PictBio

統計ソフトRを使用した次世代シーケンサーNGSデータ解析

アクシオヘリックス株式会社

AXIOHELIX

アクシオヘリックス株式会社について

本社

沖縄県那覇市

東京支社

東京都千代田区神田和泉町

代表取締役社長

シバスンタラン スハルナン

創立

2001年6月8日

事業ドメイン

ライフサイエンス

ＩＴソリューション

ビジネス
ディベロップメント

PictBio

アジェンダ

1. Rの使い方
2. RをつかったRNA-Seq解析
3. RをつかったBisulfite解析

1.Rの使い方

https://www.pictbio.com/tips/2581.html

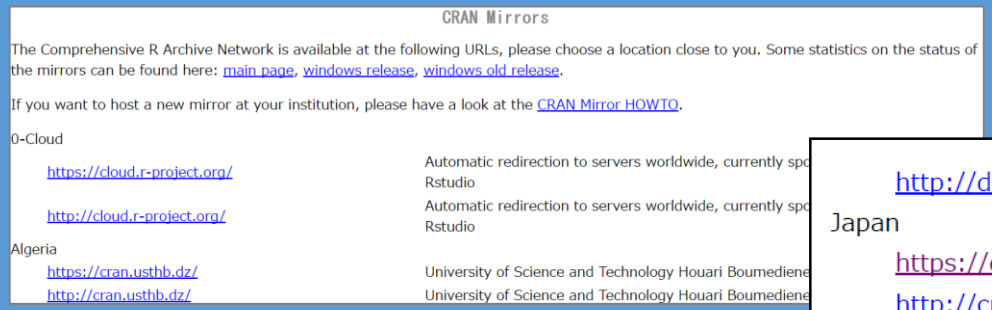
Rの導入方法

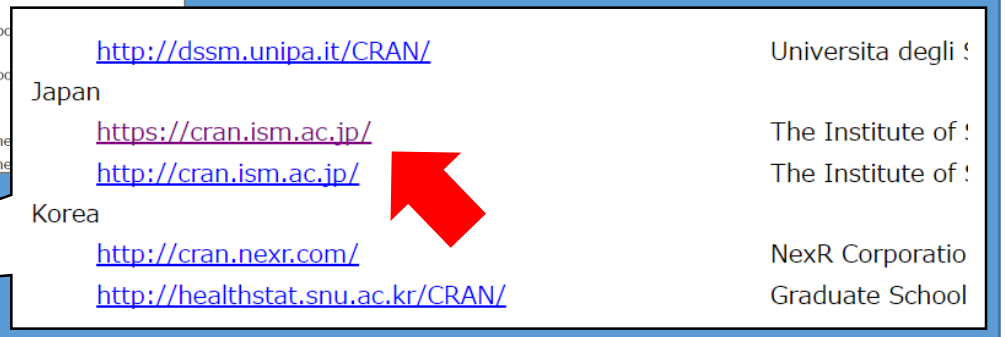
Rのダウンロード

1. HPのダウンロードリンクをクリック

https://www.r-project.org/

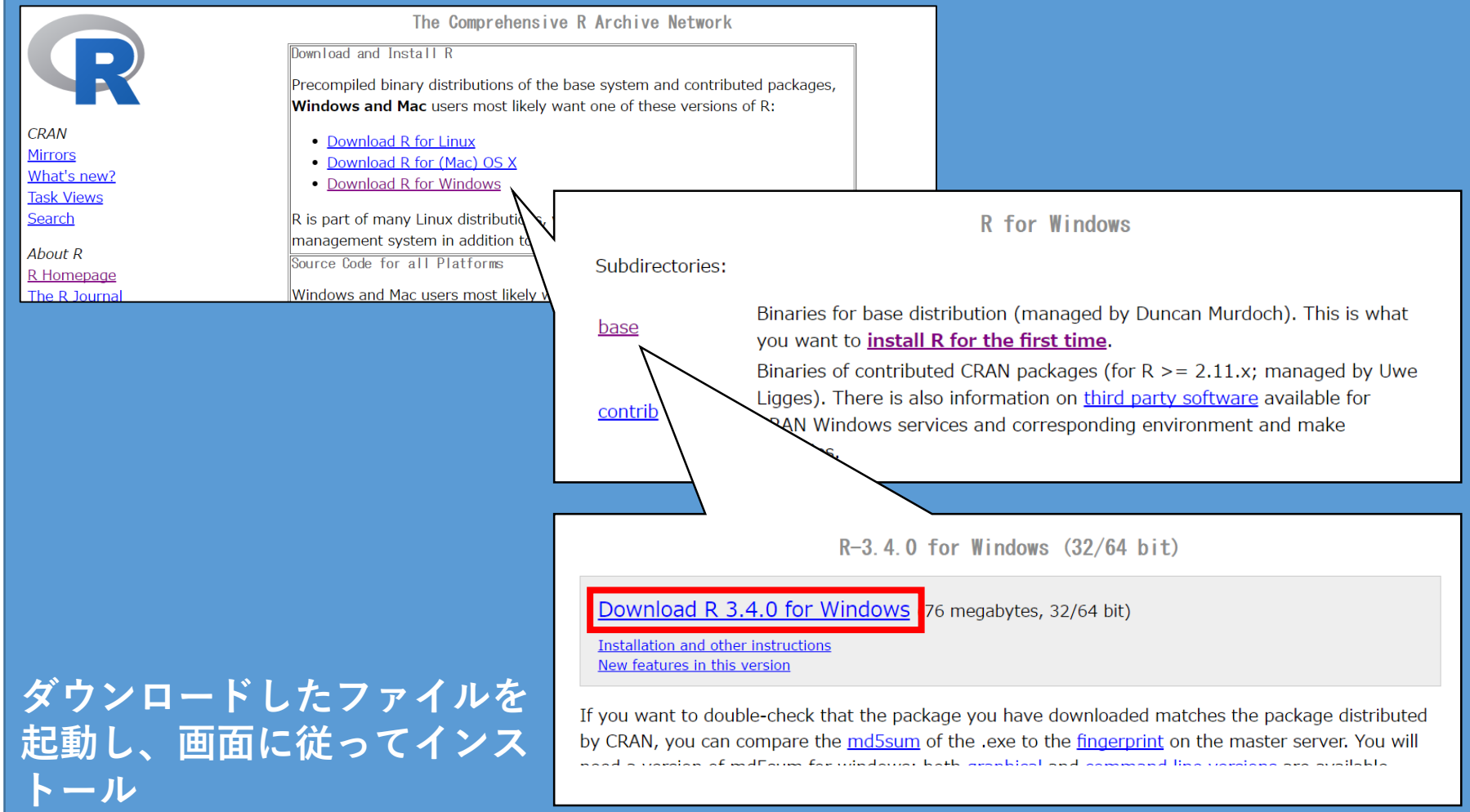
2. ミラーサイトリストからJapanのサイトを選択





Rのダウンロード～インストール

3. OSを選択して必要なファイルをダウンロード

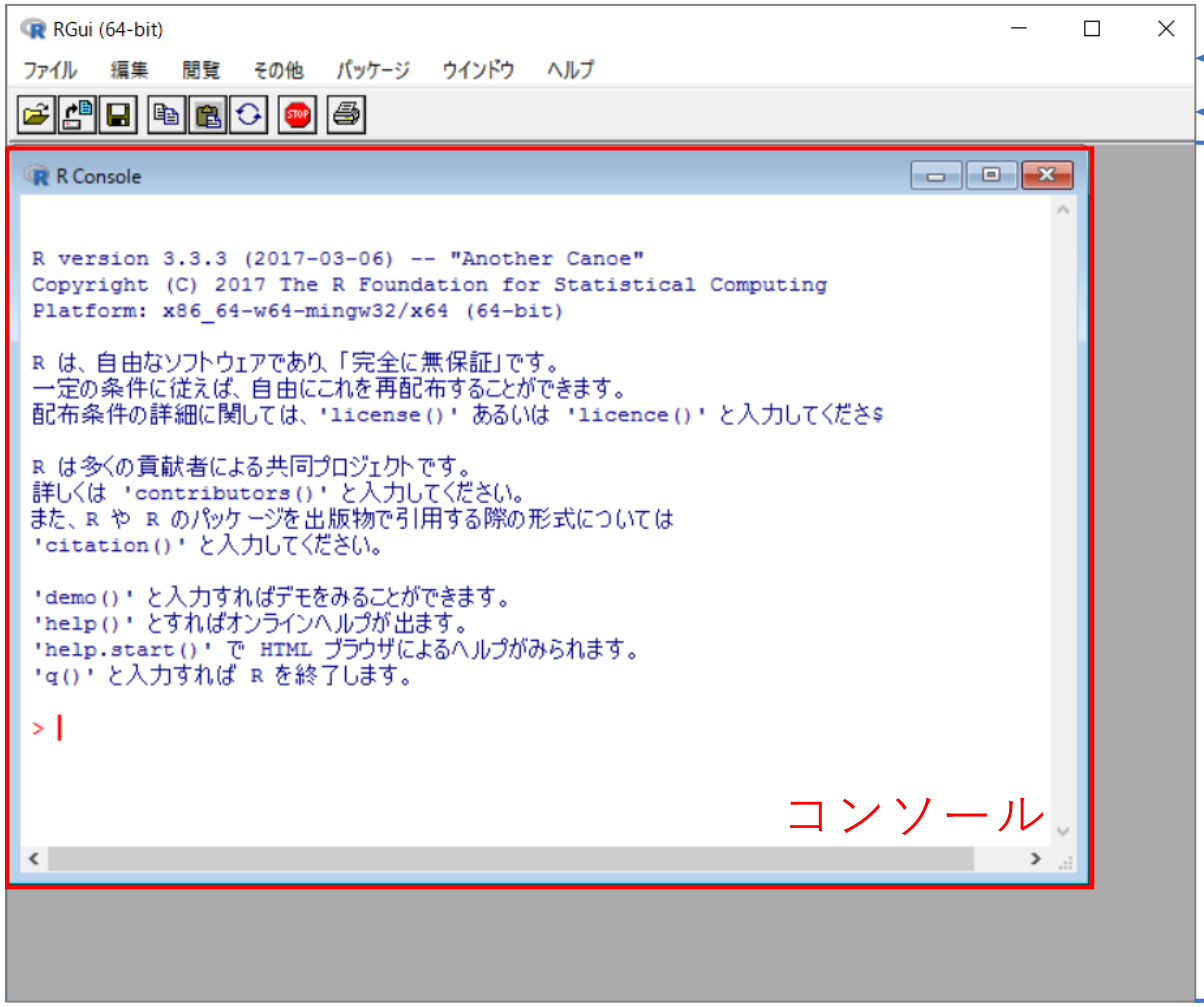


ダウンロードしたファイルを起動し、画面に従ってインストール

Rの操作

Rの起動

デスクトップのRアイコンで起動



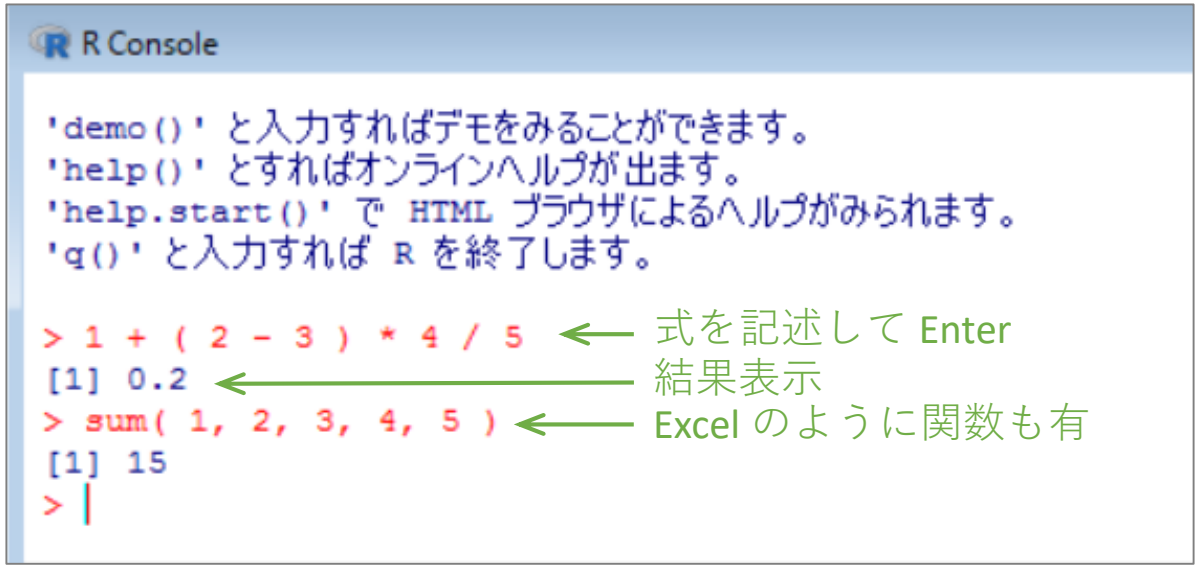
メニューバー

よく使われるメニュー機能の
クイックアクセス

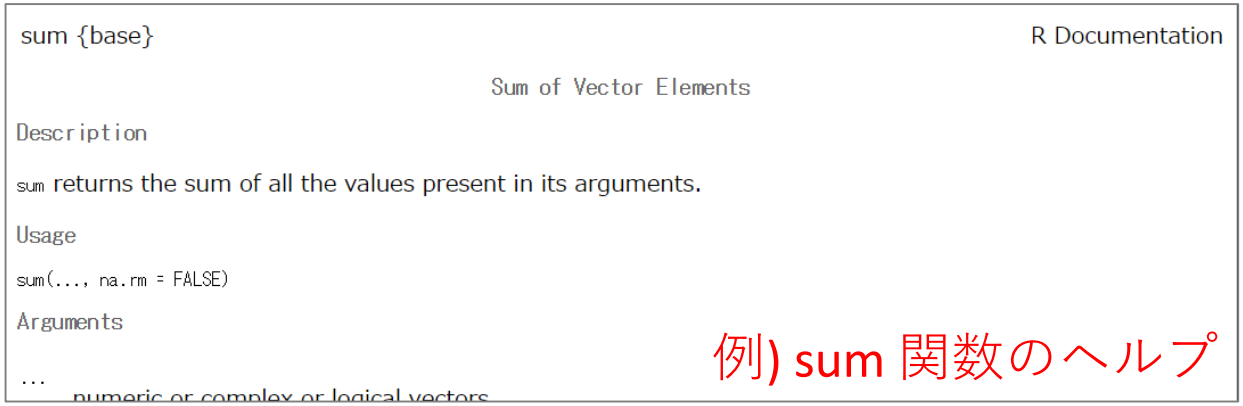
計算やプロットを行う
作業画面領域

簡単な計算

コンソールに計算式を打つことで計算を実行



関数の使い方の調べ方は、コンソールに「help(関数)」と打つ


sum {base} R Documentation

Sum of Vector Elements

Description

sum returns the sum of all the values present in its arguments.

