Galaxy Platform For NGS Data Analyses

Weihong Yan
wyan@chem.ucla.edu

Collaboratory Web Site
http://collaboratory.lifesci.ucla.edu

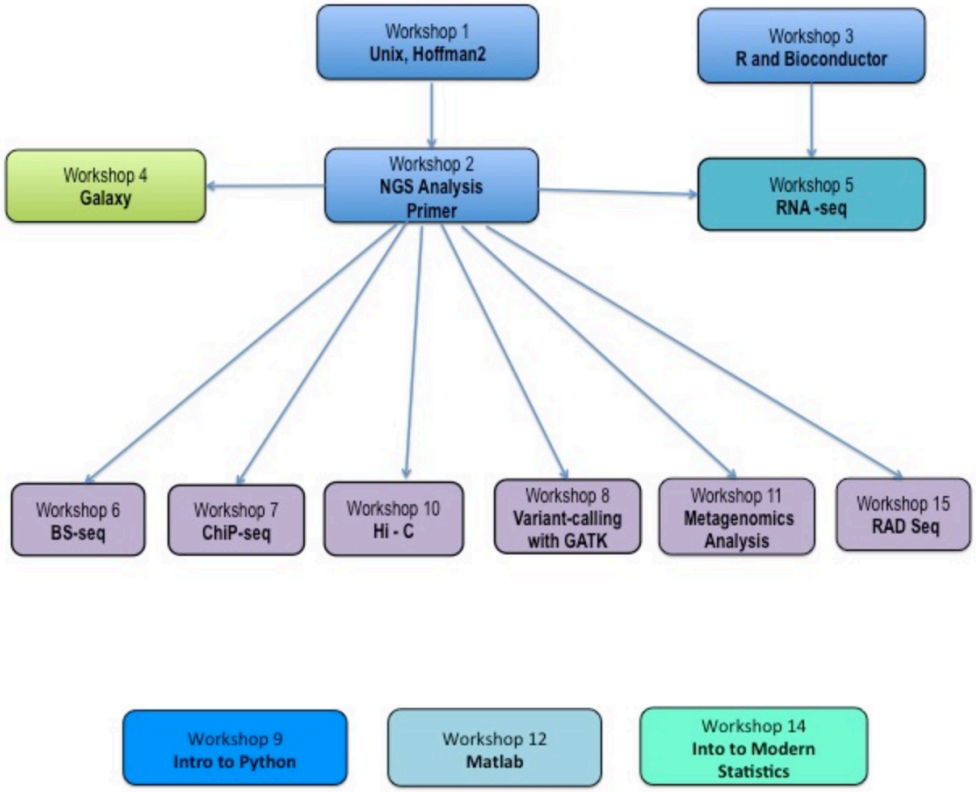
Workshop Outline

- ✓ Day 1
  - UCLA galaxy and user account
  - Galaxy web interface and management
  - Tools for NGS analyses and their application
  - Data formats
  - Build/share workflow and history
  - Q and A

- ✓ Day 2
  - Galaxy Tools for RNA-seq analysis
  - Galaxy Tools for ChIP-seq analysis
  - Galaxy Tools for annotation.
  - Q and A

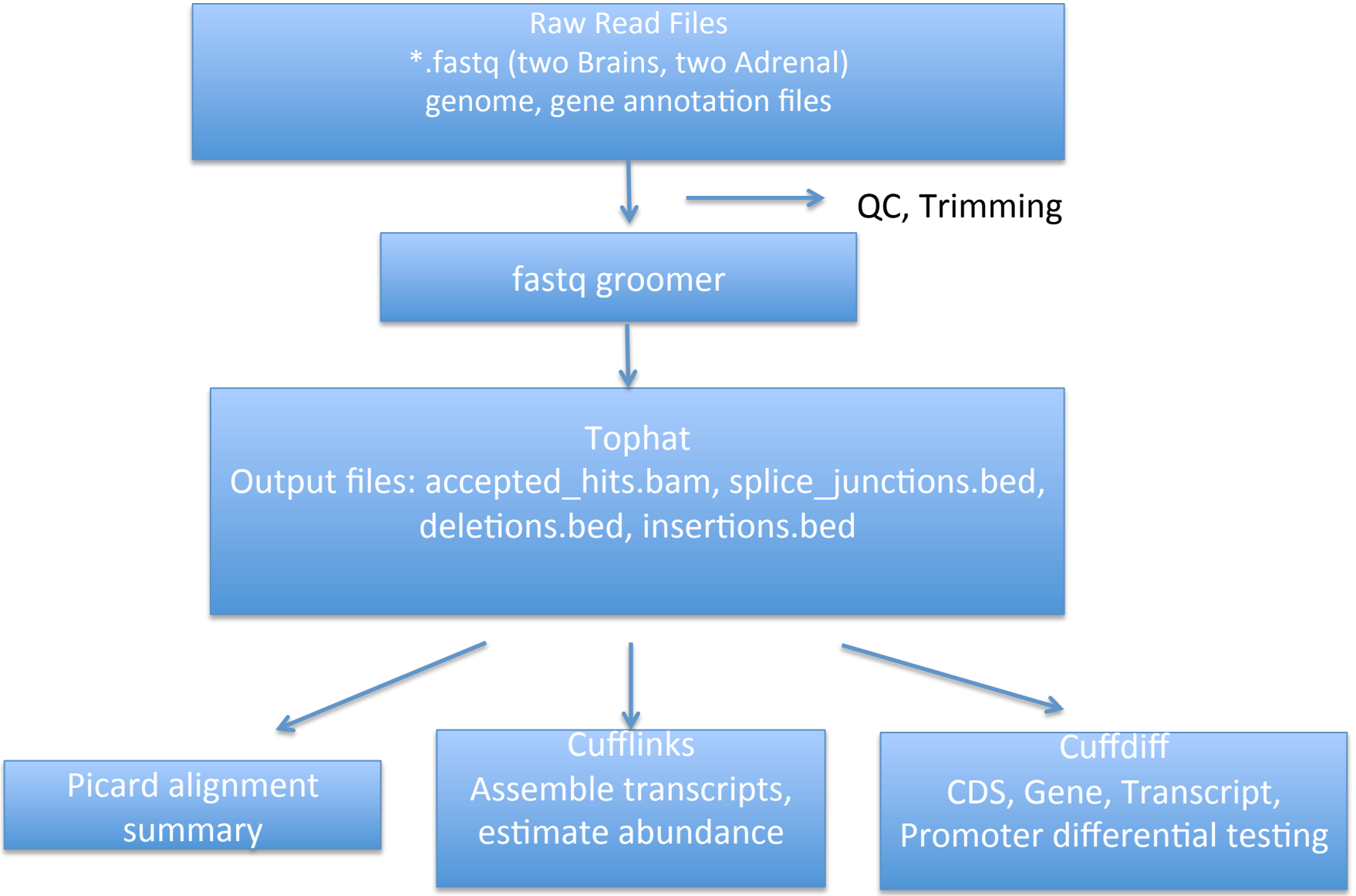
*** Published datasets/results will be used in the tutorial

Collaboratory Workshops

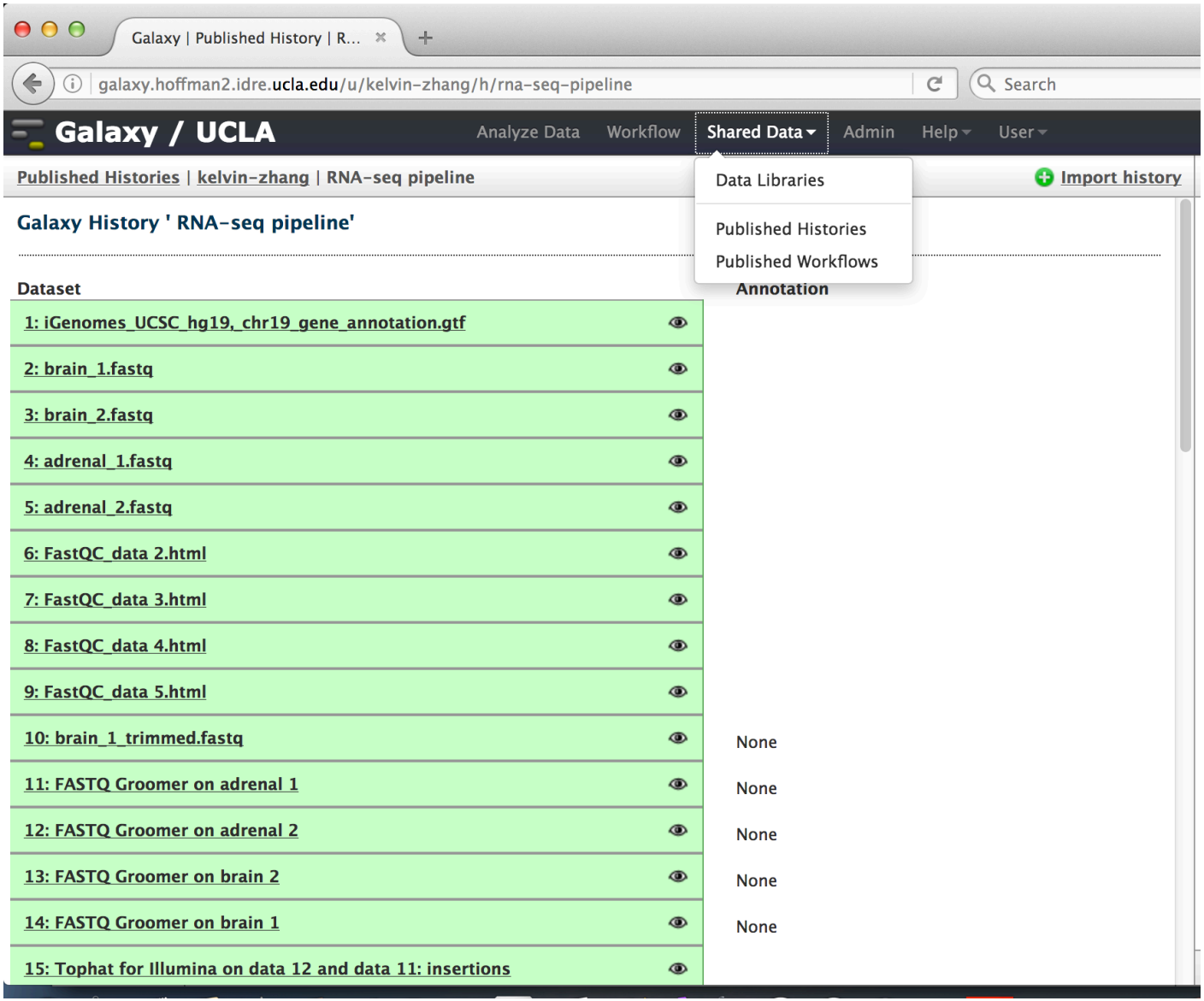


- W1: UNIX COMMAND LINE
- W2: NGS ANALYSIS
- W3: INTRO TO R
- W4: GALAXY FOR NGS DATA ANALYSIS
- W5: RNA-SEQ ANALYSIS
- W6: BS-SEQUENCING
- W7: CHIP-SEQ ANALYSIS
- W8: VARIANT DISCOVERY WITH GATK
- W9: PYTHON
- W 10: HI-C
- W 11: METAGENOMICS ANALYSIS
- W12: MATLAB
- W14: INTRODUCTION TO MODERN STATISTICS
- W15: RAD-SEQ ANALYSIS
- COLLABORATORY PYTHON USERS GROUP

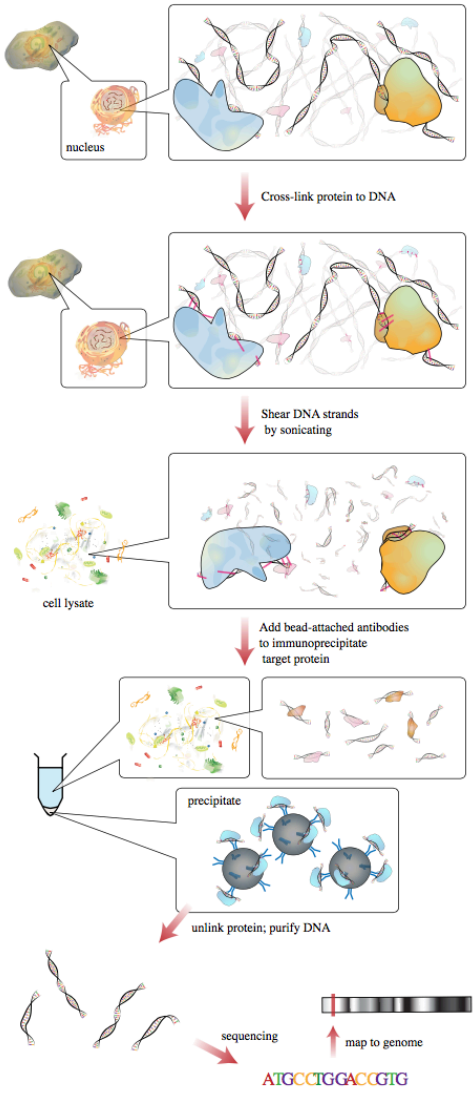
RNA-Seq Workflow



RNA-Seq Workflow



ChIP-seq Data Analysis



Applications

- ✓ Protein-DNA interaction sites
- ✓ Nucleosome position
  histone modification
- ✓ DNA methylation

http://en.wikipedia.org/wiki/ChIP-sequencing

PROTOCOL

Identifying ChIP-seq enrichment using MACS

Jianxing Feng[1,3], Tao Liu[2,3], Bo Qin[1], Yong Zhang[1] & Xiaole Shirley Liu[2]

[1]Department of Bioinformatics, School of Life Sciences and Technology, Tongji University, Shanghai, China. [2]Department of Biostatistics and Computational Biology, Harvard School of Public Health, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [3]These authors contributed equally to this work. Correspondence should be addressed to Y.Z. (yzhang@tongji.edu.cn) and X.S.L. (xsliu@jimmy.harvard.edu).

