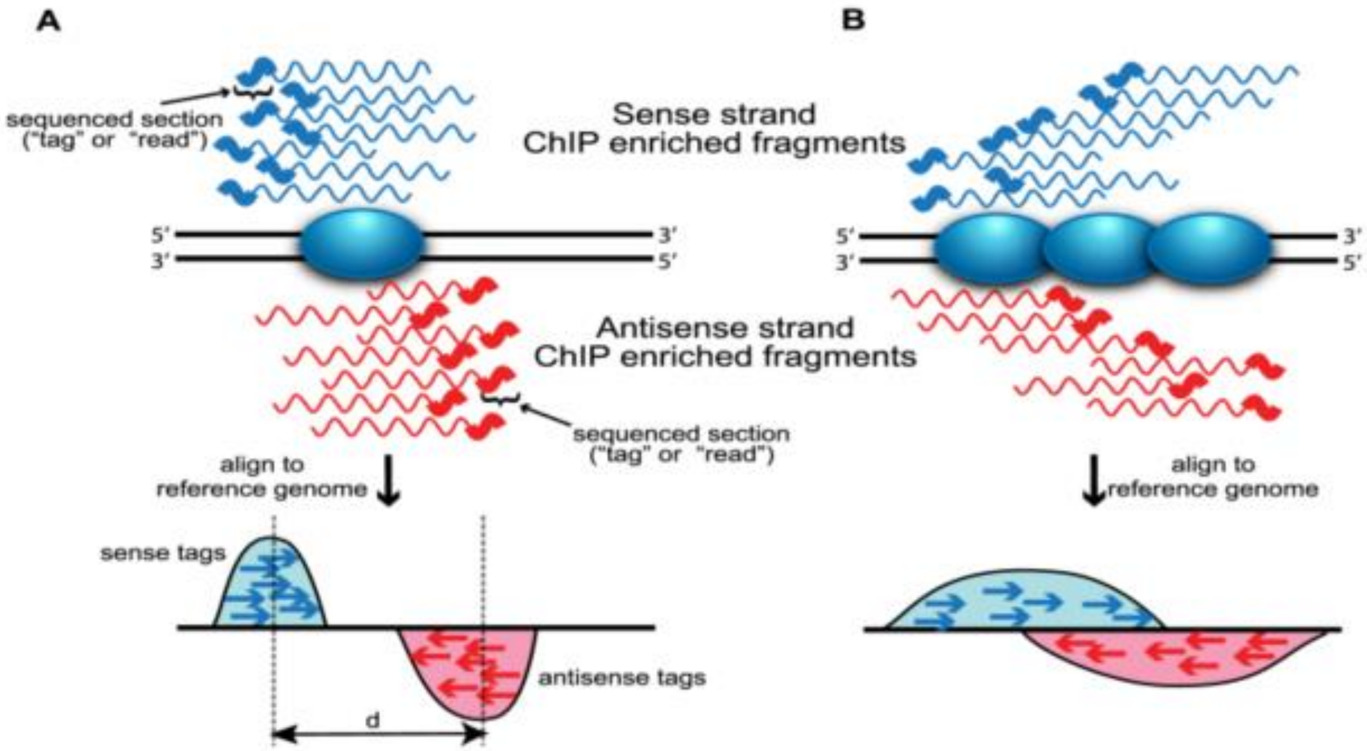
Functional annotation of ChIP-peaks

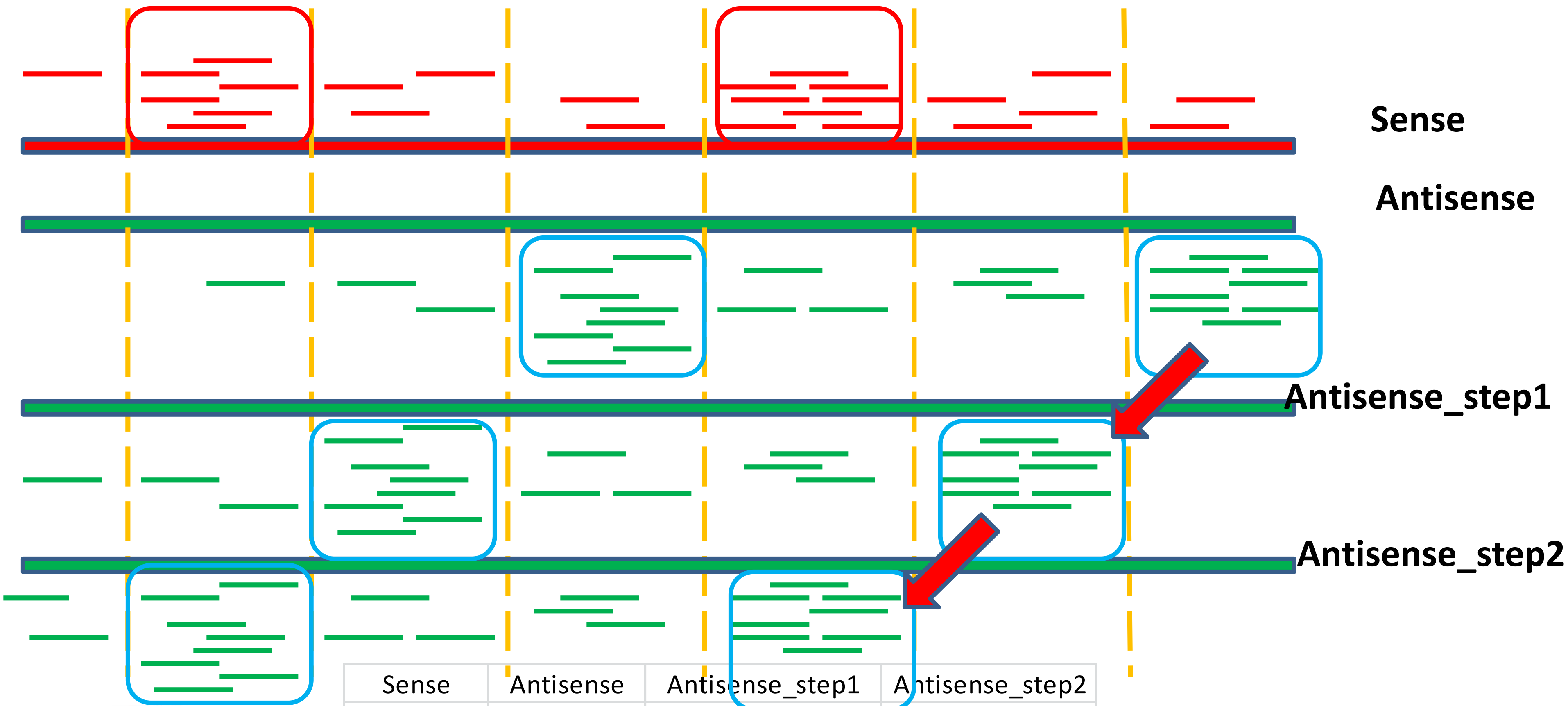
Minghui Wang, Qi Sun

Bioinformatics Facility

Institute of Biotechnology



Wilbanks et al. 2010 PLOS One



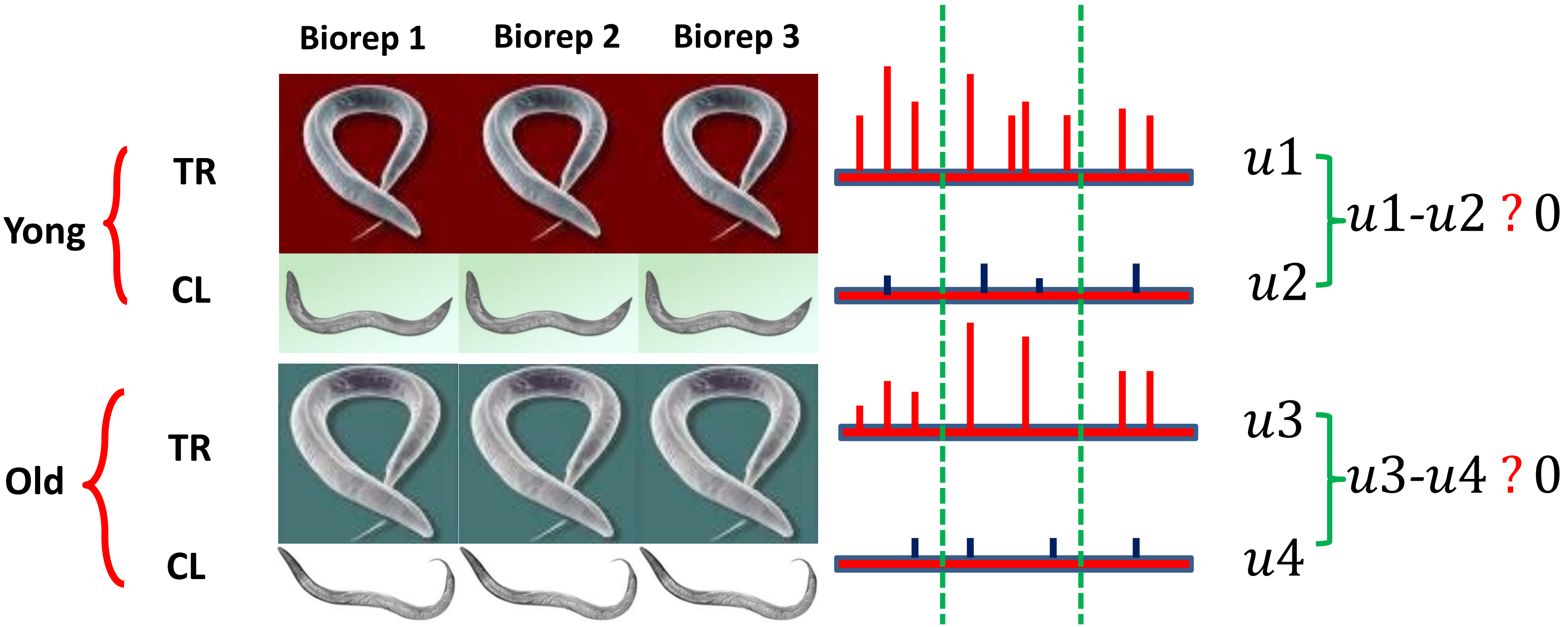
Sense	Antisense	Antisense_step1	Antisense_step2
1	0	1	2
6	1	2	8
3	2	8	3
2	8	3	3
8	3	3	8
4	3	8	0
2	8	0	0

R= -0.27

R= 0.13

R= 0.79

Experimental design



Biorep 1 Biorep 2 Biorep 3

Yong TR CL

Old TR CL

$u1$

$u2$

$u1\text{-}u2$? 0

$u3$

$u4$

$u3\text{-}u4$? 0

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

GLM (Poisson)

Reads counts

$$\underbrace{Y_i \begin{bmatrix}34\\12\\42\\18\\44\\20\\25\\10\\32\\15\\38\\14\end{bmatrix} = \begin{array}{cccc} \text{Mean} & \text{Time} & \text{IP} & \text{Time*IP} \\ \downarrow & \downarrow & \downarrow & \downarrow \\ \hline 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{array} \begin{bmatrix}\mu\\ \alpha\\ \beta\\ \alpha*\beta\end{bmatrix} + \varepsilon_i}_{Y = XB + E}$$

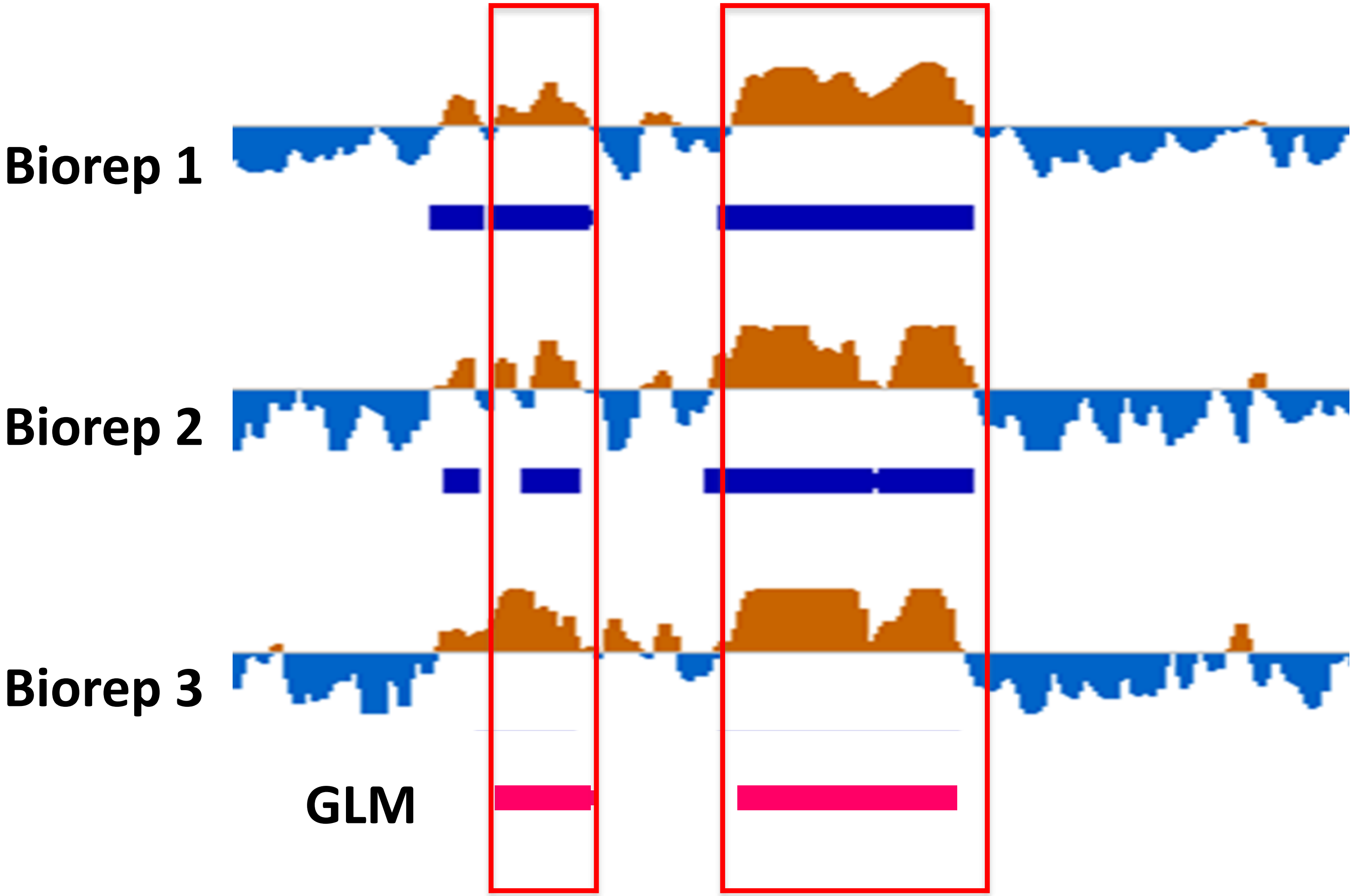
Design Matrix

Yong

Old

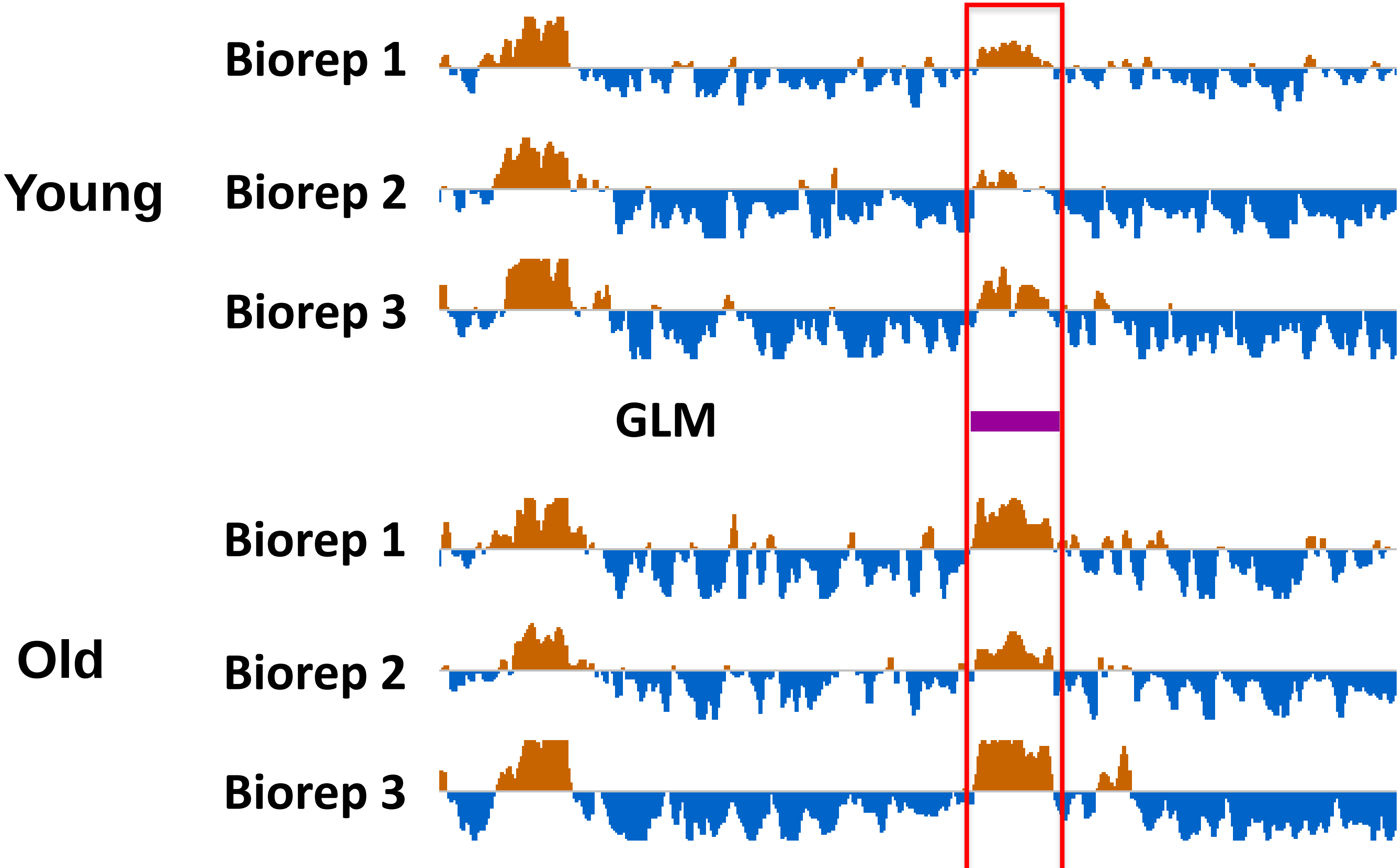
```
out<-glm(Y ~ Time*IP , family = poisson, log(offset=c(library size)))
```

