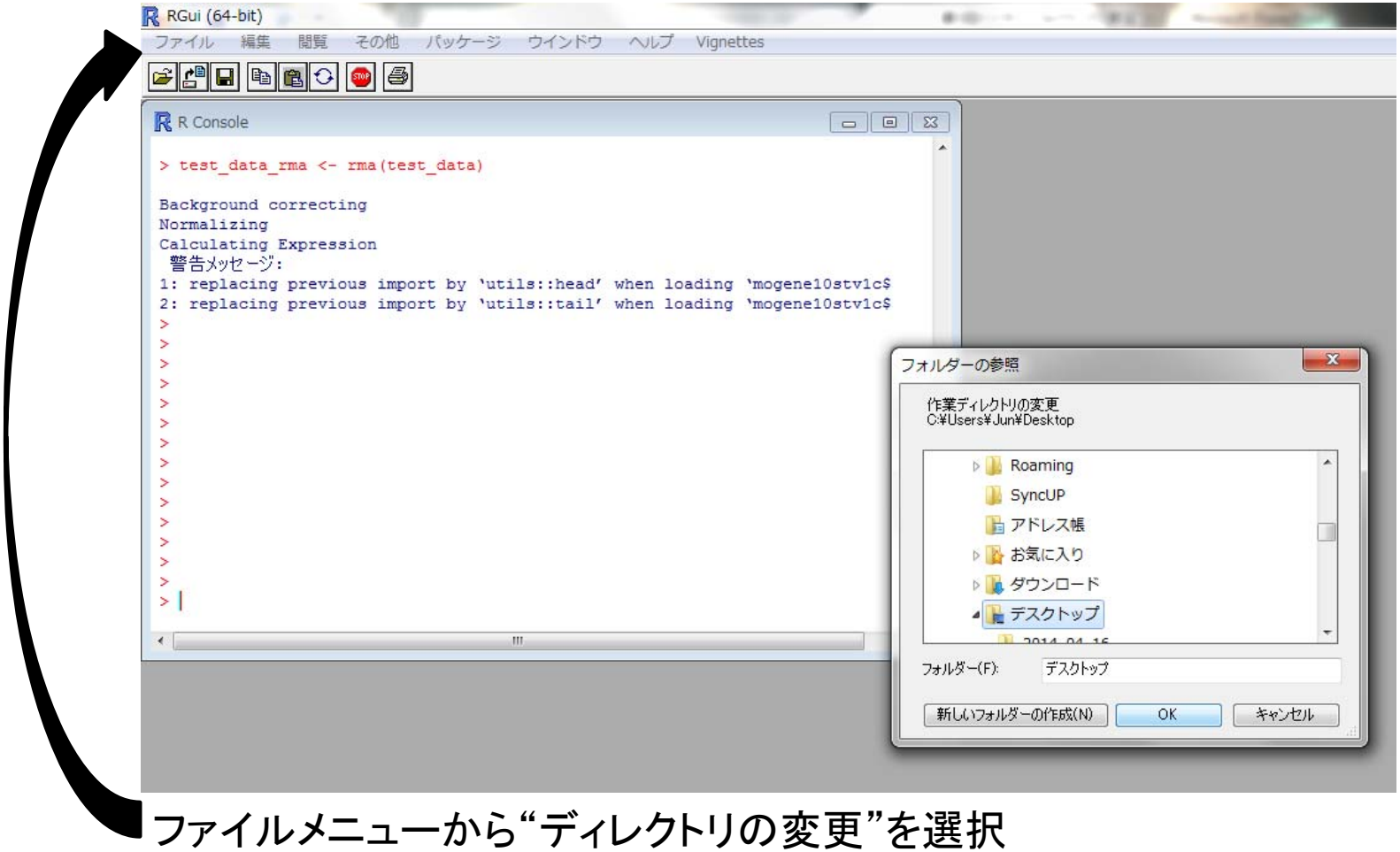
biocLite.R
バイオインフォマティクス関連のパッケージをインストールするインストーラ
パッケージ間の依存関係やバージョンの整合性を調整してくれる。

Package のインストール

R console

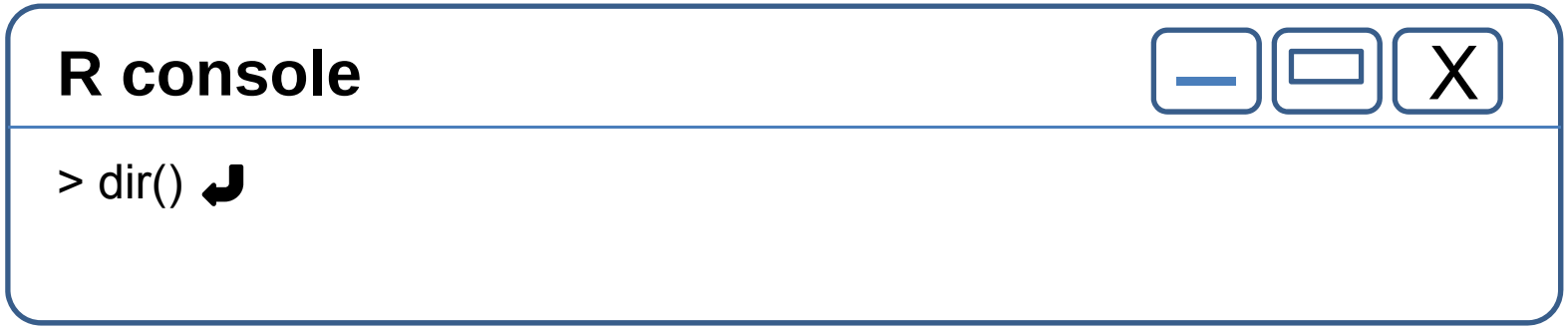
```
> biocLite("affy")
> library(affy)
> biocLite("genefilter")
> library(genefilter)
> biocLite("mygene")
> library(mygene)
```

作業ディレクトリに移動



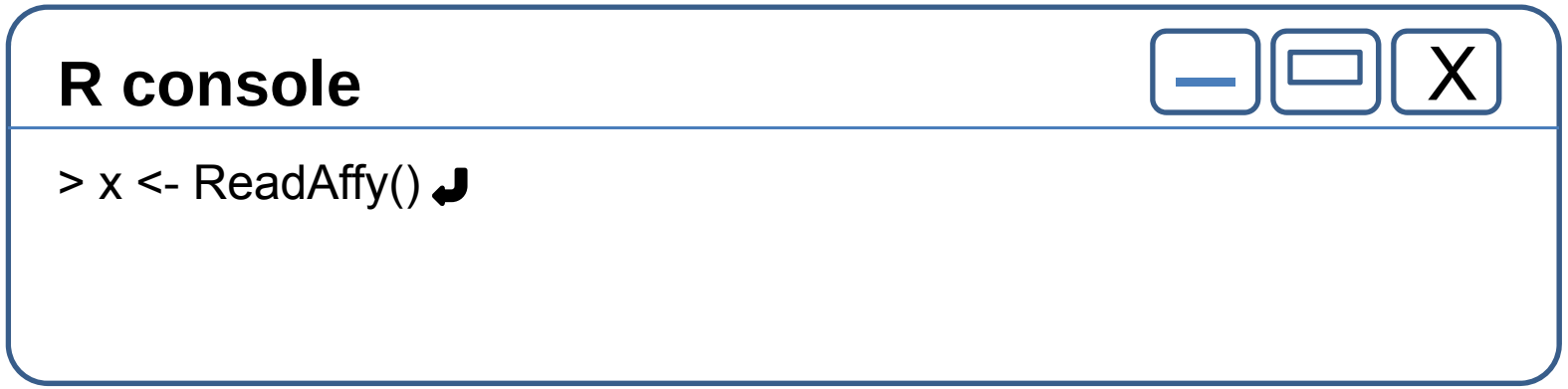
ファイルメニューから“ディレクトリの変更”を選択
作業ディレクトリを選択（読み書き可能な各自のアカウントフォルダを選択）
次回ログイン時に、今回書き込んだデータが保持されます。

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



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”)
```

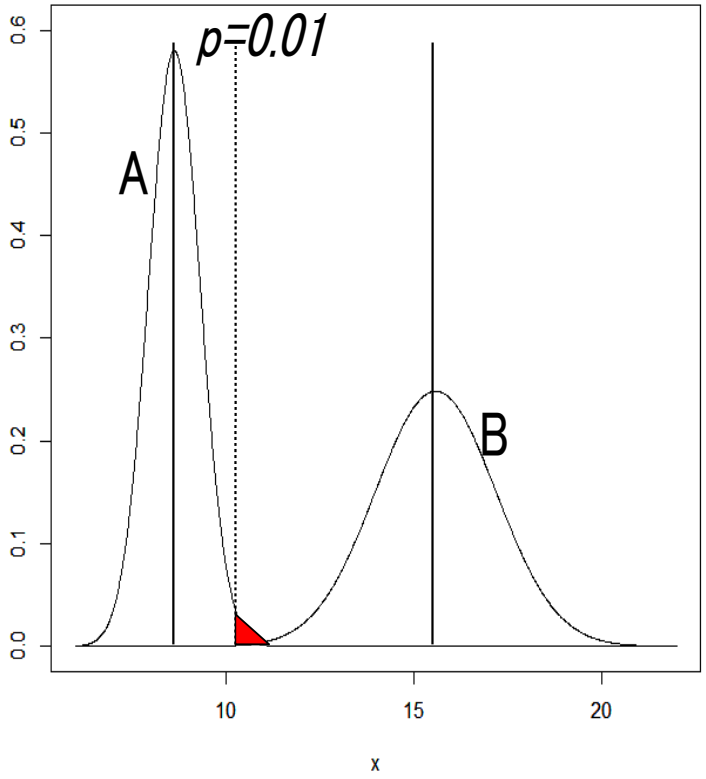
rma() : rma法で正規化
exprs() : 発現量の$\log_2$値を抽出
write.exprs() : 発現量の$\log_2$値をファイルに出力

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

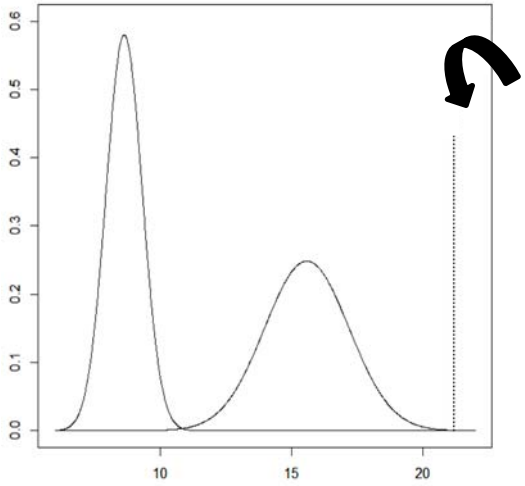
検定による過誤と多重検定問題

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

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

$p_{(i)}$ x N: 有意水準$p_{(i)}$で全ての帰無仮を棄却したときの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,…,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。