HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2011 December 07.

Published in final edited form as:

Nat Biotechnol. ; 29(6): 480–483. doi:10.1038/nbt.1893.

Quantitative analysis demonstrates most transcription factors require only simple models of specificity

Yue Zhao and

Washington University Medical School, Department of Genetics, 4444 Forest Park Parkway, Campus Box 8510, Room 5401, St. Louis, MO 63110, United States

Gary D. Stormo

Washington University Medical School, Department of Genetics, 660 S Euclid Box 8232, St. Louis, MO 63110, United States, stormo@wustl.edu

Determining the specificity of transcription factors (TFs) is an important step in understanding regulatory networks and the effects of genetic variations on those networks. In recent years several high-throughput approaches have been developed to rapidly and efficiently determine the specificity of TFs[1]. One important issue that arises in the analysis of binding data is the complexity of the specificity model needed. It has important implications for both the characterization of specificity and for the prediction of the consequences of mutations. If the recognition mechanism is simple, then the specificity of a TF can be modeled by a small number of parameters and the effects of mutations are easily predictable. If recognition is complex, then models of TF specificity will require a large number of parameters and the effects of mutations will be difficult to predict. In the worst case, recognition is so complex that no patterns exist and predictions cannot be made. Structurally, TF-DNA interactions are complex with a wide variety of interactions between the protein and DNA making a simple recognition code impossible[2]. But energetically the situation appears much simpler, with individual base pairs often contributing approximately independently to the total binding energy. Although deviations from strict independence are common, the non-independent contributions tend to be of smaller magnitude compared to the independent contributions. This allows for simple models of interactions, such as position weight matrices (PWM)[3], to be good approximations to the true binding energies. The physical intuition is that TF-DNA recognition is primarily based on complementarity between the sequence dependent positioning of hydrogen bond donors and acceptors in the grooves of the double helix and those on surface to the amino acid side chains of the TF. Since most mutations change the shape of this network of hydrogen bond donors and acceptors locally, their effects are also mostly local.

Protein binding microarray (PBM) is a technique that measures the binding of TFs to double-stranded DNA arrays that currently contain all possible 10-long binding sites and so provides enormous information about the specificity of the TF[4,5]. In a recent PBM study of

Correspondence to: Gary D. Stormo.

mouse TFs, Badis et al.[6] observed that the energetics of TF-DNA recognition appears to be highly complex: 41 out of the 104 TFs studied had clear secondary binding preferences not captured by the primary PWM and 89 out of 104 TFs were better represented by a linear combination of multiple PWMs than a single PWM. However, Badis et al.[6] used three different methods to obtain PWMs and showed that each method was superior to the others on some datasets, indicating that none of the methods can be optimal at determining the PWM parameters. As noted by Badis et al.[6] it is possible that the insufficiency of their PWMs is not due to the complexity of TF-DNA recognition, but rather the algorithms used for parameter estimation. Before abandoning the idea that specificity can be largely explained with simple models, it is critical to assess the fitness of optimal PWMs.

In a typical PBM experiment, a purified, epitope-tagged TF is applied to a double-strand DNA microarray. The degree of binding to each probe on the microarray is quantified by the application of a labeled antibody specific to the epitope tag. In theory, signal intensity of a probe should be directly proportional to the probability of TF binding to the sequence of that probe. In practice, however, the relationship is not so straightforward due to a number of factors such as background signal, position effect and influence of flanking sequences. We have found that these factors significantly confound current analysis methods, such as 8mer enrichment analysis[5] used by Badis et al.[6] (see supplemental figures S2 and associated text for details).