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



Young

Biorep 1

Biorep 2

Biorep 3

GLM

Old

Biorep 1

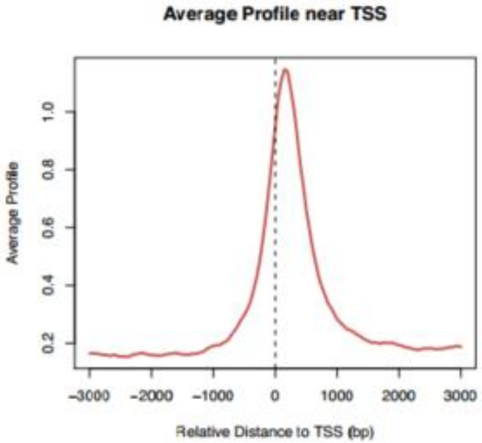
Biorep 2

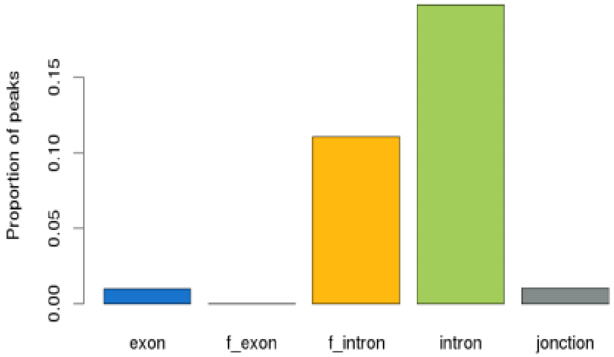
Biorep 3

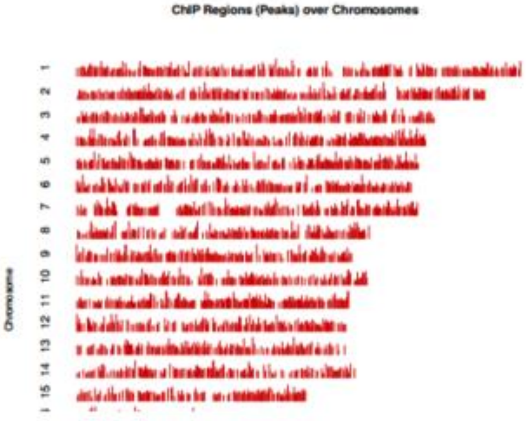
M pu (2015) Trimethylation of Lys36 on H3 restricts gene expression change during aging and impacts life span. Genes Dev 1;29(7):718-31

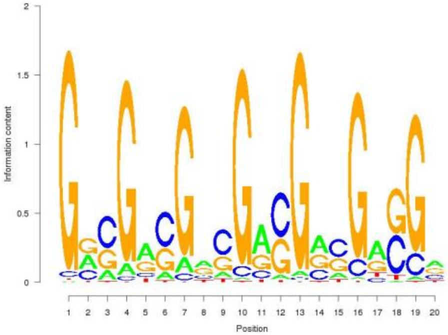
Visualization & Annotation

Enrichment profiles

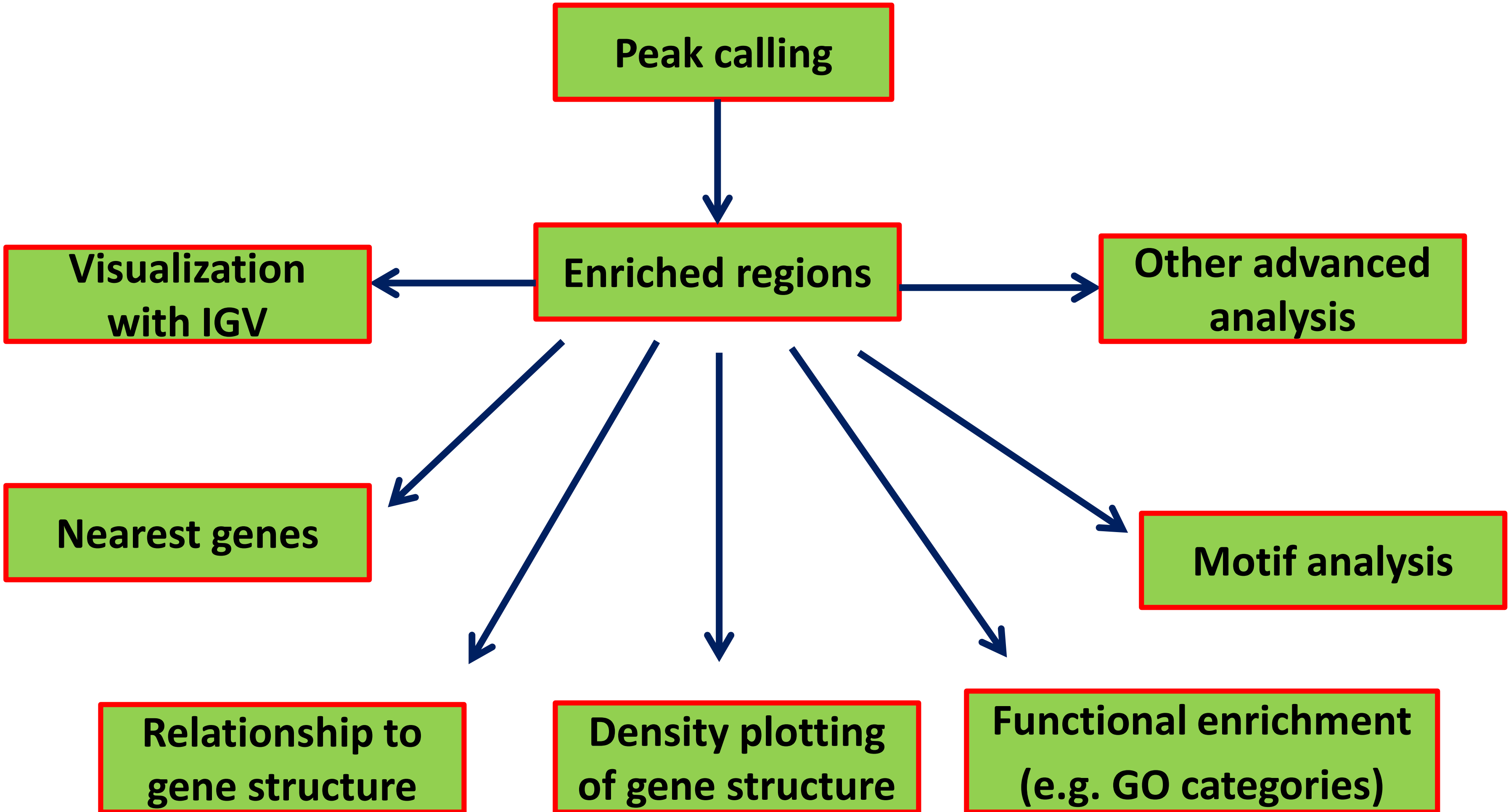








Functional annotation workflow



Peak calling

Enriched regions

Visualization with IGV

Other advanced analysis

Nearest genes

Motif analysis

Relationship to gene structure

Density plotting of gene structure

Functional enrichment (e.g. GO categories)

Annotating Peaks

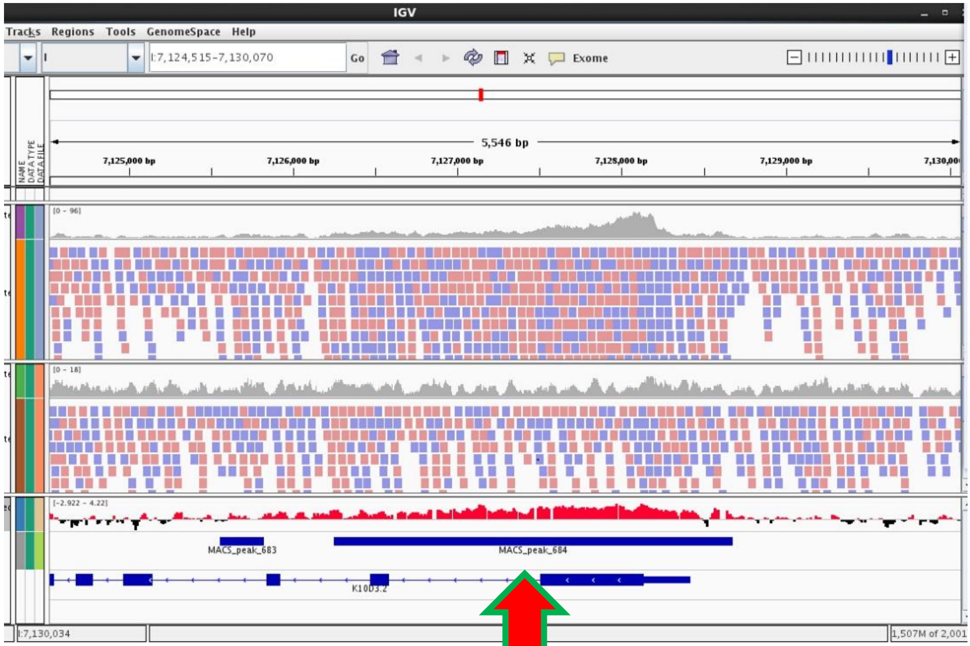
- Homer
- PeakAnalyzer
- ChIPpeakAnno
- ChIPseeker
- …

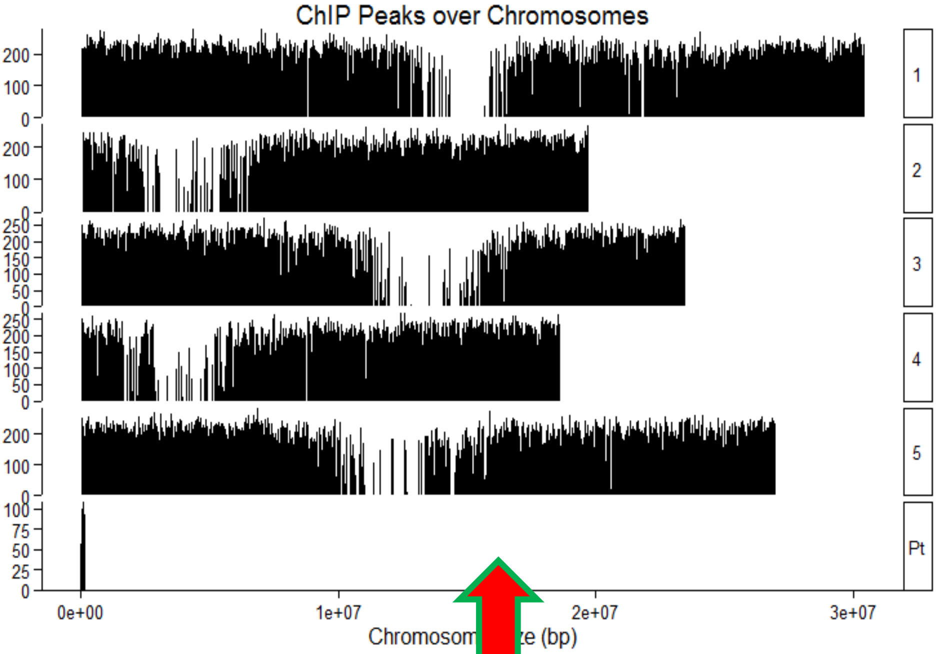
Peak region file

GFF (General Feature Format)
Genome annotation

R & Bioconductor

Visualization

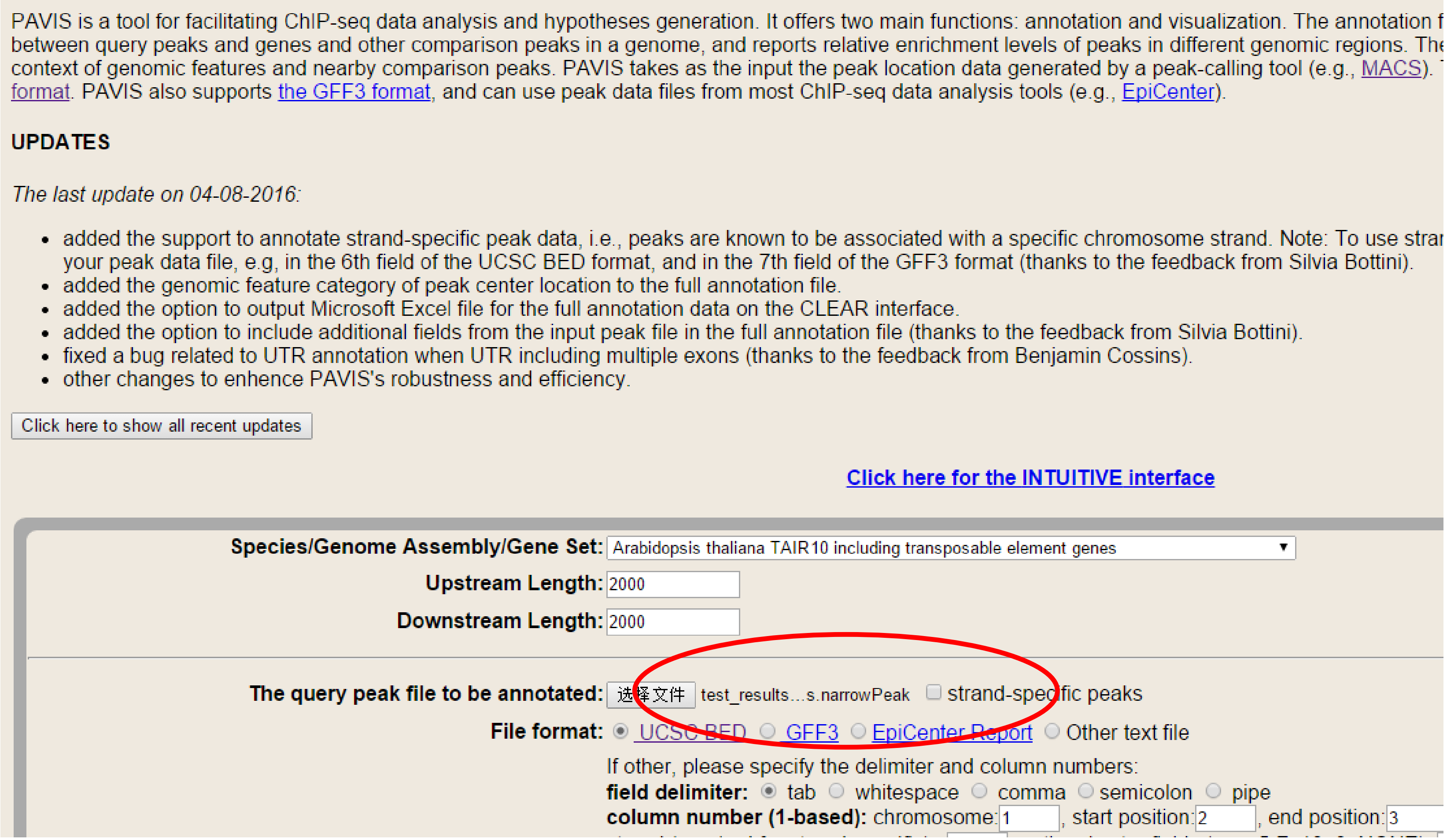




```
bamCompare --bamfile1 ChIP.bam --bamfile2 Input.bam \
--binSize 25 --fragmentLength 200 --missingDataAsZero no \
--ratio log2 --scaleFactorsMethod SES -o log2ratio_ChIP_vs_Input.bw
```

```
peak<-readPeakFile(“test_results_summits.bed”)
 covplot(peak, weightCol=“V5”)
```

PAVIS



PAVIS is a tool for facilitating ChIP-seq data analysis and hypotheses generation. It offers two main functions: annotation and visualization. The annotation f
between query peaks and genes and other comparison peaks in a genome, and reports relative enrichment levels of peaks in different genomic regions. The
context of genomic features and nearby comparison peaks. PAVIS takes as the input the peak location data generated by a peak-calling tool (e.g., MACS).
format. PAVIS also supports the GFF3 format, and can use peak data files from most ChIP-seq data analysis tools (e.g., EpiCenter).

UPDATES

The last update on 04-08-2016:

- added the support to annotate strand-specific peak data, i.e., peaks are known to be associated with a specific chromosome strand. Note: To use stra
your peak data file, e.g, in the 6th field of the UCSC BED format, and in the 7th field of the GFF3 format (thanks to the feedback from Silvia Bottini).
- added the genomic feature category of peak center location to the full annotation file.
- added the option to output Microsoft Excel file for the full annotation data on the CLEAR interface.
- added the option to include additional fields from the input peak file in the full annotation file (thanks to the feedback from Silvia Bottini).
- fixed a bug related to UTR annotation when UTR including multiple exons (thanks to the feedback from Benjamin Cossins).
- other changes to enhence PAVIS's robustness and efficiency.

