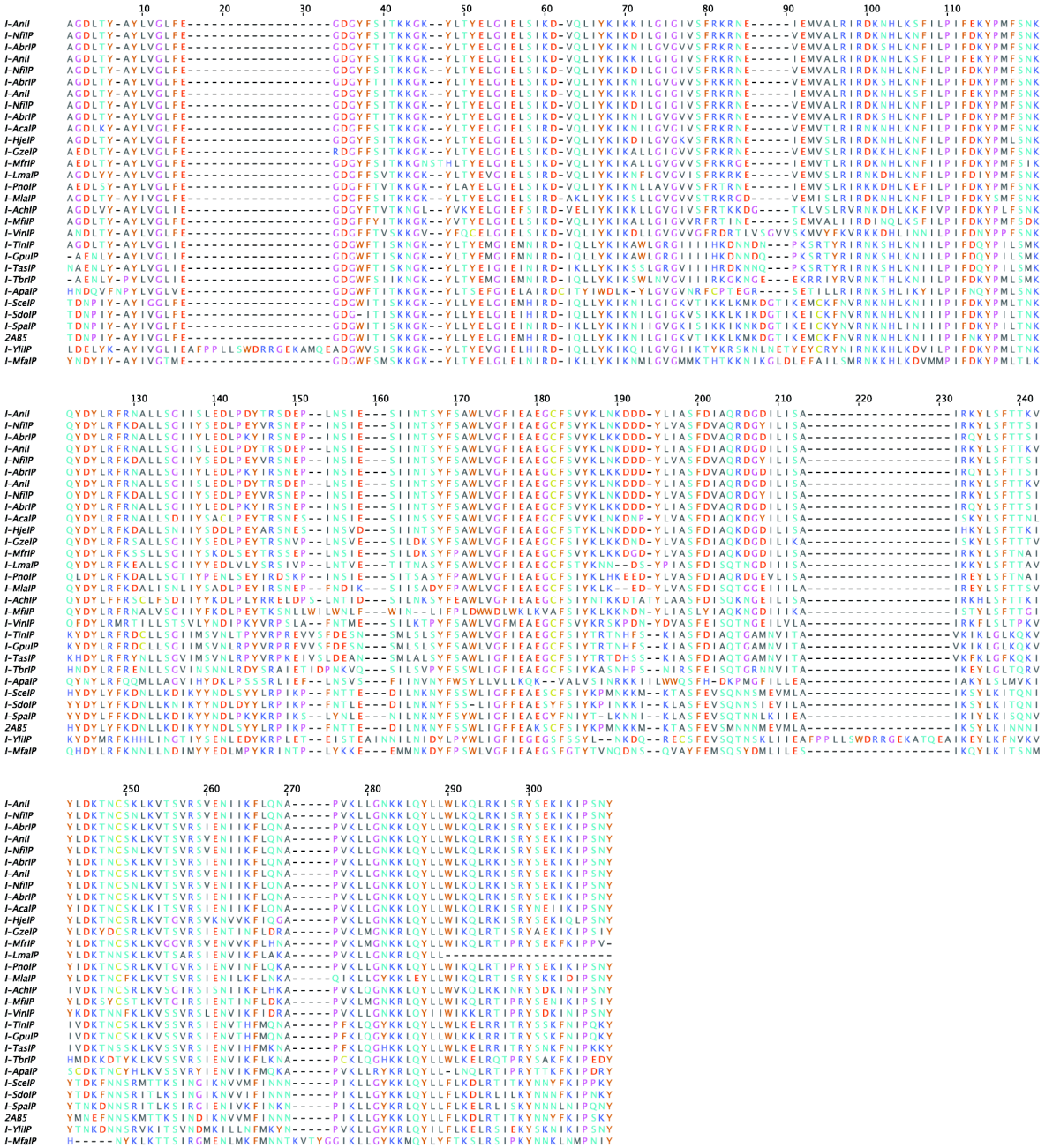
Supplementary Information

Contents:

Supplementary Figure S1. Alignment of proteins homologous to I-AniI.
Multiple sequence alignment of proteins homologous to I-AniI found using NCBI BLAST's blastp and tblastn (1) that have putative (denoted by suffix P (2)) endonuclease activity as of March 2011. The repository of data from new protein sequences and their characterizations is continuously growing (e.g. the recently elucidated maturase activity of the I-AniI homologue 2AB5 (3)). The most distant homologue in this alignment, I-MfaIP, has 47% sequence identity and 70% similarity. Alignments with less than 40% identity were predicted to have significantly divergent putative target site sequences (data not shown) and were not chosen for examination.



Supplementary Table S1. Percent identity and similarity to I-AniI for each homologue previously identified and aligned in Figure S1.

Homologue Name	Accession Number	Percent Identity to I-AniI	Percent Similarity to I-AniI
I-AniI	pdb|2QOJ|Z	100%	100%
I-NfiIP	gb|AAX39426.1	93%	98%
I-AbrIP	dbj|AB026120.1	92%	97%
I-AcaIP	gb|EER36338.1	88%	95%
I-HjeIP	gb|AF447590.1	86%	95%
I-GzeIP	gb|ABC86634.1	84%	93%
I-MfrIP	gb|GU952815.1	81%	90%
I-LmaIP	emb|CBX89979.1	79%	89%
I-PnoIP	gb|ABU49435.1	78%	91%
I-MlaIP	gb|GU933642.1	72%	86%
I-AchIP	gb|AAX34413.1	71%	86%
I-MfiIP	gb|AF343070.1	69%	79%
I-VinIP	gb|AAB95258.1	64%	79%
I-TinIP	dbj|AB175764.1	59%	75%
I-GpuIP	dbj|AB175779.1	59%	75%
I-TasIP	dbj|AB175746.1	57%	74%
I-TbrIP	dbj|AB175750.1	57%	76%
I-ApaIP	gb|AF538047.1	52%	67%
I-SceIP	emb|CAA32785.1	51%	70%
I-SdoIP	emb|X59280.1	50%	72%
I-SpaIP	gb|EU852811.1	50%	72%
2AB5 (maturase)	pdb|2AB5|A	49%	67%
I-YliIP	emb|AJ307410.1	49%	66%
I-MfaIP	emb|CAY39274.1	47%	70%

