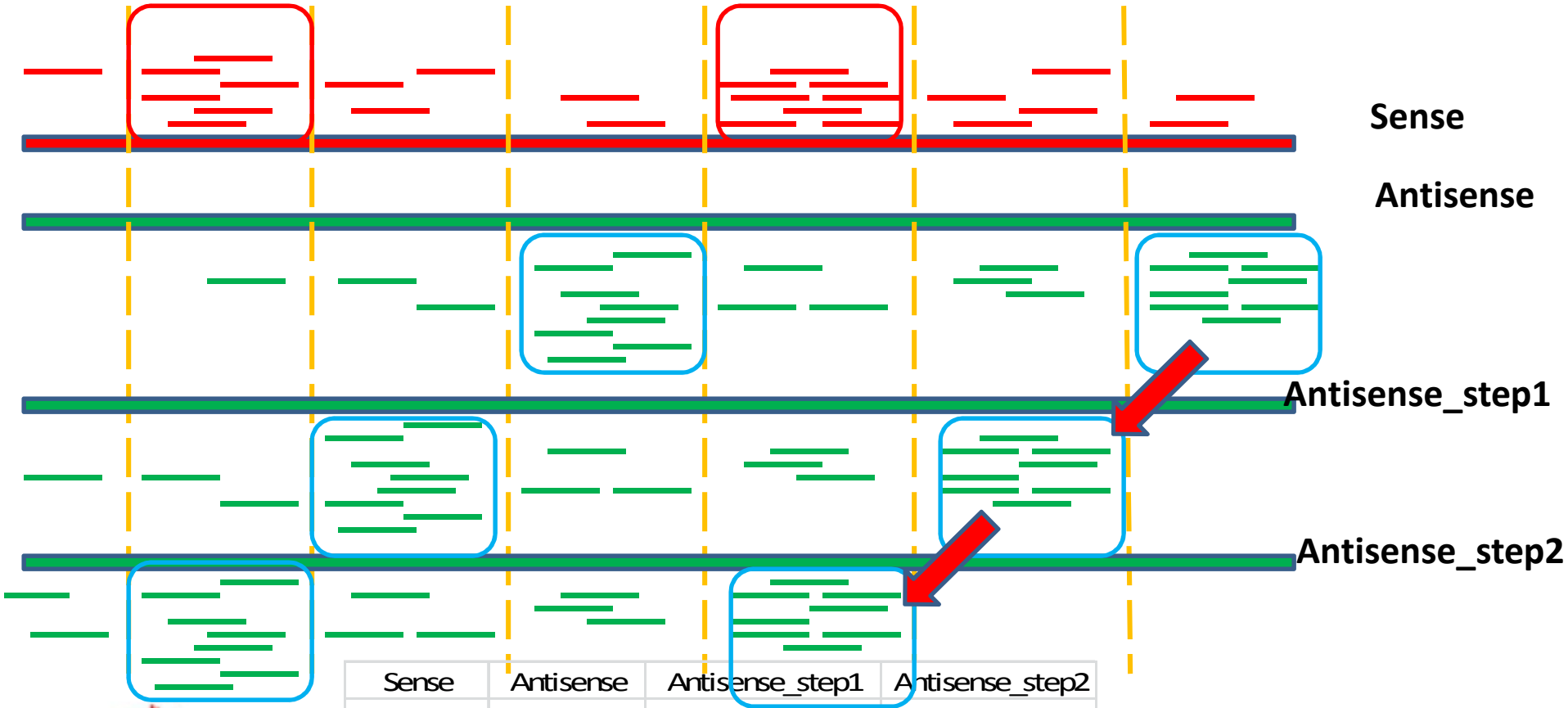
# Functional annotation of ChIP-peaks

Minghui Wang, Qi Sun
Bioinformatics Facility
Institute of Biotechnology



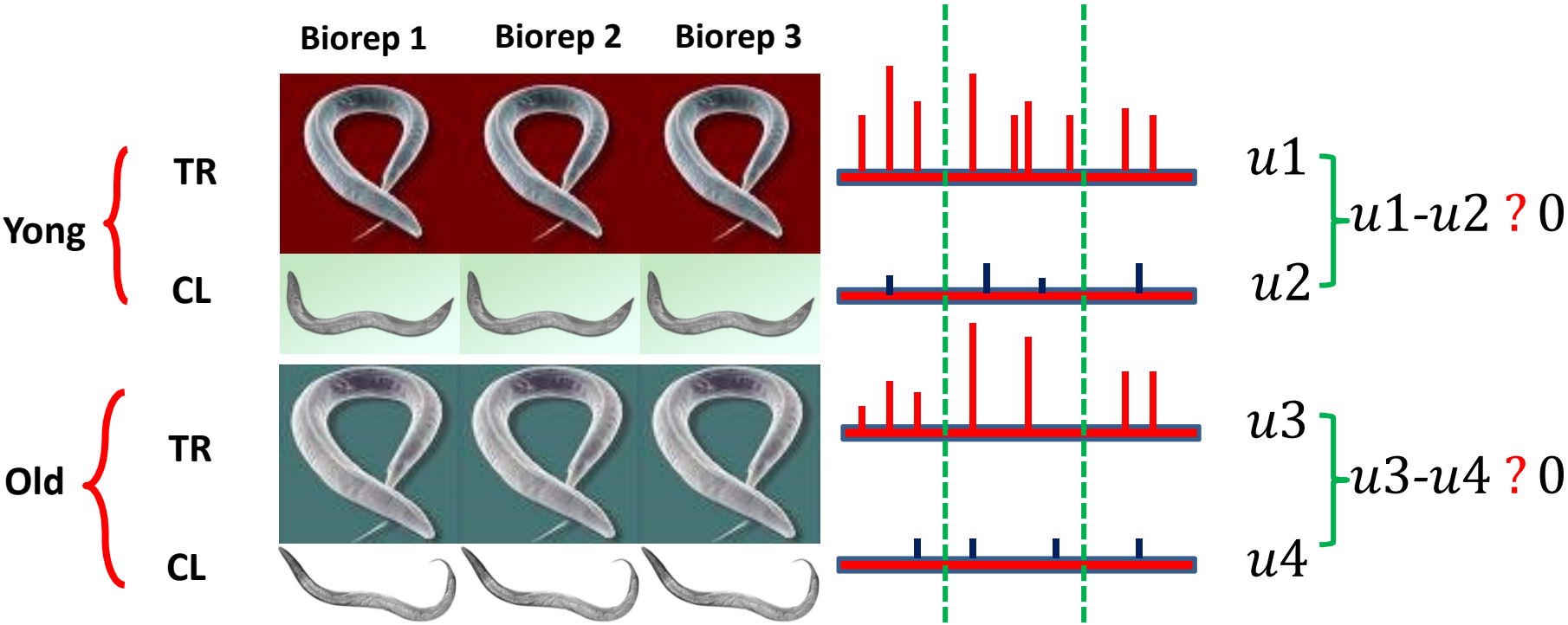
| Sense | Antisense | Antisense_step1 | Antisense_step2 |
|---|---|---|---|
| 1 | 0 | 1 | 2 |
| 6 | 1 | 2 | 8 |
| 3 | 2 | 8 | 3 |
| 2 | 8 | 3 | 3 |
| 8 | 3 | 3 | 8 |
| 4 | 3 | 8 | 0 |
| 2 | 8 | 0 | 0 |

R= -0.27

R= 0.13

R= 0.79

# Experimental design




$((\boldsymbol{u1}\text{-}\boldsymbol{u2}) - (\boldsymbol{u3} - \boldsymbol{u4}))$? 0 is for ????

# GLM (Poisson)

**Reads counts**

$$Y_i \begin{bmatrix} 34 \\ 12 \\ 42 \\ 18 \\ 44 \\ 20 \\ 25 \\ 10 \\ 32 \\ 15 \\ 38 \\ 14 \end{bmatrix} = \begin{array}{c} \begin{array}{cccc} \textbf{Mean} & \textbf{Time} & \textbf{IP} & \textbf{Time*IP} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{bmatrix} \mu \\ \alpha \\ \beta \\ \alpha*\beta \end{bmatrix} + \varepsilon_i$$

