2019.07.05版

Rを起動し、library(recount)と打ち込んで、recountパッケージがインストールされていることを確認しておいてください。
参考と書いてあるスライドは飛ばします

農学生命情報科学特論I 第2回

[1]大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究プログラム
[2]微生物科学イノベーション連携研究機構
門田幸二（かどた　こうじ）
kadota@iu.a.u-tokyo.ac.jp
http://www.iu.a.u-tokyo.ac.jp/~kadota/

予定です

講義予定

- 第1回（2019年07月01日）
  - カウント情報取得の続き
  - データの正規化（RPK, RPM, RPKM/FPKM）
- 第2回（2019年07月08日）
  - サンプル間クラスタリング、Rのクラスオブジェクト
  - RのReference Manualの読み解き方、データセットの連結
- 第3回（2019年07月22日）
  - 発現変動解析（多重比較問題とFDR）、各種プロット（M-A plot）
  - 発現変動解析（デザイン行列や3群間比較）
- 第4回（2019年07月29日）
  - 機能解析（発現変動遺伝子セット解析）、GSEA、MSigDB
  - GSVAの実行

Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、例題4
    - 例題5、例題6、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

クラスタリング

①TCCパッケージを用いて、サンプル間クラスタリングを行う。②例題7。③入力ファイルは20,689遺伝子×36サンプルのカウントデータファイル。ヒト(HS)、チンパンジー(PT)、アカゲザル(RM)の3生物種の肝臓(Liver)データ。各12サンプル。コピペ実行

- 解析 | 前処理 | scRNA-seq | について (last modified 2019/06/26) NEW
- 解析 | クラスタリング | RNA-seq | について (last modified 2019/04/04)
- 解析 | クラスタリング | RNA-seq | サンプル間 | hclust (last modified 2015/
- 解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013) ① modified 2018/08/06)
- 解析 | クラスタリング | RNA-seq | 遺伝子間(基礎) | MBCluster.Seq (... 2014) (last modified 2018/09/23)
- 解析
- 解析
- 解析

解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013)

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。 clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン掲載論文 (Tang et al., BMC Bioinformatics ...

② 7. サンプルデータ41のリアルデータ(sample_blekhman_36.txt)の場合:

Blekhman et al., Genome Res., 2010の 20,689 genes×36 samplesのカウントデータです。 ③

```
in_f <- "sample_blekhman_36.txt"          #入力ファイル名を指定してin_fに格納
out_f <- "hoge7.png"                      #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400)                  #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC)                              #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの読み込み
dim(data)                                 #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman",#クラスタリング実行結果をoutに格納
            hclust.method="average", unique.pattern=TRUE)#クラスタリング実行結果をoutに格納

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイルの各種パラメータを指定
par(mar=c(0, 4, 1, 0))                    #下、左、上、右の順で余白（行）を指定
plot(out, sub="", xlab="", cex.lab=1.2,#樹形図(デンドログラム)の表示
  cex=1.3,
dev.off()
```

TCC(Sun et al., *BMC Bioinformatics*, **14**: 219, 2013)

クラスタリング

①出力は、hoge7.pngという名前のPNGファイル。②サイズは、700×400ピクセル。これは論文の図としても使えるレベル。

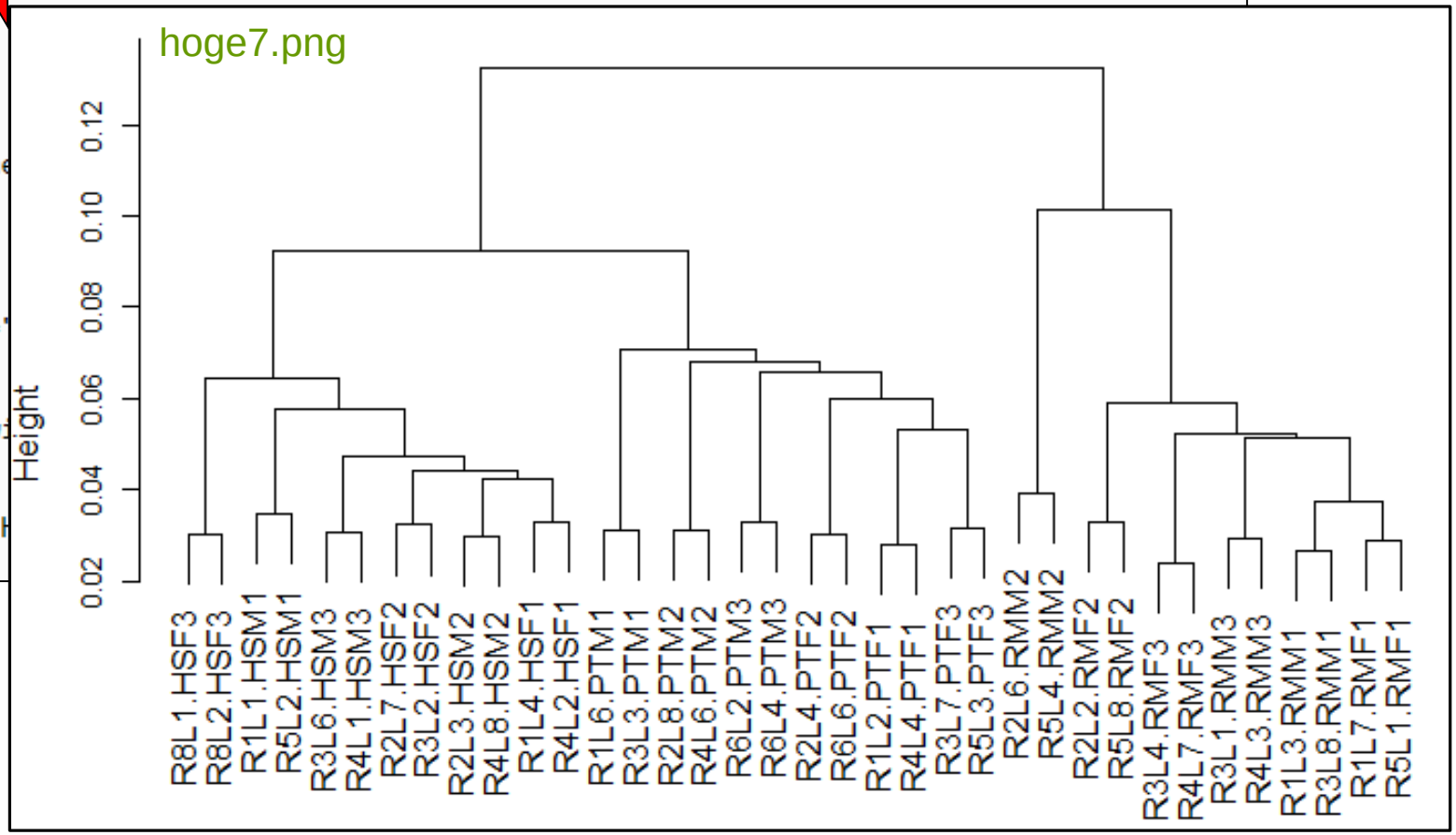
7. サンプルデータ41のリアルデータ(sample_blekhman_36.txt)の場合:

Blekhman et al., Genome Res., 2010の 20,689 genes×36 samplesのカウントデータです。

①

②

```
in_f <- "sample_blekhman_36.txt"          #入力ファイル名を指定してin_fに格納
out_f <- "hoge7.png"                      #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400)                  #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC)

#入力ファイルの読み込み
data <- read.table(in_f, he
dim(data)

#本番
out <- clusterSample(data,
            hclust.method="

#ファイルに保存
png(out_f, pointsize=13, wi
par(mar=c(0, 4, 1, 0))
plot(out, sub="", xlab="",
  cex=1.3, main="", ylab="H
dev.off()
```



クラスタリング

①出力は、hoge7.pngという名前のPNGファイル。②サイズは、700×400ピクセル。これは論文の図としても使えるレベル。③実際我々の論文中でも使っている。

7. サンプルデータ41のリアルデータ(sample_blekhman_36.txt)の場合:

Blekhman et al., Genome Res., 2010の 20,689 genes×36 samplesのカウントデータです。

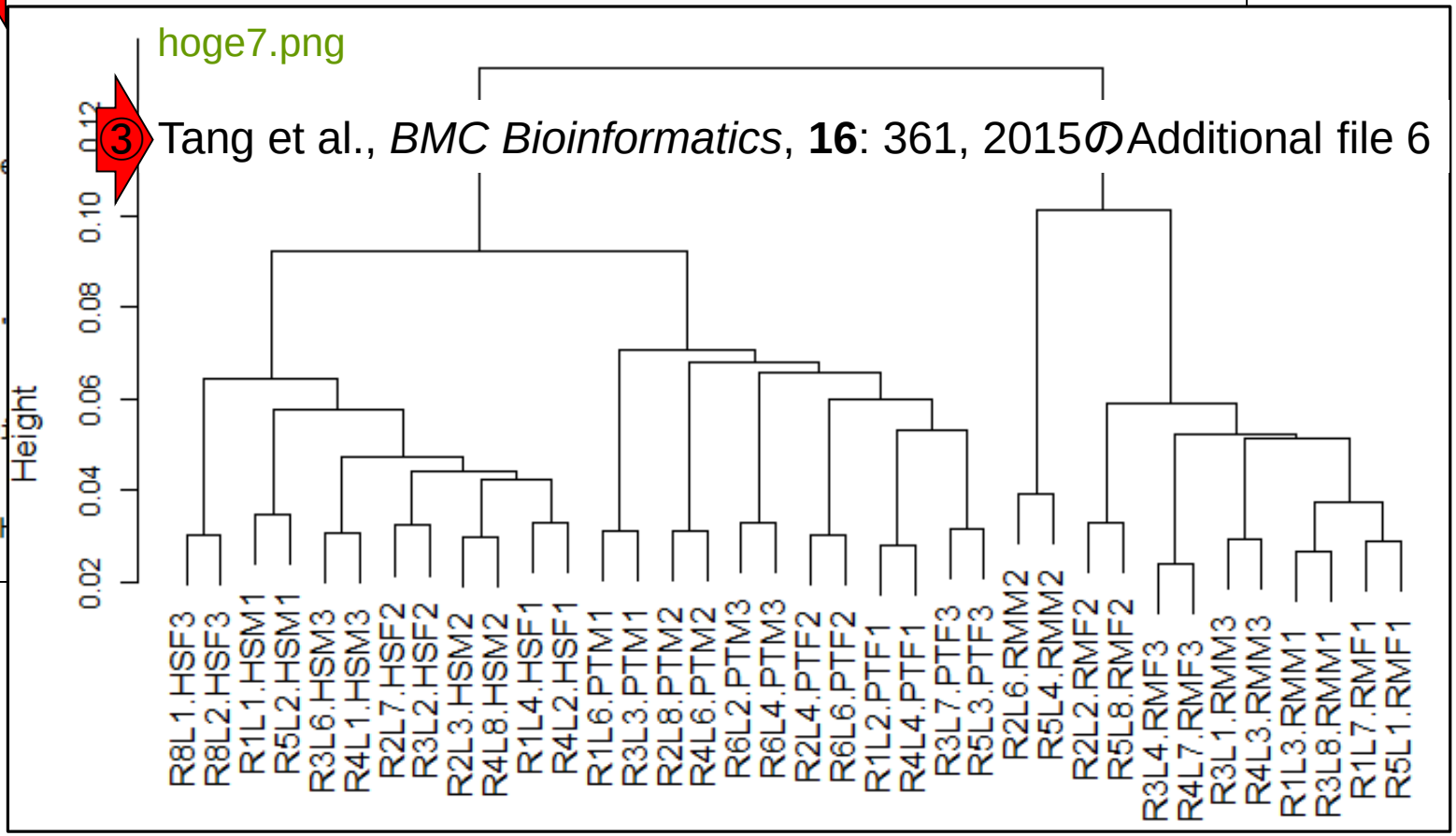
①

```
in_f <- "sample_blekhman_36.txt"          #入力ファイル名を指定してin_fに格納
out_f <- "hoge7.png"                      #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400)                  #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC)

#入力ファイルの読み込み
data <- read.table(in_f, he
dim(data)

#本番
out <- clusterSample(data,
            hclust.method="

#ファイルに保存
png(out_f, pointsize=13, wi
par(mar=c(0, 4, 1, 0))
plot(out, sub="", xlab="",
  cex=1.3, main="", ylab="H
dev.off()
```

②

hoge7.png

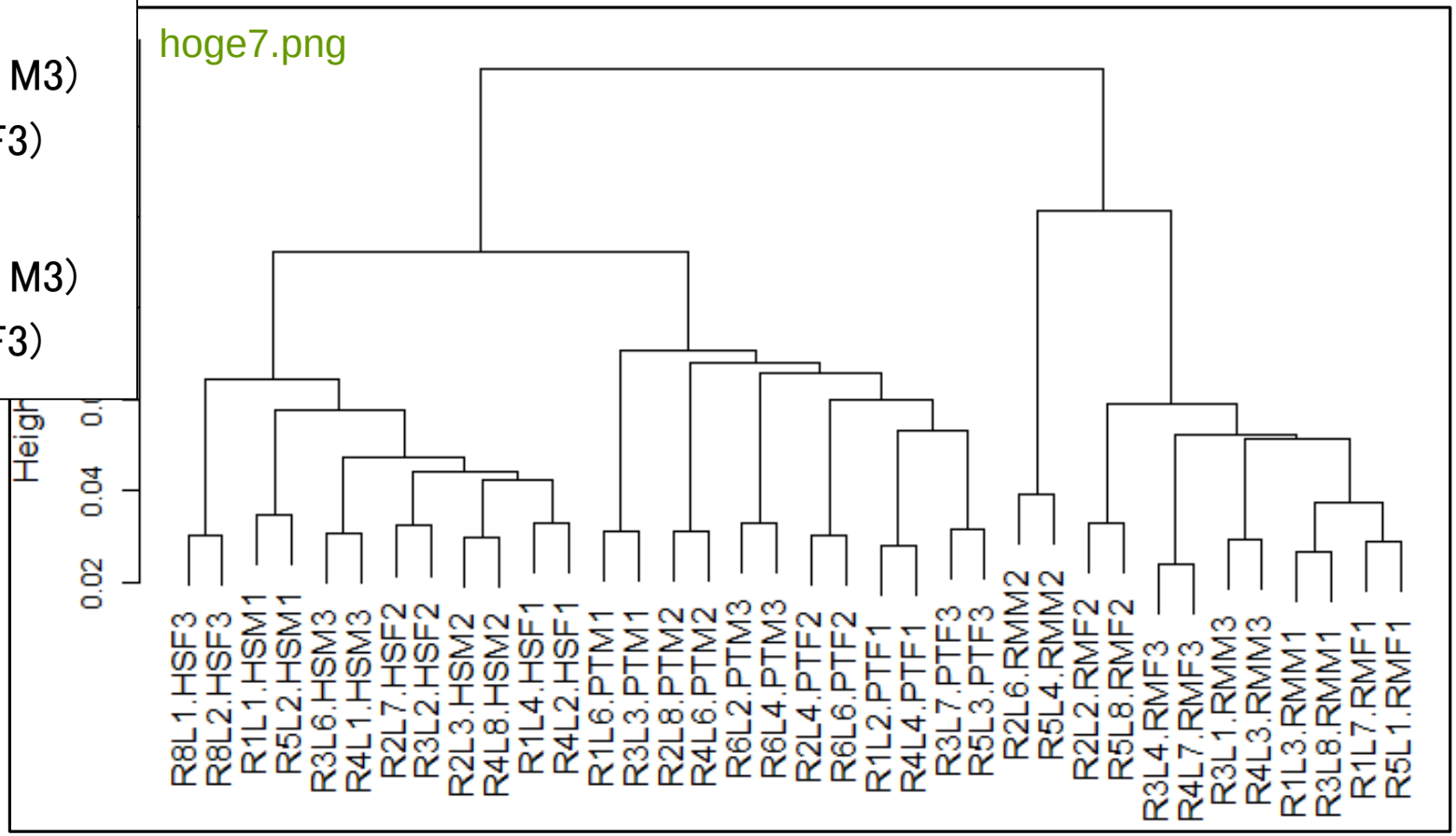
③ Tang et al., *BMC Bioinformatics*, **16**: 361, 2015のAdditional file 6



実験デザイン

入力ファイルは20,689遺伝子×36サンプルのカウントデータファイル。ヒト(HS)、チンパンジー(PT)、アカゲザル(RM)の3生物種の肝臓(Liver)データ。各12サンプル。

- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)

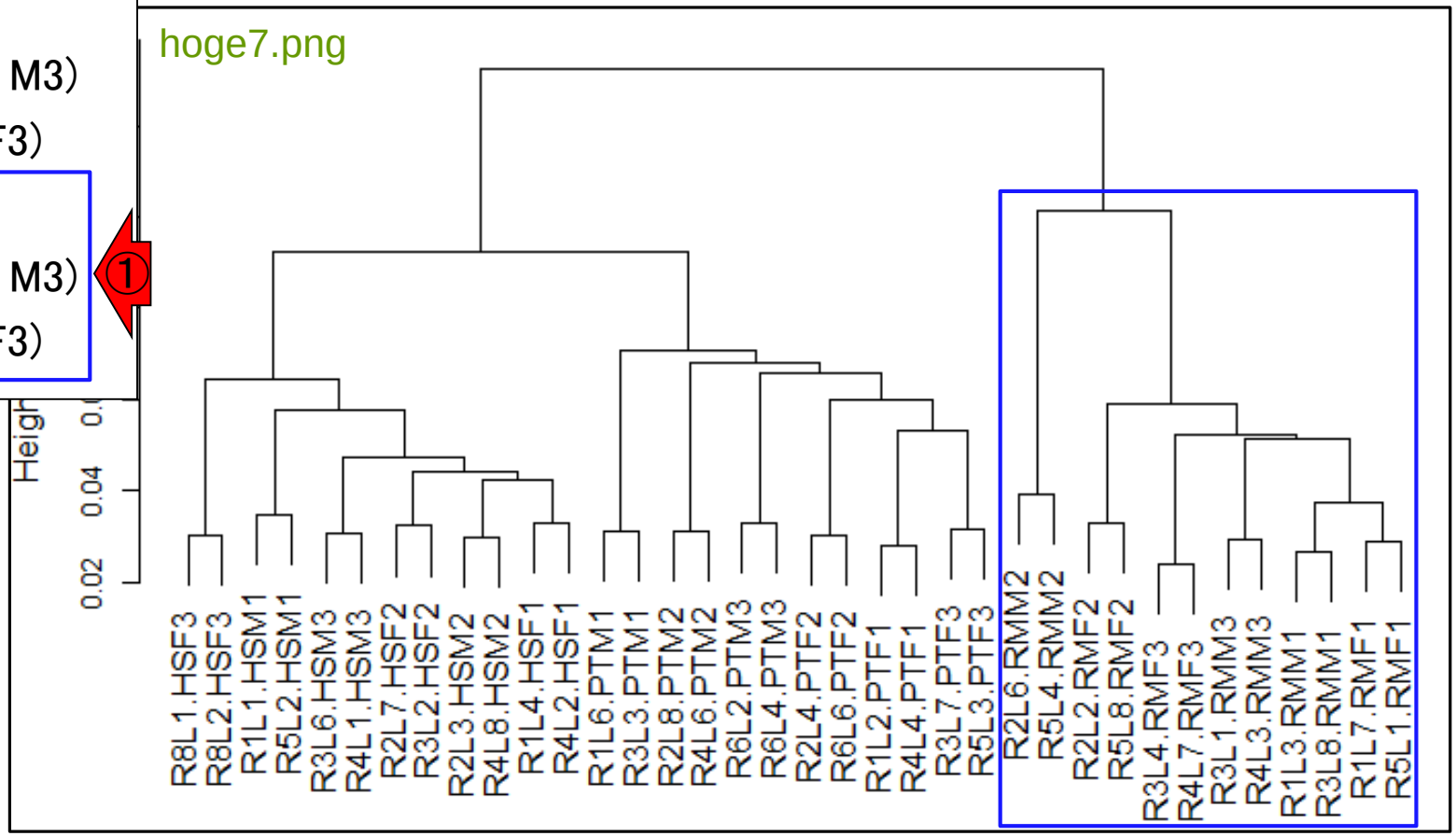


6個体×2反復

例えば、①アカゲザル(RM)の個体数は6。内訳はオス3匹とメス3匹。各個体につき2反復(technical replicatesは2)とっているので、6個体×2反復の計12サンプル。

- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)

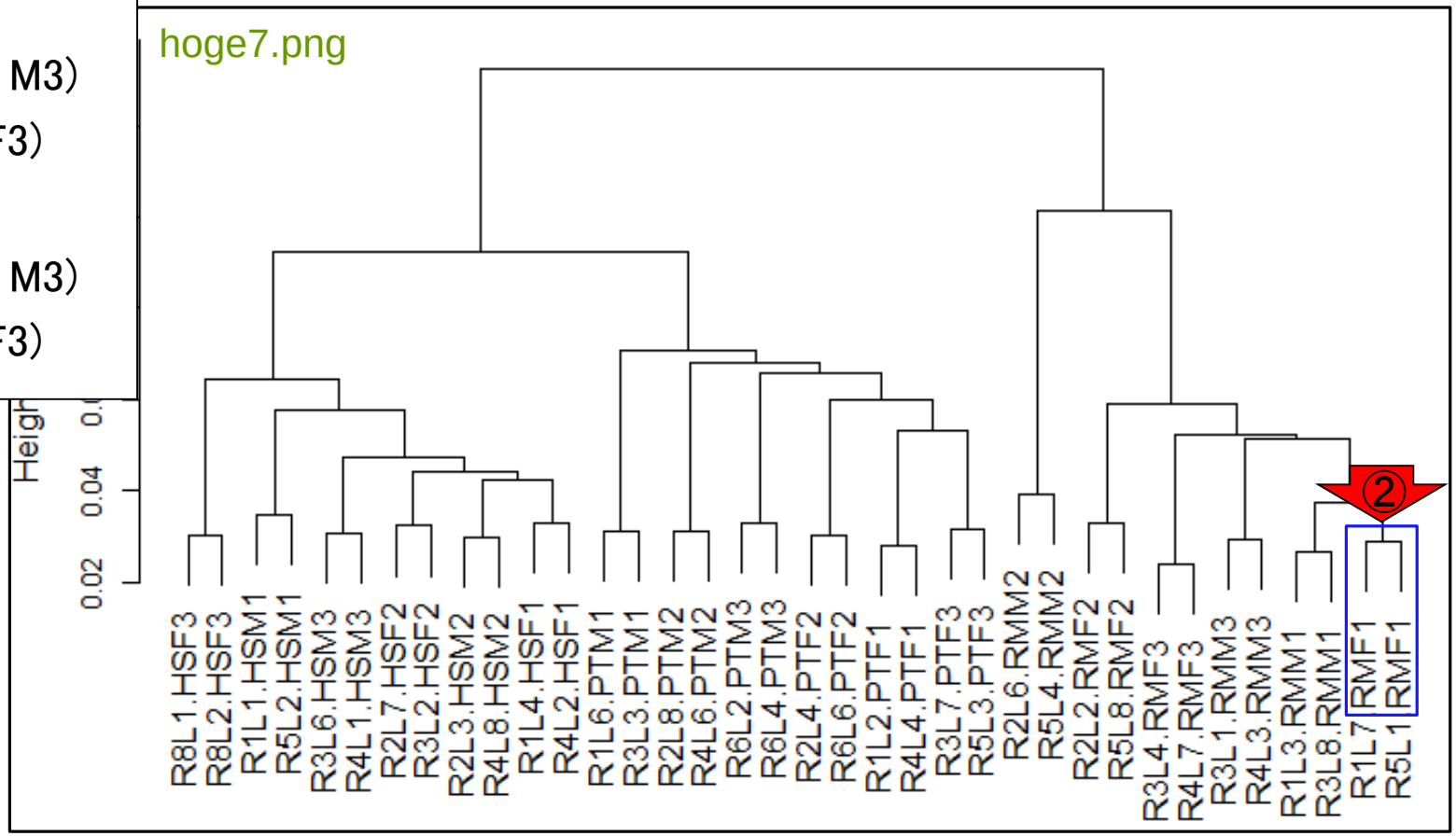
①



technical replicates

- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)

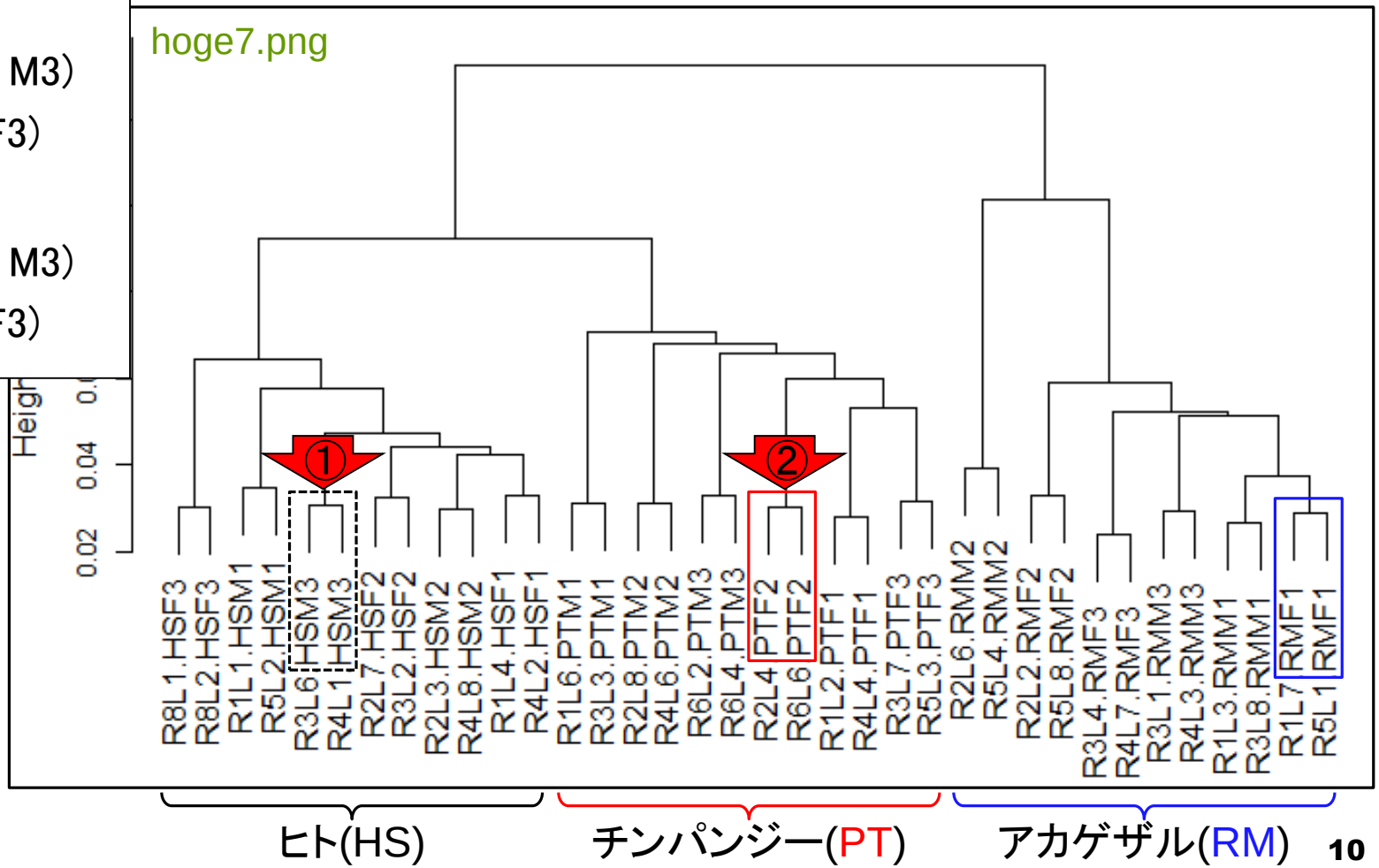
①アカゲザルのメス1個体（RMF1）の、②デンドログラム上の位置。同一個体の反復データ（technical replicates）で末端のクラスターを形成していることが分かる。これはtechnical replicates同士の類似度が非常に高いことを意味します。



HSM3やPTF2も

他の例として、①ヒトのオス（HSM3）と、②チンパンジーのメス（PTF2）も同様の結果です。全個体についてそのようになっており、妥当ですね。

- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)

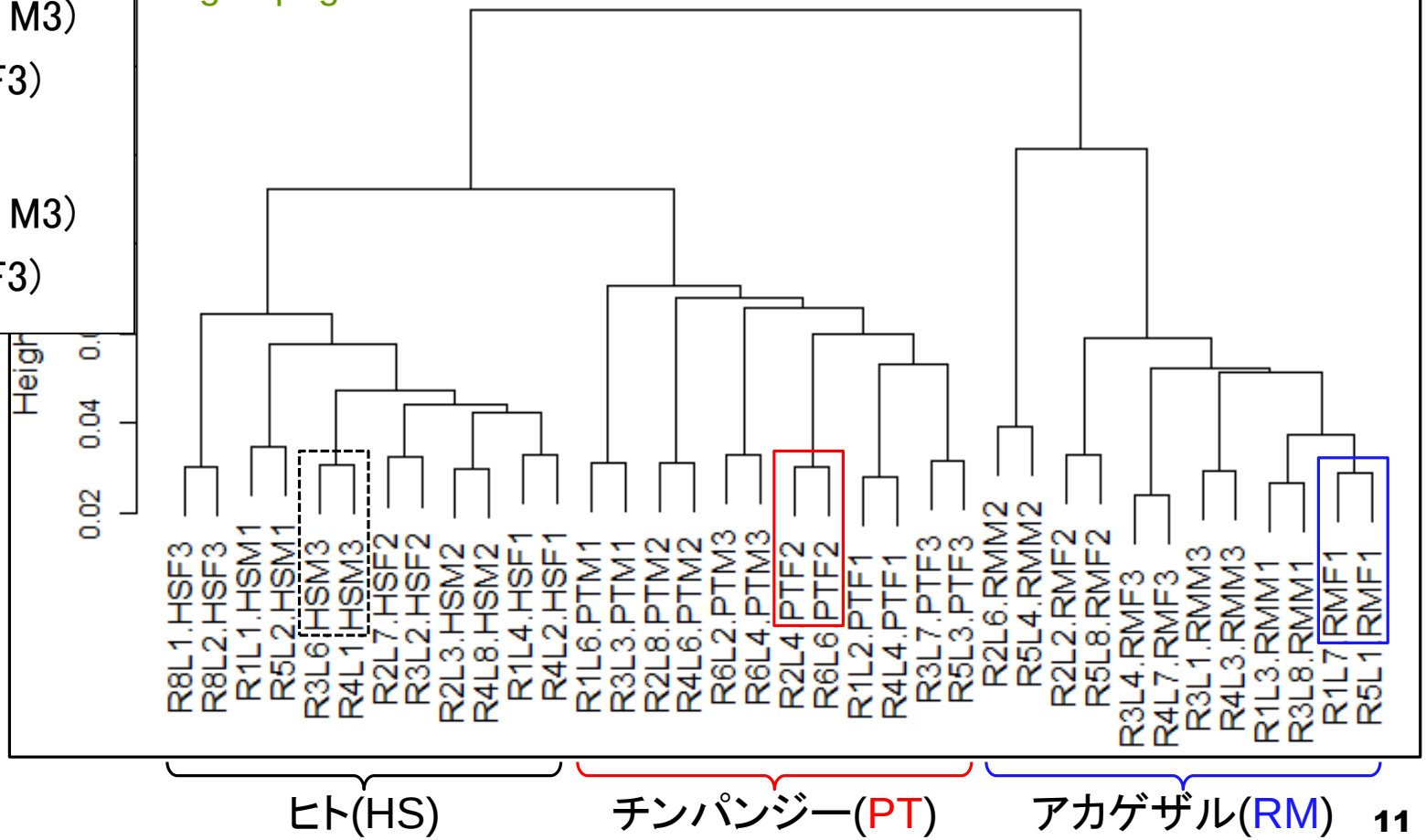


統計的手法

- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)

統計的手法で2群間比較(例えばMales vs. Females)をする目的は、同一群内の別個体(biological replicates)のばらつきの程度を見積もっておき(モデル構築)、比較する2群間で発現に変動がないという前提(帰無仮説)からどれだけ離れているのかをp値で評価することである。p値が低ければ低いほど「発現変動していない(帰無仮説に従う)」とは考えにくく、帰無仮説を棄却して「発現変動している(DEGである)」と判定することになる

hoge7.png



Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、例題4
    - 例題5、例題6、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

サブセット抽出と整形

①サンプルデータの、②例題42。統計的手法の多くは、biological replicatesのデータを前提としている。technical replicatesのデータをマージ(merge; collapseともいうらしい)したものを作成。③出力ファイルはsample_blekhman_18.txt。サンプル名部分は必要最小限の情報にしている。見るだけ。やらない。

- インストール | Rパッケージ | 個別(2018年11月以前) (last mo
- 基本的な利用法 (last modified 2019/03/12)
- サンプルデータ ① modified 2018/06/09)
- イントロ | 一般 | ランダムに行を抽出 (last modified 2014/07/
- イン
- イン
- イン

サンプルデータ NEW

② 42. Blekhman et al., Genome Res., 2010のリアルカウントデータです。
1つ前の例題41とは違って、technical replicatesの2列分のデータは足して1列分のデータとしています。20,689 genes×18 samplesのカウントデータ(sample_blekhman_18.txt)です。

```
#in_f <- "http://genome.cshlp.org/content/suppl/2009/12/16/gr.099226.109.DC1/suppTable1.xls"#入力ファイ
in_f <- "suppTable1.xls"                    #入力ファイル名を指定してin_fに格納
out_f <- "sample_blekhman_18.txt"           #出力ファイル名を指定してout_fに格納

#入力ファイルの読み込み
hoge <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの読み込み
dim(hoge)                                   #行数と列数を表示

#サブセットの取得
data <- cbind(                              #必要な列名の情報を取得したい列の順番で結合した結果をdataに格納
  hoge$R1L4.HSF1 + hoge$R4L2.HSF1, hoge$R2L7.HSF2 + hoge$R3L2.HSF2, hoge$R8L1.HSF3 + hoge$R8L2.HSF3,
  hoge$R1L1.HSM1 + hoge$R5L2.HSM1, hoge$R2L3.HSM2 + hoge$R4L8.HSM2, hoge$R3L6.HSM3 + hoge$R4L1.HSM3,
  hoge$R1L2.PTF1 + hoge$R4L4.PTF1, hoge$R2L4.PTF2 + hoge$R6L6.PTF2, hoge$R3L7.PTF3 + hoge$R5L3.PTF3,
  hoge$R1L6.PTM1 + hoge$R3L3.PTM1, hoge$R2L8.PTM2 + hoge$R4L6.PTM2, hoge$R6L2.PTM3 + hoge$R6L4.PTM3,
  hoge$R1L7.RMF1 + hoge$R5L1.RMF1, hoge$R2L2.RMF2 + hoge$R5L8.RMF2, hoge$R3L4.RMF3 + hoge$R4L7.RMF3,
  hoge$R1L3.RMM1 + hoge$R3L8.RMM1, hoge$R2L6.RMM2 + hoge$R5L4.RMM2, hoge$R3L1.RMM3 + hoge$R4L3.RMM3)
colnames(data) <- c(                        #列名を付加
  "HSF1", "HSF2", "HSF3", "HSM1", "HSM2", "HSM3",
  "PTF1", "PTF2", "PTF3", "PTM1", "PTM2", "PTM3",
  "RMF1", "RMF2", "RMF3", "RMM1", "RMM2", "RMM3")
rownames(data)<- rownames(hoge)             #行名を付加
dim(data)                                   #行数と列数を表示
```

出力ファイル

出力ファイルは、20,689遺伝子×18サンプルのbiological replicatesのみからなる、3生物種間比較用カウントデータ。ヒト(*Homo sapiens*; HS)、チンパンジー(*Pan troglodytes*; PT)、アカゲザル(*Rhesus macaque*; RM)。生物種ごとにメス3匹、オス3匹。雄雌を考慮しなければbiological replicates (生物学的な反復)は6

ヒト (*Homo sapiens*; HS)　チンパンジー (*Pan troglodytes*; PT)　アカゲザル (*Rhesus macaque*; RM)

メス(Female)　オス(Male)　メス　オス　メス　オス

20,689 genes

	HSF1	HSF2	HSF3	HSM1	HSM2	HSM3	PTF1	PTF2	PTF3	PTM1	PTM2	PTM3	RMF1	RMF2	RMF3	RMM1	RMM2	RMM3
ENSG00000000003	329	300	168	121	421	359	574	429	386	409	685	428	511	464	480	424	1348	705
ENSG00000000005	0	0	0	0	1	0	1	4	1	0	1	1	0	1	2	2	0	0
ENSG00000000419	81	61	56	39	78	62	100	66	65	59	58	93	67	72	57	49	82	90
ENSG00000000457	91	62	76	114	73	95	131	229	87	274	239	149	89	69	118	117	114	163
ENSG00000000460	6	17	12	15	7	17	8	8	5	12	7	10	4	4	10	7	3	4
ENSG00000000938	44	65	210	73	43	65	84	104	76	198	31	58	73	28	54	80	34	72
ENSG00000000971	4765	7225	3405	3600	6383	5546	5382	8331	4335	2568	5019	2653	13566	9964	18247	14236	5196	11834
ENSG00000001036	297	251	189	200	234	249	305	301	313	254	151	331	292	106	379	201	88	140
ENSG00000001084	630	737	306	336	984	459	417	328	885	298	569	218	1062	786	1110	873	664	1752
ENSG00000001167	36	30	36	29	33	28	63	80	25	69	74	41	62	34	108	97	35	61
ENSG00000001460	3	1	5	1	4	2	0	1	1	1	1	3	1	1	1	0	1	3
ENSG00000001461	49	37	34	28	62	32	75	69	40	90	69	60	210	92	176	247	81	117
ENSG00000001497	117	93	89	80	131	110	125	99	75	108	130	131	139	95	197	137	159	172

②例題8をコピペで実行

クラスタリング

- 解析 | 前処理 | scRNA-seq | について (last modified 2019/06/26) NEW
- 解析 | クラスタリング | RNA-seq | について (last modified 2019/04/04)
- 解析 | クラスタリング | RNA-seq | サンプル間 | hclust (last modified 2015/02/26)
- 解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013) (last modified 2018/08/06)
- 解析 | クラスタリング | RNA-seq | 遺伝子間(基礎) | MBCluster.Seq (Li, 2014) (last modified 2018/09/23)
- 解析
- 解析
- 解析

①

解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013)

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。 clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 (Tang et al., BMC Bioinformatics

②

8. サンプルデータ42のリアルデータ(sample_blekhman_18.txt)の場合:

Blekhman et al., Genome Res., 2010の 20,689 genes×18 samplesのカウントデータです。

```
in_f <- "sample_blekhman_18.txt"          #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.png"                      #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400)                  #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC)                              #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイ
dim(data)                                 #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman",#クラスタリング実行結果をoutに格納
            hclust.method="average", unique.pattern=TRUE)#クラスタリング実行結果をoutに格納

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイルの各種パラメー
par(mar=c(0, 4, 1, 0))                    #下、左、上、右の順で余白（行）を指定
plot(out, sub="", xlab="", cex.lab=1.2 #樹形図(デンドログラム)の表示
```

