# CEAS (Cis-regulatory Element Annotation System)

**Summary:** We present a tool designed to characterize genome-wide protein-DNA interaction patterns from ChIP-chip and ChIP-Seq data of both sharply and broadly binding factors. This stand-alone extension of our web application CEAS (*Cis*-regulatory Element Annotation System) provides summary statistics on ChIP enrichment in important genomic regions such as individual chromosomes, promoters, gene bodies, or exons, and infers the genes most likely to be regulated by the binding factor under study. CEAS also enables biologists to visualize the average ChIP enrichment signals over specific genomic regions, particularly allowing observation of continuous and broad ChIP enrichment that 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 that analyzes ChIP regions by evaluating GC content and evolutionary conservation, conducting sequence motif search, and mapping the regions to their nearest genes (Ji, et al., 2006). However, additional analysis functions are needed to provide biologists with a more complete perspective. For example, in addition to 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, visual overviews of ChIP peaks' intensity distributions across chromosomes are helpful to biologists. Such analysis functions often require the ability to manipulate large continuous ChIP enrichment signal files (*e.g.* WIG files of hundreds of mega bytes in size), which are difficult to transfer to a web server. Therefore, in order to extend 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 genomic features such as promoters, immediate downstream regions of genes and exons, as well as a report on how individual genes are associated with their proximal ChIP regions.

## Installation

## *Requirements*

1. CEAS was developed based on Python 2.5.1; thus, Python higher than 2.5.1 is recommended to be installed for running CEAS w/o any potential version mismatch problems.
2. MySQLdb Python package is optional; however, if users want to directly use gene annotation tables from UCSC, MySQL and MySQLdb must be installed before installing CEAS.

## *Installation steps*

### Unix-like OS (UNIX, Linux, or Mac OS X):

1. Extract CEAS-X.X.X.tar.gz by typing on the command line;

   $ tar xvf CEAS-X.X.X.tar.gz

2. Enter the folder created by tar (*i.e.* ./CEAS-X.X.X) and type the following on the command line.

   $ cd CEAS-X.X.X
   $ sudo python setup.py install

   Note that the installer must have the proper security privilege to install CEAS at **/usr/local/bin**; otherwise, the user can install CEAS at the home as follows.

   $ python setup.py install --prefix=$HOME

   where $HOME stands for the user's home.

   Then, add a line to the log-in shell script (*i.e.* .bashrc in case of bash shell)

   export $PYTHONPATH=$PYTHONPATH:$HOME/lib/python2.5/site-packages

   where 'python2.5' should be changed to the Python version that the user uses.

# Get started with CEAS

## *CEAS Overview*

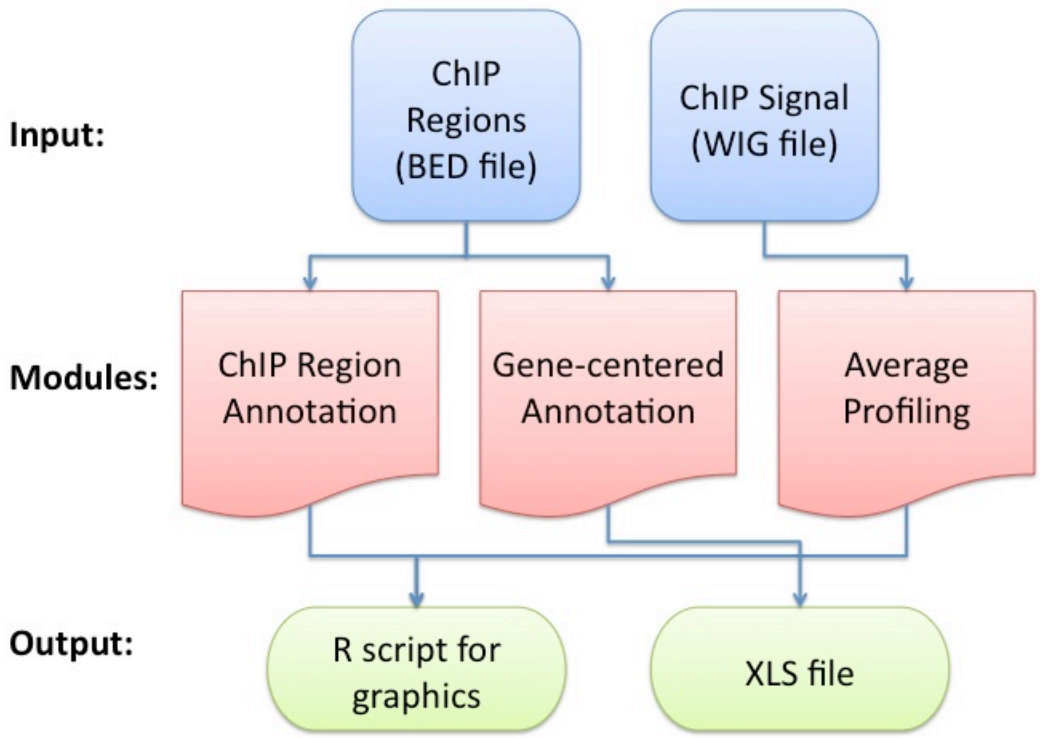



**Figure 1 The work-flow of CEAS. A gene annotation table, a BED file with the ChIP regions, and a WIG file with the ChIP enrichment signal are required. CEAS consists of three modules: ChIP region annotation, gene-centered annotation, and average profiling within/near important genomic features. As output, CEAS produces an R script of graphical results and a tab-delimited with XLS extension of gene-centered annotation.**

## *Inputs*

CEAS needs three major input files:

- Gene annotation table file
- BED file with ChIP regions
- WIG file with ChiP enrichment signal

### Gene annotation table

We provide gene annotation tables in sqlite3 files for several genomes (ce4 and ce6 for worm, dm2 and dm3 for fly, mm8 and mm9 for mouse, hg18 and hg19 for human). Although the user could also use the tables for the other genomes from the UCSC genome-browser through MySQLdb function included in CEAS, using our local sqlite3 files gives some advantages over the direct use of UCSC tables. First, since we included

pre-compiled genome background annotation for ChIP region annotation in every our gene annotation table file, CEAS can run faster. If a local gene annotation table file is not given, CEAS calculates the genome background annotation based on the gene annotation table downloaded from UCSC and the input WIG file, which usually takes longer. Moreover, if the input WIG file does not enclose the entire tiling or mappable genomic regions (*e.g.* the sequencing depth is not deep enough in case of ChIP-Seq), the genome background annotation may not be accurate because a lot of regions supposed to be considered are included by the WIG file.

### BED file with ChIP regions

The BED file are required to contain the chromosomes, start, and end locations of ChIP regions identified by a peak-caller (*e.g.* MAT for affy ChIP-chip, MA2C for Nimblegen ChIP-chip, MACS for ChIP-Seq). Other fields such as name or score are optional. An example BED file is given below. The ChIP regions do not have to be sorted in advance because CEAS sorts the regions by default.

*e.g.*)

```
chr1    779600      780954
chr1    874824      876507
chr1    1147745     1148979
chr1    1576270     1576887
chr1    2325039     2326135
chr1    2429436     2430692
```

### WIG file with ChIP enricment signal

The WIG file contains a continuous ChIP enrichment signal whereas the BED file shows discrete ChIP regions. Currently, CEAS supports only variableStep WIG file, which basically has two columns of genomic locations and corresponding signal values with a header where every chromosome starts (see the below example).

*e.g.)*

```
track type=wiggle_0
variableStep chrom=chr1 span=1
6131  3
6141  4
6151  4
6161  4
6171  4
6181  4
6191  4
6201  1
6211  0
6221  1
6231  3
```