Usage

sum(..., na.rm = FALSE)

Arguments

... numeric or complex or logical vectors

例) sum 関数のヘルプ

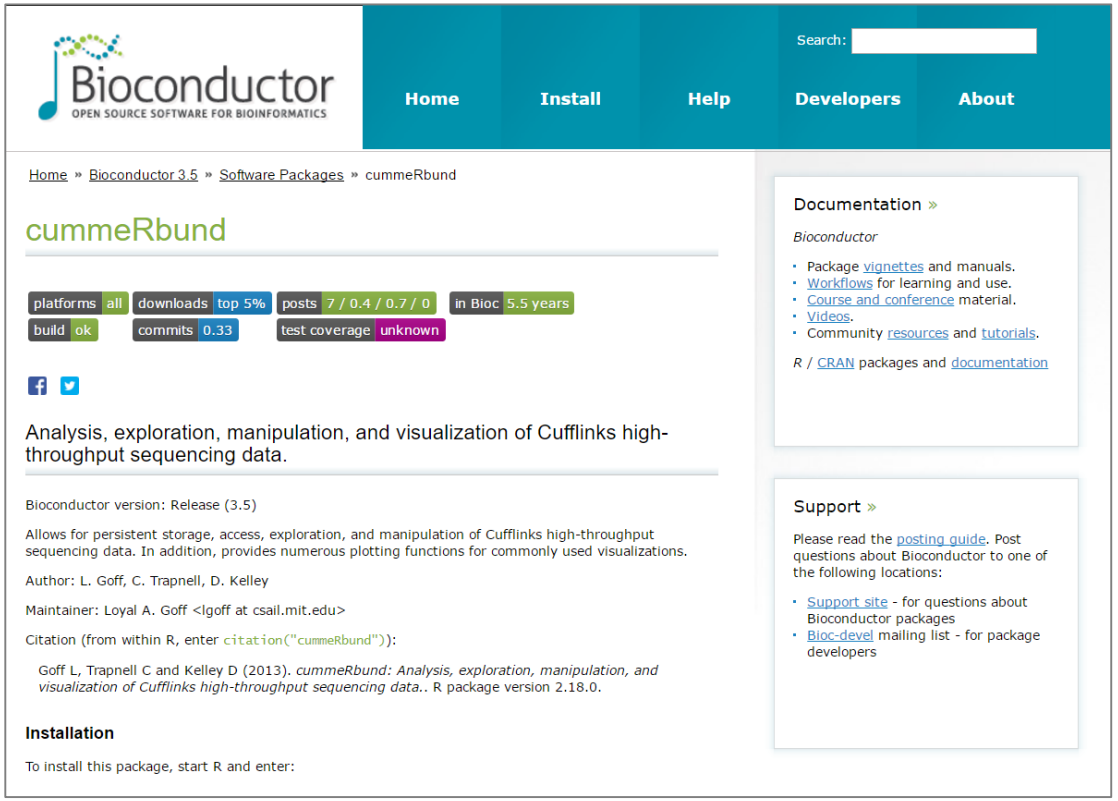
パッケージ

Rは有志が作成した関数セット（パッケージ）を読み込んで使うことができる。※

バイオインフォマティクスに関するパッケージはBioConductorで提供されている

※ CRANで提供されているパッケージ
メニューバーの「パッケージ」→「パッケージのインストール」でインストールすることも可能

例) BioconductorのcummeRbundページ



BioConductorパッケージのインストール

パッケージはインストールが必要

```
R Console

> source("https://bioconductor.org/biocLite.R")
 install.packages("BiocInstaller", repos = a["BioCs
  'lib = "C:/Program Files/R/R-3.3.3/library"' is r
 URL 'https://bioconductor.org/packages/3.4/bioc/bi
Content type 'application/zip' length 126079 bytes
downloaded 123 KB

package 'BiocInstaller' successfully unpacked and M

The downloaded binary packages are in
        C:\Users\nishi\AppData\Local\Temp\RtmpCEkbb
Bioconductor version 3.4 (BiocInstaller 1.24.0), ?b
A new version of Bioconductor is available after in
  version of R; see http://bioconductor.org/install
> biocLite("cummeRbund")
```

