Identification of DNA motifs

Hands-on-training



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.



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

Specify your credentials and click Login.

🌍 💿 🙆 platfo	platform.genexplain.com/bioumlweb/#										
	Login X										
	Platform: geneXplain web edition 3.0										
	Enter e-mail and password:										
	E-mail: <your username=""></your>										
	Password:										
	Register										



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**.





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.



Sequence collection: databases/EnsemblHuman75/Sequences/chromosomes GRCh37 Site count: 500



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.

Databases Lata Analyses	Start page 🛛 🍕 ChIPMunk	x							
Galaxy JavaScript Methods Admin Colon cancer signature Data manipulation Data normalization FBC Functional classification Genome Enhancer Inport Moreular networks NGS Bowtie ChIP-seq peak profile ChIPHorde ChIPHunk	Show expert options >>	Run			(select 16 10 (select el	ement)			
Search Info	Default	~ 🗎 🛛		My description	Graph search	Script	Clipboard	Tasks	
ID: <u>ChIPMunk</u> Complete name: analyses/Methods/NGS/ChIPMunk Description:		•	User description i	s not available					
It identifies the motif with the maximum Discrete Information Cont	ent in a set of DNA sequenc	es. 🗸							



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**.





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

Input sequences	…10847_NfkbTnfa_Top500.narrowPeak
Start length	16
Stop length	10
Number of threads	6
Step limit	
	10
	100
GC percent	0.5
D ZOOPS factor	1.0
🖸 Motif shape	flat
🗋 Use peak profiles	
🖸 Output matrix library	ing_platform/Data/2_Motifs/Matrix libra
Matrix name	NfkbTnfa
c c Hide evener entires Dun	
Kun	
	Completed
NFO - Analysis 'ChIPMunk' added to queue	
NFO - Analysis 'ChIPMunk' started	
NFO - Fetching sequences NEO - PROGINU autosome ChTPMunk V5 110/2013	
NEO - PROG ru.autosome.ChTPAct V5 11042013	
NFO - OUTC/ru.autosome.ChIPAct	
NFO - DIAG gapless local multiple alignment o	of length 16
NEO - KDTC 10 531336/1951/078	5
NFO - KDIC 10.00100419014070	
NFO - TIME 164.462	
NFO - TIME 164.462 NFO - DUMP ru.autosome.ChIPAct V5 11042013	

gene

olain

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.



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.

Databases Data Analyses	Start page 🛛 🥙 ChIPMunk X 👯 NfkbTnfa X 🌺 Create profile from matr	x
Unitional classification Genome Enhancer	Input matrix library	(select element)
🕨 🛺 Import	🔤 Core-cutoff	0.75
Molecular networks	Template for cutoffs	Custom
	Cutoff	0.8
Simulation	Nucleotide distribution template	Flat
Site analysis	Line 🖸 Output profile	(select element)
- Change profile cutoffs		
🧧 Cluster track	Run	
Compute profile thresholds		
Construct composite modules		
- 🕱 Construct composite modules on track (correlation)		
- Construct composite modules on tracks		
Construct composite modules on tracks with keynodes		
- Continute CMA		
🚰 Convert I site search summary		
💦 Create IPS model		
Create Match model		
- Create profile from CMA model		
- Create profile from gene table		
Create profile from matrix library		
Create profile from site model table		



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.





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.





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





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

🕨 🏫 🗔 🚨 🕾 💿 📩 📟		
Databases Data Analyses		
Databases Data Analyses Site analysis Apply C in model to tracks Apply C in model to tracks Cluster track Compare TFBS mutations Compute profile thresholds Construct IPS CisModule Construct composite modules on track (correlation) Construct composite modules on tracks Construct composite modules on tracks Construct composite modules on tracks with keynodes Construct composite modules on tracks Construct composite modules on tracks Construct composite modules on tracks Construct composite modules on tracks with keynodes Construct composite modules on tracks Construct composite modules on tracks Construct composite modules Create profile from gene table Create profile from site model table	e Site search on track X ack equences source ofile itput name	(select element) Ensembl 100.38 Human (hg38) enriched motifs_TRANSFAC(R))/Profile (select element)
 Search for enriched TFBSs (genes) Search for enriched TFBSs (tracks) Site search on gene set Site search on track 		



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.





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.





Part 3

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



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.





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





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

Click **Run** to start the workflow.





