5390–5402 *Nucleic Acids Research, 2014, Vol. 42, No. 8* *Published online 24 February 2014*
doi:10.1093/nar/gku155

Comprehensive analysis of the specificity of transcription activator-like effector nucleases

Alexandre Juillerat[1,*], Gwendoline Dubois[1], Julien Valton[1], Séverine Thomas[1], Stefano Stella[2], Alan Maréchal[1], Stéphanie Langevin[1], Nassima Benomari[1], Claudia Bertonati[1], George H. Silva[1], Fayza Daboussi[1], Jean-Charles Epinat[1], Guillermo Montoya[2,3], Aymeric Duclert[1] and Philippe Duchateau[1,*]

[1]Cellectis S.A., 8 Rue de la Croix Jarry, 75013 Paris, France, [2]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre, Macromolecular Crystallography Group, c/Melchor Fdez. Almagro 3, 28029 Madrid, Spain and [3]Structural Biology Group, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen, Denmark

Received October 18, 2013; Revised and Accepted February 5, 2014

ABSTRACT

A key issue when designing and using DNA-targeting nucleases is specificity. Ideally, an optimal DNA-targeting tool has only one recognition site within a genomic sequence. In practice, however, almost all designer nucleases available today can accommodate one to several mutations within their target site. The ability to predict the specificity of targeting is thus highly desirable. Here, we describe the first comprehensive experimental study focused on the specificity of the four commonly used repeat variable diresidues (RVDs; NI:A, HD:C, NN:G and NG:T) incorporated in transcription activator-like effector nucleases (TALEN). The analysis of >15 500 unique TALEN/DNA cleavage profiles allowed us to monitor the specificity gradient of the RVDs along a TALEN/DNA binding array and to present a specificity scoring matrix for RVD/nucleotide association. Furthermore, we report that TALEN can only accommodate a relatively small number of position-dependent mismatches while maintaining a detectable activity at endogenous loci *in vivo*, demonstrating the high specificity of these molecular tools. We thus envision that the results we provide will allow for more deliberate choices of DNA binding arrays and/or DNA targets, extending our engineering capabilities.

INTRODUCTION

The DNA-binding domain derived from transcription activator-like effectors (TALE) has emerged in the past few years as a scaffold of choice to develop tailor-made DNA-binding fusion proteins (1–3). The sequence specificity of this family of proteins, involved in the natural infection process of the plant pathogens of the *Xanthomonas* genus, is driven by a domain composed of repeated motifs of 33–35 amino acids. The specificity results from two polymorphic amino acids, the so-called repeat variable diresidues (RVDs) (4,5), located at positions 12 and 13 of a repeated unit. The recent achievement of the high-resolution structure of TALEs bound to DNA confirmed that each single base of the same strand of the DNA target is contacted by a single repeated unit in a 5′–3′ direction (in line with the protein N-terminal to C-terminal). These structural studies also pointed out that the amino acid at position 13 contacts, in the major groove, the top DNA strand base, whereas the amino acid at position 12 participates in the stabilization of the repeated units (4,5). In addition to the central core mediating the sequence-specific DNA interaction, natural TALEs are composed of two additional domains. The N-terminal translocation domain is responsible for the preferential requirement of a first thymine base (the so-called T_0) in the targeted sequence, and the C-terminal domain contains nuclear localization signals (NLS) and a transcriptional activation domain.

By analyzing sequences of known TALEs and corresponding DNA targets in rice promoters, two groups identified a code governing the preferential pairing of

*To whom correspondence should be addressed. Tel: +33 1 81 69 16 90; Fax: +33 1 81 69 16 90; Email: alexandre.juillerat@cellectis.com
Correspondence may also be addressed to Philippe Duchateau. Tel: +33 1 81 69 16 90; Fax: +33 1 81 69 16 90;
Email: philippe.duchateau@cellectis.com

RVDs with DNA bases (6,7). With this straightforward one-to-one RVD/nucleotide association code (NI:A, HD:C, NN:G and NG:T), TALE DNA-binding domains have rapidly become a promising platform for the creation of molecular tools such as nucleases (TALEN) (8–11), recombinases (12), transcription activators (13–17) and repressors (18,19). Despite the increasing number of publications reporting array assembly methods and successes in targeting the desired endogenous DNA sequences, little is known on the intrinsic RVD specificity along TALE DNA binding arrays (18,20–22).

Defining precisely TALEN specificity and being able to predict potential off-site targeting is of crucial importance to fully assess the potential of this technology, notably for therapeutic applications. In this study, we report an extensive analysis of the degeneracy of the RVD/nucleotide associations in the context of TALEN. We focus on the four commonly used conventional RVDs (NI, HD, NN and NG) incorporated in >350 TALE DNA binding arrays. We developed two model systems, in yeast (extrachromosomal, high throughput) and mammalian cells (intrachromosomal, medium throughput), that allow comprehensive studies of specificity and activity of TALEN at either the levels of single RVD/nucleotide association or the complete array. The analysis of the nuclease profiles resulting from >15 500 TALEN/DNA pairs in yeast allowed us to define an experimental model of the specificity of RVD/nucleotide associations, further validated to score the outcome of TALEN/targets mismatches in mammalian cells. In addition, the minimum number of mismatches required within a TALEN/target pair to significantly abolish activity at endogenous loci *in vivo* was determined. Consequently, our results set the stage for a more rational design of TALEN and contribute to a better understanding of the impact of the number, positions and types of mismatches within a TALEN/DNA binding array and DNA target.

MATERIALS AND METHODS

TALE arrays

All TALE or TALEN arrays were obtained from Cellectis Bioresearch (Paris, France). TALEN™ is a trademark owned by Cellectis Bioresearch. Sequences of TALEN backbones, TALEN RVD array composition and/or relevant targets are presented in Supplementary Tables S1, S2 and S5–S8.

Extrachromosomal SSA assay in yeast

Mutants (TALEN-containing yeast strain) were gridded at high gridding density (~20 spots/cm^2) on nylon filters placed on solid agar-containing YP-glycerol plates, using a colony gridder (QpixII, Genetix). A second layer, consisting of reporter-harboring yeast strains, was gridded on the same filter for each target. Membranes were incubated overnight at 30°C to allow mating. To select diploids, filters were then placed and incubated for 2 days at 30°C on a medium lacking leucine (for the mutant) and tryptophan (for the target) with glucose (2%) as the carbon source. To induce the expression of the TALEN, filters were transferred on YP-galactose-rich medium for 24–48 h at 30 or 37°C. To monitor TALEN activity, through the β-galactosidase activity, filters were finally placed on solid agarose medium containing 0.02% X-Gal in 0.5 M sodium phosphate buffer, pH 7.0, 0.1% SDS, 6% dimethyl formamide, 7 mM β-mercaptoethanol and 1% agarose and incubated at 37°C for up to 48 h.

Filters were scanned and each spot was quantified using the median values of the pixels constituting the spot. We attribute the arbitrary values 0 and 1 to white and dark pixels, respectively. β-Galactosidase activity is directly associated with the efficiency of homologous recombination, thus with the cleavage efficiency of the TALEN. Any value >0 is considered as the consequence of cleavage. For all our large-scale analyses, we considered a robust nuclease activity when above a threshold (t) of activity equal to 0.45. This value corresponds to the mean values of negative controls (m) plus three times the standard deviation (s) ($t = m + 3s$).

Endogenous green fluorescent protein activity assay

CHO-K1 (CGPS-CHOK1, Cellectis Bioresearch) cells containing the chromosomally integrated green fluorescent protein (GFP) reporter gene including the TALEN recognition sequence (TGAACCGCATCGAGCTG aagggcatcgacTTCAAGGAGGACGGCAA) were cultured at 37°C with 5% CO_2 in a F12-K complete medium supplemented with 2 mM l-glutamine, penicillin (100 IU/ml), streptomycin (100 µg/ml), amphotericin B (Fongizone: 0.25 µg/ml, Life Technologies) and 10% fetal bovine serum (FBS). Cell transfection was performed according to the manufacturer's instructions using the Nucleofector apparatus (Amaxa, Lonza). Adherent CHO-K1 cells were harvested at Day 1 of culture, washed with phosphate-buffered saline (PBS), trypsinized and resuspended in T nucleofection solution to a concentration of 1×10^6 cells/100 µl. Subsequently, 5 µg of each of the two TALEN expression vector pairs (10 µg final DNA amount) was mixed with 0.1 ml of the CHO-K1 cell suspension, transferred to a 2.0-mm electroporation cuvette and nucleofected using program U_023 of the Amaxa Nucleofector apparatus. Maximum 20 min after nucleofection, 0.5 ml of prewarmed F12-K medium was added to the electroporation cuvette. Cells were then transferred to a Petri dish containing 10 ml F12-K medium and cultured at 37°C with 5% CO_2, as previously described. On Day 3 post-transfection, cells were washed with PBS, trypsinized, resuspended in 5 ml and the percentage of GFP-negative cells was monitored by flow cytometry (Guava EasyCyte, Merck Millipore).

