# TFBS-Finder and TFBS-Mutator: Two scripts for mapping and mutating transcription factor binding sites to study gene regulation

Michael Bennet[1], Miranda Bigler[2], Naureen Aslam Khattak[2] Esdras Simervil[1], Landon Carre[1], Jeff Kinne[1,3], Shaad M. Ahmad[2,3]

[1]Department of Computer Science , [2]Department of Biology, [3]The Center for Genomic Advocacy, Indiana State University, Terre Haute, IN 47809

## Introduction

Enhancers are stretches of DNA that are recognized and bound by particular combinations of sequence-specific DNA-binding transcription factors (TFs) to regulate cell-specific or tissue-specific expression of enhancer-associated genes [1] as shown in Figure 1..

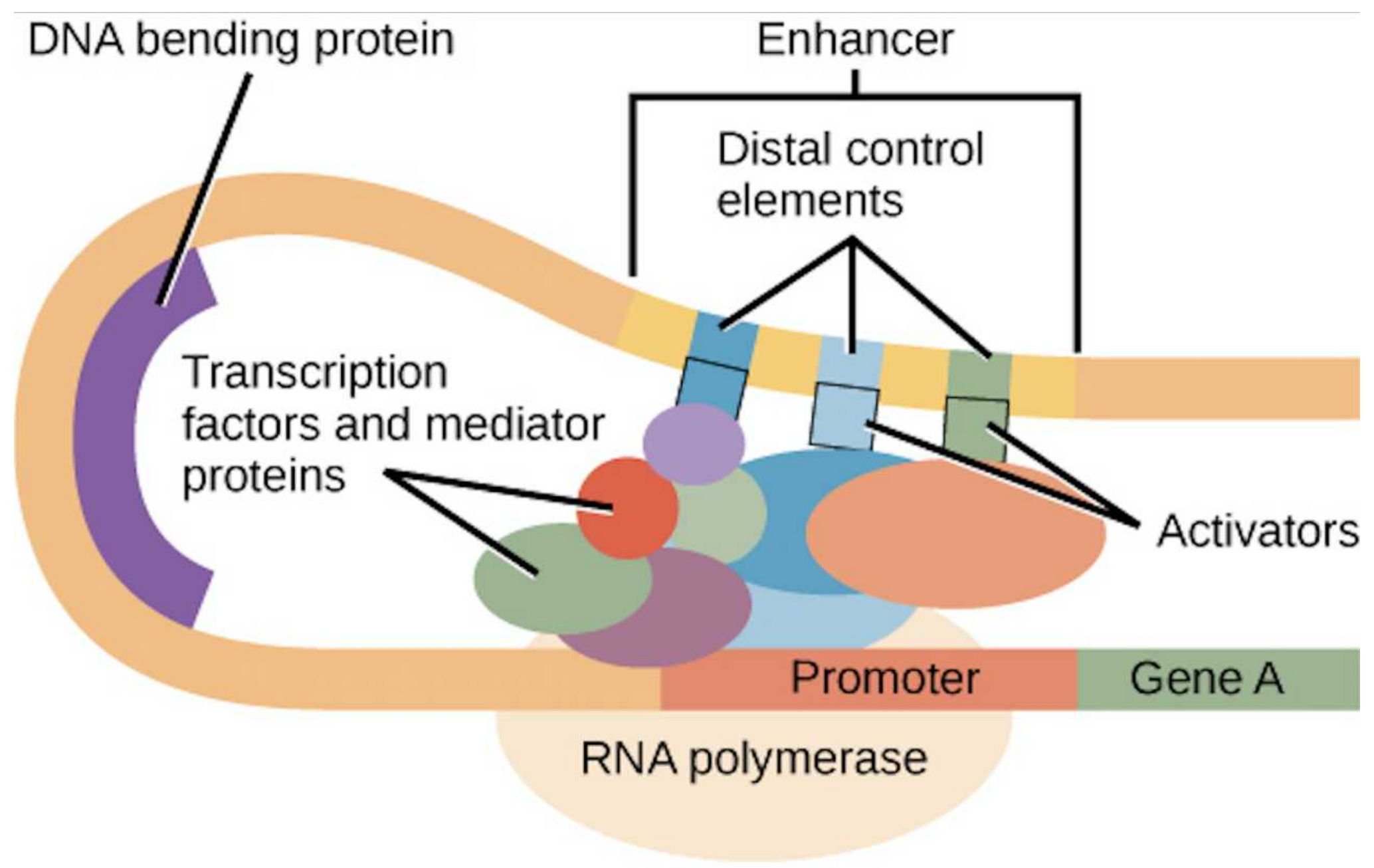


Figure 1 : The schematic representation of Enhancer and Transcription factor binding proteins

## Aims and Objective

- Design a script, TFBS-Finder, that maps potential TF binding sites in any given enhancer.
- Develop a second script, TFBS-Mutator, to design mutations to a binding site that could prevent the binding of one TF without impacting the binding of other TFs at the site.

## Methods

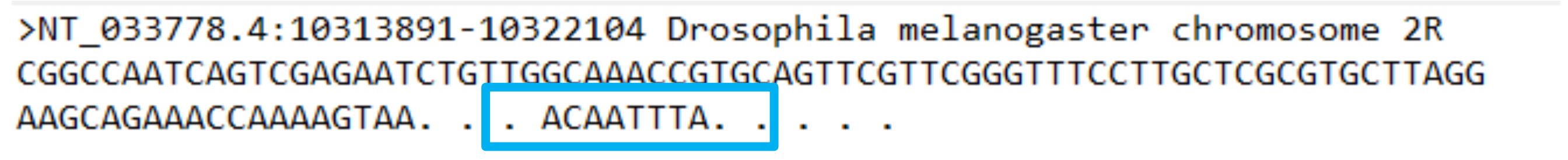
Protein Binding Microarray (PBM) data [2] is used as the input to the programming scripts, which are written in R. A PBM data file for a particular TF contains a row for each possible string of 8 nucleotides, together with their reverse complements; each row records the binding affinities to of the TF to the 8-mer in terms of E-score, Median and Z-score as shown in Table 1.

| 8-mer | 8-mer | E-score | Median | Z-score |
|---|---|---|---|---|
| AAAAAAAA | TTTTTTTT | 0.32779 | 4276.12 | 2.0578 |
| AAAAAAAC | GTTTTTTT | 0.27145 | 3864.99 | 1.8001 |
