



Published in final edited form as:

Nat Commun. ; 6: 7440. doi:10.1038/ncomms8440.

Context Influences on TALE-DNA Binding Revealed by Quantitative Profiling

Julia M. Rogers^{1,2,*}, Luis A. Barrera^{1,4,*}, Deepak Reyon^{5,8}, Jeffrey D. Sander^{5,8}, Manolis Kellis⁴, J. Keith Joung^{5,8}, and Martha L. Bulyk^{1,3,9,2}

¹Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

²Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138, USA

³Harvard-MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115, USA

⁴Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts 02139, USA

⁵Molecular Pathology Unit, Massachusetts General Hospital, Charlestown, MA 02129, USA

⁶Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, MA 02129, USA

⁷Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA 02129.

⁸Department of Pathology, Harvard Medical School, Boston, MA 02115, USA

⁹Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

Abstract

Transcription Activator-Like Effector (TALE) proteins recognize DNA using a seemingly simple DNA-binding code, which make them attractive for use in genome engineering technologies that require precise targeting. While this code is used successfully to design TALEs to target specific sequences, off-target binding has been observed and is difficult to predict. Here we explore TALE-DNA interactions comprehensively by quantitatively assaying the DNA binding

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

Correspondence should be addressed to M.L.B. (mlbulyk@receptor.med.harvard.edu).

*These authors contributed equally to this work.

AUTHOR CONTRIBUTIONS

M.L.B., J.M.R. and L.A.B. designed the study, D.R. and J.D.S. assembled TALEs, J.M.R. and L.A.B. designed custom oligonucleotide arrays, J.M.R. cloned and expressed TALEs and performed PBM experiments, J.M.R. and L.A.B. performed data analysis, L.A.B. developed and fitted models, performed statistical analyses, and created the SIFTED web tool, M.K., J.K.J. and M.L.B. supervised research, J.M.R., L.A.B. and M.L.B. wrote the manuscript.

Additional Information

All analyzed microarray data and array designs have been deposited in NCBI GEO under Series ID 56978.

Competing financial interests: J.K.J. is a consultant for Horizon Discovery. J.K.J. has financial interests in Editas Medicine, Hera Testing Laboratories, Poseida Therapeutics, and Transposagen Biopharmaceuticals. J.K.J.'s interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies.

specificities of 21 representative TALEs to ~5,000-20,000 unique DNA sequences per protein using custom-designed protein binding microarrays (PBMs). We find that protein context features exert significant influences on binding. Thus, the canonical recognition code does not fully capture the complexity of TALE-DNA binding. We used the PBM data to develop a computational model, Specificity Inference for TALE-Effector Design (SIFTED), to predict the DNA-binding specificity of any TALE. We provide SIFTED as a publicly available web tool that predicts potential genomic off-target sites for improved TALE design.

Introduction

The discovery of Transcription Activator-Like Effector (TALE) proteins has enabled the development of a host of genome and epigenome editing technologies¹⁻⁸. Naturally occurring as bacterial virulence factors, TALE proteins harbor an array of repeats, each 33 or 34 amino acids in length^{9,10}. The sequence of the repeats is highly conserved except at the hypervariable positions 12 and 13, termed the repeat variable diresidues (RVDs). The amino acids at the RVD positions determine which DNA base is preferred, and each repeat in the TALE contacts one base in the target site. This led to a simple one-to-one “TALE code” that uniquely predicts the optimal DNA target from the sequence of RVDs within the repeat array^{9,10}. The most commonly used RVDs are NI, HD, NN, and NG, used to target A, C, G, and T, respectively. Co-crystal structures have shown the mechanism of this one-to-one code, in which the TALE protein wraps around the DNA in a helical structure with each repeat contacting a single base^{11,12}. Additionally, contacts between the N-terminal region (NTR) of the TALE protein and DNA specify a preference for a thymine base 5' to the DNA target site¹³.

This simple TALE recognition code allows for any DNA site preceded by a T to be targeted by a TALE protein designed with the corresponding repeat sequence. Therefore, the TALE DNA binding domain has been adapted for use in many technologies that require precise targeting of genomic loci. For example, dimeric TALE nucleases (TALENs) have been used in various organisms and cell lines to knock out genes by the introduction of indels or to create specific mutations². Fusions of TALE monomers to transcriptional activation or repression domains can create artificial transcription factors, which have been shown to strongly and cooperatively modulate gene expression^{4,6,8}. Monomeric TALE fusions to chromatin-modifying enzymes can introduce specific DNA or histone modifications at target loci, resulting in changes in expression of the associated genes^{3,5}. TALEs can also be used to pull down specific genomic regions to identify bound proteins¹. Additionally, TALEs fused to fluorescent proteins can be used to visualize chromatin dynamics in live cells^{1,7}. While other technologies, (*e.g.*, CRISPR-Cas9) have also been developed for some of these targeting applications¹⁴, TALE versus dCas9 fusions might be more effective in different applications and having both technologies in the toolkit for genome engineering is likely optimal.

Despite these successes in genome editing, off-target activities of TALE fusions have been described but have proven difficult to predict¹⁵⁻²¹. Experimental approaches have identified off-target TALEN effects²⁰, but no technology has directly measured off-target binding for

Author Manuscript

Author Manuscript

Author Manuscript

monomeric TALE fusions^{15-17,22}. Here, we define TALE protein specificity as the relative binding energies of the protein to different DNA sequences. Computational tools that use the specificities of the individual repeats to predict the specificity of the whole protein have been developed to predict off-target binding sites^{23,24}; these approaches assume that each repeat independently contributes to the specificity of the whole protein and that each instance of a given repeat RVD type has the same preference for its intended base. However, a quantitative analysis of TALE affinity indicated that repeat position within the repeat array affects RVD specificity, indicating a potential role for repeat context in predicting specificity²⁵. Other studies have also found that total protein length affects specificity²⁰. Additionally, particular repeat types may contribute differentially to overall protein specificity. One study showed that some repeats are more active when assembled into a TALE activator, leading to the distinction between strong (NN and HD) and weak (NI and NG) repeats, although the relationship between RVD strength and specificity is unclear²⁶. Altogether, these findings suggest that TALE-DNA binding specificity may be more complex than previously thought, but these effects have yet to be assayed comprehensively and quantitatively.

Tools used to predict TALE specificity and to identify likely genomic targets have not kept pace with these increasing, albeit qualitative, reports on TALE-DNA recognition. Some computational tools, such as PROGNOS and Talvez, have incorporated context effects qualitatively in predicting TALEN pair off-target sites, but assume all repeat types are affected identically by context^{27,28}. A recently described approach used a selection-based cleavage assay to characterize a TALEN pair's specificity profile in order to identify potential TALEN off-target sites; however, that study did not provide a predictive model, but instead required that the specificity of each TALEN pair be determined experimentally²⁰. As such, there remains a need for a purely computational tool that quantitatively incorporates these context effects in predicting TALE specificity, and thus, off-target binding sites.

Author Manuscript

Author Manuscript

In this study, we perform a quantitative, in-depth examination of context effects on RVD specificity in order to infer general rules for highly accurate prediction of the DNA sequence-specificity of any TALE protein. We design custom protein binding microarrays (PBMs) to investigate the DNA binding specificities of 21 TALE proteins that comprise all possible pairs of repeat types. The custom PBMs contain probes in which all possible mono- and di-nucleotide substitutions within the TALE target sites are represented. The resulting quantitative binding data for the TALE proteins to ~20,000 unique DNA sequences allows us to quantify the effects of TALE repeat array length, repeat position, and neighboring repeat types on the specificity of each RVD, henceforth referred to as RVD specificity. We use the PBM-derived quantitative binding data to develop a computational model (Specificity Inference for TAL-Effector Design or SIFTED) that incorporates these context effects to predict both the DNA binding specificity and the potential off-target sites of any TALE protein without requiring any additional PBM experiments. We implement this model in a publicly available, user-friendly suite of web tools at <http://thebrain.bwh.harvard.edu/sifted.html>.

