bPeaks: a R package to perform ChIP-seq peak calling

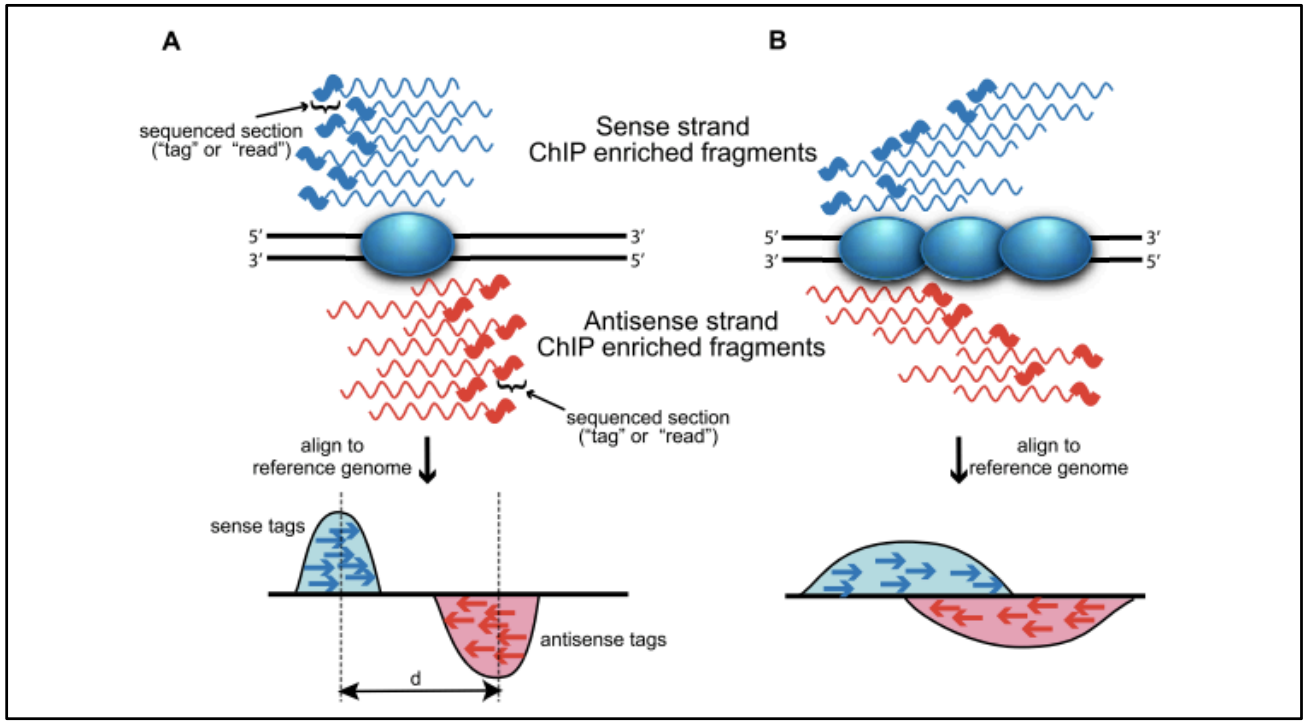
How to detect transcription factor binding sites from ChIP-seq data in small eukaryotic genomes ?

Merhej et al.

Supplementary Data

What is “peak calling” ?

➥ Location of DNA binding sites of proteins (transcription factors, histones, etc.)



Wilbanks et al., Plos One (2010)

➥ “Peak calling” (ChIP-seq data) = identification of genomic regions with a high density of sequences (reads)

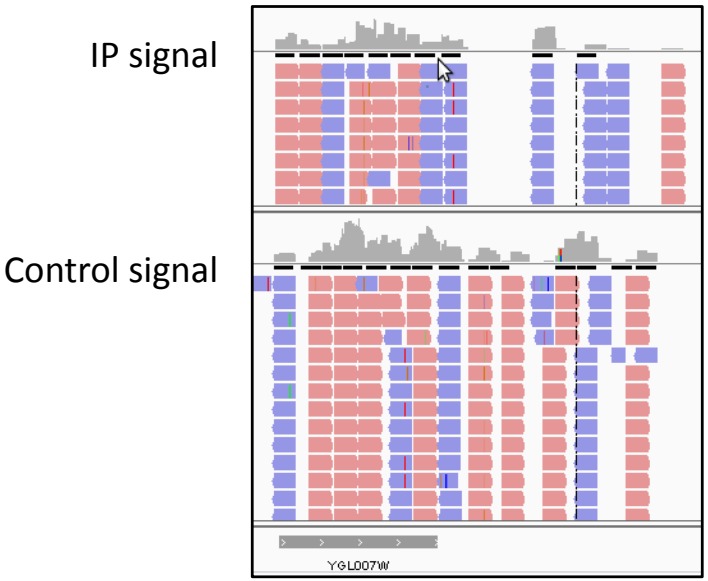
What is bPeaks?

