ライフサイエンス分野データの可視化と共有化 at AJACS安芸

情報・システム研究機構 データサイエンス共同利用基盤施設 ライフサイエンス統合データベースセンター
坊農 秀雅 http://bonohu.jp/ bono@dbcls.rois.ac.jp mailto:bono@dbcls.rois.ac.jp
2016年7月6日

これは統合データベース講習会AJACS安芸「ライフサイエンス分野データの可視化と共有化」の資料です。

概要

本講習は、データ可視化とそれらを共有化する手段に関して、さまざまなツールを紹介します。とくに、大量の塩基配列データを可視化する手段、ならびに遺伝子発現量などの数値データからそれらの特徴を抽出する方法について説明、実習します。

講習の流れ

今回の講習では、以下の内容について説明します。【実習】と書かれた項目を皆さんと一緒にやって参ります。【発展】はこの場ではやりませんが、早く出来てしまった人は是非取り組んでみてください。

- 研究現場で頻繁に使われるDBやツールを知る: 統合TV
- 配列データの可視化と共有化
    - 多重配列アラインメント
    - ゲノムブラウザ
    - 配列レポジトリ
- 数値データの共有化: 公共データレポジトリ
- 数値データの可視化
    - パイチャート(pie chart)
    - 主成分分析(PCA: Principal Component Analysis)
    - 階層クラスタリング(Hierachical Clustering)

    - ヒートマップ(heatmap)
- ビジネスインテリジェンス(BI)ツールの活用

研究現場で頻繁に使われるデータベースやツールを知る

統合TV http://togotv.dbcls.jp/ja/

- 生命科学分野の有用なデータベースやツールの使い方を動画で紹介するウェブサイト
    - http://togotv.dbcls.jp/ http://togotv.dbcls.jp/

https://gyazo.com/edbabee661757e2a50f6f8ee77c3e778

    - YouTube版もあります　http://youtube.com/togotv/ http://youtube.com/togotv/
    - ウェブサイトへのアクセスから結果の見方まで、操作の一挙手一投足がわかります。
        - 講義・講習などの参考資料や後輩指導の教材として利用できます。
        - 本講習中、本家サイトが繋がらない時は、統合TVのYouTube版を見ればおおよその内容がわかるようになっています。
        - 今回の講習に関連する内容の多くは、「発現解析」タグのついた動画にあります。
        - 過去の講習会の内容はそのほとんどが統合TVに収録されており、いつでもどこでも繰り返し復習できるようになっています。
    - お探しの動画が見つからない or 統合TV未掲載の場合は、統合TV番組リクエストフォーム http://togotv.dbcls.jp/ja/contact.htmlへどうぞ!!

- 統合TVを作ってくれる方、募集中!!
  https://twitter.com/bonohu/status/747954940157399040

配列データの可視化と共有化

多重配列アラインメント

- Jalview http://www.jalview.org/ http://www.jalview.org/

【発展1】Jalviewでアミノ酸配列群を可視化する

1. DesktopとApplet版がありますが、Desktop版をJalviewのウェブサイトからインストールしましょう
2. 例として、さまざまな生物種で *PARK7*(*DJ-1*)のタンパク質配列(`PARK7.fa` https://raw.githubusercontent.com/AJACS-training/AJACS60/master/bono2/PARK7.fa)を File -> Input Alignment -> From File から読み込んでみましょう。
3. 読み込んで現れたWindowの中のメニューの Web Service -> Alignment -> ClustalO -> withDefaults を選んでみましょう。しばらく待つと多重配列アラインメントが計算されて返って来ます。
4. さらに Calculate -> Calculate Tree -> Neighbor Joining Using… を選ぶと系統樹が描かれます
5. 配列名がIDばかりで無味乾燥? `PARK7.out2.fa` https://raw.githubusercontent.com/AJACS-training/AJACS60/master/bono2/PARK7.out2.fa)を代わりの読み込んでみましょう

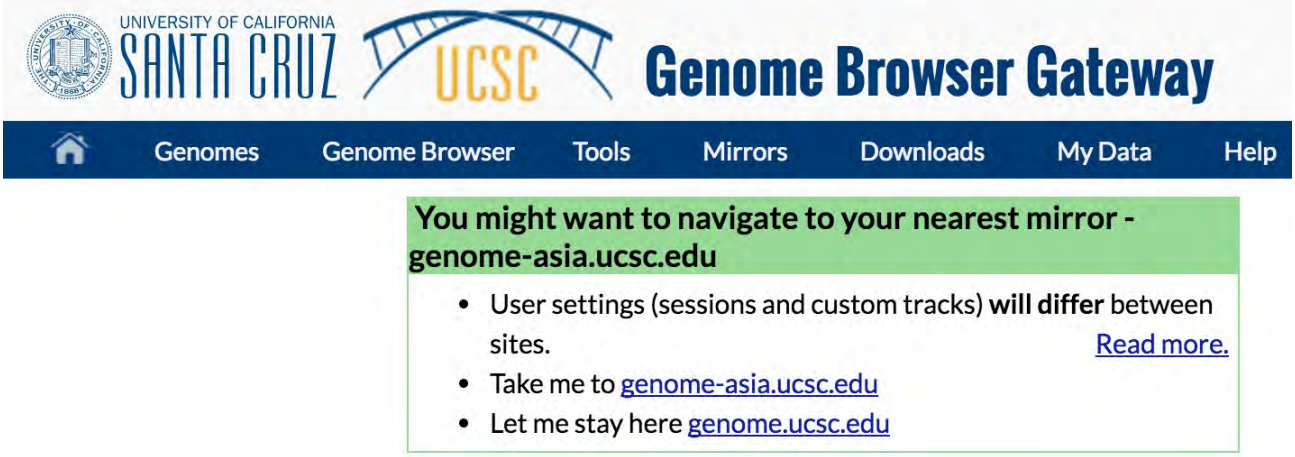
【復習用 統合TV】 Jalviewを使って配列解析・系統樹作成をする2013
http://doi.org/10.7875/togotv.2013.049

