Received: 17 October 2019 | Revised: 27 November 2019 | Accepted: 2 December 2019

DOI: 10.1002/pro.3801

ARTICLE

THE PROTEIN SOCIETY WILEY

Context and number of noncanonical repeat variable diresidues impede the design of TALE proteins with improved DNA targeting

James T. Anderson[1] | Julia M. Rogers[1,2] | Luis A. Barrera[1,2,3] | Martha L. Bulyk[1,2,3,4]

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts

[2]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, Massachusetts

[3]Harvard-MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, Massachusetts

[4]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts

Correspondence
Martha L. Bulyk, Division of Genetics, Department of Medicine, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115.
Email: mlbulyk@genetics.med.harvard.edu

Funding information
National Human Genome Research Institute, Grant/Award Numbers: R21 HG007573, T32 HG002295

Abstract

Transcription activator-like effector (TALE) proteins have been used extensively for targeted binding of fusion proteins to loci of interest in (epi)genome engineering. Such approaches typically utilize four canonical TALE repeat variable diresidue (RVD) types, corresponding to the identities of two key amino acids, to target each nucleotide. Alternate RVDs with improved specificity are desired. Here, we focused on seven noncanonical RVDs that have been suggested to have improved specificity for their target nucleotides. We used custom protein binding microarrays to characterize the DNA-binding activity of 65 TALEs containing these alternate or corresponding canonical RVDs at multiple positions to ~5,000 unique DNA sequences per protein. We found that none of the noncanonical thymine-targeting RVDs displayed stronger preference for thymine than did the canonical RVD. Of the noncanonical RVDs with putatively improved specificity for guanine, only EN and NH showed greater discrimination of guanine over adenine. This improved specificity, however, comes at a cost: more substitutions of a noncanonical RVD for a canonical RVD generally decreased the protein's DNA-binding activity. Our results highlight the need to investigate RVD-nucleotide specificities in multiple protein contexts and suggest that a balance between canonical and noncanonical RVDs is needed to build TALEs with improved specificity.

KEYWORDS

genome engineering, nucleotide specificity, protein binding microarray, RVDs, TALEs

1 | INTRODUCTION

Transcription activator-like effector (TALE) proteins evolved as virulence factors secreted by the bacteria *Xanthomonas* via their Type III secretion system to alter expression of plant host genes beneficial to the bacteria's proliferation.[1–3] In recent years, the ability to customize TALEs that recognize specific sequences has allowed for broad applications in genome and epigenome engineering, including genome editing, transcriptional activation, transcriptional repression, live visualization of chromatin dynamics, and chromatin affinity purification of a targeted native genomic locus.[4–11] The underlying DNA interaction of this transcriptional reprogramming is mediated by the TALE repeat array, a domain containing repeating subunits (TALE repeats), with each subunit comprising 33 or 34 amino acid residues. These residues are highly conserved except for the hypervariable

residues 12 and 13, the so-called repeat variable diresidue (RVD). This RVD is primarily responsible for the specific protein–nucleotide interaction of the repeat with its target base.[1–3]

Early evidence suggested a simple "TALE code," whereby the four nucleotides adenine, cytosine, guanine, and thymine are bound by the canonical RVDs NI (Asparagine, Isoleucine), HD (Histidine, Aspartic Acid), NN (Asparagine, Asparagine), and NG (Asparagine, Glycine), respectively.[12,13] To target a custom TALE protein to a desired DNA sequence, one could assemble a TALE repeat array with the order of RVDs from N-terminus to C-terminus corresponding to the order of target nucleotides from 5′ to 3′. The DNA sequence predicted to be bound by the fully-assembled TALE protein is also immediately preceded by a requisite 5′T, since the N-terminal region adjacent to the TALE repeat array binds thymine with high specificity.[14]

Emerging evidence, however, has revealed unexpected complexities underlying this seemingly simple TALE code. Rogers *et al.* suggested a context effect, whereby the specificity and affinity of any particular RVD-nucleotide interaction may be significantly affected by the identities of the immediately neighboring RVDs/nucleotides. They furthermore reported a position and length effect, whereby the specificity of an RVD for its target nucleotide is influenced both by the position of the given RVD within the repeat array, and by the total length of the repeat array, with increased TALE protein length diminishing RVD specificity.[15] Rinaldi *et al.* similarly observed that specificity for target versus nontarget DNA varies with TALE length, decaying at longer lengths.[16] Other evidence has suggested that TALE-DNA specificity might be improved by altering TALE repeat residues aside from just the RVD.[17] Beyond the repeat array, the TALE C-terminal domain (CTD) has been shown to increase nonspecific TALE-DNA interactions, suggesting that re-engineering this domain might improve on-target specificity.[18] A comprehensive characterization of these emerging complexities would allow for TALE proteins to be applied for genome and epigenome engineering with greater precision.

The RVD repertoire employed by *Xanthomonas* extends well beyond just the four canonical RVD types, and natural TALE repeat arrays often utilize serial combinations of both canonical and noncanonical RVDs[19] in a manner which is not fully understood. Attempts to characterize these noncanonical RVDs have utilized various approaches, in terms of both TALE design and experimental methodology. Some studies have considered only TALEs of particular biological relevance.[20] Others have looked at large numbers of different RVDs within a fixed repeat-array context.[21] Methods for inferring the DNA-binding activity of these TALEs have included sequencing of cleaved DNA fragments from TALE-nuclease fusions,[18] ELISA assays for TALE-oligonucleotide binding,[21] reporter assays for TALEs' transcriptional activation effects,[22] and in silico modeling.[23] Notably, most prior studies of noncanonical RVD-DNA binding examined RVD-DNA specificity only at a particular position within a TALE protein.[21,24] Moreover, the most comprehensive studies of the full range of RVDs tested them at just one specific position within one TALE protein context.[21]

The DNA binding specificities of the canonical RVDs are not pure.[1,2] The canonical RVD, NN, for guanine binds nearly as well to adenine.[15] Among the remaining nucleotides, adenine, and cytosine appear to be specified better by NI and HD, respectively, than is thymine by NG. Degeneracy in RVD-nucleotide specificity leads to off-target recognition by custom-designed TALE proteins. Therefore, alternate RVDs with more specific recognition have been sought.