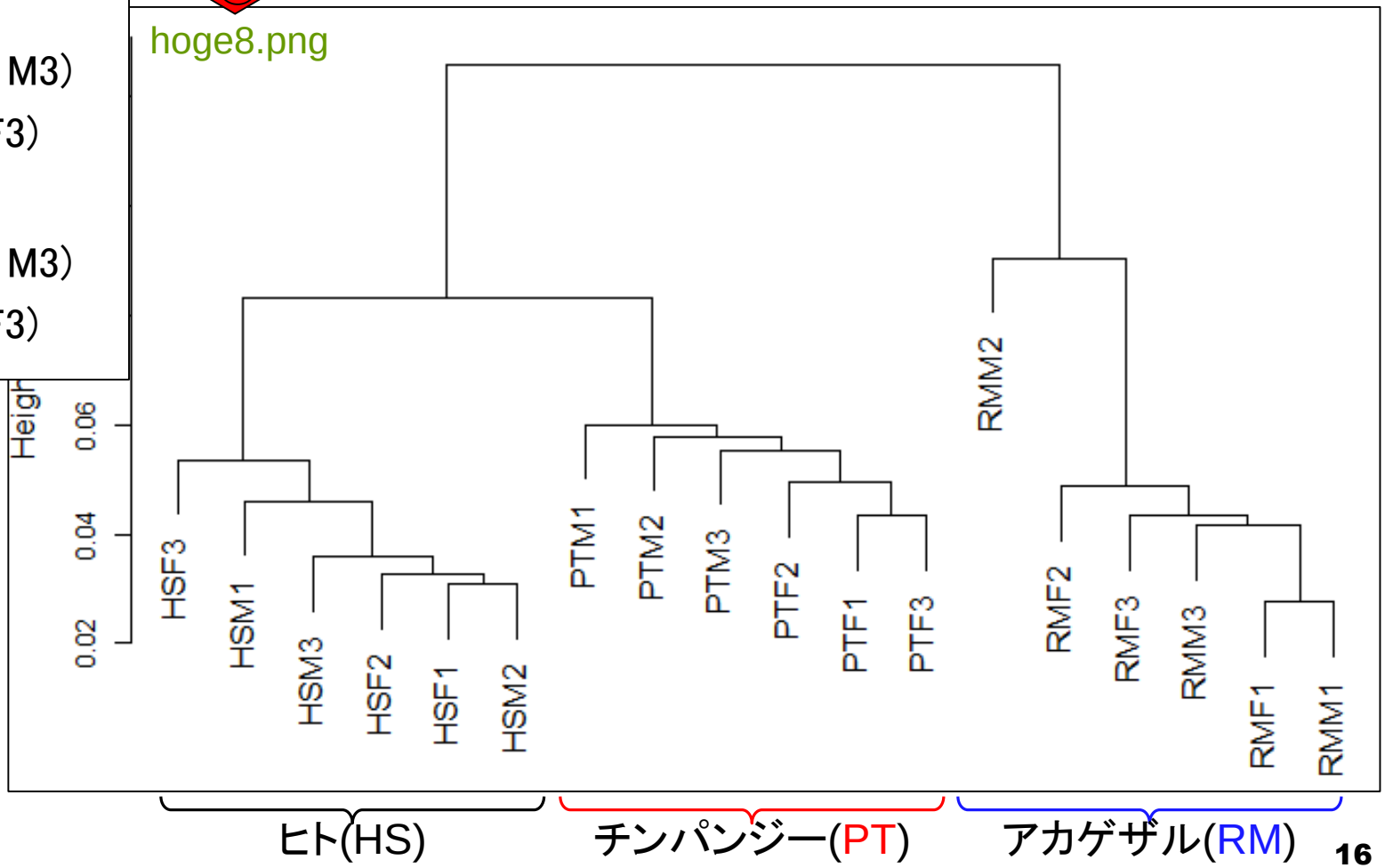
結果の解釈

- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)

①コピペ実行結果ファイル(hoge8.png)。これは肝臓の発現データでクラスタリングした結果。全体を生物種間比較という観点で眺める。

①

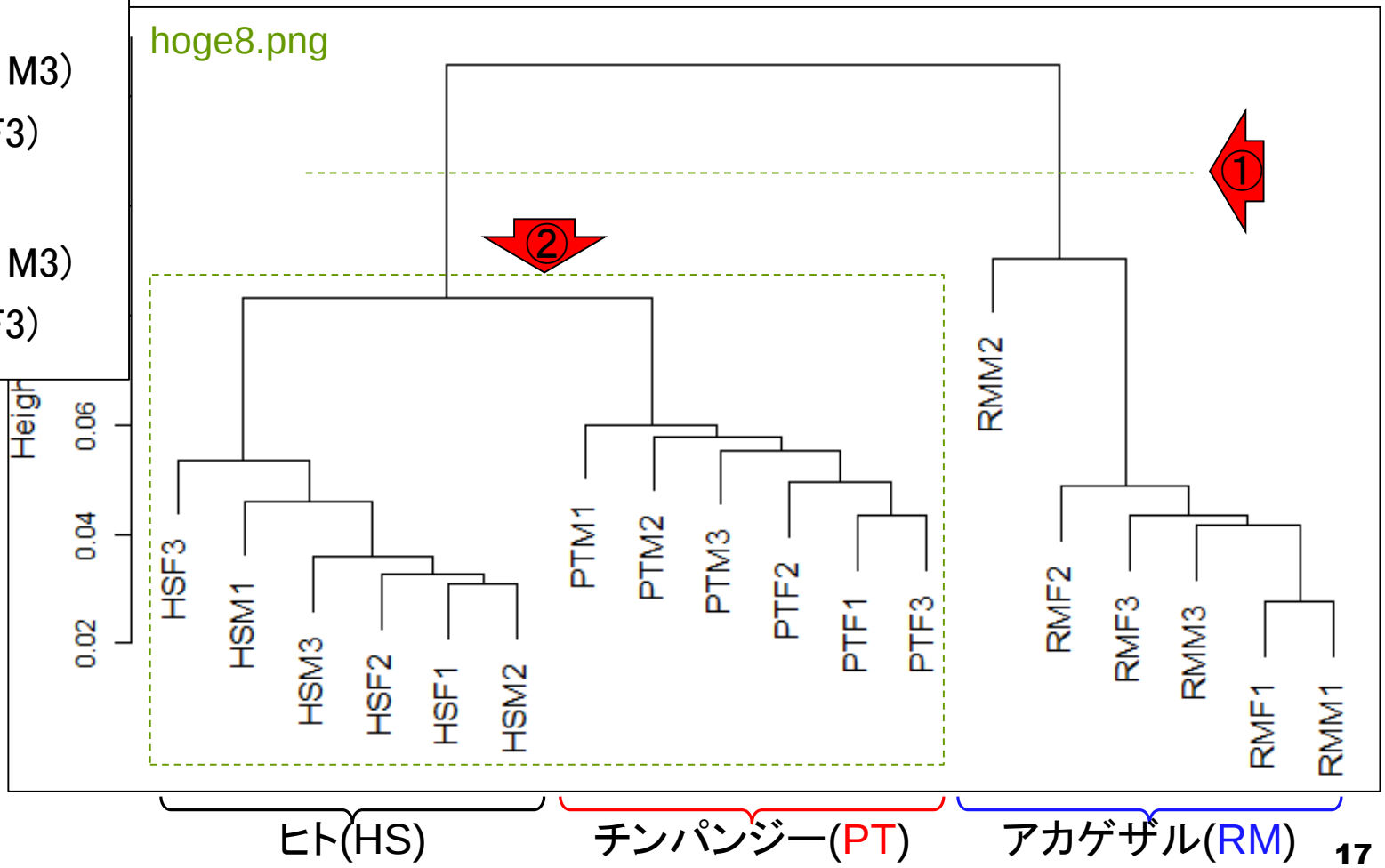
hoge8.png



HSとPTは似てる

①の部分で2つのグループに分けると…、②ヒト(HS)とチンパンジー(PT)は、アカゲザル(RM)と比べて似ている。

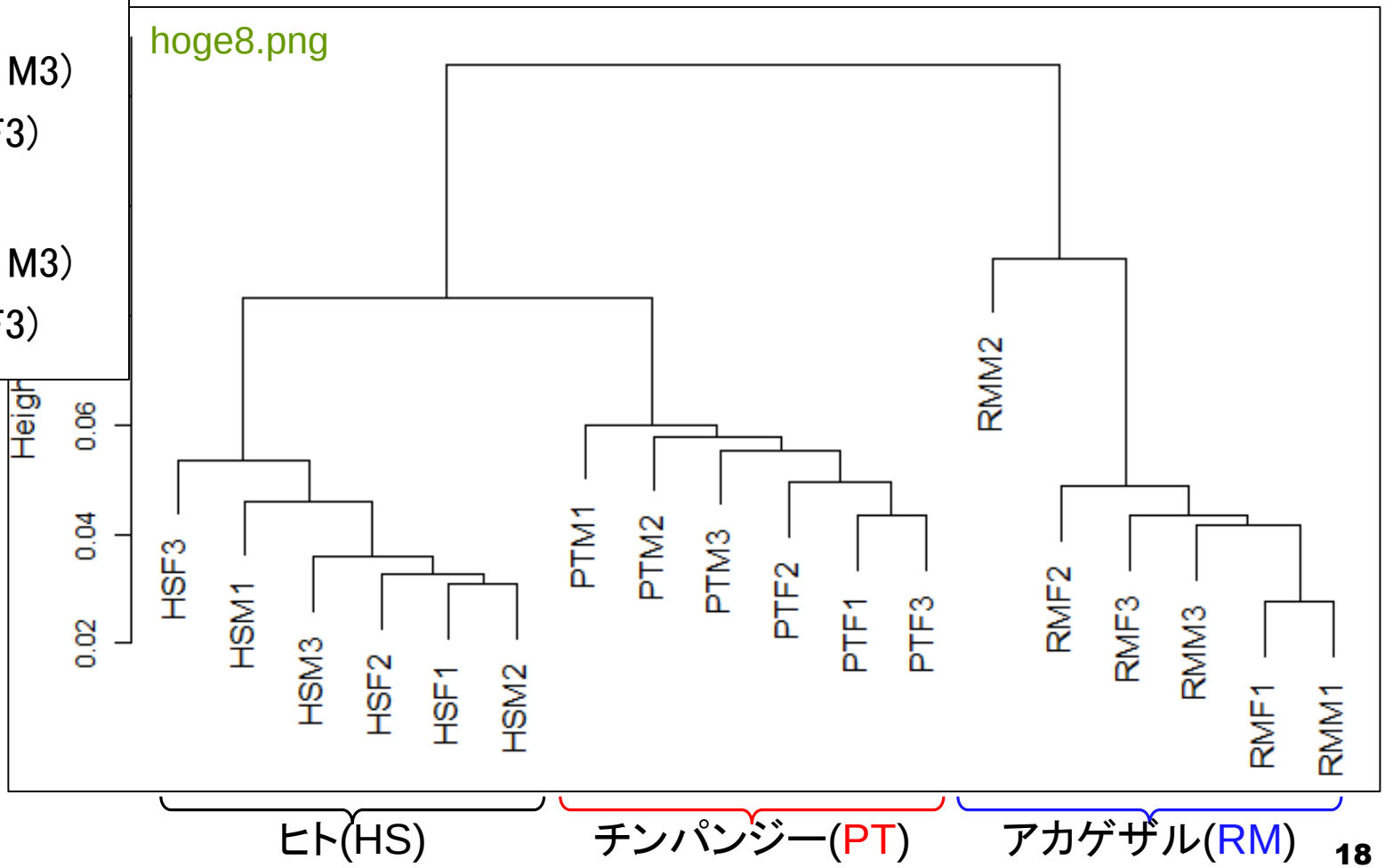
- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)



DEG検出結果の予想

2群間比較(発現変動遺伝子検出; DEG検出)を行うと、「HS vs. RMで得られるDEG数」のほうが「HS vs. PTで得られるDEG数」よりも多そう。

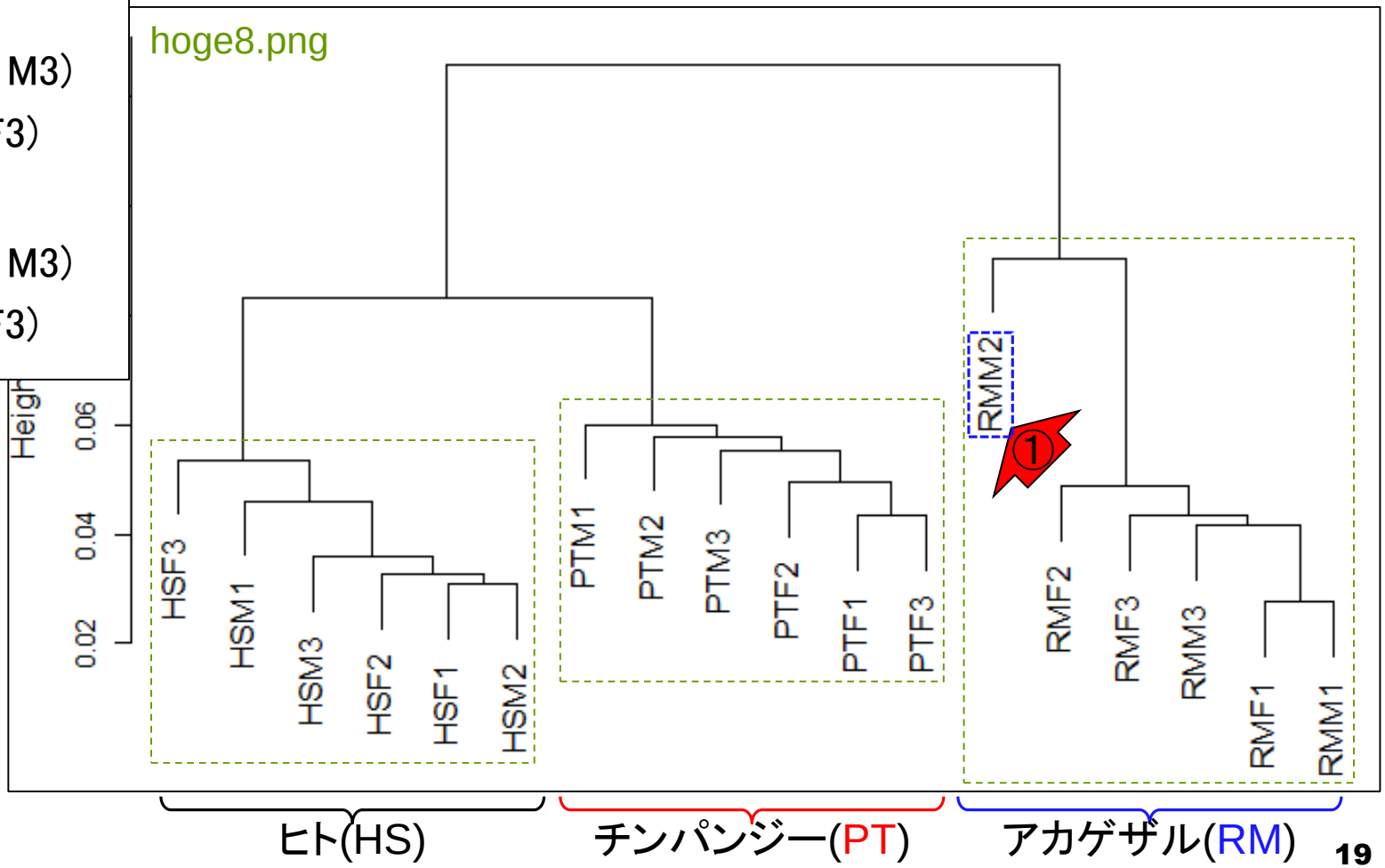
- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)



生物種内でクラスター形成

同一生物種でクラスターを形成している。①RMM2は「外れサンプル」っぽい

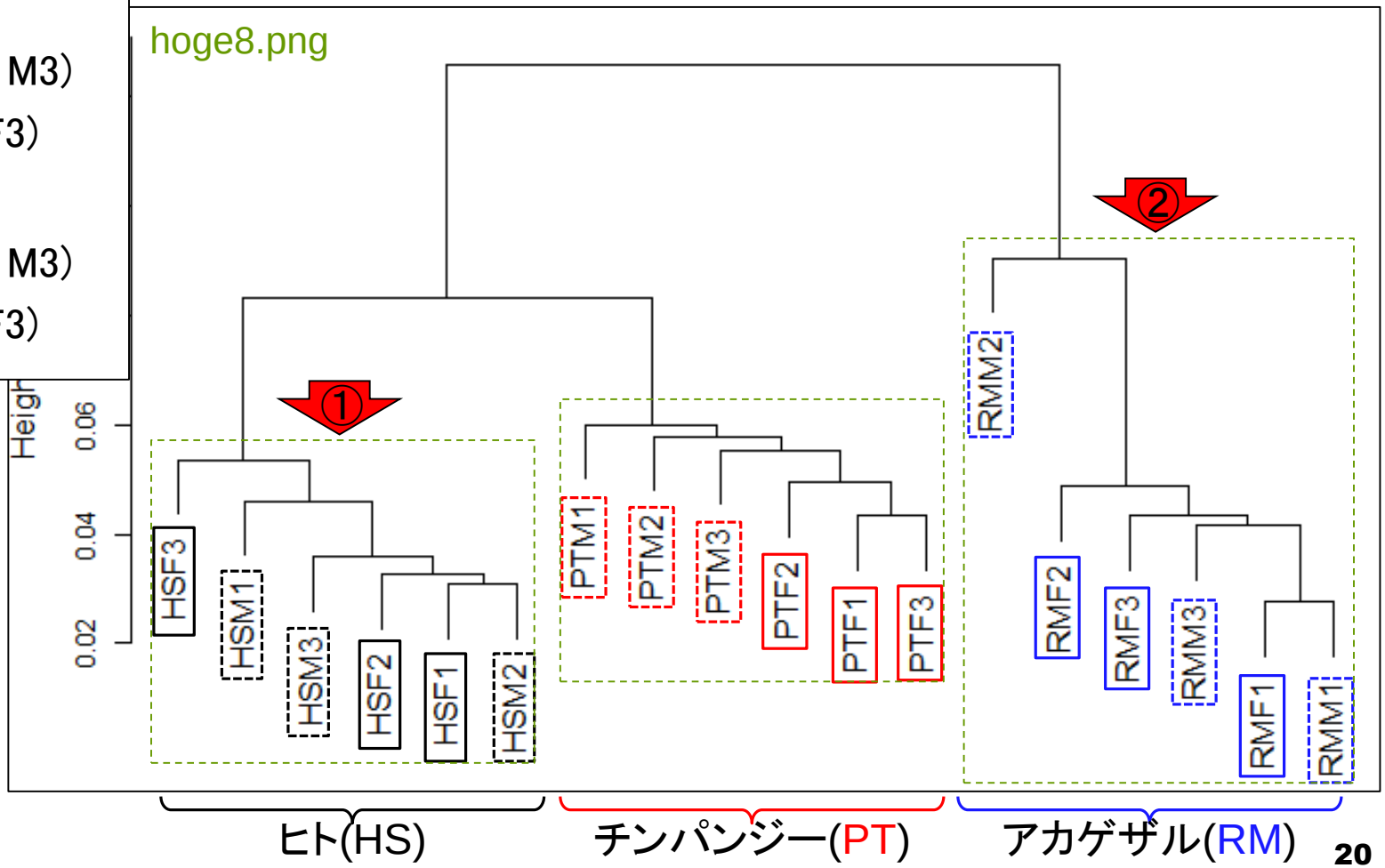
- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)



雌雄差はなさそう

- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)

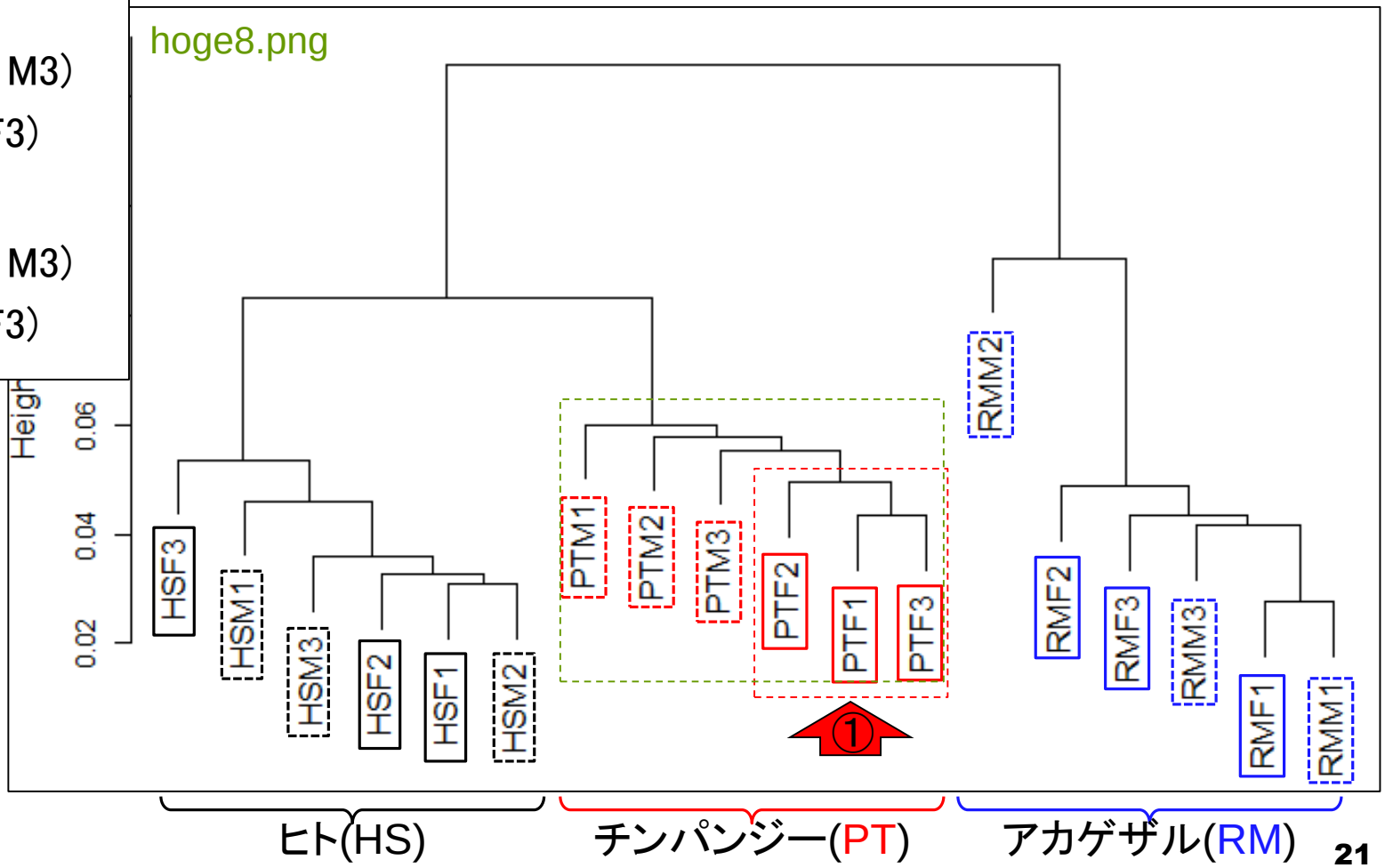
①ヒト(HS)と②アカゲザル(RM)は、メスとオスのサンプルが入り混じっている。これらの生物種内で、「メス群 vs. オス群」の2群間比較を行ってもDEGはほとんど検出されないだろう



結果の解釈

チンパンジー(PT)に限っていえば、①メス3匹がクラスターを形成しているので、「メス群 vs. オス群」の2群間比較結果として、多少なりともDEGが検出されるだろう

- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)



参考

樹形図とDEG数の関係

クラスタリング結果（樹形図；dendrogram）とDEG数の関係性に関する主観的な評価は、①のあたりに書いてます。このようにクラスタリング結果の解釈は往々にして主観的。

BMC Bioinformatics. 2015 Nov 4;16:361. doi: 10.1186/s12859-015-0794-7.

Evaluation of methods for differential expression analysis on multi-group RNA-seq count data.

Tang M[1], Sun J[2], Shimizu K[3], Kadota K[4].

Author information

Abstract

BACKGROUND: RNA-seq is a powerful tool for measuring transcriptomes, especially for identifying differentially expressed genes or transcripts (DEGs) between sample groups. A number of methods have been developed for this task, and several evaluation studies have also been reported. However, those evaluations so far have been restricted to two-group comparisons. Accumulations of comparative studies for multi-group data are also desired.

METHODS: We compare 12 pipelines available in nine R packages for detecting differential expressions (DE) from multi-group RNA-seq count data, focusing on three-group data with or without replicates. We evaluate those pipelines on the basis of both simulation data and real count data.

RESULTS: As a result, the pipelines in the TCC package performed comparably to or better than other pipelines under various simulation scenarios. TCC implements a multi-step normalization strategy (called DEGES) that internally uses functions provided by other representative packages (edgeR, DESeq2, and so on). We found considerably different numbers of identified DEGs (18.5 ~ 45.7% of all genes) among the pipelines for the same real dataset but similar distributions of the classified expression patterns. We also found that DE results can roughly be estimated by the hierarchical dendrogram of sample clustering for the raw count data.

CONCLUSION: We confirmed the DEGES-based pipelines implemented in TCC performed well in a three-group comparison as well as a two-group comparison. We recommend using the DEGES-based pipeline that internally uses edgeR (here called the EEE-E pipeline) for count data with replicates (especially for small sample sizes). For data without replicates, the DEGES-based pipeline with DESeq2 (called SSS-S) can be recommended.

PMID: 26538400 PMCID: PMC4634584 DOI: 10.1186/s12859-015-0794-7

[Indexed for MEDLINE] Free PMC Article

参考

樹形図とDEG数の

クラスタリング結果（樹形図）を眺めて、興味あるグループ間の関係性（特にDEG検出結果）を客観的に評価する指標として、シルエットスコア（Silhouette score）が有用だということを示した論文。

Biol Proced Online. 2018 Mar 1;20:5. doi: 10.1186/s12575-018-0067-8. eCollection 2018.

Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results.

Zhao S[1], Sun J[1], Shimizu K[1], Kadota K[1].

Author information

Abstract

BACKGROUND: Hierarchical Sample clustering (HSC) is widely performed to examine associations within expression data obtained from microarrays and RNA sequencing (RNA-seq). Researchers have investigated the HSC results with several possible criteria for grouping (e.g., sex, age, and disease types). However, the evaluation of arbitrary defined groups still counts in subjective visual inspection.

RESULTS: To objectively evaluate the degree of separation between groups of interest in the HSC dendrogram, we propose to use *Silhouette* scores. Silhouettes was originally developed as a graphical aid for the validation of data clusters. It provides a measure of how well a sample is classified when it was assigned to a cluster by according to both the tightness of the clusters and the separation between them. It ranges from 1.0 to - 1.0, and a larger value for the average silhouette (AS) over all samples to be analyzed indicates a higher degree of *cluster* separation. The basic idea to use an AS is to replace the term *cluster* by *group* when calculating the scores. We investigated the validity of this score using simulated and real data designed for differential expression (DE) analysis. We found that larger (or smaller) AS values agreed well with both higher (or lower) degrees of separation between different groups and higher percentages of differentially expressed genes (P_{DEG}). We also found that the AS values were generally independent on the number of replicates (N_{rep}). Although the P_{DEG} values depended on N_{rep}, we confirmed that both AS and P_{DEG} values were close to zero when samples in the data showed an intermingled nature between the groups in the HSC dendrogram.

CONCLUSION: Silhouettes is useful for exploring data with predefined group labels. It would help provide both an objective evaluation of HSC dendrograms and insights into the DE results with regard to the compared groups.

KEYWORDS: Bioinformatics; Differential expression analysis; Hierarchical sample clustering; Silhouettes

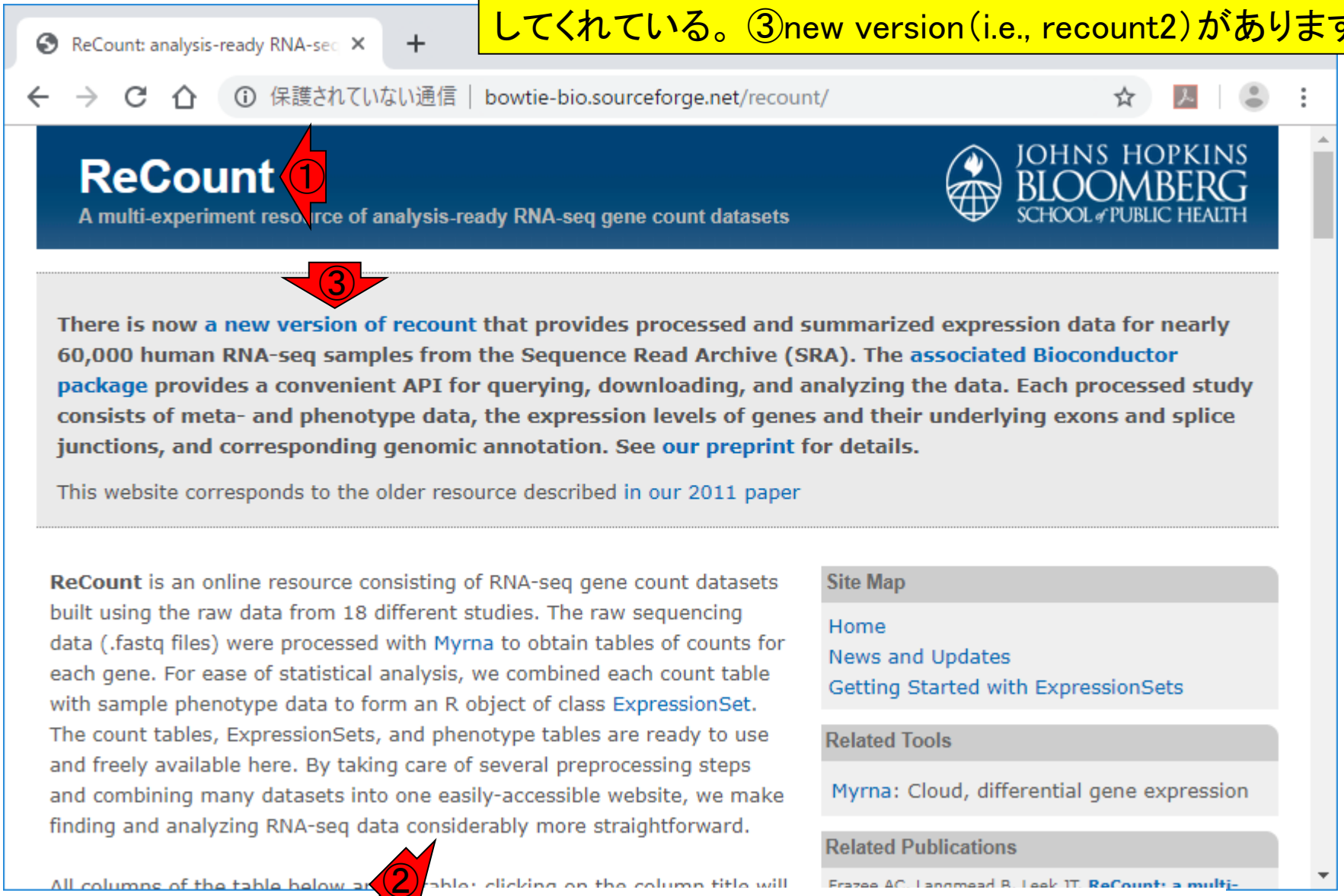
PMID: 29507534 PMCID: PMC5831220 DOI: 10.1186/s12575-018-0067-8

Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、例題4
    - 例題5、例題6、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

ReCount

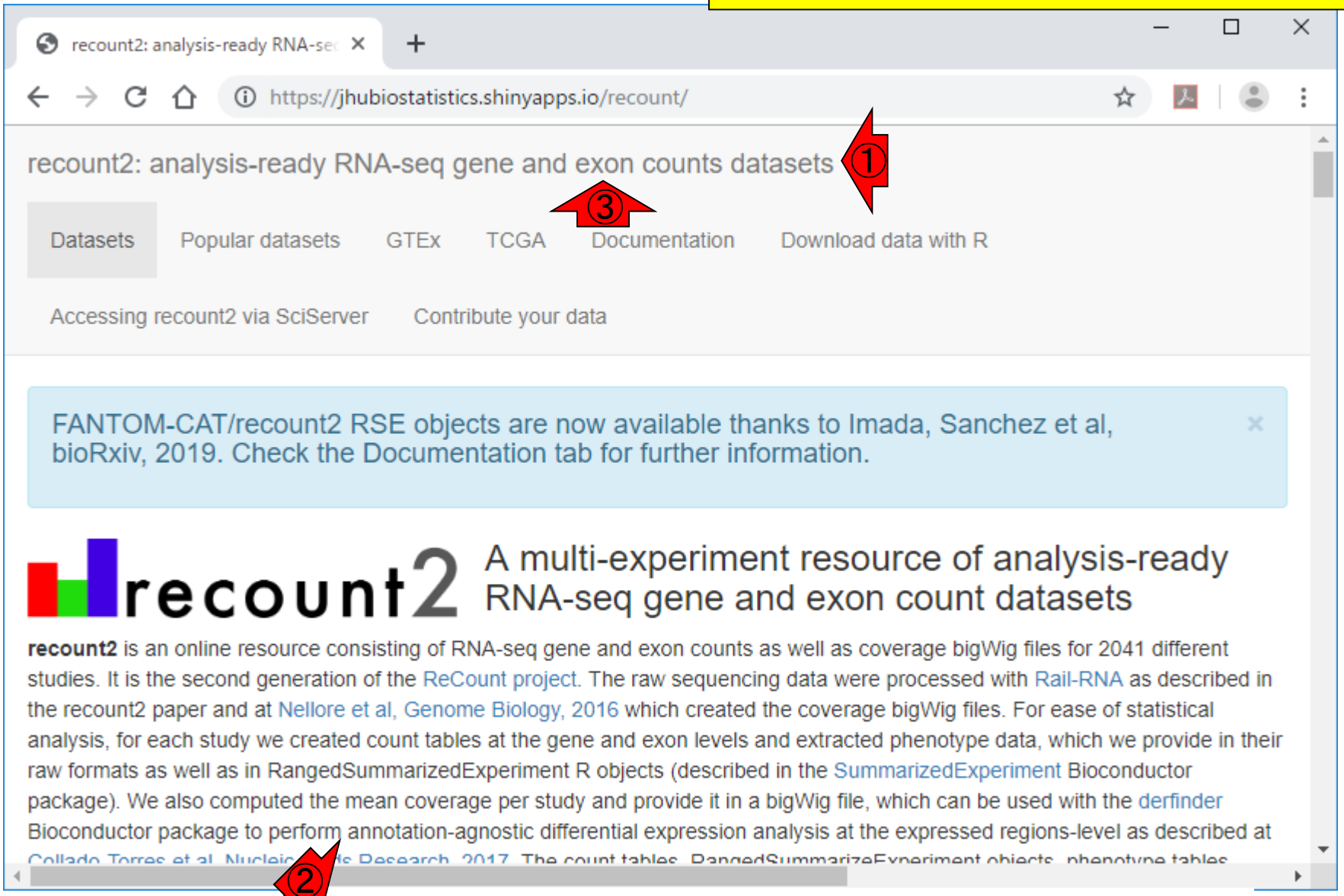
①ReCountは、カウントデータを提供しているサイト。②原著論文。自分でマッピングから行わずに済むので便利。technical replicatesのデータセットについては、biological replicatesにマージしたものも提供してくれている。③new version（i.e., recount2）があります。



Frazee et al., *BMC Bioinformatics*, **12**: 449, 2011

recount2

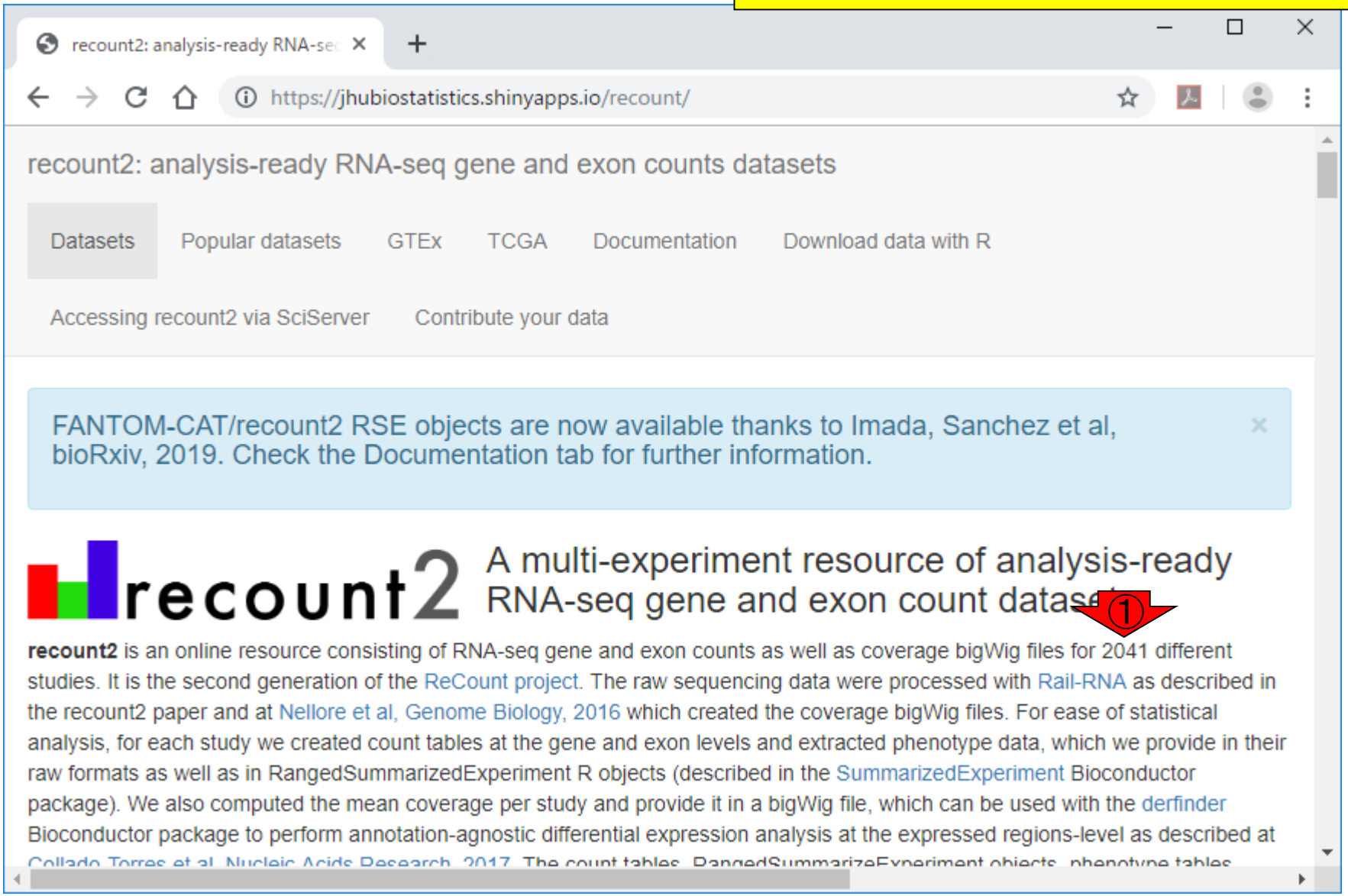
①recount2のウェブサイト。②原著論文。前のバージョン（ReCount）ではgeneレベルのみでしたが、recount2では③exonレベルのカウントデータも利用可能です。



Collado-Torres et al., *Nat Biotechnol.*, **35**: 319-321, 2017

recount2

前のバージョン（ReCount）では18個しかありませんでしたが、recount2では①2,041個もあるようです。この数値は大まかにカウントデータセット数に相当します。

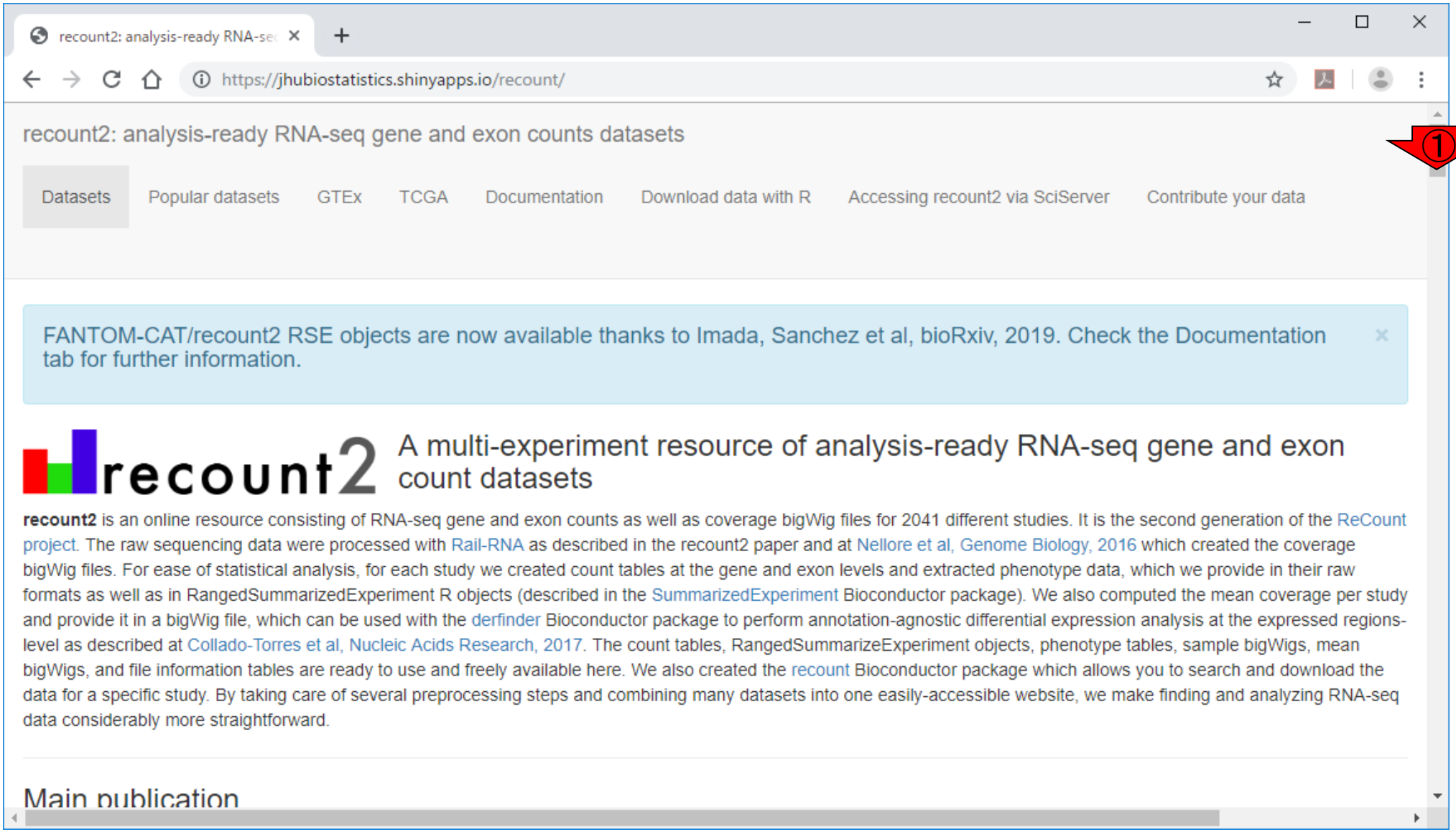


Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、例題4
    - 例題5、例題6、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