## *Modules*

### ChIP region annotation

CEAS estimates the relative enrichment level of ChIP regions in each genomic feature with respect to the whole genome. To do this, it first calculates the percentages of the ChIP regions that reside in the following four categories: (a) promoters, (b) bidirectional promoters, (c) downstream regions of genes, and (d) gene bodies (3'UTRs, 5'UTRs, coding exons, and introns). In addition to these categories, the user can add another user-specified extra category (*e.g.* non-coding regions) as an optional input BED file. 'Promoters' correspond to the upstream regions of the transcription start sites (TSSs) of genes. The user specifies three promoter sizes (1kb, 2kb, and 3kb by default) to be used in annotation. For instance, if the user sets the promoter sizes to be 1kb, 3kb, and 10kb upstream of the TSS, CEAS computes the cumulative percentages of ChIP regions that fall in <=1kb, <=3kb, and <=10kb upstream of the TSSs of genes. 'Bidirectional promoters' are promoter regions between divergently transcribed genes whose TSSs are closer in proximity than user-defined distances (two options, genes, equal in length to the sizes specified for 'promoters.' 'Gene bodies' are divided into UTR regions (3' and 5' UTRs), coding exons and introns. After the percentages of ChIP regions residing within the above categories are obtained, they are compared against the genome background percentages for the same categories and *p*-values for the significance of the relative enrichment with respect to the background are calculated using one-sided binomial test. In order to summarize the ChIP region annotation, CEAS draws a pie chart displaying the distribution of ChIP regions across the genomic features. 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 by proximity is important to infer the direct regulatory gene targets of the binding factor under study. 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 under study. In case a broad ChIP peak covers all or part of a gene body, it is useful to know how much of the gene, including its promoter or downstream region, is occupied by the ChIP region. To this end, CEAS divides every gene into three equal fractions and calculates the percentage of the area covered by ChIP regions. CEAS also estimates the percentages of the promoter and downstream region 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 convenient use with Excel and contain 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 used in peak calling. Therefore, CEAS displays the continuous ChIP enrichment signal within and near important genomic features for biologists to visualize the average binding patterns in these regions. CEAS draws the

average signals around TSSs and TTSs in a user-defined range (+/- 3kb from the TSS and TTS by default). In addition, CEAS computes average signals on 'meta-gene,' 'meta-concatenated-exon,' 'meta-concatenated-intron,' 'meta-exons' and 'meta-introns,' where the prefix 'meta' indicates that every element (*e.g.* every gene) is normalized to have the same length (*e.g.* 3kb for the meta-gene). The difference between meta-concatenated-exon and meta-exons is that the first concatenates all exons of a gene (like a meta-cDNA) before calculating the average gene 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 that draws the average ChIP signals of multiple user-specified sub-groups of genes, allowing the user to compare the signals 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 user-provided list of sites (specified in a BED) to enable biologists to visualize the average signals in any arbitrary regions (*e.g.* transcription factor binding sites) in addition to the pre-defined genomic regions.

## *Outputs*

CEAS generates two files as output.

- R script
- Tab-delimited TXT file with XLS extension

The R file contains the script to generate the graphical results of ChIP region annotation and average signal profiling within/near important genomic regions. CEAS writes the result of gene-centered annotation in a tab-delimited txt file with XLS extension for easy XLS visualization.

### R script

The user can obtain the PDF of graphical results by typing the following on the command line.

$ R --vanilla < *file.R*

, where *file.R* stands for the R script produced by CEAS. As a result of running the R script, file.pdf will be generated.

### XLS file

CEAS generates a tab-delimited TXT file with XLS extension for gene-centered annotation. See ***Example use*** for a detailed description of the XLS file.

## *Easy run*

If the user wants to run CEAS in the default mode, simply type on the command line:

$ **ceas -g** *gdb* **-b** *bed* **-w** *wig*

where *gdb, bed*, and *wig* stands for a gene annotation table file, a BED file with ChIP regions, and a WIG file with ChIP enrichment signal file, respectively. If no local gene annotation table is used, *gdb* will be a genome number (*e.g.* hg18)

For more sophisticated use of CEAS, see **Use CEAS**.

# Use CEAS

## *Running modes*

The CEAS running mode is determined by input file combination. As can be seen in Figure 1, ChIP region annotation and gene-centered annotation requires a gene annotation table file and a BED file whereas average signal profiling needs a gene annotation table and a WIG file. Therefore, CEAS adpatively operates the modules according to which input files are given. Of course, when all the input files are given, all of the three modules will run.

Assuming that default parameters values are selected for all other parameters,

- When the user wants to run only ChIP region annotation and gene-centered annotation

  $ **ceas** [*options*] **-g** *gdb* **-b** *bed*

- When the user wants to run only average signal profiling

  $ **ceas** [*options*] **-g** *gdb* **-w** *wig*

However, it should be noted that if the user needs to re-do genome background annotation with their own WIG file or run ChIP region annotation on a genome with no pre-compiled genome background annotation available, CEAS requires the user to input a WIG file and set **--bg** as follows. See ***Arguments*** for more details.

  $ **ceas** [*options*] **--bg -g** *gdb* **-b** *bed* **-w** *wig*

## *Arguments*

Usage: **ceas** [*options*] **-g** *gdb* (**-b** *bed* | **-w** *wig* )

The following is a detailed description of the options used to control CEAS.

| | |
|---|---|
| --version | Show program's version number and exit. |
| -h, --help | Show this help message and exit. |
| -b, --bed | BED file with ChIP regions. |
| -w, --wig | WIG file for either wig profiling or genome background annotation. WARNING: CEAS only accepts variableStep WIG file. The user must set --bg flag for genome background annotation. |
| -e, --ebed | BED file of extra regions of interest (*e.g.* non-coding regions) |

| | |
|---|---|
| -g, --gdb | Gene annotation table (a local sqlite3 db file provided by CEAS or a BED file or genome name used by UCSC). CEAS searches for a local db file. If not find, it looks up UCSC for the table. WARNING: When using UCSC, MySQLdb package must be installed! |
| --name | Experiment name. This will be used to name the output files (R script and XLS file). If an experiment name is not given, the stem of the input BED file name will be used instead. (*e.g.* if BED is peaks.bed, 'peaks' will be used as a name.) If a BED file is not given, the input WIG file name will be used. |
| --sizes | Promoter (also downstream) sizes for ChIP region annotation. Comma-separated three integer numbers or a single number will be accepted. If a single integer is given, it will be segmented into three equal fractions (*i.e.* 3000 is equivalent to 1000,2000,3000). DEFAULT: 1000,2000,3000. WARNING: numbers > 10000bp are automatically fixed to 10000bp. |
| --bisizes | Bidirectional-promoter sizes for ChIP region annotation. The user can choose two numbers to define bidirectional promoters. Comma-separated two values or a single value can be given. If a single value is given, it will be segmented into two equal fractions (*i.e.* 5000 is equivalent to 2500,5000) DEFAULT: 2500,5000bp. WARNING: numbers > 20000bp are automatically fixed to 20000bp. |
| --bg | Run genome BG annotation. WARNING: this flag is effective only if a WIG file is given through -w (--wig). Otherwise, ignored. |
| --span | Span from TSS and TTS in the gene-centered annotation. ChIP regions within this range from TSS and TTS are considered when calculating the coverage rates of promoter and downstream by ChIP regions. DEFAULT=3000bp |
| --pf-res | Wig profiling resolution, DEFAULT: 50bp. WARNING: a number smaller than the wig interval (resolution) may cause aliasing error. |
| --rel-dist | Relative distance to TSS/TTS in wig profiling. DEFAULT: 3000bp |
| --gn-groups | Gene-groups of particular interest in wig profiling. Each gene group file must have gene names in the 1s column. The file names are separated by commas w/ no space (*e.g.* --gn-groups=top10.txt,bottom10.txt) |
| --gn-group-names | The names of the gene groups in --gn-groups. The gene group names are separated by commas. (*e.g.* --gn-group-names='top 10%,bottom 10%'). These group names appear in the legends of the wig profiling plots. If no group names given, the groups are represented as 'Group 1, Group2,...Group n'. |
| --no-refseq | Whether RefSeq accession IDs (*i.e.* 'name' in refGene of UCSC) or gene IDs (*i.e.* 'name2' in refGene of UCSC) are used in --gn-groups. This flag is meaningful only if --gn-groups is set. |

## *Example use*

Here, we included two examples of CEAS runs. The first example is $CD4T^+$ cell H3K36me3 ChIP-Seq data and the second C.elegans SDC-3 ChIP-chip (Nimblegen).

### Human CD4T+ cell H3K36me3 ChIP-Seq

An example command line is as follows.

$ **ceas --name=**H3K36me3_ceas **--pf-res**=20 **--gn-groups**=top10.txt,bottom10.txt **--gn-group-names**='Top 10%,Bottom 10%' **-g** hg18 **-b** H3K36me3_MACS_pval1e-5_peaks.bed **-w** H3K36me3.wig

The name of this run is 'H3K36me3_ceas' (after **--name**). Two files of gene groups, top10.txt and bottom10.txt after **--gn-groups**, contain the lists of RefSeq genes that are highly expressed (top 10%) and lowly expressed (bottom 10%) in the human $CD4T^+$ cell, respectively. The average profiles of ChIP enrichment signal on these two gene groups, which will be referred to as *Top 10%* and *Bottom 10%* (after **--gn-group-names**), will appear along with the average profile over all genes in the average profile plots. In this example, it is assumed that the local sqlite3 db file of hg18 (after **-g**) was used and it is located in the working directory. If the user wants to use the gene table in UCSC through MySQLdb, type '**-g** hg18 **--bg**' after moving the local db file from the working directory because CEAS first searches for a local file by default. The option, **--bg,** forces CEAS to run genome background annotation using H3K36me3.wig. However, in this special case, using the local db file is more recommendable because H3K36me3 is believed to be transcription elongation mark (gene body mark) and the WIG file may contain only gene bodies.

