BMC Bioinformatics

RESEARCH Open Access

The evaluation of transcription factor binding site prediction tools in human and Arabidopsis genomes

Dinithi V. Wanniarachchi[1], Sameera Viswakula[2] and Anushka M. Wickramasuriya[1*]

*Correspondence:
anushka@pts.cmb.ac.lk

[1] Department of Plant Sciences, Faculty of Science, University of Colombo, Colombo 03, Sri Lanka
[2] Department of Statistics, Faculty of Science, University of Colombo, Colombo 03, Sri Lanka

Abstract

Background: The precise prediction of transcription factor binding sites (TFBSs) is pivotal for unraveling the gene regulatory networks underlying biological processes. While numerous tools have emerged for in silico TFBS prediction in recent years, the evolving landscape of computational biology necessitates thorough assessments of tool performance to ensure accuracy and reliability. Only a limited number of studies have been conducted to evaluate the performance of TFBS prediction tools comprehensively. Thus, the present study focused on assessing twelve widely used TFBS prediction tools and four de novo motif discovery tools using a benchmark dataset comprising real, generic, Markov, and negative sequences. TFBSs of *Arabidopsis thaliana* and *Homo sapiens* genomes downloaded from the JASPAR database were implanted in these sequences and the performance of tools was evaluated using several statistical parameters at different overlap percentages between the lengths of known and predicted binding sites.

Results: Overall, the Multiple Cluster Alignment and Search Tool (MCAST) emerged as the best TFBS prediction tool, followed by Find Individual Motif Occurrences (FIMO) and MOtif Occurrence Detection Suite (MOODS). In addition, MotEvo and Dinucleotide Weight Tensor Toolbox (DWT-toolbox) demonstrated the highest sensitivity in identifying TFBSs at 90% and 80% overlap. Further, MCAST and DWT-toolbox managed to demonstrate the highest sensitivity across all three data types real, generic, and Markov. Among the de novo motif discovery tools, the Multiple Em for Motif Elicitation (MEME) emerged as the best performer. An analysis of the promoter regions of genes involved in the anthocyanin biosynthesis pathway in plants and the pentose phosphate pathway in humans, using the three best-performing tools, revealed considerable variation among the top 20 motifs identified by these tools.

Conclusion: The findings of this study lay a robust groundwork for selecting optimal TFBS prediction tools for future research. Given the variability observed in tool performance, employing multiple tools for identifying TFBSs in a set of sequences is highly recommended. In addition, further studies are recommended to develop an integrated toolbox that incorporates TFBS prediction or motif discovery tools, aiming to streamline result precision and accuracy.

Keywords: Transcription factor binding sites, Bioinformatics tools, Performance evaluation

Background

The regulation of gene expression is governed by various factors, including transcription factors (TFs). Each cell type or tissue, at a specific developmental stage or in response to extracellular signals, exhibits a distinctive pattern of activated TFs. This repertoire of TFs is a key determinant for defining cellular identity and function. Studies have identified 274,633 TF-encoding genes from 183 animal genomes [1], including >1600 TFs from the human genome [1] and 320,370 TFs from the genomes of 165 plant species [2]; more than 1500 TFs have been reported from the Arabidopsis genome [3].

TFs modulate gene expression by binding to specific regulatory regions of the genes (upstream and basal promoter elements, enhancers, and silencers) [4, 5]. An important step in unraveling regulatory mechanisms of gene expression is the identification of these binding sites in DNA for TFs termed TF binding sites (TFBSs) [6, 7]. These are typically short and often degenerate stretches, ranging from 5 to 20 base pairs (bp) in length [8]. TFs possess specific DNA binding domains that recognize and bind to particular sequences within TFBSs. These DNA binding domains allow TFs to directly interact with the DNA molecule, facilitating the regulation of gene expression. Alternatively, TFs can bind to DNA indirectly by interacting with another TF [9].

There are many in vitro and in vivo experimental approaches to determine the interaction of TFs and their potential binding sites on genomic sequences. These methods include the Electro-Mobility Shift Assay [10], DNase I footprinting/protection assay [11], Systematic Evolution of Ligands by EXponential enrichment (SELEX) [12], and Chromatin ImmunoPrecipitation (ChIP) Assay [13]. Advances in these experimental technologies, particularly ChIP followed by sequencing (ChIP-seq), have led to the generation of large-scale datasets of TFBSs across the genome. These comprehensive datasets provide a valuable resource for scientists to develop and refine computational approaches for predicting TFBSs. Experimental validation of TFBSs is a complex and resource-intensive process. Consequently, in recent years, in silico prediction of TFBSs has emerged as an efficient alternative to these time-consuming experimental methods [5, 14–17].

Numerous algorithms have been developed to predict TFBSs. Many of the methods originally used for predicting TFBSs from sequences were based on position weight matrices (PWMs), also known as position-specific scoring matrices (PSSMs) [18]. These are the most widely used and well-established mathematical models for predicting TFBSs in DNA sequences [14, 19–22]. PWMs are computed from the multiple sequence alignment of an experimentally validated set of TFBSs, quantitatively scoring binding motifs based on the frequency and positions of the nucleotides within the binding sites for the corresponding TF [19, 23]. Several databases, such as JASPAR [24] and TRANSFAC [25], maintain comprehensive collections of TF-binding profiles as PWMs or position frequency matrices (PFMs).

JASPAR is an open-access resource that provides access to a manually curated, non-redundant set of TF-binding profiles in the form of PFMs. These PFMs are derived from experimentally defined TFBSs and can be converted into PWMs, which can then be used to scan DNA sequences to predict TFBSs. TRANSFAC, on the other hand, is a commercial database that includes PWMs derived from published collections of experimentally validated binding sites. In addition to these databases, several other resources provide access to collections of PWMs, such as HOmo sapiens

COmprehensive MOdel Collection (HOCOMOCO) [26], UniBind [27], UniPROBE [28], and Factorbook [29]. Several tools are available for predicting TFBSs by scanning for PWMs against DNA sequences. These include Find Individual Motif Occurrences (FIMO) [30], MATCH [31], MOtif Occurrence Detection Suite (MOODS) [32], Ciiider [33], Hypergeometric optimization of motif enrichment (HOMER) [34], INSECT [35], Matrix-scan [36], Morpheus [37], PWMscan [38], TFBStools [39], Contra V3 [40], TFBS-Discovery Tool Hub (TDThub [41]), TFBIND [42], PROMO [43], Cluster-Buster [44], rVISTA [45], LASAGNA-search [46], and PscanChIP [47]. Furthermore, several freely available de novo motif discovery tools have been developed to handle the large volumes of data generated from high-throughput technologies. These de novo discovery tools, such as AlignAce [48], Weeder [49], Improbizer [50], MotifSampler [51], rGADEM [52], Sensitive, Thorough, Rapid, Enriched Motif Elicitation (STREME) [53], and Multiple Em for Motif Elicitation (MEME) [54], are capable of identifying novel TFBS motifs without prior knowledge of the binding sites.

In addition to PWM-based models, more complex models have recently been proposed for modeling TFBSs with increased accuracy [15], in particular, hidden Markov models (HMMs) [23, 55–58], hierarchical mixture models [59], Bayesian network-based methods [60] and deep learning based neural networks [61–64]. Tools such as Motif cluster alignment and search tool (MCAST) [65], TFBSpred [17], and FABIAN-variant [66] utilize HMM-based methods while VOMABT [67], and MotEvo [68], tools use methods based on Bayesian network models. Tools based on deep learning methods include DeepBind [69], DeeperBind [70], DeepGRN [61], DeepSNR [71], DeepSTF [62], Desso [72], MaResNet [73], TAMC [74] and TFImpute [75]. While a variety of computational tools are available for predicting TFBSs, as shown in Fig. 1, PWM-based methods remain widely used for their simplicity in construction from a set of sequences, and the availability of several curated databases of PWMs applicable to several species.

Accurate prediction of TFBSs in silico is important for understanding gene regulation. However, only a limited number of independent assessments have been conducted to evaluate the performance of TFBS prediction and motif discovery tools [14, 76–81], with the latest comprehensive analysis dating back to 2016. As the field of computational biology continues to evolve and new tools are developed, assessment of tool performance is essential to ensure accurate and reliable predictions. The findings of such studies will guide the researchers to choose the most appropriate tool(s) for their analyses. Therefore, the present study aimed to evaluate the performance of several widely used TFBS prediction and de novo motif discovery tools, including both previously evaluated and untested tools. The specific objectives of this study were to (1) create a benchmark dataset to evaluate the performance of tools, (2) assess the performance quality of tools using numerous statistics to identify the best tool(s), and (3) conduct case studies involving known genes from key biological pathways in Arabidopsis and humans, using the identified best tool(s) to predict TFBSs.

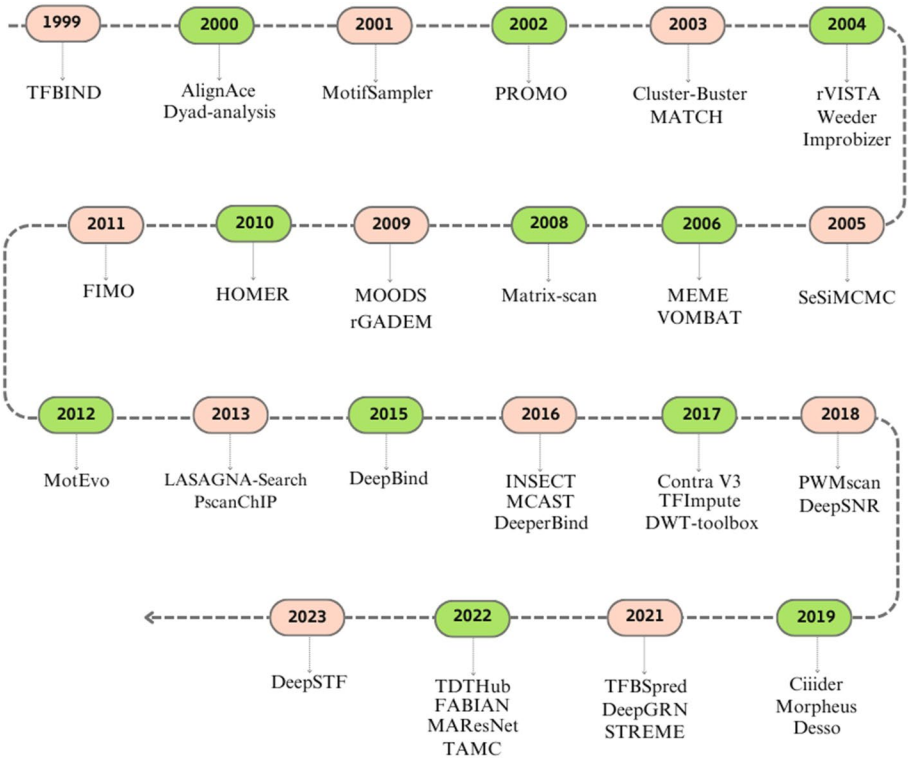

Fig. 1 Tools available for modeling transcription factor binding sites

Table 1 Transcription factors used in the present study

Organism	Transcription factor name	JASPAR ID
Arabidopsis thaliana	APETALA3 (AP3)	MA0556.1
	G-box binding factor 3 (GBF3)	MA1351.1
Homo sapiens	Transcription factor AP-2 alpha (TFAP2A)	MA0003.2
	Paired box 5 (PAX5)	MA0014.2