➥ Simple approach for detection of basic peaks (bPeaks) from ChIP-seq data

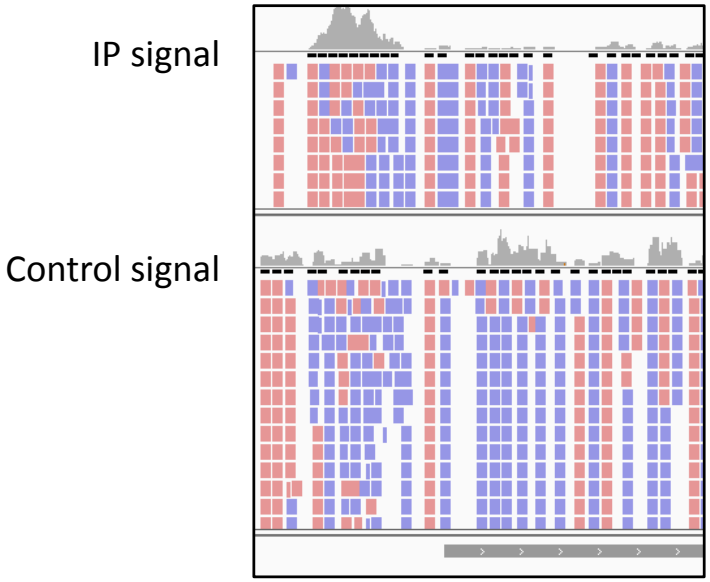
General philosophy

Easy-to-use tool based on an intuitive definition of peaks by a biologist who visually inspects the ChIP-seq data on a genome browser

