



第1回バイオインフォマティクス実習コース  
横浜市大 先端医科学研究センター  
バイオインフォマティクス研究室

室長 田村智彦

准教授 中林潤

免疫学 小泉真一

- データベースからデータの取得
- 正規化
- ファイルへ出力

# M402LL教室のPC環境

- YCUアカウントでログイン  
読み書き可能フォルダ  
Z:/ユーザ名  
読み込み可能フォルダ  
課題配布/BioInfoJishu
- インターネット接続可  
proxyサーバ経由
- R ver 3.0.2がインストール済

# 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 Canceled | ni.2987.pdf | nihms374495.pdf | 12月マネ会議日程調...xlsx | Show all downloads...

# Rの起動

Download R-3.1.2 for Windows (32/64 bit)

[Download R 3.1.2 for Windows](#) (54 megabytes, 32/64 bit)  
[Installation and other instructions](#)  
[New features in this version](#)

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the [md5sum](#) of the .exe to the [true fingerprint](#). You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- [How do I install R when using Windows Vista?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

[Windows FAQ](#) for Windows-specific information.

Other builds

- [snapshot build](#).
- ... (the next major release of R) is available in the [r-devel snapshot build](#).
- ... (the next Windows binary release is

Windows Live フォト ギャラリー  
Microsoft Office PowerPoint 2007  
**R x64 3.1.0**  
Microsoft Office Excel 2007  
Mozilla Thunderbird  
Cygwin64 Terminal  
ワードパッド  
Microsoft Office Word 2007  
Java Treeview  
Lhaplus  
3D Vision を有効にする  
すべてのプログラム

プログラムとファイルの検索

シャットダウン

18:17  
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. On the left, the 'R Console' window shows a table of data and several R commands. The data table has 6 rows and 4 columns of numerical values. Below the table, the console shows the execution of several R commands, including plotting and hierarchical clustering. A red error message is visible: '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'. The 'R Graphics: Device 2 (ACTIVE)' window on the right shows a scatter plot of two variables, with both axes ranging from 4 to 14. The plot shows a strong positive correlation between the two variables.

```
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)
>
```

↑  
コンソール  
実行させる“コマンド”を入力  
enter キーで実行

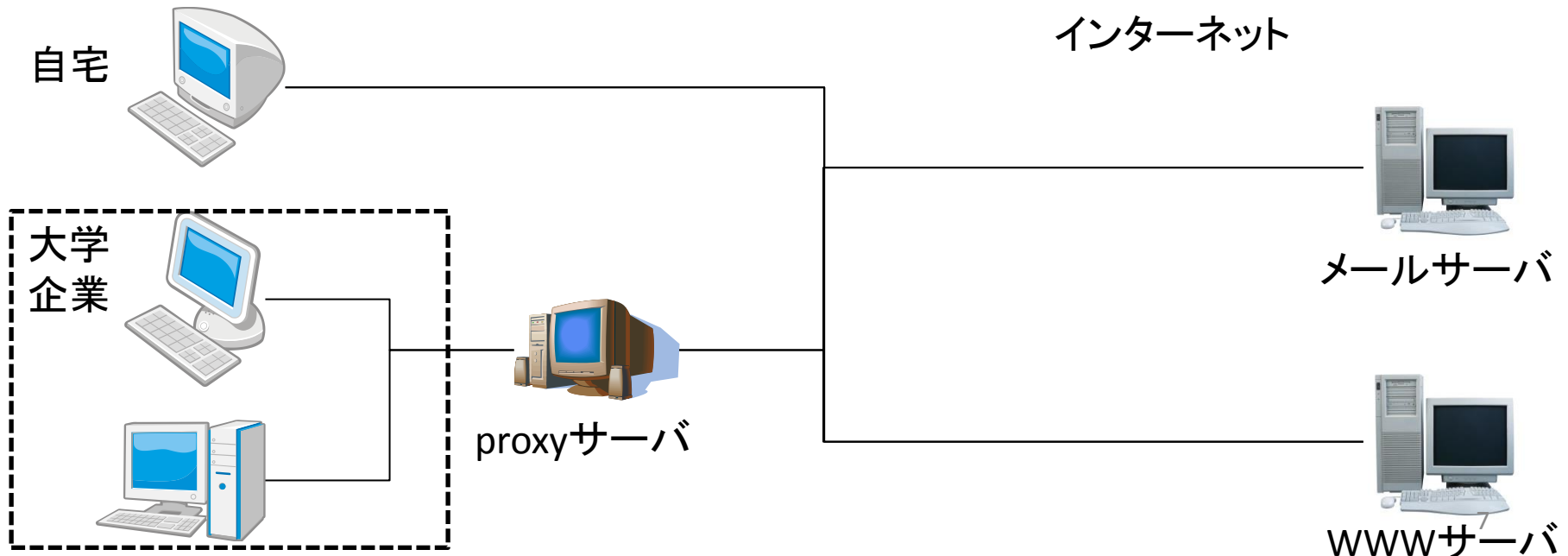
# proxyの設定(横浜市大の場合)

