Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity

John P Guilinger1,2, Vikram Pattanayak1,2, Deepak Reyon3,4, Shengdar Q Tsai3,4, Jeffry D Sander3,4, J Keith Joung3,4 & David R Liu1,2

Although transcription activator–like effector nucleases (TALENs) can be designed to cleave chosen DNA sequences, TALENs have activity against related off-target sequences. To better understand TALEN specificity, we profiled 30 unique TALENs with different target sites, array length and domain sequences for their abilities to cleave any of 1012 potential off-target DNA sequences using in vitro selection and high-throughput sequencing. Computational analysis of the selection results predicted 76 off-target substrates in the human genome, 16 of which were accessible and modified by TALENs in human cells. The results suggest that (i) TALE repeats bind DNA relatively independently; (ii) longer TALENs are more tolerant of mismatches yet are more specific in a genomic context; and (iii) excessive DNA-binding energy can lead to reduced TALEN specificity in cells. Based on these findings, we engineered a TALEN variant that exhibits equal on-target cleavage activity but tenfold lower average off-target activity in human cells.

The ability to engineer site-specific changes in genomes is a powerful capability with important research and therapeutic implications. TALENs are fusions of the FokI restriction endonuclease cleavage domain with a DNA-binding TALE repeat array (Fig. 1a). These arrays consist of multiple 34-amino acid TALE repeats, each of which uses a repeat variable diresidue (RVD), the amino acids at positions 12 and 13, to recognize each of the four DNA nucleotides1,2. Thus, one can construct a TALE repeat to bind virtually any DNA sequence. TALENs can be engineered to be active only as heterodimers using obligate TALE repeat to bind virtually any DNA sequence. TALENs can also be engineered to be active only as heterodimers using obligate TALE repeat to bind virtually any DNA sequence. TALENs can also be engineered to be active only as heterodimers using obligate TALE repeat to bind virtually any DNA sequence. TALENs can be designed to cleave chosen DNA sequences, TALENs have activity against related off-target sequences. To better understand TALEN specificity, we profiled 30 unique TALENs with different target sites, array length and domain sequences for their abilities to cleave any of 1012 potential off-target DNA sequences using in vitro selection and high-throughput sequencing. Computational analysis of the selection results predicted 76 off-target substrates in the human genome, 16 of which were accessible and modified by TALENs in human cells. The results suggest that (i) TALE repeats bind DNA relatively independently; (ii) longer TALENs are more tolerant of mismatches yet are more specific in a genomic context; and (iii) excessive DNA-binding energy can lead to reduced TALEN specificity in cells. Based on these findings, we engineered a TALEN variant that exhibits equal on-target cleavage activity but tenfold lower average off-target activity in human cells.

The ability to engineer site-specific changes in genomes is a powerful capability with important research and therapeutic implications. TALENs are fusions of the FokI restriction endonuclease cleavage domain with a DNA-binding TALE repeat array (Fig. 1a). These arrays consist of multiple 34-amino acid TALE repeats, each of which uses a repeat variable diresidue (RVD), the amino acids at positions 12 and 13, to recognize each of the four DNA nucleotides1,2. Thus, one can construct a TALE repeat to bind virtually any DNA sequence. TALENs can be engineered to be active only as heterodimers using obligate TALE repeat to bind virtually any DNA sequence. TALENs can be designed to cleave chosen DNA sequences, TALENs have activity against related off-target sequences. To better understand TALEN specificity, we profiled 30 unique TALENs with different target sites, array length and domain sequences for their abilities to cleave any of 1012 potential off-target DNA sequences using in vitro selection and high-throughput sequencing. Computational analysis of the selection results predicted 76 off-target substrates in the human genome, 16 of which were accessible and modified by TALENs in human cells. The results suggest that (i) TALE repeats bind DNA relatively independently; (ii) longer TALENs are more tolerant of mismatches yet are more specific in a genomic context; and (iii) excessive DNA-binding energy can lead to reduced TALEN specificity in cells. Based on these findings, we engineered a TALEN variant that exhibits equal on-target cleavage activity but tenfold lower average off-target activity in human cells.

The ability to engineer site-specific changes in genomes is a powerful capability with important research and therapeutic implications. TALENs are fusions of the FokI restriction endonuclease cleavage domain with a DNA-binding TALE repeat array (Fig. 1a). These arrays consist of multiple 34-amino acid TALE repeats, each of which uses a repeat variable diresidue (RVD), the amino acids at positions 12 and 13, to recognize each of the four DNA nucleotides1,2. Thus, one can construct a TALE repeat to bind virtually any DNA sequence. TALENs can be engineered to be active only as heterodimers using obligate TALE repeat to bind virtually any DNA sequence. TALENs can be designed to cleave chosen DNA sequences, TALENs have activity against related off-target sequences. To better understand TALEN specificity, we profiled 30 unique TALENs with different target sites, array length and domain sequences for their abilities to cleave any of 1012 potential off-target DNA sequences using in vitro selection and high-throughput sequencing. Computational analysis of the selection results predicted 76 off-target substrates in the human genome, 16 of which were accessible and modified by TALENs in human cells. The results suggest that (i) TALE repeats bind DNA relatively independently; (ii) longer TALENs are more tolerant of mismatches yet are more specific in a genomic context; and (iii) excessive DNA-binding energy can lead to reduced TALEN specificity in cells. Based on these findings, we engineered a TALEN variant that exhibits equal on-target cleavage activity but tenfold lower average off-target activity in human cells.