Supplementary Figure S2. Example putative target site identification for I-PnoIP, an I-AniI homologue. An intron (approximately spanning positions 8073 through 9092) within the cytochrome B gene of the complete annotated Phaeosphaeria nodorum SN15 mitochondrion genome (gb|EU053989.1) is observed to encode a LAGLIDADG endonuclease from position 8108 through 9037 (bold text), the homologue ORF. The flanking nucleotide sequences were examined and aligned as shown to each half of the native I-AniI target site. The resulting predicted sequence of the target site for I-PnoIP is colored yellow.

```
Pno genome 8041 aggacaagatattgttgaatttatttgaggtggtttaaacacagttggac  8090
                                     |||||.|||||
                ------------------------TGAGGAGGTTT-------------- ← (-) half of I-AniI site

           8091 cacattacggtgacgta**atgttaaaaattttgcttaatgctggaaaatcc**  8140

           8141 **cccaatttagtatttgcatacgatttattcttgattttaacaacaattat**  8190

           8191 **ttacgtaaaaattgcattgacatggggacaatcagcaggggtgagaagta**  8240

           8241 **tacacacttcagaagcctctcagagactacatgcagaagatctctcatat**  8290

           8291 **gcttatttagtaggtttatttgaaggtgacggtttttttactgttacaaa**  8340

           8341 **aaaaggtaaatatctagcctatgaattaggtattgaattgtctattaaag**  8390

           8391 **acgttcaattgatatataaaattaaaaatcttttagctgtaggtgtagta**  8440

           8441 **agttttagaacaagaaatgaaattgaaatggtatctttaagaattagaaa**  8490

           8491 **aaaagaccatttaaaagaatttattctacccatatttgataaatacccta**  8540

           8541 **tgttttctaataaacaacttgattacttaagatttaaagacgcactatta**  8590

           8591 **tctggtactatatatccagagaatttatctgaatatattagagatagtaa**  8640

           8641 **acctataaattcgatagaatctattacaagtgcttcttattttcccgctt**  8690

           8691 **gattagtaggatttatagaagctgaaggttgtttcagtatttacaaatta**  8740

           8741 **cacaaagaggaagattatttagtggctagtttcgatattgctcaaagaga**  8790

           8791 **tggagaggtattaatatctgctattcgtgagtatttatcttttactaatg**  8840

           8841 **ctatatacatagataaaactaattgttccagactgaaagttacaggtgta**  8890

           8891 **agatctatagaaaatgttattaatttttacaaaaggctcctgtaaaatt**  8940

           8941 **attgggtaataaaaaattacaatatttattatgaattaaacaattacgta**  8990

           8991 **ctatacctagatattcagagaaaattaagataccttcaaattactaaa**ga  9040

           9041 gagatcaagatatagtccgatcaataaagaaatttattgagcgtaacgat  9090

           9091 agtttgtttcaaatgaagttaccaacacaaatgctctgttaacaatgcta  9140
                                               ||||||.||
                ---------------------------------CTCTGTAAA-------- ← (+) half of I-AniI site
```

Supplementary Table S2. $EC_{0.5max}$ (nM) cleavage efficiencies and cleavage plateaus (f_{max} , the maximal fraction of site cleavage) for homologue-based I-AniI variants tested on singly-substituted target sites. Variants are grouped into categories (C-terminal loops, K200, Central 4 loops, Core Mutations) dependent on the location and theorized role of the mutations transferred to the I-AniI scaffold. All variants include the F80K and L233K mutations for solubility, with the exception of the K24N/T29K variant that only includes F80K. The base activity column indicates whether the variant was made with the activating F13Y or S111Y mutations. $EC_{0.5max}$ values are the mean (nM) ± coefficient of variation (%CV) of two independently determined enzyme cleavage profiles chosen (via inspection for substrate degradation and experimental error) from at least two separate *in vitro* cleavage assays on plasmid DNA substrates containing single base-pair substitutions from the I-AniI wild-type target site. The CV is given as a percentage measure of variability estimated through dividing the standard error from the mean (SEM) by the mean. Values of $EC_{0.5max}$ >750nM are too high to allow accurate quantitative determination, and therefore no cleavage plateau is reported. $f_{max} \approx 1$ designates a cleavage plateau at its greatest allowable value, meaning the substrate can be completely cleaved with no remaining uncut fraction.

