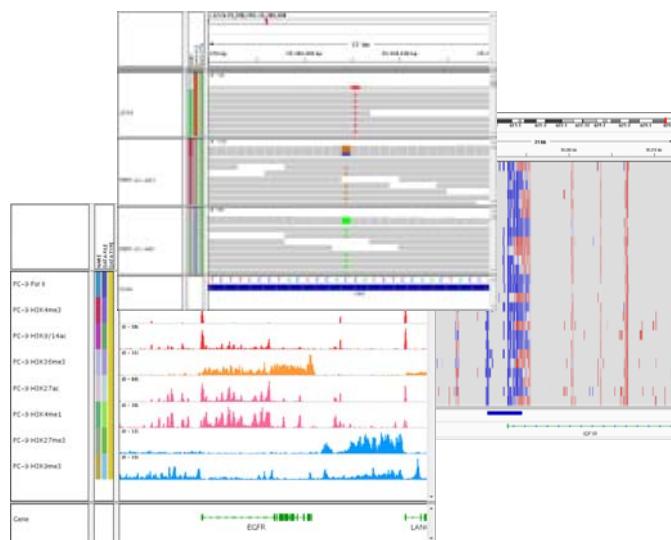


2015.08.04 AJACS米子

# 次世代シークエンサーを用いた がん細胞のオミクス解析



国立がん研究センター 先端医療開発センター  
トランスレーショナルリサーチ分野  
鈴木 純子

# 本日の予定

## Session 1

シークエンスタグのマッピングと可視化  
(RNA-Seq)

## Session 2

データベースDBTSSの紹介

# 東京大学 医科学研究所 ヒトゲノムセンター スーパーコンピュータ

<https://supcom.hgc.jp/japanese/>

Human Genome Center  
the Institute of Medical Science, the University of Tokyo

企業の方もスパコンを利用できます  
利用料金の詳細

トップページ 解析ツール データベース ダウンロード スーパーコンピュータ HGC とは

TOP > スーパーコンピュータ

## スーパーコンピュータ

アカウントを取得された方は、まず[こちら](#)をご覧ください。

### ニュース

- 2012年09月04日 [Genomon-fusion を公開しました](#)
- 2012年06月18日 [Shirokane2 が TOP500 で 183 位になりました](#)
- 2012年05月11日 [Genomon を公開しました](#)

### お知らせ

- 2014年09月04日 [遺伝子ネットワーク解析実習講習会 9月 26 日開催](#)
- 2014年08月11日 [GeneCards バージョンアップ](#)
- 2014年05月27日 [統計処理言語 R 講習会 6月 24,25 日開催](#)
- 2014年04月24日 [HGC スパコンユーザー会 5月 20 日開催](#)
- 2014年04月15日 [HGC 解析ツール講習会のお知らせ \(2回目\) 2014年 5月 13 日開催](#)
- 2014年03月28日 [HGC 解析ツール講習会のお知らせ \(1回目\) 2014年 4月 22 日開催](#)

ツイート

フォローする

9月25日

Supercomputer@HGC @schgc share3 障害に対する戻し作業は終了しました。

9月25日

Supercomputer@HGC @schgc 昨日に発生した share3 の障害に対する戻し作業を本日 16 時より実施します。この作業により、一時的に share3 の一部の領域へのアクセスができなくなります。

お問い合わせ



スパコンサポート係の方に  
ご助力いただきました。ありがとうございます。

[https://supcom.hgc.jp/japanese/sys\\_const/system-main.html](https://supcom.hgc.jp/japanese/sys_const/system-main.html)

# スペコンのアカウント

ユーザ名: lect-2 ~ lect-50  
パスワード: \*\*\*\*\*

今から一人一人にアカウントを割り当てます。  
自分のユーザ名をメモしてください。

パスワードは講習会当日にお知らせいたします。

パスワードは変更しないでください。  
作業は、自分のホームで行ってください。

# 本日用いるソフトウェアの準備

## ターミナルソフト TeraTerm

- Windows PCからLinux環境のマシンへアクセスするのに必要  
※Macはターミナルでリモートアクセスできます  
※Cygwinをインストール済みの方も不要です
- インストールは基本的にデフォルトの設定で問題ありません

<http://osdn.jp/projects/ttssh2/releases/63335>

## ファイル転送ソフト WinSCP

Windows PC←→Linux環境のデータのやり取りに必要

※Macはターミナルでデータのアップロード/ダウンロードできます  
※Cygwinインストール済みの方は不要です

<http://winscp.net/eng/download.php>

## ゲノムビューアー IGV (Integrative Genomics Viewer)

リファレンスゲノム上にアライメントされたシークエンスタグを可視化するツール

[ for Win ]

[http://dbtss.hgc.jp/cgi-bin/downloader2.cgi/IGV\\_2.3.32.zip](http://dbtss.hgc.jp/cgi-bin/downloader2.cgi/IGV_2.3.32.zip)

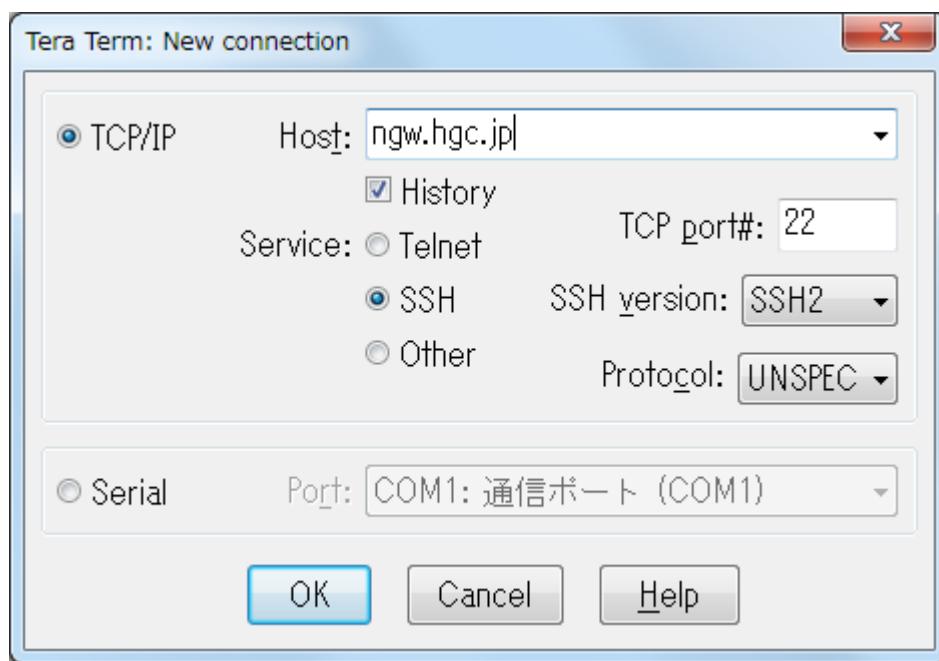
[ for Mac ]

[http://dbtss.hgc.jp/cgi-bin/downloader2.cgi/IGV\\_2.3.32.app.zip](http://dbtss.hgc.jp/cgi-bin/downloader2.cgi/IGV_2.3.32.app.zip)

すでにインストールされています。起動するか確認してください。

# Windows PCから解析サーバへのログイン①

- TeraTermを開きます。
- 接続先の設定画面にて、ホストのボックスにIPアドレスngw.hgc.jpを入力し、OKをクリックします。



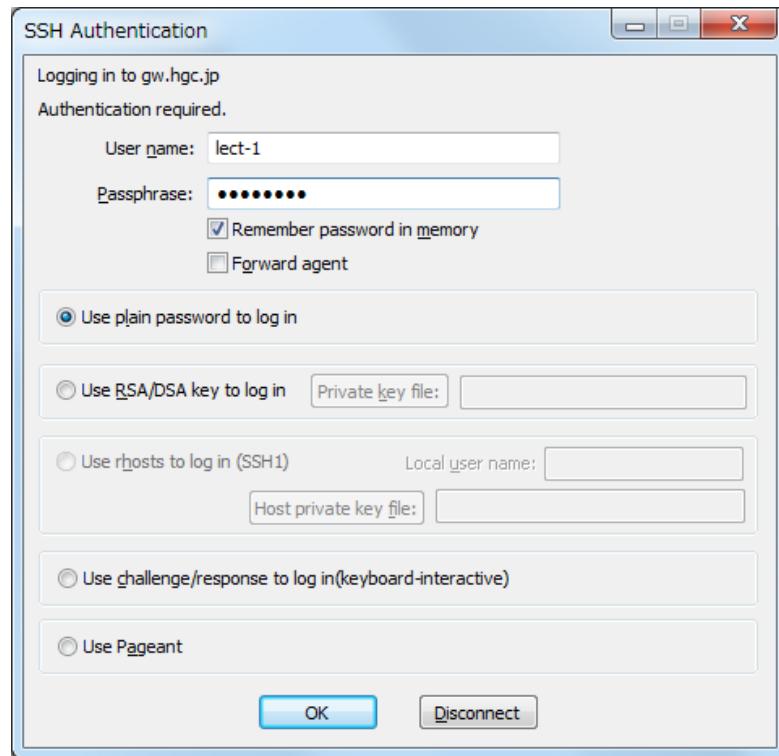
# Windows PCから解析サーバへのログイン②

- ユーザ名とパスワードを入力します。

ユーザ名: lect-2 ~ lect-50  
パスワード: \*\*\*\*

自分のユーザ名(lect-2 ~ lect-50)を入れてください。

- OKを押します
- 正しくログインできると、  
[lect-1@xxxxx ~]\$  
と表示されます。

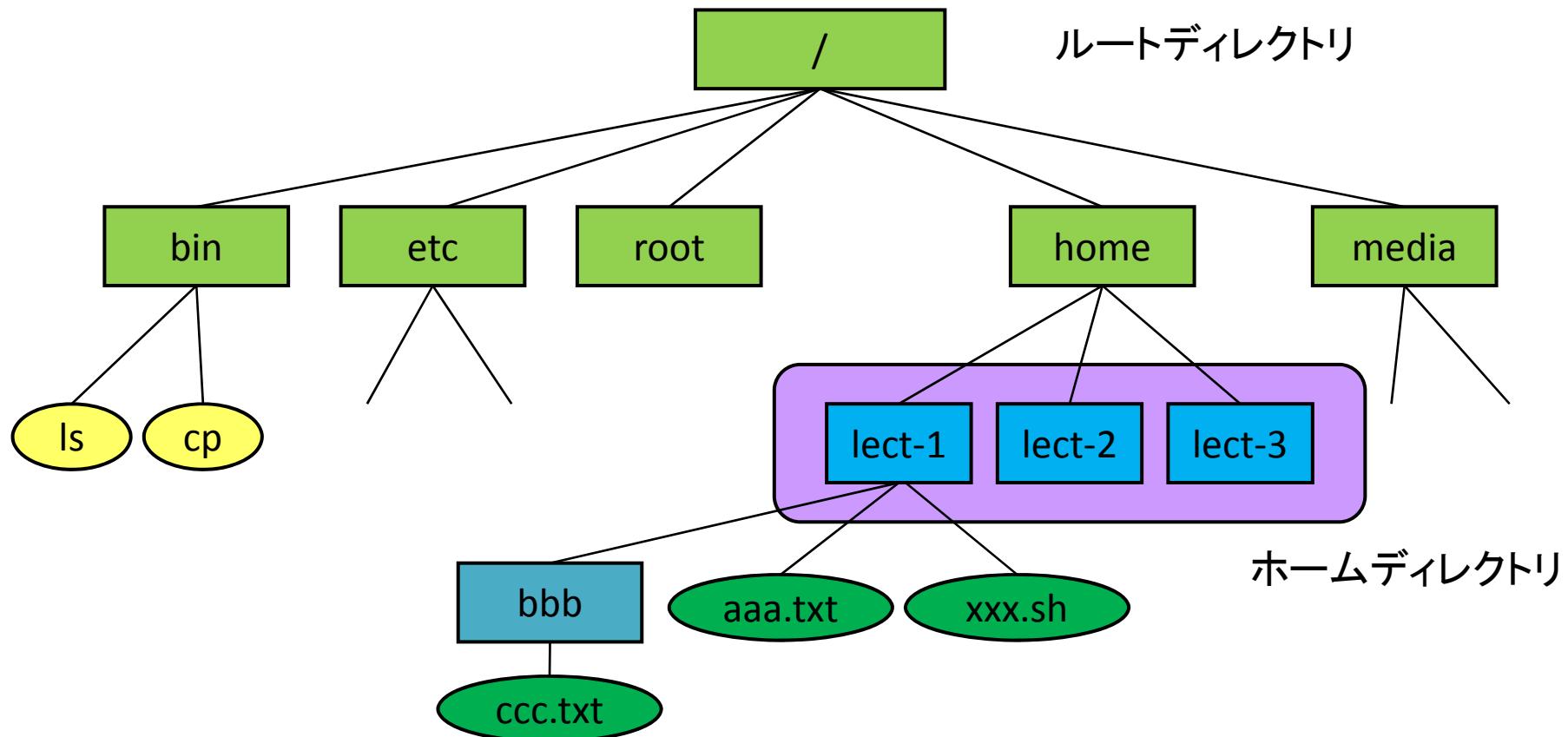


A screenshot of the Tera Term VT window titled 'gw.hgc.jp:22 - Tera Term VT'. The window shows the following text in the terminal:

```
Last login: Thu Oct 2 10:33:30 2014 from 157.82.238.193
[lect-1@sapphire ~]$
```

The window has a standard Windows-style title bar and menu bar (File, Edit, Setup, Control, Window, Help).

# ディレクトリ構造



- ルートディレクトリ: ツリー構造のトップとなるディレクトリ
- ホームディレクトリ: 各ユーザがログインしたときの最初のディレクトリ(/home/[user ID])
- カレントディレクトリ: 現在、作業しているディレクトリ

# ディレクトリ関係の基本コマンド

- cd: 指定のディレクトリに移動する

- ディレクトリAに移動する場合

```
$ cd A
```

cdのみ入力した場合はホームディレクトリへ移動する。

cd ..と入力した場合は、一つ上のディレクトリに移動する。

- pwd: 現在のディレクトリを表示する

```
$ pwd
```

- ls: 指定のディレクトリ内のディレクトリ、ファイルを表示する

- ディレクトリAのファイルを表示する。

```
$ ls A
```

lsのみを入力した場合は、カレントディレクトリのファイルが表示される。

• -lオプションをつけると詳細表示になり、-tオプションをつけると更新日時でソートされる。

## 【実習】実際にコマンドを入力しましょう

```
$ pwd  
$ ls  
$ ls -lt  
$ cd jishu  
$ pwd  
$ ls  
$ ls -lt  
$ cd  
$ pwd
```

The screenshot shows a terminal window titled 'slogin.hgc.jp:22 - lect-2@ngw04:~ VT'. The window contains the following command history:

```
[lect-2@ngw04 ~]$ pwd  
/home/lect-2  
[lect-2@ngw04 ~]$ ls  
jishu  
[lect-2@ngw04 ~]$ ls -lt  
合計 4  
drwxrwxr-x 2 lect-2 lect 4096 7月 30 08:44 2015 jishu  
[lect-2@ngw04 ~]$ cd jishu  
[lect-2@ngw04 jishu]$ pwd  
/home/lect-2/jishu  
[lect-2@ngw04 jishu]$ ls  
LC2ad fq MAPPING.sh  
[lect-2@ngw04 jishu]$ ls -lt  
合計 127196  
-rwxr--r-- 1 lect-2 lect 354 7月 30 08:44 2015 MAPPING.sh  
-rw-r--r-- 1 lect-2 lect 130238229 7月 27 12:18 2015 LC2ad fq  
[lect-2@ngw04 jishu]$ cd  
[lect-2@ngw04 ~]$ pwd  
/home/lect-2  
[lect-2@ngw04 ~]$
```

# ファイル内容を表示するコマンド

- cat: ファイルの全内容を表示する

```
$ cat a.txt
```

一度に全部表示されるため、大きいサイズのファイルには不向き

- head, tail: ファイルの先頭、末端を表示(デフォルトでは10行)

```
$ head a.txt
```

```
$ tail a.txt
```

100行表示したい場合は-nオプションを付ける

```
$ head -n 100 a.txt
```

- more, less: ファイルの内容をコマ送りで表示する

```
$ more a.txt
```

```
$ less a.txt
```

Enterもしくはspaceキーで進む

lessコマンドは↑↓キーでファイルを自由に見ることができる

qと打つと、more, lessコマンドを途中で中断できる

【実習】実際にコマンドを入力しましょう

```
$ cat /home/lect-1/readme.txt
```

```
$ cat /home/lect-1/news.txt
```

```
$ head /home/lect-1/news.txt
```

```
$ tail /home/lect-1/news.txt
```

```
$ more /home/lect-1/news.txt
```

```
$ less /home/lect-1/news.txt
```