The ability to engineer site-specific changes in genomes is a powerful capability with important research and therapeutic implications. TALENs are fusions of the FokI restriction endonuclease cleavage domain with a DNA-binding TALE repeat array (Fig. 1a). These arrays consist of multiple 34-amino acid TALE repeats, each of which uses a repeat variable diresidue (RVD), the amino acids at positions 12 and 13, to recognize each of the four DNA nucleotides1,2. Thus, one can construct a TALE repeat to bind virtually any DNA sequence. TALENs can be engineered to be active only as heterodimers using obligate TALE repeat to bind virtually any DNA sequence. TALENs can be designed to cleave chosen DNA sequences, TALENs have activity against related off-target sequences. To better understand TALEN specificity, we profiled 30 unique TALENs with different target sites, array length and domain sequences for their abilities to cleave any of 1012 potential off-target DNA sequences using in vitro selection and high-throughput sequencing. Computational analysis of the selection results predicted 76 off-target substrates in the human genome, 16 of which were accessible and modified by TALENs in human cells. The results suggest that (i) TALE repeats bind DNA relatively independently; (ii) longer TALENs are more tolerant of mismatches yet are more specific in a genomic context; and (iii) excessive DNA-binding energy can lead to reduced TALEN specificity in cells. Based on these findings, we engineered a TALEN variant that exhibits equal on-target cleavage activity but tenfold lower average off-target activity in human cells.

Although TALENs do not cause widespread genomic off-target modification14–17, cleavage at off-target sites can result in unintended mutations at genomic loci. Whereas recent studies have identified closely related off-target sites containing two or fewer mismatches in zebrafish18 and in human cell lines13, more distantly related off-target sites are of particular interest because one would expect a typical 36-bp target site to be approximately eight or more mutations away from any sequence in the human genome. In previous studies, two distant genomic off-target sites have been verified to be cleaved in human cell lines from 19 potential off-target sites predicted using systematic evolution of ligands by exponential enrichment (SELEX)7, an in vitro method to identify binding sites of DNA-binding domains in isolation. Only three off-target sites have been identified using an integrase-deficient lentiviral vector-based approach19,20 to capture off-target double-strand break sites in cells. The limited number of off-target TALEN sites identified in previous studies suggests that further research is needed both to better understand the extent of TALEN-induced genomic off-target mutations and to improve TALEN specificity to minimize these unwanted effects.

Principles that determine specificities of TALEN proteins remain poorly characterized. Although SELEX experiments and a high-throughput study of TALE activator specificity have described the DNA-binding specificities of monomeric TALE proteins5,7,9 and a single TALE activator21, respectively, the DNA-cleavage specificities of active, dimeric nucleases can differ from the specificities of their component monomeric DNA-binding domains22. For example, zinc-finger nucleases, which are different engineered dimeric nucleases, demonstrate compensation effects between monomers22. Cellular methods to study off-target genomic modification such as whole-genome sequencing or integrase-deficient lentiviral vector–based capture can be complicated by DNA accessibility, which varies from site to site and between cell

1Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, USA. 2Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts, USA. 3Molecular Pathology Unit, Center for Cancer Research, and Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, Massachusetts, USA. 4Department of Pathology, Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to D.R.L. (drliu@fas.harvard.edu).

RECEIVED 25 SEPTEMBER 2013; ACCEPTED 11 JANUARY 2014; PUBLISHED ONLINE 16 FEBRUARY 2014; DOI:10.1038/NMETH.2845

NATURE METHODS | VOL.11 NO.4 | APRIL 2014 | 429
types, or by DNA repair and integration pathways after cleavage that could obscure the determination of intrinsic TALEN specificity. Purely cellular studies are also limited to the stochastic handful of off-target sites in a given genome that are similar to the target sequence and thus cannot be used to evaluate the ability of TALENs to cleave a very large number of off-target sites necessary for a broad and in-depth study of TALEN specificity.

Using a previously described in vitro selection method, we interrogated TALENs for their abilities to each cleave $10^{12}$ potential off-target DNA substrates related to their intended target sequences. The resulting data are, to our knowledge, the first comprehensive profiles of TALEN cleavage specificities. The selection results suggest a model in which excess nonspecific DNA-binding energy gives rise to greater off-target cleavage relative to on-target cleavage. Based on this model, we engineered TALENs with a modified architecture and substantially improved specificities of DNA cleavage in vitro. In human cells, these modified TALENs exhibited 24-fold to more than 120-fold greater specificity for the most readily cleaved off-target sites than currently used TALEN constructs. Our results demonstrate four key findings: (i) TALENs are highly specific for their intended target base pair at 103 of the 104 positions profiled, with specificity increasing near the N-terminal TALEN end of each TALE repeat array (corresponding to the 5’ end of the bound DNA); (ii) longer TALENs are more specific in a genomic context, whereas shorter TALENs have higher specificity per nucleotide; (iii) TALE repeat each bind their respective base pairs relatively independently; and (iv) excess DNA-binding affinity leads to increased TALEN activity against off-target sites and therefore decreased specificity.

RESULTS

Specificity profiling of CCR5- and ATM-targeted TALENs

We profiled the specificities of 30 unique heterodimeric TALEN pairs (hereafter referred to as TALENs) harboring different C-terminal, N-terminal and FokI domain variants, and targeted to half-sites of various lengths. The number of base pairs recognized by each half-site that we list in this paper includes the 5’ thymine (T) recognized by the N-terminal domain. Most of the TALENs we tested were obligate heterodimers with FokI Q105E, I118L in one TALEN and FokI E109K, I157K for the
other (EL/KK); we also used a more active heterodimeric variant with FokI Q105E,I118L,N115D in one TALEN and FokI E109K,I157K,H156R in the other (ELD/KKR), and homodimeric FokI nuclease domains, as specified below.

We designed TALENs as previously reported to target one of three distinct sequences, CCR5A, CCR5B or ATM, in two different human genes, CCR5 and ATM (Supplementary Fig. 1).