| AAAAAAAG | CTTTTTTT | 0.274 | 3201.76 | 1.3202 |
| AAAAAAAT | ATTTTTTT | 0.20289 | 2688.5 | 0.8747 |
| AAAAAACA | TGTTTTTT | 0.35522 | 5967.84 | 2.9076 |
| AAAAAACC | GGTTTTTT | 0.02376 | 2390.99 | 0.5758 |

Table 1: E-scores range from -0.5 to +0.5. Higher E-scores represent TFs binding with higher affinity to an 8mer. We can identify possible binding sites in a longer genetic sequence by locating these high binding affinity 8mers in a sequence.

## TFBS-Finder

TFBS-Finder takes input files for an enhancer sequence in FASTA or plain text format [3] together with PBM data flies for TFs of interest. For each TF PBM file, the script finds all 8-mers with E-score binding affinity above a user-defined cutoff. The script results can be output as a file in BED or GFF format. The file can further be visualized via R Studio or a genome browser to locate the TF binding sites in the respective genome.



```
>NT_033778.4:10313891-10322104 Drosophila melanogaster chromosome 2R
CGGCCAATCAGTCGAGAATCTGTTGGCAAACCGTGCAGTTCGTTCGGGTTTCCTTGCTCGCGTGCTTAGG
AAGCAGAAACCAAAAGTAA. . . ACAATTTA. . . . .
```

| 8-mer | 8-mer | E-score | Median | Z-score |
|---|---|---|---|---|
| ACAATTGC | GCAATTGT | 0.1236 | 2509.3 | 0.6989 |
| ACAATTGG | CCAATTGT | 0.1365 | 2708.12 | 0.8933 |
| ACAATTGT | ACAATTGT | 0.29054 | 3706.78 | 1.6936 |
| ACAATTTA | TAAATTGT | 0.30443 | 3570.92 | 1.5984 |
| ACAATTTC | GAAATTGT | 0.18556 | 1760.42 | -0.2048 |
| ACAATTTG | CAAATTGT | 0.22964 | 3397.86 | 1.4717 |

Figure 2: Top: Excerpt from the *Ndg* release 5 sequence FASTA containing a particular 8mer. Bottom: Excerpt from CHES-1-like PBM file highlighting the same 8mer as one with high binding affinity [2].

