

Software Documentation

Version 0.9

January 2019

Linden Gearing, Helen Cumming, Ross Chapman, Alex Finkel, Isaac Woodhouse, Kevin Luu, Sam Forster and Paul Hertzog

Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research

# Table of Contents

Table of Contents ii

Table of figures iv

Introduction 1

Scan 2

Enrichment analysis 4

Background gene list selection 4

Enrichment calculations 4

Proximal enrichment analysis 6

Acquisition and Installation 7

System requirements 7

Prerequisite installation 7

Windows/Mac/Linux 7

Java installation 7

Program Acquisition 7

Installing the Program 8

Windows 8

Linux 8

Mac 8

GUI workflow 9

Scan 10

Importing a gene list 10

Importing transcription factor models 11

Selecting a deficit 11

The user interface 12

Site interface 14

Transcription factor panel 15

Sliders 15

Saving data and images 16

Enrichment analysis 17

Enrichment interface 17

Enrichment plot 18

Saving data and images 19

Subsequent enrichments 19

Proximal enrichment analysis 20

Selecting sites of interest 20

Performing a new scan and enrichment 20

Command line workflow 21

General parameters 21

Required: 21

Optional: 21

Scan parameters 21

Input files: 21

Parameters: 22

Output files: 22

Enrichment parameters 22

Input files: 22

Parameters: 22

Output files: 22

References 23

# Table of figures

Figure 1: Program workflow 1

Figure 2: Transcription factor motif 2

Figure 3: Site prediction 3

Figure 4: Enrichment statistics 5

Figure 5: Proximal enrichment analysis 6

Figure 9: The start panel 8

Figure 10: Scan load box 9

Figure 11: JASPAR matrix format 10

Figure 12: The promoter panel user interface 11

Figure 13: The site interface 13

Figure 14: Options for transcription factor and site display. 14

Figure 15: The enrichment load box 16

Figure 16: Enrichment interface 17

Figure 17: Interactive enrichment scatter plot 18

Figure 18: The proximal enrichment load box 19

# Introduction

Welcome to **CiiiDER**, a software package for predicting and analysing transcription factor binding sites.

The **CiiiDER** workflow consists of two types of analysis (Figure 1):

1. Identification of potential transcription factor binding sites in regulatory regions (Scan)
2. Identification of over- or under-represented transcription factors compared to a background (Enrichment)

This user guide gives some background information about these analyses, how to interact with the graphical user interface (GUI) and run analyses from the command line.

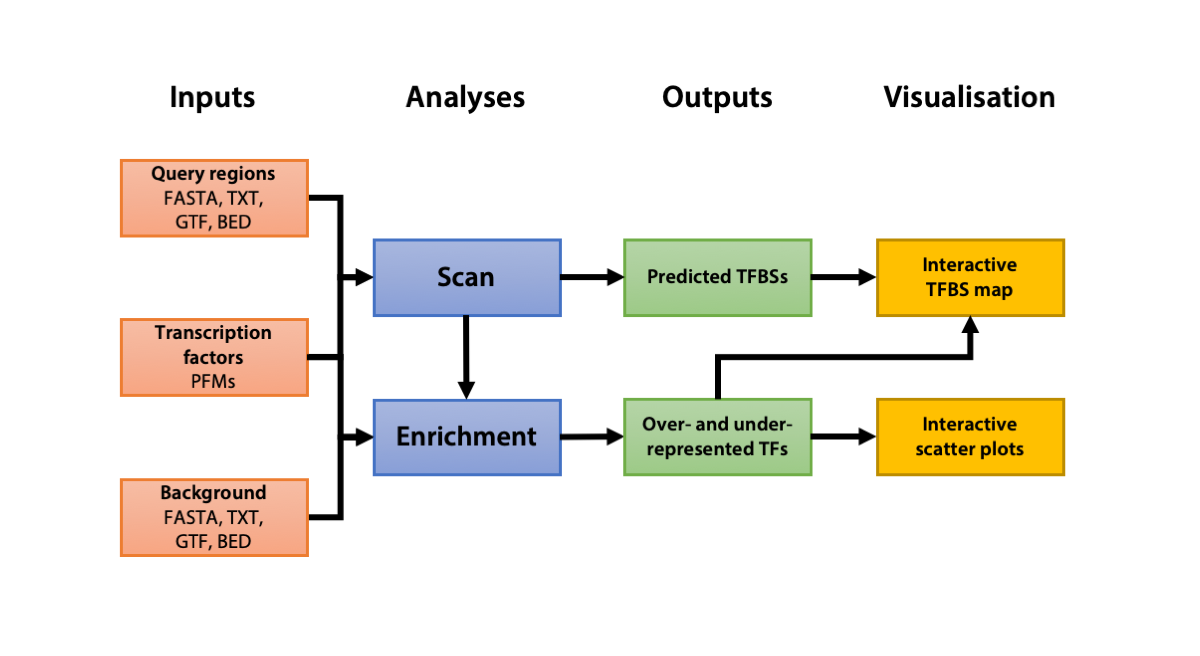


Figure : Program workflow

Analyses can be performed using the graphical user interface or from the command line.

## Scan

**CiiiDER** uses an implementation of the MATCH algorithm (Kel et al., 2003) to predict transcription factor binding sites in a query set of DNA sequences. The scan produces a map showing the location of all potential sites across the sequences.

Transcription factors are represented by position frequency matrices (PFMs; Figure 2) in either JASPAR (Khan et al., 2018) or TRANSFAC (Matys et al., 2006) format. The mapping of matrix to sequence is performed with a user-defined deficit between 0 and 1, which determines the rigour of the mapping (0 for a perfect match; 1 for no match).

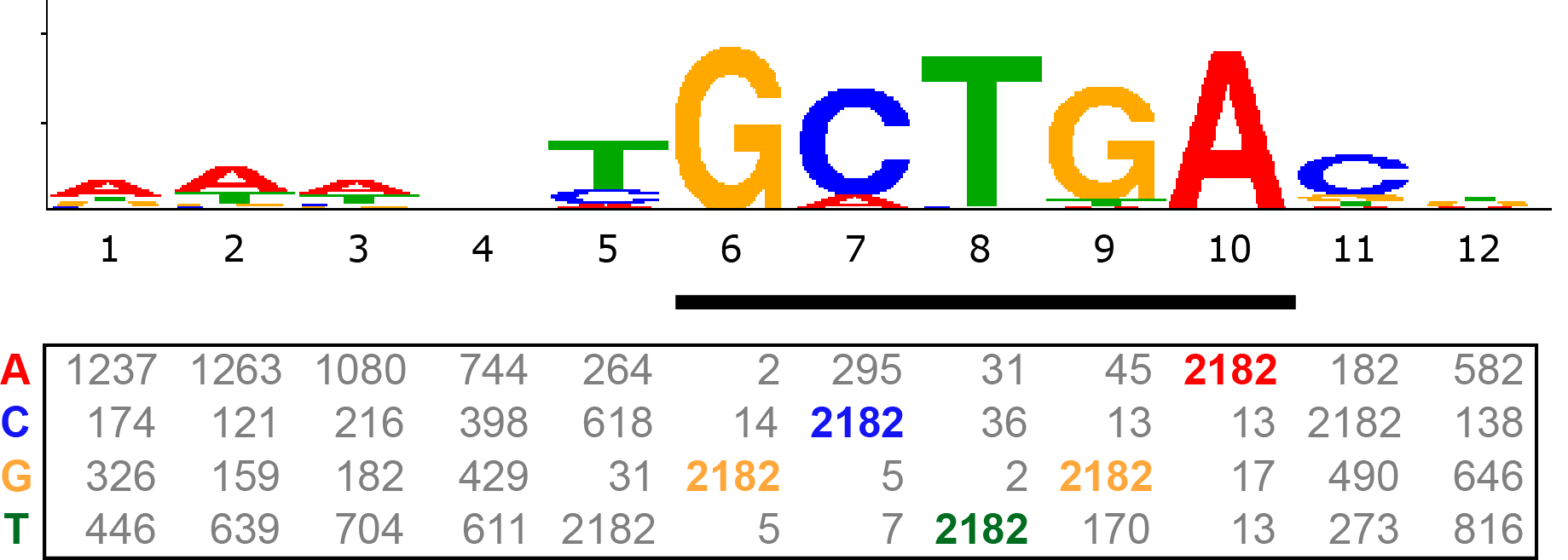


Figure : Transcription factor motif

Position frequency matrix and sequence logo representation of a Mafb transcription factor motif from JASPAR (MA0117.2), showing the highly conserved core region (black bar) and the variable flanking regions.

Transcription factors often have a core binding region that is highly conserved, which is flanked by areas of higher variability. **CiiiDER** defines the core as the five most conserved consecutive bases (which is calculated using the sum of information vector values; see (Kel et al., 2003)).

To predict binding sites, **CiiiDER** splits sequences into overlapping regions of five bases and first matches these smaller regions with the core of the transcription factor model. If a core match is found, then the window is increased to incorporate the full length of the transcription factor binding site matrix and a whole matrix match is calculated. A potential site must have a core match and matrix match below the deficit cut-off.

**CiiiDER** performs these calculations simultaneously for DNA sequences using threading techniques to employ multiple processors, which greatly increases the overall analysis speed.