When running CEAS, it prints the following summary of argument selection to stdout, which helps the user to trace their parameter selections, particularly in case that the user wants to run CEAS multiple times with different parameter sets.

```
# PARAMETER LIST:
# name = H3K36me3_ceas
# gene annotation table = hg18
# BED file = H3K36me3_MACS_pval1e-5_peaks.bed
# WIG file = H3K36me3.wig
# extra BED file = None
# ChIP annotation = True
# gene-centered annotation =  True
# average profiling = True
# re-annotation for genome background (ChIP region annotation) = False
# promoter sizes (ChIP region annotation) = 1000,2000,3000 bp
# downstream sizes (ChIP region annotation) = 1000,2000,3000 bp
# bidirectional promoter sizes (ChIP region annotation) = 2500,5000 bp
# span size (gene-centered annotation) = 3000 bp
# profiling resolution (average profiling) = 20 bp
# relative distance wrt TSS and TTS (average profiling) = 3000 bp
```

# gene groups (average profiling) = CD4T_top10.txt, CD4T_bottom10.txt

- ChIP region annotation and average profiling within/near important genomic features -

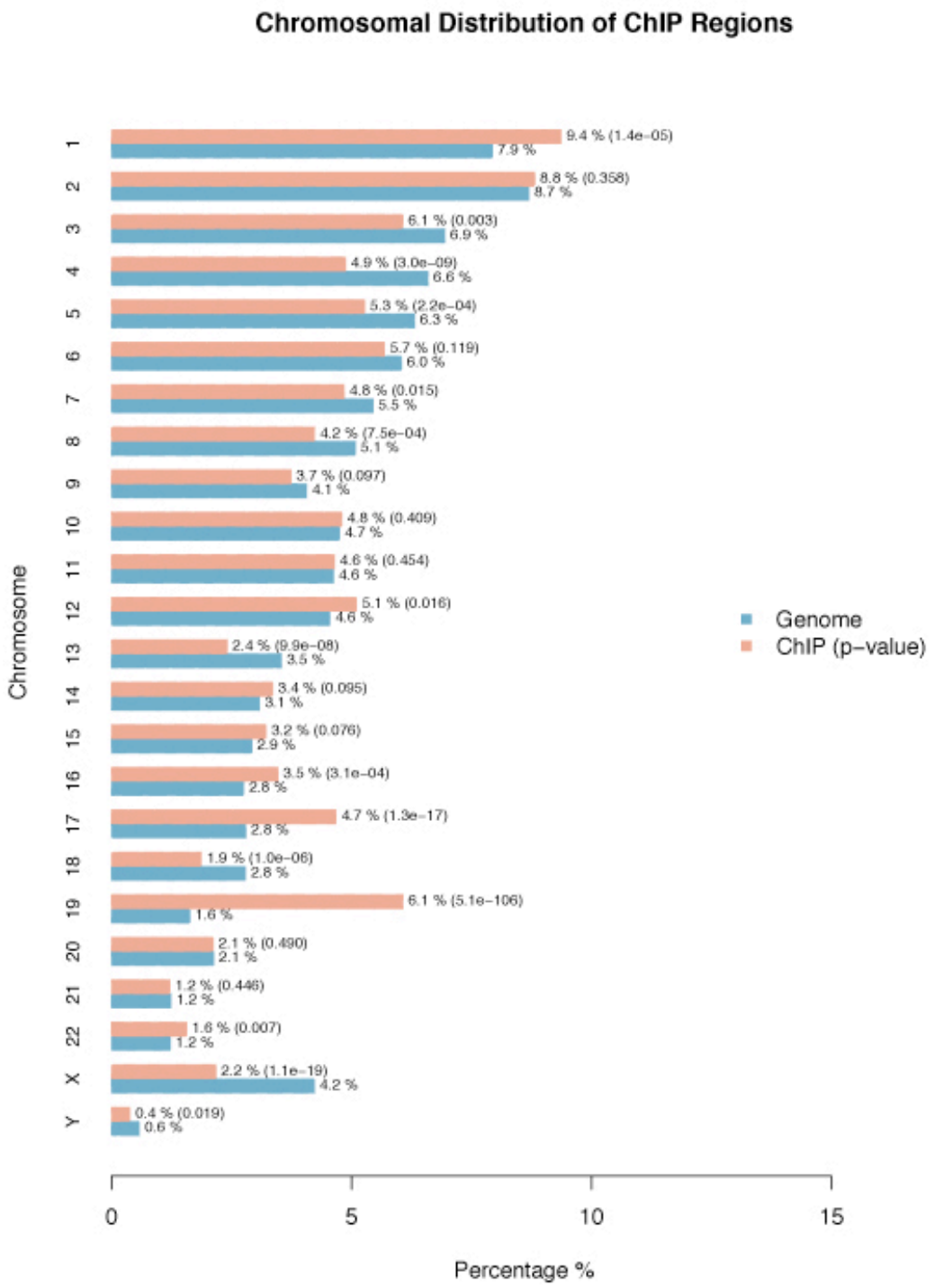



**Figure 2 The first page shows the distribution of ChIP regions over chromosomes. The blue bars represent the percentages of the whole tiled or mappable regions in the chromosomes (genome background) and the red bars the percentages of the whole ChIP. These percentages are also marked right next to the bars. *P*-values for the significance of the relative enrichment of ChIP regions with respect to the gnome background are shown in parentheses next to the percentages of the red bars. For example, in the above figure,**

**9.4 % of ChIP regions reside in chr1 whereas 7.9 % of the whole tiled (or mappable regions) occupy chr1 with a *P*-value of 1.4e-5. The sum of percentages of the red bars (or blue bars, equivalently) is 100 %.**

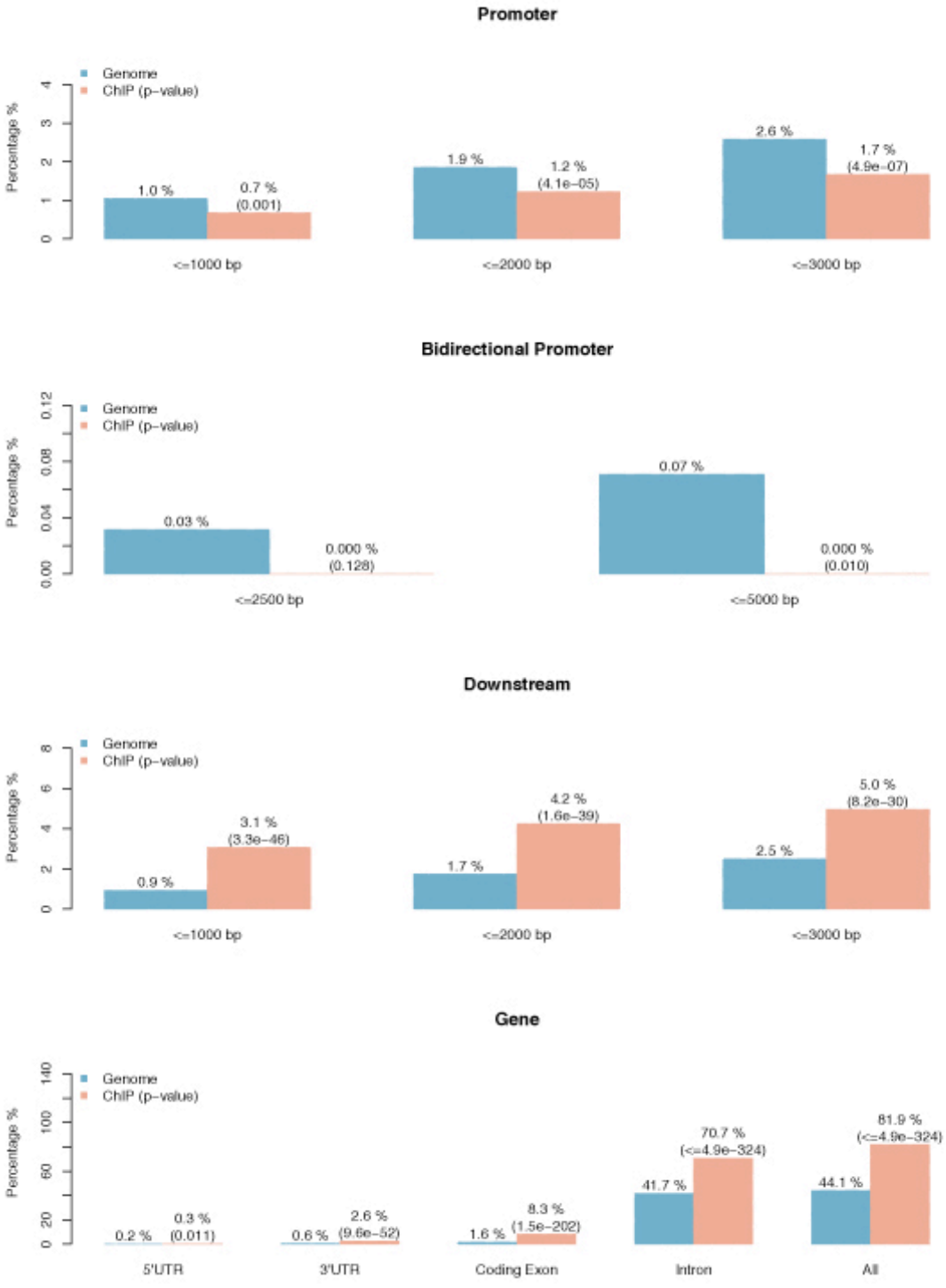



**Figure 3 The second page shows the relative enrichments of ChIP regions in important genomic features, such as promoters, immediate downstreams of genes, and gene bodies, with respect to the genome background. As on the first page, the blue bars represent the percentages of the tiled or mappable regions that are located in such genomic regions (genome background) and the red bars the percentages of ChIP regions. Since H3K36me3 is a transcriptional elongation mark, it shows very high relative**

