Improved specificity of TALE-based genome editing using an expanded RVD repertoire

Jeffrey C Miller, Lei Zhang, Danny F Xia, John J Campo, Irina V Ankoudinova, Dmitry Y Guschin, Joshua E Babiarz, Xiangdong Meng, Sarah J Hinkley, Stephen C Lam, David E Paschon, Anna I Vincent, Gladys P Dulay, Kyle A Barlow, David A Shivak, Elo Leung, Jinwon D Kim, Rainier Amora, Fyodor D Urnov, Philip D Gregory & Edward J Rebar

Transcription activator–like effector (TALE) proteins have gained broad appeal as a platform for targeted DNA recognition, largely owing to their simple rules for design. These rules relate the base specified by a single TALE repeat to the identity of two key residues (the repeat variable diresidue, or RVD) and enable design for new sequence targets via modular shuffling of these units. A key limitation of these rules is that their simplicity precludes options for improving designs that are insufficiently active or specific. Here we address this limitation by developing an expanded set of RVDs and applying them to improve the performance of previously described TALEs. As an extreme example, total conversion of a TALE nuclease to new RVDs substantially reduced off-target cleavage in cellular studies. By providing new RVDs and design strategies, these studies establish options for developing improved TALEs for broader application across medicine and biotechnology.

Designed TALE proteins have rapidly gained prominence as versatile reagents for targeting user-chosen DNA sequences. Since the first description of rules for TALE-DNA recognition in 2009 (refs. 1,2), successive studies have established design principles[1–4], assembly methods[3,5–7] and a structural framework[8,9] for engineering this motif. Related efforts have also shown that TALEs can perform new functions when linked to exogenous domains, with particular attention focused on TALE-nuclease hybrids[10,11] (TALENs) because of their utility for genome editing in higher eukaryotes. Since the development of a highly active TALEN architecture[12], TALEN-mediated genome editing has been demonstrated in diverse species[13] and cell types, including human primary and stem cells[14–16]. These studies have established TALENs as attractive reagents for genome editing that are somewhat easier to engineer than zinc-finger nucleases yet offer substantially higher targeting densities (up to tenfold) than systems based on clustered, regularly interspaced, short palindromic repeats (CRISPR)-Cas9 (refs. 5,17,18).

The initial interest in designed TALEs—as well as the rapid progress of this field—has been motivated by the apparent simplicity and adaptability of TALE-DNA recognition. When bound to DNA, TALE proteins identify base sequences via contacts from a central array of TALE repeat units[8,9] with each unit specifying one base[1,2]. Repeats exhibit little diversity except at the RVD (positions 12 and 13), which recognizes the targeted base[1,2]. Critically, the base preference of a TALE repeat is substantially determined by the identity of its resident RVD. This enables the generation of TALEs with new, user-defined specificities via simple tandem assembly of repeats bearing the appropriate RVDs. In natural TALEs, the four most common RVDs[19]—NI, HD, NN and NG—tend to specify bases A, C, G/A and T, respectively[1,2]. With rare exception, TALE engineering efforts have relied on the RVD-base correspondences provided by this natural code (**Fig. 1a**,**b**).

Although the natural code provides a straightforward means for engineering new sequence preferences, the very simplicity of this correspondence may serve to limit the performance of TALEs created exclusively via this method. While simplifying the engineering process, the rigidity of these rules also restricts options for improving TALE behavior should a given design prove inadequate. In this respect, TALEs differ from other DNA-binding motifs that exhibit a greater functional complexity and that in turn present richer opportunities for fine-tuning recognition. As created using the natural code, TALENs can specify unintended bases in their binding sites[12,14,20,21] and also cleave nontargeted cellular sequences[14,15,20,22,23]. Although levels of off-target activity have not approached those described for the CRISPR-Cas9 platform[21,24–26], strategies for eliminating such behavior will likely be required to achieve the full potential of this motif, especially for highly sensitive applications such as human therapeutics.

In this study we have developed an expanded repertoire of RVDs for use in design and demonstrated the utility of these new RVDs for enhancing performance. We proceeded in four stages. First, in order to better understand the performance of current design methods, we characterized the binding specificity of a large panel of TALEs that use only the four canonical RVDs (NI, HD, NN and NG) to recognize DNA. This study revealed previously unappreciated complexities to TALE-DNA recognition, including

Sangamo BioSciences Inc., Richmond, California, USA. Correspondence should be addressed to E.J.R. (erebar@sangamo.com).

RECEIVED 8 MAY 2014; ACCEPTED 30 JANUARY 2015; PUBLISHED ONLINE 23 MARCH 2015; DOI:10.1038/NMETH.3330

discrete positional and sequence contexts that inhibit effective base sensing by each RVD. Second, in order to establish a foundation for new design methods, we mapped binding properties for the full RVD landscape by assessing the affinity and specificity of all 400 possible residue combinations. We identified dozens of non-natural RVDs that mediated efficient binding, many of which were as avid and specific for DNA recognition as the canonical RVDs. Next, we verified function by applying these new RVDs in a context-specific manner to improve the activities and specificities of previously described TALENs. Finally, we demonstrated that our new RVDs may be combined with other, mechanistically distinct TALEN design strategies, including a highly active, truncated architecture, to further improve genome-editing performance.

RESULTS

Reliability of the natural TALE code

Although the natural TALE code is often depicted as a binary matching of each RVD to its base target, previous studies had suggested that these preferences were not quantitatively absolute[12,21,27] and that they might vary with context[12]. In order to characterize this behavior, we used systematic evolution of ligands by exponential enrichment (SELEX)[28] to determine the consensus binding preferences for a large panel of synthetic TALEs designed using canonical RVDs (76 proteins, >250 examples of each RVD; **Supplementary Table 1**). We then examined these data for RVD-base fidelity. Our results revealed considerable variation in the ability of each RVD to discriminate target sequence. Base compositions selected by NN, for example, spanned a wide quantitative range, from predominantly adenine to almost exclusively guanine (for example, **Fig. 1c**). Likewise, base preferences for the other RVDs also varied (**Supplementary Fig. 1a**). Some of this variability could be attributed to identifiable elements of context. For example, the N-terminal repeat frequently selected adenine irrespective of its resident RVD (**Fig. 1c**,**d**), whereas NI exhibited a reduced preference for adenine in the context of the C-terminal half repeat (**Supplementary Figs. 1b** and **2**). Neighboring bases and repeats also influenced specificity (**Fig. 1e** and **Supplementary Figs. 1b** and **3**). In many cases, however, the reasons for degraded or alternative base preferences were obscure. Such behavior limits the predictability of TALEs built using the canonical RVDs and underscores the need for additional design options.