Methods

Development of a benchmark dataset

A benchmark dataset was generated following the recommendations of Tompa et al. 2005 [77], to assess the performance of TFBS prediction tools. This dataset comprised four types of sequences: real, generic, Markov, and negative. Real sequences, containing experimentally validated binding sites, were collected from the JASPAR database [24, 82] for the TFs listed in Table 1 for both *Arabidopsis thaliana* and *Homo sapiens*. Five sequences of 115 bp length were randomly selected per TF, resulting in a total of 20 real sequences. The choice of 115 bp was based on the typical length of known binding sites from the JASPAR database, which are derived from ChIP-seq experiments and generally fall within this range. Generic sequences were generated by randomly selecting promoter sequences from the genomes of *A. thaliana* and *H. sapiens*. Initially, gff3 files of *H. sapiens* and *A. thaliana* reference genome sequences were obtained from GENCODE [83] and NCBI [84], respectively. Ten genes were randomly extracted per organism, and

115 bp long promoter sequences (upstream from the transcription start site (TSS)) were obtained for each gene using the Ensembl genome browser [85], resulting in 20 generic sequences. Markov (or artificial) sequences (20 sequences) were generated using a third-order Markov model. Experimentally validated binding sites obtained from the JASPAR database were then implanted into the generic sequences and Markov sequences. Negative sequences, lacking binding sites for the selected four TFs, were also acquired from the JASPAR database; sequences containing similar sites were removed from the dataset, resulting in 20 negative sequences. The real, generic, Markov, and negative sequences, totaling 80 sequences compiled into a single file in FASTA format to create the benchmark dataset (Additional file 1).

Selection of tools for performance evaluation

As shown in Fig. 1, several tools have been released since 2019. However, the currently available deep learning approaches for predicting TFBSs have primarily been trained on human datasets [61, 62, 72–74]. Consequently, their direct applicability and performance on plant-specific data remain uncertain without significant retraining and validation [86–88]. Therefore, in the current study, we opted to exclude deep learning models from our evaluation.

Given that this study involves TFs from both human and Arabidopsis genomes, we focused our evaluation on tools mainly based on PWMs and HMMs. However, we did not include the recently released FABIAN-variant tool [66] in our analysis. Although it uses transcription factor flexible models (an extension of HMMs) and PWMs, FABIAN-variant is specifically designed to predict the effects of DNA variants on human TFs [66]. Additionally, we excluded two PWM-based tools developed after 2019, TDThub and TFBSPred. TDThub is a web server that relies on a set of pre-computed TFBS in plant species [41], and TFBSPred is designed for human and mouse TFs [17], limiting their relevance to our study, which includes both human and Arabidopsis TFs.

In our study, we evaluated the performance of 12 TFBS prediction tools and four de novo motif discovery tools commonly used in research. This focused evaluation ensures that our findings are robust and relevant to both human and plant TFs. Table 2 provides an overview of the tools employed in this study. PWMs for the selected TFs were obtained from the JASPAR database in JASPAR, MEME, and TRANSFAC matrix formats; the list of PWMs used in this study, along with related information is provided in Additional file 2. Default parameters were applied to all tools to enable comparisons among them.

Performance metrics

Several statistical parameters were used to assess the performance quality of the tools. For each tool, the number of true positives (nTPs), false positives (nFPs), and false negatives (nFNs) were calculated for the benchmark dataset, considering different overlap percentages between the lengths of predicted and known binding sites or motifs (i.e., 80, 90, and 100%). True positives (TPs) denote the number of nucleotide positions in both known and predicted binding sites, false positives (FPs) represent the number of nucleotide positions not in known binding sites but in predicted binding sites, and false negatives (FNs) indicate the number of nucleotide positions in known binding sites

Table 2 Brief description of 16 analyzed tools

Tool	Description	URL	References
TFBS prediction tools			
Ciiider	A software which uses a combination of PWMs and MATCH algorithm to predict TFBSs	[89]	[33]
DWT-toolbox	A regulatory motif model that incorporates arbitrary pairwise dependencies between positions in binding sites using a Bayesian framework along with PSWMs	[90]	[106]
FIMO	Part of MEME-Suite package which search for individual occurrences of a motif in a DNA or protein sequence	[91]	[30]
HOMER	A software that can be implemented in UNIX operating systems, which offers a set of tools for motif discovery and analysis	[92]	[34]
INSECT 2.0	An online tool which allows the prediction of *cis*-regulatory modules (CRMs) at the genome level and long lists of genes. CRM search is done using PWMs available in databases	[93]	[35]
Matrix-scan	A part of regulatory sequence analysis tools (RSAT) which predicts TFBS using PWMs	[94]	[36]
MCAST	A part of the MEME-suite program which uses HMM with a p-value based scoring scheme to identify candidate CRMs which contain TFBSs	[95]	[65]
MOODS	A software implemented in C++ which uses advanced matrix matching algorithms to scan matrices against sequences	[96]	[32]
Morpheus	A web-based tool which allows using PWM models considering the dependencies between di or tri nucleotide positions	[97]	[37]
MotEvo	A software tool which uses a Bayesian probabilistic approach to predict TFBSs in multiple alignments of phylogenetically related DNA sequences	[98]	[68]
PWMScan	A web server which allows rapid scanning of genomes to match user-supplied or server-resident PWMs	[99]	[38]
TFBStools	A package based on R for the analysis and manipulation of TFBSs	[100]	[39]
De novo *motif discovery tools*			
MEME	A part of MEME-suite which discovers novel-ungapped, recurring motifs of fixed length	[101]	[54]
MotifSampler	A tool based on the Gibbs sampling algorithm, which allows the identification of novel motifs	[102]	[51]
rGADEM	An R package based on GADEM software, which identifies novel motifs in large-scale genomic sequence data	[103]	[52]
STREME	A part of MEME-suite which identifies novel-ungapped, recurring motifs of fixed lengths that are relatively enriched	[104]	[53]

but not in predicted binding sites [14, 77]. Employing these values, statistical parameters were calculated for each tool, including sensitivity (Sn) (Eq. (1)), positive predictive value (PPV) (Eq. (2)), performance coefficient (PC) (Eq. (3)), average site performance (ASP) (Eq. (4)), and geometric accuracy (ACCg) (Eq. (5)). The tools were individually ranked using Excel's RANK function, considering each of the above-mentioned statistical parameters. Subsequently, the average rank for each tool across all statistical parameters was computed using Excel's AVERAGE function.

$$\text{Sensitivity (Sn)} = (\text{nTPs})/(\text{nTPs} + \text{nFNs}) \tag{1}$$

$$\text{Positive predictive value (PPV)} = (\text{nTPs})/(\text{nTPs} + \text{nFPs}) \tag{2}$$

$$\text{Performance coefficient (PC)} = (\text{nTPs})/(\text{nTPs} + \text{nFNs} + \text{nFP}) \tag{3}$$

$$\text{Average site performance (ASP)} = (\text{Sn} + \text{PPV})/2 \tag{4}$$

$$\text{Geometric accuracy (ACCg)} = \sqrt{(\text{Sn} * \text{PPV})} \tag{5}$$

Evaluation of the best-performing tools

To assess the best-performing tools, we randomly selected 50 genes from both the *A. thaliana* and *H. sapiens* genomes (See Additional file 3 for the gene list). Promoter regions of these genes, defined as the 1000 bp upstream from the TSS, were retrieved from the Eukaryotic Promoter Database (EPD) [107]. These sequences were scanned for 879 vertebrate PWMs and 1,407 plant PWMs obtained from the JASPAR database (Additional file 2) using the three best-performing tools: MCAST, FIMO, and MOODS. All the predicted motifs were ranked by frequency, and the top 20 motifs from each tool were compared to evaluate the consistency of their prediction.

Case study 1: transcription factor binding sites in genes associated with the anthocyanin biosynthesis pathway in plants

A total of 149 genes involved in regulating anthocyanin biosynthesis in *A. thaliana* were obtained from Grau et al. 2022 [41] (Additional file 3). Additionally, we downloaded genes associated with anthocyanin biosynthesis in *Zea mays*, *Oryza sativa*, and *Glycine max* using TDThub [41] (Additional file 3). The promoter regions of these genes (1000 bp upstream from the TSS) were extracted from the EPD and NCBI databases [105, 107]. These sequences were then scanned against PWMs of all plant TFBSs available in the JASPAR database, using the three best-performing TFBS prediction tools identified in this study.

Case study 2: transcription factor binding sites in genes associated with the pentose phosphate pathway in humans

Thirty-one genes involved in the pentose phosphate pathway were retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Additional file 3). Promoter regions (1000 bp upstream from the TSS) were extracted from the EPD and scanned against 879 vertebrate PWMs obtained from the JASPAR database using the three best-performing tools.