# ファイル、ディレクトリの作成、移動、削除コマンド

- cp: ファイルの複製
  - ファイルa.txtをカレントディレクトリにコピーする場合  
\$ cp a.txt .
  - ファイルa.txtをディレクトリAの中にコピーする場合  
\$ cp a.txt A
- mkdir: 新規にディレクトリを作成する
  - ディレクトリAを作成する  
\$ mkdir A
- mv: ファイルの移動
  - ファイルa.txtをディレクトリAに移動する  
\$ mv a.txt A
- rm: ファイルを削除する
  - ファイルa.txtを削除する  
\$ rm a.txt
  - ディレクトリを削除するときは-rオプションを付ける  
\$ rm -r A

※一度データを削除すると復元することはできません。

【実習】実際にコマンドを入力しましょう

```
$ cd  
$ cp /home/lect-1/readme.txt .  
$ ls -lt
```

```
$ mkdir work  
$ ls -lt
```

```
$ mv readme.txt work  
$ ls -lt
```

```
$ cd work  
$ pwd  
$ ls -lt
```

```
$ rm readme.txt  
$ ls -lt
```

```
$ cd ..  
$ pwd  
$ ls -lt  
$ rm -r work  
$ ls -lt
```

```
gw.hgc.jp:22 - Tera Term VT
File Edit Setup Control Window Help
[lect-2@sapphire ~]$ cd
[lect-2@sapphire ~]$ cp /home/lect-1/readme.txt .
[lect-2@sapphire ~]$ ls -lt
合計 4
-rw-r--r-- 1 lect-2 lect 814 10月 2 20:34 readme.txt
drwxr-xr-x 2 lect-2 lect 4096 10月 2 12:23 jishu
[lect-2@sapphire ~]$
[lect-2@sapphire ~]$ mkdir work
[lect-2@sapphire ~]$ ls -lt
合計 8
drwxr-xr-x 2 lect-2 lect 4096 10月 2 20:34 work
-rw-r--r-- 1 lect-2 lect 814 10月 2 20:34 readme.txt
drwxr-xr-x 2 lect-2 lect 4096 10月 2 12:23 jishu
[lect-2@sapphire ~]$
[lect-2@sapphire ~]$ mv readme.txt work
[lect-2@sapphire ~]$ ls -lt
合計 8
drwxr-xr-x 2 lect-2 lect 4096 10月 2 20:34 work
drwxr-xr-x 2 lect-2 lect 4096 10月 2 12:23 jishu
[lect-2@sapphire ~]$
[lect-2@sapphire ~]$ cd work
[lect-2@sapphire work]$ pwd
/home/lect-2/work
[lect-2@sapphire work]$ ls -lt
合計 0
-rw-r--r-- 1 lect-2 lect 814 10月 2 20:34 readme.txt
[lect-2@sapphire work]$
[lect-2@sapphire work]$
```

ディレクトリ/home/lect-1にあるreadme.txtを  
カレントディレクトリ(ホームディレクトリ)にコピー

ディレクトリworkを作成し、  
readme.txtをディレクトリworkに移動

ディレクトリworkに移動する

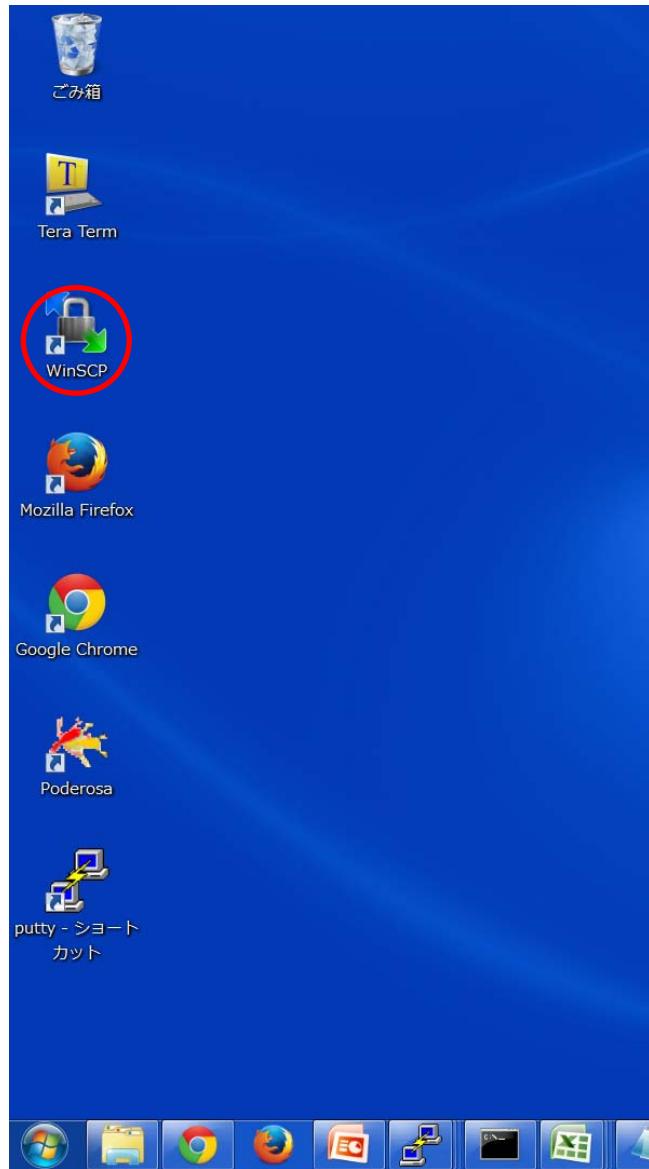
ディレクトリworkに移動させた  
readme.txtを削除

ディレクトリworkを削除

```
gw.hgc.jp:22 - Tera Term VT
File Edit Setup Control Window Help
[lect-2@sapphire work]$ ls -lt
合計 4
-rw-r--r-- 1 lect-2 lect 814 10月 2 20:34 readme.txt
[lect-2@sapphire work]$
[lect-2@sapphire work]$ rm readme.txt
[lect-2@sapphire work]$ ls -lt
合計 0
[lect-2@sapphire work]$
[lect-2@sapphire work]$ cd ..
[lect-2@sapphire ~]$ pwd
/home/lect-2
[lect-2@sapphire ~]$ ls -lt
合計 8
drwxr-xr-x 2 lect-2 lect 4096 10月 2 20:36 work
drwxr-xr-x 2 lect-2 lect 4096 10月 2 12:23 jishu
[lect-2@sapphire ~]$ rm -r work
[lect-2@sapphire ~]$ ls -lt
合計 4
drwxr-xr-x 2 lect-2 lect 4096 10月 2 12:23 jishu
[lect-2@sapphire ~]$
```

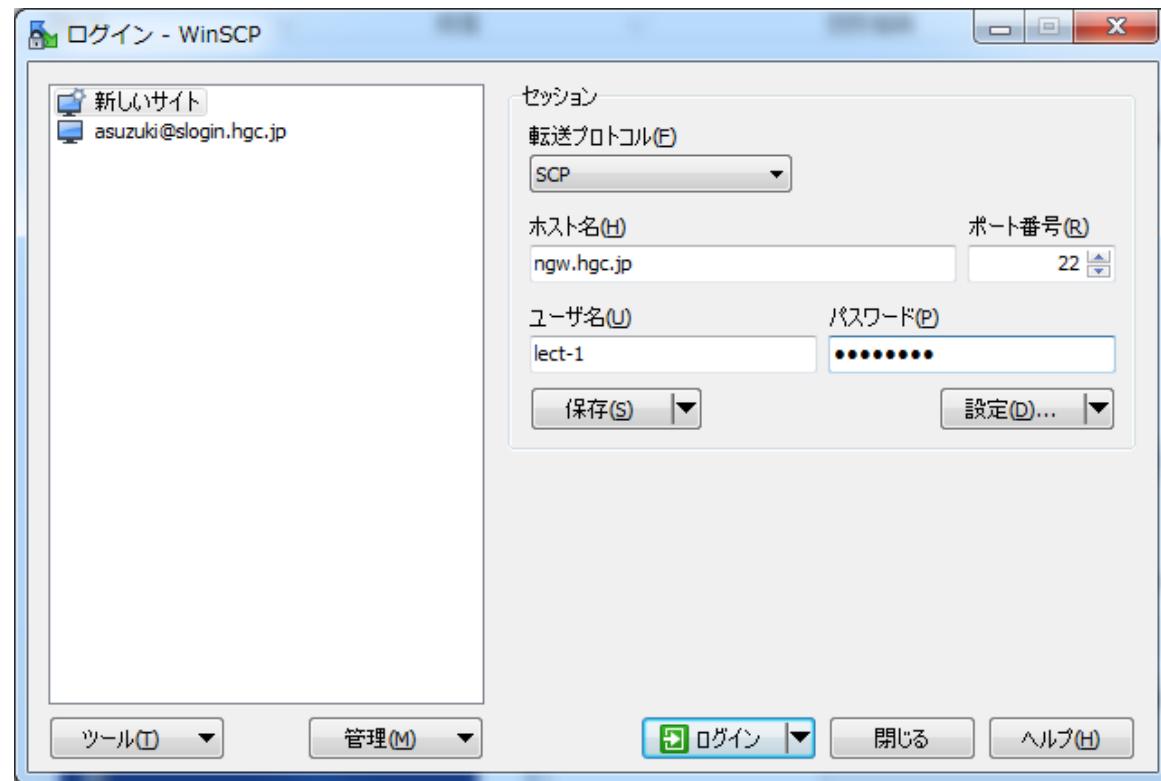
一つ上のディレクトリ  
(ホームディレクトリ)に戻る

# 解析データのWindows PCへのダウンロード



WinSCPを開き、以下の情報を入力します

- ホスト名: **ngw.hgc.jp**
- ユーザ名: **lect-2 ~ lect-50**
- パスワード: **\* \* \* \***

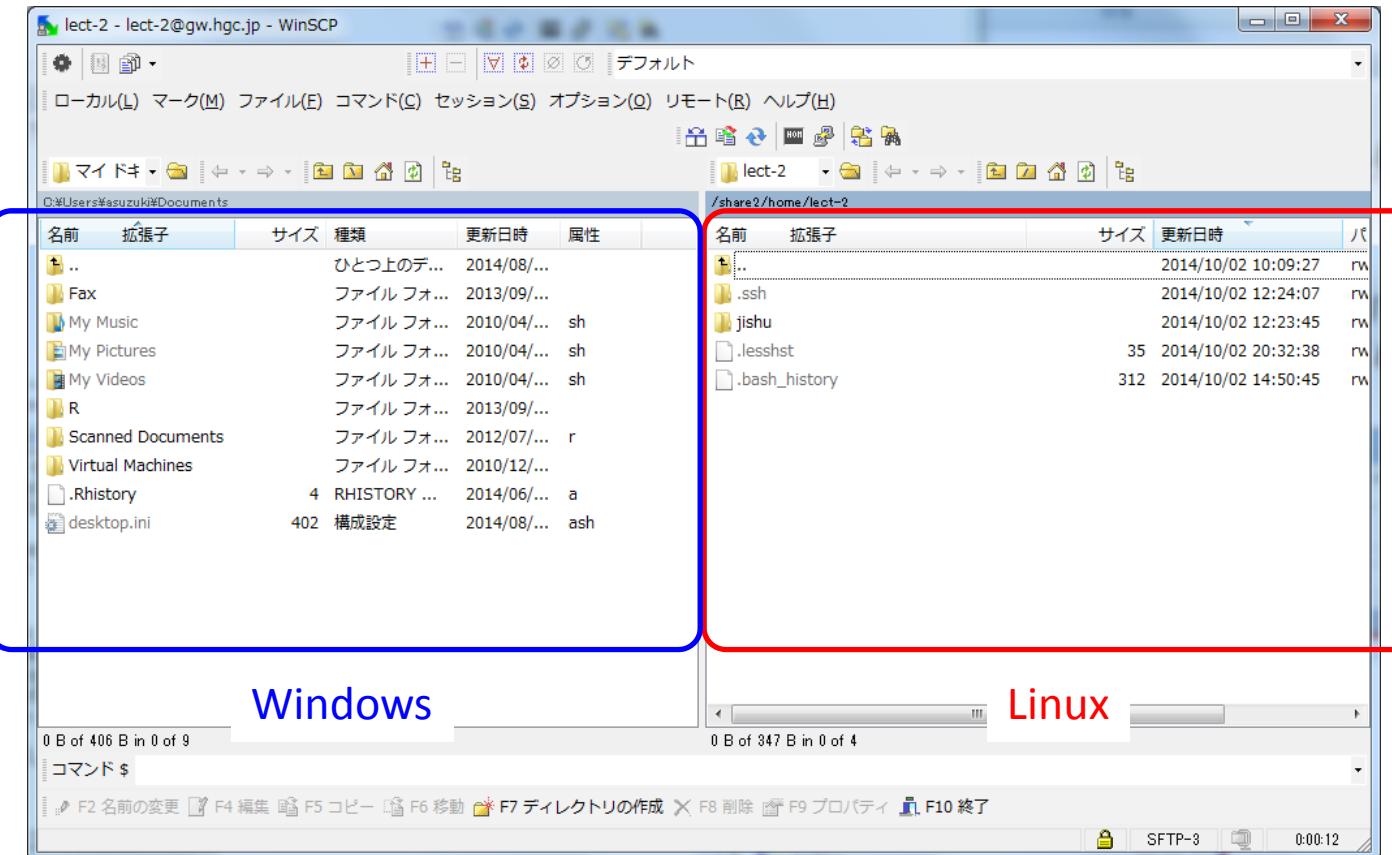


# 解析データのWindows PCへのダウンロード

LinuxからWindows PCへのデータコピー

後で使用します。

- 下図のようなウィンドウが表示されます。  
左側がWindows PC、右側が解析サーバ(Linux)です。
- ファイルあるいはフォルダをクリックし、ドラッグ & ドロップを実行すると、  
データコピーが始まります。

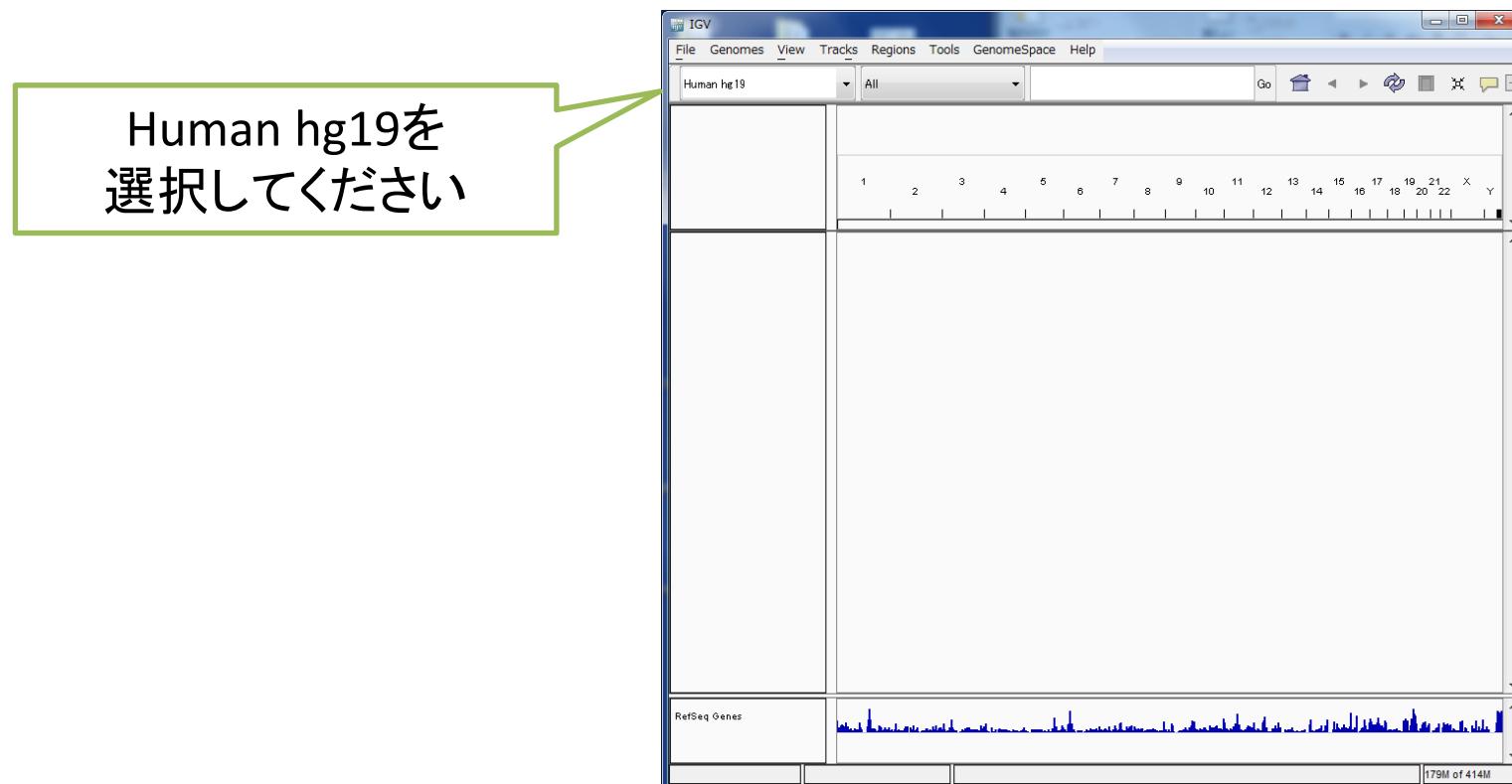


# シークエンス可視化

## Integrative Genomics Viewer(IGV)

シークエンスデータの可視化に使用します。

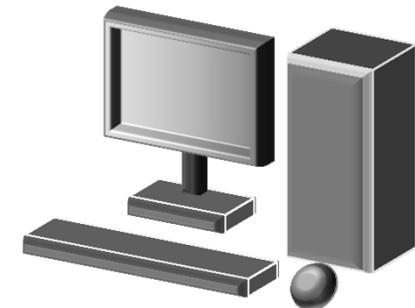
IGVを起動してみてください。



あとで使用しますので、うまく起動しない方や"Human hg19"がない方がいましたらお知らせください。

# Session 1

シークエンスタグのマッピングと可視化  
(RNA-Seq)