ゲノムブラウザ

- UCSC Genome Browser http://genome.ucsc.edu/ http://genome.ucsc.edu/

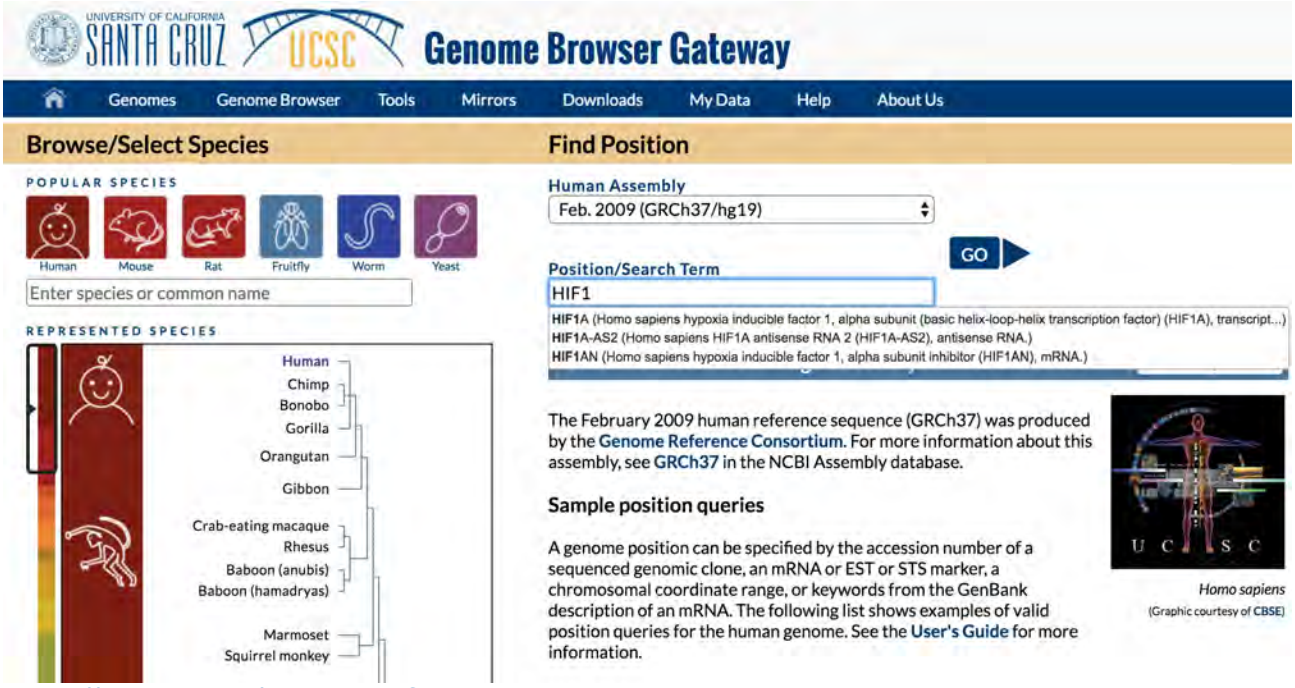
【実習1】UCSC Genome Browserで公共データを可視化する

1. http://genome.ucsc.edu/ http://genome.ucsc.edu/ にアクセスし、上部のメニューバーの `Genomes` にマウスをのせると上述のヒトとマウスのアッセンブリが選択できますが、ここでは気にせずそのままクリックします
2. mirrorサーバー選択のページがでますので、 `genome-asia.ucsc.edu` をクリック



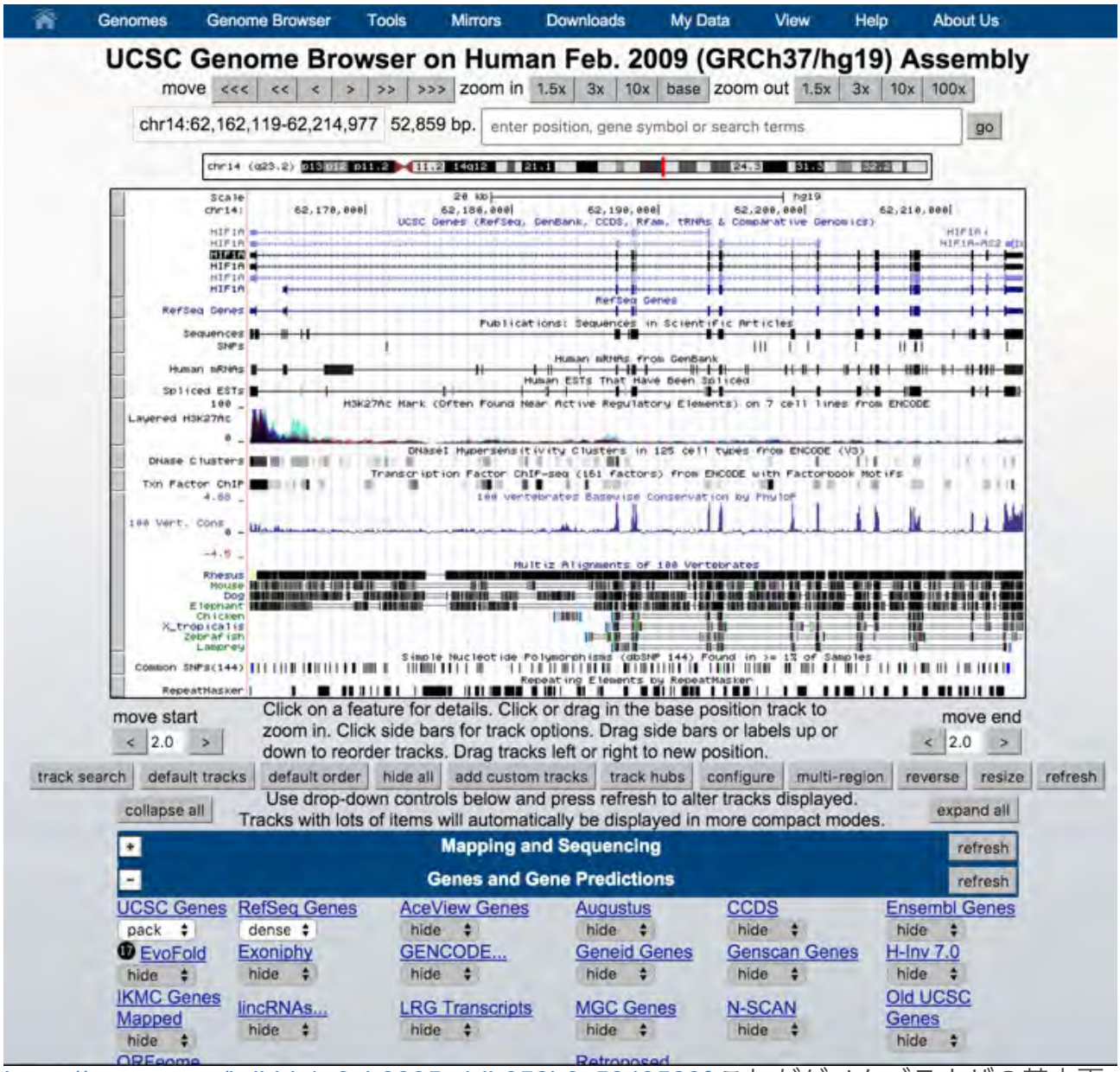
https://gyazo.com/378a8f0db4792fbe2b42b5ed46d00484