We have taken a different approach: estimate the position and background effects from the data first, then perform weighted regression to parameterize a model of binding energy, explicitly taking these biases into account (see supplemental materials for details). This offers several benefits. First, using a model drastically reduces the number of parameters required: a 10-long PWM only requires 30 parameters. This represents a 1000 fold reduction over 8mer analysis[6], which attempts to estimate TF affinity for all 8-long sequences. Second, having a model of specificity allows us to test hypotheses about the binding mechanism. For example, if the performance of the palindromic model, where the parameters of the half-sites are constrained to equal to each other, is comparable to the full model where all parameters are allowed to vary then it is likely that the TF binds DNA as a homodimer with no interactions between half-sites. An example of this analysis for yeast TF Pho4 is shown in supplemental figure S3. Third, all of the data are used to estimate each parameter, improving accuracy. Finally, by using a model to calculate TF binding probability for the entire probe, the influence of flanking sequence that confound the current analysis is explicitly included. Our algorithm, BEEML-PBM (Binding Energy Estimation by Maximum Likelihood for Protein Binding Microarrays) extends the existing algorithm BEEML[7] to estimate models of TF specificity by weighted regression on PBM data. PBM signal intensity is modeled as a convolution of background effect, position effect and equilibrium binding probability to the probe sequence. Using BEEML-PBM, we find that the simple PWM model of specificity performs very well for most transcription factors. This simplicity has important implications for our understanding of the molecular basis of TF specificity and demonstrates the importance of the analysis method in the interpretation of high-throughput data. Although only PWMs are fitted here, higher order interactions can be easily incorporated into the energy model and their significance can be assessed by standard statistical methods[8].

We evaluate PWM performance by its ability to predict TF binding preferences on a different PBM design. PBM experiments are performed using two arrays with different probe sequences, but both contain all possible 10-long binding sites. We use the PWM trained on array 1 to predict array 2 probe intensities, and vice versa (see supplemental materials for details). While this gives us confidence that the performance achieved by BEEML-PBM PWMs is not due to overfitting to the training data, the fact that the arrays do not have the same probe sequences means we do not have a direct measure of the reproducibility of variations in probe intensities. For this reason, we conduct our analysis at the level of 8mer median intensities (the median intensity of all probes containing each 8-long sequence). 8mer median intensities can be calculated for measured probe intensities of both array designs as well as PWM predicted probe intensities, which allows us to not only compare PWM predictions with experimental measurements, but also determine what fraction of reproducible variance of TF binding can be explained by the PWM model. Although 8mer median intensities are problematic as measures of binding affinity, they serve as a useful measure of how much of the observed sequence-dependent binding variation is experimentally reproducible. In supplemental materials we provide several examples of the PWMs obtained by BEEML-PBM and their assessment by various criteria. Here we focus on the finding that a single BEEML-PBM PWM is usually sufficient to provide excellent quantitative descriptions of PBM data. An example of this is shown in Fig. 1 for mouse factor Plagl1 (pleomorphic adenoma gene-like 1), where the PWM estimated from replicate 1 performs very well on replicate 2 data (Figure 1A). By contrast, the primary PWM found by Badis et al.[6] is unable to capture Plagl1 binding specificity (Figure 1B), leading them to the conclusion that multiple PWMs are required. The BEEML-PBM PWM is qualitatively different from the primary PWM identified by Badis et al.[6] (Figure 1C); given the high level of performance achieved by a single BEEML-PBM PWM it is likely that the need for multiple PWMs identified by Badis et al.[6] is due to suboptimal parameterization rather than the complexity of Plagl1 DNA recognition.

This holds true for most of the 41 TFs identified by Badis et al.[6] as having clear secondary binding preferences. Figure 2A shows that in all but 7 cases, a single PWM explains more than 90% of the experimental variability, defined as the reproducibility of 8mer median intensities (R^2) between replicates. In some cases, PWM performances are better than experimental reproducibility, likely due to different TF concentrations used in replicate PBM experiments. Figure 2B demonstrates that for these 41 TFs, a single BEEML-PBM PWM usually performs as well as, and sometimes better than, a combination of primary and secondary PWMs in the UniPROBE database[9]. Figure 2C shows that in all of the 104 PBM datasets of Badis et al.[6], the PWMs obtained by the BEEML-PBM method fit the replicate data better than the UniPROBE primary PWMs, in many cases very much better. Badis et al.[6] validated binding to secondary motifs of six TFs by electrophoretic mobility shift assay (EMSA). We find that the BEEML-PBM PWMs are usually shorter than the PWMs found by Badis et al.[6], and that those PWMs are often consistent with the EMSA results. For example, the consensus sequence of the BEEML PWM for TF Foxj3 is AAACA, which can be found on both primary (GTAAACAA) and secondary (CAAAACAA) probes. However, there are also a few cases, such as Hnf4a, where the single PWM model is clearly insufficient to capture TF binding specificity.