Prior studies found the NH RVD to be more specific for guanine than is the canonical NN RVD[9,22]; however, those results are in contrast to those from a more recent study that found NN to be more specific for guanine.[21] The NK RVD has also been found to be more specific for guanine than is NN,[25] but has reduced affinity,[26,27] resulting in lower nuclease targeting efficiency[28] and lower fold-activation as a transcriptional activator fusion[27] as compared to TALEs containing the NN RVD, and hence is less desirable for various bioengineering applications. An ELISA-based binding survey of nucleotide specificities of RVDs within one TALE protein context and within one fixed flanking DNA sequence context reported the improved specificity of GN, DN, and EN for guanine and of HG, VG, and VA for thymine, among others.[21] A separate, reporterassay-based study also surveyed RVDs within one TALE protein context and within one fixed flanking DNA sequence context, with further testing of the top candidates of interest as 6N or 12N tandem RVD multimers within a single TALE protein context, which did not support improved binding by NH for guanine or HG for thymine, among others.[24]

In this study, we focused on three main questions on how the usage of noncanonical RVDs affects TALE DNA targeting (Figure 1). First, in light of our prior study of canonical RVD-DNA binding specificities which revealed that the identity of neighboring RVDs, position of a RVD within a TALE protein, and length of the TALE protein influence the base preference of an RVD,[15] we were motivated to investigate further the DNA binding specificities of alternate RVDs across a panel of TALE protein contexts (Figure 1, Question 1). As part of this investigation, we were also interested to test whether a more highly specific TALE protein could be constructed by simply swapping out all occurrences of a canonical RVD for an

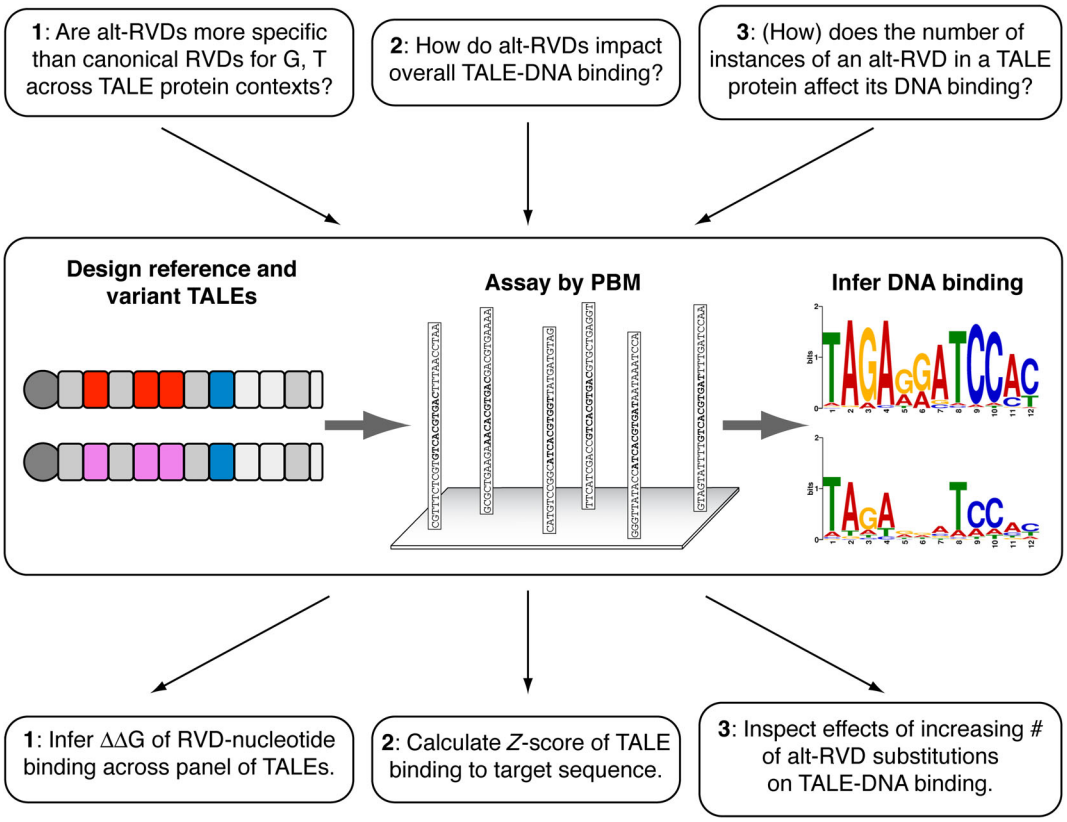

FIGURE 1 Overall schema. This study was designed to address three main questions, for each of which we used PBM experiments to assay sequence-specific DNA-binding by custom-designed TALE proteins, followed by separate data analyses designed to address each of the three questions about the impact of alternate RVDs ("alt-RVDs") on TALE DNA binding. PBM, protein binding microarray; RVD, repeat variable diresidue; TALE, transcription activator-like effector

alternate RVD reported to have improved specificity for the same target nucleotide (Figure 1, Question 2). Hence, in this study we employed a "full swap-out" approach to characterize seven noncanonical RVDs previously reported to have improved specificity for their target nucleotide: four (NH, GN, DN, and EN) for guanine and three (HG, VG, and VA) for thymine. Third, we created a "combinatorial swap-out" series of TALE proteins to investigate how increasing the number of a noncanonical RVD within a TALE protein impacts TALE DNA binding activity (Figure 1, Question 3).

In total, we assembled a library of 65 TALE proteins, comprising 11 reference TALEs bearing exclusively canonical RVDs, and 54 variant TALEs for characterizing the DNAbinding properties of the seven noncanonical RVDs of interest. Protein binding microarrays (PBMs) are a high-throughput, *in vitro* technology for rapid, highly parallel characterization of the DNA binding specificities of proteins. We previously used custom-designed PBMs to investigate the DNA binding specificities of all 11 of the reference TALE proteins[15] that we used in this present study. Here, we designed custom PBMs which bear probes for the predicted target sequence of each TALE, in addition to probes for all possible mono- and dinucleotide substitutions of these target sequences. The resulting quantitative binding data for the 65 TALE proteins to a total of ~5,000 unique DNA probes allowed us to infer the relative affinity and specificity of each noncanonical RVD to each of the four nucleotides.