Extrachromosomal assay in CHO-K1 cells

Activity in CHO-K1 cells was measured as previously reported by Valton *et al.* (23). In brief, cells were transfected (polyfect, Qiagen) with the two TALEN expression vectors and the reporter plasmid. Three days post-transfection, β-galactosidase was quantified at 420 nm using ONPG in a liquid assay. The entire process was performed using a 96-well plate format on an automated Velocity11 BioCel platform.

Endogenous targeted mutagenesis

Moreover, 293H cells were cultured at 37°C with 5% CO_2 in Dulbecco's modified Eagle's medium (DMEM) complete medium supplemented with 2 mM l-glutamine, penicillin (100 IU/ml), streptomycin (100 µg/ml), amphotericin B (Fongizone: 0.25 µg/ml, Life Technologies) and 10% FBS. Adherent 293H cells were seeded at 1.2×10^6 cells in 10 cm Petri dishes a day before transfection. Cell transfection was performed using the Lipofectamine 2000 reagent according to the manufacturer's instructions (Invitrogen). Furthermore, 2.5 µg of each of the two TALEN nuclease expression plasmids and 10 ng of GFP expression vector (5 µg final DNA amount) were mixed with 0.3 ml of DMEM without FBS. In another tube, 25 µl of Lipofectamine were mixed with 0.3 ml of DMEM without FBS. After 5 min incubation, both DNA and Lipofectamine mixes were combined and incubated for 25 min at RT. The mixture was transferred to a Petri dish containing the 293H cells in 9 ml of complete medium and then cultured at 37°C with 5% CO_2. Three days post-transfection, the cells were washed with PBS, trypsinized, resuspended in 5 ml complete medium and the percentage of GFP-positive cells was measured by flow cytometry (Guava EasyCyte) to monitor transfection efficacy. Cells were pelleted by centrifugation and genomic DNA was extracted using DNeasy Blood & Tissue Kit (Qiagen), according to the manufacturer's instructions. Polymerase chain reaction of the endogenous loci was performed using the oligonucleotide sequences presented in Supplementary Table S6 and purified using the AMPure kit (Invitrogen). Amplicons were further analyzed by deep sequencing using the 454 system (Roche).

TALE protein expression and purification

The TALE IL2RG was purified as previously described for TALE AvrBs3 (24). The clarified lysate of the cells overexpressing TALE IL2RG was loaded onto a Ni-NTA (GE-Healthcare) column and eluted with a linear gradient to 500 mM imidazole. Fractions containing the protein were loaded onto a heparin column and eluted by a linear salt gradient to 1 M NaCl. The protein was then loaded onto a Superdex 200 (GE-Healthcare) gel filtration column.

Fluorescence anisotropy

The dissociation constants between the TALE IL2RG protein and dsDNAs were estimated from the change in fluorescent polarization of complexes between protein and 6-FAM-labeled dsDNAs. The 20-bp DNAs for this assay were annealed by slow cooling in 25 mM Hepes (pH 8.0) and 150 mM NaCl at a final duplex concentration of 500 nM

The optimal concentration of 6-FAM-DNAs for the assay was empirically determined by measuring the fluorescence polarization of serially diluted 6-FAM-DNAs samples (24). The concentration of the 6-FAM DNAs ranged between 20 and 40 nM and that of the TALE IL2RG protein ranged between 0 and 1000 nM. Both proteins and dsDNAs were dialyzed in buffer containing 25 mM Hepes (pH 8), 150 mM NaCl and 0.2 mM TCEP. After incubation at 25°C for 10 min, the fluorescence polarization was measured in a black 96-well assay plate with Wallac Victor2V 1420 multilabel counter (PerkinElmer). The fitting of the data and the K_d calculations were done as previously described in (24).

Data analysis

Context dependence was analyzed taking into account only TALEN activity on their cognate targets. For each nucleotide pair (N_i, N_j), we studied how the presence of N_j just on the right (or on the left) of N_i was influencing activity. For all targets containing the subsequence N_iN_j (or N_jN_i), we computed the ratio $R(N_i,N_j) = A(N_i,N_j)/<A(N_k,N_j)>$ where $A(N_i,N_j)$ is the activity on the target and $<A(N_k,N_j)>$ is the average activity on all targets where N_i is replaced by any of the four nucleotides. If there is complete context independence, $R(N_i,N_j)$ should not depend on N_j, and for a given N_i, all values for the four N_j nucleotides should be the same. The average value and standard deviation of $R(N_i,N_j)$, N_i is indicated on the axis, and N_j represented by various colors is shown in (Supplementary Figure S7A and B).

To compute specificity matrices, we took all the activities of TALEN on targets T_m differing from one mutation compared with their cognate target T, and computed the drop of activity $R = A(T_m)/A(T)$, where $A(T_m)$ and $A(T)$ are the activities on mutated and cognate targets, respectively. By definition, R is equal to 1 for the cognate nucleotide corresponding to the code. The specificity matrix was computed by calculating the average of these R values for all the available pairs for each given position and represented by gray levels (black = 1, white = 0). An overall specificity matrix was computed from positions 1 to 7 by averaging them all.

We predicted the value on mutated targets using the overall specificity matrix by taking the value V_0 of the TALEN on its cognate targets and multiplying it by the product of the square roots of the matrix coefficients $C(RVD_i,N_i)$ corresponding to all RVD/nucleotides pairs, giving a final value $V = V_0 \times \prod_i \text{sqrt}\,[C(RVD_i,N_i)]$. The square root was taken because the specificity matrix was obtained from symmetrically mutated targets, with one mismatch on one side corresponding, in fact, to two mutations on the target.

To determine the probability to find off-site targets, we randomly drew 15 000 sequences of 200 bp each from the human genome and randomly selected a potential TALEN site available on each. For each potential TALEN site, off-site targets were determined as all sequences having binding site pairs diverging from those of the TALEN site from three mismatches or less and having a spacer length ranging from 9 to 30 bp. All combinations (left + right, left + left, right + right) were taken into account.

RESULTS

Experimental setup for the mammalian activity screening

To investigate the specificity of the NI:A, HD:C, NN:G and NG:T RVD/nucleotide pairing, we first sought to

design an experimental setup that will focus on the enzymatic activity of the nuclease while minimizing or normalizing variations due to interfering parameters such as chromatin accessibility and epigenetic modification. Toward this goal, we developed and used a model CHO-K1 cell line containing a unique integrated GFP reporter gene. Small deletions or insertions (indels) produced by the non-homologous end-joining (NHEJ) repair pathway at the double-strand break (DSB) site (generated by the TALEN) led to the gene knockout, and thus to a GFP-negative phenotype. Therefore, we monitored the extinction of GFP, due to the activity of a TALEN on a unique specific sequence within this reporter gene (Figure 1A).

Throughout this study, we used two different TALEN scaffolds, namely, a +C40 and a +C11-SGSGSGG. In our hands, these two scaffolds presented the best balance between the possibilities to obtain a very high activity while keeping a narrow spacer window to improve the specificity. We first design a highly active TALEN (further referenced as wt) within the GFP gene (up to 70% of GFP disruption), where the RVD/nucleotide association fitted the described NI:A, HD:C, NN:G and NG:T code (6,7). Along with this TALEN, we created five collections of additional TALEN that contained, in one of the two TAL monomers, alternative (further referred as mismatches) RVD/nucleotide pairings. These artificial mismatches were located at positions 1-2-3 (Collection 1), 8–9 (Collection 2), 10–11 (Collection 3), 12-13-14 (Collection 4) and 14–15 (Collection 5), as defined from the first thymine base (T_0) (Figure 1B and Supplementary Table S1). These 110 TALEN were thus tested in our CHO-K1 model cell line, and the percentage of GFP-negative cells was recorded. We first analyzed the global effect of mismatch numbers on the activity, taking into account all datasets and found a decent correlation, although this correlation could mask particular positioning effects ($r = -0.68$, $P = 4e{-}16$, Figure 1C). We thus decided to further decipher the RVD/nucleotide pairing specificity by monitoring the effect of increasing number of mismatches (1, 2 or 3) as a function of their position in the array (sliding of the experimental window from the N- to the C-terminus). The analysis of the data pointed out that while the presence of a single mismatch had a limited and similar impact regardless its position in the array ($P = 0.08$), multiple mismatches present a more pronounced effect when positioned at the N-terminal end rather than at the C-terminal end ($P = 4e{-}4$ for two mismatches and $P = 2e{-}4$ for three mismatches, Figure 1D, E and F). A near-complete loss of activity was observed when three consecutive mismatches were positioned close to the N-terminus, whereas a significant activity (up to ~40% of the wt TALEN) was detected for mismatches close the C-terminal end (Figure 1F).