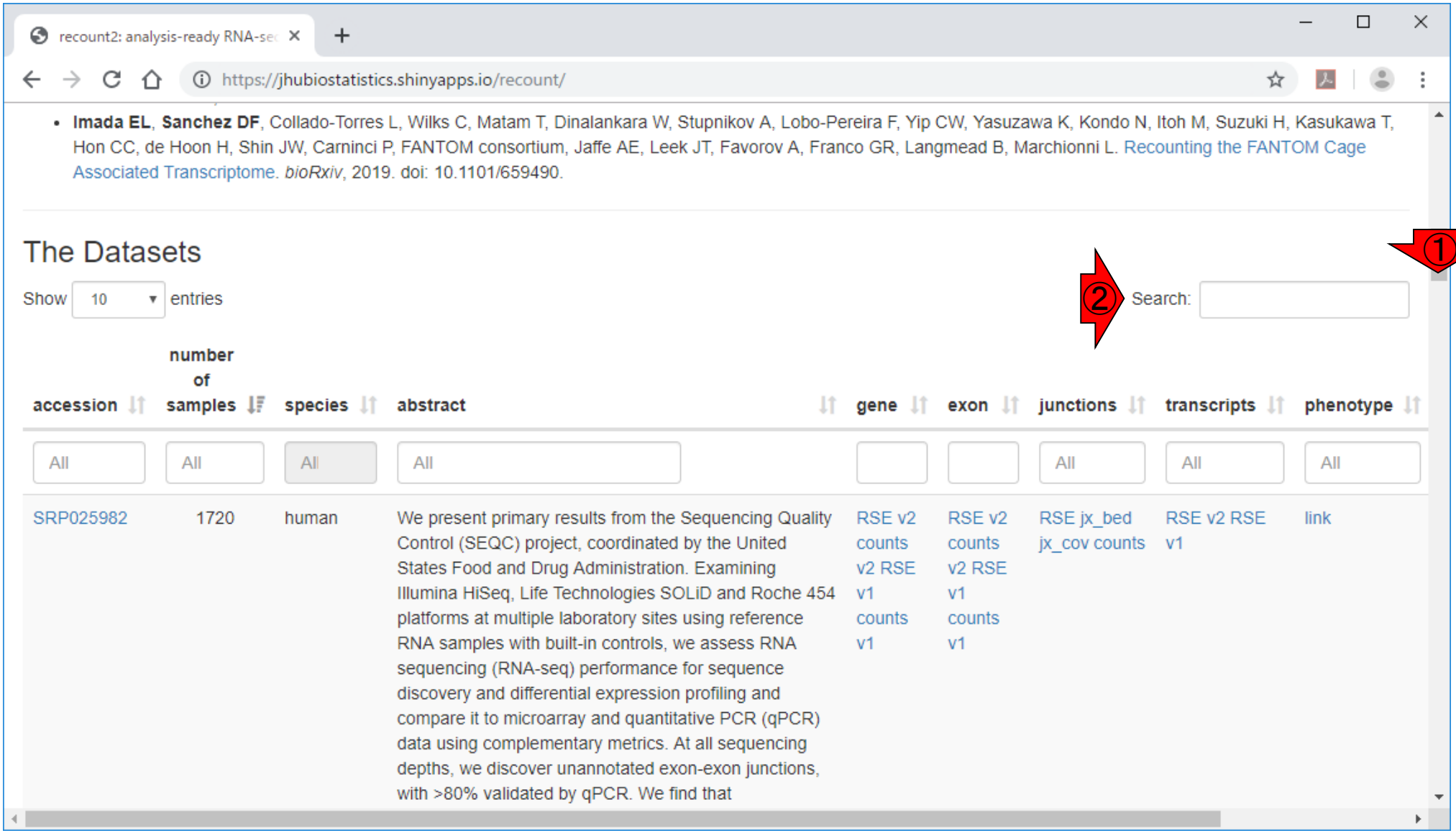
①ページ下部に移動

Search



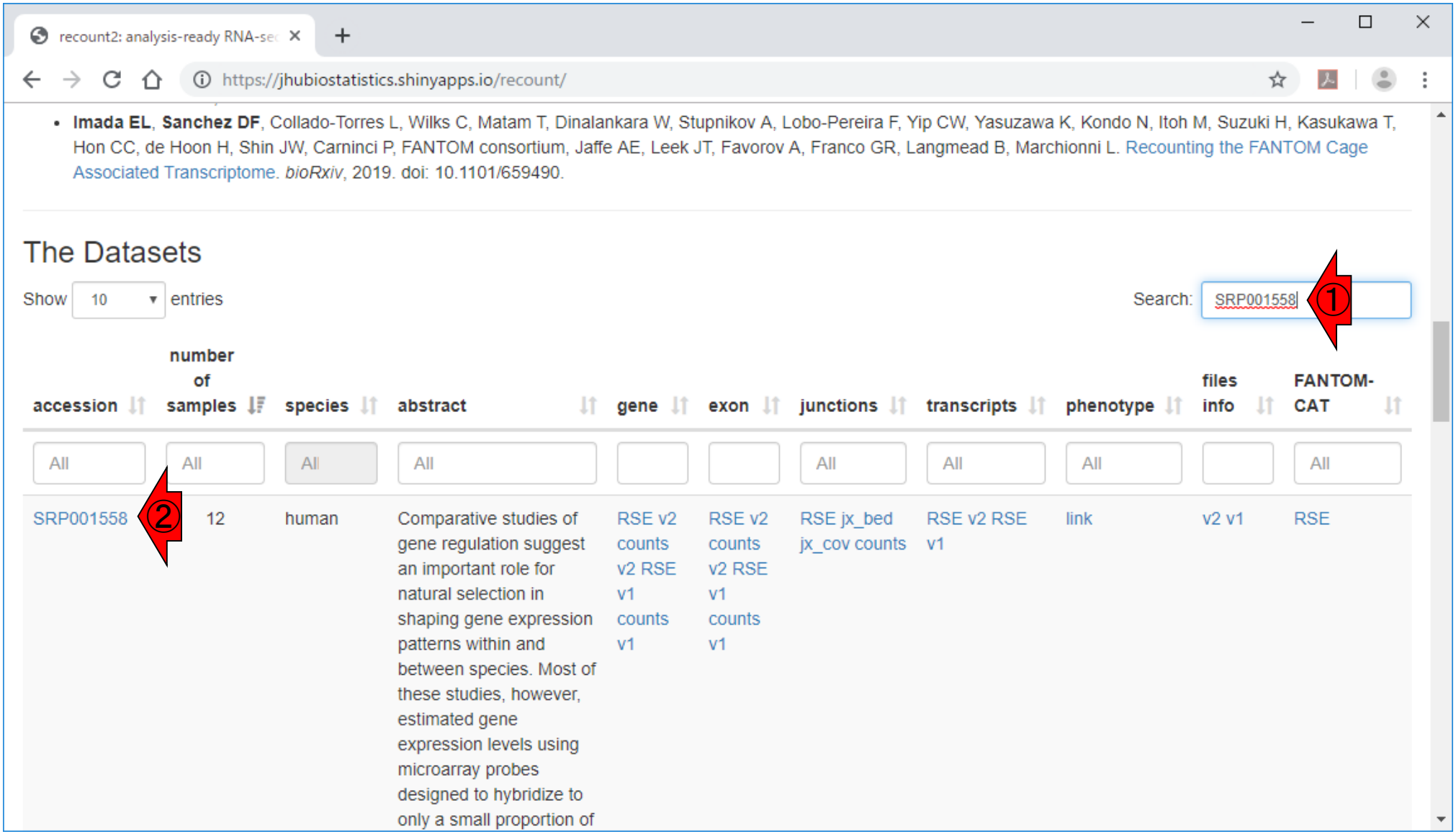
Search

①このあたりまで移動すると、②検索窓があります。



SRP001558

①SRP001558と打ち込むと、このような画面になります。②も切り替わっていることがわかります。



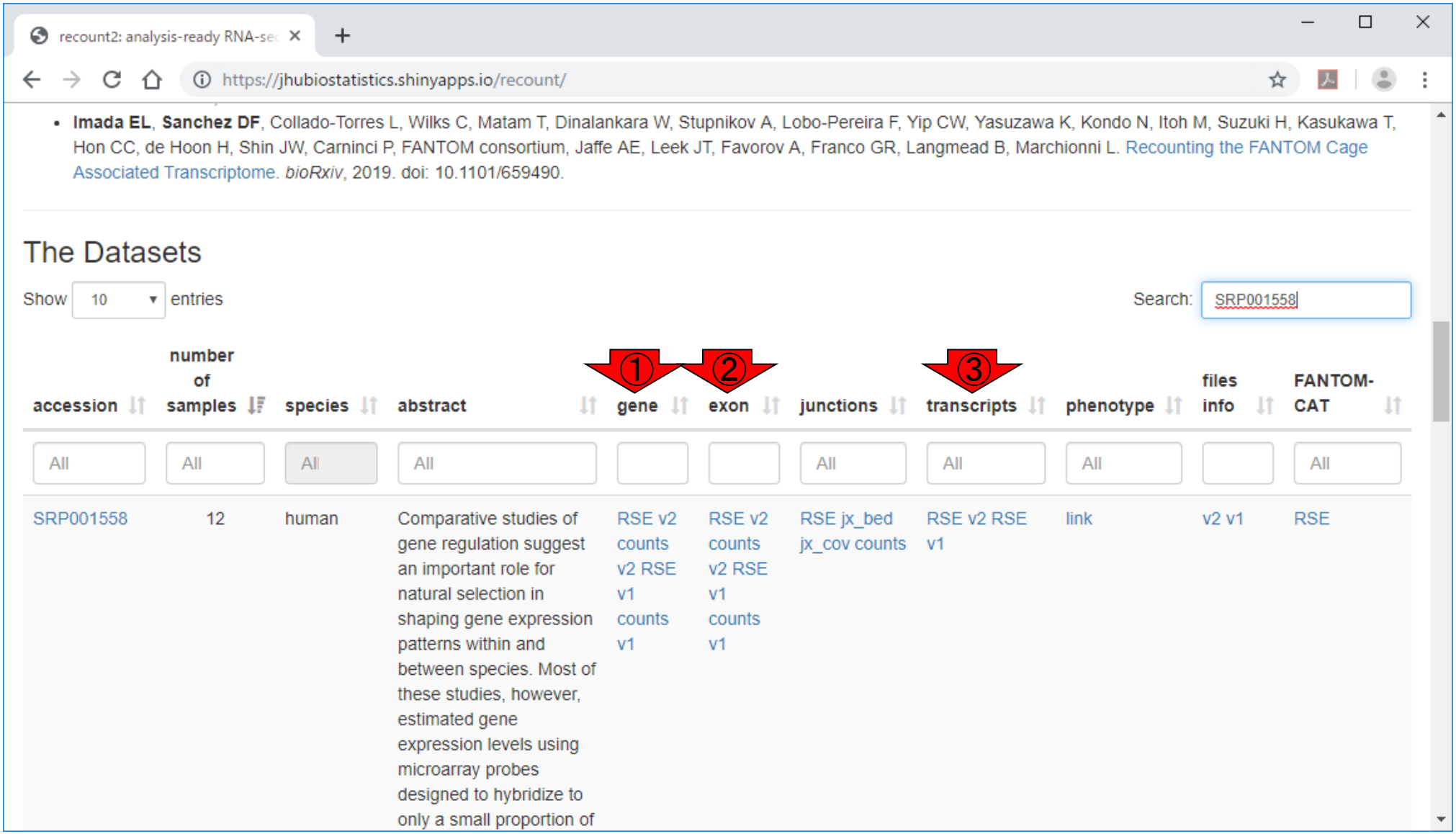
SRP001558

recount2から提供されている①のカウント情報は、②ヒト(HS)データ限定なのだろうと読み解く。③サンプル数は12と書かれている。メス(Female)3匹、オス(Male)3匹の計6個体で、各個体につき2反復(technical replicatesは2)とっているので、6個体×2反復の計12サンプルとなるのは妥当。

Blekhman et al., *Genome Res.*, **20**: 180-9, 2010

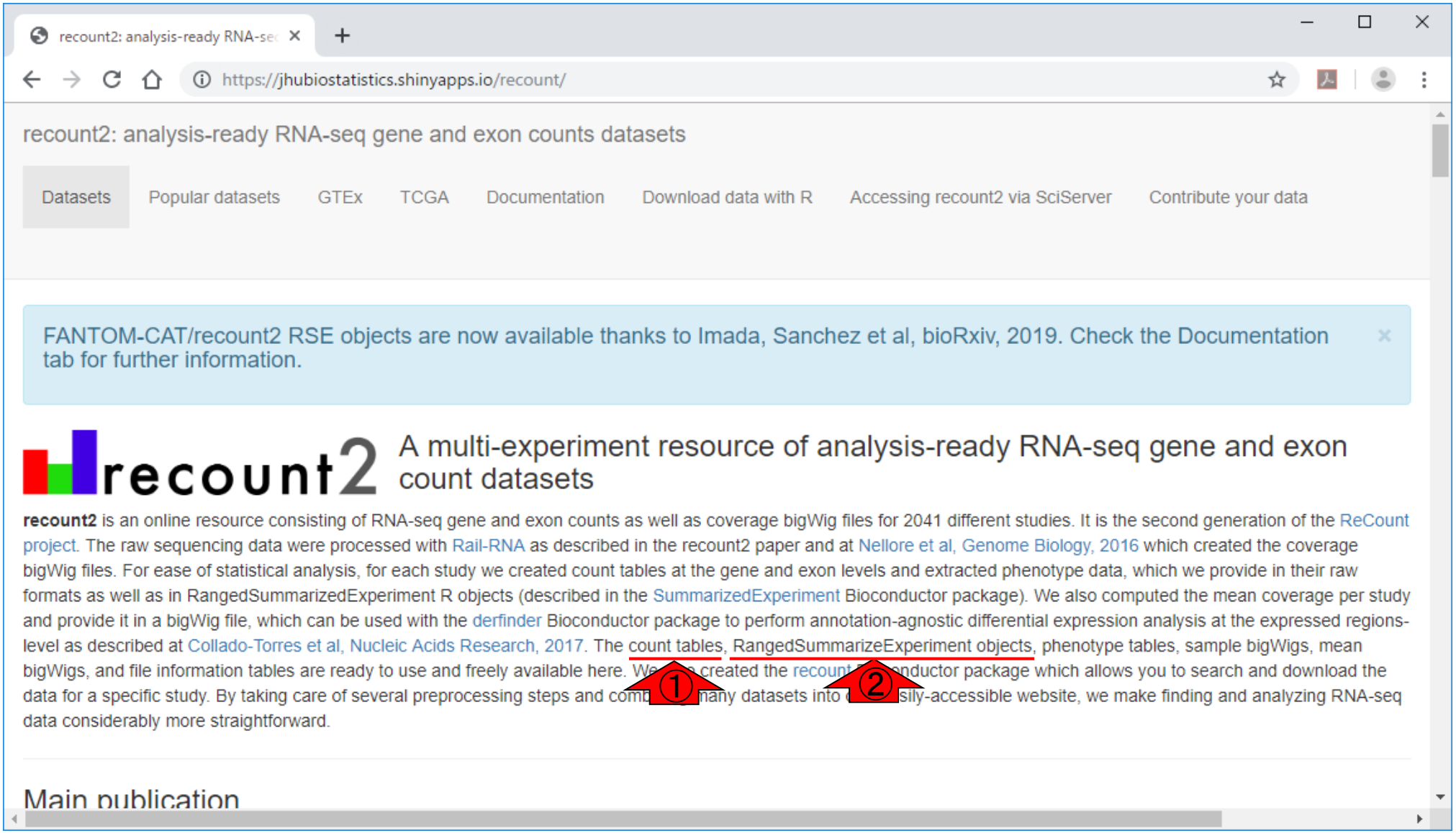
様々なfeature

①geneレベル、②exonレベル、③transcriptレベルなど、様々なfeatureのカウントデータが提供されてますね。



提供形式

提供している形式は、①生のカウントデータと、②RangedSummarizedExperimentというRオブジェクトの2種類。



RSE

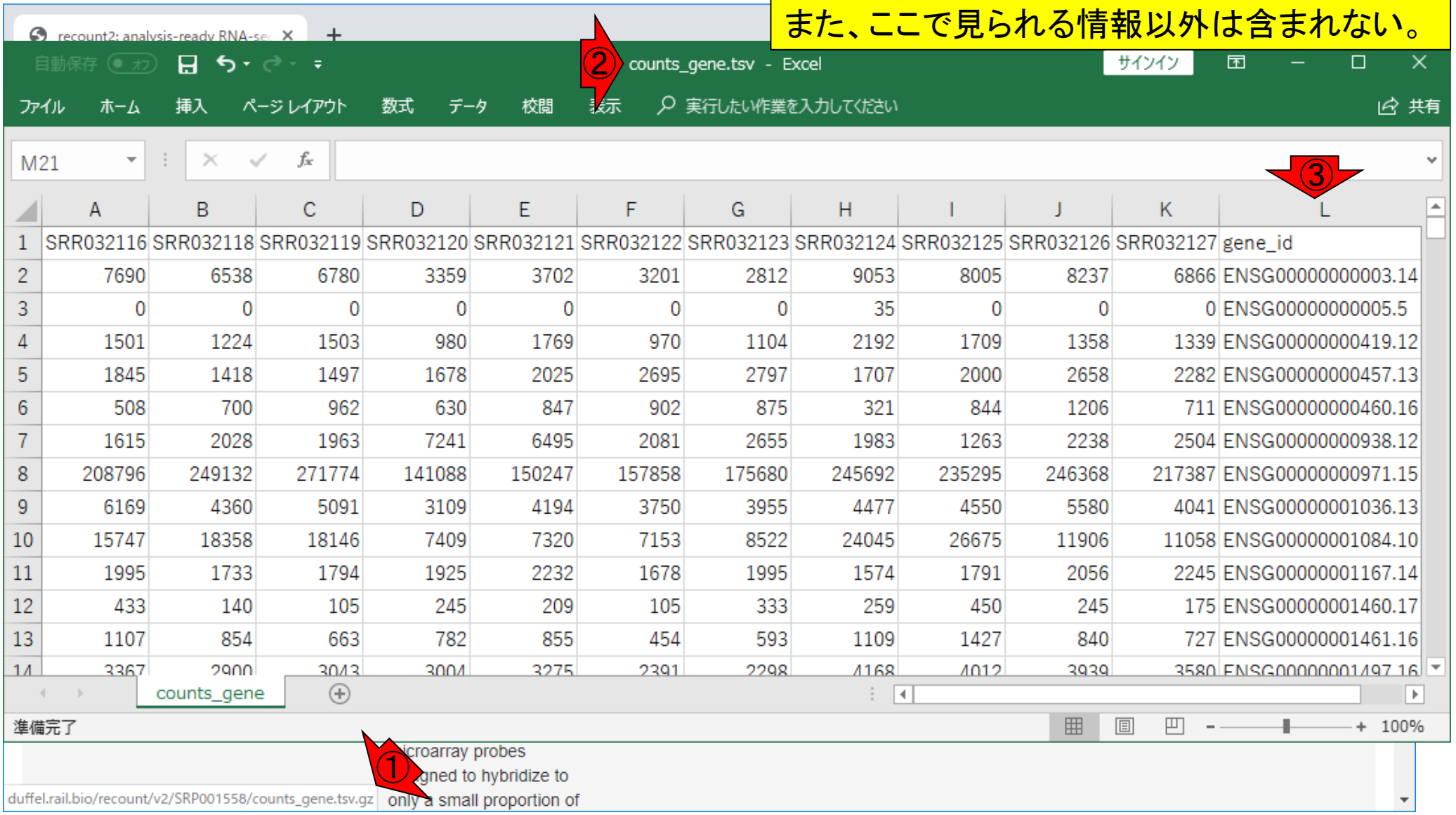
提供している形式は、①生のカウントデータと、②RangedSummarizedExperimentというRオブジェクトの2種類。
RangedSummarizedExperimentの略称の③RSE v2をクリックして得られる、④rse_gene.Rdataと…

Counts

提供している形式は、①生のカウントデータと、②RangedSummarizedExperimentというRオブジェクトの2種類。RangedSummarizedExperimentの略称の③RSE v2をクリックして得られる、④rse_gene.Rdataと、⑤生のカウントデータのcounts v2をクリックして得られる、⑥counts_gene.tsv.gzの2種類をダウンロード可能です。やったつもりでもよい。

生のカウント形式

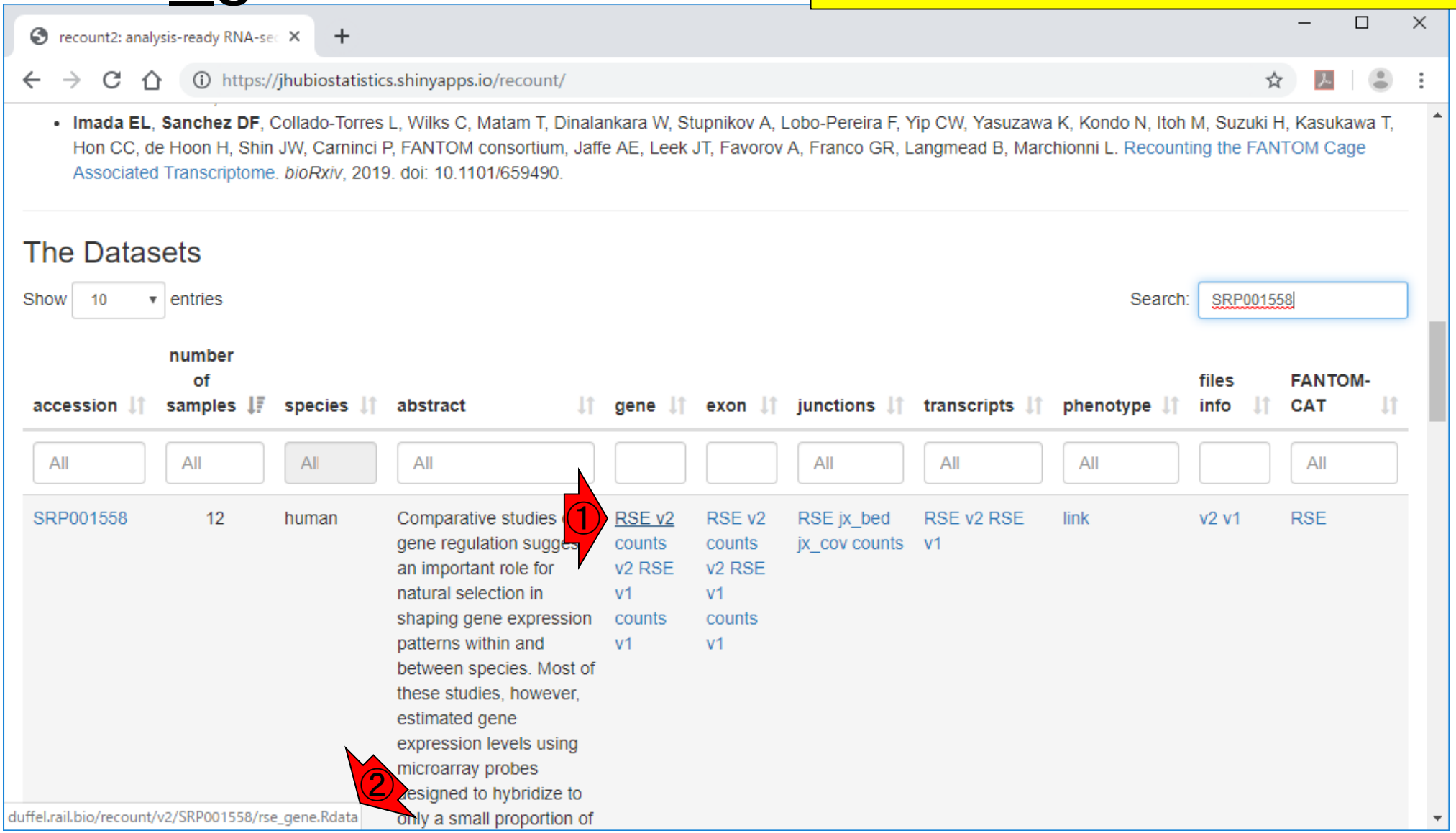
①counts_gene.tsv.gzを解凍し、Excel上で②counts_gene.tsvを眺める。なぜか③gene_id列が一番右側になっている（ので気持ち悪い）。また、ここで見られる情報以外は含まれない。



	A	B	C	D	E	F	G	H	I	J	K	L
1	SRR032116	SRR032118	SRR032119	SRR032120	SRR032121	SRR032122	SRR032123	SRR032124	SRR032125	SRR032126	SRR032127	gene_id
2	7690	6538	6780	3359	3702	3201	2812	9053	8005	8237	6866	ENSG00000000003.14
3	0	0	0	0	0	0	0	35	0	0	0	ENSG00000000005.5
4	1501	1224	1503	980	1769	970	1104	2192	1709	1358	1339	ENSG00000000419.12
5	1845	1418	1497	1678	2025	2695	2797	1707	2000	2658	2282	ENSG00000000457.13
6	508	700	962	630	847	902	875	321	844	1206	711	ENSG00000000460.16
7	1615	2028	1963	7241	6495	2081	2655	1983	1263	2238	2504	ENSG00000000938.12
8	208796	249132	271774	141088	150247	157858	175680	245692	235295	246368	217387	ENSG00000000971.15
9	6169	4360	5091	3109	4194	3750	3955	4477	4550	5580	4041	ENSG00000001036.13
10	15747	18358	18146	7409	7320	7153	8522	24045	26675	11906	11058	ENSG00000001084.10
11	1995	1733	1794	1925	2232	1678	1995	1574	1791	2056	2245	ENSG00000001167.14
12	433	140	105	245	209	105	333	259	450	245	175	ENSG00000001460.17
13	1107	854	663	782	855	454	593	1109	1427	840	727	ENSG00000001461.16
14	3367	2900	3043	3004	3275	2391	2298	4168	4012	3939	3580	ENSG00000001497.16

rse_gene.Rdata

①RSE v2をクリックすると、recount2から②rse_gene.Rdataをダウンロードできますが、迷惑をかけるのでここではやらないでください。



RSE形式を推奨

②rse_gene.Rdataをロードして得られるRangedSummarizedExperiment (RSE)形式のオブジェクトには、サンプルに付随する各種情報（メタデータ）や、geneの染色体上の位置、配列長、gene symbolsなど多くの情報が含まれているのでいろいろと便利です。なので慣れましょう。

Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、例題4
    - 例題5、例題6、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

Rパッケージrecount

①RSE v2をクリックすると、recount2から②rse_gene.Rdataをダウンロードできます。他の手段として、Rパッケージrecountを用いることで、③SRP001558の④geneレベルカウントデータ情報を含む、rse_gene.Rdataをダウンロードできます。

①がSRP001558のカウントデータを取得する項目。

Rパッケージrecount

(Rで)塩基配列解析

(last modified 2019/07/01, since 2010)

このウェブページ(
Macintosh2018.1
ています。 初心者
2018年7月に(R
(2018/07/18)

What's new? (

- マップ後 | カウント情報取得 | トランスクリプトーム | BEDファイルから (last modified 2014/06/21)
- カウント情報取得 | リアルデータ | について (last modified 2019/05/16)
- カウント情報取得 | リアルデータ | SRP061240 | recount(Collado-Torres_2017) (last modified 2018/08/28)
- カウント情報取得 | リアルデータ | SRP056295 | recount(Collado-Torres_2017) (last modified 2018/08/29)
- カウント情報取得 | リアルデータ | SRP056146 | recount(Collado-Torres_2017) (last modified 2018/10/25)
- カウント情報取得 | リアルデータ | SRP035988 | recount(Collado-Torres_2017) (last modified 2018/08/25)
- カウント情報取得 | リアルデータ | SRP026126 | recount(Collado-Torres_2017) (last modified 2018/08/30)
- カウント情報取得 | リアルデータ | SRP018853 | recount(Collado-Torres_2017) (last modified 2018/08/26)
- カウント情報取得 | リアルデータ | SRP012167 | recount(Collado-Torres_2017) (last modified 2018/08/24)
- カウント情報取得 | リアルデータ | SRP012167 | parathyroidSE(Haglund_2012) (last modified 2018/08/19)
- カウント情報取得 | リアルデータ | SRP001558 | recount(Collado-Torres_2017) ① (last modified 2018/08/08)
- カウント情報取得 | リアルデータ | SRP001540 | recount(Collado-Torres_2017) (last modified 2018/08/10)
- カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann_2013) (last modified 2018/07/03)
- カウント情報取得 | リアルデータ | ERP000546 | recount(Collado-Torres_2017) (last modified 2018/08/07)
- カウント情報取得 | シミュレーションデータ | RNA-seq | について (last modified 2019/04/05)

例題1

①例題1が、SRP001558のgeneレベルカウントデータを含むrse_gene.Rdataをダウンロードする基本形。
このあとダウンロードするので、ここではやらない。

カウント情報取得 | リアルデータ | SRP001558 | recount(Collado-Torres_2017)

recountパッケージを用いて、 SRP001558(Blekhman et al., Genome Res., 2010；ブラウザはIE以外を推奨) のカウント情報を含むRangedSummarizedExperimentクラスオブジェクトという形式の.Rdataをダウンロードしたり、カウントデータの数値行列にした状態で保存するやり方を示します。 原著論文では、3生物種(ヒト12 samples、チンパンジー12 samples、そしてアカゲザル12 samples)のカウントデータを取得しています。 ウェブサイトrecount2上でSRP001558で検索すると、 number of samplesが12、speciesがhumanとなっていることから、提供されているカウントデータはhumanに限定されていることがわかります。 例題2までで、なぜか11 samples分のデータしかないことに気づきます。これは、ウェブサイトrecount2上でSRP001558で検索し、 phenotype列のlinkをダウンロードして得られるSRP001558.tsv を眺めることでなんとなくの理由がわかります。私は、「**SRR032117のデータがおかしなことになっており、recount2で提供するクオリティに達しなかった。** このため、recount2のウェブページ上は12 samplesとなっているものの、カウントデータ自体は11 samples分となっているのだろう。」と予想しました。 また、PRJNA119135・GSE17274 ・SRA010277はENA上にリンク先がありますが、 ウェブサイトrecount2上では引っかかってきませんでした。 2018年8月7日に、recountWorkflow で推奨されているscale_counts関数実行後のカウントデータとなるように変更しました。
「ファイル」－「ディレクトリの変更」でダウンロードしたいディレクトリに移動し以下をコピペ。

1. geneレベルカウントデータ情報を得たい場合： ①

SRP001558という名前のフォルダが作成されます。中にあるrse-gene.Rdataをロードして読み込むとrse-geneというオブジェクト名で取り扱えます。 ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2をダウンロードして得られるrse_gene.Rdataと同じです。

```
param_ID <- "SRP001558"                 #IDを指定

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#本番(.Rdataをダウンロード)
download_study(param_ID, type="rse-gene", download=T)#ダウンロード
```

例題3

①例題3が、手元にあるrse_gene.Rdataを読み込んで、geneレベルカウントデータの数値行列を得る基本形。
②rse_gene.Rdataをデスクトップにダウンロード

カウント情報取得 | リアルデータ | SRP001558 | recount(Collado-Torres_2017)

recountパッケージを用いて、 SRP001558(Blekhman et al., Genome Res., 2010 ; ブラウザはIE以外を推奨) のカウント情報を含むRangedSummarizedExperimentクラスオブジェクトという形式の.Rdataをダウンロードしたり、カウントデータの数値行列にした状態で保存するやり方を示します。 原著論文では、3生物種(ヒト12 samples、チンパンジー12 samples、そしてアカゲザル12 samples)のカウントデータを取得しています。 ウェブサイトrecount2上でSRP001558で検索すると、 number of samplesが12、speciesがhumanとなっていることから、提供されているカウントデータはhumanに限定されていることがわかります。 例題2までで、なぜか11 samples分のデータしかないことができます。これは、ウェブサイトrecount2上でSRP001558で検索し、 phenoty
かります。私は、「**SRR03211**
のため、recount2のウェブページ
ろう。」と予想しました。 また、PR
recount2上では引っかかってきませ
後のカウントデータとなるように変更
「ファイル」－「ディレクトリの変更

1. geneレベルカウントデータ情報を

SRP001558という名前のフォルダ
エクト名で取り扱えます。 ウェブ
rse_gene.Rdataと同じです。

```
param_ID <- "SRP001558"

#必要なパッケージをロード
library(recount)

#本番(.Rdataをダウンロード)
download_study(param_ID, typ
```

①

②

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージの読み込み

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #in_fで指定した.Rdataをロード
rse <- rse_gene                             #rseとして取り扱う
rse                                         #確認してるだけです

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_counts実行(2018.08.07追加)
data <- assays(rse)$counts                  #カウントデータ行列を取得してdataに格納
dim(data)                                   #行数と列数を表示
head(data)                                  #確認してるだけです

#ファイルに保存
tmp <- cbind(rownames(data), data)          #保存したい情報をtmpに格納
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定したファイル名で
```

例題3

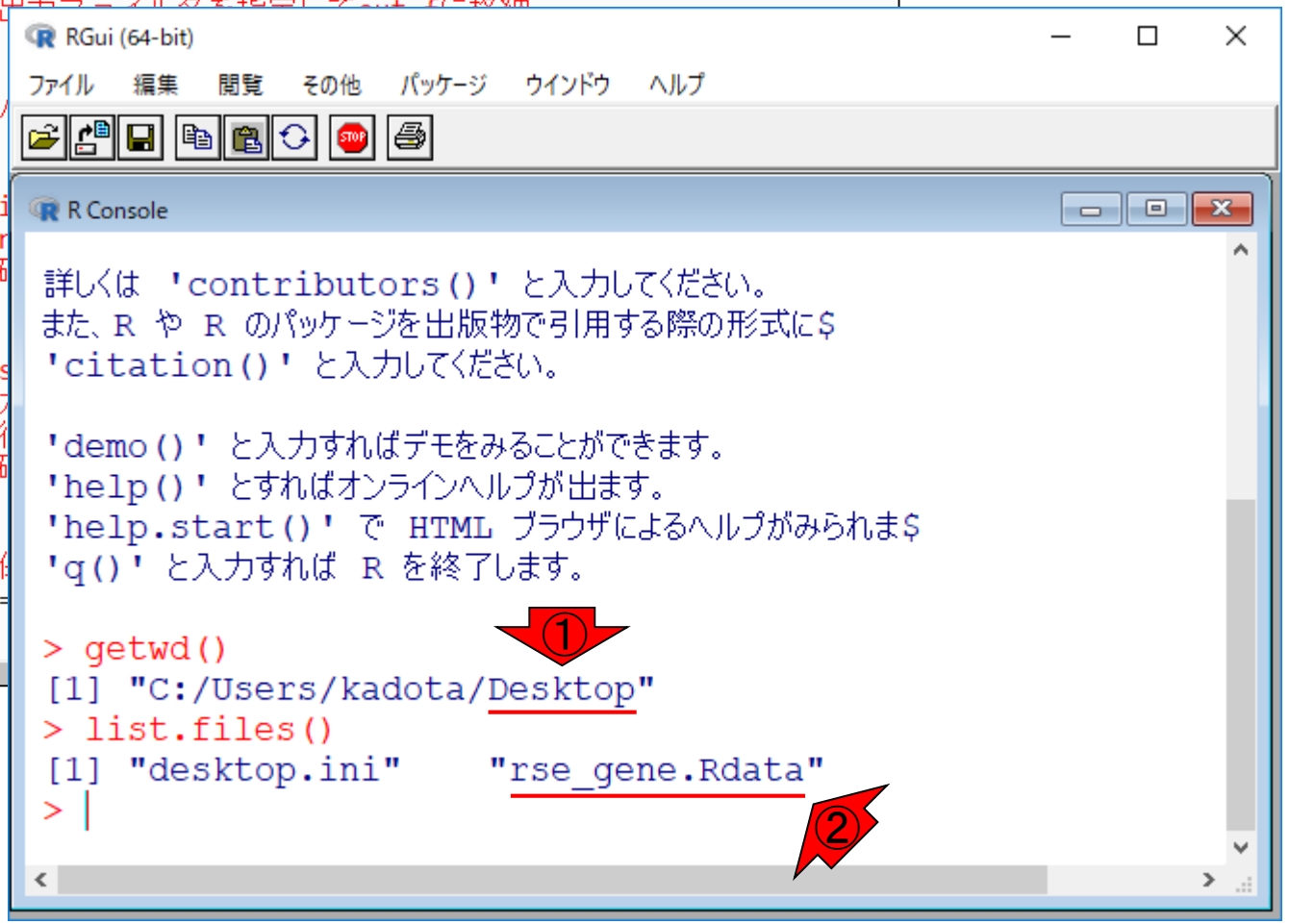
①作業ディレクトリをデスクトップにして、②rse_gene.Rdataがある状態で、とりあえず③の部分をコピペ。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)
rse <- rse_gene
rse

#本番(カウントデータ取得)
rse <- scale_counts(rse)
data <- assays(rse)$counts
dim(data)
head(data)

#ファイルに保存
tmp <- cbind(rownames(data), data)
write.table(tmp, out_f, sep="\t", append=
```

③



例題3

ここまでが、①rse_gene.Rdataを、②ロードして(取り込んで)、③オリジナルのrse_geneというオブジェクト名をrseに変更したものを、④表示した結果。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージをロード

#本番(type で指定した名前の.Rdataをロード)
load(in_f)                                  #in_fをロード
rse <- rse_gene                             #rse_geneをrseに変更
rse

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_counts関数でrseを変換
data <- assays(rse)$counts                  #assays(rse)$countsをdataに格納
dim(data)                                   #行列の大きさを確認
head(data)                                  #確認

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行名とdataを結合してtmpに格納
write.table(tmp, out_f, sep="\t", append=
```

①

②

③

④

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> rse                                       #確認してるだ$
class: RangedSummarizedExperiment
dim: 58037 11
metadata(0):
assays(1): counts
rownames(58037): ENSG00000000003.14
  ENSG00000000005.5 ... ENSG00000283698.1
  ENSG00000283699.1
rowData names(3): gene_id bp_length symbol
colnames(11): SRR032116 SRR032118 ... SRR032126
  SRR032127
colData names(21): project sample ... title
  characteristics
>
```