PBMs are an important technological development, especially in the latest implementations that include all possible 10mer binding sites. They provide an inexpensive and high-throughput method for determining binding specificities of TFs and are rapidly increasing the database of characterized TFs. To maximize the information obtained from this technique it is critical to employ optimized analysis methods. The success of the BEEML-PBM method is mainly due to the power of regression analysis and demonstrates that quantitative PBM data can be analyzed in the traditional biochemical framework of equilibrium binding to obtain accurate binding energies.

With a few exceptions, the simple PWM model performs very well, supporting the hypothesis that the energetics of TF-DNA recognition is generally simple. This simplicity has considerable practical implications. The main difficulty in the study of TF specificity is one of scale. Unlike protein-protein interactions, a single affinity is not sufficient to parameterize TF specificity. For example, there are more than a million possible sequences for a 10-long binding site. Even with high-throughput techniques, direct measurement of affinity for all sites is not practical. However, if the bases contribute to the total binding free energy independently, then a model with only 31 parameters can give accurate predictions of the million binding energies. Even if neighboring di-nucleotide interactions are important, only 112 parameters are necessary[10]. Furthermore, this simplicity can be exploited in the design of promoters with tunable induction or TFs with custom specificity.

In this correspondence, we demonstrate that simple PWMs generally give good approximations of TF specificity, up to the level of reproducibility of PBM experiments. Previous methods to determine PWMs from PBM data did not utilize a biophysical model for the binding and were based on summary statistics, such as E-scores and Z-scores, rather than maximizing the fit to the intensity data directly, taking into account the specific characteristics of PBM data. We conclude that the widespread phenomenon of secondary binding preference identified by Badis et al.[6] is not supported by the data and is likely due to suboptimal estimation of the PWMs. A support vector regression (SVR) method has also been applied to PBM data[11] that provided improved predictions compared to the UniPROBE PWMs in most, but not every, case. In contrast, the PWMs obtained by BEEML-PBM improved the predictions compared to the UniPROBE PWMs in every case and the resulting model has many fewer parameters than the SVR model and each parameter has a specific biophysical interpretation, such as a binding energy contribution of a specific base-pair to the TF-DNA interaction.

BEEML-PBM is freely available at http://ural.wustl.edu/~zhaoy/beeml/

Supplementary Material

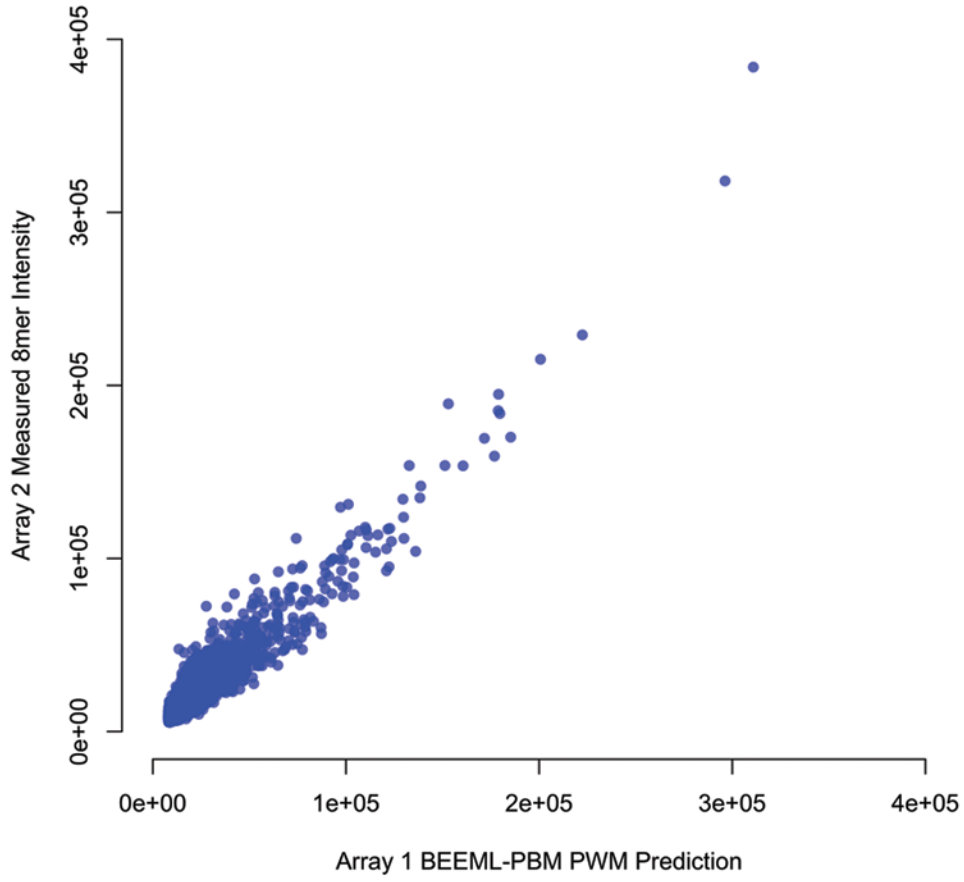
Refer to Web version on PubMed Central for supplementary material.