R console



```
>Sys.setenv(http_proxy="http://proxy.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。



# Rの基本操作

R console



```
> 34 + 58 ↵  
> 92 ↵  
> 105 / 33 ↵  
> 3.181818 ↵  
> pi ↵  
> 3.141593 ↵  
> sqrt(2) ↵  
> 1.414214 ↵  
> x <- 10 ↵  
> y <- 15 ↵  
> z <- x + y ↵  
> x <- seq(-10, 10, by=0.1) ↵  
> plot(sin(x), type="l") ↵
```





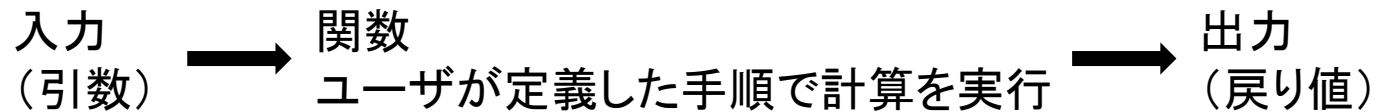
# ファイルの読み込み

R console



```
> p <- read.table("Kokonoe.txt", header=T, sep="¥t")  
> p[1,1]  
> p[1,3]
```

# ユーザ定義関数



R console



```
> bmi <- function(q){ ↵  
+ r <- p[q,4] / (p[q,3] / 100)^2 ↵  
+ return(r)} ↵  
> bmi(3) ↵  
> p <- cbind(p, p[,4] / (p[,3] / 100)^2) ↵  
> write.table(p, "Kokonoe_rev.txt", quote=F, sep="¥t") ↵
```

# edit関数を使った入力