← BioLite関数を呼び出し

← BioLite関数を使用しインストール

パッケージ関数を使う際は呼び出しが必要

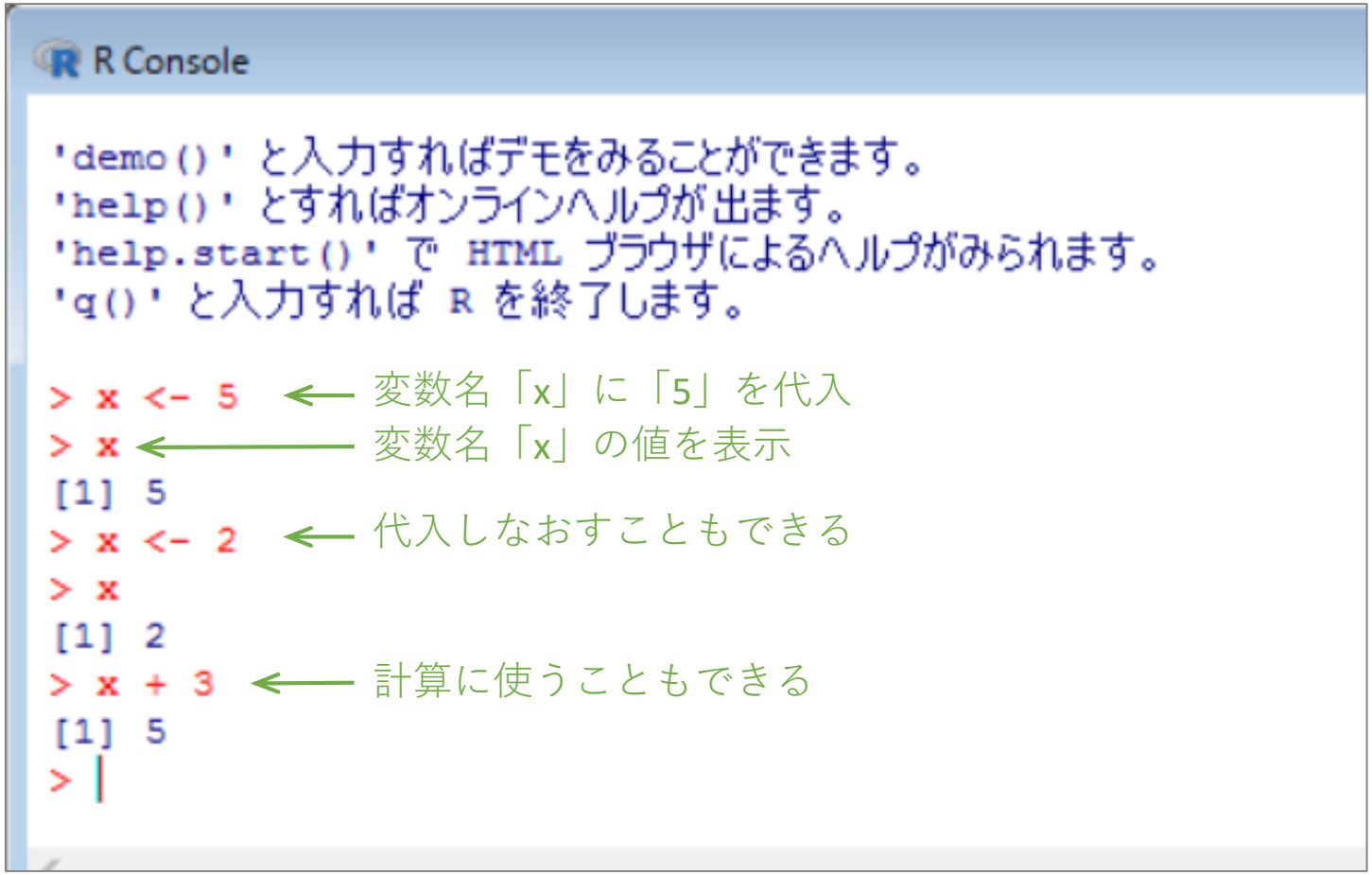
```
        C:\Users\nishi\AppData\Local\Temp\RtmpCEkbb
installation path not writeable, unable to update p
  cluster, foreign, lattice, MASS, Matrix, rpart, s
> library("cummeRbund")
```