Variant	Mutations	Base Activity	Position		$EC_{0.5max}$ (nM)	Cleavage Plateau
C-terminal loops						
I-PnoIP	V153I, N157H, D159E, D160E, I164V	F13Y	wild-type		4 ± 3%	0.980
			+8A	C	3 ± 28%	0.987
				G	19 ± 40%	0.938
				T	3 ± 41%	0.975
			+9A	C	70 ± 23%	~1
				G	41 ± 23%	~1
				T	15 ± 31%	0.987
			+10G	A	6 ± 9%	0.981
				C	8 ± 27%	0.960
				T	5 ± 42%	0.974
I-AchIP	V153I, K155N, L156T, N157K, D160T, inserted A after 160, D161T, I164A	F13Y	wild-type		8 ± 6%	0.977
			+7T	A	8 ± 31%	0.995
				C	74 ± 5%	0.965
				G	328 ± 7%	~1
			+8A	C	8 ± 4%	0.985
				G	13 ± 6%	0.948
				T	6 ± 27%	0.980
			+9A	C	10 ± 8%	0.981
				G	11 ± 8%	0.986
				T	9 ± 3%	0.981
			+10G	A	8 ± 20%	0.966
				C	8 ± 13%	0.957
				T	8 ± 23%	0.974
I-TasIP	K155T, L156R, N157T, K158 deletion, D160H, D161S, Y162S, L163K	F13Y	wild-type		86 ± 35%	0.941
			+7T	A	20 ± 32%	0.934
				C	588 ± 26%	~1
				G	459 ± 41%	0.752
			+8A	C	48 ± 18%	0.856
				G	6 ± 13%	0.942

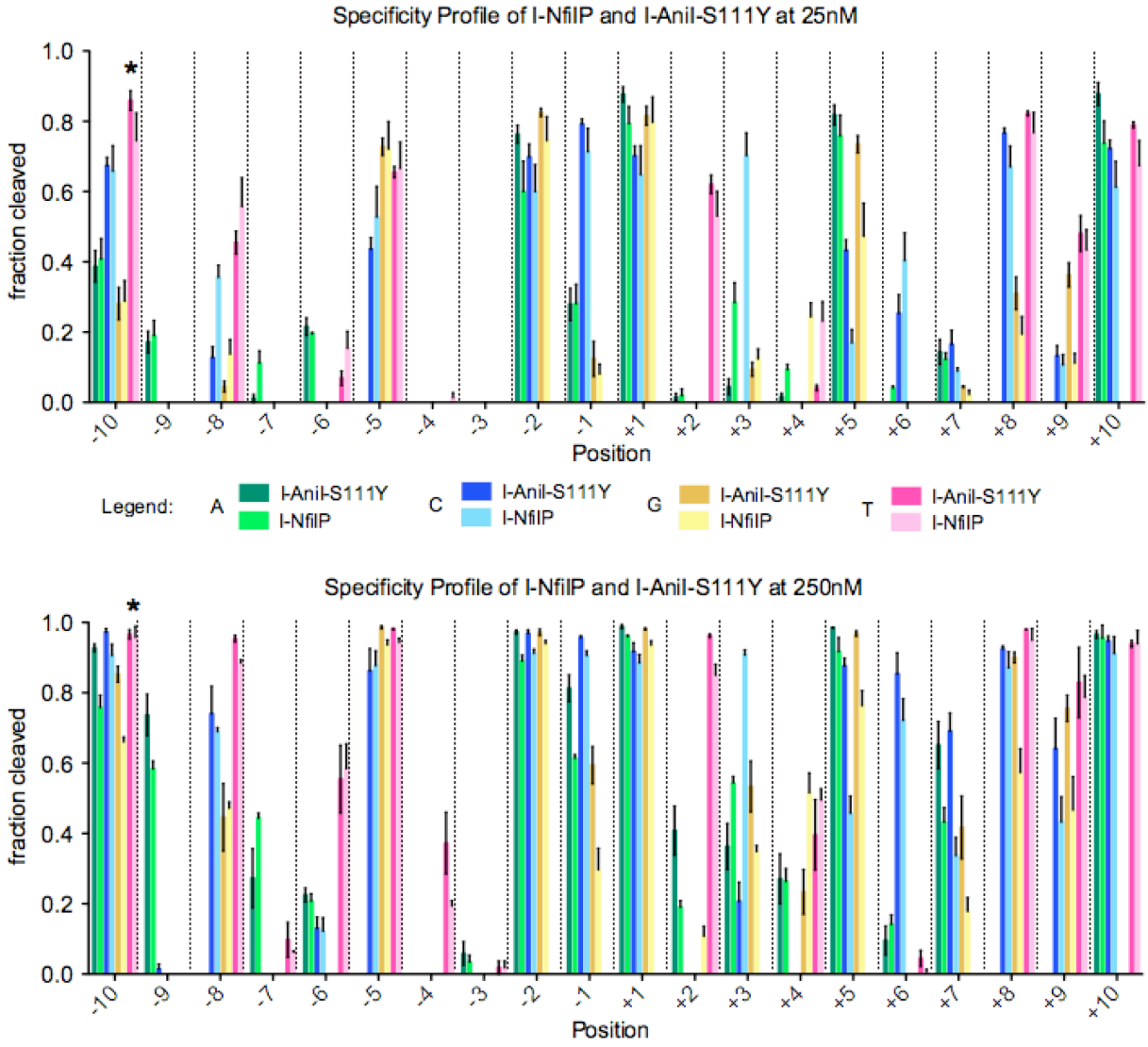
				T	35 ± 13%	0.910
			+9A	C	6 ± 1%	0.857
				G	81 ± 13%	0.864
				T	79 ± 11%	0.925
			+10G	A	139 ± 25%	~1
				C	15 ± 22%	0.928
				T	80 ± 5%	~1
K200						
I-NfiIP	I164V, K200N	F13Y	wild-type		6 ± 5%	0.981
			+3T	A	146 ± 4%	0.854
				C	14 ± 39%	0.925
				G	155 ± 8%	0.670
			+4C	A	260 ± 7%	0.565
				G	205 ± 14%	~1
				T	155 ± 32%	0.612
I-PnoIP	K200R	F13Y	wild-type		7 ± 15%	0.977
			+3T	A	140 ± 7%	~1
				C	358 ± 27%	0.929
				G	42 ± 9%	0.996
			+4C	A	118 ± 4%	~1
				G	116 ± 8%	~1
				T	18 ± 17%	0.962
			+5T	A	5 ± 28%	0.951
				C	17 ± 31%	0.950
				G	9 ± 24%	0.933
Central 4 loops						
I-GzeIP*	R172K, T196Y, N197D	S111Y	wild-type		7 ± 28%	0.919
			+2C	A	100 ± 7%	~1
				G	348 ± 7%	0.717
				T	27 ± 25%	0.954
			+3T	A	212 ± 6%	~1
				C	315 ± 6%	0.860
				G	156 ± 6%	~1
I-AchIP	K60T, R61K, N62K, E63D, inserted G after 63, I64T, E65K, M66L	S111Y	wild-type		18 ± 4%	0.956
			-2T	A	6 ± 2%	0.953
				C	40 ± 6%	0.987
				G	10 ± 16%	0.979
I-VinIP Loop	S57G, K60D, N62T , E63L, I64V, SGVVS insert after 64, E65K, A68Y, L69F, R70K, I71V	S111Y	wild-type		32 ± 9%	0.962
			-2T	A	26 ± 5%	0.927
				C	21 ± 24%	0.929
				G	23 ± 28%	0.932
Core mutations						
I-VinIP w/o core	A68Y, R70K		wild-type		>750	N/A
			-6G	A	>750	N/A
				C	>750	N/A
				T	>750	N/A
			-5A	C	>750	N/A
				G	>750	N/A
				T	>750	N/A
I-VinIP w/ core	A68Y, L69F, R70K, I71V		wild-type		428 ± 2%	~1
			-6G	A	>750	N/A