Published online 30 August 2012; doi:10.1038/nprot.2012.101

Model-based analysis of ChIP-seq (MACS) is a computational algorithm that identifies genome-wide locations of transcription/chromatin factor binding or histone modification from ChIP-seq data. MACS consists of four steps: removing redundant reads, adjusting read position, calculating peak enrichment and estimating the empirical false discovery rate (FDR). In this protocol, we provide a detailed demonstration of how to install MACS and how to use it to analyze three common types of ChIP-seq data sets with different characteristics: the sequence-specific transcription factor FoxA1, the histone modification mark H3K4me3 with sharp enrichment and the H3K36me3 mark with broad enrichment. We also explain how to interpret and visualize the results of MACS analyses. The algorithm requires ~3 GB of RAM and 1.5 h of computing time to analyze a ChIP-seq data set containing 30 million reads, an estimate that increases with sequence coverage. MACS is open source and is available from http://liulab.dfci.harvard.edu/MACS/.

INTRODUCTION

Researchers have widely used the process of ChIP-seq[1] to map transcription factor binding sites and histone modification status on a genome-wide scale[2]. ChIP comprises a few basic steps: cross-linking

version of MACS (1.4.2) to publicly available ChIP-seq data[20] on a local computer. MACS is also available at the web-based ChIP-seq analysis portal Cistrome[21], which provides a complete workflow

http://www.nature.com/nprot/journal/v7/n9/pdf/nprot.2012.101.pdf

Model-based Analysis of ChIP-seq (MACS)

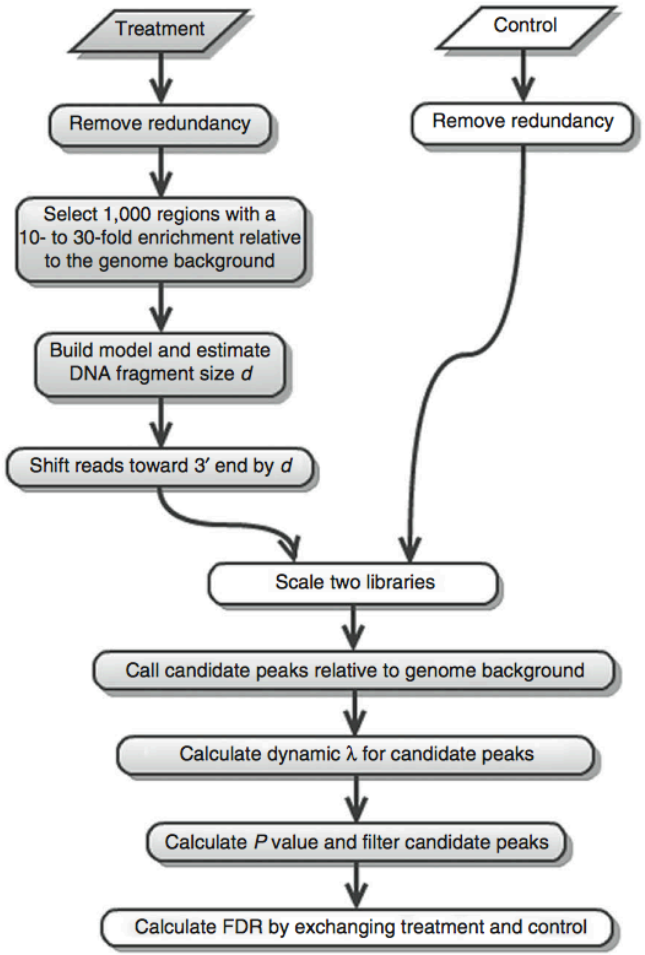


Figure 1 | Workflow of MACS 1.4.2. If the control sample is missing, then the steps shown in white boxes will be skipped (remove redundancy of the control sample, scale two libraries and calculate FDR by exchanging treatment and control).

- ✓ Removing redundant reads: MACS retains no more than one read per genomic location
- ✓ Building the peak model: Reads that map to the positive and negative strands often appear to the left and right of the protein-DNA interaction location.
- ✓ Calculate peak enrichment using local background normalization

MACS Protocol to Call Peaks

Running MACS to call peaks

6| We use four different ChIP-seq data sets to illustrate how to run MACS using varying parameters: use option A to call FoxA1 peaks; option B to call H3K4me3 peaks with fragment size estimation turned on; option C to call H3K4me3 peaks with a specified DNA fragment size; or option D to call H3K36me3 peaks. Please note that some of these data sets are very large and may take an hour or more to download.

(A) Calling FoxA1 peaks ● TIMING 90 min

(i) Locate the downloaded prebuilt index for Bowtie, and unpack the package using the following command:

```
> unzip hg19.ebwt.zip
```

This command will generate several files with names prefixed by 'hg19' in the current directory.

(ii) Download the HudsonAlpha Institute FoxA1 raw reads from http://cistrome.dfci.harvard.edu/MACSNatureProtocol/HAIB_T47D_FoxA1.tar.gz, locate the download directory, unpack the compressed file and map the raw reads to the reference genome using Bowtie by entering the following two commands:

```
> tar xvzf HAIB_T47D_FoxA1.tar.gz
> bowtie -m 1 -S -q /path_to/hg19 HAIB_T47D_FoxA1.fastq HAIB_T47D_FoxA1.sam
```

In these commands,

`-m 1` specifies that reads with only one hit on the genome are retained;

`-S` specifies the output format as SAM;

`-q` specifies the input format as FASTQ;

`/path_to/` is the directory containing the unzipped bowtie prebuilt indexes; and

`HAIB_T47D_FoxA1.fastq` contains the downloaded raw reads for FoxA1.

Please refer to the Bowtie manual for more information.

(iii) Run MACS in the same directory by entering the following command:

```
> macs14 -t HAIB_T47D_FoxA1.sam -n HAIB_T47D_FoxA1 -g hs -B -S --call-subpeaks
```

MACS Protocol to Call Peaks

```
> macs14 -t HAIB_T47D_FoxA1.sam -n HAIB_T47D_FoxA1 -g hs -B -S --call-subpeaks
```

The meanings of the parameters in this command are as follows (see also **Box 1** for further parameters that could be used):

`-t` specifies the file name for the ChIP-seq sample read alignment. MACS supports and can automatically detect any of the following file formats: SAM, BAM, BED, ELAND, ELANDMULTI, ELANDMULTIPET, ELANDEXPORT and BOWTIE. The user-specified parameter `--format` can override the automatic format detection.

`-g` specifies the genome size. The `hs` parameter is a shortcut for the approximate effective genome size of humans, which equals 2.7e9.

`-n` applies the prefix '`HAIB_T47D_FoxA1`' to the output file names.

`-B` generates signal files in the bedGraph format containing the extended read pileup at every base pair. This step is very time-consuming and memory-intensive; therefore, only specify `-B` if bedGraph output files are needed.

`-S` generates a single bedGraph file for the whole genome; otherwise, signal files will be generated for each chromosome separately.

Box 1 | Additional MACS parameters

Several additional key parameters that could be used to run MACS in Step 6A(iii) are as follows:
`--bw` sets the 'bandwidth', which is half of the sliding window size used in the model-building step.
`--mfold` specifies an interval of high-confidence enrichment ratio against the background on which to build the model. The default value '10, 30' means that a model will be built on the basis of regions having read counts that are 10- to 30-fold of the background.
`--pvalue` establishes a threshold *P* value: only peaks surpassing the threshold will be reported. The default threshold is 10^{-5}. Users can first set a loose *P* value cutoff so that a sufficient number of peaks will be reported and then select peaks having the smallest *P* values for downstream analyses.

Peak Calling Workflow

Step 1: Input dataset

Input Dataset

2: HAIB_T47D_FoxA1.fastq

type to filter

Step 2: Map with Bowtie for Illumina

Step 3: MACS14

Experiment Name
FoxA1 peaks

Paired End Sequencing
Single End

ChIP-Seq Tag File
Output dataset 'output' from step 2

ChIP-Seq Control File
Selection is Optional

Effective genome size
2700000000.0

Tag size
25

Band width
300

Pvalue cutoff for peak detection
1e-05

Select the regions within MFOLD range of high-confidence enrichment ratio against background to build model. The regions must be lower than upper limit, and higher than the lower limit
10,30

Parse xls files into into distinct interval files
False

Save shifted raw tag count at every bp into a wiggle file
Save

Resolution for saving wiggle files
10

Use fixed background lambda as local lambda for every peak region
False

The small nearby region in basepairs to calculate dynamic lambda. This is used to capture the bias near the peak summit region. Invalid if there is no control data
1000

The large nearby region in basepairs to calculate dynamic lambda. This is used to capture the surround bias
10000

Build Model
Build the shifting model

Diagnosis report
Do not produce report (faster)

Perform the new peak detection method (futurefdr)
False

Bowtie:
-m 1
-S sam format output

MACS14:
--wig saving wig file

MACS Output Files

Additional output created by MACS (FoxA1_peaks)

Additional Files:

- FoxA1_peaks_model.pdf
- FoxA1_peaks_model.r
- FoxA1_peaks_model.r.log
- FoxA1_peaks_peaks.xls
- FoxA1_peaks_summits.bed

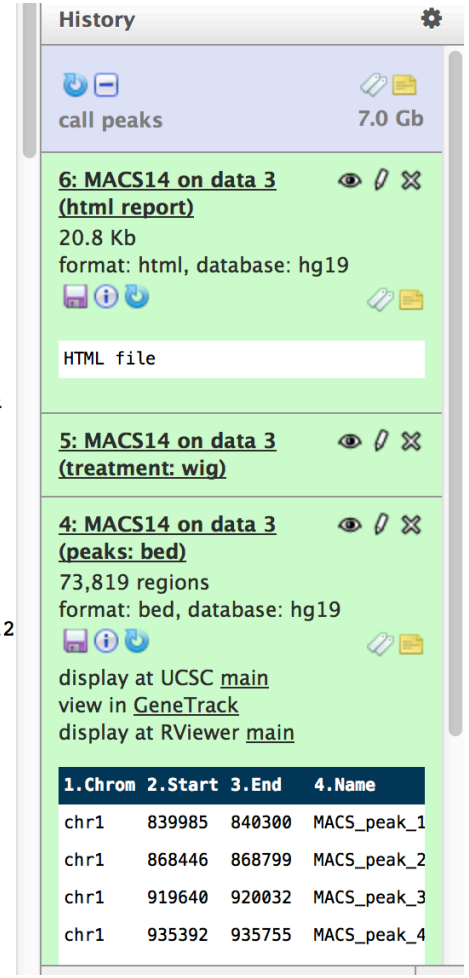
CMD Executed:

```
macs14 -t /u/home/galaxy/galaxy/galaxy-dist/database/files/000/112/dataset_112400.dat
```

