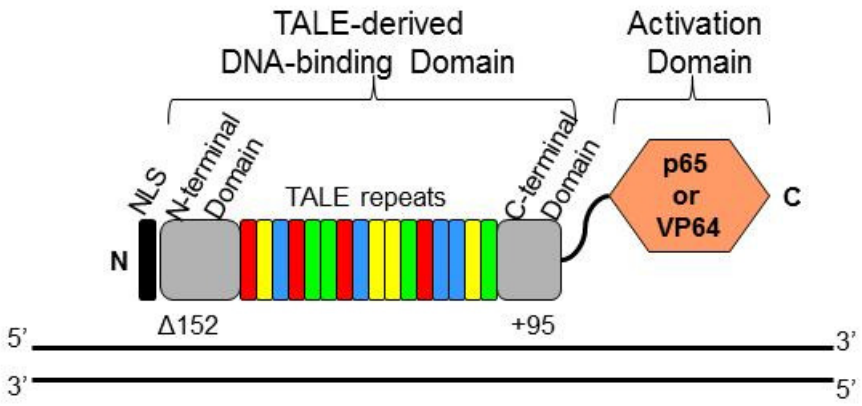
Robust, synergistic regulation of human gene expression using TALE activators

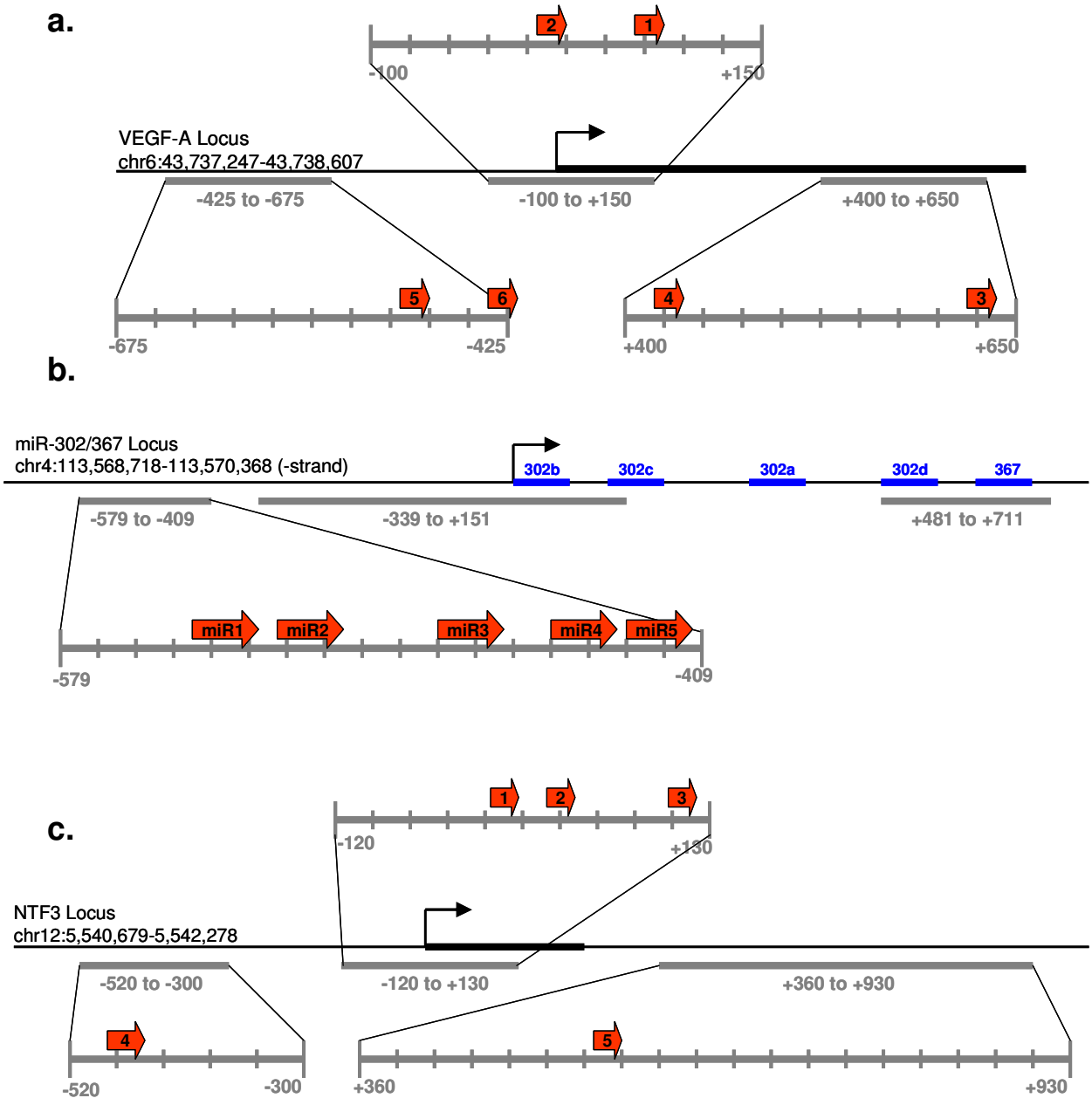
Morgan L. Maeder[1,2,3,4], Samantha J. Linder[1,2,3,*], Deepak Reyon[1,2,3,5,*], James F. Angstman[1,2,3], Yanfang Fu[1,2,3,5], Jeffry D. Sander[1,2,3,5], J. Keith Joung[1,2,3,4,5]