Bad peak ☹

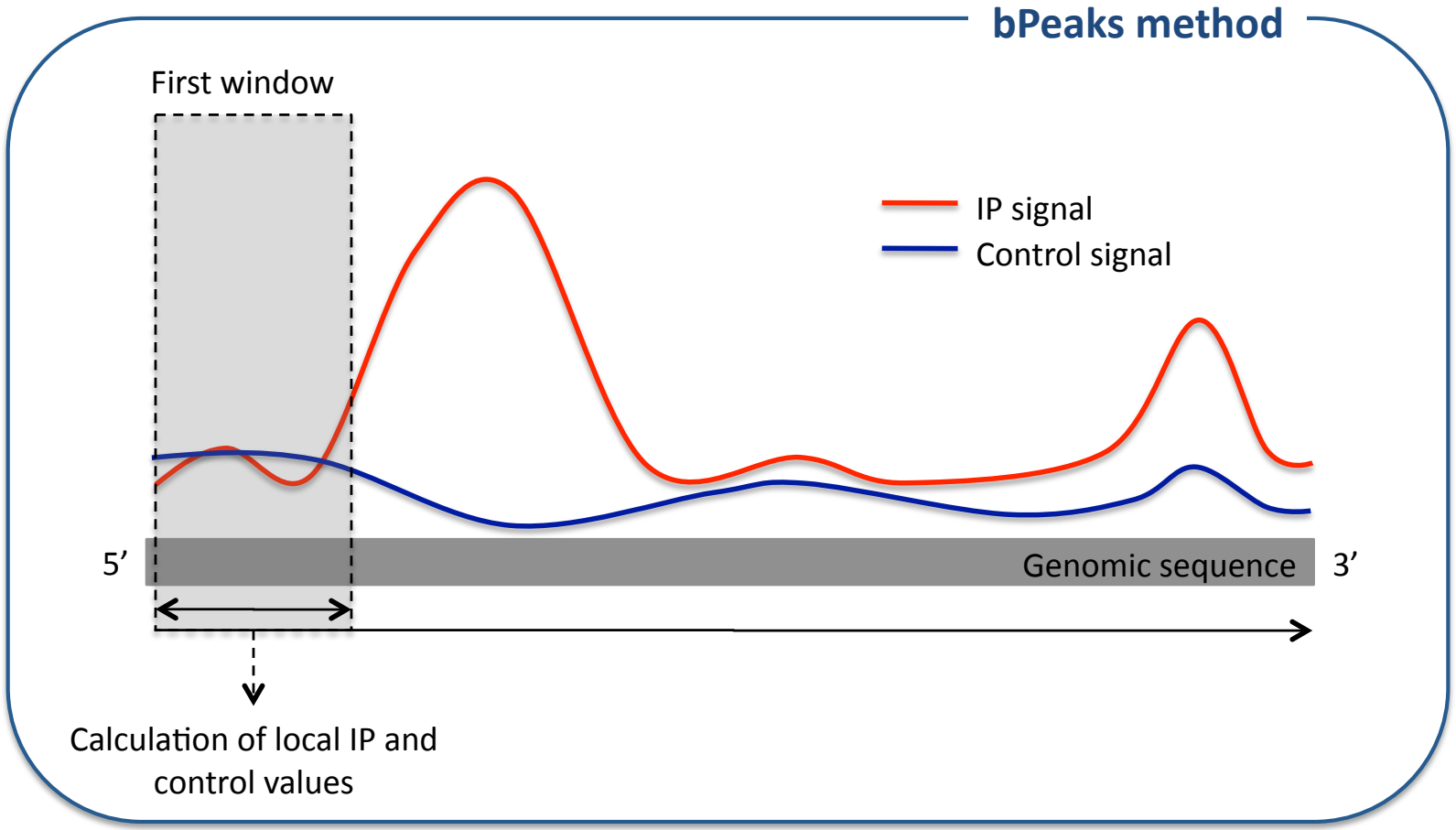


Good peak ☺



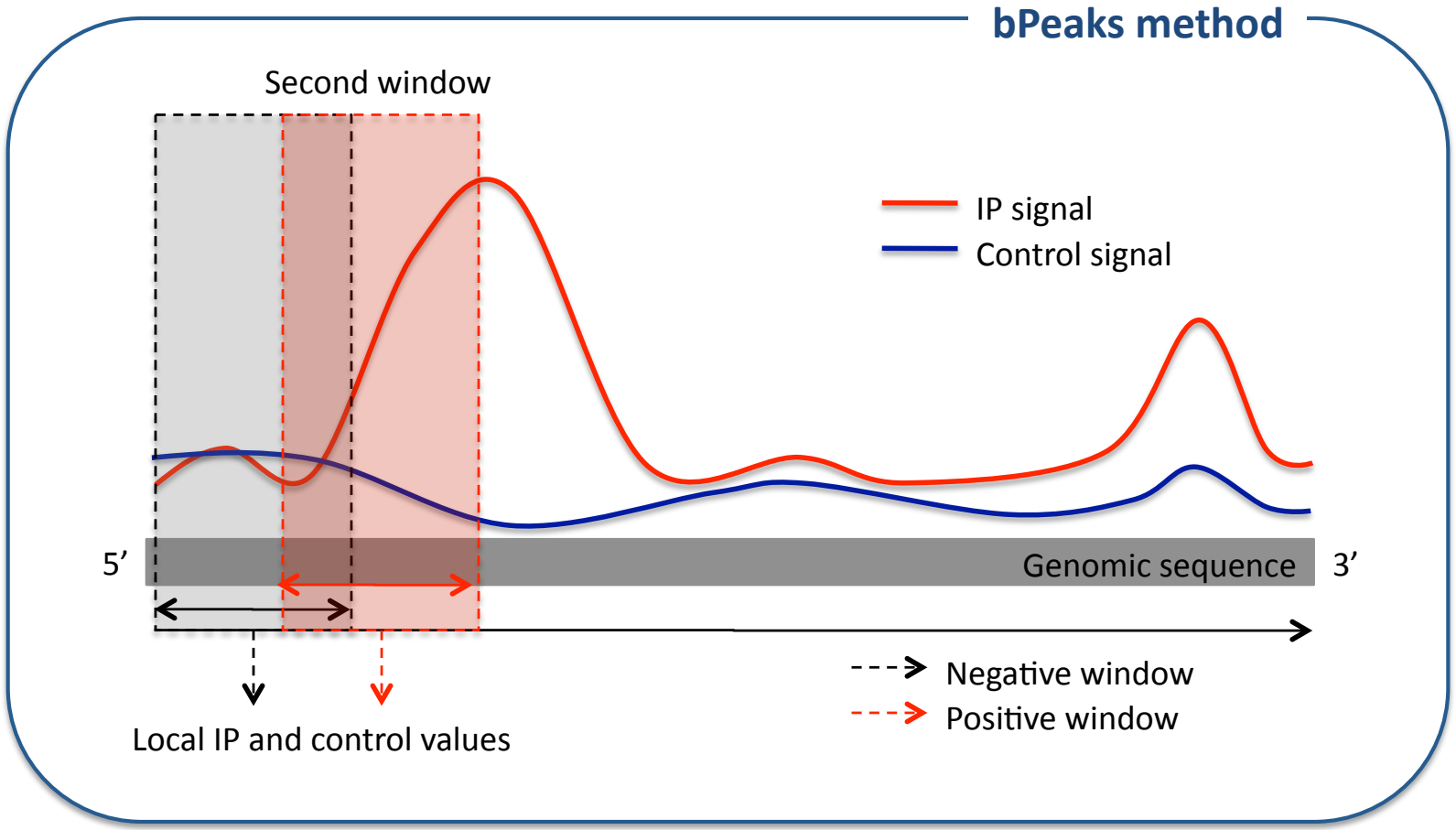
General principle (1/3)

➥ Sliding window to scan the genomic sequence and compare IP and control signals



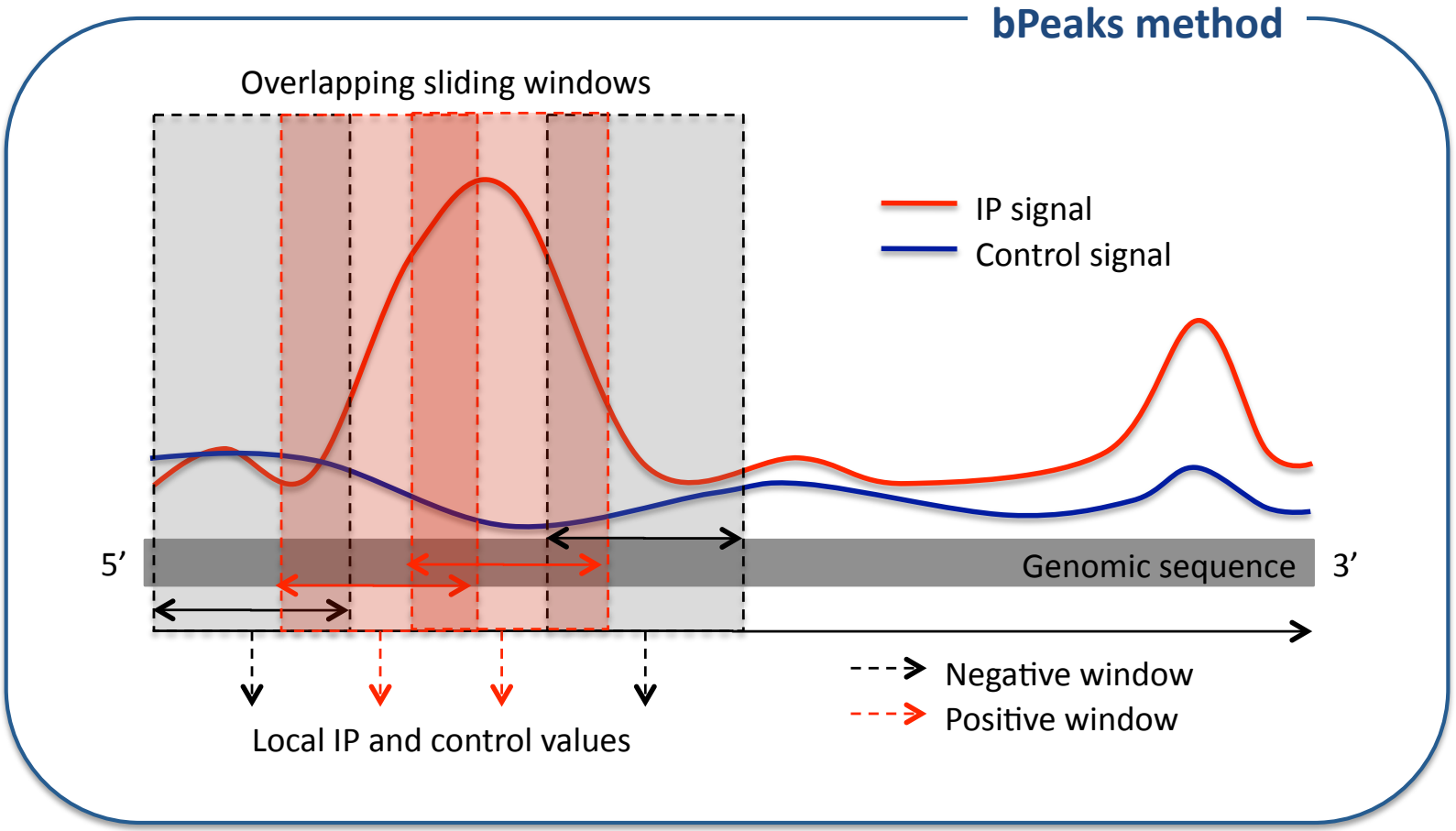
General principle (1/3)