← パッケージ名を指定して呼び出す

変数と代入

何かの値を入れた「変数」を扱うことができる

代入に使用する演算子は「<-（小なりハイフン）」や「=（イコール）」

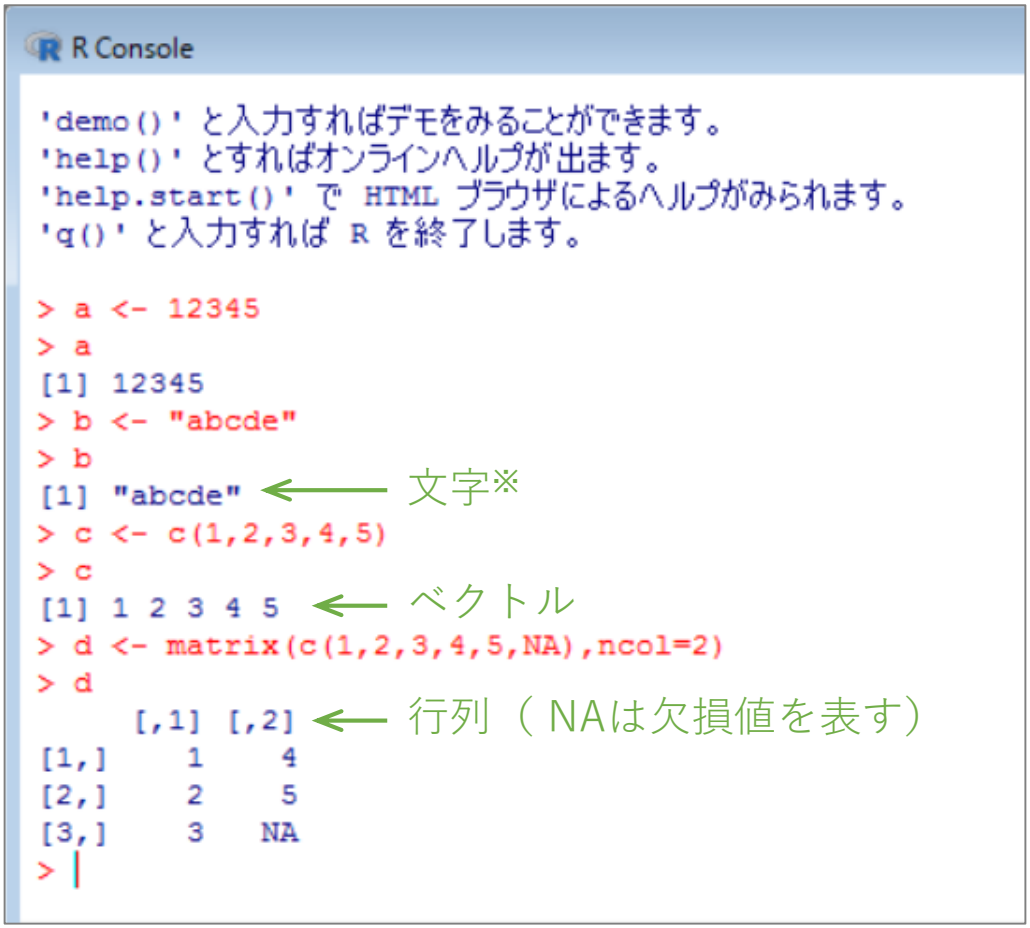


変数名には半角英数を使うことができる（数値を先頭にはできない）

データ型

Rで扱う値には「データ型」という属性が存在する

特定のデータ型を要求する関数が存在する



※ 変数に文字を代入する場合はダブルクオートで囲う

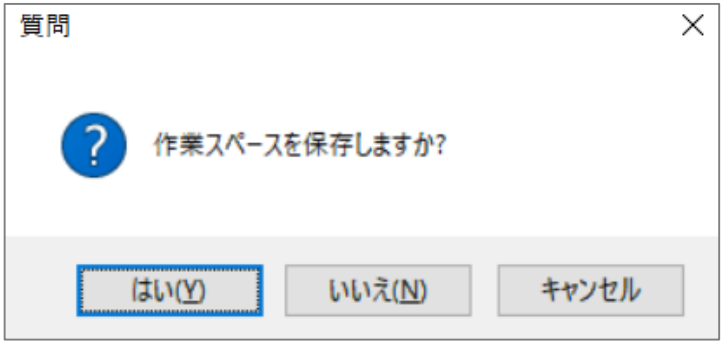
Rの終了と保存

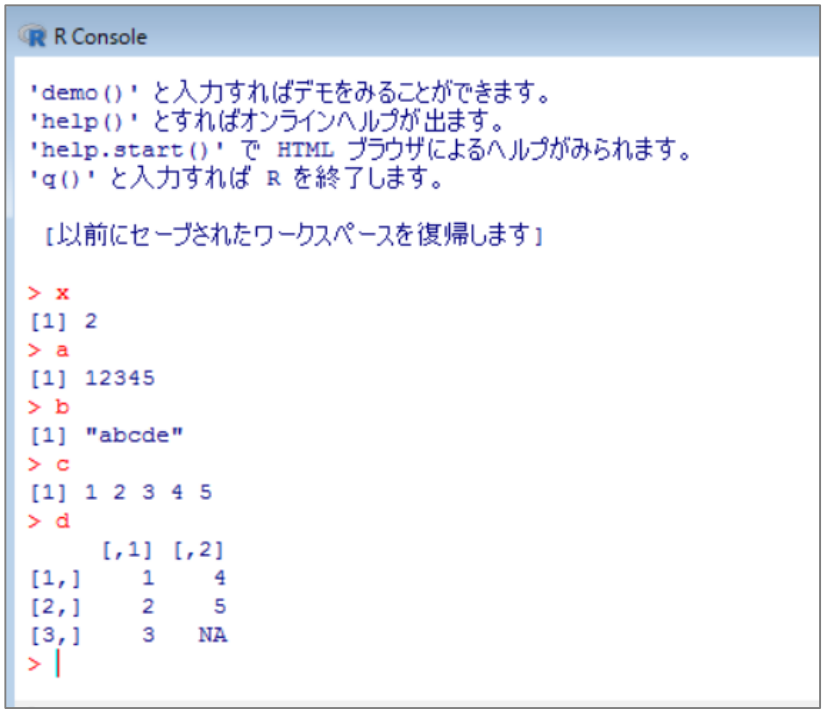
コンソールに「q()」と打つか右上の閉じるボタンで終了する

保存するか聞かれるのでどちらかを選択する

保存を選択した場合、作業フォルダ※に「.RData」ファイルが作成される

「.RData」を読み込むと保存した変数が復帰する





保存は作業中でもメニューバーから可能

Rの使い方

- 計算用の関数やパッケージがたくさんある
- 変数に数値などの値を保存できる
- 値にはデータ型という属性がある

2.Rを使った RNA-Seq解析

RNA-Seqの一般的なデータ解析フロー

シーケンスデータ

FASTQファイル

クオリティ
コントロール

マッピング

TopHatツール

発現量計算

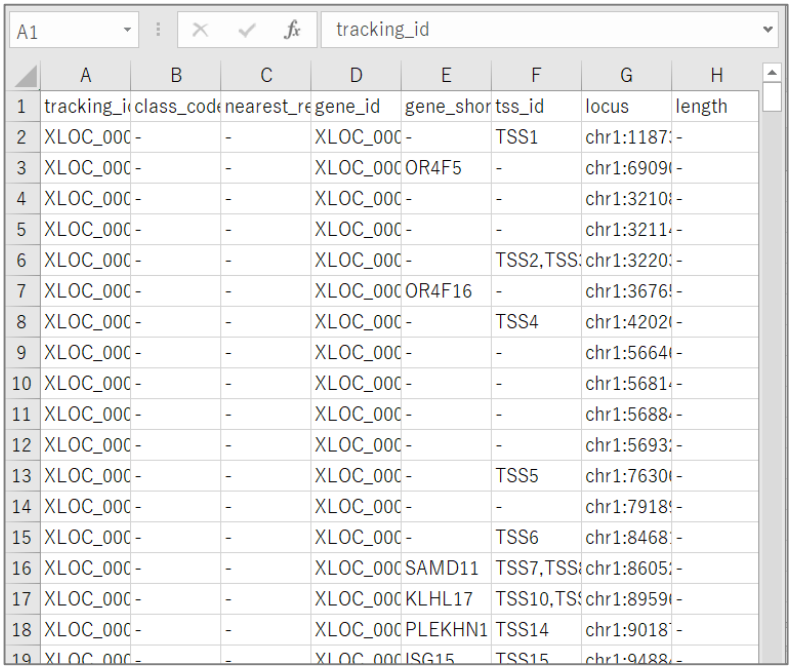
発現量比較統計

Cufflinksツール

結果の数値表

Rでプロット

エクセルに読み込んで表示できる



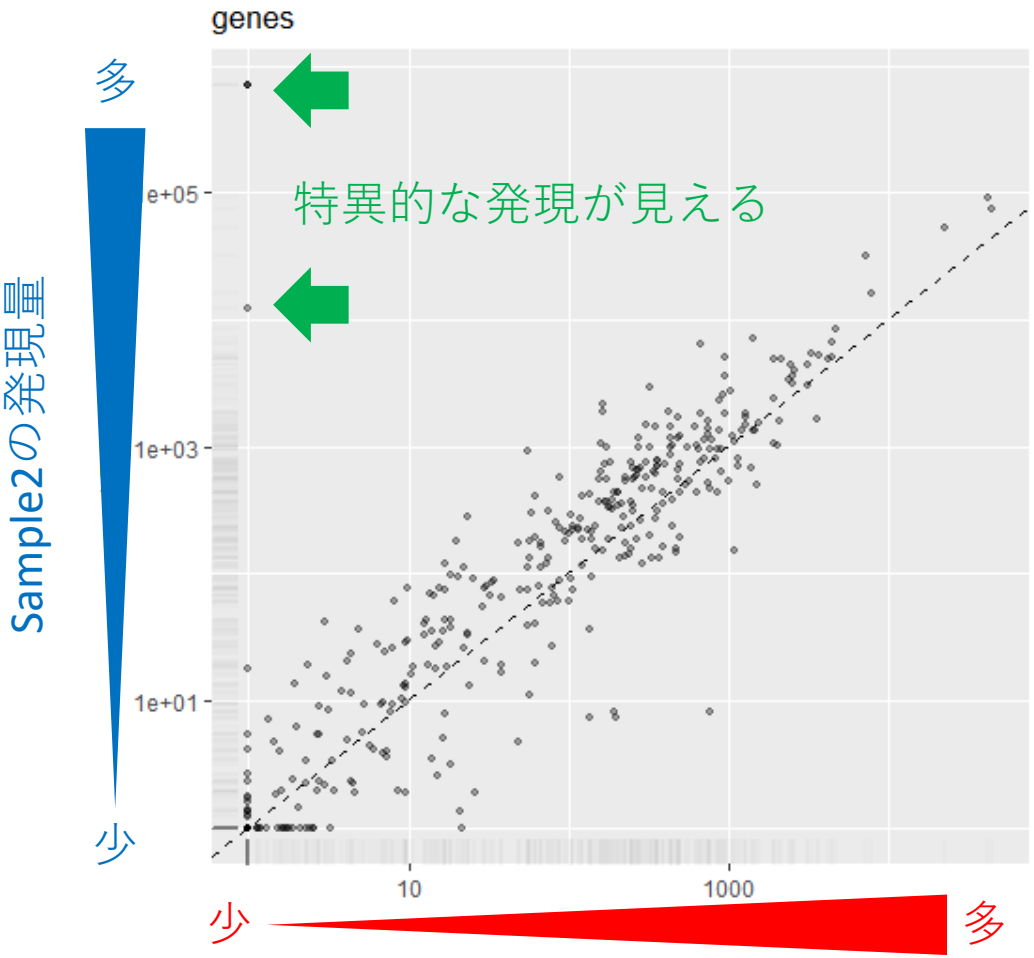
	A	B	C	D	E	F	G	H
1	tracking_i	class_cod	nearest_re	gene_id	gene_shor	tss_id	locus	length
2	XLOC_000	-	-	XLOC_000	-	TSS1	chr1:1187	-
3	XLOC_000	-	-	XLOC_000	OR4F5	-	chr1:6909	-
4	XLOC_000	-	-	XLOC_000	-	-	chr1:3210	-
5	XLOC_000	-	-	XLOC_000	-	-	chr1:3211	-
6	XLOC_000	-	-	XLOC_000	-	TSS2,TSS	chr1:3220	-
7	XLOC_000	-	-	XLOC_000	OR4F16	-	chr1:3676	-
8	XLOC_000	-	-	XLOC_000	-	TSS4	chr1:4202	-
9	XLOC_000	-	-	XLOC_000	-	-	chr1:5664	-
10	XLOC_000	-	-	XLOC_000	-	-	chr1:5681	-
11	XLOC_000	-	-	XLOC_000	-	-	chr1:5688	-
12	XLOC_000	-	-	XLOC_000	-	-	chr1:5693	-
13	XLOC_000	-	-	XLOC_000	-	TSS5	chr1:7630	-
14	XLOC_000	-	-	XLOC_000	-	-	chr1:7918	-
15	XLOC_000	-	-	XLOC_000	-	TSS6	chr1:8468	-
16	XLOC_000	-	-	XLOC_000	SAMD11	TSS7,TSS	chr1:8605	-
17	XLOC_000	-	-	XLOC_000	KLHL17	TSS10,TS	chr1:8959	-
18	XLOC_000	-	-	XLOC_000	PLEKHN1	TSS14	chr1:9018	-
19	XLOC_000	-	-	XLOC_000	ISG15	TSS15	chr1:9488	-

