

Transcription factors

Hands-on-training

Part 1: Mapping a gene list with TF classification

Part 2: Converting Matrices to molecules

Part 3: Analyzing ChIP-seq peaks

Part 1

TFClass is a classification of eukaryotic transcription factors based on the characteristics of their DNA-binding domains (DBDs).

We will use a list of genes and identify the most prominent transcription factor class with the functional classification tool and underlying statistics. This analysis allows you to classify a set of genes into TF classification groups.

Part 1 – Classify a set of genes into TF classification groups

After logging into the geneXplain platform, open the **Functional classification** tool as shown below.

The screenshot displays the geneXplain web interface. On the left, a sidebar menu under the 'Analyses' tab shows a tree structure. The 'Functional classification' tool is highlighted with a red circle, and a red dashed arrow points from it to the main configuration panel on the right. The main panel has a tab labeled 'Functional classification X'. It contains a form with the following fields:

Source data set	<input type="text" value="(select element)"/>
Species	<input type="text" value="Human (Homo sapiens)"/>
Classification	<input type="text" value="Full gene ontology classification"/>
Minimal hits to group	<input type="text" value="2"/>
P-value threshold	<input type="text" value="0.05"/>
Result name	<input type="text" value="(select element)"/>

Below the form, there is a link 'Show expert options >>' and a 'Run' button.

Part 1 – Classify a set of genes into TF classification groups

Navigate to the *UpDownReg_168_Ensembl_genes* table and drag-and-drop it to the **Source data set** field of the tool. The field **Classification** is a source directory whose information is used for the comparison (mapping) and select **TF classification**. Change **Minimal hits to group** to **1**, because our input list is quit small and we would like to have an overrepresentation from the first hit already. Click **Run** when parameters are set as shown below.

The screenshot displays the GeneXplain Functional classification tool interface. On the left, the 'Data' section is expanded, showing a list of folders and files. The file 'UpDownReg_168_Ensembl_genes' is circled in red. A dashed red arrow points from this file to the 'Source data set' field in the tool configuration panel on the right. The 'Source data set' field is also circled in red. Other fields in the panel include 'Species' (Human (Homo sapiens)), 'Classification' (TF classification), 'Minimal hits to group' (1), 'P-value threshold' (0.05), and 'Result name' (..._168_Ensembl_genes TF classification). A red arrow points to the 'Run' button, which is labeled 'RUN' in large red letters.

Part 1 – Classify a set of genes into TF classification groups

Sorting the resulting table with lowest **Adjusted P-value** on top gives four most prominent TF classes in the red box.

Start page

Functional classification X

UpDownReg_168_Ense... X

Edit

Apply

Cancel

Select all

Select page

First

Previous

Page 1 of 1

Next

Last

Showing 1 to 35 of 35 entries

Show 50 entries

ID	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
6.1	Rel homology region (RHR) factors	4	30	0.35225	3.0202E-4	0.00944	BCL3, NFKBIA, NFKBIE, RELB
6.1.2	Ankyrin domain-only factors	3	14	0.16438	4.5699E-4	0.00944	BCL3, NFKBIA, NFKBIE
6.1.2.1	IκB-related factors	3	13	0.15264	3.6172E-4	0.00944	BCL3, NFKBIA, NFKBIE
6	Immunoglobulin fold	5	67	0.78669	7.5582E-4	0.01172	BCL3, NFKBIA, NFKBIE, RELB, STAT6
1.1.1.1.2	JunB	1	1	0.01174	0.01174	0.02912	JUNB
1.1.2.2.1	ATF-3	1	1	0.01174	0.01174	0.02912	ATF3
2.1.3.5.2	COUP-TFII (NR2F2)	1	1	0.01174	0.01174	0.02912	NR2F2
2.2.1.1.3	GATA-3	1	1	0.01174	0.01174	0.02912	GATA3
2.3.4.5.2	HIV-EP2 [2+2] (MBP-2, ZNF40B)	1	1	0.01174	0.01174	0.02912	HIVEP2
3.1.2.11.1	MSX-1 (HOX7)	1	1	0.01174	0.01174	0.02912	MSX1
3.3.2.1.8	E2F-8	1	1	0.01174	0.01174	0.02912	E2F8
3.5.2.1.2	c-Ets-2	1	1	0.01174	0.01174	0.02912	ETS2
3.5.3	Interferon-regulatory factors	2	9	0.10568	0.00447	0.02912	IRF1, IRF2
3.5.3.0.1	IRF-1	1	1	0.01174	0.01174	0.02912	IRF1
3.5.3.0.2	IRF-2	1	1	0.01174	0.01174	0.02912	IRF2
5.1.1.1.3	MEF-2C	1	1	0.01174	0.01174	0.02912	MEF2C
6.1.1.2.2	RelB (I-Rel)	1	1	0.01174	0.01174	0.02912	RELB