3. Humanアイコンをクリックして、Assemblyに `Feb.2009 (GRCh37/hg19)` を選び、Position/Search Termに `HIF1` と入力すると入力補完されるので、一番上の `HIF1A` を選び、GOボタンをクリック



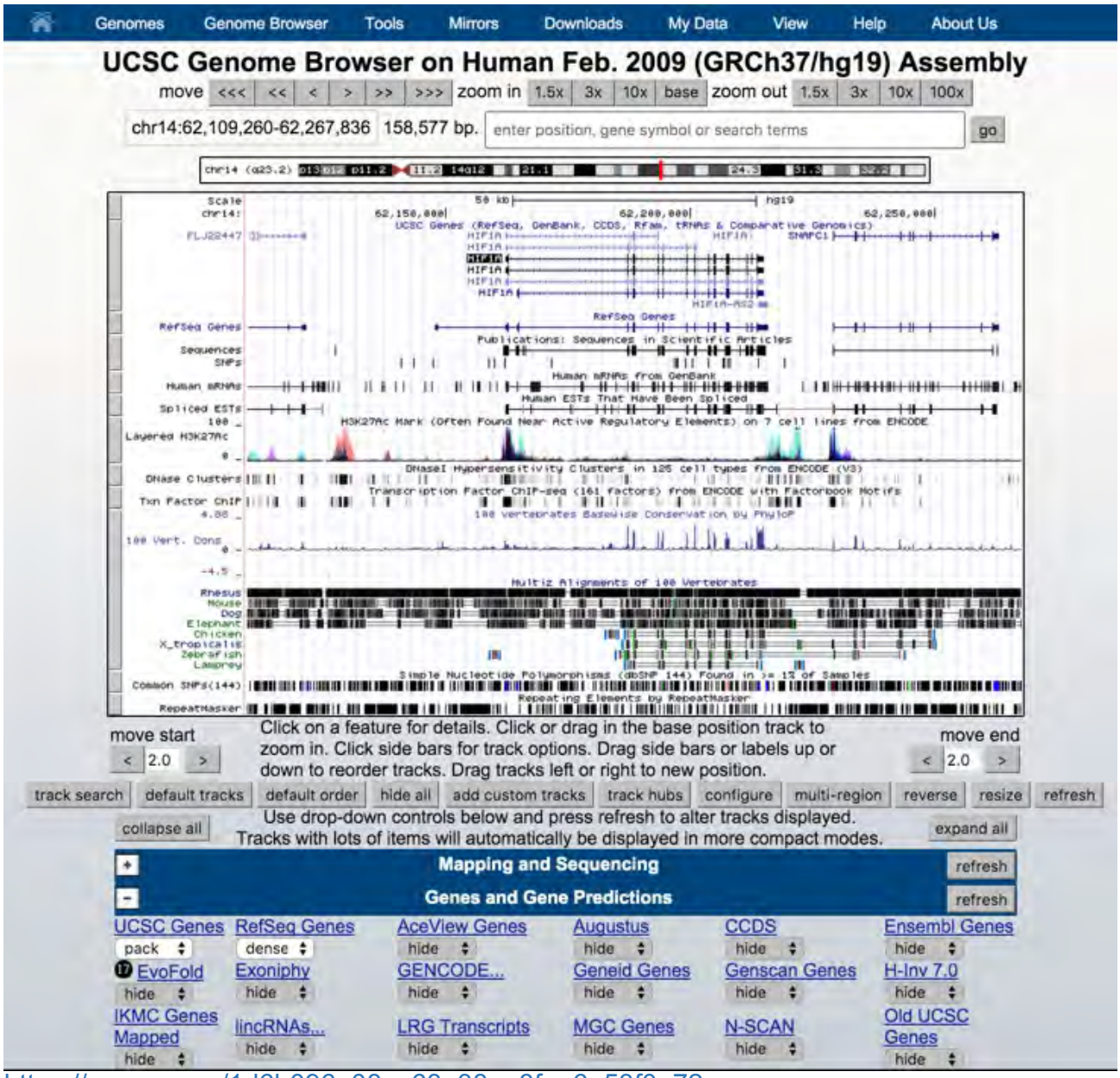
https://gyazo.com/6ede06a7f0a4ce62ac054038b6624857