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								« Start pag	e 🚜 Identif	y enriched motif	s X 📑 UpRegG	eneList X 🝖 Prof	file X 📑 Site	e search -1000 100	X 🛃 Торз ТFBS 🗴 🖪	Sites X	>>	
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)							First Previo	us Page 1	of 5 Next	Last Showin	g 1 to 50 of 208 e	entries Sites	s view	Edit Apply Ca	Cel Select a	Total count	ge v\$.	
Workflow Enriched transcription factor binding sites (TFBSs) Visualization of top 3 enriched TFBSs Potential identified transcription factors (TFs)															e e e e e e e e e e e e e e e e e e e			
Workflow: Identify enriched motifs in promoters (TRANSFAC(R))						ENSG000000	43355 ZI	C2							199	71		
Workflow path: analyses/Workflows/TRANSFAC/Identify enriched motifs in promoters (TRANSFAC(R))																		
Enriched TFBSs																		
In this study, 732 tran	nscription factor bin	ding sites we	ere identified	as enriched, with f	filter >1 by TFBS enric	hment fold.					•							
Folder: Sites Number of rows: 732											e (n (n e e) (n (n)	• •	****	100 100 4 100 100 100 100 100 4 000 0	den den den den den den den b b den	GPC1		
Complete name: da	ata/Projects/Trainin Matrix	g_platform/[. logo	Data/2_Motifs	UpRegGeneList (e	enriched motifs_TRAN	SFAC(R))/Sites	« 🛃 TI	ranscription fa	tors Ens	X Report X	chromosome:	GRCh38 X	1.5.1				»	
V\$ZFP740_05	Ç	cCCC	CCA.	Zfp740	1.2910	3.8857e-37	Sequ	yes site	4635 s4	V\$ZFP74 V\$ZN45	0_05 V\$SM	23424765 23424680 AD5_Q5 V\$E VF28_02 V\$	EGR1_19 \$ZNF341_03	V\$EGR1_Q6 V\$EGR2_Q4	V\$EGR1_06	23424740 I V\$TIEG1 /\$GM497_04	04 V V\$ZNF45	.24 \$SF
V\$EGR2_Q4	_G	<u>zgG(</u>	, ₽ ₽ ₽ ₽ ₽ ₽ ₽	Egr-2	1.2798	8.4026e-26				V\$ZFP V\$KLF	202_02 V\$T 715_Q2 V\$Z	XF12_05 V\$MC	OVOB_01 BP89_Q4_01	V\$EGR1_06 V\$ZFP281_02	VSEGR2_Q4	V\$MEQ_01 V\$EGR2_Q6	V\$BTEB2_05 V\$KLF4_0	5 03
										V\$Z	VF658 04 VSL	MAF Q2 V\$ZF	FP740 05	V\$EGR1_23	V\$ZBTB7C Q2	VSFKI	G1 04	
ID 🔶	Gene description 🖨	Gene symbol 🔶	Species 🔷		Site model ID		♦ de 10	nsity d ber 00bp 1	ensity per 000bp	Yes-No ratio	P-value 🔷	VSZNF	320_03	V\$KROX_Q6	V\$EGR3_06	V\$TIE V\$F	G1_05	
ENSG00000184937	WT1 transcription factor	WT1	Homo sapiens	V\$WT1_03,V\$WT1_Q	24,V\$WT1_Q6		4.4	7483 3	.29701	10.36413	0.00129				(_		•
ENSG00000178187	zinc finger protein 454	ZNF454	Homo sapiens	V\$ZNF454_05,V\$ZNF	454_06,V\$ZNF454_07,V\$	ZNF454_08	4.6	31564 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			0.2	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			7.3	3046 5	63823	8.47974	3.8808E- 5							
ENSG0000197647	zinc finger protein 433	ZNF433	Homo sapiens	V\$ZNF433_03			0.0	0528 0	.00623	8.47974	8.9124E- 4							
ENSG0000091010	POU class 4 homeobox 3	POU4F3	Homo sapiens	V\$POU4F3_02,V\$POU V\$POU4F3_08,V\$POU	J4F3_03,V\$POU4F3_05,V\$ J4F3_09,V\$POU4F3_10	POU4F3_06,V\$POU4F3_07	7, 1.2	0561 0	.69427	7.14497	1.7311E- 9							



Practical session completed