Part 2

Now we are using a tool to convert a list of transcription factor binding sites into a list of corresponding transcription factors.

Part 2 – Convert a list of transcription factor binding sites (TFBSs) into a list of transcription factors (TFs)

Open the **Matrices to molecules** tool as shown below.

The image shows the geneXplain web interface. On the left, the 'Analyses' menu is open, with 'Data manipulation' and 'Matrices to molecules' highlighted by red circles. A red dashed arrow points from 'Matrices to molecules' in the menu to the tool's configuration page on the right.

The 'Matrices to molecules' tool configuration page has a title bar 'Start page Matrices to molecules X'. It contains the following settings:

Parameter	Value
Sites table	(select element)
Profile	<input type="checkbox"/> (select element)
Species	Human (Homo sapiens)
Output type	Unspecified
Ignore empty values	<input checked="" type="checkbox"/>
Numerical value treatment rule	extreme
Leading column	(none)
Output table	(select element)

Part 2 – Convert a list of transcription factor binding sites (TFBSs) into a list of transcription factors (TFs)

Navigate to the *Enriched motifs top50* table and drag-and-drop it to the **Sites table** field of the tool. The field **Output type** defines the hub, which is used for the conversion mapping and defines the output type and ID. Click **Run** when parameters are set as shown below. The resulting table has Ensembl IDs added to the original Sites table.

Databases Data Analyses

Training 290414 P9
Training_platform
Data
1_Introduction
2_Motifs
3_Factors
Sites top50
UpDownReg_168_Ensembl_genes
UpDownReg_168_Ensembl_genes TF classification
4_Enhancers
5_Pathways
6_Networks
7_Targets
8_Biomarkers

Start page Matrices to molecules X

Sites table ..._platform/Data/3_Factors/Sites top50

Profile ☒ ...ched motifs_TRANSFAC(R))/Profile

Species Human (Homo sapiens)

Output type Genes: Ensembl

Ignore empty values ☒

Numerical value treatment rule extreme

Leading column P-value

Output table ...actors/Sites top50 TFs Genes Ensembl

Run **RUN**

Databases Data Analyses

Training 290414 P9
Training_platform
Data
1_Introduction
2_Motifs
3_Factors
Sites top50
Sites top50 TFs Genes Ensembl
UpDownReg_168_Ensembl_genes
UpDownReg_168_Ensembl_genes TF classification
4_Enhancers
5_Pathways
6_Networks
7_Targets
8_Biomarkers

Start page Matrices to molecules X Sites top50 TFs Genes E... X

Edit Apply Cancel Select all Select page

First Previous Page 1 of 1 Next Last Showing 1 to 39 of 39 entries

ID	Site model ID	Length	Matrix logo
ENSG00000006194	V\$FPM315_02	20	GGAGGA
ENSG00000008196	V\$AP2_Q6	12	cCc
ENSG000000028277	V\$OCT2_Q6	14	ATGCAAA