➥ Sliding window to scan the genomic sequence and compare IP and control signals



General principle (1/3)

➥ Sliding window to scan the genomic sequence and compare IP and control signals



General principle (2/3)

➥ Four criterion define interesting genomic regions

Criteria #1

High number of sequences (reads) in IP sample

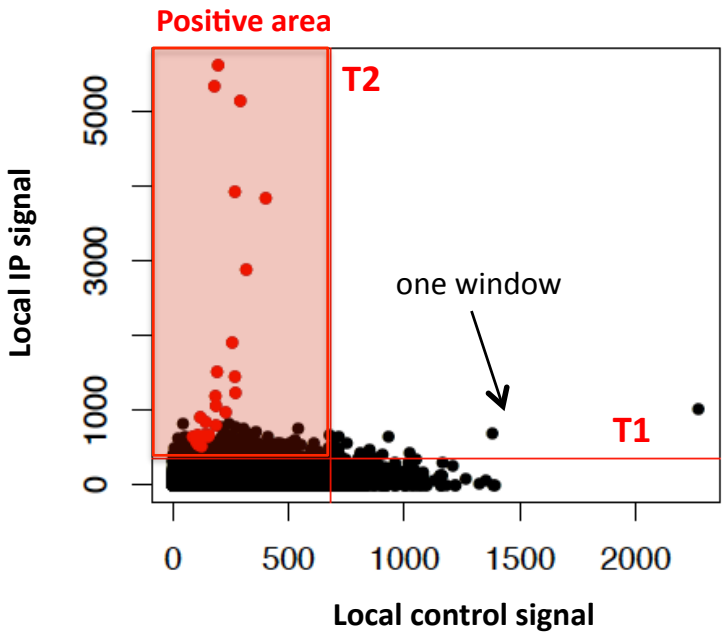
= T1

Criteria #2

Low number of sequences (reads) in control sample

= T2

Saccharomyce cerevisiae **(chromosome 15)**



General principle (2/3)

➥ Four criterion define interesting genomic regions

Criteria #1

High number of sequences (reads) in IP sample

= T1

Criteria #2

Low number of sequences (reads) in control sample

= T2

Criteria #3

High value of log(IP/control)

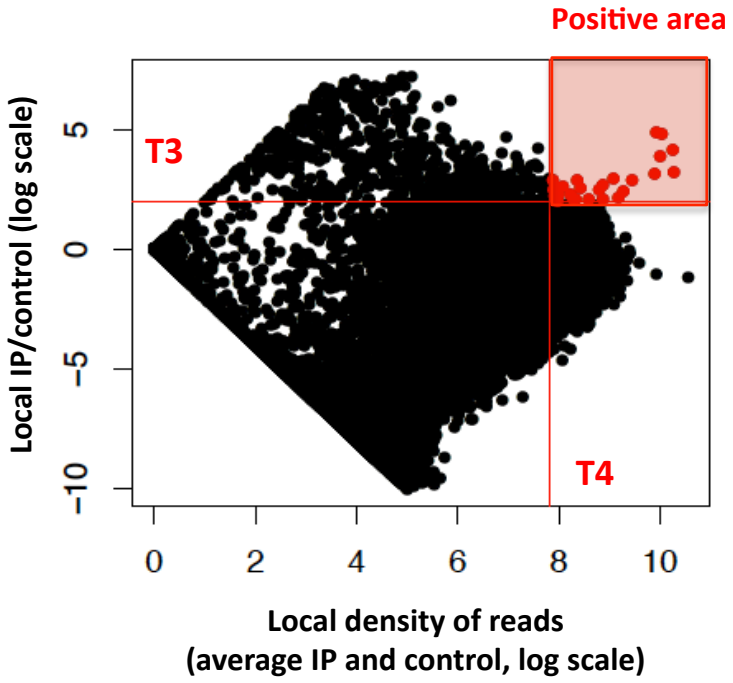
= T3

Criteria #4

High density of reads (IP and control samples)

= T4

Saccharomyce cerevisiae
(chromosome 15)



bPeaks R package - G. Lelandais

General principle (2/3)

➥ Four criterion define interesting genomic regions

Criteria #1

High number of sequences (reads) in IP sample

T1

+

Criteria #2

Low number of sequences (reads) in control sample

T2

+

Criteria #3

High value of log(IP/control)

T3

+

Criteria #4

High density of reads (IP and control samples)