This study represents an important approach to further characterize the effects of noncanonical RVDs that have been reported to exhibit improved nucleotide specificity. Our results provide valuable insights on seven such noncanonical RVDs and serve as important models for engineering TALE proteins with greater specificity.

2 | RESULTS

2.1 | Full swap-out TALE approach and custom PBM design

We selected seven noncanonical RVDs and two canonical RVDs — NH, GN, DN, EN, HG, VG, and VA, and NN and NG, respectively — for in-depth characterization of their DNA-binding specificities based on the literature. Prior

studies, some conflicting, reported the NH, GN, DN, and EN RVDs to be more specific than the canonical NN RVD for guanine,[9,21,22,24] and reported the HG, VG, and VA RVDs to be more specific than is the canonical NG RVD for thymine.[21] We will henceforth collectively refer to NH, GN, DN, and EN as "alt-G" RVDs, and HG, VG, and VA as "alt-T" RVDs, as these are alternate RVDs for targeting guanine or thymine, respectively.

To investigate these RVDs, we constructed a library of TALE proteins in which we represented the canonical or alternate RVDs within the context of: (a) six different TALE proteins ranging from 9 to 13 RVDs in length for the alt-G RVDs, comprising a total of 22 distinct TALE proteins (6 canonical plus 16 alternate); or (b) five different TALE proteins ranging from 9 to 13 RVDs in length for the alt-T RVDs, comprising a total of 20 additional distinct TALE proteins (5 canonical plus 15 alternate; Figure 2a). We refer to the set of TALE proteins in which the different RVDs targeting a particular nucleotide are represented in the context of otherwise the same TALE protein as "allelic series," or more precisely the alt-G or alt-T series. In creating these allelic series, we employed a "full swap-out" design in which the "reference" TALE bears a repeat array of entirely canonical RVDs for its target sequence. These reference TALE proteins were all selected from a panel of TALE proteins which have been studied extensively by our group[15] and others.[29] For each variant of a reference TALE, we substituted every instance in the repeat array where a particular canonical RVD appeared (*i.e.*, either NN in the alt-G series, or NG in the alt-T series) with a particular non-canonical RVD. All remaining RVDs in the allelic series remain unchanged from the canonical RVDs in the reference protein. For example, the reference TALE2016 protein ("TALE2016-Ref") bears three instances of the NN RVD in its repeat array, and the corresponding GN variant protein ("TALE2016-GN") within the alt-G series has a GN RVD in place of all three of the NN RVDs while all the remaining RVDs remain unchanged (Figure 2b). In this manner, for each of the alt-T RVDs, we assembled five TALE variants, and for the alt-G RVDs we assembled six TALE variants for GN, five for NH, three for DN, and two for EN (Figure 2a).

We used custom PBMs to assay the DNA binding specificities of each of these TALE proteins (Figure 2, middle panel). Briefly, on the custom PBMs, we designed probes

(a)

				alt-G				alt-T		
TALE Series	Target DNA Sequence	RVD Sequence in Repeat Array	Ref	GN	DN	EN	NH	HG	VG	VA
2002	TGGTGCCCAT	NN-NN-NG-NN-HD-HD-HD-NI-NG	x					x	x	x
2003	TTACGTCGCC	NG-NI-HD-NN-NG-HD-NN-HD-HD	x	x			x			
2009	TAGGTGGCATC	NI-NN-NN-NG-NN-NN-HD-NI-NG-HD	x	x			x			
2016	TAGAGGATCCAC	NI-NN-NI-NN-NN-NI-NG-HD-HD-NI-HD	x	x	x	x	x			
2017	TCGCCCTTGCTC	HD-NN-HD-HD-HD-NG-NG-NN-HD-NG-HD	x					x	x	x
2020	TGACCTACGGCG	NN-NI-HD-HD-NG-NI-HD-NN-NN-HD-NN	x					x	x	x
2023	TCGGCGAGCTGC	HD-NN-NN-HD-NN-NI-NN-HD-NG-NN-HD	x	x	x		x			
2036	TTCAGCGTGTCCGG	NG-HD-NI-NN-HD-NN-NG-NN-NG-HD-HD-NN-NN	x	x		x				
2037	TTGCCGTAGGTGGC	NG-NN-HD-HD-NN-NG-NI-NN-NN-NG-NN-NN-HD	x					x	x	x
2039	TTGAAGAAGTCGTG	NG-NN-NI-NI-NN-NI-NI-NN-NG-HD-NN-NG-NN	x	x	x		x			
2041	TTCTTCTGCTTGTC	NG-HD-NG-NG-HD-NG-NN-HD-NG-NG-NN-NG-HD	x					x	x	x

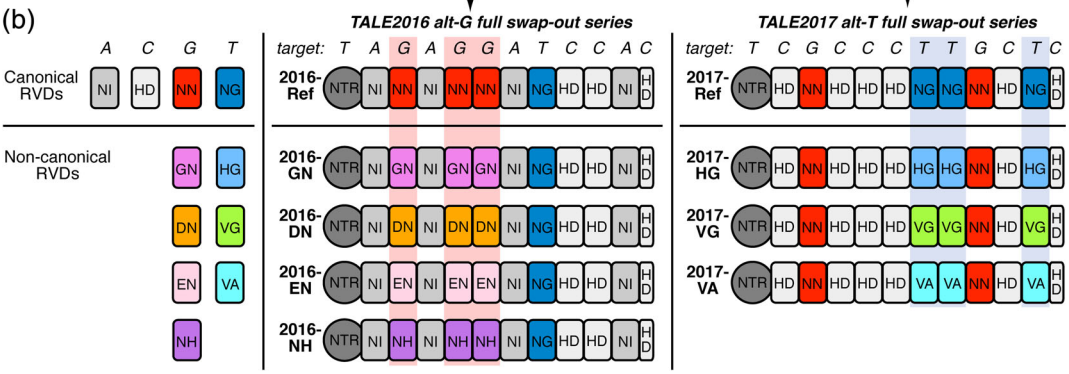

FIGURE 2 Overview of full swap-out TALE proteins. (a) Each of 11 TALE protein allelic series comprising a TALE protein assembled with only canonical RVDs ("Ref") and those assembled with a particular alternate ("alt-T" or "alt-G") RVD fully substituting all occurrences of the NG or NN, respectively, canonical RVD. "x" denotes TALE proteins that were assembled and analyzed. (b) Two representative examples of assembled full swap-out TALE proteins. RVD, repeat variable diresidue; TALE, transcription activator-like effector

that represent the target site for each TALE protein, as well as sequences representing all possible mononucleotide and consecutive dinucleotide substitutions within the target site. Target sites were predicted according to the canonical TALE code and were flanked by constant DNA sequences and situated at a fixed position within the 60-bp probes relative to the slide surface (Figure S1). The constant flanking sequences had been used in a prior study to assay these TALE proteins assembled using only canonical RVDs.[15]

