HOMER(v3.15, 8-2-2012)

Software for motif discovery and next generation sequencing analysis

Summer school in Genomic Bioinformatics
21.8.2012
User experiences
Minna Kaikkonen

HOMER presented

- = Hypergeometric Optimization of Motif EnRichment
- Suite of tools for Motif Discovery and next-generation sequencing analysis.
- It is a collection of command line programs for unix-style operating systems written in Perl and C++.
- HOMER contains many useful tools for analyzing ChIP-Seq, GRO-Seq, RNA-Seq, DNase-Seq, and numerous other types of functional genomics sequencing data sets.

ChIP-Seq: Isolation and sequencing of genomic DNA "bound" by a specific transcription factor, covalently modified histone, or other nuclear protein. This methodology provides genome-wide maps of factor binding. Most of HOMER's routines cater to the analysis of ChIP-Seq data.

MNase-Seq: Micrococcal Nuclease (MNase) is a restriction enzyme that degrades genomic DNA not wrapped around histones. The remaining DNA represents nucleosomal DNA, and can be sequencing to reveal nucleosome positions along the genome. This method can also be combined with ChIP to map nucleosomes that contain specific histone modifications.

RNA-Seq: Extraction, fragmentation, and sequencing of RNA populations within a sample. The replacement for gene expression measurements by microarray. There are many variants on this, such as Ribo-Seq (isolation of ribosomes translating RNA), small RNA-Seq (to identify miRNAs), etc.

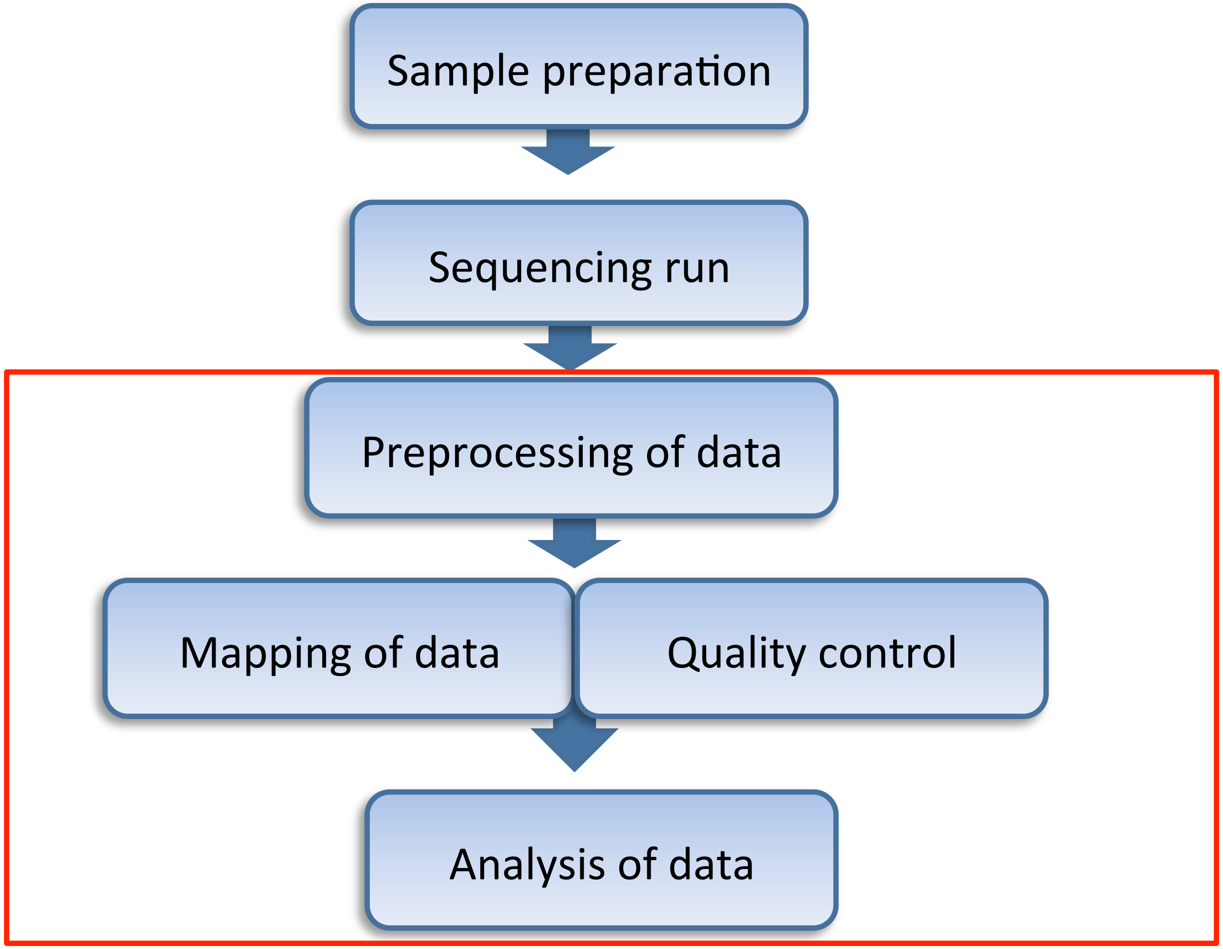
GRO-Seq: RNA-Seq of nascent RNA. Transcription is halted, nuclei are isolated, labeled nucleotides are added back, and transcription briefly restarted resulting in labeled RNA molecules. These newly created, nascent RNAs are isolated and sequenced to reveal "rates of transcription" as opposed to the total number of stable transcripts measured by normal RNA-seq.

History of HOMER

Chris Benner, UCSD

- “There was basically nothing else at the time” (in 2005)
- “We wanted to analyze the data in different ways and in these cases you need to write your own software since most tools perform very specific tasks that may not suite our specific needs.”
- “I "packaged" it up as a public resource for my collaborators so they could conduct their analysis by themselves.” –> Free software!