T4

Interesting region

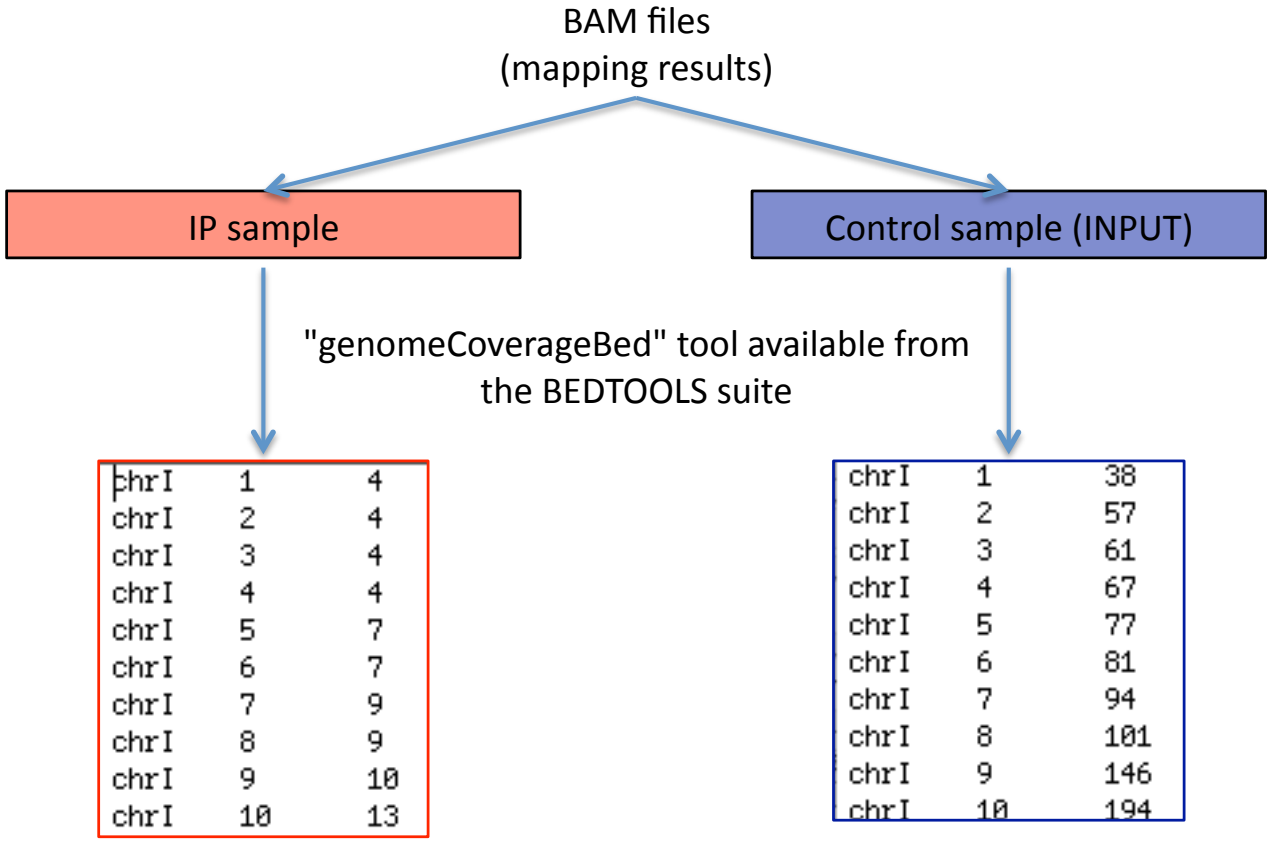
Successive positive windows

=

Basic peaks (bPeaks)

bPeaks input datasets

➥ Sequencing results should be datafiles with numbers of sequences mapped on each nucleotide in the reference genome



BEDTOOLS: https://code.google.com/p/bedtools/

Getting started with the R package

Package 'bPeaks'

July 30, 2013

Type Package

Title bPeaks: an intuitive peak-calling strategy to detect transcription factor binding sites from ChIP-seq data in small eukaryotic genomes

Version 1.2

Date 2013-07-30

Author Jawad MERHEJ and Gaelle LELANDAIS

Maintainer Gaelle LELANDAIS <gaelle.lelandais@univ-paris-diderot.fr>

Description bPeaks is a simple approach to identify transcription factor binding sites from ChIP-seq data. Our general philosophy is to provide an easy-to-use tool, well-adapted for small eukaryotic genomes (< 20 Mb). bPeaks uses a combination of 4 cut-offs (T1, T2, T3 and T4) to mimic ''good peak'' properties as described by biologists who visually inspect the ChIP-seq data on a genome browser. For yeast genomes, bPeaks calculates the proportion of peaks that fall in promoter sequences. These peaks are good candidates as transcription factor binding sites.

License GPL

Depends R (>= 2.10)

R topics documented:

1

Examples

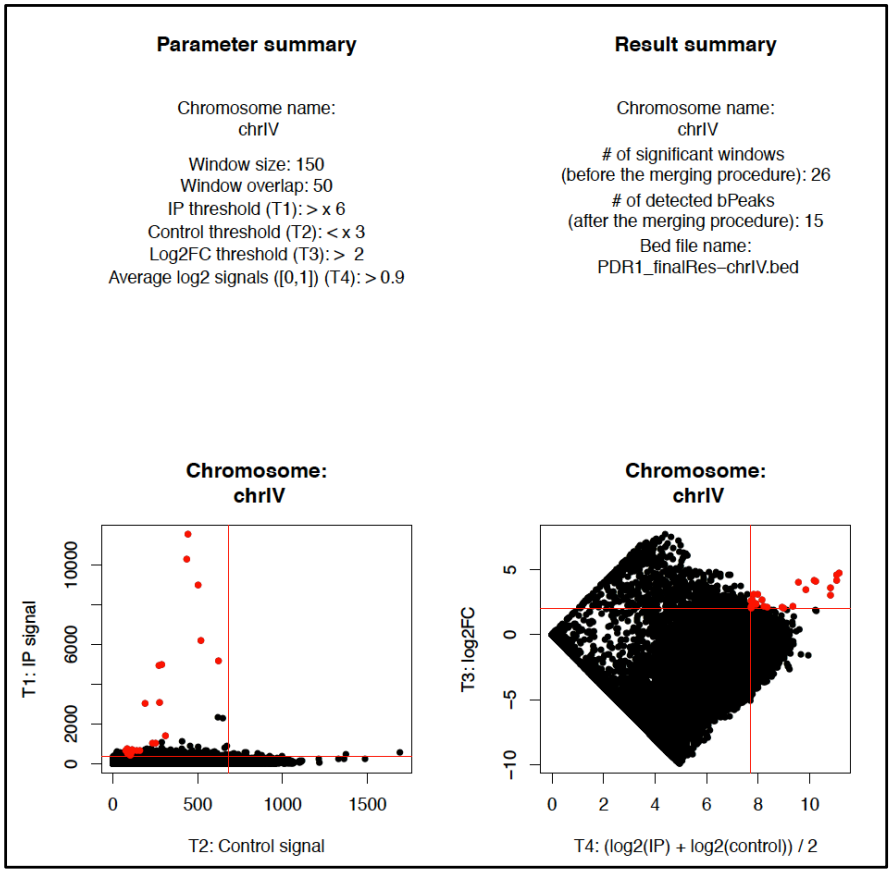
```
# get library
library(bPeaks)

# STEP 1: get PDR1 data
data(dataPDR1)

# STEP 2 : bPeaks analysis
bPeaksAnalysis(IPdata = dataPDR1$IPdata, controlData = dataPDR1$controlData,
               chromosomalFeatures = dataPDR1$chromosomalFeatures,
               smoothingValue = c(20),
               windowSize = c(150), windowOverlap = 50,
               IPcoeff = c(6), controlCoeff = c(4), log2FC = c(2),
               averageQuantiles = c(0.9),
               resultName = "bPeaks_PDR1",
               peakDrawing = TRUE, promSize = 800)
```

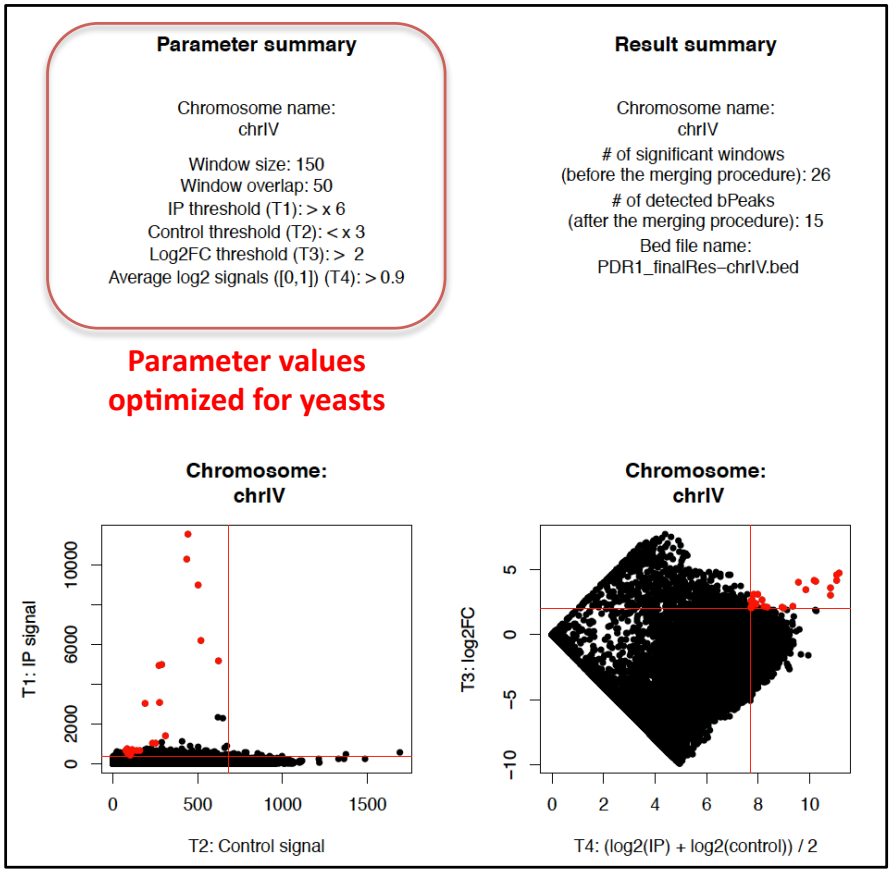
bPeaks output files (1/4)