4. HIF1Aがコードされたゲノム上の領域が表示されます



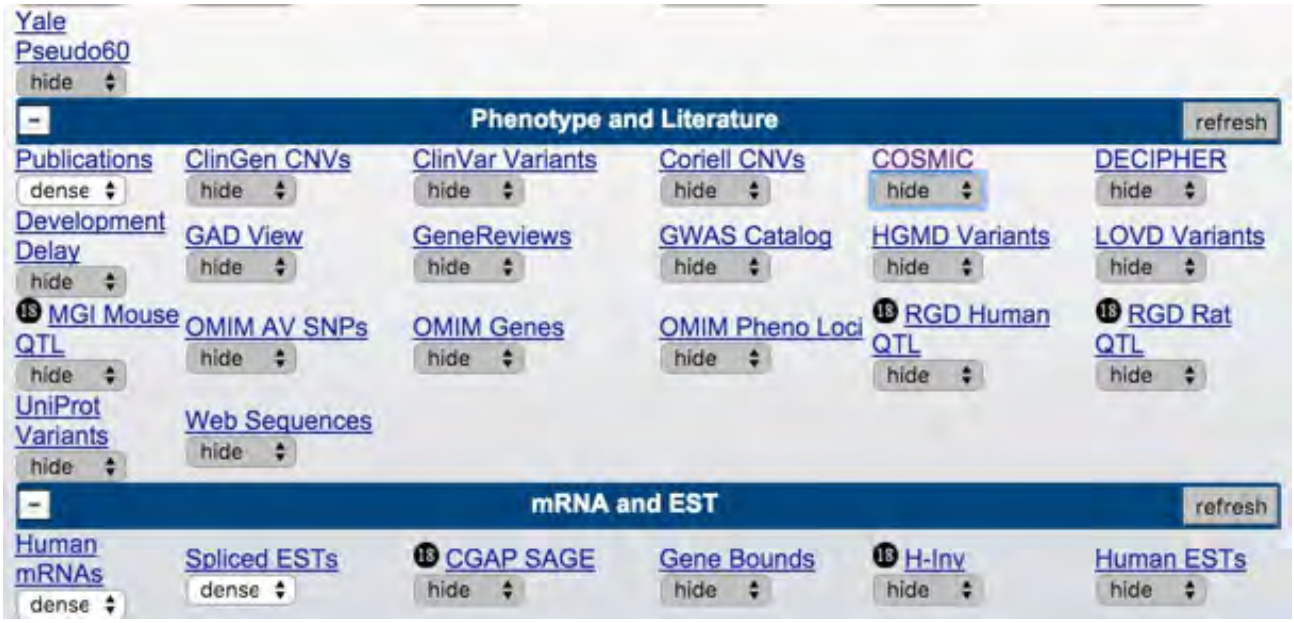
https://gyazo.com/bdbbb1a3cb3835e4db052b0c5843528f これがゲノムブラウザの基本画面です。このページにあるさまざまなボタンなどいろいろと操作して必要な情報を追加したり、見たくない情報を削除したりします。

5. 上部のnavigationボタンでmoveやzoom in/out等できますが、遺伝子名検索でたどり着いた場合、mRNAの領域に拡大されて表示されるので、zoom out `3x` しておきましょう



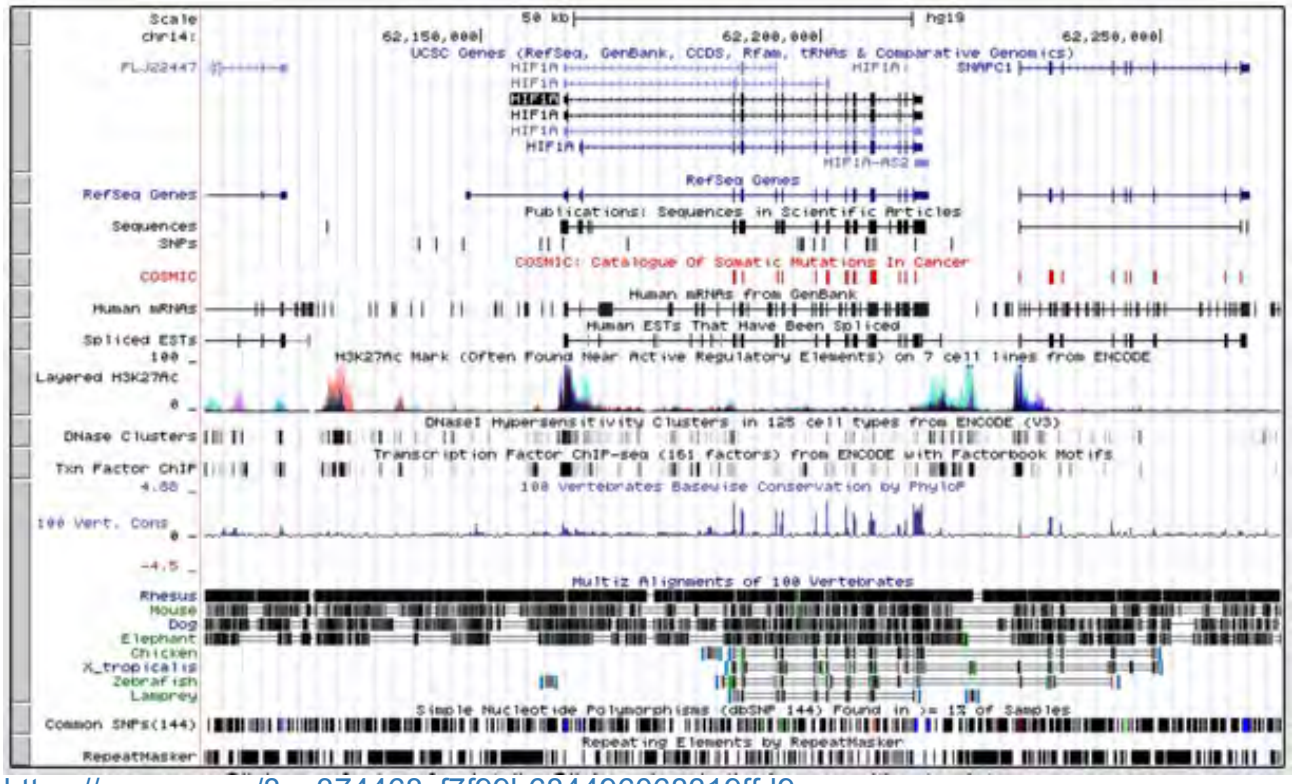
https://gyazo.com/1d6b096e08ec68e38ee8fae6e53f0e72

6. 画面下の方にあるのがアノテーションです。Phenotype and Literature カテゴリー中の `COSMIC` が'hide'になっているのを `dense` に変えて、 `refresh` ボタンを押してみましょう