We determined the specificity profiles using a previously described in vitro selection method. Briefly, we digested preselection libraries of > 10^12 DNA sequences, each theoretically containing at least 10 copies of all possible DNA sequences with six or fewer mutations relative to the on-target sequence, with 3 nM to 40 nM of an in vitro–translated TALEN (Online Methods, Supplementary Table 1, Supplementary Fig. 2 and Supplementary Results).

Cleaved library members harbored a free 5′ monophosphate that enabled them to be captured by adaptor ligation (Fig. 1b,c). We isolated DNA fragments of length corresponding to 1.5 target sites (an intact target site and a repeated half-site up to the point of TALEN-induced DNA cleavage) by gel purification. High-throughput sequencing and computational analysis of TALEN-treated or control samples that survived this selection revealed the abundance of all TALEN-cleaved sequences as well as the abundance of the corresponding sequences before selection (Supplementary Notes). In the control sample, all members of the preselection library were cleaved by a restriction endonuclease at a constant sequence to enable their capture by adaptor ligation and isolation by gel purification. We calculated the enrichment value for each library member that survived selection by dividing the abundance of its sequence after selection by the expected abundance of its sequence before selection.

For all TALEN variants and under all tested conditions, the DNA that survived the selection contained significantly fewer mean mutations in the targeted half-sites than were present in the preselection libraries (P < 10^−7; Fisher’s exact test; Fig. 2a,b and Supplementary Tables 2 and 3). For all selections, the on-target sequences were enriched 8-fold to 640-fold with a mean enrichment value of 110-fold (Supplementary Table 4). To validate our selection results in vitro, we assayed the ability of the CCR5B TALENs targeting 13-bp left and right half-sites (L13 + R13) to cleave each of 16 diverse off-target substrates (Supplementary Fig. 3). The efficiencies correlated well (r = 0.90) with the observed enrichment values from the selection (Fig. 2e).

To quantify specificities of DNA cleavage at each position in the TALEN target sites for all four possible base pairs, we calculated specificity scores as the differences between preselection and postselection base-pair frequencies, normalized to the maximum possible change of the preselection frequency from complete specificity (defined as 1.0) to complete antispecificity (defined as −1.0). For all TALENs tested, the targeted base pair at every position in both half-sites was preferred, with the sole exception of the base pair closest to the spacer for some ATM TALENs at the right half-site (Fig. 2c,d and Supplementary Figs. 4–9). The 5′ T recognized by the N-terminal domain was highly specified, and the 3′ DNA end (targeted by the C-terminal TALEN end) generally tolerated more mutations than the 5′ DNA end; both of these observations are consistent with previous reports.

All 12 of the positions targeted by the Asn-Asn (NN) RVDs in the ATM and CCR5A TALENs were enriched for guanine, confirming previous reports that the NN RVD specifies guanine.

TALEN off-target cleavage in cells

For TALENs that target 36 base pairs (bp), potential off-target sites in the human genome are expected on average to contain approximately eight or more mutations relative to the on-target site (Supplementary Table 5), more mutations than theoretically are covered in the in vitro selection. Therefore, we used a machine learning–based ‘classifier’ algorithm trained on the tens of thousands of off-target sites revealed by the in vitro selection to identify rare TALEN candidate off-target sites in the human genome (Supplementary Results). Using this algorithm, we identified the 36 best-scoring heterodimeric candidate off-target sites...
Table 1 | Cellular modification induced by TALENs at on-target and predicted off-target genomic sites

| Site   | No TALEN (%) | CCR5A EL/KK | CCR5A ELD/KKR | CCR5A Homo Foki |
|--------|--------------|-------------|--------------|----------------|
| L10    |              |             |              |                |
| L10    | 0.007        | 6.8         | 16           | 18             |
| L13    |              |             |              |                |
| L13    | 0.006        | <0.006      | 0.026        | 0.077          |
| L16    |              |             |              |                |
| L16    | 0.006        | <0.006      | 0.036        | 0.39           |
| OffA-1 |              |             |              |                |
| OffA-1 | 0.006        | <0.006      | 0.025        | <0.006         |
| OffA-16 |             |             |              |                |
| OffA-16 | 0.006        | <0.006      | <0.006       | 0.057          |
| OffA-17 |              |             |              |                |
| OffA-17 | 0.051        | <0.006      | <0.006       | 0.94           |
| OffA-23 |              |             |              |                |
| OffA-23 | 0.018        | <0.006      | 0.029        | 0.23           |
| OffA-35 |              |             |              |                |
| OffA-35 | 0.006        | <0.006      | <0.006       | 0.070          |

For cells treated with either no TALEN or CCR5A TALENs containing heterodimeric EL/KK, heterodimeric ELD/KKR or the homodimeric (Homo) Foki cleavage domain variants, cellular modification rates are shown as the percentage of observed insertions or deletions (indels) consistent with TALEN cleavage relative to the total number of sequences for on-target (On) and predicted off-target sites (Off). ND, no data were collected. Same as above for ATM TALENs. Sample sizes and P values are given in Supplementary Tables 7 and 8.