A comprehensive survey of RVD affinity and specificity

As the foundation of our efforts to develop new design strategies, we first sought to map the base recognition properties of every possible RVD. By so doing, we hoped to identify new candidates for use in design and to also discern patterns of affinity and specificity that might provide functional insights. To achieve this, we assembled each RVD (400 total) into the fifth repeat of a host TALE protein (**Fig. 2a**) and then assayed for binding to all four bases at the corresponding sequence position within the host target. Binding studies were performed using a high-throughput ELISA procedure. In choosing this test system, we sought to maximize dynamic range while reducing context effects

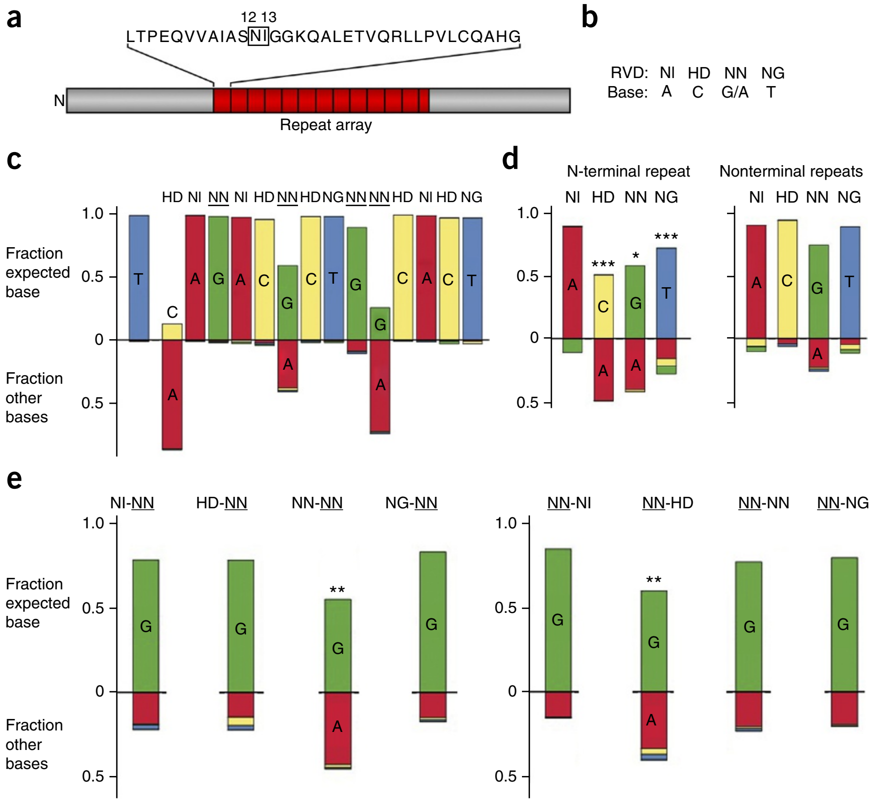


Figure 1 | Design and specificities of TALEs generated using the natural TALE code. (**a**) Sketch of a natural TALE highlighting the central repeat array that mediates DNA recognition (red boxes). A typical repeat sequence is provided above in single-letter amino acid code, with a square enclosing the RVD that determines base preference (positions 12 and 13). Flanking protein segments are shaded in gray. "N" denotes the amino terminus. (**b**) RVD-base correspondences that constitute the natural TALE code. (**c**) Graphical depiction of a SELEX-derived base-frequency matrix for a synthetic TALE. At each matrix position, the frequency of the intended target base is projected above the *x* axis, whereas remaining frequencies are plotted below the *x* axis. Matrix positions are arranged in 5′ to 3′ order with corresponding RVDs listed above (14 total). Frequencies are derived from 367 aligned sequences. Underlining highlights the NN RVD, which exhibits a variable preference for guanine vs. adenine. Note that the N-terminal flanking segment specifies the 5′ thymine base. (**d**) Average base preference for each canonical RVD when used within either the N-terminal repeat (left) or nonterminal repeats (right) of an engineered TALE. HD, NN and NG exhibit enhanced binding to adenine in the context of the N-terminal repeat (***$P < 10^{-6}$; *$P < 0.002$). (**e**) Average base preferences for NN when flanked by each possible canonical RVD, showing that neighbor identity affects selectivity for guanine vs. adenine. **$P < 0.0002$ with respect to guanine preference vs. each other context in the panel. Values in **d** and **e** are derived from SELEX studies of 76 TALEs summarized in **Supplementary Table 1**. For plotted values and sample sizes, see **Supplementary Figures 2** and **3**. All *P* values were calculated with the Mann-Whitney test corrected for false discovery rate.

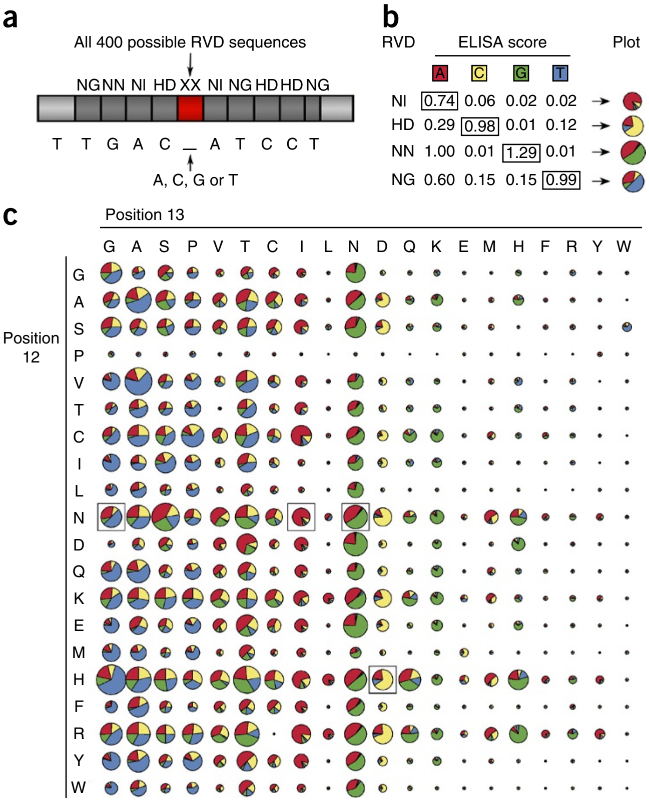


Figure 2 | Comprehensive survey of RVD-DNA binding properties. (**a**) Top, diagram of proteins generated for this study. A ten-repeat host TALE was diversified into 400 proteins bearing each possible RVD in its fifth repeat (red box; XX denotes varied RVD composition). Remaining repeats were held constant (dark gray boxes; two-letter codes indicate each RVD). Bottom, targets used for this study. Each TALE was assayed for binding to four variants of the host target in which the base contacted by the fifth repeat (underscore) was A, C, G or T. (**b**) Example of assay data for the four canonical RVDs. Four proteins, each bearing the indicated RVD in its fifth repeat, were screened by ELISA for binding to the A, C, G or T variants of the host target. ELISA signals were normalized to the average value obtained for the interaction of each code RVD with its preferred base (i.e., values on the diagonal, highlighted by boxes). Graphical depictions of binding data are provided at right. The size of each circular plot is proportional to the ELISA signal of the indicated RVD with its preferred base (i.e., affinity), and the areas of each colored wedge indicate relative binding activities for each base target (i.e., specificity). (**c**) Graphical summary of binding data for all RVDs and targets. RVDs are identified by residue type at position 12 and 13, with residues listed in order of increasing volume. ELISA data for each RVD are summarized by a circular plot at the corresponding location in the 20 × 20 matrix. Each value is the average of four measurements. Boxes highlight data corresponding to the natural code RVDs (NI, HD, NN and NG). For plotted values, see **Supplementary Table 2**.

