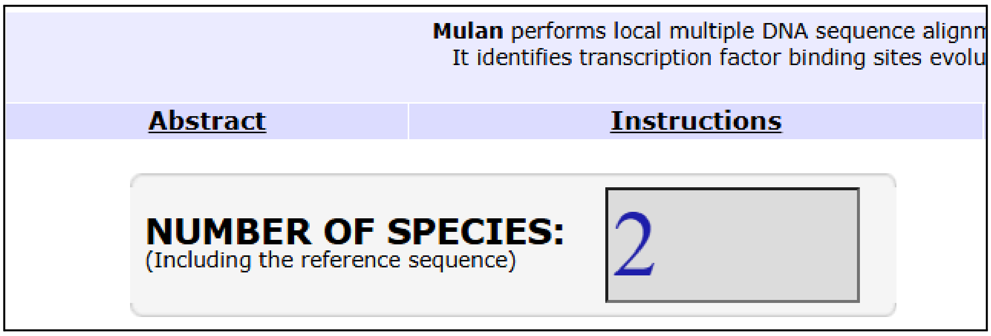
Supplementary Document 1. Detailed workflow for predicting Transcription Factor Binding Sites (TFBS) using the MultiTF tool within the Mulan platform (MUltiple sequence Local AligNment and conservation visualization).

---

1. Visit Mulan website (https://mulan.dcode.org/).
2. Select the desired number of species for the alignment (see Figure 1).

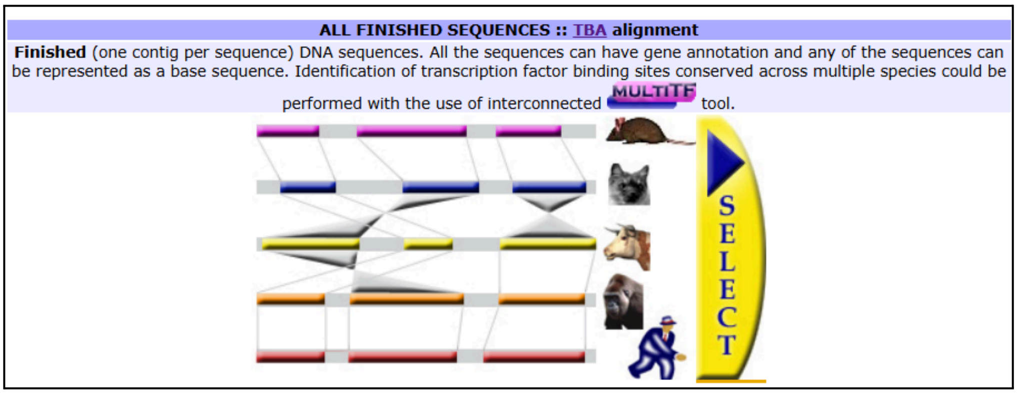


Mulan performs local multiple DNA sequence alignm
It identifies transcription factor binding sites evolu

NUMBER OF SPECIES: 2
(Including the reference sequence)

Figure 1

3. In the "ALL FINISHED SEQUENCES :: TBA alignment" section, click on the “**SELECT**” option to access the MultiTF tool (see Figure 2).