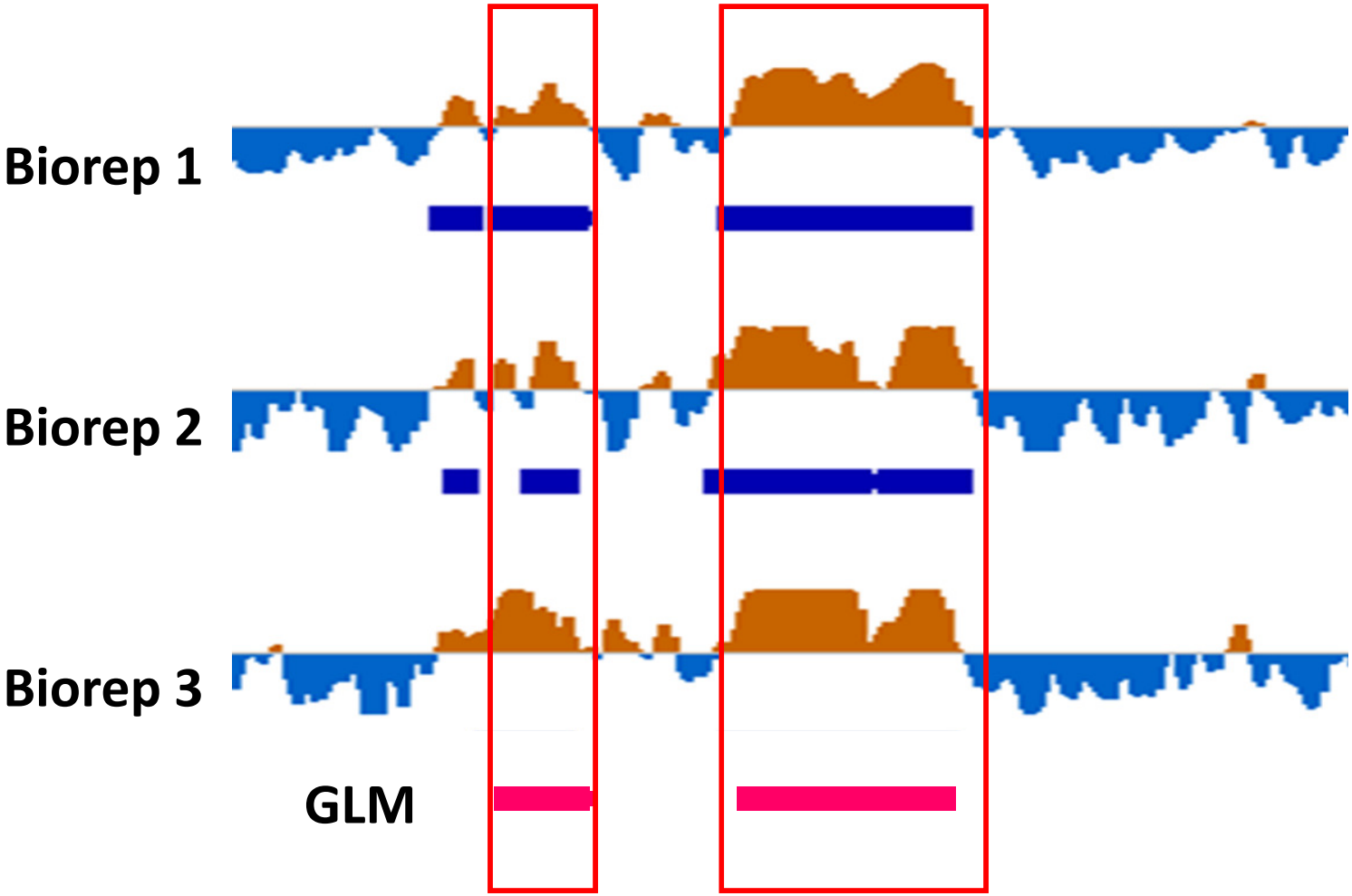
**Yong**

**Old**

$$Y = XB + E$$

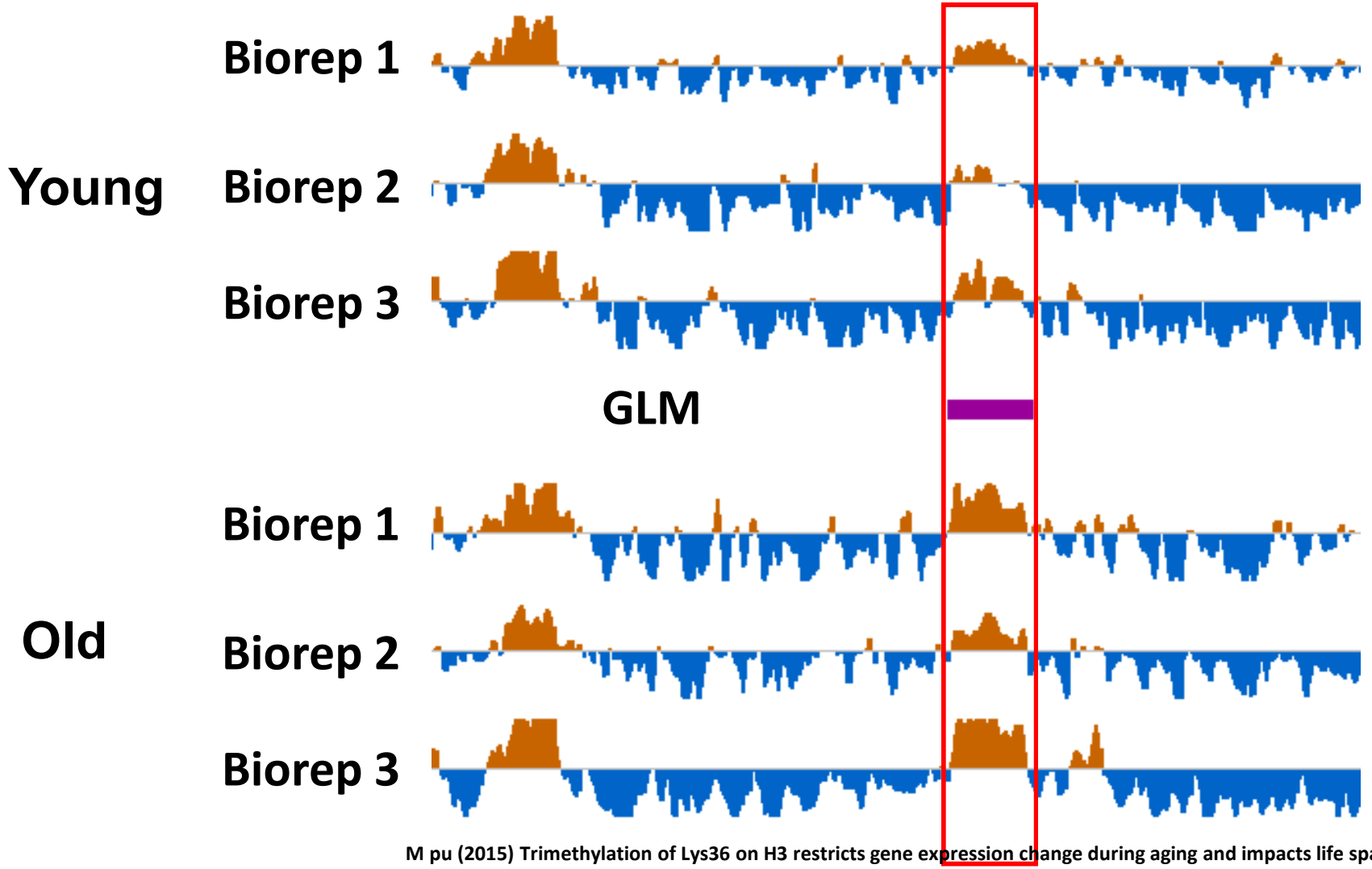
**out<-glm(Y ~ Time*IP , family = poisson, log(offset=c(library size)))**

# Identify enriched regions within Yong or Old



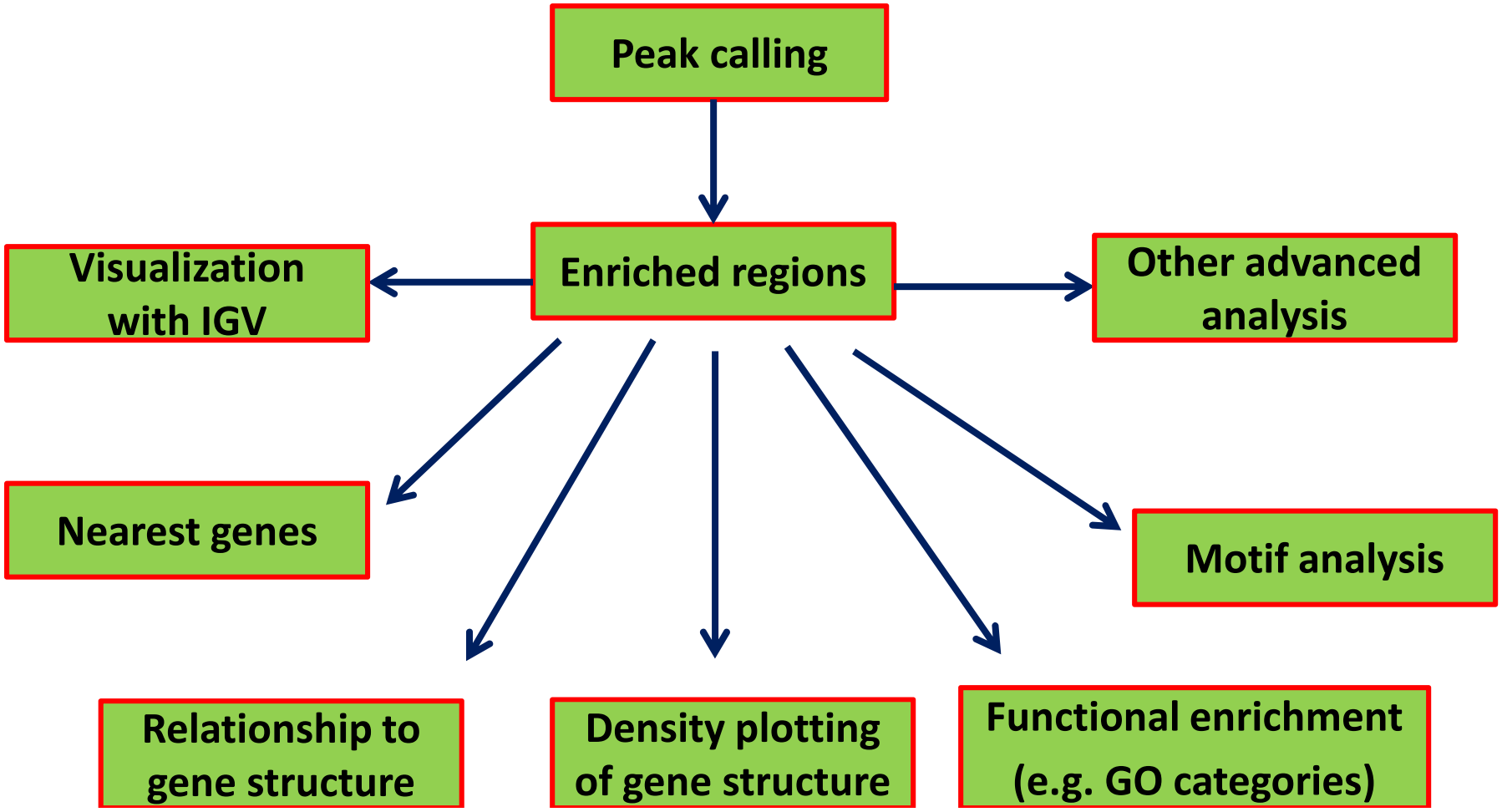

**M pu (2015) Trimethylation of Lys36 on H3 restricts gene expression change during aging and impacts life span.**
**Genes Dev** 1;29(7):718-31

# Identify enrichment regions between young and old stages



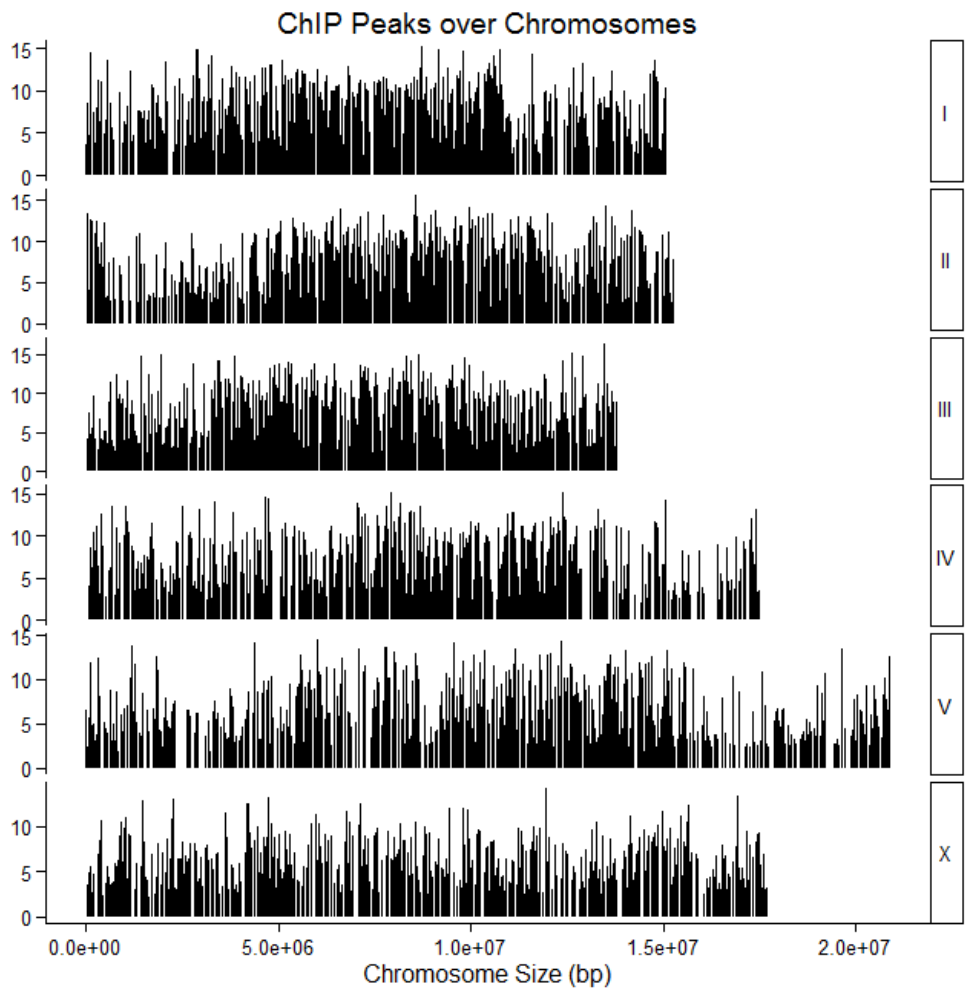

**M pu (2015) Trimethylation of Lys36 on H3 restricts gene expression change during aging and impacts life span. Genes Dev** 1;29(7):718-31