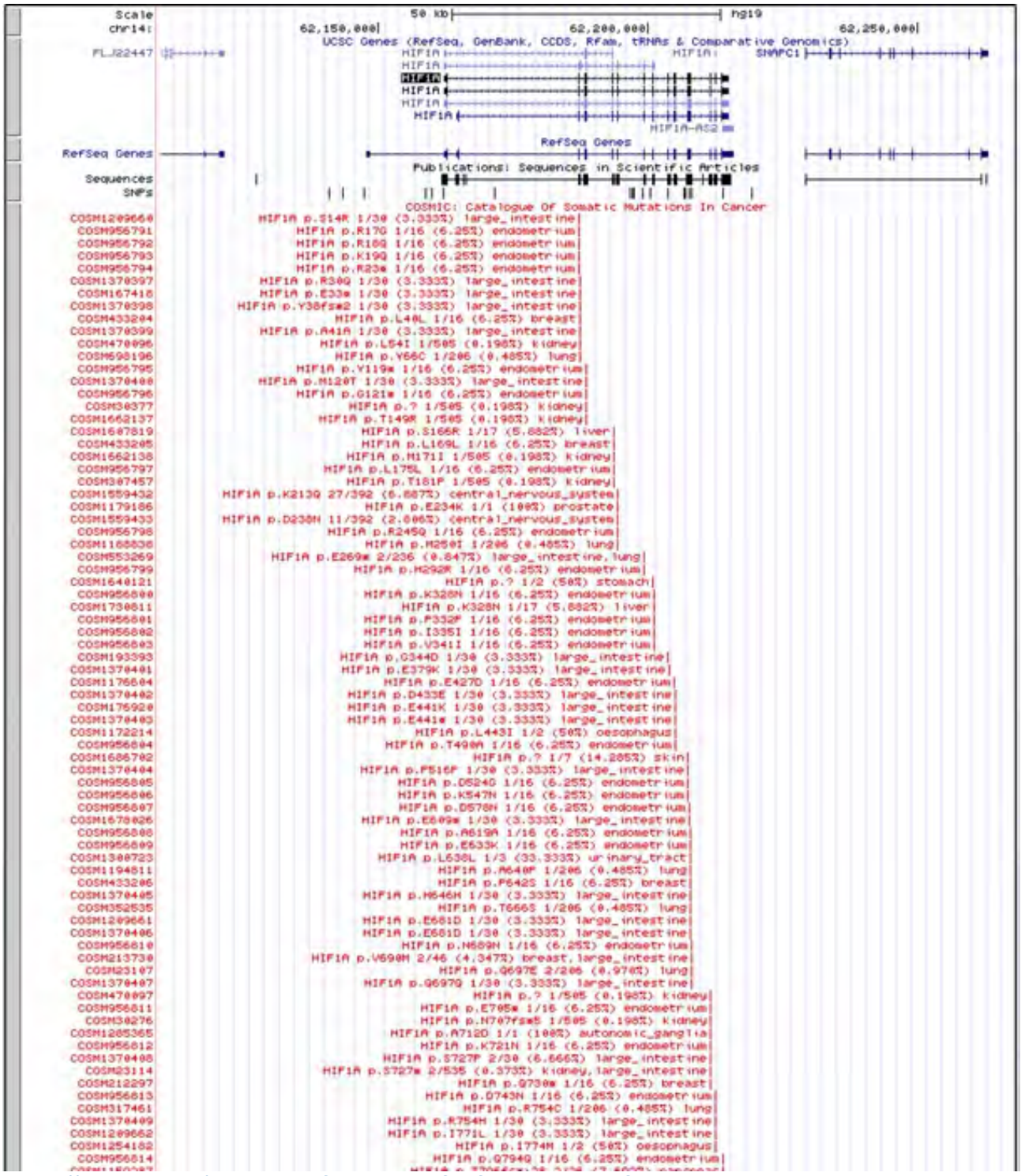
https://gyazo.com/a4130a4abc68b93803584ce5aa2ff8bb

7. そうすると、上部のゲノム領域にこのゲノムアノテーション(COSMIC)が付加されて表示されます(画像中央、赤色)



