

2012.02.22

Rでつなぐ次世代オミックス情報統合解析研究会

# R + Bioconductor によるChIP-seq解析の基礎

Itoshi NIKAIDO, Ph.D.

RIKEN CDB@Kobe

# はじめに

講義で使用する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。

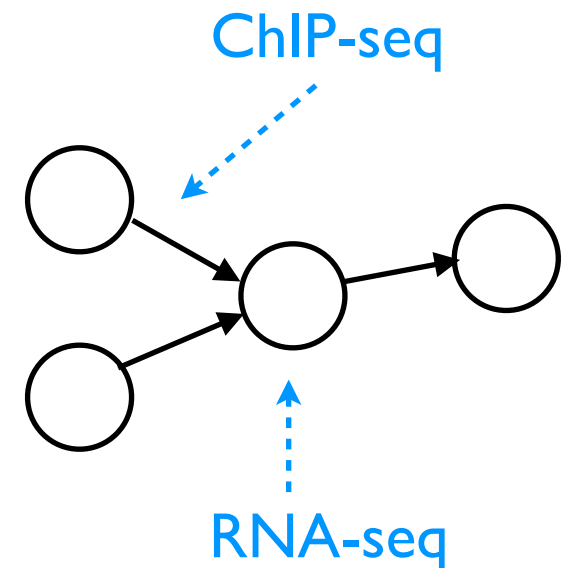
[http://github.com/dritoshi/jsbi\\_chipseq/wiki](http://github.com/dritoshi/jsbi_chipseq/wiki)



この作品は [クリエイティブ・コモンズ 表示 - 非営利 2.1 日本 ライセンス](#)の下に提供されています。

# 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



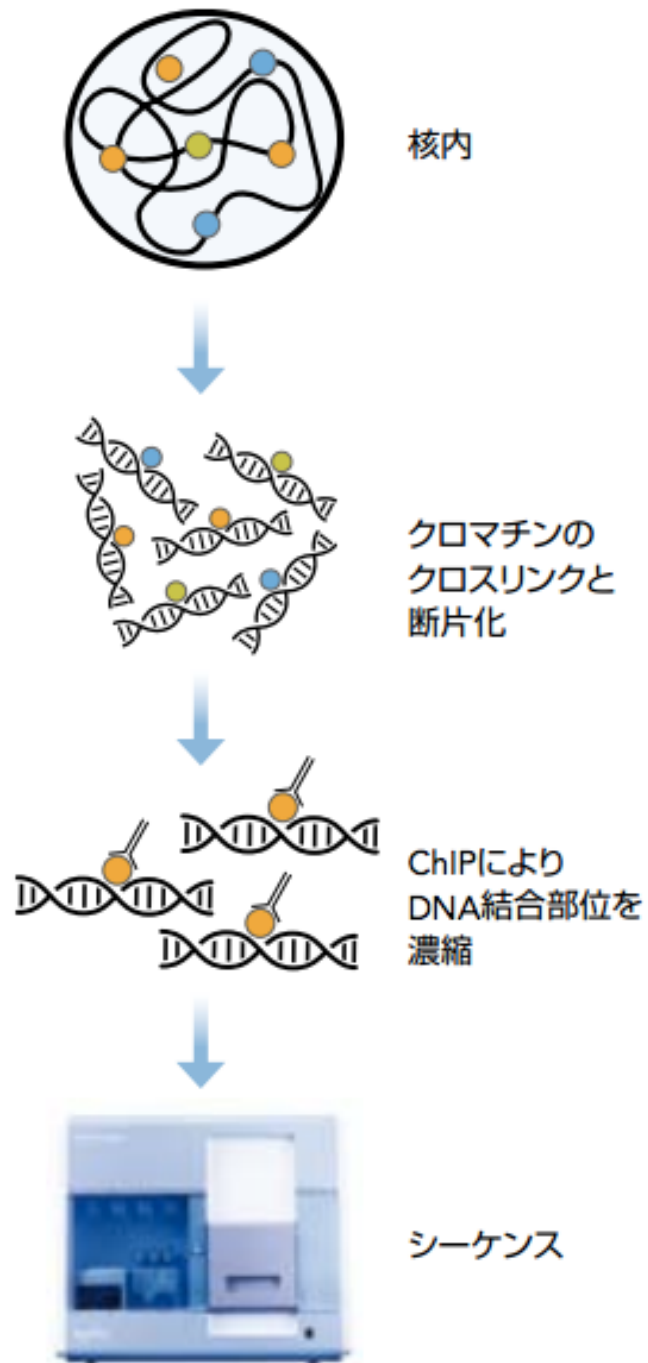
# ChIP-seq

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

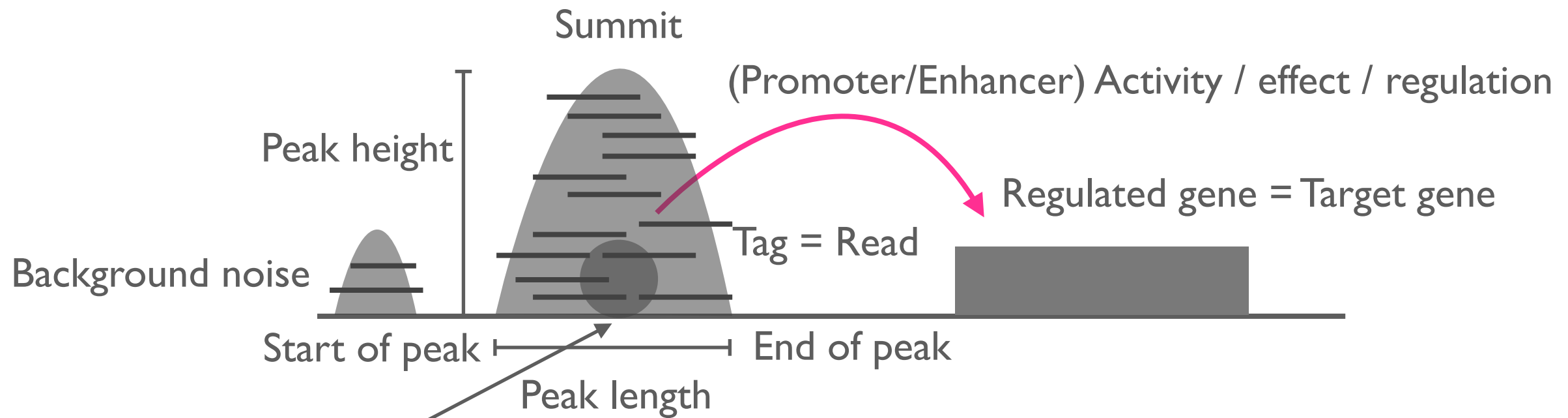


# ChIP-seqの一般化

- 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、 $n \ll p$  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

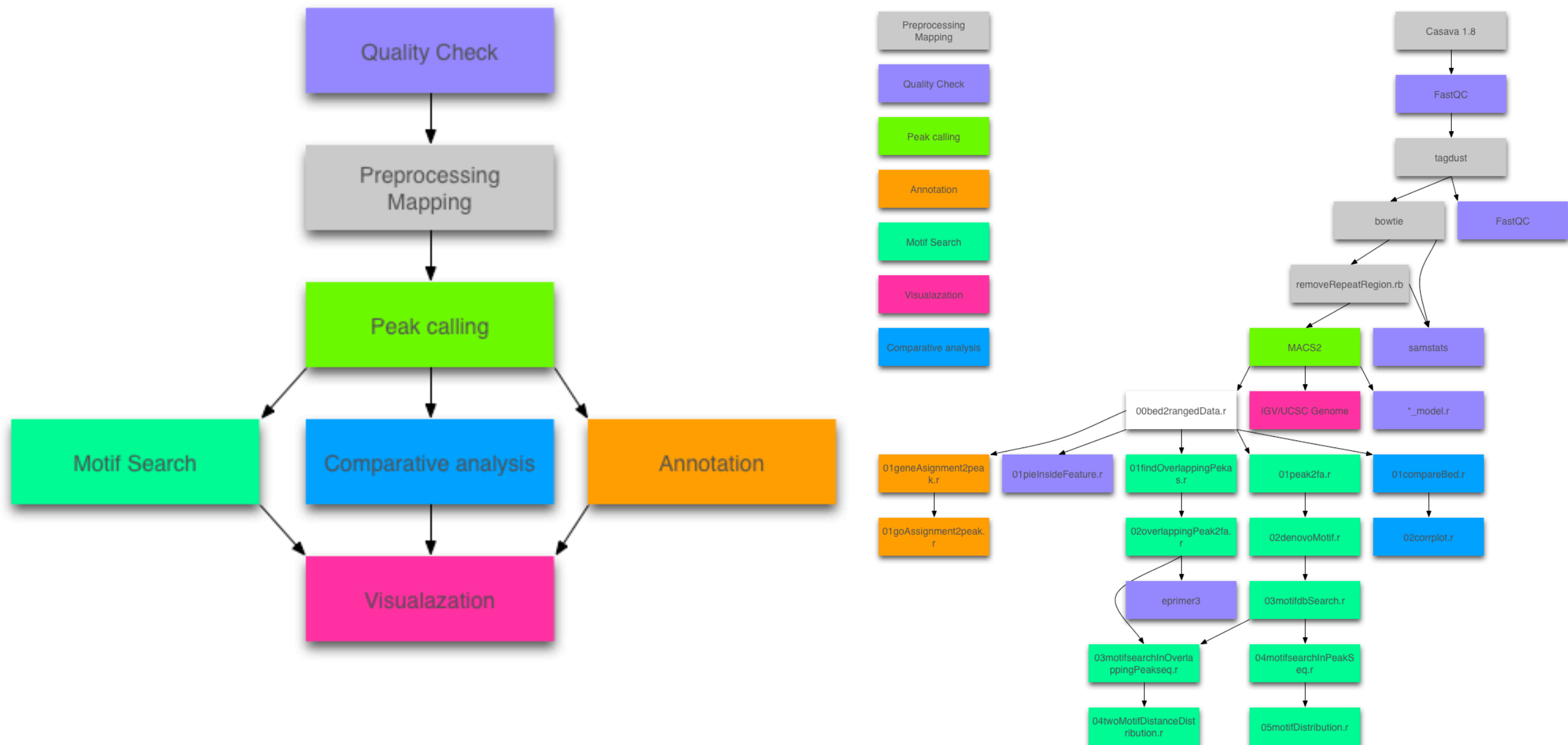
	1	2	3	4	5	6	7	8	9	10	11
A	0.0651	0.0706	0.4631	0.0067	0.0575	0.1146	0.3751	0.1744	0.9098	0.0081	0.0012
C	0.4672	0.5806	0.0397	0.0356	0.0046	0.2761	0.0129	0.2541	0.0046	0.0012	0.0032
G	0.1985	0.0596	0.0294	0.0218	0.0967	0.5559	0.0204	0.1173	0.0204	0.0081	0.8803
T	0.2692	0.2892	0.4679	0.9360	0.8411	0.0534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



# 目的

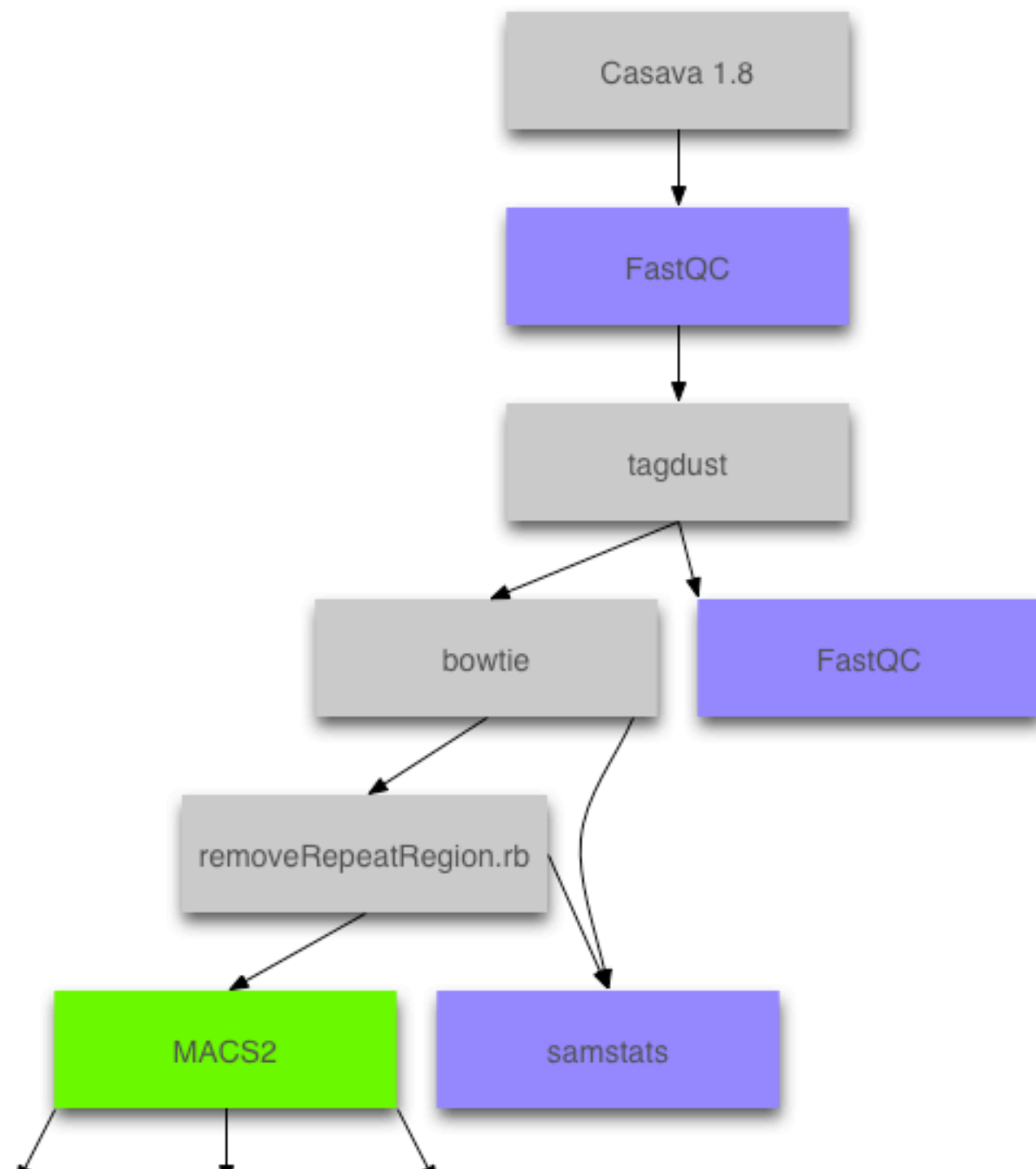
1. ChIP-seq解析の流れとポイントを理解する
2. 実際に利用されているRのコードを読む
3. 自分でパイプラインを構成できるようになる