## 課題1

ホームディレクトリに入り、

```
$ qlogin -l s_vmem=5G -l mem_req=5
```

```
$ pwd
```

```
$ cd jishu
```

```
$ ls
```

```
$ /bin/sh MAPPING.sh
```

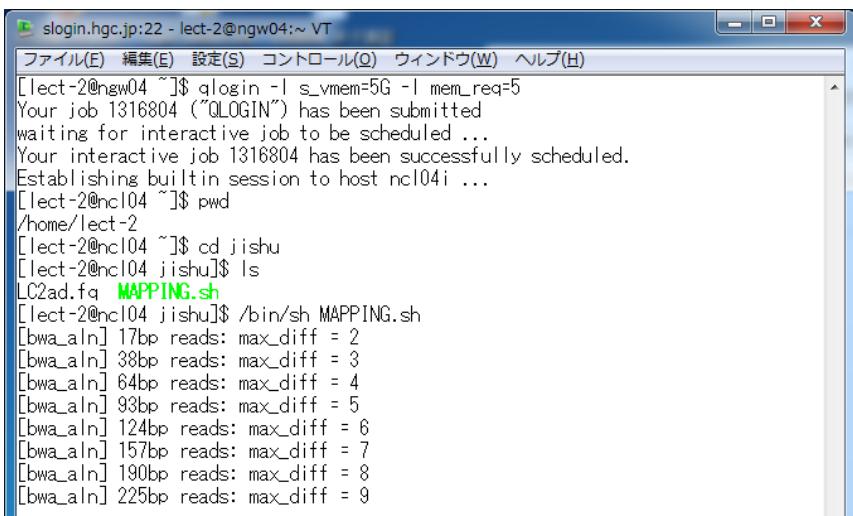
5Gのメモリが必要なので要求する  
(サーバによって異なる)

ディレクトリjishuに移動してファイルを確認

MAPPING.shを実行

と打ってください。

内容は、あとで説明いたします。



The screenshot shows a terminal window titled 'slogin.hgc.jp:22 - lect-2@ngw04:~ VT'. The session details are as follows:

```
[lect-2@ngw04 ~]$ qlogin -l s_vmem=5G -l mem_req=5
Your job 1316804 ("QLOGIN") has been submitted
waiting for interactive job to be scheduled ...
Establishing built-in session to host nc104i ...
[lect-2@nc104 ~]$ pwd
/home/lect-2
[lect-2@nc104 ~]$ cd jishu
[lect-2@nc104 jishu]$ ls
LC2ad.fq MAPPING.sh
[lect-2@nc104 jishu]$ /bin/sh MAPPING.sh
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_aln] 225bp reads: max_diff = 9
```

# 次世代シークエンサー

## Illumina MiSeq / HiSeqシリーズ

リード長: 短鎖

主たる用途: ゲノムシークエンシング,  
エキソームシークエンシング,  
トランск립トームシークエンシング



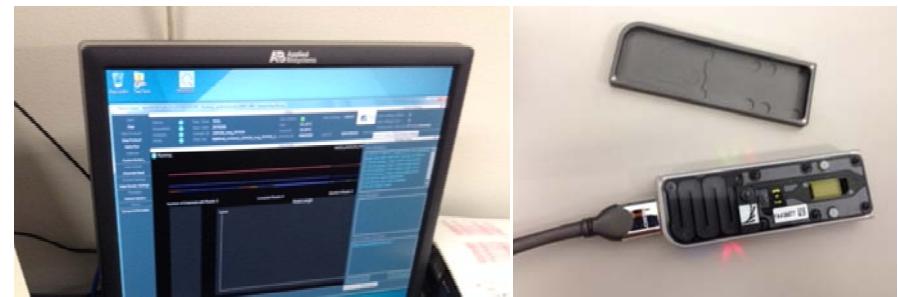
## Ion PGM / Ion Proton

デスクトップ型シークエンサー

リード長: 短鎖

主たる用途: アンプリコンシークエンシング,  
エキソームシークエンシング

HiSeq2500 (東大・新領域・鈴木研)



## PacBio RS II

1分子リアルタイムシークエンサー

リード長: 最長>20 kbの長鎖リード

主たる用途: De novoアセンブル,  
細菌ゲノムのシークエンシング, 構造多型の解析

## ONT MinION / GridION / PromethION

ナノポアシークエンサー

リード長: 短鎖～数kbの長鎖

主たる用途: DNA・RNAシークエンシング



MinION  
(東大・新領域・鈴木研)

# 用途も様々

## Whole-genome/exome sequencing

DNA配列を解読し、SNP/SNVs やindel等を同定する

## RNA-Seq, small RNA-Seq

mRNAやsmall RNAをシークエンスし、発現量の計算や新規転写産物の同定を行う

## ChIP-Seq

ヒストン修飾や転写因子の結合部位を同定する

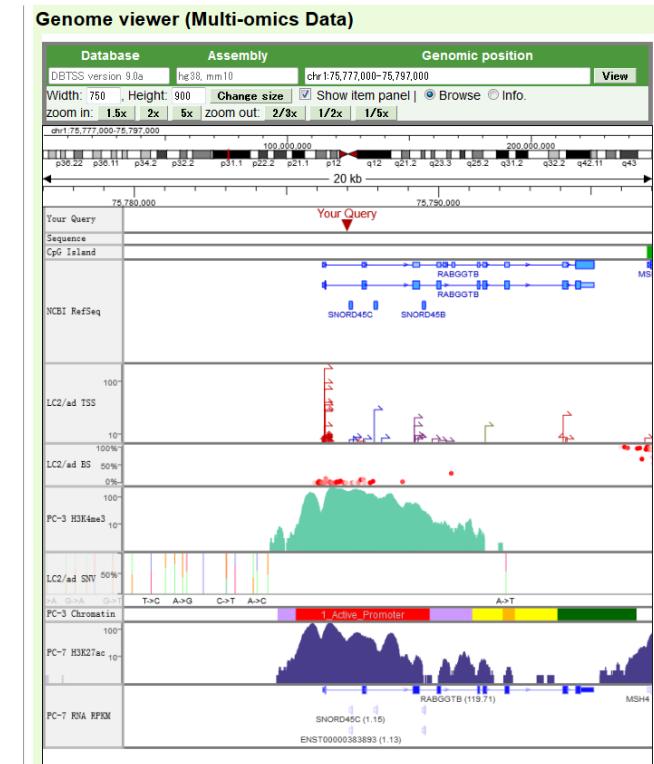
## Bisulfite sequencing

DNAのメチル化のパターンを検出する

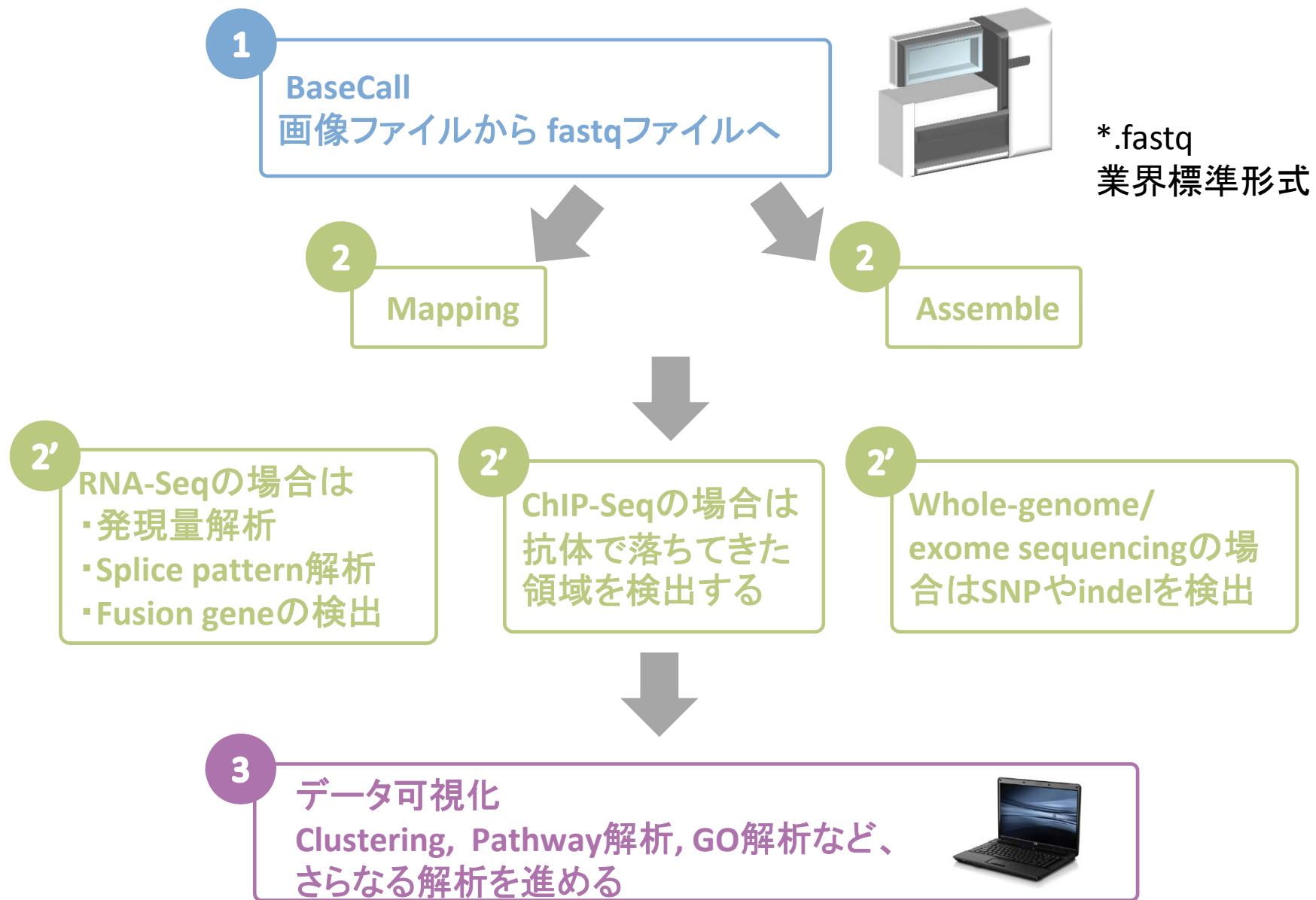
## TSS-Seq

転写開始点を同定する

など



# 鑄型に併せて解析フローも様々



# fastqファイル（シークエンスファイル）

## Format [\[edit\]](#)

A FASTQ file normally uses four lines per sequence.

- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional*/description (like a [FASTA](#) title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

A FASTQ file containing a single sequence might look like this:

```
@SEQ_ID
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT
+
!''*((((***+))%%%+)(%%%).1***-+*''))**55CCF>>>>CCCCCCCC65
```

The character '?' represents the lowest quality while '^' is the highest. Here are the quality value characters in left-to-right increasing order of quality ([ASCII](#)):

```
!"#$%&'()*+,./0123456789:;=>?@ABCDEFGHIJKLMNPQRSTUVWXYZ[¥]^_`abcdefghijklmnopqrstuvwxyz{|}~
```



参照 : [http://en.wikipedia.org/wiki/FASTQ\\_format](http://en.wikipedia.org/wiki/FASTQ_format)

# fastqファイル（シークエンスファイル）

シークエンスファイルを見てみましょう。  
新しくTeraTermを立ち上げて、  
ログインしてください。

【実習】実際にコマンドを入力しましょう

```
$ less /home/lect-1/sample_1.fq  
$ less /home/lect-1/sample_2.fq
```

The screenshot shows a terminal window titled "gw.hgc.jp:22 - Tera Term VT". The window has a menu bar with File, Edit, Setup, Control, Window, and Help. The main area displays a fastq sequence file. The first few lines of the sequence are:

```
@DGTZQN1:313:C1DGWACXX:7:1101:2385:1929 1:Y:0:ACTTGAA  
GGCAGGAGGGCAGGCAGATGAGAGAAAAGAAGAGGAGTTGGGAAGGGGACTACAAGACACAAGGAGAGATGGCTAGGGAAAGGCAGAACCGGGCCAAAAGA  
+  
#####  
@DGTZQN1:313:C1DGWACXX:7:1101:2443:1984 1:Y:0:ACTTGAA  
GGACGGGGAGGCCGGGAGACAAAACCTATAAGGTTTTGGGAAATGAGGTAGGGATCTATGAAAGAGGGTAAAGACTAGGGAAATCATGCTGGTTGGT  
+  
#####  
@DGTZQN1:313:C1DGWACXX:7:1101:2802:1948 1:Y:0:ACTTGAA  
TTTAATCTGGGAAGCTAATTCTGAGGGCTAAACTGGGATGCTCTTCTGAGATATTGGCTTTTTGTTGCTTGCCTATGGAGCTGGTAACAG  
+  
++:7+?,,7++<+10+<+74,<3>@@<=AA)?1111?)0*0)0*0=?4?*99(88AAA=A#####  
(END)
```

lessコマンドは  
qで終了します。

wcコマンドでシークエンスのリード数を  
数えてみましょう。

【実習】実際にコマンドを入力しましょう

```
$ wc /home/lect-1/sample_1.fq
```

The screenshot shows a terminal window titled "slogin.hgc.jp:22 - lect-1@ngw04:~ VT". The window has a menu bar with ファイル(E), 編集(E), 設定(S), コントロール(O), ウィンドウ(W), ヘルプ(H). The command "wc /home/lect-1/sample\_1.fq" was entered, and the output was:

```
[lect-1@nc104 ~]$ wc /home/lect-1/sample_1.fq  
15095664 18869580 983181490 /home/lect-1/sample_1.fq  
[lect-1@nc104 ~]$
```

A green box highlights the numbers "15095664 18869580 983181490" under the heading "行数 単語数 バイト数".