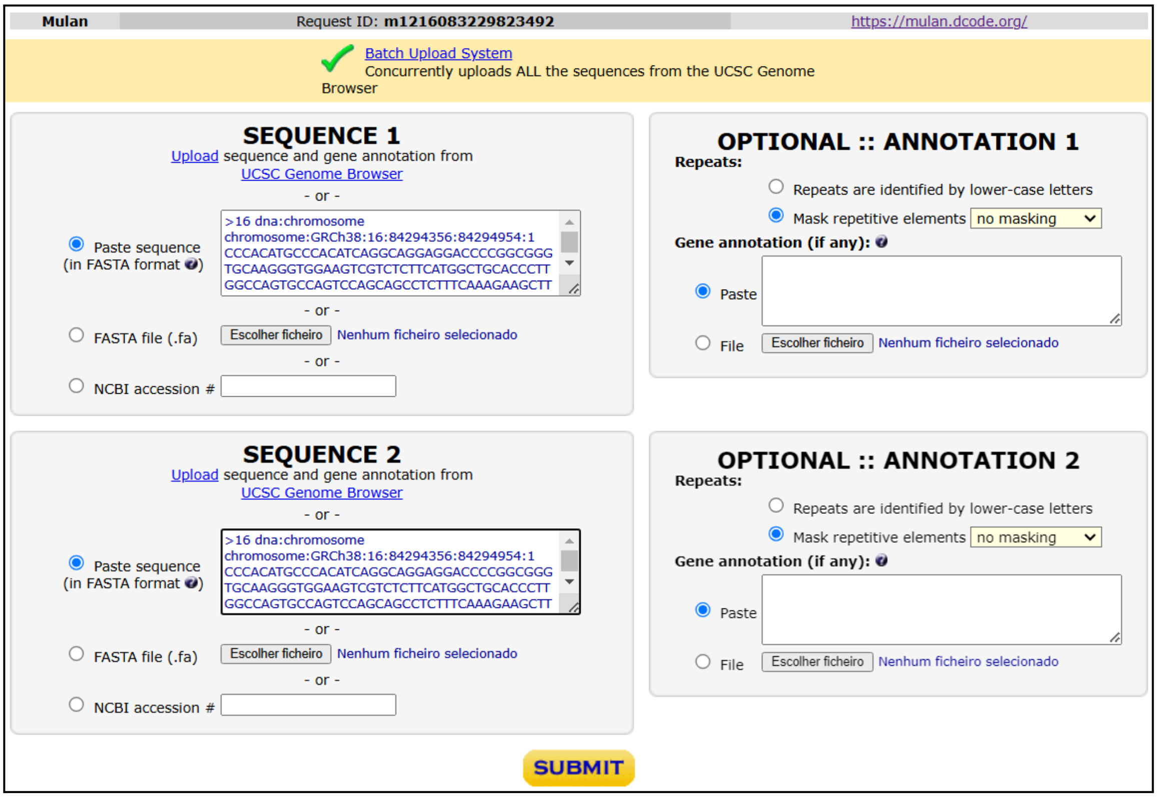


ALL FINISHED SEQUENCES :: TBA alignment

Finished (one contig per sequence) DNA sequences. All the sequences can have gene annotation and any of the sequences can be represented as a base sequence. Identification of transcription factor binding sites conserved across multiple species could be performed with the use of interconnected MULTITF tool.

Figure 2

4. Upload the sequences in FASTA format into the designated fields (see Figure 3), then click “**SUBMIT**”. For regulatory element analysis, providing an “Annotation” file is not required.



Mulan Request ID: **m1216083229823492** https://mulan.dcode.org/

Batch Upload System
Concurrently uploads ALL the sequences from the UCSC Genome Browser

SEQUENCE 1
Upload sequence and gene annotation from UCSC Genome Browser
- or -

● Paste sequence (in FASTA format)

```
>16 dna:chromosome
chromosome:GRCh38:16:84294356:84294954:1
CCCACATGCCCACATCAGGCAGGAGGACCCCGGCGGG
TGCAAGGGTGGAAGTCGTCTCTTCATGGCTGCACCCTT
GGCCAGTGCCAGTCCAGCAGCCTCTTTCAAAGAAGCTT
```

- or -

○ FASTA file (.fa) Escolher ficheiro Nenhum ficheiro selecionado

- or -

○ NCBI accession #

OPTIONAL :: ANNOTATION 1

Repeats:
○ Repeats are identified by lower-case letters
● Mask repetitive elements no masking

Gene annotation (if any):
● Paste
○ File Escolher ficheiro Nenhum ficheiro selecionado

SEQUENCE 2
Upload sequence and gene annotation from UCSC Genome Browser
- or -

● Paste sequence (in FASTA format)

```
>16 dna:chromosome
chromosome:GRCh38:16:84294356:84294954:1
CCCACATGCCCACATCAGGCAGGAGGACCCCGGCGGG
TGCAAGGGTGGAAGTCGTCTCTTCATGGCTGCACCCTT
GGCCAGTGCCAGTCCAGCAGCCTCTTTCAAAGAAGCTT
```