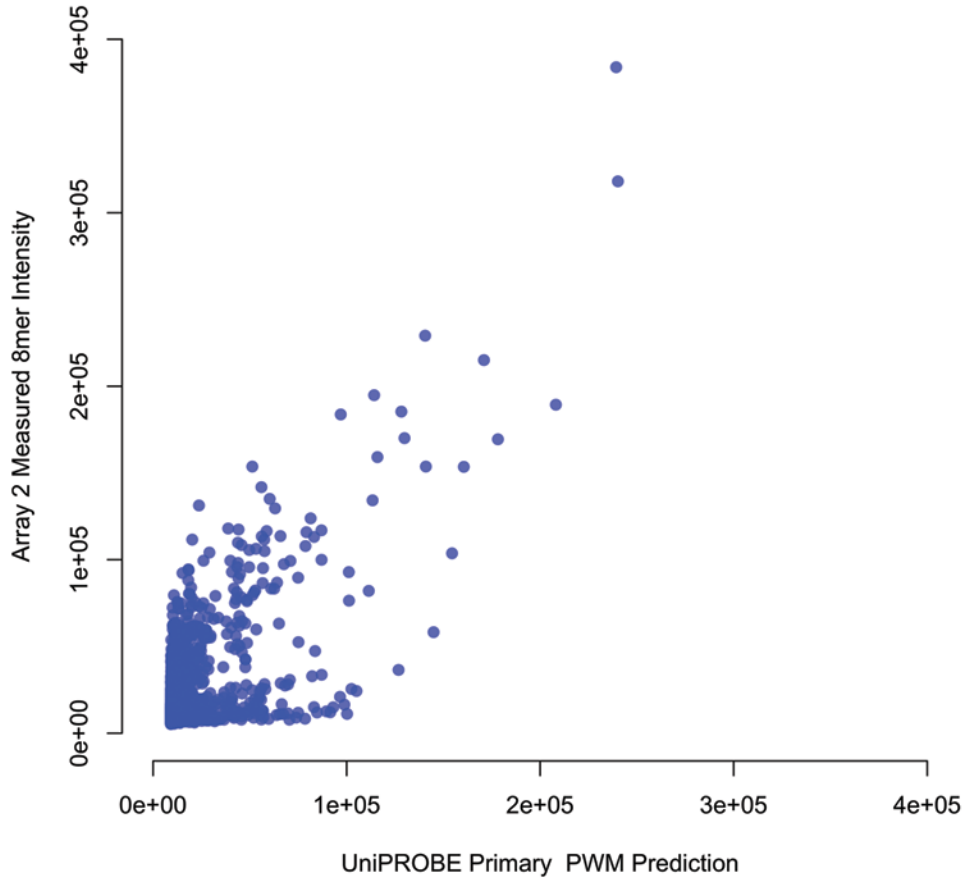
Acknowledgments

We thank Tim Hughes, Martha Bulyk and Quaid Morris for very helpful comments on the manuscript. We also thank members of the Stormo lab for their comments and suggestions throughout the course of this work. This work was supported by NIH grant R01 HG00249 to GDS and NIH training grant T32 HG00045 to YZ.