```
R console [- □ X  
> bmi <- edit(bmi) ␣
```

別ウインドウにテキストエディタが開くので、そこで入力の訂正を行う。

# 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 website. The browser tab is labeled "Bioconductor - Home". The page features a teal header with the Bioconductor logo (a stylized 'B' with a DNA helix) and the text "Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS". To the right of the logo is a search bar and a navigation menu with links for "Home", "Install", "Help", "Developers", and "About".

The main content area is divided into several sections:

- About Bioconductor:** A paragraph describing the software's purpose for analyzing high-throughput genomic data, its use of the R programming language, and its open-source nature. It mentions two releases per year and an Amazon Machine Image (AMI).
- Install >:** A section titled "Get started with Bioconductor" with links for "Install Bioconductor", "Explore packages", "Support", "Latest newsletter", "Follow us on Twitter", and "Using R".
- Learn >:** A section titled "Master Bioconductor tools" with links for "Courses", "Support site", "Package vignettes", "Literature citations", "Common work flows", "FAQ", and "Community resources".
- Use >:** A section titled "Create bioinformatic solutions with Bioconductor" with links for "Software, Annotation, and Experiment packages", "Amazon Machine Image", and "Latest release announcement".
- Develop >:** A section titled "Contribute to Bioconductor" with links for "Use Bioc 'devel'", "'Devel' Software, Annotation and Experiment packages", "Package guidelines", and "New package submission".
- News:** A section with bullet points about updated literature citations, the release of Bioconductor 3.0, and learning resources for R and Bioconductor.

# Packageのインストール

今回使用するpackage

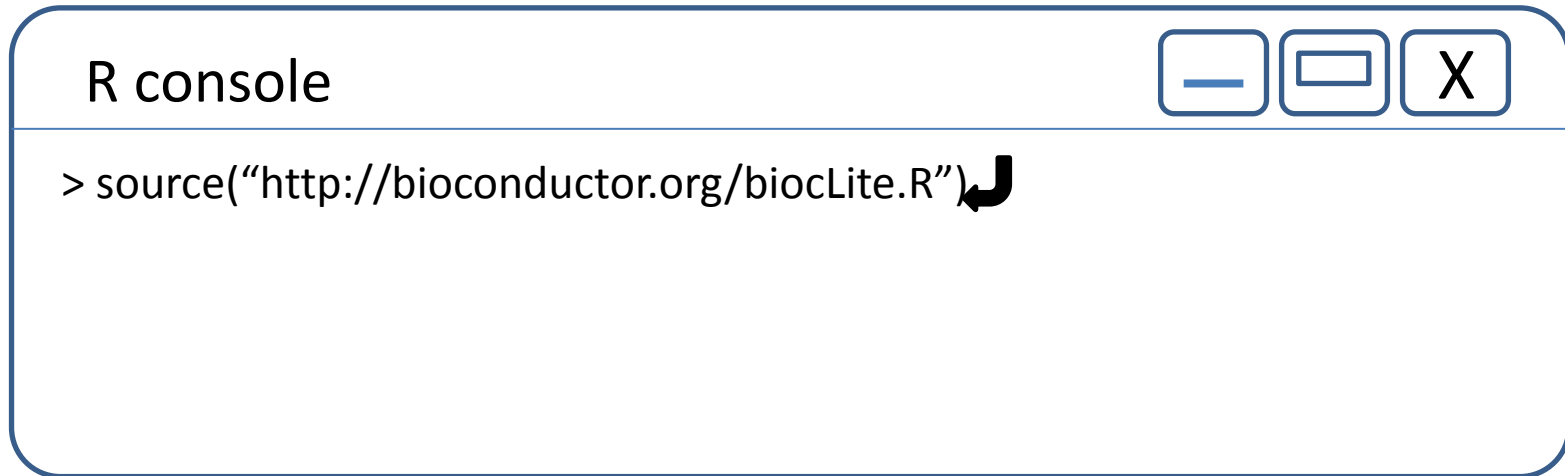
- “affy”

Affymetrixデータ処理用パッケージ

- “AnnotationDbi”

アノテーション用ゲノムインデックス

# Bioconductor, biocLiteの設定



R console

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

The image shows a simulated R console window with a title bar containing the text 'R console' and three window control buttons (minimize, maximize, close). The main area of the window contains the R command `> source("http://bioconductor.org/biocLite.R")` followed by a carriage return symbol.

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 “affy”

## Package “AnnotationDbi”

### Package “mogene10stv1cdf”

R console