➥ PDF file with IP and control signal information for each chromosome



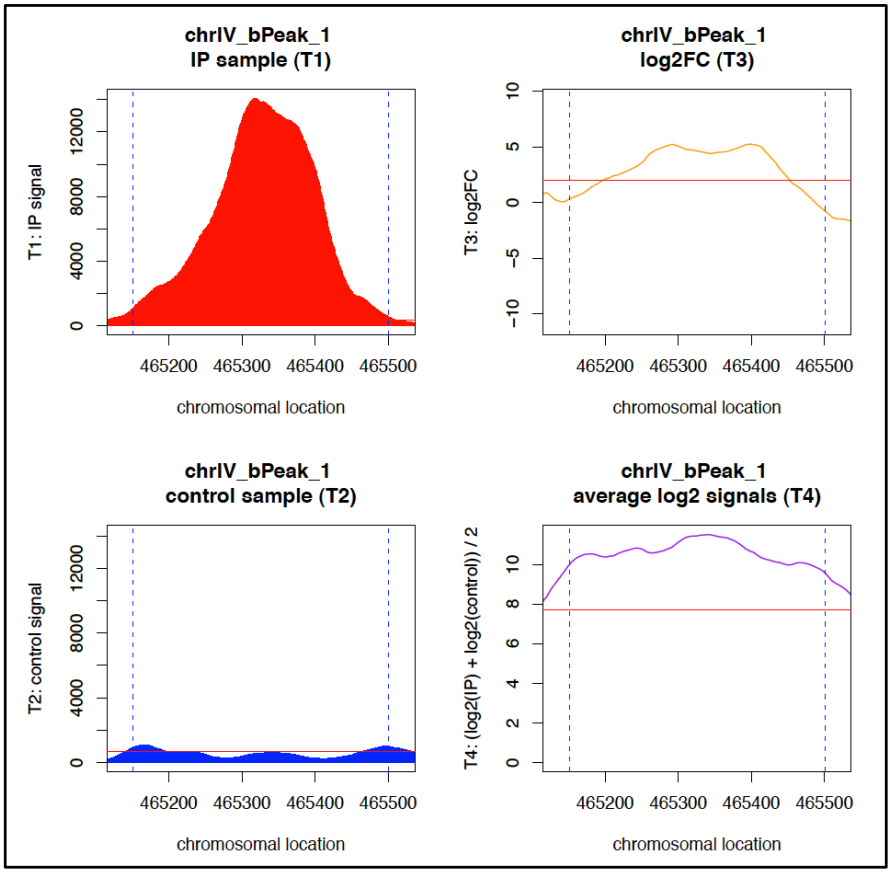
bPeaks output files (1/4)

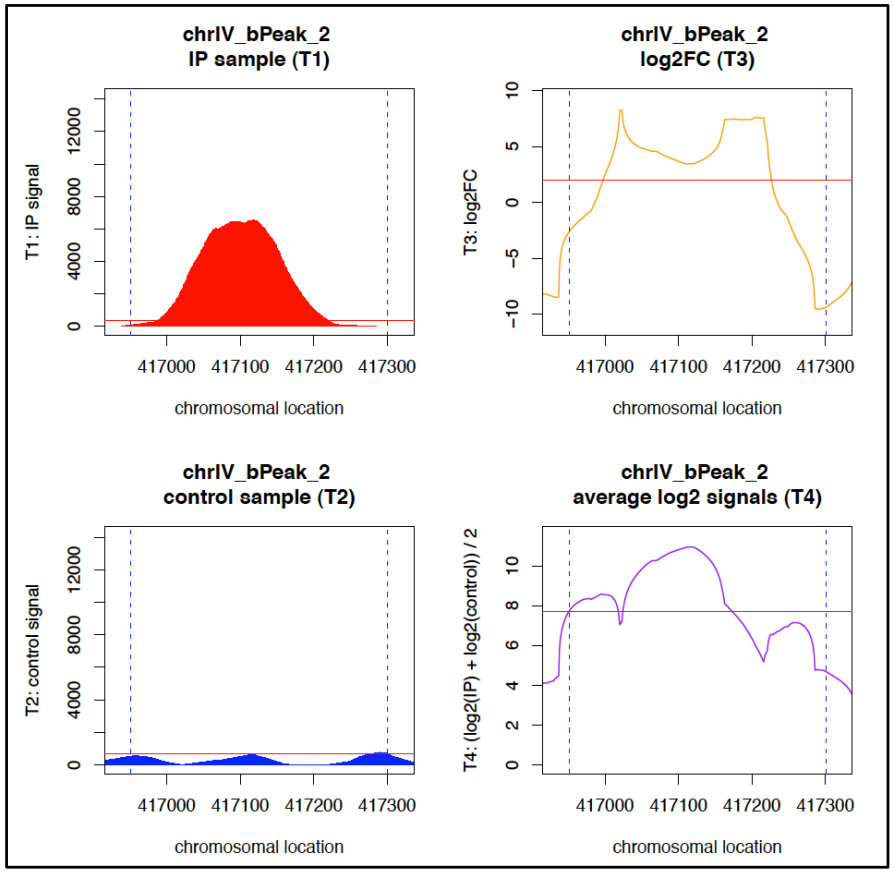
➥ PDF file with IP and control signal information for each chromosome



bPeaks output files (2/4)

➥ PDF files with threshold information for all detected regions



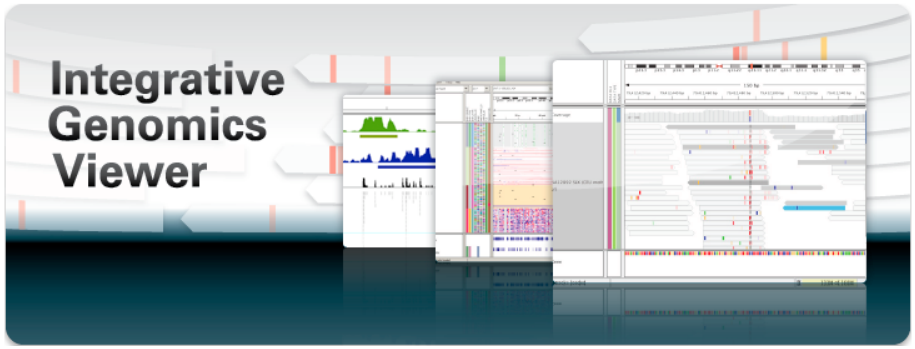


bPeaks output files (3/4)

