NGSハンズオンセミナー

ChIP-seqの基礎実習 (初級)

理化学研究所 情報基盤センター
森岡勝樹
msmorioka-tky@umin.ac.jp

はじめに

本実習の達成目標は、NGS解析の初心者がNGSデータの一つであるChIP-seqデータに触れ、ChIP-seq解析の流れを概要として掴み、自ら解析するときの足がかりとなることを目指します。

本実習で利用する方法は、一般的に利用される方法ではありますが、Biolinux8という限られた環境で行うという性質上、あくまでも“練習”であり、自分で実際に臨むときは最新のバージョンやさらに良いソフトウェアを利用することをオススメします。

ppt講義中にデータをDLしてもらいます

指示するまでDLしないでください

DLサイト1: http://tinyurl.com/npewxs7
DLサイト2: http://tinyurl.com/pomlke3

(このパワーポイントも入っています)

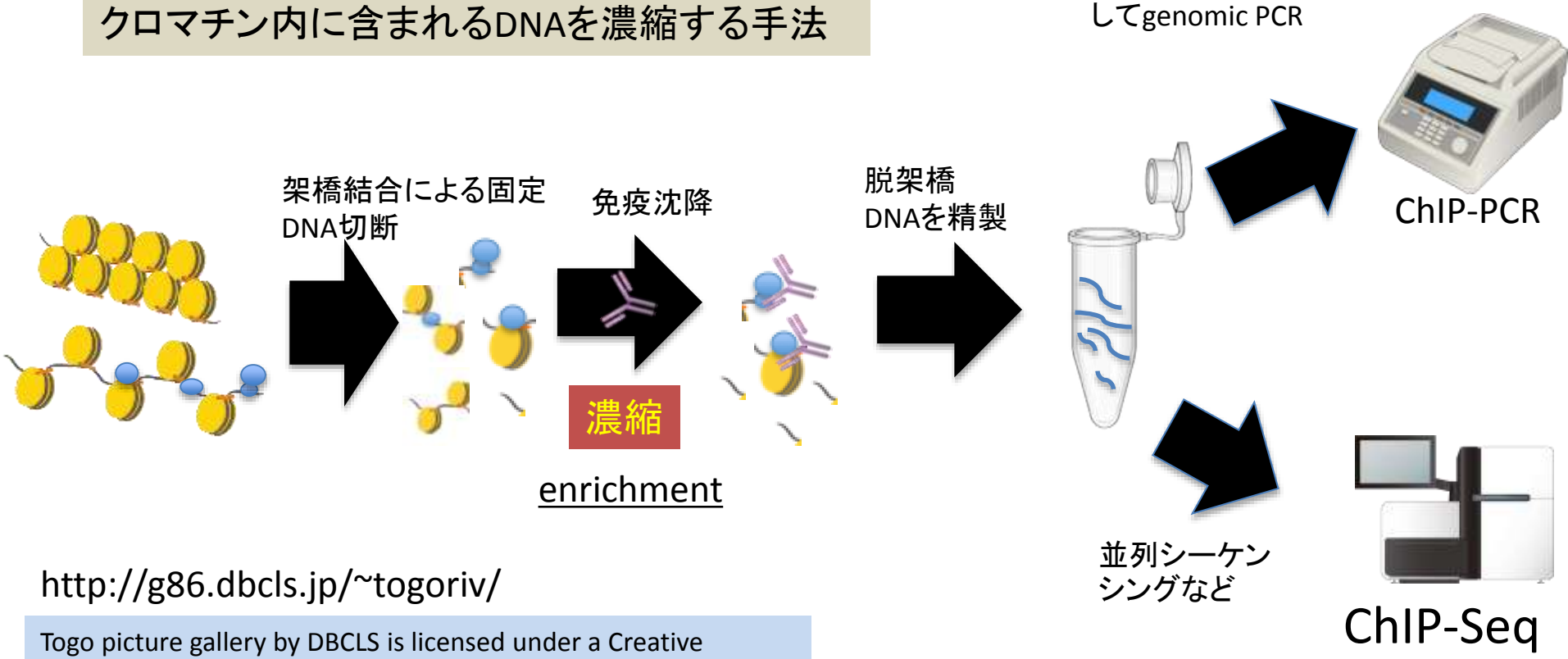
実習資料 (コピペ用に開いておく)

https://github.com/suimye/NGS_handson2015

ChIP-Seqとは

ChIP: chromatin Immuno precipitaton

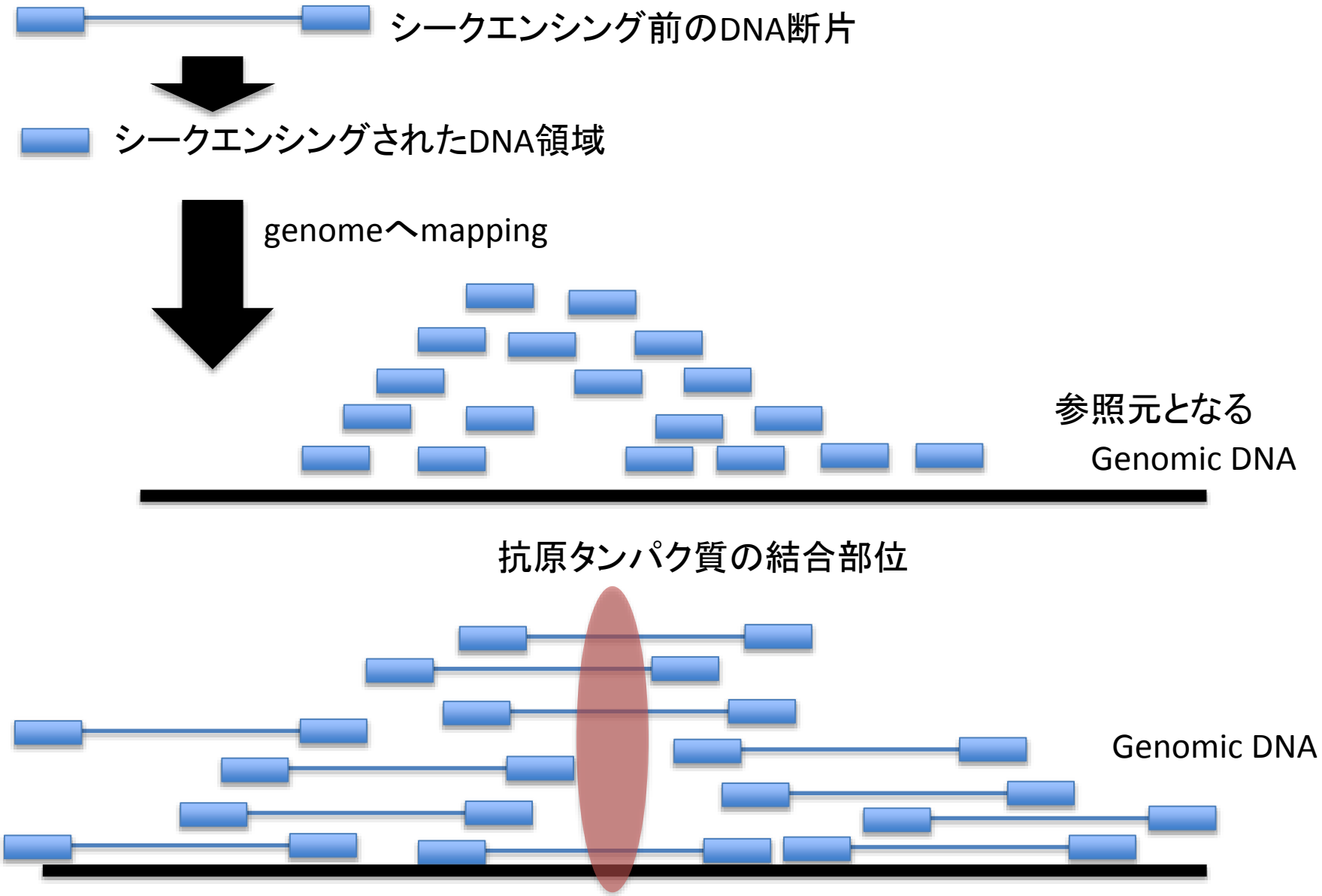
抗原抗体反応を利用して、抗原タンパク質が結合しているクロマチン構造を免疫沈降させ、クロマチン内に含まれるDNAを濃縮する手法