15095664行ありました。  
fastqファイルは4行で1リード分なので、  
このファイルには3773916リード入ってい  
ることになります

# 課題2

実際に今回使ったシークエンスファイル(fastqファイル)の中身をみてみましょう。

### 【実習】実際にコマンドを入力しましょう

```
$ cd jishu  
$ ls -lt
```

\$ less LC2ad.fq

```
gw.hgc.jp:22 - Tera Term VT
File Edit Setup Control Window Help
@DGTQZQ1:313:C1DGWACXX:7:1101:2385:1929 1:Y:0:ACTTGAA
GGCAGGAGGGCAGGCAGATCAGAGAAAAGAGGACTTGGGAAGGGGACTACAAGACACAAGGAGAGATGGCTAGGGAAAGGCAGAACCGGGCCAAAAGA
+
#####
@DGTQZQ1:313:C1DGWACXX:7:1101:2448:1984 1:Y:0:ACTTGAA
GGACGGGGAGGCCGGGAGACAAACCTATAAGGTTTGGGAAAGACTAGGGTAGATCTATGAAAGAGGGTAAAAGACTAGGGAAATCATGCTGGTTGGT
+
#####
@DGTQZQ1:313:C1DGWACXX:7:1101:2802:1948 1:Y:0:ACTTGAA
TTTTAACTGGGAAAGCTAACTTCGAGGGCTAAACTGGGATGCTCTTGAGATATTGGCTTTTTTGTGCTTTGCAATGTGGAGCTGGTAACAG
+
++:7?+,..7++<+1@+<+74,<3>@0<=AA)?1111?)0*0)0*0=24?*99(88AAA=A#####
@DGTQZQ1:313:C1DGWACXX:7:1101:3143:1935 1:N:0:ACTTGAA
CGATGCAGGATCCTATGAAATGTAACAGAACCCAGCCAGTGCACCCAGTCACCTGAATGTCATCTATGGCAAGATGTCCCCACC
+
?@D2DBDDHDH>D>B9<4CF,3>FA3+2CC<EEF11?)@060@82??AFGG150=C32EB..;B@>A6>>;355((;55;A?C3<<:@>0@3?<5
@DGTQZQ1:313:C1DGWACXX:7:1101:3011:1945 1:Y:0:ACTTGAA
ATTTCACAAAGTACATGTTCAAAGTAATAGAAACCTGGCCTTCATCAAAAAAAACTGCACAAAGAGAGTACATGCACTAGTCCTGTTACAATTTAATA
+
??BD,,=-!<,22<AED4,,<AE,<,_2AE<C>*:);C*<DDEB>?9)0*0/<299?*>=@A#####
@DGTQZQ1:313:C1DGWACXX:7:1101:3064:1958 1:N:0:ACTTGAA
CAAGACTGAGATAAAATTAATGTTGAAATGAATTAAAGCATTITGAAGAGGTATTITATGAGGTACTTGCTACTTTGCAATGTGATATGACTGCTAA
+
=?D?DFDHHP?EBG11DCF<EFHGAD?F?BCEDAEC1CHH1EE8?BDGG3?@FH:<F?DF38=BFHG1EE>CGEGG>CDHHFFCHFCFCFD@>
@DGTQZQ1:313:C1DGWACXX:7:1101:3061:1997 1:N:0:ACTTGAA
AAACGCCCCCTTGGCTGATCCGTATAATCACACCCAGTCCTACTCTCCCTATCTCTCCAGTCTAGTGTGGCATACTATACTAAACAGACCC
+
@@@DDDD)ADDFO@A1CF<E4EFFG>?CBK<DD9D2B03BF49=B8,8)>=7))@0,-=?A3(.7));;3;?@AA35>@A55>B@AA###
@DGTQZQ1:313:C1DGWACXX:7:1101:3361:1936 1:N:0:ACTTGAA
LC2ad.fq
```

# 課題3

fastqファイルは、4行で、1リード分ということでしたが、この LC2ad.fq のファイル中には、何本のリードがあるか数えてみましょう

【実習】実際にコマンドを入力しましょう

```
$ wc LC2ad fq
```

# シークエンスをヒトゲノムにマッピング

見てもらったfastqデータは、  
先ほどの課題1で、ヒトの参照ゲノム配列に対して、  
マッピングを実行中です。

投げたjobの中身を、みてみましょう！

【実習】実際にコマンドを入力しましょう

\$ less MAPPING.sh

```
slogin.hgc.jp:22 - lect-1@ngw04:~ VT
[lect-1@nc104 jishu]$ less MAPPING.sh
#!/bin/sh
## -S /bin/sh おまじない

#BWA aln
bwa aln /home/lect-1/reference/all_hg19.fa LC2ad.fq > LC2ad.sai

#BWA sampe
bwa samse /home/lect-1/reference/all_hg19.fa LC2ad.sai LC2ad.fq > LC2ad.sam

#BWA sam -> bam
samtools view -bS -o LC2ad.bam LC2ad.sam

## samtools sort
samtools sort LC2ad.bam LC2ad_sorted

## samtools sort
samtools index LC2ad_sorted.bam
```

BWAというソフトウェアでマッピング  
bwa aln → アライメント  
bwa samse → SAMファイル作成

SAMtoolsというソフトウェアで  
SAMファイルをBAMファイルにする

# Mapping software も様々

- ELAND  
Illumina社のソフトウェア
- BWA  
indelのマッピングに強く、  
ゲノム・エキソーム解析に適している
- Bowtie  
少ないメモリで高速にマッピングする(indelに弱い)
- TopHat  
スプライスを考慮してマッピングする  
RNA-Seqに適している
- など



ツール名	アルゴリズム	発表年次
Eland	read hash	2007
RMAP	read hash	2008
MAQ	read hash	2008
ZOOM	read hash	2008
SeqMap	read hash	2008
CloudBurst	read hash	2009
SHRIMP	read hash	-
SOAPv1	genome hash	2008
PASS	genome hash	2009
MOM	genome hash	2009
ProveMatch	genome hash	2009
ReSEQ	genome hash	-
Mosaik	genome hash	-
BFAST	genome hash	-
slider	merge sort	2009
SOAP2	BWT	2009
Bowtie	BWT	2009
BWA	BWT	2009



# 参照ゲノム配列

Human, mouse などの主なモデル生物の  
リファレンスゲノムや遺伝子モデル等のアノテーションデータは、  
UCSCやNCBIより取得できます。

The screenshot shows the UCSC Genome Bioinformatics website. The header includes links for Genomes, Blat, Tables, Gene Sorter, PCR, VisiGene, Session, FAQ, and Help. A sidebar on the left lists various tools: Genome Browser, ENCODE, Neandertal, Blat, Table Browser, Gene Sorter, In Silico PCR, Genome Graphs, Galaxy, VisiGene, Utilities, Downloads, Release Log, Custom Tracks, and Cancer Browser. The main content area features a section titled 'About the UCSC Genome Bioinformatics Site' with information about the genome browser, its tools like Gene Sorter and Blat, and its development by the Genome Bioinformatics Group at UCSC. It also includes a 'DONATE NOW' button and a 'News' section with a link to the 'News Archives'.

<http://hgdownload.soe.ucsc.edu/downloads.html>

今回、参照ゲノム配列として用いたUCSC hg19です。

先ほどの課題1のマッピングはおわりましたでしょうか？？

下記の出力ファイルが出てきているか確認してください。

出力データ

LC2ad.fq

Raw データ(シークエンスタグ)

LC2ad.sai  
LC2ad.sam  
LC2ad.bam  
LC2ad\_sort.bam  
LC2ad\_sort.bam.bai

マッピング結果

【実習】実際にコマンドを入力しましょう  
\$ ls -lt

```
gw.hgc.jp:22 - Tera Term VT
File Edit Setup Control Window Help
-bash-3.2$ ls -lt
合計 764144
-rw-r--r-- 1 lect-1 lect 2394296 10月 2 21:36 PC9_sorted.bam.bai
-rw-r--r-- 1 lect-1 lect 40773121 10月 2 21:36 PC9_sorted.bam
-rw-r--r-- 1 lect-1 lect 47809007 10月 2 21:36 PC9.bam
-rw-r--r-- 1 lect-1 lect 158056918 10月 2 21:36 PC9.sam
-rw-r--r-- 1 lect-1 lect 12131112 10月 2 21:35 PC9.sai
-rw-r--r-- 1 lect-1 lect 2350328 10月 2 21:28 LC2ad_sorted.bam.bai
-rw-r--r-- 1 lect-1 lect 40059303 10月 2 21:28 LC2ad_sorted.bam
-rw-r--r-- 1 lect-1 lect 47574593 10月 2 21:28 LC2ad.bam
-rw-r--r-- 1 lect-1 lect 158480436 10月 2 21:27 LC2ad.sam
-rw-r--r-- 1 lect-1 lect 12296880 10月 2 21:27 LC2ad.sai
-rwxr-xr-x 1 lect-1 lect 772 10月 2 12:22 MAPPING.sh
-rw-r--r-- 1 lect-1 lect 130238229 10月 1 16:38 LC2ad.fq
-rw-r--r-- 1 lect-1 lect 130249245 10月 1 16:37 PC9.fq
-bash-3.2$
```

# SAM (BAM) 形式データ

<http://samtools.sourceforge.net/samtools.shtml>

SAMファイルの中身を眺めてみましょう。

【実習】実際にコマンドを入力しましょう

```
$ less LC2ad.sam  
$ samtools view LC2ad.bam | more
```

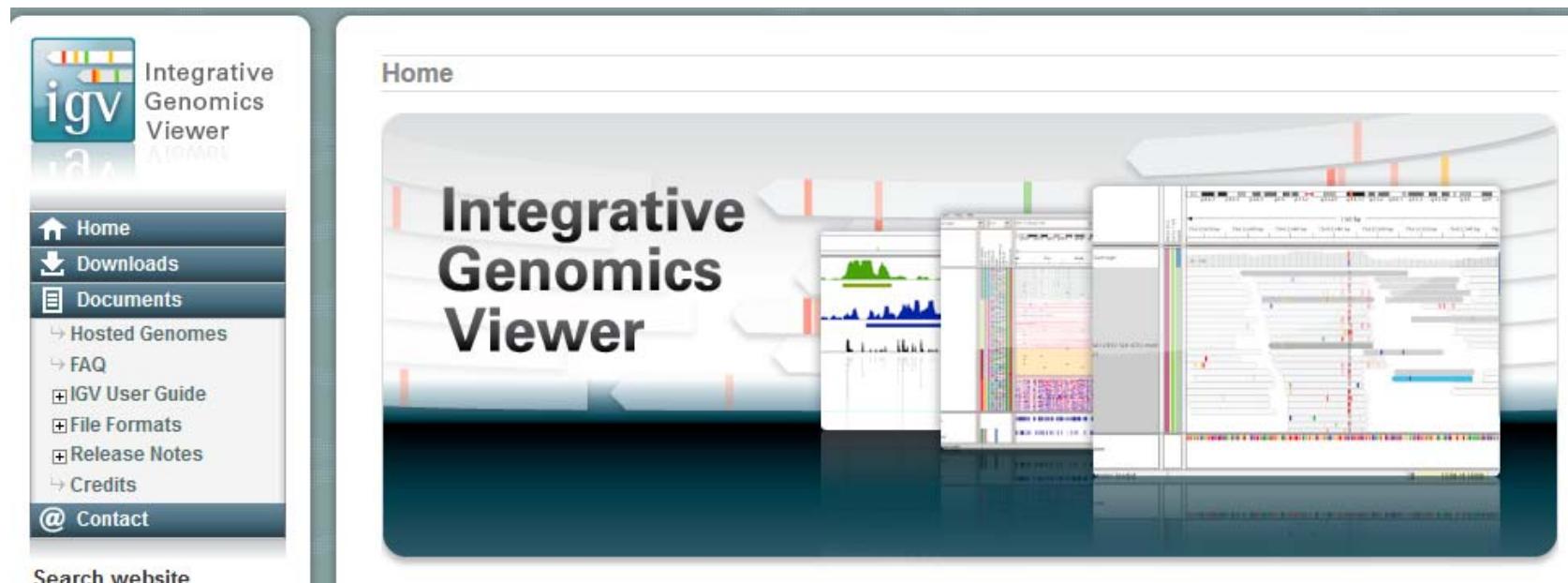
## 標準フィールド

#	略号	意味	例
1	QNAME	リード名	SRR015293.3
2	FLAG	フラグ	16
3	RNAME	リファレンス名	chr3
4	POS	スタート位置	186338939
5	MAPQ	マッピングクオリティ	25
6	CIGAR	CIGAR	32M
7	RNEXT	ペアリファレンス名	*
8	PNEXT	ペアリードのスタート位置	0
9	TLEN	総断片長 (インサートサイズ+両リード長)	0
10	SEQ	リード配列	TTGTGATGATTCGACGGTAAGCCACCATGAT
11	QUAL	クオリティ	KVNKHYYQYYJSCHQYYYYYYTYYYYYYYYYY
12	-	オプショナルフィールド(タグ)	XT:A:U NM:i:2 X0:i:1 X1:i:0 XM:i:2 X0:i:0 XG:i:0 MD:Z:4C7T19