参考

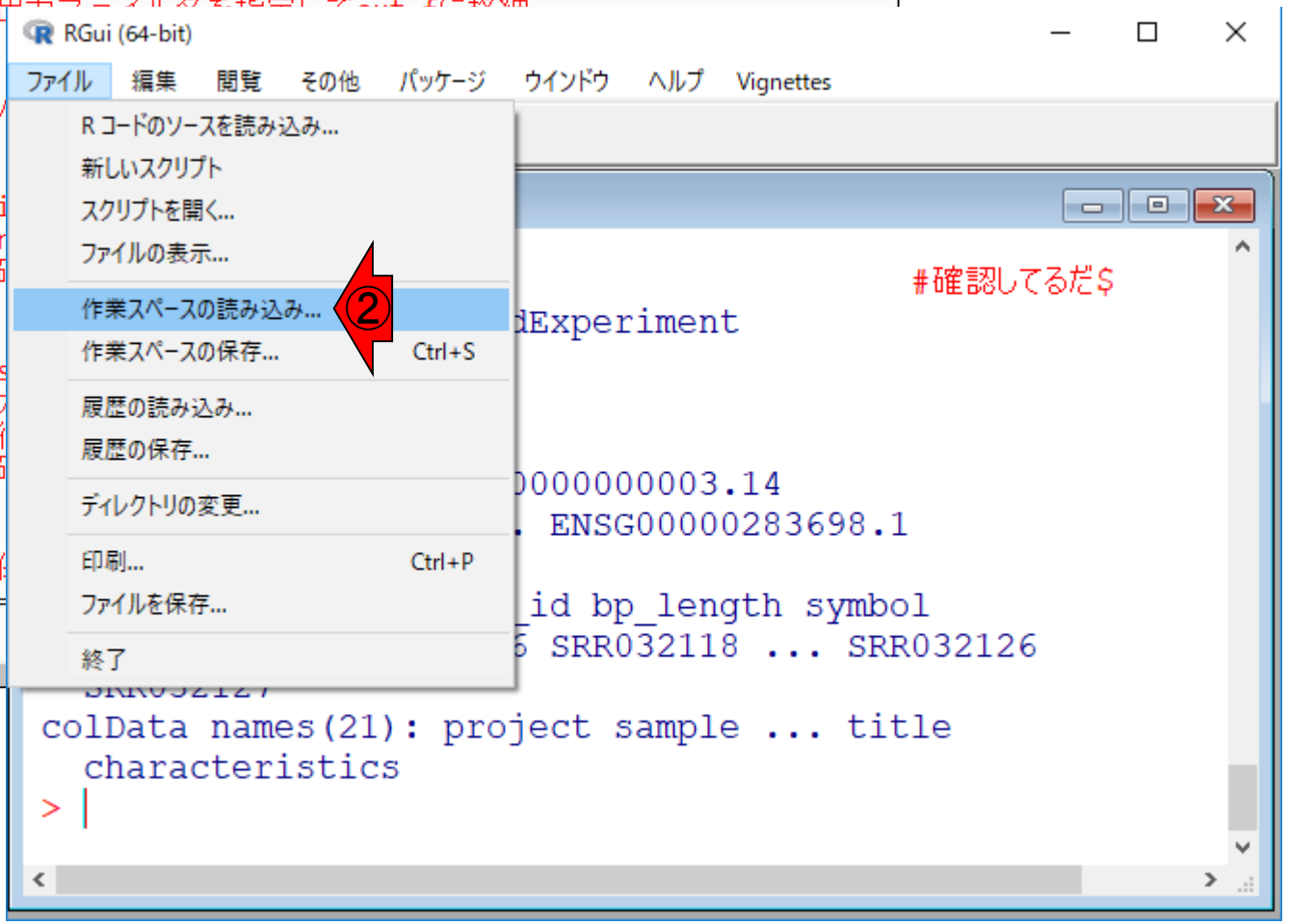
Tips：load関数

①load関数は、②「ファイル – 作業スペースの読み込み」で③rse_gene.Rdataを選択することと同義です。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パ

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #i
rse <- rse_gene                             #r
rse                                         #確

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #s
data <- assays(rse)$counts                  #カ
dim(data)                                   #行
head(data)                                  #確

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行
write.table(tmp, out_f, sep="\t", append=
```



Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ(technical replicatesマージ前)
  - Liverの3生物種間比較データ(technical replicatesマージ後)
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明(前半)
    - RangedSummarizedExperimentオブジェクトの説明(後半)、例題4
    - 例題5、例題6、ヒト(計6人分)のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5(19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得)
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行(課題)

RSE

①rseが、②RangedSummarizedExperiment (RSE)形式のオブジェクトです。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージをロード

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #in_fをロード
rse <- rse_gene                             #rse_geneをrseに格納
rse                                         #rseを確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_counts関数でrseのカウントデータを取得
data <- assays(rse)$counts                  #assays関数でrseのcountsをdataに格納
dim(data)                                   #dataの行数と列数を確認
head(data)                                  #dataの先頭を確認

#ファイルに保存
tmp <- cbind(rownames(data), data)          #dataの行名を1列目に追加してtmpに格納
write.table(tmp, out_f, sep="\t", append=
```

①

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> rse                                       #確認してるだ$
class: RangedSummarizedExperiment
dim: 58037 11
metadata(0):
assays(1): counts
rownames(58037): ENSG00000000003.14
  ENSG00000000005.5 ... ENSG00000283698.1
  ENSG00000283699.1
rowData names(3): gene_id bp_length symbol
colnames(11): SRR032116 SRR032118 ... SRR032126
  SRR032127
colData names(21): project sample ... title
  characteristics
>
```

②

カウントデータ格納部分

今とりあえず欲しいのは、カウントデータの数値行列情報。①countsという文字列をたよりに、②assaysというところに格納されているのだな、などと判断する。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージをロード

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #in_fをロード
rse <- rse_gene                             #rse_geneをrseに格納
rse                                         #確認してるだ$

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_countsでrseを変換
data <- assays(rse)$counts                  #assaysのcountsをdataに格納
dim(data)
head(data)

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行名とdataを結合してtmpに格納
write.table(tmp, out_f, sep="\t", append=
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> rse                                       #確認してるだ$
class: RangedSummarizedExperiment
dim: 58037 11
metadata(0):
assays(1): counts
rownames(58037): ENSG00000000003.14
  ENSG00000000005.5 ... ENSG00000283698.1
  ENSG00000283699.1
rowData names(3): gene_id bp_length symbol
colnames(11): SRR032116 SRR032118 ... SRR032126
  SRR032127
colData names(21): project sample ... title
  characteristics
>
```

①

②

カウントデータ格

私はある程度①classオブジェクトの概念やノリに慣れているので、まずは②str関数実行結果を眺める。そして、rseオブジェクト内から必要な情報をどのように得るかを試行錯誤する。そして大抵数回程度のトライアルで③のような書き方でよいという結論に至る。慣れないうちは、②の結果に加えてrecountパッケージのマニュアルを眺める必要もあるかも。

3. ダウンロード済みのrse_gene.Rdataを入力として読み

ウェブサイトrecount2上でSRP001558で検索し、 gene
geneレベルカウントデータ(rse_gene.Rdata; 約3MB)を
genes×11 samples)のみをタブ区切りテキストファイル

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パ

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #i
rse <- rse_gene                             #r
rse                                         #確

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #
data <- assays(rse)$counts                  #カ
dim(data)                                   #行
head(data)                                  #確

#ファイルに保存
tmp <- cbind(rownames(data), data)          #作
write.table(tmp, out_f, sep="\t", append=
```

③

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> rse                                       #確認してるだ$
class: RangedSummarizedExperiment
dim: 58037 11
metadata(0):
assays(1): counts
rownames(58037): ENSG00000000003.14
  ENSG00000000005.5 ... ENSG00000283698.1
  ENSG00000283699.1
rowData names(3): gene_id bp_length symbol
colnames(11): SRR032116 SRR032118 ... SRR032126
  SRR032127
colData names(21): project sample ... title
  characteristics
> str(rse)
```

①

②

参考

str実行結果

これが、②str(rse)実行結果の最後のほうの画面。画面がざ～っと流れる。慣れると表示結果からどんな情報が格納されているかの全貌を知ることができて便利。例えば、③ではassays(rse)$countsと書いているが、経験を積んでいくことで、④の部分を見た段階で「rse@assaysでもOKかも…」と思えるようになる。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE
geneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込ん
genes×11 samples)のみをタブ区切りテキストファイルで保存す

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージをロード

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #in_fをロード
rse <- rse_gene                             #rse_geneをrseに格納
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_countsでrseを変換
data <- assays(rse)$counts                  #assays(rse)$countsをdataに格納
dim(data)                                   #行数と列数を確認
head(data)                                  #確認

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行名とdataを結合してtmpに格納
write.table(tmp, out_f, sep="\t", append=
```

③

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

④

```
.. .. ..@ metadata       : list()
..@ assays          :Reference class 'ShallowSimpleLi$
.. ..$ data: NULL
.. ..and 14 methods.
..@ NAMES           : NULL
..@ elementMetadata:Formal class 'DataFrame' [packag$
.. .. ..@ rownames       : NULL
.. .. ..@ nrows          : int 58037
.. .. ..@ listData       : Named list()
.. .. ..@ elementType    : chr "ANY"
.. .. ..@ elementMetadata: NULL
.. .. ..@ metadata       : list()
..@ metadata        : list()
警告メッセージ:
```

①rse@assaysと②assays(rse)の結果。似てますね。

rse@assays

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パ

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #i
rse <- rse_gene                             #r
rse                                         #確

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #s
data <- assays(rse)$counts                  #デ
dim(data)                                   #
head(data)

#ファイルに保存
tmp <- cbind(rownames(data), data)          #作
write.table(tmp, out_f, sep="\t", append=
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
  .. .. ..@ elementMetadata: NULL
  .. .. ..@ metadata       : list()
  ..@ metadata       : list()
 警告メッセージ:
Not a validObject():  名前 "elementType" というスロッ$
> rse@assays
Reference class object of class "ShallowSimpleListAssa$
Field "data":
List of length 1
names(1): counts
> assays(rse)
List of length 1
names(1): counts
>
```

①

②

str(rse@assays)

再度①strでrse@assays内部の構造(structure)を眺める。②$ dataと書かれているので、rse@assays$dataをやってみようという思考回路になる。キーボードの上下左右の矢印キーを駆使して効率的に打ち込んでいますよね?!

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のと
geneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カ
genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パ

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #i
rse <- rse_gene                             #r
rse                                         #確

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #s
data <- assays(rse)$counts                  #カ
dim(data)                                   #行
head(data)                                  #確

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行
write.table(tmp, out_f, sep="\t", append
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
Not a validObject():  名前 "elementType" というスロッ$
> rse@assays
Reference class object of class "ShallowSimpleListAssa$
Field "data":
List of length 1
names(1): counts
> assays(rse)
List of length 1
names(1): counts
> str(rse@assays)
Reference class 'ShallowSimpleListAssays' [package "Su$
 $ data: NULL
 and 14 methods.
>
```

①

②

rse@assays$data

①rse@assays$dataの結果は、
②assays(rse)と同じですね。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                    #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                        #パッケージをロード

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                              #in_fをロード
rse <- rse_gene                         #rse_geneをrseに格納
rse                                     #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                #
data <- assays(rse)$counts              #rseからcountsを取得してdataに格納
dim(data)                               #
head(data)                              #h

#ファイルに保存
tmp <- cbind(rownames(data), data)      #作
write.table(tmp, out_f, sep="\t", append=
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
Field "data":
List of length 1
names(1): counts
> assays(rse)
List of length 1
names(1): counts
> str(rse@assays)
Reference class 'ShallowSimpleListAssays' [package "Su$
 $ data: NULL
 and 14 methods.
> rse@assays$data
List of length 1
names(1): counts
>
```

②

①

str(rse@assays$data)

①str(rse@assays$data)の結果より、②$ countsからrse@assays$data$countsでも、③と同じ意味だろうと想像したり、④dim関数で行数と列数を確認したりする

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージrecountをロード

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #in_fをロード
rse <- rse_gene                             #rse_geneをrseに格納
rse                                         #rseを確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #
data <- assays(rse)$counts                  #③
dim(data)                                   #
head(data)                                  #h

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行名を1列目に追加
write.table(tmp, out_f, sep="\t", append=
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> rse@assays$data
List of length 1
names(1): counts
> str(rse@assays$data)
Formal class 'SimpleList' [package "S4Vectors"] with 4$
  ..@ listData       :List of 1
.. ..$ counts: num [1:58037, 1:11] 7690 0 1501 1845 $
  .. .. ..- attr(*, "dimnames")=List of 2
  .. .. .. ..$ : chr [1:58037] "ENSG00000000003.14" "E$
  .. .. .. ..$ : chr [1:11] "SRR032116" "SRR032118" "S$
  ..@ elementType    : chr "ANY"
  ..@ elementMetadata: NULL
  ..@ metadata       : list()
> dim(rse@assays$data$counts)
```

①

②

④

様々な表現方法がある

①と②の実行結果は、確かに一致していますね。伝授したいことは、いろんな書き方がありますがうろたえることはありません、ということです。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージをロード

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #in_fをロード
rse <- rse_gene                             #rse_geneをrseに格納
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_counts()でrseを変換
data <- assays(rse)$counts                  #assays()$countsでdataに格納
dim(data)                                   #行列のサイズを確認
head(data)                                  #確認

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行名を1列目に追加
write.table(tmp, out_f, sep="\t", append
```

③

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
Formal class 'SimpleList' [package "S4Vectors"] with 4$
  ..@ listData       :List of 1
  .. ..$ counts: num [1:58037, 1:11] 7690 0 1501 1845 $
  .. .. ..- attr(*, "dimnames")=List of 2
  .. .. .. ..$ : chr [1:58037] "ENSG00000000003.14" "E$
  .. .. .. ..$ : chr [1:11] "SRR032116" "SRR032118" "S$
  ..@ elementType    : chr "ANY"
  ..@ elementMetadata: NULL
  ..@ metadata       : list()
> dim(rse@assays$data$counts)
[1] 58037    11
> dim(assays(rse)$counts)
[1] 58037    11
>
```

①

②

最後までコピペ実行した結果。

最後までコピペ

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パ

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #i
rse <- rse_gene                             #r
rse                                         #確

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #s
data <- assays(rse)$counts                  #カ
dim(data)                                   #行
head(data)                                  #確

#ファイルに保存
tmp <- cbind(rownames(data), data)          #保
write.table(tmp, out_f, sep="\t", append=
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
ENSG00000000460.16       210         53        142
ENSG00000000938.12       637        328        213
                    SRR032126 SRR032127
ENSG00000000003.14      1350       1220
ENSG00000000005.5          0          0
ENSG00000000419.12       223        238
ENSG00000000457.13       435        405
ENSG00000000460.16       198        126
ENSG00000000938.12       367        445
>
> #ファイルに保存
> tmp <- cbind(rownames(data), data)      #保存したい情$
> write.table(tmp, out_f, sep="\t", append=F, quote=F,$
>
```

最後までコピペ

58,037行×11列からなるカウント行列の、①最初の6行分の行名(rownames)と、②最後の2列分の列名(colnames)。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージをロード

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #in_fをロード
rse <- rse_gene                             #rse_geneをrseに格納
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_counts()でカウントデータ取得
data <- assays(rse)$counts                  #assays(rse)$countsをdataに格納
dim(data)                                   #行数と列数を確認
head(data)                                  #データの先頭を確認

#ファイルに保存
tmp <- cbind(rownames(data), data)          #作業用tmpに行名とdataを結合
write.table(tmp, out_f, sep="\t", append=F, quote=F, col.names=NA)
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
ENSG00000000460.16       210        53        142
ENSG00000000938.12       637       328        213
                    SRR032126 SRR032127
ENSG00000000003.14       1350      1220
ENSG00000000005.5           0         0
ENSG00000000419.12       223       238
ENSG00000000457.13       435       405
ENSG00000000460.16       198       126
ENSG00000000938.12       367       445
>
> #ファイルに保存
> tmp <- cbind(rownames(data), data)      #保存したい情$
> write.table(tmp, out_f, sep="\t", append=F, quote=F,$
>
```

①

②

Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、例題4
    - 例題5、例題6、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

RSE

①例題3で、②RangedSummarizedExperiment（RSE）というクラスオブジェクトである③rseを再度表示。
RSE形式からRSEクラスという表現に変えているが、他にもRSE containerやRSE objectなどいろんな呼び方をする。細かいことは気にしなくてよい。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合

①
ブサイトrecount2上でSRP001558で検索し、 gene列のRSE
geneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで
genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パ

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #i
rse <- rse_gene                             #
rse

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #s
data <- assays(rse)$counts                  #
dim(data)                                   #
head(data)                                  #

#ファイルに保存
tmp <- cbind(rownames(data), data)          #
write.table(tmp, out_f, sep="\t", append=
```

③

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> rse
class: RangedSummarizedExperiment
dim: 58037 11
metadata(0):
assays(1): counts
rownames(58037): ENSG00000000003.14
  ENSG00000000005.5 ... ENSG00000283698.1
  ENSG00000283699.1
rowData names(3): gene_id bp_length symbol
colnames(11): SRR032116 SRR032118 ... SRR032126
  SRR032127
colData names(21): project sample ... title
  characteristics
>
```

②

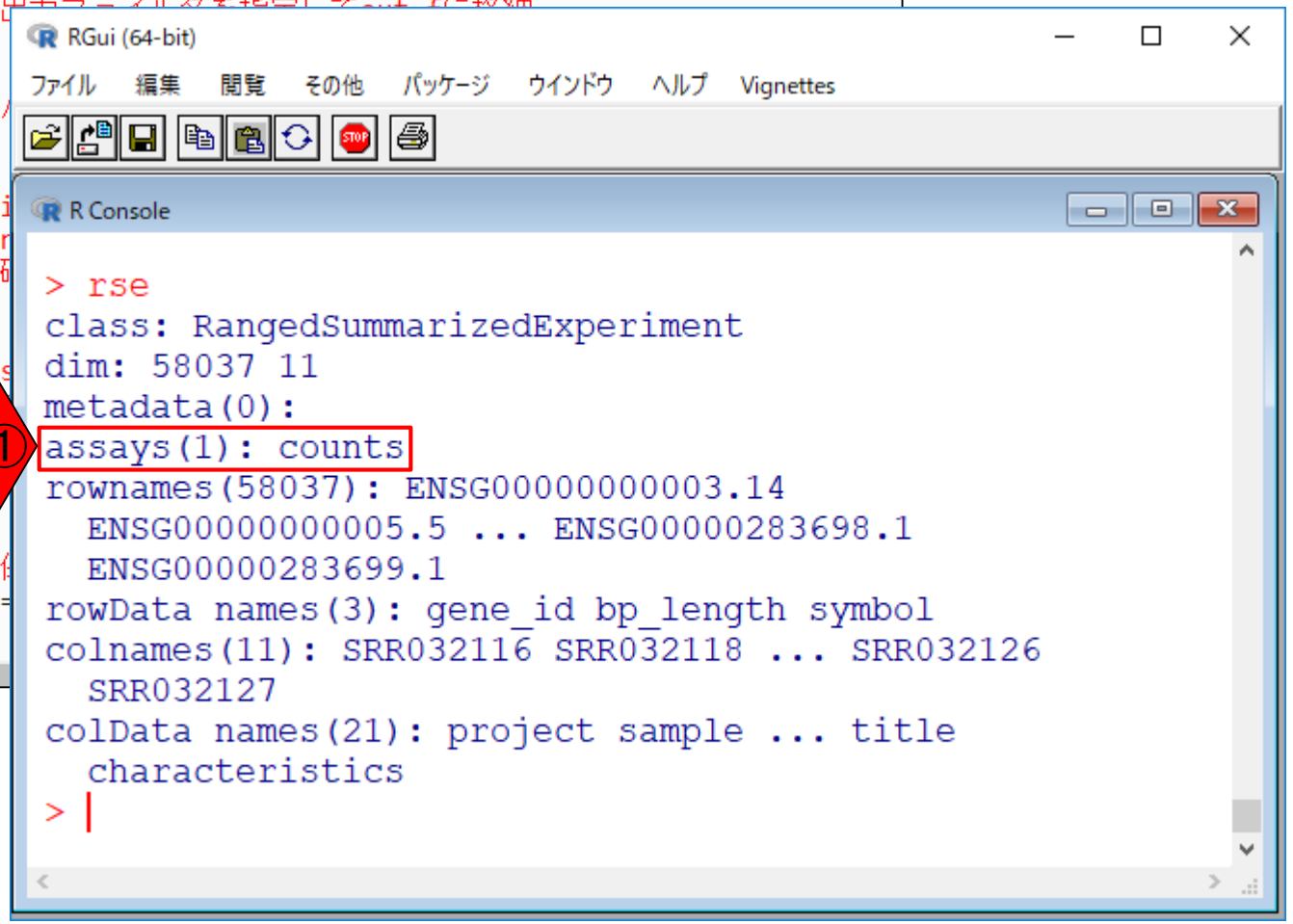
RSE

これまで主に着目していたのは、①カウントデータ取得に関するものであった。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージrecountをロード

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #in_fをロード
rse <- rse_gene                             #rse_geneをrseに格納
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_counts関数でrseを変換
data <- assays(rse)$counts                  #rseからcountsを抽出してdataに格納
dim(data)
head(data)

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行名を1列目に追加
write.table(tmp, out_f, sep="\t", append=
```



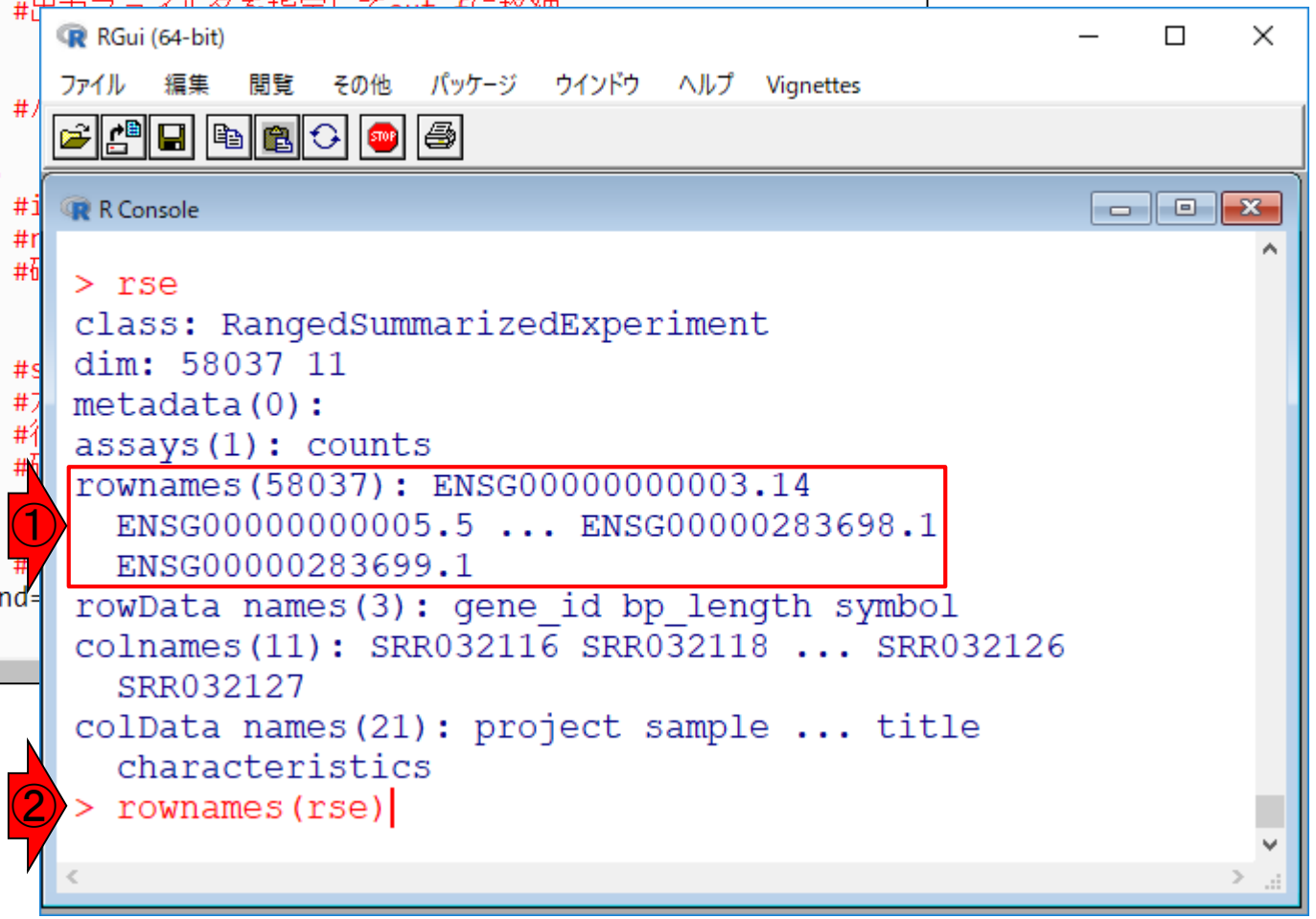
rownames

①はカウントデータ行列の行名情報。②rownames(rse)で取り出せる。58,037個の行名が一気に表示される。やらなくてもよい。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パ

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #i
rse <- rse_gene                             #r
rse                                         #確

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #s
data <- assays(rse)$counts                  #カ
dim(data)                                   #行
head(data)                                  #カ

#ファイルに保存
tmp <- cbind(rownames(data), data)          #
write.table(tmp, out_f, sep="\t", append=
```



rowData

①はgeneレベルカウントデータ行列の行(gene)ごとの付随情報としてどのようなものがあるかを示している。gene_id, bp_length, symbolの3種類の情報を、②rowData(rse)で取り出せる。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージrecountをロード

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #in_fをロード
rse <- rse_gene                             #rse_geneをrseに格納
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_counts関数でrseを変換
data <- assays(rse)$counts                  #assays(rse)$countsをdataに格納
dim(data)                                   #行数と列数を確認
head(data)                                  #確認

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行名とdataを結合してtmpに格納
write.table(tmp, out_f, sep="\t", apper
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> rse
class: RangedSummarizedExperiment
dim: 58037 11
metadata(0):
assays(1): counts
rownames(58037): ENSG00000000003.14
  ENSG00000000005.5 ... ENSG00000283698.1
  ENSG00000283699.1
rowData names(3): gene_id bp_length symbol
colnames(11): SRR032116 SRR032118 ... SRR032126
  SRR032127
colData names(21): project sample ... title
  characteristics
> rowData(rse)
```

①

②

①rowData(rse)実行結果。

rowData(rse)

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"

#必要なパッケージをロード
library(recount)

#本番(typeで指定した名前の.Rd
load(in_f)
rse <- rse_gene
rse

#本番(カウントデータ取得)
rse <- scale_counts(rse)
data <- assays(rse)$counts
dim(data)
head(data)

#ファイルに保存
tmp <- cbind(rownames(data),
write.table(tmp, out_f, sep=
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

①

```
> rowData(rse)
DataFrame with 58037 rows and 3 columns
                          gene_id bp_length          symbol
                      <character> <integer> <CharacterList>
ENSG00000000003.14 ENSG00000000003.14      4535          TSPAN6
ENSG00000000005.5   ENSG00000000005.5      1610            TNMD
ENSG00000000419.12 ENSG00000000419.12      1207            DPM1
ENSG00000000457.13 ENSG00000000457.13      6883           SCYL3
ENSG00000000460.16 ENSG00000000460.16      5967        C1orf112
...                                 ...       ...             ...
ENSG00000283695.1   ENSG00000283695.1        61              NA
ENSG00000283696.1   ENSG00000283696.1       997              NA
ENSG00000283697.1   ENSG00000283697.1      1184   LOC101928917
ENSG00000283698.1   ENSG00000283698.1       940              NA
```

rowData(rse)

①gene_idが抽出したカウントデータ行列の行名。②bp_lengthが配列長。RPKM/FPKM/TPM値を得る際の基礎情報として使います。③symbolがgene symbol情報です。機能解析（GO解析やパスウェイ解析）を行う際には、gene symbol情報で対応付けを行う必要があります。このような情報を保持しているRangedSummarizedExperimentオブジェクトを使いこなせると大変便利。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込

ウェブサイトrecount2上でSRP001558で検索し、 gene列の
geneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み
genes×11 samples)のみをタブ区切りテキストファイルで保

```
in_f <- "rse_gene.Rdata"                    #入力ファイ
out_f <- "hoge3.txt"

#必要なパッケージをロード
library(recount)

#本番(typeで指定した名前の.Rd
load(in_f)
rse <- rse_gene
rse

#本番(カウントデータ取得)
rse <- scale_counts(rse)
data <- assays(rse)$counts
dim(data)
head(data)

#ファイルに保存
tmp <- cbind(rownames(data),
write.table(tmp, out_f, sep=
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> rowData(rse)
DataFrame with 58037 rows and 3 columns
                          gene_id bp_length          symbol
                      <character> <integer> <CharacterList>
ENSG00000000003.14 ENSG00000000003.14      4535          TSPAN6
ENSG00000000005.5   ENSG00000000005.5      1610            TNMD
ENSG00000000419.12 ENSG00000000419.12      1207            DPM1
ENSG00000000457.13 ENSG00000000457.13      6883           SCYL3
ENSG00000000460.16 ENSG00000000460.16      5967        C1orf112
...                               ...       ...             ...
ENSG00000283695.1   ENSG00000283695.1        61              NA
ENSG00000283696.1   ENSG00000283696.1       997              NA
ENSG00000283697.1   ENSG00000283697.1      1184   LOC101928917
ENSG00000283698.1   ENSG00000283698.1       940              NA
```

colnames

①もう一度rseを表示。②colnamesという名前と赤枠内に表示されている情報から、カウントデータ行列の列名部分に相当するものだということが分かる。③で確認できるがやらなくてもよい。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージrecountをロード

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #in_fをロード
rse <- rse_gene                             #rse_geneをrseに格納
rse

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_counts関数でrseを変換
data <- assays(rse)$counts                  #assays(rse)$countsをdataに格納
dim(data)                                   #行列dataの行数と列数を表示
head(data)                                  #行列dataの先頭6行を表示

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行列dataの行名と行列dataを結合
write.table(tmp, out_f, sep="\t", append=
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> rse
class: RangedSummarizedExperiment
dim: 58037 11
metadata(0):
assays(1): counts
rownames(58037): ENSG00000000003.14
  ENSG00000000005.5 ... ENSG00000283698.1
  ENSG00000283699.1
rowData names(3): gene_id bp_length symbol
colnames(11): SRR032116 SRR032118 ... SRR032126
  SRR032127
colData names(21): project sample ... title
  characteristics
> colnames(rse)
```

①

②

③

colData

これまでのノリから、①サンプルに相当する列ごとに②21個の付随情報があるのではないかと予想する。③を実行すると一気に画面が流れるがやってみる。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パ

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #i
rse <- rse_gene                             #r
rse                                         #確

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #s
data <- assays(rse)$counts                  #カ
dim(data)                                   #行
head(data)                                  #確

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行
write.table(tmp, out_f, sep="\t", append=
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> rse
class: RangedSummarizedExperiment
dim: 58037 11
metadata(0):
assays(1): counts
rownames(58037): ENSG00000000003.14
  ENSG00000000005.5 ... ENSG00000283698.1
  ENSG00000283699.1
rowData names(3): gene_id bp_length symbol
colnames(11): SRR032116 SRR032118 ... SRR032126
  SRR032127
colData names(21): project sample ... title
  characteristics
> colData(rse)
```

①

②

③

colData(rse)

colData(rse)実行結果。この画面サイズだと、最後の①characteristics列しか見えていない。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パ

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #i
rse <- rse_gene                             #r
rse                                         #確

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #s
data <- assays(rse)$counts                  #カ
dim(data)                                   #行
head(data)                                  #確

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行
write.table(tmp, out_f, sep="\t", append=
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

①

```
                    characteristics
                    <CharacterList>
SRR032116 sex: female,tissue: liver
SRR032118 sex: female,tissue: liver
SRR032119 sex: female,tissue: liver
SRR032120 sex: female,tissue: liver
SRR032121 sex: female,tissue: liver
SRR032122   sex: male,tissue: liver
SRR032123   sex: male,tissue: liver
SRR032124   sex: male,tissue: liver
SRR032125   sex: male,tissue: liver
SRR032126   sex: male,tissue: liver
SRR032127   sex: male,tissue: liver
> 
```

colData(rse)

このあたり(じゃなくてもよいが)で、元々このデータは計12 samplesのはずだったのに、なぜ11 samplesしかないのだろう?!と思い始める。①femaleが5 samplesしかないので、femaleサンプルのうちの1つがカウントデータに含まれていないのだろうと判断する。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE
geneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込ん
genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パ

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #i
rse <- rse_gene                             #r
rse                                         #確

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #s
data <- assays(rse)$counts                  #カ
dim(data)                                   #行
head(data)                                  #確

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行
write.table(tmp, out_f, sep="\t", append=
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
                    characteristics
                    <CharacterList>
SRR032116 sex: female,tissue: liver
SRR032118 sex: female,tissue: liver
SRR032119 sex: female,tissue: liver
SRR032120 sex: female,tissue: liver
SRR032121 sex: female,tissue: liver
SRR032122   sex: male,tissue: liver
SRR032123   sex: male,tissue: liver
SRR032124   sex: male,tissue: liver
SRR032125   sex: male,tissue: liver
SRR032126   sex: male,tissue: liver
SRR032127   sex: male,tissue: liver
> |
```

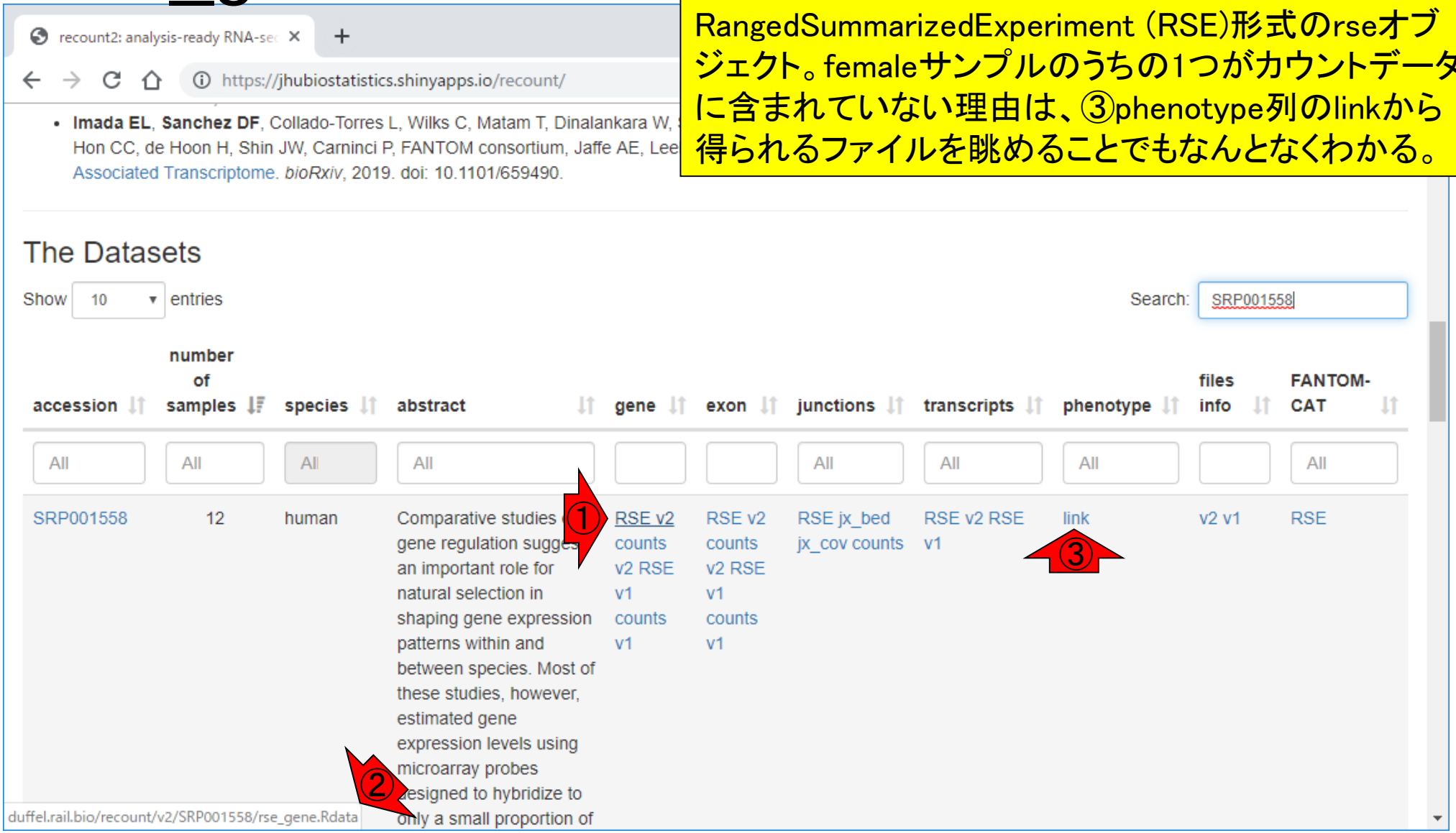
①

②

参考

rse_gene.Rdata

おさらい。ず～っと説明しているのは、ウェブサイトrecount2から①RSE v2をクリックして得られた、②rse_gene.Rdataを読み込んで得られたRangedSummarizedExperiment (RSE)形式のrseオブジェクト。femaleサンプルのうちの1つがカウントデータに含まれていない理由は、③phenotype列のlinkから得られるファイルを眺めることでもなんとなくわかる。



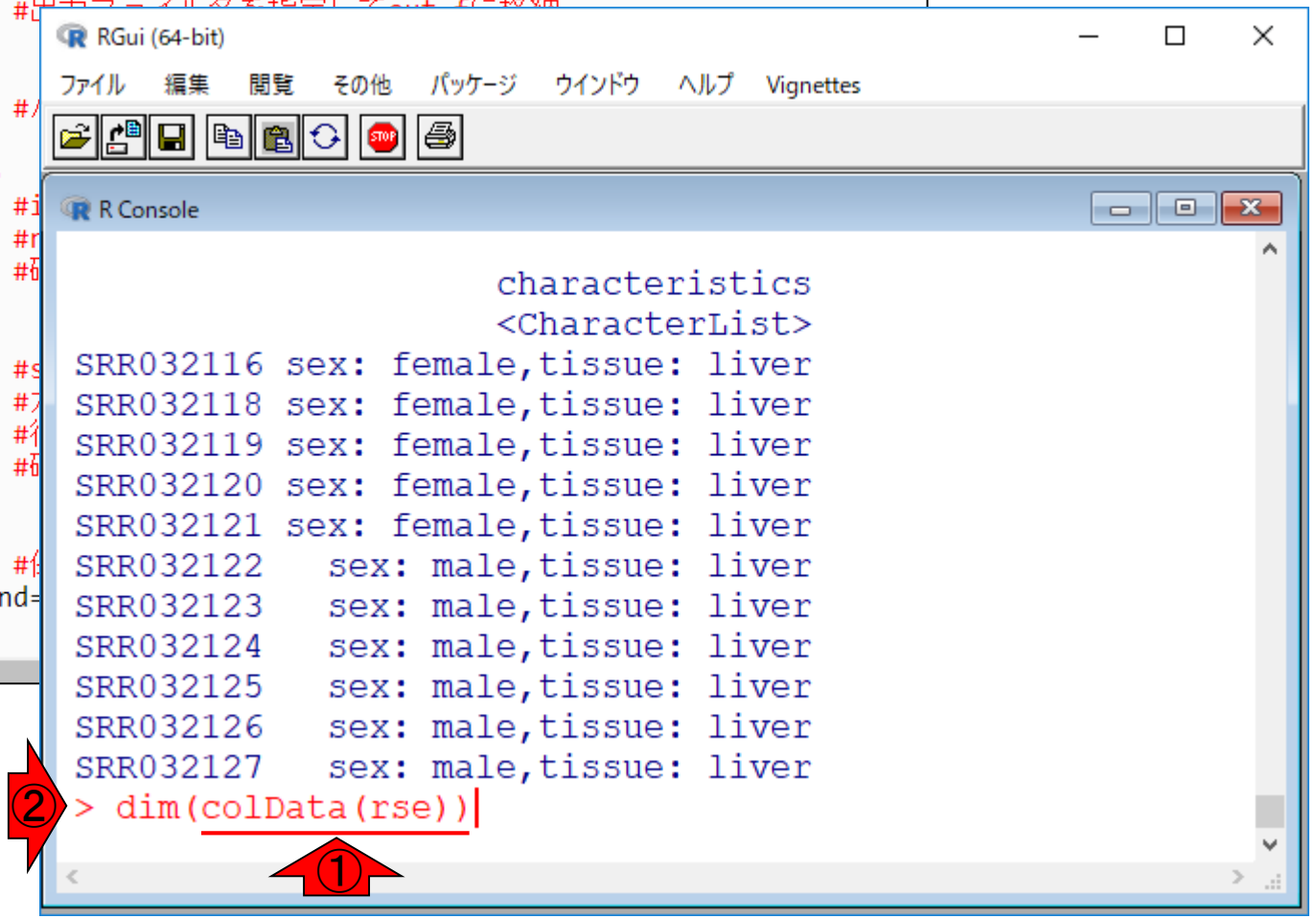
colData(rse)

①colData(rse)の行数と列数を把握すべく、②dimを実行。キーボードの上下左右の矢印キーを駆使して効率的に打ち込んでいますよね?!

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージをロード

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #in_fをロード
rse <- rse_gene                             #rse_geneをrseに格納
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #スケールカウントデータ取得
data <- assays(rse)$counts                  #カウントデータ取得
dim(data)                                   #行数と列数を確認
head(data)                                  #確認

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行名を1列目に追加
write.table(tmp, out_f, sep="\t", append=
```