Results

Custom-designed PBMs to assay TALE DNA-binding specificity

In order to develop a more in-depth, quantitative understanding of TALE-DNA recognition, we determined the DNA-binding specificities of 21 representative TALE proteins using custom-designed PBMs²⁹⁻³¹ (Fig. 1a, Fig. 1b, Supplementary Table 1, Supplementary Data 1). We selected these proteins to allow us to examine the effects of different protein features on specificity. In particular, these proteins represent all possible consecutive repeat pairs and thus allow us to assay all possible direct neighbor effects on RVD specificity (Fig. 1a)³². In addition, this set spans protein lengths from 8.5 to 18.5 repeats (targeting sites 10 to 20 base pairs in length); these lengths typically have been used in the design of monomeric TALE fusion proteins for genomic applications⁴.

PBMs are double-stranded DNA microarrays that permit rapid, highly parallel measurement of the binding of a protein of interest to tens of thousands of unique DNA sequences in replicate, allowing for a much richer picture of TALE-DNA recognition than has resulted from prior studies. Since the vast majority of our selected TALE proteins were designed to recognize sequences longer than those on the previously designed ‘all 10-mer’ universal PBM design³⁰, we designed custom TALE-PBMs for this study. Each probe sequence was represented on at least eight replicate spots on the arrays. The initial custom array was designed to broadly assay the binding preferences of our representative set of TALE proteins. Subsequently, additional arrays were designed to validate particular observations about TALE specificity, as discussed below (Supplementary Fig. 1; Supplementary Note 1).

We determined the DNA binding specificities of each TALE protein using probe sets that contain each protein’s target site as predicted by the canonical TALE code⁹, as well as variants thereof, flanked by constant DNA sequence and situated at a fixed position within the probe relative to the slide surface (Supplementary Note 1). The constant flanking sequence was predicted to not be bound by any of the TALEs tested in this study. For each protein, we measured binding to between 160 and 320 variant target sites that cover all possible adjacent dinucleotide substitutions. Although the absolute K_d of a protein-DNA interaction cannot be determined from a single PBM experiment³³, by measuring how much each substitution changes protein binding to the DNA probe, we can infer changes in binding free energy (ΔG values) for each possible substitution within the target site.

From these ΔG values, we derived a position weight matrix (PWM) for the protein (Fig. 2a; Supplementary Fig. 2). The inferred PWMs were consistent across experimental replicates and across PBM experiments performed at different concentrations of TALE proteins (Supplementary Fig. 3). Our PWMs accurately predict the 60-base-pair probe signal intensities, with a median R^2 of 0.959 (Fig. 2b; Supplementary Fig. 4), indicating that they perform well as accurate models of TALE DNA-binding specificity.

The fact that our PWMs explain binding well suggests that an additive binding model with independence between the nucleotides in the TALE target site is quite accurate. To test if this nucleotide independence extends beyond two adjacent mismatches, we designed a probe set that contains up to five nonadjacent mismatches in the target site (Supplementary Note 1,

Supplementary Fig 1). The PWM models derived from the dinucleotide substitution probes accurately predicted binding to these sequences with additional mismatches (median R^2 greater than 0.9 for all numbers of substitutions tested), indicating that the simple PWM models with mononucleotide independence perform well in modeling TALE DNA-binding specificity (Supplementary Fig. 5). These results are roughly consistent with a recent study of TALEN pair specificity determined by a selection-based cleavage assay, in which general independence in DNA recognition was observed; however, our data support a fully independent model of TALE-DNA binding, rather than a model with slightly increased tolerance for adjacent mismatches²⁰.

Modeling repeat context improves specificity prediction

Although we observed mononucleotide independence within TALE target sites, we found that the protein-DNA interactions of a given repeat are influenced by its context. In other words, the energetic parameters of a given TALE-DNA contact are not affected by neighboring nucleotide changes, but they are affected by the repeat context. Intriguingly, even within a single TALE protein, different occurrences of the same repeat type can exhibit very different specificities. For example, in TAL2009, repeats 7 and 10 were both designed with the HD RVD to target C, but within the context of the TAL2009 protein each exhibits substantially different relative preferences for C as compared to other nucleotides (Figure 2a). Typically, the highest scoring probe corresponded to the target sequence predicted by the canonical TALE code; however, we observed multiple cases (*e.g.*, TAL2024) where a TALE protein bound mismatched sequences with comparable binding strength, hereafter referred to as affinity. Moreover, some TALEs (*e.g.*, TAL2009) even preferred a mismatched sequence to the predicted optimal target sequence; this most frequently involved an NN RVD, which can target both a G and an A in different contexts (for example, see repeats 3 and 6 in Fig. 2a)¹⁰. Altogether, these results highlight that the simple one-to-one TALE code is not sufficient to accurately predict DNA binding specificity

Since our results suggested that interactions between repeats modulate their individual RVD specificities, we modeled the PBM data to predict TALE specificity considering the context of each repeat in a TALE protein (Fig. 1c). We named our model and its associated software tools SIFTED (Specificity Inference For TAL-Effector Design). In addition to modeling the intrinsic specificity of each RVD, SIFTED considers a variety of repeat context features, including the number of repeats in the protein, each repeat's position within the repeat array, and the immediately adjacent N- and C-terminal neighboring repeat types. The NTR, which specifies the preference for the 5' T in the binding site, was also included in the model and was treated equivalently to a repeat, except for the omission of its position and length features.

We trained the SIFTED model by performing a linear regression with Elastic Net regularization, using the ΔG values inferred for each protein as the input data³⁴. To prevent overfitting and to assess performance, we used a nested leave-one-out cross-validation strategy. Briefly, one protein was held out from the dataset in an iterative fashion. The remaining proteins were divided into training and test sets, which were used to derive parameter values and to control the complexity of the model (Supplementary Fig. 6). The

predicted PWM for each of the 21 TALE proteins was obtained from the model trained on data from the remaining 20 proteins in our dataset (Fig. 1a). For specificity predictions of proteins not in our dataset (*e.g.*, TALEN pairs), the regression was performed on the full dataset (no proteins excluded) and the resulting model was used to make PWM predictions.

To assess how well our model explains binding, we used the PWMs obtained from the cross-validated SIFTED model to predict PBM probe signal intensities. The SIFTED PWMs accurately predict the probe-level PBM binding data (median $R^2 = 0.877$). Additionally, SIFTED outperformed the specificity models from other available computational tools designed to predict off-target sites in explaining the PBM data ($P < 10^{-6}$, Wilcoxon signed-rank test) (Fig. 3a). Two of these tools, TALE-NT 2.0²³ and TALgetter²⁴, do not consider any context effects. Others, such as PROGNOS²⁷ and Talvez²⁸, include context effects on an RVD's specificity only as discrete penalties. In contrast, SIFTED models context effects quantitatively and also allows each repeat type (*i.e.*, NI, HD, NN, and NG) to be influenced differently by its context. These detailed context parameters in our model are keys to its success; the full model predictions from SIFTED are more accurate ($P < 10^{-6}$, Wilcoxon signed-rank test) than those of an RVD-only model that represents the canonical, one-to-one TALE-DNA recognition code (median $R^2 = 0.798$) (Fig. 4).