# Functional annotation workflow

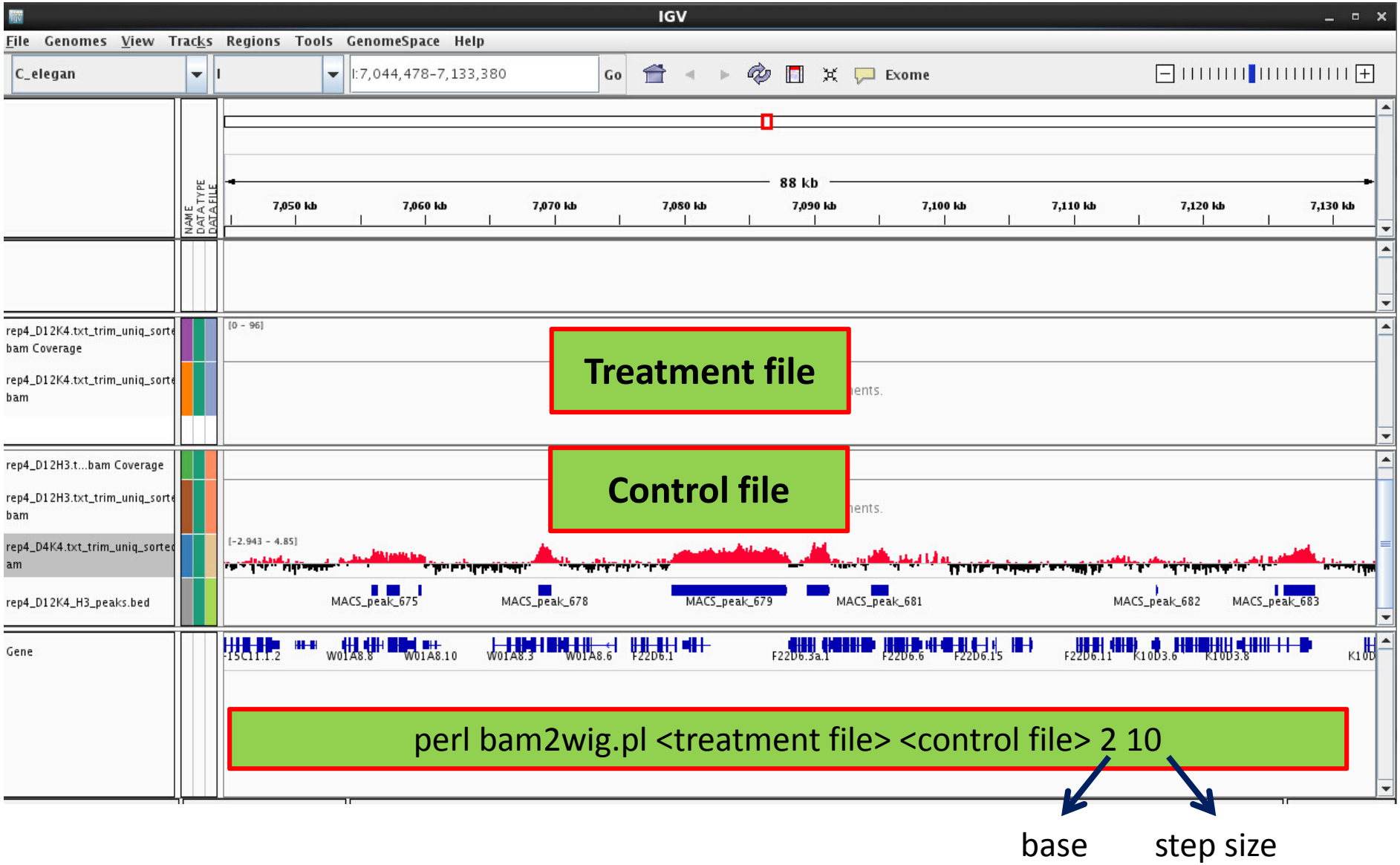


# Peaks distribution across chromosome

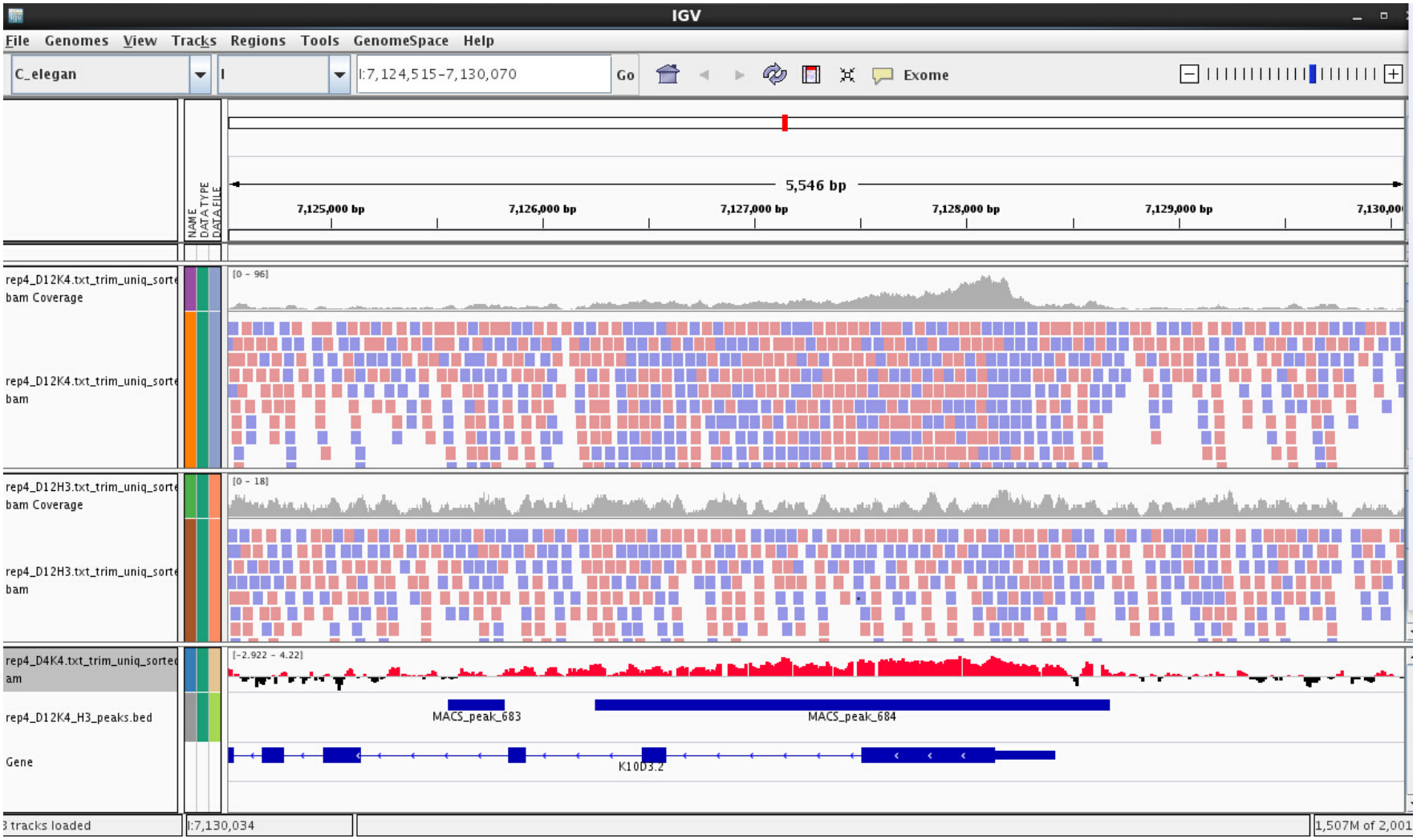
```
peak<- readPeakFile("D12_peak_rep4.txt", as="GRanges")
plotChrCov(peak, weightCol=5)
```



# IGV


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
C_elegan
I:7,044,478-7,133,380
Go
Exome
88 kb
7,050 kb
7,060 kb
7,070 kb
7,080 kb
7,090 kb
7,100 kb
7,110 kb
7,120 kb
7,130 kb
NAME
DATA TYPE
DATA FILE
rep4_D12K4.txt_trim_uniq_sorte
bam Coverage
rep4_D12K4.txt_trim_uniq_sorte
bam
[0 - 96]
Treatment file
rep4_D12H3.t...bam Coverage
rep4_D12H3.txt_trim_uniq_sorte
bam
Control file
rep4_D4K4.txt_trim_uniq_sorte
am
[-2.943 - 4.85]
rep4_D12K4_H3_peaks.bed
MACS_peak_675
MACS_peak_678
MACS_peak_679
MACS_peak_681
MACS_peak_682
MACS_peak_683
Gene
W01A8.8
W01A8.10
W01A8.3
W01A8.6
F22D6.1
F22D6.3a.1
F22D6.6
F22D6.15
F22D6.11
K10D3.6
K10D3.8
perl bam2wig.pl <treatment file> <control file> 2 10
base
step size

# IGV


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
C_elegan
I
I:7,124,515-7,130,070
Go
Exome
5,546 bp
7,125,000 bp
7,126,000 bp
7,127,000 bp
7,128,000 bp
7,129,000 bp
NAME
DATA TYPE
DATA FILE
rep4_D12K4.txt_trim_uniq_sorte
bam Coverage
[0 - 96]
rep4_D12K4.txt_trim_uniq_sorte
bam
rep4_D12H3.txt_trim_uniq_sorte
bam Coverage
[0 - 18]
rep4_D12H3.txt_trim_uniq_sorte
bam
rep4_D4K4.txt_trim_uniq_sorted
am
[-2.922 - 4.22]
rep4_D12K4_H3_peaks.bed
MACS_peak_683
MACS_peak_684
Gene
K10D3.2
3 tracks loaded
I:7,130,034
1,507M of 2,001

# Annotating Peaks

- Homer
- PeakAnalyzer
- ChIPpeakAnno
- ChIPseeker
- …

Peak region file

GFF (General Feature Format)
Genome annotation

R & Bioconductor

# Homer

Usage: perl /path-to-homer/configureHomer.pl -install mm8

Usage: annotatePeaks.pl <peak file | tss> <genome version> [additional options...]

```
Available Genomes (required argument): (name,org,directory,default promoter set)
        ce10    worm    /workdir/homer/.//data/genomes/ce10/    default
        tair10  arabidopsis     /workdir/homer/.//data/genomes/tair10/  default
        hg19    human   /workdir/homer/.//data/genomes/hg19/    default
        ce6     worm    /workdir/homer/.//data/genomes/ce6/     default
        hg18    human   /workdir/homer/.//data/genomes/hg18/    default
                -- or --
        Custom: provide the path to genome FASTA files (directory or single file)

User defined annotation files (default is UCSC refGene annotation):
        annotatePeaks.pl accepts GTF (gene transfer formatted) files to annotate positions relative
        to custom annotations, such as those from de novo transcript discovery or Gencode.
        -gtf <gtf format file> (-gff and -gff3 can work for those files, but GTF is better)

Peak vs. tss/tts/rna mode (works with custom GTF file):
        If the first argument is "tss" (i.e. annotatePeaks.pl tss hg18 ...) then a TSS centric
        analysis will be carried out.  Tag counts and motifs will be found relative to the TSS.
        (no position file needed) ["tts" now works too - e.g. 3' end of gene]
        ["rna" specifies gene bodies, will automaticall set "-size given"]
        NOTE: The default TSS peak size is 4000 bp, i.e. +/- 2kb (change with -size option)
        -list <gene id list> (subset of genes to perform analysis [unigene, gene id, accession,
                 probe, etc.], default = all promoters)
        -cTSS <promoter position file i.e. peak file> (should be centered on TSS)

Primary Annotation Options:
        -mask (Masked repeats, can also add 'r' to end of genome name)
        -m <motif file 1> [motif file 2] ... (list of motifs to find in peaks)
                -mscore (reports the highest log-odds score within the peak)
                -nmotifs (reports the number of motifs per peak)
                -mdist (reports distance to closest motif)
                -mfasta <filename> (reports sites in a fasta file - for building new motifs)
                -fm <motif file 1> [motif file 2] (list of motifs to filter from above)
                -rmrevopp <#> (only count sites found within <#> on both strands once, i.e. palindromic)
                -matrix <prefix> (outputs a motif co-occurrence files:
                        prefix.count.matrix.txt - number of peaks with motif co-occurrence
                        prefix.ratio.matrix.txt - ratio of observed vs. expected  co-occurrence
                        prefix.logPvalue.matrix.txt - co-occurrence enrichment
                        prefix.stats.txt - table of pair-wise motif co-occurrence statistics
                        additional options:
                        -matrixMinDist <#> (minimum distance between motif pairs - to avoid overlap)
                        -matrixMaxDist <#> (maximum distance between motif pairs)
                -mbed <filename> (Output motif positions to a BED file to load at UCSC (or -mpeak))
```

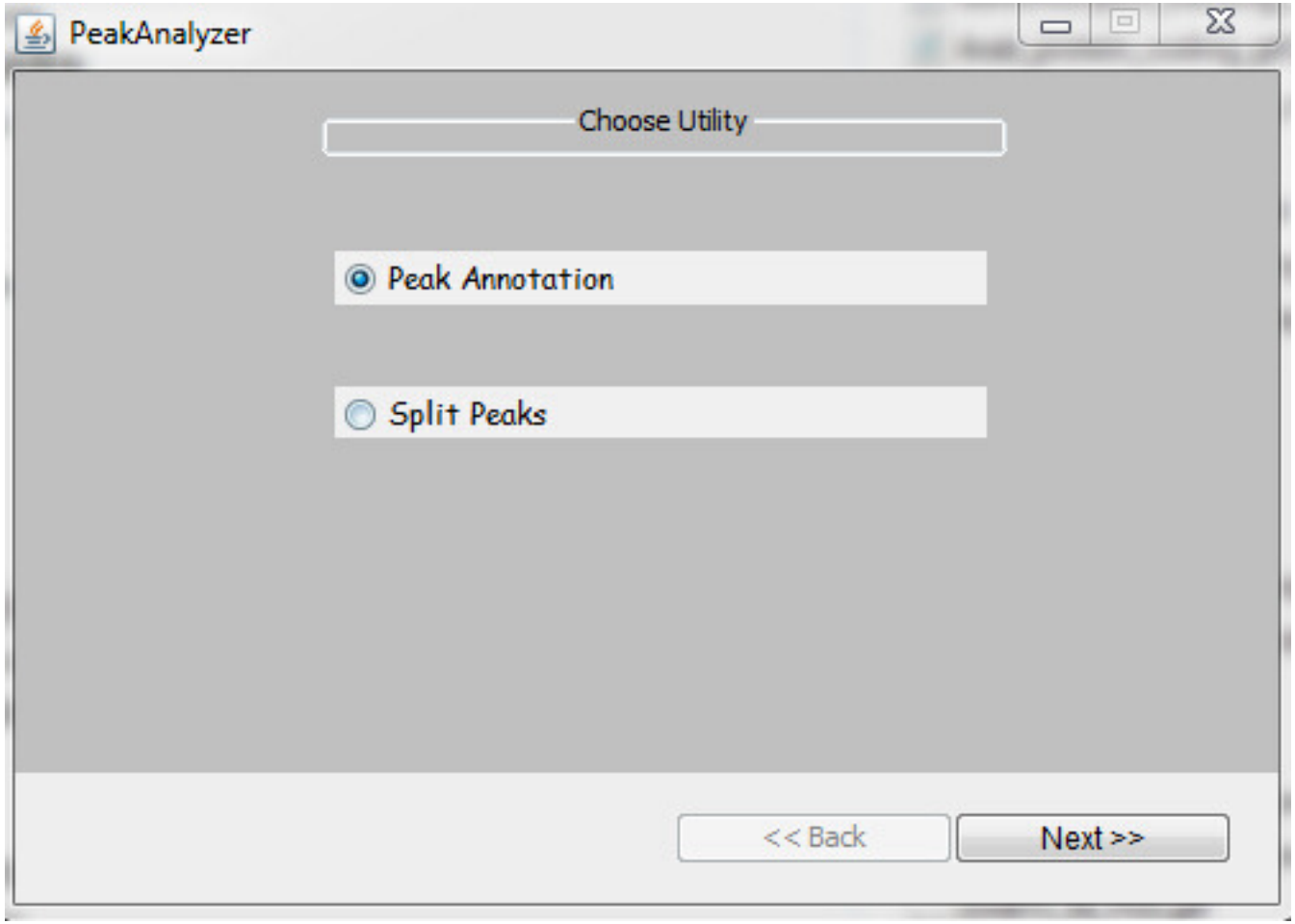
http://homer.salk.edu/homer/download.html

# Homer

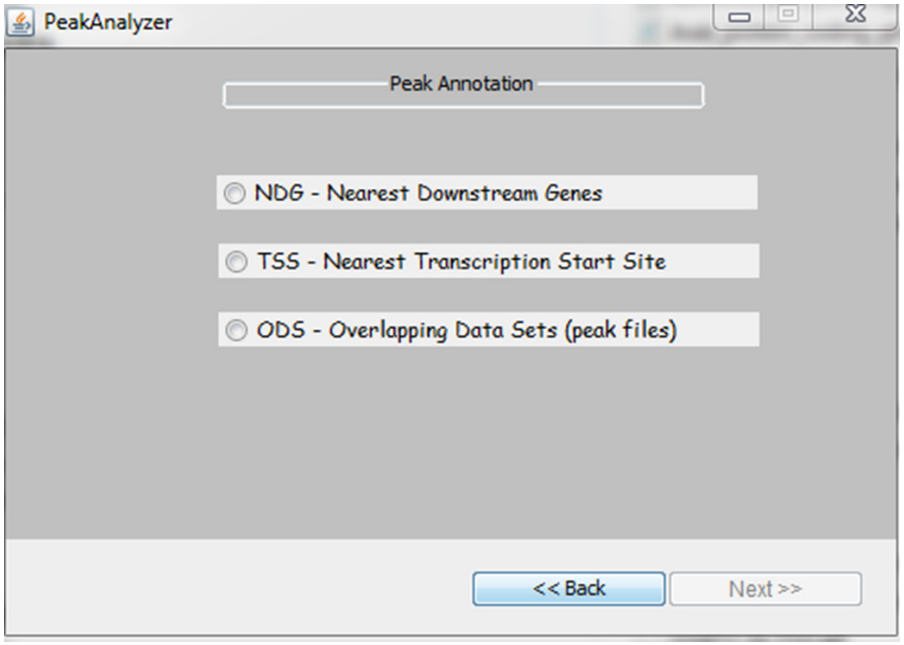
```
annotatePeaks.pl  test_peak.txt ce10 >test_peaks.out
```