Nevertheless, variations in activities were clearly dependent on multiple parameters, such as the total number of mismatches, their position in the array or their identity (RVD/nucleotide association). This latter parameter was only partially taken into account in our experimental setup, as the targeted sequence remained constant for all TALEN. We thus envisioned performing a comprehensive analysis of the RVD/nucleotide pairing focused especially on the N-terminal of the array, as this end showed the higher specificity. For such large-scale analyses requiring the screening of collections of mutants versus collections of targets, our current mammalian cell model system turned out inadequate. To address this technical limitation, we used a plasmid-based assay in yeast cells to perform high-throughput nuclease activity screenings (25).

To validate a single-strand annealing (SSA) assay as readout, we performed a comparative nuclease activity study of a subset of 13 TALEN from the original GFP dataset (Collections 1–5). We found a particularly good correlation between results from an extrachromosomal SSA assay in CHO-K1 and our previous chromosomal disruption experiments ($r = 0.88$ with $P = 3.8e{-}05$, Supplementary Figure S1). In addition, taking into account that we have previously reported a very good correlation between the yeast and CHO-K1 extrachromosomal SSA assays (26), we anticipated that the yeast model system could serve as an appropriate and representative high-throughput assay to study TALEN activity and specificity.

Experimental setup of the yeast high-throughput nuclease activity screening

The yeast nuclease activity assay, a yeast strain expressing the nuclease of interest is crossed with another strain harboring a reporter plasmid containing the target sequence. This target sequence is flanked by overlapping truncated LacZ genes. On target cleavage, the restoration of the LacZ marker through the SSA pathway of recombination restores a functional LacZ gene, which can be quantified and related to the nuclease efficiency. In addition, to minimize bias, amplify potential effects and simplify subsequent analysis, we performed initial experiments using particular homodimeric TALEN architecture. In this architecture, the targeted sequence is composed of two duplicated sequences in inverse orientation facing each other (separated by the so-called sequence spacer) on both DNA strands. This setup implies, *inter alia*, that a specific mismatch (between an RVD and a nucleotide of the target) on one-half TALEN arm will be found symmetrically in the other half TALEN arm. In addition, as most naturally occurring TALE, including our AvrBs3 scaffold, bind to targets starting with a T (corresponding to the so-called T_0), we kept this feature in all our designs of experiments.

As a first experiment, we determined an optimal TALEN repeat-array length to assure an adequate dynamic range for activity and sensitivity in our yeast screening assay (25). In this experiment, 52 TALEN, containing 9.5–15.5 repeats, were tested on their cognate targets (Figure 2A and Supplementary Table S2). As previously observed in other studies (11), no significant correlation between the size of the array and the nuclease activity could be determined ($r = 0.27$, $P = 0.05$). Second, based on our previous results in CHO-KI showing a decrease of specificity from the N- to the C-terminal end of the array, we monitored the effect of

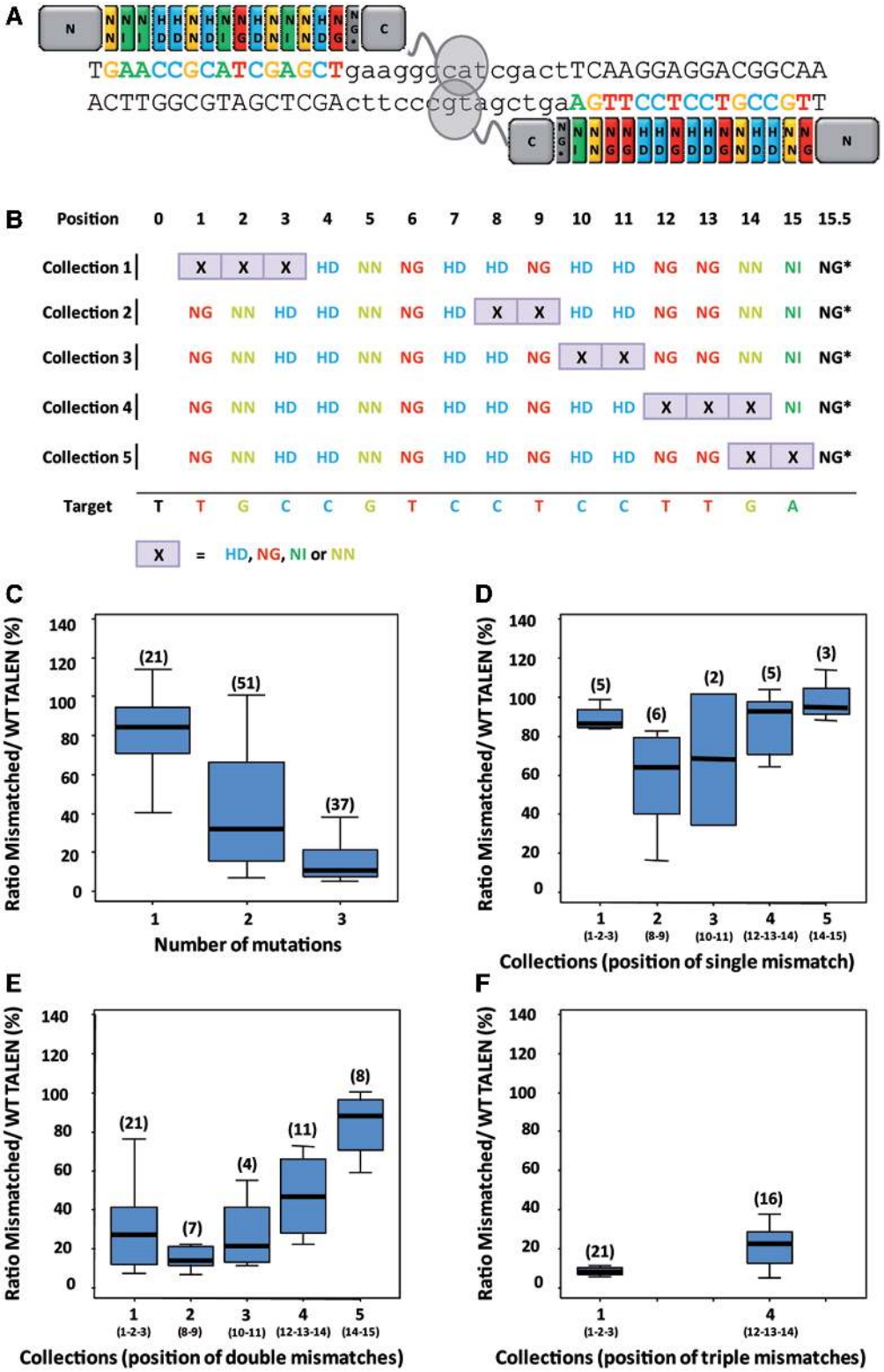

Figure 1. Gene disruption-relative activities of collection of TALEN at the integrated GFP locus. (**A**) Schematic representation of the WT GFP TALEN on the chromosomal target. (**B**) Collections of TALE used at the chromosomal GFP locus derived by mutation of the right DNA binding domain. X represents any of the four, namely, NI, HD, NN and NG RVD. Positions are numbered relative to the first thymine of the target (T_0). (**C**) Influence of the number of mismatches on the GFP disruption. Activity ratio between the mismatched and the WT TALEN is represented on a boxplot, indicating the median (thick bar), quartiles (box) and extreme values ($r = -0.68$, $P = 4e{-}16$). Mismatches are defined relative to the NI:A, HD:C, NN:G and NG:T codes. The size of the sample is indicated in brackets. (**D**) Boxplot representation, including the median (thick bar), quartiles (box) and extreme values, of the activity ratio between the mismatched and the WT TALEN in function of the collections for one mismatch. $P = 0.08$ (Kruskal–Wallis test). (**E**) Same as for (D) but for two mismatches. $P = 4e{-}4$. (**F**) Same as in (D) but for three mismatches. $P = 2e{-}4$.

