

2016/09/02 統合データベース講習会：AJACS京都2

# NGSデータから新たな知識を導出するための高次解析

**尾崎遼** 理化学研究所 情報基盤センター バイオインフォマティクス研究開発ユニット 基礎科学特別研究員

haruka.ozaki@riken.jp

<http://yuifu.github.io>

# はじめに

本日の資料: [https://github.com/yuifu/AJACS Kyoto 2](https://github.com/yuifu/AJACS_Kyoto_2)

アクセスできるか確認してみよう

11:10～15:20 「NGSデータから新たな知識を導出するための高次解析」（紙配付資料有）  
→解析用コード  
。尾崎 遼（理化学研究所情報基盤センター）

## 本日の流れ

11:10-12:20: 講義

13:30-15:20: 講義+実習

\* 講習の途中に適宜ハンズオンで手を動かします

いつでも質問を受け付けます

よくわからなかったらいつでも止めてください

ただ座っているだけではもったいない



chrome

# 講習の目的・到達目標

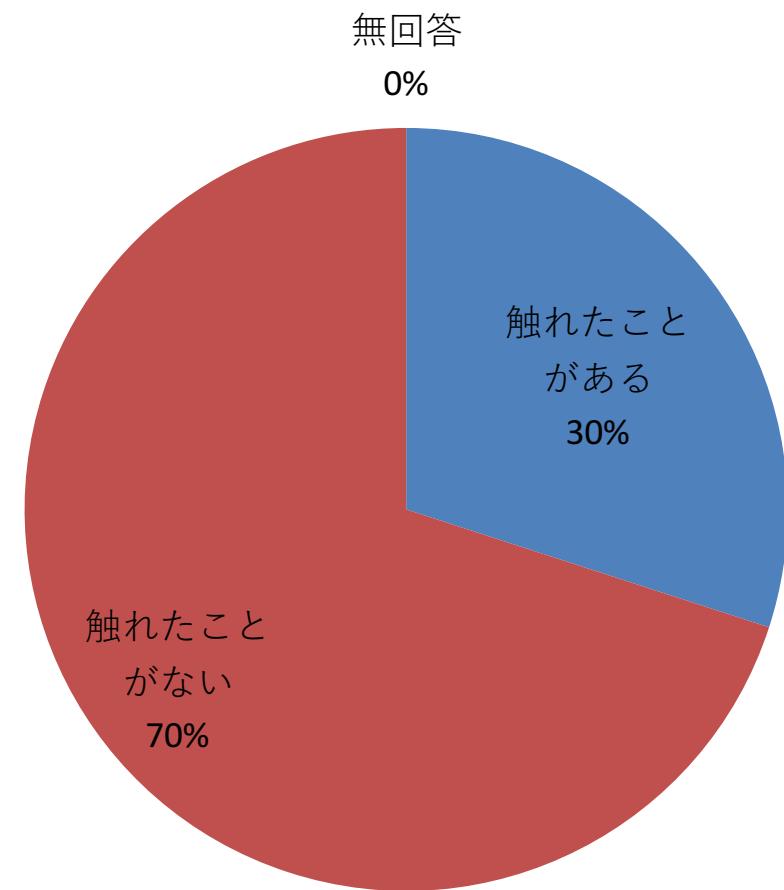
## 目的

自ら取得した（または公共データベースから取得した）  
NGS低次解析済みデータの高次解析の方法を学ぶ

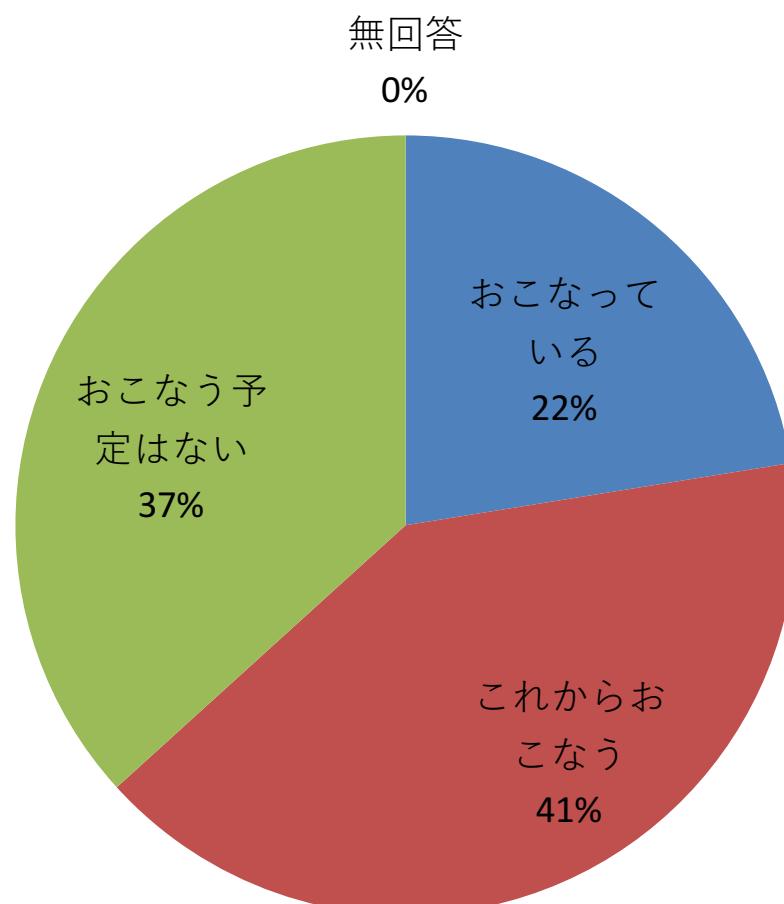
## 到達目標

NGSデータの形式を理解する  
NGSデータ高次解析の基礎知識を理解する  
自分でコマンドを打って高次解析を体験する  
自身の問題意識に応じて解析の方法を調べられるようになる

【設問5-1】NGSデータに触れたことはありますか。



【設問6-1】NGS解析をおこなっている、もしくは、おこなう予定がありますか。



# この講習の内容

NGSデータの高次解析とは

メリット

NGSデータの低次解析とは

低次解析済みのデータを取得する

NGSデータの高次解析

データフォーマット

ツール

統計手法

NGSデータの高次解析の体験

NGS高次解析の自主学習の方法

# NGSデータ解析の選択肢

方法	導入	どこまでできるか	価格
オープンソース ソフトウェア	慣れていないと時間 かかる	世の中にあるソフト ウェア次第	無償
ウェブサービス	簡単	提供されている機能 次第	無償（一部有償）
有償ソフトウェア	簡単	提供されている機能 次第	有償
ドライの研究者との 共同研究	人次第	基本的なこと～オー ダーメイド（人次第）	無償
企業に解析を受託	簡単	基本的なことが中心 (業者による)	有償