We validated that our SIFTED model can predict off-array binding affinity measurements (K_d values) more accurately than other published tools³⁵ (Fig. 3b). While PWMs cannot be used to predict absolute dissociation constants, they are able to predict the affinity of a sequence relative to that of the optimal binding site (*i.e.*, relative K_d values)³⁶. The full SIFTED model performed significantly better than PROGNOS, TALE-NT 2.0, TALgetter, Talvez, or a reduced SIFTED model with no context effects in predicting relative K_d values for one protein and 18 DNA sequences³⁵.

Quantitative modeling of context effects on RVD specificity

Since context effects contributed significantly to the predictive power of our model, we investigated in greater depth how length, position, and neighboring repeats each affect specificity. While our baseline RVD specificities (Fig. 5a) largely agree with previous studies⁹ (*e.g.*, NN is the least specific RVD and can target both G and A), in the SIFTED model these specificities are modulated by the protein context of each instance of the repeat.

Our data are consistent with previous reports that longer proteins tolerate more mismatches in their target sites²⁰ (Figure 5b). Our comprehensive profiling also revealed that NN and NG repeats are affected more strongly by protein length than are either NI or HD. Additionally, our set of proteins included two proteins of different lengths designed to target overlapping sites. The longer protein (TAL2073) is less specific overall (*i.e.*, lower total information content) than the shorter protein (TAL2043) (Supplementary Fig. 2), directly supporting our overall finding that increased TALE protein length diminishes RVD specificity.

Repeat position within the repeat array also affects the specificity of C-terminal repeats that target the 3' end of the DNA binding site, resulting in their being more tolerant to substitutions than N-terminal RVDs. To test this modeling result, we designed a custom

PBM that included probes containing clusters of three nucleotide substitutions located at either the 5' or 3' end of the target site (Supplementary Fig. 1, Supplementary Note 1). In general, substitutions at the 5' end impaired binding more than substitutions at the 3' end ($P < 0.05$, Wilcoxon signed-rank test) (Supplementary Fig. 7), supporting prior observations from reporter assays^{25,37}. Talvez and PROGNOS model this polarity effect discretely as a constant decrease in specificity after a certain position in the repeat array for all repeat types^{27,28}. In contrast, SIFTED continuously models the decrease in specificity over the length of the protein and allows different repeat types to be affected to different extents (Fig. 5b).

Lastly, we observed that a repeat's specificity is impacted by the identity of the immediately adjacent N- or C- terminal repeat (Fig. 5c). Such local context effects previously have been observed only for the 5' T preference, which is more important for binding when the first repeat is an HD³⁸. We also observed the influence of HD in the first position, but found an even stronger effect when the first repeat is an NN. Additionally, we observed neighbor context effects between repeats within the protein. For example, the NN repeat is more specific for targeting a G when the NI repeat is either N- or C-terminal to it; however, it is much less specific for G when it is positioned at the C-terminal end of a TALE repeat array.

We found that a particular repeat type can exert different effects as an N- or C- terminal neighbor (Fig. 5c). PROGNOS includes a parameter to reduce an RVD's specificity when it is next to a strong RVD (NN or HD), positing that a stronger neighboring interaction may allow for greater mismatch tolerance^{26,27}; however, it does not distinguish between N- and C- terminal neighbors. The neighbor effects we found are more complex, and in fact, the strong RVDs do not always decrease specificity. The complexities of the neighboring effects are captured quantitatively in SIFTED; each of the four RVDs as well as the 5' T preference are modeled as being affected differently by its N- and C- terminal neighboring repeats.

These observations of context effects can be condensed into some simple guidelines for TALE design (Table 1). Certain repeat combinations (*e.g.* NI-NI) are predicted to have increased specificity, while others (*e.g.* NG as the N-terminal repeat) can make an RVD more tolerant to mismatch and therefore should be avoided. However, when designing TALE proteins, one must ultimately consider all the context effects in the protein, as well as the prevalence of potential off-target sites in the genome. As such, we tested if the SIFTED model could accurately predict genomic off-target sites, and therefore could be used to guide TALE protein design.

Predicting TALE off-target sites using SIFTED

To assess whether SIFTED can predict genomic off-target sites for TALE proteins that have not been assayed by PBMs, we examined a dataset of *in vivo* TALE reporter activity²². SIFTED had the highest median performance of the five tools tested (Fig. 3c).

Although SIFTED was designed to predict TALE monomer specificity, we also tested its ability to predict TALEN binding by examining a large dataset of TALEN activity in cells²⁰. We derived the specificities of TALEN pairs from the specificities of the component monomers predicted by SIFTED. The PWMs from SIFTED resulted in better sensitivity and

specificity than those from any of the other models in distinguishing genomic target sites that showed nuclease activity from those that did not (Fig. 3d, Supplementary Data 2). The area under the receiver operating characteristic (AUROC) curve statistic was used to quantify the ability of the five tools to distinguish target from non-target sites across all possible score thresholds. SIFTED demonstrated superior sensitivity and specificity across most thresholds.

Additionally, we considered that a typical TALE user might investigate about 20 off-target sites when analyzing the specificity of their designed protein in their genome of interest. To provide a performance comparison for this typical use case, we investigated how many of the top 20 off-target sites predicted by these tools have been identified as TALEN pair off-targets *in vivo* (Supplementary Fig. 8a, Supplementary Fig 8b), or were among the 20 off-targets with the highest measured *in vivo* activity (Supplementary Fig 8c). Again, SIFTED performed better than the other tools, demonstrating higher sensitivity by predicting more of the true off-targets than the other tools (Supplementary Fig. 8b, Supplementary Data 2).

Prediction of genomic off-targets with SIFTED web tool

SIFTED was the top-performing model overall, highlighting the value of incorporating repeat context effects in predicting specificity. While other tools may perform comparably to SIFTED in a specific application, SIFTED was the only tool that was consistently a top performer across the wide range of benchmarks of predictive performance (Fig. 3, Supplementary Fig. 8). Given the success of SIFTED in predicting off-target binding, we developed it into a web-based suite of tools to aid in TALE design implemented on the Galaxy platform³⁹⁻⁴¹ at <http://thebrain.bwh.harvard.edu/sifted.html>. We provide stand-alone tools for individual tasks, such as predicting the specificity and genomic binding sites of a user-specified TALE, as well as a pipeline that combines various tools to automate the process of designing a TALE to target a particular genomic region. The complete pipeline takes a user-defined genomic target region as input, and then (1) identifies candidate TALEs to target that input region, (2) predicts the candidates' specificities, (3) finds instances of off-target sites in a user-specified genome and (4) outputs a list of candidate TALE proteins ranked by their off-target binding potential, thus allowing the user to select the best candidate protein.

Discussion

By analyzing TALE proteins of different lengths and containing all possible consecutive pairs of repeats, we were able to identify the influence of repeat context on DNA binding specificity. In contrast to other studies that used cell-based TALEN activity as a measurement of TALE specificity³⁵, our experimental design allowed us to directly assay the intrinsic binding properties of TALE monomers. We measured a total of ~200,000 binding interactions between 21 TALE proteins and ~5,000-20,000 unique DNA sequences per protein using custom-designed PBMs. Importantly, the resulting dataset allowed us to develop a model to predict TALE specificity for any candidate TALE protein without requiring any additional experimental analysis.

Our results highlight that RVD specificity is not determined simply by what base a particular RVD will bind, but also which bases it strongly disfavors. This information could be useful in designing TALEs for allele-specific applications, such as rapid, spatially resolved genotyping of patient samples through binding of fluorescently tagged, allele-specific TALEs. The HD RVD has the greatest power to discriminate between two alleles: it prefers binding to a C and strongly disfavors binding to a G. Therefore, targeting an allele where there is a C/G SNP may lead to stronger discrimination between the two alleles.