Click here for the INTUITIVE interface

Species/Genome Assembly/Gene Set: Arabidopsis thaliana TAIR10 including transposable element genes ▼

Upstream Length: 2000

Downstream Length: 2000

The query peak file to be annotated: 选择文件 test_results...s.narrowPeak □ strand-specific peaks

File format: ● UCSC BED ○ GFF3 ○ EpiCenter Report ○ Other text file

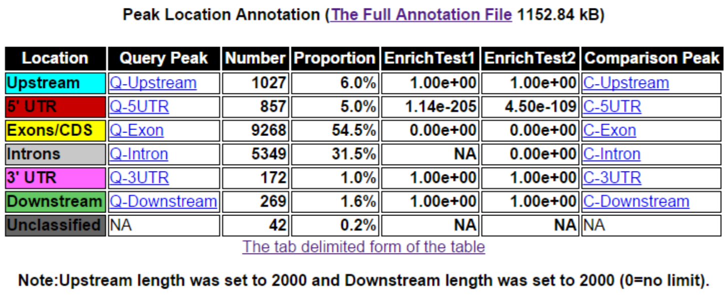
If other, please specify the delimiter and column numbers:

field delimiter: ● tab ○ whitespace ○ comma ○ semicolon ○ pipe

column number (1-based): chromosome: 1 , start position: 2 , end position: 3

http://manticore.niehs.nih.gov/pavis2/

PAVIS

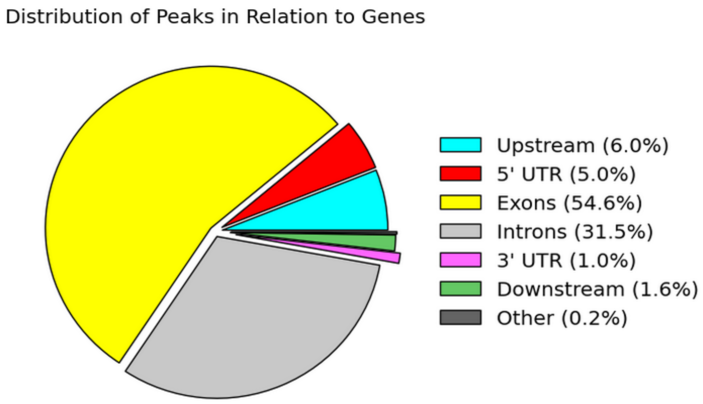


Peak Location Annotation (The Full Annotation File 1152.84 kB)

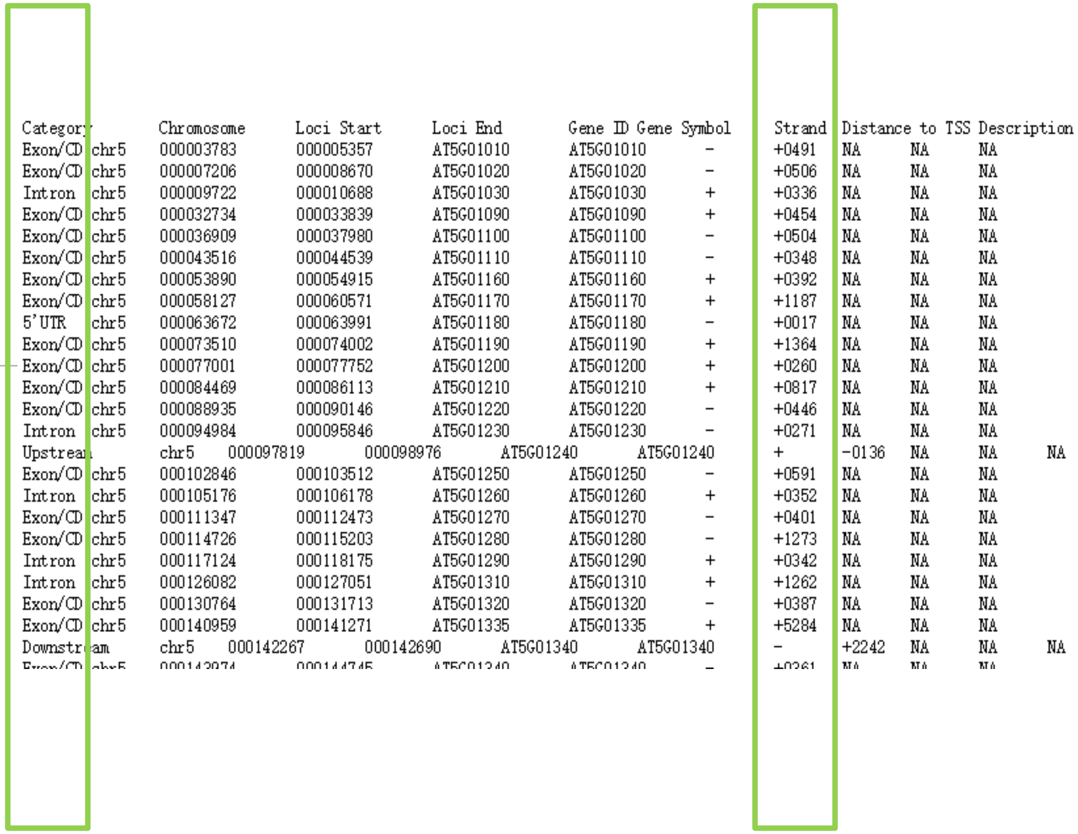
Location	Query Peak	Number	Proportion	EnrichTest1	EnrichTest2	Comparison Peak
Upstream	Q-Upstream	1027	6.0%	1.00e+00	1.00e+00	C-Upstream
5' UTR	Q-5UTR	857	5.0%	1.14e-205	4.50e-109	C-5UTR
Exons/CDS	Q-Exon	9268	54.5%	0.00e+00	0.00e+00	C-Exon
Introns	Q-Intron	5349	31.5%	NA	0.00e+00	C-Intron
3' UTR	Q-3UTR	172	1.0%	1.00e+00	1.00e+00	C-3UTR
Downstream	Q-Downstream	269	1.6%	1.00e+00	1.00e+00	C-Downstream
Unclassified	NA	42	0.2%	NA	NA	NA

The tab delimited form of the table

Note:Upstream length was set to 2000 and Downstream length was set to 2000 (0=no limit).



Distribution of Peaks in Relation to Genes

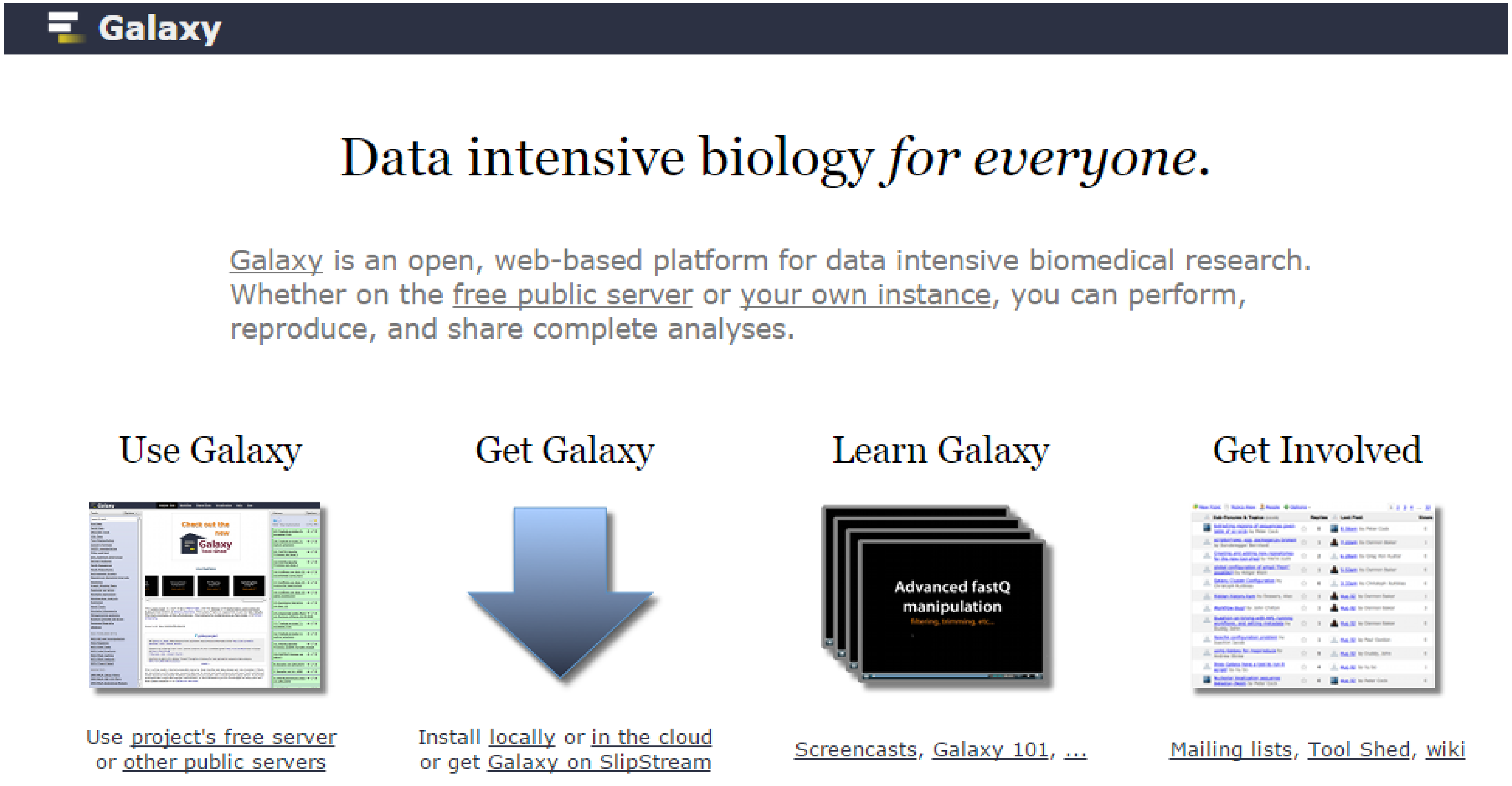


```
Category      Chromosome    Loci Start    Loci End      Gene ID Gene Symbol    Strand Distance to TSS Description
Exon/CD chr5  000003783     000005357     AT5G01010     AT5G01010     -        +0491  NA     NA     NA
Exon/CD chr5  000007206     000008670     AT5G01020     AT5G01020     -        +0506  NA     NA     NA
Intron  chr5  000009722     000010688     AT5G01030     AT5G01030     +        +0336  NA     NA     NA
Exon/CD chr5  000032734     000033839     AT5G01090     AT5G01090     +        +0454  NA     NA     NA
Exon/CD chr5  000036909     000037980     AT5G01100     AT5G01100     -        +0504  NA     NA     NA
Exon/CD chr5  000043516     000044539     AT5G01110     AT5G01110     -        +0348  NA     NA     NA
Exon/CD chr5  000053890     000054915     AT5G01160     AT5G01160     +        +0392  NA     NA     NA
Exon/CD chr5  000058127     000060571     AT5G01170     AT5G01170     +        +1187  NA     NA     NA
5'UTR   chr5  000063672     000063991     AT5G01180     AT5G01180     -        +0017  NA     NA     NA
Exon/CD chr5  000073510     000074002     AT5G01190     AT5G01190     +        +1364  NA     NA     NA
Exon/CD chr5  000077001     000077752     AT5G01200     AT5G01200     +        +0260  NA     NA     NA
Exon/CD chr5  000084469     000086113     AT5G01210     AT5G01210     +        +0817  NA     NA     NA
Exon/CD chr5  000088935     000090146     AT5G01220     AT5G01220     -        +0446  NA     NA     NA
Intron  chr5  000094984     000095846     AT5G01230     AT5G01230     -        +0271  NA     NA     NA
Upstream      chr5    000097819     000098976     AT5G01240     AT5G01240    +      -0136  NA     NA     NA
Exon/CD chr5  000102846     000103512     AT5G01250     AT5G01250     -        +0591  NA     NA     NA
Intron  chr5  000105176     000106178     AT5G01260     AT5G01260     +        +0352  NA     NA     NA
Exon/CD chr5  000111347     000112473     AT5G01270     AT5G01270     -        +0401  NA     NA     NA
Exon/CD chr5  000114726     000115203     AT5G01280     AT5G01280     -        +1273  NA     NA     NA
Intron  chr5  000117124     000118175     AT5G01290     AT5G01290     +        +0342  NA     NA     NA
Intron  chr5  000126082     000127051     AT5G01310     AT5G01310     +        +1262  NA     NA     NA
Exon/CD chr5  000130764     000131713     AT5G01320     AT5G01320     -        +0387  NA     NA     NA
Exon/CD chr5  000140959     000141271     AT5G01335     AT5G01335     +        +5284  NA     NA     NA
Downstream    chr5    000142267     000142690     AT5G01340     AT5G01340    -      +2242  NA     NA     NA
Exon/CD chr5  000143974     000144745     AT5G01340     AT5G01340     -        +0361  NA     NA     NA
```

Features

Distance

Galaxy



Galaxy

Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy

Use project's free server or other public servers

Get Galaxy

Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy

Screencasts, Galaxy 101, ...

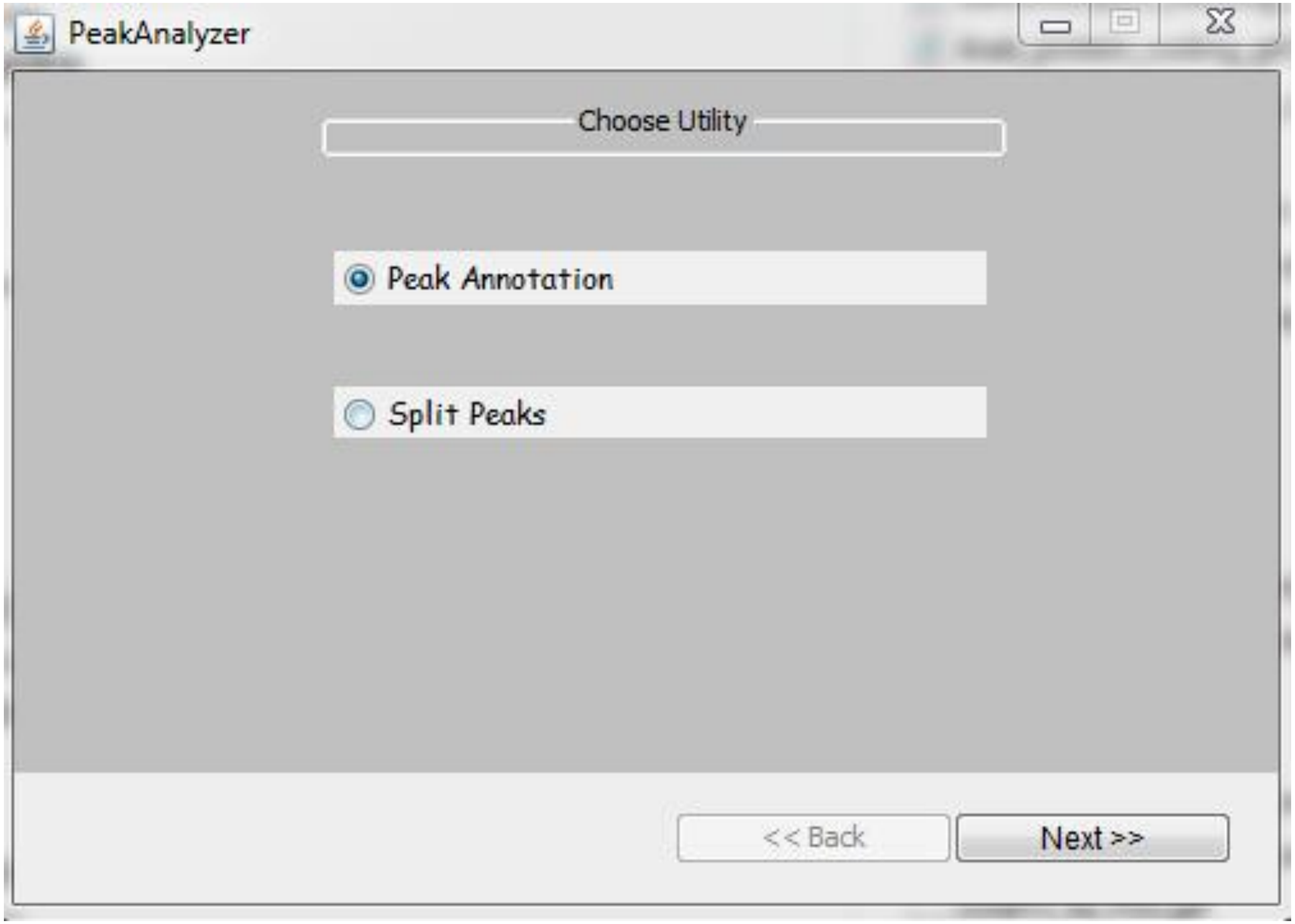
Get Involved

Mailing lists, Tool Shed, wiki

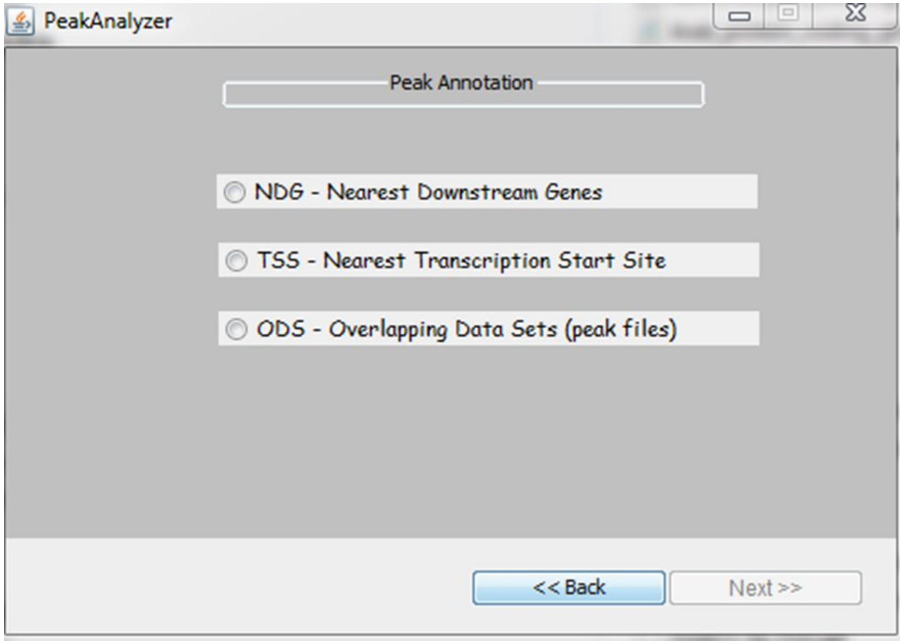
Search all resources

https://galaxyproject.org/

PeakAnalyzer



PeakAnalyzer



PeakAnalyzer

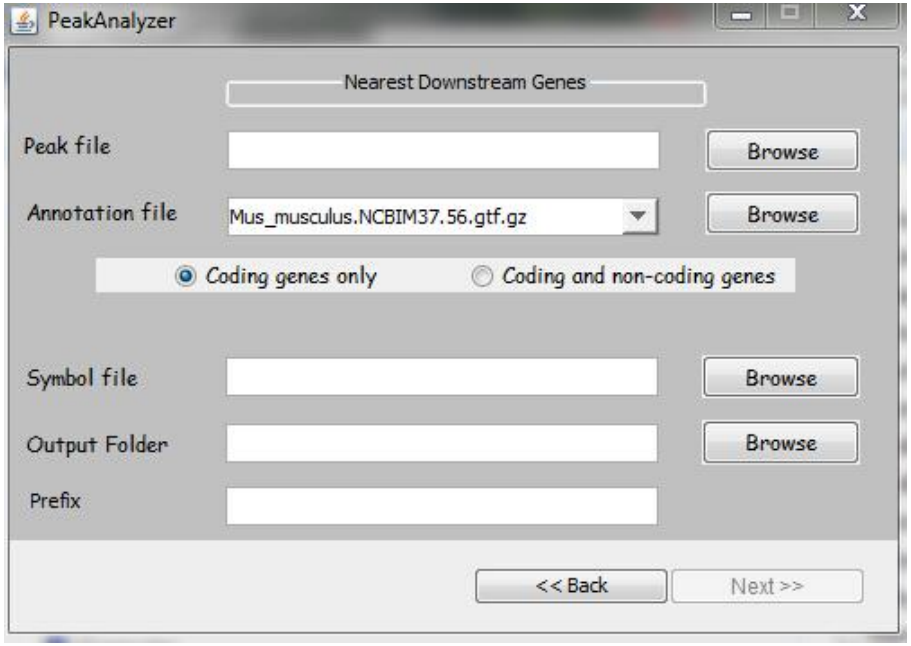
Peak Annotation

○ NDG - Nearest Downstream Genes

○ TSS - Nearest Transcription Start Site

○ ODS - Overlapping Data Sets (peak files)