**enrichment in gene bodies (81.9% of ChIP regions within gene bodies compared to 44.1 % of the genome background) but very low enrichment in promoters including bidirectional ones (only 1.7 % of ChIP regions within 3 kb upstreams of TSS). In addition, it was also observed that H3K36me3 has a long tail after TTS considering its high relative enrichment in the immediate downstream (see the third bar plot).**

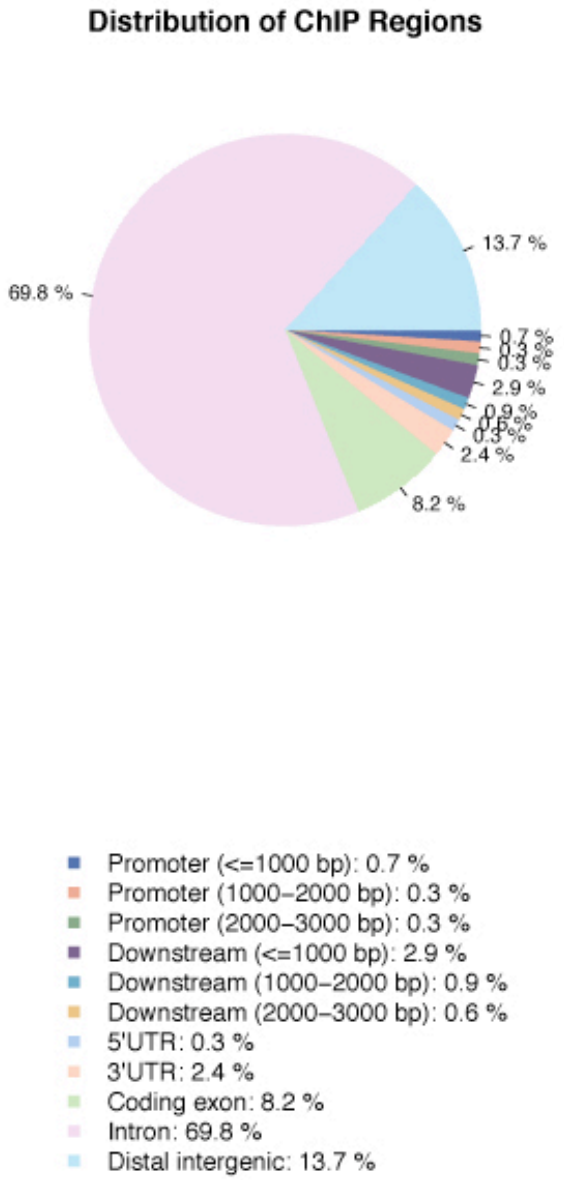



**Figure 4 The third page of the CEAS graphical results illustrates how ChIP regions are distributed over important genomic features. Unlike the bar plots on the second page, the genomic features in the above pie chart are mutually exclusive (no overlaps); thus, the sum of the percentage values is 100 %. 'Distal intergenic' represents the percentage of ChIP regions that do not belong in any of other genomic features.**

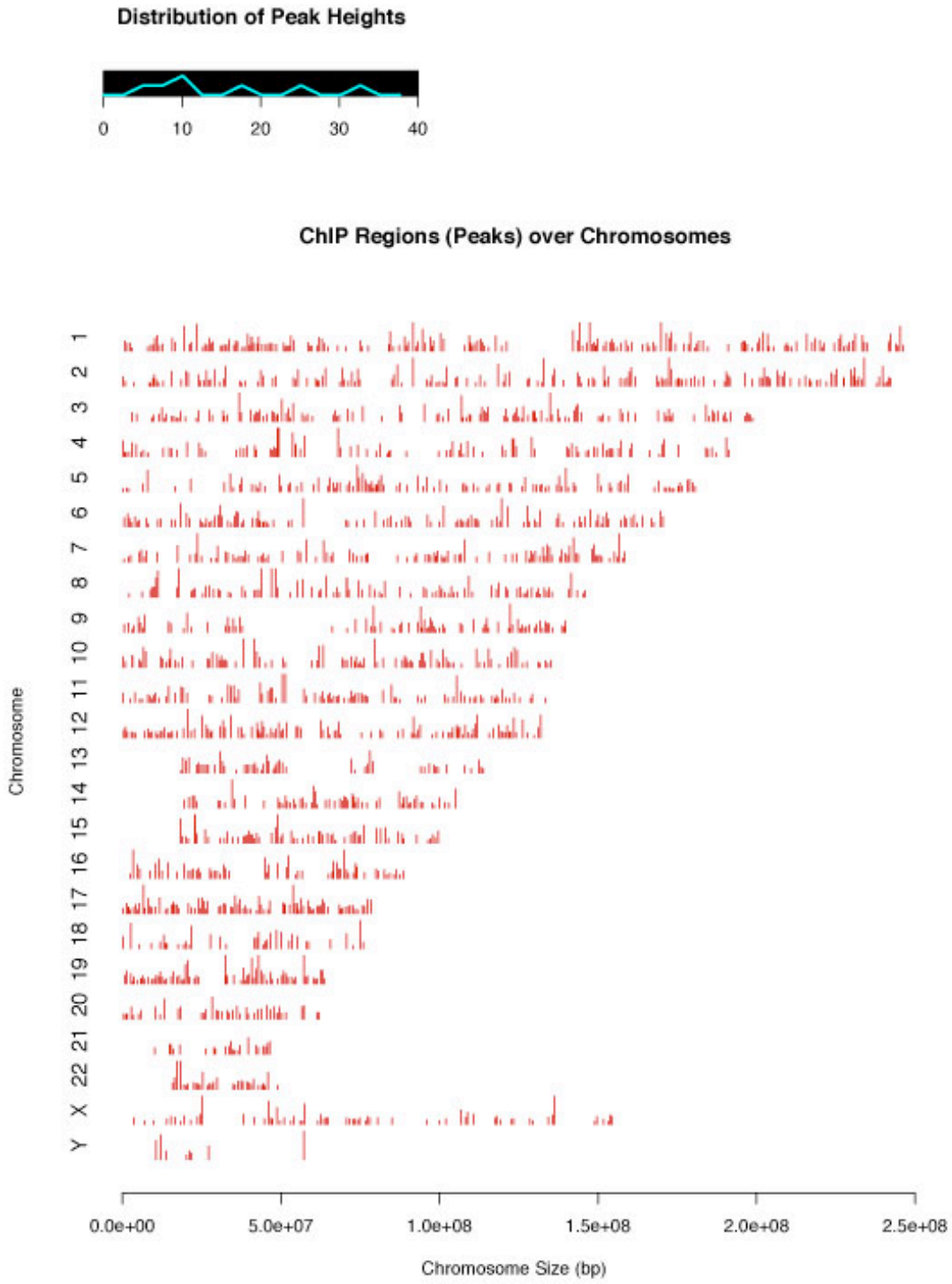



**Figure 5 The fourth page visualizes how ChIP regions are distributed over the genome along with their scores or peak heights. The line graph on the top left corner illustrates the distribution of peak heights (or scores). The red bars in the main plot ChIP regions in the input BED file. In this particular example, the maximum peak height is 40 (the largest value on the x-axis of the line). The x-axis of the main plot represents the actual chromosome sizes.**