http://g86.dbcls.jp/~togoriv/

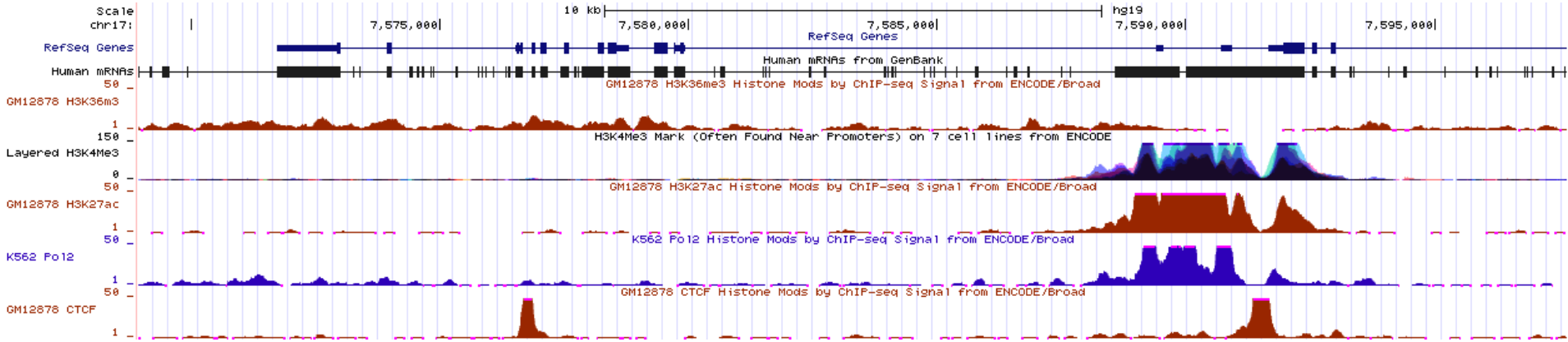
ChIP-Seqとは

注意: あくまでも模式図



ChIP-seqで得られるデータの例

TP53の遺伝子領域



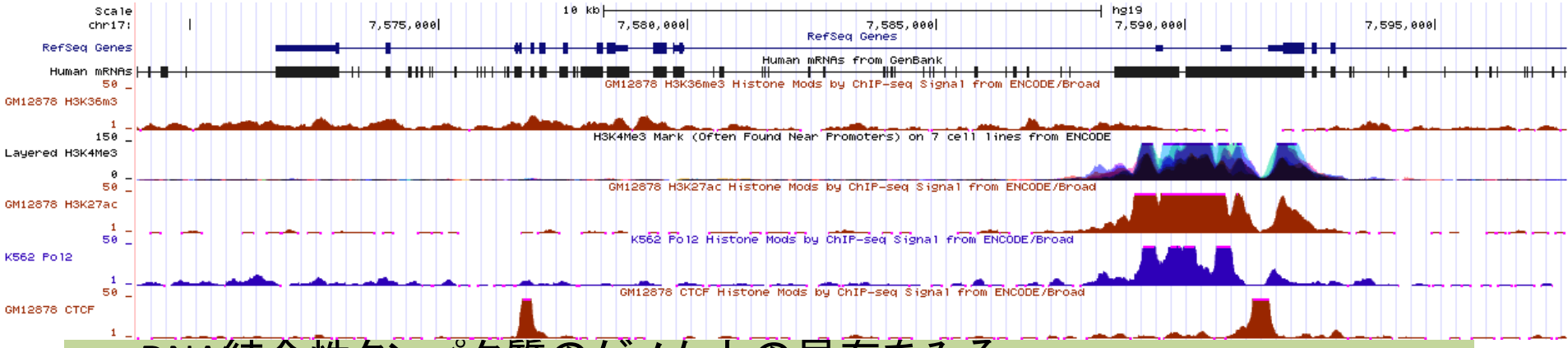
400bpぐらいのシャープな山から1000bpを越えるブロードな山まで、抗体の特性によって様々な「山」が得られる