➥ BED files for result visualization with genome browsers

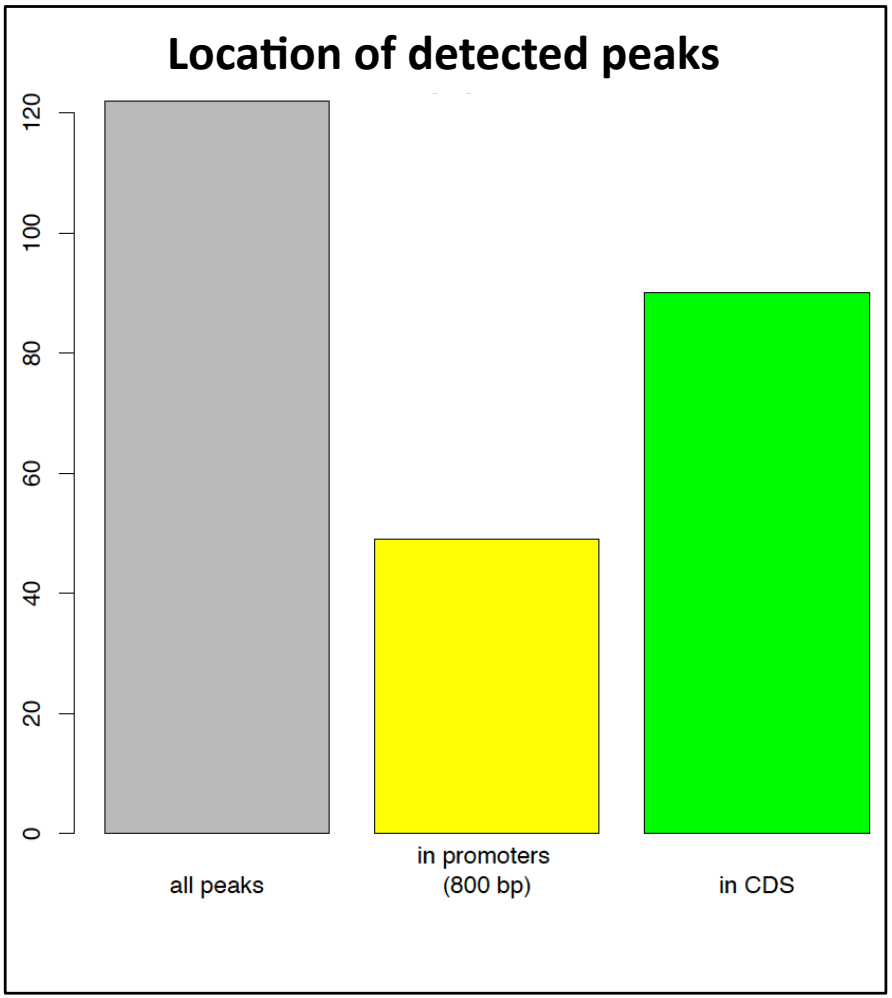
bPeaks_PDR1-chrXVI.bed

chrXVI	481451	481601	chrXVI_bPeak_1	823.109675335164|75.0573413018898|3.43767693557772|7.96785407096343
chrXVI	645451	645751	chrXVI_bPeak_2	2736.48570505573|318.769665643676|3.12817829272333|9.82056447352065
chrXVI	281001	281201	chrXVI_bPeak_3	703.313519625262|86.5303666612825|3.00057324519693|7.95199286682064
chrXVI	425201	425451	chrXVI_bPeak_4	1420.82016906262|252.974317557745|2.49941858581822|9.21997075486276
chrXVI	111751	111901	chrXVI_bPeak_5	608.152802455177|128.580197060249|2.23295890831877|8.13418090084553
chrXVI	86501	86651	chrXVI_bPeak_6	496.148441285738|106.799386205783|2.20532772022789|7.85486901350419
chrXVI	399901	400051	chrXVI_bPeak_7	509.60216443224|118.881440801163|2.09059132514139|7.9507601805171
chrXVI	727851	728001	chrXVI_bPeak_8	565.056372153125|134.798578581812|2.0594773558212|8.11506324650433
chrXVI	303101	303251	chrXVI_bPeak_9	467.096430302051|112.29946696818|2.04666468928495|7.84732960824842
chrXVI	193601	193751	chrXVI_bPeak_10	473.451461799386|114.447100629947|2.03902856024137|7.87060241122466
chrXVI	803501	803651	chrXVI_bPeak_11	576.339848166694|139.940397350993|2.03433561746313|8.15610918506283
chrXVI	433351	433501	chrXVI_bPeak_12	617.266677434986|151.683572928445|2.01768439873837|8.26324324152417



IGV: http://www.broadinstitute.org/igv/

bPeaks output files (4/4)



- For yeast genomes, bPeaks calculates the proportion of peaks falling in promoter regions.
- These peaks are good candidates as potential transcription factor binding sites.
- Annotations of genes positions for ten different yeasts species are available:
  - *Saccharomyces cerevisiae,*
  - *Candida albicans,*
  - *Candida glabrata,*
  - *Debaryomyces hansenii,*
  - *Eremothecium gossypii,*
  - *Kluyveromyces lactis,*
  - *Pichia sorbitophila,*
  - *Saccharomyces kluyveri,*
  - *Yarrowia lipolytica*
  - *Zygosaccharomyces rouxii*.

bPeaks summary

R package

•Easy-to-use tool based on an intuitive definition of peaks by a biologist who visually inspects the ChIP-seq data on a genome browser

•Graphical outputs and BED files are provided to rapidly assess the relevance of the chosen parameters (T1, T2, T3 and T4)

Method performances

•Average size of peaks detected with bPeaks is remarkably small

•Whereas bPeaks does not use elaborate statistical model, the detected peaks are relevant according to others biological information

•Visual inspection of the peaks and motif discovery show that bPeaks are well focused and centered around the protein DNA binding sites

More information?

➥ Supplementary data together with detailed tutorials (R programming and bPeaks use) are available online: http://bpeaks.gene-networks.net