# Pipeline for ChIP-seq analysis





# Preprocessing/mapping/ Quality check



Check Points:

- 0. Quality value of sequence
- 1. Mapping Rate
- 2. Adapter/Primer contamination
- 3. Read duplication rate

特に3はPCRバイアスの評価に繋がる大切な指標

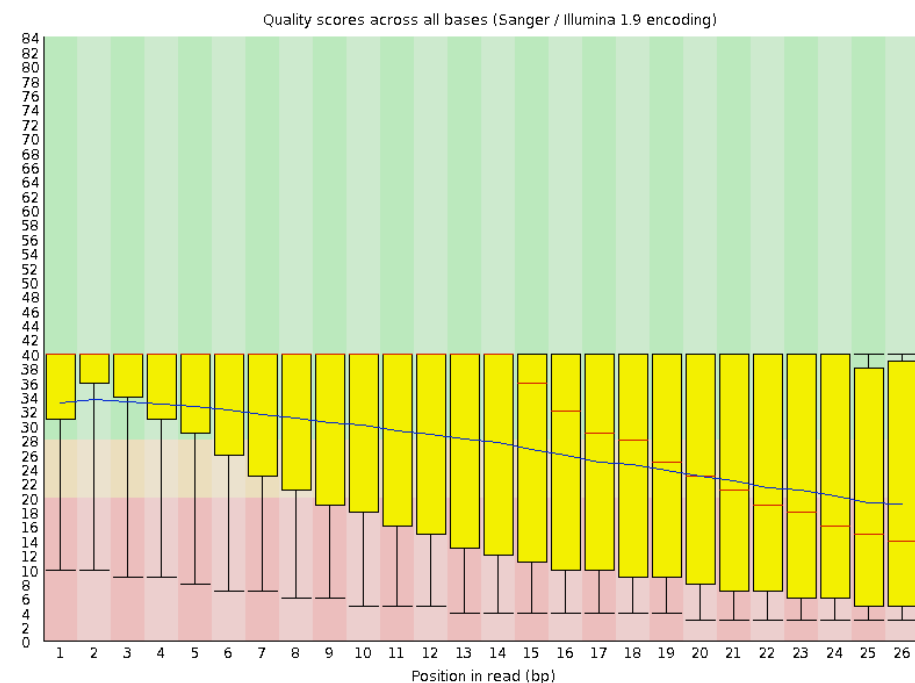
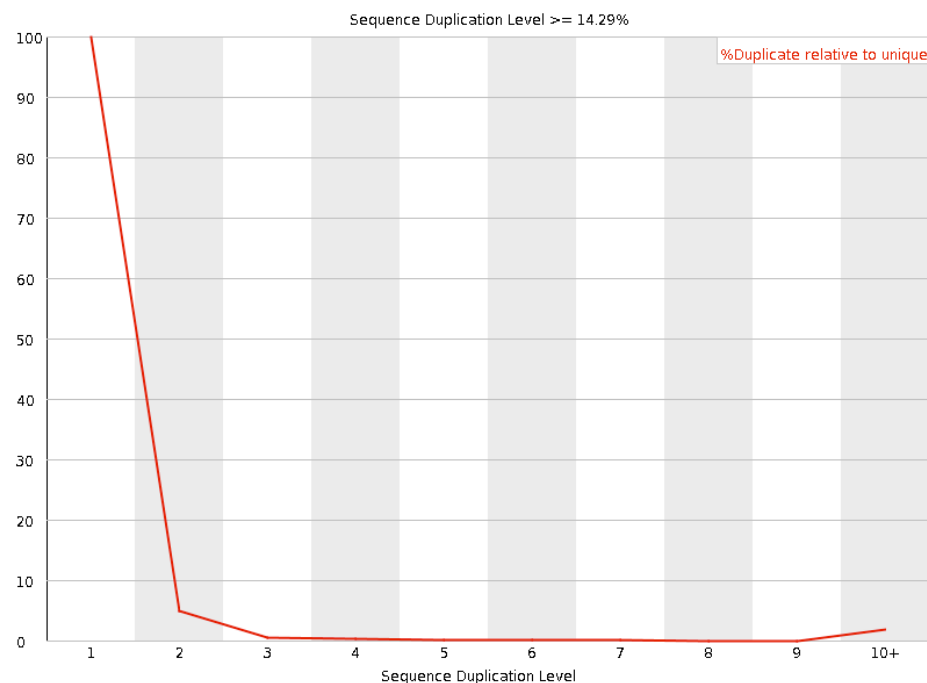
# FastQC

## # インストール

```
$ curl -O http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/fastqc\_v0.10.0.zip
$ unzip fastqc_v0.10.0.zip
$ sudo cp fastqc_v0.10.0 /opt
$ sudo ln -s /opt/fastqc_v0.10.0 /opt/fastqc
$ emacs -nw ~/.zshenv
export PATH=$PATH:/opt/fastqc
```

## # FastQC実行

```
$ fastqc -t 8 results/fastq/Oct4.fastq -o results/fastqc/Oct4
```



# tagdust

## ## インストール

```
$ curl -O http://genome.gsc.riken.jp/osc/english/software/src/tagdust.tgz zxvf tagdust.tgz
```

```
$ cd tagdust/
```

```
$ make
```

```
$ sudo make install
```

```
$ rehash
```

```
$ emacs adapter.fasta
```

```
>Adapter 1
```

```
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
```

```
>Adapters 2
```

```
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
```

```
>PCR Primers 1
```

```
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
```

```
>PCR Primers 2
```

```
CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT
```

```
>Genomic DNA Sequencing Primer
```

```
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
```

# tagdust

```
## tagdust の実行
```

```
## illumina ChIP-seq Prep. Kit の adapter/primer sequence を除く
```

```
$ tagdust adapter.fasta Oct4.pre.fastq -fdr 0.05 -o Oct4.fastq -a  
Oct4.artifactual.fastq
```

```
TagDust version 1.13, Copyright (C) 2009 Timo Lassmann
```

```
<timolassmann@gmail.com>
```

```
Creating Library
```

```
Generating Background
```

```
51          done      (73270919)
```

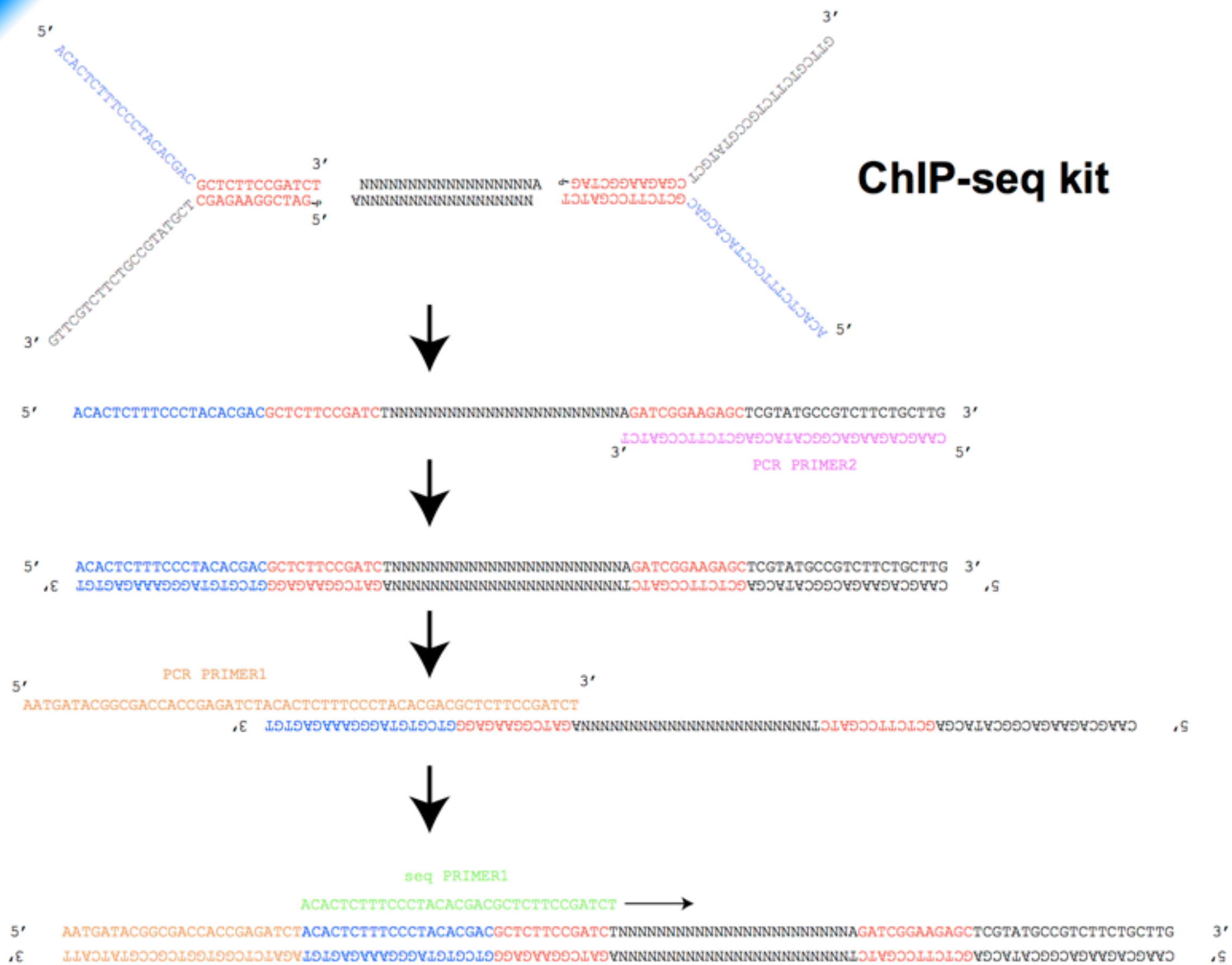
```
73270919          Sequences
```

```
2103129 Tags rejected (2.9%) at 25.0% coverage cutoff (0.050000 FDR).
```

```
Elapsed time:
```

```
720 seconds (668.20 CPU+SYS seconds)
```

Advance



Provided by Dr. Yohei Sasagawa @ RIKEN CDB

# Bowtie

# Bowtieの実行

```
$ bowtie -t -p 8 -n 3 -m 1 -a --best --strata --sam mm9 -q Oct4.fastq > Oct4.bowtie.sam
```

# bam に変換

```
$ samtools view -bS Oct4.bowtie.sam > Oct4.bowtie.bam
$ samtools sort      Oct4.bowtie.bam      Oct4.bowtie.sort
$ samtools index      Oct4.bowtie.sort.bam
$ rm Oct4.bowtie.bam
```

# repeat region にマップされたreadsを除去

```
$ intersectBed -abam Oct4.bowtie.sort.bam -b input/repeats.bed -v > Oct4.bowtie.rmRepeat.bam
$ samtools index Oct4.bowtie.rmRepeat.bam
```

# mapping rate を見る

```
$ samtools flagstat Oct4.bowtie.rmRepeat.bam > Oct4.bowtie.rmRepeat.summary.txt
$ samtools flagstat Oct4.bowtie.sort.bam > Oct4.bowtie.sort.summary.txt
$ cat Oct4.bowtie.sort.summary.txt
24021520 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
9667632 + 0 mapped (40.25%:nan%)
```

# MACS2

## numpy のインストール:

```
$ wget "http://downloads.sourceforge.net/project/numpy/NumPy/1.6.1/numpy-1.6.1.tar.gz"
```

```
$ tar zxvf numpy-1.6.1.tar.gz
```

```
$ cd numpy-1.6.1
```

```
$ python setup.py build --fcompiler=gnu
```

```
$ sudo python setup.py install
```

## MACS2のインストール:

```
$ w3m http://github.com/downloads/taoliu/MACS/MACS-2.0.9-1.tar.gz
```

```
$ tar zxvf MACS-2.0.9.tar.gz
```

```
$ cd MACS-2.0.9/
```

```
$ sudo python setup.py install --prefix=/opt
```

```
$ export PYTHONPATH=/opt/lib/python2.6/site-packages/:$PYTHONPATH
```

## MACS2の実行

```
$ macs2 -t results/bowtie/Sox2/Sox2.bowtie.sort.rmRepeat.bam -c results/bowtie/GFP/GFP.bowtie.sort.rmRepeat.bam -f BAM -g mm -n Sox2 -B -q 0.01
```