良い例(figure)の一つとして

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

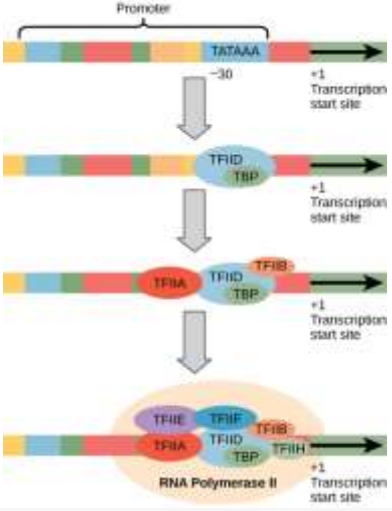
ChIP-seqの目的

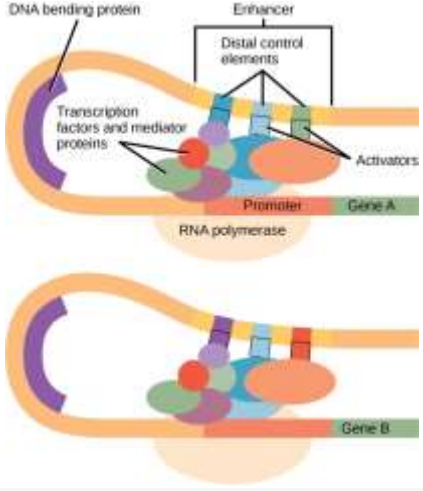
TP53の遺伝子領域



- DNA結合性タンパク質のゲノム上の局在をみる

- 転写因子
- ヒストン

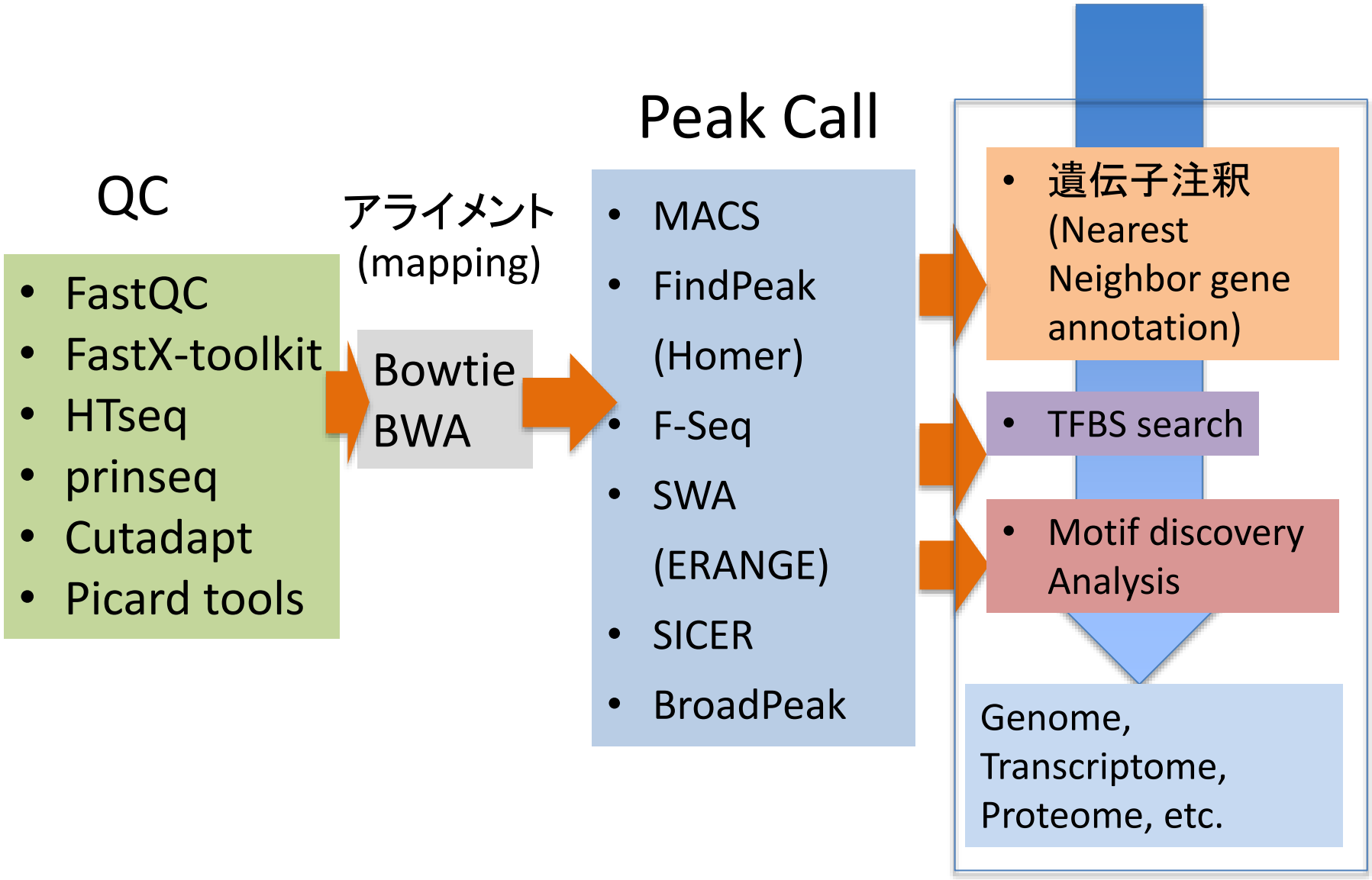




Enhancers

View on Boundless.com

ChIP-Seqデータ解析の流れ



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

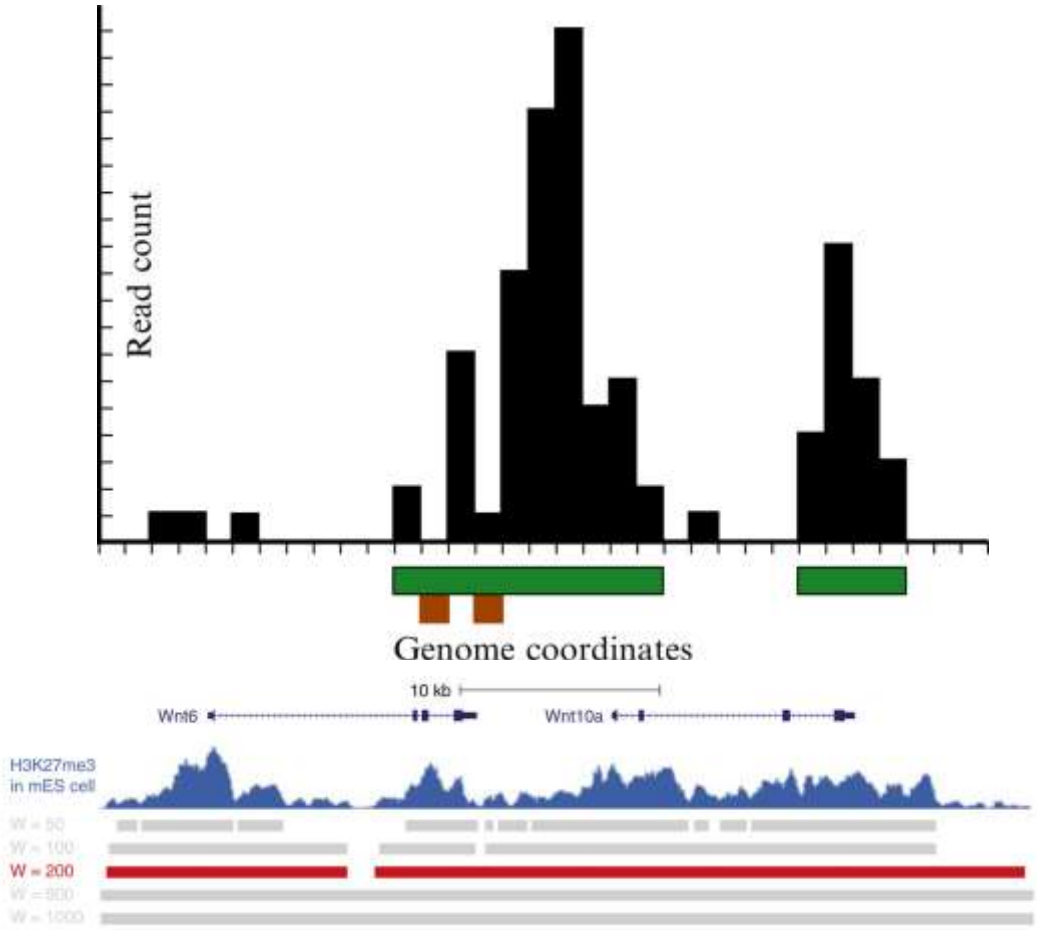
S/N is non-good or low peak (FAIRE-Seq, DNase-Seq)

F-Seq, ZIMBA

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…)