Results

Performance evaluation of different TFBS prediction tools

The benchmark dataset comprised a total of 60 TPs, indicating that the maximum nTPs that any tool could predict in this study was limited to 60. Among the 12 TFBS prediction tools examined, none of them achieved 100% accuracy in predicting all 60 binding sites. However, both MotEvo and DWT-toolbox successfully identified all TPs at both 80% and 90% overlap (Table 3). Interestingly, while the nTPs remained consistent at 80% and 90% overlap with a few exceptions, there was a notable decrease in the nTPs identified at 100% overlap across all TFBS prediction tools, resulting in higher nFNs (Table 3). Particularly, TFBStools predicted fewer TPs at 80% overlap compared to other tools and failed to predict any binding sites at 90% and 100%

Table 3 Number of true positives (nTPs), false negatives (nFNs), and false positives (nFPs) detected by different TFBS prediction tools

Tool	Overlap percentage*								
	80%			90%			100%		
	nTPs	nFNs	nFPs	nTPs	nFNs	nFPs	nTPs	nFNs	nFPs
Ciiider	43	17	10	43	17	10	34	26	9
DWT-toolbox	60	0	65	60	0	65	45	15	65
FIMO	57	3	17	57	3	17	45	15	17
HOMER	58	2	1002	58	2	1002	43	17	1002
INSECT 2.0	45	15	12	45	15	12	36	24	12
Matrix scan	50	10	13	50	10	13	41	19	13
MCAST	51	9	0	51	9	0	42	18	0
MOODS	52	8	12	52	8	12	43	17	12
Morpheus	54	6	142	51	9	142	46	14	142
MotEvo	60	0	35	60	0	35	45	15	35
PWMScan	57	3	20	57	3	20	45	15	20
TFBStools	9	51	2	0	60	2	0	60	2

* Overlap percentage: the percentage overlap between the length of predicted and known binding sites

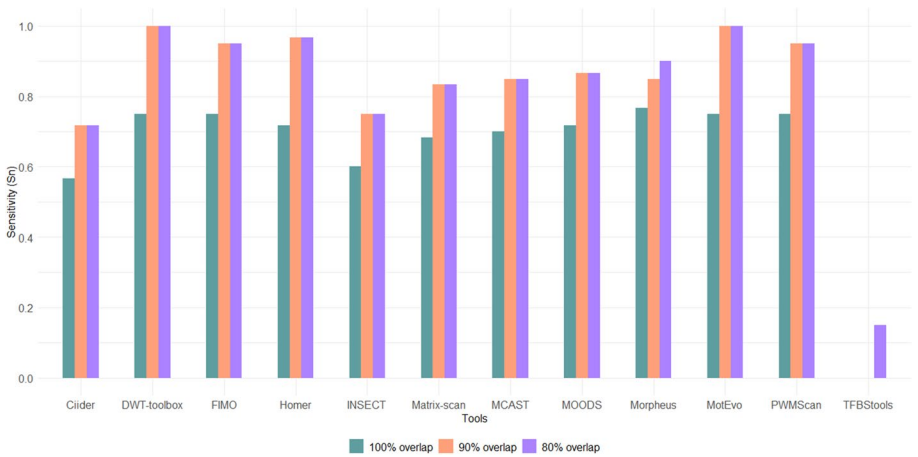

Fig. 2 Sensitivity of TFBS prediction tools at different overlap percentages

overlap. Additionally, the nFPs identified by the HOMER tool was relatively high compared to the other TFBS prediction tools examined (Table 3).

At both 80% and 90% overlap, Sn was generally higher compared to 100% for all the tools except for TFBStools (Fig. 2). Although Morpheus showed the highest sensitivity at 100% overlap with a sensitivity value of 0.767, DWT-toolbox, FIMO, HOMER, MotEvo, and PWMScan surpassed Morpheus at both 80% and 90% overlap. Moreover, TFBStools, which was unable to predict any binding sites at both 90% and 100% overlap, was able to predict binding sites at 80% overlap. However, the Sn was very low (0.150). Overall, it was evident that DWT-toolbox, MotEvo, Morpheus, PWMScan, and FIMO consistently delivered $Sn \geq 0.750$ across all three overlap percentages.

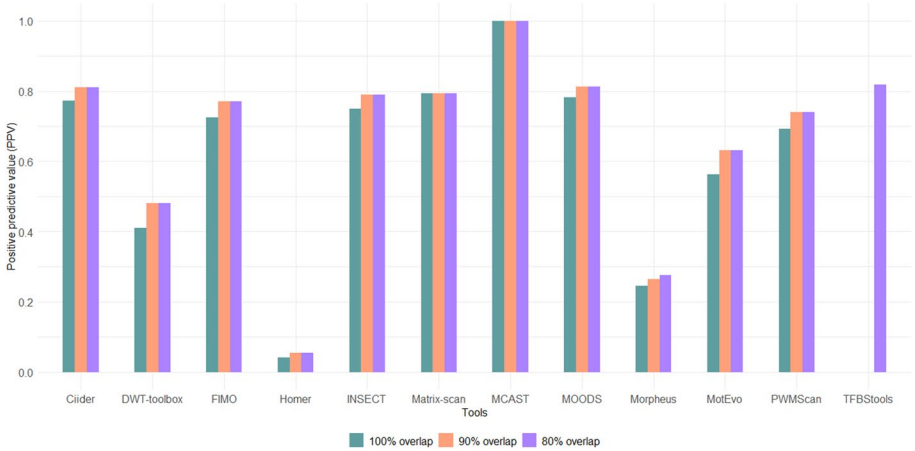

Fig. 3 Positive predictive values of TFBS prediction tools at three different overlap percentages

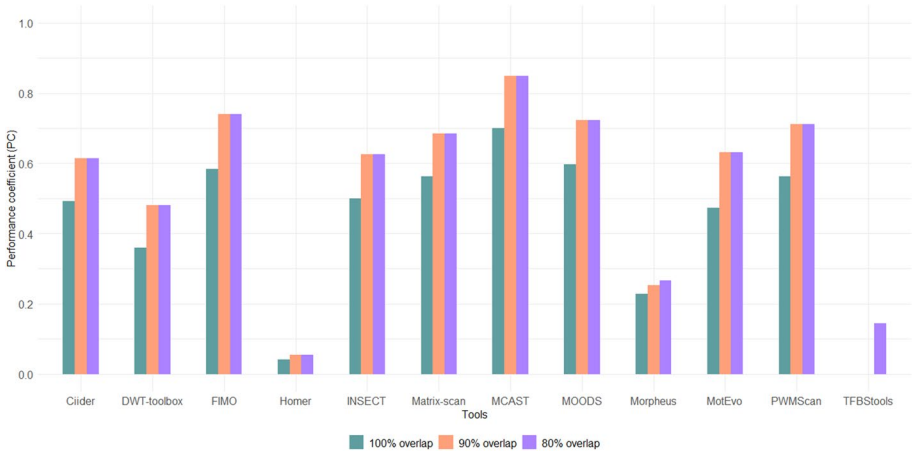

Fig. 4 Performance coefficient of TFBS prediction tools at three different overlap percentages

When examining the PPVs obtained for the three different overlap percentages, it was observed that the tools Ciiider, INSECT 2.0, Matrix-scan, MCAST, and MOODS consistently provided the best PPVs (0.750 or higher) throughout all three overlap percentages; of these, MCAST tool exhibited the highest PPV (1.000) at all overlap percentages (Fig. 3). Although the PPV of TFBStools was null at 100% and 90% overlap, a significant increase was observed at 80% overlap, making it the second highest (Fig. 3). Additionally, the PPV of the DWT-toolbox, HOMER and Morpheus tools was relatively low compared to other tools examined, at all three overlap percentages (PPV < 0.500).

Assessing the performance of TFBS tools based on the PC values revealed that the MCAST tool exhibits the best performance across all three overlap percentages (Fig. 4). Following MCAST, the tools MOODS and FIMO, secured the second and third-highest PC values, respectively, at 100% overlap (Fig. 4). However, at 90% and 80% overlap, FIMO emerged as the tool with the second-highest PC value, while MOODS exhibited the third-highest PC value (Fig. 4). Similar to statistical parameters: Sn, and PPVs, the TFBStools showed zero PC at 100% and 90% overlap but exhibited a value of 0.145 at

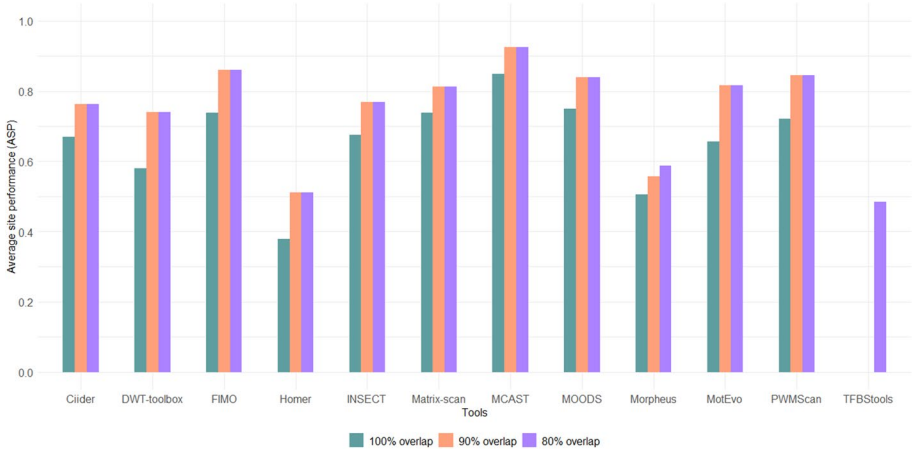

Fig. 5 Average site performance of TFBS prediction tools at three different overlap percentages

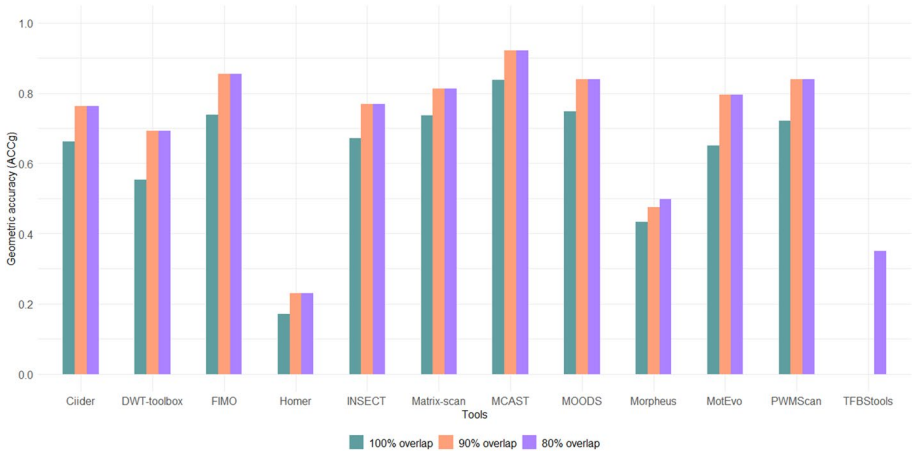

Fig. 6 Geometric accuracy of TFBS prediction tools at three different overlap percentages

80% overlap (Fig. 4). Additionally, the PC of DWT-toolbox, HOMER, and Morpheus tools were considerably low across all three overlap percentages (PC < 0.500).

As shown in Fig. 5, MCAST demonstrated the best performance across all three overlap percentages based on the estimated ASP measures. Similar to the tool performances based on PCs, the MOODS tool exhibited the second-highest ASP and the FIMO tool showed the third-highest ASP at 100% overlap (Fig. 5). Furthermore, at 90% and 80% overlap, the second-best and third-best ASPs were shown by FIMO and MOODS tools, respectively. It was also noted that TFBStools showed ASP value only at 80% overlap. Similar to PCs and ASP measures, a consistent trend in tool performance was observed when considering the ACCg of the TFBS prediction tools (Fig. 6). The MCAST demonstrated the highest ACCg across all overlap percentages, followed by FIMO and MOODS tools. HOMER and Morpheus tools exhibited relatively low ACCg compared to other tools examined, at all three overlap percentages (ACCg < 0.500). Taking into account the PC (Fig. 4), ASP (Fig. 5), and ACCg (Fig. 6), it became evident that the best-performing tools were MCAST, MOODS, and FIMO.