cummeRbund

Cufflinksの結果を読み込んで加工することができる関数を持ったパッケージ

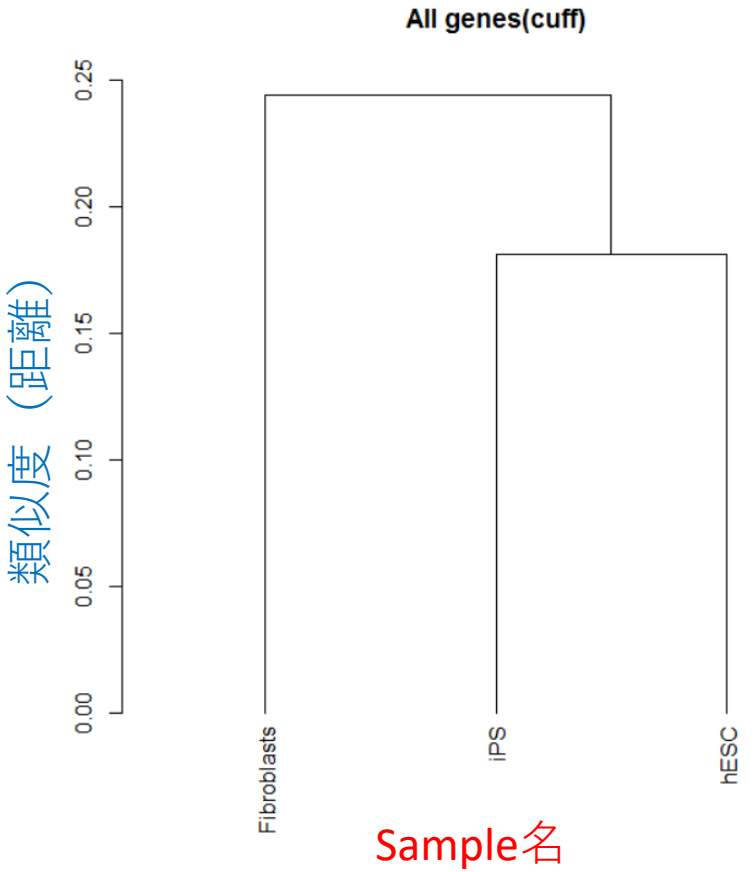
RNA-Seq解析のプロット [2群間比較]

遺伝子発現量（FPKM値）の散布図



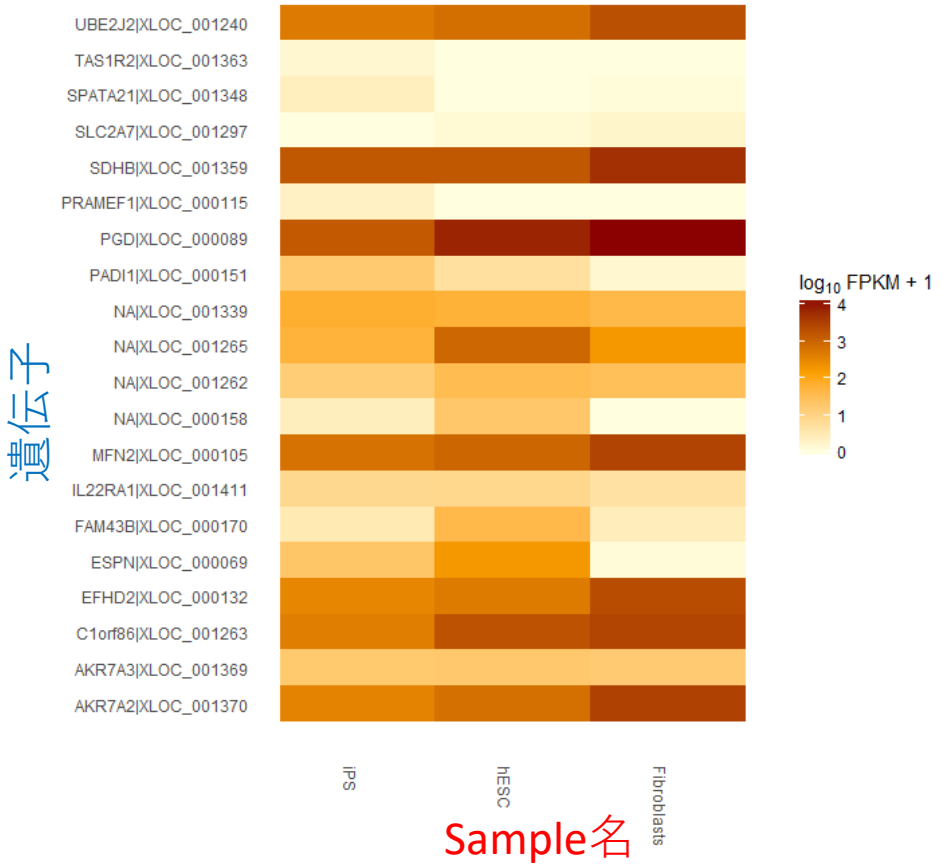
RNA-Seq解析のプロット [N群間比較]

サンプル間クラスタリングツリー



統計的に距離が近い群がわかる

遺伝子発現量（FPKM値）のヒートマップ



発現パターンが似ている群が
目検できる

解析データの読み込み

パッケージの呼び出し

```
> library(cummeRbund) #パッケージの呼び出し
```

サンプルデータの読み込み※

```
> file <- system.file("extdata",package="cummeRbund") #Rのファイル情報

> cuff <- readCufflinks(dir=file)
```

データのあるフォルダ

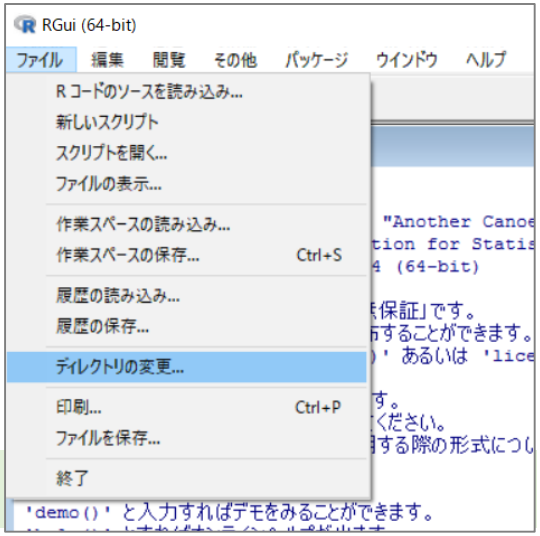
※ cummeRbundが用意している

自分のデータを読み込みたい時

作業フォルダの変更 → データの読み込み

作業フォルダ変更

1. メニューバーからファイルを選択
2. ディレクトリの変更
3. データを格納しているフォルダを選択




```
> cuff <- readCufflinks()
```

2群間比較

データの確認

```
> replicates(cuff)
```

```
> replicates(cuff)
           file sample_name replicate      rep_name total_mass norm_mass internal_scale external_scale
1  iPS_rep1.bam         iPS         0         iPS_0     173431    706934       0.958068       0.584877
2  iPS_rep2.bam         iPS         1         iPS_1     173007    706934       1.037970       0.584877
3   H1_rep1.bam        hESC         0        hESC_0     754749    706934       0.989851       1.513060
4   H1_rep3.bam        hESC         1        hESC_1     762643    706934       1.010250       1.513060
5 NHLF_rep1.bam Fibroblasts         0 Fibroblasts_0     876775    706934       0.840416       1.223240
6 NHLF_rep2.bam Fibroblasts         1 Fibroblasts_1    1412130    706934       1.198470       1.223240
```

グループ名　　各データ（replicate）の名前

遺伝子発現量（FPKM値）の散布図

```
> cuffgene <- genes(cuff) #遺伝子レベルのデータの取り出し

> csScatter(cuffgene, "iPS", "hESC")
```

プロットする2つを選択

プロットの保存

1. プロットのウインドウを選択
2. メニューバーのファイルを選択
3. 別名で保存を選択
4. 好きな画像フォーマットを選択

N群間比較

サンプル間クラスタリング

```
> csDendro(cuffgene)
```

遺伝子発現量（FPKM値）のヒートマップ

対象遺伝子データの取り出し → プロット

```
> data(sampleData) #サンプルデータセットの呼び出し
> gene_vct <- sampleIDs #サンプルデータセットの対象遺伝子ベクトル
> cuffget <- getGenes(cuff,gene_vct) #指定した遺伝子のデータの取り出し
> csHeatmap(cuffget)
```