We found that longer TALEs are generally less specific than shorter TALEs. This effect could be due to excess DNA binding energy in TALE proteins with many repeats²⁰. The mechanism of the context effects on RVD specificity remains to be determined. An ability to tolerate some binding site mismatches may allow a TALE protein from xanthomonad pathogens to overcome mutations in host genomic target sites, as the plant host may be under selection to escape xanthomonad infection. However, TALEs with very low specificity may lead to potential negative effects on virulence due to additional binding in the host genome⁴². Thus, the specificity of TALE proteins may have been strongly shaped by the complex interactions between host and pathogen.

SIFTED predicts that some DNA sequences should be targeted with greater specificity, which could be interpreted as guidelines for TALE design (Table 1). Interestingly, some of these guidelines would contradict published guidelines that were developed as part of the SAPTA tool for designing more active TALEN pairs⁴³. For example, while we predict that A-runs can be targeted with high specificity by TALE monomers, SAPTA predicts that TALENs targeting A-runs will have lower nuclease activity. The discrepancies in these guidelines and results might reflect different rules affecting the binding of monomeric TALEs versus dimeric TALENs. Alternatively, it is possible that a trade-off exists between optimizing activity and specificity in designing TALENs. Previous reports have found no correlation between activity and affinity³⁵. This lack of correlation between *in vitro* binding and different cell-based activity measurements might be due to other genomic features in cells, such as the chromatin state and competition with other transcription factors at the target and off-target sites. Ultimately, in designing TALEs, the intrinsic specificity of the protein must be considered in light of its potential off-target binding sequences in the genome. For example, the decreasing specificity of longer TALEs may be compensated by longer target sites being more rare in the genome, thus increasing the effective specificity of a protein²⁰. SIFTED can both model protein specificity as well as identify genomic off-target sites, revealing the effective specificity of a TALE, so users can choose the most specific TALE protein for their particular application.

Future studies will be required to identify chromatin features that might modulate binding specificity *in vivo*. Additionally, the specificities of other alternative RVDs (*e.g.*, NH to target G) could be studied to enable design of TALE proteins with higher sequence specificity. An improved understanding of TALE-DNA binding should allow for development of more precise genome engineering tools.

METHODS

Cloning of TALE proteins

TALEN expression vectors³² were digested with SacII and BamHI to obtain the DNA-binding domain comprising the 152 N-terminal domain (NTD), the RVD repeats, and the +63 C-terminal domain (CTD). This fragment was ligated into a modified pDONR221 vector (Invitrogen), with SacII and BamHI restriction sites internal to attL recombination sites, to create Gateway-compatible TALE Entry clones. The TALE constructs were then transferred by Gateway recombinational cloning into the pDEST15 expression vector, which adds an N-terminal glutathione S-transferase (GST) tag (Invitrogen), by an LR reaction. All clones were full-length sequence-verified (Supplementary Data 1).

Custom PBM Design

Target sites for each TALE protein were determined using the canonical TALE code (NI: A, HD: C, NN: G, NG: T), and are preceded by the 5' T to create the full target site. The constant flanking region was the same as that used in a prior custom PBM design and does not contain binding sites for any of the TALE proteins in this study⁴⁴. Probe set descriptions, including the array design versions on which they are included, are provided in Supplementary Note 1.

Protein binding microarray experiments

Proteins were expressed using the PURExpress *In Vitro* Transcription and Translation Kit (New England Biolabs). Protein concentrations were determined by anti-GST Western blots with a dilution series of recombinant GST (Sigma). Proteins were stored at +4 °C until being used in protein binding microarray (PBM) assays. PBMs were performed as follows:²⁹. Briefly, custom designed microarrays were first double-stranded by an on-slide primer extension reaction. In the PBM assay, arrays were blocked with 2% milk in phosphate buffered saline (PBS) for one hour, washed with 0.1% Tween-20 and 0.01% TX-100, then incubated with protein mixture (PBS, 2% milk, 0.2 mg ml⁻¹ BSA, and 0.3 µg ml⁻¹ salmon testes DNA) for one hour. The final concentration of TALE protein in the PBM binding reactions was 200 nM, unless otherwise indicated (Supplementary Table 1). Arrays were washed with 0.5% Tween-20 and 0.01% TX-100. Lastly, the array was incubated for 30 minutes with an Alexa488-conjugated anti-GST antibody (Invitrogen A-11131), and washed with 0.05% Tween and PBS..

PBM data quantification

PBM arrays were scanned using a GenePix 4400A Microarray Scanner (Molecular Devices), and scan images were analyzed by GenePix Pro (Molecular Devices). Raw data files were processed using the same general approach as used for universal PBMs²⁹. Briefly, masliner software⁴⁵ was used to combine Alexa488 scans at three different laser power levels and resolve the signal intensity in spots that are saturated at high laser power settings. Cy3 scans were performed at a single laser power level. If a dataset had any negative background-subtracted intensity (BSI) values (which can occur if the region surrounding a spot is brighter than the spot itself), a pseudocount was added to all BSI values for that

experiment such that all values were then positive. The adjusted BSI data were then normalized by the corresponding double-stranded DNA content of the spots and their position on the array using the same approach as described for universal PBMs²⁹. In order to normalize by the relative amount of double-stranded DNA per array spot, small quantities of Cy3-dUTP were added to the nucleotide pool during the double-stranding process. The BSIs on the Cy3 channel can therefore be used to estimate relative DNA abundance at each spot. However, because Cy3 incorporation depends on the local sequence context, we used a linear regression over the trinucleotides present in a given probe to calculate the expected Cy3 BSI and obtain the expected-to-observed ratio³⁰. This ratio is then used to normalize the Alexa488 BSIs to account for difference in relative amounts of double-stranded DNA. Any probes with BSIs that were corrected by more than 2-fold or for which the adjustment would lead to a negative BSI value were removed from the data.

All PBM designs include at least 8 replicate probes for each sequence. For each experiment and for each set of probes with identical sequences, we calculated the median adjusted BSI, median absolute deviation (MAD), and the robust standard deviation estimate from the MAD. Any individual replicate probe with a normalized adjusted BSI value more than 3 standard deviations away from the median of the replicate probes was omitted from subsequent analysis, to avoid confounding statistical tests or incorrect choice of parameter settings in model fitting.

For each TALE protein we defined a background set of probes that comprises all the probes on the array designed to represent binding sites for other TALE proteins (not the one being assayed in a given experiment). The array median level was then calculated as the median normalized adjusted BSI of all probes in the background set. The standard deviation of the background set SIs was calculated robustly using the asymptotic approximation $\sigma = 1.4826 * MAD$. The z-score for each probe was calculated relative to the median and standard deviation of its corresponding background probes. These z-scores represent a linear transformation of the median SIs for each probe, and therefore facilitate interpretation but do not affect the PWM fitting procedure, which performs its own linear scaling adjustments.

Position weight matrix model fitting

We developed a Bayesian Markov chain Monte Carlo (MCMC) method to infer free energy parameters of TALE-DNA interactions from PBM data. We relied on the theoretical framework developed for the BEEML-PBM algorithm³⁶, which can accurately derive G values for protein-DNA contacts from universal PBM experiments. The BEEML-PBM framework estimates G values for each possible nucleotide substitution in a protein's DNA binding site motif. These values can be assembled to construct an energy matrix (EM), in which each column represents a position within the binding site and each row represents a nucleotide. The EM values can be converted to probabilities using the Boltzmann distribution, creating a position weight matrix (PWM).