For each protein, we measured binding to between 144 and 208 unique variant target sites that cover all possible mononucleotide and adjacent dinucleotide substitutions. Array-bound, GST-tagged TALE protein was detected with the use of a fluorescent anti-GST antibody. By measuring how much each substitution changed the fluorescence signal intensity of TALE protein binding to the DNA probe, we inferred changes in binding free energy ($\Delta\Delta G$ values) for each possible substitution within the target site, as described previously.[15] From these ΔΔG values, we derived a position weight matrix (PWM) for each protein (Methods); probe *z*-scores predicted using the derived PWMs generally correlated well with the observed *z*-scores from the PBM data, particularly for proteins with higher observed *z*-scores (Figure S2). The inferred PWMs were consistent across experimental replicates (Figure S3) and across PBM experiments performed at different concentrations of TALE proteins (Figure S4).

2.2 | Inferring nucleotide specificities of noncanonical RVDs

To investigate the nucleotide specificities of each of the alternate RVDs across the panel of TALE protein contexts within which we assayed them, we evaluated their $\Delta\Delta G$ values (Figure 1, Analysis 1). Briefly, we calculated the mean inferred $\Delta\Delta G$ value for each RVD-nucleotide pair across all TALE contexts in which the RVD is represented, for both canonical and noncanonical RVDs, for all full swap-out TALEs and single combinatorial swap-in TALEs (see below) for which we detected significant sequence-specific DNA-binding activity. As positive controls of our assays and analyses, we compared the specificities of the four canonical RVDs inferred from the PBM data in this study with those reported in a prior study of the nucleotide specificities of canonical RVDs, each assayed in the context of a larger set of TALE proteins on PBMs, and found that they largely agree (Figure S5).[15]

Strikingly, none of the noncanonical RVDs HG, VG, or VA displayed stronger preference for thymine than did the canonical RVD NG (Figure 3). Among the guanine-specifying RVDs, the canonical RVD NN recognized A in addition to G, with a strong aversion to C and T, as reported previously.[15] A/G specificity and C/T aversion were also exhibited by the alt-G RVDs GN and DN. Importantly, we found that EN and NH also strongly disfavor C and T, but do so while better discriminating G from A (Figure 3), supporting the potential benefit of the EN and NH RVDs in TALE-based genomic applications that require more stringent guanine-specificity than is afforded by the canonical NN RVD. The larger variation observed for EN and NH (Figure 3) suggests that their specificities might exhibit greater dependence on the TALE protein context.

2.3 | Impact of noncanonical RVDs on target specificity depends on TALE context

To characterize the effects of the noncanonical RVDs on TALE-DNA binding specificity beyond the individual RVDs' nucleotide specificities, we investigated how the noncanonical RVDs impacted the relative binding

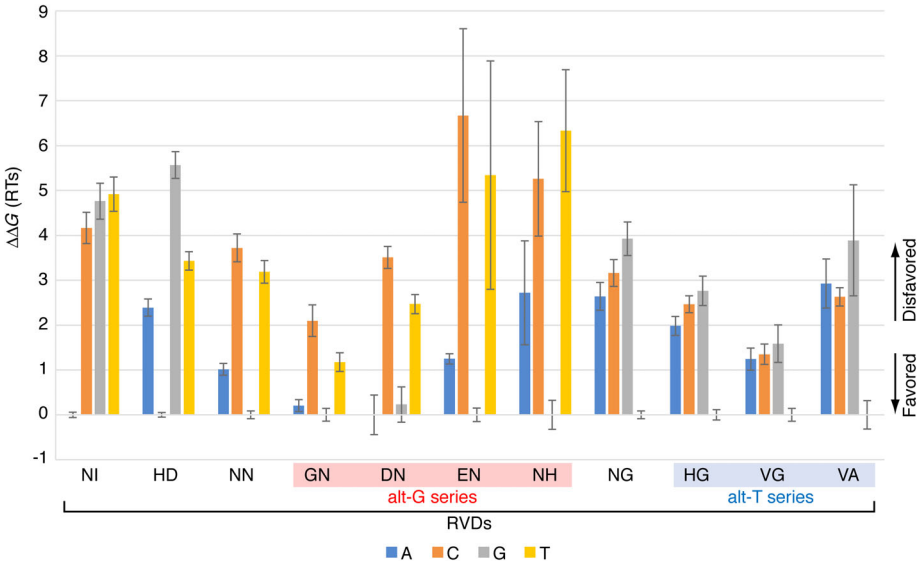

FIGURE 3 RVD-nucleotide specificities derived from TALE PBM data. Mean $\Delta\Delta G$ values for each RVD-nucleotide pair calculated from measured PBM data. Error bars represent 1 *SEM*. RVDs are ordered such that RVDs with the same targeted nucleotide are adjacent. PBM, protein binding microarray; RVD, repeat variable diresidue

preferences of the assembled TALE proteins (Figure 1, Analysis 2). To assess the significance of TALE-DNA binding, we calculated the *z*-score of each alt-G and alt-T TALE variant for its target sequence from the median signal intensity (SI) of the predicted target sequence for each TALE protein, using the median SIs from all the probes designed to represent binding sites for the other TALE allelic series as a background distribution. Thus, our custom PBM platform provides an approach to investigate not just individual RVDs' nucleotide sequence specificities, but also their higher-order effects on the relative binding affinities of TALE protein allelic series.