- or -

○ FASTA file (.fa) Escolher ficheiro Nenhum ficheiro selecionado

- or -

○ NCBI accession #

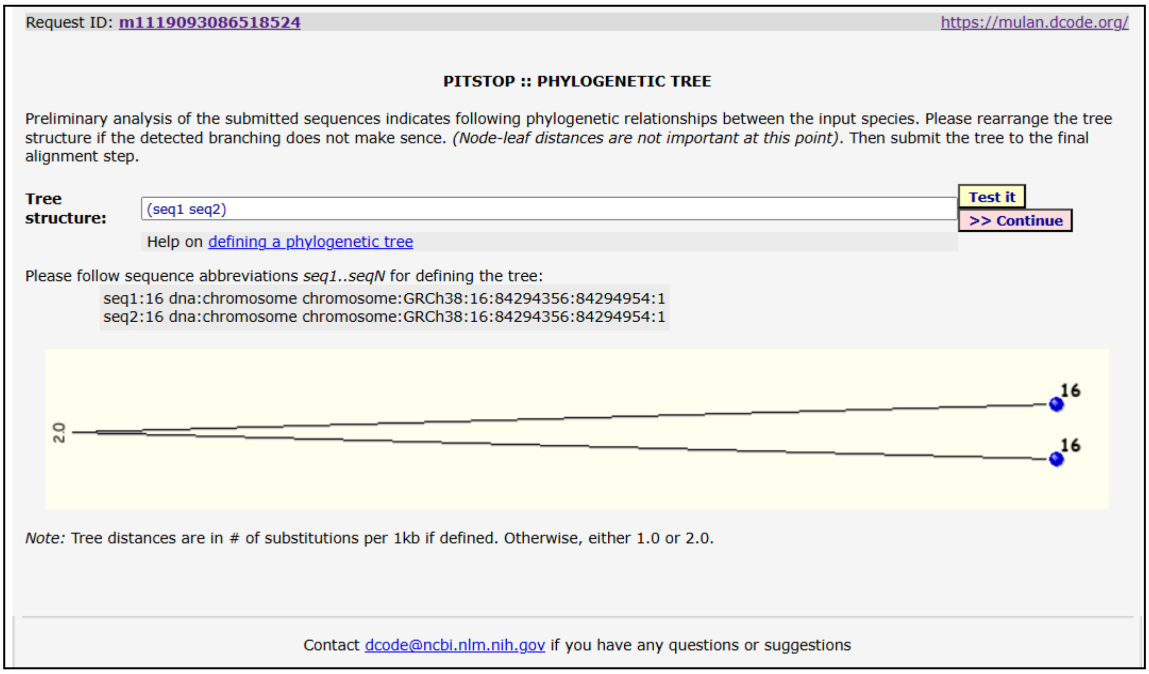
OPTIONAL :: ANNOTATION 2

Repeats:
○ Repeats are identified by lower-case letters
● Mask repetitive elements no masking

Gene annotation (if any):
● Paste
○ File Escolher ficheiro Nenhum ficheiro selecionado

Figure 3

5. Once the sequences are submitted, the system generates a “pitstop” page (Figure 4) that presents the phylogenetic relationships among the compared sequences. Press “**Continue**” on this page (Figure 4).



Request ID: **m119093086518524** https://mulan.dcode.org/

PITSTOP :: PHYLOGENETIC TREE

Preliminary analysis of the submitted sequences indicates following phylogenetic relationships between the input species. Please rearrange the tree structure if the detected branching does not make sence. *(Node-leaf distances are not important at this point)*. Then submit the tree to the final alignment step.

Tree structure: (seq1 seq2)

Help on defining a phylogenetic tree

Please follow sequence abbreviations *seq1..seqN* for defining the tree:

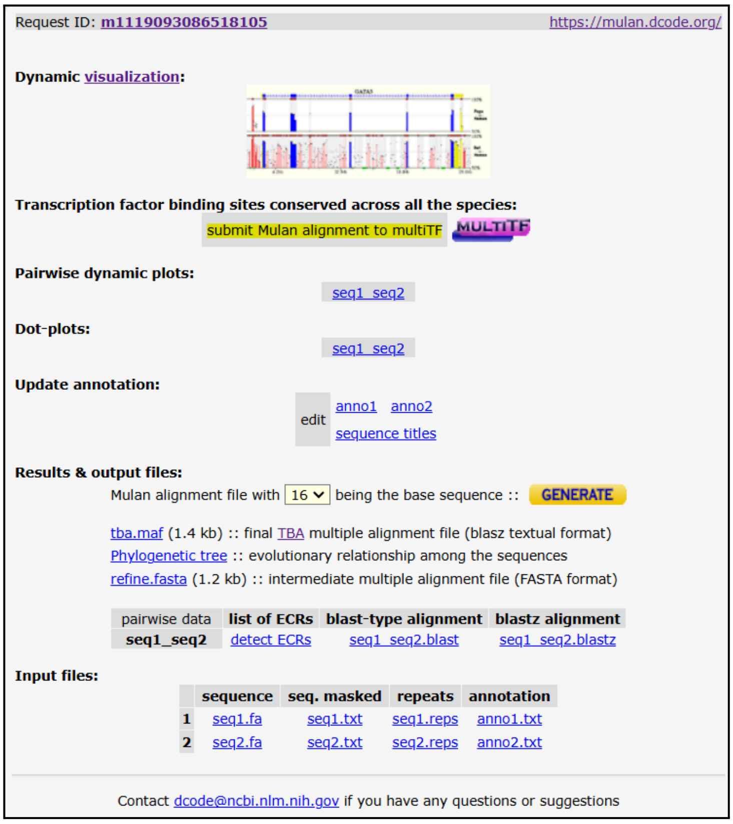
```
seq1:16 dna:chromosome chromosome:GRCh38:16:84294356:84294954:1
seq2:16 dna:chromosome chromosome:GRCh38:16:84294356:84294954:1
```

Note: Tree distances are in # of substitutions per 1kb if defined. Otherwise, either 1.0 or 2.0.

Contact dcode@ncbi.nlm.nih.gov if you have any questions or suggestions

Figure 4

6. Select the "**MultiTF**" option to proceed with TFBS prediction (Figure 5).



Request ID: m1119093086518105 https://mulan.dcode.org/

Dynamic visualization:

Transcription factor binding sites conserved across all the species:
submit Mulan alignment to multiTF MULTITF

Pairwise dynamic plots:
seq1_seq2

Dot-plots:
seq1_seq2

Update annotation:
edit anno1 anno2
sequence titles

Results & output files:
Mulan alignment file with 16 being the base sequence :: GENERATE

tba.maf (1.4 kb) :: final TBA multiple alignment file (blasz textual format)
Phylogenetic tree :: evolutionary relationship among the sequences
refine.fasta (1.2 kb) :: intermediate multiple alignment file (FASTA format)

pairwise data	list of ECRs	blast-type alignment	blastz alignment
seq1_seq2	detect ECRs	seq1_seq2.blast	seq1_seq2.blastz

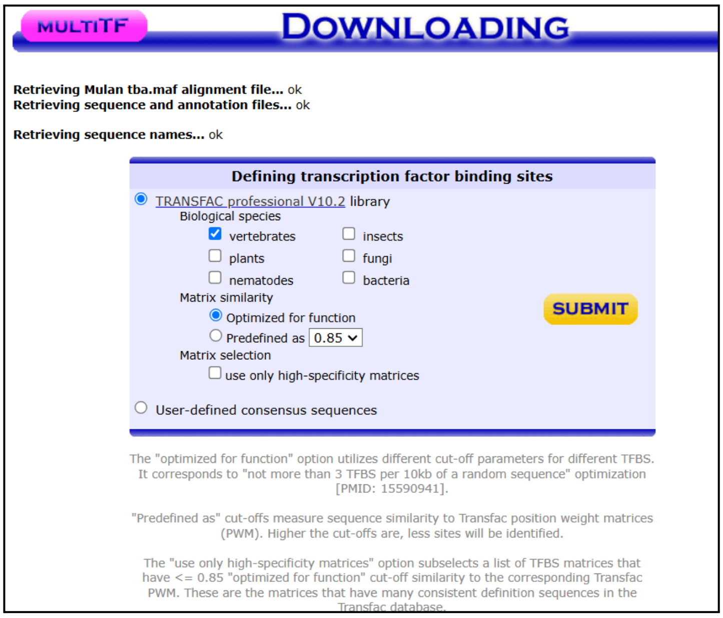
Input files:

	sequence	seq. masked	repeats	annotation
1	seq1.fa	seq1.txt	seq1.reps	anno1.txt
2	seq2.fa	seq2.txt	seq2.reps	anno2.txt

Contact dcode@ncbi.nlm.nih.gov if you have any questions or suggestions

Figure 5

7. On the following page, select "TRANSFAC professional V10.2 library" > "vertebrates" > "Optimized for function" (Figure 6), then click "**SUBMIT**".