that might confound analysis. Accordingly, we limited the length and affinity of the host TALE (ten repeats) and tested new RVDs individually given previous reports of unexpectedly poor binding of TALEs bearing multiple tandem identical repeats[4,29]. We also avoided flanking repeats that appeared likely to interact with the queried RVD as determined by our analysis of neighbor effects (**Supplementary Fig. 3**). The resulting data set (**Fig. 2b**,**c** and **Supplementary Table 2**) mapped the relative affinity for every combination of RVD and base (1,600 total) onto a scale spanning a 100-fold dynamic range.

Our results revealed an RVD landscape that is more crowded with binding-competent residue combinations than might be expected from RVD frequencies observed in natural TALEs[19] (**Fig. 2c**). Many residue combinations that are rare or absent in nature matched the activity or specificity of the canonical RVDs, suggesting that additional criteria determine natural prevalence. Our results also illuminated distinct functional roles for each RVD position suggested by structural studies[8,9]. Position 13, for example, largely determines base preference. This is most evident for residues G, I, D and N, which specify thymine, adenine, cytosine and guanine/adenine (respectively) throughout each of their respective RVD families. Residues H and K also tend to specify guanine, albeit with generally weaker affinities, whereas A and P exhibit a variable preference for thymine. Position 12, in contrast, primarily modulates binding strength, in some cases over a more than 50-fold range of relative ELISA signal, with modest or minimal effects on base preference. Overall these results revealed an RVD landscape populated with residue combinations that exhibit a variety of binding properties that could be applied for designing new TALEs with tailored affinities or base preferences.

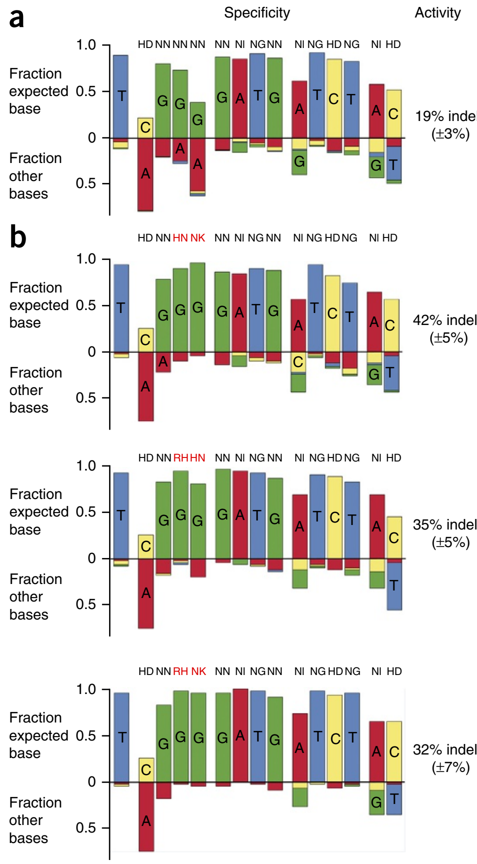
Cellular activity of noncanonical RVDs

Having identified non-natural RVDs with encouraging biochemical properties, we next sought to verify function in a more relevant context for TALE design: a chromosomal locus in a human cell. We pursued this via two complementary studies. In the first, noncanonical RVDs were swapped into key repeats of a previously described TALEN[14] in an effort to improve activity and specificity. Substitutions focused on the two least specific guanine-targeted repeats as gauged by SELEX. Cellular screens identified three variants that substantially improved gene modification activity and specificity relative to the parent TALEN (**Fig. 3** and **Supplementary Fig. 4**). This work verified cellular function of our new RVDs as well as their utility for improving an otherwise canonical TALEN. In a second study, we sought to gauge performance of the new RVDs when deployed as alternative codes. To this end, we redesigned a previously characterized TALEN (L538 (ref. 12), hereafter referred to as L) targeted to the human *CCR5* gene using variations of the natural code that employed progressively more noncanonical RVDs, and we tested the resultant proteins for cellular function. This study identified diverse alternative codes that yield highly active TALENs, including several that use exclusively noncanonical RVDs (**Supplementary Fig. 5**). These results demonstrated that canonical RVDs, although highly represented in nature, are not obligate components of the TALE-DNA interface and that alternative RVDs can be deployed using code-based rules to generate highly active TALEs.

Using new RVDs for context-specific design

Although an RVD code—natural or otherwise—may provide a straightforward means for designing new TALEs, it will not necessarily yield the most active or specific TALE for a given target. This is because all codes inherently trade design complexity for ease of use, and there is no reason to expect that the best TALE for a target will belong to the minor fraction of possible designs that conform to a simple code. The availability of an expanded set of RVDs provides an opportunity to improve upon simple codes in the design of new TALEs. Through use of richer arrays of

Figure 3 | Improvement of TALE properties via substitution of key RVDs. (**a**) SELEX-derived base-frequency matrix and gene-modification activity of an all-canonical TALEN targeted to human *PITX3* (ref. 14). (**b**) SELEX profiles and gene-modification activities of variant TALENs in which key repeats bear noncanonical RVDs (red). RVD identities and base frequency data are presented graphically as in **Figure 1c**. Base frequencies are derived from at least 45 aligned sequences. "Activity" indicates the percentage of insertions or deletions, or indels (±s.d. derived from three biological replicates), induced upon delivery to K562 cells with its partner TALEN[14]. For full screening data, see **Supplementary Figure 4**.



RVDs for target recognition, with choices optimized for context-specific performance, it should be possible to construct TALEs that exhibit improved binding properties relative to those of their code-derived counterparts.