Strikingly, almost every alt-G TALE variant in all six alt-G allelic series (*i.e.*, in the context of TALE2003, TALE2009, TALE2016, TALE2023, TALE2036, and TALE2039) displayed either dramatically reduced or abrogated binding to its target sequence as compared to the reference TALE protein (Figure 4a). In contrast, we observed different degrees of context-dependent effects (*i.e.*, in the context of TALE2002, TALE2017, TALE2020, TALE2037, and TALE2041) for the different alt-T RVDs (Figure 4b). Comparing the *z*-score of each alt-T variant protein to that of the corresponding reference protein across all five alt-T allelic series, three of the five HG TALE variants displayed similar preferences for their predicted target sequences as did the reference TALE proteins, while in one context (TALE2017) HG resulted in a much better binding TALE protein and in a different context (TALE2041) the canonical RVD NG yielded a much better binding TALE protein. In contrast, the VG and the VA TALE variants displayed much more variable, context-dependent effects on target sequence preference, with generally strongly reduced binding to their target sequences as compared to the corresponding reference TALEs, with the sole exception of TALE2020-VA, which bound its target sequence much better than did TALE2020-Ref. Notably, while VG impeded DNA binding in the context of all five TALE proteins, it had the least effect on TALE2020. Intriguingly, TALE2020 has only one T-specifying RVD and consequently TALE2020-VA has just one VA substitution, whereas TALE2017, TALE2020, TALE2037, and TALE2041 have multiple T-specifying RVDs. These results suggest that the potential beneficial effect of noncanonical RVDs with putatively improved nucleotide specificities on TALE-DNA binding may be lost with the incorporation of multiple noncanonical RVDs in a repeat array.

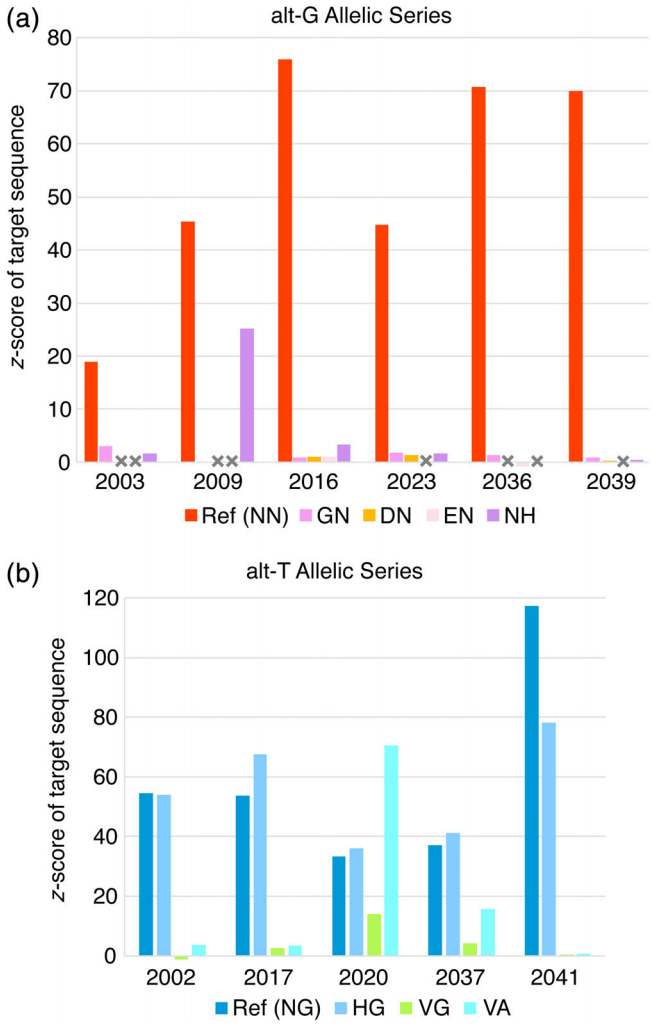

FIGURE 4 Relative preference of assembled transcription activator-like effector (TALEs) for binding predicted target sequence. A, Alt-G allelic series. *Z*-score of each TALE protein's target sequence calculated from measured PBM data. Light gray "x" in place of a bar signifies that the corresponding TALE was not assembled. (b) As in (a), but for the alt-T allelic series

2.4 | Increasing numbers of noncanonical alt-G RVDs within a TALE protein diminish TALE DNA-binding activity

In light of the results from our "full swap-out" allelic series, we sought to further investigate whether increasing numbers of a noncanonical RVD with putatively improved specificity within a TALE protein do indeed diminish the TALE protein's binding to its target sequence (Figure 1, Analysis 3). Motivated by the need to understand the effects of G-specifying RVDs on DNA binding given the relatively poor specificity of the canonical G-specifying RVD NN, we focused on further characterizing the alt-G RVDs. Here, we implemented a new approach which we term "combinatorial swap-out." Unlike the full swap-out approach, in which every instance of a particular nucleotide-specifying canonical RVD (*e.g.*, NN for guanine) is replaced by one particular