Retrieving Mulan tba.maf alignment file... ok
Retrieving sequence and annotation files... ok

Retrieving sequence names... ok

Defining transcription factor binding sites

● TRANSFAC professional V10.2 library
Biological species
☑ vertebrates ☐ insects
☐ plants ☐ fungi
☐ nematodes ☐ bacteria
Matrix similarity
● Optimized for function
○ Predefined as 0.85
Matrix selection
☐ use only high-specificity matrices

○ User-defined consensus sequences

SUBMIT

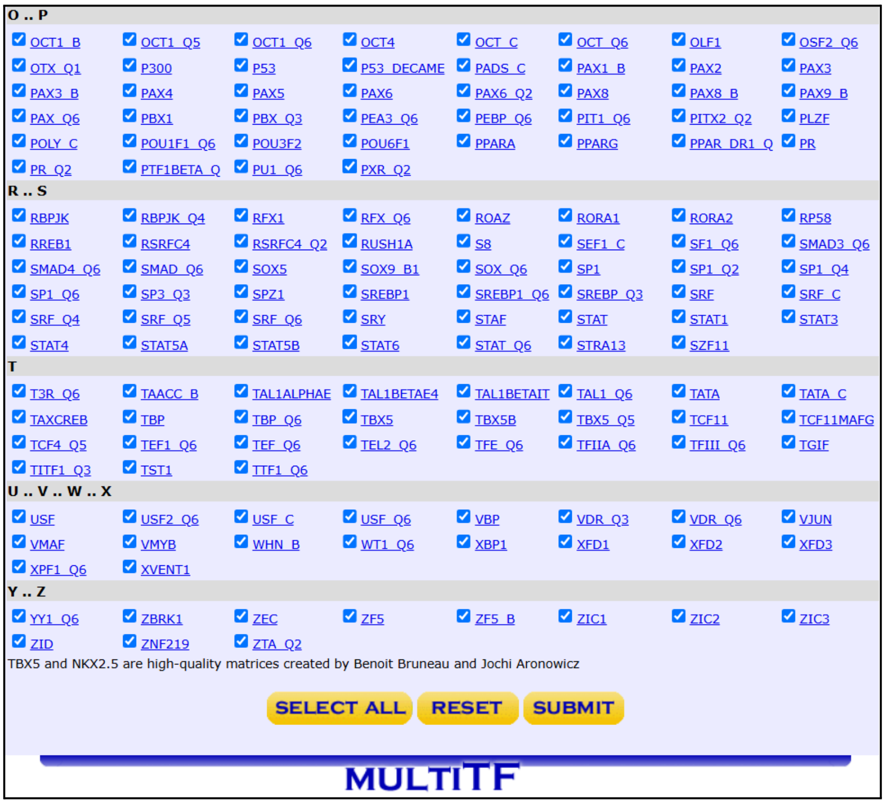
The "optimized for function" option utilizes different cut-off parameters for different TFBS. It corresponds to "not more than 3 TFBS per 10kb of a random sequence" optimization [PMID: 15590941].

"Predefined as" cut-offs measure sequence similarity to Transfac position weight matrices (PWM). Higher the cut-offs are, less sites will be identified.

The "use only high-specificity matrices" option subselects a list of TFBS matrices that have <= 0.85 "optimized for function" cut-off similarity to the corresponding Transfac PWM. These are the matrices that have many consistent definition sequences in the Transfac database.