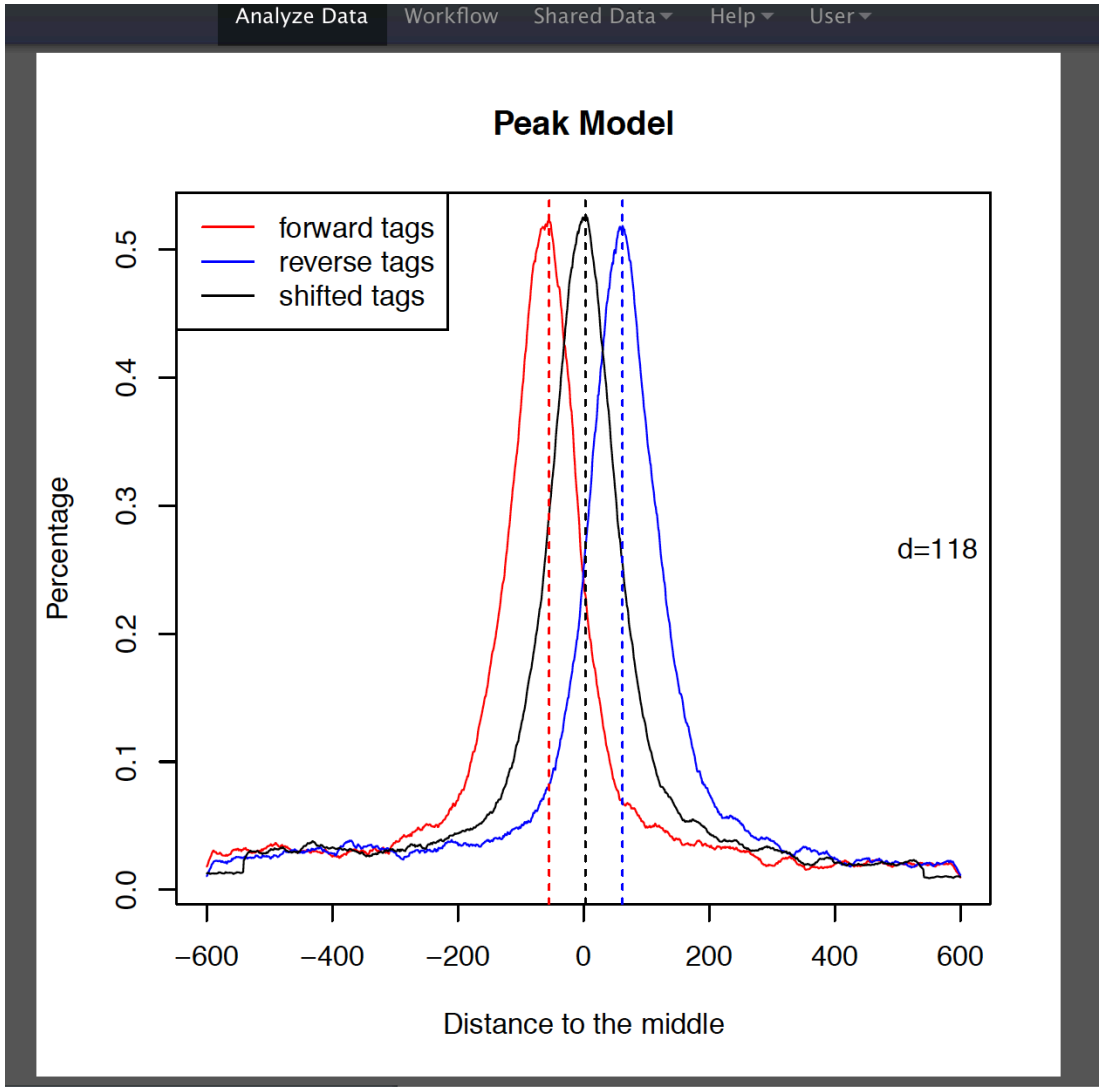
Messages from MACS:

```
INFO  @ Thu, 29 Jan 2015 23:40:00:
# ARGUMENTS LIST:
# name = FoxA1_peaks
# format = SAM
# ChIP-seq file = /u/home/galaxy/galaxy/galaxy-dist/database/files/000/112/dataset_112
# control file = None
# effective genome size = 2.70e+09
# band width = 300
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Large dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 10000 bps

INFO  @ Thu, 29 Jan 2015 23:40:00: #1 read tag files...
INFO  @ Thu, 29 Jan 2015 23:40:00: #1 read treatment tags...
INFO  @ Thu, 29 Jan 2015 23:40:07:  1000000
INFO  @ Thu, 29 Jan 2015 23:40:15:  2000000
INFO  @ Thu, 29 Jan 2015 23:40:22:  3000000
```



MACS Output Files



- ✓ FoxA1_peaks_model.pdf

Red: the percentage of reads from + strand at each pair.

Blue: the percentage of reads from – strand at each pair

d: d is the distance between the summits of read and blue curves

Black: the distribution of reads after shifting them by d/2 towards the 3'nds

MACS Output Files

chr1	839985	840300	MACS_peak_1	59.56
chr1	868446	868799	MACS_peak_2	380.36
chr1	919640	920032	MACS_peak_3	78.49
chr1	935392	935755	MACS_peak_4	76.39
chr1	959953	960300	MACS_peak_5	392.77
chr1	1009086	1009828	MACS_peak_6	130.80
chr1	1051395	1051651	MACS_peak_7	57.79
chr1	1089220	1089605	MACS_peak_8	70.12
chr1	1141650	1141865	MACS_peak_9	104.62
chr1	1172827	1173133	MACS_peak_10	151.72
chr1	1176294	1176669	MACS_peak_11	65.85
chr1	1227247	1227584	MACS_peak_12	66.79
chr1	1342522	1342879	MACS_peak_13	64.35
chr1	1365092	1365389	MACS_peak_14	83.57
chr1	1368550	1368953	MACS_peak_15	56.66
chr1	1440211	1440484	MACS_peak_16	65.79
chr1	1479097	1479413	MACS_peak_17	60.12
chr1	1510126	1510527	MACS_peak_18	107.85
chr1	1550540	1550800	MACS_peak_19	57.86
chr1	1590353	1590763	MACS_peak_20	58.58
chr1	1613625	1613937	MACS_peak_21	154.94
chr1	1677767	1678022	MACS_peak_22	100.56
chr1	1714457	1714739	MACS_peak_23	224.42
chr1	1784949	1785573	MACS_peak_24	155.43

Col1: chromosome
Col2: peak start position
Col3: peak end position
Col4: peak name
Col5: -10*log10(pvalue)

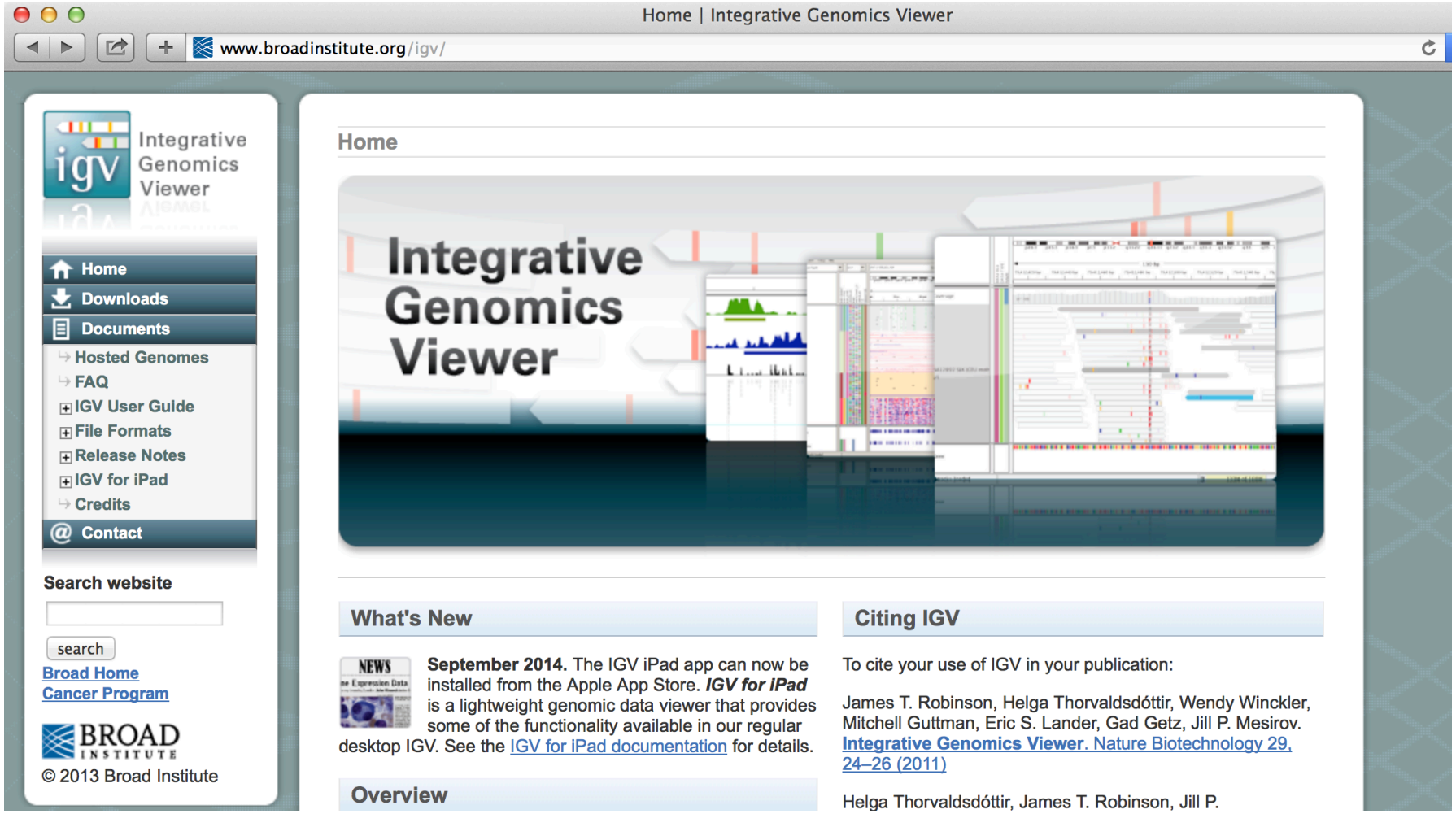
MACS Output Files

4	# format = SAM							
5	# ChIP-seq file = /u/home/galaxy/galaxy/galaxy-dist/database/files/000/112/dataset_112400.dat							
6	# control file = None							
7	# effective genome size = 2.70e+09							
8	# band width = 300							
9	# model fold = 10,30							
10	# pvalue cutoff = 1.00e-05							
11	# Large dataset will be scaled towards smaller dataset.							
12	# Range for calculating regional lambda is: 10000 bps							
13								
14	# tag size is determined as 25 bps							
15	# total tags in treatment: 12468731							
16	# tags after filtering in treatment: 10876598							
17	# maximum duplicate tags at the same position in treatment = 1							
18	# Redundant rate in treatment: 0.13							
19	# d = 118							
20	chr	start	end	length	summit	tags	#NAME?	fold_enrichment
21	chr1	839986	840300	315	169	9	59.56	12.4
22	chr1	868447	868799	353	190	46	380.36	37.45
23	chr1	919641	920032	392	289	12	78.49	21.04
24	chr1	935393	935755	363	108	13	76.39	15.69
25	chr1	959954	960300	347	214	47	392.77	42.37
26	chr1	1009087	1009828	742	155	24	130.8	18.03
27	chr1	1051396	1051651	256	140	8	57.79	12.4
28	chr1	1089221	1089605	385	269	11	70.12	14.73
29	chr1	1141651	1141865	215	107	11	104.62	23.14
30	chr1	1172828	1173133	306	152	21	151.72	24.4
31	chr1	1176295	1176669	375	101	13	65.85	13.24
32	chr1	1227248	1227584	337	101	10	66.79	12.62
33	chr1	1342523	1342879	357	253	10	64.35	12.62
34	chr1	1365093	1365389	297	180	14	83.57	13.71
35	chr1	1368551	1368953	403	243	14	56.66	8.27

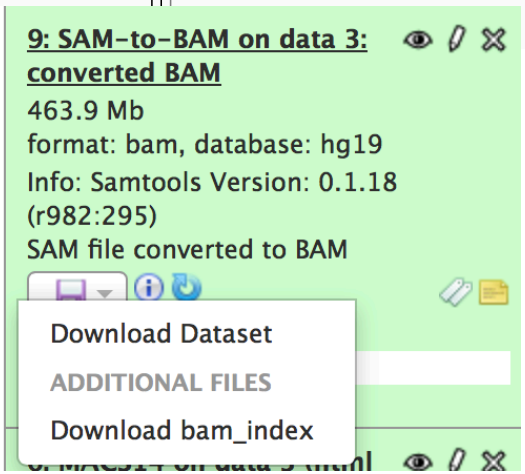
Fold_enrichment:
Compared to the expectation from Poission distribution with local lambda

False discovery rate will be reported if a control sample is used

Integrative Genomics Viewer
http://www.broadinstitute.org/igv



IGV Viewer



Download peaks bed file and bam and bam_index files from history

CEAS

liulab.dfci.harvard.edu/CEAS/ Reader

Cis-regulatory Element Annotation System

User Manual
Install
Download

Summary

We present a tool designed to characterize genome-wide protein-DNA interaction patterns from ChIP-chip and ChIP-Seq of both sharp and broad binding factors. As a stand-alone extension of our web application CEAS (Cis-regulatory Element Annotation System), it provides statistics on ChIP enrichment at important genome features such as specific chromosome, promoters, gene bodies, or exons, and infers genes most likely to be regulated by a binding factor. CEAS also enables biologists to visualize the average ChIP enrichment signals over specific genomic features, allowing continuous and broad ChIP enrichment to be perceived which might be too subtle to detect from ChIP peaks alone.