Figure 3 | In vitro specificity as a function of TALEN length. Enrichment value of on-target (zero mutation) and off-target sequences containing 1–6 mutations for CCR5B TALENs of varying TALE repeat array lengths with EL/KK Foki domains. Each selection was performed once with more than 34,900 sequences analyzed per selection.
Figure 4 | In vitro selection specificity and discrete cleavage efficiencies of TALENs containing canonical or engineered C-terminal domains. (a,b) On-target enrichment values for selections of CCR5A TALENs (a) and ATM TALENs (b) containing indicated domains with EL/KK FokI cleavage domains (Supplementary Table 4a,b). 28-aa indicates 28-aa–truncated variant. Each selection was performed once with more than 4,622 sequences analyzed per selection. (c) CCR5A on-target sequence (OnC) and double-mutant sequences with mutations highlighted in red. (d) ATM on-target sequence (OnA), single-mutant sequences and double-mutant sequences with mutations highlighted in red. (e,f) Discrete in vitro cleavage efficiency of DNA sequences listed in (c) with CCR5A TALENs (e) and of DNA sequences listed in (d) with ATM TALENs (f) containing indicated domains with EL/KK FokI domains. Error bars, s.d. from three biological replicates. Average is shown for C4 from two replicates. See Supplementary Results for P values.

enabling the cleavage of sequences with more mutations, without a corresponding increase in the cleavage of sequences with fewer mutations, because the latter are already nearly completely cleaved. Indeed, the in vitro cleavage efficiencies of discrete DNA sequences for these longer TALENs were independent of the presence of a small number of mutations in the target site (Fig. 4c–f), suggesting there was nearly complete binding and cleavage of sequences containing few mutations. Likewise, higher TALEN concentrations also resulted in decreased enrichment values of sequences with few mutations and increased enrichment values of sequences with many mutations (Supplementary Table 4). These results together support a model in which excessive TALEN binding arising from either long TALE arrays or high TALEN concentrations decreases the observed TALEN DNA-cleavage specificity for each recognized base pair. Despite the fact that TALENs designed to cleave longer target sites are less specific per base pair, this model predicts that such TALENs have higher overall specificity than those that target shorter sites, when considering the number of potential off-target sites in the human genome (Supplementary Fig. 13 and Supplementary Results).

Engineering TALENs with improved specificity

The findings above suggest that TALEN specificity could be improved by reducing non-specific DNA binding energy to only support efficient on-target cleavage. We hypothesized that reducing the cationic charge of the canonical 63-amino-acid (aa) TALE C-terminal domain, which contains ten cationic residues, would decrease nonspecific DNA binding and improve the specificity of TALENs.

We constructed two variants in which we changed three (K788Q, R792Q and R801Q; named Q3) or seven (K777Q, K778Q, K788Q, R789Q, R792Q, R793Q and R801Q; named Q7) cationic arginine or lysine residues in the canonical 63-aa C-terminal domain to glutamine (Fig. 1a). We performed in vitro selections on CCR5A and ATM TALENs containing the canonical C-terminal domain, the engineered Q3 domain and the engineered Q7 domain as well as a previously reported 28-aa–truncated C-terminal domain with a theoretical net charge (−1) identical to that of the Q7 C-terminal domain. The enrichment values of the on-target sequence in the CCR5A and ATM selections increased substantially as the net charge of the C-terminal domain decreased (Fig. 4a,b). For example, the enrichment values of the on-target sequences in ATM selections were 510, 50 and 20 for the Q7, Q3 and canonical 63-aa C-terminal domain variants, respectively. Similarly, substituting one, two or three cationic residues in the TALEN N terminus with glutamine also increased cleavage specificity (Supplementary Table 4, Supplementary Fig. 14 and Supplementary Results). Consistent with the selection results, TALENs containing Q7 C-terminal domains showed about fourfold or greater specificity of DNA cleavage in vitro for 11 of the 16 CCR5A and ATM off-target sites containing one or two mutations (Fig. 4c–f and Supplementary Results).

Improved specificity of engineered TALENs in human cells

To determine whether the increased specificity of the engineered TALENs observed in vitro also occurs in human cells, we measured TALEN-induced modification rates of the on-target and top 36 predicted off-target sites for CCR5A and ATM TALENs containing all six possible combinations of the canonical 63-aa, Q3 or Q7 C-terminal domains and the EL/KK or ELD/KKR FokI domains (12 TALENs total). We did not analyze TALENs containing a 28-aa C-terminal domain in these experiments because both the ATM and CCR5A on-target sites have DNA spacer lengths of 18 bp, which is outside the 28-aa C-terminal domain’s preferred DNA spacer length range (Supplementary Figs. 15 and 16, and Supplementary Results). For both FokI variants, the TALENs with Q3 C-terminal domains demonstrated substantial on-target activities ranging from 8% to 24% modification, comparable to the activity of TALENs with the canonical 63-aa C-terminal domains. TALENs with canonical 63-aa or Q3 C-terminal domains and the ELD/KKR FokI domain were both fivefold to ninefold more active in modifying the CCR5A and ATM on-target site in cells than the corresponding TALENs with the Q7 C-terminal domain (Fig. 5 and Supplementary Table 7).

Compared to the canonical 63-aa C-terminal domains, TALENs with Q3 C-terminal domains demonstrated a mean increase in
on-target/off-target activity ratio of more than 12-fold and more than ninefold for CCR5A and ATM sites, respectively, with the ELD/KKR FokI domain (Fig. 5, and Supplementary Tables 7 and 10a,b). These mean improvements can only be expressed as lower limits owing to the absence or near-absence of observed cleavage events by the engineered TALENs for many off-target sequences. For the ATM TALENs containing Q7 C-terminal domains, the cleavage efficiency of both the on-target and off-target sites was so low that we could not determine their specificity (Fig. 5 and Supplementary Tables 7 and 10a,b). For the most abundantly cleaved off-target site (CCR5A off-target site 5), the Q3 C-terminal domain was 24-fold more specific, and the Q7 C-terminal domain was >120-fold more specific (Fig. 5), than the canonical 63-aa C-terminal domain. To determine whether the increased specificity of the engineered TALENs observed for CCR5A and ATM TALENs applies more generally, we constructed three additional TALENs, targeting sequences in the PMS2, SDHD and HDAC1 genes12, using the canonical 63-aa, Q3 or Q7 C-terminal domains and ELD/KKR FokI domains. Of the 64 TALENs reported previously12, TALENs targeting these three genes had target sequences with closely homologous genomic off-target sites containing one to five mutations. For each of these TALENs, we measured modification rates for genomic on-target and off-target sites. PMS2, SDHD and HDAC1 TALENs with Q3 C-terminal domains demonstrated on-target activities ranging from 6% to 28% modification, comparable to the activity of TALENs with the canonical 63-aa C-terminal domain (Supplementary Table 10c and 11). Although the PMS2, SDHD and HDAC1 TALENs with Q3 C-terminal domains had similar on-target activity levels as canonical TALENs, they demonstrated a fivefold to sevenfold increase in on-target/off-target activity ratio. For the PMS2 TALENs, the Q7 C-terminal domains demonstrated a 53-fold and 64-fold increase in on-target/off-target activity ratio in cells, although as observed above, the Q7 TALENs were less active on the target site than TALENs containing the canonical or Q3 C-terminal domains (Supplementary Tables 10c and 11).