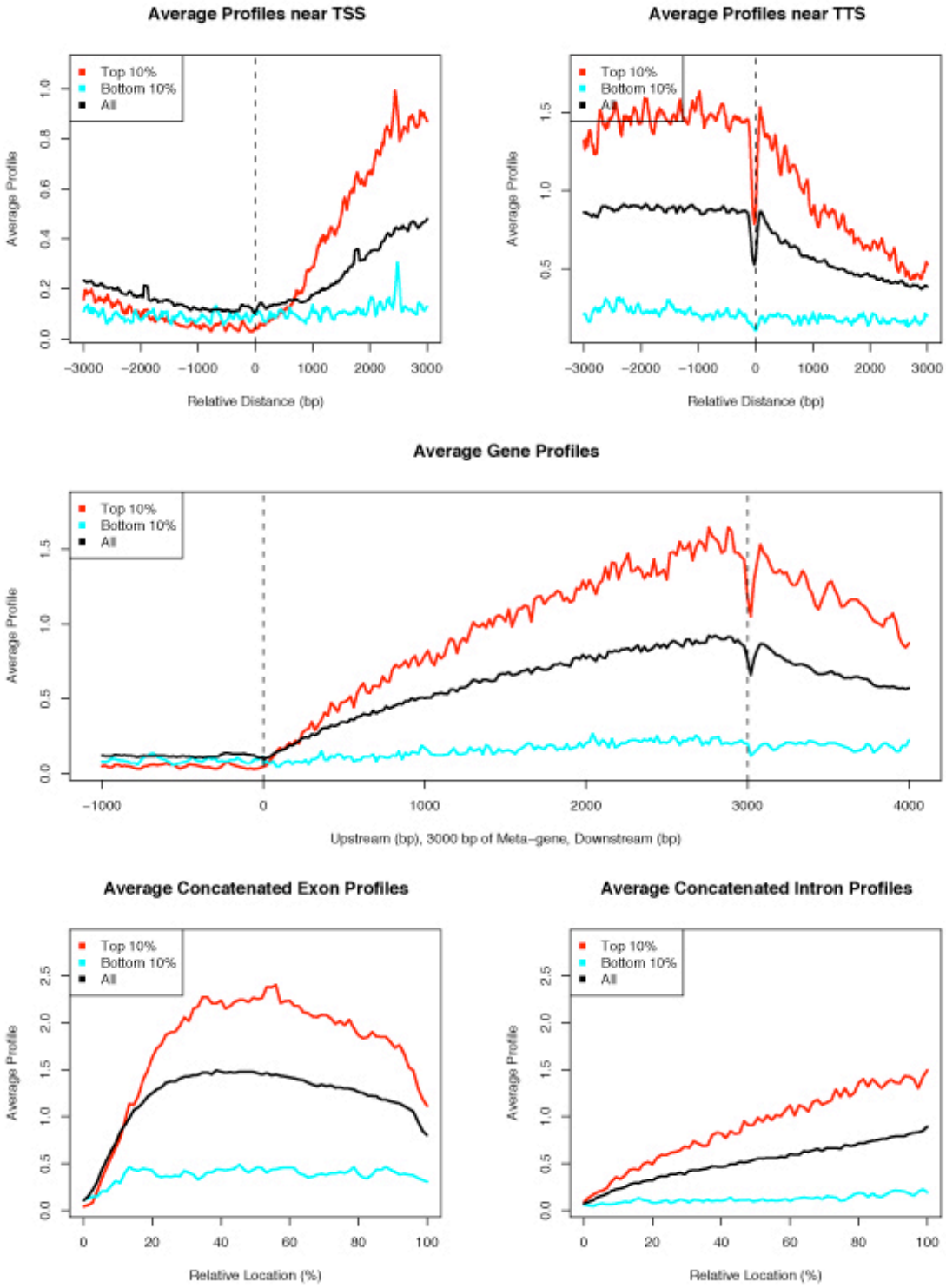



**Figure 6 The fifth and sixth pages show the results of average profiling within/near important genomic features. The panels on the first row display the average ChIP enrichment signals around TSS and TTS of genes, respectively. The red, cyan, and black colors indicate the average ChIP enrichment signals of the top 10 % of the expressed genes, bottom 10%, and all RefSeq genes. On the right panel, it is observed that H3K36me3 has a tail after 3' end of genes with a nucleosome depletion at TTS, which also**

**agrees with the observation made in Figure 3 (the bar plot of downstream). The middle panel (on the second row) represents the average ChIP signals on the meta-gene of 3 kb, which shows that H3K36me3 enriches gene bodies and increases towards the 3' end. In addition, CEAS concatenates exons of a gene before calculating the average gene (like meta-cDNA) (the left panel on the bottom row). CEAS produces a similar plot for concatenated introns as well (the right panel on the bottom row).**

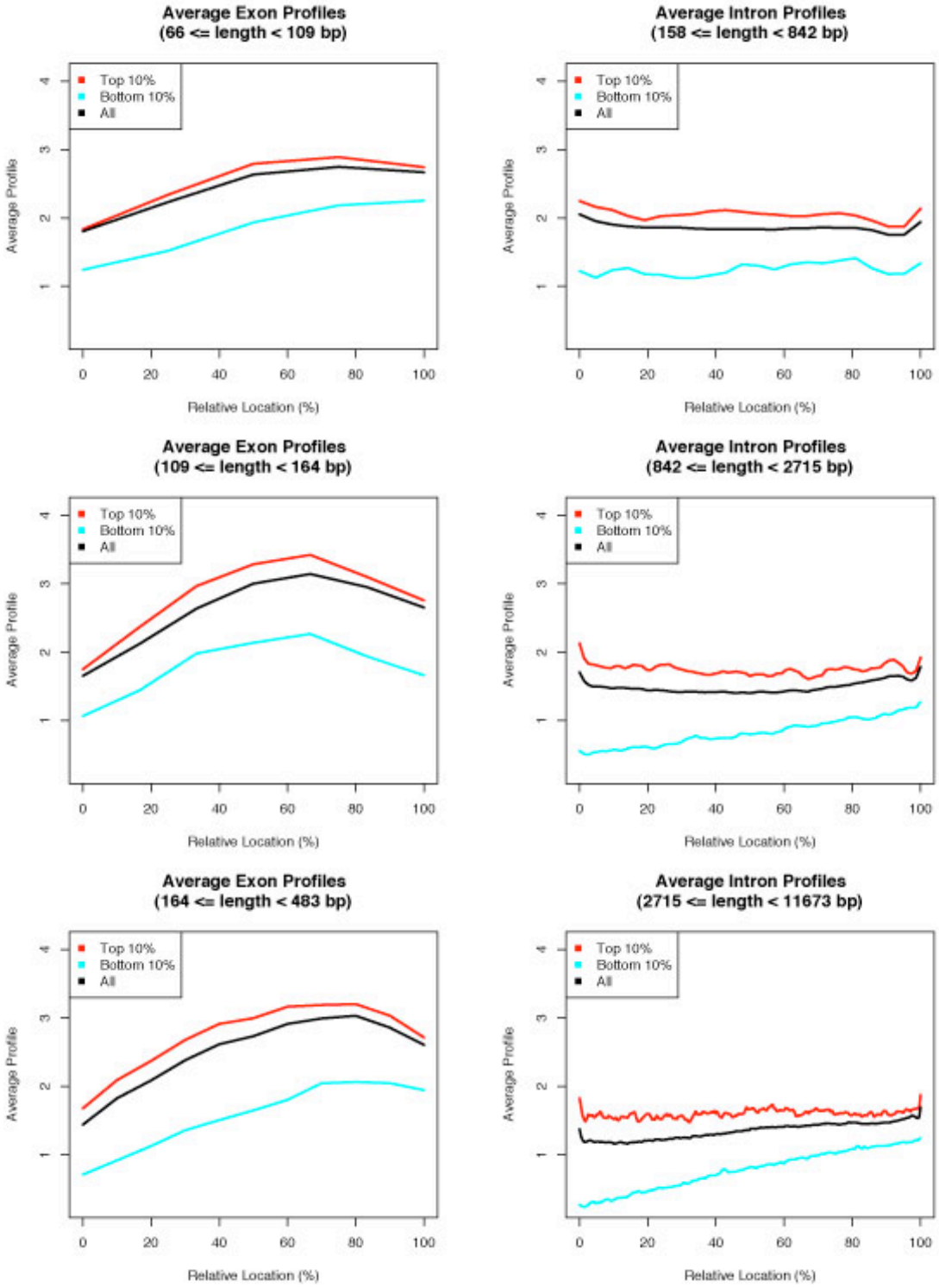
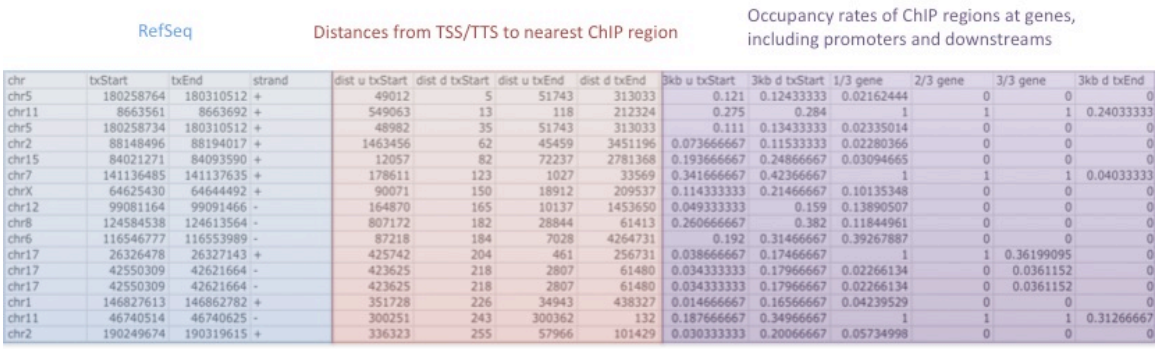



