TRANSFAC Plugin

USER MANUAL

User manual for

TRANSFAC 2.4

Windows, Mac OS X and Linux

February 24, 2017

This software is for research purposes only.

QIAGEN Aarhus
Silkeborgvej 2
Prismet
DK-8000 Aarhus C
Denmark

Contents

Chapter 1

Introduction to the 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.[1] 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.

[1] 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

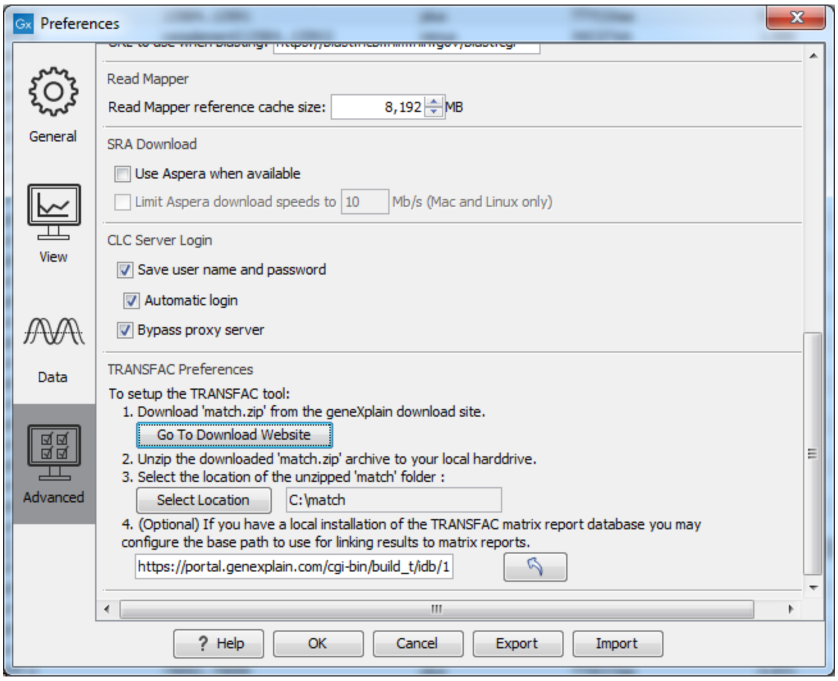
In this panel, click the **Select Location** button and select the Match folder within the downloaded TRANSFAC files. The Workbench will warn you if the files do not exist in this folder.

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?`.

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



Read Mapper

Read Mapper reference cache size: 8,192 MB

SRA Download

□ Use Aspera when available

□ Limit Aspera download speeds to 10 Mb/s (Mac and Linux only)

CLC Server Login

☑ Save user name and password

☑ Automatic login

☑ Bypass proxy server

TRANSFAC Preferences

To setup the TRANSFAC tool:

1. Download 'match.zip' from the geneXplain download site.

Go To Download Website

2. Unzip the downloaded 'match.zip' archive to your local harddrive.
3. Select the location of the unzipped 'match' folder :

Select Location C:\match

4. (Optional) If you have a local installation of the TRANSFAC matrix report database you may configure the base path to use for linking results to matrix reports.

https://portal.genexplain.com/cgi-bin/build_t/idb/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

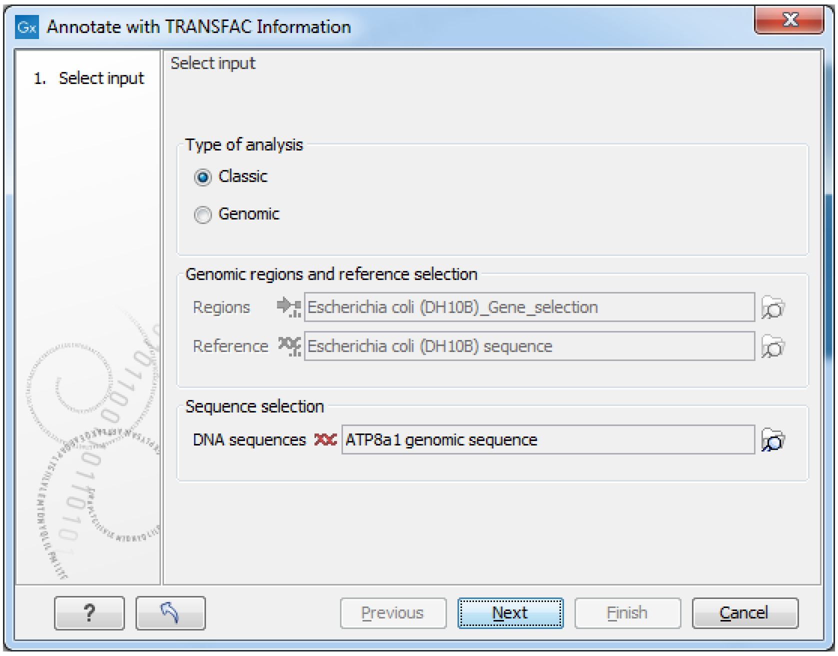
Once the plugin is installed and the preferences have been set, you can start searching for transcription factor binding sites:

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.