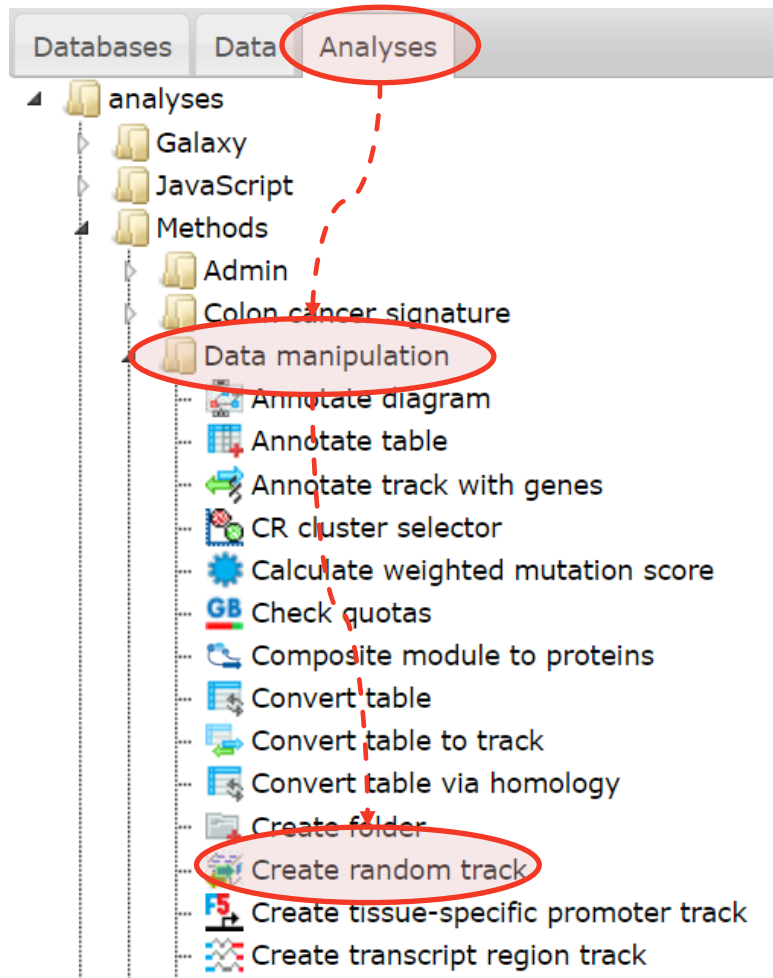
Part 3

This part demonstrates the analysis of ChIP-seq data in the geneXplain platform. We will show how to work with genomic intervals loaded from BED files. The ChIP-seq peaks from Jurkat cells were bound by TAL1 in a study by Palii et al. [1]. In the first part, we will analyze binding sites in TAL1-bound regions in Jurkat cells.

1. Palii CG, Perez-Iratxeta C, Yao Z, Cao Y, Dai F *et al.* (2011) Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. EMBO J. 30:494-509.

Part 3 – Enriched motifs in ChIP-seq peaks

Open the **Create random track tool**. This tool randomly samples upstream regions of genes that do not overlap with and has the same distribution of lengths as a specified set of intervals. The random intervals will present our background set of genomic sequences for subsequent binding site analysis.



Part 3 – Enriched motifs in ChIP-seq peaks

Navigate to the *jurkat* cell track and drag-and-drop the item onto the **Input track** field of the tool. Change the Sequence source to Ensembl 75.37 Human (hg19). Click **Run** when parameters are set as shown below.

The screenshot displays the GeneXplain web interface. On the left, the 'Data' tab is active, showing a hierarchical tree structure. The 'Training_Input_Data' folder is expanded, revealing '3_Factors_Input', which contains 'GSM614003_jurkat.tal1'. This item is circled in red. A red dashed line indicates it is being dragged to the 'Input track' field of a tool configuration panel on the right. The tool configuration panel has the following settings:

- Input track: [...ors_Input/GSM614003_jurkat.tal1]
- Sequence source: Ensembl 75.37 Human (hg19) (circled in red)
- Species: Human (Homo sapiens)
- Standard chromosomes: [checked]
- Sequence number: 1000
- Sequence length: 0
- Allow overlap: [unchecked]
- Output track: [...t_Data/3_Factors_Input/Random track]

At the bottom of the tool configuration panel, there is a 'Show expert options >' link and a 'Run' button (circled in red).