Presentation of tools flowchart



Sample preparation

Sequencing run

Preprocessing of data

Mapping of data

Quality control

Analysis of data

Sequence manipulation and mapping of NGS data

- **homerTools [command] [command specific options]**
  - barcodes - separate FASTQ file by barcodes
  - trim - trim adapter sequences or fixed sizes from FASTQ files(also splits)
  - freq - calculate position-dependent nucleotide/ dinucleotide frequencies
  - decontaminate - remove bad tags from a contaminated tag directory
  - extract - extract specific sequences from FASTA file(s)
  - cluster - hierarchical clustering of a NxN distance matrix
- For mapping uses **Bowtie**

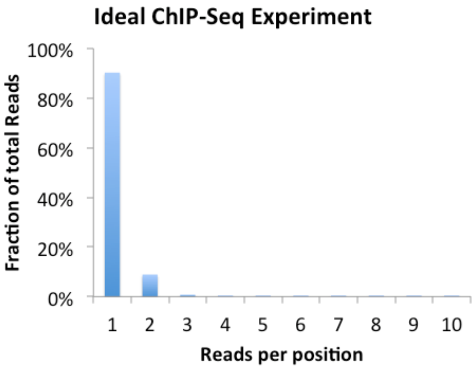
Making tag directories and quality control

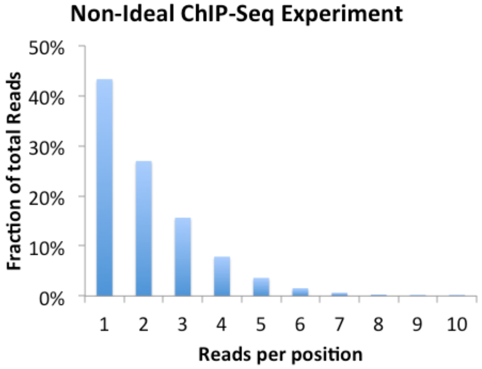
- Parses through the alignment file and splits the tags into separate files based on their chromosome. As a result, several *.tags.tsv files are created in the output directory. These are made to very efficiently return to the data during downstream analysis.

```
makeTagDirectory <Output Directory Name> [options] <alignment file1>
```

- At the same time performs basic quality control analysis

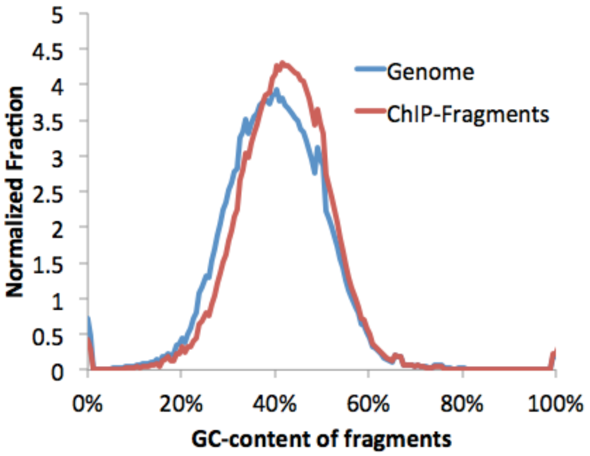
1. **tagInfo.txt**
2. **tagLengthDistribution.txt**
3. **tagCountDistribution.txt**
4. **tagAutocorrelation.txt**





Making tag directories and quality control

makeTagDirectory <Output Directory Name> [options] –genome –checkGC <alignment file1>

