



CORRESPONDENCE

REVISED Positional weight matrices have sufficient prediction power for analysis of noncoding variants

[version 3; peer review: 2 approved]

Alexandr Boytsov^{1,2}, Sergey Abramov^{id}^{1,2}, Vsevolod J. Makeev^{1,2},
Ivan V. Kulakovskiy^{id}^{1,3}

¹Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, 119991, Russian Federation²Moscow Institute of Physics and Technology, Dolgoprudny, 141700, Russian Federation³Institute of Protein Research, Russian Academy of Sciences, Pushchino, 142290, Russian Federation

v3 First published: 12 Jan 2022, 11:33
<https://doi.org/10.12688/f1000research.75471.1>
 Second version: 23 Jun 2022, 11:33
<https://doi.org/10.12688/f1000research.75471.2>
 Latest published: 04 Jul 2022, 11:33
<https://doi.org/10.12688/f1000research.75471.3>

Abstract

The position weight matrix, also called the position-specific scoring matrix, is the commonly accepted model to quantify the specificity of transcription factor binding to DNA. Position weight matrices are used in thousands of projects and software tools in regulatory genomics, including computational prediction of the regulatory impact of single-nucleotide variants. Yet, recently Yan et al. reported that "the position weight matrices of most transcription factors lack sufficient predictive power" if applied to the analysis of regulatory variants studied with a newly developed experimental method, SNP-SELEX. Here, we re-analyze the rich experimental dataset obtained by Yan et al. and show that appropriately selected position weight matrices in fact can adequately quantify transcription factor binding to alternative alleles.

Keywords

Transcriptional regulation, rSNP, TF-DNA binding, SNP-SELEX, PWM, PSSM



This article is included in the **Bioinformatics** gateway.

Open Peer Review**Approval Status** ✓ ✓

	1	2
version 3		
(revision)		
04 Jul 2022		
version 2		
(revision)	✓	✓
23 Jun 2022	view	view
	↑	↑
version 1	?	?
12 Jan 2022	view	view

1. **Victor G. Levitsky**, Institute of Cytology and Genetics, Novosibirsk, Russian Federation

2. **Philip Machanick**^{id}, Rhodes University, Makhanda, South Africa

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Alexandr Boytsov (boytsovs.av@phystech.edu), Sergey Abramov (abramov.sa@phystech.edu)

Author roles: **Boytsov A:** Formal Analysis, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Abramov S:** Formal Analysis, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Makeev VJ:** Conceptualization, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Kulakovskiy IV:** Conceptualization, Formal Analysis, Funding Acquisition, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study was supported by the Russian Science Foundation (RSF) grant 20-74-10075 to IVK.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Boytsov A *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Boytsov A, Abramov S, Makeev VJ and Kulakovskiy IV. **Positional weight matrices have sufficient prediction power for analysis of noncoding variants [version 3; peer review: 2 approved]** F1000Research 2022, 11:33 <https://doi.org/10.12688/f1000research.75471.3>

First published: 12 Jan 2022, 11:33 <https://doi.org/10.12688/f1000research.75471.1>

REVISED Amendments from Version 2

The reviewer has drawn our attention to the missing citation of one of the key PWM benchmarking papers. We also updated the link to supplementary materials.

Any further responses from the reviewers can be found at the end of the article

Introduction

Gene regulatory regions constitute an important part of non-coding DNA which defines both the global development program of a mammal and individual traits of a particular organism. Specific recognition of DNA sites by transcription factors (TFs) provides the gear system linking individual genomic variants to phenotypes.¹ The commonly accepted model to quantify the specificity of transcription factor binding to various DNA sites is the position weight matrix (PWM), which specifies additive contributions of individual nucleotides to the protein-DNA binding energy.^{2,3} Recently Yan *et al.*⁴ presented a powerful high-throughput experimental technique, SNP-SELEX, which allows measuring differential TF binding to alternative alleles *in vitro*. Yan *et al.* used the experimental data they had obtained for many TFs to assess the performance of PWM in predicting differential TF binding to alternative alleles and compare it to that of deltaSVM, a more complex method based on machine learning. As a result, they reported that in this setting “the position weight matrices of most transcription factors lack sufficient predictive power”. Keeping in mind that PWMs are extensively used for prediction of the regulatory potential of single-nucleotide variants^{5–8} the finding of Yan *et al.* could be devastating for a vast array of research projects and software tools.

Yan *et al.* tend to explain the poor performance of PWMs by model limitations, primarily, arising from the oversimplistic assumption that nucleotides occupying different positions in the binding site provide independent contributions to the binding energy. Here we re-analyze the dataset of Yan *et al.* and argue that the poor PWM performance in predicting differential transcription factor binding to alternative alleles detected by SNP-SELEX is to a major extent explained not by the principle limitations of PWM as a mathematical construction but rather by particular inadequate PWMs for TFs under study. We show that the careful selection of PWMs of many TFs from a public database quantitatively explains the differential TF binding to allelic variants with reliability comparable to deltaSVM.

Results

To re-assess PWM performance, we used PWMs stored in the CIS-BP database,^{9,10} which contains PWMs constructed from data obtained with different experimental techniques for thousands of TFs for different species. With the objective of selecting the PWM appropriate for quantifying differential allele binding of a TF, for each of 129 TFs assessed in Yan *et al.* we extracted an extended set of candidate PWMs, with a median of 32 PWMs per TF. The overall distribution was non-uniform e.g. there were only 2 candidate PWMs for ZNF396 and over a thousand for FOXA2, see *Extended data*, Supplementary Table S1.

Through cross-validation on the 1st batch of SNP-SELEX data following the strategy of Yan *et al.*, we selected the best PWM^{CIS-BP} for each TF (see “Selecting the best PWMs and estimating PWM performance with SNP-SELEX data” in the Methods). There was no correlation between the prediction performance (area under precision-recall curve, AUPRC) and the number of tested PWMs per TF ($r = -0.07$, $P = 0.425$). Many of the best-performing PWMs were originally constructed from the data related not to the target TF but to other TFs sharing the same DNA-binding domain as the TF of interest. Some PWMs were based upon the TF binding data from different species. We denote by $\Delta\text{PWM}^{\text{CIS-BP}}$ the difference of the allelic scores predicted with PWM^{CIS-BP} and by $\Delta\text{PWM}^{\text{Mult}}$ the results of PWMs obtained from HT-SELEX data using the multinomial algorithm¹¹ and assessed by Yan *et al.*⁴

Yan *et al.* provide the experimental differential binding scores for each SNP (the “pbSNP” scores). We compared these scores with $\Delta\text{PWM}^{\text{CIS-BP}}$ predictions. For a complete set of 816594 TF-SNP pairs (Figure 1a), the Pearson correlation coefficient was comparable to that observed by Yan *et al.* (0.531 vs 0.534 in Yan *et al.*, see their Figure 2a). Yet, if only SNPs with a strong predicted binding are included ($P < 10^{-4}$ for the PWM score of a stronger bound allele) then a much higher correlation ($r \sim 0.828$) is achieved, see Figure 1b. Finally, if only the most strongly bound TF is considered for each SNP (as in Figure 3a of Yan *et al.*), the respective correlation reaches 0.711, comparable to -0.777 reported for deltaSVM in Yan *et al.* (compare to their Figure 1c).