## Testing with *Nidogen* (*Ndg*) Enhancer

TFBS-Finder was tested using the *Ndg* gene enhancer in *D. melanogaster*. We evaluated the accuracy of the script by utilizing the already known binding Fkh1 sites for a protein family known as Forkhead TFs. Proteins Biniou, Jumeau, and CHES-1-like are known to bind at Fkh1, therefore, we used the PBM files for those TFs [3] as input to the scripts.

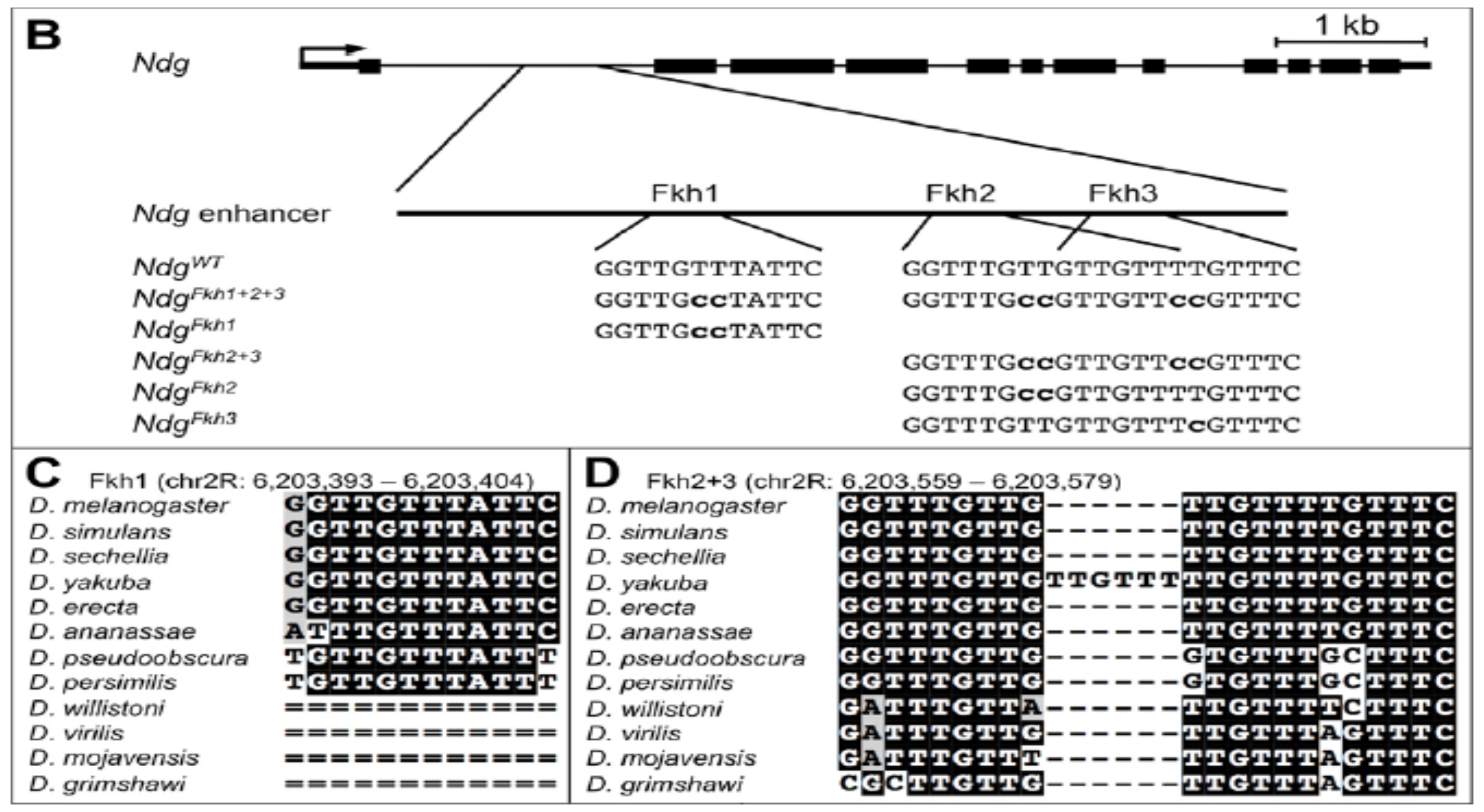

B
Ndg
1 kb
Ndg enhancer
Fkh1
Fkh2
Fkh3
Ndg^WT GGTTGTTTATTC GGTTTGTTGTTGTTTTGTTTC
Ndg^Fkh1+2+3 GGTTGccTATTC GGTTTGccGTTGTTccGTTTC
Ndg^Fkh1 GGTTGccTATTC
Ndg^Fkh2+3 GGTTTGccGTTGTTccGTTTC
Ndg^Fkh2 GGTTTGccGTTGTTTTGTTTC
Ndg^Fkh3 GGTTTGTTGTTGTTTcGTTTC
C
Fkh1 (chr2R: 6,203,393 – 6,203,404)
D
Fkh2+3 (chr2R: 6,203,559 – 6,203,579)
D. melanogaster
D. simulans
D. sechellia
D. yakuba
D. erecta
D. ananassae
D. pseudoobscura
D. persimilis
D. willistoni
D. virilis
D. mojavensis
D. grimshawi

Figure 3 : (B) Positions of Fkh (1,2,3) binding sites on the enhancer *Ndg* (C) comparison of Fkh1 for various species, (D) comparison of Fkh2 and Fkh3 for different species. *Zhu et al., 2012*

## Reproducibility

| nmer | comp | escore | median | zscore | index | pbm |