<https://cell-innovation.nig.ac.jp/wiki/tiki-index.php?page=SAM>

# データ可視化

<https://www.broadinstitute.org/igv/home>



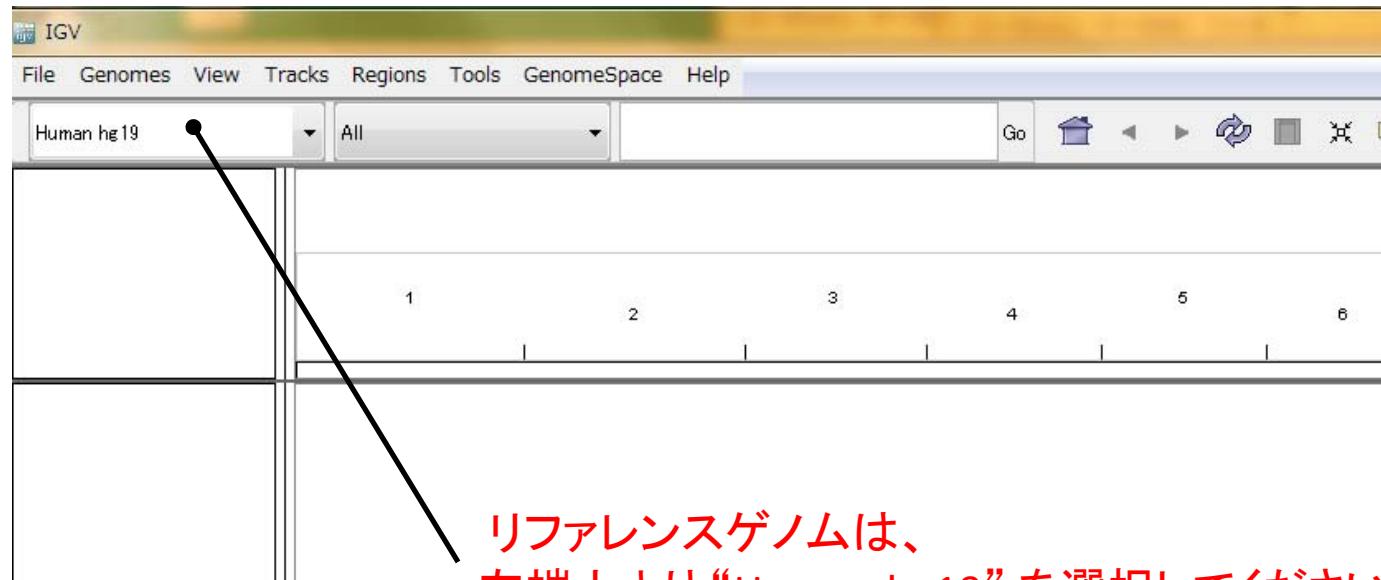
## 課題4

マッピングしたデータを可視化ツール(IGV)でみてみましょう。

WinSCPを用いて、Windows PCのデスクトップに

[LC2ad\\_sort.bam](#) および [LC2ad\\_sort.bam.bai](#) をダウンロードしてください。

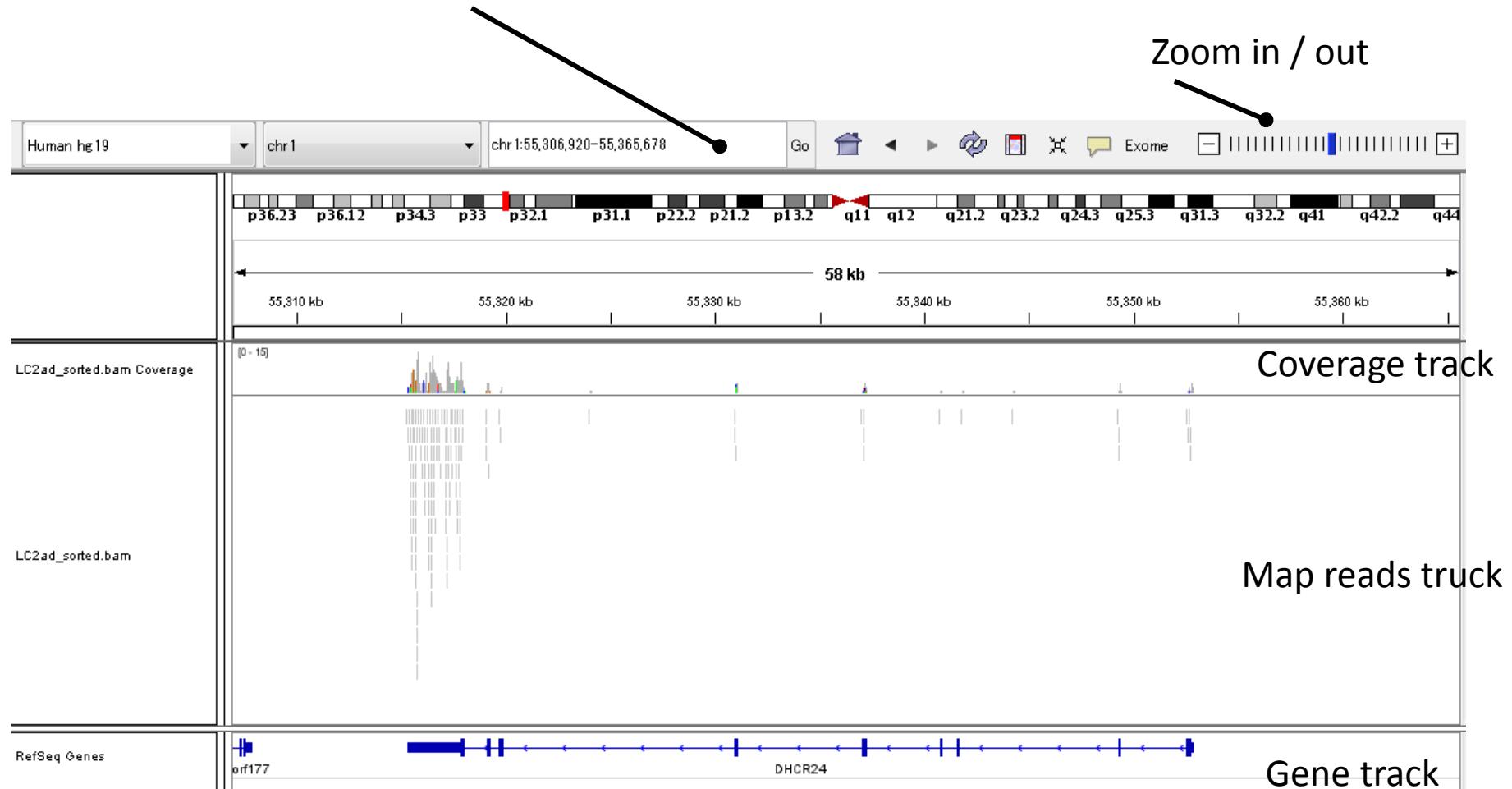
IGVを起動させて、



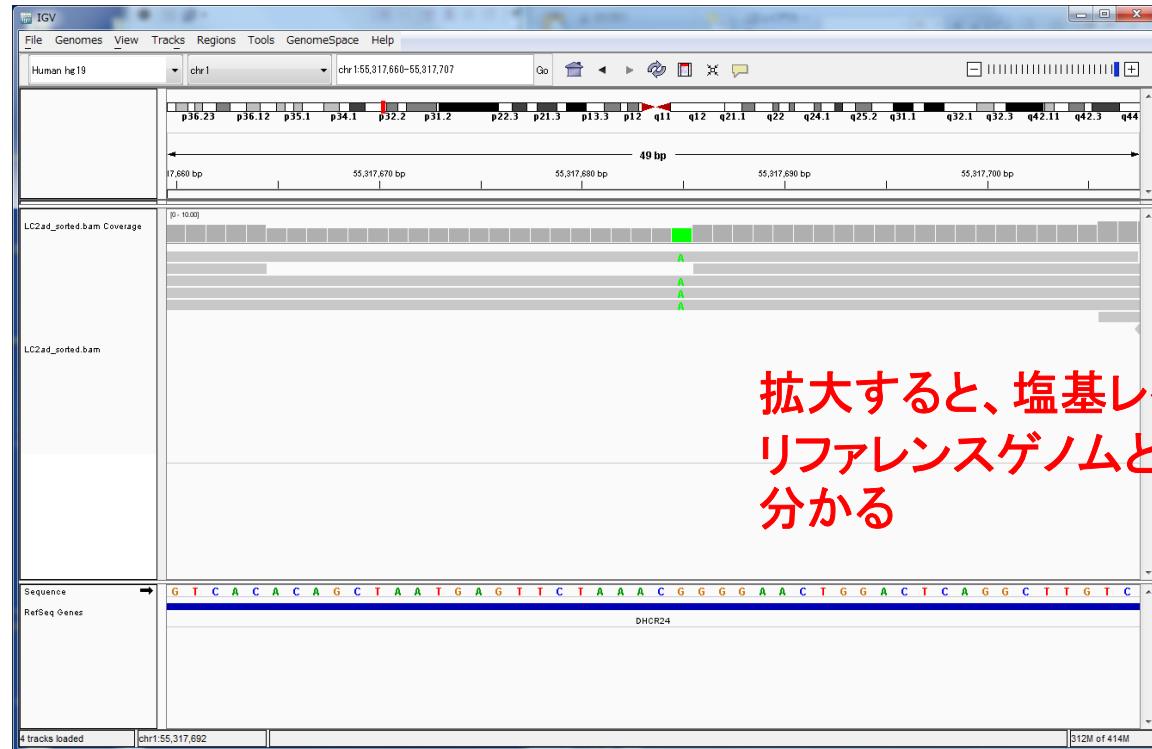
IGVのFile → Load from File... で [LC2ad\\_sorted.bam](#) を開いてください。

# IGV 表示内容

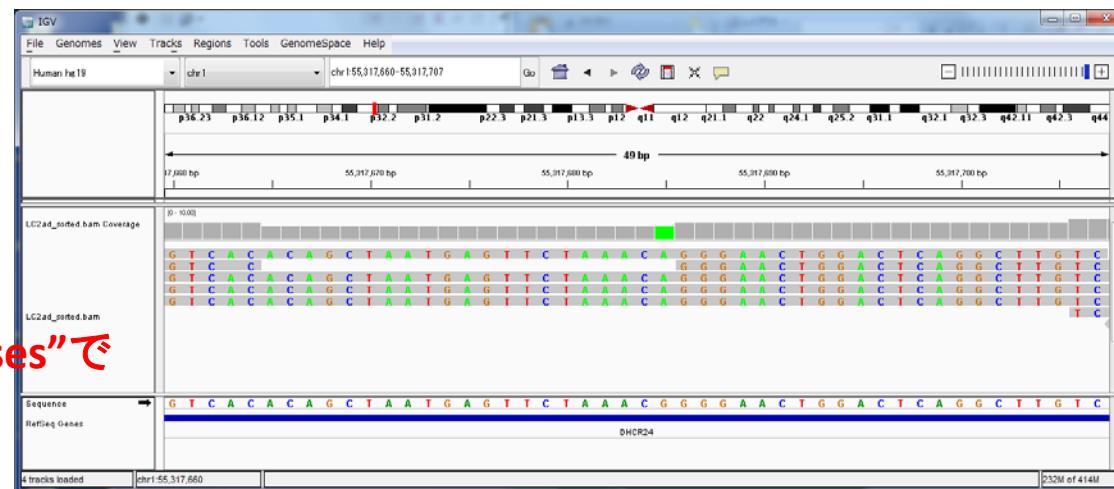
Symbol や座標で検索可能



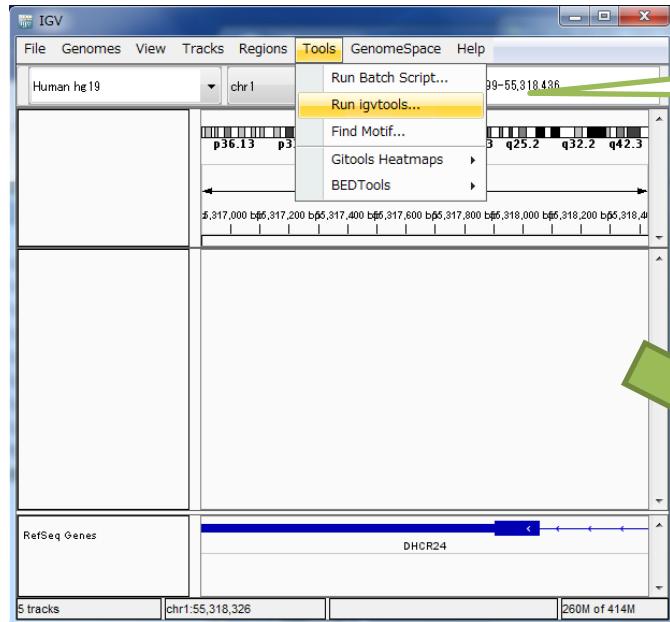
DHCR24 と検索してみましょう。



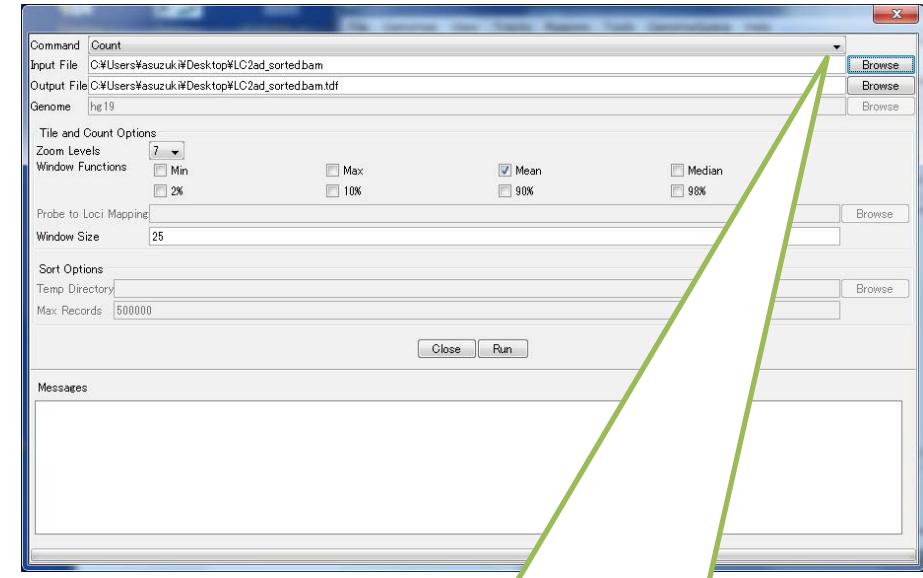
拡大すると、塩基レベルまで見られる。  
リファレンスゲノムと違う部分(多型や変異)が  
分かる



右クリック→"show all bases"で  
全ての塩基を表示できる



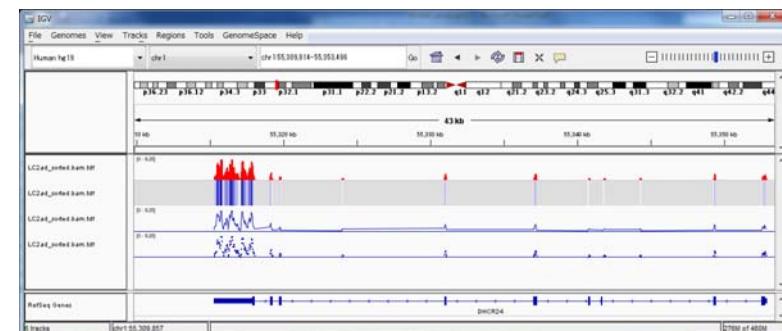
Tools → Run igvtools



Input FileにBAMファイル(LC2ad.bam)を選択  
→ Run



TDFファイルをIGVで開いてみましょう



色や表示を変えることができます

おまけ

## 課題5

IGVを見ていただきましたが、

各遺伝子ごとにリードを目視でcountするのも結構大変です。

我々は、Perlプログラムを書いて、プログラムに処理させてます。

Tera Termを開いて、LC2ad.samファイルを使って

chr1: 55,315,300 ~ 55,352,921 (symbol : DHCR24) にマップされる

リードをカウントしてみましょう。

Perlでも、Cでも、Pythonでも、Rubyでも、目視でも。

ディレクトリjishuの中に入って、LC2ad.samを使います。  
chr1にマップされているものだけ抽出：

```
$ perl -F"\t" -ane 'if($F[2] eq 'chr1'){print "$_";}' LC2ad.sam | less  
$ perl -F"\t" -ane 'if($F[2] eq 'chr1'){print "$_";}' LC2ad.sam | wc (何件あるか確認)
```

55,315,300-55,352,921 にあるという条件を足す：

```
$ perl -F"\t" -ane 'if($F[2] eq 'chr1' && $F[3]>= 55315300 && $F[3]<=55352921){print "$_";}' LC2ad.sam | less  
$ perl -F"\t" -ane 'if($F[2] eq 'chr1' && $F[3]>= 55315300 && $F[3]<=55352921){print "$_";}' LC2ad.sam | wc
```

---

-F “\t” : 読み込むファイルはtab 区切り形式

-ane

a: splitしてFに入れる

n: ファイルを1行づつみてくれる

e: one linerにしてくれる

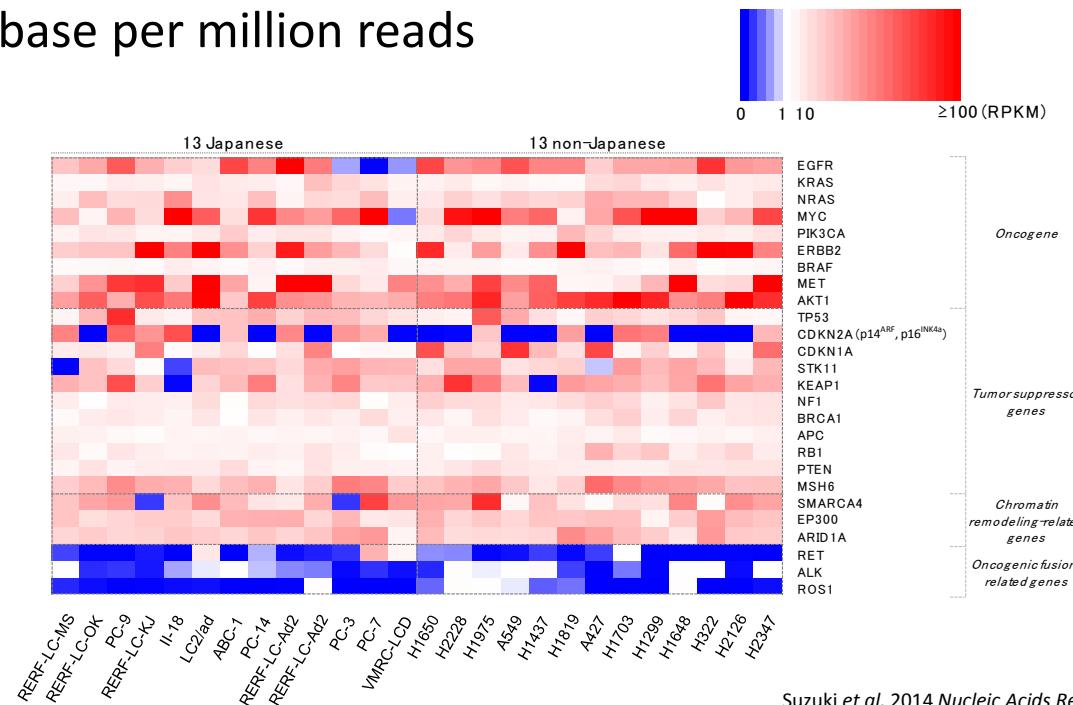
# 発現量(RNA-Seqデータ)

発現量の算出:

1. 各遺伝子領域にマップされたシークエンスタグを数える→タグ数
2. 全シークエンスタグ数でノーマライズ → ppm
3. 遺伝子の長さでノーマライズ → rpk

ppm: parts per million reads

rpk: reads per kilobase per million reads



# RNA-Seqデータの解析

TopHat-CuffLinksを用いた例

TopHat

参照ゲノム配列へマッピング

↓

CuffLinks

アセンブルによる転写産物の検出

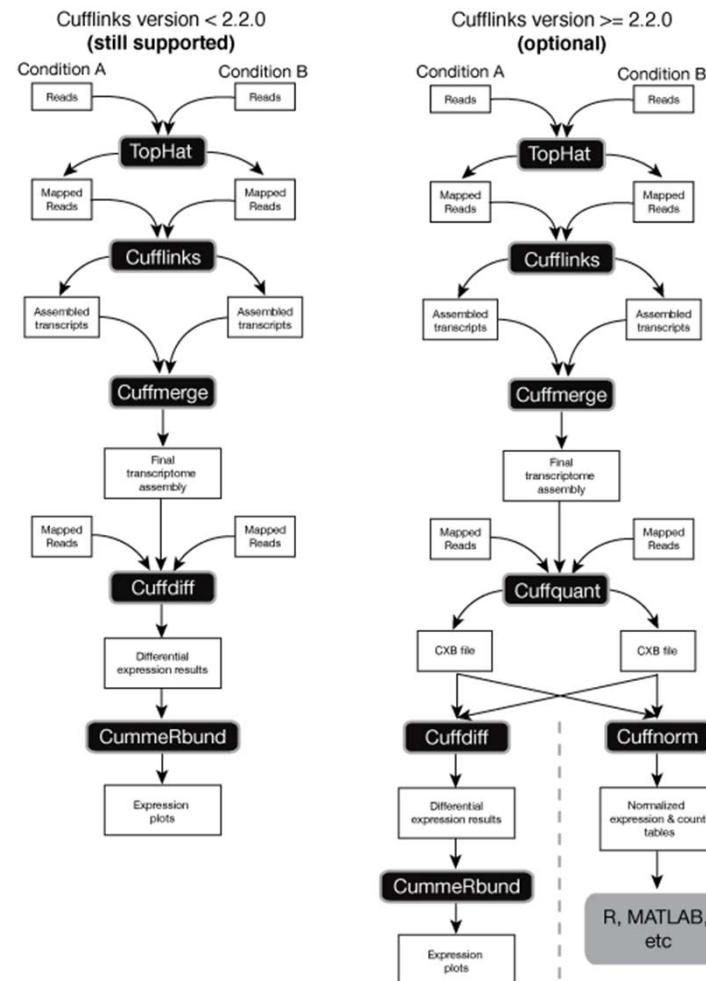
発現量算出

↓

Cuffdiff

サンプル(群)間の発現量の比較

Differentially expressed geneの検出



[http://cole-trapnell-lab.github.io/cufflinks/getting\\_started/](http://cole-trapnell-lab.github.io/cufflinks/getting_started/)

たとえば、

正常組織と癌組織で有意に発現量が異なる遺伝子群を探索する

↓

Rでヒートマップを描く・クラスタリングする・

DAVID(<https://david.ncifcrf.gov/>)等でどのような遺伝子群(GO/KEGGなど)が濃縮されているのか調べる  
などと進んでいく。