The best TFBS prediction tool/tools could not be solely identified from the values of statistical parameters, as various tools dominated different parameters. Hence, to determine the best-performing tool/tools across all the statistical parameters considered in the present analysis, the tools were ranked under each parameter, and the average rank was calculated. The average ranks obtained at different overlap percentages are provided in Additional file 4. It was revealed that the MCAST tool outperformed others at 100% overlap, while MOODS and FIMO tools secured the second and third positions, respectively. The average ranks obtained at 90% overlap (Additional file 4), further showed that the MCAST tool performed the best at 90% overlap, following FIMO and MOODS tools. Consistent with the findings at 100% and 90% overlaps, it was revealed that the MCAST tool exhibited the best performance across all statistical parameters at 80% overlap. Additionally, the FIMO and MOODS tools were observed to secure the second-best and third-best rankings, respectively at 80% overlap (Additional file 4).

Furthermore, each data type of the benchmark dataset was analyzed independently, irrespective of species, to assess the performance of TFBS prediction tools with an 80% overlap between the length of predicted and known binding sites. DWT-toolbox and MotEvo identified the maximum nTPs in all three datasets. In addition to these tools, HOMER and PWMScan identified all 20 binding sites present in the real dataset (Table 4). However, HOMER resulted in a considerably higher number of FPs across all data types. Furthermore, TFBStools resulted in a notably lower number of TPs across all data types.

When assessing the Sn of tools in predicting binding sites, it was observed that generally, all tools except TFBStools exhibited Sn greater than 0.500 across all data types (Fig. 7a). Among these, two tools (DWT-toolbox and MotEvo) exhibited maximum Sn across all three data types. Based on the results obtained for the PPVs of the tools, it was observed that only the MCAST tool consistently performed well across all three data types (Fig. 7b). All tools, except HOMER and Morpheus, exhibited PPVs greater

Table 4 Number of true positives (nTPs), false negatives (nFNs), and false positives (nFPs) detected by each TFBS prediction tool, categorized by data type with an 80% overlap

Tool	Data type								
	Real			Generic			Markov		
	nTPs	nFNs	nFPs	nTPs	nFNs	nFPs	nTPs	nFNs	nFPs
Ciiider	15	5	4	14	6	0	14	6	1
DWT toolbox	20	0	10	20	0	4	20	0	11
FIMO	16	4	4	17	3	0	17	3	0
HOMER	20	0	267	19	1	262	19	1	264
INSECT 2.0	15	5	4	15	5	0	15	5	0
Matrix scan	16	4	4	17	3	0	17	3	0
MCAST	17	3	0	17	3	0	17	3	0
MOODS	17	3	4	18	2	0	17	3	0
Morpheus	18	2	33	18	2	30	18	2	29
MotEvo	20	0	4	20	0	2	20	0	2
PWMScan	20	0	3	19	1	1	18	2	3
TFBStools	3	17	1	3	17	0	3	17	0

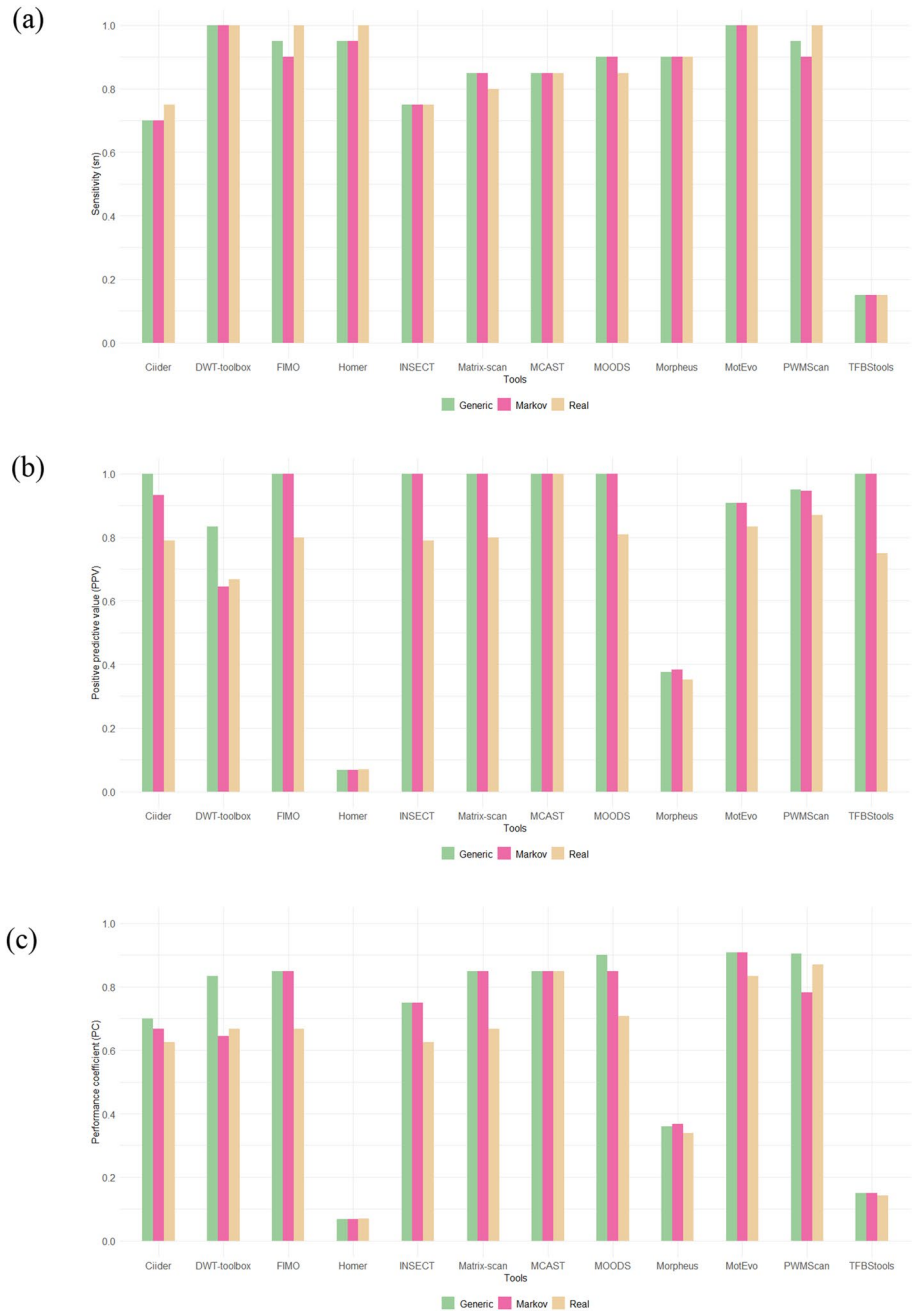

Fig. 7 Tool performance matrices across different data types. **a** Sensitivity, **b** Positive predictive values, and **c** Performance coefficient

than 0.500 for all data types. Similarly, the comparison of PC values of the tools across the three data types revealed that the MCAST tool performed equally well across all data types (Fig. 7c). In addition, MotEvo showed greater performance across all data types. It was also observed that the TFBStools, Morpheus, and HOMER tools exhibited comparatively lower performance (PC < 0.500).

Table 5 Number of true positives (nTPs), false negatives (nFNs), and false positives (nFPs) detected by de novo motif discovery tools at 70% overlap

Tool	nTPs	nFNs	nFPs
MEME	59	01	20
MotifSampler	39	21	18
STREME	39	21	14
rGADEM	19	41	08

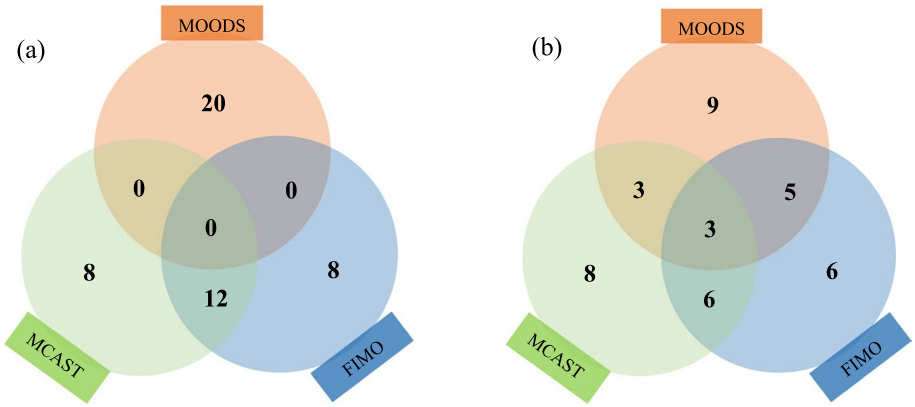

Fig. 8 Distribution of the top 20 motifs predicted by MCAST, FIMO, and MOODS tools for randomly selected promoter sequences. **a** *Arabidopsis thaliana*, **b** *Homo sapiens*

Performance evaluation of de novo *motif discovery tools*

The results of our independent assessment of four commonly used de novo motif discovery tools provide valuable insights into their performance. The evaluation was conducted as described by Jayaram et al. [14], using a 70% overlap between the known binding site and the binding site identified by the tool. Among the tools assessed, MEME emerged as the top performer, identifying 59 binding sites of the 60 binding sites in the benchmark dataset. MotifSampler and STREME both identified 39 binding sites, indicating a comparable level of performance between these two tools (Table 5). When evaluating the tools based on Sn, PC, ASP, and ACCg values, MEME consistently demonstrated the highest performance across these metrics.

Evaluation of the best-performing tools

To further evaluate the performance of the top three TFBS prediction tools identified in the present study—MCAST, FIMO, and MOODS, a comparative analysis was conducted using the top 20 motifs identified from randomly selected promoter regions of *A. thaliana* and *H. sapiens*. In the *A. thaliana* sequences, none of the predicted motifs were shared across the three tools (Fig. 8a). However, 12 motifs were commonly identified by both MCAST and FIMO, while MOODS identified a distinct set of motifs within its top 20 results (Fig. 8a). In the human promoter sequences, three motifs were predicted by all three tools. Additionally, six motifs were commonly identified by both MCAST and FIMO, while five motifs were shared between FIMO and MOODS (Fig. 8b).