				C	>750	N/A
				T	>750	N/A
			-5A	C	>750	N/A
				G	721 ± 5%	~1
				T	>750	N/A
I-VinIP-S111Y w/o core	A68Y, R70K	S111Y	wild-type		68 ± 16%	~1
			-6G	A	182 ± 45%	0.229
				C	>750	N/A
				T	245 ± 15%	~1
			-5A	C	523 ± 3%	~1
				G	111 ± 23%	~1
				T	424 ± 4%	0.955
I-VinIP-S111Y w/ core	A68Y, L69F, R70K, I71V	S111Y	wild-type		31 ± 34%	~1
			-6G	A	176 ± 24%	0.310
				C	>750	N/A
				T	135 ± 13%	0.983
			-5A	C	362 ± 0%	~1
				G	89 ± 10%	0.996
				T	404 ± 1%	~1
I-VinIP Loop	S57G, K60D, N62T , E63L, I64V, SGVVS insert after 64, E65K, A68Y, L69F, R70K, I71V	S111Y	wild-type		32 ± 9%	0.962
			-6G	A	208 ± 22%	0.995
				C	112 ± 19%	0.996
				T	28 ± 13%	0.994
			-5A	C	67 ± 1%	0.989
				G	31 ± 8%	0.957
				T	38 ± 15%	0.969
K24N/L28V/T29K	K24N, L28V, T29K		wild-type		>750	N/A
			-8A	C	>750	N/A
				G	23 ± 15%	0.939
				T	750 ± 9%	~1
K24N/T29K	K24N, T29K, lacking L233K		wild-type		>750	N/A
			-8A	C	>750	N/A
				G	23 ± 8%	0.873
				T	>750	N/A
I-AniI Base Activity Comparison						
I-AniI-F13Y	F13Y	F13Y	wild-type		10 ± 31%	0.996
			+2C	A	194 ± 35%	0.982
				G	608 ± 11%	0.310
				T	17 ± 23%	0.900
			+3T	A	343 ± 26%	0.950
				C	503 ± 20%	0.869
				G	214 ± 7%	~1
			+4C	A	393 ± 13%	0.921
				G	437 ± 25%	0.898
				T	236 ± 19%	0.943
			+5T	A	11 ± 24%	0.934
				C	60 ± 35%	0.967
				G	27 ± 40%	0.987
			+7T	A	119 ± 28%	0.949
				C	140 ± 21%	~1
				G	189 ± 26%	~1