# NGSデータの高次解析とは

## NGSを利用した研究の一般的な流れ

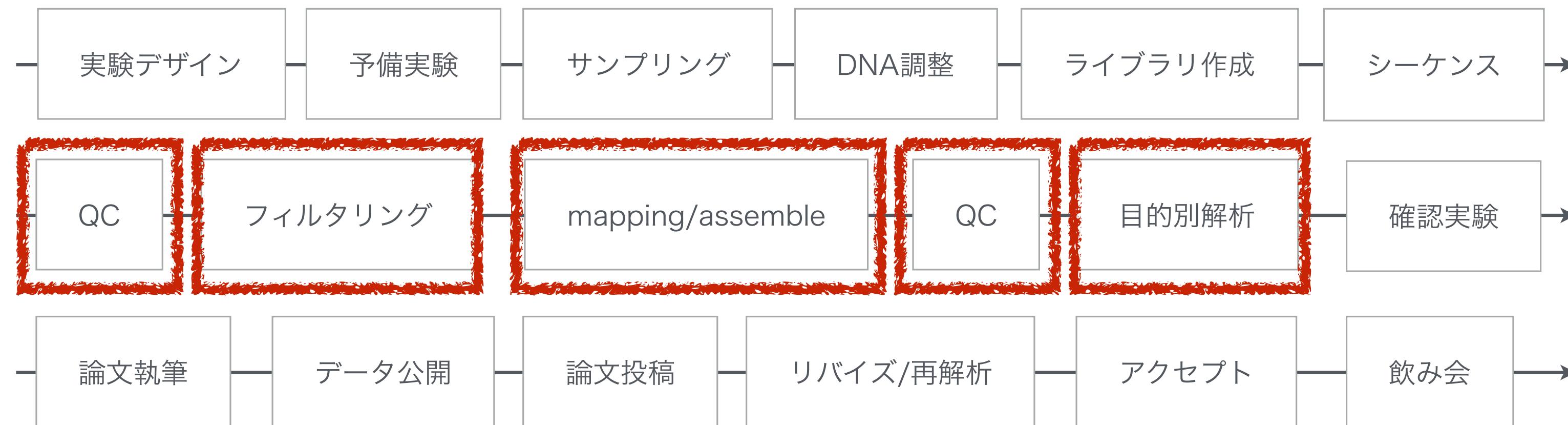
NGSを使う研究は何が大変なのか



- ・ イメージ
  - ・ 機械が高い
  - ・ データが沢山出る
  - ・ データ解析がよくわからない

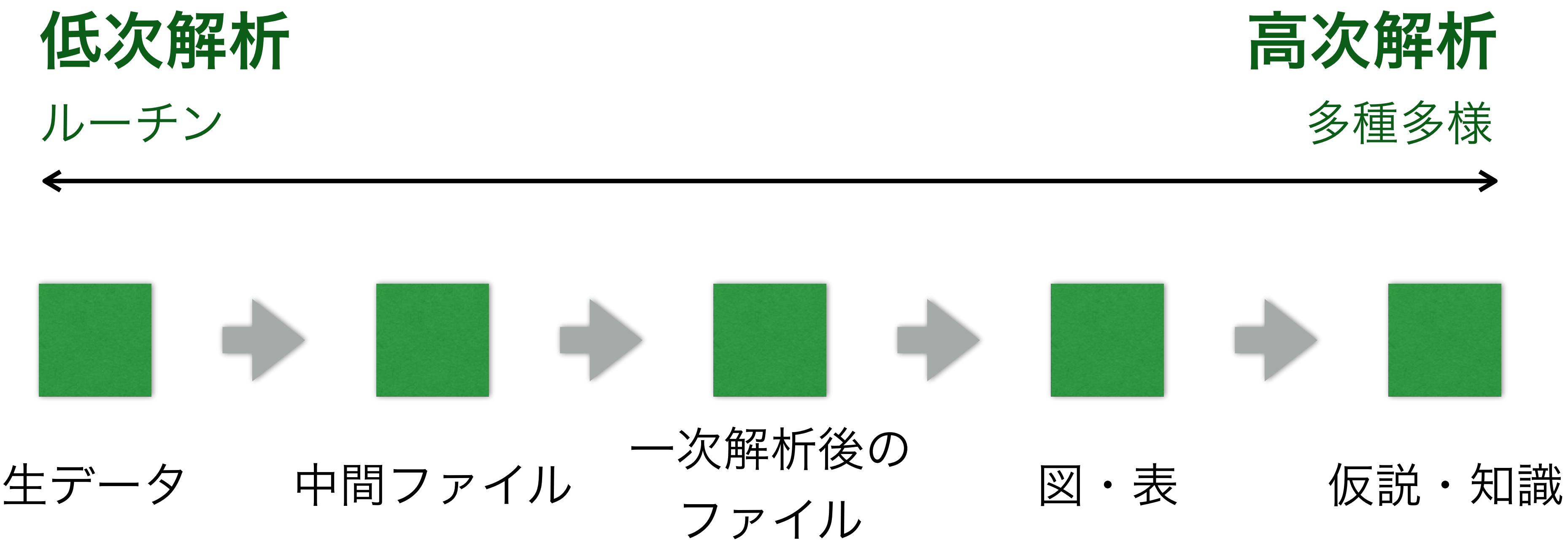
## NGSを利用した研究の一般的な流れ

NGSを使う研究は何が大変なのか



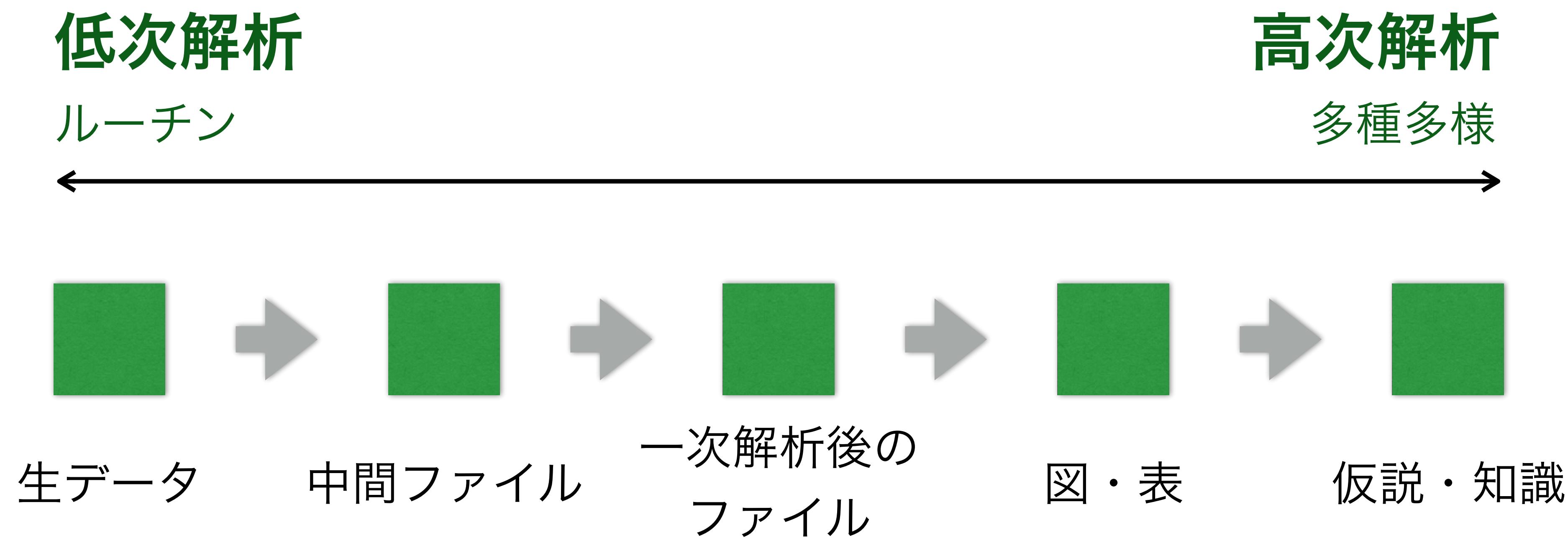
- 実際
  - 飲み会が遠い

# 高次解析？



# 知識・仮説を導出するには高次解析が必須

低次解析（マッピング、発現量定量、発現変動遺伝子検出、ピーク検出）はルーチンだが、そこから知識（＝論文）に結実させるために試行錯誤が始まる



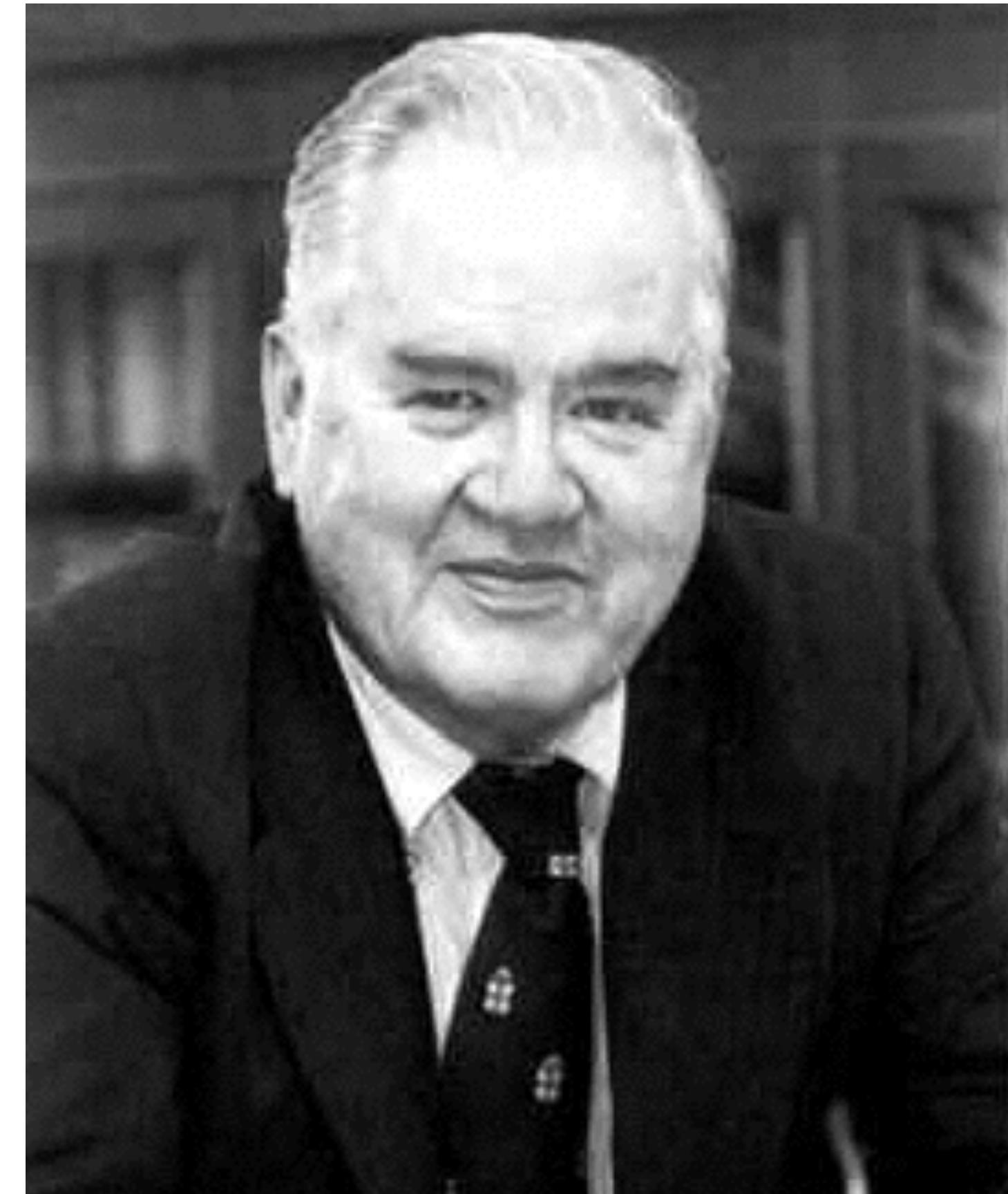
# 探索的データ解析 Exploratory data analysis

探索的データ解析 (=データから仮説を導出する) を提唱

確証的データ解析 (confirmatory data analysis) = 仮説検証への偏重を批判

まずデータを眺める

可視化テクニックでデータの特徴を観察する重要性を説く



**John Wilder Tukey**

# 探索的データ解析としてのNGSデータ高次解析

低次解析済みのデータを得るところから始まる

マッピング、発現量定量、発現変動遺伝子検出、ピーク検出などは準備段階

データを観察する

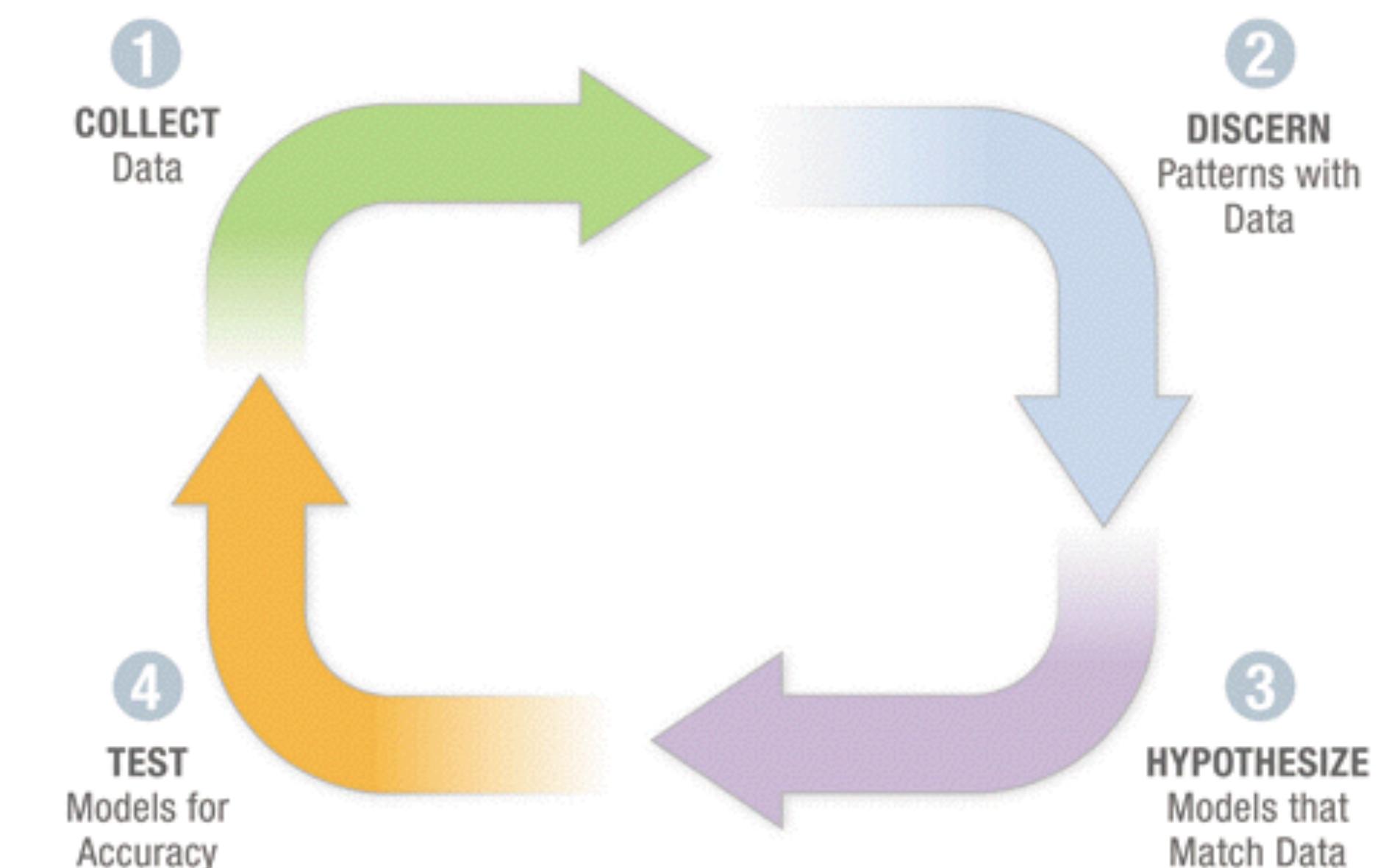
要約、可視化、濃縮

データを分類する

クラスタリング、層別、他のデータと統合

仮説を導出・検証して知識を引き出す

何度も試行錯誤できる技術を身につける



# 2016年のNGSをとりまく状況

NGSでデータを取得するのは当たり前 = 高次解析で差がつく

NGSを使っただけでトップジャーナルに載る（もっと言うと論文になる）時代ではない

\*-Seq技術開発は日進月歩

キーワード：微量化（一細胞）・同時化（RNAとメチル化など）・大規模化（数万細胞サンプルを取得）・多様化（多様な修飾塩基、翻訳後修飾）

→ ルーチンの解析（処理）+個々の実験手法・テーマに適した解析が必要

多様なNGS解析ソフトウェアがリリースされて続いている

どれを使ったらいいかキャッチアップするだけで大変

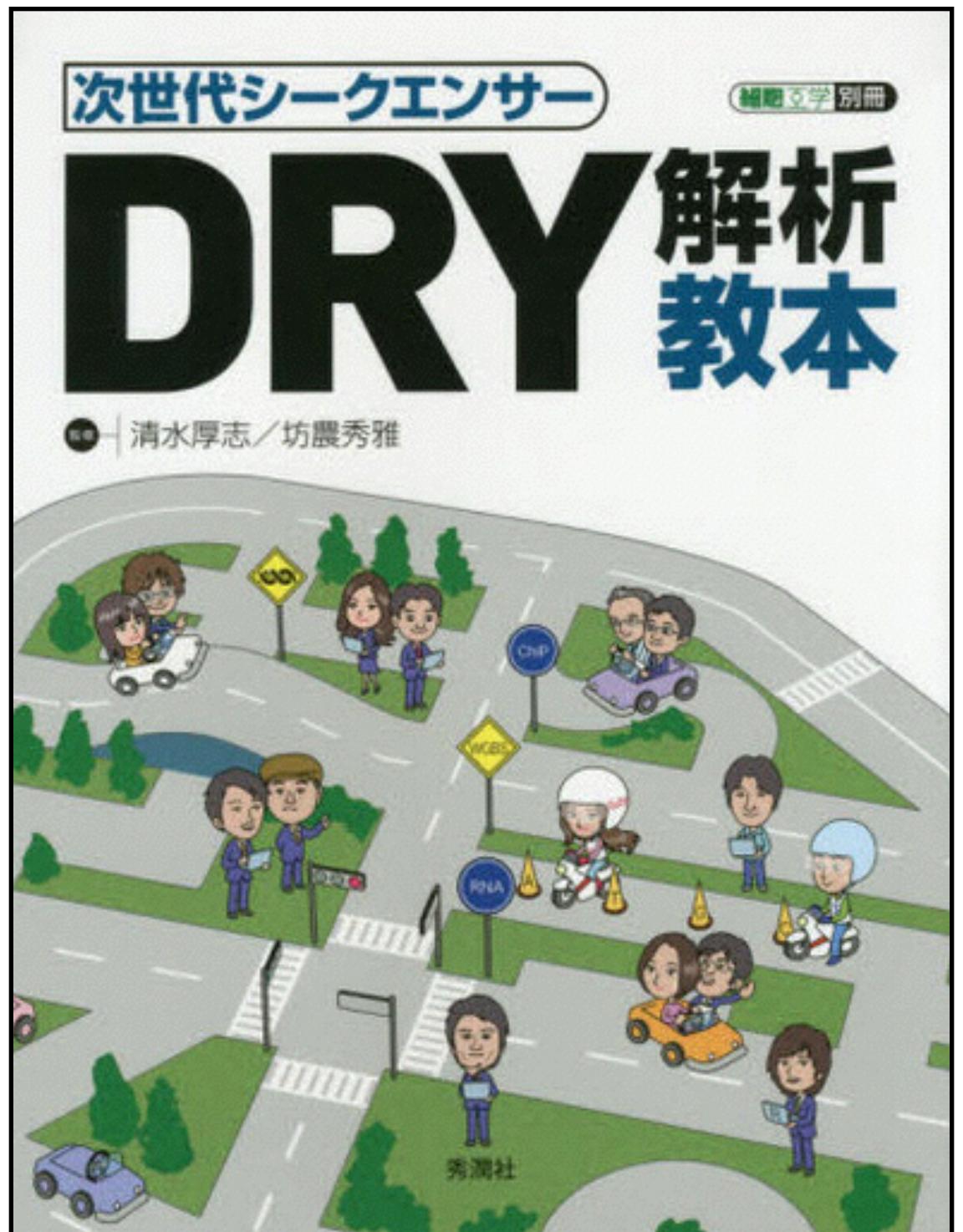
# 2016年のNGSをとりまく状況: 資料の充実

## 次世代シークエンサーDRY解析教本

## 平成28年度NGSハンズオン講習会

<http://biosciencedbc.jp/human/human-resources/workshop/h28-2>

The screenshot shows the NBDC (National Bioscience Database Center) website. At the top, there's a banner for the '次世代シークエンサーDRY解析教本'. Below it, the main navigation menu includes links for Home, NBDCについて, 研究開発, 公募情報, 採用情報, 広報, 人材支援, お問い合わせ, and リンク. The '人材支援' link is highlighted. The page content is about the '平成27年度NGSハンズオン講習会', which was held to improve on the previous year's events. It includes sections for 'H27年度概要', 'H27年度講義日程・参考資料', 'H26年度講習会の情報', 'H27年度実施報告書・講義資料・動画等', and download links for '講習会実施報告書 (PDF: 2.17MB) および受講者アンケート集計結果 (データ集) (PDF: 662KB)' and '講義資料・動画'. A note at the bottom says that clicking on file names will download PDF files.



# 2016年のNGSをとりまく状況: データの蓄積

公共レポジトリ (NGSデータをみんなが登録するところ)

SRA, ENA, DRA

大規模プロジェクトのポータルサイト

ENCODE Project <https://www.encodeproject.org>

Roadmap Epigenomics project <http://www.roadmapepigenomics.org>

modENCODE <http://www.modencode.org>

FANTOM <http://fantom.gsc.riken.jp>

二次データベース (まとめサイトみたいなもの)

Cistrome <http://cistrome.org>

ChIP-Atlas <http://chip-atlas.org>

EBI metagenomics <https://www.ebi.ac.uk/metagenomics/>

# NGSデータの低次解析とは

# 高次解析のためのファイルを作るのが低次解析

低次解析の終わりは高次解析の始まり

RNA-Seqだったら、サンプル × 遺伝子の発現量の行列

ChIP-Seqだったら、転写因子結合部位の位置とピーグ強度の表

メタゲノム解析だったら、サンプル × taxon のリードカウントの行列

高次解析は目的別

データをこねくり回すことで見えてくることがある

可視化が大切

# 一般的なNGSデータの低次解析

マッピング前

リードのクオリティチェック (QC)

リードのフィルタリング

リードのマッピング

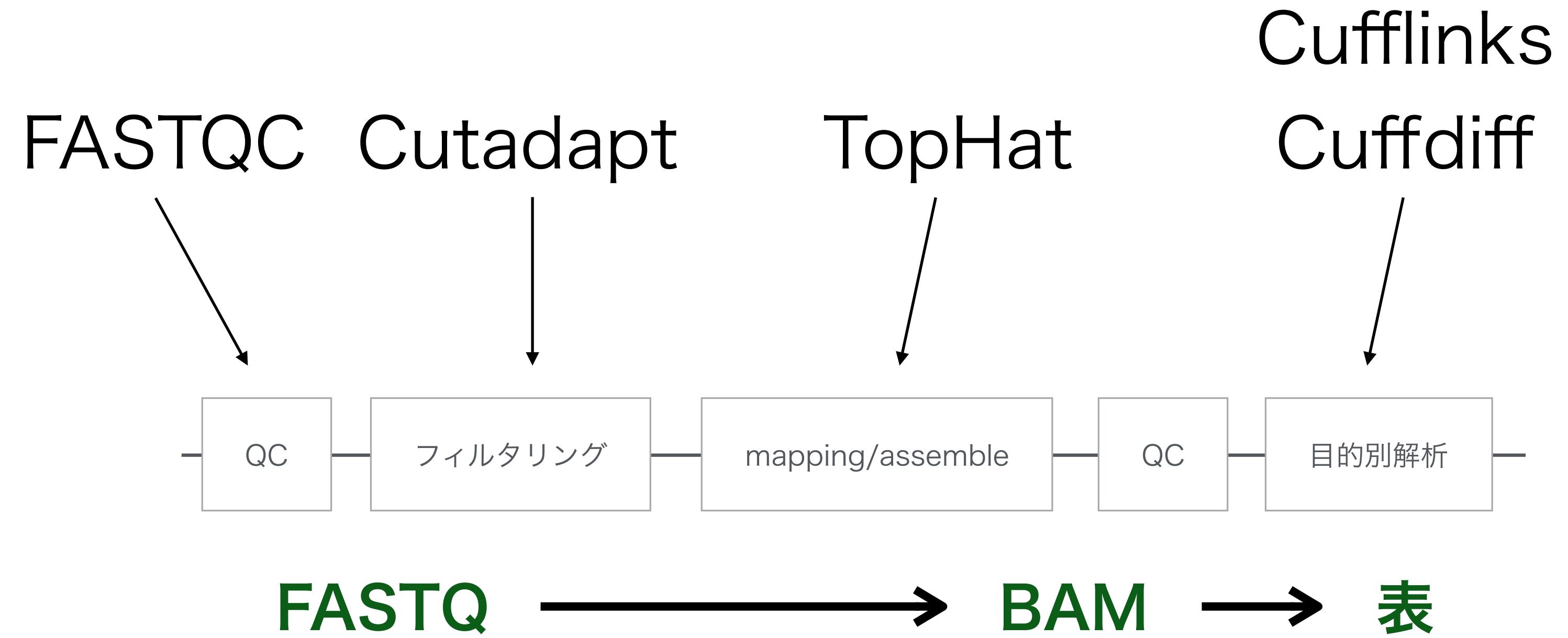
ゲノム/トランスクリプトームのインデックスの構築

マッピング

マッピング結果のクオリティチェック (QC)

発現量定量 (RNA-Seq) / ピーク検出 (ChIP-Seq)

# RNA-Seqデータの低次解析（一例）

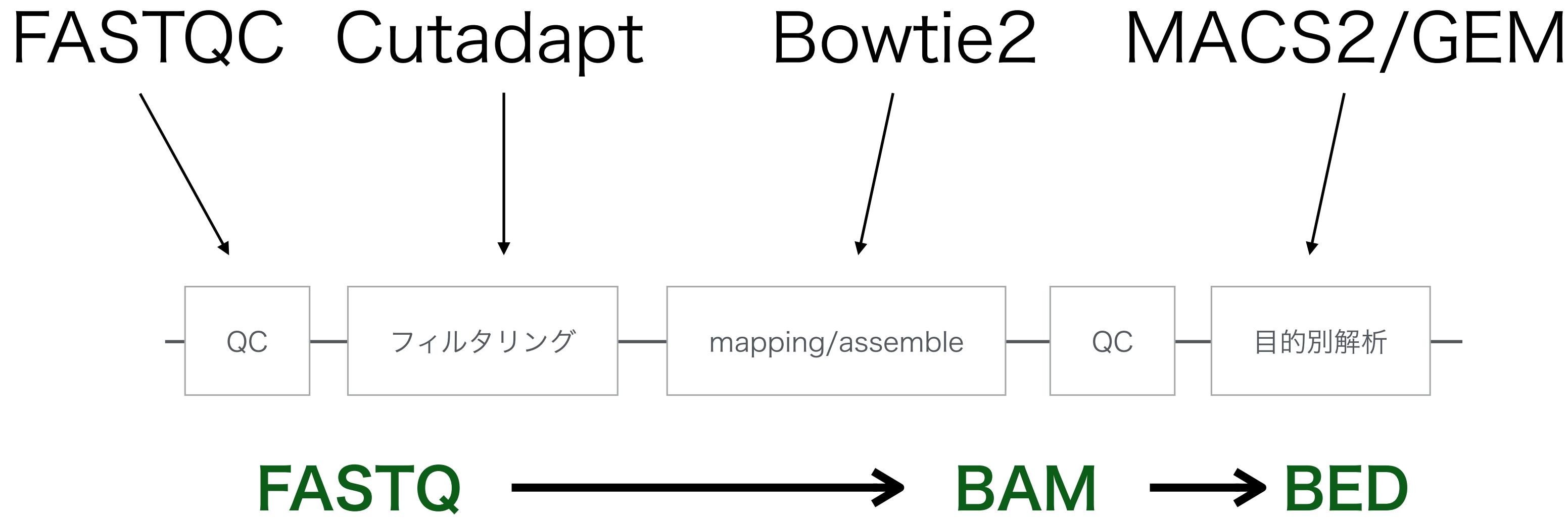


\*これがベストプラクティスとは限らない

研究目的とデータに依る

ソフトウェアの発展と共に変わる

# ChIP-Seqデータの低次解析（一例）



\*これがベストプラクティスとは限らない

研究目的とデータに依る

ソフトウェアの発展と共に変わる

低次解析済みのデータを取得する

# 低次解析済みのデータを取得する

## 自分のデータ

自分で低次解析をやる (CUI or GUI)

共同研究者・テクニシャンに頼む

企業に頼む

Galaxy、BaseSpaceなどウェブサービスで低次解析をする

## 公開データ

生データをダウンロードしたあとは上記と同じ方法で解析する

解析済みデータをダウンロードする

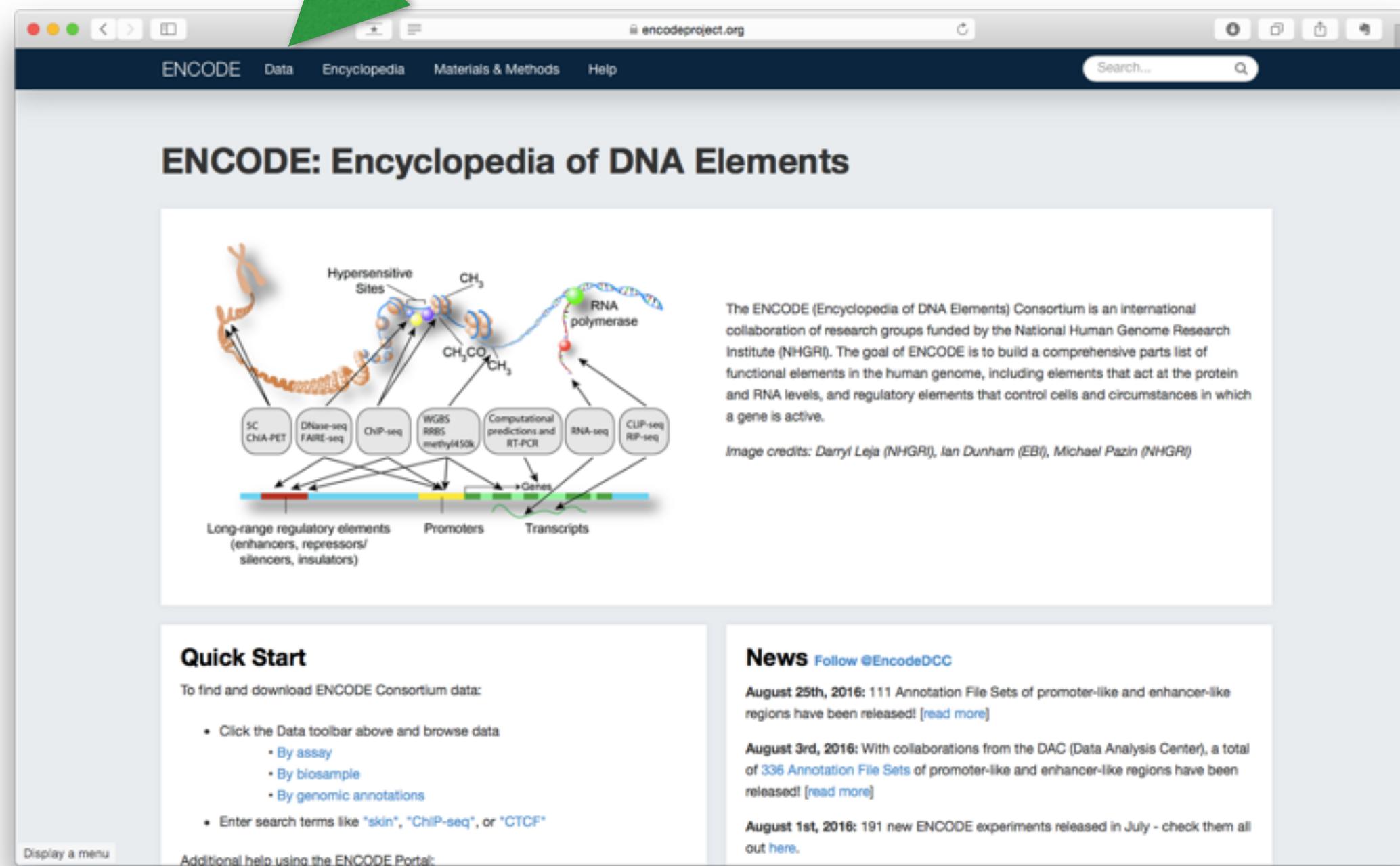
# **"Data Use Policy for External Users" of the ENCODE Project**

**External data users may freely download, analyze and publish results based on any ENCODE data without restrictions as soon as they are released.** This applies to all datasets, regardless of type or size, and includes no grace period for ENCODE data producers, either as individual members or as part of the Consortium. Researchers using unpublished ENCODE data are encouraged to contact the data producers to discuss possible coordinated publications; however, this is optional. The Consortium will continue to publish the results of its own analysis efforts in independent publications.