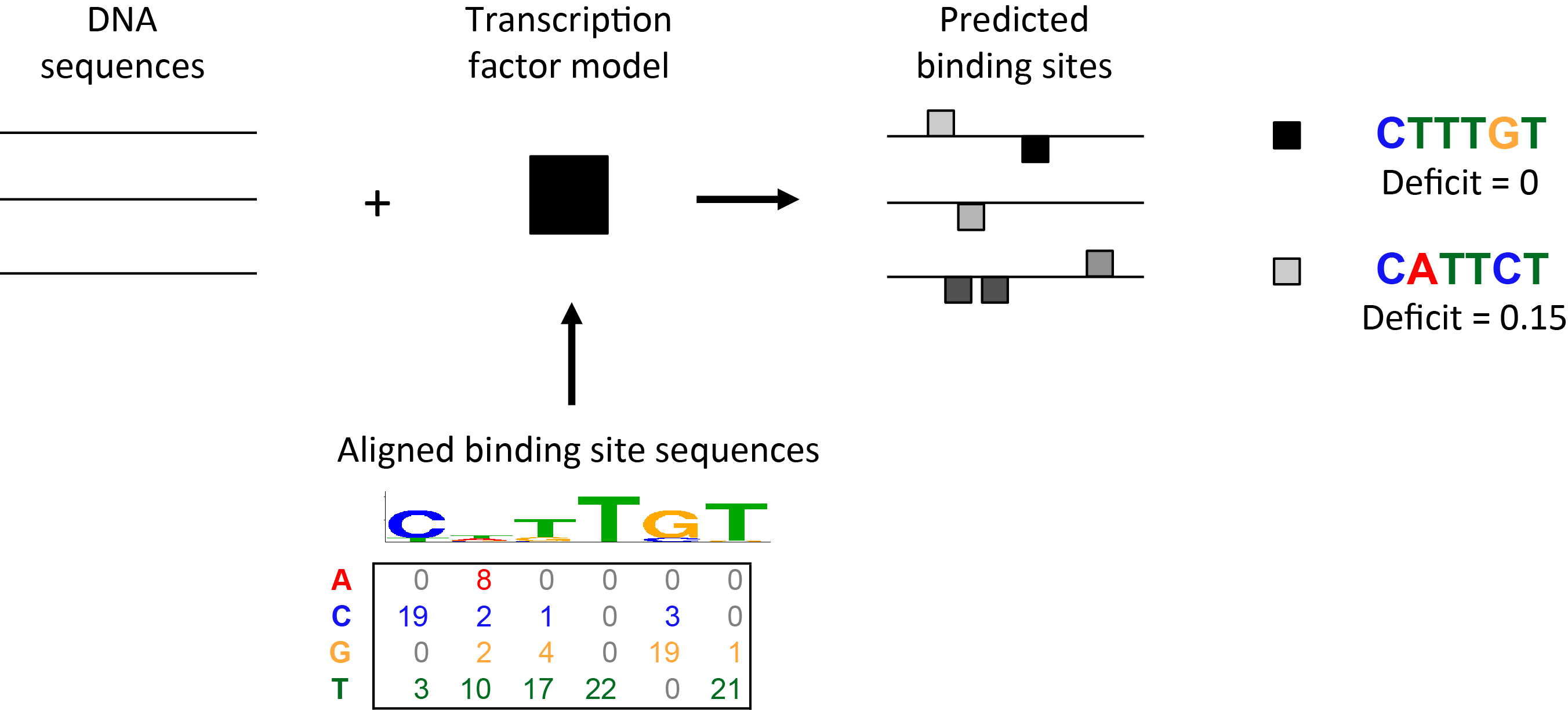


Figure : Site prediction

The JASPAR Sox10 transcription factor PFM (MA0442.1) is shown.

## Enrichment analysis

Although **CiiiDER** can efficiently identify putative transcription factor binding sites, there are many binding sites present in any given DNA sequence. Some of these are false positives, randomly generated in the DNA sequence, that are not bound by the transcription factor. Others may be true binding sites in certain biological contexts, but not others. By comparing the predicted binding sites present in a list of co-regulated genes to those found in an appropriate background list of genes, **CiiiDER** can identify which transcription factors are over- or under- represented and are therefore may be playing important roles in regulating the genes of interest.

### Background gene list selection

Selection of the background gene set is important as it can vastly alter the specificity of the answers that **CiiiDER** can provide. Multiple enrichment analyses can be performed using different background to ask different questions.

It is best practice to ensure that the background is as close to the co-regulated gene set as possible. As an example, when analysing genes that change in macrophages in response to bacterial infection, the best background would be genes that were expressed in macrophages but were unchanged during infection; the transcription factors identified are likely to be solely involved in regulating those genes altered in response to infection. A background containing genes randomly selected from the whole genome would be less appropriate, as the enriched transcription factors identified would contain those required for normal gene expression in the macrophage as well as those factors involved in response to infection.

If the co-regulated gene set comes from a microarray or RNA-seq experiment, then it is best practice to construct the background from genes that are expressed and have very low fold change.

Larger gene sets increase the power of the calculations. In general, we recommend using at least 100 genes of interest. The background gene list should be at least as large as the size of the co-regulated list.

### Enrichment calculations

#### P-values

Significantly over- and under-represented transcription factors can be found using two statistics. The main test is a Fisher’s exact test, which gives the gene coverage *P*-value (Figure 4A). This test compares the numbers of sequences that are bound and unbound by a transcription factor.

This statistic is important as it distinguishes situations where a single gene in the co-regulated gene set contains many copies of the transcription factor and the majority of the genes do not contain site for the transcription factor and the more interesting situation where the majority of the genes contain binding sites.

*P*-values are calculated for each transcription factor at different deficit cut-offs.

The other test is a Mann-Whitney *U* test to compare the distributions of the number of sites in the search and background gene set (Figure 4B). The resulting *P*-value is referred to as the site count *P*-value.

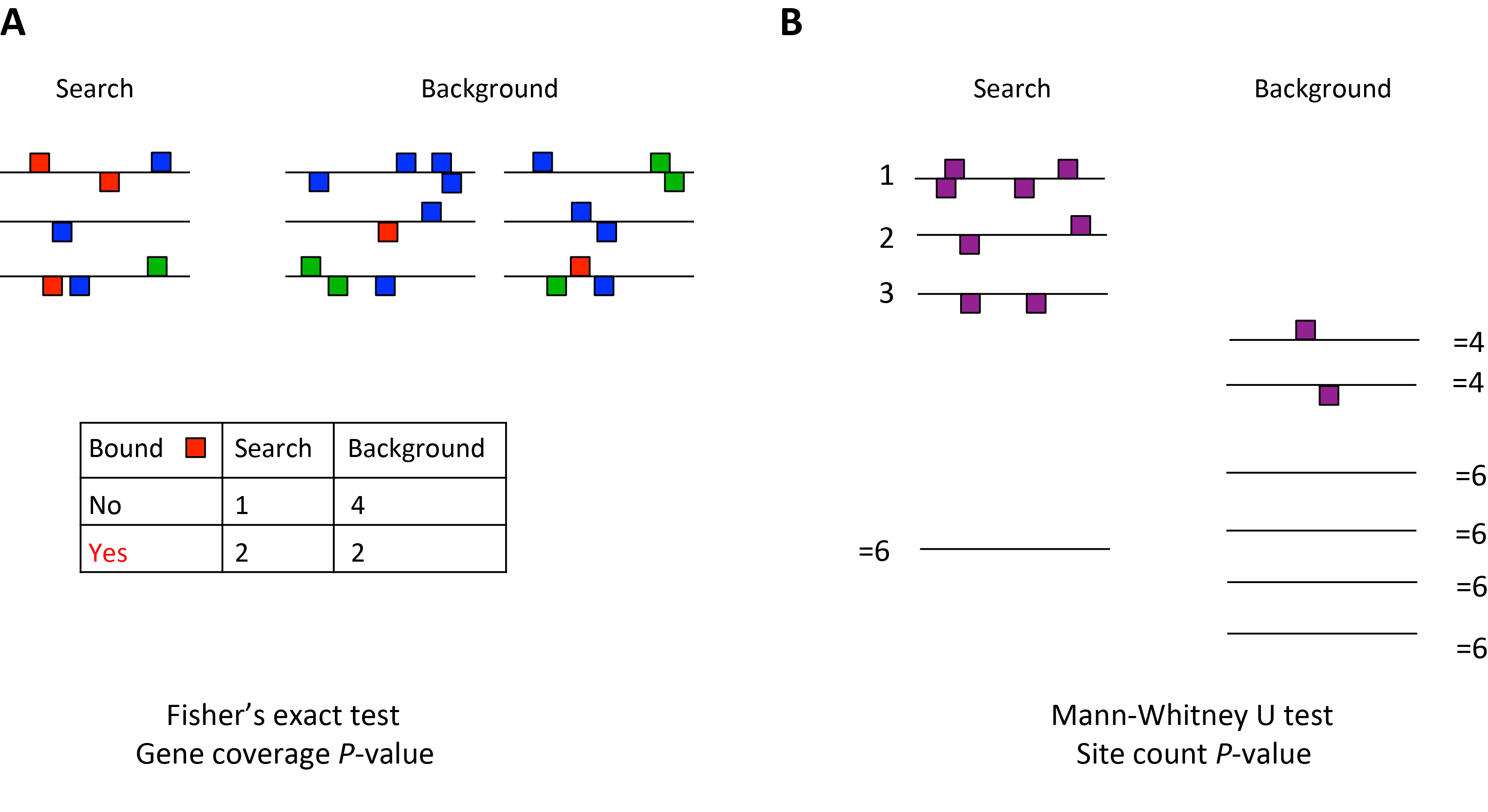


Figure : Enrichment statistics

Two statistics can be used to compare search and background genes. The gene coverage P-value uses a Fisher’s exact test on the numbers of bound and unbound genes. The site count P-value uses a Mann-Whitney U test on the distribution of sites per gene.