# 肺腺癌細胞株のRNA-Seqデータ

今回用いたRNA-Seqデータは実習用の小さいファイルサイズのデータでした。

実際に解析に用いているファイルが以下にあります。

データベースDBTSSダウンロードページ：

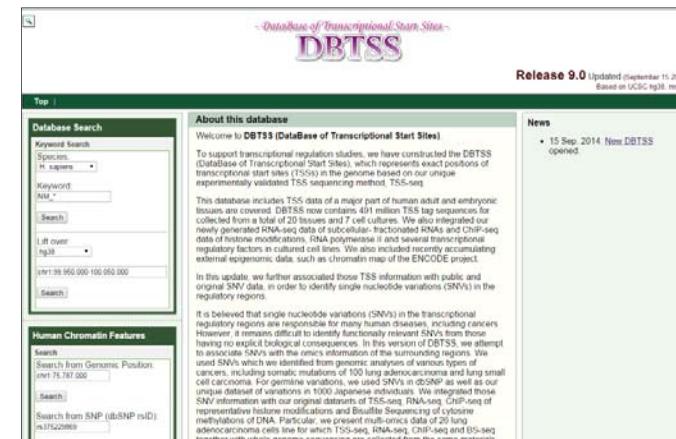
[ftp://ftp.hgc.jp/pub/hgc/db/dbtss/dbtss\\_ver9/hg38/RNAseq/original\\_hg19\\_tophat2mapped\\_data/](ftp://ftp.hgc.jp/pub/hgc/db/dbtss/dbtss_ver9/hg38/RNAseq/original_hg19_tophat2mapped_data/)

26種類の肺腺癌細胞株のRNA-Seqデータです。

各フォルダの下に、RNA-SeqのfastqファイルをTopHat2でマッピングした結果があります。

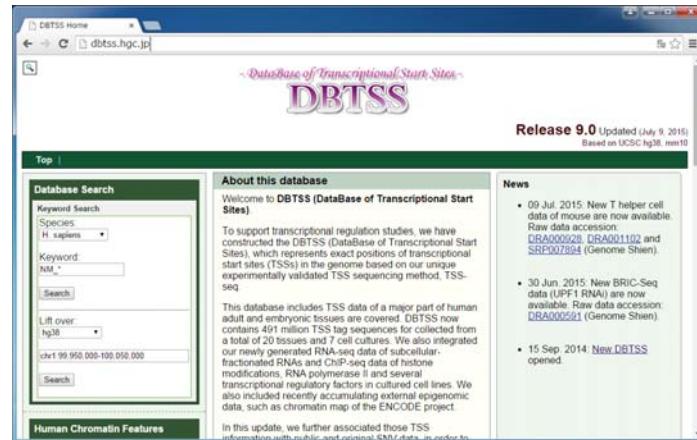
rpkms\_based\_on\_hg19\_genome.xlsxは、各遺伝子のタグ数、ppm、rpkmを算出したものになります。

DBTSSでは、TSS-Seq, RNA-Seqをはじめ、ゲノム、エピゲノム、トランск립トームのデータが見られます。

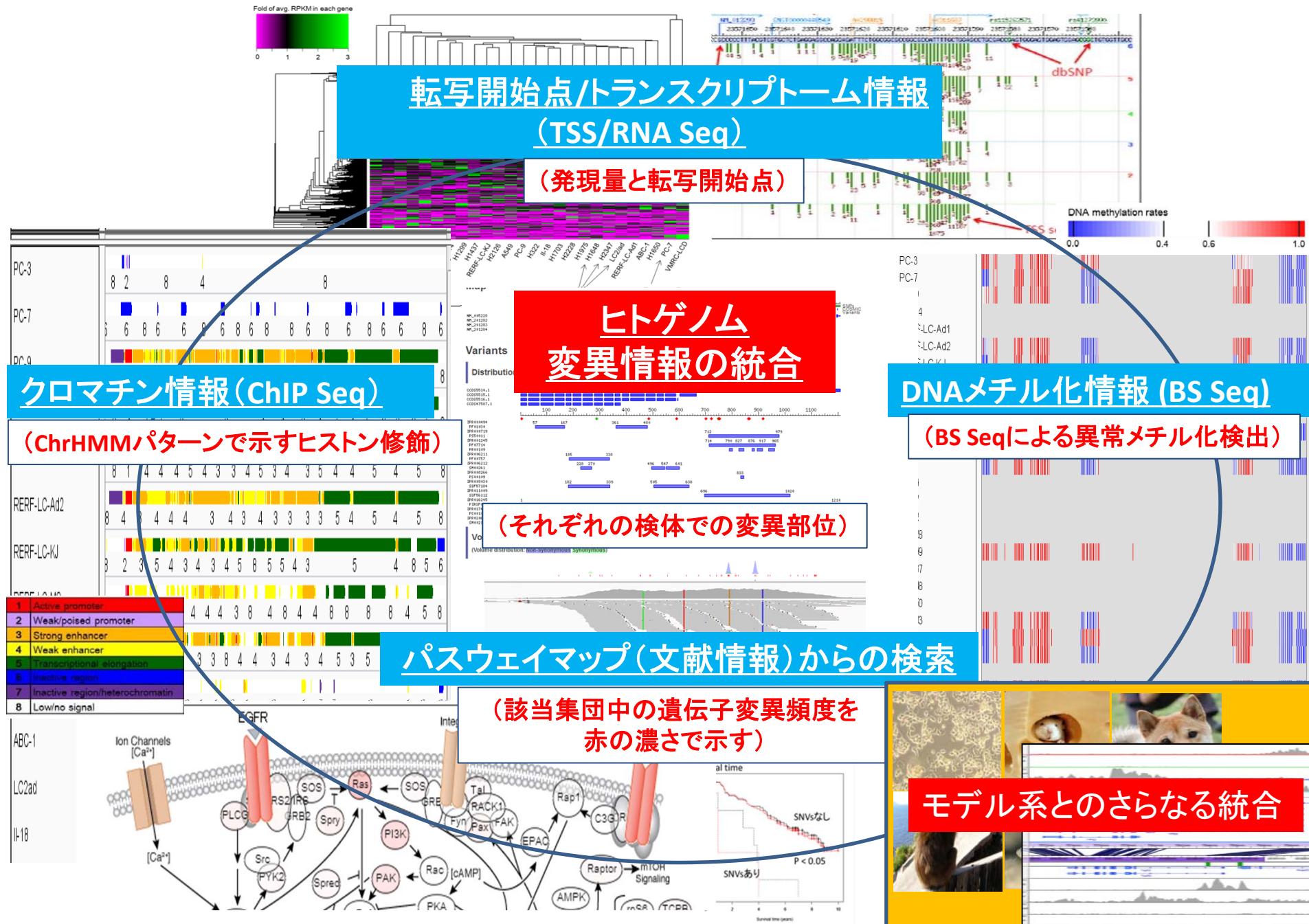


# Session 2

## データベースDBTSSの紹介



# ヒト応用研究を志向したオミクス情報の統合



# データベースDBTSS

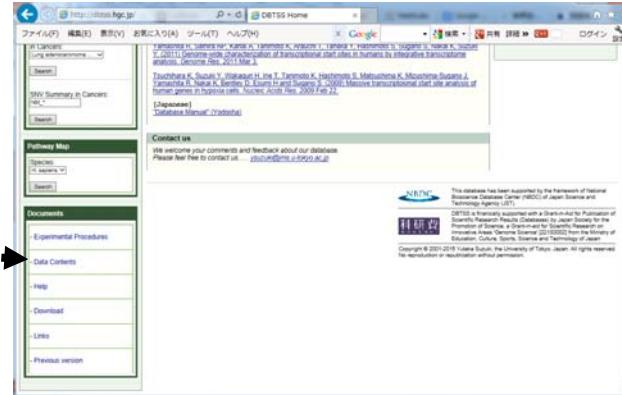
DBTSS(<http://dbtss.hgc.jp/>)

The screenshot shows the DBTSS homepage with a dark blue header bar. The title "DBTSS" is prominently displayed in the center. Below the title, it says "Release 9.0 Updated (July 9, 2015) Based on UCSC hg38, mm10". On the left, there is a "Database Search" panel with fields for "Species" (set to "H. sapiens"), "Keyword" (set to "NM\_\*"), and "Lift over" (set to "hg38"). Below this is a "Human Chromatin Features" section. The main content area has a green header "About this database" which includes a welcome message and a detailed description of the database's features and data sources. To the right, there is a "News" section listing recent updates:

- 09 Jul. 2015: New T helper cell data of mouse are now available. Raw data accession: [DRA000928](#), [DRA001102](#) and [SRP07894](#) (Genome Shien).
- 30 Jun. 2015: New BRIC-Seq data (UPF1 RNAi) are now available. Raw data accession: [DRA000591](#) (Genome Shien).
- 15 Sep. 2014: New DBTSS opened.

# データコンテンツ

Topページ → “Data Contents” ----->



## Data contents

	Number of dataset				Data contents			
	Human	Mouse	Malaria	Chyzon	Rat	Chimpanzee	Macaque	
TSS-seq	73	7	1	1	1	1	2	
RNA-seq	42	0	0	0	0	0	0	
ChIP-seq	255	0	0	~	~	~	~	
RIP-seq	12	0	0					
BS-seq	26	0	0					
ChromHMM	36	0	0					
SNV	49	0	0					

Cell lines															
Cell line			General information*				Cell culture		Sequencing dataset				Memo		
Name	Type	in-house analysis number	Distributor	Catalogue number	Ethnicity	Gender	Age	Smoking status	Medium	Dish	Whole- genome Seq	RNA-Seq	BS-Seq	CNP-Seq	TSS-Seq
SAEC (control)	Small airway epithelial cell	s_35	TAKARA (Lonza)	-	-	M	-	-	-	collagen Type I-coated	-	♦	-	♦	♦
PC-9	lung adenocarcinoma	s_9	RIKEN BRC	RCB4455	Japanese	-	-	-	RPMI	-	♦	♦	♦	♦	-
PC-14	lung adenocarcinoma	s_10	IBL	-	Japanese	-	-	-	RPMI	-	♦	♦	♦	♦	-
RERF-LC-KJ	lung adenocarcinoma	s_13	RIKEN BRC	RCB1313	Japanese	M	78	-	RPMI	-	♦	♦	♦	♦	-
RERF-LC-Ad1	lung adenocarcinoma	s_11	JCRB	JCRB1020	Japanese	M	70	-	RPMI	-	♦	♦	♦	♦	-
RERF-LC-Ad2	lung adenocarcinoma	s_12	JCRB	JCRB1021	Japanese	M	-	-	RPMI	-	♦	♦	♦	♦	-
LC2/ad	lung adenocarcinoma	s_18	RIKEN BRC	RCB0440	Japanese	F	51	-	DMEM	collagen Type I-coated	♦	♦	♦	♦	-
RERF-LC-MS	lung adenocarcinoma	s_14	JCRB	JCRB0081	Japanese	-	-	-	EMEM	-	♦	♦	♦	♦	-
VMRC-LCD	lung adenocarcinoma	s_16	JCRB	JCRB0814	Japanese	M	-	-	EMEM	-	♦	♦	♦	♦	-
ABC-1	lung adenocarcinoma	s_17	JCRB	JCRB0815	Japanese	M	47	-	EMEM	-	♦	♦	♦	♦	-
PC-7	lung adenocarcinoma	s_8	IBL	-	Japanese	-	-	-	RPMI	-	♦	♦	♦	♦	Non-adherent
PC-3	lung adenocarcinoma	s_7	JCRB	JCRB0077	Japanese	F	48	-	RPMI	collagen Type I-coated	♦	♦	♦	♦	-
II-18	lung adenocarcinoma	s_19	RIKEN BRC	RCB2093	Japanese	-	-	-	RPMI	-	♦	♦	♦	♦	-
RERF-LC-OK	lung adenocarcinoma	s_15	JCRB	JCRB0811	Japanese	-	-	-	RPMI	-	♦	♦	♦	♦	-
A549	lung adenocarcinoma	s_1	ATCC	CCL-185	Caucasian	M	58	-	DMEM	-	♦	♦	♦	♦	-
A427	lung adenocarcinoma	s_20	ATCC	HTB-53	Caucasian	M	52	-	RPMI	-	♦	♦	♦	♦	-
H322	lung adenocarcinoma	s_21	ATCC	CRL-5806	Caucasian	-	-	-	RPMI	-	♦	♦	♦	♦	-
H1648	lung adenocarcinoma	s_25	ATCC	CRL-5882	Black	M	39	Y	RPMI	collagen Type I-coated	♦	♦	♦	♦	-
H1650	lung adenocarcinoma	s_26	ATCC	CRL-5883	Caucasian	M	27	Y	RPMI	-	♦	♦	♦	♦	-
H1975	lung adenocarcinoma	s_29	ATCC	CRL-5908	-	F	-	N	RPMI	-	♦	♦	♦	♦	-

細胞株を中心に。

# 26種類の肺腺癌細胞株のオミクスデータ

## Whole-genome sequencing

Sequencing: illumina HiSeq2500

## Bisulfite sequencing (BS-Seq)

Capture: Agilent SureSelect Methyl-Seq Target Enrichment System (84 Mb)

Sequencing: illumina HiSeq2500

## ChIP-Seq for histone modifications and RNA Polymerase II

Sequencing: HiSeq2500

IP: H3K4me3, H3K9/14ac, H3K36me3, H3K27me3,

H3K9me3, H3K4me1, H3K27ac, Pol II

Control: whole cell extract (WCE)

## TSS-Seq

Sequencing: illumina HiSeq2500

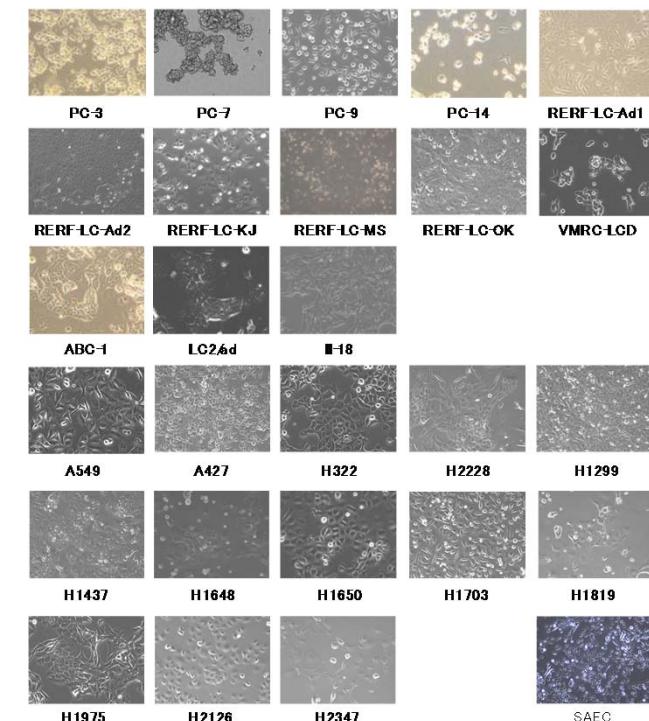
## RNA-Seq

Sequencing: illumina HiSeq2500

## Small RNA-Seq

Sequencing: illumina HiSeq2500

赤文字はシークエンスデータもDDBJから公開済



Suzuki *et al.* 2014 Nucleic Acids Research

# ヒトの多型・変異情報も格納(>15000)

## Cancer genomes

ICGC: International Cancer Genome Consortium

TCGA: The Cancer Genome Atlas

Gastric adenocarcinoma, Urothelial bladder carcinoma, Glioblastoma,  
Clear cell renal cell carcinoma, Endometrial carcinoma,  
Acute myeloid leukemia, Breast tumors, Squamous cell lung cancers,  
Colorectal cancer, Ovarian carcinoma...

National Cancer Center Hospital East:

Lung adenocarcinoma (2013 *PLoS ONE*),  
Small cell lung cancer (2014 *J Thorac Oncol*)

Others:

Myelodysplasia (2011 *Nature*), Clear-cell renal cell carcinoma (2013 *Nat Genet*),  
Lung adenocarcinoma (2012 *Cell*)...

## Japanese genomes (SNPs)

HGVD: The Human Genetic Variation Database

ToMMo: Tohoku Medical Megabank Organization

JPDSC: The Japan PGx Data Science Consortium

Database of Transcriptional Start Sites

DBTSS

Release 9.0 Updated 1

Based on UCSC Genome Browser

**Top ↑**

**Database Search**

Keyword Search

Species:

H\_sapiens

Keyword:

not \*

Search

Lift over:

highlight

Search

chr1 99 950 000-100 000

Search

Search

**Human Chromatin Features**

Search

Search from Genomic Position

chr1 71 717 000

Search

Search from SNP (dbSNP rsID)

rs17422969

Search

Search from SVN (COSMIC, somatic mutation)

BRAF

Search

Search from SVN-enriched Gene in Cancers

lung\_stomach

Search

SVN Summary in Cancers

NA\_\*

Search

**Pathway Map**

Species:

H\_sapiens

Search

**Documents**

- Experimental Procedures

- Data Contents

- Help

Download

- Previous version

**About this database**

Welcome to DBTSS (Database of Transcriptional Start Sites)

To support transcriptional regulation studies, we have constructed the DBTSS (Database of Transcriptional Start Sites), which represents exact positions of transcriptional start sites (TSSs) in the genome based on our unique experimentally validated method, TSS-seq.