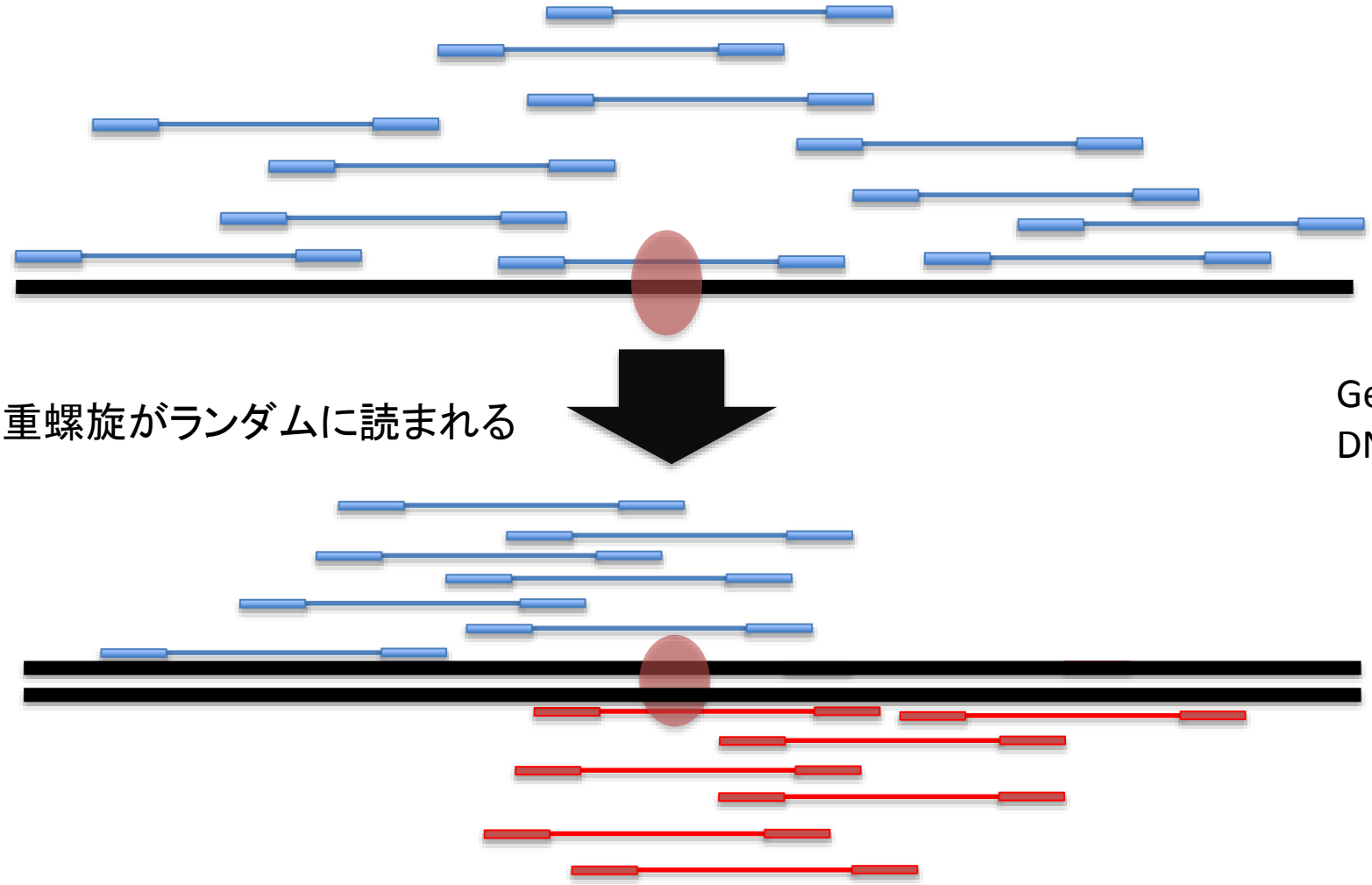


Shiliyang Xu et al. *Methods Mol Biol*. 2014 ; 1150: 97–111. doi:10.1007/978-1-4939-0512-6_5.

2. MACS

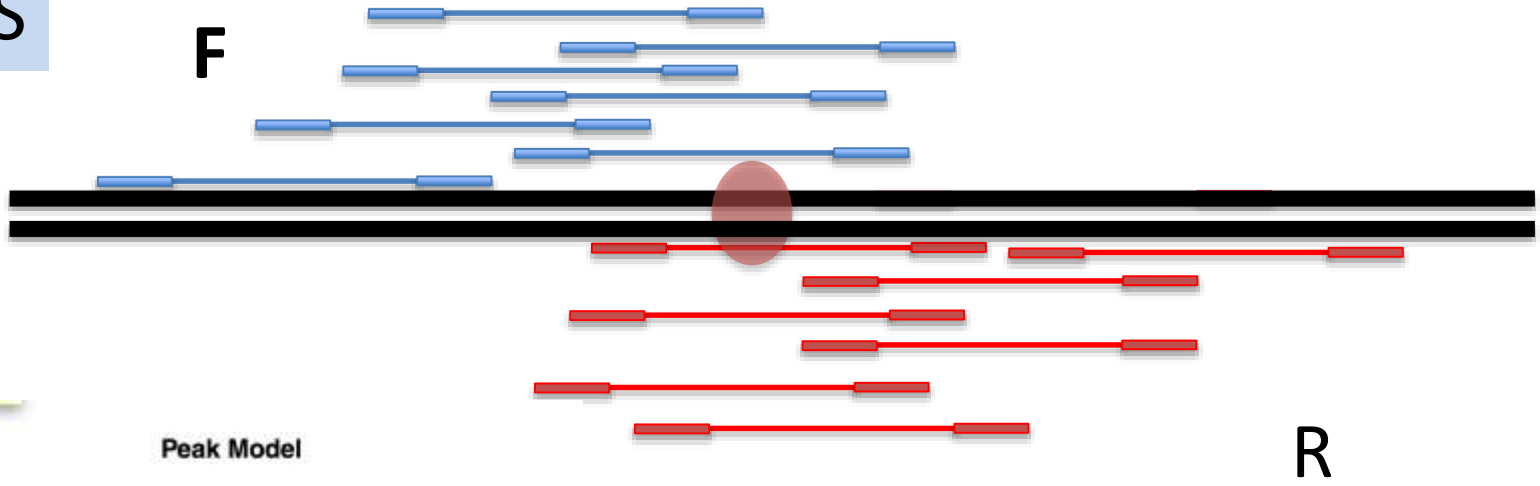
MACS

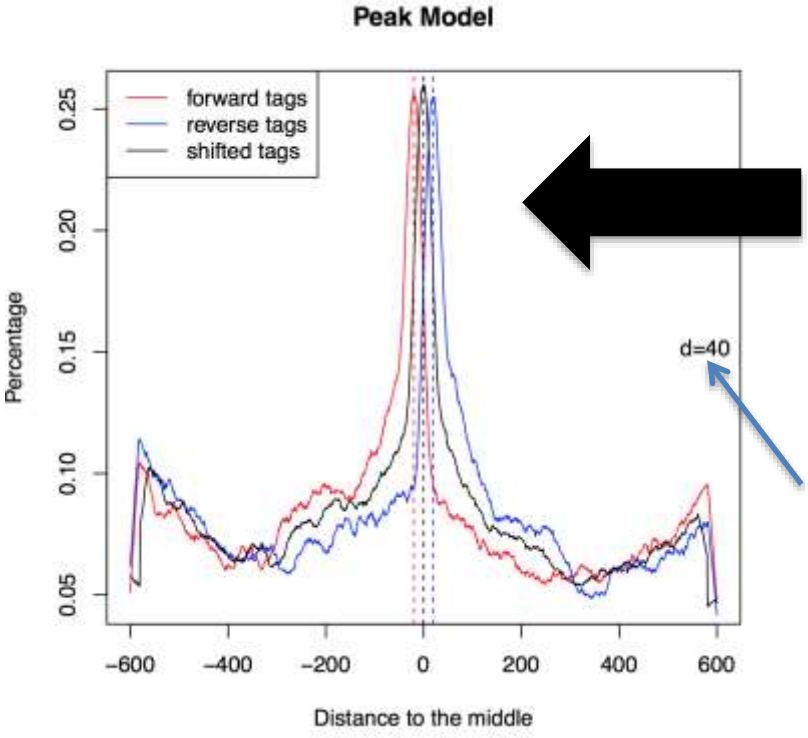
Sequencingによって得られるタグの分布形状を加味したモデルに基づいて、Tag-shiftを行いPeak Callする。



2. Tag shift

MACS

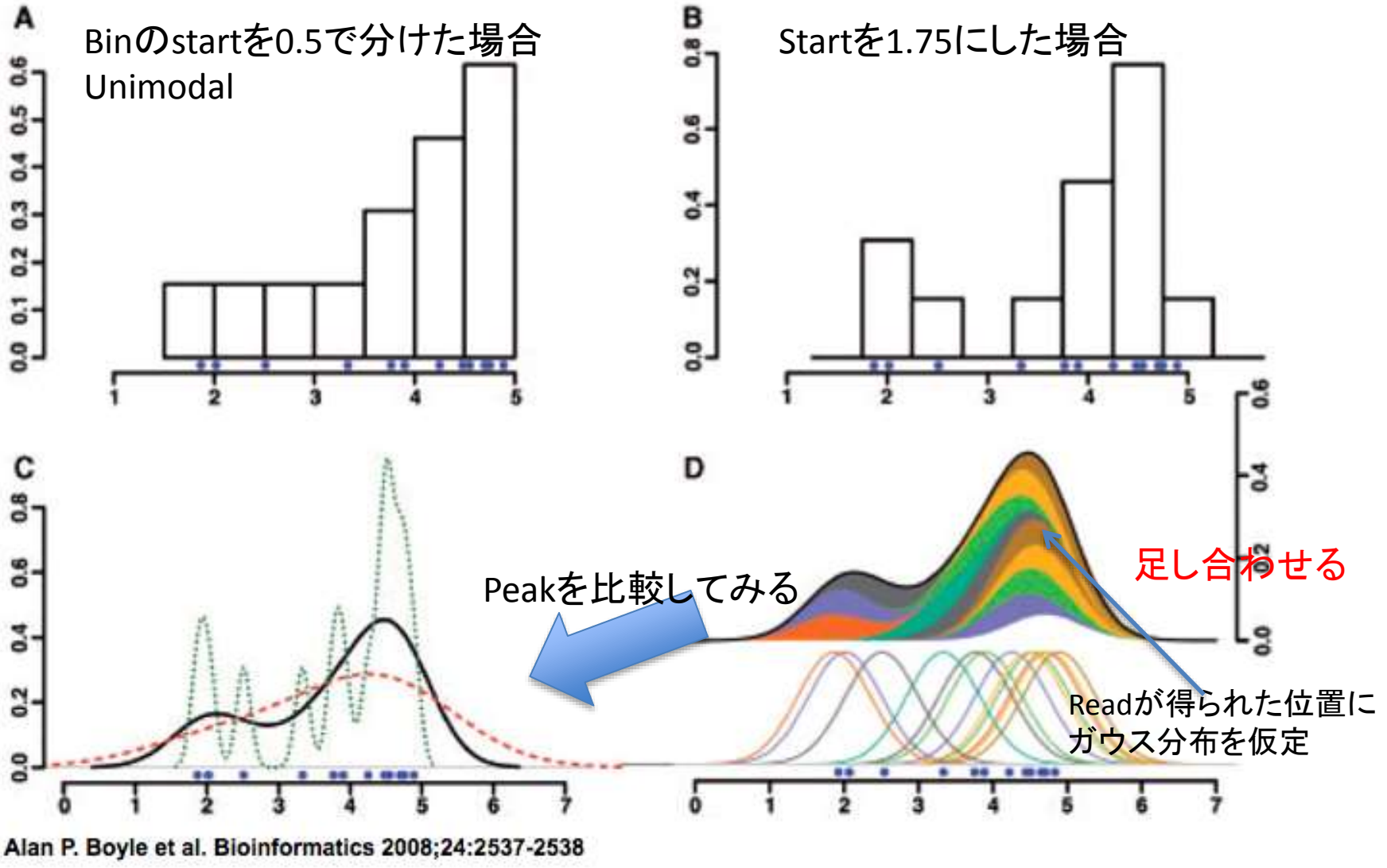




F, Rのそれぞれの頂点からcenterになる位置をPeakとする

d: distance between the summits of red and blue curves

3. F-Seq



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

Bioinformatics

Table 1 | Publicly available ChIP-seq software packages discussed in this review

	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based/ duplicate[e]	Refs.
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	10
ERANGE v3.1	Tag aggregation	1: Height cutoff High quality peak estimate, per-region estimate, or input	High quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally *P* values	*P* value	1: None 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Optional peak height, ratio to background	Yes / No	4,18
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes	19
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No	14
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Target FDR, number nearest neighbors for clustering	No / No	17
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	*P*-value threshold, tag length, mfold for shift estimate	No / Yes	13
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No	5
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation	KDE for enrichment and empirical FDR estimation	*q* value	1: NA 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$ as a function of profile threshold	KDE bandwidth, peak height, subpeak valley depth, ratio to background	Yes / Yes	9
SICER v1.02	Window scan with gaps allowed	*P* value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and *P* values	*q* value	1: None 2: From Poisson *P* values	Window length, gap size, FDR (with control) or *E*-value (no control)	No / Yes	15
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest paired tag distance	Used to compute fold-enrichment distribution	*P* value	1: Poisson 2: control distribution	1: FDR 1,2: $N_+ + N_-$ threshold	Yes / Yes	11
spp v1.0	Strand specific window scan	Poisson *P* value (paired peaks only)	Maximal strand cross-correlation	Subtracted before peak calling	*P* value	1: Monte Carlo simulation 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Ratio to background	Yes / No	12
USeq v4.2	Window scan	Binomial *P* value	Estimated or user specified	Subtracted before peak calling	*q* value	1, 2: binomial 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Target FDR	No / Yes	20

Peak Call法の種類は沢山ありますが、2009年から現在までそれほど進化していません。

Shirley Pepke et al. Nature REVIEW 2009

Peak Call後の解析

- 遺伝子注釈 (Nearest Neighbor gene annotation)
- TFBS search
- Motif discovery Analysis

Genome, Transcriptome, Proteome, etc.

HOMERがオススメ

ChIP-seq: Pipelineが一通り揃ったもの

http://homer.salk.edu/homer/chipseq/

Background:

- Introduction to ChIP-Seq
- Aligning ChIP-Seq tags

Standard ChIP-Seq analysis with HOMER:

1. Creating a "Tag Directory" from aligned sequences
2. Basic quality control (sequence bias, fragment length estimation)
3. Creating files to view your data in the UCSC Genome Browser
4. Finding Peaks (ChIP-enriched regions) in the genome
5. Finding enriched motifs in ChIP-Seq peaks
6. Annotating Peaks (and cross referencing other experiments and motifs)

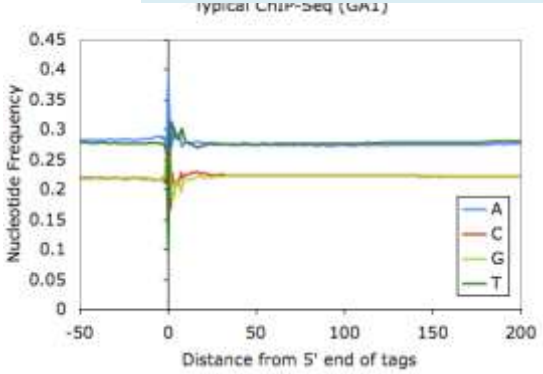
Automating standard ChIP-Seq analysis with analyzeChIP-Seq.pl

Advanced ChIP-Seq Analysis with HOMER:

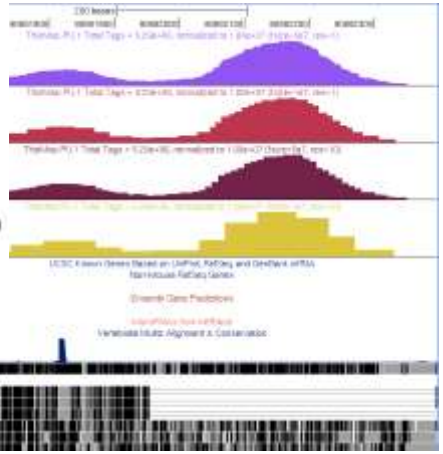
- Finding overlapping or differentially bound peaks
- Creating histograms with sequencing data
- Creating heatmaps with sequencing data
- Re-centering peaks on motifs