<< Back   Next >>



PeakAnalyzer

Nearest Downstream Genes

Peak file:

Browse

Annotation file: Mus_musculus.NCBIM37.56.gtf.gz

Browse

● Coding genes only   ○ Coding and non-coding genes

Symbol file:

Browse

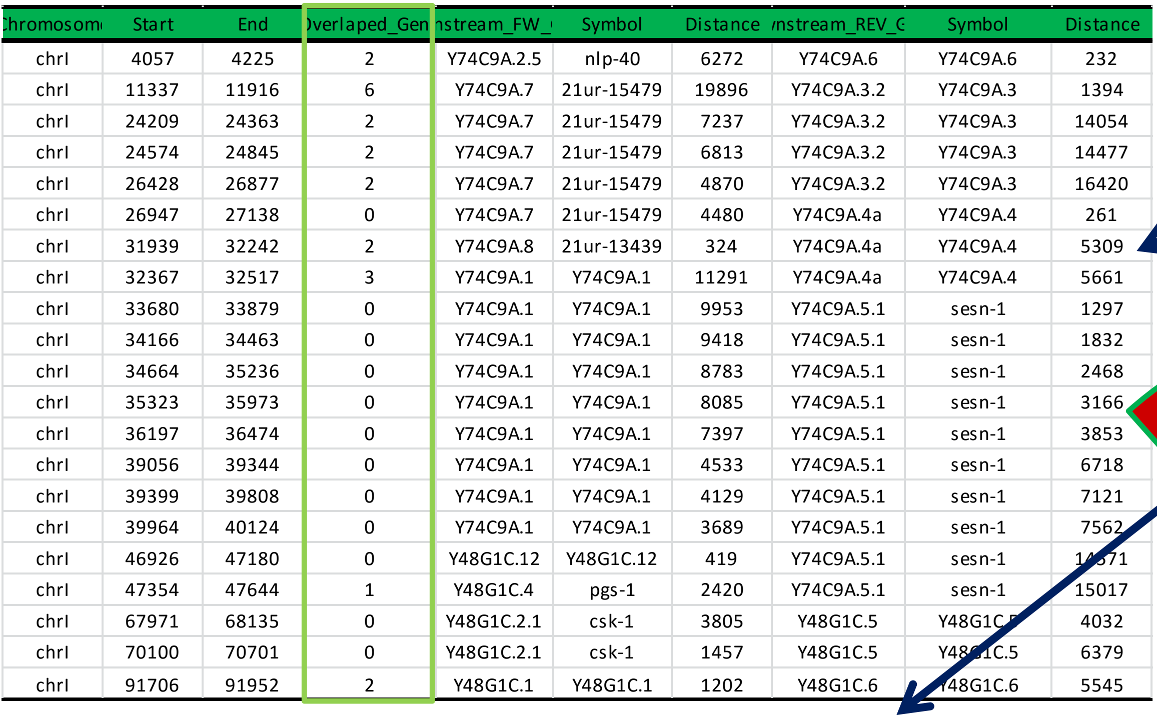
Output Folder:

Browse

Prefix:

<< Back   Next >>

Chromosomes nominate

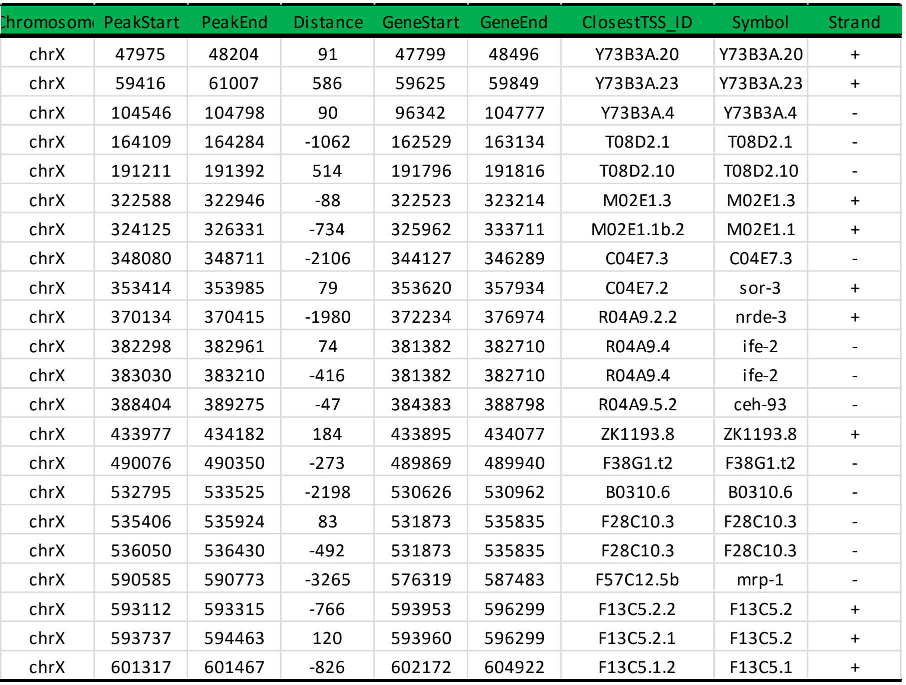


hromosom	Start	End	Overlaped_Gen	nstream_FW_(	Symbol	Distance	nstream_REV_G	Symbol	Distance
chrI	4057	4225	2	Y74C9A.2.5	nlp-40	6272	Y74C9A.6	Y74C9A.6	232
chrI	11337	11916	6	Y74C9A.7	21ur-15479	19896	Y74C9A.3.2	Y74C9A.3	1394
chrI	24209	24363	2	Y74C9A.7	21ur-15479	7237	Y74C9A.3.2	Y74C9A.3	14054
chrI	24574	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

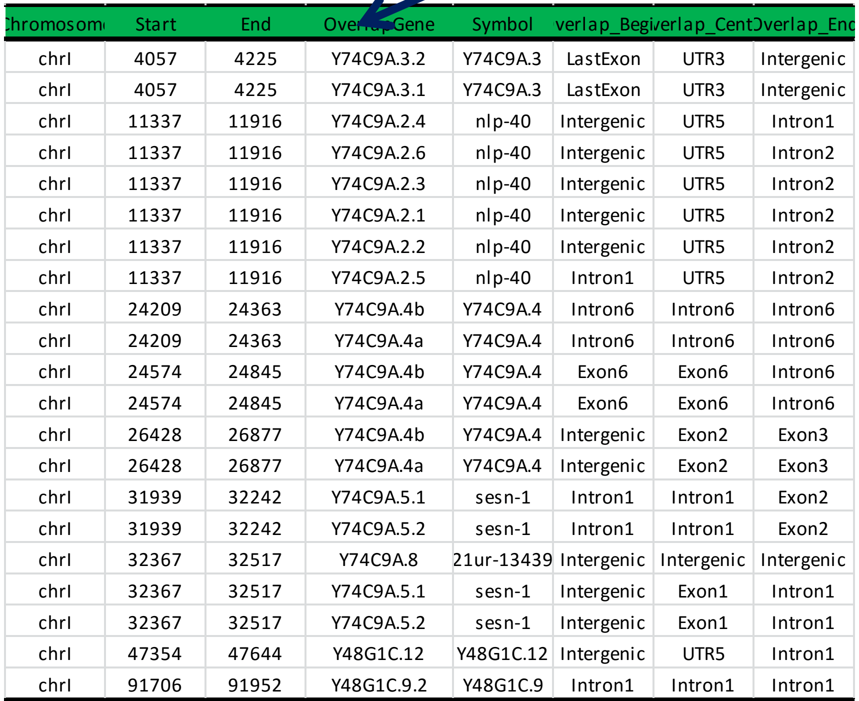
Nearest downstream genes

Nearest TSS

erlapped gene features



hromosom	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	+



hromosom	Start	End	Overlap_Gene	Symbol	verlap_Begi	verlap_Cent	Overlap_Enc
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

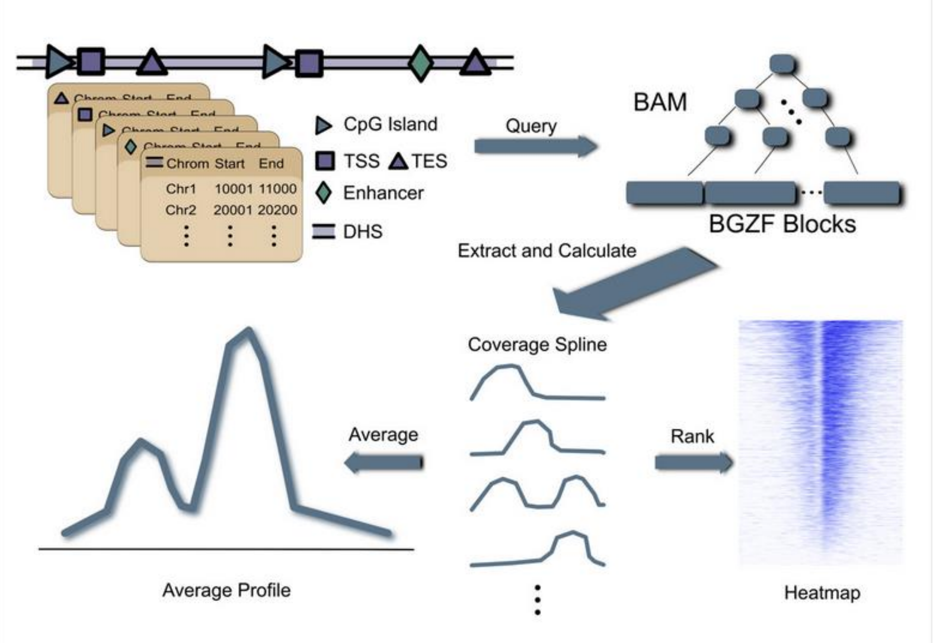
config file normalized by control

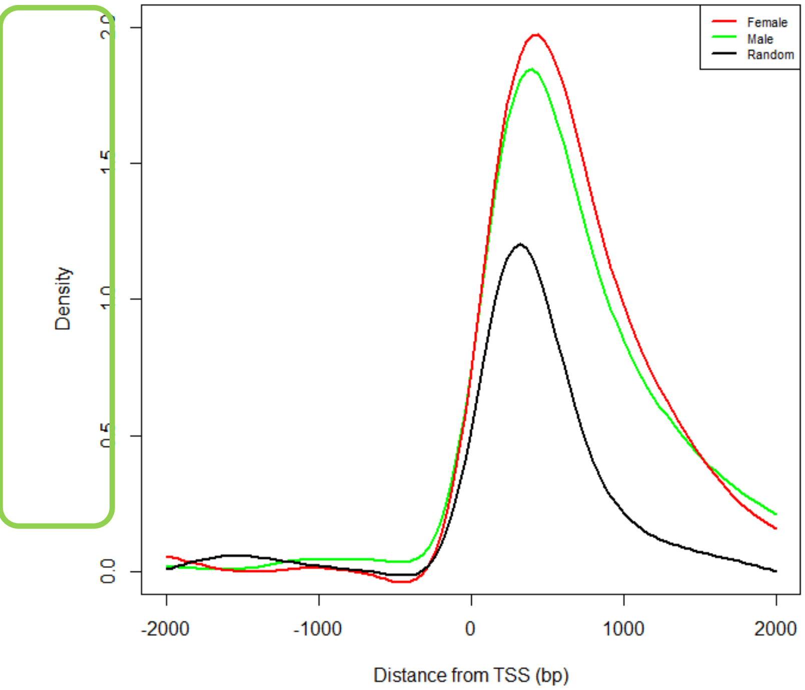
```
M_H3K4_sorted.bam:I_H3K4_sorted.bam   Male_TSS_nearest_transcripts.txt       male_H3K4_Vs_Input
M_H3K4_sorted.bam:I_H3K4_sorted.bam   Female_TSS_nearest_transcripts.txt     female_H3K4_Vs_Input
```

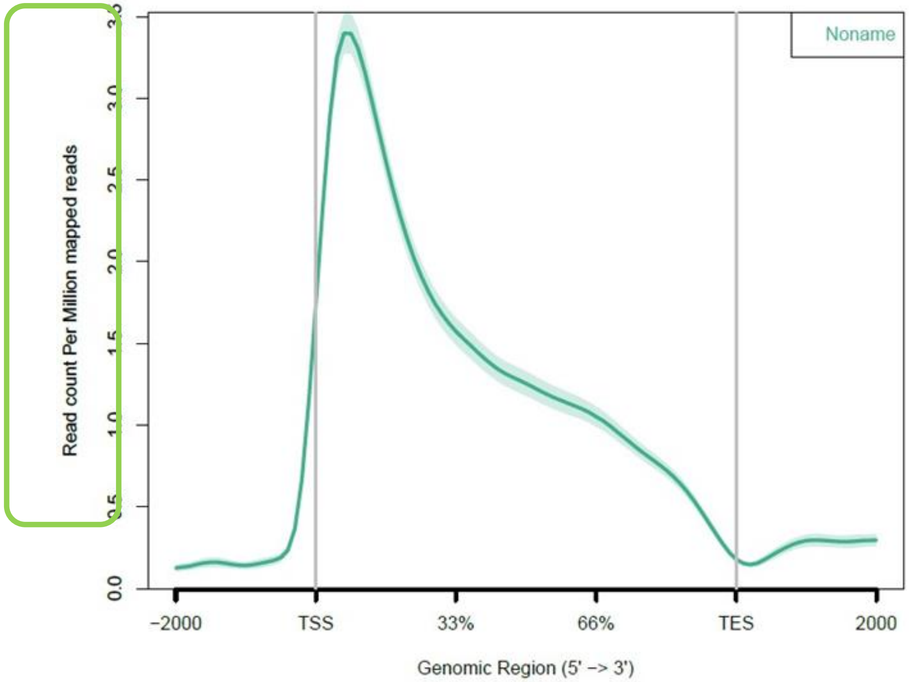
config file with only treatment bams

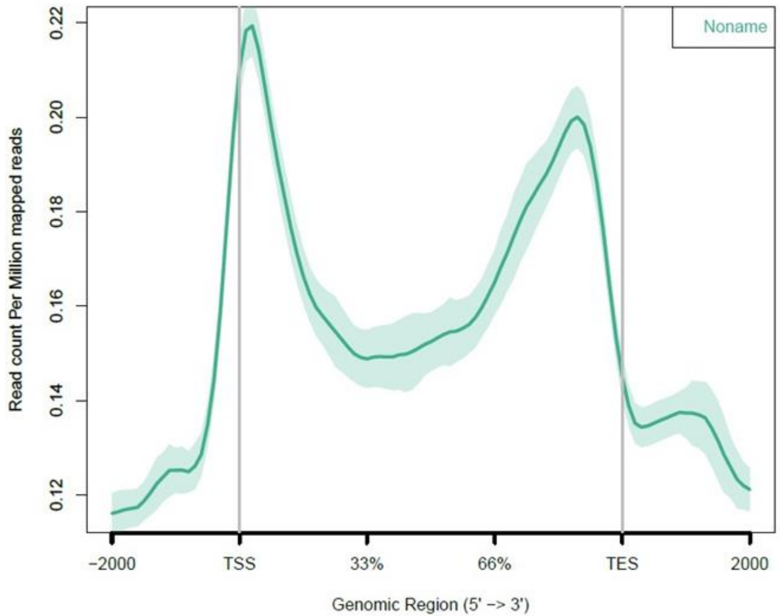
```
M_H3K4_sorted.bam   Male_TSS_nearest_transcripts.txt       male_H3K4_Vs_Input
M_H3K4_sorted.bam   Female_TSS_nearest_transcripts.txt     female_H3K4_Vs_Input
```

```
ngs.plot.r -G genome -R region -C [cov|config]file
           -O name [Options]
-G   Genome name. Use ngsplotdb.py list to show
available genomes.
 -R   Genomic regions to plot: tss, tes, genebody,
exon, cgi, enhancer, dhs or bed
 -C   Indexed bam file or a configuration file for
multiplot
 -O   Name for output: multiple files will be generated
```



