Predicting genetic variants impact on transcription factor binding sites using *k*-mer-based motif models

Manuel Mangoni[1], Dorotea Savariso[1*], Manuel Tognon[1], Rosalba Giugno[1#], Luca Pinello[2#]*

(1) Department of Computer Science, University of Verona, Verona, Italy; (2) Molecular Pathology Unit, Center for Cancer Research, Massachusetts General Hospital, Department of Pathology, Harvard Medical School, Boston, Massachusetts 02129, USA; (*) Equal Contribution; (#) Corresponding authors

k-mer Models for SNPs effect on TFBS

- **Transcription Factors** (TFs) are key regulatory proteins in the regulation of transcription in the cell
- **Motif-Raptor** (MR) is an open-source command line tool to investigate how genetic variants impact TFs function using TF binding motif databases, cell type-specific chromatin accessibility and gene expression
- In this work, we extend MR in Motif-Raptor2, to improve MR predictions of SNP effects on TFBS using ***k*-mer-based TFBS models**, instead of PWMs

Transcription Factor Overview: PWM to represent Binding Sites

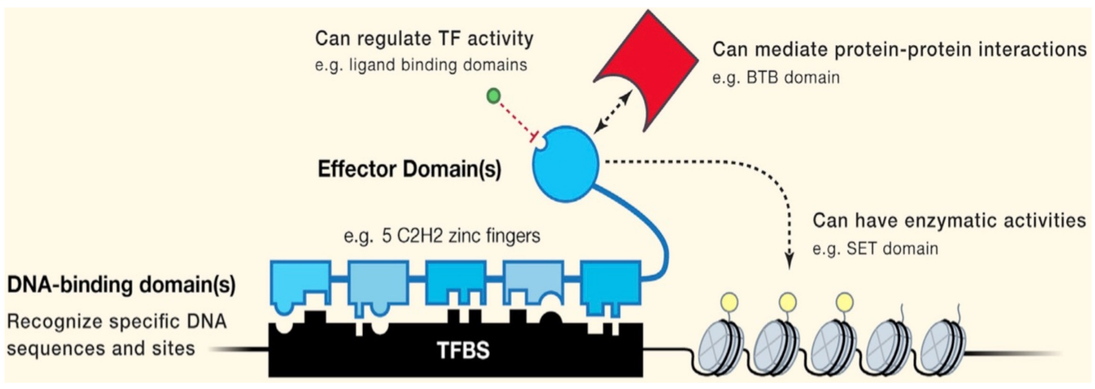

Figure 1. TFs are proteins capable of binding DNA in a sequence-specific manner and regulating transcription through many different mechanisms of action. TF DNA-binding specificities are known as "motifs"

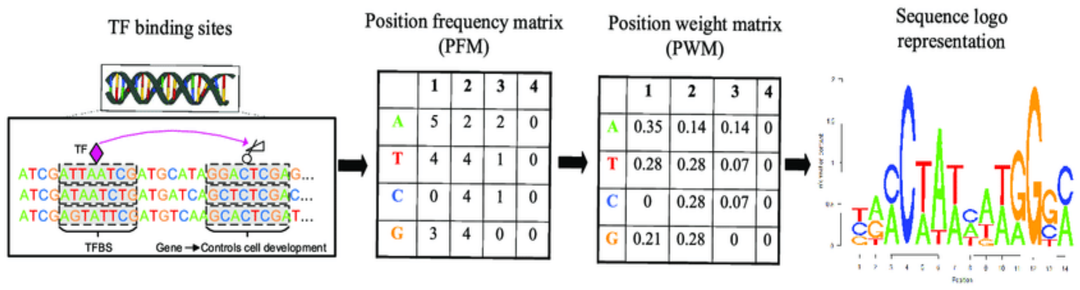

Figure 2. TF motifs are typically displayed as a sequence logo, which in turn represents a "position weight matrix" (PWM). A matrix-based PWM model accounts for the preference for each of the four nucleotides at each position in the motif.

Diseases and Traits associated SNPs and TFBS

- Several studies have reported that **genetic variants can enhance or disrupt TF-DNA binding affinity**
- For common diseases and complex traits, the great majority (>90%) of associated **SNPs correspond to functionally relevant non-coding regions** such as enhancers and promoters
- Genome-wide association studies (GWAS) have implicated hundreds of thousands of single-nucleotide polymorphisms (SNPs) in human diseases and traits