This database includes TSS data in a major part of human adult and embryonic tissues are covered. DBTSS now contains 491 million TSS tag sequences for collected from a total of 20 tissues and 7 cell cultures. We also integrated our newly generated RNA-seq data of subcellular- fractionated RNAs and ChIP-seq data of histone modifications, RNA polymerase II and several transcriptional regulatory factors in cultured cell lines. We also included recently accumulating external genomic data, such as chromatin map of the ENCODE project.

In this update, we further associated these TSS information with public and original RNA-seq data, in order to identify single nucleotide variations (SNVs) in the regulatory regions.

It is believed that single nucleotide variations (SNVs) in the transcriptional regulatory regions are responsible for many human diseases, including cancers. However, it remains difficult to identify biologically relevant SNVs from those having no explicit biological consequences. In this version of DBTSS, we attempt to associate SNVs with the entire information of the surrounding regions. We used SNVs which were identified from genomic analyses of various types of cancers, including somatic mutations in 100 cancer genomes, and 100 cancer cell lines, as well as cell carcinomas. For germline variations, we used SNVs in dbSNP as well as our unique dataset of variations in 1000 Japanese individuals. We integrated those SNV information with our original datasets of TSS-seq, RNA-seq, ChIP-seq of representative cell lines, and RNA-seq data of the same cell lines separated by cytosine methylation of DNA. Particular, we present multi-data set of TSS-seq, RNA-seq, adenocarcinoma cell lines for which TSS-seq, RNA-seq, ChIP-seq and SNV-seq together with whole genome sequencing are collected from the same cell lines. We further connected the multi-data sets of model organisms by genome-wide comparative analysis. We provide a unique data resource to investigate what genomic features are observed in a particular genome; coordinates in a wide variety of samples.

These data can be browsed in our new viewer which also supports versatile search conditions of users. We believe new DBTSS is helpful to understand biological consequences of the massively identified TSSs and identify human genetic valuations which are associated with disordered transcriptional regulations.

**References**

Suzuki A, Wakaguri H, Yamashita R, Kawano S, Tsuchihara K, Sugano S, Suzuki T, Nakai K. (2013) An Integrative Database for Transcriptional Start Site (TSS)-based genomic sequence variation data. *Nucleic Acids Res* 2013 (Database issue): D101-D108.

Suzuki A, Minaki S, Yamane Y, Kawano A, Methylumura K, Suzuki T, Sugano S, Suzuki H, Suzuki Y, Suzuki Y. (2011) Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis and genome browser. *Genome Res* 2011 Mar 3.

Tsuchihara K, Suzuki H, Kawano A, Tsumori K, Hashimoto M, Methylumura K, Miyazawa-Sugano Y, Yamashita R, Suzuki H, Kawano A, Sugano S. (2009) Massive transcriptional start site analysis of *in vitro* hypoxia cells. *Nucleic Acids Res* 2009 Feb 22.

[Japanese] "Database Manual" (Yodobashi)

**Contact us**

We welcome your comments and feedback about our database.  
Please feel free to contact us: [yodobashi@yodobashi.u-tokyo.ac.jp](mailto:yodobashi@yodobashi.u-tokyo.ac.jp)

Copyright © 2014

**News**

- 30 Jun 2015: New BRCA1 RNA-seq data (U1111 RNA) are now available. Raw data access: [DBA000028](#) | [Geneome Sh](#)
- 30 Jun 2015: New BRCA1 RNA-seq data (U1111 RNA) are now available. Raw data access: [DBA000028](#) | [Geneome Sh](#)
- 15 Sep 2014: New DBTSS opened.

Download

D

**Top**

**Database Search**

Search

Search from Genomic Position

chr1 71 717 000

Search

Search from SNP (dbSNP rsID)

rs17422969

Search

**Human Chromatin Features**

Search

Search from Genomic Position

chr1 99 950 000-100 000

Search

Search from SVN (COSMIC, somatic mutation)

BRAF

Search

Search from SVN-enriched Gene in Cancers

lung\_stomach

Search

Search from SVN Summary in Cancers

NA\_\*

Search

**Top**

**Genome Viewer (Multi-chromosome)**

Search

Search from Genomic Position

chr1 71 717 000

Search

**TSS Seq Data**

Search

Search from SNP (dbSNP rsID)

rs17422969

Search

**Human Chromatin Features**

Search

Search from Genomic Position

chr1 99 950 000-100 000

Search

Search from SVN (COSMIC, somatic mutation)

BRAF

Search

Search from SVN-enriched Gene in Cancers

lung\_stomach

Search

**Add Issues**

Search

Search from Genomic Position

chr1 71 717 000

Search

Search from SVN Summary in Cancers

NA\_\*

Search

Database Search

Human Chromatin Features

Genome viewer

Human Variation (hg19座標後リンク)

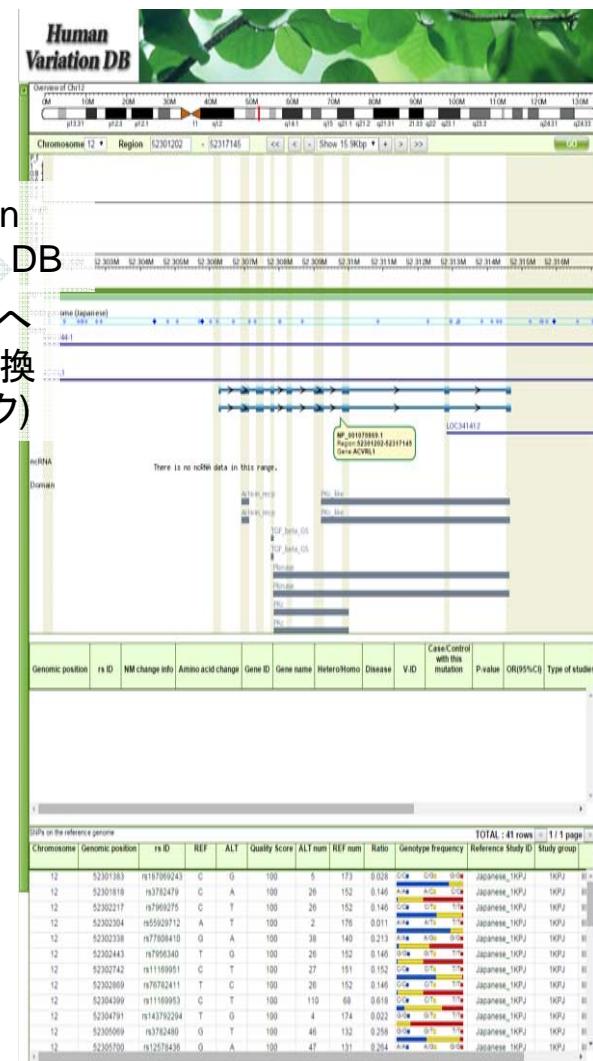
Release 9.0 Updated July 9, 2016  
Based on UCSC hg18, ver 10

Database Search

Human Chromatin Features

Genome viewer

Human Variation (hg19座標後リンク)



# TSS viewer

**Top |**

### Database Search

**Keyword Search**

Species:  
H. sapiens

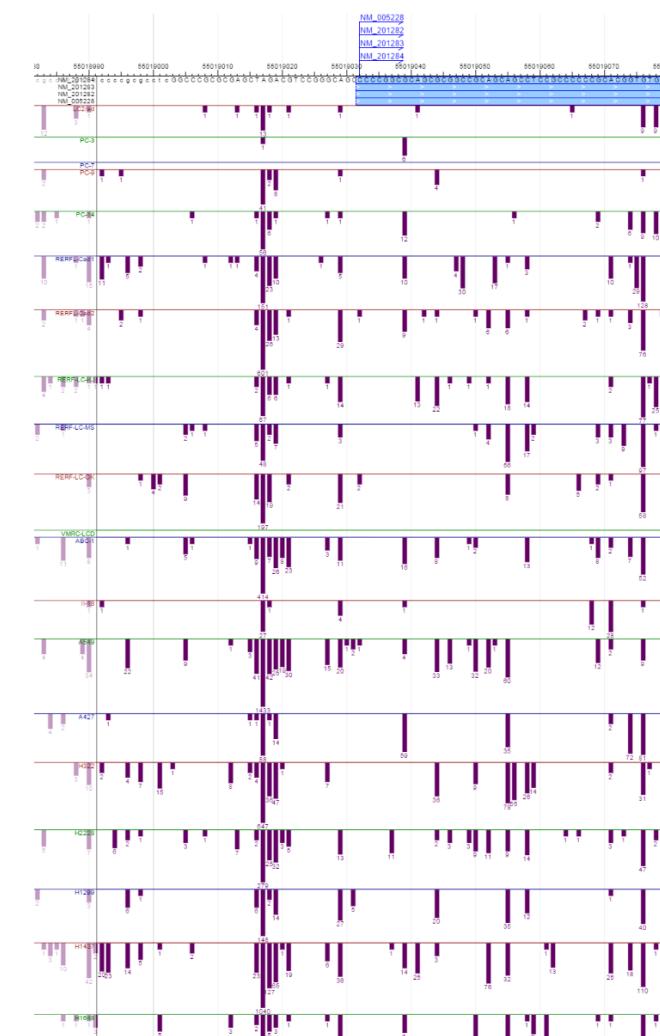
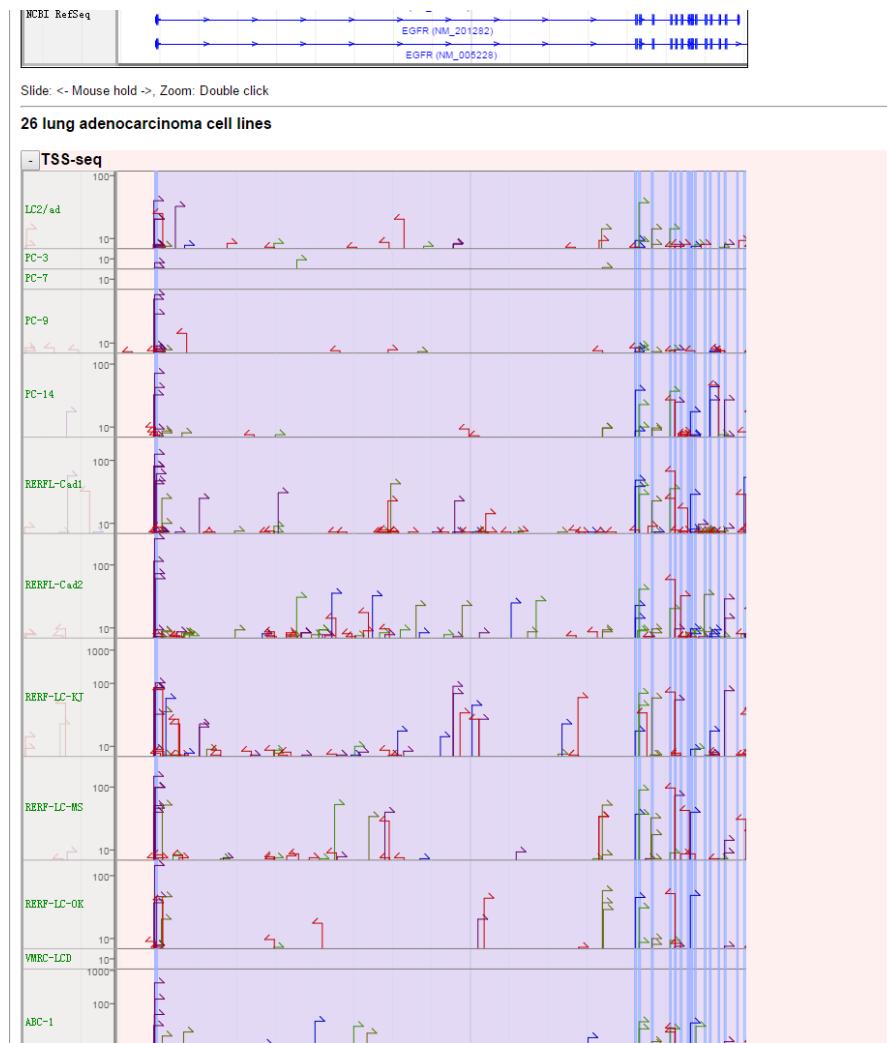
Keyword:  
EGFR\*

**Search**

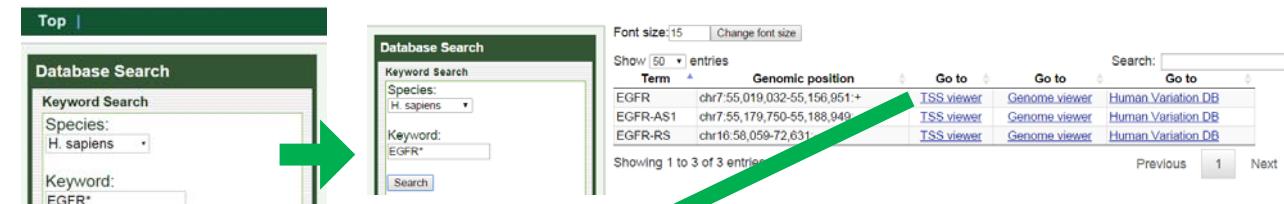
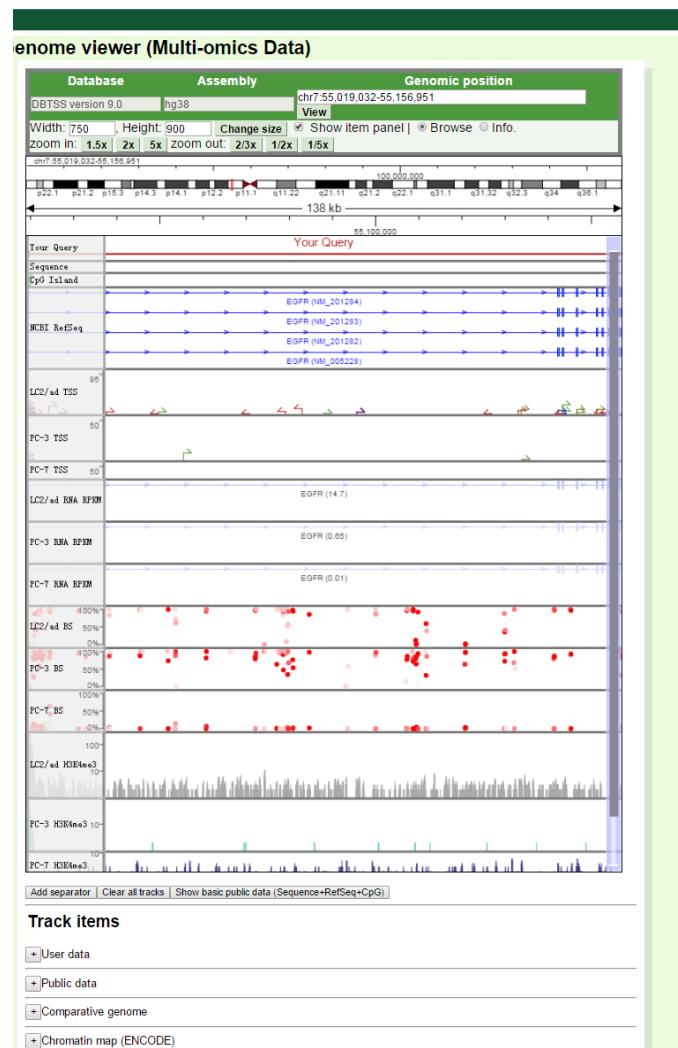
EGFR\*で検索

Database Search				Font size: 15	Change font size
Show 50 ▾ entries				Search: <input type="text"/>	
Term	Genomic position	Go to	Go to	Go to	Go to
EGFR	chr7:55,019,032-55,156,951+	TSS viewer	Genome viewer	Human Variation DB	
EGFR-AS1	chr7:55,179,750-55,188,949-	TSS viewer	Genome viewer	Human Variation DB	
EGFR-RS	chr16:58,059-72,631-	TSS viewer	Genome viewer	Human Variation DB	

## TSS viewerをクリック

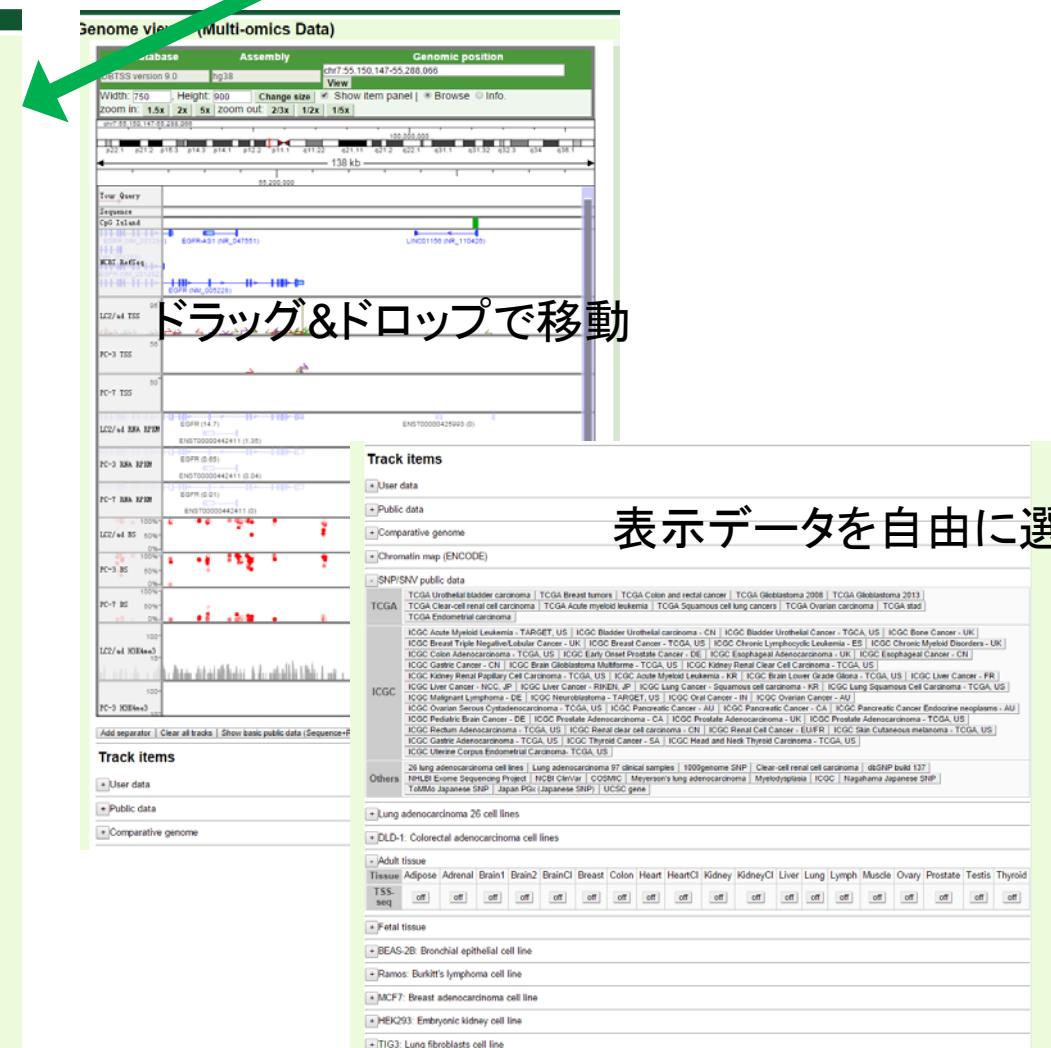


# Genome viewer



EGFR\*で検索

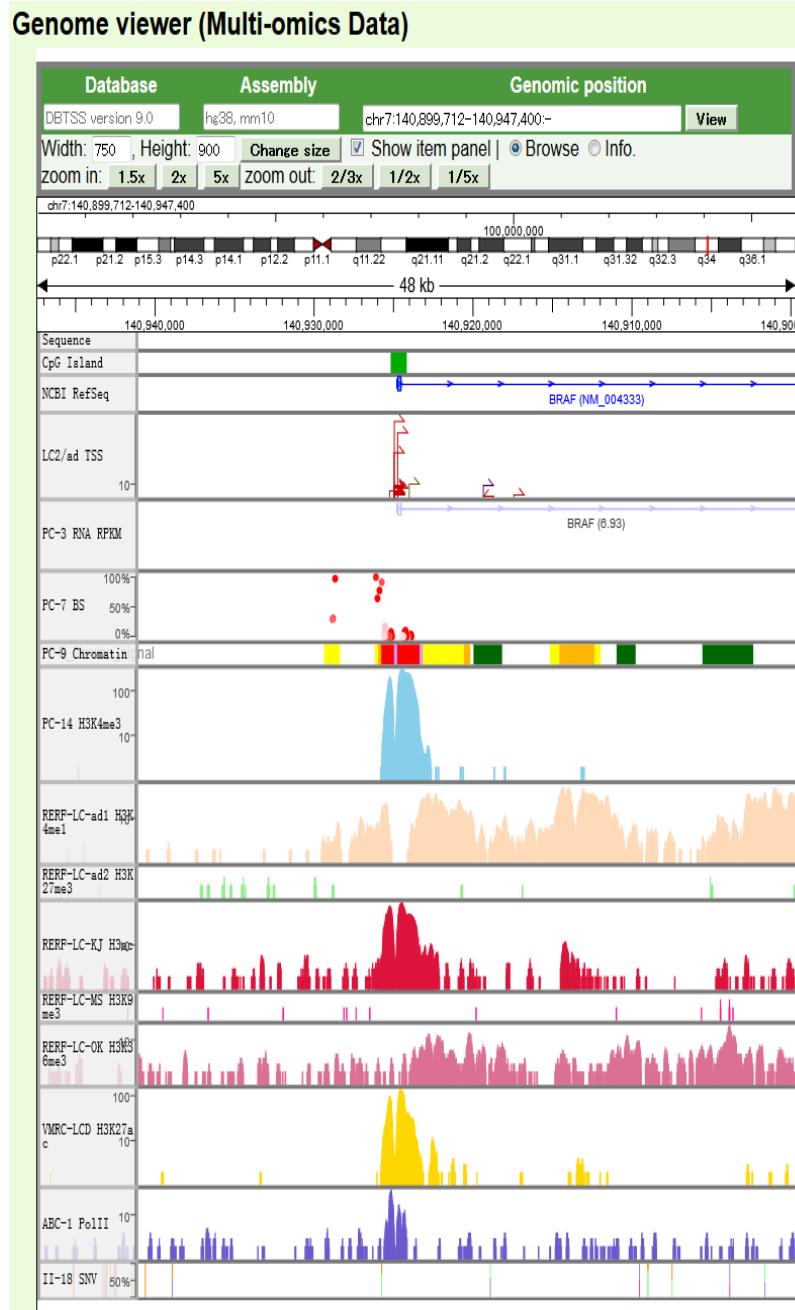
Genome viewerをクリック



ドラッグ&ドロップで移動

表示データを自由に選択

# Genome viewer (BRAF遺伝子を例に)



## Viewer control

## Gene model

## TSS-Seq

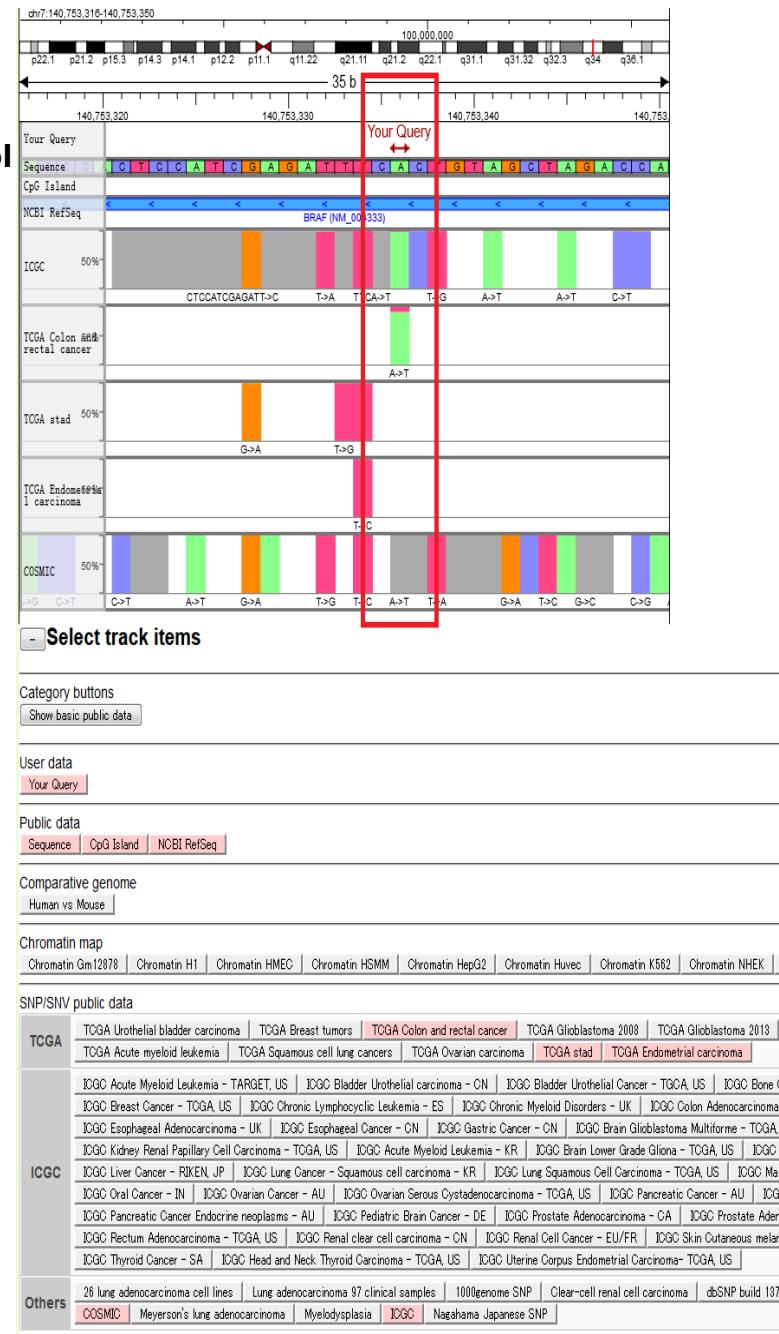
## RNA-Seq

# BS-Seq

# ChromHMM

## ChIP-Seq

-SNV

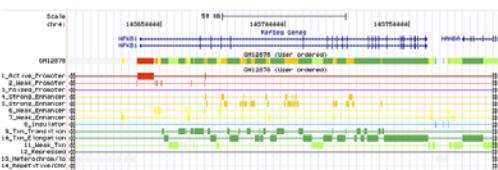


少し脱線

# ChromHMMとは

ChromHMM: Chromatin state discovery and characterization  
<http://compbio.mit.edu/ChromHMM/>

ChromHMM: Chromatin state discovery and characterization



ChromHMM is software for learning and characterizing chromatin states. ChromHMM can integrate multiple chromatin datasets such as ChIP-seq data of various histone modifications to discover de novo the major reoccurring combinatorial and spatial patterns of marks. ChromHMM is based on a multivariate Hidden Markov Model that explicitly models the presence or absence of each chromatin mark. The resulting model can then be used to systematically annotate a genome in one or more cell types. By automatically computing state enrichments for large-scale functional and annotation datasets ChromHMM facilitates the biological characterization of such states. ChromHMM also produces files with genome-wide maps of chromatin state annotations that can be directly visualized in a genome browser.

- [ChromHMM software v1.11 \(version log\)](#)
- [ChromHMM manual](#)

Quick instructions on running ChromHMM:  
1. Install Java 1.5 or later if not already installed.  
2. Unzip the file ChromHMM.zip  
3. To try out ChromHMM learning a 10-state model on the sample data enter from a command line in the directory with the ChromHMMjar file the command:

```
java -mx1600M -jar ChromHMMjar LearnModel SAMPLEDATA_HG18 OUTPUTSAMPLE 10 hg18
```

After termination in ~5-10 minutes a file in OUTUTSAMPLE/webpage.10.html will be created showing output images and linking to all the output files created. If a web browser is found on the computer the webpage will automatically be opened in it.  
In general binarized input for the *LearnModel* command can be generated by first running the *BinarizeBed* command on bed files with coordinates of aligned reads or the *BinarizeBam* command on bam files with the same alignment quality requirements.

New in version 1.11: ChromHMM has a *BinarizeBam* command which allows binarizing bam files of aligned reads.  
New in version 1.10: ChromHMM has the option for parallel training with multiprocessors leading to significantly reduced training times. Add the *-p 0* option to the *LearnModel* command to have ChromHMM to try to use as many processors as available or specify the maximum it should use.

- The ChromHMM software is described in:  
Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9:215-216, 2012.
- Here are links to some existing ChromHMM annotations in hg19 available for [127 Reference Epigenomes \(Roadmap Epigenomics\)](#), [9-ENCODE cell types \(from Ernst et al. Nature 2011\)](#), and [8-ENCODE cell types \(from ENCODE Integrative Analysis\)](#).
- Contact Jason Ernst ([jason.ernst@ucdavis.edu](mailto:jason.ernst@ucdavis.edu)) with any questions, comments, or bug reports.
- Subscribe to a [mailing list for announcements of new versions](#).
- ChromHMM is released under a [GPL 3 license](#).
- ChromHMM source code is available on [Github here](#).
- Funding for ChromHMM provided by NSF Postdoctoral Fellowship 0905969 to JE and grants from the National Institutes of Health (NIH 1-R01-HG005334 and NIH 1-U54 HG004570).

様々なヒストン修飾のChIP-Seqデータなどから、クロマチンの状態をパターン化してくれるソフトウェア

Ernst J and Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9:215-216, 2012.

ENCODEやRoadmap Epigenomics projectでも ChromHMMデータが公開されている。

```
$ cd  
$ cp /home/lect-1/ChromHMM/LC2ad_dense.bed .
```

でホームディレクトリにLC2/adのChromHMMの結果がコピーされます。  
WinSCPでPCにダウンロードして、IGVで見てみてください。

# Pathway Map

DBTSSに戻ります

**Top |**

---

### Database Search

Keyword Search  
Species:  
 ▾

Keyword:

Lift over:  
 ▾

---

### Human Chromatin Features

Search  
Search from Genomic Position:

Search from SNP (dbSNP rsID):

Search from SNV (COSMIC: somatic mutation):

---

Search from SNV-enriched Gene in Cancers:  
 ▾

SNV Summary in Cancers:

---

### Pathway Map

Species:  
 ▾

Pathway M  
Searchをク

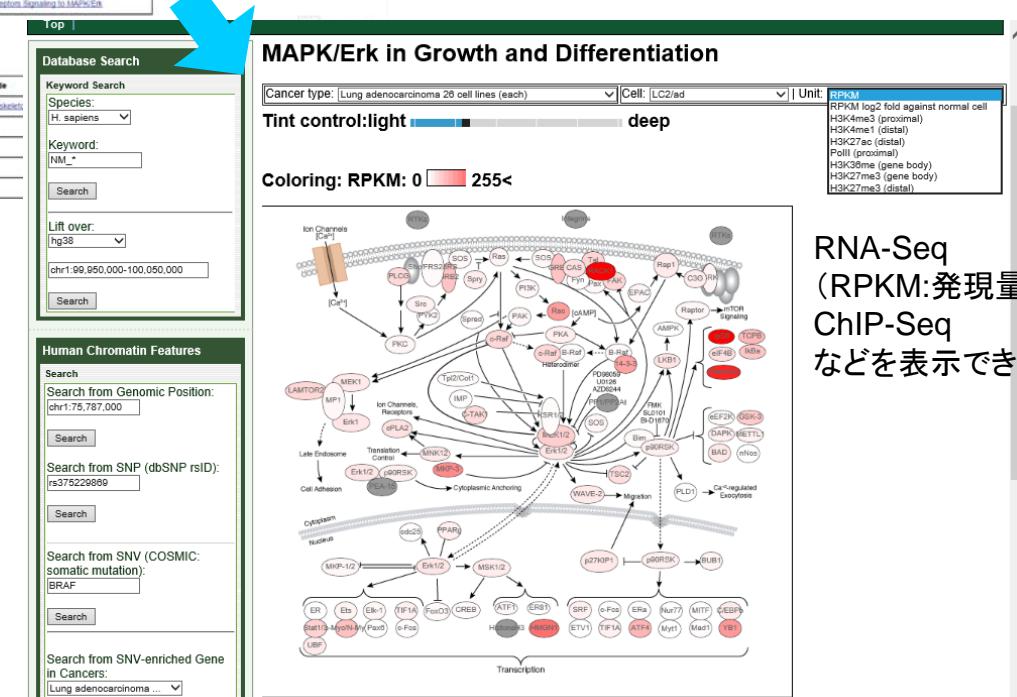
---

### Documents

- Experimental Procedures
- Data Contents

## CSTやKEGGのpathwayリスト

## MAPK/Erk in Growth and Differentiationをクリック



RNA-Seq  
(RPKM:発現量)  
ChIP-Seq  
などを表示できる

# ぜひDBTSSを使ってください

# Acknowledgements

東京大学

スパコンサポート係の皆様

中井謙太先生

宮野悟先生

鈴木穰先生

東京大学医科学研究所ヒトゲノム解析センターの  
スパコンのアカウントを手配してくださいました。

The screenshot shows the HGC website's Supercomputer page. At the top, there are links for "Top500", "Supercomputer", "Database", "Download", and "Supercomputer". A search bar for "HGC-Site Search" is also present. The main content area features a blue banner with the text "Supercomputer". Below the banner, there is a news section with several entries in Japanese. To the right, there are two columns of links under "Utilization Case" and "Utilization Method, Seminar". At the bottom, there is a "Twitter" section with a tweet from "Supercomputer@HGC" (@schgc) dated September 25, 2014.

鳥取大学の皆様

スパコン接続のためにご助力くださりありがとうございました。

本日はありがとうございました  
ご質問・コメント等がありましたら  
遠慮なくお願いします。