DISCUSSION
The 16 confirmed TALEN off-target sites containing 8–12 mutations identified from the 76 predicted sites assayed in this study represent more bona fide genomic off-target sites in the human genome than have been revealed collectively to date by other methods. These 16 sites were modified with 0.03–2.3% efficiency in human cells, which demonstrated that TALENs can have appreciable off-target activities in human cells even at sites that are eight or more mutations away from the on-target sequence. Site accessibility in cells, influenced by histone proteins, transcription factors and DNA modification23, likely accounts for at least some of the difference between our in vitro, computational and cell-based results. We compared our method with other methods for characterizing TALEN specificity and identifying genomic TALEN off-target sites in Supplementary Discussion.

The observed decrease in specificity for TALENs with more TALE repeats or more cationic residues in the C-terminal domain or N terminus is consistent with a model in which excess TALEN binding affinity leads to increased promiscuity. This excess DNA-binding energy model may explain reports that NN RVDs bind either A or G (refs. 2,27,33). Those studies used TALE arrays of more than 14 RVDs, which may have created a scenario in which excess DNA-binding energy permits a suboptimal NN RDF interaction with A compared to G. We observed NN RVDs can discriminate between A and G, consistent with reports using shorter TALE arrays of 13 RVDs28. Excess DNA-binding energy could also explain the previously reported promiscuity at the 5′ terminal T of TALENs with longer C-terminal domains34 and is consistent with observations of higher TALEN protein concentrations inducing more off-target cleavage28. Although decreasing TALEN protein expression in theory could reduce off-target cleavage, TALE arrays are reported with on-target DNA binding affinities as high as 2.8 nM (dissociation constant, Kd)26, sufficient to theoretically saturate target sites even when expressed at modest, mid-nanomolar concentrations in a cell. The difficulty of improving the specificity of such TALENs by lowering their expression, coupled with the need to maintain sufficient TALEN concentrations to effect desired levels of on-target cleavage, highlight the value of engineering TALENs with higher intrinsic specificity.

Our findings suggest that mutant C-terminal domains with reduced nonspecific DNA binding may be used to alter the DNA-binding affinity of TALENs such that on-target sequences are cleaved efficiently but with minimal excess DNA-binding energy, which results in better discrimination between on-target and off-target sites. As TALENs targeting up to 46 bp have been shown to be active in cells14, it may be possible to further improve specificity by engineering TALENs with a combination of variant N-terminal and C-terminal domains that impart reduced nonspecific DNA binding, a greater number of TALE repeats to contribute additional on-target DNA binding and lower-affinity RVDs such as the NK RVD to recognize G27,28. It is tempting to speculate that the strategy of substituting residues that contribute to nonspecific DNA binding to improve DNA specificity may also apply to other proteins used for genome engineering, including Cas9 and zinc-finger nucleases.

Our findings and the resulting improved TALENs would have been difficult to generate using purely cellular off-target cleavage methods. The ability of our profiling method to reveal the broad, unobscured DNA-cleavage specificity of TALENs in the absence...
of cellular complications enabled the elucidation of the inherent DNA-cleavage specificity of TALENs. The small number of genomic sequences that are closely dimeric to a target sequence also intrinsically limits studies of cellular off-target cleavage. In contrast, we evaluated each active, dimeric TALEN in this study for its ability to cleave any of 10^12 close variants of its on-target sequence, a library size several orders of magnitude greater than the number of different sequences in a mammalian genome. This dense coverage of off-target sequence space enabled the elucidation of detailed relationships between DNA-cleavage specificity and target base pair position, TALEN concentration, mismatch location and TALEN domain composition. These results collectively reveal principles for characterizing and improving TALENs with greater specificity that may enable a wider range of genome-engineering applications.

METHODS

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

Accession codes. Sequence Read Archive: SRP035232.

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

ACKNOWLEDGMENTS

J.P.G., V.P. and D.R.L. were supported by Defense Advanced Research Projects Agency HR0011-11-2-0003 and N00010-12-C-4207, US National Institutes of Health (NIH) GM095501 (D.R.L.), and the Howard Hughes Medical Institute (HHMI). D.R.L. was supported as a HHMI Investigator. V.P. was supported by award T32GM007753 from US National Institute of General Medicine (HHMI). D.R.L. was supported as a HHMI Investigator. V.P. was supported by award T32GM007753 from NIH National Institute of General Medicine (HHMI). D.R.L. was supported as a HHMI Investigator. V.P. was supported by award T32GM007753 from NIH National Institute of General Medicine (HHMI).