対象遺伝子ベクトルを自分で作成する時

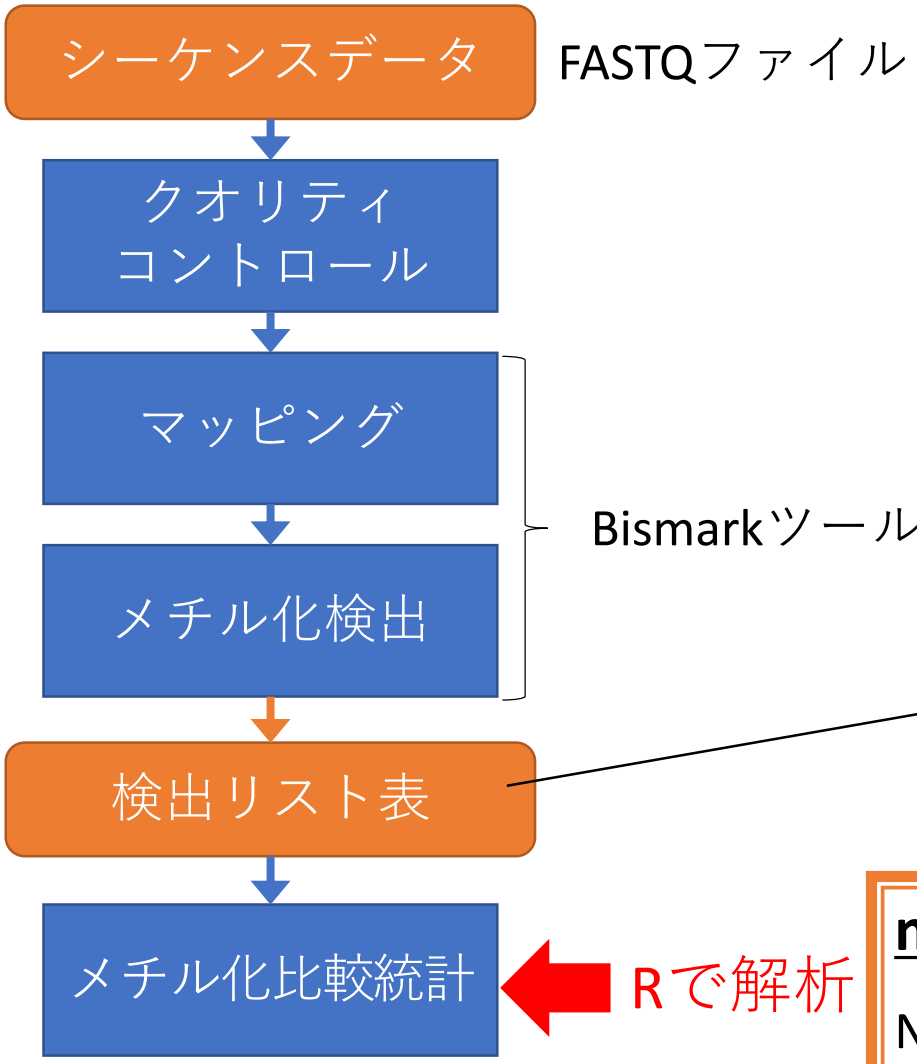
```
> gene_vct <- c("XLOC_000004","XLOC_000005","XLOC_000008","XLOC_000009","XLOC_000011")
```

Rを使ったRNA-Seq解析

- Cufflinksは統計まで行ってくれる
- cummeRbundパッケージはサンプルのクラスタリングやプロット等、RNA-Seq解析に便利な関数がそろっている
- RNA-Seq解析の方法は多岐にわたるためパッケージが豊富

3.Rを使った Bisulfite解析

Bisulfiteの一般的なデータ解析フロー



メチル化量まで得られる

A1 | CpG context

	A	B	C	D	E
1	CpG context				
2	==========				
3	position	count met	count unm	% methyla	coverage
4	1	7	3	70	10
5	2	0	0		0
6	3	0	0		0
7	4	0	0		0
8	5	0	0		0
9	6	0	0		0
10	7	0	0		0
11	8	0	0		0
12	9	0	0		0
13	10	0	0		0
14	11	0	0		0
15	12	0	0		0
16	13	0	0		0
17	14	9	1	90	10
18	15	0	0		0

methylKit

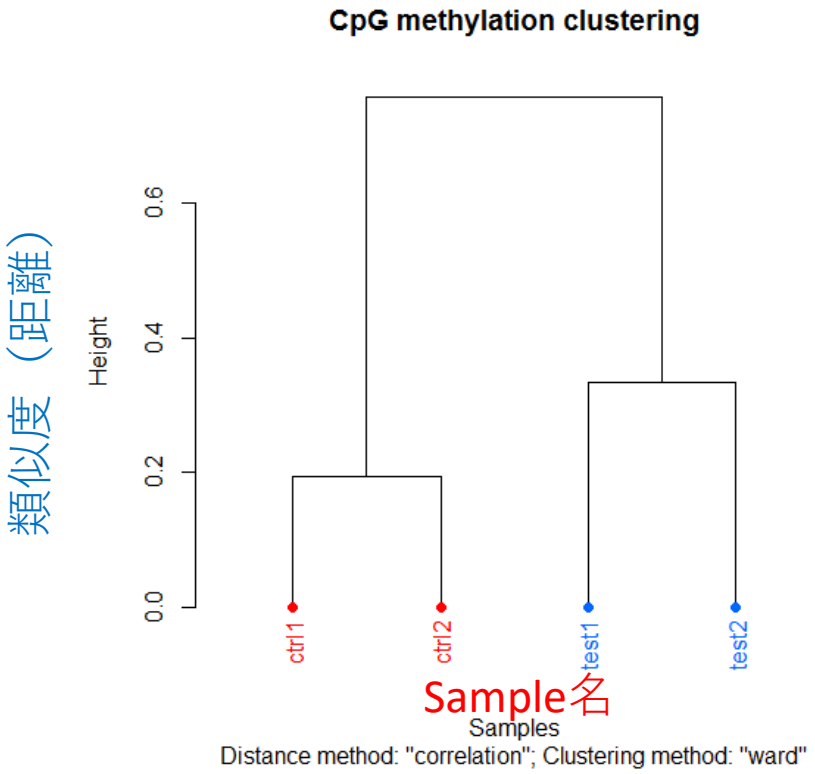
NGSを使ったBisulfite解析の結果を読み込んで統計解析ができるパッケージ

Bisulfite解析の比較

◆検出したメチル化の有意差検定

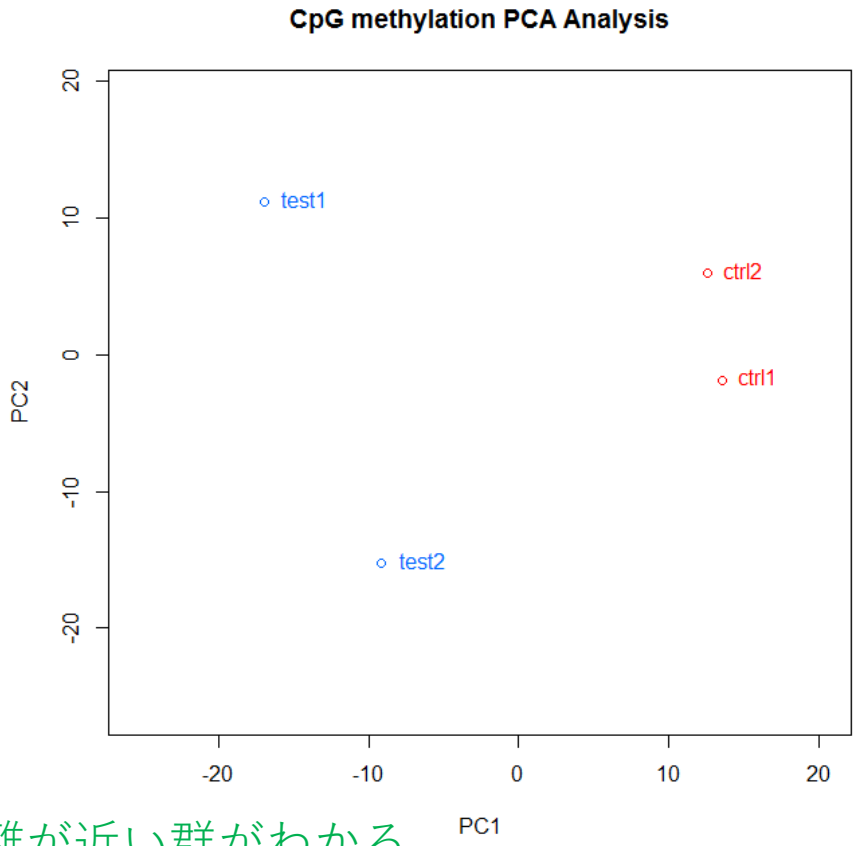
◆サンプル間の類似度検定

サンプル間クラスタリングツリー



類似度（距離）

PCA解析



統計的に距離が近い群がわかる

解析データの読み込み [単一]

パッケージの呼び出し

```
> library(methylKit) #パッケージの呼び出し
```

サンプルデータを読み込む※

※ methylKitが用意している

```
> file <- system.file("extdata","test.fastq_bismark.sorted.min.sam",
                      package="methylKit") #Rのファイル情報
> file_list <- list(file)
> sample_list <- list("test") #ファイルに対応したサンプル名
> group_vct <- c(1) #ファイルに対応したグループナンバー※
> bism <- processBismarkAln(file_list,sample_list,treatment= group_vct,
                            assembly="hg18",read.context="CpG")
```