Introduction

In analysis of cis-regulatory elements using genome-wide ChIP-chip or ChIP-Seq, it is essential to characterize the ChIP signals and identify potential association of ChIP regions with functionally important genomic regions such as gene promoters or exons. Previously, we developed a web server to evaluate GC content and evolutionary conservation of the ChIP regions, conduct sequence motif finding, and map ChIP regions to their nearest genes (Ji, et al., 2006). However, more analytic functions are needed to provide biologists with a more complete perspective. For example, in addition to merely analyzing the identified ChIP regions of a factor, displaying the average ChIP enrichment signal within/near genes helps biologists better visualize the functional loci of factors, especially for broad histone modifications. In addition, biologists often like a visual overview of ChIP peaks' intensity distributions across chromosomes. Nonetheless, such analysis functions often require the ability of manipulate large continuous ChIP enrichment signal files (e.g. WIG files of hundreds of mega bases in size), which are difficult to transfer to a web server. Therefore, as an extension of our current successful web-based CEAS (over 35K analysis queries processed in 2008), we present a stand-alone CEAS extension package with more analysis functions, including drawing average ChIP signal profiles at genes or user-specified loci from a WIG file. This stand-alone CEAS package also provides summary statistics about how the ChIP regions are distributed over important gene features such as promoters, immediate downstream of genes, and exons, and a report on how individual genes are enriched near ChIP regions.

CEAS Overview

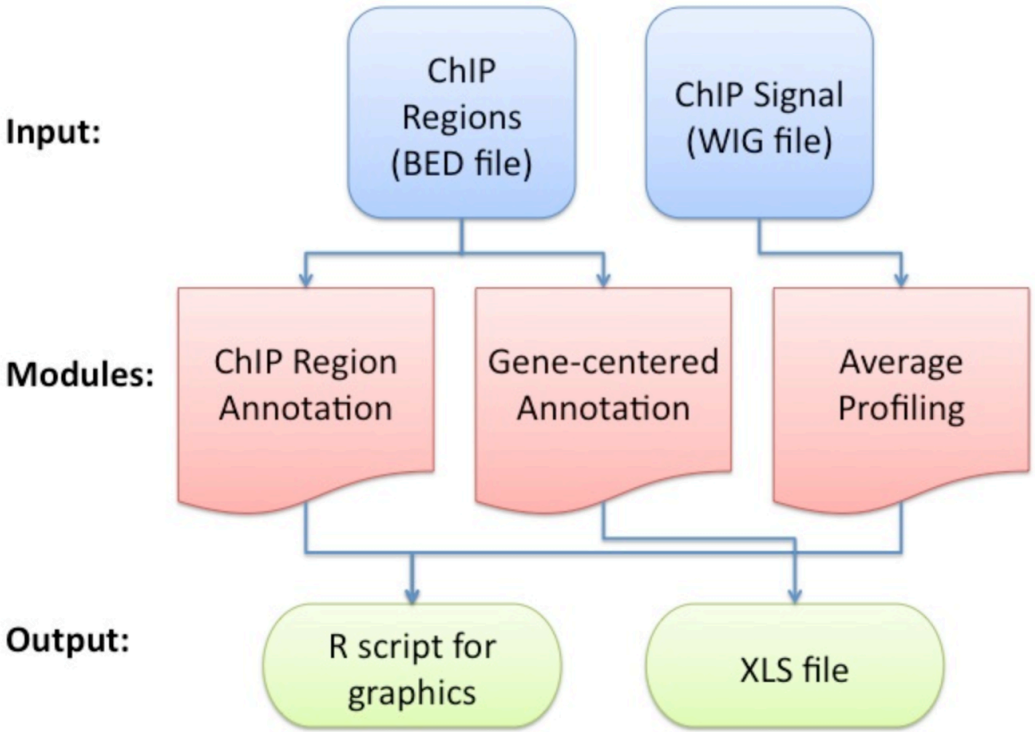


Figure 1 The work-flow of CEAS. A gene annotation table file, a BED file with ChIP regions, and a WIG file with ChIP enrichment signal are required. CEAS consists of three modules: ChIP region annotation, gene-centered annotation, and average profiling within/near important genomic features. As output, CEAS produces an R script of graphical results (or a PDF file if R can be directly called in the same environment as Python) and a tab-delimited with XLS extension of gene-centered annotation.

CEAS Modules

ChIP region annotation
CEAS estimates the relative enrichment level of ChIP regions in each gene feature with respect to the whole genome. For this, it first calculates the percentages of the ChIP regions that reside in the following four categories: (a) promoters, (b) bidirectional promoters, (c) downstreams of a gene, and (d) gene bodies (3'UTRs, 5'UTRs, coding exons, and introns). In addition to these categories, the user can add another extra category of interest such as non-coding regions as an optional input parameter. 'Promoters' correspond to the upstream regions of the transcription start site (TSS) of genes. Three promoter sizes can be specified by the user (1kb, 2kb, and 3kb by default). For instance, if the user set promoter to be the 1kb, 3kb, and 10kb upstream of TSS, CEAS computes the cumulative percentages of ChIP regions that fall in <=1kb, <=3kb, and <=10kb upstream of the TSS of genes, respectively. `Bidirectional promoters' are promoter regions between divergently transcribed genes whose TSS are closer in proximity than user-defined distances (two options, 2.5kb and 5kb by default). `Downstreams' refer to the regions immediately downstream of genes, spanning up to the same search range as in `promoters' from the transcription termination site (TTS). `Gene bodies' are further categorized into UTR regions (3' and 5' UTRs), coding exons and introns. After the percentages of ChIP regions respective categories, P-values for the significance of the relative enrichment with respect to the background are calculated using one-sided binomial test.
As a final summary of ChIP region annotation, CEAS draws a pie chart of how ChIP regions spread over the categories. If ChIP regions do not fall into any of the categories, they are considered to be `distal intergenic'.

Gene-centered annotation
Identifying genes associated with ChIP regions is important to infer the direct regulatory gene targets of the binding factor of interest. CEAS provides the distances to the centers of the nearest ChIP regions upstream and downstream of every RefSeq gene's TSS, allowing biologists to determine the potential target genes of the binding factor based on the distances. Moreover, in case that a broad ChIP peak covers the whole or part of a gene body, it is useful to know how much of the gene, including its promoter or downstream, is occupied by the ChIP region in addition to how far the TSS of the gene is from the ChIP region center. To this end, CEAS divides every gene into three equal fractions and, for each fraction, calculates the percentage of the area covered by ChIP regions. It also estimates the percentages of the promoter and downstream of the gene (3kb upstream of TSS and downstream of TTS by default) that are covered by ChIP regions. The results are saved as a tab-delimited text file with XLS extension for easy Excel visualization, which contains a row of annotations for every RefSeq gene.

Average signal profiling within/near important genomic features
Since ChIP region and gene-centered annotation operate on discrete ChIP regions identified by a peak-calling algorithm, some subtle binding patterns may fail to be captured, depending on the cut-off of peak calling. Therefore, CEAS displays continuous ChIP enrichment signal within/around important gene features to help biologists visualize the average binding patterns. It draws the average signals around TSS and TTS in a user-defined range (+/- 3kb from TSS and TTS by default). In addition, CEAS gives average signals on meta-gene, meta-concatenated-exon, meta-concatenated-intron, meta-exons and meta-intron, where the prefix `meta' indicates that every element (e.g every gene) is normalized residing within the above categories are obtained, they are compared against the genome background percentages for those to have the same length (e.g., 3kb for meta-gene). The difference between meta-concatenated-exon and meta-exons is that the first concatenates all exons of a gene before calculating the average gene (like a meta-cDNA) profile, whereas the latter calculates the average exon profile of all exons. These plots allow biologists to gain insight on how ChIP enrichment varies over the gene body (or exons and introns). CEAS provides an additional function of drawing the average ChIP signals of multiple sub-groups of genes (e.g. up vs. down genes), allowing the user to compare between the gene groups. In addition, we provide a separate script, named `sitepro', in our CEAS package, which draws the average signal (from a WIG) in a given list of sites of interest (from a BED). This script enables biologists to visualize the average signals in any arbitrary regions (e.g. transcription factor binding sites) in addition to the pre-defined genomic regions.

Galaxy CEAS Package

- ✓ CEAS: Enrichment on chromosome and annotation Annotate the given intervals and scores with genome features such as gene body
- ✓ SitePro: Aggregation plot tool for signal profiling Draw the score profile near a given interval
- ✓ GCA: Gene centered annotation Find the nearest interval in the given intervals set for every annotated coding gene
- ✓ peak2gene: Peak Center Annotation Input a peak file, and It will search each peak on UCSC GeneTable to get the refGenes near the peak center. (for ce4, ce6, dm2, dm3, hg18, hg19, mm8, mm9, and sacCer3)

CEAS Input Files

- ✓ Gene annotation table file (sqlite3)

  Available gene annotation tables in galaxy:

  Worm: ce4 and ce6

  Fly: dm2 and dm3

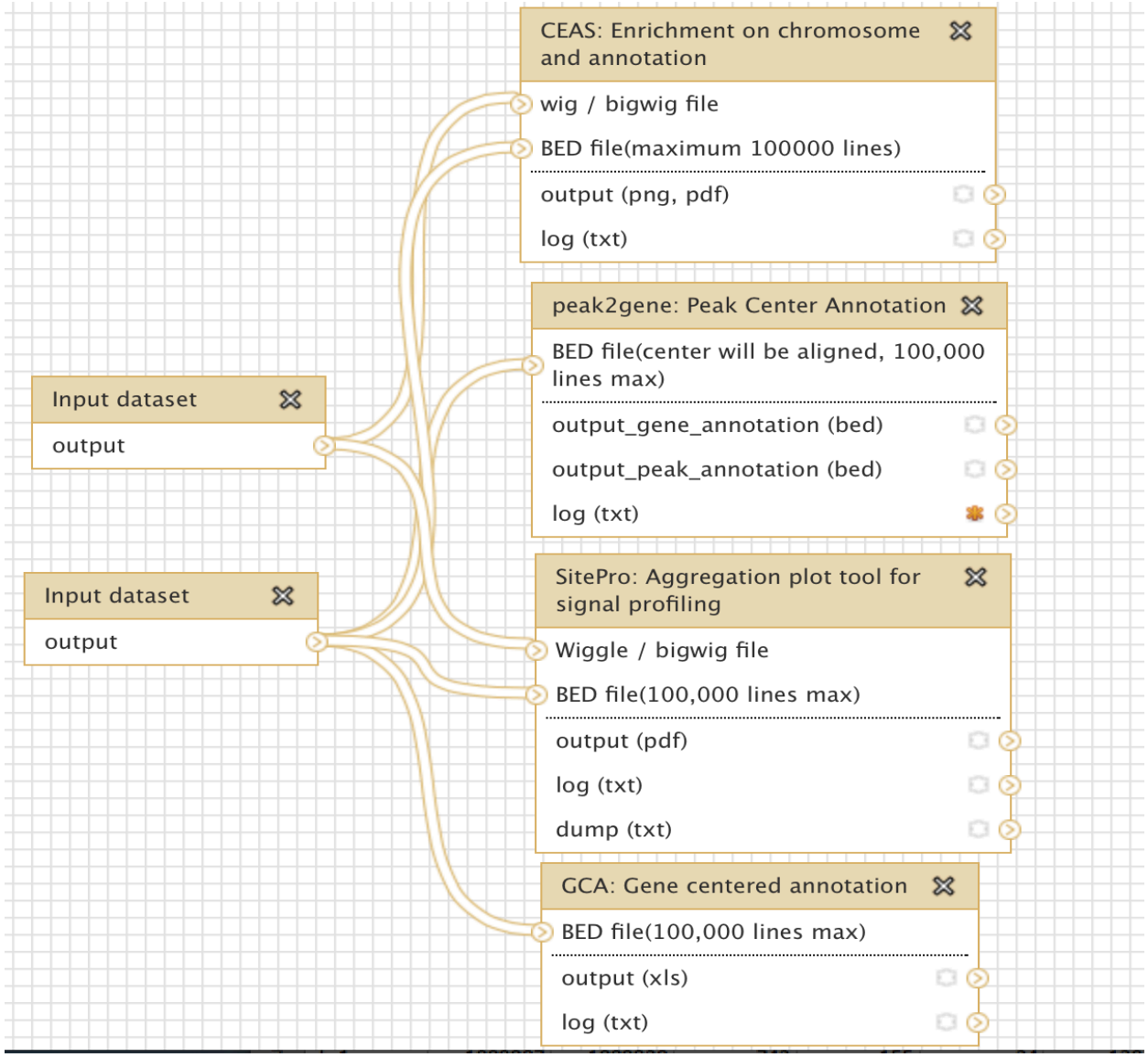
  Mouse: mm8 and mm9

  Human: hg18 and hg19

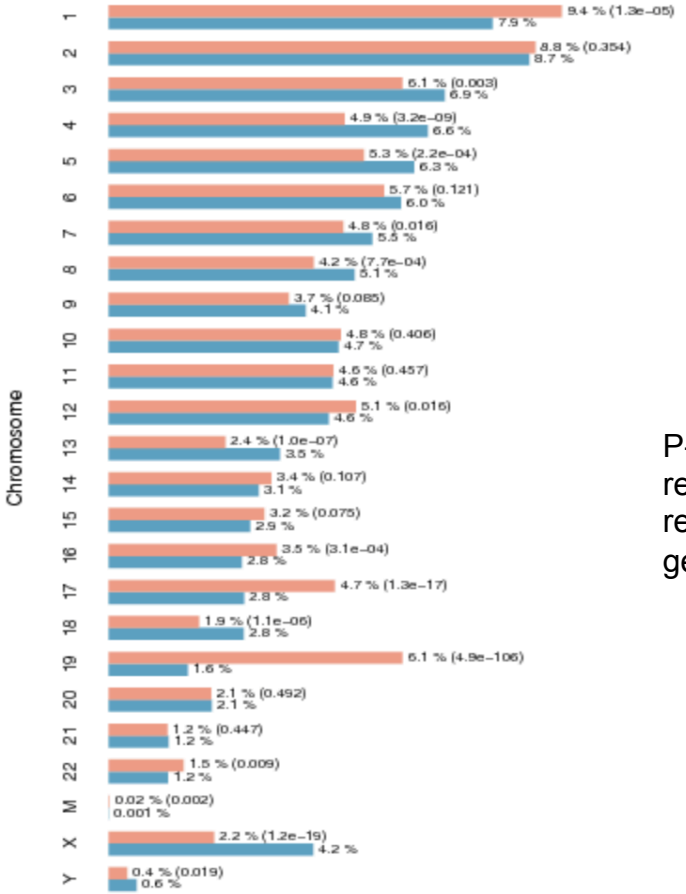
- ✓ BED file with ChIP regions (TXT)

  a bed file contains chromsome, start and end locations of ChIP regions (chr1 779600 780954)

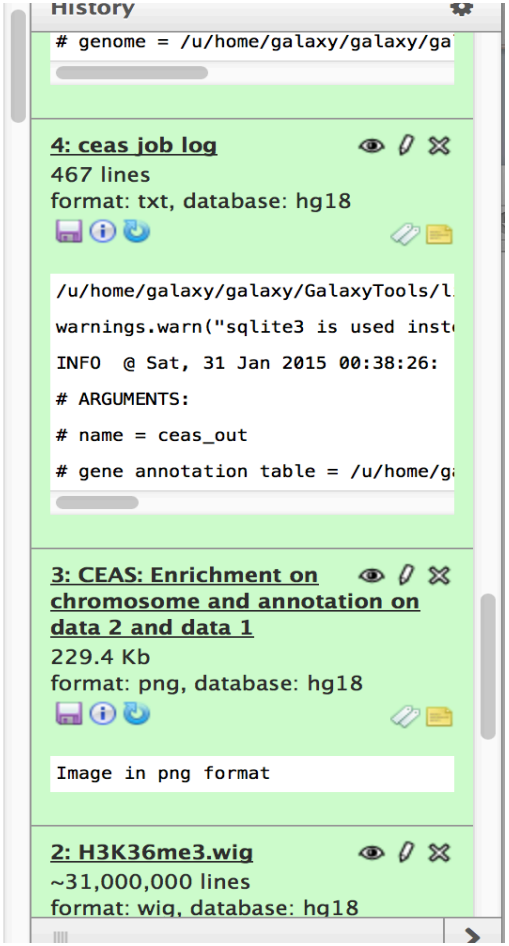
- ✓ WIG file with ChiP enrichment signal (TXT)

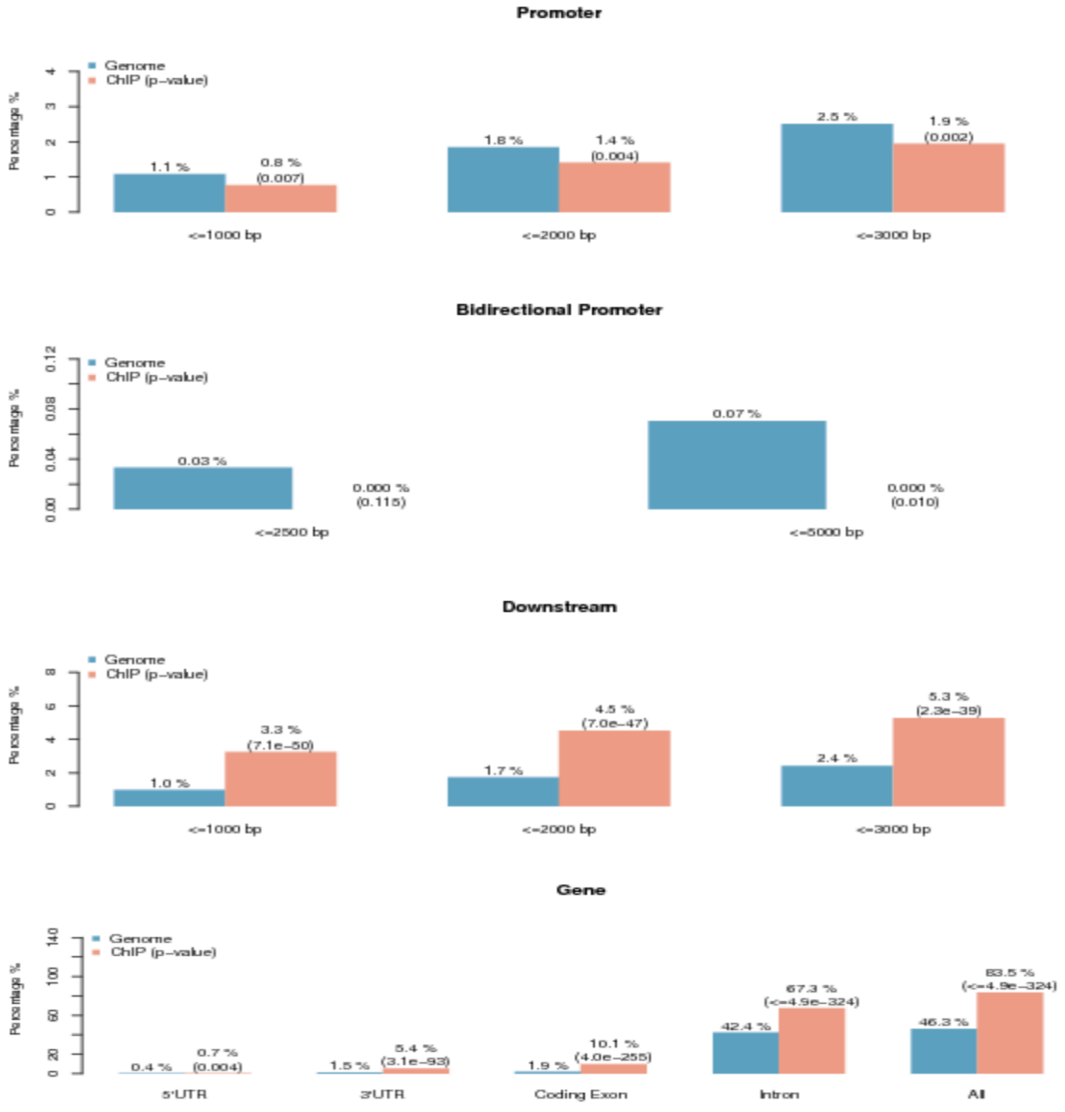

CEAS: Enrichment on chromosome and annotation
wig / bigwig file
BED file(maximum 100000 lines)
output (png, pdf)
log (txt)
peak2gene: Peak Center Annotation
BED file(center will be aligned, 100,000 lines max)
output_gene_annotation (bed)
output_peak_annotation (bed)
log (txt)
Input dataset
output
Input dataset
output
SitePro: Aggregation plot tool for signal profiling
Wiggle / bigwig file
BED file(100,000 lines max)
output (pdf)
log (txt)
dump (txt)
GCA: Gene centered annotation
BED file(100,000 lines max)
output (xls)
log (txt)

CEAS Output Files



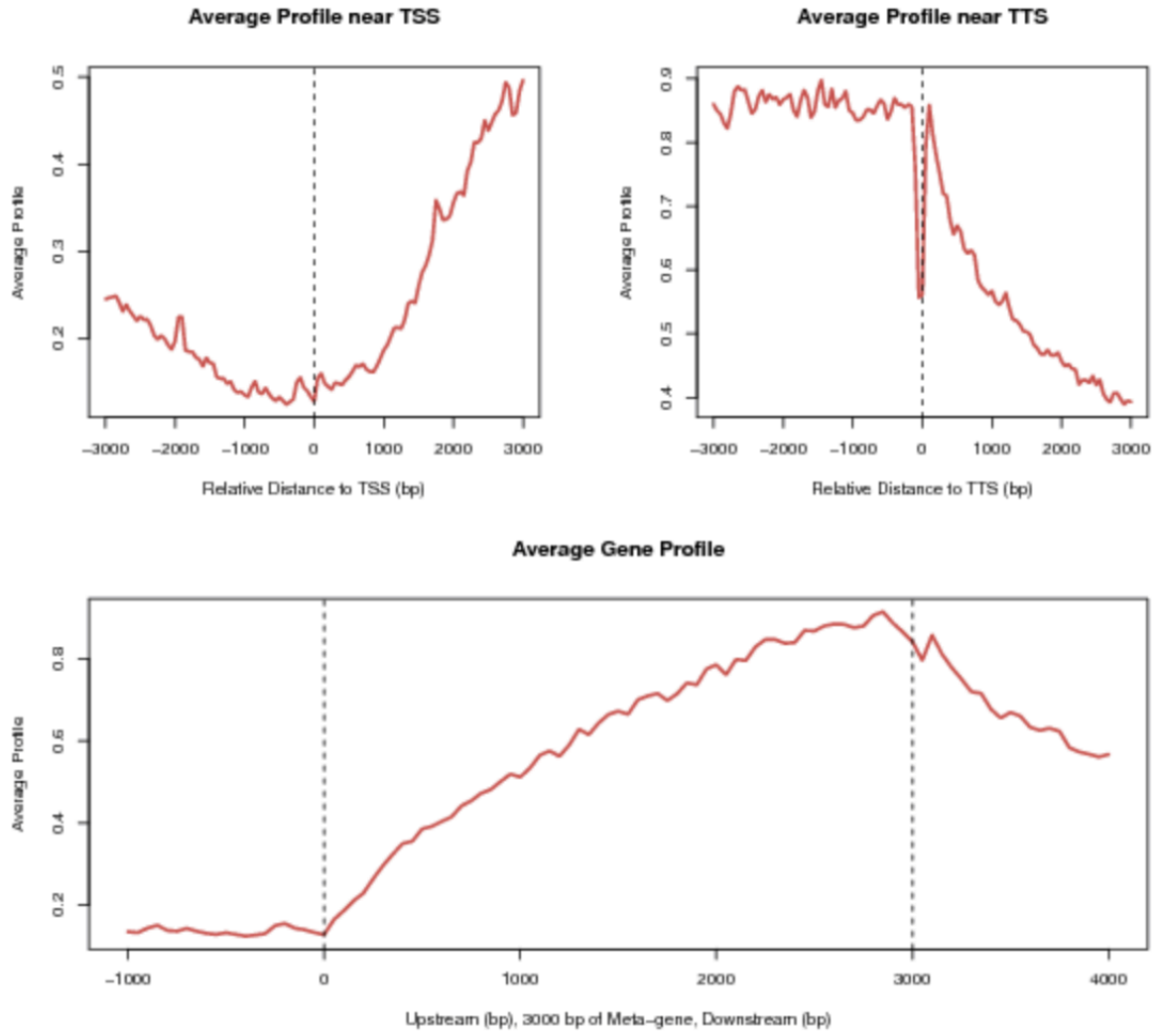
P-value: significance of the relative enrichment of ChIP regions compared to the genome background