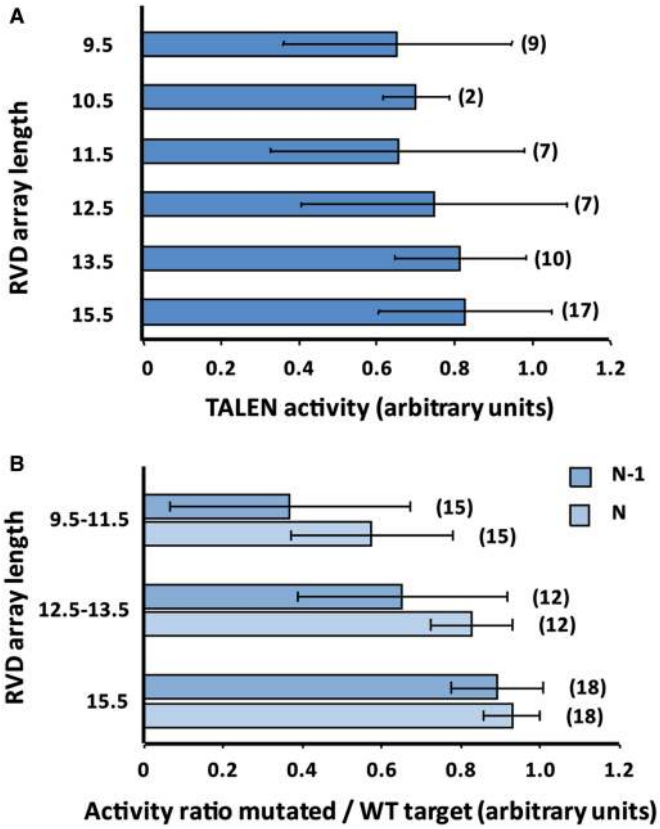

Figure 2. Design of experimental setup and optimal TALEN array length for the nuclease activity screening. (**A**) Activities of 52 TALEN pairs measured *in vivo,* in function of the repeat array length. Error bars represent standard errors. (**B**) Activities of 15 TALEN pairs on targets randomized in their positions N and N-1. Error bars represent standard errors. An analysis of variance demonstrates a significant effect of the TALEN length on specificity, position N: $r = 0.68$, $P = 4e{-}6$ (Kruskal–Wallis test); position N-1: $r = 0.7$, $P = 7e{-}6$). The yeast activity assay is based on the single-strand annealing (SSA) pathway used after the creation of a DSB by the TALEN in the target sequence. Target sites were designed to allow TALEN use in the homodimer format. All TALEN pairs showed a significant activity. The number of TALEN/target pairs for each class is indicated in brackets.

mismatches on the last two positions (called N and N-1) of the TALEN array as a function of the array length. A subset (14 TALEN) of the previous TALEN collection was screened for activity on targets containing the four possible nucleotides at positions N and N-1. When the activity toward the mismatched targets was averaged and compared with the activity on the wild-type target, a statistically significant correlation between the TAL array length and the loss of activity due to mismatches at these positions was obtained (position N: $r = 0.68$, $P = 4e{-}6$; position N-1: $r = 0.7$, $P = 7e{-}6$, Figure 2B). Based on the previous results and data from the literature, we further focused on the shorter 9.5-repeats model to perform extensive investigation of the specificity of RVDs (from positions 1 to 7), as in this configuration all RVD/nucleotide pairings are essential for activity.

Therefore, we performed a comprehensive study of TALEN activity and specificity by systematically changing DNA-binding modules to create collections of TALE arrays (9.5 repeats) containing, at three defined consecutive positions, all 64 possible RVD triplets (combinations of HD, NG, NI and NN). These triplets were located either at positions 1-2-3 (Collection 6), 3-4-5 (Collection 7) or 5-6-7 (Collection 8) of the RVD array, as defined from the first thymine base (T_0) (Figure 3A) (6,7). To cross-validate our results with respect to differences in global affinity, an additional collection located at position 1-2-3 (Collection 9) of a longer array (18.5 RVDs) was also created. These TALEN collections were assayed against their respective 64 targets, containing all possible 4-base triplets at the adequate positions. Altogether, up to 4096 TALEN/target combinations per collection were assessed for nuclease activity (Figure 3B and Supplementary Figures S2–S4).

Analysis of global nuclease activity as a function of RVD identity

Data gathered from Collection 6 (focusing on position 1-2-3) clearly highlight that the presence of an adenine (A) or a cytosine (C) base at position 1 of the target tends to present a deleterious effect on the TALEN activity (Supplementary Figure S5A). A similar observation was noted for an adenine at position 2, as previously hypothesized in another study (27). In this analysis, the error bars were relatively high because for each observed position we aggregated values derived from targets having all possible combinations on the remaining two positions. Thereby, we further analyzed the results taking into account fixed combinations of nucleotides/RVDs for the first two positions. Among all targets tested, the DNA sequences containing AAN and CAN (corresponding to RVD pairs NI-NI-XX and HD-NI-XX) appeared to be statistically the least favorable to achieve a high *in vivo* nuclease activity (difference between AA and others: P-value = 0.005 and difference between CA and others: P-value = 0.002, Student t-test, Supplementary Figure S6A). However, we cannot exclude that the reduced activity monitored in yeast was due to not only multiple parameters involving the binding affinity but also protein stability or folding. In addition, these findings have to be tempered by the fact that such effects were strongly attenuated for the longer array (18.5 repeat) collection (difference between AA and others: P value = 0.2 and difference between CA and others: P-value = 0.04, Supplementary Figures S5B and S6B). Notably, although monitored on short array collections, no such issues were observed when sliding the experimental window along the array and target DNA from positions 1-2-3 to 3-4-5 and then 5-6-7 (Supplementary Figure S5C and D). This analysis of the data suggested that the first few N-terminal RVD/nucleotide pair may have the strongest impact on TALEN activity, consistent with our first experiments in CHO-KI and in other studies (28).

Furthermore, to assess possible context dependence at the RVD level, we systematically analyzed the impact, on a central RVD, of the neighboring two RVDs (positions− 1 and +1). The recovered uniform activity levels strongly suggest independence of the central RVD from their nearest (right and left)-neighbors

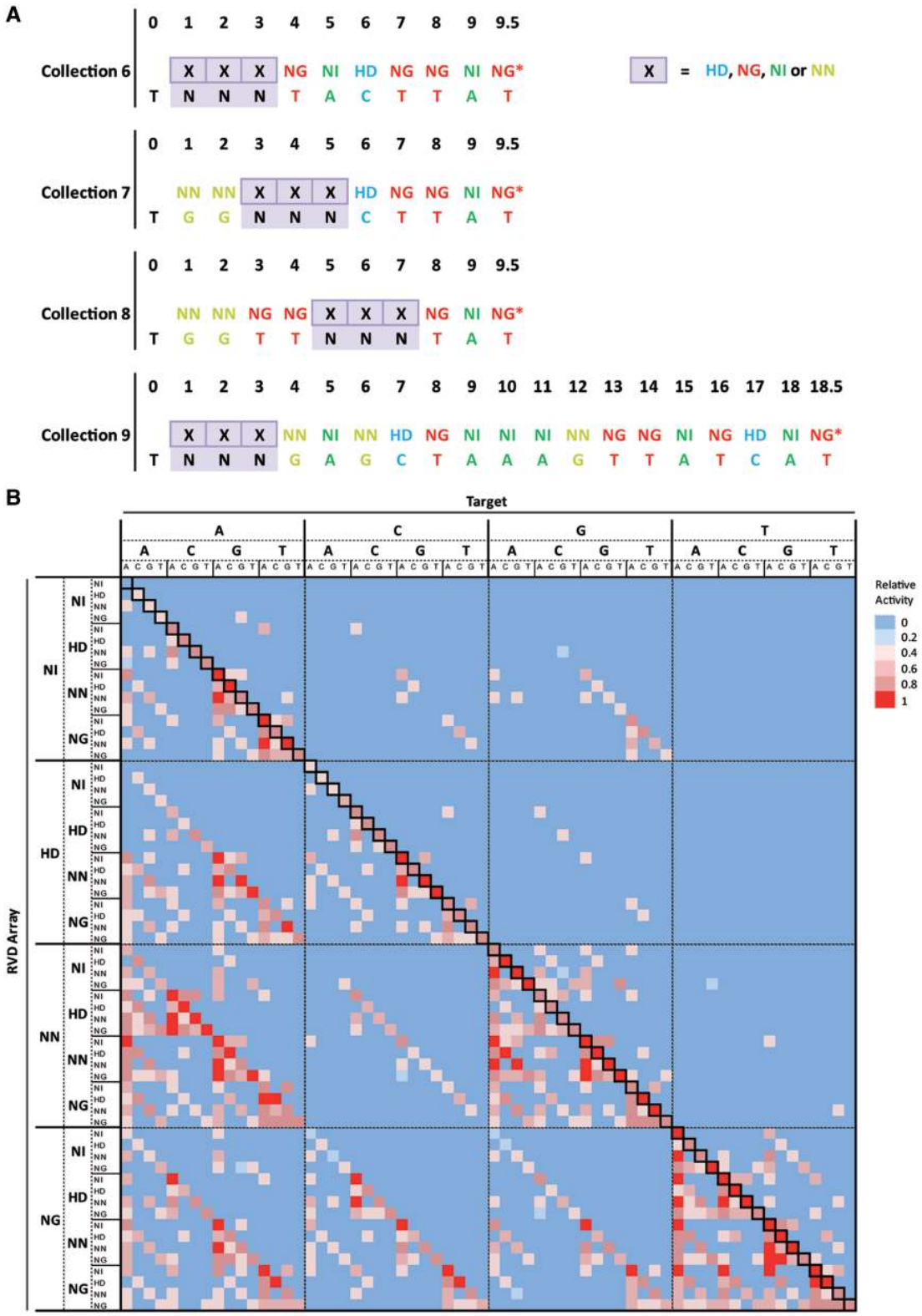

Figure 3. Setup of collections used in the large-scale experiments and graphical representation of activity results from Collection 6. (**A**) Collections of TALE and targets used for the study in yeast, where X represents any of the four NI, HD, NN and NG RVD and N any of the four A, C, G and T bases. Collections 6–8 are composed of arrays containing 9.5 repeats and Collection 9 is composed of arrays of 18.5. The TALEN collections were used in the homodimer format. (**B**) Heatmap showing the activity of the 64 TALEN of Collection 6 on the 64 corresponding targets. The outer line of text of the target (abscissa) represents the first nucleotide of the NNN triplet, the middle line of text represents the second nucleotide and the innermost line of text represents the third nucleotide. Likewise for the RVD array (XXX) on the ordinate. Red corresponds to maximum activity, whereas blue corresponds to no activity. The diagonal with framed squares represents the NI:A, HD:C, NN:G and NG:T pairings. In the case of a perfect one to one RVD/nucleotide association code, activity should be recovered on the diagonal only.