[1]Molecular Pathology Unit, Massachusetts General Hospital, Charlestown, MA 02129
[2]Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, MA 02129
[3]Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA 02129
[4]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115
[5]Department of Pathology, Harvard Medical School, Boston, MA 02115
[*]These authors contributed equally to this work
Correspondence to: JJOUNG@PARTNERS.ORG

Supplementary Figure 1	Schematic of TALE activator architecture used in this study
Supplementary Figure 2	Schematic overview of TALE activator binding sites within the VEGF-A, miR-302/367 and NTF3 gene promoter regions
Supplementary Table 1	List of previously published TALE activators and their reported activities on endogenous genes
Supplementary Table 2	List of target sites for TALE activators tested in this study and their violations of computationally-derived guidelines previously described by Bogdanove and colleagues
Supplementary Discussion	
Supplementary References	



Supplementary Figure 1. Schematic of TALE activator architecture used in this study. The TALE activator architecture we used for our experiments is similar to one described by Rebar and colleagues.[1] These proteins contain the Δ152 N-terminal domain and the +95 C-terminal domain that flank the TAL effector repeat array as well as an N-terminal nuclear localization signal (NLS) and a C-terminal activation domain (either VP64 or p65).



Supplementary Figure 2. Schematic overview of TALE activator binding sites within the (a) *VEGF-A*, (b) *miR-302/367,* and (c) *NTF3* gene promoter regions. Thick black lines indicate exons, thin black lines indicate introns or promoter regions, and black arrows indicate the start site of transcription. Thick blue lines represent microRNAs. DNase I hypersensitive sites are indicated with grey bars and those we targeted are expanded with red arrows depicting precise locations of TALE activator binding sites and orientations of the activators when bound on the DNA (the arrow indicates the direction of the protein from amino- to carboxy-terminus when bound to its target DNA site).

Gene Targeted	Organism/Cell line	TALE length (# of repeats)	Activation Domain	Approximate Fold Activation	Reference	Architecture
NTF3	Human HEK293 cells	17.5	VP16	30	1	A
SOX2	Human 293FT cells	12.5	VP64	5.5	2	B
KLF4	Human 293FT cells	12.5	VP64	2.2		
OCT4	Human 293FT cells	12.5	VP64	no activation		
C-MYC	Human 293FT cells	12.5	VP64	no activation		
PUMA	Human HEK293T-Rex cells	17.5	VP16	1.5	3	C
IFNa1	Human HEK293T-Rex cells	19.5	VP16	3.1		
IFNb1	Human HEK293T-Rex cells	17.5	VP16	3.5		
FXN	Human 293FT cells	13.5	VP64	0.9 to 1.7	4	B
				1.1 to 1.6		
				1.0 to 1.6		
				1.1 to 2.0		
				1.1 to 1.4		
				1.7 to 3.1		
				1.1 to 1.5		
OSGIN2	Human U-2OS cells	18.5	VP64	4.8	5	D
ZC3H10	Human U-2OS cells	18.5	VP64	1.3		
ROCK1	Human HeLa cells	16.5	VP64	n.d.	6	E
CACNA1C	Human 293FT cells	16.5	VP64	5.5	7	B
				2.75		
				4.5		
				6		
				3		
				1.5		
				4		
				3.5		
OCT4	Mouse ES cells	17.5	VP16	4	8	F
	Mouse neural stem cells			30*		
Bs3	pepper plants	13.5	native AvrHah1 activation domain	n.d.	9	G

* Activation observed in the presence of VPA and/or 5-azadC

Architecture Key:
A = originally referenced in Miller et al., *Nature Biotech* 2011
B = originally referenced in Zhang et al., *Nature Biotech* 2011
C = originally referenced in Geissler et al., *PLoS ONE* 2011
D = originally described in Garg et al., *NAR* 2012
E = originally described in Huang et al., *Nature Biotech* 2011
F = originally described in Morbitzer et al., *NAR* 2011
G = originally described in Cermak et al., *NAR* 2011

Supplementary Table 1. List of previously published TALE activators and their reported activities on endogenous genes.[1-9]

TALE Name	Target site	Guidelines					Total Guideline Violations
		1	2	3	4	5	
VEGF1	TCGGGAGGCGCAGCGGTT					X	1
VEGF2	TTGGGGCAGCCGGGTAGC		X		X	X	3
VEGF3	TGGAGGGGGTCGGGGCTC				X	X	2
VEGF4	TGAGTGACCTGCTTTTGGG			X	X	X	3
VEGF5	TGAGTGAGTGTGTGCGTGT			X		X	2
VEGF6	TCACTCCAGGATTCCAATA			X	X		2
Ntf3-1	TTCTGTTCACGGGACTCA		X		X		2
Ntf3-2	TCCGAACAGCTCCGCGCA				X		1
Ntf3-3	TTCCCCTGCTGGGTAGTG		X		X	X	3
Ntf3-4	TACGCCTCAGACCTGATC				X		1
Ntf3-5	TCCCTCAATCTGGGAAAG				X		1
miR1	TGGAAGCAATCTATTTAT						0
miR2	TACATTTAACATGTAGAT						0
miR3	TAGAAACACAATGCCTTT						0
* miR4	TGGGAGCACTCATTGTTA				X	X	2
miR5	TAATCTATGCCATCAAAC			X	X		2
VEGF1-1	TTGGGGGTGACCGCCG		X		X	X	3
VEGF1-2	TTGGGGGTGACCGCCGGA		X		X	X	3
VEGF1-3	TTGGGGGTGACCGCCGGAGC		X		X	X	3
VEGF1-4	TTGGGGGTGACCGCCGGAGCGC		X		X	X	3
VEGF1-5	TTGGGGGTGACCGCCGGAGCGCGG		X		X	X	3
VEGF1-6	TTGGGGGTGACCGCCGGAGCGCGGCG		X		X	X	3
VEGF2-1	TCCCGCAGCTGACCAG				X	X	2
VEGF2-2	TCCCGCAGCTGACCAGTC				X		1
VEGF2-3	TCCCGCAGCTGACCAGTCGC				X	X	2
VEGF2-4	TCCCGCAGCTGACCAGTCGCGC				X	X	2
VEGF2-5	TCCCGCAGCTGACCAGTCGCGCTG				X	X	2
VEGF2-6	TCCCGCAGCTGACCAGTCGCGCTGAC				X	X	2
VEGF3-1	TACCACCTCCTCCCCG				X	X	2
VEGF3-2	TACCACCTCCTCCCCGGC				X	X	2
VEGF3-3	TACCACCTCCTCCCCGGCCG				X	X	2
VEGF3-4	TACCACCTCCTCCCCGGCCGGC				X		1
VEGF3-5	TACCACCTCCTCCCCGGCCGGCGG				X	X	2
VEGF3-6	TACCACCTCCTCCCCGGCCGGCGGCG				X	X	2

TALE Name	Target site	Guidelines					Total Guideline Violations
		1	2	3	4	5	
VEGF4-1	TCCCCGGCCGGCGGCG				X	X	2
VEGF4-2	TCCCCGGCCGGCGGCGGA				X	X	2
VEGF4-3	TCCCCGGCCGGCGGCGGACA				X	X	2
VEGF4-4	TCCCCGGCCGGCGGCGGACAGT					X	1
VEGF4-5	TCCCCGGCCGGCGGCGGACAGTGG				X	X	2
VEGF4-6	TCCCCGGCCGGCGGCGGACAGTGGAC				X	X	2
VEGF5-1	TGGACGCGGCGGCGAG				X	X	2
VEGF5-2	TGGACGCGGCGGCGAGCC				X	X	2
VEGF5-3	TGGACGCGGCGGCGAGCCGC				X	X	2
VEGF5-4	TGGACGCGGCGGCGAGCCGCGG				X	X	2
VEGF5-5	TGGACGCGGCGGCGAGCCGCGGGC				X	X	2
VEGF5-6	TGGACGCGGCGGCGAGCCGCGGGCAG				X	X	2
VEGF6-1	TCCCAAGGGGGAGGGC				X	X	2
VEGF6-2	TCCCAAGGGGGAGGGCTC				X	X	2
VEGF6-3	TCCCAAGGGGGAGGGCTCAC				X	X	2
VEGF6-4	TCCCAAGGGGGAGGGCTCACGC				X	X	2
VEGF6-5	TCCCAAGGGGGAGGGCTCACGCCG				X	X	2
VEGF6-6	TCCCAAGGGGGAGGGCTCACGCCGCG				X	X	2
VEGF7-1	TCCGTCAGCGCGACTG				X	X	2
VEGF7-2	TCCGTCAGCGCGACTGGT					X	1
VEGF7-3	TCCGTCAGCGCGACTGGTCA				X	X	2
* VEGF7-4	TCCGTCAGCGCGACTGGTCAGC				X	X	2
VEGF7-5	TCCGTCAGCGCGACTGGTCAGCTG				X	X	2
VEGF7-6	TCCGTCAGCGCGACTGGTCAGCTGCG				X	X	2
VEGF8-1	TCCACTGTCCGCCGCC				X		1
VEGF8-2	TCCACTGTCCGCCGCCGG				X	X	2
VEGF8-3	TCCACTGTCCGCCGCCGGCC				X	X	2
VEGF8-4	TCCACTGTCCGCCGCCGGCCGG				X	X	2
VEGF8-5	TCCACTGTCCGCCGCCGGCCGGGG				X	X	2
VEGF8-6	TCCACTGTCCGCCGCCGGCCGGGGAG				X	X	2
VEGF9-1	TCCACCCCGCCTCCGG				X	X	2
VEGF9-2	TCCACCCCGCCTCCGGGC				X	X	2
VEGF9-3	TCCACCCCGCCTCCGGGCGC				X	X	2
VEGF9-4	TCCACCCCGCCTCCGGGCGCGG				X	X	2
VEGF9-5	TCCACCCCGCCTCCGGGCGCGGGC				X	X	2
VEGF9-6	TCCACCCCGCCTCCGGGCGCGGGCTC				X	X	2

Supplementary Table 2. List of target sites for TALE activators tested in this study and their violations of computationally-derived guidelines previously described by Bogdanove and colleagues.[10] Guidelines are numbered as described in Supplementary Discussion with an X indicating violation. Asterisks indicate target sites for which we did not obtain functional TALE activators.

Supplementary Discussion

Potential explanations for more robust activity of TALE activators observed in this study

In contrast to previously published literature, our results demonstrate a high success rate for engineering TALE activators capable of inducing robust activation of their target genes. We found that 62 of 65 (~95%) VP64 TALE activators (for which we could calculate a fold-activation) increased expression of their target genes by five-fold or more. By comparison, previously published studies found that only 4 of 26 (~15%) TALE activators increased expression of their endogenous gene target by five-fold or more. We theorize that our targeting of sites in DNase I HSSs, as has been reported for engineered zinc finger transcription factors,[11] may be an important factor in our higher success rate because these regions not only represent open chromatin but may also encompass the binding sites for endogenous transcription factors. However, this hypothesis remains to be formally tested in future experiments. Additionally, as others have recently shown, the sequence architecture of the TALE repeats themselves (i.e.—the amino acids present at non-hypervariable positions within the repeats) may also influence DNA-binding activity.[12] In this regard, we note that all of the TALE repeat arrays tested in this study were constructed using our FLASH assembly method[13] on a single common architecture.

Expanded targeting range of monomeric TALE activators

Our results strongly suggest that restrictive targeting guidelines published in a previous report[10] need not be followed to obtain active TALE activators. Cermak et al. originally

proposed five guidelines for identifying optimal target sites of engineered dimeric TALENs.[9] These guidelines were computationally derived from data on the binding preferences of naturally occurring TAL effectors but were not prospectively tested experimentally. As summarized previously,[13] these five guidelines can be stated as follows:

1. The nucleotide just 5' to the first nucleotide in the TAL effector repeat array binding site should be a thymine.
2. The first nucleotide of the TAL effector repeat array binding site should not be a thymine.
3. The second nucleotide of the TAL effector repeat array binding site should not be an adenosine.
4. The 3' most nucleotide of the TAL effector repeat array binding site should be a thymine.
5. The base composition of the TAL effector repeat array binding site should not vary from the observed percent composition of naturally occurring binding sites by more than 2 standard deviations. The percent composition of naturally occurring TAL effector repeat array binding sites was determined to be: A = 31±16%, C = 37±13%, G = 9±8%, T = 22±10%. Therefore, the base composition of TALE binding sites should be: A = 0-63%, C = 11-63%, G = 0-25%, T = 2-42%.

In a large-scale study, our group recently demonstrated that highly active dimeric TALENs can be made for target sites that violate one or more of guidelines 2 through 5 (none of the sites we targeted violated guideline 1) and also showed that no significant correlation exists between the number of guideline violations and the activities of the engineered TALENs.[13]

More recently, Doyle et al. suggested that target site selection for monomeric TAL effector-based proteins should be limited by these same five guidelines.[10] The TALE-NT 2.0 web-based software tool (https://boglab.plp.iastate.edu/)[10] also applies these five guidelines in its default settings when choosing target sites for monomeric TAL effector repeat arrays used in TALE activators.

However, the implementation of these guidelines has the effect of substantially limiting the targeting range of engineered monomeric TAL effector repeat arrays. For example, application of the five guidelines restricts the identification of a targetable 18 bp site (bound by a 16.5 TAL effector repeat array) to once in every 27 bps of random DNA sequence.

We used our data on 68 sites for which we were able to make active VP64 TALE activators to test the importance of following five computationally-derived guidelines for target site choice. 65 of these 68 sites fail to meet one or more of these five guidelines with 56 of these sites violating two or more guidelines (note that all of the sites did meet the guideline requiring a 5' T) (**Supplementary Table 2**). Our ability to successfully

obtain active TALE activators for all 68 of these sites clearly demonstrates that there is no absolute requirement to follow four of the five design guidelines. We conclude that highly active monomeric TALE activators can be made without meeting four of the five design guidelines. The ability to relax these restrictions improves the targeting range of TALE activators by more than ten-fold – for example, enabling proteins consisting of 16.5 TAL effector repeats to be made for a site once in every two bps of random DNA sequence, a more than 13-fold improvement in targeting range.

This resulting capability to make TALE activators for a broader range of DNA sequences is important because there are at least two factors that may potentially limit where one targets within a given gene: (1) a previous study[11] suggests that artificial transcriptional activators should ideally be designed to sites within DNase I HSSs and (2) the choice of the specific sequence targeted may help to minimize off-target effects (see below). Given this expanded targeting range of TALE activators, we have updated our Zinc Finger and TALE (**ZiFiT**) Targeter software (http://zifit.partners.org) to assist potential users with identifying target sites in their genes of interest.

Potential strategies for varying the expression of endogenous genes using TALE activators

In this study, we used TALE activators to induce target genes across a wide range of expression levels. Our experiments suggest three potential approaches that could be used to fine-tune the level of gene expression induced by TALE activators. (1) Varying the position of TALE activator binding within a single DNase I HSS. We do not currently

understand why targeting to different sites within a particular DNase I HSS leads to variable levels of activation (or why some TALE activators fail to show activity). Our results at the *VEGF-A* locus do not suggest any obvious correlation between activity and distance, location, and/or orientation of the binding site relative to the transcription start site. Potential explanations include the displacement of naturally occurring transcription factors and/or the methylation status of the target DNA site. Nonetheless, although it is currently not possible to predict the level of activation induced from any given site, the high success rate and broad targeting range of TALE activators make it straightforward for an investigator to construct and empirically characterize a series of proteins to find one that induces the desired level of target gene expression; (2) Varying the transcriptional activation domain in a TALE activator. For example, in the two human cell lines we examined, VP64 TALE activators generally induced higher levels of gene expression than matched counterparts bearing a NF-KB p65 activation domain; and (3) Combining multiple TALE activators can lead to even greater increases in target gene expression. The finding that artificial TALE activators can function synergistically to activate transcription further broadens the range of gene expression that can be achieved with this platform.