Briefly, the goal is to predict the observed probe signal intensity z-scores as a function of the binding site sequence within the probe. As an intermediate step, the G values are used to predict occupancy of the TALE protein on its binding site. The predicted occupancy is then scaled linearly to optimally scale with the observed z-scores. The chemical potential

($\log([TF]/K_d)$) was also included in the model and can account for differences in TALE protein concentration and affinity to the optimal binding site. The statistical model is described in full in Supplementary Methods. At each sampling step, the probe z-scores are predicted given the current parameter values, which can be used to derive 95% credible intervals⁴⁶, as shown in Fig. 2a. The priors on ΔG values were set as exponential distributions with mean 10.0, to cause the preferred base to adopt values close to 0 but to not significantly penalize larger parameter values for other bases. The rest of the parameters were given a uniform prior.

To perform Markov chain Monte Carlo (MCMC) sampling, we used the No-U-Turn Sampler (NUTS) included in Rstan v2.0 with default parameter settings. The ΔG parameters were initialized following a simple TALE code: $\Delta G = 0.0$ for the predicted optimal base at a given position, $\Delta G = 3.0$ otherwise (in units of kT/RT). For each dataset, we obtained 500 parameter samples in the burn-in period followed by 2000 samples that were used to approximate the posterior distributions of all parameters. Four MCMC chains were run in parallel for each dataset; the samples from each chain were then used to verify convergence of all ΔG parameters (Gelman-Rubin convergence statistic for all four chains < 1.05). Note that sampling is more efficient in Hamiltonian MCMC methods (such as NUTS) and thus fewer iterations are required than in standard MCMC methods, such as Gibbs sampling⁴⁷.

SIFTED predictive model for ΔG values

The ΔG values inferred from the TALE PBM experiments were used to train a predictive model using an Elastic Net regression³⁴. The energy term for each inferred protein-DNA contact (i.e., each repeat contacting each of the four possible nucleotides) represents a single observation. However, each column in the energy matrix has only three degrees of freedom, since adding a constant value to all terms does not change the resulting PWM. Therefore, each energy matrix derived from the data was first adjusted by adding a constant value to each column such that the preferred base has a ΔG of exactly 0 (because of the exponential prior described above, these values are already close to 0 when the repeat binds its expected base preferentially). These zero-valued ΔG terms were then removed from the dataset, leaving only the values for the non-preferred bases as input.

The full predictive feature matrix was normalized such that each column had mean zero and unit variance. Numerical features (e.g., total length of the target site) are included directly in the feature matrix. In the case of categorical features (e.g., RVD identity), we created binary indicator variable columns ('dummy variables') representing each potential categorical value. We used regression weights to reduce the biases that could be created by having an unequal number of proteins of different lengths. Each squared error term in the Elastic Net objective function was multiplied by the weight corresponding to that observation. The observations corresponding to a given protein were assigned a weight of $1/(\# \text{ of proteins of the same length})$.

We used the Elastic Net implementation in the glmnet v1.9-5 R package to train our model. The Elastic Net is a regularized regression method that seeks to penalize models that are too complex (i.e., have too many parameters) and thus prevent overfitting³⁴. The Elastic Net

objective combines the penalty terms used in L_1 (or LASSO) and L_2 (or Ridge) regressions. Here, we set the balance between the two terms to 95% L_1 penalty and 5% L_2 penalty, favoring the sparseness of the L_1 method but also keeping some of the advantages of the L_2 method, such as the uniqueness of solutions.

Each G in the dataset is paired with a vector of predictive features to create the feature matrix, in which each row is an independent observation, and each column is a different feature. The features include repeat identity, position, neighboring repeat identity, and total length of the target site. Numerical features (e.g., total length of the target site) are included directly in the feature matrix. In the case of categorical features (e.g., RVD identity), we created binary indicator variable columns ('dummy variables') representing each potential categorical value. To allow for non-linear position and length effects, we also included the natural logarithm of each as a feature. The full predictive feature matrix was normalized such that each column had mean zero and unit variance. We used regression weights to reduce the biases that could be created by having an unequal number of proteins of different lengths. Each squared error term in the Elastic Net objective function was multiplied by the weight corresponding to that observation. The observations corresponding to a given protein were assigned a weight of $1/(\# \text{ of proteins of the same length})$.

To prevent over fitting and to accurately assess the model's performance, we used a cross-validation scheme consisting of two nested levels. On the outer level, we used leave-one-out cross-validation to form a validation set by excluding a single protein in each iteration. Once a protein is excluded, the inner level performs 5-fold cross-validation on the remaining twenty proteins. For 100 different values of the Elastic Net penalty term λ , we calculate the mean-squared error (MSE) on the test set for the model obtained from the training set (Supplementary Fig. 6). For a given value of λ , the average MSE over all test sets is calculated. The λ value that minimizes the overall average MSE is then used for all subsequent predictions. This is achieved by creating bootstrap estimates of the MSE at each value of lambda and picking the simplest (i.e., most penalized) model that performs within one standard deviation of the model with the lowest average MSE (dashed vertical lines in Supplementary Fig. 6). The model associated with the best λ value was then used to make predictions on the protein excluded in the outer loop; the same λ value was used for all training sets. This entire process is repeated for each protein, leading to cross-validated predictions for the entire dataset. These predictions were then used for all model evaluation purposes.

Predicting probe signal intensities and K_d values from PWMs

The predictions of probe signal intensities were obtained using the same mathematical framework as for fitting PWMs (Supplementary Methods). However, in this case the G parameters are known and the only parameters that need to be fitted to predict probe intensities are the chemical potential μ and the scaling terms a and b . To determine these parameters, we used the implementation of the Levenberg-Marquardt algorithm in the SciPy v0.12 package with default convergence parameters. The model parameters were initialized as follows: a = minimum z-score in input data, b = maximum z-score in input data, $\mu = -1.0$. After these parameters were fitted from the observed z-scores, the predicted z-scores were

obtained by using the total ΔG for the binding site in each probe and the fitted variables as input.

In order to validate SIFTED predictions with measured K_d values³⁵, relative K_d values for target and off-target sites were predicted from SIFTED PWMs. Relative K_d values were predicted by setting the K_d of the optimal site to 1. The predicted K_d for off-target sequences were obtained through the equation $e^{-\Delta G/RT}$, where ΔG represents the difference in total free energy between the optimal binding site sequence and the sequence of the off-target site. The measured relative K_d values were similarly adjusted so that the optimal site had a K_d of 1. Because K_d values span many orders of magnitude, the correlation coefficient was computed after taking the natural logarithm of the K_d values, which prevents the calculation from being dominated by the extreme values.