To assess $\Delta\text{PWM}^{\text{CIS-BP}}$ performance at varying binding affinity ranges we, similarly to Yan *et al.*, categorized the SNPs into five quantiles based on their observed affinities (“OBS” scores) and assessed the performance of $\Delta\text{PWM}^{\text{CIS-BP}}$ separately for each quantile. For all quantiles but the lowest (the weakest bound sites) $\Delta\text{PWM}^{\text{CIS-BP}}$ outperformed $\Delta\text{PWM}^{\text{Mult}}$ of Yan *et al.* Notably, the performance of $\Delta\text{PWM}^{\text{CIS-BP}}$ was especially high for the middle quantiles and for

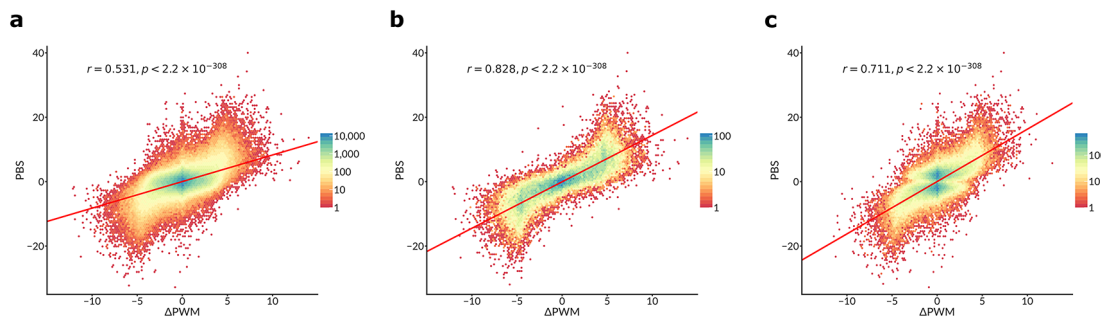


Figure 1. PWM predictions of differential TF binding to SNPs correlate with SNP-SELEX estimates. Hex-plots of PBS scores (Y-axis) vs $\Delta\text{PWM}^{\text{CIS-BP}}$ predictions (X-axis) for different sets of SNPs, analogous to Figure 2a and Figure 3a in Yan *et al.*, with PCC and two-sided t-test *P*-values displayed. a. Complete set of 816594 TF-SNP pairs for 129 TFs that were used to compare the performance of deltaSVM vs ΔPWM in Yan *et al.* b. A subset of 35837 TF-SNP pairs overlapping strong PWM hits ($P < 10^{-4}$ for a stronger bound allele). c. A subset of 84962 TF-SNP pairs obtained by considering only the TF with the highest PBS score for each SNP (as in Figure 3a of Yan *et al.*).

the highest quantile was on par with deltaSVM (see Supplementary Figure S1, compare with Extended data Figure 7 in Yan *et al.*). Particularly, for strongly bound SNPs from high quantiles in the first SNP-SELEX batch, $\Delta\text{PWM}^{\text{CIS-BP}}$ did not display any TFs with a very small AUPRC (i.e. prediction failures), the other metric for which deltaSVM dramatically outperformed $\Delta\text{PWM}^{\text{Mult}}$.

Next, we compared the overall performance of $\Delta\text{PWM}^{\text{CIS-BP}}$ for different TFs at the 1st SNP-SELEX batch. For more than a half (72 of 129) of transcription factors $\Delta\text{PWM}^{\text{CIS-BP}}$ achieved reliable predictions fulfilling the same criterion as in Yan *et al.* of the AUPRC > 0.75 , see Figure 2a. This is 3 times more transcription factors with reliable PWM predictions than reported in Yan *et al.* for $\Delta\text{PWM}^{\text{Mult}}$ (only 24 out of 129). Notably, we obtained good predictions in some cases reported as markedly underperforming such as FOXA2 (compare Figure 2b with Figure 2b of Yan *et al.*). Another TF performing markedly poorly for PWM^{Mult} was IRF3, but the best $\text{PWM}^{\text{CIS-BP}}$ performed better than both $\Delta\text{PWM}^{\text{Mult}}$ and deltaSVM (AUPRC for $\Delta\text{PWM}^{\text{CIS-BP}}$ of 0.298 as compared to 0.184 of deltaSVM). In some cases, the predictive power of $\Delta\text{PWM}^{\text{CIS-BP}}$ went in line with that of $\Delta\text{PWM}^{\text{Mult}}$, for instance, TFAP transcription factors in both cases displayed outstanding performance (AUPRC of 0.9 for $\Delta\text{PWM}^{\text{Mult}}$ and 0.92 for $\Delta\text{PWM}^{\text{CIS-BP}}$) whereas E2F family transcription factors in both cases performed worse (AUPRC of 0.4 for $\Delta\text{PWM}^{\text{Mult}}$ and 0.42 for $\Delta\text{PWM}^{\text{CIS-BP}}$).