To test this possibility we generated new versions of the L TALEN and its partner R (referred to as R557 in a prior study[12]) that use a total of 16 distinct RVDs for DNA recognition (**Fig. 4a**,**b**, **Supplementary Figs. 6** and **7** and **Supplementary Tables 3–6**). These new TALENs (hereafter called L* and R*) induced high levels of modification at their cellular target (**Fig. 4c**) without employing any canonical RVDs. We then evaluated the cellular specificity of these new TALENs and their canonical parents via deep sequencing of 23 candidate off-target sites from nuclease-treated K562 cells (**Supplementary Fig. 8** and **Supplementary Tables 7** and **8**). This study included a post-transfection cold shock[30] in order to drive higher levels of cleavage for analysis. Under these conditions both nuclease pairs modified the intended target to similarly high levels (**Fig. 5a**). However, the new TALENs produced substantially lower off-target cleavage (**Fig. 5b**,**c**). Whereas the canonical L/R pair detectably modified 13 off-target loci with an aggregate frequency of 11%, the new L*/R* TALENs modified fewer loci, with an aggregate frequency of 0.55% (**Supplementary Fig. 9** and **Supplementary Table 9**). Individual loci exhibited up to 180-fold reduced cleavage by L*/R* (**Fig. 5b**). Moreover, mixed dimers (L/R* and L*/R) yielded intermediate behaviors (**Fig. 5d**), indicating that both L* and R* contributed to the reduction in off-target cleavage. As an additional test of specificity, all TALENs were submitted for SELEX analysis. This assay revealed specific binding for all four proteins, but with L* and R* exhibiting a higher average percent match to their intended targets (**Supplementary Fig. 10** and **Supplementary Table 10**).

Combining TALEN improvement strategies

The development of improved TALENs should benefit from combining RVD substitutions with other, mechanistically distinct, strategies for enhancing performance. As an initial test of this possibility, we sought to deploy our new RVDs in the context of a TALEN architecture bearing a shorter C-terminal region. As previously shown[12], C-terminal truncations can improve gene-editing activity and increase cleavage specificity for a more limited range of dimer configurations. For this study we reduced the C-terminal region to 17 residues because this length supported peak activity in a prior truncation scan (**Supplementary Fig. 11**). This '+17' architecture also excludes a region of positive charge that was recently implicated in nonspecific binding[23] (**Supplementary Fig. 11a**). Accordingly, we combined the L* and L repeat arrays with the +17 architecture and then tested the resulting TALENs (L*+17 and L+17) for cellular activity and specificity when paired with a partner (R2+17) that enables maximal activity (**Supplementary Fig. 11c**). We observed that L*+17 and L+17 exhibited comparable levels of on-target activity (>85% indels),

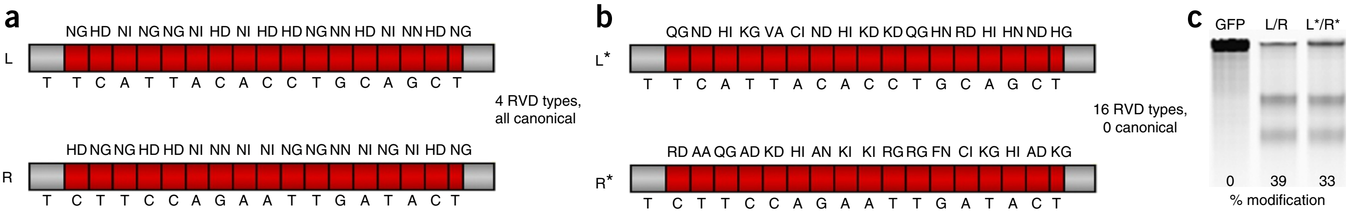

a
NG HD NI NG NG NI HD NI HD HD NG NN HD NI NN HD NG
L
T T C A T T A C A C C T G C A G C T
4 RVD types, all canonical
HD NG NG HD HD NI NN NI NI NG NG NN NI NG NI HD NG
R
T C T T C C A G A A T T G A T A C T
b
QG ND HI KG VA CI ND HI KD KD QG HN RD HI HN ND HG
L*
T T C A T T A C A C C T G C A G C T
16 RVD types, 0 canonical
RD AA QG AD KD HI AN KI KI RG RG FN CI KG HI AD KG
R*
T C T T C C A G A A T T G A T A C T
c
GFP
L/R
L*/R*
0
39
33
% modification

Figure 4 | Design and cleavage activities of *CCR5*-targeted TALENs. (**a**) Canonical TALENs L and R (L538 and R557, respectively, from a prior study[12]), which use standard RVDs for target recognition. TALENs are depicted as in **Figure 1a**, except the gray flanking segments have been shortened for clarity. DNA target sequences are provided below each TALEN. RVD identities are indicated above each repeat. (**b**) Noncanonical TALENs L* and R*, which use exclusively noncanonical RVDs for target recognition. The generation of these TALENs is described in **Supplementary Figures 6** and **7**. (**c**) Gene-modification activities. The L/R or L*/R* dimers were delivered to K562 cells, and then gene-modification activity was assessed via the Surveyor assay. This comparison was performed 11 times; a representative result is shown.