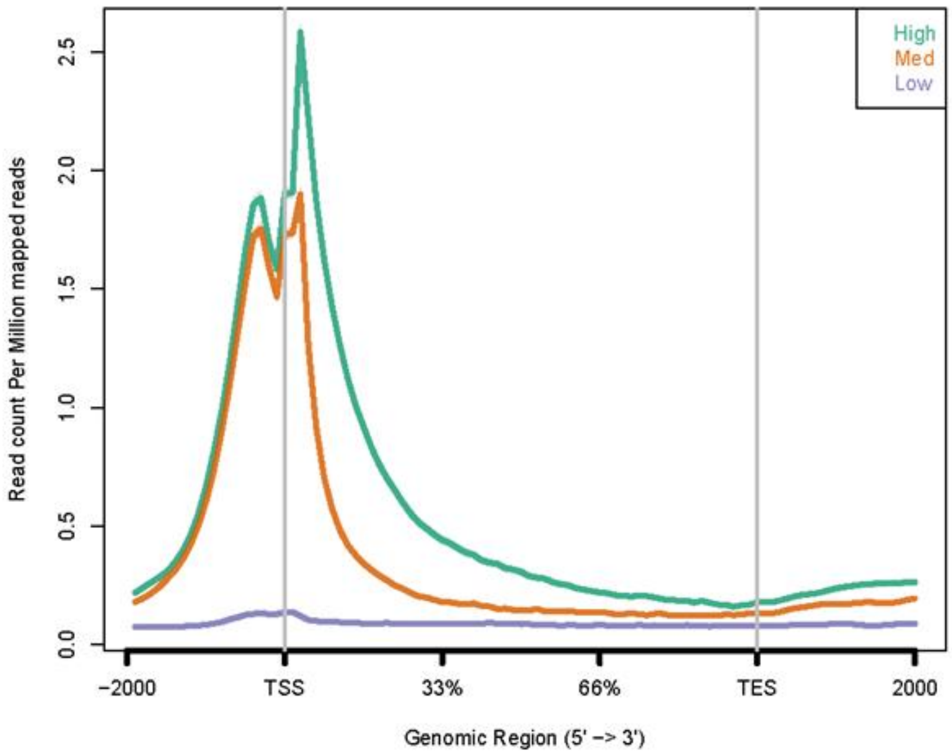


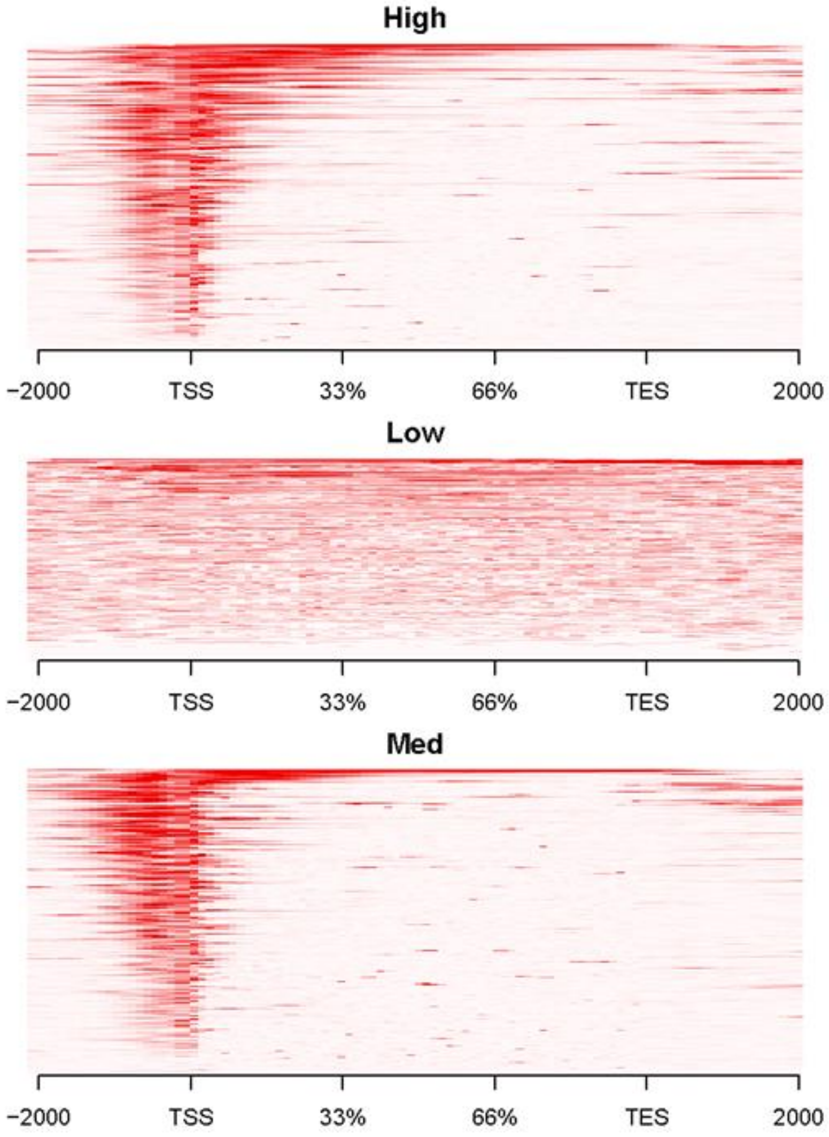




```
ngs.plot.r -G hg19 -R tss -C treatment.bam -O \
output_name -T H3K4me3 -L 3000
```

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





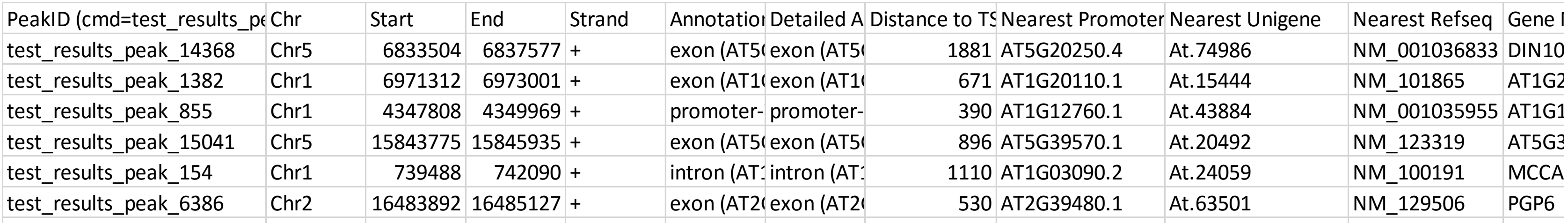
https://github.com/shenlab-sinai/ngsplot

HOMER

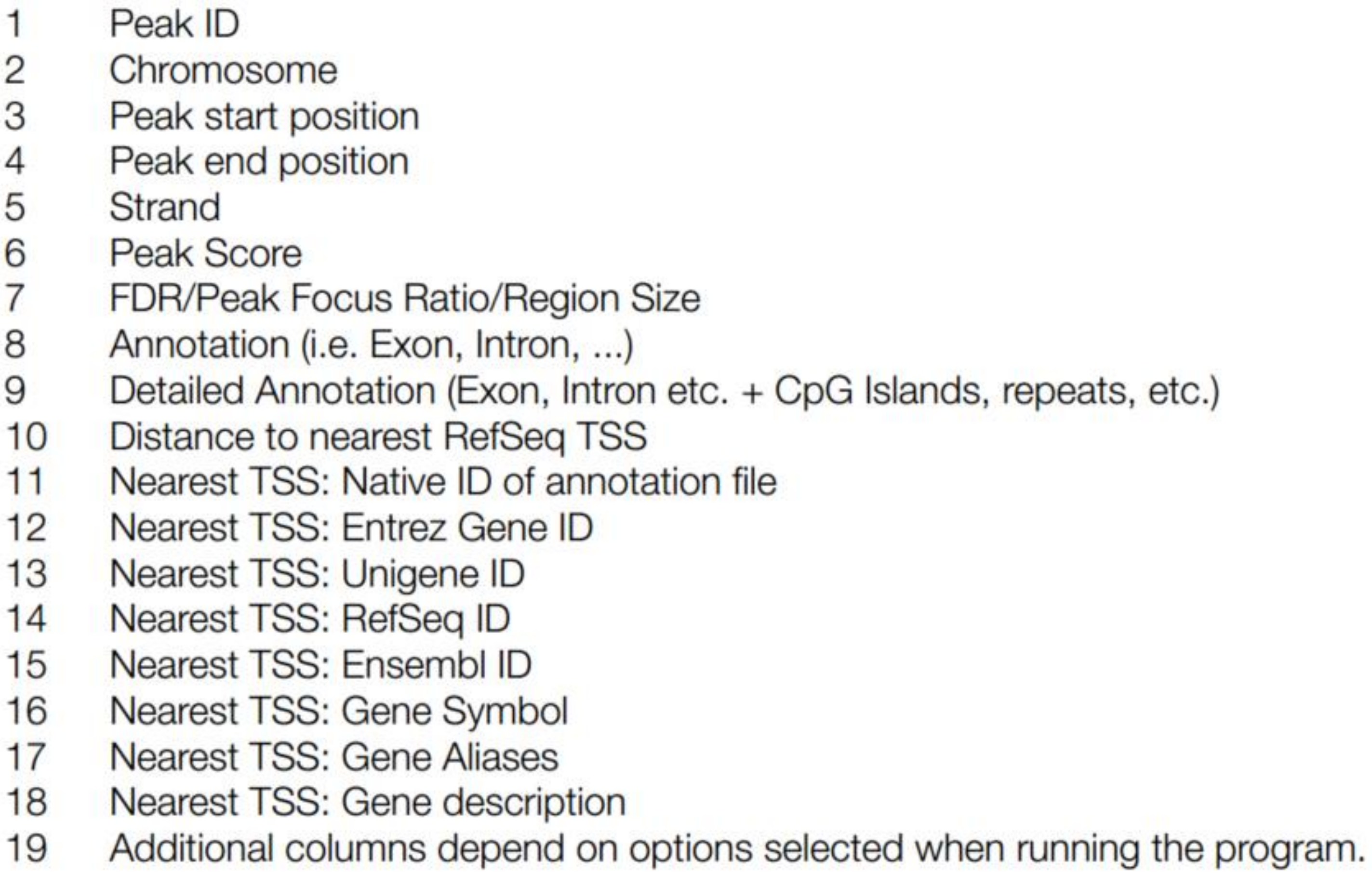
(Hypergeometric Optimization of Motif EnRichment)

- Mapping to the genome (NOT performed by HOMER, but important to understand)
- Creation Tag directories, quality control, and normalization. (**makeTagDirectory**)
- UCSC visualization (**makeUCSCfile**, **makeBigWig.pl**)
- Peak finding / Transcript detection / Feature identification (**findPeaks**)
- Motif analysis (**findMotifsGenome.pl**)
- Annotation of Peaks (**annotatePeaks.pl**)
- Quantification of Transcripts (**analyzeRNA.pl**)

- Additional analysis strategies:
- General sequence manipulation tools (**homerTools**)
- Miscellaneous Tools for Sharing Data between programs, etc. (**tagDir2bed.pl, bed2pos.pl, pos2bed.pl** ...)
- Finding overlapping or differentially bound peaks (**mergePeaks**, **getDifferentialPeaks**)
- ChIP-Seq analysis automation (**analyzeChIP-Seq.pl**)
- Description of file formats



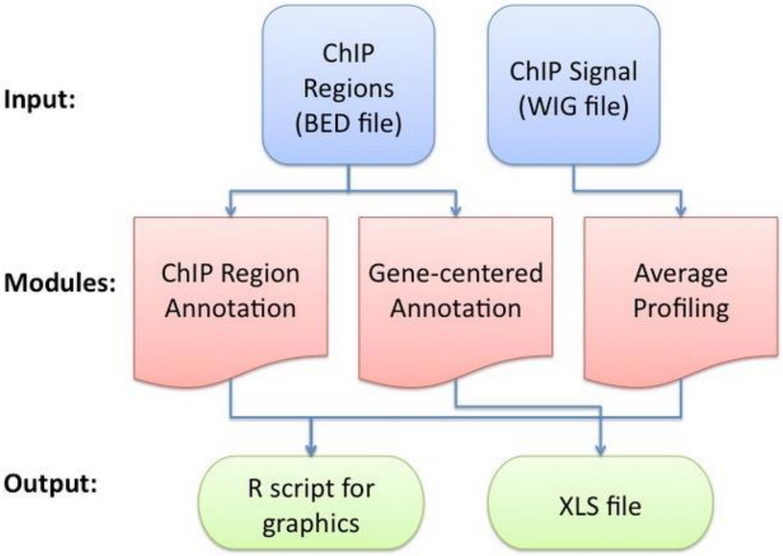
PeakID (cmd=test_results_pe	Chr	Start	End	Strand	Annotatio	Detailed A	Distance to TS	Nearest Promoter	Nearest Unigene	Nearest Refseq	Gene I
test_results_peak_14368	Chr5	6833504	6837577	+	exon (AT5	exon (AT5	1881	AT5G20250.4	At.74986	NM_001036833	DIN10
test_results_peak_1382	Chr1	6971312	6973001	+	exon (AT1	exon (AT1	671	AT1G20110.1	At.15444	NM_101865	AT1G2
test_results_peak_855	Chr1	4347808	4349969	+	promoter-	promoter-	390	AT1G12760.1	At.43884	NM_001035955	AT1G1
test_results_peak_15041	Chr5	15843775	15845935	+	exon (AT5	exon (AT5	896	AT5G39570.1	At.20492	NM_123319	AT5G3
test_results_peak_154	Chr1	739488	742090	+	intron (AT1	intron (AT1	1110	AT1G03090.2	At.24059	NM_100191	MCCA
test_results_peak_6386	Chr2	16483892	16485127	+	exon (AT2	exon (AT2	530	AT2G39480.1	At.63501	NM_129506	PGP6



1 Peak ID
2 Chromosome
3 Peak start position
4 Peak end position
5 Strand
6 Peak Score
7 FDR/Peak Focus Ratio/Region Size
8 Annotation (i.e. Exon, Intron, ...)
9 Detailed Annotation (Exon, Intron etc. + CpG Islands, repeats, etc.)
10 Distance to nearest RefSeq TSS
11 Nearest TSS: Native ID of annotation file
12 Nearest TSS: Entrez Gene ID
13 Nearest TSS: Unigene ID
14 Nearest TSS: RefSeq ID
15 Nearest TSS: Ensembl ID
16 Nearest TSS: Gene Symbol
17 Nearest TSS: Gene Aliases
18 Nearest TSS: Gene description
19 Additional columns depend on options selected when running the program.

```
annotatePeaks.pl test_results_peaks.narrowPeak_chr  tair10  >out
```

CEAS
(Cis-regulatory Element Annotation System)



Input:

ChIP Regions (BED file)

ChIP Signal (WIG file)

Modules:

ChIP Region Annotation

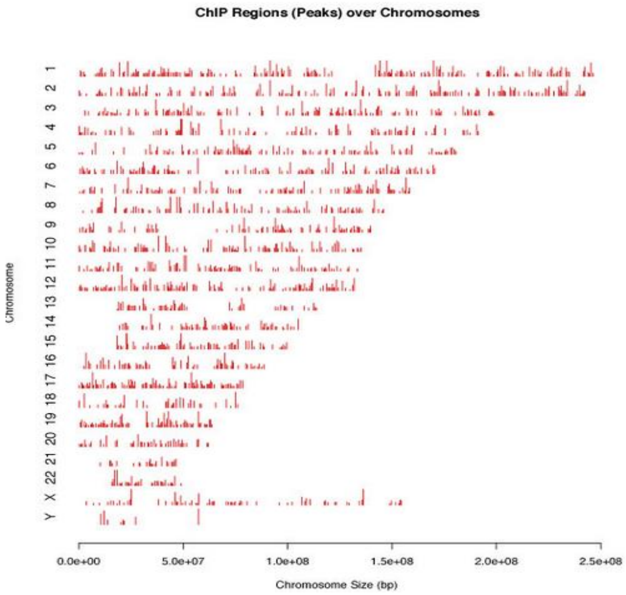
Gene-centered Annotation

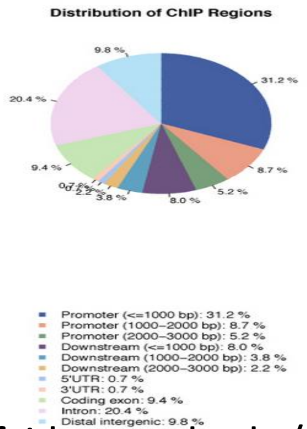
Average Profiling

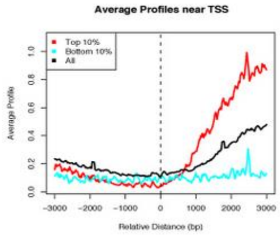
Output:

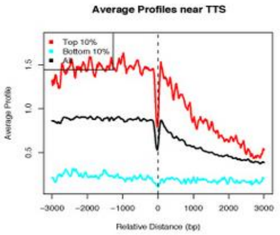
R script for graphics

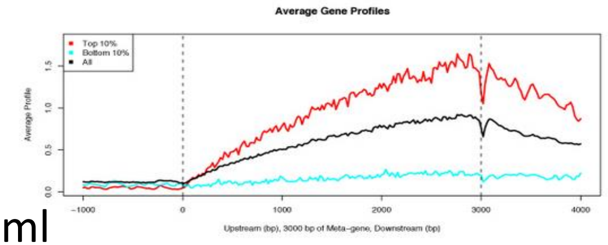
XLS file





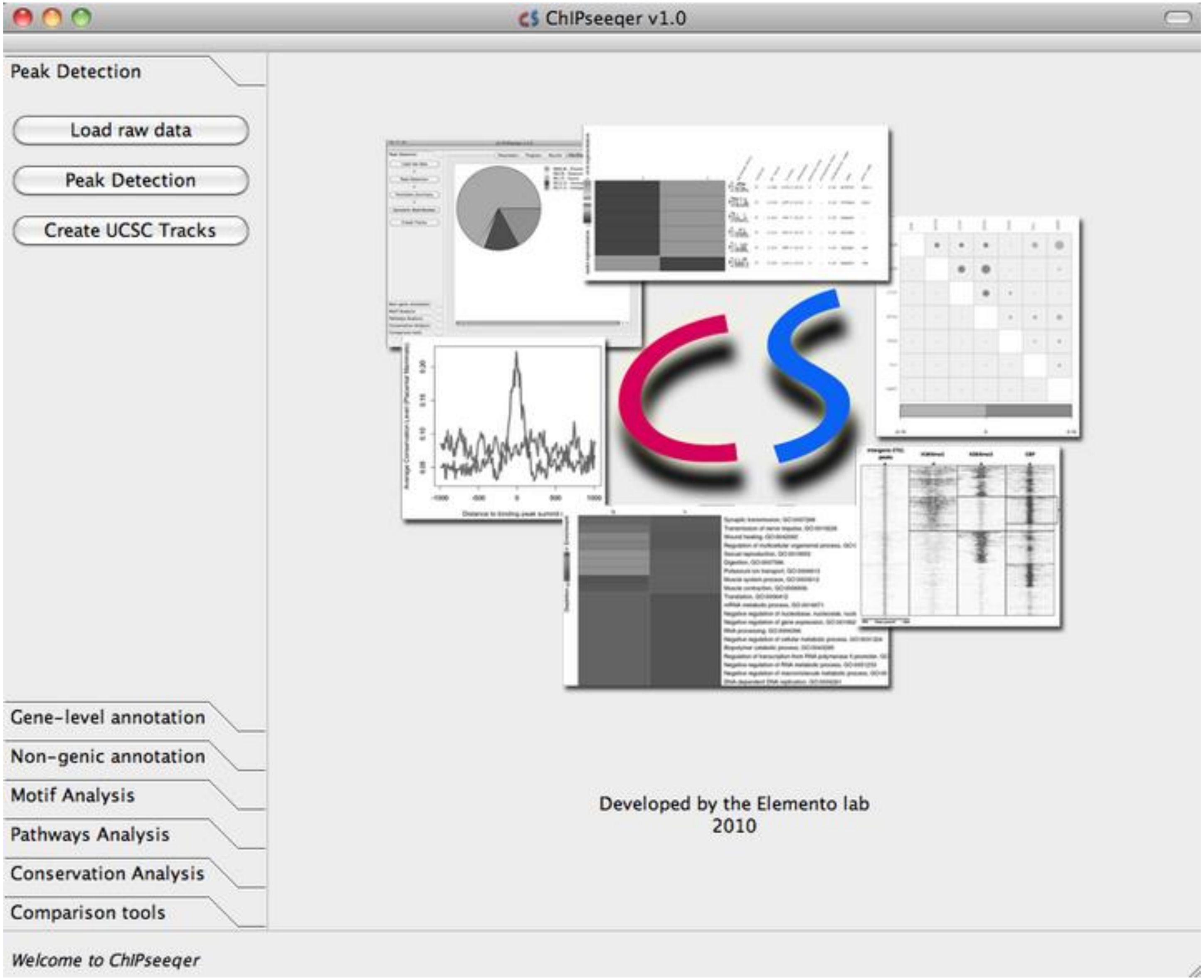






http://liulab.dfci.harvard.edu/CEAS/usermanual.html

ChIPseeqer



Developed by the Elemento lab
2010

Welcome to ChIPseeqer

Functional enrichment

- Over-represented functional annotations of nearest genes of peaks
  - Gene Ontology
  - Biological Pathways