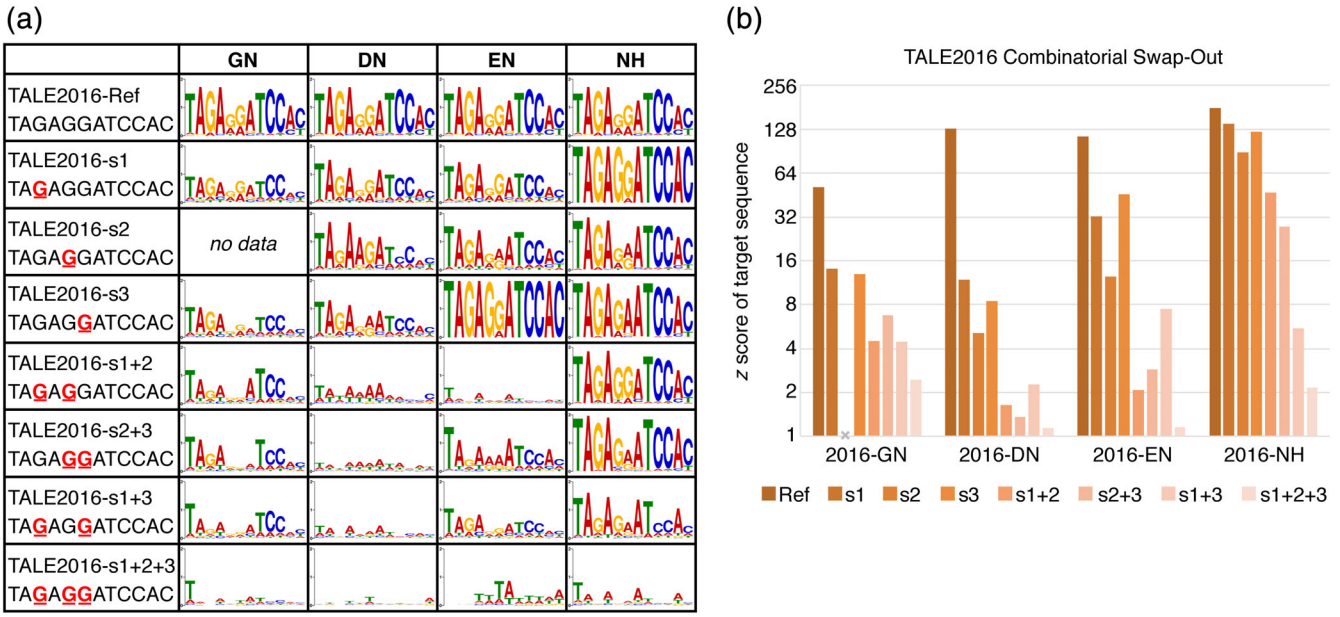

FIGURE 5 PBM analysis of alt-G combinatorial swap-out TALE proteins. (a) 23 TALE proteins with partial swap-out of the canonical G-specifying RVD NN for either the RVD GN, DN, EN, or NH, were constructed and analyzed by PBMs, and compared to PBM data for the corresponding reference alleles or full swap-out alleles. Sequence logos represent the DNA binding specificities derived from the PBM data. In the left column, red underlined font in the target DNA sequence indicates the nucleotide targeted by each substitution of a canonical RVD for a noncanonical RVD. The TALE2016-s2 GN TALE ("no data") was not assembled. (b) *Z*-score of each TALE protein's target sequence calculated from the measured PBM data. Light gray "x" in place of a bar signifies that the corresponding TALE was not assembled. PBM, protein binding microarray; RVD, repeat variable diresidue; TALE, transcription activator-like effector

noncanonical RVD (*e.g.*, NH), in the combinatorial swap-out approach we generated all possible single, double, and triple substitutions for each particular noncanonical RVD (*i.e.*, GN, DN, EN, and NH).

We applied this new combinatorial swap-out approach on the TALE2016 allelic series as a model system. Since TALE2016 has three instances of the NN RVD in its TALE repeat array, we created seven TALE2016 variants per alternate RVD: three single substitutions, three double substitutions, and one triple substitution, with the triple substitution being the "full swap-out" variant described earlier. In total, this comprised 23 additional, distinct TALE proteins (Figure 5a, left column).

We then assayed the TALE2016 combinatorial swap-out alt-G allelic series by PBM as we had done previously for the full swap-out allelic series. Strikingly, across all alt-G RVDs, increasing numbers of an alt-G RVD within a TALE protein generally resulted in both an increasingly degenerate motif indicating diminished sequence specificity (Figure 5a) and also decreased binding to its target sequence (Figure 5b). Of the four alt-G RVDs that we assayed, the motif degeneracy of TALEs incorporating NH was somewhat more robust to the number of substituted NH RVDs, although the trend for increased degeneracy with a greater number of substitutions was still observed.

3 | DISCUSSION

Our study highlights that the nucleotide specificities of noncanonical RVDs considered in an isolated context — that is, one position within the context of a TALE protein — do not necessarily reflect the specificity of the RVD in different protein contexts. For example, although the noncanonical thymine-targeting RVDs HG, VG, or VA had been described as having improved specificity for thymine when tested at a single, fixed position in the context of a single TALE protein, in our survey across positions in a variety of TALE protein contexts, none of them displayed a broadly stronger preference for thymine than did the canonical RVD NG. Similarly, our results provide further support for EN and NH, but not for GN and DN, having improved specificity for guanine.

The results from our TALE2016 combinatorial swap-out allelic series are consistent with a prior study that found the canonical NN RVD to be a "strong RVD," that is, a particularly avid, albeit not very specific, binder of guanine.[22] Consequently, Streubel *et al.* recommended that NN should comprise a certain minimal portion of the RVDs within a TALE repeat array, as the total strength of the NN-guanine hydrogen bonds might boost the overall TALE-DNA interaction to sufficient affinity for desired levels of binding.[22] Results from our

combinatorial swap-out PBM data support these TALE design guidelines, while also suggesting that single and possibly even double substitutions of a noncanonical alternative to NN is not necessarily prohibitive for recognizing the target sequence, although it may not improve its target specificity. Future studies will need to examine the effects of varying positions and numbers of substitutions in other TALE protein contexts.

Our study also demonstrates the importance of substituting canonical for noncanonical RVDs at multiple positions within a TALE protein when assaying the effects of the RVD on DNAbinding activity, in order to capture potential higher-order, including positional, effects on TALE DNA-binding activity that are not apparent when testing RVDs at a single position within a TALE protein. Although Miller *et al.*[21] used noncanonical RVDs to design TALE-FokI dimers with reduced off-target cleavage, only two out of a total of 16 RVDs within the TALE proteins were substituted with noncanonical 16 RVDs; furthermore, one of the two substituted RVDs was near the C-terminus, where RVDs exert a less substantial effect on TALE DNAbinding affinity.[15] Our results support previous characterization of NN as an "anchor" RVD[22] and suggest that the optimal approach for engineering high-specificity TALEs may be a hybrid design comprising both canonical and noncanonical RVDs, with the noncanonical RVDs being incorporated sparingly at certain key positions for improved specificity while maintaining overall DNA binding affinity. Further studies across a wider range of TALE protein sequences, TALE array lengths, and combinatorial substitutions of multiple different RVDs within the same TALE protein are needed to be able to develop a predictive model[15] of the effects of these various parameters on the nucleotide specificities of noncanonical RVDs and the overall DNA-binding affinities of TALEs containing noncanonical RVDs.

4 | MATERIALS AND METHODS

4.1 | Cloning of TALE proteins

TALE vectors were assembled by a combination of REAL assembly and REAL-Fast assembly.[30] Since the REAL assembly and REAL-Fast plasmid vectors do not include plasmids encoding noncanonical RVDs, we created such plasmids by site-directed mutagenesis of the canonical RVD plasmids. Plasmid vectors expressing one-, two-, or four-long TALE repeats (within the pUC57-ΔBsaI backbone) were ligated in a serial, hierarchical progression to assemble full TALE repeat arrays bearing the proper sequence of RVDs to target the DNA sequence of interest. Generally, this assembly involved restriction enzyme digest of each N-terminal TALE repeat vector with BsaI and BamHI, followed by digestion of its neighboring C-terminal TALE repeat vector with BsaI and BamHI, and finally ligation of these neighboring repeats by T4 DNA ligase.

We used a Gateway-compatible TALE expression vector that we had created previously[15] that allows cloning of RVD repeats within the context of the Δ152 N-terminal domain, the final 0.5-repeat, and the +63 CTD. The resulting TALE Entry clones 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 (Table S1).

4.2 | Custom PBM design

The target site for each TALE protein was determined using the canonical TALE code (NI: A; HD: C; NN: G; NG: T), and was preceded by a 5′ T to create the full target site. In addition, all mononucleotide and consecutive dinucleotide substitutions within the target site were represented on probes, as described previously.[15] All target and nontarget sites were positioned within constant flanking regions that were used in prior custom PBM designs[15,31] and do not contain binding sites for any of the TALE proteins analyzed in this study. Each probe sequence was represented on at least 10 replicate spots on the array. Probes were synthesized in 8×60K (Agilent Technologies; AMADID #084120) array format. This array design has been deposited in NCBI GEO under accession number GPL26374.

4.3 | PBM experiments

Proteins were expressed using the PURExpress In vitro Transcription and Translation Kit (New England Biolabs). GST-TALE protein concentrations were estimated by anti-GST western blots with a dilution series of recombinant GST (Sigma). Proteins were stored at 4°C until being used in PBM assays. The duration of storage at 4°C between protein expression and PBM experiments was typically 1 day, but never greater than 3 days.

PBM experiments were performed essentially as described previously.[32] 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 PBS for 1 hr, washed with 0.1% Tween-20 in PBS and 0.01% TX-100 in PBS, then incubated with protein mixture (TALE protein in PBS, 2% milk, 0.2 mg ml^{-1} BSA, and 0.3 ug ml^{-1} salmon testes DNA) for 1 hr. The final concentration of TALE protein in the PBM reactions was 200 nM,

unless otherwise indicated (Table S2). Arrays were washed with 0.5% Tween-20 in PBS and 0.01% TX-100 in PBS. Lastly, the array was incubated for 20 min with an Alexa488-conjugated anti-GST antibody (Invitrogen A-11131), and washed with 0.05% Tween in PBS, and PBS. To minimize potential batch effects, each TALE allelic series was assayed in separate "chambers" on the same 8 x 60K format PBM slides, using proteins that were expressed in the same batch of IVT reactions (except for the TALE2016 and TALE2017 allelic series, which each include one protein—TALE2016-NH and TAL2017-VA, respectively—that was expressed and assayed in a different batch) and diluted to achieve equal TALE protein concentrations across an allelic series.

4.4 | 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 a similar approach as described previously in a prior custom TALE PBM study.[15] Briefly, masliner software[33] was used to combine Alexa488 scans at three different laser power levels. If a data set had any spots with negative background-subtracted intensity (BSI) values, a pseudocount was added to all BSI values for that experiment such that all values were then positive. 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.[34] Any individual replicate probe with a normalized adjusted BSI value more than 3 standard deviations (SD) from the median of the replicate probes was omitted from subsequent analysis. 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 all other TALE proteins not 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 SD of the background set SIs was calculated robustly using the asymptotic approximation $\sigma = 1.4826$ x MAD. The z-score for each probe was calculated relative to the median and SD of its corresponding background probes.

4.5 | PWM model fitting

We employed a previously described Bayesian Markov chain Monte Carlo (MCMC) method[35] to infer-free energy parameters of TALE–DNA interactions from PBM data.[15] Briefly, this method estimates $\Delta\Delta G$ values for each possible nucleotide substitution in a protein's DNA binding site motif. The $\Delta\Delta G$ values represent the difference in binding free energy relative to the preferred base, with lower $\Delta\Delta G$ values for any RVD-nucleotide pair representing more preferred interactions. These values are assembled to construct an energy matrix, in which each column represents a position within the binding site and each row represents a nucleotide. The energy matrix values are then converted to probabilities using the Boltzmann distribution, creating a PWM. The $\Delta\Delta G$ 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 for each probe (Figure S2). At each sampling step, the probe z-scores are predicted given the current parameter values, which we used to derive 95% credible intervals.[36] The priors on $\Delta\Delta G$ 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 MCMC sampling, we used the No-U-Turn Sampler included in Rstan v2.0 with default parameter settings. The $\Delta\Delta G$ parameters were initialized following a simple TALE code: $\Delta\Delta G = 0.0$ for the predicted optimal base at a given position, otherwise $\Delta\Delta G = 3.0$ RT. For each data set, we obtained 500 parameter samples in the burn-in period followed by 2,000 samples that were used to approximate the posterior distributions of all parameters. Four MCMC chains were run in parallel for each data set; the samples from each chain were then used to verify convergence of all $\Delta\Delta G$ parameters (Gelman-Rubin convergence statistic for all four chains <1.05).[37]

4.6 | Predicting probe signal intensity z-scores from PWMs

Probe signal intensities were predicted as described previously.[15] Briefly, the chemical potential m and the scaling terms a and b were fitted using 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 probe z-scores, the predicted probe z-scores were obtained by using the total $\Delta\Delta G$ for the binding site in each probe and the fitted variables as input.

5 | ACCESSION CODES

All analyzed microarray data and array designs have been deposited in NCBI GEO under Series ID GSE129193.