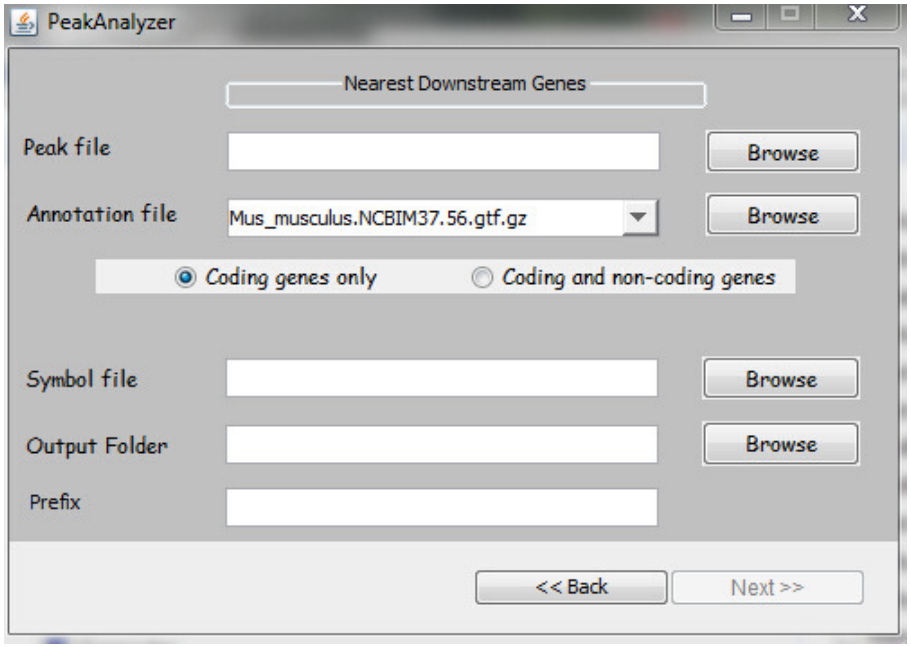
| ep4_D12K4_H3_peaks_ | Chr | Start | End | Detailed Annotation | Distance to TSS | Nearest PromoterID | Entrez ID | Nearest Refseq | Gene Name | Gene Description | Gene Type |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MACS_peak_4251 | chrIII | 13462360 | 13465118 | TTS (NM_067382) | 282 | NM_001268275 | 176773 | NM_001027628 | ant-1.1 | Protein ANT-1.1 | protein-coding |
| MACS_peak_7983 | chrX | 16929113 | 16932064 | oding (NR_000951, exon 1 | 747 | NR_000951 | 181740 | NR_000950 | C30E1.9 | ncRNA | ncRNA |
| MACS_peak_2231 | chrII | 8559913 | 8561702 | romoter-TSS (NR_003409 | -86 | NR_003409 | 4927029 | NR_003409 | F54C9.12 | ncRNA | ncRNA |
| MACS_peak_2592 | chrII | 11976667 | 11980155 | romoter-TSS (NM_182170 | -831 | NM_182170 | 259466 | NM_182170 | Y17G7B.21 | Protein Y17G7B.21 | protein-coding |
| MACS_peak_3410 | chrIII | 5938476 | 5944174 | on (NM_171868, intron 1 | 381 | NM_171868 | 175840 | NM_171139 | ubq-1 | Protein UBQ-1 | protein-coding |
| MACS_peak_875 | chrI | 8727917 | 8729277 | romoter-TSS (NM_059960 | 105 | NR_050329 | 13181245 | NR_050329 | C36B1.16 | ncRNA | ncRNA |
| MACS_peak_3642 | chrIII | 7972642 | 7974525 | romoter-TSS (NM_171200 | -77 | NM_171200 | 176121 | NM_066337 | aco-2 | Protein ACO-2 | protein-coding |
| MACS_peak_3165 | chrIII | 4193620 | 4197659 | TTS (NM_001083177) | -1633 | NM_065440 | 175542 | NM_065440 | asd-1 | Protein ASD-1 | protein-coding |
| MACS_peak_4169 | chrIII | 12606526 | 12610005 | romoter-TSS (NM_067241 | -8 | NM_067241 | 176679 | NM_067241 | epc-1 | Protein EPC-1 | protein-coding |
| MACS_peak_4897 | chrIV | 7925658 | 7927002 | romoter-TSS (NR_056256 | 29 | NM_068921 | 177583 | NM_068921 | rps-2 | Protein RPS-2 | protein-coding |
| MACS_peak_5299 | chrIV | 12390231 | 12392020 | romoter-TSS (NM_069965 | -101 | NM_069964 | 178188 | NM_069964 | rps-23 | Protein RPS-23 | protein-coding |
| MACS_peak_2956 | chrIII | 1950511 | 1951168 | omoter-TSS (NM_0011292 | -140 | NM_001129226 | 175338 | NM_001129226 | rps-22 | Protein RPS-22 | protein-coding |
| MACS_peak_1030 | chrI | 9772036 | 9773394 | romoter-TSS (NM_060195 | -164 | NM_060194 | 259393 | NM_060194 | F26E4.4 | Protein F26E4.4 | protein-coding |
| MACS_peak_3728 | chrIII | 8630742 | 8631769 | TTS (NM_066499) | 160 | NM_066501 | 176211 | NM_066500 | dnj-10 | Protein DNJ-10 | protein-coding |
| MACS_peak_4805 | chrIV | 7083367 | 7085404 | romoter-TSS (NR_003459 | 265 | NM_068702 | 177481 | NM_068702 | rps-4 | Protein RPS-4 | protein-coding |
| MACS_peak_226 | chrI | 2875156 | 2878187 | romoter-TSS (NM_058662 | -115 | NR_050012 | 13179637 | NR_050012 | Y71F9AL.19 | ncRNA | ncRNA |
| MACS_peak_3697 | chrIII | 8331282 | 8332686 | romoter-TSS (NM_181969 | -160 | NM_066434 | 176176 | NM_066434 | mrps-18A | Protein MRPS-18A | protein-coding |
| MACS_peak_1159 | chrI | 10732833 | 10734695 | romoter-TSS (NM_060409 | -117 | NM_060409 | 181833 | NM_060409 | mrpl-9 | Protein MRPL-9 | protein-coding |
| MACS_peak_2918 | chrIII | 1418790 | 1420441 | romoter-TSS (NR_052079 | 406 | NM_001027795 | 175299 | NM_001027795 | Y82E9BR.3 | Protein Y82E9BR.3 | protein-coding |
| MACS_peak_4195 | chrIII | 12904879 | 12906163 | on (NM_067280, exon 1 of | 126 | NM_067280 | 176707 | NM_067280 | cco-2 | Protein CCO-2 | protein-coding |
| MACS_peak_3118 | chrIII | 3867319 | 3869820 | on (NR_002368, intron 2 o | 723 | NR_002368 | 175501 | NM_001026083 | rpl-3 | Protein RPL-3 | protein-coding |
| MACS_peak_944 | chrI | 9163145 | 9166300 | (NM_001264082, exon 3 | 747 | NM_001264082 | 172743 | NM_001264081 | eef-2 | Protein EEF-2 | protein-coding |

# PeakAnalyzer



# PeakAnalyzer




Chromosomes nominate

| Chromosom | Start | End | Overlaped_Gene | nstream_FW_ | Symbol | Distance | nstream_REV_G | Symbol | Distance |
|---|---|---|---|---|---|---|---|---|---|
| chrI | 4057 | 4225 | 2 | Y74C9A.2.5 | nlp-40 | 6272 | Y74C9A.6 | Y74C9A.6 | 232 |
| chrI | 11337 | 11916 | 6 | Y74C9A.7 | 21ur-15479 | 19896 | Y74C9A.3.2 | Y74C9A.3 | 1394 |
| chrI | 24209 | 24363 | 2 | Y74C9A.7 | 21ur-15479 | 7237 | Y74C9A.3.2 | Y74C9A.3 | 14054 |
| chrI | 24574 | 24845 | 2 | Y74C9A.7 | 21ur-15479 | 6813 | Y74C9A.3.2 | Y74C9A.3 | 14477 |
| chrI | 26428 | 26877 | 2 | Y74C9A.7 | 21ur-15479 | 4870 | Y74C9A.3.2 | Y74C9A.3 | 16420 |
| chrI | 26947 | 27138 | 0 | Y74C9A.7 | 21ur-15479 | 4480 | Y74C9A.4a | Y74C9A.4 | 261 |
| chrI | 31939 | 32242 | 2 | Y74C9A.8 | 21ur-13439 | 324 | Y74C9A.4a | Y74C9A.4 | 5309 |
| chrI | 32367 | 32517 | 3 | Y74C9A.1 | Y74C9A.1 | 11291 | Y74C9A.4a | Y74C9A.4 | 5661 |
| chrI | 33680 | 33879 | 0 | Y74C9A.1 | Y74C9A.1 | 9953 | Y74C9A.5.1 | sesn-1 | 1297 |
| chrI | 34166 | 34463 | 0 | Y74C9A.1 | Y74C9A.1 | 9418 | Y74C9A.5.1 | sesn-1 | 1832 |
| chrI | 34664 | 35236 | 0 | Y74C9A.1 | Y74C9A.1 | 8783 | Y74C9A.5.1 | sesn-1 | 2468 |
| chrI | 35323 | 35973 | 0 | Y74C9A.1 | Y74C9A.1 | 8085 | Y74C9A.5.1 | sesn-1 | 3166 |
| chrI | 36197 | 36474 | 0 | Y74C9A.1 | Y74C9A.1 | 7397 | Y74C9A.5.1 | sesn-1 | 3853 |
| chrI | 39056 | 39344 | 0 | Y74C9A.1 | Y74C9A.1 | 4533 | Y74C9A.5.1 | sesn-1 | 6718 |
| chrI | 39399 | 39808 | 0 | Y74C9A.1 | Y74C9A.1 | 4129 | Y74C9A.5.1 | sesn-1 | 7121 |
| chrI | 39964 | 40124 | 0 | Y74C9A.1 | Y74C9A.1 | 3689 | Y74C9A.5.1 | sesn-1 | 7562 |
| chrI | 46926 | 47180 | 0 | Y48G1C.12 | Y48G1C.12 | 419 | Y74C9A.5.1 | sesn-1 | 14571 |
| chrI | 47354 | 47644 | 1 | Y48G1C.4 | pgs-1 | 2420 | Y74C9A.5.1 | sesn-1 | 15017 |
| chrI | 67971 | 68135 | 0 | Y48G1C.2.1 | csk-1 | 3805 | Y48G1C.5 | Y48G1C.5 | 4032 |
| chrI | 70100 | 70701 | 0 | Y48G1C.2.1 | csk-1 | 1457 | Y48G1C.5 | Y48G1C.5 | 6379 |
| chrI | 91706 | 91952 | 2 | Y48G1C.1 | Y48G1C.1 | 1202 | Y48G1C.6 | Y48G1C.6 | 5545 |

Nearest downstream genes

Nearest TSS

| Chromosom | PeakStart | PeakEnd | Distance | GeneStart | GeneEnd | ClosestTSS_ID | Symbol | Strand |
|---|---|---|---|---|---|---|---|---|
| chrX | 47975 | 48204 | 91 | 47799 | 48496 | Y73B3A.20 | Y73B3A.20 | + |
| chrX | 59416 | 61007 | 586 | 59625 | 59849 | Y73B3A.23 | Y73B3A.23 | + |
| chrX | 104546 | 104798 | 90 | 96342 | 104777 | Y73B3A.4 | Y73B3A.4 | - |
| chrX | 164109 | 164284 | -1062 | 162529 | 163134 | T08D2.1 | T08D2.1 | - |
| chrX | 191211 | 191392 | 514 | 191796 | 191816 | T08D2.10 | T08D2.10 | - |
| chrX | 322588 | 322946 | -88 | 322523 | 323214 | M02E1.3 | M02E1.3 | + |
| chrX | 324125 | 326331 | -734 | 325962 | 333711 | M02E1.1b.2 | M02E1.1 | + |
| chrX | 348080 | 348711 | -2106 | 344127 | 346289 | C04E7.3 | C04E7.3 | - |
| chrX | 353414 | 353985 | 79 | 353620 | 357934 | C04E7.2 | sor-3 | + |
| chrX | 370134 | 370415 | -1980 | 372234 | 376974 | R04A9.2.2 | nrde-3 | + |
| chrX | 382298 | 382961 | 74 | 381382 | 382710 | R04A9.4 | ife-2 | - |
| chrX | 383030 | 383210 | -416 | 381382 | 382710 | R04A9.4 | ife-2 | - |
| chrX | 388404 | 389275 | -47 | 384383 | 388798 | R04A9.5.2 | ceh-93 | - |
| chrX | 433977 | 434182 | 184 | 433895 | 434077 | ZK1193.8 | ZK1193.8 | + |
| chrX | 490076 | 490350 | -273 | 489869 | 489940 | F38G1.t2 | F38G1.t2 | - |
| chrX | 532795 | 533525 | -2198 | 530626 | 530962 | B0310.6 | B0310.6 | - |
| chrX | 535406 | 535924 | 83 | 531873 | 535835 | F28C10.3 | F28C10.3 | - |
| chrX | 536050 | 536430 | -492 | 531873 | 535835 | F28C10.3 | F28C10.3 | - |
| chrX | 590585 | 590773 | -3265 | 576319 | 587483 | F57C12.5b | mrp-1 | - |
| chrX | 593112 | 593315 | -766 | 593953 | 596299 | F13C5.2.2 | F13C5.2 | + |
| chrX | 593737 | 594463 | 120 | 593960 | 596299 | F13C5.2.1 | F13C5.2 | + |
| chrX | 601317 | 601467 | -826 | 602172 | 604922 | F13C5.1.2 | F13C5.1 | + |