Case study 1: assessing transcription factor binding sites in the promoters of the genes involved in the anthocyanin biosynthesis pathway

The anthocyanin biosynthesis is one of the most extensively studied pathways in plants, making it an ideal target for analyzing TFBSs in the promoter regions of related genes. In this study, we analyzed the promoter regions of 149 *A. thaliana*, 129 *G. max*, 74 *O. sativa*, and 70 *Z. mays* genes associated with anthocyanin biosynthesis using the three best-performing TFBS prediction tools: MCAST, FIMO, and MOODS. The analysis revealed that none of the top 20 motifs were shared across all three tools in any of the species analyzed. However, a notable overlap was observed between the motifs identified by MCAST and FIMO. Specifically, these two tools shared eight motifs in *A. thaliana*, ten motifs in *G. max*, ten motifs in *O. sativa*, and eight motifs in *Z. mays*, respectively (Additional file 5).

MYB (v-Myb myeloblastosis viral oncogene homolog) TFs are key regulators of anthocyanin biosynthesis in plants [108–112]. In our analysis, MCAST identified only a few MYB binding sites among the top 20 motifs for *A. thaliana* (10%), *G. max* (5%), and *O. sativa* (5%) (Additional file 5). In contrast, neither FIMO nor MOODS detected any MYB motifs in the top 20 motifs for these species. Moreover, in the promoter sequences of *Z. mays*, none of the tools detected MYB or MYB-related motifs among the top 20 motifs. As the number of analyzed motifs increased, the detection of MYB and/or MYB-related motifs increased. In *A. thaliana*, both FIMO and MOODS began detecting MYB and/or MYB-related bindings sites within the top 40 motifs (Fig. 9a). Similarly, in *G. max*, FIMO detected MYB motifs within the top 60 motifs, while MOODS detected MYB-related motifs starting at the top 40 (Fig. 9b). For *O. sativa* sequences, MOODS failed to identify any MYB binding sites even within the top 100 motifs (Fig. 9c). However, for *Z. mays*, MYB binding sites were detected by MCAST, FIMO and MOODS starting from the top 40, 80 and 60, respectively (Fig. 9d).

Notably, in the promoter regions of *A. thaliana* and *G. max*, MCAST primarily identified cystein-rich polycomb-like protein (CPP) motifs as the most abundant among the top 20 motifs, followed by DNA-binding with one finger (DOF) motifs. In contrast, FIMO predicted DOF motifs as the most abundant, followed by CPP motifs (Additional file 6). Furthermore, all of the top 20 motifs identified by MOODS in both species were DOF binding sites (Additional file 6). In the monocot species *O. sativa* and *Z. mays*, MCAST and FIMO primarily detected dehydration-responsive element-binding protein (DREB) binding sites within the top 20 motifs, while MOODS predominately identified DOF binding sites, followed by DREB binding sites (Additional file 6).

These findings highlight the variability in motif detection across different tools, even when analyzing the same gene list within a species. This underscores the importance of carefully selecting the appropriate tools for TFBS discovery.

Case study 2: assessing transcription factor binding sites in the promoters of the genes involved in the pentose phosphate pathway in Homo sapiens

The pentose phosphate pathway (PPP) is a fundamental component of cellular metabolism and plays a significant role in various human diseases [113]. An analysis of the promoter regions of 31 genes involved in the PPP using the TFBS prediction tools MCAST,

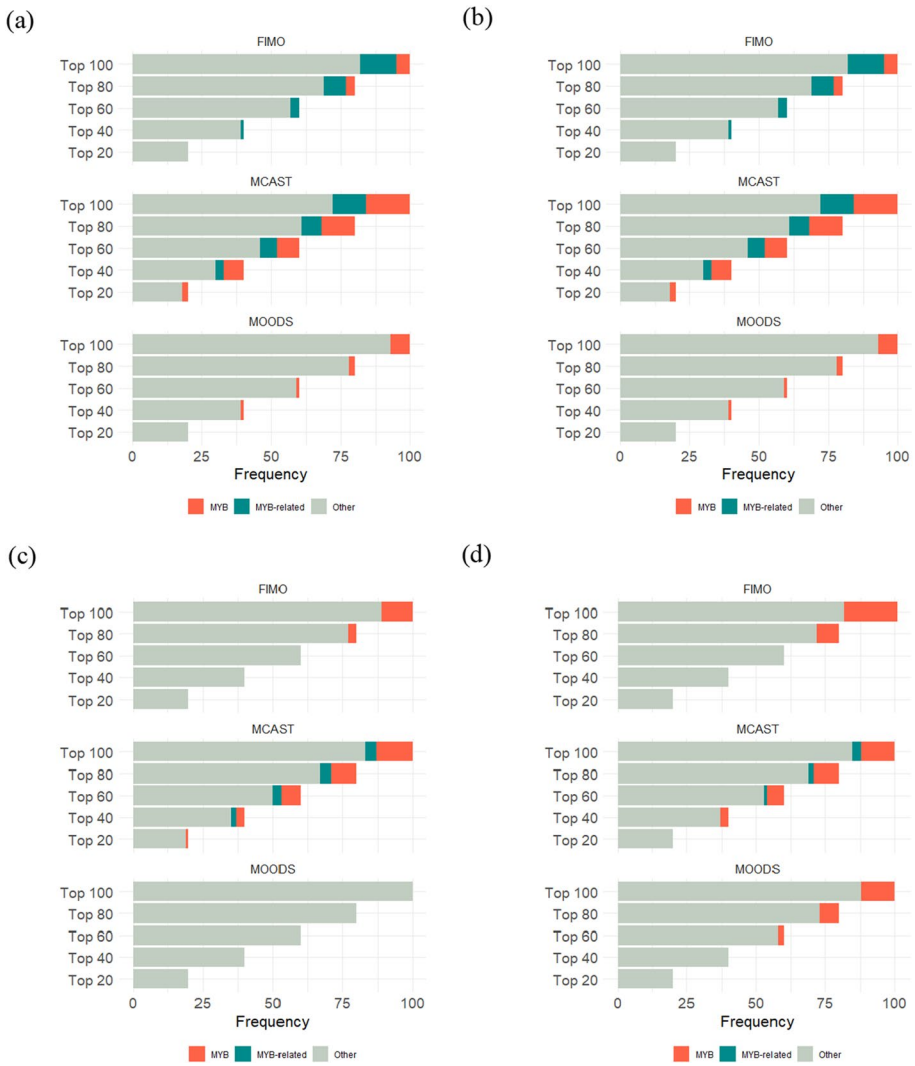

Fig. 9 Frequency of motifs detected in promoter sequences of genes involved in the anthocyanin biosynthesis pathway in plants. **a** *Arabidopsis thaliana*, **b** *Glycine max*, **c** *Oryza sativa*, **d** *Zea mays*

FIMO, and MOODS revealed that six motifs were consistently detected among the top 20 motifs by all three tools (Fig. 10). Additionally, FIMO and MOODS shared eight motifs, while MCAST and FIMO shared five motifs, highlighting both the overlap and variation in motif detection across these tools. Notably, most of the top 20 motifs identified by all three tools corresponded to binding sites for TFs belonging to the C2H2 zinc finger family (C2H2-ZF), highlighting the significance of this TF family in PPP regulation (Additional file 6).

Discussion

Advances in bioinformatics have revolutionized biological research by providing scientists with a range of computer tools for retrieval, analysis, and visualization of omics data, thereby uncovering the complexities of living organisms. However, this also presents a challenge for researchers, as the findings and conclusions may

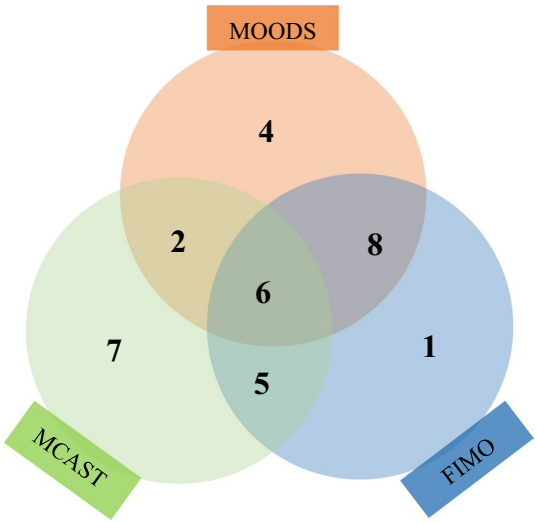

Fig. 10 Distribution of the top 20 motifs identified by MCAST, FIMO, and MOODS across the promoter regions of the genes involved in the pentose phosphate pathway in *Homo sapiens*

significantly depend on the choice of tools used for the analysis. To address this, benchmarking studies are often conducted to rigorously compare the performance of different tools, identify their strengths, and provide recommendations on the most suitable tools for specific analyses [114]. Designing a robust performance evaluation of computational tools is challenging due to various factors, such as selecting appropriate benchmark datasets, determining optimal tool parameters, and choosing suitable statistical measures for performance evaluation. Therefore, it is essential to design benchmarking studies carefully to ensure accurate, unbiased, and informative results [114].

To date, only a limited number of assessments have been conducted to evaluate the performance quality of TFBS prediction tools. For instance, Tompa et al. [77], provide a comparative analysis of several TFBS discovery tools, including AlignACE, ANN-Spec, Consensus, GLAM, Improbizer, MEME, MITRA, MotifSampler, oligo/dyad-analysis, QuickScore, SesiMCMC, Weeder, and YMF. Additionally, Jayaram et al. [14], evaluate a set of TFBS prediction tools (i.e., MCast, Baycis, Cister, Cluster-Buster, Comet, FIMO, Clover, Matrix-Scan, Patser and Possum-Search), as well as de novo motif discovery tools (i.e., MEME-ChIP, HOMER, ChIPMunk and rGADEM). Among the 16 tools assessed in the present study, the performance of INSECT 2.0, Morpheus, Ciiider, PWMScan, MOODS, MotEvo, DWT-toolbox, and TFBStool has not been previously evaluated. Our study was conducted from a general user's perspective, employing default settings wherever possible or the recommended settings provided in the corresponding user manuals. If no recommendations were available, settings were chosen based on personal preference.

The selection of the reference dataset plays a crucial role in benchmarking computational methods [114]. Previous studies evaluating TFBS prediction tools often relied on reference datasets consisting of preprocessed ChIP-seq peaks or annotated TFBSs from the TRANSFAC database (14, 77, 81]. However, in our study, we opted to utilize

the information stored in the JASPAR database which provides annotated TFBSs alongside sequences obtained from ChIP-seq experiments and is freely accessible, unlike TRANSFAC. To construct the reference or benchmark dataset, we incorporated four types of data: real, generic, Markov, and negative sequences. This approach aimed to minimize biases that could potentially favor certain tools over others [77, 114].