# MACS2

## 結果ファイル

```
$ cd results/mac2
```

```
$ less Oct4_peaks.bed
```

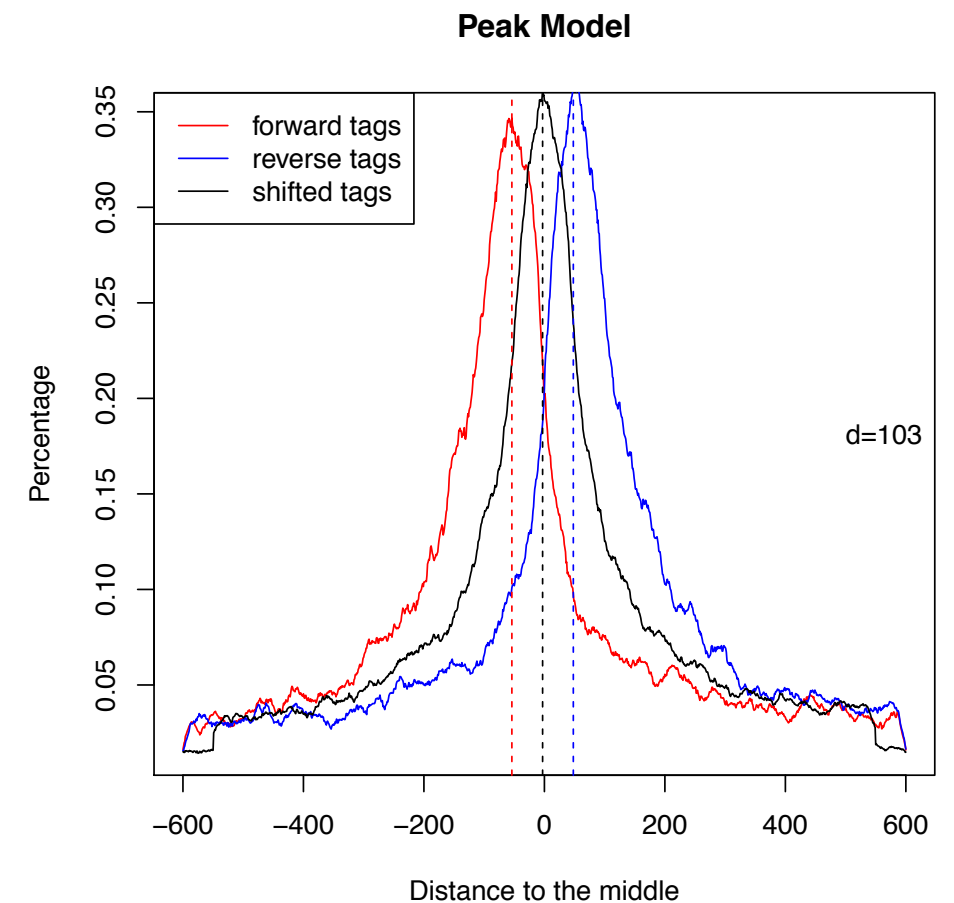
chr1	6448151	6448293	MACS_peak_1	11.91
chr1	7037487	7037628	MACS_peak_2	14.86
chr1	7303701	7303804	MACS_peak_3	14.42

```
$ 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

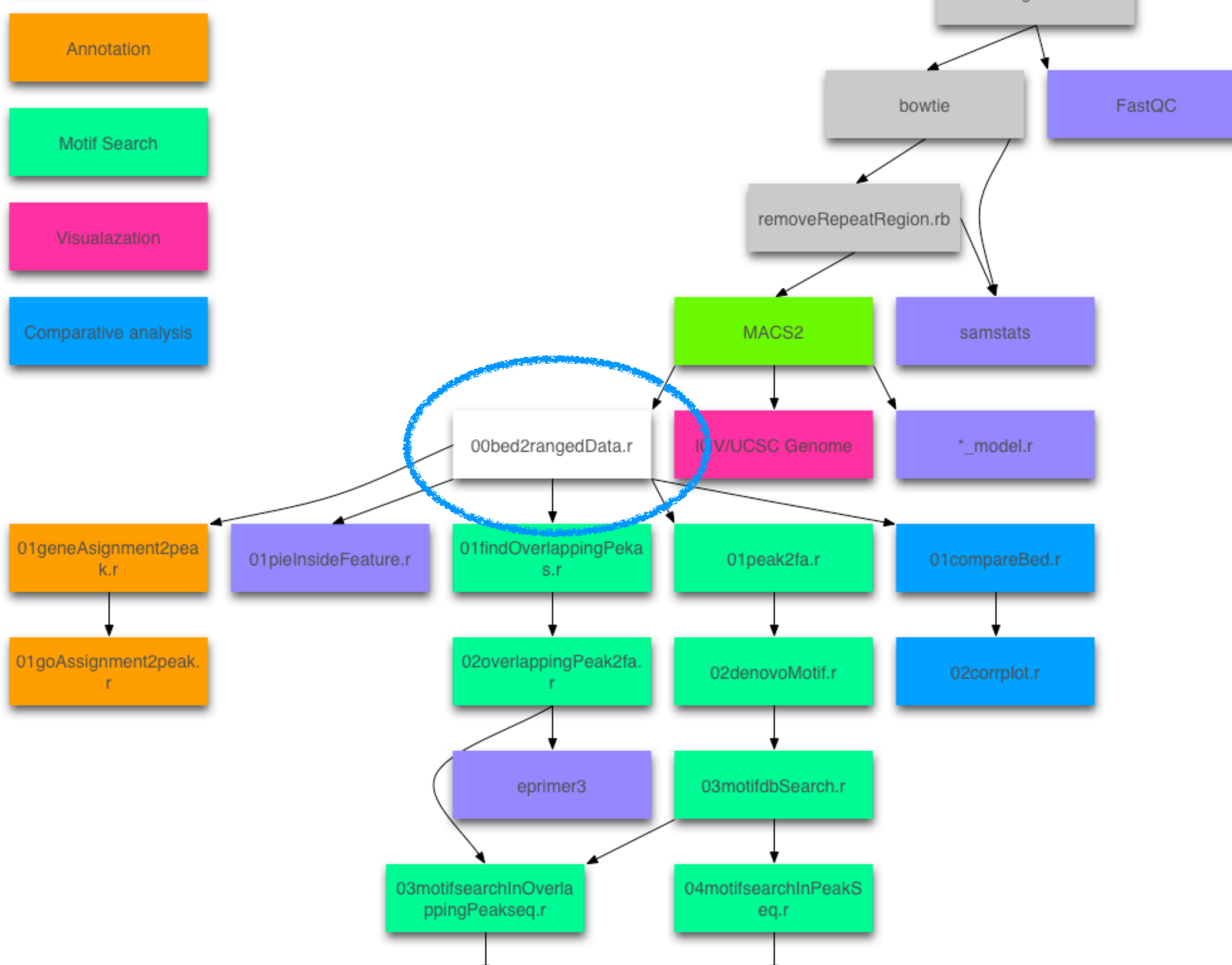
Peak caller. 入力は BED file

<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>
Peak ↓	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

space = chromosome

Start of peaks

End of peaks

# From BED file to RangedData Object

```
library("ChIPpeakAnno")
```

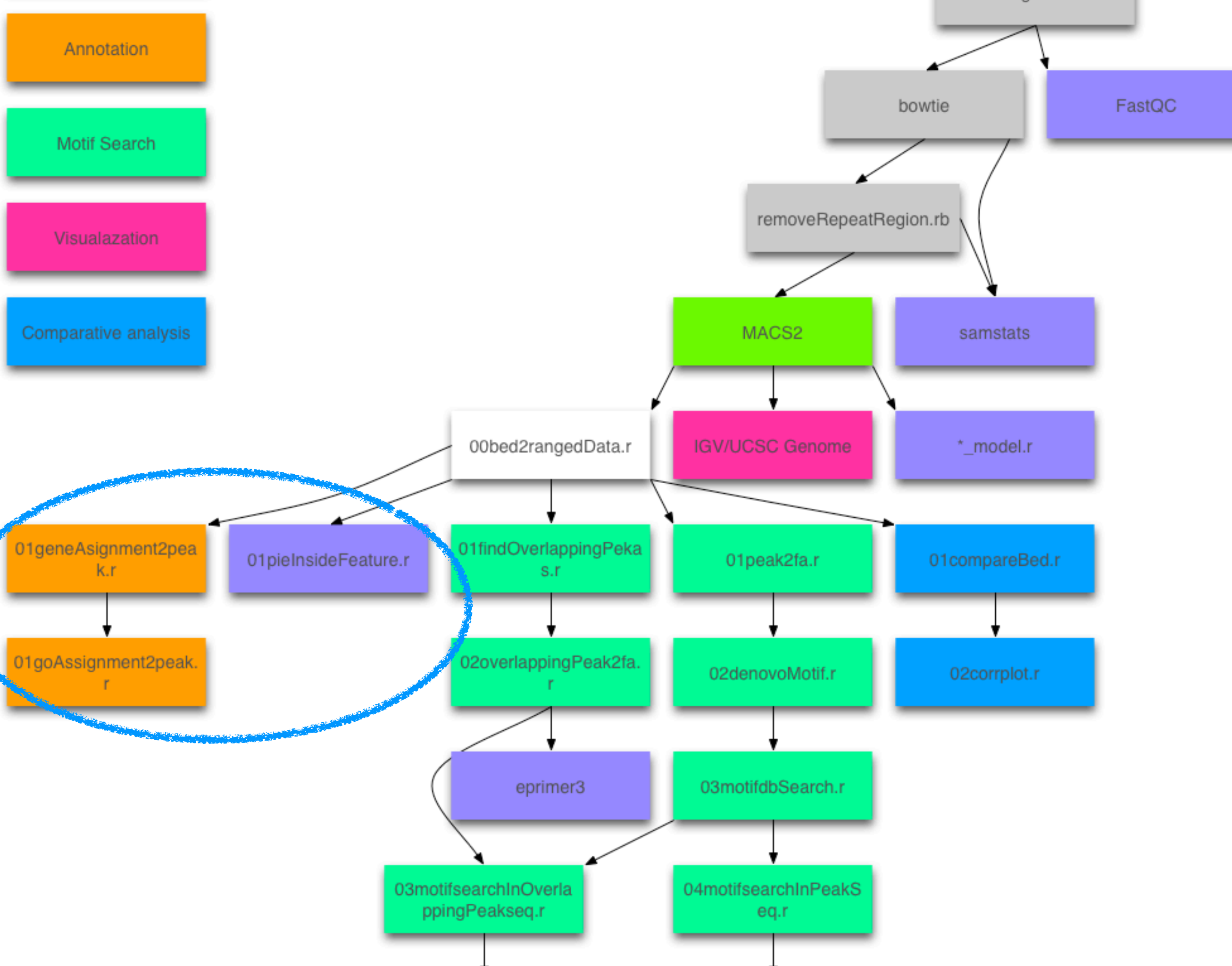
```
sox2.df <- read.table("results/macs2/Sox2_peaks.bed", header = FALSE)
```

```
oct4.df <- read.table("results/macs2/Oct4_peaks.bed", header = FALSE)
```

```
sox2.gr <- BED2RangedData(sox2.df, header = FALSE)
```

```
oct4.gr <- BED2RangedData(oct4.df, header = FALSE)
```

```
save(list=ls(), file = "results/3rd/00bed2rangedData.rdat")
```



# Gene assignment and Pie chart of peaks inside feature

```
library("ChIPpeakAnno")
data(TSS.mouse.NCBIM37)

load("results/3rd/00bed2rangedData.rdat")

oct4.anno <- annotatePeakInBatch(
  oct4.gr,
  AnnotationData = TSS.mouse.NCBIM37,
  output = "both"
)

write.table(
  as.data.frame(oct4.anno),
  file = "results/3rd/Oct4_peaks_anno.txt",
  row.names = F,
  col.names = T,
  quote = F,
  sep = "\t"
)

save(list=ls(), file = "results/3rd/01geneAssignment2peak.rdat")
```

# Gene assignment and Pie chart of peaks inside feature

```
library("ChIPpeakAnno")
```

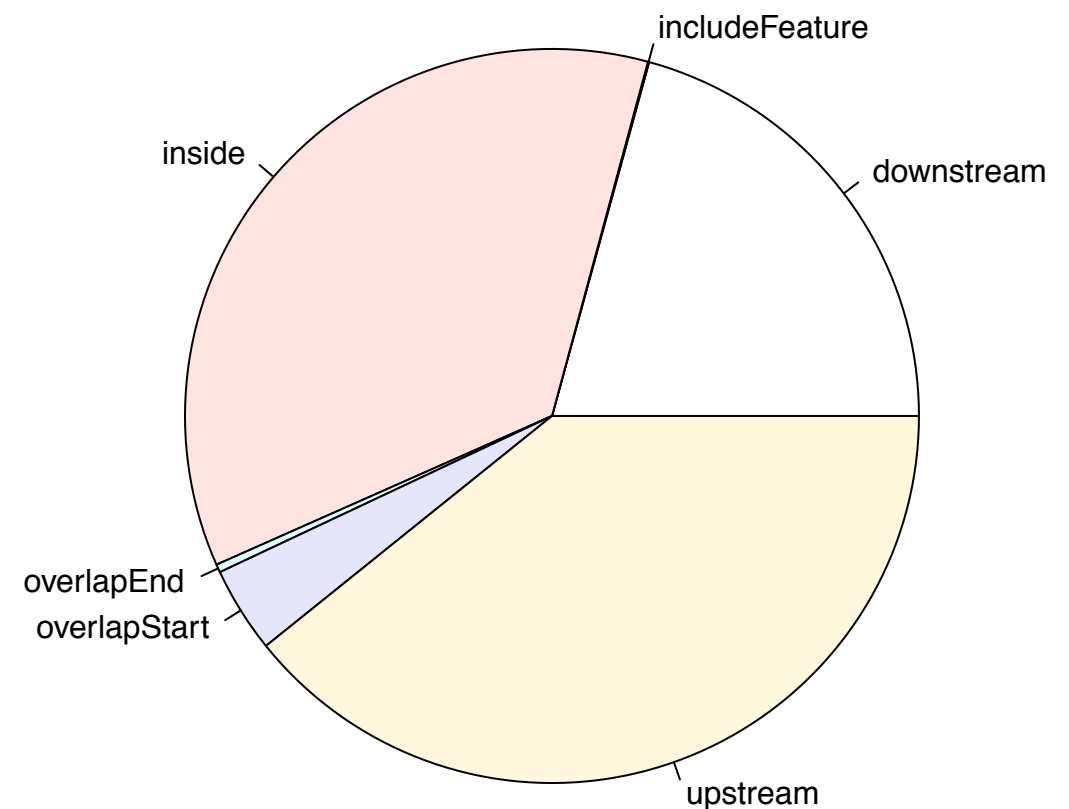
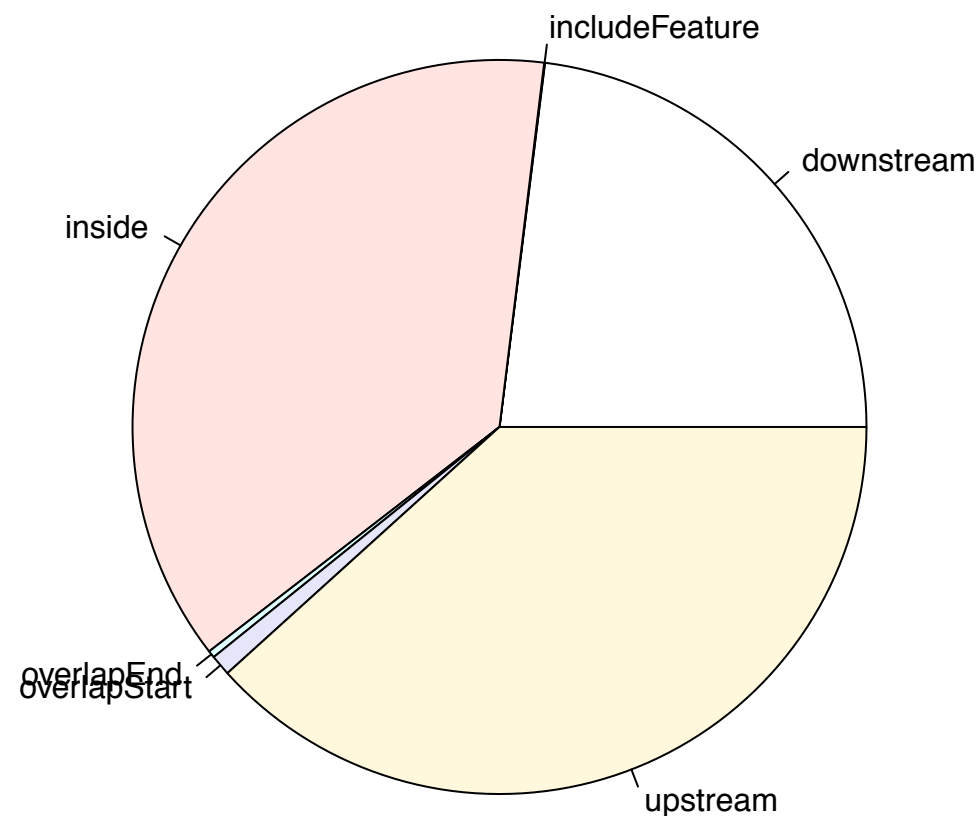
```
load("results/3rd/01geneAssignment2peak.rdat")
```

```
pdf("results/3rd/01pieInsideFeature.pdf")
```

```
pie(table(as.data.frame(sox2.anno)$insideFeature))
```

```
pie(table(as.data.frame(oct4.anno)$insideFeature))
```

```
dev.off()
```



# Gene Ontology assignment

```
library("ChIPpeakAnno")  
library(org.Mm.eg.db)
```

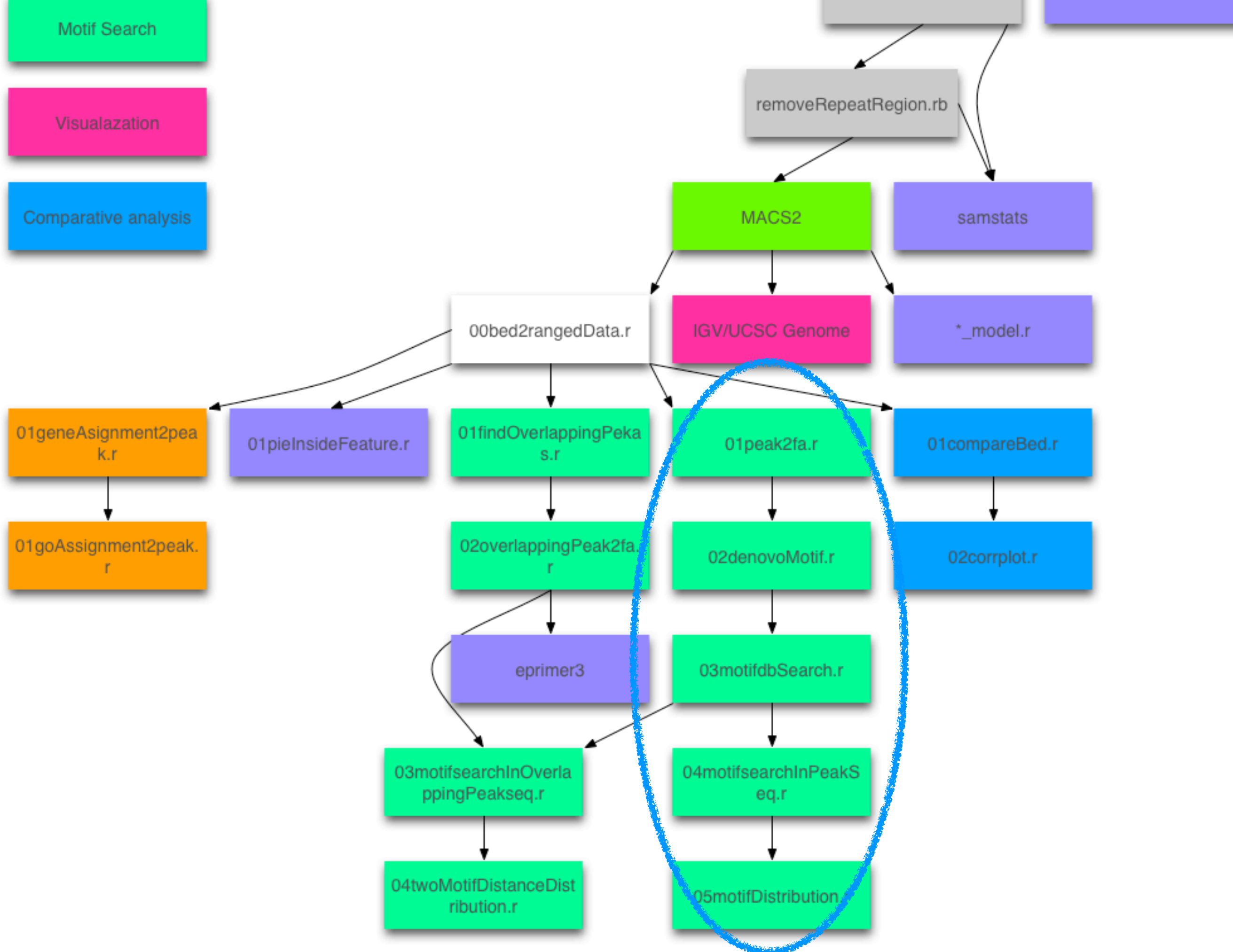
```
load("results/3rd/01geneAssignment2peak.rdat")
```

```
oct4.go <- getEnrichedGO(  
  oct4.anno,  
  orgAnn      = "org.Mm.eg.db",  
  maxP        = 0.01,  
  multiAdj    = TRUE,  
  minGOTerm   = 10,  
  multiAdjMethod = "BH"  
)
```

```
oct4.bp.goterm <- unique(oct4.go$bp[order(oct4.go$bp[,10]), c(2,10)])  
oct4.cc.goterm <- unique(oct4.go$cc[order(oct4.go$cc[,10]), c(2,10)])  
oct4.mf.goterm <- unique(oct4.go$mf[order(oct4.go$mf[,10]), c(2,10)])
```

```
save(list=ls(), file = "results/3rd/03goAssignment2peak.rdat")
```





# Get all peak sequences

```
library("ChIPpeakAnno")
library("BSgenome.Mmusculus.UCSC.mm9")

load("results/3rd/00bed2rangedData.rdat")

oct4.peaksWithSeqs <- getAllPeakSequence(
  oct4.gr,
  upstream    = 0,
  downstream  = 0,
  genome      = Mmusculus
)

write2FASTA(oct4.peaksWithSeqs, file="results/3rd/oct4.peaksWithSeqs.fa")
save(list=ls(), file = "results/3rd/01peak_fa.rdat")

$ less results/3rd/oct4.peaksWithSeqs.fa
>MACS_peak_1
TTCTTTCCTCCTTTGTACCCTGGGCGCTATAGGAATTCAACTTTACAAGCTGTGAGGAAATGGTGATTCTTGTGCA
AAGT
GAACAGCTGGGTCTGTCAACAGAAGGTAGCATTCTTTGATACTGAGCCTTCCTGGTGTGGCA
>MACS_peak_2
```

# De novo motif discovery

```
library("rGADEM")
library("BSgenome.Mmusculus.UCSC.mm9")

oct4.seqs <- read.DNAStringSet("results/3rd/oct4.peaksWithSeqs.fa", "fasta")
oct4.motif <- GADEM(oct4.seqs, verbose=1, genome = Mmusculus)

save(list=ls(), file = "results/3rd/02denovoMotif.rdat")
```

## Data structure:

### 0. parameters

### 1. motiflists: list of motif object (PWM, consensus)

```
## number of motifs
nMotifs(oct4.motif)
## get position weight matrix
getPWM(oct4.motif)
## get consensus sequences
consensus(oct4.motif)
```

# Motif database search

```
library("ChIPpeakAnno")
library("MotIV")

load("results/3rd/02denovoMotif.rdat")

## prep. database
path <- system.file(package="MotIV")
jaspar      <- readPWMfile(file.path(path, "extdata/jaspar2010.txt"))
jaspar.scores <- readDBScores(
  file.path(path, "extdata/jaspar2010_PCC_SWU.scores")
)
```

# Motif database search

```
## search motifs
oct4.motif.pwms <- getPWM(oct4.motif)
oct4.jaspar <- motifMatch(
  inputPWM = oct4.motif.pwms,
  align = "SWU",
  cc = "PCC",
  database = jaspar,
  DBscores = jaspar.scores,
  top = 5
)

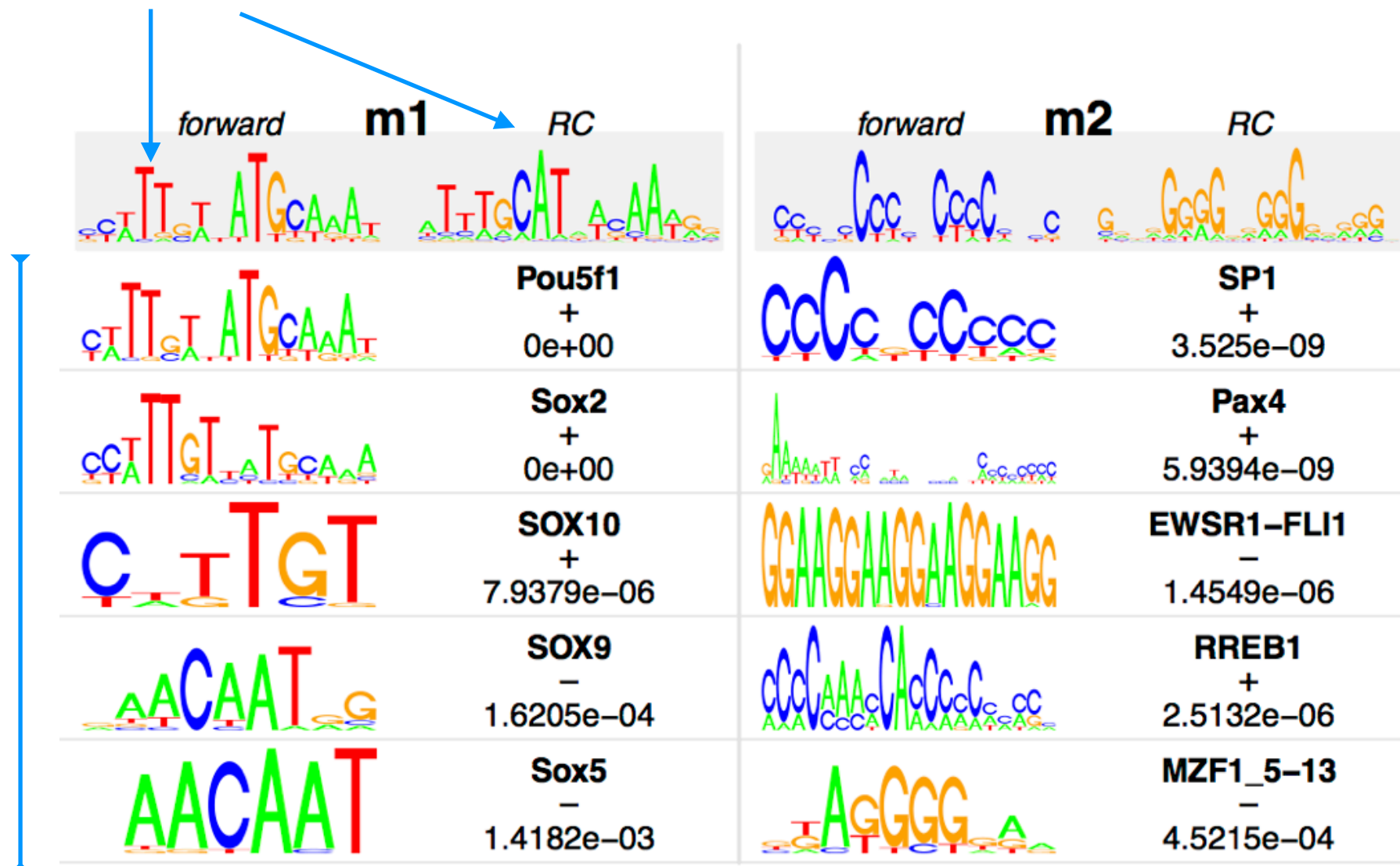
save(list=ls(), file="results/3rd/03motifdbSearch.rdat")

## output results with sequence logos
pdf("results/3rd/03motifdbSearch.pdf")
plot(oct4.jaspar, ncol = 2, top = 5, rev = FALSE,
     main = "Oct4 ChIP-seq", bysim = TRUE)
dev.off()
```

# Motif database search

## Oct4 ChIP-seq

Query motif (de novo motif)



Subject motifs in motif database

# Motif search in peak sequences

```
library("ChIPpeakAnno")
library("MotIV")
load("results/3rd/03motifdbSearch.rdat")
```

```
oct4.motif.pwms <- getPWM(oct4.motif)
```

```
oct4.motif.search.fwd <- lapply(
  oct4.seqs,
  function(x) {
    matchPWM(oct4.motif.pwms$m1, x)
  }
)
```

```
oct4.motif.search.fwd.score <- lapply(
  oct4.motif.search.fwd,
  function(hits) {
    PWMscoreStartingAt(
      oct4.motif.pwms$m1,
      subject(hits),
      start(hits)
    )
  }
)
```

```
save(list = ls(), file = "results/3rd/04motifsearchInPeakseq.rdat")
```

注意: `matchPWM` は片方の鎖しか検索しないので、PWMを `reverseComplement()` して検索する必要がある。

# Motif distribution

```
library("ChIPpeakAnno")
library("MotIV")

load("results/3rd/01peak_fa.rdat")
load("results/3rd/04motifsearchInPeakseq.rdat")

## Load summit data
oct4.sm.df <- read.table("results/macs2/Oct4_summits.bed", header=F)
oct4.sm.gr <- BED2RangedData(oct4.sm.df, header=FALSE)

## calc. distance between motif event and summit
num.peaks <- length(oct4.motif.search.fwd)
oct4.motif.summit.dist <- rep(0, num.peaks)

summit.on.peak <- start(oct4.sm.gr) - start(oct4.gr)
```



# Motif distribution

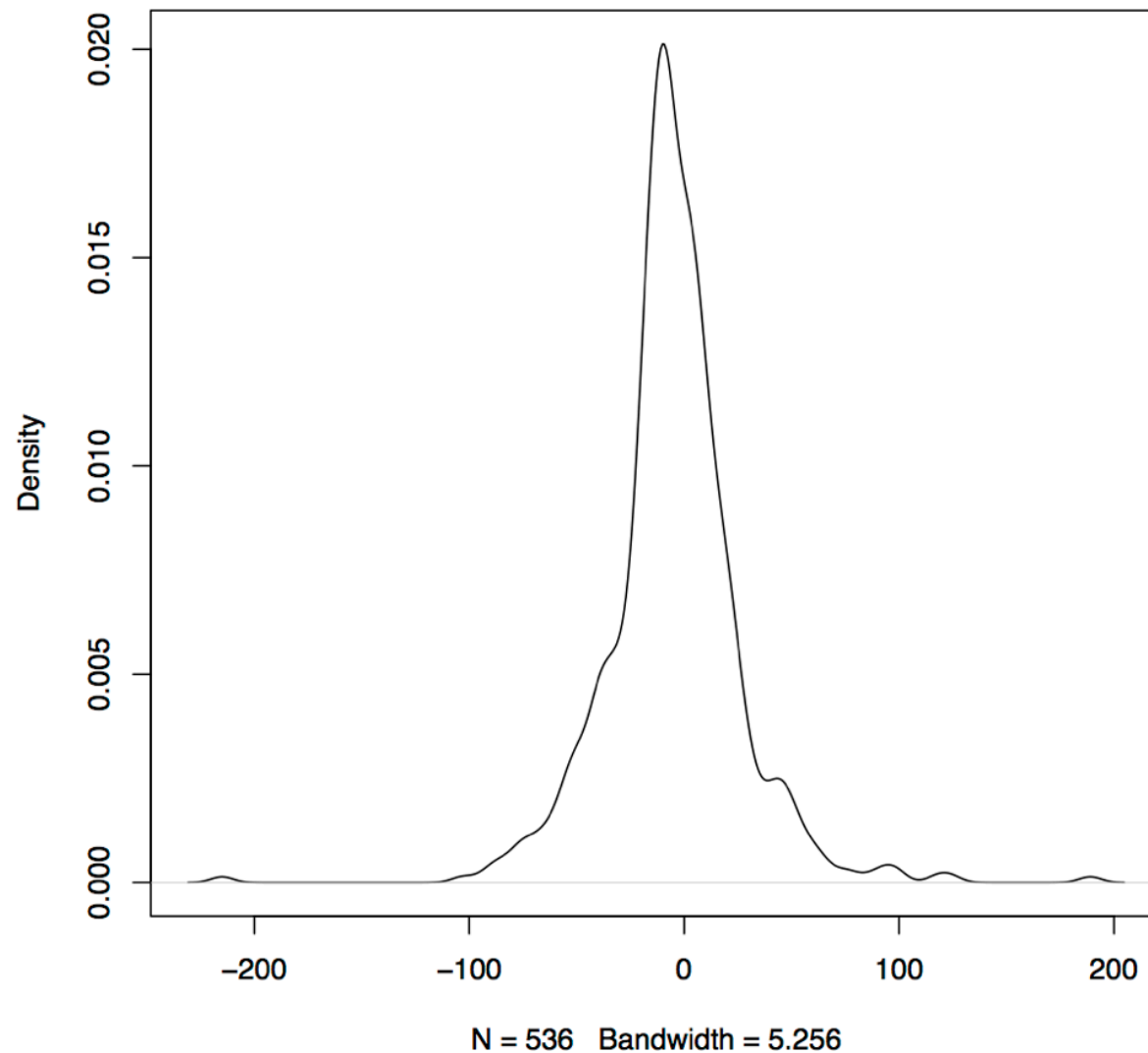
```
num.peaks <- length(oct4.motif.search.fwd)
num <- 0
oct4.motif.summit.dist <- 0
for (peak in 1:num.peaks) {
  if (! identical(start(oct4.motif.search.fwd[[peak]]), integer(0))) {
    dists <- start(oct4.motif.search.fwd[[peak]]) - summit.on.peak[peak]

    if (length(dists) == 1) {
      num <- num + 1
      oct4.motif.summit.dist[num] <- dists
    } else if (length(dists) > 1) {
      for(i in 1:length(dists)) {
        num <- num + 1
        oct4.motif.summit.dist[num] <- dists[i]
      }
    }
  }
}
save(list = ls(), file = "results/3rd/05motifDistribution.rdat")

pdf("results/3rd/05motifDistribution.pdf")
plot(density(oct4.motif.summit.dist), main="Oct4 / Motif distribution (m1)")
dev.off()
```

# Motif distribution

Oct4 / Motif distribution (m1)

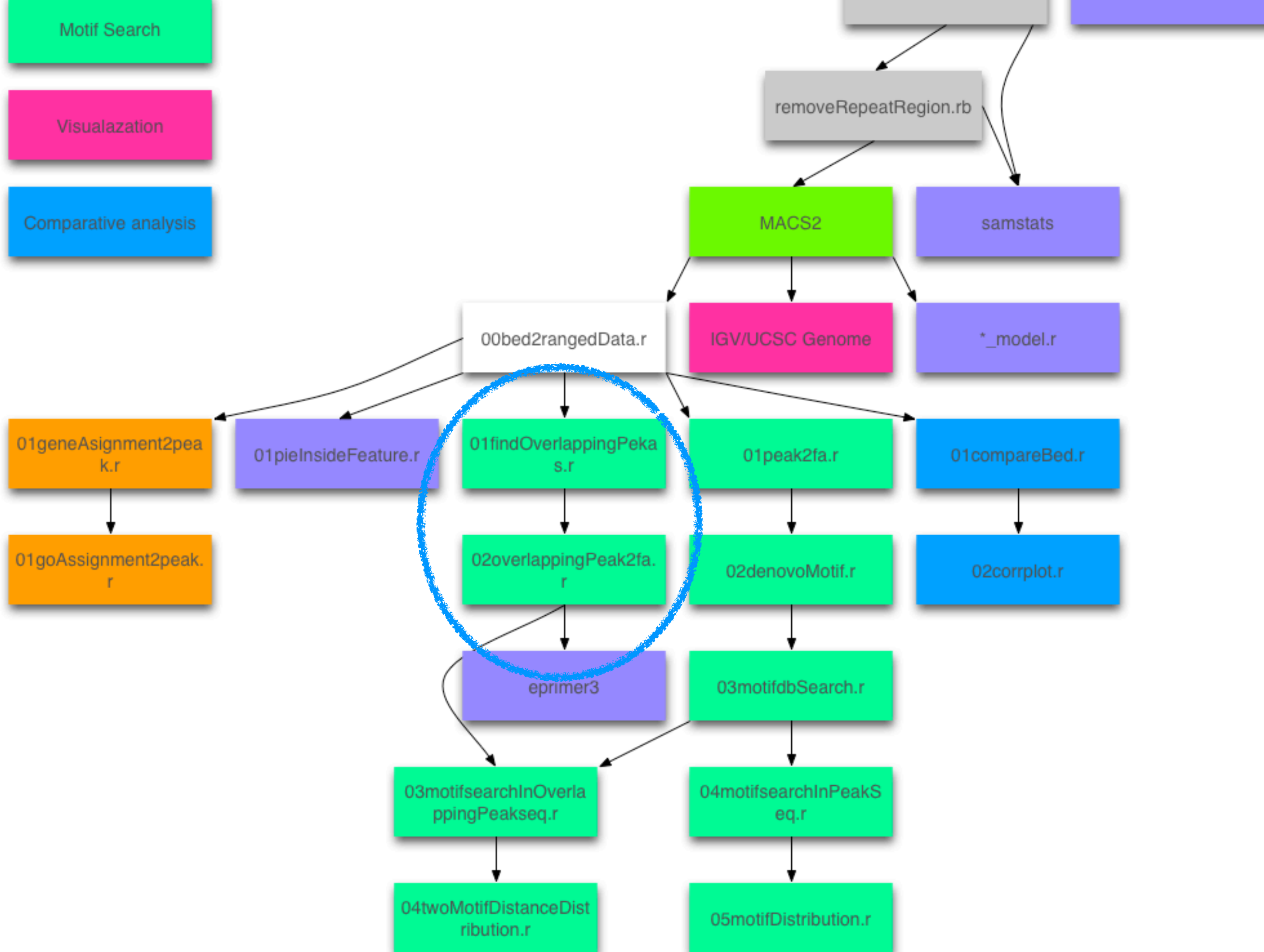


プロットの見方：  
もし発見されたモチーフが陽性の場合、0付近にピークを持つ分布になる。偽陽性だった場合は、一様分布になる。

異なる転写因子のChIP-seqから得られたモチーフ間の距離をプロットした場合は二峰性になる。ただし2つの転写因子がヘテロダイマーとして働く場合は、単峰性に近くなる

summit から motif までの距離

モチーフが発見された頻度



# Finding overlapping peaks between two different TF ChIP-seq

```
library("ChIPpeakAnno")

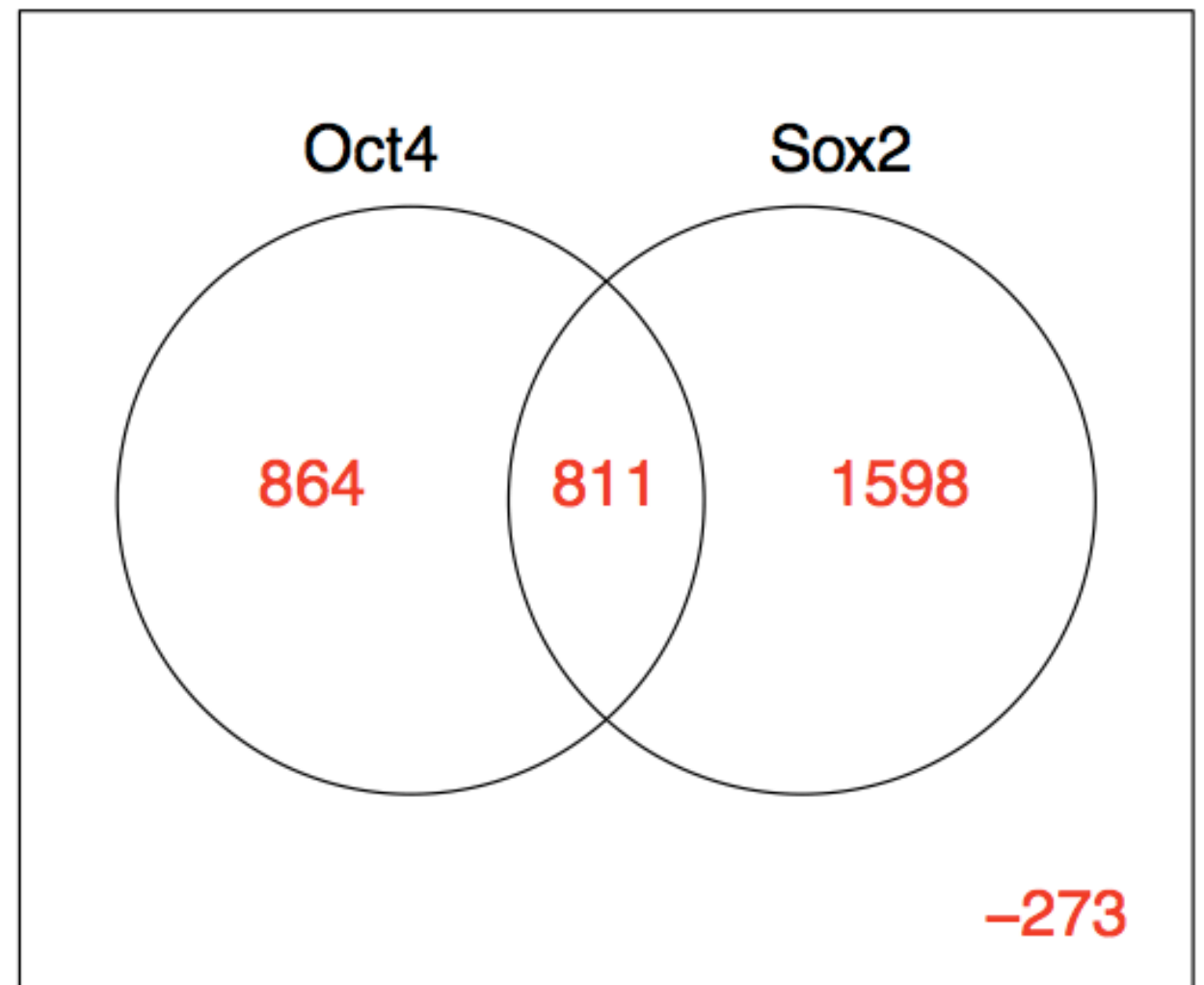
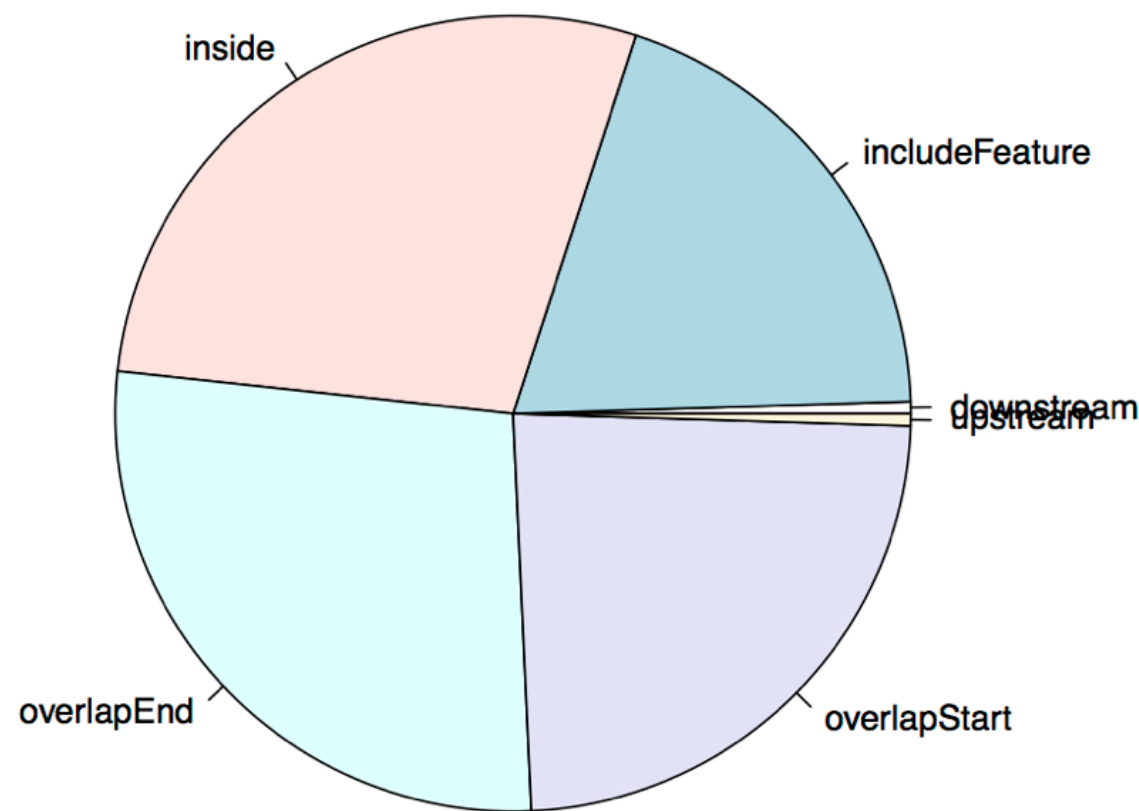
load("results/3rd/00bed2rangedData.rdat")

oct4.sox2.overlap <- findOverlappingPeaks(oct4.gr, sox2.gr, multiple=T)

save(list=ls(), file = "results/3rd/01findOverlappingPeaks.rdat")

pdf("results/3rd/01findOverlappingPeaks.pdf")
pie( table(oct4.sox2.overlap$OverlappingPeaks$overlapFeature) )
oct4.sox2.overlap.count <- makeVennDiagram(
  RangedDataList(oct4.gr, sox2.gr),
  NameOfPeaks = c("Oct4", "Sox2"),
  totalTest    = 3000,
)
dev.off()
```

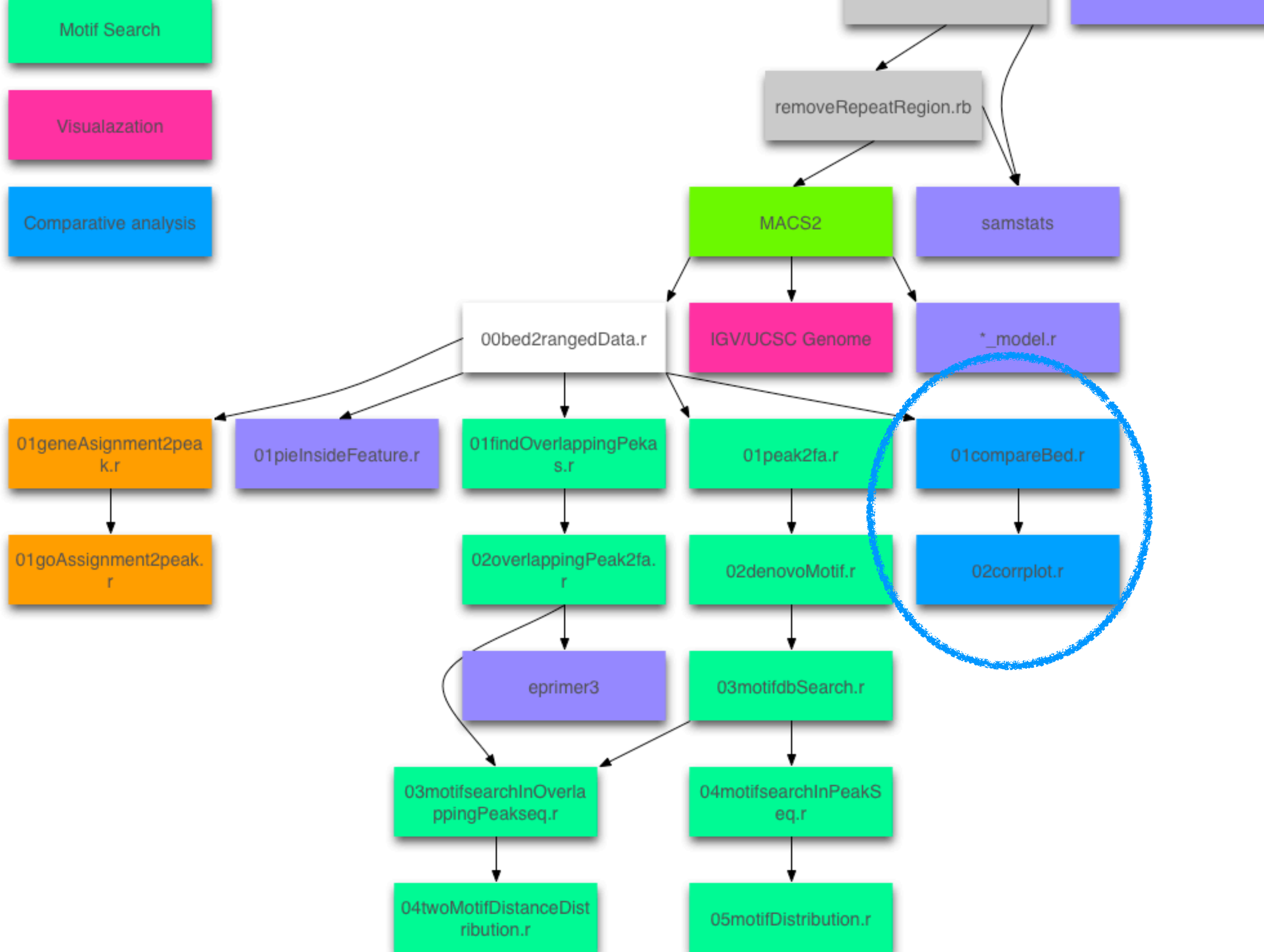
# Finding overlapping peaks between two different TF ChIP-seq



Venn diagram の描画には、R package の VennDiagram や Google Chart API の利用をお薦めする。

<http://cran.r-project.org/web/packages/VennDiagram/>

<http://code.google.com/intl/ja/apis/chart/>



# Quantitative comparison of different TF ChIP-seq experiments

```
library("QuGAcomp")
library("corrplot")

genome.length.file <- file.path(system.file(package="QuGAcomp"), "data",
"mm9.info")

oct4.bed.file <- file.path(
  system.file(package="QuGAcomp"),
  "data",
  "GSM288346_ES_Oct4.mm9.header.bed"
)

oct4.gr <- loadBedFile(oct4.bed.file, genome.length.file)

oct4.fat <- fat(oct4.gr, 200)
oct4.unistd <- unifyStrand(oct4.fat)
oct4.cov <- coverage(oct4.unistd)
oct4.bin500 <- lapply( oct4.cov, function(x) rleBinning(x, 500) )
oct4.bin500 <- flatRleList(oct4.bin500)
```

# Quantitative comparison of different TF ChIP-seq experiments

```
quga.oct4.soxx2 <- qugacomp(oct4.bin500, soxx2.bin500)
quga.oct4.nanog <- qugacomp(oct4.bin500, nanog.bin500)
quga.soxx2.nanog <- qugacomp(soxx2.bin500, nanog.bin500)
```

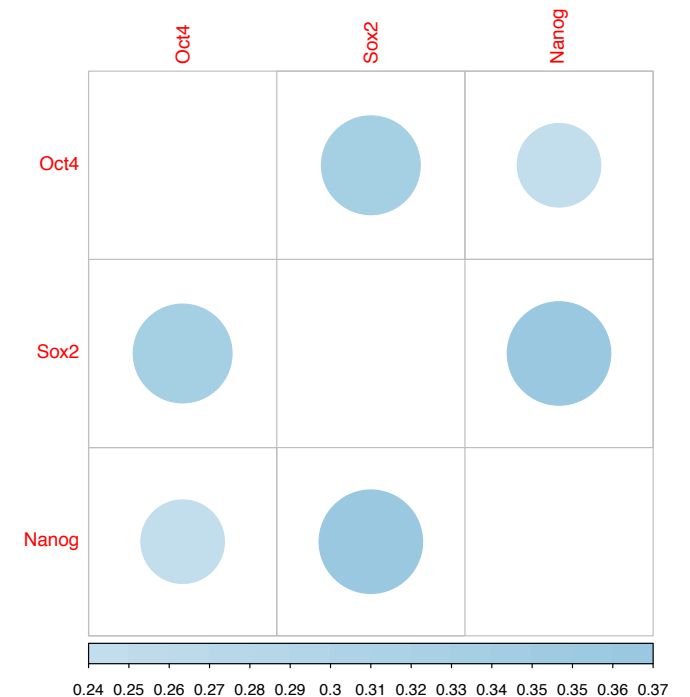
```
num <- 3
mat.cor <- matrix(0, nrow=num, ncol=num)
rownames(mat.cor) <- c("Oct4", "Soxx2", "Nanog")
colnames(mat.cor) <- c("Oct4", "Soxx2", "Nanog")
```

```
diag(mat.cor) <- rep(1, num)
```

```
mat.cor[1,2] <- mat.cor[2,1] <- pearsonCoef(quga.oct4.soxx2)
mat.cor[1,3] <- mat.cor[3,1] <- pearsonCoef(quga.oct4.nanog)
mat.cor[2,3] <- mat.cor[3,2] <- pearsonCoef(quga.soxx2.nanog)
```

```
mat.cor.max <- max( mat.cor[upper.tri(mat.cor, diag=F)] )
mat.cor.min <- min( mat.cor[upper.tri(mat.cor, diag=F)] )
```