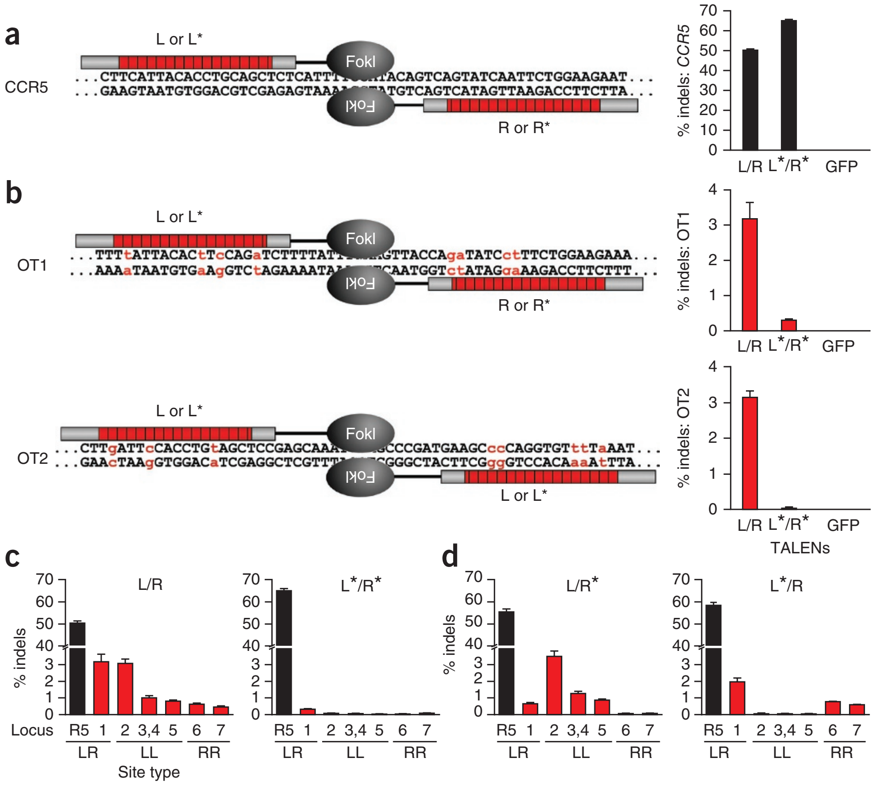


Figure 5 | Reduced modification of off-target loci in cells exposed to RVD-diversified TALENs. (**a**) Sketch of the TALENs used for this study in complex with the intended cleavage target in the *CCR5* gene. *CCR5* modification levels induced by L/R, L*/R* or a GFP negative control are plotted at right. (**b**) Sketch of TALENs in complex with the two most active off-target loci examined in this study (OT1 and OT2). Red letters indicate bases that diverge from the intended target sequence. Note that OT2 is a cleavage target for a homodimer of 'left' TALENs (i.e., L/L or L*/L*). Modification levels observed at each locus are plotted at right. (**c**) Modification levels at seven of the more active off-target sites (OT1–OT7, labeled 1–7) and the intended *CCR5* target (R5) in cells exposed to L/R and L*/R*. LR, LL and RR indicate the hetero- or homodimer TALEN species predicted to cleave each target. Note that data for OT3 and OT4 were merged because these loci differ by one single-nucleotide polymorphism and frequently cannot be distinguished after acquisition of an insertion or deletion. (**d**) Modification levels at OT1–OT7 in cells exposed to the mixed dimers L/R* and L*/R. In all panels, error bars indicate the s.d. of three determinations of modification level via deep sequencing of genomic DNA pooled from 12 replicate transfections. For complete study results, see **Supplementary Figure 9**.

but L*+17 had a >100-fold reduction in cleavage at the sole active L+17 off-target site (**Supplementary Fig. 12** and **Supplementary Tables 7** and **11–13**). This study underscores the potential for combining new RVDs with other strategies to improve TALEN performance. Consistent with this finding, in a second study we observed that three previously described specificity-enhancing mutations (the 'Q3' subsitutions[23]) could be introduced into the C-terminal regions of L* and R* to yield an additional fivefold drop in aggregate off-target cleavage (**Supplementary Fig. 13**).

DISCUSSION

Although TALEs provide a versatile platform for designing new DNA-binding proteins, most applications to date have relied on several largely untested assumptions, namely repeat modularity, code-based design, and the superiority of the four most prevalent natural RVDs. In this study, we have examined these premises via large-scale assembly and testing of new TALE proteins and RVD repeat units. These efforts have revealed not only previously unappreciated complexities to TALE design but also strategies for improving performance. Our findings should enable the generation of TALE proteins with improved properties, both via a more sophisticated application of the natural code and through the use of more complex design methods.

Our SELEX studies, for example, have provided insights into positional and sequence contexts that inhibit base sensing by each canonical RVD. While broadly confirming the predictability of TALE-DNA recognition, these studies showed that 6% of queried repeats selected a majority of one or more unintended bases, and a further 7% exhibited relaxed specificity (aggregate base preference of <2:1). Notably, these instances of poor performance often occurred in contexts that were readily identifiable and therefore avoidable via prescreening of targets. We estimate that the frequency of repeats with unintended or relaxed base preferences could be reduced by half by choosing sites in which the first repeat recognizes adenine and also by avoiding use of the NN RVD in its least favorable contexts. Further reductions may be achieved by avoiding targets bearing a 3′ adenine or runs of three or more successive thymines. In cases where target availability is not limiting, application of these simple filters should broadly reduce the chance that a code-based design will exhibit an undesired base preference.

As a complementary approach to improving outcomes, we also sought to expand the repertoire of binding-validated RVDs available for TALE design. To this end, we characterized the binding properties of all 400 possible RVDs. This effort identified dozens of binding-competent RVDs, including at least 31 whose biological or biochemical properties recommend them as useful alternatives to canonical RVDs (**Table 1**). Twenty of these combinations are new in that they have been neither observed in natural TALEs nor

Table 1 | Alternative RVDs for TALE design

Target base	RVDs with improved properties characterized in this study
Thymine	HG[a–e], VG[b], IG[b], EG[b], MG[b], YG[b,d], AA[a,c,e], EP[b], VA[a–c], QG[b,c], KG[c,e], RG[c,e]
Guanine	GN[b], SN[b,d,f], VN[b], LN[b], DN[b], QN[b], EN[b], HN[c–g], RH[b,g], NK[b,g,h], AN[c], FN[c]
Adenine	CI[c], HI[c,d], KI[c]
Cytosine	RD[c], KD[c], ND[b–e], AD[c]

[a]ELISA study indicates improved affinity. [b]ELISA study indicates improved target base preference. [c]Used in improved *CCR5*-targeted TALENs. [d]Previously highlighted in ref. 19. [e]Previously highlighted in ref. 29. [f]Previously highlighted in ref. 4. [g]Study summarized in **Figure 3** shows improved target preference and cellular activity. [h]Previously highlighted in multiple studies.

identified as design alternatives in prior RVD studies[4,19,29]. These new RVDs could provide useful alternatives for base recognition in contexts where canonical RVDs are insufficiently specific. In this regard, recent studies seeking enhanced guanine selectivity[4,19,31] could be viewed as the first applications of alternative RVDs for addressing the most well-recognized shortcoming of canonical TALEs: the context-dependent inability of NN to resolve G from A.

Our studies identified a considerable excess of active RVDs beyond those observed in natural TALEs. Because these new RVDs were validated in biologically active and specific TALENs, our results suggest that additional factors determine natural prevalence. In light of *Xanthomonas* biology, one explanation may be that natural selection has favored modularity over specificity during TALE evolution. If natural TALEs emerge via recombination of existing repeat arrays[32,33], then RVD qualities that increase the productivity of such events—such as modularity—would provide a selective advantage in countering adversities such as target loss in a resistant host. Conversely, selection pressure for specificity should be reduced given that natural TALEs operate exclusively within the host genome. The natural code may therefore represent the most modular subset of active RVDs.

These considerations suggest that it should be possible to improve upon canonical TALEs via use of an expanded RVD repertoire for target recognition. To demonstrate this concept, we identified variants of a well characterized TALEN dimer[12] that used exclusively noncanonical RVDs for DNA recognition and exhibited improved specificity. As an additional test, we replaced two poorly performing RVDs in a previously published TALEN and observed improvements in cellular activity and biochemical specificity. This latter result was achieved via a single-step screen of a small panel of new designs, and we envision that this study represents a typical example of how our new RVDs will be applied.