|---|---|---|---|---|---|---|
| GGTTGTTT | AAACAACC | 0.42496 | 73246.16 | 4.4391 | 6203394 | Bin |
| GTTGTTTA | TAAACAAC | 0.49246 | 260707.9 | 12.5957 | 6203395 | Bin |
| TTGTTTAT | ATAAACAA | 0.49789 | 372886.4 | 14.895 | 6203396 | Bin |
| TGTTTATT | AATAAACA | 0.49501 | 310009.4 | 13.7085 | 6203397 | Bin |
| GTTTATTC | GAATAAAC | 0.45683 | 90724.5 | 5.814 | 6203398 | Bin |
| TTTATTCA | TGAATAAA | 0.40734 | 70464.31 | 4.1903 | 6203399 | Bin |
| GTTGTTTA | TAAACAAC | 0.47691 | 9816.27 | 4.1764 | 6203395 | CHES-1-like |
| TTGTTTAT | ATAAACAA | 0.49306 | 18963.47 | 5.8551 | 6203396 | CHES-1-like |
| TGTTTATT | AATAAACA | 0.46619 | 9134.61 | 3.9929 | 6203397 | CHES-1-like |
| TTGTTTAT | ATAAACAA | 0.35581 | 38249.05 | 2.4132 | 6203396 | Jumeau |

Table 2 : TFBS-Finder detects Biniou, CHES-1-like, and Jumeau binding at the Fkh1 binding sites.

## TFBS-Mutator

The TFBS-Mutator script is designed to take PBM and sequence data and find optimal mutations for the sequence of interest. The ideal mutations should prevent a specific TFs from binding to a site while minimally disturbing the binding activity of other TFs in the mutated region.