# ENCODEプロジェクトのChIP-SeqデータをBED形式でダウンロードしてみる

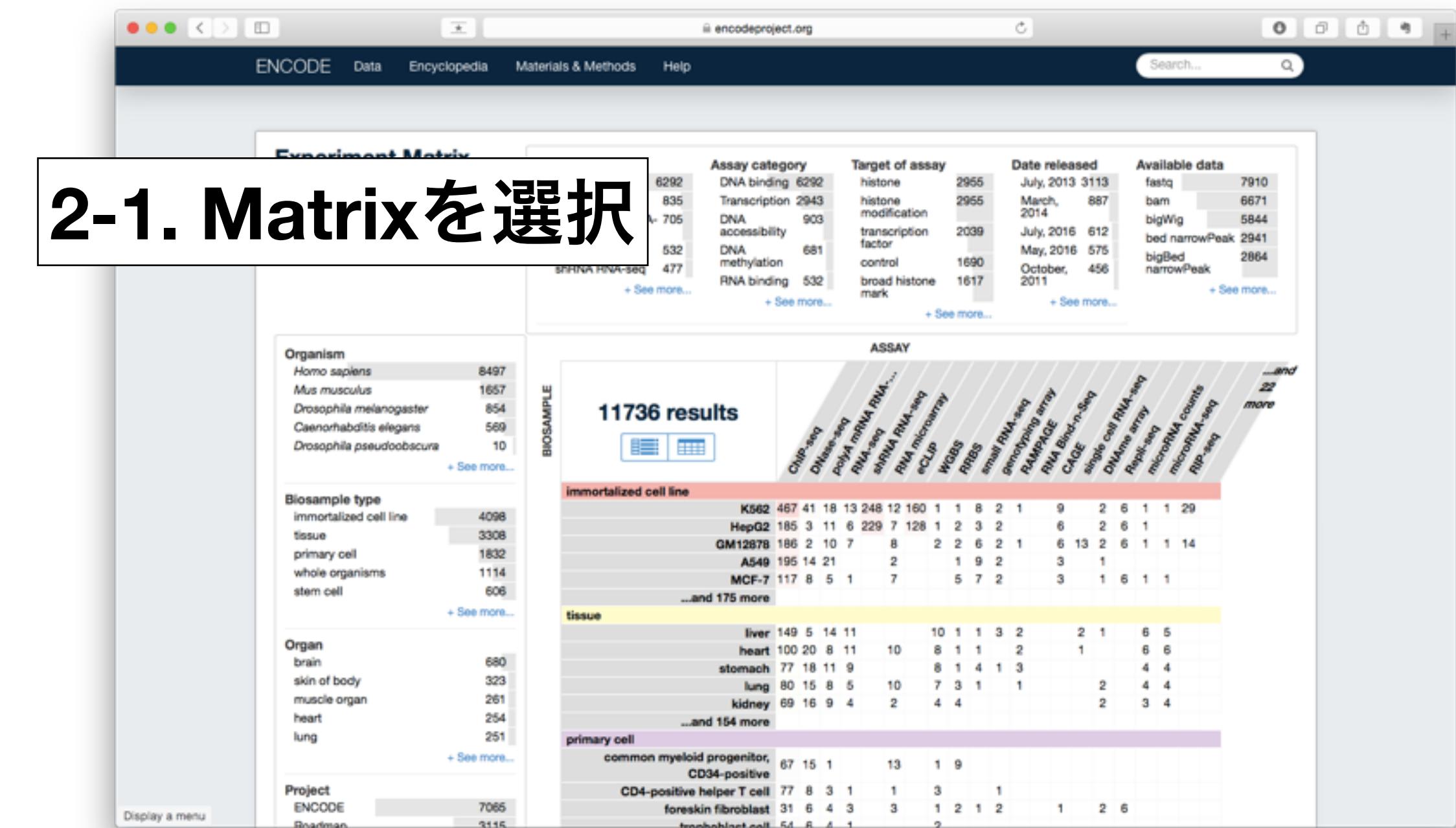
<https://www.encodeproject.org>

## 1. Data をクリック



The screenshot shows the ENCODE homepage with a green arrow pointing to the 'Data' link in the top menu. The main content area displays a diagram of the ENCODE project's workflow, showing various sequencing and computational methods like ChIP-seq, DNase-seq, and RNA-seq, all centered around genes and transcripts.

## 2-1. Matrixを選択



The screenshot shows the ENCODE Experiment Matrix page. A green box highlights the 'Experiment Matrix' section. The matrix table displays experimental data for 11736 results, categorized by assay (e.g., ChIP-seq, DNase-seq, RNA-seq) and biosample (e.g., immortalized cell line, primary cell, tissue). The table includes columns for Assay category, Target of assay, Date released, and Available data.

Assay category	Target of assay	Date released	Available data
DNA binding	histone	July, 2013	fastq
Transcription	histone	March, 2014	bam
DNA accessibility	modification	2014	bigWig
methylation	transcription factor	July, 2016	bed narrowPeak
control	control	May, 2016	bigBed
broad histone mark	broad histone mark	October, 2011	narrowPeak

# ENCODEプロジェクトのChIP-SeqデータをBED形式でダウンロードしてみる

## 3. Searchを選択

The screenshot shows a web browser displaying the ENCODE project website at encodeproject.org. The main content area is titled "Showing 25 of 11736 results" and lists several ChIP-seq experiments. Each experiment entry includes the target protein, lab, and project information. On the left side of the page, there is a sidebar with filters for "Assay category", "Assay", "Project", "RFA", "Experiment status", and "Genome assembly (visualization)". The "Assay category" filter is currently set to "ChIP-seq". The "Assay" filter is also set to "ChIP-seq". The "Experiment status" filter is set to "released". The "Genome assembly (visualization)" filter is set to "hg19". At the top of the page, there is a navigation bar with links for ENCODE, Data, Encyclopedia, Materials & Methods, and Help. A search bar is located at the top right.

## フィルターでデータを絞り込む

1. Assay categoryからDNA binding を選択
2. Available dataからBED narrowPeakを選択
3. OrganismからHomo sapiensを選択
4. Biosample type からstem cellを選択
5. Target of assayからTranscription factorを選択

# ENCODEプロジェクトのChIP-SeqデータをBED形式でダウンロードしてみる

4. クリック

The screenshot shows the ENCODE ChIP-seq experiment summary for H1-hESC. At the top, it displays the experiment title "ChIP-seq of H1-hESC", species "Homo sapiens H1-hESC stem cell", target gene "SUZ12", lab "Peggy Farnham, USC", and project "ENCODE". Below this is the "Experiment summary for ENCSR000EUR" section, which includes a table of experimental details like assay type (ChIP-seq), target gene (SUZ12), and biosample summary (Homo sapiens H1-hESC stem cell). The table also lists attribution information, including the lab (Peggy Farnham, USC), award PI (Michael Snyder, Stanford), project (ENCODE), and external resources (UCSC-ENCODE-hg19:wgEncodeEH001752, GEO:GSM935352). The experiment was released on 2011-10-29.

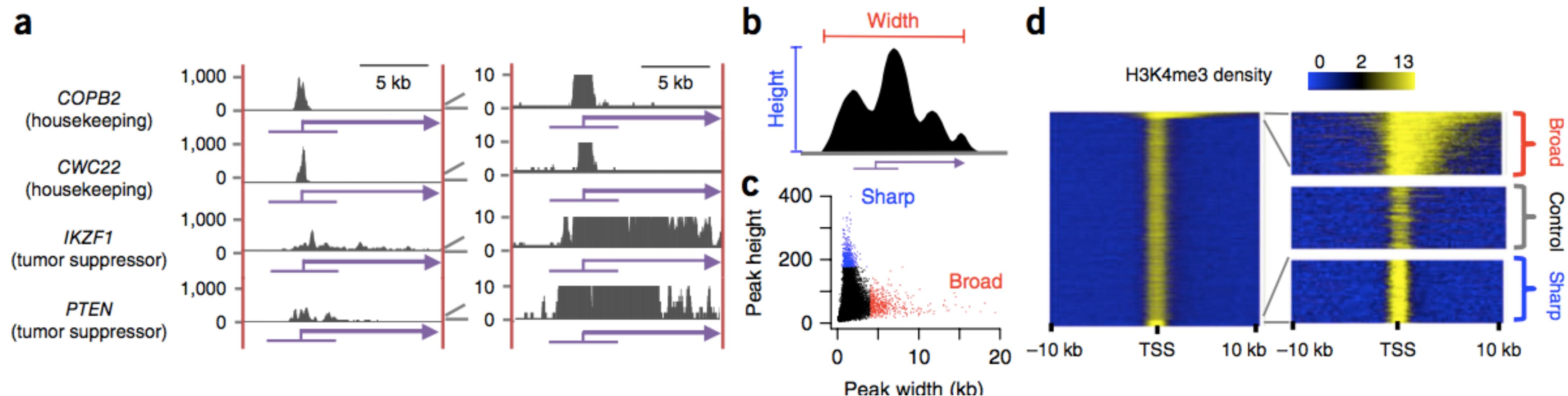
下の方にスクロールしてBEDを  
ダウンロードできる  
bed narrowPeak, optimal id thresholded  
peaks をクリック

5. ダウンロード

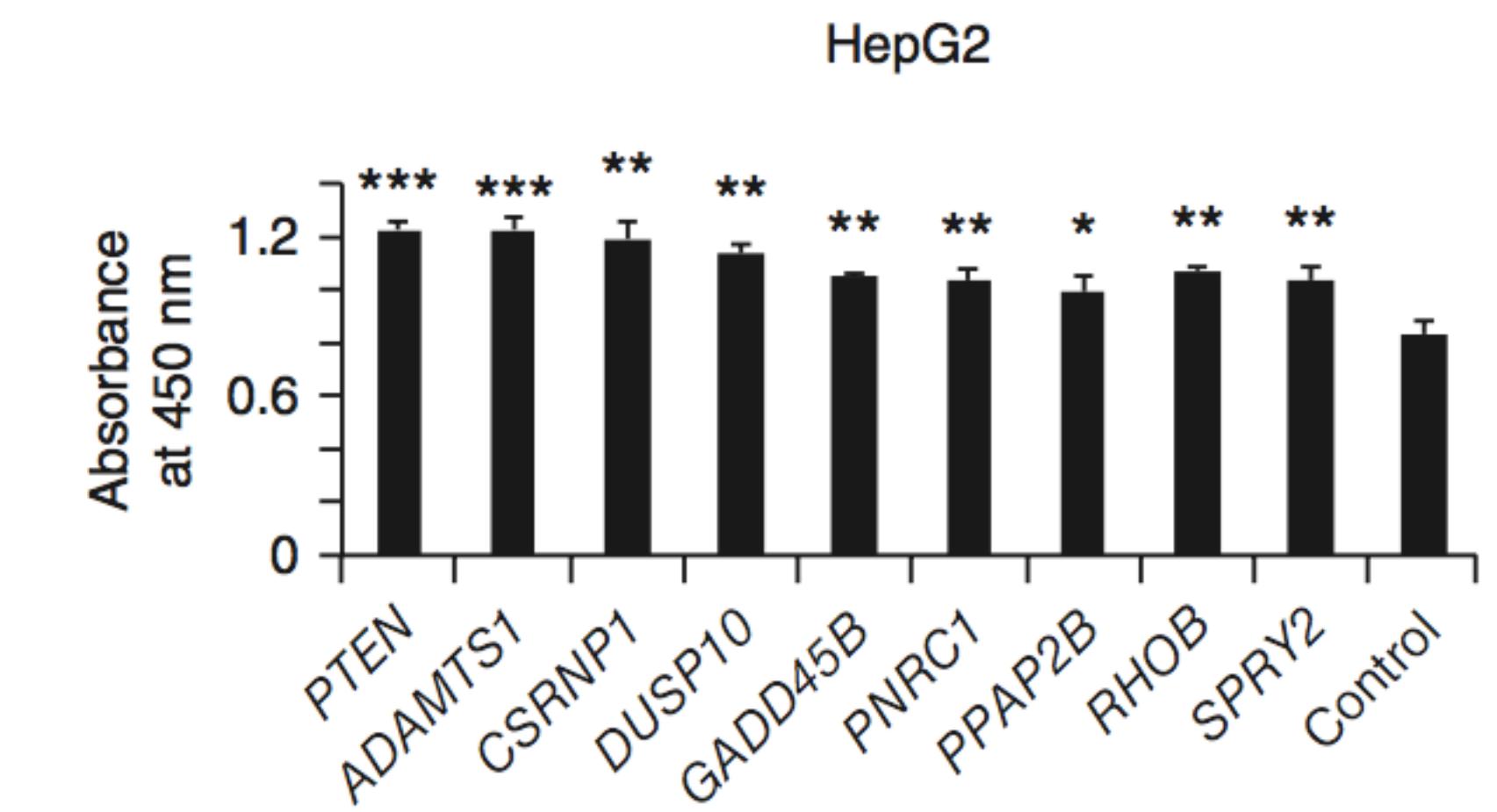
This screenshot shows the ENCODE ChIP-seq data download page for the same experiment. It features a large green arrow pointing from the "5. ダウンロード" text in the previous slide down to the download links. Below the arrow, there are two download options: "ENCFF002CRG" (with a blue download icon) and "bed narrowPeak". To the right of these links, the genomic coordinate system "hg19" is specified. Further to the right, the file details are listed: "ENCODE Consortium Analysis Working Group", "2014-06-06", "118 kB", a green checkmark indicating it's released, and a green "released" button.