Part 3 – Enriched motifs in ChIP-seq peaks

The next goal is to extract a subset of intervals, as we are not going to analyze binding sites in all peaks. Open the interval table and follow the steps described below to extract an interval subset.

Databases | **Data** | **Analyses**

Training 290414 P9
Training_platform
Data
1_Introduction
2_Motifs
3_Factors
GSM614003_jurkat.tal1
Open track
Open sequence
Open table
Expand/collapse
Export
Save a copy
Set data Omics type
Remove

Start page GSM614003_jurkat.tal1 X

Filter: Property_name > 25

Showing 1 to 451 of 451 entries

ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: name
1	1	3797506	3798136	631	?	unsure	41
2	1	6377096	6377399	304	?	unsure	39
6	1	9486386	9486877	492	?	unsure	40
9	1	10977931	10978572	642	?	unsure	51
12	1	12270557	12270974	418	?	unsure	42
17	1	16575310	16575588	279	?	unsure	36
19	1	19803414	19803729	316	?	unsure	71
20	1	21952326	21952623	298	?	unsure	37
21	1	21958479	21959258	780	?	unsure	32
24	1	24222945	24223469	525	?	unsure	47
25	1	24224259	24224490	232	?	unsure	28
28	1	26704744	26705302	559	?	unsure	61
32	1	29211463	29211790	328	?	unsure	40
35	1	31241846	31242046	201	?	unsure	30
47	1	38353684	38354329	646	?	unsure	63

Search | Info | SQL track | **Filters** | My description | Graph search | Script | Clipboard | Tasks

Track GSM614003_jurkat.tal1

Sequence collection: databases/EnsemblHuman75/Sequences/chromosomes GRCh37
Site count: 2238

platform.genexplain.com/bioulweb/#

bZIP_dimer_bound...jpg | TFBS.png

Alle anzeigen

Template to construct the filtering expression:
Select template -

Columns (double-click to paste):
To
Length
Strand
Type
Property_name

Expression in JavaScript language:
Property_name > 25

Part 3 – Enriched motifs in ChIP-seq peaks

We are now ready to compare binding sites in TAL1-bound regions with those in the sampled genomic background. Here we will apply the MEALR tool which finds a weighted set of discriminating motifs using sparse logistic regression.

Use the steps shown below to open the MEALR tool.

The screenshot shows the GeneXplain web interface. On the left is a vertical navigation menu with tabs for 'Databases', 'Data', and 'Analyses'. The 'Analyses' tab is selected, and a sub-menu 'Site analysis' is open. In this menu, 'MEALR (tracks)' is highlighted with a red circle. Red arrows indicate the navigation path from 'Analyses' to 'Site analysis' and then to 'MEALR (tracks)'. The main content area on the right shows the 'MEALR (tracks)' configuration page. It has a title bar with 'Start page' and 'MEALR (tracks) X'. Below this is a table with configuration options:

Yes set	[?]	(select element)
No set	[?]	(select element)
Sequence source	[?]	Ensembl 100.38 Human (hg38)
Input motif profile	[?]	...hed motifs_TRANSFAC(R))_new/Profile
Output path	[?]	(select element)

Part 3 – Enriched motifs in ChIP-seq peaks

In MEALR, specify **Yes** and **No** sets by dragging the interval items of the TAL-1 peak subset and of the random track onto respective fields. Navigate to the **vertebrate recommended specific** profile of **TRANSFAC 2020.3** to specify the set of motifs (a file navigator opens when clicking on the field **Input motif profile**). Click **Run** when ready.