Establishing criteria for defining an actual binding site (true positive) is crucial for comparing different TFBS or motif prediction tools. Employing a 100% overlap percentage between the length of actual and predicted binding sites to classify them as TPs may introduce bias, given the difficulty in precisely predicting TFBSs by many tools. In the study conducted by Jayaram et al. [14], an overlap percentage of 70% was used, while Tompa et al. [77], considered an overlap by at least one-quarter of the length of the known site as a TP. Consequently, in this study, different overlap percentages were applied: 100%, 90%, and 80% for TFBS prediction tools. No significant changes in results were observed beyond an 80% overlap, and thus, lower overlap percentages were not considered. Additionally, for de novo motif discovery tools, a 70% overlap was employed, following the approach described by Jayaram et al. [14].

Among the tools evaluated, MCAST stands out for its unique ability to avoid FPs across all three overlap percentages. This remarkable performance can be attributed to its algorithm, which is based on HMMs. MCAST is specifically designed to identify clusters of non-overlapping motifs, commonly referred to as *cis*-regulatory modules (CRMs) [65]. By focusing on clusters of motifs rather than individual occurrences, MCAST effectively reduces the likelihood of FPs, enhancing the accuracy of motif detection. However, while MCAST excels in minimizing FPs, its sensitivity was below that of many of the tools including FIMO and MOODS. Sensitivity is essential for ensuring that true binding sites are not overlooked. In applications where detecting all possible binding sites is crucial, this limitation of MCAST should be taken into account. Lower sensitivity means that some true binding sites may not be identified, which can be a significant drawback in comprehensive motif discovery studies.

The findings of the present study align with those of Jayaram et al. [14], where MCAST and FIMO emerged as the best best-performing tools. However, in our study, MOODS outperformed FIMO at a 100% overlap percentage, while FIMO surpassed MOODS at the other two overlap percentages. Both of these tools are designed to identify individual binding sites, unlike MCAST. FIMO offers both web-based and command-line interfaces, with the web tool being particularly user-friendly [30]. In contrast, MOODS operates solely via the command line on Linux systems, though it boasts a faster runtime compared to FIMO [32].

Comparing the performance quality of TFBS prediction tools poses a challenge due to their reliance on different algorithms for motif discovery. Our benchmark dataset included only one binding site per sequence and therefore, was not affected by these algorithmic differences. However, in real datasets, multiple binding sites may be present in a given DNA sequence. Hence, it is more reliable to conclude that for an average-end user aiming to identify CRMs with minimal FPs, MCAST stands out as the best option. Conversely, for identifying individual binding sites, FIMO and MOODS

emerge as the best options. The choice between FIMO and MOODS may depend on factors such as run-time, user-friendliness, and the expected level of accuracy.

Although MCAST, FIMO, and MOODS outperformed other tools overall, DWT-toolbox and MotEvo excelled in identifying the highest number of TPs at both 80% and 90% overlaps. DWT-toolbox, a command line tool employing a Bayesian network model, effectively addresses a critical issue in TFBS prediction tools: pairwise dependency [15, 115]. Unlike the PWM method used in most tools, which assumes independence between positions, the DWT-toolbox considers all possible pairwise dependencies within a rigourous probabilistic framework devoid of tunable parameters, automatically avoiding overfitting [106]. MotEvo, another command line tool, employs a Bayesian probabilistic model-based algorithm specifically designed for analyzing TFBSs across phylogenetically related multiple organisms [68], but can also handle single species sequences. However, considering its file requirements and parameters, it is conceivable that MotEvo could have performed even better with phylogenetically related sequences.

Consistent with previous findings [77], our evaluation of TFBS prediction tools across three different data types: real, generic, and Markov, further underscores the critical role of experimental design in performance evaluation research. Notably, a considerable increase in the number of FPs was observed in the real dataset compared to the other data types. This suggests the presence of putative binding sites for other TFs beyond those known and utilized in the current experimental design. Consequently, tools that successfully identify these additional binding sites may be unfairly penalized due to the experimental setup. To mitigate this bias, generic and Markov datasets were included in the benchmark, ensuring a more balanced evaluation across different tools. MCAST demonstrated consistent performance across all three data types and all statistical parameters assessed, reinforcing its reliability as a TFBS prediction tool.

When evaluating the overall performance of TFBS prediction tools across various overlap percentages and data types, most tools demonstrated satisfactory performance, with some emerging as the best performers. However, both HOMER and TFBStools exhibited notably poor performances. HOMER, a command-line tool, ranked among the lowest performers due to its tendency to identify a high number of FPs, despite its ability to accurately recognize TPs. This high FP rate reduces its overall reliability, highlighting the need for more precise algorithms or improved parameter settings. Furthermore, TFBStools, an R package, performed poorly in identifying TPs in our study, especially at 100% and 90% overlaps, with only a few TPs detected at 80% overlap. While TFBStools did not generate many FPs, its limited ability to predict most binding sites, even after parameter adjustments, raises concerns regarding its effectiveness as a TFBS prediction tool. This limited predictive capability suggests that TFBStools may require significant algorithmic improvements or more sophisticated parameter tuning to enhance its performance.

Among the de novo motif discovery tools examined, the comprehensive dominance of MEME underscores its reliability and efficiency in accurately identifying binding motifs. MEME directly identifies motifs within user input sequences, making it highly effective for straightforward motif discovery tasks. Although both MEME and STREME are part of the MEME suite, they utilize different algorithms. STREME compares user sequences with a set of control sequences to detect relatively enriched motifs [53]. This

methodological distinction suggests that STREME may offer superior performance for certain types of analyses, particularly those involving differential motif enrichment. MotifSampler's performance, while notable, may be impacted by a lack of updates and maintenance. This could affect its usability and reliability. Furthermore, our findings present a contrast to the results reported by Jayaram et al. [14], where rGADEM was shown to outperform tools such as ChIPMunk, HOMER, and MEME-ChIP. In our study, rGADEM exhibited the poorest performance across most metrics, except for PPV. This discrepancy could be attributed to the parameter settings used during the motif discovery process. While Jayaram et al. [14], manually optimized rGADEM's parameters to achieve better results, our study utilized the default settings. This highlights the importance of fine-tuning parameters for motif discovery tools to maximize their efficacy. Default settings, while convenient, may not always capture the nuances required for accurate motif identification. Users of rGADEM and similar tools should consider customizing parameters based on the specific characteristics of their datasets to improve performance outcomes.

To further evaluate the performance of the best-performing tools—MCAST, FIMO, and MOODS, we focused on two distinct biological pathways: the anthocyanin biosynthesis pathway in plants and the PPP in humans. In addition, we analyzed a set of random promoter sequences from *A. thaliana* and *H. sapiens* genomes to assess the consistency in motif detection across the tools. Overall, our comparative analysis revealed considerable variability in motif detection across the tools, highlighting the importance of selecting the appropriate TFBS prediction tool(s) for different biological contexts to enhance prediction accuracy.

In plants, MYB TFs are one of the key regulators of anthocyanin biosynthesis [110, 116]. Our analysis revealed inconsistencies in detecting MYB binding sites among different tools when analysing the same set of promoter sequences and focusing on the top 20 motifs. Specifically, MCAST identified MYB motifs within the top 20 motifs in the promoters of anthocyanin biosynthesis-related genes in *A. thaliana*, *G. max*, and *O. sativa*. In contrast, neither FIMO nor MOODS detected MYB motifs within this range for these species. This suggests that MCAST may be more sensitive in detecting motifs based on their frequency within the top 20 motifs. Nevertheless, only a few MYB motifs were identified among the top 20 motifs: MYB23 and MYB58 binding sites were detected in *A. thaliana*, and MYB31 binding sites were identified in both *G. max* and *O. sativa*. As this analysis was expanded to include the top 40–100 motifs in increments of 20, both FIMO and MOODS began to detect MYB motifs, indicating that these tools may perform better with a broader search range.

A recent study analyzing TFBSs in Arabidopsis anthocyanin biosynthesis-related genes using FIMO, sorted by a user-defined Significance Score (S-Score), reported a prevalence of MYB TFs within the top 20 motifs [41]. In contrast, our present study, which also analyzed the same Arabidopsis anthocyanin gene list using FIMO but sorted based on motif frequencies, did not detect any MYB binding sites within the top 20 motifs. However, binding sites for MYB51, MYB80, and MYB93, as identified in Grau and Franco-Zorrilla [41], were detected among the top 100 motifs in our analysis. This discrepancy in motif detection may be due to differences in the TFBS sorting criteria, and/or the set of PWMs employed during promoter scanning. For instance, in the

present analysis, we utilized PWMs from the JASPAR database, which do not include certain PWMs such as MYB6, MYB37, and MYB38, which were considered in the analysis by Grau and Franco-Zorrilla [41].

Furthermore, the PWMs for several important MYB TFs that regulate anthocyanin biosynthesis in plants, specifically, MYB11, MYB12, MYB75, MYB90, and MYB114 [117–119] were not included in either the JASPAR database or the study by Grau and Franco-Zorrilla [41]. This could limit the detection of important binding sites in the anthocyanin biosynthesis-related genes of the plant species if those binding sites are present. Therefore, as new TFs are identified and characterized, it is essential to continuously update freely accessible databases like JASPAR to enhance the accuracy of TFBS detection.

In addition to the MYB TF family, several other TF families are known to be involved in anthocyanin biosynthesis [120–122]. However, the role of DREB TFs in this pathway has not been extensively studied. Interestingly, our analysis revealed a higher prevalence of DREB binding sites in the anthocyanin biosynthesis-related genes of monocotyledon species such as *O. sativa* and *Z. mays*, as detected by MCAST and FIMO. This finding suggests that DREB TFs may play a regulatory role in the anthocyanin pathway in monocots. Therefore, future studies exploring the potential involvement of DREB TFs in mediating anthocyanin biosynthesis in monocots could provide valuable insights into the regulation of this important secondary metabolic pathway in plants. Moreover, we observed a predominant presence of C2H2-ZF family binding sites in the genes of the human PPP. This family of TFs represents the largest class of multifunctional regulators in both humans and plants [123, 124]. However, there is currently no information on the involvement of these TFs in the human PPP, opening avenues for future research.

Conclusion

As various TFBS prediction tools become available, understanding the performance of different tools in terms of accuracy, sensitivity, and specificity is crucial in genomic research. This study identified MCAST, FIMO, and MOODS as the top-performing TFBS prediction tools, while DWT-toolbox and MotEvo demonstrated increased efficacy in predicting a higher number of true binding sites, indicating superior sensitivity. MEME emerged as the most reliable de novo motif discovery tool.

The observed performance variations in this study emphasize the necessity for continuous benchmarking and independent assessments of TFBS prediction tools. Such evaluations help to identify the strengths and weaknesses of tools, guiding users in selecting the most appropriate tool(s) for their specific research needs. Moreover, tool developers can leverage these insights to enhance algorithm robustness and usability, ensuring that default settings are more universally effective or providing clear guidance on parameter customization. Further studies should explore the impact of parameter optimization across different datasets to provide a more comprehensive understanding of each tool's capabilities and limitations.