In fact, for 34 transcription factors, $\text{PWM}^{\text{CIS-BP}}$ outperformed advanced models of deltaSVM (Figure 2c). 5-fold cross-validation showed that models reaching higher AUPRC simultaneously had a lower variance in prediction quality across individual folds (Figure 2d). Furthermore, we tested the PWMs on the independent 2nd batch data (Figure 2e, compare with Figure 3d of Yan *et al.*), and it also showed competitive albeit lower performance, with 36 of 124 transcription factors passing 0.75 AUPRC. Finally, we tested if the PWM predictions agreed with the allelic binding ratios in HepG2 ChIP-seq data and found a small but marginally significant correlation (Figure 2f, $r = 0.194$, $P = 0.052$) for 101 SNPs tested in Yan *et al.* and reaching $r = 0.235$ ($P = 0.047$) for a subset of 72 SNPs with significant $\text{PWM}^{\text{CIS-BP}}$ hits (motif *P*-value < 0.005), in contrast to almost zero correlation for $\Delta\text{PWM}^{\text{Mult}}$ reported in Yan *et al.*

Discussion

Our approach mimics a machine learning setup, where the best model is selected ("trained") through cross-validation on a first experimental data set (1st batch of the SNP-SELEX data), and then additionally independently validated on the second experimental data set (2nd batch of SNP-SELEX data). As we select from a finite and typically small set of candidate PWMs, the risks of overfitting are minimized, and the resulting performance was not correlated with the number of 'candidate' PWMs. The utilized layout allowed us to pick up the best suited PWM independently from the original data or motif discovery method used for PWM construction, yet maintaining the main PWM limitations, such as the assumption of the independent contributions of nucleotides at different TFBS positions.

The lower performance of PWMs for TFBS recognition as compared with more complex models was reported in many publications.^{12–14} The popular explanation blames the assumption of positional independence of PWM scores, which comes short of taking into account the marked correlations of nucleotides located at neighboring or even distant positions of binding sites.^{15–17} This shortcoming is also considered over-restrictive for PWM applications in predicting the effects of single nucleotide variants on TF binding,^{4,18} which has recently come into the spotlight of modern genetics where the advent of complete genome sequencing brought about the need for interpretation of phenotypes associated with

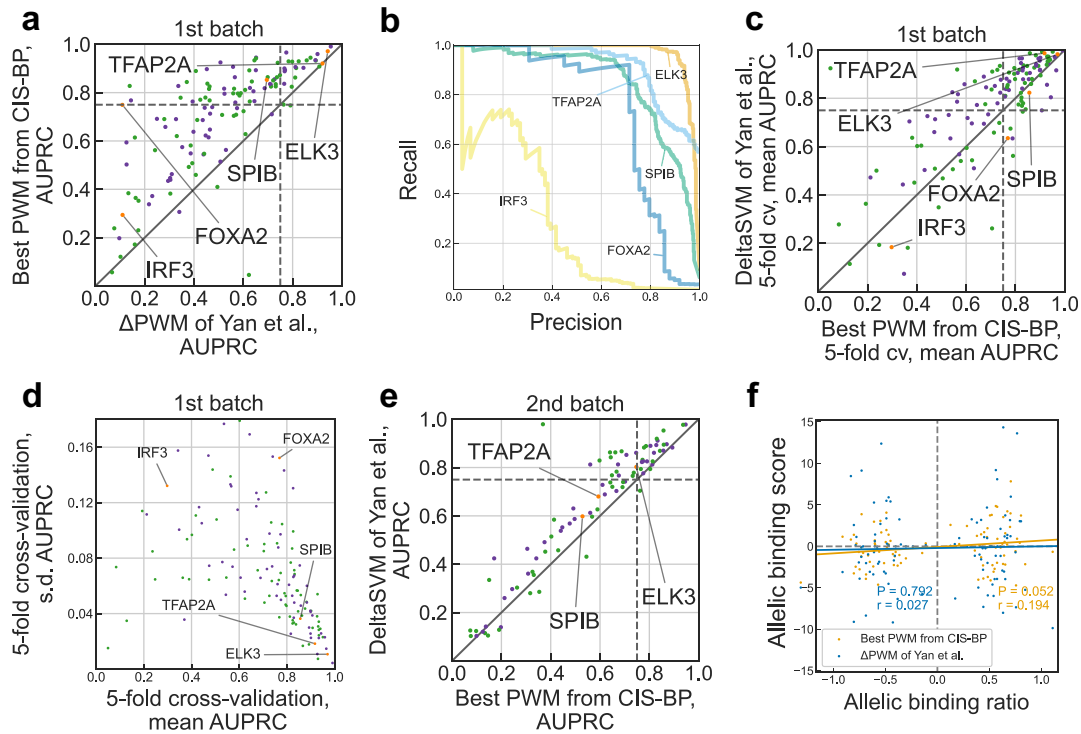


Figure 2. Re-evaluation of position weight matrices with the SNP-SELEX data. a. Comparison of performance of Yan *et al.* Δ PWM (x-axis) and best CIS-BP position weight matrices (PWMs) in predicting preferential binding SNPs in the 1st batch on the SNP-SELEX data. Each point denotes one of 129 TFs, violet and green points denote inferred and direct PWMs, respectively (see the Methods). Both axes show area under the precision-recall curve (AUPRC) values. Transcription factors (TFs) shown in Figure 2b of Yan *et al.* are highlighted in orange and labeled. Dashed lines denote AUPRC of 0.75. b. Examples of the precision-recall curves showing performance of different PWM models in predicting preferential binding SNPs (single-nucleotide polymorphisms) as in Figure 2b of Yan *et al.* c. Comparison of performance of deltaSVM (y-axis) and best CIS-BP PWMs (x-axis) in predicting preferential binding SNPs identified in the 1st batch of SNP-SELEX. Each point denotes one of 129 TFs, violet and green points denote inferred and direct PWMs, respectively. Both axes show mean AUPRC values obtained by 5-fold cross-validation (cv). Dashed lines denote AUPRC of 0.75. d. Variance of performance of CIS-BP PWMs (x-axis: mean AUPRC, y-axis: s.d.) in 5-fold cross-validation using the complete data of the 1st batch of SNP-SELEX. Each point denotes one of 129 TFs, violet and green points denote inferred and direct PWMs, respectively. e. Comparison of performance of deltaSVM (y-axis) and best CIS-BP PWMs (x-axis) in predicting preferential binding SNPs identified in the 2nd batch of SNP-SELEX. Each point denotes one of 87 TFs, violet and green points denote inferred and direct PWMs, respectively. Both axes show AUPRC values. Dashed lines denote AUPRC of 0.75. f. Correlation of allelic biases of DNA binding detected from ChIP-Seq experiments in HepG2 cells by Yan *et al.* and those predicted by Δ PWM of Yan *et al.* (blue) and best CIS-BP PWMs (orange). Pearson correlation coefficient (r) and the respective P -value are shown. The allelic binding ratio is computed as in Yan *et al.*; 101 transcription factor-SNP pairs involving 68 unique SNPs and 6 transcription factors (ATF2, FOXA2, HLF, MAFG, YBX1, and FOXA1) are shown.

regulatory variants.^{19,20} Here we suggest an alternative explanation of inadequate PWM performance in predicting the effects of single-nucleotide variants on TF binding. In many cases, the reason is an inadequate PWM construction or selection procedure.