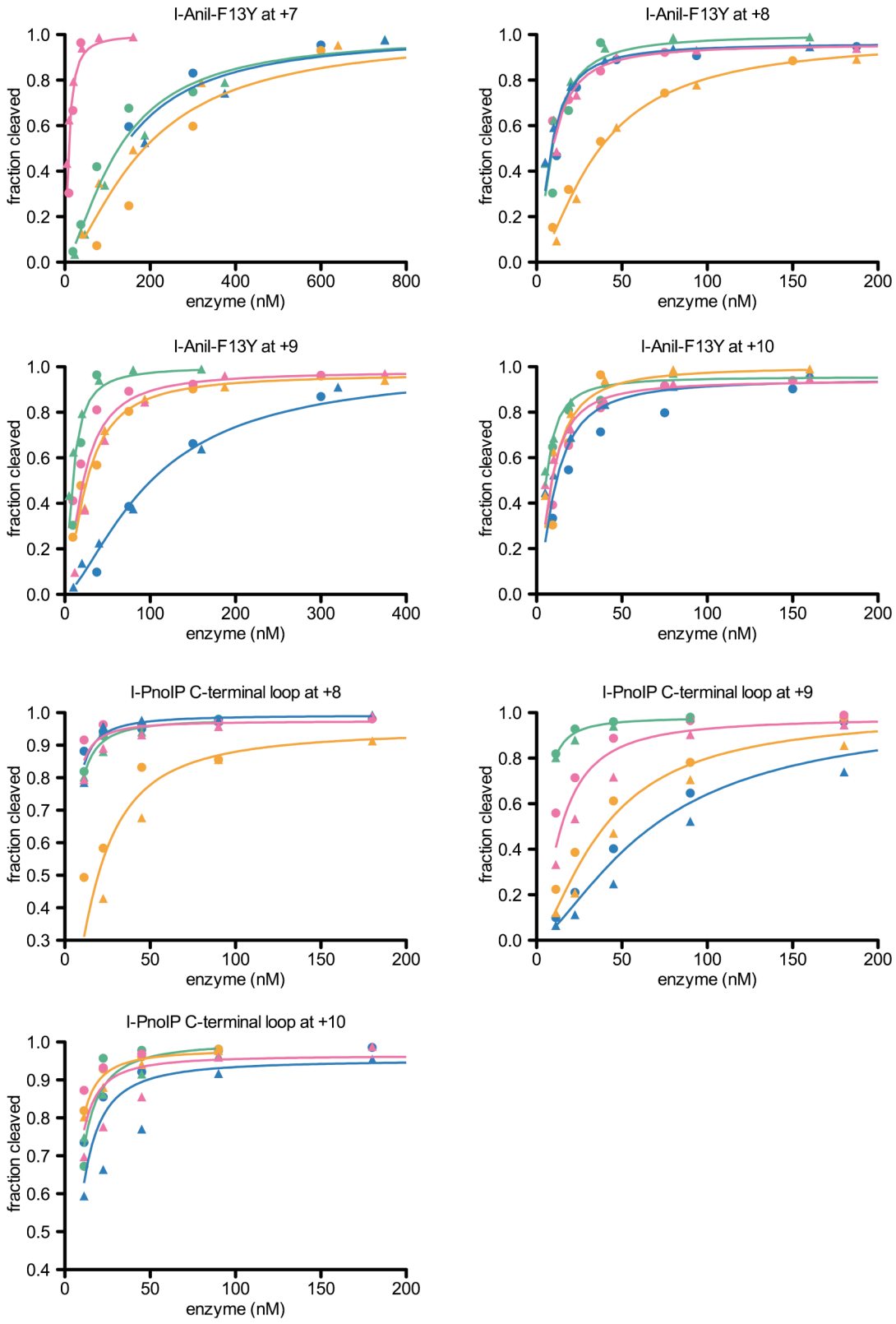
			+8A	C	9 ± 28%	0.964
				G	36 ± 11%	0.983
				T	9 ± 26%	0.956
			+9A	C	101 ± 1%	~1
				G	25 ± 10%	0.971
				T	23 ± 44%	0.981
			+10G	A	5 ± 9%	0.959
				C	12 ± 35%	0.950
				T	9 ± 36%	0.944
I-AniI-S111Y	S111Y	S111Y	wild type		9 ± 21%	0.969
			-6G	A	>750	N/A
				C	379 ± 21%	~1
				T	173 ± 5%	~1
			-5A	C	55 ± 2%	~1
				G	21 ± 21%	0.989
				T	27 ± 25%	0.985
			-2T	A	7 ± 20%	0.959
				C	12 ± 1%	0.967
				G	11 ± 3%	0.975

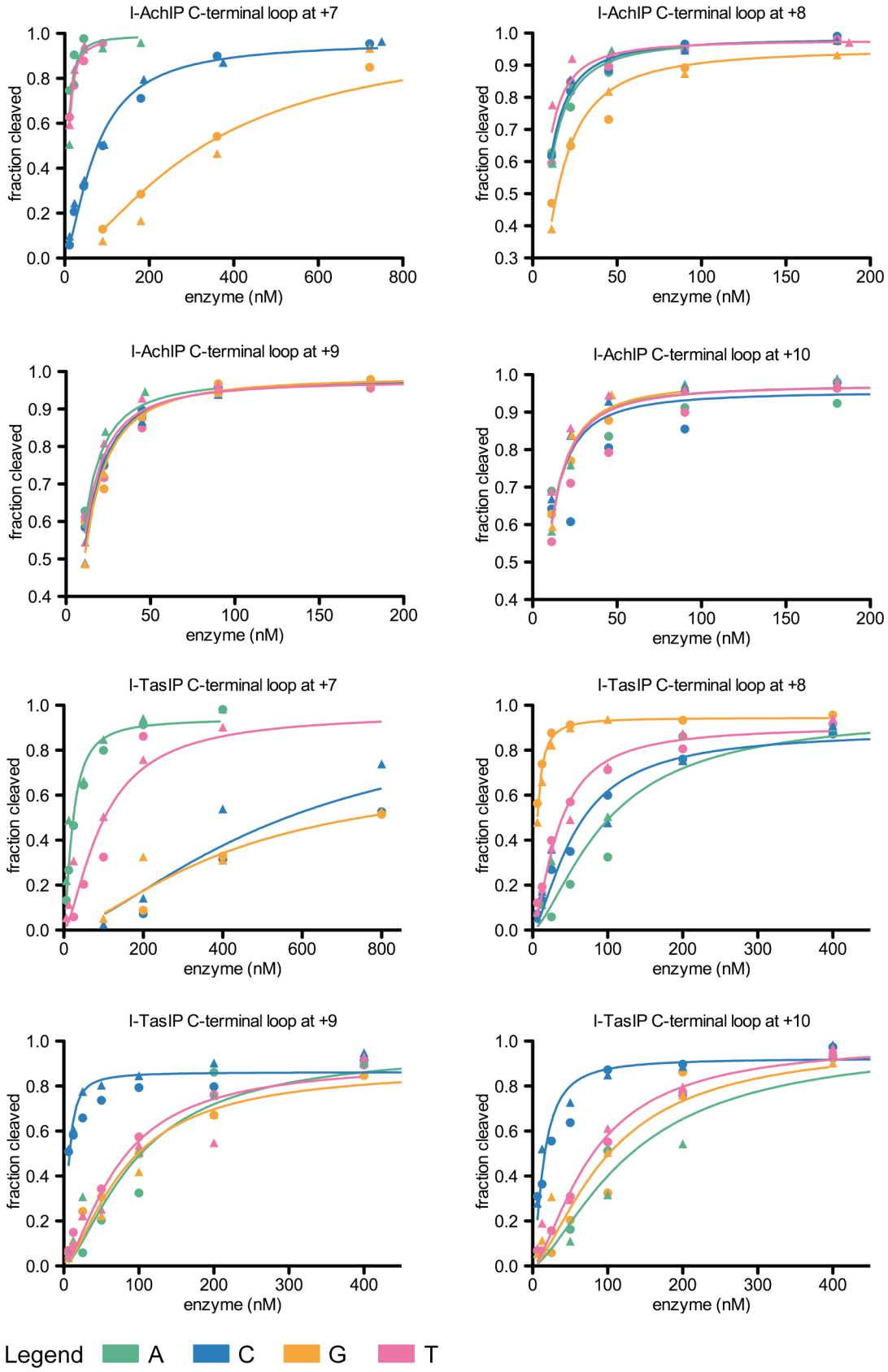
* This variant contains an additional I248V mutation that does not affect cleavage efficiency, as it is at the end of the C-terminal domain far from the interface.