- Typical tools

DAVID	https://david.ncifcrf.gov/
GREAT	http://bejerano.stanford.edu/great/public/html/
Blast2go	https://www.blast2go.com/

Blast2GO

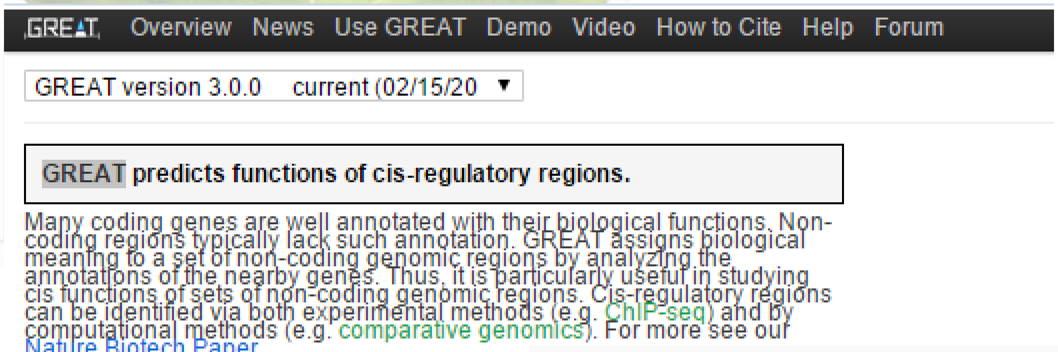
HIGH-QUALITY FUNCTIONAL ANNOTATION

UP AND RUNNING WITHIN NO TIME

Read more

Request a Free PRO Trial

Experience all advantages of a PRO account for one week



GREAT Overview News Use GREAT Demo Video How to Cite Help Forum

GREAT version 3.0.0 current (02/15/20 ▼

GREAT predicts functions of cis-regulatory regions.

Many coding genes are well annotated with their biological functions. Non-coding regions typically lack such annotation. GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying cis functions of sets of non-coding genomic regions. Cis-regulatory regions can be identified via both experimental methods (e.g. ChIP-seq) and by computational methods (e.g. comparative genomics). For more see our Nature Biotech Paper.

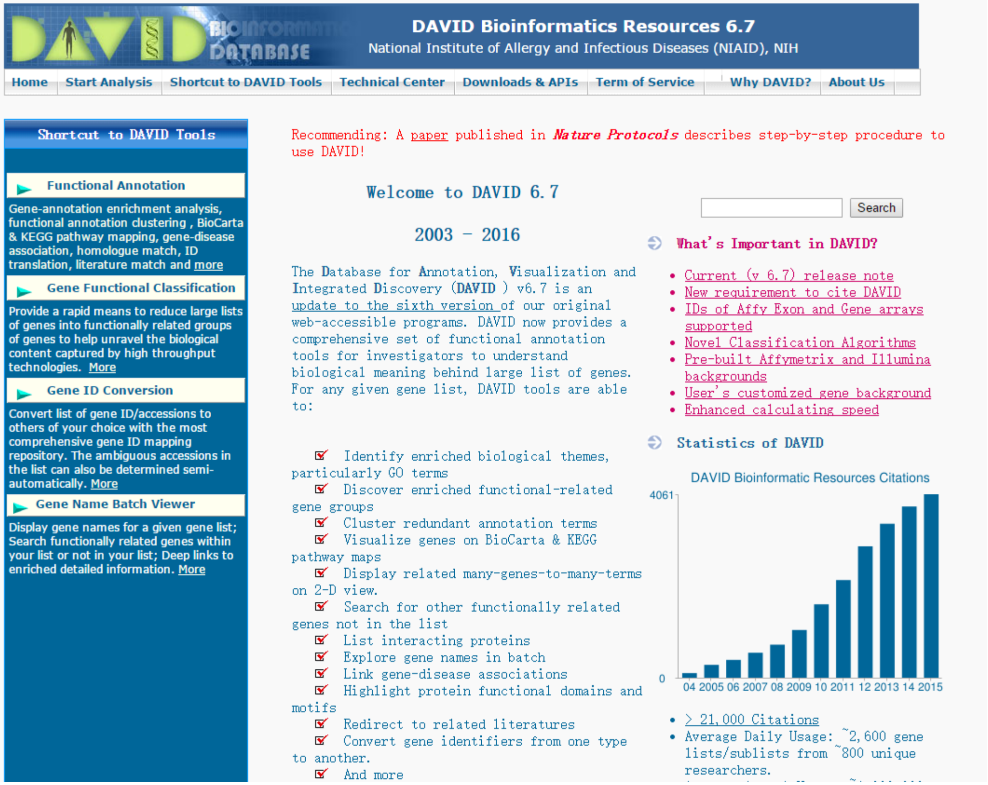
News

- Feb 15, 2015: GREAT version 3.0 switche the mouse mm10 assembly, and adds new c
- Apr 3, 2012: GREAT version 2.0 adds new a mouse ontologies and visualization tools for
- Feb 18, 2012: The GREAT forums are releas to-user interaction

More news items...

Species Assembly
- Human: GRCh37 (UCSC hg19
- Mouse: NCBI build 37 (UCSC r
- Mouse: NCBI build 38 (UCSC r
- Zebrafish: Wellcome Trust Zv9 Jul/2010) Zebrafish CNE set

Can I use a different species or a



DAVID BIOINFORMATICS DATABASE

DAVID Bioinformatics Resources 6.7

National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home | Start Analysis | Shortcut to DAVID Tools | Technical Center | Downloads & APIs | Term of Service | Why DAVID? | About Us

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification

Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion

Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer

Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Recommending: A paper published in ***Nature Protocols*** describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7

2003 - 2016

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 is an update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

☑ Identify enriched biological themes, particularly GO terms
☑ Discover enriched functional-related gene groups
☑ Cluster redundant annotation terms
☑ Visualize genes on BioCarta & KEGG pathway maps
☑ Display related many-genes-to-many-terms on 2-D view.
☑ Search for other functionally related genes not in the list
☑ List interacting proteins
☑ Explore gene names in batch
☑ Link gene-disease associations
☑ Highlight protein functional domains and motifs
☑ Redirect to related literatures
☑ Convert gene identifiers from one type to another.
☑ And more

Search

What's Important in DAVID?

- Current (v 6.7) release note
- New requirement to cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID

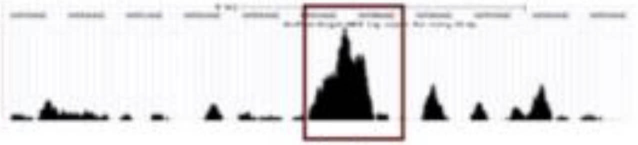
DAVID Bioinformatic Resources Citations

- > 21,000 Citations
- Average Daily Usage: ~2,600 gene lists/sublists from ~800 unique researchers.

Homer

geneOntology.html

GenomeOntology.html



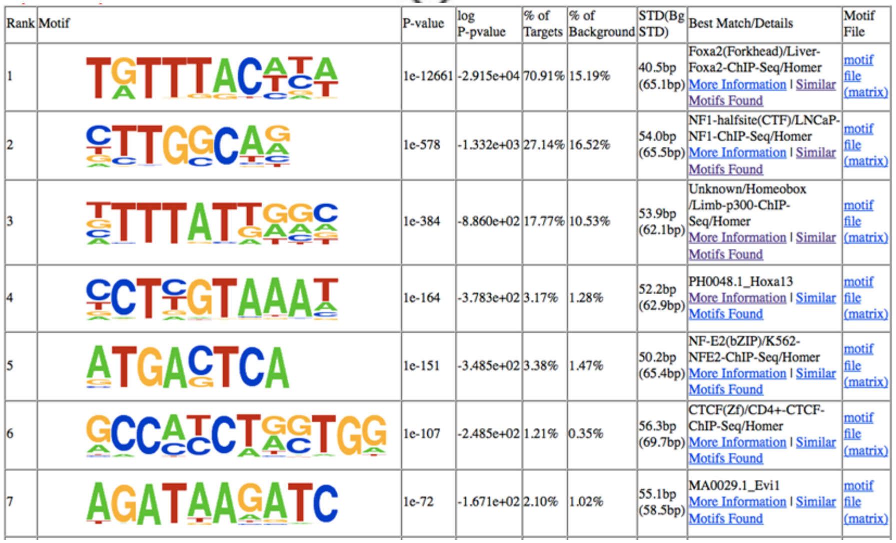
ChIP-seq peaks

```
>mm9_chr1_39249116_39251316_+
gagaggaaggggggagaaagagggagggggagGGTGATAGGTAGCCAGGAG
CCAATGGGGGCGTTTTTCCTTGTCCAGGCCACTTGCTGGAATGTGAGATGT
AGAATGACCCAAAGAGAGCTGCCAAGACAGAGCTCTGCCCCAGGAATTGA
ACTCAAAGGGTGTCAGAAAGCAGGTGGCCTTTTGTGCACCTGGCGCGGGGA
CGTGGCTCCCCTCTTCCGGCTGGTCTAGCCAGGtgcctgcctgcctgcct
gccGTGATCTCTGGACGCCAGTAGAGGGTTGTTGTGGGTTTGGGTGAAAC
ACGCCACCCCTGAGCTCTTCCGCGGGGCTAGCAATCTCCCCATCACCCCA
TTCGCGCTCAGAACCCCCTCAGCGAGTCTAACAGCAGGCCTGGTTCCCCG
```

DNA sequence

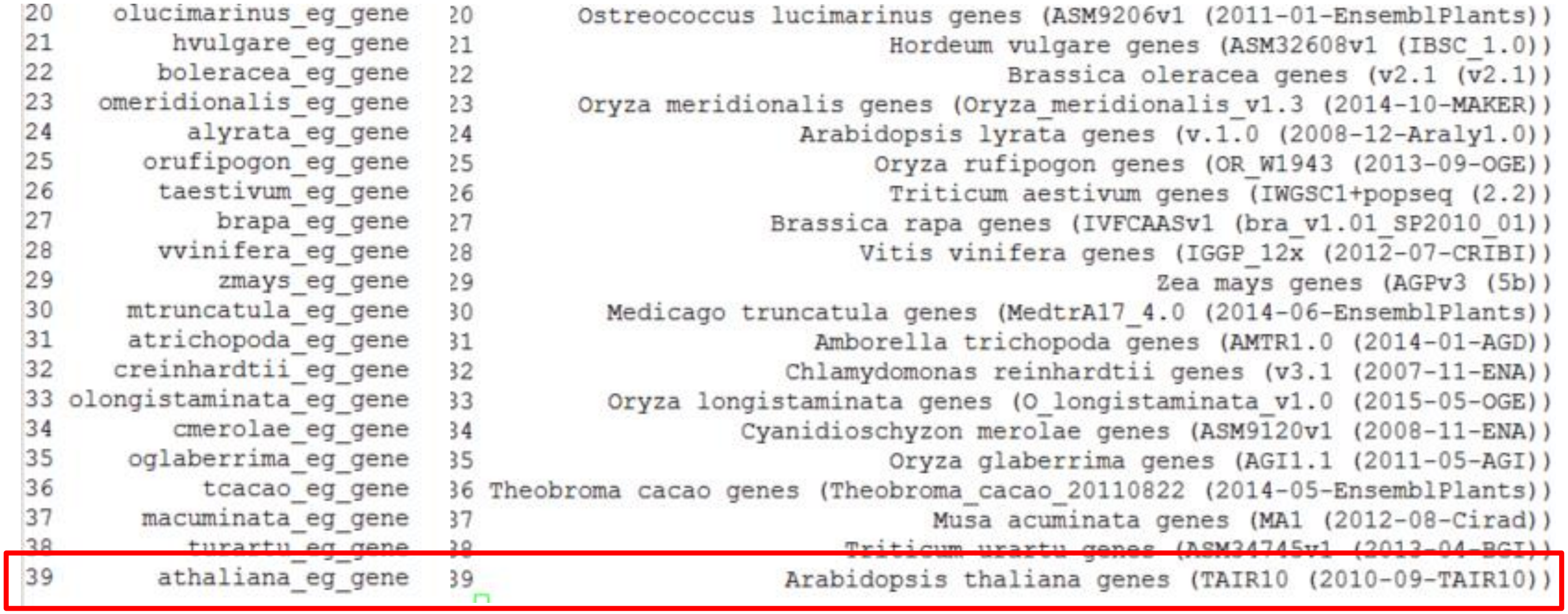
```
A  [24 54 59  0 65 71  4 24  9 ]
C  [ 7  6  4 72  4  2  0  6  9 ]
G  [31  7  0  2  0  1  1 38 55 ]
T  [14  9 13  2  7  2 71  8  3 ]
```

Discovered motif



Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1	TGTTTACATA	1e-12661	-2.915e+04	70.91%	15.19%	40.5bp (65.1bp)	Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
2	CTTGGCAG	1e-578	-1.332e+03	27.14%	16.52%	54.0bp (65.5bp)	NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
3	TTTTATTGGC	1e-384	-8.860e+02	17.77%	10.53%	53.9bp (62.1bp)	Unknown/Homeobox/Limb-p300-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
4	CCTCGTAAAT	1e-164	-3.783e+02	3.17%	1.28%	52.2bp (62.9bp)	PH0048.1_Hoxa13 More Information | Similar Motifs Found	motif file (matrix)
5	ATGACTCA	1e-151	-3.485e+02	3.38%	1.47%	50.2bp (65.4bp)	NF-E2(bZIP)/K562-NFE2-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
6	GCCATCTGGTGG	1e-107	-2.485e+02	1.21%	0.35%	56.3bp (69.7bp)	CTCF(Zf)/CD4+-CTCF-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
7	AGATAAGATC	1e-72	-1.671e+02	2.10%	1.02%	55.1bp (58.5bp)	MA0029.1_Evi1 More Information | Similar Motifs Found	motif file (matrix)