https://gyazo.com/8ae374468cf7f99b684496068319ffd9

8. dense,squich, pack, fullとなるに連れて、情報量が多くなります。`full`にしてみると…

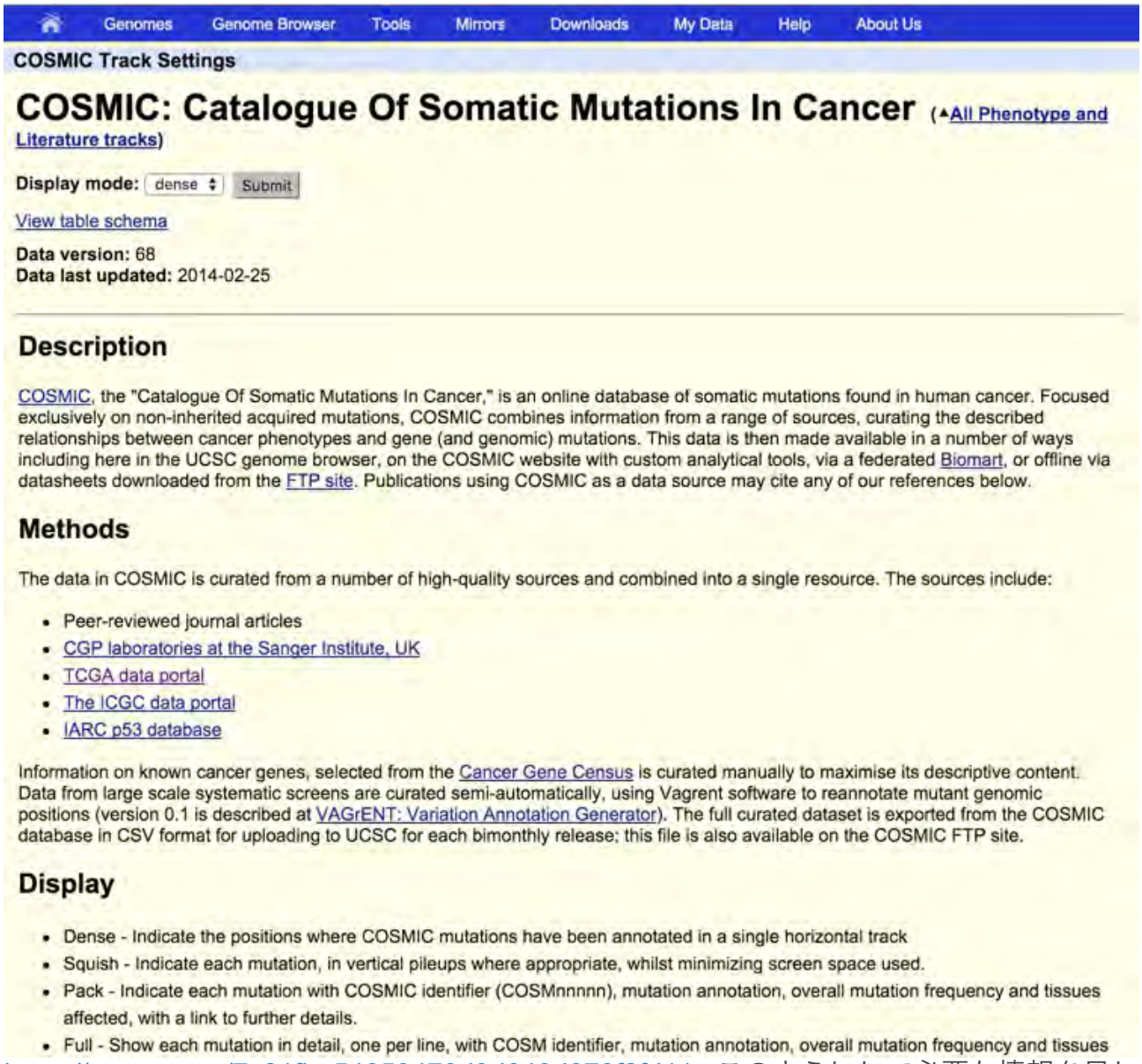


https://gyazo.com/4375db31f539aba1af1af1777546123f

9. この図は見たい情報のところをクリックすると詳細情報が得られます。たとえば、COSMICの気になる部分のIDをクリックすると、その個別のアノテーションの詳細情報が見れます。

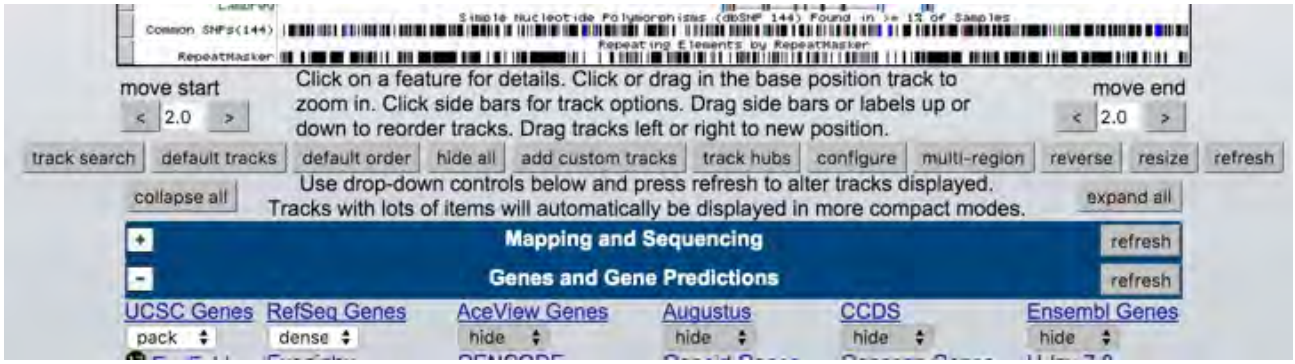
https://gyazo.com/2b52bd8f59a6b4256f1d069c3b69eff5

10. このゲノムアノテーションそのものについて詳しく知りたい場合、さきほどshowに切り替えた選択画面の上にあったリンクをクリックしてみましょう。詳しい説明が得られます



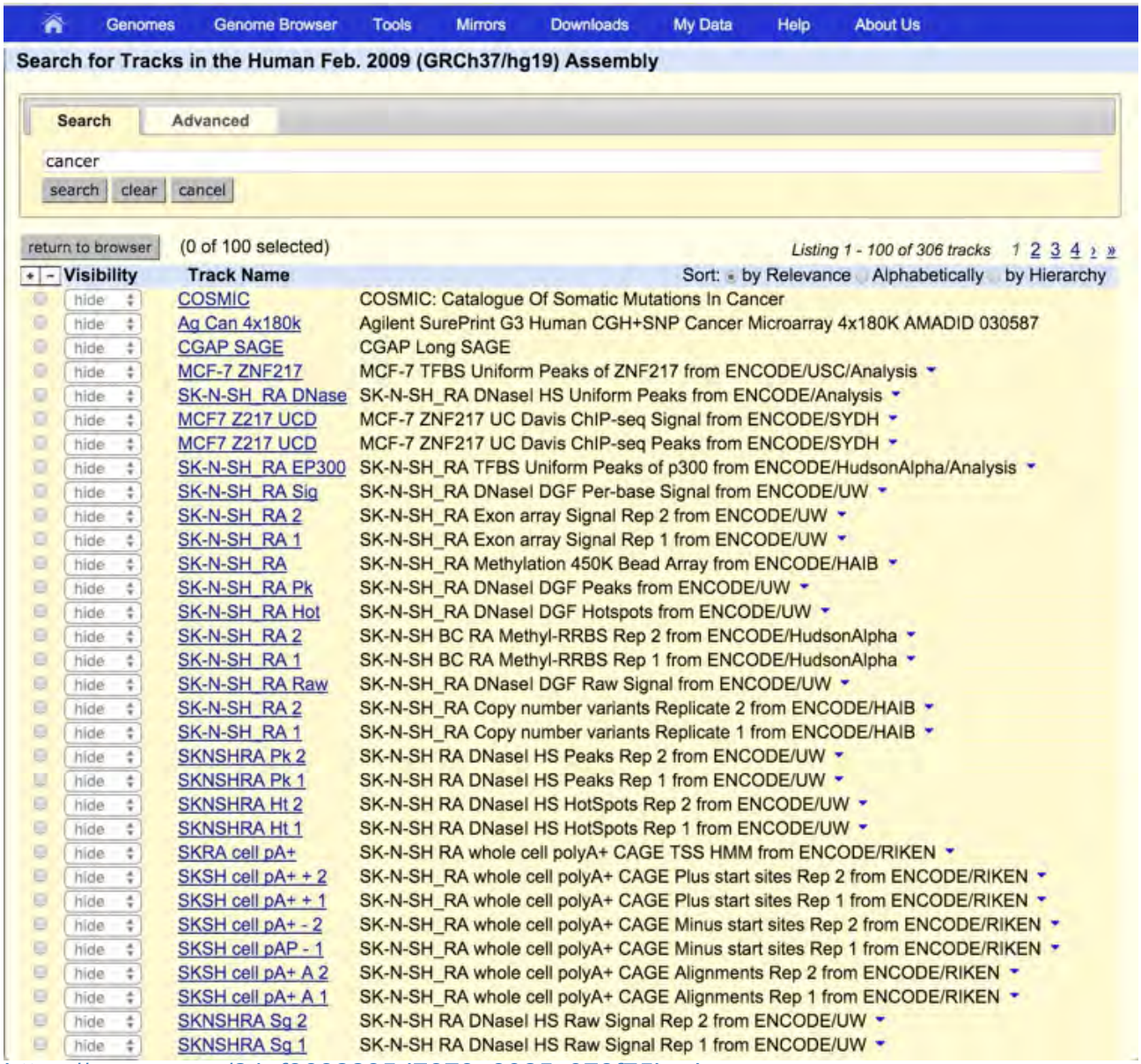
https://gyazo.com/7c21fba54853479d3d648d878f2fd11c このようにして必要な情報を足していって、自分のほしい情報を得ます。