# NGSデータ高次解析の実例

# 例: ヒストン修飾領域の長さとがん抑制遺伝子の関係



がん抑制遺伝子ではH3K4me3にカバーされる領域が長く、転写伸長活性と相關  
H3K4me3領域の長い遺伝子をノックダウンにより細胞増殖が亢進



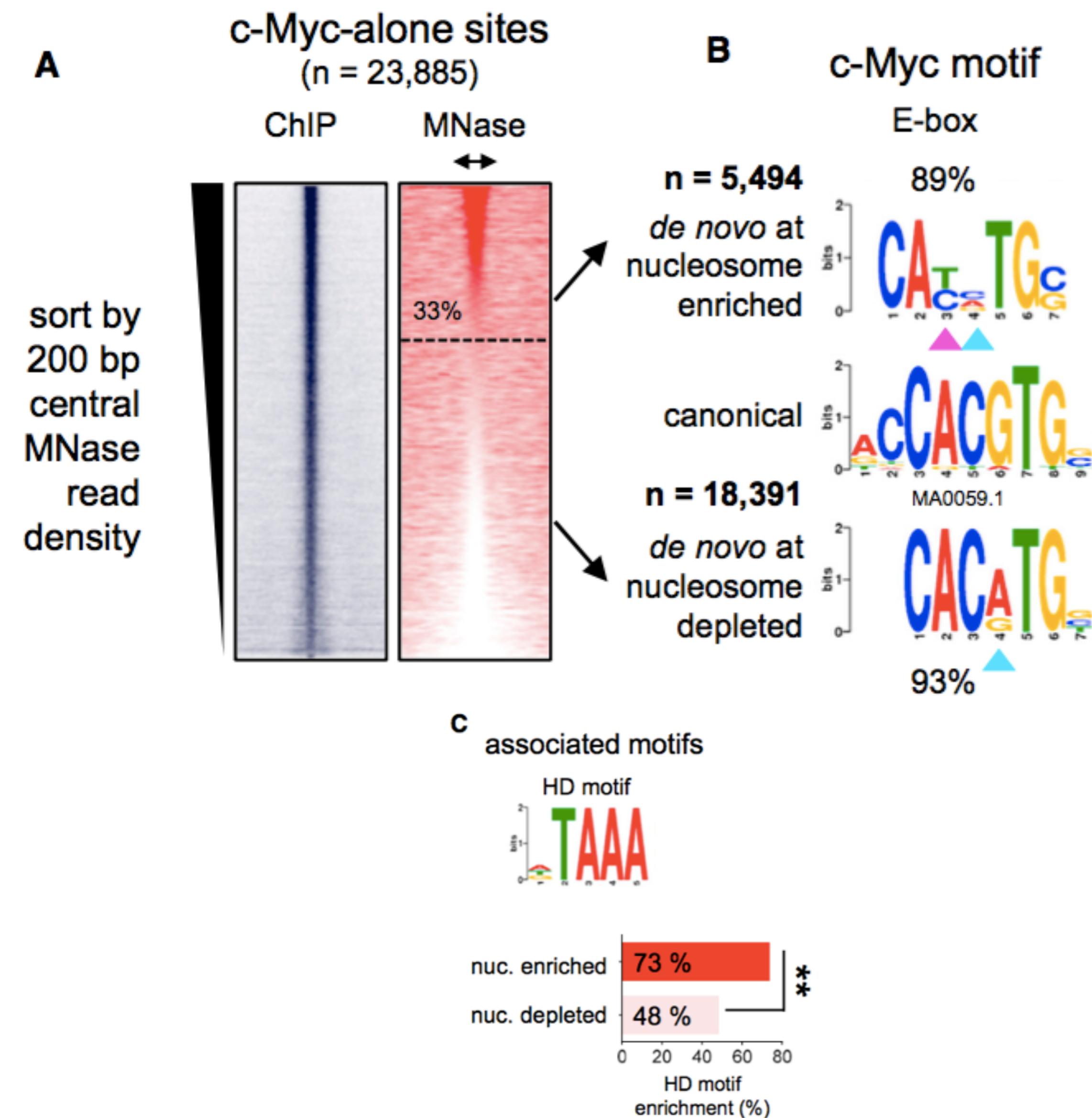
# 例: パイオニア転写因子の結合様式

## パイオニア転写因子

ヌクレオソームに結合してクロマチンを  
緩める転写因子

c-MycのChIP-Seqピークを  
MNase (ヌクレオソームの位置  
を検出) に重ね、分類

既知モチーフがdegenerateしていた  
Homeodomainが濃縮 → cofactor

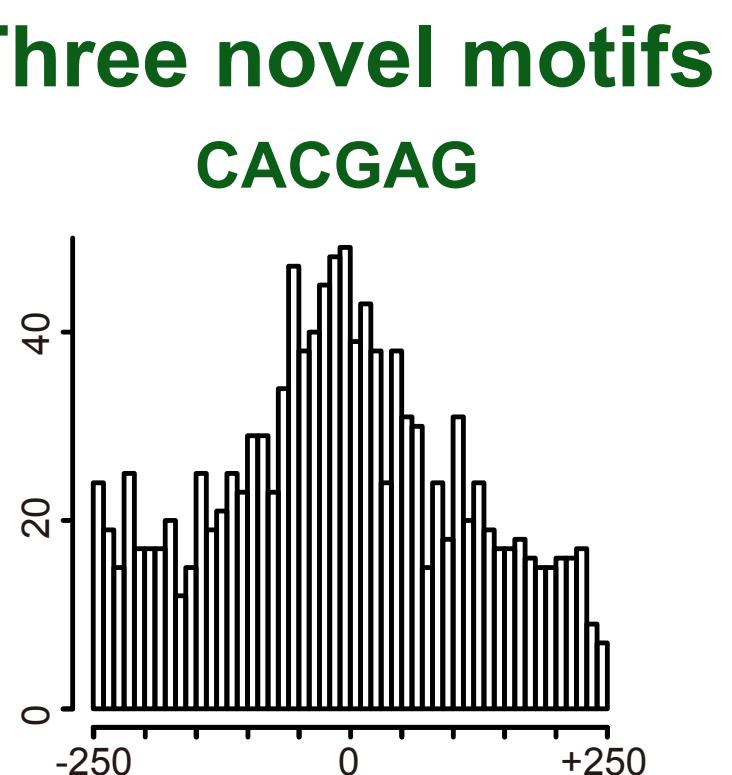
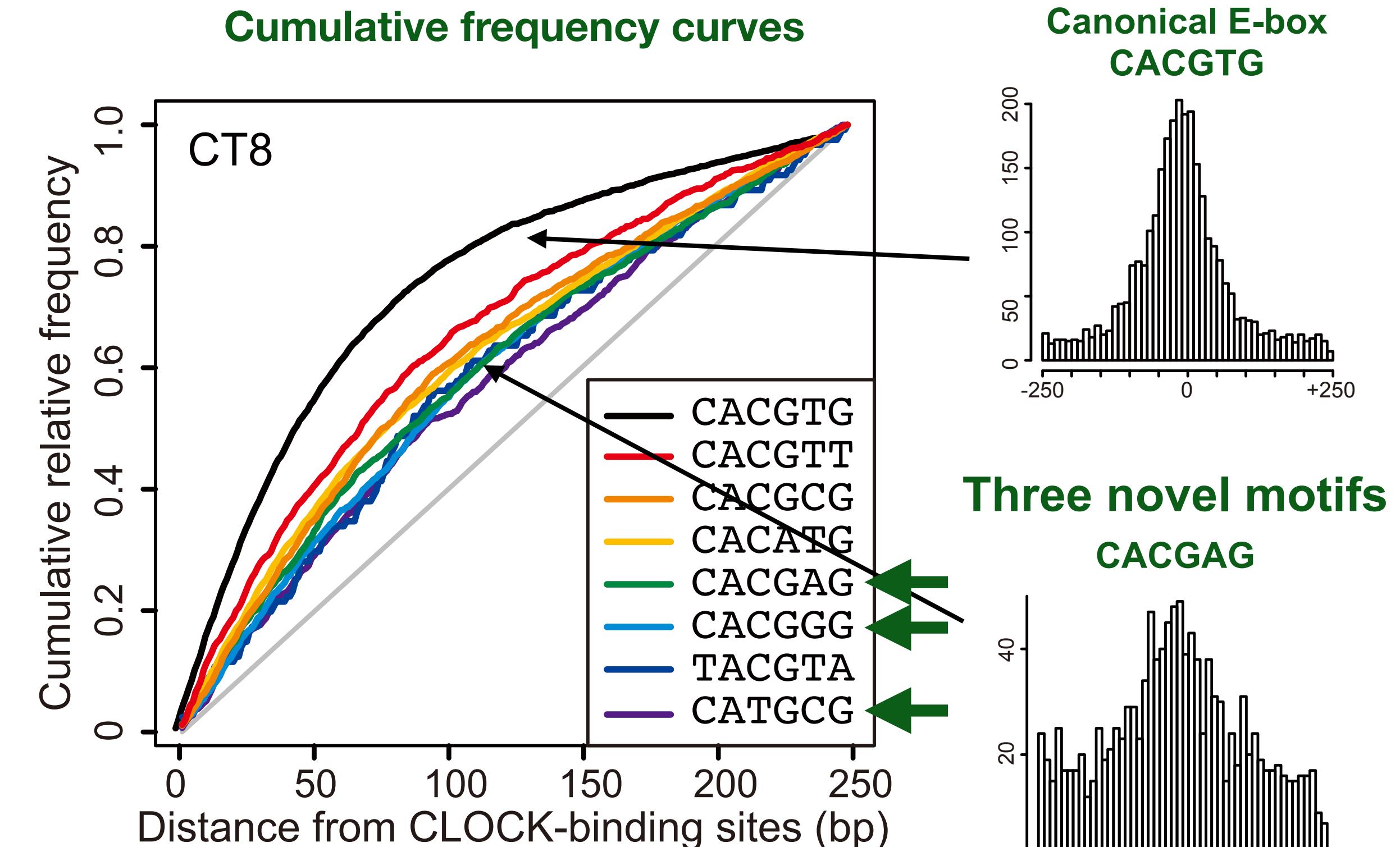
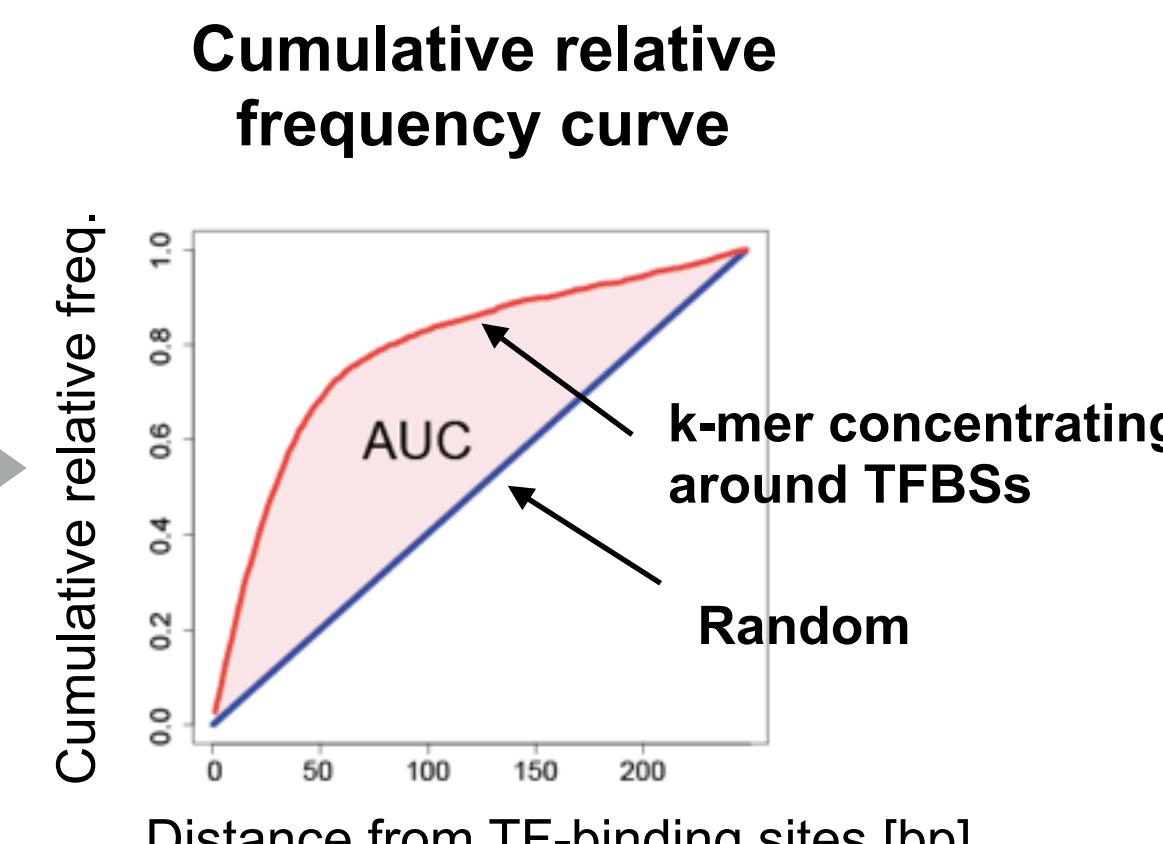
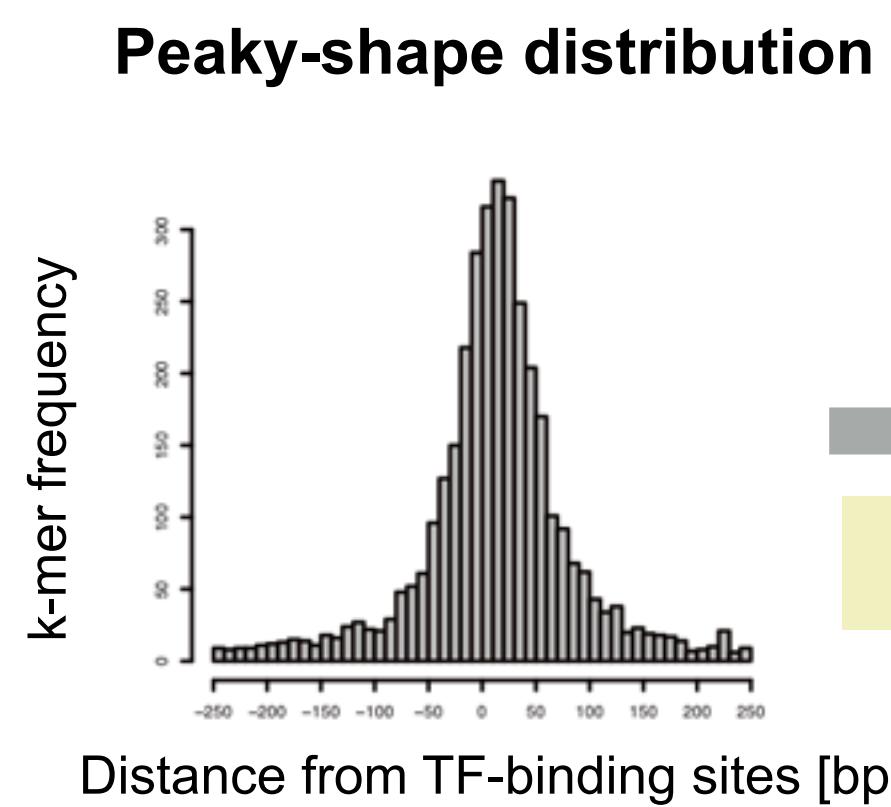


# 例: MOCCS (Motif Centrality Analysis of ChIP-Seq)

CLOCK ChIP-Seqデータから

新規モチーフを発見

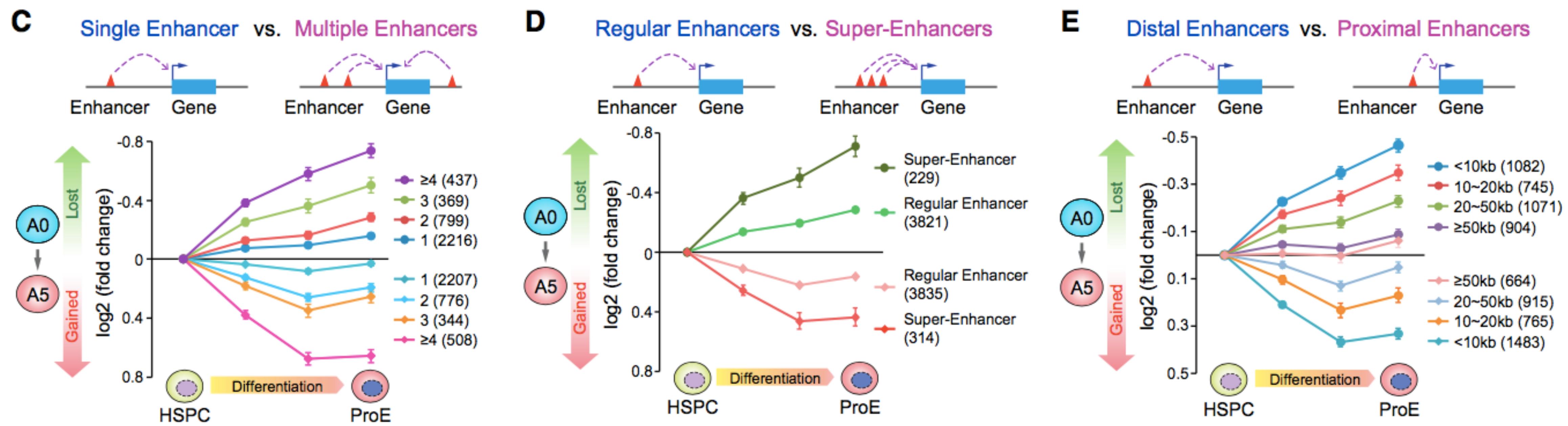
<https://github.com/yuifu/moccs>



# 例: エンハンサーの性質と遺伝子発現の関係

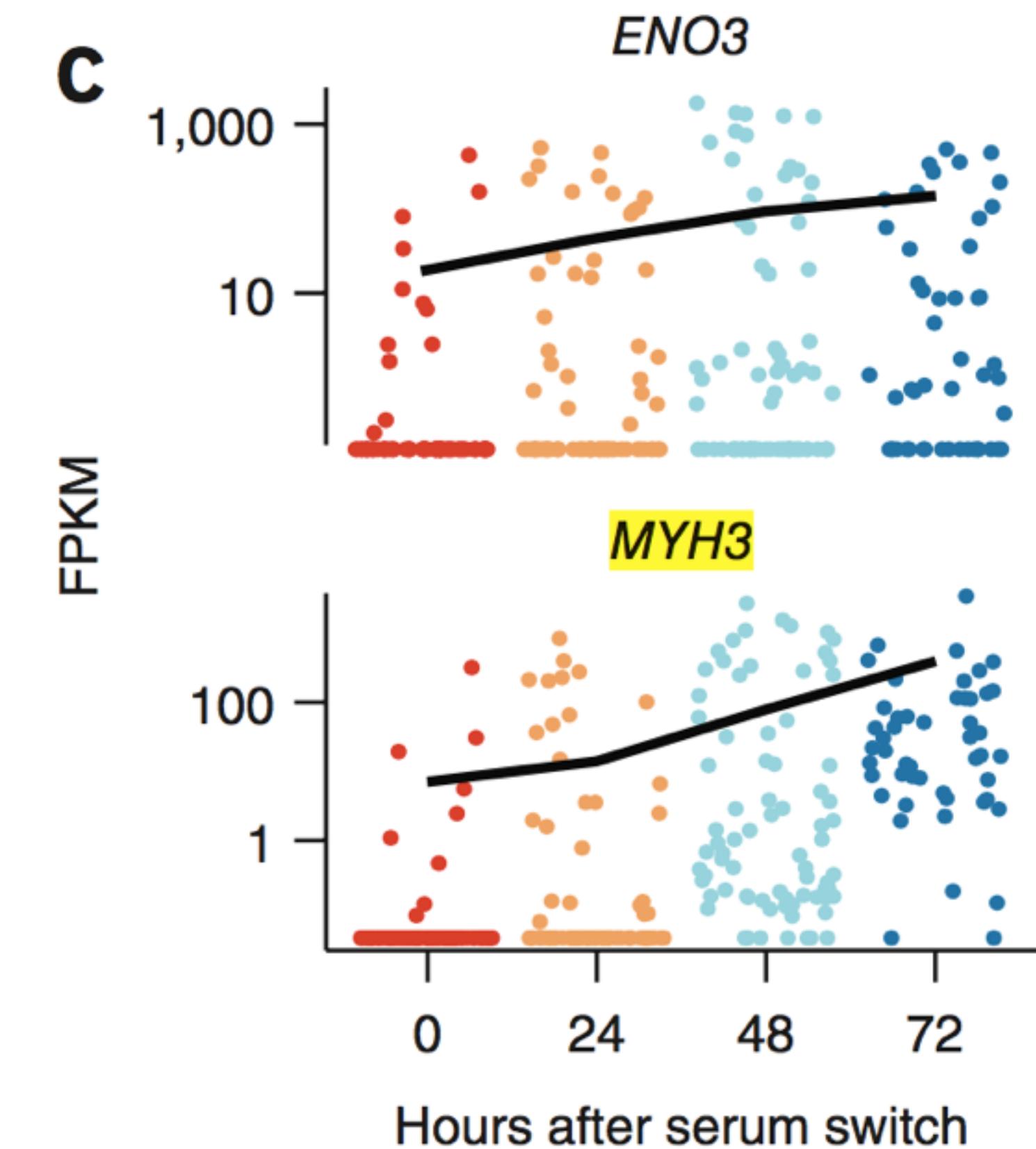
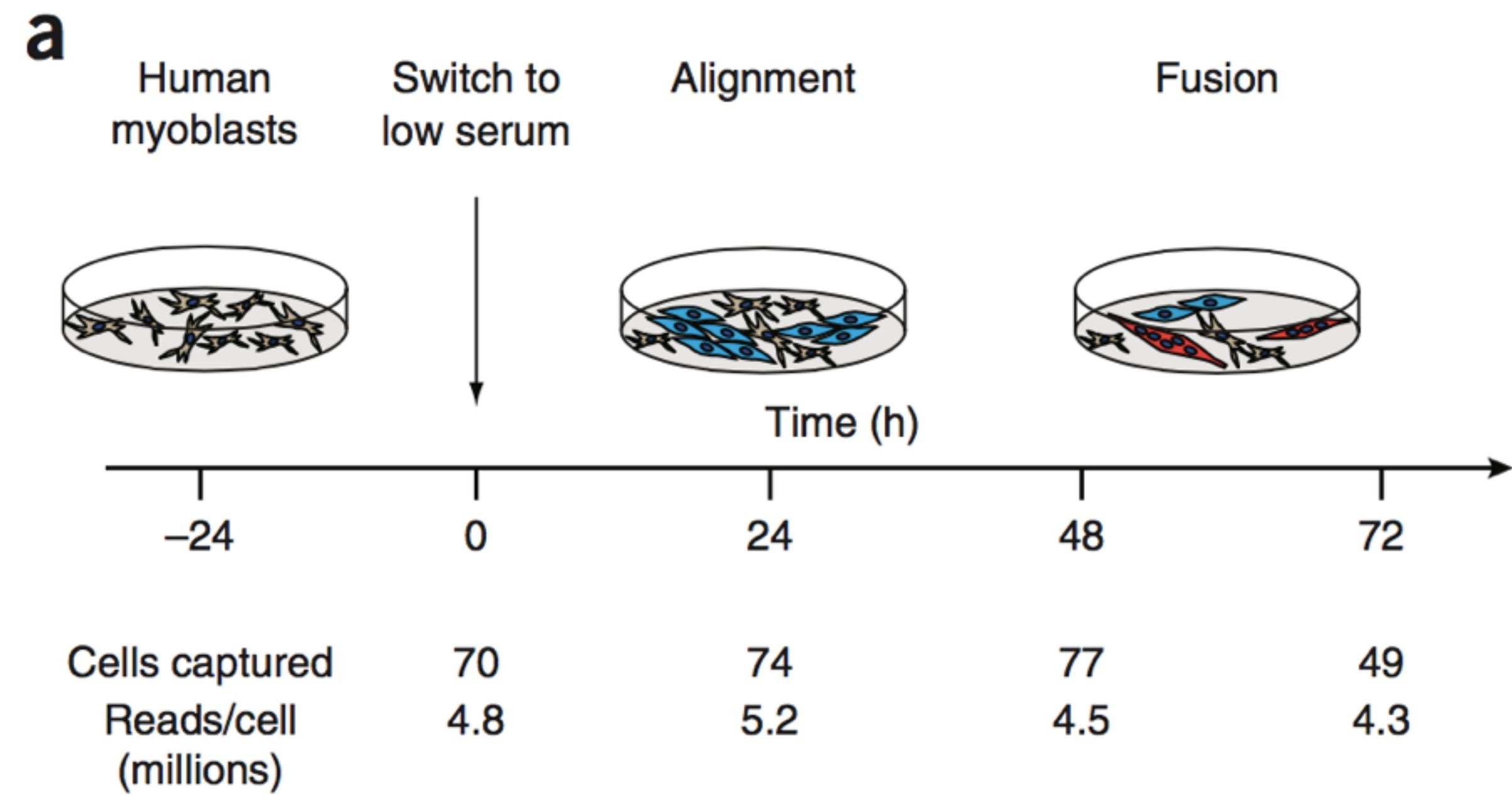
エンハンサーの性質の違いと細胞分化過程での発現変動の大きさの関係を示す