Besides technical difficulties in proper “training” a PWM through motif discovery from different types of experimental data, the particular experimental context may influence the applicability of the resulting PWM. Careful selection of PWMs from a pool of alternative existing models results in an apparent improvement of the quantitative assessment of preferential binding to single-nucleotide variants, especially for high-scoring TFBS. In many cases, the prediction power becomes comparable to that of the significantly more complex model such as deltaSVM. We specifically emphasize that in our study all selected PWM^{CIS-BP} were genuine PWMs following the classic assumption of the independent contribution of positional scores.

Summing up, our results do not compromise the high performance of deltaSVM,¹² used by Yan *et al.* as an advanced substitution of position weight matrices (PWMs). However, properly selected PWMs achieve performance that is very close and in some cases even better than that of deltaSVM. Despite the simplicity of the PWM model, its construction is

not trivial and its success depends both on the motif discovery algorithm and reliability of the training data. In our case, almost half of the best PWMs were derived from related TFs, including 8 cases of PWMs based on experimental data from other species. The experiments used to obtain the best PWMs were also of different types, including ChIP-Seq, protein-binding microarrays, and SMiLE-Seq data, see *Extended data*, Supplementary Table S1.²¹ Thus, it is important to consider various sources of PWMs and select those the most suitable by proper benchmarking. In the context of applying PWMs to analyze regulatory variants, SNP-SELEX of Yan *et al.* provides rich, unique, and practically useful data.

The objective of our study is by no means to undermine the necessity of complex TFBS models with dependent positional contributions. Advanced multiparametric and alignment-free approaches such as deltaSVM appear very likely to shape the oncoming future of transcription factor binding site models. Rather, we want to underline that the prediction performance of transcription factor binding sites in its current stage is more influenced by model training protocols than by model structure restrictions. PWMs still can deliver a solid standard in representation and bioinformatics analysis of the transcription factor binding sites, including assessment of the functional impact of single nucleotide variants in gene regulatory regions. In addition, we underline that better defined ‘baseline’ PWMs or PWM selection procedures are required for the proper evaluation of advanced models. It is important that such ‘baseline’ TFBS models, while certainly being handicapped by design, still reach meaningful prediction quality. These are good news for thousands of researchers who still use the ‘legacy’ PWM scoring for practical applications in regulatory genomics and bioinformatics.

Methods

PWMs used in the study

As a source of candidate PWMs we used the CIS-BP (Catalog of Inferred Sequence Binding Preferences) collection^{9,10} of pre-made matrices. For each TF, we gathered all PWMs assigned to the TF and added PWMs for related proteins sharing similar DNA binding domains. This was motivated by the results of the benchmarking study of Ambrosini *et al.*² where a PWM for some TF often displayed poorer TFBS recognition power than a PWM for some different TF but with the same DNA-binding domains.

The starting set of position frequency matrices was extracted from *TF_Information_all_motifs.txt* of CIS-BP 2.0 that includes models derived from direct experimental data for each TF and models that can be inferred given the TF family-specific threshold on DNA-binding domain similarity, see Ref. 11. In *Figure 2* such PWMs are referred to as ‘direct’ and ‘inferred’. All position frequency matrices were converted to log-odds PWMs as in Ref. 22 with an arbitrarily selected word count of 100, a pseudocount of 1, and uniform background nucleotide probabilities. For each TF, the set of PWMs was additionally extended by considering related TFs, i.e. PWMs for all ETV* TFs were added to the ETV1 PWM set, all FOX* (Forkhead box) PWMs were added to the FOXA2 PWM set, etc. (e.g. YY1 and YY2 PWM sets were identical). This procedure was not performed for ZNF* (zinc finger) TFs as these TFs can recognize very dissimilar motifs and thus additional PWMs of other ZNFs would unlikely provide any benefit.

Determination of transcription factor binding preference using PWMs

To assess with a particular PWM whether an SNV affects transcription factor binding, we used PERFECTOS-APE⁶ that estimates the log-fold change of motif *P*-values computed for best PWM hits detected among sites overlapping the first and the second of two alternative alleles. To use the prediction as a binary classifier, we treated the cases with *P* > 0.005 at both alleles as predicted negatives and used the log-fold change as the prediction score in the remaining cases. The auc function of the sklearn.metrics Python package was used to estimate the area under the precision-recall curve (AUPRC).

Estimating PWM performance with SNP-SELEX data

To provide a fair assessment, we mimicked the benchmarking protocol of Yan *et al.* Particularly, true positives and true negatives were selected from the SNP-SELEX data as follows. 1st batch data positives: PBS *P*-value < 0.01 and OBS *P*-value < 0.05; negatives: PBS *P*-value > 0.5 and OBS *P*-value < 0.05. 2nd batch data positives: PBS *P*-value < 0.01, negatives: PBS *P*-value > 0.5. For each TF, we tested each CIS-BP PWM from its PWM set. For each TF, the PWM reaching the highest AUPRC on the 1st batch data was selected for evaluation against the best PWM on the 1st batch (*Figure 2a*) and against deltaSVM on the 2nd batch of SNP-SELEX data (*Figure 2e*). Performance estimates for deltaSVM models (used in *Figure 2c,e*) were extracted from Supplementary Table S7 of Yan *et al.* Performance estimates of $\Delta\text{PWM}^{\text{Mult}}$ (used in *Figure 2a*) were kindly shared on our request by the authors.⁴ We also mimicked the stratified five-fold cross-validation procedure used by Yan *et al.* The mean of AUPRC across the folds was used to compare the performance of $\Delta\text{PWM}^{\text{CIS-BP}}$ with deltaSVM of Yan *et al.* at the first batch of SNP-SELEX data (*Figure 2c*).

Applying PWMs for analysis of allele-specific binding

The data on allelic binding ratios at individual SNPs and respective ΔPWM predictions of Yan *et al.* (*Figure 2f*, compare to *Figure 2d* of Yan *et al.*) were kindly shared on our request by the authors. The data included 193 TF-SNP pairs demonstrating allelic imbalance with 101 of 193 pairs annotated with the ΔPWM predictions. For these SNPs,

we obtained PWM predictions with the same protocol as for the SNP-SELEX data using the best PWMs selected with the 1st batch of the SNP-SELEX data.

