Predicting *In Vivo* Transcription Factor Occupancy from *In Vitro* Binding

by

Rumen Stamatov

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date:_______________________
Approved:

____
Raluca Gordân, Supervisor

____
Alexander Hartemink

____
Lingchong You

____
Tim Reddy

Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in the Graduate Program in
Computational Biology and Bioinformatics in the Graduate School
of Duke University

2014

UMI Number: 1555724

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI
Dissertation Publishing

UMI 1555724

Published by ProQuest LLC (2014). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.
All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code

ProQuest

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ABSTRACT

Predicting *In Vivo* Transcription Factor Occupancy from *In Vitro* Binding

by

Rumen Stamatov

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date:_______________________
Approved:

____
Raluca Gordân, Supervisor

____
Alexander Hartemink

____
Lingchong You

____
Tim Reddy

An abstract of a thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in the Graduate Program in
Computational Biology and Bioinformatics in the Graduate School
of Duke University

2014

Copyright by
Rumen Stamatov
2014

Abstract

The spatial pattern of transcription factor (TF) binding and the level of TF occupancy at individual sites across the genome determine how a TF regulates its targets. Consequently, predicting the location and level of TF binding genome-wide is of great importance and has received much attention recently. The protein-binding microarray (PBM) technology has become the gold standard for studying TF-DNA interactions *in vitro*, while chromatin immunoprecipitation followed by DNA Sequencing (ChIP-seq) is the standard method for inferring TF binding *in vivo*. However, direct interpretation of *in vitro* results in an *in vivo* context is challenging and to-date remains scarce. In this study, we focus on the E2F family of TFs, whose mechanism of binding to DNA has been controversial. Previous studies have shown that E2F factors bind to the TTTSSCGCG motif, where S can be a C or a G. Still, only a small fraction of *in vivo* targets are reported to contain this motif, hinting at indirect recruitment of E2F proteins. We observed that the genomic occupancy of E2F factors directly correlates with their *in vitro* binding affinities. By using data from universal PBM experiments, we show that E2F factors likely bind to DNA through direct sequence recognition and not through cofactor interaction. Furthermore, we developed a kinetic binding model using the PBM data to describe the competition between different members of the E2F family and we successfully distinguished between their unique *in vivo* targets. The model was validated experimentally for a different pair of competing factors – c-Myc and Mxi1. Overall, these results demonstrate how the straightforward and simple *in vitro* PBM experiments can be used for inferring the complex *in vivo* landscape of TF binding and elucidate the mechanism of E2F-DNA interaction.

Contents

List of Tables

List of Figures

1. Introduction

Transcription factors (TFs) are proteins that regulate the expression of genes in a sequence-specific manner. They can have either positive or negative effects on transcription. TFs typically contain at least one DNA-binding domain that recognizes specific features of the DNA. For instance, the TF c-Myc recognizes the consensus sequence CACGTG (called "E-box") or variations of this sequence [1]. This intrinsic preference for particular sites is called *specificity*. The TF binding to the DNA is also described by its *affinity* – how strongly a factor associates with a site. Measuring both specificity and affinity is important for understanding what the TF targets are across the genome and how strongly TFs bind to these targets. Ultimately, this knowledge leads to elucidating the target genes and the way these genes are regulated.

1.1 In vivo methods for studying transcription factor-DNA binding

A standard technique for studying TF binding *in vivo* is chromatin immunoprecipitation followed by sequencing (ChIP-seq). Briefly, proteins are cross-linked to chromatin with formaldehyde and then the DNA is sonicated and broken into fragments. The protein of interest, together with the fragments it was bound to, is isolated and the fragments are detached and sequenced [10]. The sequencing reads are aligned to the reference genome. Stronger TF binding to a location results in more fragments being mapped to this location. Thus, ChIP-seq data is represented as a one-dimensional linear pileup of reads. Computational algorithms are then used to call peaks that are statistically significant compared to the background [11]. This way the location of TF binding can be determined with around 50 base pair resolution. A more accurate variation, ChIP-exo, has been developed, which allows for more accurate observation of the TF binding positions [12]. It uses an exonuclease to degrade the DNA strands around the bound protein, so that only a small fragment remains associated.

A parallel method for pinpointing the location of TF binding is DNase-footprinting [13]. In the standard DNase-sequencing, the DNA is digested with a 5'-to-3' nuclease (DNase). Sites bound by nucleosomes and large proteins are protected from digestion. Thus, the regions of least protection correspond to open chromatin. The footprinting method relies on the fact that even within an open chromatin region, binding of a TF is enough to decrease the digestion rate, leaving a "footprint", resulting in a dip in the open chromatin peak. Statistical and computational methods have been developed to search for binding sites for a TF of interest that match such a DNase footprint [14, 15]. With this method, TF binding can be determined with decent accuracy. An advantage of DNase footprinting over the traditional ChIP-seq is that the DNase experiment can be done only once for a cell type, while ChIP-seq requires separate cell preparations and different antibodies for every TF of interest.

Because these *in vivo* methods require a very large number of cells and are very expensive and laborious, significant efforts have been devoted to building computational models to predict TF binding *in vivo*. In addition, many confounding factors interfere with our ability to understand the behavior of a TF. For example, only a fraction of the binding sites for a particular TF are actually accessible, due to obstruction by nucleosomes or other competing factors. Furthermore, the resolution of the ChIP data is too low to pinpoint the exact locations of TF binding. Therefore, the general paradigm in the field has been to study TF binding *in vitro* by using high-throughput methods and then use this data to build computational models that can accurately identify the locations of TF binding sites in an *in vivo* setting. Some of the most popular *in vitro* methods and computational models are outlined below.

1.2 In vitro methods for studying transcription factor-DNA binding

Popular methods for studying DNA-protein interactions *in vitro* are the Electrophoretic Mobility Shift Assay (EMSA), Systematic Evolution of Ligands by Experimental Enrichment (SELEX) and Protein Binding Microarrays (PBMs). These three methods are briefly reviewed below.

The EMSA technique relies on electrophoresis to separate bound from unbound DNA (RNA) fragments. Because the speed of migration through an agarose gel depends on the charge and the size of the molecule, DNA fragments bound to a protein will travel at a slower rate and result in a shifted band on the gel relative to a control, unbound DNA fragment. While this method is capable of reliable identification of the fragments bound by a TF, only a small number of potential binding sites can be tested. Thus, this technique is often facilitated by choosing candidate sites directly from the genome, such as sequences from within ChIP peaks for a protein of interest or sequences similar to the ones identified in previous studies.

SELEX is a much more high-throughput method for identifying protein-DNA interactions. A random pool of oligonucleotides is exposed to binding by the TF of interest. The sites that are bound with the highest affinity are amplified by PCR and mutated, to generate further candidate oligonucleotides to repeat the cycle. In this way the highest affinity sites are selected for.

Another reliable and high-throughput method for measuring TF-DNA interaction affinity *in vitro* is a Protein Binding Microarray (PBM). Developed by Berger et al., 2006 [8], a universal PBM typically contains all possible 8-mers of DNA immobilized to a glass slide. The protein of interest, tagged with a fluorescent dye, is then added in solution and after incubation, is expected to be bound at different levels to every spot on the array. Quantifying the fluorescence of each spot will thus reflect the affinity of the protein to every possible 8-mer. The results are reported either as the raw fluorescence intensity or a more robust, rank-based statistic called “E-score” (enrichment score), which ranges from -0.5 to 0.5, with 0.5 indicating the strongest binding [9].