```
> biocLite("affy") ↵
```

```
> library(affy) ↵
```

```
> biocLite("AnnotationDbi") ↵
```

```
> library(AnnotationDbi) ↵
```

```
> biocLite("mogene10stv1cdf") ↵
```

```
> library(mogene10stv1cdf) ↵
```

# GEOデータベース検索

http://ncbi.nlm.nih.gov

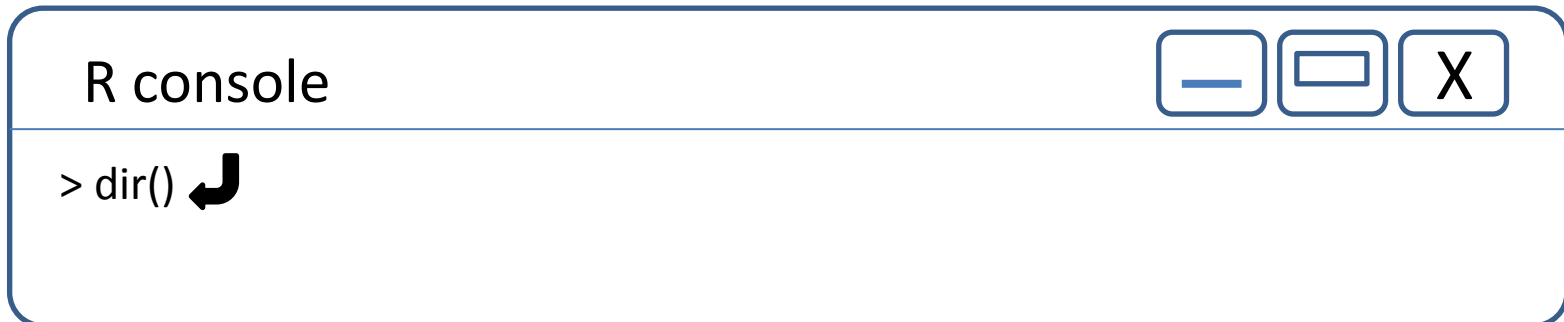
キーワードを入力

The screenshot shows the NCBI website interface. At the top, there is a search bar with the text "キーワードを入力" (Enter keyword) and a "Search" button. Below the search bar, a dropdown menu is open, showing "All Databases" and "Recent" sections. The "All Databases" section lists various databases, with "GEO DataSets" highlighted in blue. An arrow points from the text "キーワードを入力" to the search bar, and another arrow points from the text "GEO Datasets を選択" (Select GEO Datasets) to the "GEO DataSets" option in the dropdown menu. The website header includes "NCBI Resources" and "How To" menus. The main content area features a "Popular Resources" section with links to PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Below this is an "NCBI Facebook page" widget and an "NCBI Announcements" section with news about BLAST+ 2.2.30 and a webinar.

GEO Datasets  
を選択

# データの取得

- 課題配布→BioInfoJishuフォルダから  
GSE40493フォルダをZ:/デスクトップに移動
- Rの作業フォルダをZ:/デスクトップ/GSE40493  
に変更



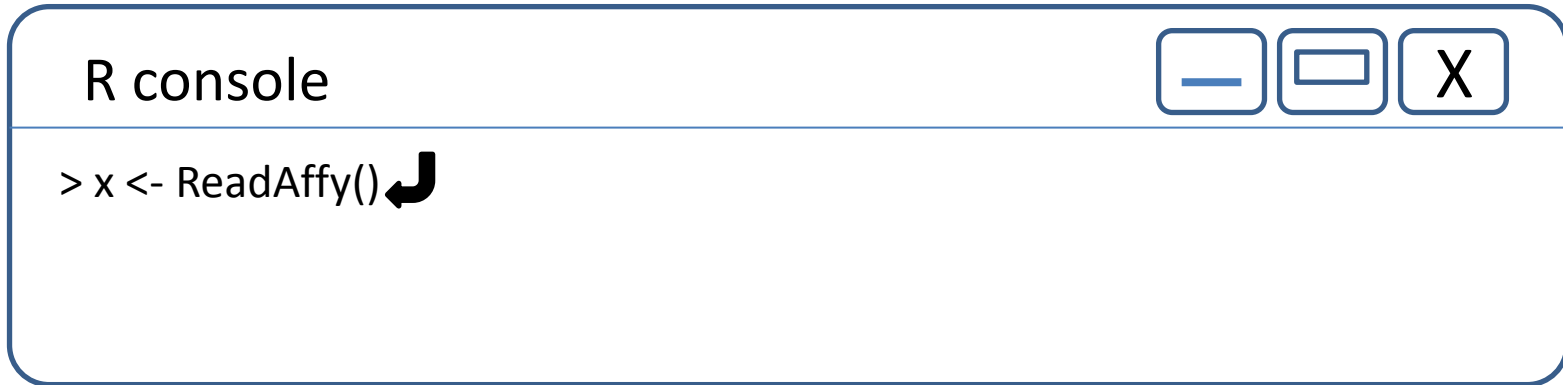
R console