Supplementary Figure S3. Specificity profile of I-NfiIP and I-AniI-S111Y at 25nM and 250nM enzyme concentration. Cleavage fractions of wild-type (denoted by asterisk) and singly-substituted target sites for all 20 positions are shown. Assays were performed in triplicate, with error bars representing the SEM. A fraction of 1 indicates complete cleavage of the substrate with no remaining uncut fraction. Since the sequence of homologue I-NfiIP contains the activating mutation S111Y, its profile is compared to that of I-AniI-S111Y. Though the specificity pattern is similar overall, the observed novel +3C specificity was later confirmed in a variant containing the I-NfiIP-derived K200N transfer (Figure 2b).

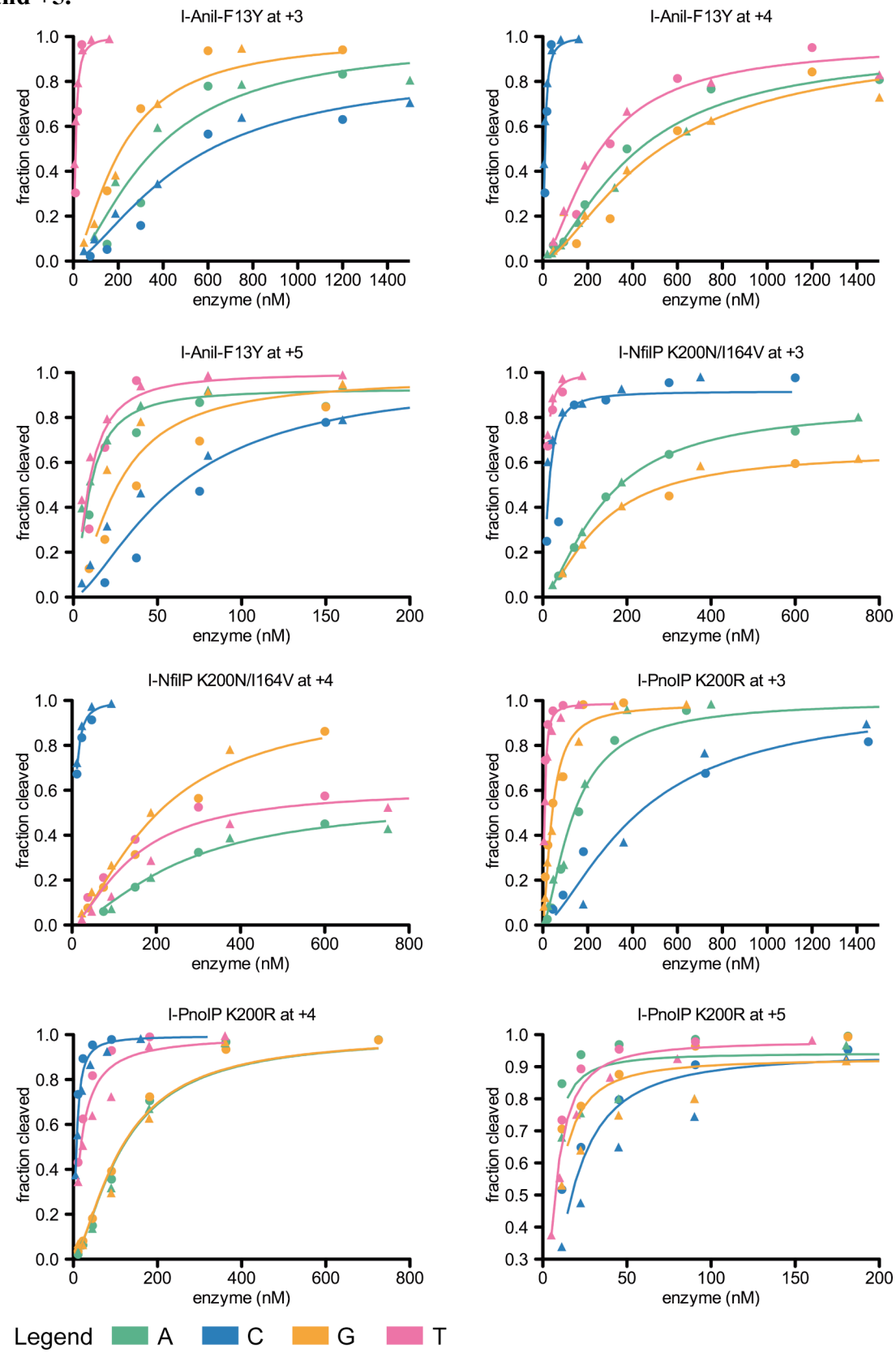


Supplementary Figure S4. Cleavage profiles for the C-terminal loop transfers and I-AniI-F13Y at positions +7 through +10.