It is recommended to further evaluate the performance quality of TFBS predicting tools using diverse benchmark datasets derived from multiple organisms, containing a range of TFBSs. Furthermore, relying solely on a single type of binding site per sequence, as done in the present study, may not provide a comprehensive evaluation of a tool's

performance. Utilizing benchmark datasets from various organisms with multiple binding sites per sequence can provide valuable insights into the tool's performance across a broader spectrum of genomic contexts, better reflecting the complexity of regulatory regions in biological systems. Additionally, this approach will enhance the accuracy and reliability of the evaluations.

In summary, due to variations in sensitivity and specificity among TFBS prediction tools, it is recommended to integrate predictions from multiple tools to mitigate the limitations of individual tools, leading to a more thorough analysis of TFBSs. Thus, the development of a comprehensive toolbox that integrates TFBS prediction tools and de novo motif discovery tools would enable users to utilize a diverse array of tools, thereby increasing the likelihood of identifying significant binding sites/motifs. This integrative approach would facilitate a more robust and detailed understanding of transcriptional regulatory mechanisms, ultimately advancing genomic research.

Abbreviations

ACCg	Geometric accuracy
ASP	Average site performance
C2H2-ZF	C2H2 zinc finger
ChIP	Chromatin immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CPP	Cystein-rich polycomb-like protein
CRM	*Cis*-Regulatory module
DOF	DNA binding with one finger
DREB	Dehydration-responsive element-binding protein
DWT	Dinucleotide weigh sensor
EPD	Eukaryotic promoter database
FIMO	Find individual motif occurrences
FNs	False negatives
FPs	False positives
HMM	Hidden Markov model
HOMER	Hypergeometric optimization of motif enrichment
MCAST	Motif cluster alignment and search tool
MEME	Multiple Em for Motif Elicitation
MOODS	MOotif occurrence detection suite
MYB	V-Myb myeloblastosis viral oncogene homolog
nFNs	Number of false negatives
nFPs	Number of false positives
nTPs	Number of true positives
PC	Performance coefficient
PFM	Position frequency matrix
PPP	Pentose phosphate pathway
PPV	Positive predictive value
PWM	Position weight matrix
Sn	Sensitivity
TFBS	Transcription factor binding site
TF	Transcription factor
TPs	True positives
TSS	Transcription start site

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05995-0.

Additional file1 (TXT 11 KB)

Additional file2 (XLSX 60 KB)

Additional file3 (XLSX 25 KB)

Additional file4 (XLSX 14 KB)

Additional file5 (PDF 205 KB)

Additional file6 (XLSX 30 KB)

Acknowledgements
Not applicable

Author contributions
A.M.W. and S.V. conceived the study, participated in its design, and assisted in drafting the manuscript. D.V.W. contributed to the study design, performed analysis and interpreted data, and drafted the manuscript. All authors read and approved the final manuscript.

Funding
Not applicable.

Data availability
All data generated or analysed during this study are included in this published article [and its supplementary information files]. No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 23 June 2024 Accepted: 21 November 2024
Published online: 02 December 2024

References

1. Shen WK, Chen SY, Gan ZQ, et al. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. Nucleic Acids Res. 2023;51(D1):D39–45.
2. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2017;45(D1):D1040–5.
3. Guo A, He K, Liu D, Bai S, Gu X, Wei L, Luo J. DATF: a database of Arabidopsis transcription factors. Bioinformatics. 2005;21(10):2568–9.
4. Khamis AM, Motwalli O, Oliva R, Jankovic BR, Medvedeva YA, Ashoor H, Essack M, Gao X, Bajic VB. A novel method for improved accuracy of transcription factor binding site prediction. Nucleic Acids Res. 2018;46(12): e72.
5. Lai X, Stigliani A, Vachon G, Carles C, Smaczniak C, Zubieta C, Kaufmann K, Parcy F. Building transcription factor binding site models to understand gene regulation in plants. Mol Plant. 2019;12(6):743–63.
6. Narlikar L, Ovcharenko I. Identifying regulatory elements in eukaryotic genomes. Brief Funct Genom Proteom. 2009;8(4):215–30.
7. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018;19(10):621–37.
8. Reid JE, Evans KJ, Dyer N, Wernisch L, Ott S. Variable structure motifs for transcription factor binding sites. BMC Genom. 2010;11:30.
9. Boeva V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. Front Genet. 2016;7:24.
10. Garner MM, Revzin A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. Nucleic Acids Res. 1981. https://doi.org/10.1093/nar/9.13.3047.
11. Galas DJ, Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 1978;5(9):3157–70.
12. Riley TR, Slattery M, Abe N, Rastogi C, Liu D, Mann RS, Bussemaker HJ. SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. Methods Mol Biol. 2014;1196:255–78.
13. Elnitski L, Jin VX, Farnham PJ, Jones SJ. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res. 2006;16(12):1455–64.
14. Jayaram N, Usvyat D, Martin ACR. Evaluating tools for transcription factor binding site prediction. BMC Bioinform. 2016;17(1):547.
15. Zeng Y, Gong M, Lin M, Gao D, Zhang Y. A review about transcription factor binding sites prediction based on deep learning. IEEE Access. 2020;8:219256–74.
16. Zhang S, Ma A, Zhao J, Xu D, Ma Q, Wang Y. Assessing deep learning methods in *cis*-regulatory motif finding based on genomic sequencing data. Brief Bioinform. 2022;23(1):bbab374.
17. Zogopoulos VL, Spaho K, Ntouka C, Lappas GA, Kyranis I, Bagos PG, Spandidos DA, Michalopoulos I. TFBSPred: a functional transcription factor binding site prediction webtool for humans and mice. Int J Epigen. 2021;1:9.
18. Stormo GD. Modeling the specificity of protein-DNA interactions. Quant Biol. 2013;1(2):115–30.
19. Ali O, Farooq A, Yang M, Jin VX, Bjørås M, Wang J. abc4pwm: affinity based clustering for position weight matrices in applications of DNA sequence analysis. BMC Bioinform. 2022;23(1):83.

20. Gershenzon NI, Stormo GD, Ioshikhes IP. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. Nucleic Acids Res. 2005;33(7):2290–301.
21. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000;16(1):16–23.
22. Stormo GD, Zhao Y. Determining the specificity of protein-DNA interactions. Nat Rev Genet. 2010;11(11):751–60.
23. Mathelier A, Wasserman WW. The next generation of transcription factor binding site prediction. PLoS Comput Biol. 2013;9(9): e1003214.
24. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon JA, Ferenc K, Kumar V, et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2024;52(D1):D174–82.
25. Wingender E, Dietze P, Karas H, Knüppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. 1996;24(1):238–41.
26. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018;46(D1):D252–9.
27. Puig RR, Boddie P, Khan A, Castro-Mondragon JA, Mathelier A. UniBind: maps of high-confidence direct TF-DNA interactions across nine species. BMC Genom. 2022;22:482.
28. Newburger DE, Bulyk ML. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res. 2009;37:D77–82.
29. Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D, et al. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res. 2013;41:D171–6.
30. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017–8.
31. Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 2003;31(13):3576–9.
32. Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics. 2009;25(23):3181–2.
33. Gearing LJ, Cumming HE, Chapman R, Finkel AM, Woodhouse IB, Luu K, Gould JA, Forster SC, Hertzog PJ. CiiiDER: a tool for predicting and analysing transcription factor binding sites. PLoS ONE. 2019;14(9): e0215495.
34. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576–89.
35. Parra RG, Rohr CO, Koile D, Perez-Castro C, Yankilevich P. INSECT 2.0: a web-server for genome-wide *cis*-regulatory modules prediction. Bioinformatics. 2016;32(8):1229–31.
36. Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J. Using RSAT to scan genome sequences for transcription factor binding sites and *cis*-regulatory modules. Nat Protoc. 2008;3(10):1578–88.
37. Minguet EG, Segard S, Charavay C, Parcy F. MORPHEUS, a webtool for transcription factor binding analysis using position weight matrices with dependency. PLoS ONE. 2015;10(8): e0135586.
38. Ambrosini G, Groux R, Bucher P. PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. Bioinformatics. 2018;34(14):2483–4.
39. Tan G, Lenhard B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinformatics. 2016;32(10):1555–6.
40. Kreft L, Soete A, Hulpiau P, Botzki A, Saeys Y, De Bleser P. ConTra v3: a tool to identify transcription factor binding sites across species, update 2017. Nucleic Acids Res. 2017;45(W1):W490–4.
41. Grau J, Franco-Zorrilla JM. TDTHub, a web server tool for the analysis of transcription factor binding sites in plants. Plant J. 2022;111(4):1203–15.
42. Tsunoda T, Takagi T. Estimating transcription factor bindability on DNA. Bioinformatics. 1999;15(7):622–30.
43. Messeguer X, Escudero R, Farré D, Núñez O, Martínez J, Albà MM. PROMO: detection of known transcription regulatory elements using species-tailored searches. Bioinformatics. 2002;18(2):333–4.
44. Frith MC, Li MC, Weng Z. Cluster-Buster: finding dense clusters of motifs in DNA sequences. Nucleic Acids Res. 2003;31(13):3666–8.
45. Loots GG, Ovcharenko I. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Res. 2004;32:W217–21.
46. Lee C, Huang CH. LASAGNA-Search: an integrated web tool for transcription factor binding site search and visualization. Biotechniques. 2013;54(3):141–53.
47. Zambelli F, Pesole G, Pavesi G. PscanChIP: finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. Nucleic Acids Res. 2013;41:W535–43.
48. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. J Mol Biol. 2000;296(5):1205–14.
49. Pavesi G, Mereghetti P, Mauri G, Pesole G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 2004;32:W199–203.
50. Ao W, Gaudet J, Kent WJ, Muttumu S, Mango SE. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. Science. 2004;305(5691):1743–6.
51. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouzé P, Moreau Y. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics. 2001;17(12):1113–22.
52. Mercier E, Droit A, Li L, Robertson G, Zhang X, Gottardo R. An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. PLoS ONE. 2011;6(2): e16432.
53. Bailey TL. STREME: accurate and versatile sequence motif discovery. Bioinformatics. 2021;37(18):2834–40.
54. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 2006;34:W369–73.
55. Mehta P, Schwab DJ, Sengupta AM. Statistical mechanics of transcription-factor binding site discovery using Hidden Markov Models. J Stat Phys. 2011;142(6):1187–205.