Annotate with TRANSFAC Information

1. Select input

Select input

Type of analysis

● Classic

○ Genomic

Genomic regions and reference selection

Regions: Escherichia coli (DH10B)_Gene_selection

Reference: Escherichia coli (DH10B) sequence

Sequence selection

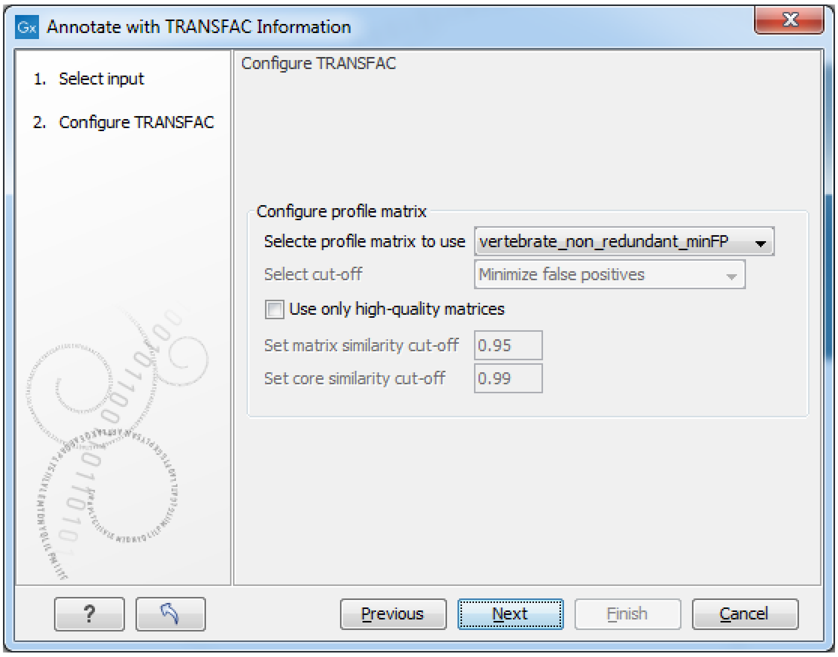
DNA sequences: ATP8a1 genomic sequence

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.



Annotate with TRANSFAC Information

1. Select input
2. Configure TRANSFAC

Configure TRANSFAC

Configure profile matrix

Selecte profile matrix to use: vertebrate_non_redundant_minFP

Select cut-off: Minimize false positives

□ Use only high-quality matrices

Set matrix similarity cut-off: 0.95

Set core similarity cut-off: 0.99

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.

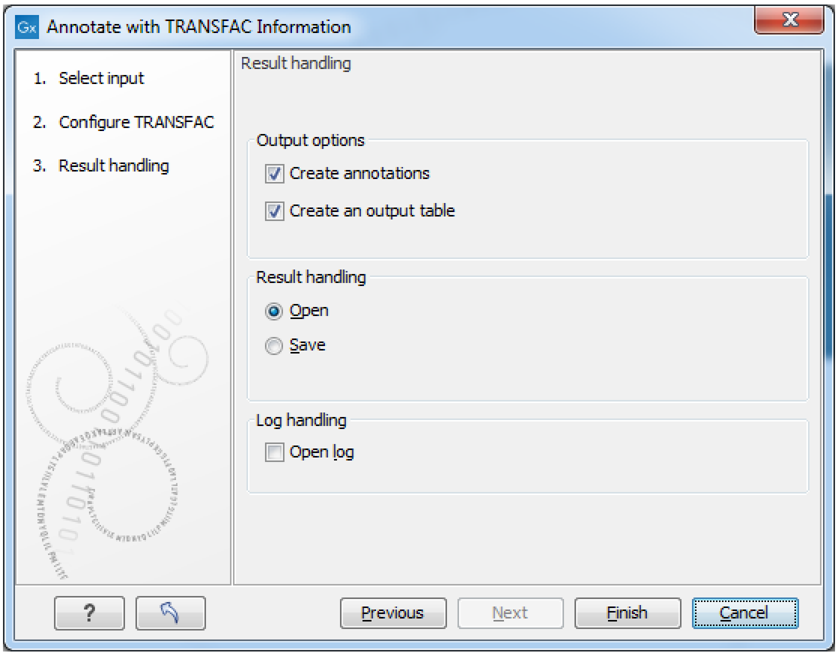
  - **Define tresholds:** This will enable similarity score thresholding.
  - **Minimum matrix similarity score:** The threshold to use for the similarity score computed for the whole matrix.

- **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.



Annotate with TRANSFAC Information

1. Select input
2. Configure TRANSFAC
3. Result handling

Result handling

Output options

☑ Create annotations

☑ Create an output table

Result handling

● Open

○ Save

Log handling

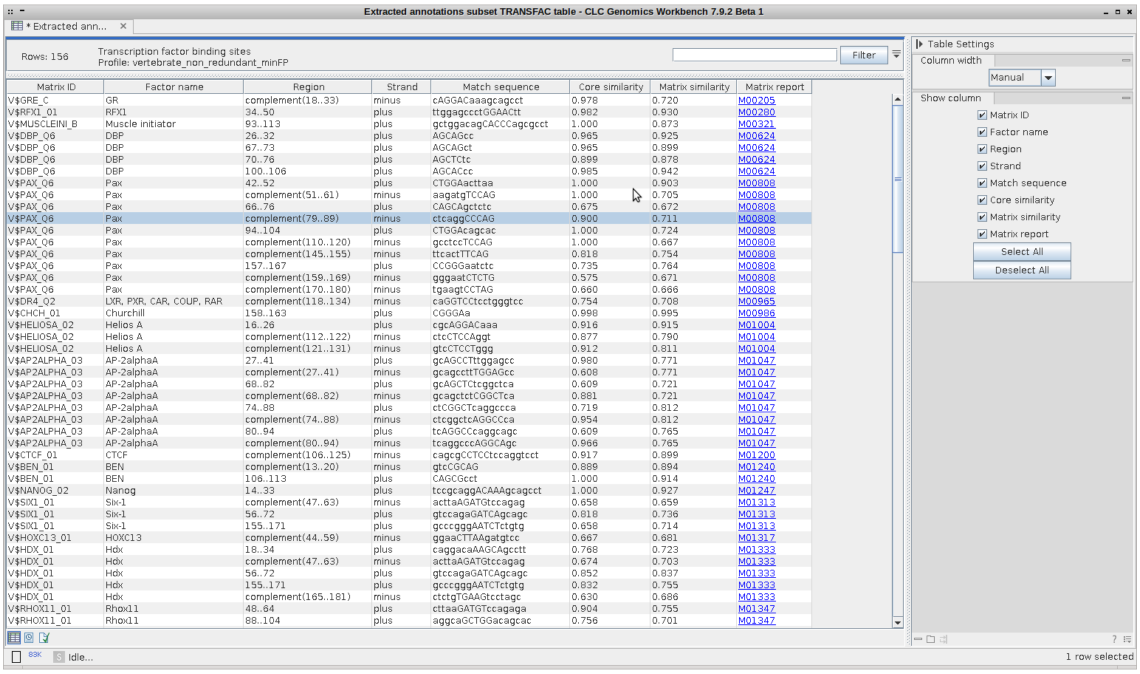
□ Open log

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.



Extracted annotations subset TRANSFAC table - CLC Genomics Workbench 7.9.2 Beta 1

Rows: 156

Transcription factor binding sites
Profile: vertebrate_non_redundant_minFP

Matrix ID	Factor name	Region	Strand	Match sequence	Core similarity	Matrix similarity	Matrix report
V$GRE_C	GR	complement(18..33)	minus	cAGGACaaagcagcct	0.978	0.720	M00205
V$RFX1_01	RFX1	34..50	plus	ttggagccctGGAACtt	0.982	0.930	M00280
V$MUSCLEINI_B	Muscle initiator	93..113	plus	gctggacagCACCCagcgcct	1.000	0.873	M00321
V$DBP_Q6	DBP	26..32	plus	AGCAGcc	0.965	0.925	M00624
V$DBP_Q6	DBP	67..73	plus	AGCAGct	0.965	0.899	M00624
V$DBP_Q6	DBP	70..76	plus	AGCTCtc	0.899	0.878	M00624
V$DBP_Q6	DBP	100..106	plus	AGCACcc	0.985	0.942	M00624
V$PAX_Q6	Pax	42..52	plus	CTGGAacttaa	1.000	0.903	M00808
V$PAX_Q6	Pax	complement(51..61)	minus	aagatgTCCAG	1.000	0.705	M00808
V$PAX_Q6	Pax	66..76	plus	CAGCAgctctc	0.675	0.672	M00808
V$PAX_Q6	Pax	complement(79..89)	minus	ctcaggCCCAG	0.900	0.711	M00808
V$PAX_Q6	Pax	94..104	plus	CTGGAcagcac	1.000	0.724	M00808
V$PAX_Q6	Pax	complement(110..120)	minus	gcctccTCCAG	1.000	0.667	M00808
V$PAX_Q6	Pax	complement(145..155)	minus	ttcactTTCAG	0.818	0.754	M00808
V$PAX_Q6	Pax	157..167	plus	CCGGGaatctc	0.735	0.764	M00808
V$PAX_Q6	Pax	complement(159..169)	minus	gggaatCTCTG	0.575	0.671	M00808
V$PAX_Q6	Pax	complement(170..180)	minus	tgaagtCCTAG	0.660	0.666	M00808
V$DR4_Q2	LXR, PXR, CAR, COUP, RAR	complement(118..134)	minus	caGGTCCtcctgggtcc	0.754	0.708	M00965
V$CHCH_01	Churchill	158..163	plus	CGGGAa	0.998	0.995	M00986
V$HELIOSA_02	Helios A	16..26	plus	cgcAGGACaaa	0.916	0.915	M01004
V$HELIOSA_02	Helios A	complement(112..122)	minus	ctcCTCCAggt	0.877	0.790	M01004
V$HELIOSA_02	Helios A	complement(121..131)	minus	gtcCTCCTggg	0.912	0.811	M01004
V$AP2ALPHA_03	AP-2alphaA	27..41	plus	gcAGCCTttggagcc	0.980	0.771	M01047
V$AP2ALPHA_03	AP-2alphaA	complement(27..41)	minus	gcagcctttTGGAGcc	0.608	0.771	M01047
V$AP2ALPHA_03	AP-2alphaA	68..82	plus	gcAGCTCtcggctca	0.609	0.721	M01047
V$AP2ALPHA_03	AP-2alphaA	complement(68..82)	minus	gcagctctCGGCTca	0.881	0.721	M01047
V$AP2ALPHA_03	AP-2alphaA	74..88	plus	ctCGGCTcaggccca	0.719	0.812	M01047
V$AP2ALPHA_03	AP-2alphaA	complement(74..88)	minus	ctcggctcAGGCCca	0.954	0.812	M01047
V$AP2ALPHA_03	AP-2alphaA	80..94	plus	tcAGGCCcaggcagc	0.609	0.765	M01047
V$AP2ALPHA_03	AP-2alphaA	complement(80..94)	minus	tcaggcccAGGCAgc	0.966	0.765	M01047
V$CTCF_01	CTCF	complement(106..125)	minus	cagcgCCTCCtccaggtcct	0.917	0.899	M01200
V$BEN_01	BEN	complement(13..20)	minus	gtcCGCAG	0.889	0.894	M01240
V$BEN_01	BEN	106..113	plus	CAGCGcct	1.000	0.914	M01240
V$NANOG_02	Nanog	14..33	plus	tccgcaggACAAAgcagcct	1.000	0.927	M01247
V$SIX1_01	Six-1	complement(47..63)	minus	acttaAGATGtccagag	0.658	0.659	M01313
V$SIX1_01	Six-1	56..72	plus	gtccagaGATCAgcagc	0.818	0.736	M01313
V$SIX1_01	Six-1	155..171	plus	gcccgggAATCTctgtg	0.658	0.714	M01313
V$HOXC13_01	HOXC13	complement(44..59)	minus	ggaaCTTAAgatgtcc	0.667	0.681	M01317
V$HDX_01	Hdx	18..34	plus	caggacaAAGCAgcctt	0.768	0.723	M01333
V$HDX_01	Hdx	complement(47..63)	minus	acttaAGATGtccagag	0.674	0.703	M01333
V$HDX_01	Hdx	56..72	plus	gtccagaGATCAgcagc	0.852	0.837	M01333
V$HDX_01	Hdx	155..171	plus	gcccgggAATCTctgtg	0.832	0.755	M01333
V$HDX_01	Hdx	complement(165..181)	minus	ctctgTGAAGtcctagc	0.630	0.686	M01333
V$RHOX11_01	Rhox11	48..64	plus	cttaaGATGTccagaga	0.904	0.755	M01347
V$RHOX11_01	Rhox11	88..104	plus	aggcaGCTGGacagcac	0.756	0.701	M01347

Table Settings

Column width: Manual

Show column

- ☑ Matrix ID
- ☑ Factor name
- ☑ Region
- ☑ Strand
- ☑ Match sequence
- ☑ Core similarity
- ☑ Matrix similarity
- ☑ Matrix report

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