AUTHOR CONTRIBUTIONS

J.P.G., V.P., D.R.L. and J.D.S. and S.O.T. performed the experiments, designed the research, analyzed the data and wrote the manuscript. J.K.J. and D.R.L. designed the research, analyzed the data and 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. Doyon, Y. et al. Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. Nat. Methods 8, 74–79 (2011).
4. Cade, L. et al. Highly efficient generation of heritable zebrafish gene mutations using homo- and heterodimeric TALENs. Nucleic Acids Res. 40, 8001–8010 (2012).
5. Miller, J.C. et al. A TALE nuclease architecture for efficient genome editing. Nat. Biotechnol. 29, 143–148 (2011).
6. Bedell, V.M. et al. In vivo genome editing using a high-efficiency TALEN system. Nature 491, 114–118 (2012).
7. Hockemeyer, D. et al. Genetic engineering of human pluripotent cells using TALE nucleases. Nat. Biotechnol. 29, 731–734 (2011).
8. 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).
9. Tesson, L. et al. Knockout rats generated by embryo microinjection of TALENs. Nat. Biotechnol. 29, 695–696 (2011).
10. Moore, F.E. et al. Improved somatic mutagenesis in zebrafish using transcription activation-like effector nucleases (TALENs). PLoS One 7, e37877 (2012).
11. Wood, A.J. et al. Targeted genome editing across species using ZFNs and TALENs. Science 333, 307 (2011).
12. Rayon, D. et al. FLASH assembly of TALENs for high-throughput genome editing. Nat. Biotechnol. 30, 460–465 (2012).
13. 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).
14. Li, T. et al. Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. Nucleic Acids Res. 39, 6315–6325 (2011).
15. Ding, Q. et al. A TALEN genome-editing system for generating human stem cell-based disease models. Cell Stem Cell 12, 238–251 (2013).
16. Lei, Y. et al. Efficient targeted gene disruption in Xenopus embryos using engineered transcription activation-like effector nucleases (TALENs). Proc. Natl. Acad. Sci. USA 109, 17484–17489 (2012).
17. Kim, Y. et al. A library of TAL effector nucleases spanning the human genome. Nat. Biotechnol. 31, 251–258 (2013).
18. Dahlem, T.J. et al. Simple methods for generating and detecting locus-specific modifications induced with TALENs in the zebrafish genome. PLoS Genet. 8, e1002861 (2012).
19. Gabriel, R. et al. An unbiased genome-wide analysis of zinc-finger nuclease specificity. Nat. Biotechnol. 29, 816–823 (2011).
20. Osborn, M.J. et al. TALEN-based gene correction for epidermolysis bullosa. Mol. Ther. 21, 1151–1159 (2013).
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. Pattanayak, V., Ramirez, C.L., Joung, J.K. & Liu, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. Nat. Methods 8, 765–770 (2011).
23. Maeder, M.L. et al. Rapid “open-source” engineering of customized zinc-finger nucleases for highly efficient gene modification. Mol. Cell 31, 294–301 (2008).
24. Pattanayak, V. et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat. Biotechnol. 31, 839–843 (2013).
25. Miller, J.C. et al. An improved zinc-finger nuclease architecture for highly specific genome editing. Nat. Biotechnol. 25, 778–785 (2007).
26. Meckler, J.F. et al. Quantitative analysis of TALE-DNA interactions suggests polarity effects. Nucleic Acids Res. 41, 4118–4128 (2013).
27. 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).
28. 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, 988 (2012).
29. Sander, J.D. et al. In silico abstraction of zinc finger nuclease cleavage profiles reveals an expanded landscape of off-target sites. Nucleic Acids Res. 41, e181 (2013).
30. Kim, Y., Kweon, J. & Kim, J.S. TALENs and ZFNs are associated with different mutation signatures. Nat. Methods 10, 185 (2013).
31. Grau, J., Boch, J. & Posch, S. TALENfoffer: genome-wide TALEN off-target prediction. Bioinformatics 29, 2931–2932 (2013).
32. McNaughton, B.R., Cronican, J.J., Thompson, D.B. & Liu, D.R. Mammalian cell penetration, siRNA transfection, and DNA transfection by supercharged proteins. Proc. Natl. Acad. Sci. USA 106, 6111–6116 (2009).
33. Streubel, J., Blucher, C., Landgraf, A. & Boch, J. TAL effector RVD specificities and efficiencies. Nat. Biotechnol. 30, 593–595 (2012).
34. Sun, N., Liang, J., Abil, Z. & Zhao, H. Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. Mol. Biosyst. 8, 1255–1263 (2012).
ONLINE METHODS

Oligonucleotides, PCR and DNA purification. All oligonu-
cleotides were purchased from Integrated DNA Technolo-
gies (IDT). Oligonucleotide sequences are listed in Supplementary
Notes. PCR was performed with 0.4 µl of 2 U/µl Phusion Hot
Start II DNA polymerase (Thermo-Fisher) in 50 µl with 1×
HF buffer, 0.2 mM dNTP mix (0.2 mM dATP, 0.2 mM dCTP,
0.2 mM dGTP and 0.2 mM dTTP) (NEB), 0.5 µM to 1 µM of
each primer and a program of 98 °C, 1 min; and 35 cycles of
(98 °C, 15 s; 62 °C, 15 s; 72 °C, 1 min) unless otherwise noted.
Many DNA reactions were purified with a QIAquick PCR
Purification Kit (Qiagen), referred to below as ‘Q-column
purification’, or MinElute PCR Purification Kit (Qiagen) referred
to below as ‘M-column purification’.