リファレンスを設定　CpGデータを指定

※ 同じグループは同じ数字にする

自分のデータを読み込みたい時

作業フォルダの変更 → データの読み込み

```
> file_list <- list("test1.bam") #BAM（SAM）ファイルのみ指定
```

データの読み込み [複数]

サンプルデータを読み込む※

ファイル、サンプル、グループ変数に追加、読み込む際の関数はmethReadを使う

```
> file1 <- system.file("extdata","test1.myCpG.txt",package="methylKit")
> file2 <- system.file("extdata","test2.myCpG.txt",package="methylKit")
> file3 <- system.file("extdata","control1.myCpG.txt",package="methylKit")
> file4 <- system.file("extdata","control2.myCpG.txt",package="methylKit")
> file_list <- list(file1,file2,file3,file4)
> sample_list <- list("test1","test2","ctrl1","ctrl2")
> group_vct <- c(1,1,0,0) #test1と2がreplicate、control1と2がreplicate
> bism <- methRead(file_list,sample_list,treatment=group_vct,
                                    assembly="hg18",context="CpG")
```

※ Bismarkの複数データサンプルは用意されていない

自分のデータを読み込みたい時

同様に変数に追加、読み込む際の関数はそのままprocessBismarkAlnを使う

```
> file_list <- list("test1.bam","test2.bam","control1.bam","control2.bam")
```

メチル化比較

検出したメチル化の有意差検定

```
> meth <- unite(bism) #統計用データの取り出し
> diff <- calculateDiffMeth(meth)
> methdiff <- getData(diff) #表データの取り出し
> write.csv(methdiff, "sample.csv") #CSVファイルフォーマットで書き出し
```

	A	B	C	D	E	F	G	H
1		chr	start	end	strand	pvalue	qvalue	meth.diff
2	1	chr21	9853296	9853296	+	0.009908	0.021566	-7.01211
3	2	chr21	9853326	9853326	+	0.947355	0.592173	0.209136
4	3	chr21	9860126	9860126	+	0.041604	0.069781	-4.11158
5	4	chr21	9906604	9906604	+	0.210652	0.259453	-7.34637
6	5	chr21	9906616	9906616	+	0.001576	0.004322	18.8175
7	6	chr21	9906619	9906619	+	0.002825	0.007322	14.19373

サンプル間の類似度検定

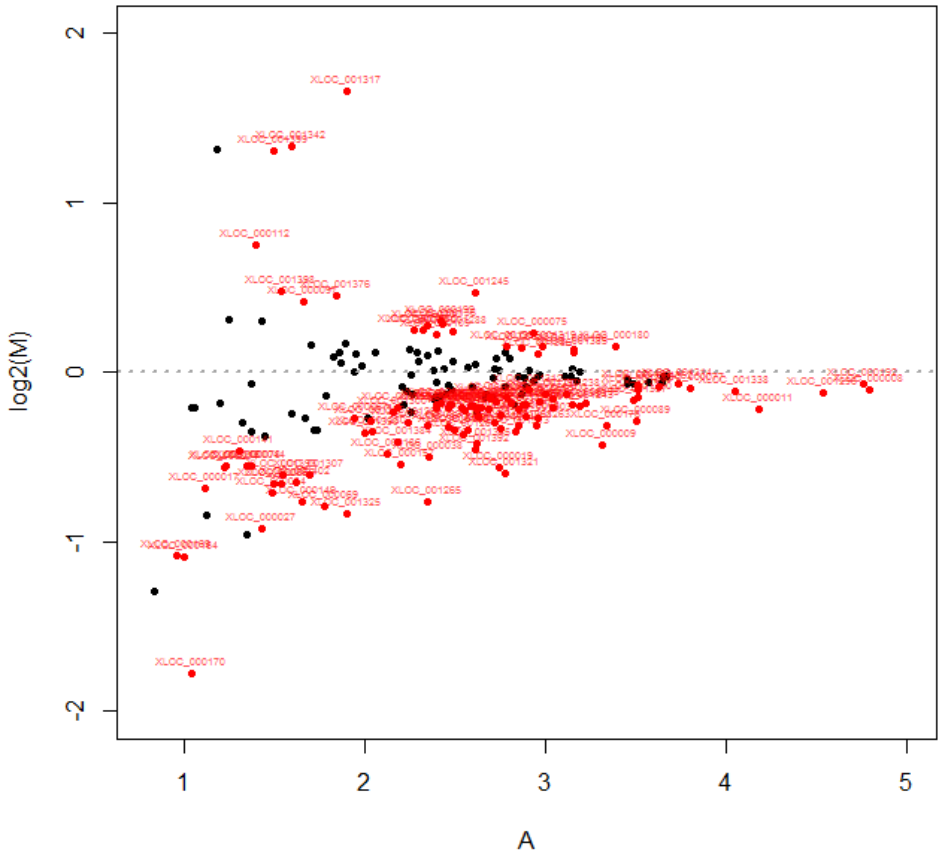
```
> clusterSamples(meth) # サンプル間クラスタリングツリー
> PCASamples(meth) # PCA解析
```

Rを使ったBisulfite解析

- Bismarkはメチル化のリストアップまで行ってくれる
- methylKitパッケージはメチル化の有意差検定などの比較統計やサンプルのクラスタリング、プロット等、Bisulfite解析に便利な関数がそろっている
- Bisulfite解析の方法はBismarkが多用されているためパッケージは少ない