**Figure 7 In addition to the average profiles on concatenated exons and concatenated introns, CEAS shows the average ChIP signals on all of exons and introns as well. Since exon and intron lengths highly vary from gene to gene, CEAS groups exons (or introns) into three classes by length and calculates the respect average profiles in order to avoid any potential graphical artifacts due to length-normalization.**

**- Gene-centered annotation -**



| | RefSeq | | | Distances from TSS/TTS to nearest ChIP region | | | | Occupancy rates of ChIP regions at genes, including promoters and downstreams | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| chr | txStart | txEnd | strand | dist u txStart | dist d txStart | dist u txEnd | dist d txEnd | 3kb u txStart | 3kb d txStart | 1/3 gene | 2/3 gene | 3/3 gene | 3kb d txEnd |
| chr5 | 180258764 | 180310512 | + | 49012 | 5 | 51743 | 313033 | 0.121 | 0.12433333 | 0.02162444 | 0 | 0 | 0 |
| chr11 | 8663561 | 8663692 | + | 549063 | 13 | 118 | 212324 | 0.275 | 0.284 | 1 | 1 | 1 | 0.24033333 |
| chr5 | 180258734 | 180310512 | + | 48982 | 35 | 51743 | 313033 | 0.111 | 0.13433333 | 0.02335014 | 0 | 0 | 0 |
| chr2 | 88148496 | 88194017 | + | 1463456 | 62 | 45459 | 3451196 | 0.073666667 | 0.11533333 | 0.02280366 | 0 | 0 | 0 |
| chr15 | 84021271 | 84093590 | + | 12057 | 82 | 72237 | 2781368 | 0.193666667 | 0.24866667 | 0.03094665 | 0 | 0 | 0 |
| chr7 | 141136485 | 141137635 | + | 178611 | 123 | 1027 | 33569 | 0.341666667 | 0.42366667 | 1 | 1 | 1 | 0.04033333 |
| chrX | 64625430 | 64644492 | + | 90071 | 150 | 18912 | 209537 | 0.114333333 | 0.21466667 | 0.10135348 | 0 | 0 | 0 |
| chr12 | 99081164 | 99091466 | - | 164870 | 165 | 10137 | 1453650 | 0.049333333 | 0.159 | 0.13890507 | 0 | 0 | 0 |
| chr8 | 124584538 | 124613564 | - | 807172 | 182 | 28844 | 61413 | 0.260666667 | 0.382 | 0.11844961 | 0 | 0 | 0 |
| chr6 | 116546777 | 116553989 | - | 87218 | 184 | 7028 | 4264731 | 0.192 | 0.31466667 | 0.39267887 | 0 | 0 | 0 |
| chr17 | 26326478 | 26327143 | + | 425742 | 204 | 461 | 256731 | 0.038666667 | 0.17466667 | 1 | 1 | 0.36199095 | 0 |
| chr17 | 42550309 | 42621664 | - | 423625 | 218 | 2807 | 61480 | 0.034333333 | 0.17966667 | 0.02266134 | 0 | 0.0361152 | 0 |
| chr17 | 42550309 | 42621664 | - | 423625 | 218 | 2807 | 61480 | 0.034333333 | 0.17966667 | 0.02266134 | 0 | 0.0361152 | 0 |
| chr1 | 146827613 | 146862782 | + | 351728 | 226 | 34943 | 438327 | 0.014666667 | 0.16566667 | 0.04239529 | 0 | 0 | 0 |
| chr11 | 46740514 | 46740625 | - | 300251 | 243 | 300362 | 132 | 0.187666667 | 0.34966667 | 1 | 1 | 1 | 0.31266667 |
| chr2 | 190249674 | 190319615 | + | 336323 | 255 | 57966 | 101429 | 0.030333333 | 0.20066667 | 0.05734998 | 0 | 0 | 0 |

**Figure 8 This figure is a screen shot of the XLS file from gene-centered annotation that was process for better visualization. The fields of the XLS file are described below.**

| 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 region in 3kb upstream of txStart (0.0 - 1.0) |
| 3kb d txStart | Occupancy rate of ChIP region in 3kb downstream of txStart (0.0 - 1.0) |
| 1/3 gene | Occupancy rate of ChIP region in the 1$^{st}$ third of a gene (0.0 - 1.0) |
| 2/3 gene | Occupancy rate of ChIP region in the 2$^{nd}$ third of a gene (0.0 - 1.0) |
| 3/3 gene | Occupancy rate of ChIP region in the 3$^{rd}$ third of a gene (0.0 - 1.0) |
| 3kb d txEnd | Occupancy rate of ChIP region in 3kb downstream of txEnd (0.0 - 1.0) |

For example, the first RefSeq gene in Figure 8 is located at chr5 180258764:180310512 at the sense strand, and the distance to the nearest ChIP region upstream of its txStart is 49,012 bp and the distance downstream of its txStart is 5 bp. About 12 % (0.121) of the 3,000 bp upstream region of the gene is occupied by ChIP region(s), and the occupancy rate decreases towards the 3' end of the gene (2 % of 1/3 gene, 0 % of the others).

## C.elegans SDC3 ChIP-chip (Nimblegen)

Another example of CEAS is C.elegans SDC3 ChIP-chip (Nimblegen).

$ **ceas --name=**SDC3_ceas **--bg --no-refseq --pf-res**=86 **--gn-groups**=ce_top10.txt,ce_middle10.txt,ce_bottom10.txt **--gn-group-names**='Top 10%,Middle 10%,Bottom 10%' **-g** ce4 **-b** SDC3_MA2C_peaks.bed **-w** SDC3_MA2Cscore.wig **-e** nc_regions.bed

In this case, we decided to re-annotate the genome background (**--bg**) based on SDC3_MA2Cscore.wig file because the input WIG file contains all tiled locations. The option, **--no-refseq**, was set since gene IDs instead of RefSeq IDs were used in the gene group files (ce_top10.txt,ce_middle10.txt,ce_bottom10.txt after **--gn-groups**). It should be noted that a wrong choice of profiling resolution (**--pf-res**) causes artifacts in average gene profiling. In this example, the WIG file resolution (tiling space) is 86 bp while the default is 50 bp. An extra BED file, nc_regions.bed, is also given after **-e** (or **--ebed**) to perform ChIP region annotation on non-coding regions of C. elegans.

The following refers to the parameter selections of this CEAS run.

# PARAMETER LIST:
# name = SDC3_ceas
# gene annotation table = ce4
# BED file = SDC3_MA2C_peaks.bed
# WIG file = SDC3_MA2Cscore.wig
# extra BED file = nc_regions.bed
# ChIP annotation = True
# gene-centered annotation = True
# average profiling = True
# re-annotation for genome background (ChIP region annotation) = True
# promoter sizes (ChIP region annotation) = 1000,2000,3000 bp
# downstream sizes (ChIP region annotation) = 1000,2000,3000 bp
# bidirectional promoter sizes (ChIP region annotation) = 2500,5000 bp
# span size (gene-centered annotation) = 3000 bp
# profiling resolution (average profiling) = 86 bp
# relative distance wrt TSS and TTS (average profiling) = 3000 bp
# gene groups (average profiling) = ce_top10.txt, ce_middle10.txt, ce_bottom10.txt