TALEN construction. The canonical TALEN plasmids were
constructed by the fast ligation-based automatable solid-phase
high-throughput (FLASH) method12 with each TALEN targeting
10–18 bp. Sequences encoding proteins with substitutions in
the N termini were cloned by PCR with Q5 Hot Start Master
Mix (NEB) (98 °C, 22 s; 62 °C, 15 s; 72 °C, 7 min)) using phos-
phorylated TAL-N1fwd (for N1), phosphorylated TAL-N2fwd
(for N2), or phosphorylated TAL-N3fwd (for N3) and phos-
phorylated TAL-Nrev as primers. 1 µl DpnI (NEB) was added, and
the reaction was incubated at 37 °C for 30 min and then subjected
to M-column purification. ~25 ng of eluted DNA was blunt-end
ligated intramolecularly in 10 µl 2× Quick Ligase buffer, 1 µl of
Quick Ligase (NEB) in a total volume of 20 µl at room temperature
(21 °C) for 15 min. 1 µl of this ligation reaction was transformed
to Top10 chemically competent cells (Invitrogen). Sequences
encoding proteins with C-terminal domain substitutions were
cloned by PCR using TAL-Cifwd and TAL-Cirev primers, and
then Q-column–purified. ~1 ng of this eluted DNA was used as the
template for PCR with TAL-Cifwd and either TAL-Q3 (for Q3)
or TAL-Q7 (for Q7) for primers and then Q-column–purified. ~1 ng of this eluted DNA was used as the template for PCR with
TAL-Cifwd and TAL-Cirev primers, and then Q-column–purified. ~1 ng of this eluted DNA was used as the template for PCR with
TAL-Cifwd and TAL-Cirev primers, and then Q-column–purified.

In vitro TALEN expression. TALEN proteins, all containing a
3× Flag tag, were expressed by in vitro transcription–translation.
800 ng of TALEN-encoding plasmid or no plasmid (‘empty lysate’
control) was added to an in vitro transcription–translation reac-
tion using the TNT Quick Coupled Transcription-Translation
System, T7 Variant (Promega) in a final volume of 20 µl at 30 °C
for 1.5 h. Western blots were used to visualize protein using a
1 µl of anti-Flag M2 monoclonal antibody (Sigma-Aldrich, SKU
F3165). TALEN concentrations were calculated by comparison to
standard curve of 1 ng to 16 ng N-terminally Flag-tagged bacterial
alkaline phosphatase (Sigma-Aldrich).

In vitro selection for DNA cleavage. Preselection libraries were prepared with 10 pmol of oligo libraries containing
partially randomized target half-site sequences (CCR5A, ATM or
CCR5B) and fully randomized 10-bp to 24-bp spacer sequences
( Supplementary Notes). Oligonucleotide libraries were sepa-
rately circularized by incubation with 100 units of CircLigase II
ssDNA ligase (Epigenome) in 1× CircLigase II Reaction buffer
(33 mM Tris-acetate, 66 mM potassium acetate and 0.5 mM
dithiothreitol, pH 7.5) supplemented with 2.5 mM MnCl2 in
20 µl total for 16 h at 60 °C then incubated at 80 °C for 10 min.
2.5 µl of each circularization reaction was used as a substrate
for rolling-circle amplification at 30 °C for 16 h in a 50-µl
reaction using the Illustra TempliPhi 100 Amplification Kit
(GE Healthcare). The resulting concatenated libraries were
quantified with Quant-iT PicoGreen dsDNA Kit (Invitrogen),
and libraries with different spacer lengths were combined in an
equilomolar ratio.

For selections on the CCR5B sequence libraries, 500 ng of
preselection library was digested for 2 h at 37 °C in 1× NEBuffer 3
with in vitro transcribed/translated TALEN plus empty lysate
(30 µl total). For all CCR5B TALENs, concentrations of in vitro–
transcribed/translated TALENs were quantified by western blot
(during the blot, TALENs were stored for 16 h at 4 °C) and then
TALEN was added to 40 nM final concentration per monomer. For
selections on CCR5A and ATM sequence libraries, the combined
preselection library was further purified in a 300,000 molecular
weight cutoff spin column (Sartorius) with three 500-µl washes
in 1× NEBuffer 3. 125 ng of preselection library was digested for
30 min at 37 °C in 1× NEBuffer 3 with a total 24 µl of fresh in vitro–
transcribed/translated TALENs and empty lysate. For all CCR5A
and ATM TALENs, 6 µl of in vitro transcription/translation
left TALEN and 6 µl of right TALEN were used, corresponding
to a final concentration in a cleavage reaction ranging from
14 nM to 18 nM for CCR5A TALENs or from 10.5 nM to 13.5 nM
for ATM TALENs. These TALEN concentrations were quantified
by western blot performed in parallel with digestion.

For all selections, the TALEN-digested library was incubated with 1 µl of 100 µg/µl RNase A (Qiagen) for 2 min and then
Q-column–purified. 50 µl of purified DNA was incubated with
3 µl of 10 mM dNTP mix (10 mM dATP, 10 mM dCTP, 10 mM
dGTP and 10 mM dTTP) (NEB), 6 µl of 10× NEBuffer 2 and
1 µl of 5 U/µl Klenow Fragment DNA polymerase (NEB) for
30 min at room temperature and Q-column–purified. 50 µl of
the eluted DNA was ligated with 2 pmol of heated and cooled
#1 adaptors containing barcodes corresponding to each sample
(selections with different TALEN concentrations or constructs;
Supplementary Notes). Ligation was performed in 1× T4 DNA
ligase buffer (50 mM Tris-HCl, 10 mM MgCl2, 1 mM ATP and
10 mM DTT, pH 7.5) with 1 µl of 400 U/µl T4 DNA ligase (NEB)
in 60 µl total volume for 16 h at room temperature, and then
Q-column–purified.