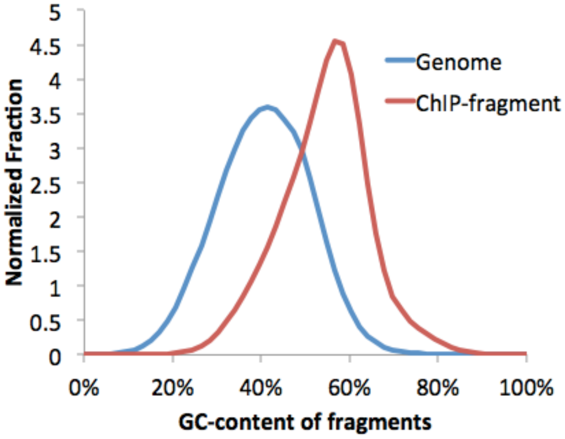


5.tagFreq.txt

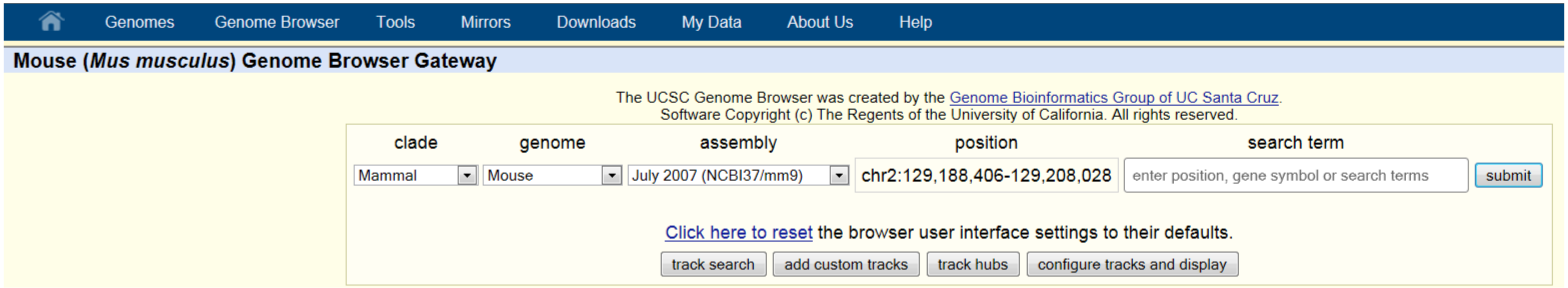
6.tagFreqUniq.txt

7.tagGCcontent.txt

8.genomeGCcontent.txt



Visualization of data



Mouse (*Mus musculus*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

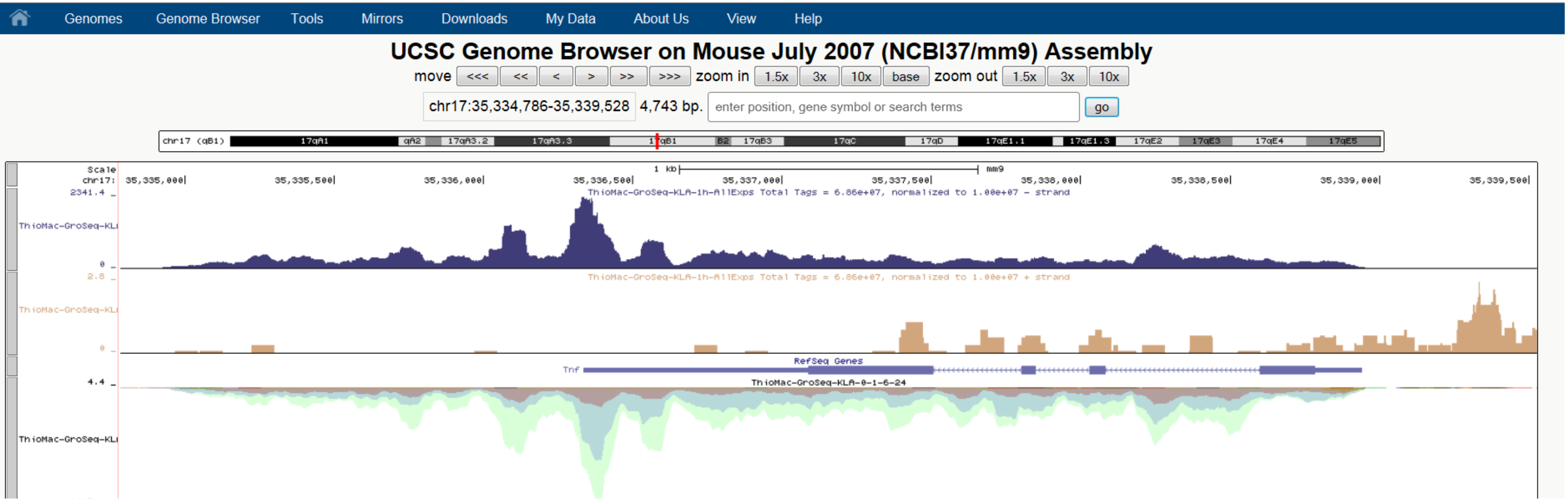
clade	genome	assembly	position	search term
Mammal	Mouse	July 2007 (NCBI37/mm9)	chr2:129,188,406-129,208,028	enter position, gene symbol or search terms

Click here to reset the browser user interface settings to their defaults.

makeUCSCfile

makeBigWig.pl
makeMultiWigHub.pl
(Dowload bedGraphToBigWig from UCSC)

Visualization of data



UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly

chr17:35,334,786-35,339,528 4,743 bp.

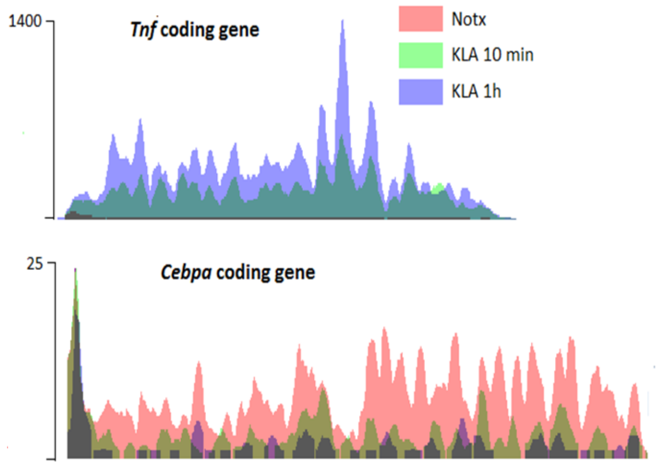
ThioMac-GroSeq-KLA-1h-AllExps Total Tags = 6.86e+07, normalized to 1.00e+07 - strand

ThioMac-GroSeq-KLA-1h-AllExps Total Tags = 6.86e+07, normalized to 1.00e+07 + strand