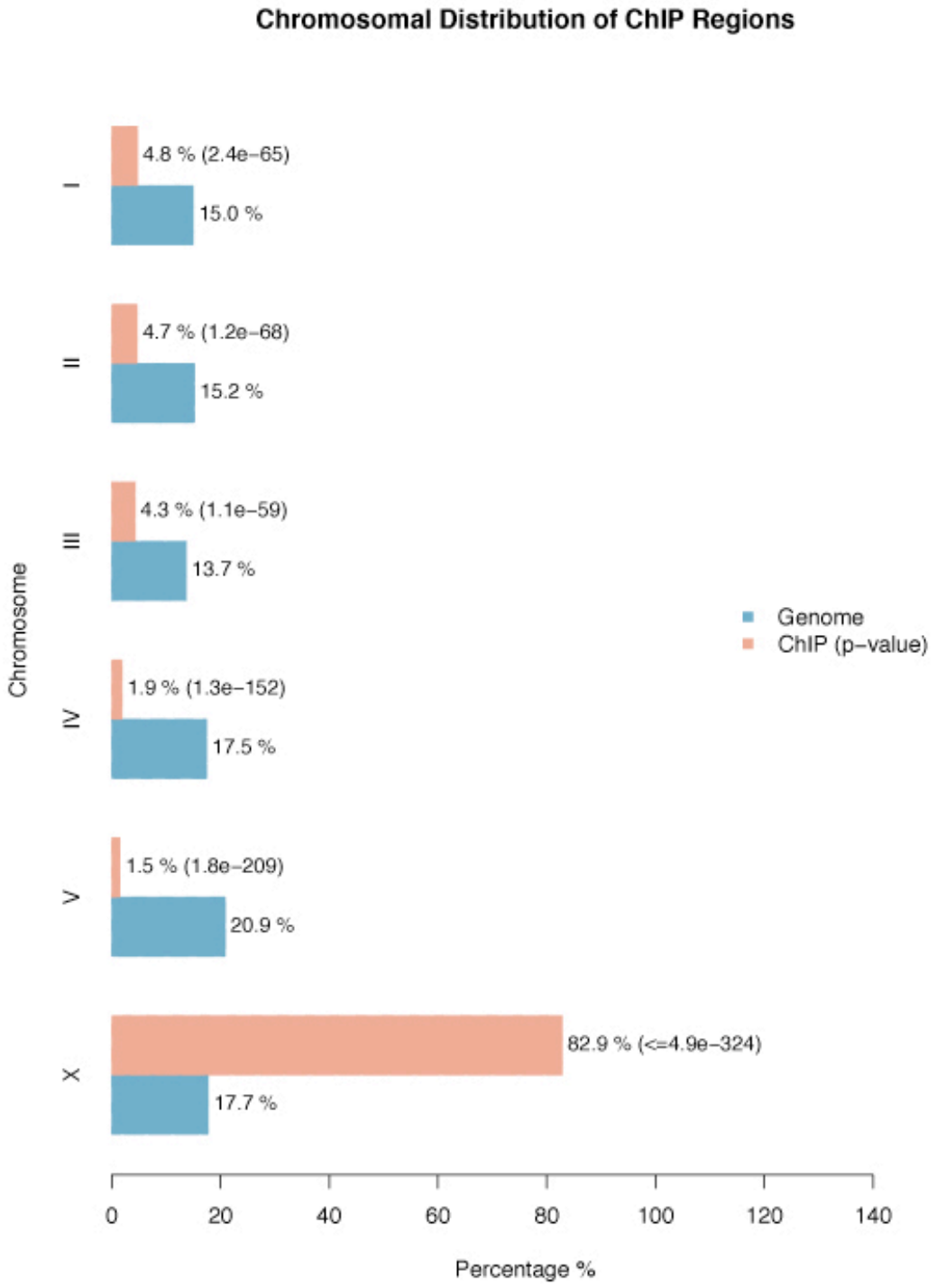



**Figure 9 SDC-3 is a transcription factor that regulates sex determination and dosage compensation in C. elegans. Therefore, it tends to bind more on X chromosome than other autosomes.**

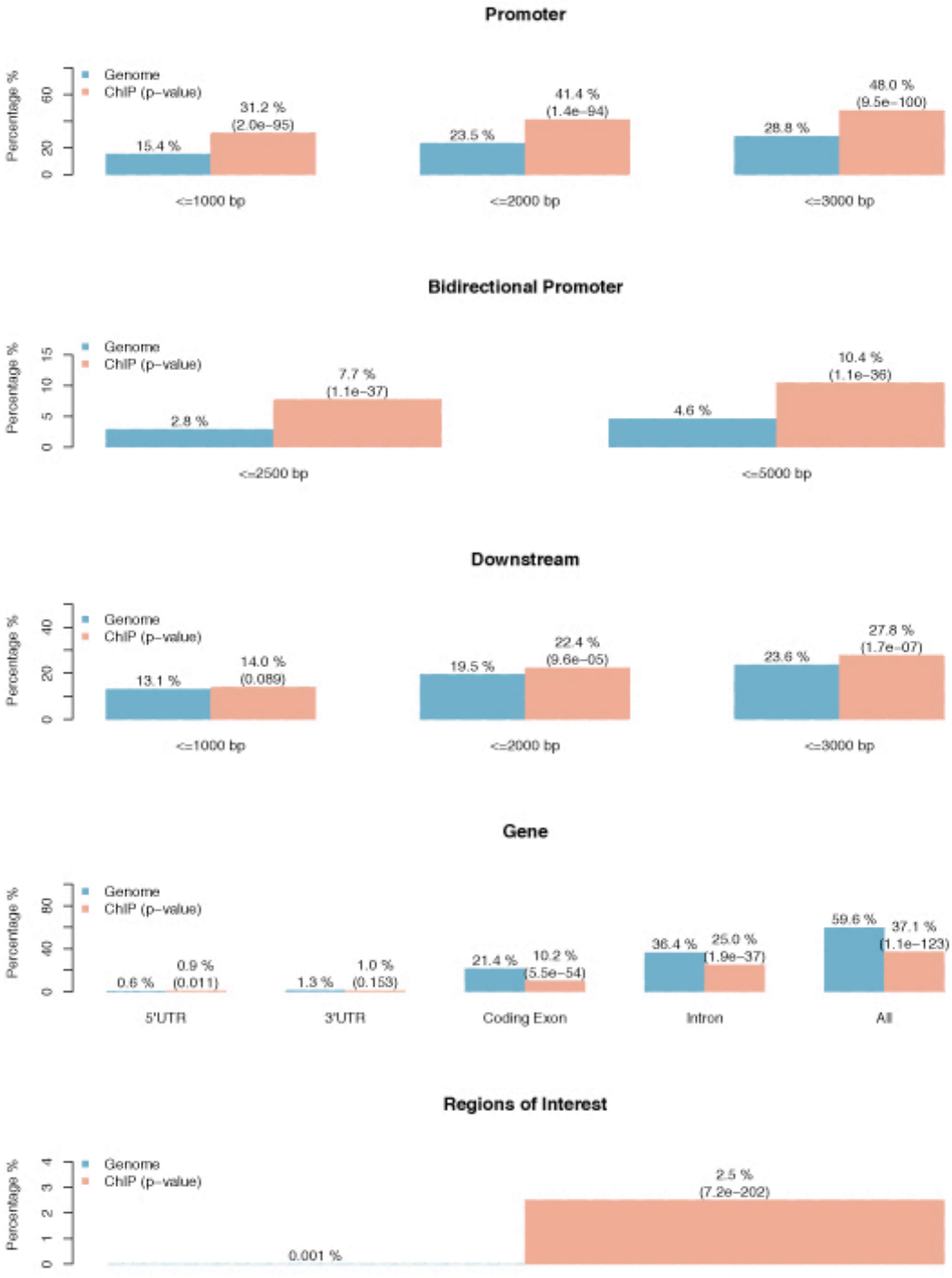



**Figure 10 In this run, an extra BED file of non-coding regions was added. The bottom row (Regions of Interest) shows a high relative enrichment level of SDC-3 ChIP regions in the non-coding regions with respect to the genome background (*P*-value of 7.2e-202).**

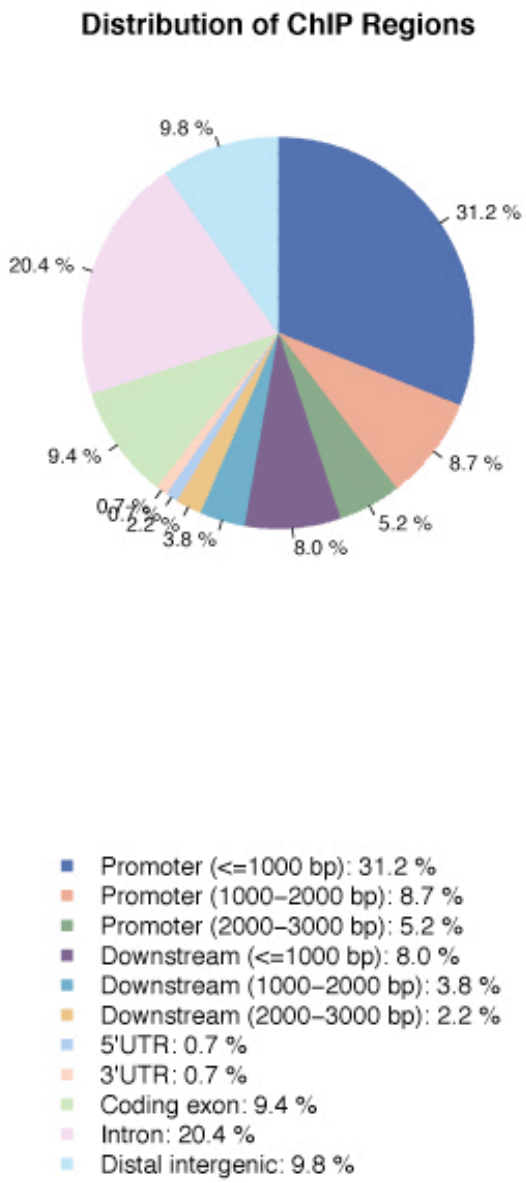



**Figure 11 Also, this pie chart clearly shows that SDC-3 tends to bind to promoters of genes to regulate the genes (about 45 % of ChIP regions are in promoters).**