Although it provides a powerful approach for engineering new TALEs, the natural code—by its very simplicity—limits options for improving performance. Perhaps as a consequence, most efforts to improve TALEN specificity have focused on other aspects of design such as removing excess charge[23], truncating the C-terminal region[7,12,31,34], eliminating homodimer activity[35,36] and engineering the N-terminal region to alter its specificity for thymine[37,38]. (Other studies directed at base-specific contacts either have not examined cleavage specificity[29] or have been more limited in scope, for example, searching for improved guanine preference[4,19,31].) As our approach focuses on a distinct region of the TALEN structure, we expect that it will be combinable with other strategies for optimizing TALEN properties. To test this possibility, we showed that our new RVDs may be combined with key substitutions in the C-terminal region[23] to improve specificity. We also demonstrated improved performance in the context of a more highly truncated Fok attachment point (carboxy-residue 17). This latter result is especially notable as this +17 architecture may prove to be ideal for genome editing given its improved activity and reduced range of dimer binding configurations that allow productive cleavage compared to the current standard architecture[12]. The results presented here should broadly facilitate the application of designed TALEs to more diverse and challenging fields.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Abrahamson for critically reviewing this manuscript. We thank M. Holmes and H. Zhang for helpful discussions.

AUTHOR CONTRIBUTIONS

J.C.M., L.Z. and E.J.R. designed experiments and supervised studies. J.J.C., J.E.B., D.F.X., I.V.A., K.A.B., J.D.K. and R.A. performed studies. J.C.M., D.A.S. and E.L. analyzed data. J.J.C., D.F.X., I.V.A., D.Y.G., X.M., S.J.H., S.C.L., D.E.P., A.I.V., G.P.D. and L.Z. developed new procedures and constructs. J.C.M., P.D.G., F.D.U. and E.J.R. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Moscou, M.J. & Bogdanove, A.J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326**, 1501 (2009).
2. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509–1512 (2009).
3. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* **39**, e82 (2011).
4. Streubel, J., Blücher, C., Landgraf, A. & Boch, J. TAL effector RVD specificities and efficiencies. *Nat. Biotechnol.* **30**, 593–595 (2012).
5. Reyon, D. *et al.* FLASH assembly of TALENs for high-throughput genome editing. *Nat. Biotechnol.* **30**, 460–465 (2012).
6. Schmid-Burgk, J.L., Schmidt, T., Kaiser, V., Höning, K. & Hornung, V. A ligation-independent cloning technique for high-throughput assembly of transcription activator–like effector genes. *Nat. Biotechnol.* **31**, 76–81 (2013).
7. Kim, Y. *et al.* A library of TAL effector nucleases spanning the human genome. *Nat. Biotechnol.* **31**, 251–258 (2013).
8. Deng, D. *et al.* Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **335**, 720–723 (2012).
9. Mak, A.N., Bradley, P., Cernadas, R.A., Bogdanove, A.J. & Stoddard, B.L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335**, 716–719 (2012).
10. Christian, M. *et al.* Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**, 757–761 (2010).
11. Li, T. *et al.* TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res.* **39**, 359–372 (2011).
12. Miller, J.C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29**, 143–148 (2011).
13. Sun, N. & Zhao, H. Transcription activator-like effector nucleases (TALENs): a highly efficient and versatile tool for genome editing. *Biotechnol. Bioeng.* **110**, 1811–1821 (2013).
14. Hockemeyer, D. *et al.* Genetic engineering of human pluripotent cells using TALE nucleases. *Nat. Biotechnol.* **29**, 731–734 (2011).
15. Osborn, M.J. *et al.* TALEN-based gene correction for epidermolysis bullosa. *Mol. Ther.* **21**, 1151–1159 (2013).
16. Ding, Q. *et al.* A TALEN genome-editing system for generating human stem cell-based disease models. *Cell Stem Cell* **12**, 238–251 (2013).
17. Tsai, S.Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* **32**, 569–576 (2014).
18. Guilinger, J.P., Thompson, D.B. & Liu, D.R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat. Biotechnol.* **32**, 577–582 (2014).
19. Cong, L., Zhou, R., Kuo, Y.C., Cunniff, M. & Zhang, F. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nat. Commun.* **3**, 968 (2012).
20. Tesson, L. *et al.* Knockout rats generated by embryo microinjection of TALENs. *Nat. Biotechnol.* **29**, 695–696 (2011).

21. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31**, 833–838 (2013).
22. Juillerat, A. *et al.* Comprehensive analysis of the specificity of transcription activator-like effector nucleases. *Nucleic Acids Res.* **42**, 5390–5402 (2014).
23. Guilinger, J.P. *et al.* Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. *Nat. Methods* **11**, 429–435 (2014).
24. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
25. Hsu, P.D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
26. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
27. Bogdanove, A.J. & Voytas, D.F. TAL effectors: customizable proteins for DNA targeting. *Science* **333**, 1843–1846 (2011).
28. Blackwell, T.K. & Weintraub, H. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science* **250**, 1104–1110 (1990).
29. Yang, J. *et al.* Complete decoding of TAL effectors for DNA recognition. *Cell Res.* **24**, 628–631 (2014).
30. Doyon, Y. *et al.* Transient cold shock enhances zinc-finger nuclease-mediated gene disruption. *Nat. Methods* **7**, 459–460 (2010).
31. Christian, M.L. *et al.* Targeting G with TAL effectors: a comparison of activities of TALENs constructed with NN and NK repeat variable di-residues. *PLoS ONE* **7**, e45383 (2012).
32. Yang, Y. & Gabriel, D.W. Intragenic recombination of a single plant pathogen gene provides a mechanism for the evolution of new host specificities. *J. Bacteriol.* **177**, 4963–4968 (1995).
33. Yang, B., Sugio, A. & White, F.F. Avoidance of host recognition by alterations in the repetitive and C-terminal regions of AvrXa7, a type III effector of *Xanthomonas oryzae* pv. *oryzae*. *Mol. Plant Microbe Interact.* **18**, 142–149 (2005).
34. Mussolino, C. *et al.* A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res.* **39**, 9283–9293 (2011).
35. Doyon, Y. *et al.* Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *Nat. Methods* **8**, 74–79 (2011).
36. Miller, J.C. *et al.* Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *Nat. Biotechnol.* **25**, 778–785 (2007).
37. Lamb, B.M., Mercer, A.C. & Barbas, C.F. III. Directed evolution of the TALE N-terminal domain for recognition of all 5′ bases. *Nucleic Acids Res.* **41**, 9779–9785 (2013).
38. Doyle, E.L. *et al.* TAL effector specificity for base 0 of the DNA target is altered in a complex, effector- and assay-dependent manner by substitutions for the tryptophan in cryptic repeat-1. *PLoS ONE* **8**, e82120 (2013).

ONLINE METHODS

TALE constructs. The sequence of each TALE examined in the large scale SELEX study is provided in **Supplementary Table 14**, along with sequences of plasmid constructs used for this study. Unless otherwise noted, all constructs were generated using standard molecular biology methods.

For large scale RVD ELISA studies (**Fig. 2** and **Supplementary Table 2**), each TALE construct was assembled by annealing two complementary oligonucleotides (5′-AATCGCGTCGxxxxxxGGGGGAAAG and 5′-CTTGCTTTCCCCCyyyyyyCGACGC, where the varied region encodes each possible RVD) and ligating the resultant duplex into the vector pVAXt-TALE-TGAC-Bsa2-ATCC-C63-Fok, which had been linearized via digestion with BsaI. The sequence of pVAXt-TALE-TGAC-Bsa2-ATCC-C63-Fok is provided in **Supplementary Table 14**. The sequences of all TALEs examined in this study are summarized in the legend of **Supplementary Table 2**.

Constructs for tetramer ELISA studies (**Supplementary Table 3**) were generated in two steps. First, TALE repeat monomer vectors, encoding the noncanonical RVDs highlighted in red in **Supplementary Figure 7**, were created by ligating the annealed oligonucleotides shown in **Supplementary Table 15** into the position-specific monomer hosting vectors, pTAL-Bsa-BsmBI-M1, pTAL-Bsa-BsmBI-M2, pTAL-Bsa-BsmBI-M3 or pTAL-Bsa-BsmBI-M4 (**Supplementary Table 14**), which had been linearized via digestion with BsmBI. Next, amplicons encoding each monomer repeat were amplified from these vectors using the primers TMF (5′-CGGCCTGCGTCGACGAATTCG) and TMR (5′-CACAGCCTGAGCTCTCTAGAG), pooled as shown in **Supplementary Figure 7**, digested with BsaI and ligated into a linearized host vector. Tetramer pools corresponding to the TCAT and CTTC targets were ligated into their host vector as repeats 1–4 of a nine-repeat TALE, and tetramer pools corresponding to the TACA, CCTG, CAGC, CAGA, ATTG and ATAC targets were ligated into a different host vector as repeats 5–8 of a nine-repeat TALE. Ligations were transformed into TOP10 *Escherichia coli* competent cells (Life Technologies), and individual clones were randomly picked, sequenced and archived for analysis in the tetramer ELISA studies summarized in **Supplementary Table 3**. An overview of the resulting vector sequences is provided in **Supplementary Table 14**, and TALE sequences examined in the ELISA studies are summarized in the legend of **Supplementary Table 3**.

To generate *CCR5*-targeted TALEN constructs bearing exclusively noncanonical RVDs, DNA fragments encoding high-affinity tetramers from **Supplementary Table 3** were PCR amplified using primers specific to the ultimate tetramer position in the TALEN transgene (**Supplementary Table 15**). PCR products were pooled as appropriate, digested with BsaI and ligated into the TALEN expression vector, pVAXt-3Flag-NLS-TALE-Bsa3-C63-Fok (provided in **Supplementary Table 14**), which had been linearized via digest with BsaI. Ligations were transformed into TOP10 *E. coli* competent cells, and single clones were randomly picked and sequenced. A summary of the resulting encoded TALEN sequences is provided in **Supplementary Tables 4** and **6**.

A similar procedure was used to generate the constructs summarized in **Supplementary Table 5** and **Supplementary Figure 5**, except that (i) tetramers coding for unique TALENs were used instead of mixed pools, and (ii) a discrete reverse primer—which encodes the RVD of the final half repeat—was used for the tetramer 4 PCR. Primer sequences and the resulting RVD encoded by the final half repeat can be found in **Supplementary Table 15**. A summary of the encoded TALEN sequences is provided in **Supplementary Tables 5** and **14**.

SELEX studies. Experimental conditions used to characterize TALENs in **Figure 3** were essentially as described[12]. Briefly, an oligonucleotide target library was synthesized bearing the sequence: 5′-CAGGGATCCATGCACTGTACGCCCNNNNNNNNNNNNNNNNNNNNNNNNNGGGCCACTTGACTGCGGATCCTGG, where N denotes a mixture of all four bases. The library was converted to double-stranded duplex by annealing 2 nmol of library oligo with 16 nmol of 3′ library primer (5′-CCAGGATCCGCAGTCAAGTGG) in 100 μl of 1× PCR Master (Roche) supplemented to 2.5 mM of each dNTP and 5 mM $MgSO_4$. This was followed by incubation at 95 °C for 2 min, 94 °C for 5 min, 58 °C for 5 min and 72 °C for 15 min.

DNA fragments encoding TALE proteins were amplified via PCR using a 5′ primer bearing a T7 promoter: 5′-GCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACGAATTCACCACCATGGTGGATCTACGCACG CTCG-3′ and a 3′ primer encoding an in-frame hemagglutinin (HA)-epitope tag: 5′-CACGTACTTCAGCTTTTATTAGGCGTAGT CGGGCACGTCGTAGGGGTAGCCCGCGACTCGATGGGAAGTTC-3′. Protein was then expressed by adding 2 μl of the PCR product to 10 μl TnT coupled transcription-translation system (Promega). After incubation at 30 °C for 80 min, the resulting reaction was mixed with 200 pmol of library duplex and SELEX buffer (0.01% BSA, 0.05% Tween 20, 0.5 mM $MgCl_2$, 20 μg/ml poly dIdC in calcium-free PBS) in a final volume of 90 μl and incubated for 30 min at room temperature. Next, 3 μl Roche anti-HA-biotin clone 3F10 (diluted in a total volume of 1 ml H_2O) and 7 μl SELEX buffer were added to each reaction. After an additional incubation for 20 min, protein-DNA-antibody complexes were captured on streptavidin-coated magnetic beads (Invitrogen). Bound DNA target was then amplified via PCR using the 3′ library primer and 5′ library primer (5′-CAGGGATCCATGCACTGTACG) and subjected to additional cycles of enrichment. A total of three rounds of enrichment were used before cloning and sequencing selected sequences.

For characterization of the TALENs in **Figure 1**, **Supplementary Figure 1** and **Supplementary Table 1**, an alternative SELEX protocol was used that had been optimized for performance at the larger scale of these studies. The randomized duplex library was generated essentially as described above, except that 6 nmol of 3′ library primer and 1.2 mM of each dNTP were used. For the first assay cycle, TALENs were expressed directly from plasmid templates using a TnT coupled transcription-translation system (Promega) and the manufacturer's recommended conditions with buffers supplemented to 10 mM $ZnCl_2$. Expressed TALENs contained a triple Flag tag fused to their N termini. 12 μl of TnT reaction mix was then mixed with 200 pmol of library duplex in a total volume of 100 μl of binding buffer (50 mM DTT, 10 μM $ZnCl_2$, 5 mM $MgCl_2$, 0.01% BSA fraction V, 100 mM NaCl in PBS (calcium free)). After incubation for 50 min, protein-DNA complexes were captured on anti-Flag M2 magnetic beads (Sigma) and washed five times with wash buffer (5 mM DTT, 10 μM $ZnCl_2$, 5 mM $MgCl_2$, 0.01% BSA fraction V, 100 mM NaCl in PBS (calcium free)). Bound target was PCR amplified using the 3′ library primer (above) and a 5′ library primer

doi:10.1038/nmeth.3330

(5′-CAGGGATCCATGCACTGTACG), and the resulting amplicon was used as input for additional cycles of enrichment. Protein expression and binding conditions for these subsequent cycles were identical to the conditions used in the first round. After three cycles, recovered DNA fragments were sequenced using an Illumina MiSeq system. The protocol for adding the Illumina sequencing primers and sequencing is as described in the section "Off-target analysis."