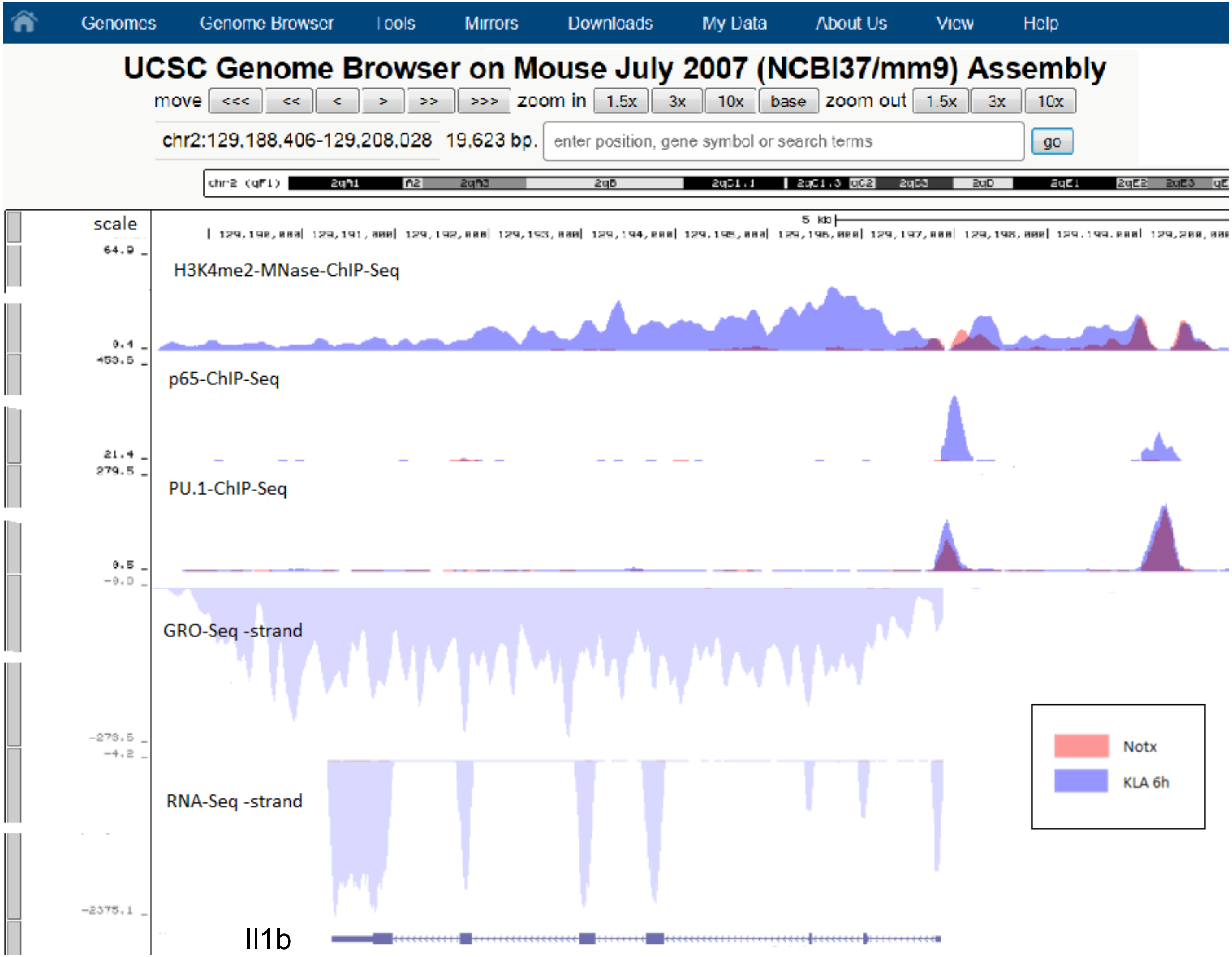
ThioMac-GroSeq-KLA-0-1-6-24

makeUCSCfile

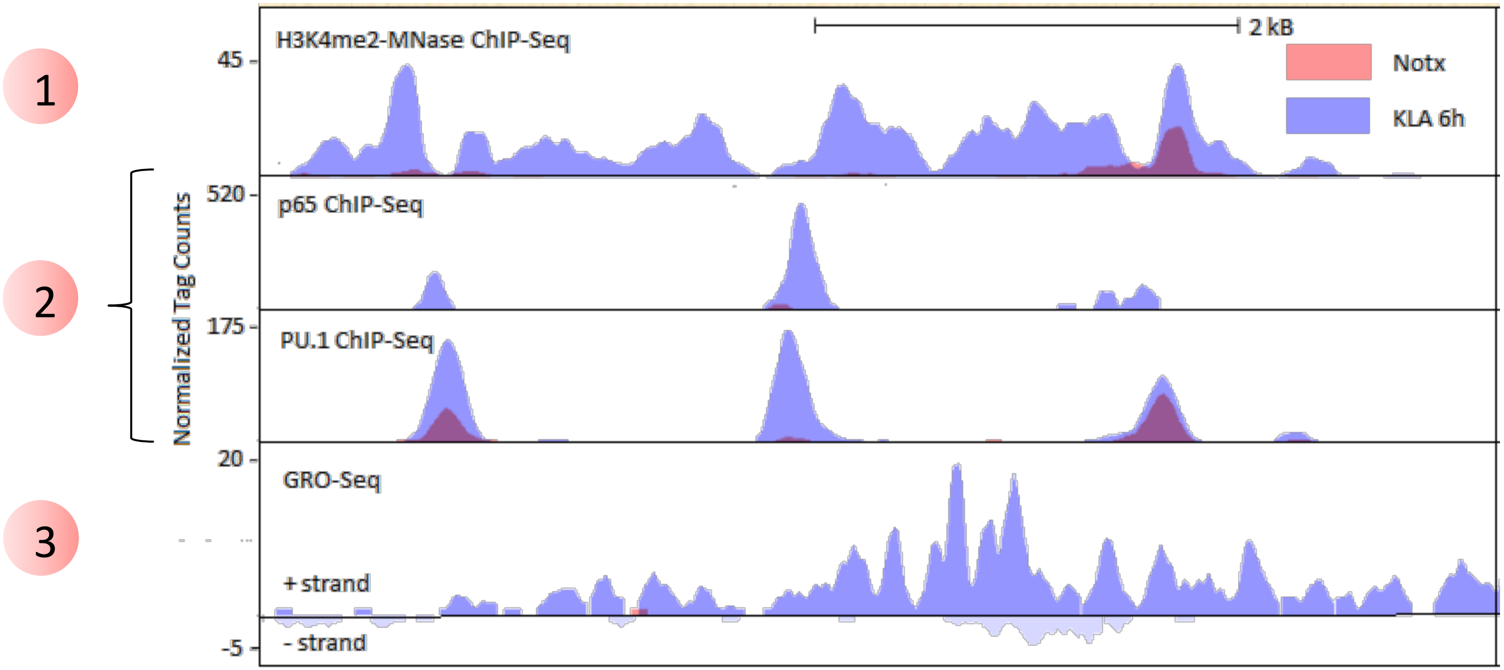
makeBigWig.pl
makeMultiWigHub.pl
(Dowload bedGraphToBigWig from UCSC)



Using HOMER for analysis of macrophage responses to inflammatory stimuli


Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
View
Help
UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
chr2:129,188,406-129,208,028 19,623 bp.
enter position, gene symbol or search terms
go
chr2 (qF1)
scale
5 kb
H3K4me2-MNase-ChIP-Seq
p65-ChIP-Seq
PU.1-ChIP-Seq
GRO-Seq -strand
RNA-Seq -strand
Il1b
Notx
KLA 6h

findPeaks



1

```
findPeaks <tag directory> -i <input> -style histone > H3K4me2-Regions.txt
```

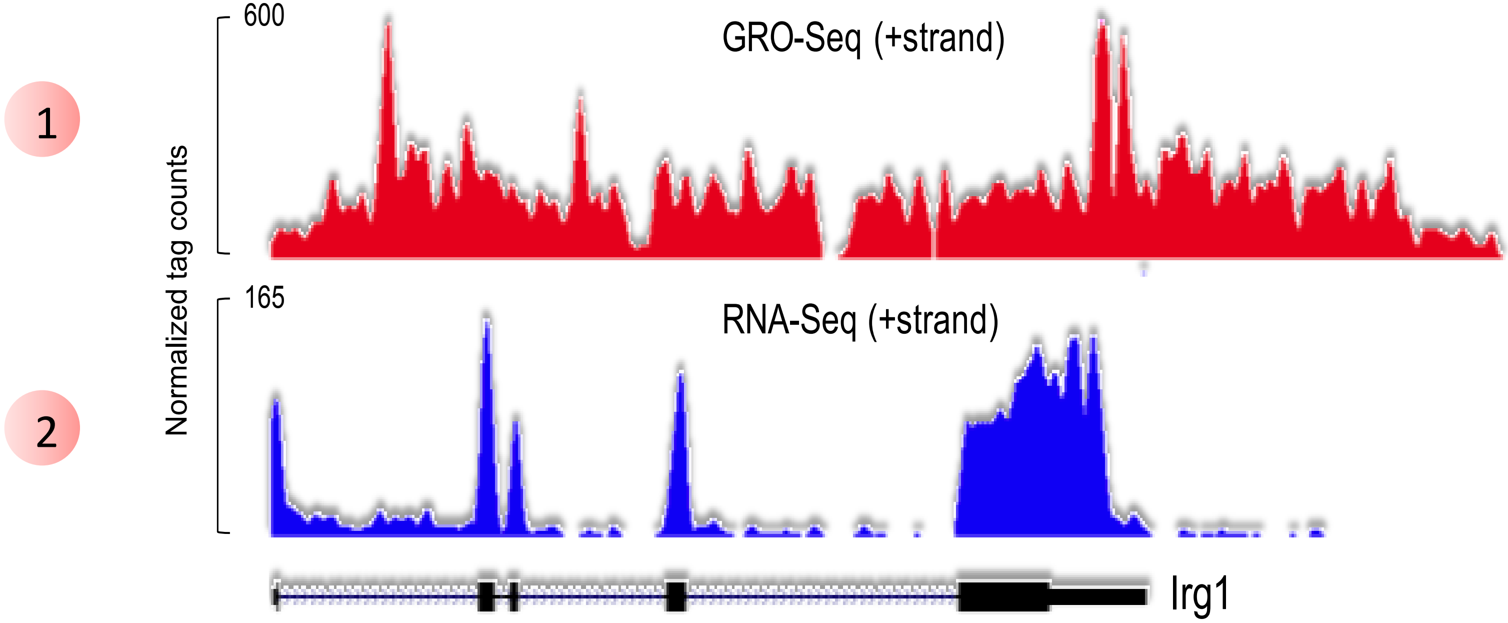
2

```
findPeaks <tag directory> -i <input> -style factor > PU.1/P65-Peaks.txt
```

3

```
findPeaks <tag directory> -style groseq > GRO-regions.txt
```

analyzeRNA.pl



```
1   analyzeRNA.pl rna mm9  -d <Directories> –count genes > GROSeq.txt

2   analyzeRNA.pl rna mm9  -d <Directories> –count exons > RNASeq.txt
```

De novo motif analysis

Regions of interest (ChIP/-GRO-Seq)

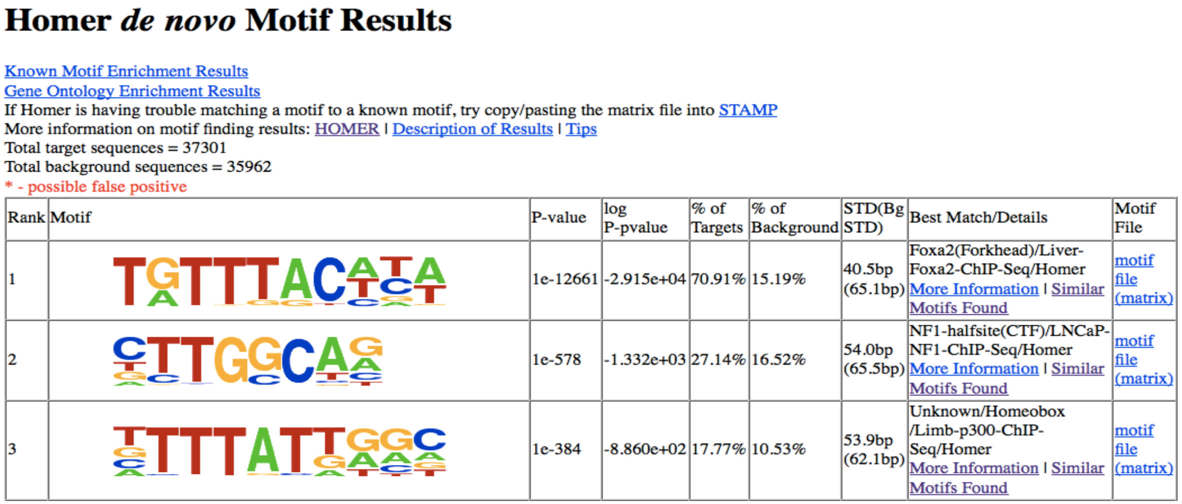
#PeakID	chr	start	end	strand	focus ratio	Peaks Score	Total Tags
chr13-1	chr13	58605058	58605210	+	0,8	76	48,9
chr19-1	chr19	5291667	5291819	+	0,969	75	45,4
chr13-3	chr13	1,09E+08	1,09E+08	+	0,923	73	47,4
chr5-3	chr5	36905894	36906046	+	0,946	70	41,4
chr9-17	chr9	70108221	70108373	+	0,869	70	40,9
chr9-19	chr9	48366927	48367079	+	0,95	68	42,4
chr1-2	chr1	1,93E+08	1,93E+08	+	0,905	63	36,4
chr2-14	chr2	50697589	50697741	+	0,84	61	32,4
chr6-3	chr6	89230645	89230797	+	0,854	61	35,4

Genes of interest (RNA-/GRO-Seq)

Gene ID	chr	start	end	strand	RNA Length(av	ne Length(av	Symbol
NM_027835	chr2	62433849	62484312	-	5519	50463	Ifih1
NM_009834	chr3	51028368	51055576	+	3069	27208	Ccrn4l
NM_009452	chr1	1,63E+08	1,63E+08	+	1609	22769	Tnfsf4
NM_008689	chr3	1,35E+08	1,35E+08	-	4118	106893	Nfkb1
NM_009140	chr5	91332924	91334964	+	1083	2040	Cxcl2
NM_021384	chr12	27127607	27141317	-	3785	13710	Rsad2
NM_011157	chr10	61957175	61970503	-	938	13328	Srgn
NM_009404	chr17	57244807	57247180	+	1230	2373	Tnfsf9
NM_013652	chr11	83476085	83478185	+	660	2100	Ccl4

findMotifsGenome.pl
<peakfile> mm9 <Output_name>

findMotifs.pl
<peakfile> mouse <Output_name>



Homer *de novo* Motif Results

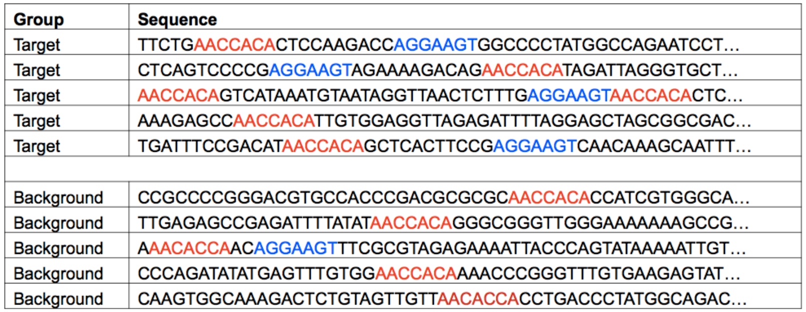
Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 37301
Total background sequences = 35962
* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		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		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		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)

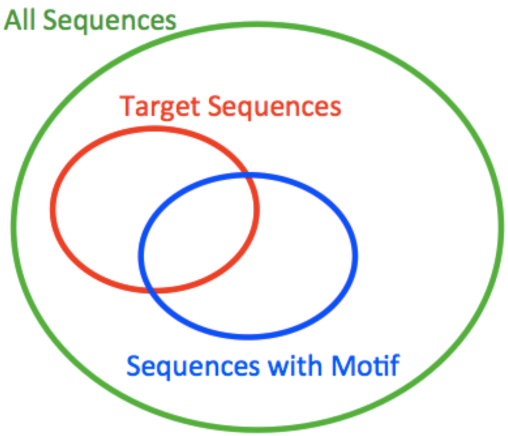
De novo Motif Discovery in HOMER

HOMER was primarily written as a *de novo* motif discovery algorithm and is well suited for finding 8-20 bp motifs in large scale genomics data.

Differential Motif Discovery: Finds sequences that are specifically enriched in the target set



Group	Sequence
Target	TTCTGAACCACACTCCAAGACCAGGAAGTGGCCCCTATGGCCAGAATCCT…
Target	CTCAGTCCCCGAGGAAGTAGAAAAGACAGAACCACATAGATTAGGGTGCT…
Target	AACCACAGTCATAAATGTAATAGGTTAACTCTTTGAGGAAGTAACCACACTC…
Target	AAAGAGCCAACCACATTGTGGAGGTTAGAGATTTTAGGAGCTAGCGGCGAC…
Target	TGATTTCCGACATAACCACAGCTCACTTCCGAGGAAGTCAACAAAGCAATTT…
Background	CCGCCCCGGGACGTGCCACCCGACGCGCGCAACCACACCATCGTGGGCA…
Background	TTGAGAGCCGAGATTTTATATAACCACAGGGCGGGTTGGGAAAAAAAGCCG…
Background	AAACACCAACAGGAAGTTTCGCGTAGAGAAAATTACCCAGTATAAAAATTGT…
Background	CCCAGATATATGAGTTTGTGGAACCACAAAACCCGGGTTTGTGAAGAGTAT…
Background	CAAGTGGCAAAGACTCTGTAGTTGTTAACACCACCTGACCCTATGGCAGAC…



Objective: Find the motifsthat are highly ***enriched*** in target sequences

(Maximize the overlap)

Pre-processing Phase:

- Remove redundant sequences
- Normalize GC-content

Exhaustive Search Phase:

- Screen all possible oligos for enrichment

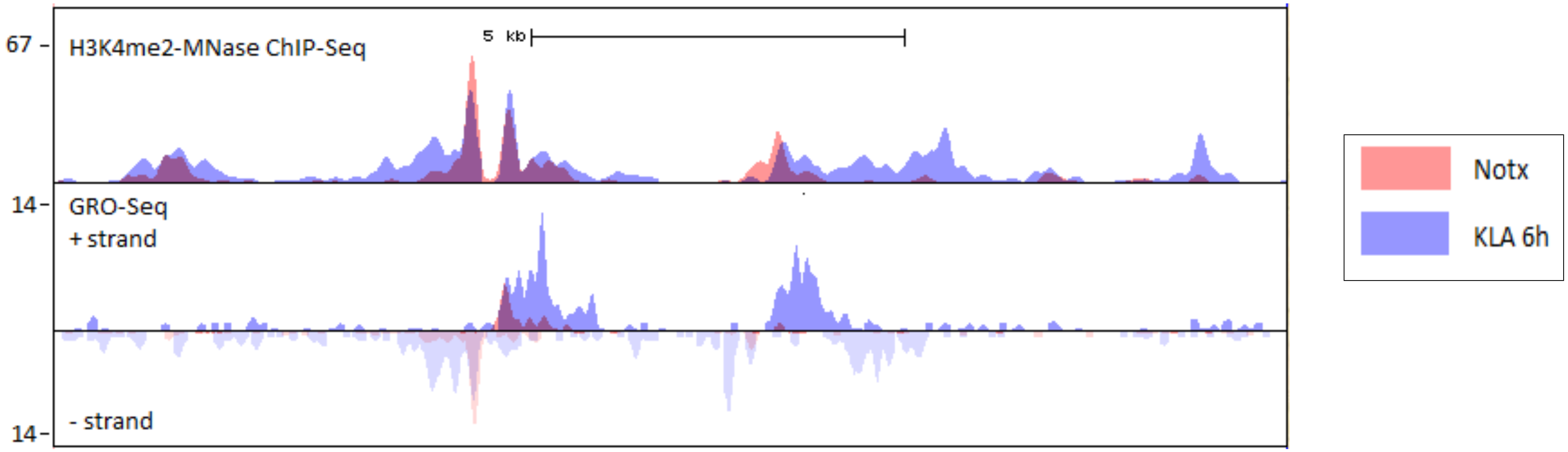
Local Optimization Phase:

- Expand promising oligos into probability matrices
- Iteratively improve matrices by considering individual contributions from different oligos

AnnotatePeaks.pl

- Associates peaks with nearby genes
- Calculates tag densities from different exps. (-d)
- Combines with gene expression data (-gene)
- Performs Gene Ontology Analysis (-go)
- Performs Genome Ontology Anal.(genomeOntology)
- Finds motif occurrences in peaks (-m)
- Creates histograms (-hist) or heatmaps (-ghist)

AnnotatePeaks.pl; example



```
annotatePeaks.pl <H3K4me2 peaks>  mm9  -d Directory 1 Directory 2 –size 2000 -log > analysis.txt
```

PeakID	Chr	Start	End	Strand	Annotation	Detailed Annotation	Distance to TSS	Nearest PromoterID	Entrez ID	Unigene	Refseq	Gene Name	Directory 1	Directory 2
Merged-chr19-23567309-2	chr19	23566309	23568309	+	Intergenic	Intergenic	-44497	NM_174857	71738	Mm.50841	NM_174857	Mamdc2	3,59	1,75
5-chrX-2536	chrX	1,37E+08	1,37E+08	+	Intergenic	B1_Mur3|SINE|Alu	21633	NM_010286	14605	Mm.485388	NM_001077364	Tsc22d3	1,44	1,89
Merged-chr6-115257312-4	chr6	1,15E+08	1,15E+08	+	Intergenic	PB1D10|SINE|Alu	-53927	NM_001127330	19016	Mm.3020	NM_001127330	Pparg	0,96	0,66
Merged-chr3-79668902-5	chr3	79667902	79669902	+	Intergenic	Intergenic	-20950	NM_133187	68659	Mm.460017	NM_133187	Fam198b	0,07	0,54
Merged-chrX-150099684-4	chrX	1,5E+08	1,5E+08	+	Intergenic	Intergenic	31139	NM_175429	207474	Mm.271572	NM_175429	Kctd12b	0,56	0,51
5-chr13-11712	chr13	49346769	49348769	+	Intergenic	RMER15|LTR|ERVL	-3694	NM_028340	66329	Mm.34308	NM_025491	Susd3	1,75	0,42
2-chr17-19182	chr17	87084660	87086660	+	Intergenic	ORR1A2|LTR|MaLR	-67544	NM_010137	13819	Mm.1415	NM_010137	Epas1	3,37	2,24
Merged-chr6-88627558-3	chr6	88626558	88628558	+	Intergenic	MTD-int|LTR|MaLR	-46848	NM_001166249	23945	Mm.272197	NM_001166249	Mgll	2,03	3,84
3-chr10-26936	chr10	1,28E+08	1,28E+08	+	Intergenic	Intergenic	-11266	NM_001164197	58223	Mm.131266	NM_001164197	Mmp19	3,58	4,36

annotatePeaks.pl

Regions of interest (ChIP/-GRO-Seq)

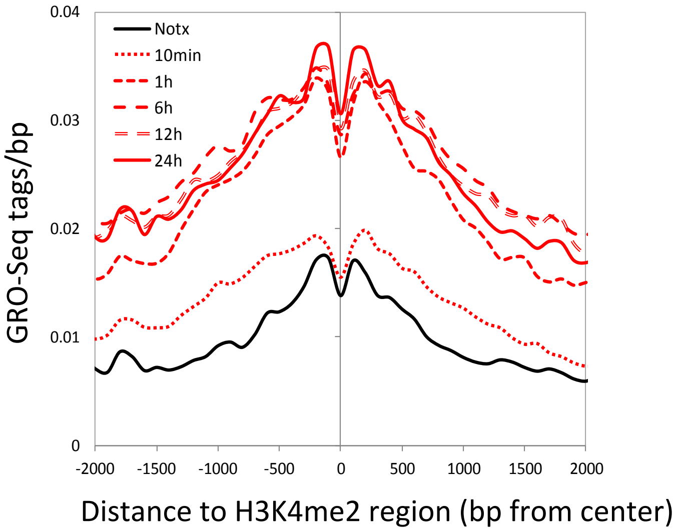
#PeakID	chr	start	end	strand	focus ratio	Peaks Score	Total Tags
chr13-1	chr13	58605058	58605210	+	0,8	76	48,9
chr19-1	chr19	5291667	5291819	+	0,969	75	45,4
chr13-3	chr13	1,09E+08	1,09E+08	+	0,923	73	47,4
chr5-3	chr5	36905894	36906046	+	0,946	70	41,4
chr9-17	chr9	70108221	70108373	+	0,869	70	40,9
chr9-19	chr9	48366927	48367079	+	0,95	68	42,4
chr1-2	chr1	1,93E+08	1,93E+08	+	0,905	63	36,4
chr2-14	chr2	50697589	50697741	+	0,84	61	32,4
chr6-3	chr6	89230645	89230797	+	0,854	61	35,4