**Figure 12 This plot also confirms that SDC-3 has more and strong binding sites on the X chromosome than autosomes.**

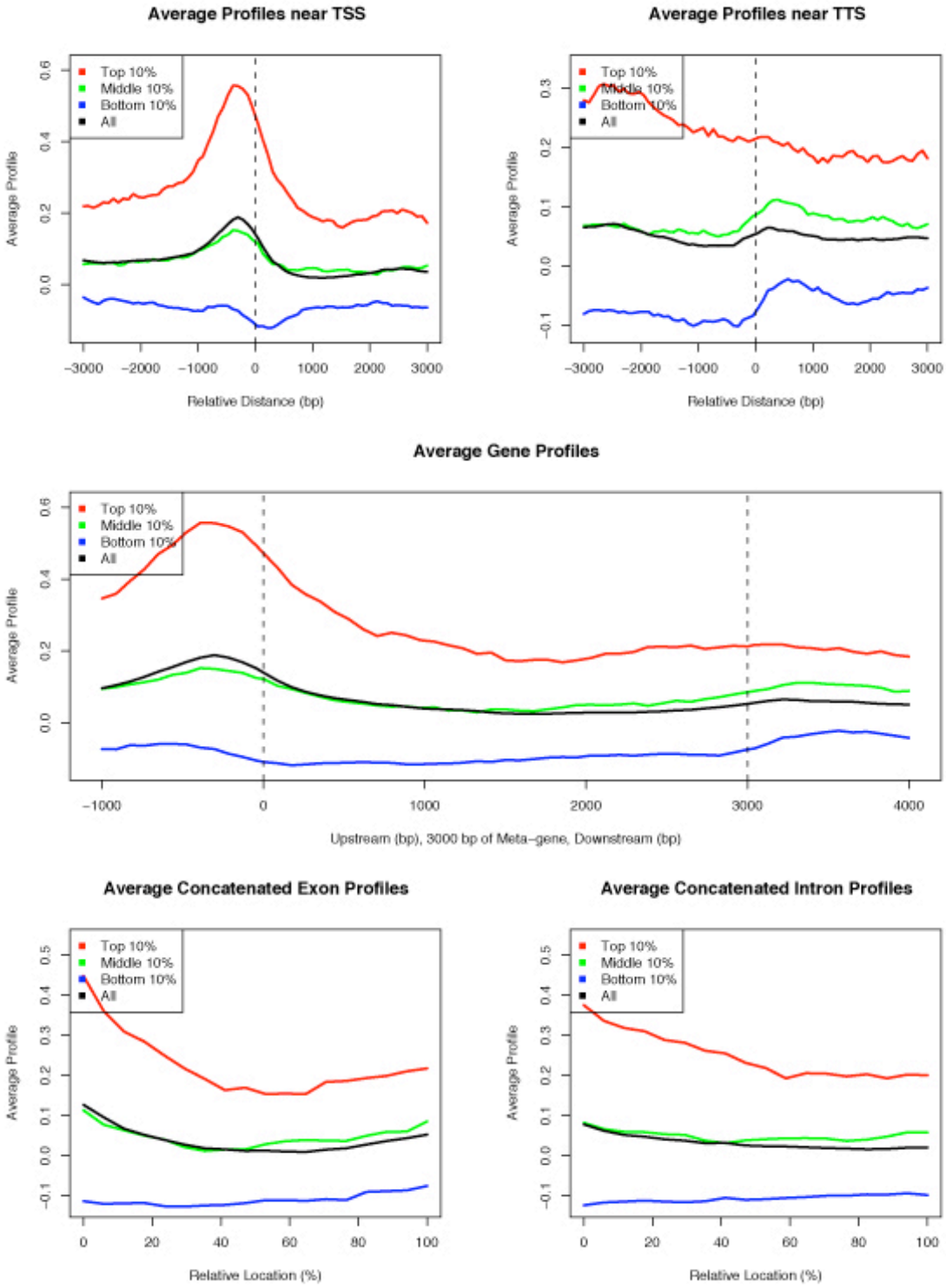



**Figure 13 The red, green, and blue colors represent the average ChIP signal profiles on top 10 % , middle 10 %, and bottom 10 % of expressed genes in C. elegans while the black represents the average profile on all genes. As observed in the prior plots, it is shown that SDC-3 highly enriches promoters.**

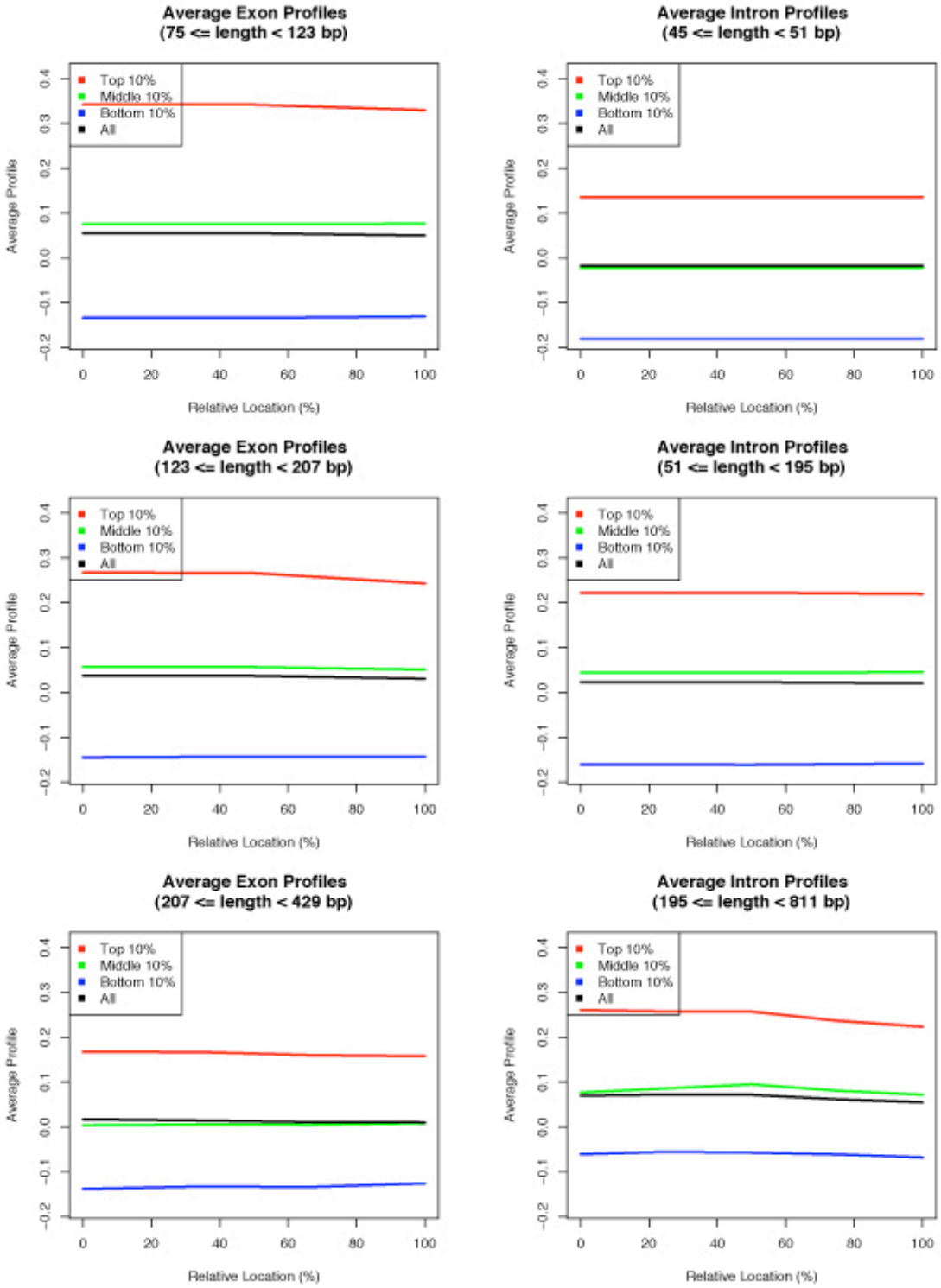



**Figure 14 The above plots illustrate the average ChIP signals over all exons and introns.**