Data availability

Source data

Original data on preferential binding SNPs as well as Δ PWM and deltaSVM predictions are provided in the supplementary materials section of the Yan *et al.* paper.⁴

CISBP Human PWMs collection was extracted from CIS-BP 2.0.^{9,10}

Extended data

Figshare: PWM-evaluation-using-SNP-SELEX, <https://doi.org/10.6084/m9.figshare.16906789.v1>.²¹

This project contains the following extended data:

- **Supplementary table S1** (Overview of PWMs and their performance in recognizing SNPs affecting transcription factor binding in SNP-SELEX data.)
- **Supplementary figure S1** (Performance of Δ PWM^{CIS-BP} in predicting weak and strong TF binding sites.)

Data are available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

Acknowledgements

This study was supported by Russian Science Foundation grant 20-74-10075 to IVK.

References

1. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat. Rev. Genet.* 2004; **5**: 276–287.
[Publisher Full Text](#)
2. Ambrosini G, *et al.*: **Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study.** *Genome Biol.* 2020; **21**: 114.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Kibet CK, Machanick P: **Transcription factor motif quality assessment requires systematic comparative analysis.** *F1000Research.* 2015; **4**(ISCB Comm J): 1429.
[Publisher Full Text](#)
4. Yan J, *et al.*: **Systematic analysis of binding of transcription factors to noncoding variants.** *Nature* 2021; **591**: 147–151.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Macintyre G, Bailey J, Haviv I, *et al.*: **is-rSNP: a novel technique for in silico regulatory SNP detection.** *Bioinformatics* 2010; **26**: i524–i530.
[PubMed Abstract](#) | [Publisher Full Text](#)
6. Vorontsov IE, Kulakovskiy IV, Khimulya G, *et al.*: **PERFECTOS-APE - Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation.** *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms 102–108 (SCITEPRESS - Science and Technology Publications* 2015.
[Publisher Full Text](#)
7. Coetzee SG, Coetzee GA, Hazelett DJ: **motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites.** *Bioinformatics* 2015; **31**: btv470–bt3849.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Deplancke B, Alpern D, Gardeux V: **The Genetics of Transcription Factor DNA Binding Variation.** *Cell* 2016; **166**: 538–554.
[Publisher Full Text](#)
9. Lambert SA, *et al.*: **The Human Transcription Factors.** *Cell* 2018; **172**: 650–665.
[Publisher Full Text](#)
10. Weirauch MT, *et al.*: **Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity.** *Cell* 2014; **158**: 1431–1443.
[PubMed Abstract](#) | [Publisher Full Text](#)
11. Yin Y, *et al.*: **Impact of cytosine methylation on DNA binding specificities of human transcription factors.** *Science* 2017; **356**: eaaj2239.
[PubMed Abstract](#) | [Publisher Full Text](#)
12. Grau J, Posch S, Grosse I, *et al.*: **A general approach for discriminative de novo motif discovery from high-throughput data.** *Nucleic Acids Res.* 2013; **41**(21): e197.
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Siebert M, Söding J: **Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences.** *Nucleic Acids Res.* 2016; **44**(13): 6055–6069.
[PubMed Abstract](#) | [Publisher Full Text](#)
14. Guo Y, Tian K, Zeng H, *et al.*: **A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction.** *Genome Res.* 2018; **28**(6): 891–900.
[PubMed Abstract](#) | [Publisher Full Text](#)
15. Mordelet F, Horton J, Hartemink AJ, *et al.*: **Stability selection for regression-based models of transcription factor-DNA binding specificity.** *Bioinformatics (Oxford, England).* 2013; **29**(13): i117–i125.
[PubMed Abstract](#) | [Publisher Full Text](#)
16. Le DD, Shimko TC, Aditham AK, *et al.*: **Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding.** *Proc. Natl. Acad. Sci. U. S. A.* 2018; **115**(16): E3702–E3711.
[PubMed Abstract](#) | [Publisher Full Text](#)
17. Dresch JM, Zellers RG, Bork DK, *et al.*: **Nucleotide Interdependency in Transcription Factor Binding Sites in the Drosophila Genome.** *Gene Regul. Syst. Biol.* 2016; **10**: 21–33.
[PubMed Abstract](#) | [Publisher Full Text](#)
18. Lee D, *et al.*: **A method to predict the impact of regulatory variants from DNA sequence.** *Nat. Genet.* 2015; **47**: 955–961.
[PubMed Abstract](#) | [Publisher Full Text](#)

19. Degtyareva AO, Antontseva EV, & Merkulova TI: **Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases.** *Int. J. Mol. Sci.* 2021;**22**(12): 6454.
[PubMed Abstract](#) | [Publisher Full Text](#)
20. Huo Y, Li S, Liu J, *et al.*: **Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk.** *Nat. Commun.* 2019; **10**(1): 670.
[PubMed Abstract](#) | [Publisher Full Text](#)
21. Abramov S: **PWM evaluation using SNP-SELEX.** *figshare.*
[Publisher Full Text](#)
22. Lifanov AP, Makeev VJ, Nazina AG, *et al.*: **Homotypic Regulatory Clusters in Drosophila.** *Genome Res.* 2003; **13**: 579–588.
[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:



Version 2

Reviewer Report 27 June 2022

<https://doi.org/10.5256/f1000research.135316.r141836>

© 2022 Machanick P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Philip Machanick 

¹ Computer Science, Rhodes University, Makhanda, South Africa

² Computer Science, Rhodes University, Makhanda, South Africa

Thank you for the updates.

They are a sufficient response to my initial review.

I include citation of one of my papers ¹ as further reading. Motif assessment is a difficult topic and there is no one solution that fits all cases.

References

1. Kibet CK, Machanick P: Transcription factor motif quality assessment requires systematic comparative analysis. *F1000Res*. 2015; **4**. [PubMed Abstract](#) | [Publisher Full Text](#)

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transcription factor binding specificity (computer science perspective).