(Supplementary Figure S7A and B). These data represent an unambiguous indication that no short distance (1 base) strong context dependence between RVDs is present in the N-terminal region of TALEN array.

Analysis of the global effect of mismatches in RVD/nucleotide pairs

We then systematically analyzed the impact on TALEN activity of RVD/nucleotide mismatches with respect to the described code (NI:A, HD:C, NN:G and NG:T). Despite the fact that the NN RVD has been described to target both guanine and adenine nucleotides, in this analysis we consider as canonical only the NN:G pairing, as nearly all researchers are using this RVD uniquely in association with a G. When considering TALEN from Collection 6 that follows the recognition code, 89% of the molecules displayed robust activity on their cognate targets (Supplementary Table S3). Introducing two, four or six mismatches within the TALEN/target pairing led to a drop in active molecules to 30, 6 and 1%, respectively (Supplementary Table S3). In contrast, increasing the repeat array length to 18.5 (Collection 9) allowed accommodating up to six mismatches in 16% of the tested molecules while still maintaining a robust nuclease activity (Supplementary Table S3). The effect of mismatch number was position-independent from positions 1 to 7 with similar decreases observed when the experimental window was shifted to positions 3-4-5 (Collection 7) or 5-6-7 (Collection 8) when compared with Collection 6 (Supplementary Table S3).

RVD specificity and prediction of effect of mismatches on TALEN activity

Having observed an absence of context dependence between RVDs, we next performed an in-depth analysis of our entire dataset to further decipher the individual specificity of each RVD. For each position, from 1 to 7, we computed an experimental matrix describing the specificity of each RVD on the four DNA bases (Figure 4A and Supplementary Table S4). Interestingly, these matrices appear to be similar (mean standard deviation: 0.1), indicating the absence of positioning effects on specificity. By taking advantage of the conserved specificity of RVDs along the TALEN array, we then computed a global matrix (or logo) (Figure 4B and C and Supplementary Table S2) representing the global specificity of each RVD of a TALEN. The global matrix confirms that the currently used RVD/nucleotide pairing code represents the most appropriate solution to generate highly active TALEN. However, taken individually each RVD can tolerate to varying extents the three other bases, although at the expense of a reduction in activity (Figure 4B and C).

We next wanted to determine whether a direct correlation between the score given by our matrix (based on *in vivo* activity measurements) and the sole protein/DNA binding parameter (based on *in vitro* measurements) could be found. Toward this goal, we designed a single TALE DNA binding array that targets an endogenous sequence of 17 bp (in line with predominant natural-size TALEs) within the human IL2RG gene. The corresponding protein (lacking the FokI catalytic domain, Supplementary Table S5) was produced as soluble protein in *E. coli* and dissociation constant (K_d) for several targets containing various mismatches with respect to number (1–3), type and position were determined *in vitro* (24). We only found a moderate correlation between our scoring and the K_d values ($r = -0.54$ with $P = 0.022$, Supplementary Figure S8 and Supplementary Table S5). We believe that this discrepancy between the two variables possibly reflects key differences between the two experimental setups. Indeed, *in vitro* experiments only characterize the intrinsic DNA binding properties of a TALE DNA binding domain, whereas the *in vivo* experiments characterized the overall TALEN activity that involved not only the direct binding properties of the TALE array but also the catalytic properties of FokI catalytic domain (DNA binding affinity, dimerization and rate of dsDNA cleavage) and DNA repair mechanism (single-strand annealing or non-homologous end joining). Although, our results stressed that the binding affinity (TALE/DNA) may be an important contributor to the final nuclease output in living cells (creation and repair of the DSB), we believe that additional detailed *in vitro* studies would be desirable to precisely decipher the individual contribution of each parameter to the final output. One could also hypothesize that the *in vitro* measurements we obtained may better correlate with specificity of engineered TALE (where a single molecule is involved).

Finally, we scored, using our global specificity matrix, the relative loss of activity due to the mismatches present in our collections from the previous CHO-KI experiments (Collections 1–5). Based on our previous data (Figure 1D, E and F), we first divided our collection into two subsets of TALEN: (i) Collections 1–3 representing the 'N-terminal specificity constant' part of the array and (ii) Collections 4 and 5 representing the C-terminal part of the array with gradual loss of specificity. The use of the global specificity matrix allowed accurate prediction of the loss of activity for both subsets of TALEN (Collections 1–3: $r = 0.81$, $P = 4e{-}16$; Collections 4–5: $r = 0.84$, $P = 3e{-}12$, Figure 5A). Furthermore, the variation of the slope of the two regressions is in accordance with the gradual loss of specificity along the TAL DNA binding array we previously observed (Figure 1E).

Evaluation of off-site targeting on sequences containing low number of mismatches

The total number of TALEN sites on the human genome was estimated using our criteria (Materials and Methods) to be in the order of 500 millions, making it computationally intractable to calculate the potential off-target sites for every TALEN. Thereby, to statistically evaluate the theoretical specificity of TALEN, a set of 15 000 potential TALEN target sites (composed of 16 bases each) were randomly picked throughout the human genome. All off-site target sequences were then computationally determined for each of these 15 000 TALEN sequences. As TALEN molecules result from the co-expression of

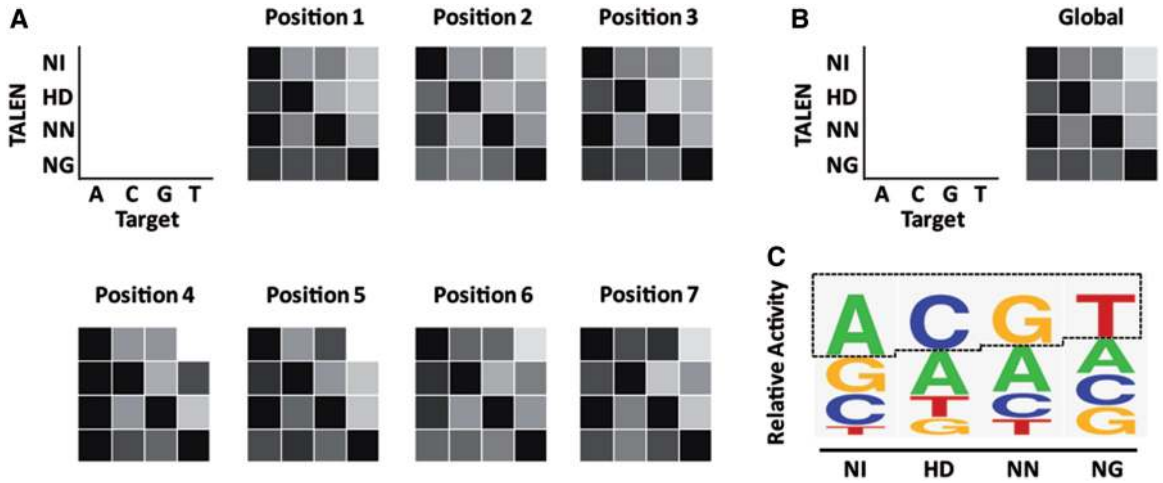

Figure 4. Graphical representation of RVD specificities. (**A**) Specificity measured for positions 1–7. The level of gray (black = 1, white = 0) represents the relative activity compared with the HD:C, NG:T NI:A and NN:G RVD/nucleotide pairing code. (**B**) Average specificity of the four HD, NG, NI and NN RVDs on the first 7 positions. (**C**) Logo representation of the global specificity matrix. Logo was generated using WebLogo (http://weblogo.berkeley.edu/logo.cgi). Values for relative specificities are presented in Supplementary Table S4.