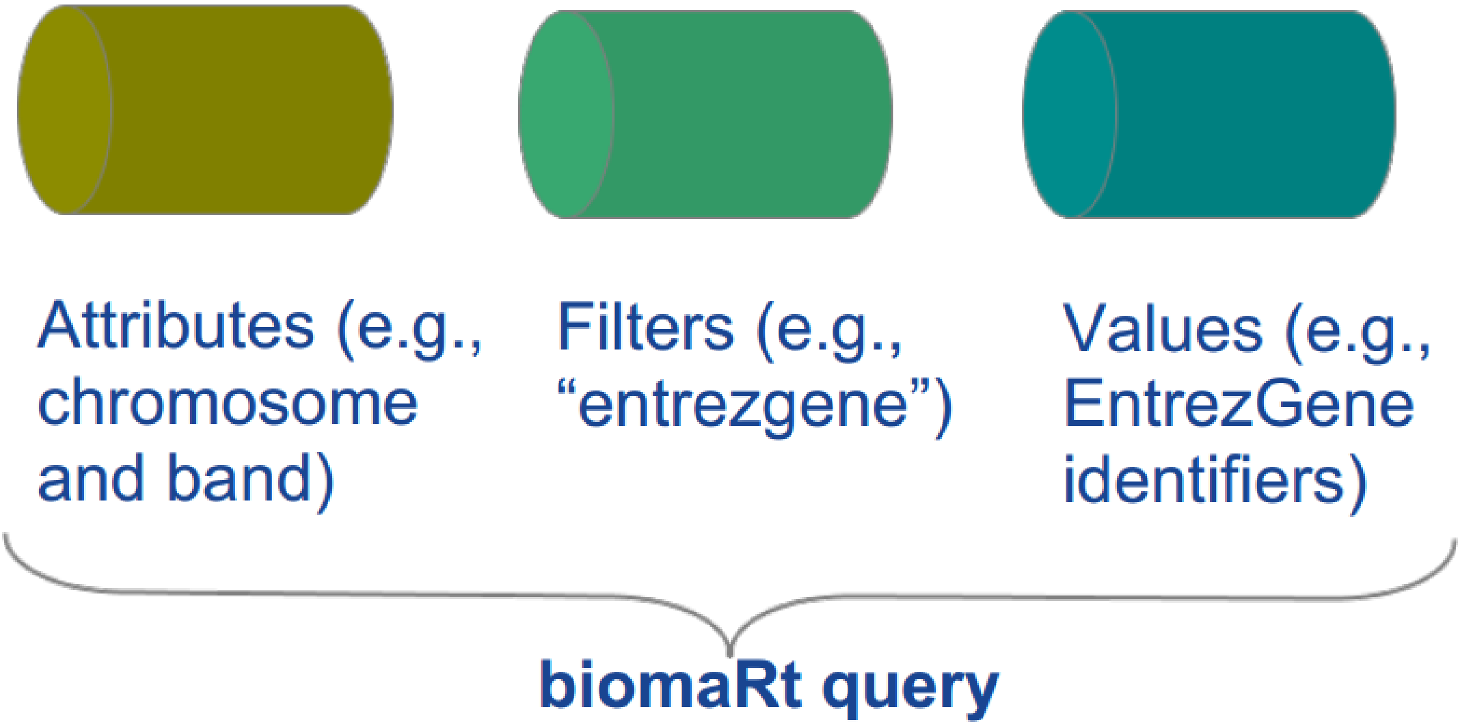
```
source("http://bioconductor.org/biocLite.R")
biocLite("biomaRt")
library (biomart)
# head(listMarts(host = "www.ensembl.org"), 10)
listMarts(host="plants.ensembl.org")
listDatasets(useMart(biomart="plants_mart",host="plants.ensembl.org"))
```



```
20      olucimarinus_eg_gene  20        Ostreococcus lucimarinus genes (ASM9206v1 (2011-01-EnsemblPlants))
21           hvulgare_eg_gene  21                           Hordeum vulgare genes (ASM32608v1 (IBSC_1.0))
22         boleracea_eg_gene  22                                  Brassica oleracea genes (v2.1 (v2.1))
23    omeridionalis_eg_gene  23      Oryza meridionalis genes (Oryza_meridionalis_v1.3 (2014-10-MAKER))
24            alyrata_eg_gene  24                      Arabidopsis lyrata genes (v.1.0 (2008-12-Araly1.0))
25        orufipogon_eg_gene  25                           Oryza rufipogon genes (OR_W1943 (2013-09-OGE))
26         taestivum_eg_gene  26                      Triticum aestivum genes (IWGSC1+popseq (2.2))
27              brapa_eg_gene  27                  Brassica rapa genes (IVFCAASv1 (bra_v1.01_SP2010_01))
28          vvinifera_eg_gene  28                        Vitis vinifera genes (IGGP_12x (2012-07-CRIBI))
29              zmays_eg_gene  29                                        Zea mays genes (AGPv3 (5b))
30       mtruncatula_eg_gene  30       Medicago truncatula genes (MedtrA17_4.0 (2014-06-EnsemblPlants))
31        atrichopoda_eg_gene  31                     Amborella trichopoda genes (AMTR1.0 (2014-01-AGD))
32       creinhardtii_eg_gene  32                    Chlamydomonas reinhardtii genes (v3.1 (2007-11-ENA))
33 olongistaminata_eg_gene  33       Oryza longistaminata genes (O_longistaminata_v1.0 (2015-05-OGE))
34           cmerolae_eg_gene  34                Cyanidioschyzon merolae genes (ASM9120v1 (2008-11-ENA))
35       oglaberrima_eg_gene  35                           Oryza glaberrima genes (AGI1.1 (2011-05-AGI))
36             tcacao_eg_gene  36 Theobroma cacao genes (Theobroma_cacao_20110822 (2014-05-EnsemblPlants))
37        macuminata_eg_gene  37                              Musa acuminata genes (MA1 (2012-08-Cirad))
38            turartu_eg_gene  38                    Triticum urartu genes (ASM34745v1 (2013-04-BGI))
39          athaliana_eg_gene  39                    Arabidopsis thaliana genes (TAIR10 (2010-09-TAIR10))
```

```
arabidopsis
=useDataset("athaliana_eg_gene",mart=useMart("plants_mart",host="plan
ts.ensembl.org"))
```

biomaRt & Bioconductor



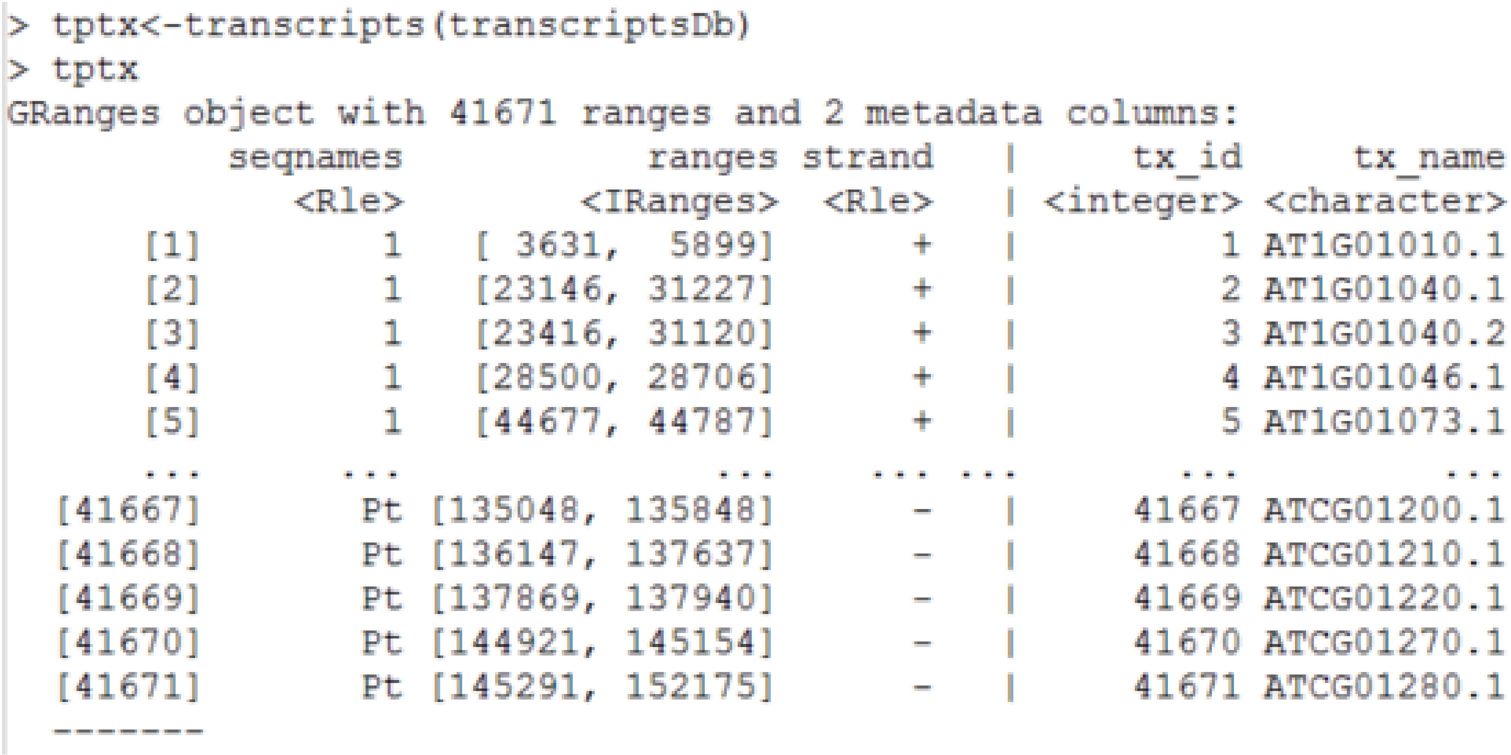
Attributes (e.g., chromosome and band)

Filters (e.g., “entrezgene”)

Values (e.g., EntrezGene identifiers)

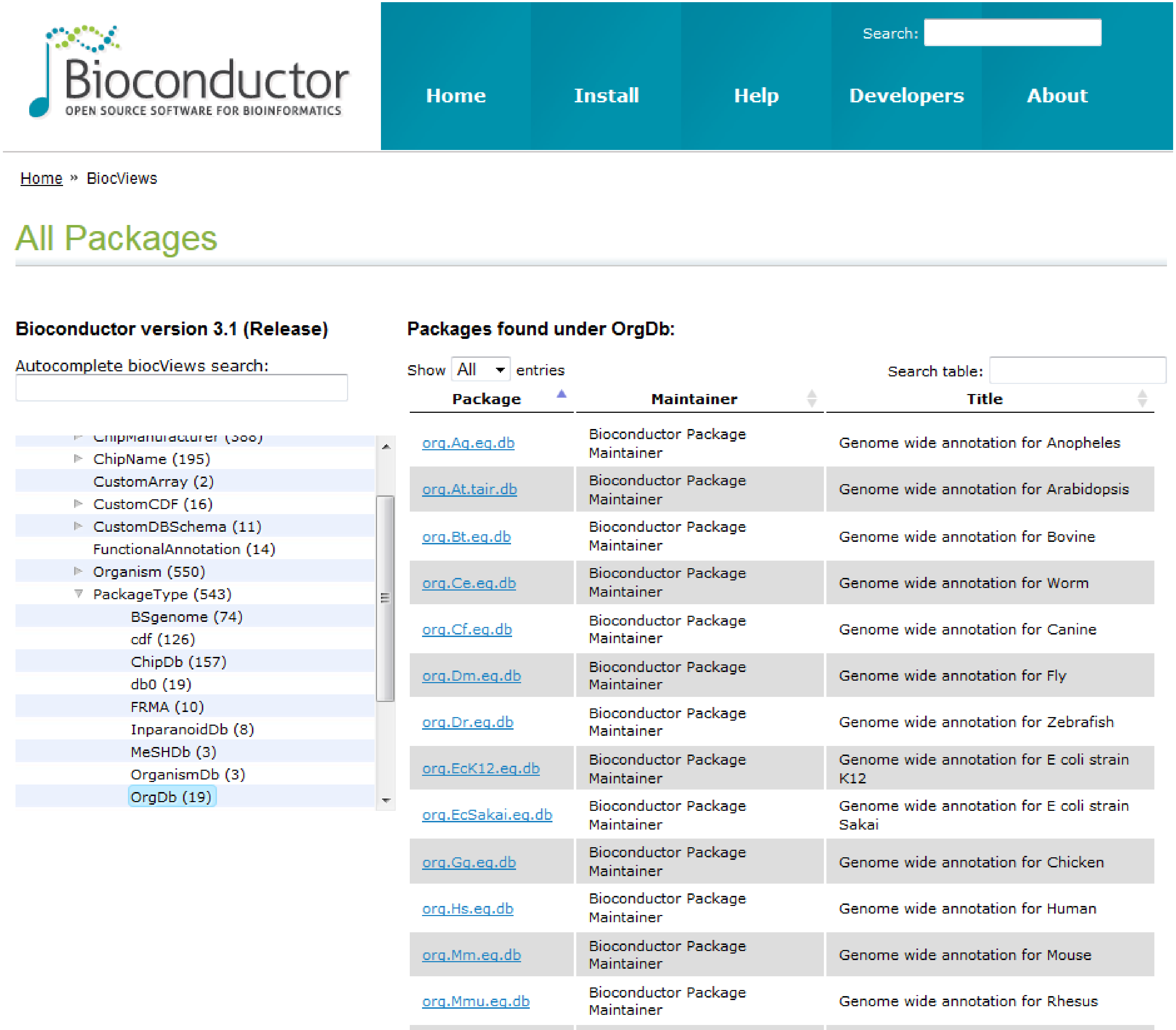
biomaRt query

```
transcriptsDb <- makeTxDbFromBiomart(biomart="plants_mart",
host="plants.ensembl.org" ,dataset="athaliana_eg_gene")
tptx<-transcripts(transcriptsDb)
```



```
> tptx<-transcripts(transcriptsDb)
> tptx
GRanges object with 41671 ranges and 2 metadata columns:
          seqnames           ranges strand   |     tx_id     tx_name
             <Rle>        <IRanges>  <Rle>   | <integer> <character>
      [1]        1   [ 3631,  5899]      +   |         1 AT1G01010.1
      [2]        1   [23146, 31227]      +   |         2 AT1G01040.1
      [3]        1   [23416, 31120]      +   |         3 AT1G01040.2
      [4]        1   [28500, 28706]      +   |         4 AT1G01046.1
      [5]        1   [44677, 44787]      +   |         5 AT1G01073.1
      ...      ...              ...    ... ...       ...         ...
  [41667]       Pt [135048, 135848]      -   |     41667 ATCG01200.1
  [41668]       Pt [136147, 137637]      -   |     41668 ATCG01210.1
  [41669]       Pt [137869, 137940]      -   |     41669 ATCG01220.1
  [41670]       Pt [144921, 145154]      -   |     41670 ATCG01270.1
  [41671]       Pt [145291, 152175]      -   |     41671 ATCG01280.1
  -------
```

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



Home » BiocViews

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:

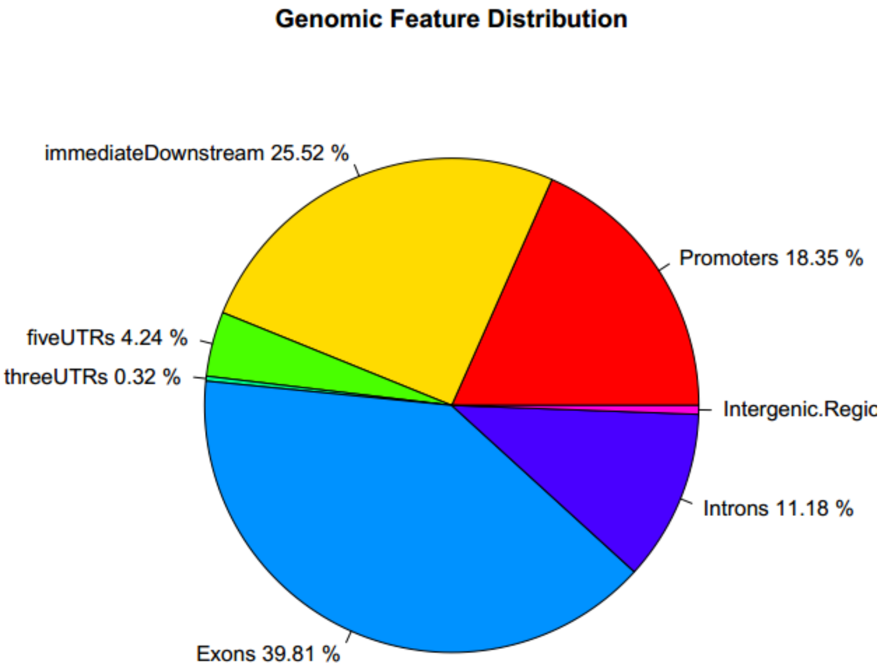
Show All entries

Search table:

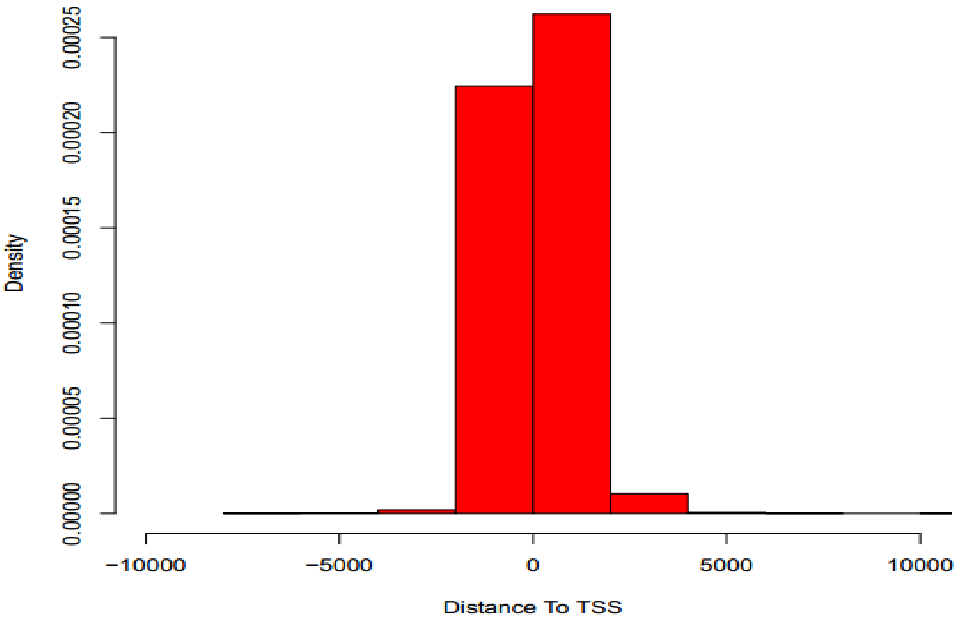
Package	Maintainer	Title
org.Ag.eg.db	Bioconductor Package Maintainer	Genome wide annotation for Anopheles
org.At.tair.db	Bioconductor Package Maintainer	Genome wide annotation for Arabidopsis
org.Bt.eg.db	Bioconductor Package Maintainer	Genome wide annotation for Bovine
org.Ce.eg.db	Bioconductor Package 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

ChIPpeakAnno

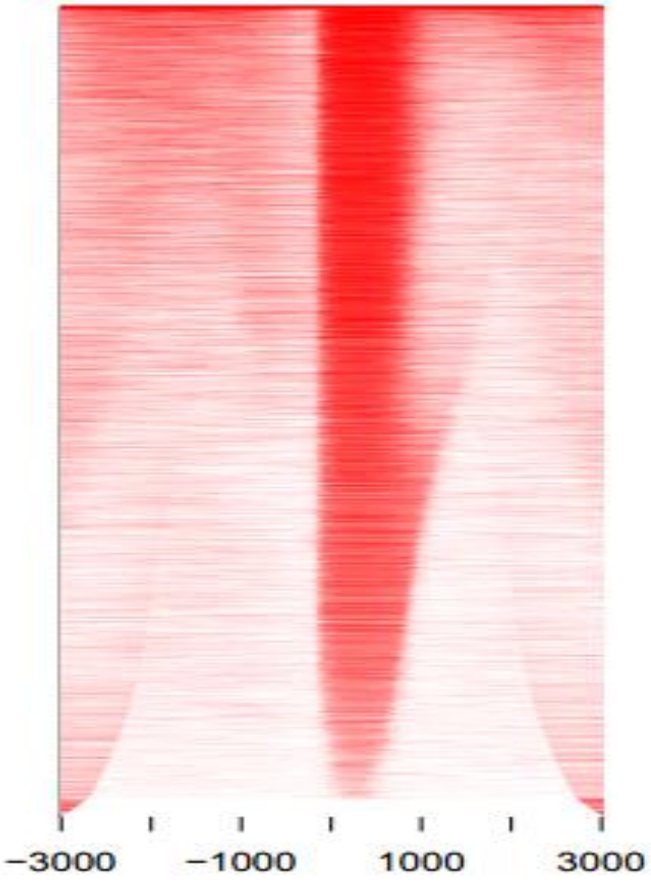
```
peak<- readPeakFile("test_results_summits.bed", as="GRanges")
aCR<-assignChromosomeRegion(peak, nucleotideLevel=FALSE,
precedence=c("Promoters", "immediateDownstream", "fiveUTRs",
"threeUTRs",   "Exons", "Introns"), TxDb=txdb)
```