The screenshot displays the MEALR (tracks) interface. On the left, the 'Data' sidebar shows a tree structure with '3_Factors' expanded, containing 'GSM614003_jurkat.tal1 subset' and 'Random track' (both circled in red). The main panel shows the 'MEALR (tracks)' configuration with the following fields:

- Yes set**: `...Factors/GSM614003_jurkat.tal1 subset`
- No set**: `...platform/Data/3_Factors/Random track`
- Sequence source**: `Ensembl 75.37 Human (hg19)`
- Input motif profile**: `...C(R) 2020.3/Data/profiles/vertebrates` (circled in red)
- Output path**: `...ata/3_Factors/Enriched_motifs_MEALR`

The **Run** button is circled in red. Below the configuration, the 'Input motif profile' dialog is open, showing a file navigator with the folder `databases/TRANSFAC(R) 2020.3/Data/profiles` selected. The file `vertebrate_recommended_specific` is highlighted in blue. The 'Name' field at the bottom contains `vertebrate_recommended_specific`. The **Ok** button is circled in red.

Part 1 – Enriched motifs in ChIP-seq peaks

Like the original study, MEALR identifies GATA, Runx-type, and ETS motifs as dominant patterns. In addition, it proposes also an MYB motif at the top of the list.

Start page MEALR (tracks) X Enriched_motifs_MEALR X

Edit Apply Cancel

First Previous Page 1 of 2 Next Last Showing 1 to 50 of 77 entries Show 50 entries

ID	Coefficient
V\$TAL1_04	0.35132
V\$AML2_03	0.16593
V\$GATA1_10	0.11608
V\$NGN2_02	0.11212
V\$ETV2_03	0.09975
V\$FLI1TCF3_01	0.09506
V\$AML1_02	0.06003
V\$HAND2_02	0.05756
V\$COREBINDINGFACTOR_Q6	0.0532
V\$GATA2_11	0.04998
V\$ZNF563_04	0.04472

A Jurkat versus control

	Class	Motif	Score
1	Runx	VCCACA*	49
2	Gata	(C)TTATCT*	45
3	E-box(GC)	CAGCTG	40
4	Ets	CAGGAAR	28
5	Gata	AGATAA	19
6	Runx	AACCACA	17
7	?	GCAGVC	17

<< Enriched_motifs_MEALR ... X

ENSG00000188227	zinc finger protein 793	ZNF793	V\$ZNF793_01	0.02102
ENSG00000141510	tumor protein p53	TP53	V\$P53_03	0.02173
ENSG00000144792	zinc finger protein 660	ZNF660	V\$ZNF660_01	0.02558
ENSG00000090447	transcription factor AP-4	TFAP4	V\$TFAP4_03	0.02621
ENSG00000178403	neurogenin 2	NEUROG2	V\$NEUROG2_03	0.02759
ENSG00000175691	zinc finger protein 77	ZNF77	V\$ZNF77_02	0.02845
ENSG00000187987	zinc finger and SCAN domain containing 23	ZSCAN23	V\$ZNF390_02	0.03276
ENSG00000148737	transcription factor 7 like 2	TCF7L2	V\$TCF4_07	0.04458
ENSG00000188868	zinc finger protein 563	ZNF563	V\$ZNF563_04	0.04472
ENSG00000164107	heart and neural crest derivatives expressed 2	HAND2	V\$HAND2_02	0.05756
ENSG00000159216	RUNX family transcription factor 1	RUNX1	V\$AML1_02	0.06003
ENSG00000071564	transcription factor 3	TCF3	V\$FLI1TCF3_01	0.09506
ENSG00000151702	Fli-1 proto-oncogene, ETS transcription factor	FLI1	V\$FLI1TCF3_01	0.09506
ENSG00000105672	ETS variant transcription factor 2	ETV2	V\$ETV2_03	0.09975
ENSG00000102145	GATA binding protein 1	GATA1	V\$GATA1_10	0.11608
ENSG00000020633	RUNX family transcription factor 3	RUNX3	V\$AML2_03	0.16593
ENSG00000162367	TAL bHLH transcription factor 1, erythroid differentiation factor	TAL1	V\$TAL1_04	0.35132

From Figure 6, Palii et al., EMBO J. 2011, 30:494-509

Practical session completed