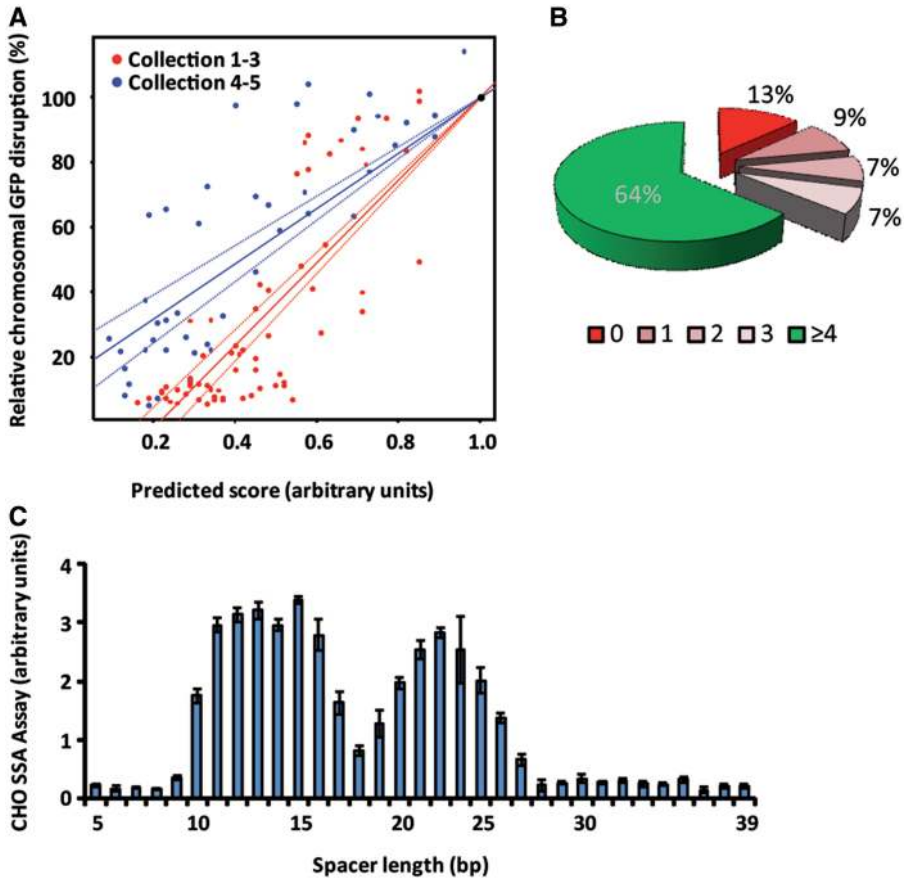

Figure 5. Effect of target mutations on TALEN activity in mammalian cells. (**A**) Correlation between experimental relative activities represented by the percentage of GFP-negative cells in the mammalian gene-targeting assay and scoring using the matrix presented in Figure 4B (Collections 1, 2 and 3 are represented in red: $r = 0.81$, $P = 4e{-}16$, and Collections 4 and 5 are represented in blue: $r = 0.84$, $P = 3e{-}12$). Linear regressions are presented for both subsets and 95% confidence intervals are represented by dashed lines. (**B**) Pie chart representation of the percentage of TALEN composed of 15.5 RVDs that will have, in the human genome, potential off-site targets containing no, one, two, three or four and more mismatches when considering a test set of 15 000 putative TALEN. All possible combinations of half TALEN (left + right, left + left, right + right) with a spacer length ranging from 9 to 30 bp were taken into account. (**C**) A collection of 33 targets comprising two AvrBs3 target sequences facing each other on both DNA strand with spacer length (between the two targets) ranging from 5 to 40 bp were designed and assayed in the CHO-K1 SSA assay to determine the optimal cleavage conditions. Targets containing a spacer of 21 and 35 bp were absent from the study.

two monomers, we also searched for off-site targets derived from the two potential palindromic TALEN target sites. Moreover, as the length of the DNA spacer between the two TALE recognition sites is known to tolerate a degree of flexibility(8–10,29), we included in our search any DNA spacer size from 9 to 30 bp. Using these criteria, TALEN can be considered extremely specific as we found that for nearly two-thirds (64%) of those chosen TALEN, the number of RVD/nucleotide pairing mismatches had to be increased to four or more to find potential off-site targets (Figure 5B). In addition, the majority of these off-site targets should have most of their mismatches in the first 2/3 of DNA binding array (representing the ''N-terminal specificity constant'' part, Figure 1). For instance, when considering off-site targets with three mismatches, only 6% had all their mismatches after position 10 and may therefore present the highest level of off-site processing. Although localization of the off-site sequence in the genome (e.g. essential genes) should also be carefully taken into consideration, the specificity data presented above indicated that most of the TALEN should only present low ratio of off-site/in-site activities.

To confirm this hypothesis, we designed six TALEN that present at least one potential off-target sequence containing between one and four mismatches. For each of these TALEN, we measured by deep sequencing the frequency of indel events generated by the non-homologous end-joining (NHEJ) repair pathway at the possible DSB sites. The percent of indels induced by these TALEN at their respective target sites was monitored to range from 1 to 23.8% (Table 1). We first determined whether such events could be detected at alternative endogenous off-target site containing four mismatches. Substantial off-target processing frequencies (>0.1%) were only detected at two loci (OS2-B, 0.4%; and OS3-A, 0.5%, Table 1). Noteworthy, as expected from our previous experiments, the two off-target sites presenting the highest processing contained most mismatches in the last third of the array (OS2-B, OS3-A, Table 1). Similar trends were obtained when considering three mismatches (OS1-A, OS4-A and OS6-B, Table 1). Worthwhile is also the observation that TALEN could have an unexpectedly low activity on off-site targets, even when mismatches were mainly positioned at the C-terminal end of the array when spacer length was unfavored (e.g. Locus2, OS1-A, OS2-A or OS2-C; Table 1 and Figure 5C).

Although a larger *in vivo* data set would be desirable to precisely quantify the trends we underlined, taken together our data indicate that TALEN can accommodate only a relatively small (<3–4) number of mismatches relative to the currently used code while retaining a significant nuclease activity.

DISCUSSION

Although TALEs appear to be one of the most promising DNA-targeting platforms, as evidenced by the increasing number of reports, limited information is currently available regarding detailed control of their activity and specificity (6,7,16,18,30). *In vitro* techniques [e.g. SELEX (8) or Bind-n-Seq technologies (28)] dedicated to measurement of affinity and specificity of such proteins are mainly limited to variation in the target sequence, as expression and purification of high numbers of proteins still remains a major bottleneck. To address these limitations and to additionally include the nuclease enzymatic activity parameter, we used a combination of two *in vivo* methods to analyze the specificity/activity of TALEN. We relied on both, an endogenous integrated reporter system in a

Table 1. Activities of TALEN on their endogenous cognate target (bold) and potential off-target sequences

Locus	Number of mismatches	Indels (%)	Ratio (%) off-site/in-site	Spacer length (bp)	Left target (5′–3′)	Right target (5′–3′)	Location	Indels (%)	Reads
L1	**0**	**23,8**	**100**	**15**			**chr7:148,544,235–148,544,283**	**424**	**1785**
OS1-A	3	0	0	27	ttaattgtatattGat	ttaattAtatattTat	chr6:67,836,053–67,836,113	0	12420
OS1-B	4	0	0	10	ttaattTtatattcat	tTaTgtaaaggAataa	chr2:167,063,216–167,063,259	0	12081
OS1-C	4	0	0	22	tgaagAaaaggAataa	ttTattgtatattAat	chr21:22,558,167–22,558,222	0	7223
OS1-D	4	0,01	0,04	16	tTaagtaaaAAtataa	ttaattTtatattcat	chr1:80,191,202–80,191,251	1	7243
OS1-E	4	0,05	0,2	29	ttaattTtatattcat	ttTaGtTtatattcat	chr2:167,063,216–167,063,278	7	12902
L2	**0**	**10,7**	**100**	**15**			**chr7:116,335,791–116,335,839**	**599**	**5601**
OS2-A	4	0	0	20	tccttcttcGcTgggC	tccttcttcacaTggt	chr19:11,753,683–11,753,736	0	763
OS2-B	4	0,5	5	12	tccttcttcacaAggt	tcctCcttcacaCgCt	chr7:155,697,642–155,697,687	14	2774
OS2-C	4	0	0	17	tccttcttcacTgggA	tccttcttcacaTggC	chr9:28,655,171–28,655,221	0	1003
L3	**0**	**13**	**100**	**15**			**chr18:45,423,236–45,423,284**	**78**	**599**
OS3-A	4	0,4	3	16	tttcaactGatAGtag	tCtcaacttatcatag	chr12:93,895,410–93,895,459	4	933
L4	**0**	**1**	**100**	**15**			**chr7:57,659,395–57,659,443**	**42**	**4171**
OS4-A	3	0,4	40	15	ttctaggaaaccaCct	tcttaCtaattctGtt	chr17:16,097,940–16,097,988	12	3141
OS4-B	1	1	100	15	tcttattaattctatt	ttctaggaaaccaCct	chr17:21,535,905–21,535,953	11	1099
L5	**0**	**17,3**	**100**	**15**			**chr12:58,145,360–58,145,408**	**877**	**5081**
OS5-A	4	0,05	0,3	20	tcctccaTctcTAcct	tccAccacctcctcct	chr11:47,746,108–47,746,161	1	2192
OS5-B	4	0	0	27	tcctccTcctcctcct	tcctTcTTctcctcct	chr2:67,777,895–67,777,955	0	2228
L6	**0**	**6,6**	**100**	**15**			**chr7:26,242,732–26,242,780**	**223**	**3357**
OS6-A	4	0,07	1	21	ttttcATctgtaattt	ttaTatccTcatattt	chr5:74,301,559–74,301,613	1	1485
OS6-B	3	0	0	21	ttttcccctgtaattt	tAacaGTcacatattt	chrX:71,283,522–71,283,580	0	565