References

1. Stormo GD, Zhao Y. Nat Rev Genet. 2010; 11:751–760. [PubMed: 20877328]
2. Luscombe NM, Thornton JM. J Mol Biol. 2002; 320:991–1009. [PubMed: 12126620]
3. Stormo GD. Bioinformatics. 2000; 16:16–23. [PubMed: 10812473]
4. Bulyk ML, Huang X, Choo Y, Church GM. Proc Natl Acad Sci USA. 2001; 98:7158–7163. [PubMed: 11404456]
5. Berger MF, et al. Nat Biotechnol. 2006; 24:1429–1435. [PubMed: 16998473]
6. Badis G, et al. Science. 2009; 324:1720–1723. [PubMed: 19443739]
7. Zhao Y, Granas D, Stormo GD. PLoS Comput Biol. 2009; 5:e1000590. [PubMed: 19997485]
8. Benos PV, Bulyk ML, Stormo GD. Nucl Acids Res. 2002; 30:4442–4451. [PubMed: 12384591]
9. Newburger DE, Bulyk ML. Nucl Acids Res. 2009; 37:D77–82. [PubMed: 18842628]
10. Stormo GD. Genetics. Advance online publication. Feb 4.2011 10.1534/genetics.110.126052
11. Agius P, Arvey A, Chang W, Nobel WS, Leslie C. PLoS Comput Biol. 2010; 6:e1000916. [PubMed: 20838582]