Overlapped gene features

| Chromosom | Start | End | Overlap_Gene | Symbol | verlap_Begi | verlap_Cent | Overlap_End |
|---|---|---|---|---|---|---|---|
| chrI | 4057 | 4225 | Y74C9A.3.2 | Y74C9A.3 | LastExon | UTR3 | Intergenic |
| chrI | 4057 | 4225 | Y74C9A.3.1 | Y74C9A.3 | LastExon | UTR3 | Intergenic |
| chrI | 11337 | 11916 | Y74C9A.2.4 | nlp-40 | Intergenic | UTR5 | Intron1 |
| chrI | 11337 | 11916 | Y74C9A.2.6 | nlp-40 | Intergenic | UTR5 | Intron2 |
| chrI | 11337 | 11916 | Y74C9A.2.3 | nlp-40 | Intergenic | UTR5 | Intron2 |
| chrI | 11337 | 11916 | Y74C9A.2.1 | nlp-40 | Intergenic | UTR5 | Intron2 |
| chrI | 11337 | 11916 | Y74C9A.2.2 | nlp-40 | Intergenic | UTR5 | Intron2 |
| chrI | 11337 | 11916 | Y74C9A.2.5 | nlp-40 | Intron1 | UTR5 | Intron2 |
| chrI | 24209 | 24363 | Y74C9A.4b | Y74C9A.4 | Intron6 | Intron6 | Intron6 |
| chrI | 24209 | 24363 | Y74C9A.4a | Y74C9A.4 | Intron6 | Intron6 | Intron6 |
| chrI | 24574 | 24845 | Y74C9A.4b | Y74C9A.4 | Exon6 | Exon6 | Intron6 |
| chrI | 24574 | 24845 | Y74C9A.4a | Y74C9A.4 | Exon6 | Exon6 | Intron6 |
| chrI | 26428 | 26877 | Y74C9A.4b | Y74C9A.4 | Intergenic | Exon2 | Exon3 |
| chrI | 26428 | 26877 | Y74C9A.4a | Y74C9A.4 | Intergenic | Exon2 | Exon3 |
| chrI | 31939 | 32242 | Y74C9A.5.1 | sesn-1 | Intron1 | Intron1 | Exon2 |
| chrI | 31939 | 32242 | Y74C9A.5.2 | sesn-1 | Intron1 | Intron1 | Exon2 |
| chrI | 32367 | 32517 | Y74C9A.8 | 21ur-13439 | Intergenic | Intergenic | Intergenic |
| chrI | 32367 | 32517 | Y74C9A.5.1 | sesn-1 | Intergenic | Exon1 | Intron1 |
| chrI | 32367 | 32517 | Y74C9A.5.2 | sesn-1 | Intergenic | Exon1 | Intron1 |
| chrI | 47354 | 47644 | Y48G1C.12 | Y48G1C.12 | Intergenic | UTR5 | Intron1 |
| chrI | 91706 | 91952 | Y48G1C.9.2 | Y48G1C.9 | Intron1 | Intron1 | Intron1 |

# ChIPseeker

- Construct transcript database

```
> listMarts()
                         biomart                                                                  version
1                        ensembl                                                   ENSEMBL GENES 79 (SANGER UK)
2                            snp                                                 ENSEMBL VARIATION 79 (SANGER UK)
3                     regulation                                                ENSEMBL REGULATION 79 (SANGER UK)
4                           vega                                                            VEGA 59  (SANGER UK)
5                   fungi_mart_26                                                     ENSEMBL FUNGI 26 (EBI UK)
6              fungi_variations_26                                            ENSEMBL FUNGI VARIATION 26 (EBI UK)
7                  metazoa_mart_26                                                   ENSEMBL METAZOA 26 (EBI UK)
8            metazoa_variations_26                                          ENSEMBL METAZOA VARIATION 26 (EBI UK)
9                   plants_mart_26                                                    ENSEMBL PLANTS 26 (EBI UK)
10            plants_variations_26                                           ENSEMBL PLANTS VARIATION 26 (EBI UK)
11               protists_mart_25                                                  ENSEMBL PROTISTS 25 (EBI UK)
12          protists_variations_26                                         ENSEMBL PROTISTS VARIATION 26 (EBI UK)
13                             msd                                                                 MSD (EBI UK)
14                     cg_mart_02                                     PROTEOMICS (UNIVERSITY OF CAMBRIDGE - UK)
15                           htgt                                       WTSI MOUSE GENETICS PROJECT (SANGER UK)
16                          WS220                                                      WORMBASE 220 (CSHL US)
17                   parasite_mart                                                      ParaSite Mart (EBI UK)
18                         biomart                                                 MGI (JACKSON LABORATORY US)
19                         example                                            FANTOM5 phase1.1 (RIKEN CSLST Japan)
20                           pride                                                            PRIDE (EBI UK)
21                 prod-intermart_1                                                        INTERPRO (EBI UK)
```

# ChIPseeker

- Construct transcript database

```
> species_version<-useMart ("ensembl")
> listDatasets(species_version)
                          dataset                                description         version
1          oanatinus_gene_ensembl      Ornithorhynchus anatinus genes (OANA5)           OANA5
2         cporcellus_gene_ensembl             Cavia porcellus genes (cavPor3)         cavPor3
3         gaculeatus_gene_ensembl      Gasterosteus aculeatus genes (BROADS1)         BROADS1
4          lafricana_gene_ensembl          Loxodonta africana genes (loxAfr3)         loxAfr3
5  itridecemlineatus_gene_ensembl  Ictidomys tridecemlineatus genes (spetri2)         spetri2
6         choffmanni_gene_ensembl         Choloepus hoffmanni genes (choHof1)         choHof1
7          csavignyi_gene_ensembl              Ciona savignyi genes (CSAV2.0)         CSAV2.0
8             fcatus_gene_ensembl         Felis catus genes (Felis_catus_6.2) Felis_catus_6.2
9        rnorvegicus_gene_ensembl          Rattus norvegicus genes (Rnor_5.0)        Rnor_5.0
10         psinensis_gene_ensembl      Pelodiscus sinensis genes (PelSin_1.0)      PelSin_1.0
11          cjacchus_gene_ensembl   Callithrix jacchus genes (C_jacchus3.2.1)  C_jacchus3.2.1
12        ttruncatus_gene_ensembl          Tursiops truncatus genes (turTru1)         turTru1
13       scerevisiae_gene_ensembl   Saccharomyces cerevisiae genes (R64-1-1)        R64-1-1
14         celegans_gene_ensembl    Caenorhabditis elegans genes (WBcel235)       WBcel235
15          csabaeus_gene_ensembl       Chlorocebus sabaeus genes (ChlSab1.1)       ChlSab1.1
16        oniloticus_gene_ensembl     Oreochromis niloticus genes (Oreni11.0)       Oreni11.0
17         trubripes_gene_ensembl           Takifugu rubripes genes (FUGU4.0)         FUGU4.0
18        amexicanus_gene_ensembl        Astyanax mexicanus genes (AstMex102)       AstMex102
19          pmarinus_gene_ensembl     Petromyzon marinus genes (Pmarinus_7.0)    Pmarinus_7.0
20        eeuropaeus_gene_ensembl         Erinaceus europaeus genes (eriEur1)         eriEur1
21       falbicollis_gene_ensembl      Ficedula albicollis genes (FicAlb_1.4)      FicAlb_1.4
22      ptroglodytes_gene_ensembl          Pan troglodytes genes (CHIMP2.1.4)      CHIMP2.1.4
23         etelfairi_gene_ensembl            Echinops telfairi genes (TENREC)          TENREC
24     cintestinalis_gene_ensembl               Ciona intestinalis genes (KH)              KH
```

# ChIPseeker

```
transcriptsDb <- makeTranscriptDbFromBiomart(biomart="ensembl",
circ_seqs=NULL ,dataset="celegans_gene_ensembl")
```

```
> transcriptsDb <- makeTranscriptDbFromBiomart(biomart="ensembl", circ_seqs=NULL ,dataset="celegans_gene_ensembl")
Download and preprocess the 'transcripts' data frame ... OK
Download and preprocess the 'chrominfo' data frame ... OK
Download and preprocess the 'splicings' data frame ... OK
Download and preprocess the 'genes' data frame ... OK
Prepare the 'metadata' data frame ... OK
```

```
Txpts <- transcripts(transcriptsDb)
```

```
> Txpts <- transcripts(transcriptsDb)
> Txpts
GRanges with 57834 ranges and 2 metadata columns:
          seqnames               ranges strand   |     tx_id     tx_name
             <Rle>            <IRanges>  <Rle>   | <integer> <character>
      [1]        I       [10413, 16842]      +   |         1 Y74C9A.2a.2
      [2]        I       [11495, 16793]      +   |         2 Y74C9A.2b.2
      [3]        I       [11499, 16842]      +   |         3 Y74C9A.2b.1
      [4]        I       [11499, 16842]      +   |         4 Y74C9A.2b.4
      [5]        I       [11505, 16842]      +   |         5 Y74C9A.2a.3
      ...      ...                  ...    ... ...       ...         ...
  [57830]        X [17670595, 17670737]      -   |     57794     T23E7.9
  [57831]        X [17673476, 17673496]      -   |     57795     T23E7.7
  [57832]        X [17680911, 17682294]      -   |     57796     T23E7.6
  [57833]        X [17697239, 17697346]      -   |     57797     F20B4.8
  [57834]        X [17710486, 17711674]      -   |     57798    H11L12.1
  ---
  seqlengths:
          I       II      III       IV        V        X    MtDNA
   15072434 15279421 13783801 17493829 20924180 17718942    13794
```

```
saveDb(transcriptsDb,file="Cel235DB.sqlite")
txdb<-loadDb('Cel235DB.sqlite')
```

# All Packages

**Bioconductor version 3.1 (Release)**

Autocomplete biocViews search:

ChipManufacturer (388)
ChipName (195)
CustomArray (2)
CustomCDF (16)
CustomDBSchema (11)
FunctionalAnnotation (14)
Organism (550)
PackageType (543)
BSgenome (74)
cdf (126)
ChipDb (157)
db0 (19)
FRMA (10)
InparanoidDb (8)
MeSHDb (3)
OrganismDb (3)
OrgDb (19)

**Packages found under OrgDb:**

Show All entries

Search table:

| Package | Maintainer | Title |
|---|---|---|
| org.Ag.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Anopheles |
| org.At.tair.db | Bioconductor Package Maintainer | Genome wide annotation for Arabidopsis |
| org.Bt.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Bovine |
| org.Ce.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Worm |
| org.Cf.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Canine |
| org.Dm.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Fly |
| org.Dr.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Zebrafish |
| org.EcK12.eg.db | Bioconductor Package Maintainer | Genome wide annotation for E coli strain K12 |
| org.EcSakai.eg.db | Bioconductor Package Maintainer | Genome wide annotation for E coli strain Sakai |
| org.Gg.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Chicken |
| org.Hs.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Human |
| org.Mm.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Mouse |
| org.Mmu.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Rhesus |