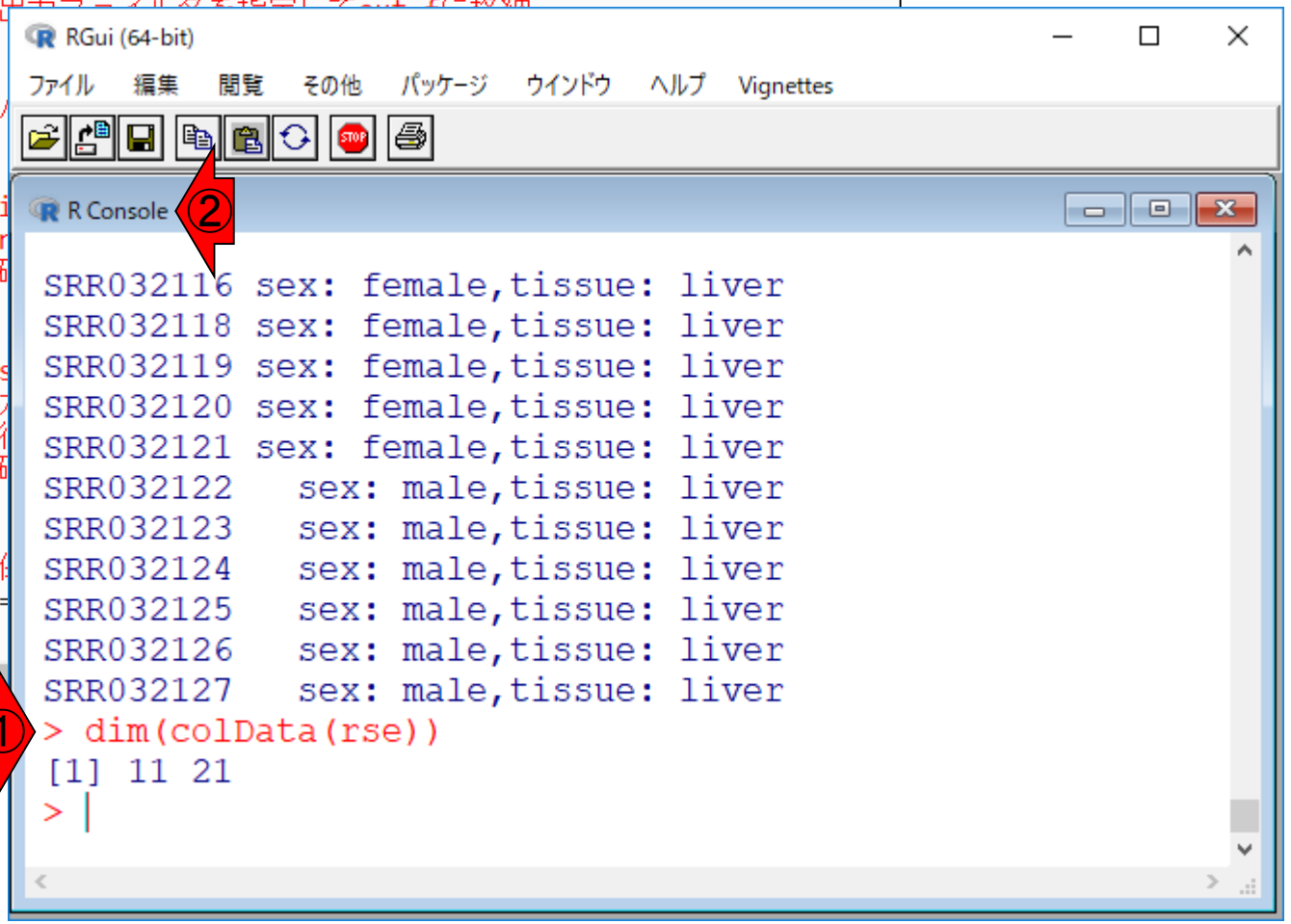
dim(colData(rse))

①colData(rse)は、11行×21列の情報からなる。これだけの情報量になると②R Console画面上で判断するのは難しいのでファイルに保存してExcelで眺めることにする。それが例題4。

3. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

ウェブサイトrecount2上でSRP001558で検索し、 gene列のRSE v2のところからダウンロードして得られたgeneレベルカウントデータ(rse_gene.Rdata; 約3MB)を読み込んで、 カウントの数値行列情報(58,037 genes×11 samples)のみをタブ区切りテキストファイルで保存するやり方です。出力ファイルはhoge3.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パ

#本番(typeで指定した名前の.Rdataをロード)
load(in_f)                                  #i
rse <- rse_gene                             #r
rse                                         #確

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #s
data <- assays(rse)$counts                  #カ
dim(data)                                   #行
head(data)                                  #確

#ファイルに保存
tmp <- cbind(rownames(data), data)          #行
write.table(tmp, out_f, sep="\t", append=
```



Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、**例題4**
    - 例題5、例題6、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

例題4

ここで着目してほしいのは、①RSE形式のrseオブジェクトから、②サンプルのメタデータ情報をファイル（srp001558_meta_samples.txt）に落とすところのみ。③コードの下部に移動。

4. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題3とベースとして、さらにサンプルのメタデータ情報ファイル(srp001558_meta_samples.txt)と、 遺伝子(features)のメタデータ情報ファイル(srp001558_meta_features.txt)も出力するやり方です。 58,037 genes×11 samplesからなるカウントデータファイル(hoge4_counts.txt)は列名をSRR...からSRS...に変更しています。 このデータセットの場合は、なぜかtechnical replicatesのサンプルに対して別々のSRS IDが付与されているので、列名変更はほぼ無意味です。

```
in_f <- "rse_gene.Rdata"                  #入力ファイル名を指定してin_fに格納
out_f1 <- "hoge4_counts.txt"              #入力ファイル名を指定してout_f1に格納(カウントデータ)
out_f2 <- "srp001558_meta_samples.txt"    #入力ファイル名を指定してout_f2に格納(samplesメタデータ)
out_f3 <- "srp001558_meta_features.txt"   #入力ファイル名を指定してout_f3に格納(featuresメタデータ)

#必要なパッケージをロード
library(recount)                          #パッケージの読み込み

#入力ファイルの読み込み
load(in_f)                                #in_fで指定した.Rdataをロード
rse <- rse_gene                           #rseとして取り扱う
rse                                       #確認してるだけです

#本来はカウントデータ取得)
rse <- scale_counts(rse)                  #scale_counts実行(2018.08.07追加)
data <- assays(rse)$counts                #カウントデータ行列を取得してdataに格納
colnames(data) <- colData(rse)$sample     #列名をERR...からERS...に変更
dim(data)                                 #行数と列数を表示
head(data)                                #確認してるだけです

#ファイルに保存(カウントデータ)
tmp <- cbind(rownames(data), data)        #保存したい情報をtmpに格納
```

①コードの下部に移動したところ。

例題4

4. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題3とベースとして、さらにサンプルのメタデータ情報ファイル(srp001558_meta_samples.txt)と、 遺伝子(features)のメタデータ情報ファイル(srp001558_meta_features.txt)も出力するやり方です。 58,037 genes×11 samplesからなるカウントデータファイル(hoge4_counts.txt)は列名をSRR...からSRS...に変更しています。 このデータセットの場合は、なぜかtechnical replicatesのサンプルに対して別々のSRS IDが付与されているので、列名変更はほぼ無意味です。

```
rse <- rse_gene                          #rseとして取り扱う
rse                                      #確認してるだけです

#本番(カウントデータ取得)
rse <- scale_counts(rse)                 #scale_counts実行(2018.08.07追加)
data <- assays(rse)$counts               #カウントデータ行列を取得してdataに格納
colnames(data) <- colData(rse)$sample    #列名をERR...からERS...に変更
dim(data)                                #行数と列数を表示
head(data)                               #確認してるだけです

#ファイルに保存(カウントデータ)
tmp <- cbind(rownames(data), data)       #保存したい情報をtmpに格納
write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定したファイル名

#ファイルに保存(サンプルのメタデータ情報)
tmp <- colData(rse)                      #保存したい情報をtmpに格納
write.table(tmp, out_f2, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定したファイル名

#ファイルに保存(featuresのメタデータ情報)
tmp <- rowData(rse)                      #保存したい情報をtmpに格納
write.table(tmp, out_f3, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定したファイル名
```

①

例題4

①赤枠部分が今着目してもらいたいところ。②rseは、rse_gene.Rdataを読み込んで得られたRangedSummarizedExperiment (RSE)形式のオブジェクト。③colData(rse)の中身を、そのまま④out_f2（srp001558_meta_samples.txtのこと）に⑤タブ区切りテキスト形式で保存している。コード全体をコピペ実行してください。

4. ダウンロード済みのrse_gene.Rdataを入力

例題3とベースとして、さらにサンプルのメタ
(features)のメタデータ情報ファイル(srp001
genes×11 samplesからなるカウントデータファイル(hoge4_counts.txt)は列名をSRR...からSRS...に変更しています。 このデータセットの場合は、なぜかtechnical replicatesのサンプルに対して別々のSRS IDが付与されているので、列名変更はほぼ無意味です。

```
rse <- rse_gene                          #rseとして取り扱う
rse                                      #確認してるだけです

#本                カウントデータ取得)
rse <- scale_counts(rse)                 #scale_counts実行(2018.08.07追加)
data <- assays(rse)$counts               #カウントデータ行列を取得してdataに格納
colnames(data) <- colData(rse)$sample    #列名をERR...からERS...に変更
dim(data)                                #行数と列数を表示
head(data)                               #確認してるだけです

#ファイルに保存(カウントデータ)
tmp <- cbind(rownames(data), data)       #保存したい情報をtmpに格納
write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定したファイル名

#ファイルに保存(サン          のメタデータ情報)
tmp <- colData(rse)                      #保存したい情報をtmpに格納
write.table(tmp, out_f2, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定したファイル名

#ファイルに保存(fea          のメタ          情報)
tmp <- rowData(rse)                      #保存したい情報をtmpに格納
write.table(tmp, out_f3, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定したファイル名
```

colData(rse)

赤枠部分が、ヘッダー行を除くと11行×21列のcolData(rse)の中身を、①(srp001558_meta_samples.txt)に保存した結果。Excelで読み込ませて、今は注目に値しない列の幅を狭めて表示させたものです。

4. ダウンロード済みのrse_gene.Rdataを入力として読み込む場

例題3とベースとして、さらにサンプルのメタデータ情報ファイル(srp001558_meta_samples.txt)と、遺伝子(features)のメタデータ情報ファイル(srp001558_meta_features.txt)も出力するやり方です。 58,037 genes×11 samplesからなるカウントデータファイル(hoge4_counts.txt)は列名をSRR...からSRS...に変更しています。 このデータセットの場合は、なぜかtechnical replicatesのサンプルに対して別々のSRS IDが付与されているので、列名変更はほぼ無意味です。

```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定してin_fに格納
out_f1 <- "hoge4_counts.txt"            #出力ファイル名を指定してout_f1に格納(カウントデータ)
out_f2 <- "srp001558_meta_samples.txt"  #出力ファイル名を指定してout_f2に格納(samplesメタデータ)
out_f3 <- "srp001558_meta_features.txt" #出力ファイル名を指定してout_f3に格納(featuresメタデータ)

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#入力フ
load(in
rse <-
rse

#本番(カ
rse <-
data <-
colname
dim(dat
head(da

#ファイ
tmp <-
```

①

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	pro	sample	exp	run	rea	rea	pro	pai	sra	ma	au	sh	sh	bio	bio	bio	avg	ge	big	title	characteristics
2	SRF	SRS009313	SR	SRR032116	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 1 rep1	c("sex: female", "tissue: liver")
3	SRF	SRS009315	SR	SRR032118	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 2 rep1	c("sex: female", "tissue: liver")
4	SRF	SRS009316	SR	SRR032119	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 2 rep2	c("sex: female", "tissue: liver")
5	SRF	SRS009317	SR	SRR032120	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 3 rep1	c("sex: female", "tissue: liver")
6	SRF	SRS009318	SR	SRR032121	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 3 rep2	c("sex: female", "tissue: liver")
7	SRF	SRS009319	SR	SRR032122	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 1 rep1	c("sex: male", "tissue: liver")
8	SRF	SRS009320	SR	SRR032123	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 1 rep2	c("sex: male", "tissue: liver")
9	SRF	SRS009321	SR	SRR032124	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 2 rep1	c("sex: male", "tissue: liver")
10	SRF	SRS009322	SR	SRR032125	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 2 rep2	c("sex: male", "tissue: liver")
11	SRF	SRS009323	SR	SRR032126	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 3 rep1	c("sex: male", "tissue: liver")
12	SRF	SRS009324	SR	SRR032127	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 3 rep2	c("sex: male", "tissue: liver")

colData(rse)

赤枠部分が、ヘッダー行を除くと11行×21列のcolData(rse)の中身を、①（srp001558_meta_samples.txt）に保存した結果。Excelで読み込ませて、今は注目に値しない列の幅を狭めて表示させたものです。②がcolData(rse)$sample、③がcolData(rse)$run、そして④がcolData(rse)$titleで取り出せる情報となる。

4. ダウンロード済みのrse_gene.Rdataを入力として読み込む場

例題3とベースとして、さらにサンプルのメタデータ情報ファイル
(features)のメタデータ情報ファイル(srp001558_meta_featu
genes×11 samplesからなるカウントデータファイル(hoge4_c
ます。 このデータセットの場合は、なぜかtechnical replicates
ので、列名変更はほぼ無意味です。

```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定してin_fに格納
out_f1 <- "hoge4_counts.txt"            #入力ファイル名を指定してout_f1に格納(カウントデータ)
out_f2 <- "srp001558_meta_samples.txt"  #入力ファイル名を指定してout_f2に格納(samplesメタデータ)
out_f3 <- "srp001558_meta_features.txt" #入力ファイル名を指定してout_f3に格納(featuresメタデータ)

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#入力フ
load(in
rse <-
rse

#本番(
rse <-
data <-
colname
dim(da
head(da

#ファイ
tmp <-
```

①

②

③

④

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	pro	sample	exp	run	rea	rea	pro	pai	sra	ma	au	sh	sh	bio	bio	bio	avg	ge	big	title	characteristics
2	SRF	SRS009313	SR	SRR032116	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 1 rep1	c("sex: female", "tissue: liver")
3	SRF	SRS009315	SR	SRR032118	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 2 rep1	c("sex: female", "tissue: liver")
4	SRF	SRS009316	SR	SRR032119	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 2 rep2	c("sex: female", "tissue: liver")
5	SRF	SRS009317	SR	SRR032120	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 3 rep1	c("sex: female", "tissue: liver")
6	SRF	SRS009318	SR	SRR032121	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 3 rep2	c("sex: female", "tissue: liver")
7	SRF	SRS009319	SR	SRR032122	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 1 rep1	c("sex: male", "tissue: liver")
8	SRF	SRS009320	SR	SRR032123	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 1 rep2	c("sex: male", "tissue: liver")
9	SRF	SRS009321	SR	SRR032124	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 2 rep1	c("sex: male", "tissue: liver")
10	SRF	SRS009322	SR	SRR032125	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 2 rep2	c("sex: male", "tissue: liver")
11	SRF	SRS009323	SR	SRR032126	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 3 rep1	c("sex: male", "tissue: liver")
12	SRF	SRS009324	SR	SRR032127	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 3 rep2	c("sex: male", "tissue: liver")

colData(rse)$sample

例えば、①colData(rse)$sampleの実行結果は、確かに②列の情報と同じです。これは、例題4のカウントデータファイル③hoge4_counts.txtの列名として使われています。

4. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題3とベースとして、さらにサンプルのメタデータ情報ファイル(srp001558_meta_samples.txt)と、遺伝子(features)のメタデータ情報ファイル(srp001558_meta_features.txt)も出力するやり方です。 58,037 genes×11 samplesからなるカウントデータファイル(hoge4_counts.txt)は列名をSRR...からSRS...に変更しています。 このデータセットの場合は、なぜかtechnical replicatesのサンプルに対して別々のSRS IDが付与されているので、列名変更はほぼ無意味です。

```
in_f <- "rse_gene.Rdata"                    #入力
out_f1 <- "hoge4_counts.txt"                #出力
out_f2 <- "srp001558_meta_samples.txt" #出力
out_f3 <- "srp001558_meta_features.txt"#出力

#必要なパッケージをロード
library(recount)                            #パッ

#入力フ
load(in
rse <-
rse

#本番(
rse <-
data <-
colname
dim(da
head(d

#ファイ
tmp <-
```

③

②

	A	B	C	D	E	F
1	pro	sample	exp	run	rea	rea
2	SRF	SRS009313	SR)	SRR032116	##	##
3	SRF	SRS009315	SR)	SRR032118	##	##
4	SRF	SRS009316	SR)	SRR032119	##	##
5	SRF	SRS009317	SR)	SRR032120	##	##
6	SRF	SRS009318	SR)	SRR032121	##	##
7	SRF	SRS009319	SR)	SRR032122	##	##
8	SRF	SRS009320	SR)	SRR032123	##	#
9	SRF	SRS009321	SR)	SRR032124	##	##
10	SRF	SRS009322	SR)	SRR032125	##	##
11	SRF	SRS009323	SR)	SRR032126	##	##
12	SRF	SRS009324	SR)	SRR032127	##	##

①

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
SRR032121 sex: female,tissue: liver
SRR032122   sex: male,tissue: liver
SRR032123   sex: male,tissue: liver
SRR032124   sex: male,tissue: liver
SRR032125   sex: male,tissue: liver
SRR032126   sex: male,tissue: liver
SRR032127   sex: male,tissue: liver
> dim(colData(rse))
[1] 11 21
> colData(rse)$sample
 [1] "SRS009313" "SRS009315" "SRS009316" "SRS009317"
 [5] "SRS009318" "SRS009319" "SRS009320" "SRS009321"
 [9] "SRS009322" "SRS009323" "SRS009324"
>
```

参考

カウントデータの列名

①の情報で、行列dataの列名に相当するcolnames(data)を置換しているので…

4. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題3とベースとして、さらにサンプルのメタデータ情報ファイル(srp001558_meta_samples.txt)と、 遺伝子(features)のメタデータ情報ファイル(srp001558_meta_features.txt)も出力するやり方です。 58,037 genes×11 samplesからなるカウントデータファイル(hoge4_counts.txt)は列名をSRR…からSRS…に変更しています。 このデータセットの場合は、なぜかtechnical replicatesのサンプルに対して別々のSRS IDが付与されているので、列名変更はほぼ無意味です。

```
in_f <- "rse_gene.Rdata"                    #入力
out_f1 <- "hoge4_counts.txt"                #出力
out_f2 <- "srp001558_meta_samples.txt" #出力
out_f3 <- "srp001558_meta_features.txt"#出力

#必要なパッケージをロード
library(recount)                            #パッ

#入力ファイルの読み込み
load(in_f)                                  #in_
rse <- rse_gene                             #rse
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #sca
data <- assays(rse)$counts                  #カウ
colnames(data) <- colData(rse)$sample       #列名
dim(data)                                   #行数
head(data)                                  #確認

#ファイルに保存(カウントデータ)
tmp <- cbind(rownames(data), data)          #保
```

①

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
SRR032121 sex: female,tissue: liver
SRR032122   sex: male,tissue: liver
SRR032123   sex: male,tissue: liver
SRR032124   sex: male,tissue: liver
SRR032125   sex: male,tissue: liver
SRR032126   sex: male,tissue: liver
SRR032127   sex: male,tissue: liver
> dim(colData(rse))
[1] 11 21
> colData(rse)$sample
 [1] "SRS009313" "SRS009315" "SRS009316" "SRS009317"
 [5] "SRS009318" "SRS009319" "SRS009320" "SRS009321"
 [9] "SRS009322" "SRS009323" "SRS009324"
>
```

①

参考

カウントデータの列名

①の情報で、行列dataの列名に相当するcolnames(data)を置換しているので…②列名変更後の行列dataの最初の2行分で見えているような状態に、③がなっています。

4. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題3とベースとして、さらにサンプルのメタデータ情報ファイル(srp001558_meta_samples.txt)と、遺伝子(features)のメタデータ情報ファイル(srp001558_meta_features.txt)も出力するやり方です。 58,037 genes×11 samplesからなるカウントデータファイル(hoge4_counts.txt)は列名をSRR…からSRS…に変更しています。 このデータセットの場合は、なぜかtechnical replicatesのサンプルに対して別々のSRS IDが付与されているので、列名変更はほぼ無意味です。

```
in_f <- "rse_gene.Rdata"                          #入力
out_f1 <- "hoge4_counts.txt"                      #出力
out_f2 <- "srp001558_meta_samples.txt" #出力
out_f3 <- "srp001558_meta_features.txt"#出力

#必要なパッケージをロード
library(recount)                                  #パッ

#入力ファイルの読み込み
load(in_f)                                        #i
rse <- rse_gene                                   #rse
rse                                               #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                          #sca
data <- assays(rse)$counts                        #カウ
colnames(data) <- colData(rse)$sample             #列名
dim(data)                                         #行数
head(data)                                        #確認

#ファイルに保存(カウントデータ)
tmp <- cbind(rownames(data), data)                #保存
```

①

②

③

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> head(data, n=2)
                    SRS009313 SRS009315 SRS009316
ENSG00000000003.14       1724      1561      1440
ENSG00000000005.5           0         0         0
                    SRS009317 SRS009318 SRS009319
ENSG00000000003.14        616       654       851
ENSG00000000005.5           0         0         0
                    SRS009320 SRS009321 SRS009322
ENSG00000000003.14        675      1498      1351
ENSG00000000005.5           0         6         0
                    SRS009323 SRS009324
ENSG00000000003.14       1350      1220
ENSG00000000005.5           0         0
>
```

Tips

①の列をみればわかるが、例えば②SRS009315とSRS009316は、③Human female 2（HSF2）に相当する同一サンプル、つまりtechnical replicatesである。

4. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題3とベースとして、さらにサンプルのメタデータ情報ファイル(srp001558_meta_samples.txt)と、遺伝子(features)のメタデータ情報ファイル(srp001558_meta_features.txt)も出力するやり方です。 58,037 genes×11 samplesからなるカウントデータファイル(hoge4_counts.txt)は列名をSRR...からSRS...に変更しています。 このデータセットの場合は、なぜかtechnical replicatesのサンプルに対して別々のSRS IDが付与されているので、列名変更はほぼ無意味です。

```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定してin_fに格納
out_f1 <- "hoge4_counts.txt"            #出力ファイル名を指定してout_f1に格納(カウントデータ)
out_f2 <- "srp001558_meta_samples.txt" #出力ファイル名を指定してout_f2に格納(samplesメタデータ)
out_f3 <- "srp001558_meta_features.txt"#出力ファイル名を指定してout_f3に格納(featuresメタデータ)

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#入力フ
load(in
rse <-
rse

#本番(
rse <-
data <-
colname
dim(dat
head(da

#ファイ
tmp <-
```

①

①

②

③

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	pro	sample	exp	run	rea	rea	pro	pai	sra	ma	au	sh	sh	bio	bio	bio	avg	ge	big	title	characteristics
2	SRF	SRS009313	SR	SRR032116	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 1 rep1	c("sex: female", "tissue: liver")
3	SRF	SRS009315	S	SRR032118	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 2 rep1	c("sex: female", "tissue: liver")
4	SRF	SRS009316	S	SRR032119	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 2 rep2	c("sex: female", "tissue: liver")
5	SRF	SRS009317	SR	SRR032120	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 3 rep1	c("sex: female", "tissue: liver")
6	SRF	SRS009318	SR	SRR032121	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human female 3 rep2	c("sex: female", "tissue: liver")
7	SRF	SRS009319	SR	SRR032122	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 1 rep1	c("sex: male", "tissue: liver")
8	SRF	SRS009320	SR	SRR032123	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 1 rep2	c("sex: male", "tissue: liver")
9	SRF	SRS009321	SR	SRR032124	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 2 rep1	c("sex: male", "tissue: liver")
10	SRF	SRS009322	SR	SRR032125	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 2 rep2	c("sex: male", "tissue: liver")
11	SRF	SRS009323	SR	SRR032126	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 3 rep1	c("sex: male", "tissue: liver")
12	SRF	SRS009324	SR	SRR032127	##	##	1	##	##	##	##	live	cor	20	20	20	35	GS	SR	Human male 3 rep2	c("sex: male", "tissue: liver")

colData(rse)$title

①colData(rse)$titleの情報が、サンプル間クラスタリングを行った際にわかりやすいと判断したので、それを行っているのが例題5。

4. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題3とベースとして、さらにサンプルのメタデータ情報ファイル(srp001558_meta_samples.txt)と、 遺伝子(features)のメタデータ情報ファイル(srp001558_meta_features.txt)も出力するやり方です。 58,037 genes×11 samplesからなるカウントデータファイル(hoge4_counts.txt)は列名をSRR...からSRS...に変更しています。 このデータセットの場合は、なぜかtechnical replicatesのサンプルに対して別々のSRS IDが付与されているので、列名変更はほぼ無意味です。

```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定してin_fに格納
out_f1 <- "hoge4_counts.txt"            #出力ファイル名を指定してout_f1に格納(カウントデータ)
out_f2 <- "srp001558_meta_samples.txt" #出力ファイル名を指定してout_f2に格納(samplesメタデータ)
out_f3 <- "srp001558_meta_features.txt"#出力ファイル名を指定してout_f3に格納(featuresメタデータ)

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#入力フ
load(in
rse <-
rse

#本番(
rse <-
data <-
colname
dim(da
head(d

#ファイ
tmp <-
```

①

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	pro	sample	exp	run	rea	rea	pro	pai	sra	ma	au	sha	sha	bio	bio	bio	avg	gec	big	title	characteristics
2	SRF	SRS009313	SR)	SRR032116	##	##	1	##	##	##	##	live	cor	20(	20(	20	35	GS	SR	Human female 1 rep1	c("sex: female", "tissue: liver")
3	SRF	SRS009315	SR)	SRR032118	##	##	1	##	##	##	##	live	cor	20(	20(	20	35	GS	SR	Human female 2 rep1	c("sex: female", "tissue: liver")
4	SRF	SRS009316	SR)	SRR032119	##	##	1	##	##	##	##	live	cor	20(	20(	20	35	GS	SR	Human female 2 rep2	c("sex: female", "tissue: liver")
5	SRF	SRS009317	SR)	SRR032120	##	##	1	##	##	##	##	live	cor	20(	20(	20	35	GS	SR	Human female 3 rep1	c("sex: female", "tissue: liver")
6	SRF	SRS009318	SR)	SRR032121	##	##	1	##	##	##	##	live	cor	20(	20(	20	35	GS	SR	Human female 3 rep2	c("sex: female", "tissue: liver")
7	SRF	SRS009319	SR)	SRR032122	##	##	1	##	##	##	##	live	cor	20(	20(	20	35	GS	SR	Human male 1 rep1	c("sex: male", "tissue: liver")
8	SRF	SRS009320	SR)	SRR032123	##	##	1	##	##	##	##	live	cor	20(	20(	20	35	GS	SR	Human male 1 rep2	c("sex: male", "tissue: liver")
9	SRF	SRS009321	SR)	SRR032124	##	##	1	##	##	##	##	live	cor	20(	20(	20	35	GS	SR	Human male 2 rep1	c("sex: male", "tissue: liver")
10	SRF	SRS009322	SR)	SRR032125	##	##	1	##	##	##	##	live	cor	20(	20(	20	35	GS	SR	Human male 2 rep2	c("sex: male", "tissue: liver")
11	SRF	SRS009323	SR)	SRR032126	##	##	1	##	##	##	##	live	cor	20(	20(	20	35	GS	SR	Human male 3 rep1	c("sex: male", "tissue: liver")
12	SRF	SRS009324	SR)	SRR032127	##	##	1	##	##	##	##	live	cor	20(	20(	20	35	GS	SR	Human male 3 rep2	c("sex: male", "tissue: liver")

Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、例題4
    - **例題5、例題6**、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

例題5

例題5では①のようにしてtitle列の情報を採用したが…②でも書いているように、いつでもここに有意義な情報があるとは限らないので注意。

5. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題4で得られたサンプルのメタデータ情報ファイル(hoge4_meta_samples.txt)中のtitle列に相当する情報で置き換えています。 これは、hoge4_meta_samples.txtをExcelで眺めたときに、たまたまtitle列情報がdiscriminable(容易に識別可能である)だと主観的に判断したためです。 このあたりの情報のクオリティというかどのような情報が提供されているかは、submitter依存です。したがって、一筋縄ではいきません。まるで有益な情報のない残念なものも結構あるからです。 58,037 genes×11 samplesからなる出力ファイルはhoge5.txtです。

②

```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.txt"                    #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#入力ファイルの読み込み(.Rdata)
load(in_f)                              #in_fで指定した.Rdataをロード
rse <- rse_gene                         #rseとして取り扱う
rse                                     #確認してるだけです

#本番(カウントデータ取得)
rse <- scale_counts(rse)                #scale_counts実行(2018.08.07追加)
data <- assays(rse)$counts              #カウントデータ行列を取得してdataに格納
dim(data)                               #行数と列数を表示
head(data)                              #確認してるだけです

#後処理(列名を変更)
colnames(data)<- colData(rse)$title     #列名を変更
head(data)                              #確認してるだけです

#ファイルに保存(カウントデータ)
```

①

例題5をコピペ実行

例題5をコピペ実行して、①のような列名になった、②hoge5.txtを得ておきましょう。

5. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題4で得られたサンプルのメタデータ情報ファイル(hoge4_meta_samples.txt)中のtitle列に相当する情報で置き換えています。 これは、hoge4_meta_samples.txtをExcelで眺めたときに、たまたまtitle列情報がdiscriminable(容易に識別可能である)だと主観的に判断したためです。 このあたりの情報のクオリティというかどのような情報が提供されているかは、submitter依存です。したがって、一筋縄ではいきません。まるで有益な情報のない残念なものも結構あるからです。 58,037 genes×11 samplesからなる出力ファイルはhoge5.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力
out_f <- "hoge5.txt"                        #出力

#必要なパッケージをロード
library(recount)                            #パッ

#入力ファイルの読み込み(.Rdata)
load(in_f)                                  #in_
rse <- rse_gene                             #rse
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #sca
data <- assays(rse)$counts                  #カウ
dim(data)                                   #行数
head(data)                                  #確認

#後処理(列名を変更)
colnames(data)<- colData(rse)$title         #列名
head(data)                                  #確認

#ファイルに保存(カウントデータ)
```

②

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
ENSG00000000460.16                    53                    142
ENSG00000000938.12                   328                    213
                    Human male 3 rep1 Human male 3 rep2
ENSG00000000003.14                  1350                   1220
ENSG00000000005.5                      0                      0
ENSG00000000419.12                   223                    238
ENSG00000000457.13                   435                    405
ENSG00000000460.16                   198                    126
ENSG00000000938.12                   367                    445
>
> #ファイルに保存(カウントデータ)
> tmp <- cbind(rownames(data), data)      #保存したい情$
> write.table(tmp, out_f, sep="\t", append=F, quote=F,$
>
```

①

例題6

①例題6は、②technical replicates(同一個体の反復データ)をマージして、③58,037 genes × 6 samplesのカウントデータ行列にするコード。コピペ実行。

6. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題5の続きのようなものですが、technical replicatesのデータをマージした結果を出力しています。 例えば、"Human female 2 rep1"列と"Human female 2 rep2"列のカウント数の和をとり、列名を"HSF2"のようにしています。 この列名の表記法は、「サンプルデータ42の20,689 genes×18 samplesのリアルカウントデータ(sample_blekhman_18.txt)」中のヒトサンプル名と同じにしています。 58,037 genes×6 samplesからなる出力ファイルはsrp001558_count_hoge6.txtです。

```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定してin_fに格納
out_f <- "srp001558_count_hoge6.txt"    #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#入力ファイルの読み込み(.Rdata)
load(in_f)                              #in_fで指定した.Rdataをロード
rse <- rse_gene                         #rseとして取り扱う
rse                                     #確認してるだけです

#本番(カウントデータ取得)
rse <- scale_counts(rse)                #scale_counts実行(2018.08.07追加)
uge <- assays(rse)$counts               #カウントデータ行列を取得してugeに格納
dim(uge)                                #行数と列数を表示
head(uge)                               #確認してるだけです

#後処理(technical replicatesの列をマージ)
uge <- as.data.frame(uge)               #行列形式からデータフレーム形式に変更
data <- cbind(                          #必要な列名の情報を取得したい列の順番で結合した結果をdata
  uge$SRR032116,                        #HSF1
  uge$SRR032118 + uge$SRR032119,        #HSF2
```

例題6

例題6をコピペ実行して、①のような列名になった、②出力ファイルを得ておきましょう。

6. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題5の続きのようなものですが、technical replicatesのデータをマージした結果を出力しています。 例えば、"Human female 2 rep1"列と"Human female 2 rep2"列のカウント数の和をとり、列名を"HSF2"のようにしています。 この列名の表記法は、「サンプルデータ42の20,689 genes×18 samplesのリアルカウントデータ(sample_blekhman_18.txt)」中のヒトサンプル名と同じにしています。 58,037 genes×6 samplesからなる出力ファイルはsrp001558_count_hoge6.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力
out_f <- "srp001558_count_hoge6.txt"        #出力

#必要なパッケージをロード
library(recount)                            #パッ

#入力ファイルの読み込み(.Rdata)
load(in_f)                                  #in_
rse <- rse_gene                             #rse
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #sca
uge <- assays(rse)$counts                   #カウ
dim(uge)                                    #行数
head(uge)                                   #確認

#後処理(technical replicatesの列をマージ)
uge <- as.data.frame(uge)                   #行列
data <- cbind(                              #必要
  uge$SRR032116,                            #HSF
  uge$SRR032118 + uge$SRR032119,            #HSF
```

② 出力

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
[1] 58037     6
> head(data)                                        #確認してるだ
                    HSF1 HSF2 HSF3 HSM1 HSM2 HSM3
ENSG00000000003.14 1724 3001 1270 1526 2849 2570
ENSG00000000005.5     0    0    0    0    6    0
ENSG00000000419.12  337  611  493  523  651  461
ENSG00000000457.13  414  657  666 1388  620  840
ENSG00000000460.16  114  371  266  450  195  324
ENSG00000000938.12  362  901 2476 1191  541  812
>
> #ファイルに保存(カウントデータ)
> tmp <- cbind(rownames(data), data)      #保存したい情$
> write.table(tmp, out_f, sep="\t", append=F, quote=F,$
>
```

① 確認

遺言

利用したいRパッケージのマニュアルでは、サンプルデータの列名変更などは最初のほうに説明されている。本来①のあたりは本質的なところではない。しかしながら、慣れないと非常に難解であり、しかもマニュアル中の説明はそれほど丁寧ではない。それゆえ、今回詳述したようなcolData(rse)を自分で眺めてうまく対処するノリに慣れるのが重要です！

5. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題4で得られたサンプルのメタデータ情報ファイル(hoge4_meta_sam
き換えています。 これは、hoge4_meta_samples.txtをExcelで眺めた
discriminable(容易に識別可能である)だと主観的に判断したためです。
のような情報が提供されているかは、submitter依存です。したがって、
報のない残念なものも結構あるからです。 58,037 genes×11 samples

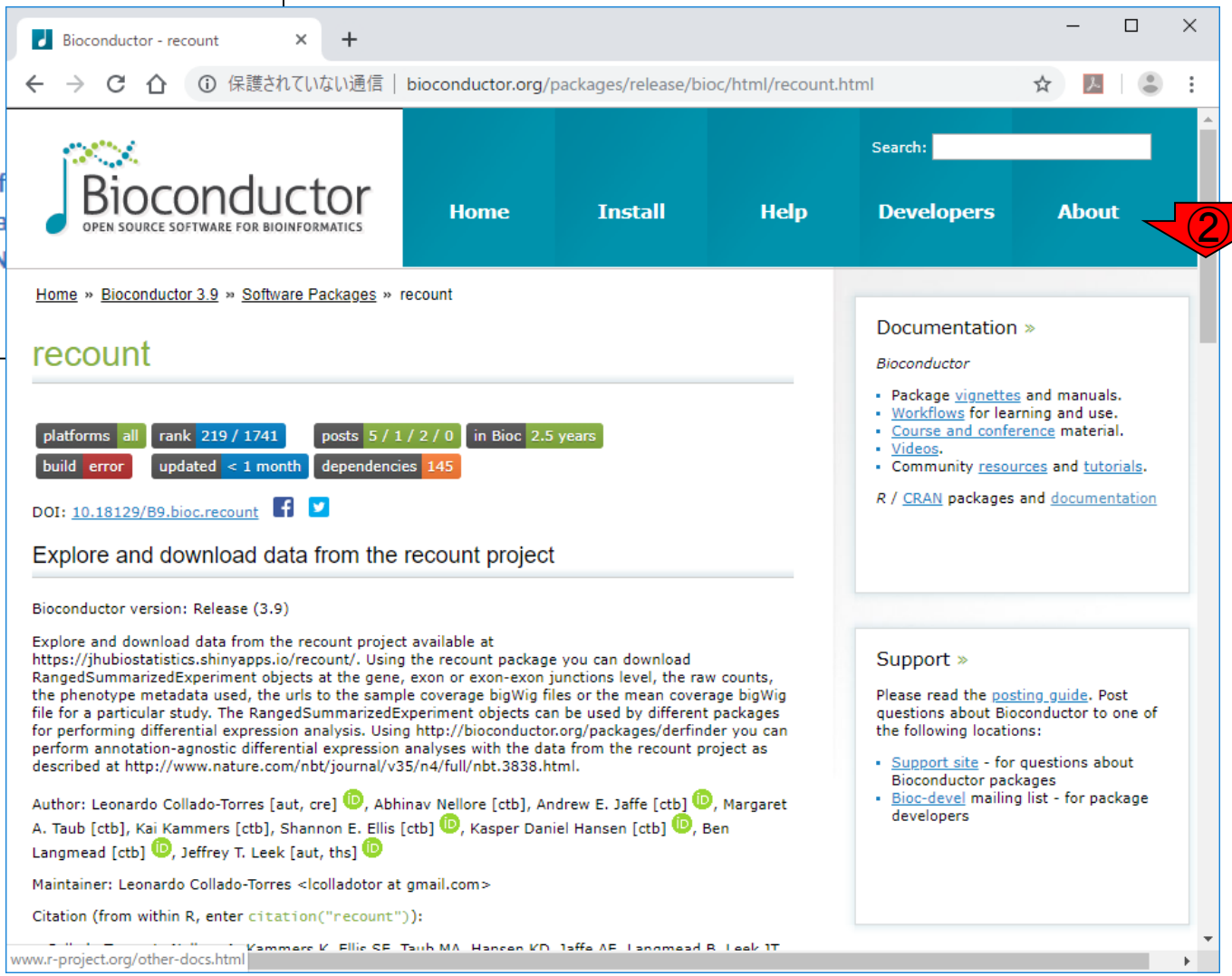
```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.txt"                    #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#入力ファイルの読み込み(.Rdata)
load(in_f)                              #in_fで指定した.Rdataをロード
rse <- rse_gene                         #rseとして取り扱う
rse                                     #確認してるだけです

#本番(カウントデータ取得)
rse <- scale_counts(rse)                #scale_counts実行(2018.08.07追加)
data <- assays(rse)$counts              #カウントデータ行列を取得してdataに格納
dim(data)                               #行数と列数を表示
head(data)                              #確認してるだけです

#後処理(列名を変更)
colnames(data)<- colData(rse)$title     #列名を変更
head(data)                              #確認してるだけです

#ファイルに保存(カウントデータ)
```

①

Rパッケージrecount

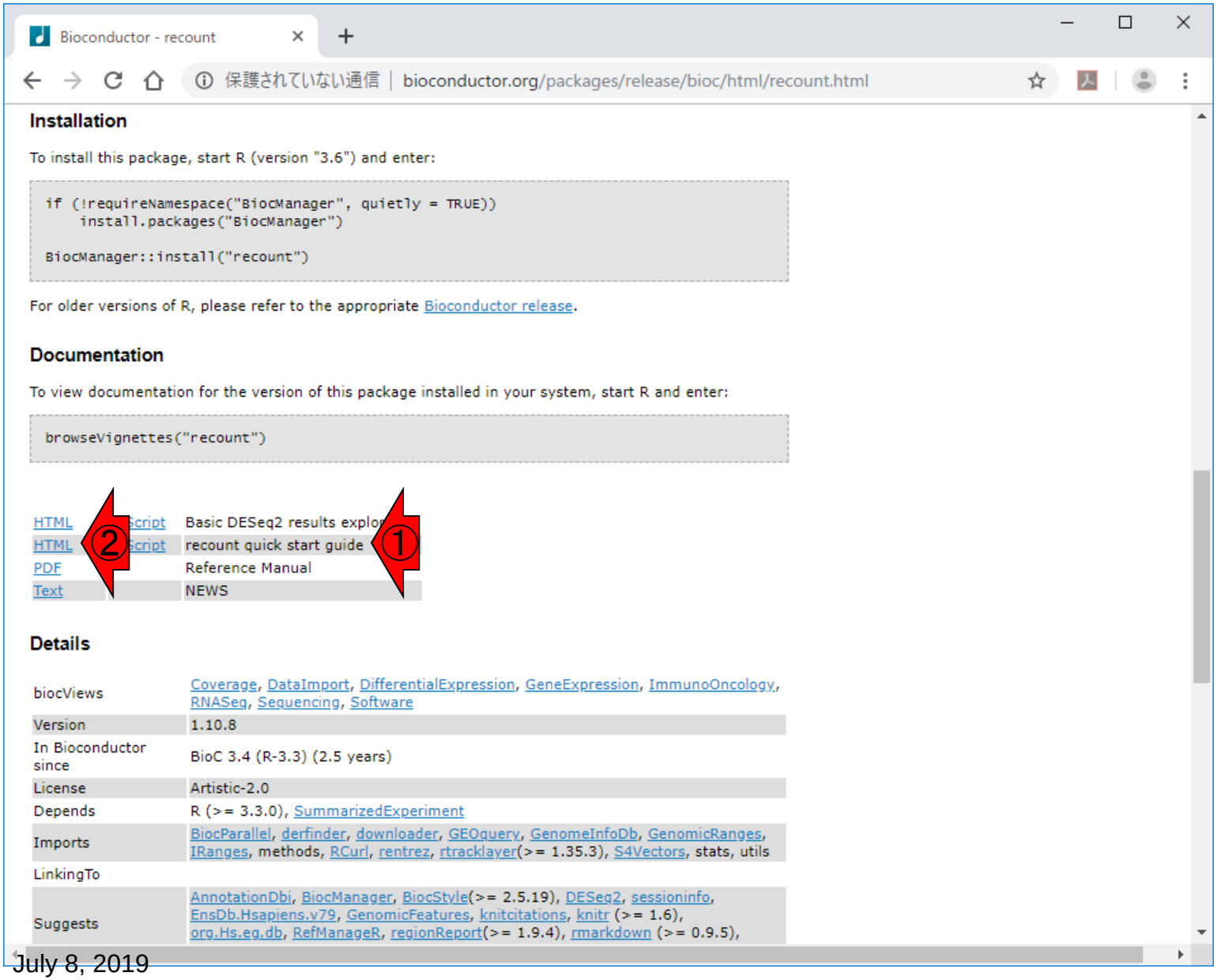
パッケージのマニュアルの読み解きが難解である例を示します。①(Rパッケージの)recount。②ページ下部に移動。