```
pdf("corrplot.pdf")
corrplot(
  mat.cor, method="circle", type="full", diag=FALSE, outline=FALSE,
  addcolorlabel="bottom", cl.lim=c(mat.cor.min, mat.cor.max), cl.ratio=0.2
)
dev.off()
```





# 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

# 課題

- このコードでは転写因子の数が増えるたびにコードが重複する。任意のデータを指定して、実行できるようプログラムを修正せよ。ヒント: `function()`, `commandArgs()`
- モチーフ間距離の分布を描画せよ (githubに回答あり)
- Oct4, Sox2 以外の転写因子のデータを利用し、モチーフ間距離の分布を計算しプロットせよ。Sox2のように Oct4とヘテロダイマーで動作することが疑われる転写因子があるか?
- これらのRのプログラムを順々に実行するプログラムを作成せよ。ヒント: `Make`, `Rake`, シェルスクリプトなどを利用する
- 紹介したパイプラインに足りないステップはなにか? `repeatmask`
- 異なるPeak caller の結果を比較せよ。ヒント: `findOverlappingPeaks()`
- RNAの転写量と結合量を調べよ。ヒント: `plot()`, `cor()`



## RとBioconductorを用いたバイオインフォマティクス [単行本]

R.ジェントルマン (著), R. ジェントルマン (編集), V.J. カリー (編集), W. フーバー (編集), R.A. イリザリー (編集), S. ドュドイト (編集), 荒川和晴 (翻訳), 粕川雄也 (翻訳), 川路英哉 (翻訳), 河野 信 (翻訳), 神田将和 (翻訳), 鈴木治夫 (翻訳), 田中伸也 (翻訳), 中尾光輝 (翻訳), 長嶋剛史 (翻訳), 二階堂 愛 (翻訳), 宮本真理 (翻訳)

★★★★★ ☒ (1 カスタマーレビュー) いいね (0)

この本は現在お取り扱いできません。 [在庫状況](#)について



## オープンソースで学ぶバイオインフォマティクス [単行本]

オープンバイオ研究会 ☒ (編集)

★★★★★ ☒ (3件のカスタマーレビュー) いいね (2)

[出品者](#)からお求めいただけます。

新品の出品 : 1¥ 4,095より    中古品の出品 : 8¥ 3,323より

<http://blog.hackingisbelieving.org/>

# NGS現場の会

## 第二回研究会 in 大阪

NGS-Genbanokai II  
next-generation →



2012年 5/24 – 25

+ 前日チュートリアル (5/23)

ホテル阪急エキスポパーク

(大阪府吹田市)

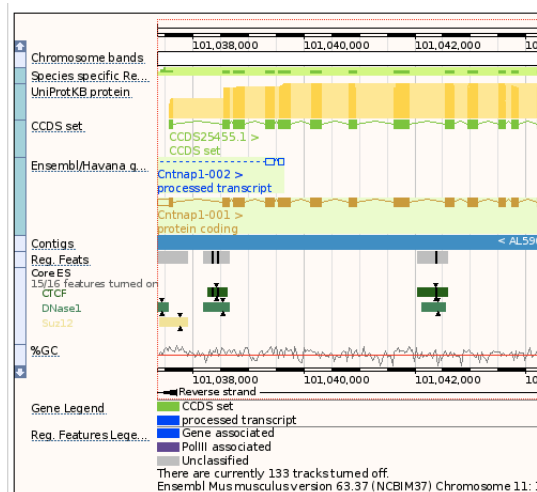
3月から申込受付開始

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

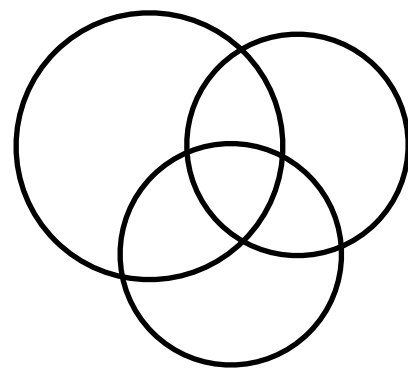
# QuGAcomp: 定量的なゲノムアノテーション比較ツール

Itoshi NIKAIDO, Ph.D.  
RIKEN CDB@Kobe

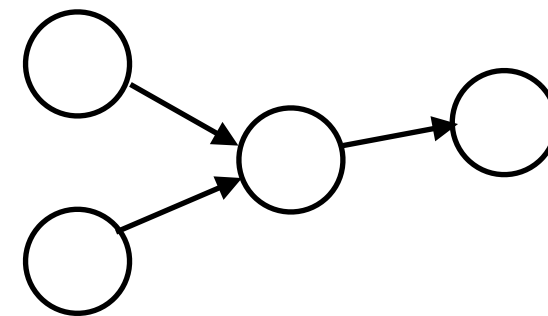
# 比較・統合の現状



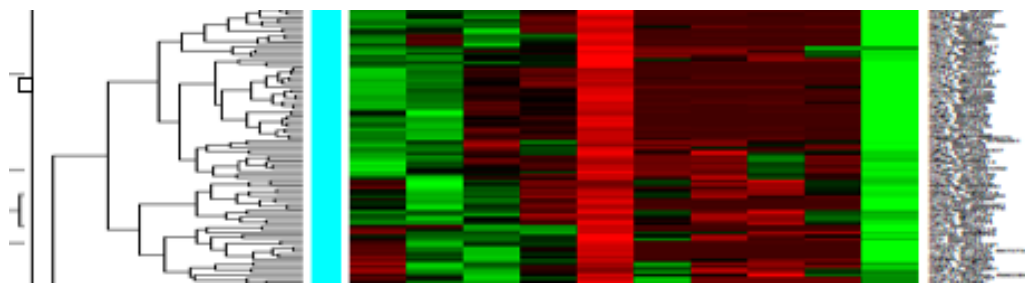
Hyperlink



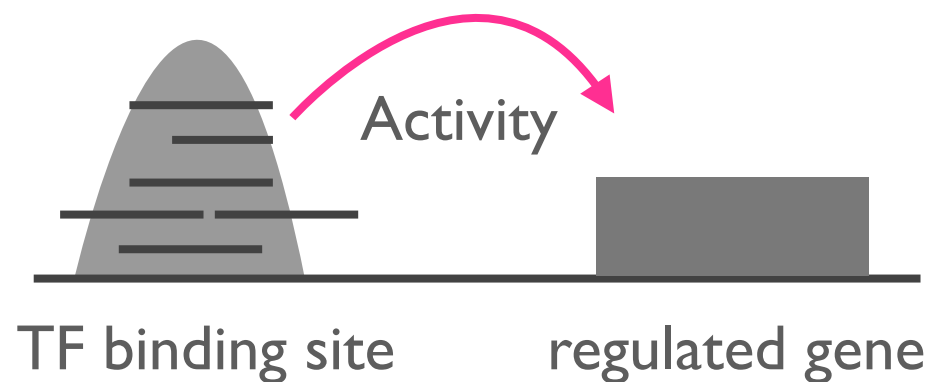
venn diagram



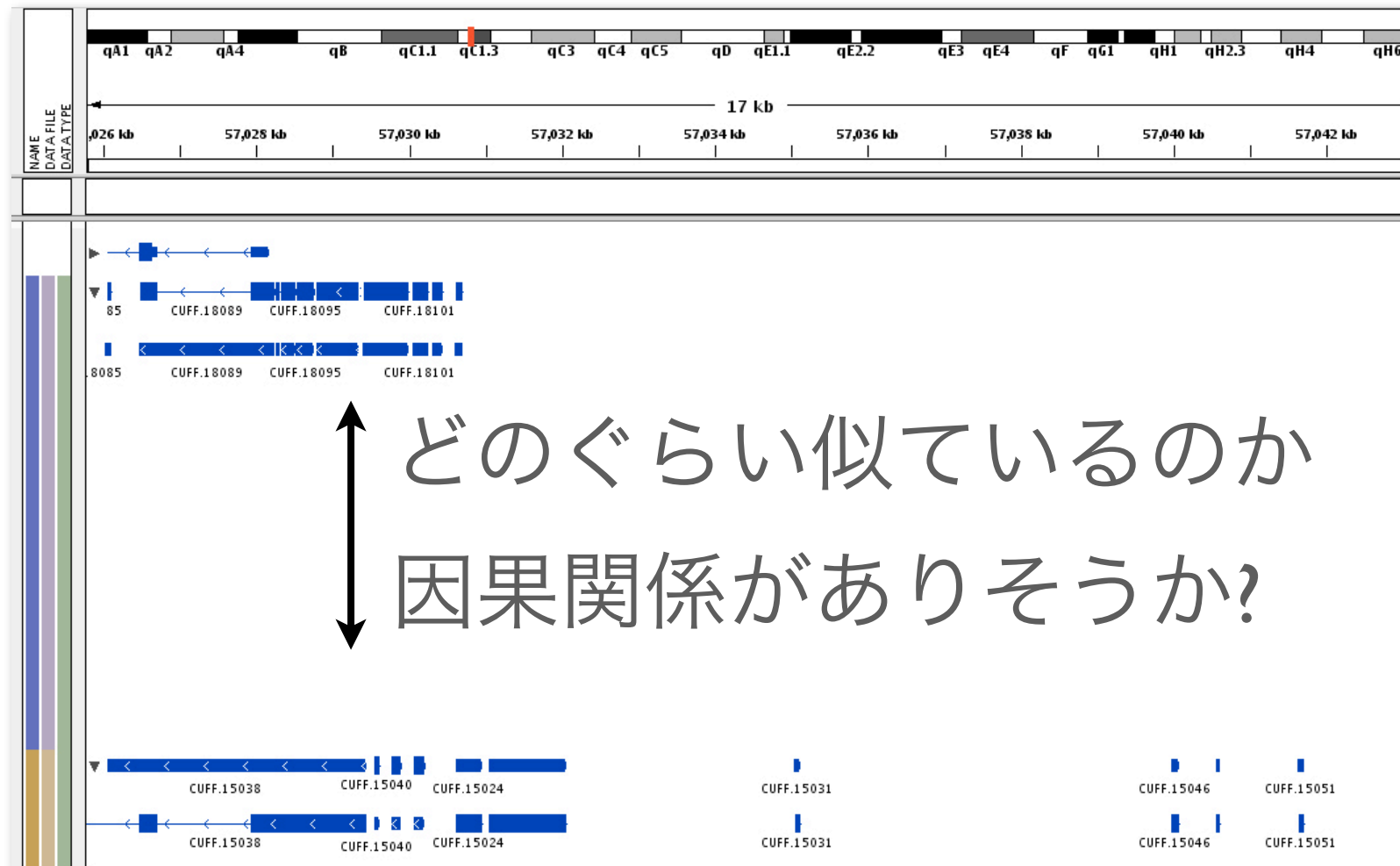
directed graph



heatmap



# ゲノムアノテーションの類似の定量化 ツールの開発



例:  
転写因子結合と発現  
メチル化と発現  
転写因子群

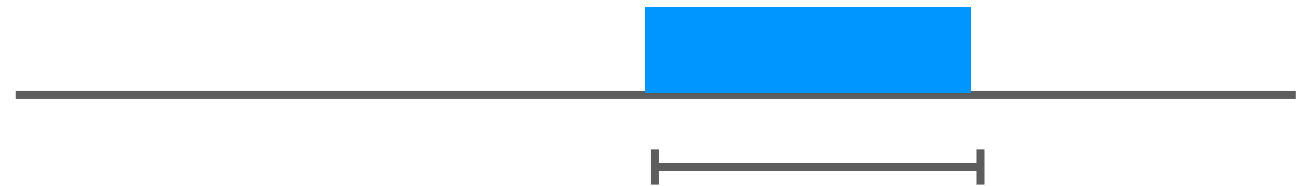
```
track name=Oct4KO_Day0 description="Day0, Sox2 ChIP-seq" useScore=0
chr1 3052831 3053414 s_1_s_8_187941 9.691 + 3053119 3053120
chr1 3472777 3473463 s_1_s_8_188786 13.277 + 3473103 3473104
chr1 4140902 4141904 s_1_s_8_180770 15.66 + 4141197 4141198
chr1 4588454 4588854 s_1_s_8_183450 9.953 + 4588647 4588648
chr1 4617652 4618346 s_1_s_8_183614 10.84 + 4617934 4617935
chr1 4792575 4793008 s_1_s_8_184913 25.682 + 4792755 4792756
```

# Two type of genome annotation

gene / region-centric comparison

gene-centric = 転写単位で比較すればよい

RNA-seq/Exome



region-centric = 比較する領域を決める必要がある

BS-seq



ChIP/MBD-seq





# Three topics

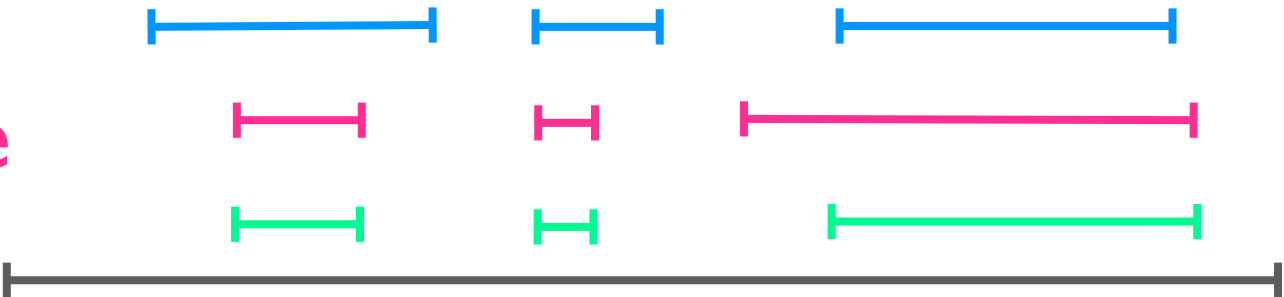
- ゲノムアノテーションのデータ構造
- データ比較のための距離・統計量
- データ比較のためのツール

Range data  
(Interval data)

domain	4	8	query1	9
domain	10	12	query2	3
domain	14	20	query3	5
domain	5	7	ref1	8
domain	10	10	ref2	2
domain	13	29	ref3	9

query

reference



overlap

domain

localization vector

00011111011101111110

00001110010011111111

co-localization vector

000011100100001111110

Contingency Table

query

reference




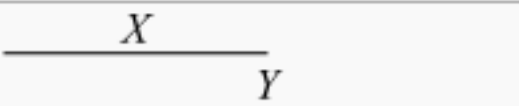
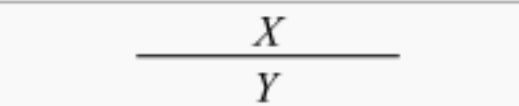
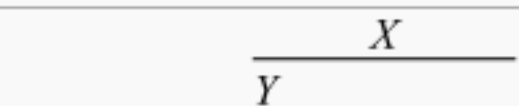
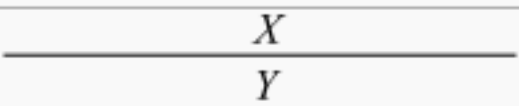
	0	1	sum
0	4	1	6
1	4	10	14
sum	8	12	20

# Allen's interval algebra

## 区間代数

ブール代数の一種

- 区間の演算
- 2つの区間の関係
- 時空間推論

Relation	Illustration	Interpretation
$X < Y$ $Y > X$		X takes place before Y
$X m Y$ $Y m i X$		X meets Y ( <i>i</i> stands for <i>inverse</i> )
$X o Y$ $Y o i X$		X overlaps with Y
$X s Y$ $Y s i X$		X starts Y
$X d Y$ $Y d i X$		X during Y
$X f Y$ $Y f i X$		X finishes Y
$X = Y$		X is equal to Y

# Run length encoding

連長圧縮

00011111011101111110 = 20 bit

0{3}1{5}01{3}1{7}0 = 10 bit

Human genome = 3G base = 3G bit

```
> query.rle <- rle(query)
> query.rle
Run Length Encoding
  lengths: int [1:7] 3 5 1 3 1 6 1
  values  : num [1:7] 0 1 0 1 0 1 0
> inverse.rle(query.rle)
[1] 0 0 0 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 0
```

# Correlation

## Pearson / Spearman correlation coefficient

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

## localization vector

```
00011111011101111110
00001110010011111111
```

## co-localization vector

```
0000111001000011111110
```

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

			sum
			6
			14
sum	8	12	20

# Edit distance

Levenshtein distance

domain	4	8	query1	9.6
domain	10	12	query2	13.2
domain	14	20	query3	8.0
domain	5	7	ref1	14.0
domain	9	10	ref2	3.2
domain	13	29	ref3	9.8



localization vector

```
00011111011101111110
00001110010011111111
```

co-localization vector

```
000011100100001111110
```

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

欠損・挿入・置換などが考えられる場合には有効

重みをつけることで、non-localization region にひっぱられない

			sum
		1	6
	1	10	14
sum	8	12	20

		-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
			0	0	0	0	1	1	1	0	0	1	0	0	1	1	1	1	1	1	1	1
-1		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	0	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
3	1	4	3	2	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
4	1	5	4	3	2	2	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	1	6	5	4	3	3	2	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14
6	1	7	6	5	4	4	3	2	1	2	3	3	4	5	6	7	8	9	10	11	12	13
7	1	8	7	6	5	5	4	3	2	2	3	3	4	5	5	6	7	8	9	10	11	12
8	0	9	8	7	6	5	5	4	3	2	2	3	3	4	5	6	7	8	9	10	11	12
9	1	10	9	8	7	6	5	5	4	3	3	2	3	4	4	5	6	7	8	9	10	11
10	1	11	10	9	8	7	6	5	5	4	4	3	3	4	4	4	5	6	7	8	9	10
11	1	12	11	10	9	8	7	6	5	5	5	4	4	4	4	4	4	5	6	7	8	9
12	0	13	12	11	10	9	8	7	6	5	5	5	4	4	5	5	5	5	6	7	8	9
13	1	14	13	12	11	10	9	8	7	6	6	5	5	5	4	5	5	5	5	6	7	8
14	1	15	14	13	12	11	10	9	8	7	7	6	6	6	5	4	5	5	5	5	6	7
15	1	16	15	14	13	12	11	10	9	8	8	7	7	7	6	5	4	5	5	5	5	6
16	1	17	16	15	14	13	12	11	10	9	9	8	8	8	7	6	5	4	5	5	5	5
17	1	18	17	16	15	14	13	12	11	10	10	9	9	9	8	7	6	5	4	5	5	5
18	1	19	18	17	16	15	14	13	12	11	11	10	10	10	9	8	7	6	5	4	5	5
19	0	20	19	18	17	16	15	14	13	12	11	11	10	10	10	9	8	7	6	5	5	6

# Correlation of contingency

Phi / Contingency C / Cramer V coefficient

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

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

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

$t = \min(k, m)$

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



# Test of independence

Chi-square / Fisher's exact test / Hypergeometric test

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

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

カイ二乗値がカイ二乗分布に従う  
localization vector の相関は高くないので注意

関連を言いたいので、帰無仮説を棄却する。P値が高いと関連する  
棄却した場合に、線形の関係があるとは限らない

期待値が1未満が1マス、5未満のマスが20%以下の場合には使うべきでない  
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

# Test of independence

Chi-square / Fisher's exact test / Hypergeometric test

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

	0	1	sum
0	a	b	a + b
1	c	d	c + d
sum	a + c	b + d	n

$$p = \Pr(T \geq 10 | H_0, m = 12, n = 8, k = 14) = 0.14$$

関連を言いたいので、帰無仮説を棄却する。P値が高いと関連する  
棄却した場合に、線形の関係があるとは限らない

Hypergeometric test = one-tail Fisher's exact test  
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

# Similarity of sample sets

Dice / Jaccard / Simpson index

共起頻度

$$F(X,Y)=|X \cap Y|$$

計算が容易で高速

相互情報量

$$-\log \frac{N|X \cap Y|}{|X||Y|}$$

空間情報を利用していない overlap

Dice 係数

$$\frac{|X \cap Y|}{|X| + |Y|}$$

Jaccard 係数

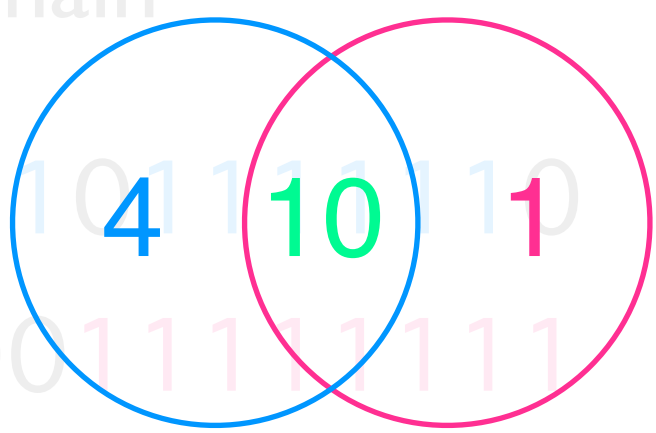
$$\frac{|X \cap Y|}{|X \cup Y|}$$

Simpson 係数

$$\frac{|X \cap Y|}{\min(|X|, |Y|)}$$

Cosine 係数

$$\frac{|X \cap Y|}{\sqrt{|X||Y|}}$$



Contingency Table

query

reference

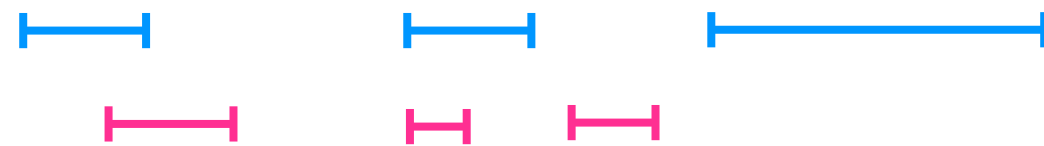
	0	1	sum
0	4	1	6
1	4	10	14
sum	8	12	20

# Preprocessing

## Peak / window-based comparison

Peak-based

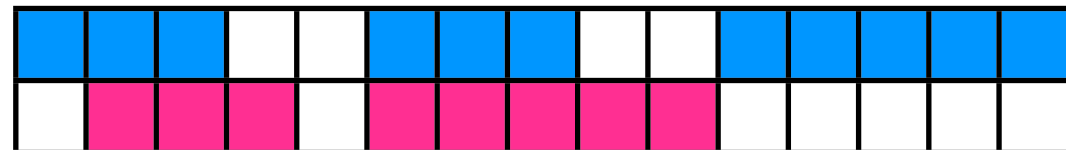
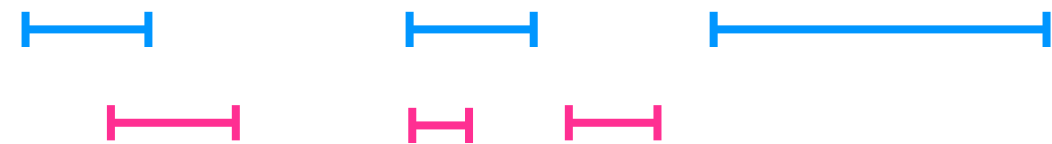
query  
reference



Peak の数や長さの影響を受ける

Window-based

query  
reference

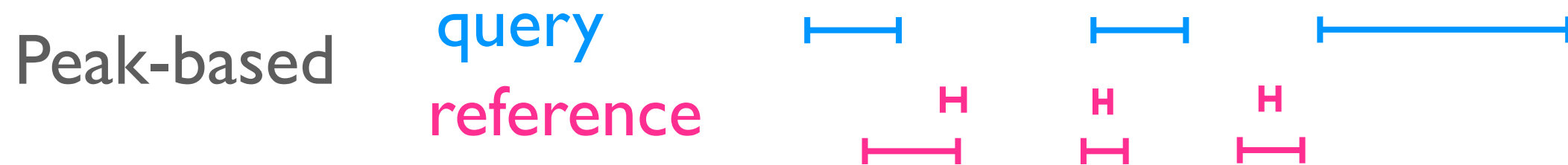


平滑化される

# 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

00099999033305555550

00008880020099999999

localization vector

00011111011101111110

00001110010011111111

# Implementation

User-friendly and comprehensive tool

- 既存ツール
  - Bedtools = Overlapping のカウントしかできない
  - ChIPseeqer = Jaccard Index のみ
- Quantitative Genome Annotation Comparison Tool (QuGAcomp) in R + Bioconductor
  - 多くの統計量・距離を用意
  - 様々な pre-processing が可能
  - 可視化

# Quantitative comparison of different TF ChIP-seq experiments

```
library("QuGAcomp")
library("corrplot")

genome.length.file <- file.path(system.file(package="QuGAcomp"), "data",
"mm9.info")

oct4.bed.file <- file.path(
  system.file(package="QuGAcomp"),
  "data",
  "GSM288346_ES_Oct4.mm9.header.bed"
)

oct4.gr <- loadBedFile(oct4.bed.file, genome.length.file)

oct4.fat <- fat(oct4.gr, 200)
oct4.unistd <- unifyStrand(oct4.fat)
oct4.cov <- coverage(oct4.unistd)
oct4.bin500 <- lapply( oct4.cov, function(x) rleBinning(x, 500) )
oct4.bin500 <- flatRleList(oct4.bin500)
```



# Quantitative comparison of different TF ChIP-seq experiments

```
quga.oct4.soxx2 <- qugacomp(oct4.bin500, soxx2.bin500)
quga.oct4.nanog <- qugacomp(oct4.bin500, nanog.bin500)
quga.soxx2.nanog <- qugacomp(soxx2.bin500, nanog.bin500)
```

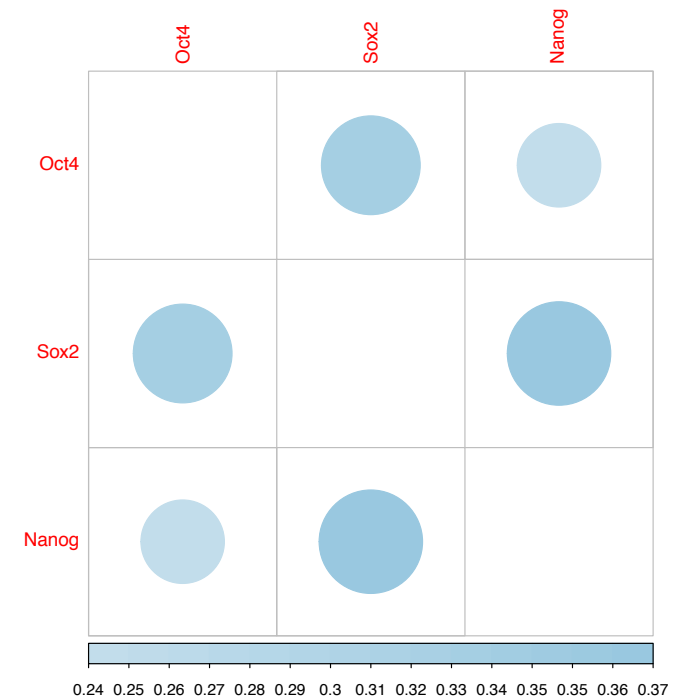
```
num <- 3
mat.cor <- matrix(0, nrow=num, ncol=num)
rownames(mat.cor) <- c("Oct4", "Soxx2", "Nanog")
colnames(mat.cor) <- c("Oct4", "Soxx2", "Nanog")
```

```
diag(mat.cor) <- rep(1, num)
```

```
mat.cor[1,2] <- mat.cor[2,1] <- pearsonCoef(quga.oct4.soxx2)
mat.cor[1,3] <- mat.cor[3,1] <- pearsonCoef(quga.oct4.nanog)
mat.cor[2,3] <- mat.cor[3,2] <- pearsonCoef(quga.soxx2.nanog)
```

```
mat.cor.max <- max( mat.cor[upper.tri(mat.cor, diag=F)] )
mat.cor.min <- min( mat.cor[upper.tri(mat.cor, diag=F)] )
```

```
pdf("corrplot.pdf")
corrplot(
  mat.cor, method="circle", type="full", diag=FALSE, outline=FALSE,
  addcolorlabel="bottom", cl.lim=c(mat.cor.min, mat.cor.max), cl.ratio=0.2
)
dev.off()
```



# ToDo

1. metrics 関数群をベクトル化
2. metrics 関数群のラッパーを実装
3. 可視化ツールのインテグレーション