```
> dir() ↵
```

The image shows a screenshot of an R console window. The title bar reads "R console" and has standard window control buttons (minimize, maximize, close). The main area of the console shows the command prompt "> dir()" followed by a black arrow cursor pointing to the right, indicating that the command has been entered and is ready to be executed.

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。

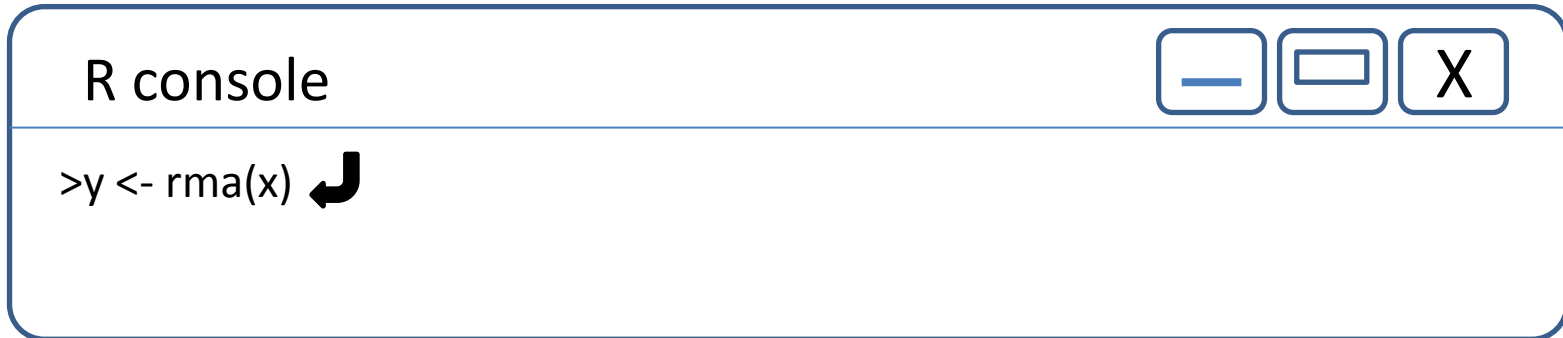
# データの読み込み

A screenshot of an R console window. The title bar reads "R console" and contains standard window control buttons (minimize, maximize, close). The main area shows the command "> x <- ReadAffy()" followed by a cursor and a carriage return symbol.

