Identification of DNA motifs

Hands-on-training

info@genexplain.com

Part 1

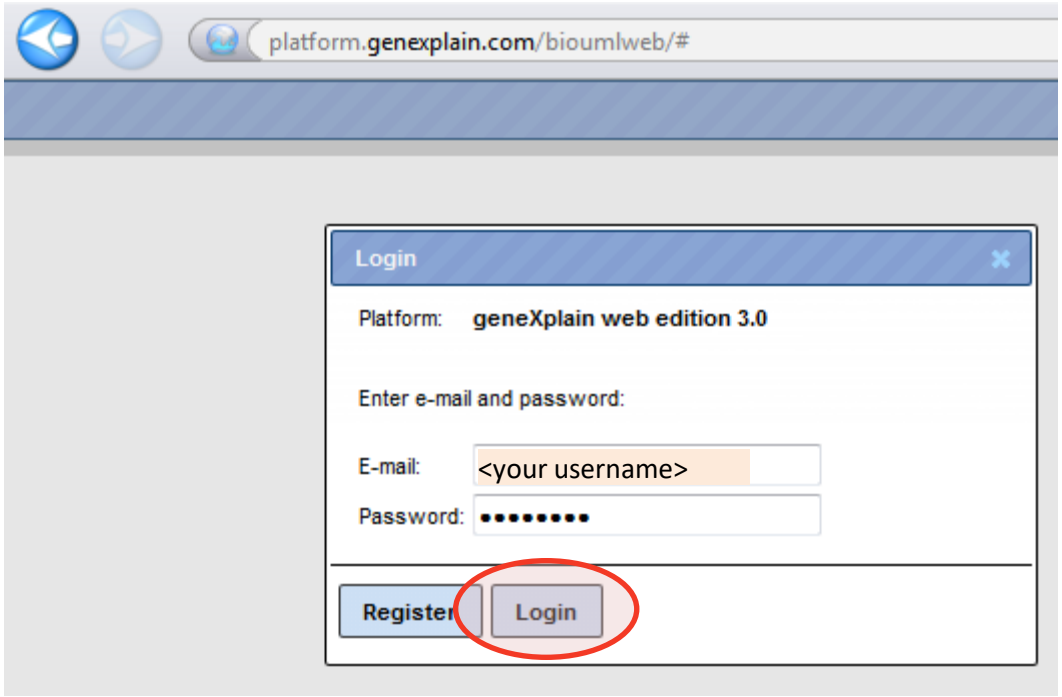
This part conducts a multi-step analysis to predict binding sites in sequences of ChIP-seq peaks with the geneXplain platform.

We will start of with loading a *narrowPeak*-file (a BED-like format used by Encode) describing locations of NF-kappaB-bound regions after TNF-alpha stimulation in Gm10847 cells and find the dominant motif in these sequences using ChIPMunk.

Part 1 – ChIPMunk in the geneXplain platform

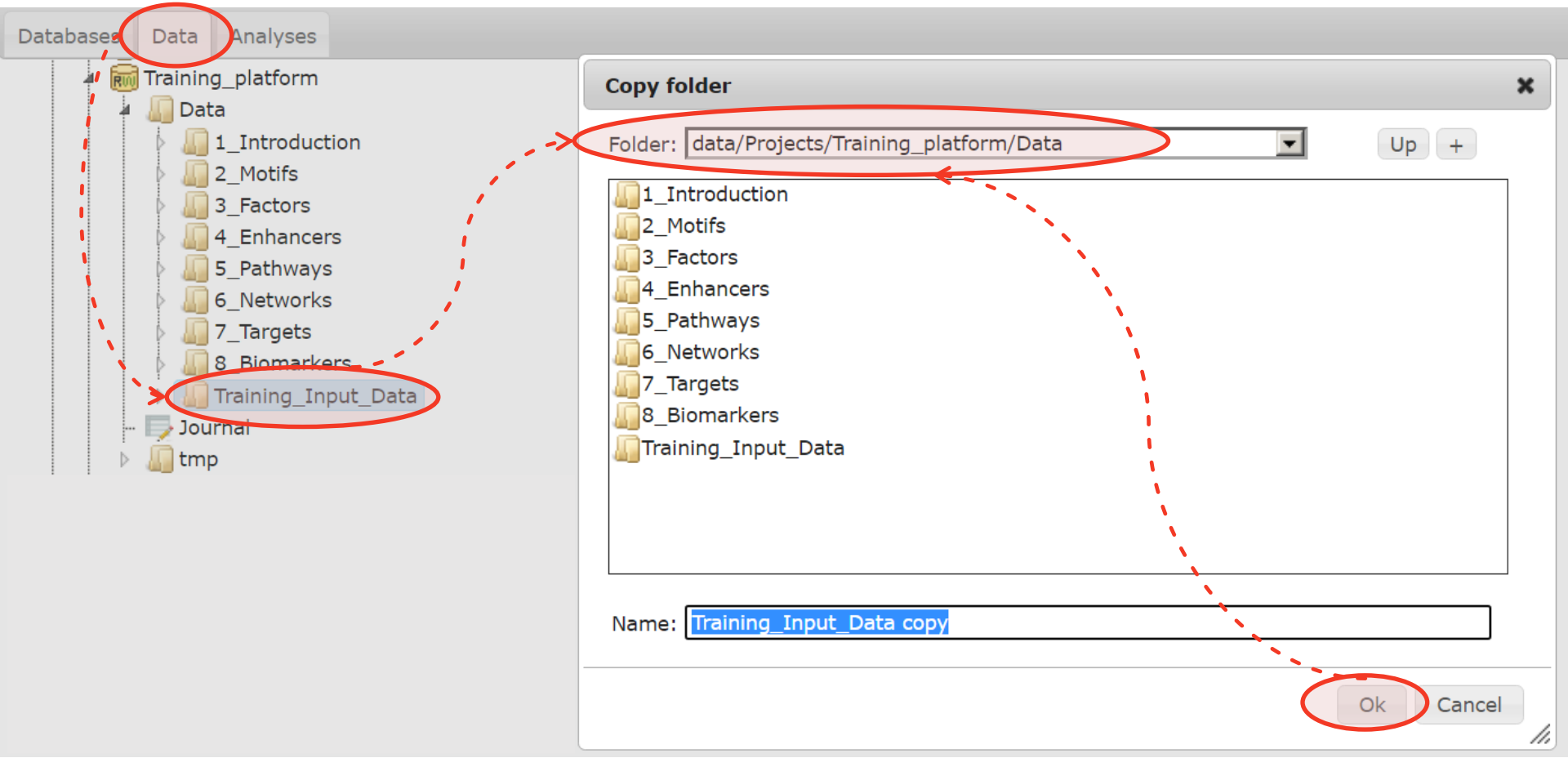
Open the geneXplain platform page in your browser: *https://platform.genexplain.com/bioumlweb*

Specify your credentials and click **Login**.



Part 1 – ChIPMunk in the geneXplain platform

Navigate to the **Training_platform** data project and further to the subfolder **Training_Input_Data.** To copy this folder, choose **Copy** with right-click on it. In the **Copy folder** window, provide a path to your data project and click **OK**.

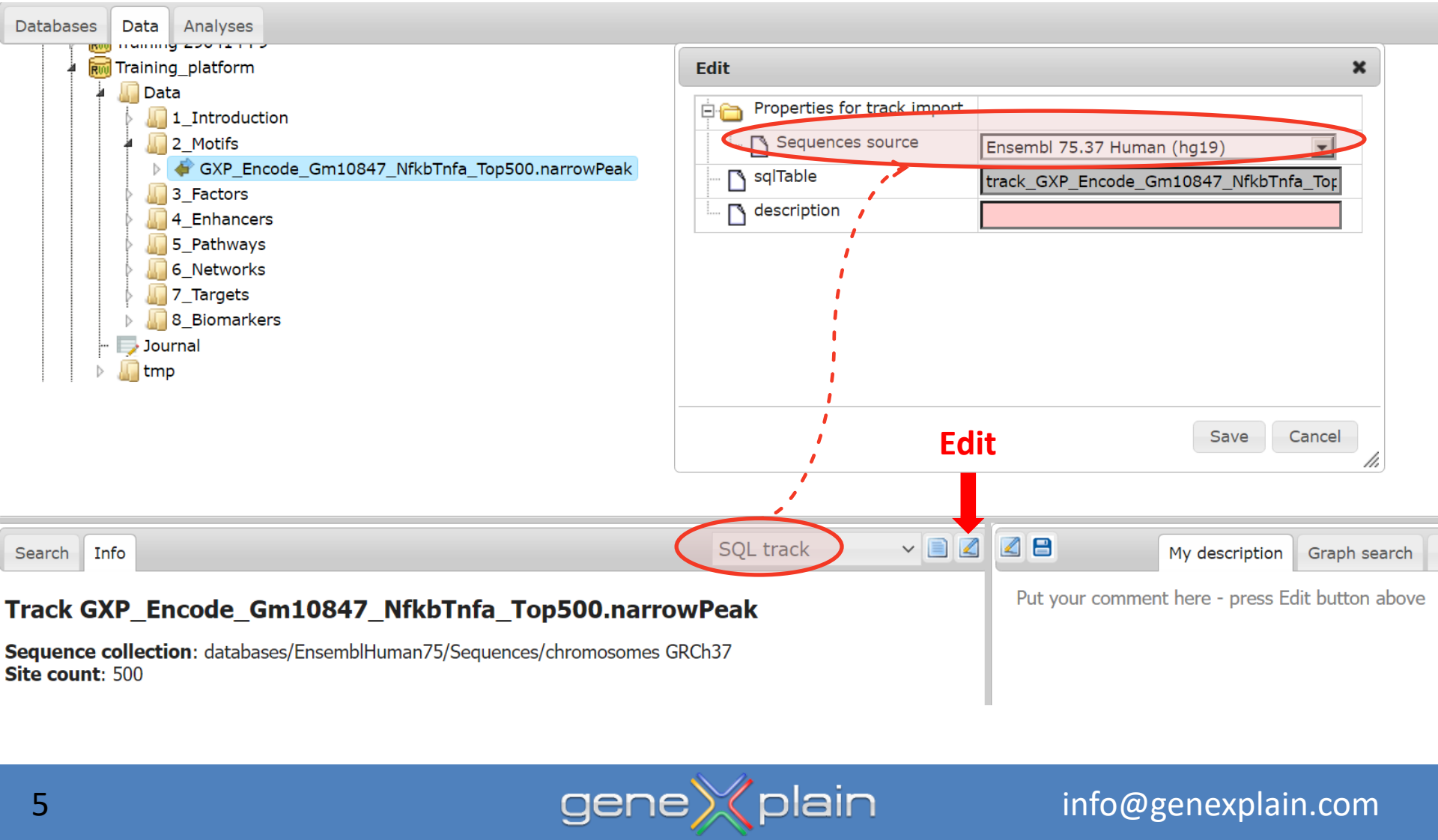


Part 2 – ChIPMunk in the geneXplain platform

We will start with the file *GXP_Encode_Gm10847_NfkbTnfa_Top500.narrowPeak.*

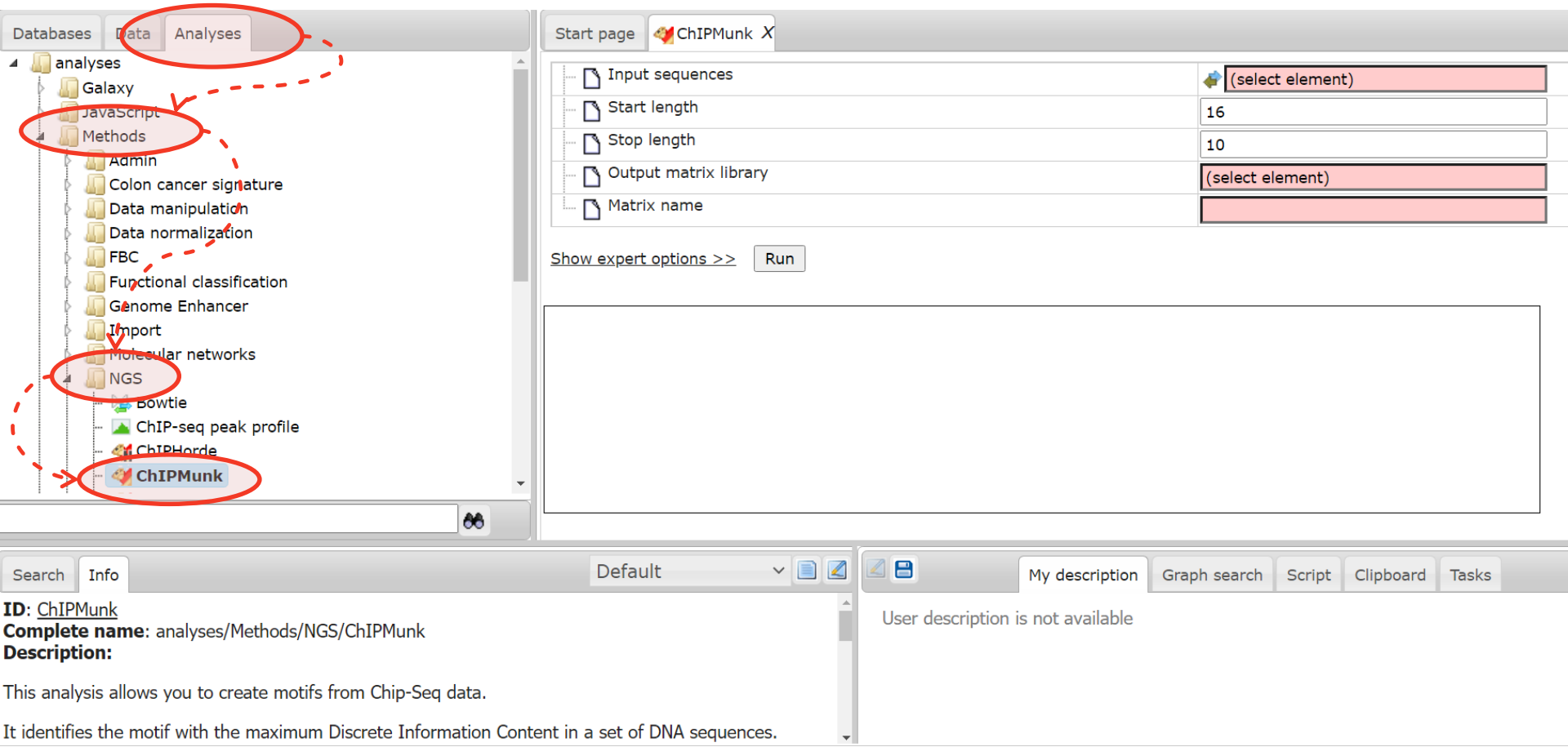
Select the file (turns blue marked) and press the **Edit** button to get more information of the file.

The format of the file *GXP_Encode_Gm10847_NfkbTnfa_Top500.narrowPeak* is a **track** and **hg19** genome is the *Sequence source*.



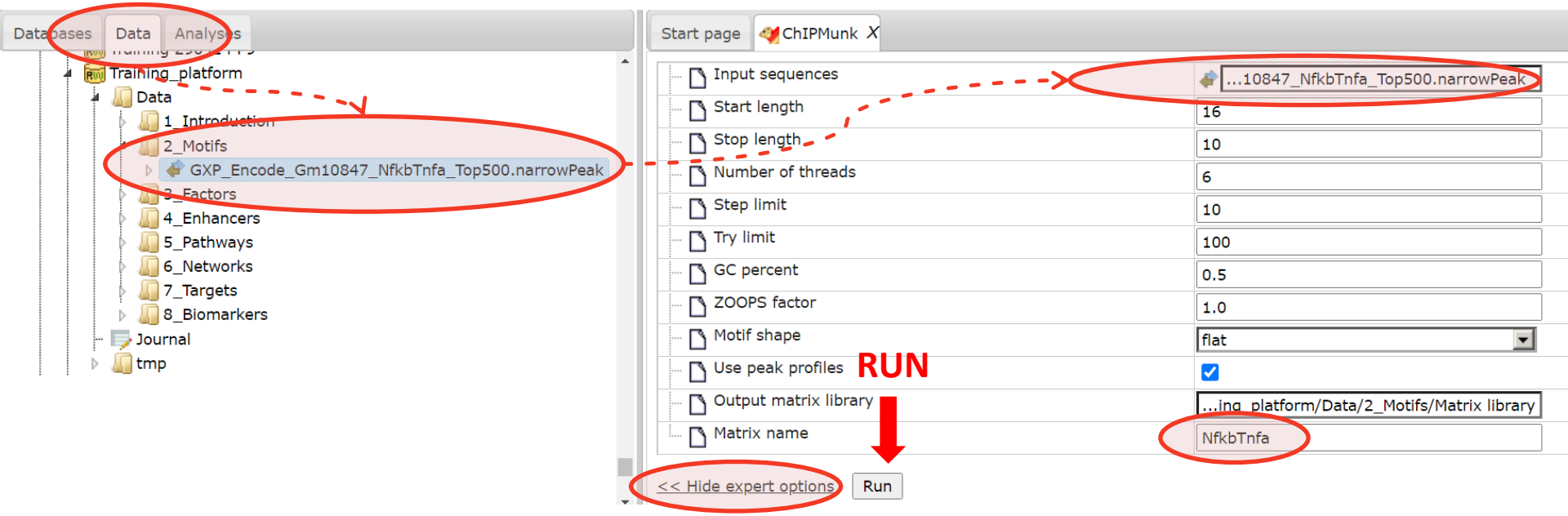
Part 2 – ChIPMunk in the geneXplain platform

When the ChIP-seq intervals have been loaded, navigate to the **ChIPMunk** tool as shown below and double-click on its name in the Working Tree to open the tool's interface.



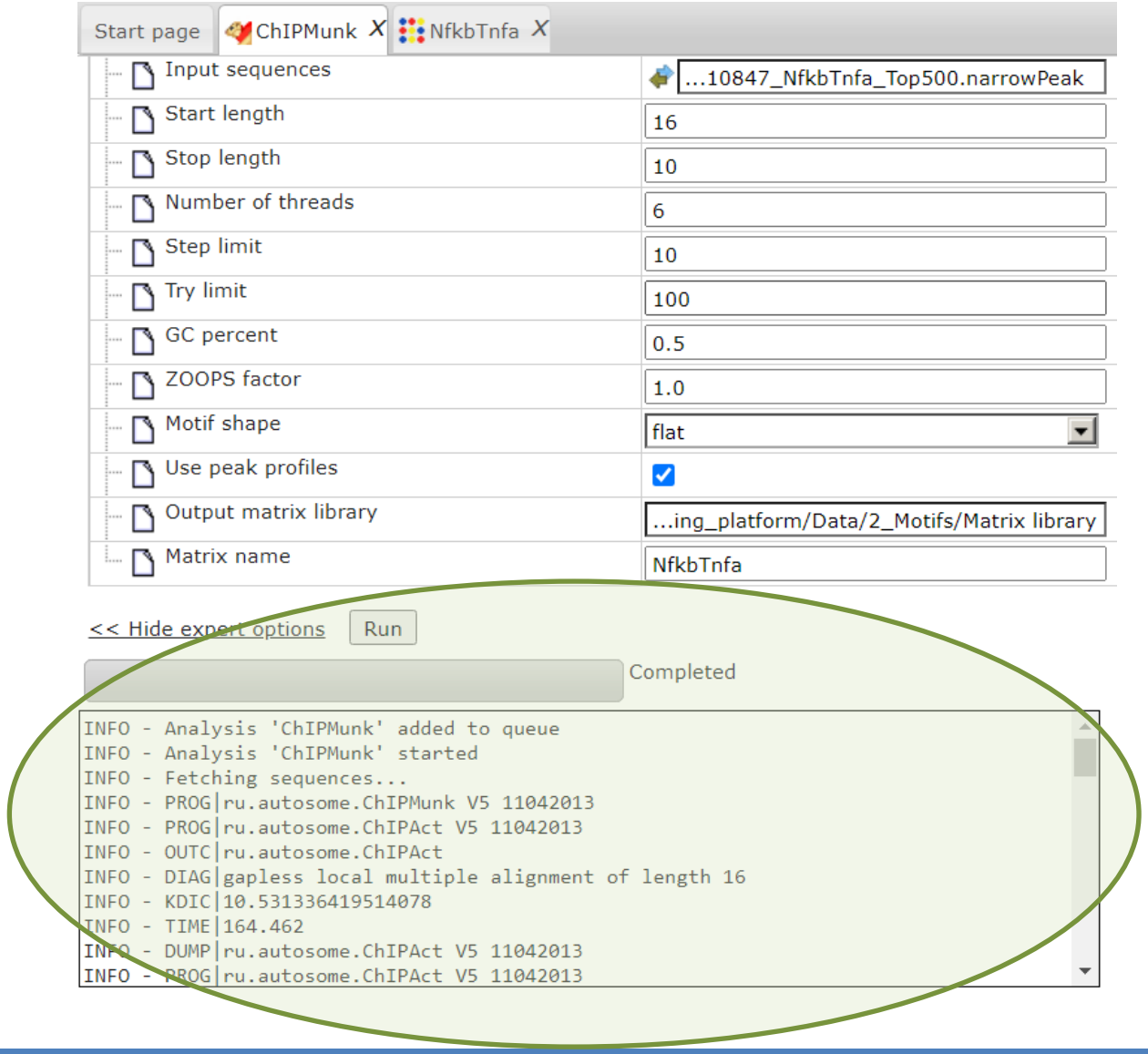
Part 2 – ChIPMunk in the geneXplain platform

Open the copied folder **2_Motifs**, click on the track in the Working Tree (blue marked) and drag-and-drop it to the **Input sequences** field in the tool interface. Expand the **Expert options** and see additional parameters of the tool, which can be modified. You can specify a special name for the output Matrix name (here: NfkbTnfa). When ready, click **Run**.



Part 2 – ChIPMunk in the geneXplain platform

It may take ChIPMunk a few minutes to complete ...

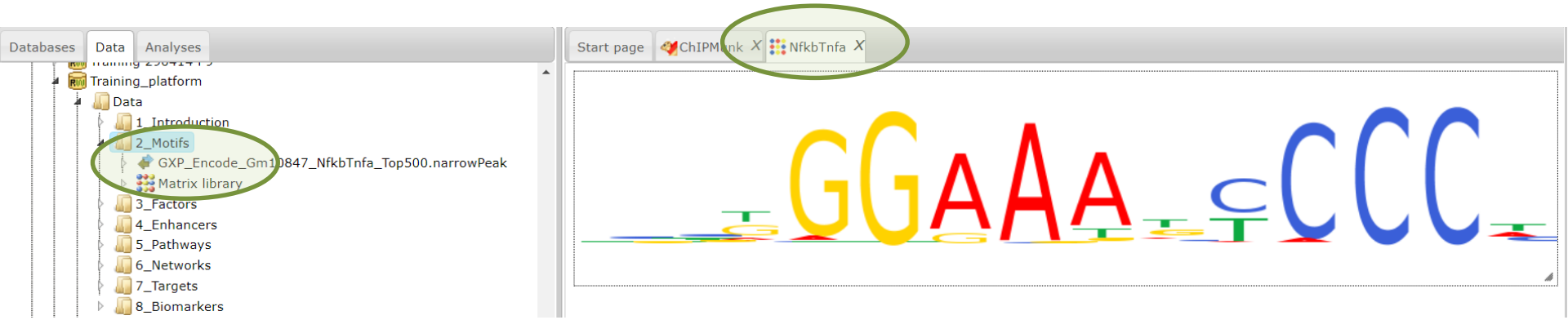


Part 1 – ChIPMunk in the geneXplain platform

When the task has finished, the platform presents a logo of the ChIPMunk motif and a new item named **Matrix library** (unless another had been specified) has been added to the tree.

The motif discovery part is complete at this point. If time permits you can

- *try out this analysis with DiChIPMunk which produces dinucleotide matrices (take care not to overwrite any of the results just produced, since they are required for the next part),*
- *or create different sorts of pretty motif logos for the ChIPMunk motif using the tool named "Create profile from matrix library" in the group of "Site analysis" tools.*



Part 2

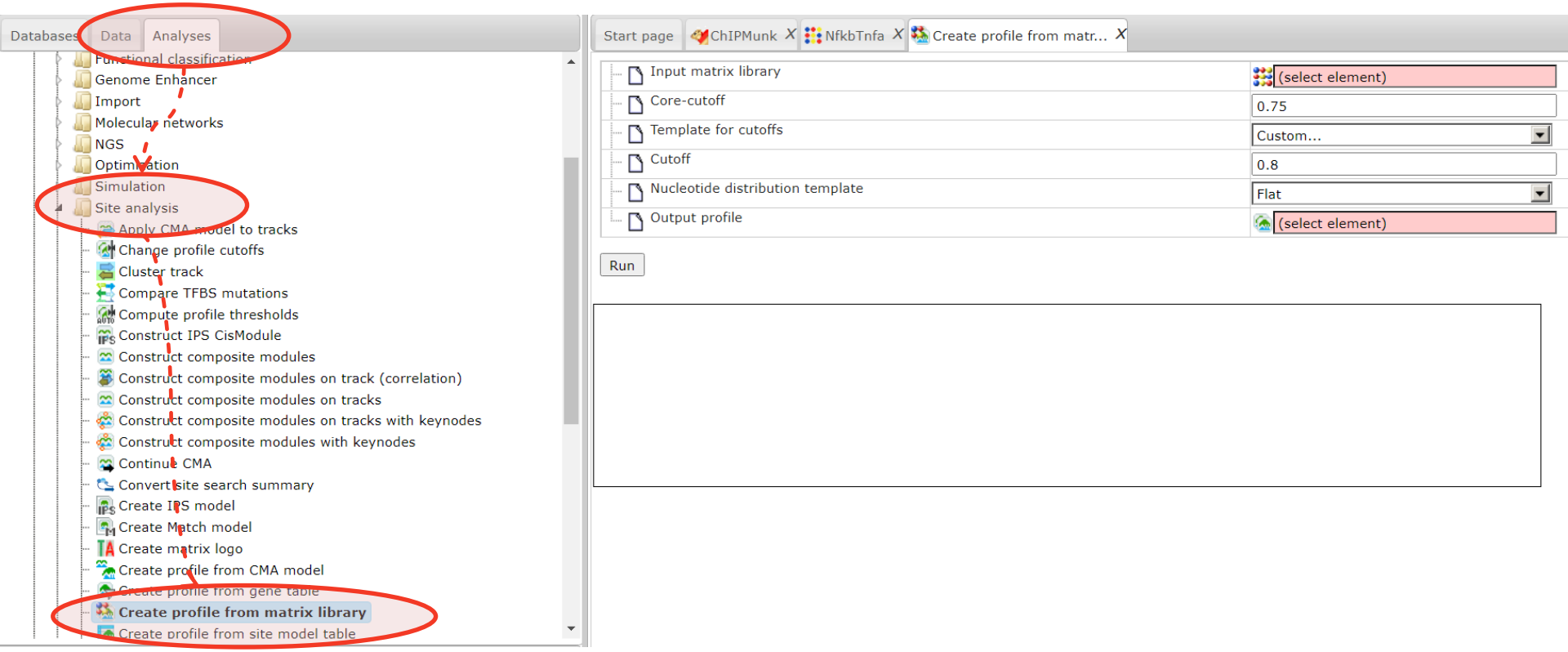
Here we will use the motif found by ChIPMunk in the previous part to predict binding sites in peak sequences uploaded as FASTA file.

info@genexplain.com

Part 2 – Binding site prediction in the geneXplain platform

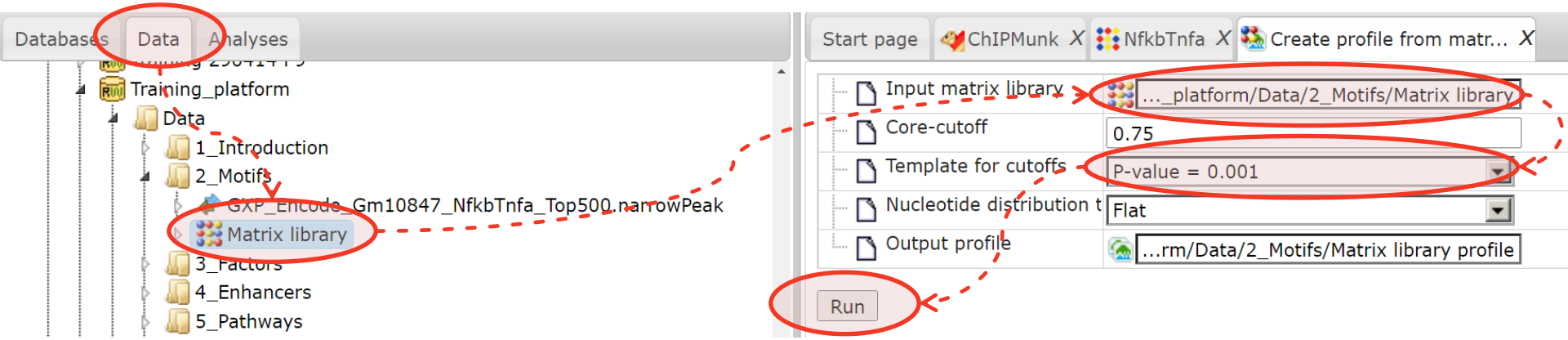
We need to create a *profile* for the matrix to use it in binding site search. A profile is a simple data structure to specify matrices and corresponding score cut-offs.

Navigate to the **Create profile from matrix library** tool as shown below and double-click on the item to open the tool interface.



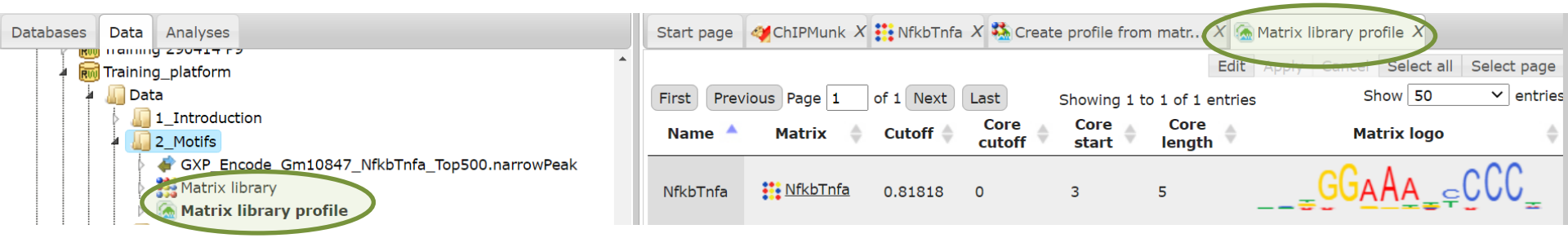
Part 2 – Binding site prediction in the geneXplain platform

Navigate to the matrix library created by ChIPMunk, click on the item and drag-and-drop it on the **Input matrix library** field of the profile tool. Specify a custom P-value of 0.001 for the score threshold. When ready, click **Run** to create the profile.



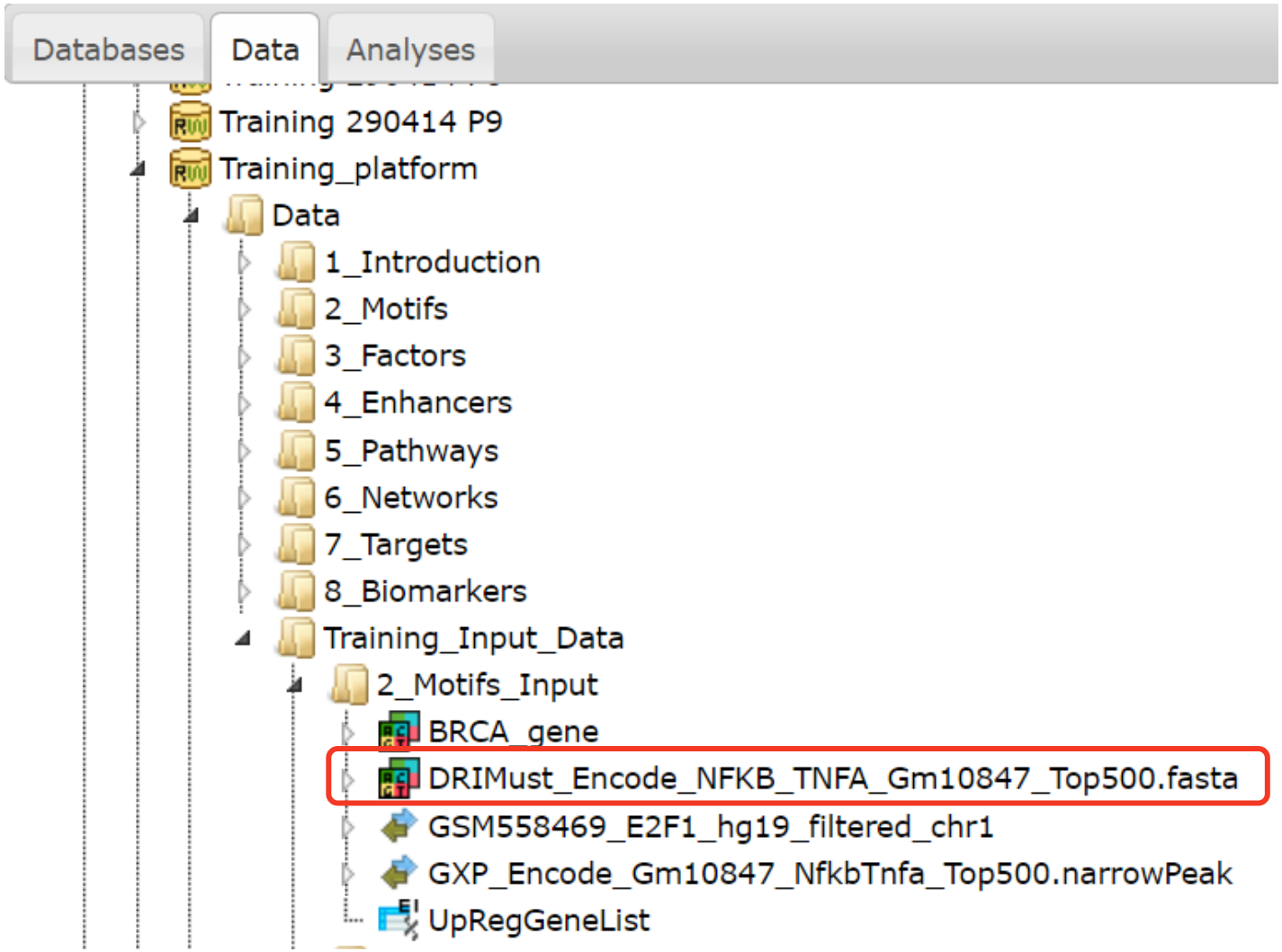
Part 2 – Binding site prediction in the geneXplain platform

A matrix library item is added to the tree and a view of the profile is shown, when the profile is ready. Note that a matrix profile can encompass parameters for several motifs.



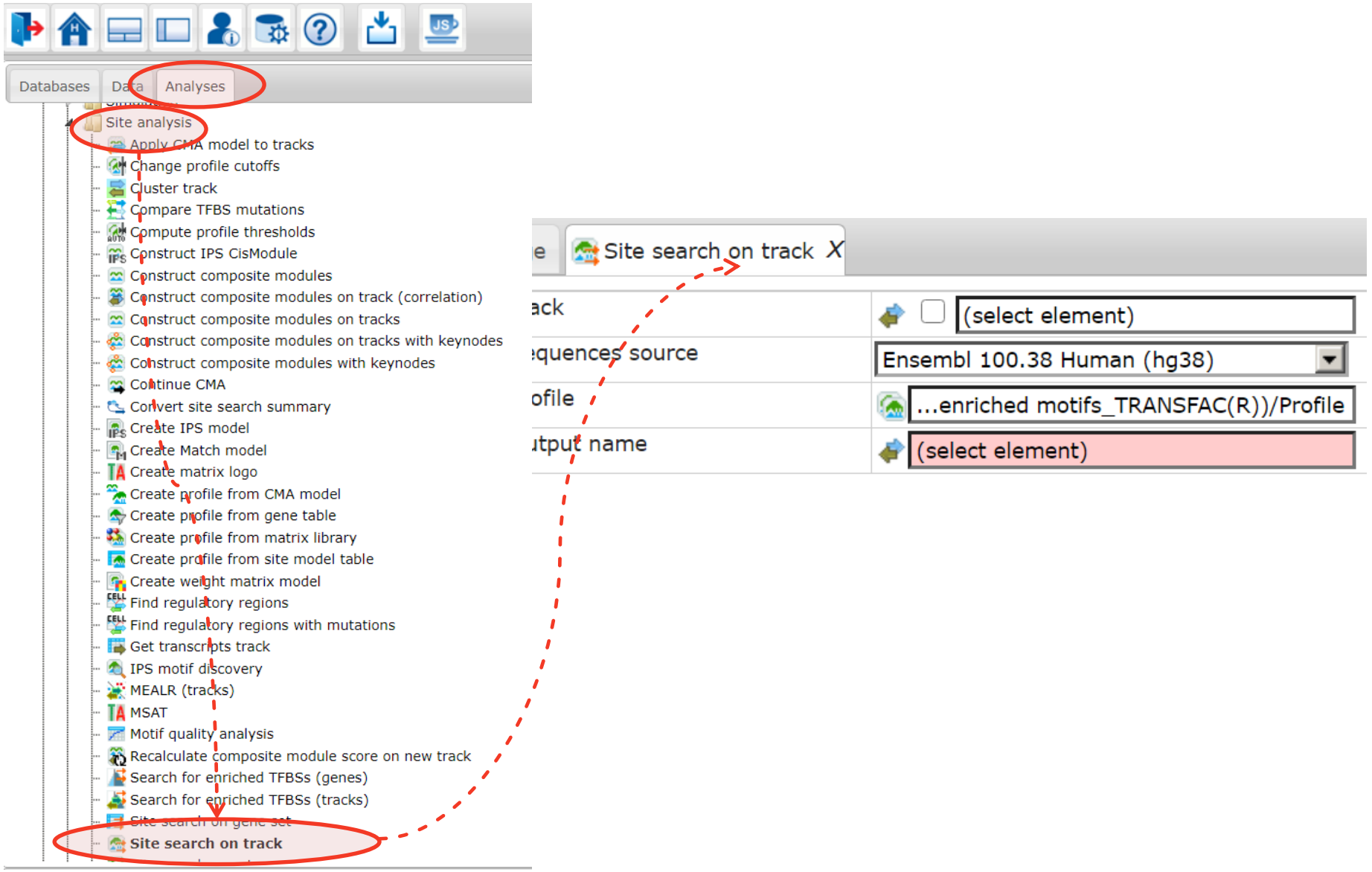
Part 2 – Binding site prediction in the geneXplain platform

We will now use the FASTA file named *GXP_Encode_Gm10847_NfkbTnfa_Top500.fasta and will predict enriched transcription factor binding sites (TFBSs).*



Part 2 – Binding site prediction in the geneXplain platform

Navigate to the **Site search on track** tool as shown below, double click and open the tool interface.

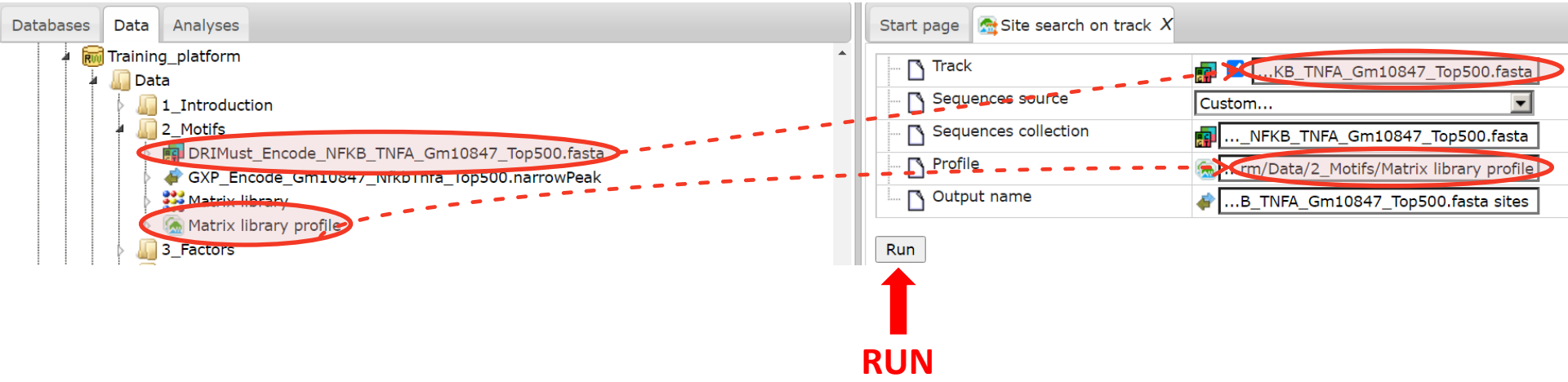


Part 2 – Binding site prediction in the geneXplain platform

In the tool interface drag-and-drop DRIMust_Encode_NFKB_TNFA_Gm10847_Top500.fasta into the field **Track.**

Drag-and-drop the recent created Matrix library file into the **Profile** field. You can specify a custom output name.

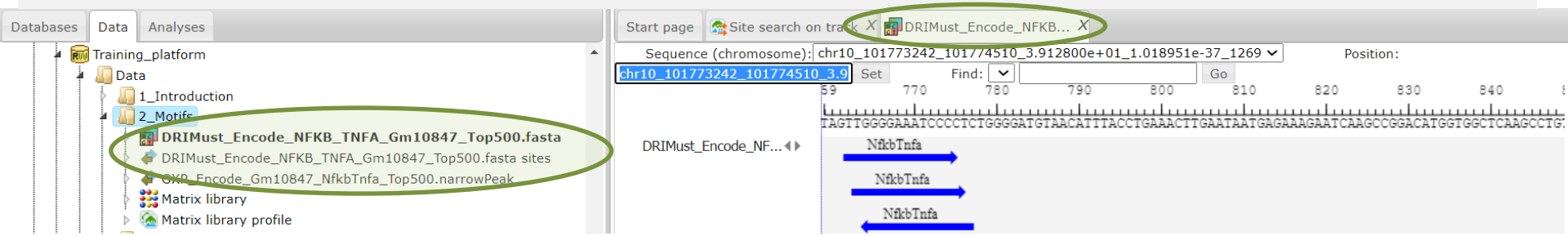
When ready, click **Run** to invoke the binding site search.

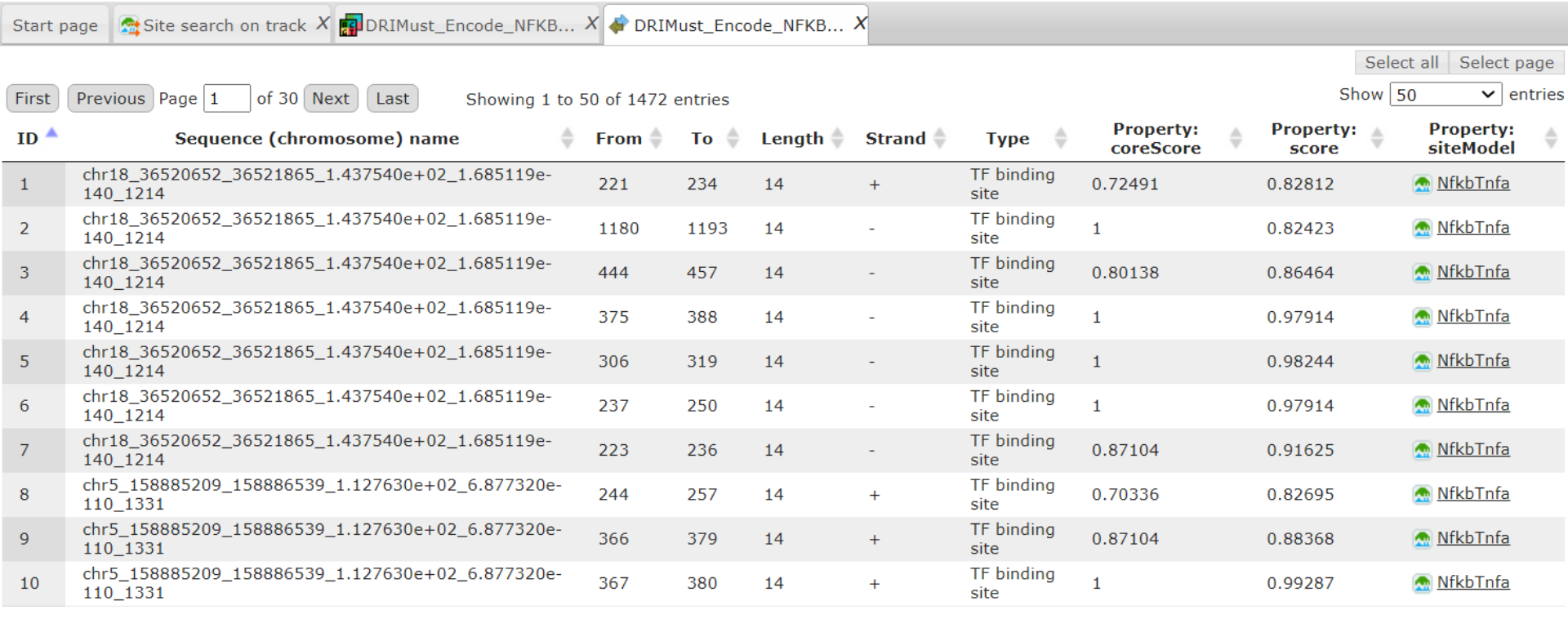


Part 2 – Binding site prediction in the geneXplain platform

When the binding site prediction has finished, you have a corresponding item in the Tree Area and a view of the sequences with binding sites is presented in the genome browser.

You can right-click on the **sites** track and open the corresponding table as an alternative to the visual presentation.





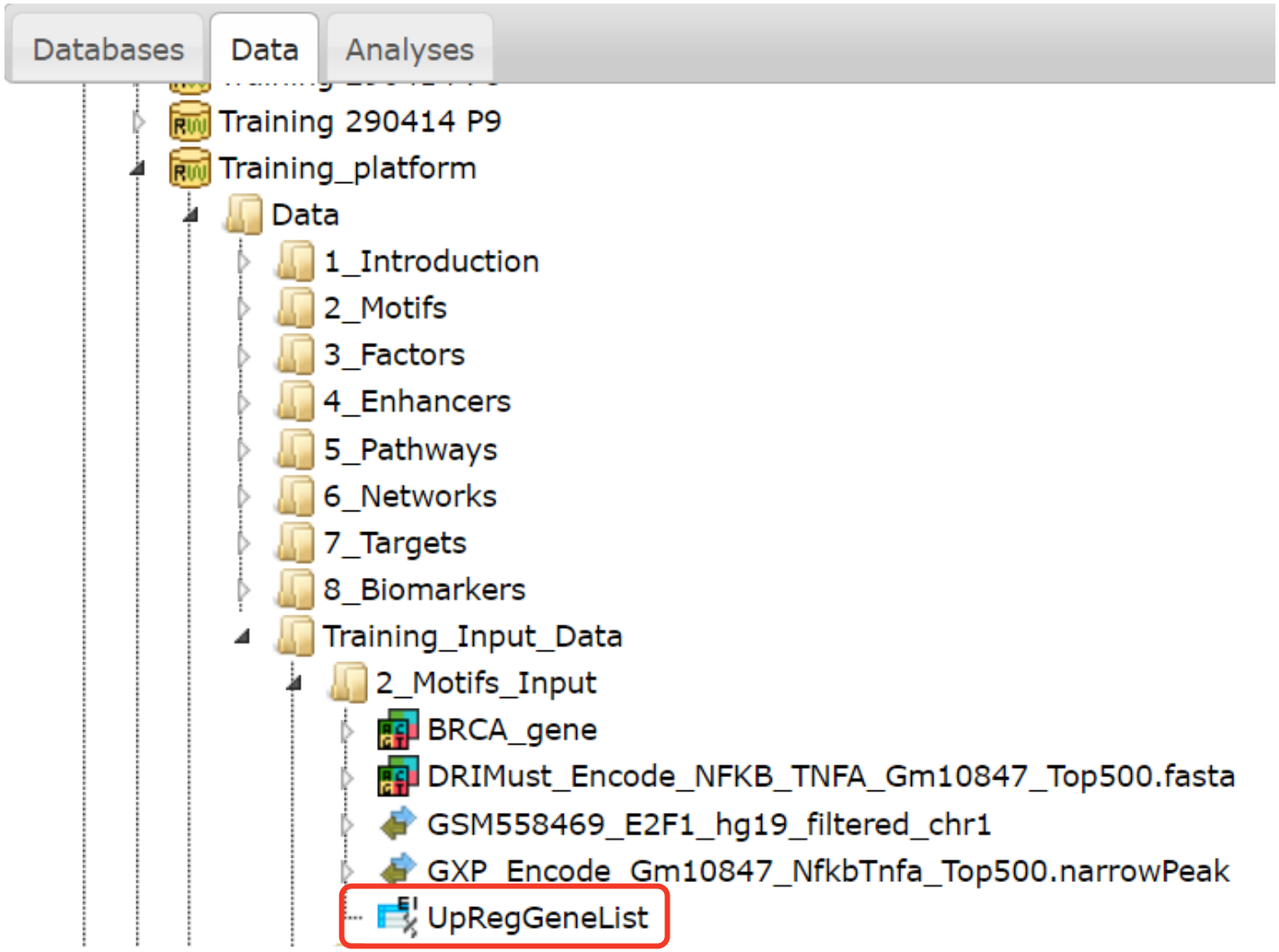
ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: coreScore	Property: score	Property: siteModel
1	chr18_36520652_36521865_1.437540e+02_1.685119e-140_1214	221	234	14	+	TF binding site	0.72491	0.82812	NfkbTnfa
2	chr18_36520652_36521865_1.437540e+02_1.685119e-140_1214	1180	1193	14	-	TF binding site	1	0.82423	NfkbTnfa
3	chr18_36520652_36521865_1.437540e+02_1.685119e-140_1214	444	457	14	-	TF binding site	0.80138	0.86464	NfkbTnfa
4	chr18_36520652_36521865_1.437540e+02_1.685119e-140_1214	375	388	14	-	TF binding site	1	0.97914	NfkbTnfa
5	chr18_36520652_36521865_1.437540e+02_1.685119e-140_1214	306	319	14	-	TF binding site	1	0.98244	NfkbTnfa
6	chr18_36520652_36521865_1.437540e+02_1.685119e-140_1214	237	250	14	-	TF binding site	1	0.97914	NfkbTnfa
7	chr18_36520652_36521865_1.437540e+02_1.685119e-140_1214	223	236	14	-	TF binding site	0.87104	0.91625	NfkbTnfa
8	chr5_158885209_158886539_1.127630e+02_6.877320e-110_1331	244	257	14	+	TF binding site	0.70336	0.82695	NfkbTnfa
9	chr5_158885209_158886539_1.127630e+02_6.877320e-110_1331	366	379	14	+	TF binding site	0.87104	0.88368	NfkbTnfa
10	chr5_158885209_158886539_1.127630e+02_6.877320e-110_1331	367	380	14	+	TF binding site	1	0.99287	NfkbTnfa

Part 3

Here we will use a workflow to find enriched TFBSs in promoters of a set of genes.

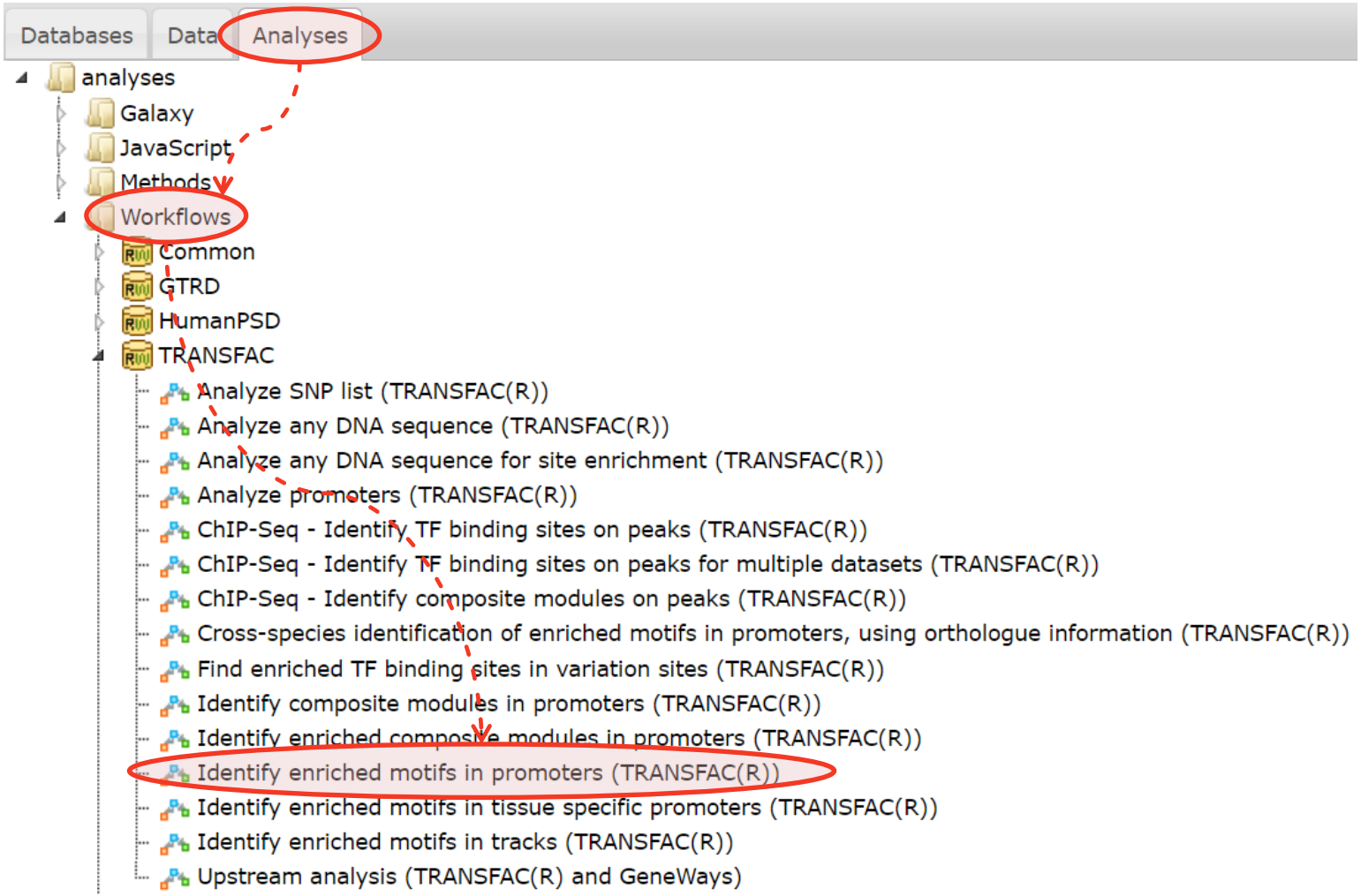
Part 3 – Binding site prediction in the geneXplain platform

We will now use the gene list named UpRegGeneList *and will predict enriched transcription factor binding sites (TFBSs) within the promoters of the gene list.*



Part 3 – Binding site prediction in the geneXplain platform

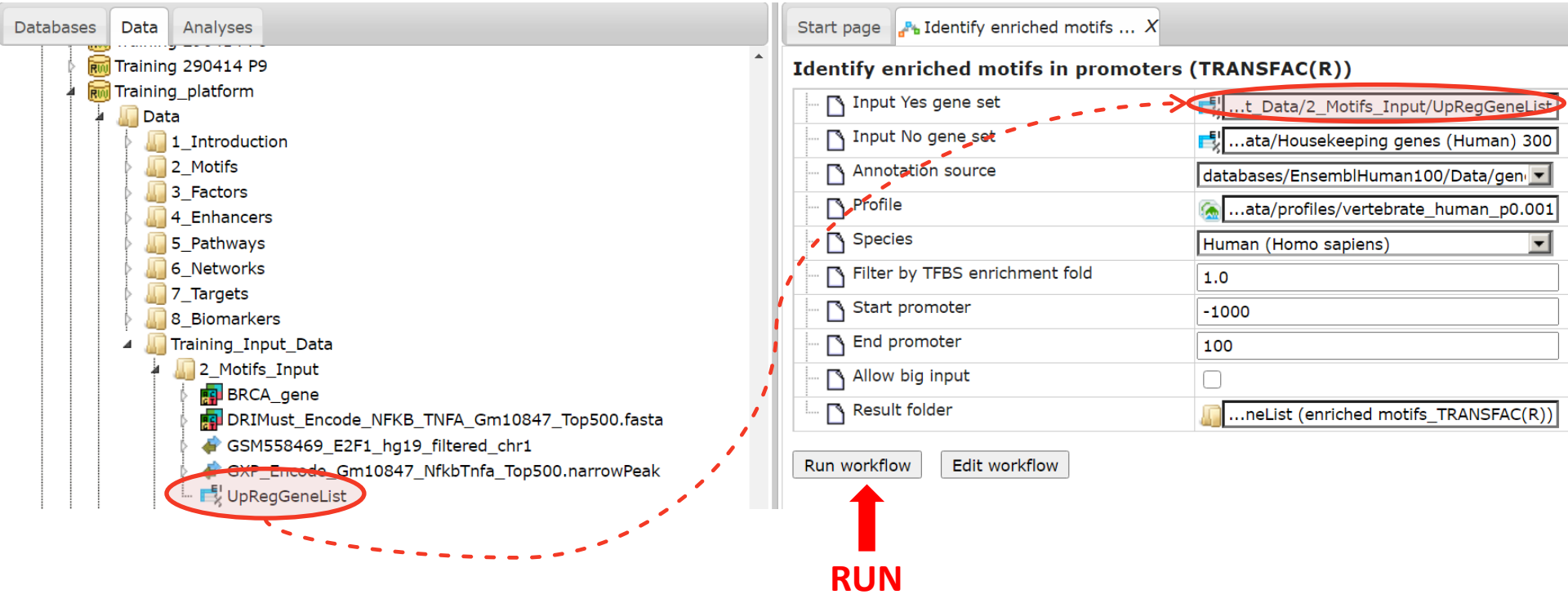
Navigate to the **Workflows** and further to **Identify enriched motifs in promoters (TRANSFAC®)** as shown below, double click and open the workflow interface.



Part 3 – Binding site prediction in the geneXplain platform

In the workflow interface drag-and-drop UpRegGeneList into the field **Input Yes gene set.**

Click **Run** to start the workflow.



Part 3 – Binding site prediction in the geneXplain platform

When the workflow has finished, several resulting tables will open automatically and finally a summary HTML **Report** as well. All results getting stored in an output folder and you can have a view of **Top 3 TFBS**, which are the three most enriched binding sites visualized in the gene promoters or all identified binding **yes sites** presented in the genome browser. A list of potential **Transcription factors** is given, which are postulated to regulate the genes from the input list.

REPORT

Data analysis is done with the geneXplain platform release 6.2
Project: data/Projects/00_jko_test
Date: Mon Nov 23 2020 13:20:51 GMT-0000 (UTC)

Workflow
Enriched transcription factor binding sites (TFBSs)
Visualization of top 3 enriched TFBSs
Potential identified transcription factors (TFs)

Workflow: Identify enriched motifs in promoters (TRANSFAC(R))

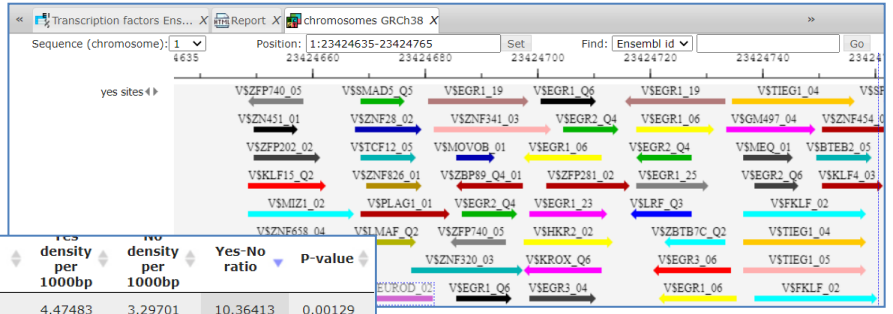
Workflow path: analyses/Workflows/TRANSFAC/Identify enriched motifs in promoters (TRANSFAC(R))

Enriched TFBSs

In this study, 732 transcription factor binding sites were identified as enriched, with filter >1 by TFBS enrichment fold.

Folder: Sites
Number of rows: 732
Complete name: data/Projects/Training_platform/Data/2_Motifs/UpRegGeneList (enriched motifs_TRANSFAC(R))/Sites

TRANSFAC ID	Matrix logo	Transcription factor	Yes-No ratio	P-value
V$ZFP740_05		Zfp740	1.2910	3.8857e-37
V$EGR2_Q4		Egr-2	1.2798	8.4026e-26



ID	Gene description	Gene symbol	Species	Site model ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	P-value
ENSG00000184937	WT1 transcription factor	WT1	Homo sapiens	V$WT1_03,V$WT1_Q4,V$WT1_Q6	4.47483	3.29701	10.36413	0.00129
ENSG00000178187	zinc finger protein 454	ZNF454	Homo sapiens	V$ZNF454_05,V$ZNF454_06,V$ZNF454_07,V$ZNF454_08	4.61564	3.57721	9.89304	6.6097E-5
ENSG00000100968	nuclear factor of activated T cells 4	NFATC4	Homo sapiens	V$NFATC4_03,V$NFAT_Q4_01	0.2112	0.09651	8.47974	8.9124E-4
ENSG00000163884	Kruppel like factor 15	KLF15	Homo sapiens	V$KLF15_01,V$KLF15_03,V$KLF15_05,V$KLF15_07,V$KLF15_Q2	7.33046	5.63823	8.47974	3.8808E-5
ENSG00000197647	zinc finger protein 433	ZNF433	Homo sapiens	V$ZNF433_03	0.0528	0.00623	8.47974	8.9124E-4
ENSG00000091010	POU class 4 homeobox 3	POU4F3	Homo sapiens	V$POU4F3_02,V$POU4F3_03,V$POU4F3_05,V$POU4F3_06,V$POU4F3_07,V$POU4F3_08,V$POU4F3_09,V$POU4F3_10	1.20561	0.69427	7.14497	1.7311E-9

info@genexplain.com

Practical session completed

info@genexplain.com