11. いろいろいじってしまうと元に戻したい時があります。その場合は、`default tracks` ボタンを押すとResetされ、元のゲノムアノテーションに簡単に戻せます



https://gyazo.com/e37518349806c036f070a079b6dfa9cb

12. このページにあるゲノムアノテーションの検索は、その左の `track search` ボタンから可能です。*cancer*をキーワードに検索してみると…



https://gyazo.com/24ef8682285d7879a8025c078f75bede

【発展2】 さらに、UCSC Genome Browserのサーバー上ではなく、外部で管理されているデータをゲノムブラウザ上に表示することもできます。上部のメニューの My Data の中にあるTrack Hubsをクリックした画面で `cancer` で検索すると以下の様な外部データが利用可能とわかります。

https://gyazo.com/0b434ffd362d5f9a846c1de2632a2c4c

```
- 【復習用 統合TV】[UCSC Genome Browserの使い方〜配列取得編〜2013](http://doi.org/10.7875/togotv.2013.087)
- [【復習用】UCSC Genome Browserの使い方〜表示+ENCODE編〜 2012(統合TV)](http://togotv.dbcls.jp/ja/20120528.html)
```

- Ensembl Genome Browser http://www.ensembl.org/ http://www.ensembl.org/
- IGV(Integrative Genomics Viewer) https://www.broadinstitute.org/igv/ https://www.broadinstitute.org/igv/
    - 【復習用 統合TV】 Integrative Genomics Viewer IGVを使い倒す 〜基本編〜 http://doi.org/10.7875/togotv.2014.027
    - 【復習用 統合TV】 Integrative Genomics Viewer IGVを使い倒す 〜マッピングデータを可視化する〜 http://doi.org/10.7875/togotv.2014.038

配列レポジトリ

- DBCLS SRA http://sra.dbcls.jp/ http://sra.dbcls.jp/
- AOE http://aoe.dbcls.jp/ http://aoe.dbcls.jp/

数値データの共有化: 公共データレポジトリ

- Figshare http://figshare.com/ http://figshare.com/
- DRYAD http://datadryad.org/ http://datadryad.org/

数値データの可視化

1. パイチャート(pie chart)

【実習2】Rを起動し以下のファイルに記述されたRのコマンド(`01piechart.r`)を実行しなさい。#(シャープ)から始まる行は、コメント行で、プログラムの実行には関係ありません(=入力しなくてよい)。データファイルとして必要な `srabystudy.txt` をgithubのサイトからダウンロードし、現在作業しているディレクトリにおいておく必要があります。実行後、そのディレクトリに `pie1.png` というファイルが新たに生成されていることを確認しなさい。

```
#出力する画像のファイル名を指定します
png("pie1.png")
#タブ区切りのデータを読み込みます
dat <- read.delim2("srabystudy.txt", header=F)
#columnにお名前を
names(dat) <- c("Study", "freq")
#列を結合
dat <- cbind(dat, serial=seq(dim(dat)[1]))
#値が0じゃないデータだけを使います
dat2 <- dat[ dat$freq != 0, ]
#パイチャートを描きます
pie(dat2$freq, labels=paste(dat2$Study, dat2$freq),
col=rainbow(dim(dat[1])[dat2$serial]), main="SRA by Study")
#画像を完成させるおまじない。忘れずに!
dev.off()
```

先ほど説明のあったSRA(Sequence Read Archive)のメタデータを取りまとめてStudyのタイプで分けた結果がこのデータ(`srabystudy.txt`)で、それを元にpie chartを描いてみるのが本課題でした。

【発展3】上記の `srabystudy.txt` を別のデータに変えて実行してみましょう。

- `srabystudy_orig.txt` : SRAのStudyの分類をoriginalデータ

- `srabyorganism.txt` : SRAに登録されている生物種で分類したデータ

> これらの結果は、お配りしたファイル群の `results` というフォルダの中においてあります(それぞれ、 `pie1.png` , `pie2.png` , `pie3.png`)。実は、このスクリプトはBono H et al. Genome Res 2003 http://genome.cshlp.org/content/13/6b/1318.fullの図1を作成するために使ったものです。実際には、このRコードを生成するPerlのプログラムを作成してからそれをRで実行していました。現在よく使われているGSEA(Gene Set Enrichment Analysis)のハシリで、遺伝子にアノテーションされたGene Ontologyの高次(根っこに近い)タームで集計してその分布を見ていました

2. 階層的クラスタリング(hierachical clustering)

【実習3】Rを起動し以下のファイルに記述されたRのコマンド(`02hclust.r`)を実行しなさい。データファイルとして必要な `matrix.txt` をgithubのサイトからダウンロードし、現在作業しているディレクトリにおいておく必要があります。実行後、そのディレクトリに `hclust.png` というファイルが新たに生成されていることを確認しなさい。

```
#出力する画像のファイル名を指定します
png("hclust.png")
#データを読み込みます(スペース区切り)
d <- read.table('matrix.txt')
#UPGMA法で階層的クラスタリングを実行
c <- hclust(as.dist(d), method = 'average')
#結果をプロット
plot(c, hang=-1)
#画像を完成させるおまじない。忘れずに!
dev.off()
```

このプログラムはかつての蛋白質・核酸・酵素に寄稿したレビューで紹介したもので、当時階層的クラスタリングをフリーウェアで実行する手段として重宝された(はず)。Bono H and Nakao MC, PNE 2003 http://www.ncbi.nlm.nih.gov/pubmed/12638180

3. 主成分分析(PCA)