Comparison using PWMs from other tools

PROGNOS, TALgetter, Talvez, and TALE-NT 2.0, the publicly available tools against which we compared SIFTED, do not explicitly provide the user with predicted PWMs^{23,24,27,28}. However, with the exception of TALgetter, each tool uses an internal scoring scheme that is mathematically equivalent to a PWM (i.e., the score for a site represents the sum of an independent score for each nucleotide position). Therefore, in the comparisons with PROGNOS, Talvez and TALE-NT 2.0, we predicted PWMs based on the scheme described by each paper and the associated parameters^{23,27,28}. To predict TALgetter scores, we instead used the downloadable TALgetter software tool to compute log-odds values for all binding site sequences in a given experiment²⁴. These binding scores can then be compared directly to PWM log-odds scores, even if the underlying scoring scheme is distinct. For comparisons using TALEN activity data, we combined the values predicted by PWMs for each TALE in a TALEN pair using the same scoring scheme as PROGNOS²⁷. Here, the scores S are obtained by taking the negative natural logarithm of each value in the PWM, creating a value that becomes larger the more disfavored a particular nucleotide is at a particular position. Then, we compute the ratio of the score S summed over the optimal target site and the score S summed over the potential off-target site being analyzed. These ratios are elevated to an exponent (0.6, as determined to be optimal by Fine *et al.*²⁷). Finally, the partial score for each member of the TALE pair is added to create a final score, as in the equation below.

$$Pair\ score = \left(\frac{S_{left}(optimal\ site)}{S_{left}(target\ site)} \right)^{0.6} + \left(\frac{S_{right}(optimal\ site)}{S_{right}(target\ site)} \right)^{0.6} \quad (1)$$

We analyzed the TALEN target sites reported by Guillinger *et al.*²⁰. We scored each reported target site that contained only NN, NI, HD, and NG RVDs using the TALEN Pair Score derived from the PWMs obtained from SIFTED, PROGNOS and TALE-NT 2.0. We summarized the performance of each tool as a receiver operating characteristic (ROC) curve, which shows the sensitivity and specificity values achieved by each tool when predicting sites that were targeted by the TALEN pairs. The different sensitivity and specificity values represent different Pair Score thresholds, above which a locus is predicted to show evidence

of nuclease activity (indels). We also compared against the TALE activator reported by Mali et al²². All of the reported binding sites up to three mismatches away from the predicted site were scored as described above. These scores were then compared to a normalized expression score (the ratio of barcode tags for that binding site relative to a control experiment) associated with that binding site – TALE combination. Since we expect the relationship between TALE occupancy and expression to be nonlinear, we compared the results using Spearman correlation.

Algorithmic approach of SIFTED web tool

The overall approach of the entire pipeline to identify and score candidate TALEs to target a genomic region is as follows. First, candidate TALE binding sites within the user-input DNA sequence are identified. For each site found, the protein that targets that sequence is determined using the TALE code, and its PWM is predicted. For each protein, the PWM is used to enumerate all putative binding site sequences (both target and off-target sequences) with a relative K_d threshold (by default, set to 10), using a bounded breadth-first search. All genomic instances of the putative binding site sequences are found using a short read aligner (*bowtie*). Finally, a summary score is calculated for each protein that describes the overall number and strength of genomic target sequences. Under default parameter settings (*e.g.*, 13.5 repeat TALE, 1-kb region), the SIFTED pipeline typically identifies optimal TALE candidates within minutes. Additionally, a user can input a TALE with a defined RVD sequence, and SIFTED will predict its specificity and identify potential genomic off-target sites. Tutorials are hosted on the SIFTED website for designing TALEs to target a region, and for predicting the specificity of a pre-designed TALE, and include additional guidelines for setting parameters and troubleshooting. For more details on the algorithmic approach, see Supplementary Methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project was supported in part by National Science Foundation Graduate Research Fellowships to J.M.R. and L.A.B., grant R21 HG007573 from NIH/NHGRI to M.L.B., an NIH Director's Pioneer Award (DP1 GM105378) to J.K.J., and The Jim and Ann Orr MGH Research Scholar Award to J.K.J. We thank Alexandre Palagi and Kian Hong Kock for helpful discussion.

References

1. Byrum SD, Taverna SD, Tackett AJ. Purification of a specific native genomic locus for proteomic analysis. *Nucleic Acids Res.* 2013; 41:e195. [PubMed: 24030711]
2. Joung JK, Sander JD. TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol.* 2013; 14:49, 55. [PubMed: 23169466]
3. Maeder ML, et al. Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat Biotechnol.* 2013; 31:1137, 1142. [PubMed: 24108092]
4. Maeder ML, et al. Robust, synergistic regulation of human gene expression using TALE activators. *Nat Methods.* 2013; 10:243, 245. [PubMed: 23396285]

5. Mendenhall EM, et al. Locus-specific editing of histone modifications at endogenous enhancers. *Nat Biotechnol.* 2013; 31:1133, 1136. [PubMed: 24013198]
6. Cong L, Zhou R, Kuo YC, Cunniff M, Zhang F. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nat Commun.* 2012; 3:968. [PubMed: 22828628]
7. Miyanari Y, Ziegler-Birling C, Torres-Padilla ME. Live visualization of chromatin dynamics with fluorescent TALEs. *Nat Struct Mol Biol.* 2013; 20:1321, 1324. [PubMed: 24096363]
8. Perez-Pinera P, et al. Synergistic and tunable human gene activation by combinations of synthetic transcription factors. *Nat Methods.* 2013; 10:239, 242. [PubMed: 23377379]
9. Moscou MJ, Bogdanove AJ. A Simple Cipher Governs DNA Recognition by TAL Effectors. *Science.* 2009; 326:1501
10. Boch J, et al. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science.* 2009; 326:1509, 1512. [PubMed: 19933107]
11. Mak AN, Bradley P, Cernadas RA, Bogdanove AJ, Stoddard BL. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science.* 2012; 335:716, 719. [PubMed: 22223736]
12. Deng D, et al. Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science.* 2012; 335:720, 723. [PubMed: 22223738]
13. Stella S, et al. Structure of the AvrBs3-DNA complex provides new insights into the initial thymine-recognition mechanism. *Acta Crystallogr D Biol Crystallogr.* 2013; 69:1707, 1716. [PubMed: 23999294]
14. Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science.* 2014; 346:1258096
15. Ding Q, et al. A TALEN genome-editing system for generating human stem cell-based disease models. *Cell Stem Cell.* 2013; 12:238, 251. [PubMed: 23246482]
16. Osborn MJ, et al. TALEN-based gene correction for epidermolysis bullosa. *Mol Ther.* 2013; 21:1151, 1159. [PubMed: 23546300]
17. Tesson L, et al. Knockout rats generated by embryo microinjection of TALENs. *Nat Biotechnol.* 2011; 29:695, 696. [PubMed: 21822240]
18. Hockemeyer D, et al. Genetic engineering of human pluripotent cells using TALE nucleases. *Nat Biotechnol.* 2011; 29:731, 734. [PubMed: 21738127]
19. Mussolino C, et al. A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res.* 2011; 39:9283, 9293. [PubMed: 21813459]
20. Guilinger JP, et al. Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. *Nat Methods.* 2014; 11:429, 435. [PubMed: 24531420]
21. Morbitzer R, Römer P, Boch J, Lahaye T. Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors. *Proc Natl Acad Sci U S A.* 2010; 107:21617, 21622. [PubMed: 21106758]
22. Mali P, et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol.* 2013; 31:833, 838. [PubMed: 23907171]
23. Doyle EL, et al. TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Res.* 2012; 40:W117–122. [PubMed: 22693217]
24. Grau J, et al. Computational predictions provide insights into the biology of TAL effector target sites. *PLoS Comput Biol.* 2013; 9:e1002962. [PubMed: 23526890]
25. Meckler JF, et al. Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Res.* 2013; 41:4118, 4128. [PubMed: 23408851]
26. Streubel J, Blucher C, Landgraf A, Boch J. TAL effector RVD specificities and efficiencies. *Nat Biotechnol.* 2012; 30:593, 595. [PubMed: 22781676]
27. Fine EJ, Cradick TJ, Zhao CL, Lin Y, Bao G. An online bioinformatics tool predicts zinc finger and TALE nuclease off-target cleavage. *Nucleic Acids Res.* 2014; 42:e42. [PubMed: 24381193]
28. Perez-Quintero AL, et al. An improved method for TAL effectors DNA-binding sites prediction reveals functional convergence in TAL repertoires of *Xanthomonas oryzae* strains. *PLoS One.* 2013; 8:e68464. [PubMed: 23869221]

29. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc.* 2009; 4:393, 411. [PubMed: 19265799]
30. Berger MF, et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol.* 2006; 24:1429, 1435. [PubMed: 16998473]
31. Mukherjee S, et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet.* 2004; 36:1331, 1339. [PubMed: 15543148]
32. Reyon D, et al. FLASH assembly of TALENs for high-throughput genome editing. *Nat Biotechnol.* 2012; 30:460, 465. [PubMed: 22484455]
33. Siggers T, Duyzend MH, Reddy J, Khan S, Bulyk ML. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol Syst Biol.* 2011; 7:555
34. Zou HH. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B.* 2005; 67:301–320. T.
35. Juillerat A, et al. Comprehensive analysis of the specificity of transcription activator-like effector nucleases. *Nucleic Acids Res.* 2014; 42:5390, 5402. [PubMed: 24569350]
36. Zhao YS. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol.* 2011; 29 G.D.
37. Garg A, Lohmueller JJ, Silver PA, Armel TZ. Engineering synthetic TAL effectors with orthogonal target sites. *Nucleic Acids Res.* 2012; 40:7584, 7595. [PubMed: 22581776]
38. Schreiber T, Bonas U. Repeat 1 of TAL effectors affects target specificity for the base at position zero. *Nucleic Acids Res.* . 2014
39. Blankenberg D, et al. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol.* 2010:11–21. Chapter 19, Unit 19 10. [PubMed: 20069539]
40. Giardine B, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 2005; 15:1451, 1455. [PubMed: 16169926]
41. Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010; 11:R86. [PubMed: 20738864]
42. Gu K, et al. R gene expression induced by a type-III effector triggers disease resistance in rice. *Nature.* 2005; 435:1122, 1125. [PubMed: 15973413]
43. Lin Y, et al. SAPTA: a new design tool for improving TALE nuclease activity. *Nucleic Acids Res.* 2014; 42:e47. [PubMed: 24442582]
44. Siggers T, et al. Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF-kappaB family DNA binding. *Nat Immunol.* 2012; 13:95, 102. [PubMed: 22101729]
45. Dudley AM, Aach J, Steffen MA, Church GM. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc Natl Acad Sci U S A.* 2002; 99:7554, 7559. [PubMed: 12032321]
46. Lee PM. *Bayesian Statistics: An Introduction.* 1997
47. Hoffman MDG. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. 2011 A. arXiv:1111.4246 [stat.CO].

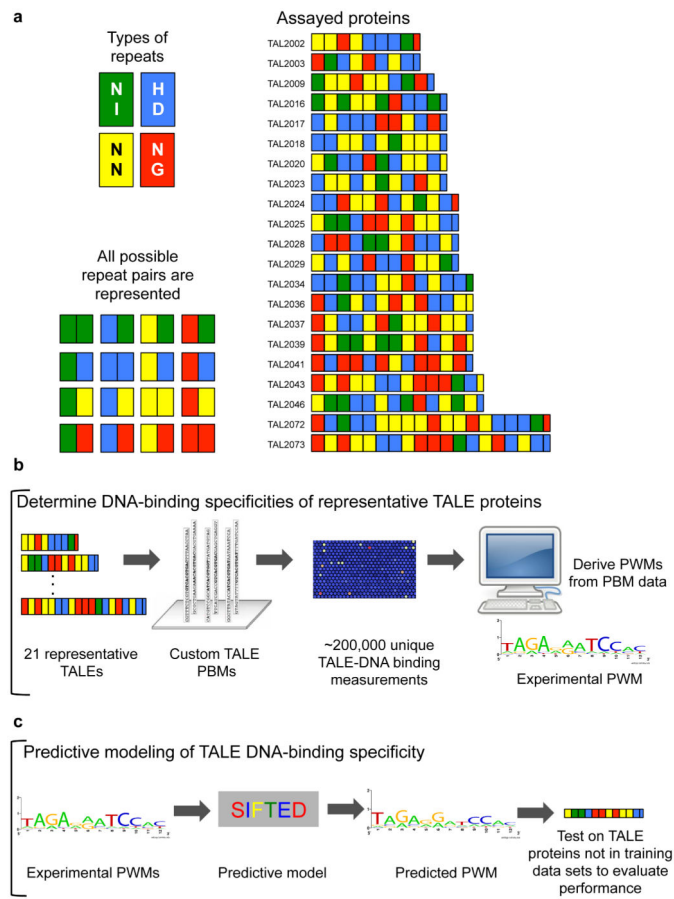


Figure 1. Overall experimental design and analysis scheme

(a) 21 Representative TALE proteins used in this study. Repeats are indicated by colored rectangles, and C-terminal half-repeats are indicated by smaller rectangles. RVD identities are indicated by letters. The set was chosen to include all possible repeat pairs and to cover a range of repeat lengths from 8.5 to 18.5 repeats. (b) Custom-designed PBMs were used to determine the specificity of representative TALE proteins. (c) These specificity profiles were used to learn features of TALE-DNA recognition and to train a predictive TALE specificity model, SIFTED (Specificity Inference for TAL-Effector Design).

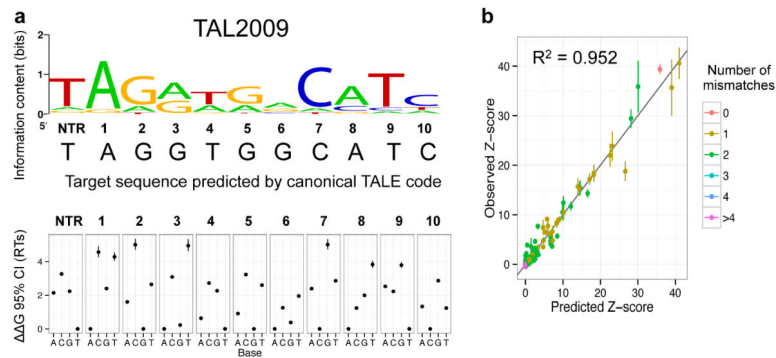


Figure 2. Determining PWMs from custom-designed PBMs

(a) Representative logo and $\Delta\Delta G$ estimates. The vertical bars represent the 95% credible interval (CI) and the points show the mean of the posterior distribution, in units of RT. The base predicted for each position by the TALE code is indicated below the logo. **(b)** Representative comparison between the probe z-scores measured in PBMs and the z-scores predicted by the derived PWM. Points represent the mean and vertical bars show its 95% confidence interval. Points are colored by the number of mismatches between the sequence in the probe and the consensus sequence predicted from RVD identities using the canonical TALE code.