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 27 June 2022

<https://doi.org/10.5256/f1000research.135316.r141837>

© 2022 Levitsky V. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Victor G. Levitsky

¹ Department of System Biology, Institute of Cytology and Genetics, Novosibirsk, Russian Federation

² Department of System Biology, Institute of Cytology and Genetics, Novosibirsk, Russian Federation

Though I still disagree with authors on the principal motivation of study, i.e. I mean that the traditional PWM and the alternative models neither good or bad, they predict sites of different structure. But let the current paper reflects the advantages of PWMs, so I agree that the manuscript could be accepted

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, massive analysis of genome data

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 07 April 2022

<https://doi.org/10.5256/f1000research.79349.r128877>

© 2022 Machanick P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Philip Machanick**

¹ Computer Science, Rhodes University, Makhanda, South Africa

² Computer Science, Rhodes University, Makhanda, South Africa

³ Computer Science, Rhodes University, Makhanda, South Africa

The Yan *et al.* article is a useful addition to the literature so questioning the validity of their results is also useful. However, this article makes an exaggerated claim of the extent to which Yan *et al.* reduce the utility of PWMs.

PWMs are already known to be potentially flawed models, varying from very accurately predicting DNA binding specificity to poorly doing so, with potential confounders like cofactors ¹ and indirect binding ³. The extent to which these issues apply can depend on the method used to determine the PWM. For example, using ChIP-seq data can create a composite motif incorporating part of a cofactor. Using PBM may eliminate this effect, but can produce poor binding specificity in some cases, possibly because either the specificity is mediated by cofactor requirements or because binding is indirect ^{4,5}.

While I am specifically mentioning older approaches like PBMs here, any *in vitro* or *in silico* method potentially has similar issues.

For this reason, I advocate using a range of methods to assess motif quality ².

As regards SNPs and other variability, this sort of issue has to be taken into account, otherwise any variation in specificity may not correspond to *in vivo* reality.

So, back to the approach of this paper: selecting PWMs that match specific criteria for reliability. It is not clear to me that this in any way invalidates the results of Yan *et al.* as there is variability in the predictive quality of PWMs, given the potential for confounders.

I would like to see a clearer explanation of the extent to which Yan *et al.* actually diminish the utility of PWMs (noting this is in a specific context, assessing small genomic variants) and the extent to which this review generalises beyond carefully selected PWMs..

References

1. Gao Z, Ruan J: A structure-based Multiple-Instance Learning approach to predicting *in vitro* transcription factor-DNA interaction. *BMC Genomics*. 2015; **16 Suppl 4**: S3 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Kibet CK, Machanick P: Transcription factor motif quality assessment requires systematic comparative analysis. *F1000Res*. 2015; **4**. [PubMed Abstract](#) | [Publisher Full Text](#)
3. Ambrosini G, Vorontsov I, Penzar D, Groux R, et al.: Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biology*. 2020; **21** (1). [Publisher Full Text](#)
4. Kerschgens J, Renaud S, Schütz F, Grasso L, et al.: Protein-binding microarray analysis of tumor suppressor AP2α target gene specificity. *PLoS One*. 2011; **6** (8): e22895 [PubMed Abstract](#) | [Publisher Full Text](#)
5. Siggers T, Duyzend MH, Reddy J, Khan S, et al.: Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol Syst Biol*. 2011; **7**: 555 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the rationale for commenting on the previous publication clearly described?

Partly

Are any opinions stated well-argued, clear and cogent?

Partly

Are arguments sufficiently supported by evidence from the published literature or by new data and results?

Partly

Is the conclusion balanced and justified on the basis of the presented arguments?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transcription factor binding specificity (computer science perspective).

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 21 Jun 2022

Sergey Abramov

The Yan et al. article is a useful addition to the literature so questioning the validity of their results is also useful. However, this article makes an exaggerated claim of the extent to which Yan et al. reduce the utility of PWMs.

We did our best to clarify our claim as it was also questioned by Dr. Levitsky.

PWMs are already known to be potentially flawed models, varying from very accurately predicting DNA binding specificity to poorly doing so, with potential confounders like cofactors 1 and indirect binding 3. The extent to which these issues apply can depend on the method used to determine the PWM. For example, using ChIP-seq data can create a composite motif incorporating part of a cofactor. Using PBM may eliminate this effect, but can produce poor binding specificity in some cases, possibly because either the specificity is mediated by cofactor requirements or because binding is indirect 4,5.

We fully agree that the data source and the computational procedure used to derive the TFBS model would significantly affect the result in terms of whether it reflects the genuine TF binding specificity or significantly depends on confounding factors. In this paper we restricted ourselves to a more specific context of using PWMs for quantifying the variants identified with the SNP-SELEX, which is an in vitro assay, so indirect binding and cofactors do not influence the outcome. To make it clear, we have revised the Introduction section of our manuscript.

For this reason, I advocate using a range of methods to assess motif quality 2.

Indeed. A comprehensive assessment of motif models using different types of experimental data was performed e.g. in Ambrosini et al. 2022. In this study, we did not focus on selecting the optimal PWMs for a wide range of practical applications or in terms of representing in vivo binding. Our aim was to demonstrate that PWMs provide the type of a model, which is able to show a reasonable performance in classifying differentially bound oligonucleotides with single-nucleotide substitutions.

So, back to the approach of this paper: selecting PWMs that match specific criteria for reliability. It is not clear to me that this in any way invalidates the results of Yan et al. as there is variability in the predictive quality of PWMs, given the potential for

confounders. I would like to see a clearer explanation of the extent to which Yan et al. actually diminish the utility of PWMs (noting this is in a specific context, assessing small genomic variants) and the extent to which this review generalises beyond carefully selected PWMs.

The same issue was pointed out by Dr. Levitsky and we did our best to clarify the aim of the study in the revised version of the manuscript. We believe, that quantifying the effects of single nucleotide variants on TF binding is an important practical problem emerging in the increasingly influential field of personalized genomics, as according to the recent reports up to 80% of causal variants are found in the regulatory regions [see e.g. <https://www.medrxiv.org/content/10.1101/2021.06.08.21258515v2>]. Even though this is a limited problem, it is worth clarifying the PWM performance for this particular application. We have added the necessary information in the Introduction section.

Competing Interests: No competing interests were disclosed.

Reviewer Report 14 February 2022

<https://doi.org/10.5256/f1000research.79349.r123085>

© 2022 Levitsky V. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Victor G. Levitsky

¹ Department of System Biology, Institute of Cytology and Genetics, Novosibirsk, Russian Federation

² Department of System Biology, Institute of Cytology and Genetics, Novosibirsk, Russian Federation

³ Department of System Biology, Institute of Cytology and Genetics, Novosibirsk, Russian Federation

1.
Boytsov et al. in the Abstract of their correspondence cited Yan et al. paper
Yet, recently Yan et al. presented new experimental method for analysis of regulatory variants and, based on its results, reported that "the position weight matrices of most transcription factors lack sufficient predictive power". Here, we reanalyze the rich experimental dataset obtained by Yan et al. and show that appropriately selected position weight matrices in fact can successfully quantify transcription factor binding to alternative alleles...