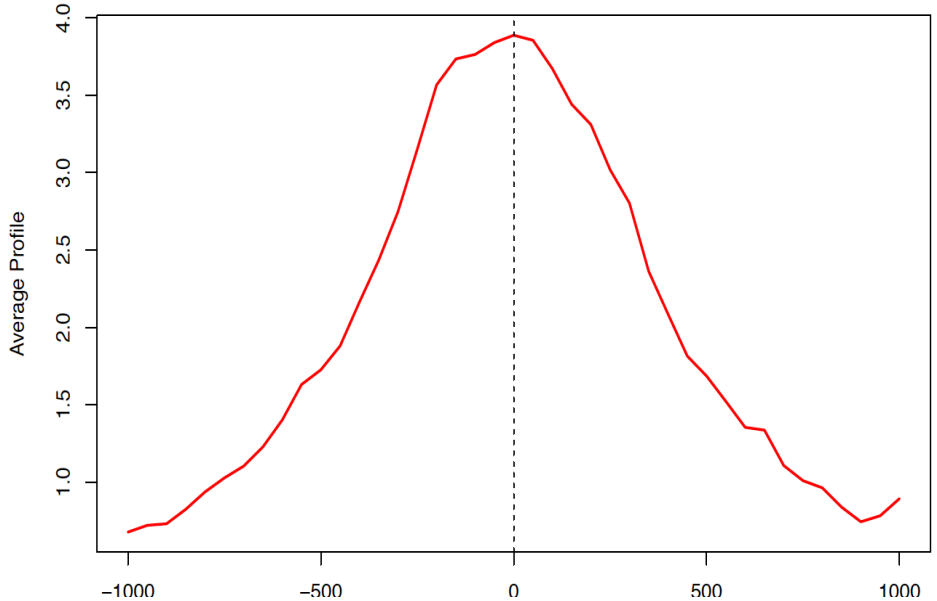
Relative enrichments of ChIP regions in promoters, downstreams of genes, and gene bodies (intron, exon, 5'UTR, 3'UTR)

Bidirectional promoters are promoter regions between divergently transcribed genes whose TSS are closer than defined distances



Top: average ChIP enrichment signals near transcription starting site (left) and transcription termination site (right)
Bottom: average ChIP signals on the meta-gene of 3kb. H3K36me3 enriches gene bodies and increases towards the 3’ end.

SitePro Output




```
chr1    779600  780954
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,3.0,5.0,2.0,1.0,0.0,0.0,3.0,1.0,1.0,1.0,
1.0,0.0,5.0,13.0,2.0,4.0,8.0,4.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
chr1    874824  876507
0.0,0.0,0.0,0.0,0.0,4.0,2.0,2.0,2.0,1.0,2.0,5.0,1.0,1.0,0.0,5.0,3.0,2.0,6.0,6.0,9.0,
4.0,12.0,18.0,4.0,9.0,12.0,2.0,6.0,5.0,3.0,1.0,1.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0,
chr1    1147745 1148979
1.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,2.0,3.0,1.0,1.0,0.0,2.0,1.0,3.0,10.0,6.0,14.0,21
.0,5.0,3.0,4.0,3.0,3.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
chr1    1576270 1576887
5.0,3.0,2.0,2.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,2.0,1.0,1.0,6.0,10.0,7.0
,6.0,2.0,0.0,3.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
chr1    2325039 2326135
2.0,6.0,2.0,0.0,1.0,1.0,1.0,0.0,0.0,0.0,2.0,7.0,0.0,1.0,4.0,4.0,1.0,1.0,1.0,2.0,8.0,
11.0,3.0,0.0,1.0,2.0,1.0,1.0,4.0,3.0,1.0,0.0,0.0,0.0,0.0,1.0,1.0,1.0,1.0,3.0,0.0,
chr1    2429436 2430692
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,3.0,2.0,1.0,4.0,16.0,3.0
,0.0,1.0,1.0,0.0,1.0,2.0,1.0,3.0,3.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
chr1    2965061 2965772
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
4.0,4.0,2.0,1.0,3.0,3.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
chr1    7979109 7980053
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,3.0,1.0,2.0,2.0,4.0,
4.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
chr1    7986682 7987747
0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,3.0,2.0,2.0,4.0,5.0,1.0,7.0,14.0,4.0,0.0
,1.0,2.0,2.0,4.0,2.0,0.0,0.0,1.0,1.0,0.0,0.0,1.0,0.0,3.0,4.0,2.0,2.0,3.0,0.0,0.0,0.0,0.0,
```

SitePro measures average signal (defined in the wig file) for the regions defined in the input file (bed file) and then draws the average signal in any arbitrary regions (e.g transcription binding sites)

peak2gene: peak annotation

```
# Argument List:
# Name = output
# peak file = /u/home/galaxy/galaxy/galaxy-dist/database/files/000/112/dataset_112435.dat
# gene pos to peak = up,down
# distance = 30000 bp
# genome = /u/home/galaxy/galaxy/galaxy-dist/tool-data/ceaslib/GeneTable/hg18
# Output symbol as gene name = True
```

#chrom	pStart	pEnd	pName	pScore	NA	gene
chr1	100242425	100243660	NA	0	0	
chr1	100443471	100445867	NA	0	0	LRRC39
chr1	100474727	100476120	NA	0	0	RTCD1|DBT|RTCD1
chr1	100621559	100623180	NA	0	0	
chr1	101238173	101239546	NA	0	0	DPH5|DPH5|DPH5
chr1	10147067	10149295	NA	0	0	
chr1	101476462	101479482	NA	0	0	S1PR1
chr1	10161585	10163471	NA	0	0	
chr1	10343478	10344980	NA	0	0	
chr1	10356265	10358380	NA	0	0	PGD
chr1	10542049	10542666	NA	0	0	
chr1	108026938	108027851	NA	0	0	VAV3
chr1	108486216	108488640	NA	0	0	
chr1	108505151	108505968	NA	0	0	
chr1	109314652	109315850	NA	0	0	CLCC1|CLCC1
chr1	109354595	109356299	NA	0	0	
chr1	109839864	109841826	NA	0	0	CYB561D1|CYB561D1|CYB561D1|CYB561D1
chr1	110340288	110341491	NA	0	0	AHCYL1
chr1	11047508	11048584	NA	0	0	SRM|MASP2|MASP2
chr1	11063147	11065911	NA	0	0	EXOSC10|EXOSC10|SRM

Nearest genes near the peak

Gene centered annotation (GCA)

#name	chr	txStart	txEnd	strand	dist u TSS	dist d TSS	dist u TTS	dist d TTS	3000bp u TSS	3000bp d TSS	1/3 gene	2/3 gene	3/3 gene	3000bp d TTS	exons
NR_024540	chr1	4224	19233	-	761044	NA	776053	NA	0	0	0	0	0	0	0
NR_028269	chr1	4224	7502	-	772775	NA	776053	NA	0	0	0	0	0	0	0
NR_026820	chr1	24474	25944	-	754333	NA	755803	NA	0	0	0	0	0	0	0
NR_026818	chr1	24474	25944	-	754333	NA	755803	NA	0	0	0	0	0	0	0
NR_026822	chr1	24474	25944	-	754333	NA	755803	NA	0	0	0	0	0	0	0
NM_001005484	chr1	58953	59871	+	NA	721324	NA	720406	0	0	0	0	0	0	0
NR_028325	chr1	313754	318443	+	NA	466523	NA	461834	0	0	0	0	0	0	0
NR_028327	chr1	313754	318443	+	NA	466523	NA	461834	0	0	0	0	0	0	0
NR_028322	chr1	313754	318443	+	NA	466523	NA	461834	0	0	0	0	0	0	0
NM_001005277	chr1	357521	358460	+	NA	422756	NA	421817	0	0	0	0	0	0	0
NM_001005224	chr1	357521	358460	+	NA	422756	NA	421817	0	0	0	0	0	0	0
NM_001005221	chr1	357521	358460	+	NA	422756	NA	421817	0	0	0	0	0	0	0

Field	Description
chr	Chromosome of a RefSeq gene
txStart	Transcription starting site (TSS) of a RefSeq gene
txEnd	Transcription terminating site (TTS) of a RefSeq gene
strand	Strand of a RefSeq Gene
dist u txStart	Distance to the nearest ChIP region (center) upstream of txStart (bp)
dist d txStart	Distance to the nearest ChIP region (center) downstream of txStart (bp)
dist u txEnd	Distance to the nearest ChIP region (center) upstream of txEnd (bp)
dist d txEnd	Distance to the nearest ChIP region (center) downstream of txEnd (bp)
3kb u txStart	Occupancy rate of ChIP regions in 3kb upstream of txStart (0.0 - 1.0)
3kb d txStart	Occupancy rate of ChIP regions in 3kb downstream of txStart (0.0 - 1.0)
1/3 gene	Occupancy rate of ChIP regions in the 1st third of a gene (0.0 - 1.0)
2/3 gene	Occupancy rate of ChIP regions in the 2nd third of a gene (0.0 - 1.0)
3/3 gene	Occupancy rate of ChIP regions in the 3rd third of a gene (0.0 - 1.0)
3kb d txEnd	Occupancy rate of ChIP regions in 3kb downstream of txEnd (0.0 - 1.0)
exons	Occupancy rate of ChIP regions in the exons (0.0-1.0)

Other Tools in CEAS Package

- ✓ Conservation Plot

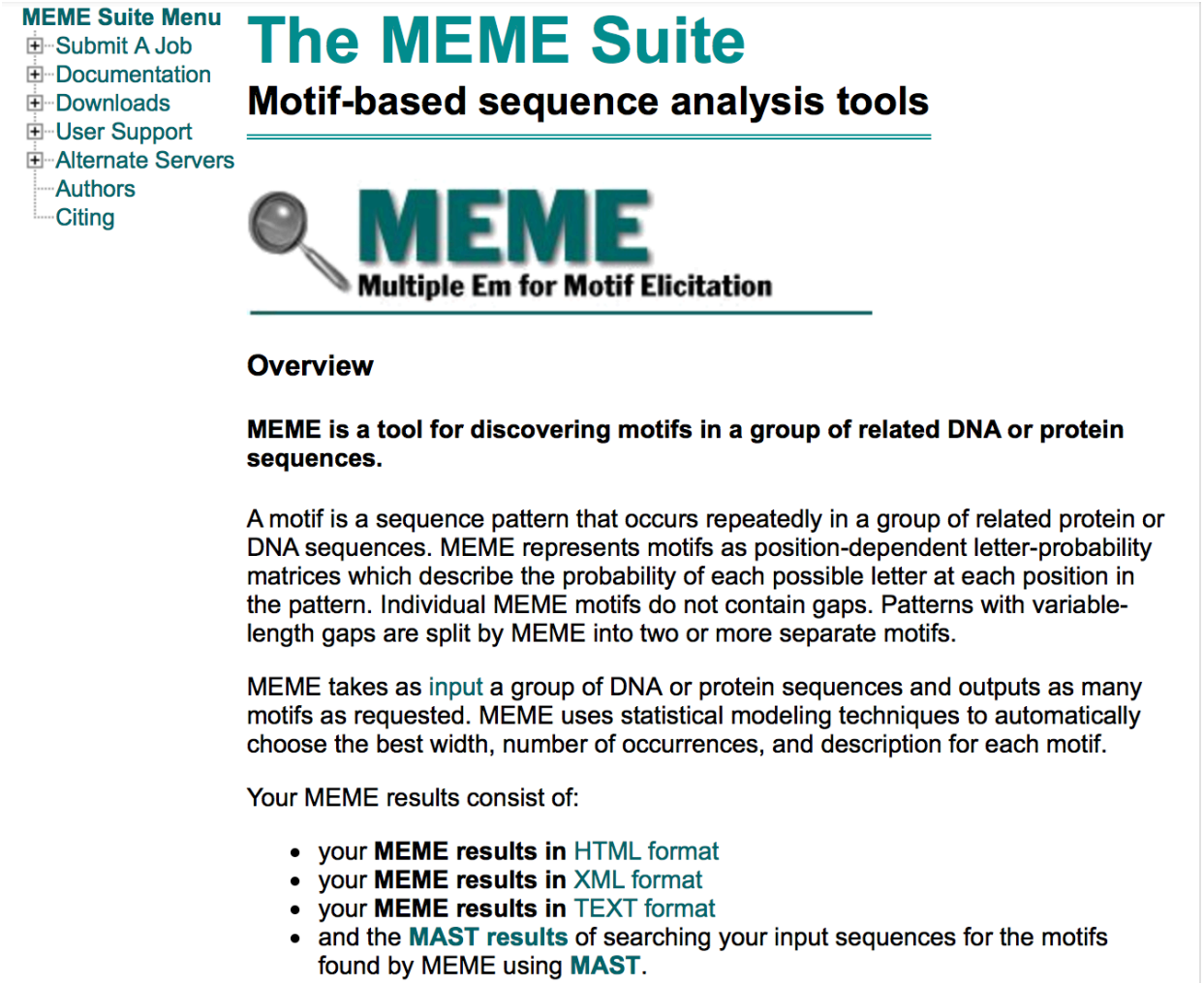
  use UCSC PhastCons conservation scores to produce a figure showing the average conservation score profiles around the peak centers. Useful as an indicator of data quality