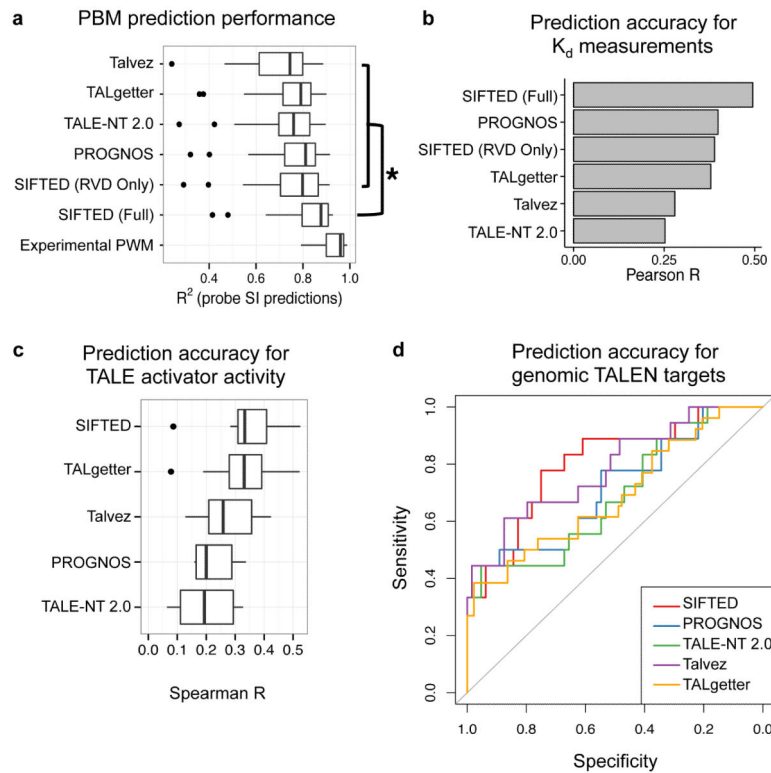


Figure 3. SIFTED predictive model performance

(a) Comparison of prediction accuracy of PWMs derived by different methods. The box plot shows how well the PBM probe intensities for each protein are predicted by the PWMs generated by SIFTED and other methods. Two versions of SIFTED are shown: one that only models repeats independently (“SIFTED (RVDs Only)”) and one that considers all repeat context features (“SIFTED (Full)”). Experimental PWMs are those derived from the PBM data. (*) The brackets highlight a subset of statistically significant differences ($P < 10^{-6}$, Wilcoxon signed-rank test). The box plots shows the median and the first and third quartiles. Whiskers extend to data points not considered outliers, while outliers are shown as individual points. Data are considered outliers when they are 1.5 times the interquartile range (IQR) higher than the third quartile, or 1.5 * IQR lower than the first quartile. (b) Prediction accuracy for relative binding affinity. PWMs derived from existing tools or from SIFTED (as in (a)) were used to predict relative K_d values for a single TALE protein^{27,35}. The bars display the Pearson correlation coefficient between observed and predicted $\log(K_d)$ values. (c) Validation of TALE activator binding specificity predictions by comparison to TALE activator activity data reported in Mali *et al*²². The five predictive methods were used to score all reported binding sites up to three mismatches away from the predicted target. These scores were compared to an expression score associated with that binding site using Spearman correlation. (d) Validation of TALEN binding specificity predictions by comparison to cell-based TALEN activity data, reported in Guilinger *et al*²⁰. The five methods shown were used to predict the binding of TALEN pairs to genomic target sites. The ROC curves show the sensitivity and specificity of each method for distinguishing genomic sites that showed nuclease activity (i.e., indels) and those that did not.

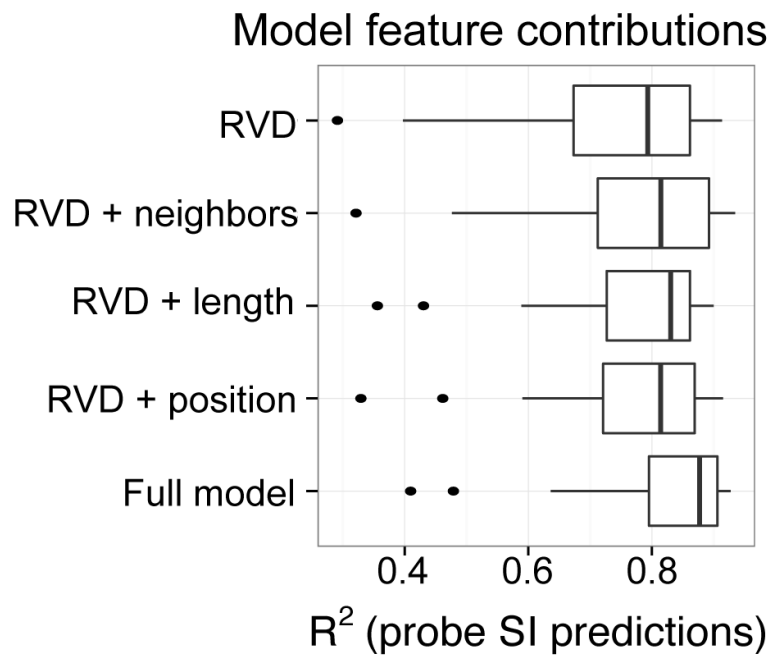


Figure 4. Contribution of model features

The plot shows the accuracy at predicting PBM probe intensities of a PWM predicted with no context features (top), with one single context feature added (middle) or with all context features included (bottom). Box plots are formatted as in Fig. 3a.

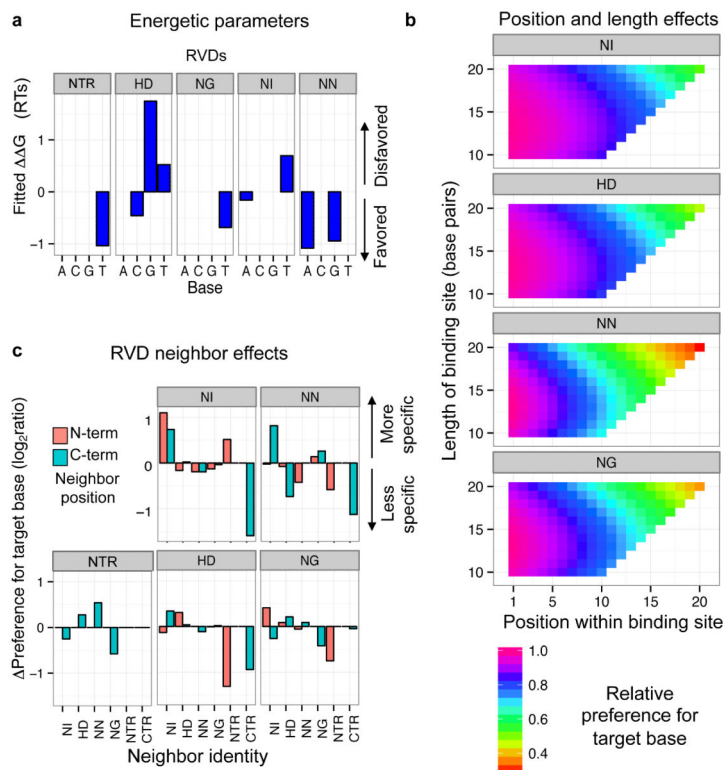


Figure 5. Protein features that affect repeat specificity

(a) RVD identity. G values from the model are indicated for each repeat type with each base. Additionally, the Gs for the four bases at the 5' T position, which are contacted by the NTR, are shown. **(b) Length and position.** The effects of protein length and repeat position on the specificity of each repeat type are shown. **(c) Effect of neighboring repeats or terminal regions on specificity.** For each repeat type and the NTR, the bar heights display the effect on specificity for different neighbors in the N- or C-terminal direction (orange and teal, respectively). The quantity shown is the log₂ ratio between the PWM frequency predicted with and without the presence of a given neighbor in the model. NTR refers to the N-terminal region of the protein. CTR refers to the C-terminal region; repeats with the CTR as the C-terminal neighbor are the half-repeats in the final repeat position.

Table 1

Target site guidelines for TALE design.

Target site guideline	Rationale
Target A runs	The NI repeat is more specific with NI as its N- or C-terminal neighbor
Avoid 3' A, C, or G	The NI, HD, and NN repeats are less specific at the C terminal end
Avoid T in first position	Both the 5' T preference and the NG repeat are less specific if the first repeat is NG
Use the SIFTED web tool to identify off-targets	The web tool incorporates all context effects, and can evaluate effective specificity in the genome.

The observed context effects were used to create simple guidelines to incorporate when designing TALEs. However, we recommend using the SIFTED web tool to predict specificity and locate potential off-target sites when designing a TALE protein to target a genomic region

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript