2012.08.24
R/Bioconductor によるNGS解析: Part. 3

ChIP-seq データ解析の基礎

Itoshi NIKAIDO, Ph.D.
RIKEN CDB@Kobe

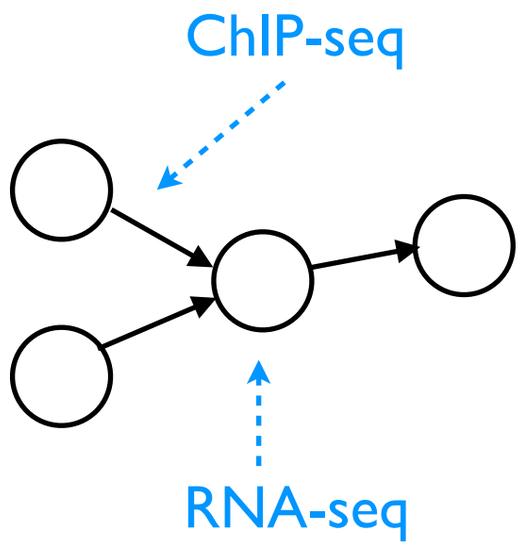
はじめに

講義で使用するレジュメ、ソースコードとデータはすべて以下からダウンロードできます。

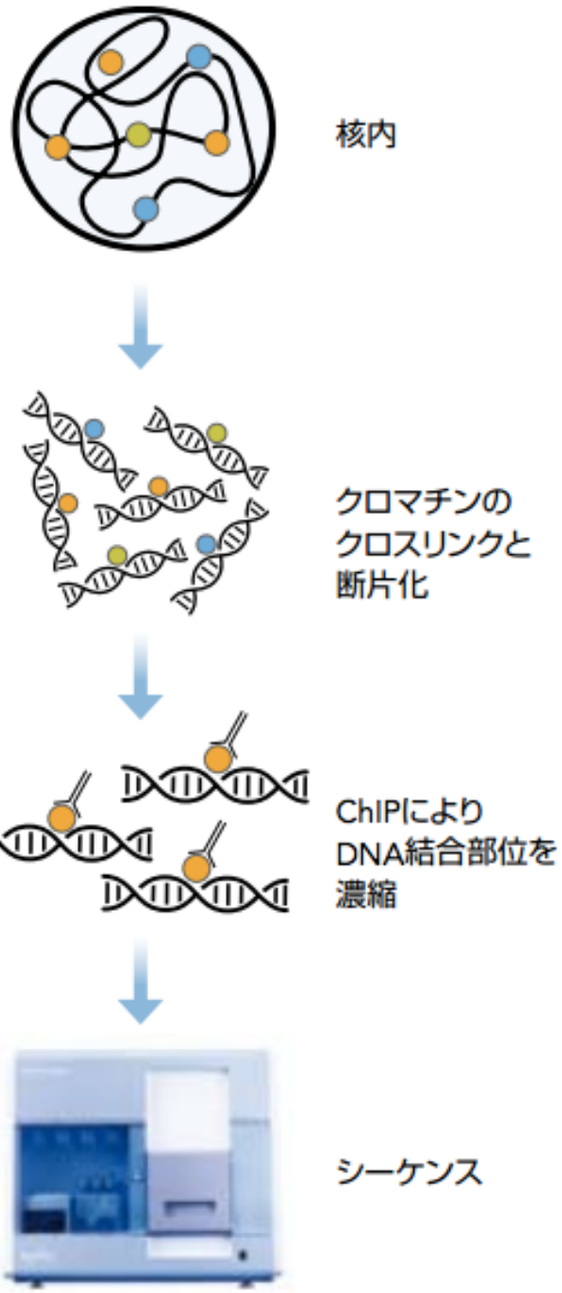
http://cat.HackingIsBelieving.org/lecture/

生命科学とChIP-seq

- 生命現象は複数の因子が相互作用する複雑なプロセス
  - 因子の量の変化
    - RNA-seq, CAGE-seq
  - 因子の相互作用
    - ChIP-seq



ChIP-seq



http://www.illuminakk.co.jp/pdf/ChIP-Seq_DS.pdf より引用

タンパク質が結合しているゲノム領域の地図を描く

タンパク質が結合しているゲノム領域のDNAを enrichment させる

× purification

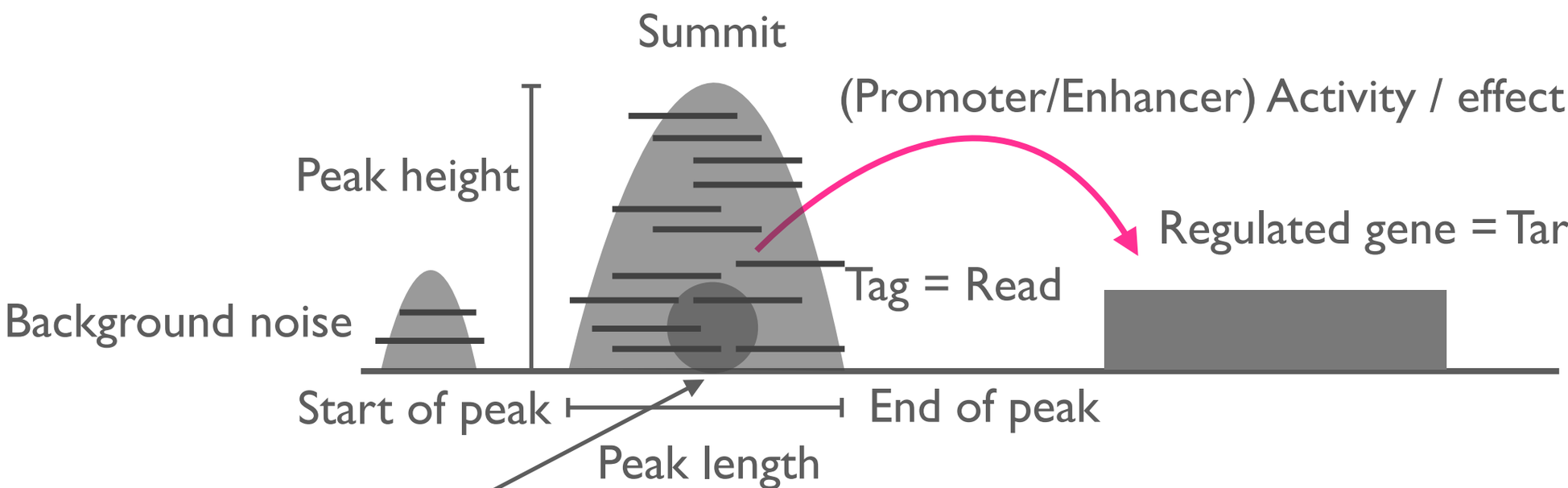
※ タンパク質-DNA結合が転写などの減少の因果を示すとは限らないことに注意

ChIP-seqの一般化

- illumina HiSeq 2000
  - 14 lanes x 3 multiplex = 42 samples
  - 4.7万円/sample
    - = 200万円/42 samples
- Tilling Array
  - 70万円/samples

低コスト化、n<<p 問題の緩和

Terminology of ChIP-seq



Binding event (binding site) → Consensus sequence of DNA (binding) motif

yywTTswyATGCAAAw

Position weight matrix → Sequence Logo of DNA Motif

	1	2	3	4	5	6	7	8	9	10	11
A	0.0651	0.0706	0.4631	0.0067	0.0575	0.1146	0.3751	0.1744	0.9098	0.0081	0.0012
C	0.4672	0.5806	0.0397	0.0356	0.0046	0.2761	0.0129	0.2541	0.0046	0.0012	0.0032
G	0.1985	0.0596	0.0294	0.0218	0.0967	0.5559	0.0204	0.1173	0.0204	0.0081	0.8803
T	0.2692	0.2892	0.4679	0.9360	0.8411	0.0534	0.5916	0.4541	0.0651	0.9827	0.1153

目的

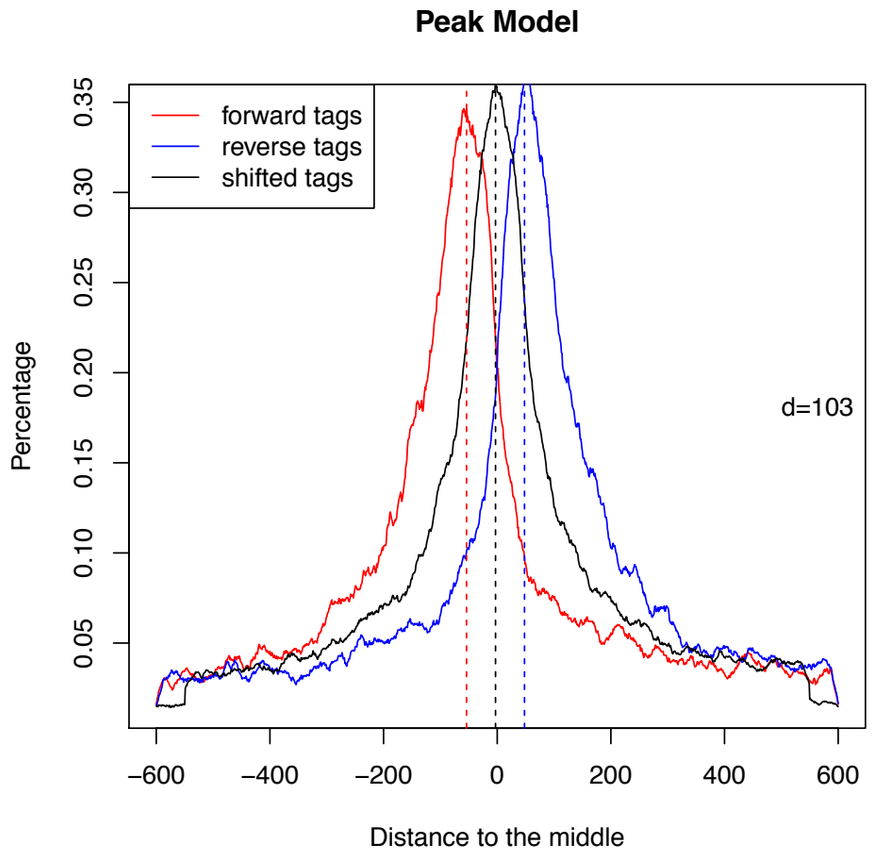
1. ChIP-seq解析の流れとポイントを理解する
2. 実際に利用されているRのコードを読む
3. 自分でパイプラインを構成できるようになる

Pipeline for ChIP-seq analysis

MACS2

```
## 結果ファイル
$ cd results/macs2
$ less Oct4_peaks.bed
chr1    6448151 6448293 MACS_peak_1     11.91
chr1    7037487 7037628 MACS_peak_2     14.86
chr1    7303701 7303804 MACS_peak_3     14.42
$ less Oct4_summit.bed
chr1    6448196 6448197 MACS_summit_1   11.91
chr1    7037538 7037539 MACS_summit_2   14.86
chr1    7303769 7303770 MACS_summit_3   14.42

## Peak model distribution を描画する
$ R -q -f Oct4_model.r
```



peak calling on R

iSeq: Bayesian Hierarchical Modeling of ChIP-seq Data Through Hidden Ising Models
隠れイジングモデルを使った binding site の同定。手法の元論文は、Q Mo, 2011. A fully Bayesian hidden Ising model for ChIP-seq data analysis, Biostat.
http://www.bioconductor.org/packages/2.9/bioc/html/iSeq.html

CSAR: Statistical tools for the analysis of ChIP-seq data
いわゆる peak caller で正規化・サンプル間比較などもできる。有意差はFDRで。C++
http://www.bioconductor.org/packages/release/bioc/html/CSAR.html

BayesPeak: Bayesian Analysis of ChIP-seq Data
Peak caller. 入力は BED file
http://www.bioconductor.org/packages/release/bioc/html/BayesPeak.html

PICS: Probabilistic inference of ChIP-seq
Empirical Bayes mixture model による peak calling。snow で分散計算することが推奨されている。
http://www.bioconductor.org/packages/release/bioc/html/PICS.html

RangedData Object

Data structure:

0. IRanges data of Peaks
1. Factor of space (chromosome)
2. additional information (score, strand)

```
> oct4.gr
RangedData with 1675 rows and 2 value columns across 21 spaces
                 space                 ranges |     strand      score
             <factor>              <IRanges> | <numeric> <numeric>
MACS_peak_1         1   [ 6448151,  6448293] |         1     11.91
MACS_peak_2         1   [ 7037487,  7037628] |         1     14.86
MACS_peak_3         1   [ 7303701,  7303804] |         1     14.42
MACS_peak_4         1   [ 7722943,  7723046] |         1      6.29
MACS_peak_5         1   [12734705, 12734815] |         1      8.33
MACS_peak_6         1   [12734855, 12734958] |         1      3.66
MACS_peak_7         1   [12826211, 12826358] |         1     22.40
MACS_peak_8         1   [14302765, 14302906] |         1      9.58
MACS_peak_9         1   [16120140, 16120296] |         1     20.94
```