- ✓ Heatmap

  extract the signals centered at every given genomic location, perform k-means clustering and draw a heatmap

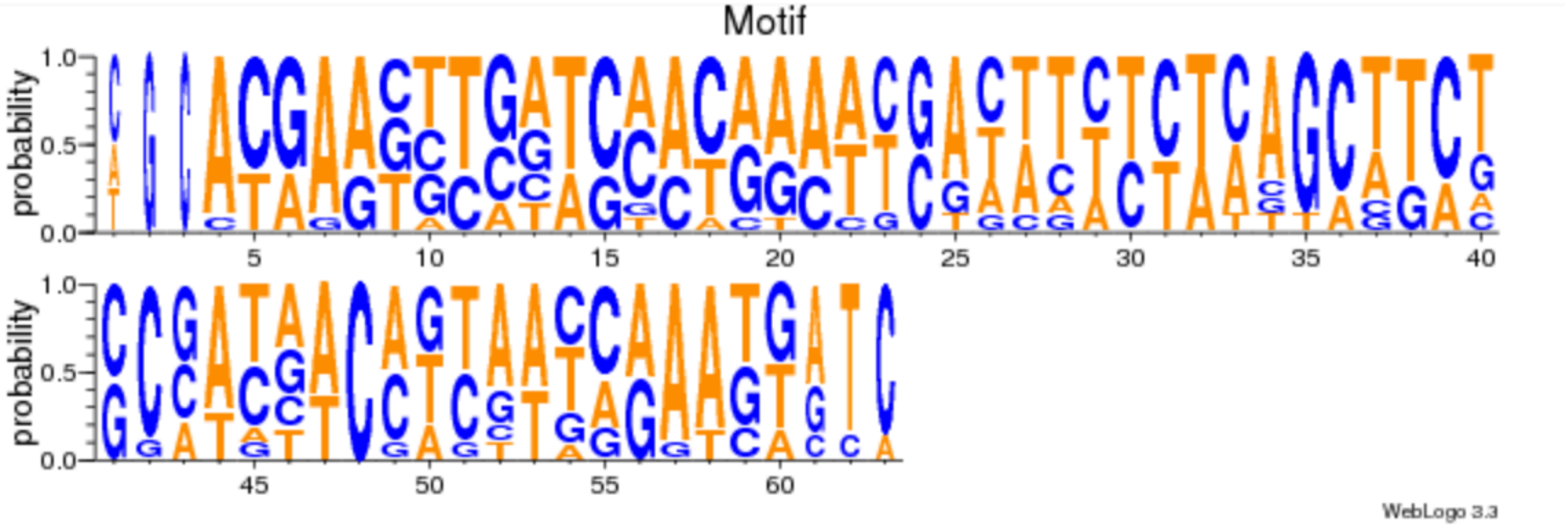
Downstream Analysis Tools

- ✓ Motif Tools – MEME and FIMO



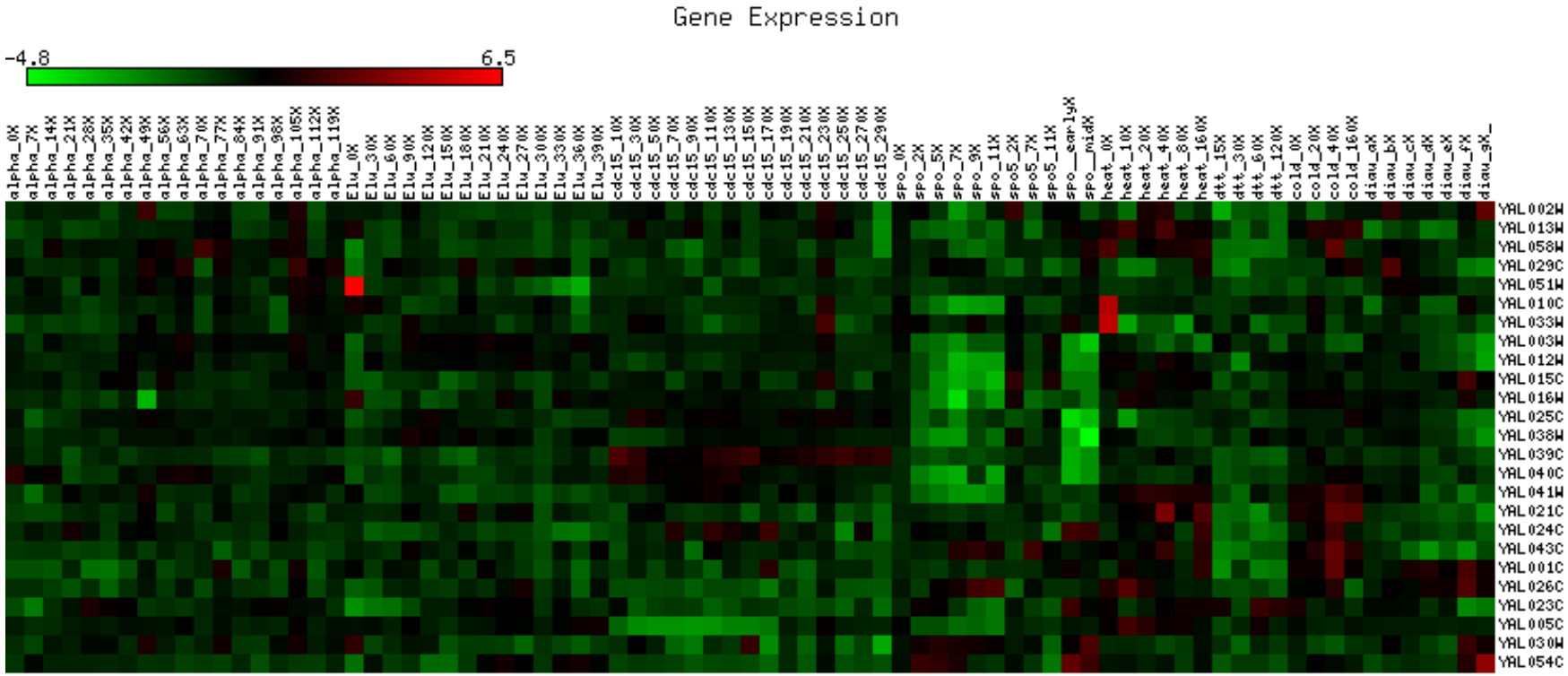
Downstream Analysis Tools

- ✓ Motif tools – sequence logo



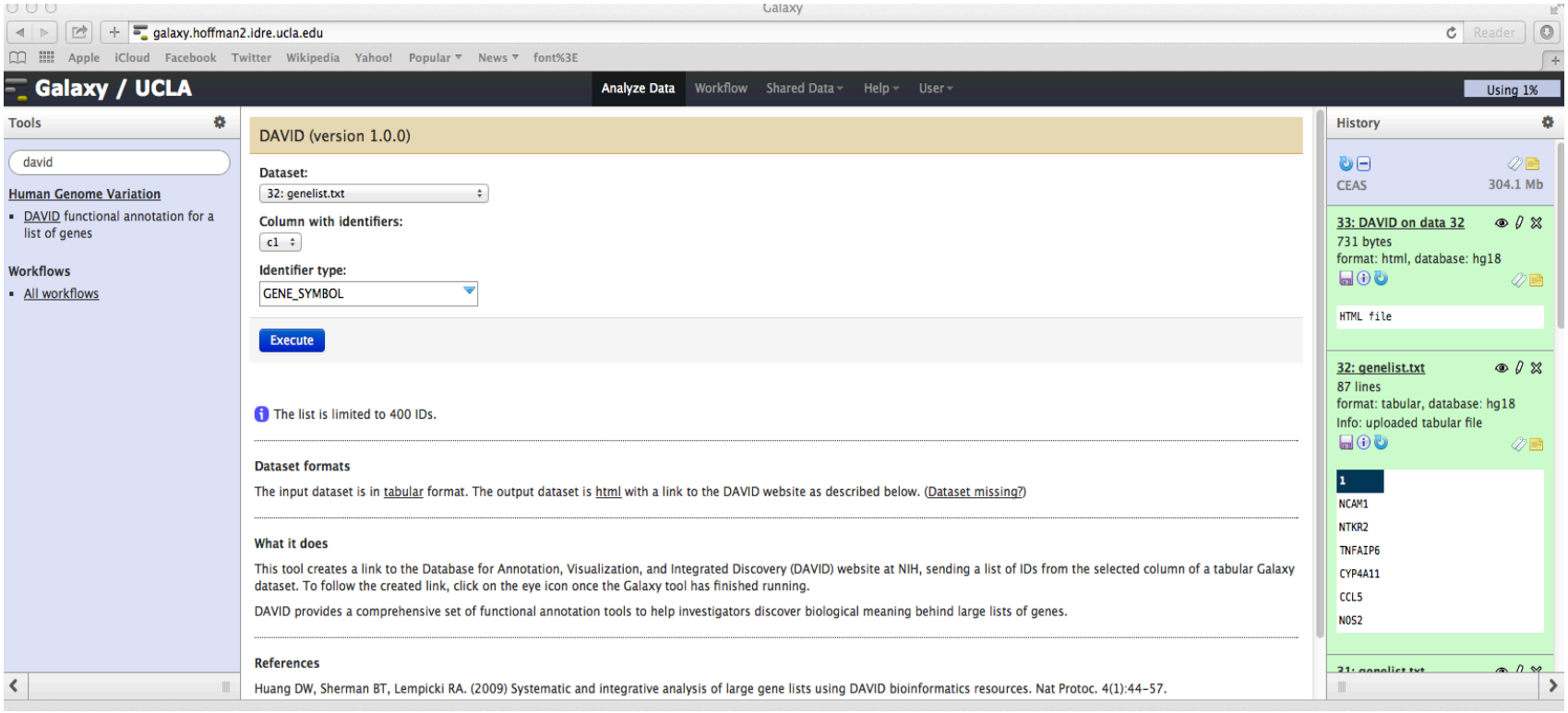
Downstream Analysis Tools

- ✓ Kmeans cluster and Heatmap

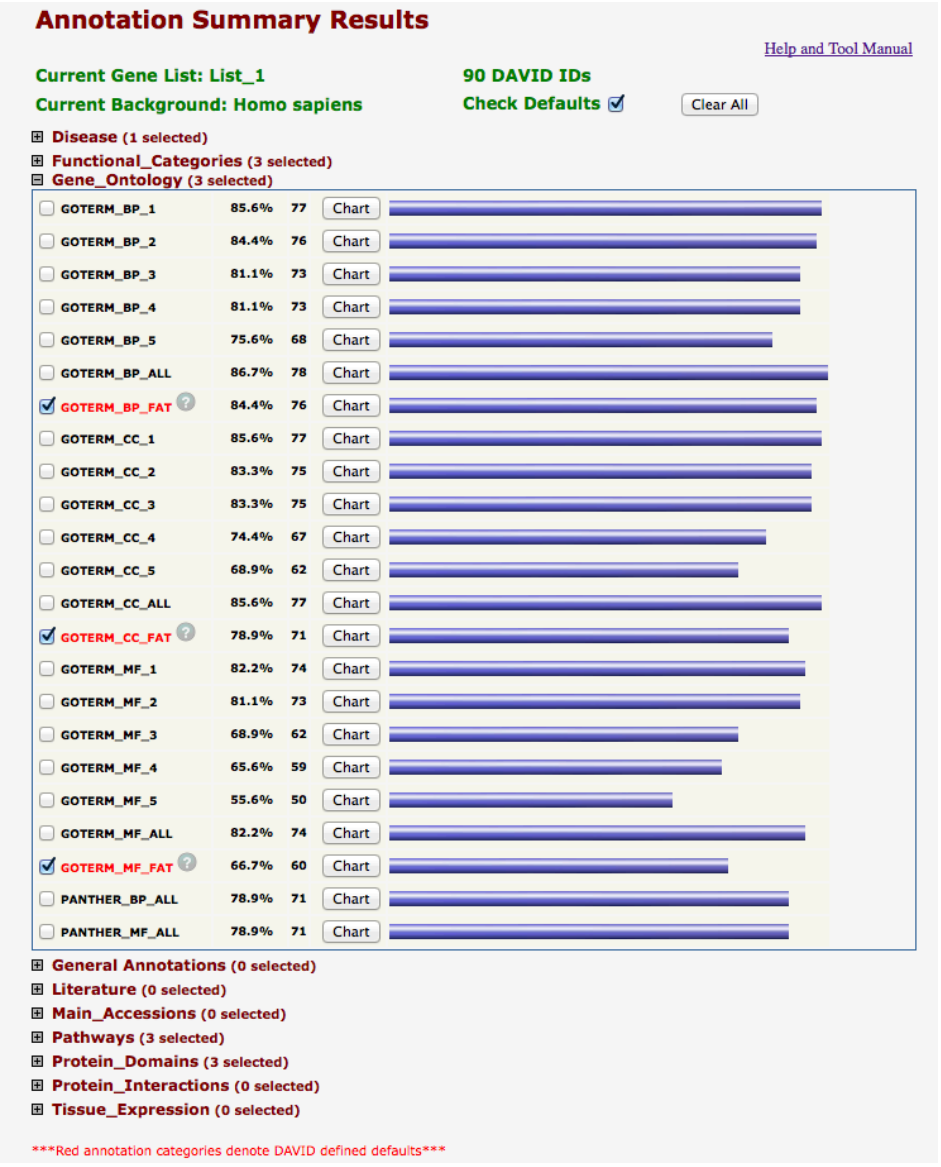


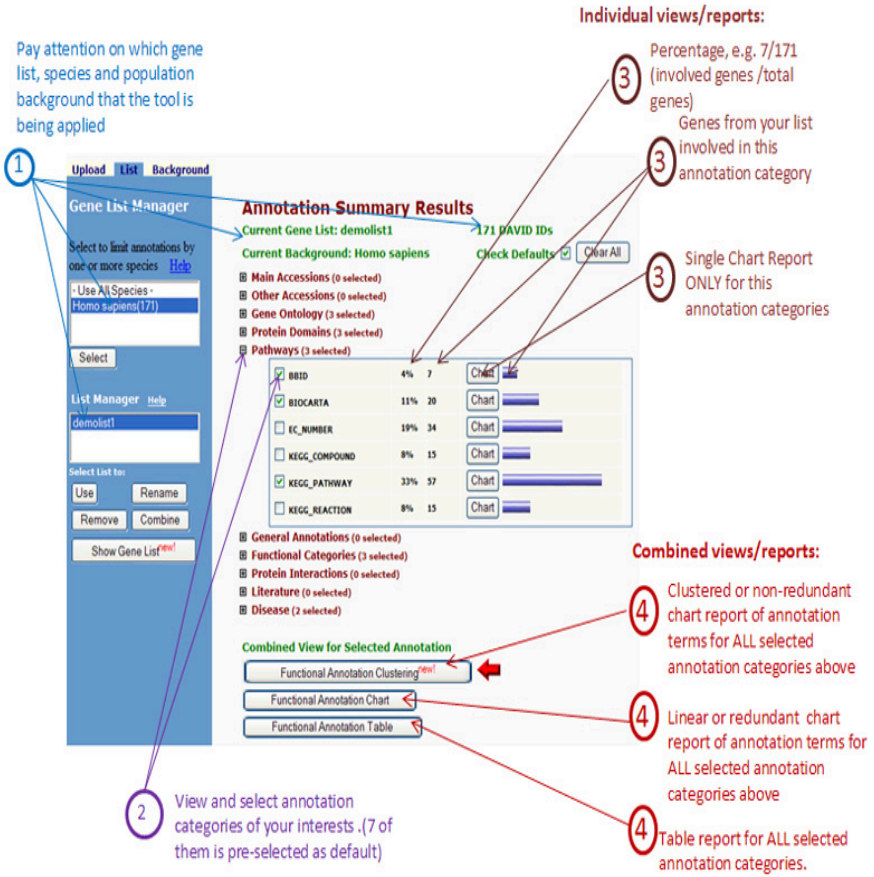
Downstream Analysis Tools

- ✓ Gene Ontology Analysis – Database for Annotation, Visualization, and Integrated Discovery (DAVID)



Downstream Analysis Tools





Downstream Analysis Tools

Functional Annotation Table

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
2797 DAVID IDs

946 record(s) Download File

AKAP7	A kinase (PRKA) anchor protein 7	Related Genes	Mus musculus
GOTERM_BP_FAT	cell surface receptor linked signal transduction, enzyme linked receptor protein signaling pathway, transmembrane receptor protein serine/threonine kinase signaling pathway, protein localization,		
GOTERM_CC_FAT	intrinsic to membrane, anchored to membrane,		
GOTERM_MF_FAT	enzyme binding, kinase binding, protein kinase binding, protein kinase A binding,		
INTERPRO	Protein kinase A anchor protein, nuclear localisation signal domain, Protein kinase A anchor protein, RI-RII subunit-binding domain,		
SP_PIR_KEYWORDS	kinase, lipoprotein, membrane, myristate, palmitate,		
UP_SEQ_FEATURE	chain:A-kinase anchor protein 7, lipid moiety-binding region:N-myristoyl glycine, lipid moiety-binding region:S-palmitoyl cysteine, region of interest:Required for membrane localization, region of interest:RII-binding,		
AKAP7	**A kinase (PRKA) anchor protein 7**	**Related Genes**	**Homo sapiens**
GOTERM_BP_FAT	ion transport, intracellular signaling cascade, protein localization,		
GOTERM_CC_FAT	cytosol, plasma membrane, apical plasma membrane, lateral plasma membrane, intrinsic to membrane, anchored to membrane, plasma membrane part, apical part of cell,		
GOTERM_MF_FAT	enzyme binding, kinase binding, protein kinase binding, protein kinase A binding,		
INTERPRO	Protein kinase A anchor protein, nuclear localisation signal domain, Protein kinase A anchor protein, RI-RII subunit-binding domain,		
SP_PIR_KEYWORDS	alternative splicing, cell membrane, complete proteome, cytoplasm, kinase, lipoprotein, membrane, myristate, palmitate, polymorphism,		
UP_SEQ_FEATURE	chain:A-kinase anchor protein 7 isoform gamma, chain:A-kinase anchor protein 7 isoforms alpha and beta, lipid moiety-binding region:N-myristoyl glycine, lipid moiety-binding region:S-palmitoyl cysteine, region of interest:Required for apical membrane localization, region of interest:Required for membrane localization, region of interest:RII-binding, sequence variant, splice variant,		
AKAP7	**A kinase (PRKA) anchor protein 7**	**Related Genes**	**Bos taurus**
GOTERM_MF_FAT	enzyme binding, kinase binding, protein kinase binding,		
INTERPRO	Protein kinase A anchor protein, nuclear localisation signal domain,		
AKAP7	**A kinase (PRKA) anchor protein 7**	**Related Genes**	**Rattus norvegicus**
GOTERM_BP_FAT	protein localization, regulation of protein kinase cascade, regulation of protein kinase A signaling cascade,		
GOTERM_CC_FAT	cell fraction, membrane fraction, soluble fraction, insoluble fraction, endoplasmic reticulum, cytosol, plasma membrane, cytoplasmic membrane-bounded vesicle, apical plasma membrane, lateral plasma membrane, sarcoplasm, sarcoplasmic reticulum, transport vesicle, T-tubule, cytoplasmic vesicle, vesicle, membrane-bounded vesicle, sarcolemma, plasma membrane part, apical part of cell, exocytic vesicle,		