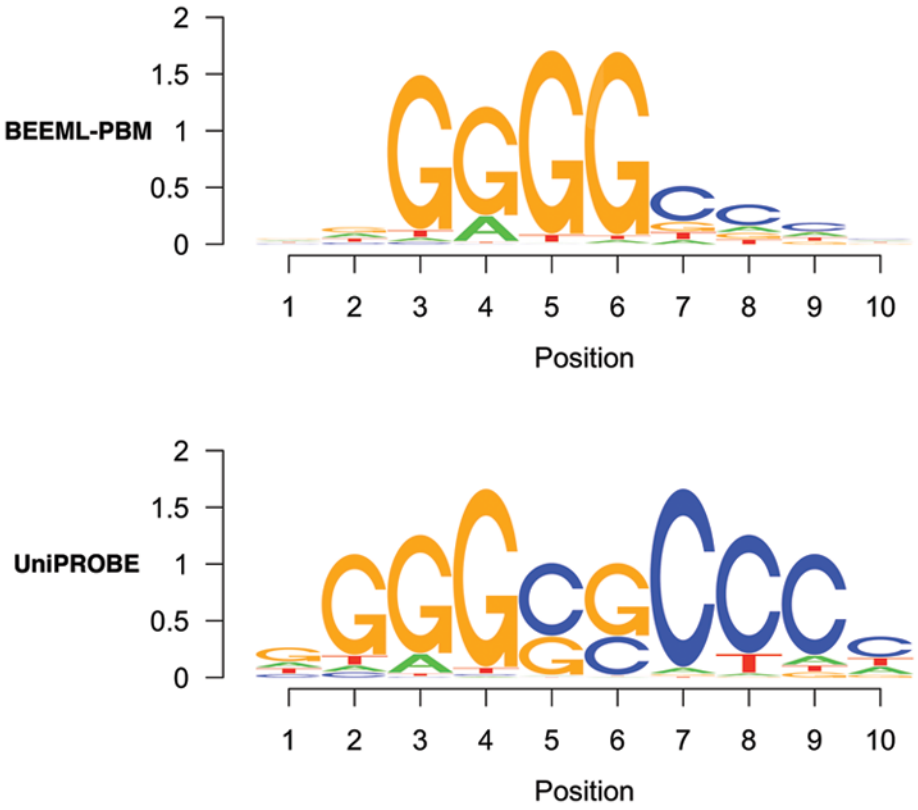
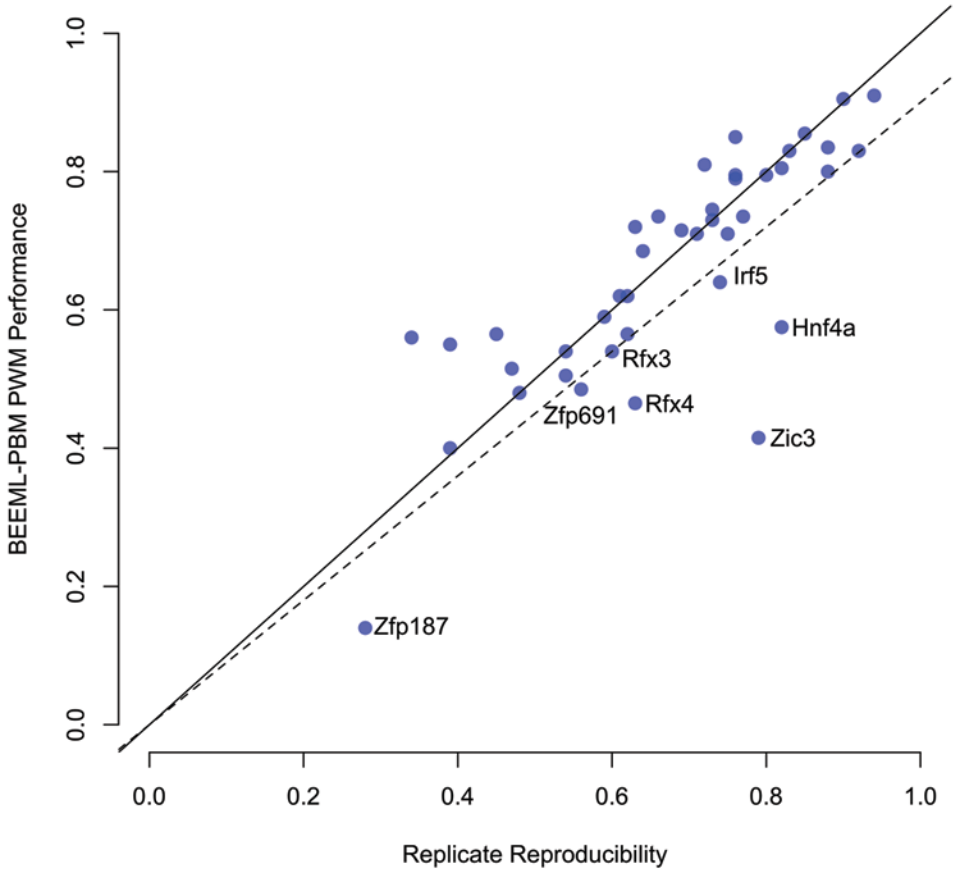
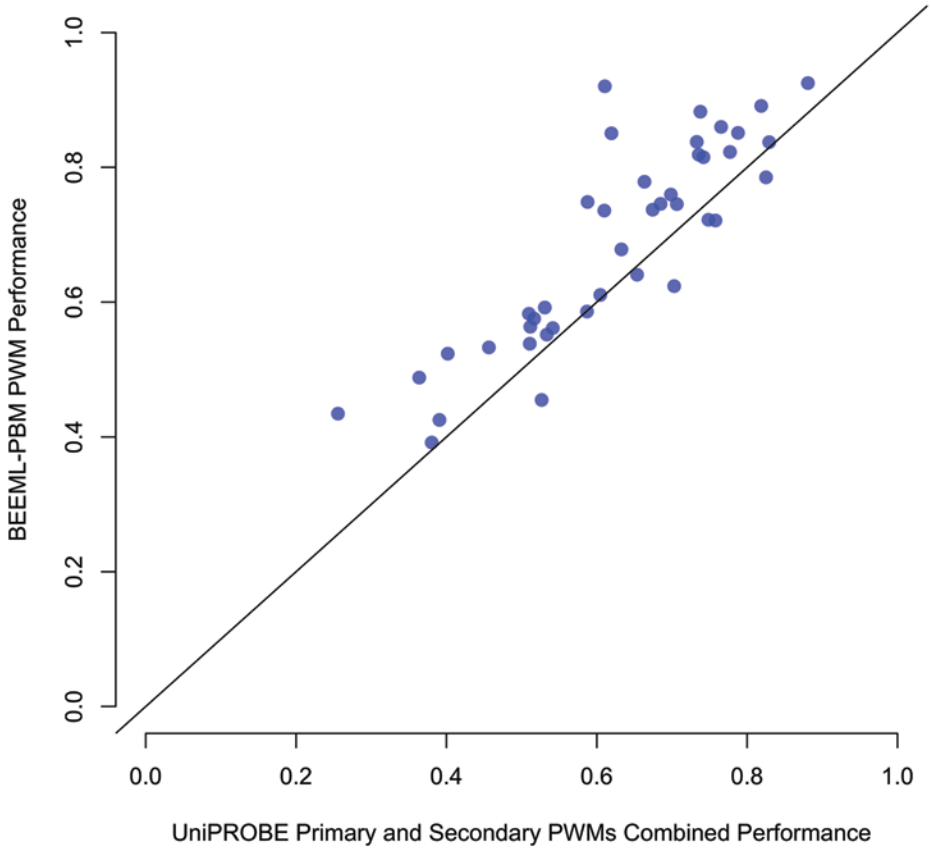


Fig. 1.