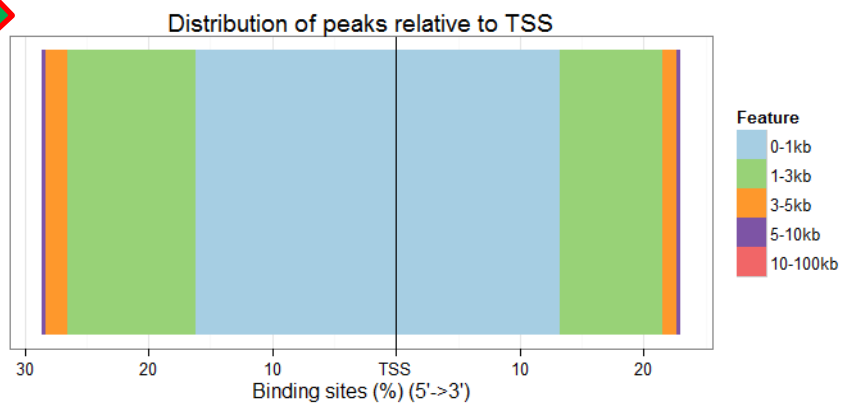
# ChIPseeker

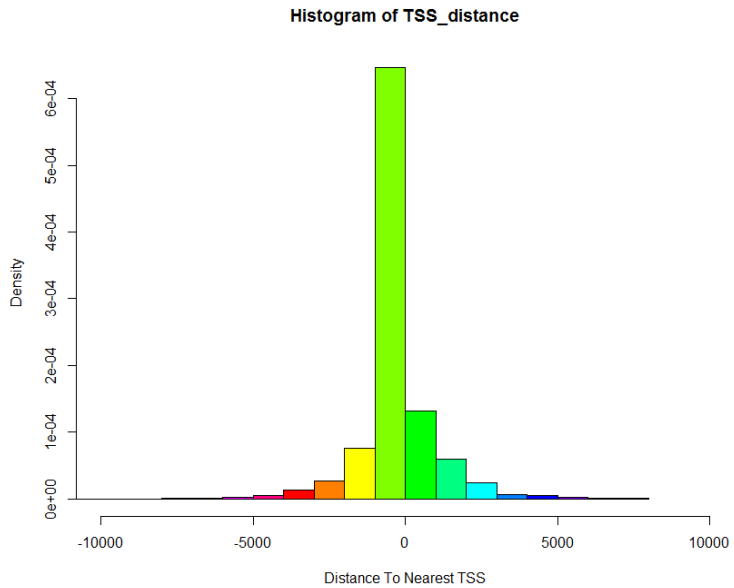
```
peak<- readPeakFile("test_peak.txt", as="GRanges")
peakAnno <- annotatePeak(peak, tssRegion=c(-3000,3000), TranscriptDb = txdb,
annoDb="org.Ce.eg.db")
```

| seqnames | start | end | width | strand | length | summit | tags | .log10.pval | _enrichm | FDR... | annotation | geneChr | geneStart | geneEnd | eneLengt | neStrar | geneId | transcriptId | distanceToTSS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I | 4058 | 4225 | 168 | * | 168 | 4136 | 46 | 12.7 | 3.57 | 10.99 | Promoter (<=1kb) | I | 3747 | 3909 | 163 | - | WBGene00023193 | Y74C9A.6 | -316 |
| I | 11338 | 11916 | 579 | * | 579 | 11448 | 58 | 20.14 | 4.5 | 18.26 | Promoter (<=1kb) | I | 11495 | 16793 | 5299 | + | WBGene00022276 | Y74C9A.2b.2 | 0 |
| I | 24210 | 24363 | 154 | * | 154 | 24269 | 39 | 8.94 | 3.03 | 7.34 | Promoter (<=1kb) | I | 17483 | 24796 | 7314 | - | WBGene00022278 | Y74C9A.4d | 586 |
| I | 24575 | 24845 | 271 | * | 271 | 24707 | 54 | 17.52 | 4.19 | 15.7 | Promoter (<=1kb) | I | 17483 | 24796 | 7314 | - | WBGene00022278 | Y74C9A.4d | 0 |
| I | 26429 | 26877 | 449 | * | 449 | 26696 | 111 | 63.35 | 8.58 | 60.77 | Promoter (<=1kb) | I | 17483 | 26643 | 9161 | - | WBGene00022278 | Y74C9A.4c | 0 |
| I | 26948 | 27138 | 191 | * | 191 | 27019 | 46 | 11.21 | 3.23 | 9.54 | Promoter (<=1kb) | I | 17483 | 26781 | 9299 | - | WBGene00022278 | Y74C9A.4a | -357 |
| I | 31940 | 32242 | 303 | * | 303 | 32133 | 39 | 6.41 | 2.42 | 4.91 | Promoter (<=1kb) | I | 32415 | 32435 | 21 | + | WBGene00173569 | Y74C9A.8 | -475 |
| I | 32368 | 32517 | 150 | * | 150 | 32460 | 32 | 5.7 | 2.48 | 4.23 | Promoter (<=1kb) | I | 27595 | 32482 | 4888 | - | WBGene00022279 | Y74C9A.5.2 | 0 |
| I | 33681 | 33879 | 199 | * | 199 | 33771 | 53 | 15.81 | 3.87 | 14.02 | Promoter (1-2kb) | I | 27595 | 32482 | 4888 | - | WBGene00022279 | Y74C9A.5.2 | -1397 |
| I | 34167 | 34463 | 297 | * | 297 | 34337 | 46 | 10.97 | 3.17 | 9.31 | Promoter (1-2kb) | I | 27595 | 32482 | 4888 | - | WBGene00022279 | Y74C9A.5.2 | -1981 |
| I | 34665 | 35236 | 572 | * | 572 | 35010 | 60 | 19.04 | 4.14 | 17.18 | Promoter (2-3kb) | I | 27595 | 32482 | 4888 | - | WBGene00022279 | Y74C9A.5.2 | -2754 |
| I | 35324 | 35973 | 650 | * | 650 | 35434 | 68 | 24.29 | 4.69 | 22.33 | Downstream (2-3kb) | I | 27595 | 32482 | 4888 | - | WBGene00022279 | Y74C9A.5.2 | -3491 |
| I | 36198 | 36474 | 277 | * | 277 | 36321 | 50 | 15.05 | 3.88 | 13.28 | Distal Intergenic | I | 27595 | 32482 | 4888 | - | WBGene00022279 | Y74C9A.5.2 | -3992 |
| I | 39057 | 39344 | 288 | * | 288 | 39151 | 36 | 7.27 | 2.74 | 5.74 | Distal Intergenic | I | 43733 | 44677 | 945 | + | WBGene00022275 | Y74C9A.1 | -4676 |
| I | 39400 | 39808 | 409 | * | 409 | 39535 | 50 | 14.74 | 3.81 | 12.98 | Distal Intergenic | I | 43733 | 44677 | 945 | + | WBGene00022275 | Y74C9A.1 | -4333 |
| I | 39965 | 40124 | 160 | * | 160 | 40040 | 41 | 8.9 | 2.92 | 7.31 | Distal Intergenic | I | 43733 | 44677 | 945 | + | WBGene00022275 | Y74C9A.1 | -3768 |
| I | 46927 | 47180 | 254 | * | 254 | 47068 | 59 | 20.81 | 4.58 | 18.92 | Promoter (<=1kb) | I | 47472 | 49819 | 2348 | + | WBGene00044345 | Y48G1C.12 | -545 |
| I | 47355 | 47644 | 290 | * | 290 | 47517 | 68 | 27.15 | 5.28 | 25.14 | Promoter (<=1kb) | I | 47472 | 49819 | 2348 | + | WBGene00044345 | Y48G1C.12 | 0 |
| I | 67972 | 68135 | 164 | * | 164 | 68030 | 40 | 8.94 | 2.98 | 7.34 | Distal Intergenic | I | 71425 | 81063 | 9639 | + | WBGene00000812 | Y48G1C.2b.2 | -3453 |
| I | 70101 | 70701 | 601 | * | 601 | 70314 | 77 | 27.79 | 4.78 | 25.77 | Promoter (1-2kb) | I | 71425 | 81063 | 9639 | + | WBGene00000812 | Y48G1C.2b.2 | -1324 |
| I | 91707 | 91952 | 246 | * | 246 | 91886 | 33 | 5.56 | 2.41 | 4.11 | Promoter (<=1kb) | I | 91404 | 92877 | 1474 | + | WBGene00021682 | Y48G1C.9.1 | 548 |
| I | 96148 | 96732 | 585 | * | 585 | 96498 | 56 | 18.82 | 4.35 | 16.97 | Promoter (1-2kb) | I | 94834 | 95367 | 534 | - | WBGene00189949 | Y48G1C.13 | -1365 |

# ChIPseeker

```
plotDistToTSS(peakAnno,title="Distribution of peaks relative to TSS")
```

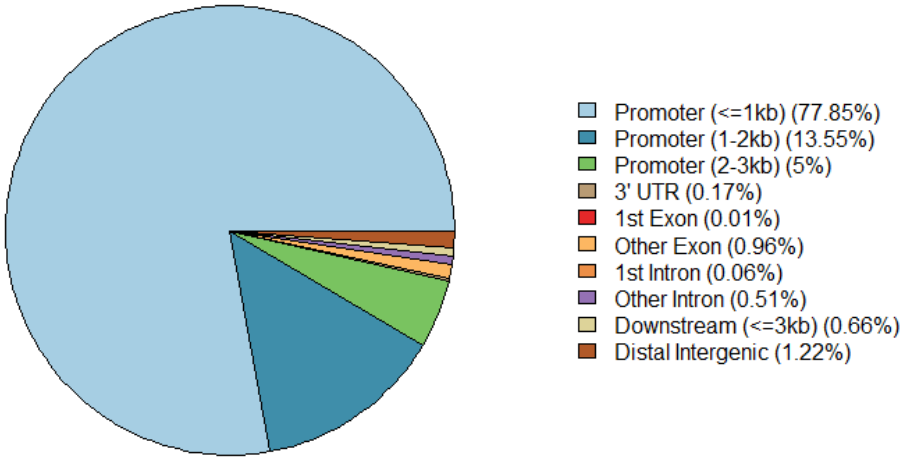


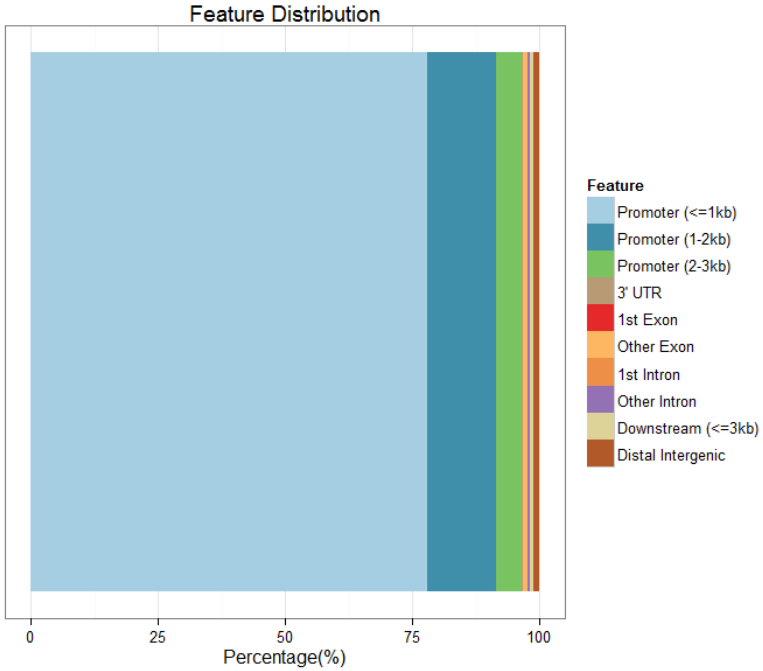

```
TSS_distance = peakAnno $distanceToTSS [!is.na(peakAnno $distanceToTSS)]
hist(TSS_distance, xlab = "Distance To Nearest TSS",prob=T, breaks = 20, xlim = c(-10000,10000),col="red")
```

# ChIPseeker

plotAnnoPie(peakAnno)

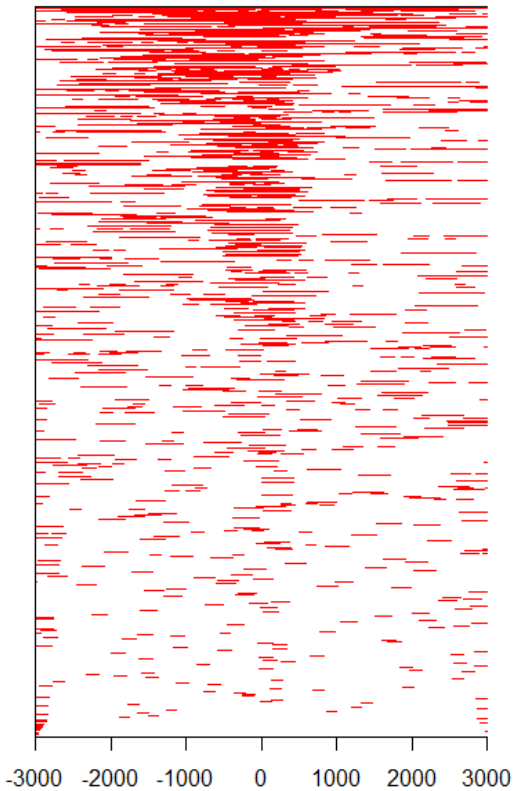


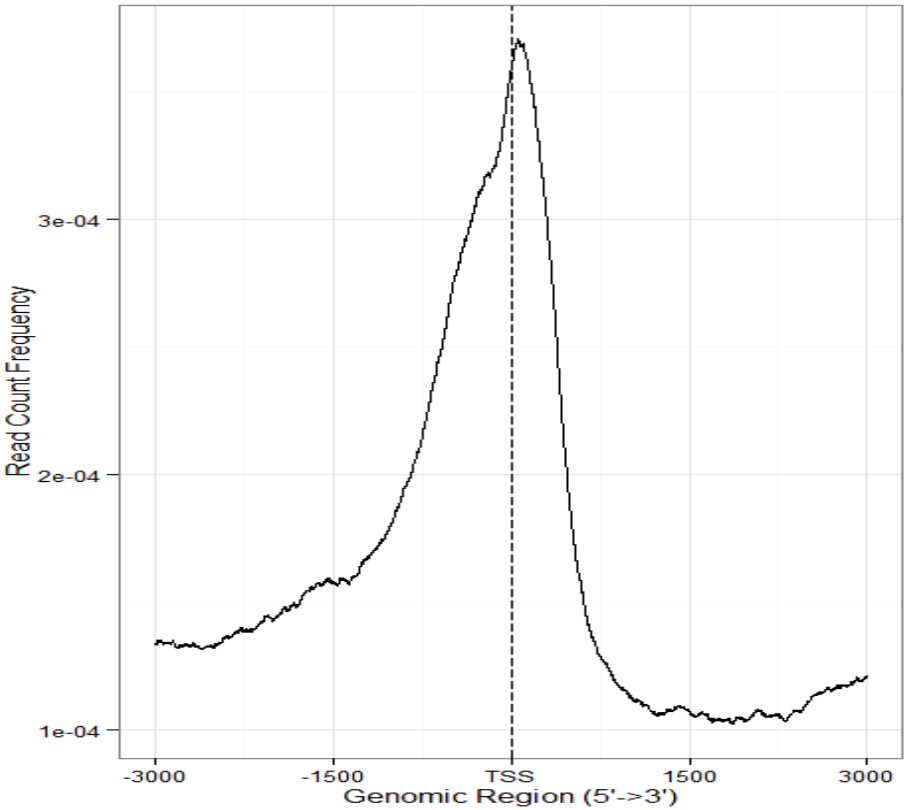

plotAnnoBar(peakAnno)