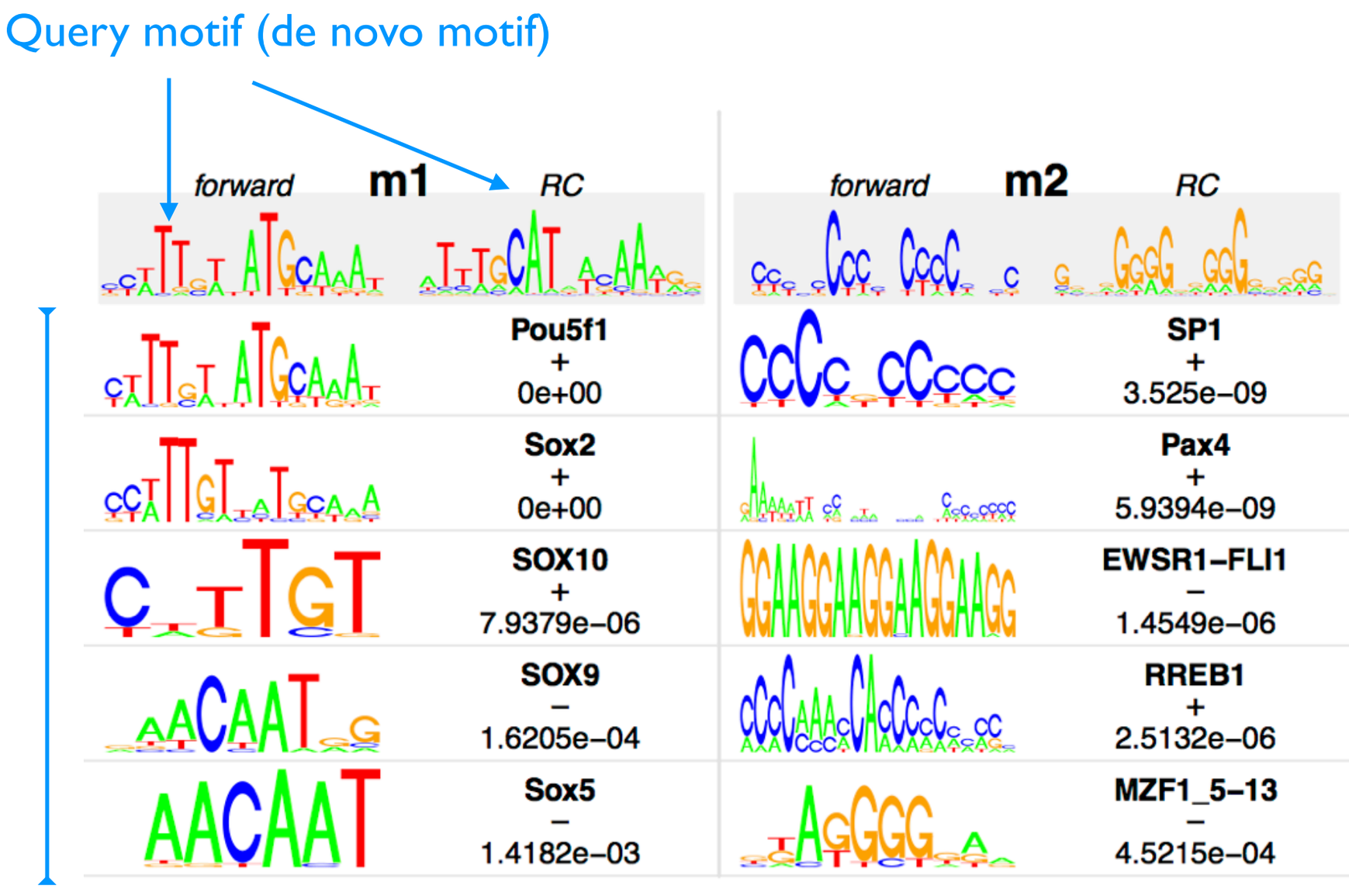
Peak

space = chromosome

Start of peaks

End of peaks

Motif database search



Correlation

Pearson / Spearman correlation coefficient

```
> query <- c(0,0,0,1,1,1,1,1,0,1,1,1,0,1,1,1,1,1,1,0)
> ref   <- c(0,0,0,0,1,1,1,0,0,1,0,0,1,1,1,1,1,1,1,1)
> cor(query, ref)
[1] 0.3563483
cor(query, ref, method="spearman")
[1] 0.3563483
```

localization vector

```
00011111101110111111110
00001110010011111111111
```

co-localization vector 00001110010001111110

全ゲノムの localization vector 間の相関を計算するのは難しい (メモリ)

