

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

アクセスできるか確認してみよう

本日の流れ

講義の途中に適宜ハンズオンで手を動かす

いつでも質問を受け付けます

よくわからなかつ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

講義の目的・到達目標

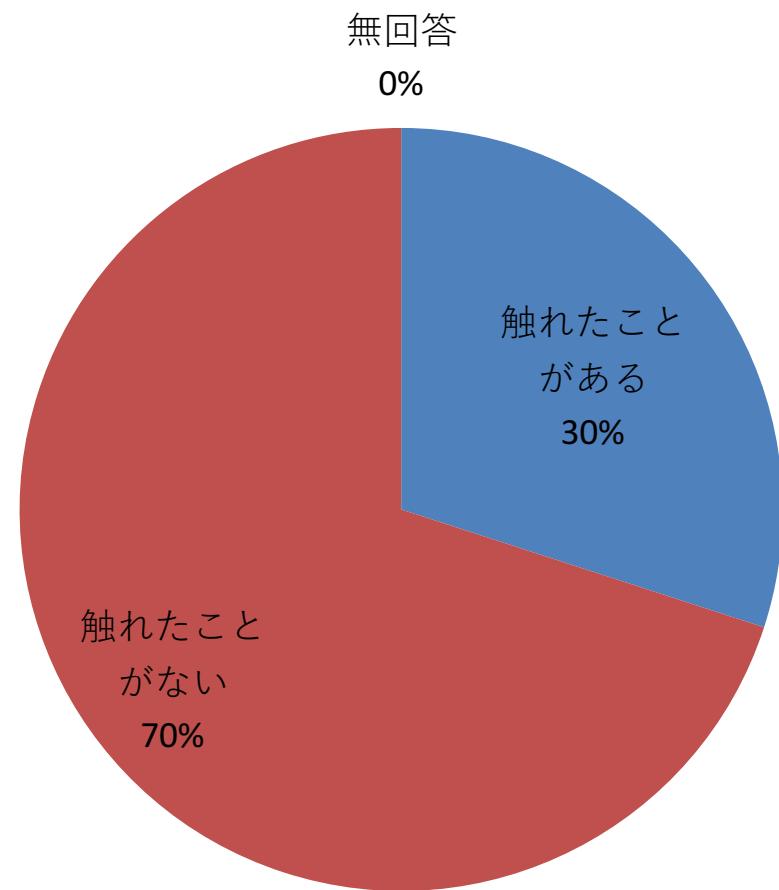
目的

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

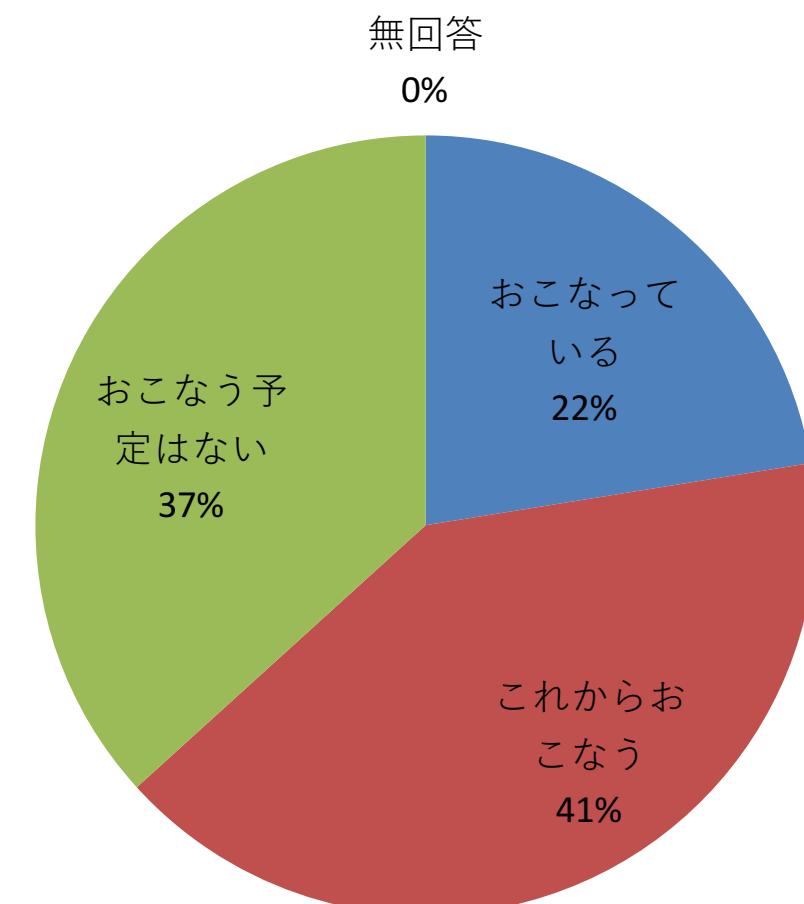
到達目標

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。



この講義の内容

NGSの基礎

NGSについて

NGSデータ解析の基礎

NGSデータ高次解析の基礎

NGSデータ高次解析の実例

NGSデータ高次解析で使われるツール・統計手法

NGSデータ高次解析の体験

さらに学習を進めるために

NGSについて

NGSとは

次世代シーケンサー (Next Generation Sequencer)

DNA塩基配列を大量・高速に読み取る機械

CACGTGGG
ATGCATGC TTGGAACC
CGCATCGA
TACTCTAA AGATGGGA
TTGGAACC

DNA



NGS

データ

第一世代と次世代シークエンサーの違い

塩基配列決定時に電気泳動を必要としない。

第二世代と第三世代シークエンサーの違い

錆型のPCR増幅が必要ない。

第三世代と第四世代シークエンサーの違い

蛍光色素を使わない。

次世代シークエンサー

第二世代

Roche



illumina



Life Tech



Complete Genomics



QIAGEN



第三世代

Helicos

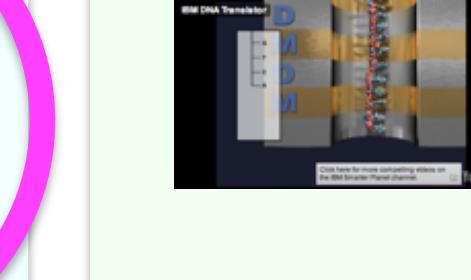


第四世代

Oxford



PacBio



IBM

NGSは多様な生命現象をDNAの形で読むのがすごい

*-Seq=「どんなDNAをNGSに入れるか」

ChIP-seq: あるDNA結合タンパク質が結合しているDNAを免疫沈降で集めたDNA

RNA-seq: RNAを逆転写してつくったcDNA

○○-seqの解析=そのデータが由来する生命現象を反映した解析が必要

ChIP-seq: ピーク検出

RNA-seq: 発現量定量

DNA/Genome

WGS
Exome
RC-Seq
Tn-Seq
TC-Seq

DNA-タンパク質
相互作用

DNase-Seq
MAINE-Seq
ChIP-Seq
FAIRE-Seq
ATAC-Seq
ChIA-PET
Hi-C/3C-Seq
4-C or 4C-Seq
5-C

RNA

発現
ChIRP-Seq
GRO-Seq
Ribo-Seq
RIP-Seq
HITS-CLIP
CLIP-Seq
PAR-CLIP
iCLIP
NET-Seq
TRAP-Seq
CLASH-Seq
PARE-Seq
GMUCT
TIF-Seq
PEAT

構造

SHAPE-Seq
PARS-Seq
FRAG-Seq
CAP-Seq
CIP-TAP
ICE
MeRIP-Seq

DNA Methylation

BS-Seq
PBAT
T-WGBS
oxBS-Seq
TAB-Seq
MeDIP-Seq
MethylCap
MBDCap
RRBS-Seq

微量検体検出

RNA	DNA
Quartz-Seq	smMIP
DP-Seq	MDA
Smart-Seq	MALBAC
Smart-Seq2	OS-Seq
UMI	Duplex-Seq
CEL-Seq	
STRT-Seq	

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使っただけでトップジャーナルに載る時代ではない

*-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）

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

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	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

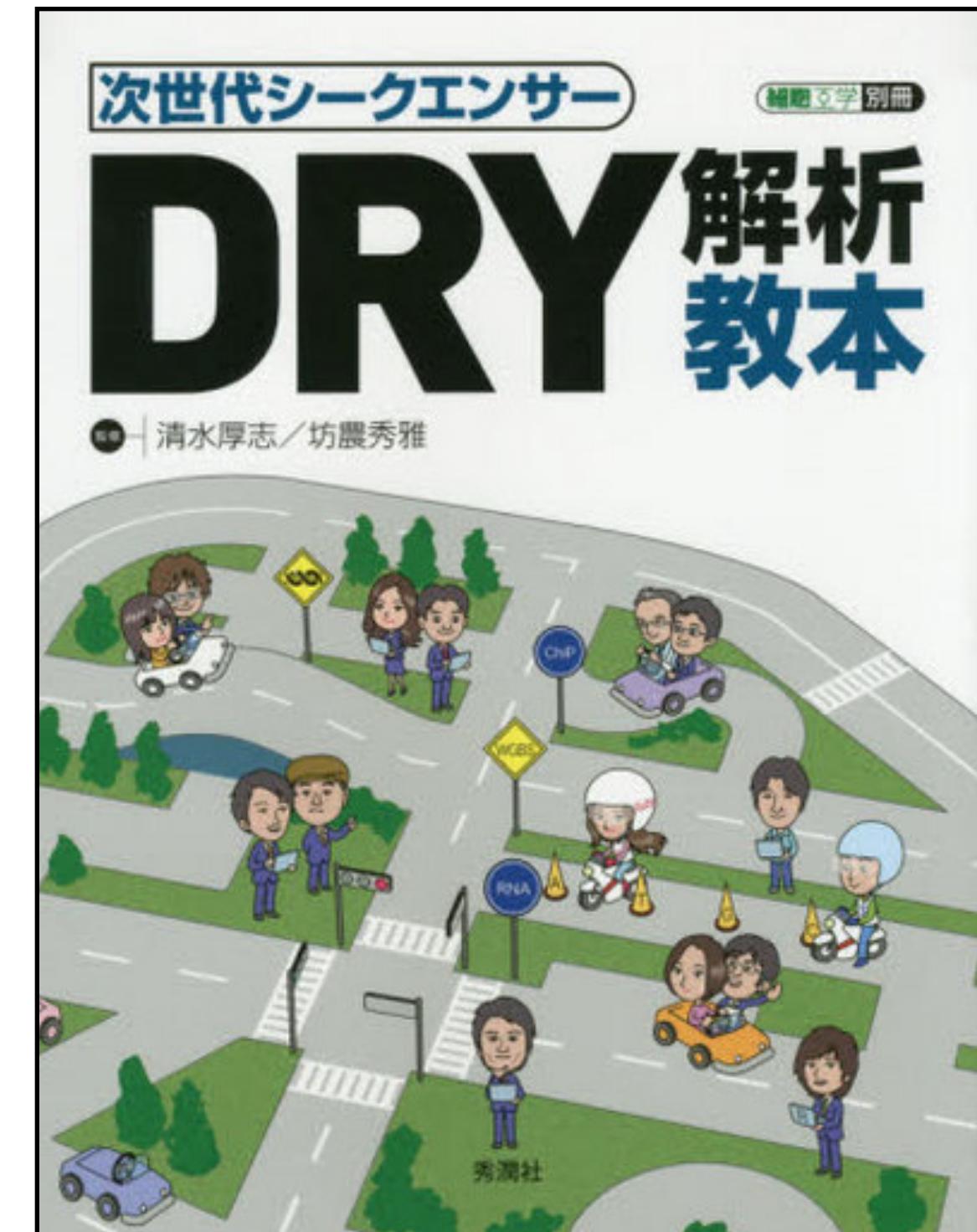
NGSデータ解析の資料が充実しつつある

次世代シーケンサー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 "H27年度NGSハンズオン講習会". Below it, a navigation bar 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 main content area displays a summary of the workshop, including 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 the implementation report and survey results.



大量のNGSデータが蓄積・公開されている

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>

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

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

0-basedと1-based

塩基配列

A, T, G, Cから成る文字列

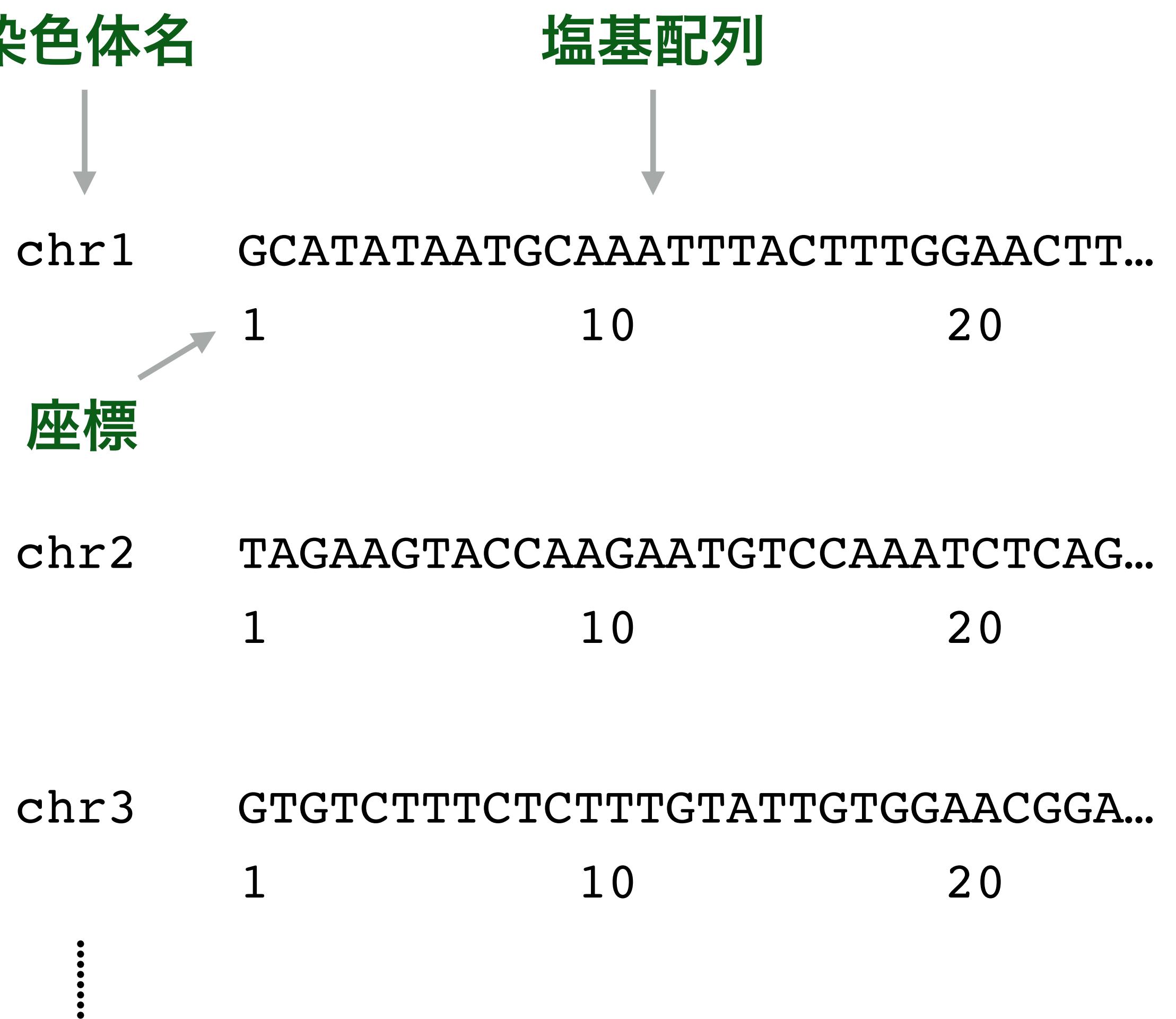
GCATATAATGCAAATTACTTGGAACTT...

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



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

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



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）

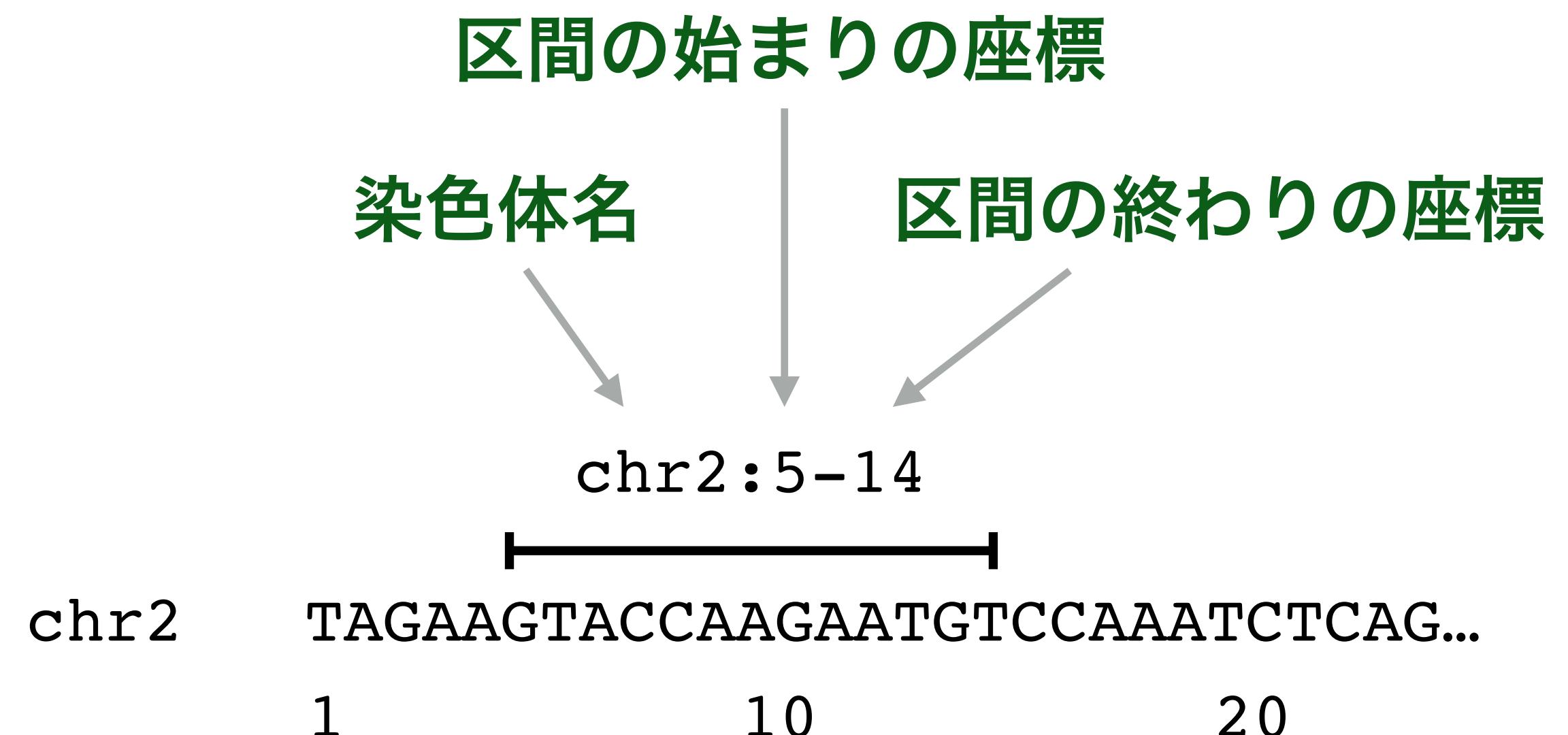
終点の座標（End）

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

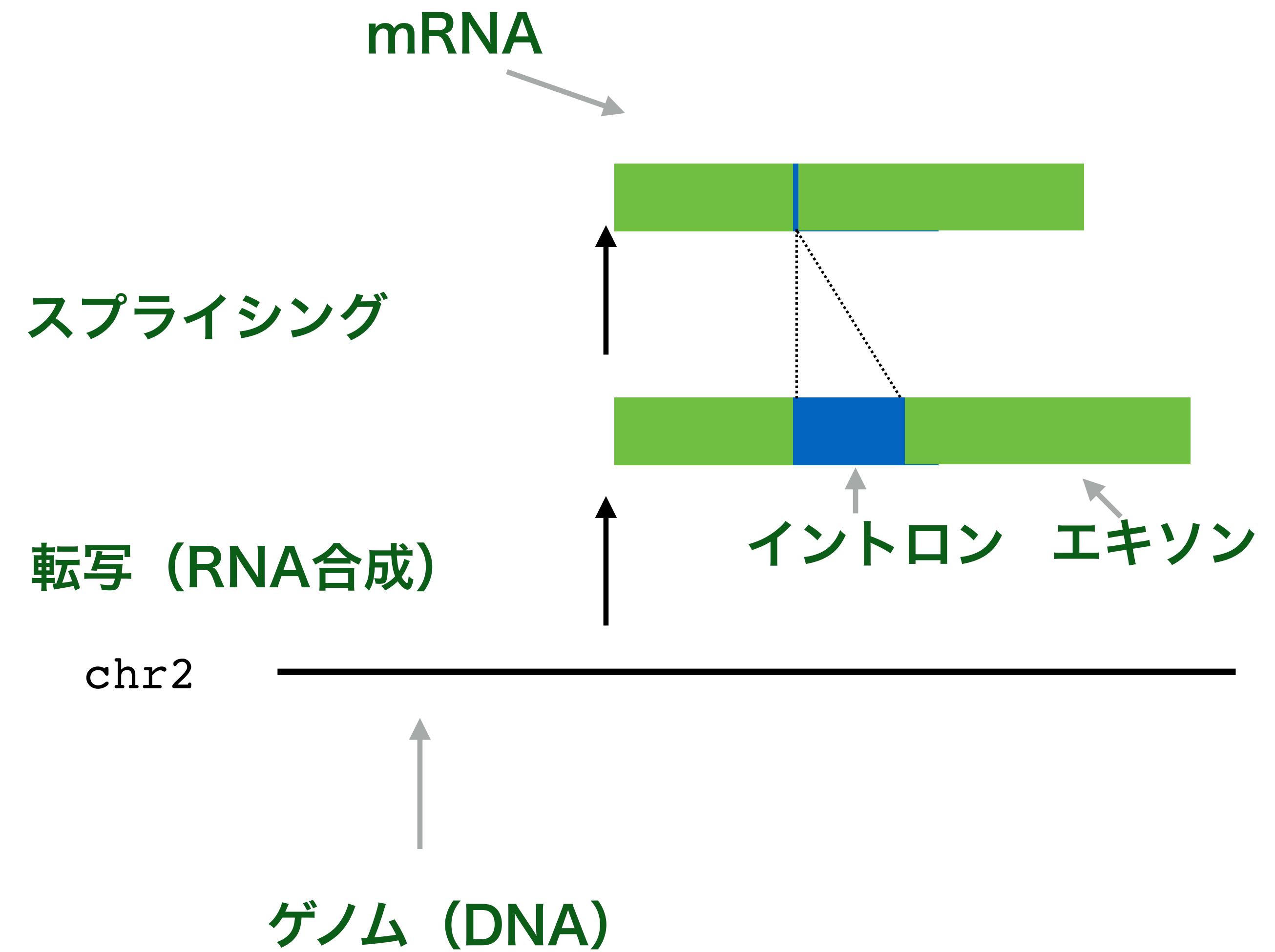
ゲノムの中でたんぱく質の情報を持っている部分

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。



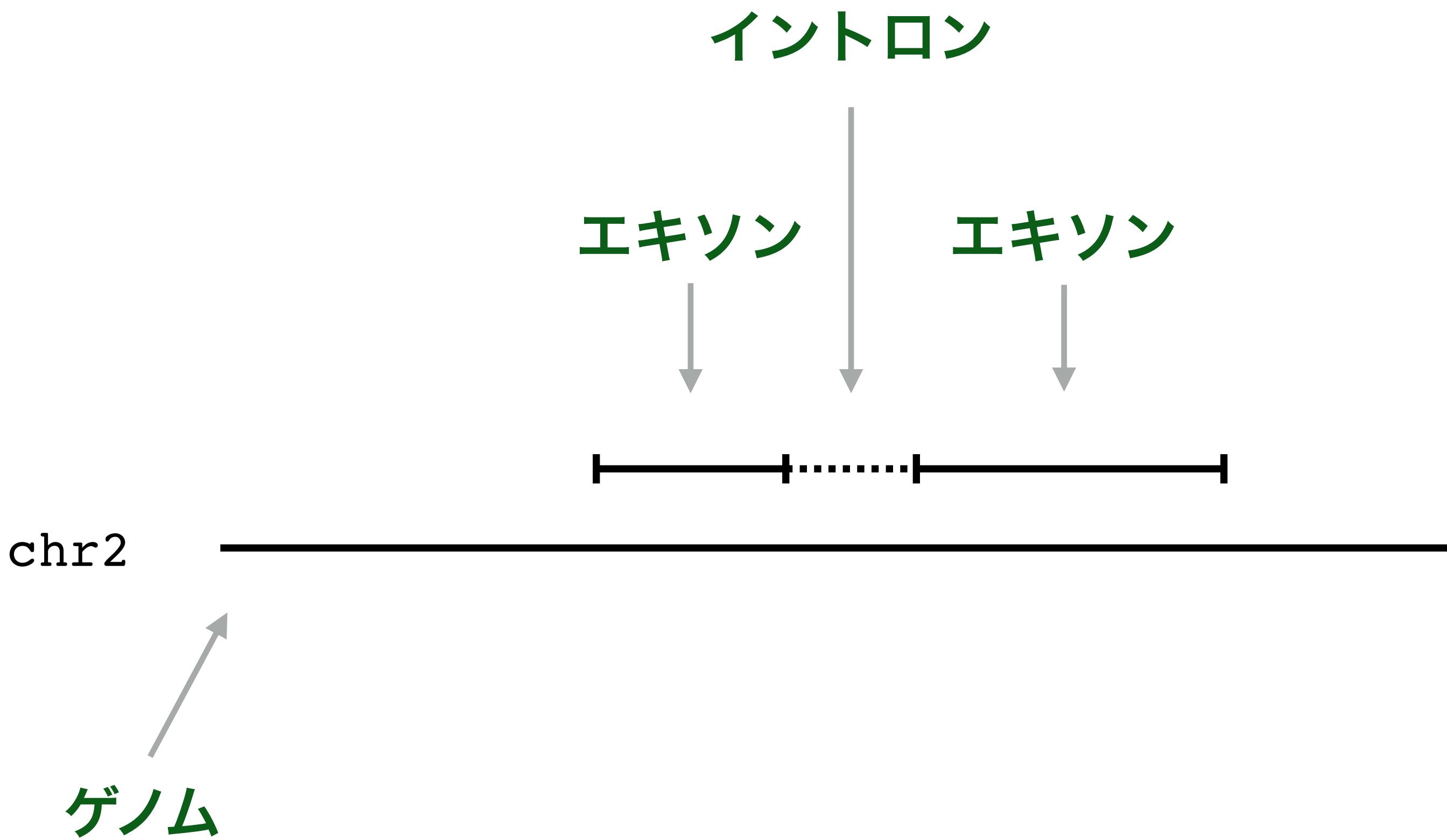
区間の集合としての遺伝子

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

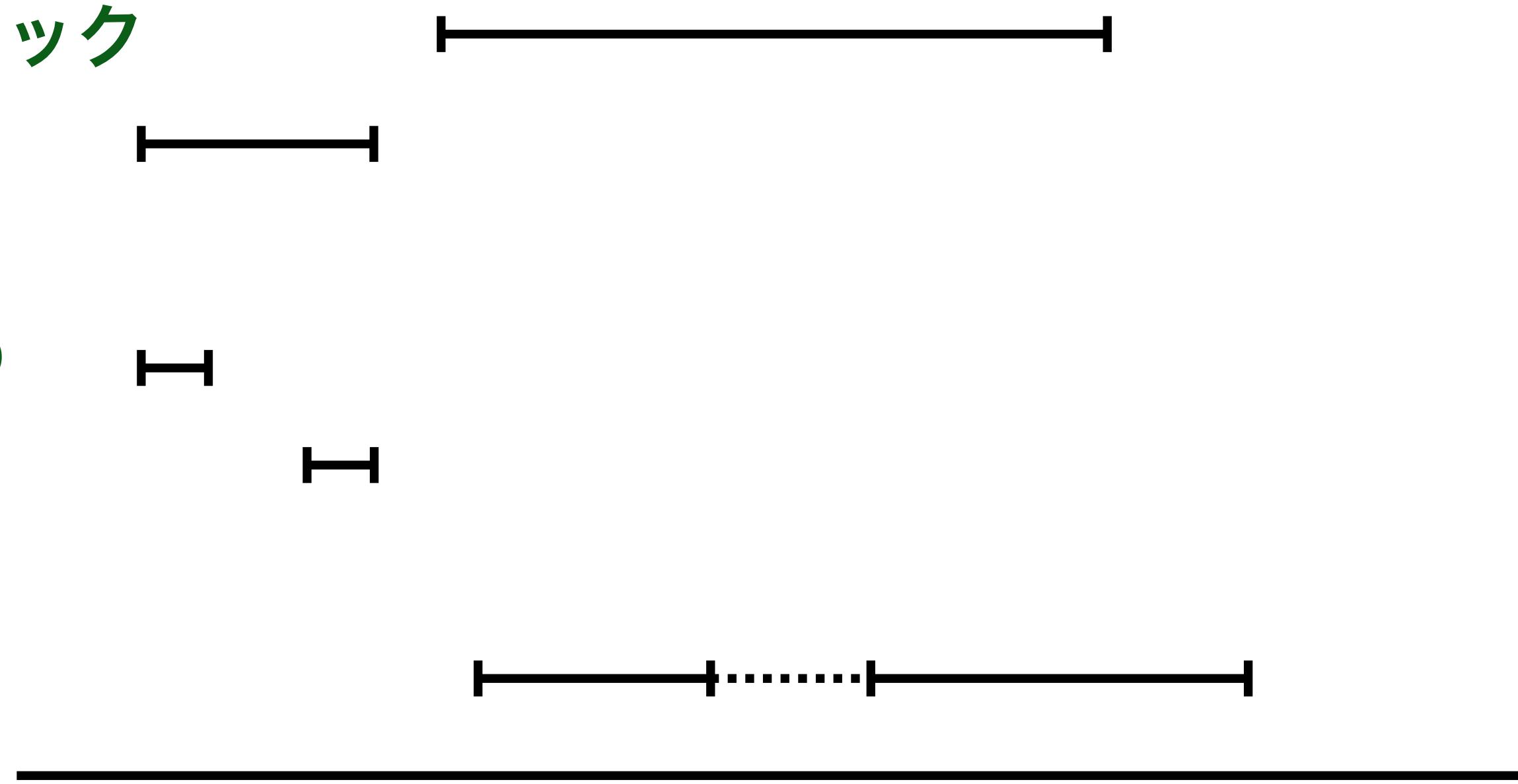
DNAが巻きつくヒストンたん
ぱく質の化学修飾やCのメチ
ル化

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>

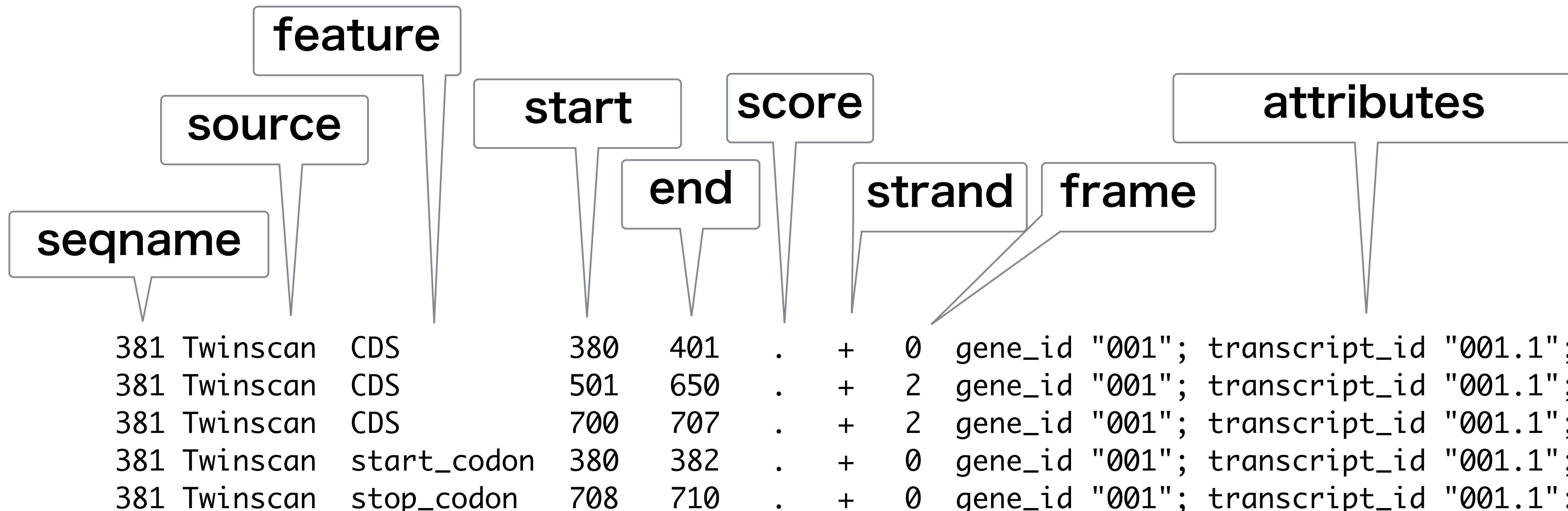
より複雑な表現が可能なBED12フォーマットもある

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>



座標には1-basedと0-basedがある

1-based coordinate system

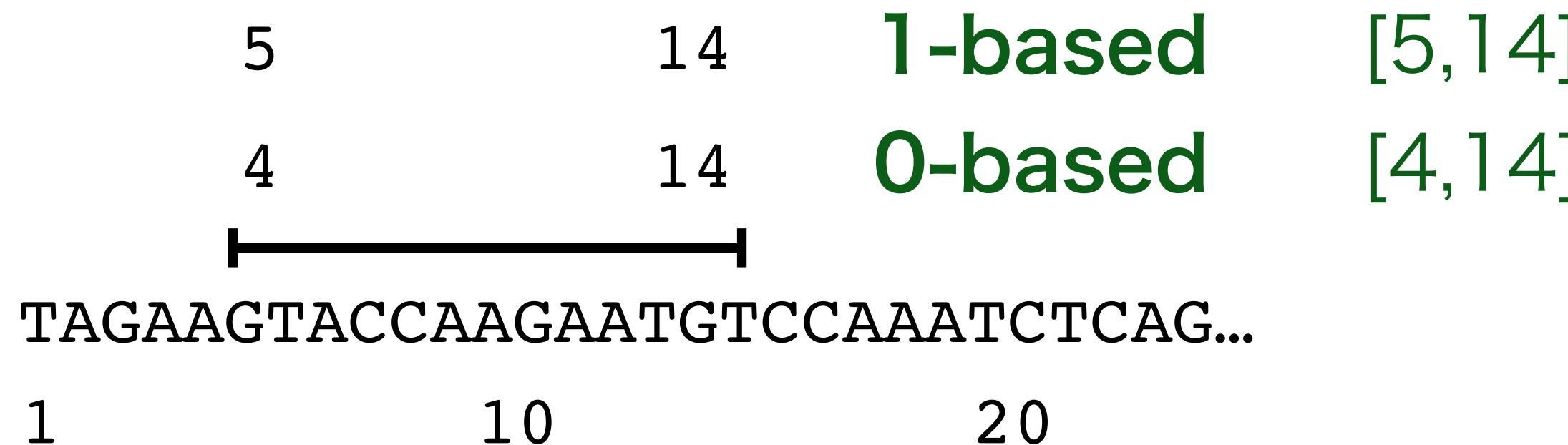
Closed interval

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

0-basedと1-based

NGSデータ解析の基礎

NGSデータ解析の流れ

NGSを利用した研究の一般的な流れ

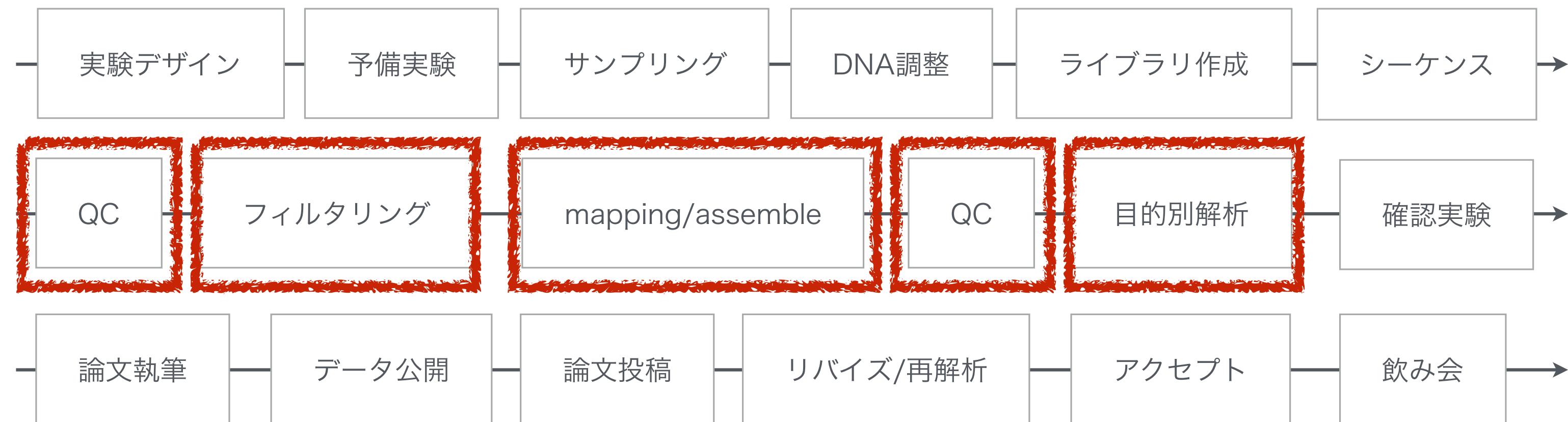
NGSを使う研究は何が大変なのか



- ・ 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

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



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

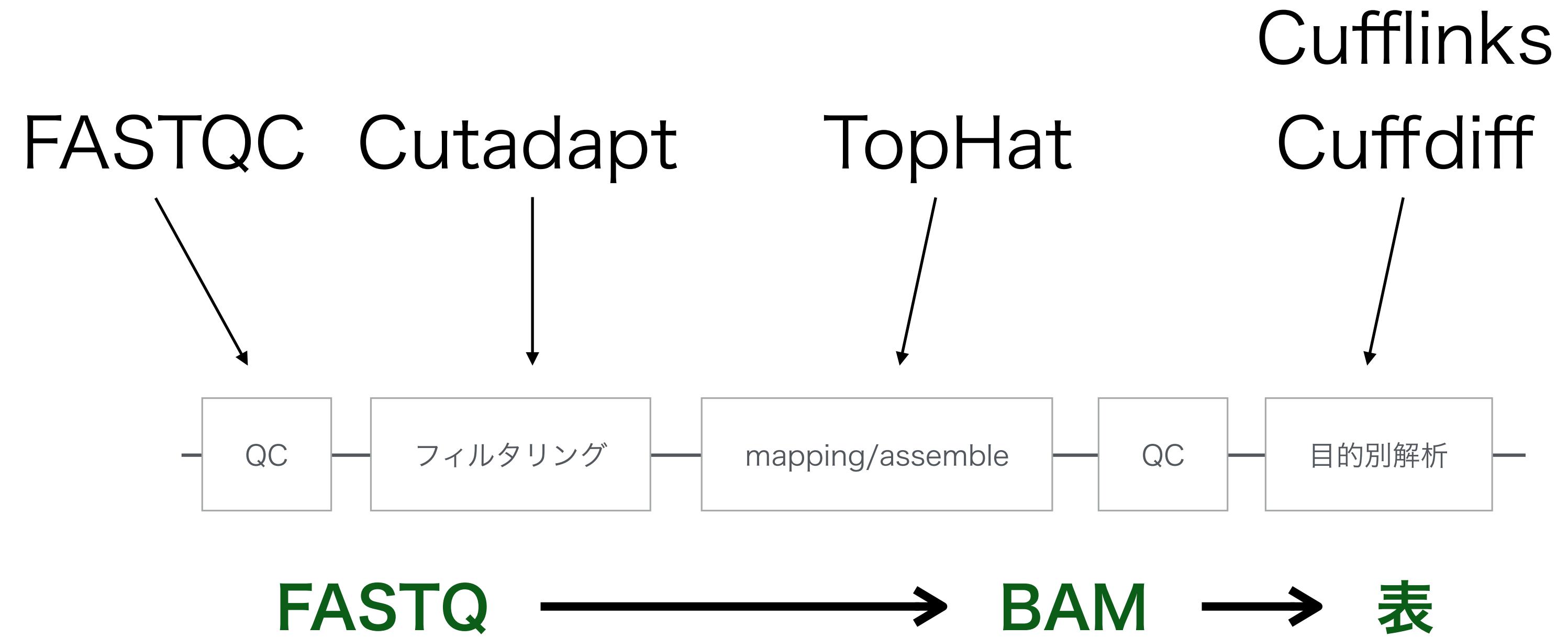
ゲノム/トランスクリプトームのインデックスの構築

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）

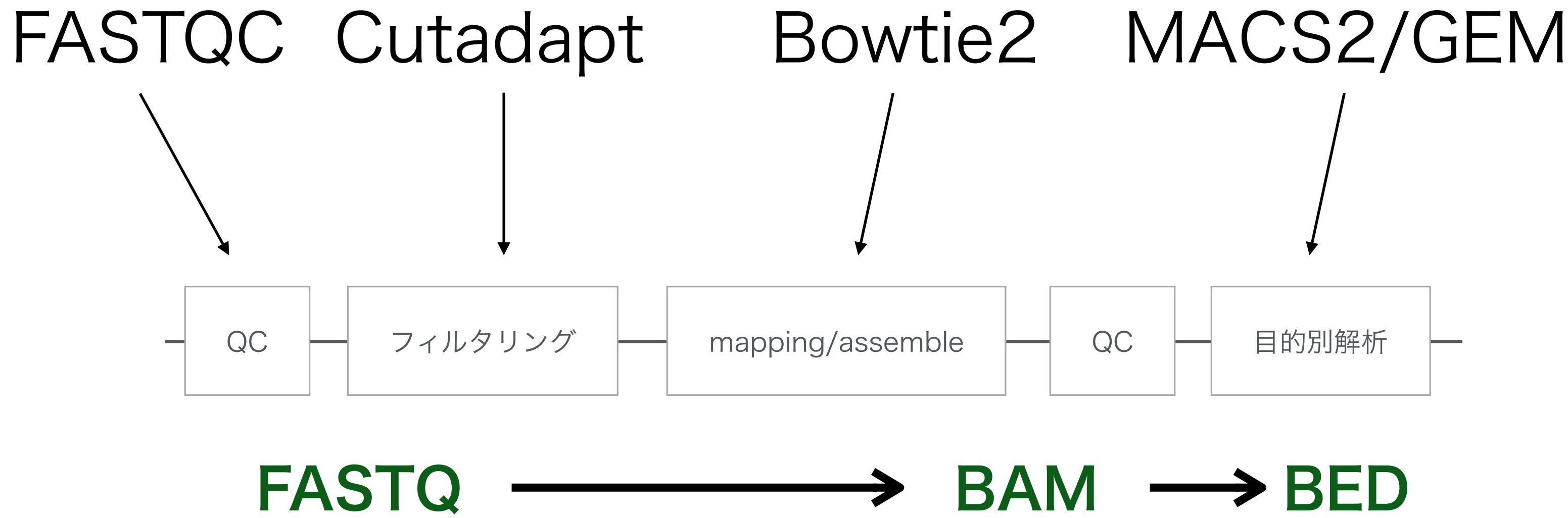


*これがベストプラクティスとは限らない

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

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

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

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

ENCODE: Encyclopedia of DNA Elements

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Quick Start

To find and download ENCODE Consortium data:

- Click the Data toolbar above and browse data
 - By assay
 - By biosample
 - By genomic annotations
- Enter search terms like "skin", "ChIP-seq", or "CTCF"

Additional help using the ENCODE Portal:

2-1. Matrixを選択

Experiment Matrix

Assay category	Target of assay	Date released	Available data
DNA binding	histone	July, 2013	fastq
Transcription	histone	March, 2014	bam
DNA accessibility	transcription factor	July, 2016	bigWig
WGRS	control	May, 2016	bed narrowPeak
methyl450k	broad histone mark	October, 2011	bigBed
Computational predictions and RT-PCR			narrowPeak
RNA-seq			
CUT-seq			
RIP-seq			

11736 results

ASSAY	...and 22 more
ChIP-seq	
DNase-seq	
DNase mRNA RNA-seq	
DNase RNA-seq	
dnRNA-seq	
RNA microarray	
ECU	
WGRS	
small RNA-seq	
genotyping array	
RAMPAGE	
RNA Bind-n-Sq	
CAGE	
single cell RNA-seq	
Dnase array	
Replic-seq	
microRNA counts	
microRNA-seq	
Bip-seq	

BIOSAMPLE

Organism	...and 175 more
Homo sapiens	8497
Mus musculus	1657
Drosophila melanogaster	854
Caenorhabditis elegans	569
Drosophila pseudoobscura	10

immortalized cell line

Organ	...and 175 more
K562	467 41 18 13 248 12 160 1 1 8 2 1 9 2 6 1 1 29
HepG2	185 3 11 6 229 7 128 1 2 3 2 6 2 6 1
GM12878	186 2 10 7 8 2 2 6 2 1 6 13 2 6 1 1 14
A549	195 14 21 2 1 9 2 3 1
MCF-7	117 8 5 1 7 5 7 2 3 1 6 1 1

tissue

Organ	...and 154 more
liver	149 5 14 11 10 1 1 3 2 2 1 6 5
heart	100 20 8 11 8 1 1 2 1 6 6
stomach	77 18 11 9 8 1 4 1 3 4 4
lung	80 15 8 5 10 7 3 1 1 2 4 4
kidney	69 16 9 4 2 4 4 2 3 4

primary cell

Project	...and 154 more
ENCODE	67 15 1 13 1 9
Roadmap	7065 3115

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 the ENCODE project website at encodeproject.org. The search results page displays 25 of 11736 results. On the left, there is a sidebar with filters for Assay category, Assay, Project, RFA, Experiment status, and Genome assembly. The main area lists experiment details for various targets and assays, such as RNA Bind-n-Seq experiments for RBMS3, ZFP36, and UNK targets. Each entry includes the experiment ID, target, lab, and project information.

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, target gene (SUZ12), lab (Peggy Farnham, USC), and project (ENCODE). Below this is a detailed experiment summary table with various parameters like assay type (ChIP-seq), target gene (SUZ12), and biosample type (Homo sapiens H1-hESC stem cell). The table also includes attribution information such as the lab (Peggy Farnham, USC), award PI (Michael Snyder, Stanford), and project (ENCODE). External resources like UCSC-ENCODE-hg19:wgEncodeEH001752 and GEO:GSM935352 are listed, along with the date released (2011-10-29). At the bottom, there's a section for isogenic replicates.

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 same ENCODE ChIP-seq experiment summary page as above, but with a green arrow pointing to the 'bed narrowPeak' download link. The download link is labeled 'ENCFF002CRG' with a blue download icon. Below the download links, the file details are shown: hg19 genome build, ENCODE Consortium Analysis Working Group, release date 2014-06-06, size 118 kB, and a green checkmark indicating it is released. A green button labeled 'released' is also visible.

NGSデータ高次解析の実例

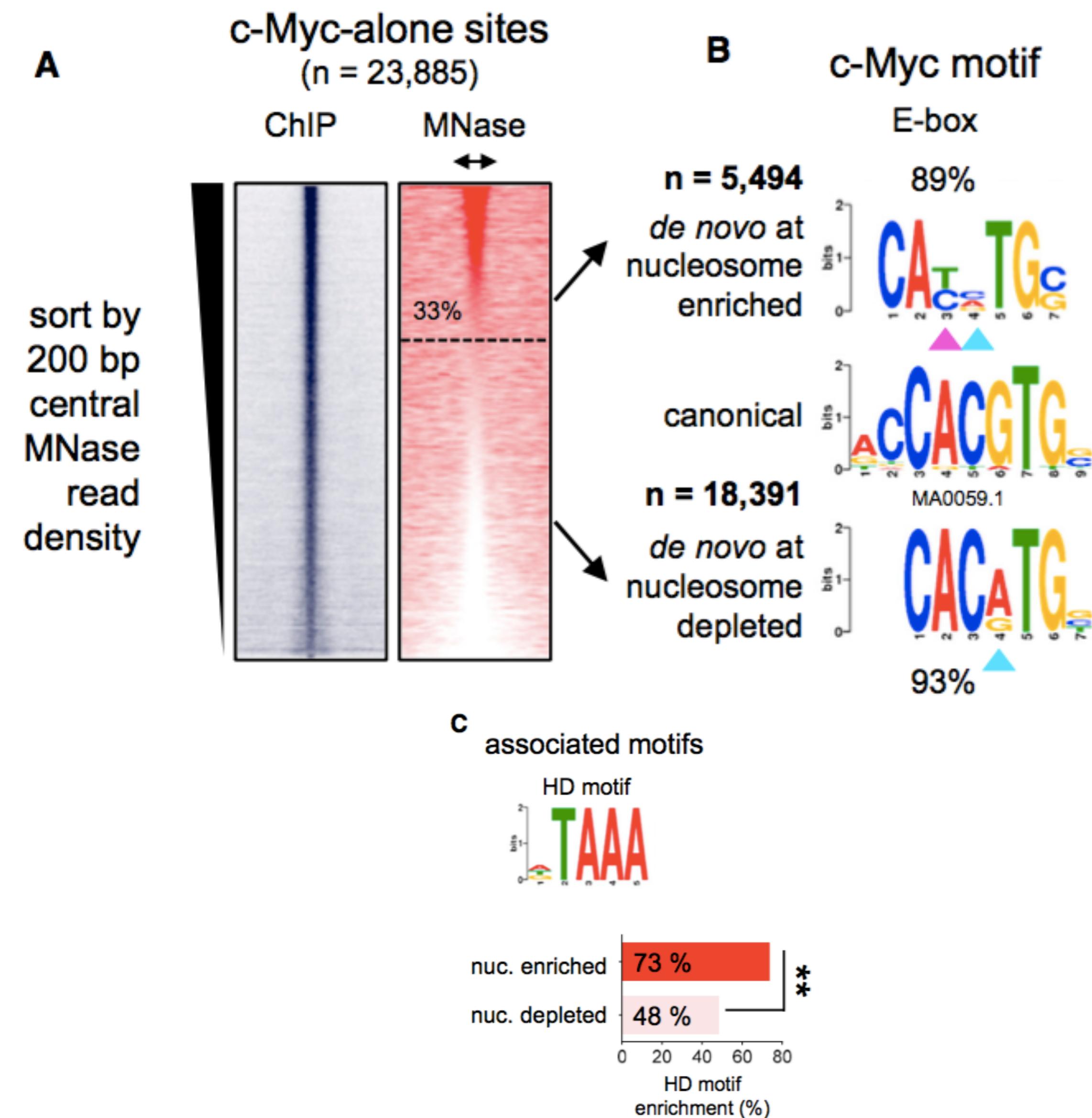
例: パイオニア転写因子の結合様式

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

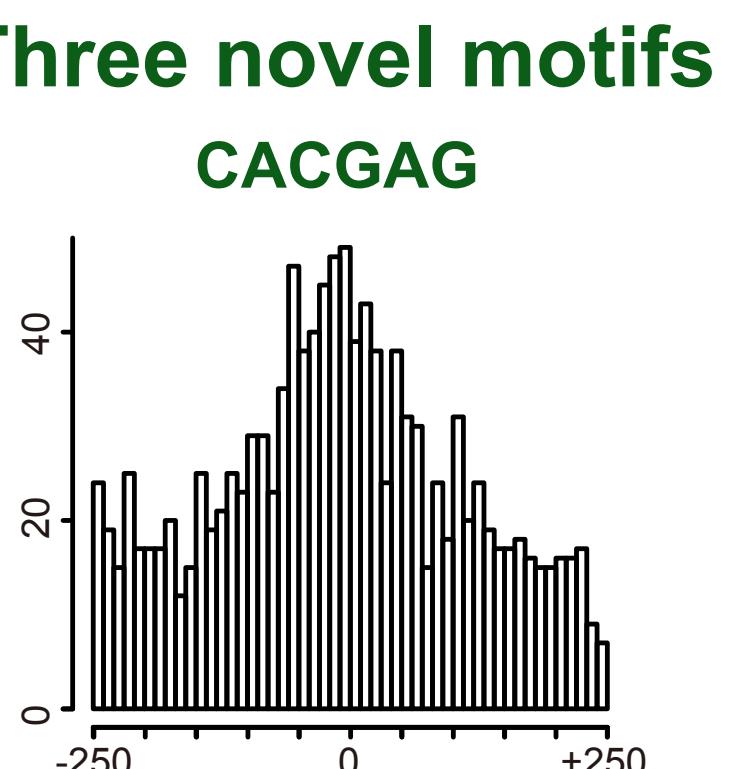
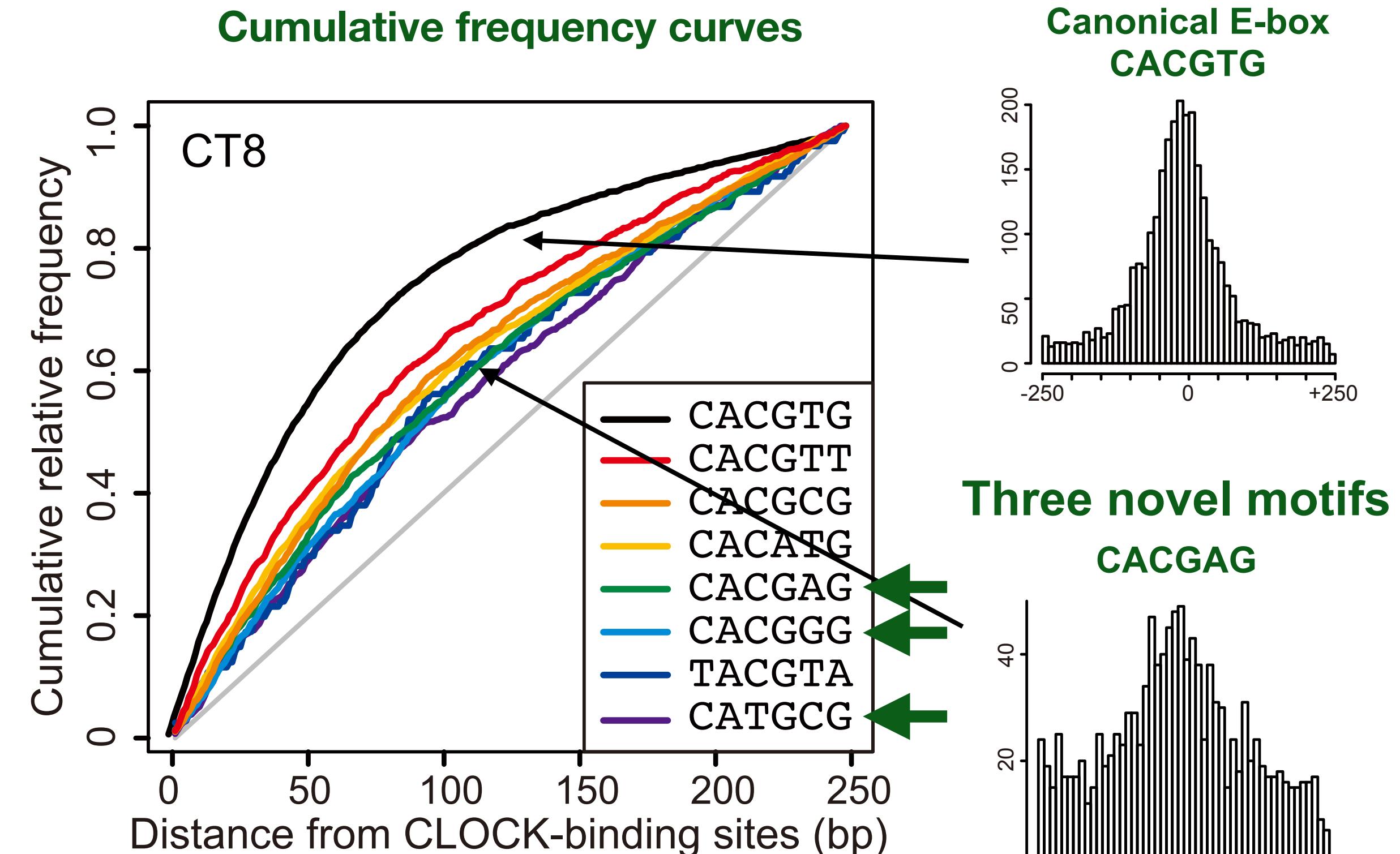
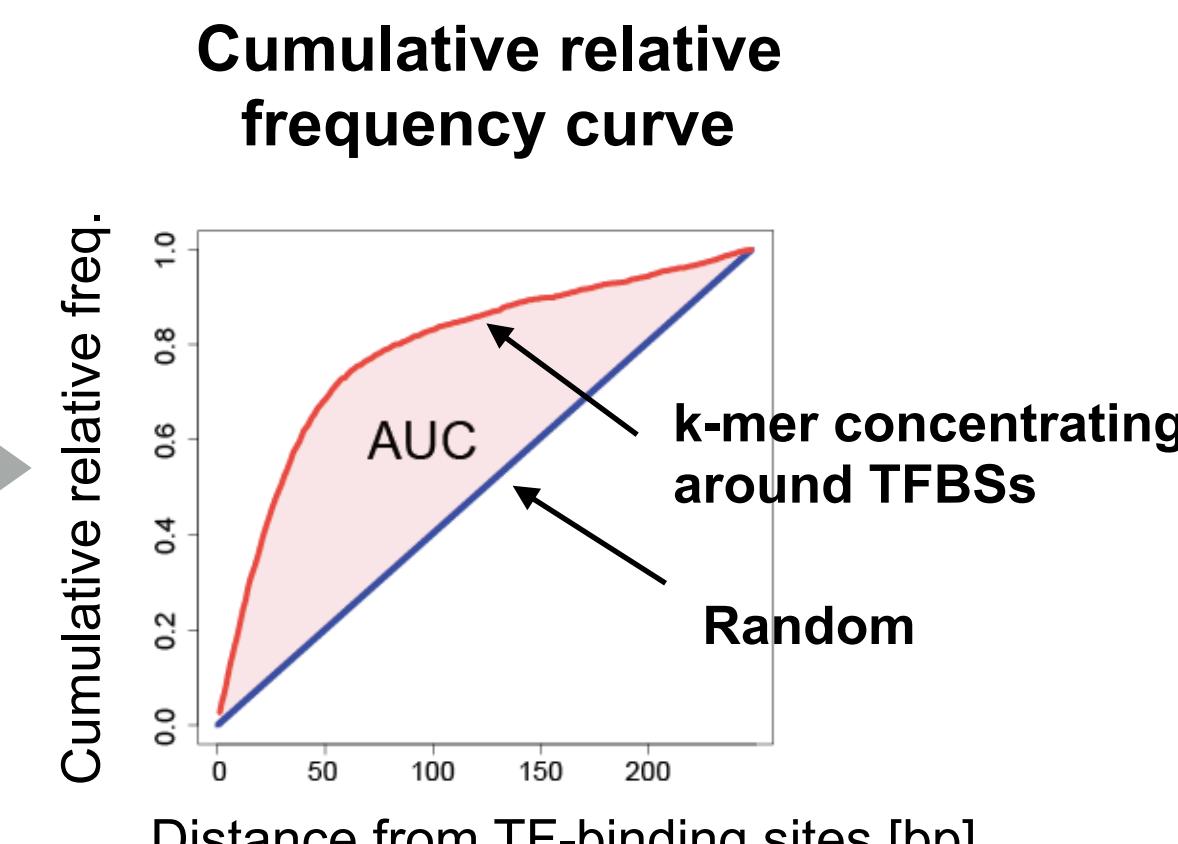
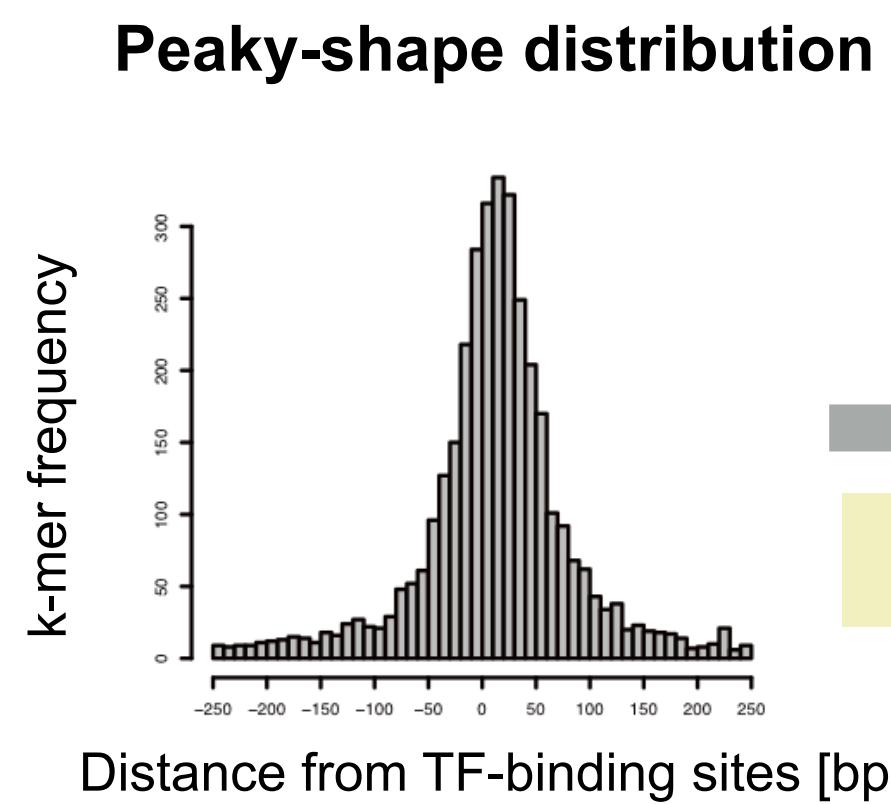


例: MOCCS (Motif Centrality Analysis of ChIP-Seq)

CLOCK ChIP-Seqデータから

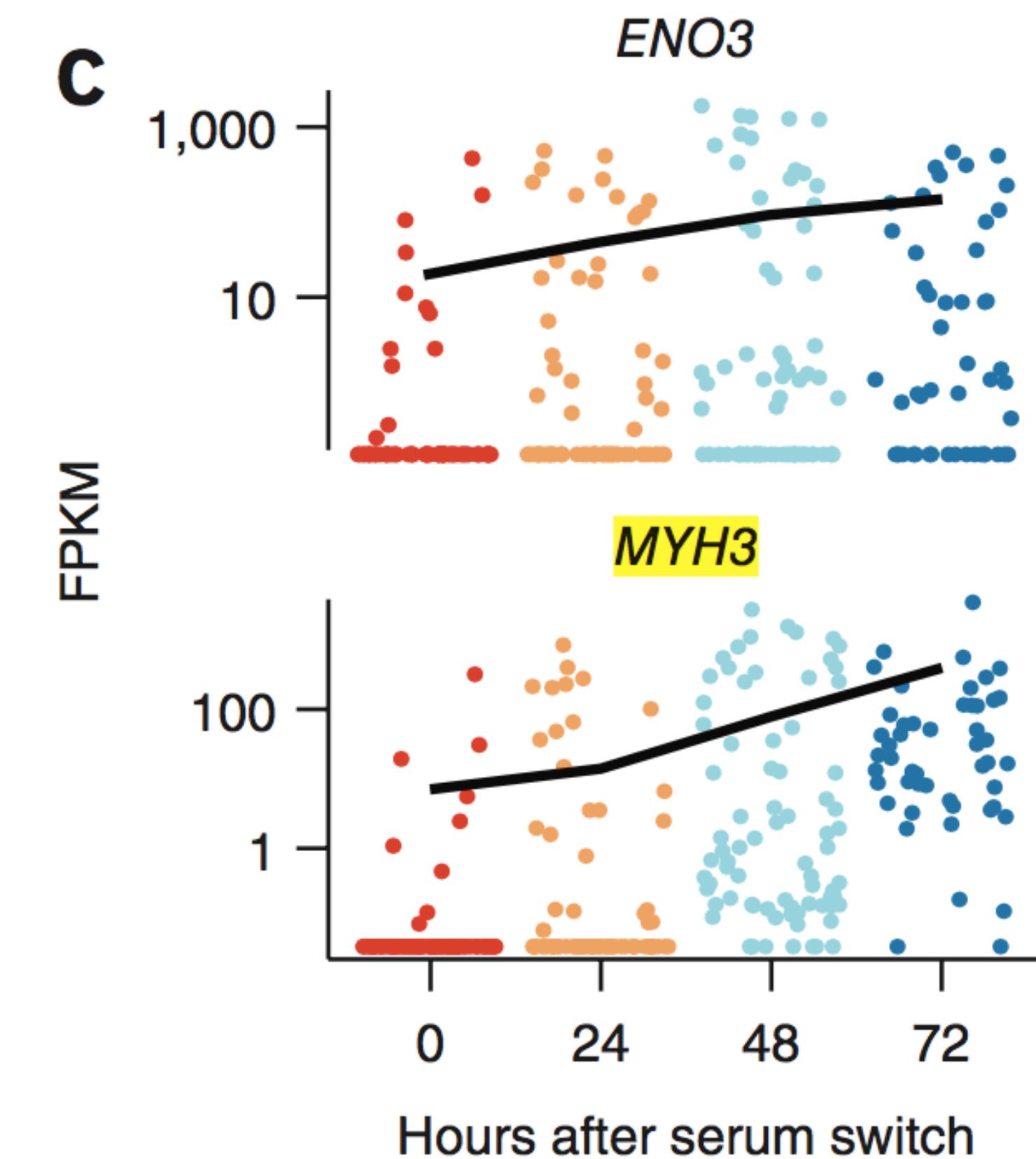
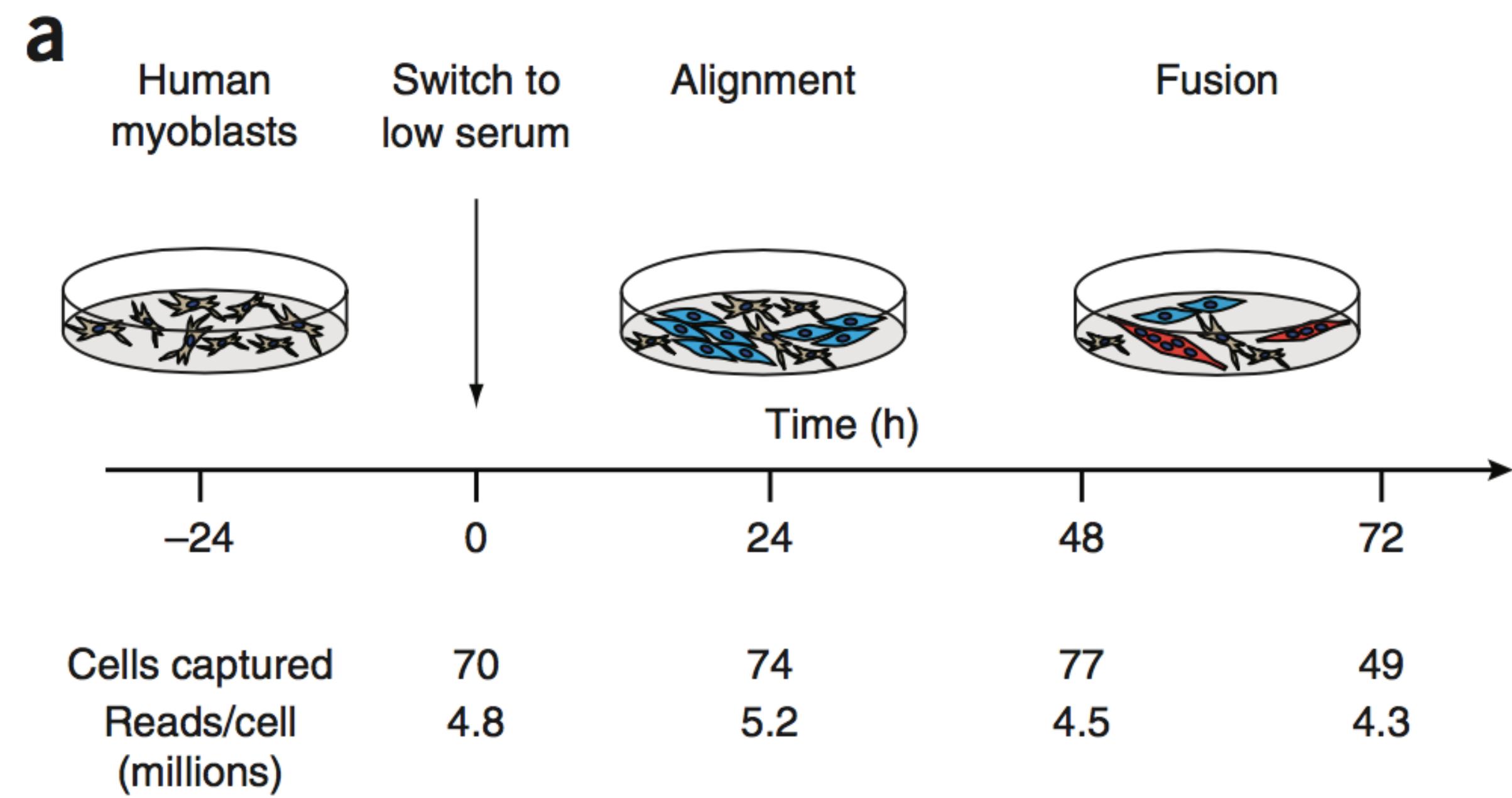
新規モチーフを発見

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



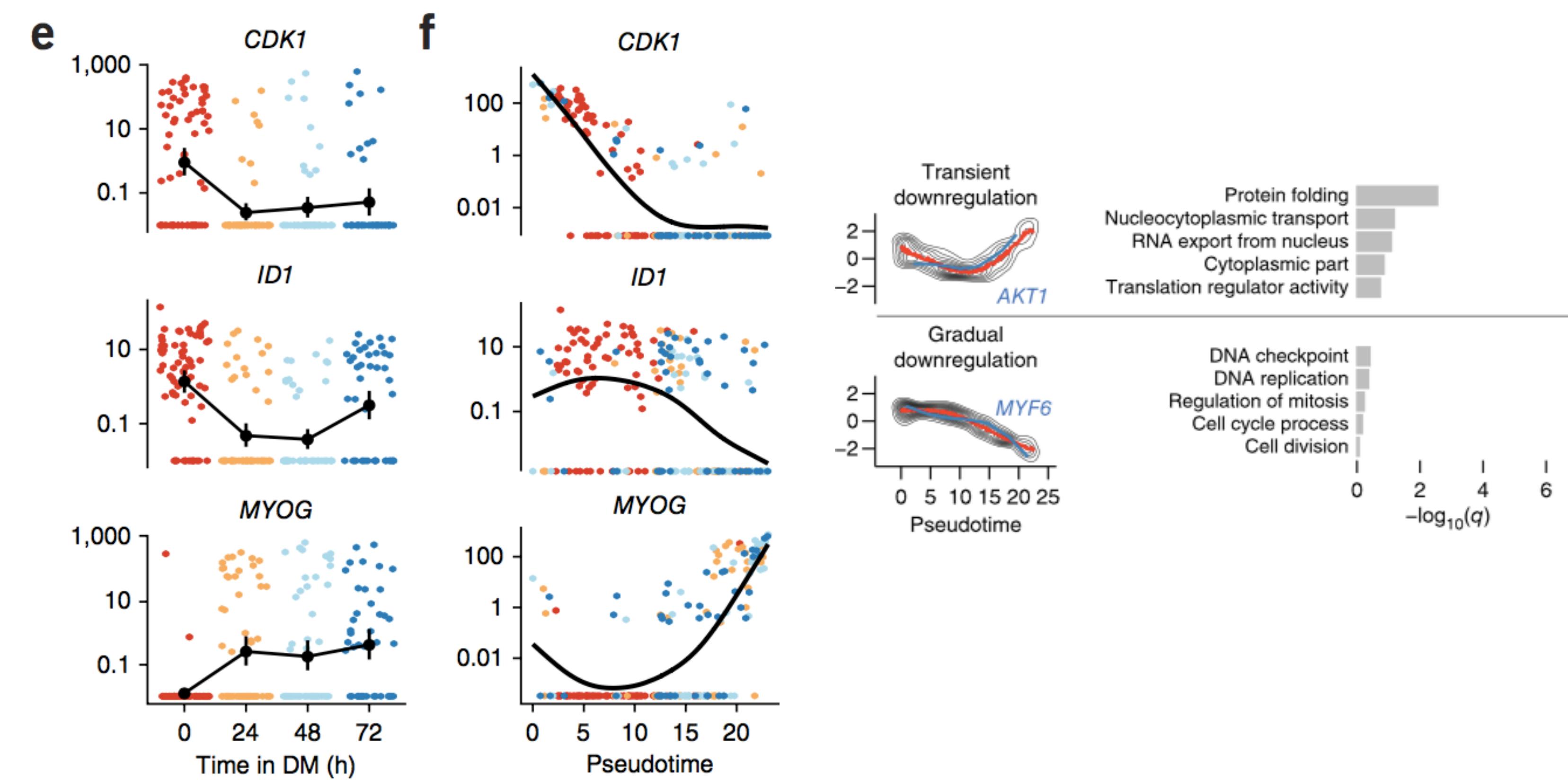
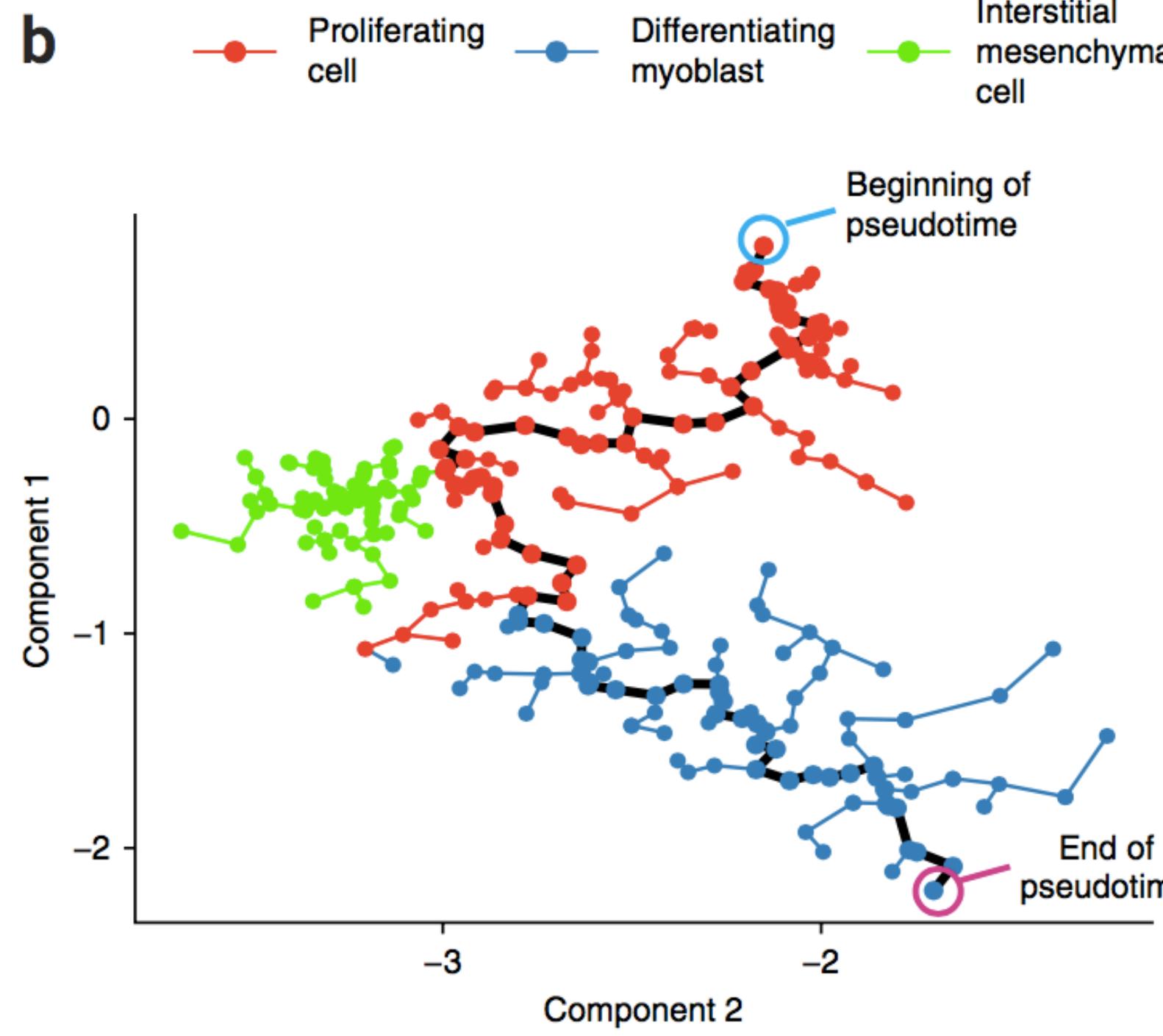
例: 擬時間推定 Pseudotime estimation

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データ高次解析の基礎

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

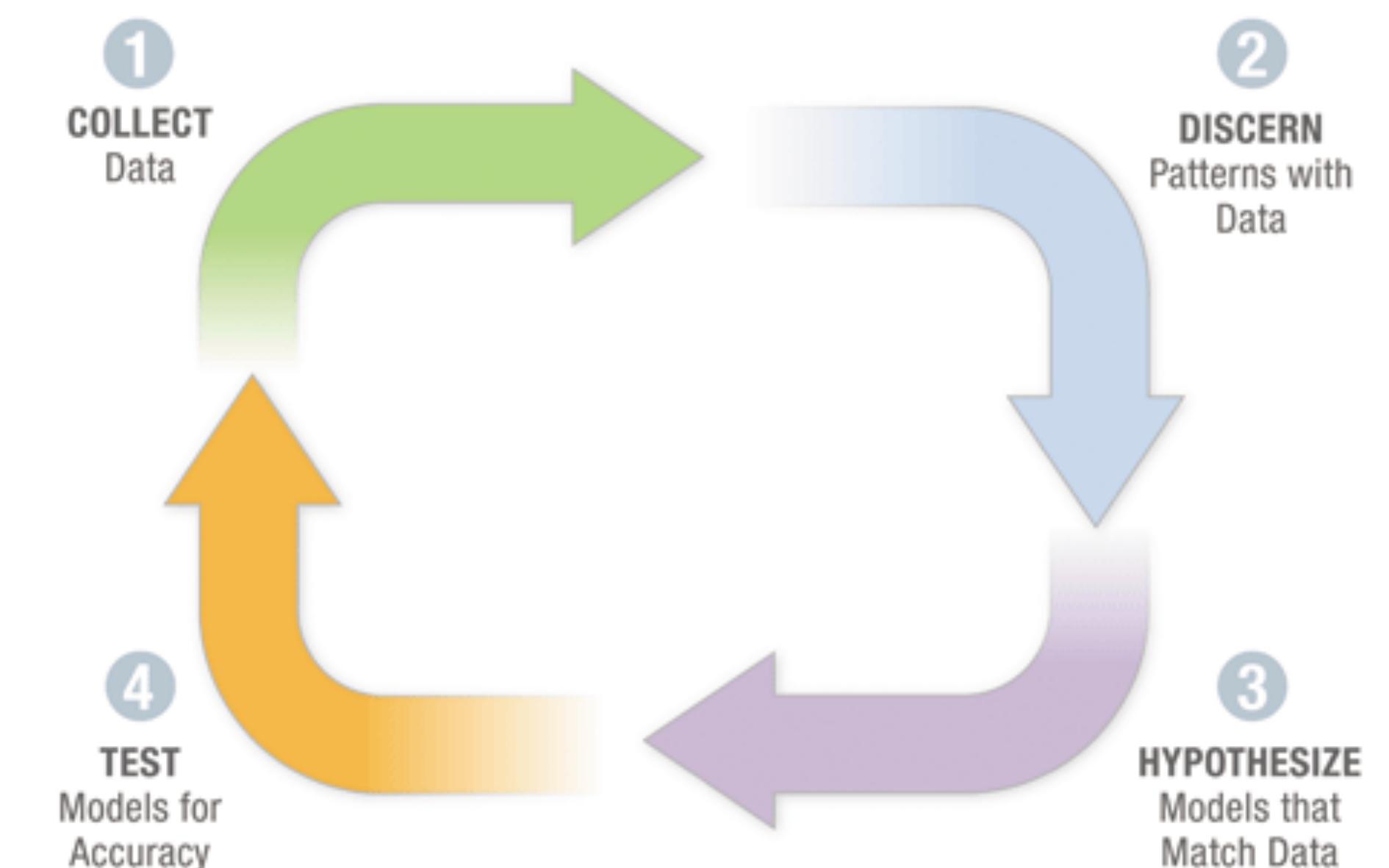
要約、可視化、次元圧縮

データを分類する

クラスタリング、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



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

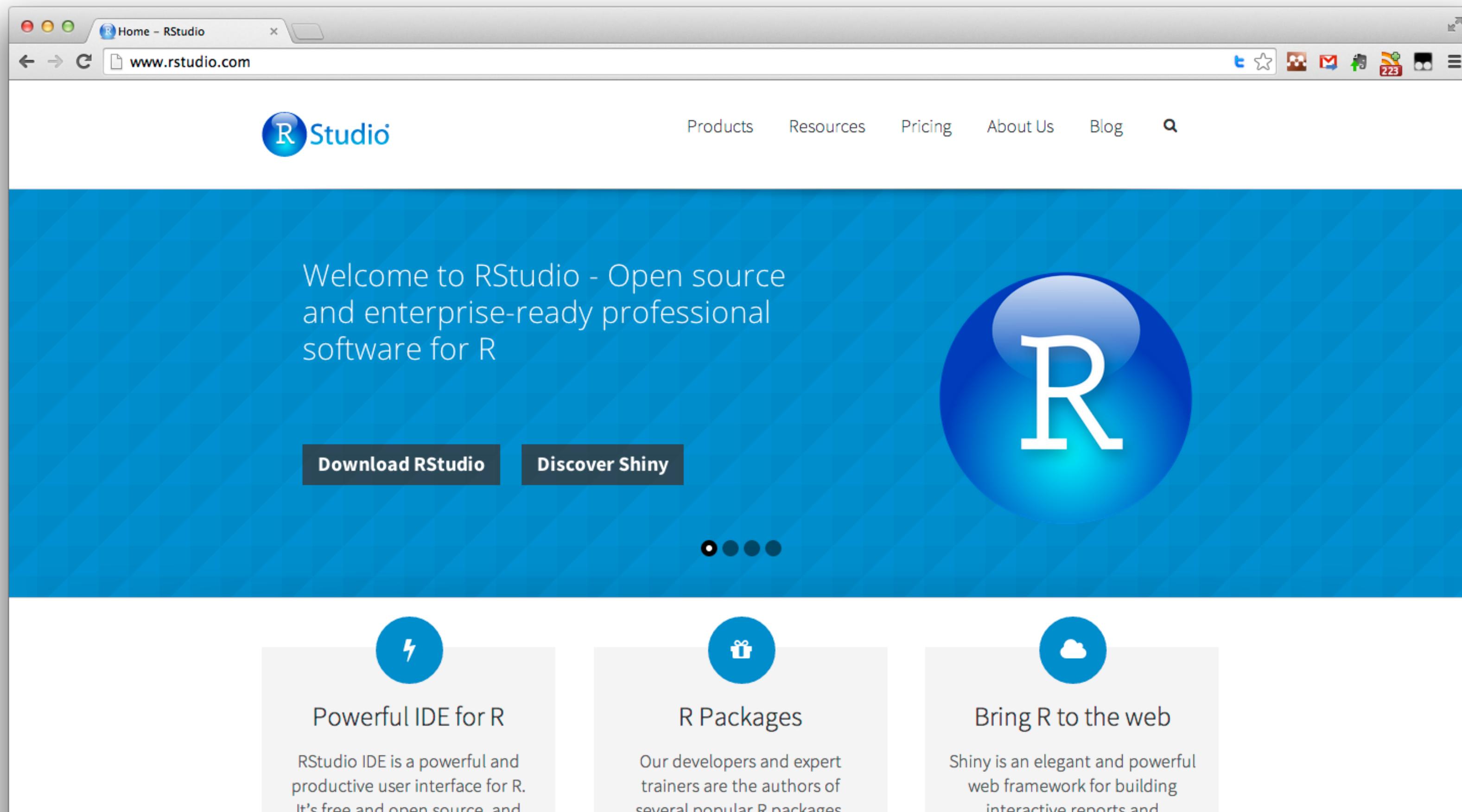
<http://bioconductor.org>



RStudio

RをGUIで使うための統合解析環境

<http://www.rstudio.com>

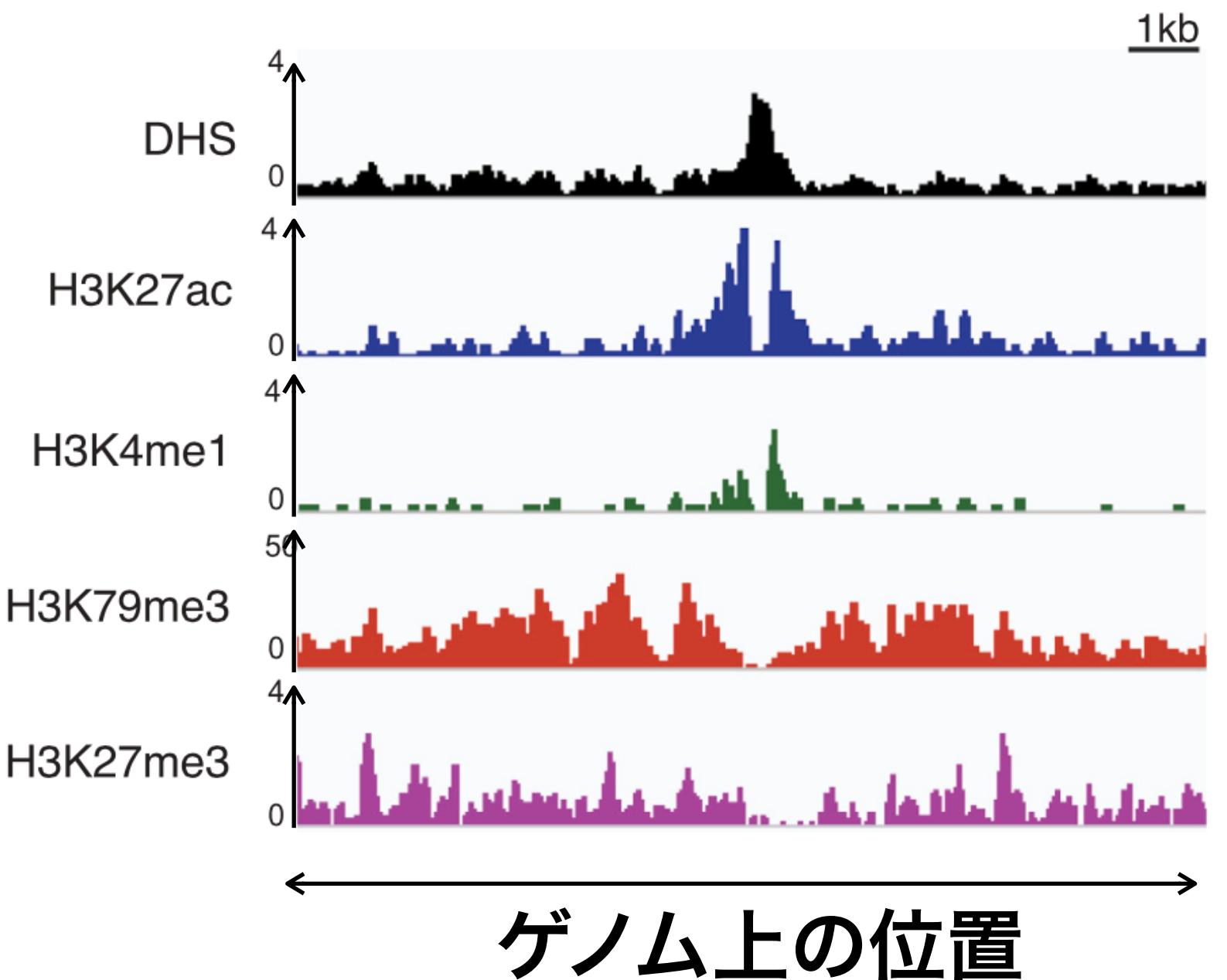


ゲノムブラウザ

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/>

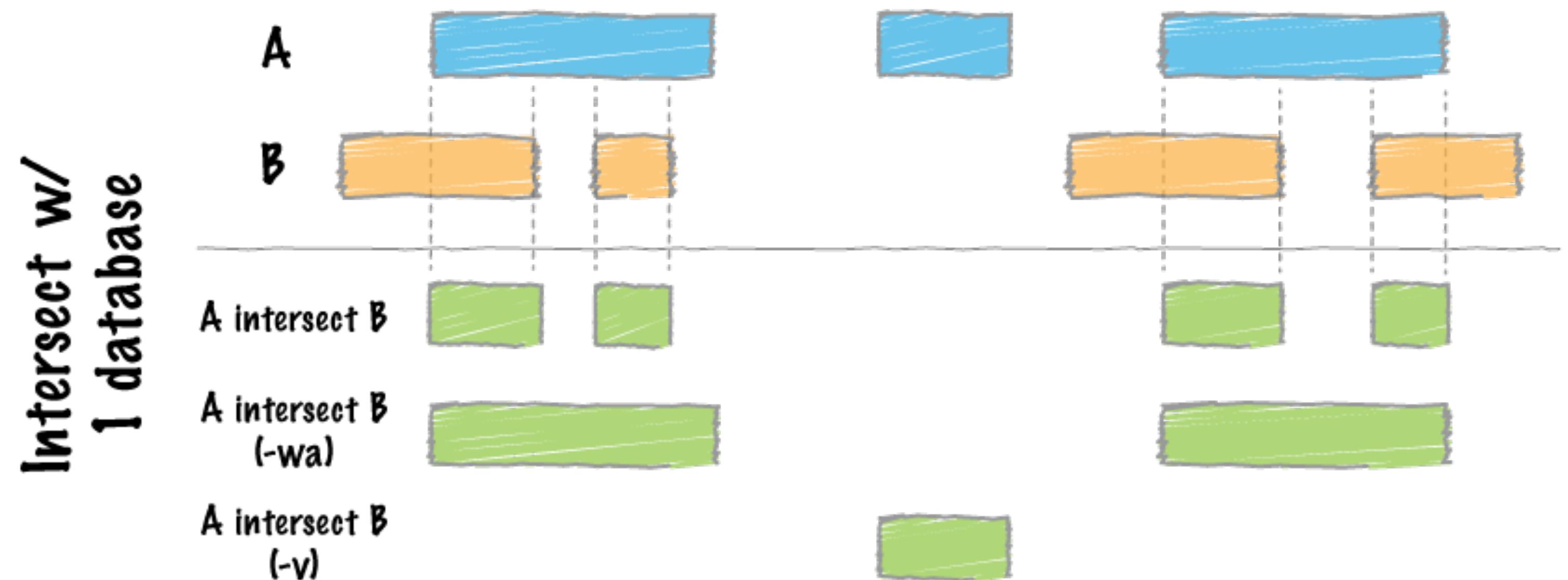
The screenshot shows the homepage of the IGV website. The header reads "Home | Integrative Genomics Viewer". The main content area features a large image of the IGV software interface, which displays multiple tracks of genomic data. Below this, there are sections for "What's New" (mentioning the iPad app release in September 2014), "Citing IGV" (with a citation to a 2012 paper in Briefings in Bioinformatics), and "Overview" and "Funding" sections. The left sidebar contains links for Home, Downloads, Documents, Hosted Genomes, FAQ, IGV User Guide, File Formats, Release Notes, IGV for iPad, Credits, Contact, and a search bar. The bottom left corner includes the Broad Institute logo and copyright information.



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 header navigation bar containing links for Home, Download, Workflows, Documentation, and Support. The main content area features a large title "Samtools" and a brief description of the suite's purpose and components: Samtools, BCFtools, and HTSlib. Below this, a note states that Samtools and BCFtools use HTSlib internally. The page is divided into four main sections: "Download", "Workflows", "Documentation", and "Support".

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 htlib 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

ngs.plot

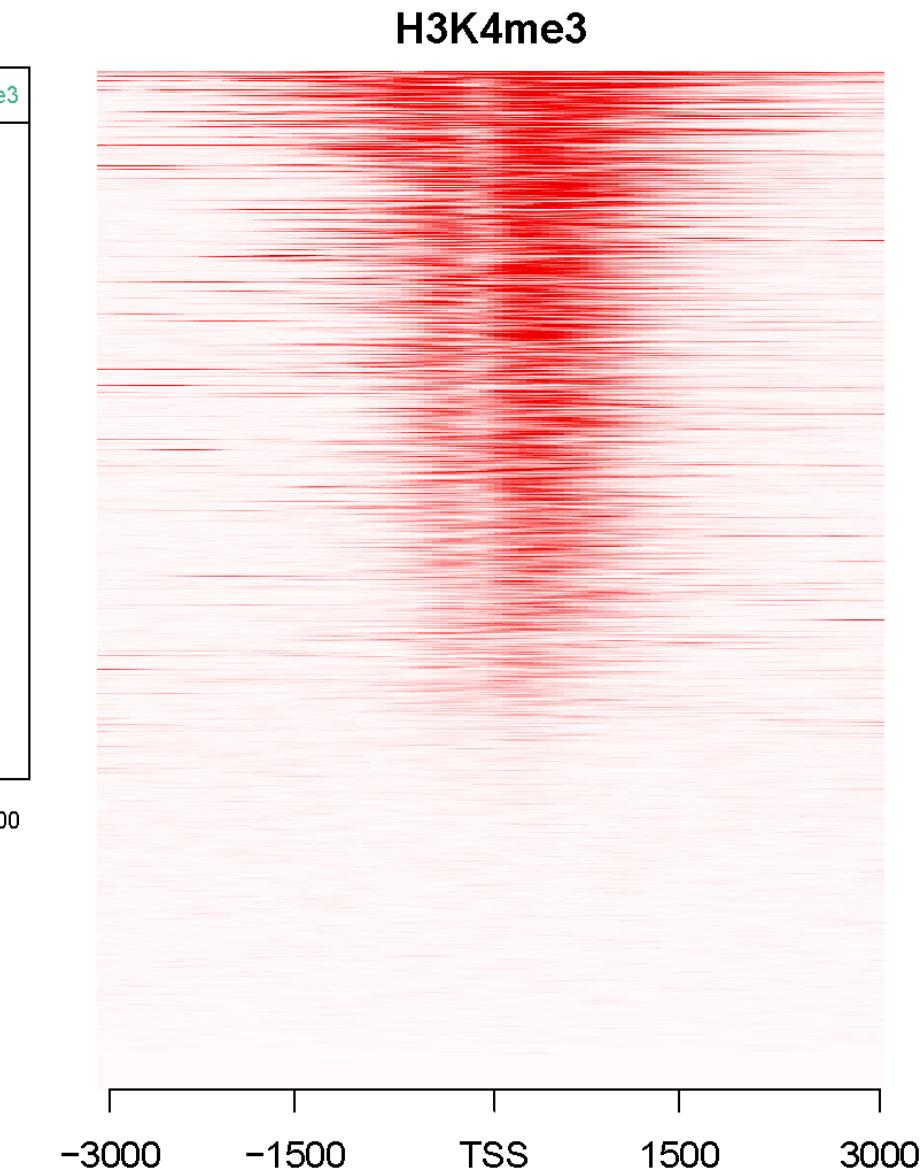
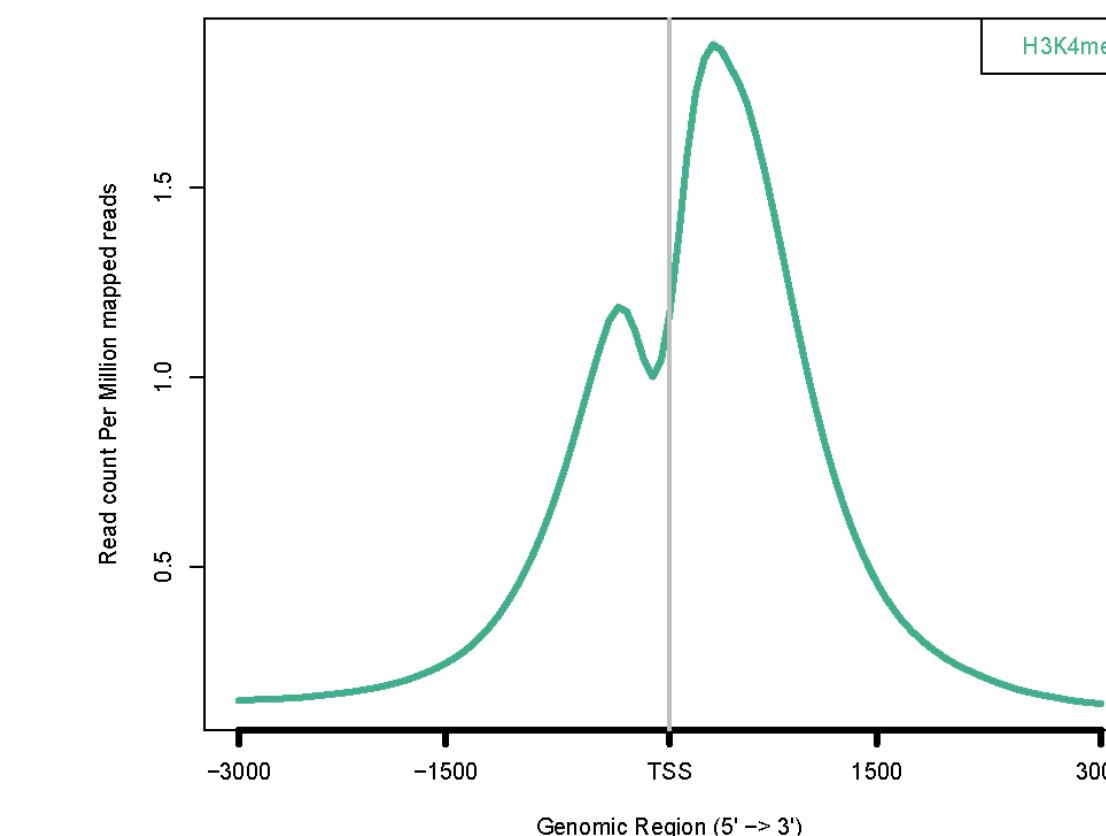
機能ゲノミクス関連のプロット

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

Aggregation plot

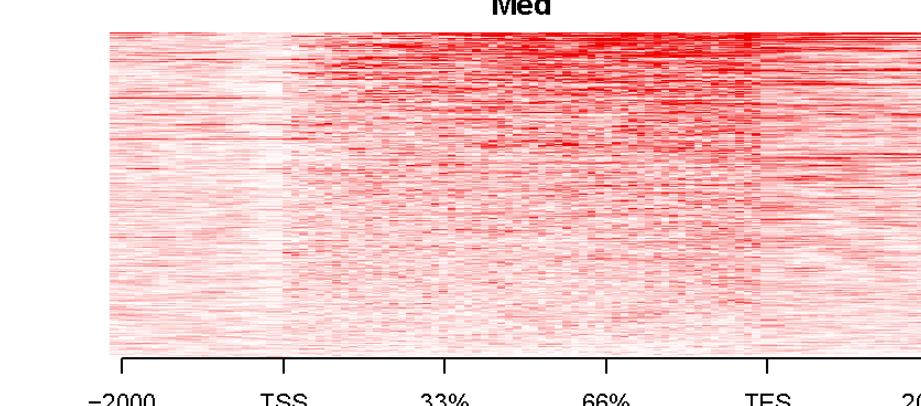
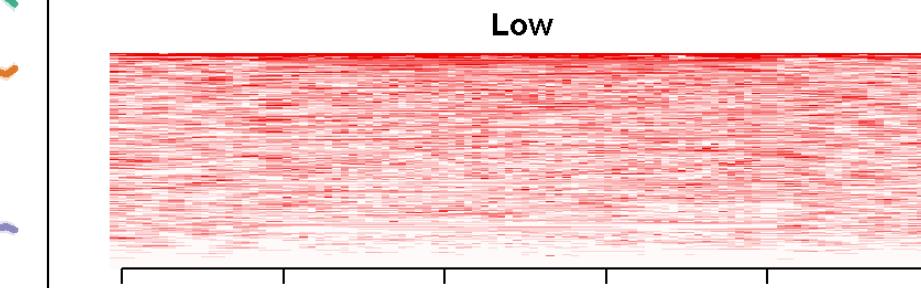
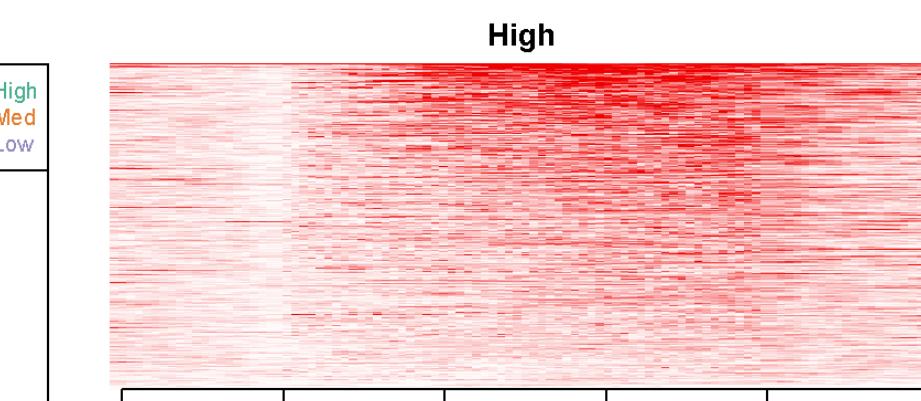
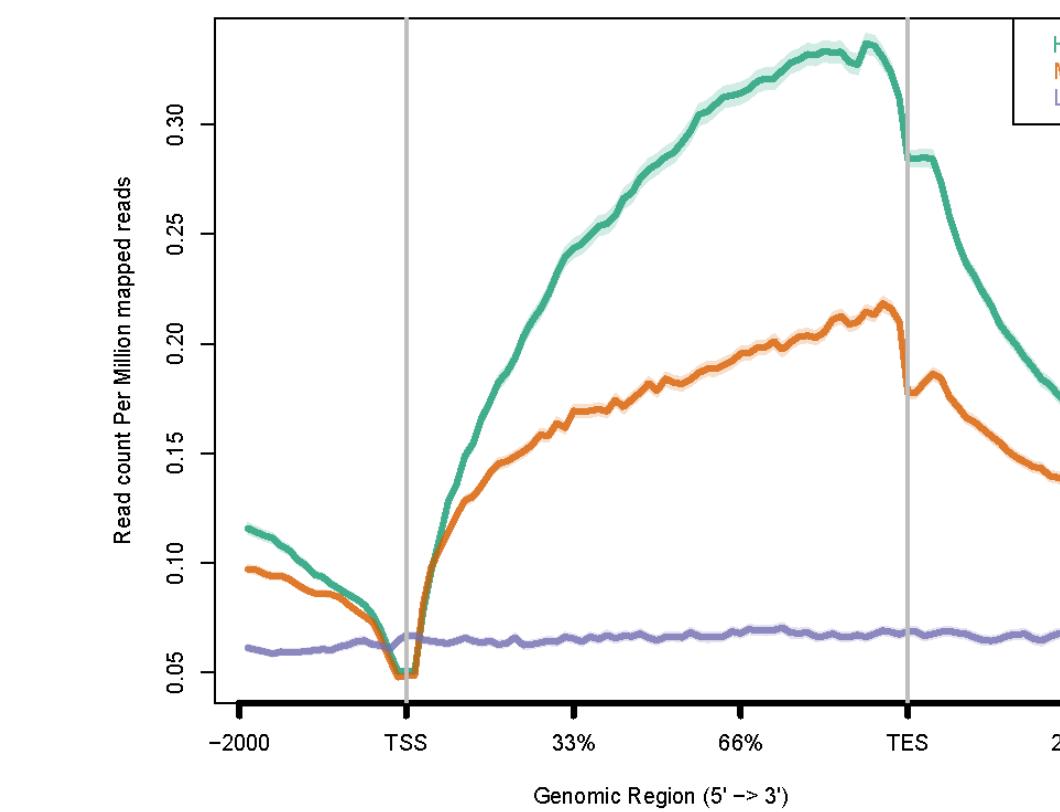
ゲノム上の点の集合（例: TSSs）に対する
NGSリードの分布

Aggregation plot



Meta-gene plot

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



Heatmap

<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

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

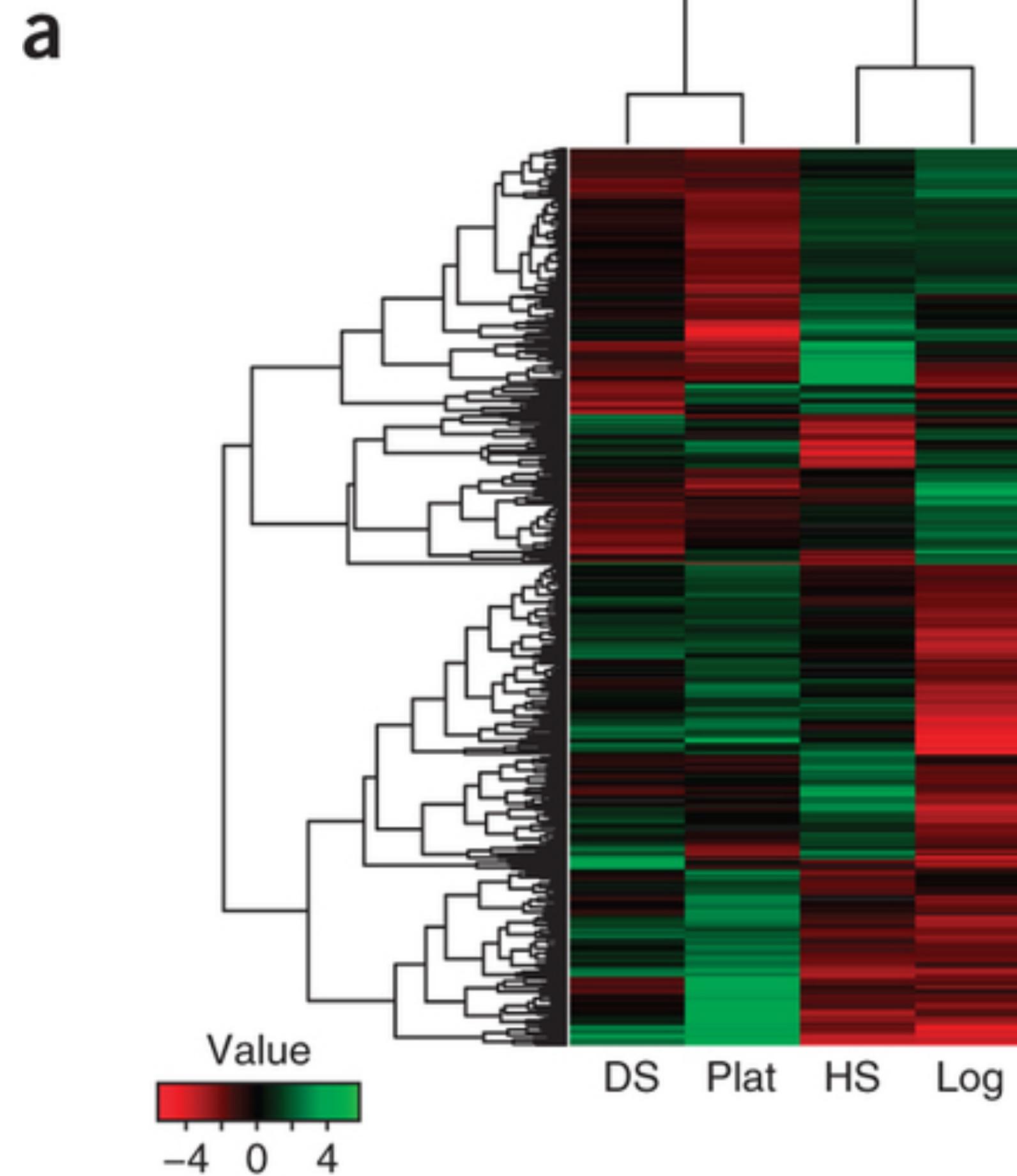
Spectral clustering

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

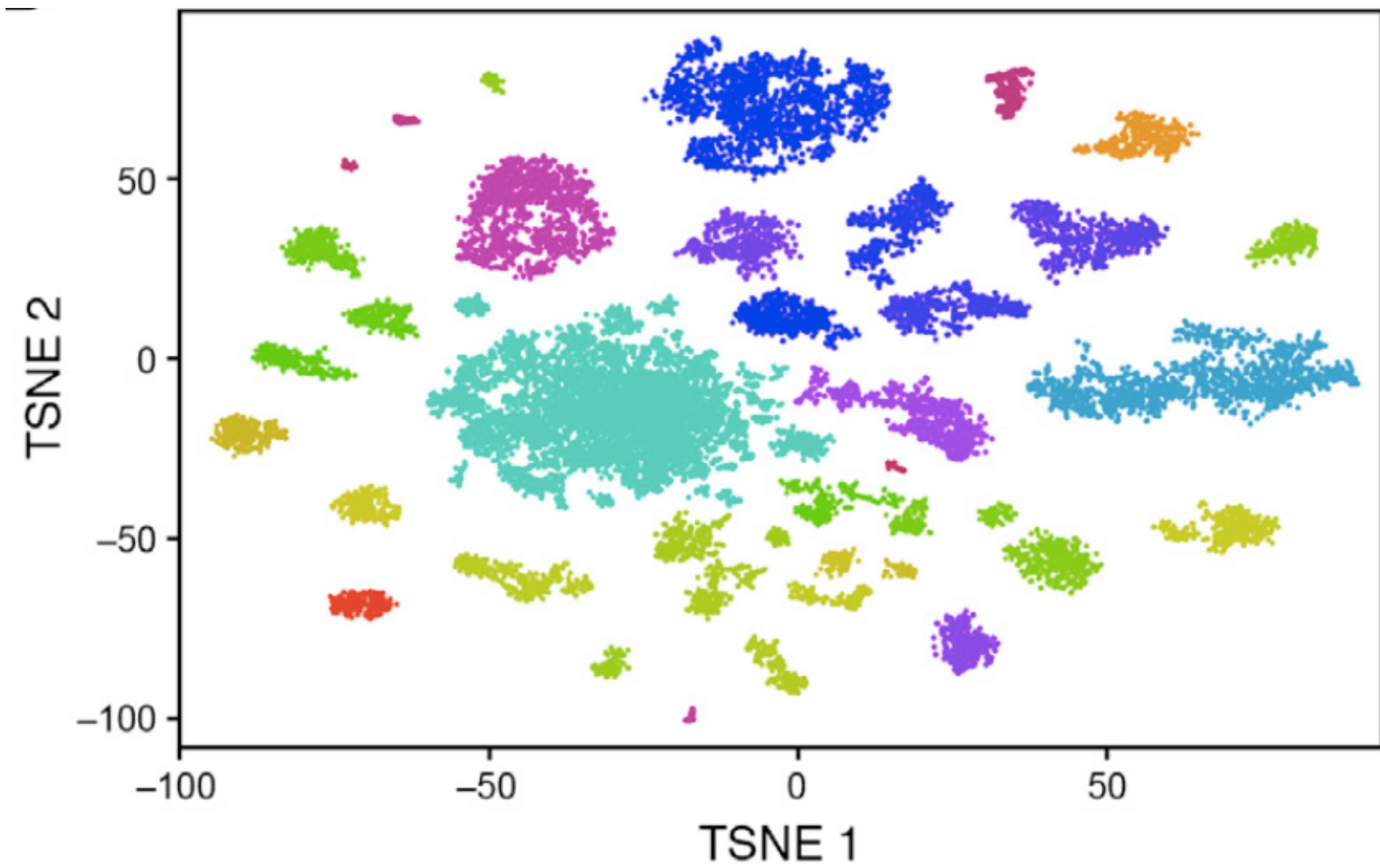
PCA、ICAなど

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

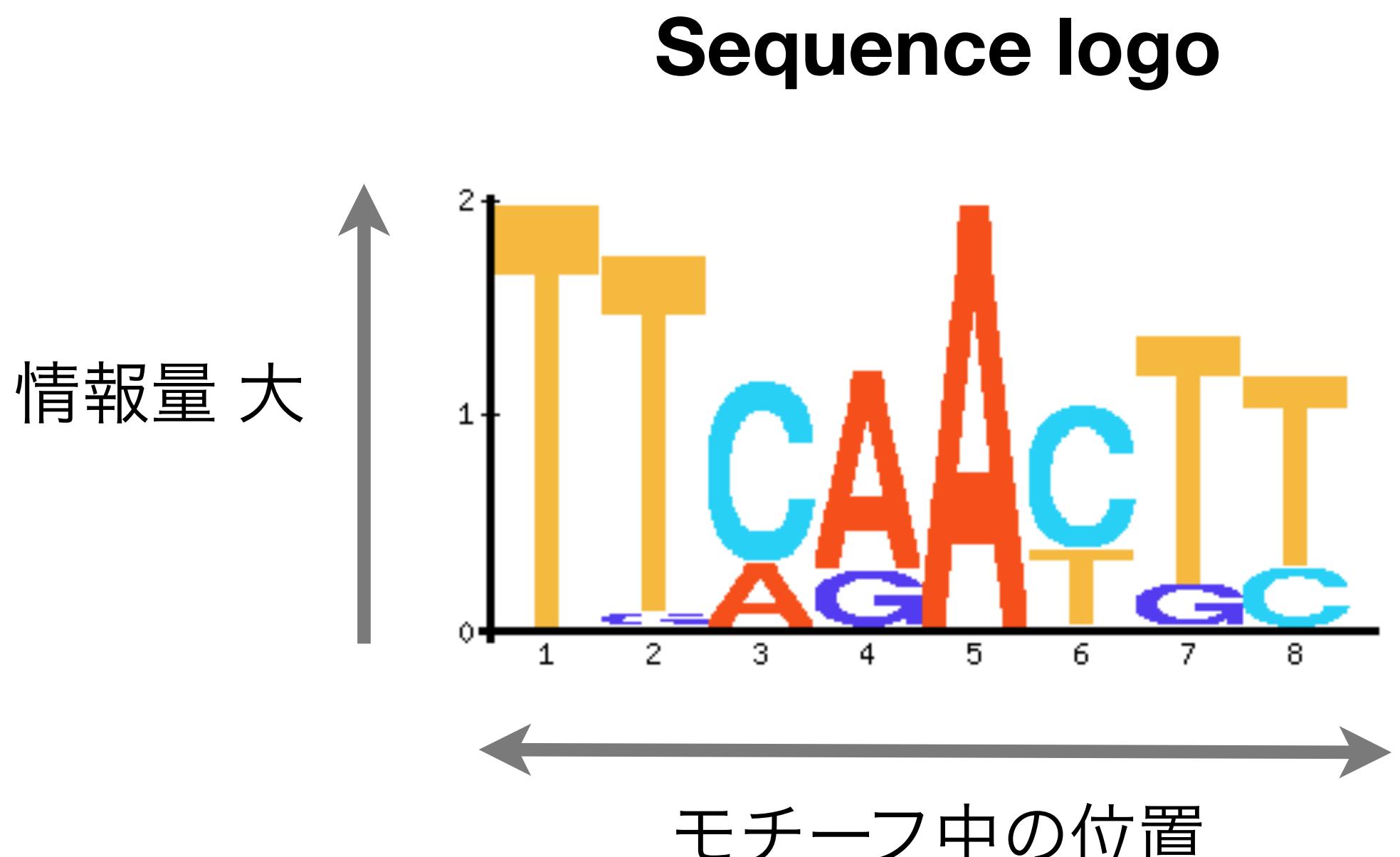
<http://meme-suite.org/doc/dreme.html>

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

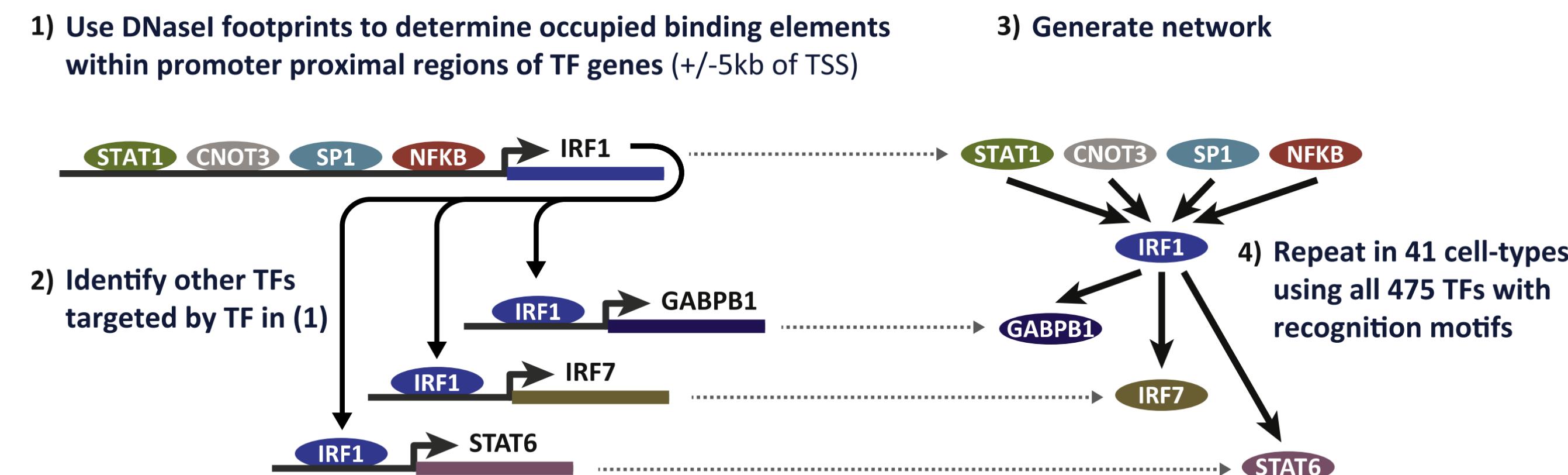
転写因子ネットワーク

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_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

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

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～

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

