

# **TRANSFAC** Plugin

USER MANUAL

# User manual for TRANSFAC 2.5

Windows, Mac OS X and Linux

November 7, 2017

**This software is for research purposes only.**

QIAGEN Aarhus  
Silkeborgvej 2  
Prismet  
DK-8000 Aarhus C  
Denmark



# Contents

<b>1 Introduction to TRANSFAC plugin</b>	<b>4</b>
<b>2 Installing the TRANSFAC plugin and files</b>	<b>5</b>
2.1 Download and install the TRANSFAC files on your computer . . . . .	5
2.2 Installation of the TRANSFAC plugin . . . . .	5
2.3 Configuring the Workbench Preferences . . . . .	5
<b>3 Searching for transcription factor binding sites</b>	<b>7</b>
<b>4 Output of TRANSFAC plugin</b>	<b>11</b>
<b>Bibliography</b>	<b>13</b>

# Chapter 1

## Introduction to TRANSFAC plugin

The TRANSFAC plugin can be used to search for putative transcription factor binding sites in DNA sequences.

The binding site predictions are done by the Match™ tool, which uses the positional weight matrix library from TRANSFAC® to analyze your sequences.<sup>1</sup> For each of these matrices, Match contains optimized parameters, for minimization of error rates (false negative rate and false positive rate).

As a special feature, Match provides the option to use profiles that are specific subsets of matrices with optimized cut-offs. These profiles allow users to adapt Match to their specific interests. For example, promoters of a certain tissue can be searched with tissue-specific profiles. In addition to several pre-defined profiles provided in the TRANSFAC plugin, you can create your own profiles in the online version of TRANSFAC. The matrices in the result list of the TRANSFAC plugin are hyperlinked to the respective Matrix Reports in the TRANSFAC online version.

In order to be able to use the TRANSFAC plugin you need to:

- Download and install the files relevant for TRANSFAC from geneXplain on your computer. You will need a download subscription of TRANSFAC from geneXplain to do this (see <http://genexplain.com/transfac/>).
- Install the TRANSFAC plugin into your workbench.
- Configure the Preferences in your workbench so that the TRANSFAC plugin knows where to find the TRANSFAC data and binaries.

We cover the above topics first in this manual, and following that, provide information on carrying out searches of TRANSFAC via the CLC Workbench.

---

<sup>1</sup>TRANSFAC is a registered trademark of QIAGEN

## Chapter 2

# Installing the TRANSFAC plugin and files

### 2.1 Download and install the TRANSFAC files on your computer

Once you have a subscription to the flat file release download of TRANSFAC, you can download the **match.zip** file from: <https://portal.genexplain.com/download/transfac/>

Decompress the zip file and place the extracted folder somewhere on your machine. Remember where you extracted the folder as you need to configure the Workbench preferences later on with this location. Please do not change the relative location or names of the files within this folder.

### 2.2 Installation of the TRANSFAC plugin

TRANSFAC plugin is installed using the plugin manager. In order to install plugins, the workbench must be run with administrator privileges.

**Help in the Menu Bar | Plugins... (  )**

or **Plugins (  ) in the Toolbar**

The plugin manager has two tabs at the top:

- **Manage Plugins.** This is an overview of plugins that are installed.
- **Download Plugins.** This is an overview of available plugins.

Install the plugin by clicking the **Download Plugins** tab and browsing for the plugin called "TRANSFAC".

When you close the dialog, you will be asked whether you wish to restart CLC Workbench. The plugin will not be ready for use before you have restarted the workbench.

Note that you may wish to click on the button labeled "No" and restart the Workbench yourself. If you choose "Yes", the workbench restarts using the administrator privileges you were using so that you could install the plugin.

### 2.3 Configuring the Workbench Preferences

Open up the Workbench Preferences by going to:

**Edit | Preferences | Advanced | TRANSFAC Preferences**

In this panel, click the **Select Location** button and select the Match folder within the downloaded TRANSFAC files. The Workbench will warn you if the files do not exist in this folder.

If you have a local installation of TRANSFAC available via a web interface, you can specify an alternative base-url pointing to your local installation.

If you do not have a local installation, just leave this at the default setting at [https://portal.genexplain.com/cgi-bin/build\\_t/idb/1.0/get.cgi?](https://portal.genexplain.com/cgi-bin/build_t/idb/1.0/get.cgi?).

In case your geneXplain online subscription is not restricted to TRANSFAC (*build\_t*), but also includes other geneXplain databases (e.g. *build\_ghptywl*), please replace the *build\_t* in the default base-url by *build\_ghptywl* (or as appropriate to your subscription).

The preferences panel is shown in figure 2.1

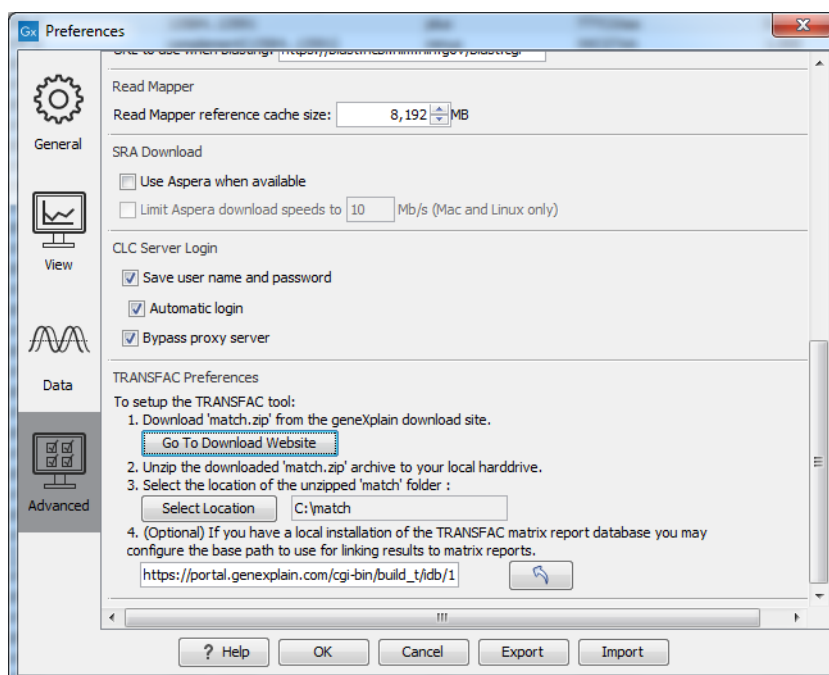


Figure 2.1: Pointing the Workbench to the data files downloaded from the geneXplain web site.

## Chapter 3

# Searching for transcription factor binding sites

Once the plugin is installed and the preferences have been set, you can start searching for transcription factor binding sites:

**Toolbox | Epigenomics Analysis (📁) | Annotate with TRANSFAC Information (⚠️)**

or (in Main Workbench):

**Toolbox | Nucleotide Analysis (📁) | Annotate with TRANSFAC Information (⚠️)**

This opens the wizard, the first step shown in figure 3.1 and explained below.

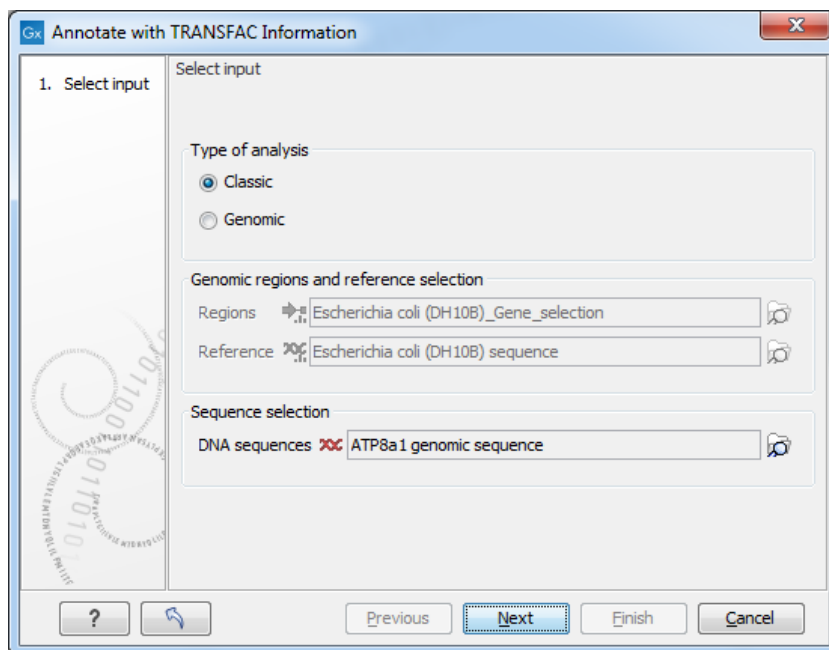


Figure 3.1: Select type of analysis and input data to search for transcription factor binding sites.

**Select Input** In this wizard step you specify the input data.

- **Type of analysis** First you need to select between the two modes of the analysis, each explained below.
  - **Classic** This is the legacy mode of analysis where you have raw sequences without genomic information and wish to search for transcription factor binding sites in these sequences.
  - **Genomic** Use this mode if genomic information is available in the form of genomic regions for which you wish to search for transcription factor binding sites. For example, if you have done peak detection from a ChIP-Seq analysis, the regions where peaks are found then are the genomic regions you wish to investigate further.
- **Genomic regions and reference selection:** This section will be enabled in the **Genomic** mode of analysis. Data is selected by clicking the corresponding folder icon (📁) in the right-hand side of the wizard.
  - **regions:** The regions of interest as a track of annotations.
  - **reference:** The reference genome as a track of symbols.
- **Sequence selection:** This section will be enabled in the **Classic** mode of analysis. Data is selected by clicking the corresponding folder icon (📁) in the right-hand side of the wizard.
  - **DNA sequences:** The raw DNA sequences of interest.

Clicking **Next** now opens the wizard step shown in figure 3.2.

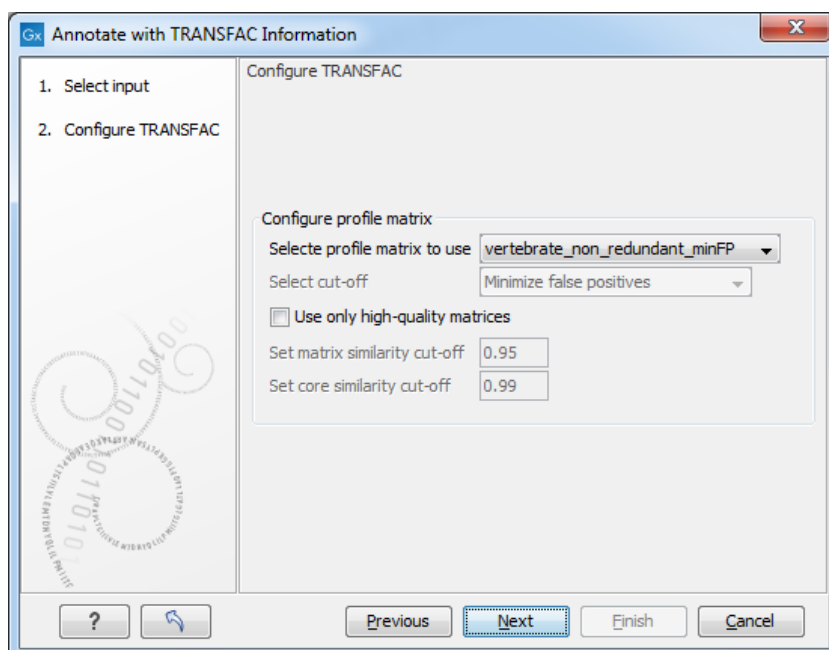


Figure 3.2: Selecting profiles and settings.

### Configure TRANSFAC

- **Select profile matrix:**



- **Select profile matrix to use:** From the dropdown menu, you specify the profile (set of position weight matrices) for a taxonomic group, tissue, etc. with which you want to scan your sequences/regions for putative transcription factor binding sites. You can also generate your own profiles (matrix sets) in the Match interface of the TRANSFAC online version. During profile generation in Match you are given the option to export/download the profiles. If you place profiles which you generated into the prfs folder of Match of your TRANSFAC flat file download which is used by the TRANSFAC plugin, your profiles will also be available for selection in this dialog box.
- **Configure precalculated cut-off:** This menu section is only enabled if the 'all' profile was selected.
  - **Use only high quality matrices:** When enabled this will exclude highly abundant matrices which produce at minSUM (see below) more than 10 hits (false positives) per 1000 nucleotides. "High quality matrices" are defined as matrices producing less than 10 hits (FP) per 1000 nucleotides (in sequences, 10.000 to 5.000 nucleotides upstream of the transcription start sites) at minSUM. About 5% of the current matrices producing higher FP rate, can be excluded as "highly abundant" / "low quality"; these 5% of matrices produce about 50% of all FP hits.
  - **Select cut-off:** The combobox has the following options:
    - \* **Minimize false positives (minFP).** The minFP cut-off can be used to reduce the number of false positives. The false positive rate is estimated by applying the Match algorithm to upstream sequences. The minFP cut-off is defined as the score that gives one percent of hits in the used sequences relative to the number of hits received at the minFN cut-off.
    - \* **Minimize false negatives (minFN).** The minFN cut-off can be used to reduce the number of false negatives. The false negative rate is measured, as far as available, on known genomic binding sites for the transcription factors. In case not sufficient (less than 10) genomic binding sites are available, SELEX sites or sets of oligonucleotides based on the nucleotide distribution in the weight matrix are used for estimating the minFN cut-off. The minFN cut-off is defined as that score at which at least 90% of the positive test set are recognized, i.e. it equals a false negative rate of 10%.
    - \* **Minimize the sum of both error rates (minSUM).** The minSUM cut-off can be used to minimize the sum of both error rates. The sum of corresponding percentages for false positives and false negatives is computed for every cut-off ranging from minFN to minFP, whereby the false positive rate at minFN (10% false negative rate) is defined as 100%. The score at which this sum is minimal is used for the minSUM cut-off.
- **Configure similarity score thresholds:** Here it is possible to further limit the number of matches reported in the results by defining a threshold that is applied to all matches in the results. The similarity score measures the quality of a match (see [Kel et al., 2003] for details). The range is from 0.0 to 1.0 where a score of 1.0 is given to the exact match. Try a more restrictive (higher) setting if you experience running out of memory.
  - **Define tresholds:** This will enable similarity score thresholding.
  - **Minimum matrix similarity score:** The threshold to use for the similarity score computed for the whole matrix.

- **Minimum core similarity score:** The threshold to use on the similarity score computed only on the core of the matrix (the 5 consecutive positions with highest similarity score).

## Chapter 4

# Output of TRANSFAC plugin

Clicking **Next** allows you to specify the output options as shown in figure 4.1.

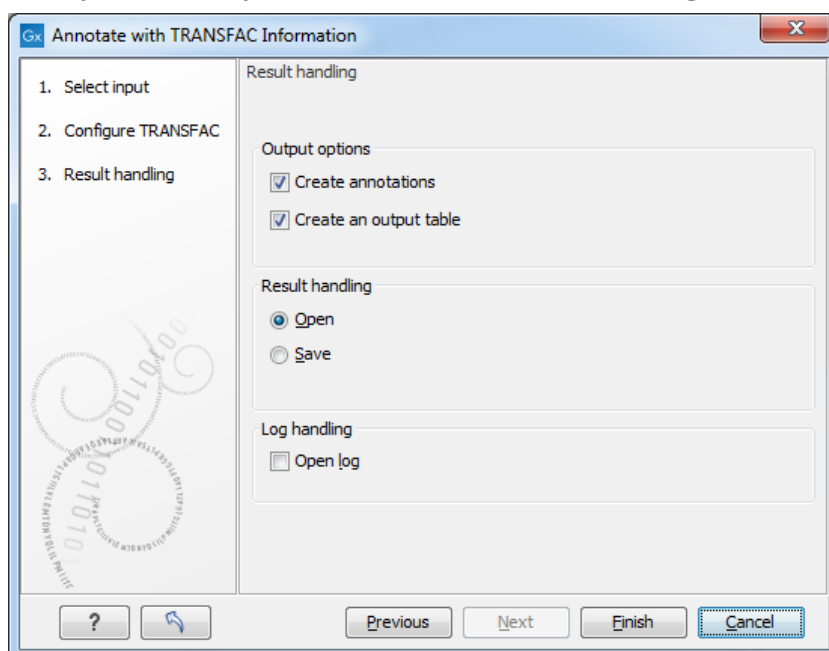


Figure 4.1: Selecting profiles and settings.

Based on the settings in figure 3.2, the Workbench uses the Match program to search for transcription factor binding sites. The binding sites can be reported as annotations that are either added to the input sequences or reported as a track of annotations (depending on the analysis type, Classic or Genomic, see Chapter 3) and/or as an output table as shown in figure 4.2.

The table includes the following information:

- **Matrix ID.** Identifier for the matrix with which the putative binding site was found.
- **Factor name.** Name of the binding factor, represented by the matrix. If a group of factors is assigned to a matrix, only representative factor(s) are given. For a complete list of linked binding factors, please see the Matrix Report in TRANSFAC.
- **Region.** Position of the matrix match (putative binding site) within the analyzed sequence.

Matrix ID	Factor name	Region	Strand	Match sequence	Core similarity	Matrix similarity	Matrix report
V\$GRE_C	GR	complement(18..33)	minus	cAGGACaaagcagcct	0.978	0.720	<a href="#">M00205</a>
V\$RFX1_01	RFX1	34..50	plus	ttggagcctGGAAcctt	0.982	0.930	<a href="#">M00280</a>
V\$MUSCLEINI_B	Muscle initiator	93..113	plus	gtggagcagCACCcagcct	1.000	0.873	<a href="#">M00321</a>
V\$DBP_06	DBP	26..32	plus	AGGAGGc	0.965	0.925	<a href="#">M00624</a>
V\$DBP_06	DBP	67..73	plus	AGCAGct	0.965	0.899	<a href="#">M00624</a>
V\$DBP_06	DBP	70..76	plus	AGCTctc	0.899	0.878	<a href="#">M00624</a>
V\$DBP_06	DBP	100..106	plus	AGCAGcct	0.965	0.942	<a href="#">M00624</a>
V\$PAX_06	Pax	42..52	plus	CTGGkatttaa	1.000	0.803	<a href="#">M00808</a>
V\$PAX_06	Pax	complement(51..61)	minus	aagatTCCAG	1.000	0.705	<a href="#">M00808</a>
V\$PAX_06	Pax	66..76	plus	CAGCAGctctc	0.675	0.672	<a href="#">M00808</a>
V\$PAX_06	Pax	complement(79..89)	minus	ctcaggGCGAG	0.900	0.711	<a href="#">M00808</a>
V\$PAX_06	Pax	84..104	plus	ctGGAcagac	1.000	0.724	<a href="#">M00808</a>
V\$PAX_06	Pax	complement(110..120)	minus	gtctctTCCAG	1.000	0.667	<a href="#">M00808</a>
V\$PAX_06	Pax	complement(145..155)	minus	ttaactTTCAG	0.818	0.754	<a href="#">M00808</a>
V\$PAX_06	Pax	157..167	plus	CCGGGaatctc	0.735	0.764	<a href="#">M00808</a>
V\$PAX_06	Pax	complement(159..169)	minus	gggaatCTCTG	0.575	0.671	<a href="#">M00808</a>
V\$PAX_06	Pax	complement(170..180)	minus	tgaagtCCTAG	0.660	0.666	<a href="#">M00808</a>
V\$DR4_02	LXR, FXR, CAR, COUP, RAR	complement(118..134)	minus	caGGTCctctgggctc	0.754	0.708	<a href="#">M00965</a>
V\$CH1_01	Churchill	158..163	plus	CGGGaa	0.999	0.995	<a href="#">M01004</a>
V\$HELOSA_02	Helios A	16..26	plus	cqcAGGACaaa	0.916	0.915	<a href="#">M01004</a>
V\$HELOSA_02	Helios A	complement(112..122)	minus	ctcCTCCAggt	0.877	0.790	<a href="#">M01004</a>
V\$HELOSA_02	Helios A	complement(121..131)	minus	gtcCTCTggg	0.812	0.811	<a href="#">M01004</a>
V\$AP2ALPHA_03	AP-2alpha	27..41	plus	gtAGCTTgagacc	0.990	0.771	<a href="#">M01047</a>
V\$AP2ALPHA_03	AP-2alpha	complement(27..41)	minus	gcagcctTGGAGcc	0.608	0.771	<a href="#">M01047</a>
V\$AP2ALPHA_03	AP-2alpha	68..82	plus	gcAGCTCctgcctca	0.609	0.721	<a href="#">M01047</a>
V\$AP2ALPHA_03	AP-2alpha	complement(68..82)	minus	gcagctctCGGCTca	0.881	0.721	<a href="#">M01047</a>
V\$AP2ALPHA_03	AP-2alpha	74..88	plus	ctcCTCcaagcca	0.719	0.812	<a href="#">M01047</a>
V\$AP2ALPHA_03	AP-2alpha	complement(74..88)	minus	ctcggctcAGGCGca	0.954	0.812	<a href="#">M01047</a>
V\$AP2ALPHA_03	AP-2alpha	80..84	plus	tcAGGCcaagcagc	0.609	0.765	<a href="#">M01047</a>
V\$AP2ALPHA_03	AP-2alpha	complement(80..84)	minus	tcaggcccAGGAGc	0.966	0.765	<a href="#">M01047</a>
V\$CTCF_01	CTCF	complement(106..125)	minus	cagagcCTctccagctct	0.917	0.889	<a href="#">M01200</a>
V\$BEN_01	BEN	complement(13..20)	minus	gtcCGAG	0.889	0.884	<a href="#">M01240</a>
V\$BEN_01	BEN	108..113	plus	CAGCGcct	1.000	0.914	<a href="#">M01240</a>
V\$NANOG_02	Nanog	14..33	plus	tctcagagCAAGcagcct	1.000	0.927	<a href="#">M01247</a>
V\$SIK1_01	SiK-1	complement(47..63)	minus	acttaGATGtccagag	0.659	0.659	<a href="#">M01313</a>
V\$SIK1_01	SiK-1	56..72	plus	gtccagaGATCagcagc	0.818	0.736	<a href="#">M01313</a>
V\$SIK1_01	SiK-1	155..171	plus	gtccgggAATctctgtg	0.659	0.714	<a href="#">M01313</a>
V\$HOX13_01	HOX13	complement(44..59)	minus	ggagcCTTAAagatgctc	0.967	0.681	<a href="#">M01317</a>
V\$HDX_01	Hdk	18..34	plus	caggacaAAGCAGcctt	0.768	0.723	<a href="#">M01333</a>
V\$HDX_01	Hdk	complement(47..63)	minus	acttaGATGtccagag	0.674	0.703	<a href="#">M01333</a>
V\$HDX_01	Hdk	56..72	plus	gtccagaGATCagcagc	0.852	0.837	<a href="#">M01333</a>
V\$HDX_01	Hdk	155..171	plus	gtccgggAATctctgtg	0.832	0.755	<a href="#">M01333</a>
V\$HDX_01	Hdk	complement(165..181)	minus	ctctGTCAAGtcttagc	0.630	0.686	<a href="#">M01333</a>
V\$RHOK11_01	Rhok1	48..64	plus	cttaaGATGtccagaga	0.904	0.755	<a href="#">M01347</a>
V\$RHOK11_01	Rhok1	68..104	plus	aggcaGCTGcaagcagc	0.756	0.701	<a href="#">M01347</a>

Figure 4.2: Result table.

- **Strand.** Plus/minus. The strand on which the putative site was found depends on the orientation in which the matrix is given in TRANSFAC.
- **Match sequence.** Shows the matching sequence. Capital letters indicate the positions in the sequence that match with the core sequence of the matrix, while the lower case letters refer to positions that match to the remaining part of the matrix.
- **Core similarity.** The core similarity score for the matrix match. (The matrix core is defined as the five consecutive most conserved nucleotides within the matrix.)
- **Matrix similarity.** The matrix similarity score for the matrix match. The Match score can vary from 0 to 1, with 0 for the lowest similarity and 1 for the highest similarity of the match to the matrix. Only those matches are listed in the result, for which the core and matrix similarity are higher than the chosen cut-offs.
- **Matrix Report.** Matrix accession with a hyperlink to the Matrix Report on the TRANSFAC web page, where you can find details on the matrix and links to its binding factors.

You can **Export**  the table in csv or Excel format.

# Bibliography

[Kel et al., 2003] Kel, A. E., Ling, E. G., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003). MATCHM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, 31(13):3576–3579.