56. Rabiner LA. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. 1989;77(2):257–86.
57. Wu J, Xie J. Hidden Markov model and its applications in motif findings. Methods Mol Biol. 2010;620:405–16.
58. Xu D, Liu HJ, Wang YF. BSS-HMM3s: an improved HMM method for identifying transcription factor binding sites. DNA Seq. 2005;16(6):403–11.
59. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011;21(3):447–55.
60. Barash Y, Elidan G, Friedman N, Kaplan T. Modeling dependencies in protein-DNA binding sites. Annu Int Conf Res Comput Mol Biol. 2013. https://doi.org/10.1145/640075.640079.
61. Chen C, Hou J, Shi X, Yang H, Birchler JA, Cheng J. DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks. BMC Bioinform. 2021;22(1):38.
62. Ding P, Wang Y, Zhang X, Gao X, Liu G, Yu B. DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape. Brief Bioinform. 2023;24(4):bbad231.
63. Ghosh N, Santoni D, Saha I, Felici G. Predicting transcription factor binding sites with deep learning. Int J Mol Sci. 2024;25(9):4990.
64. Wang W, Jiao X, Sun B, Liang S, Wang X, Zhou Y. DeepGenBind: a novel deep learning model for predicting transcription factor binding sites. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. 2022;3629–3635.
65. Grant CE, Johnson J, Bailey TL, Noble WS. MCAST: scanning for *cis*-regulatory motif clusters. Bioinformatics. 2016;32(8):1217–9.
66. Steinhaus R, Robinson PN, Seelow D. FABIAN-variant: predicting the effects of DNA variants on transcription factor binding. Nucleic Acids Res. 2022;50(W1):W322–9.
67. Grau J, Ben-Gal I, Posch S, Grosse I. VOMBAT: prediction of transcription factor binding sites using variable order Bayesian trees. Nucleic Acids Res. 2006;34:W529–33.
68. Arnold P, Erb I, Pachkov M, Molina N, van Nimwegen E. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. Bioinformatics. 2012;28(4):487–94.
69. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8.
70. Hassanzadeh HR, Wang MD. DeeperBind: enhancing prediction of sequence specificities of DNA binding proteins. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. 2016;178–183.
71. Salekin S, Zhang JM, Huang Y. Base-pair resolution detection of transcription factor binding site by deep deconvolutional network. Bioinformatics. 2018;34(20):3446–53.
72. Yang J, Ma A, Hoppe AD, Wang C, Li Y, Zhang C, Wang Y, Liu B, Ma Q. Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. Nucleic Acids Res. 2019;47(15):7809–24.
73. Han K, Shen LC, Zhu YH, Xu J, Song J, Yu DJ. MAResNet: predicting transcription factor binding sites by combining multi-scale bottom-up and top-down attention and residual network. Brief Bioinform. 2022;23(1):bbab445.
74. Yang T, Henao R. TAMC: a deep-learning approach to predict motif-centric transcriptional factor binding activity based on ATAC-seq profile. PLoS Comput Biol. 2022;18(9): e1009921.
75. Qin Q, Feng J. Imputation for transcription factor binding predictions based on deep learning. PLoS Comput Biol. 2017;13(2): e1005403.
76. Roulet E, Fisch I, Junier T, Bucher P, Mermod N. Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. Silico Biol. 1998;1(1):21–8.
77. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol. 2005;23(1):137–44.
78. Hu J, Li B, Kihara D. Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res. 2005;33(15):4899–913.
79. Medina-Rivera A, Abreu-Goodger C, Thomas-Chollier M, Salgado H, Collado-Vides J, van Helden J. Theoretical and empirical quality assessment of transcription factor-binding motifs. Nucleic Acids Res. 2011;39(3):808–24.
80. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, Saez-Rodriguez J, Cokelaer T, et al. Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol. 2013;31(2):126–34.
81. Tran NT, Huang CH. A survey of motif finding web tools for detecting binding site motifs in ChIP-Seq data. Biol Direct. 2014;9(4):1–22.
82. JASPAR Database. https://jaspar.elixir.no/. Accessed 23rd August 2023
83. GENCODE. https://www.gencodegenes.org/. Accessed 25th August 2023
84. NCBI. https://www.ncbi.nlm.nih.gov/. Accessed 25th August 2023
85. Ensembl genome browser. https://asia.ensembl.org/index.html. Accessed 27th August 2023
86. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017;18(5):851–69.
87. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12–8.
88. Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JWR, Fajardo-Flores SB, et al. A review of deep learning applications for genomic selection. BMC Genom. 2021;22(1):19.
89. Ciiider. https://gitlab.erc.monash.edu.au/ciiid/ciiider. Accessed 4th December 2023
90. DWT-toolbox. https://swissregulon.unibas.ch/sr/software. Accessed 15th January 2024
91. FIMO. https://meme-suite.org/meme/tools/fimo. Accessed 30th October 2023
92. HOMER. http://homer.ucsd.edu/homer/introduction/install.html. Accessed 12th December 2023
93. INSECT 2.0. http://bioinformatics.ibioba-mpsp-conicet.gov.ar/INSECT2/index.php. Accessed 2nd November 2023
94. Matrix-scan. https://rsat01.biologie.ens.fr/rsat/matrix-scan_form.cgi. Accessed 27th November 2023
95. MCAST. https://meme-suite.org/meme/tools/mcast. Accessed 12th January 2024
96. MOODS. https://www.cs.helsinki.fi/group/pssmfind/. Accessed 7th December 2023
97. Morpheus. http://biodev.cea.fr/morpheus/. Accessed 29th November 2023
98. MoteEvo. https://swissregulon.unibas.ch/sr/software. Accessed 6th January 2024

99. PWMScan. https://sourceforge.net/projects/pwmscan/. Accessed 30th November 2023
100. TFBStools. https://bioconductor.org/packages/release/bioc/html/TFBSTools.html. Accessed 20th December 2023
101. MEME. https://meme-suite.org/meme/tools/meme. Accessed 20th January 2024
102. MotifSampler. https://homes.esat.kuleuven.be/~sistawww/bioi/thijs/Work/MotifSampler.html Accessed on 27th January 2024
103. rGADEM. https://bioconductor.org/packages/release/bioc/html/rGADEM.html. Accessed 13th February 2024
104. STREME. https://meme-suite.org/meme/tools/streme. Accessed 2nd February 2024
105. Eukaryotic Promoter Database. https://epd.expasy.org/epd/. Accessed on 19th February
106. Omidi S, Zavolan M, Pachkov M, Breda J, Berger S, van Nimwegen E. Automated incorporation of pairwise dependency in transcription factor binding site prediction using dinucleotide weight tensors. PLoS Comput Biol. 2017;13(7): e1005176.
107. Périer RC, Praz V, Junier T, Bonnard C, Bucher P. The eukaryotic promoter database (EPD). Nucleic Acids Res. 2000;28(1):302–3.
108. Yan H, Pei X, Zhang H, Li X, Zhang X, Zhao M, Chiang VL, Sederoff RR, Zhao X. MYB-mediated regulation of anthocyanin biosynthesis. Int J Mol Sci. 2022;22(6):3103.
109. He G, Zhang R, Jiang S, Wang H, Ming F. The MYB transcription factor RcMYB1 plays a central role in rose anthocyanin biosynthesis. Hortic Res. 2023;10(6):uhad080.
110. Li C, Yu W, Xu J, Lu X, Liu Y. Anthocyanin biosynthesis induced by MYB transcription factors in plants. Int J Mol Sci. 2022;23(19):11701.
111. Busche M, Pucker B, Weisshaar B, Stracke R. Three R2R3-MYB transcription factors from banana (*Musa acuminata*) activate structural anthocyanin biosynthesis genes as part of an MBW complex. BMC Res Notes. 2023;16:103.
112. Menconi J, Perata P, Gonzali S. Novel R2R3 MYB transcription factors regulate anthocyanin synthesis in Aubergine tomato plants. BMC Plant Biol. 2023;23(1):148.
113. Stincone A, Prigione A, Cramer T, Wamelink MM, Campbell K, Cheung E, et al. The return of metabolism: biochemistry and physiology of the pentose phosphate pathway. Biol Rev Camb Philos Soc. 2015;90(3):927–63.
114. Weber LM, Saelens W, Cannoodt R, Soneson C, Hapfelmeier A, Gardner PP, et al. Essential guidelines for computational method benchmarking. Genome Biol. 2019;20:125.
115. Bi Y, Kim H, Gupta R, Davuluri RV. Tree-based position weight matrix approach to model transcription factor binding site profiles. PLoS ONE. 2011;6(9): e24210.
116. Cao Y, Li K, Li Y, Zhao X, Wang L. MYB Transcription factors as regulators of secondary metabolism in plants. Biology (Basel). 2020;9(3):61.
117. Chaves-Silva S, Santos ALD, Chalfun-Júnior A, Zhao J, Peres LEP, Benedito VA. Understanding the genetic regulation of anthocyanin biosynthesis in plants – tools for breeding purple varieties of fruits and vegetables. Phytochemistry. 2018;153:11–27.
118. Liu Y, Tikunov Y, Schouten RE, Marcelis LFM, Visser RGF, Bovy A. Anthocyanin biosynthesis and degradation mechanisms in *Solanaceous* vegetables: a review. Front Chem. 2018;9(6):52.
119. Shi MZ, Xie DY. Biosynthesis and metabolic engineering of anthocyanins in *Arabidopsis thaliana*. Recent Pat Biotechnol. 2014;8(1):47–60.
120. Zhang F, Gonzalez A, Zhao M, Payne CT, Lloyd A. A network of redundant bHLH proteins functions in all TTG1-dependent pathways of Arabidopsis. Development. 2003;130(20):4859–69.
121. Xu W, Grain D, Bobet S, Le Gourrierec J, Thévenin J, Kelemen Z, Lepiniec L, Dubos C. Complexity and robustness of the flavonoid transcriptional regulatory network revealed by comprehensive analyses of MYB-bHLH-WDR complexes and their targets in Arabidopsis seed. New Phytol. 2014;202(1):132–44.
122. Wang J, Lian W, Cao Y, Wang X, Wang G, Oi C, Liu L, Oin S, et al. Overexpression of BoNAC019, a NAC transcription factor from *Brassica oleracea*, negatively regulates the dehydration response and anthocyanin biosynthesis in *Arabidopsis*. Sci Rep. 2018;8:13349.
123. Schmitges FW, Radovani E, Najafabadi HS, Barazandeh M, Campitelli LF, Yin Y, Jolma A, Zhong G, Guo H, Kanagalingam T, Dai WF, Taipale J, Emili A, Greenblatt JF, Hughes TR. Multiparameter functional diversity of human C2H2 zinc finger proteins. Genome Res. 2016;26(12):1742–52.
124. Li C, Xia Y, Jin K. The C2H2 zinc finger Protein MaNCP1 contributes to conidiation through governing the nitrate assimilation pathway in the entomopathogenic fungus *Metarhizium acridum*. J Fungi (Basel). 2022;8(9):942.

Publisher's Note