Plag1 can be modeled well by a single PWM. (**A**) BEEML-PBM PWM trained on Plagl1 replicate 1 predicts replicate 2 8mer median intensities well with R^2=0.91. (**B**) Performance of Plagl1 primary PWM from UniPROBE database[9] has only R^2=0.47. (**C**) Comparison of Plagl1 BEEML-PBM PWM with primary PWM from UniPROBE database[9].





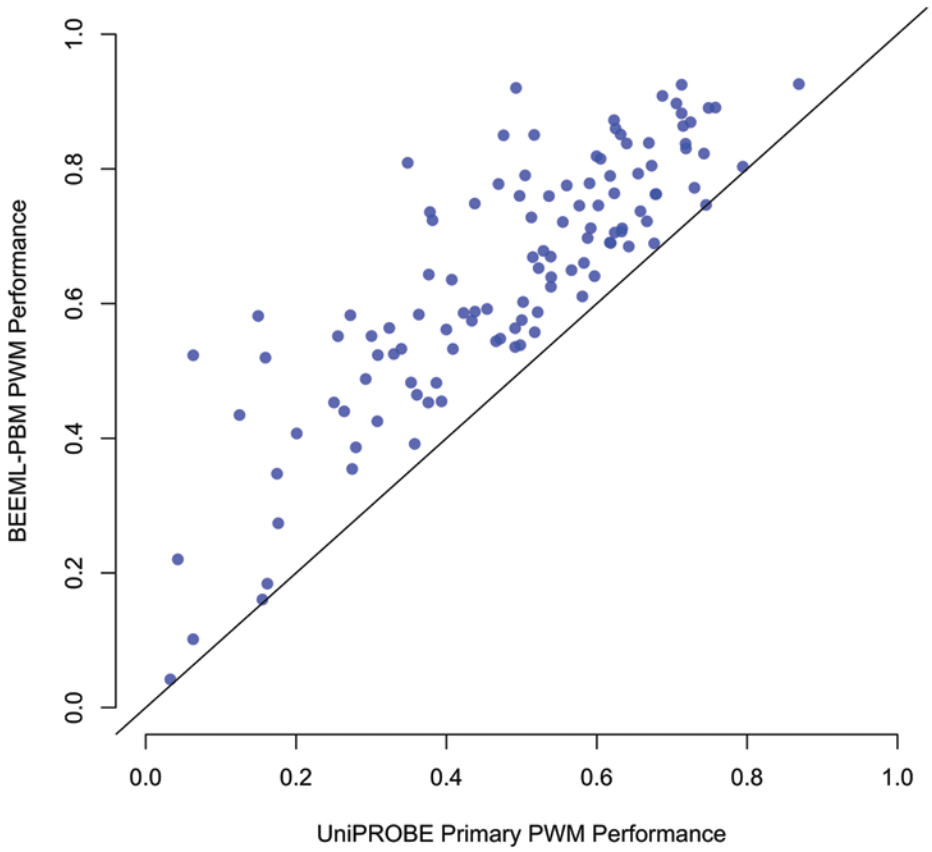


Fig. 2.

A single BEEML-PBM PWM explains "secondary motif" phenomenon (**A**) In all but 7 cases, BEEML-PBM PWM captured more than 90% of experimentally reproducible variability. Dashed line marks 90% variability. (**B**) A single BEEML-PBM PWM usually outperforms a combination of primary and secondary PWMs from Badis et al.[8]. (**C**) BEEML-PBM PWMs outperforms primary PWMs from UniPROBE database[9] for all TFs studied by Badis et al.[6]. The BEEML-PBM PWM from the replicate that gives the best fit is used.