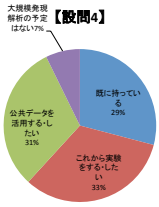
# 統合データベース講習会

In 群馬大学

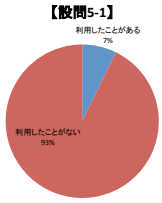
Masaki Suimye Morioka

## アンケート結果から

【設問4】自分で実験して得た、数十?数千の遺伝子からなる「遺伝子リスト」(例: 発現差のあった遺伝子など) を持っていますか。




【設問5-1】NGS解析ソフトウェアの利用経験について




2

## 本日の講習内容

1. NGSについての簡単な説明 (illumina型中心)
   - （スライドで約1時間ほどの予定）
2. NGS解析のチュートリアル（ChIP-seq）

利用した画像について

Togo picture gallery by DBCLS is licensed under a Creative Commons Attribution 2.1 Japan license (c)

参考にさせていただいたweb site

DBCLS 大田さんの「GalaxyによるNGS解析」(https://github.com/inutano/training/tree/master/ajacs-advanced-01)
株式会社ジナリス、竹田さんの「いまさら聞けないNGS超!入門」
http://www.slideshare.net/Genaris/ngs-46348501

3

## HiSeq X Ten System

$1,000 genome sequencer, illumina HiSeq X ten

- Ten HiSeq X per one system.
- $1,000/genome (including chemistry, personnel costs)

Jan. 2014. Press in illumina seminar

実物とは異なります

Togo picture gallery by DBCLS is licensed under a Creative Commons Attribution 2.1 Japan license (c)

4

## ゲノムの塩基数とカバー率

Remember this....

1 billion = 100Mb　~~Byte~~ Base!

A human genome: 30 billion base = 3Gb

Human whole genome sequencing:
commonly required over x20-30 coverage.

3Gb x 30 = 90 Gb / human

5

## Performances of HiSeq X ten*

HiSeq 2500 (Rapid mode, v3)

TP: 100Gb x3
Reads: 最大10億
Time: 27時間 x3

High coverage mode
11days...

(high coverage mode)

HiSeq X ten system

TP: 1.6Tb-1.8Tb
Reads: maximum 60 billion
Time: 3日間

1 week after...

320 peoples

x20

* Based on Illumina web site information

6

## HiSeq SBS Kit v4 and HiSeq Cluster Kit v4

HiSeq SBS Kit v4 (May 2014)

- 167Gb/Day
- 1Tb / 6 Day, 125 bp, max 40 billion reads (max 500Gb / flow-cell)
- Cluster density is increased 33% (High coverage mode )

- 10 samples in human whole genome sequencing
- 150 samples in Exome-seq
- 80 samples of transcriptome.

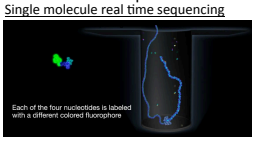
From illumina web site information

7

## NGSの最近の動向

より長く塩基配列を決定したい

Pacbio: SMRT-Seq
Single molecule real time sequencing



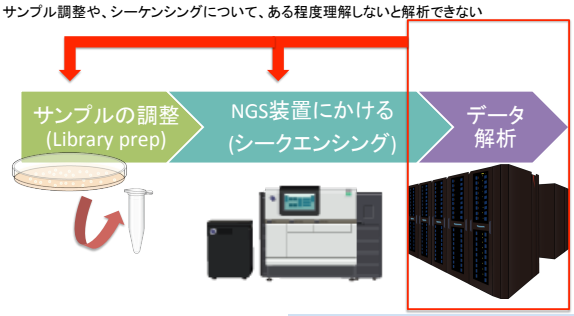

★Oxford NANOPORE


10X genomics


各社のwebサイトより転載させていただきました。

## NGS解析の流れ

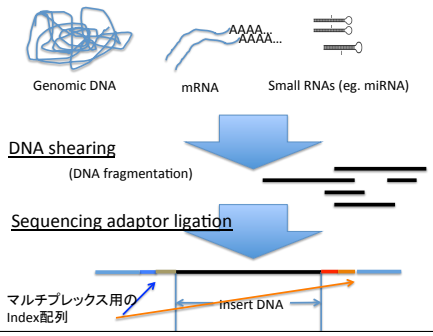
サンプル調整や、シーケンシングについて、ある程度理解しないと解析できない





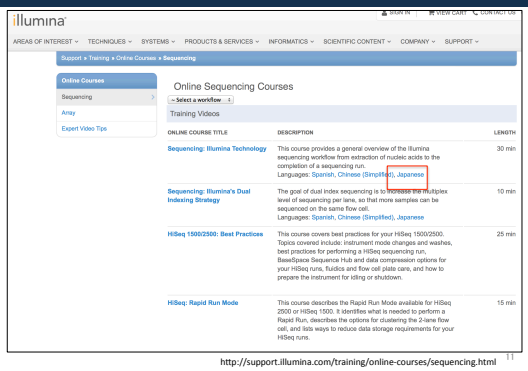

## NGS解析のサンプル調整

DNA断片化を介したsequencing用DNA libraryの調整




10

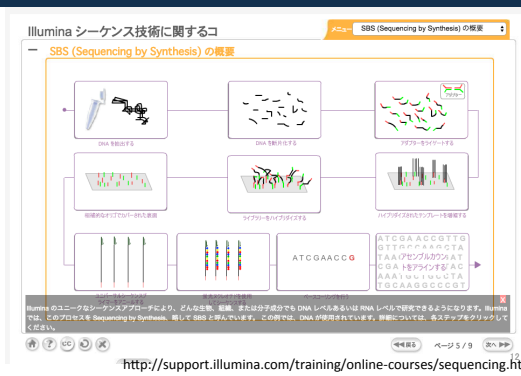
## Illumina社のonline学習




http://support.illumina.com/training/online-courses/sequencing.html

11

## Illumina社のonline学習




http://support.illumina.com/training/online-courses/sequencing.html

12

# NGS解析のサンプル調整

DNA断片化を介したsequencing用DNA libraryの調整

Sequencing DNA library

Library内のDNA lengthは大体一定のサンプルを用いる事が多い

決定された塩基配列 (read)

templateDNA

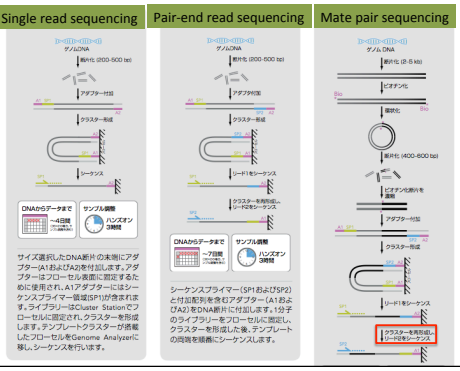
総リード数 x リード長 = 塩基配列の決定塩基数

13

# Differences in Sequencing strategies

On illumina platform

| Single read sequencing | Pair-end read sequencing | Mate pair sequencing |
|---|---|---|



14

# Sequencingの原理

Sequencing qualityを理解するためには原理を理解しておかなくてはいけない

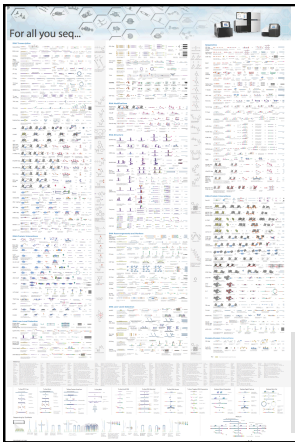
15

イルミナ社のSBS法を例に: https://www.youtube.com/watch?v=HMyCqWhwB8E

# NGSのアプリケーション

- WGSの簡単な説明
- ChIP-seq
- RNA-seq

Masaki Suimye Morioka




# 核酸検出をどう工夫して、調整するか

http://www.illumina.com/techniques/sequencing/ngs-library-prep/library-prep-methods.html?scid=2014249PT1

# Whole genome sequencing

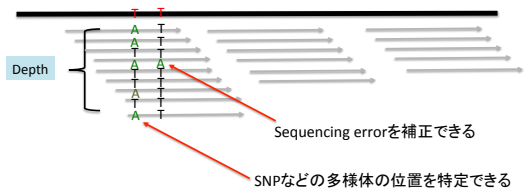
サンプルの全ゲノムをシーケンシングを用いて決定する

- De novo assembly
  - 参照ゲノムへのmappingを行わず得られた塩基配列のみで、全ゲノム構造の再構成を行う
- whole genome resequencing
  - 参照ゲノムへのmappingを行う事で、一塩基対バリアント (single nucleotide variant)や、CNV (copy number variant)などの変異検出を目的とする

18

## Whole genome sequencing

リファレンスへのマッピングに基づく解析

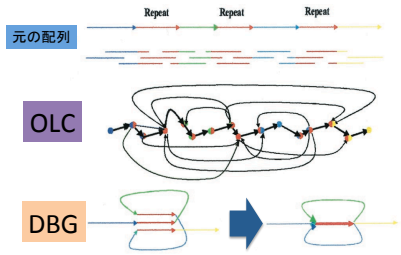



30x: マップされたリードの合計長の割合が参照ゲノムの全ゲノムの長さの30倍

19

## アライメントアルゴリズムの違い

事例: 繰り返し配列が多いゲノム領域




Pavel A. Pevzner et al. PNAS 2001;98:9748-9753

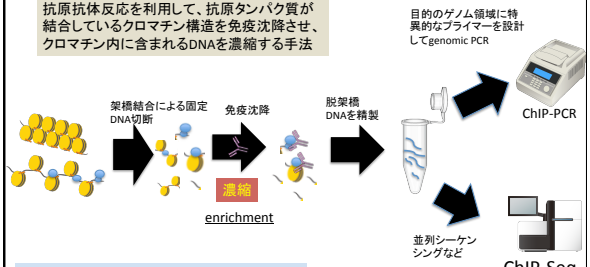
PNAS

20

## ChIP-Seqとは

ChIP: chromatin Immuno precipitaton

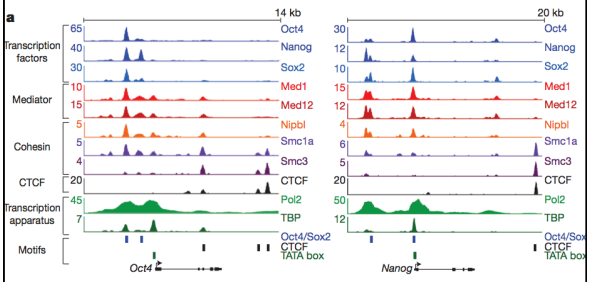
抗原抗体反応を利用して、抗原タンパク質が結合しているクロマチン構造を免疫沈降させ、クロマチン内に含まれるDNAを濃縮する手法






21

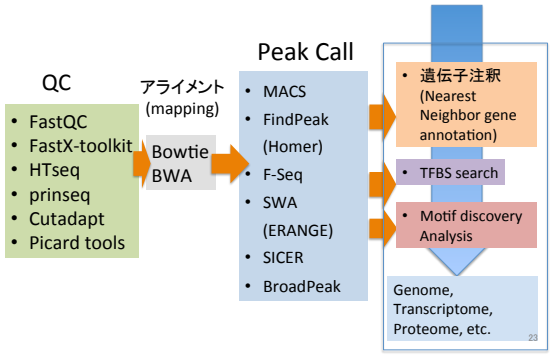
## ChIP-seq analysis of Mediator complex




Mediator and cohesin connect gene expression and chromatin architecture. Michael H. Kagey et al. (nature 2010).

22

## ChIP-Seqデータ解析の流れ




23

## Peak Call

Enrichされた領域を山の頂点としてみたい

Broad peak (eg. H3K27Ac, K3K9… )


SICER, BroadPeak

Narrow and sharp peak with low noise (TFs)


MACS, F-Seq, SWAs

S/N is non-good or low peak (FAIRE-Seq, DNase-Seq)

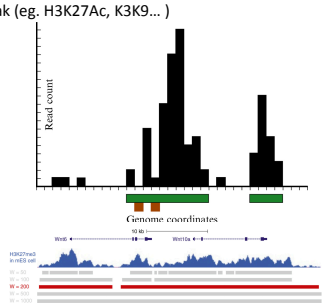
F-Seq, ZIMBA

24

## 1. SICER

SICER ヒストン修飾をはじめとするBroad peak用のpeak caller

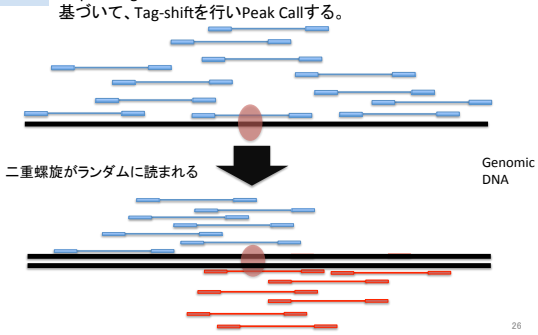
Broad peak (eg. H3K27Ac, K3K9… )




Shiliyang Xu et al. *Methods Mol Biol*. 2014 ; 1150: 97–111. doi:10.1007/978-1-4939-0512-6_5.

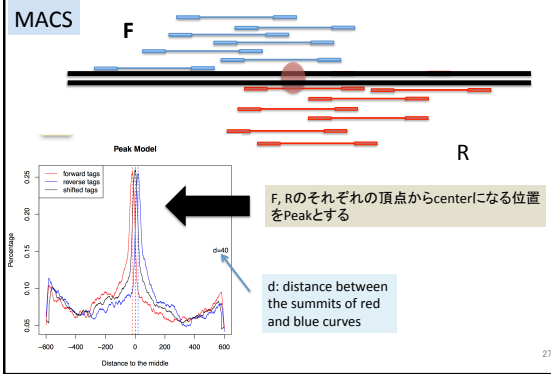
## 2. MACS

MACS Sequencingによって得られるタグの分布形状を加味したモデルに基づいて、Tag-shiftを行いPeak Callする。




## 2. Tag shift

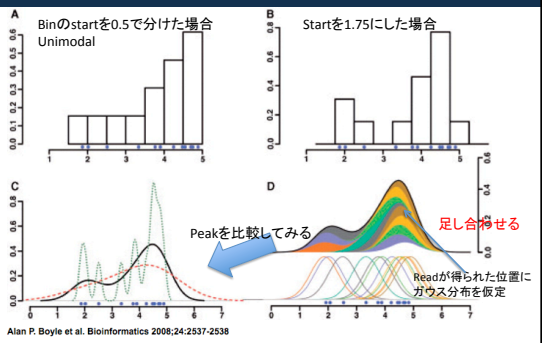
MACS




F, Rのそれぞれの頂点からcenterになる位置をPeakとする

d: distance between the summits of red and blue curves

## 3. F-Seq

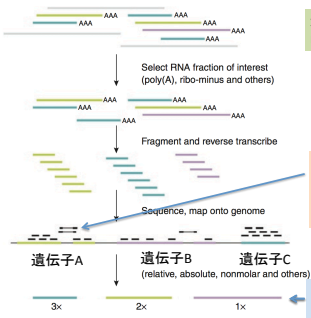



Alan P. Boyle et al. Bioinformatics 2008;24:2537-2538



Bioinformatics

## RNA-seqとmRNA発現量




オリゴdT beadsを使った RNA-Seq (PolyA-Seq)

Read配列が、ジャンクションを挟んでいる場合があるので、ギャップを入れてマッピングをおこなう junction mappingを行う

この例の場合の発現データはReadをタグとしたカウント数となる

この例でのTotal Read数: 34

## RPKM値

RPKM: Reads Per Kilo base-pair per Million mapped reads

ある転写産物(*A*)の1kb当たりの発現量として、試料中の核酸を百万回読んだ時に相当するリード数を発現量とする

$$RPKM(A) = n_a \frac{10^3 \cdot 10^6}{l_a \cdot N}$$

転写産物Aが読まれた数$n_a$

転写産物Aの長さ$l_a$を1.000 basesに規格化

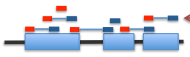
総リード数を100万回に正規化

補足: RPKM値は、導出可能である

## FPKM値

FPKM: Fragments Per Kilo base-pair per Million mapped reads

ペアエンドのデータに対応したRPKM値と考えてください



ペアエンドのデータを1つのフラグメント（DNAライブラリ作成時のインサートフラグメント）として捉えてカウントする

31

## NGSデータの特徴のまとめ

Introduction of NGS analysis

1. 塩基配列である
2. Read (読み取った配列)が体積した山になる
3. 試料のプールからシーケンサでランダムに読まれていく
4. シーケンシングされるread数は、決まったread数にはならない (試料ごとにTotalのread数が異なる)

1. アライメントする、アッセンブリ — 1. Genome mapping, Genome assembler
2. 山をみつける — 2. Peak Caller
3. ランダムサンプリングされていることを前提にする — 3. Poisson Distribution
4. 全体のリード数を整えてから、NGSデータ間の比較 — 4. RPKM FPKM

32

## NGSのファイル形式

- Sequencing quality
- Mapping quality

33

## NGS解析から得られるデータ

Raw data

例
- Illumina型: 画像
- PacBio型: 動画

Band Pathfilter
BaseCall

配列データ

例 Fastq形式 (.fastq)
HDF5形式 (.h5)

シークエンサから出力された情報をもとに塩基が決定され、その塩基配列情報の品質評価値とともに文字情報として出力される。

シークエンサから出力された塩基情報のもっともらしさ

Phred Scale

34

## Fastqファイル

(sanger形式)

```
head sample.fastq
```

ファイルの中身を先頭から数行抜き出すunixコマンド

|1行目: リードID
|2行目: 塩基配列
|3行目: +（とコメント）
|4行目: クオリティスコア

```
@SRR445816.18948743 HWUSI-EAS366_145:3:39:5673:1812/1
ACTCCTGCCTCAGGTGATCCATCCGCCTCAGCCTCT
+
BDGGG@GDGGBG@DDE@ABFGGGB<CFEAE<:???=
@SRR445816.28703832 HWUSI-EAS366_145:3:61:8637:6699/1
ACTGACTCAAATGTTAATCTCCTTTGGCAACACTCT
+
IIIIIIIIIIIIIIIIIIIIIIIEIIIIIIIIIIIII
@SRR445816.43211059 HWUSI-EAS366_145:3:94:3572:7097/1
TAAGCCCCTTCTCTTAGGATTTATAACCTCATCACT
```

35

## Phred Score

General NGS data format

$$Q_{phred} = -10\log_{10}(P)$$

| Phred Score | 塩基が違う確率(P) | 正確性 |
|---|---|---|
| 10 | 1/10 | 90% |
| 20 | 1/100 | 99% |
| 30 | 1/1000 | 99.9% |
| 40 | 1/10000 | 99.99% |

Phred Scoreは、アスキーコードで書かれており、1文字に対応する数字が記載される。
例えば、文字 'a' は 97に相当する。
サンガー形式のクオリティ値は、アスキーコードの数字から33引いた値（文字'a'は65に相当する）

```
@HWUSI-EAS1869:80:FC66STUAAXX:4:1:1329:1035 1:N:0:
GTCCCTTAAGTACTGGTCAATGAGGGCCTCGTGATCCA
+
GG>>EEEEBEG@GGGDEEEEEDE=>GGGBBGGEGGGG<
```

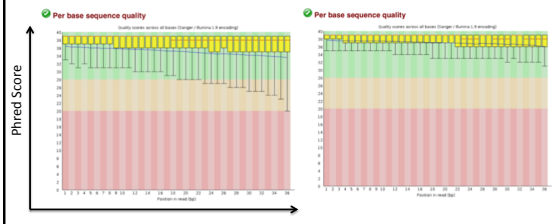
大文字"G"は、71なので、71 − 33 = 38
文字">"は、62なので、62 − 33 = 29

36

# 読み取り長とPhredScore

イルミナ型のシークエンサを用いた時に得られた36塩基を例に

FastQCを用いた品質評価の結果の一例

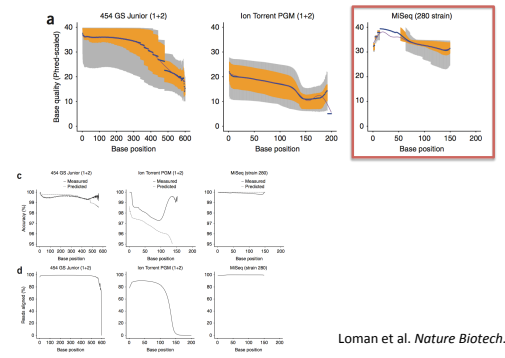



キャピラリーシーケンサのころと同様に、読まれた塩基の長さが長くなるとともに、塩基配列決定の精度が低下する傾向がある

37

# 読み取り長と塩基配列決定の品質

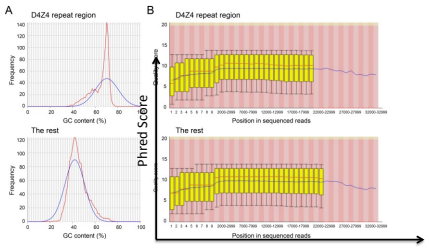
古い情報ですが、分かりやすい例として




Loman et al. *Nature Biotech.* 2012.

# 読み取り長と塩基配列決定の品質

PacBioを用いた時に得られた塩基を例に

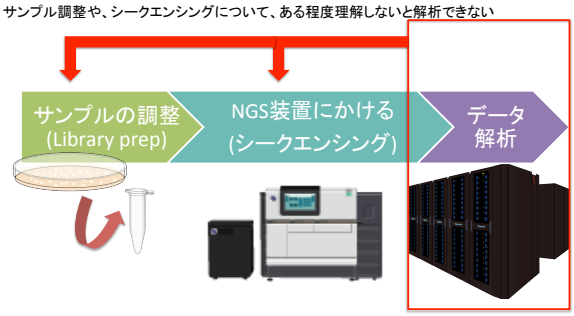



PacBio型では、塩基配列決定の精度は、GC含有率や、シーケンシング長にほとんど影響されない。

Masaki Suimye Morioka et al. (*PLOS One* 2016)

# NGS解析の流れ

サンプル調整や、シークエンシングについて、ある程度理解しないと解析できない




40

# NGSデータ解析のロードマップ

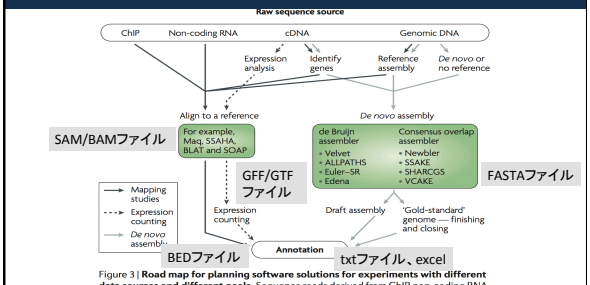



Figure 3 | **Road map for planning software solutions for experiments with different data sources and different goals.** Sequence reads derived from ChIP, non-coding RNA ... sequencing experiments are aligned to a reference sequence before ... counting and final annotation. Sometimes, a cDNA sequence can be ... *de novo* before these steps. Genome sequence reads may be aligned if a ... is available, but if not assembly *de novo* can still be carried out.

可視化
BigWigファイル
tdfファイル
(bam, bed, gff etc..)

Daniel MacLean et al. Nature reviews Microbiology 2009

41

# SAM/BAMファイル

リファレンスにアライメントされていないリードの情報も含む

SAM (Sequence Alignment/Map) ファイルは、シークエンサから得られたリード配列がゲノムDNAのどの領域に由来するのかを調べるために、既に読まれたゲノム配列をリファレンスとしたアライメントを行った時に得られるアライメント結果のファイル。

```
iu@bio[tutorial150806] head -n5 sample.sam          [12:12午前]
```

Header行

```
@HD VN:1.0  SO:unsorted
@SQ SN:chr17    LN:81195210
@PG ID:bowtie2  PN:bowtie2  VN:2.2.4    CL:"/usr/bin/../lib/bowtie2/bin/bowtie2-align-s --wr
SRR445816.15448471  0   chr17   5102645 1   36M *0  0   CTCCCAGGCTGGAGTGCAGTGGTGCAATCTTGGCTC
SRR445816.28703832  0   chr17   10949619    1   36M *   0   0   ACTGACTCAAATGTTAATCTCCTTTGGC
iu@bio[tutorial150806]
```

例では、36塩基全てがmatchという意味

- Read Name
- SAM flag
- chromosome (if read is has no alignment, there will be a "*" here)
- position (1-based index, "left end of read")
- MAPQ (マッピングクオリティ、0:non-unique, >10 probably unique)
- CIGAR string (insertion (i.e. introns)がリードの中にあるか否かなどに重要。例えば、splicing junctionの決定などにも利用)
- Name of mate (mate pair information for paired-end sequencing, often "=")
- Position of mate (mate pair information)
- Template length (always zero for me)
- Read Sequence

http://samtools.github.io/hts-specs/SAMv1.pdf

42

# BEDファイル

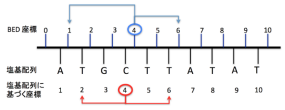
リファレンスにアライメントされていないリードの情報も含む

領域の座標情報を保持しているファイル（最低3列）
Sam/bamファイルから作成可能。

染色体一番の1塩基目：c

```
browser position chr7:127471196-127495720
browser hide all
track name="ColorByStrandDemo" description="Color by strand demonstration"
visibility=2 colorByStrand="255,0,0 0,0,255"
chr7    127471196  127472363  Pos1  0  +
chr7    127472363  127473530  Pos2  0  +
chr7    127473530  127474697  Pos3  0  +
chr7    127474697  127475864  Pos4  0  +
chr7    127475864  127477031  Neg1  0  -
chr7    127477031  127478198  Neg2  0  -
chr7    127478198  127479365  Neg3  0  -
chr7    127479365  127480532  Pos5  0  +
chr7    127480532  127481699  Neg4  0  -
```

1列目：染色体やリファレンス配列の名前
2列目：region start
3列目：end
4列目：領域のidや名前
5行目：何らかのスコアなど
6行目：strand




https://genome.ucsc.edu/FAQ/FAQformat.html 43
http://qiita.com/suimye/items/e4d067d04d89146220 7e

# GFFファイル

リファレンスにアライメントされていないリードの情報も含む

General Feature Format (GFF)

ゲノムの特徴について一行で記載されているファイル。現行は、GFF3。GTF (General Transfer Format)は、GFF2と同一。

大きく3つの構成要素
- Fields
- Track lines
- More information

Sample GTF output from Ensembl data dump:

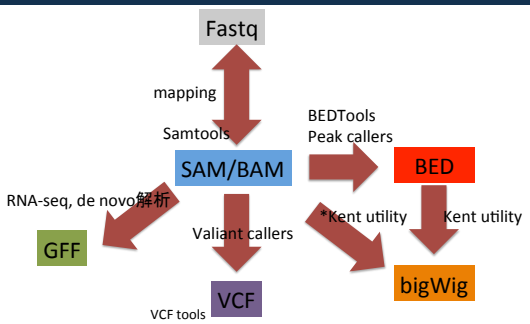
```
1 transcribed_unprocessed_pseudogene  gene        11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"
1 processed_transcript                transcript  11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST
```