HOMER

プログラミングなしで解析できるおおよそ全てのChIP-seq解析内容がパイプライン化されている。カスタマイズもRとperlの知識で簡単。

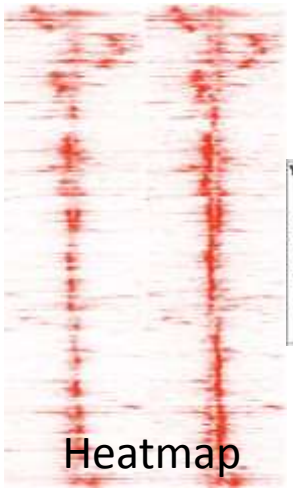


QC



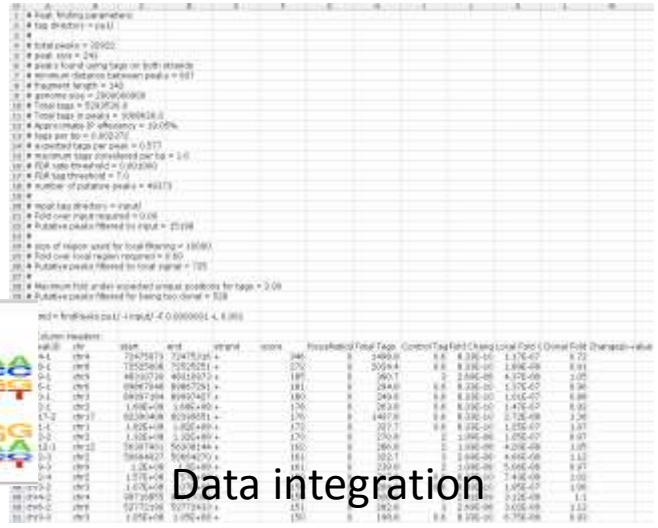
Visualize

Peak Call



Heatmap

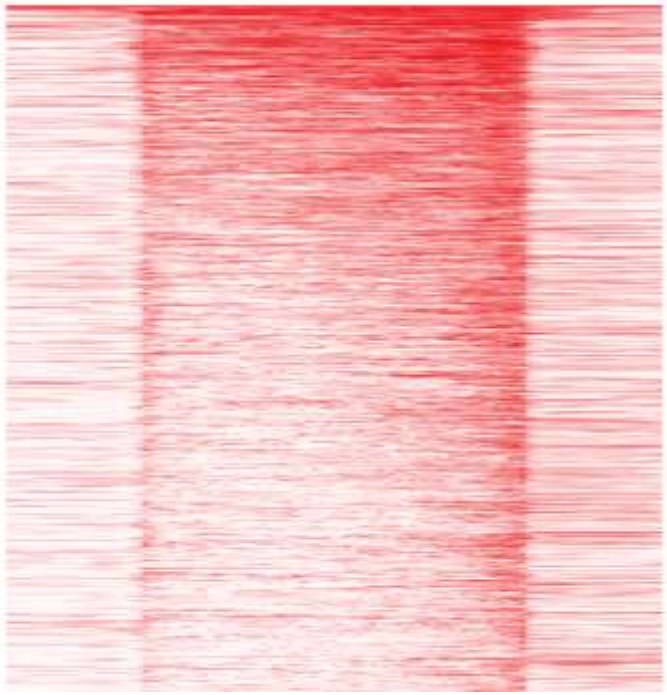
Motif analysis

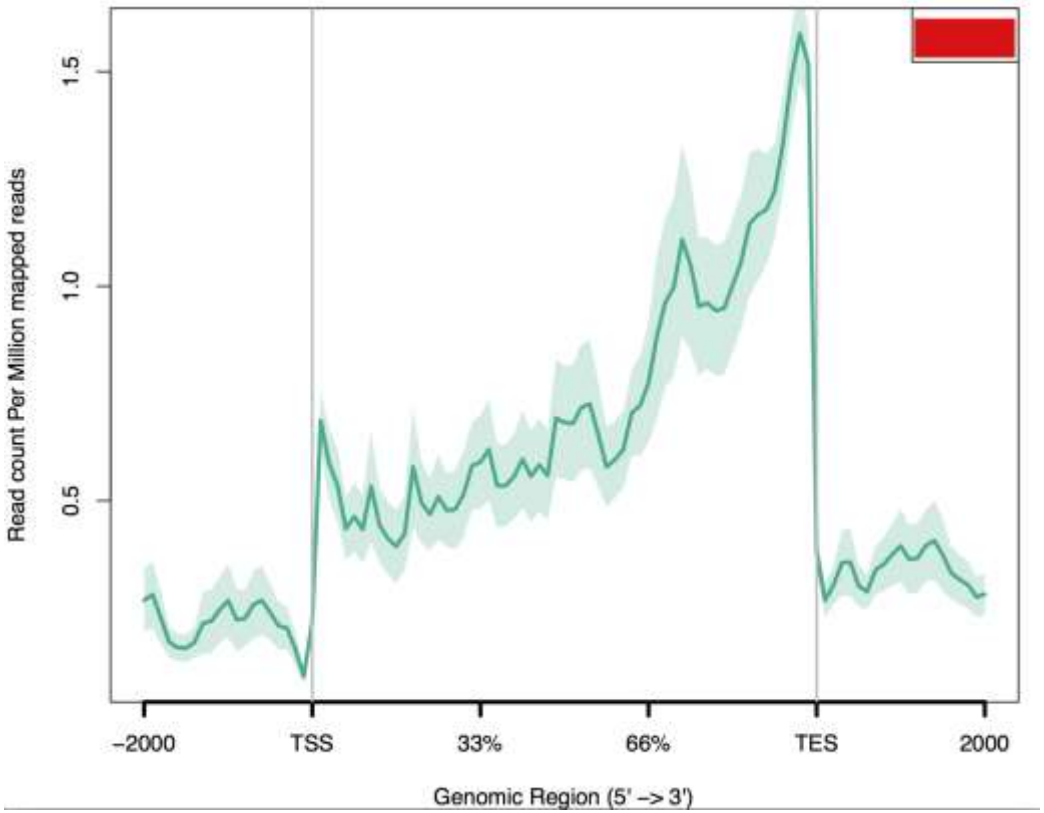


Data integration

NGS plot

Rを用いて、データの可視化が行える。
特に、gene bodyでの分布や、データ間の分布の違いを表示するのに便利





統合解析環境
galaxy

https://use**galaxy**.org

http://galaxy.dbcls.jp

ChIP-seqデータをとにかく可視化

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

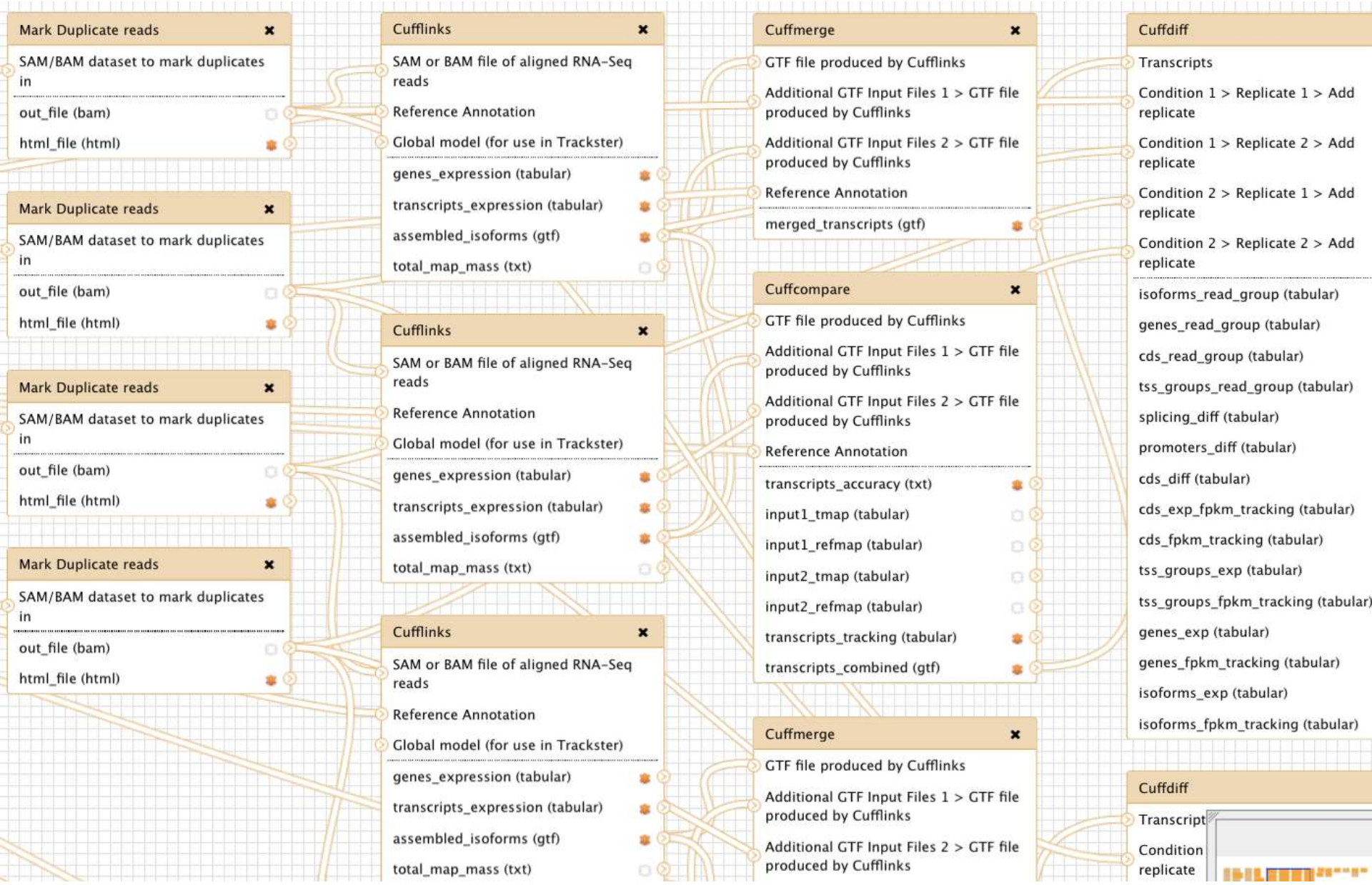
Sratailer

[presentation] 既存のChIP-seqデータを全て可視化する

本日の統合TVは、2014年12月22日にライフサイエンス統合データベースセンター (DBCLS) にて行われた 沖 真弥 助教 (九州大学大学院
学研究院 発生再生医学分野)によるセミナー「既存のChIP-seqデータを全て可視化する」をお送りします。ChIP-seqデータの可視化ツー
SraTailorに関する話題を中心にお話いただきました。約40分です。

YouTube版はこちらです。

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


Mark Duplicate reads
SAM/BAM dataset to mark duplicates in
out_file (bam)
html_file (html)
Mark Duplicate reads
SAM/BAM dataset to mark duplicates in
out_file (bam)
html_file (html)
Mark Duplicate reads
SAM/BAM dataset to mark duplicates in
out_file (bam)
html_file (html)
Mark Duplicate reads
SAM/BAM dataset to mark duplicates in
out_file (bam)
html_file (html)
Cufflinks
SAM or BAM file of aligned RNA-Seq reads
Reference Annotation
Global model (for use in Trackster)
genes_expression (tabular)
transcripts_expression (tabular)
assembled_isoforms (gtf)
total_map_mass (txt)
Cufflinks
SAM or BAM file of aligned RNA-Seq reads
Reference Annotation
Global model (for use in Trackster)
genes_expression (tabular)
transcripts_expression (tabular)
assembled_isoforms (gtf)
total_map_mass (txt)
Cufflinks
SAM or BAM file of aligned RNA-Seq reads
Reference Annotation
Global model (for use in Trackster)
genes_expression (tabular)
transcripts_expression (tabular)
assembled_isoforms (gtf)
total_map_mass (txt)
Cuffmerge
GTF file produced by Cufflinks
Additional GTF Input Files 1 > GTF file produced by Cufflinks
Additional GTF Input Files 2 > GTF file produced by Cufflinks
Reference Annotation
merged_transcripts (gtf)
Cuffcompare
GTF file produced by Cufflinks
Additional GTF Input Files 1 > GTF file produced by Cufflinks
Additional GTF Input Files 2 > GTF file produced by Cufflinks
Reference Annotation
transcripts_accuracy (txt)
input1_tmap (tabular)
input1_refmap (tabular)
input2_tmap (tabular)
input2_refmap (tabular)
transcripts_tracking (tabular)
transcripts_combined (gtf)
Cuffmerge
GTF file produced by Cufflinks
Additional GTF Input Files 1 > GTF file produced by Cufflinks
Additional GTF Input Files 2 > GTF file produced by Cufflinks
Cuffdiff
Transcripts
Condition 1 > Replicate 1 > Add replicate
Condition 1 > Replicate 2 > Add replicate
Condition 2 > Replicate 1 > Add replicate
Condition 2 > Replicate 2 > Add replicate
isoforms_read_group (tabular)
genes_read_group (tabular)
cds_read_group (tabular)
tss_groups_read_group (tabular)
splicing_diff (tabular)
promoters_diff (tabular)
cds_diff (tabular)
cds_exp_fpkm_tracking (tabular)
cds_fpkm_tracking (tabular)
tss_groups_exp (tabular)
tss_groups_fpkm_tracking (tabular)
genes_exp (tabular)
genes_fpkm_tracking (tabular)
isoforms_exp (tabular)
isoforms_fpkm_tracking (tabular)
Cuffdiff
Transcript
Condition
replicate