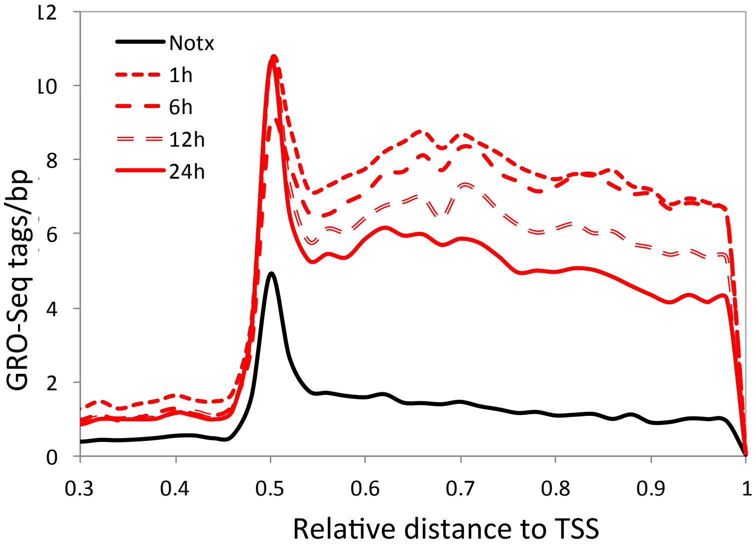
Genes of interest (RNA-/GRO-Seq)

Gene ID	chr	start	end	strand	RNA Length(a	ne Length(av	Symbol
NM_027835	chr2	62433849	62484312	-	5519	50463	Ifih1
NM_009834	chr3	51028368	51055576	+	3069	27208	Ccrn4l
NM_009452	chr1	1,63E+08	1,63E+08	+	1609	22769	Tnfsf4
NM_008689	chr3	1,35E+08	1,35E+08	-	4118	106893	Nfkb1
NM_009140	chr5	91332924	91334964	+	1083	2040	Cxcl2
NM_021384	chr12	27127607	27141317	-	3785	13710	Rsad2
NM_011157	chr10	61957175	61970503	-	938	13328	Srgn
NM_009404	chr17	57244807	57247180	+	1230	2373	Tnfsf9
NM_013652	chr11	83476085	83478185	+	660	2100	Ccl4

```
annotatePeaks.pl <Peakfile>
mm9  -d <Directories 0-24 h>  –size 4000
–hist 50 > histogramGRO.txt
```

```
annotatePeaks.pl tss mm9  -d <Directories 0-24h>
–size given –hist 50 > histogramTSS.txt
```

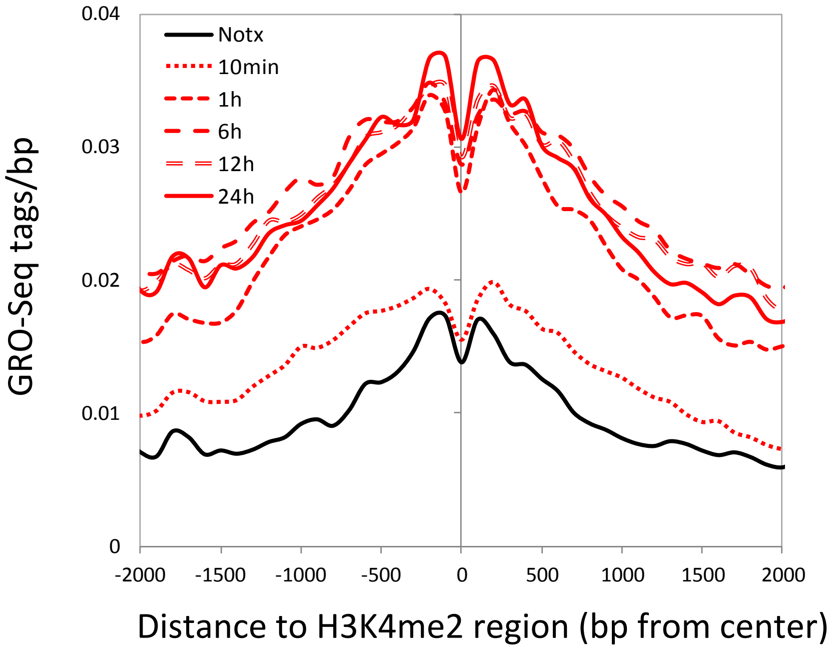


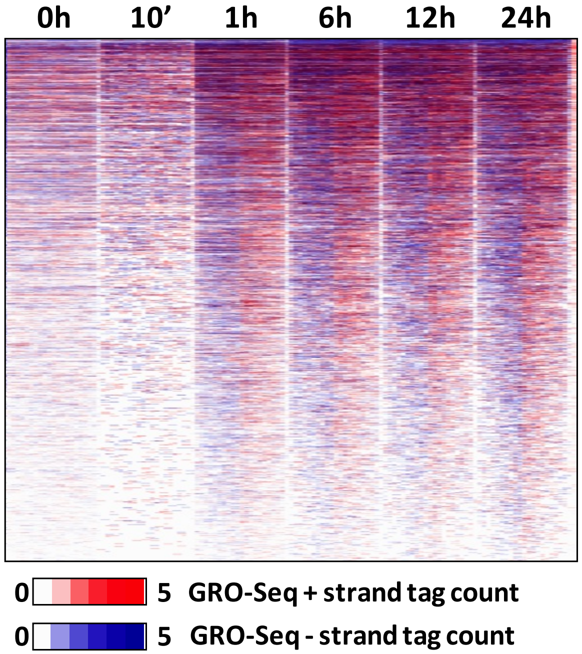


AnnotatePeaks.pl; example

```
annotatePeaks.pl <H3K4me2 peaks>
mm9  -d <Directories 0-24 h>  –size 4000
 –hist 50 > histogram.txt
```

```
annotatePeaks.pl <H3K4me2 peaks>
 mm9  -d <Directories 0-24 h> –size 4000
–hist 50 –ghist > heatmap.txt
```





Analysis of data

ChIP-/GRO-Seq

findPeaks
-style factor/histone/groseq

findMotifsGenome.pl

annotatePeaks.pl

RNA-/GRO-Seq

analyzeRNA.pl
–count exons/genes

findMotifs.pl

annotatePeaks.pl

Questions