But actually, Yan et al. in the Abstract wrote:
...the position weight matrices of most transcription factors lack sufficient predictive power, whereas the support vector machine combined with the gapped k-mer representation show much improved performance, when assessed on results from independent SNP-SELEX experiments involving a new set of 61,020 sequence variants....

I think that Yan et al. are not wrong

Since, in particular, Yan et al. also wrote that

...(1) We reasoned that the poor performance of many PWMs was probably because they did not take into account dinucleotide interdependency in transcription factor–DNA interactions and the influence of flanking DNA sequences^{11,12}. Previous studies have shown that dinucleotide interdependency exists for some transcription factor dimers⁴. For example, according to the PWM model, the SNP rs79124498—located within a binding site of HLF, a bZIP family transcription factor that binds DNA as homodimers—would have little effect on HLF binding. However, SNP-SELEX indicated that the G allele bound more strongly than the T allele to HLF. This could be caused by the dinucleotide interdependency between position 2 (the SNP position) and position 10 in the binding site (Fisher’s exact test $P < 2.2 \times 10^{-16}$, odds ratio = 3.34)...

...(2) PWM performed poorly for SNPs located in low-affinity binding sites of transcription factors. However, this limitation could be overcome by using deltaSVM. When we categorized SNPs into five quantiles on the basis of their binding affinities as measured by OBS, and assessed the performance of PWM and deltaSVM in predicting their allelic binding by fivefold cross-validation or using the novel batch of SNP-SELEX experimental results (Extended Data Fig. 7), deltaSVM outperformed PWM in all quantiles, particularly in the lower quantiles corresponding to weak transcription factor binding sites.....

Does Boytsov et al. not agree with Yan et al. in these two points?

As I know, the alternative model can outperform the standard PWM model (e.g. BaMM, Siebert, M. and Söding, J. 2016 Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, 44, 6055–6069). At least this should be if an alternative model incorporates a PWM model (as BaMM does).

Hence, Yan et al. compared PWM (e.g. in BEESEM realization) and deltaSVM, and proved that PWMs are worse than deltaSVM. Boytsov et al. used additional PWMs from public databases such as CIS-BP to select the best performed PWM. This actually proves that ready PWMs (that respects to the same family, i.e. to other TFs with the same DNA binding domain) may be quite successive, but this does not prove that PWMs are better than deltaSVMs. This also does not imply that PWMs are good or bad. We should develop a special pipeline to compare PWMs and deltaSVMs (this is out of scope of paper). Boytsov et al. did not try to incorporate the non-traditional model deltaSVMs and data from CIS-BP to potentiate the performance of deltaSVMs. So about what the Boytsov paper? Hence, Boytsov et al. proved that BEESEM realization may be better if we incorporate CIS-BP data or what?

Boytsov et al. concluded

...However, properly selected PWMs achieve performance that is very close and in some cases even better than that of deltaSVM. Despite the simplicity of the PWM model, its construction is not trivial and its success depends both on the motif discovery algorithm and reliability of the training data...

Any motif discovery algorithm does not use any motif library on the process of de novo search. Usually, motif libraries are applied to interpret enriched motifs (e.g. STREME and Tomtom in meme suite, <https://meme-suite.org/meme/index.html>) Hence, application of motifs library is not a step in de novo process. At this step, I again does not understand why Boytsov et al. compared Figure 1b with Fig. 2b of Yan et al.

Overall, Boytsov et al. should draw attention to the point of disagreement with data or conclusion of Yan et al. paper.

2.

The TF classification by family is wrongly described.

...For each TF, the set of PWMs was additionally extended by considering related TFs, i.e. PWMs for all ETV* TFs were added to the ETV1 PWM set, all FOX* (Forkhead box) PWMs were added to the FOXA2 PWM set, etc. (e.g. YY1 and YY2 PWM sets were identical). This procedure was not performed for ZNF* (zinc finger) TFs as these TFs can recognize very dissimilar motifs and thus additional PWMs of other ZNFs would unlikely provide any benefit...

This description does not explain several pairs from Supplementary Data (Overview of best CIS-BP PWMs), e.g. ETV2 & FLI1

The correct and default approach was described in the previous publication (Ambrosini G, et al. *Genome Biol.* 2020: Matrices were manually mapped to gene symbols and TF families from TFclass [Wingender E, et al. *Nucleic Acids Res.* 2018] and CIS-BP). Moreover, the CIS-BP database contains TF PWMs that were already classified by families.

3.

Currently, links to Figure 1 are contained in the Introduction section. Although the format of correspondence paper is flexible, I propose that authors should either do not use various sections, or apply the standard sections, Introduction, Methods, Results, Conclusions/Discussion.

References

1. Yan J, Qiu Y, Ribeiro Dos Santos AM, Yin Y, et al.: Systematic analysis of binding of transcription factors to noncoding variants. *Nature*. **591** (7848): 147-151 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Siebert M, Söding J: Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*. 2016; **44** (13): 6055-6069 [Publisher Full Text](#)
3. Ambrosini G, Vorontsov I, Penzar D, Groux R, et al.: Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biology*. 2020; **21** (1). [Publisher Full Text](#)

Is the rationale for commenting on the previous publication clearly described?

Partly

Are any opinions stated well-argued, clear and cogent?

No

Are arguments sufficiently supported by evidence from the published literature or by new data and results?

Partly

Is the conclusion balanced and justified on the basis of the presented arguments?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, massive analysis of genome data

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 21 Jun 2022

Sergey Abramov

... But actually, Yan et al. in the Abstract wrote: ...the position weight matrices of most transcription factors lack sufficient predictive power, whereas the support vector machine combined with the gapped k-mer representation show much improved performance, when assessed on results from independent SNP-SELEX experiments involving a new set of 61,020 sequence variants....

I think that Yan et al. are not wrong...

In fact, we do not challenge the authors' statement regarding the high performance of deltaSVM in predicting the SNP-SELEX results by sequence analysis. Particularly, we explicitly state that

"... our results do not compromise the high performance of deltaSVM, used by Yan et al. as an advanced substitution of position weight matrices (PWMs)" (paragraph 4 of Discussion).