ACKNOWLEDGMENTS

The authors thank Deepak Reyon for technical assistance with the REAL assembly protocol, Stephen Gisselbrecht for assistance in figure preparations, and Kian Hong Kock and Timothy Read for critical reading of the manuscript.

CONFLICT OF INTEREST

M.L.B. is a co-inventor on patents on PBM technology.

ORCID

Martha L. Bulyk https://orcid.org/0000-0002-3456-4555

REFERENCES

1. Moscou MJ, Bogdanove AJ. A simple cipher governs DNA recognition by TAL effectors. Science. 2009;326:1501.
2. Boch J, Scholze H, Schornack S, et al. Breaking the code of DNA binding specificity of TAL-type III effectors. Science. 2009;326:1509–1512.
3. Bogdanove AJ, Koebnik R, Lu H, et al. Two new complete genome sequences offer insight into host and tissue specificity of plant pathogenic Xanthomonas spp. J Bacteriol. 2011;193:5450–5464.
4. Byrum SD, Taverna SD, Tackett AJ. Purification of a specific native genomic locus for proteomic analysis. Nucleic Acids Res. 2013;41:e195.
5. Joung JK, Sander JD. TALENs: A widely applicable technology for targeted genome editing. Nat Rev Mol Cell Biol. 2013;14:49–55.
6. Maeder ML, Angstman JF, Richardson ME, et al. Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. Nat Biotechnol. 2013;31:1137–1142.
7. Maeder ML, Linder SJ, Reyon D, et al. Robust, synergistic regulation of human gene expression using TALE activators. Nat Methods. 2013;10:243–245.
8. Mendenhall EM, Williamson KE, Reyon D, et al. Locus-specific editing of histone modifications at endogenous enhancers. Nat Biotechnol. 2013;31:1133–1136.
9. 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.
10. Miyanari Y, Ziegler-Birling C, Torres-Padilla ME. Live visualization of chromatin dynamics with fluorescent TALEs. Nat Struct Mol Biol. 2013;20:1321–1324.
11. Perez-Pinera P, Ousterout DG, Brunger JM, et al. Synergistic and tunable human gene activation by combinations of synthetic transcription factors. Nat Methods. 2013;10:239–242.
12. 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.
13. Deng D, Yan C, Pan X, et al. Structural basis for sequence-specific recognition of DNA by TAL effectors. Science. 2012;335:720–723.
14. Stella S, Molina R, Yefimenko I, et al. Structure of the AvrBs3-DNA complex provides new insights into the initial thymine-recognition mechanism. Acta Crystallogr. 2013;D69:1707–1716.
15. Rogers JM, Barrera LA, Reyon D, et al. Context influences on TALE-DNA binding revealed by quantitative profiling. Nat Commun. 2015;6:7440.
16. Rinaldi FC, Doyle LA, Stoddard BL, Bogdanove AJ. The effect of increasing numbers of repeats on TAL effector DNA binding specificity. Nucleic Acids Res. 2017;45:6960–6970.
17. Tochio N, Umehara K, Uewaki JI, et al. Non-RVD mutations that enhance the dynamics of the TAL repeat array along the superhelical axis improve TALEN genome editing efficacy. Sci Rep. 2016;6:37887.
18. Guilinger JP, Pattanayak V, Reyon D, et al. Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. Nat Methods. 2014;11:429–435.
19. Erkes A, Reschke M, Boch J, Grau J. Evolution of transcription activator-like effectors in *Xanthomonas oryzae*. Genome Biol Evol. 2017;9:1599–1615.
20. Kim Y, Kweon J, Kim A, et al. A library of TAL effector nucleases spanning the human genome. Nat Biotechnol. 2013;31:251–258.
21. Miller JC, Zhang L, Xia DF, et al. Improved specificity of TALE-based genome editing using an expanded RVD repertoire. Nat Methods. 2015;12:465–471.
22. Streubel J, Blucher C, Landgraf A, Boch J. TAL effector RVD specificities and efficiencies. Nat Biotechnol. 2012;30:593–595.
23. Doyle EL, Booher NJ, Standage DS, et al. TAL effector-nucleotide targeter (TALE-NT) 2.0: Tools for TAL effector design and target prediction. Nucleic Acids Res. 2012;40:W117–W122.
24. Yang J, Zhang Y, Yuan P, et al. Complete decoding of TAL effectors for DNA recognition. Cell Res. 2014;24:628–631.
25. Morbitzer R, Romer 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.
26. Christian ML, Demorest ZL, Starker CG, et al. Targeting G with TAL effectors: A comparison of activities of TALENs constructed with NN and NK repeat variable di-residues. PLoS One. 2012;7:e45383.
27. Meckler JF, Bhakta MS, Kim MS, et al. Quantitative analysis of TALE-DNA interactions suggests polarity effects. Nucleic Acids Res. 2013;41:4118–4128.
28. Huang P, Xiao A, Zhou M, Zhu Z, Lin S, Zhang B. Heritable gene targeting in zebrafish using customized TALENs. Nat Biotechnol. 2011;29:699–700.
29. Reyon D, Tsai SQ, Khayter C, Foden JA, Sander JD, Joung JK. FLASH assembly of TALENs for high-throughput genome editing. Nat Biotechnol. 2012;30:460–465.
30. Reyon D, Khayter C, Regan MR, Joung JK, Sander JD. Engineering designer transcription activator-like effector nucleases (TALENs) by REAL or REAL-fast assembly. Curr Protoc Mol Biol. 2012;100:12.15.11–12.15.14.
31. Siggers T, Chang AB, Teixeira A, et al. Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF-kappaB family DNA binding. Nat Immunol. 2011;13:95–102.
32. 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.

33. 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.
34. Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. J Am Statist Assoc. 1993;88:1273–1283.
35. Zhao Y, Stormo GD. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat Biotechnol. 2011;29:480–483.
36. Lee PM. Bayesian Statistics: An Introduction. Hoboken, NJ: Wiley, 1997.
37. Hoffman MD, Gelman A. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. 2011;15: 1593-1623.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Anderson JT, Rogers JM, Barrera LA, Bulyk ML. Context and number of noncanonical repeat variable diresidues impede the design of TALE proteins with improved DNA targeting. *Protein Science*. 2020;29:606–616. https://doi.org/10.1002/pro.3801