Sample GFF output from Ensembl export:

```
X       Ensembl Repeat  2419108 2419128 42      .       .       hid=trf; hstart=1; hend=21
X       Ensembl Repeat  2419108 2419410 2502    -       .       hid=AluSx; hstart=1; hend=303
X       Ensembl Repeat  2419108 2419128 0       .       .       hid=dust; hstart=2419108; hend=2419128
X       Ensembl Pred.trans.     2416676 2418760 450.19  -       2       genscan=GENSCAN00000019335
X       Ensembl Variation       2413425 2413425 .       +       .
X       Ensembl Variation       2413805 2413805 .       +       .
```

http://www.ensembl.org/info/website/upload/gff.html

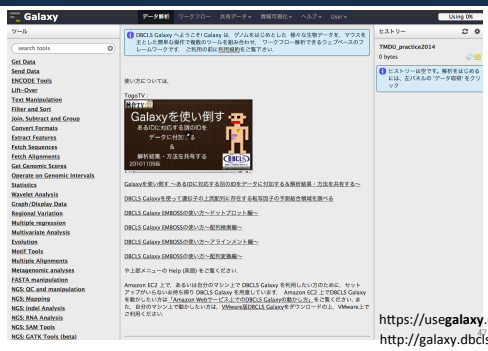
# NGS データ変換




*http://hgdownload.soe.ucsc.edu/admin/exe/

# 解析環境の選択

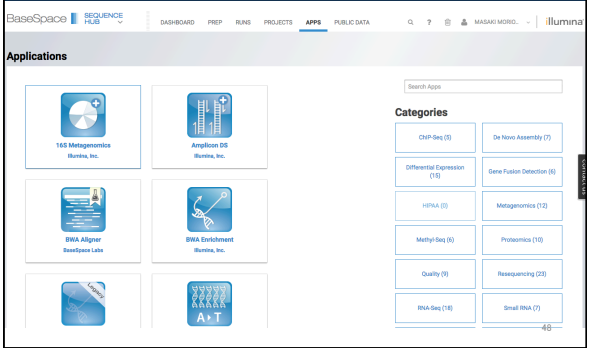
# 統合解析環境
# galaxy



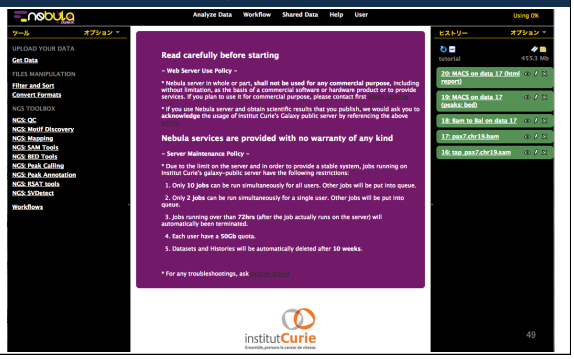

https://usegalaxy.org
http://galaxy.dbcls.jp

# Illumina BaseSpace

https://basespace.illumina.com/apps/



## GalaxyからChIP-seq専用に作成された nebula (http://nebula.curie.fr)




## ChIP-seqデータをとにかく可視化

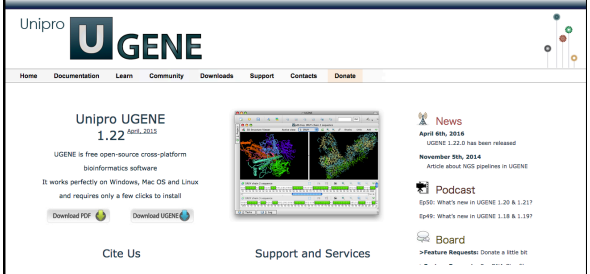
http://www.devbio.med.kyushu-u.ac.jp/sra_tailor/

Sratailer


http://togotv.dbcls.jp/20150106.html

Asperaを利用しているので、proxyのsettingを忘れずに

## 統合解析環境UGENE




NGS解析には別途、BWA, Tuxedo Toolsなどのinstallが必要であるが、piplineを作れる