# ChIPseeker
# Density plotting

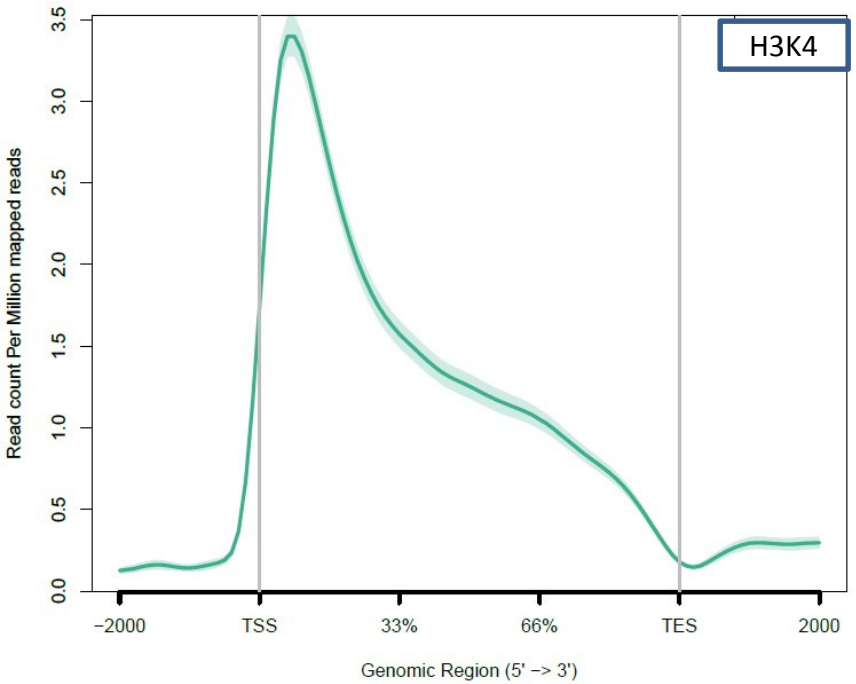
```
promoter <- getPromoters(TranscriptDb=txdb, upstream=3000, downstream=3000)
tagMatrix <- getTagMatrix(peak, weightCol=NULL, windows=promoter)
tagHeatmap(tagMatrix, xlim=c(-3000, 3000), color="red")
plotAvgProf(tagMatrix, xlim=c(-3000, 3000), xlab="Genomic Region (5'->3')", ylab = "Read Count Frequency")
```

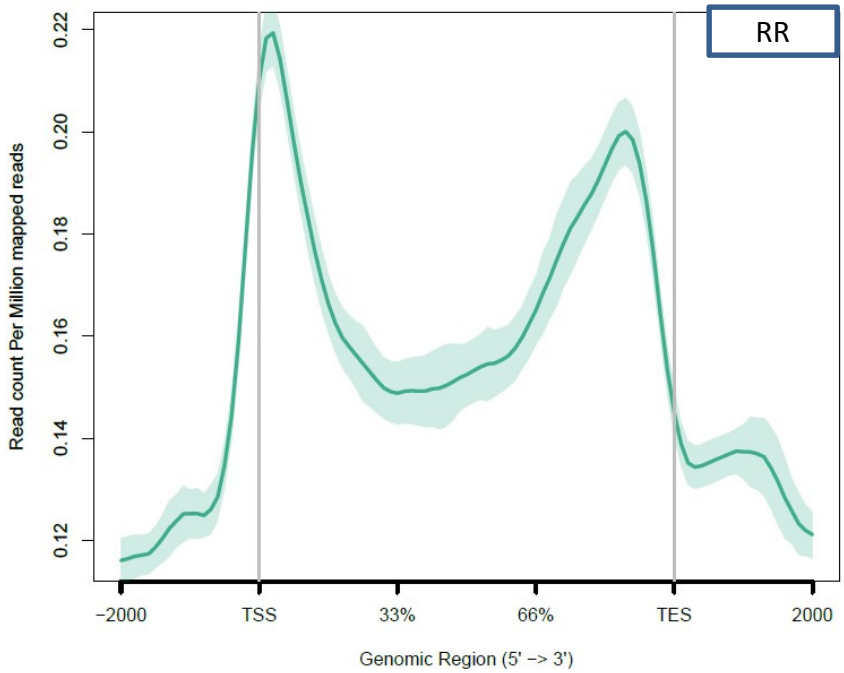


# ngsplot
# Density plotting

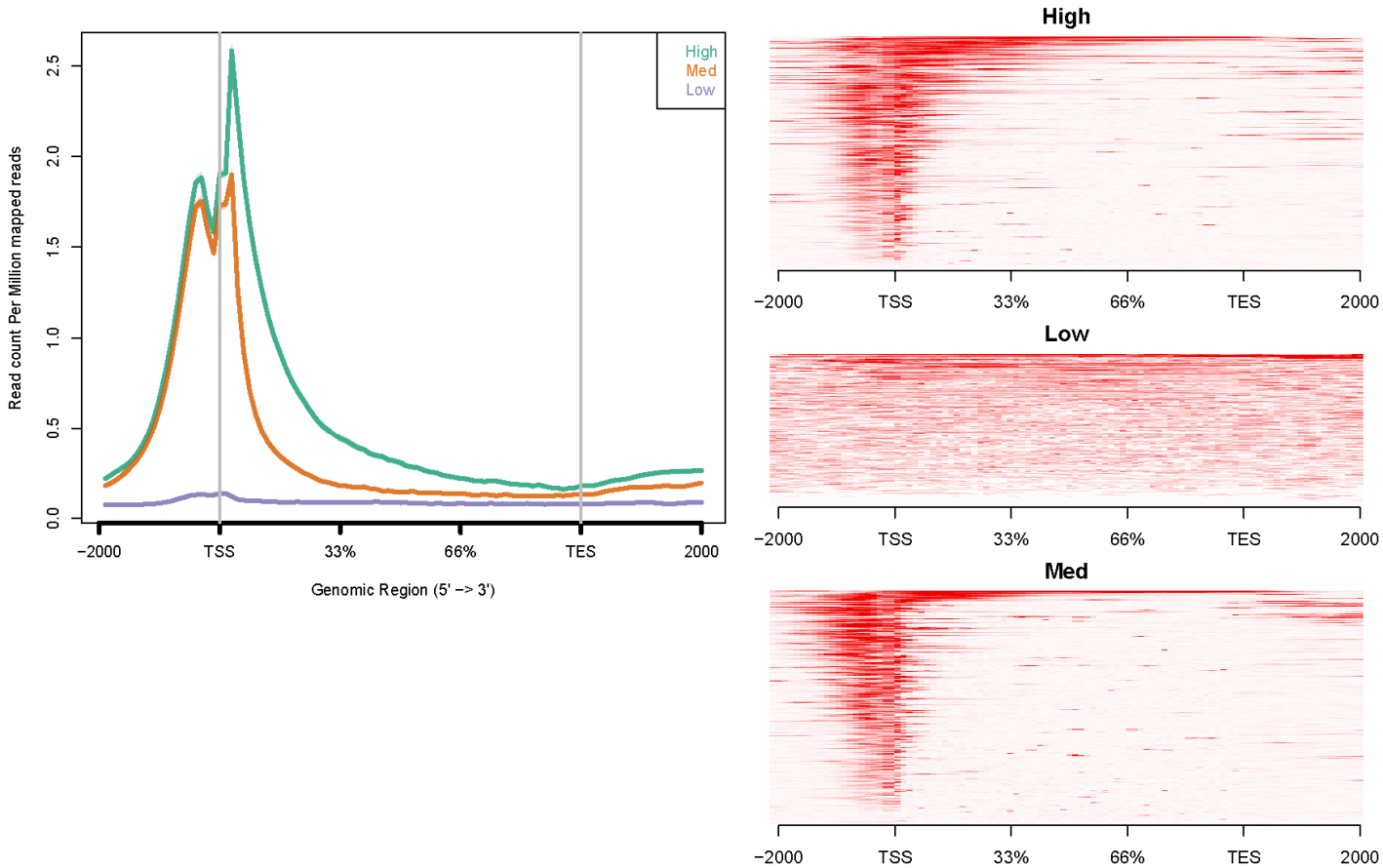
```
ngs.plot.r -G genome -R region -C [cov|config]file  -O name [Options]
```




https://code.google.com/p/ngsplot/

# ngsplot




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

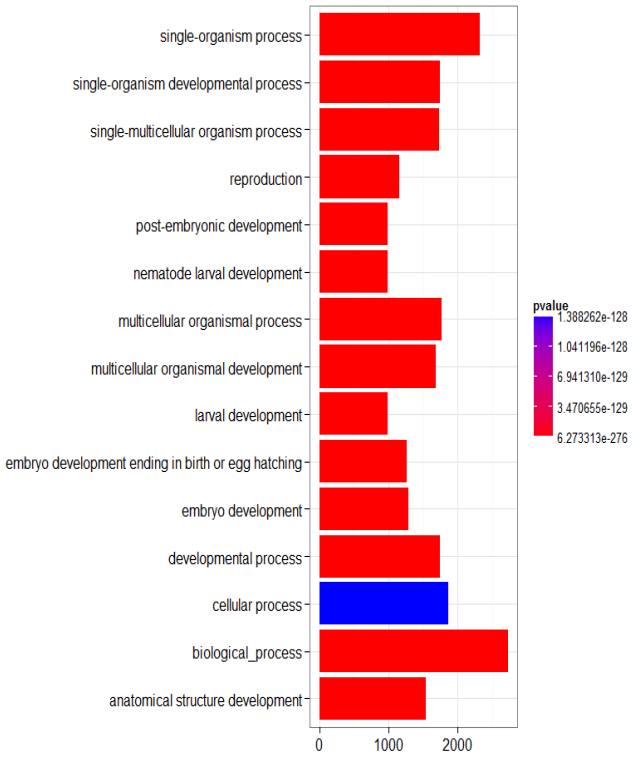
# ChIPseeker
# GO categories

```
ensemblID<- as.list(org.Ce.egENSEMBL2EG)
ensemblID_subset<-ensemblID [unique(as.data.frame(peakAnno)$geneId)]
entrez_ID<-matrix(unlist(ensemblID_subset),ncol=1,byrow=TRUE)
gene<-as.vector(entrez_ID)
ego<-enrichGO(gene,organism ="worm", ont = "MF", pvalueCutoff = 0.01,pAdjustMethod = "BH",qvalueCutoff = 0.05, minGSSize = 5, readable = TRUE)
```

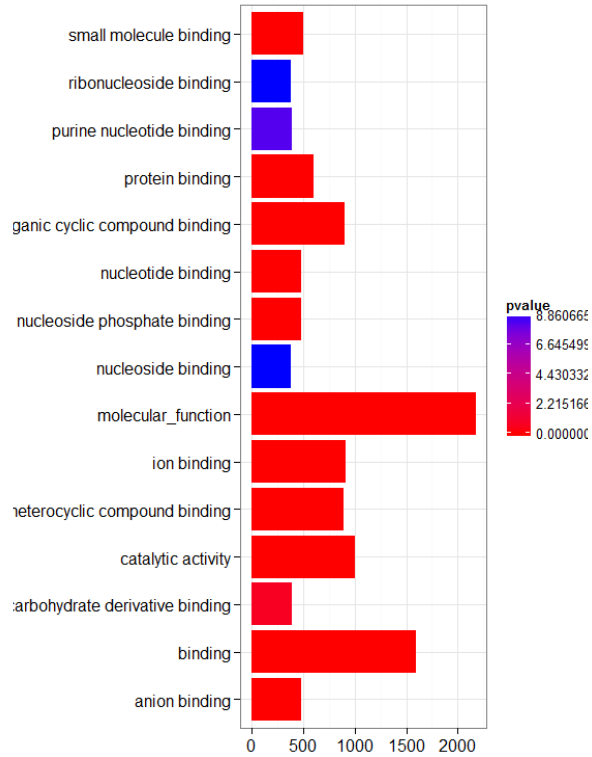
```
> head(summary(ego),n=3)
                   ID        Description GeneRatio   BgRatio        pvalue      p.adjust        qvalue
GO:0003674 GO:0003674 molecular_function 2172/2172 6566/9690  0.000000e+00  0.000000e+00  0.000000e+00
GO:0005488 GO:0005488            binding 1591/2172 4041/9690 2.907499e-252 4.550236e-250 1.361934e-250
GO:0043167 GO:0043167        ion binding  911/2172 2258/9690 2.054867e-110 2.143912e-108 6.416954e-109

GO:0003674 CELE_F17C8.9/C02D5.4/C28G1.10/rab-11.1/ztf-3/C53D5.5/xpo-2/CELE_R119.5/haf-4/acs-13/Y65B4BR.5/
GO:0005488
GO:0043167
           Count
GO:0003674  2172
GO:0005488  1591
GO:0043167   911
```