1.3 Computational models for studying transcription factor-DNA binding

In the simplest model, the affinity of a TF to a sequence can be approximated with how closely this sequence matches a consensus site, derived from an *in vitro* experiment, such as SELEX. For every site in the genome, one can count the number of matches to this consensus sequence and call the site a true binding site if the number of matches is higher than a pre-defined threshold. However, most TFs bind to highly degenerate sites rather than to a unique motif and therefore a model based on matching to the consensus site is a weak predictor of *in vivo* binding.

A more accurate binding model, taking into account the possible flexibility in TF-DNA recognition, is a Position Weight Matrix (PWM) – a 4-by-k table representation of a k-nucleotide long binding site, in which every column corresponds to a fixed position in this k-mer, and every row specifies the probability that this position is an A, C, G, or T. PWMs can be visualized with motif logos where the height of the letters corresponds to the probability of the specific nucleotide being present at the current position [3]. Clearly, a PWM is a more accurate binding model than a single consensus sequence because it allows for the intrinsic flexibility in the binding. Still, such a model has a major drawback, as it assumes that each position within the site contributes independently to the binding. While the landmark work of Berg and von Hippel [3] suggested that this independence assumption is approximately justified in some cases, numerous studies have shown that in general independence between positions cannot be assumed [4, 5, 6]. For example, the dinucleotide CG may be highly favored but neither C, nor G alone would have any contribution on its own.

The most accurate model to-date of TF-DNA binding is a regression model in which the input features are not only individual nucleotides at specific positions, but also di- and tri-nucleotides [7]. Modeling binding in this way explains *in vitro* binding results with very high accuracy because it captures the intrinsic dependencies in the contribution of neighboring bases to the TF binding.

Different computational methods for predicting TF binding sites have been systematically compared by Weirauch et al. [27]. This study evaluated the performance of 26 models on predicting the measured binding in PBM experiments as well as ChIP data. These methods fall in one of several classes: some are based on learning a PWM, others rely on using k-mers as regression features, and others use Markov chain representations. Simple models, such as those based on a PWM, were reported to have better performance on predicting *in vivo* data, while more complex models trained on k-mer features or using Markov chains were better at predicting *in vitro* PBM binding. All these models, however, have a common property – they try to represent binding information in a compact way, by learning a common trend. As we are going to show later, using binding scores for every 8-mer directly *does* predict *in vivo* data consistently better than the simple PWM models. This approach is different from the models tested by Weirauch et al. because using 8-mer data directly does not represent binding in a succinct way and thus has the advantage of not discarding any relevant information (but has the drawback of being difficult to visualize and interpret).

1.4 Challenges in applying in vitro measurements and computational models to in vivo data

How TFs are recruited to their target sites is a central question in the field of gene regulation, but the mechanisms of protein-DNA recognition are quite complex. Several distinct mechanisms of binding have been proposed. The DNA-binding domain of a TF can interact with the DNA directly – by recognizing a specific DNA motif (dubbed "base readout") or a specific DNA conformation ("shape readout") – reviewed in [23]. Alternatively, the TF can be recruited indirectly by interacting with another protein [24].

Therefore, it is not surprising that a direct correlation between *in vivo* binding (determined from ChIP or DNase footprinting) and either *in vitro* binding (determined from PBM experiments) or computational models (e.g. PWM) has been difficult to establish. Many ChIP-seq peaks are thought to reflect indirect binding – a TF targeting a site not through sequence specificity but through interactions with another DNA-binding protein. Hesselberth et al, 2010 listed several TFs according to what fraction of their ChIP-seq peaks is due to indirect binding. The fraction ranged from 20 % for CTCF to 80 % for E2F4. However, they used only computational tools to predict the binding of a TF and did not incorporate the highly accurate PBM binding data already available for most of these proteins.

In this study we address this issue and clearly demonstrate that the PBM binding data vastly outperforms a PWM model in explaining ChIP-seq data. Our results suggest that close to 100 % of E2F ChIP-seq peaks reflect sequence specificity, compared to only 20 % inferred by Hesselberth et al.

Another difficulty in mapping PBM data to ChIP peaks directly is the competition between paralogous TFs – closely related factors that bind similar sequences but often have spatially non-overlapping occupancy *in vivo* [17]. For example, all E2F family members recognize very similar sequences [18, 19] but some of them bind to mostly unique targets [20]. There is very little overlap between the ChIP-seq peaks of E2F1 and E2F4, consistent with the fact that E2F1 typically functions as an activator and E2F4 as a repressor [21, 22]. Although distinguishing the regions bound by E2F1 from random open chromatin sites is possible, as shown later in the Results section, differentiating between E2F1 and E2F4 peaks is impossible with any of the existing methods described above. We hypothesize that competition between E2F1 and E2F4 for the same sites determines their relative occupancy. To test this theory, we develop a computational model for the competition of these two TFs. Using PBM data to estimate the model parameters, we show that unique E2F1 and E2F4 targets can be predicted with high accuracy, thus providing some insight into the long-standing question of how paralogous TFs recognize different targets.

1.5 Previous studies on competition between DNA-binding proteins

Competition between TFs has been studied previously to some extent. Segal et al., 2008 constructed a thermodynamic model to account for cooperation and competition between TFs with overlapping target sites. They demonstrated that their approach predicts the spatial and temporal patterns of development with very high accuracy. Still, competition was found to play no significant role in affecting the prediction of these patterns. A computational approach to assess the effect of competition was also adopted by Roider et al., 2006 and competition was also found to be non-essential. A possible reason for this inconsequential effect of competition on the conclusions is that these studies focused on unrelated TFs whose binding sites rarely overlap and as a result the places where they compete are very few compared to the places where they bind uniquely. This setting is in sharp contrast with the situation when paralogous TFs are considered, because their target sites are practically the same and it is presumably the intricate differences in specificity for the same sites that determine their different binding profiles.

A comprehensive insight into the competition of DNA-binding proteins was contributed by Wasson et al., 2010. This study challenged the previously existing binary view of proteins as being either bound or unbound, instead providing evidence that the genomic occupancy of a protein can be modeled much more accurately as a continuous profile. By modeling competition between TFs and nucleosomes with a Boltzmann chain approach, the authors demonstrated that TF binding can stabilize or displace nucleosomes, while nucleosome binding in turn mediates the TF occupancy. Their model further suggests that cooperation can arise naturally from competition – as long as several TFs compete for the same binding site and this binding leads to a combined effect on displacing the nucleosomes from this region, thus providing an advantage to both. This study showed in a convincing way that competition for binding is a major factor of *in vivo* occupancy but it did not consider competition between paralogous TFs, which is the focus of the current investigation.

2. Methods

2.1 Universal PBM experiments

A Protein Binding Microarray (PBM) contains 60-bp long oligonucleotides arranged in such a way that every possible 10-mer appears exactly once; thus, every possible 8-mer is represented at least 16 times. Using these 16 replicates, one can determine a robust measure of the binding affinity of a TF. PBM experiments were performed as described previously [7]. Briefly, standard 4 x 44K probe Agilent arrays were first double stranded, blocked with 2% Milk in PBS and subsequently incubated with the protein of interest (E2F1, E2F4, c-Myc or Mxi1). The array was then incubated with a primary antibody against the native protein or a GST/His tag if the protein was tagged. Finally, the array was incubated with a secondary antibody conjugated to a fluorescent dye. In some cases the primary antibody was conjugated to a fluorescent dye, so no secondary antibody was used. Detection was performed by scanning with a laser at the corresponding wavelength. Table 1 summarizes the primary and (where applicable) secondary antibodies for each protein. The data was analyzed with custom software and the median intensity values for the 16 replicates for each 8-mer was recorded. These values were also converted to enrichment scores (E-scores), as described previously [9].

In order to study the competition between two TFs for the same binding sites, we designed a system where two proteins are recognized with antibodies conjugated to different fluorescent dyes, thus allowing simultaneous binding and detection. This experiment is identical to the universal PBM described above with the only modification that two TFs are used instead of one. It is also important that the primary antibodies come from different organisms, so the binding signal of each protein can be separated. In the case of E2F1/E2F4 competition, we used primary antibodies against E2F1 and E2F4 from rabbit and mouse, respectively. They are recognized by the corresponding secondary antibodies goat anti-rabbit and goat anti-mouse, conjugated to a blue or a red fluorescent protein, respectively.

Table 1: Antibodies and fluorescent dyes used in PBM experiments

Proteins	Primary antibody	Secondary antibody	Wavelength [nm]
His::E2F1	Rabbit anti-E2F1	Goat anti Rabbit-Alexa 647	647
His::E2F4	Mouse anti-E2F4	Goat anti-Mouse-Alexa 488	488
His::c-Myc	Rabbit anti c-Myc	Goat anti-Rabbit-Alexa 647	647
His::Mxi1	anti-His-Alexa 488	None	488

2.2 Classification of bound and unbound regions by using an ROC curve

All ChIP peaks for a particular TF were used as a positive set. The peaks were called with the MACS2 software at significance threshold q = 0.001. Using a less stringent cutoff threshold of 0.01 did not change the number of peaks significantly. For the negative set, we selected a random sample of the same size from the DNase-seq peak regions not overlapping with the ChIP-seq peaks. Both sets were taken from the ENCODE data on the same cell lines. For each set, we extracted the DNA sequences in a 150 bp region around the peak summit. Using a sliding window of size 8, we listed all overlapping 8-mers in the region and mapped each 8-mer to the corresponding E-score from the universal PBM. The highest of these scores was assigned to the region. Then an arbitrary threshold was set and all regions with a score above were called "bound". By varying this threshold across its possible range, from -0.5 to 0.5 (the possible range of the PBM E-score), we plotted the false positive rate and the true positive rate, thus generating a Receiver Operating Characteristic (ROC) curve. The area under the curve was calculated and used as an evaluation of the model. PBM data for E2F1, E2F4, c-Myc, MAX and Mxi1 was used from recent experiments performed in our lab; PBM data for all other proteins was taken from the Uniprobe database.

2.3 Visualizing the in vitro binding profile of genomic regions

The binding score profile plots were generated by using a sliding window of size 8 bp and taking the corresponding PBM score for every such 8-mer. The resulting signal was smoothed with a Savitzky-Golay filter of power 2 and length 301 to remove high-frequency noise. The ChIP data was visualized in a similar way, by smoothing the raw read counts for every position in the region of interest.

2.4 Enrichment analysis of 8-mers in ChIP peaks

For every possible 8-mer, we counted the number of times it appears in all ChIP-seq peaks, in a 100 bp region around the peak summit. This number was divided by the number of times the 8-mer appears in the 400 bp region around the peak summit but excluding the central 100. The latter represents the background distribution of the 8-mer. The ratio of the two numbers represents the enrichment of the 8-mer in ChIP peaks relative to the background. This ratio was plotted against the corresponding PBM score.

2.5 Kinetic model for TF competition and application to the E2F family

The idea behind using a competition *in vitro* system is described in Fig. 1. Paralogous TFs (Fig. 1a) bind different genomic targets for reasons that are currently unknown. Fig. 1b pictures TF competition, a possible reason for explaining the different *in vivo* specificity, explored in this project. In this case study, E2F1 and E2F4 are paralogous TFs taht can be either bound to DNA, or free in solution. We model their competition to sets of PBM sequences and in the Results and Discussion section we explain how it can be adapted to an *in vivo* setting. For each spot on the array, let [Ifree] represent the concentration of binding sites (equivalent to the number of oligonucleotides per unit area) that are not bound by any protein. By analogy, the concentrations of bound E2F1 and bound E2F4 to site I would be [E2F1bound_i], and [E2F4bound_i]. For simplicity, we are going to drop the index *i* and model the binding process for each of the 4^8 spots on the universal PBM array individually, assuming that the binding at site i does not affect the binding at site j for any i and j. This assumption is justified if the TF concentration of both factors is in excess, so that the amount of bound protein at any site is negligible and does not limit the amount of free protein available for binding at the rest of the sites. Thus, we are going to have a system of 4 equations for each spot on the array. Since independence was assumed, we can solve this system for every spot individually.

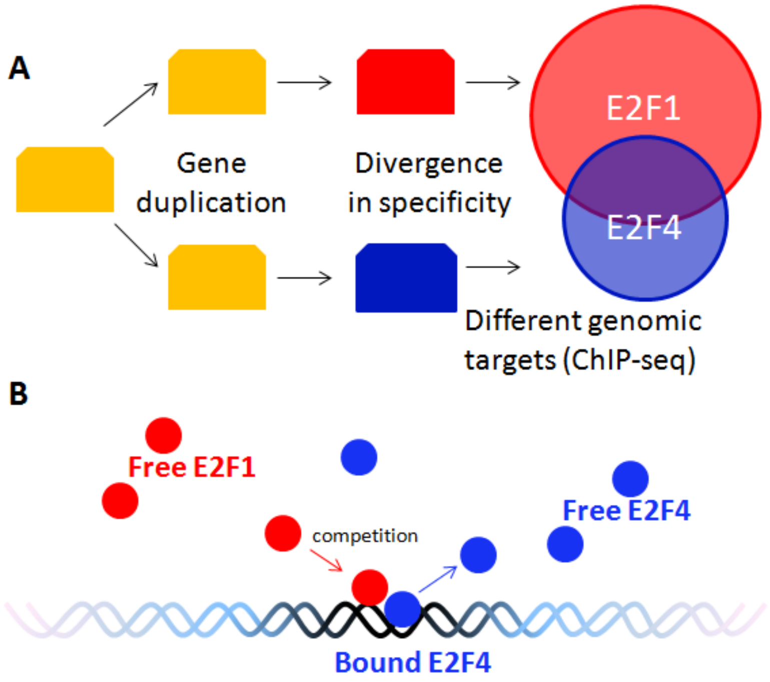

Figure 1: Competition between paralogous transcription factors

First, let us define a simple case when only one protein (E2F1) is available for binding. By definition, the dissociation constant of E2F1 to site I is the following:

$$K_{d1i} = \frac{\left[E2F1_{free}\right]\left[I_{free}\right]}{\left[E2F1_{bound}\right]} \tag{1}$$

Mass conservation for both E2F1 and the binding site yields these two equalities:

$$[\mathrm{E2F1}_{free}] + \left[E2F1_{bound}\right] = [\mathrm{E2F1}_{total}] \tag{2}$$

$$\left[I_{free}\right] + \left[E2F1_{bound}\right] = [I_{total}] \tag{3}$$

Therefore, the fraction of bound E2F1 at equilibrium, which also represents the probability of binding of E2F1 to site I, is defined as follows:

$$P_{E2F1bound} = \frac{\left[E2F1_{bound}\right]}{\left[E2F1_{total}\right]} \tag{4}$$

Similarly, the same 4 equations can be written for E2F4, in a simple case when it is the only protein available:

$$K_{d4i} = \frac{\left[E2F4_{free}\right]\left[I_{free}\right]}{\left[E2F4_{bound}\right]} \tag{5}$$

$$[\mathrm{E2F4}_{free}] + \left[E2F4_{bound}\right] = [\mathrm{E2F4}_{total}] \tag{6}$$

$$\left[I_{free}\right] + \left[E2F4_{bound}\right] = [I_{total}] \tag{7}$$

$$P_{E2F4bound} = \frac{\left[E2F4_{bound}\right]}{\left[E2F4_{total}\right]} \tag{8}$$

These two systems (eq. 1-4 on one hand and eq. 5-8 on the other hand) are analogous to the universal PBM experiments with either E2F1 or E2F4. In these experiments, the binding score for each spot is directly related to the probability of binding of the TF to this spot. Thus, we are going to assume that the probability of binding of E2F1 (or E2F4) to site I, in the case of this simple non-competing setting, can be represented by the area under the ROC curve describing binding to site I relative to the background binding (described in detail in [9]). By definition, the standard E-score statistic is equal to the difference between the site-specific area under the ROC curve and the area under the background binding ROC curve. Because the background binding can be described by a random (non-specific) model, the area under the ROC curve in that case will be around 0.5. Therefore, we will assume that the fraction of bound E2F1 (or E2F4) equals the E-score + 0.5. Because the E-score ranges from -0.5 to 0.5, the fraction of bound protein will range between 0 and 1.

Expanding equation (1) by using (3), (4) and (5) and making substitutions where possible yields the following:

$$
\begin{aligned}
K_{d1i} &= \frac{\left[E2F1_{free}\right]\left[I_{free}\right]}{\left[E2F1_{bound}\right]} = \\
&= \frac{([\mathrm{E2F1}_{total}]-\left[E2F1_{bound}\right])([I_{total}]-\left[E2F1_{bound}\right])}{\left[E2F1_{bound}\right]} = \\
&= \frac{([\mathrm{E2F1}_{total}]-P_{E2F1bound}\left[I_{total}\right])([I_{total}]-P_{E2F1bound}\left[I_{total}\right])}{P_{E2F1bound}\left[I_{total}\right]} = \\
&= \frac{([\mathrm{E2F1}_{total}]-P_{E2F1bound}\left[I_{total}\right])(1-P_{E2F1bound})}{P_{E2F1bound}} \approx \\
&\approx \frac{([\mathrm{E2F1}_{total}])(1-P_{E2F1bound})}{P_{E2F1bound}}
\end{aligned}
\qquad (10)
$$

The approximation on the last line follows from the assumption that the concentration of E2F1 is in excess and the bound E2F1 is negligible compared to the total. This derivation shows that the dissociation constant of E2F1 to any site can be approximately calculated by knowing the probability of binding (directly observed in the PBM experiment) and the concentration of the TF (also controlled in the experiment). By analogy, we can derive a similar equation for the dissociation constant of E2F4 to every site I:

$$
K_{d4i} \approx \frac{([\mathrm{E2F4}_{total}])(1-P_{E2F4bound})}{P_{E2F4bound}} \qquad (11)
$$

Now we can proceed to modeling the situation when both E2F1 and E2F4 are present. The system will consist of equations (1), (2), (4), (5), (6) and (8), as well as equation (3) but modified to account for the possibility of both E2F1 and E2F4 binding to the same site:

$$\left[I'_{free}\right]+\left[E2F1'_{bound}\right]+\left[E2F4'_{bound}\right]=[I_{total}] \qquad (12)$$

We note that this system is independent of the two simplified cases we considered above. Thus, although we use the same variable names, the concentrations will be different and we will denote the concentrations in the double case with a prime (E2F1' instead of E2F1) to distinguish them from the single case. However, the dissociation constants will be the same, since they describe intrinsic physical properties of the proteins and the DNA and do not change depending on the protein concentration or the conditions. Therefore, we can determine these constants for every site by using equations (10) and (11), and we can apply them to the competition setting. We are interested in the following fraction:

$$E2F1_{preference}=\frac{\left[E2F1'_{bound}\right]}{\left[E2F4'_{bound}\right]} \qquad (13)$$

This represents the ratio of bound E2F1 to bound E2F4 at the same site.

$$\begin{aligned} E2F1_{preference} &= \frac{\left[E2F1'_{bound}\right]}{\left[E2F4'_{bound}\right]} = \\ &= \frac{\left[E2F1'_{free}\right]\left[I'_{free}\right]K_{d4i}}{K_{d1i}\left[E2F4'_{free}\right]\left[I'_{free}\right]} = \\ &= \frac{\left[E2F1'_{free}\right]K_{d4i}}{\left[E2F4'_{free}\right]K_{d1i}} \end{aligned} \qquad (14)$$

Using equations 10 and 11 and making the appropriate substitutions yields the following:

$$E2F1_{preference} = \frac{\left[E2F1'_{free}\right]}{\left[E2F4'_{free}\right]} \frac{([\mathrm{E\,2\,F\,4}_{total}])(1-P_{E2F4bound})P_{E2F1bound}}{([\mathrm{E\,2\,F\,1}_{total}])(1-P_{E2F1bound})P_{E2F4bound}} \quad (15)$$

If the concentrations of E2F1 and E2F4 are equal, then we have:

$$E2F1_{preference} = \frac{P_{E2F1bound}(1-P_{E2F4bound})}{P_{E2F4bound}(1-P_{E2F1bound})} \quad (16)$$

3. Results and discussion

3.1 In vitro binding data outperforms the PWM in predicting the genomic locations of TF binding

Position Weight Matrices (PWMs) are the standard model for TF binding site preferences but they suffer from the assumption that every position contributes independently to the binding. Therefore, we sought to compare the performance of this model against the more accurate in vitro PBM binding data on predicting the locations of TF binding genome-wide. We used ChIP-seq peaks for E2F1 called with the MACS2 software at significance level q = 0.001 as a positive set, and an equally-sized set of random regions of open chromatin determined from DNase-seq, not overlapping with the E2F1 regions, as a negative set. For both the positive and the negative set, we recorded the sequence with the closest match to the PWM and the top-scoring 8-mer from the PBM experiment for the TF of interest (PWMs were taken from the TRANSFAC database and PBM experiments from Uniprobe [26, 27] or data generated in this study). Then we set an arbitrary threshold and classified a region as "bound" if its PWM match or 8-mer score exceeds the threshold, and "unbound" otherwise. Thus, the fraction of ChIP-seq peaks correctly identified as "bound" is the true positive rate, and the fraction of incorrectly called negative DNase peaks as "bound" is the false positive rate. By varying the threshold and recording the false positive rate and the true positive rate, we generated a receiver operating characteristic (ROC) curve and used the area under the curve (AUC) as a measure of the model accuracy. An area of 1.0 is equivalent to a flawless predictive model and 0.5 corresponds to random prediction.

We chose the TFs from [16] – based on their DNase footprinting and motif match results, the authors reported that the majority of these factors bind to DNA indirectly – through protein-protein interactions and not by direct sequence recognition.

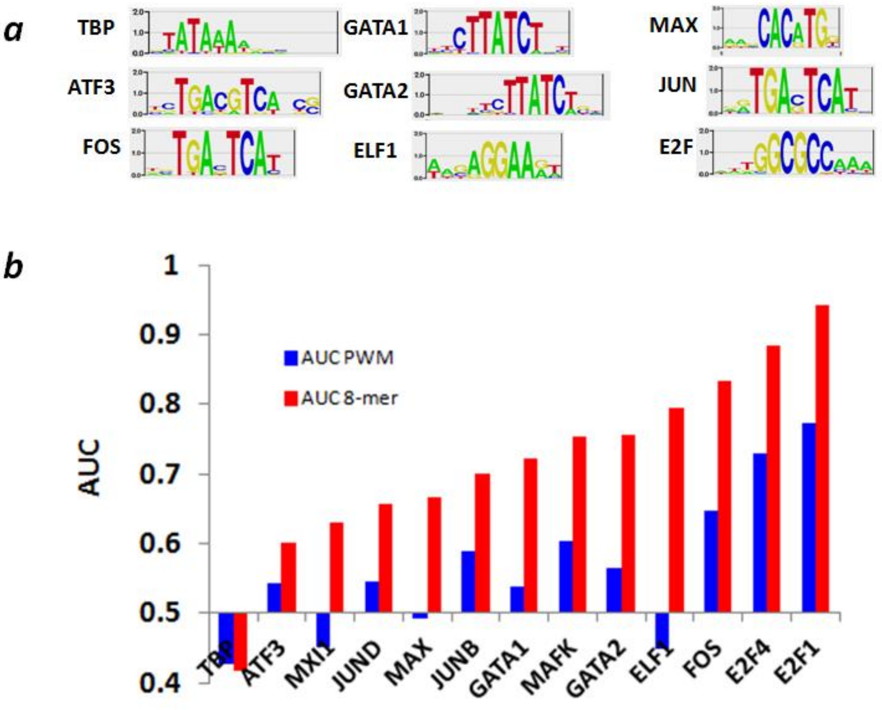

Figure 2: PBM binding data outperforms the standard PWM model in predicting TF binding. (a) Motif logos of the PWMs for TFs with different sequence specificities. (b) Comparison of AUC enrichment of a classifier distinguishing bound from unbound regions on the basis of the PWM model (blue) or the PBM 8-mer scores (red).

The motif logos for these TFs are shown in Fig. 2a and the AUC results in Fig. 2b. For all factors tested, the PBM score vastly outperformed the PWM model in predicting bound regions, which implies that the PBM score is a much more accurate representation of TF-DNA recognition than the standard PWM model and conclusions based on PWM matches should be reevaluated in light of this high-throughput and now widely available technology.

3.2 Accurate prediction of in vivo DNA binding by E2F proteins

As discussed above, TF-DNA binding can occur through direct sequence/shape recognition or indirect binding through another DNA-binding protein. Which of these modes is utilized by E2F has been a controversial issue. According to the popular model, E2F is primarily recruited through protein-protein interactions. On the other hand, a recent molecular biology study argues against this idea and supports the direct DNA recognition hypothesis [24].

To elucidate the mechanism of E2F1 recruitment, we developed a classification model based solely on *in vitro* Protein Binding Microarray (PBM) experiments and validated it on publicly available ChIP-seq data. If *in vitro* binding alone is sufficient for accurate prediction of *in vivo* occupancy, then the direct recognition hypothesis is valid. The model was constructed as discussed in section 3.1 and evaluated with an AUC score.

First, we used the *in vitro* enrichment score (E-score) from the universal PBM of the top-scoring 8-mer in the 300 bp window. This feature alone achieved an AUC = 0.93. The high accuracy of the model might be due to the fact that E2F binds mostly in promoter regions, which may be easier to distinguish from general DNase hypersensitive sites, without reflecting the E2F specificity. To address this concern, we selected only negative regions that are in promoters, and still achieved the same high prediction accuracy, implying that the model indeed reflects E2F specificity (Fig. 3).

Next, we took the best match score to the canonical E2F motif – TTTssCGCG, where s can be C or G. This resulted in a significantly lower AUC of 0.82, underlying the accuracy and high quality of the PBM data compared to the PWM matching. This difference also demonstrates the plasticity of the E2F recognition site, which cannot be captured in a single consensus motif, but still can be uniquely determined from the sequence.

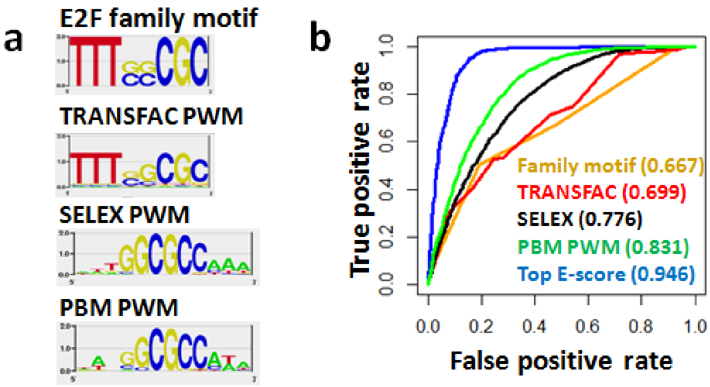

Figure 3: In vitro binding agrees with in vivo occupancy. A series of receiver operating characteristic (ROC) curves were calculated to show the performance of a classification model for predicting bound sequences *in vivo* using either PWMs, or PBM E-scores.

The high AUC of the classification model provides strong evidence that E2F recognizes its DNA binding sites directly and is not recruited by protein-protein interactions.

In order to further demonstrate that E2F binding is not confined to one specific site, we observed the E-score profile, custom PBM score profile and PWM match profile along 10-30 Kbp of DNA stretches of open chromatin. Then we smoothed the profiles with a 2nd degree Savitzky-Golay filter of range 300. This filter removes high frequency noise in the data and thus weakens the effect of any single high-scoring sequence, while emphasizing the density of high-scoring sequences as a whole. These profiles were then compared to ChIP-seq pileup data of the same regions, smoothed as above (Fig. 4).

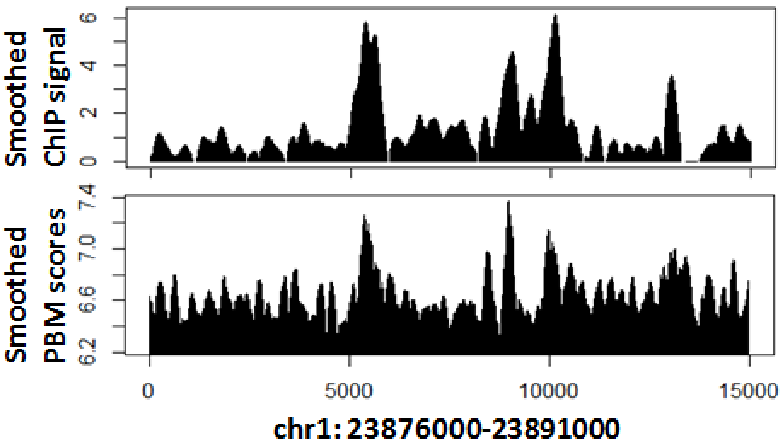

Figure 4: Peaks of 8-mer scores in the genome correlated well with peaks from *in vivo* ChIP-seq experiments. Shown is a region of chromosome 1 of the human genome. Data were smoothed with a Savitzky-Golay filter (power = 2, n = 301).

Strikingly, visual inspection shows that the smoothed *in vitro* scores and the *in vivo* ChIP-seq data match. This result solidifies the argument that single high affinity sites are not all-or-none determinants of E2F occupancy; rather, strong and weak binding sites tend to cluster together in a dense array that attracts E2F as a whole. One can speculate about the significance of such clustering. It may be important for more efficient gene regulation – multiple E2F proteins bound together might be more efficient in assembling the transcription pre-initiation complex. The clustering may also provide a robust firewall against random mutations – in the case of a single binding site a mutation would have a detrimental effect on regulating a gene, while in the multiple binding hypothesis, other sites will make up for the lowered affinity.

3.3 Enrichment of high-scoring 8-mers in ChIP-seq data

We observed that high scoring 8-mers from the PBM were enriched in the center of ChIP peaks compared to the flanking sequences (Fig. 5). This observation provides another line of evidence that the PBM results are biologically relevant *in vivo*.

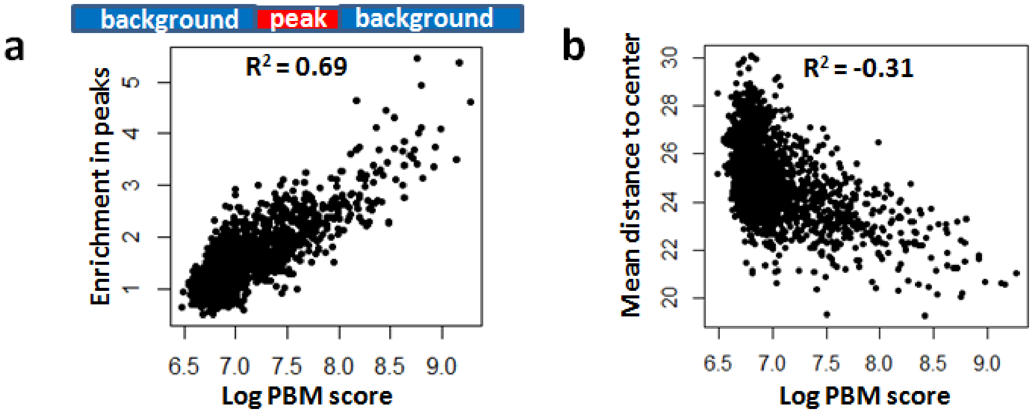

Figure 5: (a) High scoring 8-mers are enriched in ChIP-seq peaks and are located closer to the peak center. The occurrences of each 8-mer were counted in a 100 bp region centered at every ChIP-seq peak summit and the sum was divided by the sum of occurrences in the flanking 200 bp regions. The 8-mers with at least 100 total occurrences in the peaks are shown. This enrichment ratio is well correlated with the log PBM score. The mean distance to the peak center is inversely correlated with the PBM score (b).

3.4 Computational model of TF competition for DNA binding

After having successfully applied the PBM results to differentiating between bound and unbound regions by E2F, we undertook the more difficult task of differentiating between unique E2F1 and unique E2F4 targets. The unique targets of these factors have important significance *in vivo*, as E2F1 usually functions as a transcriptional activator and E2F4 as a transcriptional repressor. Yet, the highest scoring 8-mers in the E2F1 PBM and E2F4 PBM are the same. Therefore, we hypothesized that the unique targets are determined by competition of the two factors for lower affinity sites. This idea is supported by several lines of evidence. First, unique ChIP peaks for either E2F1 or E2F4 are quantitatively less significant than common peaks between the two factors, suggesting that the unique peaks may not contain high-affinity sites. Second, a typical ChIP peak contains many potential E2F binding sites – E-score profile shown in Fig. 6. And finally, some 8-mers with lower affinity show significant difference between their *in vitro* E2F1 PBM score and the E2F4 PBM score.

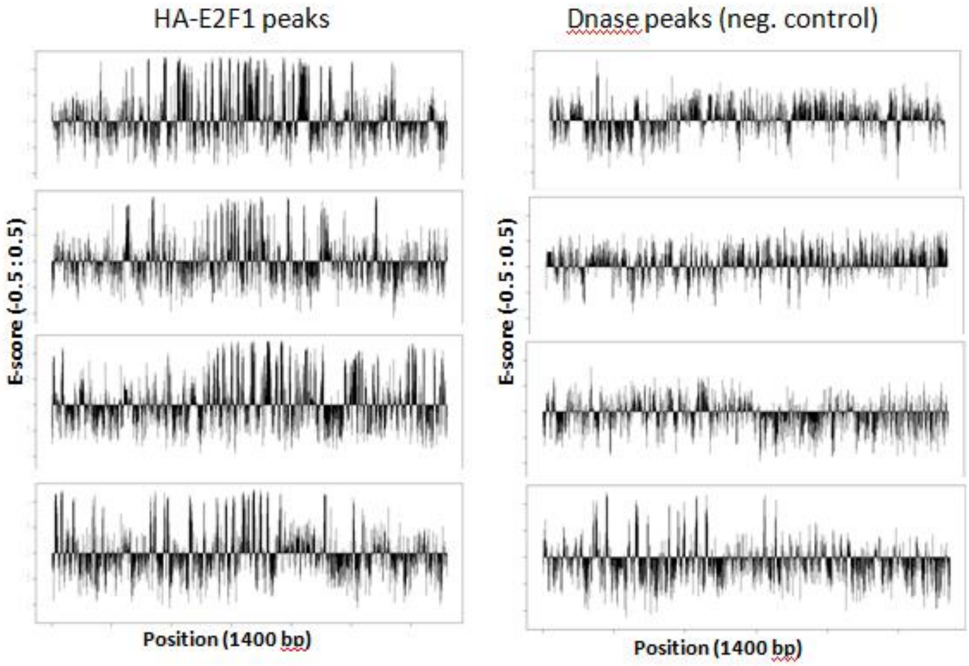

Figure 6: Comparison between the 8-mer profiles of ChIP peaks and non-ChIP peaks

Therefore, we developed a model to describe binding of E2F1 and E2F4 to DNA, consisting of two parts. In the first part, it describes the binding of a single protein – either E2F1 or E2F4 – to different sites on DNA. In the second part, it captures the binding when both proteins are present at the same time and compete for the same sites. We used the universal PBM experiments to estimate the values of the dissociation constants for E2F1 and E2F4 to all possible 8-mers. Then using the second part of the model, we predicted the ratio of E2F1 occupancy to E2F4 occupancy for every 8-mer.

The derivation of the model equations are explained in detail in the Methods section. We validated the model on independently published *in vitro* data for the human transcription factor Max. Maerkl et al., 2007 measured the dissociation constants of the human Max TF to different 10-mers of the form GGNNNGTGGG, where NNN were varied. This was achieved with a technique called mechanically induced trapping of molecular interactions (MITOMI). We compared these Kd values to the ones obtained from our model by using PBM data for the 8-mers NNNGTGGG – matching the last 8 nucleotides of the 10-mers used by Maerkl et al. They agree very well within a constant factor (Fig. 7). Therefore, while we cannot get the exact value of the Kd from the model, the relative values are meaningful. This does not pose a problem for the conclusions because calculating the preferences for E2F1 vs. E2F4 requires taking the ratio of the Kd-s and not their exact values.

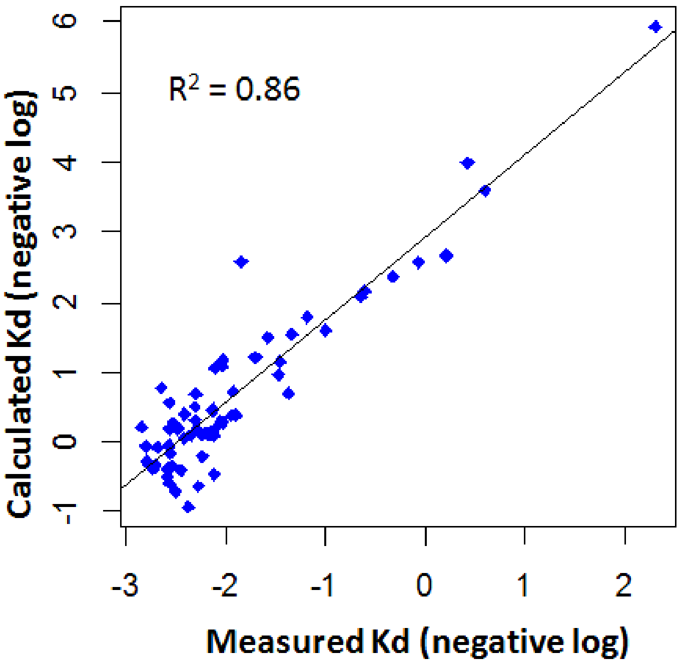

Figure 7: Predicted dissociation constants for the human transcription factor Max agree well, within a constant factor, with those measured with MITOMI in vitro.

The results are shown in Fig. 8. Indeed, small differences in the binding affinities between the factors when they bind individually are sufficient to result in a significant difference in binding when they are together. The predicted ratio of E2F1 to E2F4 varies between 1/10 and 10 times, as demonstrated by the color-coded plot (Fig. 8). As expected, the 8-mers with the greatest differential binding (darkest color in Fig. 8) are the ones with E-scores of 0.35 – 0.4, which are not the highest affinity binding sites, consistent with the above reasoning.

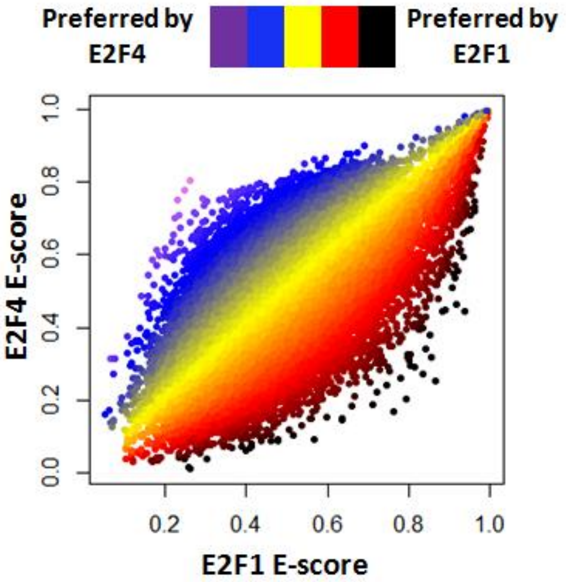

Figure 8: Visualization of the competition results in vitro. Shown is a plot of the E-scores + 0.5 to every 8-mer from the PBM experiments for E2F1 and E2F4. We assume that these scores approximate the fraction of bound protein to every 8-mer. The colors indicate the predicted ratio of E2F1 to E2F4 occupancy, logarithmic scale ranging from -4 (purple) to +4 (black).

We then asked if we can use the competition model to computationally differentiate unique E2F1 from unique E2F4 peaks. We defined unique peaks as those where the ratio of the sum of ChIP-seq reads in a 150-bp region of E2F1 to E2F4 (or vice versa) is larger than 1.5. Then for each such peak we calculated an integrated competition score – the sum of the competition scores of 10 most pronounced 8-mers in terms of E2F1/E2F4 differential binding minus the sum of the scores of the 10 most pronounced 8-mers for E2F4/E2F1 differential binding. Using this integrated competition score as a threshold for classifying peaks as "E2F1 unique" or "E2F4 unique", we built an ROC as explained previously. The area under the ROC curve (AUC) for E2F1 vs. E2F4 unique peaks was 0.865 and the AUC for E2F4 vs. E2F1 was 0.795 (Fig. 9). These results strongly support the previous conclusion that E2F binding to DNA is sequence-specific instead of co-factor mediated. They further indicate that paralogous TF competition *in* *vivo* plays a significant role for the determination of unique targets and that these targets in the case of E2F can be directly predicted from *in vitro* binding specificity.

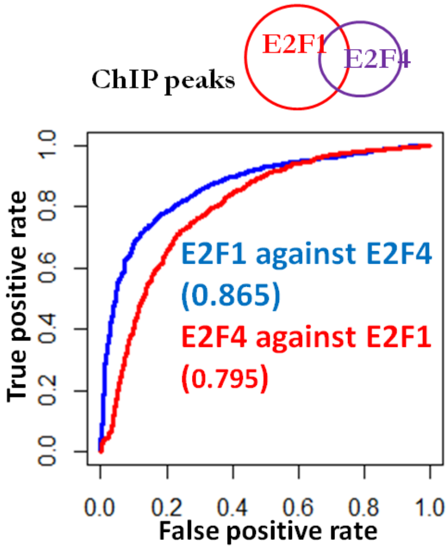

Figure 9: Distinguishing between unique E2F1 and unique E2F4 peaks. Top: a schematic Venn diagram shows the unique and overlapping ChIP-seq peaks between E2F1 and E2F4. Bottom: ROC curves for predicting E2F1 against E2F4 unique peaks (blue) or E2F4 against E2F1 unique peaks (red).

A competition PBM experiment provides another line of evidence in favor of the hypothesis that differences in the binding specificity of paralogous TFs are a major factor in determining their different genomic targets. We designed a system to observe TF competition directly by allowing two proteins, having different tags, to bind to the microarray at the same time. For practical reasons, we chose the c-Myc (henceforth referred to as Myc) and Mad2 (or Mxi1; henceforth referred to as Mad) factors, involved in differentiation, cell proliferation and apoptosis. Both factors are known to heterodimerize with another protein, Max. Thus, we planned to observe the differences in binding between Myc:Max and Mad:Max heterodimers, for simplicity referred to as just Myc and Mad.

The experimental design is shown in Fig. 10a-c. Both Myc and Mad were tagged with a His tag, which can be recognized by an antibody conjugated to a blue fluorescent dye. Myc can be recognized by another antibody, specific to the protein, which is conjugated to a red fluorescent dye. The three proteins (including Max, required for binding) were mixed together and allowed to compete for the same sites on the microarray. Detecting the fluorescence in the blue channel yielded the amount of total bound protein, while the fluorescence in the red channel reflected the bound Myc. We then asked whether the fraction of bound Myc in the competition system agrees with the one predicted from the model, based on separate PBMs for Myc and Mad. Fig. 10d shows the plot of the enrichment scores of Myc and Mad determined from separate PBM experiments, color-coded to reflect the Myc or Mad preference, as determined from the model. A similar plot was generated for the competition PBM (Fig. 10e) and also color-coded according to the model. Interestingly, we notice that both the shapes and the colors of these plots are almost identical. The sequences preferred by Myc (or Mad) in the competition experiment are the same as the ones predicted to be preferred by Myc (or Mad) from the model based on the separate PBMs. Together, these results suggest that separate universal PBM experiments, combined with the computational model, provide a powerful tool to understand the differences in genomic binding of paralogous TFs.