Yet, we strongly disagree with the authors' conclusion on the PWM performance ("lack sufficient predictive power") and believe that their comparison of the performance of the PWM and deltaSVM became one-sided in favor of deltaSVM due to an accidental selection of particular PWMs in their study. We believe that public databases contain PWM which can display much better performance in quantifying allele-specific binding and put it explicitly in the Introduction that

"We show that the careful selection of PWMs of many TFs from a public database quantitatively explains the differential TF binding to allelic variants with reliability comparable to that of deltaSVM."

Again, we put into Discussion that

"Summing up, our results do not compromise the high performance of deltaSVM, used by Yan et al. as an advanced substitution of position weight matrices (PWMs). However, properly selected PWMs achieve performance that is very close and in some cases even better than that of deltaSVM."

We have also added a detailed discussion on the subject into the Discussion section (see the response to the next question of the reviewer).

Since, in particular, Yan et al. also wrote that ...(1) We reasoned that the poor performance of many PWMs was probably because they did not take into account dinucleotide interdependency in transcription factor-DNA interactions and the

influence of flanking DNA sequences...

...(2) PWM performed poorly for SNPs located in low-affinity binding sites of transcription factors.

Does Boytsov et al. not agree with Yan et al. in these two points?

We agree with the theoretical limitations of position weight matrices regarding their inability to account for non-additive contributions of particular nucleotides into protein affinities. Yet, it is important to distinguish between the general and well-known limitations of the PWM as a model and the low performance of particular matrices. In our opinion the poor performance of PWMs used in the study of Yan et al. was not due to the intrinsic inability of PWMs to classify TF binding preferences at particular TFBS, but rather due to the way the PWMs used in this study were selected/constructed. We prove it by providing the alternative PWMs that belong to the same class of mononucleotide models but perform better than the PWMs of Yan et al. and comparably to deltaSVM. We have added an explicit statement on this matter in the revised version of the manuscript.

"The objective of our study is by no means to undermine the necessity of complex TFBS models with dependent positional contributions. Advanced multiparametric and alignment-free approaches such as deltaSVM appear very likely to shape the oncoming future of transcription factor binding site models. Rather, we want to underline that the prediction performance for transcription factor binding sites in its current stage is more influenced by model training protocols than by model structure restrictions. PWMs still can deliver a solid standard in representation and bioinformatics analysis of the transcription factor binding sites, including assessment of the functional impact of single nucleotide variants in gene regulatory regions. In addition, we underline that better defined 'baseline' PWMs or PWM selection procedures are required for the proper evaluation of advanced models. It is important that such 'baseline' TFBS models, while certainly being handicapped by design, still reach meaningful prediction quality."

... ready PWMs (that respects to the same family, i.e. to other TFs with the same DNA binding domain) may be quite successive, but this does not prove that PWMs are better than deltaSVMs.

We used PWMs with the same DNA binding domains to increase the repertoire of candidate PWMs, from which the best PWM for assessing variants identified by SNP-SELEX experiments can be selected. To avoid confusion we have added two subsections to the Methods section: "PWMs used in the study" and "Selection of the best PWM for a TF". In fact, carefully selected PWMs outperformed deltaSVM models for 34 of 129 TFs (see paragraph 3 of Introduction in the manuscript and Figure 2), and many of these PWMs were initially constructed for different TFs and even different species (see Supplementary Table S1). This does not compromise better deltaSVM performance for other TFs (see Fig. 1e).

Hence, Boytsov et al. proved that BEESEM realization may be better if we incorporate CIS-BP data or what?

We did not test BEESEM or other types of motif discovery software or alternative PWM-like

motif representations, and thus don't know if they provide even better PWMs than we found in CIS-BP. We only used the existing published PWMs available in the CIS-BP database. SNP-SELEX provides a rich data source to test various types of models in the task of predicting rSNP effects on transcription factor binding, but such testing does not fit the scope of our manuscript.

Hence, application of motifs library is not a step in de novo process. At this step, I again does not understand why Boytsov et al. compared Figure 1b with Fig. 2b of Yan et al.

We did not discuss de novo motif discovery. The idea of our study was to verify whether the inadequate performance of PWMs reported in Yan et al. was related to the type of the model or if it characterized the particular PWMs they used.

In fact, Yan et al. also did not construct PWMs through de novo motif discovery but used the pre-made PWMs of Yin et al. Similarly, we followed the suit and avoided de novo motif discovery in favor of reusing existing PWMs from CIS-BP. Selection of a single PWM from the pool of related PWMs can be considered as "training" of the model, and we fully replicated the approach of Yan et al. i.e. the cross-validation on the 1st batch of SNP-SELEX data.

Overall, Boytsov et al. should draw attention to the point of disagreement with data or conclusion of Yan et al. paper.

We did our best to better highlight the key idea of the study in the revised version of the manuscript and added an extensive Discussion section.

2. The TF classification by family is wrongly described. ... This description does not explain several pair from Supplementary Data (Overview of best CIS-BP PWMs), e.g.ETV2 & FLI1

In the pool of possible PWMs for each TF, we included CIS-BP 'inferred' PWMs (as described in Methods) which belonged to TFs with a similar DNA-binding domain, hence there is no contradiction. We revised the Methods section, see the subsection "PWMs used in the study."

The correct and default approach was described in the previous publication (Ambrosini G , et al. Genome Biol. 2020: Matrices were manually mapped to gene symbols and TF families from TFclass [Wingender E,et al.. Nucleic Acids Res. 2018] and CIS-BP). Moreover, the CIS-BP database contains TF PWMs that were already classified by families.

CIS-BP classification of DNA-binding domains is very general and leads to very wide sets of PWMs potentially applicable to a particular TF, if all PWMs across the TF family are

considered. To reduce computational complexity, we made a compromise of including 'inferred' motifs (see above) but only for related proteins by matching gene names and not relying on the detailed TF family annotation. Even with this simplification in the PWM selection, which greatly reduced the number of available PWMs, the resulting performance of the best PWMs was significantly better than the PWM performance reported by Yan et al. for the same TF.

Of note, in Ambrosini et al. (2020) we used all-vs-all testing strategy and reported cross-family applicability of PWMs, although this is computationally ineffective in the practical selection of the best-performing matrices.

3. Currently, links to Figure 1 are contained in the Introduction section. Although the format of correspondence paper is flexible, I propose that authors should either do not user various sections, or apply the standard sections, Introduction, Methods, Results, Conclusions/Discussion.

We have revised the manuscript structure according to the reviewer's suggestion.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research