本日のデータ

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

GM12878 CTCF Histone Mods by ChIP-seq Peaks from ENCODE/Broad

Position: chr21:33055346-33055603
Score: 559
Signal value: 12.229
P-value (-log10): 9.600

View table: schema, downloads, metadata

Go to Broad Histone track controls

Data version: ENCODE Jan 2011 Freeze
Data last updated: 2010-11-05

- ヒトリンパ芽球様**細胞の細胞株**
- **CTCF抗体**

Ram O, et al. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell*. 2011 Dec 23;147(7):1628-39.

Description

This track displays maps of chromatin state generated by the Broad/MGH ENCODE group using ChIP-seq. Chemical modifications (methylation, acetylation) to the histone proteins present in chromatin influence gene expression by changing how accessible the chromatin is to transcription.

The ChIP-seq method involves first using formaldehyde to cross-link histones and other DNA-associated proteins to genomic DNA within cells. The cross-linked chromatin is subsequently extracted, mechanically sheared, and immunoprecipitated using specific antibodies. After reversal of cross-links, the immunoprecipitated DNA is sequenced and mapped to the human reference genome. The relative enrichment of each antibody-target (epitope) across the genome is inferred from the density of mapped fragments.

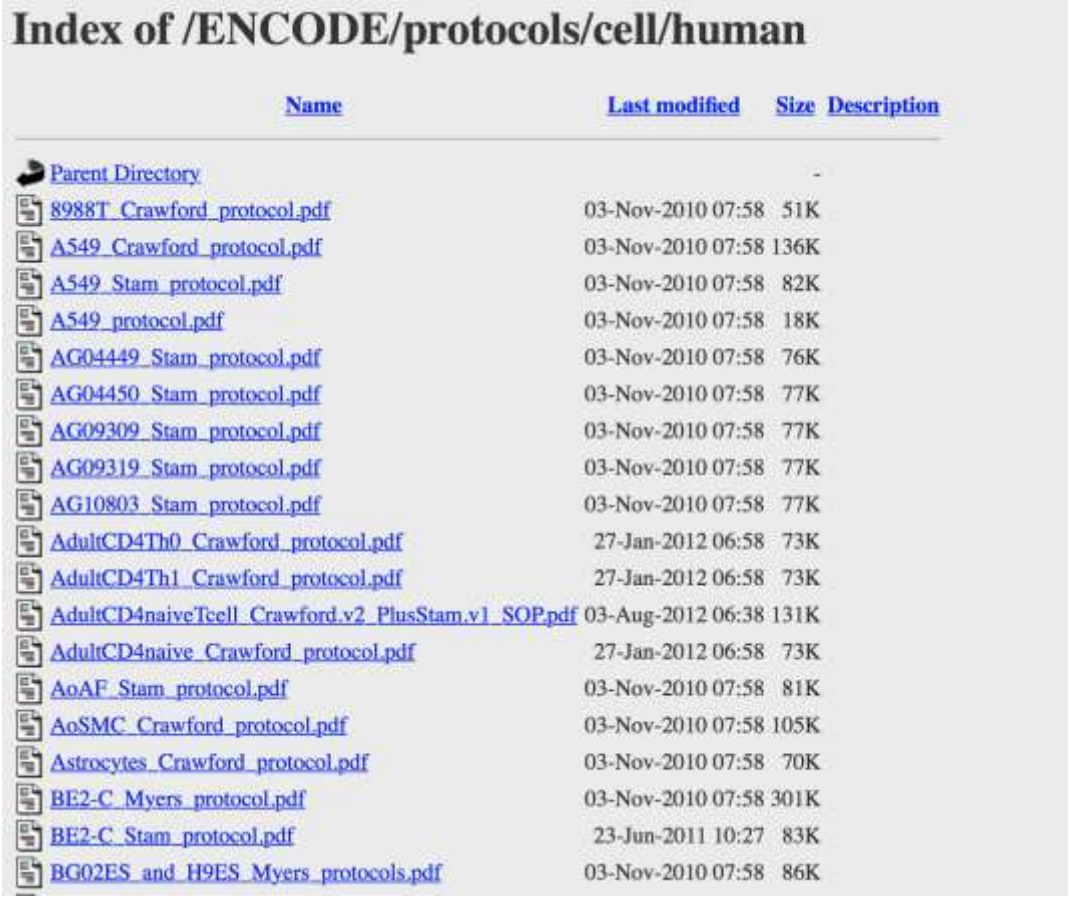
本日のデータ

ENCODEプロジェクトのNGSデータは、
実験プロトコルが公開されているので、
ChIP-seqをはじめNGS用の実験をはじめる
場合の参考にもなる

https://genome.ucsc.edu/ENCODE/protocols/

今回のデータ

https://genome.ucsc.edu/ENCODE/protocols/cell/human/GM12878_protocol.pdf



Index of /ENCODE/protocols/cell/human

Name	Last modified	Size	Description
Parent Directory		-	
8988T_Crawford_protocol.pdf	03-Nov-2010 07:58	51K	
A549_Crawford_protocol.pdf	03-Nov-2010 07:58	136K	
A549_Stam_protocol.pdf	03-Nov-2010 07:58	82K	
A549_protocol.pdf	03-Nov-2010 07:58	18K	
AG04449_Stam_protocol.pdf	03-Nov-2010 07:58	76K	
AG04450_Stam_protocol.pdf	03-Nov-2010 07:58	77K	
AG09309_Stam_protocol.pdf	03-Nov-2010 07:58	77K	
AG09319_Stam_protocol.pdf	03-Nov-2010 07:58	77K	
AG10803_Stam_protocol.pdf	03-Nov-2010 07:58	77K	
AdultCD4Th0_Crawford_protocol.pdf	27-Jan-2012 06:58	73K	
AdultCD4Th1_Crawford_protocol.pdf	27-Jan-2012 06:58	73K	
AdultCD4naiveTcell_Crawford.v2_PlusStam.v1_SOP.pdf	03-Aug-2012 06:38	131K	
AdultCD4naive_Crawford_protocol.pdf	27-Jan-2012 06:58	73K	
AoAF_Stam_protocol.pdf	03-Nov-2010 07:58	81K	
AoSMC_Crawford_protocol.pdf	03-Nov-2010 07:58	105K	
Astrocytes_Crawford_protocol.pdf	03-Nov-2010 07:58	70K	
BE2-C_Myers_protocol.pdf	03-Nov-2010 07:58	301K	
BE2-C_Stam_protocol.pdf	23-Jun-2011 10:27	83K	
BG02ES_and_H9ES_Myers_protocols.pdf	03-Nov-2010 07:58	86K	