H3K4me1とH3K27acでエンハンサーを定義



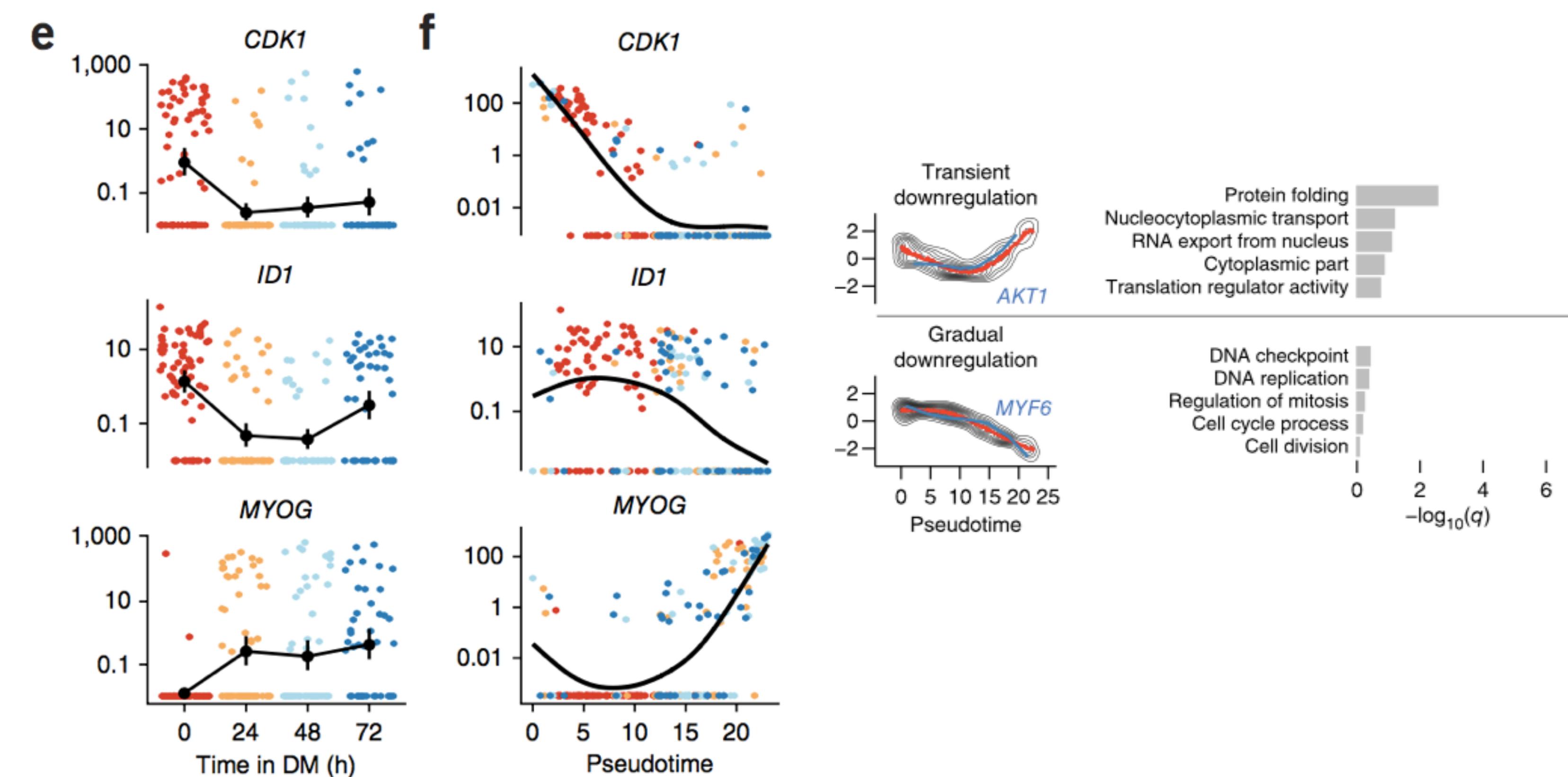
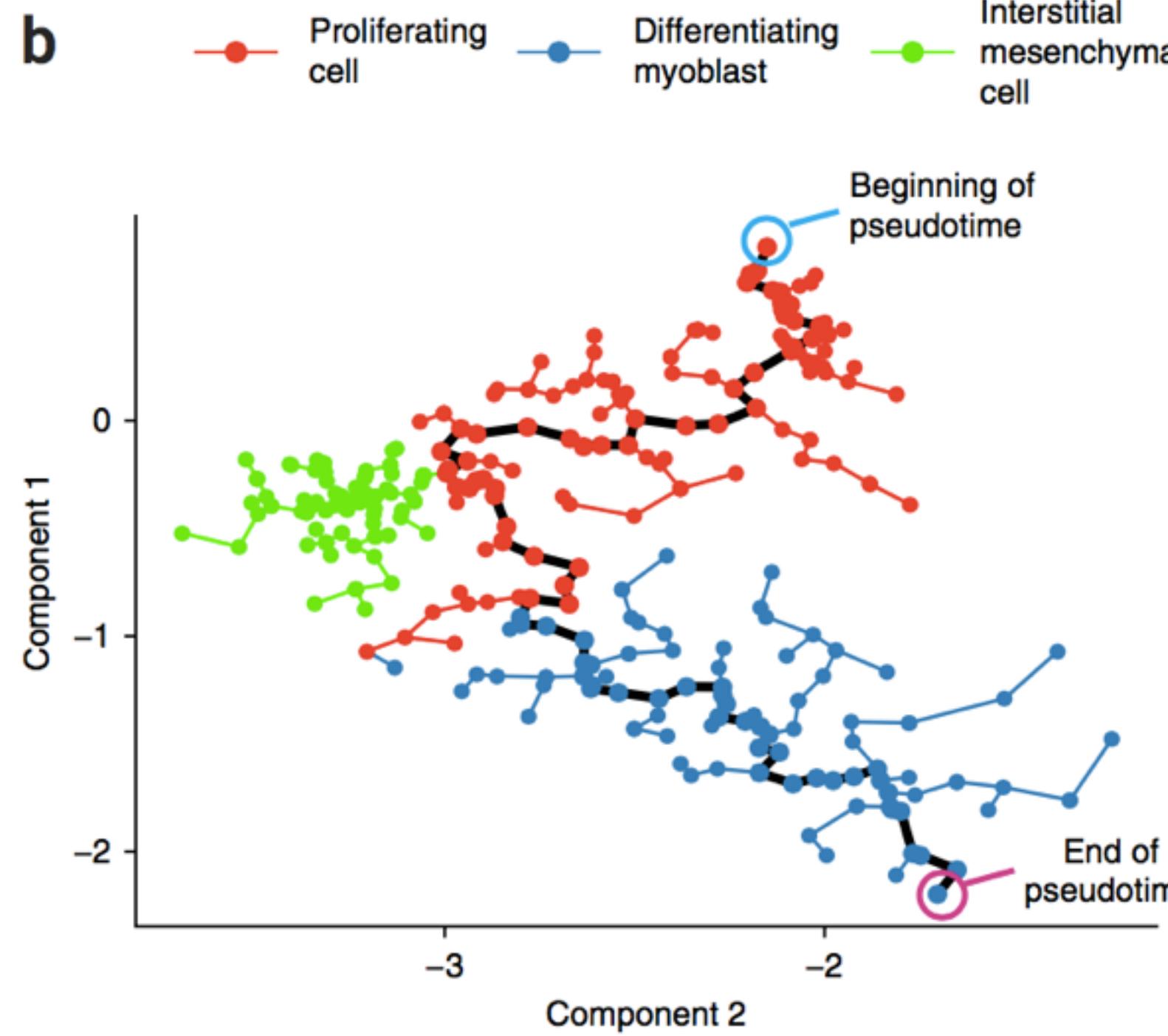
# 例: 擬時間推定 Pseudotime estimation

非同期的な細胞分化



# 例: 擬時間推定 Pseudotime estimation

細胞の"主観的"時間に応じて変動する遺伝子群



# NGSデータ解析の基礎

## ファイルフォーマット

# ファイルフォーマット File format

データを記述するためのルール（仕様 Specification）

目的に応じてさまざまなフォーマットがある

アクセスしてみよう <https://genome.ucsc.edu/FAQ/FAQformat.html>

[https://en.wikipedia.org/wiki/Biological\\_data](https://en.wikipedia.org/wiki/Biological_data)

ソフトウェアは特定のフォーマットを想定して設計されている  
別のフォーマットだったり、フォーマット通りに書かれていないファイルを入れるとエラーが出る

# ファイルフォーマットのまとめ

塩基配列

FASTA、FASTQ

アラインメント

BAM/SAM、CRAM

区間

BED、GTF

ゲノム座標上の数値データ

Wig、BigWig

0-basedと1-based

# 塩基配列

A, T, G, Cから成る文字列

GCATATAATGCAAATTACTTGGAACTT...

# ゲノム配列

A, T, G, Cから成る文字列（の集合）

個々の文字列に染色体名がある

座標で位置を表す

染色体名  
↓  
chr1    GCATATAATGCAAATTACTTGGAACTT...

            1                10                20

座標  
↑

chr2    TAGAAGTACCAAGAACGTCCAAATCTCAG...

            1                10                20

chr3    GTGTCTTCCTTTGTATTGTGGAACCGGA...

            1                10                20

⋮

# FASTAフォーマット

一般的な塩基配列の情報

[https://en.wikipedia.org/wiki/FASTA\\_format](https://en.wikipedia.org/wiki/FASTA_format)

Label

Sequence

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKSQI
```

# FASTQフォーマット

NGSリードの配列情報とクオリティスコア

<http://ja.wikipedia.org/wiki/Fastq>

[https://en.wikipedia.org/wiki/Phred\\_quality\\_score](https://en.wikipedia.org/wiki/Phred_quality_score)



# Phred quality score

塩基読み取りがエラーである確率  $p$  を変換した値（大きいほど信頼度が高い）

よく使われるのはSangerの式

$$Q_{\text{sanger}} = -10 \log_{10} p$$

ASCIIコードによるエンコーディング

Sanger形式では0から93の値をASCIIコードでの33から126の間の文字として表現

ASCII codeについて <https://en.wikipedia.org/wiki/ASCII>

より詳しくは Nucl. Acids Res. (2010) 38 (6): 1767-1771を参照

<http://nar.oxfordjournals.org/content/38/6/1767.full>

# アラインメント

## 塩基配列同士の対応付け

NGS解析では、リードをゲノム配列やコンティグにアラインメント（リードの塩基配列と各塩基とゲノムの各塩基を対応づける）することをマッピングともいう  
ゲノムをリファレンス（Reference）、リードをクエリ（Query）と呼ぶこともある  
対応のパターン：マッチ、ミスマッチ、挿入、欠失など

	ミスマッチ	欠失	挿入	
	AGCACCA	GTCCAA-TC	AGGTGCC	リード
chr2	TAGAAGTACCAAGAATGTCCAAATCTCAGAGGT-CCCCAATG...			ゲノム配列
	1	10	20	

# BAM/SAMフォーマット

## リードのゲノム等に対するマッピング情報

リードとゲノムの両方に関する情報が書いてある

詳しくは <http://samtools.github.io/hts-specs/SAMv1.pdf>

```
QHD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAACGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

1. QNAME (リードのラベル)
2. FLAG
3. RNAME (染色体名)
4. POS (マッピング開始位置)
5. MAPQ (Mapping quality)
6. CIGAR (塩基の対応)
7. RNEXT (mate readの染色体名)
8. PNEXT (mate readのマッピング開始位置)
9. TLEN (符号付リード長)
10. SEQ (リードの塩基配列)
11. QUAL (リードのquality score)

# ゲノム上の区間

区間を座標で定義できる

染色体名（コンティグ名）

始点の座標（Start）

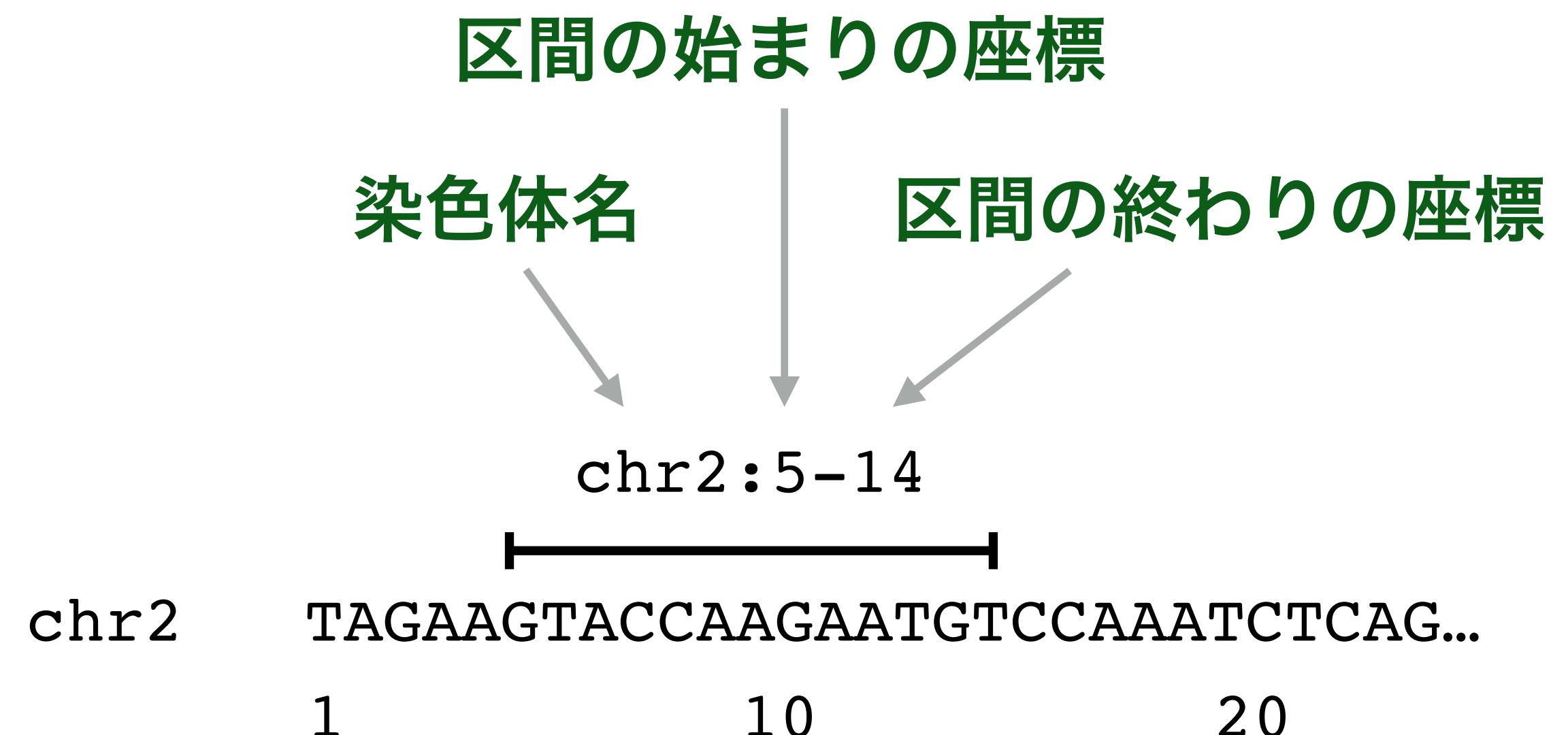
終点の座標（End）

プラス鎖かマイナス鎖か

Strand information

通常 +/- か 1/-1 で表現される

Strand情報がない場合（ChIP-Seqのピークなど）はピリオド（.）で表記することが多い



# 遺伝子

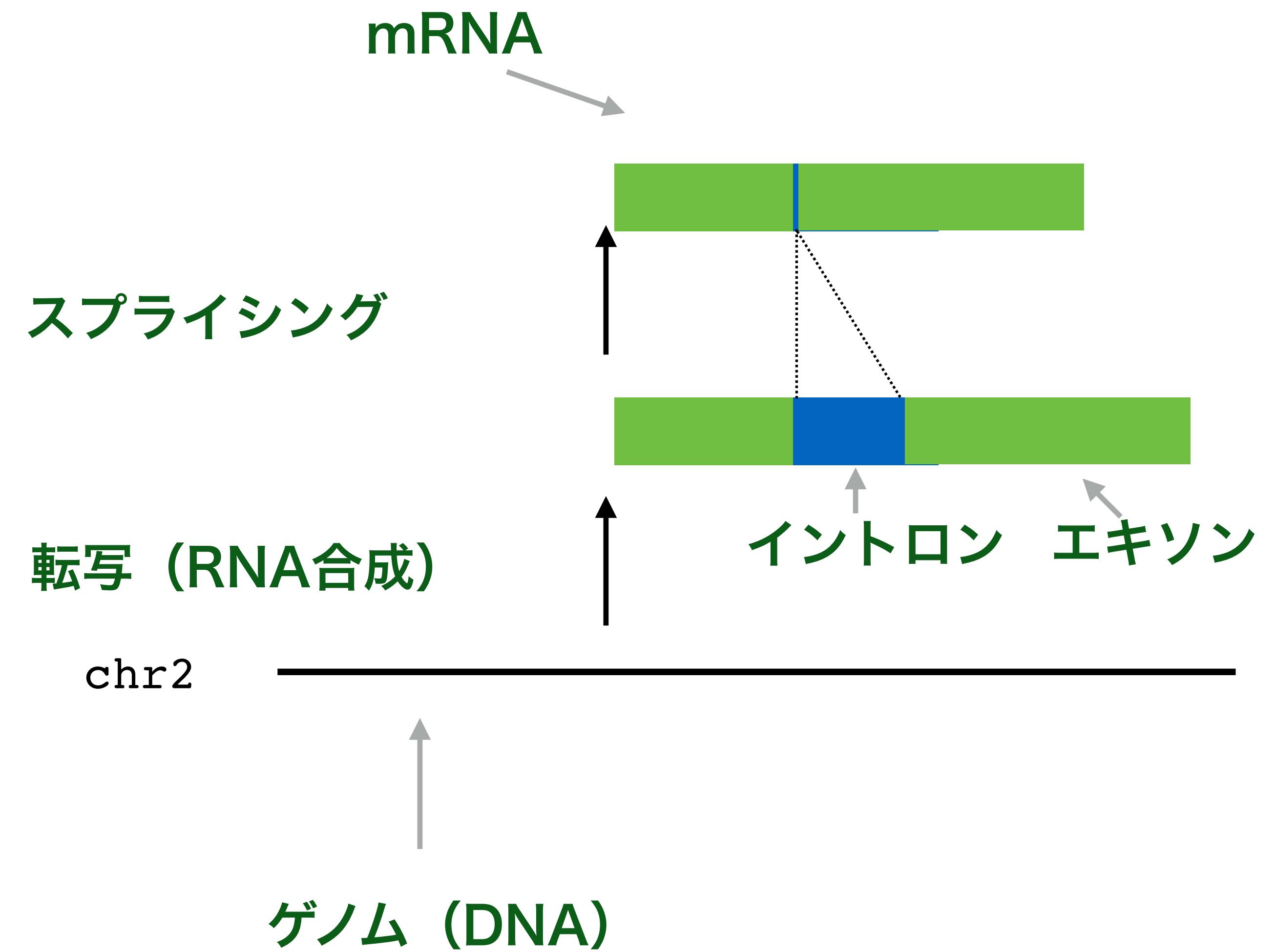
ゲノムの中でたんぱく質の情報を持っている部分

転写される

その部分をテンプレートとして  
RNAが合成される

スプライシングされることがある

※簡単のため、問題のある説明です。また、非コードRNAについて  
は割愛しています。



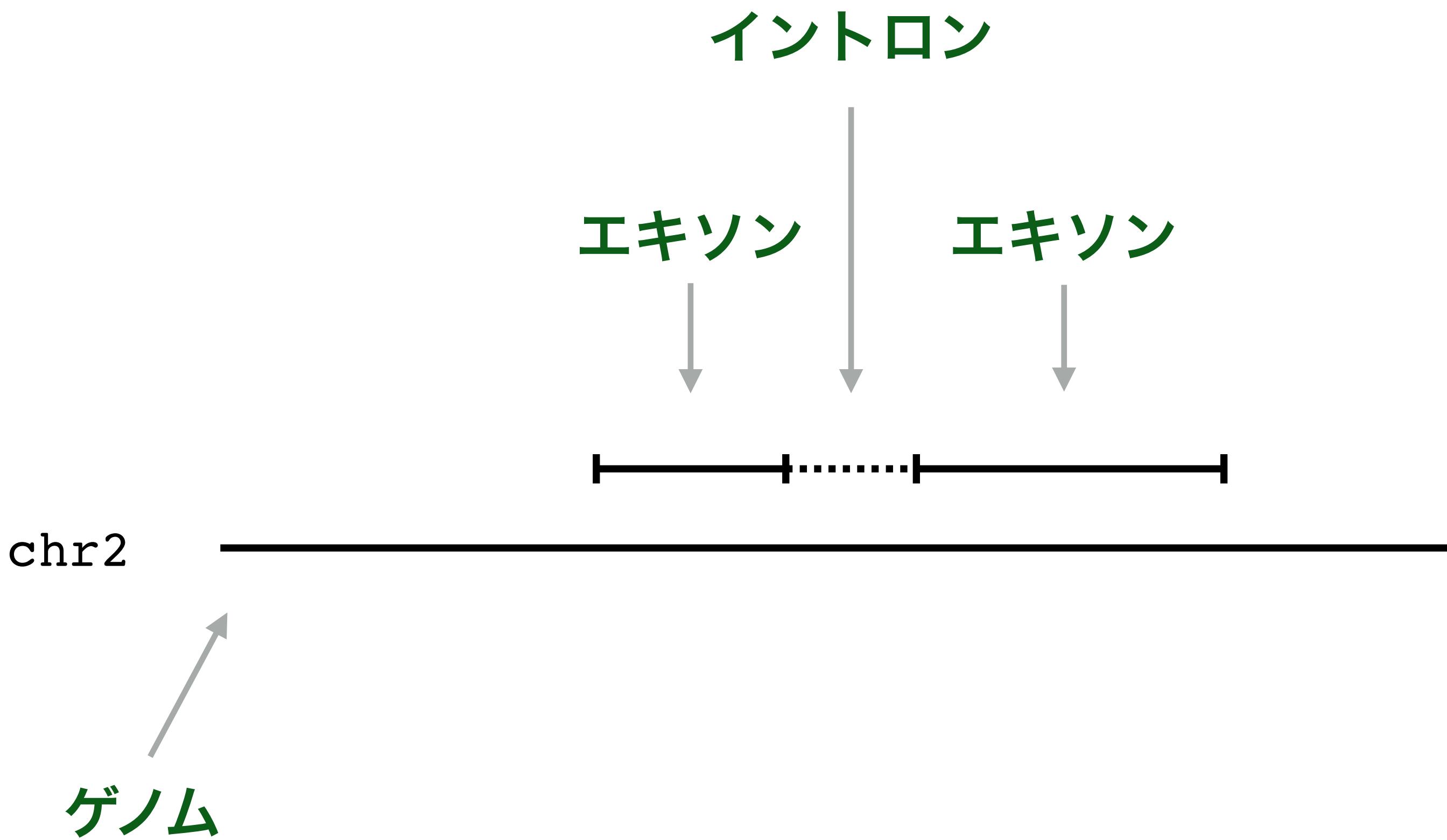
# 区間の集合としての遺伝子

遺伝子 = 区間の集合

一つのエキソンが一つの区間

イントロンは陽に定義しない

エキソンに挟まれた区間



# 様々な生命現象がゲノム上の区間として表現される

DNA結合タンパク質

転写因子などの結合

エピジェネティック修  
飾

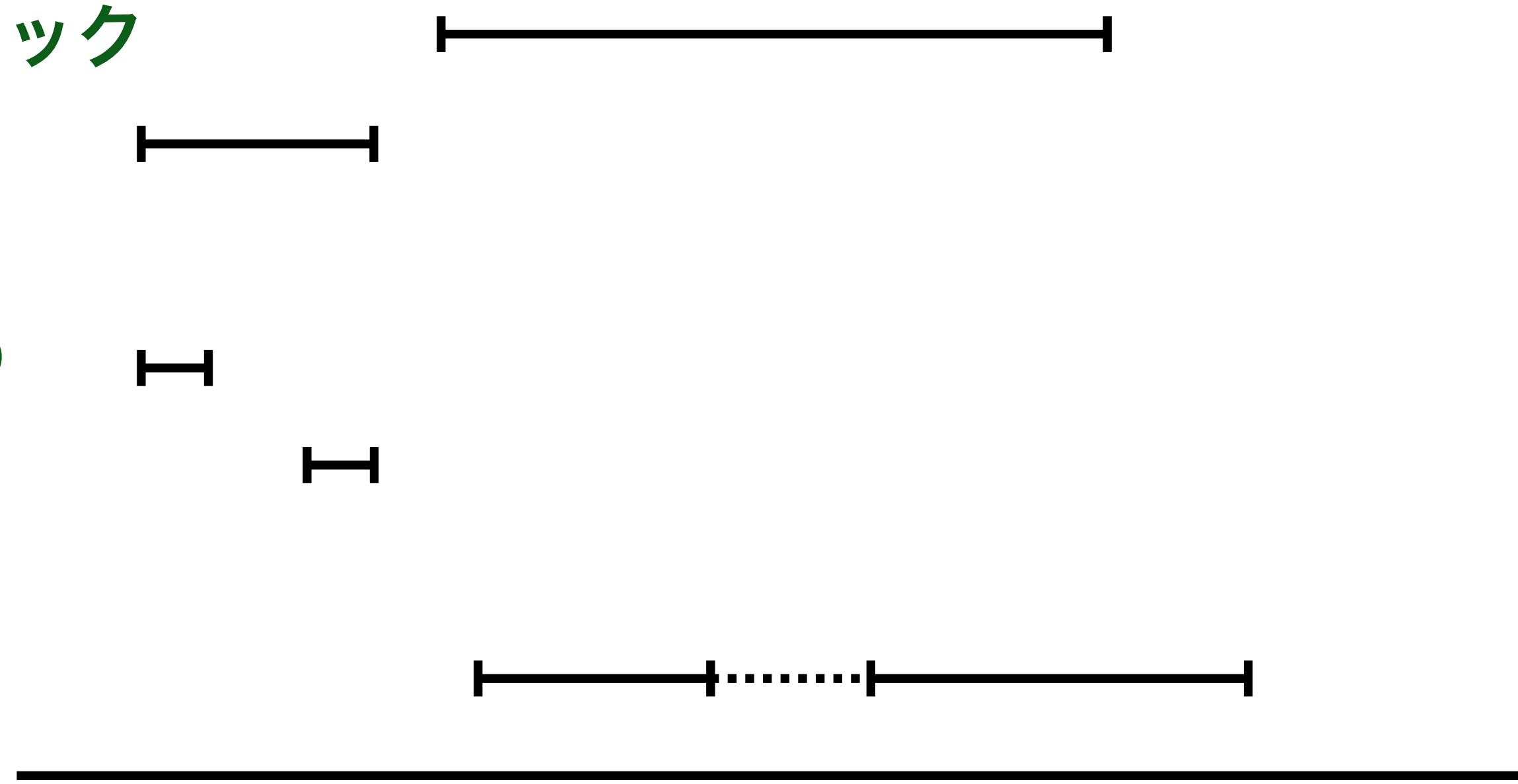
DNAが巻きつくヒストンたん  
ぱく質の化学修飾やCのメチ  
ル化

エピジェネティック  
修飾

転写因子の  
結合

遺伝子  
chr2

ゲノム



# BEDフォーマット

ゲノムやコンティグ上の区間を表す

<https://genome.ucsc.edu/FAQ/FAQformat.html#format1>

より複雑な表現が可能なBED12フォーマットもある

NGSのリードがマップされた位置を表せる

Chromosome	Start	End	Label	Score	Strand
chr7	127473530	127474697	Pos3	0	+
chr7	127474697	127475864	Pos4	0	+
chr7	127475864	127477031	Neg1	0	-

# GTFフォーマット

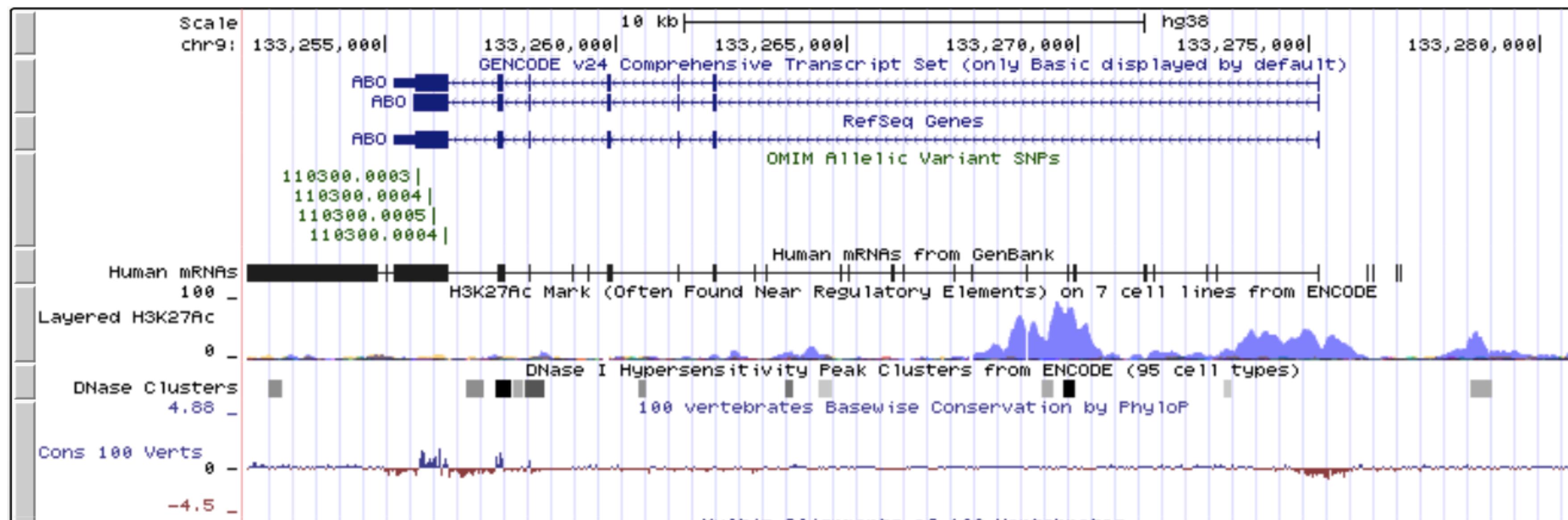
遺伝子の位置（遺伝子アノテーション）を表現する場合など

<https://genome.ucsc.edu/FAQ/FAQformat.html>



# ゲノム座標上の数値データ

例: NGSのリードのカバレッジ、進化的保存性のスコア  
(phastConsなど)、GC%



ゲノム座標

リードカバレッジ

スコア

# Wig / BigWigフォーマット

ゲノム座標上の大規模な数値データ（例: GC percent）

バリエーションがある

<https://genome.ucsc.edu/goldenpath/help/wiggle.html>

<https://genome.ucsc.edu/goldenpath/help/bigWig.html>

BigWigはバイナリ

データの記述方法

Chromosome

座標

```
variableStep chrom=chr2  
300701 12.5  
300702 12.5  
300703 12.5  
300704 12.5  
300705 12.5
```

スコア

# 座標には1-basedと0-basedがある

## 1-based coordinate system

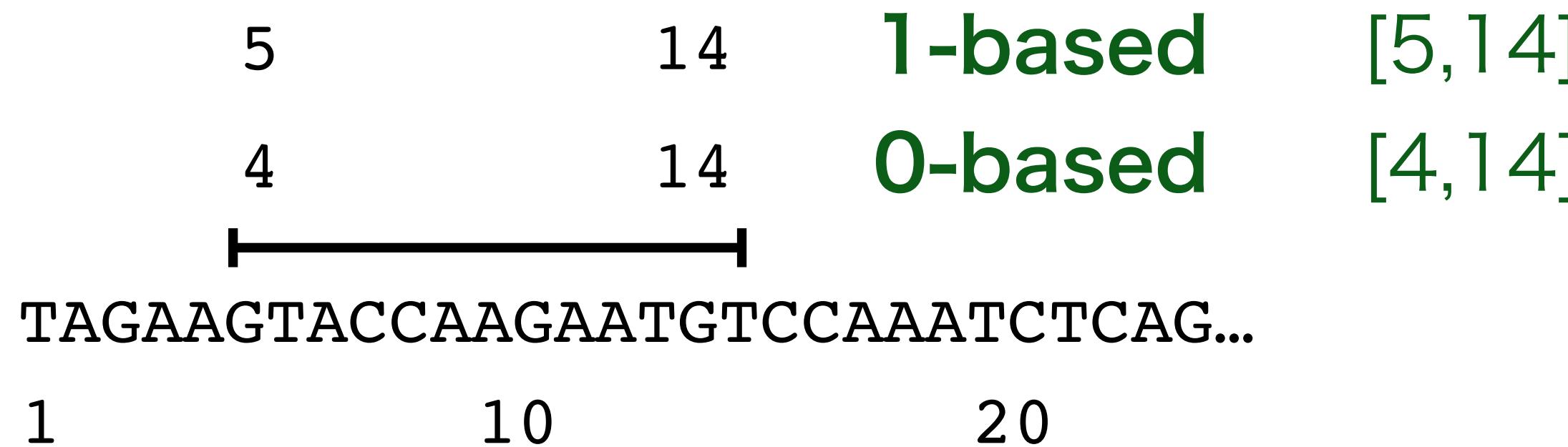
Closed interval

例: SAM, VCF, GFF and Wiggle

## 0-based coordinate system

Half-closed-half-open interval

例: BAM, BCFv2, BED, and PSL



# ファイルフォーマットのまとめ

塩基配列

FASTA、FASTQ

アラインメント

BAM/SAM、CRAM

区間

BED、GTF

カバレッジ

Wig、BigWig

0-basedと1-based

# NGSデータ高次解析の基礎

# NGSデータ高次解析の基礎

よく使われる解析環境・ツール

# **UNIX**

コマンドラインツールを使えるようにする

Linux / Mac OSX

ターミナル

Windows

cygwin

# Linuxに関する知識

## ディレクトリの移動、ファイル操作

cd, pwd, ls, mv, cp, rm, mkdir

## PATHを通す

ターミナルでソフトウェアを動かすときに、どこにそのソフトウェアがあるかを記述しておくこと

`~/.bashrc` または `~/.bash_profile` に記述する

編集した後は `source ~/.bashrc` または `source ~/.bash_profile`

## 詳しくは

<http://biosciencedbc.jp/human/human-resources/workshop/h28-2>

<http://www.lpi.or.jp/linuxtext/text.shtml>

# スクリプト言語

データの前処理などに便利な道具

Perl, Python, Ruby, Juliaなどどれでもいいから一つ  
慣れが必要

全角文字は使わないようにする

大文字と小文字は別物

# R/Bioconducor

R

統計言語

可視化、統計検定、クラスタリングなど  
基本的なデータ解析ができる

<https://cran.r-project.org>



## Bioconductor

バイオインフォマティクス向けのRパッ  
ケージ群

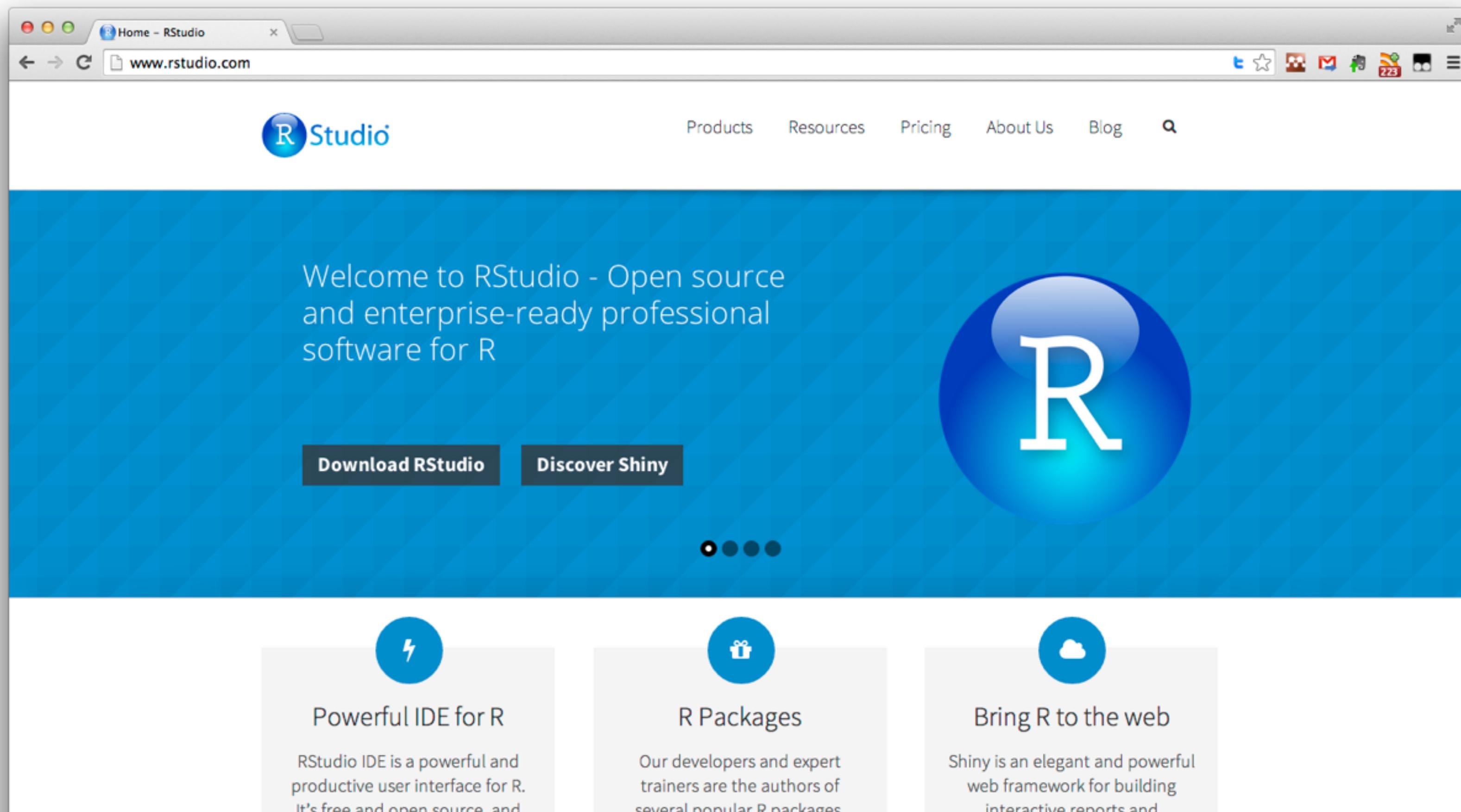
<http://bioconductor.org>



# RStudio

RをGUIで使うための統合解析環境

<http://www.rstudio.com>

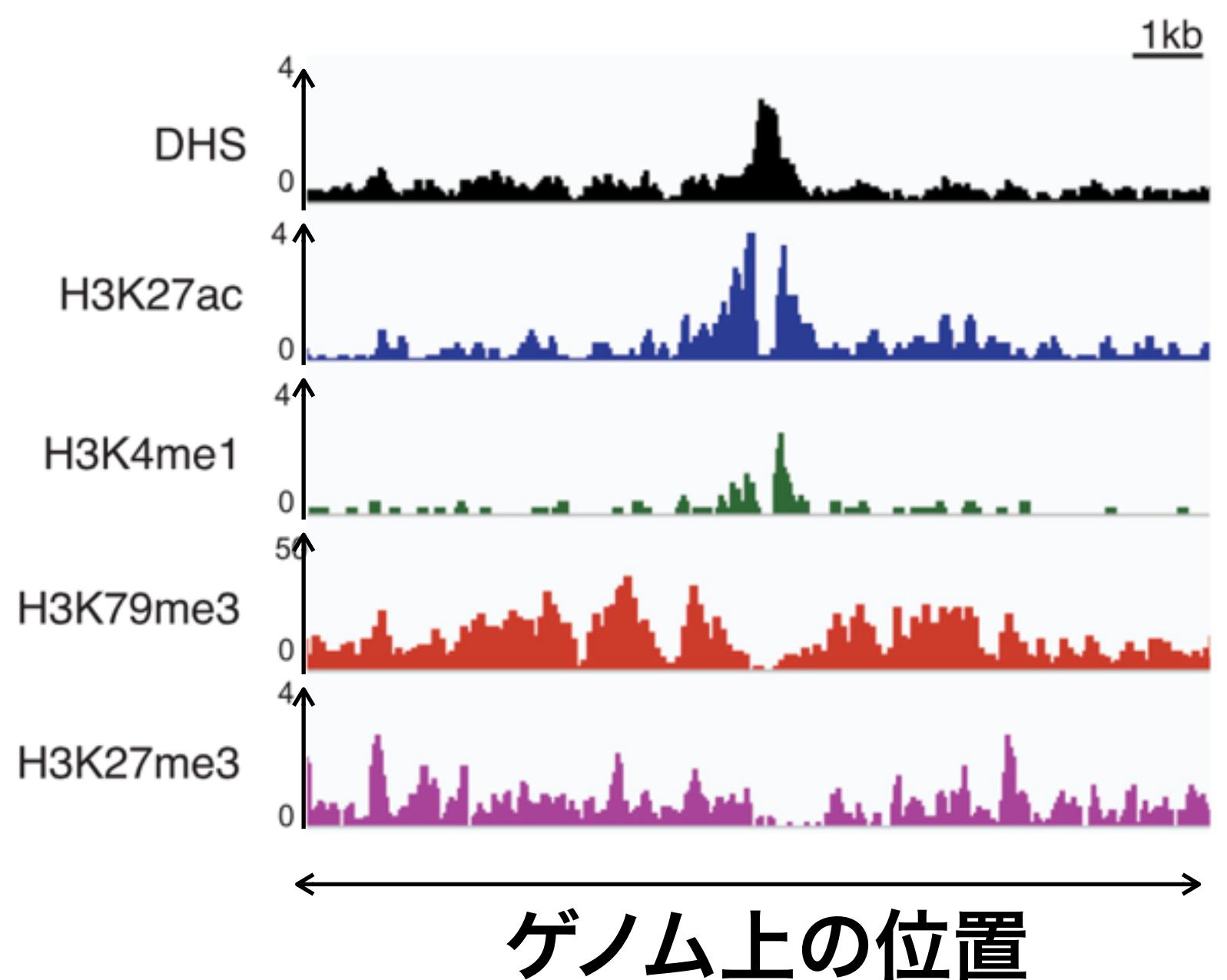


# ゲノムブラウザ

ゲノム上に表現されたデータを可視化できる

IGVが有名 <http://www.broadinstitute.org/igv/>

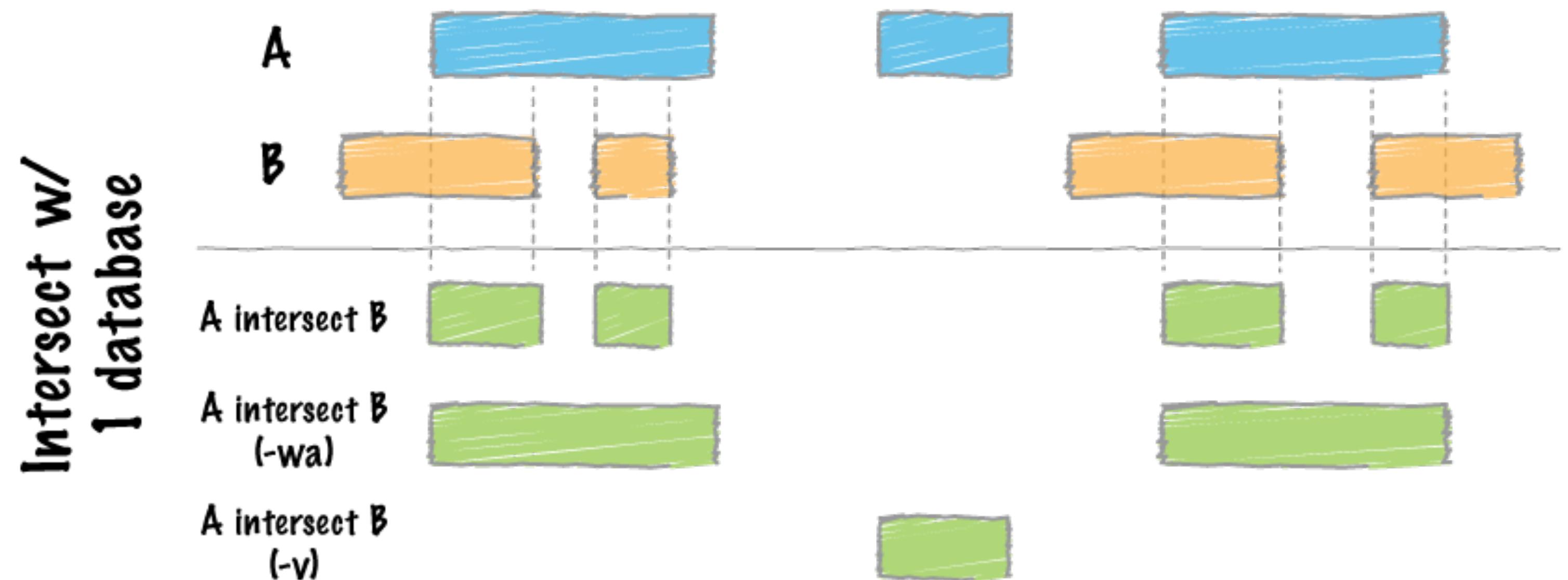
The screenshot shows the homepage of the Integrative Genomics Viewer (IGV) at [www.broadinstitute.org/igv/](http://www.broadinstitute.org/igv/). The page features a large central image showing a screenshot of the IGV software interface with multiple tracks of genomic data. To the left is a sidebar with links to Home, Downloads, Documents, Hosted Genomes, FAQ, IGV User Guide, File Formats, Release Notes, IGV for iPad, Credits, and Contact. Below the sidebar is a search bar and links to Broad Home and Cancer Program. At the bottom left is the BROAD INSTITUTE logo and copyright information for 2013.



# bedtools

BEDフォーマットを操作するコマンドラインツール群

<http://bedtools.readthedocs.io>



# samtools

BAM/SAMの操作ができる

<http://www.htslib.org>

The screenshot shows the homepage of the Samtools website at [htslib.org](http://www.htslib.org). The page has a clean, modern design with a light gray background. At the top, there's a navigation bar with tabs for Home, Download, Workflows, Documentation, and Support. The main content area features a large title "Samtools" and a brief description of what it is: "Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories: Samtools, BCFtools, and HTSlib." Below this, there's a note that "Samtools and BCFtools both use HTSlib internally, but these source packages contain their own copies of htslib so they can be built independently." At the bottom of the page, there are four sidebar boxes: "Download" (with a download icon), "Workflows" (with a workflow icon), "Documentation" (with a document icon), and "Support" (with a support icon). Each sidebar box contains some descriptive text and links.

**Samtools**

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

- Samtools** Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
- BCFtools** Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
- HTSlib** A C library for reading/writing high-throughput sequencing data

Samtools and BCFtools both use HTSlib internally, but these source packages contain their own copies of htslib so they can be built independently.

**Download**

Source code releases can be downloaded from [GitHub](#) or [Sourceforge](#):

[Source release details](#)

**Workflows**

We have described some standard workflows using Samtools:

- WGS/WES Mapping to Variant Calls
- Using CRAM within Samtools

**Documentation**

- Manuals
- Specifications
- Zlib Benchmarks
- CRAM Benchmarks
- Publications

**Support**

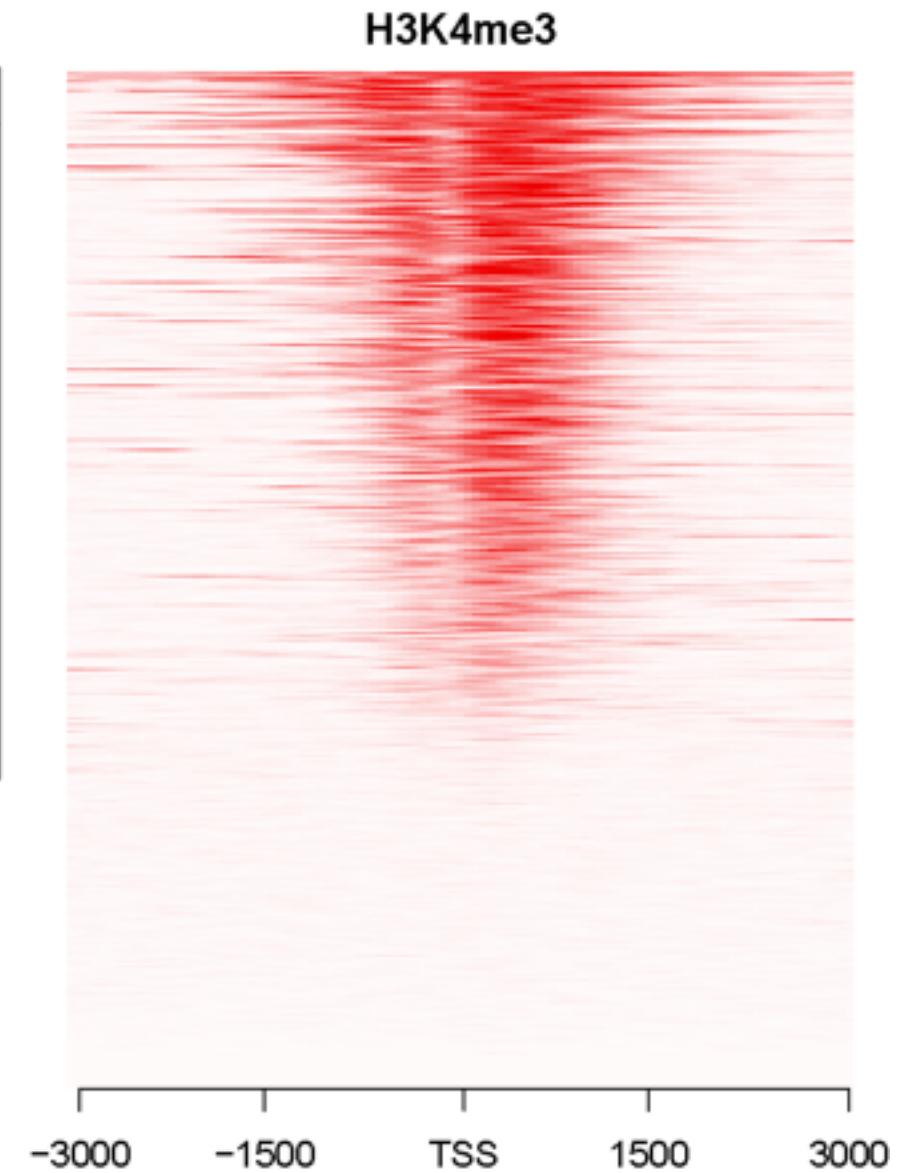
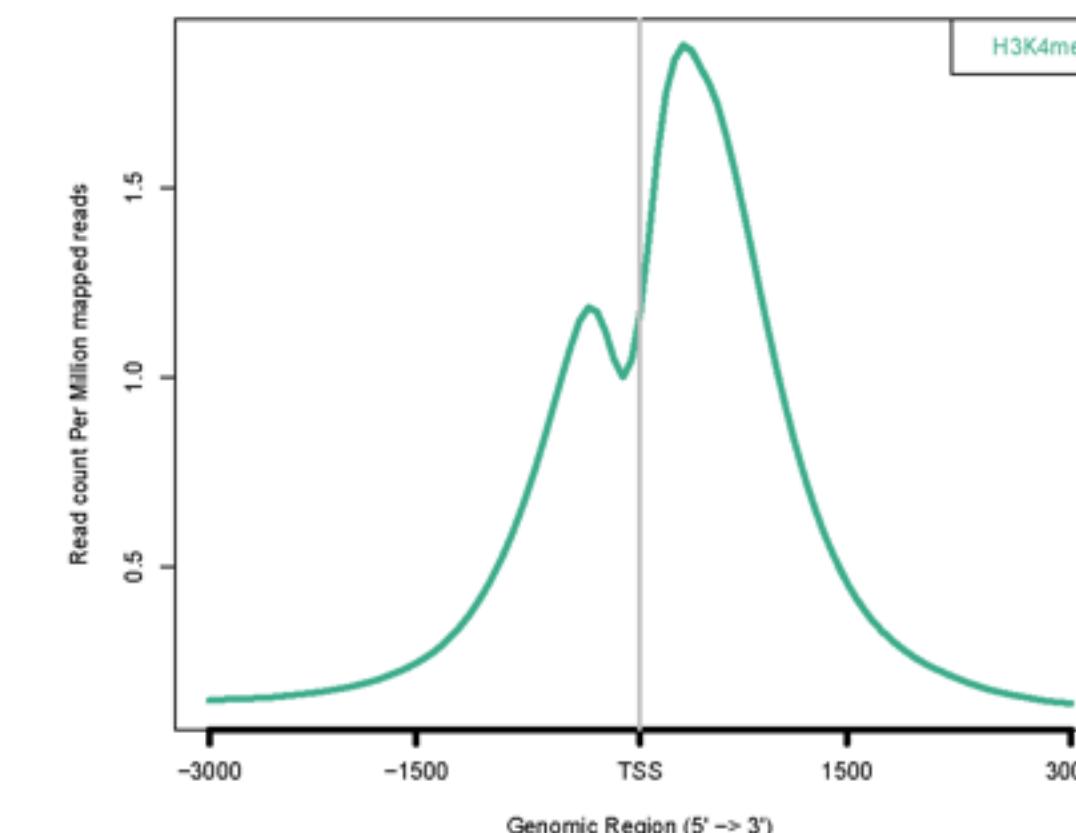
- Mailing Lists

# deepTools, ngs.plot

## Aggregation plot

ゲノム上の点の集合（例: TSSs）に対するNGSリードの分布

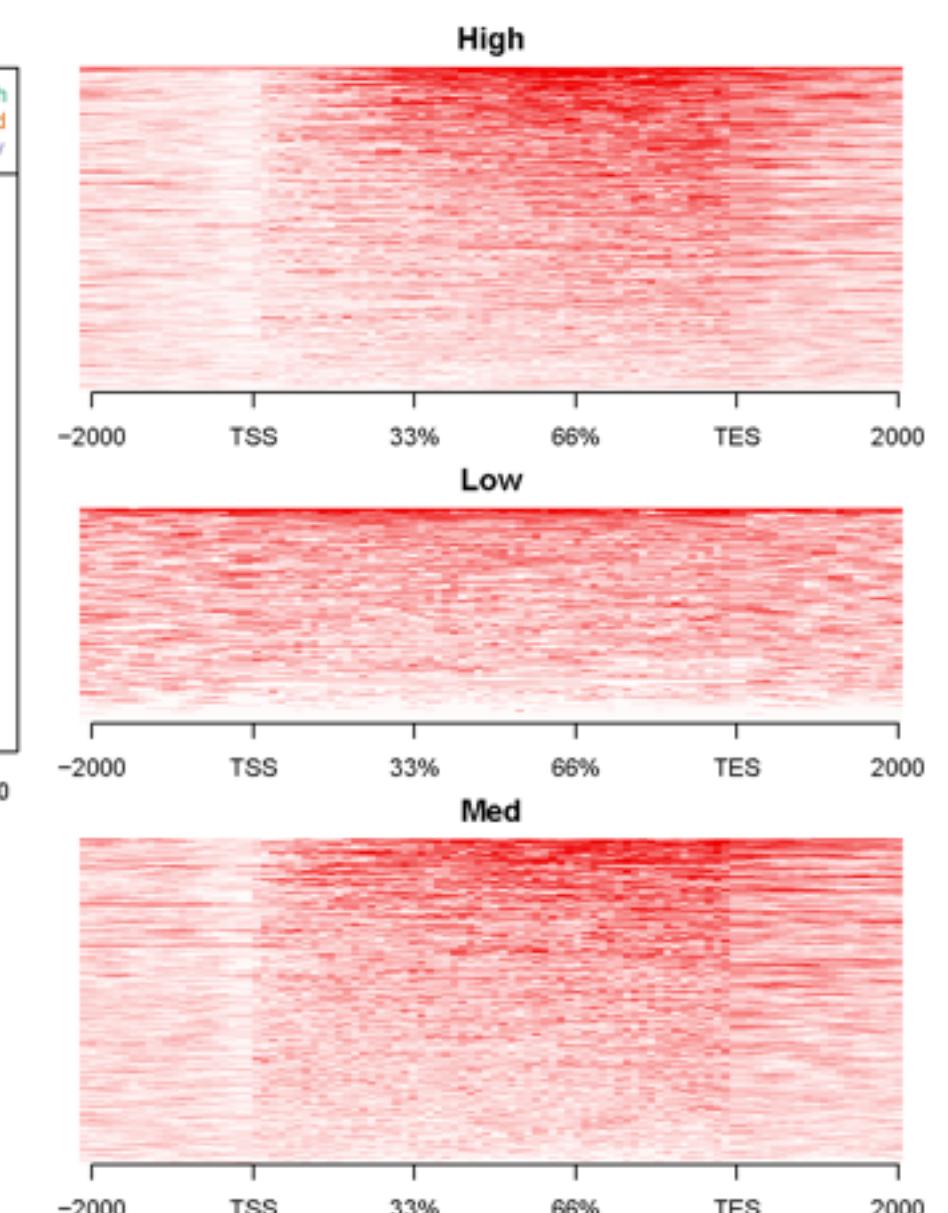
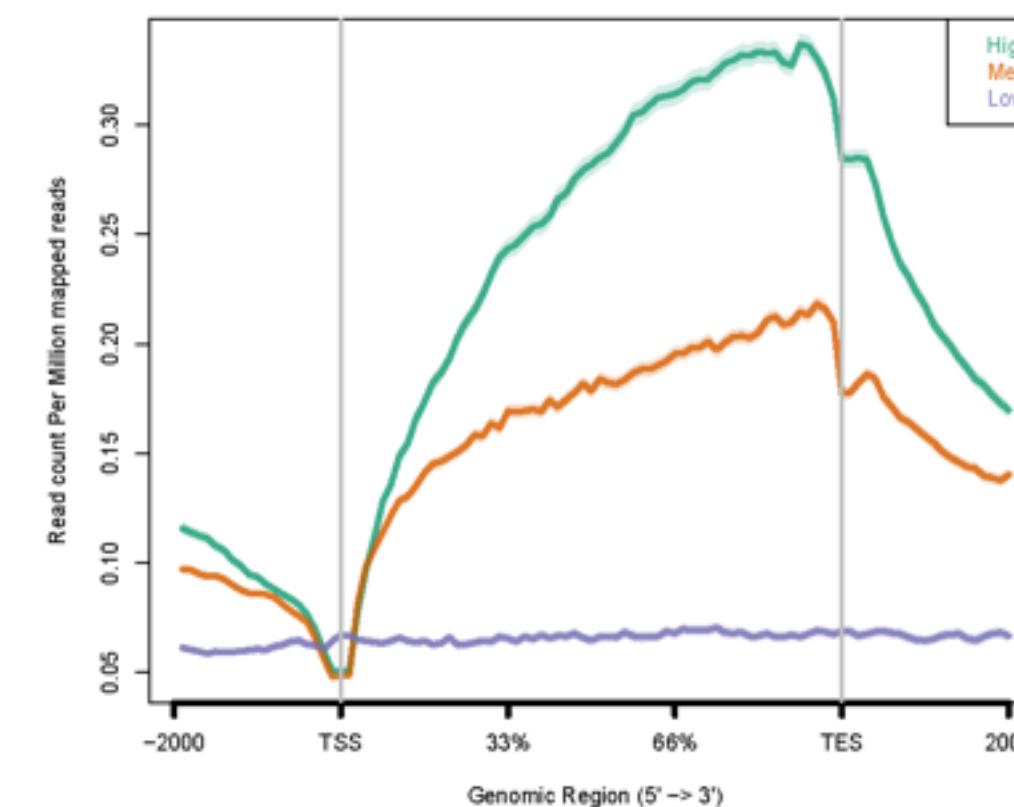
## Aggregation plot



## Meta-gene plot

ゲノム上の区間の集合（例: 遺伝子領域）に対するNGSリードの分布

## Meta-gene plot



## Heatmap

## deepTools

<http://deeptools.readthedocs.io/en/latest/>

## ngs.plot

<https://github.com/shenlab-sinai/ngsplot>

<https://github.com/shenlab-sinai/ngsplot>

# NGSデータ高次解析の基礎

よく使われる統計手法

# 基本的な可視化・要約

量を見る

棒グラフ

分布を見る

ヒストグラム、箱ヒゲ図、Violin plot

変数間の関係を見る

散布図、Density plot

層別



# クラスタリング

階層的手法

凝聚型

分割型

非階層的手法

k-means clustering

k-nearest neighbors algorithm

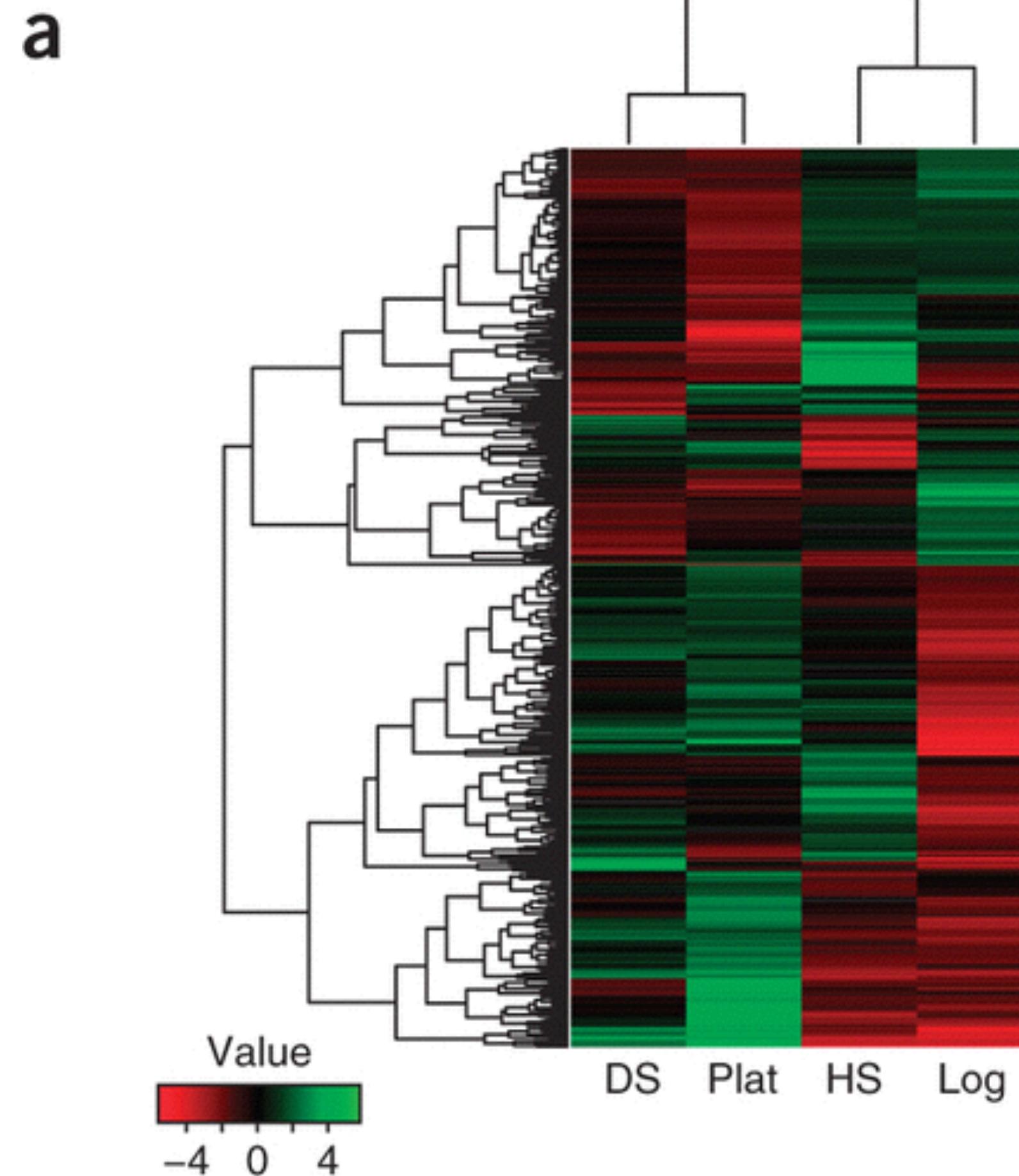
Spectral clustering

Gaussian mixture model

Biclustering

階層的

biclusteringによる  
遺伝子とサンプルのクラスタリング



# 次元圧縮 Dimensional reduction

データを低次元に射影してデータ全体の概略をつかむ

線形

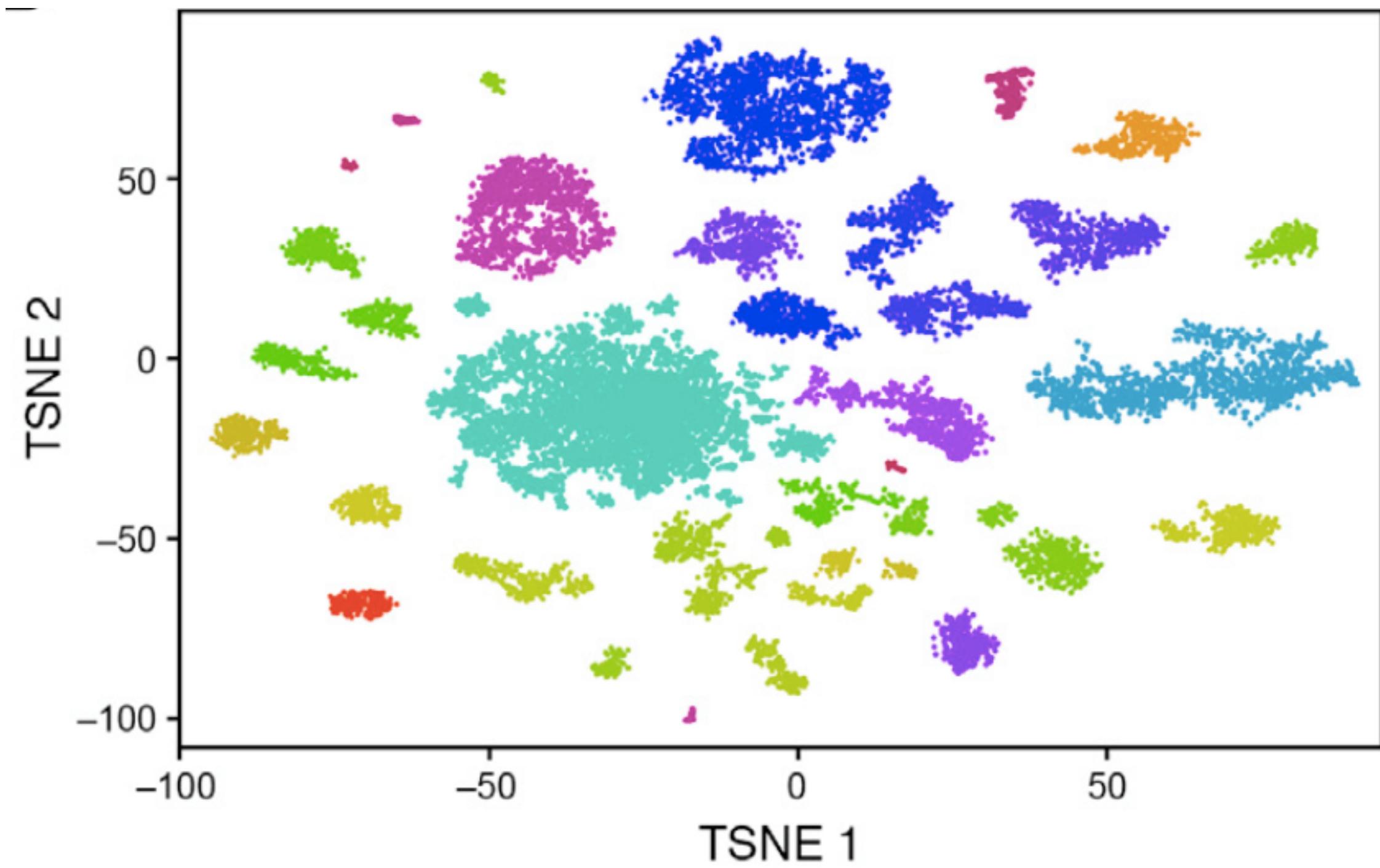
PCA、ICAなど

非線形

MDS、Diffusion map、t-SNEなど

詳しくは <http://www.slideshare.net/mikayoshimura50/150905-wacode-2nd>

t-SNEによるsubpopulationの可視化



# DNAモチーフ解析

配列群に濃縮した塩基パターンをモチーフとして抽出する

多数のソフトウェアが存在

MEMEが有名だが多くの配列は扱えない

DREME

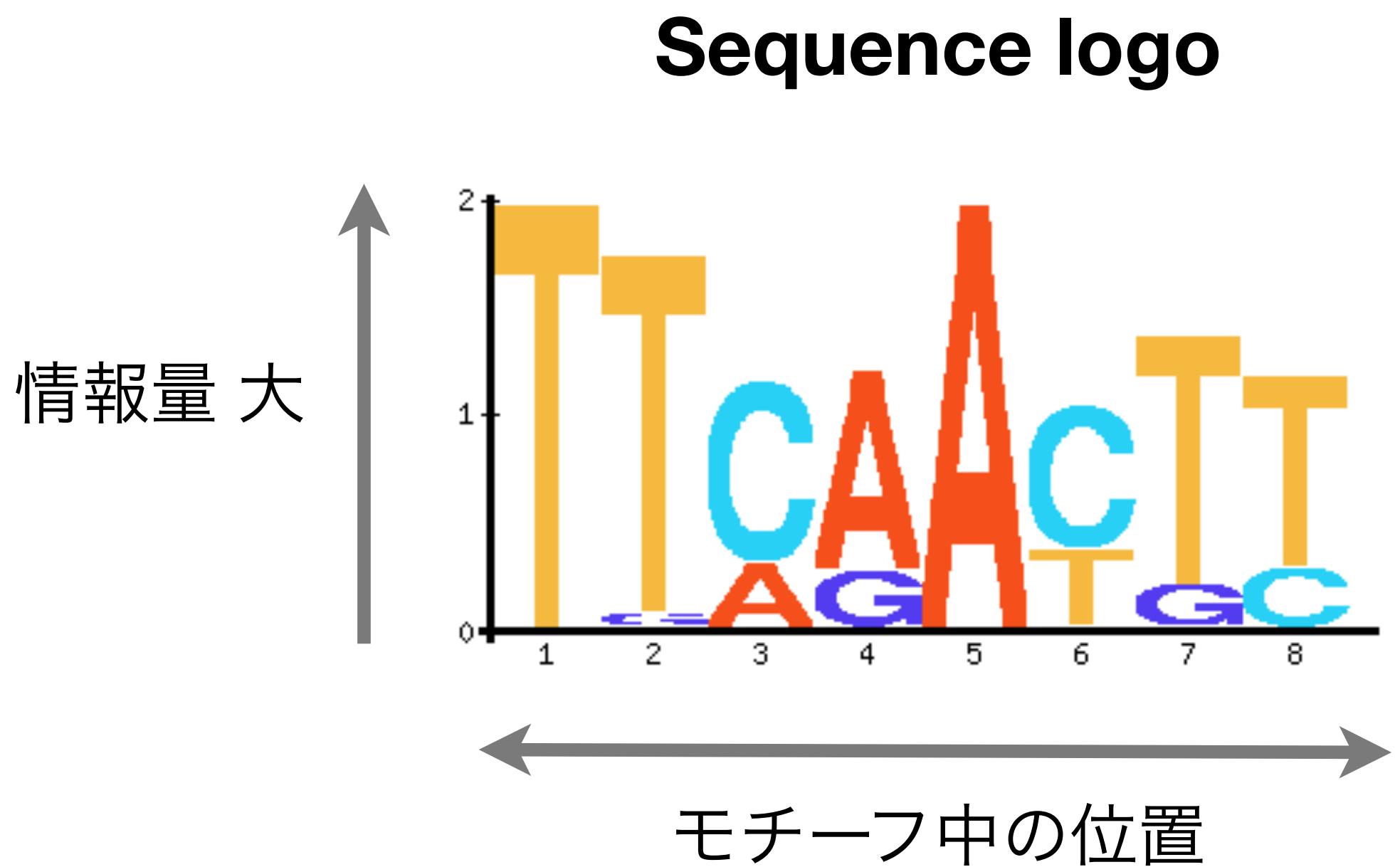
<http://meme-suite.org/doc/dreme.html>

HOMER

<http://homer.salk.edu/homer/ngs/>

MOCCS

<https://github.com/yuifu/moccs>



<http://molbio.mgh.harvard.edu/sheenweb/PromoterATAK&MZ06.html>

# ネットワーク解析

分子間の関係を表現

遺伝子制御ネットワーク

共発現遺伝子ネットワーク

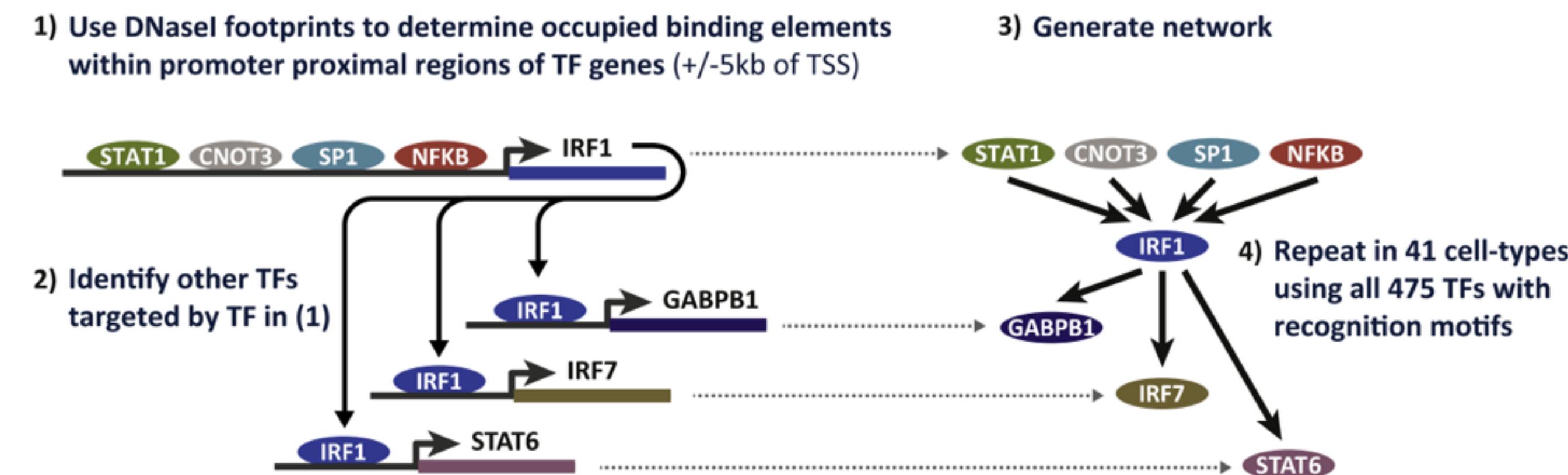
転写因子ネットワーク

タンパク質相互作用ネットワーク

様々なネットワーク構築手

法が提案されている

## DNase-Seqの解析データを利用した 転写因子ネットワーク構築の例



# 多重検定補正 Multiple testing correction

多重検定問題=NGS解析ではそれ以前の分子生物学ではない規模に検定する数が多いため、そのままでは偽陽性が生じやすい

例：遺伝子の数、転写因子結合領域の数

Bonferroni補正、Benjamini-Hochberg法、Sorey法など

詳しくは

FDRの使い方 <http://www.slideshare.net/yuifu/fdr-kashiwar-3>

[http://www.mbsj.jp/admins/ethics\\_and\\_edu/PNE/5\\_article.pdf](http://www.mbsj.jp/admins/ethics_and_edu/PNE/5_article.pdf)

[http://www.mbsj.jp/admins/ethics\\_and\\_edu/PNE/5\\_QandA.pdf](http://www.mbsj.jp/admins/ethics_and_edu/PNE/5_QandA.pdf)

# NGSデータ高次解析の体験

こちらにアクセス: [https://github.com/yuifu/AJACS\\_Kyoto\\_2](https://github.com/yuifu/AJACS_Kyoto_2)

# NGS高次解析の自主学習の方法

# 自分で解析環境をセットアップ

A. 次世代シークエンサーDRY解析教本

<https://www.amazon.co.jp/dp/B0185JENAK/>

B. NGSハンズオン講習会の資料

<http://biosciencedbc.jp/human/human-resources/workshop/h28-2>

# 積極的に調べる

検索する



英語の方が情報が多い場合が多い

QAサイト

ライフサイエンスQA (β) (日本語) <http://qa.lifesciencedb.jp>



SEQanswers (英語) <http://seqanswers.com>



BioStar (英語) <http://www.biostars.org>



誰かに聞く (オンライン・メーリングリスト)

NGS現場の会 <http://www.ngs-field.org/top-page/join/>



誰かに任せる (共同研究・受託解析)

# NGS現場の会に参加する

NGSを軸に「現場」の人間が交流する

## 研究コミュニティ

初心者～ベテラン、学生～教授、研究者・技術者・  
営業職、大学・研究所・産業界、医学・農学・薬  
学・工学から基礎科学まで

## 研究会

現場の人間が一堂に会し、オープンでフラットな  
交流を行う

## 情報共有

Wiki、メーリングリスト、QAサイトなど

# NGS 現場の会

<http://www.ngs-field.org>



## 第5回研究会

会期：2017年5月22日 - 24日

会場：仙台国際センター 展示棟

# ソフトウェア・解析プロトコルを調べる: どれを選べばよいのか?

## 性能がよいソフトウェア

性能はそのソフトウェアの論文や評価論文を読んで調べる

たいていソフトウェア論文を出していて、他のツールとの比較をしている

評価論文は“Evaluation”とか”Assessment”で検索すると出てくる

自分でベンチマーク（評価）する

## みんなが使っているソフトウェア（≠最高性能のソフトウェア）

ノウハウが豊富なため、エラーが出たときなどに対処しやすい

あまり使われていないソフトウェアをあえて選ぶと、論文を書くときに「なぜわざわざそれを選んだか」を説明しないいけないこともある

# 本日のまとめ

NGSデータのフォーマット

NGSデータ高次解析で使われるツール・統計手法

NGSデータ高次解析のツール

解析の方法を調べるためのノウハウ

# Further reading: 書籍



次世代シーケンサーDRY解析教本



次世代シーケンス解析スタンダード

# Further reading: NGSデータ解析のチュートリアル

平成28年度NGSハンズオン講習会

<http://biosciencedbc.jp/human/human-resources/workshop/h28-2>

(Rで)塩基配列解析

[http://www.iu.a.u-tokyo.ac.jp/~kadota/r\\_seq.html](http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html)

biopapyrus

<http://biopapyrus.net>

統合TV (NGS解析だけでなくDBなども)

<http://togotv.dbcls.jp/>

# Further reading: NGSデータ解析の資料（英語）

RNA-seqlopedia

<http://rnaseq.uoregon.edu/>

EMBL-EBI のオンライントレーニング

<http://www.ebi.ac.uk/training/online/>

R Bioconductor のチュートリアル

<http://bioconductor.org/help/course-materials/2016/BioC2016/>

# Further reading: Linux関連

Linux環境でのデータ解析：JavaやRの利用法

<http://biosciencedbc.jp/human/human-resources/workshop/h28-2>

Linux標準教科書

<http://www.lpi.or.jp/linuxtext/text.shtml>

# 実験系の方からよく聞かれる質問

「バイオインフォを勉強するのにプログラミングってどのくらい必要なんですか」

既存のプロトコルを（コピペ）でなぞるだけなら特にプログラミングは必要ない

実験に例えると、キットを使った実験だけができる（プロトコルの改変はできない）

既存のソフトウェアを組み合わせて、適宜

実験に例えると、実験系を組んだり、プロトコルを改変できるイメージ

# バイオインフォレベル2

#NGLSBI

## バイオインフォマティクス研究者の分類(改) ～富山城の天守に喩えて～

DBCLS Database Center for Life Sciences

3. ガチ系  
2. コマンドライン系  
1. コピペ系  
0. 他力本願

21

## 2. コマンドライン系バイオインフォマティクス

- UNIXのコマンドライン上で、既存のツールを組み合わせて解析をする
  - Command line User Interface(CUI) (cf. GUI)
- たまに捨てコードを書く
- 武器
  - shell script
  - Perl, Ruby
  - Python
  - R

23

© 2013 DBCLS Licensed under CC 表示 2.1 日本

# バイオインフォレベル2

【分類表】

No.	カテゴリー	能力
1	基礎/応用研究者 (ドライ)	自分で生物の問題を発見し、定式化し、必要に応じて新規のアルゴリズム、情報技術やDBを開発し、問題を解くことができる。
2	基礎/応用研究者 (ドライ)	新しい情報技術、DB、アルゴリズムを開発できる。 生物系の研究者と共同研究して問題を解ける。
3	基礎/応用研究者 (ドライ)	既存の情報技術、DBを使って問題を解ける。 生物系の研究者と共同研究して問題を解ける。
4	基礎/応用研究者 (ドライ+ウェット)	自分でウェットの研究開発を行い、新しい情報技術、DB、アルゴリズムを開発できる。
5	基礎/応用研究者 (ドライ+ウェット)	自分でウェットの研究開発を行い、既存の情報技術、DBを使って問題を解ける。
6	基礎/応用研究者 (ウェット)	自分で生物の問題を発見したり、定式化したりできる。 情報系の研究者と共同研究して問題を解ける。
7	基礎/応用研究者 (ウェット)	自分で生物の問題を発見したり、定式化したりできる。 情報系の企業にデータの解析を依頼して問題を解ける。
8	支援的研究者 (プログラマー)	カテゴリー1, 2, 3, 4, 5の研究者と協力して、プログラムを作り、支援的な研究開発ができる。
9	支援的研究者	ツールやDBを使ってカテゴリー4, 5, 6, 7の研究者の支援的研究ができる。
10	支援的研究者 (アノテータ、キュレータ)	カテゴリー1, 2, 3, 4, 5, 6, 7の研究者と協力して、データのアノテーション、DBのキュレーションなどの研究開発ができる。
11	支援者(SE)	DBや情報インフラの管理を通じて研究支援ができる。
12	その他	現時点ではバイオインフォマティクスとの関わりは特になし。

このへん

# 再現性を担保するために

## ソフトウェア・アノテーションのバージョンを記録

バージョンによって結果が異なる場合があるため

## コマンドラインに入力したことを記録

ソフトウェア実行時のパラメタなど

できれば、スクリプトファイルとして残しておく

ディレクトリ名を日付にすると後々便利（個人の感想）

# 心構え

エラーが出たら

落ち着いてエラーメッセージ（英語）を読む

エラーメッセージをGoogle検索する

分からなかったら人に聞く（対面 or オンライン）

実験と違ってやり直せる