Identical conditions were used to characterize L, L*, R and R* (data in **Supplementary Fig. 10**) except that 48 μl of TnT extract was used for each binding reaction.

SELEX FASTQ sequences from the MiSeq were adaptor trimmed using SeqPrep (J. St. John, unpublished, https://github.com/jstjohn/SeqPrep). SELEX library sequences were further filtered by custom Python scripts for correct length and fixed flanking region composition (exact match). 1,000 randomly sampled filtered sequences were used as input to the GADEM motif discovery program with options maskR=0 fullscan=0 gen=3. Position frequency matrices discovered by GADEM[39] were then aligned to the intended sequence and reverse complemented if necessary. Matrices longer than the intended sequence were trimmed to only those regions overlapping the intended sequence according to the highest-scoring alignment, yielding the final matrices provided in **Supplementary Table 1**.

Large-scale ELISA study. Full sequences of the DNA binding site duplexes used in this assay are provided in **Supplementary Table 15**. To generate the binding-site premix used in each ELISA reaction, we annealed 100 pmol of the corresponding oligonucleotide to 12.5 pmol of 5′ biotinylated primer (5′-Biotin-GACGTGTGGACTGACTGTGA) and then incubated this under the following conditions with 25 μl 1× Accuprime buffer (Invitrogen)/Accuprime *Taq* enzyme: 94 °C, 3 min (1×); 94 °C, 30 s (1×); 55 °C, 30 s and 68 °C, 2 min (3×); 68 °C, 3 min (1×). Biotinylated binding sites (0.375 pmol) were combined with 165 ng salmon sperm (Invitrogen) and 2.28 mU of anti-HA-peroxidase high affinity (Roche) in a total volume of 55 μl of ELISA binding buffer (0.5 mM $MgCl_2$, 10 μm $ZnCl_2$, 0.5% Tween 20 and 0.01% BSA).

To run the assay, we amplified fragments encoding TALE proteins by PCR using a 5′-primer sequence (5′-GCAGAGC TCTCTGGCTAACTAGAG-3′) and a 3′ primer encoding an in-frame HA-epitope tag (5′-GCGTAAAGCTTAGGCGTAG TCGGGCACGTCGTAGGGGTAGCCGGGCACCAGCTGG GATCCCCG CAGGTG). Proteins were expressed by adding the PCR product to a TnT coupled transcription-translation system (Promega) supplemented with methionine and $ZnCl_2$ (final concentrations of 20 μM and 330 μM, respectively) and incubating as specified by the manufacturer. The binding-site premix was then combined with the TnT reaction and incubated for 40 min. Streptavidin-coated high–binding capacity black 96-well plates (Pierce) were washed with PBS solution (supplemented with $MgCl_2$ and $ZnCl_2$ to final concentrations of 0.5 mM and 10 μM, respectively) using an ELx405 automated plate washer. DNA-protein-antibody complexes were captured by transferring the TnT–binding site mixture to the plate, incubating for 40 min and washing with ELISA binding buffer using the ELx405. Immediately following the wash, 45 μl of QuantaBlue substrate solution (Pierce) and 5 μl of QuantaBlue stable peroxidase solution (Pierce) were added to each well and developed for 30 min. Plates were read using a Gemini XS plate reader following the QuantaBlu protocol. Data were exported from the plate reader and analyzed in Microsoft Excel.

TALE tetramer ELISA study. TALE tetramer constructs were screened via an identical set of procedures. Target sequences for each tetramer are shown in **Supplementary Table 15**.

Gene modification of endogenous *CCR5* and *PITX3*. In order to screen TALEN pairs for NHEJ-mediated gene modification, we cultured K562 cells in RPMI1640 medium (Invitrogen) supplemented with 10% (v/v) FBS, 2 mM L-glutamine, 100 U/ml penicillin and 100 mg/ml streptomycin. Cells (1×10^5 to 2×10^5) were nucleofected with TALEN expression plasmids (400 ng each) using the Amaxa 96-well shuttle system (Amaxa Biosystems/Lonza) according to manufacturer instructions. Cells were collected 3 d post-transfection, and genomic DNA was extracted using the QuickExtract DNA extraction solution (Epicentre Biotechnologies) according to supplier instructions. Cells were obtained from the ATCC and were not tested for mycoplasma contamination. Frequency of gene modification by NHEJ was evaluated by the Surveyor nuclease assay as described previously[12,40] as well as deep sequencing using an Illumina MiSeq.

Off-target analysis. Off-target loci were determined computationally by identifying all genomic sites with up to eight mismatches relative to the intended TALEN binding sites with spacing between TALEN binding sites of 10–24 nt. PCR primers were then designed using Primer3 with the following optimal conditions: amplicon size of 200 nt, a T_m of 60 °C, primer length of 20 nt and G+C content of 50%. Adaptors were added for a second PCR reaction to add the Illumina library sequences (ACACGACGCTCTTCCGATCT forward primer and GACGTGTGCTCTTCCGAT reverse primer). See **Supplementary Table 7** for a list of primers.

Genomic DNA was purified with the Qiagen DNeasy Blood and Tissue Kit. Regions of interest were amplified in 50 μl using 250 ng of genomic DNA with Phusion (NEB) in Buffer GC with 200 μM dNTPs. Amplification of OT20 required the addition of DMSO to a final concentration of 3%. Primers were used at a final concentration of 0.5 μM and the following cycling conditions: initial melt of 98 °C for 30 s, followed by 30 cycles of 98 °C for 10 s, 60 °C for 30 s and 72 °C for 15 s, followed by a final extension at 72 °C for 10 min. PCR products were diluted 1:200 in H_2O. 1 μl diluted PCR product was used in a 10-μl PCR reaction to add the Illumina library sequences with Phusion (NEB) in Buffer GC with 200 μM dNTPs. Primers were used at a final concentration of 0.5 μM and the following conditions: initial melt of 98 °C for 30 s, followed by 12 cycles of 98 °C for 10 s, 60 °C for 30 s and 72 °C for 15 s, followed by a final extension at 72 °C for 10 min. PCR products were pooled and purified using the Qiagen Qiaquick PCR Purification Kit. Samples were quantitated with the Qubit dsDNA HS Assay Kit (Life Technologies). Samples were diluted to 2 nM and sequenced on an Illumina MiSeq Instrument with a 300-cycle sequencing kit.

[39] Li, L. GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *J. Comput. Biol.* **16**, 317–329 (2009).

[40] Perez, E.E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat. Biotechnol.* **26**, 808–816 (2008).

doi:10.1038/nmeth.3330