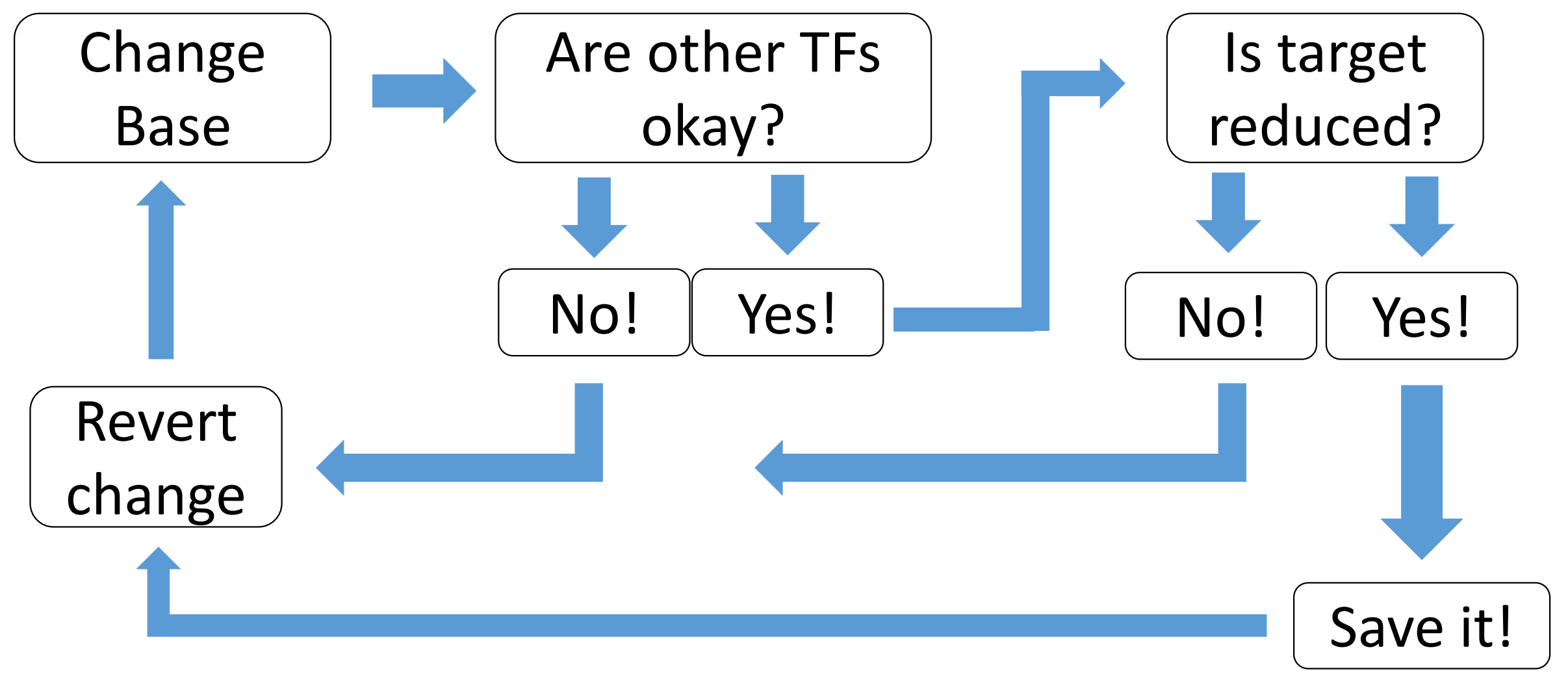


Figure 4: Work flow of TFBS-Mutator as it searches for ideal mutations to block one TF from binding.

## Conclusion

TFBS-Finder detects a known binding site on the *Ndg* enhancer in *D. melanogaster*, indicating that our script is functional. However, the coordinates on the detected site were off by one base pair, indicating that the script needs improvements. TFBS-Mutator currently produces a list of mutations and filters out undesirable ones, but is not yet capable of choosing the best mutations to use on its own.

## Acknowledgements

This work was supported by NIH grant 5R25MD011712-03 (J.K.), AHA grant 16SDG31390005 (S.M.A.), and the Indiana State University Research Committee.

## References

1. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. *Science.* 281 (5373): 60–3
2. UniPROBE database
3. NCBI GenBank
4. Zhu et al., Differential regulation of mesodermal gene expression by Drosophila cell type-specific Forkhead transcription factors. *Development.* 139, 1457-1466