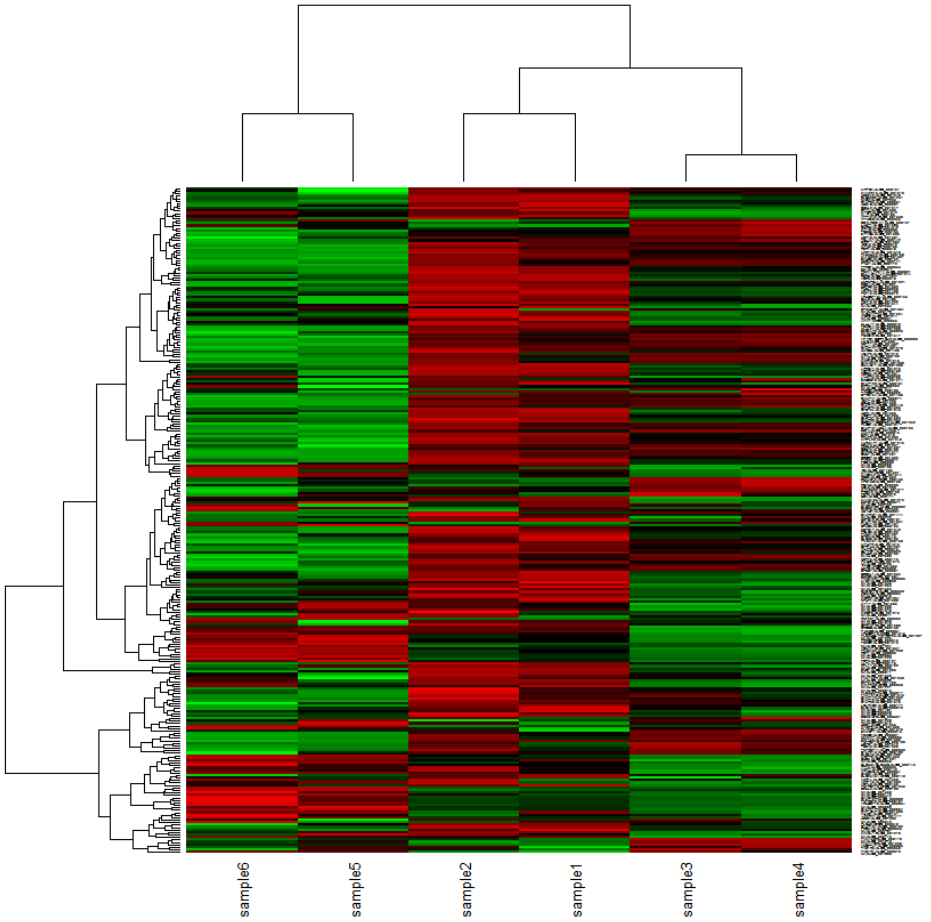
弊社作成サンプルプロット（１）

MAプロット



有意差のある発現変動が見られた遺伝子に色を付け、IDを表示

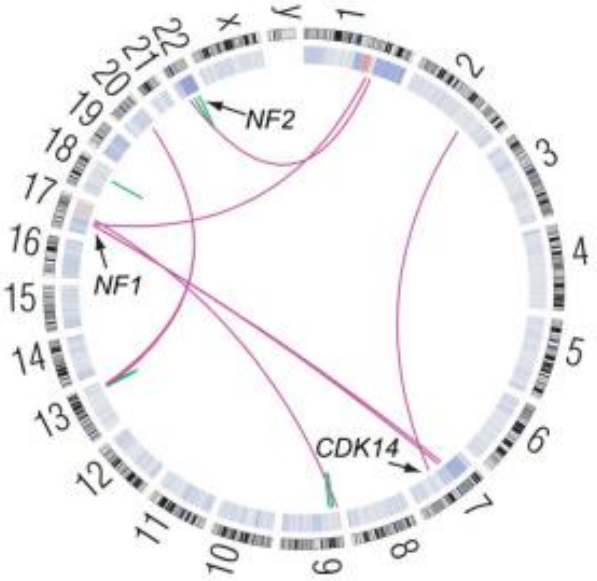
デンドログラム付きヒートマップ



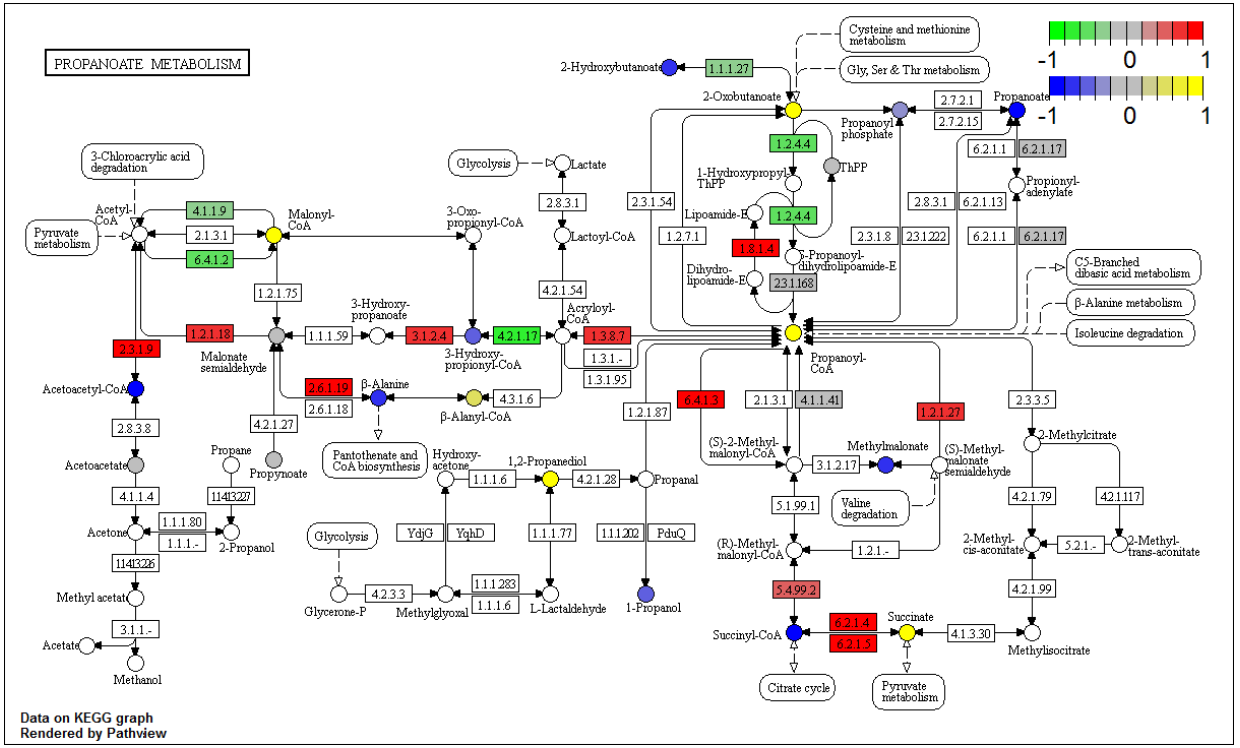
全遺伝子を表示、カラーを選択

弊社作成サンプルプロット（2）

染色体構造変異の
サーコスプロット



Brastianos et al.(2013)



色でスケール等を表したパスウェイ図

PictBioサービスの紹介

NGS解析サービスの種類と特徴

- シーケンス

  NGSを使ったシーケンス解析を行う。

  フラグメントの状態を送付することが多い。

  生物種と解析の種類によって、機器や調整手法を選んでくれるところもある。

- データ解析

  - パッケージ

    メニュー化された解析で低コスト。納品物は一定の形式に従う。フォローサービスは少ない。

  - オーダーメイド

    研究目的に沿った解析フローから作成。または、依頼フローに従い解析。フォローサービスが充実。

PictBioのご紹介

- シーケンス、データ解析（パッケージ、オーダーメイド）**すべてお取り扱い**
- シーケンス前の**計画段階からご相談可**
- 研究者の方の**ご都合に合わせたご提案**

ご清聴、有難うございました。

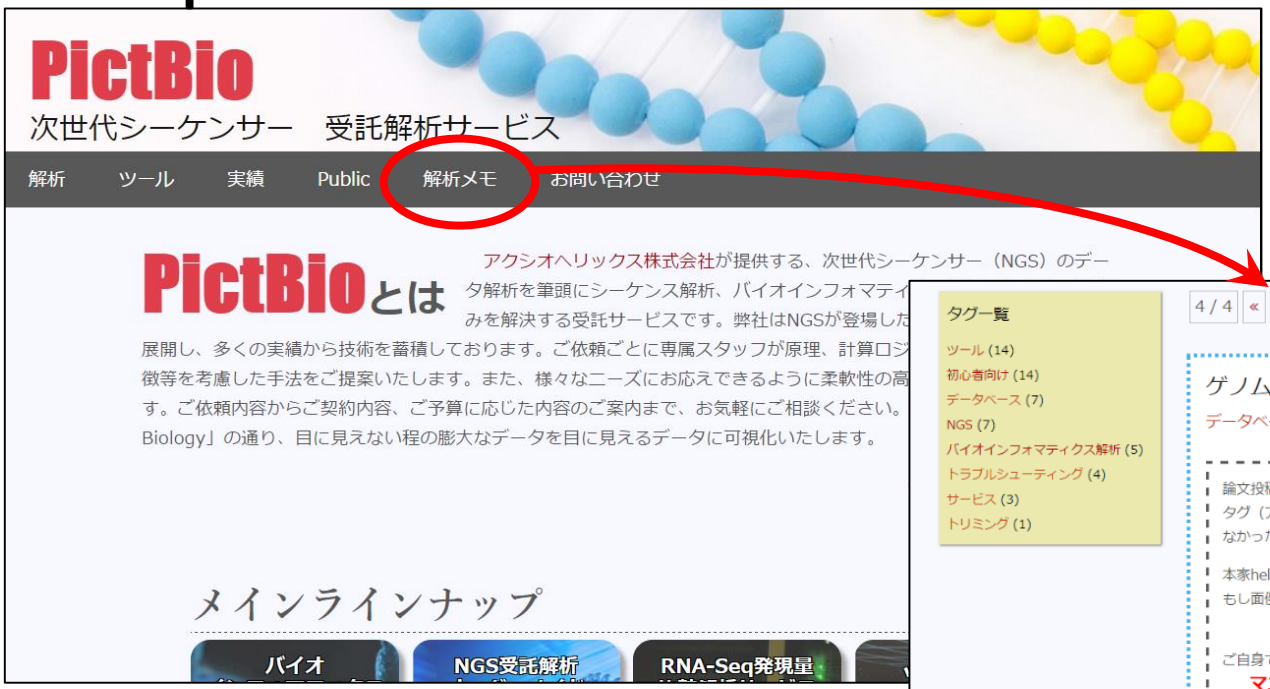
こちらに資料に関するお問い合わせは

pictbio@axiohelix.com

へお願いいたします。

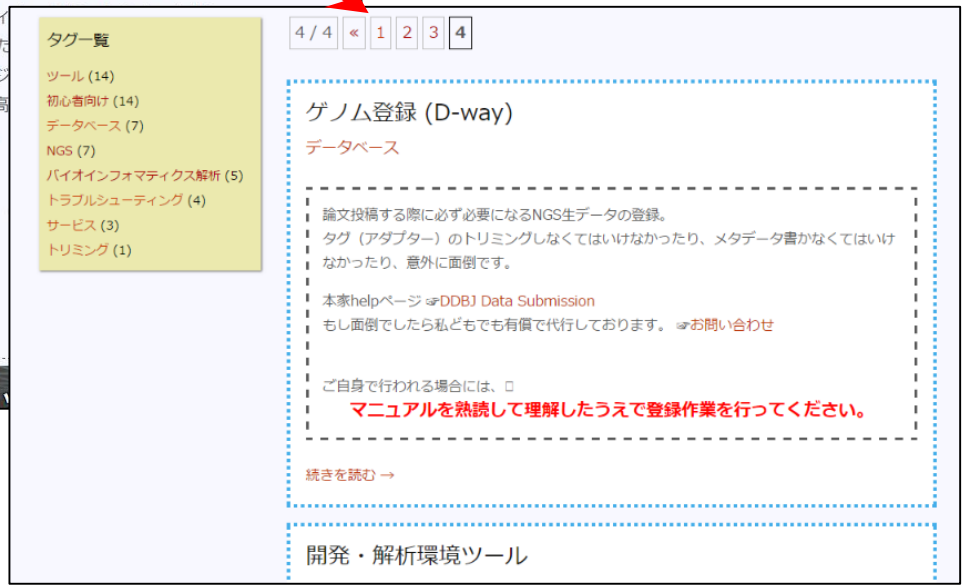
ホームページの紹介

www.pictbio.com



おすすめコンテンツ

解析メモ



Twitter account : @PictBio