ex3: Count data import and scatter plot

''arab2.txt''は、6 libraries (2 groups x 3 biological replicates) のシロイヌナズナRNA-seqのデータである。すでにマッピング済みで遺伝子毎のリードカウントがタブ区切りテキストとして提供されている。このexerciseでは、テーブルの中身を確認しデータの概要を把握する基本テクニックを習得する。

Data

(~/data/SS/以下にある)

- arab2.txt : count table

Inspect table with MS Excel

１）表計算ソフトMS Excelを使って ''arab2.txt'' の中身を確認しよう。

２）MS Excelで、m1とm2のscatter plot（散布図）を書いてみよう。このふたつは同一コンディションのbiological replicateなので、発現パターンは両者で良く似ているはずである。次にm1とh1を比較しよう。このふたつはコントロールと実験群の比較なので有る程度の発現パターンの違い（すくなくともm1 vs m2よりも大きい違い）が期待される。

ヒント：xy軸ともに対数をとること。

コメント：ノーマライズなどを施していない生データでもこれだけ豊富な情報が得られることを認識して欲しい。

Inspect table with R

R でテーブルの確認とscatter plotを書いてみよう。

Data import

```
> dat <- read.delim("arab2.txt", row.names=1, head=T)  # read tab-delimited text
> head(dat)
          m1 m2 m3 h1 h2 h3
AT1G01010 35 77 40 46 64 60
AT1G01020 43 45 32 43 39 49
AT1G01030 16 24 26 27 35 20
AT1G01040 72 43 64 66 25 90
AT1G01050 49 78 90 67 45 60
AT1G01060  0 15  2  0 21  8
```

Inspect table

```
> dim(dat)
[1] 26221     6
```

Q1：dimコマンドは何をするものですか？

Q2：また、その結果得られた、26221 と 6　は何を意味しますか？

Inspect data by column

それぞれのライブラリの、リードカウント合計は重要な基礎情報である。計算してみよう。

Q3: それぞれのライブラリのリードカウント合計を求めなさい。

例としてm1カラムの合計を計算する。

```
> sum(dat$m1)
[1] 1902032
```

他のカラムも合計を計算しよう。また、これらは基礎情報として重要なので記録しておこう。

Q4 (やや難): 約2万5千遺伝子にはカウント０のものから非常にたくさんのカウントをもつものがある。カウントの、i)最大値、最小値、平均値、中央値はいくつか調べよう。ii) ヒストグラムを書きなさい。

m1 を例に実行例を示す。

i) 最大値、最小値、平均値、中央値

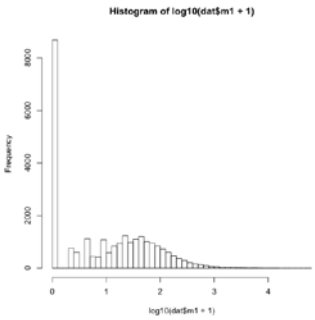
```
> sum(dat$m1)
[1] 1902032
> max(dat$m1)
[1] 61791
> min(dat$m1)
[1] 0
> mean(dat$m1)
[1] 72.5385
> median(dat$m1)
[1] 9
```

ii) ヒストグラム

```
> hist(dat$m1)
```

しかし、このグラフではあまり特徴がつかめないと思う。対数をとってみよう。

```
> hist(log10(dat$m1 + 1), breaks="Scott")
```



Scatter plot

Q5: m1 vs m2 をscatter plotで比較しよう。

```
> plot(dat$m1 + 1, dat$m2 + 1, log="xy")
```

それぞれ+1しているのは、log0は計算できないため。+1して下駄を履かせている。

Q6: ほかのライブラリどうしもscatter plotを描いて比較しよう。

Play with scatter plot

plotなどのグラフィックス関数には、様々な引数を与えることによって非常に多くの描画パラメータを変更でき、グラフの見栄えを変更することが出来る。scatter plotの色、形、などを変更する練習をしてみよう。

以下のコマンドをテンプレートにして、col（色）, pch（点の形状）, cex（点の大きさ）, main（グラフのタイトル）, xlab|ylab（x軸やy軸のラベル）を変更して、変化を確認しよう。

```
> plot(log2(dat$m1)+1, log2(dat$h1)+1, col="DarkBlue", pch=16, cex=0.6,
       main="MA plot", xlab="m1", ylab="h1")
```