binning (= window analysis, smoothing) して計算する必要がある

localization vector の相関は一般的に高くならないので注意

non-localization region にひっぱられて相関が不当に高くなる

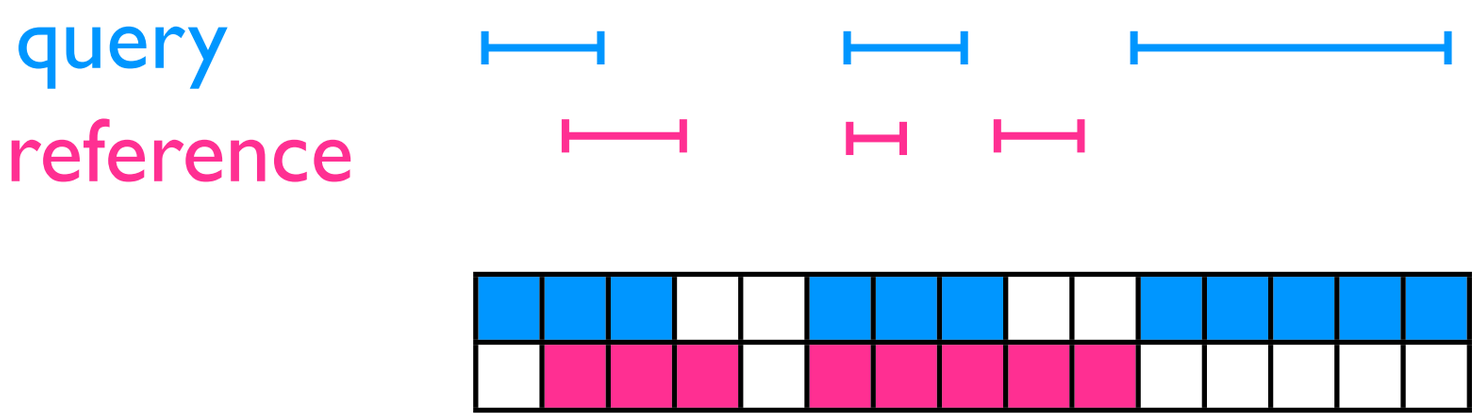
Preprocessing

Peak / window-based comparison

Peak-based

Peak の数や長さがの影響を受ける

Window-based

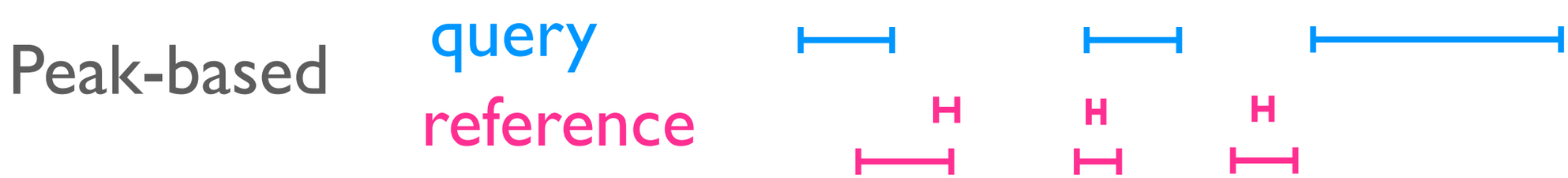


平滑化される

Preprocessing

Extension of peak length

Single / several base resolution (summit) のデータに対して、前後 200bp を加える処理をする



Single / several base resolution のデータに対して、前後 200bp を加える処理をする

Preprocessing

From score to localization vector

Peak height / FPKM のような score vector を localization vector に変換する

スコアの正規化の必要がない

score vector

000999990333055555500000888002009999999

localization vector

000111110111011111110000111100100111111111

Correlation of contingency

Phi / Contingency C / Cramer V coefficient

$$E_{ij} = n_{i.}\, n_{.j} \,/\, n$$

$$\chi_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} (O_{ij} - E_{ij})^2 \,/\, E_{ij}$$

係数	定義	とる値の範囲
ϕ	$\sqrt{\chi_0^2 / n}$	$0 \sim \sqrt{t-1}$
C	$\sqrt{\chi_0^2 / (n+\chi_0^2)}$	$0 \sim \sqrt{(t-1)/t}$
V	$\phi / \sqrt{t-1}$	$0 \sim 1$

$t = \min(k, m)$

2 x 2 contingency table に対する phi coefficient = Pearson correlation

contingency table の要素の大きさに影響を受ける

空間情報を利用していない

Contingency Table

query \ reference	0	1	sum
0	4	1	6
1	4	10	14
sum	8	12	20

まとめ

- 簡単な前処理について
- R + Bioconductor を利用してChIP-seqのデータを操作する
  - アノテーション
  - モチーフ検索
  - 簡単な比較