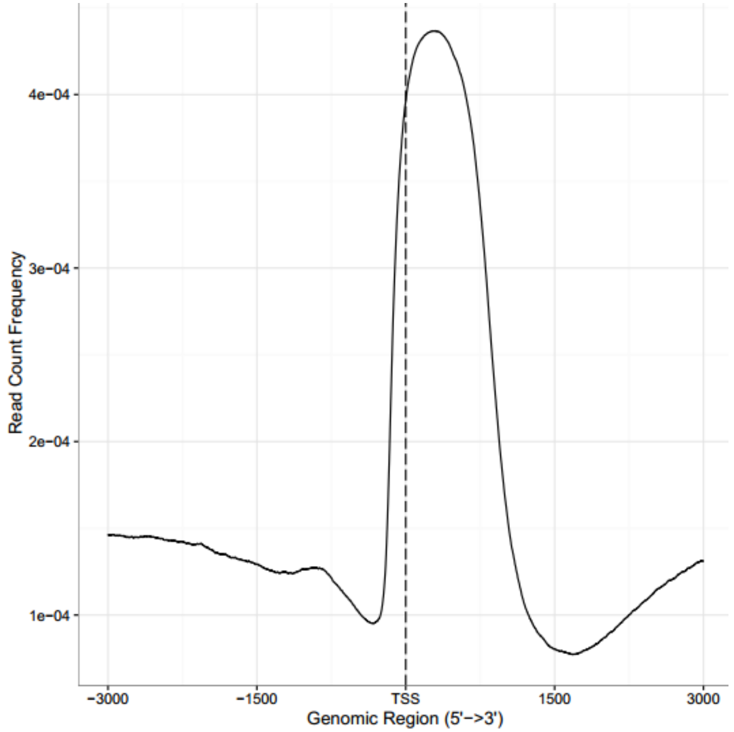


```
tx_by_gn <- transcriptsBy(txdb, by="gene")
unlisted <- unlist(tx_by_gn)
TSS <- ifelse(strand(unlisted) == "+", start(unlisted), end(unlisted))
TSS <- GRanges(seqnames(unlisted), IRanges(TSS, width=1), strand(unlisted))
………..
macs.anno <- annotatePeakInBatch(peak, AnnotationData=unlisted_TSS)
hist(macs.anno$distancetoFeature,xlab="Distance To TSS", main="",
xlim=c(-10000,10000),breaks=20,prob=T,col="red")
```



```
promoter <- getPromoters(TxDb=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")
```



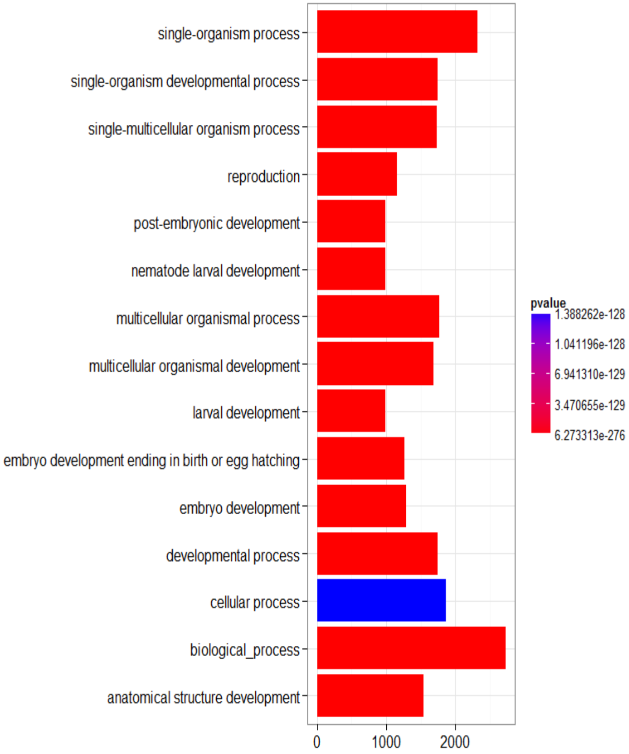


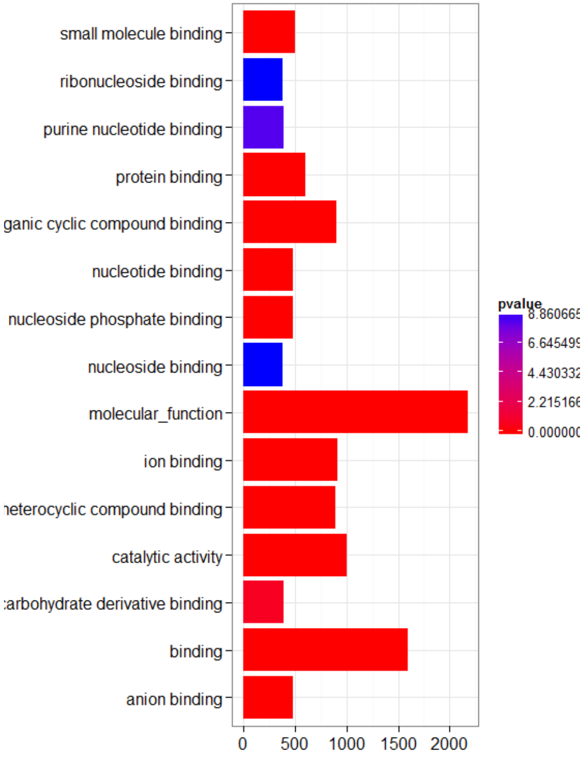
GO & Pathway

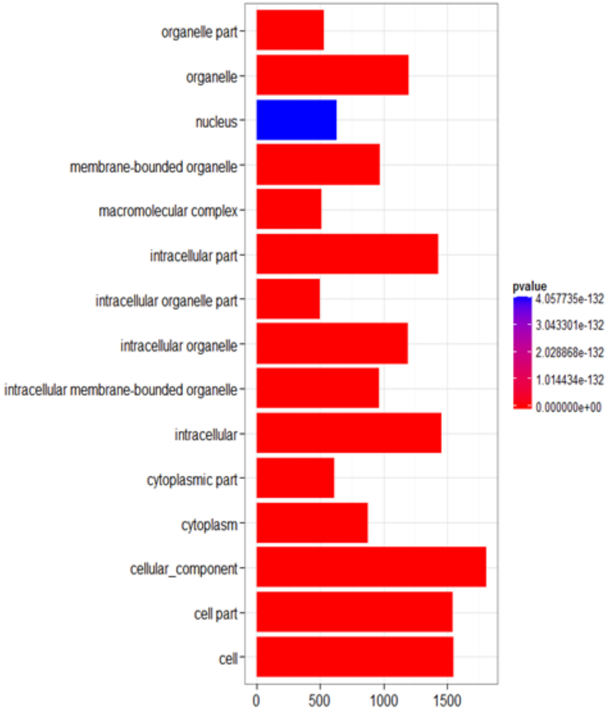
```
library(org.Hs.eg.db)
over <- getEnrichedGO(annotatedPeak[1:500], orgAnn="org.Hs.eg.db",
             maxP=0.01, minGOterm=10,
             multiAdjMethod="BH",
             condense=FALSE)
```

```
library(org.Hs.eg.db)
library(reactome.db)
enriched.PATH = getEnrichedPATH(annotatedPeak, orgAnn="org.Hs.eg.db",
pathAnn="reactome.db", maxP=0.01, minPATHterm=10,
multiAdjMethod=NULL)
```

GO enrichment







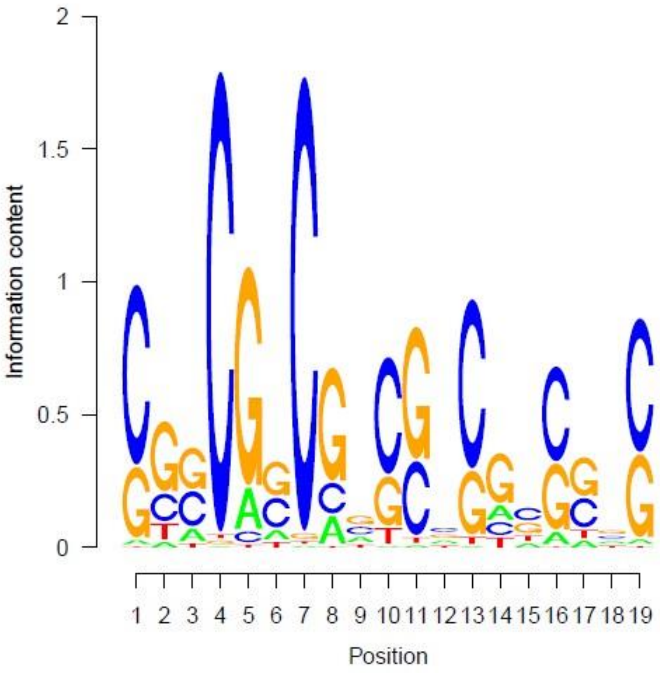
Biological process

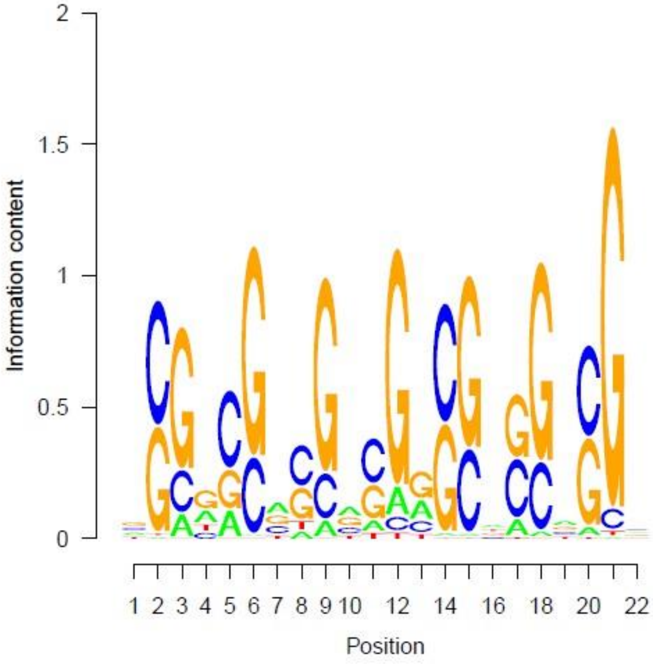
Molecular function

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)
```

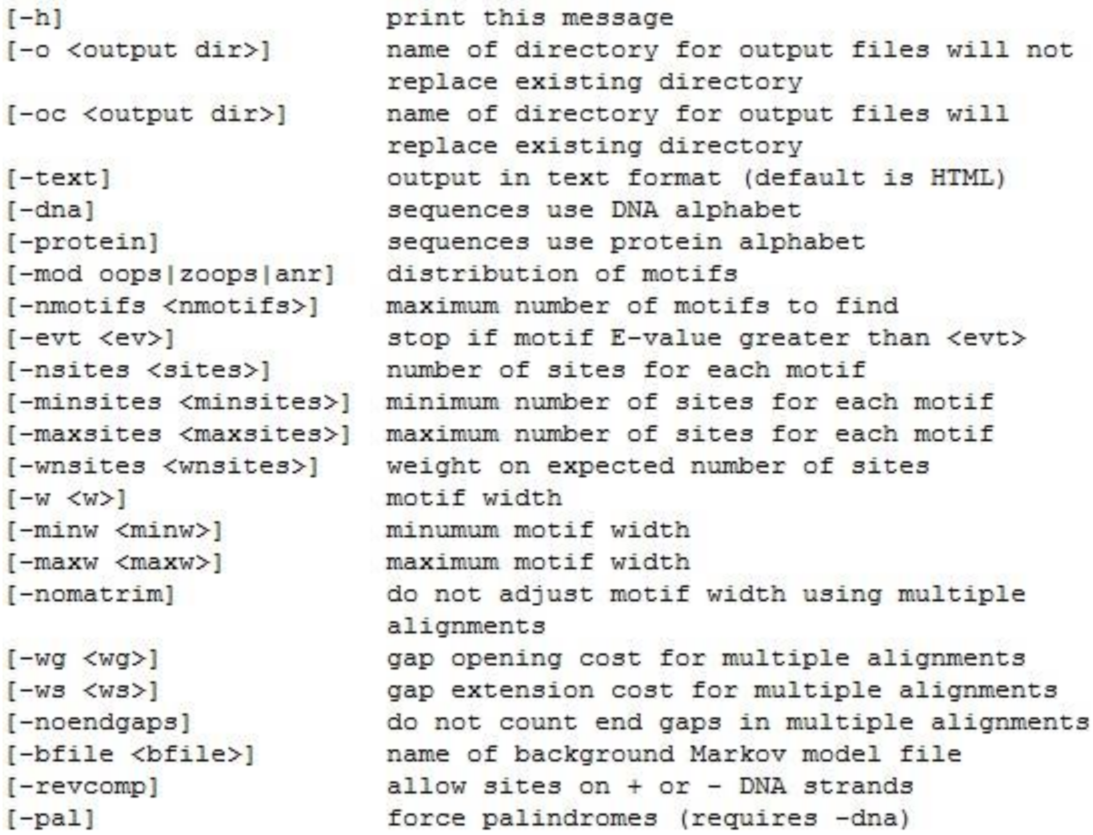




- rGADEM -motif discovery
- MotifRG -motif discovery
- MotIV -map motif to known TFBS, visualize logos
- motifStack -plot sequence logos
- MotifDb -motif database
- PWMenrich -motif enrichment analysis
- TFBSTools – R interface to the JASPAR database

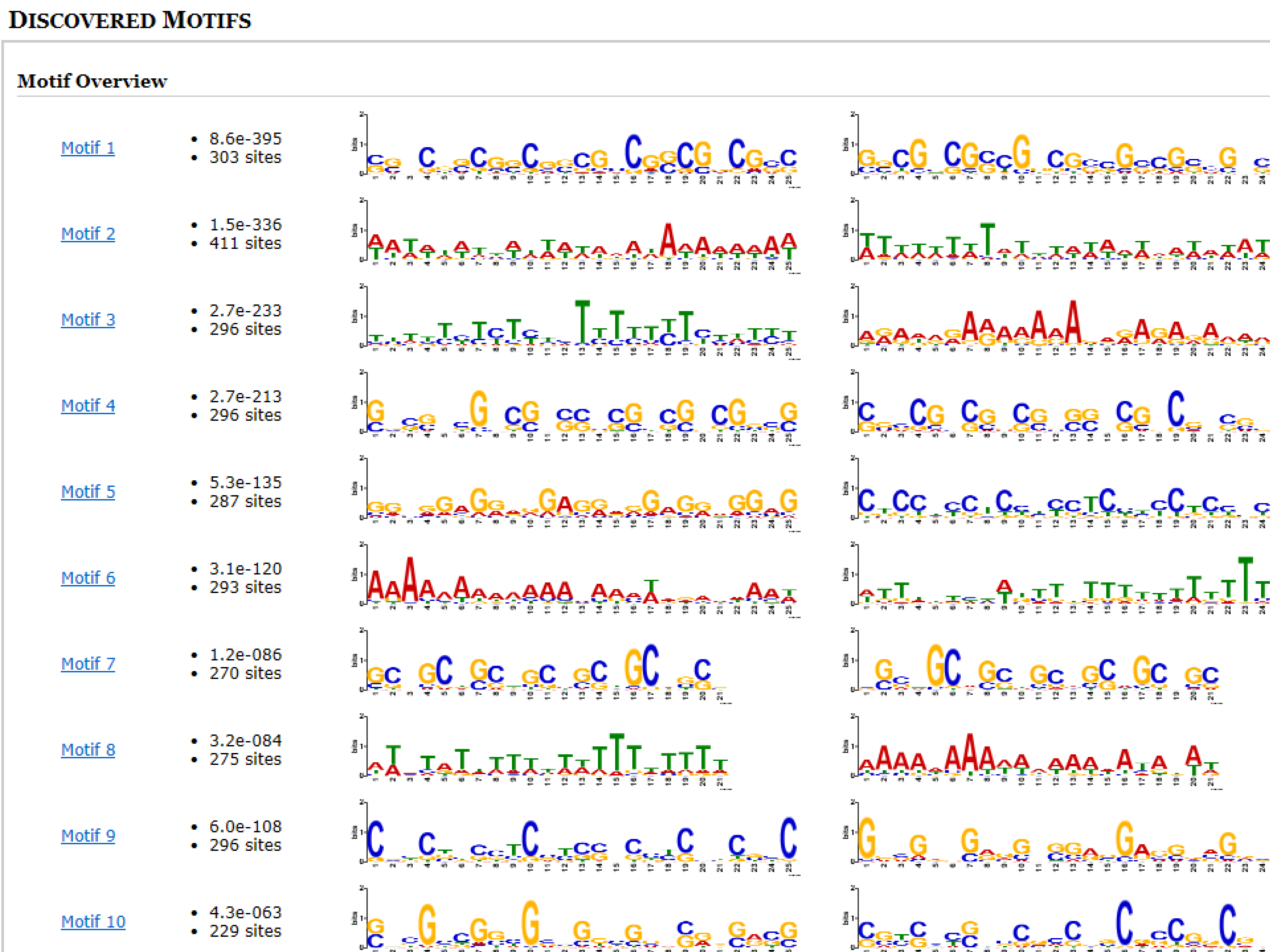
Motif analysis

meme <in.fas> option



```
[-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)



Discovered Motifs

Motif Overview

Motif 1

- 8.6e-395
- 303 sites

Motif 2

- 1.5e-336
- 411 sites

Motif 3

- 2.7e-233
- 296 sites

Motif 4

- 2.7e-213
- 296 sites

Motif 5

- 5.3e-135
- 287 sites

Motif 6

- 3.1e-120
- 293 sites

Motif 7

- 1.2e-086
- 270 sites

Motif 8

- 3.2e-084
- 275 sites

Motif 9

- 6.0e-108
- 296 sites

Motif 10

- 4.3e-063
- 229 sites