Figure 6

8. On the “SELECT TRANSCRIPTION FACTORS” page, click “**SELECT ALL**” and then “**SUBMIT**” (Figure 7).

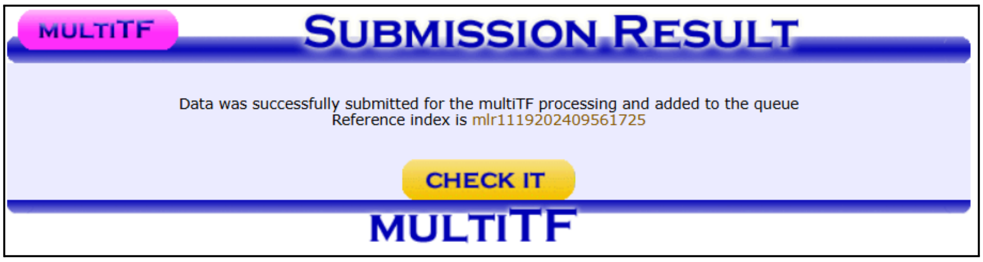


O .. P							
☑ OCT1_B	☑ OCT1_Q5	☑ OCT1_Q6	☑ OCT4	☑ OCT_C	☑ OCT_Q6	☑ OLF1	☑ OSF2_Q6
☑ OTX_Q1	☑ P300	☑ P53	☑ P53_DECAME	☑ PADS_C	☑ PAX1_B	☑ PAX2	☑ PAX3
☑ PAX3_B	☑ PAX4	☑ PAX5	☑ PAX6	☑ PAX6_Q2	☑ PAX8	☑ PAX8_B	☑ PAX9_B
☑ PAX_Q6	☑ PBX1	☑ PBX_Q3	☑ PEA3_Q6	☑ PEBP_Q6	☑ PIT1_Q6	☑ PITX2_Q2	☑ PLZF
☑ POLY_C	☑ POU1F1_Q6	☑ POU3F2	☑ POU6F1	☑ PPARA	☑ PPARG	☑ PPAR_DR1_Q	☑ PR
☑ PR_Q2	☑ PTF1BETA_Q	☑ PU1_Q6	☑ PXR_Q2				
R .. S							
☑ RBPJK	☑ RBPJK_Q4	☑ RFX1	☑ RFX_Q6	☑ ROAZ	☑ RORA1	☑ RORA2	☑ RP58
☑ RREB1	☑ RSRFC4	☑ RSRFC4_Q2	☑ RUSH1A	☑ S8	☑ SEF1_C	☑ SF1_Q6	☑ SMAD3_Q6
☑ SMAD4_Q6	☑ SMAD_Q6	☑ SOX5	☑ SOX9_B1	☑ SOX_Q6	☑ SP1	☑ SP1_Q2	☑ SP1_Q4
☑ SP1_Q6	☑ SP3_Q3	☑ SPZ1	☑ SREBP1	☑ SREBP1_Q6	☑ SREBP_Q3	☑ SRF	☑ SRF_C
☑ SRF_Q4	☑ SRF_Q5	☑ SRF_Q6	☑ SRY	☑ STAF	☑ STAT	☑ STAT1	☑ STAT3
☑ STAT4	☑ STAT5A	☑ STAT5B	☑ STAT6	☑ STAT_Q6	☑ STRA13	☑ SZF11	
T							
☑ T3R_Q6	☑ TAACC_B	☑ TAL1ALPHAE	☑ TAL1BETAE4	☑ TAL1BETAIT	☑ TAL1_Q6	☑ TATA	☑ TATA_C
☑ TAXCREB	☑ TBP	☑ TBP_Q6	☑ TBX5	☑ TBX5B	☑ TBX5_Q5	☑ TCF11	☑ TCF11MAFG
☑ TCF4_Q5	☑ TEF1_Q6	☑ TEF_Q6	☑ TEL2_Q6	☑ TFE_Q6	☑ TFIIA_Q6	☑ TFIII_Q6	☑ TGIF
☑ TITF1_Q3	☑ TST1	☑ TTF1_Q6					
U .. V .. W .. X							
☑ USF	☑ USF2_Q6	☑ USF_C	☑ USF_Q6	☑ VBP	☑ VDR_Q3	☑ VDR_Q6	☑ VJUN
☑ VMAF	☑ VMYB	☑ WHN_B	☑ WT1_Q6	☑ XBP1	☑ XFD1	☑ XFD2	☑ XFD3
☑ XPF1_Q6	☑ XVENT1						
Y .. Z							
☑ YY1_Q6	☑ ZBRK1	☑ ZEC	☑ ZF5	☑ ZF5_B	☑ ZIC1	☑ ZIC2	☑ ZIC3
☑ ZID	☑ ZNF219	☑ ZTA_Q2					