Off-target effects of TALE activators

Although we have demonstrated that the FLASH platform can be used to construct highly active TALE activators, future work will also need to address the specificities of such proteins. At present, relatively little is understood about the specificities of TALE activators but we note that proteins harboring 16.5 or more repeats should bind to

sequences of sufficient length (18 or more bps) to have a high probability of being unique in a complex genome. Our demonstration that the targeting range of TALE activators is substantially greater than previously suggested will also provide greater flexibility for choosing target sites to minimize potential off-target effects (once methods for determining these undesired binding events have been developed). In addition, two reports have recently suggested that a TAL effector repeat bearing hypervariable residues NH may be more specific for G than the repeat bearing NN (which in some contexts can also bind to A) that we used in all of our TALE activators.[7, 12] One of these reports suggested that this NH repeat may bind more weakly than the NN repeat and therefore described recommendations (based on transient transfection reporter assays in plants) for how and when to use this NH repeat without compromising activity.[12] Construction and testing of variants based on our 68 successful VP64 activators could provide a large-scale prospective test of these guidelines and of the effect of substituting NN repeats with NH repeats on the activities and specificities of TALE activators at their endogenous gene targets in human cells.

Web-based ZiFiT Targeter software for TALE activator design

We have updated our publicly available Zinc Finger and TALE (ZiFiT) Targeter webserver to include tools designed to assist users interested in assembling monomeric TAL effector repeat arrays using our FLASH[13] assembly method. Support is also provided for making arrays using our lower-throughput REAL[14, 15] and REAL-Fast[15] methods because these approaches yield TAL effector repeat arrays that are identical in amino acid sequence to those made by FLASH. Our updated ZiFiT Targeter version 4.2

is currently available without registration at: http://zifit.partners.org Identification of potential target sites using ZiFiT Targeter is performed by entering a sequence of interest into a query box. ZiFiT Targeter will, by default, identify TAL effector repeat arrays composed of 16.5 repeats that bind to sites 18 bp in length. Users can change this length constraint by entering a new value in the length input box.

Depending on the mode of assembly chosen (FLASH or REAL/REAL-Fast), ZiFiT will provide users with the names of plasmids required for assembly, and in the case of REAL or REAL-Fast assembly, a printable graphical guide. All plasmids required to practice REAL are available through the non-profit plasmid distribution service Addgene (http://www.addgene.org/talengineering/). The archive of 376 plasmids required to practice FLASH and REAL-Fast are available by request from the Joung Lab (http://www.TALengineering.org).

ZiFiT Targeter also provides tools to help users verify their TALE activator plasmids after they have been assembled. As noted above, the DNA and amino acid sequences of TAL effector repeat arrays assembled by FLASH, REAL, and REAL-Fast are all identical. Users can download sequences for their specific engineered TAL effector repeat array using ZiFiT Targeter. Because alignment of TAL effector repeat array sequences can be challenging due to their highly repetitive nature, ZiFiT Targeter also provides a sequence alignment tool that attempts to align DNA sequence reads to the consensus sequence after anchoring sequences encoding the non-repetitive amino- and carboxy-terminal TAL effector-derived sequences that flank the repeat array.

Supplementary References

1. Miller, J.C. et al. *Nat Biotechnol* **29**, 143-148 (2011).
2. Zhang, F. et al. *Nat Biotechnol* **29**, 149-153 (2011).
3. Geissler, R. et al. *PLoS One* **6**, e19509 (2011).
4. Tremblay, J.P., Chapdelaine, P., Coulombe, Z. & Rousseau, J. *Hum Gene Ther* **23**, 883-890 (2012).
5. Garg, A., Lohmueller, J.J., Silver, P.A. & Armel, T.Z. *Nucleic Acids Res* **40**, 7584-7595 (2012).
6. Wang, Z. et al. *Angew Chem Int Ed Engl* **51**, 8505-8508 (2012).
7. Cong, L., Zhou, R., Kuo, Y.C., Cunniff, M. & Zhang, F. *Nat Commun* **3**, 968 (2012).
8. Bultmann, S. et al. *Nucleic Acids Res* **40**, 5368-5377 (2012).
9. Cermak, T. et al. *Nucleic Acids Res* **39**, e82 (2011).
10. Doyle, E.L. et al. *Nucleic Acids Res* **40**, W117-122 (2012).
11. Liu, P.Q. et al. *J Biol Chem* **276**, 11323-11334 (2001).
12. Streubel, J., Blucher, C., Landgraf, A. & Boch, J. *Nat Biotechnol* **30**, 593-595 (2012).
13. Reyon, D. et al. *Nat Biotechnol* **30**, 460-465 (2012).
14. Sander, J.D. et al. *Nat Biotechnol* **29**, 697-698 (2011).
15. Reyon, D., Khayter, C., Regan, M.R., Joung, J.K. & Sander, J.D. *Curr Protoc Mol Biol* **Chapter 12**, Unit12 15 (2012).