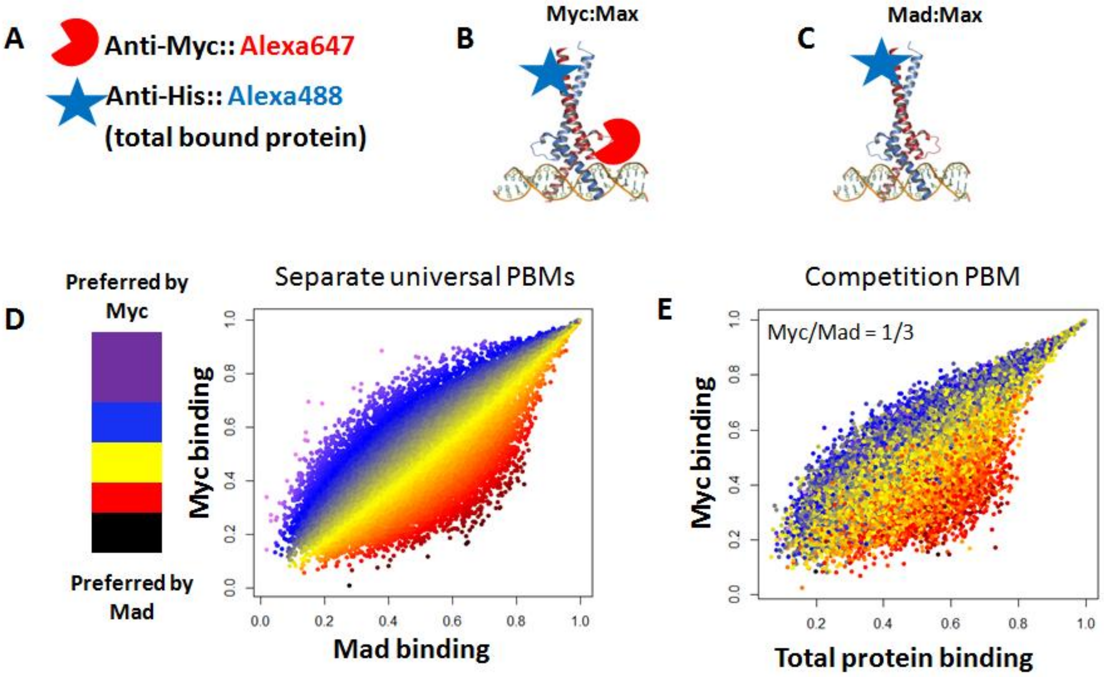

Figure 10: The competition PBM experiment between Myc:Max and Mad:Max heterodimers. (a) Fluorescently conjugated antibodies allow us to measure either bound Myc or the total bound protein. (b, c) Schematic representation of the Myc:Max and Mad:Max heterodimer recognition by the antibodies. (d) The enrichment scores of separate PBM experiments are color-coded according to the computational model to reflect the different preferences between Myc and Mad. (e) The enrichment scores for Myc and total protein binding in the competition PBM agree well with those obtained from the separate universal PBMs.

4. Conclusion

The results obtained in this project will possibly help to answer a few essential questions about gene regulation. First, it was unclear to what extent TF binding is due to direct sequence recognition. Our observations strongly suggest that the contribution of sequence specificity has been vastly underestimated. By using quantitative PBM experimental data, we were able to demonstrate that sequence specificity is sufficient to distinguish between bound and unbound regions of E2F1. Second, these results suggest a method for direct comparison between *in vitro* and *in vivo* TF binding data. This comparison is important for understanding gene regulation because of the technical difficulties in obtaining high quality *in vivo* data. A very large number of cells are required for performing a ChIP experiment. This disadvantage is especially problematic for the E2F family of factors because it is known that they are regulators of the cell cycle; however, unless the cells are synchronized, the ChIP data represents an average of the whole population of cells, which are at different stages of the cell cycle. Synchronization of such large cell cultures is inefficient and such an experiment has still not been performed. Consequently, predicting E2F binding *in vivo* will help us understand the way it regulates its targets and so will advance our knowledge of cell division. Finally, these results provide the first demonstration of competition between paralogous TFs and successful application of a quantitative competition model in discriminating between unique targets of E2F family members. While some studies have explored competition between unrelated factors, competition between paralogous ones has been overlooked, resting on the assumption that they recognize the same sequences and their function is redundant. Our data shows that this is not true and, in the case of E2F, competition is necessary and sufficient to explain the *in vivo* preferences of these paralogous factors, despite their very small differences in specificity. The next step in studying the competition between factors of the same family is to determine the effect of varying their relative concentration. This is a biologically important question especially for the E2F family because these TFs are expressed at different times in the cell cycle and they have non-overlapping function.

References

1. Kim J, Lee J-h, Iyer, *Global Identification of Myc Target Genes Reveals Its Direct Role in Mitochondrial Biogenesis and Its E-Box Usage In Vivo*. PLoS ONE, 2008.

2. Stormo, G. D. *DNA binding sites: representation and discovery*. Bioinformatics, 2000.

3. Berg OG, von Hippel PH, *Selection of DNA binding sites by regulatory proteins : Statistical-mechanical theory and application to operators and promoters*, Journal of Molecular Biology, 1987.

4. Siddharthan R, *Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix*. PLoS ONE, 2010.

5. Bulyk ML, et al., *Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors.* Nucleic Acids Res. *2002.*

6. Man TK, Stormo GD*., Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay.* Nucleic Acids Res. 2001.

7. Gordân R, Shen N, Dror I, Zhou T, Horton J, Rohs R, Bulyk ML *Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape*, Cell Reports 3(4):1093–1104.

8. Berger MF, Philippakis AA, Qureshi A, He FS, Estep PW 3rd, Bulyk ML. *Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities*, Nature Biotechnology, 2006.

9. Bulyk M.L.**,** *Protein binding microarrays for the characterization of DNA-protein interactions*., Adv Biochem Eng Biotechnol, 2007.

10. Aparicio, O., Geisberg, J. V., Sekinger, E., Yang, A., Moqtaderi, Z. and Struhl, K., *Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences In Vivo*. Current Protocols in Molecular Biology, 2005.

11. Zhang et al., *Model-based Analysis of ChIP-Seq (MACS)*. Genome Biol, 2008.

12. Rhee, Ho Sung; BJ Pugh, *Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution*, Cell, 2011.

13. Galas D.J. and Schmitz A., *DNAse footprinting: a simple method for the detection of protein-DNA binding specificity*, Nucleic Acids Res., 1978.

14. Pique-Regi et al., *Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data*, Genome Res., 2011.

15. Luo K. and Hartemink A., *Using DNase digestion data to accurately identify transcription factor binding sites*, Pac Symp Biocomput., 2013.

16. Hesselberth et al., *Global mapping of protein-DNA interactions in vivo by digital genomic footprinting*, Nat Methods, 2009

17. Munteanu A. and Gordan R., *Distinguishing between Genomic Regions Bound by Paralogous Transcription Factors*, Lecture Notes in Computer Science Volume 7821, 2013.

18. Slansky J.E., Farnham P.J. *Introduction to the E2F family: Protein structure and gene regulation*. Springer-Verlag; New York: 1996.

19. Rabinovich A. et al., *E2F in vivo binding specificity: Comparison of consensus versus nonconsensus binding sites*, Genome Res., 2008.

20. Xu, X., Bieda, M., Jin, V.X., et al.: *A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members*. Genome Research, 2007.

21. Wu, Z., Zheng, S., Yu, Q.: *The E2F family and the role of E2F1 in apoptosis*. Int. J., 2009.

22. Attwooll C. et al., *The E2F family: specific functions and overlapping interests*, EMBO J., 2004.

23. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS, *Origins of specificity in protein-DNA recognition,* Annu. Rev. Biochem. 2010

24. Cao AR, Rabinovich R, Xu M, Xu X, Jin VX, Farnham PJ. *Genome-wide analysis of transcription factor E2F1 mutant proteins reveals that N- and C-terminal protein interaction domains do not participate in targeting E2F1 to the human genome*, J Biol Chem., 2011.

25. Matys, V., Kel-Margoulis, O.V., Fricke, E., et al.: *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes*. Nucleic Acids Research 34, D108–D110 (2006)

26. ENCODE Project Consortium, Bernstein, B., Birney, E., Dunham, I., Green, E., Gunter, C., Snyder, M.: An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012)

27. Todd Wasson and Alexander Hartemink, *An ensemble model for multi-factor binding to the genome*, Genome Research, 2009.

28. Sebastian J. Maerkl and Stephen R. Quake, A systems approach to measure the binding energy landscapes of transcription factors, Science, 2007.