2. 2019年07月08日 (PC使用)
講義資料PDF(最終更新：2019.07.01)
(Rで)塩基配列解析
Blekhman et al., Genome Res., 2010
TCC : Sun et al., BMC Bioinformatics, 2013
Tang et al., BMC Bioinformatics, 2015
Zhao et al., Biol. Proc. Online, 2018
ReCount(website) : Frazee et al., BMC Bioinf
recount2(website) : Collado-Torres et al., Na
recount(R package) : Collado-Torres et al., N
①
rse_gene.Rdata(SRP001558)
rse_gene.Rdata(ERP000546)



①recount quick start guideの②HTML。

Rパッケージrecount



Bioconductor - recount

保護されていない通信 | bioconductor.org/packages/release/bioc/html/recount.html

Installation

To install this package, start R (version "3.6") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("recount")
```

For older versions of R, please refer to the appropriate Bioconductor release.

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("recount")
```

HTML	Script	Basic DESeq2 results explor
HTML	Script	recount quick start guide
PDF		Reference Manual
Text		NEWS

②

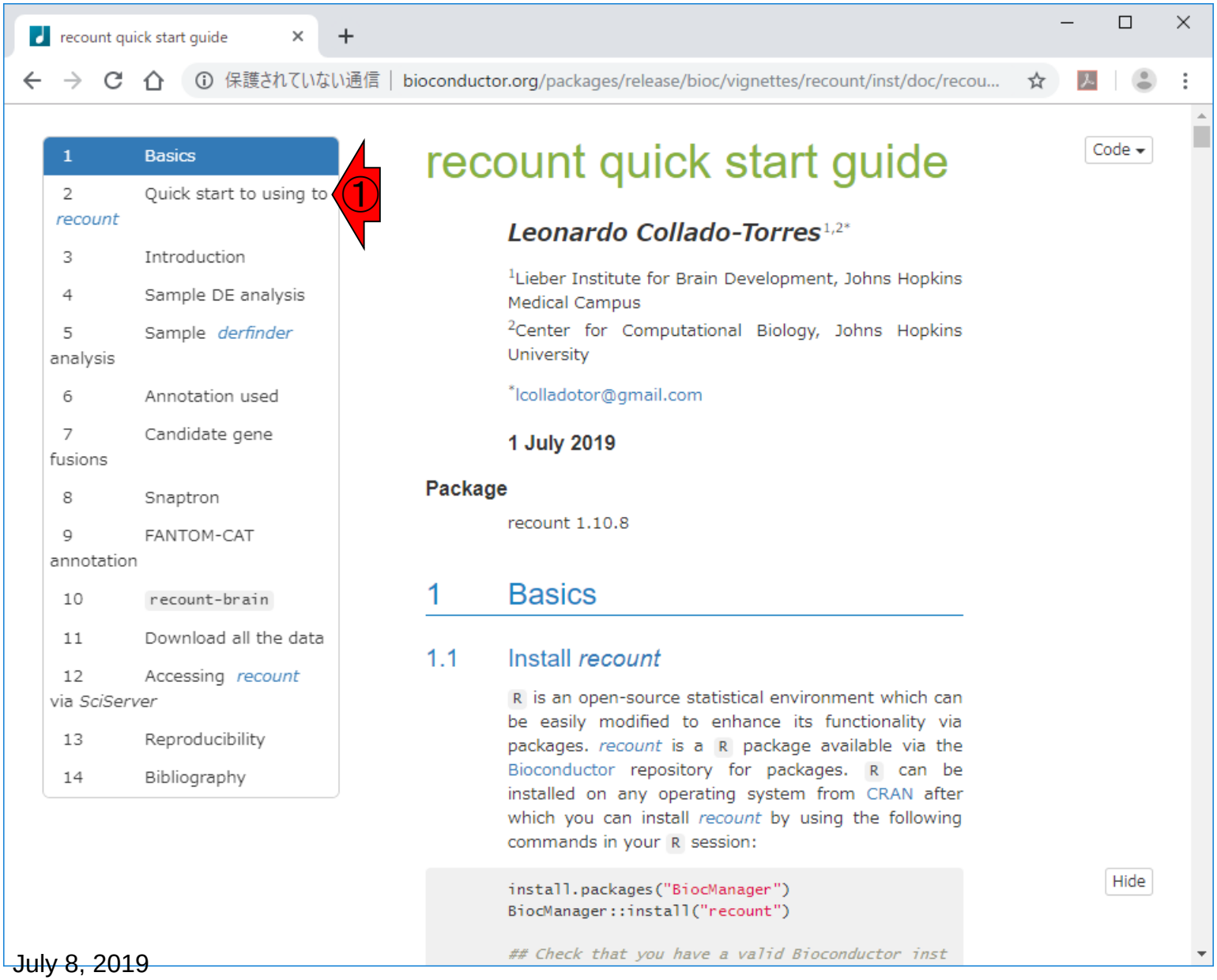
①

Details

biocViews	Coverage, DataImport, DifferentialExpression, GeneExpression, ImmunoOncology, RNASeq, Sequencing, Software
Version	1.10.8
In Bioconductor since	BioC 3.4 (R-3.3) (2.5 years)
License	Artistic-2.0
Depends	R (>= 3.3.0), SummarizedExperiment
Imports	BiocParallel, derfinder, downloader, GEOquery, GenomeInfoDb, GenomicRanges, IRanges, methods, RCurl, rentrez, rtracklayer(>= 1.35.3), S4Vectors, stats, utils
LinkingTo	
Suggests	AnnotationDbi, BiocManager, BiocStyle(>= 2.5.19), DESeq2, sessioninfo, EnsDb.Hsapiens.v79, GenomicFeatures, knitcitations, knitr (>= 1.6), org.Hs.eg.db, RefManageR, regionReport(>= 1.9.4), rmarkdown (>= 0.9.5),

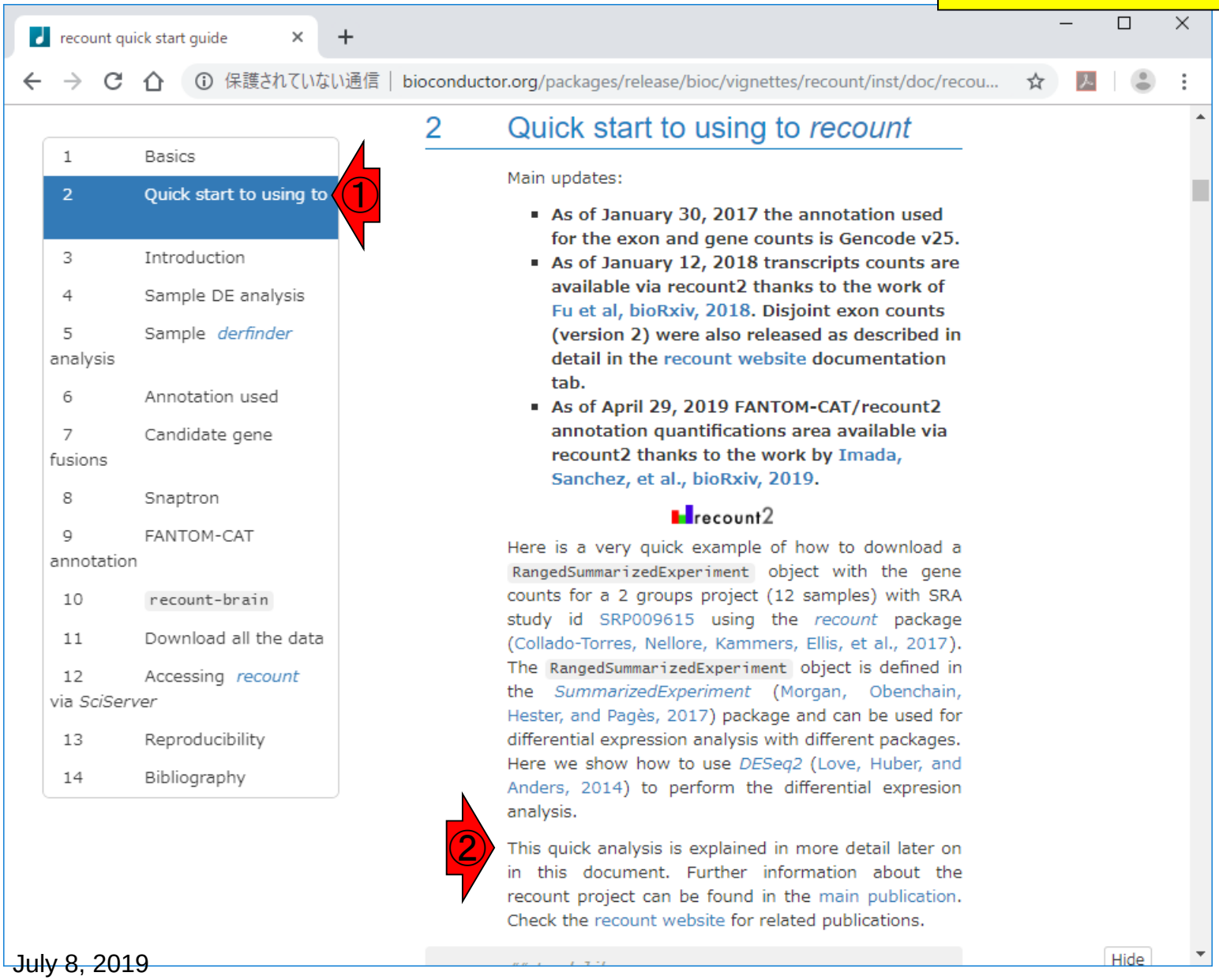
①2 Quick start to using to recount

Rパッケージrecount



Rパッケージrecount

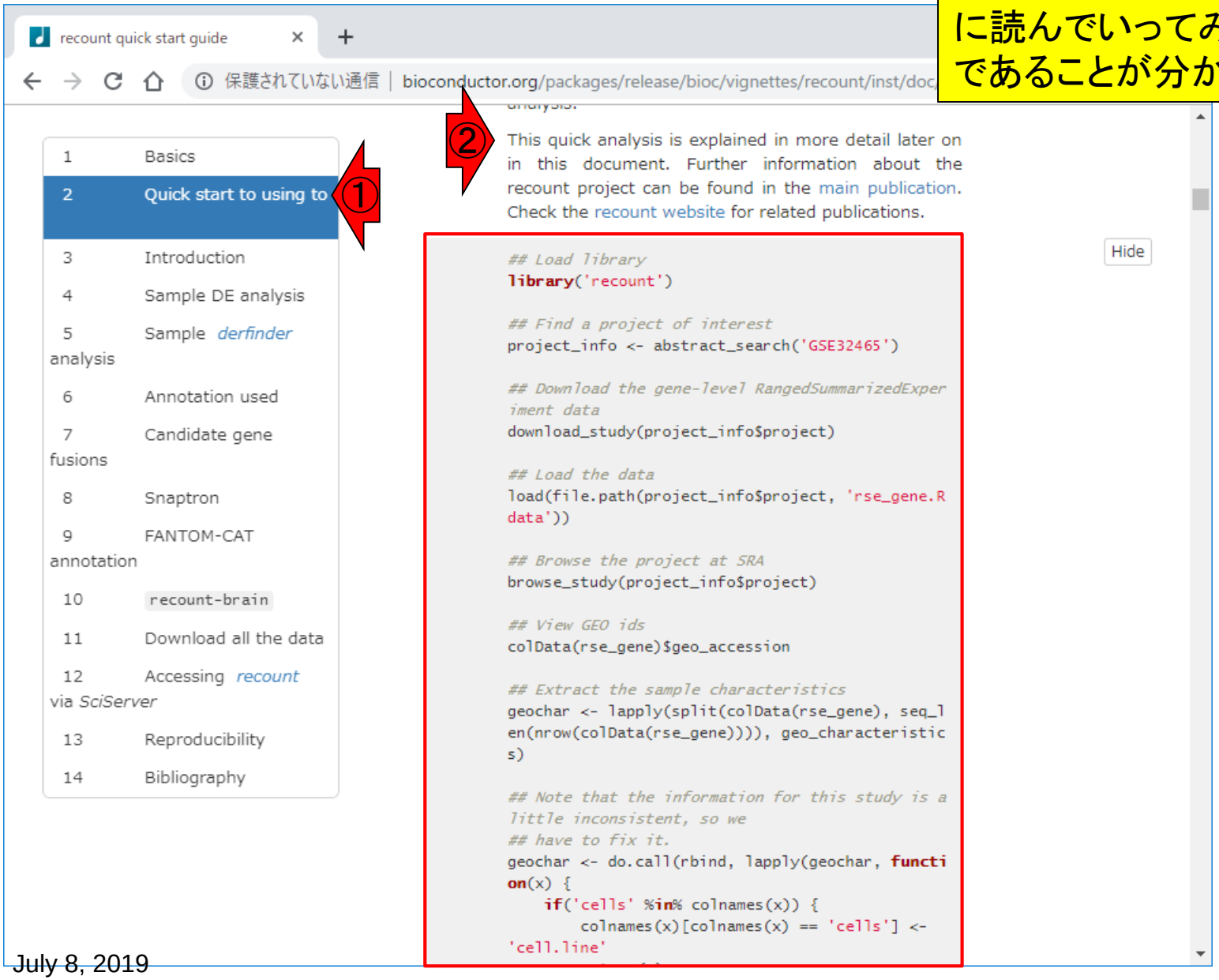
①2 Quick start to using to recountに移動したことがわかります。②がページ上部になるように移動。



参考

Rパッケージrecount

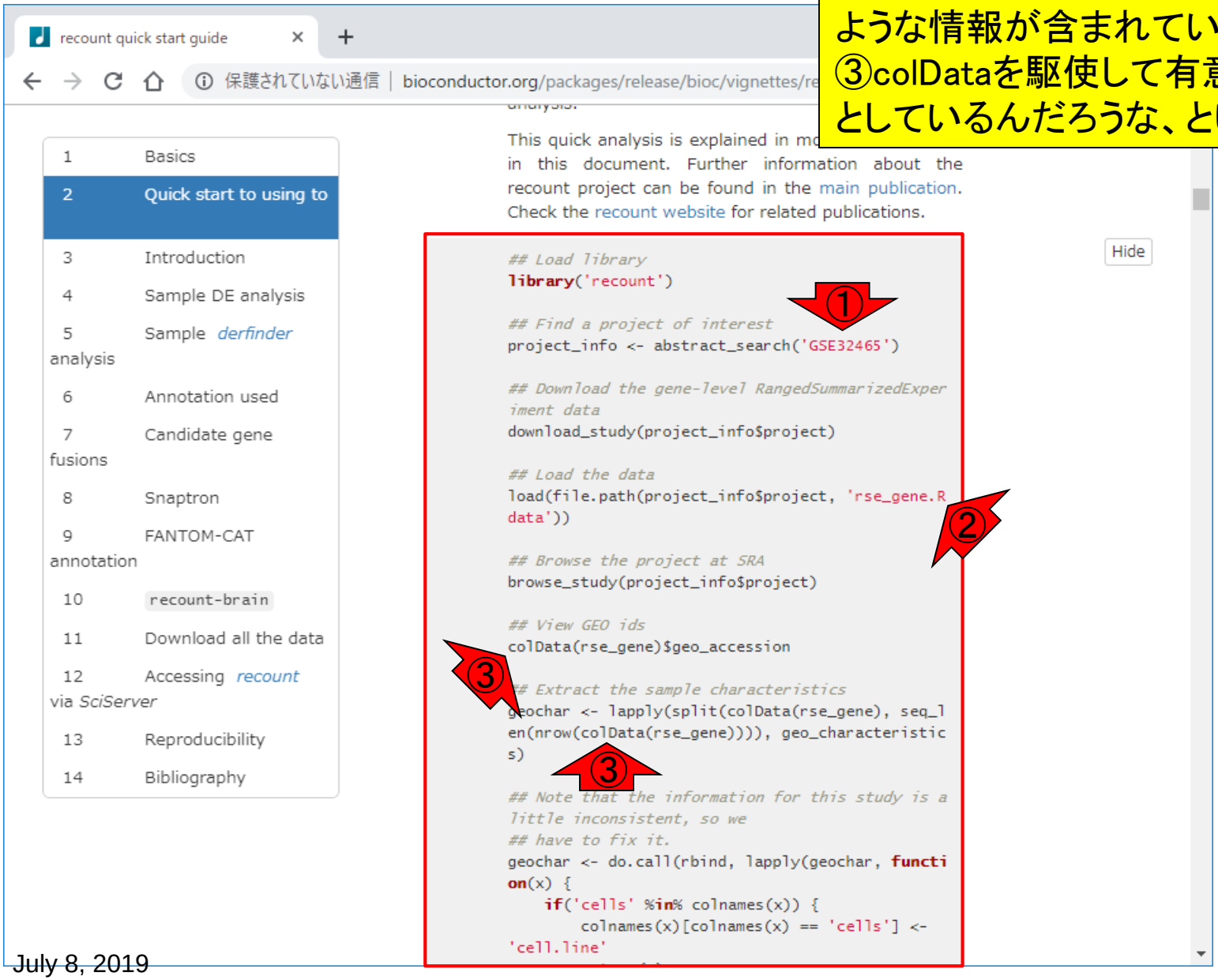
①2 Quick start to using to recountに移動したことがわかります。②がページ上部になるように移動。こんな感じ。赤枠を順に読んでいってみてください。非常に難解であることが分かります。



参考

Rパッケージrecount

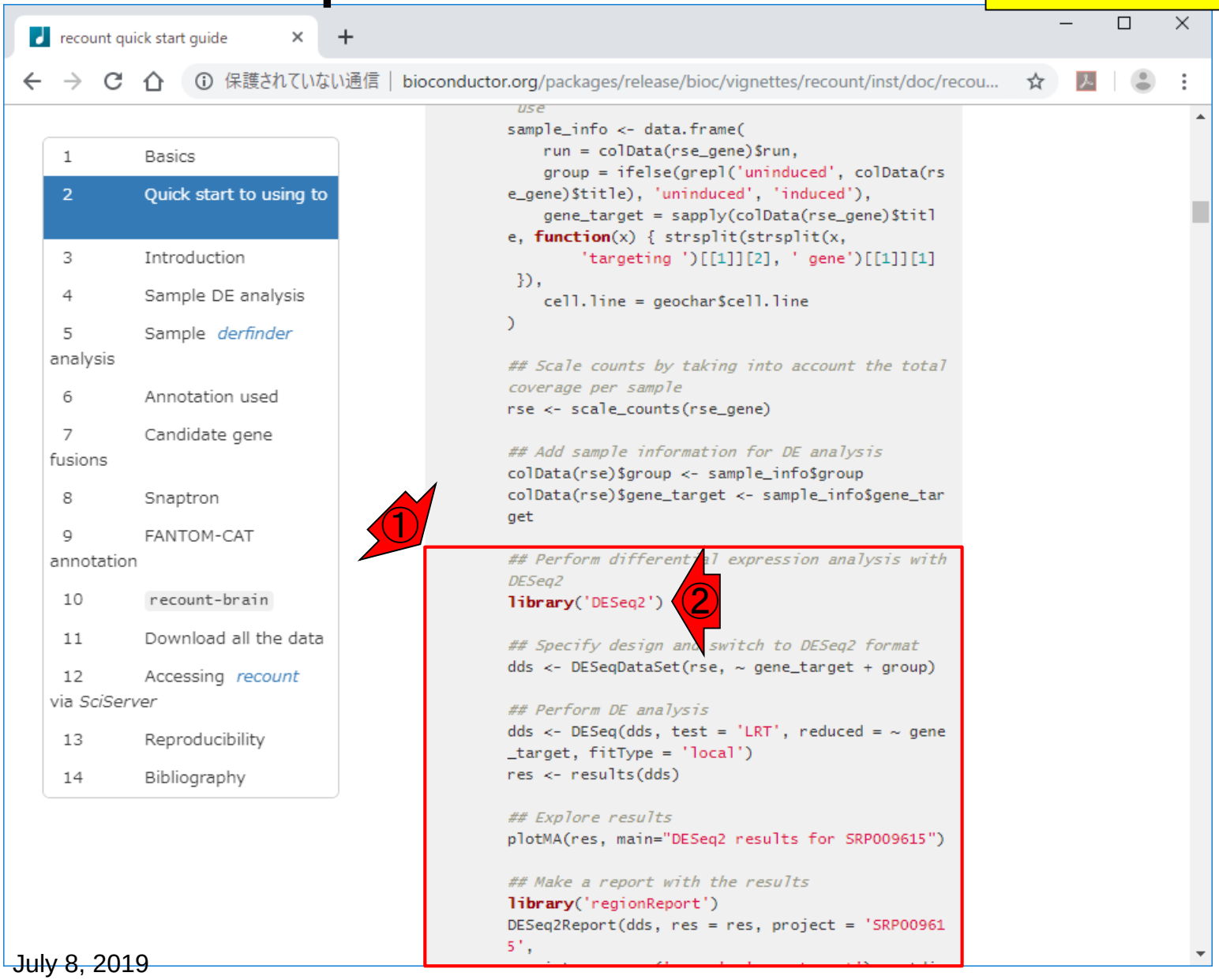
落ち着いてよく眺めると、①GSE32465というIDの、②rse_gene.Rdataを取得しているんだろうな、というのはわかります。また、colData実行結果にどのような情報が含まれているかがわかっていれば、③colDataを駆使して有意義な列名情報を得ようとしているんだろうな、という程度はわかります。



参考

DESeq2との連結

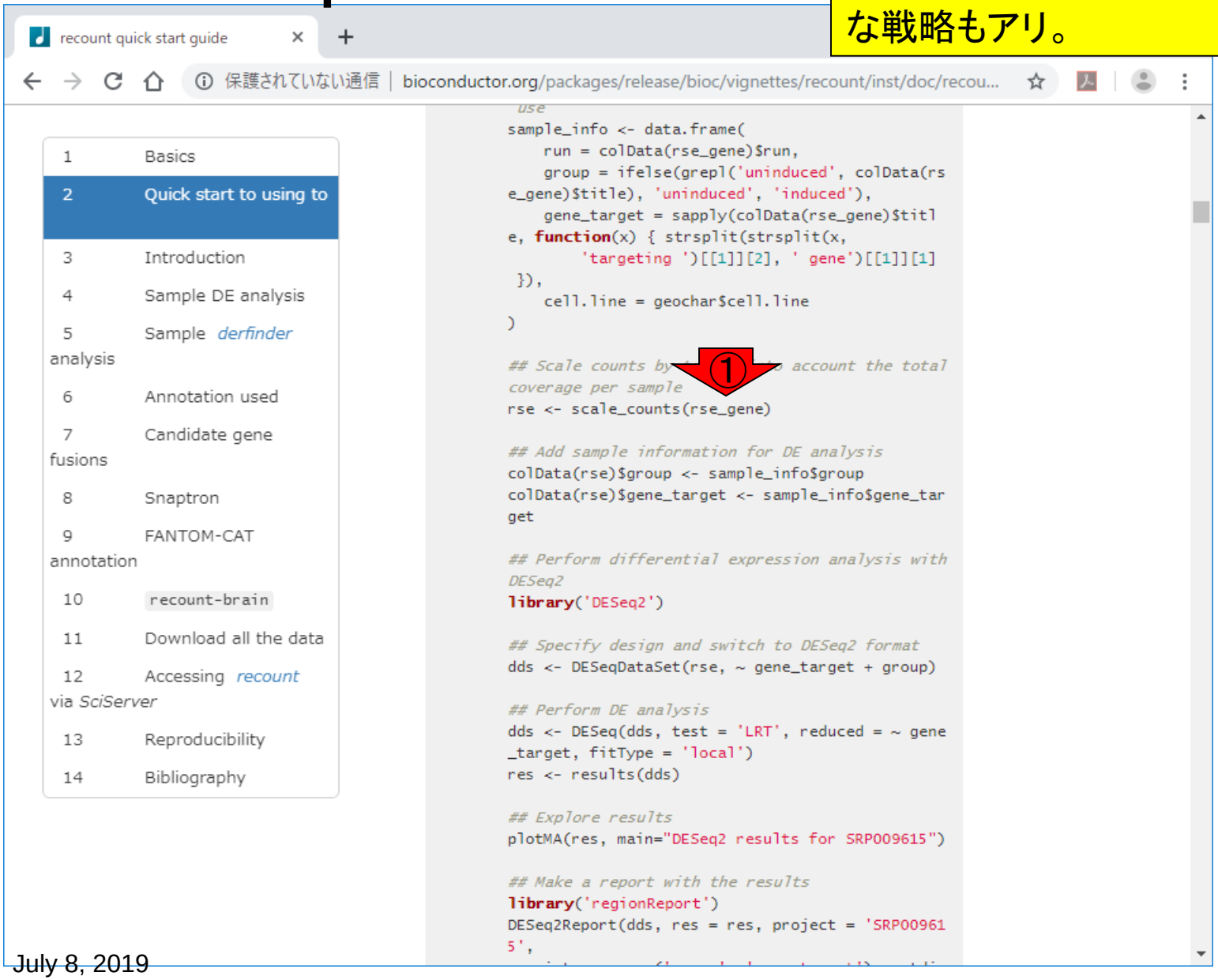
半ページ分ほど下部に移動。①が②
DESeq2という有名なRパッケージを利用した発現変動解析部分。



参考

DESeq2との連結

まずは①rse_geneを得るところまでコピペ実行して、rse_geneの中身を様々な視点で眺め、これより上の行で一体何をやっているかを解読するような戦略もアリ。



参考

scale_counts

まずは①rse_geneを得るところまでコピペ実行して、rse_geneの中身を様々な視点で眺め、これより上の行で一体何をやっているかを解読するような戦略もアリ。また、②DESeq2への受け渡し前に③scale_countsを実行している点も見逃してはいけない。私も挙動を完全に掌握できているわけではないが、マニュアル中に明記されているので…

recount quick start guide

保護されていない通信 | bioconductor.org/packages/release/bioc/vignettes/rec


```
use
sample_info <- data.frame(
    run = colData(rse_gene)$run,
    group = ifelse(grepl('uninduced
e_gene)$title), 'uninduced', 'induc
    gene_target = sapply(colData(rs
e, function(x) { strsplit(strsplit(
        'targeting ')[[1]][2], ' gene')[[1]][1]
 }),
    cell.line = geochar$cell.line
)

## Scale counts by ... account the total
coverage per sample
rse <- scale_counts(rse_gene)

## Add ... rmation for DE analysis
colData(rse)$group <- sample_info$group
colData(rse)$gene_target <- sample_info$gene_tar
get

## Perform differential expression analysis with
DESeq2
library('DESeq2')

## Specify design and switch to DESeq2 format
dds <- DESeqDataSet(rse, ~ gene_target + group)

## Perform DE analysis
dds <- DESeq(dds, test = 'LRT', reduced = ~ gene
_target, fitType = 'local')
res <- results(dds)

## Explore results
plotMA(res, main="DESeq2 results for SRP009615")

## Make a report with the results
library('regionReport')
DESeq2Report(dds, res = res, project = 'SRP00961
5',
```

scale_counts

まずは①rse_geneを得るところまでコピペ実行して、rse_geneの中身を様々な視点で眺め、これより上の行で一体何をやっているかを解読するような戦略もアリ。また、②DESeq2への受け渡し前に③scale_countsを実行している点も見逃してはいけない。私も挙動を完全に掌握できているわけではないが、マニュアル中に明記されているので…とりあえず④の部分にこの関数を入れているのです。

6. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題5の続きのようなものですが、technical replicatesのデータをマージした
ば、"Human female 2 rep1"列と"Human female 2 rep2"列のカウント数の
しています。 この列名の表記法は、「サンプルデータ42の20,689 genes×1
(sample_blekhman_18.txt)」中のヒトサンプル名と同じにしています。 58
力ファイルはsrp001558_count_hoge6.txtです。

④

```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定して
out_f <- "srp001558_count_hoge6.txt"    #出力ファイル名を指定して

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#入力ファイルの読み込み(.Rdata)
load(in_f)                              #in_fで指定した.Rdataをロード
rse <- rse_gene                         #rseとして取り扱う
rse                                     #確認してるだけです

#本番(カウントデータ取得)
rse <- scale_counts(rse)                #scale_counts実行(2018.08.07追加)
uge <- assays(rse)$counts               #カウントデータ行列を取得してugeに格納
dim(uge)                                #行数と列数を表示
head(uge)                               #確認してるだけです

#後処理(technical replicatesの列をマージ)
uge <- as.data.frame(uge)               #行列形式からデータフレーム形式に変更
data <- cbind(                          #必要な列名の情報を取得したい列の順番で結合した結果をdata
  uge$SRR032116,                        #HSF1
  uge$SRR032118 + uge$SRR032119,        #HSF2
```

Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、例題4
    - 例題5、例題6、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

これまでコピペ実行したものは、①SRP001558の…

今手元にあるのは…

(Rで)塩基配列解析

(last modified 2019/07/01, since 2010)

このウェブページ(
Macintosh2018.1
ています。 初心者
2018年7月に(R
(2018/07/18)

What's new? (

- マップ後 | カウント情報取得 | トランスクリプトーム | BEDファイルから (last modified 2014/06/21)
- カウント情報取得 | リアルデータ | について (last modified 2019/05/16)
- カウント情報取得 | リアルデータ | SRP061240 | recount(Collado-Torres_2017) (last modified 2018/08/28)
- カウント情報取得 | リアルデータ | SRP056295 | recount(Collado-Torres_2017) (last modified 2018/08/29)
- カウント情報取得 | リアルデータ | SRP056146 | recount(Collado-Torres_2017) (last modified 2018/10/25)
- カウント情報取得 | リアルデータ | SRP035988 | recount(Collado-Torres_2017) (last modified 2018/08/25)
- カウント情報取得 | リアルデータ | SRP026126 | recount(Collado-Torres_2017) (last modified 2018/08/30)
- カウント情報取得 | リアルデータ | SRP018853 | recount(Collado-Torres_2017) (last modified 2018/08/26)
- カウント情報取得 | リアルデータ | SRP012167 | recount(Collado-Torres_2017) (last modified 2018/08/24)
- カウント情報取得 | リアルデータ | SRP012167 | parathyroidSE(Haglund_2012) (last modified 2018/08/19)
- カウント情報取得 | リアルデータ | SRP001558 | recount(Collado-Torres_2017) (last modified 2018/08/08)
- カウント情報取得 | リアルデータ | SRP001540 | recount(Collado-Torres_2017) (last modified 2018/08/10)
- カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann_2013) (last modified 2018/07/03)
- カウント情報取得 | リアルデータ | ERP000546 | recount(Collado-Torres_2017) (last modified 2018/08/07)
- カウント情報取得 | シミュレーションデータ | RNA-seq | について (last modified 2019/04/05)

①

今手元にあるの

これまでコピペ実行したものは、①SRP001558の…②例題5を実行して得られた58,037 genes × 11 samplesからなる③hoge5.txtと、④例題6を実行して得られた58,037 genes × 6 samplesからなる⑤srp001558_count_hoge6.txt。これらを入力としてサンプル間クラスタリングをやってみましょう。

カウント情報取得 | リアルデータ | SRP001558 |

recountパッケージを用いて、 SRP001558(Blekhman et al., Genome Res
を含むRangedSummarizedExperimentクラスオブジェクトという形式の.R
行列にした状態で保存するやり方を示します。 原著論文では、3生物種(ヒト12 samples、チンパンジー12 samples、そしてア
カゲザル12 samples)のカウントデータを取得しています。 ウェブサイトrecount2上でSRP001558で検索すると、 number of
samplesが12、speciesがhuma
ります。 例題2までで、なぜか1
SRP001558で検索し、 phenot
かります。私は、「SRR032117
のため、recount2のウェブページ上は
ろう。」と予想しました。 また、PRJ
recount2上では引っかかってきません
後のカウントデータとなるように変更
「ファイル」−「ディレクトリの変更」

1. geneレベルカウントデータ情報を

SRP001558という名前のフォルダが
ェクト名で取り扱えます。 ウェブサ
rse_gene.Rdataと同じです。

```
param_ID <- "SRP001558"

#必要なパッケージをロード
library(recount)

#本番(.Rdataをダウンロード)
download_study(param_ID, type
```

②

5. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題4で得られたサンプルのメタデータ情報ファイル(hoge4_meta_samples.txt)中のtitle列に相当する情報で置き換えています。 これは、hoge4_meta_samples.txtをExcelで眺めたときに、たまたまtitle列情報がdiscriminable(容易に識別可能である)だと主観的に判断したためです。 このあたりの情報のクオリティというかどのような情報が提供されているかは、submitter依存です。したがって、一筋縄ではいきません。まるで有益な情報のない残念なものも結構あるからです。 58,037 genes×11 samplesからなる出力ファイルはhoge5.txtです。

```
in_f <- "rse_gene.Rdat
out_f <- "hoge5.txt"

#必要なパッケージをロード
library(recount)

#入力ファイルの読み込み(.Rdata)
load(in_f)
rse <- rse_gene
rse

#本番(カウントデータ取得)
rse <- scale_counts(rse)
data <- assays(rse)$counts
dim(data)
head(data)

#後処理(列名を変更)
colnames(data)<- colData(rse)$t
head(data)

#ファイルに保存(カウントデータ)
```

③ ④

6. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題5の続きのようなものですが、technical replicatesのデータをマージした結果を出力しています。 例えば、"Human female 2 rep1"列と"Human female 2 rep2"列のカウント数の和をとり、列名を"HSF2"のようにしています。 この列名の表記法は、「サンプルデータ42の20,689 genes×18 samplesのリアルカウントデータ(sample_blekhman_18.txt)」中のヒトサンプル名と同じにしています。 58,037 genes×6 samplesからなる出力ファイルはsrp001558_count_hoge6.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "srp001558_count_hoge6.txt"        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッケージの読み込み

#入力ファイルの読み込み(.Rdata)
load(in_f)                                  #in_fで指定した.Rdataをロード
rse <- rse_gene                             #rseとして取り扱う
rse                                         #確認してるだけです

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_counts実行(2018.08.07追加)
uge <- assays(rse)$counts                   #カウントデータ行列を取得してugeに格納
dim(uge)                                    #行数と列数を表示
head(uge)                                   #確認してるだけです

#後処理(technical replicatesの列をマージ)
uge <- as.data.frame(uge)                   #行列形式からデータフレーム形式に変更
data <- cbind(                              #必要な列名の情報を取得したい列の順番で結合した結果をdata
  uge$SRR032116,                            #HSF1
  uge$SRR032118 + uge$SRR032119,            #HSF2
```

⑤

サンプル間クラスタリング

①の、②例題5をテンプレートとして利用して、サンプル間クラスタリングをやってみましょう。

- 解析 | 前処理 | scRNA-seq | について (last modified 2019/06/26) NEW
- 解析 | クラスタリング | RNA-seq | について (last modified 2019/04/04)
- 解析 | クラスタリング | RNA-seq | サンプル間 | hclust (last modified 2015/02/26)
- 解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013) (last modified 2018/08/06)
- 解析 | クラスタリング | RNA-seq | 遺伝子間(基礎) | MBCluster.Seq(Li_2014) (last modified 2018/09/23)
- 解析
- 解析
- 解析

①

解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013)

②

TCCパッケージを用い
クラスタリング結果を返し
2015)中でも この関数を用
(2015/11/15)。
「ファイル」－「ディレク

1. 59,857 genes×6 sar

Neyret-Kahn et al., Ger
RNA-seqカウントデータ
SRP017142(Neyret-Kah

```
in_f <- "srp017142_c
out_f <- "hoge1.png"
param_fig <- c(500,

#必要なパッケージをロ
library(TCC)

#入力ファイルの読み込
data <- read.table(i
dim(data)
```

5. 60,234 genes×6 samplesのリアルデータ(hoge9_count_gene.txt)の場合：

Neyret-Kahn et al., Genome Res., 2013のgene-levelの2群間比較用(3 proliferative samples vs. 3 Ras samples)ヒトRNA-seqカウントデータです。 マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015)から得られます。

```
in_f <- "hoge9_count_gene.txt"          #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.png"                    #出力ファイル名を指定してout_fに格納
param_fig <- c(500, 400)                #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの読
dim(data)                               #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman",#クラスタリング実行結果をoutに格納
            hclust.method="average", unique.pattern=TRUE)#クラスタリング実行結果をoutに格納