| | ID | Description | GeneRatio | BgRatio | pvalue | p.adjust | qvalue | geneID | Count |
|---|---|---|---|---|---|---|---|---|---|
| GO:0003674 | GO:0003674 | molecular_function | 2172/2172 | 6566/9690 | 0 | 0 | 0 | p-11/unc-108/guk-1/hcp | 2172 |
| GO:0005488 | GO:0005488 | binding | 1591/2172 | 4041/9690 | 2.91E-252 | 4.55E-250 | 1.36E-250 | ıp-1/chd-7/W09C3.4/let- | 1591 |
| GO:0043167 | GO:0043167 | ion binding | 911/2172 | 2258/9690 | 2.05E-110 | 2.14E-108 | 6.42E-109 | rp-1/drh-3/pck-2/nmy-2/ | 911 |
| GO:0097159 | GO:0097159 | organic cyclic compound binding | 902/2172 | 2236/9690 | 6.72E-109 | 5.26E-107 | 1.57E-107 | a-1/prg-1/drh-3/pck-2/n | 902 |
| GO:1901363 | GO:1901363 | heterocyclic compound binding | 897/2172 | 2229/9690 | 2.57E-107 | 1.61E-105 | 4.81E-106 | -1/prg-1/drh-3/pck-2/nn | 897 |
| GO:0005515 | GO:0005515 | protein binding | 601/2172 | 1253/9690 | 1.75E-103 | 9.13E-102 | 2.73E-102 | ab-5/ppfr-1/skr-1/klp-16 | 601 |
| GO:0003824 | GO:0003824 | catalytic activity | 999/2172 | 2709/9690 | 2.23E-94 | 9.98E-93 | 2.99E-93 | 'bli-4/gld-2/acdh-3/cdc-2 | 999 |
| GO:0036094 | GO:0036094 | small molecule binding | 502/2172 | 1119/9690 | 2.12E-71 | 8.31E-70 | 2.49E-70 | 5/CELE_Y106G6D.7/csnk- | 502 |
| GO:0000166 | GO:0000166 | nucleotide binding | 485/2172 | 1085/9690 | 4.75E-68 | 1.49E-66 | 4.45E-67 | Y106G6D.7/csnk-1/wts-1 | 485 |
| GO:1901265 | GO:1901265 | nucleoside phosphate binding | 485/2172 | 1085/9690 | 4.75E-68 | 1.49E-66 | 4.45E-67 | Y106G6D.7/csnk-1/wts-1 | 485 |
| GO:0043168 | GO:0043168 | anion binding | 477/2172 | 1071/9690 | 3.65E-66 | 1.04E-64 | 3.11E-65 | '/ndk-1/C34B2.6/kin-3/B | 477 |

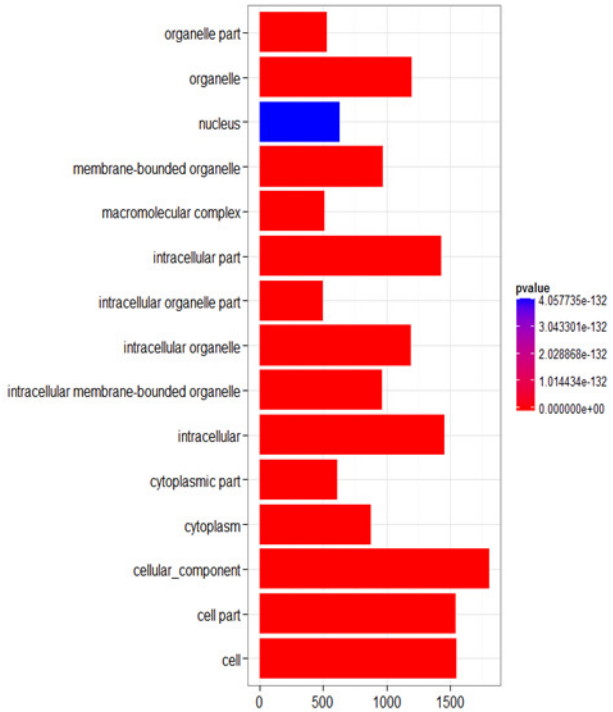
# GO enrichment




Biological process



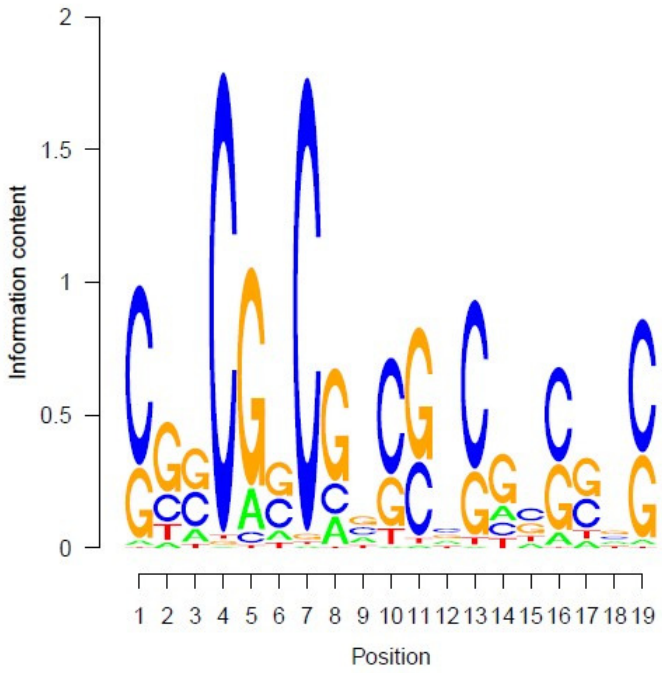

Molecular function

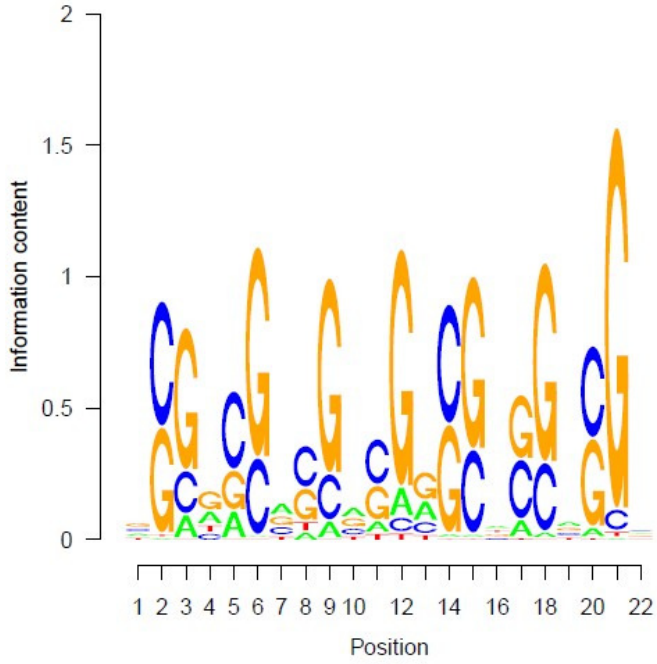



Cellular component

# Motif analysis

```
/programs/R-2.15.0/bin/R
library(BSgenome)
available.genomes()
library(MotIV)
library(ShortRead)
library(rGADEM)
library(rtracklayer)
library("BSgenome.Celegans.UCSC.ce10”)
sequences<- read.DNAStringSet("test_peak.fa","fasta")
motifs_macs_female=GADEM(sequences, genome=Celegans,verbose=TRUE,pValue=0.0002,eValue=-5,numGeneration=500)
```



# Motif analysis

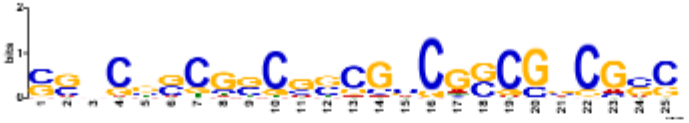
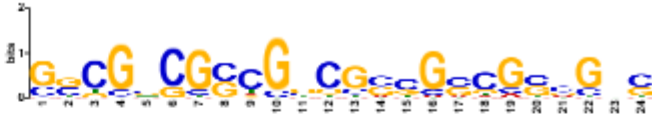
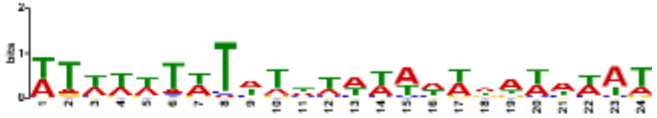
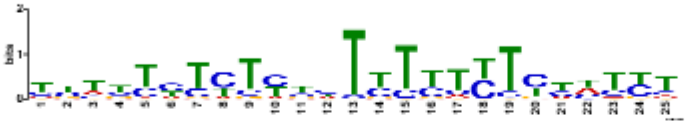
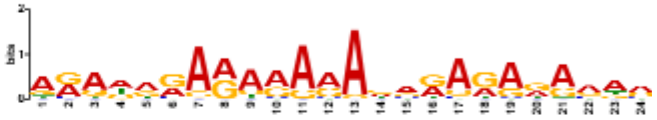
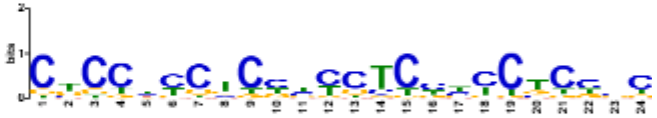
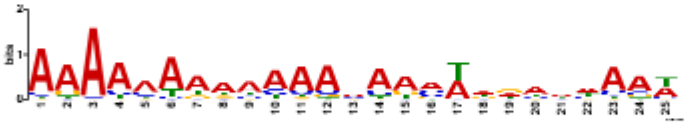
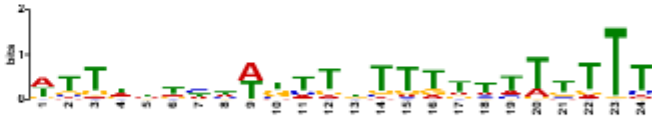
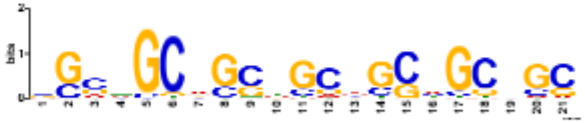
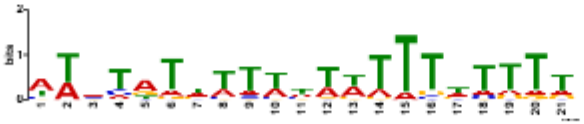
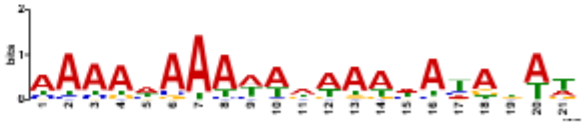
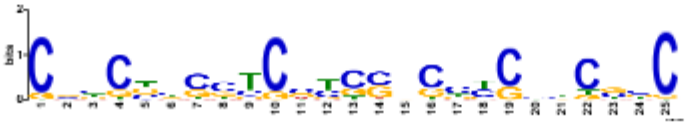
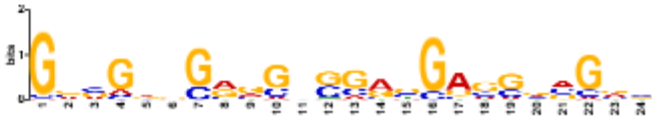
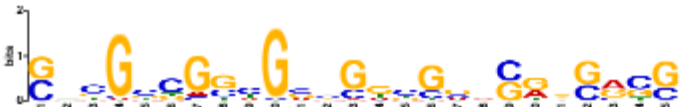
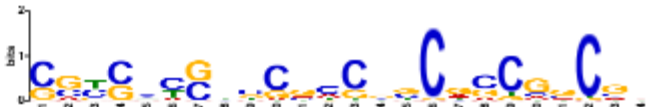
meme <input.fa> [optional arguments]

```
[-h]                      print this message
[-o <output dir>]         name of directory for output files will not
                          replace existing directory
[-oc <output dir>]        name of directory for output files will
                          replace existing directory
[-text]                   output in text format (default is HTML)
[-dna]                    sequences use DNA alphabet
[-protein]                sequences use protein alphabet
[-mod oops|zoops|anr]     distribution of motifs
[-nmotifs <nmotifs>]      maximum number of motifs to find
[-evt <ev>]               stop if motif E-value greater than <evt>
[-nsites <sites>]         number of sites for each motif
[-minsites <minsites>]    minimum number of sites for each motif
[-maxsites <maxsites>]    maximum number of sites for each motif
[-wnsites <wnsites>]      weight on expected number of sites
[-w <w>]                  motif width
[-minw <minw>]            minumum motif width
[-maxw <maxw>]            maximum motif width
[-nomatrim]               do not adjust motif width using multiple
                          alignments
[-wg <wg>]                gap opening cost for multiple alignments
[-ws <ws>]                gap extension cost for multiple alignments
[-noendgaps]              do not count end gaps in multiple alignments
[-bfile <bfile>]          name of background Markov model file
[-revcomp]                allow sites on + or - DNA strands
[-pal]                    force palindromes (requires -dna)
```

MEME ([http://meme.sdsc.edu/meme/cgi-bin/meme.cgi](http://meme.sdsc.edu/meme/cgi-bin/meme.cgi))

## DISCOVERED MOTIFS

### Motif Overview

| | | | |
|---|---|---|---|
| Motif 1 | • 8.6e-395<br>• 303 sites |  |  |
| Motif 2 | • 1.5e-336<br>• 411 sites | |  |
| Motif 3 | • 2.7e-233<br>• 296 sites |  |  |
| Motif 4 | • 2.7e-213<br>• 296 sites | | |
| Motif 5 | • 5.3e-135<br>• 287 sites | |  |
| Motif 6 | • 3.1e-120<br>• 293 sites |  |  |
| Motif 7 | • 1.2e-086<br>• 270 sites | |  |
| Motif 8 | • 3.2e-084<br>• 275 sites |  |  |
| Motif 9 | • 6.0e-108<br>• 296 sites |  |  |
| Motif 10 | • 4.3e-063<br>• 229 sites |  |  |