CHO-KI mammalian cell line and a plasmid-based nuclease activity in yeast. These two approaches had the major advantages to be up-scalable to medium or high throughput and to minimize or normalize bias from epigenetic variations. To extend our knowledge on activity and specificity of RVDs along TALEN arrays, we thus analyzed the cleavage profiles of >15 500 TALEN/target combinations, leading to the most exhaustive analysis of this new class of DNA-targeting tools available today.

Although guidelines based on computational analysis of a small subset (20) of natural effectors have previously been published (27), recent studies showed that failing to follow these 'guidelines' had little or no effect on TALEN activity (11,31). Our systematic approach, based on investigating windows of RVD triplets along the TALE array, reveals the positional effects of unfavorable TALE DNA pairings in the N-terminal region of the DNA binding array. In agreement with Cermak *et al.* (27), we found that TALE designs should avoid targeting an Adenine (A) residue at position 2. Notably, we also show that the presence of AA or CA at the first two positions of the target sequence leads to a decrease in activity and should therefore be removed from TALE hit-search engines, especially when designing short arrays (in the range size of 9.5–12.5 repeats). Currently, the commonly used array lengths (15.5–20.5 repeats) should only be weakly impacted by the use of these combinations, although a larger data set composed of TALEN of common size (e.g. 15.5 repeats) harboring these two particular N-terminal RVD compositions would be desirable to confirm or infirm the validity of these findings for current TALEN designs. However, the use of shorter arrays could be of interest, as, when rationally and carefully designed (e.g. by the use of our specificity matrices), these arrays are more sensitive to mismatches and the resulting TALEN should potentially reveal a higher specificity as recently described (32,33).

In a recent large-scale study, Church and coworkers aimed to evaluate the landscape of targeting specificity of not only the clustered regularly interspaced short palindromic repeats (CRISPR)-Cas system but also of TALEs (32). They found, in accordance to previous results, that N-terminal repeats are more sensitive to mismatches, a trend that was also evidenced in this study for TALEN. They also reported that long arrays containing 18 repeats can tolerate up to two or three mismatches and shorter arrays (14 and 10 repeats) are much less tolerant to mismatches. The total number of tolerated mismatches in single TALE was roughly half of the one we reported in this study for TALEN. However, their study and ours raised a few differences between the behaviors of TALE and TALEN. In particular, they noticed a decrease in the activity on par with the reduced size of the array (18–10 repeats), a characteristic that we and others have not observed for TALEN (11). This feature might result from the specific architecture of TALEN that requires two binding monomers (versus only one for the TALE) and the dimerization of the FokI catalytic domains. A second observation that was not confirmed in our study using TALEN is the fact that mutations in the middle of the array can lead to higher activities (32).

The TALE 'code' currently used by most researchers is based on statistical analyses of a limited number of natural effectors and target gene promoters leading to the analysis of only a fraction of all pairing possibilities. In addition, when looking at a naturally occurring pairing between TALE and a plant target promoter, most, if not all, contain mismatches relative to a perfect one-to-one code (NI:A, HD:C, NN:G and NG:T). For instance, AvrBs3 (the TALE scaffold used in this study) contains, when bound to its target, three mismatches (two HD:A and one NG:C at positions 1, 15 and 17.5, respectively) indicating a certain degree of liberty relative to a perfect one-to-one association. In this work, we aimed at providing an extensive knowledge of the specificity of each TALE binding module (NI, HD, NN and NG) along TALEN/DNA binding arrays. Interestingly, although individual RVDs can tolerate mismatches, the cumulative effects of multiple mismatches within an array rapidly out-balances the overall TALE DNA-binding affinity. Additionally, our experimental designs allowed us to report in particular (i) the absence of short distance context dependence between RVDs, (ii) the gradual decrease of specificity within the C-terminal half array and (iii) a predictive model of RVD specificity scoring.

The combination of our experimental results with a large-scale computational analysis of 15 000 randomly chosen potential target sites and TALEN indicated that two-thirds of the nucleases (composed of two DNA-targeting cores of 15.5 repeats) should show a strong preference for the designed targets over possible off-site sequences, especially when the mismatches are present at the N-terminal end of the DNA-targeting core (Figure 5B, Table 1). Consistent with this statistical analysis, the experimental characterization of off-site mutations on putative targeted sites confirmed that TALEN having genomic off-site with less than four mismatches should be proscribed. The finding that TALEN have such high specificity is also coherent with previous studies reporting the absence of observed toxicity in HEK293 cells (29,34) and undetected or very low off-site targeting frequencies in rat (35), stem cells (36), *Xenopus* embryos (37) and the yeast genome (38). However, despite that this cutoff of four mismatches should be sufficient for a majority of designs; specific applications may require additional levels of safety (e.g. higher mismatches cutoff, use of obligatory FokI heterodimers). As reported by Hockemeyer *et al.,* TALEN containing up to nine mismatches (for two DNA binding cores of 16.5 repeats) can still show indel mutations at detectable frequencies (ratio off-site/in-site: 0.5%) (36).

Nonetheless, a potential caveat to our findings is that we did not take in consideration epigenetic factors that will inescapably be present at endogenous loci of most, if not all, organisms. For example, we (23) and others (13,31,39) recently reported the sensitivity of TALE-based arrays to cytosine methylation, a complexity currently not addressed in our experimental setup, that should however further decrease off-site targeting.

Nevertheless, additional experiments with larger sampling that combines epigenetic factors and mismatches would be desirable to fully assess the potential of this technology. We further envision that the constant accumulation of experimental data on TALEN activity and specificity will allow, in the near future, to rapidly fulfill these open questions.

Although we believe that our process is a step toward improving automated design methods for TALE-based molecular tools, additional issues such as the use of rare or unnatural RVDs may still be addressed to further extend the targeting possibilities. However, we envision that, together with the advances in, and access to, genome sequencing and epigenetic information, the implementation of our experimental model will permit a more rigorous and educated design of TALE-based tools. Beyond off-site prediction, the precise knowledge of alternative RVD/nucleotide pairing opens new possibilities to discriminate between sequences for application requiring, for example, allele-specific targeting. In conclusion, we anticipate that the provided results will expand our engineering capabilities by increasing our level of experimental control of TALE-based molecular tools, notably for applications in synthetic biology (22,40,41).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Andy Scharenberg for critical reading of the manuscript. They acknowledge the contribution of Frederic Cedrone, the Cellectis Nuclease Production Platform and the Cellectis Bioinformatics department.

FUNDING

Funding for open access charge: Cellectis.

Conflict of interest statement. All co-authors except S.S. and G.M. are Cellectis employees.