```
R console  
> x <- ReadAffy()↵
```

作業フォルダ内のCELファイルの内容を変数xに格納する。

# rma法で正規化



```
R console
>y <- rma(x) ↵
```

正規化したデータを変数yに格納する

RMA (Robust Multi-Array Average) 法

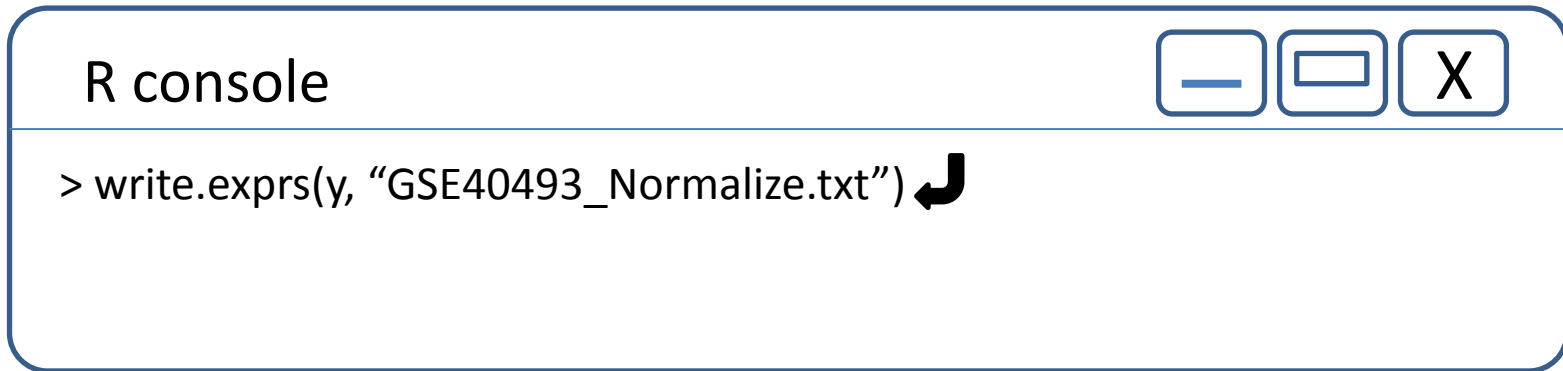
Exploration, normalization, and summaries of high density oligonucleotide array probe level data.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP  
Biostatistics 2003 4(2):249-64

Affymetrixマイクロアレイデータの正規化法として良く用いられている手法の一つ。  
以下の3ステップでデータを正規化する。

- 1) バックグラウンド補正
- 2) quantile normalization
- 3) median polish法でsamalize

# write.exprsでファイルに出力



R console

```
> write.exprs(y, "GSE40493_Normalize.txt") ↵
```

The image shows a simulated R console window with a title bar containing the text 'R console' and three window control buttons (minimize, maximize, close). The main area of the window contains the R command `> write.exprs(y, "GSE40493_Normalize.txt")` followed by a black arrow pointing downwards, representing the execution of the command.

yの内容をタブ区切りテキストファイルとして出力。  
GSE40493フォルダにGSE40493\_Normalizeと言う名前のテキストファイルが  
できていれば作業が完了しています。

# タブ区切りテキストファイルをExcelで開く

The screenshot shows Microsoft Excel with a tab-separated text file opened. The data is organized into a table with 36 rows and 11 columns (A-K). The first column (A) contains IDs, and the following columns (B-K) contain numerical values. The interface includes the ribbon with tabs for Home, Insert, Page Layout, Formulas, Data, Review, and Send. The taskbar at the bottom shows various applications and the system clock.

	A	B	C	D	E	F	G	H	I	J	K
1		GSM995228_AD01 M008	GSM995227_AD01 M007	GSM995226_AD01 M006	GSM995225_AD01 M005	GSM995224_AD01 M004	GSM995223_AD01 M003	GSM995222_AD01 M002	GSM995221_AD01 M001		
2	10338001	12.91442548	12.8671738	12.78268659	13.03053132	12.82905147	12.86226274	12.35716522	12.2326946		
3	10338003	11.28114413	11.30709222	11.10473015	11.49628584	11.19508745	11.25554372	10.69346842	10.65796845		
4	10338004	10.24385947	10.21488992	10.14135173	10.58159823	10.20326957	10.57304408	9.881962815	9.80914948		
5	10338017	13.63967267	13.61198487	13.68939676	13.79937742	13.684571	13.79471508	13.40530929	13.12198731		
6	10338025	9.908501846	9.938011267	9.649013557	10.19238936	9.964045748	10.00988712	9.472846696	9.491247593		
7	10338026	13.93510379	13.96977431	13.99477282	14.10127816	13.98974595	14.07891251	13.86303949	13.81183269		
8	10338029	10.67000607	10.70176467	10.60177371	10.98229195	10.59413637	10.88909846	10.20845142	10.14011654		
9	10338035	9.879418746	9.949708688	9.802723235	10.15610795	9.854045301	9.934284622	9.507837646	9.538623679		
10	10338036	10.43533886	10.4920886	10.41617879	10.98431064	10.49502855	10.79873636	10.12200472	10.01245256		
11	10338037	4.290409006	4.303970035	4.275305621	4.252870391	4.252870391	4.235840949	4.270563079	4.296982084		
12	10338041	12.13131044	12.08674964	12.06145042	12.4410089	12.10317752	12.24704861	11.55415508	11.44279685		
13	10338042	11.12622997	11.24062048	11.08221186	11.63495726	11.20597002	11.31158685	10.62185688	10.59196757		
14	10338044	12.96715264	12.74866645	12.74866645	13.13863544	12.93614621	12.98696313	12.4284398	12.26979821		
15	10338047	6.763653772	6.901370033	6.9107405	6.724367429	6.718937144	6.697530492	6.711506397	6.779435792		
16	10338056	4.159647488	4.173418903	4.131062695	4.13045013	4.165768879	4.104994121	4.154426975	4.164225438		
17	10338059	13.95348944	13.93705037	13.96089114	13.97393993	13.9276401	13.93559217	13.766423	13.67270388		
18	10338060	4.295996842	4.294548503	4.266691578	4.266335562	4.283269153	4.24607745	4.281992206	4.30901929		
19	10338063	4.225048553	4.259822836	4.212457699	4.2056999	4.214572437	4.174237693	4.19236055	4.211607208		
20	10338064	5.819037388	6.043330703	6.201893486	5.774275043	5.768607083	5.793544556	6.022835011	6.120632578		
21	10338065	6.352138566	6.633064097	6.782333391	6.326170161	6.306145322	6.285299134	6.42007228	6.655293247		
22	10338066	4.910927396	5.06311953	5.106761426	4.914891606	4.903922725	4.948643348	5.090043501	5.236642491		
23	10338067	7.714886054	8.058697609	8.160952442	7.675949218	7.618620174	7.699587081	7.6943987	7.785893965		
24	10338068	8.638172472	8.61654494	8.40195936	8.751163569	8.604806493	8.523999215	7.863958602	7.707226069		
25	10338069	5.455789648	4.96948286	5.959576471	5.370045266	4.970349606	5.308601645	5.828781389	5.546517292		
26	10338070	4.774529463	4.903033703	4.623856664	4.596480622	4.715788631	5.12415018	4.730363926	4.474340236		
27	10338071	4.75308505	4.79812396	4.647918368	4.417310338	4.500078278	4.342114367	4.532176037	4.467510215		
28	10338072	4.842275309	5.01043487	4.982517004	4.891179128	5.304482953	4.932346088	4.755200176	5.175230878		
29	10338073	7.247722666	6.520938602	6.640031052	7.070276903	7.014413961	6.456815841	5.605120557	6.315544201		
30	10338074	8.104924033	8.408530411	8.401969636	8.18223739	7.404345862	8.010868736	8.534637082	8.541650866		
31	10338075	6.175731245	6.4971469	6.284237422	5.769584555	6.348672189	5.980803102	7.774690082	7.895733236		
32	10338076	6.169678012	6.581742519	6.752830992	5.887165144	6.295685281	5.750888821	6.699114267	6.485506814		
33	10338077	9.377591073	9.164535392	9.734262878	8.698280769	9.302784155	9.557002796	9.557002796	10.264837369		
34	10338078	5.784767782	5.640911444	5.436511457	5.823124266	5.709017733	5.988510278	5.395258931	5.007097466		
35	10338079	5.26842573	5.595138549	5.540079643	5.613896933	5.286217301	5.069353539	5.490831462	5.360131623		
36	10338080	6.00730114	6.00730114	6.00730114	6.00730114	6.00730114	6.00730114	6.00730114	6.00730114		

# 課題

- GEOデータベース、アクセション番号 GSE26910のデータを取得し、正規化して結果をテキストファイルに出力してください。
- 今回の実習で使ったパッケージでは足りないものがあります。
- ヒント: サンプルを採取した細胞は？



# 第2回

- 正規化後のデータを可視化
- 散布図
- ヒートマップ
- その他