TBX5 and NKX2.5 are high-quality matrices created by Benoit Bruneau and Jochi Aronowicz

Figure 7

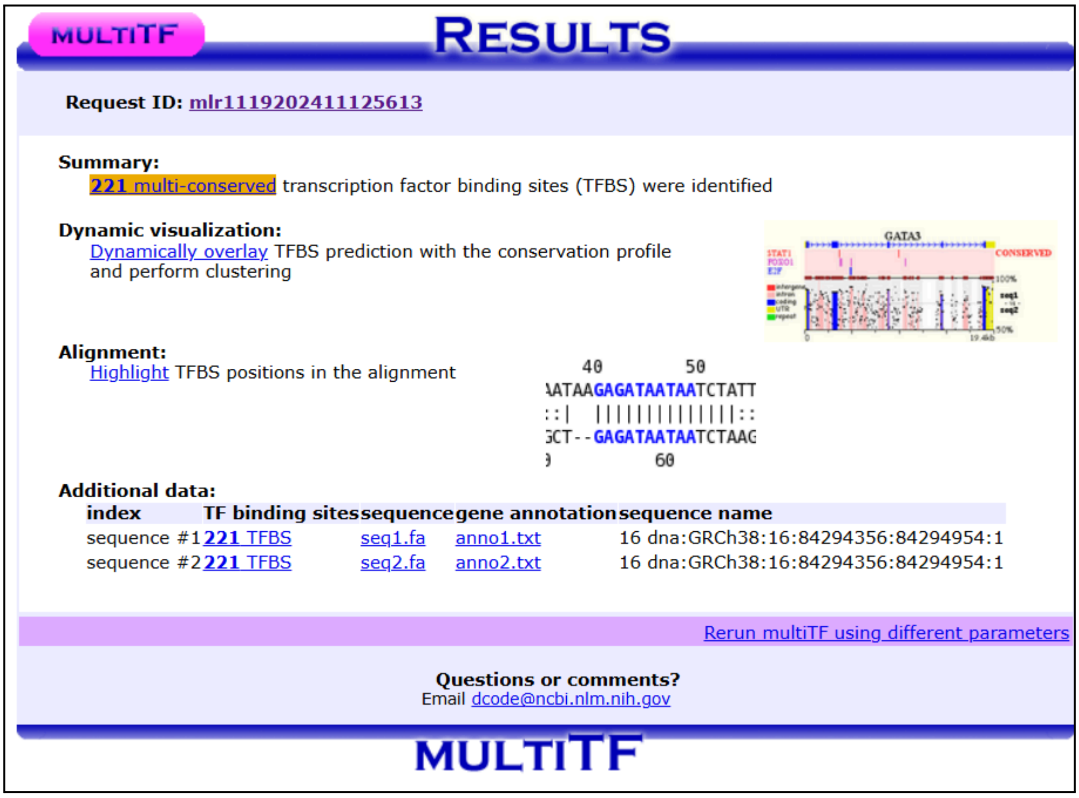
9. Press “**CHECK IT**” (Figure 8).



Data was successfully submitted for the multiTF processing and added to the queue
Reference index is mlr1119202409561725

Figure 8

10. Once the analysis is complete, the system generates the results. To view the list of predicted TFBS, click the link in the “**Summary**” section (highlighted in orange, Figure 9). The summary includes the transcription factor name, binding site sequence, DNA strand orientation, and genomic position (Figure 10).



Request ID: mlr1119202411125613

Summary:
221 multi-conserved transcription factor binding sites (TFBS) were identified

Dynamic visualization:
Dynamically overlay TFBS prediction with the conservation profile and perform clustering

Alignment:
Highlight TFBS positions in the alignment

```
  40        50
AATAAGAGATAATAATCTATT
::|  |||||||||||||::
GCT--GAGATAATAATCTAAG
9          60
```