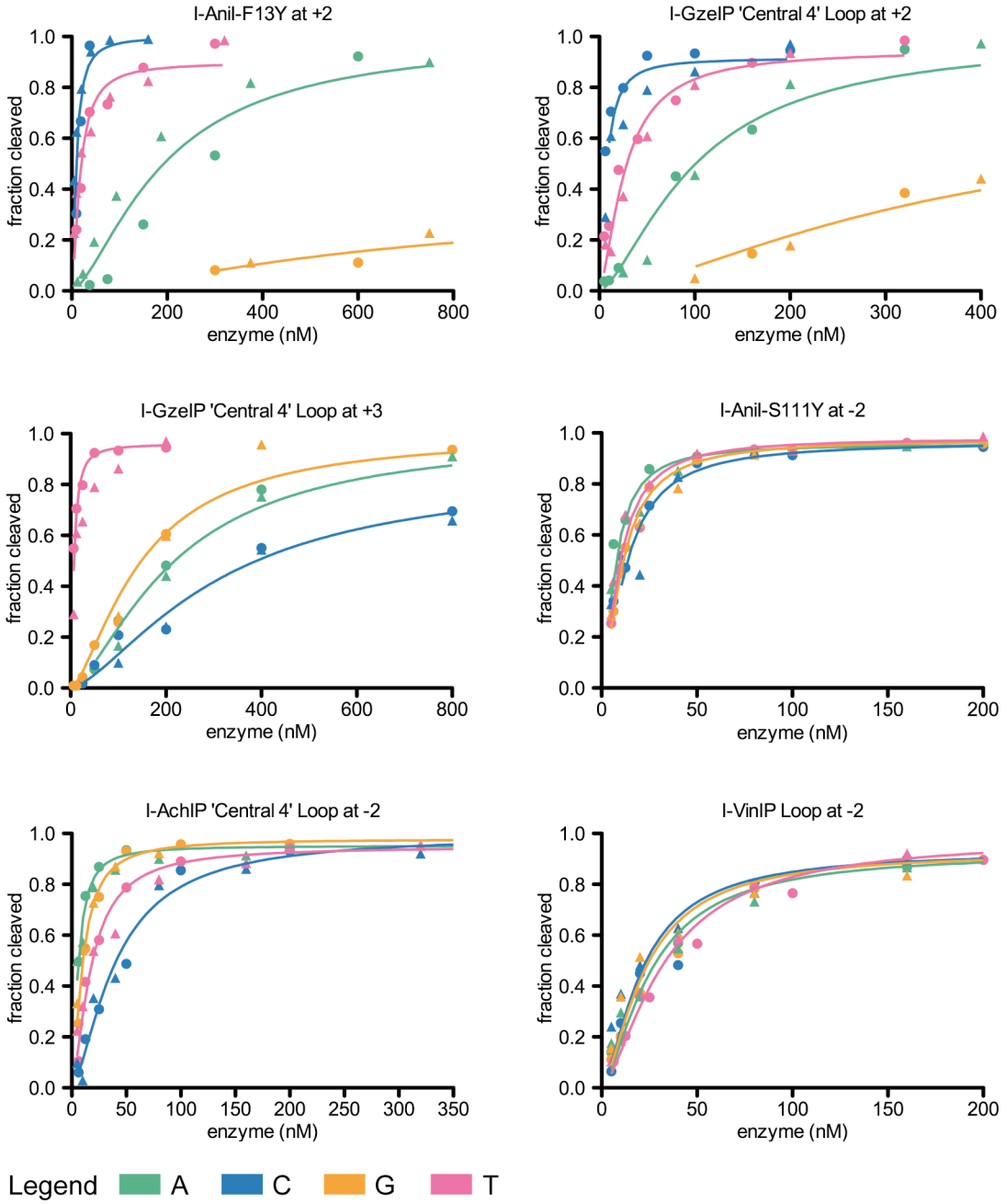



I-AchIP C-terminal loop at +7
I-AchIP C-terminal loop at +8
I-AchIP C-terminal loop at +9
I-AchIP C-terminal loop at +10
I-TasIP C-terminal loop at +7
I-TasIP C-terminal loop at +8
I-TasIP C-terminal loop at +9
I-TasIP C-terminal loop at +10
fraction cleaved
enzyme (nM)
Legend
A
C
G
T

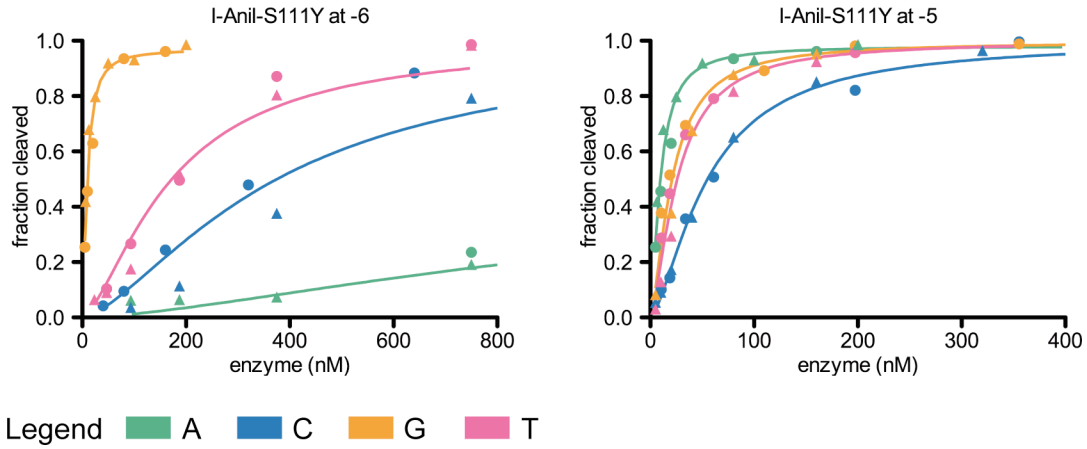
Supplementary Figure S5. Cleavage profiles for the K200 variants and I-AniI-F13Y at +3, +4, and +5.



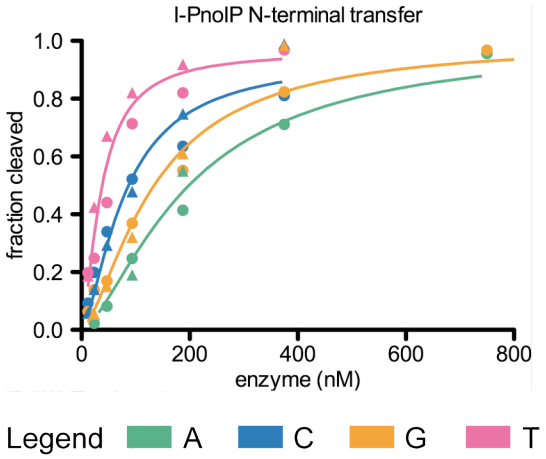
Supplementary Figure S6. Cleavage profiles of variants affecting the central four target site positions. Results for I-AniI-F13Y at position +2 and I-AniI-S111Y at -2 are provided here for comparison; the plot for I-AniI-F13Y at +3 can be seen in Supplementary Figure S5.



Supplementary Figure S7. Cleavage profiles for I-AniI-S111Y at positions -6 and -5. The corresponding $EC_{0.5max}$ (nM) were shown previously (Figure 3b) along with the cleavage plots for the I-VinIP derived variants (Figure 3a) evidencing the important role of core mutations.



Supplementary Figure S8. Cleavage profile for I-PnoIP N-terminal transfer at the -5 position. The -5T substitution is preferred over the wild-type -5A. This result is in accordance with the -5T of the predicted I-PnoIP target site given in Figure 1c.



Supplementary Table S3. Sequences of target sites near disease-causing genes that contain the novel +3C, +7A, and -6T specificities.

Feasibly targetable cleavage sites in DNA sequences of disease-causing genes were identified computationally according to previously validated methods (4). For the site sequences given below, bases in lower case lettering differ from the native I-AniI target site. Colored bases indicate the substitutions that could not be cleaved prior to this study: -6T (yellow), +3C (blue), and +7A (green).

Human CCR5, a coreceptor for HIV-1, has been successfully targeted and modified by zinc-finger nucleases (5) and is now undergoing clinical trials (6), but there is potential for further modification with homing endonucleases. Nine target site positions differ from that of I-AniI, and we found homologue-derived mutations that can be grafted onto the I-AniI scaffold to target the thymine substitution at position -6 and the adenine at +7. Two target sites found in the murine fumaryl acetoacetate hydrolase gene (FAH, mutations can lead to tyrosinemia (7)) are also shown to have sets of substitutions relative to I-AniI that contain +3C, +7A, and -6T. Lastly, recent work has addressed gene delivery in large animal models that shows great potential in preclinical studies, especially canine models. Pyruvate kinase (PK) deficiency in dogs is associated with severe hemolytic anemia that can be treated via allogeneic transplantation, and *in vivo* gene delivery through transduced cells can sufficiently correct the XSCID disease phenotype (8). Both genes contain targetable sites that can benefit from our novel specificities.

Gene	Target Site
CCR5	ccAGtAtGTTgCcCacaAAA
FAH (Target Site 1)	TGAGccctTTcCcCccTAgc
FAH (Target Site 2)	TGAGtAGcTTTCTCataAgt
PK	cGtGGccaTTgCcCTGgAcA
XSCID	gctGctGcTTctTCTGaccA

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *J. Mol. Biol.* **215**, 403-410
2. Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., Bitinaite, J., Blumenthal, R.M., Degtyarev, S.Kh., Dryden, D.T., Dybvig, K., et al. (2003) *Nucleic Acids Res.* **31**, 1805-1812
3. Longo, A., Leonard, C.W., Bassi, G.S., Berndt, D., Krahn J., Hall, T.M., and Weeks, K.M. (2005) *Nat. Struct. Mol. Biol.* **12**, 779-787
4. Ashworth, J., Taylor, G. K., Havranek, J.J., Quadri, S.A., Stoddard, B.L., and Baker, D. (2010) *Nucleic Acids Res.* **38**, 5601-5608
5. Guo, H., Karberg, M., Long, M., Jones, J.P. 3rd., Sullenger, B., and Lambowitz, A.M. (2000) *Science* **289**, 452-457
6. University of Pennsylvania; Sangamo Biosciences. Autologous T-Cells Genetically Modified at the CCR5 Gene by Zinc Finger Nucleases SB-728 for HIV (Zinc-Finger). In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000- [cited 2011 Apr 19]. Available from: http://clinicaltrials.gov/show/NCT00842634 NLM Identifier: NCT00842634
7. Aponte, J.L., Sega, G.A., Hauser, L.J., Dhar, M.S., Withrow, C.M., Carpenter, D.A., Rinchik, E.M., Culiat, C.T., and Johnson, D.K. (2001) *Proc. Natl. Acad. Sci. U.S.A.* **98**, 641-645
8. Trobridge, G.D., and Kiem, H-P. (2010) *Gene Ther.* **17**, 939–948