REFERENCES

1. DeFrancesco,L. (2011) Move over ZFNs. *Nat. Biotechnol.*, **29**, 681–684.
2. Bogdanove,A.J. and Voytas,D.F. (2011) TAL effectors: customizable proteins for DNA targeting. *Science*, **333**, 1843–1846.
3. Perez-Pinera,P., Ousterout,D.G. and Gersbach,C.A. (2012) Advances in targeted genome editing. *Curr. Opin. Chem. Biol.*, **16**, 268–277.
4. Mak,A.N., Bradley,P., Cernadas,R.A., Bogdanove,A.J. and Stoddard,B.L. (2012) The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science*, **335**, 716–719.
5. Deng,D., Yan,C., Pan,X., Mahfouz,M., Wang,J., Zhu,J.K., Shi,Y. and Yan,N. (2012) Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science*, **335**, 720–723.
6. Boch,J., Scholze,H., Schornack,S., Landgraf,A., Hahn,S., Kay,S., Lahaye,T., Nickstadt,A. and Bonas,U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509–1512.
7. Moscou,M.J. and Bogdanove,A.J. (2009) A simple cipher governs DNA recognition by TAL effectors. *Science*, **326**, 1501.
8. Miller,J.C., Tan,S., Qiao,G., Barlow,K.A., Wang,J., Xia,D.F., Meng,X., Paschon,D.E., Leung,E., Hinkley,S.J. *et al.* (2011) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.*, **29**, 143–148.
9. Christian,M., Cermak,T., Doyle,E.L., Schmidt,C., Zhang,F., Hummel,A., Bogdanove,A.J. and Voytas,D.F. (2010) Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics*, **186**, 757–761.
10. Li,T., Huang,S., Jiang,W.Z., Wright,D., Spalding,M.H., Weeks,D.P. and Yang,B. (2010) TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res.*, **39**, 359–372.
11. Reyon,D., Tsai,S.Q., Khayter,C., Foden,J.A., Sander,J.D. and Joung,J.K. (2012) FLASH assembly of TALENs for high-throughput genome editing. *Nat. Biotechnol.*, **30**, 460–465.
12. Mercer,A.C., Gaj,T., Fuller,R.P. and Barbas,C.F. III (2012) Chimeric TALE recombinases with programmable DNA sequence specificity. *Nucleic Acids Res.*, **40**, 11163–11172.
13. Bultmann,S., Morbitzer,R., Schmidt,C.S., Thanisch,K., Spada,F., Elsaesser,J., Lahaye,T. and Leonhardt,H. (2012) Targeted transcriptional activation of silent oct4 pluripotency gene by combining designer TALEs and inhibition of epigenetic modifiers. *Nucleic Acids Res.*, **40**, 5368–5377.
14. Geissler,R., Scholze,H., Hahn,S., Streubel,J., Bonas,U., Behrens,S.E. and Boch,J. (2011) Transcriptional activators of human genes with programmable DNA-specificity. *PLoS One*, **6**, e19509.
15. Morbitzer,R., Romer,P., Boch,J. and Lahaye,T. (2010) Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors. *Proc. Natl Acad. Sci. USA*, **107**, 21617–21622.
16. Zhang,F., Cong,L., Lodato,S., Kosuri,S., Church,G.M. and Arlotta,P. (2011) Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat. Biotechnol.*, **29**, 149–153.
17. Scholze,H. and Boch,J. (2011) TAL effectors are remote controls for gene activation. *Curr. Opin. Microbiol.*, **14**, 47–53.
18. Cong,L., Zhou,R., Kuo,Y.C., Cunniff,M. and Zhang,F. (2012) Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nat. Commun.*, **3**, 968.
19. Mahfouz,M.M., Li,L., Piatek,M., Fang,X., Mansour,H., Bangarusamy,D.K. and Zhu,J.K. (2012) Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein. *Plant Mol. Biol.*, **78**, 311–321.
20. Grau,J., Wolf,A., Reschke,M., Bonas,U., Posch,S. and Boch,J. (2013) Computational predictions provide insights into the biology of TAL effector target sites. *PLoS Comput. Biol.*, **9**, e1002962.
21. Streubel,J., Blucher,C., Landgraf,A. and Boch,J. (2012) TAL effector RVD specificities and efficiencies. *Nat. Biotechnol.*, **30**, 593–595.
22. Garg,A., Lohmueller,J.J., Silver,P.A. and Armel,T.Z. (2012) Engineering synthetic TAL effectors with orthogonal target sites. *Nucleic Acids Res.*, **40**, 7584–7595.
23. Valton,J., Dupuy,A., Daboussi,F., Thomas,S., Marechal,A., Macmaster,R., Melliand,K., Juillerat,A. and Duchateau,P. (2012) Overcoming TALE DNA binding domain sensitivity to cytosine methylation. *J. Biol. Chem.*, **287**, 38427–38432.
24. Stella,S., Molina,R., Yefimenko,I., Prieto,J., Silva,G., Bertonati,C., Juillerat,A., Duchateau,P. and Montoya,G. (2013) Structure of the AvrBs3-DNA complex provides new insights into the initial thymine-recognition mechanism. *Acta Crystallogr. D Biol. Crystallogr.*, **69**, 1707–1716.
25. Arnould,S., Chames,P., Perez,C., Lacroix,E., Duclert,A., Epinat,J.C., Stricher,F., Petit,A.S., Patin,A., Guillier,S. *et al.* (2006) Engineering of large numbers of highly specific homing endonucleases that induce recombination on novel DNA targets. *J. Mol. Biol.*, **355**, 443–458.
26. Daboussi,F., Zaslavskiy,M., Poirot,L., Loperfido,M., Gouble,A., Guyot,V., Leduc,S., Galetto,R., Grizot,S., Oficjalska,D. *et al.* (2012) Chromosomal context and epigenetic mechanisms control the efficacy of genome editing by rare-cutting designer endonucleases. *Nucleic Acids Res.*, **40**, 6367–6379.

27. Cermak,T., Doyle,E.L., Christian,M., Wang,L., Zhang,Y., Schmidt,C., Baller,J.A., Somia,N.V., Bogdanove,A.J. and Voytas,D.F. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.*, **39**, e82.
28. Meckler,J.F., Bhakta,M.S., Kim,M.S., Ovadia,R., Habrian,C.H., Zykovich,A., Yu,A., Lockwood,S.H., Morbitzer,R., Elsaesser,J. *et al.* (2013) Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Res.*, **41**, 4118–41128.
29. Mussolino,C., Morbitzer,R., Lutge,F., Dannemann,N., Lahaye,T. and Cathomen,T. (2011) A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res.*, **39**, 9283–9293.
30. Scholze,H. and Boch,J. (2011) TAL effector-DNA specificity. *Virulence*, **1**, 428–432.
31. Chen,S., Oikonomou,G., Chiu,C.N., Niles,B.J., Liu,J., Lee,D.A., Antoshechkin,I. and Prober,D.A. (2013) A large-scale *in vivo* analysis reveals that TALENs are significantly more mutagenic than ZFNs generated using context-dependent assembly. *Nucleic Acids Res.*, **41**, 2769–2778.
32. Mali,P., Aach,J., Stranges,P.B., Esvelt,K.M., Moosburner,M., Kosuri,S., Yang,L. and Church,G.M. (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.*, **31**, 833–838.
33. Bacman,S.R., Williams,S.L., Pinto,M., Peralta,S. and Moraes,C.T. (2013) Specific elimination of mutant mitochondrial genomes in patient-derived cells by mitoTALENs. *Nat. Med.*, **19**, 1111–1113.
34. Kim,Y., Kweon,J., Kim,A., Chon,J.K., Yoo,J.Y., Kim,H.J., Kim,S., Lee,C., Jeong,E., Chung,E. *et al.* (2013) A library of TAL effector nucleases spanning the human genome. *Nat. Biotechnol.*, **31**, 251–258.
35. Tesson,L., Usal,C., Menoret,S., Leung,E., Niles,B.J., Remy,S., Santiago,Y., Vincent,A.I., Meng,X., Zhang,L. *et al.* (2011) Knockout rats generated by embryo microinjection of TALENs. *Nat. Biotechnol.*, **29**, 695–696.
36. Hockemeyer,D., Wang,H., Kiani,S., Lai,C.S., Gao,Q., Cassady,J.P., Cost,G.J., Zhang,L., Santiago,Y., Miller,J.C. *et al.* (2011) Genetic engineering of human pluripotent cells using TALE nucleases. *Nat. Biotechnol.*, **29**, 731–734.
37. Lei,Y., Guo,X., Liu,Y., Cao,Y., Deng,Y., Chen,X., Cheng,C.H., Dawid,I.B., Chen,Y. and Zhao,H. (2012) Efficient targeted gene disruption in Xenopus embryos using engineered transcription activator-like effector nucleases (TALENs). *Proc. Natl Acad. Sci. USA*, **109**, 17484–17489.
38. Li,T., Huang,S., Zhao,X., Wright,D.A., Carpenter,S., Spalding,M.H., Weeks,D.P. and Yang,B. (2011) Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic Acids Res.*, **39**, 6315–6325.
39. Deng,D., Yin,P., Yan,C., Pan,X., Gong,X., Qi,S., Xie,T., Mahfouz,M., Zhu,J.K., Yan,N. *et al.* (2012) Recognition of methylated DNA by TAL effectors. *Cell Res.*, **22**, 1502–1504.
40. Blount,B.A., Weenink,T., Vasylechko,S. and Ellis,T. (2012) Rational diversification of a promoter providing fine-tuned expression and orthogonal regulation for synthetic biology. *PLoS One*, **7**, e33279.
41. Purnick,P.E. and Weiss,R. (2009) The second wave of synthetic biology: from modules to systems. *Nat. Rev. Mol. Cell Biol.*, **10**, 410–422.