Motif-Raptor: Investigating disease associated SNPs effect on TFBS

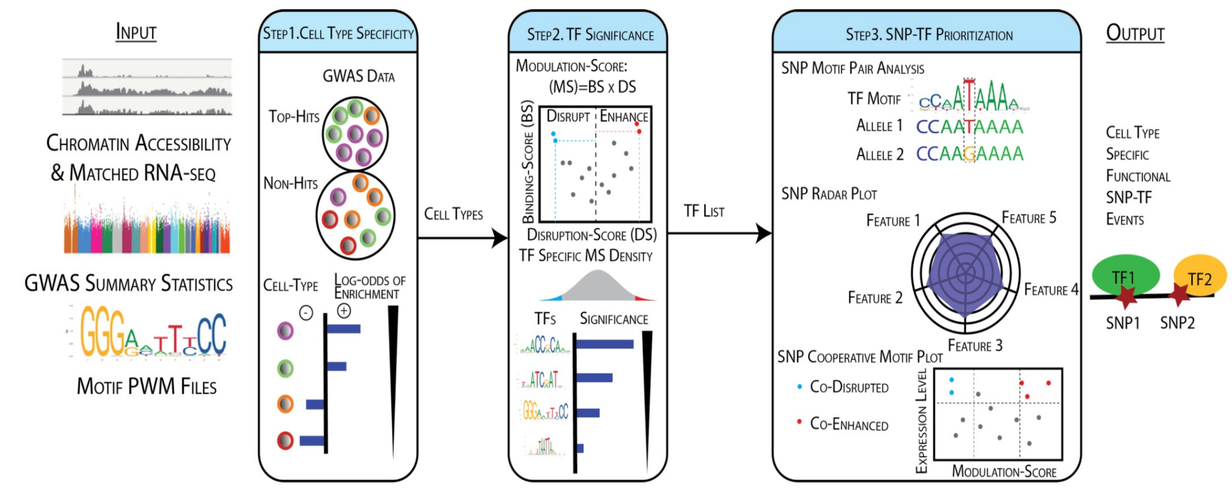

Figure 3. Summary of Motif-Raptor analysis workflow. Three steps are performed: (1) characterize relevant cell types based on the enrichment of phenotype associated SNPs in chromatin accessible sites, (2) find TFs with binding sites that are significantly modulated by genetic variants in these cell types and (3) identify and visualize individual TF-SNP regulation events

Using gkm-SVM Models to represent TFBS

- *K*-mers are features that are either present or not, so **it doesn't require large amount of data**
- *K*-mers can capture the ***k*-th order dependencies** between the bases

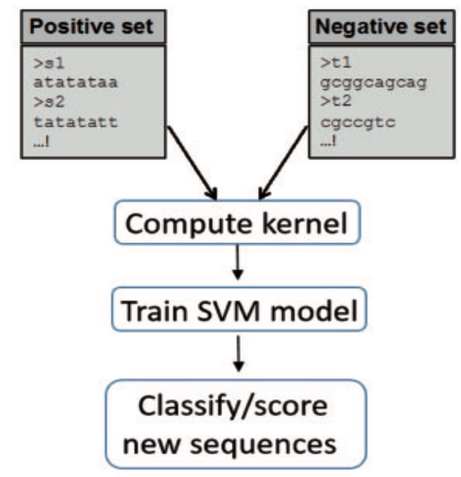

Figure 4. General pipeline to generate gkm-SVM Models and use them to score new sequences

k-mer models show better discriminative power and were integrated in Motif Raptor

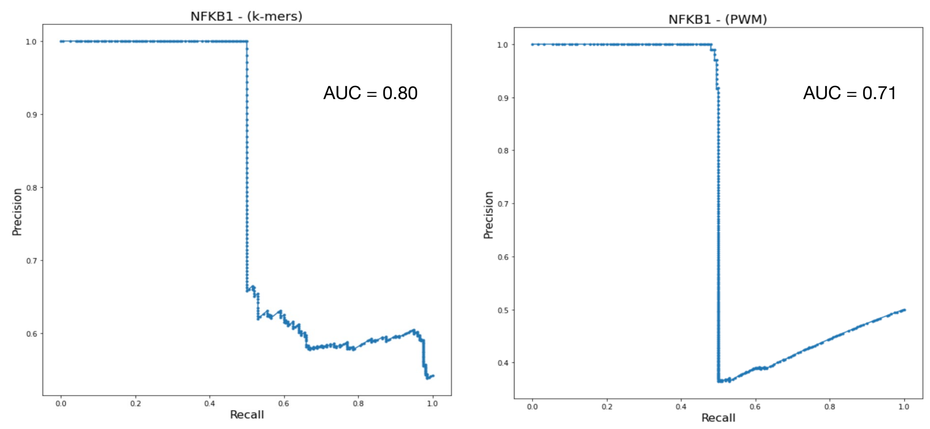

Figure 5. Precision-Recall plot for NFKB1 models (*k*-mers and PWM) and AUC results

- We observed that ***k*-mer-based motif models show a higher discriminative power than PWM** when identifying bound (ChIP-seq peaks) and unbound sequences (random genomic sequences)
- We integrated gkm-SVM models into the Motif-Raptor pipeline in order to have more reliable predictions for genetic variants impact on TFBS

Conclusions and future perspectives

- gkm-SVM models show higher discriminative power with respect to PWMs
- We plan on developing an **extension of Motif-Raptor** considering *k*-mer based models to improve the prediction of genetic variants effects on TFBS
- In the future, we aim to do complete tests on **MR2** and check if the prediction of genetic variants effects on TFBS is improved