6 µl of the eluted DNA was amplified by PCR in 150 µl total
reaction volume (divided into three 50-µl reactions) for 14 to 22
cycles using the #2A adaptor primers (Supplementary Notes).
The PCR products were purified by Q column. Each DNA sample
was quantified with Quant-iT PicoGreen dsDNA Kit (Invitrogen)
and then pooled into an equimolar mixture. 500 ng of pooled
DNA was run on a 5% TBE 18-well Criterion PAGE gel (Bio-Rad)
for 30 min at 200 V and DNAs of length ~230 bp (corresponding to
1.5 target site repeats plus adaptor sequences) were isolated and
purified by Q column. ~2 ng of eluted DNA was amplified by PCR

NATURE METHODS
doi:10.1038/nmeth.2845
for 5–8 cycles with #2B adaptor primers (Supplementary Notes) and purified by M column.

10 µl of eluted DNA was purified using 12 µl of AMPure XP beads (Agencourt) and quantified with an Illumina/Universal Library Quantification Kit (Kapa Biosystems). DNA was prepared for high-throughput DNA sequencing according to Illumina instructions and sequenced using a MiSeq DNA Sequencer (Illumina) using a 12 pm final solution and 156-bp paired-end reads. To prepare the preselection library for sequencing, the preselection library was digested with 1–4 µl of appropriate restriction enzyme (CCR5A, CCR5B, Tsp45I; ATM, Acc65I; CCR5B, AvalI (NEB)) for 1 h at 37 °C then ligated as described above with 2 pmol of heated and cooled #1 library adaptors. Preselection library DNA was prepared as described above using #2A library adaptor primers and #2B library adaptor primers in place of #2A adaptor primers and #2B adaptor primers, respectively (Supplementary Notes). The resulting preselection library DNA was sequenced together with the TALEN-digested samples.

Discrete in vitro TALEN cleavage assays. Discrete DNA substrates for TALEN digestion were constructed by combining pairs of oligonucleotides as specified in Supplementary Notes with restriction cloning into pUC19 (NEB). Corresponding cloned plasmids were amplified by PCR (59 °C annealing for 15 s) for 24 cycles with pUC19Ofwd and pUC19Orev primers (Supplementary Notes) and Q-column–purified. 50 ng of amplified DNAs were digested in 1× NEBuffer 3 with 3 µl each of in vitro–transcribed–translated TALEN left and right monomers (corresponding to a ~16 nM to ~12 nM final TALEN concentration), and 6 µl of 'empty lysate' in a total reaction volume of 120 µl. The digestion reaction was incubated for 30 min at 37 °C, then incubated with 1 µl of 100 µg/µl RNase A (Qiagen) for 2 min and purified by M column. The entire 10 µl of eluted DNA with glyc erol added to 15% was analyzed on a 5% TBE 18-well Criterion PAGE gel (Bio-Rad) for 45 min at 200 V, then stained with 1× SYBR Gold (Invitrogen) for 10 min. Bands were visualized and quantified on an AlphaImager HP (Alpha Innotech).

Cellular TALEN cleavage assays. TALENs were cloned into mammalian expression vectors and the resulting TALEN vectors transfected into U2OS-EGFP cells, a clonal U2OS human cell line with an integrated construct that constitutively expresses an EGFP-PEST fusion protein, as previously described. Genomic DNA was isolated after 2 d as previously described. For each assay, 50 ng of isolated genomic DNA was amplified by PCR (98 °C, 15 s 67.5 °C, 15 s; 72 °C, 22s) for 35 cycles with pairs of primers with or without 4% DMSO as specified in Supplementary Notes. Two PCR reactions were performed for OffC-5 to improve the limit of detection. The relative dsDNA content of the PCR reaction for each genomic site was quantified with Quant-iT PicoGreen dsDNA Kit (Invitrogen) and then pooled into an equimolar mixture, keeping no-TALEN and all TALEN-treated samples separate. DNA corresponding to 150–350 bp was purified by PAGE as described above.

44 µl of eluted DNA was incubated with 5 µl of 1× T4 DNA ligase buffer and 1 µl of 10 U/µl polynucleotide kinase (NEB) for 30 min at 37 °C and Q-column–purified. 43 µl of eluted DNA was incubated with 1 µl of 10 mM dATP (NEB), 5 µl of 10× NEBuffer 2, and 1 µl of 5 U/µl DNA Klenow fragment (3′→ 5′ exo−; NEB) for 30 min at 37 °C and purified by M column. 10 µl of eluted DNA was ligated as above with 10 pmol of heated and cooled genomic (G) adaptors (Supplementary Notes), and purified by Q column. 8 µl of eluted DNA was amplified by PCR for 6–8 cycles with 'G-B' primers containing barcodes corresponding to each sample. Each sample DNA was quantified with Quant-iT PicoGreen dsDNA Kit (Invitrogen) and then pooled into an equimolar mixture. The combined DNA was subjected to high-throughput sequencing using MiSeq as described above.

Data analysis. Illumina sequencing reads were filtered and parsed with scripts written in Unix Bash as outlined in Supplementary Notes. DNA sequences are available upon request. Source code is available as Supplementary Software. Specificity scores were calculated as previously described. Sample sizes for sequencing experiments were maximized (within practical experimental considerations) to ensure greatest power to detect effects. Statistical analysis of the distribution of number of mutations in various TALEN selections in Supplementary Table 2 was performed as previously described. Statistical analysis of TALEN modified genomic sites in Supplementary Tables 7, 8 and 10 was performed as previously described with multiple comparison correction using the Benjamini–Hochberg method.

To determine extrapolated mean enrichment curves mutation enrichment value as function of mutation number were fit to an exponential function, a × e^b, with R^2 reported using the nonlinear least-squares method. The a, b and R^2 values and the mutation range for these fits are reported in Supplementary Tables. The exponential decrease, b, was used to extrapolate all mean enrichment values beyond five mutations to determine the extrapolated mean enrichment.