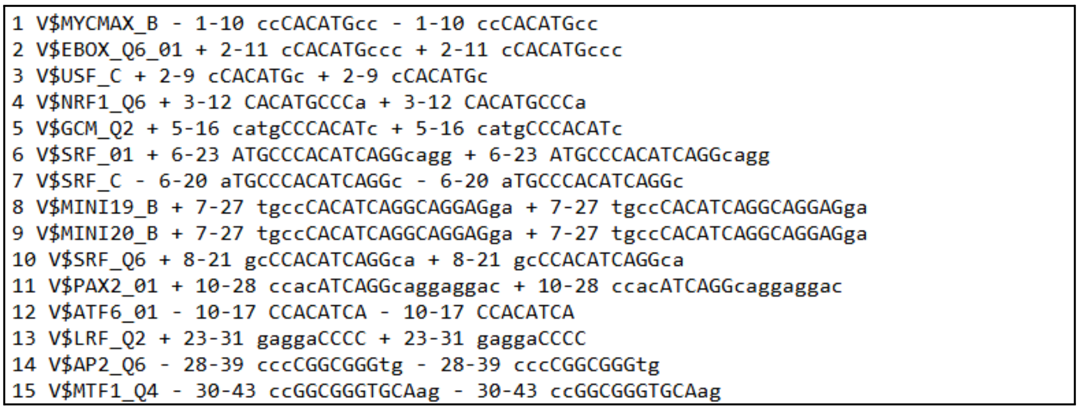
Additional data:

index	TF binding sites	sequence	gene annotation	sequence name
sequence #1	221 TFBS	seq1.fa	anno1.txt	16 dna:GRCh38:16:84294356:84294954:1
sequence #2	221 TFBS	seq2.fa	anno2.txt	16 dna:GRCh38:16:84294356:84294954:1

Rerun multiTF using different parameters

Questions or comments?
Email dcode@ncbi.nlm.nih.gov

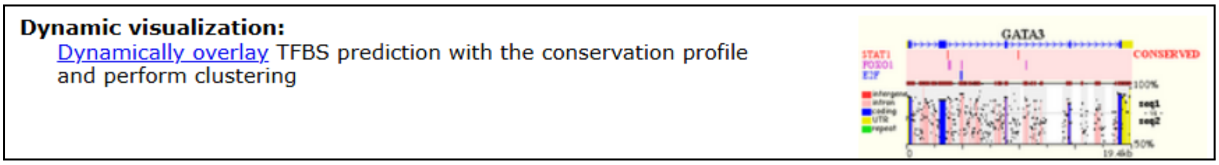
Figure 9



```
1 V$MYCMAX_B - 1-10 ccCACATGcc - 1-10 ccCACATGcc
2 V$EBOX_Q6_01 + 2-11 cCACATGccc + 2-11 cCACATGccc
3 V$USF_C + 2-9 cCACATGc + 2-9 cCACATGc
4 V$NRF1_Q6 + 3-12 CACATGCCCa + 3-12 CACATGCCCa
5 V$GCM_Q2 + 5-16 catgCCCACATc + 5-16 catgCCCACATc
6 V$SRF_01 + 6-23 ATGCCCACATCAGGcagg + 6-23 ATGCCCACATCAGGcagg
7 V$SRF_C - 6-20 aTGCCCACATCAGGc - 6-20 aTGCCCACATCAGGc
8 V$MINI19_B + 7-27 tgccCACATCAGGCAGGAGga + 7-27 tgccCACATCAGGCAGGAGga
9 V$MINI20_B + 7-27 tgccCACATCAGGCAGGAGga + 7-27 tgccCACATCAGGCAGGAGga
10 V$SRF_Q6 + 8-21 gcCCACATCAGGca + 8-21 gcCCACATCAGGca
11 V$PAX2_01 + 10-28 ccacATCAGGcaggaggac + 10-28 ccacATCAGGcaggaggac
12 V$ATF6_01 - 10-17 CCACATCA - 10-17 CCACATCA
13 V$LRF_Q2 + 23-31 gaggaCCCC + 23-31 gaggaCCCC
14 V$AP2_Q6 - 28-39 cccCGGCGGGtg - 28-39 cccCGGCGGGtg
15 V$MTF1_Q4 - 30-43 ccGGCGGGTGCAag - 30-43 ccGGCGGGTGCAag
```

Figure 10

11. Additional information, such as the conservation profile of TFBS across species (if multiple species were analyzed), can be accessed on the Results page by using the “**Dynamically overlay**” feature (Figure 11).



Dynamic visualization:
Dynamically overlay TFBS prediction with the conservation profile and perform clustering

Figure 11