#### Enrichment and proportion bound

Two additional calculations are made for the enrichment results, either at a defined deficit or at the deficit that gives the most significant gene coverage *P*-value. If a given transcription factor has binding sites in out of genes and in out of background genes then:

This is greater than zero if the transcription factor is over-represented, occurring in a greater proportion of search genes than of background genes, and less than zero if the transcription factor is under-represented. The larger the value, the greater the level of over- or under-representation. Note that 1/2 is added to the numerators and denominators to avoid zeroes.

This is equal to zero if the transcription factor site is predicted in every gene in both gene sets, otherwise it is less than zero (e.g. −1 if sites are present in 1/2 genes, −2 if 1/4 genes).

### Proximal enrichment analysis

It is possible to extend the analysis described above to examine whether transcription factor sites are enriched near other transcription factor sites (Figure 5), which may indicate that they are acting co-operatively. Transcription factor sites of interest are identified in a search gene set and a background gene set. The regions surrounding these sites are then scanned with different transcription factor models and an enrichment analysis is performed to find over- or under-represented factors in these windows. Note that this analysis is currently only available from the GUI, not the command line.

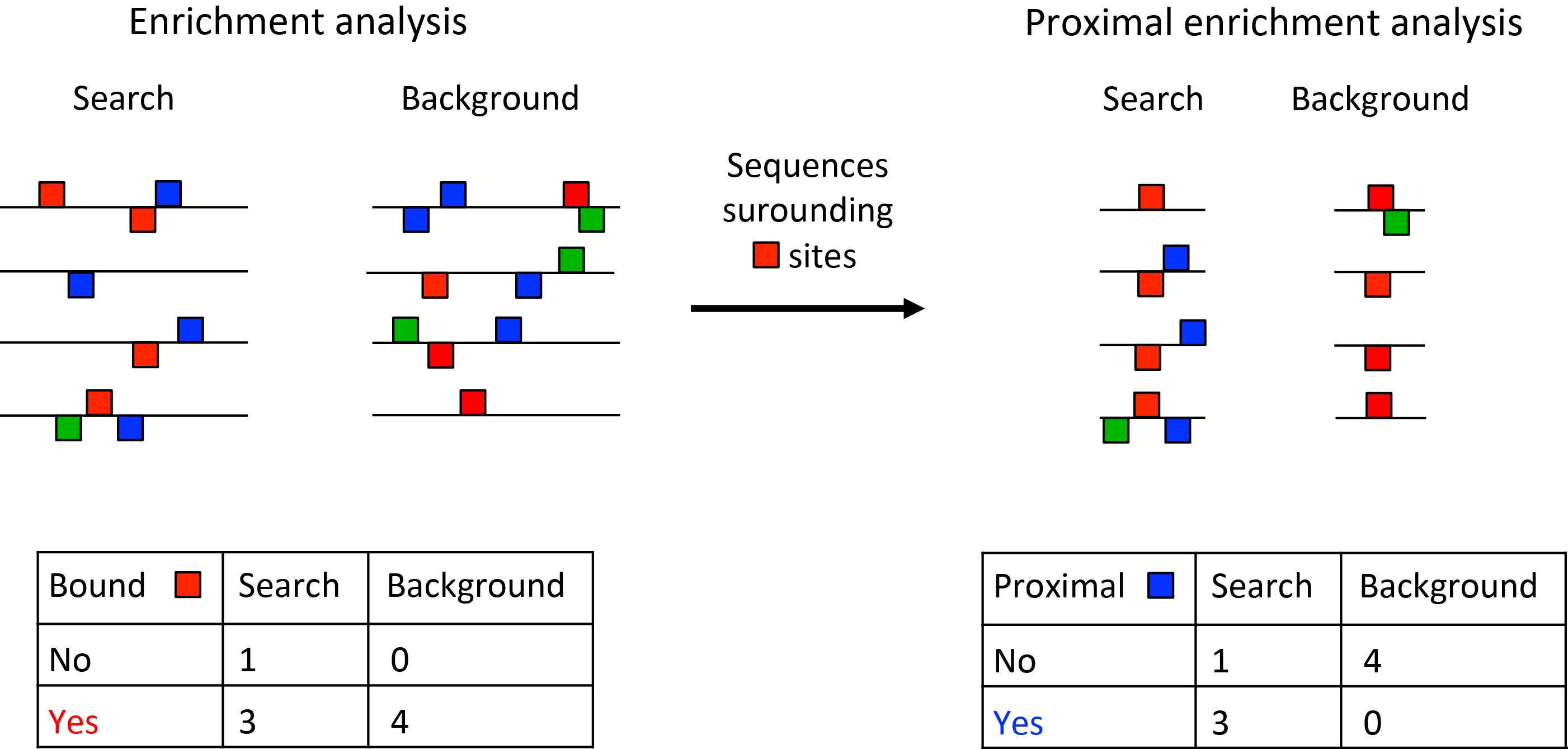


Figure : Proximal enrichment analysis

Regions surrounding transcription factor binding sites of interest are assessed for enrichment of other transcription factors, compared to a background.

# Acquisition and Installation

## System requirements

**CiiiDER** works best on desktop computers.

## Prerequisite installation

In order to run the **CiiiDER** JAR file, the computer must have the Java Runtime Environment (JRE) (most computers will already have this installed). **CiiiDER** requires JRE version 1.6 or above. To determine which version is installed on your computer, try the Java website version test:

<http://www.java.com/en/download/installed.jsp?detect=jre&try=1>

Note that this test does not work with some browsers (such as Google Chrome).

### Windows/Mac/Linux

Alternatively, if you are comfortable using the command line then simply type into your terminal or command line window:

java –version

If Java is installed you will see the output:

java version "1.8.0\_60"

Java(TM) SE Runtime Environment (build 1.8.0\_60-b27)

Java HotSpot(TM) 64-Bit Server VM (build 25.60-b23, mixed mode)

To open command prompt in windows: Open command prompt by pressing ‘Windows’ + R shortcut to bring up the ‘Run’ dialogue, then type in ‘cmd’ and click OK. In command prompt, type “java -version”.

### Java installation

If you do not already have JRE or the correct version of JRE running on your machine, then it can be downloaded from the Java website ([www.java.com/en)](http://www.java.com/en)), along with installation instructions.

## Program Acquisition

Download from [www.ciiider.org](http://www.ciiider.org). Mouse and human genomes can also be pre-packaged with **CiiiDER**.

# GUI workflow

This section provides a guide to using the graphical user interface. Open the **CiiiDER** program (or run the JAR file) and press the “START” button to commence a new analysis.

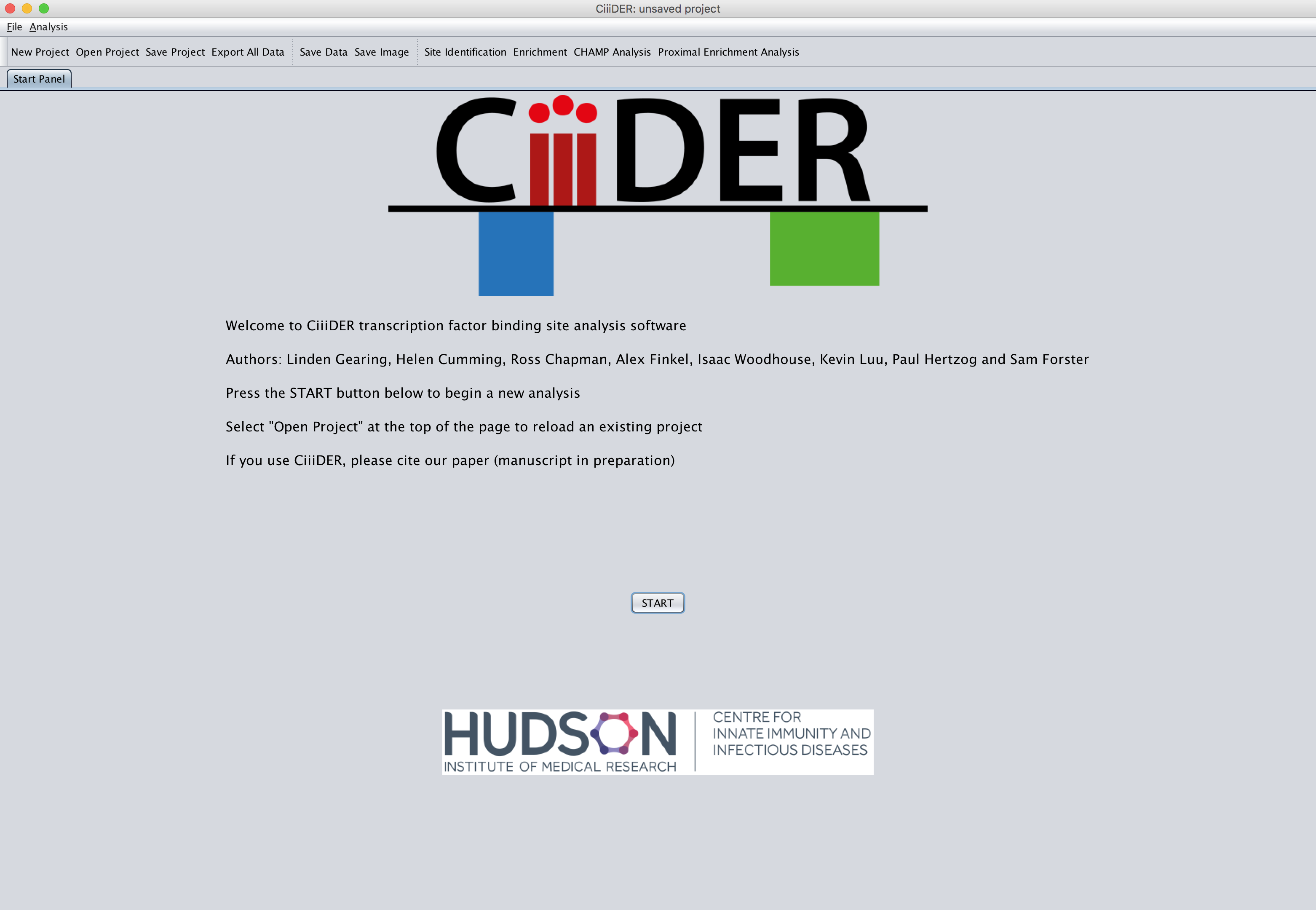


Figure : The start panel

## Scan

A new dialog box will open requesting a gene list, transcription factors and a deficit. The “Run Scan” button will begin the site prediction for the chosen gene list; the “Run Enrichment” button can be used to queue an enrichment analysis to immediately follow this site prediction, but this is not essential (see page 20).



Figure : Scan load box

### Importing a gene list

Gene lists can be loaded from a file or pasted into the text box. FASTA sequences, gene symbols, Ensembl IDs, GTF and BED files are all acceptable. DNA sequences in FASTA format require no further input; otherwise, the program requires a genome to extract DNA sequences flanking the transcription start site of the gene (or sequences surrounding the start position given in a GTF file or the midpoint of BED file regions). The default sequence extraction takes 1,500 bases upstream and 500 bases downstream.

Available genomes are stored in the Genomes folder within the **CiiiDER** program folder and appear in a dropdown menu. The most recent Ensembl genomes and annotations for key species will be available to download from the **CiiiDER** [website](http://www.ciiider.org). If no files are present in the Genomes folder or the “Advanced selection…” option is chosen, a genome file (FASTA format) and an annotation file (GTF or GLM format) must be loaded (these can be obtained from the Ensembl FTP website).

### Importing transcription factor models

PFM transcription factor models are required to perform the scan. **CiiiDER** provides the JASPAR CORE non-redundant vertebrate transcription factors (Mathelier et al., 2016) and PFMs from a large SELEX experimental dataset (Jolma et al., 2010). These files are stored in the Matrices folder within the **CiiiDER** program folder and are available for selection in a dropdown menu.

Alternatively, any file can be loaded containing transcription factors in JASPAR or TRANSFAC format. Other transcription factors can be obtained from the JASPAR website (<http://jaspar.genereg.net/html/DOWNLOAD/JASPAR_CORE/pfm/>), including redundant vertebrate models and PFMs for different phyla. The current TRANSFAC models require a licence, although older versions are freely available (<http://www.gene-regulation.de/pub/databases.html>).

#### Other transcription factor models

Transcription factors from other sources may need to be reformatted to work with **CiiiDER**. PFMs in JASPAR format are preceded by annotation information beginning with the “>” greater-than symbol (as for FASTA sequences), which is immediately followed by a unique transcription factor identifier, then a space (or tab) and the transcription factor name.

Each row of the PFM represents a base (ordered A, C, G and T) and each column represents the count or frequency of that base at that position within the binding site (columns are separated by a space or a tab). Only the numbers are essential; the square brackets and the bases shown are optional.

>MA0004.1 Arnt

A [ 4 19 0 0 0 0 ]

C [16 0 20 0 0 0 ]

G [ 0 1 0 20 0 20 ]

T [ 0 0 0 0 20 0 ]

Figure : JASPAR matrix format

### Selecting a deficit

In order to predict whether a sequence contains a site for a transcription factor, **CiiiDER** compares the DNA sequence to the PFM and generates a score of similarity. If the DNA sequence matches the PFM perfectly then the deficit value is 0. Transcription factor binding sites are variable and binding sites rarely match the models perfectly, so it is useful to allow imperfect matches as well.

Any deficit cut-off between 0 and 1 is accepted. For general scans we recommend using deficit values between 0.1 and 0.2, especially for large datasets (with hundreds of genes and transcription factors). It is important to note that for higher deficits, it is less likely that a sequence contains a true site, so it is a balance to maximise the number of true positive sites, while restricting the number of false positive sites.

## The user interface

Once the analysis is complete, the promoter panel will be displayed. The basic layout of the user interface is the same for the different stages of analysis. At the top of the screen there are a number of options to help navigate the project and save data and images at any stage during the various analyses.

It is interactive allowing the user to change parameters and perform new analyses quickly and easily, as well as compare multiple analyses by clicking between separate panels.

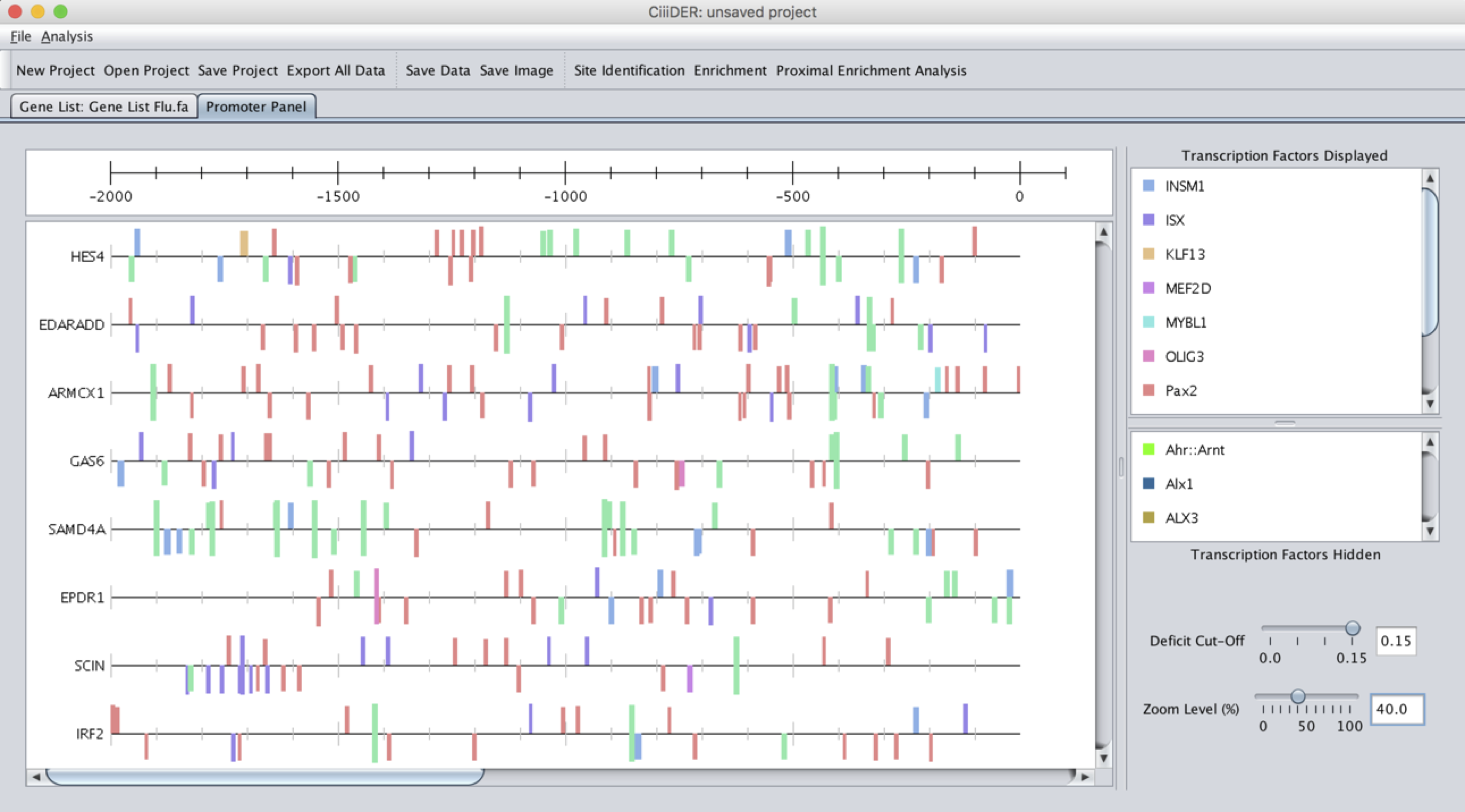
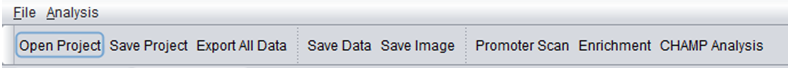


Figure : The promoter panel user interface

#### Menus and toolbar

The toolbar provides the user with the tools to manage their projects, to save data and images, and begin new analyses.



File. Contains options to begin new projects, open existing projects, to save data and images as well as alter general properties.



New Project begins a new project.

Open Project opens an existing saved project from a file chosen by the user.

Save Project saves the project in its current form in a “.cdr” file that can be opened again later.

Export Project saves all the data from the existing project in its current state, including all images and analysis reports and parameters used in the analysis. Importantly it does not save the project.

Save Image saves the current image.

Project Information shows files and parameters used in the current project.

Settings this is also where general visual properties are set including the number of computer processes you wish to use (by default this is one fewer than the number available), so this option is only useful if you wish to limit the processing power CiiiDER can access). The user can also specify how tall each promoter appears in the image (the default value is set to 30 pixels) and how many transcription factors can be visualised at the one time (the default value is set to 10).

Exit closes the program, with an option to save the current project

Several of these options are available as short cuts on the toolbar, as well as links to commence further analyses, Scan and Enrichment, as well as Proximal Enrichment Analysis (see later).

#### Tab Panels

These tabs represent each of the current analyses that are part of the project. Note that although there can be multiple Enrichment and Proximal Enrichment analyses, there can be only one gene list and one promoter panel. Performing site identification again (with new matrices or with a different deficit) removes all other panels.

#### Gene List Panel

This panel includes the list of genes and their sequences that are currently being used in the site identification and enrichment analyses. If the user has simply provided a list of gene symbols then they can view the complete DNA sequences for the regions they have specified. At the bottom of the list are any genes for which the sequence could not be found.

### Site interface

This is where the predicted transcription factor binding sites are displayed on the current gene list (Figure 13). The top of the panel contains a scale bar to show the position of each transcription factor binding site, either relative to the position of the transcription start site (if applicable) or the 5’ position from the end of the sequence. It is interactive to allow the creation of an appropriate image for publication or display.

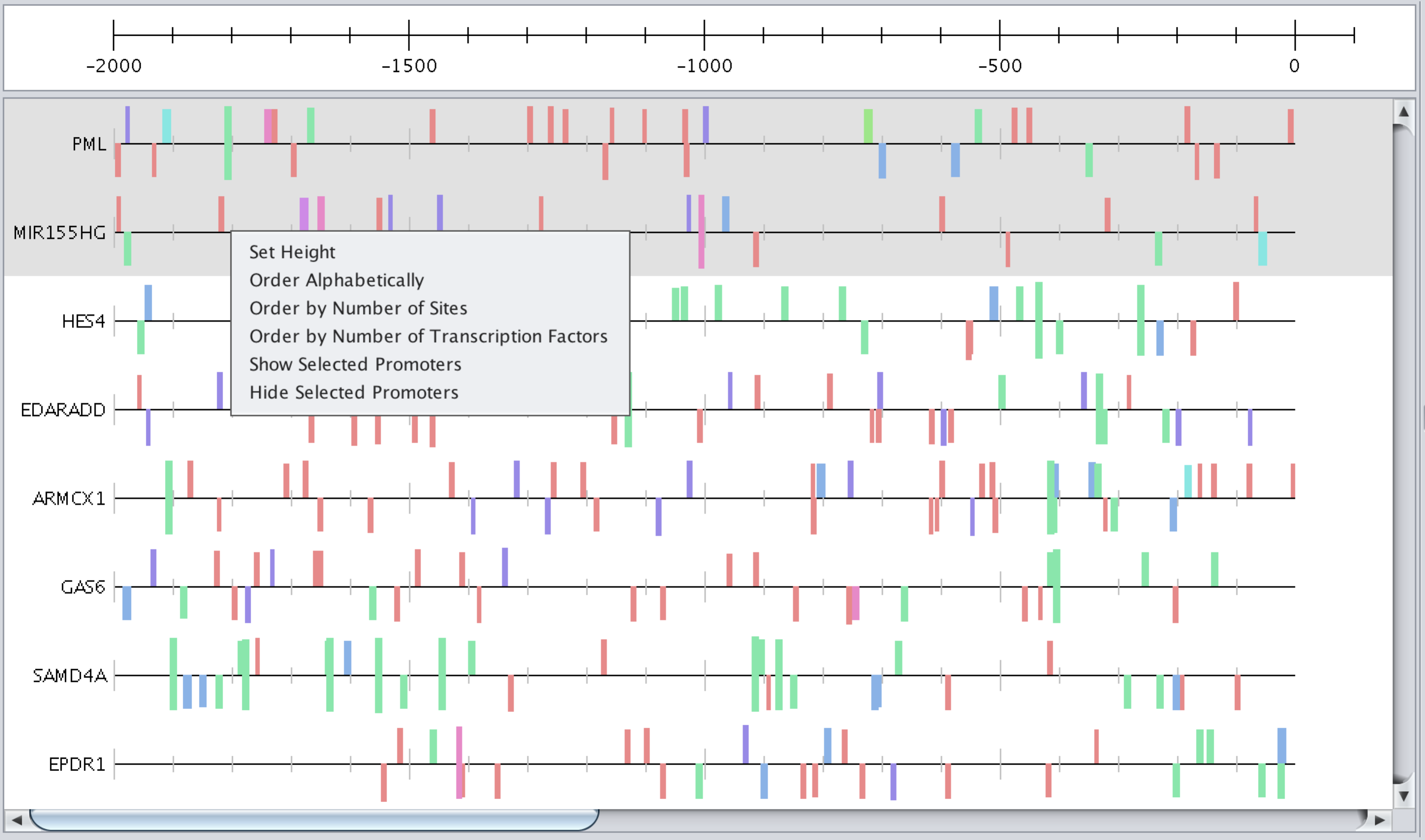


Figure : The site interface

There are several options for displaying or reordering sequences:

* Click – select a sequence
* Press and drag – move a single sequence up or down
* Ctrl or shift click – select multiple sequences
* Right click – show options
  + Set height
  + Order alphabetically
  + Order by number of (displayed) sites (descending)
  + Order by number of (displayed) transcription factors (descending)
  + Show or hide selected promoters
  + Show all promoters

### Transcription factor panel

This is where the predicted transcription factors are displayed (Figure 14). The top section contains the transcription factors that are currently being shown on the site interface, while those below are currently hidden from view. Ten transcription factors are displayed by default (this can be changed under File: Properties). Double-clicking on a transcription factor sends it from one list to the other.

Several options are available by right-clicking on a displayed transcription factor:

* Choose colour
* Hide – moves the TF to the hidden list
* Hide all other transcription factors – only sites for the selected transcription factor are displayed
* Save genes with this transcription factor – a text file containing the names of all genes containing predicted binding sites
* Show only promoters with this transcription factor – to show all promoters again, right click on the site interface

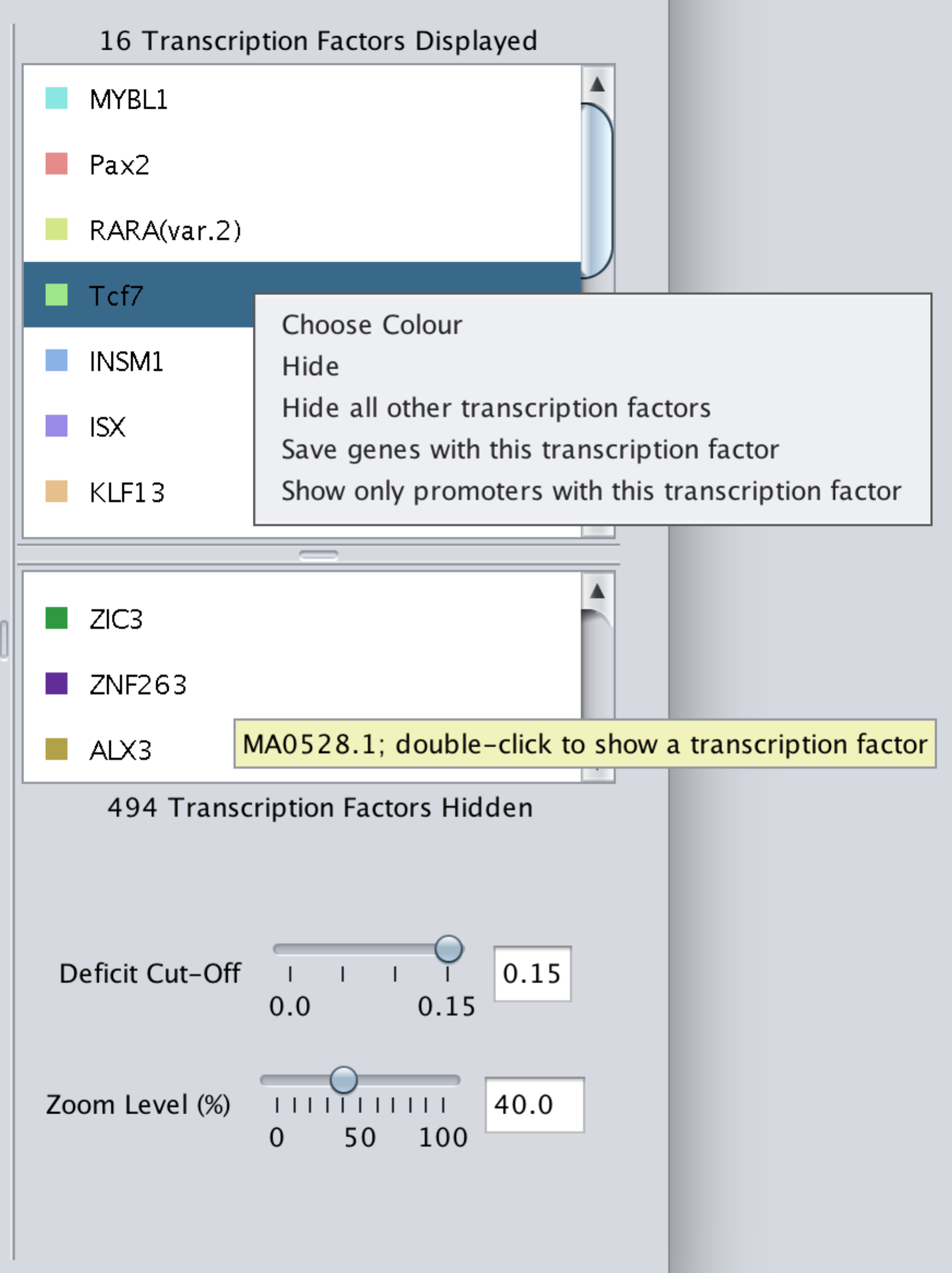


Figure : Options for transcription factor and site display.

### Sliders

This transcription factor panel contains two sliders and text boxes for changing the image (Figure 14). For each slider the user can enter the value in the text box or select on the slider using the mouse.

Zoom level slider: This changes the size of the promoters as they appear on the screen allowing the user to see the overall patterns or zoom in to closely inspect a single site.

Deficit cut-off slider: This slider can be used to restrict the displayed transcription factor sites to those with a lower deficit, i.e. those sites that better match their transcription factor model. Note that a new scan must be performed to view sites at a higher deficit.

### Saving data and images

The “Save Data” and “Save Image” buttons are used to save information related to the currently selected panel. After a scan, the positions and scores for every predicted site can be saved in CSV format and an image of the current state of the site interface can be saved in JPG, PNG or GIF format.

## Enrichment analysis

To begin an enrichment analysis, select the “Enrichment” button in the toolbar; this opens the enrichment load box for selecting a background gene list (Figure 15). This background can be provided in the same formats as the current gene list (see the scan load box). If necessary, the genome, upstream and downstream parameters will be populated with the previous values.

It is possible to enter a name for the analysis, since multiple enrichments may be run within the project; by default, the analysis will be named after the background gene list file or numbered sequentially. At this stage a gene coverage enrichment *P*-value can also be set, which determines the transcription factor sites that will be displayed (the default value is 0.05, but this can also be changed later). The “Run” button will commence the analysis.

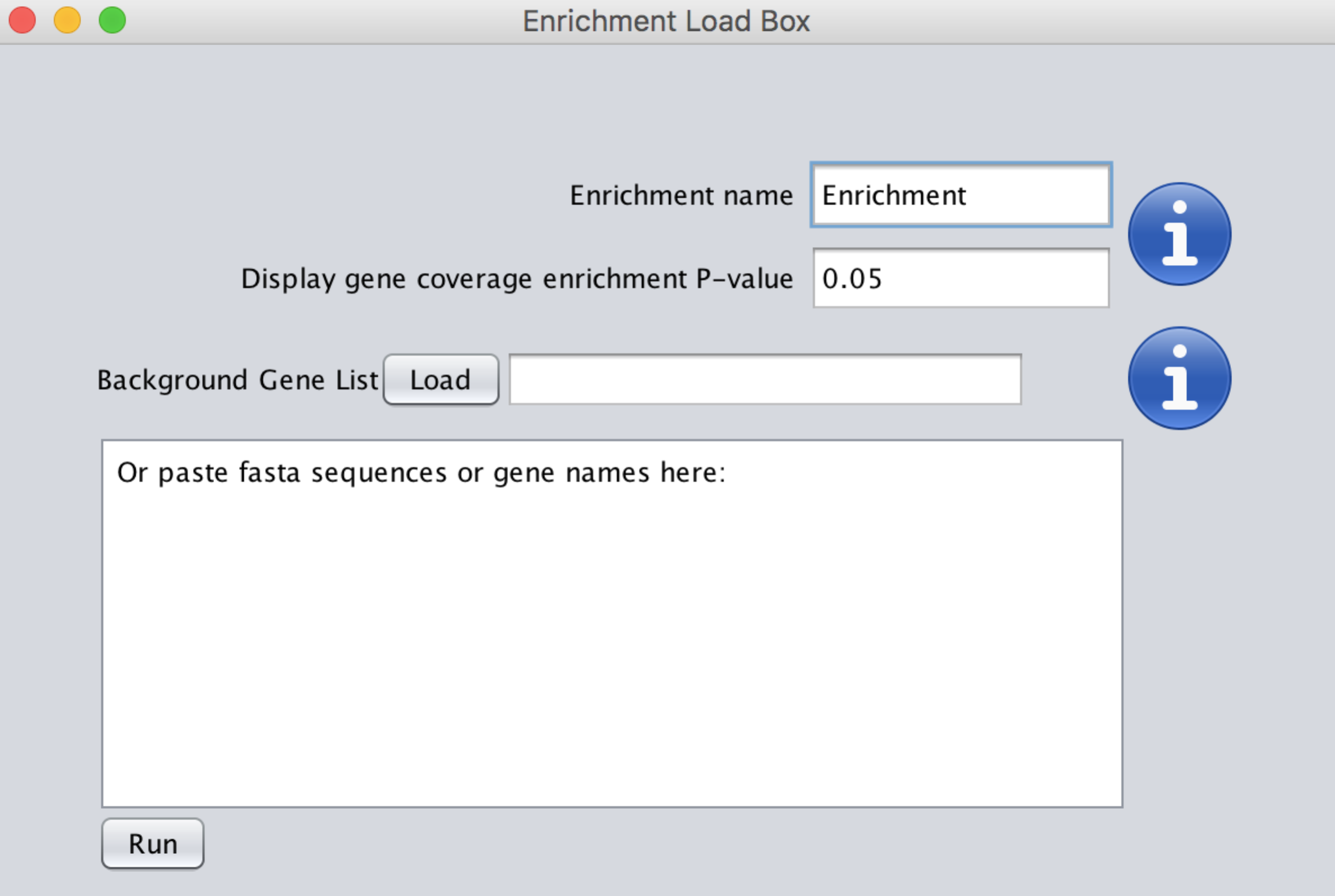


Figure : The enrichment load box

Used for supplying a background gene list for an enrichment analysis.

### Enrichment interface

The enrichment interface is very similar to the scan interface, with all the same options; however, there is an extra set of sliders and check boxes (Figure 16) that are used to filter the sites that are displayed.

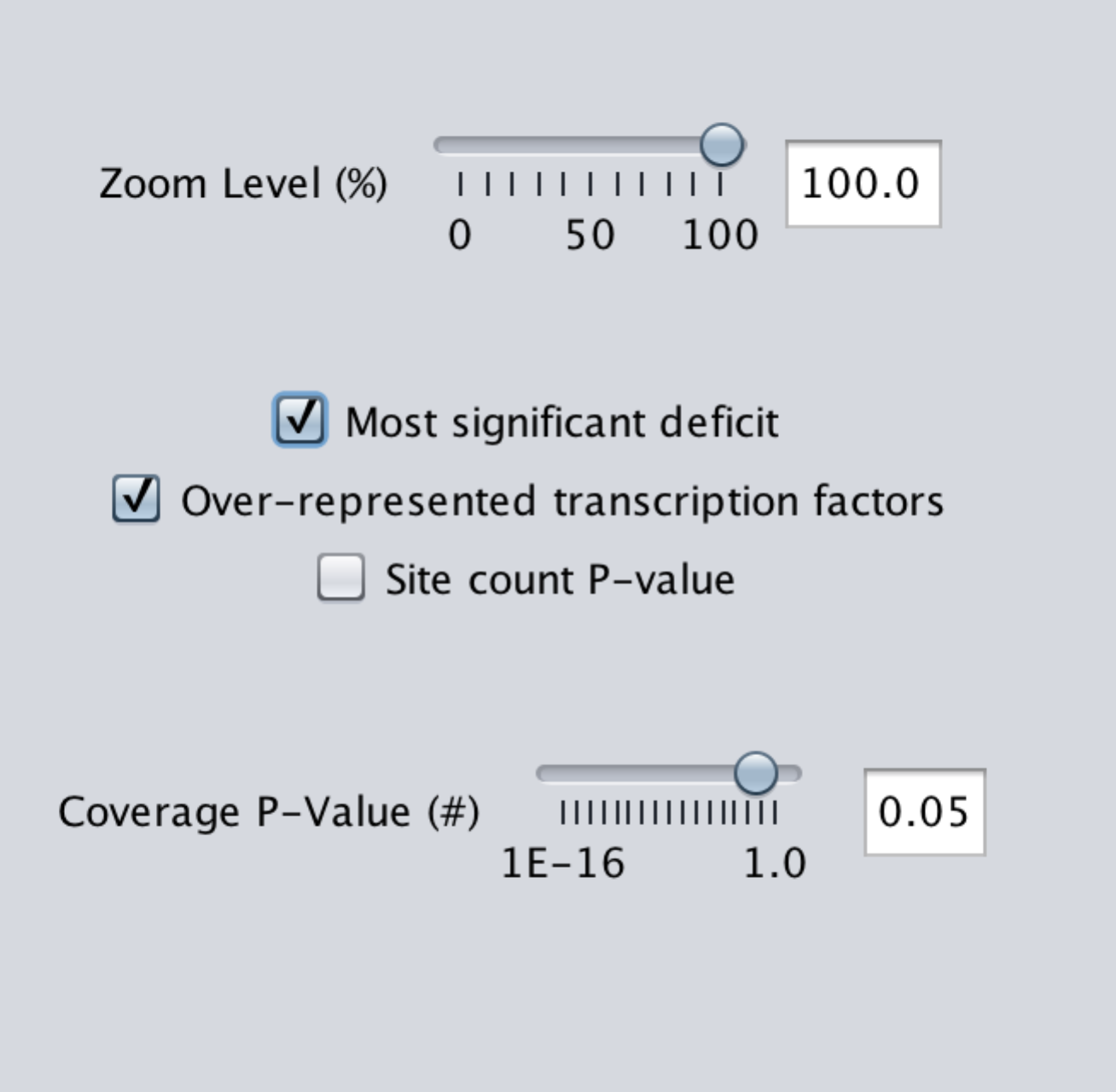
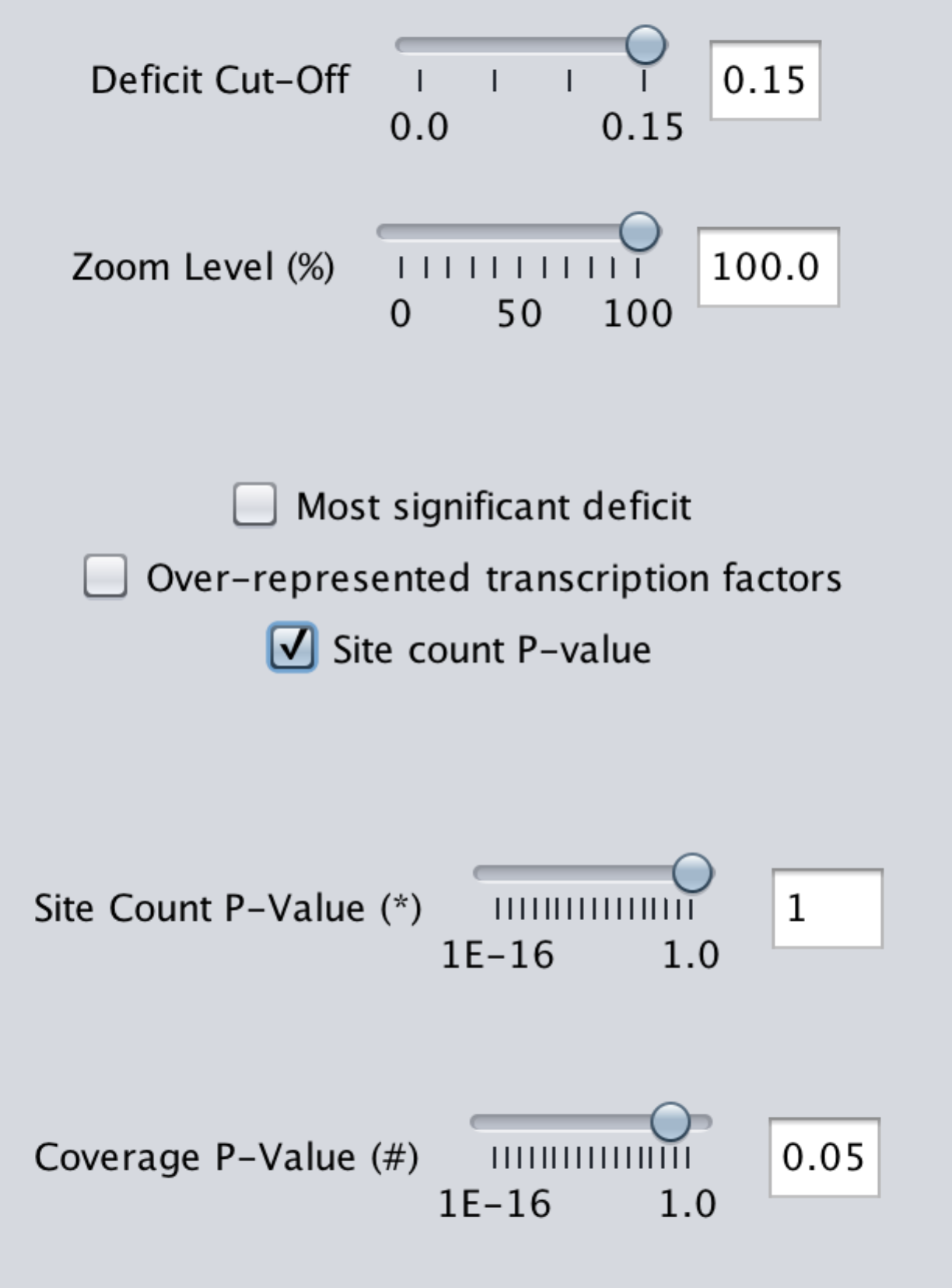
 

Figure : Enrichment interface

The default enrichment interface and optional sliders.

Most significant deficit: When this check box is ticked, all transcription factors are displayed at the deficit at which they are most significantly enriched (minimising the gene coverage *P*-value) and the deficit cut-off slider disappears. This is the default.

Over-represented transcription factors: Transcription factors may be over-represented or under-represented in the search compared to the background. By default, only over-represented factors are displayed.

Coverage P-value slider: Transcription factors will be displayed if their gene coverage *P*-Value is less than the chosen value. (For descriptions of *P*-values see page 4.)

Site count P-value slider: When the Site Count *P*-value check box is ticked, transcription factors must also meet the site count *P*-value threshold.

### Enrichment plot

The enrichment results for all transcription factors can also be viewed as an interactive HTML graph, created using the plotly javascript API (Figure 17). These plots show log2(Enrichment) versus Average Log2(Proportion Bound).

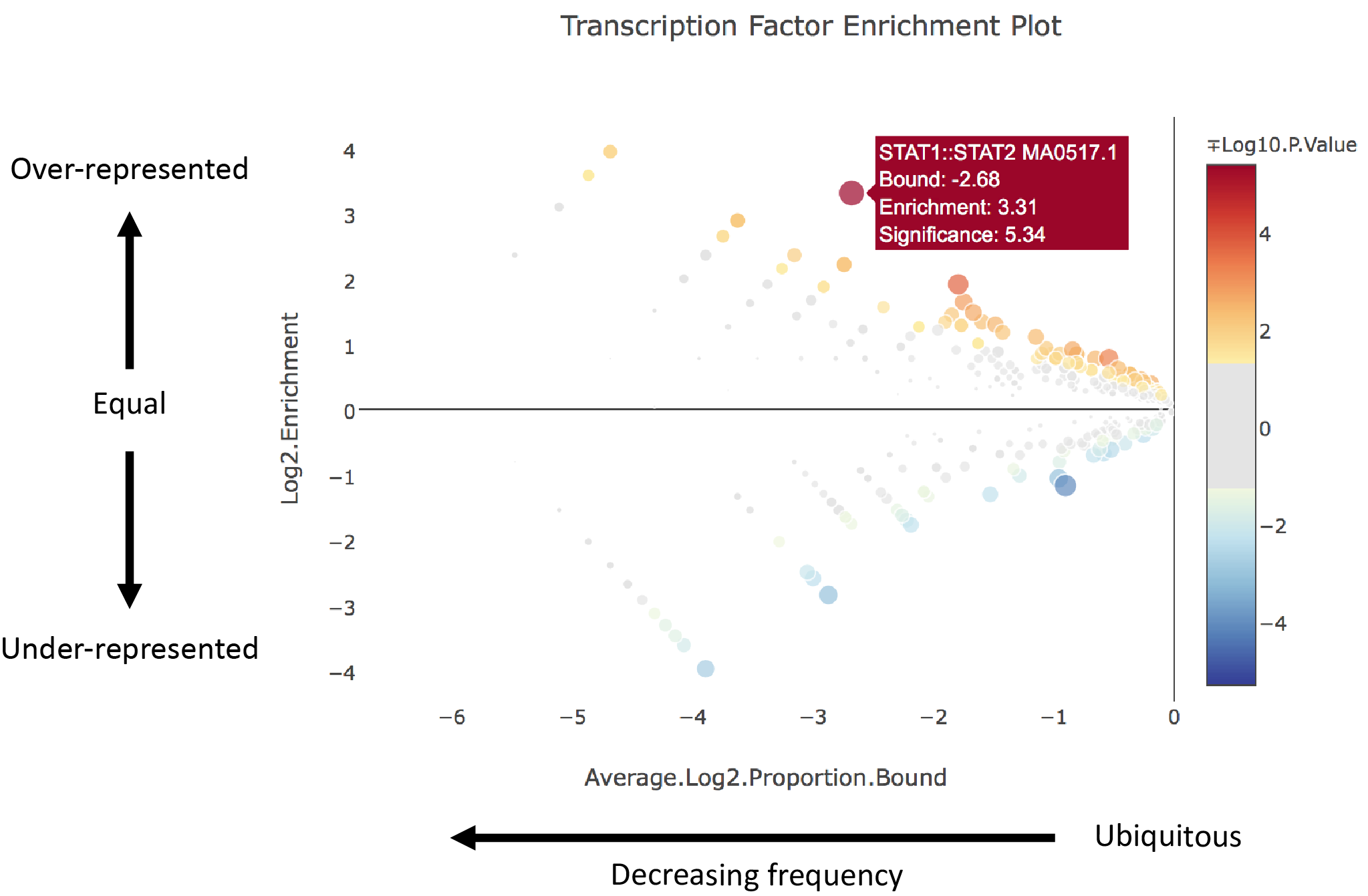


Figure : Interactive enrichment scatter plot

Illustrates the degree of over- or under-representation of each transcription factor and the proportion of genes that contain binding sites. Transcription factors are coloured according to their gene coverage P-value and whether they are over- or under-represented (while transcription factors that do not meet the chosen P-value threshold are in grey); the size of each point is also proportional to log10(P-value). Hovering over each point displays its annotation information.

### Saving data and images

The following data can be saved from the enrichment panel in CSV format:

* Statistics at the most significant deficit (an HTML graph is also saved)
* Statistics at the current deficit (an HTML graph is also saved)
* Information about the currently displayed (and hidden) sites
* Information about all predicted binding sites

An image of the current state of the site interface can be saved in JPG, PNG or GIF format.

### Subsequent enrichments

To run another analysis using a different background, click the “Enrichment” button in the toolbar again. The results for each enrichment are stored in different tabs.

## Proximal enrichment analysis

The “Proximal Enrichment Analysis” button on toolbar launches the proximal enrichment load box (Figure 18).

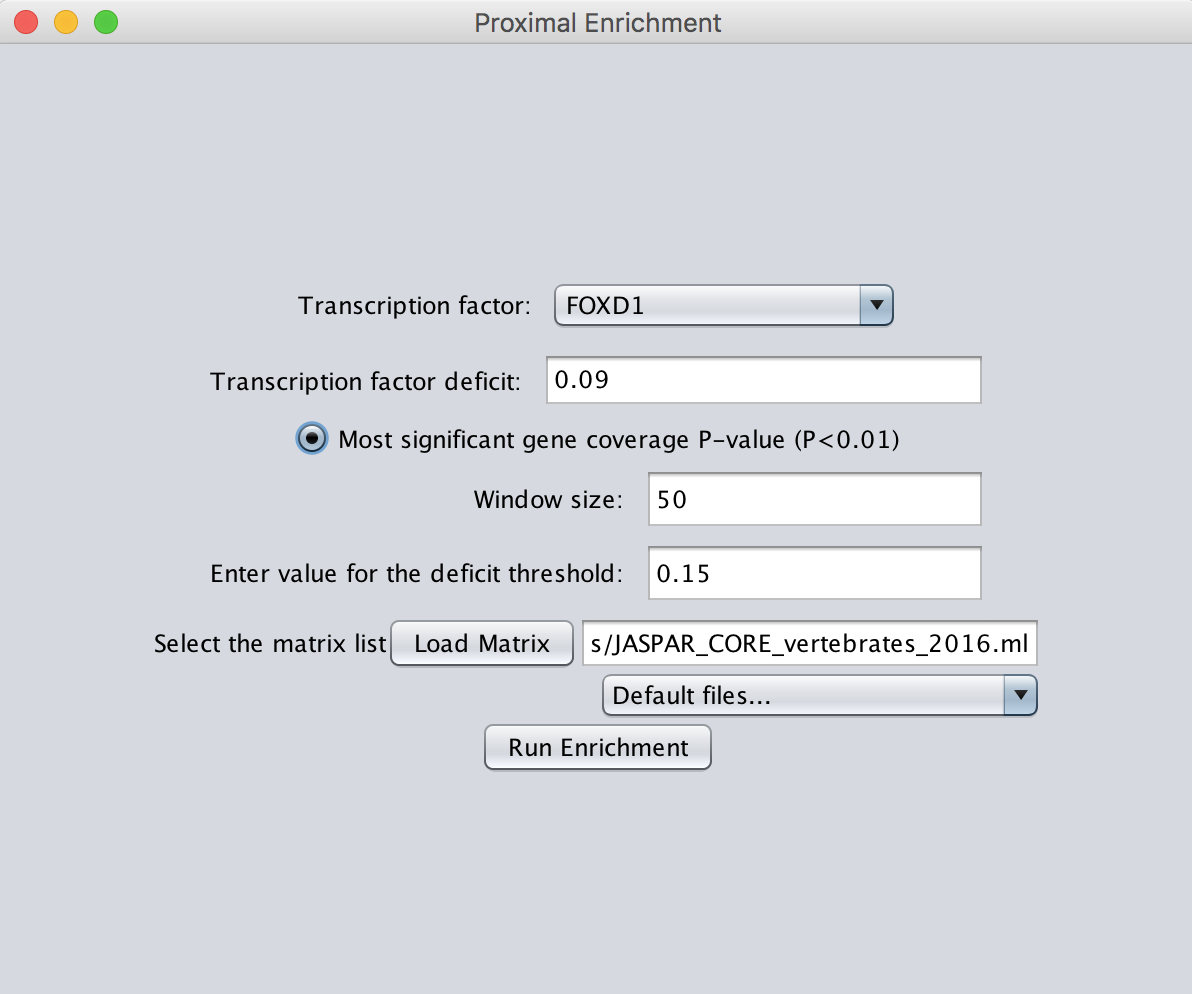


Figure : The proximal enrichment load box

### Selecting sites of interest

The first step is to choose a transcription factor of interest. The sequences around the predicted binding sites for this transcription are used for a new scan; the window size option defines the length of these sequences. The transcription factor deficit is the cut-off for inclusion of sites in the analysis. If an enrichment analysis has been performed, it is possible to choose the most significant deficit to select the sites.

### Performing a new scan and enrichment

A set of transcription factor models are then used with the new scan at the specified deficit. These default to those which were used for the original scan. Once these options have been selected and “Run Enrichment” is pressed, another window opens for selection of the background gene list (this is identical to the usual enrichment load box). The background is first scanned with the transcription factor of interest to find predicted sites, then sequences are extracted around these sites, to be used as background sequences for the enrichment analysis.

The proximal enrichment analysis panel is identical to the enrichment panel, with the same options for saving data, images and generating plots.

# Command line workflow

**CiiiDER** can be run using Terminal (Mac, Linux) or Command Prompt (Windows) without the GUI. Run the JAR using ‘-n’ and supply a configuration (config) file with a “.ini” extension:

java –jar CiiiDER.jar –n config.ini

The config file provides input file paths, parameters and output file paths for **CiiiDER**.

Note that the proximal enrichment analysis is not currently implemented in the command line workflow.

## General parameters

### Required:

STARTPOINT: Start point of the analysis, either 1 (scan) or 2 (enrichment analysis).

ENDPOINT: End point of the analysis. 1, (scan) or 2 (enrichment analysis).

### Optional:

PROCESSORS: Number of processors to use for multithreaded analyses. Defaults to the maximum number of the computer.

INPUTFOLDER: File path of a folder containing input files.

OUTPUTFOLDER: File path of a folder for saving results files.

DEBUGLOGFILE: File path or name for debug file for storing progress and errors. Defaults to run.log in the output folder.

PROJECTOUTPUTFILE: File path to save a **CiiiDER** project file, which can be opened in the GUI (.cdr extension).

## Scan parameters

The scan requires always requires *GENELISTFILENAME* and *MATRIXFILE*. If the scan is also the *ENDPOINT*, then *GENESCANRESULTS* is also required.

### Input files:

GENELISTFILENAME: File path for the gene list in FASTA, GTF, BED or TXT format.

REFERENCEFASTA: Genome sequence file in FASTA format (.fa or .fasta extension). Required for gene names, Ensembl IDs, GTF files and BED files.

GENELOOKUPMANAGER: Transcription start site reference file (.gtf or .glm extension). Required for gene names and Ensembl IDs.

MATRIXFILE: Input file path for the transcription factor models in TRANSFAC or JASPAR format.

### Parameters:

DEFICIT: Cut-off for site prediction. Default 0.15.

UPSTREAMOFFSET: For GTF, BED or TXT format gene lists. Default 1500.

DOWNSTREAMOFFSET: For GTF, BED or TXT format gene lists. Default 500.

### Output files:

GENESCANRESULTS: Output file path for binding site data. BSL extension to save for further analysis or TXT extension to obtain a list of sites.

## Enrichment parameters

The enrichment analysis always requires *GENELISTFILENAME* and *MATRIXFILE*. If enrichment is the *STARTPOINT*, then results of scans must also be provided as *BINDSITEFILENAME* and *BGBINDSITEFILENAME*. If enrichment is the *ENDPOINT*, then *ENRICHMENTOUTPUTFILE* is required.

### Input files:

BINDSITEFILENAME: Optional file path for the containing existing binding site data in BSL format.

BGGENELISTFILENAME: Input file path for the background gene list in FASTA, GTF or TXT format.

BGBINDSITEFILENAME: Optional file path for the containing existing binding site data in CSL format.

### Parameters:

ENRICHMENTCOVERAGEPVALUE: Default 0.05

ENRICHMENTSITEPVALUE: Default 1.0.

### Output files:

ENRICHMENTOUTPUTFILE: Output file path for enrichment results.

BGGENESCANRESULTS: Optional output file path for the background binding site data in CSL format.

# References

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpaa, M.J.*, et al.* (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res *20*, 861-873.

Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res *31*, 3576-3579.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S.R., Tan, G.*, et al.* (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res *46*, D260-D266.

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R.*, et al.* (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res *44*, D110-115.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K.*, et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res *34*, D108-110.