R/Bioconductorのaffyパッケージで発現データを定量化して数値データにして、それを使ってPCAを実行します。

Affymetrixデータ正規化

RMAの計算が重く、パソコンによってはメモリ不足となり実行不可能となることが予想されますので、できれば。

【参考】ぼうのブログ: justRMAでnormalize http://bonohu.jp/blog/2013/06/17/justrma/

【発展4】Bioconductorを利用します。Rを起動して以下のコマンドでaffy libraryをインストールしなさい。

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

そして、affy library中のjustRMA関数を利用して、RMA(Robust Multichip Average)正規化してみましょう。自らのAffymetrix Genechip データ(CELファイル)がある人はそれを、ない人は `GSE17264` Comparative transcriptome analysis of dedifferentiation in porcine mature adipocytes and follicular granulosa cellsを例に実行しましょう。GEO(`GSE17264`) https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17264もしくはArrayExpress(`E-GEOD-17264`) https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-17264/から生データ(CELファイル)をダウンロードできます。 実行する際、Session → Set Working Directory を、CELファイルをダウンロードしてきたディレクトリに指定することがポイントです。それができたら、以下のコード(`03justrma.r`)を実行します。

```
#Bioconductorを使うときの呪文
source("http://bioconductor.org/biocLite.R")
#affyライブラリをインストール
biocLite("affy")
#affyライブラリを召喚
library(affy)
#justRMAを実行、RMA.txtというファイルに出力
write.exprs(justRMA(), file="RMA.txt")
```

その結果生成される `RMA.txt` ファイルがRMAによって正規化された遺伝子発現データになります。

【発展5】上記の実習のサンプルデータは **GPL3533 [Porcine] Affymetrix Porcine Genome Array** と呼ばれる(カタログ)マイクロアレイ(プラットフォーム)を使ったデータでした。同じマイクロアレイ(プラットフォーム)のデータであればjustRMAを実行することが出来ます。同じプラットフォームのデータの中からiPS細胞のものを探しだして上記のデータに混ぜてjustRMAを実行しなさい。

> 解答例: GSE15472 Induced Pluripotent Stem Cells from the Pig Somatic Cells http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15472 の3つのサンプル(GSM388154,GSM388155,GSM388156)がそれです。 CELファイルをダウンロードして同じdirectoryに置き、再度justRMAを実行してみましょう

> このマイクロアレイデータは以下の論文で我々が発表したもので、全く同じ手段で正規化を行いました。また、発展課題にある公共遺伝子発現データを足したデータ解釈はやはりこの論文で実際に発表したもので、我々のマイクロアレイデータの生物学的解釈(DFATがMA(Mature AdipocyteよりもiPSに近い)に役だっています。 Ono H, Oki Y, Bono

> H, Kano K. BBRC 2011 http://www.ncbi.nlm.nih.gov/pubmed/21419102

主成分分析(PCA)

PCA(Primary Component Analysis)。 このパートはぼうのブログ「RでPCA」
http://bonohu.jp/blog/2013/07/13/pcaonr/を参考に。

【実習4】Rを起動し以下のファイルに記述されたRのコマンド(`03pca.r`)を実行しなさい。データファイルとして必要な `RMA.txt` は【発展4】で作成したものを利用すればよいのですが、現在作業しているディレクトリにおいておく必要があります。
この発展課題をやらなくても実行できるように用意してあり、 `RMA.txt.zip` というファイルをダブルクリックすることで解凍(圧縮を解く)すると `RMA.txt` というファイルが生成されるようになっています。
PCAを実行するには、以下のRスクリプトを実行します。

```
#タブ区切りのデータを読み込み
data <- read.table("RMA.txt", header=TRUE, row.names=1, sep="\t",
quote="")
#主成分分析を実行
data.pca <- prcomp(t(data))
#名前を得る
names(data.pca)
#標準偏差をプロット
plot(data.pca$sdev, type="h", main="PCA s.d.")
#行列を転置
data.pca.sample <- t(data) %*% data.pca$rotation[,1:2]
#PCAの結果をプロット
plot(data.pca.sample, main="PCA")
#ラベルに色付け
text(data.pca.sample, colnames(data), col = c(rep("red", 3),
rep("blue",3),rep("green",3),rep("black",3)))
```

図に.CEL.gzの文字がいっぱい入っていて見づらい?そう思った方はこちらを参照
http://bonohu.jp/blog/2014/09/10/afterjustrma/

4. ヒートマップ(cuffdiffの結果可視化)

RNA-seqデータ解析プログラムの1つであるcufflinksパッケージに `cuffdiff` という発現データの差分を計算するプログラムがあります。それを実行したあとのデータを可視化する手段として使うBioconductorのパッケージに `cummeRbund` があります(cuffdiffの使用例
http://dx.doi.org/10.1371%2Fjournal.pone.0104283)。

- 参考: bonohuの発表資料「NGS解析(RNAseq)」
  http://dx.doi.org/10.6084/m9.figshare.1216717

【発展6】Rを起動し以下のファイルに記述されたRのコマンドを実行しなさい。データファイル

として必要なcuffdiffディレクトリ以下のファイルは手持ちのものか、なければサンプルをUSBメモリでお渡しします。現在作業しているディレクトリにおいておく必要があります。

```
#Bioconductorを使うときの呪文
source("http://bioconductor.org/biocLite.R")
#cummeRbundをインストール
biocLite("cummeRbund")
#cummeRbundを召喚
library("cummeRbund")
#cuffdiffの結果のディレクトリを指定
cuff.dir <- "cuffdiff"
#cuffdiffの結果の読み込む
cuff <- readCufflinks(dir=cuff.dir)
```

準備はここまでで、

```
cuff
```

で、読み込んだデータのサマリが見れます。以下、解析事例です。

```
#発現密度分布のプロット
dens <- csDensity(genes(cuff))
dens

# サンプル方向のデンドログラム
dend <- csDendro(genes(cuff))

#CSV形式でFPKM値を出力
gene.matrix <- fpkmMatrix(genes(cuff))
write.csv(gene.matrix, file="fpkm.csv")
```

詳しくは、マニュアルを読みましょう。

```
?cummeRbund
```

ビジネスインテリジェンス(BI)ツールの活用

- Spotfire http://spotfire.tibco.jp/ http://spotfire.tibco.jp/
- Yellowfin http://yellowfin.jp/ http://yellowfin.jp/
- Tableau http://www.tableau.com/ http://www.tableau.com/