GOTERM_MF_FAT	nucleotide binding, nucleoside binding, purine nucleoside binding, structural molecule activity, protein C-terminus binding, AMP binding, purine nucleotide binding, enzyme binding, kinase binding, protein kinase binding, protein domain specific binding, adenyl nucleotide binding, ribonucleotide binding, purine ribonucleotide binding, adenyl ribonucleotide binding, protein complex scaffold, protein kinase A regulatory subunit binding, protein kinase A binding,		
INTERPRO	Protein kinase A anchor protein, nuclear localisation signal domain, Protein kinase A anchor protein, RI-RII subunit-binding domain,		
SP_PIR_KEYWORDS	kinase,		
ADAM23	**ADAM metallopeptidase domain 23**	**Related Genes**	**Gallus gallus**
GOTERM_BP_FAT	proteolysis, cell surface receptor linked signal transduction, integrin-mediated signaling pathway,		
GOTERM_CC_FAT	integral to membrane, intrinsic to membrane,		
GOTERM_MF_FAT	endopeptidase activity, metalloendopeptidase activity, peptidase activity, metallopeptidase activity, zinc ion binding, ion binding, cation binding, metal ion binding, transition metal ion binding, peptidase activity, acting on L-amino acid peptides,		
INTERPRO	EGF-like, type 3, Peptidase M12B, ADAM/reprolysin, Blood coagulation inhibitor, Disintegrin, Peptidase M12B, propeptide, EGF-like, ADAM, cysteine-rich, EGF-like region, conserved site, EGF, extracellular, Disintegrin, conserved site,		
SMART	DISIN, EGF, ACR,		
SP_PIR_KEYWORDS	integrin, metalloprotease, Protease,		
ADAM23	**ADAM metallopeptidase domain 23**	**Related Genes**	**Bos taurus**
GOTERM_BP_FAT	proteolysis,		
GOTERM_CC_FAT	integral to membrane, intrinsic to membrane,		
GOTERM_MF_FAT	endopeptidase activity, metalloendopeptidase activity, peptidase activity, metallopeptidase activity, zinc ion binding, ion binding, cation binding, metal ion binding, transition metal ion binding, peptidase activity, acting on L-amino acid peptides,		
INTERPRO	EGF-like, type 3, Peptidase M12B, ADAM/reprolysin, Blood coagulation inhibitor, Disintegrin, Peptidase M12B, propeptide, EGF-like, ADAM, cysteine-rich, EGF-like region, conserved site, EGF, extracellular,		
SMART	DISIN, EGF, ACR,		

Functional Annotation Chart

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
90 DAVID IDs

Options

Rerun Using Options | Create Sublist

149 chart records Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	UP_SEQ_FEATURE	signal peptide	RT		31	34.4	2.8E-5	1.1E-2
	SP_PIR_KEYWORDS	signal	RT		31	34.4	3.2E-5	7.3E-3
	GOTERM_CC_FAT	extracellular region	RT		26	28.9	3.3E-5	5.0E-3
	SP_PIR_KEYWORDS	Secreted	RT		21	23.3	3.6E-5	4.1E-3
	GOTERM_BP_FAT	response to bacterium	RT		8	8.9	9.4E-5	9.8E-2
	GOTERM_CC_FAT	extracellular region part	RT		16	17.8	1.8E-4	1.3E-2
	INTERPRO	Alpha-defensin	RT		3	3.3	2.3E-4	6.0E-2
	INTERPRO	Defensin propeptide	RT		3	3.3	2.3E-4	6.0E-2
	INTERPRO	Alpha defensin	RT		3	3.3	2.3E-4	6.0E-2
	SP_PIR_KEYWORDS	antibiotic	RT		5	5.6	2.3E-4	1.8E-2
	SP_PIR_KEYWORDS	disulfide bond	RT		27	30.0	2.5E-4	1.4E-2
	SP_PIR_KEYWORDS	Antimicrobial	RT		5	5.6	2.8E-4	1.3E-2
	GOTERM_CC_FAT	extracellular space	RT		13	14.4	3.1E-4	1.5E-2
	UP_SEQ_FEATURE	disulfide bond	RT		26	28.9	3.3E-4	6.1E-2
	PIR_SUPERFAMILY	PIRSF001875:alpha-defensin	RT		3	3.3	3.6E-4	2.0E-2
	GOTERM_BP_FAT	defense response to bacterium	RT		6	6.7	3.9E-4	1.9E-1
	GOTERM_BP_FAT	defense response	RT		12	13.3	5.4E-4	1.8E-1
	INTERPRO	Mammalian defensin	RT		3	3.3	6.4E-4	8.2E-2
	SP_PIR_KEYWORDS	fungicide	RT		3	3.3	1.0E-3	3.9E-2
	SMART	DEFSN	RT		3	3.3	1.1E-3	6.9E-2
	GOTERM_BP_FAT	response to drug	RT		7	7.8	1.2E-3	2.9E-1
	SP_PIR_KEYWORDS	lipoprotein	RT		10	11.1	2.1E-3	6.6E-2
	GOTERM_BP_FAT	defense response to fungus	RT		3	3.3	2.3E-3	3.9E-1
	SP_PIR_KEYWORDS	glycoprotein	RT		32	35.6	2.3E-3	6.4E-2
	SP_PIR_KEYWORDS	palmitate	RT		6	6.7	2.5E-3	6.3E-2
	GOTERM_BP_FAT	killing of cells of another organism	RT		3	3.3	2.6E-3	3.8E-1