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイルの各種パラメータを指
par(mar=c(0, 4, 1, 0))                  #下、左、上、右の順で余白（行）を指定
plot(out, sub="", xlab="", cex.lab=1.2,#樹形図(デンドログラム)の表示
  cex=1.3, main="", ylab="Height")      #樹形図(デンドログラム)の表示
dev.off()                               #おまじない
```

サンプル間クラスタリング

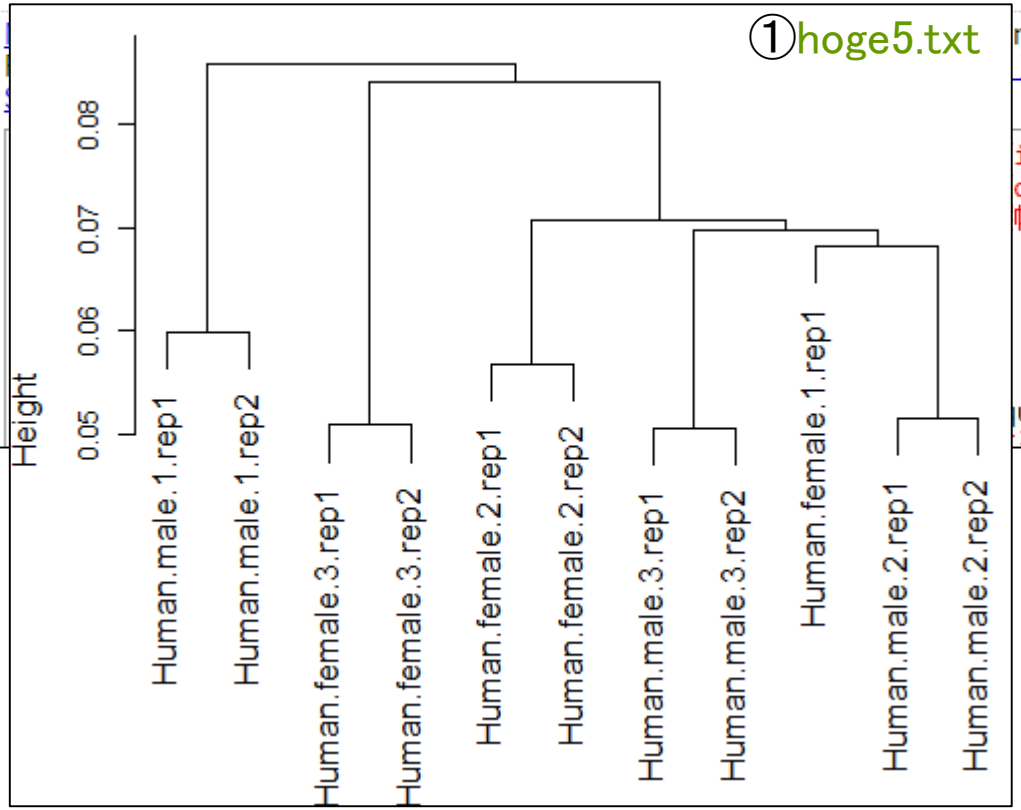
例題5をテンプレートとして、①hoge5.txt、②srp001558_count_hoge6.txtを実行した結果。

解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013)

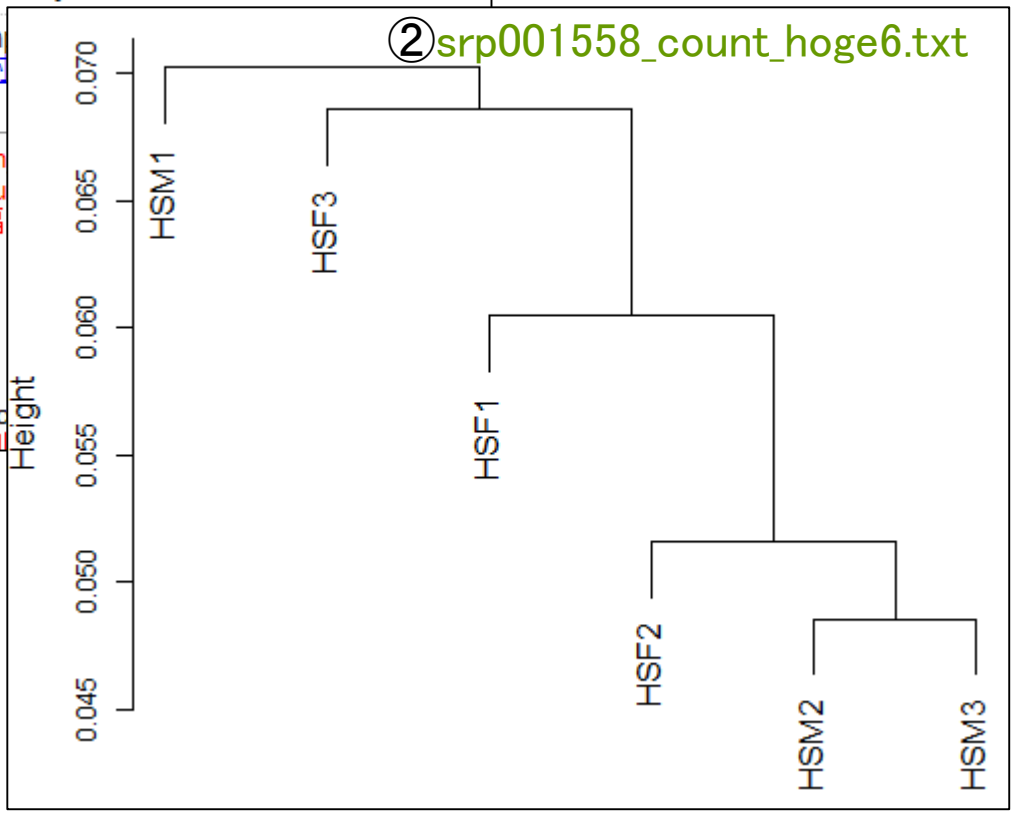
TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。 clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 (Tang et al., BMC Bioinformatics, 2015)中でも この関数を用いています(2015/11/05追加)。xlsx形式ファイルを入力とするやり方も追加しました(2015/11/15)。
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. 59,857 genes×6 samplesのリアルデータ(srp017142_count_bowtie.txt)の場合：

①hoge5.txt



②srp001558_count_hoge6.txt



サンプル間クラ

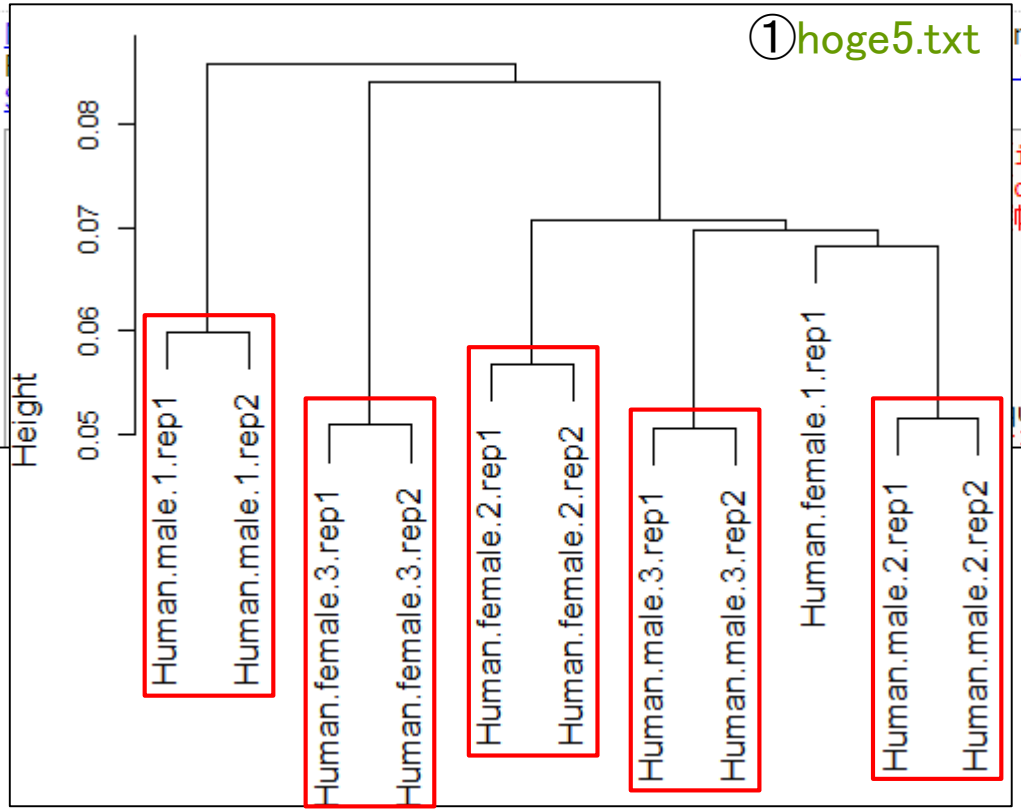
赤枠より、同一個体の反復データ（technical replicates）で末端のクラスターを形成していることが分かる。これはtechnical replicates同士の類似度が非常に高いことを意味しており、妥当。

解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013)

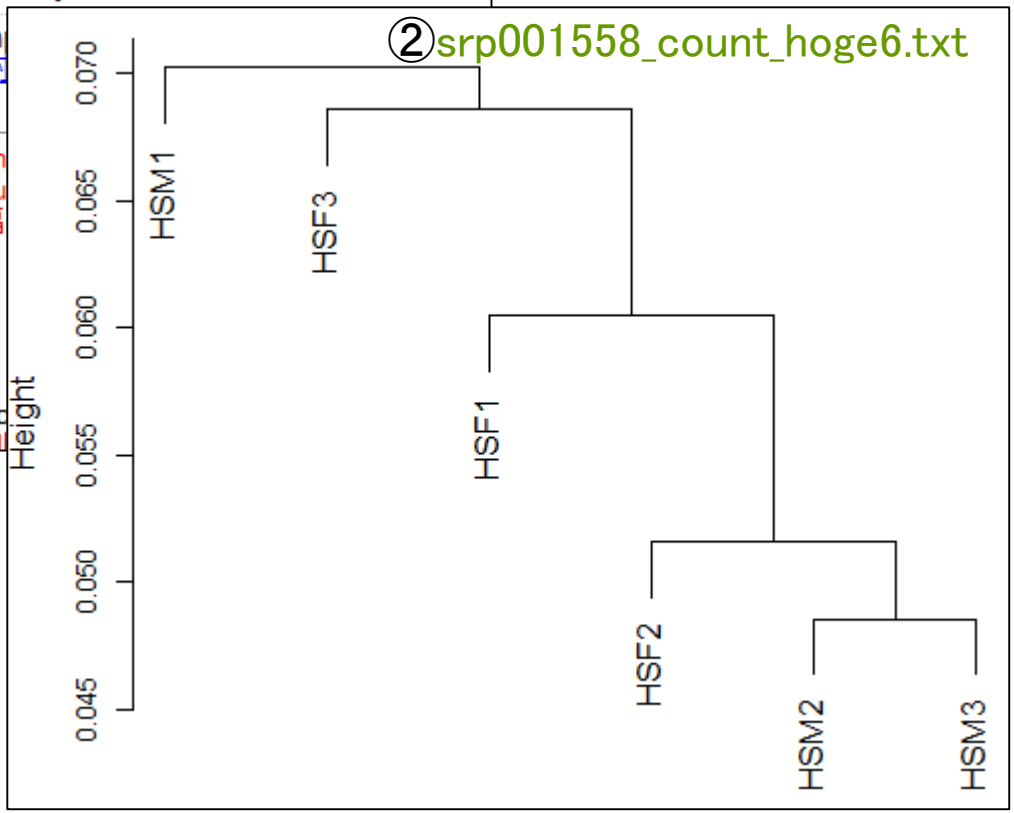
TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。 clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 (Tang et al., BMC Bioinformatics, 2015)中でも この関数を用いています(2015/11/05追加)。xlsx形式ファイルを入力とするやり方も追加しました(2015/11/15)。
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. 59,857 genes×6 samplesのリアルデータ(srp017142_count_bowtie.txt)の場合：

①hoge5.txt



②srp001558_count_hoge6.txt



サンプル間クラスタリング

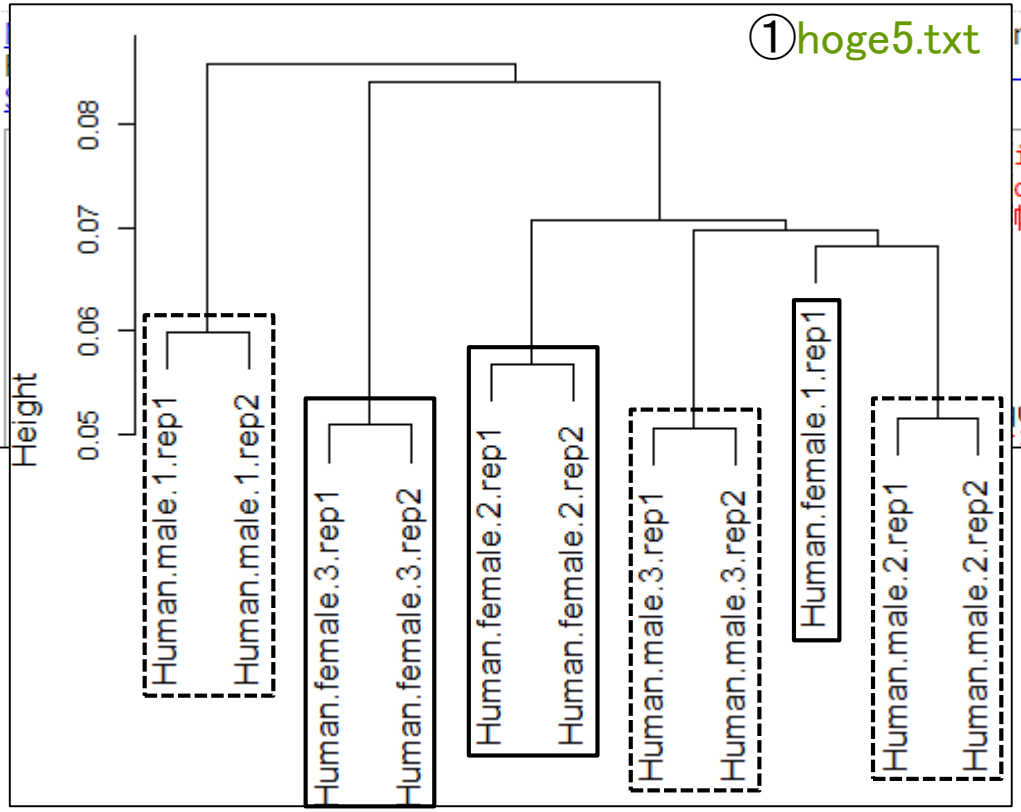
ヒトのメス（HSF）とヒトのオス（HSM）では、肝臓(Liver)の発現パターンに差がないことがわかる。つまり雌雄差はなさそう。

解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013)

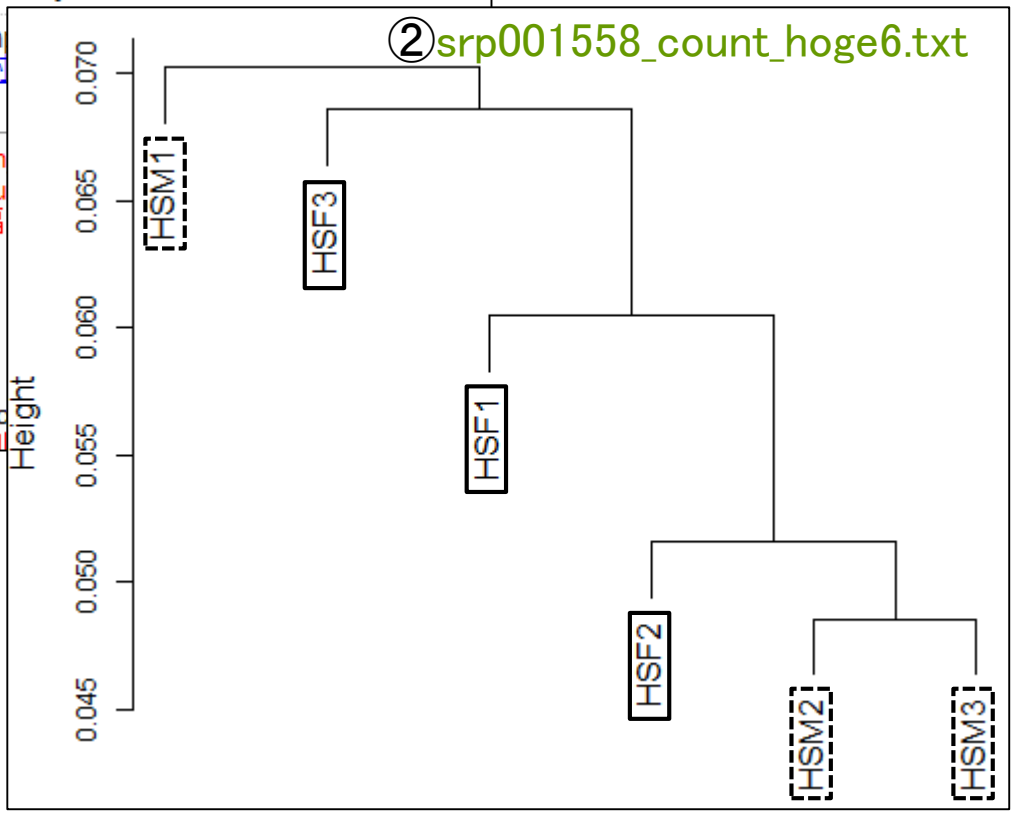
TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。 clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 (Tang et al., BMC Bioinformatics, 2015)中でも この関数を用いています(2015/11/05追加)。xlsx形式ファイルを入力とするやり方も追加しました(2015/11/15)。
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. 59,857 genes×6 samplesのリアルデータ(srp017142_count_bowtie.txt)の場合：

①hoge5.txt



②srp001558_count_hoge6.txt



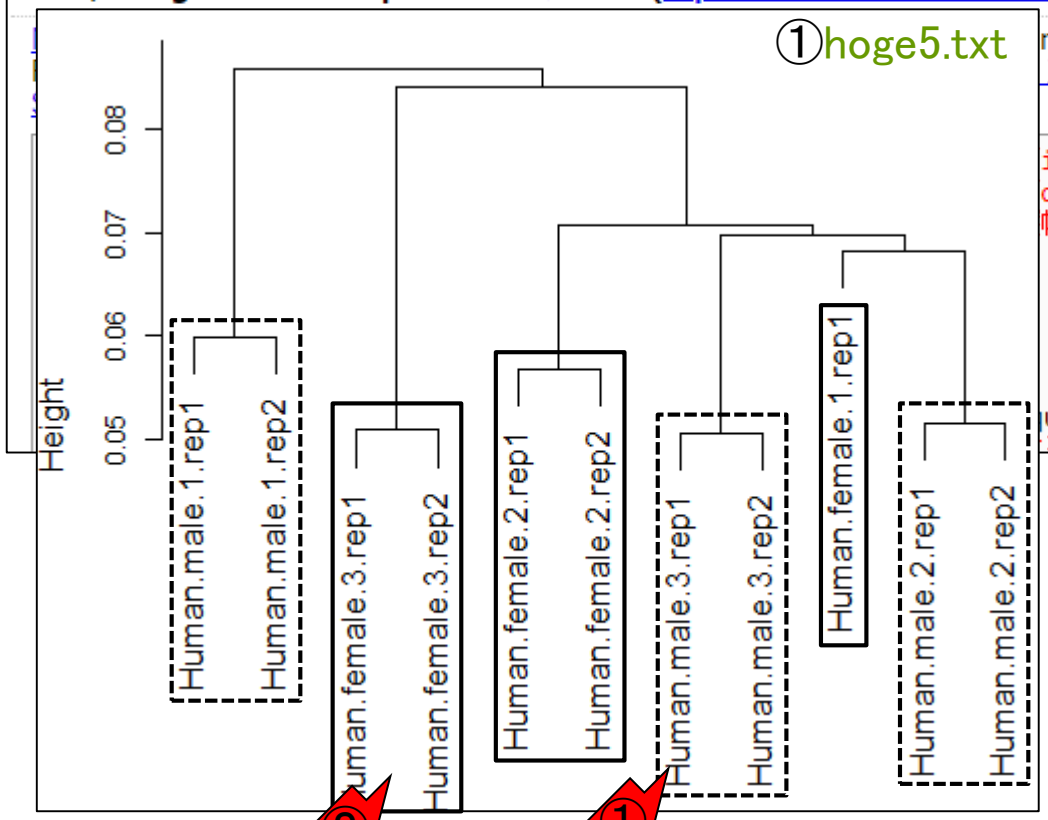
おさらい

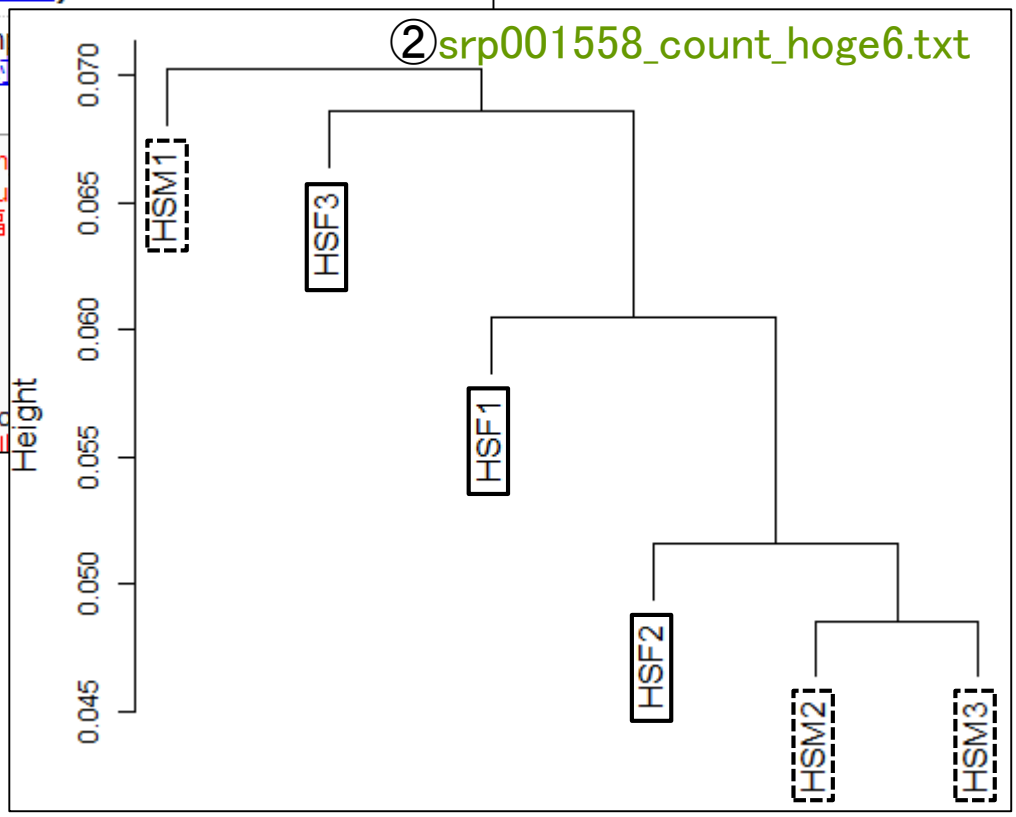
このクラスタリング結果の元データ（58,037 genesからなるカウントデータ）は、recountのグループが①この原著論文の、②公共DBのIDであるSRP001558（の生リードデータ）を独自のパイプラインで実行した結果をRangedSummarizedExperiment(RSE)形式のオブジェクトとしてrse_gene.Rdataとして提供しているもの。サンプルのメタデータ情報もSRP001558の記載内容をベースとしている。

解析 | クラスタリング | RNA-seq | サン

TCCパッケージを用いてサンプル間クラスタリングを行うやり方
クラスタリング結果を返します。多群間比較用の推奨ガイドライ
2015)中でも この関数を用いています(2015/11/05追加)。xlsx
(2015/11/15)。
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. 59,857 genes×6 samplesのリアルデータ(srp017142_count_bowtie.txt)の場合：





②

①

SRP001558（Blekhman et al., *Genome Res.*, **20**: 180-9, 2010）

おさらい

スライド20あたりまでで取り扱っていた20,689 genesからなるカウントデータは、①この原著論文の、②Supplementary Table 1で提供されているものであり、③から辿れます。遺伝子数も異なるうえ、当時とはアノテーション（遺伝子の座標）情報も異なると思われるので、クラスタリング結果の単純な比較はできない。

- マップ後 | カウント情報取得 | paired-end | トランスクリプ
- マップ後 | カウント情報取得 | トランスクリプトーム | BED
- カウント情報取得 | リアルデータ | について ③ modified
- カウント情報取得 | リアルデータ | SRP06124 | recount(C
- カウント情報取得 | リアルデー
- カウント情報取得 | リアルデー
- カウント情報取得 | リアルデー
- カウント情報取得 | リアルデー
- カウント情報取得 | リアルデー
- カウント情報取得 | リアルデー
- カウント情報取得 | リアルデー
- カウント情報取得 | リアルデー
- カウント情報取得 | リアルデー

カウント情報取得 | リアルデータ | について

ここではSAM/BAMなどのマッピング結果ファイルからのカウント情報取得ではなく、**最初からカウント情報になっているもののありかや、 それらを提供しているデータベースから取得するプログラム**を示します。

R用：

- GSVAdata：Hänzelmann et al., BMC Bioinformatics, 2013
- recount：Collado-Torres et al., Nat Biotechnol., 2017

  recountWorkflowでは、 raw countsではなくscale_counts関数実行結果をその後の解析に利用することを推奨しているので、2018年8月7日にrecountパッケージ経由でカウントデータを取得する際には デフォルトでscale_counts実行結果とするように変更しました。

R以外：

② Supplementary Table1(suppTable1.xls)：Blekhman et al., Genome Res., 2010

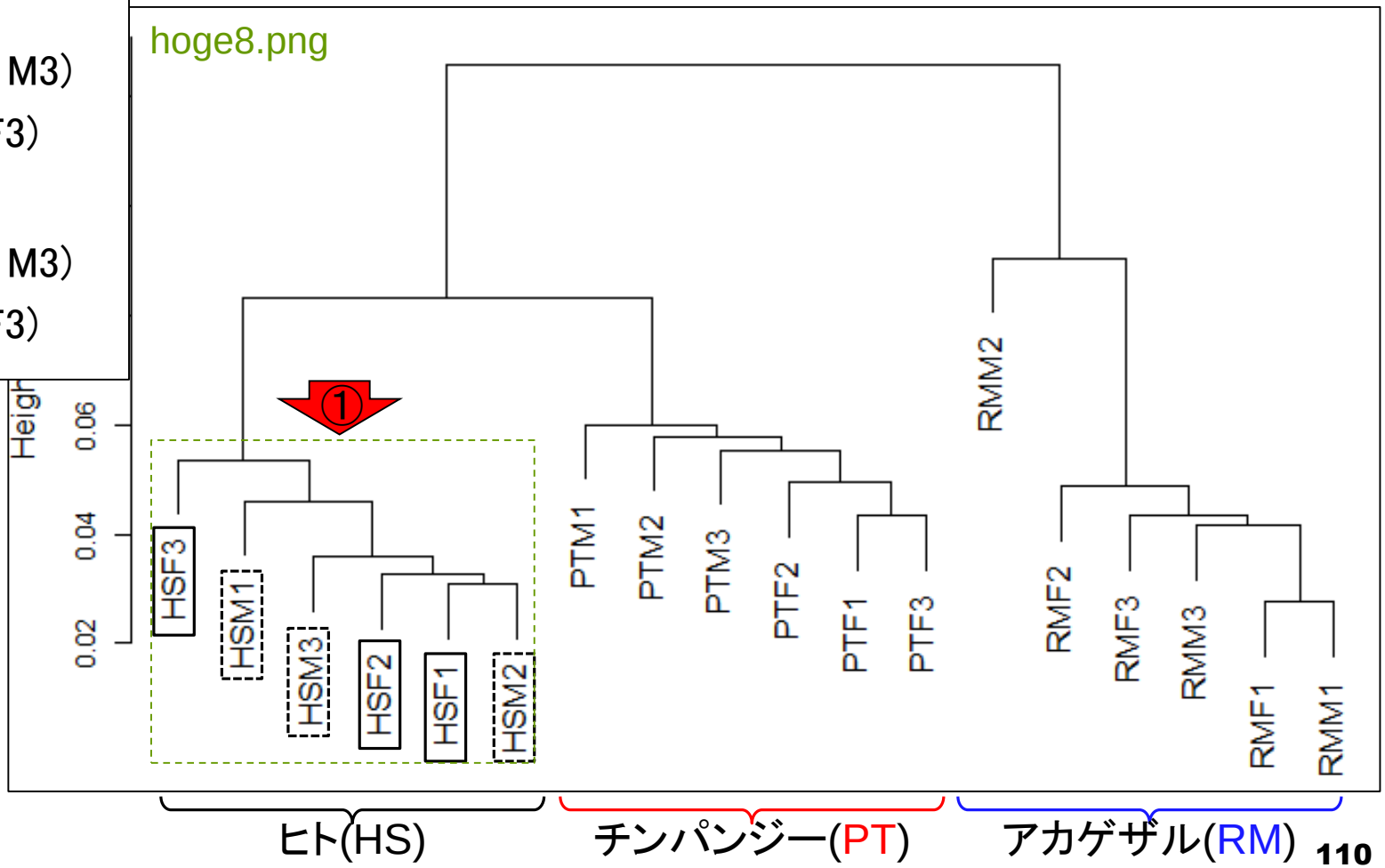
  この公共DB内のIDはGSE17274やSRP001558です。以下に整形したデータもあります：
  - サンプルデータ41で作成した20,689 genes×36 samplesのカウントデータ(sample_blekhman_36.txt)
  - サンプルデータ42で作成した20,689 genes×18 samplesのカウントデータ(sample_blekhman_18.txt)
- ReCount(website)：Frazee et al., BMC Bioinformatics, 2011
- recount2(website)：Collado-Torres et al., Nat Biotechnol., 2017
- DEE2(website)：Ziemann et al., Gigascience, 2019

① SRP001558（Blekhman et al., *Genome Res.*, **20**: 180-9, 2010）

おさらい

サンプルデータの例題42で作成した20,689 genes × 18 samplesのカウントデータのクラスタリング結果（スライド15–21）。
①メスとオスのサンプルが入り混じっており、雌雄差はなさそうという結論は、recountの58,037 genesのときと変わらない。

- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)



Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、例題4
    - 例題5、例題6、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

①ERP000546（ヒトの様々な器官由来のRNA-seqカウントデータ）。

ERP000546

(Rで)塩基配列解析

(last modified 2019/07/01, since 2010)

このウェブページ(
Macintosh2018.1
ています。 初心者
2018年7月に(R
(2018/07/18)

What's new? (

- マップ後 | カウント情報取得 | トランスクリプトーム | BEDファイルから (last modified 2014/06/21)
- カウント情報取得 | リアルデータ | について (last modified 2019/05/16)
- カウント情報取得 | リアルデータ | SRP061240 | recount(Collado-Torres_2017) (last modified 2018/08/28)
- カウント情報取得 | リアルデータ | SRP056295 | recount(Collado-Torres_2017) (last modified 2018/08/29)
- カウント情報取得 | リアルデータ | SRP056146 | recount(Collado-Torres_2017) (last modified 2018/10/25)
- カウント情報取得 | リアルデータ | SRP035988 | recount(Collado-Torres_2017) (last modified 2018/08/25)
- カウント情報取得 | リアルデータ | SRP026126 | recount(Collado-Torres_2017) (last modified 2018/08/30)
- カウント情報取得 | リアルデータ | SRP018853 | recount(Collado-Torres_2017) (last modified 2018/08/26)
- カウント情報取得 | リアルデータ | SRP012167 | recount(Collado-Torres_2017) (last modified 2018/08/24)
- カウント情報取得 | リアルデータ | SRP012167 | parathyroidSE(Haglund_2012) (last modified 2018/08/19)
- カウント情報取得 | リアルデータ | SRP001558 | recount(Collado-Torres_2017) (last modified 2018/08/08)
- カウント情報取得 | リアルデータ | SRP001540 | recount(Collado-Torres_2017) (last modified 2018/08/10)
- カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann_2013) (last modified 2018/07/03)
- カウント情報取得 | リアルデータ | ERP000546 | recount(Collado-Torres_2017) (last modified 2018/08/07)
- カウント情報取得 | シミュレーションデータ | RNA-seq | について (last modified 2019/04/05)

①

ERP000546

①ERP000546（ヒトの様々な器官由来のRNA-seqカウントデータ）。
②例題5は、technical replicatesのデータをマージし、サンプル名を理解しやすいものに変更して出力するコードです。

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#count_real_ERP000546_rec...

カウント情報取得 | リアルデータ | ERP000546 | recount(Collado-Torres_2017)

recountパッケージを用いて、 ERP000546(原著論文なし；ブラウザはIE以外を推奨)のカウント情報を含む RangedSummarizedExperimentクラスオブジェクトという形式の.Rdataをダウンロードしたり、 カウントデータの数値行列に した状態で保存するやり方を示します。 gedSummarizedExperimentというのがよくわからないとは思いますが、この中に Ensemblなどのgene IDだけでなく gen に、recountWorkflow で推奨されてい 「ファイル」－「ディレクトリの変更」で

1. geneレベルカウントデータ情報を得たい場

ERP000546という名前のフォルダが作成さ エクト名で取り扱えます。 ウェブサイトrec rse_gene.Rdataと同じです。

```
param_ID <- "ERP000546"

#必要なパッケージをロード
library(recount)

#本番(.Rdataをダウンロード)
download_study(param_ID, type="rse-
```

2. geneレベルカウントデータ情報を得たい場

1.の発展形として、ダウンロードも行い、さ ストファイルで保存するやり方です。 出力フ

```
out_f <- "hoge2.txt"
param_ID <- "ERP000546"

#必要なパッケージをロード
library(recount)

#本番(.Rdataのダウンロードとロード)
download_study(param_ID, type="rse-
```

②

5. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題4で得られたサンプルのメタデータ情報ファイル(erp000546_meta_samples.txt)中の ERR...からERS...の情報を手がかりにして、erp000546_meta_samples_added.txtの1番右の列で示すような 「これがheartで、これがkidneyで...」という対応関係をENAで1つ1つ調べたもので置き換えています。 出力ファイルはhoge5.txtです。

```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.txt"                    #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#入力ファイルの読み込み(.Rdata)
load(in_f)                              #in_fで指定した.Rdataをロード
rse <- rse_gene                         #rseとして取り扱う
rse                                     #確認してるだけです

#本番(カウントデータ取得)
rse <- scale_counts(rse)                #scale_counts実行(2018.08.07追加)
uge <- assays(rse)$counts               #カウントデータ行列を取得してugeに格納
dim(uge)                                #行数と列数を表示
head(uge)                               #確認してるだけです

#後処理(同一ERS IDの列をマージ)
uge <- as.data.frame(uge)               #行列形式からデータフレーム形式に変更
data <- cbind(                          #必要な列名の情報を取得したい列の順番で結合した結果をdataに
  uge$ERR030885 + uge$ERR030893,        #kidney(ERS025081)
  uge$ERR030894 + uge$ERR030886,        #heart(ERS025082)
```

ERP000546の例題5

入力の①rse_gene.RdataはERP000546用です。
改めてダウンロードしてから、コピペ実行。

5. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題4で得られたサンプルのメタデータ(ファイル(erp000546_meta_samples.txt)中の ERR...からERS...の情報を手がかりにして、erp000546_meta_samples_added.txtの1番右の列で示すような 「これがheartで、これがkidneyで...」という対応関係をENAで1つ1つ調べたもので置き換えています。 出力ファイルはhoge5.txtです。

①

```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.txt"                    #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#入力ファイルの読み込み(.Rdata)
load(in_f)                              #in_fで指定した.Rdataをロード
rse <- rse_gene                         #rseとして取り扱う
rse                                     #確認してるだけです

#本番(カウントデータ取得)
rse <- scale_counts(rse)                #scale_counts実行(2018.08.07追加)
uge <- assays(rse)$counts               #カウントデータ行列を取得してugeに格納
dim(uge)                                #行数と列数を表示
head(uge)                               #確認してるだけです

#後処理(同一ERS IDの列をマージ)
uge <- as.data.frame(uge)               #行列形式からデータフレーム形式に変更
data <- cbind(                          #必要な列名の情報を取得したい列の順番で結合した結果をdataに
  uge$ERR030885 + uge$ERR030893,        #kidney(ERS025081)
  uge$ERR030894 + uge$ERR030886,        #heart(ERS025082)
```

ERP000546の例題5

①出力ファイル（hoge5.txt）は、②58,037 genes × 19 samplesからなるカウントデータ。

5. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題4で得られたサンプルのメタデータ情報ファイル(erp000546_meta_samples.txt)中の ERR...からERS...の情報を手がかりにして、erp000546_meta_samples_added.txtの1番右の列で示すような「これがheartで、これがkidneyで...」という対応関係をENAで1つ1つ調べたもので置き換えています。 出力ファイルはhoge5.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッ

#入力ファイルの読み込み(.Rdata)
load(in_f)                                  #in_
rse <- rse_gene                             #rse
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #sca
uge <- assays(rse)$counts                   #カウ
dim(uge)                                    #行数
head(uge)                                   #確認

#後処理(同一ERS IDの列をマージ)
uge <- as.data.frame(uge)                   #行列
data <- cbind(                              #必要
  uge$ERR030885 + uge$ERR030893,            #ki
  uge$ERR030894 + uge$ERR030886,            #he
```

①

RGui (64-bit)

ファイル　編集　閲覧　その他　パッケージ　ウインドウ　ヘルプ　Vignettes

R Console

```
+    "mixture1", "brain", "lymphnode",     #列名を付加
+    "mixture2", "breast", "colon",        #列名を付加
+    "thyroid", "white_blood_cells",       #列名を付加
+    "adrenal", "mixture3", "testes",      #列名を付加
+    "prostate", "liver",                  #列名を付加
+    "skeletal_muscle", "adipose", "lung") #列名を付加
> rownames(data)<- rownames(uge)          #行名を付加
> dim(data)                               #行数と列数を$
[1] 58037    19
>
> #ファイルに保存(カウントデータ)
> tmp <- cbind(rownames(data), data)      #保存したい情$
> write.table(tmp, out_f, sep="\t", append=F, quote=F,$
>
```

②

ERP000546の例題5

①出力ファイル（hoge5.txt）は、②58,037 genes × 19 samplesからなるカウントデータ。③ここで見えているようなヒト組織のデータです。④58,037という数字に着目！Recountは生のリードデータから統一的な手順でカウントデータを得ているので、他のデータ（例：SRP001558）とほぼ直接的な比較ができます!

5. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題4で得られたサンプルのメタデータ情報ファイル(erp000546_meta_samples.t
報を手がかりにして、erp000546_meta_samples_added.txtの1番右の列で示すよ
がkidneyで...」という対応関係をENAで1つ1つ調べたもので置き換えています。 出

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに
out_f <- "hoge5.txt"                        #出力ファイル名を指定してout_fに

#必要なパッケージをロード
library(recount)                            #パッ

#入力ファイルの読み込み(.Rdata)
load(in_f)                                  #in_
rse <- rse_gene                             #rse
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #sc
uge <- assays(rse)$counts                   #カ
dim(uge)                                    #行
head(uge)                                   #確認

#後処理(同一ERS IDの列をマージ)
uge <- as.data.frame(uge)                   #行列
data <- cbind(                              #必要
  uge$ERR030885 + uge$ERR030893,            #kid
  uge$ERR030894 + uge$ERR030886,            #hea
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
+    "mixture1", "brain", "lymphnode",      #列名を付加
+    "mixture2", "breast", "colon",         #列名を付加
+    "thyroid", "white_blood_cells",        #列名を付加
+    "adrenal", "mixture3", "testes",       #列名を付加
+    "prostate", "liver",                   #列名を付加
+    "skeletal_muscle", "adipose", "lung")  #列名を付加
> rownames(data)<- rownames(uge)           #行名を付加
> dim(data)                                #行数と列数を$
[1] 58037    19
>
> #ファイルに保存(カウントデータ)
> tmp <- cbind(rownames(data), data)       #保存したい情$
> write.table(tmp, out_f, sep="\t", append=F, quote=F,$
>
```

Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ(technical replicatesマージ前)
  - Liverの3生物種間比較データ(technical replicatesマージ後)
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明(前半)
    - RangedSummarizedExperimentオブジェクトの説明(後半)、例題4
    - 例題5、例題6、ヒト(計6人分)のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5(19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得)
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行(課題)

サンプル間クラスタリング

②の例題5をテンプレートにして、③さきほど得たカウントデータファイル（hoge5.txt）を入力として実行。

- 解析 | 前処理 | scRNA-seq | について (last modified 2019/06/26) NEW
- 解析 | クラスタリング | RNA-seq | について (last modified 2019/04/04)
- 解析 | クラスタリング | RNA-seq | サンプル間 | hclust (last modified 2015/02/26)
- 解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013) (last modified 2018/08/06)
- 解析 | クラスタリング | RNA-seq | 遺伝子間(基礎) | MBCluster.Seq (Si_2014) (last modified 2018/09/23)
- 解析
- 解析
- 解析

①

解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013)

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。 clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 (Tang et al., BMC Bioinformatics, 2015)中でも この関数を用いています(2015/11/05追加)。xlsx形式ファイルを入力とするやり方も追加しました(2015/11/15)。
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ

②

1. 59,857 genes×6 sa

Neyret-Kahn et al., Ge
RNA-seqカウントデータ
SRP017142(Neyret-Kah

```
in_f <- "srp017142_c
out_f <- "hoge1.png"
param_fig <- c(500,

#必要なパッケージをロ
library(TCC)

#入力ファイルの読み込
data <- read.table(i
dim(data)
```

5. 60,234 genes×6 samplesのリアルデータ(hoge9_count_gene.txt)の場合：

Neyret-Kahn et al., Genome Res., 2013のgene-levelの2群間比較用(3 proliferative samples vs. 3 Ras samples)ヒトRNA-seqカウントデータです。 マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015)から得られます。

③

```
in_f <- "hoge9_count_gene.txt"          #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.png"                    #出力ファイル名を指定してout_fに格納
param_fig <- c(500, 400)                #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの読
dim(data)                               #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman",#クラスタリング実行結果をoutに格納
            hclust.method="average", unique.pattern=TRUE)#クラスタリング実行結果をoutに格納

#ファイルに保存
```

サンプル間クラスタリング

②の例題5をテンプレートにして、③さきほど得たカウントデータファイル（hoge5.txt）を入力として実行。④実行結果（hoge5.png）。⑤liverはここに位置しています。

- 解析 | 前処理 | scRNA-seq | について (last modified 2019/06/26) NEW
- 解析 | クラスタリング | RNA-seq | について (last modified 2019/04/04)
- 解析 | クラスタリング | RNA-seq | サンプル間 | hclust (last modified 2015/02/26)
- 解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun 2013) (last modified 20
- 解析 | クラスタリング | RNA-seq | 遺伝子間(基礎) | MBCluster.Seq(Si 2014) (last
- 解析
- 解析
- 解析

解析 | クラスタリング | RNA-seq | サンプル間

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示し
クラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱
2015)中でも この関数を用いています(2015/11/05追加)。xlsx形式フ
(2015/11/15)。
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いて

1. 59,857 genes×6 sa

Neyret-Kahn et al., Ge
RNA-seqカウントデータ
SRP017142(Neyret-Ka

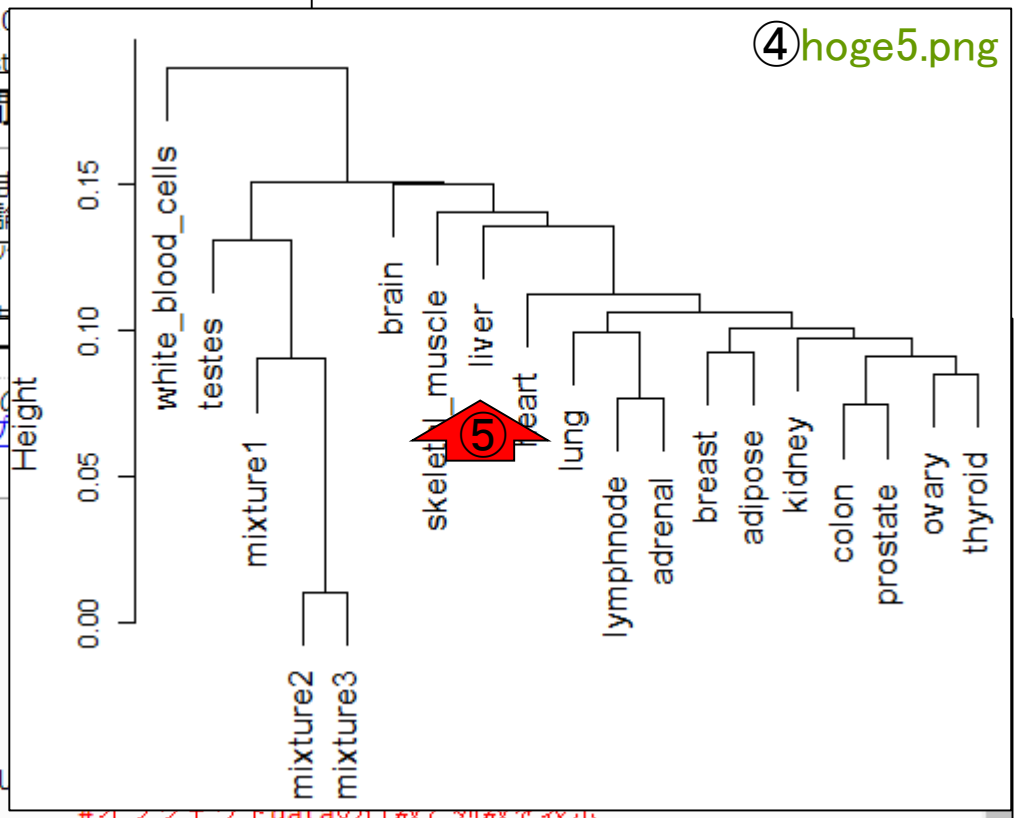
```
in_f <- "srp017142_c
out_f <- "hoge1.png"
param_fig <- c(500,

#必要なパッケージをロ
library(TCC)

#入力ファイルの読み込
data <- read.table(i
dim(data)
```

5. 60,234 genes×6 samplesのリアルデ

Neyret-Kahn et al., Genome Res., 2013
ヒトRNA-seqカウントデータです。 マップ
QuasR(Gaidatzis 2015)から得られます。

```
in_f <- "hoge9_count_gene.txt"
out_f <- "hoge5.png"
param_fig <- c(500, 400)

#必要なパッケージをロード
library(TCC)

#入力ファイルの読み込み
data <- read.table(in_f, header=TRU
dim(data)

#本番
out <- clusterSample(data, dist.method="spearman",#クラスタリング実行結果をoutに格納
            hclust.method="average", unique.pattern=TRUE)#クラスタリング実行結果をoutに格納

#ファイルに保存
```

④hoge5.png



Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、例題4
    - 例題5、例題6、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

おさらい1

①Blekhmanらのヒト肝臓（Liver）カウントデータを、②の例題6を実行してRecountデータベースから取得したファイルが…

(Rで)塩基配列解析

(last modified 2019/07/01, since 2010)

このウェブページ(
Macintosh2018.1
ています。 初心者
2018年7月に(R
(2018/07/18)

What's new? (

- マップ後 | カウント情報取得 | トランスクリプトーム | BEDファイルから (last modified 2014/06/21)
- カウント情報取得 | リアルデータ | について (last modified 2019/05/16)
- カウント情報取得 | リアルデータ | SRP061240 | recount(Collado-Torres_2017) (last modified 2018/08/28)
- カウント情報取得 | リアルデータ | SRP056295 | recount(Collado-Torres_2017) (last modified 2018/08/29)
- カウント情報取得 | リアルデータ | SRP056146 | recount(Collado-Torres_2017) (last modified 2018/10/25)
- カウント情報取得 | リアルデータ | SRP035988 | recount(Collado-Torres_2017) (last modified 2018/08/25)
- カウント情報取得 | リアルデータ | SRP026126 | recount(Collado-Torres_2017) (last modified 2018/08/30)
- カウント情報取得 | リアルデータ | SRP018853 | recount(Collado-Torres_2017) (last modified 2018/08/26)
- カウント情報取得 | リアルデータ | SRP012167 | recount(Collado-Torres_2017) (last modified 2018/08/24)
- カウント情報取得 | リアルデータ | SRP012167 | parathyroidSE(Haglund_2012) (last modified 2018/08/19)
- カウント情報取得 | リアルデータ | SRP001558 | recount(Collado-Torres_2017) (last modified 2018/08/08)
- カウント情報取得 | リアルデータ | SRP001540 | recount(Collado-Torres_2017) (last modified 2018/08/10)
- カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann_2013) (last modified 2018/07/03)
- カウント情報取得 | リアルデータ | ERP000546 | recount(Collado-Torres_2017) (last modified 2018/08/07)
- カウント情報取得 | シミュレーションデータ | RNA-seq | について (last modified 2019/04/05)

②

①

SRP001558（Blekhman et al., *Genome Res.*, **20**: 180-9, 2010）

おさらい1

①Blekhmanらのヒト肝臓（Liver）カウントデータを、②の例題6を実行してRecountデータベースから取得したファイルが…③のような列名になった、④計6サンプルからなる⑤srp001558_count_hoge6.txtでした。

6. ダウンロード済みのrse_gene.Rdataを入力として読

例題5の続きのようなものですが、technical replicatesのデータをマージした結果を出力しています。 例えば、"Human female 2 rep1"列と"Human female 2 rep2"列のカウント数の和をとり、列名を"HSF2"のようにしています。 この列名の表記法は、「サンプルデータ42の20,689 genes×18 samplesのリアルカウントデータ(sample_blekhman_18.txt)」中のヒトサンプル名と同じにしています。 58,037 genes×6 samplesからなる出力ファイルはsrp001558_count_hoge6.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力
out_f <- "srp001558_count_hoge6.txt"        #出力

#必要なパッケージをロード
library(recount)                            #パッ

#入力ファイルの読み込み(.Rdata)
load(in_f)                                  #in_
rse <- rse_gene                             #rse
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #sca
uge <- assays(rse)$counts                   #カウ
dim(uge)                                    #行数
head(uge)                                   #確認

#後処理(technical replicatesの列をマージ)
uge <- as.data.frame(uge)                   #行列
data <- cbind(                              #必要
  uge$SRR032116,                            #HSF
  uge$SRR032118 + uge$SRR032119,            #HSF
```

⑤

RGui (64-bit)

ファイル　編集　閲覧　その他　パッケージ　ウインドウ　ヘルプ　Vignettes

R Console

④

```
[1] 58037     6
> head(data)                                #確認してるだ
                  HSF1 HSF2 HSF3 HSM1 HSM2 HSM3
ENSG00000000003.14 1724 3001 1270 1526 2849 2570
ENSG00000000005.5     0    0    0    0    6    0
ENSG00000000419.12  337  611  493  523  651  461
ENSG00000000457.13  414  657  666 1388  620  840
ENSG00000000460.16  114  371  266  450  195  324
ENSG00000000938.12  362  901 2476 1191  541  812
>
> #ファイルに保存(カウントデータ)
> tmp <- cbind(rownames(data), data)      #保存したい情$
> write.table(tmp, out_f, sep="\t", append=F, quote=F,$
> |
```

③

おさらい1

①srp001558_count_hoge6.txtを入力として、②の例題5をテンプレートとしてクラスタリングした結果。

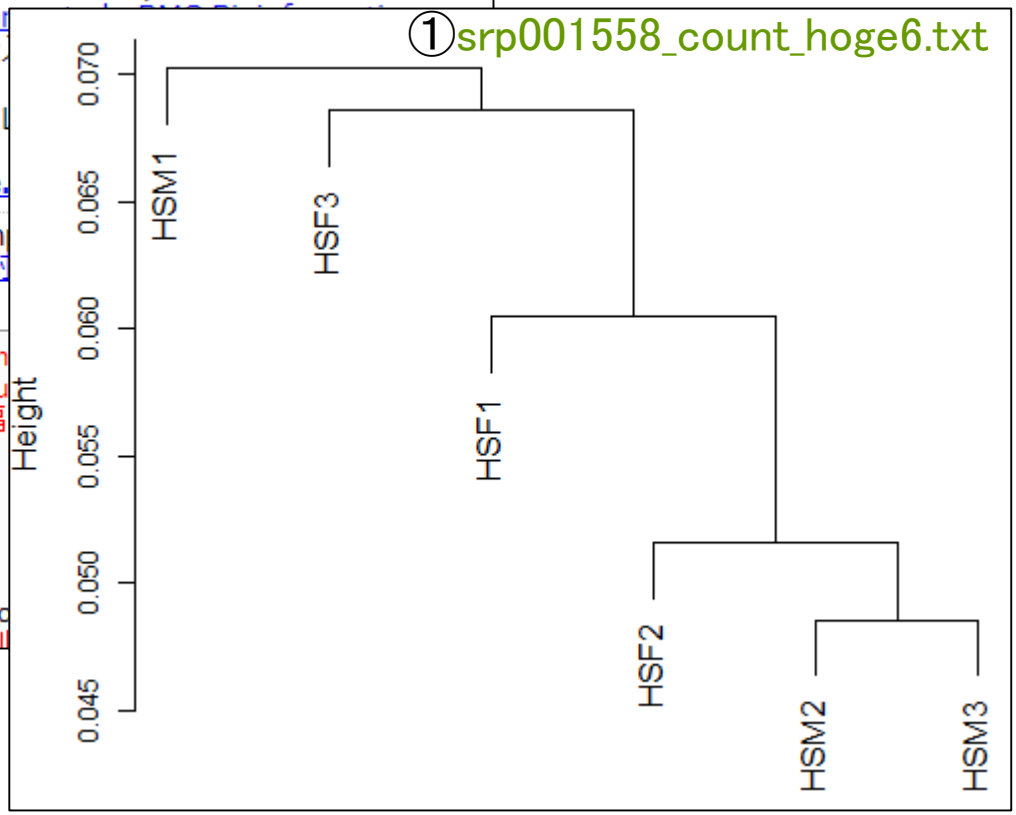
解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013) ②

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。 clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 (Tar
2015)中でも この関数を用いています(2015/11/05追加)。xlsx形式ファイルを
(2015/11/15)。
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディ

1. 59,857 genes×6 samplesのリアルデータ(srp017142_count_bowtie.

Neyret-Kahn et al., Genome Res., 2013の2群間比較用(3 proliferative sam
RNA-seqカウントデータです。 パイプライン | ゲノム | 発現変動 | 2群間 | 対
SRP017142(Neyret-Kahn_2013)から得られます。

```
in_f <- "srp017142_count_bowtie.txt"    #入力ファイル名を指定してin
out_f <- "hoge1.png"                    #出力ファイル名を指定してou
param_fig <- c(500, 400)                #ファイル出力時の横幅と縦幅

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quo
dim(data)                               #オブジェクトdataの行数と列
```



①ERP000546（様々な器官由来）のカウントデータを、

おさらい2

(Rで)塩基配列解析

(last modified 2019/07/01, since 2010)

このウェブページ(
Macintosh2018.1
ています。 初心者
2018年7月に(R
(2018/07/18)

What's new? (

- マップ後 | カウント情報取得 | トランスクリプトーム | BEDファイルから (last modified 2014/06/21)
- カウント情報取得 | リアルデータ | について (last modified 2019/05/16)
- カウント情報取得 | リアルデータ | SRP061240 | recount(Collado-Torres_2017) (last modified 2018/08/28)
- カウント情報取得 | リアルデータ | SRP056295 | recount(Collado-Torres_2017) (last modified 2018/08/29)
- カウント情報取得 | リアルデータ | SRP056146 | recount(Collado-Torres_2017) (last modified 2018/10/25)
- カウント情報取得 | リアルデータ | SRP035988 | recount(Collado-Torres_2017) (last modified 2018/08/25)
- カウント情報取得 | リアルデータ | SRP026126 | recount(Collado-Torres_2017) (last modified 2018/08/30)
- カウント情報取得 | リアルデータ | SRP018853 | recount(Collado-Torres_2017) (last modified 2018/08/26)
- カウント情報取得 | リアルデータ | SRP012167 | recount(Collado-Torres_2017) (last modified 2018/08/24)
- カウント情報取得 | リアルデータ | SRP012167 | parathyroidSE(Haglund_2012) (last modified 2018/08/19)
- カウント情報取得 | リアルデータ | SRP001558 | recount(Collado-Torres_2017) (last modified 2018/08/08)
- カウント情報取得 | リアルデータ | SRP001540 | recount(Collado-Torres_2017) (last modified 2018/08/10)
- カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann_2013) (last modified 2018/07/03)
- カウント情報取得 | リアルデータ | ERP000546 | recount(Collado-Torres_2017) (last modified 2018/08/07)
- カウント情報取得 | シミュレーションデータ | RNA-seq | について (last modified 2019/04/05)

①

おさらい2

①ERP000546（様々な器官由来）のカウントデータを、②例題5を実行してRecountデータベースから取得したファイルが、③19サンプルからなる、④hoge5.txtでした。

②

5. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題4で得られたサンプルのメタデータ情報ファイル(erp000546_meta_samples.txt)中の ERR...からERS...の情報を手がかりにして、erp000546_meta_samples_added.txtの1番右の列で示すような「これがheartで、これがkidneyで...」という対応関係をENAで1つ1つ調べたもので置き換えています。 出力ファイルはhoge5.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッ

#入力ファイルの読み込み(.Rdata)
load(in_f)                                  #in_
rse <- rse_gene                             #rse
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #sca
uge <- assays(rse)$counts                   #カウ
dim(uge)                                    #行数
head(uge)                                   #確認

#後処理(同一ERS IDの列をマージ)
uge <- as.data.frame(uge)                   #行列
data <- cbind(                              #必要
  uge$ERR030885 + uge$ERR030893,            #kidn
  uge$ERR030894 + uge$ERR030886,            #hear
```

④

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
+    "mixture1", "brain", "lymphnode",      #列名を付加
+    "mixture2", "breast", "colon",         #列名を付加
+    "thyroid", "white_blood_cells",        #列名を付加
+    "adrenal", "mixture3", "testes",       #列名を付加
+    "prostate", "liver",                   #列名を付加
+    "skeletal_muscle", "adipose", "lung")#列名を付加
> rownames(data)<- rownames(uge)           #行名を付加
> dim(data)                                #行数と列数を$
[1] 58037    19
>
> #ファイルに保存(カウントデータ)
> tmp <- cbind(rownames(data), data)       #保存したい情$
> write.table(tmp, out_f, sep="\t", append=F, quote=F,$
>
```

③

おさらい2

①ERP000546（様々な器官由来）のカウントデータを、②例題5を実行してRecountデータベースから取得したファイルが、③19サンプルからなる、④hoge5.txtでした。Recountは生のリードデータから統一的な手順でカウントデータを得ているので、⑤58,037 genesの並びは完全に同じ。つまり、他のデータをマージして解析可能です。

②

5. ダウンロード済みのrse_gene.Rdataを入力とし

例題4で得られたサンプルのメタデータ情報ファイル
報を手がかりにして、erp000546_meta_samples
がkidneyで…」という対応関係をENAで1つ1つ調べたもので置き換えています。 出力ファイルはhoge5.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.txt"                        #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(recount)                            #パッ

#入力ファイルの読み込み(.Rdata)
load(in_f)                                  #in_
rse <- rse_gene                             #rse
rse                                         #確認

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #sca
uge <- assays(rse)$counts                   #カウ
dim(uge)                                    #行数
head(uge)                                   #確認

#後処理(同一ERS IDの列をマージ)
uge <- as.data.frame(uge)                   #行列
data <- cbind(                              #必要
  uge$ERR030885 + uge$ERR030893,            #kidn
  uge$ERR030894 + uge$ERR030886,            #hear
```

④

RGui (64-bit)

ファイル　編集　閲覧　その他　パッケージ　ウインドウ　ヘルプ　Vignettes

R Console

```
+    "mixture1", "brain", "lymphnode",      #列名を付加
+    "mixture2", "breast", "colon",         #列名を付加
+    "thyroid", "white_blood_cells",        #列名を付加
+    "adrenal", "mixture3", "testes",       #列名を付加
+    "prostate", "liver",                   #列名を付加
+    "skeletal_muscle", "adipose", "lung")#列名を付加
> rownames(data)<- rownames(uge)           #行名を付加
> dim(data)                                #行数と列数を$
[1] 58037    19
>
> #ファイルに保存(カウントデータ)
> tmp <- cbind(rownames(data), data)       #保存したい情$
> write.table(tmp, out_f, sep="\t", append=F, quote=F,$
>
```

③

⑤

課題のイントロ

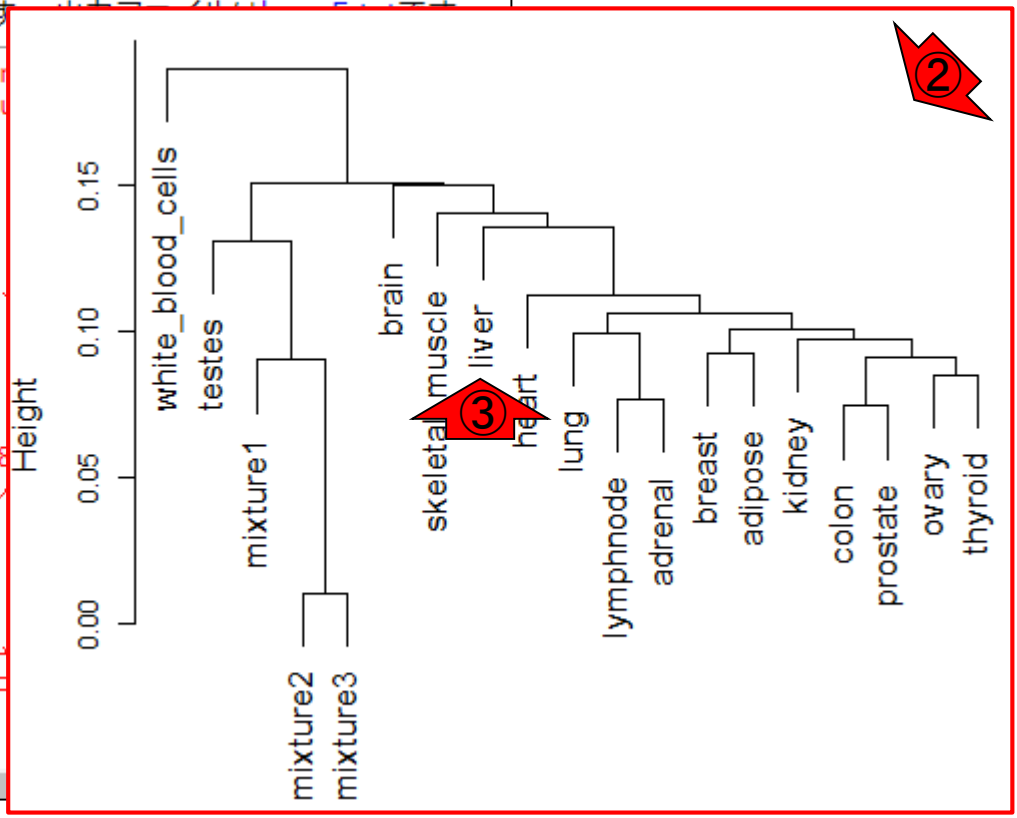
ERP000546（様々な器官由来）の計19サンプルからなるカウントデータである、①hoge5.txtを入力として、②サンプル間クラスタリングを実行した結果。③liverはこの位置にある。

5. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題4で得られたサンプルのメタデータ情報ファイル(erp000546_meta_samples.txt)中の ERR...からERS...の情報を手がかりにして、erp000546_meta_samples_added.txtの1番右の列で示すような 「これがheartで、これがkidneyで...」という対応関係をENAで1つ1つ調べたもので置き換えています

```
in_f <- "rse_gene.Rdata"                #入力ファイル名を指定してin
out_f <- "hoge5.txt"                    #出力ファイル名を指定してou

#必要なパッケージをロード
library(recount)                        #パッケージの読み込み

#入力ファイルの読み込み(.Rdata)
load(in_f)                              #in_fで指定した.Rdataをロ
rse <- rse_gene                         #rseとして取り扱う
rse                                     #確認してるだけです

#本番(カウントデータ取得)
rse <- scale_counts(rse)                #scale_counts実行(2018.08
uge <- assays(rse)$counts               #カウントデータ行列を取得し
dim(uge)                                #行数と列数を表示
head(uge)                               #確認してるだけです

#後処理(同一ERS IDの列をマージ)
uge <- as.data.frame(uge)               #行列形式からデータフレーム
data <- cbind(                          #必要な列名の情報を取得した
  uge$ERR030885 + uge$ERR030893,        #kidney(ERS025081)
  uge$ERR030894 + uge$ERR030886,        #heart(ERS025082)
```

①



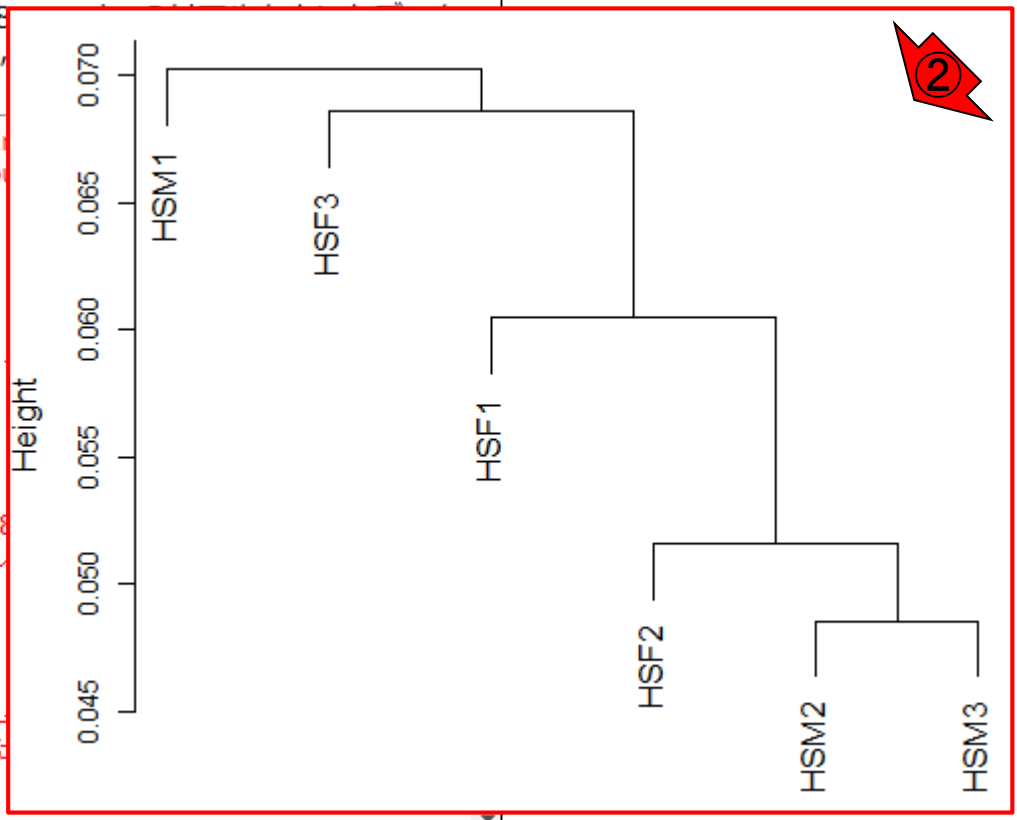
課題のイントロ

SRP001558（ヒト肝臓；Liver）の計6サンプルからなるカウントデータである、①srp001558_count_hoge6.txtを入力として、②サンプル間クラスタリングを実行した結果。

6. ダウンロード済みのrse_gene.Rdataを入力として読み込む場合：

例題5の続きのようなものですが、technical replicatesのデータをマージした結果を出力しています。 例えば、"Human female 2 rep1"列と"Human female 2 rep2"列のカウント数の和をとり、列名を"HSF2"のようにしています。 この列名の表記法は、「サンプルデータ42の20,689 genes×18
(sample_blekhman_18.txt)」中のヒトサンプル名と同じにしています。 58,
力ファイルはsrp001558_count_hoge6.txtです。

```
in_f <- "rse_gene.Rdata"                    #入力ファイル名を指定してi
out_f <- "srp001558_count_hoge6.txt"        ①出力ファイル名を指定してo

#必要なパッケージをロード
library(recount)                            #パッケージの読み込み

#入力ファイルの読み込み(.Rdata)
load(in_f)                                  #in_fで指定した.Rdataをロ
rse <- rse_gene                             #rseとして取り扱う
rse                                         #確認してるだけです

#本番(カウントデータ取得)
rse <- scale_counts(rse)                    #scale_counts実行(2018.08
uge <- assays(rse)$counts                   #カウントデータ行列を取得し
dim(uge)                                    #行数と列数を表示
head(uge)                                   #確認してるだけです

#後処理(technical replicatesの列をマージ)
uge <- as.data.frame(uge)                   #行列形式からデータフレーム
data <- cbind(                              #必要な列名の情報を取得した
  uge$SRR032116,                            #HSF1
  uge$SRR032118 + uge$SRR032119,            #HSF2
```



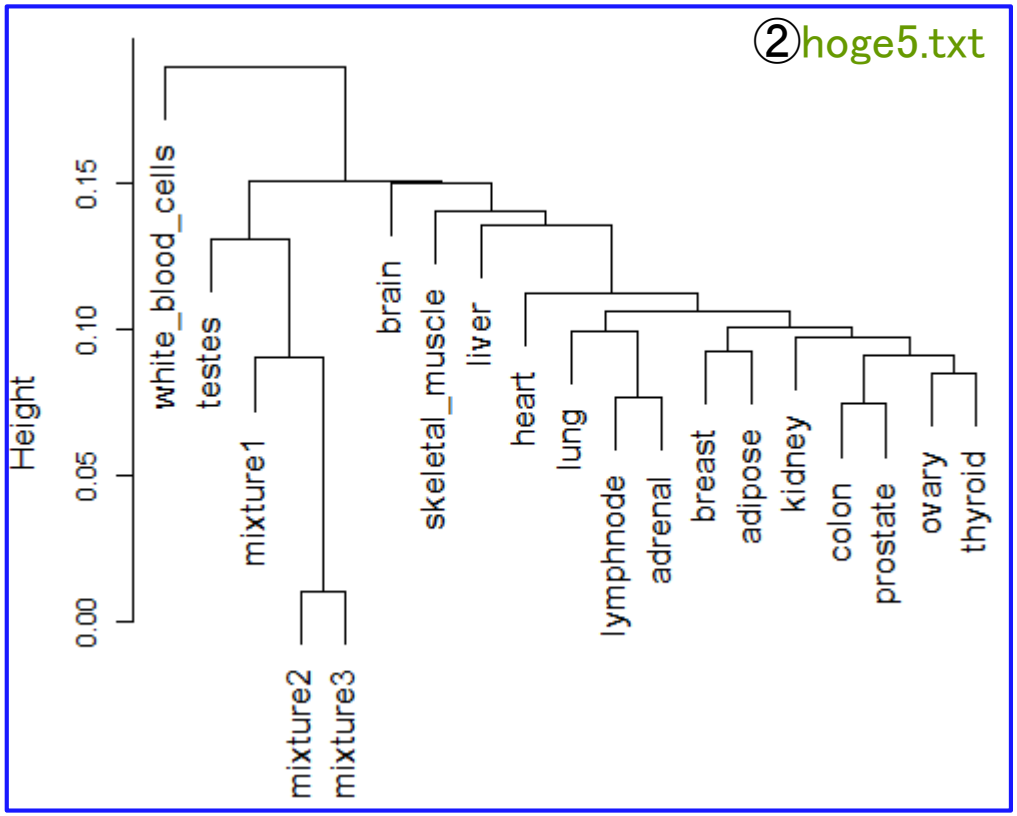
Contents

- サンプル間クラスタリング
  - Liverの3生物種間比較データ（technical replicatesマージ前）
  - Liverの3生物種間比較データ（technical replicatesマージ後）
- 公共?!カウントデータセット
  - Recount、recount2
  - Liverの3生物種間比較データ(SRP001558)をrecount2で眺める
  - SRP001558のrse_gene.Rdataからの情報抽出
    - 例題1と3、RangedSummarizedExperimentオブジェクトの説明（前半）
    - RangedSummarizedExperimentオブジェクトの説明（後半）、例題4
    - 例題5、例題6、ヒト（計6人分）のデータのみでサンプル間クラスタリング
  - ERP000546のrse_gene.Rdataからの情報抽出
    - 例題5（19サンプルからなるヒトの様々な器官由来のカウントデータファイルの取得）
    - サンプル間クラスタリングの実行
  - SRP001558とERP000546をマージしてクラスタリング
    - おさらい、実行（課題）

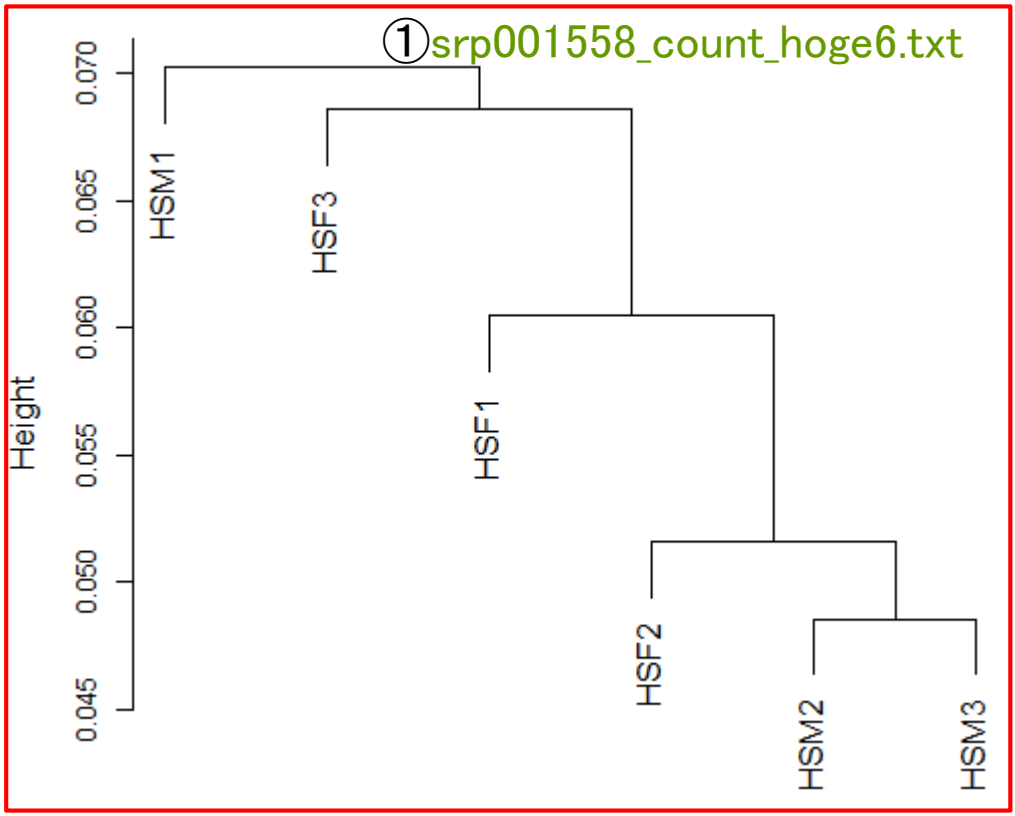
課題

SRP001558（ヒト肝臓；Liver）の計6サンプルからなるカウントデータである①srp001558_count_hoge6.txtと、ERP000546（様々な器官由来）の計19サンプルからなるカウントデータである②hoge5.txtをマージ（連結）したデータを入力として、サンプル間クラスタリングを実行し、結果を考察せよ。

②hoge5.txt



①srp001558_count_hoge6.txt



課題のヒント

プログラミングスキルを高めるのが目的のヒトは、見ずに自力で行ってください。

```
in_f1 <- "hoge5.txt"                          #入力ファイル名を指定してin_fに格納↓
in_f2 <- "srp001558_count_hoge6.txt"          #入力ファイル名を指定してin_fに格納↓
out_f <- "hoge.png"                           #出力ファイル名を指定してout_fに格納↓
param_fig <- c(600, 400)                      #ファイル出力時の横幅と縦幅を指定(単位はピクセル)↓
↓
#必要なパッケージをロード↓
library(TCC)                                  #パッケージの読み込み↓
↓
#入力ファイルの読み込み↓
data1 <- read.table(in_f1, header=TRUE, row.names=1, sep="¥t", quote="")#in_fで指定したファイルの読み込み↓
data2 <- read.table(in_f2, header=TRUE, row.names=1, sep="¥t", quote="")#in_fで指定したファイルの読み込み↓
data <- cbind(data1, data2)                   #列方向で連結した結果をdataに格納↓
dim(data)                                     #オブジェクトdataの行数と列数を表示↓
↓
#本番↓
out <- clusterSample(data, dist.method="spearman",#クラスタリング実行結果をoutに格納↓
            hclust.method="average", unique.pattern=TRUE)#クラスタリング実行結果をoutに格納↓
↓
#ファイルに保存↓
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイルの各種パラメータを指定↓
par(mar=c(0, 4, 1, 0))                        #下、左、上、右の順で余白（行）を指定↓
plot(out, sub="", xlab="", cex.lab=1.2,#樹形図(デンドログラム)の表示↓
  cex=1.3, main="", ylab="Height")            #樹形図(デンドログラム)の表示↓
dev.off()                                     #おまじない↓
```

例えば、①例題5のRコードをテンプレートとして利用しましょう。

課題のヒント

- 解析 | 前処理 | scRNA-seq | について (last modified 2019/06/26) NEW
- 解析 | クラスタリング | RNA-seq | について (last modified 2019/04/04)
- 解析 | クラスタリング | RNA-seq | サンプル間 | hclust (last modified 2015/02/26)
- 解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013) (last modified 2018/08/06)
- 解析 | クラスタリング | RNA-seq | 遺伝子間(基礎) | MBCluster.Seq(Si_2014) (last modified 2018/09/23)
- 解析
- 解析
- 解析

解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013)

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。 clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 (Tang et al., BMC Bioinformatics, 2015)中でも この関数を用いています(2015/11/05追加)。xlsx形式ファイルを入力とするやり方も追加しました(2015/11/05)

「ファイル

①

1. 59,857 g

Neyret-Kah
RNA-seqカ
SRP017142

```
in_f <- "
out_f <-
param_fig

#必要なパッ
library(TC

#入力ファイ
data <- re
dim(data)
```

5. 60,234 genes×6 samplesのリアルデータ(hoge9_count_gene.txt)の場合：

Neyret-Kahn et al., Genome Res., 2013のgene-levelの2群間比較用(3 proliferative samples vs. 3 Ras samples)ヒトRNA-seqカウントデータです。 マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015)から得られます。

```
in_f <- "hoge9_count_gene.txt"          #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.png"                    #出力ファイル名を指定してout_fに格納
param_fig <- c(500, 400)                #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの読
dim(data)                               #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman",#クラスタリング実行結果をoutに格納
            hclust.method="average", unique.pattern=TRUE)#クラスタリング実行結果をoutに格納

#ファイルに保存
```

課題のヒント

例えば、①例題5のRコードをテンプレートとして利用しましょう。
変更するのは、基本的に②の部分のみでよい。2つのデータをマージした後の数値行列のオブジェクト名を③dataとすれば、④本番のところ以降のコードは変更する必要がなくなります。

- 解析 | 前処理 | scRNA-seq | について (last modified 2019
- 解析 | クラスタリング | RNA-seq | について (last modified
- 解析 | クラスタリング | RNA-seq | サンプル間 | hclust (last modified 2015/02/26)
- 解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013) (last modified 2018/08/06)
- 解析 | クラスタリング | RNA-seq | 遺伝子間(基礎) | MBCluster.Seq(Si_2014) (last modified 2018/09/23)
- 解析
- 解析
- 解析

解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun_2013)

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。 clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 (Tang et al., BMC Bioinformatics, 2015)中でも この関数を用いています(2015/11/05追加)。xlsx形式ファイルを入力とするやり方も追加しました(2015/11/05)

「ファイル

1. 59,857 g

Neyret-Kah
RNA-seqカ
SRP017142

①

5. 60,234 genes×6 samplesのリアルデータ(hoge9_count_gene.txt)の場合：

Neyret-Kahn et al., Genome Res., 2013のgene-levelの2群間比較用(3 proliferative samples vs. 3 Ras samples)ヒトRNA-seqカウントデータです。 マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015)から得られます。

```
in_f <- "hoge9_count_gene.txt"          #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.png"                    #出力ファイル名を指定してout_fに格納
param_fig <- c(500, 400)                #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの読
dim(data)                               #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman",#クラスタリング実行結果をoutに格納
            hclust.method="average", unique.pattern=TRUE)#クラスタリング実行結果をoutに格納

#ファイルに保存
```

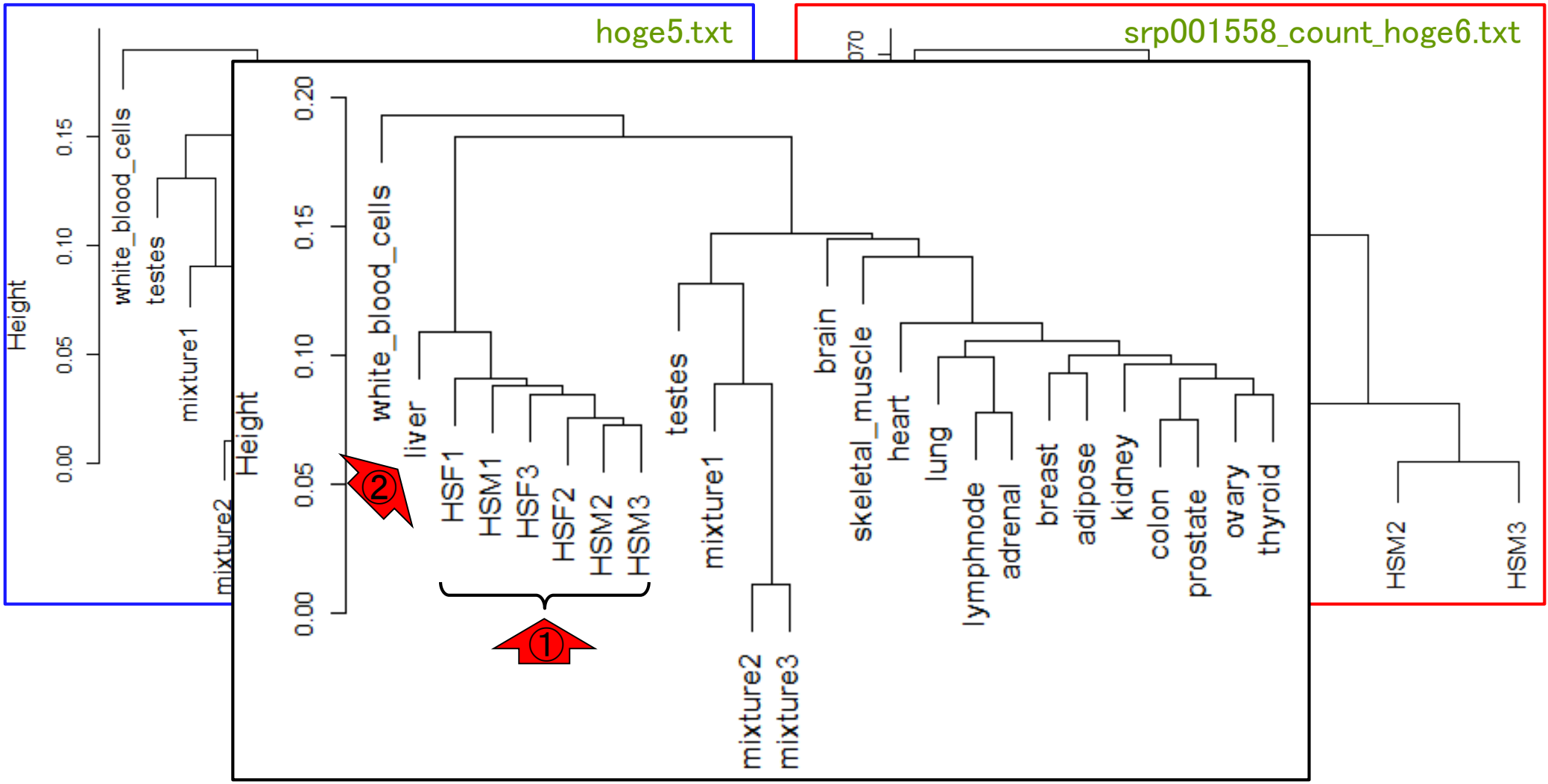
②

③

④

課題のヒント

SRP001558（ヒト肝臓；Liver）の①6サンプルは、ERP000546（様々な器官由来）の計19サンプルからなるカウントデータ中の②liverと似た発現パターンになっていますね。



課題のヒント

他のデータセットを組み合わせて眺めることを通じて、SRP001558（ヒト肝臓；Liver）の6サンプルのみのクラスタリング結果の縦軸（Height；数値が0に近いほど類似度が高い）が非常に小さい値があったことに気づくことができます。